Final Project Report
Introduction to Data Analytics

Project Title:
Prediction/Analysis of Popularity of Top 50 Spotify Songs 2019

Prepared by:
Spencer Standish: N01576620
Keziah Thomas: N01541155

ITE 5201 – Winter 2023
Humber College

1. Problem Statement

=> Prediction/Analysis of Popularity of Top 50 Spotify Songs 2019

2. Dataset Description

=> This dataset contains the following variables/columns:

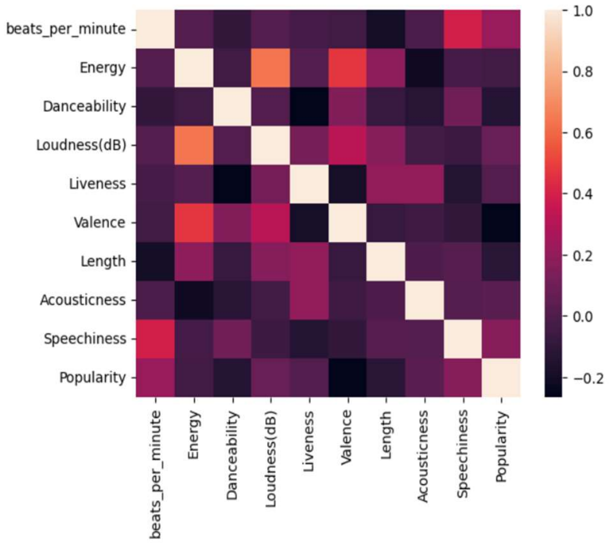
- **track_name** - Name of the Track
- **artist_name** - Name of the Artist
- **Genre** - The genre of the track
- **beats_per_minute** - The tempo of the song.
- **Energy** - The energy of a song - the higher the value, the more energetic. song
- **Danceability** - The higher the value, the easier it is to dance to this song.
- **Loudness(dB)** - The higher the value, the louder the song.
- **Liveness** - The higher the value, the more likely the song is a live recording.
- **Valence** - The higher the value, the more positive mood for the song.
- **Length** - The duration of the song.
- **Acousticness** - The higher the value the more acoustic the song is.
- **Speechiness** - The higher the value the more spoken words the song contains.
- **Popularity** - The higher the value the more popular the song is.

We train the model using beats_per_minute, Energy, Danceability, Loudness(dB), Liveness, Valence, Length, Acousticness and Speechiness to predict the popularity.

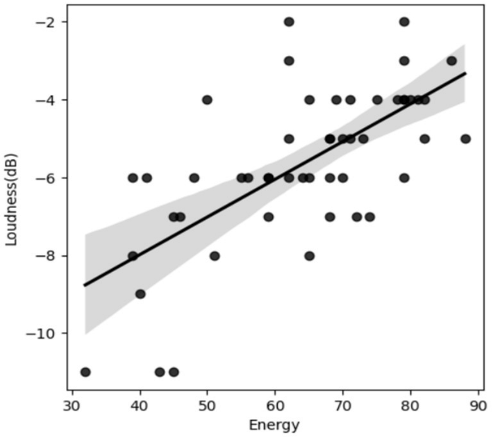
3. Dataset Analysis and Observations

=> For dataset analysis, we plot the heatmap plot shown below using spearman method to identify the correlation coefficient rank (R) between different variables.

- R = 1 indicates strong positive relationship
- R = 0 indicates that they are not linearly correlated
- R = -1 indicates strong negative relationship



=> **Observation:** Based on the heatmap where lighter colors indicate high correlation and darker colors indicate low correlation, we can conclude that Energy and Loudness have a positive correlation compared to the rest of the variables.



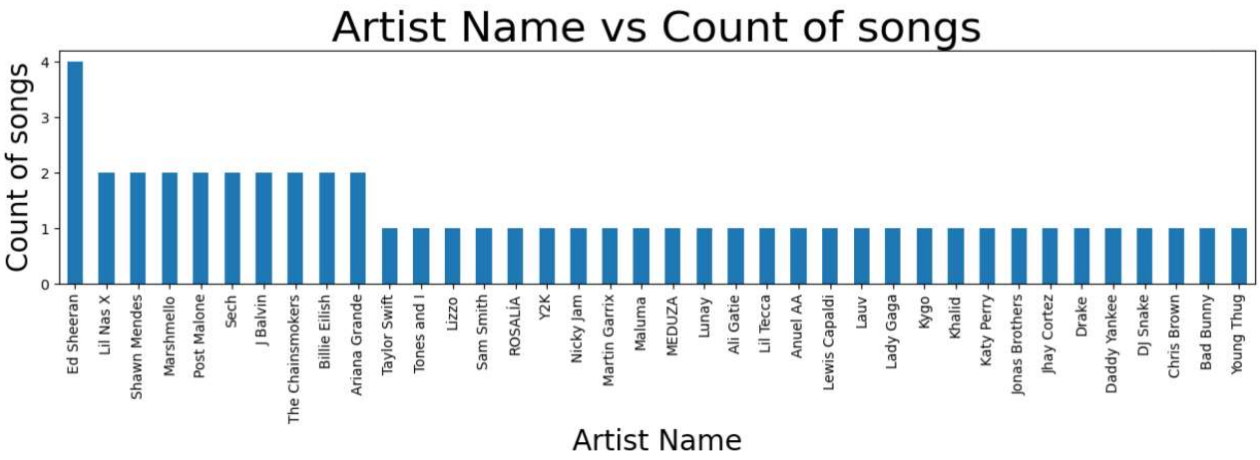
4. Proposed Analytical/Prediction Model

=> As above analysis, the dataset has dependent variables for calculating popularity.

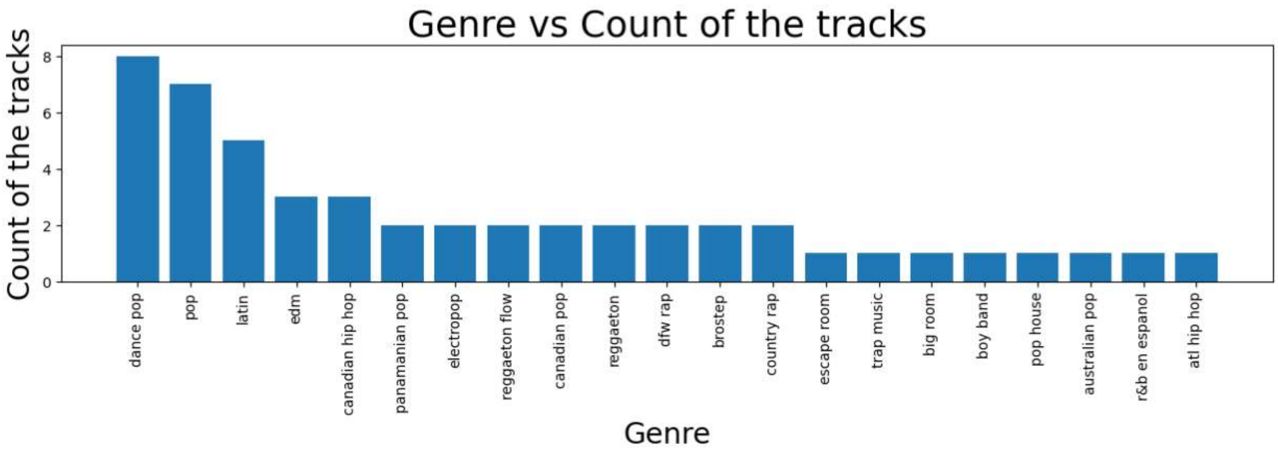
=> We used a linear regression model for this dataset as we work with numerical data. We don't use logistic regression model or K-Nearest Neighbor model as we have to analyze the reason behind the popularity of songs based on numerical measures and not whether the song is popular or not.

5. Results and Discussions

=> The bar chart shown below plots the number of popular songs against each artist. Using this visualization, can we say that the artist with the most popular songs is likely to be the top artist for the #1 song?



=> Or, can we state that the top #1 song is likely to belong to the dance pop genre as this genre has the highest number of popular songs as per the bar chart visualized below?

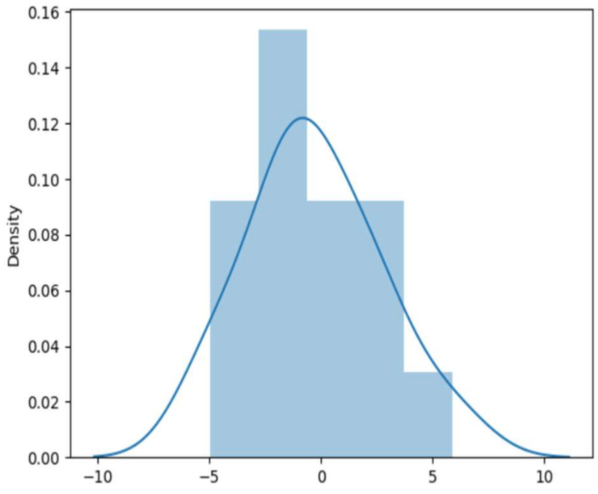


=> Popularity of the song doesn't depend on the most popular genre or the most popular artist. As per our analysis, we can observe that popularity actually depends on: beats_per_minute, Energy, Danceability, Loudness(dB), Liveness, Valence, Length, Acousticness and Speechiness.

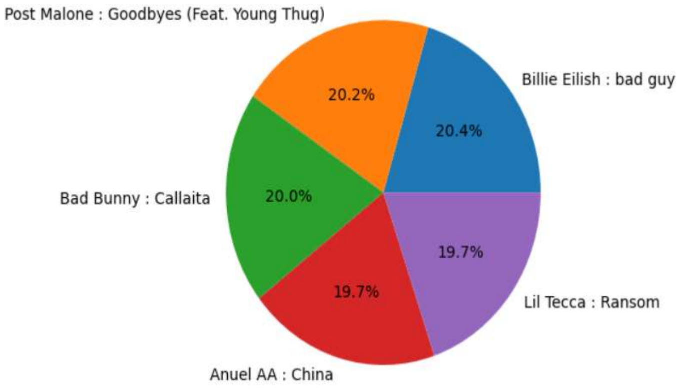
=> We get the following metric values with our linear regression model:

- ⇨ **Mean Absolute Error:** 2.39197463658397
- ⇨ **Mean Squared Error:** 8.362163216805305
- ⇨ **Root Mean Squared Error:** 2.891740516852317

=> Since the error terms resemble closely to a normal distribution, we can move ahead and make predictions using the model in the test dataset.



=> **Conclusion:** We achieved very low MAE, MSE and RMSE values when we compared our actual and predicted values. Our linear regression model is highly accurate and we are able to understand and predict the popularity of a song based on beats_per_minute, Energy, Danceability, Loudness(dB), Liveness, Valence, Length, Acousticness and Speechiness. The pie chart plots the top 5 popular songs and artists for the Spotify 2019 dataset.


Post Malone : Goodbyes (Feat. Young Thug)
Billie Eilish : bad guy
Bad Bunny : Callaita
Lil Tecca : Ransom
Anuel AA : China