Music Industry Analysis with Spotify

CS-GY 6513 Big Data Spring 2022 Final Project Report

Team Members
Abhishek Nandan Mishra (anm9189)
Lakshana Kolur (lk2719)
Shreeraj Pawar (srp8095)

Table of Contents

Introduction and Problem Statement

Imagine wanting to get started in the music industry. You believe you have a skill for spotting great talent and promoting them to become stars.

Yet, apart from your ingenious ability, you need to be familiar with facts.

Almost all artists in the modern age have a presence on Spotify and an important aspect of their fame, popularity is their reach and engagement on this brilliant music application.

In this project we will explore the top charts of Spotify across various regions around the world and try to glean some insights into the Music industry by performing distributed operations on the charts dataset.

We will also attempt to apply some Machine Learning techniques using the very powerful Spark ML libraries which can help us in building insightful ML models which would help in predictions, classification and regression.

About the Data

Spotify releases Top 200 Charts & Viral 50 Charts every 2 days. This can be accessed using Spotify APIs. It gives us information about:

- Regions where the charts are topping - Argentina, Paraguay, Global, United States
- Various artists along with their titles
- Total amount of data: 3.48GB

Why Big Data ?

- Data is updated every 2 days, it keeps accumulating and it's not possible to use the single node solutions for analysis.

- As the number of years increase, the scale of the data increases

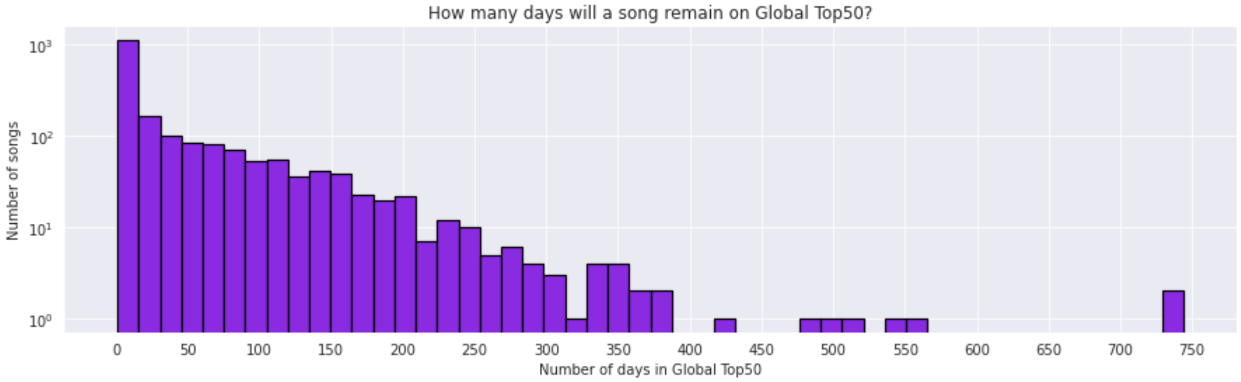
- To perform analysis with machine learning, with a larger number of features, we would need Big Data infrastructure.

Architecture

To perform scalable analytics and machine learning we used PySpark and the libraries it comes with as well as other visualization libraries like plotly, matplotlib and seaborn. The aforementioned dataset can be ingested into PySpark whenever the visualization and prediction needs to be performed.

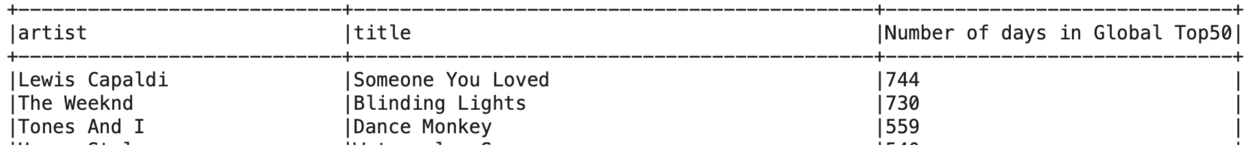
I am on top. But for how long?

We wanted to ask the question: How many days does a song stay in the Global Top50?



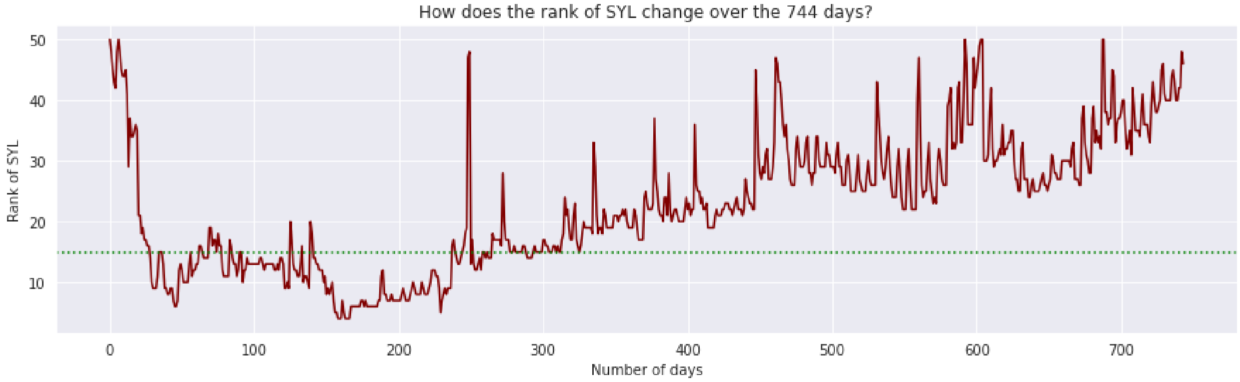
As we expected most songs were there for just about one or two months (notice the log scale on the y-axis). There are some songs which were on the Global Top50 for over two years.

So let's see what were these top songs -

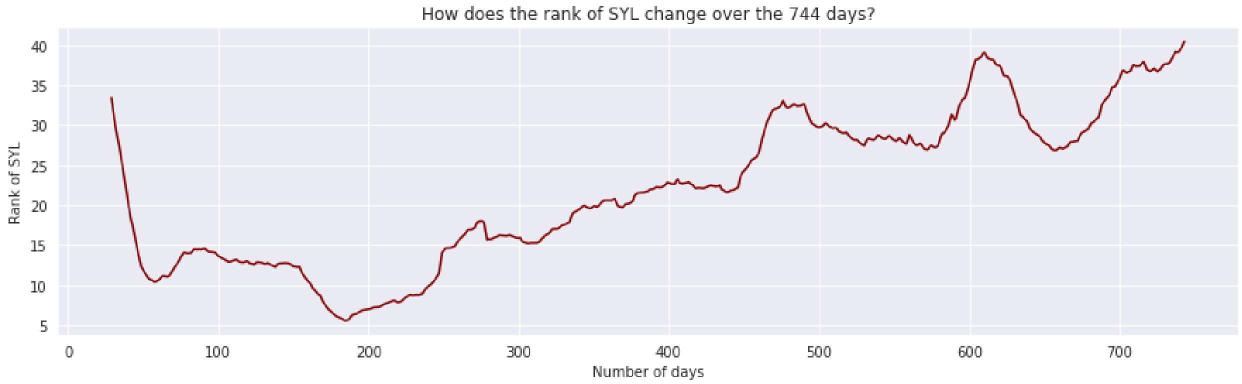


```
+-----------------------------+---------------------------------------------+------------------------------+
|artist                       |title                                        |Number of days in Global Top50|
+-----------------------------+---------------------------------------------+------------------------------+
|Lewis Capaldi                |Someone You Loved                            |744                           |
|The Weeknd                   |Blinding Lights                              |730                           |
|Tones And I                  |Dance Monkey                                 |559                           |
```

It’s been in the top charts for over 2 years - but how does the rank change over time?

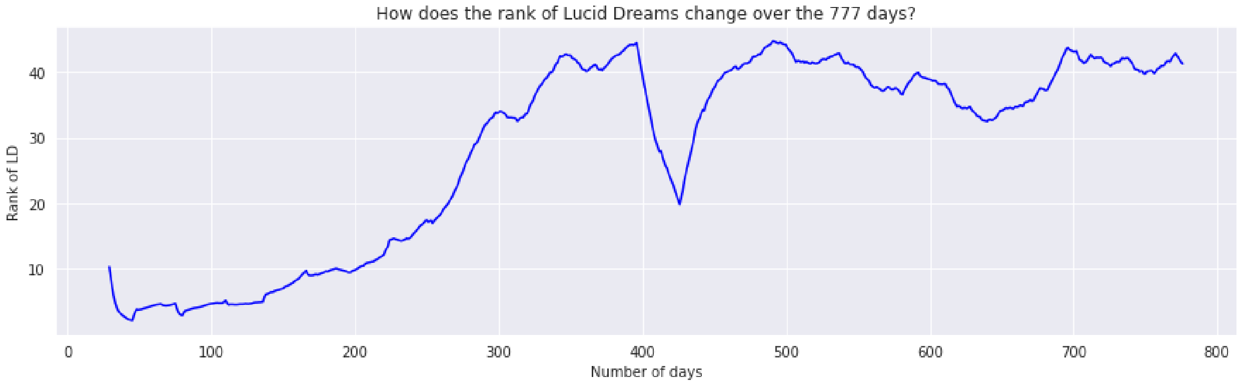


- At about 325-350 days (~ 1 year), the song's rank started dropping to mid-range ranks - around 25
- But this plot looks messy. Perhaps we could look at moving monthly averages for rank, instead of daily rankings.



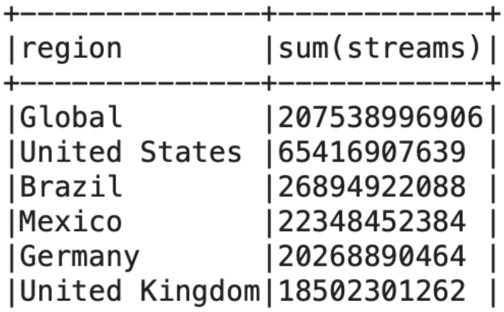
Look at different regions - how does this compare to Global Charts?

Let's look at United States next :



How the same song performed globally -

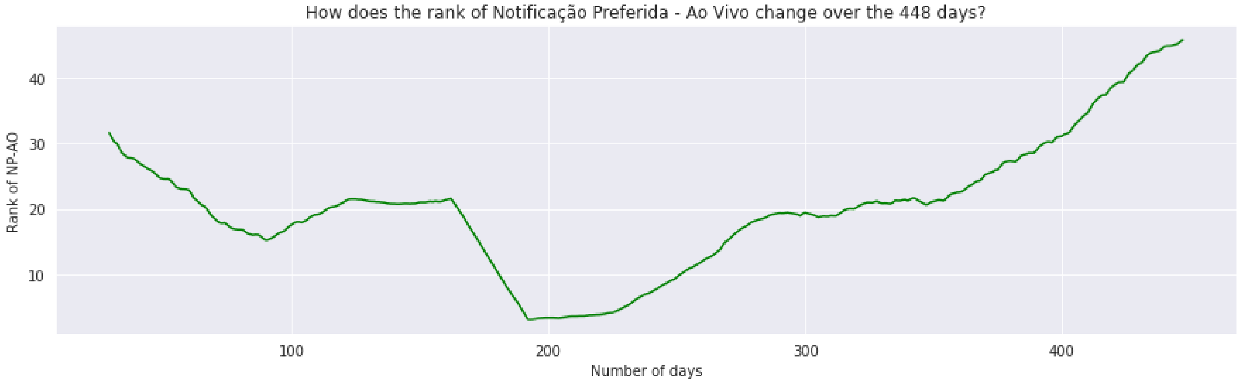
Started off with lower ranks, reached peak early. Compared to Global peak ~200day



```
+--------------+------------+
|region        |sum(streams)|
+--------------+------------+
|Global        |207538996906|
|United States |65416907639 |
|Brazil        |26894922088 |
|Mexico        |22348452384 |
|Germany       |20268890464 |
|United Kingdom|18502301262 |
```

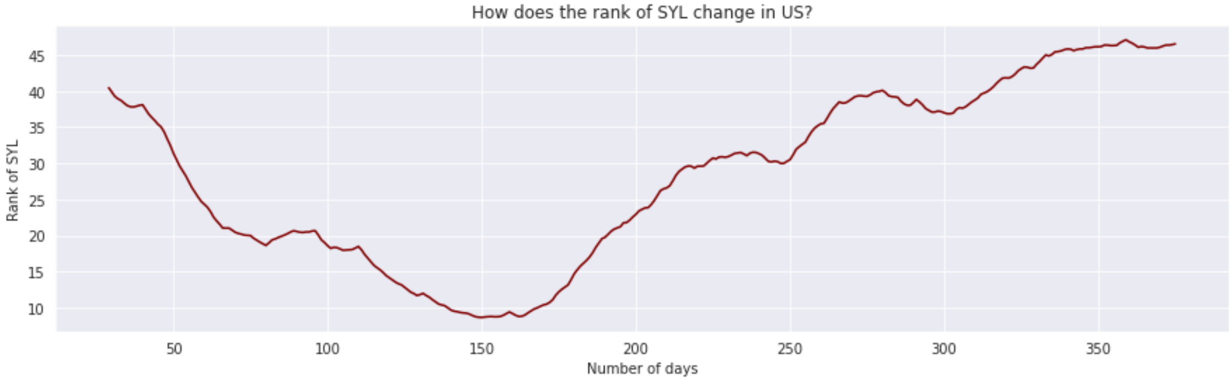
Songs become popular more quickly locally, than globally.

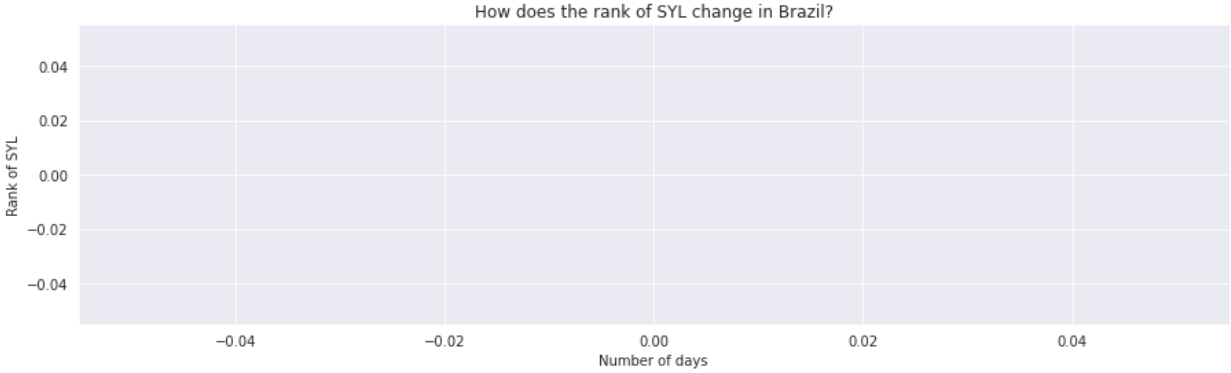
Look at different regions - how does this compare to Global Charts?
Let's look at Brazil next :



- Portuguese song is top (as expected)
- 448 days top charts soaring
- Same as Global peak ~200day

Conversely, how does ‘Someone You Loved’ do in the US and Brazil?





The song was not stream at all in Brazil.

SYL stayed on top of the global charts for over 2 years. But the popular songs vary based on countries. US and Brazil had very different songs that were most listened to. Hence, popularity is local.

In addition to that, even though a song is in the top charts, that doesn’t imply that it is consistent. The trajectory of the song rank varies based on local charts or global. Specific to a country, songs become popular very quickly and their rank has a steady decrease, whereas for global ranking, the songs come into these charts after they have performed well in all the country-centric charts and hence, reach their peak after approx.. 6 months.

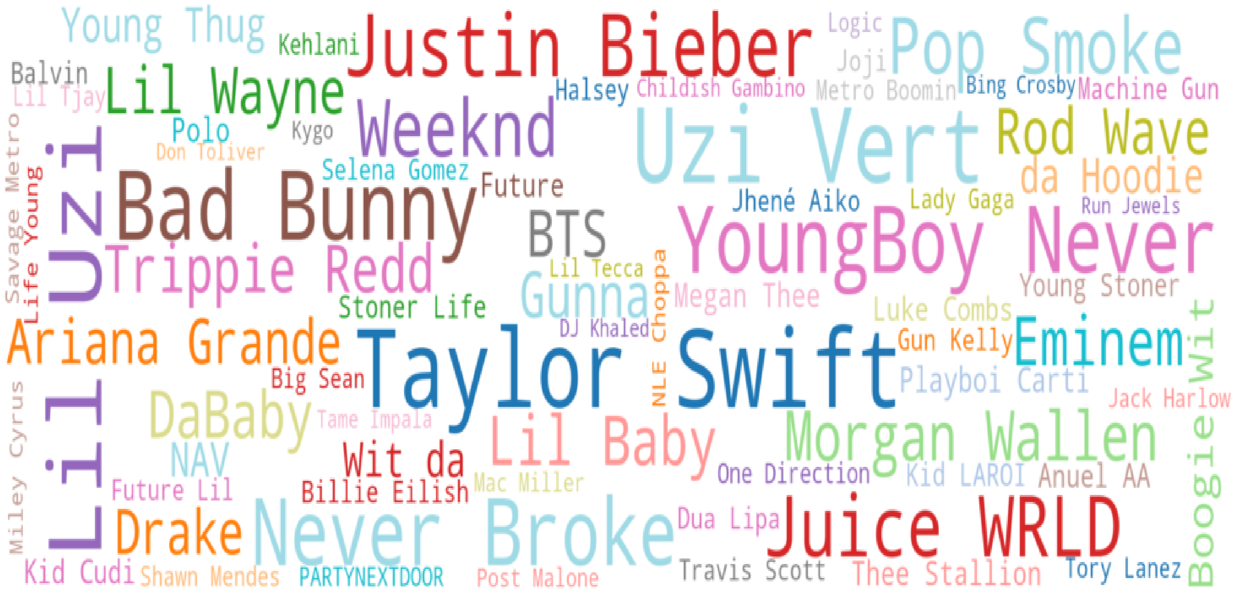
How is the Pandemic affecting popularity in the US?

- We wanted to explore the effects of the COVID 19 Pandemic on the top charts and how it has affected the popularity of songs and artists in the US.

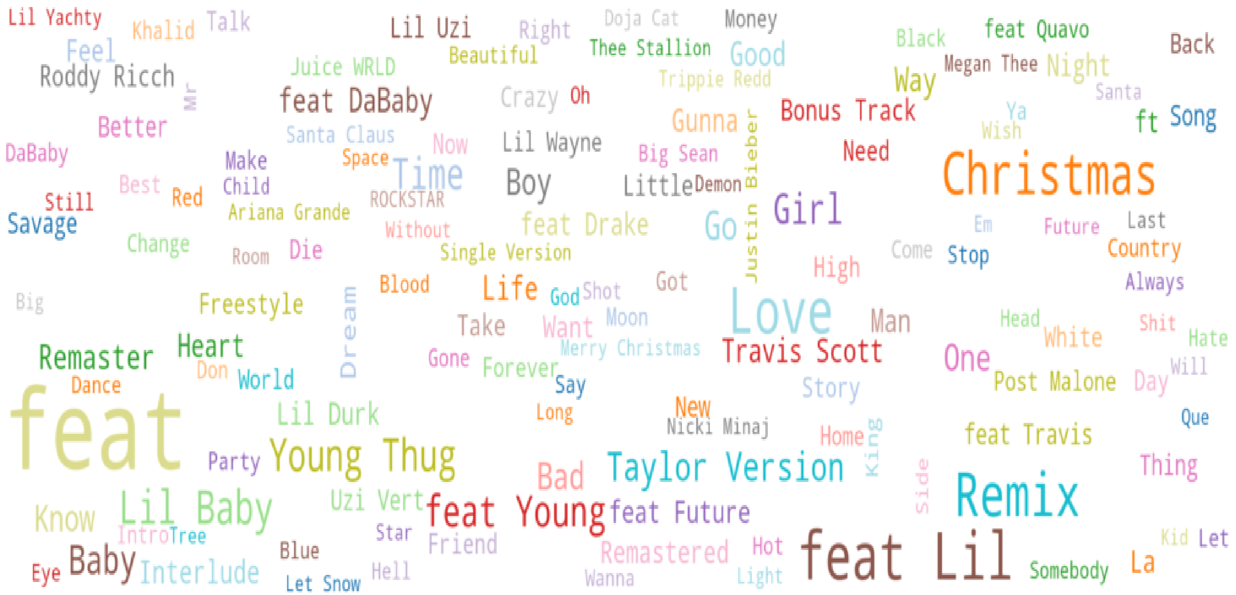
- For this we followed an approach where we examine the trends of populars songs during the pandemic.

- We see the trend of the top 3 songs in the US in three different time periods - pre-pandemic, post-pandemic and during pandemic.

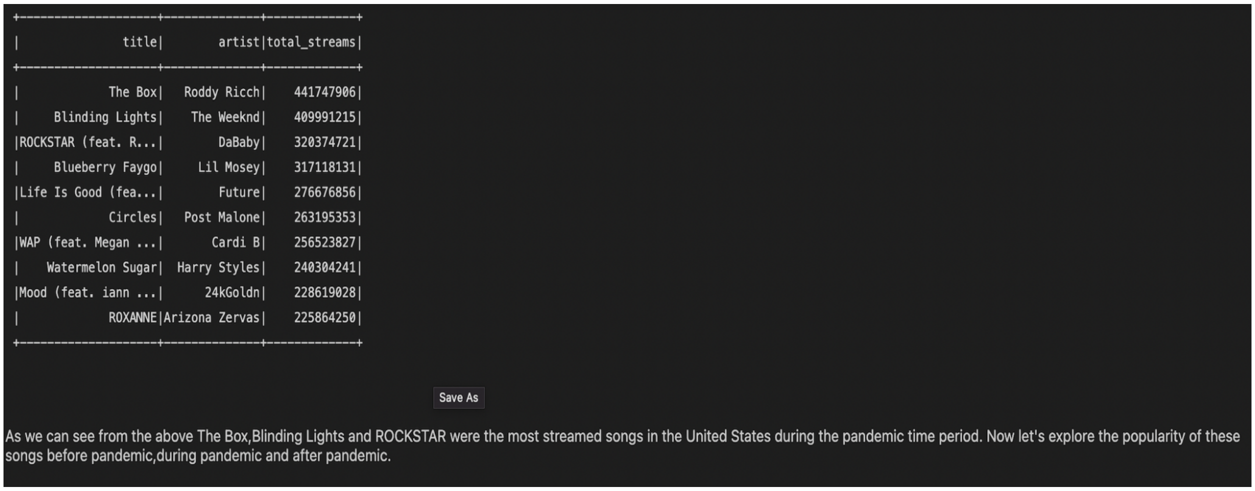
We used the pyspark world cloud API to generate a world cloud for songs that were popular in the pandemic in US. Larger font in the world cloud represents more number of streams.



Artist WordCloud



Song WordCloud

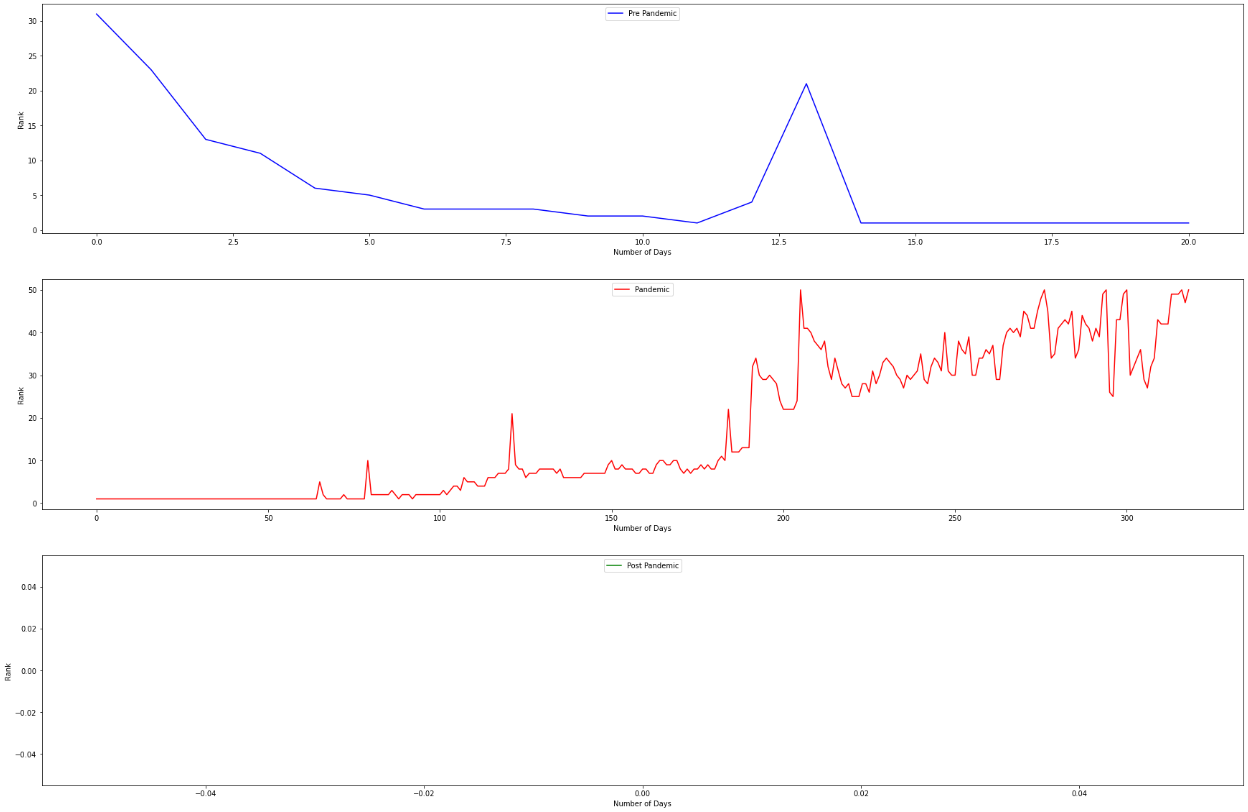


```
+--------------------+--------------+-------------+
|               title|        artist|total_streams|
+--------------------+--------------+-------------+
|             The Box|   Roddy Ricch|    441747906|
|     Blinding Lights|    The Weeknd|    409991215|
|ROCKSTAR (feat. R...|        DaBaby|    320374721|
|     Blueberry Faygo|     Lil Mosey|    317118131|
|Life Is Good (fea...|        Future|    276676856|
|             Circles|   Post Malone|    263195353|
|WAP (feat. Megan ...|       Cardi B|    256523827|
|    Watermelon Sugar|  Harry Styles|    240304241|
|Mood (feat. iann ...|      24kGoldn|    228619028|
|             ROXANNE|Arizona Zervas|    225864250|
+--------------------+--------------+-------------+
```

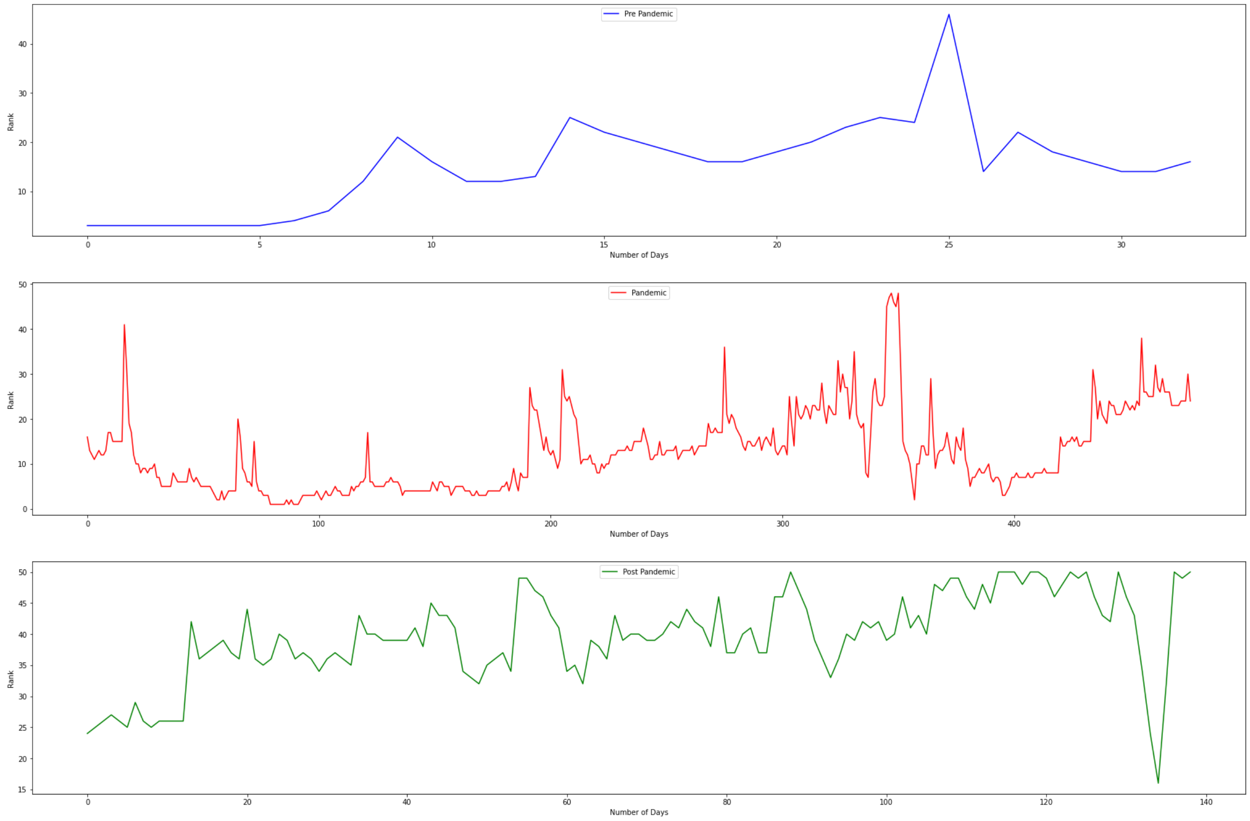
As we can see from the above The Box,Blinding Lights and ROCKSTAR were the most streamed songs in the United States during the pandemic time period. Now let's explore the popularity of these songs before pandemic,during pandemic and after pandemic.

As we can see from above, The Box and Blinding Lights and ROCKSTAR were the most streamed songs in the United States

Now, let’s explore the popularity of these songs before pandemic, during the pandemic and after pandemic.



Trends for The Box By Roddy Ricch



Trends for Blinding Lights by The Weeknd

The previous analysis was for the United States and we noticed that a lot of the singers were rappers and highly streamed songs during this time was the rap and hip hop genre. This hints at a possible return to rap and hip hop culture during the tough times of the pandemic in the United States.

Let's see the case globally as to which artists are dominating and where they are most popular. Initially let's compare the top 3 pandemic artists and get their global ranks in terms of number of streams.

Artist	Global Rank
The Weeknd	1
Dababy	15
Roddy Ricch	20

Around the World

Here we explore what's happening around the world and for an artist does the origin or language used to write the song matter. To do so we start with an american pop rapper Post Malone and compare these songs based on number of streams and global 200 chart to see which parts of the world does the artist have.

We then use a choropleth map to visualize the results. The figure below shows Post Malone has more influence or success in America. The red color indicates more streams and popularity as compared to Blue.

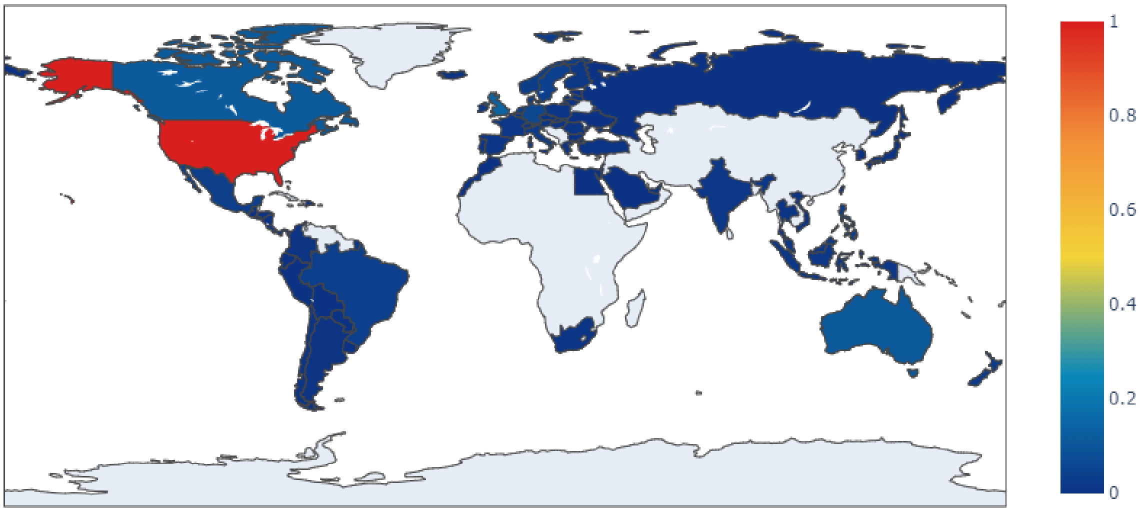


Figure 5

Post Malone being an American born artist, it is possible that the distribution above is more concentrated in America. So now we try some other artists from another country.

The artist of choice this time was One Direction. We chose this artist as the band originates from London, UK with English-Irish band members. Contrary to Figure 5 above we see a different trend with One Direction having more influence over Asian countries especially in Thailand - Malaysia region (South Asia)

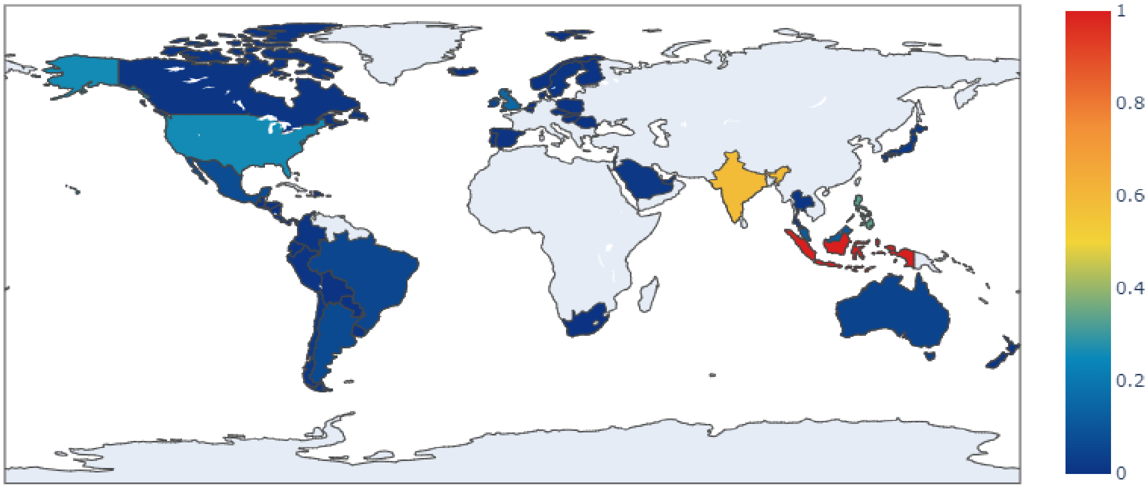


Figure 6

Till now we looked into artists that produce songs in English language. To see if language is a factor in play we looked into the korean boy band BTS. Hailing from South Korea, BTS produces the majority of its songs in Korean.

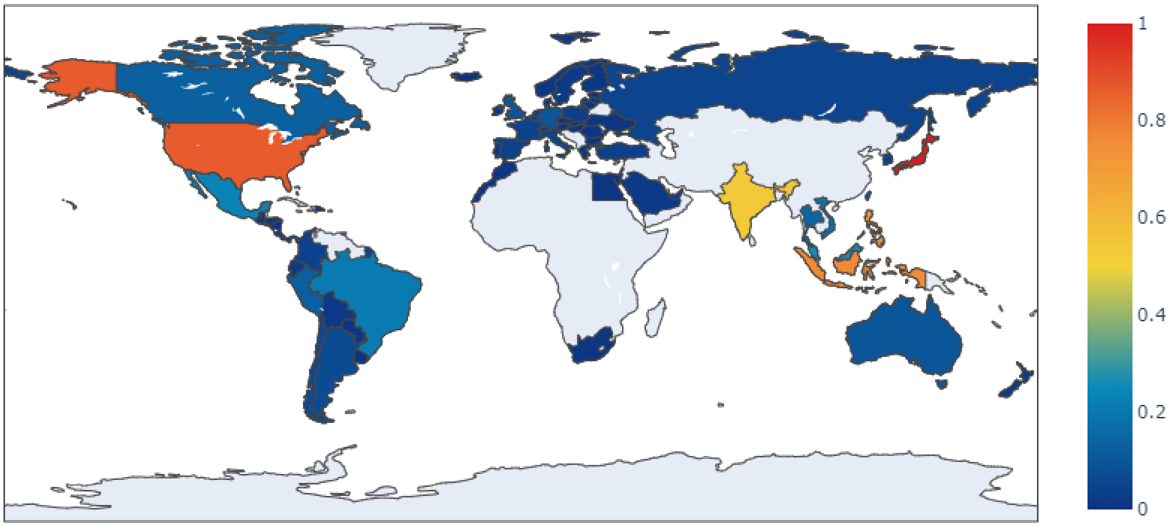


Figure 7

Figure 7 shows us the global dominance of BTS. Even though their songs are in Korean it hasn't hindered their influence over the world. We can conclude that language or country of origin is not a barrier for success or popularity of an artist.

What is the trend of songs in the Top Charts?

- Leveraged Spark ML Library to apply ML techniques.
- One of the use cases for ML : Using Logistic Regression to do multi- class classification on the trend column.
- Trains the model based on the rank and number of streams column while using trend (MOVE_UP, MOVE_DOWN and SAME_POSITION) as a label.
- The model can be used to predict future values on the trend of a particular song based on the rank on that date and number of times it was streamed on that day.

ML Model

- We used the Logistic Regression with LBFGS API in PySpark. It is used to leverage multinomial/binary logistic regression using Limited Memory BFGS.
- It uses standard feature scaling and L2 regularization methods.

Mariah Carey Christmas Spirit

- Mariah Carey's famous song “All I Want for Christmas is You” is an all time holiday classic and we wished to explore the trends based on our trained model.
- For our first prediction we assume our song has a rank of number 63 and streams as 68087 and we get a trend of “MOVE_DOWN”.
- For values of rank number 4 and stream number 3489570 we get a trend of “SAME_POSITION”

Future Work -

- We can incorporate region into the prediction as we notice that trend depends on the region too.
- In the Christmas of 2018 Mariah Carey’s “All I want For Christmas is You” was ranked number 63 in India as Christmas is not such a big festival in India and access to Spotify was limited at that time.
- However at the same time in the United States we obtain a rank of 4 and a trend of “SAME_POSITION” as the Christmas Spirit is in full swing during this time.

References

1. https://www.kaggle.com/datasets/dhruvildave/spotify-charts
2. https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35