Context-based User Playlist Analysis for Music Recommendation

Master Thesis, 25 ECTS
Human Computer Interaction

Author: Joey Moes (6258255)

Project supervisor: Eelco herder
Second supervisor: Almila Akdag

October 4th, 2023

Abstract

Convenient access to music through streaming platforms has given rise to an insurmountable amount of choice when it comes to listening to music. These platforms have turned to music recommender systems to keep the user engaged by giving personalized recommendations. In recent years these algorithms have made great strides and seen huge improvement. However, these music recommender systems can enforce certain biases and cause a lack of diversity within their recommendations. Research has focused on countering these problems with the use of context-dependent recommender systems. Interestingly, there has been a lack of focus on activity based music listening behavior. This study uses different analysis methods to research the correlation between user activity context and musical preferences. Results show that there are significant differences between different activities and the musical features that are contained within a song. Thereby suggesting a use for activity context within music recommender systems. Contrastingly, results from the clustering, classification and the user survey show that it remains difficult to determine which songs are listened to in which contexts of activity. On top of showing that musical taste can not solely be determined by activity, these results show that musical preference remains distinctly subjective and recommendation algorithms will forever struggle in determining the right music for the right person at the right time. Concluding, while activity context shows promise in being useful in recommending music and helping overcome biases and lack of diversity within recommendations, an activity based method should be combined with other algorithms such as content based recommenders. Thereby helping to adhere to users' broad and expansive musical preferences while ensuring relevant and personal recommendations.

1 TABLE OF CONTENTS

2 INTRODUCTION

Listening to music has a rich history dating back to ancient times (Minnix, 2016), and its benefits encompass increased communion, focus, and reduced stress (Črnčec, 2006; Lehmberg, 2010). The emergence of recorded music in 1877 (Burgess, 2014) brought about convenient access to favorite music without the presence of live artists. Over time, recorded music has evolved through various formats, including records, cassettes, CDs, and MP3 files, culminating in the current era of streamable music. Streaming platforms like Spotify, Apple Music, and SoundCloud have established their own platforms to retain user engagement, necessitating a vast library of available music.

As music availability expands, the choice of what to listen to becomes increasingly vast. Previously, individuals could only listen to one album at a time (Whitman, n.d.), but modern music streaming platforms allow for unlimited combinations in users' listening habits. While users can curate personalized playlists from familiar songs, there is an evident desire to discover new music (Garcia-Gathright, 2018). Consequently, streaming platforms have turned to music recommendation software as a means of sustaining user engagement.

Music recommendation systems hold a prominent position in computer science and musicology, employing algorithms to suggest songs or playlists based on users' past listening behavior, similar users, demographic data, and other relevant factors. These systems aim to suggest music that users are likely to enjoy, thereby maintaining their engagement with the platform. The importance of music recommendation systems has grown significantly due to the exponential growth of online music streaming services, which have made vast collections of songs accessible to listeners worldwide. By utilizing music recommendation algorithms, these services aid users in discovering new music that aligns with their preferences, ultimately enhancing the overall user experience and engagement (Garcia-Gathright, 2018).

While the field of music recommendation systems witnesses continuous efforts to enhance accuracy, it is notable that many studies in the related works section seem to focus on novel ideas without a clear foundation rooted in user behavior studies. They often opt for a top-down approach where a suggestion is implemented which could improve accuracy of a recommender system, without first basing the motivation on a predetermined analysis. Popular recommender systems can also perpetuate several biases and can inherently be a cause for non-diverse music recommendations.

Understanding the factors that influence music consumption is paramount when it comes to music recommendation. Cultural background, age, gender, and even emotional state (Chamorro-Premuzic, 2012; Hargreaves, 1995; Gurpinar, 2012) can significantly shape music preferences and consumption patterns. For example, an individual's mood can impact their music preferences, with upbeat music being favored during moments of happiness. Cultural background also plays a role, as individuals may have a propensity to listen to music from their own culture, country, or political background (Fox, 1974). Incorporating these factors into music recommendation systems can enhance their accuracy by accounting for contextual information when making personalized recommendations.

Therefore, recently much emphasis has been placed on context based recommender systems, contexts such as demographics, time of day, or emotional state. However, some of these systems still categorize users as having one particular sort of musical preference, while most people have a large range of genres that they listen to, thereby limiting the diversity of recommendations. Being able to understand context of listening would require an analysis of what users might listen to when during certain activities, since overall, listening to music is a secondary activity, meaning it is mostly done during a different main activity. There is a need to bridge this gap by understanding if people listen to different kinds of music during different activity contexts and if there are similarities between people in terms of these contexts.

An understanding of user behavior during different contexts of activity in music listening can help counter biases in music consumption. These biases refer to a user's inclination towards specific types of music or artists, sometimes due to the platform's recommender system, potentially limiting the diversity of recommendations. For instance, users may exhibit a bias toward a particular genre, reinforced by modern recommender systems, resulting in recommendations that align closely with their listening history, thereby restricting exposure to new and diverse music (Ospítia-Medina, 2022). By not adhering to these biases of a person and instead relying on the context of a situation, music recommendation systems can provide personalized recommendations that are both tailored and diverse, thereby enriching the overall music-listening experience.

In conclusion, delving into user behavior in music listening holds the potential to enhance the accuracy and effectiveness of music recommendation systems. Analyzing patterns of music consumption during certain contexts and addressing biases contribute to the development of personalized and diverse recommendations. This research seeks to explore these aspects further, with the ultimate aim of enhancing the user experience and broadening the horizons of music discovery. We will do so by answering the question of "What insights can be gained from analyzing playlist names in understanding the different kinds of music that users listen to during diverse activities?".

In the following section the related works and the reasons for this research will be motivated. These will be summarized and expanded on further in section 4, where we will delve deeper into the research question. Section 5 and 6 will show the methodology and results, after which we will be discussing what the future might hold for music recommender systems.

3 RELATED WORK

3.1 USER BEHAVIOR IN MUSIC LISTENING

Music is an integral part of our lives, and it has a profound impact on our emotions and behavior. With the advancement in technology, music has become easily accessible, and people can listen to their favorite songs anytime and anywhere. The widespread use of smartphones and music streaming platforms has revolutionized the way people listen to music. Music listening behavior and preference is influenced by a multitude of factors. We will get into what these factors are in the following section.

3.1.1 Personality

Most people would assume that each person is unique and there music taste is based on their own personality. Some have explored the relationship between music preferences and personality (Rentfrow, 2003), using a survey with 3.500 participants. During the survey, the participants were asked to rate their preference for 25 different genres of music and perform a personality test as well. This personality test was based on the Big Five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. They found evidence for three distinct dimensions of music preference: Reflective and complex, intense and rebellious, and finally Upbeat and Conventional. Each of these dimensions coincided with certain personality traits. For example, Reflective and Complex music was preferred by people that had high levels of Openness and low levels of Conscientiousness and Extraversion, whereas Upbeat and Conventional music preferences were associated with high levels of Extraversion and low levels of Openness.

Later research revealed that instead of these three dimensions, a total of 5 dimensions of music preferences could be found (Rentfrow, 2011). These 5 dimensions are as follows: Mellow (smooth and relaxing), Unpretentious (country and singer-songwriter), Sophisticated (complex, intelligent and inspiring), Intense (loud, forceful and energetic) and Contemporary (rhythmic and percussive music). Just like in the previous study, they found that these dimensions of music coincided with personality traits in predictable ways. For instance, the Mellow dimension positively correlated with Agreeableness and negatively correlated with Neuroticism.

There has been evidence of personality affecting the use we have for music as well (Chamorro-Premuzic, 2007). By using over 500 participants, they found that people with specific personality traits, tend to use music in different ways. People with high levels of extraversion seemed more likely to use music as a means of social bonding. Those who used music to regulate their mood were more likely to have high levels of neuroticism.

3.1.2 Age and gender

Others have shown that personality traits, such as emotional intelligence, neuroticism, extraversion, and openness were not good indicators to use when predicting music consumption (Chamorro-Premuzic, 2012). Instead, they found that age and use of music to be most useful when predicting music consumption.

To be precise, they found that people of different ages used music in different ways. For instance, younger people seemed to primarily use music for mood regulation as well as social bonding. In contrast, older people seemed to use music mainly to relax or for intellectual stimulation. Not only that, but people that use music for mood regulation listen to music more than people who don't, those that use it for intellectual stimulation usually preferred a wider variety of music genres.

An older study already found differences in age and gender in regards to music preference. With a sample of 381 students, they had the subjects make a questionnaire regarding their favorite musical styles, performers, as well as questions about their musical background and training. Results showed that younger students were more likely to prefer pop and rock music, while older students usually

preferred classical and jazz music. Boys seemed to be more into rock music than girls, since girls tended to like popular music more than their counterparts.

They also found that musical training and background played a role in shaping musical preferences. Students who had received formal musical training were more likely to prefer classical music than those who had not received training. However, the effect of training on musical preferences was not as strong as the effects of age and gender.

3.1.3 Demographic differences

Although age and gender can also be seen as a demographic, there have been some studies suggesting that culture, country of origin (Schedl, 2021), or even political orientation can have an influence on musical preference.

A research investigation conducted by Fox (1974) examined the association between political orientation and music preferences among a sample of 730 college students enrolled at a prominent university in the southeastern region of the United States. The participants were administered a comprehensive survey encompassing inquiries about their preferred music genres and artists, as well as their political beliefs and affiliations.

The study's outcomes unveiled a significant association between the political orientation of college students and their music preferences. It was observed that liberal or left-leaning students displayed a heightened affinity for rock music, gravitating towards artists like Bob Dylan. Conversely, conservative or right-leaning students exhibited a stronger inclination towards country and western music, favoring artists such as Elvis Presley. Remarkably, the study indicated that political orientation emerged as a more robust predictor of music preferences compared to other demographic factors, including gender, race, and socioeconomic status.

Others have shown that preferences for western music can be categorized into five factors that are consistent across 53 countries around the world (Greenberg, 2022). This would suggest that there are fundamental dimensions underlying in musical preference which are shared across diverse populations. The same study found similar results as previously discussed regarding personality using the Big Five personality traits in all the different countries.

An important thing to note is that there is does not seem to be any research that is not geared toward western music and culture. Therefore it is impossible to say with a certainty that the results we have discussed would be similar with non-western music.

3.1.4 Mood

The intricate relationship between mood and music poses a complex subject of inquiry in the realm of user behavior. Several scholarly papers have delved into the correlation between mood and music listening habits, uncovering intriguing insights. In an illuminating study conducted by Gurpinar (2012), it was observed that individuals in positive emotional moods displayed a heightened inclination towards genres like pop, rock, and electronic music, whereas those experiencing negative emotional states were more drawn to classical and jazz music.

In contrast, Friedman (2012) unearthed contrasting findings, indicating that individuals in a negative mood exhibited a pronounced aversion to happy songs, perceiving them as discordant with their emotional state. However, their findings did not uncover a distinct inclination towards listening to sad songs when individuals were in a negative mood.

Notably, contrasting evidence has emerged in more recent studies, such as the work by Xue (2018), suggesting that individuals in a happy state may indeed demonstrate an aversion to sad music, whereas individuals in a sad state may exhibit a greater inclination towards listening to sad songs. These findings challenge the notion of recommending music solely based on mood as an inherently consistent approach.

Moreover, studies have also explored the reciprocal influence of music on mood. Campbell (2021) conducted research indicating that listening to positive music caused an increase in positive affect, while exposure to negative music corresponded to a decrease in positive affect and an upsurge in negative affect. Listening to neutral music, on the other hand, yielded no significant impact on mood or affect.

Collectively, these studies offer intriguing and sometimes conflicting insights into the intricate interplay between mood and music, emphasizing the need for nuanced considerations when developing music recommendation methods solely based on mood.

3.1.5 Social Influence

Many of us have seen the classic examples of social influence and music. Such as the forming of subcultures based on the music style. Think of punks, hippies, metalheads, etc. (Gelder, 2005). An important question to ask is, do these subcultures come together based on their music taste or did they form and adopt a music taste afterwards? The essays in Gelder's book (2005) do not provide a definitive answer to this.

However, research has shown that social influence can have an effect on music taste in different manners. Most of the influence on music taste comes from parents, peers, or media (Thompson, 2014).

The first way that these sources can influence a person's musical taste is through social comparison. As people tend to compare their own music taste to others in order to establish some sense of social identity and belonging. People are more likely to listen to music that their peers listen to in order to fit in with the group.

Conformity is another of these mechanisms through which a person's taste can be influenced. While similar to the last example, conformity seems to be rooted in uncertainty and insecurity. People want to feel included and are afraid of social disapproval or criticism, therefore they adopt a music preference that fits in with their peers.

Thirdly, identification, parents can have a some influence on their children's preferences. Since parents expose children to music that coincides with their own personal preferences. Children may adopt those preferences as a way of expressing their loyalty and admiration for their parents. This can either be conscious or unconscious but can shape a child's musical tastes for years to come.

A fourth and final mechanism that has an influence on someone's music taste is media. Radio and television seem to have a significant impact on the music that is being listened to. Media outlets shape the public's opinion as well, since they can provide positive or negative coverage of certain music or artists, this in turn affects people's perception of those songs or artists and can (re)shape their preferences.

3.2 Different Approaches to Music Recommendation Systems

This section will go over the many methods that are used to when recommending music. Lots of these methods are not specifically made for music recommendation but can be used for any recommendation system. Starting with naming the most popular methods and their problems, then moving on to more niche methods that try to improve upon the more basic ones. This section is meant to provide a basic understanding of the different recommender systems, problems they might face and biases they might create.

3.2.1 Collaborative filtering approach

Recommender systems use different kinds of recommending methods. The most popular being a collaborative filtering (CF) method (Su, 2009, Mathadil, 2017). This method compares the user to other users that listen to the same songs or artists, and then suggest new music based on the other songs that one user listens to and the other does not yet listen to. These CF methods therefore need to compare users with each other. Sometimes users are linked together based on location, such that popular music in a certain country will also be recommended to users from the same country or region. However, most times users are linked together based on the music they each listen to. If two users listen to the same songs, then the songs that only one of them listens to will be recommended to the other person that doesn't yet listen to that song.

There are several problems that are caused by using a collaborative filtering method. The main one being that of echo chambers and filter bubbles (Pariser, 2011). Due to the influence that similar users have on each other, none of them are likely to find novel music or artist due to this method. When one user only listens to pop, and is linked to other users that only listen to pop, it will be unlikely that the system suggests anything other than pop. This will lead to lower user satisfaction because users desire novel and diverse experiences (Anderson, 2020). Due to the CF method, popular music is also heavily enforced. Since music will be recommended more that is already popular. CF methods therefore do not create access to the long tail of music in streaming services that use this method. This makes new and niche artists less likely to become popular and it encourages lots of marketing from music labels since music that is listened to often will also be recommended more.

However, recommender systems that use CF approaches are still very common and do lead to high accuracies. Exactly how CF approaches work can vary from system to system and efforts are being made to keep improving the CF approach. Research (Sánchez-Moreno, 2016) has shown that when a logarithmic transformation is applied to play counts and normalizing the playing coefficients for each user, these coefficients can be used to measure the similarity between users and between artists. In turn, those similarity values are used to recommend new music to the user. Methods such as these are being used to ensure that CF approaches can still maintain an above baseline accuracy. However, this does not solve the problems that the CF methods face as we have mentioned. Research should focus on tackling these problems while maintaining a high level of accuracy.

A method that improves upon basic CF approaches is one that directly tackles the problem of low user satisfaction that is caused by a lack of diversity that CF approaches are known for. Xing and colleagues (2014) use the million song dataset to improve upon the basic CF approach. They do so by balancing recommending popular songs (exploration) and recommending less popular but potentially interesting songs (exploitation). A modified version of Thompson Sampling algorithm is used to calculate the expected reward, which is based on the item's popularity and the similarity between users who have listened to the item. The modified algorithm also calculates the uncertainty, which is based on the variance of the similarity values. Results show that this modified version of the CF approach outperforms basic approaches in the intended categories, namely recommendation diversity and user satisfaction.

3.2.2 Content based approach

A different recommending method is called the content based (CB) method. This method does not compare users, instead it analyzes components of the music itself and recommends music to the user that is similar to what the user has already listened to (Pazzani, 2007). This is mostly done with information such as genre, artist, album etc. However, content based approaches can also use more advanced content such as pitch, tempo or rhythm. It would be easy to write an entire paper about the intricacies of content-based approaches and the benefits of using this method, it is more important for this study that the downsides of these approaches are discussed as well.

When using more complex variables of a musical piece, it can become hard to discern whether the similarities of two songs actually coincide with the users musical preference. Since there are many options to consider when deciding what features are most important. One person might only like specific songs because they were performed by their favorite artist, whilst other might not care at all by whom the songs they listen to were performed. They might instead be more inclined to listen to a certain genre or some other abstract aspect of the musical piece.

Similarly to CF based approaches, this approach can also result in echo chambers, a lack of diversity, and this method also struggles with the cold-start problem. Users only receive recommendations that are more of what they are already listening to, there is no way to diversify recommended content by strictly using this method. The cold-start problem arises when the user has not listened to enough songs to let the method run its calculations.

A lot of CB methods also lack the awareness of why users listen to the specific music that they listen to, reasons for listening to a song can range from the artist, genre, tempo, the lyrics, etc. Recommending a song based on artist can result in very different kinds of songs. Moreover, people tend to listen to a lot of different artists and genres, and their musical interest may vary widely, therefore these content based predictions already face some initial difficulty.

As mentioned, both the CF and the CB methods result in a low diversity of content. A low diversity of content has been shown to be detrimental in user satisfaction (Anderson, 2020). Ensuring diversity in music recommendations correlates strongly with user conversion and retention. “Generalist users are much more likely to remain on Spotify than specialist users.” (Anderson, 2020)

Since most companies use a combination of CF and CB approaches, this might explain why so many people still mainly have other means of finding new music. Although this research is outdated, in a study by Tepper (2009), results show that people mostly find new music through peer recommendations and their own social network.

3.2.3 Hybrid approaches

There are hybrid approaches that combine CF and CB methods as well (McFee, 2011). In the study by McFee (2011), they optimized content-based audio similarity by learning from a sample of collaborative data. Therefore, “recommendations can be made where no collaborative filter data is available”. Since collaborative filters cannot directly form recommendations without the items being ranked or consumed by users, this new method provides a simple solution to the cold-start problem that plagues collaborative filtering approaches.

Another hybrid approach which tackles some of the problems that CF and CB methods face (Yoshii, 2006) uses a Bayesian network called an aspect model. This method introduces latent variables, which are statistically estimated, in turn these represent unobservable user preferences. Effectively being able to simultaneously consider user ratings as well as content similarity. Thereby tackling the CF problem of the cold-start and the CB problem of content-similarity not directly reflecting user preference.

Whilst hybrid approaches can more accurately give recommendations, it does not solve all of the problems that the original approaches struggle with. Mainly the lack of diversity remains an issue.

3.2.4 Context based approaches

Efforts have been made to develop more advanced recommendation methods, with the most prominent being the context-based approach. Context-based approaches encompass various methods. For instance, Schedl and colleagues (2021) found that considering the user's country of residence leads to better recommendations compared to relying solely on popularity-based approaches.

Another context-based recommendation method involves analyzing the user's sentiments from their online social media posts and suggesting songs based on that (Rosa, 2015). However, such solutions are flawed and raise privacy concerns, even more so than normal collaborative filtering (CF) approaches (Polat, 2003).

In a study conducted by Moscato and colleagues (2020), a context-based recommendation method was developed, taking into account personality and mood. This research is still in its early stages, as it currently only crudely accounts for personality based on the "big five" traits. Nevertheless, it shows promising results. We have seen in the previous section that personality can be a big influence for the type of music that people listen to. Context based methods like this try to make use of those more abstract influences of music listening behavior.

However, these methods also raise questions about how these traits would be measured. Initially, users would have to specify their personality type, and the program would require a way to gauge the user's mood. A personality-based method would result in a system similar to collaborative filtering. Yet, basing recommendations on mood might be counterintuitive, as users would need to continually update the system on their mood. We have also seen that music can influence mood as well. Although there might be other methods of doing this without direct user involvement (LiKamWa, 2013), it is challenging to conceive of a non-invasive way to measure the user's current mood.

In conclusion, a method must be developed that is independent of other users to avoid filter bubbles and echo chambers (Pariser, 2011). This method should enable users to discover new music while still maintaining sufficient personalization to ensure user satisfaction (Garcia, 2018). As a context-based approach, it should be non-invasive and minimize the burden on the user.

People's musical preferences cannot be easily categorized into a single genre (Greasley, 2006). Individuals listen to different types of music at different times of the day or when performing different tasks. A potential solution that incorporates this phenomenon is a temporal context-based method (Herrera, 2010). People tend to have routines where they engage in certain tasks at specific times, such as driving to work, having dinner, walking the dog, etc. Therefore, adhering to the times at which users listen to certain types of music would cause diverse but relevant music recommendations.

Herrera and colleagues (2010) found evidence of temporal patterns in music listening, indicating that "for certain users, artists, and genres, temporal patterns of listening can be used to predict music listening selections with above-chance accuracy." These findings could be applied to music recommendation and playlist generation to offer music suggestions at the opportune moment.

As evident, there is extensive ongoing research in the field of recommender systems, with each study building upon one another. While improvements have been made, some solutions may not be able to solve some of the foundational issues that coincide with the method of user categorization that is inherent in most recommender systems, therefore not battling the biases and non-diversity that they create.

3.2.5 Biases of the recommender systems

Music recommender systems, like other machine learning models, are susceptible to biases originating from the data, algorithms, and relevant features used for recommendation. Collaborative filtering methods, which rely on other users' data to suggest music, introduce a bias towards popular music known as the "popularity bias." (Kowald, 2020) As popular songs are recommended more frequently, this bias reinforces their popularity, benefiting already popular artists and well-funded labels. Consequently, users may receive a limited range of music and their individual preferences may be overlooked.

Collaborative filtering methods also suffer from "user-item bias" as they tend to recommend items similar to what a user has previously liked, potentially limiting the scope of recommendations (Koren, 2021). For instance, if a user has only listened to one genre, collaborative filtering methods may neglect to suggest music from other genres that might be of interest to the user.

Content-based methods recommend music items based on their similarity to those previously liked by a user, focusing on features like genre, tempo, or mood. However, content-based methods can exhibit biases such as "feature bias," where certain features are overemphasized at the expense of other relevant features important to the user's preferences. Additionally, "genre bias" may lead the model to recommend music solely based on genre, disregarding other influential factors like mood or tempo.

Hybrid methods combining collaborative filtering and content-based approaches aim to provide more accurate recommendations but can also be subject to the same biases. Additionally, the "cold-start" problem arises when there is insufficient data to make accurate recommendations for new users or items, leading to a bias towards recommending popular items.

Context-based methods for music recommendation utilize the user's present context, encompassing factors such as location, time of day, or weather. However, these methods are not immune to biases. One notable bias is the "location bias," which occurs when context-based recommendations prioritize music items popular in a specific geographic region without taking into account the user's unique preferences. Another bias is the "temporal bias," whereby recommendations focus on music items popular during a particular time of day, disregarding the user's individual preferences and current mood.

Moreover, biases in music recommender systems can arise from the lack of diversity in the training data. If the data primarily represents a specific genre or culture, the model may struggle to make accurate recommendations for users with different preferences or from diverse cultural backgrounds.

In conclusion, music recommender systems exhibit various biases depending on the method, data, and features used for training. Awareness of these biases is crucial in the design and evaluation of music recommender systems. Steps should be taken to mitigate biases, such as incorporating more diverse training data or utilizing advanced algorithms.

3.2.6 Where to go from here?

Previous sections of this paper have discussed various types of music recommender systems and the factors that influence music listening behavior, including age, gender, demographics, personality, and social influence. These systems have shown promising results in improving music discovery and user satisfaction.

Many of the recommender systems that are being researched nowadays seem to have two main goals: increase the accuracy of predictions and maintain or improve user satisfaction. Each new method produced may cause better recommendation accuracy than baseline, however, they all seem to be a shot in the dark, trying a new method and hoping it succeeds. While most methods produce better than baseline results, it is hard to see where research should focus on.

Moreover, many of these methods do not directly tackle the problems of older recommender systems such as a lack of diversity, tackling biases and adhering to peoples' broad range of musical preferences. Even context based recommender systems do not seem immune to these problems.

Except for the collaborative filtering and content-based methods, most of the recommender systems that we have discussed need a lot of active information to work, such as mood, geographical data, etc. Asking this information from the user is quite invasive. There is a growing need for non-invasive methods to analyze music listening behavior without relying on explicit user input. Such methods could help understand how people consume music in different contexts, how their listening behavior changes over time, and how it is influenced by external factors such as cultural trends and events.

As we have seen in the previous section, temporal information may be a good variable to keep in mind when recommending music. Depending on the time of day, a user might listen to different types of music (Herrerra, 2010). This method is non-invasive as well. Another study found that the kind of activity that is performed during music listening has some influence on the decision of a classifier of what genre or artist a person wants to listen to, they combined a lot of different features to achieve a high accuracy in prediction. Their system performed well when recommending genres and artists, with a 60% and a 55% percent accuracy respectively (Gillhofer & Schedl, 2015). However, suggesting a specific songs based on their features did not seem to be fruitful at all with 1.5% accuracy. In this study, no emphasis was placed on activity context in terms of their analysis, it was just a small part of the overall context in their research, which only relied on classifiers.

The fact that most studies do not reach amazingly high accuracy on recommendations goes to show that people's tastes are unpredictable. If a person likes a certain artist or genre, is it enough to recommend another song based solely on those variables? It is important to study the similarities and differences between songs, artists, and genres and find out how some of these factors can appeal to a person while others do not.

4 INVESTIGATING USER MUSIC PREFERENCE THROUGH PLAYLIST ANALYSIS

In the previous sections we have gone over the extensive research that has been conducted to develop novel approaches which aim to enhance the accuracy of music recommendations. While these approaches have shown improvements over baseline methods, a notable limitation lies in their limited understanding of how users' preferences may vary across different activities. This knowledge gap calls for a deeper exploration of user preferences and the contextual factors influencing their music choices.

We build upon the work of previous researchers who have contributed to the understanding of why people choose specific music and how these choices vary across diverse contexts. However, our approach specifically focuses on investigating the intricate relationship between activities or contexts and music preferences. This includes aspects such as tempo, vocal versus instrumental compositions, and other musical attributes that play a pivotal role in shaping individuals' musical tastes based on their daily activities.

By leveraging the names of playlists, which often reflect specific activities or moods, it becomes possible to gain valuable insights into the relationship between music and activities, and how they intertwine to shape individuals' listening habits.

In light of these considerations, this research aims to conduct a comprehensive analysis of user preferences by exploring the correlation between playlist names and the corresponding activities during which individuals listen to music. By investigating the association between specific song characteristics and playlist names, we seek to uncover patterns that shed light on the music listening behavior of people using their music for specific purposes during different activities.

By delving into playlists we are able to directly find out in what way users use their music. Specifically, the goal of this research is to find patterns in listening behavior, by studying context dependent patterns. We may confirm or disprove the current assumptions that music recommender system studies adhere to. The playlist names will say something about the use of the music users listen to, but further insights can be gained from the music that they put in their playlists. The analysis will uncover basic and complex information about music listening behavior. For instance, do users listen to different kinds of music during different kinds of activities, or do specific contexts for playlists have distinct musical attribute values that correspond to them?

The primary research question guiding this study is as follows:

What insights can be gained from analyzing playlist names in understanding the different kinds of music that users listen to during diverse activities?

By focusing on playlist names as a window into users' music listening behaviors, we strive to move beyond traditional genre-based or artist-based approaches and provide a more holistic understanding of how individuals engage with music in various contexts. Specifically, we will be focusing on the on the interplay between context of activity and music preferences such as happy/sad, fast/slow, instrumental or vocal etc.

This research aims to fill the existing gap in knowledge regarding user music preferences by adopting a bottom-up approach that investigates the relationship between playlist names and activities associated with music consumption. By gaining insights into the activities during which people listen to specific genres or artists, we can contribute to the development of more context-aware and user-centric music recommender systems.

5 METHODS

5.1 DATA SOURCES

We have decided to use an existing dataset that provides user data, such as the songs that a user has listened to or the playlist in which these songs are located (Larxel). The data from this dataset has been expanded with the music data from the Spotify API[1] to receive more information about the songs themselves. The data from the Larxel dataset provides information which makes it possible to see in what context people listen to what kinds of songs. We have done so with the use of keywords, analyzing the playlist names, for instance if a playlist is called running or workout it will be categorized as such. Then, we will analyze the songs that users put into a playlist made for such specific purposes.

To perform the analysis, data will be collected mainly from Spotify with the use of the Spotify API. With the API we can find out a lot of information about songs, artists and genres. Information such as pitch, tempo and rhythm, but also key, danceability, speechiness, acousticness etc.

5.2 SAMPLE SELECTION

The Larxel[2] dataset does not provide a lot of information about the user and the circumstances of their listening behavior. However, it has provided a large sample size of 15.918 unique users, 290.001 artists, 2.036.734 tracks and 161.529 playlists. Those were the only columns in the original Larxel dataset: “user ID”, “songname”, “artistname” and “playlistname”. The dataset provides a mean of 810 listened-to tracks per user. By using the information available from the Spotify API we can add large amounts of information about the song attributes to the existing Larxel dataset.

5.3 DATA CLEANING

Before the start of the analysis the data was cleaned. Ensuring that we don’t waste time during the gathering of information from the Spotify API. Firstly, all the rows were removed that contained partial information. Secondly, keywords were used to find the playlist that gave an indication of context. The following contexts were chosen at first: ‘studying’, ‘vacation’, ‘running/sports’, ‘partying’, ‘relaxing’, ‘cleaning’, ‘gaming’ and ‘romance’. Examples of keywords are, for sports: “run”, “running”, “jogging”, “training” etc., or for relaxing: “peace”, “calm”, “sleep” etc. These keywords were also translated into English, Spanish, French, German, Chinese, Japanese, Korean, Hindi and Bengali. A new column was made and the perceived purpose of each playlist was assigned and added to that column. All playlists that did not contain any of the chosen keywords, were removed, as well as the songs that were added to the same playlist twice. When viewing the distribution of playlists among the different contexts that were initially chosen it was clear that the amount of ‘cleaning’, ‘gaming’ and ‘romance’ playlist contexts were negligible, Therefore those categories were removed and ‘studying’, ‘vacation’, ‘relaxing’, ‘running/sports’ and ‘partying’ contexts were the ones left over for the analysis.

After cleaning of the data, we were left with only 35.557 songs in 538 playlists of 453 users. A lot less than the initial amount of data that the Larxel dataset provided. Mostly due to the fact that users tend not to name their playlists after the activity they use it for. Most playlists were called after an artist or genre, or had their own special and unique name. While the data that we can work with is a lot less than the initial data, there is enough information available to conduct the study. The following information for each song was collected using the Spotify API: Acousticness, Danceability, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, Valence and Popularity. These are

[1] https://developer.spotify.com/documentation/web-api
[2] https://www.kaggle.com/datasets/andrewmvd/spotify-playlists?select=spotify_dataset.csv

characteristics that are unique in every song, and therefore they allow for a deep analysis of different song types and their correlation to the context categories in which they are being listened to.

5.4 STATISTICAL ANALYSIS

We will be conducting several different types of analysis throughout this study to answer the research question: “What insights can be gained from analyzing playlist names in understanding the diverse activities in which individuals listen to different kinds of music?” In this subsection we will go over each of these analyses.

Frequency Distribution: The frequency distribution of playlist categories will be calculated to determine the prevalence of each activity type among the collected playlists. This analysis will provide a comprehensive view of users' preferences for different activities. By examining the distribution, it will be possible to identify the most and least popular activity types, offering a snapshot of the overall playlist landscape. This information will help our understanding of the general trends and patterns in music choices, enabling tailoring recommendations and assess the popularity of different activity categories among users.

Diversity analysis: Our diversity analysis primarily focuses on evaluating the diversity within playlists, emphasizing both the composition of tracks and the distribution of artists within each playlist category. To quantify these aspects, we employed two key diversity metrics: Shannon Entropy and Gini Index.

Shannon entropy was utilized to measure the uncertainty or randomness in the track composition of each playlist type. A higher Shannon entropy value indicates a greater degree of diversity in song selection within a given playlist category. This metric enables us to assess the variety and randomness in the songs chosen for different purposes.

The Gini index was employed to assess the inequality in the distribution of tracks and artists within each playlist type. A lower Gini index value suggests a higher level of diversity in song and artist selection.

Correlation Analysis: A correlation analysis will be performed to explore potential relationships between playlist categories and different features of a song. The features that we will be analyzing are: Acousticness, Danceability, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, Valence and Popularity. Which, as previously mentioned, were gathered with the use of the Spotify API They will show us whether or not people listen to different music depending on the activity context. The objective of this analysis is to uncover significant correlations that can enrich our understanding of music listening behavior. By examining the connections between playlist categories and these variables, we may gain insights into users' preferences and tendencies. For example, we might discover that playlists labeled as "workout" often contain energetic and fast-paced songs. These associations will provide deeper insights into the nuanced relationships between context and musical attributes.

As a first step, it is important to see if any significant difference between the categories of playlist and every variable that we have information on can be found. To test the significance of the difference between each category for each variable, an ANOVA test will be conducted. In addition, a post hoc test, the Tukey test will be used to view the significances in the differences between each category. The results of these tests will be shown in the results section. The results from every song property will be shown on a different page. To not skew the results due to outliers, since there will be many of those, the differences in the median results will be analyzed instead of the mean.

Clustering: To further explore the underlying patterns within the playlist data, a cluster analysis technique, k-means clustering will be employed. The goal of this analysis is to identify groups of songs that exhibit similar characteristics, enabling us to uncover distinct user segments based on their music preferences.

Clustering will be used to find clusters using the properties of each song that have been found using the Spotify API. While these clusters might not initially tell us anything about the activity category of the playlists, interesting findings can be obtained when researching these clusters. The clusters themselves will contain specific songs, these songs will then be cross referenced with the playlist they were found in. This way, it is possible to see if the clusters that were found somehow correlate to the activity category of the playlists or not.

Specifically, k-means clustering methods (Lloyd, 1982) will be used to group users based on the aforementioned variables. K-means clustering is a widely-used algorithm that autonomously divides data into k clusters, assigning each data point to the cluster with the closest mean value. By applying this method, it will be possible to identify natural groupings of songs and gain insights into the preferences of different user segments. In our case, we decided to use 5 clusters, to hopefully align them with the 5 different activity contexts that we are studying.

Classification: In the literature section, a large number of recommender systems have been discussed. In this thesis, we are studying the relationship between the activity that was meant for the playlist and the songs that the playlist contains. To study whether this relationship correlates enough to make accurate predictions about what category of playlist contains which songs, different classification methods will be used. If these classification methods end up with a high accuracy, it can be said that the songs differ enough between each playlist such that it would be easy to determine the playlist which should contain that particular song. A lower accuracy will tell us that there are more factors that determine why a song is put into a context or it might tell us that users are very different from each other when it come to listening to songs for a specific context.

4 different classification methods will be used to test this: Logistic regression, Decision Tree classifier, Random Forest and lastly a SVM Classifier.

5.5 Survey

We decided to perform a user survey based on the results that we gathered from the classification and the clustering process in order to compare how people would perform while classifying songs into different activity contexts themselves.

A total of $N = 63$ participants were recruited using convenient sampling. 12 responses were excluded because of unfilled questions, resulting in $N = 51$. Out of these responses there were 27 men and 24 women. The ages of the participants ranged from 22 to 61 (M_{age}= 26.88, SD = 11.41).

The survey took around 10 minutes to complete. It was created using the online survey platform Qualtrics. Participants were first asked to read the informed consent. The informed consent stated that participation is voluntary, and the research was completely anonymous. All participants were aged 18 years or older. The researchers' email address and telephone number were added in the informed consent, in case of questions about rights as a participant. Participants could complete the survey on any device. Participants could agree or not agree to consent, in case of the latter participants were immediately directed to the end of the experiment.

The survey included 10 questions which included a picture of a Spotify user's playlist, a video with audio where they could listen to the songs in the playlist and the question itself. Each question was asked the same way: "For what purpose do you think this playlist was created?". An example of what a survey question looked like can be found in appendix A. The survey included 2 playlists of each of the categories that we have researched in this study: relaxing, studying, vacation, running/sports and partying. The answers were given using a multiple choice selection with these categories as possible answers. The questions were shown in a random order for each participant, so as to rule out a learning effect that might take place with the later answers.

After each question, the participants were shown whether they had answered correctly or not. They were also shown their eventual score in terms of how many percent of answers they had correct. While scoring high does not necessarily mean they performed better, since that is up to interpretation, this allowed for a more fun experience for the participants, which made participant recruitment easier.

5.6 Ethical Considerations

When analyzing the data, anonymity will be ensured. By translating keywords into other languages it will be ensured that the analysis is not biased towards the English language and the users that name their playlists in English. The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences classified this research as low-risk with no fuller ethics review or privacy assessment required.

6 Results

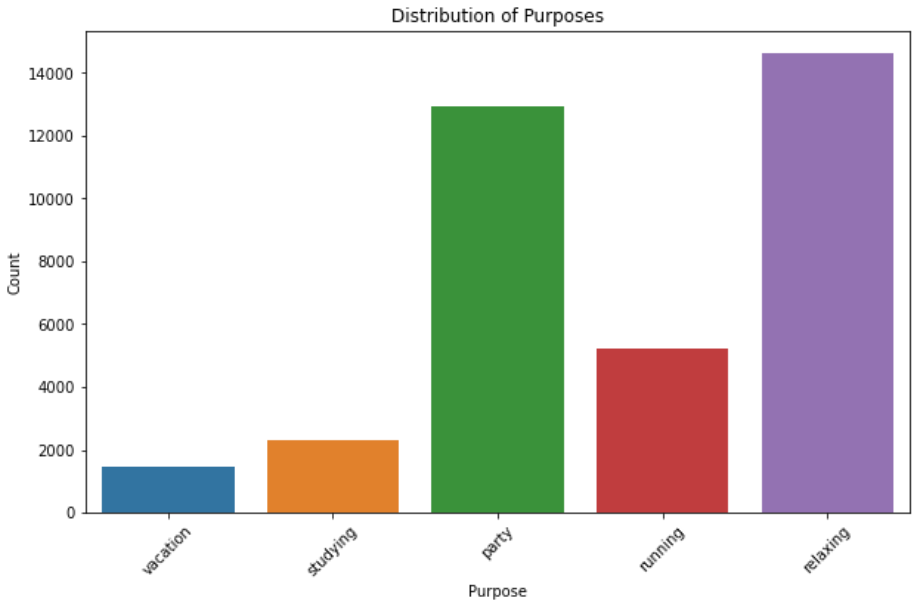


Figure 1: Distribution of the playlist categories in terms of songs

In this section, we will be going over the results gathered from the statistical analysis and the user survey. Beginning with a quick frequency distribution analysis and a diversity analysis. After, we will have a large section showing the correlation analysis where the differences between the playlist contexts and the musical features are analyzed. After, the clustering and classification analyses are discussed. Lastly, the survey and its results will be shown. In this section, the results are mostly only shown and most of their interpretation will be in the discussion section.

6.1 Frequency distribution

Figure 1 shows us the distribution of the playlists. We can see that most people make playlists for partying or for relaxing. While there is not a lot of data for vacation, studying and running playlists relatively, there still seems to be a enough data to gather some information from.

6.2 DIVERSITY ANALYSIS

In this section, we delve into the diversity analysis of playlists, to answer the question that we have alluded to before: how much do users playlists differ even when the playlist has a similar goal? We will be focusing on both track composition and artist distribution within each playlist type. To quantify these aspects, we employ two key diversity metrics: Shannon Entropy and Gini Index. Shannon entropy measures the uncertainty or randomness in a playlist type's track composition. Higher entropy indicates greater diversity. The Gini index measures the inequality in the distribution of tracks within a playlist type. Lower Gini index values suggest higher diversity.

Diversity of Songs:

We begin by examining the diversity of songs within each playlist category, measured by the Gini Index. Lower Gini Index values indicate higher diversity in song selection. Our findings reveal that for the "party," "relaxing," and "running" playlists, Gini Index values are approximately 0.61, 0.57, and 0.58, respectively, signifying moderate inequality in track distribution. In contrast, the "studying" and "vacation" playlists exhibit Gini Index values of approximately 0.52, implying a more even distribution of songs than other playlist categories.

Additionally, we employ Shannon Entropy to quantify the uncertainty or randomness in song composition. Higher entropy values suggest greater diversity. The "party" and "relaxing" playlists exhibit Shannon Entropy values of approximately 13.01 and 13.45, respectively, indicating a relatively high degree of randomness in song selection. Conversely, the "studying" and "vacation" playlists have Shannon Entropy values of approximately 11.04 and 9.32, respectively, suggesting a more structured track composition.

Diversity of Artists:

Turning our attention to the diversity of artists within each playlist category, we observe similar trends. The Gini Index, which measures inequality in artist distribution, reveals that the "party," "relaxing," and "running" playlists exhibit Gini Index values of approximately 0.79, 0.79, and 0.76, respectively, indicating moderate diversity in artist selection. The "studying" and "vacation" playlists display Gini Index values of approximately 0.81 and 0.75.

Shannon Entropy for artists further corroborates these findings. The "party" and "relaxing" playlists have Shannon Entropy values of approximately 10.85 and 10.88, respectively, indicating a moderate level of randomness in artist selection. Meanwhile, the "studying" and "vacation" playlists showcase Shannon Entropy values of approximately 7.77 and 7.32, implying a more structured artist composition. These metrics collectively offer insights into the variety and balance of both songs and artists, enriching our understanding of playlist diversity for various purposes.

6.3 CORRELATION ANALYSIS

In this subsection, the different characteristics of songs will be analyzed. For each of the features, an ANOVA will be conducted together with a post hoc Tukey test to see if there are significant differences between the playlist activity context categories. In the tables on every page, the results of the Tukey test are shown. The right most column shows whether the two activities being compared are significantly different from each other, indicated by ‘True’ or ‘False’.

Acousticness:

Acousticness is a musical attribute that indicates the measure of acoustic or electric components in the song. A high acousticness value suggests that the song has a predominantly acoustic sound, while a low value indicates a more electronic or synthetic sound. The ANOVA and Tukey post-hoc test show that for each of the categories the difference in accousticness is significant.

In figure 2 we can see the clear trends that acousticness has in relation to playlist categories. Partying and running playlists have very low levels of acousticness. With median values of 0.037 and 0.030 respectively. Parties usually have electronic dance music.

Vacation and studying playlists are very similar to each other, though still significantly different, see Table 1. They both have high values of acousticness, indicating a preference for acoustic music during those specific activities. Relaxing playlists strike a balance between the two, the upper quartile value of 0.652 shows that a significant portion of songs in these playlists lean towards acoustic instrumentation. a majority of songs having some acoustic element, though with a median value of 0.209, relaxation playlists seem to tend toward less acoustic music than studying or vacation playlists, but more than party and running playlists.

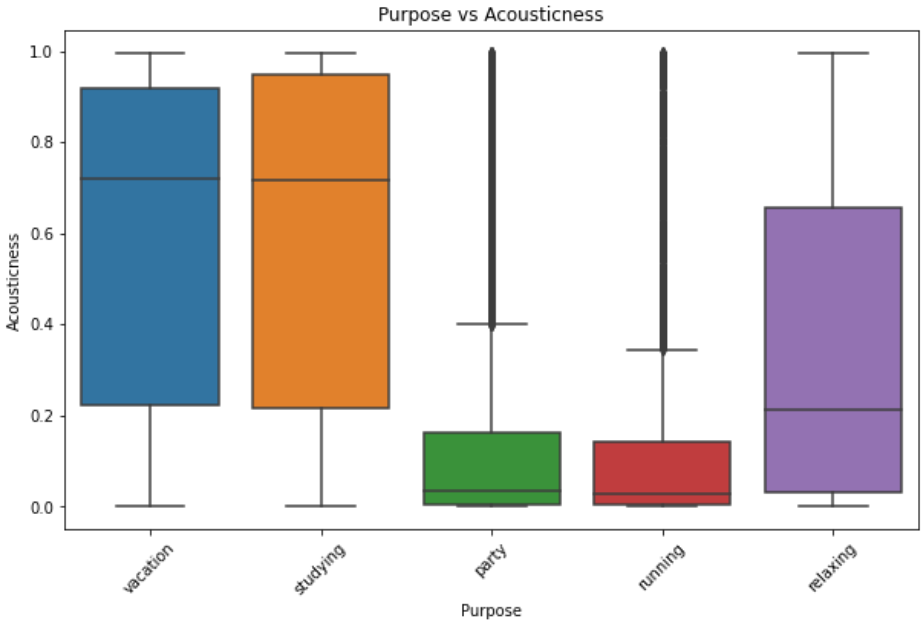


Figure 2: Boxplot of Acousticness for each playlist category

Acousticness: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	0.204	0	0.195	0.214	True
Party	Running	-0.016	0.005	-0.029	-0.003	True
Party	Studying	0.456	0	0.439	0.474	True
Party	Vacation	0.292	0	0.261	0.323	True
Relaxing	Running	-0.220	0	-0.231	-0.208	True
Relaxing	Studying	0.252	0	0.235	0.270	True
Relaxing	Vacation	0.088	0	0.057	0.118	True
Running	Studying	0.473	0	0.453	0.492	True
Running	Vacation	0.308	0	0.276	0.340	True
Studying	Vacation	-0.165	0	-0.199	-0.130	True

Table 1: Tukey test results for Acousticness

Danceability:

Danceability is a musical attribute that quantifies the suitability of a song for dancing based on elements such as tempo, rhythm, and beat strength. A higher danceability value indicates that a song is more likely

to be danceable, while a lower value suggests a less dance-friendly composition. The ANOVA and Tukey post-hoc test, as seen in table 2, show that for each of the categories the difference is significant.

Figure 3 shows the (dis)similarities of each of the context categories in terms of danceability. The highest mean and median value is that of the party playlist which does not surprise much. Many parties have an element of dancing, therefore party playlists have higher danceability values.

In contrast the lowest values of danceability, with a median value of 0.443, are in the studying playlists. This also does not come as a surprise. Users will want their studying playlists to have the least distracting songs in their library in order to make learning more effective. Hearing songs that make you want to dance will not help with the learning process.

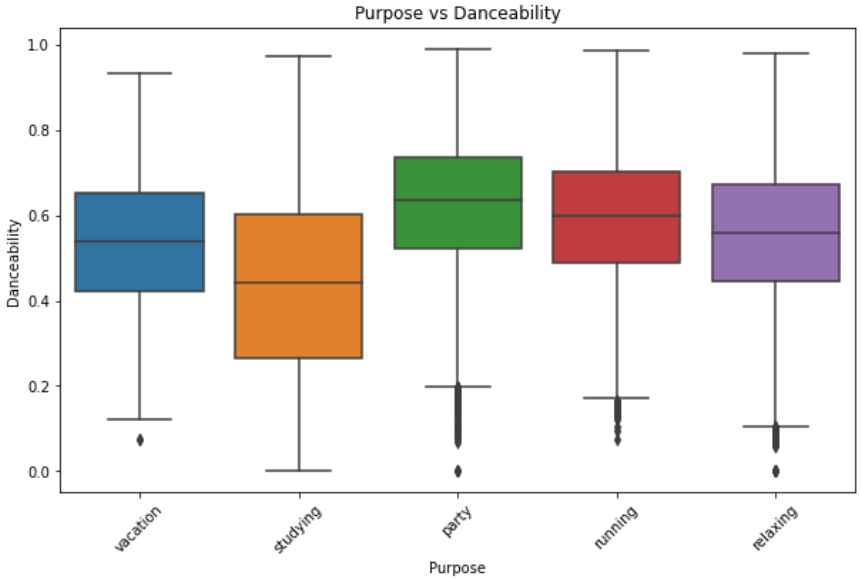


Figure 3: Boxplot of Danceability for each playlist category

Danceability: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.068	0	-0.0732	-0.062	True
Party	Running	-0.031	0	-0.039	-0.024	True
Party	Studying	-0.184	0	-0.194	-0.174	True
Party	Vacation	-0.114	0	-0.132	-0.097	True
Relaxing	Running	0.037	0	0.029	-0.044	True
Relaxing	Studying	-0.116	0	-0.126	-0.107	True
Relaxing	Vacation	-0.047	0	-0.064	-0.029	True
Running	Studying	-0.153	0	-0.164	-0.142	True
Running	Vacation	-0.083	0	-0.102	-0.065	True
Studying	Vacation	0.070	0	0.05	0.089	True

Table 2: Tukey test results for Danceability

Energy:

Energy, in the context of music analysis, refers to the intensity and activity level present in a song. High energy values indicate energetic and fast-paced tracks, while low values suggest more subdued and calm compositions. The ANOVA and Tukey post-hoc test show that for each of the categories the difference is significant.

With energy there is a very similar trend as with acousticness, see figure 4. Vacation and studying playlists show a distinct similarity, as do partying and running playlists, with relaxing playlists standing in the middle. Vacation and studying playlists have median values of 0.302 and 0.307 respectively, though their means still differ significantly, as seen in the Tukey test shown in table 3. Vacation playlist might want to induce a calm atmosphere, to fully being able to rest on you vacation. Studying playlists show that energy levels are preferably low during studying sessions, which might have a similar reasoning as with danceability where a high energy song might be distracting enough to not being able to focus on the learning material.

In contrast running and party playlist have high levels of energy. Though some parties can have a relaxed atmosphere, running as well as parties are active activities and therefore would want to lean into music that gives listeners some more energy or at least are in line with the high energy atmosphere that are expected during these two activities.

The most interesting finding in terms of energy is that relaxing playlists, with a median value of 0.585, do not seem to indicate a high necessity for low energy songs. Instead, the relaxing playlist have a balance of high and low energy songs. The term relaxing semantically indicates that people want to rest during this activity, which might be best done with low energy songs. However, apparently the authors of these playlist seem to value a low energy atmosphere from their music less when it come to playlists that are meant for relaxing.

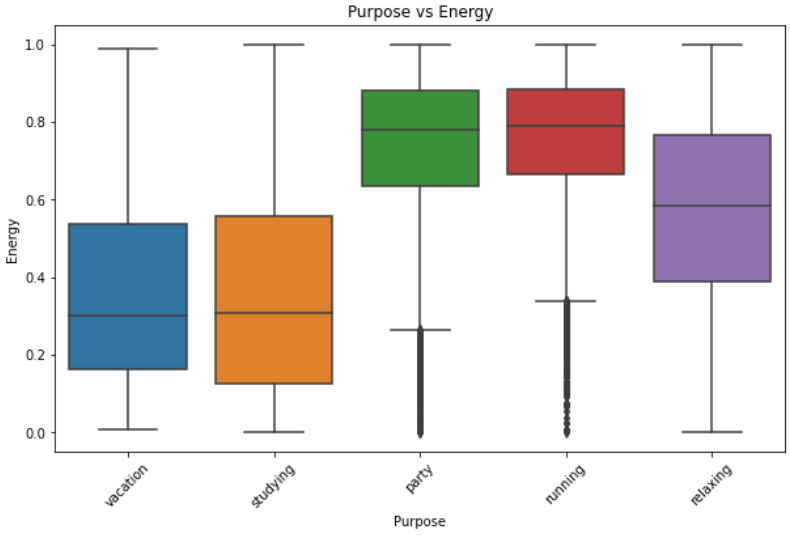


Figure 4: Boxplot of Energy for each playlist category

Energy: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.172	0	-0.179	-0.167	True
Party	Running	0.014	0.001	0.005	0.024	True
Party	Studying	-0.383	0	-0.396	-0.370	True
Party	Vacation	-0.209	0	-0.233	-0.185	True
Relaxing	Running	0.186	0	0.177	0.196	True
Relaxing	Studying	-0.211	0	-0.224	-0.198	True
Relaxing	Vacation	-0.037	0	-0.061	-0.014	True
Running	Studying	-0.397	0	-0.412	-0.382	True
Running	Vacation	-0.223	0	-0.248	-0.199	True
Studying	Vacation	0.174	0	0.148	0.200	True

Table 3: Tukey test results for Energy

Instrumentalness:

Instrumentalness is a musical attribute that measures the presence of vocals in a song. A high instrumentalness value suggests that a song is primarily instrumental (without vocals), while a low value indicates the presence of vocals.

Though the ANOVA test reveals a significant difference in the instrumentalness variable between the playlist categories (P= 0.0), the Tukey test, seen in table 4, shows that instrumentalness shows no significant difference between the running and party categories, the party and vacation categories, and the running and vacation categories. The rest of the categories do have significant differences.

From figure 5 we can see that 4 categories of playlists, exhibit low instrumentalness values. The only exception being studying playlists with a median value of 0.710 which soars above the others. A high intstrumentalness means that those songs have little to no vocals. With studying playlists vocals might stimulate distraction, perhaps by wanting to sing along. While listening to music can be a background activity, listening to someone and understanding the words is a lot harder while also trying to read and study. Therefore, users might want the songs that are in their studying playlist to hardly have any vocals and lyrics.

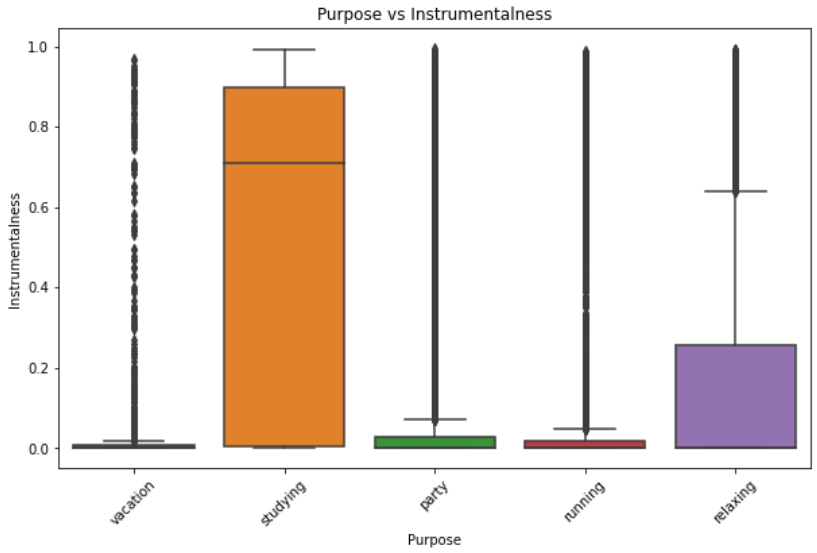


Figure 5: Boxplot of Instrumentalness for each playlist category

Instrumentalness: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	0.077	0	0.067	0.086	True
Party	Running	-0.01	0.733	-0.019	0.01	False
Party	Studying	0.409	0	0.391	0.427	True
Party	Vacation	-0.026	0.171	-0.057	0.006	False
Relaxing	Running	-0.082	0	-0.095	-0.070	True
Relaxing	Studying	0.333	0	0.315	0.350	True
Relaxing	Vacation	-0.102	0	-0.134	-0.071	True
Running	Studying	0.415	0	0.395	0.0435	True
Running	Vacation	-0.020	0.464	-0.052	0.013	False
Studying	Vacation	-0.435	0	-0.470	-0.4	True

Table 4: Tukey test results for Instrumentalness

Popularity:

Popularity, in the context of music analysis, refers to the measure of a song's popularity based on factors like its play count, user interactions, and chart performance. High popularity values indicate widely liked and frequently played songs, while low values suggest lesser-known tracks. The ANOVA and Tukey post-hoc test, seen in table 5, show that for each of the categories the difference is significant.

Interestingly, we can see a similar relation with popularity and the categories as we did with energy and acousticness. Where the studying and vacation playlists are very closely related, as well as the running and partying playlists, with the relaxing playlist category taking the middle ground again.

With popularity, it is harder to guess where these distinct differences come from. A party might want to have popular songs which allow for feelings of nostalgia or sing along moments during a party, inducing a feeling of closeness with the group. While studying playlist will want to reduce the amount of popular songs so as to not be distracted by the recognition of lyrics or melodies. However, this is speculation, since (un)popular songs do not really tell us anything of how well the song is recognized by the user themselves. Other interesting findings are that running seems to have the most popular songs with a median of 54.0, while vacation playlists have the least popular songs with a median of 21.0

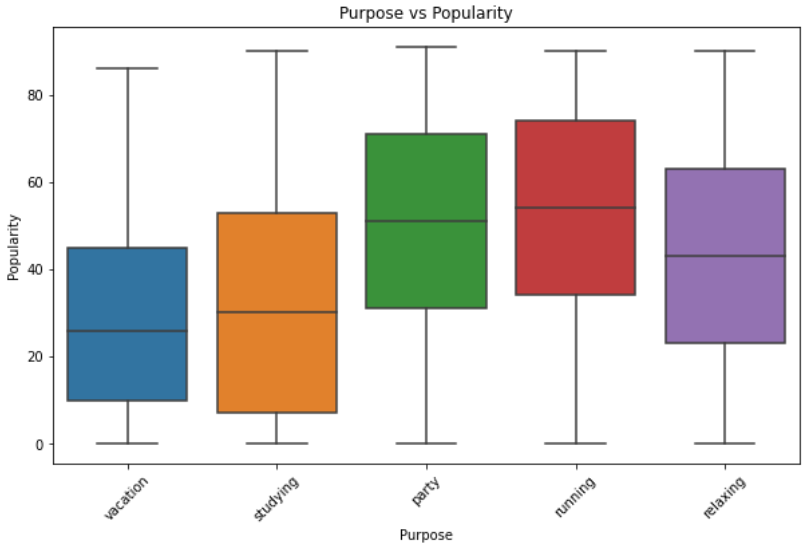


Figure 6: Boxplot of Popularity for each playlist category

Popularity: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-6.635	0	-7.443	-5.827	True
Party	Running	2.148	0	1.048	3.248	True
Party	Studying	-17.502	0	-19.006	-15.999	True
Party	Vacation	-22.410	0	-25.057	-19.763	True
Relaxing	Running	8.783	0	7.702	9.864	True
Relaxing	Studying	-10.867	0	-12.357	-9.377	True
Relaxing	Vacation	-15.775	0	-18.414	-13.136	True
Running	Studying	-19.650	0	-21.316	-17.983	True
Running	Vacation	-24.558	0	-27.300	-21.815	True
Studying	Vacation	-4.910	0	-7.836	-1.980	True

Table 5: Tukey test results for Popularity

Valence:

Valence, in music analysis, represents the musical attribute of emotional positivity conveyed by a song. High valence values indicate positive and uplifting tracks, while low values suggest more negative or subdued emotions. The ANOVA and Tukey post-hoc test, seen in table 6, show that for each of the categories the difference is significant. Let's explore the differences between the categories of playlists based on the provided valence results.

Valence also shows clear differences between all context categories, seen in figure 6. In the related works section, a paper (Gurpinar, 2012, Friedman 2012) was discussed about mood and music. Where studies showed that people are likely to want to listen to music that is congruent with their current mood. Since valence shows a degree of emotional positivity, this relates to that. People that are on vacation, on a party or are relaxing will most likely have a good mood. Therefore, it shows that in these contexts, users are most likely to want to listen to positive songs.

Studying on the other hand is most likely not a “fun” activity to most people. Therefore having very emotionally positive songs would likely be in opposition to their mood while studying.

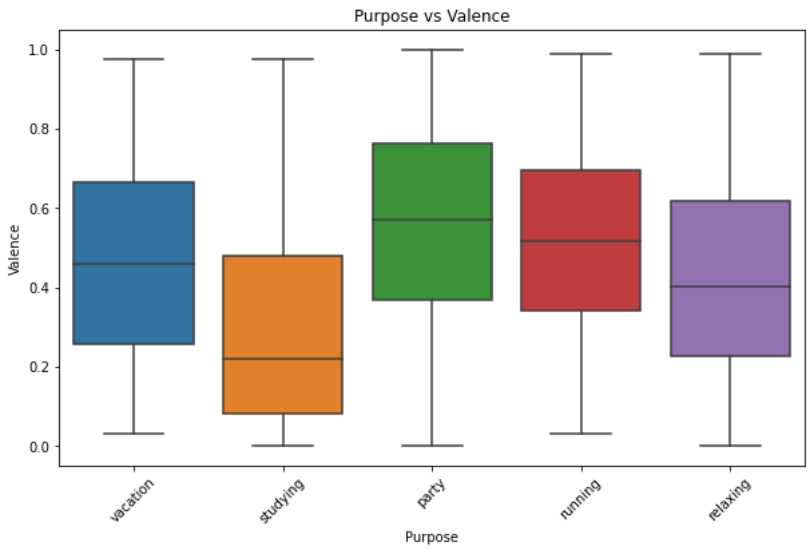


Figure 7: Boxplot of Valence for each playlist category

Valence: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.127	0	-0.135	-0.119	True
Party	Running	-0.045	0	-0.056	-0.034	True
Party	Studying	-0.252	0	-0.268	-0.237	True
Party	Vacation	-0.091	0	-0.117	-0.064	True
Relaxing	Running	0.082	0	0.071	0.093	True
Relaxing	Studying	-0.126	0	-0.141	-0.111	True
Relaxing	Vacation	0.036	0.002	-0.010	0.063	True
Running	Studying	-0.208	0	-0.225	-0.191	True
Running	Vacation	-0.046	0	-0.074	-0.018	True
Studying	Vacation	0.162	0	0.132	0.191	True

Table 6: Tukey test results for Valence

Liveness:

Liveness, in music analysis, refers to the attribute that conveys the perception of a live performance in a song. Higher liveness values suggest that a song sounds more like a live performance, while lower values indicate a more studio-recorded or synthesized sound. While the ANOVA test shows a significant difference (P = 0.0) the Tukey test results show that liveness might be the least accurate in describing the differences between the context categories, since 4 relations do not show significant differences.

The different categories have liveness values that are all incredibly close to each other, the lowest being 0.176 and the highest being 0.206. This is probably due to the fact that songs on Spotify are mostly studio recorded.

In summary, even though the ANOVA test came to a significant difference, liveness might be the least helpful musical feature when differentiating between contexts of song listening behavior. Presumably this is due to the uneven distribution of studio recorded songs and live performances that are on the Spotify platform.

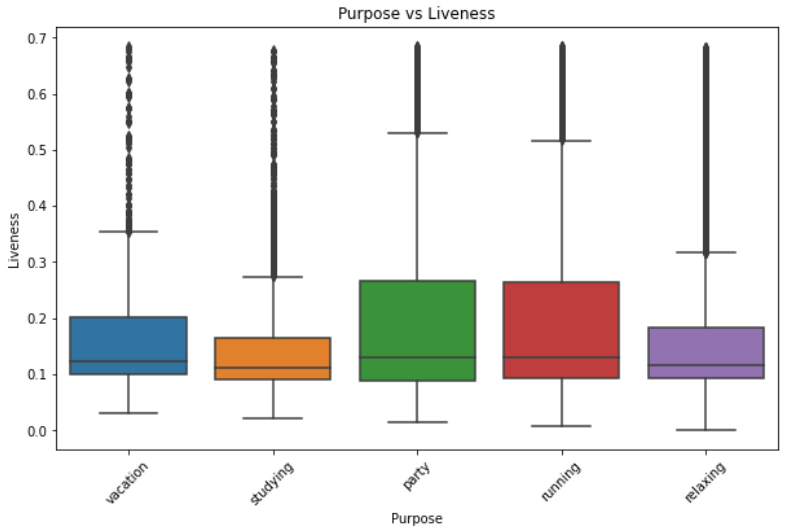


Figure 8: Boxplot of Liveness for each playlist category

Liveness: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.029	0	-0.034	-0.023	True
Party	Running	0.001	0.998	-0.007	0.008	False
Party	Studying	-0.034	0	-0.045	-0.024	True
Party	Vacation	-0.022	0.005	-0.040	-0.005	True
Relaxing	Running	0.030	0	0.022	0.037	True
Relaxing	Studying	-0.006	0.516	-0.016	0.004	False
Relaxing	Vacation	0.006	0.871	-0.011	0.024	False
Running	Studying	-0.035	0	-0.046	-0.024	True
Running	Vacation	-0.023	0.005	-0.047	-0.005	True
Studying	Vacation	0.012	0.453	-0.008	0.032	False

Table 7: Tukey test results for Liveness

Tempo

Tempo, in music analysis, measures the speed or pace of a song, often quantified in beats per minute (BPM). The Tukey test results for the "tempo" variable reveal interesting differences and similarities between the various playlist categories. It shows that all different categories differ significantly from each other in terms of tempo, except for vacation category with the relaxing category and the party category. Which is interesting since those two categories are very most polarized, at least semantically. Vacation playlists seem to be a distinct middle ground of context in terms of tempo.

Interestingly all the tempos of the playlist categories are very close, see figure 9. With the highest being 125.96 BPM for running playlists and 114.02 BPM the lowest tempo for the studying playlists. While the values of tempo are very close for each context, there are still clear differences between the categories in relative terms. Difference enough to warrant the rejection of the null hypothesis between most of the categories.

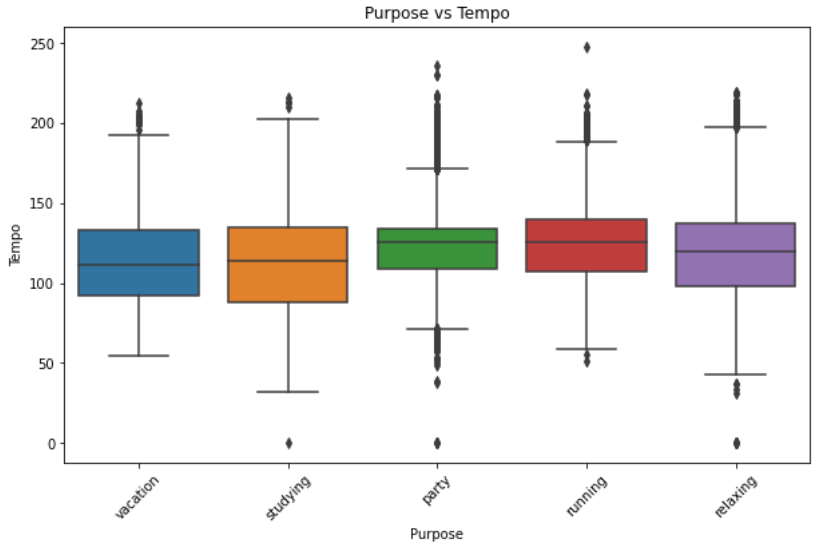


Figure 9: Boxplot of Tempo for each playlist category

Tempo: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-4.233	0	-5.125	-3.341	True
Party	Running	1.953	0	0.738	3.168	True
Party	Studying	-10.003	0	-11.663	-8.342	True
Party	Vacation	-2.789	0.070	-5.711	0.134	False
Relaxing	Running	6.186	0	4.992	7.380	True
Relaxing	Studying	-5.770	0	-7.715	-4.134	True
Relaxing	Vacation	1.445	0.658	-1.470	4.359	False
Running	Studying	-11.955	0	-13.796	-10.115	True
Running	Vacation	-4.741	0	-7.77	-1.713	True
Studying	Vacation	7.214	0	3.981	10.447	True

Table 8: Tukey test results for Tempo

Speechiness

Speechiness, in music analysis, measures the presence of spoken words or vocal elements in a song. A speechiness value above 0.66 typically indicates that a song is primarily composed of spoken words, making it highly speech-oriented. Conversely, lower speechiness values suggest a higher proportion of instrumental or non-vocal elements in the music. Let's explore the differences between the categories of playlists based on the provided speechiness results.

Speechiness, just as tempo and liveness exhibits less clear differences between the categories, as seen in figure 10, at least relative to the other features. We can see that party and running playlists have the highest speechiness values, with a mean of 0.085 and 0.086 respectively. They do not differ significantly from each other, though the reasoning for their higher speechiness values may differ significantly. It would not be far fetched to say that party playlists most likely have higher speechiness values due to wanting to sing along during a party. While users with running playlists might want more spoken words in their songs because they can be a motivating factor while working out.

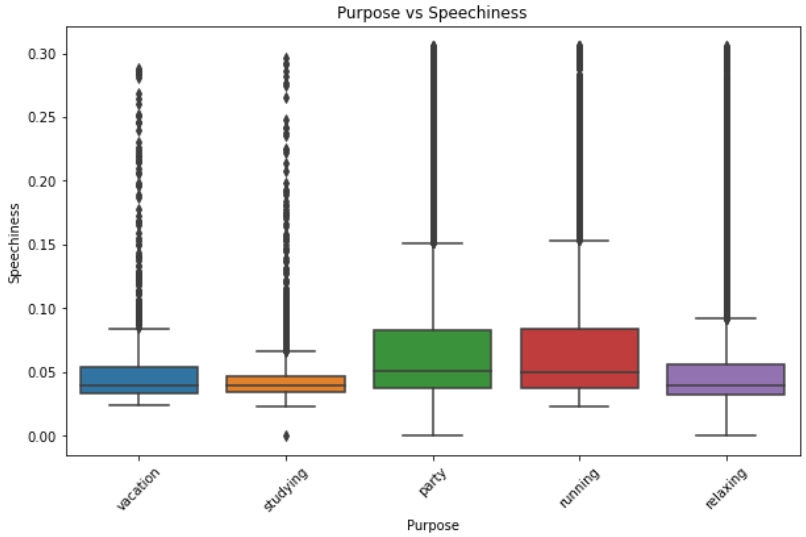


Figure 10: Boxplot of Speechiness for each playlist category

Speechiness: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.023	0	-0.026	-0.021	True
Party	Running	0.001	0.973	-0.003	0.004	False
Party	Studying	-0.034	0	-0.038	-0.029	True
Party	Vacation	-0.034	0	-0.038	-0.022	True
Relaxing	Running	0.024	0	0.020	0.028	True
Relaxing	Studying	-0.011	0	-0.015	-0.006	True
Relaxing	Vacation	-0.007	0.141	-0.015	0.001	False
Running	Studying	-0.034	0	-0.040	-0.029	True
Running	Vacation	-0.031	0	-0.039	-0.022	True
Studying	Vacation	0.004	0.826	-0.006	0.013	False

Table 9: Tukey test results for Speechiness

Loudness

Loudness, in music analysis, represents the attribute related to the volume or amplitude of a song. Higher loudness values suggest louder and more intense music, while lower values indicate softer or more subdued sound. While the ANOVA test yielded a significant difference (P = 0.0), the Tukey test in figure 10 showed that in terms of loudness, the vacation and relaxing playlist categories are not significantly different from each other.

Though party and running playlists seem very similar to each other in the boxplot of figure 11, they are significantly different from each other as shown in table 10. They exhibit the highest values for loudness, possibly due to the louder nature of these activities. The other activities have a more relaxed ambiance or in the case of studying, the necessity for the music to not be on the foreground to concentrate better. This is most likely why the loudness values are slightly lower than those of running and party playlists.

However, overall, loudness also does not look like it has the biggest differences between the playlists, therefore it might not be productive to use loudness as a means of classifying songs for activity context.

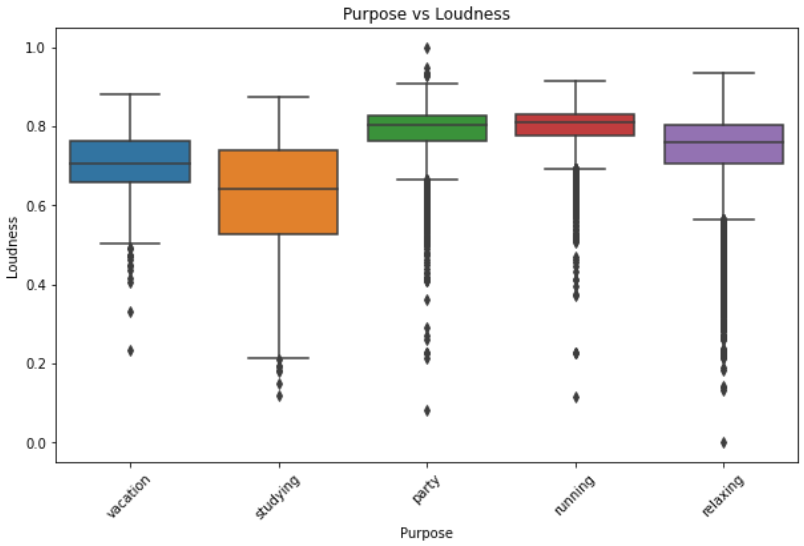


Figure 11: Boxplot of Loudness for each playlist category

Loudness: *Multiple Comparison of Means – Tukey HSD, FWER = 0.05*

Group 1	Group 2	Meandiff	p-adj	Lower	Upper	reject
Party	Relaxing	-0.045	0	-0.047	-0.042	True
Party	Running	0.008	0	0.004	0.011	True
Party	Studying	-0.162	0	-0.167	-0.158	True
Party	Vacation	-0.048	0	-0.056	-0.039	True
Relaxing	Running	0.053	0	0.049	0.056	True
Relaxing	Studying	-0.118	0	-0.122	-0.113	True
Relaxing	Vacation	-0.003	0.883	-0.011	0.006	False
Running	Studying	-0.170	0	-0.176	-0.165	True
Running	Vacation	-0.056	0	-0.064	-0.047	True
Studying	Vacation	0.115	0	0.105	0.123	True

Table 10: Tukey test results for Loudness

Correlation Analysis summary

Our correlation analysis demonstrates that all musical attributes within our current dataset exhibit significant differences across different playlist categories. The results reveal intriguing insights into the relationship between musical attributes and playlist categories.

Firstly, danceability varies significantly across these categories. Party playlists emerge as the most danceable, aligning with the expectation of lively and energetic music. Conversely, studying playlists have the lowest danceability scores, reflecting a preference for less distracting music during study sessions.

Acousticness, which measures the presence of non-electronic instruments, showcases distinct patterns. Party and running playlists lean towards a more electronic or synthetic sound, while vacation and studying playlists prefer acoustic music. This divergence in acousticness suggests that users tailor their music choices to the nature of the activity, seeking high energy beats for parties and tranquil acoustics for relaxation.

Instrumentalness, indicating the absence of vocals in songs, is notably high in studying playlists. This preference for instrumental tracks suggests a desire to minimize distractions while studying.

Popularity in songs differs among playlist categories, with running playlists featuring the most popular songs and vacation playlists having the least. The reasons behind these variations remain speculative.

Valence, representing the emotional tone of songs, correlates with different activities. Vacation, party, and relaxing playlists tend to favor positive, uplifting songs, while studying playlists opt for a less positive tone, likely aligning with the focused and serious nature of studying.

Liveness and loudness exhibit minor variations across categories, possibly due to the prevalence of studio-recorded songs on music platforms.

Tempo varies only slightly among categories, with vacation playlists serving as a midpoint in terms of tempo.

Finally, speechiness is higher in party and running playlists, potentially reflecting the desire to sing along or have the music act a motivational force during workouts.

What this tells us about the contexts of music listening behavior is that there are clear differences between the different activities of people listening to music and the musical attributes that constitutes songs. While some features are very distinct, such as valence, popularity and energy for example. Others are less distinct, such as liveness, tempo, speechiness and loudness. While the latter features show significant differences amongst most of the contexts, it would be best if we continue our investigation leaving these behind.

The following features show a lot more promise in classifying the songs for a playlists based on song features: ‘acousticness’, ‘danceability’, ‘energy’, ‘instrumentalness’, ‘valence’ and popularity’. Therefore we will be using these features moving forward with the clustering and classification methods.

6.4 CLUSTERING

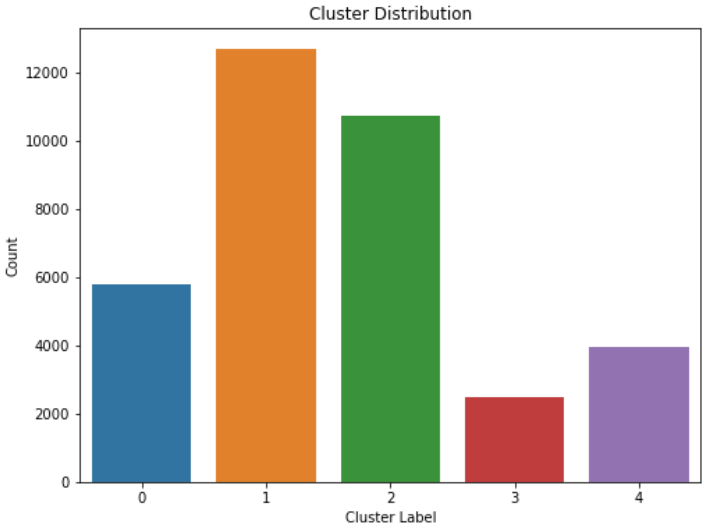


Figure 12: Barplot showing the distribution of the clusters

In this section we will be discussing the results from the clustering process. To keep in line with the playlist contexts, we made 5 different clusters, to find out if these would somehow correlate to the 5 categories of playlists that have been studied in the correlation analysis section. In figure 12 we can see that the distribution is very similar to the playlist distribution that was shown in the methods section, where 2 groups are dominant and the other 3 groups are smaller. However, so far, it is difficult to say what exactly defines these clusters. These clusters are based on the 6 most defining features that we have already discussed, which are: 'acousticness', 'danceability', 'energy', 'instrumentalness', 'valence' and popularity. What the clusters represent is up for interpretation, they could also represent different genres of music, different demographics or different personalities of the user as we have seen in the literature section..

For the clustering we used the 6 most significant variables of the 10. Since the other 4 variables that were shown in the previous subsection showed signs of non-significance in between some playlists. Therefore, 'liveness', 'speechiness', 'tempo' and 'loudness' were not included in the clustering process. This will also make additional analysis of the clusters easier.

In figure 13 you can see the distribution of the variables within each cluster. For example: cluster 3 shows the highest values in the instrumentalness and acousticness variables. So we can assume that cluster 3 contains the least amount of electronic and vocal songs. In figure 14 the distribution of the playlist categories within the different clusters can be seen. Unfortunately, there does not seem to be a clear correlation between the k-means clusters and the playlist categories. Though some categories do seem a lot more prominent in some clusters than in other clusters. For example, cluster 3 contains the most songs that are in the study category. This coincides with our previous correlation analysis of the studying playlists, those largely contained non-electronic songs.

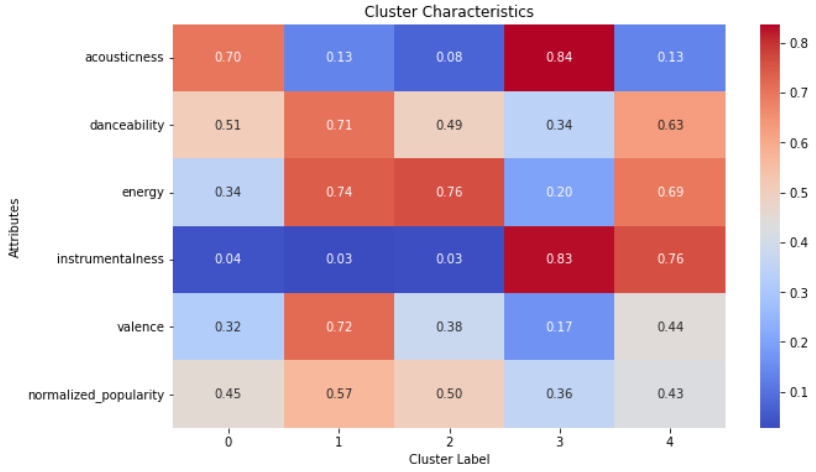


Figure 13: Heatmap of the Features in the clusters

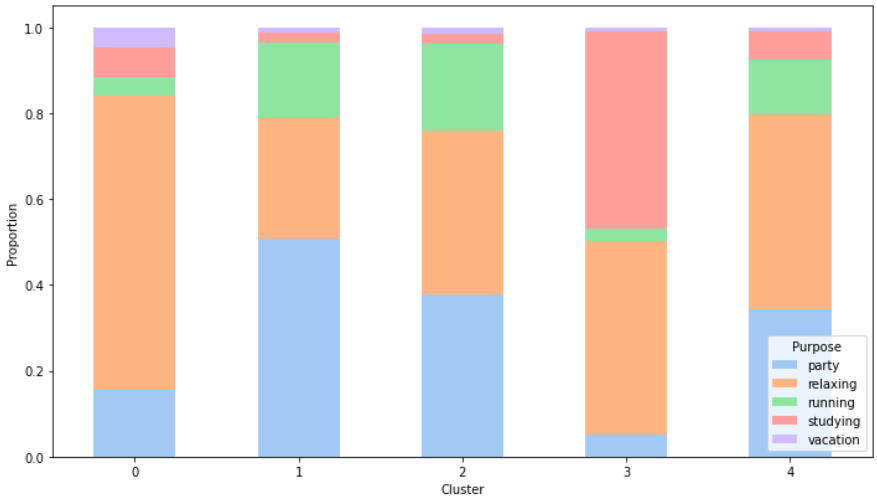


Figure 14: Distribution of Context Categories within different clusters

If not the playlist categories from our users, then what do these clusters represent. Here is a short summary of how the different clusters might be interpreted:

Cluster 0: "Acoustic Bliss"

This cluster earns its name, "Acoustic Bliss," due to its high acousticness (0.70). The songs here rely heavily on acoustic instruments, creating a warm and organic sound. Danceability is moderate (0.51), allowing for a balance between dancing and relaxed listening. Energy levels are relatively low (0.34), setting a mellow and subdued vibe. Instrumentalness is low (0.04), emphasizing the presence of vocals. Valence is moderate (0.32), offering a mix of positive and negative emotions. Popularity is moderate (0.45), making these songs appealing to a broad audience.

Cluster 1: "Dance Floor Anthems"

Named "Dance Floor Anthems," this cluster shows a preference for electronic elements, resulting in low acousticness (0.13). Danceability soars with a high score (0.71), making these songs ideal for energetic dancing. Energy levels are high (0.74), setting an upbeat and lively mood. Instrumentalness is low (0.03), emphasizing vocal prominence. Valence is highly positive (0.72), creating an uplifting and positive emotional tone. Popularity is relatively more significant (0.57), indicating a higher level of recognition.

Cluster 2: "Electronic Vibes"

"Electronic Vibes" leans strongly toward electronic and non-acoustic sounds, with a low acousticness value (0.08). Danceability is moderate (0.49), offering versatility for both dancing and relaxed listening. Energy levels are high (0.76), infusing the music with vibrancy. Instrumentalness is low (0.03), with a preference for vocals. Valence is moderate (0.38), providing a balanced mix of emotions. Popularity is moderate (0.50), appealing to a diverse audience.

Cluster 3: "Acoustic Serenity"

"Acoustic Serenity" stands out with a very high acousticness value (0.84), emphasizing acoustic instruments. Danceability is low (0.34), making these songs less suitable for dancing. Energy levels are very low (0.20), creating a calm and relaxed atmosphere. Instrumentalness is high (0.83), as these songs are predominantly instrumental. Valence is low (0.17), reflecting a more somber and introspective emotional tone. Normalized popularity is lower (0.36), indicating lower overall recognition.

Cluster 4: "Versatile Grooves"

"Versatile Grooves" shows a preference for electronic elements, resulting in a low acousticness value (0.13). Danceability is moderately high (0.63), offering a good potential for dancing. Energy levels are relatively high too (0.69), ensuring a energetic listening experience. Instrumentalness is high (0.76), suggesting a significant instrumental component. Valence is moderate (0.44), providing a mix of emotions. Popularity is moderate (0.43), appealing to a broad range of listeners.

Concluding, context categories do not coincide directly with the clusters that were found. The different interpretations say something about different trends in music, however they do not say anything that is useful to this study.

6.5 CLASSIFICATION

Commencing the classification procedure, our initial approach encompassed all available features. Subsequently, we endeavored to enhance classification accuracy through a refined feature selection process. The following features are the baseline for the classification process: 'popularity,' 'valence,' 'tempo,' 'loudness,' 'speechiness,' 'liveness,' 'key,' 'instrumentalness,' 'energy,' 'danceability,' and 'acousticness.'

The accuracy outcomes of various classification methods are as follows:

Logistic Regression Accuracy: 53.15%

Decision Tree Classifier Accuracy: 47.87%

Random Forest Classifier Accuracy: 56.47%

SVM Classifier Accuracy: 55.39%

It is noteworthy that the achieved accuracy rates remain modest, though the results are above baseline chance. This suggests that the classification models employed may have limitations in accurately categorizing playlists based on the designated features. Further exploration and refinement of the classification approach may be warranted to attain more robust results.

6.5.1 Feature selection

In our previous analysis, we noted that each feature exhibited significant variations among the playlist categories, as demonstrated by the ANOVA test. However, upon conducting a more granular examination using the post hoc Tukey test, it became evident that certain features, namely 'liveness, 'tempo,' 'speechiness,' and 'loudness,' displayed limited capacity to differentiate between all of the playlist purpose categories.

To bolster classification accuracy, we opted to exclude these less discriminative features and focused our attention on a curated set of attributes for analysis. Notably, the selected features for this refined analysis included 'acousticness,' 'danceability,' 'energy,' 'valence', 'instrumentalness' and 'popularity.' Our aim was to optimize classification performance by refining the feature set.

The classification outcomes, post-refinement, revealed varying accuracies across different algorithms. The algorithm with the highest accuracy was the Random Forest Classifier, achieving an accuracy of 55.28%. Despite this feature refinement, it is apparent that the classification process did not experience a significant enhancement in accuracy, implying that the challenge of precisely categorizing songs into specific purposes may extend beyond feature selection alone.

To gain further insights, we turned our attention to the boxplots representing the feature distributions.. Notably, in a previous subsection 'valence,' 'popularity,' 'danceability' and 'energy' emerged as features that potentially held more promise for effective classification, as they exhibited pronounced distinctions across all categories, rather than just a few.

In our pursuit of higher accuracy, we decided to further streamline the feature set, retaining only 'valence,' 'popularity,' 'energy,' and 'danceability.' However, the classification results, while showing some fluctuations, remained relatively stable, with the highest accuracy achieved by the Logistic Regression algorithm at 51.2%.

In conclusion, our classification efforts did not yield substantial improvements in accurately categorizing songs based on their intended purposes, particularly within the constraints of the existing playlist categories.

Exploring alternative approaches, we conducted experiments by removing specific playlist categories to assess potential enhancements in accuracy. Notably, excluding the "studying" or "vacation" categories

resulted in a slight accuracy boost across multiple algorithms. The highest accuracy attained, 60.22%, was achieved by the Random Forest Classifier. Upon removing the "running" category, an even greater accuracy was achieved. This time, the random forest classifier came to an accuracy of 71.82%. This outcome suggests that distinguishing between the "party" and "relaxing" categories proved more straightforward in selecting songs for their respective playlists, as evident from the improved accuracy above the 70% threshold—a notable milestone.

6.6 SURVEY

In this subsection, we present the outcomes derived from our user survey, which sought to assess participants' perceptions regarding the underlying purposes of various playlists. As mentioned before, the motivation behind conducting the survey was to compare its results to the previous results of classification. This might tell us whether the classification results were lackluster or if it is an inherently difficult task to classify songs into a specific context category.

Our survey consisted of ten questions, each featuring five potential response choices. Participants were tasked with gauging the intended use of a given playlist by selecting from the following options: "Relaxing," "Partying," "Vacation," "Studying," and "Running/Sports." The distribution of the responses is graphically illustrated below.

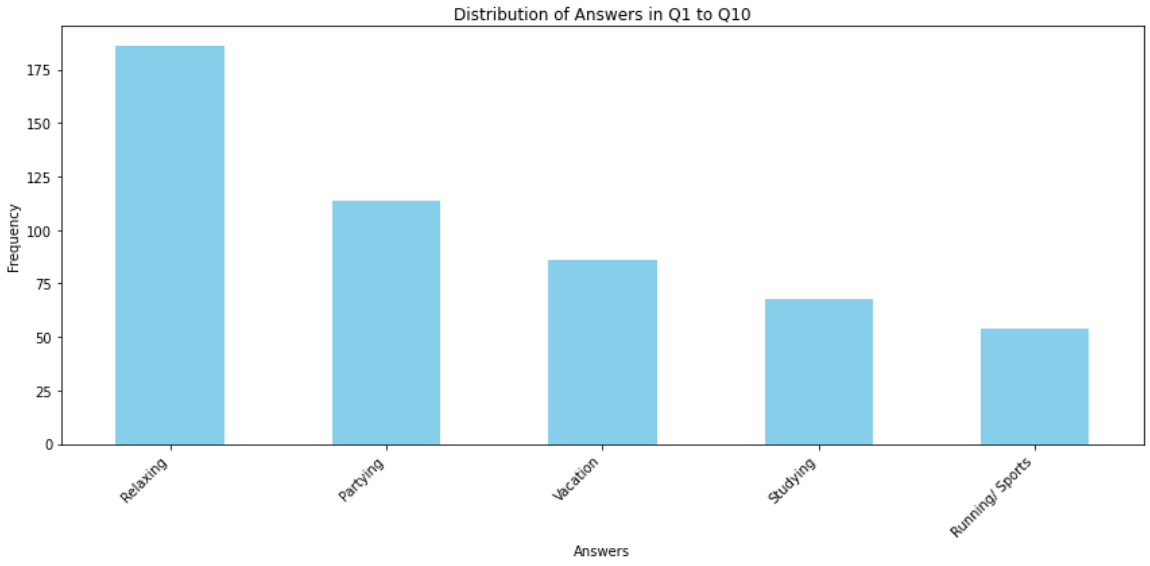


Figure 14: Distribution of Total Answers

As depicted in Figure 15, a substantial portion of the survey participants speculated that the playlists in question were primarily curated for "Relaxing" or "Partying" purposes. These two categories stood out prominently, closely trailed by "Vacation," "Studying," and "Running/Sports," though with noticeably fewer selections. This distribution of answers closely mirrors the prevalence of playlists within our initial dataset, as explored in earlier sections, where playlists catering to relaxation and partying were evidently the most popular.

It is evident, however, that not all respondents were able to accurately decipher the intended purpose of these playlists, despite the fact that we included two playlists within each category for comparison. This observation raises intriguing questions about the factors influencing participants' choices, which will be addressed in more detail.

To quantify the accuracy of participants' responses, we calculated several statistical measures. The average number of correct responses, represented by the mean (M = 5.51), demonstrates that, on average, participants accurately identified just over half of the playlist purposes. The standard deviation (SD) of responses (SD = 1.56) signifies the extent to which participants' responses varied or spread around this mean value.

Purpose	Studying	Partying	Vacation	Running	Relaxing
Studying	41	1	9	8	43
Partying	0	83	8	11	0
Vacation	5	2	49	10	30
Running	7	23	17	25	30
Relaxing	15	0	3	0	83

Table 11: Heatmap of All Survey Answers

To offer a more comprehensive view of these results, Table 11 presents a heatmap that showcases all the responses provided by the participants. The x-axis of the table reflects the answers given by the participants, while the y-axis indicates the correct answers.

When the correct answer was "Studying," participants tended to provide a significant number of accurate responses. However, there was notable confusion among participants when it came to distinguishing between "Relaxing" and "Studying" playlists, which aligns with the earlier discussions of the musical similarities between these categories.

Furthermore, participants exhibited some confusion in discerning the purpose of "Vacation" playlists, with a substantial number erroneously associating them with "Relaxing." This finding resonates with our previous analyses, wherein we uncovered parallels between the musical features of "Vacation" and "Relaxing" playlists.

Interestingly, the survey revealed that "Running/Sports" playlists were occasionally mistaken for "Relaxing" playlists, with 30 respondents attributing the purpose of relaxation to what should have been "Running/Sports" playlists. This observation is intriguing and warrants further exploration.

It is worth noting that some of these survey findings overlap with those derived from our earlier dataset analysis. However, from the survey data alone, it remains challenging to determine the specific criteria upon which participants based their choices. These findings and their implications will be explored more extensively in the forthcoming discussion and limitations sections.

7 DISCUSSION

In this section the results will be discussed, what do the results mean and what can they tell us about a potential future for music recommender systems.

In the related works section, different music recommender systems were discussed as well as the problems they currently face. These problems are caused by the biases that the algorithms create. For instance, many music recommender systems can create popularity biases (Kowald, 2020), where popular songs are more likely to be recommended since those are listened to the most. This process only enforces the popularity of these songs. Other biases that were discussed include “user-item bias”, where recommender systems only recommend songs based on the category of songs that the user listens to most, as well as “feature bias” and “genre bias”. These biases directly contribute to a limited scope of recommendations and therefore enforce a lack of diversity, which causes lower user satisfaction (Anderson, 2020).

Recommender systems based on context might be a solution to some of these biases. Adhering to context allows for a diverse range of recommendations while still being able to personalize the recommendations based on the user. First, temporal based recommendations, where time of day would be a factor in the recommendation process, was discussed. (Herrera, 2010). The research showed promise but using time of day might be too restrictive. Research from Gillhofer and Schedl (2015) showed activities might have some influencing factor on music listening behavior. Therefore we decided to study the correlation between activities and musical preferences further. By performing different analyses on the contents of playlists which are being used for different activities we were able to study this correlation.

7.1 INTERPRETATION OF RESULTS

Summarizing the results, most people tend to listen to songs in playlists geared toward relaxing and partying, followed by running/sports, studying, and vacation. The current playlist contexts that we are analyzing have low diversity overall in terms of songs and artists. Party and relaxing playlists have the lowest song and artist diversity.

The correlation analysis was meant to see if using the features of the dataset, we would be able to differentiate significantly between the activity contexts, to be able to use activity context in the recommendation process. All of the features that were analyzed during the correlation analysis showed a significant difference with the ANOVA test. However, the post hoc Tukey test showed insignificant differences between some of the different context categories, such as with the following features: ‘liveness’, ‘instrumentalness’, ‘tempo’, ‘loudness’ and ‘speechiness’. However, 'acousticness,' 'danceability,' 'energy,' 'popularity,' and 'valence,' showed significant differences between playlist categories, indicating potential for activity-based song recommendations, which, again, would directly help overcome the biases and struggles of modern music recommender systems.

The k-means clustering process did not directly correlate with predefined playlist categories, suggesting that song categorization may rely on more nuanced factors. In light of the lack of direct correlation with playlist categories, we explored alternative interpretations for the clusters, this resulted in a description of the 5 clusters, which, while telling something about the distribution and characteristics of the song types within the clusters, did not say anything useful for our research.

Classification models achieved modest accuracy rates when categorizing songs into specific playlist purposes, with 56.47% the random forest classifier showed the best results. Though, 56.47% is an above chance result, and should not be taken lightly, it seems insufficient to determine activity context

based on the features used. To research whether this result is caused by inadequate features and classification algorithms or whether this result highlights the complexity of playlist classification and user variability, a user survey was created.

The survey revealed that participants were not consistently accurate in identifying playlist purposes, indicating significant user variability even within the same context. Therefore we can conclude that classifying activity context is an inherently difficult task, because users have dissimilar preferences within each different context of activity.

7.2 LIMITATIONS

An important limitation of this study is the lack of data, although we managed to acquire information on 35.557 songs in 538 playlists of 453 users for our data analysis, some information could not be gathered. Mainly information on the users, for example it would have been very useful to be able to compare contexts of playlists of every user to see if users individually listen to other music during different contexts. We were not able to directly analyze how diverse user's music taste is per playlist category, since we were working with a mean of 1.19 playlists per user. The basis of our conclusion that users listen to different kinds of music during different activities is therefore solely based on the correlation analysis which showed significant differences between the activities in terms of features of a song.

Secondly, we were missing time information about when the users listened to specific songs or made certain playlists, and lastly, we were not able to gather any information about the genres of songs from the Spotify API. This put a stop to some of the analyses that initially we wanted to perform. The data gathered might also be specific to certain regions of the world, we tried to combat this by using keywords that were translated from different languages, however, it was impossible to know where the data was gathered from and from which countries the users originated from.

A different problem more speculative problem might be that the playlists that are being studied, could have been put together with the help of a recommender system, this would be counterproductive. However, the playlists will most likely contain music that users have handpicked themselves, even if that's with help from a recommender system. Also, as a non-expert in the field of music, I have looked at patterns from a strictly objective perspective without spending much time getting into the why and how of specific music characteristics create these user behaviors.

In terms of the user survey that we made, though it did provide us with some interesting information, we are unable to know what exactly the users based their choices on. We also don't know if they knew any of the songs. They might not have listened to the video that was provided either.

7.3 FUTURE RESEARCH

Future research should focus on ensuring diversity within music recommendations. A promising avenue for ensuring this is activity based recommendations. However, where some conclusions can be drawn from the analyses discussed within this paper, some limitations exist. Data is needed that can directly study the differences in music listening behavior in terms of activity per user, whereas we have studied the differences using a between-subject study design.

The clustering process, classification methods and the user survey have shown us that there is a high degree of user variability within each of the activities that have been studied, highlighting the complexity of playlist context classification. Therefore, activity context alone will not be enough to recommend music to users and should be used in tandem with other methods that ensure relevant recommendations per user. The combination of collaborative filtering and content based methods together with an activity based method might result in high recommendation accuracies whilst still adhering to users' broad music taste and ensuring diversity within recommendations.

8 CONCLUSION

Concluding, we have discussed music recommender systems and their pitfalls. The related works section paints a picture of the current research in the field of music recommender systems. An important downfall of music recommender systems is the lack of understanding of the underlying principles that make them more accurate, especially in the field of music recommendation. Furthermore, music recommender systems often contribute to a lack of diversity in listening experience for the user and they reinforce popular songs and artists, therein obstructing access to the long tail of unpopular or niche artists and music. Last but not least, music recommender systems do not adhere to peoples broad music taste.

A proposed solution would be context based approaches for music recommendation systems. In this study we have opted for a bottom-up approach by analyzing the features that could influence a context of listening behavior in users, thereby motivating further analysis using clustering and classification. We have determined the context of listening behavior by analyzing the names of playlists that user have constructed. Some of these names were very clearly oriented to 5 different contexts of use.

Results from the correlation analysis show that there are a lot of features with significant differences between the features and activity types. The patterns found during the correlation analysis were mostly unsurprising, such as party playlists having high danceability values. Others were interesting to see, such as studying playlists having the highest instrumentalness values by far, and relaxing playlists not having the lowest values of energy. Some features seem more important than others in terms of the scale of the differences between playlists. However, the results show distinct patterns and differences between playlist types nonetheless. Thereby proving that users listen to different kinds of music during different kinds of activities. Therefore, activity context is usable and useful with a context based music recommender system.

Our clustering process yielded no immediate correlation to playlist categories for the features that we have analyzed. Unsurprisingly, we can conclude that activity context is not the only characteristic that differentiates music choice and what does differentiate musical preference is not directly related to activity context. Through further analysis, we managed to find a different pattern within these clusters, these patterns are more based on feeling and genre. This motivates the use of other recommender methods in combination with activity context.

The classification process yielded similar results. Different stages of feature selection did not seem to improve upon the initial accuracy of around 56%, though this result is above chance. A score of 56% is still too low to rely fully on activity context for recommendations and it shows that users differ between each other in terms of their musical preference during activities, with the features that we have used. Suggesting that, when using context activity in the recommendation process, a user should not be recommended music based on what others tend to listen to in those specific contexts, such as collaborative filtering approaches would. Rather, using content based approaches in combination with using activity context might be the way forward. Thereby allowing personalized recommendations whilst ensuring diversity and respecting the user's diverse musical preferences.

This suggestion is further motivated by the results of the user survey, wherein participants were asked to determine for what contexts certain playlists were constructed. The survey provided similar results as the classification process, with a mean of (M=5.51) showing that participants were correct slightly more than half of the time. Though it is unclear what the participants based their results on, this shows us that using the features we have selected in the classification process, perform as well as human participants,. Both being above chance whilst also being low enough to show that users are diverse in the music they listen to during the contexts that we have been researching. Furthermore, these results show that the

musical preference remains highly subjective and might never be algorithmically solved, though strides can be made to keep improving music recommendation methods.

Answering our research question, with the use of playlist names we can derive accurate contexts of listening. The features of songs can in turn provide significant information about the preferred context for listening behavior and show that there are clear differences between activity contexts. However, features and context do not have a significant enough correlation to provide accurate recommendations. Therefore it should be used in combination with other existing methods of recommendation. Using the contexts of listening, together with content based approaches, relevant but diverse recommendations would be ensured. Future research should focus on combining methods such as these to develop more accurate recommendations while keeping user satisfaction high by maintaining diversity, adhering to users' broad musical preference.

9 LITERATURE

Črnčec, R., Wilson, S. J., & Prior, M. (2006). The cognitive and academic benefits of music to children: Facts and fiction. *Educational Psychology*, *26*(4), 579-594.

Minnix, W. (2016). The Mystical Pentatonic Scale and Ancient Instruments, Part I: Bone Flutes.

Lehmberg, L. J., & Fung, C. V. (2010). Benefits of music participation for senior citizens: A review of the literature. *Music Education Research International*, *4*(1), 19-30.

Burgess, R. J. (2014). *The history of music production*. Oxford University Press.

Madathil, M. (2017). Music recommendation system spotify-collaborative filtering. *Reports in Computer Music. Aachen University, Germany*.

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

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. penguin UK.

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020, April). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020* (pp. 2155-2165).

Tepper, S. J., & Hargittai, E. (2009). Pathways to music exploration in a digital age. *Poetics*, *37*(3), 227-249.

Moscato, V., Picariello, A., & Sperli, G. (2020). An emotional recommender system for music. *IEEE Intelligent Systems*, *36*(5), 57-68.

LiKamWa, R., Liu, Y., Lane, N. D., & Zhong, L. (2013, June). Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services* (pp. 389-402).

Herrera, P., Resa, Z., & Sordo, M. (2010, September). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In *1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain*.

Schedl, M., Bauer, C., Reisinger, W., Kowald, D., & Lex, E. (2021). Listener modeling and context-aware music recommendation based on country archetypes. *Frontiers in Artificial Intelligence*, *3*, 508725.

Polat, H., & Du, W. (2003, November). Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE international conference on data mining* (pp. 625-628). IEEE.

Greasley, A. E., & Lamont, A. M. (2006, August). Music preference in adulthood: Why do we like the music we do. In *Proceedings of the 9th international conference on music perception and cognition* (pp. 960-966). Bologna, Italy: University of Bologna.

Garcia-Gathright, J., St. Thomas, B., Hosey, C., Nazari, Z., & Diaz, F. (2018, June). Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 55-64).

Ryczkowska, A. (2022). Positive mood induction through music: The significance of listener age and musical timbre. *Psychology of Music*, *50*(6), 1961-1975.

Olsen, K. N., Stevens, C., & Tardieu, J. (2007). A perceptual bias for increasing loudness: loudness change and its role in music and mood. *Proceedings of ICoMCS December*, 111.

Balch, W. R., & Lewis, B. S. (1996). Music-dependent memory: The roles of tempo change and mood mediation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1354.

Rosa, R. L., Rodriguez, D. Z., & Bressan, G. (2015). Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, *61*(3), 359-367.

Hargreaves, D. J., Comber, C., & Colley, A. (1995). Effects of age, gender, and training on musical preferences of British secondary school students. *Journal of Research in Music Education*, *43*(3), 242-250.

Chamorro-Premuzic, T., Swami, V., & Cermakova, B. (2012). Individual differences in music consumption are predicted by uses of music and age rather than emotional intelligence, neuroticism, extraversion or openness. *Psychology of Music*, *40*(3), 285-300.

Fox, W. S., & Williams, J. D. (1974). Political orientation and music preferences among college students. *Public Opinion Quarterly*, *38*(3), 352-371.

GURPINAR, E., ZAHAL, O., TASDEMIR, T., & YURGA, M. C. (2022). Musical Preferences of High School Students: The Roles of Emotional Moods and Demographic Characteristics. *International Online Journal of Educational Sciences*, *14*(1).

Pereira, C. S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S. L., & Brattico, E. (2011). Music and emotions in the brain: familiarity matters. *PloS one*, *6*(11), e27241.

Campbell, E. A., Berezina, E., & Gill, C. H. D. (2021). The effects of music induction on mood and affect in an Asian context. *Psychology of Music*, *49*(5), 1132-1144.

Whitman, B., & Jehan, T. (n.d.). The evolution of music consumption: How we got here. Echo Nest. Retrieved from: https://www.makeuseof.com/tag/the-evolution-of-music-consumption-how-we-got-here/

Garcia-Gathright, J., St. Thomas, B., Hosey, C., Nazari, Z., & Diaz, F. (2018, June). Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 55-64).

Schedl, M., Knees, P., McFee, B., Bogdanov, D., & Kaminskas, M. (2015). Music recommender systems. *Recommender systems handbook*, 453-492.

Ospitia-Medina, Y., Baldassarri, S., Sanz, C., & Beltrán, J. R. (2022). Music Recommender Systems: A Review Centered on Biases. *Advances in Speech and Music Technology: Computational Aspects and Applications*, 71-90.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, *84*(6), 1236.

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: a five-factor model. *Journal of personality and social psychology*, *100*(6), 1139.

Sánchez-Moreno, D., González, A. B. G., Vicente, M. D. M., Batista, V. F. L., & García, M. N. M. (2016). A collaborative filtering method for music recommendation using playing coefficients for artists and users. *Expert Systems with Applications*, *66*, 234-244.

Xing, Z., Wang, X., & Wang, Y. (2014, October). Enhancing Collaborative Filtering Music Recommendation by Balancing Exploration and Exploitation. In *Ismir* (pp. 445-450).

Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2006, October). Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences. In *ISMIR* (Vol. 6, pp. 296-301).

Chamorro-Premuzic, T., & Furnham, A. (2007). Personality and music: Can traits explain how people use music in everyday life?. *British journal of psychology*, *98*(2), 175-185.

Greenberg, D. M., Baron-Cohen, S., Stillwell, D. J., Kosinski, M., & Rentfrow, P. J. (2015). Musical preferences are linked to cognitive styles. PloS one, 10(7), e0131151.

Gelder, K. (Ed.). (2005). *The subcultures reader*. Psychology Press.

Thompson, W. F. (Ed.). (2014). *Music in the social and behavioral sciences: an encyclopedia*. Sage Publications.

Jacobson, K., Murali, V., Newett, E., Whitman, B., & Yon, R. (2016, September). Music personalization at Spotify. In Proceedings of the 10th ACM Conference on Recommender Systems (pp. 373-373).

Gillhofer, M., & Schedl, M. (2015). Iron maiden while jogging, debussy for dinner? An analysis of music listening behavior in context. In MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II 21 (pp. 380-391). Springer International Publishing.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.

Simpson, E. H. (1949). Measurement of diversity. nature, 163(4148), 688-688.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. The adaptive web: methods and strategies of web personalization, 325-341.

Greenberg, D. M., Wride, S. J., Snowden, D. A., Spathis, D., Potter, J., & Rentfrow, P. J. (2022). Universals and variations in musical preferences: A study of preferential reactions to Western music in 53 countries. Journal of personality and social psychology, 122(2), 286.

Friedman, R. S., Gordis, E., & Förster, J. (2012). Re-exploring the influence of sad mood on music preference. *Media Psychology*, *15*(3), 249-266.

Kowald, D., Schedl, M., & Lex, E. (2020). The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42* (pp. 35-42). Springer International Publishing.

Koren, Y., Rendle, S., & Bell, R. (2021). Advances in collaborative filtering. *Recommender systems handbook*, 91-142.

Larxel dataset:

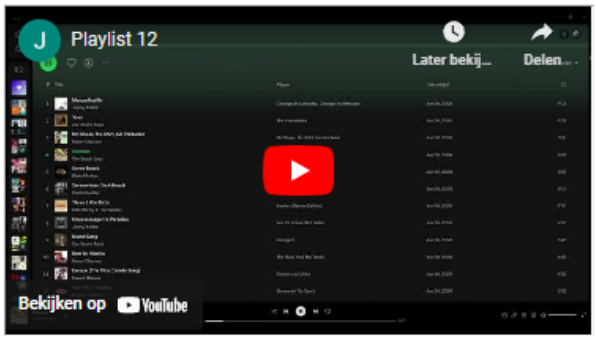
https://www.kaggle.com/datasets/andrewmvd/spotify-playlists?select=spotify_dataset.csv

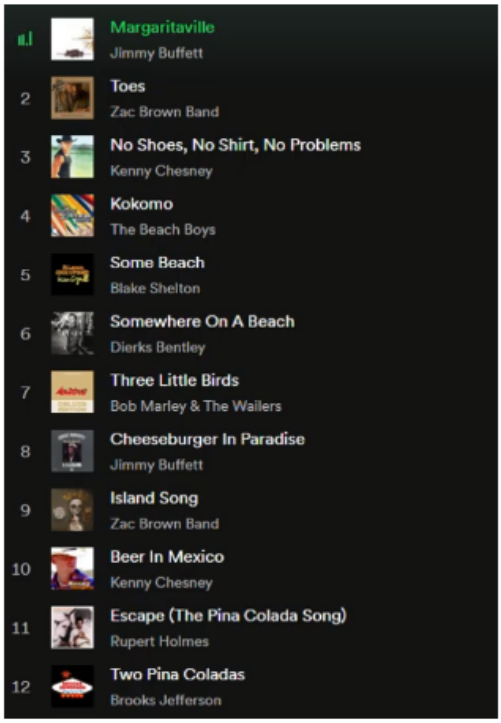
Spotify API:

https://developer.spotify.com/documentation/web-api

Appendix A

Example of a survey question





For what purpose do you think this playlist was created?

- ○ Running/ Sports
- ○ Relaxing
- ○ Studying
- ○ Vacation
- ○ Partying