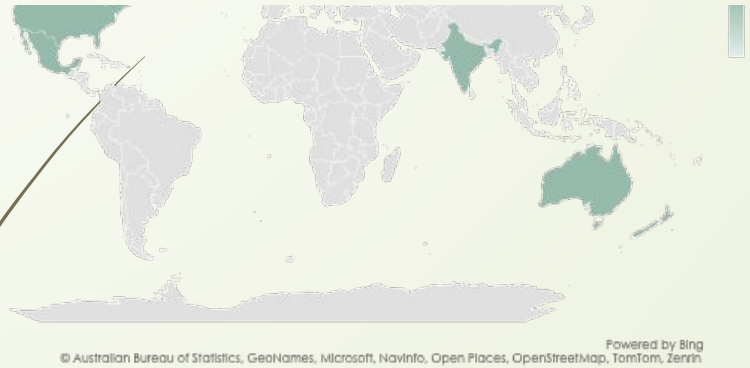
ANALYZING GLOBAL MUSIC TRENDS:

A Data Mining Exploration of Top Spotify Songs in top 7 Countries



By:

Soundarya Chandra Mohan (rp7088)

MOTIVATION

- The music industry's landscape has evolved drastically, necessitating data-driven insights for strategic decisions across the board.
- This study is poised to delve into diverse cultural nuances, unveiling the intricacies of regional music preferences and trends.
- Understanding global music trends becomes pivotal for tailoring strategies that resonate with audiences worldwide in an increasingly interconnected world.
- Embracing data mining techniques within the music realm opens doors to innovation, adaptation, and staying ahead in the competitive music industry for all the stakeholders.
- Thus, By dissecting this rich dataset, the project aims to unravel the underlying attributes contributing to the success of songs in different cultural contexts.

RESEARCH QUESTIONS

1) To what extent do the correlations between song popularity and specific musical attributes (such as danceability, energy, and acousticness) exhibit consistent patterns across diverse countries, and how does it have influence over global music trends?

2) How does the user's sentiment is being related with song's popularity and how these impact artist preference?

3) How does each model's accuracy and predictive capability compare when analyzing Spotify songs? What is the better approach?

4) Identify the most popular genres of music in all countries over time.

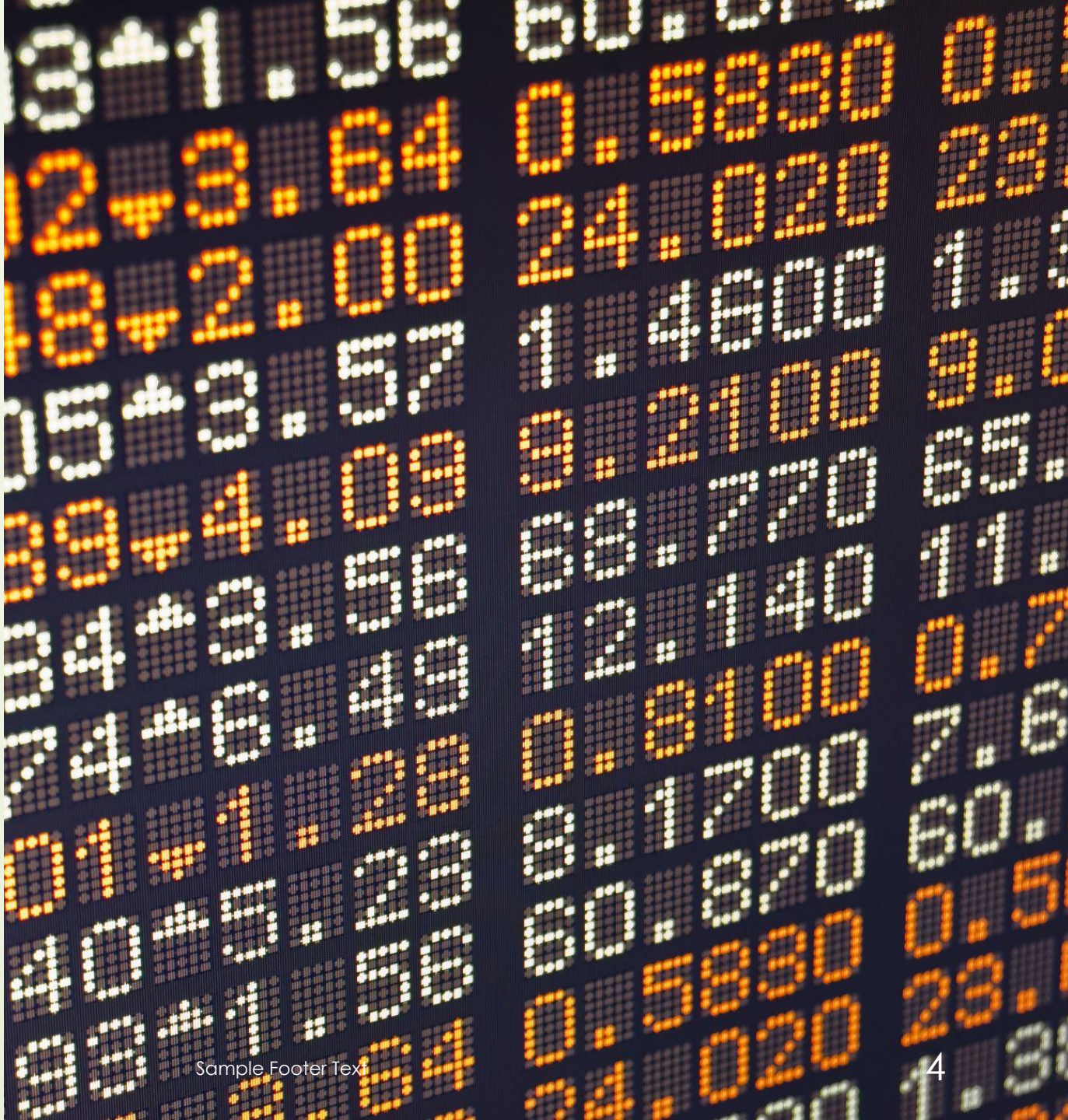
5) To Analyze the relationship between song features

DATASET

Data has been taken from the Kaggle data set of spotify which keeps on updating on daily basis.

Find the link below :

https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated



DATA SET OVERVIEW

The dataset consist of 36523 rows where there are no duplicate values but have missing values in the database.

Features:

spotify_id: The unique identifier for the song in the Spotify database. (type: str)
name: The title of the song. (type: str)
artists: The name(s) of the artist(s) associated with the song. Do split(', ') to convert to a list (type: str)
daily_rank: The daily rank of the song in the top 50 list. (type: int)
daily_movement:The change in rankings compared to the previous day. (type: int)
weekly_movement :The change in rankings compared to the previous week. (type: int)
country:The ISO code of the country of the Top 50 Playlist. If Null, then the playlist if 'Global Top 50'.
snapshot_date:The date on which the data was collected from the Spotify API.
popularity: Song ratings of spotify audience.
is_explicit:Indicates whether the song contains explicit lyrics.
duration_ms: The duration of the track in milliseconds.
album_name:contains the name of the album.
album_release_date: year when the album was actually released.
danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.

DATA SET OVERVIEW

key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation .
loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music.Values below 0.33 most likely represent music and other non-speech-like tracks.
acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
instrumentalness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

UNVEILING CULTURAL RESONANCE : UNDERSTANDING THE PULSE OF GLOBAL MUSIC

- Collection of Comprehensive Spotify data from relevant sources like Kaggle, Spotify API etc which provide daily updated information on top spotify songs in 7 countries(USA,UK,CANADA,MEXICO,INDIA,AUSTRALIA,NEW ZEALAND) including audio features, popularity metrics and regional details.
- Data PreProcessing
- Exploratory Data Analysis

  -Predictive Modeling

  -Sentiment Analysis

METHODOLOGY : PREDICTIVE MODELLING

Predictive modelling is a data driven approach that could help in feature selection,understanding music preferences, and thereby predicts user behavior.

Data Collection and Preprocessing

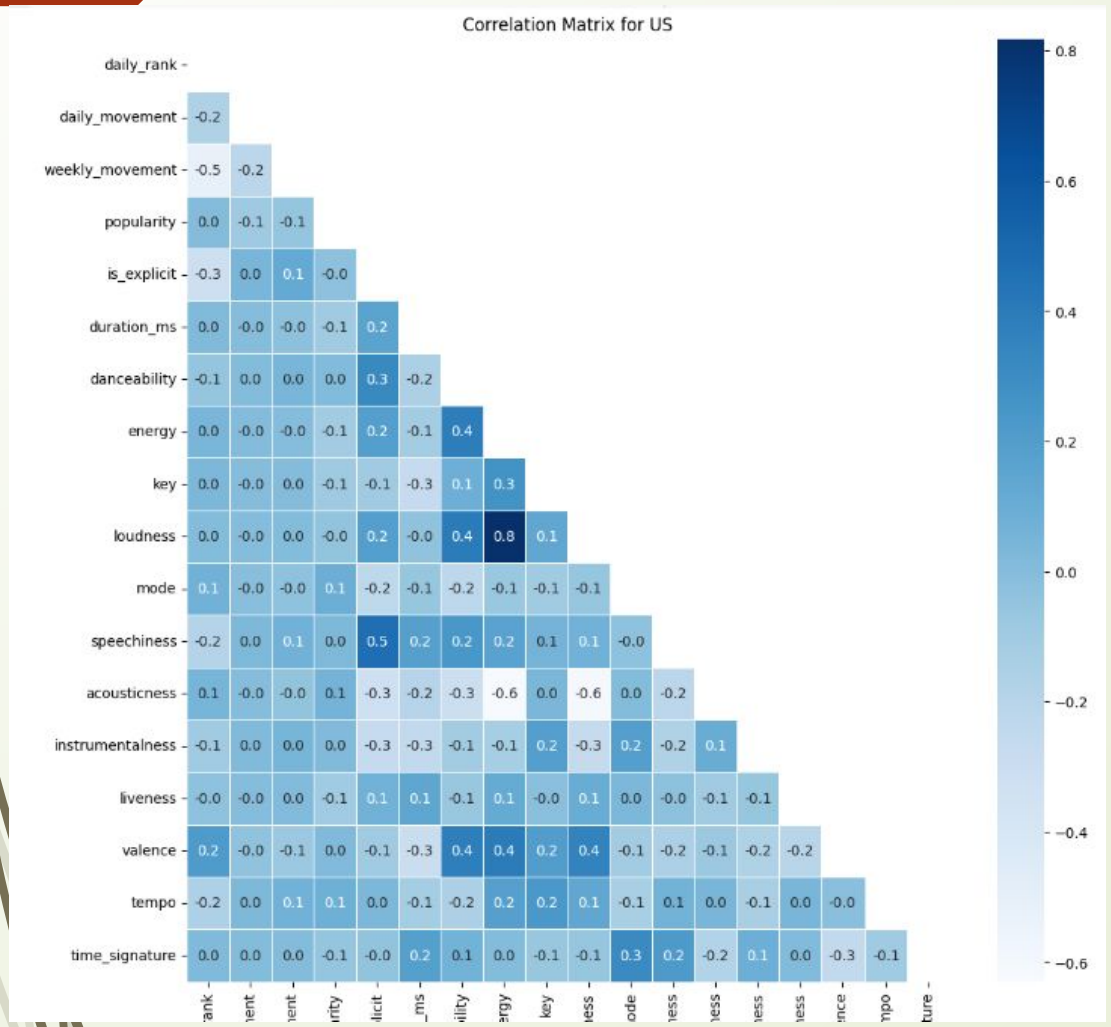
1. Cleaning raw data- Importing packages and reading the dataset.
2. Removing null values in the given data set to produce clean data.
3. Filtering data for 7 countries out of 73.

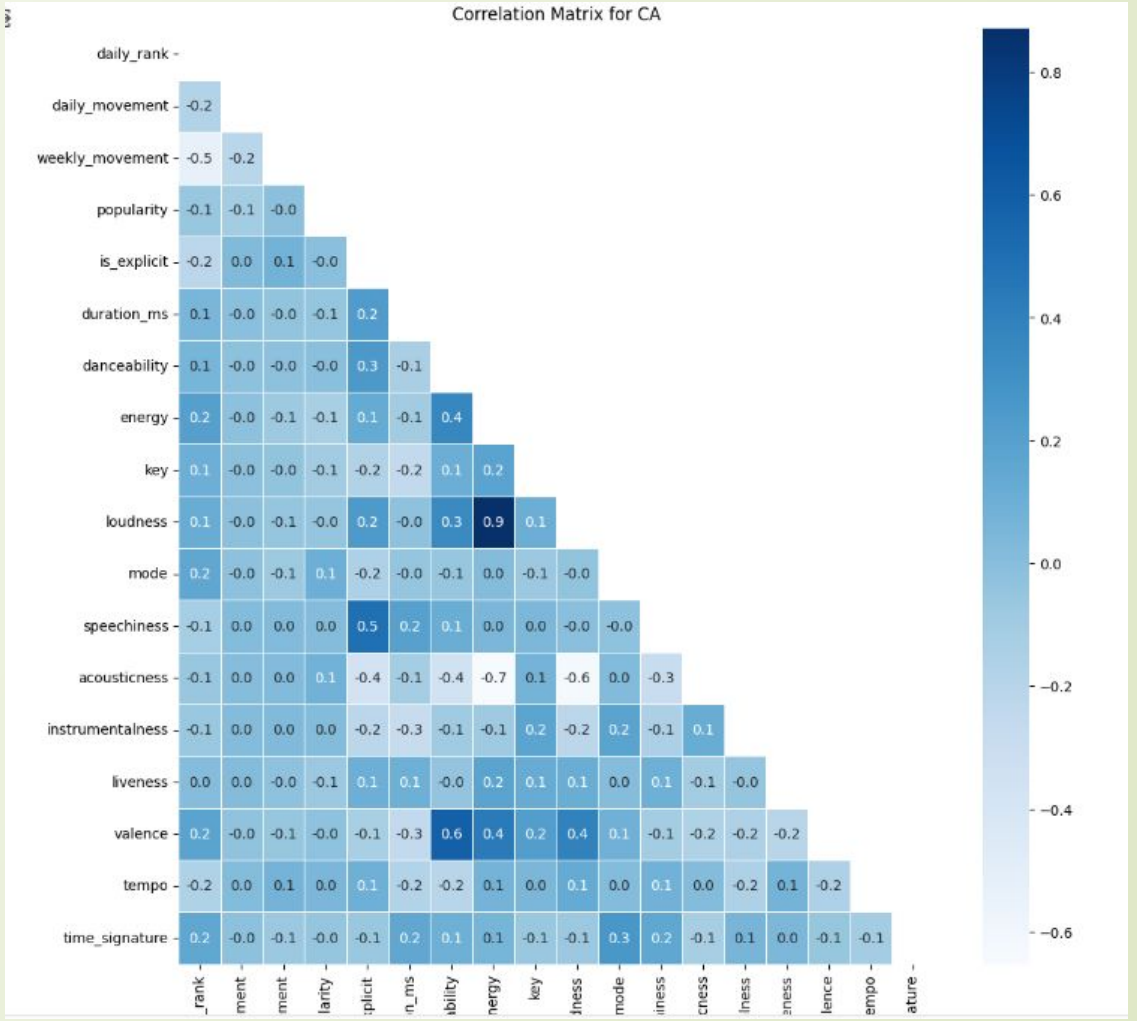
METHODOLOGY : PREDICTIVE MODELLING

Feature selection for predicting popularity of songs over 7 countries:

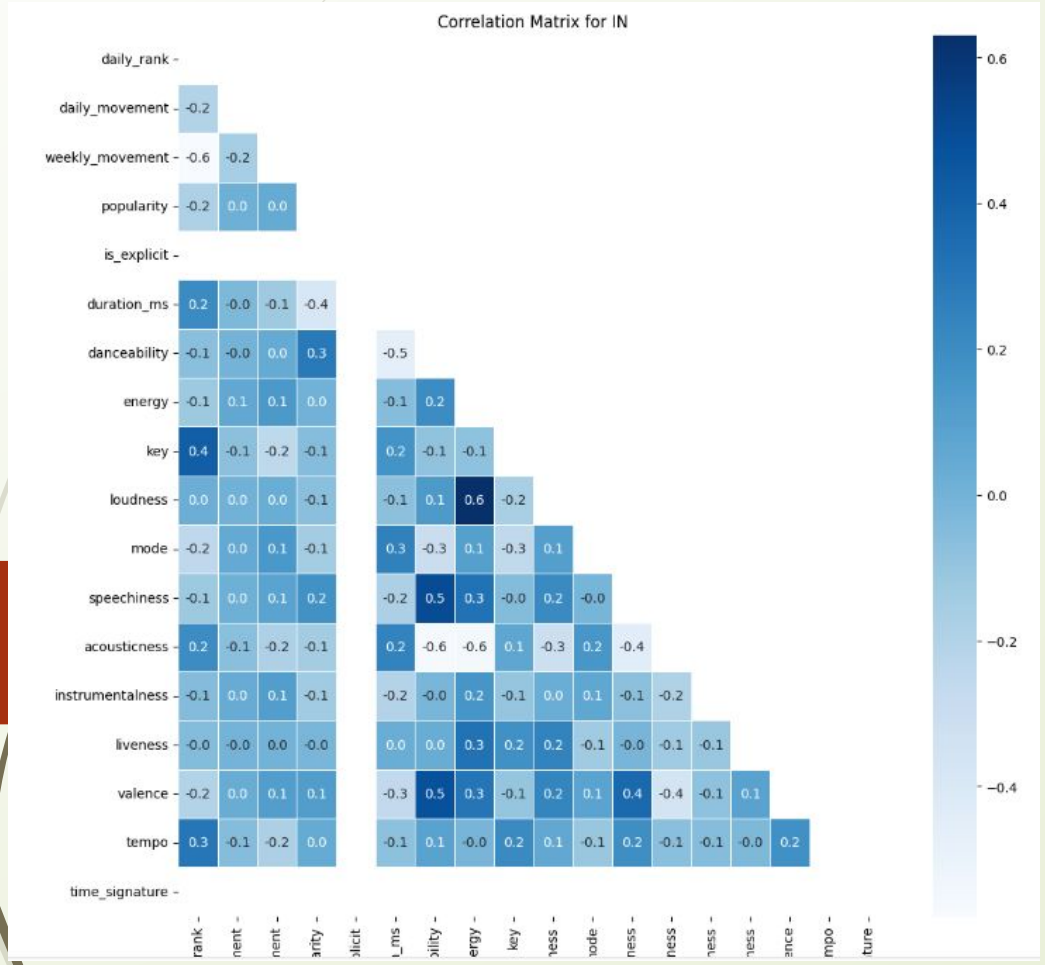
Through this modelling approach, correlating how "popularity" attribute is significant over other features.

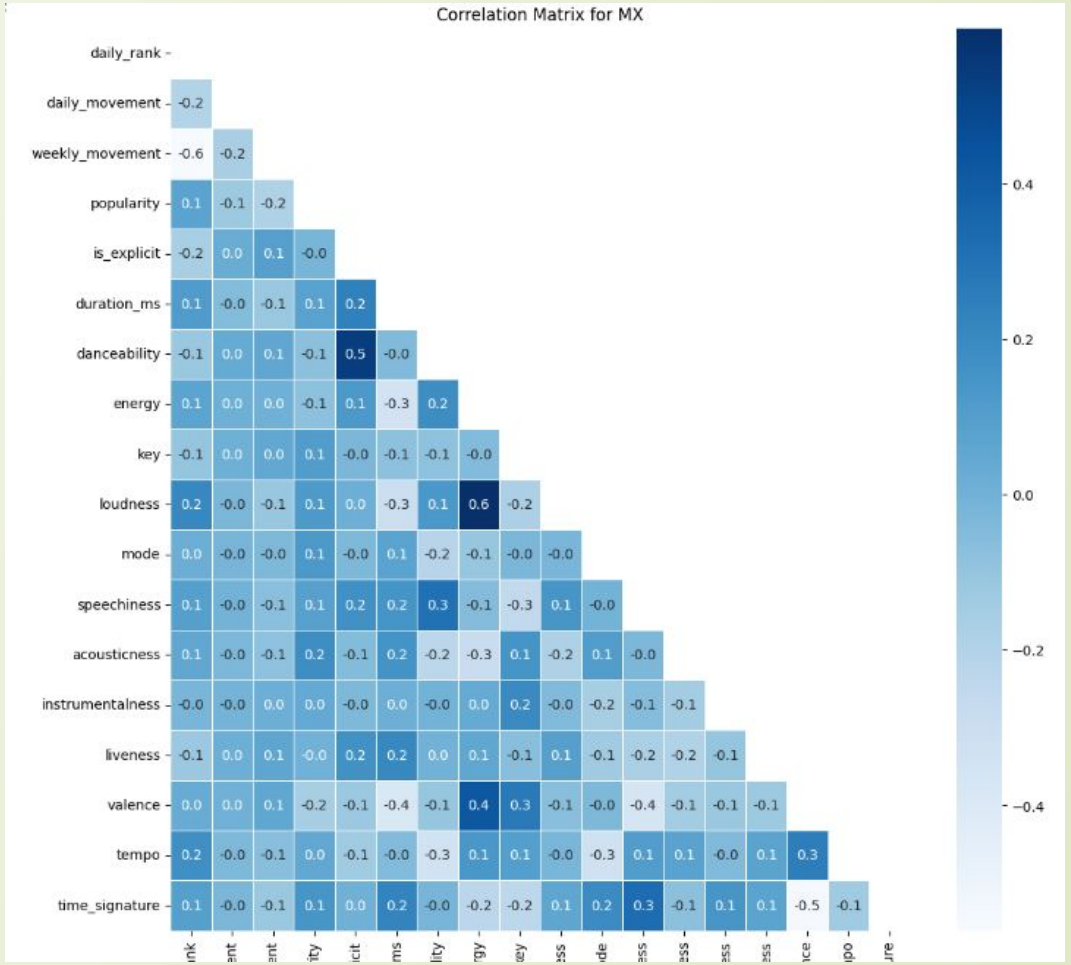
METHODOLOGY : PREDICTIVE MODELLING(US,CA)



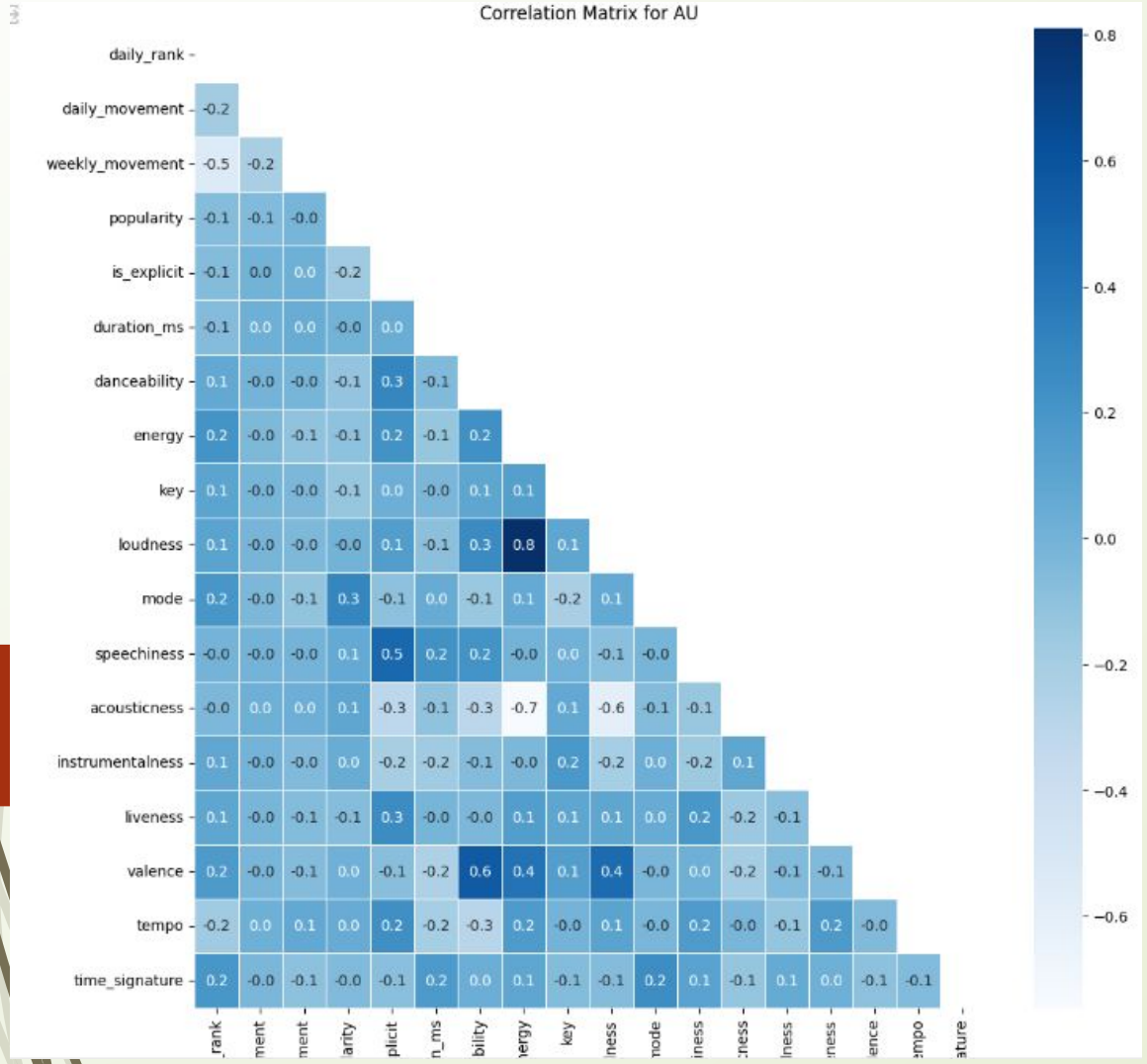


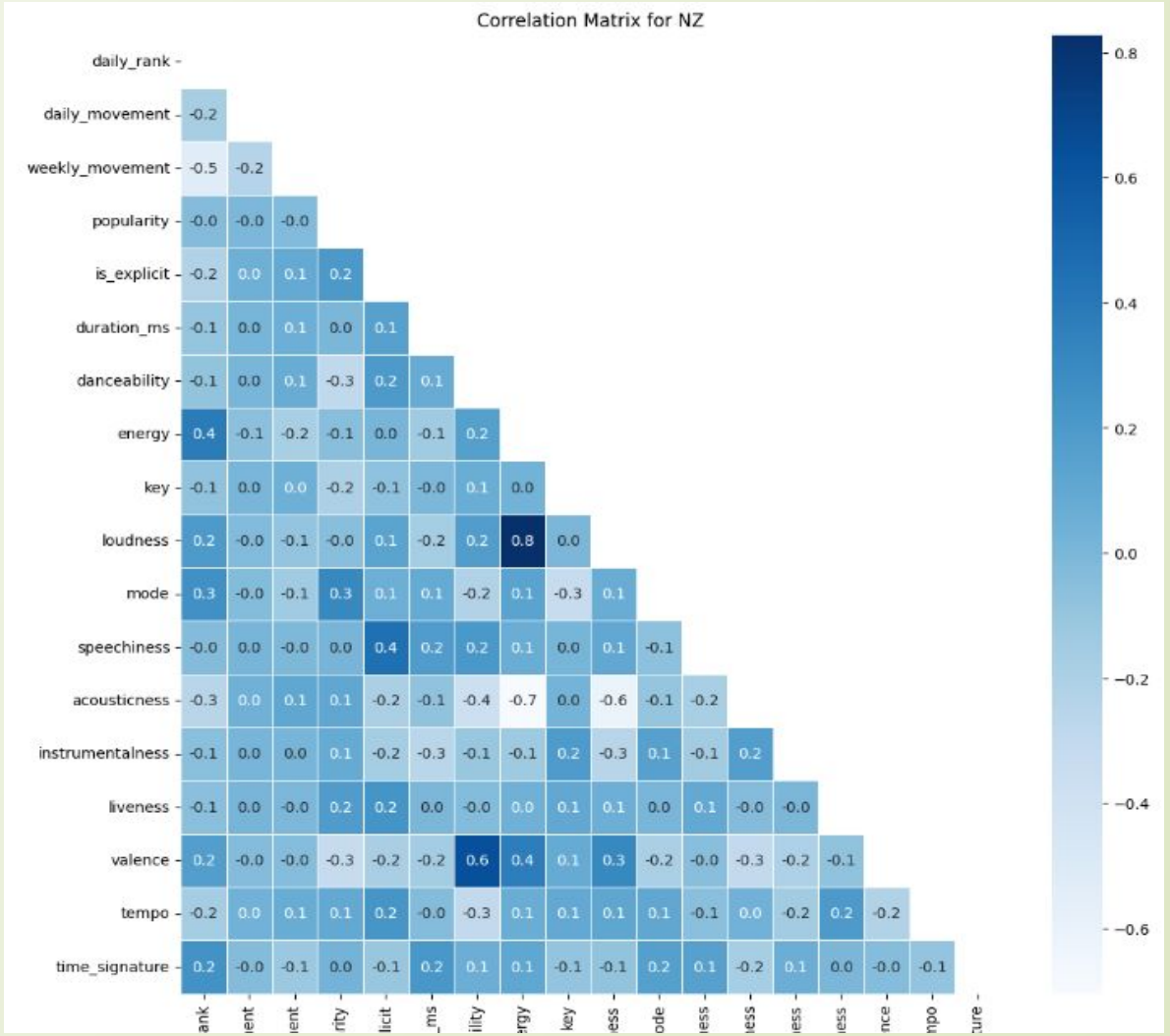
METHODOLOGY : PREDICTIVE MODELLING(IN,MX)



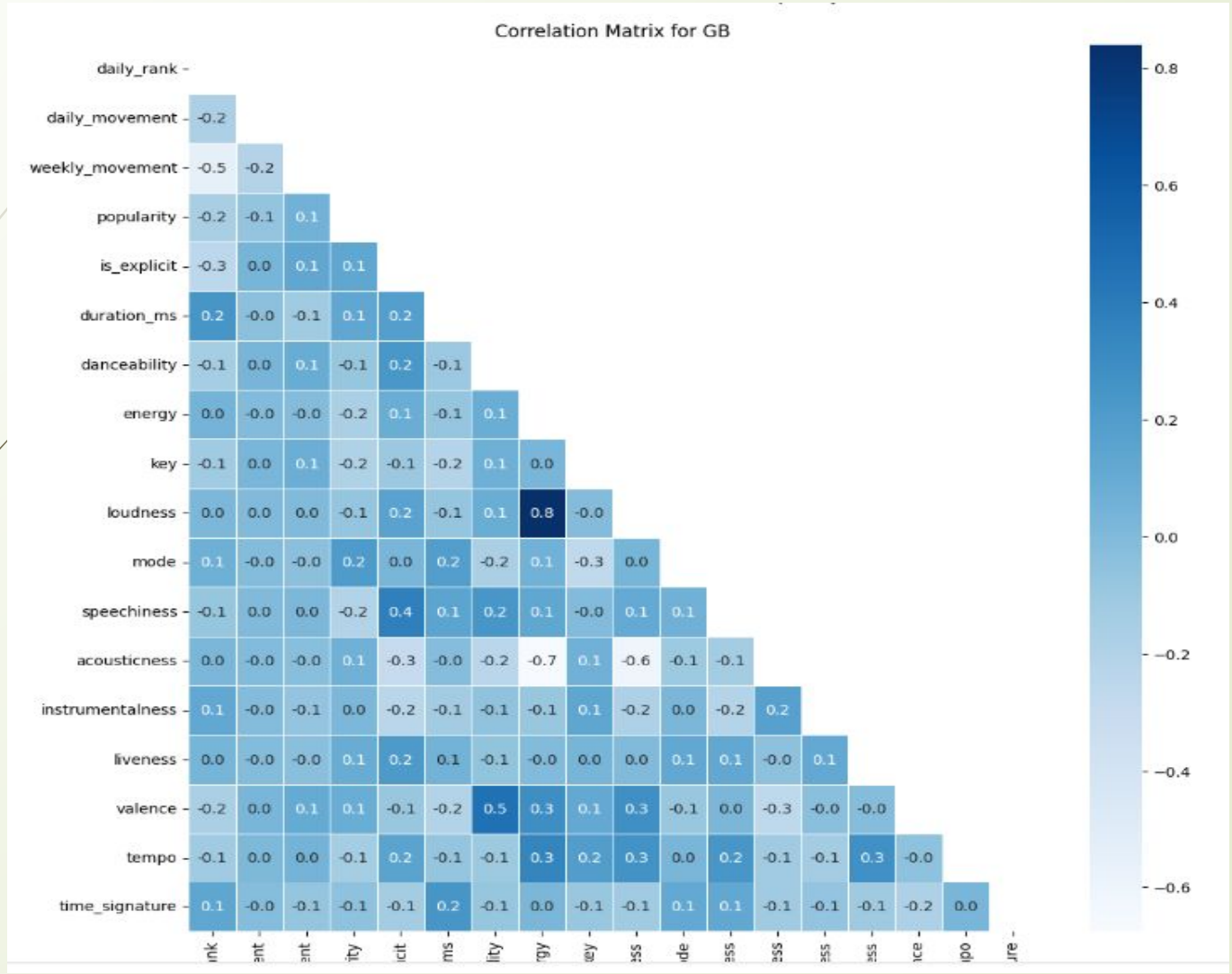


METHODOLOGY : PREDICTIVE MODELLING(AU,NZ)



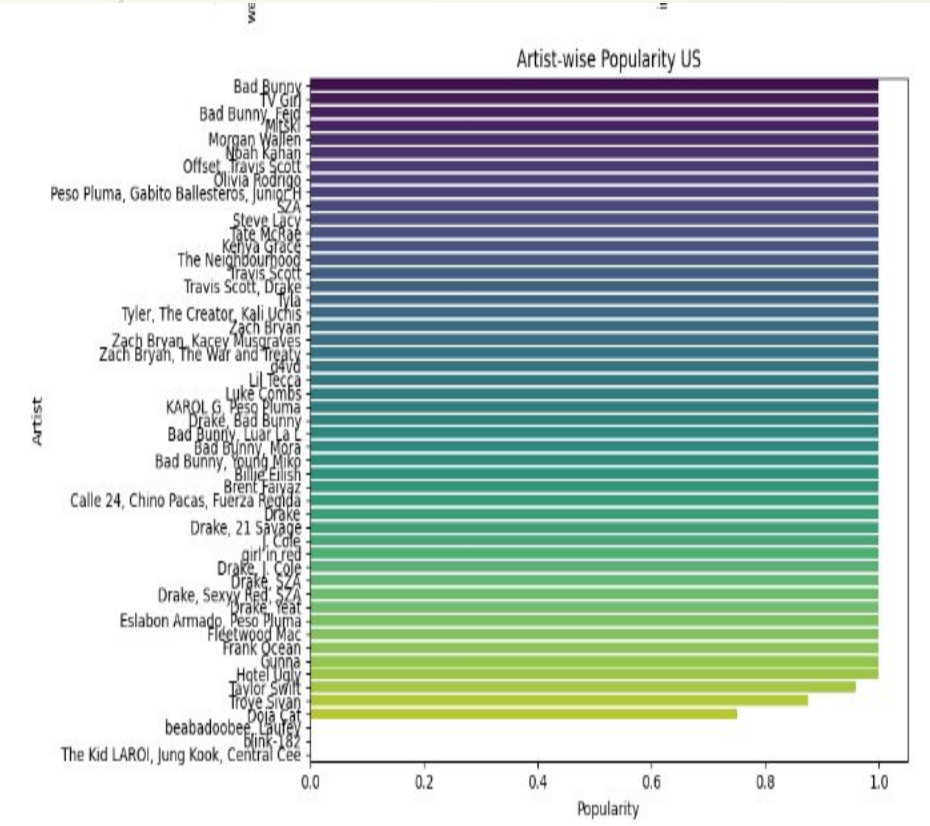


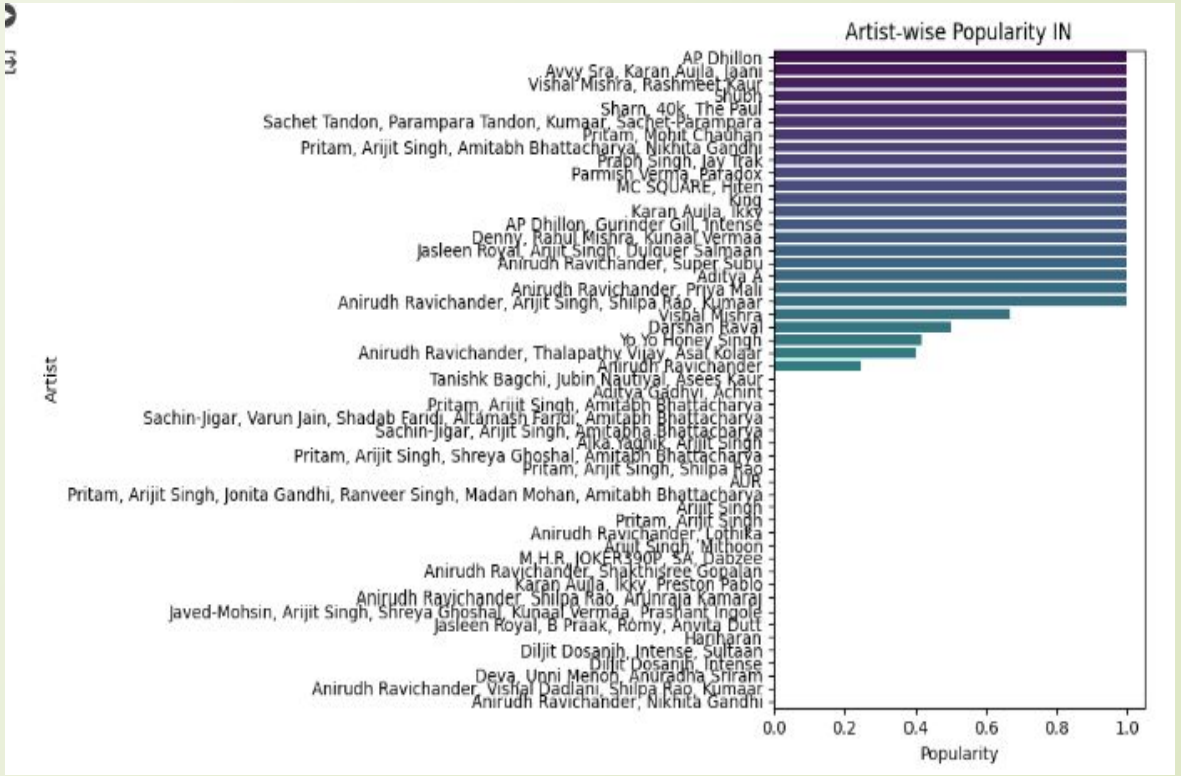
METHODOLOGY : PREDICTIVE MODELLING(GB)



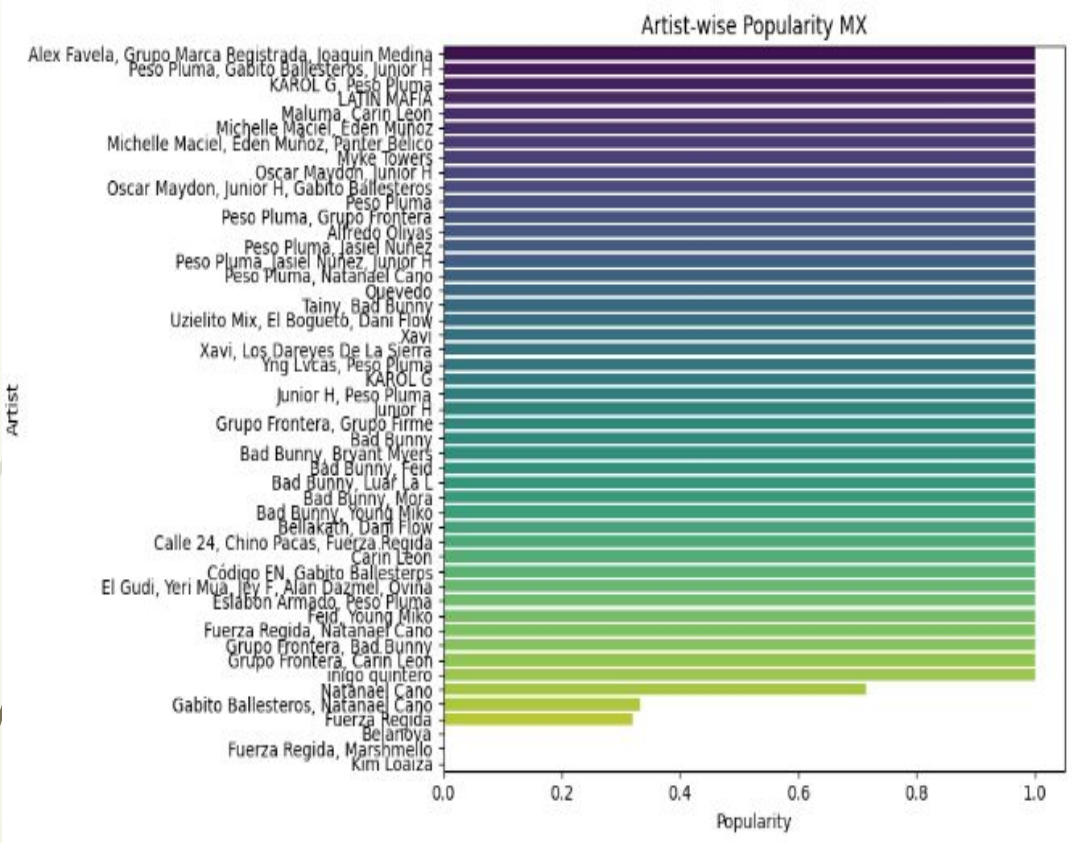
METHODOLOGY : PREDICTIVE MODELLING

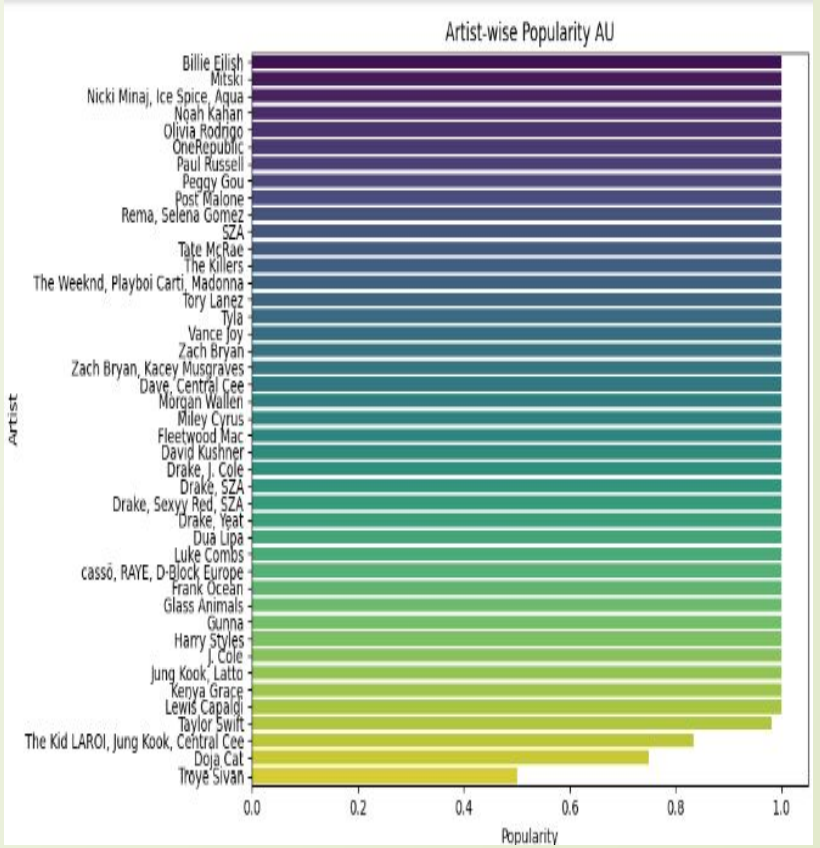
? *Through the above correlation, we can predict how popularity plays major role in artist preferences over 7 countries. Using this top 100 artist recommendation of songs among countries could be predicted.*



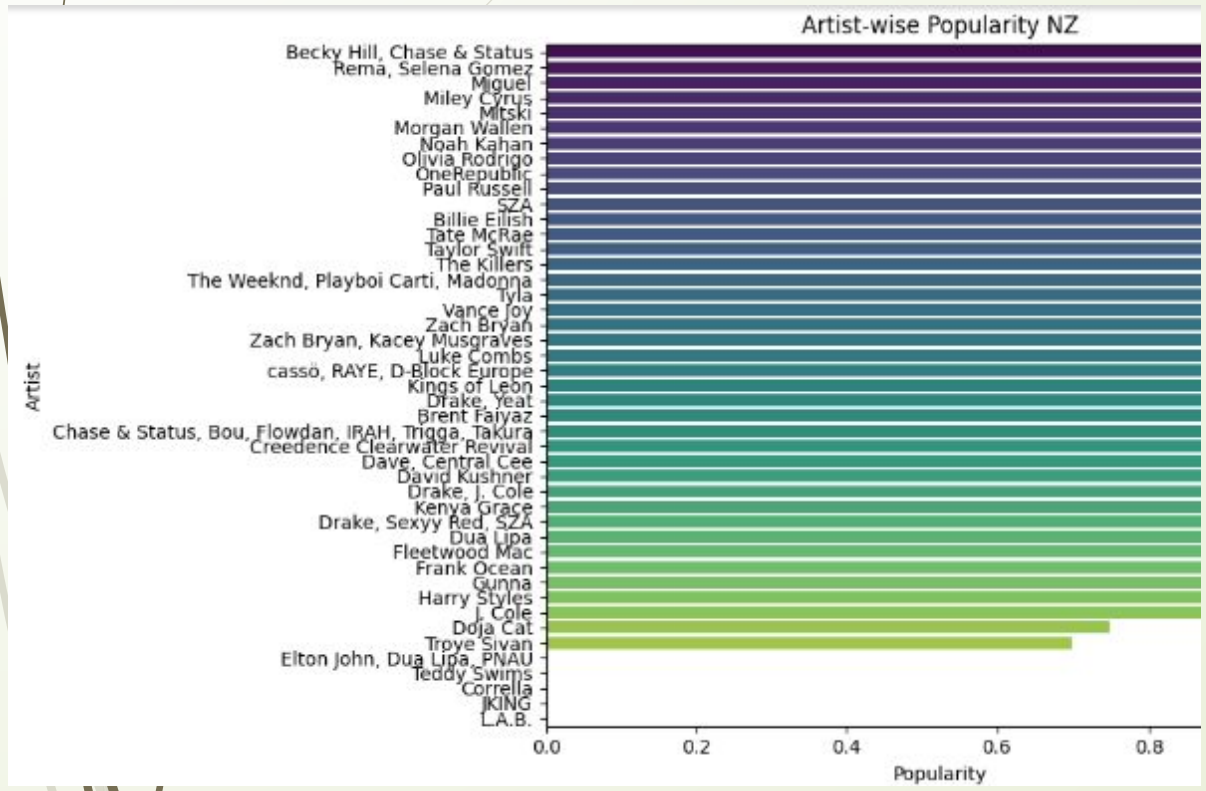


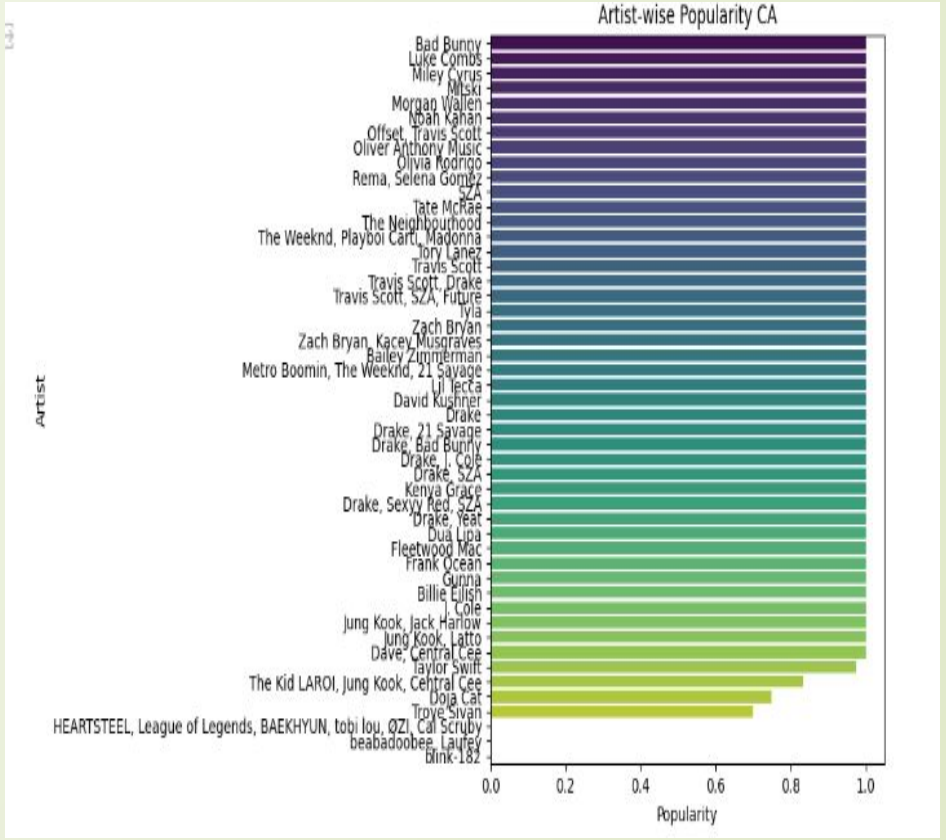
Artist popularity (MX,AU)



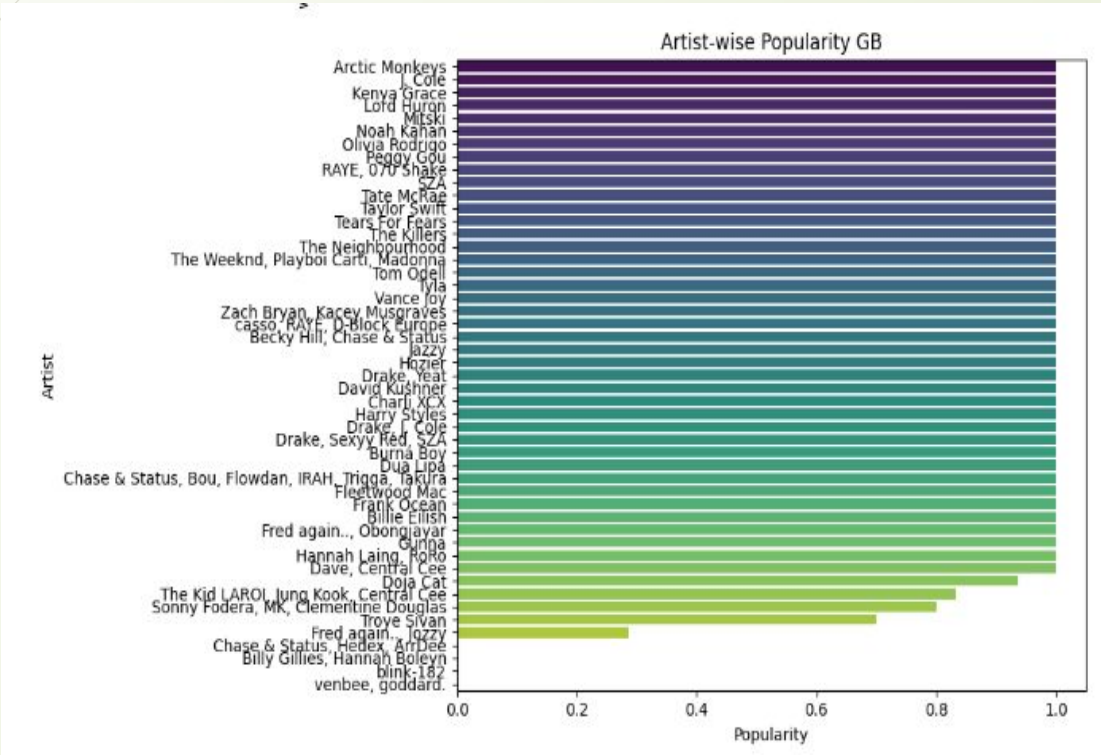


Artist popularity (NZ,CA)



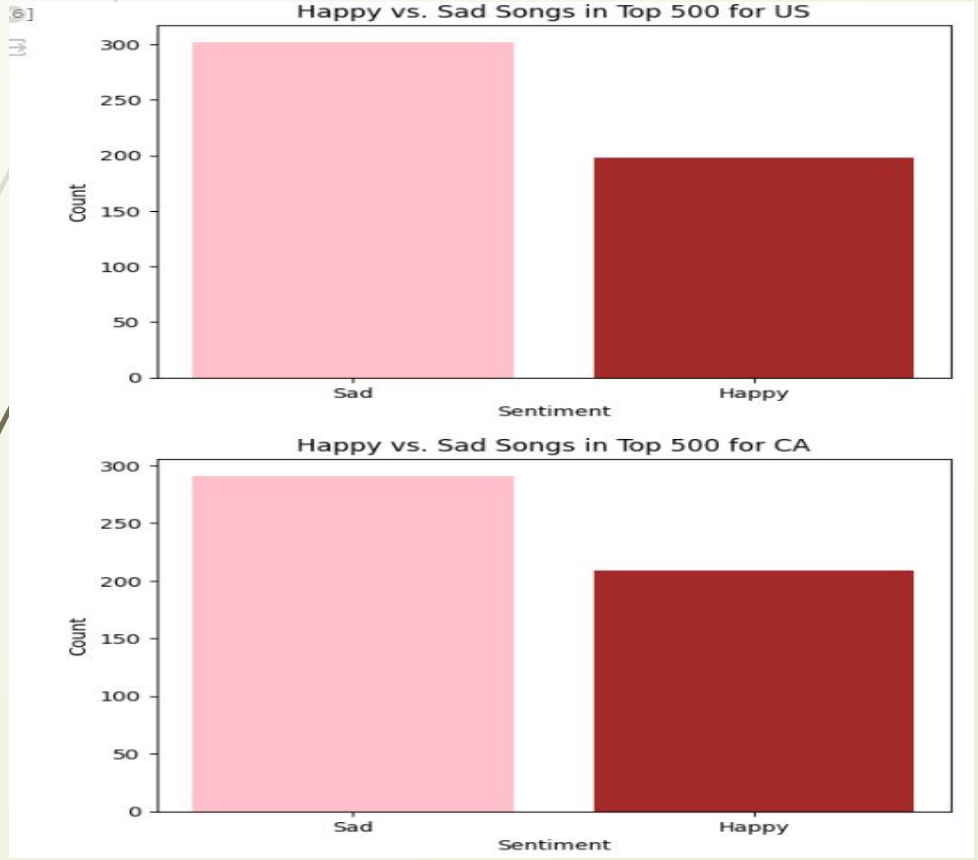


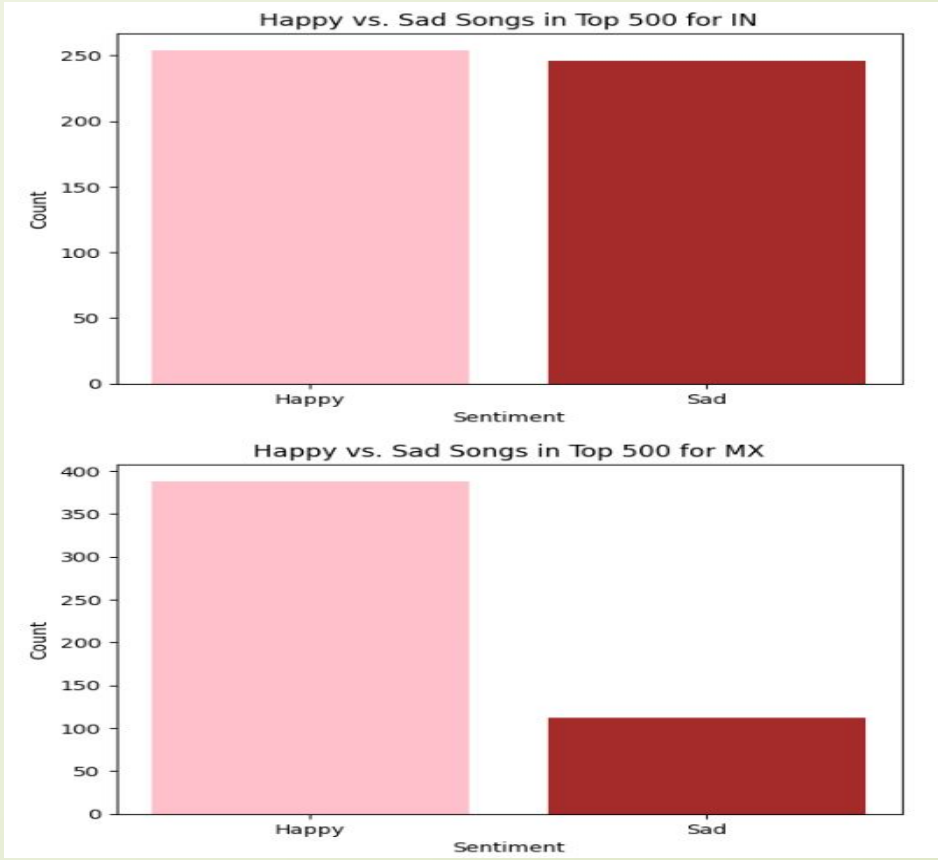
Artist popularity (GB)



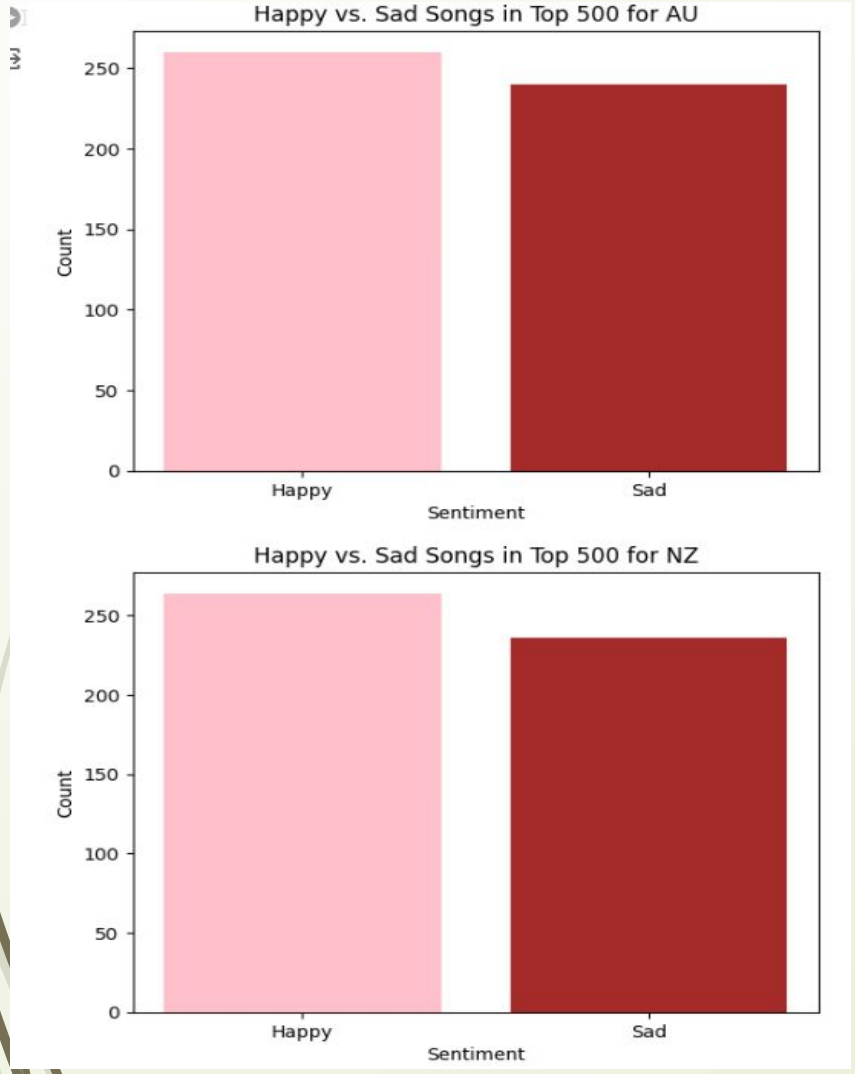
Features supporting Sentiment Analysis

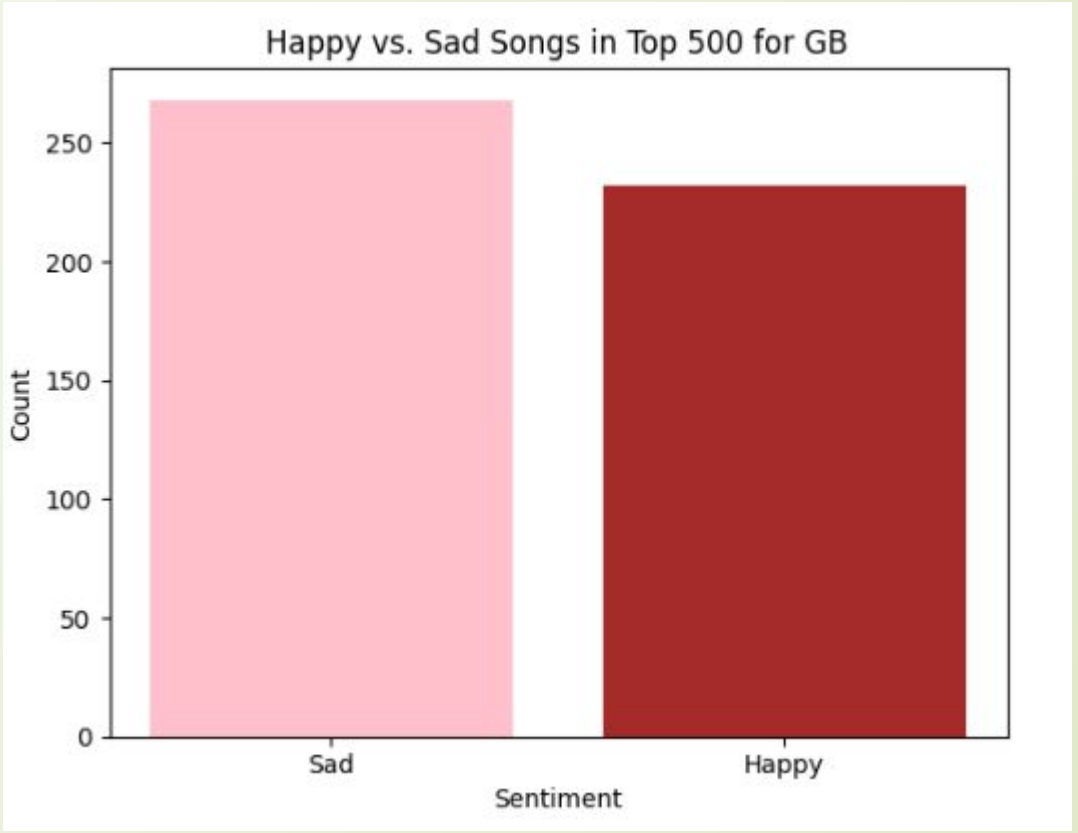
Using feature like "Valence", the mindset of user could be adhered for improving user engagement and recommendation systems. Here predictive models could utilize this feature for improving song recommendation based on emotional preferences.





Feature Engineering

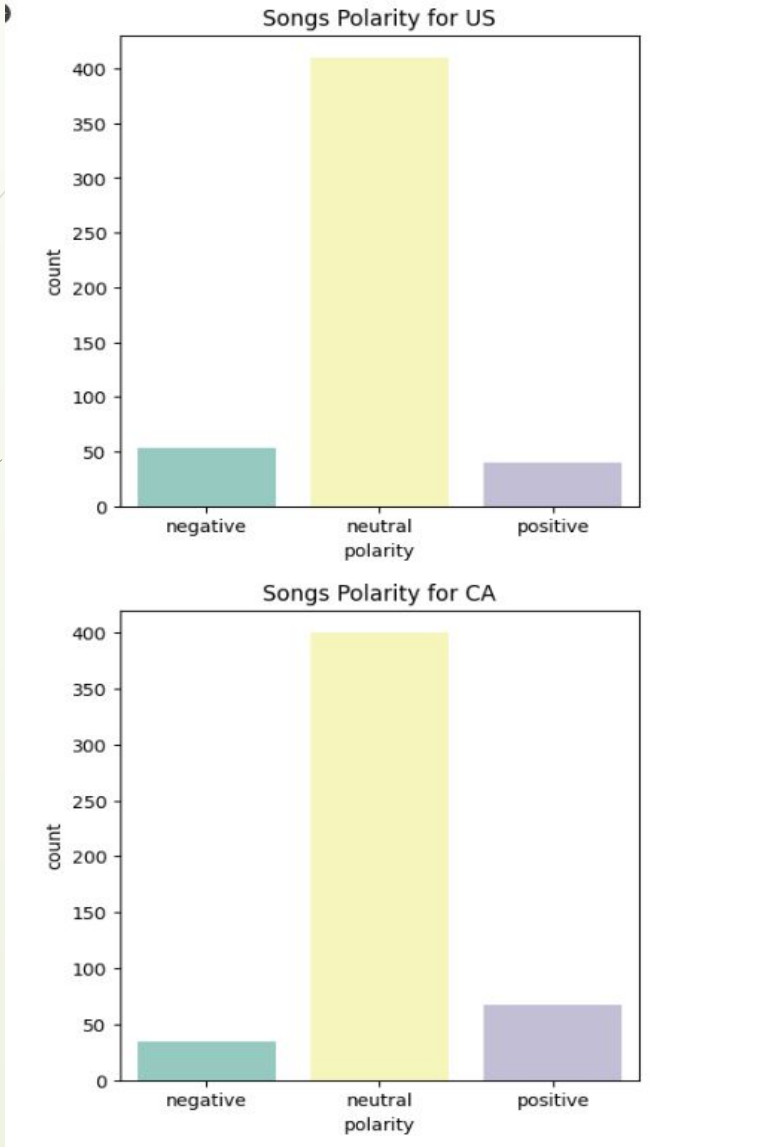


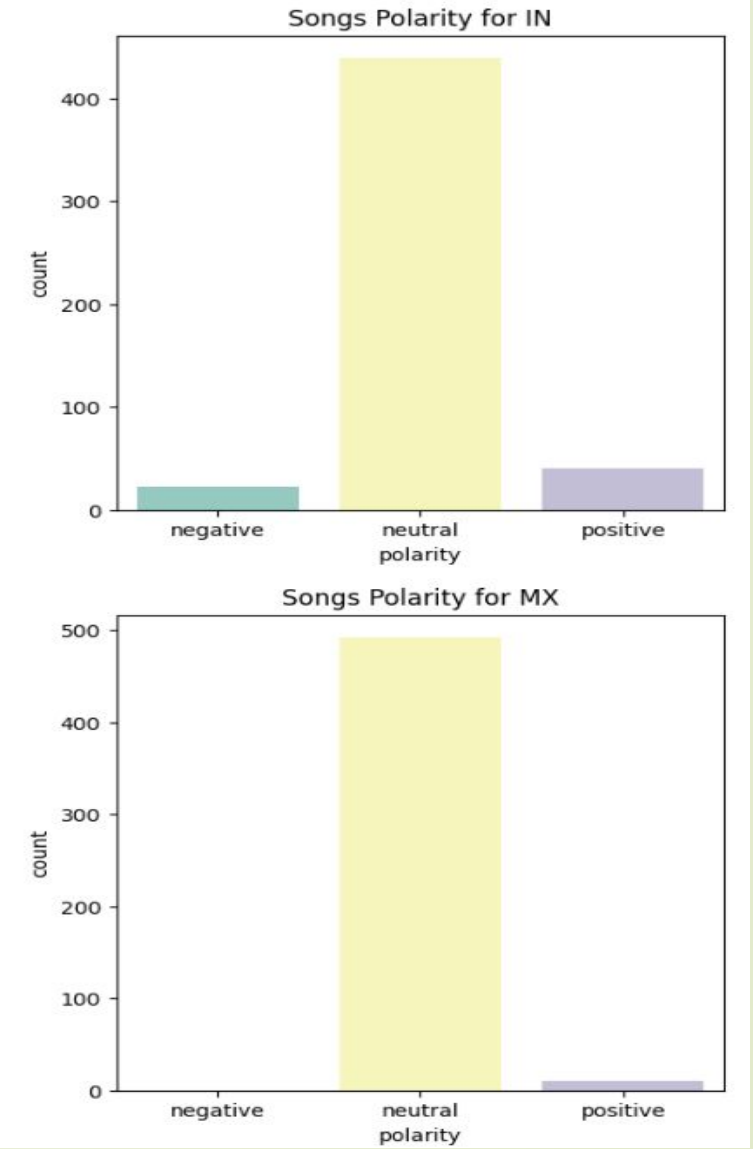


METHODOLOGY : SENTIMENT ANALYSIS

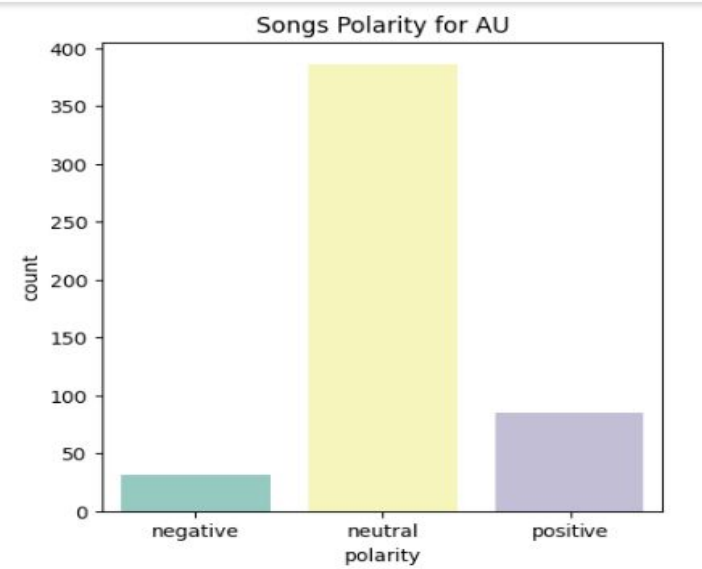
- Perform sentiment analysis on finding polarity associated with the top Spotify songs in 7 countries. The goal is to understand the sentiment expressed through comparison.
- The polarity determines the emotion of the songs through degree of positive, negative or neutral through supporting song name attribute indirectly.

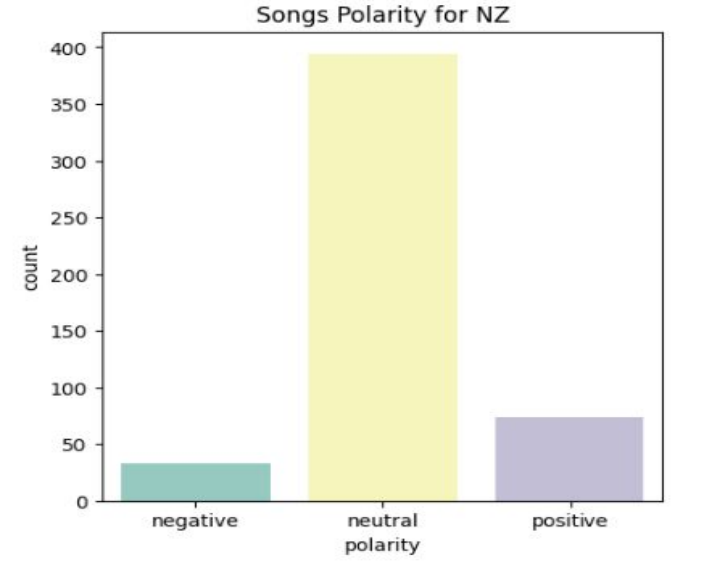
METHODOLOGY : SENTIMENT ANALYSIS

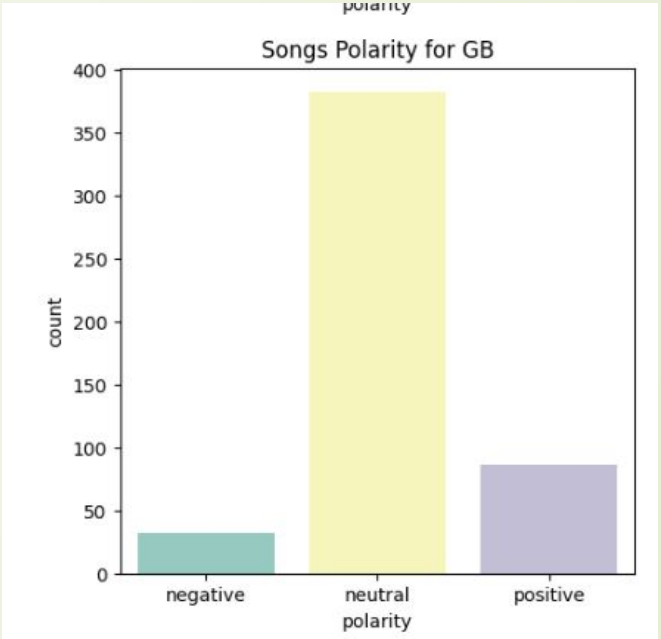




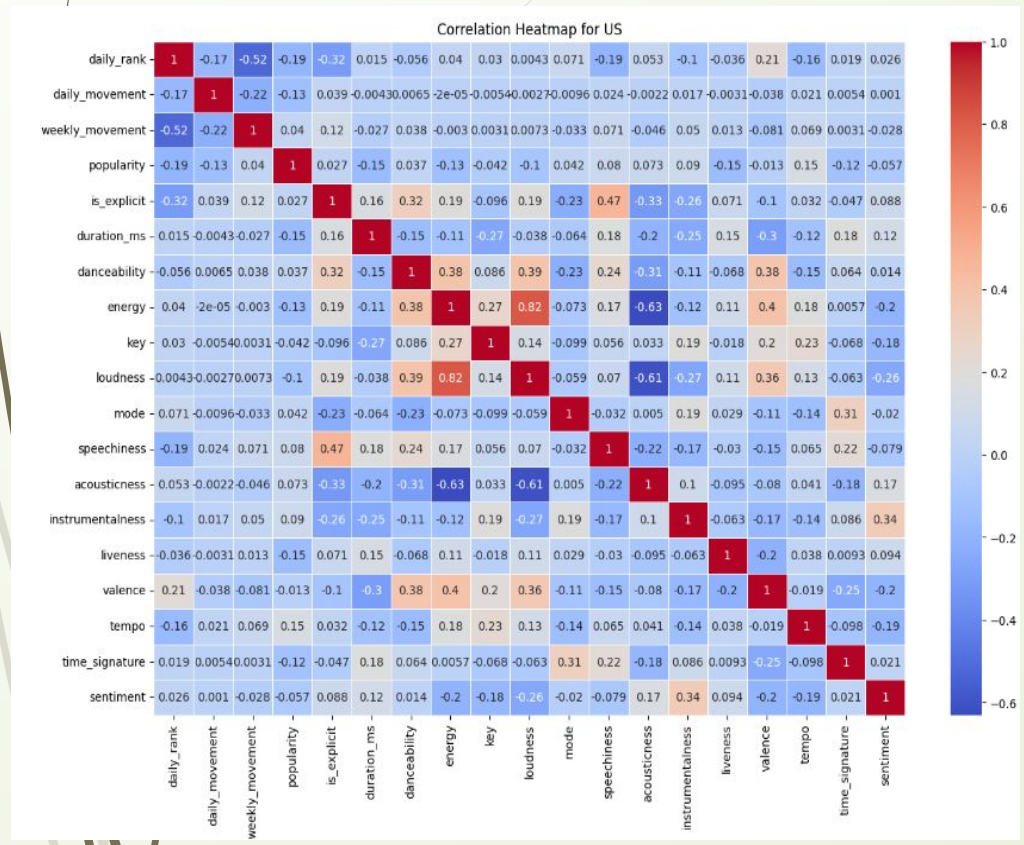
METHODOLOGY : SENTIMENT ANALYSIS

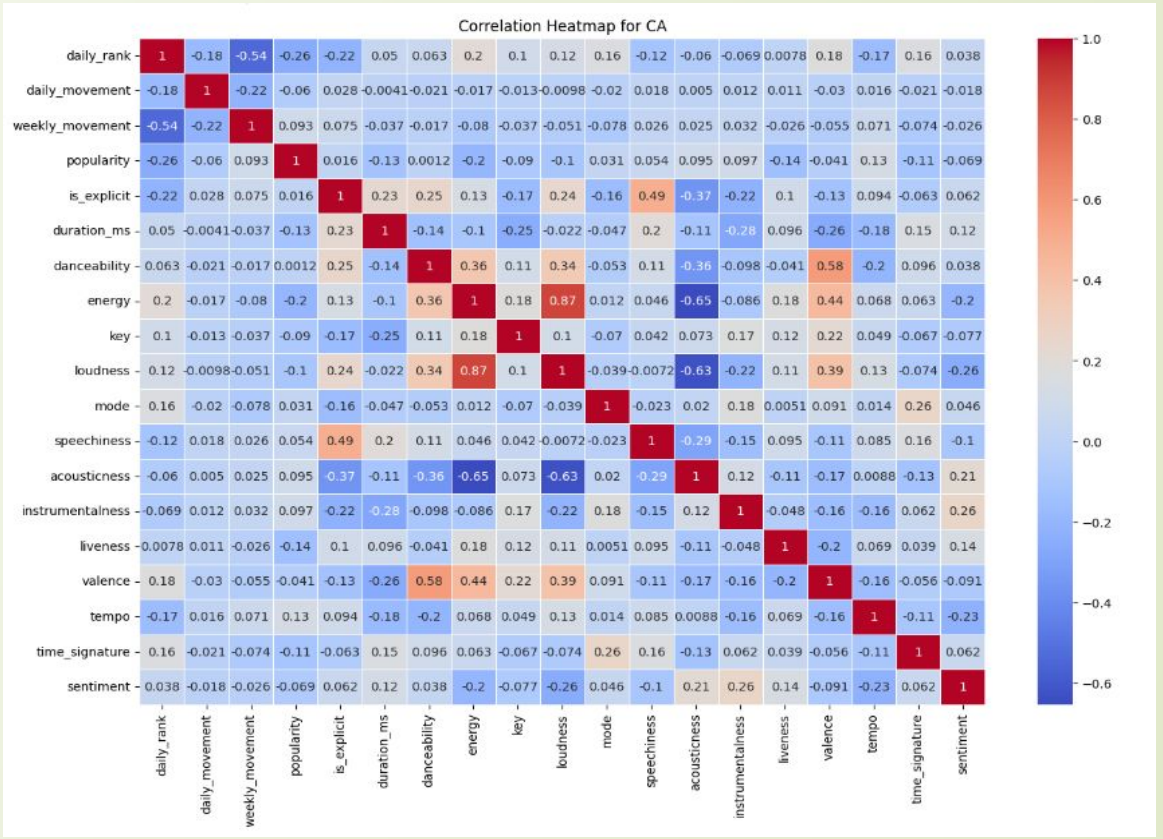




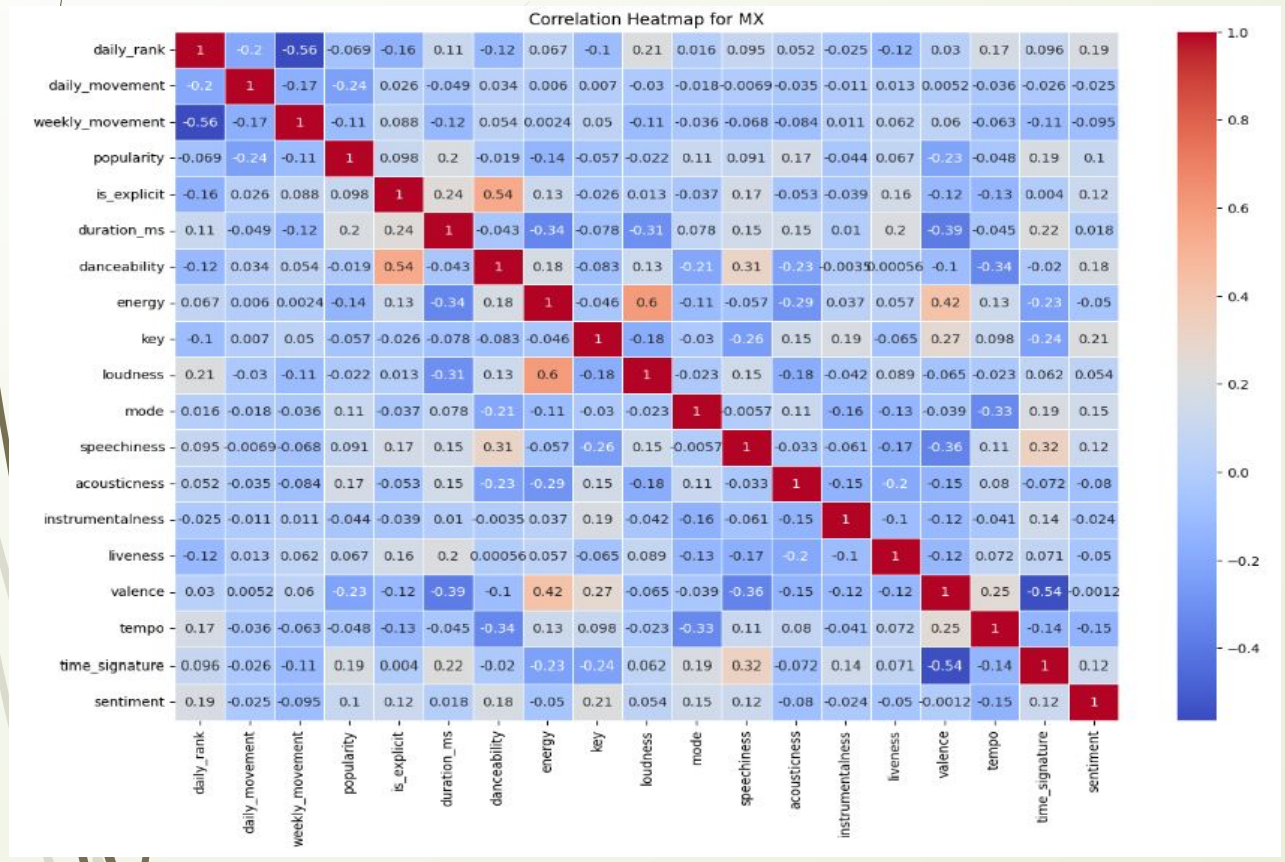


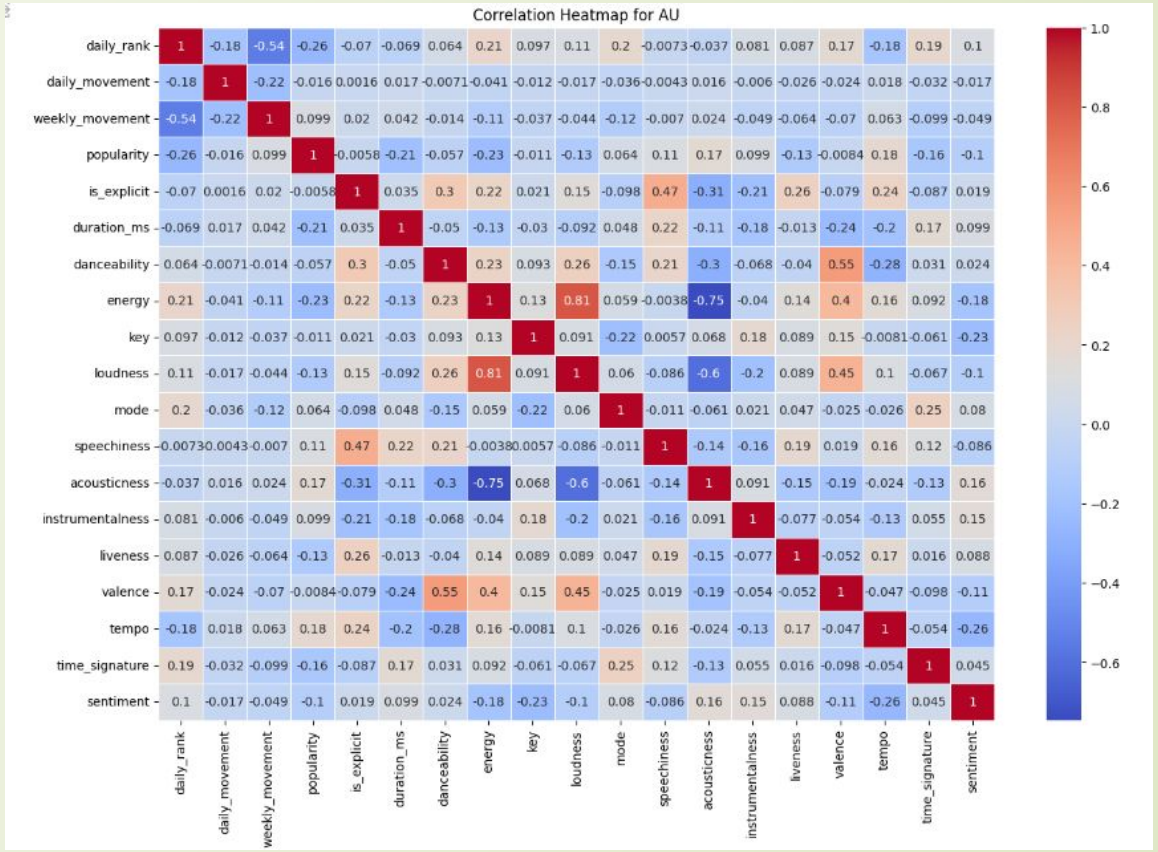
Representation of sentiment over other metrics



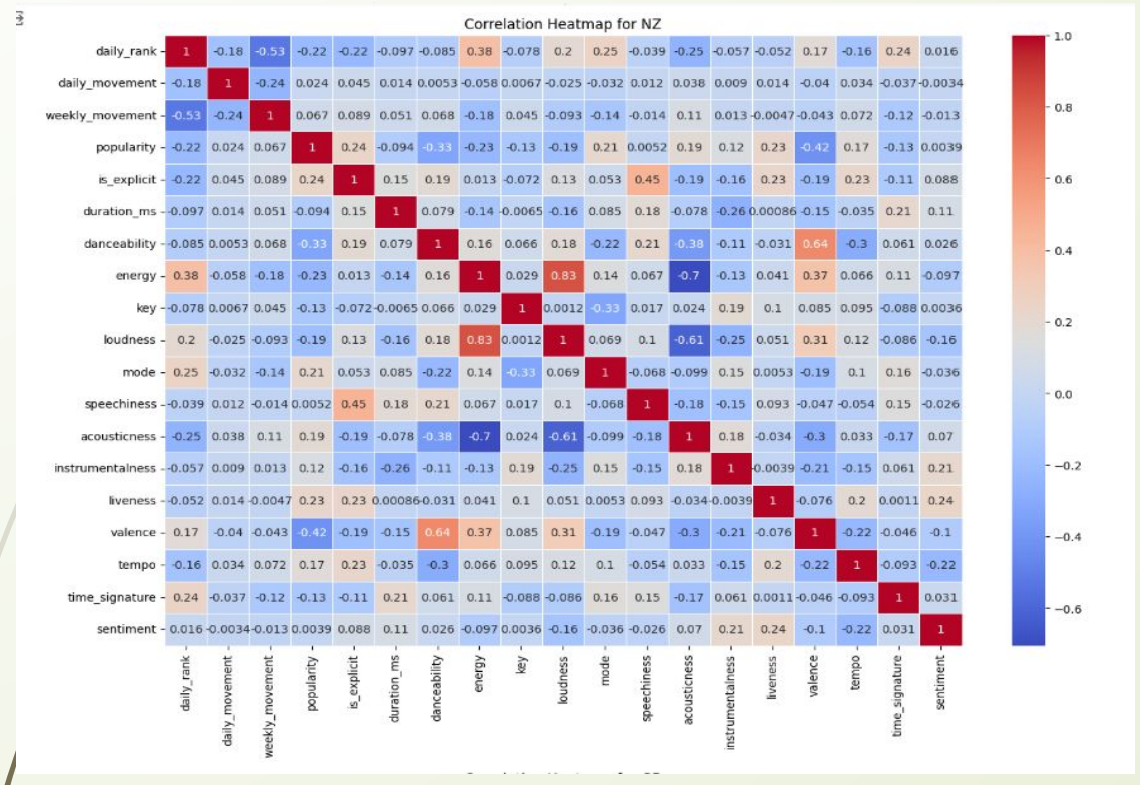


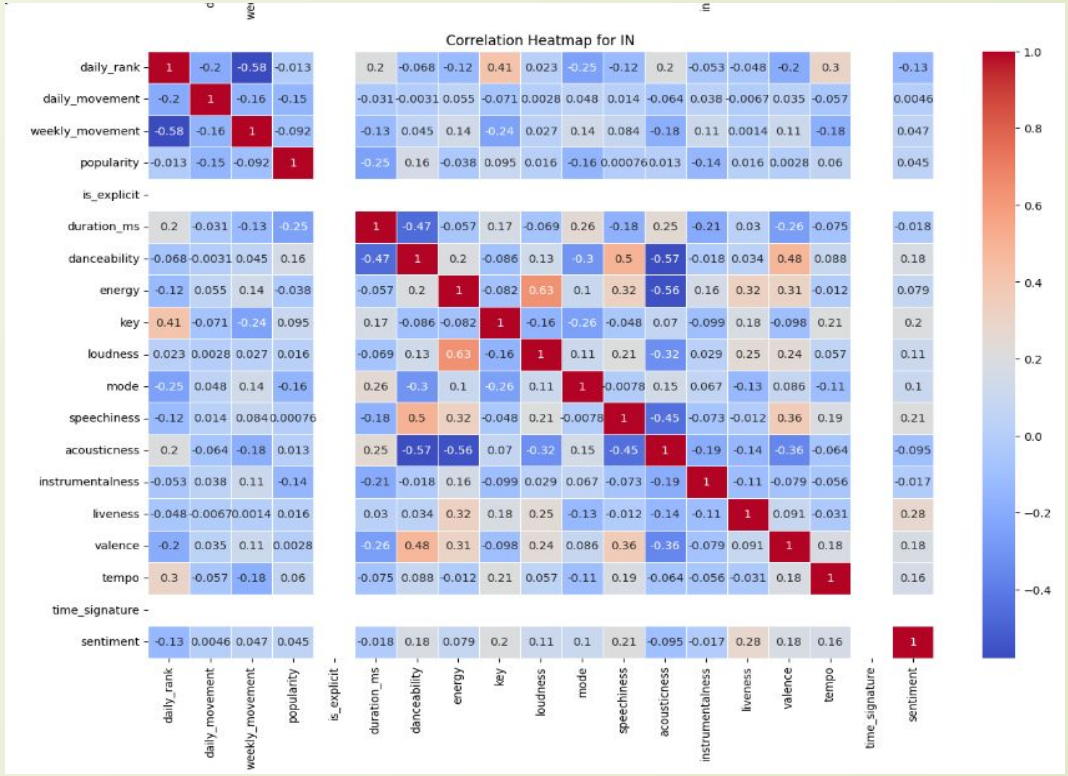
Representation of sentiment over other metrics



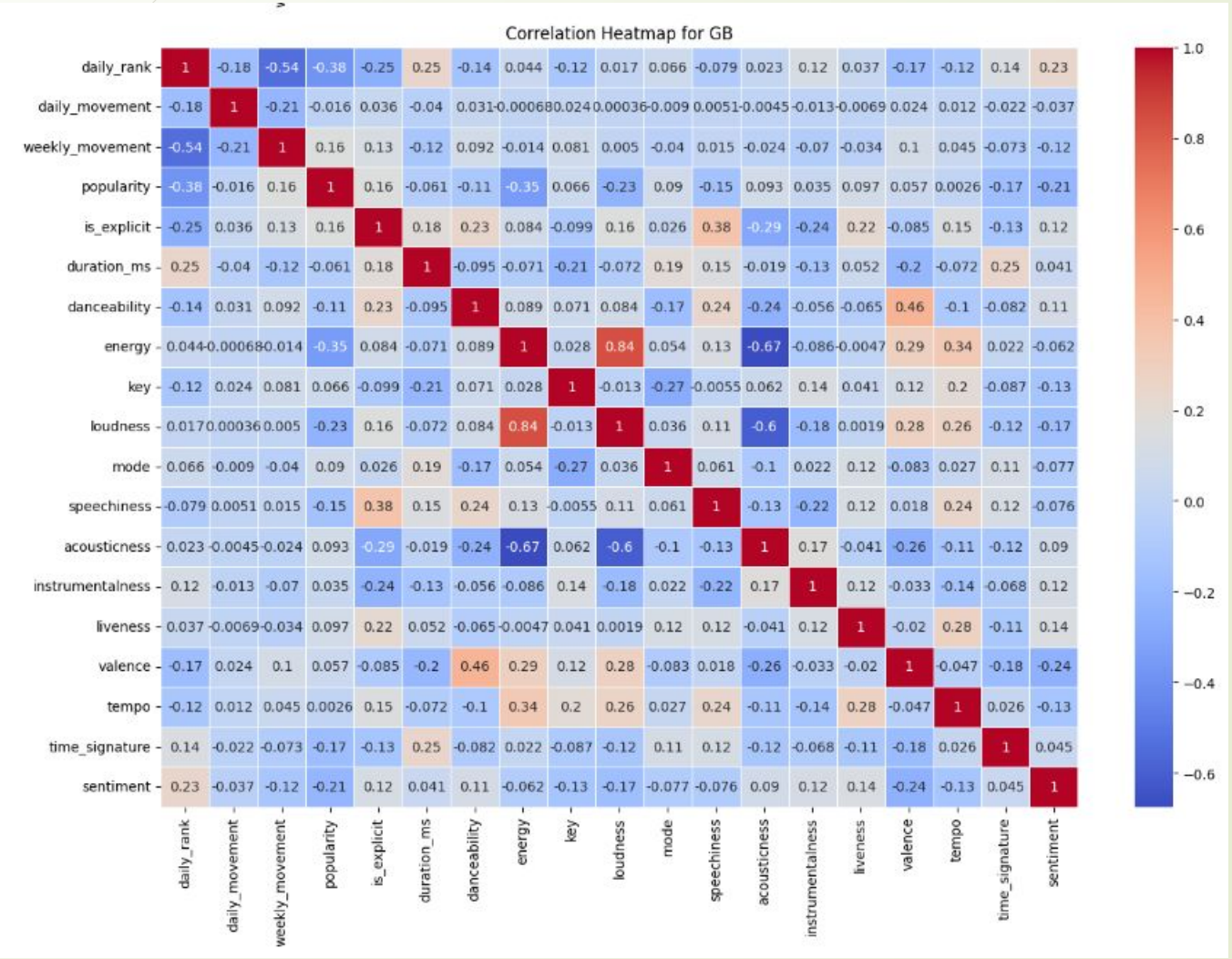


Representation of sentiment over other metrics





Representation of sentiment over other metrics



COMAPARITIVE ANALYSIS OF ML Classifiers

Classification has become a very valuable tool where a large amount of data is used on a wide range of decisions.

The main significance of classification is to classify data from our large dataset to find patterns out of it.

Nevertheless, it is very important to choose the best classification algorithm which is also called as the classifier.

All these classifiers have their own efficiency and have an important role in identifying the set of populations based on the training datasets.

To choose the best classifiers among the four classifiers, the classifiers performance is required to be evaluated based on the performance metrics.

The performance metrics of these classifiers were determined using accuracy.

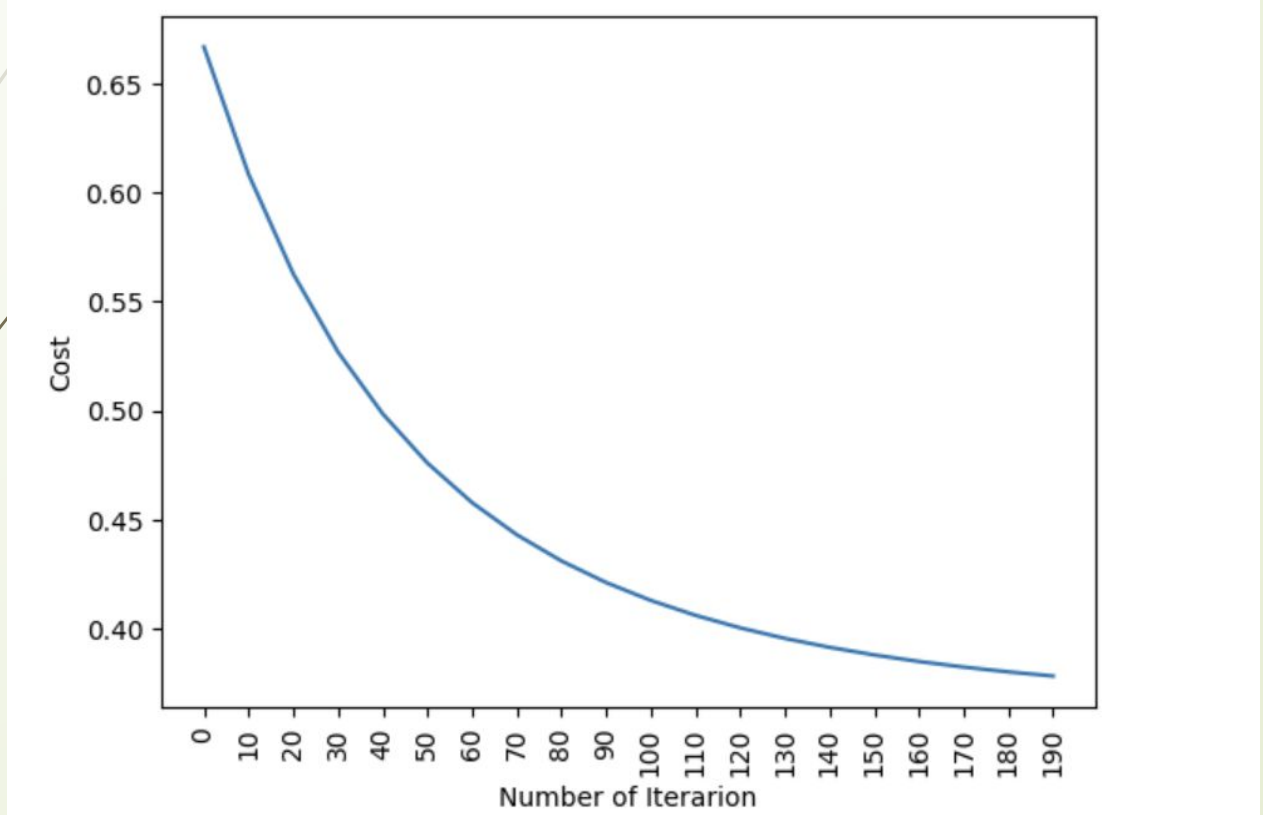
Therefore, this part of research aims to conduct comparative evaluation on our dataset between three classifiers which are

- - Logistic Regression
- - Naive Bayes
- - Random Forest

LOGISTIC REGRESSION

- LR is one of the basic classification method is used prediction of categorical variables.Our problem has two possible outputs popular(1) and unpopular(0) which is suitable for binary logistic regression. Since it is a probability value that we want to get from the problem, we obtained a value between [0,1] using the sigmoid function.
  $\alpha(z) = 1/(1+e-z)$
- Binary cross entropy is used for the loss function and gradient descent for the update the parameters.

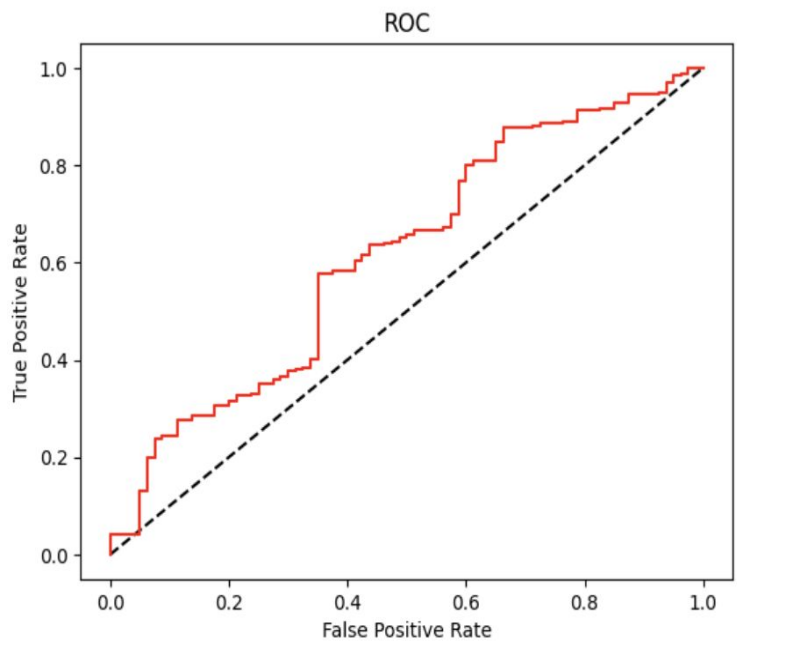
LOGISTIC REGRESSION - EVALUATION



This is the Result we got for our data set. When the trained data sets and the test data sets are calculated, the resultant tained accuracy and test accuracy are as follows.

LOGISTIC REGRESSION – ROC CURVE

Below image, is the ROC curve, which we got as per our dataset.



LOGISTIC REGRESSION – IDEAL ROC CURVE

On the plots below, the green line represents where TPR = FPR, while the blue line represents the ROC curve of the classifier. If the ROC curve is exactly on the green line, it means that the classifier has the same predictive power as flipping a coin.

On the left plot the blue line is relatively close to the green one, which means that the classifier is bad. The rightmost plot shows a good classifier, with the ROC curve closer to the axes and the "elbow" close to the coordinate (0,1). The middle one is a good enough classifier, closer to what is possible to get from real-world data.

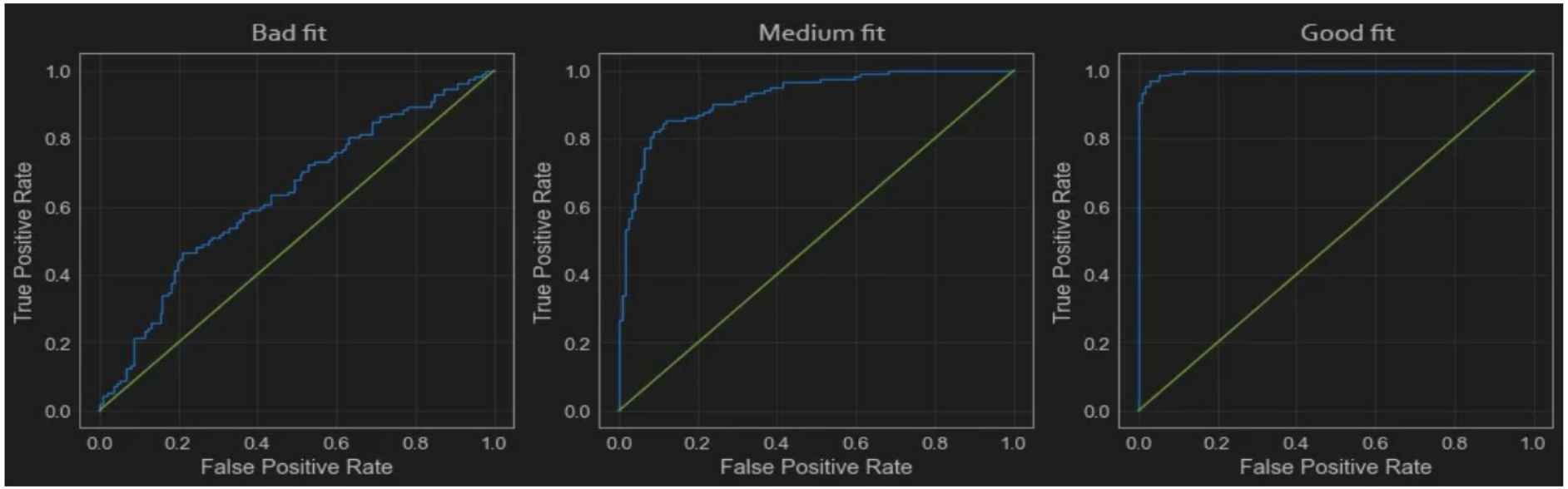
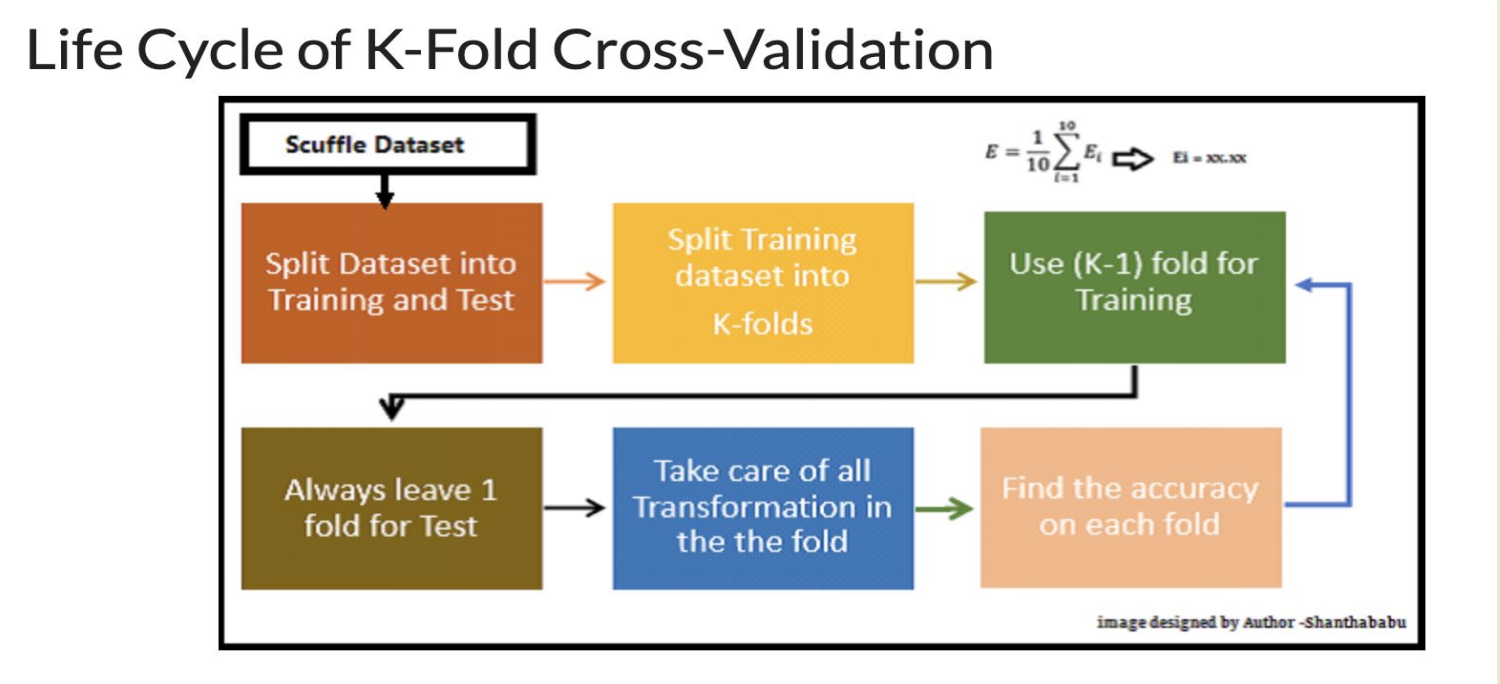


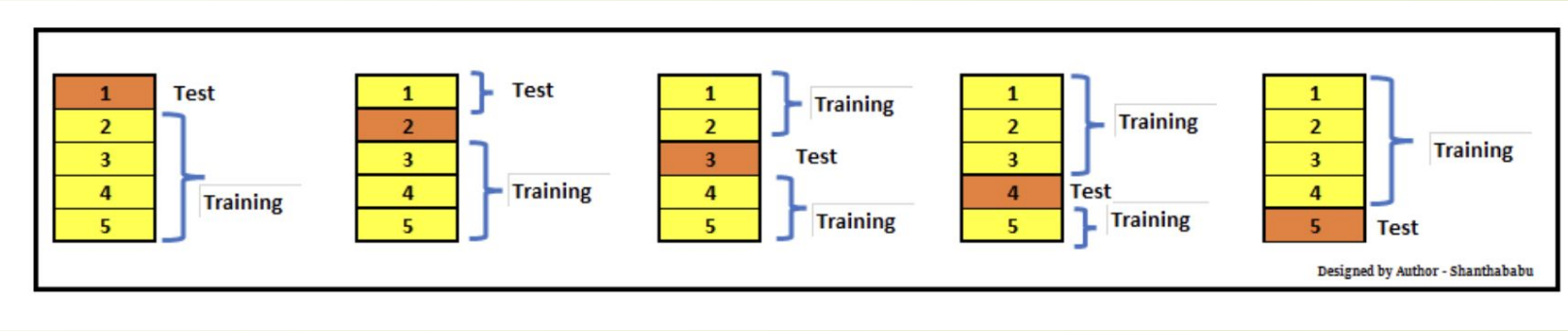
Image by author

K-Fold Cross Validation Technique

- K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time.
- This method aids in model assessment, selection, and hyperparameter tuning, providing a more reliable measure of a model's effectiveness.



K-Fold Cross Validation Technique – Logistic Regression



In the above set, 5- Testing 20 Training. In each iteration, we will get an accuracy score and have to sum them and find the mean. Here we can understand how the data is spread in a way of consistency and will make a conclusion whether to for the production with this model (or) NOT.

So we applied similar technique here in our code for logistic regression, with K= 10 and we got the accuracy as 88.24

EVALUATION OF Gaussian Naive Bayes Classifier

These accuracy scores represent the proportion of correctly predicted instances over the total number of instances in both the training and test datasets.

Evaluating on both datasets helps to understand how well the model generalizes to unseen data.

```
Train accuracy of naive bayes: 0.8782051282051282
Test accuracy of naive bayes: 0.8847795163584637
```

RANDOM FOREST MODEL

RF is one of the most popular ensemble learning method in machine learning not only gives good results even without hyperparameter optimization but also can use both classification and regression problems.

But the common problem of the traditional decision trees is over-fitting. In order to avoid overfitting, random forest models select and train hundreds of different sub-samples (multiple deep decision trees) randomly and reduce the variance.

```
Train accuracy of random forest 1.0
Test accuracy of random forest 0.9857752489331437
```

EVALUATION OF RANDOM FOREST MODEL

The classification report for the model predicting whether a song is explicit based on its audio features shows:

High precision and recall for both classes (False for non-explicit and True for explicit songs)

An overall accuracy of 1.00, which suggests a perfect prediction on the test set

This result is unusually perfect and may indicate overfitting issue, as It's rare to achieve perfect metrics in real-world scenarios, so further investigation would be prudent.

```
Train accuracy of random forest 1.0
Test accuracy of random forest 0.9857752489331437
```

Confusion matrix – RANDOM FOREST MODEL

A confusion matrix is a way to express how many of a classifier's predictions were correct, and when incorrect, where the classifier got confused (hence the name!)

In the confusion matrices below, the rows represent the true labels and the columns represent predicted labels.

This is a convenient way to spot areas where the model may need a little extra training.

Confusion matrix – RANDOM FOREST MODEL

In our dataset, the below are the results of confusion matrix and THE CLASSIFICATION REPORT.

```
Confusion matrix:
 [[ 75   5]
 [  5 618]]
Classification report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94        80
           1       0.99      0.99      0.99       623

    accuracy                           0.99       703
   macro avg       0.96      0.96      0.96       703
weighted avg       0.99      0.99      0.99       703
```

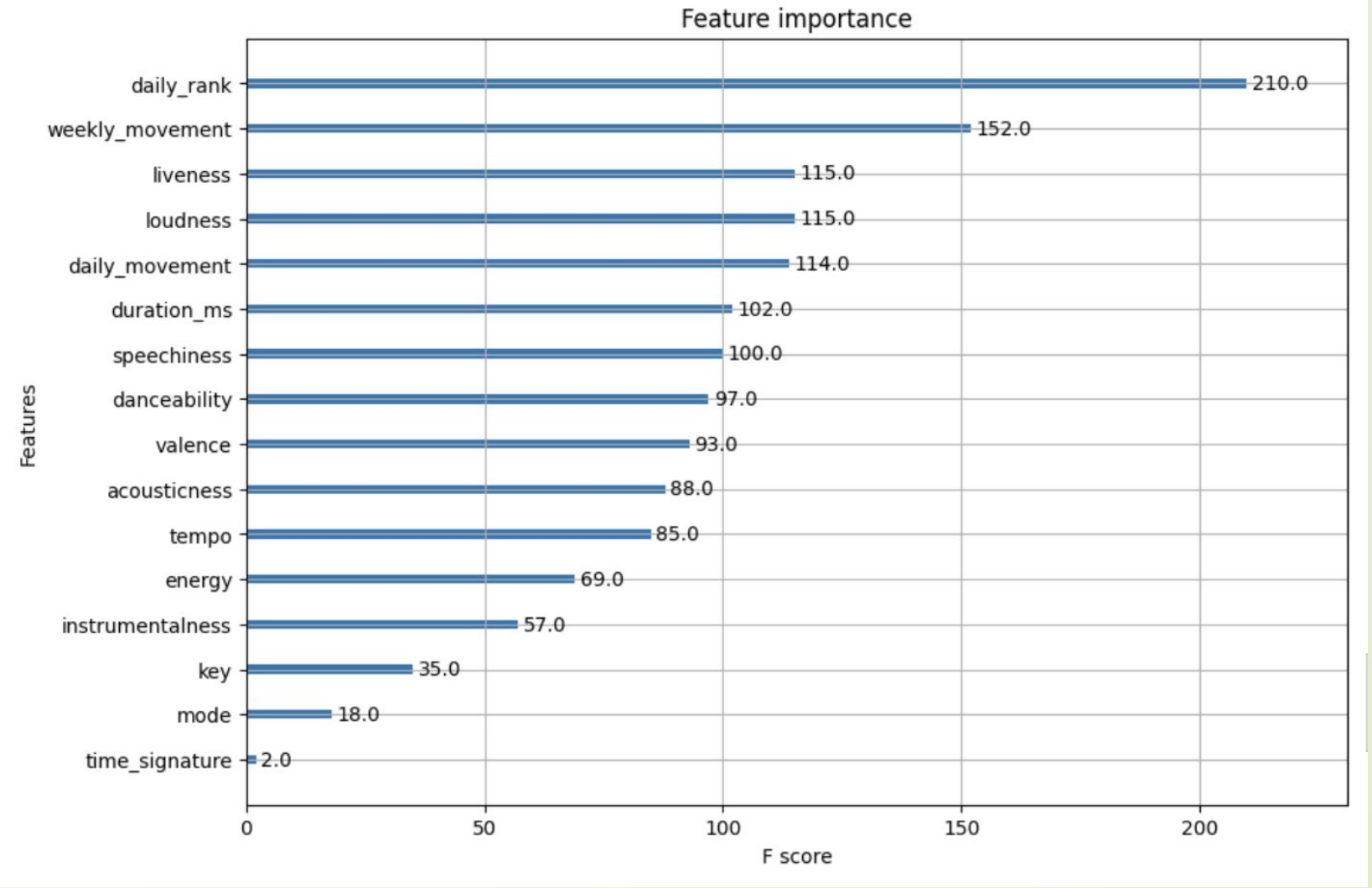
FEATURE IMPORTANCE –XG BOOST MODEL

Plotting Feature importances for an XG Boost Classifier helps in understanding which features are more influential in the XGBoost model's decision-making process.

Features with higher importance scores contribute more to the model's predictions. Adjustments or analysis can be made based on these importance scores to improve model performance or interpret the model's behavior.

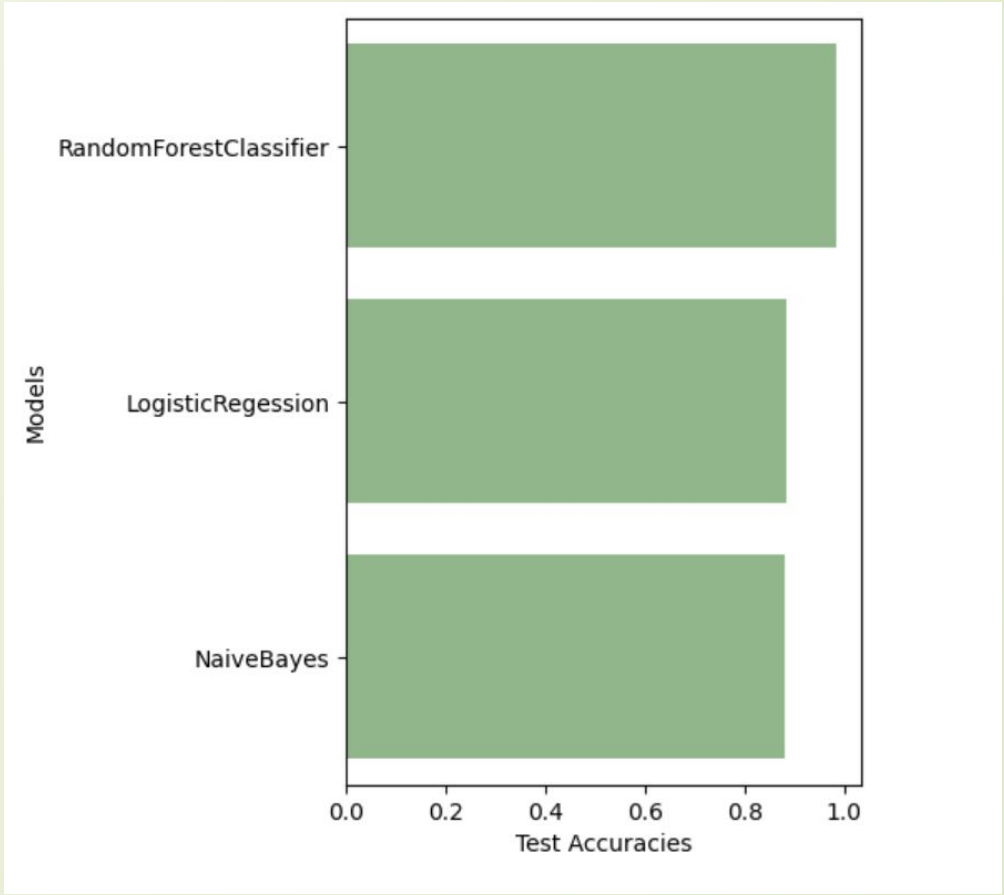
Then, it creates a bar plot using Matplotlib to visualize these importance scores for each feature.

FEATURE IMPORTANCE –XG BOOST MODEL



COMPARISON OF PERFORMANCE - CONCLUSION

	Model	Accuracy
0	RandomForestClassifier	0.985775
2	NaiveBayes	0.884780
1	LogisticRegession	0.882479



CLUSTERING TOP SPOTIFY SONGS IN 7 COUNTRIES

BACKGROUND

SPOTIFY

Spotify's platform revolutionized music listening forever

when it launched in 2008. Spotify's move into

podcasting brought innovation and a new generation of

listeners to the medium, and in 2022 Spotify entered the

next audio market primed for growth with the addition

of audiobooks.

BUSINESS OBJECTIVE

WHY IS IMPROVING USER EXPERIENCE AND ENGAGEMENT IMPORTANT?

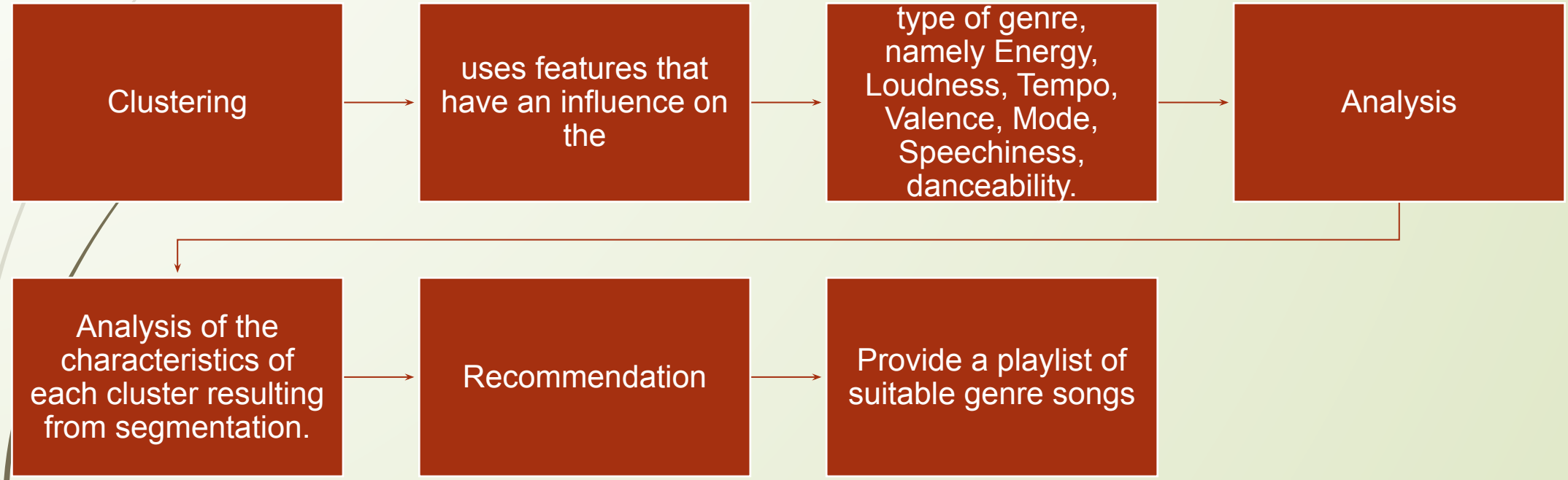
IT IS IMPORTANT TO INCREASE USER SATISFACTION AND RETENTION

WHAT EFFORTS SHOULD BE MADE?

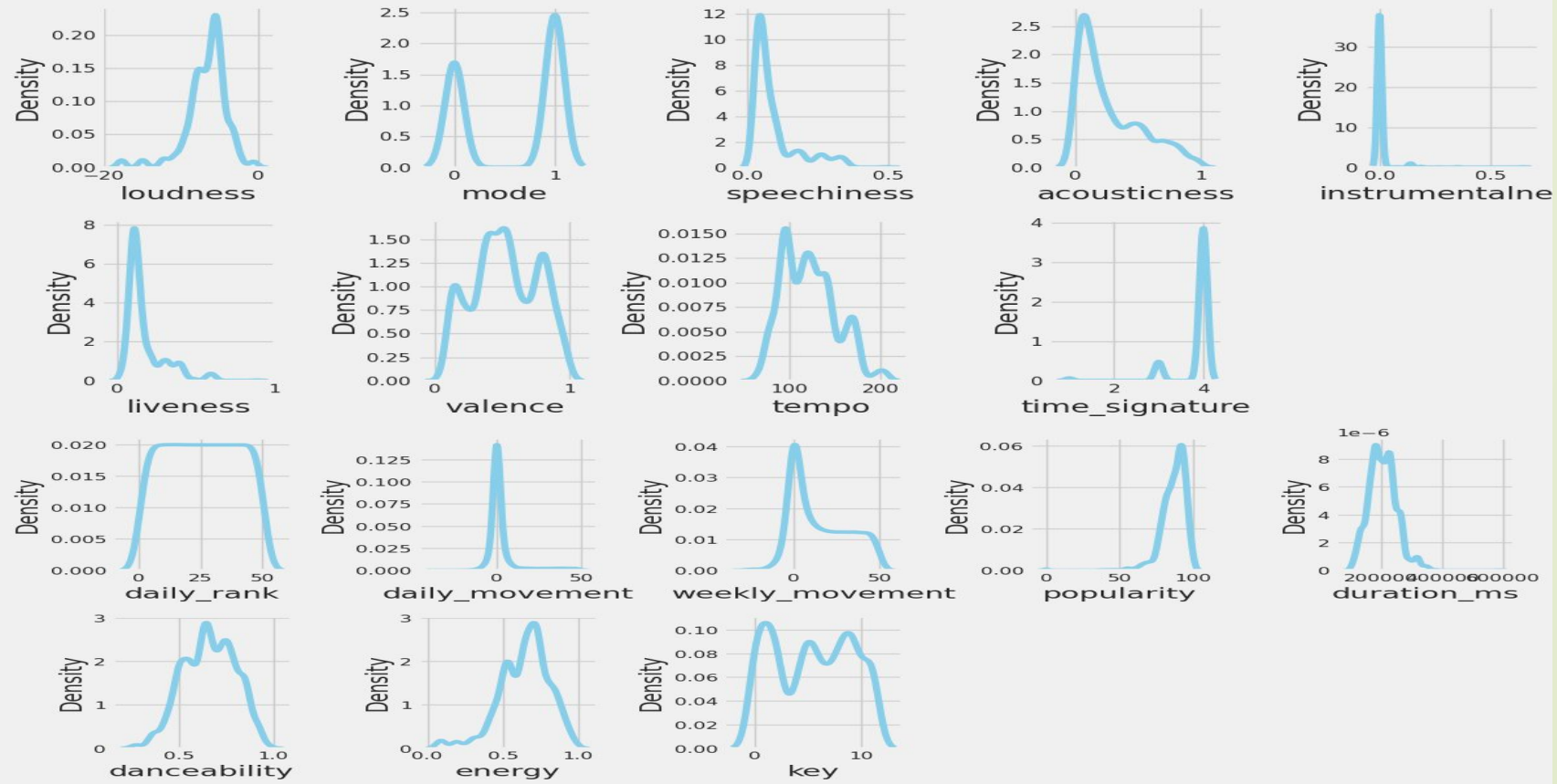
BY CLUSTERING SONGS BASED ON USER PREFERENCES AND

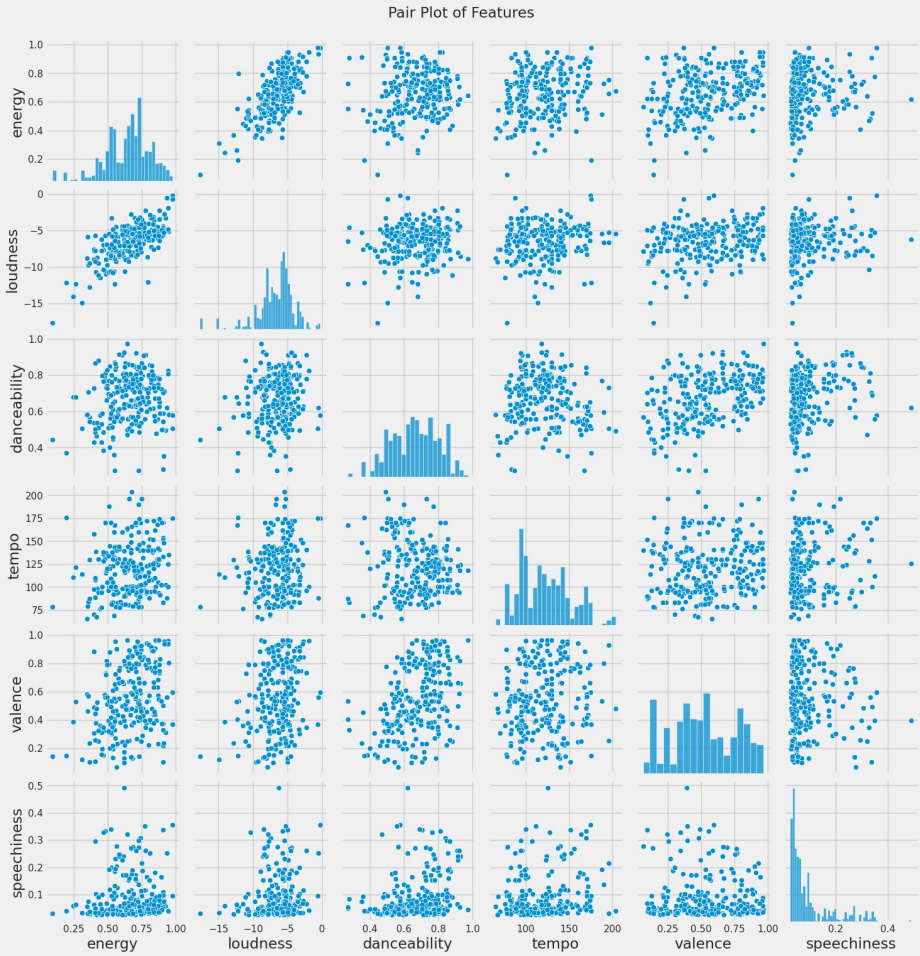
CHARACTERISTICS, PLATFORMS CAN OFFER CUSTOMIZED PLAYLISTS AND RECOMMENDATIONS.

Modeling


Clustering
uses features that have an influence on the
type of genre, namely Energy, Loudness, Tempo, Valence, Mode, Speechiness, danceability.
Analysis
Analysis of the characteristics of each cluster resulting from segmentation.
Recommendation
Provide a playlist of suitable genre songs

DATA DISTRIBUTION



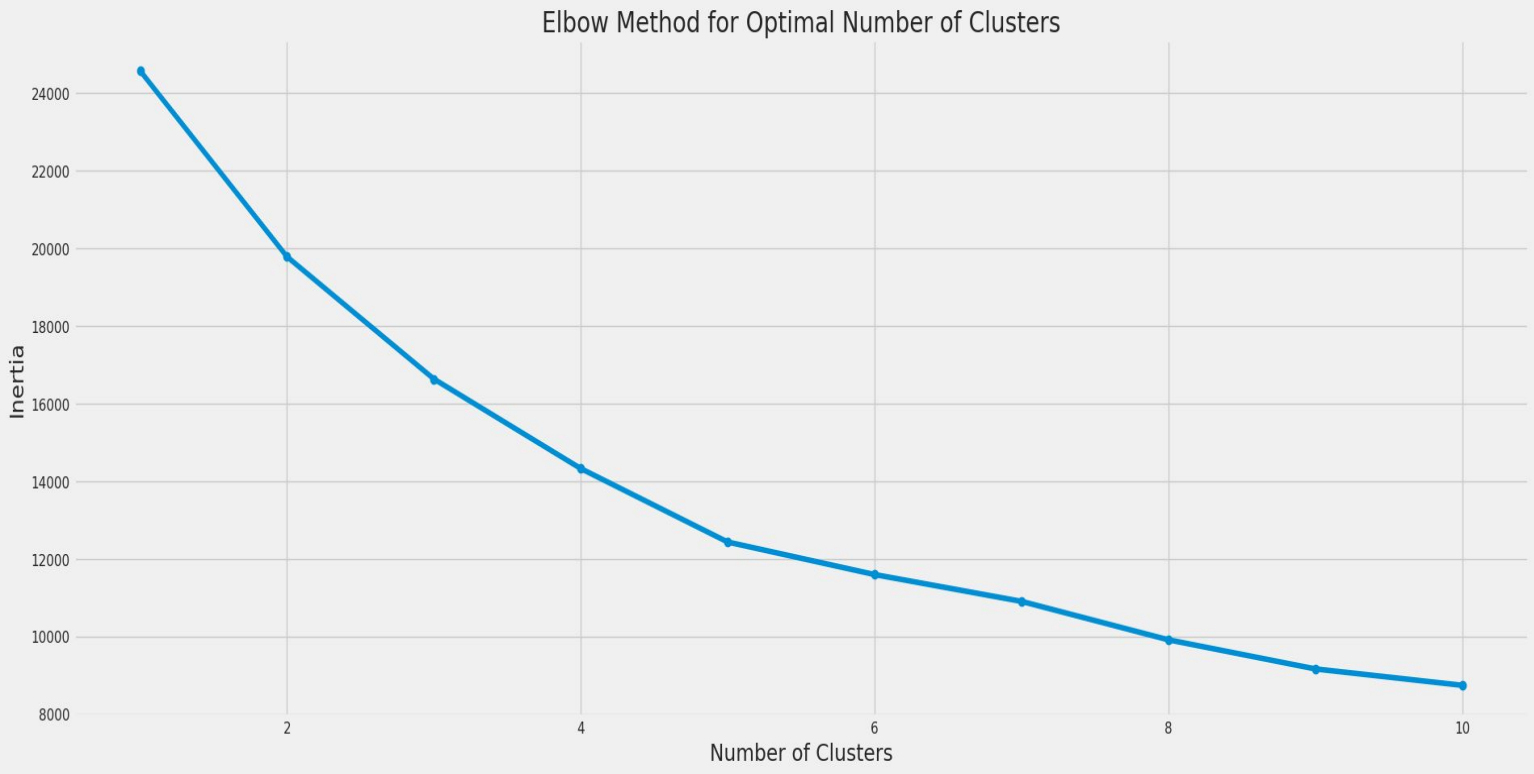


MULTIVARIATE ANALYSIS

Based on the graph, it can be seen that there is a relationship between some of the variables.

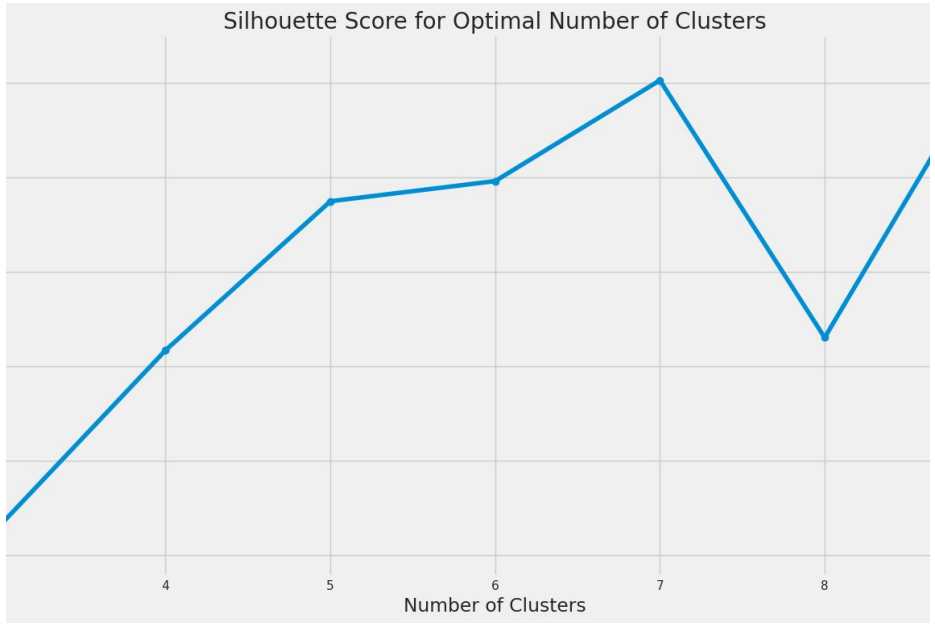
For example, energy and loudness are positively correlated, which means that songs with high energy tend to have high loudness as well. Tempo and valence are also positively correlated, which means that songs with a fast tempo tend to have high valence as well.

INERTIA



Before doing clustering, it would be better to determine the best and right number of clusters first. According to the graphic (Elbow method), the angle change starts to occur at point 4, then the correct K value for K-Means Clustering is K = 4

SILHOUETTE SCORE

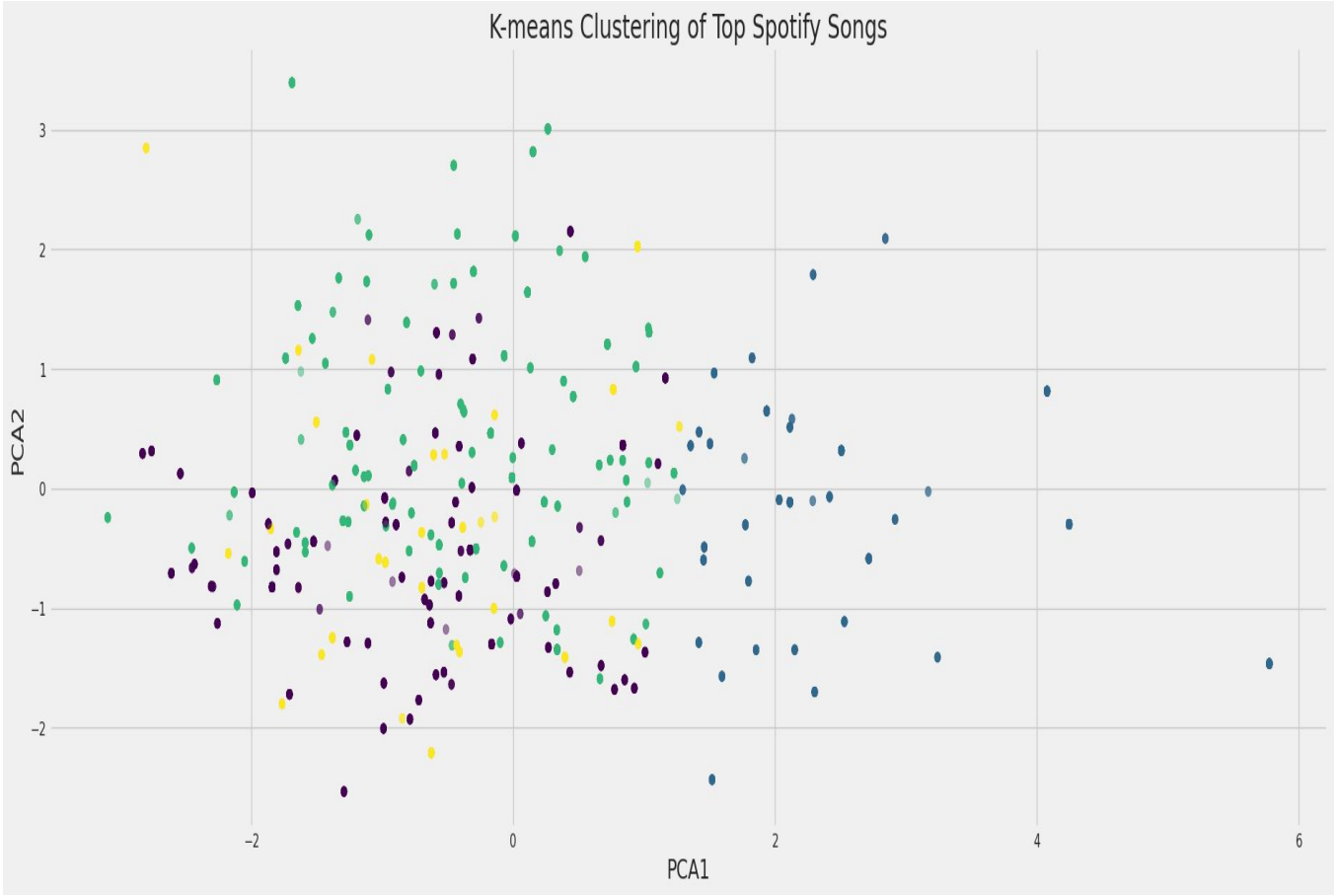


According to the graph (silhouette), the highest points occur at points 7 and 9, so I choose the right K value for K-Means Clustering is K = 4.

K MEANS CLUSTERING

From the results of this clustering and visualized with a scatterplot as shown below.

This diagram shows the distribution of data which is divided into clusters according to the K-Means Clustering algorithm.



MEDIAN PER CLUSTER

cluster	energy	loudness	danceability	tempo	valence	mode	speechiness
0	0.703	-5.6815	0.738	125.093	0.558	0.0	0.0584
1	0.417	-9.4750	0.501	109.094	0.200	1.0	0.0366
2	0.708	-5.6500	0.638	117.913	0.549	1.0	0.0527
3	0.640	-7.6830	0.802	123.061	0.494	1.0	0.2710

RECOMMENDATION

Provide a playlist of genre songs

Conclusion

Cluster 0 (Hard Rock) -> energy is quite high, loudness is quite high, tempo is very high, valence is very high, danceability is very high , speechiness is high and mode is a minor.

Cluster 1 (Ambient Music) -> very low energy, very low loudness, low tempo, very low valence, danceability is very low, speechiness is very low and mode is a major mode.

Cluster 2 (EDM) -> very high energy, very high loudness, quite low tempo, very high valence, danceability is quite low , speechiness is quite low and mode is a major mode.

Cluster 3 (POP) -> qutie low energy, quite low loudness, quite high tempo, quite low valence, danceability is very high , speechiness is very high and mode is a major mode.


Thank You!


Any Questions