Final Project Report: Spotify Exploratory Data Analysis

MSBX5420 - Unstructured and Distributed Data Modeling & Analysis

By: Patrick Xiao

Disclaimer: While a majority of the work here is my own work, ChatGPT was used in drafting some areas of the report

I. Abstract

This project aims to analyze the spotify charts dataset to gain insights into popular artists and songs, such as trends and regional differences. Using Pyspark on an AWS EMR cluster, various queries, aggregations, and visualizations were done to unravel meaningful narratives and patterns in song popularity, with a focus on Ed Sheeran. The analysis covers a range of data from 2017 to 2021, exploring questions such as rankings of songs, regional differences, and distribution of streams.

II. Background

As a frequent Spotify user that listens to a wide variety of genres and songs, I was interested in understanding the trends and patterns of the music that I frequently listen to. I often find myself making a new playlist every month and spending several hours per day with my headphones plugged in. With new artists, songs, and genres of music rising in popularity over the last couple of years, the music industry has been an area that I've been following with great interest. The motivation behind this project was to apply the skills and concepts covered in this course to perform big data analytics and to extract meaningful insights from a large dataset related to that specific area .

III. Dataset and Analysis

The dataset used in this project was pulled from Kaggle, titled "Spotify Charts". The dataset includes information on song titles, artists, ranks, dates, regions, and charts such as "Viral 50" or "Top 200". The data, approximately 3.48 GB in size with 164,807 unique values, was processed using PySpark on an AWS EMR cluster. When importing data and reading from the CSV file, the inferred data type(s) may not be fully accurate. To ensure consistency and better accuracy, I transformed the dataframe using the following format: The "Rank" column was cast to a long integer (LongType), the "Date" column was cast to a date (DateType), and the "Streams" column was cast to an integer (IntegerType).

A general overview of my analysis and the different questions I drove insights from:

- The time range of the dataset
- The number of regions in the data
- The top 10 songs that appeared in the top 200 between 2017 and 2021
- The regions with the highest number of unique song titles on the top 200 chart
- Most streamed song in each of the 6 most unique regions
- Plotting total streams of top 20 songs in the top 200 chart
- Total number of songs by Ed Sheeran that has appeared in the top 200
- Total number of times Ed Sheran appeared in the top 200
- Plotting the average streams of song by region
- Listing Ed Sheeran's top 10 songs
- Number of days each song was on the top 200 chart
- The highest ranks the songs have attained and frequency of how often they have hit that milestone
- Plotting the trends of Photograph, Shivers, and Shape of you over the course of time
- Plotting the trends of same songs in the viral 50 chart
- The average rank of the songs
- How streams are distributed across different regions for a specific song

The first part of my analysis of the Spotify Charts dataset involved running SQL queries and visualizing trends for top songs and artists on the Top 200 chart. There are many approaches to performing an analysis on the dataset, but I chose to tailor it to my personal interest by doing the second part of my analysis on Ed Sheeran and some of his top songs to analyze performance over time. For the last part of the analysis, I aggregated the number of average streams for a specific song to see how they are distributed across regions. In order to demonstrate some level of potential for horizontal scaling and whether or not my project implementation takes advantage of distributed computing, I measured the speed at which the task took to complete under the initial configuration of 5 executors. I then repeated the same task but with a new configuration of 12 executors and once again, measured the speed it took to complete the process.

IV. Results and Insights

The analysis revealed several key insights that I found to be very interesting. Between 2017 and 2021, the region with the highest number of unique song titles on the top 200 chart was Switzerland. I was interested in finding out how the total number of streams for some of the top songs on the Top 200 charts looked so I visualized the total number of streams of the top 20 songs:

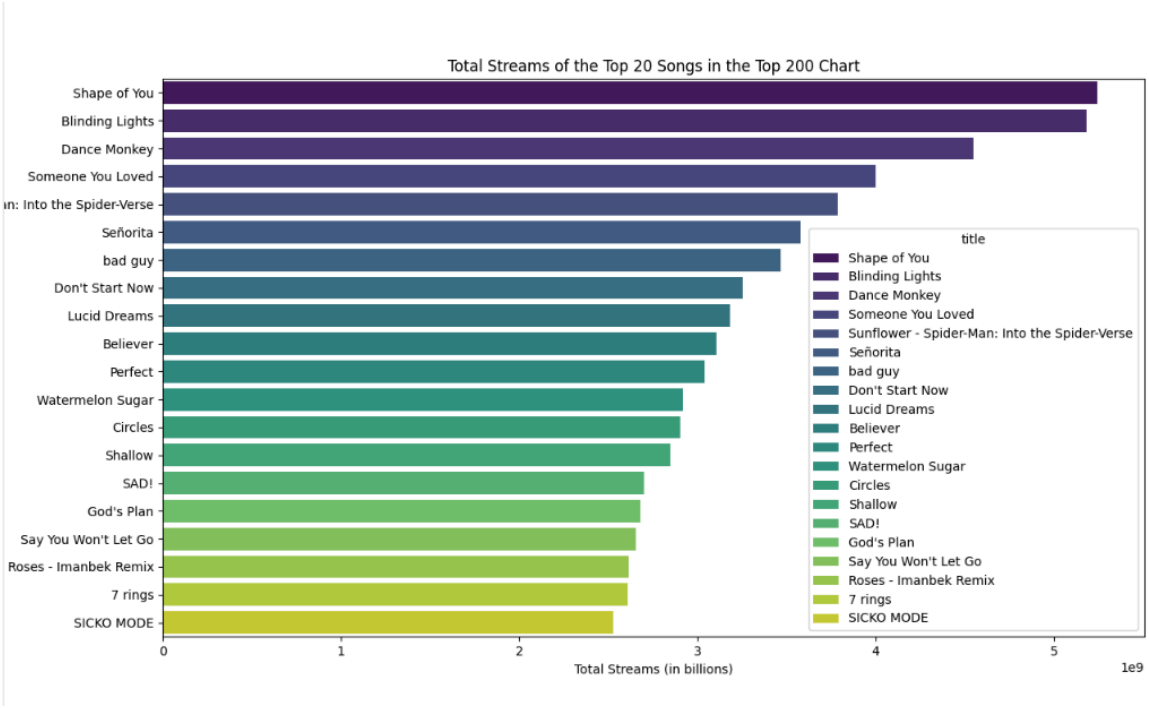

Figure 1: Visualization of the total streams of the top 20 songs in the Top 200 chart

From Figure 1, we can see that Shape of You was the most streamed song at over 5 billion total streams. Ed Sheeran has 121 songs that has appeared in the Top 200 chart while he has appeared a total of 368,388 times in the Top 200 chart. Out of those 368,388 times, 366,026 are solo appearances while the rest are with other artists such as Taylor Swift, Tori Kelly, The Weeknd, etc. What I found to be surprising (and a little disappointing) was that out of his top 10 songs, “Photograph” was relatively lower on that list, with 1,081,454,379 streams compared to his top song “Shape of You” with 5,245,740,051 streams. Additionally, “Shape of You” has been on the Top 200 chart for 65,262 days while "Photograph” has been on it for 28,605 days. Hitting rank 1 is a milestone in itself, but “Shape of You” has hit that rank a total of 8,565 times between 2017 and 2021. Below is a plot of the trends of three of my favorite songs by Ed Sheeran over the course of time.

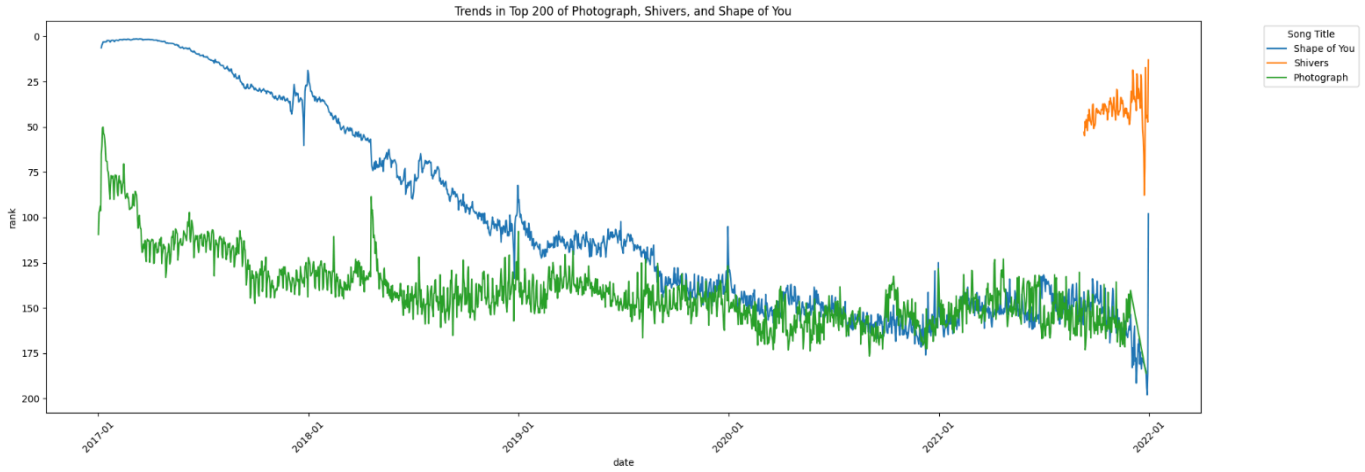

Figure 2: Visualization of the trends in the Top 200 chart for "Photograph", "Shivers", and "Shape of You"

We can draw several conclusions based on what the trends look like between 2017 and 2021. "Shape of You" has been trending in the Top 200 chart from the beginning of 2017 until the start of 2022 (dataset timeline cutoff) with a resurging rise of ~120 ranks in early 2022. "Photograph" has been trending for a similar duration as "Shape of You". Despite starting at a lower rank, it eventually evens out with "Shape of You" around early to mid 2022. Lastly, "Shivers" has been trending for a much shorter duration compared to "Shape of You" and "Photograph", but has already peaked at a higher rank than "Photograph".

After reviewing how the trends look for “Shape of You”, “Photograph”, and “Shivers”, I was curious to see how those trends might differ in the Viral 50 chart over time

.

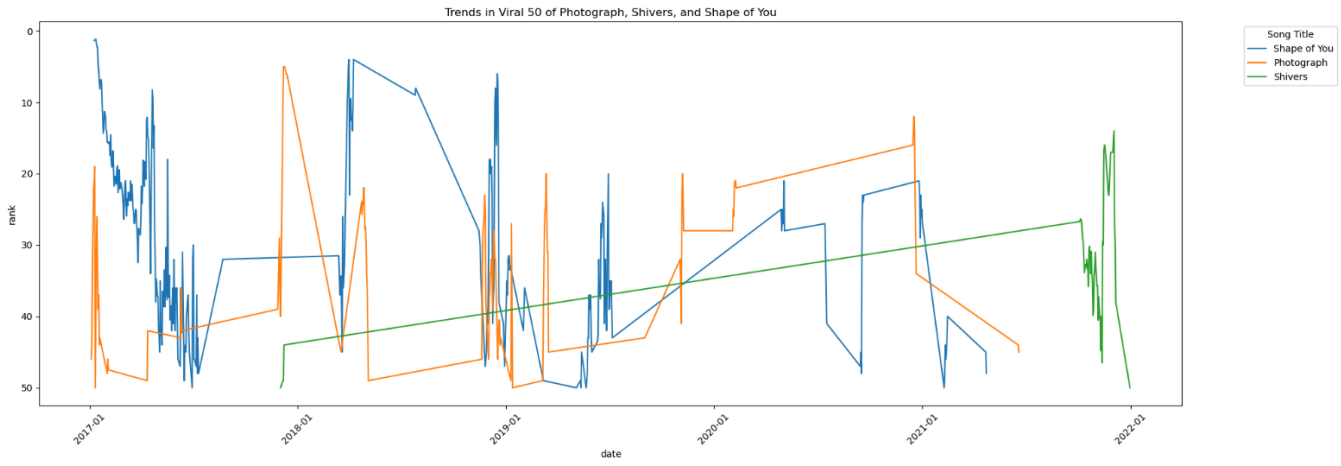

Figure 3: Visualization of trends of “Shape of You”, “Photograph”, and “Shivers” in the Viral 50 chart

From the visualization, we can infer that both “Photograph” and “Shape of You” have been trending for a similar curation in the Viral 50 chart as the Top 200 chart, fading in popularity around mid 2021. “Shivers” started to trend in the Viral 50 chart in late 2017 with a steady increase until late 2021, where it peaked at rank ~18 before dropping down to rank 50 in 2022. The Top 200 chart suggests “Shivers” has been trending for a much shorter period of time but is ranked higher than both “Shape of You” and “Photograph” around early 2022. However, the Viral 50 chart suggests that “Shivers” started trending a while later but for roughly the same total length of time as the other two songs. Diving a little further into the songs, I wanted to see how the 3 songs perform individually on the Top 200 chart on average. The results showed the following ranks, suggesting that to date (within the dataset), “Shivers” is the best performing song out of the three.

- “Shape of You” at an average rank of 80.53
- “Shivers” at an average rank of 41.90
- “Photograph” at an average rank of 128.14

The last part of my analysis was to see how streams are distributed across different regions for my favorite song by Ed Sheeran, “Photograph”. The results showed that the United States had by far the most number of average streams (351,961) and Brazil (60,168), the United Kingdom (56,740), and Mexico (49,977) following in it’s lead respectively.

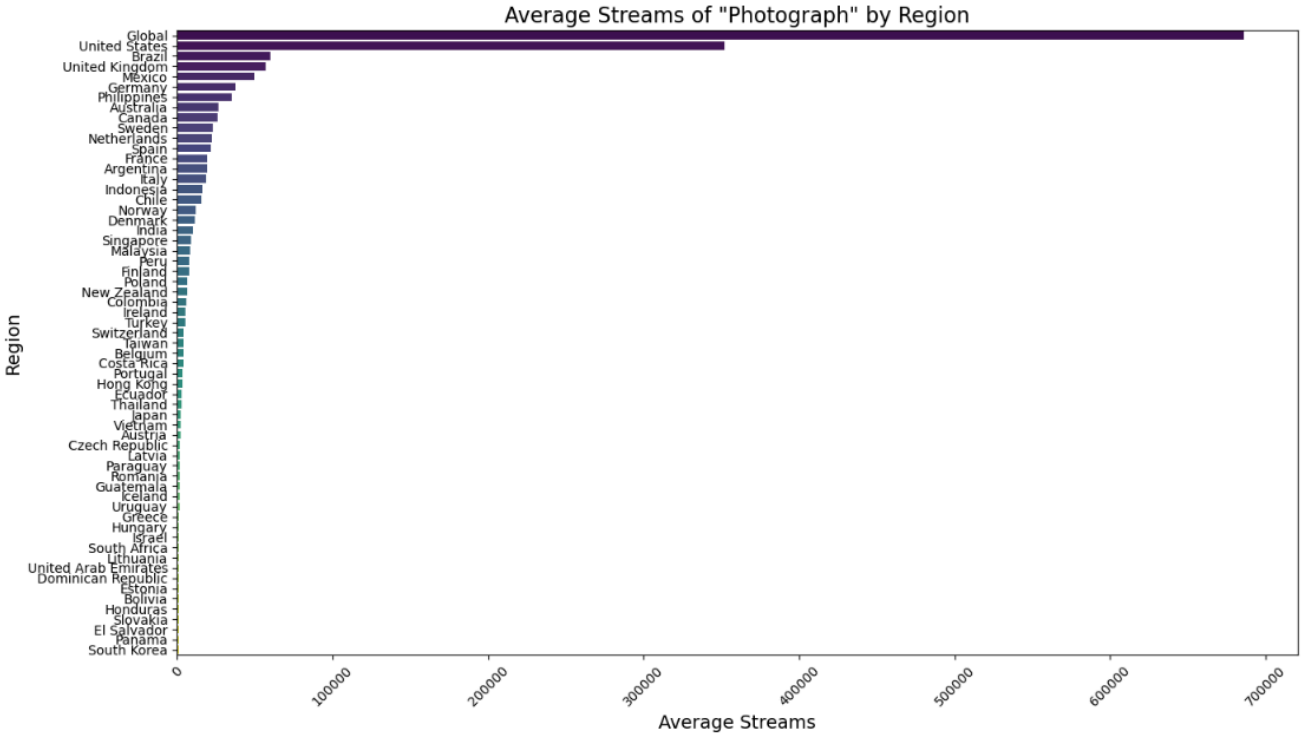

Figure 4: Visualization of the number of average streams of “Photograph” by region

V. Conclusion and Implications

To demonstrate some potential of horizontal scaling and the advantages of distributed computing, I measured the speed of the last task (distribution of streams across different regions for “Photograph”) under two different configurations. The default configuration consisted of 5 executors while the new configuration increased this number to 12 executors. Recognizing that simply increasing the number of executors does not always guarantee an increase in processing speed and time, I also increased the parameter for executor memory from 1G to 1.5G. Increasing the number of executors and the memory allocated to each executor can significantly impact the processing time of tasks in a distributed computing environment. More executors mean that more data partitions can be processed, leading to faster overall job completion times and with improved resource allocation from an increase in executor memory, each executor is allowed to handle bigger partitions of data without running into any memory issues. From the statistics we obtained after measuring the time for our task to complete, we saw an improvement from 15.82 seconds to 11.95 seconds.

There are other configurations that would have allowed for a more significant increase in speed. One improvement that could have been made is to save our data as a Parquet file. Because Parquet files are able to be split, larger files can be divided into smaller segments of data that can be processed in parallel by multiple executors. This along with the columnar storage format of Parquet files improves the efficiency and scalability of data processing tasks.

The results of this analysis provided interesting insights into the music industry and consumer behavior. Understanding regional preferences can help growing artists target and adjust their marketing strategies. In the case of Ed Sheeran, the trends observed for his songs offer a case study on how certain songs can achieve sustained popularity and how that popularity can vary geographically.