Gender-Based Music Consumption Pattern Analysis and Personalised Recommendation Using LFM-1b Dataset

Anushree Vishnoi
220756349
Supervisor Name - Jingjing Tang
Data Science and Artificial Intelligence
Queen Mary University of London

***Abstract*—User behavior is inherently complex and dynamic, posing significant challenges when it comes to capturing and accurately modeling it. Users exhibit varied behavior influenced by factors such as personal preferences, contextual and external influences. We discuss in detail the gender-specific user behaviour by analysis of the important features of music recommendation systems.**

We assess the gender-based disparities in features through significance testing and discern significance of these features through machine learning techniques using the LFM-1b dataset. These features included distribution of listening events among artists and users. We also showcase the music consumption patterns between genders across days of the week, hours of the day and music preferences in terms of mainstream and novel music. Additionally, the gender specific datasets are explored for an artist recommendations task, using the collaborative filtering algorithm to analyse the impact of gender imbalanced dataset.

***Index Terms*—Music Consumption Patterns, Gender-Based User Behavior Analysis, Music Recommendation, Collaborative Filtering, LFM-1b Dataset, Digital Marketing**

I. Introduction

Online services and digital marketing has become an integral part of our everyday lives. The content we consume is collected and stored by online platform to build and train recommendation models, that present us with customized suggestions based on our past history and searches. This process of recommending items is facilitated via algorithms and collaborative filtering, is one such popular recommendation algorithm that is implemented across platforms.

Music streaming platforms and marketers use recommendation system algorithms to their advantage (Lambrecht & Tucker 2019), leading to an uplift in the online advertising industry. For example, a fast moving consumer good (FMCG) brand, which is keen to target female audiences, will be more interested to invest monies on the content that is being consumed and recommended to females (household decision makers) than males and vice -versa. This will not only help them save their costs but at the same time improve efficiency of their campaign and increase sales. Similarly, useful insights pertaining to gender specific user-preferences, trends, and consumption patterns can be derived from recommendation systems data. A good recommendation system is one, that is efficient in providing personalized and relevant recommendations to its users.

A latest research by Hesmondhalgh et al. (2023), reveals that music streaming platforms (MSPs) and their music recommendation systems (MRS) are influenced by algorithm dynamics, utilised for making recommendations. There have also been studies in the past (Schedl et al. 2015, Melchiorre et al. 2021), which state that algorithms are unfair to user groups with different characteristics such as age, gender, race, country of origin, or personality (Abdollahpouri et al. 2019). We have conducted an analysis on the LFM-1b dataset to gain an understanding on how collaborative filtering algorithms affect the different users groups based on gender.

The paper is organised as follows: (Section 1) Importance of gender analysis for personalized recommendation systems.(Section 2) A review of collaborative filtering technique and studies related to gender-based music recommendations. (Section 3) A brief overview of the LFM -1b dataset, including its acquisition and content details. (Section 4) Exploration of the dataset to ascertain feature importance, significance testing for gender comparison, and gender analysis based on percentage, geographies, listening events, preferences and consumption behavior. (Section 5) Further exploration of the dataset for building a music recommendation system using collaborative filtering algorithms for artist recommendation and evaluation.(Section 6) A concluding summary of the paper with a summary. (Section 7) Discussion and identification of possible extensions.

II. Related Work

A. Evolution of the Music Industry and Recommendation Systems

The music industry landscape has experienced significant growth since mid 2000s with the introduction of pioneering streaming platforms such as Pandora in 2005 and Spotify in 2008 (Schedl et al. 2018). This transition represents more than just a change in format; it signifies a fundamental shift from the **"Discover + Own"** model to an **"Access"** model. In this context, recommendations systems have become integral to music streaming platforms. These systems aim to

enhance listenership, drive revenue growth, and expand online advertising. Recommending content to listeners goes beyond suggesting songs, artists, or items; it's about delivering a personalised experience that guides users into a continuous journey of music consumption.

B. Types Of Recommendation System Strategies

The most widely adopted strategies in recommendation systems are Collaborative Filtering (CF) (Ricci et al. 2011), and Content-based Filtering (CBF) (Lops et al. 2011), as well as combinations thereof Hybrid recommendation system (Aggarwal 2016), and Context aware recommendation system (CARS) (Adomavicius et al. 2011), that takes into account various contextual factors to provide personalized and relevant recommendations to the users.

C. Collaborative Filtering

CF approaches leverage user behaviour to make predictions and remain the most popular technique for personalised recommendations. It is based on the fact that relationships exist between items/products and people's interests. Many recommendation systems use CF to find these relationships and to give an accurate recommendation of an item/product that a user might like or be interested in. CF basically has two approaches: user–based and item-based; user-based CF is based on the user's similarity (user's neighbourhood) and item-based CF is based on similarity among items (item's similarity).

In user-based CF, we have an active user for whom the recommendation is aimed at. The collaborative filtering engine first looks for users who are similar, that is the users who share the active user preferences or patterns. Similarity between the active (or target) user and other users is calculated to find those with the most similar preferences. Recommendations are then made based on items that these nearest neighbours have interacted with, but the active user has not yet engaged with.

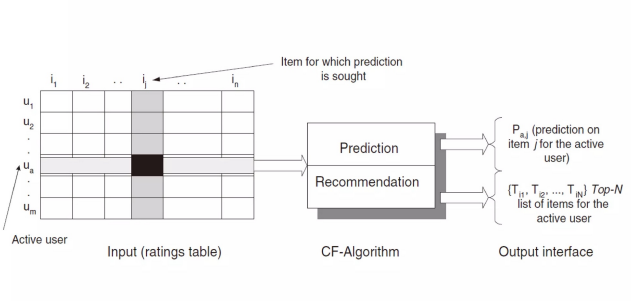

Fig. 1. Collaborative Filtering Process

The collaborative filtering algorithm is based on a matrix that represents historical interactions, such as listening events (MRS). Each entry in this matrix signifies a user's preference, which can encompass various aspects like artists, albums, tracks, and play counts, specifically for a particular item, often an artist. By utilizing similarity measures like cosine similarity or Pearson correlation, the system identifies users, who share similar preferences. These similarity measures generate high-dimensional vectors that encapsulate users' preferences across a range of items, primarily artists in this case. Let $\mathbf{i} = [i_1, i_2, \ldots, i_n]$ and $\mathbf{j} = [j_1, j_2, \ldots, j_n]$ be two vectors in a n-dimensional space. The cosine similarity between these vectors is calculated as:

$$\text{cosine\_similarity}(\mathbf{i}, \mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\|\mathbf{i}\| \cdot \|\mathbf{j}\|}$$

The algorithm calculates the similarity between the active user (the user for whom recommendations are being generated) and other users. The aim is to pinpoint those users whose preferences align most closely with the active user's tastes. Consequently, this process leads to the formulation of recommendations that cater to the active user's preferences and interests.

D. Evaluation trends and gender biases

Evaluation campaigns within the field of music information retrieval and recommendation, such as the Music Information Retrieval Evaluation eXchange[1](MIREX) and the KDD Cup 2011[2] (Dror et al. 2012), have covered a wide range of tasks. While many are task-specific (e.g., tempo estimation or genre classification), personalized approaches requiring metadata is often absent, limiting behavior predictions and analysis.

There are a few publicly available datasets (Brost et al. 2019) for these purposes, the most well-known of which is probably the Million Song Dataset (MSD) (Bertin-Mahieux et al. 2011), but it come with certain restrictions.The listening-specific information in this dataset is provided rather scarcely, on a high level, or in a summarized form only.

Researchers have also explored biases and inequalities within music recommendation systems related to user gender groups (Melchiorre et al. 2020). Those studies have revealed significant disparities, underscoring algorithmic performance differences between male and female user groups.The focus of those studies is primarily based on evaluation experiments to uncover disparities.

We have conducted a detailed analysis of gender bias and unequal gender distribution among participants using the LFM-1b dataset, which offers substantial size, metadata, and important features for the research (Schedl 2016). Emphasizing the results for different user groups while assessing algorithm performance is essential (Liem et al. 2011). It is crucial to critically consider how different genders engage with various recommendation system features, reflecting on personalized behavior (Belkin 2008). Our performance evaluation, employing collaborative filtering using K nearest neighbor similarities for artist recommendation task, dissects gender-based differences and examines them empirically. This paper aims to understand how these differences persist among genders and overall user groups.

[1]http://www.music-ir.org/mirex/wiki
[2]http://www.sigkdd.org/kdd2011/kddcup.shtml

III. THE LFM-1B DATASET

In the sections ahead, we elaborate on the rationale, significance, and acquisition process of the dataset , as well as provide an overview of its content details.

A. Rationale and Significance of the dataset

This creation of this dataset aims to provide a large-scale music recommendation and evaluation dataset for research in the field of recommender systems and music information retrieval. It was developed by Last.fm, a music streaming service and released in 2016.

This dataset stands out due to its emphasis on details specific individual listeners, which are essential for constructing personalized music retrieval systems. The unique aspect of the LFM-1b dataset lies in its comprehensive information with respect to individual listeners and their listening activities. For instance, the dataset provides personalized scores that reflect user's music preferences, including factors like the mainstream nature of their tastes are and their willingness to explore novel music. It is important to note that the dataset is considered a derivative work, as outlined in paragraph 4.1 of Last.fm's API Terms of Service.[3] It was used in accordance with the specified terms and conditions when employed for this study purposes.

B. Acquisition methodology

The LFM – 1b dataset[4], which is approximately 8GB in size was obtained by utilizing the top 250 overall tags[5] from the Last.fm API. This was done to gather information about the leading artists associated with these tags. From the fan base of these top artists, a total of 465,000 active users were extracted. From this larger group, a random subset of 120,322 users were then selected along with their listening histories[6], forming the basis of the LFM-1b dataset.

To ensure fairness, the listening histories of about 5,000 users in the dataset was capped at 20,000 listening events (cf. section). This precaution was taken to prevent an uneven distribution of user data, where a few users might have an exceptionally high number of listening events. A listening event is defined as a quintuple, combination of five components: user, artist, album, track and timestamp. The data collection period spans from January 2013 to August 2014.

C. Overview of the dataset

The dataset is composed of three distinct components. Firstly, there is the metadata, encompassing details about artists, albums, tracks, user's demographic information, and listening events. These details stored in simple text files, encoded using the UTF-8 format. Secondly, there is the additional user features file, which provides additional user features pertaining to user preferences and consumption behavior. This characteristic is unique to LFM-1b and plays a critical role, especially when designing user-aware music recommender systems aimed at delivering personalized music recommendations to individual users. It's important to note that both the metadata and the additional features file contain information specific to individual users. Thirdly, there is the user-artist-playcount matrix(UAM), which was provided as a sparse matrix in a Matlab file, and is formatted using HDF5 for ease of accessibility. The file was created by discarding users who listened to fewer than 10 distinct/unique artists, as well as artists who were listened to by less than 10 users.

It consists of three components:

[1] A 120,175-dimensional vector (idx users), where each element is linked to the user-ids in user's demographic (metadata file), user's additional features file, and the listening events file.

[2] A 585,095-dimensional vector (idx artists), with its elements linked to the artist -ids in artist file (metadata) and the details of listening events (metadata).

[3] A sparse matrix (LEs) of dimensions $120{,}175 \times 585{,}095$, where the rows correspond to users and columns correspond to artists.

IV. DATASET EXPLORATION

A. Feauture Importance

To assess the importance of dataset features for the gender prediction task within the context of recommendation system, various machine learning techniques were implemented including Logistic Regression, Random forest classifier, Support Vector Machine (SVM) and Clustering techniques. Initially, a data balancing technique was employed to rectify the imbalance in the gender ratio, ensuring a fair assessment of feature coefficients for gender prediction. Based on the results and context of the gender prediction problem, Logistic Regression and Random forest classifier emerged as the most suitable for this task (Accuracy - 65%).

On evaluating the feature coefficients, we saw that the "count of distinct tracks" feature had the highest positive importance (0.972), suggesting that it is a strong positive predictor for predicting male. On the other hand, features like "count of listening events", "count of distinct artists" and "mainstreaminess score averaged over 12 months" have negative coefficients, indicating they might be correlated with predicting females.

B. Significance testing and gender analysis across features

In this section, we provide an extensive analysis on the independent features to understand the gender behaviour and make comparisons (1) gender-wise distribution at an overall and country-specific level, (2) distribution of listening events by artists including the number of listeners each artist has,(3) distribution of listening events by users focusing on the count of distinct artists that each user listens to, (4) temporal listening habits - pattern and behaviour across days of the week and hours of the day, (5) music preferences relating to mainstreamness and novelty trends (Schedl & Hauger 2015) averaged over 12 months.

[3] http://www.last.fm/api/tos

[4] http://www.cp.jku.at/datasets/LFM-1b

[5] http://www.last.fm/api/show/tag.getTopTags

[6] http://www.last.fm/api/show/user.getRecentTracks

1) Overall versus country specific gender disparity: To begin, the data was pre-processed, computed and visualized on the basis of gender distribution across all the regions (Table 1) compared to the top countries (Table 2). In Table

TABLE I
GENDER DISTRIBUTION - OVERALL DATASET

Gender	*Number of users*	*Percentage in dataset*
Female	*34085*	*71.63%*
Male	*13501*	*28.28%*
Noise	*72736*	*60.45%*

1, we presents the gender distribution of users (overall) in the dataset. Notably the cleaned data accounts for **39.55%** of the total users. Among these users, we observed that more than two-thirds are male (**71.63%** $>$ **66.67%**), while less than one-third are female (**28.28%** $<$ **33.33%**). Additionally, during the analysis, we identified some erroneous information provided by certain users, such as negative values for age, missing values for country, gender and outliers for playcount.

In Table 2, we present the percentage distribution of female

TABLE II
GENDER DISTRIBUTION - COUNTRY WISE

Country	Percentage Distribution of gender				
	Female	*Male*	*% Female*	*% Male*	*% Total*
US	*2492*	*6233*	*5.24*	*13.10*	*18.34*
RU	*1305*	*3186*	*2.74*	*6.69*	*9.43*
PL	*1632*	*2327*	*3.43*	*3.43*	*8.32*
DE	*950*	*2974*	*2.00*	*6.25*	*8.25*
UK	*840*	*3001*	*1.76*	*6.31*	*8.07*
BR	*1201*	*2345*	*2.52*	*4.93*	*7.45*
NL	*274*	*926*	*0.57*	*1.95*	*2.52*
FI	*339*	*846*	*0.71*	*1.78*	*2.49*
ES	*297*	*753*	*0.62*	*1.58*	*2.21*
UA	*277*	*741*	*0.58*	*1.56*	*2.14*
SE	*270*	*730*	*0.57*	*1.53*	*2.10*

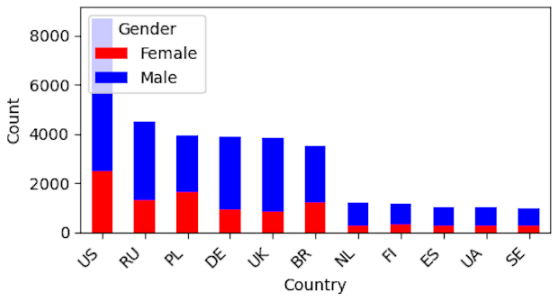

Fig. 2. Gender distribution across regions.

and male users across countries that have the highest number of users in the dataset. We have included only those countries with 1,000 and above users in the analysis. The calculation of country- specific percentages of female and male users is based only on those users who have provided their country and gender information properly. Observations based on overall and country-specific gender distribution data:

- Both distributions were similar, more than two-third was male and less than one-third distribution was female.
- Only two countries were an exception, where the percentage of female users was higher compared to the overall gender ratio - **Poland(PL)** and **Brazil(BR)**. Poland boasts a significant user base, particularly of females, even though it does not rank among the top 30 populated countries.
- The top four focus markets for female gender group was observed as the US, PL, RU and BR , each having more than 1000 female users. For the male user group, the top 7 markets observed were – US, RU, UK, DE, BR, PL, each having more than 2000 male users.

2) Distribution of listening events – by artists: Given that we are dealing with two independent groups (male and female) and that the data does not follow a normal distribution, the **Mann-Whitney U test (also known as Wilcoxon rank-sum test)** was employed for significance testing. The calculated p-value for user's listening events was **1.32e-65**, which strongly indicated that there was a statistically significant difference in the user's listening events (playcount) between the male and female groups. Post which, we conducted an analysis by filtering and plotting the listening events based on the gender groups (female, male and overall) for a comparison. The following figure depict the distribution of listening events for overall listeners in the dataset. For gender plots (see Appendix A) . The red line in the Figure 3 illustrates the sorted listening events for artists, while the blue line represents the number of listeners for each artist. Similarly, Figure 4, displays the sorted listening event counts for all users in red, while the blue line shows the number of artists each user has listened to. The axes of both figures are logarithmically scaled.

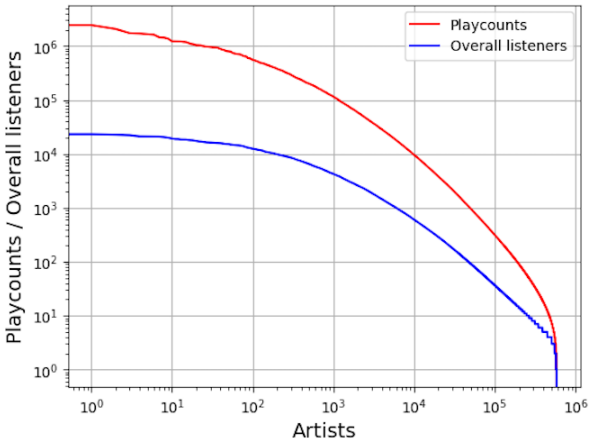

Fig. 3. Distribution of listening events by artists, log – log scaled

Across all the three male, female and overall figures (see Appendix A) and Table 3, we observed a distinct trend, particularly among artists with exceptionally high playcounts located on the left side of the figure and table. As we

TABLE III
PLAYCOUNT PER LISTENER

Gender	playcount/no. of listeners		
	Top most	10^3*th most popular*	10^5*th least popular*
Female	***110***	***27.23***	***7.9***
Male	***114***	***27.12***	***8.4***
Overall	***113***	***27.17***	***8.7***

move, slightly toward the right of the plot, we observed the number of playcounts decreased significantly faster compared to the number of listeners. This indicated that highly popular artists (even the 1000th most popular), have relatively fewer playcounts, despite having a substantial number of listeners. For instance, the top played artist is listened to an average of only 27.23 times respectively, compared to the 100,000th least popular artist, who is listened 7.9 times on an average for female user group and likewise.

These finding strongly support of what's is known as the "long tail" phenomenon in the dataset (Celma 2010). This phenomenon refers to the wise spread distribution of artists with competitively lower playcounts, forming a long tail on the right side of the popularity distribution curve. While a handful of artists enjoy high popularity and are frequently played, a large number of less popular artists collectively accumulate a significant number of plays from a diverse range of listeners.

Gender-based analysis: Analysing Table 3, we noticed that,

- On an average, male listeners have a slightly higher playcounts (114) for their top-played artists compared to female listeners (110). This suggested that male listeners might have a stronger preference for their favourite artists, leading to more frequent listens. The observation introduced the possibility of a popularity bias in music recommendation (Abdollahpouri et al. 2019).
- Music recommendation systems often employ collaborative filtering algorithms, which recommend items based on user behavior and preferences. If certain artist are highly popular among male listeners and are listened to more frequently, the recommendation system could perceive male users as having a higher affinity for those artist. Consequently, this could lead to the recommendation system unintentionally promoting the same popular artist to female listeners, creating a loop where well-known artists are frequently recommended to both genders. This can reinforce the popularity of already well-known artists while potentially overlooking lesser-known but deserving artists, ultimately affecting diversity.
- For the 100th most popular artist both female and male listeners exhibit very similar average playcounts, around 27. This indicated that there is a relatively consistent level of interest and engagement with artists at this popularity tier across genders.
- For the 100,000th least popular artist, the average playcounts per listener follow a decreasing trend, as explained through "long tail" phenomenon in datasets.

3) Distribution of listening events – by users: p-value for count of unique artist was calculated as **5.44e-69**, strongly indicating there is a statistically significant difference in count of artists between female and male groups in the dataset.

Table 4 data gives the ratio of the most, moderately and least active users derived by dividing the listening counts by the number of artists each user has.For gender plots, see Appendix A. We observed that:

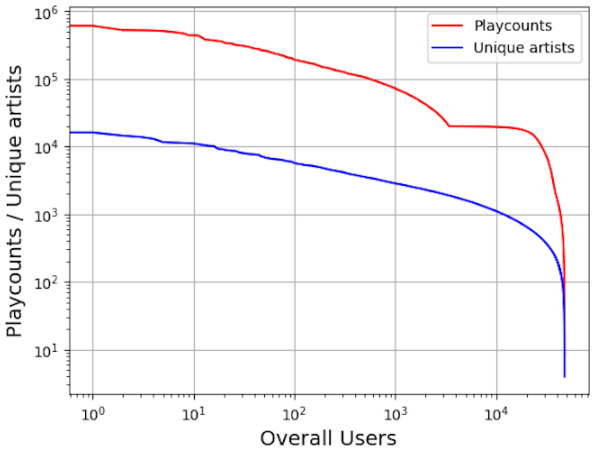

Fig. 4. Distribution of listening events by users, log – log scaled

TABLE IV
PLAYCOUNT PER UNIQUE ARTIST

Gender	playcount/no. of unique artists		
	Most Active	10^3*th highly active*	10^5*th least active*
Female	***40.94***	***30.22***	***11.68***
Male	***37.39***	***33.83***	***21.1***
Overall	***37.79***	***33.74***	***17.6***

- Highly active listeners (most and moderately active users) tend to have a stable relationship between total playcounts and the number of artists they've listened to.
- For least active, the average number of playcounts per artist strongly decreases. The ratios decreases strongly as we move from the most and moderate active user to least active users suggesting that less active users listen to fewer tracks by a smaller number of artists.
- Therefore, we concluded that highly active listeners tend to play tracks by the same artists again and again repeatedly compared to less and occasionally active listeners who tend to play only a few tracks by their preferred artists indicating that the distribution of listening events per artist is influenced by the listening habits of users in the dataset

Please note the listening histories of approximately 5,000 users have been capped at 20,000 listening events, to prevent

an imbalanced user distribution where few users might have an extremely large number of listening events which could skew the analysis and make it difficult to draw meaningful conclusions.

Gender-based analysis:

- The ratio of most active female user 37.79 is slightly higher than the most active male ratio 33.74 which suggests that females tend to tend have slightly higher level of engagement with the artists they listen to compared to males and overall user base (Artist engagement).
- This also indicates that most active female users tend to listen to tracks by a larger number of artists compared most active males.
- For the moderately active user (100th most active), female ratio (30.22) is lower than the male ratio (33.83) suggesting that moderately active males might explore a slightly wider range of artists compared to moderately active females.
- Among the least active ratio again, female user has a lower value than the male value meaning that even among the least active users, males tend to explore more artists than females

4) Descriptors of preferences - Days of the week: In this section, we analysed user-specific features focusing on temporal listening habits.

Temporal listening habits refers to the patterns and behaviors people exhibit in terms of when and how they listen to music over time. These habits can vary from person to person and can be influenced by factors such as daily routines, mood, seasons, events, and cultural practices. We have binned the relative listening events of users based on the days of the week and hours of the day computing the share of each user's listening events over the bins. Figure 5, is a

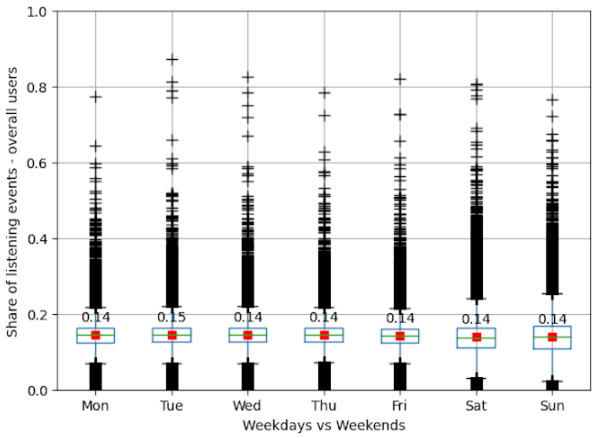

Fig. 5. Distribution of listening over days of the week

box plot illustrations of the distribution of the shares of the listening events for overall users across the days of the week, female, male plots, see Appendix A. The green horizontal line illustrates the median of the data and the two black lines, one is lower and the other upper black line of the box indicate 25-percentile and 75-percentiles. The box plot indicates two horizontal black lines, where the lower line represents the 25th percentile, and the upper line represents the 75th percentile. Additionally, there are horizontal black lines extending further above and below the box, which represent the data points that are not considered outliers. These points fall within the 1.5 times inter-quartile range. Any data points beyond this range are shown as black plus signs. The red circles in the plot indicate the arithmetic mean of the data.

Analysis of the overall users - Fig 5:

- Share of listening event does not differ significantly across weekdays (working days). The number of people listening to music is substantially the same during weekdays.
- During weekend (Saturday and Sunday), there is a much larger spread in the share of listening events. This means that the number of people listening to music varies significantly on weekends compared to working days.
- Median is lower illustrating that majority of people listen less in weekends than during weekdays
- However, the top 25% of active listeners (the most engaged music listeners) consume much more music during weekend. The 75th percentile for listening events is higher on Saturday and even higher on Sunday.
- During weekend, when individuals have more free time, the active listeners tend to consume more music in contrast to casual listeners who might reduce their music consumption during weekends. They might be occupied with other activities, taking advantage of their free time for leisure pursuits or spending time with friends and family.

Gender based analysis:

- The share of listening events (number of people listening to music) and weekend larger spread was common across female and male users
- Difference in the median value for female users – median was lower on Wednesday, Thursday, Saturday, and Sunday in contrast to male/overall plot, due to working habits on weekdays like music consumption while travelling
- In contrast to female audience, who depict a different music consumption pattern than the usual weekday/weekend temporal habits (male and overall plots, see Appendix A). They tend to listen to less music mid-week and on weekends which might be due to engagements in other household activities or responsibilities

5) Descriptors of preferences - Time of the day: Figures 6 illustrate distribution of listening events based on the hours of the day.

Variation over Hours: Distribution of listening over hours of the day exhibited more variation compared to the distribution over days of the week. It implies that there is a significant

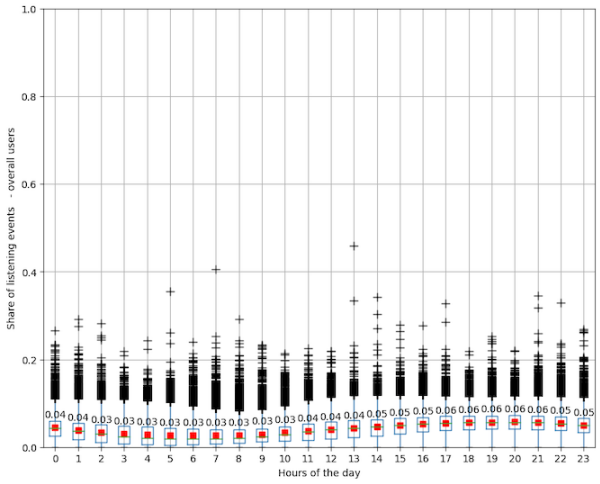

Fig. 6. Distribution of listening over hours of the day

fluctuation in the music consumption pattern throughout the day.

Early Morning Hours: Specifically between 4 AM to 7AM, the number of listening events was relatively low. This was not surprising, as most people are typically asleep during these hours, resulting in less music consumption.

Afternoon and Early Evening: Consumption gradually increased from the morning and peaks in the afternoon and early evenings, between 5PM and 10PM. Probably during this time, users have more free time to listen to music, leading to a surge in listening events.

To summarize, both female and male users, see Appendix A showcased similar music consumption pattern across hours during a day. Understanding these trends, can prove to be valuable for music platforms for tailoring their content and recommendations to align with user preferences and availability throughout the day.

6) Consumption behaviour-music preferences in terms of mainstreaminess and novelty: **Novelty score** is a term used to quantify the level of uniqueness or originality of a piece of music.**Mainstreaminess score** is a term used to quantify the degree to which a piece of music aligns with mainstream trends and popularity. It is defined as the user's distribution of listening events and the global distribution overall. The Mann-Whitney p-value for novelty score was **2.59e-11** and for mainstreaminess score was **1** hence we can say that there is a significant difference in the novelty score between male and female group but that is not true for mainstreaminess score since the p value is greater than 0.05 and hence we do not have enough evidence to reject the null hypothesis.

Table 5, provides the key statistics for the novelty and mainstreaminess scores, which were calculated based on yearly time windows for female, male and overall users. Both user groups (f – 45%, m – 46%) showcased a strong inclination towards novel music exploration as the average mean listenership of both was closer to 50%.

TABLE V
MUSIC PREFERENCES - MAINSTREAM AND NOVEL MUSIC

	Average Mean	
Gender	***Novel***	***Mainstream***
Female	***0.450***	***0.061***
Male	***0.461***	***0.059***
Overall	***0.458***	***0.059***

However, their music preference tends to be highly diverse and distinct from the mainstream, as there was only a mere 6% (f – 5.9%, m - 6.1%) overlap between the user's distribution of listening events and the global distribution on average. This indicated that most of the users prefer to explore and enjoy music outside of the popular and mainstreaminess choices for both male and female gender which could be the reason of the result obtained from the hypothesis testing proving that there is not much difference between groups.

V. MUSIC RECOMMENDATION EXPERIMENT AND RESULTS

A. Why artist recommendation experiment?

Artist recommendation is the preferred practice in Music Information retrieval MIR recommendation systems for several reasons like (1)Data sparsity but artists typically have more interactions than individual tracks leading to denser interaction matrices leading to accurate recommendations, (2) Listening patterns - user's preferences for artists tend to be more stable over time compared to individual tracks,(3) Cold start Problem, it helps mitigate the cold start problem by relying on the user's existing preferences for well-known artist (4) Individual tracks can have varying quality and popularity, leading to noisy recommendations, (5) It provides a way to aggregate preferences across multiple tracks capturing the overall music preference of a user more accurately.

B. Artist Recommendation using Collaborative Filtering

In this paper, we are evaluating a dataset gender-wise using collaborative filtering algorithm for music recommendation Isinkaye et al. (2015). The algorithm aims to suggest the top 10 artists to the female, male and overall users to make comparison as to how the results differ. We employ a standard memory-based approach that involves computing the inner product of the normalized User-Artist Matrix(UAM) after filtering the data based on gender (female users, male users and both). Additionally, we load metadata for artists and users.

For each seed user/target user, we calculate similarity between the seed user and all other users is computed using dot product of their playcount vectors from the UAM. Based on the similarity score, we select the K nearest neighbors to the seed user . Subsequently, a set of recommended artists top 10 (sorting) based on the listening history of these neighbors to the seed user.

TABLE VI
MUSIC PREFERENCES - GENDER BASED ARTIST RECOMMENDATION LIST

Gender	Top 10 artist recommendations - Index									
Female	***1***	***2***	***3***	***10***	***11***	***12***	***15***	***16***	***19***	***20***
Male	***0***	***2***	***3***	***6***	***7***	***9***	***10***	***11***	***12***	***13***
Overall	***0***	***2***	***3***	***6***	***7***	***9***	***10***	***11***	***12***	***13***

C. Results

The results and findings from the recommendation algorithm is presented in Table 6 and discussed as follows, uncovering interesting patterns and trends in gender user groups listening habits.

- Gender based preferences-Artist at indices 1,15,16,19,20 are the popular ones among female users, which are not present in the male list of top 10 artist. Likewise, artist @indices 0,6,7,9,10,11 are preferred by males but don't have any place in the top 10 of female users.
- Diversity of listening-On comparing the overlap of top recommended artists between female and male users reveal how diverse their music tastes are. Differences in the variety of recommended artists indicate distinct listening behaviors.
- Top artists-Male top artist is 0 indices and for female the top artist indices is 1
- Under-representation of female users preferences: Since the system has more data from male users, it will be better at capturing male user preferences and music interests. The recommendations will be biased towards the tastes and preference of male users, as there is more data available to understand their behavior.
- Impact on the diversity of recommendation-Repeatedly recommending popular artists can reinforce and perpetuate mainstream music preferences. As a result, lesser-known artists and genres may struggle to gain recognition and exposure, leading to a potential lack of diversity in the overall music landscape.
- As the system continues to make recommendations biased towards the majority data group (males), it reinforces the population bias in the dataset. Female users might be discouraged from using the system if they consistently receive irrelevant or uninteresting recommendations.

D. Performance Evaluation across user groups

For recommendation system, the popular evaluation strategies are offline experiments, user study and online evaluation (Zangerle & Bauer 2022). We have implemented an offline experiment using CF algorithm for artist recommendation task for female, male and overall user groups. The findings demonstrate that male user group had a higher recall than females (Fig. 7.)(see Appendix A). However, the female group, accounting for 30% of the dataset, exhibits uneven precision and recall curves with low values, highlighting the data imbalance and distinct user behaviour between both groups.

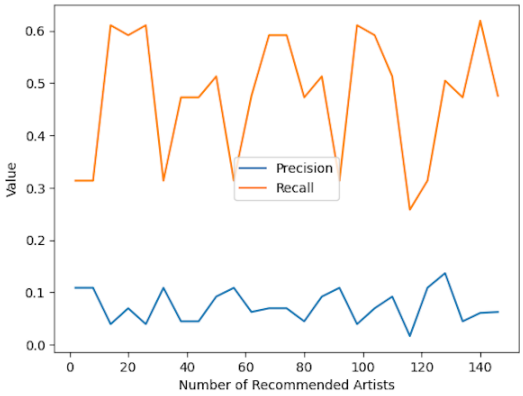

Fig. 7. Precision/Recall plot for the overall users plotted for various numbers of recommended artists N, ranging from 2 to 149 using a step size of 6.

VI. CONCLUSION

We can confidently deduce that there exists a notable distinction in the music consumption habits and behavior between the gender groups.The application of Collaborative filtering techniques present an opportunity to extract meaningful insights for creating personalised recommendations while taking care to mitigate potential biases. The outcomes of the analysis further affirm that the gender group with a more substantial volume of data points holds more influence, potentially overshadowing the user group with fewer instances.This situation can be effectively addressed by tailoring personalised recommendations.

VII. FUTURE WORK AND LIMITATIONS

We have used the memory based Collaborative filtering system (user-item) but there are model based collaborative filtering system that create mathematical models that could be used for the future analysis and dissection of gender based data.These models capture patterns and relationships in user-item interaction (latent features) data that can be used for gender analysis. These models utilize techniques such as matrix factorization, neural networks and other machine learning algorithms. These systems learn to predict user's preferences for items based on available data and the underlying patterns they have captured.

An extension of the experiment could involve deploying multiple recommendation system techniques, including content recommendation, hybrid recommenders, random baselines approaches, for a comparative study.

Lack of relevant data is a major concern in understanding, modeling and predicting how users interact with content on streaming services and has been understudied.

REFERENCES

Abdollahpouri, H., Mansoury, M., Burke, R. & Mobasher, B. (2019), 'The unfairness of popularity bias in recommendation'.

Adomavicius, G., Mobasher, B., Ricci, F. & Tuzhilin, A. (2011), 'Context-aware recommender systems', *AI Magazine* **32**, 67–80.

Aggarwal, C. C. (2016), *Ensemble-Based and Hybrid Recommender Systems*, Springer International Publishing, Cham, pp. 199–224.

Belkin, N. J. (2008), 'Some(what) grand challenges for information retrieval', *SIGIR Forum* **42**(1), 47–54.
URL: *https://doi.org/10.1145/1394251.1394261*

Bertin-Mahieux, T., Ellis, D., Whitman, B. & Lamere, P. (2011), The million song dataset., *in* 'Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)', pp. 591–596.

Brost, B., Mehrotra, R. & Jehan, T. (2019), The music streaming sessions dataset, *in* 'The World Wide Web Conference', WWW '19, Association for Computing Machinery, New York, NY, USA, p. 2594–2600.
URL: *https://doi.org/10.1145/3308558.3313641*

Celma, O. (2010), *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*, Springer.

Dror, G., Koenigstein, N., Koren, Y. & Weimer, M. (2012), 'The yahoo! music dataset and kdd-cup'11', *J Mach Learn Res* **18**.

Hesmondhalgh, D., Valverde, R. C., Kaye, D. & Li, Z. (2023), The impact of algorithmically driven recommendation systems on music consumption and production - a literature review, Report, ARRAY(0x55f77fdb14a8).
URL: *https://eprints.whiterose.ac.uk/196738/*

Isinkaye, F., Folajimi, Y. & Ojokoh, B. (2015), 'Recommendation systems: Principles, methods and evaluation', *Egyptian Informatics Journal* **16**(3), 261–273.
URL: *https://www.sciencedirect.com/science/article/pii/S1110866515000341*

Lambrecht, A. & Tucker, C. (2019), 'Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads', *Management Science* **65**.

Liem, C. C., Müller, M., Eck, D., Tzanetakis, G. & Hanjalic, A. (2011), The need for music information retrieval with user-centered and multimodal strategies, *in* 'MIRUM '11', Scottsdale, Arizona, pp. 1–6.

Lops, P., de Gemmis, M. & Semeraro, G. (2011), *Content-based Recommender Systems: State of the Art and Trends*, Springer US, Boston, MA, pp. 73–105.

Melchiorre, A. B., Zangerle, E. & Schedl, M. (2020), Personality bias of music recommendation algorithms, *in* 'Proceedings of the 14th ACM Conference on Recommender Systems', RecSys '20, Association for Computing Machinery, New York, NY, USA, p. 533–538.
URL: *https://doi.org/10.1145/3383313.3412223*

Melchiorre, A., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O. & Schedl, M. (2021), 'Investigating gender fairness of recommendation algorithms in the music domain', *Information Processing & Management* **58**, 102666.

Ricci, F., Rokach, L., Shapira, B. & Kantor, P. (2011), *Recommender Systems Handbook*, Spriner.

Schedl, M. (2016), The lfm-1b dataset for music retrieval and recommendation, *in* 'Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval', ICMR '16, Association for Computing Machinery, New York, NY, USA, p. 103–110.
URL: *https://doi.org/10.1145/2911996.2912004*

Schedl, M. & Hauger, D. (2015), Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty, *in* 'Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '15, Association for Computing Machinery, New York, NY, USA, p. 947–950.
URL: *https://doi.org/10.1145/2766462.2767763*

Schedl, M., Hauger, D., Farrahi, K. & Tkalčič, M. (2015), On the influence of user characteristics on music recommendation algorithms, *in* A. Hanbury, G. Kazai, A. Rauber & N. Fuhr, eds, 'Advances in Information Retrieval', Springer International Publishing, Cham, pp. 339–345.

Schedl, M., Knees, P. & Gouyon, F. (2018), 'Overview and new challenges of music recommendation research in 2018'.

Zangerle, E. & Bauer, C. (2022), 'Evaluating recommender systems: Survey and framework', *ACM Comput. Surv.* **55**(8).
URL: *https://doi.org/10.1145/3556536*