Kristina Jovanovska, Bsc.

Data analysis of students behavior using "Pocket Code"

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Computer Science

submitted to

Graz University of Technology

Supervisor

Denis Helic, Assoc. Prof. Dipl-Ing. Dr.techn.

Institute of Interactive Systems and Data Science

Graz, January 2019

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Acknowledgements

Being part of the master program in Computer Science at TU Graz gave me the chance to expand my knowledge in this field, to have the opportunity to learn from great professors and gave me entire new perspective on technology. I am very happy and proud to be part of this program. But the road full with challenges leading to achieving successful completion, was possible in the past 2,5 years thanks to the great support by many people, to who I want to express my gratefulness in this part.

First and foremost, I would like to express my very great appreciation to my supervisor and mentor Prof. Dipl-Ing. Dr.techn. Denis Helic, for his constant support, guidance and suggestions for successful research and development of this thesis. The meetings were very helpful and contributed greatly to this master thesis.

Secondly, I would like to express my gratitude to Dr.techn. Bernadette Spieler, for providing me with the dataset as well as for providing me with information and clarification that answered many challenges that I came across during the analysis of the data.

Thirdly, a special word of thanks to my parents, who unconditionally supported me in my idea to continue my education abroad and believed in me that I will successfully graduate this master program from the very beginning of it, till the end of the preparation of this thesis. Words can not describe my gratitude to you.

I am particularly grateful to Simon for all his patience, unreserved support during the entire time of preparing for my exams and preparing this thesis.

I would like to thank to my aunt Aleksandra who entirely supported me during my studies. Also thanks to my friends Aleksandar and Tanja with who we shared sometimes stressful, but also many great student days, and Maria for all the support. Having you all helped me always to stay motivated and encouraged me to accept new challenges.

Finally, a word of thanks to the ones not mentioned in this part who indirectly also supported me.

Kristina Jovanovska
January 28st, 2019

Contents

1 Introduction **1**
- 1.1 Abstract . . . 1
- 1.2 Motivation . . . 1
- 1.3 General Overview . . . 2
- 1.4 Description of the project, project goals, and project challenges 2
  - 1.4.1 Pocket Code . . . 2
  - 1.4.2 Available data . . . 3
  - 1.4.3 Challenges . . . 5

2 Related work **7**
- 2.1 Cleaning and tidying the data . . . 7
  - 2.1.1 "Tidy data" . . . 7
  - 2.1.2 Data Cleaning . . . 17
- 2.2 Finding patterns in the data . . . 18
  - 2.2.1 Sequential pattern mining . . . 18
  - 2.2.2 PrefixSpan . . . 22
- 2.3 Classification and defining significance . . . 24
  - 2.3.1 Gender different behaviour in programming . . . 24
  - 2.3.2 Creating games as a way of motivating girls for IT studies and programming . . . 31

3 Implementation and results **34**
- 3.1 Cleaning and tidying the NOLB data . . . 34
  - 3.1.1 Data received for analysis . . . 34
  - 3.1.2 Tidying the data . . . 36
- 3.2 Finding patterns in the received data . . . 37
  - 3.2.1 Similarity and significance . . . 37
  - 3.2.2 Pattern Mining with PrefixSpan . . . 58
- 3.3 Classification and definition of the difference in the patterns found . . . 88

4 Future work and conclusion **108**
4.1 Conclusion 108
4.2 Future Work 109

A PrefixSpan **111**

List of Figures

1.1 Lego-style building blocks principle of Pocket Code 3
1.2 Pocket code screenshots 4

2.1 Presenting variable values as column 8
2.2 Presenting variable values as rows 9
2.3 9
2.4 This is the original data representation 10
2.5 Molted representation on the same data 11
2.6 Original data from World Health Organization 11
2.7 Molted representation on the same data 12
2.8 Tidied data: the column split into gender and age 12
2.9 Original data from Global Historical Climatology Network . . 13
2.10 Molted representation on the same data 13
2.11 Tidied the original data 14
2.12 Original partial data from billboard dataset 15
2.13 First dataset: *Song* 15
2.14 Second dataset: *Rank* 16
2.15 Sequence database S 21
2.16 Algorithm PrefixSpan on the example sequence database S . . 23
2.17 Way of mapping sentences to predefined categories 26
2.18 26 card types 27
2.19 Predefined concepts and results from both male and female students 28
2.20 Mastering concepts given for both genders 29
2.21 Programming experience before collage and average results in collage 30

3.1 List of events from user "naakf001" 37
3.2 All users 38
3.3 All female users 38
3.4 All male users 38

3.5 Female users from every country 42
3.6 Male users from every country 42
3.7 Female users from every school 42
3.8 Male users from every school 42
3.9 Female users from every country with Standard Deviation . . 43
3.10 Male users from every country with Standard Deviation 43
3.11 Female users from every country 43
3.12 Male users from every country 43
3.13 Female users from every school 44
3.14 Male users from every school 44
3.15 Female users from every country with Standard Deviation . . 44
3.16 Male users from every country with Standard Deviation 44
3.17 Distribution per user for all students 45
3.18 Distribution per user for Austrian students 45
3.19 Distribution per user for Spanish students 46
3.20 Distribution per user for British students 46
3.21 Distribution per user for all female students 46
3.22 Distribution per user for Austrian female students 46
3.23 Distribution per user for Spanish female students 46
3.24 Distribution per user for British female students 46
3.25 Distribution per user for all male students 47
3.26 Distribution per user for Austrian male students 47
3.27 Distribution per user for Spanish male students 47
3.28 Distribution per user for British male students 47
3.29 Distribution per session for all students 47
3.30 Distribution per session for Austrian students 47
3.31 Distribution per session for Spanish students 48
3.32 Distribution per session for British students 48
3.33 Distribution per session for all female students 48
3.34 Distribution per session for Austrian female students 48
3.35 Distribution per session for Spanish female students 48
3.36 Distribution per session for British female students 48
3.37 Distribution per session for all male students 49
3.38 Distribution per session for Austrian male students 49
3.39 Distribution per session for Spanish male students 49
3.40 Distribution per session for British male students 49
3.41 Comparing the users between countries 50
3.42 Comparing genders between countries 50
3.43 Comparing the users between schools 51
3.44 Comparing the users between genders between schools 52
3.45 Comparing the users between countries per user 57

3.46 Comparing genders between countries per user 57
3.47 Comparing the users between schools per user 58
3.48 Comparing the users between genders between schools per user 59
3.49 Sequences of length 1 in a session for all users with PrefixSpan 61
3.50 Sequences of length 2 in a session for all users with PrefixSpan 61
3.51 Sequences of length 3 in a session for all users with PrefixSpan 62
3.52 Sequences of length 4 in a session for all users with PrefixSpan 62
3.53 Sequences of length 5 in a session for all users with PrefixSpan 62
3.54 Sequences of length 1 in a session for all female users with PrefixSpan 63
3.55 Sequences of length 5 in a session for all female users with PrefixSpan 63
3.56 Sequences of length 1 in a session for all male users with PrefixSpan 64
3.57 Sequences of length 5 in a session for all male users with PrefixSpan 64
3.58 Sequences of length 1 in a session for all Austrian users with PrefixSpan 65
3.59 Sequences of length 5 in a session for all Austrian users with PrefixSpan 65
3.60 Sequences of length 1 in a session for all Austrian female users with PrefixSpan 66
3.61 Sequences of length 5 in a session for all Austrian female users with PrefixSpan 66
3.62 Sequences of length 1 in a session for all Austrian male users with PrefixSpan 67
3.63 Sequences of length 5 in a session for all Austrian male users with PrefixSpan 67
3.64 Sequences of length 1 in a session for all Spanish users with PrefixSpan 68
3.65 Sequences of length 5 in a session for all Spanish users with PrefixSpan 68
3.66 Sequences of length 1 in a session for all Spanish female users with PrefixSpan 69
3.67 Sequences of length 5 in a session for all Spanish female users with PrefixSpan 69
3.68 Sequences of length 1 in a session for all Spanish male users with PrefixSpan 70
3.69 Sequences of length 5 in a session for all Spanish male users with PrefixSpan 71

3.70 Sequences of length 1 in a session for all British users with PrefixSpan 71
3.71 Sequences of length 5 in a session for all British users with PrefixSpan 72
3.72 Sequences of length 1 in a session for all British female users with PrefixSpan 72
3.73 Sequences of length 5 in a session for all British female users with PrefixSpan 73
3.74 Sequences of length 1 in a session for all British male users with PrefixSpan 73
3.75 Sequences of length 5 in a session for all British male users with PrefixSpan 74
3.76 Sequences of length 1 in a session for all users from GIBS with PrefixSpan 75
3.77 Sequences of length 5 in a session for all users from GIBS with PrefixSpan 75
3.78 Sequences of length 1 in a session for all female users from GIBS with PrefixSpan 76
3.79 Sequences of length 5 in a session for all female users from GIBS with PrefixSpan 76
3.80 Sequences of length 1 in a session for all male users from GIBS with PrefixSpan 77
3.81 Sequences of length 5 in a session for all male users from GIBS with PrefixSpan 77
3.82 Sequences of length 1 in a session for all users from Borg Birkfeld with PrefixSpan 78
3.83 Sequences of length 5 in a session for all users from Borg Birkfeld with PrefixSpan 78
3.84 Sequences of length 1 in a session for all female users from Borg Birkfeld with PrefixSpan 78
3.85 Sequences of length 3 in a session for all female users from Borg Birkfeld with PrefixSpan 79
3.86 Sequences of length 1 in a session for all male users from Borg Birkfeld with PrefixSpan 79
3.87 Sequences of length 5 in a session for all male users from Borg Birkfeld with PrefixSpan 79
3.88 Sequences of length 1 in a session for all users from Akademisches Gymnasium with PrefixSpan 81
3.89 Sequences of length 5 in a session for all users from Akademisches Gymnasium with PrefixSpan 81

3.90 Sequences of length 1 in a session for all female users from Akademisches Gymnasium with PrefixSpan 81
3.91 Sequences of length 3 in a session for all female users from Akademisches Gymnasium with PrefixSpan 82
3.92 Sequences of length 1 in a session for all male users from Akademisches Gymnasium with PrefixSpan 82
3.93 Sequences of length 5 in a session for all male users from Akademisches Gymnasium with PrefixSpan 82
3.94 Sequences of length 1 in a session for all users from Oakfield with PrefixSpan 83
3.95 Sequences of length 5 in a session for all users from Oakfield with PrefixSpan 83
3.96 Sequences of length 1 in a session for all female users from Oakfield with PrefixSpan 83
3.97 Sequences of length 3 in a session for all female users from Oakfield with PrefixSpan 84
3.98 Sequences of length 1 in a session for all male users from Oakfield with PrefixSpan 84
3.99 Sequences of length 5 in a session for all male users from Oakfield with PrefixSpan 84
3.100Sequences of length 1 in a session for all users from St. Stephens Primary with PrefixSpan 85
3.101Sequences of length 5 in a session for all users from St. Stephens Primary with PrefixSpan 85
3.102Sequences of length 1 in a session for all female users from St. Stephens Primary with PrefixSpan 85
3.103Sequences of length 3 in a session for all female users from St. Stephens Primary with PrefixSpan 86
3.104Sequences of length 1 in a session for all male users from St. Stephens Primary with PrefixSpan 86
3.105Sequences of length 5 in a session for all male users from St. Stephens Primary with PrefixSpan 86
3.106Sequences of length 1 in a session for all users from Christ the King with PrefixSpan 87
3.107Sequences of length 5 in a session for all users from Christ the King with PrefixSpan 87
3.108Sequences of length 1 in a session for all female users from Christ the King with PrefixSpan 87
3.109Sequences of length 1 in a session for all male users from Christ the King with PrefixSpan 88

3.110Sequences of length 1 in a session for all users from St.Peters school with PrefixSpan 89
3.111Sequences of length 5 in a session for all users from St.Peters school with PrefixSpan 89
3.112Sequences of length 1 in a session for all female users from St.Peters school with PrefixSpan 89
3.113Sequences of length 3 in a session for all female users from St.Peters school with PrefixSpan 90
3.114Sequences of length 1 in a session for all male users from St.Peters school with PrefixSpan 90
3.115Sequences of length 5 in a session for all male users from St.Peters school with PrefixSpan 90
3.116Comparing the sequences of length 1 between countries 94
3.117Comparing the sequences of length 2 between countries 94
3.118Comparing the sequences of length 3 between countries 94
3.119Comparing the sequences of length 4 between countries 94
3.120Comparing the sequences of length 5 between countries 95
3.121Comparing sequences of length 1 between genders and countries 95
3.122Comparing sequences of length 2 between genders and countries 95
3.123Comparing sequences of length 3 between genders and countries 95
3.124Comparing sequences of length 4 between genders and countries 96
3.125Comparing sequences of length 5 between genders and countries 96
3.126Comparing the sequences of length 1 between schools 96
3.127Comparing the sequences of length 2 between schools 96
3.128Comparing the sequences of length 3 between schools 97
3.129Comparing the sequences of length 4 between schools 97
3.130Comparing the sequences of length 5 between schools 97
3.131Comparing sequences of length 1 between genders and schools 97
3.132Comparing sequences of length 2 between genders and schools 98
3.133Comparing sequences of length 3 between genders and schools 98
3.134Comparing sequences of length 4 between genders and schools 98
3.135Comparing sequences of length 5 between genders and schools 98
3.136Most common sequences of length 1 between each groups and the one for all users 99
3.137Most common sequences of length 2 between each groups and the one for all users 100
3.138Most common sequences of length 3 between each groups and the one for all users 101
3.139Most common sequences of length 4 between each groups and the one for all users 102

3.140Most common sequences of length 5 between each groups and the one for all users 103
3.141The results of Spearman between countries 106
3.142The results of Spearman between genders between countries . 106
3.143The results of Spearman between schools 106
3.144The results of Spearman between genders between schools . . 106
3.145Spearman comparison for all groups 107

A.1 Sequences of length 2 in a session for all users from ak with PrefixSpan 112
A.2 Sequences of length 3 in a session for all users with PrefixSpan 112
A.3 Sequences of length 4 in a session for all users from ak with PrefixSpan 112
A.4 Sequences of length 2 in a session for all female users from ak with PrefixSpan 113
A.5 Sequences of length 3 in a session for all female users with PrefixSpan 113
A.6 Sequences of length 4 in a session for all female users with PrefixSpan 113
A.7 Sequences of length 2 in a session for all male users from ak with PrefixSpan 114
A.8 Sequences of length 3 in a session for all male users from ak with PrefixSpan 114
A.9 Sequences of length 4 in a session for all male users from ak with PrefixSpan 114
A.10 Sequences of length 2 in a session for all users from GIBS with PrefixSpan 115
A.11 Sequences of length 3 in a session for all users from GIBS with PrefixSpan 115
A.12 Sequences of length 4 in a session for all users from GIBS with PrefixSpan 115
A.13 Sequences of length 2 in a session for all female users from GIBS with PrefixSpan 116
A.14 Sequences of length 3 in a session for all female users from GIBS with PrefixSpan 116
A.15 Sequences of length 4 in a session for all female users from GIBS with PrefixSpan 116
A.16 Sequences of length 2 in a session for all male users from GIBS with PrefixSpan 117
A.17 Sequences of length 3 in a session for all male users from GIBS with PrefixSpan 117

A.18 Sequences of length 4 in a session for all male users from GIBS with PrefixSpan . . . 117
A.19 Sequences of length 2 in a session for all users from bb with PrefixSpan . . . 118
A.20 Sequences of length 3 in a session for all users from bb with PrefixSpan . . . 118
A.21 Sequences of length 4 in a session for all users from bb with PrefixSpan . . . 118
A.22 Sequences of length 2 in a session for all female users from bb with PrefixSpan . . . 119
A.23 Sequences of length 2 in a session for all male users from bb with PrefixSpan . . . 119
A.24 Sequences of length 3 in a session for all male users from bb with PrefixSpan . . . 119
A.25 Sequences of length 4 in a session for all male users from bb with PrefixSpan . . . 120
A.26 Sequences of length 2 in a session for all users from ub with PrefixSpan . . . 120
A.27 Sequences of length 3 in a session for all users from ub with PrefixSpan . . . 120
A.28 Sequences of length 4 in a session for all users from ub with PrefixSpan . . . 121
A.29 Sequences of length 2 in a session for all female users from ub with PrefixSpan . . . 121
A.30 Sequences of length 3 in a session for all female users from ub with PrefixSpan . . . 121
A.31 Sequences of length 4 in a session for all female users from ub with PrefixSpan . . . 122
A.32 Sequences of length 2 in a session for all male users from ub with PrefixSpan . . . 122
A.33 Sequences of length 3 in a session for all male users from ub with PrefixSpan . . . 122
A.34 Sequences of length 4 in a session for all male users from ub with PrefixSpan . . . 123
A.35 Sequences of length 2 in a session for all users from of with PrefixSpan . . . 123
A.36 Sequences of length 3 in a session for all users from of with PrefixSpan . . . 123
A.37 Sequences of length 4 in a session for all users from of with PrefixSpan . . . 124

A.38 Sequences of length 2 in a session for all female users from of with PrefixSpan 124
A.39 Sequences of length 3 in a session for all female users from of with PrefixSpan 124
A.40 Sequences of length 4 in a session for all female users from of with PrefixSpan 125
A.41 Sequences of length 2 in a session for all male users from of with PrefixSpan 125
A.42 Sequences of length 3 in a session for all male users from of with PrefixSpan 125
A.43 Sequences of length 4 in a session for all male users from of with PrefixSpan 126
A.44 Sequences of length 2 in a session for all users from ss with PrefixSpan 126
A.45 Sequences of length 3 in a session for all users from ss with PrefixSpan 126
A.46 Sequences of length 4 in a session for all users from ss with PrefixSpan 127
A.47 Sequences of length 2 in a session for all female users from ss with PrefixSpan 127
A.48 Sequences of length 3 in a session for all female users from ss with PrefixSpan 127
A.49 Sequences of length 4 in a session for all female users from ss with PrefixSpan 128
A.50 Sequences of length 2 in a session for all male users from ss with PrefixSpan 129
A.51 Sequences of length 3 in a session for all male users from ss with PrefixSpan 129
A.52 Sequences of length 4 in a session for all male users from ss with PrefixSpan 129
A.53 Sequences of length 2 in a session for all male users from sp with PrefixSpan 130

List of Tables

1.1 Students in respect to country and school 4

3.1 Data format 2016 35
3.2 Data format 2017 36
3.3 Number of events in the first and last session and average over all sessions from students with more than one session 39
3.4 Average number of events in the first and last session and average over all sessions for Austrian students 40
3.5 Average number of events in the first and last session and average over all sessions for Spanish students 40
3.6 Average number of events in the first and last session and average over all sessions for British students 40
3.7 Average events per user per country 41
3.8 Average events per user per school 42
3.9 Average events per session per country 43
3.10 Average events per session per school 44
3.11 Students who used one of the help events/buttons: ("openBrickHelpOption" or "hintOption") 91
3.12 Students who used one of the copy events: (including: "CopyProgram", "CopyObject", "CopyLook", "CopyLook", "CopySound", "CopyBrick") 91
3.13 Students who used one of the advanced events: (including: "backpackObject", "unpackObject", "backpackLook", "unpackLook", "backpackSound", "unpackSound", "backpackScripts", "unpackScripts", "backpackScenes", "unpackScenes") 92

Abbreviations used

NOLB - No One Left Behind (the name of the project)

IDE - Integrated Development Environment

GSP - Generalized Sequence Pattern

FP - Frequent Pattern

CS - Computer Science

AT - Austria

ES - Spain UK - The United Kingdom

ak - Akademisches Gymnasium

gi - Graz International Bilingual School

GIBS - Graz International Bilingual School

bb - Borg Birkfeld

ub - Ubeda

of - Oakfield School

ss - St. Stephan Primary

ctk - Christ the King

sp - St. Peters

Chapter 1

Introduction

1.1 Abstract

This master thesis deals mainly with data analysis. The data on which the analysis were carried out was obtained from the mobile application "Pocket Code" and was gathered from classes with children in different schools, in different countries, as part of the NOLB project. The first goal was tidying the data and creating uniform data representation from the two different data structures. Tidying the dataset made it cleaner and better prepared for analysis. In the second step, many statistical analysis were done and presented like average number of events in respect to user and session. The significance between the differences found with the statistical tests were checked with Mann Whitney U-Test. The sequences of events used by users in different groups were analyzed in between countries/schools/genders in the data with the use of PrefixSpan. For checking the significance of the results obtained with PrefixSpan 2 different tests were implemented. With the first test the overlapping was checked between the different groups and with Spearman was tested the correlation between the countries/schools/genders. Finally, the results were presented, the differences and similarities between the groups.

1.2 Motivation

The motivation for analyzing the dataset which was collected during the NOLB project, was the potential for getting more and new results.

The way the countries analyzed the data is influenced by their motivation to join the project and the results they wanted to achieve. For example, the Austrian schools had the goal of stimulating and promoting technical fields of study to girls, while the Spanish schools wanted to create a medium for

internationally diverse classes to create things together. The British schools focused more on applying this project in classes with students with cognitive disabilities and monitoring the effect they had on their learning abilities.

The experiment has a unique international aspect, where the data gathering is done in the same way. This is an opportunity for interesting comparisons and new insights which might have been overlooked when just focusing on the goals of one of the participating counties.

1.3 General Overview

In the first section the project is explained. A general description is given for the data received, the immediately known challenges are presented, as well as the motivation for working on this topic. In the second chapter, there is a description of methods and results of different relevant literature reviews and short explanations and reasons why certain ideas are relevant to be implemented and can be used for the data analysis made for this thesis. Building on the ideas from chapter 2, the results of the implementation and their analysis are presented in chapter 3. Finally, in the last - fourth chapter, the conclusion is given as well as the possibilities for future work on this topic.

1.4 Description of the project, project goals, and project challenges

This master thesis is mainly focused on data analysis. The data analyzed is obtained from the mobile application "Pocket Code" made by Slany (2014). The application was used by children in several schools spread over Austria, Spain and the United Kingdom.

1.4.1 Pocket Code

Pocket code is a mobile application created by Slany (2014). The main idea of the application is to give children a first experience with programming and allow them to intuitively create and share their own mobile applications. It provides users with a visual "gamified" integrated development environment (IDE). In that IDE the users can use a visual programming language where elements of the language are represented by bricks which can be chained together. The bricks have prewritten key coding words so called "events" and putting them together makes it possible to create real useful applications or games.

The main goal is not to train the the students' programming skills, but rather they to get the chance to see that creative use of digital tools is essential for the future society. With that the focus is put on learning logical and structural thinking, as well as understanding of daily used technologies. With the principle of Lego-like building bricks, typical traditional programming language problems are eliminated. This is done by hiding the more complex underlying syntax of programming languages, such as conditions, variables, events or parallelism. The use of a graphical and intuitive way of programming leaves space for the children to be able completely to focus on their own creative ideas.

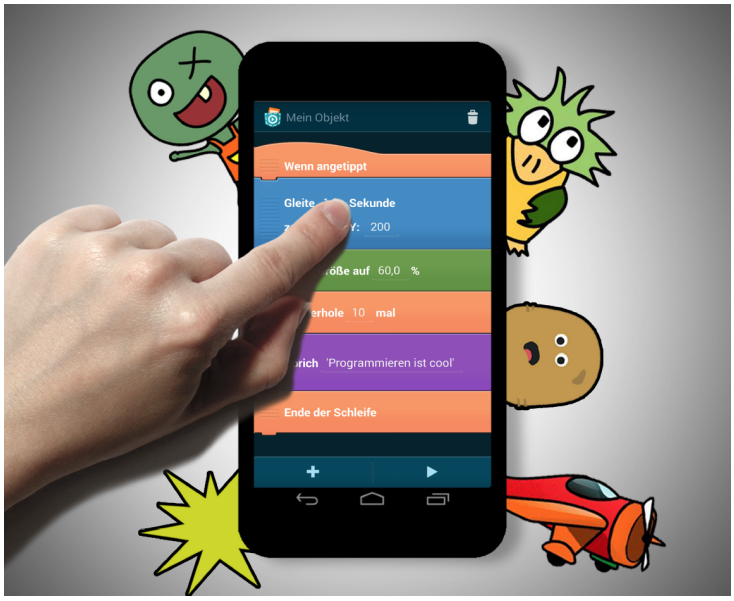

Figure 1.1: Lego-style building blocks principle of Pocket Code

Source: Slany (2014)

1.4.2 Available data

The data on which analysis has to be carried out was gathered from classes of children in different schools over different countries. Table 1.1 gives an overview of the schools and countries.

For each school, one class of children was introduced to the mobile application. During the usage of the application, key events or actions were recorded. So for each user, a set of event-timestamp pairs is available. The events that were tracked can be divided in a few groups. The group session

Figure 1.2: Pocket code screenshots

Source: Slany (2014)

Country	School	Male	Female	Total
AT		**60**	**71**	**131**
	GIBS	23	41	64
	Borg Birkfeld	23	6	29
	Akademische Gymnasium	14	24	38
ES		**59**	**26**	**85**
	Santa Maria	0	0	0
	Ubeda	59	26	85
UK		**78**	**61**	**139**
	Oakfield	10	7	17
	St. Stephens Primary	51	38	89
	Christ the King	4	3	7
	St. Peters school	13	13	26
Total students				**355**

Table 1.1: Students in respect to country and school

contains two events: “init_session” and “end_session”. The ”“init_session” event is created when the user logs in. A challenge came from the fact that users did not always remember to log out, which would create the “end_session”. Having these two events would make it easier to always be able to distinguish between different sessions. This challenge is handled by assuming that events that happen in a new day can not be part of the same session, because it is a new schoolday for students. The next group of events is “create_events” and it contains 11 different events that the students can choose from: ”createProgram”, ”createObject”, “createLook”, “editLook”, “createSound”, “addBrick”, “createList”, “createVariable”, “addScene”, and “createGroup”. The next group of events is “brick_list”, where the events are divided in two subgroups, or more precisely in simple and advanced bricks. Both subgroups contain the same events, being “Event”, “Control”, “Motion”, “Sound”, “Look”, and “Data”. The only difference between the two groups is that the advanced group contains the event ”Pen”. Next group of events is ”delete_events”, where it is possible to delete all the events that can be created with the events defined in the group ”createEvents”. The next group of events is ”copy_events”, which gives the possibility to the users to reuse already written code. The events in this group are: “copyProgram”, “copyObject”, “copyLook”, “copySound”, “copyBrick”. The group ”main_menu_events” group, where the buttons from the main menu are offered, such as: “openProgram”, “mainmenuExploreButton”, “stopExploreSession”, “mainMenuHelpButton”, “stopWebtutorialSession”. The group “other_usecases” contains the following events: “downloadProgram”, “uploadProgram”, “useTemplate”, “applyAccessabilitySetting”, “OpenBrickHelpButton”, “mergePrograms”, “OpenFormulaEditor”, “saveFormula”, “createBroadcaseMessage”, “hintsOption”. Finally, the last group of events is ”backpack_events”, which are hidden events. Users that use these events are considered to be more advanced.

The data was collected over a period of two years from 2016 to 2017. During this period some parameters were continuously tracked, but others were added or removed during the course of time.

1.4.3 Challenges

There are several main challenges that arise when analyzing this data. The first one is to find a suitable way of analyzing the data so that meaningful conclusions can be derived. The second challenge is to handle missing data. The data was recorded for a limited number of months, where one complete month of data is missing. Additionally, one out of nine schools did not track the data in a valid format. All of the three countries have a different amount

of schools participating in the project, each having a different number of students participating. This often makes it very difficult to compare and make conclusions since there are many factors that constantly have to be taken into account. Finally, the experimental setup of the data-gathering process has a very important role and influence on the results, which brings more challenges to the analysis of this dataset. The teachers had the freedom to explain to the children in the way they thought is appropriate for their students. The positive side of this approach is that the teachers adjust their presentation of the experiment to their students. The negative side is that students from different schools received a different introduction before interacting with the application. There are suggestions on the web site for "Pocket Code" and guidelines for teachers, but it is not known how much the teachers were following these suggestions.

For the schools in Austria, more information is available on how much the teachers knew about the application beforehand, how they presented it to the children and whether or not the children used the application before. Unfortunately, there is no information for the introduction classes in the schools from Spain and the United Kingdom, which prevents to get the real picture, and makes it more challenging to make conclusions after getting the results in chapter 3. Goals deal with missing data, tidy data, processing the tidy data and finding points for further deeper analysis and analyzing the data and finding beneficial outcomes.

Chapter 2

Related work

This chapter provides an overview literature related, it contains 3 sub chapters. The first one is related to cleaning data, the need to clean the data and methods for doing that. In the second sub chapter are given relevant ways of preparing the data for analysis and finding patterns. In the last sub chapter, there is relevant literature for defining and classifying of the patterns.

2.1 Cleaning and tidying the data

2.1.1 "Tidy data"

Most raw gathered datasets are not very "clean". They contain missing data, and need a lot of preprocessing in order to be usable for an application. In contrast to these raw datasets, tidy datasets lend themselves for easy analysis, manipulation and visualization because of their better defined structure.

This section describes multiple aspects of tidy data and explores several methods to transform a dirty dataset into a tidy one.

Wickham (2014a) defines a definition, structure and semantics for tidy datasets by means of a case study in his paper. In the remainder of this section, these three aspects are discussed.

Part 1 Data structure Most available statistical datasets can be represented as a table consisting of labelled rows and columns. In many cases, the same dataset can be represented in multiple tables with a different layout.

Part 2 Data semantics A dataset represents a collection of values which can be either numbers or textual. Wickham (2014a) states that every value of the dataset belongs to a *variable* and an *observation*. A variable is

the collection of values that explain an attribute of the observations done. An observation is the collection of values that were taken on the same unit.

The author demonstrates that a dataset can be represented in several ways. For example, a very common approach is to take possible variable values as column headers and units as row headers, as she showed in his example given in 2.1. Another variant is the transposed version of the first table, where variables are now row headers while observations are column headers.

Wickham (2014a) states that it is not easy to find a general definition of variable and observation, because these can be different depending on the application for which the representation will be used. To make things easier, a general rule of thumb is introduced that states that is it easier to find a functional relationship between variables than between observations. Another suggestion is that it is more intuitive to describe functional relationship between variables compared to rows and it is also easier to make comparison between groups observation rather than between groups of columns.

In the paper, 3 tables are presented that represent the same data but have the values displayed in a different ways. Table 2.1 is the usual representation that can be found the most often. The second table represents exactly the same data as the first one, with the difference being that the data is given in a transposed structure. Finally, the third table again contains the same data as in the 2 previous ones, but with the difference that the variables are the columns and observations in rows. The table 2.2 is believed to be tidier compared to table 2.1.

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Figure 2.1: Presenting variable values as column

Source: Wickham (2014a)

In the example given in the paper, there are 18 values which represent 3 variables (person, treatment and result) and 6 observations (3 names offered for the variable person, 2 treatment possibilities and five values

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Figure 2.2: Presenting variable values as rows

Source: Wickham (2014a)

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Figure 2.3

Source: Wickham (2014a)

given and one missing for the results). For this specific example, its straightforward to defined variables and observations. But, as mentioned before, finding a general definition for these two terms is not easy.

Part 3 Tidy data

From the previous explanation given in this paper, can be concluded that a dataset can be easy to read or in other words tidy or on the other hand can be as well messy and difficult to read, depending on the way how are the variables and the observations represented in rows, columns and tables.

According to Wickham (2014a), tidy data has the following structure:

1. A column is formed by each variable
2. A row is formed by each observation
3. A table is formed by each type of observational unit. And any other organization of the data is considered to be messy data

The goal is to have a tidy data because of the fact that it is easier for both analyst as well as computers to identify variables thanks to the standard way of structuring the dataset.

Distinction is made between fixed and measured variables. Fixed variables are known beforehand and give details about the experimental design. Measured variables are the ones measured during the study. The order of the columns should be in a way that the first columns are the fixed variables and then the measured variables come after.

Typical problems that exist when tidying messy datasets according to Wickham (2014a) are:

1. Column headers are values instead of variable names - This is on one hand a representation of messy dataset, but on the other hand it can be very useful representation as well. In a sense that, if the wanted operation can be represented as matrix that will result in a very effective computation. The author Wickham (2014a), gives an example of molten dataset as shown below in table 2.4, where dataset with three variables is molten.

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

Figure 2.4: This is the original data representation

Source: Wickham (2014a)

2. *Multiple variables are contained in one column* - instead of every variable having its own column. The author demonstrates this by referencing a column published by the world Health Organization on tuberculosis, shown in table 2.6. The column headers are depicting both the sex and age range.

The same data is shown in table 2.7, in a more tidy way. Each column is used for only one variable. However, the third column still holds values that can be split up, since each value consists of a letter representing the gender and an age range. The result of splitting this column is shown in table 2.8

The same data in molted version as it was done before with the previous example makes the original dataset to get closer look to what was

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

Figure 2.5: Molted representation on the same data

Source: Wickham (2014a)

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Figure 2.6: Original data from World Health Organization

Source: Wickham (2014a)

suggested that looks better. Additionally, as suggested in this part the column containing two different variables is split up on two separate columns, one for each of the two variables.

3. *Variables are present in both rows and columns* - instead of forming only columns and additionally having separate one for each variable as defined before. This is considered to be very complex type of chaotic data. In this part Wickham (2014a) has chosen example which is not so messy dataset but it has the characteristic of representing variables in both rows and columns. This data is from the Global Historical Climatology Network and it represents the different daily weather. It's char-

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

Figure 2.7: Molted representation on the same data

Source: Wickham (2014a)

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Figure 2.8: Tidied data: the column split into gender and age

Source: Wickham (2014a)

acterized as not tidy because it contains variables as separate columns

but also it contains variables as rows. As previously described the path to the getting the datasets tidy, the author first molted the data, which results in data closer to tidy data and after that it's done the second step to bring it to tidy stage. In the stage of molted data, the data still has column with variable names instead of variable. Also in the concrete example the missing data is dropped because in this case that's possible to be done, since we have the knowledge which month has how many days. In the last step, the two variables which were represented in rows are now represented in columns.

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

Figure 2.9: Original data from Global Historical Climatology Network

Source: Wickham (2014a)

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

Figure 2.10: Molted representation on the same data

Source: Wickham (2014a)

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Figure 2.11: Tidied the original data

Source: Wickham (2014a)

4. *Multiple types of observational units are represented in the same table* -like different things and types and remarks. In this part the author gives example of the Billboard dataset, which consists of two different observational components, like the *song* as well as the *ranking*, which was received for each particular week. But identical facts are repeated twice regarding the songs, like for example the *artist* and the *time*. According to Wickham (2014a) we can get tidier dataset if the given data from Billboard is divided into two separate datasets or more precisely: one for *song datasets*, where the information as artist, song and name are stored and a second one for *ranking*, where the rank of the songs is stored for every week.

5. *Single observational unit shown in multiple tables instead of one single table.* This type of problem happens regularly and it's pretty often one observational component to be found in many different tables and files. This problem can be seen in two scenarios: The first scenario is when format of representation is consistent over all tables and files and as such is considered as solvable problem. This is because it takes only few steps to read through all the files and tables, there is possibility for each table a new column to be added which is representing the actual file name and final step is to combine all of the tables into one table. The second possible scenario is however more complicated and it's a result of the structure of the datasets in general not being consistent during the course of time. Changes such as using the same variable but changing it's name, using different file formats, or making different

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

Figure 2.12: Original partial data from billboard dataset

Source: Wickham (2014a)

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6	98ˆ0	Give Me Just One Nig...	3:24
7	A*Teens	Dancing Queen	3:44
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

Figure 2.13: First dataset: *Song*

Source: Wickham (2014a)

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
1	2000-04-01	94
1	2000-04-08	99
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66

Figure 2.14: Second dataset: *Rank*

Source: Wickham (2014a)

conversions for the missing values, might result in only one solution and that is single and separate tiding of the each and every dataset. The second solution is difficult and not doable in only few steps like in the case of the first scenario of this problem.

But this are only most common issues which does not include all possible cases of having messy dataset.

Next, Wickham (2014a) explores tidy tools which as input get 1. tidy datasets. By tidy datasets is expected that the datasets are owning the characteristics given previously as definition in his paper, and giving t1. tidy dataset as output. Tidy tools are considered useful when they can be used to do data analyses more efficient as well as because the output is tidy data which can be used as input for another tidy tool. The combination of many tidy tools as a solution to messy datasets. Additionally, when datasets are tidy it is possible to make combination or to merge with other tidy datasets, and all of that can be done in a simple and fast way. According to the author: there are two causes of messy tools: the first one is that not tidy data was taken as input or the second reason is or the tool itself gives a messy data as output. Wickham (2014a) gives examples of both tidy and messy tools for three components which are of great significance for doing data analyses and

that are the following:

1. *Data manipulation-* which according to the author consists of few ways of manipulation such as filtering, transforming, aggregation and sorting. Filtering under certain conditions results in being able to focus on more important, useful parts of the dataset. Then by transformation includes changing or appending one or more variables. Aggregating - collection of several values. And finally, the sorting which means making the order of the observations different. In this part of the paper suggestions of functions are given, ideas how can this be implemented by using R, but since for my thesis I am using the programming language Python, I used corresponding functions with the ones suggested in the paper.

2. *Visualization* Visualization tools are giving visual representation of the tidy data received as input.

3. *Modeling* For modeling tools, tidy datasets are of high importance because they are needed as input. But this doesn't necessarily mean that the output will be tidy data as well, in fact there are many commonly used examples where the output is not a tidy data such as predictions or calculation of coefficients.

For the dataset used for the purposes of this thesis, this method is relevant because gives the idea to check whether any pupil used events that many other pupils didn't think of or maybe the meaning of this events was not very clear to many of them and they avoided using them. This information would be useful, to know what on this age is less clear or harder to comprehend regarding programming for children or also maybe additional explanation should be added before using the mobile application from the teachers side, like for example they can encourage more complex events to be used.

2.1.2 Data Cleaning

According to the author, Broeck *et al.* (2005), data cleaning consists of three parts, being screening, diagnosing and editing the data which is believed to have abnormalities. These three stages of getting the data clean can be done at different phases of a certain research. The authors believe that assiduous search for detecting errors normally brings more fruitful results compared to finding errors accidentally by completing other tasks. This does not mean that every temporarily inexplicable

data point is immediately categorized as an error, but it's just an indication that a further inspection is needed. One very common indication for further inspection is missing value, which can be result of different reasons. This is why the authors of this paper believe that having predefined rules for handling errors are best practice.

The errors itself can be caused at any phase of the data flow, as well as during the data cleaning. This is the reason why data cleaning has to be build on values or scopes that are likely and possible, or with other words knowledge of this values is needed before hand, which would help such errors to be avoided.

When dealing with time limitations prioritization is very crucial, in the paper it is suggested that error priority is very individual depending on the particular research.

According to the author Broeck *et al.* (2005), the screening phase is beneficial when it's objective and well structured. In order to identify potential error data points, the point of reference should be defined in advance, and then it will be easier to see which values are possible. Additionally, the screening phase can be automated to some degree but this will also mean that when a certain wrong data point appears but happens to have a value that is in the range of values which are likely to happen and acceptable as normal pattern than the screening for this type of wrong data point will be unsuccessful. There are other methods that help to notice this kind of error data points. Like suggested in the paper, this points can get uncovered when looked in respect to other variables by using scatter plots, regression analyses as well as examined consistency.

2.2 Finding patterns in the data

2.2.1 Sequential pattern mining

Sequential pattern mining plays important role in finding patterns in databases. The importance of this data mining problem is proved with the many applications where this problem is appearing. Like the examples given in this paper: customer purchase behaviour, Web access patterns, scientific experiments, disease treatments, DNA sequences and many other important applications.

According to the author Jian Pei *et al.* (n.d.), earlier proposed methods for sequence mining and most of the them were similar to the Apriori approach. The Apriori approach first checks how frequent subsequences are present.

Then the assumption is made that if a subsequence is not frequent, the containing sequence can not be frequent. Therefore, there is no need to further take the non frequent subsequence or its containing sequences into account.

The Generalized Sequence Pattern (GSP) algorithm is based on the Apriori algorithm. The similarity between Apriori and GSP is that they both make multiple passes and generates candidates. First, scanning is performed with which frequent candidates are found that are actually item frequent sequences of length one. This frequent sequences are used as so called seed set for the next pass for the purpose to find candidate frequent sequences. In every next pass the candidate sequence is build on the seed set from the previous pass and in every pass the length of the sequence increases for one item compared to the seed set of the previous pass. The algorithm stops with iterating either in the moment when there is no new sequential pattern found in a pass or when there is no candidate sequence that can be generated.

states that there are also some disadvantages to this approach, such as:

- Potentially huge set of candidate sequences - this method might generate very large set of candidate sequences even for a small seed set. The reason for this is because the approach includes permutation of all items as well as repetition.

- Multiple scans of database - Since the approach is iterative and the sequence grows in length in every next pass according to Apriori, this means that the database will be scanned at least 15 times for an example as (abc)(abc)(abc)(abc)(abc).

- Difficulties at mining long sequential patterns - The number of candidate sequences grows exponentially with the length of the sequential patterns mined, this is because every long sequential pattern is made by combining shorter sequences.

This are the reasons that show the need to for more efficient and scalable solutions. According to this paper, the goal was to develop a new method which will use the advantages of the Apriori method but will reduce or entirely bypass the disadvantages of it - like the expensive generation of frequent candidates.

As one of the possible solutions presented in this paper is Frequent pattern (FP) tree structure, which performance was tested. FP approach is exploring the maximum of shared common prefix paths within the tree by reordering the items in transactions. The problem with this method was either that items which were differently ordered can not be reordered or it broke in

sequential pattern mining. That's why the structures generated with FP tree would be huge and such a mining will not be beneficial.

As a second possible solution in this paper, FreeSpan-Frequent pattern projected Sequential pattern mining was developed. The way of working is as following: it takes the frequent items in order to recursively project sequence databases into a set of smaller ones and additionally to grow subsequence parts in each projected database. The results of testing it's performance are that FreeSpan mines the entire set of frequent items and it's faster and more efficient compared to Apriori, GSP algorithm. The disadvantage of this method is that it is costly, this is because it keeps the entire sequence in the original database without reducing the length, which is a lot when we take into consideration that the growth of the subsequences is explored at each split point within the candidate sequence.

As a third possible solution for sequential pattern mining method developed in this paper is PrefixSpan - Prefix projected Sequential pattern mining. In this method the goal is to find only the prefix subsequences, first to find length one sequence patterns, then length two and so on and then to project only their corresponding postfix subsequences into projected databases. In every projected database the local frequent patterns are explored and with that the sequential patterns are grown. For improving the efficiency there are two types of database projections presented, the first one is level by level projection and the second one is bi level projection. Or more precisely for saving the cost of projections and speeding up the processing, pseudo projection processing is used when the structure is possible to fit in the main memory. The advantages of this approach are that no candidate subsequences need to be generated, also the projection databases keep shrinking, there is no need of physically coping suffixes there is only pointer to the sequence. The results of the performance of this method in this paper show that bi level projection has better performance in the cases of large database and pseudo projection speeds up the processing when the projection database can fit in memory. This are the reasons why PrefixSpan can mine the complete set of patterns faster and in a more efficient way than Apriori based GSP method or FreeSpan.

Definition of problem with sequence mining and FreeSpan:

I = $i_1, i_2, i_3, \ldots, i_n$ with i_x being an item

Item - subset of items

Sequence s - ordered list of items $< s_1, s_2, \ldots, s_l >$

s_l - element of the sequence, denoted as $(x_1, x_2, \ldots, x_m)$

x_m - item

An item can occur maximum once within an element, but can occur many times within different elements of a sequence.

Length of a sequence - number of instances of items within a sequence l sequence - sequence of length l

sequence database S - set of tuples $< s_{id}, s >$, where s_{id}=sequence_id, s=sequence

Support of sequence= α in a sequence database S is the number of tuples in the database containing α

ξ(positive integer)=support threshold

(frequent)sequential pattern - A sequence α in a sequential database S if the sequence is contained by at least ξ tuples in the database

l pattern - a sequential pattern with length l

Example 1:

Sequence_id	Sequence
10	$\langle a(abc)(ac)d(cf)\rangle$
20	$\langle (ad)c(bc)(ae)\rangle$
30	$\langle (ef)(ab)(df)cb\rangle$
40	$\langle eg(af)cbc\rangle$

Figure 2.15: Sequence database S

Source: Jian Pei *et al.* (n.d.)

The min support threshold in this example is 2. The set of items in the database is a, b, c, d, e, f, g. The sequence in the first row has five elements, and that are: (a), (abc), (ac), (d), (c,f), where the items a and c are more frequent and appear in more different elements and also more than once. Item a occurs three times, which means that it contributes three to the length of the sequence, but the entire sequence contributes only one to support of ¡a¿. It also contains nine instances, which makes it nine-sequence. The two sequences 10 and 20 given in the table of this example contain the subsequence s=¡(ab)c¿, and this makes the subsequence s sequential pattern of length three.

Problem definition - Finding the entire complete set of sequential patterns in the database.

Example 2: If the same database is given as in example 1 but in this example it is solved with the method FreeSpan. As first step in FreeSpan scanning of the database S is done and the f_list is formed. The f_list gives the information for the support of every item. Like for example in the case of "a" it will be a:4, which is notation for a having support four. With the help of the f_list, one can find the disjoint subsets from the sequential patterns in S. This list also shows which items are not frequent and should be removed, like "g" is in the given example and for g no projected database will

be formed. But for the frequent items projected databases will be formed in order to mine of the subsets of sequential patterns. The process is recursively done on projected databases. FreeSpan is more efficient compared to GSP because FreeSpan projects the large sequence database recursively into a smaller projected sequence databases, and such a subsequent mining is limited with the relevant projected database, which is smaller set of relevant candidates. The disadvantage of the method FreeSpan is that it's costly. Because when a certain pattern appears in each sequence of the database then its projection does not get reduced. Like in the case of this exact example the only thing that was reduce was that the infrequent item "g" that was removed, but except that nothing else was reduced. Furthermore, the size of the subsequence might grow in length at any position, then for the candidate sequence every combination needs to be checked and that is also very costly.

2.2.2 PrefixSpan

According to the author Jian Pei *et al.* (n.d.), the main idea of PrefixSpan is to avoid the costly approach of FreeSpan, where projecting sequence databases were done by considering all of the frequent subsequences, but instead in the case of PrefixSpan, the projection will be only based on frequent prefixes, because if the prefix is not frequent than the rest of the subsequence is also not frequent.

Definition of prefix: If assumed that, the items in all elements are sorted alphabetically, the given sequence $\alpha =< \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \ldots \epsilon_n >$, and sequence $\beta =< \epsilon_1\prime, \epsilon_2\prime, \epsilon_3\prime, \epsilon_4\prime, \ldots \epsilon_m\prime >$,$(m \leq n)$, then β is called prefix of α if and only if the following 3 conditions are satisfied:

1. $\epsilon_i\prime = \epsilon_i$
2. $\epsilon_m\prime \subseteq \epsilon_m$
3. All the items in $(\epsilon_m\prime - \epsilon_m)$ are alphabetically after the ones in $\epsilon_m\prime$

The algorithm PrefixSpan explained consists of three steps: In step 1 the the entire sequence database is scanned and the most common sequences of length one are found. This frequent sequential patterns are considered to be frequent when satisfying the minimum support and with them f-list is created, where they are defined in the following way: ¡pattern¿ : support count, like for example ¡a¿ : 6.

In step 2 the complete sequence database is partitioned into subsets starting with one of the frequent sequences of length one, according to the prefixes found in step 1.

And finally, in step 3, subsets of sequential patterns are found. In a way that, it is only considered the subsequence after the first occurrence of the

frequent prefix sequence of length one which is defined in step one.

If this three steps of the PrefixSpan method are implemented for the previously given example in this paper, then the projected databases and sequential patterns are defined as given in the following table.

Prefix	Projected (postfix) database	Sequential patterns
$\langle a \rangle$	$\langle (abc)(ac)d(cf) \rangle$, $\langle (\_d)c(bc)(ae) \rangle$, $\langle (\_b)(df)cb \rangle$, $\langle (\_f)cbc \rangle$	$\langle a \rangle$, $\langle aa \rangle$, $\langle ab \rangle$, $\langle a(bc) \rangle$, $\langle a(bc)a \rangle$, $\langle aba \rangle$, $\langle abc \rangle$, $\langle (ab) \rangle$, $\langle (ab)c \rangle$, $\langle (ab)d \rangle$, $\langle (ab)f \rangle$, $\langle (ab)dc \rangle$, $\langle ac \rangle$, $\langle aca \rangle$, $\langle acb \rangle$, $\langle acc \rangle$, $\langle ad \rangle$, $\langle adc \rangle$, $\langle af \rangle$
$\langle b \rangle$	$\langle (\_c)(ac)d(cf) \rangle$, $\langle (\_c)(ae) \rangle$, $\langle (df)cb \rangle$, $\langle c \rangle$	$\langle b \rangle$, $\langle ba \rangle$, $\langle bc \rangle$, $\langle (bc) \rangle$, $\langle (bc)a \rangle$, $\langle bd \rangle$, $\langle bdc \rangle$, $\langle bf \rangle$
$\langle c \rangle$	$\langle (ac)d(cf) \rangle$, $\langle (bc)(ae) \rangle$, $\langle b \rangle$, $\langle bc \rangle$	$\langle c \rangle$, $\langle ca \rangle$, $\langle cb \rangle$, $\langle cc \rangle$
$\langle d \rangle$	$\langle (cf) \rangle$, $\langle c(bc)(ae) \rangle$, $\langle (\_f)cb \rangle$	$\langle d \rangle$, $\langle db \rangle$, $\langle dc \rangle$, $\langle dcb \rangle$
$\langle e \rangle$	$\langle (\_f)(ab)(df)cb \rangle$, $\langle (af)cbc \rangle$	$\langle e \rangle$, $\langle ea \rangle$, $\langle eab \rangle$, $\langle eac \rangle$, $\langle eacb \rangle$, $\langle eb \rangle$, $\langle ebc \rangle$, $\langle ec \rangle$, $\langle ecb \rangle$, $\langle ef \rangle$, $\langle efb \rangle$, $\langle efc \rangle$, $\langle efcb \rangle$.
$\langle f \rangle$	$\langle (ab)(df)cb \rangle$, $\langle cbc \rangle$	$\langle f \rangle$, $\langle fb \rangle$, $\langle fbc \rangle$, $\langle fc \rangle$, $\langle fcb \rangle$

Figure 2.16: Algorithm PrefixSpan on the example sequence database S

Source: Jian Pei *et al.* (n.d.)

This is relevant work for this thesis because with the use of the algorithm PrefixSpan and the explained steps in this paper, the most used events by the students, from the application Pocket Code were detected. Where in the first step were found the most used events by all of the students. In the second step the search space was divided, or more precisely all the events were partitioned into subsets starting with one of the most frequent events found in step one. Finally, in the third step, sequences starting with frequent event were found, or more precisely, subsets of frequent event sequences were mined by projecting the database as shown in the example of this paper and mining each sequence recursively.

As stressed in this paper, the biggest advantage of PrefixSpan when compared to other algorithms like Apriori and GSP is its efficiency. This is result of the fact that PrefixSpan does not generate or test all possible candidate sequences, but it rather grows longer frequent sequences based on the frequent shorter prefix sequences. Because the projected database is much smaller than the original one, PrefixSpan searches for frequent candidates in a much smaller space. The biggest cost in the PrefixSpan algorithm is the construction of the projected databases, which can be a great cost in cases when there is big number of sequential patterns. As a solution to costly databases projection of PrefixSpan, in this paper are offered two solutions: bi-level projection and pseudo projection. With the bi-level projection the first step is exactly the same as the fist step of the PrefixSpan algorithm with witch the most frequent sequential patterns of length one are found, but the second step is different. Namely, instead of making projections based on all

possible frequent patterns of length, in the second step is constructed nxn lower triangular matrix, where n is the number of prefixes found in the first step. In this way all the sequential patterns of length two, can be easily immediately identified and this approach is a lot more efficient because many not needed projected databases don't need to be created at all. The second solution offered, pseudo projection, where the use of pointers is presented as an alternative solution to creating physical projection by collecting all the postfixes. The idea is that the pointers can be used to point in the main memory at sequences as a pseudo projection, and with that the physical costly copying from PrefixSpan is avoided and the same results are obtained in a less costly way.

2.3 Classification and defining significance

2.3.1 Gender different behaviour in programming

The author Funke *et al.* (2015) stresses that there is great gender difference in the number of students attending studies related to Computer Science. For this purpose the authors of it collected outcomes of other studies and made their own study by using different approach. Or more precisely instead of like the rest of the relevant researches they did not only follow behaviour of students of different genders but instead they surveyed 63 experienced teachers in filed of Computer Science, and the teachers were asked to express their point of few coming from their experience with different genders in their filed and additionally with special focus on programming and what are the differences that they noticed between students from different gender so far.

There are many reasons why there is big difference in female and male student in the filed of Computer Science. In some of the references of this paper it is stated that one of the reasons are the stereotypes and that typically as a role model for Computer Science is male computer scientist and this remains for very long time. Another statement in the relevant literature is that often boys choose to study computer science because of great interest for it while many girls study this studies because they see great future perspective in it. For boys is typical that they have the interest for Computer Science and programming sooner than for girls and that made them try different things and already gain some experience and prior knowledge even before starting their studies. On the other hand for girls it is stated that are less confident in this area and it's typical for female students in case of failing to finish studies related to Computer Science to believe that is due to lack of ability.

According to the relevant literature mentioned in this paper, the learning process is influences by several different aspects, like for example the emotions affect differently the two genders. Positive emotions during the learning affects the performance of learning. Happiness has positive and anxiety has negative influence on the learning process and the motivation of female students. Girls are naturally more emphatic compared to boys. The communication with others is also very important factor with big influence on the their performance for girls. For girls group works, collaborations and discussions bring many advantages and this brings better results for their performance.

The study was including 185 Bavarian teachers from the filed of computer science, out of which 49 were women and 136 men, participate in answering questionnaire for their students and 7 of them were interviewed from which 1 women and 6 men, and on both the interview and the questionnaire the main question was: "What are the differences between girls and boys concerning programming". At the end of the study 63 responses were handed in and 57 of them contained the answer to the previously mentioned question, out of which 14 were women. The analysis of the answers were done in a way that each sentence given as a reply to the questions by the teachers was categorized. And this was done separately by two different researches and both of them got very similar results.

All the responses were divided in 14 categories and that are the following: 1. "structuredness" 2. "self-conflict" 3. "scientific curiosity" 4. "no difference" 5. "results" 6. "interests" 7. "learning receptivity" 8. "accuracy" 9. "previous knowledge" 10. "perseverance" 11. "creativity" 12. "frustration" 13. "team work" 14. "evolution".

The way how this is done it's shown on the next figure, where in the first column is given the sentence or part of it which some of the teachers wrote for their students as answer in the questionnaire, the second column is to which particular category is this sentence attached to and finally the third column is how often such a sentenced appeared among all the teachers.

The results from all the feedback of the teachers and after categorizing each sentence they gave was showing that 19 % of the teachers don't see any difference between genders regarding their programming skills. 14 % of them could see difference in students from different gender regarding the categories "learning receptivity", "results" and "interests". According to the results girls show more structured approach, lack of curiosity in programming and are less confident about their abilities. Girls have generally less interest in Computer Science and especially in programming. And finally, girls benefit more from conversations, collaborations and positive emotions.

Similarly, in the study "Woman Catch Up: Gender Differences in Learn-

category	sample	abs. freq.
structuredness	"Boys just start working, girls think about it first"	28
self-confidence	"and because there are sometimes a few girls who do not dare to"	19
scientific curiosity	"Boys want to try more, test, find things out"	18

Figure 2.17: Way of mapping sentences to predefined categories

Source: Funke *et al.* (2015)

ing programming Concepts ", Murphy *et al.* (2006), inspection was done on gender differences in graduating students in Computer Science as well as the different perception of programming concepts. In this study the goal was to check how many students are persistent in their studies in field of Computer Science and whether the fact that female students start with less pre experience in programming and less confidence for their abilities, influences on them or not in achieving the same results as male students. In this study participated 73 students from 8 different universities and colleges from United States, graduating Computer Science bachelor studies. The focus was put on specifically defined 26 programming concepts, how good their knowledge is for each of them, how much pre knowledge they had before their studies and how hard was to learn it during their studies. The study was made with methodology that didn't include questions that leave space for subjective answers regarding the knowledge of programming, because some students have less confidence than others in their knowledge. This is the reason why the approach in this study was done in a way that the questions were focuses on the 26 defined programming concepts and the students answered whether they had knowledge for them before they started their bachelor studies in Computer Science, which once they learned during their studies, and how difficult was to achieve that. This approach is more likely to bring realistic results and in such a way also for the students with less confidence makes it less difficult to answer on questions like whether they used a certain programming concepts or not, rather then how confident are they will programming in general.

The interview consisted of two types of categorization tasks: unconstrained and constraint. The unconstrained type, got the name because of the way the subjects categorize certain concepts comes from the internal representation of those concepts. The second type are called constrained because both the criteria name and the category name were predefined be-

forehand. During both types both subjects were categorized in the set of 26 cards containing 26 programming concepts. Including both, more simple concepts as well as more complicated concepts, which are shown in are listed in the table 2.18.

function	scope	type	constant
method	list	loop	boolean
procedure	recursion	expression	event
dependency	choice	tree	variable
object	state	thread	if-then-else
decomposition	encapsulation	iteration	array
abstraction	parameter		

Figure 2.18: 26 card types

Source: Murphy *et al.* (2006)

For every participant during the study given in this paper, the following data was collected: expected graduation date, age, gender, first spoken language, first and second programming language, age on which they started programming, the level of experience with other programming languages, the average grade they obtained as well as the average grade for CS subjects. Additionally,the data was collected from the answers to the constrained type of cards, which contained the following 4 questions: "1.When i was first introduced to it" - and the offered answers were: "before college, lower-level CS courses, upper level CS courses, on the job, on my own or don't know the term". The second question was "When I mastered it", the possibilities to answer were "before college, lower-level CS courses, upper level CS courses, on the job, on my own, haven't mastered it or don't know the term". The third question was defined as "Language paradigm", which could be answered as "procedural, functional, object-oriented, logic, not sure, don't know the term". And finally, the forth questions was defined as "Difficulty level", with the possible answers: "easy, intermediate, advanced or don't know the term".

The results were called from all participants, which in total were 73 students, from which 22 were female and 51 were male, with the same age in average of 24 years old. The results have shown that female students start programming later than male students, or in numbers, from the participants, the female students started programming on average on age of 18,4 years old, while male students on average 16,5 years old. This difference means that female students are more likely to start programming directly at collage, wheres male students have already stated before they attended college. Regarding the marks achieved, they were comparable and there wasn't big difference between female and male students. Next, the results from the answer on

difficulty from the constrain cards showed that both female and male students know most of the concepts, a bit bigger percentage of the girls found the concepts easy compared to guys, or in numbers, 54% of the girls found the concepts easy and 49% of the guys too. More male students found them intermediate then guys, and more boys(18%) found them "advanced" compared to girls(15%). For the difficulty ratings of every single concept, both boys and girls had similar answers, they marked the concepts "decomposition, abstraction, dependency, thread" as difficult concepts and the concepts "constant, boolean, variable" as the easiest.

From the results of this study, can be concluded that male students learn about and master many programming concepts before college and for female students that's different. This is visible in the figure 2.19, which was created for this paper where the differences are noticeable.

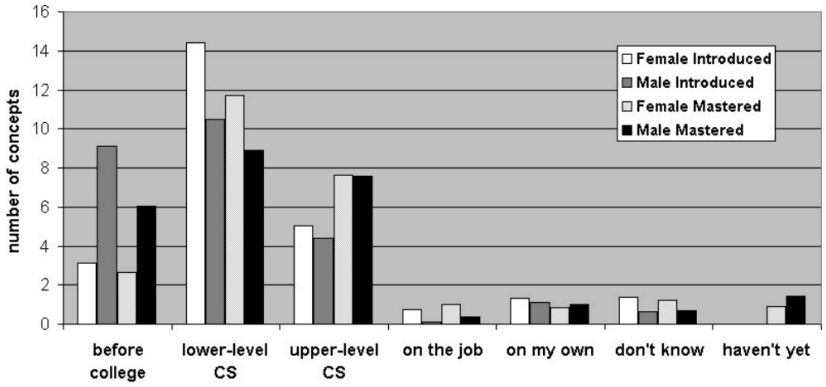

Figure 2.19: Predefined concepts and results from both male and female students

Source: Murphy *et al.* (2006)

Furthermore, there were interesting outcomes regarding the self motivation and the self learning. In general it is believed that this stronger side for male students, that they tend to be more self motivated and learn concepts on their own, but the results from this study show that both male and female students learned very few concepts on their own. The results also show that, even though female students start college with almost no prior knowledge, it looks like that they can master the concepts during their CS classes at college and reach the level of their male co-students. This is represented in the figure2.20 which was included in this study, where it can be clearly seen that girls start with less knowledge but by the time they are completing the courses as part of their CS studies they are almost as good as the boys.

According to the fact how many concepts the students knew before college or during collage and how many of them they mastered, the students in the

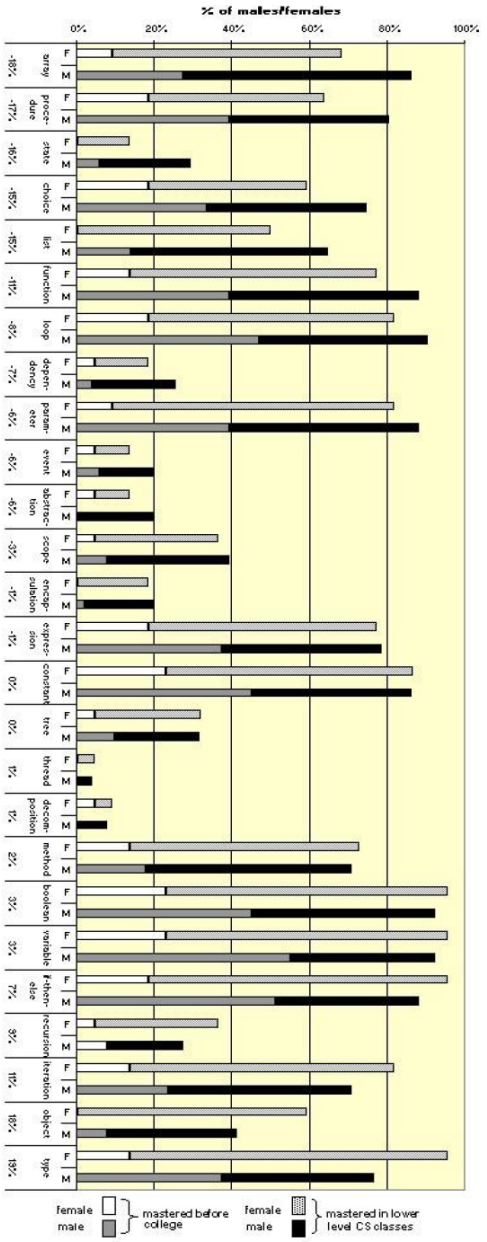

Figure 2.20: Mastering concepts given for both genders

Source: Murphy *et al.* (2006)

study were divided in 4 different groups and then a relation was made to their average success. The results of the comparison are given in the figure2.21 below, where can be seen that the average mark is slightly better when there is more pre-knowledge, but the difference is very little in the average mark

where the pre-knowledge is very different.

# of concepts introduced before college	Percentages			CS GPA		
	M	F	All	M	F	All
none	20%	64%	33%	3.11	3.20	3.16
1 to 7	22%	18%	21%	3.30	3.55	3.37
8 to 14	33%	14%	27%	3.33	3.42	3.35
15 to 26	25%	5%	19%	3.34	3.82	3.38

Figure 2.21: Programming experience before collage and average results in collage

Source: Murphy *et al.* (2006)

Which again leads to the conclusion of this paper that even though girls start with less pre-knowledge in programming they are able to start learning it in college and still achieve comparable results as guys.

The data used for this thesis (the NOLB data), was entirely collected from students before college, or more precisely from students being part of elementary or high school. If the same study is implemented to the data that I am analyzing and furthermore new data is collected after the same students go to university, especially the once that decide to study bachelor of Computer Science, then even more precise and accurate results can be found. Since in this case the students will not have to categorize their knowledge and answer to predefined questions but there will be data proving their knowledge on that age, their persistence in using more events and trying with more bricks as well as the information how soon they knew got introduced to certain concepts. In other words this study gives motivation for future work as well as gives expectation that even in the girls don't show the best results before college because of luck of pre-knowledge in programming, they can later be at least as good students as the male students. Furthermore, using the application "Pocket Code", means that both girls and boys got introduces to programming concepts before college, and there is left to investigate whether this will make the girls to have less concepts to catch up later on, or said in another way girls will also have programming pre-knowledge and then the results is unknown whether this will mean that they will show better average results later in their studies or the results will be comparable like it was in the case without pre-knowledge in programming.

2.3.2 Creating games as a way of motivating girls for IT studies and programming

In the study from Denner (2007), it's pointed out that girls and guys are using technology equally but in the creation and the design of the technology their is great gender gap. The author creates a project in which girls are creating games and his goal is to increase the number of girls that will feel comfortable to create technology. He was aiming to minimize the negative influence of the typical stereotypes for girls connected with technology and to increase the motivation and the expectations for success with computers by providing great support during this project. In this study 114 girls took part, most of them on age 11,8 , worked in pairs for creating the games, all of them were part of middle school at the time of the project, and the reason for this is author's believe that middle school is the time when everyone makes the big decisions about future profession and studying programs and he sees this part of the student's life as the right time to influence and show more opportunities to the female students than the once that they are usually considering. Many schools have the same goal for many years to decrease the gender gap and attract female students to studies in field of science, technology, engineering, math, but not so many tried to actually find out what are the reasons for this great gap and why both genders are not equally interested in the study fields mentioned. This study suggests that if girls receive support and start believing that they are capable of doing computer related studies, that increases their expectations, then they are more likely to be persistent in the decision of doing IT studies.

This study from Denner (2007) was conducted in 23 sessions of two hours over four semesters. The students had the sessions twice per week during the school semester after their lectures and four times per week during the summer holidays. The girls task was to create a game that would be useful to the new coming students in their school to adjust faster.

This study contained different parts which were having different directions of focus. In the first part named "Learning by Design" was focused on the creation of the game by using Macromedia's Flash MX software. The next part was "Scaffolding and Modeling" and during this part the focus was put on understanding, learning to think in a problem solving way and understanding the concepts needed for creating a game. The part "Knowledge about computers" was collecting the information from the female students whether they have access to computers at home, which 87% had and how much it's computer used in their family. The next part of this project was "Collaborative learning" which included included activities that helps to create community of learners. The tasks were done in pairs and the task was

divided in two parts: designing and programming the game. With this part the goal was to strengthen the communication and the mutual decision making. In the part "Social support" the female participants were asked to give the information how many of their friends that have interest in computers are female or male, as part of the pre and post survey. The next part named "Identity Formation" was the part where the female participants got the chance to get to know more information about possibilities for career in IT, additionally they also got the chance to interact with female role models who tried to change the photo of the gender stereotypes for the participants.

All the activities were design to result in positive experience of working with computers, fun and attractive. Additionally, they used this project to make connections between technology and real world everyday applications. The evaluation of the obtained results was focused on two things: 1. To find whether there is difference between the girls who participate in this project and girls who didn't, both part of middle school, regarding their expectations for success. 2. What did the participating female students liked about the learning during this project and what they didn't like and has space for improvement.

At the beginning of the project 214 girls completed the pre test, from which 126 girls who were part of the project and 88 that were part of the same school but not part of the learning sessions of this project. In the analysis to get the result were only included the students who participated in both pre test and post test, including both groups, the ones participating in the project - treatment group and the other who didn't participate - comparison group. From the total number of girls who participated in the first step of pre test, 28% of the treatment group and and 17% of the comparison group didn't participate in the post testing. And with that are not part of the final results.

T-test can be used to find out whether two sets of data are significantly different one from another. T-test was done in order to find the reason or the influence for not all girls participating in the final test. T-test represents statistical hypothesis test in which the statics for a certain student are following a certain t-distribution under defined null hypothesis. The result showed that the reason for not participating had nothing to do with the skills of the students, their age or confidence but it was connected to computer accessibility at home and computer knowledge. The once not having computers at home were less likely to take part in the post test. And regarding the knowledge, the once who were part of the comparison group and showed more knowledge at the pre test were less likely to take part in the post test too.

The results had the goal to show how the participants changed over the period of two years and if the changes that are achieved can be also seen at

the students from the comparison group. The results showed that the confidence of the girls participating in the learning part of this project increased slightly, but wasn't drastically different compared to the confidence of the students from the caparison group. For increasing the computer skills was more noticeable difference between the students participating in the learning part compared to the ones not participating in the learning part. Another thing that change for the female students part of the treatment group is there self knowledge, their perception of their knowledge for computers was different and they were more aware for their knowledge in this filed. As part of the survey all the female students received the statement "boys do better than girls with computers", the female students have to express themselves how much they agree with this statement or not before the start of this project and after it. The treatment group of girls had decreased stereotypes by the end of the project, while the stereotypes for the comparison group only increased during the period of two years. On the question connected to problem solving, which was one of the topics they received training for during the learning sessions, 60% of the treatment group answered that if there is a problem computer related they will first try to find the solutions by themselves, which was very different than the answers of the caparison group. With the aim to check whether there are changes related to social support, the students received the statement "There is no one I talk to about computers ", and they answered how much this applies to them. 10% of the students from the treatment group agreed at the beginning of the project and only 2% after the project, while at the comparison group 11% agreed at the beginning and 10% agreed again two years later, which shows big improvement for the treatment group. And finally, the students were asked to share which activities were positive experience for them and which parts they didn't like. The results of this question were that girls liked that they got the chance to use the computer regularly, for creating games, they liked playing the games when they had to do it for getting ideas for creating the new game and they like the feeling of completing and achieving. On the other hand, even though in the previous paper it was mentioned that girls benefit of working in pairs, however many had difficulties working in pairs, because they felt that the other girl didn't contribute as much, didn't accept ideas or didn't come up with ideas.

Chapter 3

Implementation and results

3.1 Cleaning and tidying the NOLB data

3.1.1 Data received for analysis

The data was collected between late 2016 (table 3.1) and the first half of 2017 (table 3.2). There are 3 files in total from 2016, one from October, second one from December and third one covering all 3 months from October until December. Additionally, there are 4 files from 2017; one file containing data from February and March, and 3 files from May 2017. There are no files for January 2017 because the data was lost.

Not only do the files contain data from different years, but the data format between files of 2016 differs from the format of 2017.

The differences in the data format from 2016 and 2017 as follows:

The first field is the username. The username in both cases (2016 and 2017) is built from 4 or 5 letters and 4 digits. The first letter is always “n” on every valid user name, because this means that is tracked by the group of the ”No one left behind-NOLB” project. The second letter gives the information from which country is the user, this letter can only be:“a”, “e” or “u”, and the meaning of it is: “a”=Austria, “e”=Spain and “u”=United Kingdom. The next 2 or 3 letters are the code from schools. All of the schools are coded with 2 letters except one which contains 3 letters, and all of this cases are separately checked in the code, in order to eliminate invalid user names. For schools the possibilities are: for Austria “gi” - for Graz International Bilingual School-GIBS, “ak” - Akademisches Gymnasium and “bb”- Bord Birkfeld. For Spain there are two options “mf”- Santa Maria and “ub” - Ubeda. And for Great Britain there are 4 different schools with the following shortcut codes added in the username: “of” - Oakfield School, “ss” - St Stephan Primary, “ctk” - Christ the King and “sp” - St. Peters.

Parameter name	Example value
Username	nabbm0002
TimeStamp	1479983102923
EventType	openProgram
Country	at
OS	Android
AppRelease	2.0
Locale	de_DE
programmer Dim1	
programmer Dim2	
Date (YYYYMMDD)	20161124

Table 3.1: Data format 2016

And finally, the last letter from the username can only be "m" or "f", which stands for the gender of the users, m=male and f=female. At the beginning in the data there were 1122 unique names tracked, but after removing all invalidly saved, 767 usernames were filtered, with that there are 355 unique valid user names for analysis.

In the 2016 version there is "Timestamp" tracked in milliseconds. The next field is the "Event Type" from the finite list of possible events that can be used in the application. The next 6 parameters, tracked separately in the old version, are omitted in the new version. These are the following parameters: "Country", "OS", "APPRelease", "Locale", "Programmer Dim1", "Programmer Dim2". In the 2017 version these fields are omitted in the data format, but are either known from the experiments or are derivable from other data fields. As an example country can be derived from the username of a user. The two parameters "Programmer Dim1" and "Programmer Dim2" were never used. The next parameter is "Date" represented in format "YYYYMMDD" and "Date (YYYY-MM-DD) Time(hh:mm:ss)" in version 2016 and 2017, respectively.

The last parameter is the "Custom Data" which is tracked in both versions. It contains an number of key-value pairs. This number differs per event. The representation of these key-value pairs also differs between the 2016 and 2017 version.

In the 2016 version the pairs are stored as {*key1:value1|key2:value2|...*}. In the 2017 version it is stored as {*"key1":"value1","key2":"value2", ...*}

In the new version there are only 4 parameters in total. The first one is the username, which is structured in the same way as in the version from 2016. The next parameter is again in this version "EventType", which is

Parameter Name	Example value
Username	neubm0042
EventTypes	openProgram
Date (DD-MM-YYY) Time (hh:mm)	2017-05-23 21:00:01
Custom Data (key1:value1 key2:value2 ...)	“VersionCode”:8, “ProgramName:salud

Table 3.2: Data format 2017

also identical. The third one is the field ”Date and Time”, where the time is tracked in the form hh:mm:ss which is easier for analysis compared to the “TimeStamp” given in milliseconds as in the previous version. And the“CustomData” field has the same keys and values as previously, only with different format.

3.1.2 Tidying the data

Because of the different data structures discussed in section 3.1.1, there is a clear need for tidying the data before it can be analyzed.

The first step is creating one dataset with an uniform data representation from the different data files. To do so, all files were converted to the latest data structure of 2017. Practically, this means that the “Timestamps” and “CustomData” were converted to the new version and some parameters omitted. All this unified data was then again stored in a CSV file.

The following step is removing invalid usernames. The data contains many entries created for testing the application by teachers or by the project organizers. This data can be identified by the username, since the username follows a strict pattern as discussed in section 3.1.1. The invalid usernames do not follow this pattern, are containing only numbers, are containing words such as ”test”, ”teacher” or ”translator” or are saved without NOLB account. This makes them not relevant for this analysis.

According to the paper “Tidy Data” Wickham (2014a), which is discussed in relevant works in chapter 2, data is considered to be tidy when:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table - this is the reason why from the column ”Date and Time” are split into 2 columns: ”Date” and

"Time". This tidied way of representation can be seen in the example on Figure 3.1, where few of the used events from user "naakf0001" are shown. This is a valid user since the username follows the username pattern. Additionally, it can be concluded it is female Austrian student studying at Akademisches Gymnasium.

```
'naakf0001'
```

	EventType	Date	Time	CustomData
0	downloadProgram	2016-11-23	14:56:25	{"programName":"MaxeneMarijana_Alice"}
1	stopExploreSession	2016-11-23	14:56:41	{"durationOfWebSessionInMillis":"90207"}
2	startExploreSession	2016-11-23	14:58:10	{}
3	stopExploreSession	2016-11-23	14:59:21	{"durationOfWebSessionInMillis":"70568"}
4	startExploreSession	2016-11-23	14:59:29	{}
5	init_session	2016-11-23	15:55:21	{}
6	startExploreSession	2016-11-23	15:55:24	{}
7	stopExploreSession	2016-11-23	15:56:13	{"durationOfWebSessionInMillis":"49871"}
8	startProgramExecutionSession	2016-11-23	15:56:33	{"sceneName":"Szene 1","programName":"AnnaHedi...

Figure 3.1: List of events from user "naakf001"

This tidy representation dataframe was created for every user. This representation contains all events used by this user during the whole tracking period. Next to dataframes for the user, also dataframes for each session were created. These dataframes contain only the events within a session of a certain user.

3.2 Finding patterns in the received data

3.2.1 Similarity and significance

In order to be able to find any patterns in the data, it is essential to get to know the data as good as possible. Therefore, many statistics were calculated from the data.

All of the users used 59 unique events in total, during the entire period of tracking, which is exactly 143 unique sample days. There are 355 users in total (Table 1.1) whose results were correctly tracked and are considered to be valid, from which 197 male ($\sim$55.5%) and 158 female ($\sim$44.5%) students. From that 131 users are Austrian students, 60 male and 71 female students.

There are 85 students from Spain, where the project was running in 2 schools but there are only valid results from one school where all of the 85 students study. From them 59 male and 26 female students. There are 139 users in total from United Kingdom, 78 male student and 61 female students.

From the total number of 355 users, 55 of them used the application only once - have only one session and the remaining 300 users have more than one session. This is the reason why analysis is done per session and per user. The maximum amount of sessions per user is 21, as is shown in figure 3.2, 3.3, and 3.4. Figure 3.2 shows the graphs for number of sessions per user for all the users from the dataset, while figure 3.3 shows this for female users and figure 3.4 shows this for male users.

There are two events that mark the beginning and the end of a session, namely "init_session" and the "end_session". However, these two sessions are not always present, making it difficult to define where a user started or stopped a session and began a new one. Students who used the application, very often forgot to log out after their class. When they use the application the next time, no new "init_session" event is recorded because the user is still logged in. This is handled by making the assumption that parts of the same session can not be present in more than one day. If there are more sessions in one day, this means they are handled as one session.

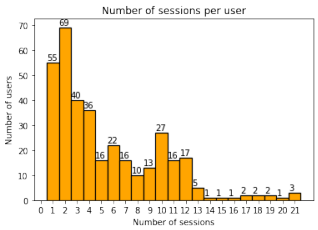

Figure 3.2: All users

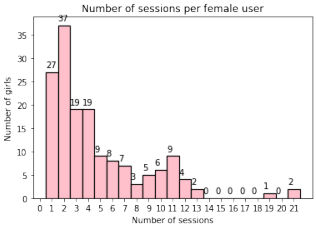

Figure 3.3: All female users

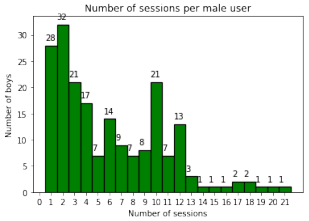

Figure 3.4: All male users

First, the average number of events per user was estimated. But since some of the users (∼15,5%) had only one session and the others had different number of more than one session, it is difficult to make conclusion out of this results. Later, the average number of events per session was estimated, which is also different between the the users and is presented later in this chapter.

First, the results per user are compared. The comparisons are made between genders and between countries are given in table 3.7 and the visual comparison is shown on the graph 3.5 and 3.6. The differences between schools are given in table 3.8 and shown on the graph 3.7 and 3.8. There

is even a significant difference between the average number of events in the first and last session of the users, as shown in table 3.3. Table 3.4, 3.5, and 3.6 shown the same statistics just for students from Austria, Spain and the UK.

The numbers that are significantly different from the other groups of users are marked in the tables with the asterisk ("*"). From Table 3.7, showing the average events per user per country, the conclusion can be made that the average number of events for Austrian students and British students are comparable. The average number of events for Spanish students is much greater.

When events per user for the female students are compared, Austrian female students have the smallest amount of events per user (252,83), then the British female students have 314,10 events and finally Spanish female users have the greatest average compared to the other two countries of 404,08 events. For the male users, United Kingdom has the smallest average of 395,51 events, Austrian male student have slightly more 419,55 events per user and finally the Spanish male users have again the greatest average of 624,46 events per user. On first sight it looks like there is a large similarity between the students from Austria and United Kingdom, because the numbers estimated as averages or the results are more comparable, and Spain shows different behaviour and is less similar to the other two countries. This is, however, a first conclusion based on a very general metric. Although this metric shows a difference, that difference is not necessarily a significant indicator for different behavior. For example, from this results the conclusion is that Spanish students have the highest average per user, but there are less Spanish students (in total 85) in the dataset, compared to Austrian(131) and British(139) students. Moreover, only 26 Spanish female students were

Country	Gender	First session	Last session	Avg. over all sessions
Austria		56,89	165,83	107,05
	female	37,40	192,90	103,00
	male	78,81	135,37	111,59
Spain		45,41	107,38	88,86
	female	38,20	100,83	81,52
	male	48,80	110,47	92,31
United Kingdom		25,82	45,59	48,78
	female	20,67	46,81	44,72
	male	29,71	44,67	51,85

Table 3.3: Number of events in the first and last session and average over all sessions from students with more than one session

School	Gender	First session	Last session	Avg over all sessions
GIBS		38,15	139,84	106,26
	female	32,75	142,42	101,78
	male	48,05	135,11	114,49
Borg Birkfeld		125,77	58,44	97,10
	female	2	3	2,5
	male	133,05	61,70	102,67
Akademische Gymnasium		48,27	254,57	113,68
	female	46,85	285,70	110,05
	male	50,46	232,07	119,26

Table 3.4: Average number of events in the first and last session and average over all sessions for Austrian students

School	Gender	First session	Last session	Avg over all sessions
Ubeda		45,41	107,38	88,86
	female	38,20	100,83	81,52
	male	48,80	110,47	92,31

Table 3.5: Average number of events in the first and last session and average over all sessions for Spanish students

School	Gender	First session	Last session	Avg over all sessions
Oakfield		5,92	39,35	23,57
	female	13,60	25,00	16,95
	male	1,66	47,33	27,25
St. Stephans Primary		28,01	46,55	54,30
	female	22,55	48,52	49,24
	male	32,07	45,07	58,06
St. Peters School		30,00	45,70	41,87
	female	17,10	51,20	41,41
	male	42,90	40,20	42,33

Table 3.6: Average number of events in the first and last session and average over all sessions for British students

tracked, compared to 71 Austrian and 61 British. The different number of samples for different countries makes it more difficult to make conclusions. The average number of events per user is so high for Spanish users, because there is one Spanish student who has 3487 events tracked during this project, then there are 17 students (20% of the Spanish students) in total, who have more than 1000 events per user and only 5 users who have less than 10 events per user. In order to get better overview on the dataset, regarding how much the average number of events per user varies in each of the different groups, the standard deviation was estimated and added to the existing plots representing the average of events in different groups. The results for the three female groups are shown on Figure 3.9 and the results for the three male groups are given on Figure 3.10. This plots prove that the average events per user varies the most for the Spanish users, both male and female, and varies less for Austrian and British students.

Additionally, the differences between schools and schools per gender are given in Table 3.8, where it can be seen that the biggest average number of events per user is from the school "Ubeda" with 557,05 events per user, which is the one of the two schools from Spain that were part of this project, but is the only one that successfully tracked the results. This means that the results from Spain are completely based on the "Ubeda" school, which is different for the other 2 countries where in Austria there are 3 different schools tracked and in United Kingdom there are 4 different schools tracked. This table gives more information for the 2 countries that have results containing more than one school. The result that stands out compared to the rest of the results in the same table 3.8, is the very little number of average events per user in general, from the school "Christ the King" in the United Kingdom. That is only 16,71 events per user and is significantly smaller than the average events per user for the other groups too.

	All users from this group	**Female**	**Male**
All users	395,7	301,3	471,4
Austria	329,19	252,83	419,55
Spain	557,05*	404,08*	624,46*
United Kingdom	359,78	314,10	395,51

Table 3.7: Average events per user per country

As written at the beginning of this chapter, the same table and the same graphs are made for representing the average events used per session. In this case every session is taken into consideration as a separate, individual session. Even if there are several sessions from the same user, each of them

	All users from this group	Female	Male
GIBS	292,62	246,68	374,52
Akademisches Gymnasium	309,92	325,42	283,36
Borg Birkfeld	435,14	4,5*	547,48
Ubeda	557,05	404,08	624,46
Oakfield School	169,88	71,00	239,10
St. Stephan Primary	492,88	444,55	528,88
Christ the King	16,71*	3,00*	27,00*
St. Peters	120,73	106,00	135,46

Table 3.8: Average events per user per school

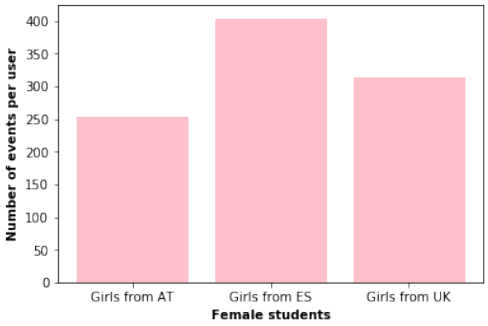

Figure 3.5: Female users from every country

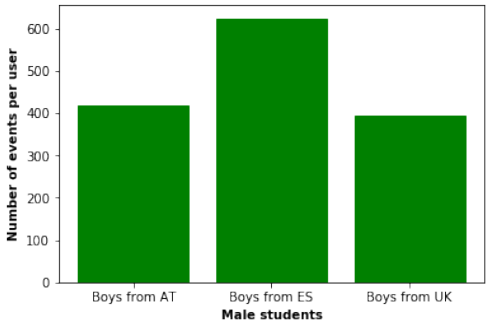

Figure 3.6: Male users from every country

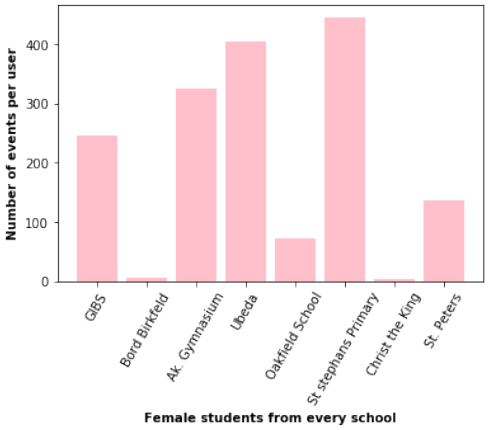

Figure 3.7: Female users from every school

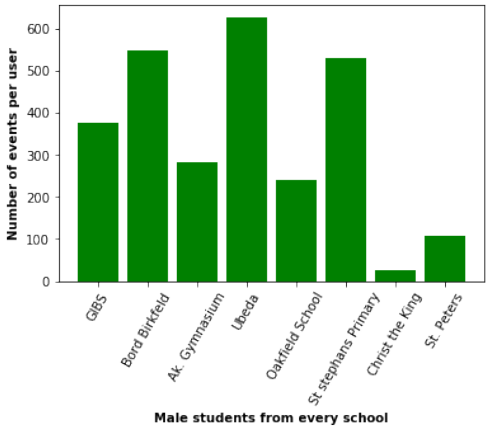

Figure 3.8: Male users from every school

is treated as individual new session. The differences between genders and between countries are again presented in table 3.9 and the visual comparison

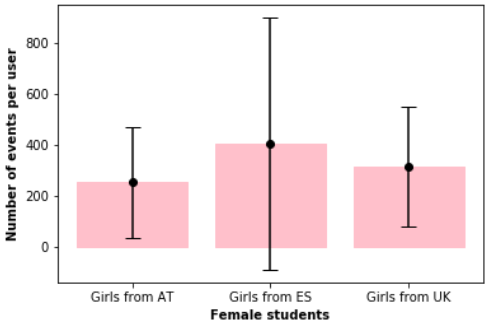

Figure 3.9: Female users from every country with Standard Deviation

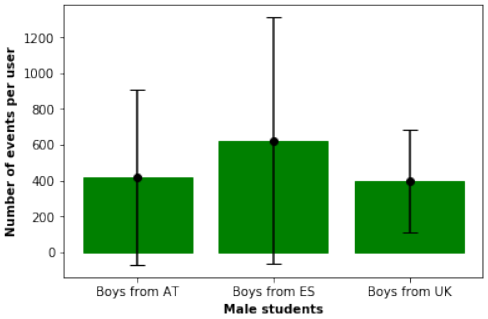

Figure 3.10: Male users from every country with Standard Deviation

is shown on the graph 3.11, 3.12 and the differences between schools are given in table 3.10 and shown on the graph 3.13, 3.14.

All users			
Austria	329,19	252,83	419,55
Spain	557,05	404,08	624,46
United Kingdom	359,78	314,10	395,51

Table 3.9: Average events per session per country

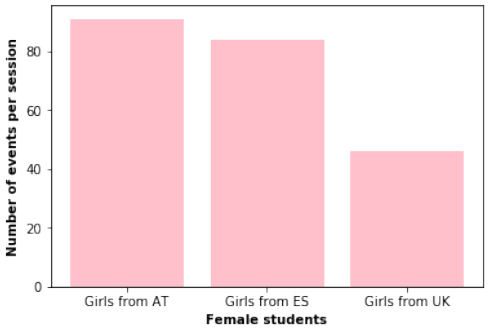

Figure 3.11: Female users from every country

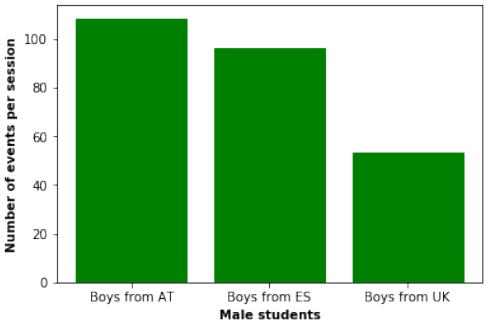

Figure 3.12: Male users from every country

In order to be able to make conclusions whether certain difference is significant or not and to be able to interpret the results in the right way, the p-value was needed to be estimated. The p-values between each two groups from the list of all groups can be obtained by implementing the (Student's) t-

	All users from this group	Female	Male
GIBS	99,09	91,12	110,44
Akademisches Gymnasium	102,41	98,86	110,19
Borg Birkfeld	100,95	3,86*	106,71
Ubeda	93,39	84,05	96,45
Oakfield School	30,40	19,88*	34,16
St. Stephan Primary	53,63	48,54	57,39
Christ the King	16,71*	3,00*	27,00*
St. Peters	43,00	42,95	43,06

Table 3.10: Average events per session per school

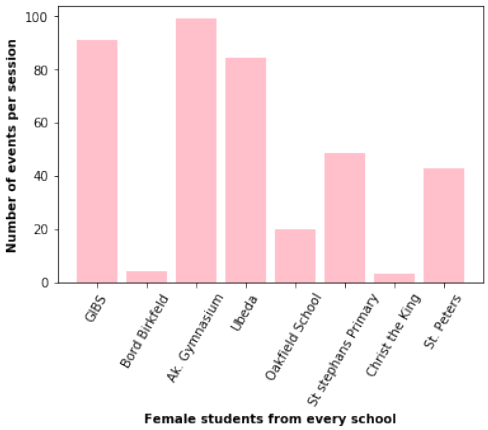

Figure 3.13: Female users from every school

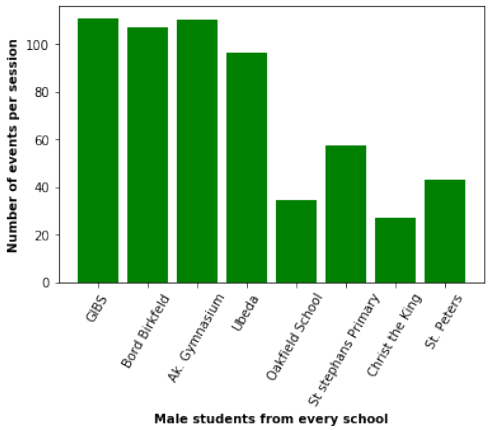

Figure 3.14: Male users from every school

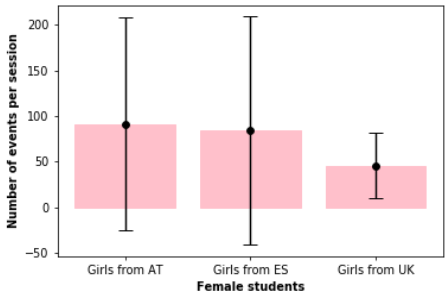

Figure 3.15: Female users from every country with Standard Deviation

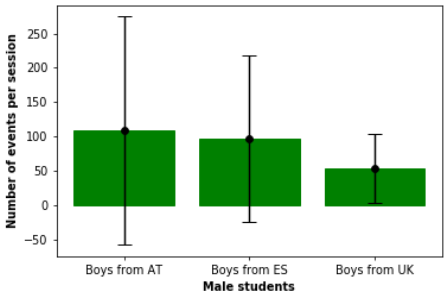

Figure 3.16: Male users from every country with Standard Deviation

test[1]. Such a test can be used for comparing two different groups. Concretely,

[1]https:en.wikipedia.orgwikiStudent's_t-test

for the need of the dataset used for this thesis, comparison can be done between genders, or between each two different countries or between each two different schools. Comparing two groups with the use of this test is done in the way that samples are taken from each group, then averages are compared between the two groups. The mean, however, does not always give enough information to be able to bring significant conclusions.

A lot depends also on the distribution, which by using the t-test can give more information whether there could be significant difference between the samples of the two groups. The test works by trying to determine if the null hypothesis defined as "There is no significant difference between the two samples" holds or not. Then, depending on the critical value this hypothesis is rejected or not rejected. Once the critical value is chosen, it is compared to the estimated value. If it is smaller, the null hypothesis is accepted. Otherwise, the null hypothesis is rejected. Rejecting the null hypothesis means that a significant difference between the two groups is possible.

When applying the t-test, some assumptions are made, such as the assumption that the distribution is normal, the samples have similar variance, and to have roughly the same number of data points in both groups. In order to check this before implementing the test for the distribution were plotted of average number of events for both per session and per user, for all the different groups, in order to check the normality. The results for all students per user are given in the Figures: 3.17,3.18, 3.19, 3.20 as well as the results for female 3.21, 3.22, 3.23, 3.24 and male students 3.25, 3.26, 3.27, 3.28.

The results for all students per session are given in the Figures 3.29, 3.30, 3.31, 3.32, 3.33, 3.34, 3.35, 3.36, 3.37, 3.38, 3.39, 3.40.

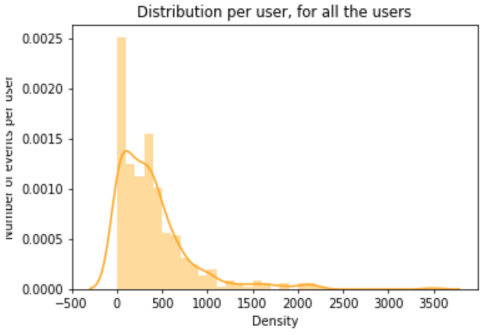

Figure 3.17: Distribution per user for all students

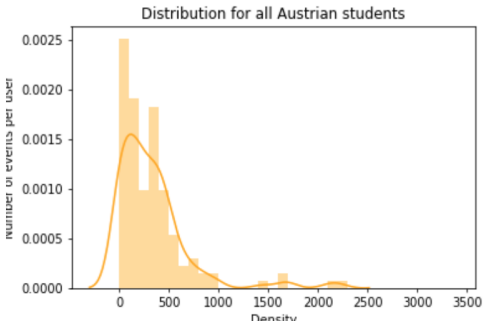

Figure 3.18: Distribution per user for Austrian students

From the results can be concluded that in this dataset the distribution is not normal and because the first requirement is not satisfied the t-test can not be used to find the significance in the differences between the groups.

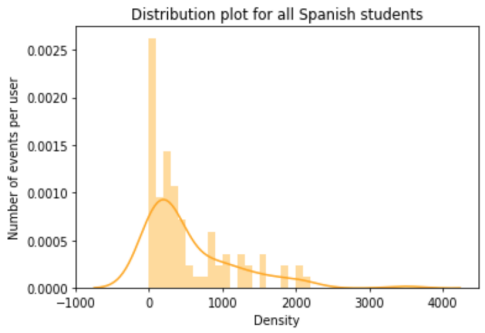

Figure 3.19: Distribution per user for Spanish students

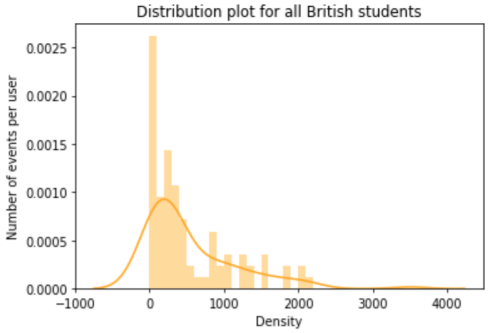

Figure 3.20: Distribution per user for British students

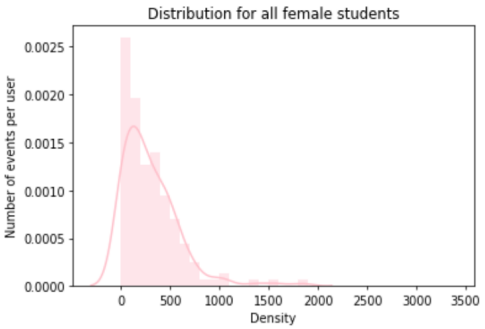

Figure 3.21: Distribution per user for all female students

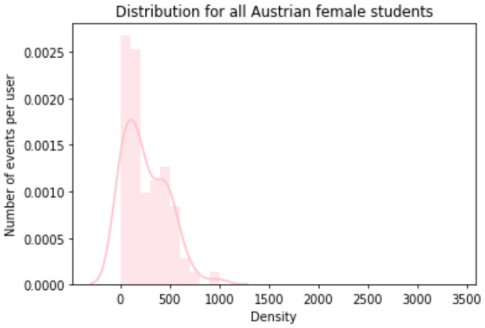

Figure 3.22: Distribution per user for Austrian female students

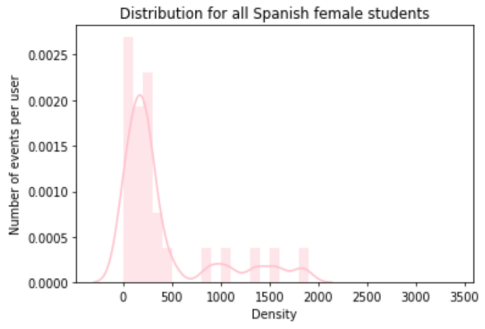

Figure 3.23: Distribution per user for Spanish female students

Figure 3.24: Distribution per user for British female students

In order to get the needed results and to check the differences between different groups, alternative tests were searched that could be used for the given

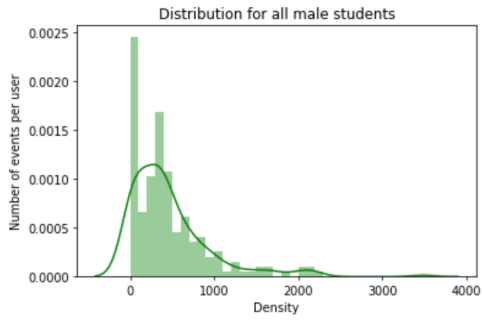

Figure 3.25: Distribution per user for all male students

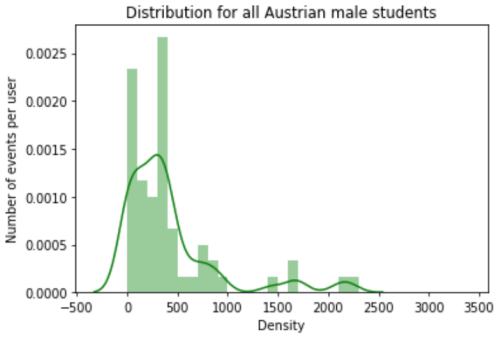

Figure 3.26: Distribution per user for Austrian male students

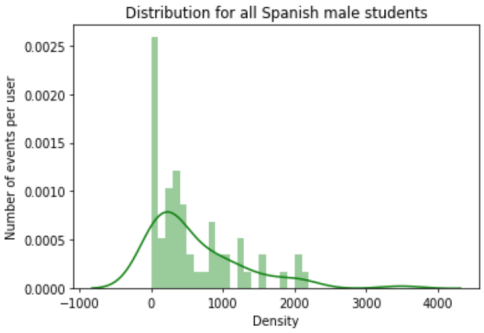

Figure 3.27: Distribution per user for Spanish male students

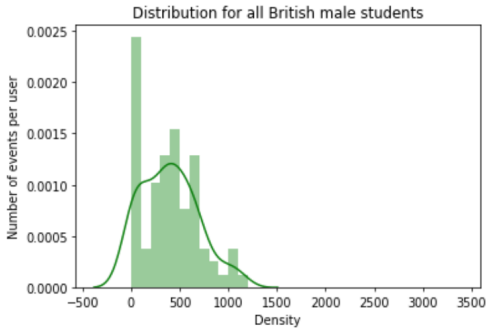

Figure 3.28: Distribution per user for British male students

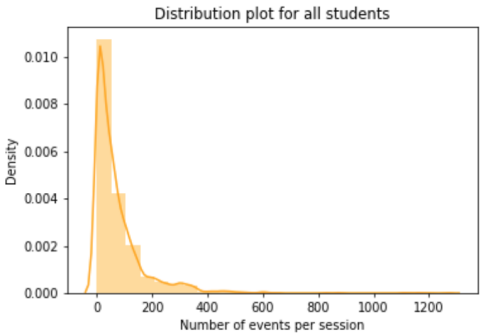

Figure 3.29: Distribution per session for all students

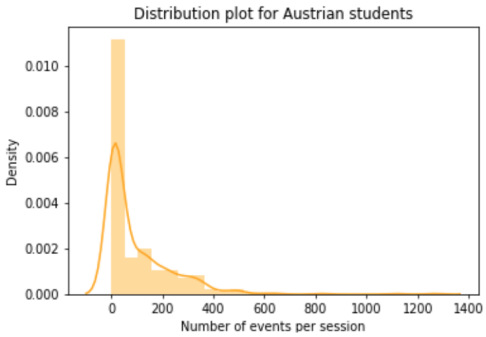

Figure 3.30: Distribution per session for Austrian students

dataset. As a possible alternative the Mann Whitney U-test https:en.wikipedia.orgwikiMann-Whitney_U_test was chosen. The reason for this is that it is a version of a t-test, with the difference that it does not assume normal distribution. This

Figure 3.31: Distribution per session for Spanish students

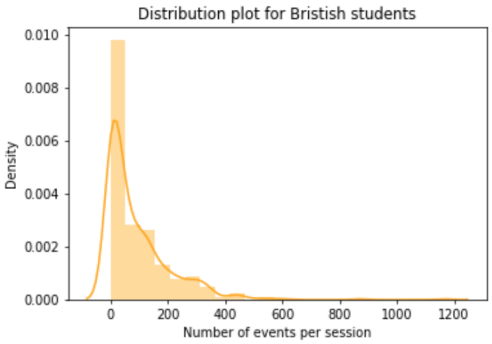

Figure 3.32: Distribution per session for British students

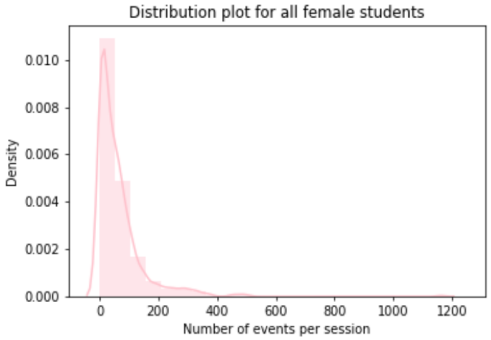

Figure 3.33: Distribution per session for all female students

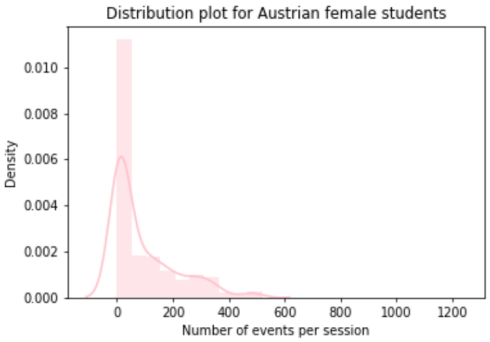

Figure 3.34: Distribution per session for Austrian female students

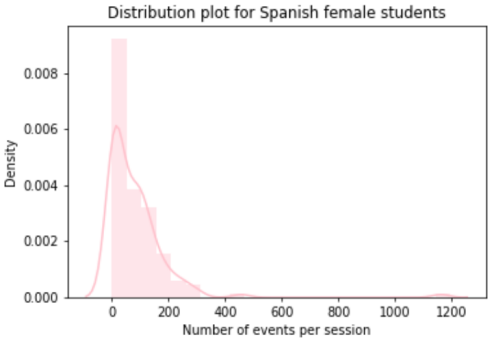

Figure 3.35: Distribution per session for Spanish female students

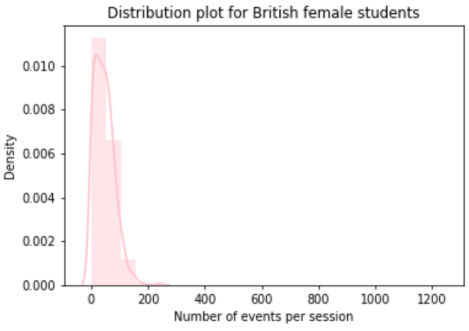

Figure 3.36: Distribution per session for British female students

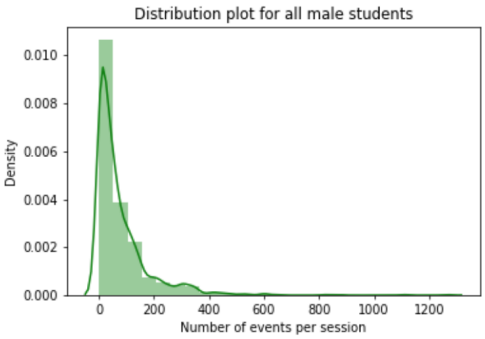

Figure 3.37: Distribution per session for all male students

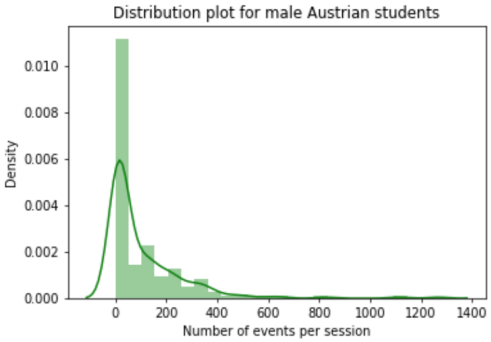

Figure 3.38: Distribution per session for Austrian male students

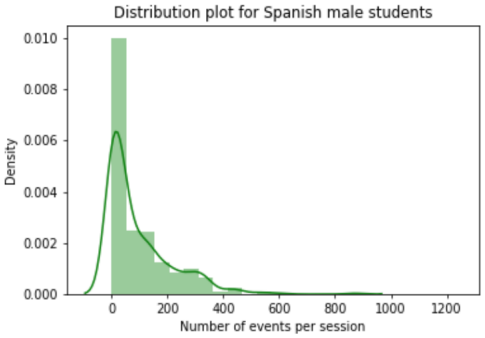

Figure 3.39: Distribution per session for Spanish male students

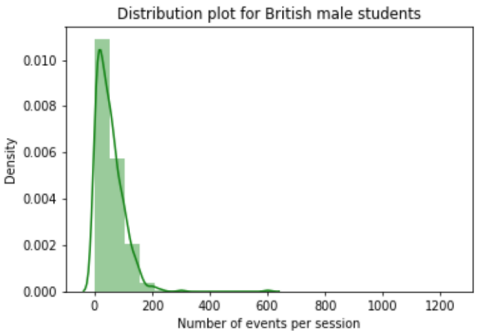

Figure 3.40: Distribution per session for British male students

fulfills the needs of the analyzing dataset. The assumptions for this test are the following:

1. The observations from both groups are independent of each other,
2. it has to be visible which of the two observations is greater,
3. the null hypothesis is defined as - "there is no significant difference between the two groups".

A decision rule follows the z distribution. The sample size has to be more than 20 samples in order this test to work properly and give correct results Nachar (2008), which is always the case in the analyzed dataset for the purpose of this thesis. The smallest group is the one with female students coming from Spain which is exactly 26. This is important to be satisfied,

because it means that the results obtained with this test are correct for the concrete dataset and from them conclusions can be build. It is possible to make one or two tail test. For this dataset two tail test was done. For the z distribution, the value of z is checked in the z-table according to the defined alpha. Z represents the critical value that is important for making the decisions and conclusions. Z needs to be estimated and then can be compared whether the estimated z is greater than the adequate z found in the z-table or smaller than it, in case of smaller the null hypothesis is rejected. For that reason the test statistics are calculated. For the test statistics rank, score, definition from which samples and points are defined. With the use of the equation for z, defined for the need of this test, the estimated z can be obtained. This z, as explained previously, is compared and with that can be decided whether for the two compared groups the null hypothesis should be rejected or accepted, by using the statistical test. Additionally, the p-value is also part of the results of this test, which shows whether the difference between the two compared groups is significant or not. Because whether there is a difference or not can be immediately concluded from the average events per session or per user, but how much significance there is in the difference can be concluded from the p-value estimated with the use of this test. The results, (the p-values) for the concrete dataset are presented with confusion matrix, for the differences between countries per session is given on Figure 3.41 and genders per session between countries on Figure 3.42, and for the differences between schools per session 3.43 and genders between schools per session is given on Figure 3.44.

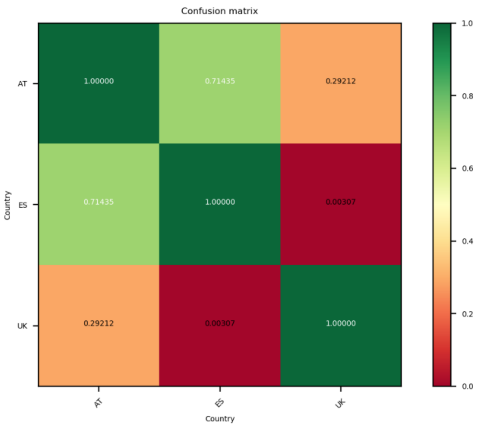

Figure 3.41: Comparing the users between countries

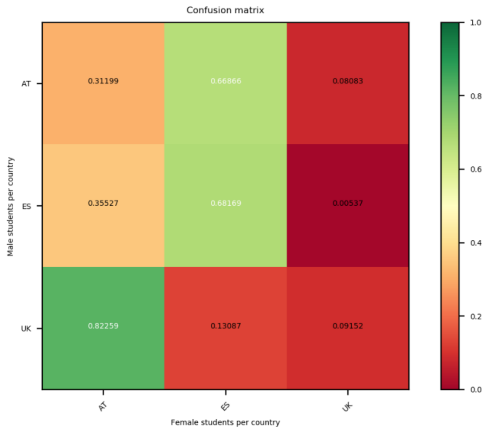

Figure 3.42: Comparing genders between countries

Comparing the students from Austria and Spain in confusion matrix 3.41

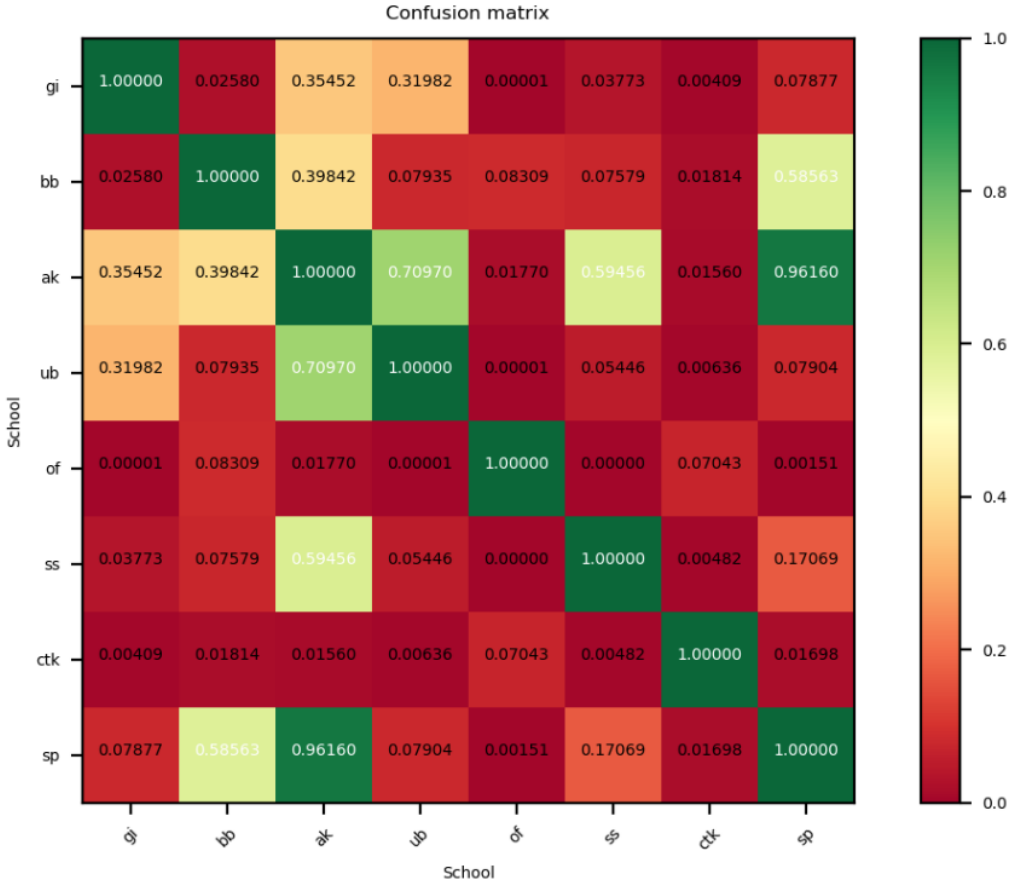

Figure 3.43: Comparing the users between schools

results in a p-value of 0,71435. Therefore, the Mann Whitney U-test showed that there is a different, but it is insignificant, hence it is marked as green in the confusion matrix. Comparing the students from Austria and United Kingdom, shows a greater difference. The p-value is not very small (0,29212). From this the conclusion can be made that there is bigger difference between the students from Austria and United Kingdom compared to the students from Austria and Spain. Finally, when British and Spanish students are compared, the p-value is very small (0,00307). This means that the difference between this two countries is significant when comparing average number of events per session, hence is marked with red on the confusion matrix. It is difficult to make a conclusion from this, since every country has a different number of schools that were tracked in the project, which means that the number of data samples is different. This shows the need for splitting into smaller groups for better comparison and more accurate results. A possibility

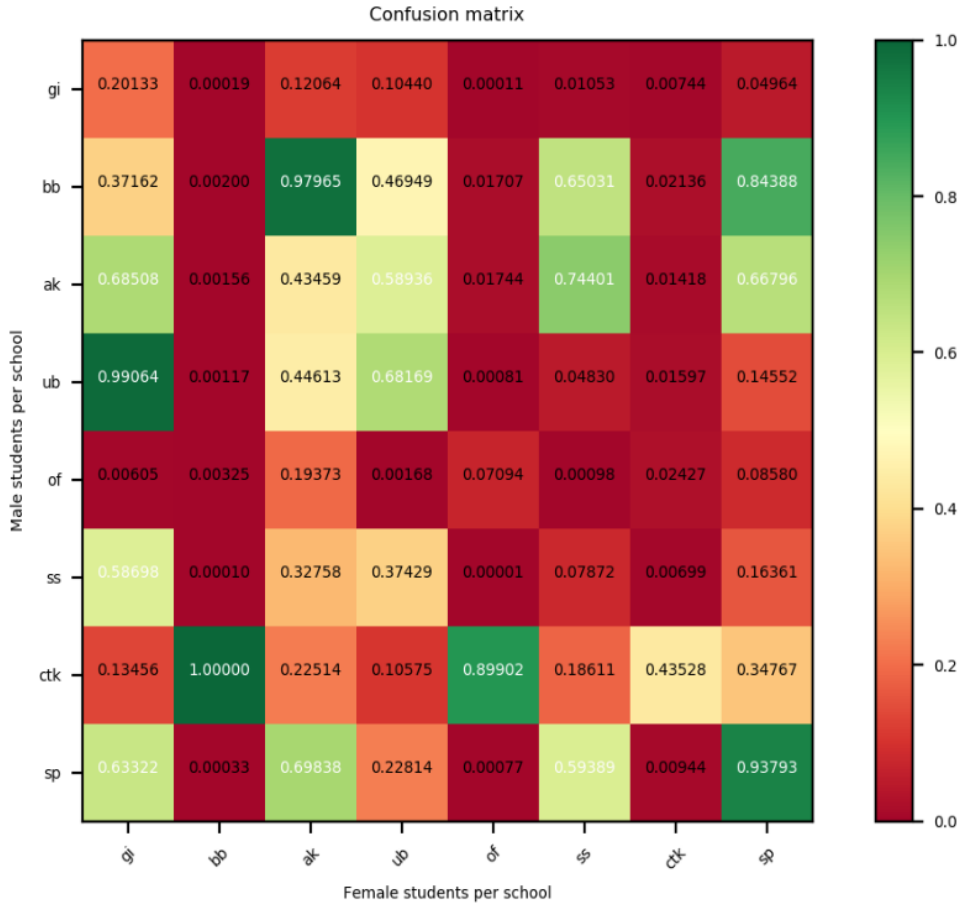

Figure 3.44: Comparing the users between genders between schools

would be to compare the difference between the genders from each of the countries. Another possibility to have even smaller groups is to make a comparison between each of the schools or even between the genders in the schools.

Figure 3.42 also shows a comparison between countries, but this time compares female students (X-axis) of each country to each group of male students (Y-axis) per country. When comparing the male and the female students from Austria, the p-value is 0,31199, which means that the difference exist but it is not very significant. The difference is even smaller when comparing the male and female students from Spain, with greater p-value=0,68169. But, the difference is significant when comparing the male and female students from United Kingdom with p-value=0,09152. Additionally, strong similarity is visible between the male students from United

Kingdom and the female students from Austria. The same conclusion for similarity between the male students from Austria and the female students from Spain can be made. On the other hand, significant differences can be seen between the male students from Spain and female students from United Kingdom, as well as the other way around, between the male students from United Kingdom and the female students from Spain.

A comparison is made between the 8 different schools, again by using the Mann Whitney U test. The results of this comparison are presented in 2 different confusion matrices. Figure 3.43 gives the results between all the different schools and figure 3.44 shows the differences between the genders in all 8 different schools. As in the previous example, p-values are colored according to their significance. Red, yellow/orange and green represent significant differences, differences that are not significant and similarity, respectively. Finally, the ones marked with green have very big p-value, which means the similarity between the compared two groups is large. The conclusion can be made that there is a significant difference between the following pairs of schools:

- Borg Birkfeld (AT) and GIBS(AT)
- Borg Birkfeld (AT) and Christ the King (UK)
- Akademisches Gymnasium (AT) and Oakfield School (UK)
- Akademisches Gymnasium (AT) and Christ the King (UK)
- Ubeda (ES) and Oakfield School (UK)
- Ubeda (ES) and Christ the King (UK)
- Oakfield School (UK) and GIBS (AT)
- Oakfield School (UK) and St. Peters (UK)
- St Stephens Primary (UK) and Christ the King (UK)
- Christ the King (UK) and GIBS (AT)
- Christ the King (UK) and St. Peters (UK)

Additionally, there is a lower significant difference (higher p-value) between the following pairs. These are marked with brighter red in figure 3.43.

- Borg Birkfeld (AT) and Ubeda (ES)
- Borg Birkfeld (AT) and Oakfield School (UK)

- Borg Birkfeld (AT) and St Stephens Primary (UK)
- Ubeda (ES) and St Stephens Primary (UK)
- Ubeda (ES) and St. Peters (UK)
- Oakfield School (UK) and Christ the King (UK)
- St Stephens Primary (UK) and GIBS (AT)
- St Stephens Primary (UK) and St. Peters (UK)

Finally, between the following pairs of schools there is high similarity:

- Akademisches Gymnasiuim (AT) and Ubeda (ES)
- Akademisches Gymnasiuim (AT) and St Stephens Primary (UK)
- St Stephens Primary (UK) and St. Peters (UK).

The reason for the visible differences between schools can be caused of different experimental setup of the different NOLB teams, which is known in details only for the Austrian schools. Other conditions that can influence is the previous experience with Pocket Code or similar applications and finally the experience of the teacher in programming applications and creating games and the way they choose to explain it to the children. Finally the cause can also be different subject in which the application was presented because on that depended the task that the children received, which was different for all of them.

On one hand there is big similarity between the one Austrian school in respect to the two other Austrian school. The similarity between Akademisches Gymnasium (AT) and GIBS (AT) can result of the fact that both started with templates. Then, the reason for the similarity between Akademisches Gymnasium (AT) and Borg Birkfeld (AT) can be the fact that the students from Akademisches Gymnasium (AT) already had experience in Pocket Code. Akademisches Gymnasium (AT) used this application before the project, while the children from Borg Birkfeld (AT) did not have it at the very beginning, but rather use it in other subjects after the start of the NOLB project, which means that they also did most of their tasks with additional pre-experience.

On the other hand there is significant difference between the two Austrian schools Borg Birkfeld and GIBS. The reason for this difference can be result of a few factors that were different for this two groups: firstly in Borg Birkfeld the teacher had previous experience with Pocket Code and this school did

not need the team of the NOLB to give instructions to the students. The children there started learning how to use Pocket code by using tutorial cards made by their teacher and their goal was to create a game. They all submitted one game for the Galaxy Game Jam competition[2], while the students had no prior knowledge for Pocket Code or similar application. In GIBS they also received explanation from their teacher, but the teacher was not experienced in using Pocket Code. They also received help from the NOLB team when they were explaining to the children, which is not the case for Borg Birkfeld. GIBS were first practising building games on paper. Another difference is that the students from GIBS had pocket code during their arts subject and gained more experience during the project compared to the students from Borg Birkfeld. Another difference is that the students from Borg Birkfeld got the task to create games while the students from GIBS were encouraged to use templates, to modify, to play and add scenes. Also for this confusion matrix (figure 3.43) we should take into account that these results are estimated based on the average events per user in the first step. So, it can be expected that a confusion matrix calculated per session per user will look different from this confusion matrix. When considering average events per user, it contains all sessions combined, while the user may evolve in its usage of the application over time in different session. When looking at individual sessions, the evolution and differences between those sessions will be easier to spot.

For the schools outside of Austria there is no detailed information for the experimental set up which is very probable to reflect in the results. What can be seen from the data is which students used the help button, which students were more reusing already made events as well as which users were using complex or hidden events considered to be advanced actions. From table 3.13 for advanced events, it can be concluded that the students from Ubeda (ES) used the advanced events more than almost every school, with the one exception of St Stephens Primary (UK). This is very different for children from Christ the King who never used advanced events and St. Peters (UK) where the advanced events were used by one student only one time.

When comparing countries, all 4 schools from UK are different between each other according to the results from the confusion matrix in figure 3.43. The same is the case for Austria. This should also be taken into account when looking at the results between countries since Spain has only results from one school and therefore does not containing different performances and different experimental setups or classes.

In the confusion matrix in figure 3.44, genders are also taken into account.

[2]www.galaxygamejam.com

It can be concluded that the female students from Borg Birkfeld (AT), Oakfield School (UK) and Christ the King (UK) are very different from the male students from 7 schools, which only excludes the boys from Christ the King (UK). The rest of the schools generally show significant difference with at most 3 other schools and less significant difference (greater p-value) with the rest of the schools.

Now the same comparison is done, but instead of taking the average number of events per user, the average number of per session is used. From figure 3.45 can be concluded that there is a significant difference between the Austrian and Spanish students and between Austrian and British students. There is no notable difference between British and Spanish students. Which is very different from the previous results is that there is also a high similarity between Austrian and Spanish students.

From figure 3.46 can be concluded that there is significant difference between male Spanish users and female Austrian, male Spanish students and British female, and also between British male students and Austrian female students. An interesting result is that there is no negligible difference between the different genders of all the three countries.

From the results given in the confusion matrix 3.47, it can be concluded that the three Austrian schools show that they are similar. This is an expected result when taking into account that all three schools are from the same city (Graz) and had support from the same team from the project NOLB during the tracking period. There was difference in the previous confusion matrix 3.43 between the Akademisches Gymnasium and GIBS. However, these results, there are similarities between any two Austrian schools. The biggest difference in this results when comparing the Austrian schools is between Akademisches Gymnasium and Borg Birkfeld with p=0,15392, which is not remarkably different.

Then Ubeda (ES) shows only similarity with Borg Birkfeld and significant difference when compared with every other school. The reason for this might be that the Spanish users used a lot of "copy" events. Like for example they copied a lot of looks, sounds, objects and programs as it can be seen in Table 3.12, or to be precise 1008 which is outstandingly more than any other school with the only exception for the St. Stephens Primary, where copied bricks were used 1408, but this school also tracked 4 more students.

Christ the King (UK) shows significant difference with any other school. This is also an expected result when taking into account that each of the students from this school only did one single session. For all of other schools all the sessions are taken into consideration, and mostly for later sessions it means more events used as shown in the tables: 3.3, 3.4, 3.5, 3.6.

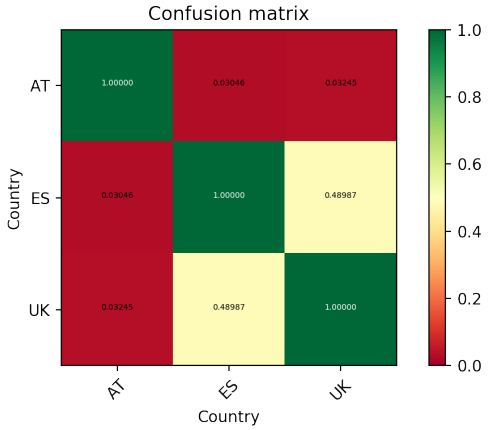

Figure 3.45: Comparing the users between countries per user

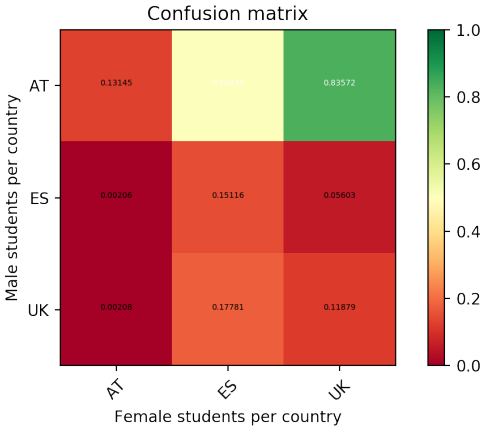

Figure 3.46: Comparing genders between countries per user

In the results there is similarity between St. Peters school (UK) with Borg Birkfeld (AT) and Oakfield (UK). The schools St. Peters school (UK) and Oakfield (UK) have comparable number of students. The similarities with Borg Birkfeld (AT) except the numbers of events session are that both schools have similar number of students who participated in this project (26 and 29), additionally the female students from both schools did not use any advance events, did not copy events and did not use the "help" buttons, which might be the reason for the same amount of events per session they used.

From the results given in the confusion matrix 3.48, can be concluded that the male and female students from Christ the King (UK) are similar defined by the p-value=0,43528 as well as the male and the female students from Akademisches Gymnasium (AT) with p-value=0,22815. Rather bigger difference can be seen between the male and female students from GIBS (AT) with p-value=0,14952, then is Ubeda (ES) with p-value=0,15116, Oakfield (UK) with p-value=0,14323 and St. Stephens Primary with p-value=0,11219. The most remarkable difference is between the male and the female students from Borg Birkfeld (AT) with p-value=0,00641. This significant difference can be a result mostly because of the fact that there were only 6 female students tracked during the project and 23 male students from Borg Birkfeld. The girls have in average only 3,86 events, whereas boys have 106,71 as shown in the table 3.10. Another difference is that the male students used "copy" events, they reused already existing Program, Look, Sound, exactly 102 times, and female students never used this type of events. The notable difference between both genders from Christ the King (UK) and any gender

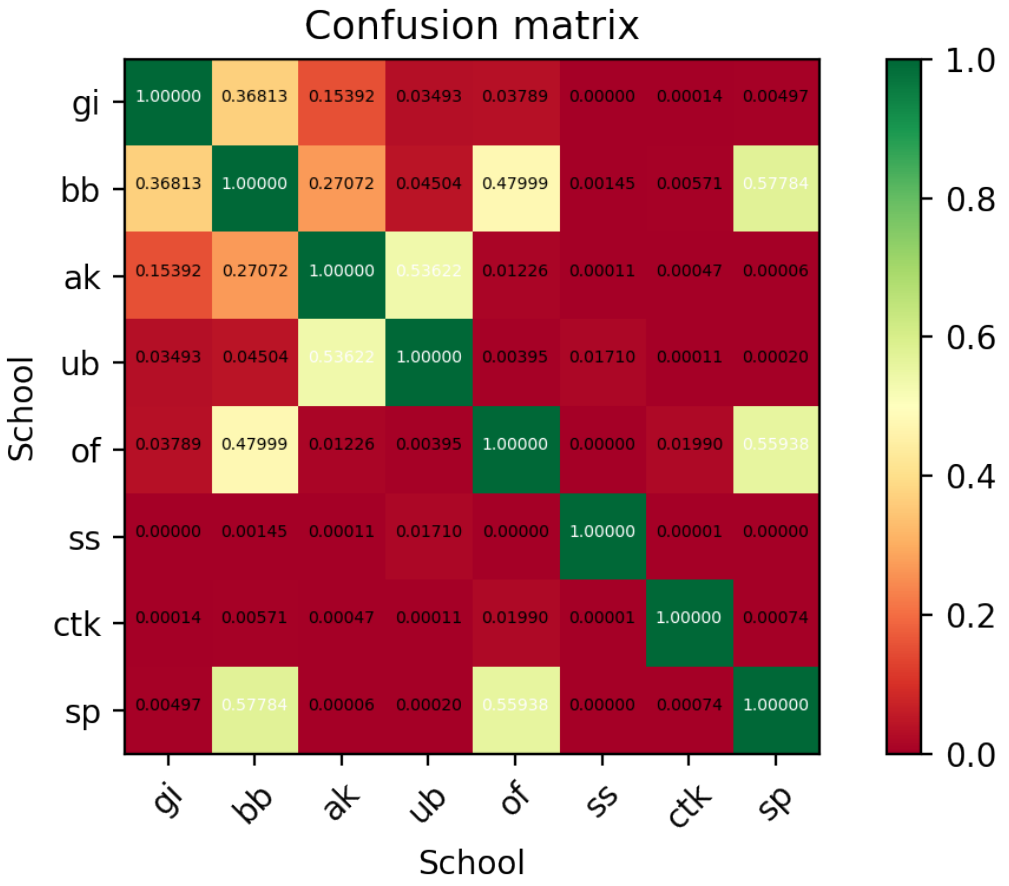

Figure 3.47: Comparing the users between schools per user

from any other school can be because of the very low number of students tracked. Only 7 in total of which 3 female and 4 male students, each of them with only one session tracked.

3.2.2 Pattern Mining with PrefixSpan

Except the average number of events per session and average number of events per user, the most common events were also taken into account when searching for patterns. Moreover, the sequence in which the most common events were used was also included in the results. In order to find this information from the dataset, first the Generalized Sequential Pattern (GSP) algorithm was used. This was giving results when comparing small groups, such different genders in the same school, but was very inefficient to be used for the entire dataset. This is the reason why another method for

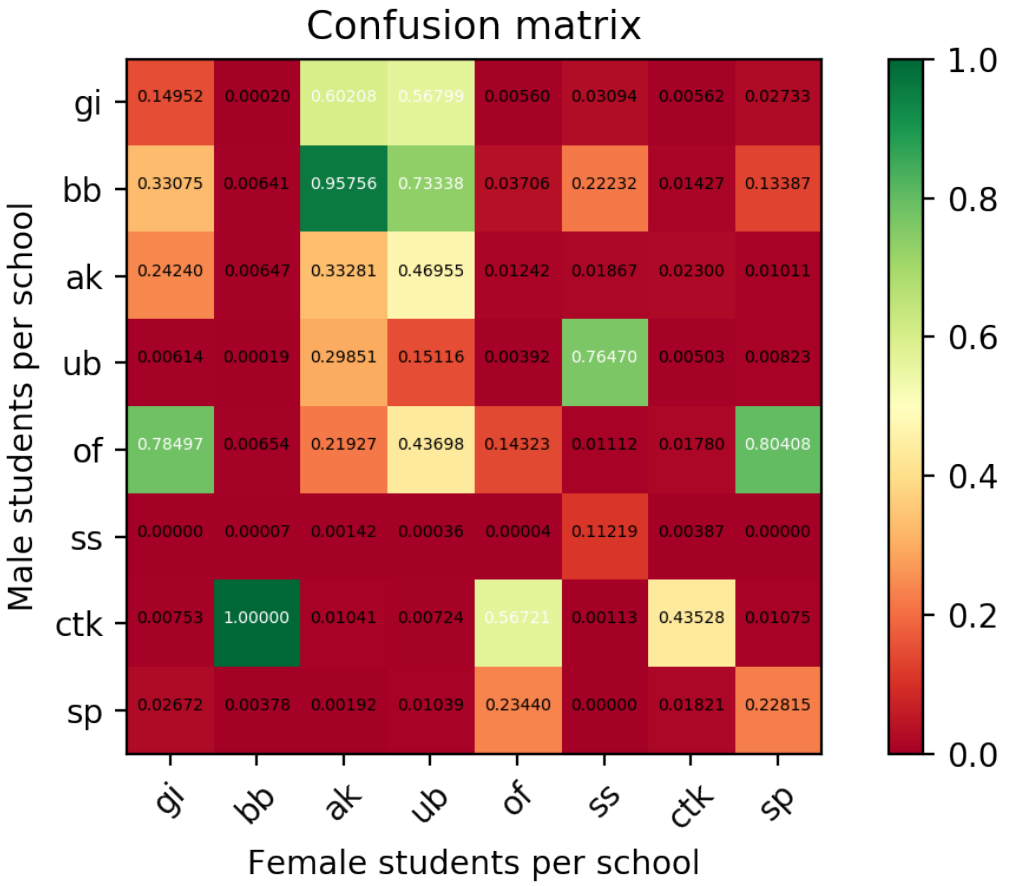

Figure 3.48: Comparing the users between genders between schools per user

sequence mining was searched as suggested in one of the papers described in the relevant literature in chapter 2, where it is shown that PrefixSpan outperforms GSP. This is the reason why PrefixSpan was used for the purpose of sequence pattern mining. The way PrefixSpan works reduces the number of candidates in a more efficient way. The frequent events are defined as prefixes and the entire search space is divided in the prefix-projected databases and their projections are then defined as suffixes (prefix based projections). This approach reduces the number of projections needed to find the most frequent sequences. The most used sequence of events was searched by limiting the PrefixSpan to the top 70 most used events by all of the users because of computing reasons. It was also tried with top 80 and top 90 and gave the same results.

According to the results of the PrefixSpan sequence mining, the most common events (sequences of length 1) per session for all of the users who

are part of the dataset are the following: “openProgram”, “stopProgramExecutionSession”, “startProgramExecutionSession”, as shown on 3.49. The “openProgram” appears in 1632 sessions in total, which is 84,60% of all the sessions. It is part of the main menu and the “program name” field can be customized by using this brick. This event appears the most frequent because it has 2 use cases when it is used. The first use case is as mentioned before from the main menu, this is done when the user sees the list of already created or downloaded programs and chooses one of them. In this case the “openProgram” event will be created. Another use case is that “openProgram” is called in combination with “ContinueButton”, whenever the user clicks on the “ContinueButton” then the ”open Program” is automatically created as next event. According to the Figure 3.49, this means that “openProgram” is used because of the first use case 372 times and 1260 (65,32% of all the sessions) times because of the second use case (“ContinueButton”). The ”ContinueButton” is also part of the main menu of Pocket Code, it appears when the user did not log out after the last time playing and it offers to continue with the same game played previously. Another reason why this event can be created is because of the known bugs, which means that if the application crashes then after refreshing the user is back at the main menu and has to click the ”ContinueButton”. After clicking the ”startProgramExecutionSession” the user is staring to play, and the time between the ”startProgramExecutionSession” and ”stopProgramExecutionSession” is the actual time that the user spent playing. The event ”startProgramExecutionSession” is used 1580 times by all of the users over all sessions, which makes it 81,91% of all sessions. The event ”stopProgramExecutionSession” is used 1535 times in total and that is 79,57% of all the sessions. Built on this prefixes sequences of length 2, 3, 4 and 5 for all users are found, as shown in figures 3.50, 3.51, 3.52, 3.53, respectively. Normally, the event ”startProgramExecutionSession” should be followed by the event ”stopProgramExecutionSession”. The results show ”stopProgramExecutionSession” followed by again the same event ”stopProgramExecutionSession”. This might be result of an error in the tracking of the events. Namely, the users could pause the program and start it again. Very likely in this situations the pause was tracked as ”stopProgramExecutionSession” and the restart is not tracked as ”startProgramExecutionSession”. However this is not explaining the sequential use of ”stopProgramExecutionSession” one after another if both the pause and restart click are not mistakenly tracked as ”stopProgramExecutionSession”. More frequent usage of the ”startProgramExecutionSession” and the ”stopProgramExecutionSession” events shows that there were several different attempts of the user in one session, which was encouraged and shows persistence from the students in playing the game or creating the pro-

gram, with the application.

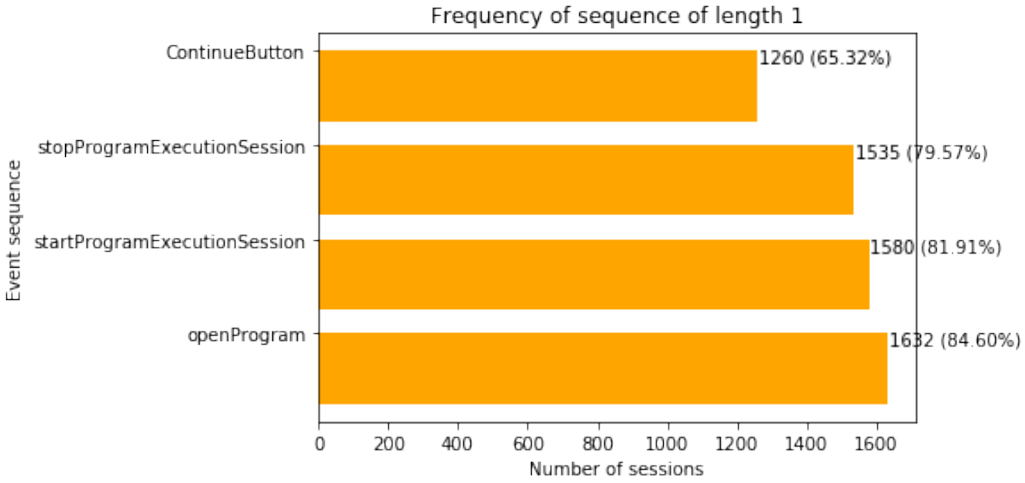

Figure 3.49: Sequences of length 1 in a session for all users with PrefixSpan

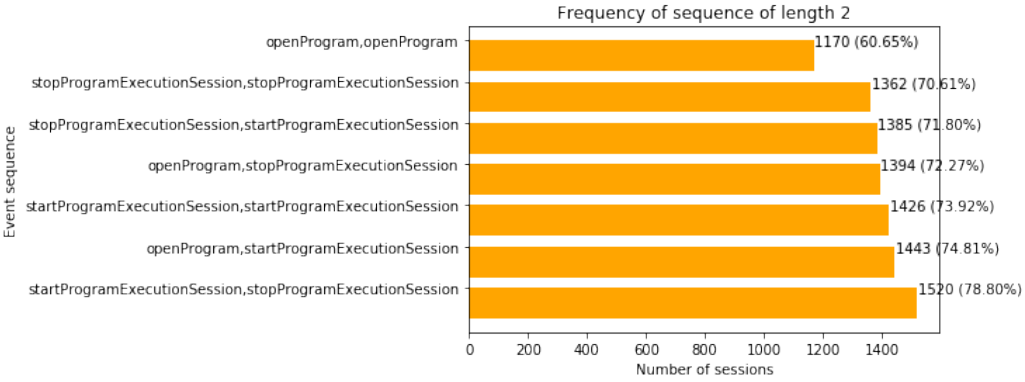

Figure 3.50: Sequences of length 2 in a session for all users with PrefixSpan

The most common used events with sequences of length 1 for female students including all of the three countries together, are the following: "startPocketPaintSessionEditLook", "ContinueButton", "openProgram", "stopProgramExecutionSession", "startProgramExecutionSession", as shown on the Figure 3.54. Only the group of girls uses the event "startPocketPaintSessionEditLook" frequently, unlike the general group or the group of boys. With this event the users can use already created object and edit its look in Pocket Paint. This event was used in 450 sessions, which is 60,89% of the sessions recorded for female users. The event "openProgram" is used in 600

Figure 3.51: Sequences of length 3 in a session for all users with PrefixSpan

Figure 3.52: Sequences of length 4 in a session for all users with PrefixSpan

Figure 3.53: Sequences of length 5 in a session for all users with PrefixSpan

sessions which is in 81,19% of the session for girls, from which 464 (62,79%) as a result of the "ContinueButton".

For brevity, only the sessions of length 1 and length 5 are included for all male and female users. The plots for other lengths are added at the end of the thesis in the appendix A. When the results of sequences of length 5 are compared to the sequences for all users, the difference is that the female students have only one sequence that is commonly used and that starts with the event "openProgram", whereas in the most common events for all the students there are three commonly used sequences starting with this event.

The most common used events (sequences of length 1) for male students including all of the three countries together, are the following: "ContinueBut-

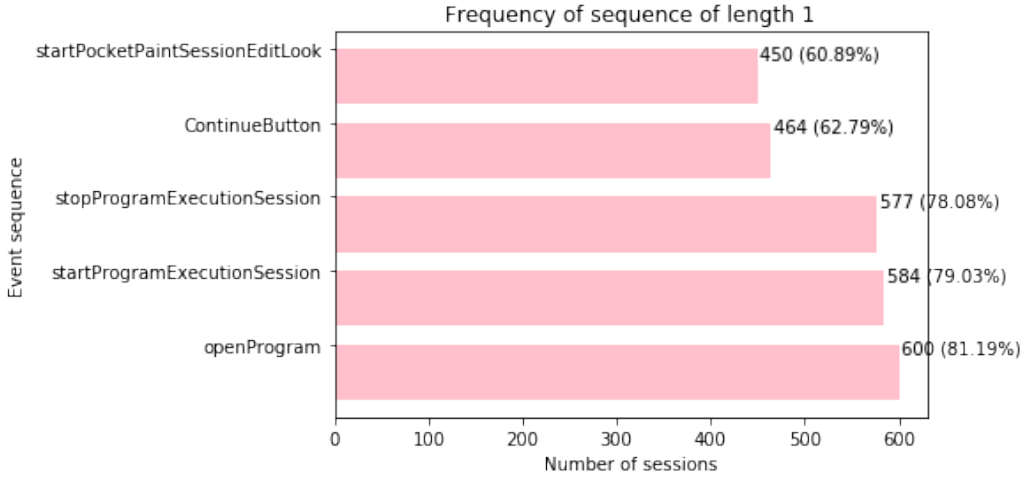

Figure 3.54: Sequences of length 1 in a session for all female users with PrefixSpan

Figure 3.55: Sequences of length 5 in a session for all female users with PrefixSpan

ton", "openProgram", "stopProgramExecutionSession", "startProgramExecutionSession", as shown on the Figure 3.56. The difference when comparing with the girls is that the event "startPocketPaintSessionEditLook" is not part of the most common events for boys. In this plot can be seen that exactly the same 4 common general events are also the most commonly used for boys. The event "openProgram" was used in 1032 sessions, from which 796 as a result of the "ContinueButton".

After that PrefixSpan was used to find the most common sequences from every country for each gender. Finally the most common sequences in every school, separately for every school, and for both male and female from every school are calculated.

For all of the students from Austria, as shown on figure 3.58 there are 3 events that were also commonly used in general for all of the students and are also commonly used by the group of all Austrian students. These are the

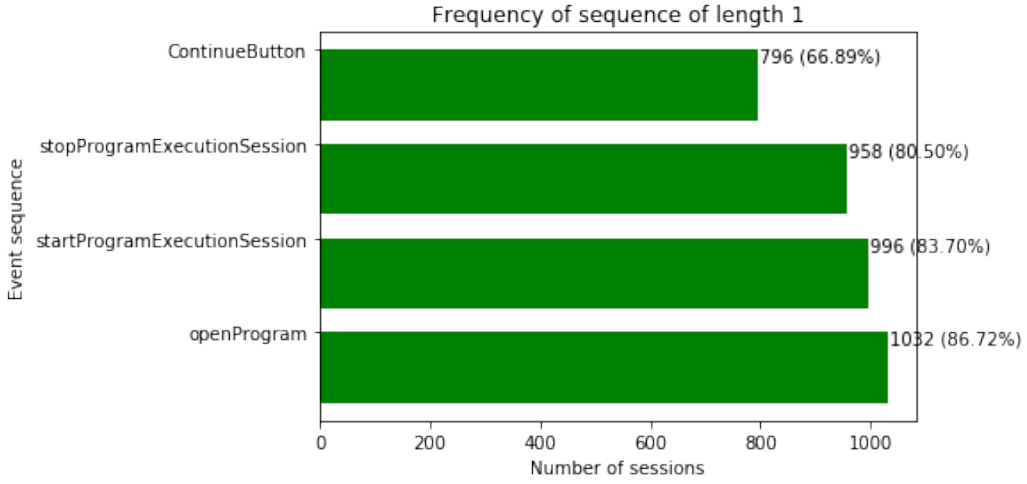

Figure 3.56: Sequences of length 1 in a session for all male users with PrefixSpan

Figure 3.57: Sequences of length 5 in a session for all male users with PrefixSpan

events: "startProgramExecutionSession", "ContinueButton" and "openProgram". The event that is not part of the commonly used sequences of length 1 for all users but is commonly used sequence of length 1 for Austrian students is the "init_session". This event is created when the user logs in. This might be an indicator that Austrian users were logging in each time they started a new session as it is supposed to be done by all of the users. The event that is frequently used by the general group but not by the Austrian users is "stopProgramExecutionSession".

When comparing the most frequently used sequences of length 5 from all students and most common used sequences of length 5 from Austrian students in figure 3.59, one noticeable difference is that the Austrian students have common sequence beginning with the event "ContinueButton", which is not the case for all the students in general. All of the sequences that do not start with "ContinueButton" or "openProgram" are starting with "startProgramExecutionSession" and none of them are starting with "stopProgramExecutionSession", which is not the case for sequences of length 5

for all the users where 4 sequences start with that event.

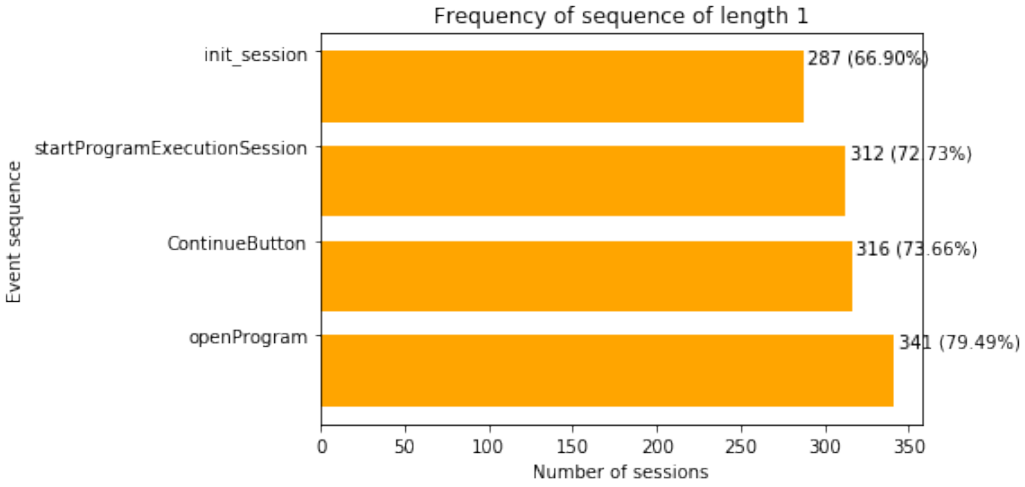

Figure 3.58: Sequences of length 1 in a session for all Austrian users with PrefixSpan

Figure 3.59: Sequences of length 5 in a session for all Austrian users with PrefixSpan

The sequences of length 1 for Austrian female students, are given in figure 3.60. The event "init_session" is present, as expected and as in the case for frequent sequences of length 1 for all Austrian students in general. They used the event "init_session" in 111 sessions which is 73,10% from qll the sessions. Additionally, there is the event "startPocketPaintSessionEditLook", which was previously not part of the sequences of length 1 for all users, but it was part of the sequences of 1 for all female users.

When comparing the sequence of events with length 5, on Figure 3.61 can be noticed that there is sequence starting with the event "init_session", which was not the case before neither in the case for all students nor in the case for all Austrian students.

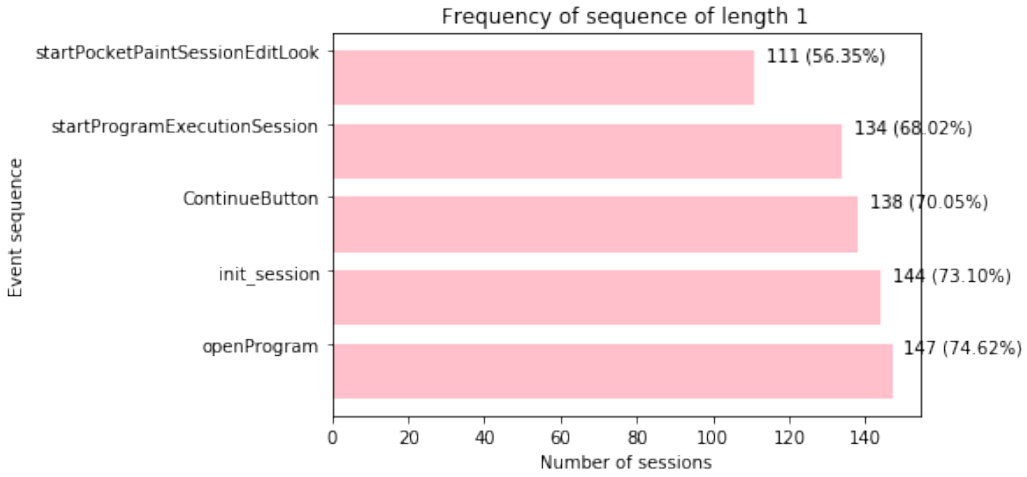

Figure 3.60: Sequences of length 1 in a session for all Austrian female users with PrefixSpan

Figure 3.61: Sequences of length 5 in a session for all Austrian female users with PrefixSpan

For Austrian male students as given on Figure 3.62, the only difference with Austrian female students is that the male students were not using the event "startPocketPaintSessionEditLook". But they did used the "init_session", which is similar behaviour with the girls from Austria. They used it in 143 session, which is in percentage less than the girls, or more precisely they used it in 61,64% of all the sessions.

For the sequences of length 5 for the male Austrian students, as shown on 3.63, there are even 5 sequences starting with the event "openProgram", one sequence with "ContinueButton" and all the rest with "startProgramExecutionSession".

From the most frequent sequences of length 1 for all Spanish students, as shown on 3.64 4 events are common sequences of length 1 for all of the users too and 3 sequences that are common sequences of length 1 for Austrian users. What is different compared to all of the results presented for sequence of 1

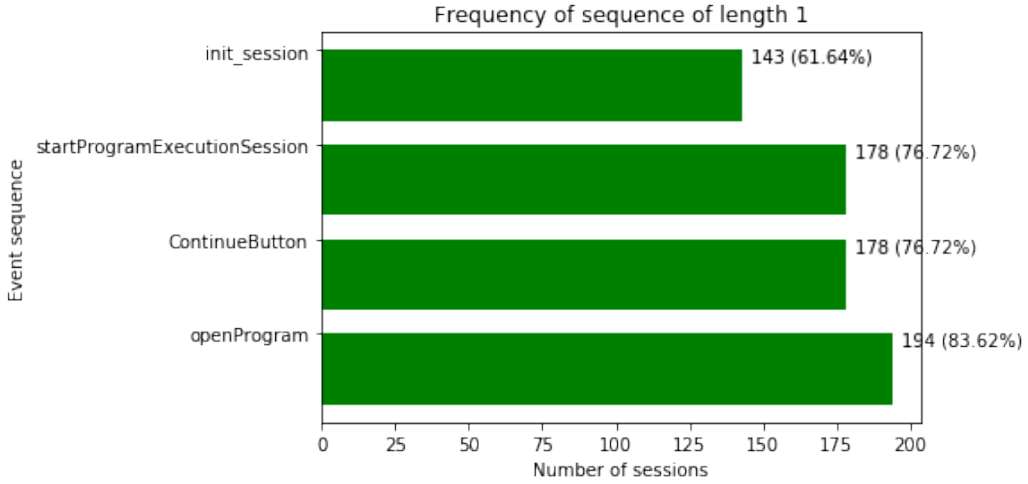

Figure 3.62: Sequences of length 1 in a session for all Austrian male users with PrefixSpan

Figure 3.63: Sequences of length 5 in a session for all Austrian male users with PrefixSpan

is the commonly used event "openFormulaEditor", which is interesting and unique result. This event can be used to calculate mathematical formulas. This event was used in 279 sessions, or in 55,03%. Another difference with the Austrian users is that Spanish users have in the sequence of length 1 the event "stopProgramExecutionSession". On the other hand, the event "init_session", which is used by the Austrian students is not in the most frequently used sequences of length 1 for the Spanish students.

For most common sequences of length 5 as shown in the Figure 3.65, there are 2 sequences of length 5 which start with the event "stopProgramExecutionSession", which was part of the results for all student but not for Austrian students. There is only one sequence starting with the event "openProgram" and all the rest with "startProgramExecutionSession". There are 19 common sequences of length 5 for the Spanish users, which is more than Austrian students (15).

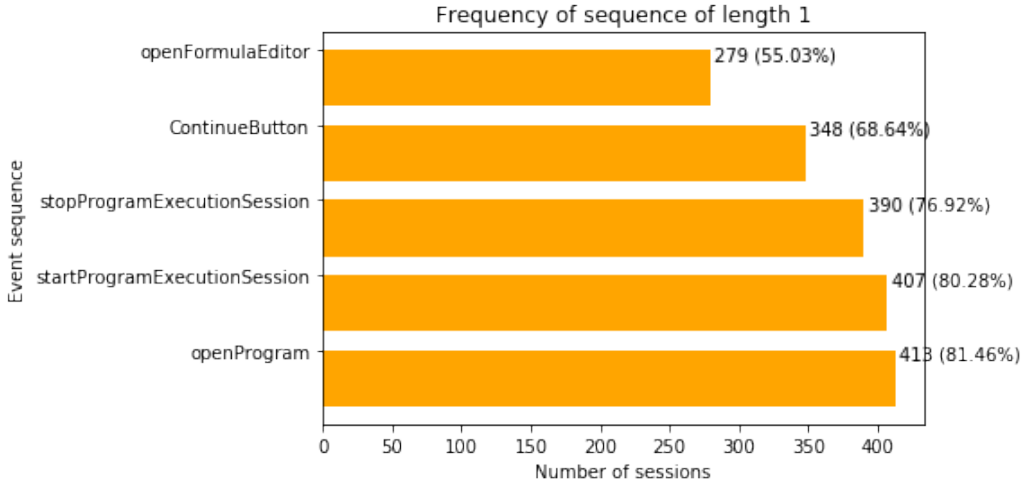

Figure 3.64: Sequences of length 1 in a session for all Spanish users with PrefixSpan

Figure 3.65: Sequences of length 5 in a session for all Spanish users with PrefixSpan

For the female Spanish users for sequence of length 1 the results can be seen on the Figure 3.66. Compared to the female Austrian students, they have 3 events in common. Sequences of length 1 that are used by Spanish female users and were not used by the Austrian students are: "addBrick" and "dropBrick" used in 66 sequences (52,80%), and "openFormulaEditor" used in 73 sequences (58,40%). The event "addBrick" is created each time a new brick is added to the script. The event is part of the group of so called "creating events". By using this event the user defines the name of the brick and the category of the brick. The category can be "Event", "Control", "Motion", "Sound", "Looks" or "Data". Depending on the choice of category the name can be, for example for "Event": "whenStartedBrick", "whenTouchedBrick", "whenScreenTouched", "broadcaseBrick". When the category "Control" is selected the BrickName can be: "WaitBrick", "CloneBrick", "ForeverBrick", "WhenCloneBrick", "DeleteThisCloneBrick". For

the other possible categories there is also a limited number of brick names that users can use. The "dropBrick" event is created automatically the same amount of times the users added a new brick.

The most common sequences of length 5 for Spanish female users are shown on the Figure 3.67. All of the sequences start with the events "startProgramExecutionSession" and "stopProgramExecutionSession". The difference with the female Austrian students is that Spanish students do not have sequences of length 5 starting with "ContinueButton", "init_session" or "openProgram".

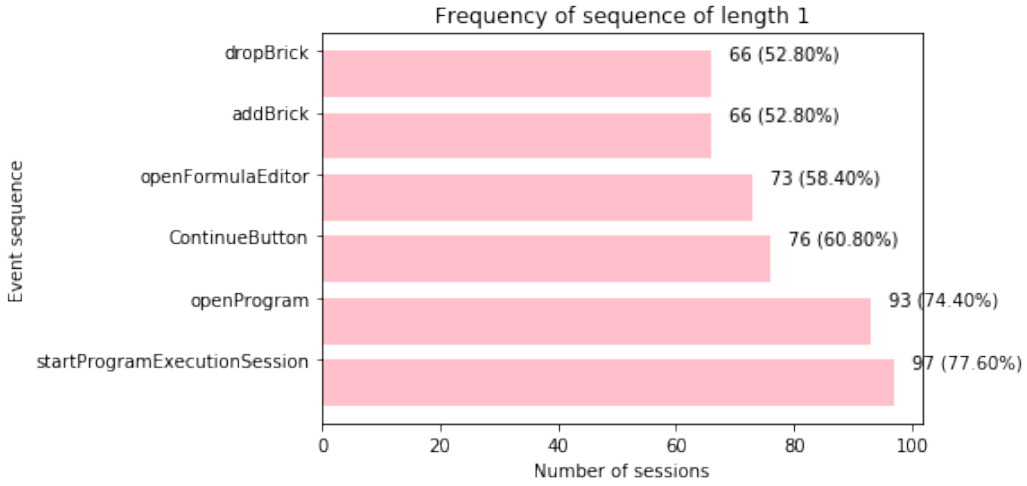

Figure 3.66: Sequences of length 1 in a session for all Spanish female users with PrefixSpan

Figure 3.67: Sequences of length 5 in a session for all Spanish female users with PrefixSpan

In the sequences of length 1 for male Spanish users, shown in the Figure 3.68, compared to the sequences of length 1 for the Austrian students shows only one difference. The Spanish male users did not log out as often after their first session, hence they did not create too many "init_session" events

in order to log in again since they were already logged in from the previous time, in contrast to the Austrian male students who logged in 61,64% of the times. Another difference is that the Spanish male users used the "stopProgramExecutionSession" event more compared to Austrian.

If the results from sequence of length 1 between the female and male students from Spain are compared, 3 out of 6 of the sequences with length 1 used by female students are also used by male students. The event "stopProgramExecutionSession" used by the male students is not part of the most common sequences of length 1 for female students. The male Spanish students did not create as many bricks as the female students and their most frequent events show that their creativity is less when compared to the female students who used it in 52,80% of the cases.

The sequences of length 5, as shown in the Figure 3.69, start with either the event "startProgramExecutionSession" (14 sequences), "stopProgramExecutionSession"(4 sequences), and only one sequence of length 5 starts with "openProgram". One of the differences with female Spanish students is that they do not have any commonly used sequences of length 5 that start with "openProgram".

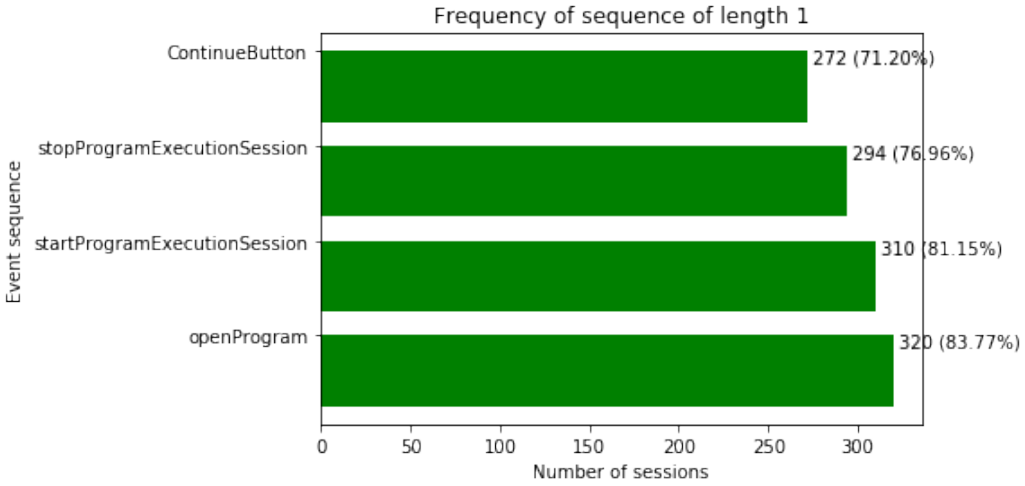

Figure 3.68: Sequences of length 1 in a session for all Spanish male users with PrefixSpan

The results for most frequently used sequences of length 1 from all British users are shown in Figure 3.70. There are 4 sequences of length one and that are: "startPocketPaintSessionEditLook" in 676 sessions (67,77%), "stopProgramExecutionSession" in 842 sessions (84,79%), "startProgramExecutionSession" in 861 sessions (86,71%) and "openProgram" in 878 sessions

Figure 3.69: Sequences of length 5 in a session for all Spanish male users with PrefixSpan

(88,42%). This results are compared with the sequences of length 1 of all Spanish students. The Spanish students have two different sequences: "openFormulaEditor" and "ContinueButton", the rest three are the same as the ones used by the British students. The sequence of length 1 "startPocketPaintSessionEditLook" is frequently used by all the British students, but not by the Spanish students. Two of the frequently used sequences of length 1 by the Austrian students: "init_session" and "ContinueButton", are not frequently used by the British students, but the rest of the frequently used events are also used by the British students.

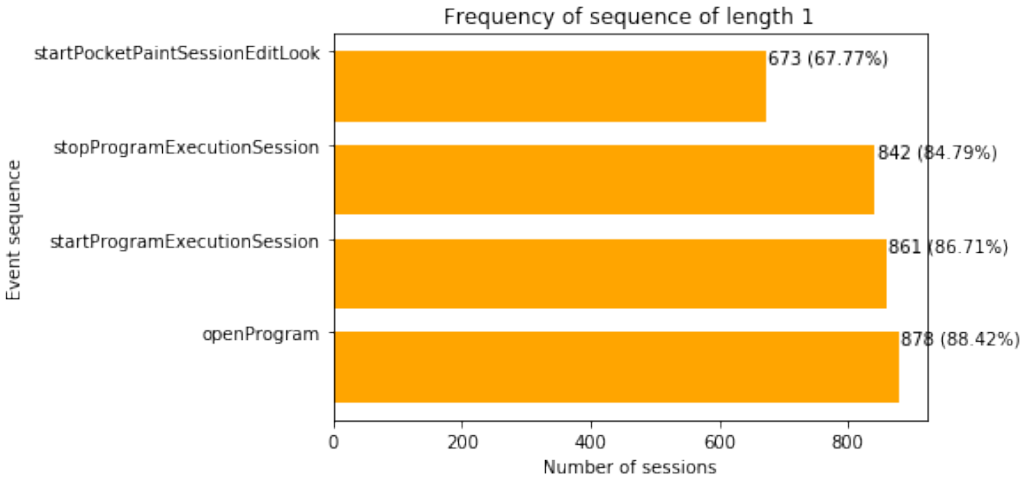

Figure 3.70: Sequences of length 1 in a session for all British users with PrefixSpan

The results for most common sequences of length 1 from all female British users are shown in Figure 3.72. The sequence startPocketPaintSessionEditLook" in 282 sessions (67,63%), then "stopProgramExecutionSession" in 349 sessions (83,69%), the event "startProgramExecutionSession" in 353 sessions 84,65% and finally "openProgram" in 360 sessions (86,33%).

Figure 3.71: Sequences of length 5 in a session for all British users with PrefixSpan

The results for most common sequences of length 5 from all female British users are shown in Figure 3.73

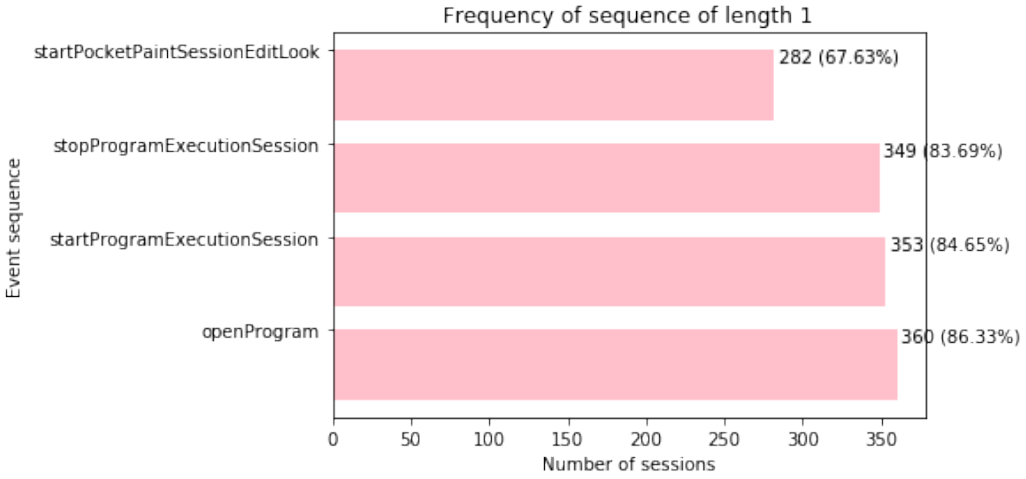

Figure 3.72: Sequences of length 1 in a session for all British female users with PrefixSpan

For all the male British users the most common sequences of length 1 are shown in figure 3.74. All 4 sequences of length 1 that are commonly used by the male British users are also used by the female users. This means that there are no differences between the sequences of length one, between the British boys and girls. The only difference is the numbers of sessions in which they used these events, but the numbers are comparable.

The results for most common sequences of length 5 from all male British users are shown in figure 3.75. There are 10 sequences of length 5 for British male students that start with the event "openProgram". These are very similar to the most frequent sequences of length 5 used by female British students, where 11 sequences start with the exactly same event. All of the rest

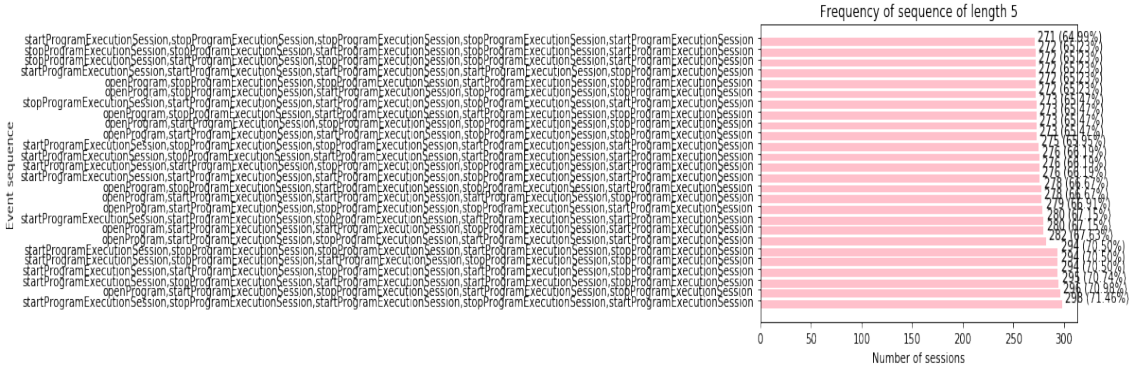

Figure 3.73: Sequences of length 5 in a session for all British female users with PrefixSpan

sequences start with "startProgramExecutionSession" and "stopProgramExecutionSession", which is also like in the case of the female British students. The only difference is that in the case of the male British students the percentage of use is slightly higher.

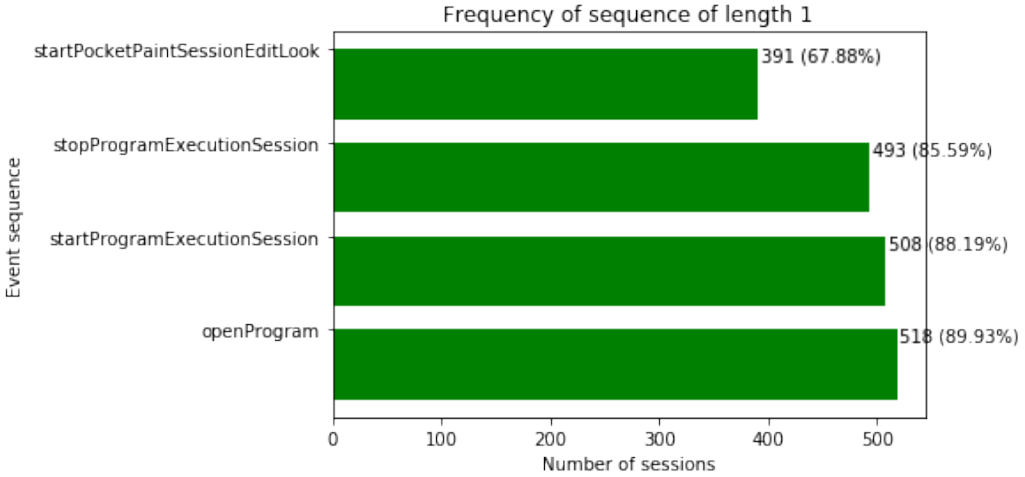

Figure 3.74: Sequences of length 1 in a session for all British male users with PrefixSpan

For the Graz International Bilingual School the results of most frequently used sequences of length 1 are given in figure 3.76 and for sequences of length 5 in the figure 3.77. The results are as it can be expected from the results of the most frequently used events for all the Austrian students.

For the students studying in the school Borg Birkfeld, the results obtained with the use of PrefixSpan are as anticipated. Especially, that is the case for

Figure 3.75: Sequences of length 5 in a session for all British male users with PrefixSpan

the female students from this school. There are only a few female students tracked and have significantly lower number of events both per user and per session, which makes it predictable that sequences of higher length are not used by the female students from this school. The maximum length of sequences that female students used frequently is of length 3, and there are 3 such a sequences in total. Two of them were used only by 14,29% of the female students, and the most common one was only used by 28,57%. As shown in the figure 3.85.

The results for frequent sequences of length 1 from all of the students being part of Akademisches Gymnasium are given in figure 3.88 and the results for frequent sequences of length 5 are given in the figure 3.89. In the results of frequently used sequences of length 1 there are 11 different events used by all of the students studying in Akademisches Gymnasium. The students studying at Akademisches Gymnasium have more frequent sequences of 1 compared to the Borg Birkfeld (5) and GIBS (4). Some of the events used only by the students studying at Akademisches Gymnasium are “startExploreSession” and “stopExploreSession”. The time between these two events represents the time spent by the users in web session. Other unusual events for the the other two Austrian schools but frequently used by the students from the Akademisches Gymnasium are “createObject”m “createLook”, “openFormulaEditor” and “saveFormula”, which are used in more than 40 % of all of the sessions as shown in 3.88. This shows that the users from Akademisches Gymnasium used more creative events compared to the other two Austrian schools. The frequent sequences of length 5 as shown in figure 3.89 are only 5 in total, which is also different in the case of Borg Birkfeld (15) and GIBS (15). The difference between the most frequent sequences of length one for female students studying in the three different

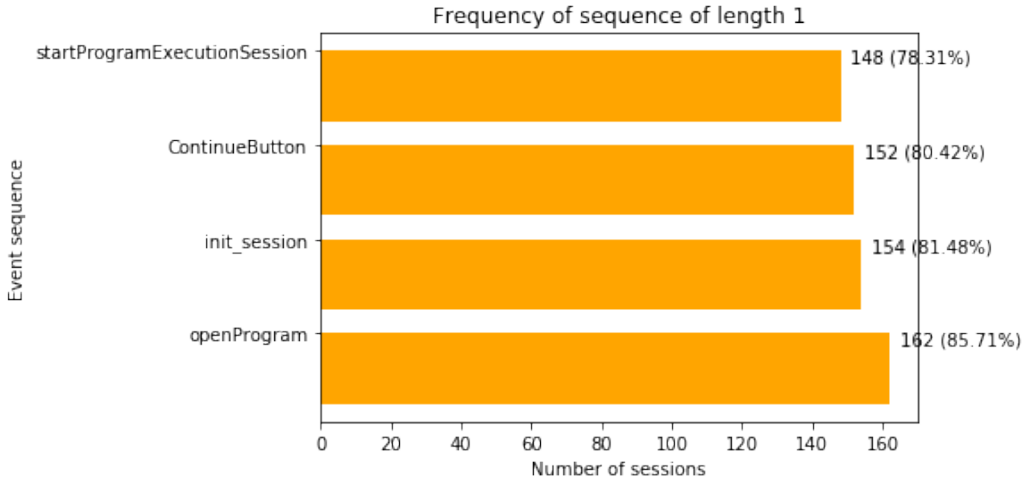

Figure 3.76: Sequences of length 1 in a session for all users from GIBS with PrefixSpan

Figure 3.77: Sequences of length 5 in a session for all users from GIBS with PrefixSpan

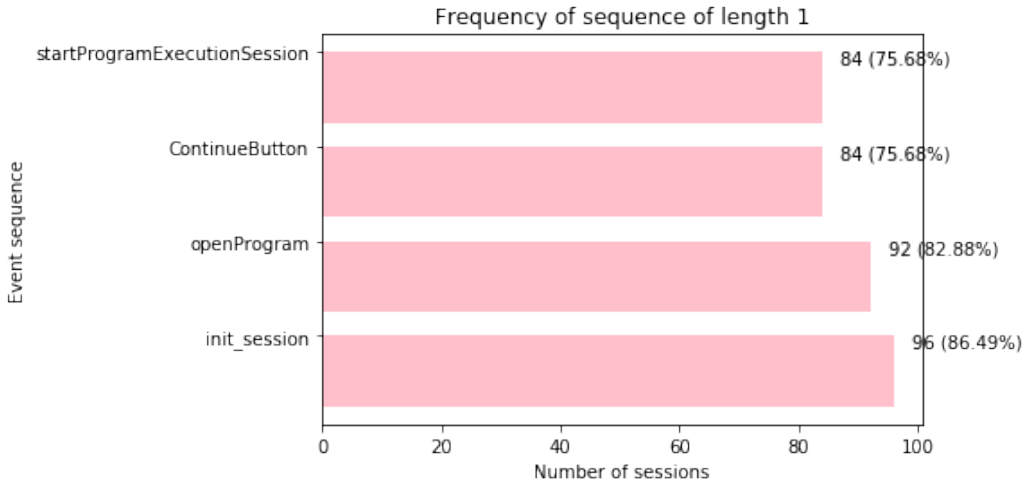

Figure 3.78: Sequences of length 1 in a session for all female users from GIBS with PrefixSpan

Figure 3.79: Sequences of length 5 in a session for all female users from GIBS with PrefixSpan

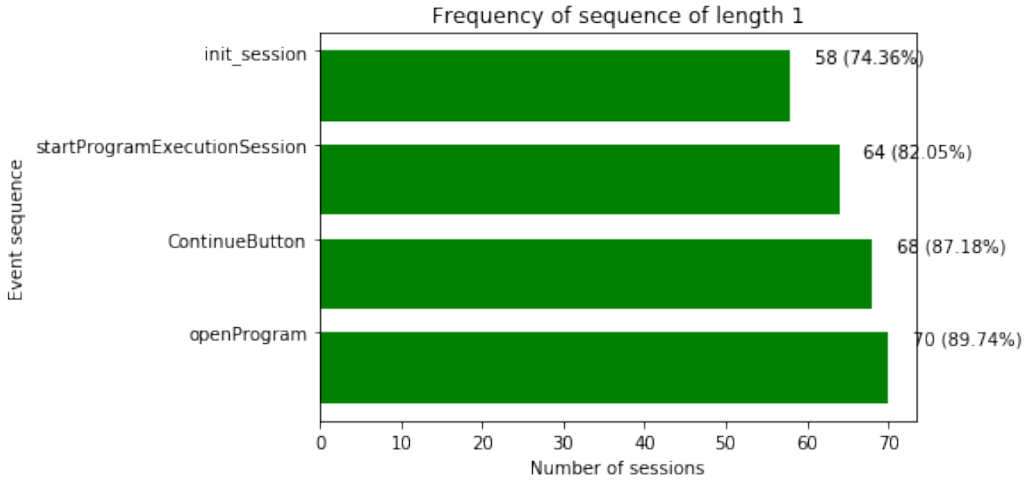

Figure 3.80: Sequences of length 1 in a session for all male users from GIBS with PrefixSpan

Figure 3.81: Sequences of length 5 in a session for all male users from GIBS with PrefixSpan

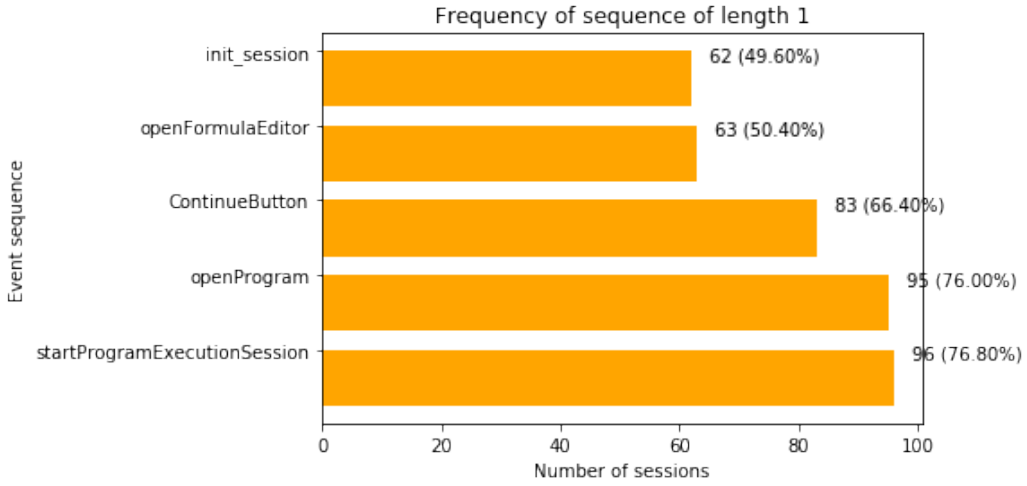

Figure 3.82: Sequences of length 1 in a session for all users from Borg Birkfeld with PrefixSpan

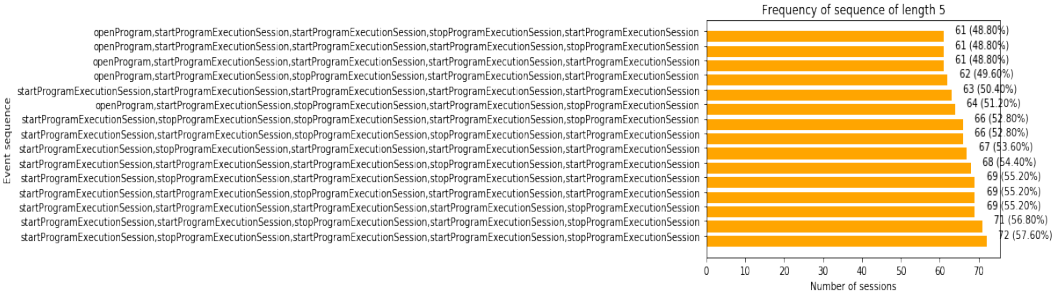

Figure 3.83: Sequences of length 5 in a session for all users from Borg Birkfeld with PrefixSpan

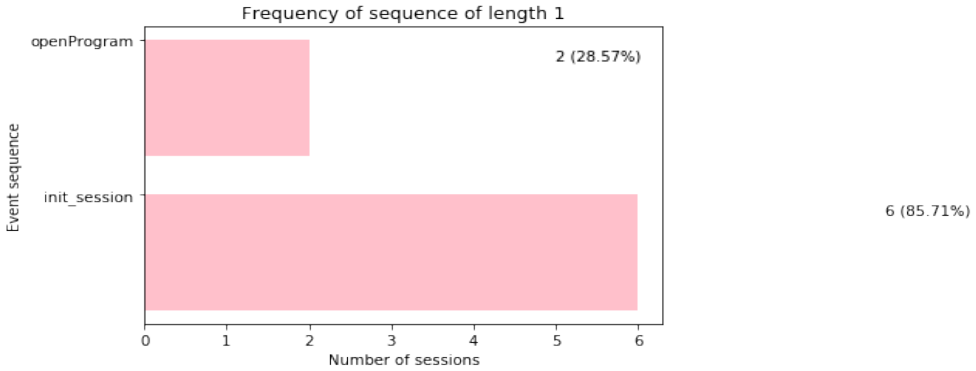

Figure 3.84: Sequences of length 1 in a session for all female users from Borg Birkfeld with PrefixSpan

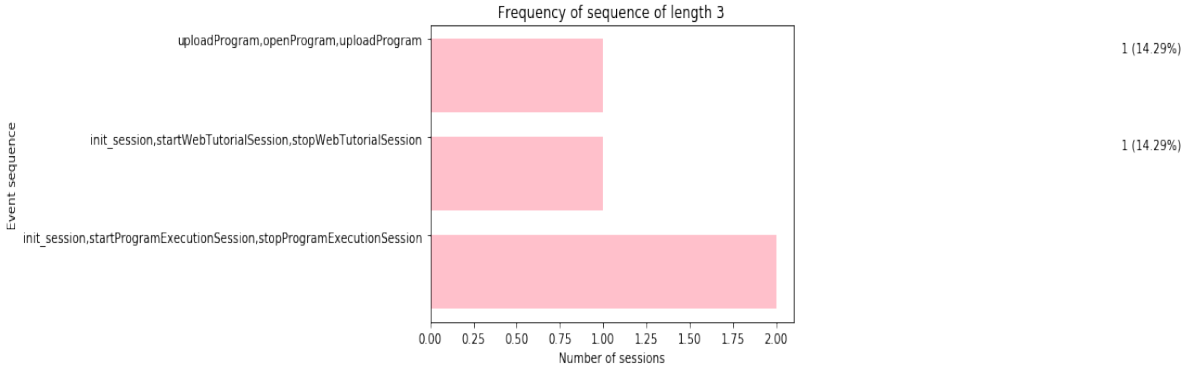

Figure 3.85: Sequences of length 3 in a session for all female users from Borg Birkfeld with PrefixSpan

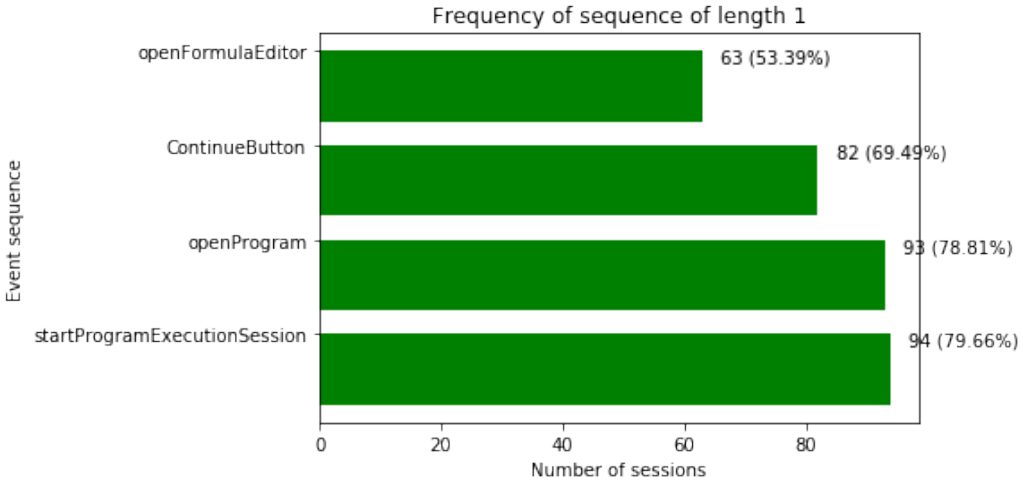

Figure 3.86: Sequences of length 1 in a session for all male users from Borg Birkfeld with PrefixSpan

Figure 3.87: Sequences of length 5 in a session for all male users from Borg Birkfeld with PrefixSpan

Austrian schools is significant. There are 10 different sequences of length 1 used by the girls studying at AK, These sequences are used in high number of sessions (between 40 % and 67 %) as shown in figure 3.90 . The number of sequences of length 1 is also different in the case for the female students studying at Borg Birkfeld (2 frequent sequences of length 1) and GIBS (4 frequent sequences of length 1).

Similarly, for the male students studying at Akademisches Gymnasium there are 9 different sequences of length 1, given in figure 3.92. These sequences are used in between 44 % and 47 % of all the sessions created by boys. The number of sequences of length 1 is again different in the case for the male students studying at Borg Birkfeld (4 frequent sequences of length 1) and GIBS (4 frequent sequences of length 1).

The results from Spain are entirely based on the results from the schools Ubeda. That is the reason why the results from this school are not separately listed again in this part.

The results for frequent sequences of length 1 from all of the students being part of Oakfield (UK) are given in figure 3.94 and the results for frequent sequences of length 5 are given in the figure 3.95. There are only 3 sequences of frequent events of length 1 and are used 58 % and 81 % of all the sessions used by the students from Oakfield. The female students used frequently 5 different sequences of length 1 shown in figure 3.96 and the male students 3 different sequences of length 1 as shown in figure 3.98.

The results for frequent sequences of length 1 from all of the students being part of St. Stephens Primary (UK) are given in figure 3.100 and the results for frequent sequences of length 5 are given in the figure 3.101. There are 4 frequent sequences of length 1 and are used in between 72 % and 89 % of all the sessions used by the students from St.Stephens Primary. There are 4 different frequently used sequences of length 1 used by the female students at St.Stephens Primary, as shown in figure 3.102. The frequent sequences of length 1 used by male students are shown in figure 3.104. The frequently used event "startPocketPaintSessionEditLook" by both female and male students at St. Stephens Primary, is not part of the frequently used sequences of length 1 for the female or male students at Oakfield school.

The results for frequent sequences of length 1 from all of the students being part of Christ the King (UK) are given in figure 3.106 and the results for frequent sequences of length 5 are given in the figure 3.107. As expected, there are no frequent sequences of higher length than 1 for boys and girls.

The results for frequent sequences of length 1 from all of the students being part of St. Peters school (UK) are given in figure 3.110 and the results for frequent sequences of length 5 are given in the figure 3.111. There are 4 frequently used sequences of length 1 by the students studying at St. Peters

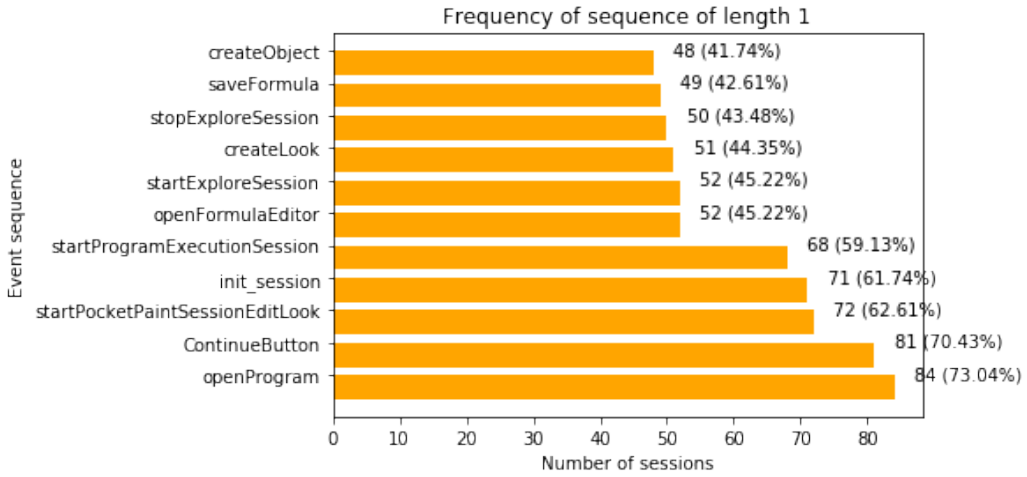

Figure 3.88: Sequences of length 1 in a session for all users from Akademisches Gymnasium with PrefixSpan

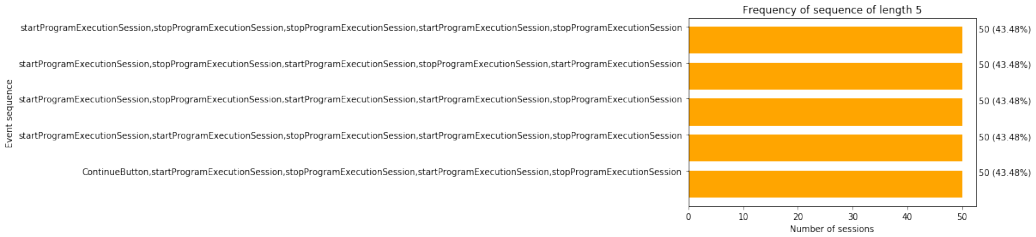

Figure 3.89: Sequences of length 5 in a session for all users from Akademisches Gymnasium with PrefixSpan

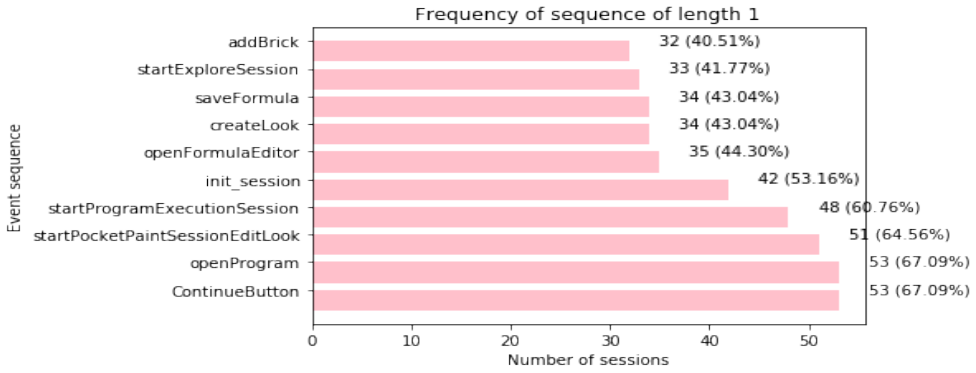

Figure 3.90: Sequences of length 1 in a session for all female users from Akademisches Gymnasium with PrefixSpan

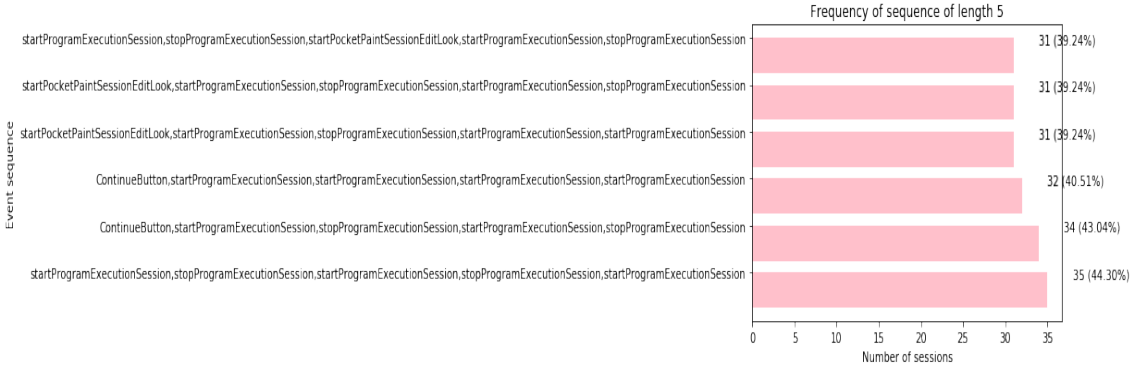

Figure 3.91: Sequences of length 3 in a session for all female users from Akademisches Gymnasium with PrefixSpan

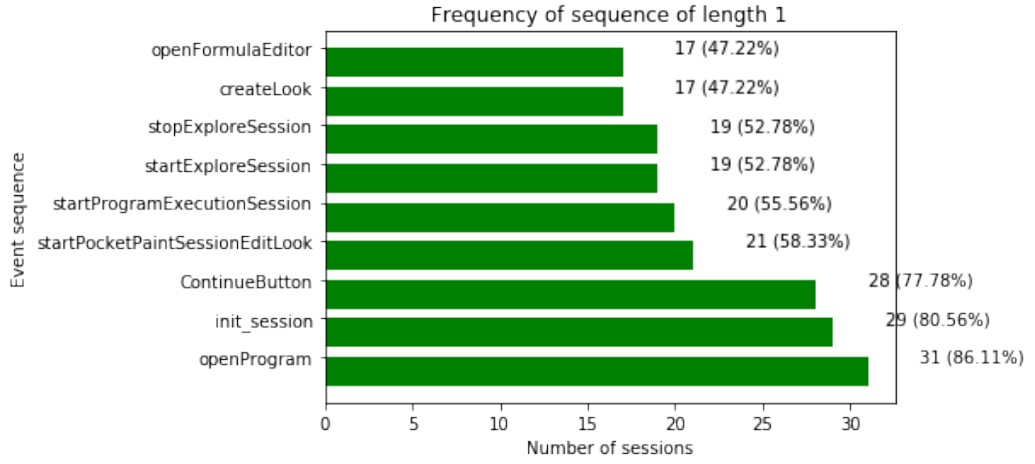

Figure 3.92: Sequences of length 1 in a session for all male users from Akademisches Gymnasium with PrefixSpan

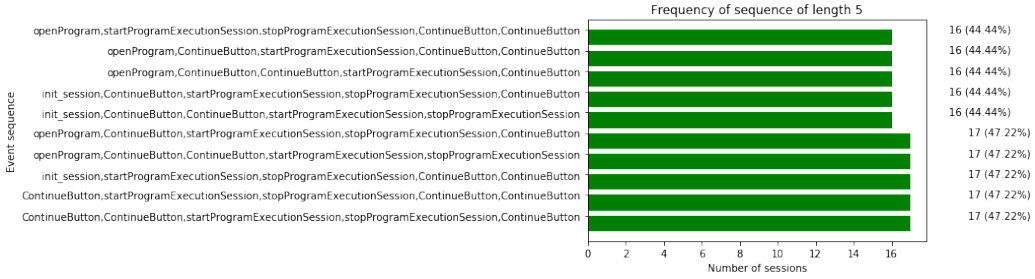

Figure 3.93: Sequences of length 5 in a session for all male users from Akademisches Gymnasium with PrefixSpan

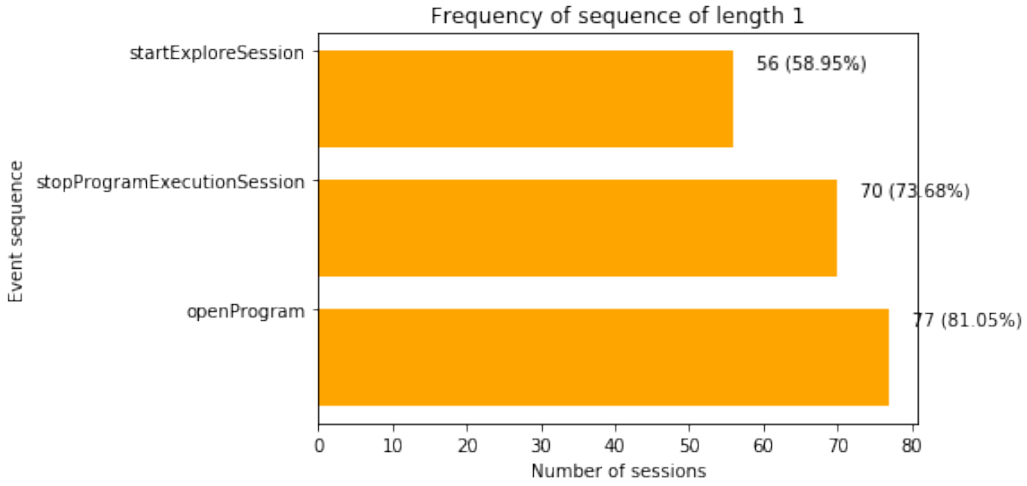

Figure 3.94: Sequences of length 1 in a session for all users from Oakfield with PrefixSpan

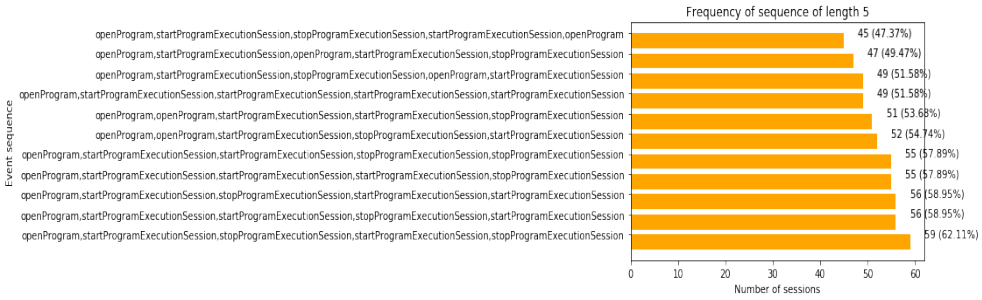

Figure 3.95: Sequences of length 5 in a session for all users from Oakfield with PrefixSpan

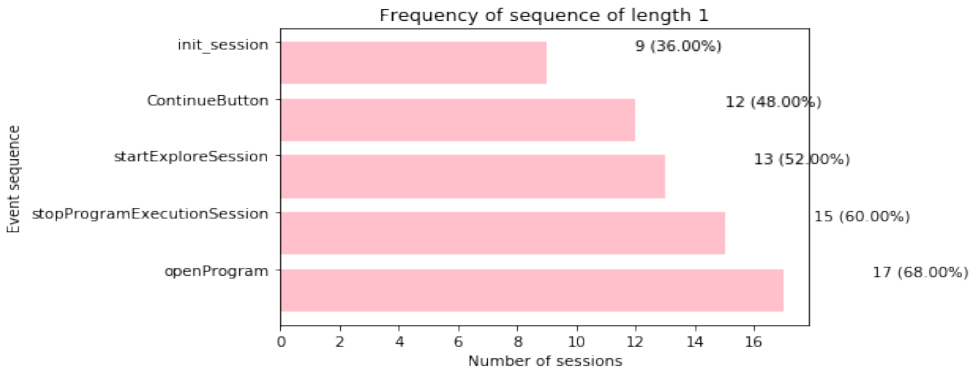

Figure 3.96: Sequences of length 1 in a session for all female users from Oakfield with PrefixSpan

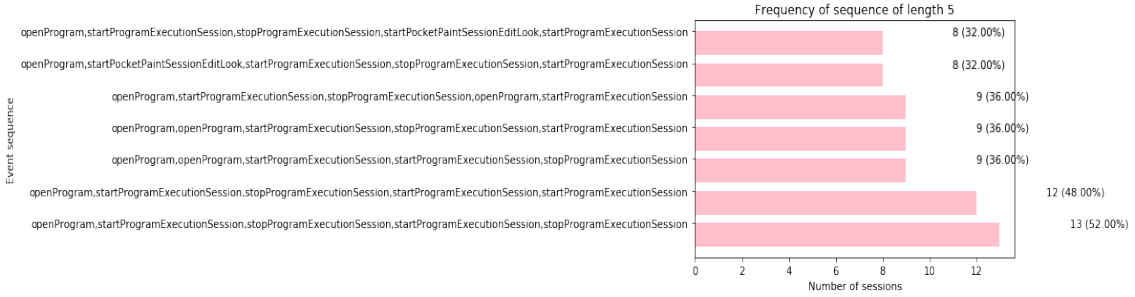

Figure 3.97: Sequences of length 3 in a session for all female users from Oakfield with PrefixSpan

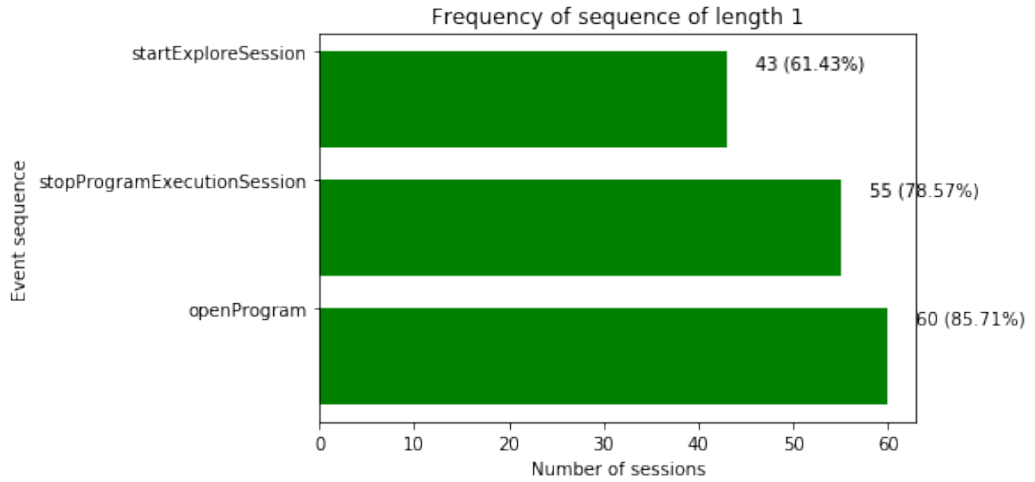

Figure 3.98: Sequences of length 1 in a session for all male users from Oakfield with PrefixSpan

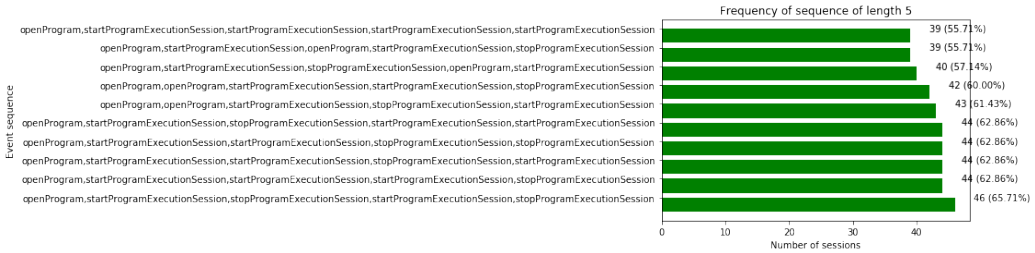

Figure 3.99: Sequences of length 5 in a session for all male users from Oakfield with PrefixSpan

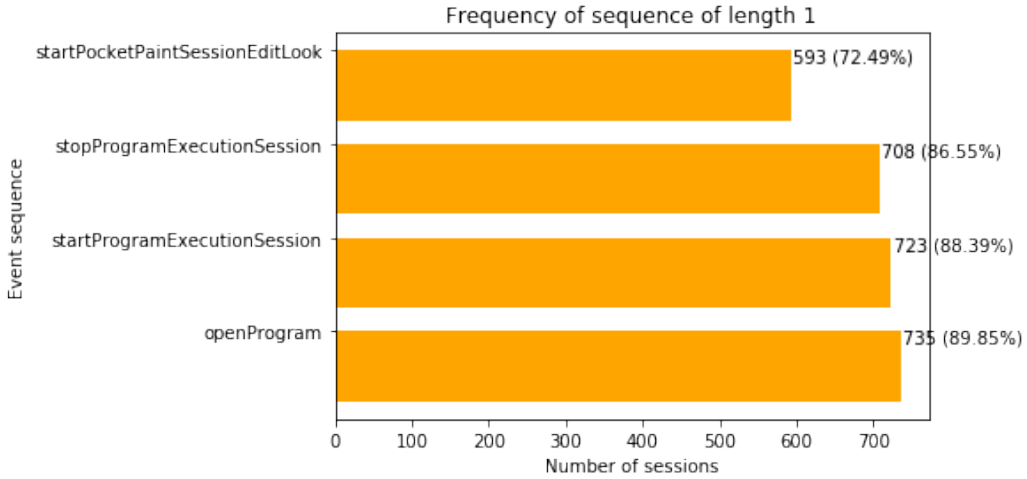

Figure 3.100: Sequences of length 1 in a session for all users from St. Stephens Primary with PrefixSpan

Figure 3.101: Sequences of length 5 in a session for all users from St. Stephens Primary with PrefixSpan

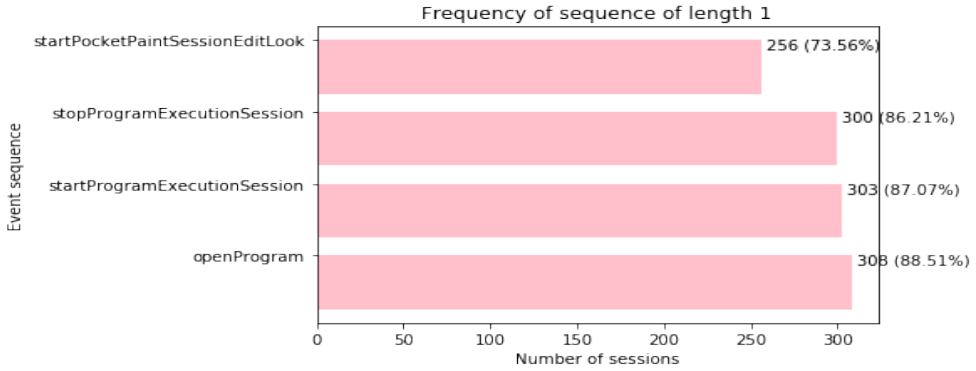

Figure 3.102: Sequences of length 1 in a session for all female users from St. Stephens Primary with PrefixSpan

Figure 3.103: Sequences of length 3 in a session for all female users from St. Stephens Primary with PrefixSpan

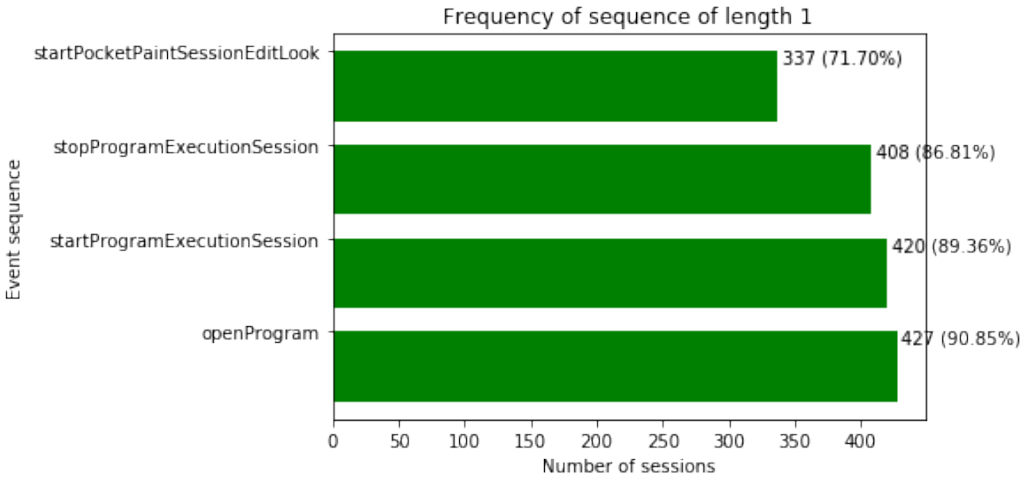

Figure 3.104: Sequences of length 1 in a session for all male users from St. Stephens Primary with PrefixSpan

Figure 3.105: Sequences of length 5 in a session for all male users from St. Stephens Primary with PrefixSpan

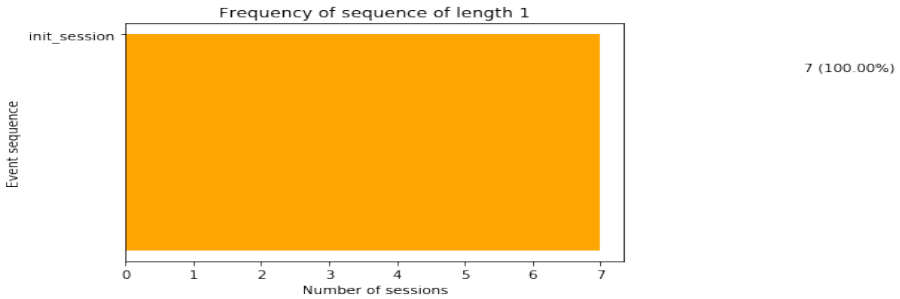

Figure 3.106: Sequences of length 1 in a session for all users from Christ the King with PrefixSpan

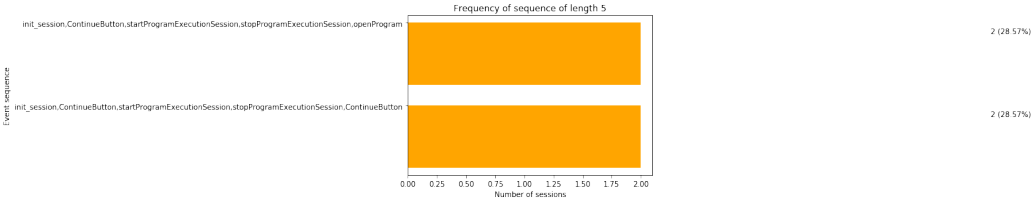

Figure 3.107: Sequences of length 5 in a session for all users from Christ the King with PrefixSpan

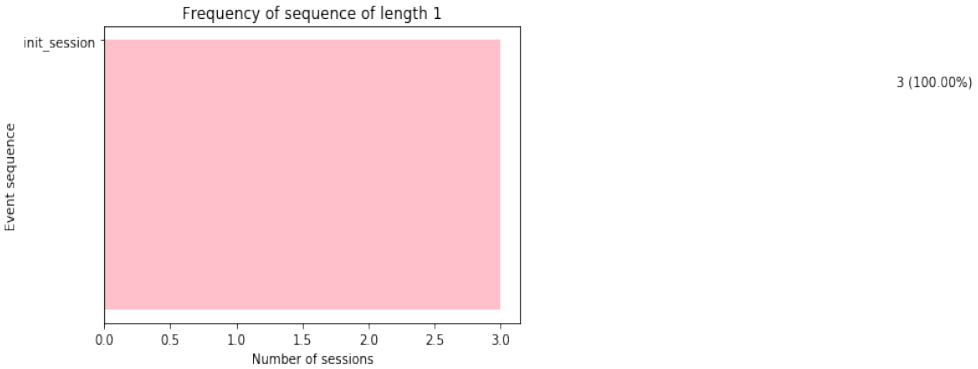

Figure 3.108: Sequences of length 1 in a session for all female users from Christ the King with PrefixSpan

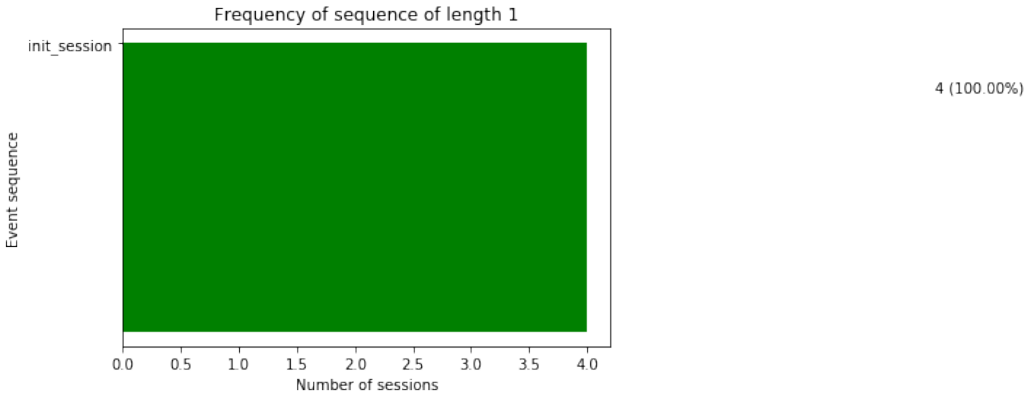

Figure 3.109: Sequences of length 1 in a session for all male users from Christ the King with PrefixSpan

school. The amount of frequent sequences is the same for all the students studying at St. Stephan, but the sequences are not the same, two of them are mutual and two are different. The two sequences that are frequent for the students from St. Peters school and are not for St. Stephan Primary are: 'openFormulaEditor" and 'ContinueButton". Frequent sequences of length 1 for female students part of the St. Peters school are 3 in total as shown in figure 3.112, and in the case for the students from St. Stephan Primary there are 4 in total as shown in figure 3.102.

3.3 Classification and definition of the difference in the patterns found

The type of events frequently used by different group of users was obtained by the use of PrefixSpan. From the results comparisons can be easily made for the sequences with length 1, but it's not trivial to make comparison and conclusions from the results of higher length 2, 3, 4 or 5. In order to check the significance in the differences for frequent sequences of events from different groups, found as results by using the PrefixSpan pattern mining in the previous step, 2 different tests for similarity calculation were done:

The first test was done in the following way: The most frequent sequences of events with length 1, 2, 3, 4 and 5 and by filtering the top 70 most frequent events in respect to every particular group (students from certain country/school/gender), from any group were compared to any other group. With the results from this test can be concluded how many of the sequences are overlapping between different groups. From the results obtained for the

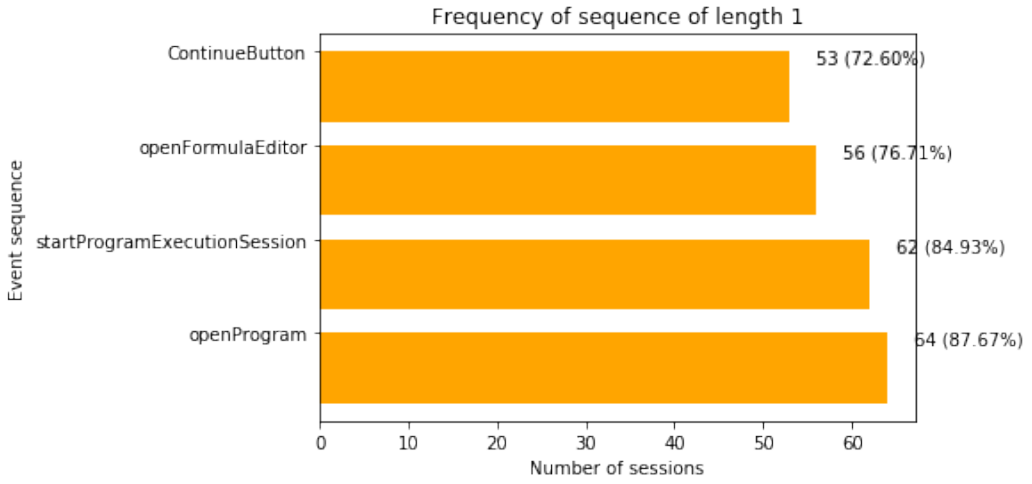

Figure 3.110: Sequences of length 1 in a session for all users from St.Peters school with PrefixSpan

Figure 3.111: Sequences of length 5 in a session for all users from St.Peters school with PrefixSpan

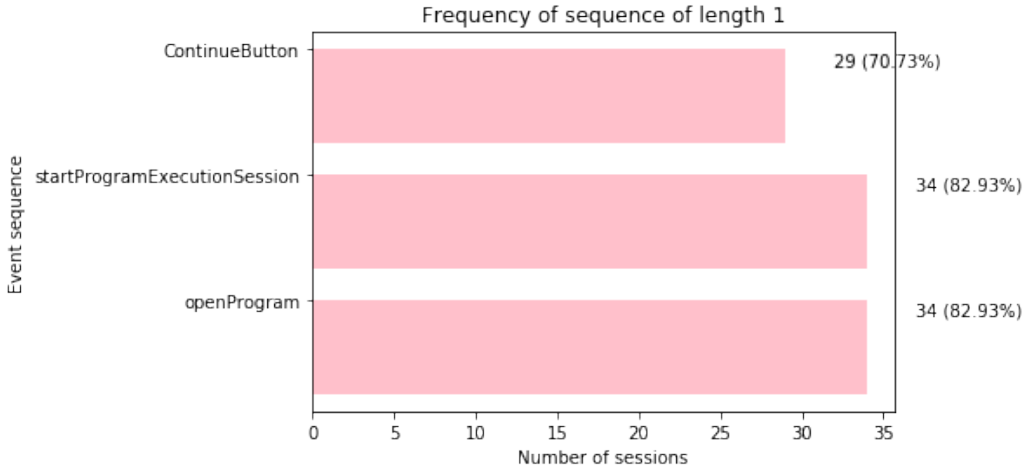

Figure 3.112: Sequences of length 1 in a session for all female users from St.Peters school with PrefixSpan

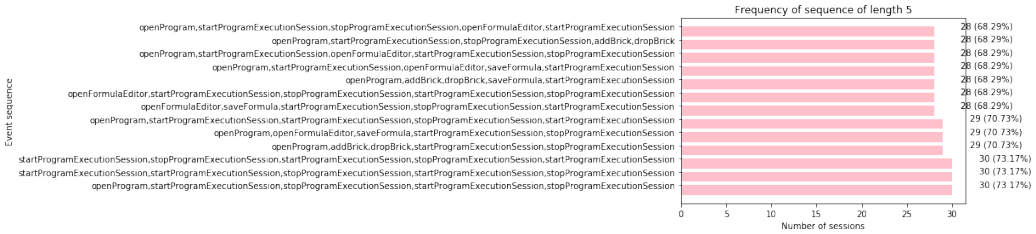

Figure 3.113: Sequences of length 3 in a session for all female users from St.Peters school with PrefixSpan

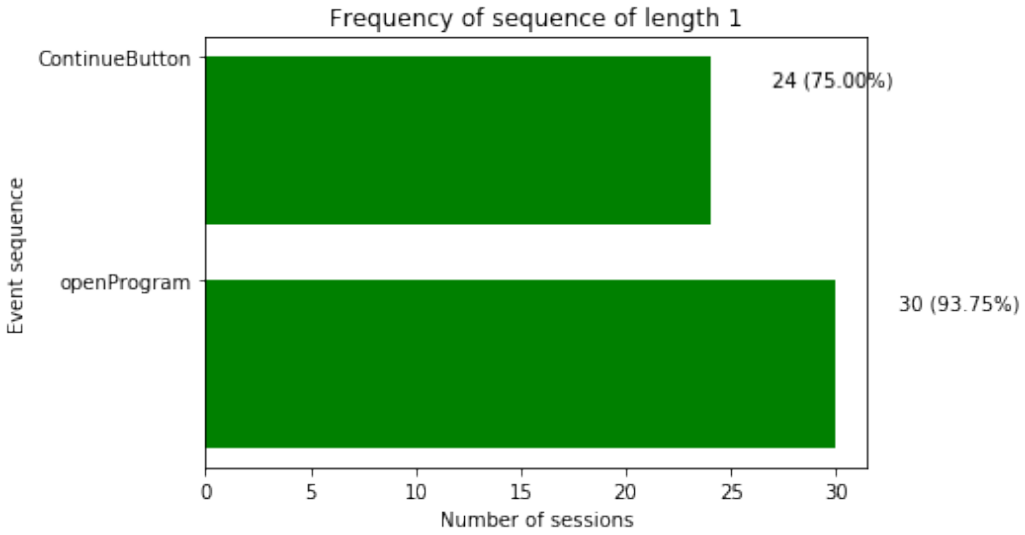

Figure 3.114: Sequences of length 1 in a session for all male users from St.Peters school with PrefixSpan

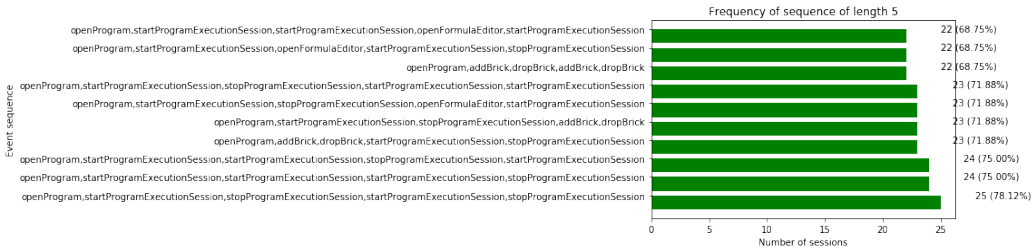

Figure 3.115: Sequences of length 5 in a session for all male users from St.Peters school with PrefixSpan

Country	School	Male	Female	Total
AT		8	2	10
	GIBS	3	1	4
	Borg Birkfeld	3	0	3
	Akademische Gymnasium	2	1	3
ES		7	5	12
	Santa Maria	0	0	0
	Ubeda	7	5	12
UK		17	6	23
	Oakfield	2	1	3
	St. Stephens Primary	15	5	20
	Christ the King	0	0	0
	St. Peters school	0	0	0
Total number of used "help" events				**45**

Table 3.11: Students who used one of the help events/buttons: ("openBrickHelpOption" or "hintOption")

Country	School	Male	Female	Total
AT		333	219	552
	GIBS	155	99	254
	Borg Birkfeld	102	0	102
	Akademische Gymnasium	76	120	196
ES		858	150	1008
	Santa Maria	0	0	0
	Ubeda	858	150	1008
UK		1210	211	1421
	Oakfield	12	0	12
	St. Stephens Primary	1198	210	1408
	Christ the King	0	0	0
	St. Peters school	0	1	1
Total number of "copy" events				**2981**

Table 3.12: Students who used one of the copy events: (including: "CopyProgram", "CopyObject", "CopyLook", "CopyLook", "CopySound", "CopyBrick")

Country	School	Male	Female	Total
AT		77	74	151
	GIBS	61	73	134
	Borg Birkfeld	15	0	15
	Akademische Gymnasium	1	1	2
ES		85	53	138
	Santa Maria	0	0	0
	Ubeda	85	53	138
UK		20	13	33
	Oakfield	10	8	18
	St. Stephens Primary	8	5	13
	Christ the King	0	0	0
	St. Peters school	3	0	3
Total number of "backpack" events				**322**

Table 3.13: Students who used one of the advanced events: (including: "backpackObject", "unpackObject", "backpackLook", "unpackLook", "backpackSound", "unpackSound", "backpackScripts", "unpackScripts", "backpackScenes", "unpackScenes")

dataset used for this thesis, confusion matrices were created. The way of representation of the results on the confusion matrix is explained with an example before the actual results are presented in order to avoid misinterpretation.

Example for clarifying the representation: Each time two different groups are compared, like for example female and male students from Austria. If they are marked as G1 and G2 respectively, and the letters given below are sequences of frequent events (like for example A="openProgram", B="addBrick", C="dropBrick"). By using PrefixSpan the following sequences of 3 are obtained for both groups:

G1: ABC, DEF

G2: XZY, DEF, YEG, IGF, UHG

The sequence of events of length 3: "DEF" is used frequently by both groups. This presented on the confusion matrix would look as following: at the (y,x) y-axis where G1 is given and the x-axis is G2, there the result will be 0.5, because half of the most frequent events of length 3 are also frequent for the G2 (1 out of the 2 possible sequences). That means that from G1 in respect to G2 there is 50% overlap. But when looking at G2 on y axis in respect to G1 on x-axis then the result is 0.2 (20%), because only one out of the 5 frequent sequences of length 3 overlaps or is also is used by G1.

The results from the first part from the actual dataset explain as in the example, are given below.

First the significance in the difference between all the countries for length 1,2,3,4 and 5 are given in the Figures: 3.116, 3.117, 3.118, 3.119, 3.120.

In the results it can be seen that for sequences of length 1 between countries there is a great overlap, hence every filed in the matrix is marked with green. Can be concluded that the Austrian students in comparison to the Spanish students are overlapping 75% (but in respect to the Austrian number of frequent sequences), then the Austrian students in comparison with the British students are overlapping 50% (again in respect to the Austrian number of frequent sequences). Spanish students in respect to Austrian students are overlapping with 60% (in respect to the Spanish number of frequent sequences), Spanish students with British students are overlapping also with 60% (again in respect to the Spanish number of frequent sequences). Finally, British students are overlapping 50% (in respect to the British) and British students with Spanish students overlap 75%(in respect to the British).

In the results for length 2 3.117, there are slightly lower percentages and the the difference between the sequences of Austrian students and British students (in the respect of the number of sequences of the Austrian students) was 50% for length 1 and it is 40% for length 2. That's the one which has the higher difference compared to length 1, the rest have rather smaller difference than before.

For length 3 and length 4 the similarity increases in percentage between the different countries, as shown in figures 3.118 and 3.119.

For sequences of length 5, the percentage of overlapping are still high, the difference is higher between the British student and Austrian students (in respect to the number of the British students), their overlap was length 3 over 88% and for length 4 over 81% and for length 5 is little over 52%. And the similarity between British and Spanish students also decreased on 56% compared to the 75% for sequences of length 4.

The difference between the genders of the different countries for common sequences of length 1, 2, 3, 4 and 5 are given in the figures: 3.121, 3.122, 3.123, 3.124, 3.125.

For sequences of length 1 between genders from different countries there is overlapping between 50 and 100%. Already for sequences of length 2 this similarities are decreasing as shown on 3.122.

But then for sequences of length 3, 4 and 5 their is again high similarity as shown on 3.123, 3.124, 3.125. This is expected result when knowing the results from the PrefixSpan, which were very similar between genders between countries and between countries in general.

The results of the significance in the difference between all the schools for sequences of length 1, 2, 3, 4 and 5 are given in the figures: 3.126, 3.127, 3.128, 3.129, 3.130 respectively. The difference between the genders of the

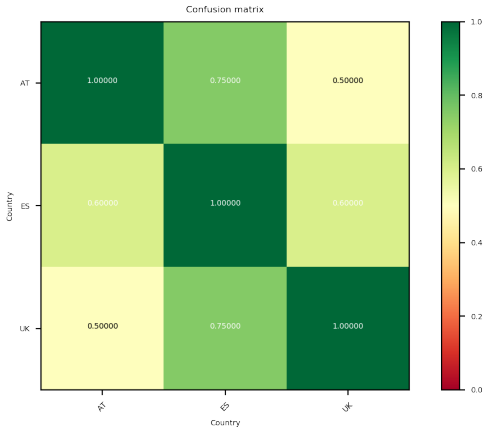

Figure 3.116: Comparing the sequences of length 1 between countries

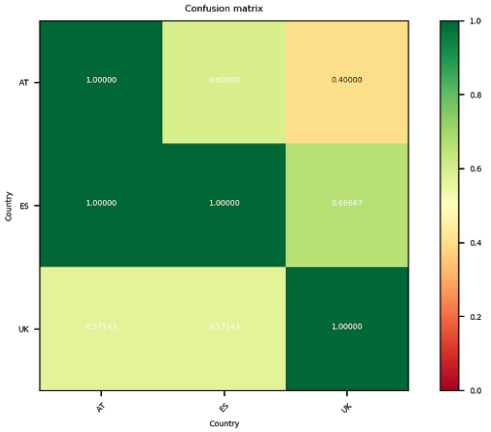

Figure 3.117: Comparing the sequences of length 2 between countries

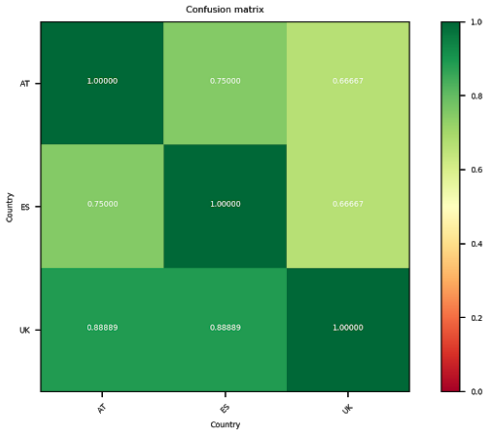

Figure 3.118: Comparing the sequences of length 3 between countries

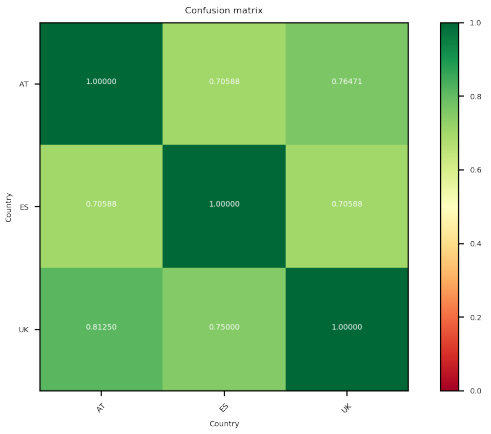

Figure 3.119: Comparing the sequences of length 4 between countries

different schools for commonly used sequences of length 1, 2, 3, 4 and 5 are given in the figures: 3.131, 3.132, 3.133, 3.134, 3.135, respectively.

Still as part of the first test, in the same way as explained in the example above, similarity was calculated between every subgroup in respect to the most commonly used events obtained with PrefixSpan for all students, again by limiting to the top 70 most used events. The goal is to check the overlapping between each single group with the groups of common sequences for all users and how much in percentage that overlaps. The comparison is done for

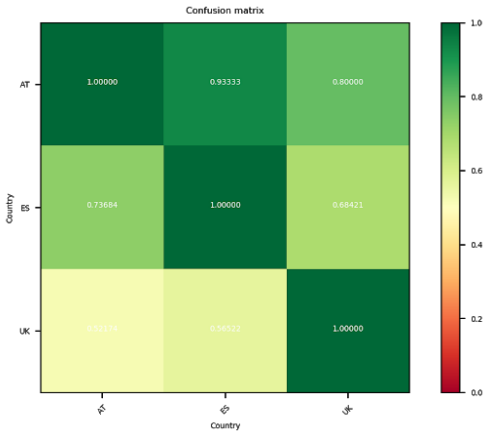

Figure 3.120: Comparing the sequences of length 5 between countries

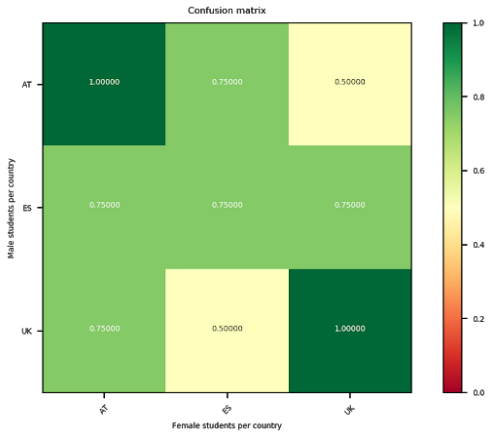

Figure 3.121: Comparing sequences of length 1 between genders and countries

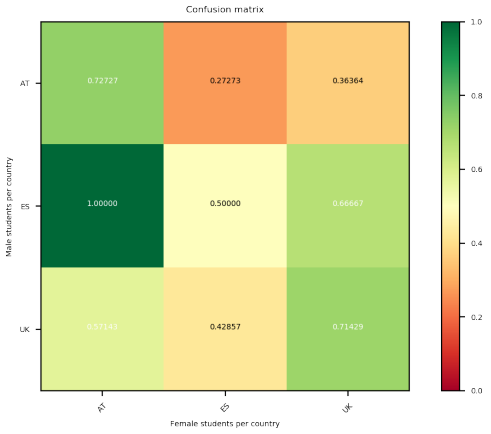

Figure 3.122: Comparing sequences of length 2 between genders and countries

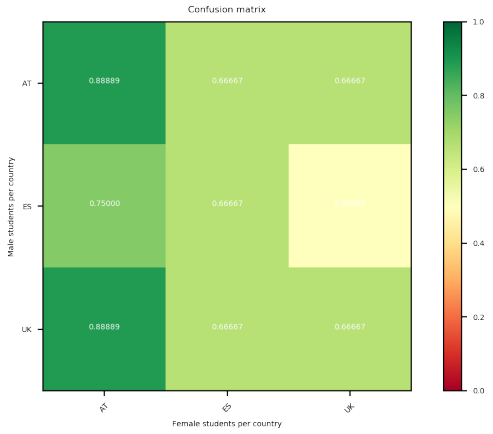

Figure 3.123: Comparing sequences of length 3 between genders and countries

length 1 in figure 3.136, for sequences of length 2 in Figure 3.137, for length 3 in Figure 3.138, for length 4 in Figure 3.139 and finally for length 5 in figure 3.140.

The results of comparing between all the schools for sequences of length 1, 2, 3, 4 and 5 are presented in the figures: 3.126, 3.127, 3.128, 3.129, 3.130 respectively. From this results can be seen that there is great overlapping between the Austrian schools. This is expected result when taking into consideration that the aim of all three Austrian students was the children to

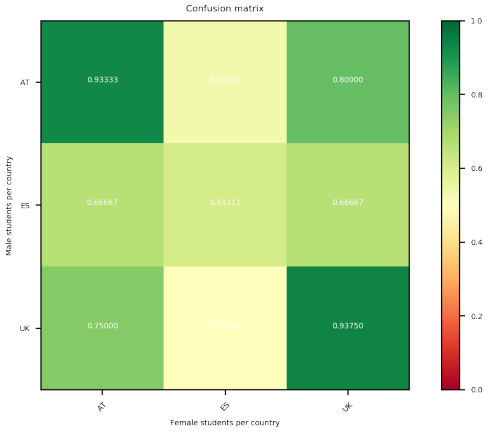

Figure 3.124: Comparing sequences of length 4 between genders and countries

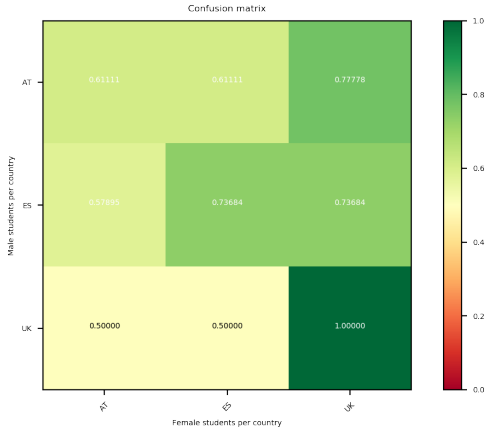

Figure 3.125: Comparing sequences of length 5 between genders and countries

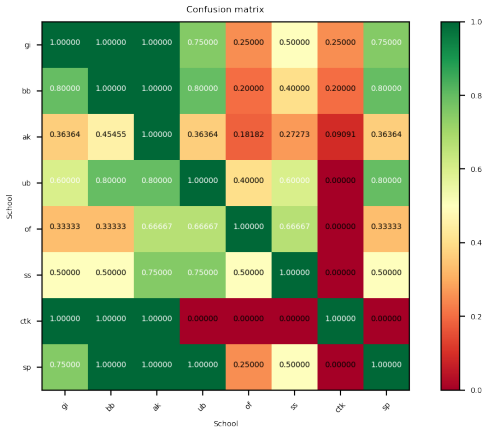

Figure 3.126: Comparing the sequences of length 1 between schools

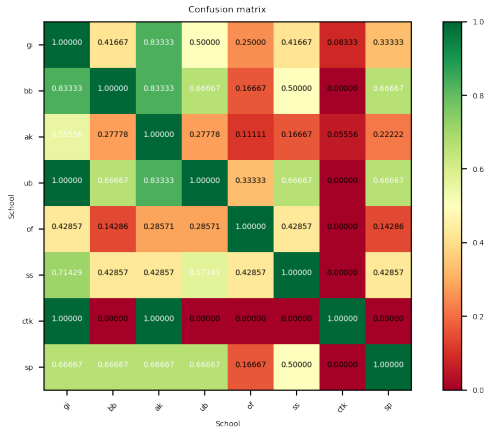

Figure 3.127: Comparing the sequences of length 2 between schools

create games. From the results, for sequences of length 5 given in the figure 3.130, is also visible that there is no great overlapping between the school Ubeda and Oakfield as well as between Ubeda and Christ the King. The reason for this is the too little number of sequences of events used by these two British schools by their students. And that is a consequence of the very little number of students tracked from both of the schools. Another conclusion from the results in the same figure 3.130, is the correlation between the two British schools St. Stephens Primary and St. Peters School. The

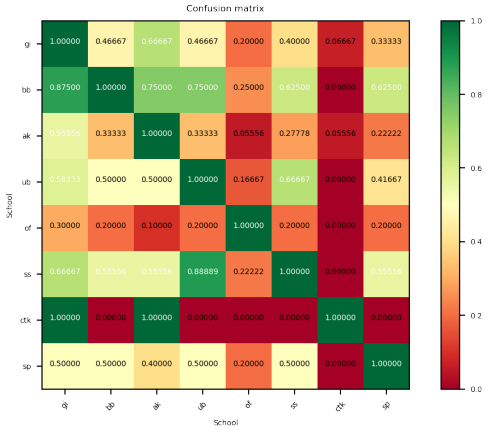

Figure 3.128: Comparing the sequences of length 3 between schools

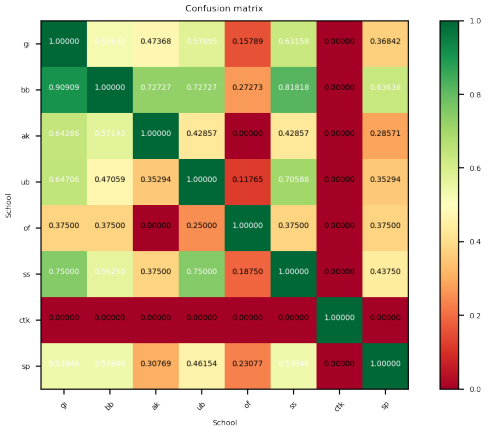

Figure 3.129: Comparing the sequences of length 4 between schools

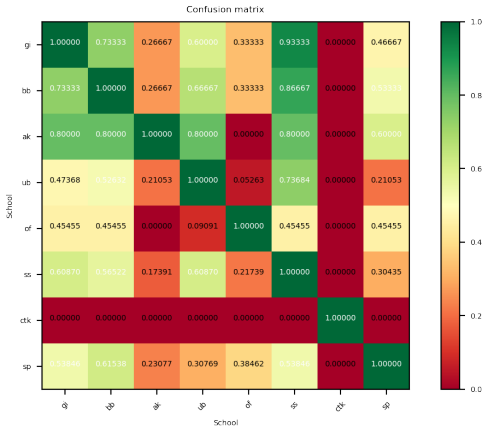

Figure 3.130: Comparing the sequences of length 5 between schools

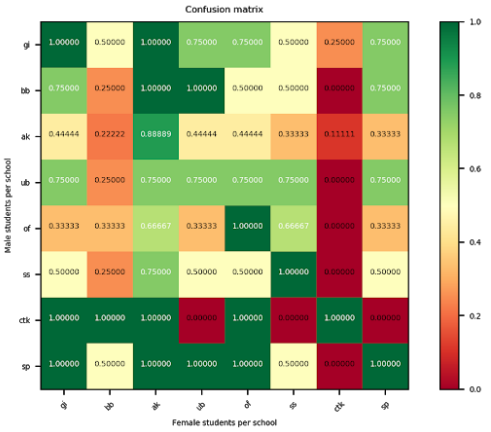

Figure 3.131: Comparing sequences of length 1 between genders and schools

reason for this is that the two schools joined the NOLB project with the same motivation and in the both of the schools study children with cognitive disabilities. Their idea was the children to benefit from the use of Pocket Code. The reason why the other two British school do not show overlapping according to the results is the very small number of tracking students and not having enough data samples, which resulted in not too many frequent sequences when using PrefixSpan.

The same comparison was done between the genders in different schools

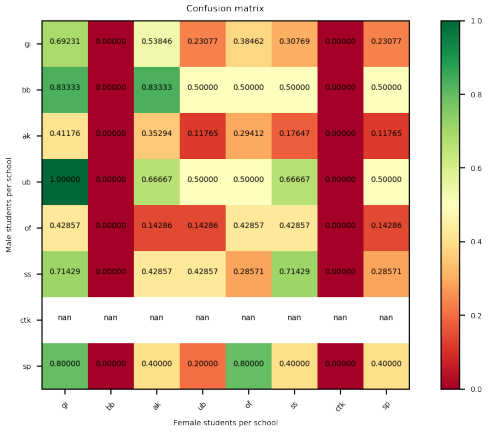

Figure 3.132: Comparing sequences of length 2 between genders and schools

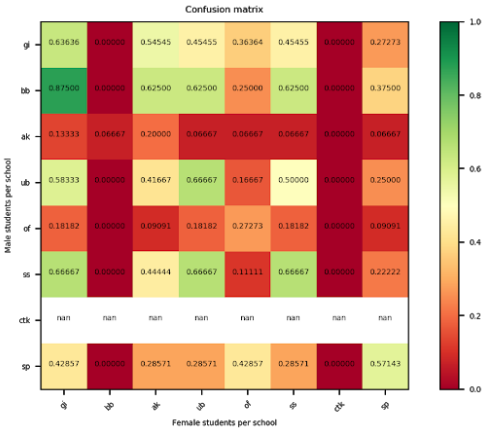

Figure 3.133: Comparing sequences of length 3 between genders and schools

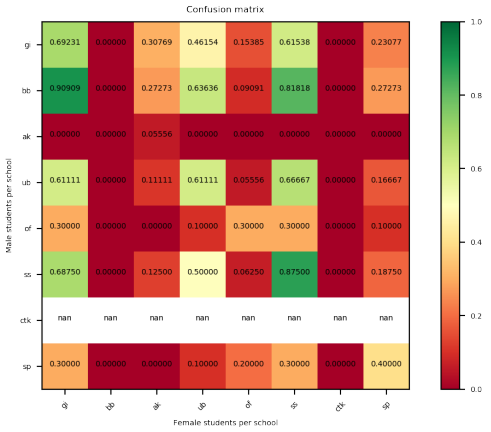

Figure 3.134: Comparing sequences of length 4 between genders and schools

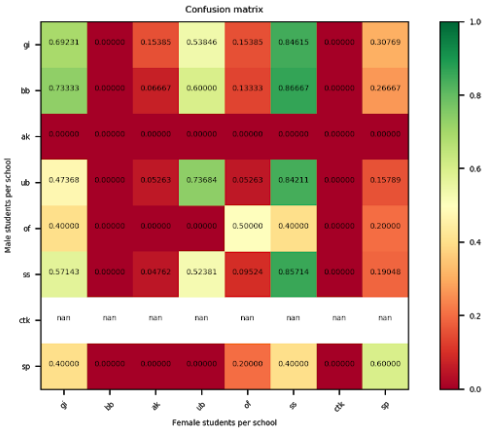

Figure 3.135: Comparing sequences of length 5 between genders and schools

again for sequences of length 1, 2, 3, 4 and 5 are given in the figures: 3.131, 3.132, 3.133, 3.134, 3.135 respectively. In the results for comparing genders between schools for frequent sequences of length 1 shown in the figure 3.131, it's visible that there is high correlation between the girls and the guys from all schools with only exception of the group of female students from the school Christ the King that don't show correlation with any other school, but only with the male students from the same school. From the results comparing sequences of length 5 between genders from different schools, as

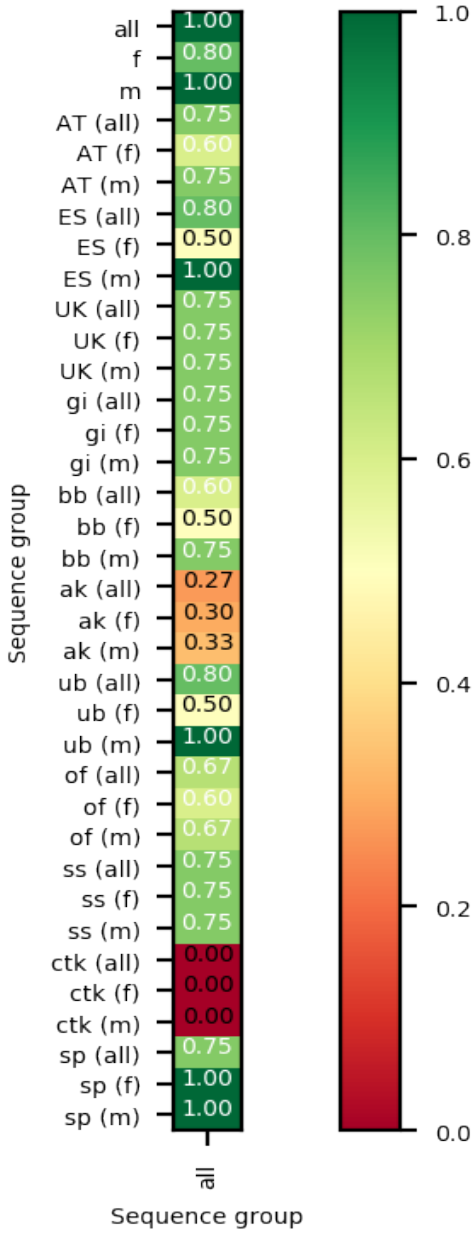

Figure 3.136: Most common sequences of length 1 between each groups and the one for all users

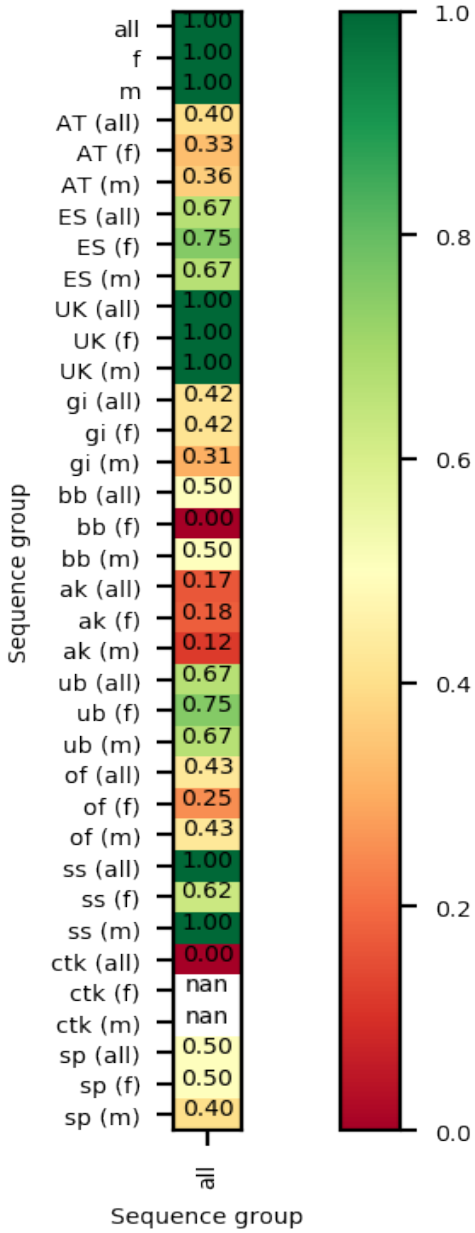

Figure 3.137: Most common sequences of length 2 between each groups and the one for all users

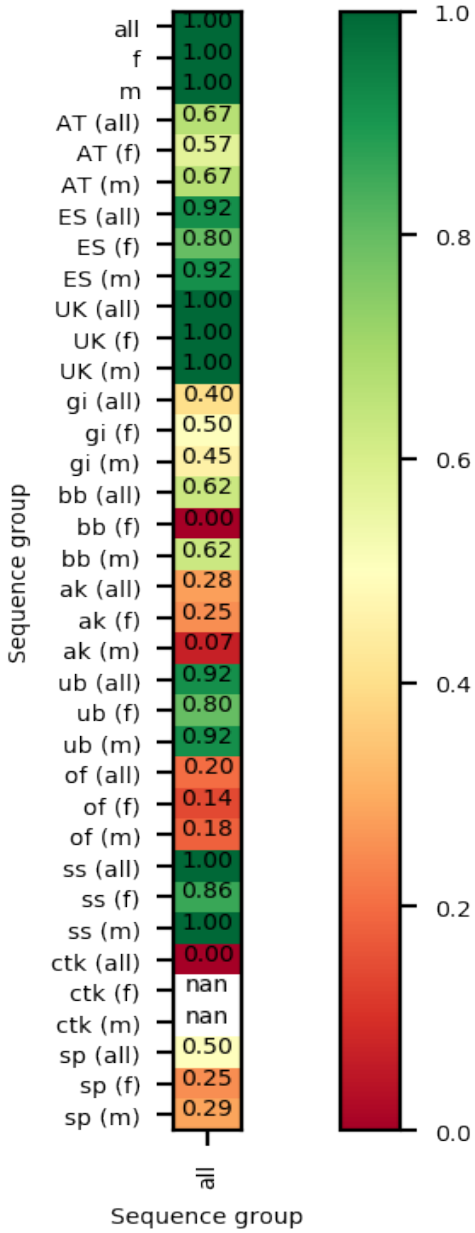

Figure 3.138: Most common sequences of length 3 between each groups and the one for all users

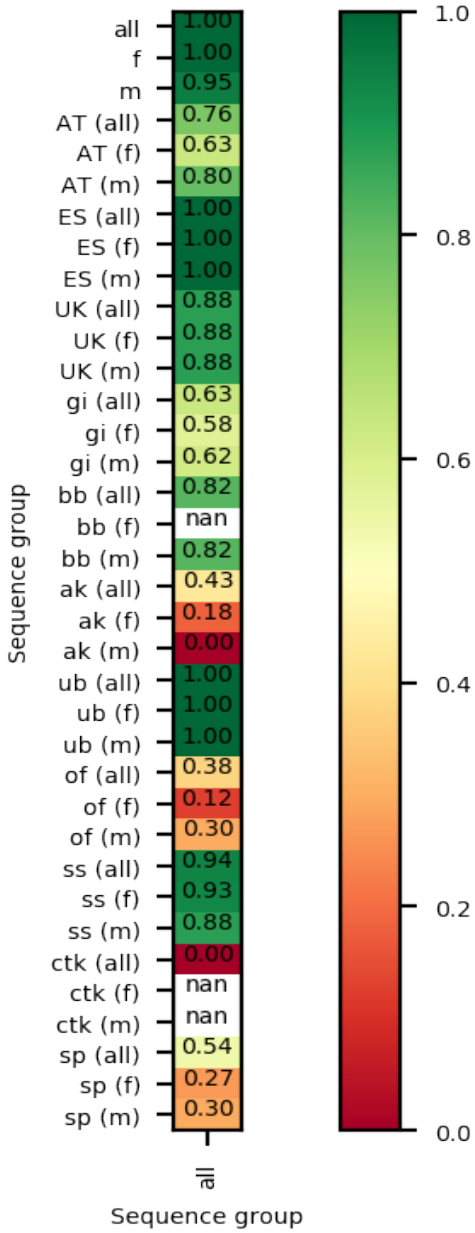

Figure 3.139: Most common sequences of length 4 between each groups and the one for all users

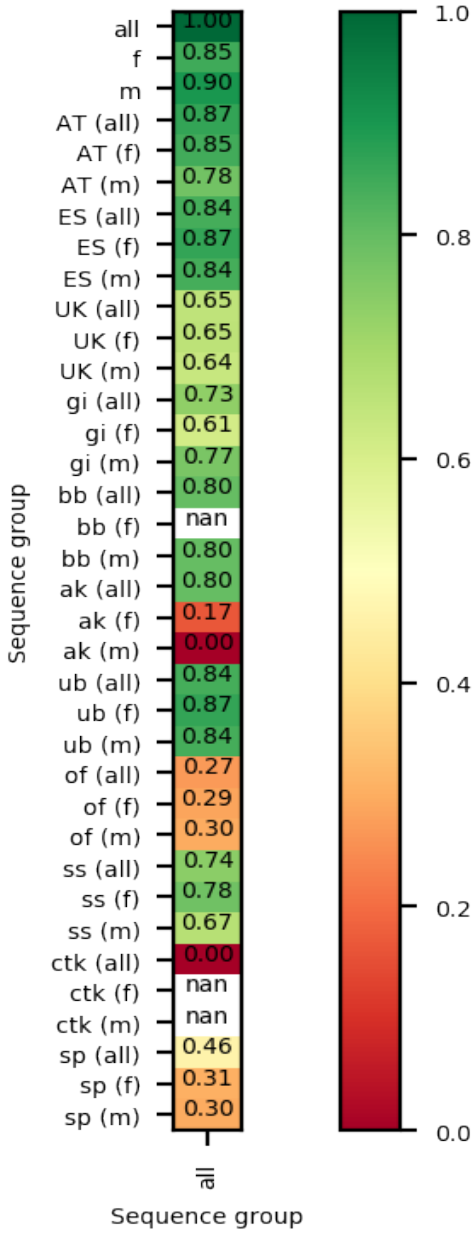

Figure 3.140: Most common sequences of length 5 between each groups and the one for all users

shown in figure 3.135, can be seen that the correlation is lower between the groups when compared with the results for length 1 shown in figure 3.131. The reason for this might be the different goals of the different countries for joining the NOLB project, that highly influenced the tasks that the children received and that plays important role in the results. For example the goal for the Austrian schools was to motivate and to raise the interest to girls for programming and creating games. So the goal of the classes in which Pocket Code was presented and used was games to be created. So the goal for the children of the three Austrian schools was clearly defined. The task and the reasons for using the application were very different in Spain. There the idea was that the application is used in schools with many international children that speak different languages. The idea was not to create something concretely like games, like in the case of the Austrian students, but rather to give opportunity to the children who can't speak to each other because of having different mother language, to be able to do something together. They had no defined task, but were only encouraged to explore the possibilities of Pocket Code together. The sequence of events used in free exploring of the application and of goal aiming are expected to be different.

For the purpose of the second test, the same input is used, or more precisely the results from the PrefixSpan. With the only difference that in this case the goal is not to see how much overlapping of sequences of events there is in percentage between the different groups, but rather to estimate the correlation between different groups or with other words to test the significance of the differences found by comparing the most frequent sequences of events from different groups. The results are obtained by estimating the Spearman's rank correlation coefficient and the p-values are presented in confusion matrices.

Spearman is implemented for this dataset in the following way: each time two different groups are compared, as for every other comparison done before, it is done between different countries/schools/genders. The most frequent sequences are listed from both groups, limited by the top 70 most used events in respect to every group. Then the numbers of occurrence of the most common sequences of both groups are considered. This number of occurrence are used to do the ranking for each of the groups for each of the most frequent sequences including the sequences from both groups. According to Spearman, the numbering starts from the highest ranking (ranked with 1) and the rank increases with the decreases of the number of times of occurrence of the sequences. In the next step the differences between the ranks between the two groups is estimated. Like for example if in the first group a certain sequence is ranked 4 and in the second group is ranked 6, then the difference

in this case is d = 4 - 6 and d is set to -2. The sum of all differences should be equal to 0. In the next step d^2 squared is estimated from all the obtained d-s. After this step is possible to estimate the Spearman's rank correlation coefficient. Then the rank r is estimated, then also n which is the number of pairs. If the result r is close to 1, then in a case of a scatter plot on xy-axis, with every increase in x, there will be an increase in y too. This means that there is strong positive correlation between the ranks of the two groups. If the result of r is close to -1 that will mean that with every decrease of x will result in decrease of y. And finally, r=0 indicates no association between the ranks. As result the z and the p value are obtained. In the results the p-values are given, for which the null hypothesis is that the two sets of data are not uncorrelated.

From the results obtained by the use of Spearman can be seen in the figures 3.141 and 3.142 that there is great correlation between the results of all the three countries in general as well as between the genders in the different countries. The results of the correlation between schools and between the genders between the schools is given in the figures 3.143 and 3.144 respectively. Finally, the correlation between each group and the the general group of all users, the results of the use of Spearman are given in figure 3.145, where as before the great correlation between two groups is marked with green and when two groups are not correlated is marked with red.

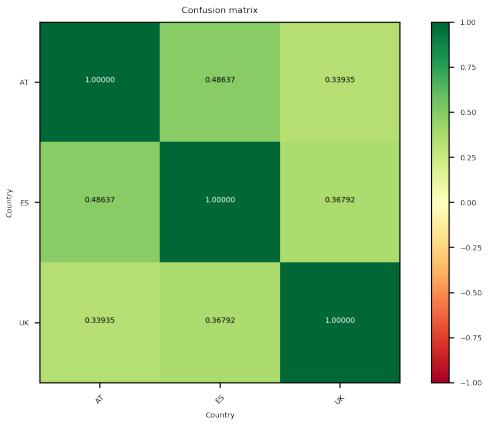

Figure 3.141: The results of Spearman between countries

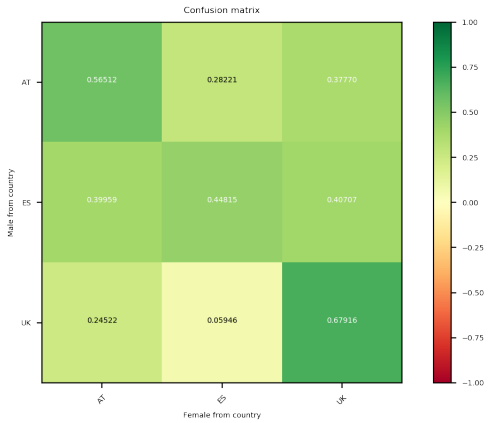

Figure 3.142: The results of Spearman between genders between countries

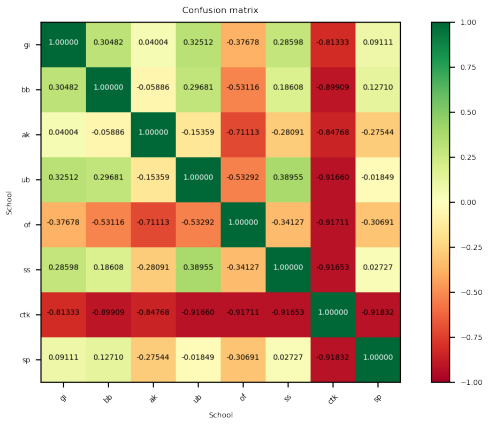

Figure 3.143: The results of Spearman between schools

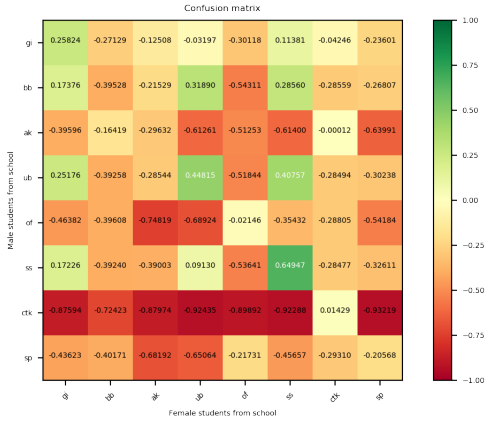

Figure 3.144: The results of Spearman between genders between schools

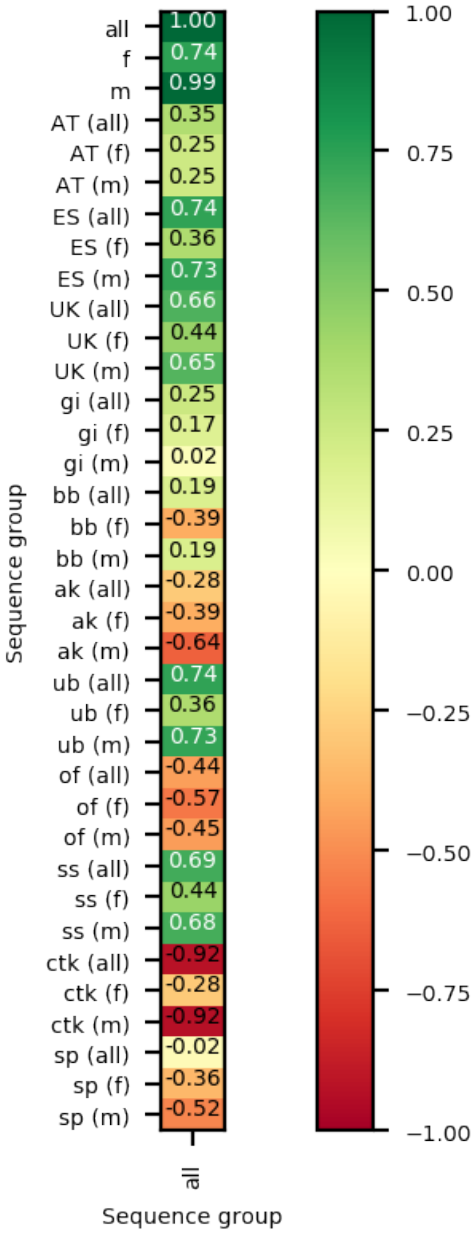

Figure 3.145: Spearman comparison for all groups

Chapter 4

Future work and conclusion

4.1 Conclusion

This master thesis was mainly focused on data analysis. First, general background information was presented regarding the available data and the NOLB project as a part of which this data was collected by the NOLB research group.

Secondly, an overview of relevant literature was presented. As part of it relevant ways to be able to make the data tidy, to analyze the data, outcomes of different ways for pattern mining and possible tests to check the significance of results, were listed.

Thirdly, all the steps and results were presented. The first step was cleaning and tidying the data. Next, the average events per user and the average events per session were estimated. From the results of both differences and similarities could be seen between different groups. After that the significance of the differences was tested by using Whitney Mann U test. Additionally, the number of times special events were used was taken into account, like 'help"" events, "advanced" events and reused "copied"" events were checked. Additionally, the difference of the number of events used in the first, compared to the last session. The number of times this events were used was also included in the results.

Since when searching for patterns it is not enough only to compare the amount of events the users use, or the amount of events the users use in a session, but it is also important which events exactly the user uses and in which sequence uses them most often. For this reason was used PrefixSpan with witch the most common sequences of 1,2 3, 4 and 5 were found for every group and subgroup. With this results was possible to compare whether the groups and the subgroups were using the same events most frequently, the

same sequences and the number of sessions in which they used them. It was trivial to compare the sequences of length one. But for the results of sequences of length 2, 3, 4 and 5 is difficult to make any comparison and conclusions. This is the reason why additional test was used, in order to find out whether there is significant difference between the most commonly used sequences of different groups, from the results obtained by the use of PrefixSpan previously. For this purpose 2 different tests were done with which was possible to see the overlapping and the correlation between the different groups. One using the similarities to estimate the overlapping and Spearman in order to estimate the correlation between different groups.

Some of the groups showed similar behaviour, others showed lower overlapping and lower correlation between each other. Sometimes the reason for the differences was the different behaviour of the students when using Pocket Code. Some of the groups showed preference over certain events compared to others, or used different amount of events per session or used them in a different order. But often the reason for differences was combination of factors that resulted in the presented differences, which does not not always necessarily mean different behaviour between the groups of users but rather different possibilities as part of the different experimental setup.

4.2 Future Work

The collection of the dataset is a unique possibility for analyzing because it contains international tracking on the same application. There is further potential for development and use of the dataset.

During this project the events were tracked during the lectures of different subjects. For example, in Austria this was mostly done during math or physics classes. Possibility for future work can be to discover how much the interest in the subject influences in the performance of the children. This might bring new conclusions and ideas for different experimental setup.

Another possibility for future work is to conduct similar study and make the students do the task both individually and in pairs. Then girls should show better results when working in group compared to working individually, according to the relevant literature in chapter 2. Then the conclusions and the tables of most commonly used events will most likely be different from the results when girls work individually.

One more opportunity is to purely focus on the data from Spain. In the chapter 2, details are given on a study that claims that the mother tongue of

the children in the past influences on how much the children are confident to start programming or to have a profession in IT. That is claimed to influence on their confidence, role model, and their ambition for the the computer science world. Since the goal for Spain was to track international students this data is available for analysis and can be used to discover whether children speaking different language at home than the one spoken in the country still today has such an influence as explained in the paper. Because the study made for the purpose of the paper is already done quite some time ago.

The same study should be done again in the future because most of the students should be in college now, and then the results can be compared whether the trend is the same as in the years 2004 and 2005 when studies were done which are included in the chapter 2. And there are compared the student knowledge before and after studying computer science, but the students had to answer according to constrain cards in order the researchers to get the idea how prepared they arrived at university. If we would manage to collect the data from the same students that are part of this dataset or at least from some of them, after they graduate bachelor in CS, we would have more accurate results and answers to this question, since the students will not need cards with defined questions to tell when and how they first met programming, but instead there will be more realistic knowledge about them and can be seen how much they improved during the years and whether presenting programming in early age as part of games really motivates students, especially female students to continue in IT studies.

Appendix A

PrefixSpan

For the coding part and use of PrefixSpan was used: https://pypi.org/project/prefixspan/

For brevity, only the results for length 1 and length 5 were included in the thesis. In this part are included the results for length 2,3 and 4 for each school.

The results from Akademische Gymnasium:(A.1-A.9)

The results from Graz International Bilingual School: (A.10-A.18)

The results from Borg Birkfeld:(A.19-A.25)

The results from Ubeda:(A.26-A.34)

The results from Oakfield School:(A.35-A.43)

The results from St.Stephan Primary:(A.44-A.52)

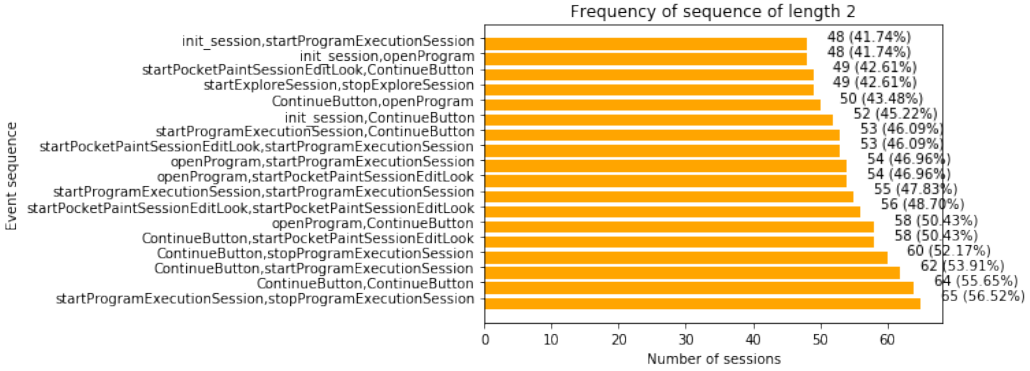

Figure A.1: Sequences of length 2 in a session for all users from ak with PrefixSpan

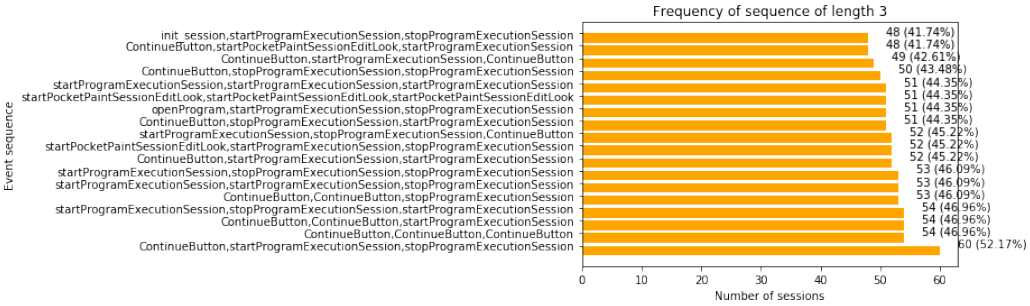

Figure A.2: Sequences of length 3 in a session for all users with PrefixSpan

Figure A.3: Sequences of length 4 in a session for all users from ak with PrefixSpan

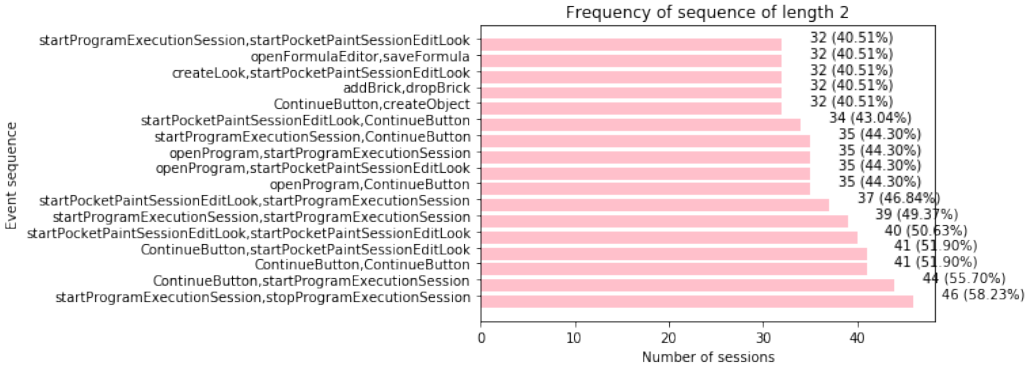

Figure A.4: Sequences of length 2 in a session for all female users from ak with PrefixSpan

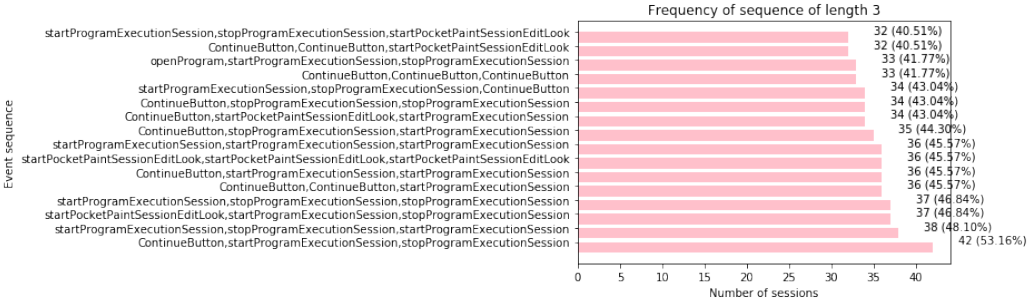

Figure A.5: Sequences of length 3 in a session for all female users with PrefixSpan

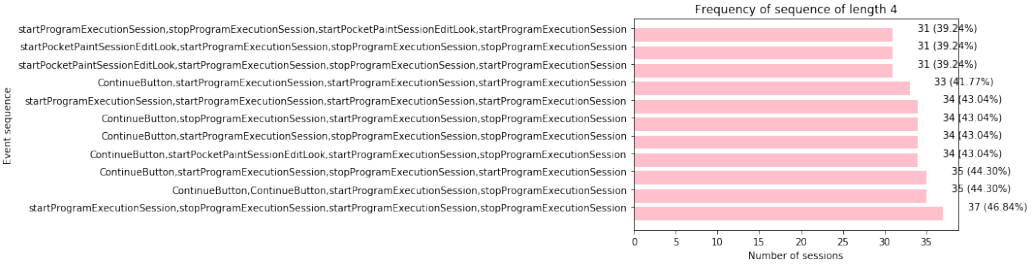

Figure A.6: Sequences of length 4 in a session for all female users with PrefixSpan

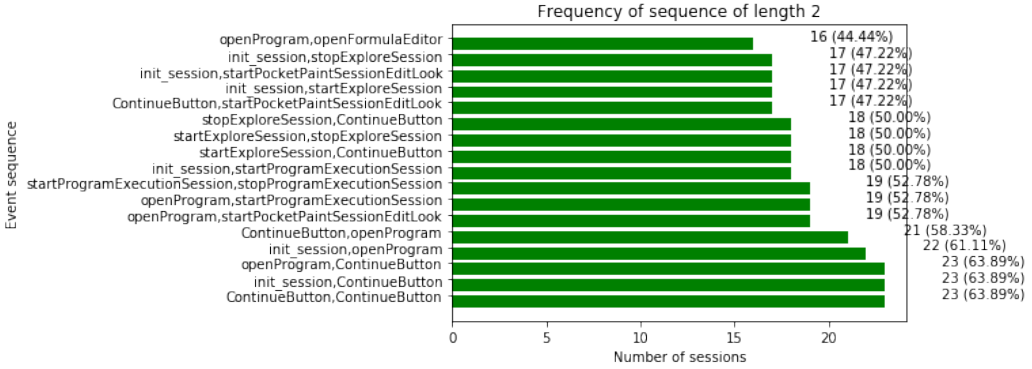

Figure A.7: Sequences of length 2 in a session for all male users from ak with PrefixSpan

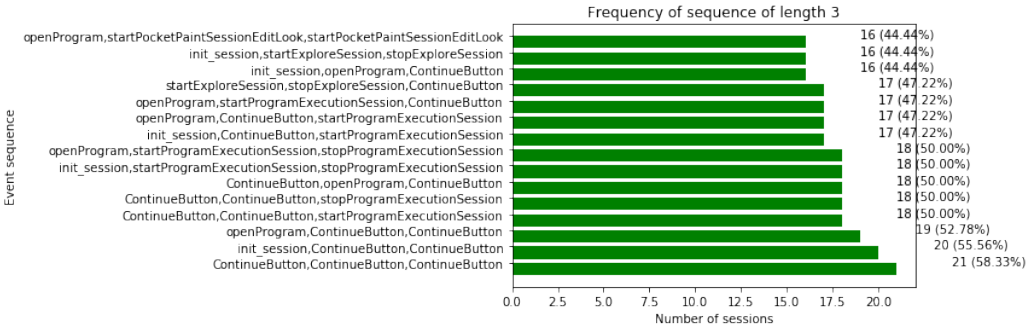

Figure A.8: Sequences of length 3 in a session for all male users from ak with PrefixSpan

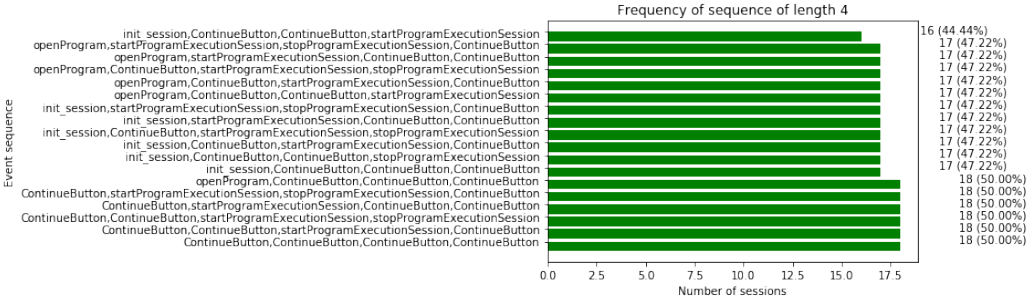

Figure A.9: Sequences of length 4 in a session for all male users from ak with PrefixSpan

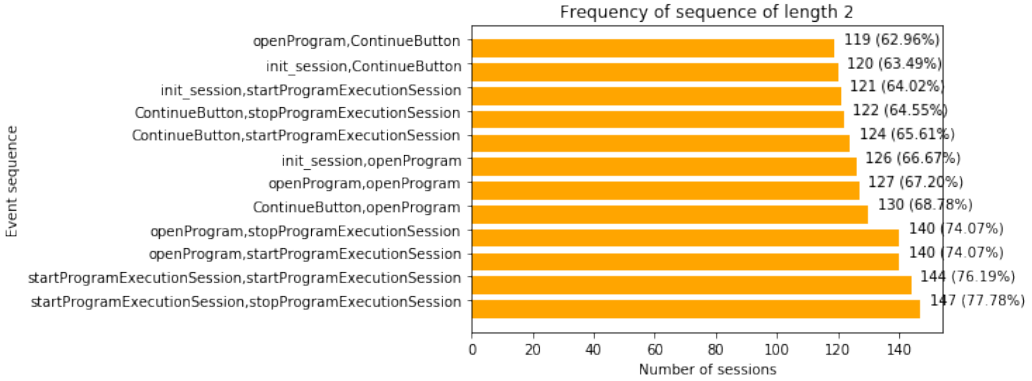

Figure A.10: Sequences of length 2 in a session for all users from GIBS with PrefixSpan

Figure A.11: Sequences of length 3 in a session for all users from GIBS with PrefixSpan

Figure A.12: Sequences of length 4 in a session for all users from GIBS with PrefixSpan

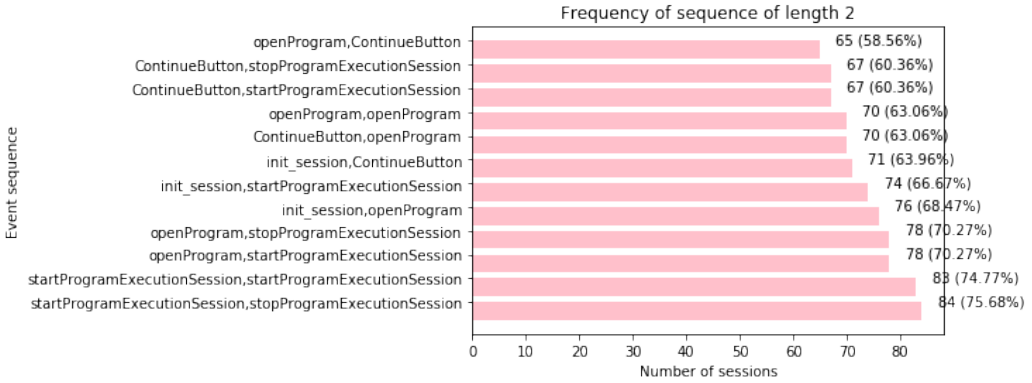

Figure A.13: Sequences of length 2 in a session for all female users from GIBS with PrefixSpan

Figure A.14: Sequences of length 3 in a session for all female users from GIBS with PrefixSpan

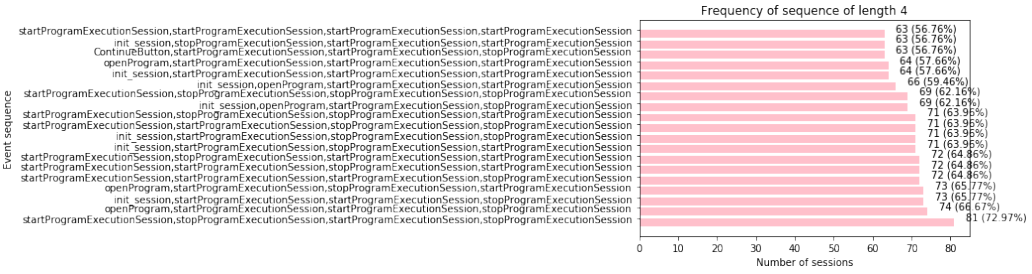

Figure A.15: Sequences of length 4 in a session for all female users from GIBS with PrefixSpan

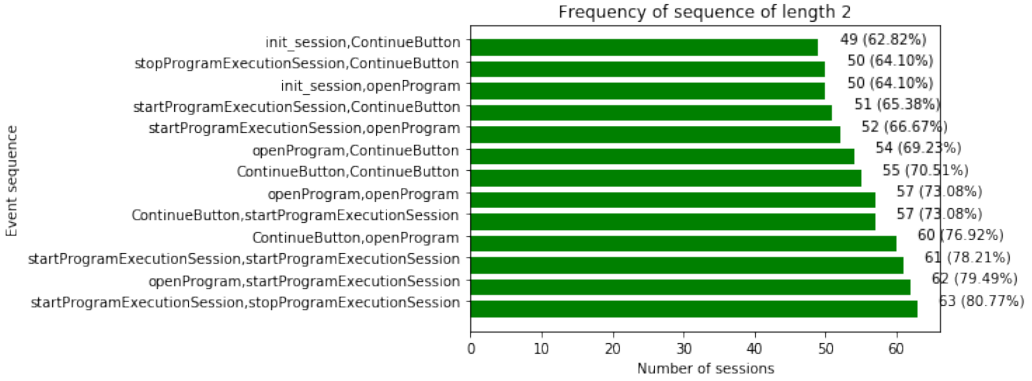

Figure A.16: Sequences of length 2 in a session for all male users from GIBS with PrefixSpan

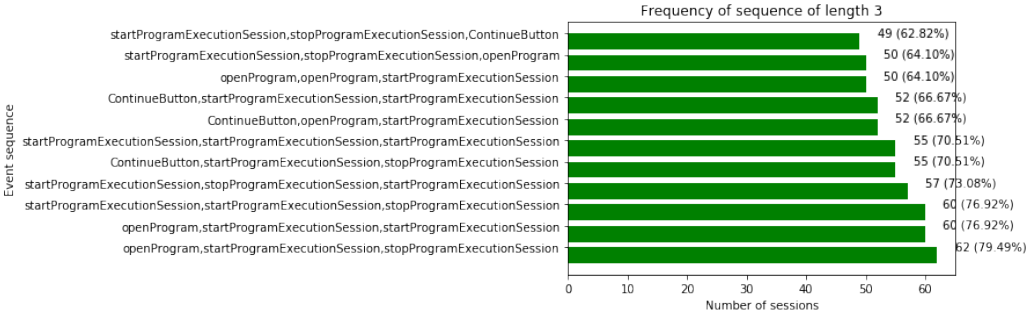

Figure A.17: Sequences of length 3 in a session for all male users from GIBS with PrefixSpan

Figure A.18: Sequences of length 4 in a session for all male users from GIBS with PrefixSpan

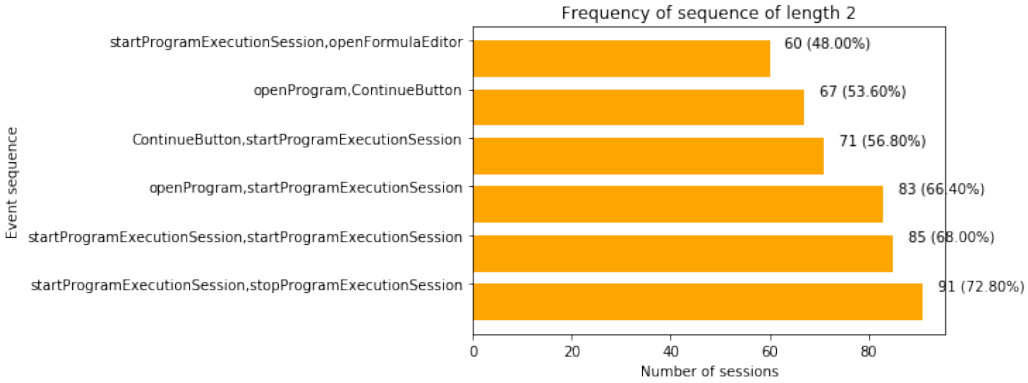

Figure A.19: Sequences of length 2 in a session for all users from bb with PrefixSpan

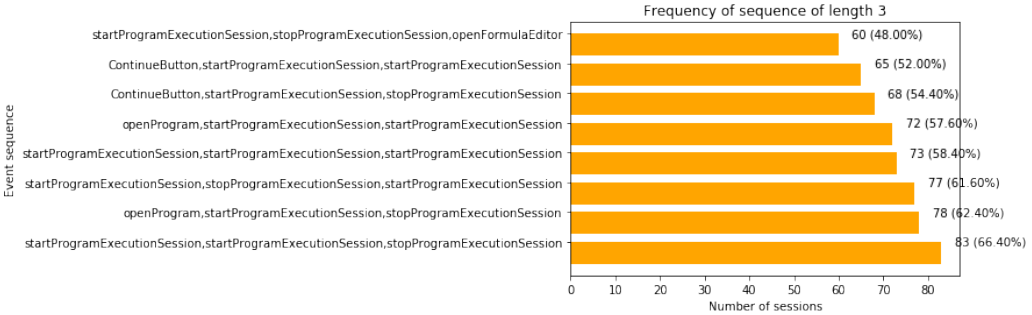

Figure A.20: Sequences of length 3 in a session for all users from bb with PrefixSpan

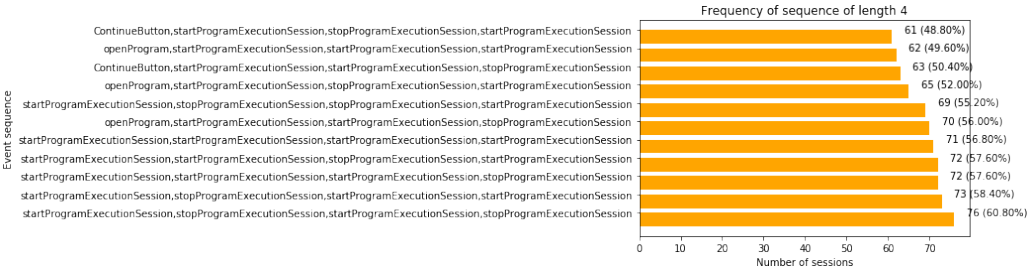

Figure A.21: Sequences of length 4 in a session for all users from bb with PrefixSpan

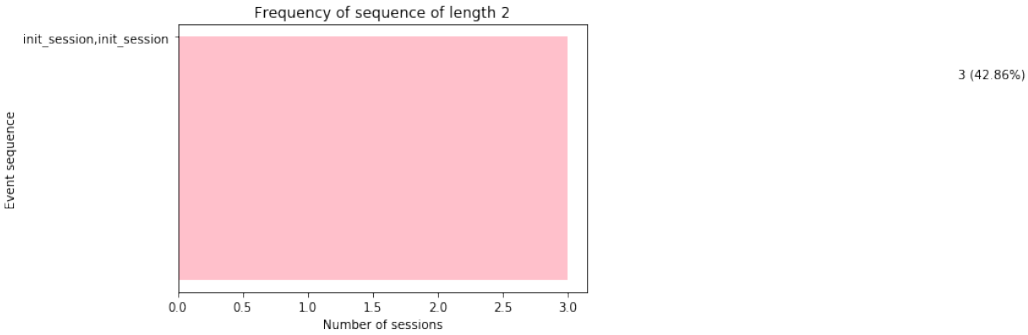

Figure A.22: Sequences of length 2 in a session for all female users from bb with PrefixSpan

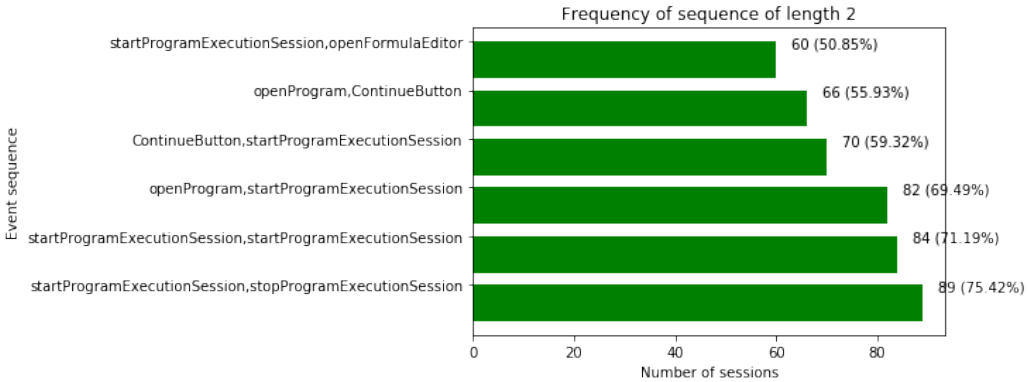

Figure A.23: Sequences of length 2 in a session for all male users from bb with PrefixSpan

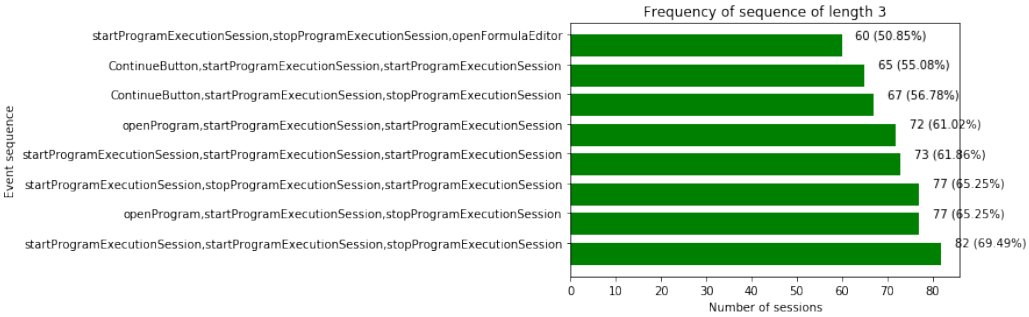

Figure A.24: Sequences of length 3 in a session for all male users from bb with PrefixSpan

Figure A.25: Sequences of length 4 in a session for all male users from bb with PrefixSpan

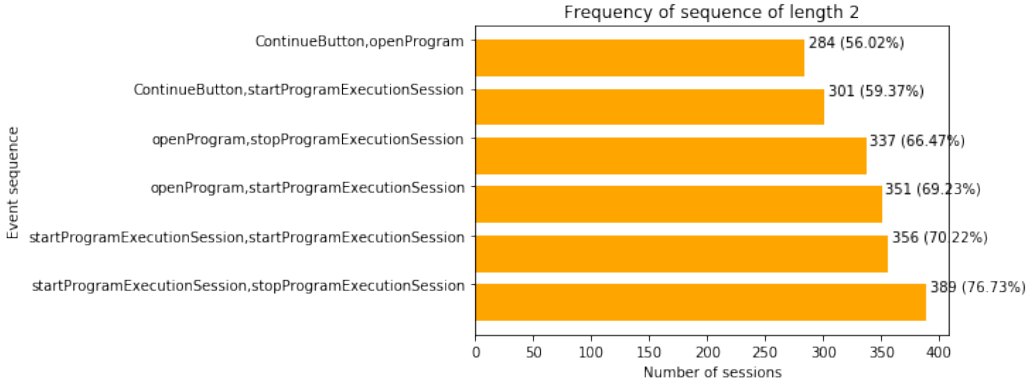

Figure A.26: Sequences of length 2 in a session for all users from ub with PrefixSpan

Figure A.27: Sequences of length 3 in a session for all users from ub with PrefixSpan

Figure A.28: Sequences of length 4 in a session for all users from ub with PrefixSpan

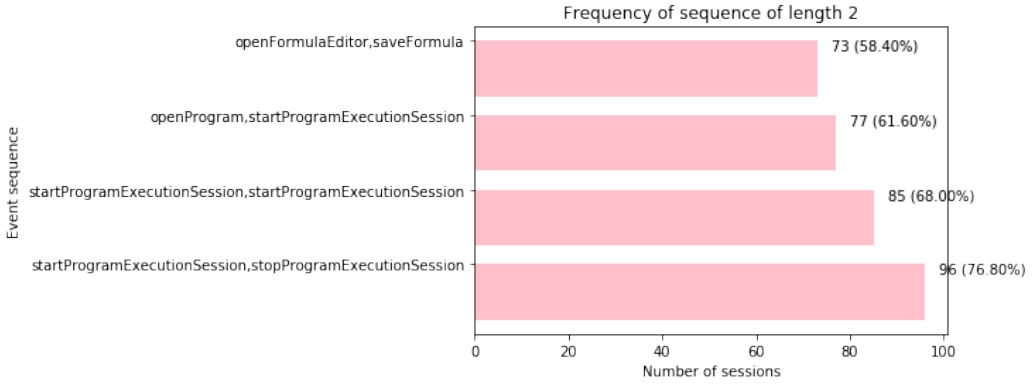

Figure A.29: Sequences of length 2 in a session for all female users from ub with PrefixSpan

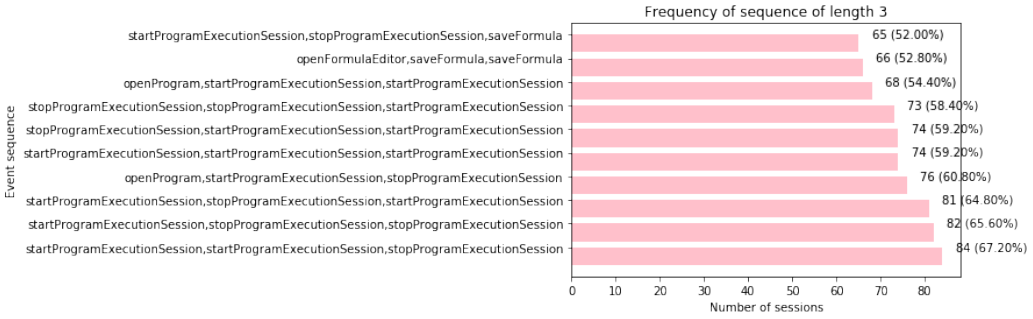

Figure A.30: Sequences of length 3 in a session for all female users from ub with PrefixSpan

Figure A.31: Sequences of length 4 in a session for all female users from ub with PrefixSpan

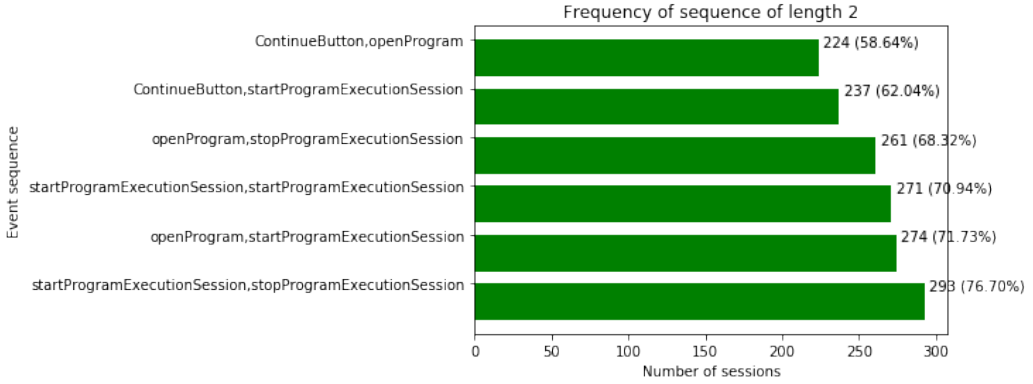

Figure A.32: Sequences of length 2 in a session for all male users from ub with PrefixSpan

Figure A.33: Sequences of length 3 in a session for all male users from ub with PrefixSpan

Figure A.34: Sequences of length 4 in a session for all male users from ub with PrefixSpan

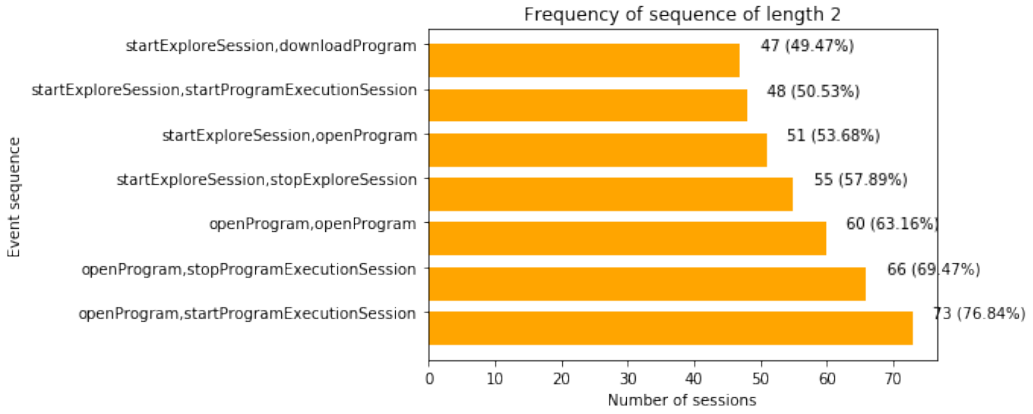

Figure A.35: Sequences of length 2 in a session for all users from of with PrefixSpan

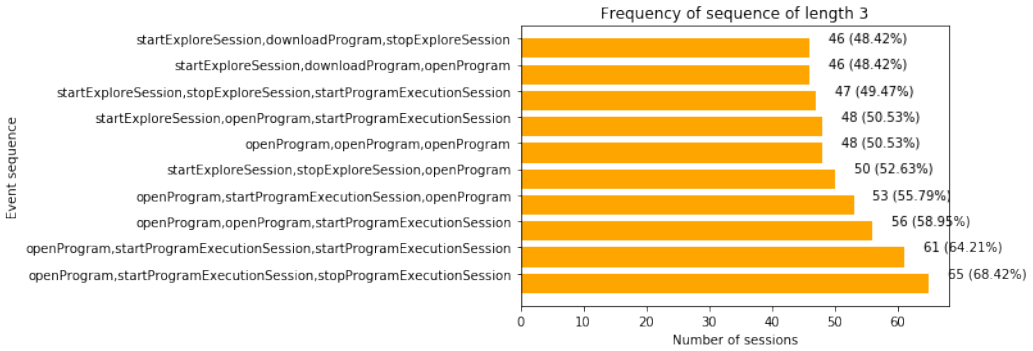

Figure A.36: Sequences of length 3 in a session for all users from of with PrefixSpan

Figure A.37: Sequences of length 4 in a session for all users from of with PrefixSpan

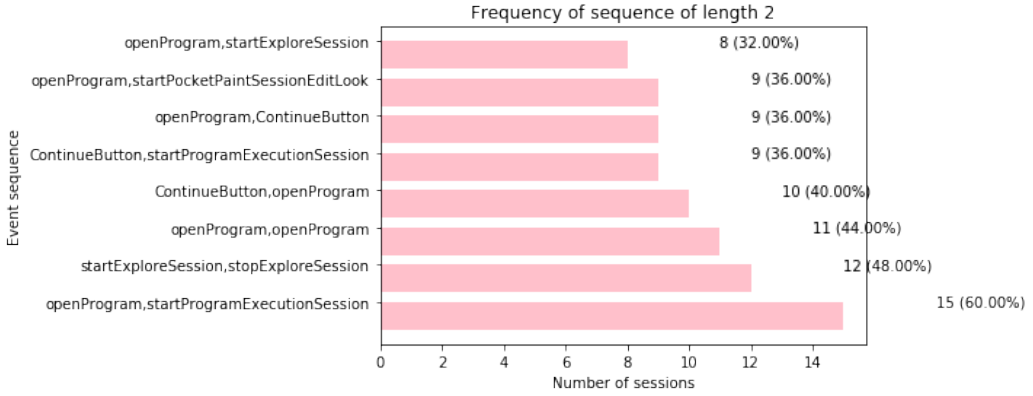

Figure A.38: Sequences of length 2 in a session for all female users from of with PrefixSpan

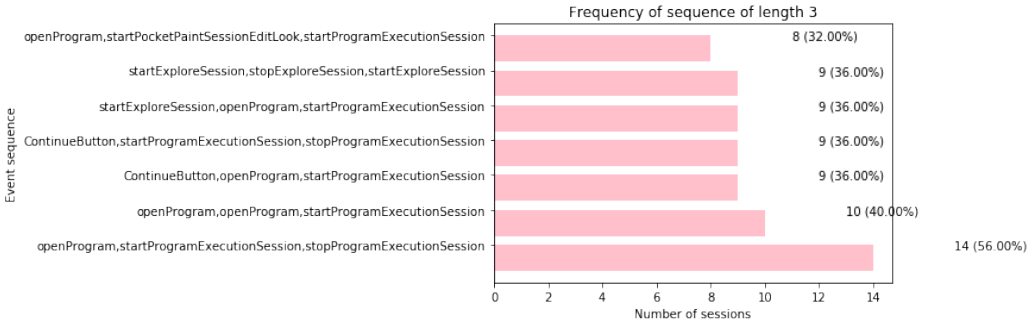

Figure A.39: Sequences of length 3 in a session for all female users from of with PrefixSpan

Figure A.40: Sequences of length 4 in a session for all female users from of with PrefixSpan

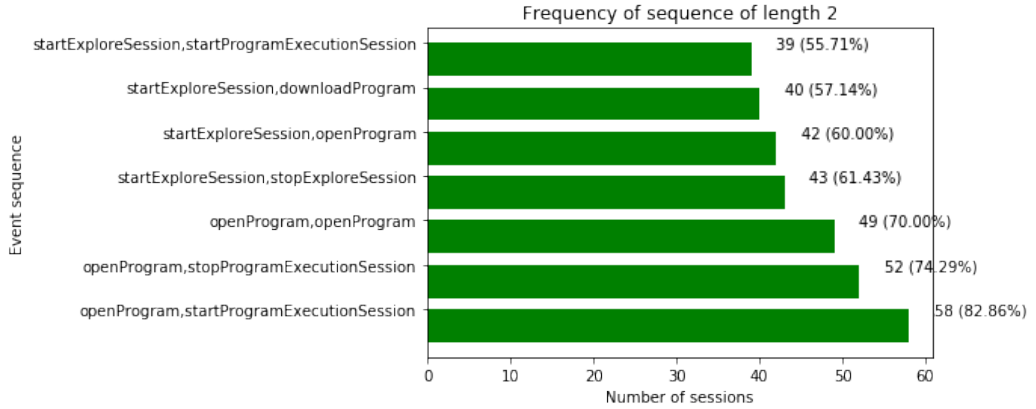

Figure A.41: Sequences of length 2 in a session for all male users from of with PrefixSpan

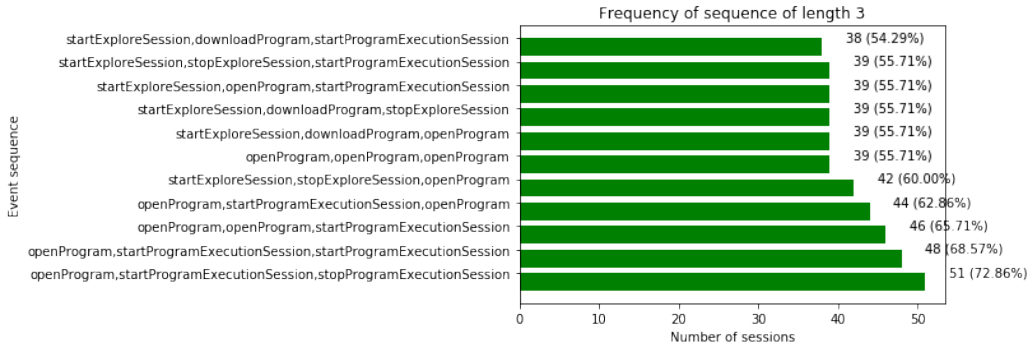

Figure A.42: Sequences of length 3 in a session for all male users from of with PrefixSpan

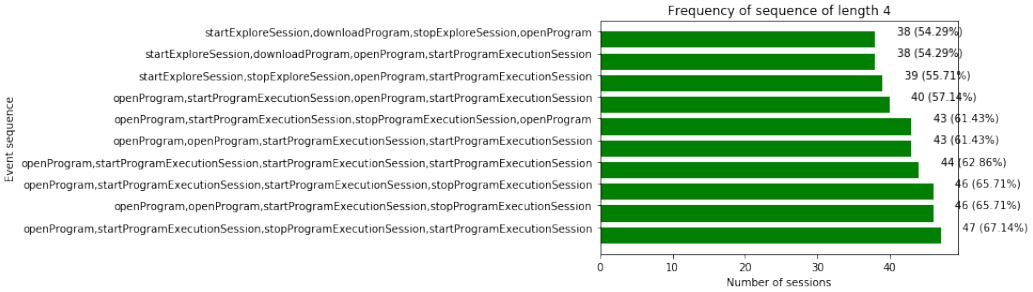

Figure A.43: Sequences of length 4 in a session for all male users from of with PrefixSpan

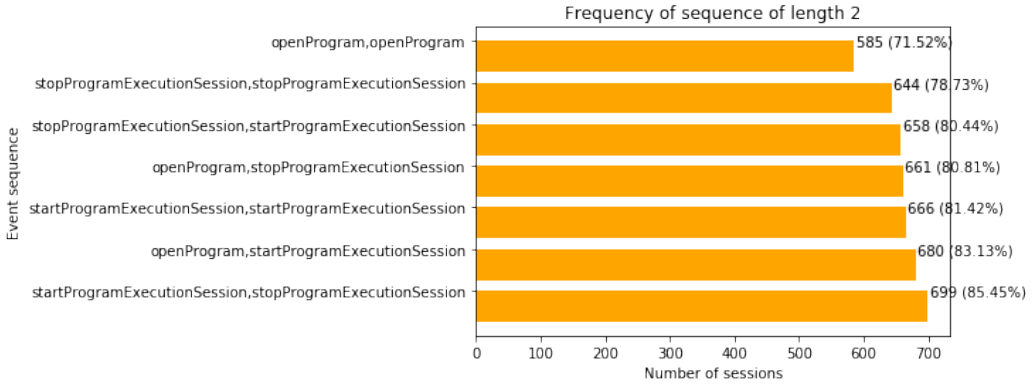

Figure A.44: Sequences of length 2 in a session for all users from ss with PrefixSpan

Figure A.45: Sequences of length 3 in a session for all users from ss with PrefixSpan

Figure A.46: Sequences of length 4 in a session for all users from ss with PrefixSpan

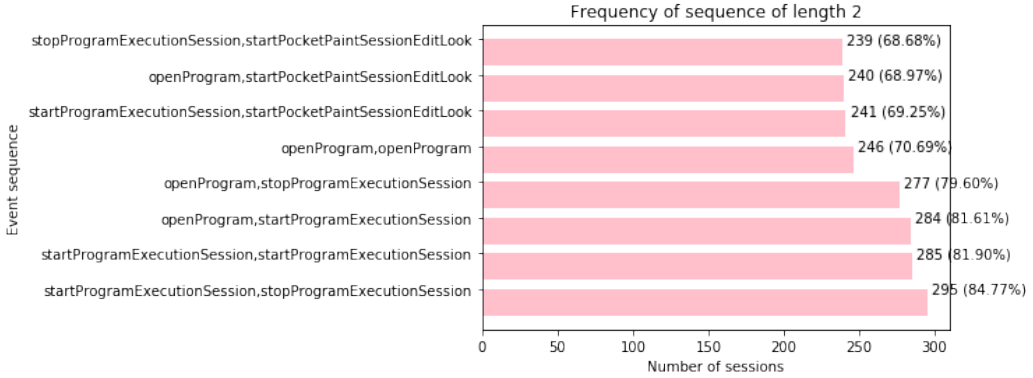

Figure A.47: Sequences of length 2 in a session for all female users from ss with PrefixSpan

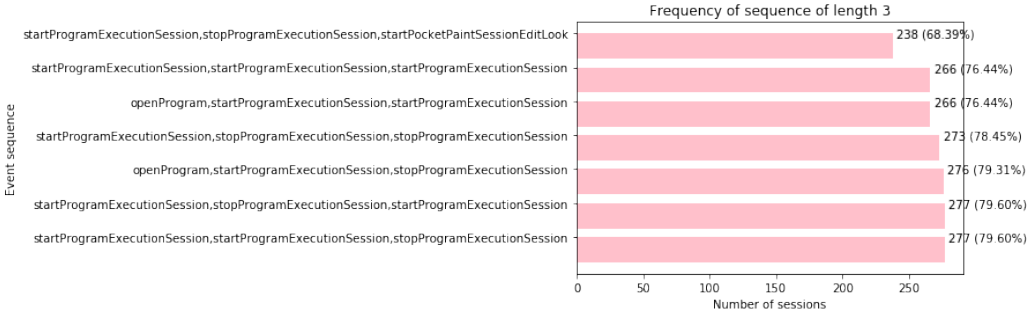

Figure A.48: Sequences of length 3 in a session for all female users from ss with PrefixSpan

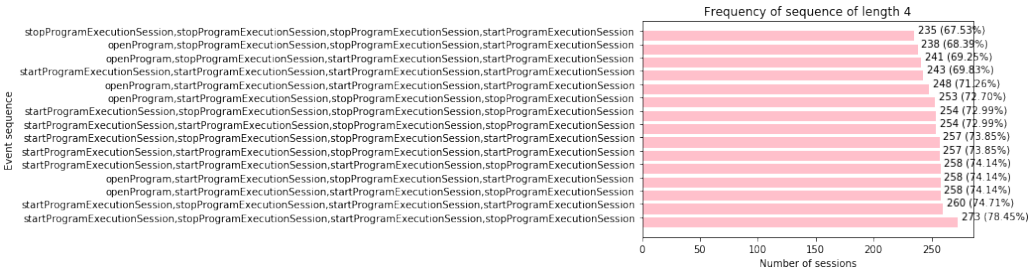

Figure A.49: Sequences of length 4 in a session for all female users from ss with PrefixSpan

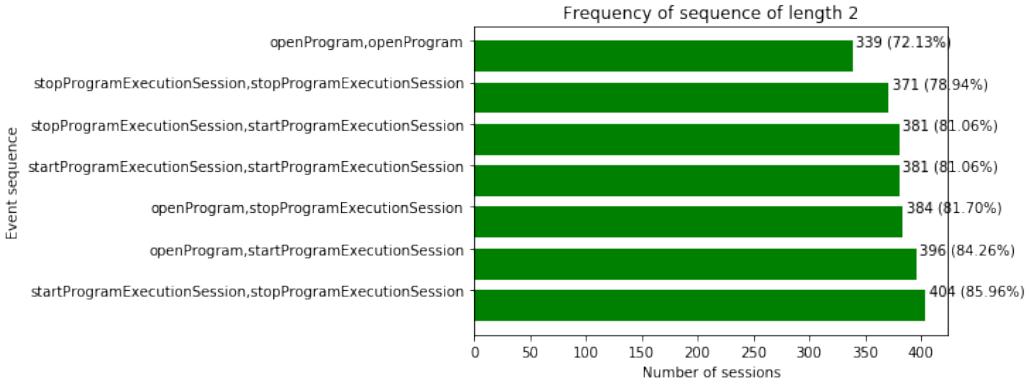

Figure A.50: Sequences of length 2 in a session for all male users from ss with PrefixSpan

Figure A.51: Sequences of length 3 in a session for all male users from ss with PrefixSpan

Figure A.52: Sequences of length 4 in a session for all male users from ss with PrefixSpan

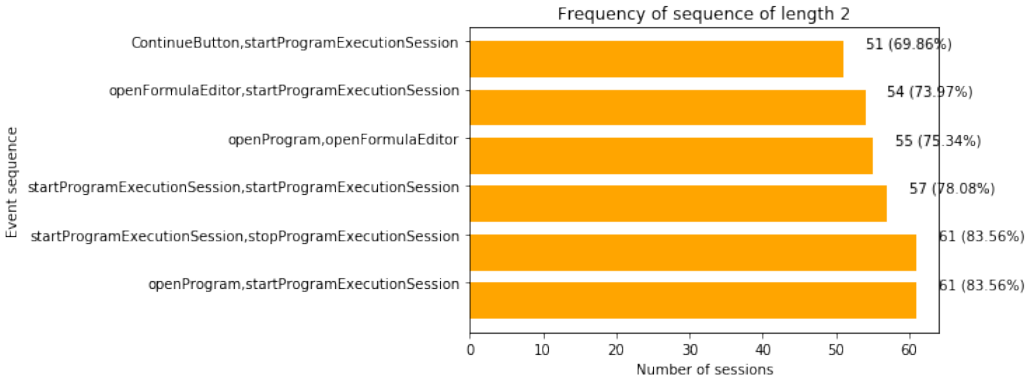

Figure A.53: Sequences of length 2 in a session for all male users from sp with PrefixSpan

Bibliography

Becta (2008). *How do boys and girls differ in their use of ICT?* Tech. rep. August.

Broeck, J. V. D., Cunningham, S. A., Eeckels, R., and Herbst, K. (2005). "Data Cleaning : Detecting , Diagnosing , and Editing Data Abnormalities". 2(10), pp. 966–970.

Cox, N. J. (2004). "Review: Exploratory Data Mining and Data Cleaning". *Journal of Statistical Software*, 11(September), pp. 1–3.

Denner, J. (2007). "The Girls Creating Games Program: An innovative approach to integrating technology into middle school". *Meridian: A Middle School Computer Technologies ...*

Ebner, M., Janisch, S., Höllerbauer, B., Grandl, M., and Slany, W. (2017). "Pocket Code – Programmieren für Alle mit einem offenen Online-Kurs". (April), pp. 16–19.

Fairlie, R. W. (2016). "Do Boys and Girls Use Computers Differently, and Does It Contribute to Why Boys do Worse in School Than Girls?" *B.E. Journal of Economic Analysis and Policy*, 16(1), pp. 59–96.

Funke, A., Berges, M., Mühling, A., and Hubwieser, P. (2015). *Gender Differences in Programming : Research Results and Teachers ' Perception.* Tech. rep., pp. 161–162.

Han, J., Pei, J., and Yin, Y. (n.d.). "Mining FrequentPatterns without Candidate Generation" (), pp. 317–318.

Hsieh, T.-y. and Wu, K.-c. (2014). "The Influence of Gender Difference on the Information-Seeking Behaviors for the Graphical Interface of Children's Digital Library." *Universal Journal of Educational Research*, 3(3), pp. 200–206.

Jian Pei, Jiawei Han, Mortazavi-Asl, B., Pinto, H., Qiming Chen, Dayal, U., and Mei-Chun Hsu (n.d.). "PrefixSpan,: mining sequential patterns

efficiently by prefix-projected pattern growth". *Proceedings 17th International Conference on Data Engineering* (), pp. 215–224.

Magon, A. J. and Blunt, A. (2009). "Gender, the Brain, and Education: Do Boys and Girls Learn Differently?" PhD thesis, p. 108.

Murphy, L., Richards, B., McCauley, R., Morrison, B. B., Westbrook, S., and Fossum, T. (2006). "Women Catch Up: Gender Differences in Learning Programming Concepts". *ACM SIGCSE Bulletin*, 38(1), p. 17.

Nachar, N. (2008). "The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution". *Tutorials in Quantitative Methods for Psychology*, 4(1), pp. 13–20.

Nasser, A. (2016). "The Difference Between Girls and Boys in Learning". *Defense Language Institute Foreign Language School*, 1(December).

O'Connor, D., McCormack, M., Robinson, C., and O'Rourke, V. (2017). "BOYS AND GIRLS COME OUT TO PLAY: GENDER DIFFERENCES IN CHILDREN'S PLAY PATTERNS". In: pp. 4713–4719.

Pei, J., Han, J., Helen, B. M.-a., Science, C., E-mail, C. V. A., Chen, Q., and Dayal, U. (n.d.). *PrefixSpan——Mining Sequential Patterns Efficiently by Prefix-Projected Pattern.pdf*.

Rahm, E. and Do, H. (2000). "Data cleaning: Problems and current approaches". In: *IEEE Data Engineering Bulletin*. Vol. 23. 4, pp. 3–13.

Resnick, M., Silverman, B., Kafai, Y., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., and Silver, J. (2009a). "Scratch". *Communications of the ACM*, 52(11), p. 60.

— (2009b). "Scratch". *Communications of the ACM*, 52(11), p. 60.

Roig, A. E. (2004). "The digital gap: gender and computer games". 9(9), pp. 1–15.

Slany, W. (2014). "Pocket Code: A Scratch-like Integrated Development Environment for Your Phone". In: *Proceedings of the Companion Publication of the 2014 ACM SIGPLAN Conference on Systems, Programming, and Applications: Software for Humanity*. SPLASH '14. Portland, Oregon, USA: ACM, pp. 35–36.

Spieler, B., Hummel, D., Herbertz, S., and Mädche, A. (2018). "Reinforcing gender equality by analysing female teenagers' performances in coding activities: a lesson learned". In: *Proceedings of the 4th Conference on Gender & IT - GenderIT '18*. August, pp. 209–219.

Srikant, R. and Agrawal, R. (1996). "Mining sequential patterns: Generalizations and performance improvements". In: pp. 1–17.

Volkovs, M. and Chiang, F. (n.d.). "Continuous Data Cleaning" ().

Wickham, H. (2007). "Reshaping Data with the reshape Package". 21(12).

— (2014a). "Tidy Data". *Journal of Statistical Software*, 59(10), pp. 1–23.

— (2014b). "Tidy Data". *Journal of Statistical Software*, 59(10), pp. 1–23.