

**TASK**

**Exploratory Data Analysis on the Spotify Top 50 Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

The purpose of this report is to discuss the findings of the Spotify’s 2019 top 50 songs data set using a preliminary exploratory data analysis (EDA) approach. The data set contains a total of 13 attributes which includes song names, artist’s names and several song characteristics. The dimension of the data set is 50 x 13.s

The data set is a sample of how Spotify categorises their songs to provide song prediction and recommendations for its users. Spotify uses audio modelling i.e. convolutional neural networks (CNN) to assign artists songs several characteristics and metrics such as loudness, energy or genre type. Subsequently, Spotify is able to group and order songs with similar characteristics. Thereafter suggest or recommend songs to their users based on their listening preference at any time.

**DATA CLEANING**

The data was checked for any duplicate rows but none was found. The data was checked whether the numerical data was in integer or float format.

MISSING DATA

No missing data was observed.

DATA STORIES AND VISUALIZATIONS

This is section investigates the following:

* The relationship between rank and genre
* Typical numerical Spotify algorithmic characteristics for Spotify top 50 songs
* Correlation between the data set’s quantitative data
* The most featured artist in data set

RANK AND GENRE

This section seeks to investigate the relationship between song rank and genre. This section initially finds the most common genres within the top 50. Thereafter, it finds the ranking distribution for each genre if there is any.

Figure 1 is a bar graph showing the total count of the various genres present within the data set. There are 21 different genres within the data set. Most of the genres encompasses subsets of pop, hip hop or rap. For example, there is dfw rap and country rap. The top three genres present in the data set in descending order is dance pop, pop and latin.

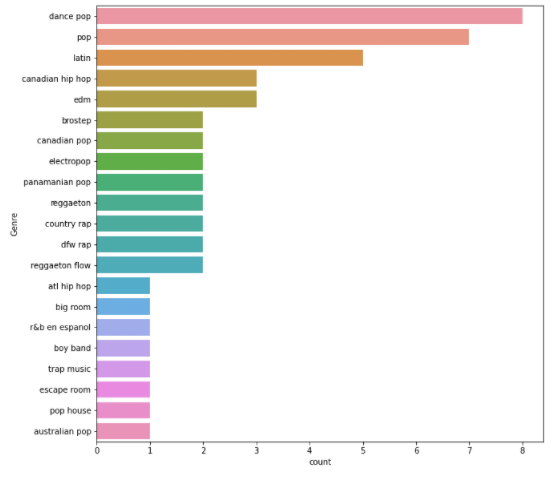


Figure : Total count of song genre in Spotify top 50

Figure 2 is a boxplot showing the rank distribution of each genre. Canadian pop songs, being the most counted genre in the data tends to be on the top half of the data of the song rankings i.e. 25 and less. Other genres that have a similar characteristic are reggaeton, dfw rap, trap music, country rap and electro pop.

Both dance pop and pop have the largest rank distribution. However, the difference when comparing their mediums. Dance pop songs tend to be on the present on the lower half of the rank. In contrast, pop songs tend to on the upper half of the rankings.

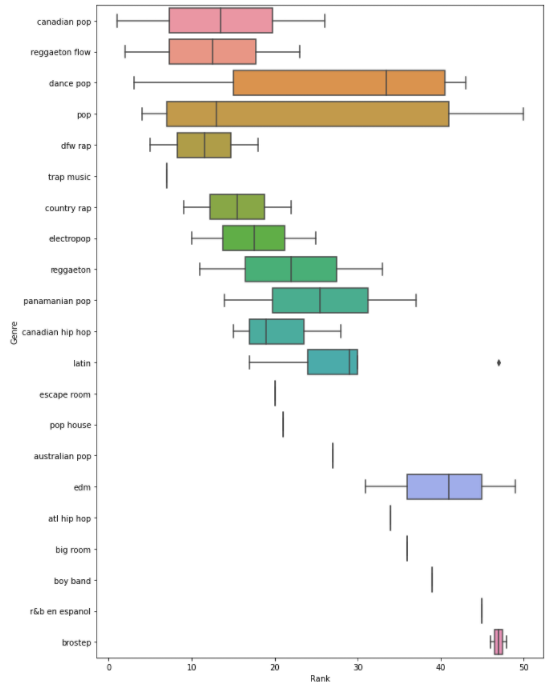


Figure : Correlation between genre and rank

RANKING

This section seeks to investigate the correlation between song rankings and all the quantitative data types within the data set. The following quantitative data were investigated:

|  |  |  |
| --- | --- | --- |
| **Data** | **Data Type** | **Range** |
| Beats per Minute | Continuous | Infinity |
| Danceability | Continuous | 0-100 |
| Loudness | Continuous | Infinity |
| Liveness | Continuous | 0-100 |
| Valence | Continuous | 0-100 |
| Song Length | Continuous | 0-Infinity |
| Acoustiness | Continuous | 0-100 |
| Speechiness | Continuous | 0-100 |

For attributes with a range from 0 to 100, the closer the value is to 100 the more significant the respective song characteristic. The song popularity was not investigation for any correlation with song ranking as these songs are expected to have high popularity values as it is in Spotify top 50 songs for 2019.

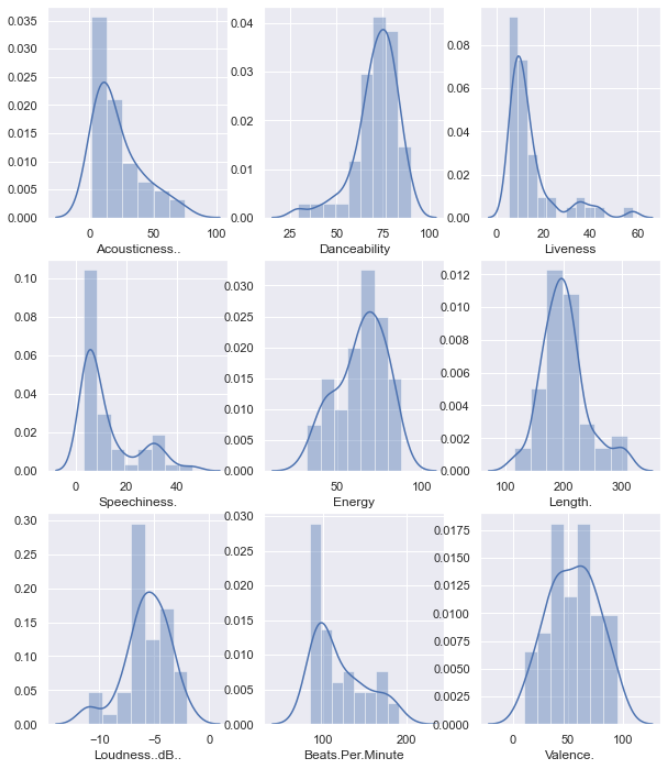


Figure : Distribution of the song characteristics

Figure 3 shows multiple frequency distribution histograms of the various song characteristics. Clearly tends in all characteristics can be seen for what is generally what is expected in terms of song characteristics in Spotify top 50. For example, songs tend have high energy and songs that people would likely dance to.

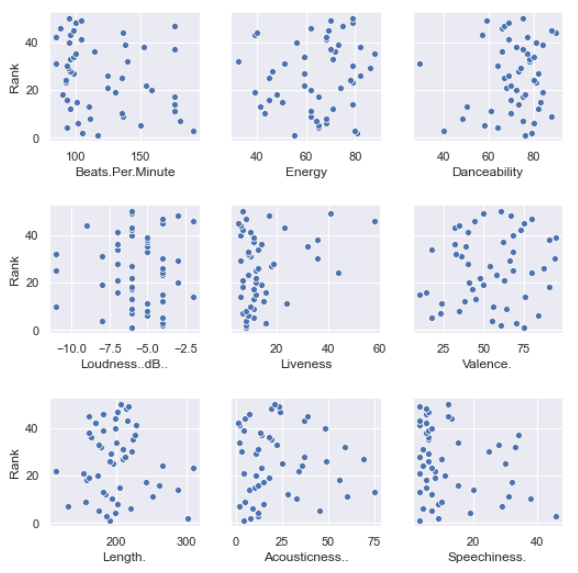


Figure : Correlation between song ranking and characteristics

Figure 4 show multiple scatterplot of the correlation between song ranking and the various characteristics. No clear linear trends can be observed for any of the graphs. This suggests that if an artist wants their song in the top 50, it does not necessarily have to highest or best characteristic in order to be the best song. A more reasonable approach is to make a song which has similar attributes to the other songs in the top 50.

QUANTITATIVE DATA CORRELATIONS

This section seeks to investigate the correlation between all the quantitative data in the data set. Figure 5 is a heat map showing linear correlations between the quantitative values. The numerical value within each block denotes the strength of linear correlation. The closer to 1 or -1 the value is, the more significant the linear correlation either positive or negative respectively. The closer the value is to 0, the less significant the correlation is.

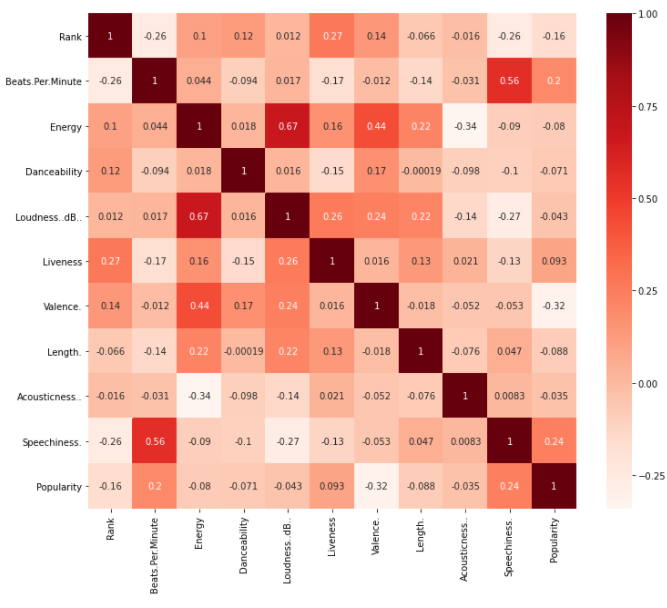


Figure 5: Linear correlation significance for all data set's quantitative values

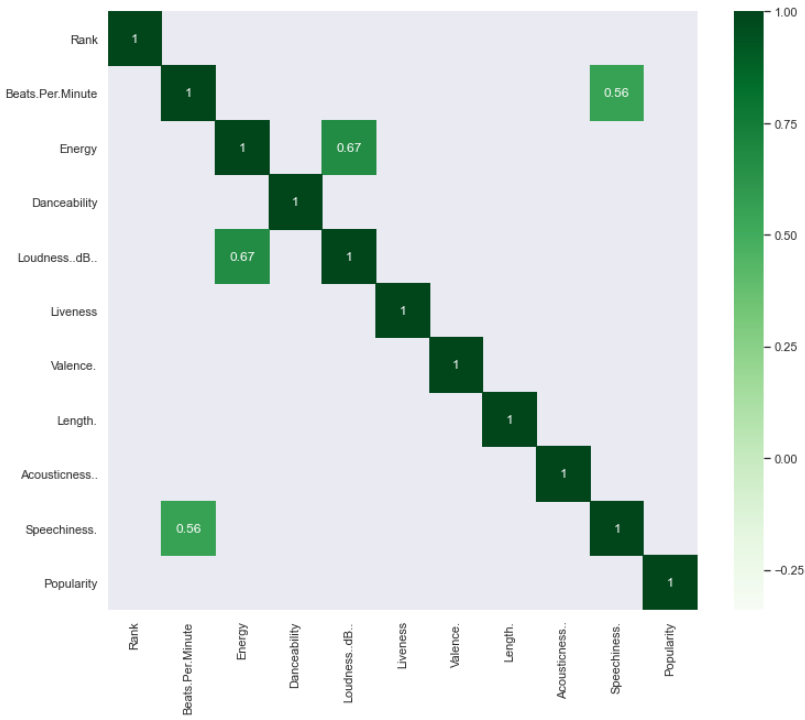


Figure 6: Quantitative data with significant positive linear correlation

Figure 6 is a heat map showing the significant positive linear correlations between the qualitative data. It was assumed that is the value was greater 0.5, it would be categorised as a reactively significant positive correlation. Only two positive correlations can be observed.

Figure 7 is a heat map showing the significant negative linear correlations between the qualitative data. It was assumed that is the value was less than -0.5, it would be categorised as a reactively significant negative correlation. No relatively significant negative correlations was observed.

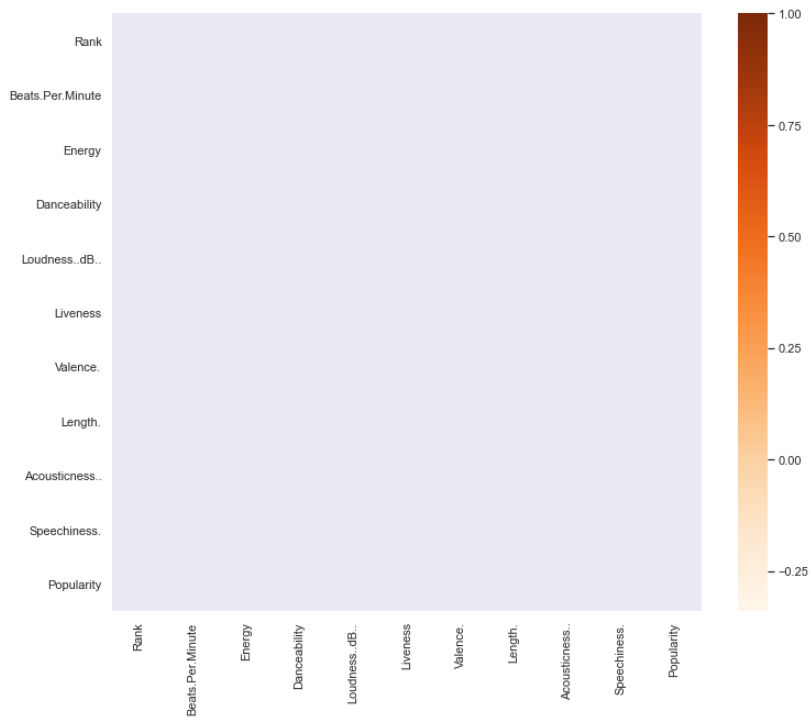


Figure 7: Quantitative data with significant negative linear correlation

COMMON ARTISTS

This section seeks to investigate the most common artists present in Spotify’s top 50 songs. The songs in the data set either only has one artist i.e. the main artist or multiple artists on one track. In a case where there are multiple artists, the Artist name column (i.e. “Artist.Name”) will have the main artist the collaborating artist/s are shown in the Track.name column. The collaborating artist name normally has the words “feat.”, “ft” or “with” in front of them (Refer to Figure 8). As a result, this section will also find the number of times an artist appears in the data set as a collaborating artist as well.

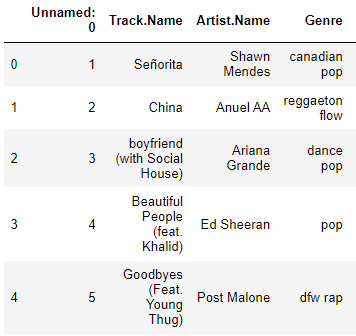


Figure : Typical songs with collaborating artists

Figure 9 is a bar graph showing the total number of times an artist is a main artist in Spotify top 50 songs. It was observed that Ed Sheeran was the most common artist in the data set. Further investigation shows that the songs made by Ed Sheeran are all pop songs.

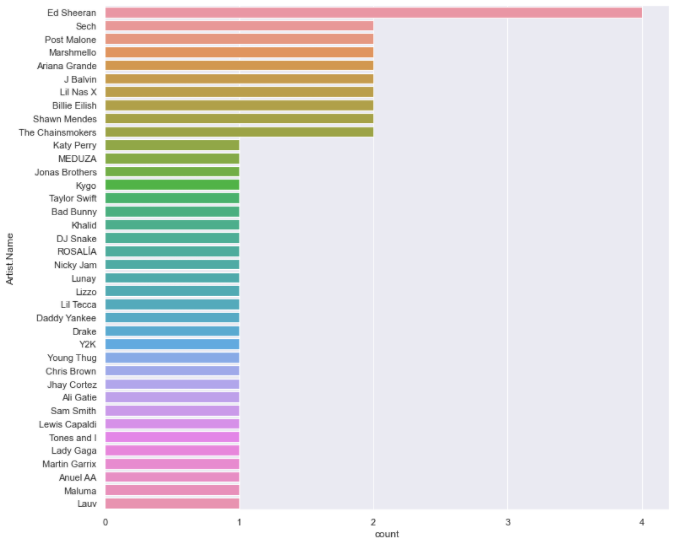


Figure 9: Count of main artists in Spotify top 50

Figure 10 is a bar graph showing the total number of times an artist is a collaborating artist in Spotify top 50 songs. Using the fuzzywuzzy package, it was determined that some of these artists are also present as main artists. Drake and Khalid were main artists in the data set.

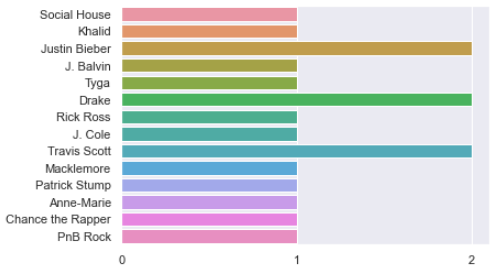


Figure : Total count of featuring artists in Spotify top 50

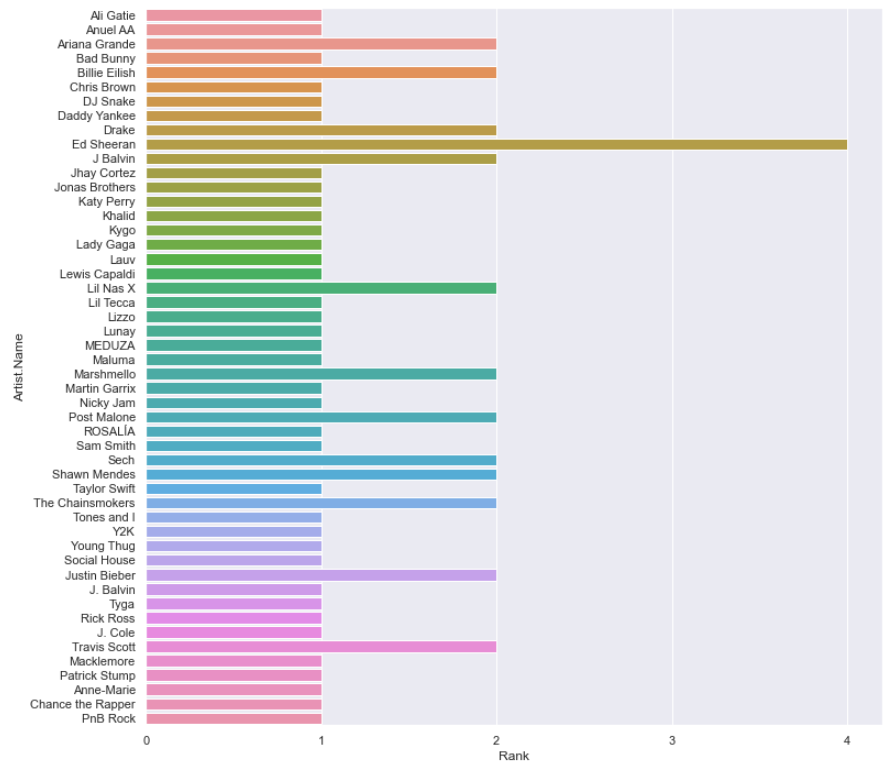


Figure : total count of artists as the main or collaborating artist in Spotify top 50

Figure 11 is a bar graph showing the total number of times an artist was both a main and a collaborating artist in Spotify top 50 songs. With the addition of the collaborating artist counts, there are more artists that show up twice in the data set. Drake shows up three times in the data set. However, Ed Sheeran is still the most present artist in the data set.

CONCLUSION

The purpose of this report is to discuss the findings of the Spotify’s 2019 top 50 songs data set using a preliminary exploratory data analysis (EDA) approach. The investigation identified trends between song rank and genre, typical song characteristics, quantitative data correlations and the most featured artist in playlist.

In conclusion, Canadian pop was the most common genre in the data set. Most of the genres encompasses subsets of pop, hip hop or rap. There are similar song characteristics in the data set. However, a higher the rank does correlate to a better characteristic value. Finally, artists in the data set can exist as either the main artist or the collaborate artist. Overall Ed Sheeran was the most present artist in the data set.

REFERENCES

https://stackoverflow.com/questions/68258626/unicodedecodeerror-utf-8-codec-cant-decode-byte-0xf1-in-position-5-invalid

https://stackoverflow.com/questions/187455/counting-array-elements-in-python

https://medium.com/swlh/spotify-song-prediction-and-recommendation-system-b3bbc71398ad

https://stackoverflow.com/questions/39359601/python-pandas-series-str-contains-whole-word

https://www.kaggle.com/leonardopena/top50spotify2019

https://www.linkedin.com/pulse/how-spotify-recommender-system-works-daniel-roy-cfa/

THIS REPORT WAS WRITTEN BY: Nabeel Omar

