Final Project - Spotify Charts Analysis using Spark (Group Beta)

**Class number & name:** ALY6110 & Data Management and Big Data

**CRN number:** 80443

**Member & Email address:** Trieu Vo [vo.trieu@northeastern.edu](mailto:vo.trieu@northeastern.edu)

Jack Brown [brown.jac@northeastern.edu](mailto:brown.jac@northeastern.edu)

Chuanzhang Tian [tian.chu@northeastern.edu](mailto:tian.chu@northeastern.edu)

Lei Fang [fang.lei1@northeastern.edu](mailto:fang.lei1@northeastern.edu)

Dongkai Yu [yu.do@northeastern.edu](mailto:yu.do@northeastern.edu)

Table of Contents

1. [**Summary 3**](#_Toc103757305)
2. [**Introduction 4**](#_Toc103757306)
3. [**Dataset 4**](#_Toc103757307)
4. [**Analysis 5**](#_Toc103757308)
5. [**Comments 16**](#_Toc103757309)
6. [**Conclusion 17**](#_Toc103757310)
7. [**Appendix 18**](#_Toc103757311)
8. [**References 19**](#_Toc103757312)

Summary

Our effort aimed to compare characteristics of the most and least popular musicians in Spotify's top 200 to provide context for an artist's popularity. Spotify, which was created in 2006, has become the world's most popular streaming network. We will utilize this platform's data collection to assess song trends among vocalists. From there, it is possible to recommend ways to increase the popularity of vocalists, so making them more competitive and successful within the music business. For instance, Ed Sheeran featured in the Top 200 three hundred times more often than Aspova, while Ed's tracks appeared six times more frequently than Aspova's. Aspova should concentrate more on the quality and promotion of his songs than on the quantity of his songs. Ed Sheeran does better in solo tracks, but the most successful song by Aspova, Susamam, is a collaboration. Aspova should work with well-known musicians to increase the popularity of his music. Ed Sheeran's songs are in English and have been successful all over the globe, but Aspova's songs are in Turkish, hence their popularity was brief. Aspova should produce additional songs in English to enter the global music business and work on long-term marketing to maintain his hits at the top of the charts.

Introduction

The music industry is growing faster than ever. Each year, new platforms and mediums skyrocket to stand out, make a name for themselves, and reshape the way audiences connect with artists. At the same time, new technologies put creative tools in the hands of people who were previously unable to access them. Obviously, this industry is growing rapidly and becoming more competitive. Companies are having to differentiate themselves by providing unique, artist-focused content or developing different pricing models. Spotify, founded in 2006, has become the most popular streaming platform worldwide. We will use the data set from this platform to analyze song trends of singers. From there, it is possible to suggest solutions to improve the popularity of singers, which in turn will help them be more competitive and successful in the music industry.

**Dataset**

We use the dataset 'Spotify Charts' from Kaggle, with a file size of 3.48 GB. It has 26,173,514 observations and 9 columns, including title, rank, date, artist, url, region, chart, trend, and stream. This dataset contains the top 200 streamed tracks on Spotify every day from Jan 1 2017 to Dec 31 2021, collected by using Spotify API. The data is refreshed daily. The original chart is shown in this link: <https://spotifycharts.com/regional>

There are 9 columns in this dataset and their description are as follows.

***title***: title of the song

***rank***: rank from 1 - 200 (1 is the most streamed track that day)

***date***: date of data

***artist***: artist name

***url***: url of the song

***region***: countries around the world

***chart***: top200 or viral50

***trend***: the position of that song on the chart compared to yesterday. It has 3 values: MOVE\_UP, MOVE\_DOWN or SAME\_POSITION

***streams***: the total number of global streams of that song in one day

Analysis

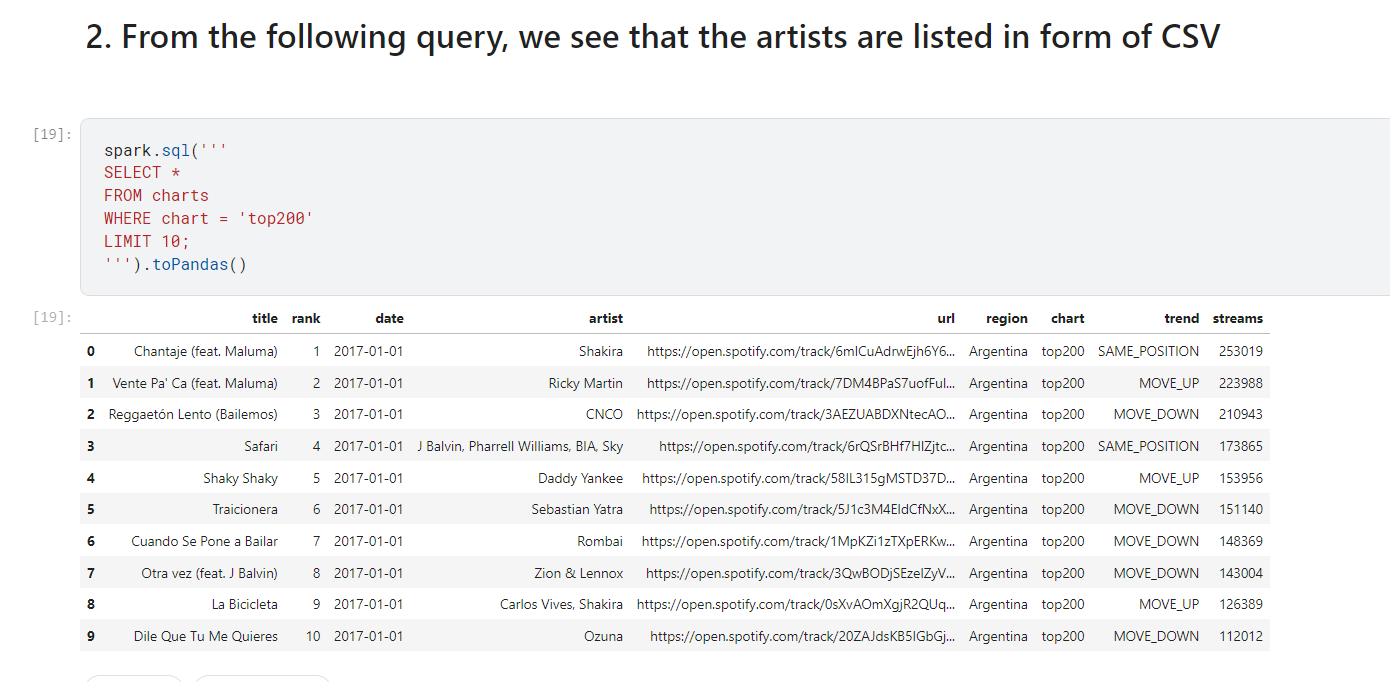
In order to examine the trends of different singers' songs, we performed Python, Pandas, Matplotlib, Spark and SQL on Kaggle to select the most popular singer 'Ed Sheeran' and the least popular singer 'Aspova' and performed data visualization analysis of three of their hottest songs.

First, we utilized Spark to perform exploratory data analysis to get the range of timeline of the data from January 1, 2017 to December 31, 2021.

Graphical user interface, text, application

Description automatically generated

Here is a sample of top 10 songs from ‘Top 200’ chart in January 1, 2017, all relevant information is presented below.



In order to get the total number of entries by all singers, we performed the ‘count ()’ function to get the number of 20,318,183.

A picture containing chart

Description automatically generated

Next, we created a word cloud visualization to compare the popularity of different singers. The size of the name is influenced by the number of times the singer’s name appeared on the ‘Top 200’ and ‘Viral 50’ charts.

Text

Description automatically generated

Afterwards, we implemented the ‘counts.head ()’ function and ‘counts.tail ()’ function to find the most popular singer and the least popular singer separately. Here we selected 'Ed Sheeran' as the most popular and 'Aspova' as the least popular one for further analysis.

Table

Description automatically generated

First, we created a pie chart to count the streams by region. The United States takes up 12.7% of streams, while Turkey only streams 1.1%. As a turkey rapper, it is hard for Aspova to achieve more streams than Ed Sheeran from the United States.

Chart, pie chart

Description automatically generated

Afterwards, we counted the number of times Ed Sheeran and Aspova appeared in the TOP 200 Charts. This can give us a preliminary understanding of the popularity of the two singers and the gap between them that Ed Sheeran appeared in the Top 200 300 times more than Aspova.

Graphical user interface, application

Description automatically generated

In the next statistics, in addition to the songs released by the singers themselves, we also counted the number of songs that the two singers cooperated with others in the TOP200 trend. To our surprise, Ed Sheeran is well-known, but the songs worked with other singers are not very popular. While the cooperation between Apsova and others can significantly improve the popularity of the song. To be more specific, Ed Sheeran and Aspova are both more successful solo and Ed Sheeran's songs are 500 times more popular than Aspova's songs.

Table

Description automatically generatedTable

Description automatically generated

Next, we counted the top 10 most popular songs of Ed Sheeran and Aspova and found that Shape of You of Ed Sheeran has 5 billion streams, while Eskimiş Senelere of Aspova only has 37 million streams. Besides, The difference between the top song of Ed Sheeran and Aspova is 135 times.

Table

Description automatically generatedTable

Description automatically generated

Then we counted the number of times top songs of Ed Sheeran and Aspova has appeared in top 200. Shape of You of Ed has appeared in the Top 200 100 times more than Aspova's Eskimiş Senelere. The top songs of Ed Sheeran meet the demands of audience better than that of Aspova since Ed’s songs last longer on the Charts. Audience prefers pop music of Ed Sheeran rather than rap music of Aspova.

Table

Description automatically generatedTable

Description automatically generated

Afterwards, the highest ranks the songs of Ed Sheeran and Aspova have attained. Shape of You is Ed Sheeran's most successful song. It stayed at #1 for 6 times longer than the second song Castle of the Hill. Aspova's Eskimis Senelere is his most popular song, but it only peaked at number 10 on one occasion. Susamam is a collaboration song of Aspova and it ranks 1, so collaboration is better for Aspova to be more popular than solo.

Graphical user interface, application

Description automatically generated

And then we calculated the highest, lowest and the mean rank of the songs of Ed Sheeran and Aspova. Ed Sheeran has much more songs in rank 1 than Aspova, and most songs of Aspova are around rank 50.

Table

Description automatically generatedTable

Description automatically generated

Finally, we created two visualizations to demonstrate the trend of ranking the three hottest songs by two singers from 2017 through 2022. We can find Songs of Ed Sheeran began in rank 1 and slowly decreased to rank 200, but still in the Top 200 chart.

Chart

Description automatically generated

However, the ranking of each song on Aspova has experienced wild fluctuations, and three songs demonstrate three different trends, which means songs with a steady increase in ranks are more likely to be popular. Although one of Aspova's songs in orange (approximately 200 in rank), all three songs by Ed Sheeran achieved similar rankings. Compared to Aspova, Ed Sheeran already has a stable of targeted audiences.

Graphical user interface, chart, line chart

Description automatically generated

Songs of Aspova began in rank under 100, then quickly increased to rank 1, then slowly decreased to rank 200 and then were out of Top 200 chart, Aspova should focus on the long-run marketing to keep his songs on top.

Comments

Some drawbacks are that Spotify may not be a good forum for sampling popular artists since Spotify is used more often in richer countries that have easy access to the internet and wifi, which means findings may only be generalized to these kinds of countries. Our analysis is also limited by time. Ideally, more than two artists (likely the 30 highest and lowest ranking artists in the top 200) in order to analyze changes in song popularity over time with a greater degree of generalizability. Challenges with the dataset primarily have to do with the limited number of fields. More sophisticated analysis could be done if attributes like genre, sub-genre, recently toured, recently promoted, recent appearance in media, etc. were included as we could then look to see which characteristics were associated with popularity. There is also some multicollinearity in the data as streams and trending are associated with each other, but this was not problematic as only streams had been used for our analysis.

Conclusion

Our project sought to compare features between the most and least popular artists in Spotify’s top 200 in order to provide some context for how popular a given artist is. We have achieved this to a certain extent as the minimum and maximum values within the top 200 dataset have been flushed out, and thus our analysis has specified the range that any artist in the dataset can take. Benefits to this approach include the ample sample size associated with the data, which makes our findings more robust. Our analysis is also straight forward which increases its accessibility and minimizes the likelihood of error.

In conclusion, Ed Sheeran appeared in the Top 200 300 times more than Aspova and Ed’s songs appeared 6 times more than Aspova’s songs. Aspova should focus more on the quality of his songs and marketing than the number of songs. Ed Sheeran performs better in solo songs, while Aspova's most popular song Susamam is a collaborated song. Aspova should collaborate with famous artists to make his music product much more popular. Songs by Ed Sheeran are in English and have been popular all over the world, but Aspova's songs are in Turkish, so they were only popular for a short period of time. Therefore, Aspova should publish more songs in English to join the global music market and focus on the long-run marketing to keep his songs on top.

Appendix

We publish our code at this link on Kaggle: <https://www.kaggle.com/code/vhtrieu/aly-6110-spotify-charts-analysis/notebook>

References

Dave, D. (2022). Spotify Charts. *Kaggle.*

<https://www.kaggle.com/datasets/dhruvildave/spotify-charts>