Costa Rica University of Technology

Computer Engineering

IC4302-Databases 2

Project 2

Music Charts Analysis

Professor

Erick Fabian Hernández Bonilla

Students

Segura Barboza Walter A./2023072838

Hidalgo Paz Carmen/2020030538

Ilama Gamboa Naomi Joseph/2021114064

Morales Vargas David Enrique/2021052762

December, 2025

Table of Contents

Figure Index

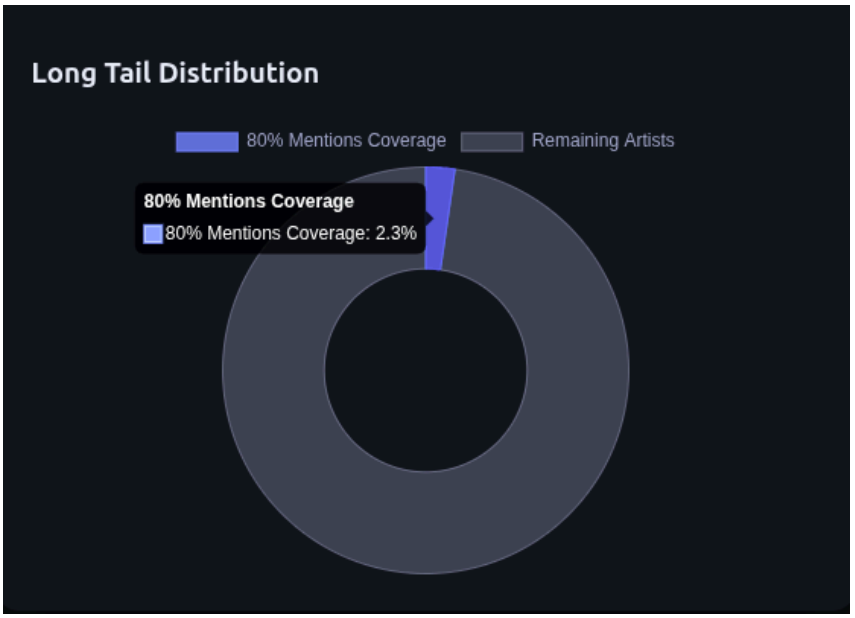
Music Charts Analysis

Conclusions

The Popularity Analysis

Our analysis demonstrates that the dataset exhibits a pronounced long tail distribution. Although over 193 thousand unique artists have been registered, out of the 97,7% that don't make up the 80% of coverage, more than 58% of them show fewer than 5 mentions.

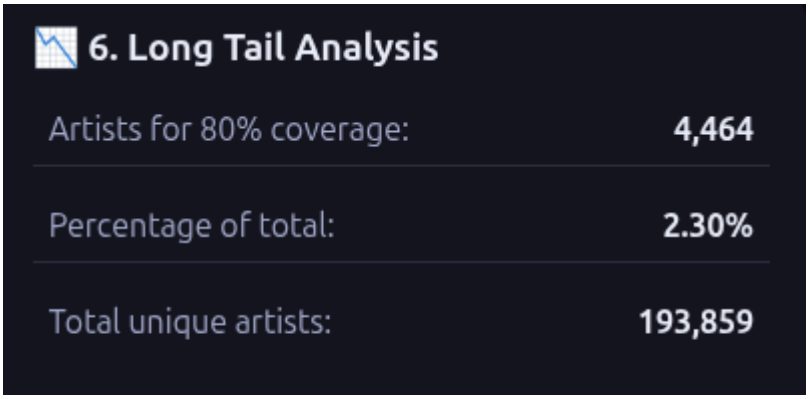
Figure 1 - Long Tail Distribution graph



Source: Our own elaboration

Figure 1 shows the Long Tail Distribution graph, where it can be appreciated the miniscule percentage that makes up the top 80% of artists users listen to.

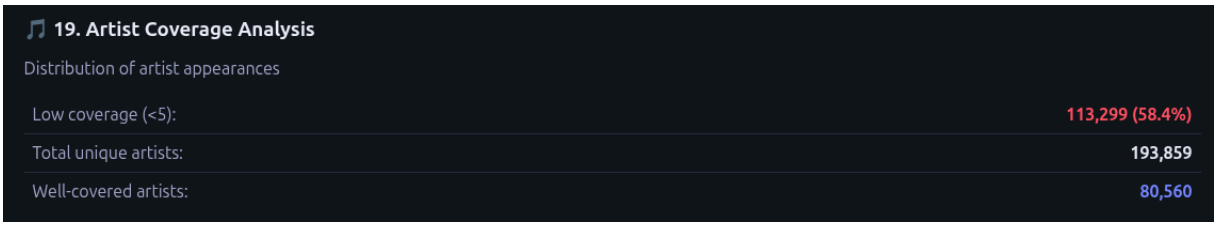
Figure 2 - Long Tail analysis



Source: Our own elaboration

Figure 2 shows the same information as figure 1 but with the added exact numbers for each section.

Figure 3 - Artist coverage analysis



Source: Our own elaboration

In Figure 3, it can be observed the percentage and the exact number of artists that are only five or less times in the users' lists. This finding confirms that the listening behavior of the population is highly skewed, where mainstream artists dominate global statistics while the majority of artists have minimal visibility. It can be seen by the top artist charts as well just how many of the total of listeners are dedicated to each of them:

Figure 4 - Fraction of the Top 20 Artists list

#	Artist	Mentions	Percentage
1	Radiohead	141,178	0.69%
2	Lana Del Rey	107,593	0.53%
3	The Beatles	99,625	0.49%
4	Kanye West	95,535	0.47%
5	Lady Gaga	91,719	0.45%
6	Taylor Swift	87,734	0.43%
7	Arctic Monkeys	87,443	0.43%
8	The Smiths	82,105	0.40%
9	Charli xcx	81,039	0.40%
10	Kendrick Lamar	79,129	0.39%

Source: Our own elaboration

This long tail effect has implications for recommendation systems, model training, and noise reduction. Ideally, every artist should be able to have an equal chance at exposure in music streaming apps, but the reality shows that being a small artist means they will get less recommended to users than the top artists. It could be concluded then, that it might be more important for artists to manage to release something that'll make them popular (whether this is considered good or bad) just so they can place on the charts and start being recommended to listeners.

User Statistics

User behavior metrics reveal significant consistency in listening patterns. More than 34% of users have their most played track belonging to their top artist and 19% maintain the same artist in both their first and second positions. Additionally, the discovery of highly recurring artist pairs and triplets indicates that there are certain groups of artists that are appreciated by the same people.

Figure 5 - User behavior metrics

📈 10. User Behavior Metrics

Metric	Count	Percentage
Users with concentrated taste (top 5 same artist)	1,853	0.37%
Song #1 matches Artist #1	110,936	34.43%
Users with #1 artist in global top 5	37,792	11.73%
Users with stable top-2 positions (same artist)	95,284	19.06%

Source: Our own elaboration

Figure 5 shows the amount of users as well as the equivalent in percentage that have characteristics like their number 1 song matching their top artists as well as having the same artist in multiple positions.

Figure 6 - Fraction of the Top 50 Artist Pairs list

👥 11. Top 50 Artist Pairs

Most frequently co-occurring artist combinations

#	Artist 1	Artist 2	Co-occurrences
1	David Bowie	The Beatles	12
2	Beyoncé	Charli xcx	11
3	Kanye West	Kendrick Lamar	10
4	Britney Spears	Lady Gaga	9
5	Kendrick Lamar	Tyler, The Creator	9
6	Charli xcx	Miley Cyrus	9
7	Charli xcx	Lana Del Rey	9
8	Lady Gaga	Madonna	9
9	Britney Spears	Katy Perry	8
10	Katy Perry	Lady Gaga	8

Source: Our own elaboration

Figure 7 - Fraction of the Top Artist Triplets list

🎯 12. Top Artist Triplets

Most common 3-artist combinations

#	Artist 1	Artist 2	Artist 3	Occurrences
1	Beyoncé	Lady Gaga	Rihanna	83
2	Ariana Grande	Lady Gaga	Taylor Swift	73
3	Beyoncé	Lady Gaga	Lana Del Rey	71
4	Kanye West	Kendrick Lamar	Tyler, The Creator	71
5	Ariana Grande	Beyoncé	Lady Gaga	71
6	Britney Spears	Lady Gaga	Rihanna	69
7	Beyoncé	Britney Spears	Lady Gaga	68
8	Katy Perry	Lady Gaga	Rihanna	67
9	Beyoncé	Lady Gaga	Miley Cyrus	66
10	Lady Gaga	Rihanna	Taylor Swift	65

Source: Our own elaboration

In Figures 6 and 7, not only can multiple names be appreciated more than once, but

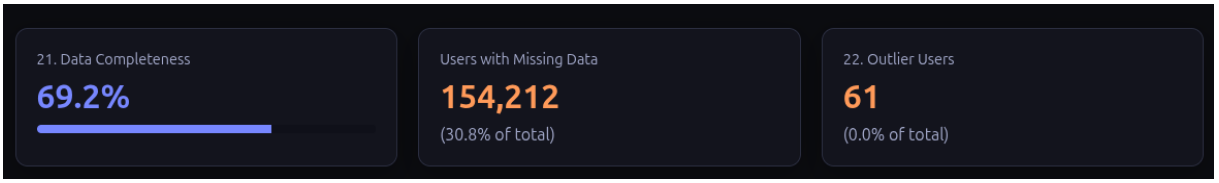
artists like Kanye West, Kendrick Lamar and Tyler, The Creator as well as Beyoncé and Lady Gaga, etc. consistently seem to be listened to and appreciated by the same people. This can be useful when trying to figure out if people have a preferred genre, if they listen to mostly women as opposed to men, if the artists they like are from the same region and more.

These patterns show that users often converge around specific genres, eras, or stylistic groups, and that a large portion of the audience maintains coherent and predictable listening habits. This correlation can be leveraged to enhance collaborative filtering and co-occurrence-based recommendation engines.

Quality Metrics

Data quality and completeness metrics confirm that the dataset maintains a high degree of reliability. Approximately 69% of user entries are structurally complete, and only 61 users fall into the 99th-percentile outlier category, effectively 0% of the users. This means that extreme listening behavior is rare and the dataset is well-behaved for statistical modeling. On the other hand, there is a worrying 154 thousand users that whilst not outliers, do still have missing data which could affect some results.

Figure 8 - Dataset Integrity



Source: Our own elaboration

Figure 8 shows how many users have all their data versus how many have an incomplete profile. This is important to know because these users could change the data collected, whether this is in a positive or negative way, if they are taken into account when their information is incomplete.

Figure 9 - Outlier detection

Source: Our own elaboration

Figure 9 shows the breakdown of these outliers. 56 of them have way too much activity whilst only 5 of them have almost no activity. It's useful to know this because this is suspicious behavior and could either be bot accounts or multiple accounts created by a user to inflate the numbers of a song or artist.

The relatively low number of suspicious users and the large, consistent distribution indicate that the data collected is stable and suitable for large scale processing using Spark. Whilst the majority of users having complete data helps validate the overall robustness of the pipeline, including data loading, HDFS storage, analytical transformations, and MariaDB result integration, it is important to note that the 30.8% of users with missing information could cause some inconsistencies or incorrect records depending on if they are taken into account or not.

Bibliography

Apache Software Foundation (n.d.). Spark Overview – Spark 0.7.0 Documentation. From https://spark.apache.org/docs//0.7.0/

Apache Software Foundation (n.d.). Apache Spark – Official Docker Image. From https://hub.docker.com/r/apache/spark/

MariaDB Foundation (n.d.). MariaDB – Official Docker Image. From https://hub.docker.com/_/mariadb

MariaDB Foundation (n.d.). Installing and Using MariaDB via Docker. From https://mariadb.com/docs/server/server-management/automated-mariadb-deployment-and-administration/docker-and-mariadb/installing-and-using-mariadb-via-docker/

Pure Storage (n.d.). What Is an Apache Parquet File? From https://www.purestorage.com/es/knowledge/what-is-parquet-file.html

Apache Parquet Project (n.d.). Apache Parquet – Overview. From https://parquet.apache.org/docs/overview/