Calvin Truong

Professor Fadi Batarseh

CIS 4321

7/1/2020

Spotify Clustering

Introduction

For as long as I can remember, I have always been passionate about music and was interested in working at a technology music industry such as Pandora, SoundCloud or Spotify. I have always wanted to share my knowledge in music with others and be able to find important trends that can help the consumers discover new songs that they can enjoy. For this report, I will be focusing on Spotify because I have been using Spotify as my daily music platform. The dataset that will be used comes from Kaggle.com. The dataset contains Spotify's top 50 songs that people like to listen the most in 2019, which includes 13 attributes and they are:

- Track_Name – Name of the track (data type: object)
- Artist_Name – Name of the artist (data type: object)
- Genre – The genre of the track (data type: object)
- Beats_Per_Minute – Tempo speed of the song (data type: int64)
- Energy – The higher the value the more energetic the song is (data type: int64)
- Danceability – The higher the value, the easier it is to dance to the song (data type: int64)
- Loudness – The higher the value, the louder the song is (data type: int64)
- Liveness – The higher the value, the more the song is a live recording (data type: int64)

- Valence – The higher the value, the more positive mood for the song (data type: int64)
- Length – The duration of the song (data type: int64)
- Acousticness – The higher the value, the more acoustic the song is (data type: int64)
- Speechiness – The higher the value, the more spoken word the song has (data type: int64)
- Popularity – The higher the value the more popular the song is (data type: int64)

Understanding what type of music consumers search for is very crucial to companies like Spotify. It tells them the genre the company should focus on and with the help of machine learning, the company can make predictions for recommended songs to the consumers based on the songs they already like and added to their library. There are some questions I want to discover and answer. What is the top song genre? What are the top five most important features? How many clusters should be implemented to achieve the best prediction? How can we determine what people like to listen the most? The dataset did not contain any messy or missing values, however, to read the csv file, the file must be encoded with a Latin script called "ISO-8859-1". After this, everything should be normal for the data analysis process.

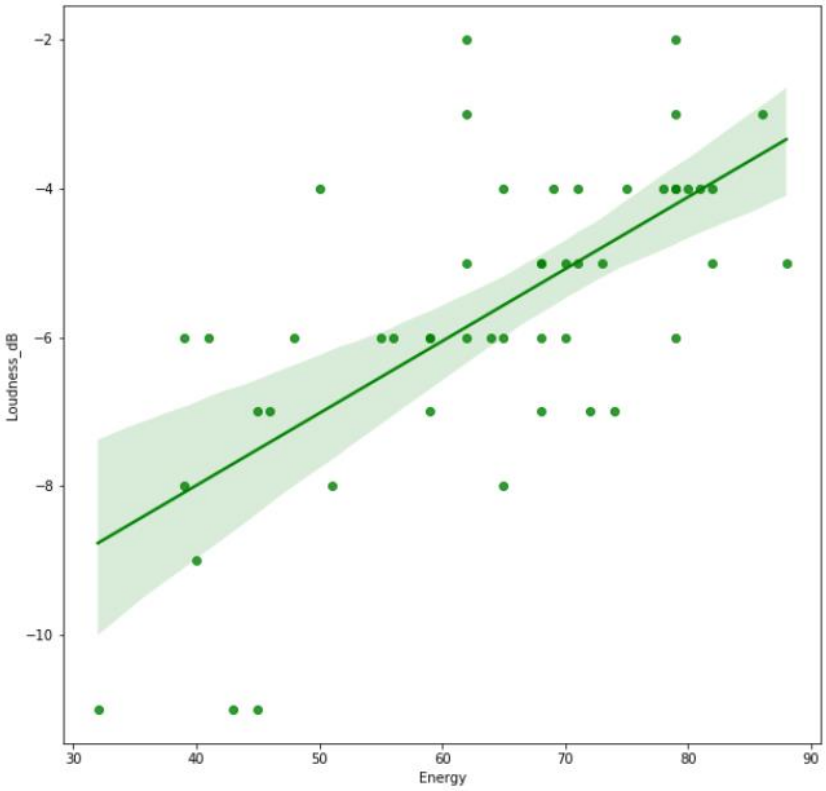
Data Analysis

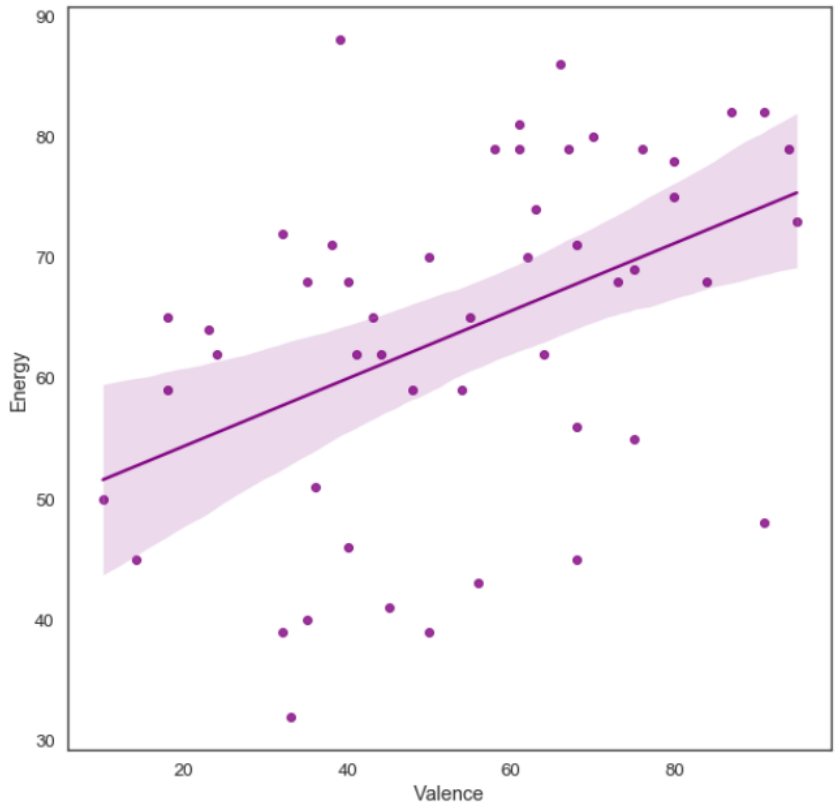
Once the file has been accessed, it is important to explore the dataset and see what is inside. It is also good to use the **.describe()** function to find out more information about the overall attributes.

	Unnamed: 0	Beats_Per_Minute	Energy	Danceability	Loudness_dB	Liveness	Valence
count	50.00000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000
mean	25.50000	120.060000	64.060000	71.38000	-5.660000	14.660000	54.600000
std	14.57738	30.898392	14.231913	11.92988	2.056448	11.118306	22.336024
min	1.00000	85.000000	32.000000	29.00000	-11.000000	5.000000	10.000000
25%	13.25000	96.000000	55.250000	67.00000	-6.750000	8.000000	38.250000
50%	25.50000	104.500000	66.500000	73.50000	-6.000000	11.000000	55.500000
75%	37.75000	137.500000	74.750000	79.75000	-4.000000	15.750000	69.500000
max	50.00000	190.000000	88.000000	90.00000	-2.000000	58.000000	95.000000

Next, is to focus on cleaning up the data and prepare it to make it easier to understand. Since there are some columns that might not be useful for the data set. Users should organize the data and drop irrelevant columns (in this case, drop the Unnamed index column).

While looking at the dataset, there have been a few interesting relationships between the attributes. The first relationship is between the energy and the loudness of the song, there is positive correlation between the two. Based on the figure, the higher the loudness, the higher the energy the song will be. Another relationship is between valence and energy. The happier the song is, the more energy the song will have.


Loudness_dB
−2
−4
−6
−8
−10
30
40
50
60
70
80
90
Energy


Energy
90
80
70
60
50
40
30
20
40
60
80
Valence

Modeling Tasks

The main modeling approach that was used to learn about the dataset was clustering. Clustering is an unsupervised machine learning technique that groups up the top 50 songs into categories based on similar patterns from the attributes and labels them accordingly. To implement the model, the dataset must first be normalized and transformed. For normalizing the data, the track name, artist name, and genre should be excluded from the X matrix, but will be used again after the preprocessing and clustering process. Then the X matrix need to be trained and tested. For this report, 70% of the X matrix will be for training while the other 30% of the X matrix will be for testing. This can used to predict the Y matrix of an attribute and can be used as a comparison to the actual targeted attribute. Lastly, the data will undergo fitting and scaling transformation to run the principal component analysis (PCA) on the scaled X matrix. Lastly

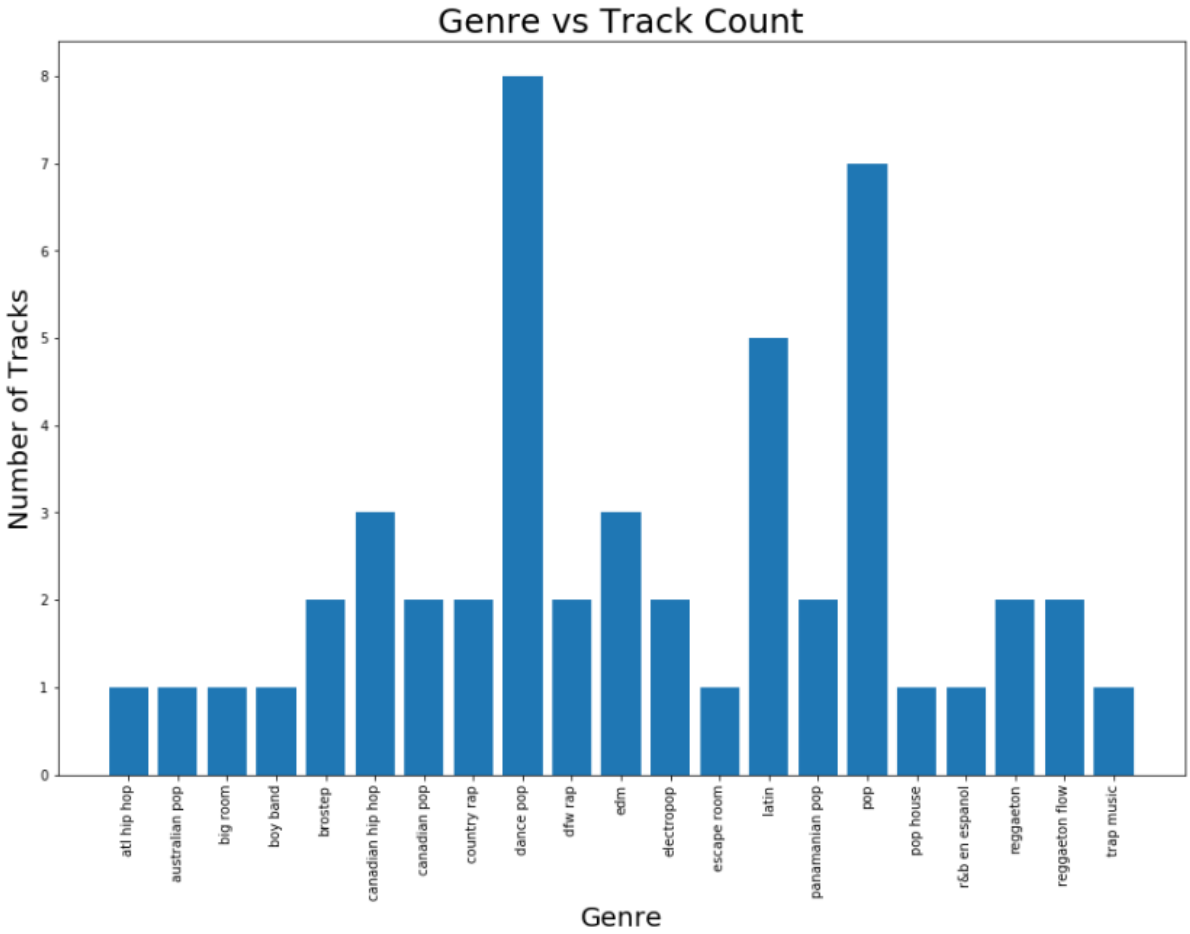
Now the user needs to determine how many cluster amounts should be optimized for this set of data so useful algorithm are to implement the Silhouette Analysis and K-Nearest Neighbor (KNN) classifier. The Silhouette function will calculate estimate percentages for each number of N clusters. The number of clusters with the highest percentage should be optimized for the KNN classifier. It is important to note that increasing the N dimensions will not always create a better performance because there is a maximum number of features before the performance starts to degrade. This is known as the curse of dimensionality.

Once the number of clusters is determined, the K-means Clustering can begin and categorize each song to a specific group. This type of data clustering will calculate the Euclidean distance between the placement and estimated data centroids to further match where the clusters

should be placed. The clusters categories will then be renamed based on the labels. This method works well because it can make predictions to what type of music people enjoy listening to most.

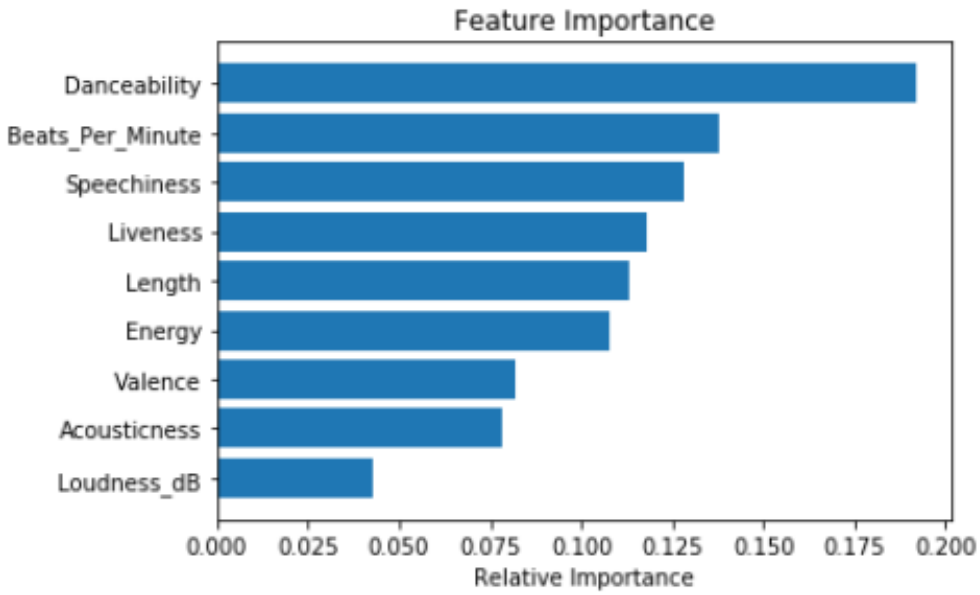
Results

Using data mining and machine learning techniques, I was able to answer the questions from the introduction section. The top song genres are dance pop with 8 tracks and pop with 7 tracks. The comparison between the prediction of popularity songs and the actual popularity songs are very similar to each other, which shows accuracy.



According to feature importance, the top five highest percentage of importance are beats per minute, danceability, speechiness, liveness, and acousticness compared to the other

attributes. This makes sense because people tend to listen more to the songs that are easily danceable and can be shared with others. This is also useful because companies can research how they can improve the features that have a low relative importance.



As for the clustering process, the best cluster amount should be 5 clusters because it has the highest average silhouette score of 16%. These 5 clusters will have the label of high energy & danceable, low energy & danceable, mid energy & lowly danceable, low energy & lowly danceable, and mid energy & highly danceable. The most popular type of music people listen is high energy & danceable at 18 counts.

```
High Energy, Danceable           18
Low Energy, Danceable            12
Mid Energy, Lowly Danceable       8
High Energy, Highly Danceable     6
Low Energy, Lowly Danceable       6
Name: Cluster Names, dtype: int64
```

Discussion

Many people enjoy listening to music that can make them feel happy and dance. They are more likely to listen to a positive song rather than a sad and chill type of song in a car ride. Companies can use this information to advertise and influence future artists to create more songs in the Pop genre to increase streaming profits. For the same reason above, feature importance heavily depends on danceability and speechiness. Due to social media platforms such as Tiktok and Instagram, more people are influence by others through dancing. If people enjoy dancing to the music, they will listen to the songs more often. Lastly, having a high energy in the song also has a huge impact in songs because it makes people want to dance and sing along. Pop music contains both a high danceability, energy, and speechiness; which proves why it is the most popular genre in the music industry.

In addition, even though EDM music did not score too well in the data for 2019, EDM may become popular in the future due the fact that it is still a brand new genre and has many different sub-genres such as house, progressive, trap, future bass, and etc. A good data mining topic that could be used is classification since it can help categorize the different types of EDM subgenres and help consumers determine which type of EDM music they like to listen.

Conclusion

The dataset provided has given many valuable insights that can help music companies such as Spotify. Some questions that can be found using this information such as. How can the company improve the important features that did not have a huge impact on the overall analysis (such as valence and loudness)? How can the company promote and advertise pop music to the consumers knowing that it is the most popular music genre? A recommendation for the future is to clean out the Unnamed index attribute and any other unnecessary columns. It is also

recommended to do data transformation and pay extra attention to the required preprocessing steps. The key component of the code is to include the encoding of the Latin script “ISO-8859-1”, otherwise the csv file cannot be accessible. Another key component is to use the Silhouette Analysis to determine which number of clusters will have the best performance and the KNN classifier to perform the K-Means clustering.

CITATIONS:

Henrique, L. (2019, August 08). Top 50 Spotify Songs - 2019. Retrieved from https://www.kaggle.com/leonardopena/top50spotify2019