Track Popularity Dataset Analysis

André Barbosa | João Alves | Marco André | Rúben Monteiro
Engenharia Informática e Computação
Faculdade de Engenharia da Universidade do Porto
e-mail: {up202007398, up202007614, up202004891, up202006478}@up.pt

***Abstract*—This study analyzes the *Track Popularity Dataset* (TPD), concentrating on track and artist similarities, music service comparability, and popularity patterns. The dataset, compiled from *Billboard*, *Last.fm*, and *Spotify*, offers insights into music rankings and listener behavior from 2004 to 2015. Methodologies used include ranking correlation analysis, track and artist similarity assessments, and temporal popularity patterns. Key findings include disparities across services, inconsistent data representation, and noticeable patterns of similarity among artists with varying popularity levels. These provide a foundation for analyzing music trends, although data constraints warrant careful interpretation.**

***Index Terms*—Track Popularity, Artist Similarity, Music Trends Analysis, Music Services Comparison**

I. Introduction

With the ever-growing importance of streaming platforms, analyzing music trends has become increasingly significant for both artists and industry professionals. This includes relationship between music popularity, artist similarity and listener behavior, which all offer valuable insights into the dynamics of the music industry.

This report explores the *Track Popularity Dataset* (TPD), a dataset designed to investigate track popularity and its determinants across different music services. Notably, it includes data from *Billboard*, *Last.fm* and *Spotify*, spanning the years 2004 to 2015. This study looks into the consistency of rankings across platforms, the similarities between tracks and artists, and the rise and fall of track and artist popularity. By addressing these issues, the research hopes to provide insights into the patterns and variables impacting music trends.

Concerning structure, the document is organized as follows: Section II details the used dataset and data pre-processing as well as the initial analysis findings; Section III outlines the methodology used; Section IV showcases the most pertinent findings; and finally, Section V summarizes key insights and future directions. The exploration source code can be found at https://github.com/m21ark/s_music, where additional graphs and tables are available.

II. Dataset

The *Track Popularity Dataset* (TPD), which can be found at http://mir.ilsp.gr/track_popularity.html, is a music-related dataset that has been curated to enable research in the field of music information retrieval (MIR) with the main objective of predicting track popularity across some of the popular music streaming services. The dataset includes data from three primary sources: *Billboard*, *Last.fm*, and *Spotify*, with a timeline that sensibly spans from 2004 to 2015. The following subsections will explore how the dataset was processed and is now structured, as well as present early findings reached from an early data exploration step.

A. Data Pre-Processing

The initial dataset was organized in the form of a relational database, being thus heavily fragmented, with 13 different tables, as illustrated in the compact schema in Figure 1. As such, the team took advantage of this relational organization and combined the tables using their foreign keys. Furthermore, there were tables concerning each individual service that were combined into a single global table.

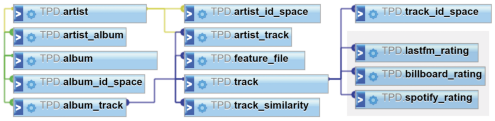

Fig. 1. Relational Database Simplified Schema

The resulting processed dataset was composed of only five relational tables:

- **Track:** Individual information of 23,385 songs (9,193 are designated as popular because they appear in relevant charts, while 14,192 tracks are included in albums containing at least one popular track but not designated as popular themselves);
  - **Main Attributes:** *track_id, title, release_date, album_id, artist_id*
- **Album:** There are 1843 total albums containing all represented tracks.
  - **Main Attributes:** *album_id, name, artist_id, lastfm_playcount, lastfm_listeners*
- **Artist:** Simple *artist_id - name* mapping of all 2557 represented artists.
- **Track Similarity:** In this table, for each *track_id_1, track_id_2* pair, it indicates the calculated similarity degree. However, these pairs exclude a lot of the tracks included in the dataset as only 7211 unique *track_ids* are evaluated.
- **Rating:** Arguably the most interesting data table. It contains the popularity records of the three main services.
  - Billboard: 57,800 weekly records.
  - Last.fm: 43,300 records.
  - Spotify: 6,500 records.

However, only 1.5% of the tracks are listed across all three sources and 5.9% across two sources, limiting ranking correlation analysis among services. Nonetheless, all available data was combined in a single timeline per track, with the columns:

- **Main Attributes:** *track_id, date, position_billboard, position_lastfm, position_spotify, no_of_listeners_lastfm, no_of_listeners_spotify*

A sixth table was created, **Weekly Rating**, which has the same attributes as Rating, plus a new attribute, *time_epoch*. This attribute represents the week number, counting from 29 December 2003, which is the time of the first entry of any of the datasets. It allows mixing all the ratings of a week in a single row, which is a benefit due to all ratings being weekly and having different days of the week.

B. Data Exploration

After the data cleaning process described earlier, the team examined the resulting dataset to derive early insights to inform the project's direction. For instance, a significant portion of artists name are actually collaborations (e.g. "Justin Bieber" technically appears in 18 different artist names), with, in total, 464 different artist names matching that description.

In terms of data distribution, the dataset was considerable imbalanced. Specifically, 63.4% of the represented artists lacked any associated albums and the number of tracks per artist was heavily skewed, with a small number of artists having a disproportionately high track count. This distribution is illustrated in Figure 2, where several artist outliers with large track counts are evident. Among these artists, *Glee Cast* had by far the greatest associated track count of 358.

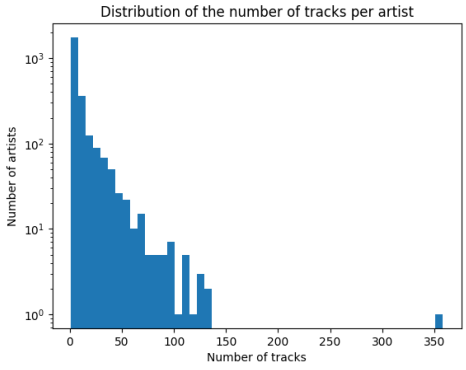

Fig. 2. Track Distribution per Artist

A similar observation can be made for the distribution of tracks per album, as shown in Figure 3. The data exhibits a left and uneven skew, with a notable imbalance and eight outlier albums containing more than 80 tracks.

Finally, regarding weekly individual track popularity metrics, the data from the three music services exhibits non-overlapping characteristics. For example, the Billboard ranking trajectory of the track Demons is depicted in Figure 4. Several noteworthy patterns emerge: first, there are gaps in the data for

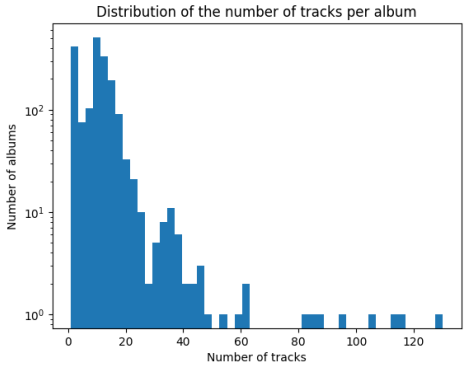

Fig. 3. Track Distribution per Album

certain dates; second, the *Last.fm* data demonstrates significant volatility, a trend observed across most tracks in the dataset.

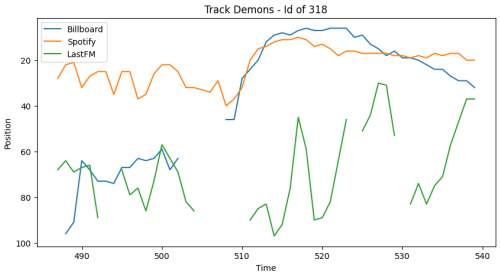

Fig. 4. Billboard Position for Track '*Demons*'

In conclusion, the popularity metrics' instability and the unbalanced dataset suggest that it may not provide an ideal representation of real-world trends. As such, the findings presented in this report should be interpreted carefully.

III. Methodology

The study investigated various aspects of music ranking and artist similarity and popularity, being divided into three primary tasks.

The **first task's** focus was on determining whether artists maintained consistent ranks across different ranking lists. To assess positional alignment, we analyzed the correlation between rankings from the three music services.

The **second task** relates to investigating similarity patterns and finding outliers, notably lesser-known artists that have strong stylistic or track-level similarities with their more popular counterparts.

Finally, the **third task** examined patterns in the rise and fall in popularity for songs and performers. This was accomplished by analyzing indicators like listener counts and chart positions, which may be expressed either numerically or as binary presence in rankings.

For all tasks, a three-stage approach was loosely used: a broad study of the dataset to find overarching patterns, a comprehensive case analysis of representative artists or songs to give nuanced insights, and finally, an assessment of outliers to analyze deviations from the norm.

IV. Results and Discussion

A. Service Popularity Metric Comparison

The team had access to three distinct music ranking services, each providing both chart positions and listener counts. Exploring how each ranking relates to the other would allow us to see if there is any correlation between values and evaluations. Two distinct methods were used to examine the correlation between rankings because the time periods at which each rating was published varied and the number of tracks featured across all services was little.

The first method involved grouping all existing ratings by track and by month, averaging the ranking values for each source. After filtering columns that would have no values, the sample was considerably reduced, but allowed to detect a slight correlation between Spotify and Billboard. Figure 5 displays correlations, demonstrating how little correlation exists between Last.fm and Spotify or Billboard.

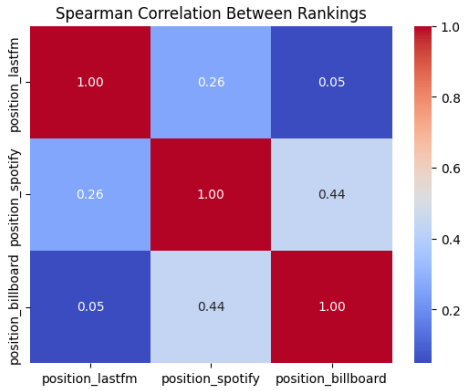

Fig. 5. Music Service Track Rating Correlation

The second method shifted focus to the Weekly Rating dataset, which grouped rankings from different services by track and week. Initially, correlation values were calculated for each track across platforms, as shown in Figure 6. This revealed that Spotify and Last.fm exhibit closer correlations for recent tracks, while Spotify and Billboard show the most significant correlations overall. To address data loss seen in the first method, a pairwise analysis was employed, comparing two services at a time without requiring non-missing values across all three. This approach preserved more data while still allowing for meaningful insights into platform relationships.

Figure 7 illustrates these pairwise comparisons, highlighting the scatter plots, distribution trends, and lines of best fit for each platform pair. The computed overall correlation values—0.50 for Spotify and Billboard, 0.19 for Spotify and Last.fm, and 0.10 for Last.fm and Billboard—confirm a moderate correlation between Spotify and Billboard and weaker relationships involving Last.fm. While it is expected that a popular track would appear on the charts of multiple platforms, the variation in positions may stem from differences in the demographic preferences represented by each service, which could influence how tracks are ranked.

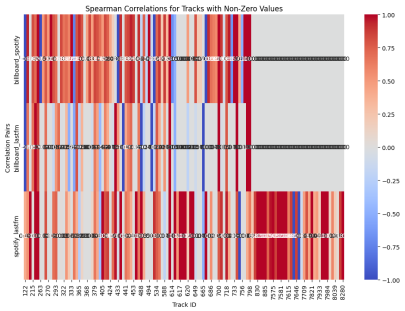

Fig. 6. Service Correlation Value per Track

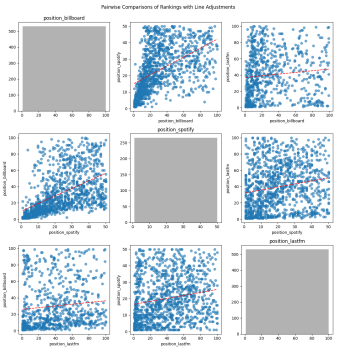

Fig. 7. Pairwise Comparison of Rankings

B. Similarity

The team had access to track similarity scores, to assess the degree of likeness across at both songs and artists levels. Notably, only 30% of the tracks in the dataset were represented in the similarity table and some artists were disproportionately represented, ultimately constraining the scope and nature of the results obtained.

Despite these limitations, the analysis produced notable findings. For example, the team visualized tracks in a graph where the relative distances between nodes were determined by similarity scores. In this representation, nodes with increasingly warmer tones corresponded to more popular tracks, as measured by the total number of listeners. This visualization, presented in Figure 8, clearly shows a heat spot for song popularity, meaning the most popular songs tend to be more similar, with an optimal center to entice listener count.

Regarding artist similarity, determined by averaging pairwise comparisons of their respective tracks, several observations emerge. First, numerous intra-artist comparisons were included, which, as anticipated, generally yielded high similarity scores due to the homogeneity of an artist's own work. Second, a notable portion of the data had to be excluded for two primary reasons: either the comparison involved artists represented by only a single track (insufficient for meaningful relational analysis) or the comparison involved collaborations that were not appropriate for this analysis.

After filtering such instances, similar to the approach applied to track analysis, the team visualized artist inter-similarity in a graph where node colors represents the total

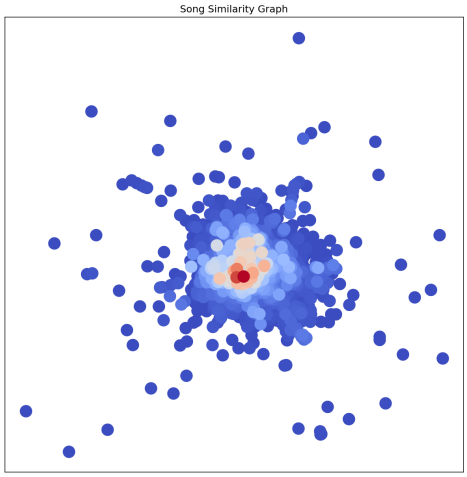

Fig. 8. Track Similarity-Popularity Relation

number of listeners across all tracks and streaming platforms for each artist. These results, illustrated in Figure 9, reveal less distinct patterns. Unlike the track-level analysis, which exhibited a clear concentration, the artist-scale analysis demonstrated a more even distribution across music styles. The graph suggests that the most popular artists at the time were fairly dispersed, with only a subtle agglomeration near the center and around those top artists. This clustering likely reflects a specific music genre and its leading figures.

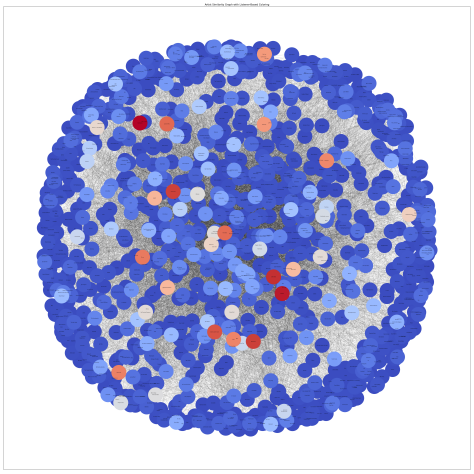

Fig. 9. Artist Similarity-Popularity Relation

Still within the context of similarity analysis, the team briefly examined the relationships between the top 100 artists in the dataset (ranked by listener count) and lesser-known artists. Due to the discography under-representation of the latter group, only 12 meaningful pairs could be extracted, as shown in Table I. In this table, the popularity ratio is calculated as the fraction of listeners for the most popular one in the dataset.

It is noteworthy, given the time frame of the data collection (early 2010s), that some less popular artists, marked in bold

Artist 1	Popularity 1	Artist 2	Popularity 2	Similarity
Arcade Fire	0.93	Muse	0.80	0.67
Becky G	0.01	Taylor Swift	0.65	0.56
Bonnie McKee	0.00	Katy Perry	0.56	0.95
Bonnie McKee	0.00	Lady Gaga	0.96	0.70
Katy Perry	0.56	Pitbull	0.04	0.89
Meghan Trainor	0.02	Taylor Swift	0.65	0.54
Passion Pit	0.26	Arcade Fire	0.93	0.72
Rita Ora	0.01	Katy Perry	0.56	0.56
Rita Ora	0.01	Taylor Swift	0.65	0.58
Sam Smith	0.02	Coldplay	0.98	0.51
The xx	0.82	**Ed Sheeran**	0.15	0.70
Troye Sivan	0.00	Taylor Swift	0.65	0.62

TABLE I
PAIR OF ARTIST POPULARITY AND SIMILARITY

in the table, displayed notable similarity to significantly more popular artists. Interestingly, several of these less popular artists have since risen in prominence, gaining millions of listeners. This trend may reflect the previously identified connection between specific music styles and their associated success. The implication is that lesser-known artists exhibiting stylistic similarity to highly popular artists are more likely to achieve subsequent popularity.

C. Popularity Analysis

The main identifiers for a track's popularity used for analysis were the number of listeners, as well as the track's position in the leaderboards. Regarding a track's popularity, both number of listeners and their position seem to follow a similar pattern of: a small period of discovery with an increase in attention until it reaches a peak, followed by a slow descent until the track falls off the leaderboard. An example of this behaviour is Figure 10 for position, and Figure 11 for number of listeners. Regarding this analysis, it is also noticeable that the three leaderboards have different rankings for any single song, mostly overlapping, but showing slight variations altogether. This can be seen also in Figure 10.

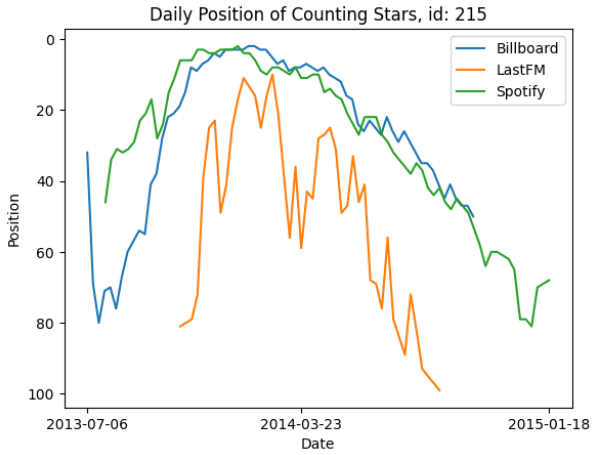

Fig. 10. Daily position of track 'Counting Stars' across leaderboards

Analyzing the most popular artists, based on how long their tracks stay in the leaderboards, its possible to conclude that

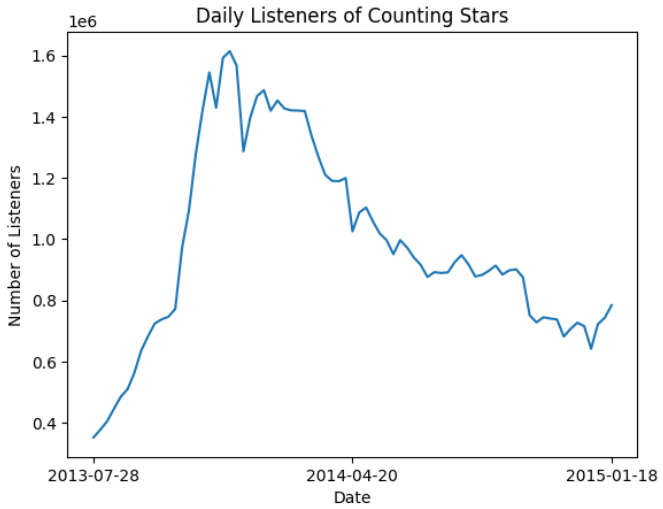

Fig. 11. Daily number of listeners of track 'Counting Stars'

the three leaderboards have very different rankings. This is shown in Figures 15, 16, 17. In a similar way, the popularity of a given artist was also evaluated by the maximum number of listeners at any single point in time, as is shown in Figure 12.

From these figures it is possible to conclude that, although there are some reappearing names, such as 'Drake', 'Eminem' and 'Kanye West', most artists only appear in one figure. This shows the lack of correlation and accordance between leaderboards, as well as between position and number of listeners, as a method of evaluating popularity.

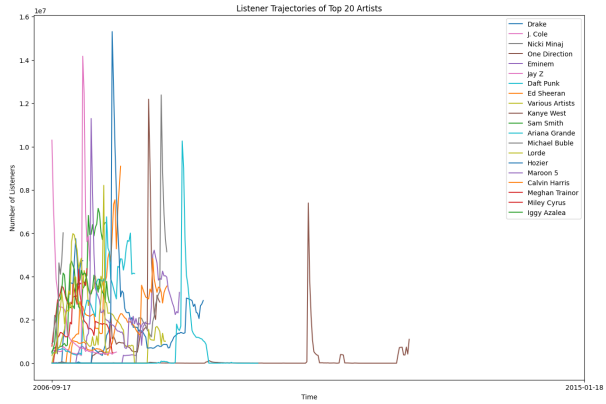

Fig. 12. Top 20 artists based on daily listeners

For each artist, an analysis was also done for their staying time, according to their track's staying time. Examples of this exploration can be seen with Figures 13 14. As argued previously, because the leaderboards contain different data, some artists show different statistics, depending on the leaderboard, while others are only present in a subset of the three leaderboards.

V. Conclusions and Future Work

This study delves into the *Track Popularity Dataset*, shedding light on key elements of music trends that took place between 2004 and 2015. Most notably, the study highlights substantial variations in music ranking systems and reveals a stylistic preference for music that resonates with a "sweet spot" for popular music. Furthermore, the findings suggests that stylistic similarities to famous musicians may impact the prospective growth of lesser-known artists.

However, due to limitations in data distribution and the inherent volatility of popularity indicators, the results should be interpreted with caution. Future research might overcome these limitations by including new datasets or improving approaches for capturing the complexity of music popularity and listener behavior.

APPENDIX

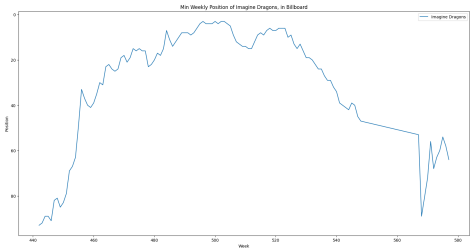

Fig. 13. Highest ranking for any track from artist 'Imagine Dragons' in Billboard

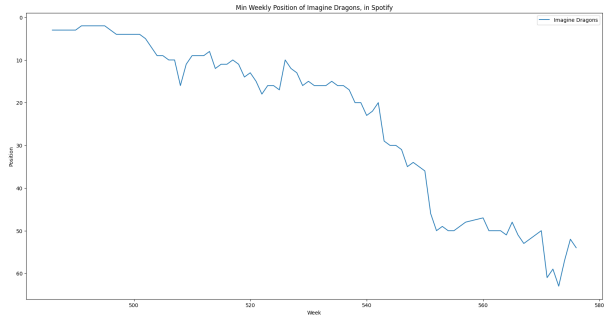

Fig. 14. Highest ranking for any track from artist 'Imagine Dragons' in Spotify

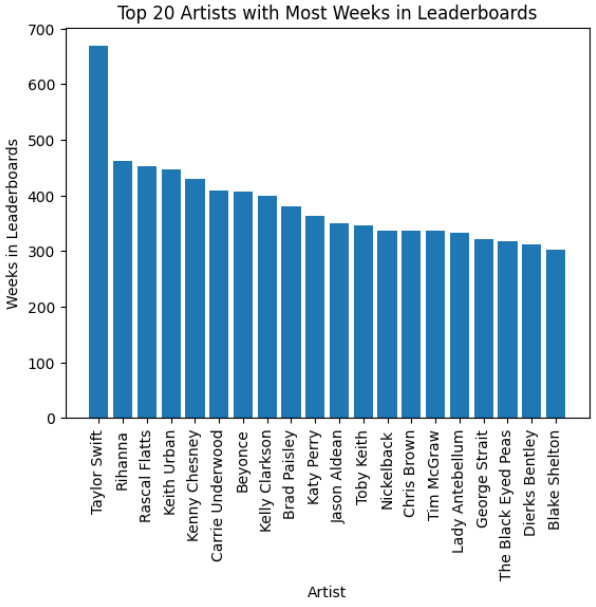

Fig. 15. Top 20 artist based on staying time of their tracks, on leaderboard Billboard

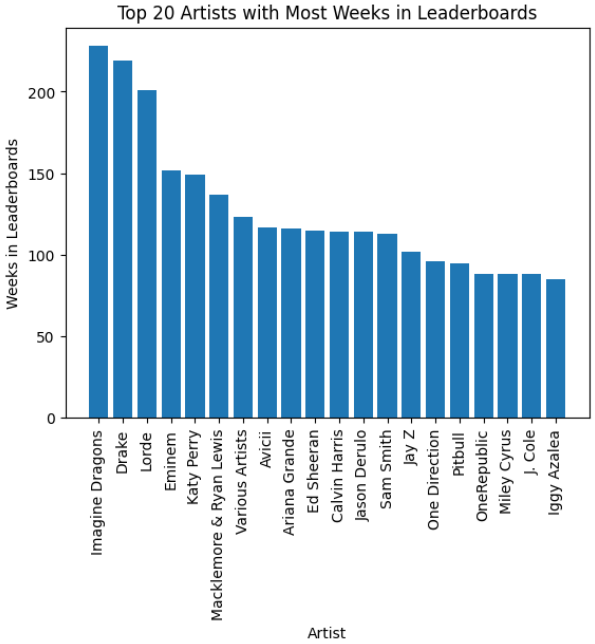

Fig. 16. Top 20 artist based on staying time of their tracks, on leaderboard Spotify

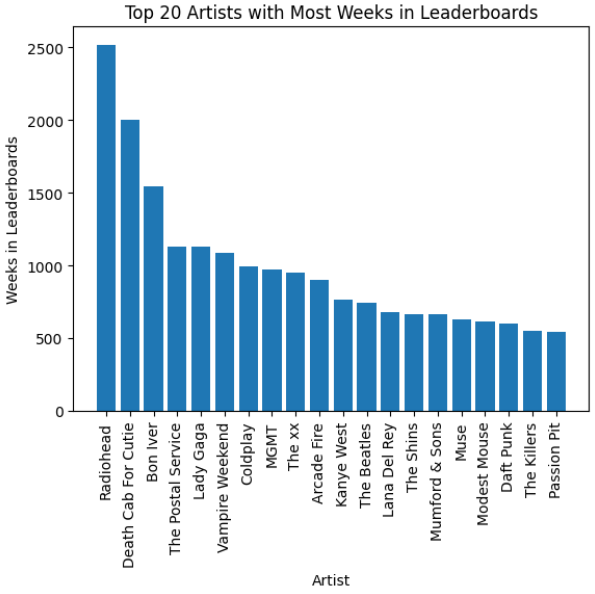

Fig. 17. Top 20 artist based on staying time of their tracks, on leaderboard Last.fm

```
Track: White Christmas - Duet With Shania Twain - Max Consecutive Weeks: 2 - Starting Week: 520.0 - Ending Week: 521.0
Track: Jingle Bells - feat. The Puppini Sisters - Max Consecutive Weeks: 2 - Starting Week: 520.0 - Ending Week: 521.0
Track: All I Want For Christmas Is You - Max Consecutive Weeks: 4 - Starting Week: 518.0 - Ending Week: 521.0
```

Fig. 18. Holiday Season Effect on Rankings