**BAIS:3200**

**Spotify Streams Project Report**

**Seth Barker, Alex Mona, Tony Mischiara, Jake Wis, Jack Gray**

**Introduction**

Streaming platforms have changed the way we listen to music and the way artists create and market ­­­their music. With Spotify being one of the most popular streaming platforms in the world, we are going to examine the factors that make a song more likely to make the Spotify top charts in the United States. We will apply data driven analysis to examine whether the artist, genre, tempo, collaboration, and valence of a song play a role in a song making the top charts in the United States in. This database application and analysis may prove useful for artists and record labels when making music or finding an artist to sign.

**Data**

\*\* Using the group8\_usaspotify\_utf8 file \*\*

The project uses data from Kaggle. The dataset contains songs from the Spotify 'Weekly Top Songs' charts for each country from the week of 2021/02/04 - 2022/07/14 ([Spotify Weekly Top 200 Songs Streaming Data (kaggle.com)](https://www.kaggle.com/datasets/yelexa/spotify200/)). The original dataset contained the weekly global top charts from 2021/02/04 – 2022/07/14 containing 1.8 million rows. We have reduced the size of the data by only selecting data from USA weekly charts from the first six months 2022, reducing the data to 8097 rows. This reduction allows us to focus on the factors that make a song more likely to appear in the USA top charts**.** The original data contains special characters such as accents and diacritical marks which got corrupted when transferring the data into Excel. Using the find and replace function in excel, we were able to keep the special characters to ensure that our data was as accurate as possible. To reduce redundancy, we deleted the country, region, and language columns since all our data is from the USA weekly charts. To make this data compatible with APEX Oracle we used the remove duplicate function to remove all duplicate values. Table 1 provides a description of the data.

*Table 1 Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| Week | Numeric | Week date |
| Rank | Numeric | Ranking of the song on the chart |
| PeakRank | Numeric | highest rank the song achieved on Spotify Charts |
| PreviousRank | Numeric | song’s rank on Spotify Charts in the previous week |
| WeeksOnCharts | Numeric | number of weeks the song was on Spotify Charts |
| TrackID | Text | Unique identifier for each track |
| TrackName | Text | Name of the song |
| Collab | Binary | name of track |
| ReleaseDate | Numeric | 0 if there is only one artist, 1 if not |
| Danceability | Numeric | Danceability describes how suitable a track is for dancing. |
| Energy | Numeric | represents a perceptual measure of intensity and activity. |
| Key | Numeric | The key the track is in. |
| Mode | Binary | modality (major or minor) of a track |
| Loudness | Numeric | The overall loudness of a track in decibels |
| Speechiness | Numeric | Speechiness detects the presence of spoken words in a track. |
| Acousticness | Numeric | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| Instrumentalness | Numeric | Instrumentalness measures from 0.0 to 1.0 on how likely a track contains vocals |
| Liveness | Numeric | Detects the presence of an audience in the recording. |
| Valence | Numeric | describes the musical positiveness conveyed by a track |
| Tempo | Numeric | The overall estimated tempo in beats per minute (BPM) |
| Duration | Numeric | The duration in milliseconds |
| AlbumNumTracks | Numeric | Number of tracks in an album |
| Label | Text | Name of label that promoted the music/artist |
| ArtistID | Text | Unique identifier for artists |
| ArtistName | Text | Spotify artist uri for the artist in artist individual |
| ArtistGenre | Text | names of all artists who participated in the song |

A close-up of a song

Description automatically generated

Fig. 1 Entity relationship diagram (ERD)

* **ERD Description:**

The central entity in the database is SONG, which is identified by TrackID. The SONG entity also contains various attributes of a song, such as the ReleaseDate, Danceability, Energy, Key, and many other attributes of a song. Every attribute except for Collab is required because not all songs are a collab every other attribute we are analyzing is required in every song. In the ERD CHARTRANK is represented as a weak entity because it does not have a primary identifier that is unique. CHARTRANK represents the varying chart positions a song holds each week. A song can have multiple Chart ranks, but each rank is unique to the combination of the song and specific week. A song can also be associated with multiple artists. We show this relationship with the SONGARTIST entity which is an associative entity used to breakdown the many-to-many relationship between song and artist. An artist can have multiple entries in SONGARTIST but only once per TrackID. The SONG entity is connected to both CHARTRANK and SONGARTIST with a one-to-many relationship that is mandatory on the song side. This ERD is a visual representation of our database’s structure and helps to give a better understanding of the relationships between entities as well as the business rules of the database.

A black screen with white text

Description automatically generated

Fig. 2 Graphical Relational Schema (Schema)

* **Schema Description:**

Based off our ERD, we normalized the data and created a relational schema with 5 tables. This includes the 4 original entities from the ERD, plus an additional table ‘ARTISTGENRE’. The song table is the parent table with trackID being the primary key. The song artist and chart rank tables are the child tables to the song table. Both tables use trackId as a foreign key that references trackID in the song table. The artist table is a parent table for the aritist genre table. In artist genre the foreign key is artistID that references the artist table.

**Database Implementation**

To implement the database in APEX, we wrote CREATE TABLE commands for each table in the relational schema.

* **Creating drop table commands**

DROP TABLE SONGARTIST;

DROP TABLE CHARTRANK;

DROP TABLE SONG;

DROP TABLE ARTISTGENRE;

DROP TABLE ARTIST;

* **Creating the 'ARTIST' table**

CREATE TABLE ARTIST (

ArtistID VARCHAR(255) NOT NULL,

ArtistName VARCHAR(255) NOT NULL,

CONSTRAINT ARTIST\_PK PRIMARY KEY (ArtistID)

);

* **Creating the 'SONG' table**

CREATE TABLE SONG (

TrackID VARCHAR(255) NOT NULL,

TrackName VARCHAR(255) NOT NULL,

Collab VARCHAR(5), -- Optional, can be NULL

ReleaseDate DATE NOT NULL,

Danceability VARCHAR(255) NOT NULL,

Energy VARCHAR(255) NOT NULL,

TrackKey VARCHAR(255) NOT NULL,

TrackMode VARCHAR(255) NOT NULL,

Loudness VARCHAR(255) NOT NULL,

Speechiness VARCHAR(255) NOT NULL,

Acousticness VARCHAR(255) NOT NULL,

Instrumentalness VARCHAR(255) NOT NULL,

Liveness VARCHAR(255) NOT NULL,

Valence VARCHAR(255) NOT NULL,

Tempo VARCHAR(255) NOT NULL,

TrackDuration VARCHAR(255) NOT NULL,

AlbumNumTracks VARCHAR(255) NOT NULL,

MainLabel VARCHAR(255) NOT NULL,

CONSTRAINT SONG\_PK PRIMARY KEY (TrackID)

);

* **Creating the 'ARTISTGENRE' table**

CREATE TABLE ARTISTGENRE (

ArtistID VARCHAR(255) NOT NULL,

ArtistGenre VARCHAR(255) NOT NULL,

CONSTRAINT ARTISTGENRE\_PK PRIMARY KEY (ArtistID, ArtistGenre),

FOREIGN KEY (ArtistID) REFERENCES ARTIST(ArtistID)

);

* **Creating the 'SONG ARTIST' associative table**

CREATE TABLE SONGARTIST (

TrackID VARCHAR(255) NOT NULL,

ArtistID VARCHAR(255) NOT NULL,

CONSTRAINT SONGARTIST\_PK PRIMARY KEY (TrackID, ArtistID),

FOREIGN KEY (TrackID) REFERENCES SONG(TrackID),

FOREIGN KEY (ArtistID) REFERENCES ARTIST(ArtistID)

);

* **Creating the 'CHART RANK' table**

CREATE TABLE CHARTRANK (

TrackID VARCHAR(255) NOT NULL,

Weekly DATE NOT NULL,

Ranks VARCHAR(255) NOT NULL,

PeakRank VARCHAR(255) NOT NULL,

PreviousRank VARCHAR(255) NULL,

WeeksOnCharts VARCHAR(255) NOT NULL,

CONSTRAINT CHARTRANK\_PK PRIMARY KEY (TrackID, Weekly),

FOREIGN KEY (TrackID) REFERENCES SONG(TrackID)

);

**Analysis**

This exploration aims to enlighten individuals aspiring to thrive in the realm of music analytics, delineating the essential factors that influence a song's ascent to the Spotify top charts in the United States. By employing a data-driven methodology, we delve into the intricate interplay of artist identity, genre, tempo, collaboration, and valence to decipher the patterns that contribute to a song's success. Furthermore, our analysis serves as a valuable resource for artists, record labels, and industry enthusiasts, offering actionable insights for refining creative strategies and making informed decisions in the ever-evolving landscape of the music industry.**Top of Form**

* **Question one:**

**What genre is most popular among top charts in 2022?** Specifically, our group wanted to see if genre shows significance in what songs would reach the top charts. Stereotypically, one may have assumed that pop would continue to dominate in the mainstream charts, but we wanted to see if that would hold true based on our accumulated data. To tackle this question, we decided to make a simple join query combining tables SONGARTIST, CHARTRANK, and ARTISTGenre while tracking an Artist’s genre and number of TrackIDs while making sure distinct is used to not double count. We then grouped this data by Artist Genre and ordered it by the aggregate data column we created called “NumberOfTopChartSongs” in descending order.**Top of Form**

SELECT ARTISTGENRE.ArtistGenre, COUNT(DISTINCT SONGARTIST.TrackID) AS NumberOfTopChartSongs

FROM ARTIST

JOIN SONGARTIST ON ARTIST.ArtistID = SONGARTIST.ArtistID

JOIN CHARTRANK ON SONGARTIST.TrackID = CHARTRANK.TrackID

JOIN ARTISTGENRE ON ARTIST.ArtistID = ARTISTGENRE.ArtistID

GROUP BY ARTISTGENRE.ArtistGenre

ORDER BY NumberOfTopChartSongs DESC;

A screenshot of a computer

Description automatically generatedThe results of the query are shown below (Figure 3). From Spotify’s data, the artist genre and its number of top chart songs display the count of songs that appeared in the top charts from the most popular genres of music in the year 2022. As we expected, pop took the throne as the top genre, but genres “rap” and “hip hop” did not fall far behind.

Figure 3: Artist Genre Count on Top Chart Songs

* **Question two:**

**Who are the artists with the greatest number of songs in the top charts in 2022?** In this question, we wanted to test both the popularity of artists as well as how listeners received these artists’ music. Our group was interested in seeing how correlated the top pop artists would be with the number of songs on the top charts, but we wanted to see a clear visualization of our dataset. The table below shows the top 10 artist with the most songs within the top charts. In 2022 a song by Kendrick Lamar had the most appearances with 33.

We made another simple join query, joining tables SONGARTIST and ARTIST on ArtistID. We select the artist’s name and aggregated count of all their songs as “NumberofSongs.” We then ordered the number of songs by each individual artists by the “NumberOfSongs” column in descending order again.

SELECT ARTIST.ArtistName, COUNT(\*) AS NumberOfSongs

FROM ARTIST

JOIN SONGARTIST sa ON ARTIST.ArtistID = sa.ArtistID

GROUP BY ARTIST.ArtistName

ORDER BY NumberOfSongs DESC;

Our results (Figure 4) exemplified the fact that artists with the greatest number of songs in the top charts correlated with artist genre. For example, rap being the second highest genre among top chart songs would imply that artist’s whose genre is rap would have a high number of songs within the top charts.

**A black and grey background with black stripes

Description automatically generated with medium confidence**

Figure 4: Artist and Their Respective Number of Songs on Top Charts

* **Question three:**

**How does longevity on the top charts relate to a songs average overall rank?** In this question we wanted to see if there was a relationship between a songs average overall rank and how long it stays in the top charts. Assuming a song has a high average, rank would leave it remaining in the top charts for an extended period compared to a lower ranked song.

SELECT CASE

WHEN WeeksOnCharts <= 52 THEN '1 Year'

WHEN WeeksOnCharts <= 104 THEN '2 Years'

WHEN WeeksOnCharts <= 156 THEN '3 Years'

WHEN WeeksOnCharts <= 208 THEN '4 Years'

WHEN WeeksOnCharts <= 260 THEN '5 Years'

ELSE 'More Than 5 Years'

END AS RankCategory, ROUND(AVG(Ranks), 0) AS AverageRank

FROM CHARTRANK

GROUP BY CASE

WHEN WeeksOnCharts <= 52 THEN '1 Year'

WHEN WeeksOnCharts <= 104 THEN '2 Years'

WHEN WeeksOnCharts <= 156 THEN '3 Years'

WHEN WeeksOnCharts <= 208 THEN '4 Years'

WHEN WeeksOnCharts <= 260 THEN '5 Years'

ELSE 'More Than 5 Years'

END

ORDER BY RankCategory ASC;

A black and white background with several horizontal lines

Description automatically generated with medium confidenceFrom the results you can see that the general trend is that the longer a song is on the top charts the lower its average rank is. Songs that are on the top charts for a shorter period often have a higher average rank. This suggests that songs often start with a rank when they first enter the charts and then slowly drop in the charts or leave the top charts.

Figure 5: Longevity in top charts and average ranking

* **Question four:**

**Which record label produces the most songs on the top charts?** In this question we wanted to see if there was a certain record label that is more likely of producing top chart songs.

SELECT SONG.MainLabel, COUNT(DISTINCT CHARTRANK.TrackID) AS NumberOfTopChartSongs

FROM SONG

JOIN CHARTRANK ON SONG.TrackID = CHARTRANK.TrackID

GROUP BY SONG.MainLabel

ORDER BY NumberOfTopChartSongs DESC;

**A screenshot of a computer

Description automatically generatedFrom the results you can see that Columbia produced the most songs in the top charts. This data can be useful to up and coming artists as they can see which record labels are producing popular songs.**

Figure 6: Label and their number of songs on the top charts

* **Question Five:**

**Does the tempo of a song affect a song’s likelihood of making it on the top charts?** In this question we are looking to see if there is any correlation between tempo and top chart placement. If a songs tempo is high or low tempo, which is more likely to do better among top chart songs? We differentiated between high and low tempo with high being greater than or equal to 100 beats per minute (BPM). We found out that there are more songs with a high tempo in the top charts.

SELECT CASE

WHEN SONG.Tempo >= 100 THEN 'High Tempo'

ELSE 'Low Tempo'

END AS TempoCategory,

COUNT(DISTINCT CHARTRANK.TrackID) AS NumberOfTopChartSongs

FROM SONG

JOIN CHARTRANK ON SONG.TrackID = CHARTRANK.TrackID

GROUP BY CASE

WHEN SONG.Tempo >= 100 THEN 'High Tempo'

ELSE 'Low Tempo'

END

ORDER BY NumberOfTopChartSongs DESC;

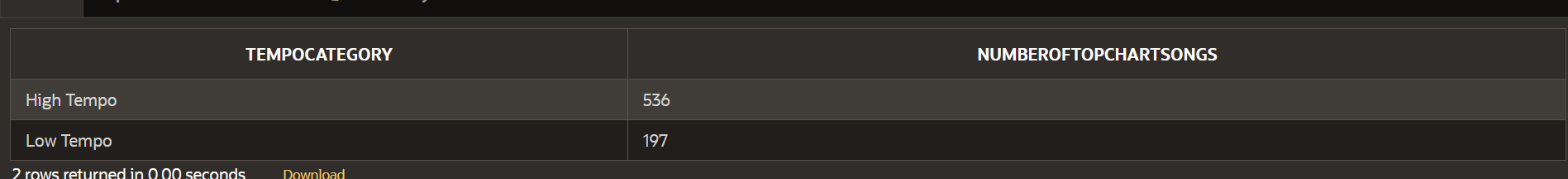
In our results (Figure 6) we found that there was a correlation between tempo and top chart placement. From our data it was apparent that songs with a higher tempo were much more likely to to place in the top charts compared to songs with a lower tempo.

Figure 7: Tempo and its effect on top chart placement

**Web Application**

* **Home Page:**

The home page to the web application contains a brief description of the project. The original data source we used is linked [here](https://www.kaggle.com/datasets/yelexa/spotify200/). Within the application is a simple navigation menu organized into the tables, charts, and a home page featuring [Spotify Charts Weekly Top Songs](https://www.google.com/url?sa=i&url=https://blackpink.cafe/charts/lisas-money-reaches-a-new-peak-of-5-on-spotifys-top-50-songs-global-chart/&psig=AOvVaw0vJykKBjAceS7UV82YeRNY&ust=1702323987580000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCJj2spjRhYMDFQAAAAAdAAAAABAD).Throughout the entire application we used colors to match Spotify’s brand and logo to create a unique and attractive appeal to users. Figure SOMETHING shows a screenshot of the home page.

A screenshot of a music app

Description automatically generated

Figure 8: Home Page

**Tables:**

* A screenshot of a computer

  Description automatically generated**Artist Table:**

Figure 9: ArtistID paired with Artist Name

* **Song Table:**

A screenshot of a computer

Description automatically generated

Figure 10: Song Table with TrackID as primary key and other attributes

* **Artist Genre Table:**

A screenshot of a computer

Description automatically generated

Figure 11: ArtistID paired with Artist Genre

Figure 11:

* **Song Artist Table:**

A screenshot of a computer

Description automatically generated

Figure 12: SongID paired with ArtistID

* **Chart Rank Table:**

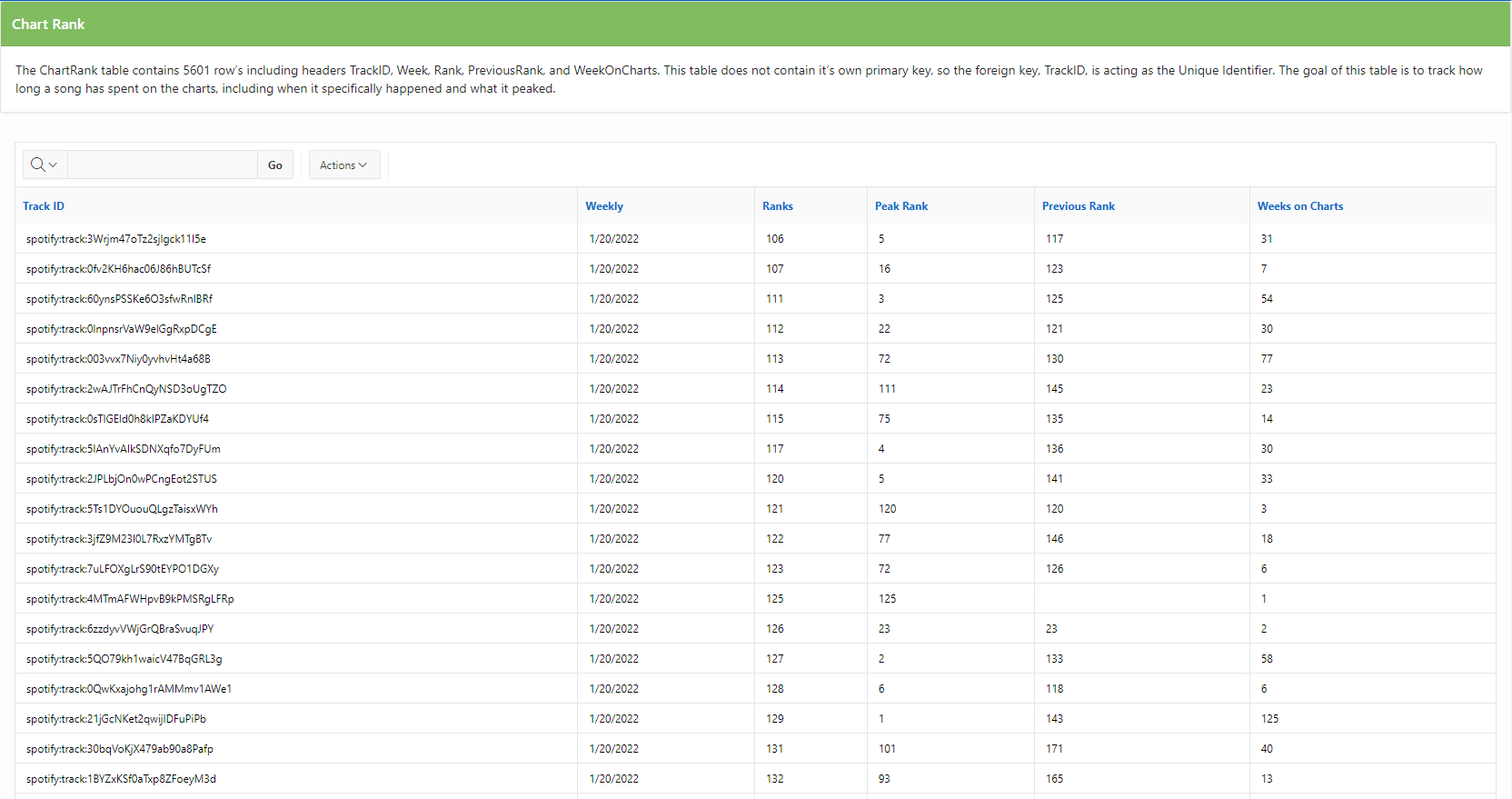


Figure 13: Table that shows the TrackID and its various chart ranks.

**Queries:**

We presented the results of the first question, regarding which genre of music is most popular among top charts as a classic report (Figure SOMETHING). In addition to the visual data we added a text box explaining the research questions and results. Below, you can see, Figure Something, which goes along with question one. This bar chart shows popularity of genres among the charts.

* **Question/Query one:**

We presented the results of the Genre question in the form of a classic report. We also added a bar chart which visualizes the number of songs of each genre that appear on the top charts.

­A close-up of a text

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a music list

Description automatically generated

Figure 14: Tempo Chart and Table

* **Question/Query Two:**

We added this section of the application to present the answer to question two in a table. This shows the artists with the most songs on the top charts in the first 6 months of 2022.

A screenshot of a computer

Description automatically generated

Figure 15: Top Artist’s Table

* **Question/Query Three:**

We added this section of the application to present the answer to question 3 in a table. The table shows the average rank of songs based on how long they appeared on the top charts.

A screenshot of a computer

Description automatically generated

Figure 16:

* **Question/Query Four:**

A screenshot of a computer

Description automatically generated

Figure 17:

* **Question/Query Five:**

We presented the results of the tempo question in the form of a classic report. We also added a pie chart which visualizes the tempo of each song on the top chart with a high tempo song being above 100BPM and a low tempo song being below 100BPM.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 18: