Spotify in Spark

By: Niclas Allison

The Problem

- For the project I performed EDA on Spotify data that I got from Kaggle. This data contains charts that show the top 200 and top 50 songs from most if not all countries from dates ranging from January 2017 to February 2022.
- My goal of this project was to find the top 20 songs and artists for each month of each year for each country that I chose as well as the top 20 for each year for the selected countries.
- Do to the size of the file being 3.48GB I would need to cut down the data so it is more manageable.

The Solution

- So when deciding to cut down the data I choose to only keep the top 200 songs that begin at January 2017 and that I would only use 10 of those countries.
- I also chose to keep an eleventh region which the data had a global value that contains a sum of the top 200 songs across all countries, and I kept it so I could compare how ten countries compares to global value.
- To view the top 20 songs a Bar graph is used after the data has been adjusted and sorted correctly so it will be able to display in descending order.

Example

This is the raw data for a specific just to show the streams despite it being a little difficult to see the streams increase and decrease so I was able to use that as the key device when determining the order of the songs.

```
dim.filter((dim.title == 'Bad and Boujee (feat. Lil Uzi Vert)') & (dim.region == 'United States') & (dim.artist == 'Migos') & (dim.date.startswith('2017-01'))).show(10)
```

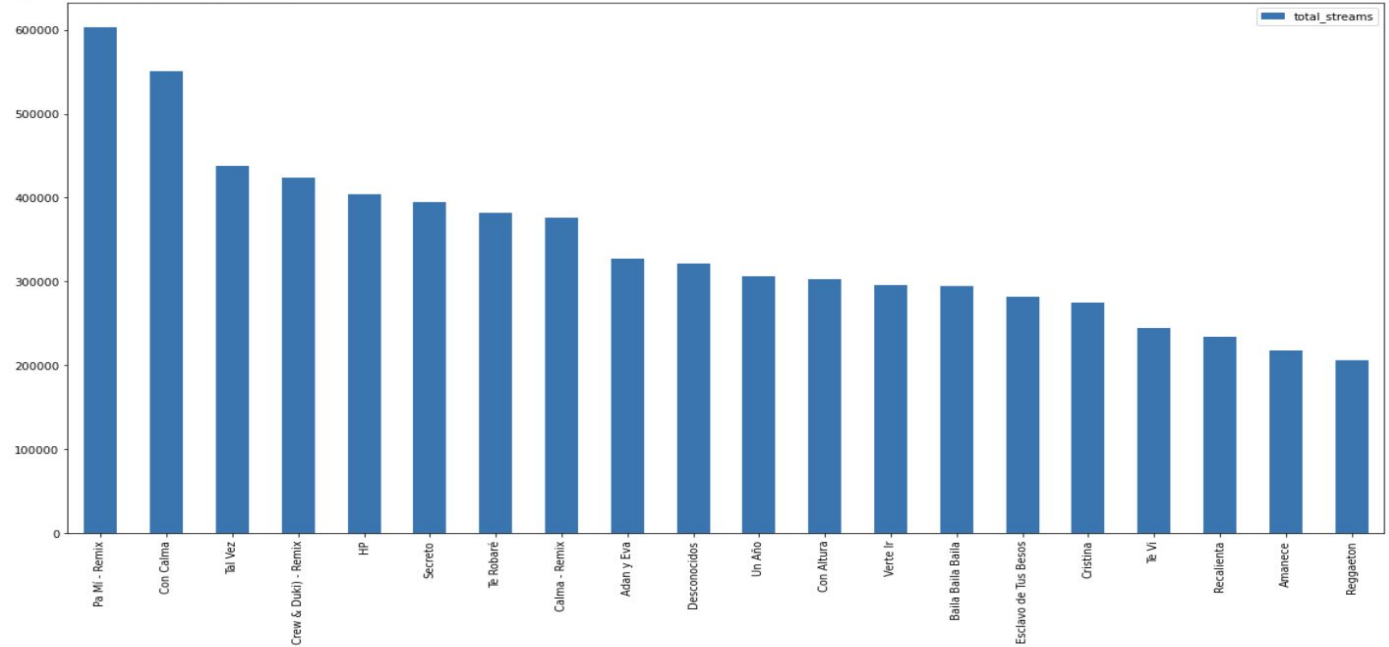
```
+--------------------+----+----------+------+--------------------+-------------+------+-------------+-------+
|               title|rank|      date|artist|                 url|       region| chart|        trend|streams|
+--------------------+----+----------+------+--------------------+-------------+------+-------------+-------+
|Bad and Boujee (f...|   1|2017-01-01| Migos|https://open.spot...|United States|top200|SAME_POSITION|1371493|
|Bad and Boujee (f...|   1|2017-01-02| Migos|https://open.spot...|United States|top200|SAME_POSITION|1161746|
|Bad and Boujee (f...|   1|2017-01-03| Migos|https://open.spot...|United States|top200|SAME_POSITION|1284891|
|Bad and Boujee (f...|   1|2017-01-04| Migos|https://open.spot...|United States|top200|SAME_POSITION|1293486|
|Bad and Boujee (f...|   1|2017-01-05| Migos|https://open.spot...|United States|top200|SAME_POSITION|1295592|
|Bad and Boujee (f...|   3|2017-01-06| Migos|https://open.spot...|United States|top200|    MOVE_DOWN|1328631|
|Bad and Boujee (f...|   2|2017-01-07| Migos|https://open.spot...|United States|top200|      MOVE_UP|1281623|
|Bad and Boujee (f...|   2|2017-01-08| Migos|https://open.spot...|United States|top200|SAME_POSITION|1189435|
|Bad and Boujee (f...|   2|2017-01-09| Migos|https://open.spot...|United States|top200|SAME_POSITION|1465579|
|Bad and Boujee (f...|   1|2017-01-10| Migos|https://open.spot...|United States|top200|      MOVE_UP|1554966|
+--------------------+----+----------+------+--------------------+-------------+------+-------------+-------+
```

The 5 Steps

1. First is loading in the csv file and filtering down to the 10 countries, top 200 songs, as well as editing the date so that month of a year and year.
2. Second is summing the data based on the streams column.
3. Third a Window is created so the data is can be reordered so that the total streams are in descending order while the date is in ascending order as well as sets up the top 5 and top 20 songs and artists.
4. Fourth the top 5 and top 20 songs are converted into Pandas and the top 5 for each dataframe is printed out and saved as a jpg.
5. Lastly there is a function to that allows you to see what every country for a given year or month of year by following its steps it will produce a bar graph that will display the top 20 songs.

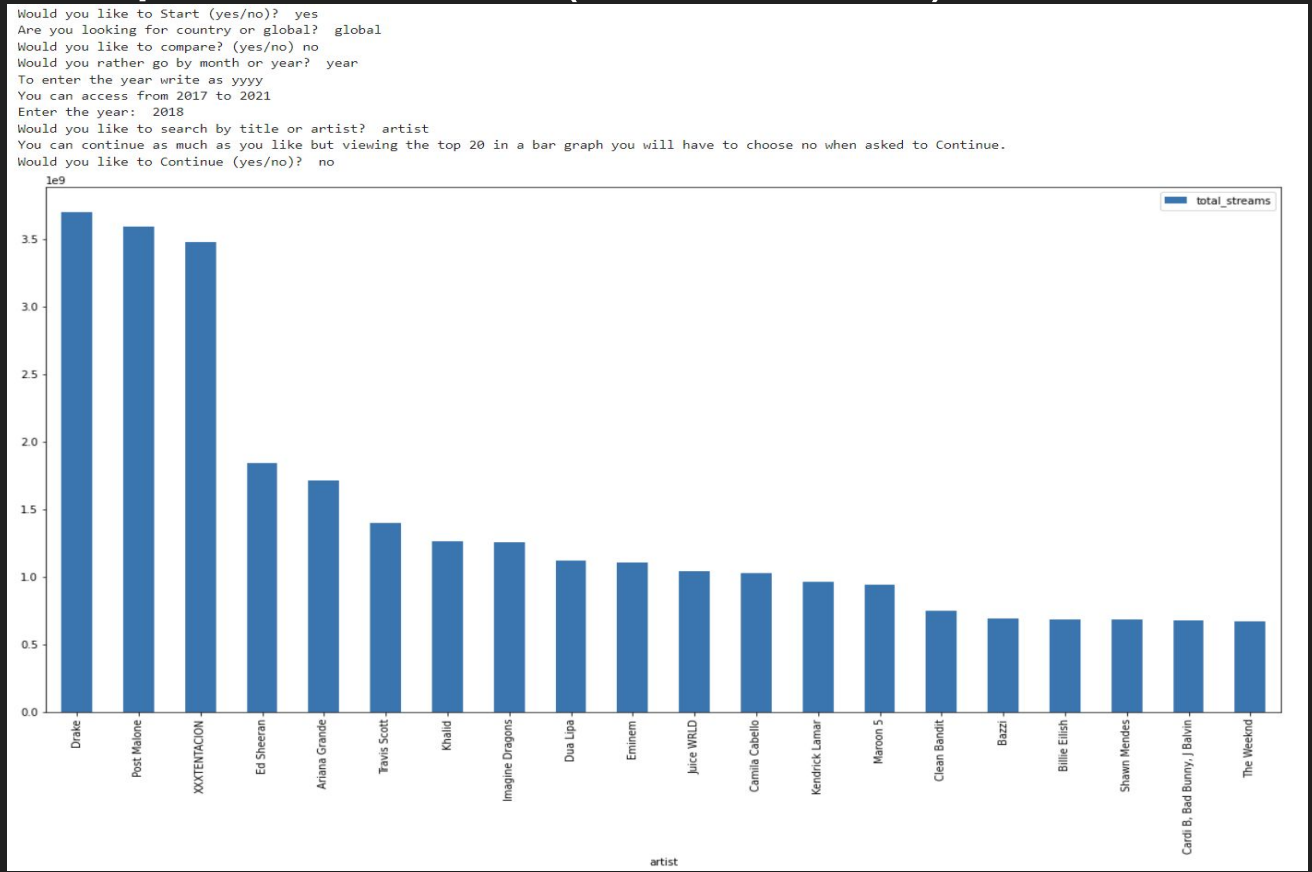
Example of the Runner (Country/Month)

```
Would you like to Start (yes/no)?  yes
Are you looking for country or global?  country
The Possible Countries are: United States, Switzerland, Australia, Brazil, Germany, United Kingdom, Sweden, Austria, Uruguay, or Chile
What Country would you like?  Uruguay
Would you rather go by month or year?  month
To enter the month write as yyyy-MM
You can access from 2017-01 to 2021-12
Enter the month:  2019-04
Would you like to search by title or artist?  title
You can continue as much as you like but viewing the top 20 in a bar graph you will have to choose no when asked to Continue.
Would you like to Continue (yes/no)?  no
```

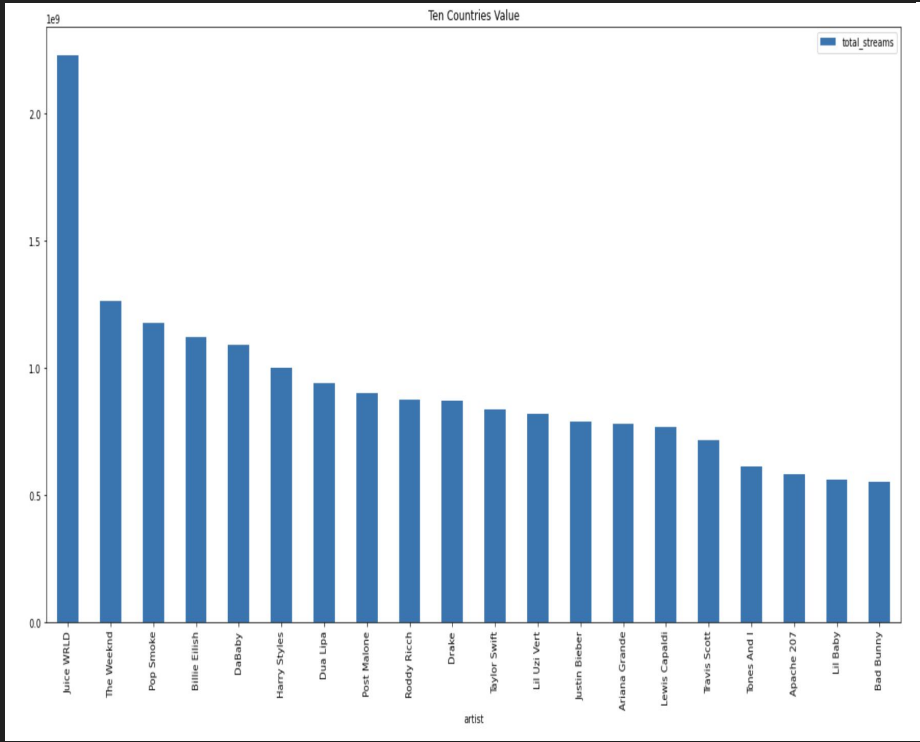


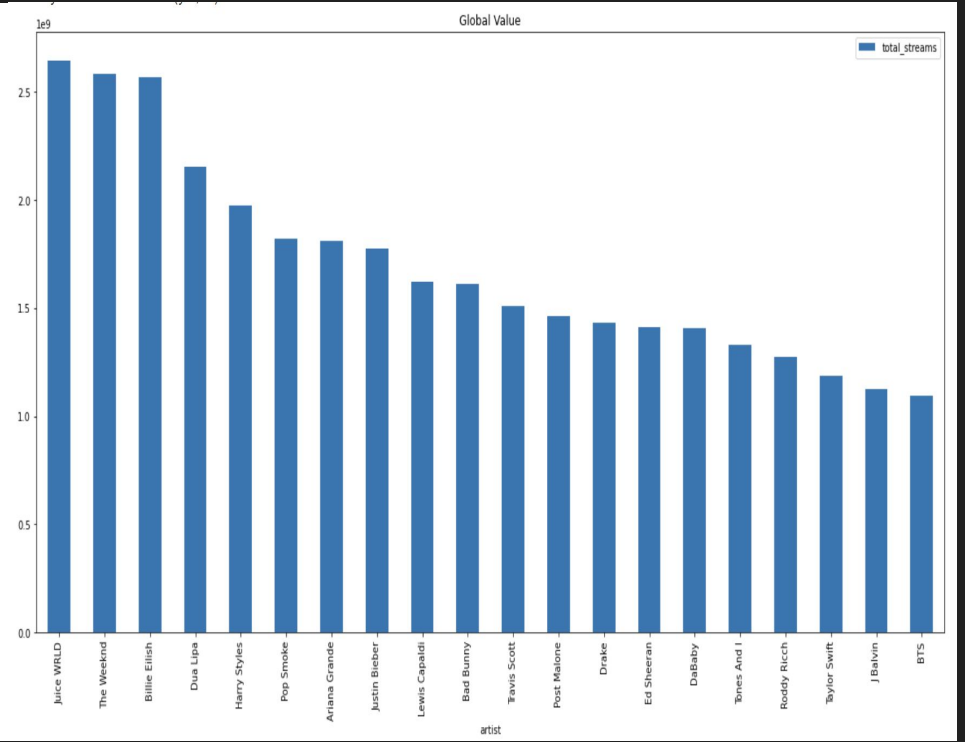
Example of Runner (Global/Year)

```
Would you like to Start (yes/no)?  yes
Are you looking for country or global?  global
Would you like to compare? (yes/no) no
Would you rather go by month or year?  year
To enter the year write as yyyy
You can access from 2017 to 2021
Enter the year:  2018
Would you like to search by title or artist?  artist
You can continue as much as you like but viewing the top 20 in a bar graph you will have to choose no when asked to Continue.
Would you like to Continue (yes/no)?  no
```



Example of Runner (Global and Comparison)





The Conclusion

- If I had more time I would have liked to have a better looking search engine for the graph and allow the graph to print out in place just after being called.
- It was really interesting to learn about the data and it was nice that the streaming value every time that it was gathered would either increase or decrease showing that the data isn’t based off the first time the data was gathered.
- That Spark is a pro it can be ready to use the 3.5GB is seconds while it takes pandas a few minutes for it to be ready.