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

Jack Brown brown.jac@northeastern.edu

Chuanzhang Tian tian.chu@northeastern.edu

Lei Fang fang.lei1@northeastern.edu

Dongkai Yu yu.do@northeastern.edu

Table of Contents

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.

1. Get the range of timeline of the data

+ Code + Markdown

[18]:
```
spark.sql('''
SELECT MIN(date) begin, MAX(date) end
FROM charts
WHERE chart = 'top200';
''').toPandas()
```

[18]:

	begin	end
0	2017-01-01	2021-12-31

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

2. From the following query, we see that the artists are listed in form of CSV

[19]:
```
spark.sql('''
SELECT *
FROM charts
WHERE chart = 'top200'
LIMIT 10;
''').toPandas()
```

[19]:

	title	rank	date	artist	url	region	chart	trend	streams
0	Chantaje (feat. Maluma)	1	2017-01-01	Shakira	https://open.spotify.com/track/6mICuAdrwEjh6Y6...	Argentina	top200	SAME_POSITION	253019
1	Vente Pa' Ca (feat. Maluma)	2	2017-01-01	Ricky Martin	https://open.spotify.com/track/7DM4BPaS7uofFul...	Argentina	top200	MOVE_UP	223988
2	Reggaetón Lento (Bailemos)	3	2017-01-01	CNCO	https://open.spotify.com/track/3AEZUABDXNtecAO...	Argentina	top200	MOVE_DOWN	210943
3	Safari	4	2017-01-01	J Balvin, Pharrell Williams, BIA, Sky	https://open.spotify.com/track/6rQSrBHf7HlZjtc...	Argentina	top200	SAME_POSITION	173865
4	Shaky Shaky	5	2017-01-01	Daddy Yankee	https://open.spotify.com/track/58IL315gMSTD37D...	Argentina	top200	MOVE_UP	153956
5	Traicionera	6	2017-01-01	Sebastian Yatra	https://open.spotify.com/track/5J1c3M4EldCfNxX...	Argentina	top200	MOVE_DOWN	151140
6	Cuando Se Pone a Bailar	7	2017-01-01	Rombai	https://open.spotify.com/track/1MpKZi1zTXpERKw...	Argentina	top200	MOVE_DOWN	148369
7	Otra vez (feat. J Balvin)	8	2017-01-01	Zion & Lennox	https://open.spotify.com/track/3QwBODjSEzelZyV...	Argentina	top200	MOVE_DOWN	143004
8	La Bicicleta	9	2017-01-01	Carlos Vives, Shakira	https://open.spotify.com/track/0sXvAOmXgjR2QUq...	Argentina	top200	MOVE_UP	126389
9	Dile Que Tu Me Quieres	10	2017-01-01	Ozuna	https://open.spotify.com/track/20ZAJdsKB5IGbGj...	Argentina	top200	MOVE_DOWN	112012

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.

3. Get total number of entries by all the singers in TOP 200

```
[20]: spark.sql('''
      SELECT COUNT(*) NoOfEntries
      FROM charts
      WHERE chart = 'top200';
      ''').toPandas().head(10)
```

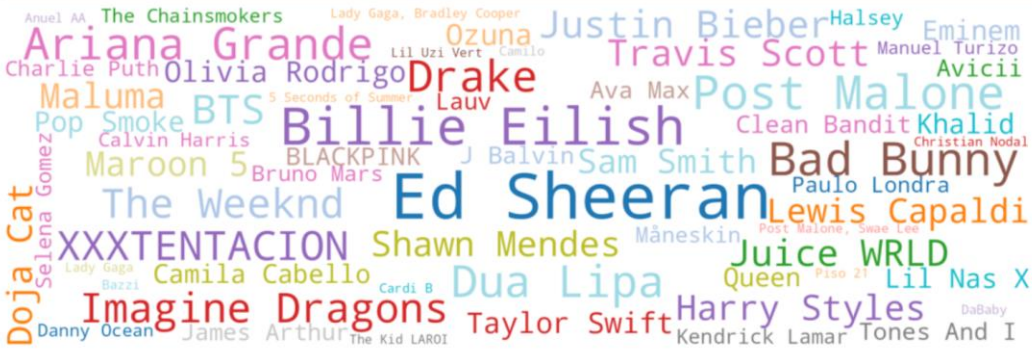
[20]:

	NoOfEntries
0	20318183

+ Code + Markdown

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.

```
visualize_word_counts(counts)
```



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.

```
[46]: counts.head(10)
```

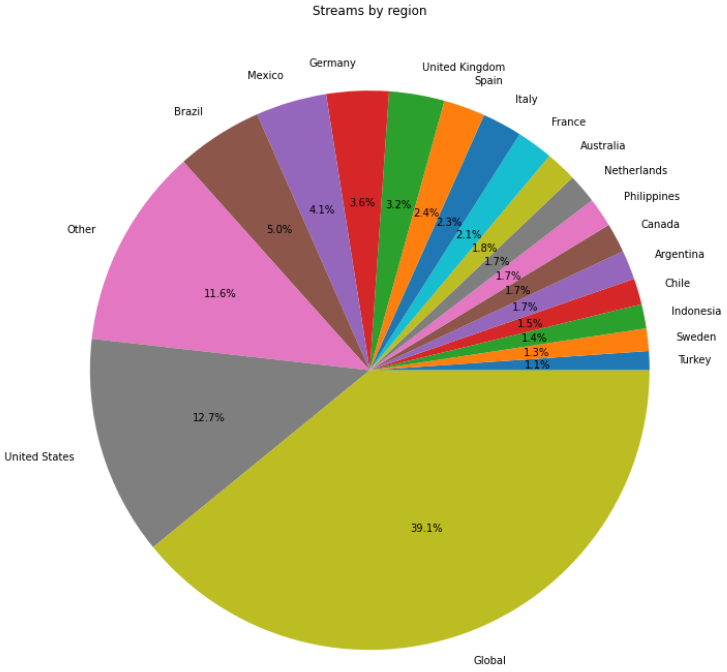
```
[46]: Ed Sheeran         387917
      Billie Eilish      251825
      Post Malone        211272
      Bad Bunny          203403
      Ariana Grande      189914
      Dua Lipa           187085
      Drake              177380
      Imagine Dragons    157900
      XXXTENTACION       155458
      BTS                154338
      Name: artist, dtype: int64
```

```
[63]: counts.tail(10)
```

```
[63]: Tymek, Kizo, Szpaku, Michał Graczyk, 2K                                1
      Aspova, Motive                                                        1
      Creditcard Reasons                                                    1
      Lips, Rose McIver                                                     1
      Aspova, Ceg                                                           1
      Ezhel, Patron, Anıl Piyancı, Sansar Salvo, Allame, Pit10, Beta, Kamufle    1
      Lips, George Mason, Rose McIver, Kimbra                               1
      Joakim Kleven, KOH                                                    1
      Lazyloxy, Maiyarap, OG-ANIC, UrboyTJ                                  1
      Luna 99                                                               1
      Name: artist, dtype: int64
```

+ Code + Markdown

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.



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.

5. Counting number of times Different singers appeared in the TOP 200 trend

+ Code + Markdown

```
spark.sql('''
SELECT COUNT(*) NoInTop200
FROM charts
WHERE artist LIKE '%Ed Sheeran%' AND chart = 'top200';
''').toPandas().head(10)
```

[23]:

	NoInTop200
0	368388

+ Code + Markdown

[67]:

```
spark.sql('''
SELECT COUNT(*) NoInTop200
FROM charts
WHERE artist LIKE '%Aspova%' AND chart = 'top200';
''').toPandas().head(10)
```

[67]:

	NoInTop200
0	1344

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.

	artist	count
0	Ed Sheeran	366026
1	Taylor Swift, Ed Sheeran, Future	1995
2	Tori Kelly, Ed Sheeran	176
3	Ed Sheeran, Elton John	125
4	Fireboy DML, Ed Sheeran	19
5	Alonestar, Rick Live, Ed Sheeran	14
6	The Weeknd, Ed Sheeran	9
7	Taylor Swift, Ed Sheeran	8
8	Alonestar, HerbertSkillz, Ed Sheeran	8
9	Foy Vance, Ed Sheeran	4

	artist	count
0	Aspova	762
1	Şanışer, Kamufle, Mert Şenel, Mirac, Aga B, De...	223
2	Aspova, Şanışer	96
3	Aspova, Tanerman	67
4	Aspova, Patron	52
5	Tuğkan, Aspova	42
6	Aspova, Şehinşah	35
7	Şanışer, Fuat, Ados, Hayki, Server Uraz, Beta ...	32
8	Vio, Aspova	8
9	Aspova, Motive, Murgs	7

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.

	title	streams
0	Shape of You	5245740051
1	Perfect	3038712776
2	I Don't Care (with Justin Bieber)	2296138118
3	Beautiful People (feat. Khalid)	1612020133
4	Bad Habits	1473943611
5	Photograph	1081454379
6	Castle on the Hill	1021200333
7	Galway Girl	1007337746
8	Thinking out Loud	853200137
9	Happier	782141998

	title	streams
0	Eskimiş Senelere	37435814
1	Susamam	20156260
2	Suç	3013264
3	Sağanak	2211168
4	Kader	2141421
5	Kanayan Yaralar	2046450
6	Ecel	2013589
7	Nude	1557447
8	TANIMIYORUM	742503
9	Içinde	361846

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.

	title	count
0	Shape of You	65262
1	Perfect	52392
2	Photograph	28605
3	I Don't Care (with Justin Bieber)	27101
4	Thinking out Loud	26014
5	Beautiful People (feat. Khalid)	21799
6	Happier	17056
7	Galway Girl	15614
8	Castle on the Hill	14653
9	Perfect Duet (Ed Sheeran & Beyoncé)	11258

	title	count
0	Eskimiş Senelere	645
1	Susamam	255
2	Suç	96
3	Sağanak	73
4	Nude	67
5	Kanayan Yaralar	52
6	Ecel	42
7	Kader	35
8	TANIMIYORUM	20
9	Dur Dedik	9

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.

10. Total how many times has Different singers been on the 1st rank

+ Code + Markdown

```
spark.sql('''
SELECT COUNT(*) NoOfRank1
FROM charts
WHERE artist LIKE '%Ed Sheeran%'
AND chart = 'top200'
AND rank = 1;
''').toPandas().head(10)
```

[29]:

	NoOfRank1
0	4810

+ Code + Markdown

[66]:

```
spark.sql('''
SELECT COUNT(*) NoOfRank1
FROM charts
WHERE artist LIKE '%Aspova%'
AND chart = 'top200'
AND rank = 1;
''').toPandas().head(10)
```

[66]:

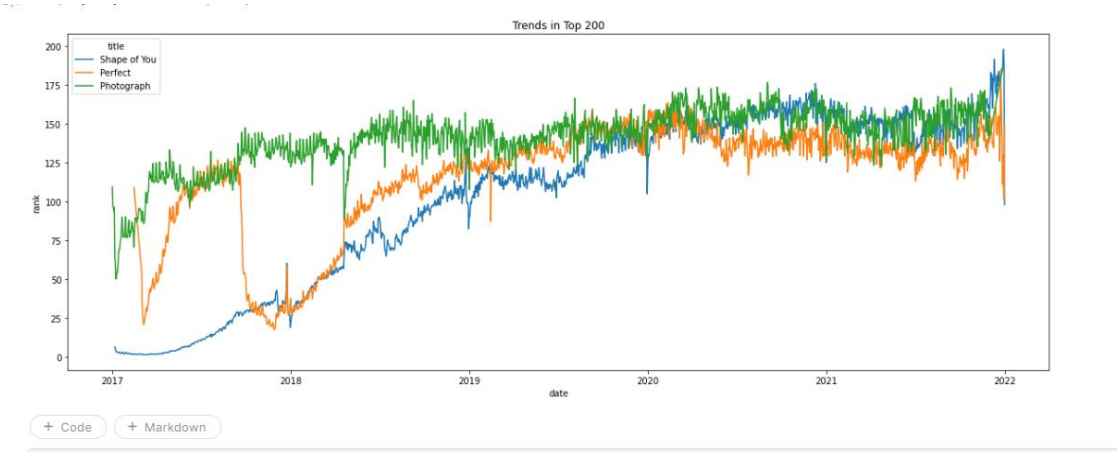
	NoOfRank1
0	14

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.

	Title	Highest	Lowest	Avg
0	Beautiful People (feat. Khalid)	1	200	77.225240
1	Galway Girl	1	200	80.686499
2	Shape of You	1	200	80.527842
3	I Don't Care (with Justin Bieber)	1	200	77.736910
4	Perfect Duet (Ed Sheeran & Beyoncé)	1	200	101.299698
5	Castle on the Hill	1	200	86.030779
6	Perfect	1	200	102.164930
7	Bad Habits	1	200	41.974160
8	Shivers	1	200	41.899490
9	Eraser	2	200	97.767938

	Title	Highest	Lowest	Avg
0	Susamam	1	198	91.403922
1	Eskimiş Senelere	10	200	68.116279
2	Kanayan Yaralar	17	198	109.826923
3	Kader	21	192	84.000000
4	Içinde	30	177	94.750000
5	Ecel	31	156	95.714286
6	Suç	43	200	126.385417
7	TANIMIYORUM	51	197	122.500000
8	Nude	72	199	162.313433
9	Sür	73	190	140.000000

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.

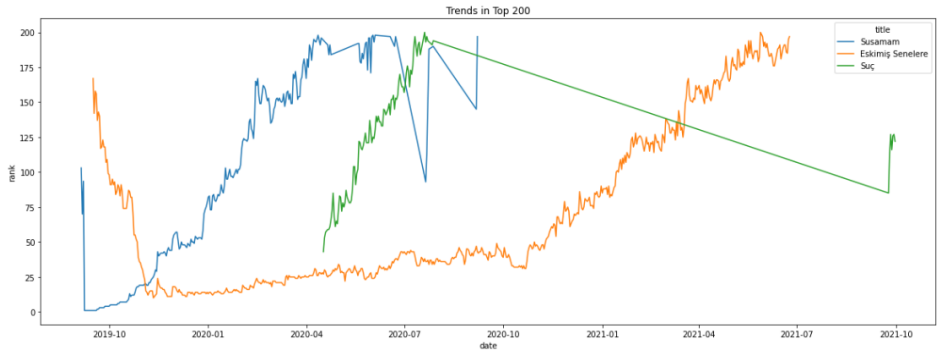


However, the ranking of each song on Aspova has experienced wild fluctuations, and three songs demonstrate three different trends, which means songs with a steady increase in ranks are more likely to be popular. Although one of Aspova's songs in orange (approximately 200 in rank), all three songs by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

```
fig, axes = plt.subplots(figsize=(20,7))
sns.lineplot('date', 'rank', data=q, hue='title', ci=None).set_title('Trends in Top 200')
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument w
will result in an error or misinterpretation.
  FutureWarning
```

[71]: `Text(0.5, 1.0, 'Trends in Top 200')`



+ Code + Markdown

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