Peter Müllner, BSc

Studying Non-Mainstream Music Listening Behavior For Fair Music Recommendations

Master Thesis

for the attainment of the degree of

Diplom-Ingenieur (Dipl.-Ing.)

submitted to

Graz University of Technology

Supervisor:
Ass.-Prof. Dr. techn. Dipl.-Ing. Elisabeth Lex

Advisor:
Dr. techn. Dipl.-Ing. Dominik Kowald, BSc

Institute for Interactive Systems and Data Science

Graz, October 2019

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit/Diplomarbeit/Dissertation identisch.

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/diploma thesis/doctoral dissertation.

_____ _____

Datum / Date Unterschrift / Signature

Abstract

Recommender systems are widely used in many domains. Despite large research efforts regarding, e.g., context-aware recommender systems, it remains a fundamental problem that recommendations are biased towards popularity. This can be mainly attributed to the long-tailed distribution of items. In particular, recommendation of music is heavily influenced by this effect, due to particularities of music. Especially listeners of non-mainstream music from the long tail perceive low recommendation quality. Our work is inspired by recent research in fair music recommendations concerned with proving non-mainstream users being disadvantaged by today's recommendation algorithms. We apply unsupervised clustering and classification to identify non-mainstream music styles. Subsequently, we link users to their favorite music style and thus, obtain user groups of different taste. This opens the opportunity to conduct in-depth analyses of non-mainstream users in regard to their music taste, demography, culture and listening behavior. We found that there are indeed different types of non-mainstream users: Complex (i.e., Blues, Soul), Festival (i.e., Punk, Hardrock), Relax (e.g., Ambient, Postrock) and Heavy Listeners (e.g., Deathmetal, Blackmetal). Among several findings, results indicate that Heavy and Festival Listeners are of younger age than others. Interestingly, Complex and Relax Listeners tend to exhibit more diverse music taste and favor future developments. Hence, our research shows that cultural aspects can be utilized to further describe this understudied group of non-mainstream users. In addition, we extend previous work in the sense that we not only illustrate the discrimination of non-mainstream users, but also provide significant evidence that recommendation quality for this subset of users varies for different types of listeners. Thus, performance of state-of-the-art recommendation algorithms in regard to non-mainstream users could be improved by considering both, the notion of mainstreaminess and music styles within the long tail.

Zusammenfassung

Empfehlungssysteme werden in vielen verschiedenen Bereichen genutzt. Trotz großem Forschungsaufwand bezüglich Empfehlungssystemen, die Kontext betrachten, besteht nach wie vor das Problem der Beeinflussung durch Popularität. Aufgrund der Besonderheiten von Musik sind Musikempfehlungen besonders stark betroffen. Die Qualität der Empfehlungen ist im Speziellen für Nutzer schlecht, die eher unpopuläre Musik hören. Forschung im Bereich der fairen Musikempfehlungen zeigt die Benachteiligung von Hörern unpopulärer Musik. Dies wirkt als Inspiration für diese Arbeit. Wir identifizieren verschiedene Musikgeschmäcker und klassifizieren Nutzer anhand ihrer Lieblingsmusik. Dies ermöglicht es uns, Hörer unpopulärer Musik in Anbetracht ihres Geschmacks, ihrer Demographie, ihrer Kultur und ihrer Interaktion mit Musik näher zu analysieren. Hörergruppen von vier Stilen konnten dabei identifiziert werden: Komplex (e.g., Blues, Soul), Festival (e.g., Punk, Hardrock), Ruhig (e.g., Ambient, Postrock) und Heavy (e.g., Deathmetal, Blackmetal). Wie wir herausfanden sind Heavyhörer und Festivalhörer jünger als andere Nutzergruppen. Weiters zeigen Hörer der Stile Komplex und Ruhig einen höheren Grad an Vielseitigkeit. Unter anderem legen die zuletzt angeführten Nutzer weniger Wert auf den Erhalt von Tradition, als Heavyhörer und Festivalhörer. Unsere Forschungsergebnisse zeigen, dass mittels kultureller Aspekte Nutzer mit unpopulärem Musikgeschmack näher beschrieben werden können. Zusätzlich erweitern wir die aktuelle Forschung, indem wir nicht nur verifizieren, dass Nutzer unpopulären Geschmacks von modernen Empfehlungsalgorithmen benachteiligt werden, sondern zeigen auch, dass die Qualität der Empfehlungen für diese Nutzer ebenso durch ihren Musikgeschmack beeinflusst wird. Kurzgefasst kann die Qualität der Empfehlungen für Nutzer unpopulären Geschmacks verbessert werden, indem man auch verschiedene unpopuläre Musikstile betrachtet.

"I found myself in the position of that child in a story who noticed a bit of string and - out of curiosity - pulled on it to discover that it was just the tip of a very long and increasingly thick string ... and kept bringing out wonders beyond reckoning."

Benoit B. Mandelbrot, Fractalist

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Music recommendation systems (MRS) aim to provide tracks, artists, genres or playlists, the user may like. Typically, contextual information [Cheng and Shen, 2014, Park et al., 2006, Levy and Sandler, 2008] or previous listening history [Zheleva et al., 2010] are exploited to find well-suited music. [Schedl et al., 2018] described several particularities of music recommendations, for instance, the abundance of music data and a user's listening intent, situation or emotions. Furthermore, there exists only a weak linkage between low-level descriptions (e.g., audio signals) and high-level descriptions (e.g., genre annotations, emotions) of music. This so called semantic gap [Celma et al., 2006, Aucouturier, 2009] poses problems in the field of music recommendations, since pieces of music similar in their abstract, high-level descriptions do not have to be similar in their low-level descriptions. That means, the notion of similarity of a system that solely relies on music data deviates from a user's notion of similarity. As a consequence, there exists a discrepancy between what a user perceives similar to her music taste and what a recommender system regards as well-suited for a user. Thus, recommender systems have to grasp contextual aspects, in order to alleviate this gap. Many factors are thought to heavily influence the perceived quality of music recommendations, but to this end, are hard to include into a recommender system. Caused by the aforementioned aspects, research strives to develop sophisticated methods [Schedl and Hauger, 2015, Kowald et al., 2019] and models [Van den Oord et al., 2013, Zheng et al., 2019]. Another strand of research is devoted to model user-behavior [Zangerle and Pichl, 2018]. Despite the large body of approaches destined to improve music recommendations, recent research shows

that (i) methods fail to give stable results over varying datasets, all representing some kind of user to item rating (e.g., Epinions[1], citeulike-a[2], Pinterest[3]) [Dacrema et al., 2019] and (ii) are uncapable of providing satisfying recommendation quality for users of non-popular, unorthodox taste [Schedl and Bauer, 2017]. The notion of mainstreaminess [Bauer and Schedl, 2019] gained importance in research concerned with studying subpopulations that do not comply with global taste. [Abdollahpouri et al., 2019b] found that recommendations for users of low mainstreaminess are biased towards popular items and hence, lack in resembling a user's non-mainstream listening behavior. Research offers several approaches to tackle this crucial issue [Steck, 2018, Abdollahpouri et al., 2019a], laying strong focus on personalized recommendations. Since music consumption is partially guided by the user's psychological profile, background and intent, practitioners likely profit from research work not only concerned with recommender systems, but with shedding light on music consumers themselves. As an example, psychologists linked music preferences to personal characteristics [Delsing et al., 2008]. Additionally, [Mulder et al., 2007] defined several groups of music listeners with distinct music taste and listening behavior.

In this work, we strive to attach to both, computer scientists' and psychologists' work. More specifically, we aim to understand the understudied group of non-mainstream users with unorthodox music taste. Those users' preferred music style is modelled by exhibiting low-level and high-level descriptions of tracks. We furthermore aim to outline non-mainstream user's characteristics, which serve as starting point for future work that is driven by providing fair recommendations for both, users of mainstream and non-mainstream taste. Additionally, this thesis intends to identify distinct groups of non-mainstream users and eventually depict their unique listening behavior, demography and culture.

[1]http://www.trustlet.org/downloaded_epinions.html
[2]http://www.cs.cmu.edu/ chongw/data/citeulike/
[3]http://sites.google.com/site/xueatalphabeta/

1.1 Research Questions

In this section we define four research questions which, as we think, are of high importance for alleviating the aforementioned issues regarding the role of non-mainstream music listeners in the field of today's music recommender systems.

RQ1: How can non-mainstream music styles be identified and concisely described? This research question is driven by the observation that music typically exhibits very different patterns in terms of audio signals. There exists an abundance of high-level descriptors of music like, e.g., genres, which - in this case - constitutes a large taxonomy of genres. Anyway, [Van den Oord et al., 2013] noted that it is hard to relate low-level descriptors to easily interpretable ones like genres or tags. The reason for this are properties of widely used metrics for measuring the similarity of tracks via their audio signals [Pohle et al., 2006]. This research question aims to alleviate (i) the issue of finding groups of similar music and (ii) producing high-level descriptions of these groups.

RQ2: Which user groups of non-mainstream music exist? Often, recommender systems are faced with the problem of popularity bias. This notion refers to the probability of a user receiving mostly popular or mainstream music being high. Here, we intend to determine user groups based on the found music styles in RQ1 and subsequently analyze them thoroughly. This clear, but yet exhaustive description of a user's music taste could possibly be used as additional input feature for a recommender system. Thus, recommendations for users of unpopular music taste may improve.

RQ3: How does the music consumption of non-mainstream users deviate from each other? Users usually listen to various kinds of music. Similarly, the set of users is very heterogeneous. Investigating the consumption behaviour could clarify why a user listens to certain types of music. For instance, a uniform distribution of listening events over time would hint that the task of music consumption is not of primary interest to the user. Furthermore, analyzing the properties of listened tracks (e.g., number of distinct genres) could constitute descriptions of users.

RQ4: How do user groups listening to non-mainstream music differ in terms of culture and demography? Recent research emphasizes the large influence of user-specific features on music consumption. In particular, psychological factors impact the way a user listens to music. Furthermore, a user's economic situation, social environment, insecurity, self-confidence, etc. are assumed to play an important role in music consumption. This research question is driven by the strive to describe certain types of users by means of their mind and psychology.

1.2 Scientific Contributions

This thesis constitutes four scientific contributions that are all linked to the aforementioned research questions:

1. We reduce the gap between low-level and high-level descriptions of music by distinguishing different music styles based on acoustic features of tracks and a subsequent explanation via genres.

2. Users of low mainstreaminess are clustered depending on their preferred music style and hence, quantitative and qualitative analyses give insights into their listening behavior, culture, demography and interactions with the long tail of music. Hence, we satisfy the need to shed light on the understudied set of users favoring less popular music from the long tail.

3. Our experiments show that unorthodox users are indeed disadvantaged by state-of-the-art recommender systems. Furthermore, we provide evidence that even if recommendation algorithms would entirely focus on unorthodox users, differences in recommendation quality do still exist for user groups preferring certain non-popular music styles.

4. We furthermore discuss the implications and consequences of our findings for fair music recommendations.

In order to answer the research questions stated in Section 1.1, we conduct several experiments. A coarse overview of the experimental setup employed in this thesis is illustrated in Figure 1.1.

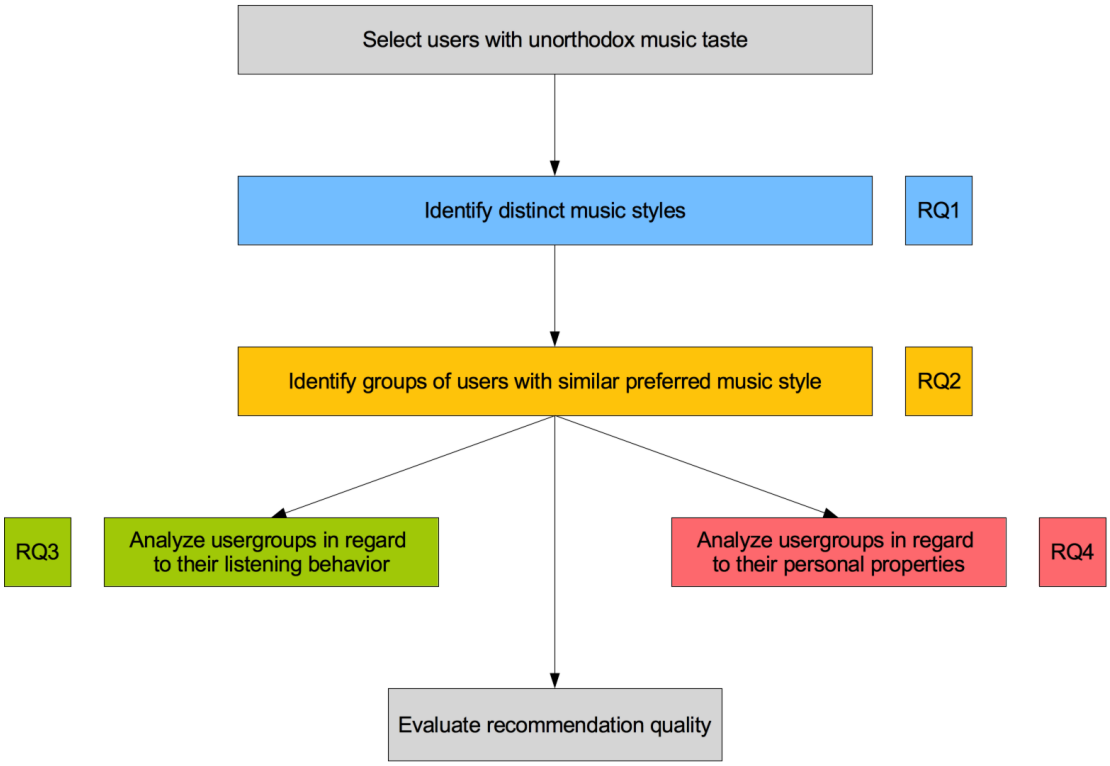


Figure 1.1: Experimental setup for this thesis. First, we select a subset of users that exhibit a low level of mainstreaminess and hence, are users of unorthodox music taste. Then we identify distinct styles of music based on acoustic properties of tracks (RQ1). Subsequently, the aforementioned subset of users is classified into several user groups of different preferred music style (RQ2). In a further step, user groups are investigated in regard to their listening behavior (RQ3). Similarly, we analyze user groups with respect to personal aspects, i.e., demography and culture (RQ4). Eventually, differences in recommendation quality between users of popular and non-popular taste are illustrated, followed by evaluating recommendation quality for user groups of different, non-popular music taste.

1.3 Structure

The course of this work is structured as follows: In Chapter 2, previous work that is related to fair music recommendations is covered. Here, we aim to provide

fundamental knowledge, which is necessary to fully understand this thesis. Furthermore, an overview of state-of-the-art research is given. Chapter 3 deals with the methodology of how we tackle the previously stated research questions. For instance, methods used for identifying different music styles and user groups are outlined. Additionally, we describe the evaluation procedure and explain several metrics. The results of the conducted experiments are given in Chapter 4, among with our interpretations and various statistical analyses. We eventually provide evidence that there is indeed a demand for further research concerned with fairness in recommender systems. In Chapter 5, our findings are summarized alongside with ideas that might be of interest for future work.

Chapter 2

Related Work

The goal of this chapter is to give an understanding of a broad body of related work. Therefore, we offer a quick overview of all key topics relevant for this thesis and give insights into contemporary research. More precisely, we cover the main principle of recommender systems, sophisticated recommender systems that include contextual information, recommendations in the music domain and finally, music recommendation to users that do not conform to mainstream. Additionally, we explain how this thesis deviates from previous research and what it contributes to the problems of today's research in the topic of fair music recommendations.

2.1 Recommender Systems

In general, recommender systems (RS) aim to provide a user with items he may like. These recommendations are the result of the RS exhibiting various informations from different sources. Typically, these informations include, e.g., attributes describing the content, ratings of items and listening counts of music tracks. Due to today's vast amount of information, these systems face the serious problem of what subset of information is best suited for providing the best recommendations. Hence, filtering approaches are used to filter information relevant to a certain user. [Bobadilla et al., 2013] highlight the main three filtering methods widely used in state-of-the-art RS.

Collaborative Filtering. This method is motivated by the fact that a user's opinions are typically influenced by friends and/or acquaintances. Hence, recommendations can be based on information a user's neighbors provide. Users of

similar taste prefer similar items. As an alternative formulation, similar items get consumed by similar users. Say, user U_1 watched movie I_1. User U_2 watched I_1 and I_2. As I_1 got watched by U_1 and U_2, but U_2 also watched I_2, U_1 may also like to watch U_2. These depiction leads to two ways of implementing collaborative filtering (CF): user-based and item-based. The recommendations of user-based CF rely on users similar to each other, whereas recommendations of item-based CF rely on items that are rated similarly by a set of users.

Additionally, two further categories can be distinguished: memory-based and model-based. The first technique exploits previous user to item interactions, e.g., ratings. It relies on information that is already in the past/memory. Therefore it is coined memory-based CF. The second technique, model-based CF, aims to construct models that identify complex patterns based on training data. For instance, it can be observed that users within a certain range of age prefer to listen to a specific kind of music. Hence, a model $M : U \mapsto I$ is learnt that maps a set of users U to a set of items I, based on the age of users.

[Su and Khoshgoftaar, 2009] highlighted several shortcomings of memory-based and model-based CF. Most importantly, memory-based CF depends on users rating - in some way - their preference for a certain item, which means that recommending items with no ratings is impossible. Similarly, new users have not defined their item preferences yet and therefore, similar users cannot be found. This certain problem is coined "cold-start" problem. Model-based CF relies on building a model mapping from users to items. Building such a model is usually very expensive in terms of computational effort. Hence, a trade-off between prediction performance and scalability has to be made.

Content-based Filtering. Assuming we can quantify a user's preferences by means of a so called user-profile. Similarly, assume that also the content of items can be described by so called item-profiles. Content-based Filtering (CBF) recommends items of which the item-profile is similar to the user-profile.

[Lops et al., 2011] provide a high level architecture for CBF. First, item-profiles have to be constructed in order to get a description of items. This can be achieved

by considering, e.g., an item's audio signals [Cano et al., 2005], tags [Cantador et al., 2010] or an already provided quantification. Secondly, a user-profile has to be learned. That means that a user should be described by his item preferences, e.g., any kind of combination of his consumed items' item-profiles. This could be realized by employing a so called Vector Space Model [Salton et al., 1975] and eventually computing a weighted average of item-profiles, which then constitutes the user-profile. Alternatively, a user could also freely provide a custom-made user-profile. This user-profile can then be finetuned iteratively by utilizing methods from the field of Information Retrieval [Rocchio, 1971]. Lastly, the system recommends items to a user that have a large degree of similarity to the user-profile.

Unfortunately, CBF is prone to overspecialization. Due to CBF's methodology, it is clear that the set of recommendations only comprises very similar items, which results in low diversity. Furthermore, it remains a major concern of how to extract meaningful features from unstructured data, in order to build a description of items.

Hybrid Recommender Systems. As shown in the previous paragraphs about CF and CBF, both have major disadvantages. Hence, research proposes the realization of RS that combine CF and CBF, coined hybrid recommender systems (HRS). HRS successfully mitigate certain shortcomings of CF and CBF, by combining user-related and content-related information. For instance, as CF is unable to conduct good recommendations for a new user, CBF can alleviate this issue under some circumstances (e.g, user manually defines preferences).

Clearly, recommendations of CF and CBF have to be combined in some way. [Burke, 2002] provides a survey of HRS and furthermore presents seven aggregation schemes:

1. Weighted: An aggregation is constructed by computing the weighted average of the recommendation scores. For instance, linear combination.

2. Cascade: In this schema, one RS is built on top of another. The recommendations of the first RS are further finetuned by the second one on top. The second RS basically serves as correction mechanism.

3. Switching: Based on some switching criterion, either the first or the second RS has the power to present its recommendations. For instance, a switching criterion could be the inability to construct good recommendations. If a certain score is too low for one RS, the other RS gets its chance to do better.

4. Mixed: Here, recommendations of multiple RS are presented simultaneously. For instance, a user buys item I_1. CF ("Persons who bought I_1 also bought I_2.") and CBF ("You like I_1? You may also like I_2.") recommendations are presented.

5. Feature Combination: Include neighborhood features typically used in CF into CBF as additional features.

6. Feature Augmentation: The recommendations of one RS serve as input for the second one. For instance, CBF finds similar items and in a subsequent step, CF takes this additional features and conducts the actual recommendations.

7. Meta-level: In this approach, a model is learnt. Hence, the model is a concise representation of a user's taste. This representation is the input for a succeeding RS.

The selection of one of these aggregation schemes heavily depends on the application domain. Furthermore, one schema only solves a certain subset of problems in CF and CBF. Thus, one may decide carefully on what problems to focus on.

2.2 Context-Aware Recommender Systems

The consumption behavior of individuals is typically heavily influenced by temporal, social, personal, demographic, situational and economic factors. According to [Schmidt et al., 1999], these factors can be split into human- and physical-induced. Human factors include the user itself (e.g., habits), the social environment (e.g., music preferences of friends) and the task (e.g., exploration of new content). Physical factors include the surrounding conditions (e.g., weather), the infrastructure (e.g., input device) and the location (e.g., music consumption while driving). The configuration of these factors defines the context. In other words, context describes the circumstances under which a user conducts certain

actions. Classic RS are context-agnostic. They lack in incorporating contextual influences of the user, as they consider the problem of providing well suited items to a user as two-dimensional. More formally, classic RS conduct a two-dimensional mapping $Ratings : Users \times Items$. Hence, performance declines for varying contexts, since the quality of recommendations changes depending on the context.

The aforementioned thoughts motivate development and research in the field of context-aware recommender systems (CARS). Contrary to classic RS, the incorporation of a user's context plays a fundamental role within these systems. They consider recommendations not as two-dimensional, but as three-dimensional problem. Formally, they incorporate context by considering the mapping $Ratings : Users \times Items \times Contexts$. In general, the performance improves, since recommendations now consider the distinct context of a user.

Hence, CARS are required to have knowledge of a user's context. [Adomavicius and Tuzhilin, 2011] classified contextual factors further into fully observable, partially observable and unobservable. This is of relevance, since CARS usually focus on leveraging contextual factors of a certain degree of observability. For instance, matrix factorization techniques [Koren et al., 2009] include unobservable factors, as they extract latent factors from the user to item relationships. Furthermore, contextual information may not stay valid over time. As, e.g., occupation likely is a relevant contextual factor for most users, it may be non-relevant after retiring. Thus, [Koren et al., 2009] defined additional properties of contextual factors: static and dynamic.

Another problem is the structure of contextual factors. RS cannot interpret raw values obtained from, e.g., sensors as human beings can do. Therefore, research strives to build abstract concepts upon contextual factors' raw values [Shin et al., 2009]. Interestingly, [Lee et al., 2010] introduced a rigid mathematical framework for context abstraction. They utilized Fuzzy Set Theory [Zadeh, 1965], as they found that several contexts do not exhibit clear boundaries and hence, fuzziness is required. Also, names of music playlists have been utilized to build groups of tracks with similar context (e.g., "summer" playlists) [Pichl et al., 2015].

In practice, it is a crucial question of how to integrate context into a RS. [Haruna et al., 2017] outlined three approaches: prefiltering, postfiltering and contextual. Prefiltering excludes data that is not relevant in the current context before computing recommendations. Recent work utilizes a grouping of users according to similar contexts [Chen et al., 2014]. Postfiltering postpones this selection by choosing items that are relevant in the current context only after recommendations are computed [Ramirez-Garcia and García-Valdez, 2014]. Contextual approaches aim to directly include contextual factors into the model. This can be achieved by utilizing matrix factorization [Baltrunas et al., 2011], tensor factorization [Karatzoglou et al., 2010], or by enhancing matrix factorization with convolutional networks [Kim et al., 2016]. Please note that research regarding directly including contextual information into the model clearly focuses on matrix factorization techniques.

Due to their rich capabilities, CARS are applied in several fields of everyday life. For instance in the learning domain [Verbert et al., 2012], in the tourist industry [Van Setten et al., 2004, Meehan et al., 2013] or regarding mobile applications [Böhmer et al., 2010, Woerndl et al., 2007]. In addition, [Haruna et al., 2017] provides an exhaustive overview of several application areas.

2.3 Music Recommendations

In music recommendation systems (MRS), different abstractions of music can be recommended to a user. For instance, tracks, albums, genres, artists or playlists. Typically, collaborative, content-based or hybrid approaches are used, whereas hybrid systems achieve good results by incorporating both, collaborative and content information into its model. In most cases, features of audio signal are used as content information.

Music Information Retrieval. This highly active research area is concerned with building content-based models that are capable of extracting meaningful information out of music data. [Typke et al., 2005] classified Music Information Retrieval (MIR) systems into two groups. The first group deals with structured music data (i.e., genre annotations). The second group tackles the problem of MIR

in unstructured music data (e.g., audio signals). As [Casey et al., 2008] noted, the so called semantic gap poses a fundamental problem in the field of MIR. This gap illustrates the differences in systems that utilize, e.g., low-level audio signals and systems that utilize high-level, abstract features like genre annotations or emotions of music. Hence, it is very hard to include both, low- and high-level features into a RS, as the connection between these two types is missing. In other words, it is very hard to infer high-level descriptions based on low-level descriptions of music data and vice versa.

Research regarding MRS is driven by the high value of this topic in industry and society. Additionally, MRS are also of high interest for research, since they introduce problems, different from other domains in RS. Challenges like the Recsys challenge 2018 for automatic music playlist continuation [Zamani et al., 2018] aim to foster research in MRS by providing an overview of contemporary state-of-the-art approaches.

[Schedl et al., 2018] lists several particularities of MRS. Music is assumed to be easier to dispose, since the duration is in general much shorter than for, e.g., movies. Music is usually consumed sequentially and successive items follow a clear pattern in terms of style. In other words, the rate of change for a sequence of listened tracks tends to be small. Different types of listeners can be distinguished. For instance, listeners, who prefer to repeatedly listen to a small set of tracks and listeners, who prefer to explore a wide variety of music. The latter depiction illustrates that music consumption is heavily influenced by the psychology and intent of a user. Music typically tries to trigger a certain type of emotion. Similarly, a user's emotions influence the preferred type of music. Interestingly, [Jin et al., 2018] investigated two personal characteristics of users, coined musical sophistication and visual memory capacity. In their work, they observed that high musical sophistication positively correlates with the perceived quality of recommendations. Anyway, the previous aspects pose problems for MRS. Next, consider music that is listened whilst at work and music that is played inside a bar. The type of music that we consume heavily depends on our current situation. Hence, music consumption differs for varying contextual factors.

MRS exploit several types of information. According to [Schedl et al., 2015], high-level descriptors like genres or low-level descriptors like acoustic features have been applied in research. As mentioned before, context, especially contextual information about the listener, can be exploited for music recommendation. For instance [Cheng and Shen, 2014] proposed a contextual method that combines the popularity of items with a user's current location. Other contextual information includes the user's emotional state, the music taste of friends, acquaintances or colleagues. Furthermore, music taste differs significantly for users of varying homecountries.

Today's research strives to incorporate the aforementioned informations into modern MRS. Early approaches to music recommendation included the utilization of statistical models modelling listening sequences of users [Zheleva et al., 2010], or contextual systems like [Park et al., 2006]. Caused by (i) the large amount of available music data and (ii) the large body of factors related to music taste, deep learning methods gained popularity in this domain. [Schedl, 2019] reviewed recent approaches in utilizing Deep Neural Networks (DNNs) for music recommendations. For instance, DNNs have been applied for the matter of learning sequential models [Zheng et al., 2019, Sachdeva et al., 2018] or generating music [Huang and Wu, 2016]. Unfortunately, current research in [Dacrema et al., 2019] conducted experiments in which classic methods like collaborative filtering or matrix factorization outperform models employing Deep Learning. They found that the good performance of DNNs can be mainly attributed to unsuited evaluation. Furthermore, hybrid models that incorporate information about both, user and content have been developed [Lee et al., 2018]. Additionally, [Van den Oord et al., 2013] inferred latent factors from audio signals in order to find the missing linkage between high-level music characteristics and low-level signals.

2.4 The Long Tail of Items in the Music Domain

On various occasions in, e.g., nature, society or economy, it can be observed that the distribution of observations follows a long-tailed distribution. Here, a small subset of, e.g, tracks forms the head and the remaining set of tracks forms the tail, where the head comprises the majority of observations. In the music domain, the head

is often referred to as mainstream resp. popular music. This uneven distribution induces a strong bias towards popular items in music recommendations. Hence, research strives to develop methods, which facilitate less popular items from the long tail. Furthermore, users exist that solely listen to non-mainstream music from the long tail.

Popularity Bias. In general, MRS aim to provide tracks, artists, genres or playlists that are of high interest to a user. For this matter, often ratings of users towards items are utilized. Unfortunately, many RS suffer from popularity bias. This phenomenon describes the observation of popular items being more likely to be recommended than less popular items. As [Celma, 2010] pointed out, only 1% of digital tracks contributed 80% to all sales in 2007. Hence, purchases of digital tracks follow a long tailed distribution. Buyers focus on a small subset of highly popular tracks, whereas they resign buying tracks from the long tail, which corresponds to tracks with only minor popularity.

Reconsidering MRS, popular tracks have been listened by a majority of users. Therefore, rating data is imbalanced, which induces a tendency of RS to recommend popular items. Those items may have been consumed by all users, but most likely they do not reflect a user's personal taste in music. [Abdollahpouri et al., 2019b] conducted an in-depth analysis of movie recommendations for different types of users. They observed that the portion of popular movies in the recommendations is far larger than the portion of popular movies in a niche-user's history. This shows the prevalence of strong popularity bias in the field of RS. To tackle this problem, [Steck, 2018] proposed to calibrate recommendations to comply with the popularity distribution of a user's consumption history. Another interesting approach is outlined in [Abdollahpouri et al., 2019a]. The authors introduce a novel objective function that steers the contribution of the long tail via a weight parameter.

Mainstreaminess in the Music Domain. Recent research in [Schedl and Bauer, 2017] has shown that MRS lack in providing satisfying recommendations for users of unorthodox taste. This problem arises from the model-tuning in contemporary RS. Typically, the evaluation of such models follows the well-known

schema of splitting the dataset into training- and testset. Both sets comprise a random selection of ratings, items and users. Eventually, the recommendation quality is assessed by utilizing accuracy measurements, which quantify the average deviation between recommendations and real information in the testset. Due to the accuracy measurement, the MRS is unaware of a possibly large deviation for a certain subset of users or items. This indicates a strong need for considering fairness in music recommendations. This group of users, for which the MRS does not perform well, are usually those of low mainstreaminess.

[Bauer and Schedl, 2019] introduced an abundance of mainstreaminess metrics. Basically, their metrics rely on the correlation in music consumption between a user and a larger population (i.e., country or global). Research in [Schedl and Hauger, 2015] suggests to conduct recommendations on user groups of different mainstreaminess. They showed that popularity-based recommendation methods perform best on users of high mainstreaminess, but fail for users of low mainstreaminess. Eventually, the authors also indicate that considering the notion of mainstreaminess can indeed improve recommendations.

2.5 Summary

In general, recommender systems aim to find a set of items that are thought to be of interest to a certain user. To filter the abundance of available data, several approaches can be applied. User-based Collaborative Filtering identifies similar users and thus, infers items to be recommended. Equivalently, item-based Collaborative Filtering identifies similar items and hence, provides a user with a set of new items that are similar the the ones he already liked. Content-based Filtering models a user's preferences by a so called user-profile. Similarly, item-profiles are built that describe properties of items. Eventually, items are recommended, for which the item-profile is similar to the user-profile. Hybrid methods utilize both, collaborative- and content-based methods in order to alleviate certain problems and improve the quality of recommendations. Often, e.g., music consumption is heavily influenced by our location, social situation, etc. Context-aware recommender systems consider the contextual configuration of a user. Hence, these systems do not only utilize the relationship between users and

items, but also the relationship towards context. Research in the field of music recommender systems strives to tackle several problems that are not as prevalent in other domains. For instance, music is easy to dispose, has short duration and successively consumed items tend to be similar. Most importantly, the perceived quality of recommendations varies for different types of listeners, as individuals' music consumption is typically influenced by psychology and context. Especially in the music domain, a long-tailed distribution of music can be observed. That means that a small subset of, e.g., tracks is very popular, whereas the majority of tracks is far less popular and hence, lives within the long tail. Thus, a few popular tracks are dominating the consumption behavior of users. Furthermore, there exists a set of users that prefers listening to less popular music from the long tail. music recommender systems face the problem of very bad recommendation quality for this certain type of users.

We identify two works similar to this thesis. [Schedl and Bauer, 2017] split users into groups of different mainstreaminess. They show that contemporary recommender systems focus on users of high mainstreaminess and provide rather poor recommendation quality for user of low mainstreaminess. Contrary to their contributions, we do not only show a disadvantage for users of low mainstreaminess, but illustrate varying recommendation quality for different non-popular music styles. [Kowald et al., 2019] introduced a novel algorithm, which outperforms widely used baselines, also on low-mainstreaminess users. Furthermore, they provide descriptive statistics of this understudied group of users. In this work, we do not aim to find well-performing recommendation methods, but contribute a further in-depth analysis of the music taste of unorthodox users.

Chapter 3

Methodology

This chapter aims to explain our approach to tackle the research questions. In order to answer research question RQ1 (*RQ1: How can non-mainstream music styles be identified and concisely described?*), we apply clustering to find collections of tracks, which represent music of similar style. Those styles are then described by acoustic features, genres and personas. Considering RQ2 (*Which user groups of non-mainstream music exist?*) we introduce a classification schema to gain knowledge about the music preferences of user groups. In order to answer RQ3 (*How does the music consumption of non-mainstream users deviate from each other?*), we employ metrics that quantify music consumption behavior. Furthermore, we examine the relation between user groups and music styles. Eventually, we utilize the rich body of attributes within the used dataset and conduct an explorative analysis of of user groups based on several cultural dimensions and personal information. This answers RQ4 (*How do user groups listening to non-mainstream music differ in terms of culture and demography?*).

3.1 Dataset

Throughout the course of this work we utilize the LFM-1b dataset, which is widely used in the area of music recommendations. Its variant, Cultural LFM-1b, is a strict subset of LFM-1b, but comprises different additional features for both, users and music. We furthermore analyze Non-Mainstream Cultural LFM-1b, which is a dataset representing only users of low mainstreaminess.

The overall goal of this thesis is to provide insights into how to handle music recommendations for user groups that do not align with common behavior. Hence, analyzing user groups via their personal attributes and their personal listening behavior is crucial. For the following properties of LFM-1b, we conclude that it is well suited for the matter of this work:

1. Freely available
2. Large amount of listening events and users
3. Demographic features of users
4. Availability of subsets with additional features
5. Already used in research

3.1.1 LFM-1b

This dataset was introduced by [Schedl, 2016] for the purpose of evaluating music recommender systems. It represents more than 120,000 Last.fm users and their more than one billion listening events. Each listening event is characterized by artist, album, track and timestamp.

Entity	Count
Users	120,322
Artists	3,190,371
Tracks	32,291,134
Listening Events (LEs)	1,088,161,692
Min. LEs per user	4
Q_1 LEs per user	999
Median LEs per user	3,410
Q_3 LEs per user	15,283
Max. LEs per user	654,936
Avg. LEs per user	8,878.762 ($\pm$ 15,962.078)

Table 3.1: Descriptive statistics of the LFM-1b dataset. The value within the parenthesis is the standard deviation.

User-specific demographic features are provided as well. This combination of features serves as playground for evaluating and analyzing personalized music recommender systems. The authors used the 250 top tags to get the corresponding artists and their top fans, which results in 465,000 active users. The detailed listening events are then obtained for a randomly chosen subset of 120,322 users. The data fetched comprises events ranging from January 2013 up to August 2014. Please find a short summary of LFM-1b in Table 3.1.

3.1.2 Cultural LFM-1b

The Cultural LFM-1b dataset [Zangerle et al., 2018] is a subset of LFM-1b and provides further features regarding music and user demographics. Tracks are described with in total eleven acoustic features and users are linked to cultural aspects in addition to their personal attributes. These acoustic features were queried utilizing the Spotify API[1]. Hence, each track is characterized by its Spotify Audio Feature Descriptions[2]. Exhaustive descriptive statistics of Cultural LFM-1b can be found in Table 3.2.

Entity	Count
Users	55,190
Artists	337,840
Tracks	3,471,884
Listening Events (LEs)	351,469,333
Min. LEs per user	1
Q_1 LEs per user	1,242
Median LEs per user	5,028
Q_3 LEs per user	8,750
Max. LEs per user	345,014
Avg. LEs per user	6,373.780 ($\pm 9,118.109$)

Table 3.2: Descriptive statistics of the Cultural LFM-1b dataset. The value within the parenthesis is the standard deviation.

Cultural aspects of users are captured by Hostede's dimensions and by the World Happiness Report 2018. Hofstede studied the cultural aspects of nations

[1] https://developer.spotify.com/web-api

[2] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features

and hence, introduced four cultural dimensions in [Hofstede et al., 2005]. Subsequently, additional two dimensions were introduced in [Hofstede, 2011]. Hofstede's work encountered severe criticism. [Baskerville, 2003] and [Jones, 2007] pointed out that Hofstede assumes the population of a nation to be cultural homogeneous and that cultures can be partitioned into nations. In the following, we explain all dimension of Hofstede:

Power distance is the degree to which individuals reckon power to be unevenly distributed within society. *Individualism* is the extent to which individual interests are of higher importance than group interests. *Masculinity* is the extent to which masculine dimensions like heroism and material reward are of higher importance than feminine dimensions like cooperation and caring for the weak. *Uncertainty avoidance* quantifies how a society prefers to neglect unorthodox beliefs and approaches. *Long-term orientation* measures how individuals of a certain society are future-driven and not tradition-oriented. *Indulgence* is the tendency of allowing gratification of basic human drives like enjoying life and having fun.

The World Happiness Report [Sachs et al., 2018] comprises a formerly called happiness index (now Life Ladder) and other happiness-related attributes, which quantify the social and economic situation of a country's population. In total, it includes life ladder, log gdp per capita, social support, healthy life expectancy at birth, freedom to make life choices, generosity, perceptions of corruption, positive affect, negative affect, confidence in national government, democratic quality, delivery quality, standard deviation of life ladder, standard deviation / mean of life ladder, gini index, gini index (average 2000-2015), gini index of household income. Due to the abundance of features, we chose to select only those, which are of primary interest to us. These dimensions are given in the following:

Social support is the degree to which individuals of a society can rely on their family and/or friends if they need them. *Perceptions of corruption* is the degree to which persons of a country think their government is corrupt. *Log GDP* is the logarithm of the gross domestic product per capita. *Life Ladder* quantifies the subjective happiness. *Life expectancy* is the expected lifespan an individual is at health. *Generosity* quantifies if individuals of a country are willing to spend money for charity. *Freedom* measures the perceived freedom to make life-relevant choices.

Users

In the following paragraphs we depict different interesting properties of the set of users. Starting with their demographic properties (i.e., age, gender, country), we continue analyzing also the corresponding listening events.

As one can observe in Table 3.3, the vast majority of users is male, whereas only a quarter of users is female. A minor portion of users decided to choose their gender as neutral, which nevertheless can be caused by not considering themselves as either female or male, or by not wanting to share this personal information. Hence, this dataset is clearly biased towards men.

Gender	Count	Percentage
Male	36,506	66.15 %
Female	13,937	25.25 %
Neutral	4,663	8.45 %

Table 3.3: Gender distribution of users. This dataset is obviously biased towards men. Furthermore, notice that it is unclear whether neutral indicates neutral gender, or users not providing gender information.

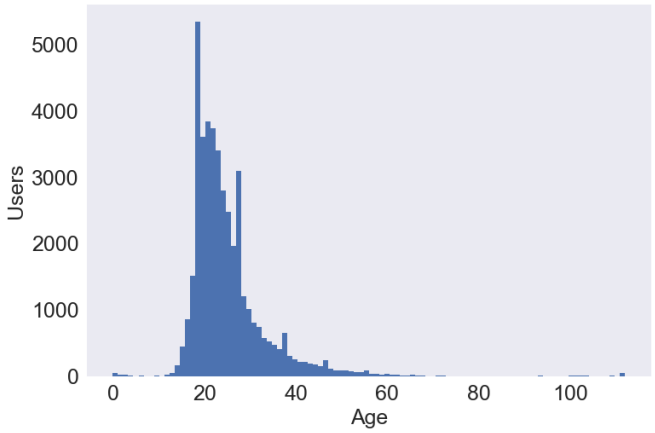


Figure 3.1: Histogram of age distribution. This dataset includes a bias towards users, whose age is in the range 20 to 28.

The illustration in Figure 3.1 shows the age distribution of users. We can see that the distribution is heavily skewed to the right, with a median of 23 years and the first and third quantile located at 20 years resp. 28 years. Hence, 50 % of users have an age between 20 and 28. Please note that 21.99 % did not specify any age. These observations indicate that this dataset is biased towards users with an age in the range of 20 to 28. Furthermore, it is questionable, whether the left- and rightmost values represent users with less than 10 or more than 100 years.

Table 3.4 shows the homecountries of users. Here, only the top ten countries with the highest number of users are listed. The dataset is clearly biased towards the United States, since they represent roughly twice the amount of Russia's users, which is the second most popular country. Interestingly, the top seven countries represent 59.18 % of all users. We identify these as countries, dominantly prevalent in this dataset. Hence, analyses based on the country distribution have to take this into account.

Country	Count	Percentage
United States	10,255	18.58 %
Russia	5,024	9.10 %
Germany	4,578	8.30 %
United Kingdom	4,534	8.20 %
Poland	4,408	8.00 %
Brazil	3,886	7.00 %
Finland	1,409	2.55 %
Netherlands	1,375	2.49 %
Spain	1,243	2.25 %
Sweden	1,231	2.23 %

Table 3.4: Distribution of users' homecountries. Only the top ten countries are listed. Three types of countries can be distinguished: (1) countries with a contribution of $\geq 10\%$, (2) countries with a contribution of $< 10\%$ and $\geq 5\%$ and (3) countries with a contribution of $< 5\%$. This yields a strong community bias towards the United States.

In general, mainstreaminess denotes how well an individual user's behavior complies with the behavior of a population. Non-mainstream subpopulations can pose serious problems even for state-of-the-art recommender systems. Therefore, a lot of research, e.g., [Bauer and Schedl, 2019] has been devoted to find suitable mainstreaminess metrics, where most rely on the correlation between the listening behavior of a reference population and a user within this population. Including such a metric enhances this dataset in a way, such that we have three major advantages:

1. Measure spread of user's music preferences.
2. Assess, how popularity-based recommenders would perform.
3. Split users into groups exhibiting different levels of mainstreaminess.

Users belong to a certain mainstreaminess group in their homecountry. If we changed the homecountry, a user does not necessarily belong to the same mainstreaminess group. Country-agnostic mainstreaminess metrics can mitigate this problem. Hence, we choose to compare a user's behavior to the behavior of the global population. Furthermore, correlation of rank ordering has been shown to work well in the context of music [Pohle et al., 2006, Schedl and Bauer, 2017]. The Cultural LFM-1b dataset conveniently comprises several mainstreaminess measurements. The aforementioned thoughts indicate that $M_{R,APC}^{global}$ is suitable for the matter of measuring mainstreaminess independent of homecountry. [Bauer and Schedl, 2019] defined mainstreaminess of a user u relative to the global population with respect to the listening events per artist as

$$M_{R,APC}^{global}(u) = \tau(ranks(APC), ranks(APC(u))) \tag{3.1}$$

where τ denotes Kendall's τ [Kendall, 1948], APC is the number of events per artist, $APC(u)$ is the events per artists of u and $ranks$ provides a ranking of artists. In this work, the term "mainstreaminess" of users refers to this certain metric.

Figure 3.2 illustrates the distribution of user's mainstreaminess. Firstly, there is only a very limited number of users, whose behavior strongly correlates positively

or negatively with global behavior. Clarifying, only few users either heavily comply with, or dislike global taste. Secondly, the majority of users lies around the mean of 0.171 ($\pm$ 0.099). Hence, most users only partially comply with global taste. Anyway, this hints that the majority of users is indeed influenced by the global popularity of music.

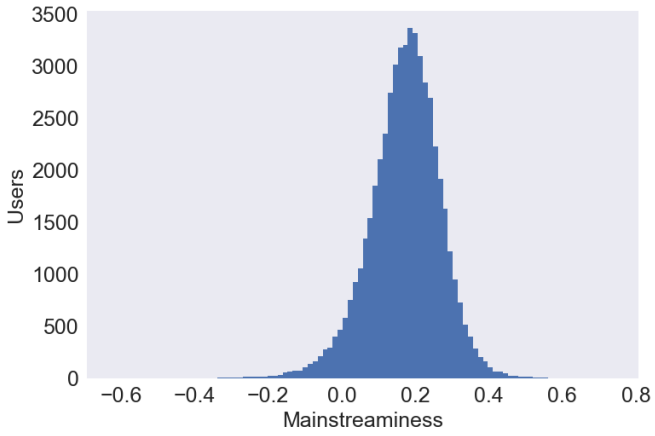


Figure 3.2: Mainstreaminess $M_{R,APC}^{global}$ distribution of users. The distribution is slightly shifted towards 0.2. On average, users' taste is positively correlated with the taste of the overall population.

Hofstede's Dimensions and World Happiness Report

In order to describe the population of users in the Cultural LFM-1b dataset with Hofstede's dimensions, we illustrate the distribution of the aforementioned dimensions over all users in Figure 3.3.

In general, we can see that Hofstede's cultural dimensions exhibit a high degree of variability. Anyway, several interesting observations can be made. For instance, users apparently tend to value their individual needs higher than the needs of others. They also believe that power is distributed among all parts of the population, rather than focused on small entities, e.g., the government. Furthermore, wealth and success seem to be favored against feminine dimensions like caring for others. Users also tend to not face disapproval of luxury and pleasure. Eventually, note that this dataset comprises users of several

homecountries. Hence, the latter observations are likely biased towards prominent countries and thus, deeper analyses have to consider this issue.

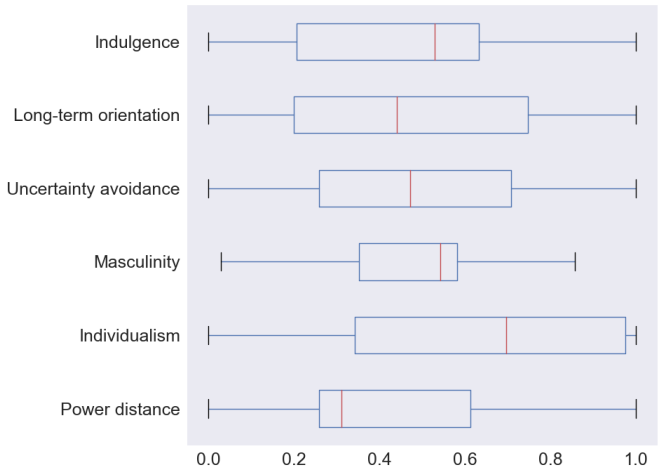


Figure 3.3: Distribution of Hofstede's cultural dimensions over users. Clear cultural tendencies of users in this dataset can be observed. Anyway, this illustration is thought to be heavily biased towards some dominant countries.

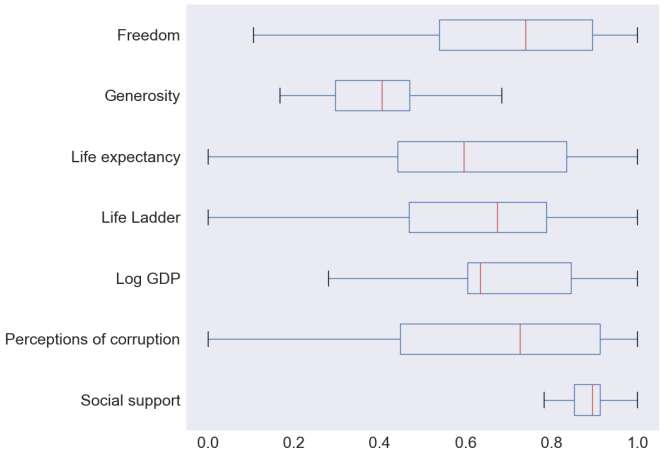


Figure 3.4: Distribution of World Happiness Report's dimensions. Only the most recent data was taken into consideration. This illustration is thought to be heavily biased towards some dominant countries. Most noteworthy, there are strong indications that this dataset represents users of different happiness.

All seven selected dimensions of the World Happiness Report are illustrated in Figure 3.4. We observe that Social Support exhibits rather limited variance. In the case of Log GDP, the distribution indicates that users of small wealth and income may be underrepresented in this dataset. The majority of users apparently has the ability to choose freely how to live their life. Interestingly, Life Ladder and Life expectancy are similarly distributed. This hints a relation between happiness and the expected number of years to live. Furthermore, governments seem to be in general accused of corruption, since most users are unconfident of their government in this respect. Again, note that all observations could be biased towards overrepresented countries and hence, in-depth analyses have to tackle this problem.

Tracks and Artists

Within the Cultural LFM-1b dataset, listening events capture the act of a user consuming music. Music can be organized hierarchically in track and artist. Each hierarchical level exhibits another granularity of music preference. Hence, an explorative analysis of the aforementioned levels can clarify what entities influence the users' music taste.

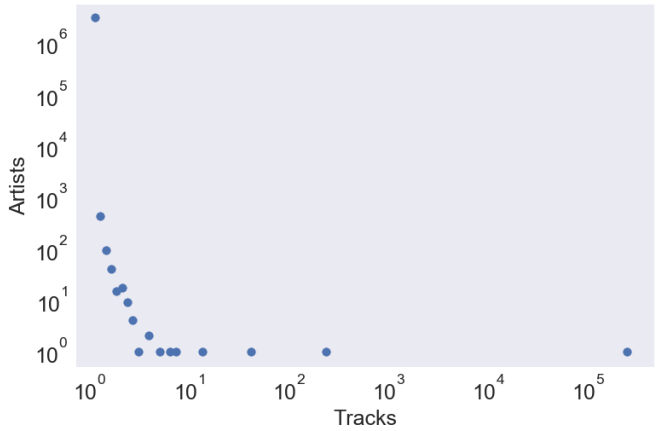


Figure 3.5: Distribution of the number of tracks per artist on a log-log scale. This very uneven distribution indicates that most artists only have few track attributed to them, whereas some artists produced an abundance of tracks.

The distribution of artists with a certain number of tracks attributed to them is shown in Figure 3.5 on a logarithmic scale. We see that the vast majority of artists

produced only a comparatively small number of tracks. Only a few artists exhibit a large amount of tracks. Therefore, high listening count of single tracks does not have to coincide with high listening count for artists and vice versa. Additionally, a user is more likely to listen to artists with a lot of tracks than listening to artists with only a small number of tracks. Interestingly, we observe one outlier with approx. 200,000 of tracks. This observation corresponds to the artist [unknown], thus, it represents tracks for which the artist information is not known.

Similar observations can be made about the distribution of the number of albums per artists in Figure 3.6. One difference lies in the slope of the curve. The relationship between artists and albums behaves much more smoothly than the relationship between artists and tracks. This hints that the effect of some artists dominating the set of albums is not as prevalent as in the artist to track distribution.

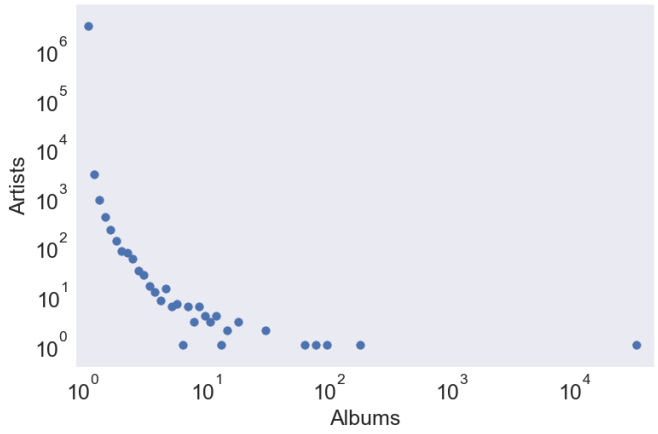


Figure 3.6: Distribution of the number of albums per artist on a log-log scale. It can be observed that most artists only produced a small number of albums, whereas few artists have an abundance of albums attributed to them.

Acoustic Features

In the Cultural LFM-1b dataset, all tracks are linked to their acoustic features. This descriptors are obtained by utilizing the Spotify API and yield acoustic information about the tracks. We utilize a selection of nine acoustic features:

Danceability quantifies how suitable the track is for dancing. *Energy* refers to

the perceptual intensity and activity of the track. *Loudness* is the loudness in decibels. *Speechiness* denotes the fraction of spoken words. *Acousticness* is the confidence whether a track is acoustic. *Instrumentalness* refers to the probability that the track is purely instrumental and contains no vocals. *Liveness* measures the presence of audience in the recording. *Valence* quantifies the musical positiveness conveyed. *Tempo* denotes the tempo of the track in beats per minute.

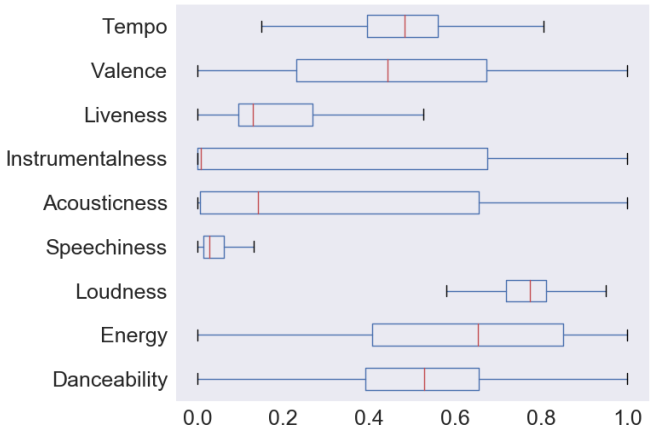


Figure 3.7: Acoustic features of tracks in Cultural LFM-1b. In general, most tracks are of low speechiness and instrumentalness. Anyway, tracks seem to exhibit a large variety of energy.

Furthermore, the acoustic features' distributions are illustrated in Figure 3.7. As one can see, the tracks in this dataset show a very low degree of instrumentalness. Hence, most tracks include vocals. Similarly, most tracks exhibit at least some contribution of instruments, as the number of mostly spoken recordings is very small. Furthermore, live-recordings seem to be in the minority. Even though all levels of energy are present in this dataset, tracks tend to be of rather high energy.

Listening Events

Analyzing listening events sheds light on a user's individual music consumption behavior. It clarifies, which tracks, albums or artists a user listened to. Furthermore, we can get insights into the temporal aspects of music consumption.

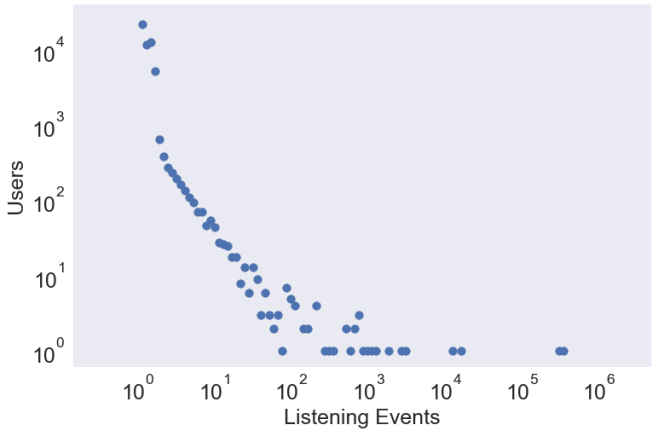
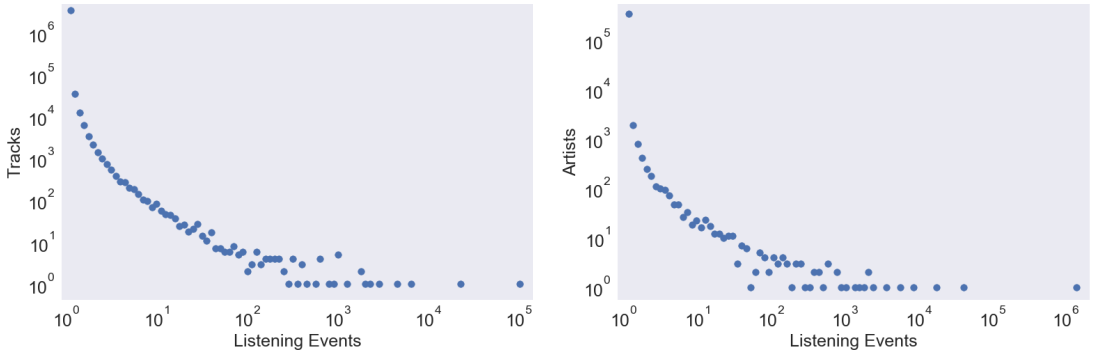


Figure 3.8: Distribution of listening events of users on a log-log scale. It can be observed that most users have less than 10,000 listening events. Anyway, a small subset of users consumes music excessively.

The distribution of the number of users with a certain amount of listening events is depicted in Figure 3.8. Both axes are scaled by the decadic logarithm. This illustration shows that the number of listening events heavily increases, as the number of users decreases. Hence, there is only a small subset of users that have a very large amount of listening events. Most users do not consume music as excessively as the latter group. Furthermore, the number of users with more than 10,000 listening events seems to be rather stable. Anyway, note that most users have less than 10,000 listening events. This also verifies the statistics in Table 3.2. There is a crucial information worth considering: The listening events per user are not normalized over time. As the number of listening events usually increases the longer a user is active, this could deteriorate our findings. Please note that, ignoring the first few observations, this distribution roughly follows a power law.

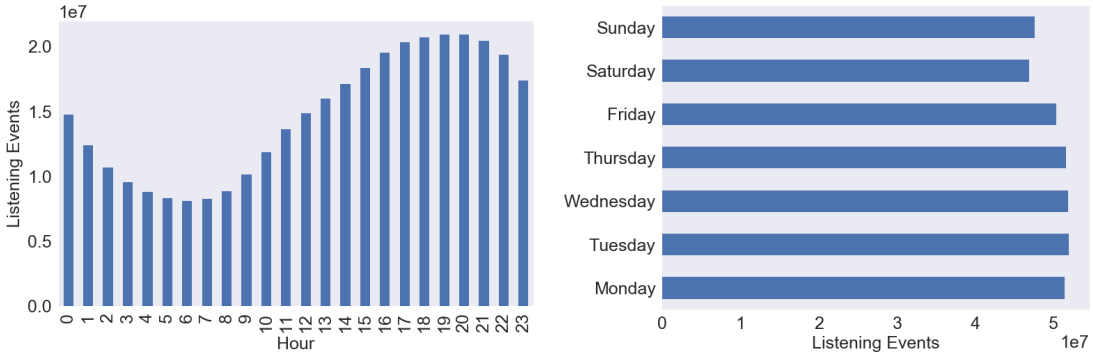
We illustrate the distributions of listening events over tracks and artists in Figure 3.9. As already outlined on several occasions, some tracks dominate the listening behavior of users. The majority of tracks has only a small number of listening events related to them. The latter findings are also valid in the case of the distribution over artists.



(a) (b)

Figure 3.9: Distribution of listening events of (a) tracks and (b) artists. Interestingly, the distributions look very similar. Hence, the presence of some dominant tracks resp. artists is prevalent on both levels.

The temporal distribution of listening events per user is depicted in Figure 3.10. One can observe the mean being in the evening at around eight p.m. Hence we can conclude that the average user tends to focus her music consumption to the evening, whilst maintaining a steady increase throughout the day. In the morning at around six a.m., listening events are at their minimum. Apparently, most users decrease their music consumption on the weekend. Furthermore, we observe an even distribution during the week.



(a) (b)

Figure 3.10: Temporal distribution of listening events. It can be observed that music consumption increases steadily throughout daytime (a). Furthermore, users tend to consume more music during the week than on Saturday and Sunday (b).

3.1.3 Non-Mainstream Cultural LFM-1b

Music preferences and listening behavior is likely to be different for users that do not behave like most of the population. Therefore, we only consider data from this low mainstreaminess group (LowMs) within the Cultural LFM-1b dataset.

Dataset generation

In a first step, we ignore users with less than approx. 5,000 an more than approx. 15,000 listening events (see Figure 3.11), causing the listening events to be quite evenly distributed. In a second step, we partition the users into groups of different mainstreaminess. As illustrated in Figure 3.12, we split them in accordance to the maximal gradient. We reckon this method to be well suited, because it indicates, at which point the area of the bulk starts. This partition results in groups of low (LowMs) and normal (NormMs) mainstreaminess, which have cardinalities of |LowMs| = 2,074 and |NormMs| = 10,740.

$$\text{LowMs} = \{u \in U : M^{global}_{R,APC}(u) \leq 0.097732\} \tag{3.2}$$

$$\text{NormMs} = \{u \in U : 0.097732 < M^{global}_{R,APC}(u)\} \tag{3.3}$$

where U denotes the set containing all users within the listening event threshold.

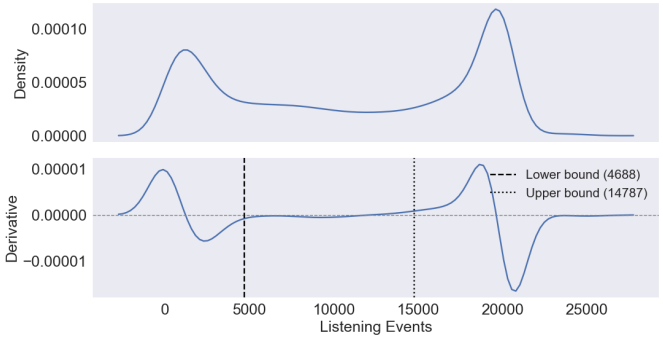


Figure 3.11: Density of listening events of the LFM-1b users, which provide country and mainstreaminess information. The upper and lower bound depict the area, in which the gradient is within ± 1e-6. This results in 12,814 users.

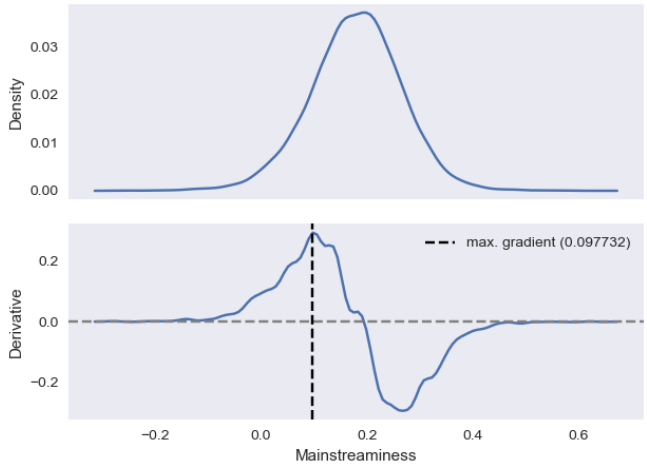


Figure 3.12: $M_{R,APC}^{global}$ distribution of the 12,814 users within the listening event threshold. This split leads to two user groups of different mainstreaminess.

The focus of this work clearly lies on LowMs. We furthermore aim to represent a user of this group by her music preferences. Hence, knowledge about the music signal itself (i.e., acoustic features) and genre information is necessary. The inclusion of a high-level descriptor, i.e., genres seems promising, since they describe tracks in a concise and clear way. In contrast to acoustic features, explaining a user's taste in terms of listened genres is easy to understand and interpret.

For these reasons, we modify the Cultural LFM-1b dataset in a way, such that tracks without genre annotations are excluded. In addition, we also omit tracks, the 2,074 users did not listen to. Genre information was retrieved by utilizing the Spotify API. We observe that a large amount of tracks, especially those listened by LowMs, is not annotated with genres. Hence, the number of tracks within this dataset decreases tremendously compared to LFM-1b and Cultural LFM-1b, as can be seen in Table 3.5.

Item	Value
Users	2,074
Tracks	147,156
Artists	14,316
Spotify Track Genres	1,191
Listening Events (LEs)	4,725,664
Min. LEs per user	1
Q_1 LEs per user	1,210
Median LEs per user	1,959
Q_3 LEs per user	3,103
Max. LEs per user	10,536
Avg. LEs per user	2,279.626 ($\pm$ 1,471.687)

Table 3.5: Descriptive statistics of LowMs (contains only tracks with genres). The value within the parenthesis is the standard deviation.

Mitigating dominant countries and genres

Heavily uneven distributions can cause a bias within the observations. This could mislead our analysis. Table 3.4 gives clear evidence that a few countries occur much more frequently than others. Hence, statistics and analyses will be biased towards them. Therefore, we assign a score to every country, such that frequent countries have a low score and rarely occurring countries have a high score. This behavior can be achieved by the Inverse Document Frequency (IDF) [Jones, 1972] measurement from the field of Information Retrieval. For users' homecountries, we have

$$IDF(c) = \log_{10}\left(\frac{\mid U \mid}{\mid U_c \mid}\right) \tag{3.4}$$

where U is the set of users and U_c is the set of users with homecountry c.

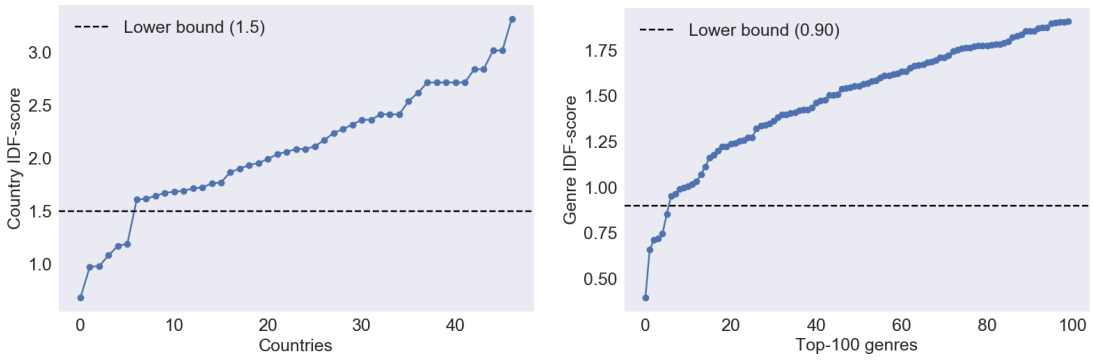
Similar to the distribution of users' homecountries, there exists also a bias in the distribution of genres, users from LowMs listen to. We found that the majority of tracks is annotated with genres like, e.g., rock and pop. At a later point in this work, we explain different collections of tracks via their genres. Since dominating

genres would mislead this explanation, we again take advantage of an IDF-scoring. For genres, we have

$$IDF(g) = \log_{10}\left(\frac{\mid T \mid}{\mid T_g \mid}\right) \tag{3.5}$$

where T is the set of tracks and T_g is the set of tracks annotated with genre g.

The results of this scoring can be seen in Figure 3.13. We ordered countries according to their IDF-scores. In the case of the country distribution, the slope is roughly constant for all countries above the threshold. As there is a nonconstant increase between countries below and above the threshold, the six countries with an IDF-score smaller than 1.5 can be considered as the reason for the bias. Therefore, country-related analyses should handle those countries with care. Similarly, we ordered genres according to their IDF-score. With the same arguments as in the country case, we conclude that the six genres below the threshold dominate. Hence, analyses of genres should take care of a bias induced by those genres.



(a) (b)

Figure 3.13: IDF-distribution of (a) countries and (b) genres. Countries below the threshold are identified as dominant countries. In ascending order, those are: US, RU, DE, UK, BR and PL. Similarly, genres below the threshold are identified as dominant genres. In ascending order, those are: rock, pop, electronic, metal, alternativerock and indierock.

3.2 Technical Details

In order to have the possibility to conveniently handle and share this large collection of data, we built a database utilizing MariaDB[3]. It only represents the Cultural LFM-1b dataset and not LFM-1b, since the acoustic features and cultural aspects are necessities for this work. We also did not select a certain group of users (e.g., LowMs) for the simple reason of having the option to compare different users of unequal mainstreaminess. The resulting database consists of in total nine tables and has a size of approx. 20 Gb. An exhaustive summary and explanation of all entities captured in the database is provided in Table 3.7. Please find further descriptive statistics and relations in Table 3.6.

Table	Attributes	Entries
acoustic_features	track_id, ...	3,446,151
albums	album_id, artist_id, ...	15,418,379
artists	artist_id, ...	3,148,535
events	user_id, artist_id, album_id, track_id, ...	349,797,888
hofstede	country, ...	47
tracks	track_id, artist_id, ...	29,840,402
user_mainstreaminess	user_id, country, ...	53,111
users	user_id, country, ...	55,176
world_happiness	country, ...	1,562

Table 3.6: Descriptive database statistics and relations. Same color indicates linkage between two attributes. Furthermore, we provide the number of a table's entries.

[3]https://mariadb.com

Entity	Attribute	Description
User	user_id	unique user identifier
	country	homecountry
	age	age in years
	gender	gender of user (M, F or N)
	playcount	number of listening events
	registered_timestamp	unix timestamp of register time
	mainstreaminess	correlation with population's taste
Artist	artist_id	unique artist identifier
	artist_name	name of artist
Album	album_id	unique album identifier
	album_name	name of album
	artist_id	unique artist identifier
Track	track_id	unique track identifier
	track_name	name of track
	artist_id	unique artist identifier
	acoustic_features	all music descriptors
Listening Event	user_id	unique user identifier
	artist_id	unique artist identifier
	album_id	unique album identifier
	track_id	unique track identifier
	timestamp	unix timestamp of event

Table 3.7: Description of entities and their attributes within our database. Please note that this database represents a structured view on the available data in the LFM-1b dataset.

3.3 Identification of User Groups

Our goal is to classify individual users into different user groups. These user groups should deviate from each other in terms of music taste and consumption behavior. As will be shown in this section, we achieve this by utilizing acoustic descriptors

in order to cluster tracks. These clusters represent collections of similar tracks. By investigating the interactions of users with these clusters, we obtain a description of users' listening behavior and taste. Subsequently, each user is assigned to a track cluster. All user groups describe different personas, which all prefer different music styles. Summarizing, we perform four steps:

1. Reduce dimensionality of tracks' acoustic features, such that we can ease the task of clustering and conduct a visual, qualitative analysis.

2. Cluster tracks, for the sake of finding clusters representing tracks of similar music and such that we can clearly distinguish clusters via their included tracks.

3. Assign each user to exactly one music cluster.

4. Study resulting user groups with respect to their listening behavior, demographics and cultural aspects.

3.3.1 Track Clustering

We found that loudness does neither improve nor deteriorate results. Therefore, we exclude loudness from our calculations. As already noted above, clusters should be explained by the genres assigned to the contained tracks. To achieve this, only genre annotated tracks are used, which LowMs listened to.

First and foremost, we transform the data into a latent representation. With this, in our case, two dimensional space, we have the advantage of less computation time. Furthermore, it gives opportunity to qualitatively inspect clustering results, as two-dimensional data is usually easy to visualize.

We empirically investigated a large body of dimensionality reduction techniques. In the following paragraph, we outline reasons, why certain techniques might not work for this type of data. Please note that these arguments only represents our subjective assessment. *Principal Component Analysis (PCA)* [Tipping and Bishop, 1999] and its nonlinear variants did not work, since they rely on the covariance matrix of all observations. Hence, only global and no

local informations are considered. Furthermore, this method is prone to noise. *Locally Linear Embedding (LLE)* [Roweis and Saul, 2000] is a manifold learning method, which relies on the assumption that the area around a reference point is linear. In our case, we expect the manifold to be nonlinear even only in small areas. *Multidimensional Scaling (MDS)* [Kruskal, 1964] also assumes distances between points to be linear. Contrary to the latter aforementioned methods *Isomap (IMAP)* [Tenenbaum et al., 2000] does not utilize a linear distance metric, but the geodesic distance. If datapoints are considered as graph, the geodesic distance is defined as the sum of edge weights in the shortest path between two points. Unfortunately, these weights are computed by the euclidean distance, hence, the similarity or distance between two points still is heavily influenced by a linear measurement. *Spectral Embedding (SE)* [Ng et al., 2002] can utilize both, a gaussian kernel or the nearest neighbors as similarity measurement. Firstly, a gaussian kernel determines the distance of points by the Mahalanobis distance. Secondly, the distance to the k nearest neighbors is computed by the Minkowski distance. Both distance metrics are linear. *t-SNE* [Maaten and Hinton, 2008] again models similarity in the high-dimensional space by means of a gaussian kernel. Summarizing, we assume that the latter methods gave unsatifsying results because of two reasons. One, *PCA* and its variants only consider the global structure of data, not aspects of finer granularity. Two, *LLE*, *MDS*, *IMAP*, *SE* and *t-SNE* rely on the assumption that the local neighborhood of a reference point is linear, as they employ linear distance metrics. Since we expect datapoints to lie on a manifold, methods modelling global structure cannot be used. We hypothesize that this manifold exhibits complex and nonlinear structures even in local areas, thus, methods utilizing linear similarity measurements are not advantageous.

We found that *Uniform Manifold Approximation and Projection (UMAP)* [McInnes et al., 2018] works best. This novel method is very similar to *t-SNE*, but (i) is motivated by Riemannian Geometry and (ii) employs a nonlinear high dimensional metric space, i.e., Fuzzy Simplical Sets [Spivak, 2009b], which is a modification of Simplical Sets [Spivak, 2009a]. Another reason is that UMAP is rather insensitive to noise, as it clamps noisy observations into very distant but dense regions. Furthermore, it has already been widely used for music datasets in [Zangerle and Pichl, 2018], [Baig et al., 2018], [Moore et al., 2012] and

[Levy and Sandler, 2008]. Our experiments entirely use the implementation of McInnes[4].

Parameter	Explanation	Value
n_neighbors	number of neighbors used to learn manifold	15
min_dist	defines how tightly points are packed together	0.1
n_components	dimensionality of latent space	2
metric	distance metric	euclidean

Table 3.8: Parameters for UMAP. The small number of neighbors enables this method to learn finer structures, which is necessary for this type of data. Furthermore, we also insisted on packing datapoints tightly together to help the succeeding clustering algorithm.

In addition, we expect only minor differences of the acoustic features between tracks. Hence, the number of neighbors is chosen rather small. Therefore, UMAP learns the manifold structure based on only a small set of neighbors. This results in a preservation of fine structures. For the sake of clustering, it is an advantage to have similar datapoints clumped together. This leads to our decision, to allow UMAP to arrange datapoints in such a way, by defining the corresponding parameter appropriately. Please find an exhaustive list of the applied parameters in Table 3.8.

UMAP fails to preserve densities, but succeeds in preserving proximities. Therefore, proximal datapoints in the original space would have a similar distance to each other in the latent space. That leads to the clusters not exhibiting the same density. Hence, we conduct a broad empirical analysis of clustering methods.

Widely used density based clustering methods, e.g., *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* [Ester et al., 1996] are not capable of dealing with multiple densities. *K-Means* [Bishop, 2006] also fails caused by varying densities. Furthermore, its assumption that clusters are centered around the mean is not true for this dataset. Hence, results are unsatisfying. Since

[4]https://umap-learn.readthedocs.io

Gaussian Mixture Models (GMM) [Reynolds, 2015] are the generalization of K-Means and thus, assume clusters to be disk-shaped, a GMM is not an approach to favor. In *Affinity Propagation (AP)* [Frey and Dueck, 2007], so called exemplars represent datapoints within a cluster. The exemplars are selected via a simple message-passing system. Anyway, AP requires large computational resources and tends to find disk-shaped clusters. *Spectral Clustering (SC)* [Shi and Malik, 2000] inherits some disadvantages of K-Means, since it includes the subsequent calculation of K-Means. We achieve good clustering results for a sample of the data when using *Hierarchical Agglomerative Clustering (HAC)* [Murtagh and Legendre, 2014]. Since this approach is heavily memory-consuming, it is infeasible to apply on the complete dataset. Furthermore, this bottom-up method merges observations, which minimize variance. Variance is not suited for measuring the goodness of clusters, if we aim to consider varying inter-point distance. *Ordering Points To Identify the Clustering Structure (OPTICS)* [Ankerst et al., 1999] is a successor of DBSCAN, aiming to tackle the issue of clusters with varying densities. To some extent, it can be seen as hierarchic version of DBCSAN. Anyway, OPTICS requires sensitive parameters to be selected. Furthermore, the selection of clusters based on the concept of reachability seems not advantageous in our problem setting. In summary, we hypothesize that hierarchic clustering methods that can handle varying densities could lead to the required quality of results.

A further improvement of OPTICS is *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)* [McInnes et al., 2017], which does not require as precise parameter tuning. This is achieved, by building a cluster hierarchy over different parameter values. Hence, it selects clusters that are stable, i.e., remain present for different parameter configurations. HDBSCAN also provides a parameter to define the conservativeness of a clustering. High conservativeness leads to tightly packed clusterings, while some datapoints between clusters may not be classified. In our case, this is a clear advantage, since clusters should yield nicely separated and nonoverlapping collections of tracks with different musical properties. Conventiently, we utilize an implementation of McInnes[5]. The parameters in Table 3.9 are obtained by evaluating the sum of squared errors within clusters via the elbow method [Thorndike, 1953].

[5]https://hdbscan.readthedocs.io

Parameter	Explanation	Value
min_cluster_size	minimal size of a resulting cluster	1,375
min_samples	defines how conservative clustering is	1,375

Table 3.9: Parameters for HDBSCAN. To focus on clearly prevalent clumps of datapoints, we require that clusters have to have a certain size. For the matter of retrieving clusters of tracks with nonoverlapping acoustic features, we employ a rather conservative clustering.

3.3.2 Classification of Users

In order to find user groups of different listening behavior and music taste, we exploit the results of the track clustering. In Section 3.3.1 we elaborated how clusters of tracks can be obtained. Every cluster represents a collection of tracks of different music style. Subsequently, users can be assigned to a single track cluster by utilizing a certain classification schema. For every user we measure how each track cluster influences the user's music taste. This can be realized by introducing weights.

$$w(u) = (w_1, w_2, \ldots, w_N) \tag{3.6}$$

where N is the number of track clusters and w_i is the weight of user u to track cluster i, where w_i is the normalized amount of listening events of u towards tracks within i.

Based on these weights, a user classification could be achieved by assigning a user to the highest weighted track cluster. Unfortunately, this classification would lead to the majority of users getting assigned to the cluster containing most tracks. Clearly, the more tracks inside a cluster, the higher the chance of an arbitrarily chosen user to have listened to tracks within this cluster. In most cases, this simple classification would not represent users' music preferences well. Hence, taking both, weight and cluster size into account, is advantageous. Similar to 3.1.3, we propose a more sophisticated method that employs IDF-scoring to reduce the influence of

clusters containing an extraordinary amount of tracks.

$$IDF(C_i) = \log_{10}\left(\frac{\mid C_1 \mid + \mid C_2 \mid + \cdots + \mid C_N \mid}{\mid C_i \mid}\right) \tag{3.7}$$

where C_i is the set of tracks in track cluster i and the number of clusters is N. The weights of a user can be modified to take also the cluster size into account by applying

$$\tilde{w}(u) = (w_1 \cdot IDF(C_1), w_2 \cdot IDF(C_2), \ldots, w_N \cdot IDF(C_N)) \tag{3.8}$$

Utilizing the new weights $\tilde{w}$, we can execute our original approach to assign each user to the highest weighted cluster, without the deterioration induced by heavily varying cluster sizes. Summarizing, u is classified to be part of user group U_{C_i}, if $\tilde{w}_i(u)$ is the maximum of the user to track cluster weights. Hence a user group is defined as

$$U_{C_i} = \{u \in U : \underset{1 \leq i \leq N}{\arg\max}\, \tilde{w}_i(u)\} \tag{3.9}$$

3.4 Recommendations

Previous scientific work like [Schedl and Hauger, 2015] shows that user groups of different mainstreaminess are faced with a varying accuracy of recommendations. Thus, we aim to (i) outline this problem and (ii) provide indications that also the preference of certain non-popular music styles causes a decline in recommendation quality. For this matter we employ several recommendation methods and measure the performance via a body of evaluation metrics. Please note that we utilize the Python-based and open-source Surprise library [Hug, 2017] for algorithms NORM, BASE, KNN and partially PL. Furthermore, TOP was motivated by recent work in [Kowald et al., 2019].

3.4.1 Algorithms

In this section we provide descriptions of multiple recommendation algorithms used in this work. Here, we focus on very basic methods, since we strive to show the aforementioned differences in recommendation quality between users of low and

normal mainstreaminess. Hence, the usage of more sophisticated algorithms would not be advantageous. As simplification, we define the set of listening events to be

$$LE = \{(u,t) : \text{user } u \text{ listened to track } t, \text{ where } u \in U, t \in T\} \tag{3.10}$$

where U is the set of users, T is the set of tracks and $u = (r_{u,t_1}, ...r_{u,t_{|T|}})$. Hence, the set of listening events between user u and track t is given by

$$LE_t(u) = \{(u,t) \in LE\} \tag{3.11}$$

where the rating of user u towards track t is $|LE_t(u)|$, which is the number of times u listened to t.

We observed in Figure 3.9 that the distribution of listening events per track is very uneven and skewed. Hence, we expect the distribution of ratings per user to be skewed. This deteriorates recommendations, since rating-scales of users exhibit major differences. Therefore it would be not possible to produce meaningful extrapolations of a user's rating based on other users or based on the average rating of all users. Motivated by [Schedl and Bauer, 2017], we scale all ratings of a user to a range of $[0, 1000]$ via

$$r_{u,t} = 1000 \cdot \frac{r_{u,t} - min(r_u)}{max(r_u) - min(r_u)} \tag{3.12}$$

Hence, for any user, the most listened track has rating 1000, whereas the least listened track has rating 0. Anyway, the large range of 1000 is chosen such that differences between listening events are still present.

Mainstream-Aware Baseline (TOP)

This method recommends k tracks with the most listening events over all users. It serves as crude baseline, based solely on the popularity of tracks.

$$\tilde{T}_k(u) = \arg\max_{t \in T}^{k} |LE_t| \tag{3.13}$$

Please be aware, that the ratings are now likely not within the range between 0 and 1000. Hence, we again apply the scaling as post-processing step.

Random Predictor (NORM)

This algorithm is a simple baseline, which models ratings via a normal distribution. It estimates mean and variance of ratings within the training set by utilizing Maximum Likelihood Estimation (MLE) [Fisher, 1922].

$$\mu_{ML} = \frac{1}{|R_{train}|} \sum_{r_{u,t} \in R_{train}} r_{u,t} \tag{3.14}$$

$$\sigma^2_{ML} = \frac{1}{|R_{train}|} \sum_{r_{u,t} \in R_{train}} (r_{u,t} - \mu_{ML})^2 \tag{3.15}$$

$$\tilde{r}_{u,t} \sim \mathcal{N}(R_{train} | \mu_{ML}, \sigma^2_{ML}) \tag{3.16}$$

Due to the normality assumption, a rating of user u to track t is a sample from a normal distribution with the maximum likelihood estimates as parameters.

Power Law Predictor (PL)

The illustration in Figure 3.8 indicates that the number of listening events behaves like a power law. Since rating $r_{u,t}$ is defined as the number of times user u listened to track t, ratings do not follow a normal distribution as assumed in the NORM baseline, but follow a power law distribution. Hence, we design a random prediction baseline, which samples ratings from a power law, fitted to the listening counts of all users in LowMs and NormMs. In general, power law distributions are given by

$$\mathcal{L}(x_1, \ldots, x_n | c, \alpha) = \prod_{i=1}^{n} c \cdot \frac{1}{x_i^{\alpha}} \tag{3.17}$$

First and foremost, scaling c can be omitted, since we generate a probability density function and thus, scaling would not give any differing results. Anyway it could happen that the fist few observations cannot be explained by a power law distribution. Hence, an additional parameter x_{min} is introduced, which serves as lower bound to the power law behavior (i.e., only points $x > x_{min}$ are considered).

The latter modifications can be denoted as

$$\mathcal{L}(x_1, \ldots, x_n|\alpha) \propto \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}} \left(\frac{x_i}{x_{min}} \right)^{-\alpha} \tag{3.18}$$

As precisely described in [Clauset et al., 2009], α is found by utilizing Maximum Likelihood Estimation. The optimal x^*_{min} is obtained via selecting the model with the highest Bayesian Information Criterion (BIC) [Schwarz et al., 1978]. Fitting a power law distribution to observations x_i was conducted by utilizing the rich library in [Alstott et al., 2014]. Eventually, this baseline is given by

$$\mathcal{L}(x_1, \ldots, x_n|\alpha) \propto \prod_{i=1}^{n} \frac{\alpha - 1}{x^*_{min}} \left(\frac{x_i}{x^*_{min}} \right)^{-\alpha} \tag{3.19}$$

$$\alpha_{ML} = 1 + n \left(\sum_{i=1}^{n} \frac{x_i}{x_{min}} \right)^{-1} \tag{3.20}$$

$$\tilde{r}_{u,t} \sim \mathcal{L}(x_1, \ldots, x_n|\alpha_{ML}) \tag{3.21}$$

for $x_i \in R_{train}$.

Baseline Estimation (BASE)

Baseline Estimation exploits the fact that users and items have an inherent bias. Furthermore, there do exist tracks that exhibit non-normal distributed ratings. The overall mean of ratings μ provides a general notion of how ratings are globally distributed. Furthermore, b_u models the the user bias (i.e., the tendency to give higher or lower ratings). The user bias can also be understood as deviation of a user's average rating to the global expectation. The deviation of a track's average rating to the global expectation is denoted as the track bias b_t.

$$\mu = \frac{1}{|R_{train}|} \sum_{r_{u,t} \in R_{train}} r_{u,t} \tag{3.22}$$

$$\tilde{r}_{u,t} = \mu + b_u + b_t \tag{3.23}$$

Please note that b_u and b_t are found via minimizing a Least Squares Optimization Problem with Alternating Least Squares (ALS).

Parameter	Explanation	Value
method	optimization algorithm	als
reg_i	item regularization	10
reg_u	user regularization	15
n_epochs	the number of iterations for the method	10

Table 3.10: Parameters for BASE.

k-nearest Neighbors (KNN)

User-based Collaborative Filtering (CF) [Shardanand and Maes, 1995] employs the idea of similar users having similar tastes. Cosine similarity is widely used in contemporary research in order to quantify similarity between two vectors

$$sim(v_1, v_2) = \frac{v_1^T v_2}{||v_1||_2 \cdot ||v_2||_2} \tag{3.24}$$

where $v_1, v_2 \in \mathbb{R}^n$ and $|| \cdot ||_2$ is the L2-norm.

In this work, the rating is defined as the number of times a user listened to a certain track. Due to the dot product, cosine similarity would be forced towards one despite our scaling procedure. Hence, we chose to define the similarity of two users as the Pearson Similarity [Freedman et al., 2007], which is, in this case, the cosine similarity of the users' u, v mean-centered ratings.

$$sim(u, v) = \frac{\sum_{t \in T}(r_{u,t} - \bar{r}_u)(r_{v,t} - \bar{r}_v)}{\sqrt{\sum_{t \in T}(r_{u,t} - \bar{r}_u)^2}\sqrt{\sum_{t \in T}(r_{v,t} - \bar{r}_v)^2}} \tag{3.25}$$

CF aims to find a set of k nearest neighbors for a certain user u. These nearest neighbors are the k users, which exhibit the highest degree of similarity towards u.

$$N_k(u) = \arg\max_{v \in U}^{k} sim(u, v) \tag{3.26}$$

As the computation of the similarity requires a set of common items between two users, we define that the similarity is zero, if there is no common item. By empirical investigations, we set the number of neighbors to 40. Please find an explanation of

parameters for this algorithm in Table 3.11.

Parameter	Explanation	Value
k	number of neighbors	40
min_k	minimal number of neighbors to compute similarity	1
metric	similarity metric	pearson

Table 3.11: Parameters for KNN.

Each user is represented by a vector, which encodes her music taste. In our case, the user-profiles are vectors containing the number of times a user listened to a distinct track. Hence, the top-k recommendations can be found by utilizing CF as follows:

$$\tilde{r}_{u,t} = \frac{1}{|N_k(u)|} \sum_{v \in N_k(u)} sim(u,v) \cdot r_{v,t} \tag{3.27}$$

The popular k nearest neighbor algorithm realizes Collaborative Filtering and improves the previously mentioned crude algorithm by introducing a normalization over the similarities.

$$\tilde{r}_{u,t} = \frac{\sum_{v \in N_k(u)} sim(u,v) \cdot r_{v,t}}{\sum_{v \in N_k(u)} sim(u,v)} \tag{3.28}$$

3.4.2 Evaluation of Recommendations

In order to answer the research questions stated in Section 1.1, exhaustive evaluation of our results is necessary. Therefore, we utilize several metrics, widely used in the field of recommender systems and information retrieval. The term "relevant item" refers to an item that is of interest to a user. Often, this term is used to denote the groundtruth in the context of developing and evaluating recommender systems. The selection of evaluation metrics is partially inspired by [Gunawardana and Shani, 2015]. For the reason of notational convenience, we define a binary function

$$\mathbb{1}_{u,i}(\mathcal{X}) = \begin{cases} 1, & \text{if } i\text{-th item in } \mathcal{X} \text{ is relevant for user } u \\ 0, & \text{otherwise} \end{cases} \tag{3.29}$$

The outlined evaluation methods in this section are designed to measure either the quality of recommended items and their ordering or the error of predicted ratings. The first class considers the top-k recommendations, whereas the second class measures the error of predicted ratings.

1. Top k recommendations: Precision, Recall, F1-Score, Mean Reciprocal Rank, Mean Average Precision, Normalized Distributed Cumulative Gain.

2. Rating prediction: Root Mean Squared Error, Mean Absolute Error, Fraction of Concordant Pairs.

Please note that predicted ratings can be easily transformed into top-k recommendations. This is achieved by selecting and recommending only the k highest rated items per user.

$$\tilde{T}_k(u) = \arg\max_{t \in T}^{k} \tilde{r}_{u,t} \tag{3.30}$$

where the predicted rating $\tilde{r}_{u,t}$ is computed by any rating prediction algorithm. Hence, the performance of rating predicting algorithms can be measured by means of top-k recommendation evaluation strategies.

Some models used in this evaluation require several parameters to be defined. One may be tempted to conduct model selection prior to evaluation. But for the matter of this works this is not necessary. We aim to show that there is a lack of recommendation quality for users of low mainstreaminess. Hence, any reasonable well working model enables us to compare recommendations between users of low mainstreaminess and other users. Thus, we resign doing a exhaustive model selection and continue with empirically selecting well-suited parameters for the models.

In particular, our experiments are aimed to show that differences in recommendation quality exist for user groups of varying mainstreaminess and preferred music style. Thus, we perform following steps:

1. Performing 5-fold cross-validation on data from LowMs and a random sample of NormMs, where the number of samples is equal to the cardinality of LowMs. This step is equivalent to five runs with distinct 80%/20% splits.

2. Report averaged error metrics on entire dataset.

3. Report averaged metrics on NormMS.

4. Report averaged metrics on LowMs.

Thus, we comply with the widely used practice of training a model on 80% and testing on 20% of ratings. Please note that the sampling of users from NormMs is motivated by having equally-sized user groups. Anyway, we verified that the mainstreaminess distribution of samples and NormMs are equivalent. After conducting the 5-fold cross-validation procedure, we average the error measurements and compute the standard deviation. Please note that this split is conducted randomly. Hence, the number of ratings in the training- resp. test set may vary per user.

Following notation is used in this section: $T_k(u)$ is the set of the k relevant tracks of user u, $\tilde{T}_k(u)$ is the set of k recommended tracks, U is the set of users, $R : Users \times Tracks$ is the real rating matrix, where $r_{u,t}$ is the rating of u for track t. Similarly, $\tilde{R} : Users \times Tracks$ is the predicted rating matrix, where $\tilde{r}_{u,t}$ is the predicted rating of user u for track t. As mentioned before, we define the rating $r_{u,t}$ to be the number of times, u listened to track t, scaled to the range of $[0, 1000]$. In the following, we explain all metrics based on tracks. Anyway, note that instead of tracks, any kind of items could be used.

Precision (P)

Precision is the fraction of recommended tracks that are relevant divided by the number of recommended tracks. It defines, how precise recommendations are. Precision for a single user, considering only the top-k recommended items, can be defined as

$$P@k(u) = \frac{|T_k(u) \cap \tilde{T}_k(u)|}{k} \tag{3.31}$$

We can aggregate the user-wise precision for a set of users with

$$P@k = \frac{1}{|U|} \sum_{u \in U} P@k(u) \tag{3.32}$$

Recall (R)

Recall is the fraction of recommended tracks that are relevant divided by all relevant tracks. It defines, what percentage of relevant tracks is recommended. Recall for a single user, considering only the top-k recommended tracks, can be defined as

$$R@k(u) = \frac{|T_k(u) \cap \tilde{T}_k(u)|}{|T_k(u)|} \tag{3.33}$$

Similar to precision, we can aggregate the user-wise recall for a set of users with

$$R@k = \frac{1}{|U|} \sum_{u \in U} R@k(u) \tag{3.34}$$

F1-Score (F1)

Usually, both, high precision and high recall are considered to be worthwhile. Unfortunately, increasing precision eventually leads to decreasing recall and vice versa. F1 offers a tradeoff between both metrics. Since both measurements represent rates, F1 is defined as their harmonic mean.

$$F1@k = \frac{1}{\frac{1}{2}\left(\frac{1}{R@k} + \frac{1}{P@k}\right)} = \frac{2}{\frac{R@k}{R@k \cdot P@k} + \frac{P@k}{R@k \cdot P@k}} = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k} \tag{3.35}$$

Mean Reciprocal Rank (MRR)

The ordering of recommended items is very important. Clearly, we want the most relevant track to occur very early within the recommendations. Mean Reciprocal Rank is the average of the reciprocal ranks of the first recommended track that is relevant. Please note, that this method only considers the first track.

$$MRR@k = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank(u)} \tag{3.36}$$

$$rank(u) = \min i, \text{ s.t. } \mathbb{1}_{u,i}(\tilde{T}_k(u)) = 1, 0 < i \leq k \tag{3.37}$$

where $rank(u)$ is the rank of the first recommended item that is relevant for u.

Mean Average Precision (MAP)

Often, it is not only wanted to recommend relevant tracks, but to place them among the first (i.e., most important) recommendations. Average Precision takes also the rank of correctly recommended tracks into account. Hence, it penalizes correctly recommended tracks, which do not lie within the first recommendations.

$$AP@k = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|T_k(u)|} \sum_{i=1}^{k} \mathbb{1}_{u,i}(\tilde{T}_k(u)) \cdot P@k(u) \tag{3.38}$$

Mean Average Precision averages Average Precision over different values of k.

$$MAP@k = \frac{1}{k} \sum_{i=1}^{k} AP@i \tag{3.39}$$

Normalized Distributed Cumulative Gain (nDCG)

Similar to MRR and MAP, Discounted Cumulative Gain [Järvelin and Kekäläinen, 2002] penalizes non-relevant tracks appearing among the first recommendations and relevant tracks not appearing among the first recommendations. Intuitively, the logarithmic term is a scaling w.r.t. the position of the recommended track. Hence, the earlier a relevant track occurs within the recommendations, the higher the DCG is. [Wang et al., 2013] provide sound justifications for the logarithmic scaling. The DCG for the first k recommendations is given by

$$DCG@k = \sum_{i=1}^{k} \frac{2^{\mathbb{1}_{u,i}(\tilde{T}_k(u))} - 1}{\log_2(i+1)} \tag{3.40}$$

The Ideal Discounted Cumulative Gain (IDCG) is equivalent to the optimal DCG. It represents the DCG of the set of relevant track. In other words, it represents the DCG of "perfect" recommendations.

$$IDCG@k = \sum_{i=1}^{k} \frac{2^{\mathbb{1}_{u,i}(T_k(u))} - 1}{\log_2(i+1)} \tag{3.41}$$

Eventually, DCG is normalized by the IDCG. Hence, we get

$$nDCG@k = \frac{DCG@k}{IDCG@k} \tag{3.42}$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error is the square root of the mean squared error between real and predicted rating. Hence, it is very similar to the standard deviation. This is advantageous, since the results of RMSE typically are easy to interpret, as they are on the same scale as the ratings. This technique tolerates small errors (i.e., $|r_{u,i} - \tilde{r}_{u,i}| < 1$) and penalizes large errors (i.e., $|r_{u,i} - \tilde{r}_{u,i}| \geq 1$) in a quadratic manner.

$$RMSE = \sqrt{\frac{1}{|R|} \sum_{\substack{r_{u,i} \in R \\ \tilde{r}_{u,i} \in \tilde{R}}} (r_{u,i} - \tilde{r}_{u,i})^2} \tag{3.43}$$

Mean Absolute Error (MAE)

In some applications, it may be interesting, to not scale errors in any way. Therefore, Mean Absolute Error only averages the absolute difference between real and predicted rating.

$$MAE = \frac{1}{|R|} \sum_{\substack{r_{u,i} \in R \\ \tilde{r}_{u,i} \in \tilde{R}}} (|r_{u,i} - \tilde{r}_{u,i}|) \tag{3.44}$$

Fraction of Concordant Pairs (FCP)

[Koren and Sill, 2013] pointed out that RMSE and MAE fail in representing individual rating scales among users. Hence, they introduced the method of calculating the Fraction of Concordant Pairs. Often, ratings are predicted in order to recommend the top rated items to a user. Hence, it is crucial that an algorithm predicts ratings that preserve the ordering of the real ratings. This is achieved by

counting the number of concordant $n_c(u)$ and disconcordant $n_d(u)$ pairs per user.

$$n_c(u) = |\{(i,j) : \tilde{r}_{u,i} > \tilde{r}_{u,j} \wedge r_{u,i} > r_{u,j}\}| \tag{3.45}$$

$$n_d(u) = |\{(i,j) : \neg(\tilde{r}_{u,i} > \tilde{r}_{u,j} \wedge r_{u,i} > r_{u,j})\}| \tag{3.46}$$

for $r_{u,i}, r_{u,j} \in R$ and $\tilde{r}_{u,i}, \tilde{r}_{u,j} \in \tilde{R}$. Subsequently, $n_c(u)$ can be normalized by

$$FCP(u) = \frac{n_c(u)}{n_c(u) + n_d(u)} \tag{3.47}$$

Of course, it holds that $n_c(u) + n_d(u) = |R \times R|$ is the number of all possible pairs. This measure can be aggregated to represent the FCP for a set of users U as

$$FCP = \sum_{u \in U} \frac{n_c(u)}{n_c(u) + n_d(u)} \tag{3.48}$$

3.5 Miscellaneous Methods

Min-max-scaling refers to scaling the values of vector x, such that the scaled vector $\bar{x}$ has its minimum at zero and its maximum at one.

$$\bar{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{3.49}$$

Adjusted cosine similarity handles items having different rating scales by subtracting the average item rating $\bar{r_i}$. Hence, it weakens the influence of highly dominant items. For users u, v and the set of items I, this method is given by

$$sim(u,v) = \frac{\sum_{i \in I}(r_{u,i} - \bar{r}_i)(r_{v,i} - \bar{r}_i)}{\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_i)^2}\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_i)^2}} \tag{3.50}$$

Chapter 4

Results and Discussion

After conducting all steps described in Chapter3, this section aims to outline some interesting observations. Our analysis is three-fold. Firstly, we investigate a clustering of tracks by means of their acoustic features and genres. In order to concisely describe the properties of a cluster, typical personas listening to a cluster's tracks are listed. Secondly, we provide exhaustive statistics of users' music taste, listening behaviour and demographics. Thirdly, we compare the results of several recommendation algorithms applied on groups of varying mainstreaminess and on user groups with distinct music tastes. The goal of this chapter is to enhance our crude explorative analysis of the Cultural LFM-1b dataset and in particular, focus on properties of low mainstreaminess users.

Since a description of tracks by their acoustic features and genres is desired, we decide to exclude tracks with any invalid acoustic features. Furthermore, dominant genres are removed, as they would deteriorate and blur results. The term "dominant genre" refers to genres below the threshold depicted in Figure 3.13. For the sake of readability, we resign explaining this pruning procedure at any reasonable point in this section. Descriptive statistics of the modified dataset introduced in this chapter can be found in Table 4.1.

Furthermore, illustrations of acoustic features (Figure 4.2), Hofstede's dimensions (Figure 3.3) and World Happiness Report (Figure 3.4) underwent min-max-scaling, which is thoroughly described in Equation 3.49. This procedure was chosen for an easy comparison between several dimensions.

Item	Value
Users	2,074
Tracks	145,131
Artists	14,243
Spotify Track Genres	1,185
Listening Events (LEs)	4,682,141
Min. LEs per user	1
Q_1 LEs per user	1,198
Median LEs per user	1,945
Q_3 LEs per user	3,078
Max. LEs per user	10,536
Avg. LEs per user	2,258.630 ($\pm$ 1,457.310)

Table 4.1: Descriptive statistics of Cultural LFM-1b for LowMs. Only tracks with nondominant genres and valid acoustic features are considered. The value within the parenthesis is the standard deviation.

4.1 Track Clusters

We aim to find clusters, where each cluster represents a collection of tracks similar in their acoustic features. In order to do so, dimensionality reduction with UMAP was conducted. Subsequently, we performed clustering with HDBSCAN.

The resulting clusters are illustrated in Figure 4.1. As can be seen, four clusters are obtained. In this work, the size of a cluster denotes the number of tracks within a cluster, not its diameter. Thus, to what we refer to as size of a cluster does not necessarily coincide with the visual extent shown, since the track clusters heavily deviate from each other in terms of density.

First and foremost, we observe that not all tracks are assigned to a cluster. In particular, 13,988 tracks could not be classified, since they lie in between clusters. Furthermore, the sizes vary heavily. Since the clustering of tracks was conducted utilizing the two dimensional latent representation of tracks' acoustic features, this

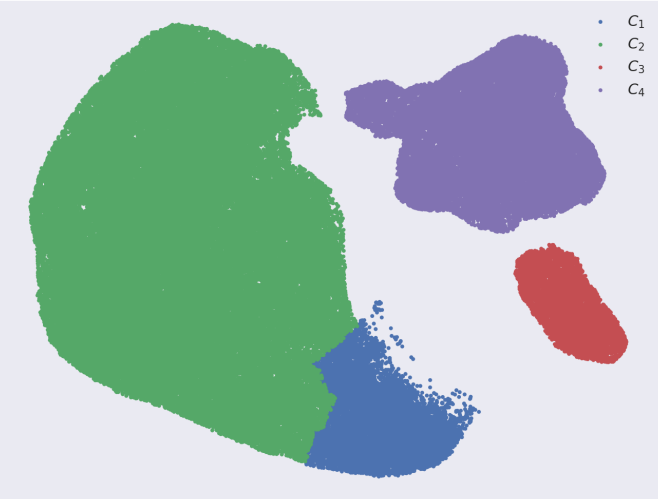


Figure 4.1: Clusters of tracks, obtained by considering tracks' acoustic features. These four clusters comprise a very different number of tracks, i.e., 11,588 (C_1), 85,663 (C_2), 6,446 (C_3) and 27,446 (C_4). Furthermore, 13,988 tracks could not be assigned to a cluster.

observation indicates that there exist certain configurations of acoustic properties that are more widely used among tracks than others. In this work, C_i is defined as the set of tracks in track cluster i. Mainstreaminess is usually related to the listening events per user. Anyway, cluster C_2 can be said to represent mainstream within the set of tracks, with user-interactions pushed aside, as it contains the vast majority of tracks.

The illustration in Figure 4.2 depicts the distributions of each clusters' acoustic features. Danceability and tempo do not show any apparent differences worth interpreting. Thus, we conclude that including danceability and tempo might not yield any advantages. Contrary to that, all remaining dimensions indicate moderate or strong differences between track clusters. Energy, speechiness, acousticness, valence and liveness give strong evidence that there is a large difference between pairs C_1, C_3 and C_2, C_4. Furthermore, they indicate that C_1 is similar to C_3 and C_2 is similar to C_4. Instrumentalness hints that C_1 and C_2 comprise tracks with a large amount of vocals. In contrast, tracks from C_3 and C_4 exhibit only a minor contribution of vocals. Thus, instrumentalness is the only dimension that makes it possible to distinguish four clusters and not only two.

Therefore, we point out the importance of instrumentalness, as it (in combination with the other dimensions) could serve as encoding for the track clusters.

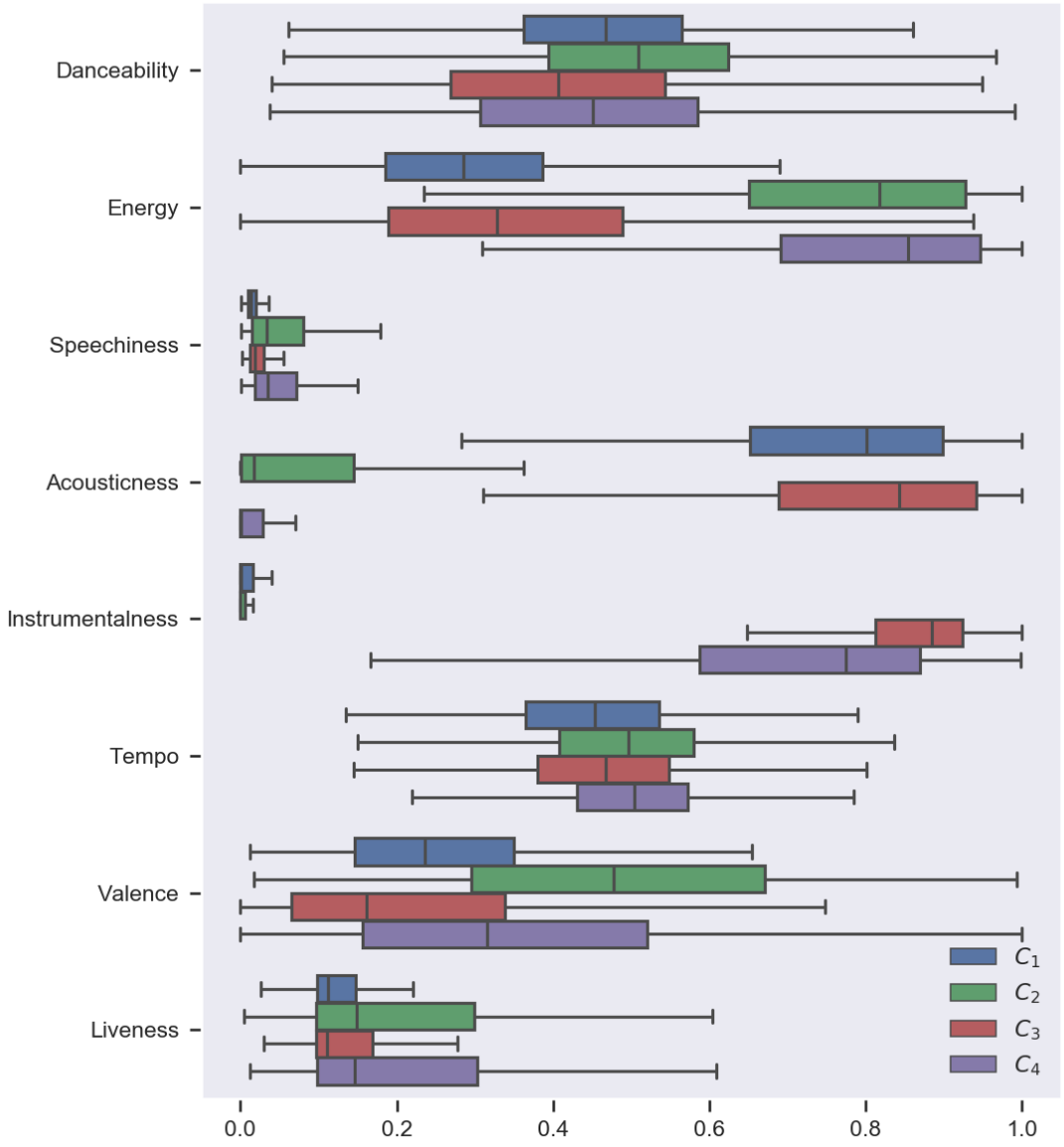


Figure 4.2: Distribution of acoustic features of all track clusters. Danceability and tempo show only minor differences. Interestingly, energy, speechiness, acousticness, valence, liveness and instrumentalness indeed show deviations between clusters. Please note that instrumentalness makes it possible to distuingish four and not only two track clusters.

Due to the large similarity between pairs of clusters in terms of acoustic features, we aim to further precise our interpretations by analyzing the genres of a cluster's tracks. Hence, let T_g be the set of tracks, annotated with genre g and C_i be the set of tracks assigned to cluster i. Then the term frequency $TF_i(g)$ refers to the number of tracks with genre g within track cluster C_i.

$$TF_i(g) = |T_g \cap C_i| \tag{4.1}$$

Additionally, let $IDF(g)$ be the idf-score of genre g. Contrary to the idf-score, $TFIDF_i(g)$ does not only consider the importance of g, but also its frequency. Thus, frequent but unimportant genres contribute to this scoring as well as less frequent but important ones. The tfidf-score of g for cluster i can be computed via

$$TFIDF_i(g) = TF_i(g) \cdot IDF(g) \tag{4.2}$$

Eventually, a tfidf-scoring of genres within each cluster is achieved. Utilizing this technique, track clusters are explained by their top scoring genres. This is advantageous over explanations via acoustic features, since genres can be more easily interpreted.

The top genres of each cluster are illustrated in Figure 4.3. These tfidf-score distributions depict the genres serving as description of each track cluster. In particular, the distributions for C_1 and C_3 show interesting behavior. These two track clusters include genres (singersongwriter and folk for C_1, ambient for C_3) with a much higher tfidf-score than other genres. This effect cannot be observed in any other track cluster. The aforementioned high-scoring genres can be said to be mainstream within their track cluster. This behavior could be caused by two things. Firstly, a genre with low importance (i.e., low idf-score) may occur much more frequent in a single cluster than it does over all tracks. Secondly, globally dominant genres do not have to be equal to the dominant genres of a single cluster. If this was the case, genres only dominating their track cluster would survive the pruning procedure proposed in Figure 3.13. As a counterexample, the distributions for C_2 and C_4 exhibit a genre distribution of roughly linear increase. Hence, C_2 and C_4 do not comprise any dominant genres.

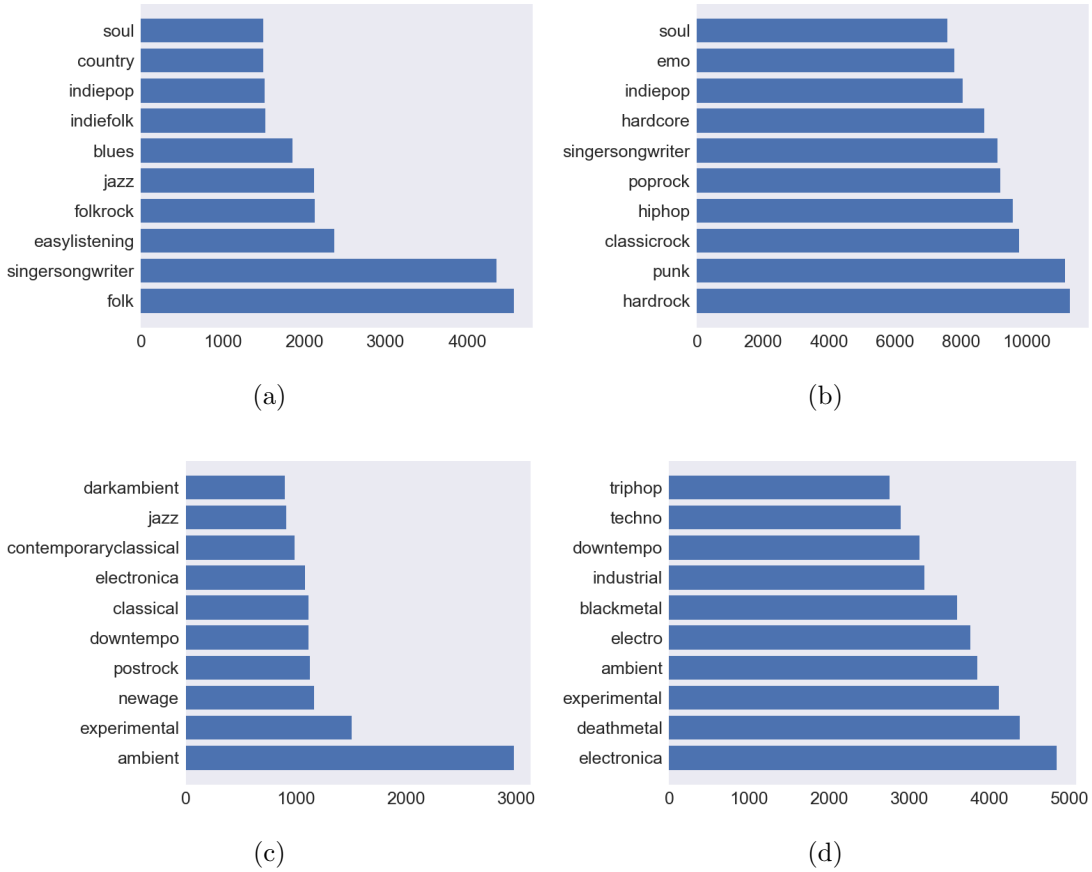


Figure 4.3: IDF-distribution of track clusters based on genres. From top left to bottom left clockwise: IDF-distribution for (a) C_1, (b) C_2, (c) C_3 and (d) C_4. We can see distributions, where the idf-score increases linearly, hence there are no dominant genres. In contrast, uneven distributions indicate the presence of dominant genres.

Based on the top genres and the acoustic features, we define several personas, where each persona represents the typical kind of music within the corresponding track cluster. Persona P_i describes the music style within track cluster i.

1. Person P_1 (Complex Music): Low-energetic, mostly acoustic and complex music. In particular folk, singersongwriter, jazz, blues.

2. Person P_2 (Festival Music): high-energetic, non-acoustic music. This is the most mainstream cluster within LowMs music. In particular hardrock, punk, hiphop.

3. Person P_3 (Relax Music): low-energetic, acoustic, slow and instrumental music. In particular ambient, experimental, newage, classical.

4. Person P_4 (Heavy Music): high-energetic, non-acoustic, fast and instrumental music. In particular electronica, deathmetal, industrial.

The depiction in Figure 4.4 illustrates the top 30 genres and their relative genre importance. Relative genre importance is defined as the fraction of listening events per genre within a track cluster.

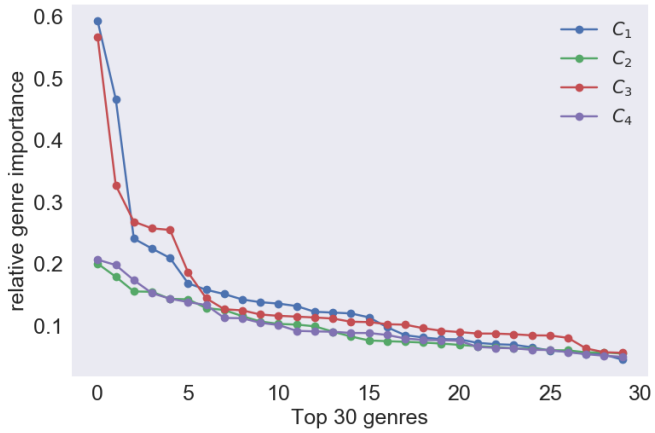


Figure 4.4: Relative genre importance denotes the fraction of listening events per genre within a track cluster. Notice the presence of dominant genres in clusters C_1 and C_3.

It can be observed that for C_2 and C_4, importance is roughly uniformely distributed. Contrary to that, C_1 and C_3 clearly exhibit a nonuniform distribution. In these two clusters, a few genres get much more attention from users than other genres. Please note that C_2 and C_4 are the largest track clusters. Removing the dominating genres (i.e., rock, pop, electronic, metal, alternativerock, indierock) as a preprocessing step did not work well for the smaller clusters C_1 and C_3. The calculation of $IDF(g)$ relies on the number of tracks, which are annotated with a certain genre g. The larger the cluster, the higher the chance of the cluster comprising some of the dominating genres. Hence, it is obvious that the dominating genres coincide with the top genres of the larger clusters and deviate from the top genres of the smaller ones. Removing the dominating genres removes

the top genres of C_2 and C_4, but not all of C_1 and C_3.

Anyway, we see in Table 4.2 that after pruning, also C_2 and C_4 exhibit very different genres. Hence, dominant genres in C_1 and C_3 are only dominant within their cluster. Another reason for this phenomenon could be the presence of subgenres linked to a main genre. Obviously, tracks of a certain subgenre are very likely to be also annotated with the main genre. Hence, the main genre gains popularity. In conclusion, dominating genres of all tracks do not entirely coincide with the top genres of each cluster, but the intersection is larger for large track clusters than for small ones.

Track Cluster	Top Genres in terms of listening events
C_1	singersongwriter, folk, folkrock, easylistening, indiepop
C_2	hardrock, punk, hardcore, poprock, hiphop
C_3	ambient, experimental, postrock, electronica, downtempo
C_4	experimental, electronica, ambient, deathmetal, postrock

Table 4.2: Top genres of each track cluster in terms of listening events. The top genres show interesting patterns for all track clusters. Except C_3 and C_4, which share a large subset of top genres.

Inspired by recent work of [Antenucci et al., 2018], we introduce $ArtistHeterogeneity$, which is aimed to measure the diversity in terms of a user's listening behavior. Let $T(u)$ be the set of distinct tracks, a user u listened to and $artist(t)$ be the artist of track t, then

$$ArtistHeterogeneity(u) = \log_2 \left(\frac{|T(u)|}{|\{artist(t) : t \in T(u)\}|} \right) \tag{4.3}$$

Similarly, we modify the latter equation to consider genres, hence, let $genre(t)$ be the genre of track t, then

$$GenreHeterogeneity(u) = \log_2 \left(\frac{|T(u)|}{|\{genre(t) : t \in T(u)\}|} \right) \tag{4.4}$$

where a low heterogeneity-score indicates high diversity.

As can be seen in Table 4.3, track clusters exhibit very different heterogeneity-scores for both, artists and genres. In general, C_2 and C_4 have the highest heterogeneity-score. Hence, those clusters exhibit a large number of distinct tracks compared to the number of distinct artists and genres. This observation could be biased by the cluster size, as the number of distinct tracks grows much faster than the number of artists and genres. The logarithmic scaling should mitigate this issue. Anyway, that is no contradiction to the intent of using these measurements, which is to measure diversity.

Track cluster	AH	GH
C_1	1.8012	4.1741
C_2	2.9388	6.2778
C_3	1.5450	3.6034
C_4	2.3341	5.0769

Table 4.3: Track clusters' artist- and genre-heterogeneity. Artist-heterogeneity (AH) indicates that clusters C_1 and C_3 comprise tracks of a variety of artists. Equivalently, genre-heterogeneity hints that C_1 and C_3 also exhibit a more diverse set of genres than C_1 and C_4.

4.2 User Groups

In this section we outline the user-focused analysis. Based on the number of tracks within a cluster that users listened to (i.e., weight), users are assigned to exactly one track cluster. In order to weaken the influence of large clusters, these weights are scaled by an idf-scoring of clusters. User u is said to belong to U_{C_i}, if cluster C_i is the highest weighted track cluster. Please note that one user from LowMs did not listen to any classified tracks, thus, this user is not assigned to any user group. As a consequence, $|U_{C_1} \cup U_{C_2} \cup U_{C_3} \cup U_{C_4}| \neq |LowMs|$. Eventually, the resulting user groups are investigated regarding their demographics, behaviour and properties of consumed music.

Analogous to the investigations concerning the top genres per track cluster in Figure 4.4, we conduct an analysis of the top genres per user group. Figure 4.5

illustrates the distribution of the top 30 genres per user group. Interestingly, the user-based distribution exhibits exactly the same properties as in the track-based case. This indicates a very strong correlation of a user group's listening events with the listening events of its corresponding track cluster. The top 5 genres of a user group U_{C_i} coincide with the top 5 genres of its linked track cluster C_i by at least 80% as Tables 4.2 and 4.4 prove. Therefore, music taste of a user group is very similar to the type of music its corresponding track cluster offers.

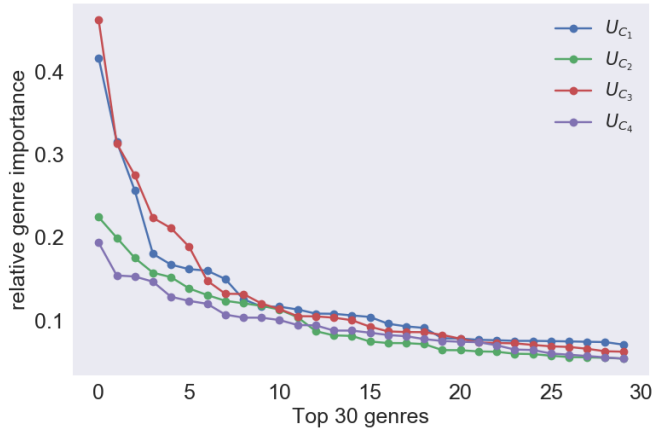


Figure 4.5: Top 30 genres of user groups. Relative genre importance denotes the fraction of listening events per genre within a user group. Notice the presence of dominant genres in clusters U_{C_1} and U_{C_3}.

User Group	Top Genres in terms of listening events
U_{C_1}	singersongwriter, folk, indiepop, folkrock, poprock
U_{C_2}	hardrock, punk, hardcore, poprock, emo
U_{C_3}	ambient, experimental, electronica, downtempo, postrock
U_{C_4}	experimental, ambient, electronica, deathmetal, hardock

Table 4.4: Top genres of each user group in terms of listening events. The top genres show interesting patterns for all user groups. Except C_3 and C_4, which share a large subset of top genres.

Note Since some paragraphs in the remainder of this chapter are concerned with the analysis of different types of music listeners, we occasionally refer to the personas introduced in Section 4.1. As a short reminder, U_{C_1}: Complex Listener, U_{C_2}: Festival Listener, U_{C_3}: Relax Listener and U_{C_4}: Heavy Listener.

The correlation matrix of user groups based on weights (i.e., listening events towards track clusters) is provided in Figure 4.6. Some correlations can be explained in more detail by the previously conducted analysis of acoustic features. Complex Listeners tend to listen to similar tracks as Relax Listeners. Furthermore, Complex Listeners do not like the genres hiphop, punk, hardrock, deathmetal and electronica. It is apparent that Festival Listeners rather stay within their preferred type of music, as they refuse to listen to deathmetal, ambient, electronica, jazz or soul. This observation does not comply with our previous analysis of acoustic features. Relax Listeners choose to additionally listen to genres like folk, jazz and blues. Surprisingly, completely different genres such as deathmetal, blackmetal and triphop also attract Relax Listeners. Heavy Listeners value relaxing music, but neglect to listen to typical festival music.

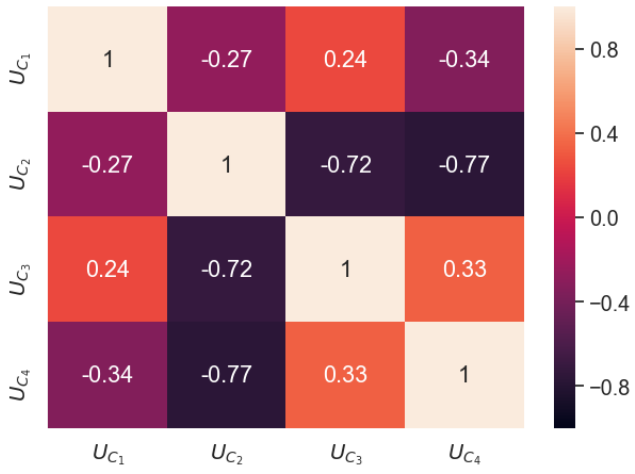


Figure 4.6: Correlation between user groups, where the correlation is based on each user's weights. Interestingly, U_{C_2} tends to heavily dislike relaxing music and metal. Furthermore, U_{C_3} and U_{C_4} apparently do not like to listen to festival music. U_{C_2} seems to not explore music styles other than festival music.

The depictions in Figure 4.7 illustrate the contribution (i.e., weight) of the track clusters to the user groups. Again, to reduce the influence of large track clusters, weights are scaled with the cluster's idf-score. It can be observed that the two largest user groups (i.e., Festival Listeners U_{C_2} and Heavy Listeners U_{C_4}) mainly prefer to listen to their favorite type of music, whilst not listening exhaustively to tracks from other track clusters. In contrast, U_{C_1} (i.e., Complex Listeners) and U_{C_3} (i.e., Relax Listeners) listen to a variety of music styles, as their consumption behavior is much more uniformly distributed over track clusters. These observations give evidence of U_{C_1} and U_{C_3} exhibiting a higher degree of omnivorousness than U_{C_2} and U_{C_4}. Omnivorousness refers to the breath of an individual's music taste [Atkinson, 2011]. Additionally, the aforementioned observations are backed by the Kullback-Leibler Divergence between the weight distribution and a uniform distribution.

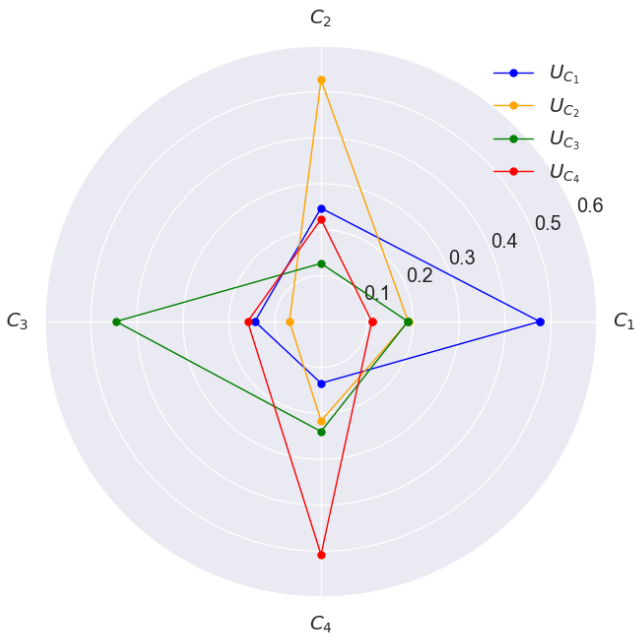


Figure 4.7: Visualization of normalized weights $\tilde{w}$, averaged over all users of a certain user group. As one can see, U_{C_2} and U_{C_4} tend to focus on their preferred style of music, whereas U_{C_1} and U_{C_3} seem to distribute their listening events more evenly on all music styles.

User Group	$KL(P_{\tilde{w}}\|\|U_4)$
U_{C_1}	0.3272
U_{C_2}	0.5181
U_{C_3}	0.2311
U_{C_4}	0.3796

Table 4.5: Kullback-Leibler Divergence between the probability distribution of normalized weights $P_{\tilde{w}}$ and a uniform distribution with four events, representing four track clusters. Numerical results indicate that U_{C_2} and U_{C_4} tend to focus on their preferred style of music and U_{C_1} and U_{C_3} seem to distribute their listening events more evenly on all music styles.

User Group	AH	GH
U_{C_1}	1.8068	13.5003
U_{C_2}	2.0110	16.3864
U_{C_3}	1.8053	12.6542
U_{C_4}	2.1539	13.7443

Table 4.6: User groups' artist- and genre-heterogeneity. Artist-heterogeneity (AH) indicates that U_{C_1} and U_{C_3} tend to listen to various artists. Furthermore, genre-heterogeneity (GH) shows evidence for U_{C_2} being the least diverse user group.

In analogy to the artist- and genre-heterogeneity of track clusters, we conduct the same measurement for user groups. The results are depcited in Table 4.6. Similar to track clusters, U_{C_1} and U_{C_3} show the highest degree of artist- and genre-heterogeneity. This gives indication of these two user groups being the most diverse ones in terms of artists and genres. More interestingly, we can compare the artist-heterogeneity of track clusters in Table 4.3 with the artist-heterogeneity of user groups. It can be observed that the difference in artist-heterogeneity between U_{C_2} and C_2 is very large. This hints that the music style resp. persona representing C_2 (i.e., Festival Listener) does deviate from the observed listening behavior of U_{C_2}. This raises concerns that the Festival Listener is not a good exemplar for U_{C_2} and does not represent its music style well. Contrary to that, e.g., U_{C_1} and C_1 do not show any difference. Hence, we conclude that the observed

listening behavior of users assigned to U_{C_1} does indeed coincide with the music style described by C_1 (i.e., Complex Listener).

In order to assess the cohesiveness of user groups, we measure the pairwise similarity within each user group. Each user is represented by a vector comprising the number of listening events per genre. Motivated by observing multiple genres with a extraordinarily high number of listening events in Figure 4.4, we choose to utilize adjusted cosine similarity as measurement, since highly popular genres would deteriorate the results. Clearly, neglecting the presence of user groups and hence, considering all users from LowMs yields the smallest similarity. U_{C_1} and U_{C_3} exhibit the highest degree of similarity. Please note that this behavior can be caused by the fact that those two user groups are far smaller than the other two. Hence, they may not comprise a large variety of usertypes. Furthermore, U_{C_2} includes pairs of users that seem to have only minor common music taste. The latter finding is also backed by the large variance of 0.1350.

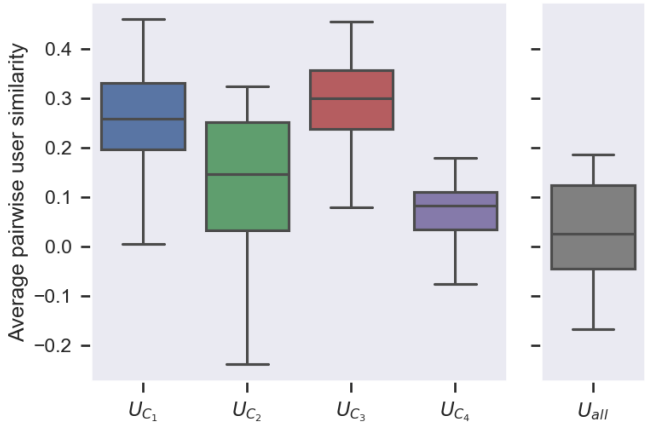


Figure 4.8: Average pairwise user group similarity based on genres. The similarity measure chosen is the adjusted cosine similarity. U_{all} denotes the set of all users. Furthermore, the average similarities between a user and all other users in the user group show variances 0.1329 (U_{C_1}), 0.1350 (U_{C_2}), 0.1146 (U_{C_3}), 0.1099 (U_{C_4}) and 0.1252 (U_{all}).

Descriptive statistics for all four user groups can be found in Table 4.7. The ordering in terms of user group size apparently corresponds to the size of the track

clusters. Anyway, we were able to significantly weaken the deterioration caused by the largest track cluster C_2 via the idf-scaling, as only approx. 43% of users are assigned to this track cluster, whereas it contains approx. 65% of tracks. Furthermore, it can be observed that the number of distinct genres, tracks and artists is not correlated to either user group or track cluster size. We have shown in Section 4.1 that C_1 and C_3 are highly specialized track clusters.

User Group	$\|U\|$	$\|A\|$	$\|T\|$	$\|G\|$	$\|LE\|$	$\|Avg.G/T\|$
U_{C_1}	396	9,673	76,520	1,085	783,090	5.061
U_{C_2}	900	11,453	104,266	1,095	2,094,082	4.590
U_{C_3}	102	5,621	32,172	918	186,921	5.772
U_{C_4}	675	11,710	111,872	1,128	1,618,048	4.792

Table 4.7: Descriptive statistics of all user groups. $|U|$ is the number of users, $|A|$ is the number of distinct artists, $|T|$ is the number of distinct tracks, $|G|$ is the number of distinct genres, $|LE|$ is the number of listening events and $|Avg.G/T|$ is the average number of genres per track.

Furthermore, U_{C_1} and U_{C_3} have only a small number of listening events compared to U_{C_2} and U_{C_4}. Therefore, it is surprising that users assigned to those clusters achieve to listen to nearly all genres. This leads to the conclusion that the majority of tracks are annotated with a variety of genres. Additionally, this behavior can be explained by a simple experiment. Assume a user chooses randomly, which genres, tracks or artists to listen to. If this decision is made independently for 50,000 times, it can be observed that the number of distinct genres and artists saturates at some point. The results of this experiment are depicted in Figure 4.9.

Additionally, $|Avg.G/T|$ indicates tracks listened by U_{C_3} being annotated with slightly more genres than tracks listened by other user groups. This observation can be caused by two things. Firstly, tracks listened by U_{C_3} may exhibit more genres and subgenres (e.g., ambient, dark ambient). Secondly, users may have a more diverse taste in music and thus, listen to more complex tracks. Please note that there is a large overlap in genres listened by different user groups.

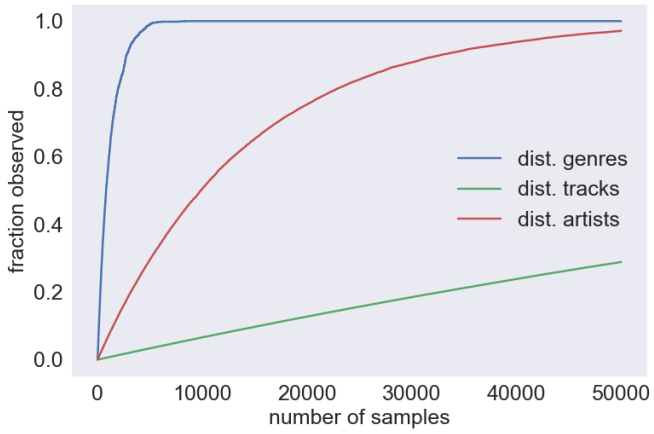


Figure 4.9: Convergence of the number of distinct genres, tracks and artists. In total, 50,000 samples from the uniform distribution over genres, tracks and artists are drawn. Fraction observed denotes the fraction of genres, tracks and artists observed during sampling.

User Group	$\|Avg.G/U\|$	$\|Avg.T/U\|$	$\|Avg.LE/U\|$	$\|Avg.LE/T/U\|$
U_{C_1}	235.987	584.692	1,977.500	0.026
U_{C_2}	212.516	552.461	2,326.758	0.022
U_{C_3}	228.775	522.010	1,832.559	0.057
U_{C_4}	249.733	729.640	2,397.108	0.021

Table 4.8: User-related statistics of all user groups. $|Avg.G/U|$ is the average number of distinct genres a user has listened to, $|Avg.T/U|$ is the average number of distinct tracks a user has listened to, $|Avg.LE/U|$ is the average number of listening events per user and $|Avg.LE/T/U|$ is the average number of listening events per track of a user.

The statistics outlined in Table 4.8 are based on the interactions between users and tracks resp. genres. The goal is to give insights into user-related properties of music consumption. It can be seen that U_{C_2} listens to less distinct genres than other user groups. Similarly, users from U_{C_3} tend to listen to fewer distinct tracks. This observation can be strengthend by the standard deviation of 345.526 being small compared to 391.666 (U_{C_1}), 383.139 (U_{C_2}) and 451.342 (U_{C_4}). With the same

arguments as before, we can conclude that users from U_{C_4} neglect listening repetitively to the same tracks. Furthermore, notice that U_{C_1} and U_{C_3} are much less active in terms of listening events than U_{C_2} and U_{C_4}. Surprisingly, such a behavior cannot be observed for the average number of tracks played. We found that U_{C_3} tends to repetitively listen to the same set of tracks, as the number of listening events per track is high for this user group. This observation could be induced by the scarcity of tracks in C_3.

User Group	LE_{work}	$\|Avg.LE_{day}/U\|$
U_{C_1}	41.073%	17.658
U_{C_2}	40.089%	25.019
U_{C_3}	42.570%	23.891
U_{C_4}	42.003%	24.478

Table 4.9: Temporal statistics of all user groups. $|LE_{work}|$ is the fraction of listening events throughout a workday (i.e., 7-18 on businessdays) and $|Avg.LE_{day}/U|$ is the average number of listening events of a user per day.

Several interesting observations can be made about the temporal listening behavior of different user groups. Statistics in Table 4.9 indicate that users from U_{C_2} tend to focus their music consumption to their spare time and not to work time. On average, users from U_{C_1} listen to less music per day than others. Digging deeper, listeners of genres like jazz, folk, blues and soul do not listen to as many pieces of music per day as other listeners. This observation could be influenced by the length of tracks since, e.g., blues tracks' duration is in general much longer than the duration of punk tracks.

As already noted in Section 3, users from some homecountries (i.e., US, RU, DE, UK, BR, PL) dominate this dataset. When analyzing demographic properties, this leads to a strong bias towards those countries. Table 4.10 verifies that country-based analyses would not make sense, since the dominant countries would deteriorate results. Hence, statistics shown in Table 4.11, Figure 4.10 and Figure 4.11 are obtained by only considering users from nondominant countries. Additionally, all users without valid age and country information are excluded. This leads to modified user groups, where each user group $\tilde{U}_{C_i}$ is a proper subset of

the unmodified user group U_{C_i}. The cardinalities (i.e., number of users) are $|\tilde{U}_{C_1}| = 180$, $|\tilde{U}_{C_2}| = 306$, $|\tilde{U}_{C_3}| = 35$ and $|\tilde{U}_{C_4}| = 245$.

User Group	Top Countries
U_{C_1}	US, DE, BR, UK, ES
U_{C_2}	US, DE, BR, UK, RU
U_{C_3}	US, RU, DE, UK, JP
U_{C_4}	US, RU, DE, UK, PL

Table 4.10: Top 5 homecountries of user groups without omitting dominant countries. Obviously, most user groups are biased towards US, DE, BR, RU and UK.

Table 4.11 indicates that $\tilde{U}_{C_4}$ has the highest degree of mainstreaminess within LowMs. Interestingly, young people value metal and festival-like music. In contrast, slower and acoustic music is clearly preferred by individuals of higher age. The male to female ratio is very large for $\tilde{U}_{C_3}$ and $\tilde{U}_{C_4}$. Please notice that these two user groups comprise tracks of high instrumentalness. Thus, males seem to favor music that exhibits less vocals. After excluding dominant countries, we perceive a much clearer picture of user groups' homecountries. We find that spanish and dutch listeners prefer rather complex music. Finns tend to listen to metal and surprisingly, to festival music. Furthermore, japanese and indian listeners focus more on relaxing music.

User Group	$\lvert Avg.MS \rvert$	$\lvert Avg.Age \rvert$	M/F	Top Countries
$\tilde{U}_{C_1}$	0.041	27.925	63%/37%	ES, NL, FR, SE, IT
$\tilde{U}_{C_2}$	0.043	23.910	67%/33%	AU, FI, ES, FR, NL
$\tilde{U}_{C_3}$	0.041	31.691	82%/18%	JP, ID, NL, TR, BE
$\tilde{U}_{C_4}$	0.048	24.538	82%/18%	UA, FI, CA, IT, AU

Table 4.11: Demographic statistics of all user groups. $|Avg.MS|$ is the average mainstreaminess, $|Avg.Age|$ is the average age, M/F is the male to female ratio and Top Countries denotes the top 5 homecountries omitting dominant countries.

An illustration of Hofstede's cultural dimensions can be found in Figure 4.10. Heavy Listeners apparently believe that power is distributed unequally within their society. Additionally, they - together with Relax Listeners - show a large degree of masculinity. Therefore, masculine properties like heroism and wealth are valued more than feminine dimensions like, e.g., caring for others. Despite hofstede's term of masculinity not being directly related to the gender, the population of these user groups is mainly male. Complex Listeners do not tolerate unorthodox beliefs and despise ambiguity and uncertainty. Long-term orientation is high for Relax Listeners. Hence, they are future-oriented rather than tradition-driven. Furthermore, they encourage adaption and pragmatic problem-solving. Interestingly, the cultural properties of Complex and Relax Listeners could be justified by the higher average age of these user groups.

Figure 4.11 illustrates the distributions of the World Happiness Report's dimensions. Complex Listeners show a higher degree of happiness than all remaining user groups. Furthermore, they are expected to have the longest span of life. Interestingly Relax Listeners tend to believe that they have the freedom to make life choices by themselves and not being under influence from others. Festival and Heavy Listeners perceive the corruption in their homecountry as high, whereas Relax and Complex Listeners do not reckon the amount of corruption to be as serious. Perceived corruption is apparently high for user groups of lower age. Please note that it was verified that no user group has more than half of its users from a single homecountry. Hence, equal medians in Generosity are a coincidence.

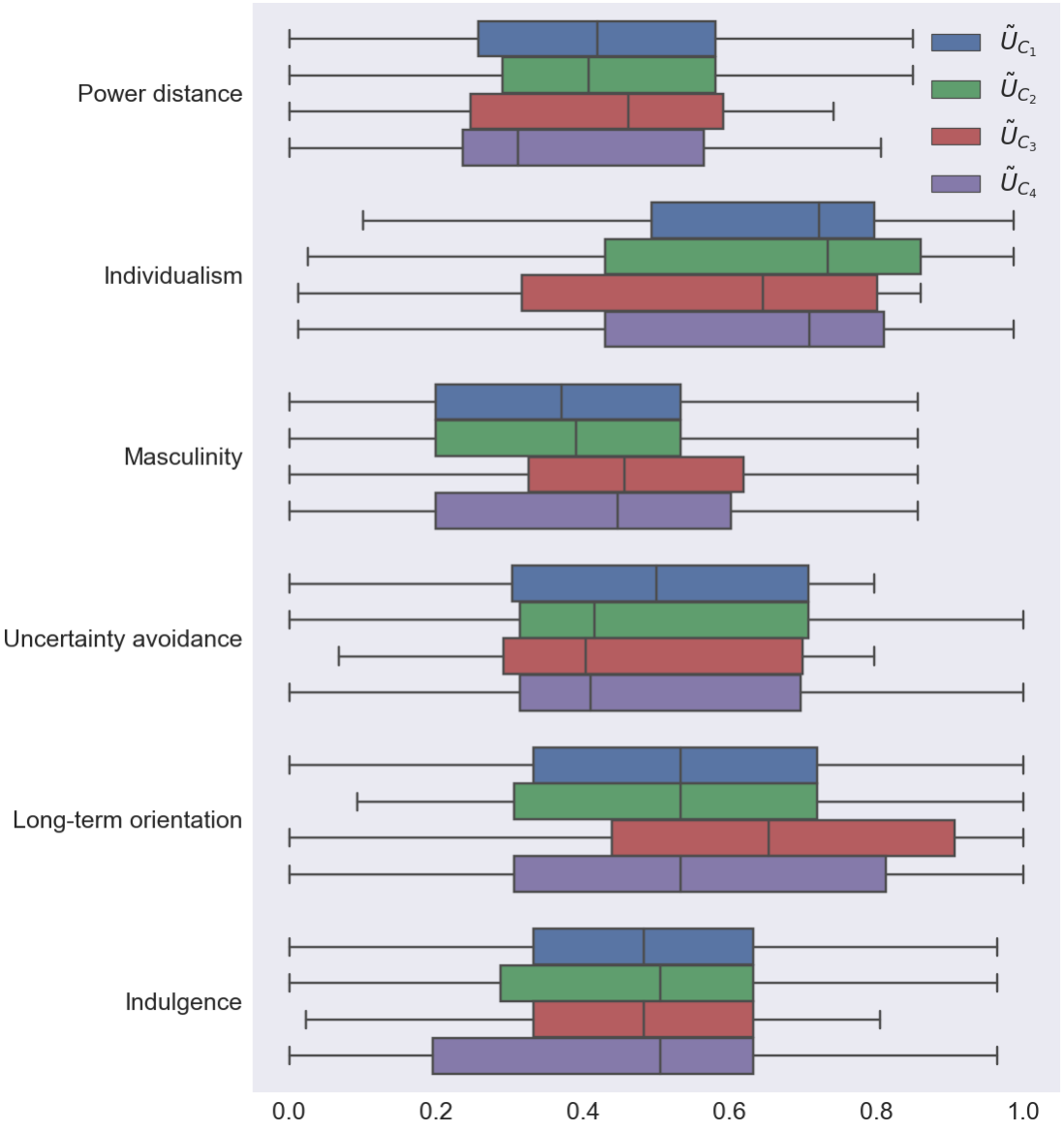


Figure 4.10: Distribution of Hofstede's cultural dimensions of user groups $\tilde{U}_{C_1}$, $\tilde{U}_{C_2}$, $\tilde{U}_{C_3}$ and $\tilde{U}_{C_4}$. It can be observed that $\tilde{U}_{C_4}$ (i.e., Heavy Listeners) tend to believe that power is not distributed equally among society. Furthermore, $\tilde{U}_{C_3}$ (i.e., Relax listeners) are much more future-oriented than other user groups.

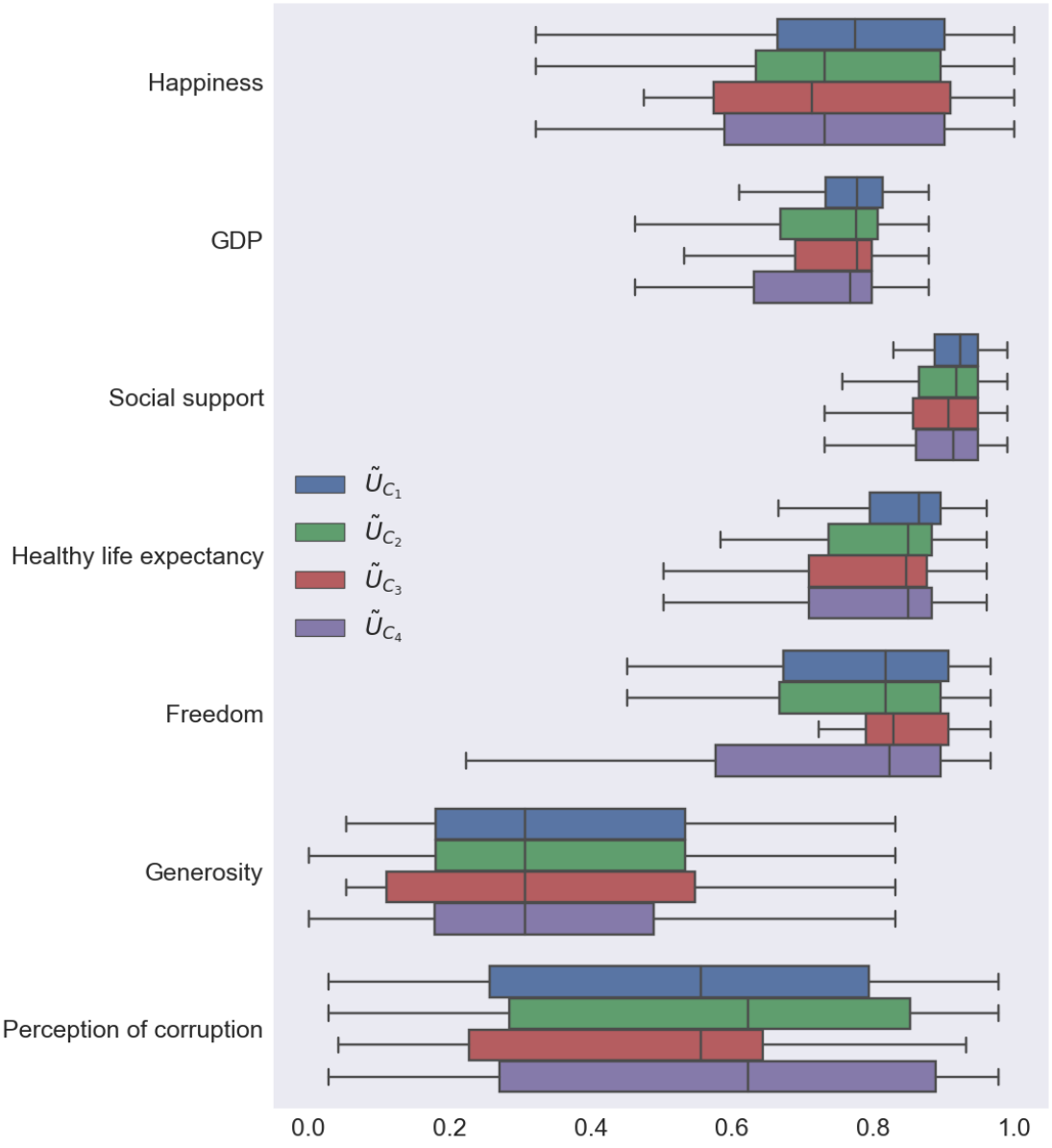


Figure 4.11: Dimensions of the World Happiness Report of user groups $\tilde{U}_{C_1}$, $\tilde{U}_{C_2}$, $\tilde{U}_{C_3}$ and $\tilde{U}_{C_4}$. This depiction indicates that $\tilde{U}_{C_1}$ (i.e., Complex Listeners) are slightly happier than other user groups. Furthermore, users from $\tilde{U}_{C_3}$ apparently have the freedom to make choices by themselves. Please note that $\tilde{U}_{C_3}$ is a user groups with rather high average age.

4.3 Recommendations

For the sake of assessing how well-known recommendation algorithms perform on groups of different mainstreaminess and other subsets of users we train a selection of algorithms and evaluate them on groups of different mainstreaminess and taste.

4.3.1 Rating Prediction

Here, we perform 5-fold cross-validation and average the error metrics. As can be observed, BASE gives the best results in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). KNN outperforms all competing methods in terms of Fraction of Concordant Pairs (FCP). The reason for KNN's problems could be the high level of sparsity (i.e., All: 99.87%, NormMs: 99.78% and LowMs: 99.86%). Additionally, a one-tailed t-test was conducted ($\alpha = 0.005$), with the nullhypothesis of the MAE for NormMs being larger or equal to the MAE for LowMs. We favor MAE over RMSE, since we aim to get a clearer picture on the rating prediction performance [Willmott and Matsuura, 2005].

Group	Metric	TOP	NORM	PL	BASE	KNN
LowMs	MAE	***63.9866	***88.7751	***472.3068	***63.4515	***69.0510
	RMSE	124.9542	136.8513	607.4750	**107.5587**	118.5350
	FCP	0.4850	0.4504	0.5002	0.5189	**0.5463**
NormMs	MAE	56.4509	83.0046	478.5724	**53.2003**	60.4767
	RMSE	108.7255	125.3239	606.9077	**92.4414**	102.1099
	FCP	0.5086	0.4509	0.5001	0.5201	**0.5472**
All	MAE	59.2702	85.1634	476.2283	**57.0354**	63.6845
	RMSE	115.0657	129.7564	605.8299	**98.3811**	108.5464
	FCP	0.5046	0.4506	0.5001	0.5199	**0.5468**

Table 4.12: Prediction errors for groups of different mainstreaminess and several recommendation methods. For all measurements marked with ***, a one-tailed t-test ($\alpha = 0.005$) indicates that the MAE is significantly higher than for NormMs. The best performing methods are written in bold. Interestingly, BASE outperforms all its competing methods in terms of MAE and RMSE. For FCP, KNN gives the best results.

We found significant evidence that the MAE for LowMs is higher than for NormMs for all methods. In other words, rating prediction is significantly worse for users of unorthodox, non-popular music taste.

Within the previous paragraphs, we showed that contemporary recommendation algorithms perform significantly worse for users of unorthodox taste than for users that like popular music. In the remainder of this section, we conduct further in-depth analyses, since we question how recommendation quality varies for different music styles preferred by non-mainstream users. Hence, a set of algorithms is trained only on data from LowMs. Again, we employ 5-fold cross-validation and average the MAE. Eventually, we evaluate each algorithm separately for each user group. The MAE for all four user groups can be found in Table 4.13. In concordance to the results of the analogous experiment for LowMs and NormMs in Table 4.12, BASE performs best. Notably, most methods seem to have problems with U_{C_2}. One reason for this finding could be the high degree of sparsity for U_{C_2}'s rating matrix. In particular, we note that there might be a relationship between MAE and sparsity, as the sparsities of user groups (U_{C_1}: 99.57%, U_{C_2}: 99.74%, U_{C_3}: 98.72% and U_{C_4}: 98.72%) are correlated to the MAE for most methods.

User Group	TOP	NORM	PL	BASE	KNN
U_{C_1}	65.6952	93.6594	482.6342	**63.9458**	72.7809
U_{C_2}	68.7394	95.3009	479.3241	**65.9453**	75.9294
U_{C_3}	62.1081	92.5584	488.0025	**60.8146**	72.5826
U_{C_4}	62.9007	92.9105	483.3543	**61.5985**	73.0778
U_{all}	65.6132	93.9407	481.9690	**63.6613**	74.0654

Table 4.13: MAE of five recommendation algorithms learnt directly on data of all user groups (i.e., LowMs). The best method for each user group is in bold. BASE clearly outperforms all other methods. Surprisingly, TOP yields better recommendations than, e.g., KNN. We hypothesize that the low MAE for U_{C_3} may by attributed to the low sparsity of this user group's rating data.

As the differences between user groups in Table 4.13 are not obvious, we perform a further statistical analysis and provide evidence that the MAE indeed

significantly deviates between pairs of user groups. Firstly, ANOVA ($\alpha = 0.05$) shows that the MAEs of all user groups are not equal. In a second step, post-hoc analysis via Tukey-HSD ($\alpha = 0.05$) is conducted. Significant pairwise differences are reported in Table 4.14. Pairs are marked with **, if both, ANOVA and Tukey-HSD are significant over all folds. All algorithms agree on pairs U_{C_2}, U_{C_3} and U_{C_2}, U_{C_4} exhibiting significant differences. This gives strong evidence for a gap in recommendation quality between one, Festival Listeners and Relax Listeners and two, Festival Listeners and Heavy Listeners. Furthermore, the two best performing methods (i.e., TOP and BASE) may be more valuable. They indentify more significant differences than NORM, PL and KNN.

	TOP				NORM				PL				BASE				KNN			
User Groups	U_{C_1}	U_{C_2}	U_{C_3}	U_{C_4}	U_{C_1}	U_{C_2}	U_{C_3}	U_{C_4}	U_{C_1}	U_{C_2}	U_{C_3}	U_{C_4}	U_{C_1}	U_{C_2}	U_{C_3}	U_{C_4}	U_{C_1}	U_{C_2}	U_{C_3}	U_{C_4}
U_{C_1}		**	**	**										**	**	**		**		
U_{C_2}	**		**	**			**	**					**		**	**	**		**	**
U_{C_3}	**	**				**							**	**				**		
U_{C_4}	**	**				**							**	**				**		

Table 4.14: Significant differences between pairs of user groups (marked with **), as determined by ANOVA ($\alpha = 0.05$) and a subsequent Tukey-HSD test ($\alpha = 0.05$). Results for PL varied among folds, hence, we are not able to provide any significant results. Anyway, note that all remaining methods agree on a significant difference between pairs U_{C_2}, U_{C_3} and U_{C_2}, U_{C_3}. Interestingly, no method found evidence that the MAEs for U_{C_3} and U_{C_4} deviate significantly from each other.

Anyway, no algorithms found evidence for the MAEs of U_{C_3} and U_{C_4} being different. Reconsidering the top genres of each user group in Table 4.4, we see that the top three genres in terms of listening events are identical. Hence, both user groups focus on a common set of genres. As the models are trained on all data from LowMs, there exist more ratings for these common genres than for other ones. In other words, U_{C_3} and U_{C_4} share a common subset of training data. This could explain the slightly smaller MAEs for both user groups. Additionally, we like to point out that the predicted ratings for U_{C_3} are influenced by ratings from U_{C_4} and vice versa. The algorithms aim to find the best model for U_{C_3} and U_{C_4} simultaneously. Thus, increasing accuracy for one group must not lead to a decrease in accuracy for the second group. As a consequence, the learnt model

attains a state, in which no group is disadvantaged. We hypothesize that this leads to an adaption of MAE for both, U_{C_3} and U_{C_4}.

4.3.2 Top-k recommendations

In a real scenario, recommender systems do not present predicted ratings to a user, but items. Hence, we recommend k tracks to each user and measure how well these recommendations fit the best rated 30 tracks in a user's testset. Similar to the previous paragraph, we conduct 5-fold cross-validation and average the results.

Group	Metric	TOP	NORM	PL	BASE	KNN
LowMs	F1@5	0.0833	0.0758	0.0746	0.0959	**0.1196**
	F1@10	0.1377	0.1255	0.1247	0.1568	**0.2036**
	MRR@10	0.0305	0.0281	0.0276	0.0348	**0.0424**
	MAP@10	0.2602	0.2347	0.2297	0.3039	**0.3842**
	nDCG@10	0.1372	0.1248	0.1231	0.1576	**0.2002**
NormMs	F1@5	0.1044	0.0832	0.0833	0.1055	**0.1209**
	F1@10	0.1617	0.1297	0.1302	0.1654	**0.2019**
	MRR@10	0.0416	0.0341	0.0342	0.0417	**0.0466**
	MAP@10	0.2968	0.2211	0.2218	0.2993	**0.4501**
	nDCG@10	0.1633	0.1285	0.1289	0.1656	**0.1959**
All	F1@5	0.0938	0.0795	0.0789	0.1007	**0.1202**
	F1@10	0.1497	0.1276	0.1274	0.1611	**0.2027**
	MRR@10	0.0361	0.0311	0.0308	0.0383	**0.0445**
	MAP@10	0.2758	0.2279	0.2258	0.2945	**0.3632**
	nDCG@10	0.1503	0.1267	0.1259	0.1616	**0.1980**

Table 4.15: Top-k recommendations evaluation metrics for groups of different mainstreaminess and several recommendation methods. The best performing methods are written in bold. KNN clearly outperforms all competing methods. Interestingly, only TOP yields major differences between LowMs and NormMs.

Contrary to the results for rating prediction, KNN clearly outperforms all other methods. Furthermore, a noteable difference in recommendation quality for LowMs and NormMs can only be observed for TOP. We partially attribute this to the suboptimal experimental setup, since the random split in train- and testset is suited for rating prediction and not for top-k recommendations. Time dependent splits may be favored [Kowald et al., 2019] in this regard. Anyway, results indicate that the recommendation of popular items does perform bad on users of unorthodox taste.

4.4 Discussion

Our observations provide insights into the understudied group of users (i.e., LowMs) that prefer to listen to non-mainstream, non-popular music. In the following, we discuss our key findings.

In a first step, we analyzed music listened by LowMs in order to identify different music styles. Hence, dimensionality reduction and subsequent clustering of tracks was performed. This posed several problems. As our empirical investigation of dimensionality reduction and clustering approaches showed, most state-of-the-art methods yield unsatisfying results. We attribute this to the nontrivial notion of similarity of tracks in their acoustic features. To tackle this problem, dimensionality reduction utilizing Riemannian Geometry and hierarchic, density-based clustering were chosen as best-working methods. Eventually, we found four clusters of tracks, each representing a distinct style of music in terms of acoustic features. Furthermore, we explained each style by high-level genres and thus, were able to concisely describe different music styles and back results in [Mulder et al., 2007] solely with properties of music.

We then defined personas, which are intended to serve as exempar for users that prefer a certain music style. The latter users are found via modelling the influence of music styles in terms of listening events and subsequently, choosing the most promising style resp. persona for each user. For this matter, we found that artificially degrading the influence of large, dominant music styles is necessary.

Several interesting observations can be made about the listening behavior of these user groups. We showed that Festival and Heavy Listeners tend to focus on their preferred music style, whereas Complex and Relax Listeners exhibit a higher degree of omnivorousness and hence, tend to also explore music of other styles. Furthermore, Complex and Relax Listeners listen to a broader body of distinct artists and genres. Despite the observation that Heavy Listeners rather focus on metal music, they surprisingly exhibit the highest level of disparity in terms of music taste among all user groups. I.e., users classified as Heavy Listeners deviate in their music taste from each other. Interestingly, Complex Listeners tend to listen to the same set of tracks over and over again. It remains an open question, whether this is induced by the scarcity of complex music or free choice. Furthermore, we did not expect listeners of complex music to consume less music per day than other types of users.

Additionally, we analyzed user groups based on their demography and culture. Noteably, users of very specialized music styles, i.e., complex and relaxing music, tend to be older than users preferring other styles of music. Furthermore, Relax Listeners show a high level of future orientation. Heavy Listeners believe that power is distributed unequally within society and also asses wealth and financial stability as being important. Festival and Heavy Listeners perceive the presence of corruption in their homecountry higher than Complex and Relax Listeners. Please note that this negatively correlates with the average age of user groups.

Experiments were conducted in which we apply state-of-the-art recommendation algorithms on users of low (LowMs) and normal (NormMs) mainstreaminess. We provide significant evidence that LowMs is disadvantaged with respect to recommendation quality. In this regard, we verified recent work in [Schedl and Bauer, 2017]. This previous work is extended by evaluating recommendations on the aforementioned user groups of different preferred music style. By rigid statistical analyses, we present strong evidence that the recommendation quality varies between user groups. Thus, we illustrate the need for not only utilizing the notion of mainstreaminess, but also non-popular music styles within LowMs, in order to improve recommendation quality for users of unorthodox taste.

Chapter 5

Conclusions and Future Work

In this work, we conducted an in-depth analysis of users of unorthodox music taste, often referred to as non-mainstream users. Based on acoustic properties of tracks, four distinct music styles have been identified within music listened by the aforementioned kind of users. To model users of different taste, we further split non-mainstream users into four user groups, where each group is associated with its favorite music style. We provide strong evidence that user groups show differences in terms of listening behavior and demographics. In particular, user groups exhibit different levels of omnivorousness in regard to music styles. Furthermore, observations hint a relationship between music taste of non-mainstream users and culture. We verified the results of previous research, in the sense that state-of-the-art recommendation algorithms significantly advantage mainstream users over users of unorthodox taste. Additionally, it has been shown that recommendation quality varies with respect to non-popular music styles. Hence, we demonstrated that the problem of providing adequate recommendations for non-mainstream users can only be solved by considering both, the notion of mainstreaminess and music styles in the long tail.

5.1 Research Questions

In this chapter, we recap the research questions guiding the analyses conducted within the course of this work and additionally outline our proposed anwers to the research questions precisely depicted in the beginning of this thesis in section 1.1.

How can non-mainstream music styles be identified and concisely described? By conducting multiple experiments we empirically found that dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP) serves as well-suited preprocessing step for a subsequent identification of similar tracks. Collections of similar tracks were obtained by utilizing Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) as clustering method applied on a selection of tracks' acoustic features. The four identified track clusters show obvious deviations in their low-level descriptions (i.e., acoustic features), where each cluster represents a certain style of music. More interestingly, investigating the genres annotated to each cluster's tracks, we obtained a concise and distinct description of music styles.

RQ2: Which user groups of non-mainstream music exist? We measured the contribution of music styles towards users by exploiting the number of a user's listening events linked to tracks within each track cluster. By mitigating the blurring induced by excessively large track clusters, we could assign each user to exactly one music style. This music style serves as description or exemplar for the preferred music taste of assigned users. Hence, we identified four user groups, which we described as Complex, Festival, Relax and Metal Listeners.

RQ3: How does the music consumption of non-mainstream users deviate from each other? Analyzing the distribution of listening events over music styles for each user group sheds light on their taste profiles. Here, we identified user groups that solely listen to music of their preferred style. Others exhibit a broader taste, as they listen to multiple music styles. Furthermore, we analyzed user groups regarding their heterogeneity in terms of listened genres and observed user groups that have indeed much more diverse taste than others.

RQ4: How do user groups listening to non-mainstream music differ in terms of culture and demography? Several interesting correlations could be found between demography and music taste. For instance, Complex and Relax Listeners are in general of higher age than Festival and Metal Listeners. Also, male-dominated user groups were identified. Hofstede's dimensions and attributes in the World Happiness Report indicate that, e.g., Relax Listeners are future oriented and have the ability to

make life choices by themselves. Interestingly, we provided evidence that Complex Listeners despise unorthodox beliefs and ideas.

5.2 Self-Assessment

We identify several limitations of our work. First of all, this thesis solely relies on the LFM-1b dataset. Furthermore, LFM-1b comprises only a sampled subset of users. As we noted in chapter 3, the distribution of users is not equivalent to the real world. In particular, this dataset is biased towards males, a few dominant countries and users in their twenties. Secondly, only selecting users who's number of listening events is within some range, clearly excludes users that may exhibit unorthodox taste, but deviate from others in terms of the amount of listening events. Thirdly, many qualitative analyses lack to be backed by quantitative results and quantitative findings are mostly based on very simple metrics. Hence, more precise statistical analyses are needed. Finally, we draw a sample from NormMs in order to make the computation of recommendations feasible. Even though the mainstreaminess distribution of the sample is equivalent to the distribution of NormMs, personalized recommendations may lack in validity.

5.3 Future Work

Our future research will be driven by the goal of developing more sophisticated methods to identify different types of users. Since observations are partially based on qualitative analyses and quantitative analyses often lack in significance, we strive to consolidate findings. We furthermore aim to take a closer look at more sophisticated algorithms such as Matrix Factorization, which is already widely used for incorporating additional contextual information into the model Here, our findings about user groups or knowledge about the track clusters could be included, for the matter of providing fair recommendations. Furthermore, considering the large body of research from the field of Music Psychology seems to be a promising future direction.

Reproducibility. This work is based on the LFM-1b[1] and Cultural LFM-1b dataset. To foster reproducibility and future research in the area of Fair Music Recommendations, implementations are freely available in our GitHub repository[2].

Acknowledgements. Additionally to Elisabeth Lex and Dominik Kowald, I would also like to thank Markus Schedl and Christine Bauer from JKU Linz and Eva Zangerle from the University of Innsbruck for providing the datasets and for guiding me into the right directions. This work is funded by the Know-Center.

[1] http://www.cp.jku.at/datasets/LFM-1b/
[2] https://github.com/pmuellner/FairMusicRecommendations

Bibliography

[Abdollahpouri et al., 2019a] Abdollahpouri, H., Burke, R., and Mobasher, B. (2019a). Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555*.

[Abdollahpouri et al., 2019b] Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. (2019b). The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

[Adomavicius and Tuzhilin, 2011] Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer.

[Alstott et al., 2014] Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777.

[Ankerst et al., 1999] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, volume 28, pages 49–60. ACM.

[Antenucci et al., 2018] Antenucci, S., Boglio, S., Chioso, E., Dervishaj, E., Kang, S., Scarlatti, T., and Dacrema, M. F. (2018). Artist-driven layering and user's behaviour impact on recommendations in a playlist continuation scenario. In *Proceedings of the ACM Recommender Systems Challenge 2018*, page 4. ACM.

[Atkinson, 2011] Atkinson, W. (2011). The context and genesis of musical tastes: Omnivorousness debunked, bourdieu buttressed. *Poetics*, 39(3):169–186.

[Aucouturier, 2009] Aucouturier, J.-J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. *Language, evolution and the brain*, pages 35–64.

[Baig et al., 2018] Baig, M. H., Varghese, J. R., and Wang, Z. (2018). Musicmapp: A deep learning based solution for music exploration and visual interaction. In *ACM Multimedia*, pages 1253–1255.

[Baltrunas et al., 2011] Baltrunas, L., Ludwig, B., and Ricci, F. (2011). Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 301–304. ACM.

[Baskerville, 2003] Baskerville, R. F. (2003). Hofstede never studied culture. *Accounting, organizations and society*, 28(1):1–14.

[Bauer and Schedl, 2019] Bauer, C. and Schedl, M. (2019). Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PloS one*, 14(6):e0217389.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[Bobadilla et al., 2013] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-based systems*, 46:109–132.

[Böhmer et al., 2010] Böhmer, M., Bauer, G., and Krüger, A. (2010). Exploring the design space of context-aware recommender systems that suggest mobile applications. In *2nd Workshop on Context-Aware Recommender Systems*, volume 5.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370.

[Cano et al., 2005] Cano, P., Koppenberger, M., and Wack, N. (2005). Content-based music audio recommendation. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 211–212. ACM.

[Cantador et al., 2010] Cantador, I., Bellogín, A., and Vallet, D. (2010). Content-based recommendation in social tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 237–240. ACM.

[Casey et al., 2008] Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

[Celma, 2010] Celma, Ò. (2010). The long tail in recommender systems. In *Music Recommendation and Discovery*, pages 87–107. Springer.

[Celma et al., 2006] Celma, O., Herrera, P., and Serra, X. (2006). Bridging the music semantic gap.

[Chen et al., 2014] Chen, C., Zheng, X., Wang, Y., Hong, F., and Lin, Z. (2014). Context-aware collaborative topic regression with social matrix factorization for recommender systems. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[Cheng and Shen, 2014] Cheng, Z. and Shen, J. (2014). Just-for-me: an adaptive personalization system for location-aware social music recommendation. In *Proceedings of international conference on multimedia retrieval*, page 185. ACM.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

[Dacrema et al., 2019] Dacrema, M. F., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109. ACM.

[Delsing et al., 2008] Delsing, M. J., Ter Bogt, T. F., Engels, R. C., and Meeus, W. H. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality: Published for the European Association of Personality Psychology*, 22(2):109–130.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[Fisher, 1922] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A,*

Containing Papers of a Mathematical or Physical Character, 222(594-604):309–368.

[Freedman et al., 2007] Freedman, D., Pisani, R., and Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

[Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

[Gunawardana and Shani, 2015] Gunawardana, A. and Shani, G. (2015). Evaluating recommender systems. In *Recommender systems handbook*, pages 265–308. Springer.

[Haruna et al., 2017] Haruna, K., Akmar Ismail, M., Suhendroyono, S., Damiasih, D., Pierewan, A. C., Chiroma, H., and Herawan, T. (2017). Context-aware recommender system: a review of recent developmental process and future research direction. *Applied Sciences*, 7(12):1211.

[Hofstede, 2011] Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

[Hofstede et al., 2005] Hofstede, G., Hofstede, G. J., and Minkov, M. (2005). *Cultures and organizations: Software of the mind*, volume 2. Citeseer.

[Huang and Wu, 2016] Huang, A. and Wu, R. (2016). Deep learning for music. *arXiv preprint arXiv:1606.04930*.

[Hug, 2017] Hug, N. (2017). Surprise, a Python library for recommender systems. `http://surpriselib.com`.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Jin et al., 2018] Jin, Y., Tintarev, N., and Verbert, K. (2018). Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 13–21. ACM.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[Jones, 2007] Jones, M. L. (2007). Hofstede-culturally questionable?

[Karatzoglou et al., 2010] Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. (2010). Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86. ACM.

[Kendall, 1948] Kendall, M. G. (1948). Rank correlation methods.

[Kim et al., 2016] Kim, D., Park, C., Oh, J., Lee, S., and Yu, H. (2016). Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 233–240. ACM.

[Koren et al., 2009] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.

[Koren and Sill, 2013] Koren, Y. and Sill, J. (2013). Collaborative filtering on ordinal user feedback. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[Kowald et al., 2019] Kowald, D., Lex, E., and Schedl, M. (2019). Modeling artist preferences of users with different music consumption patterns for fair music recommendations. *arXiv preprint arXiv:1907.09781*.

[Kruskal, 1964] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129.

[Lee et al., 2010] Lee, D., Park, S. E., Kahng, M., Lee, S., and Lee, S.-g. (2010). Exploiting contextual information from event logs for personalized recommendation. In *Computer and Information Science 2010*, pages 121–139. Springer.

[Lee et al., 2018] Lee, J., Lee, K., Park, J., Park, J., and Nam, J. (2018). Deep content-user embedding model for music recommendation. *arXiv preprint arXiv:1807.06786*.

[Levy and Sandler, 2008] Levy, M. and Sandler, M. (2008). Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37(2):137–150.

[Lops et al., 2011] Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[McInnes et al., 2017] McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[Meehan et al., 2013] Meehan, K., Lunney, T., Curran, K., and McCaughey, A. (2013). Context-aware intelligent recommendation system for tourism. In *2013 IEEE international conference on pervasive computing and communications workshops (PERCOM workshops)*, pages 328–331. IEEE.

[Moore et al., 2012] Moore, J. L., Chen, S., Joachims, T., and Turnbull, D. (2012). Learning to embed songs and tags for playlist prediction. In *ISMIR*, volume 12, pages 349–354.

[Mulder et al., 2007] Mulder, J., Ter Bogt, T., Raaijmakers, Q., and Vollebergh, W. (2007). Music taste groups and problem behavior. *Journal of youth and adolescence*, 36(3):313–324.

[Murtagh and Legendre, 2014] Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295.

[Ng et al., 2002] Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856.

[Park et al., 2006] Park, H.-S., Yoo, J.-O., and Cho, S.-B. (2006). A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In *International conference on Fuzzy systems and knowledge discovery*, pages 970–979. Springer.

[Pichl et al., 2015] Pichl, M., Zangerle, E., and Specht, G. (2015). Towards a context-aware music recommendation approach: What is hidden in the playlist name? In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1360–1365. IEEE.

[Pohle et al., 2006] Pohle, T., Knees, P., Schedl, M., and Widmer, G. (2006). Automatically adapting the structure of audio similarity spaces. In *Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS)*, pages 66–75.

[Ramirez-Garcia and García-Valdez, 2014] Ramirez-Garcia, X. and García-Valdez, M. (2014). Post-filtering for a restaurant context-aware recommender system. In *Recent Advances on Hybrid Approaches for Designing Intelligent Systems*, pages 695–707. Springer.

[Reynolds, 2015] Reynolds, D. (2015). Gaussian mixture models. *Encyclopedia of biometrics*, pages 827–832.

[Rocchio, 1971] Rocchio, J. (1971). Relevance feedback in information retrieval. *The Smart retrieval system-experiments in automatic document processing*, pages 313–323.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

[Sachdeva et al., 2018] Sachdeva, N., Gupta, K., and Pudi, V. (2018). Attentive neural architecture incorporating song features for music recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 417–421. ACM.

[Sachs et al., 2018] Sachs, J. D., Layard, R., Helliwell, J. F., et al. (2018). World happiness report 2018. Technical report.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Schedl, 2016] Schedl, M. (2016). The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 103–110. ACM.

[Schedl, 2019] Schedl, M. (2019). Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics*, 5:44.

[Schedl and Bauer, 2017] Schedl, M. and Bauer, C. (2017). Distance-and rank-based music mainstreaminess measurement. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 364–367. ACM.

[Schedl and Hauger, 2015] Schedl, M. and Hauger, D. (2015). Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 947–950. ACM.

[Schedl et al., 2015] Schedl, M., Knees, P., McFee, B., Bogdanov, D., and Kaminskas, M. (2015). Music recommender systems. In *Recommender systems handbook*, pages 453–492. Springer.

[Schedl et al., 2018] Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., and Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116.

[Schmidt et al., 1999] Schmidt, A., Beigl, M., and Gellersen, H.-W. (1999). There is more to context than location. *Computers & Graphics*, 23(6):893–901.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Shardanand and Maes, 1995] Shardanand, U. and Maes, P. (1995). Social information filtering: algorithms for automating" word of mouth". In *Chi*, volume 95, pages 210–217. Citeseer.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107.

[Shin et al., 2009] Shin, D., Lee, J.-w., Yeon, J., and Lee, S.-g. (2009). Context-aware recommendation by aggregating user context. In *2009 IEEE Conference on Commerce and Enterprise Computing*, pages 423–430. IEEE.

[Spivak, 2009a] Spivak, D. I. (2009a). Higher-dimensional models of networks. *arXiv preprint arXiv:0909.4314*.

[Spivak, 2009b] Spivak, D. I. (2009b). Metric realization of fuzzy simplicial sets. *Preprint*.

[Steck, 2018] Steck, H. (2018). Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 154–162. ACM.

[Su and Khoshgoftaar, 2009] Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009.

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

[Thorndike, 1953] Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482.

[Typke et al., 2005] Typke, R., Wiering, F., and Veltkamp, R. C. (2005). A survey of music information retrieval systems. In *Proc. 6th International Conference on Music Information Retrieval*, pages 153–160. Queen Mary, University of London.

[Van den Oord et al., 2013] Van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651.

[Van Setten et al., 2004] Van Setten, M., Pokraev, S., and Koolwaaij, J. (2004). Context-aware recommendations in the mobile tourist application compass.

In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 235–244. Springer.

[Verbert et al., 2012] Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., and Duval, E. (2012). Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335.

[Wang et al., 2013] Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6.

[Willmott and Matsuura, 2005] Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82.

[Woerndl et al., 2007] Woerndl, W., Schueller, C., and Wojtech, R. (2007). A hybrid recommender system for context-aware recommendations of mobile applications. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 871–878. IEEE.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

[Zamani et al., 2018] Zamani, H., Schedl, M., Lamere, P., and Chen, C.-W. (2018). An analysis of approaches taken in the acm recsys challenge 2018 for automatic music playlist continuation. *arXiv preprint arXiv:1810.01520*.

[Zangerle and Pichl, 2018] Zangerle, E. and Pichl, M. (2018). Content-based user models: Modeling the many faces of musical preference.

[Zangerle et al., 2018] Zangerle, E., Pichl, M., and Schedl, M. (2018). Culture-aware music recommendation. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 357–358. ACM.

[Zheleva et al., 2010] Zheleva, E., Guiver, J., Mendes Rodrigues, E., and Milić-Frayling, N. (2010). Statistical models of music-listening sessions in social media. In *Proceedings of the 19th international conference on World wide web*, pages 1019–1028. ACM.

[Zheng et al., 2019] Zheng, H.-T., Chen, J.-Y., Liang, N., Sangaiah, A. K., Jiang, Y., and Zhao, C.-Z. (2019). A deep temporal neural music recommendation model utilizing music and user metadata. *Applied Sciences*, 9(4):703.