Anvi Parikh

19 December 2023

Analysis: What Makes Music Popular!

Data pre-processing involved first checking for missing values, and none were found. Then, duplicates defined as entries with all of the same features except album name and song track ID. After identifying 4087 such duplicates, they were dropped from the dataset. There were 49221 remaining entries afterward.

Question 1

A histogram was printed for each of the 10 song features in order to visualize their distributions.

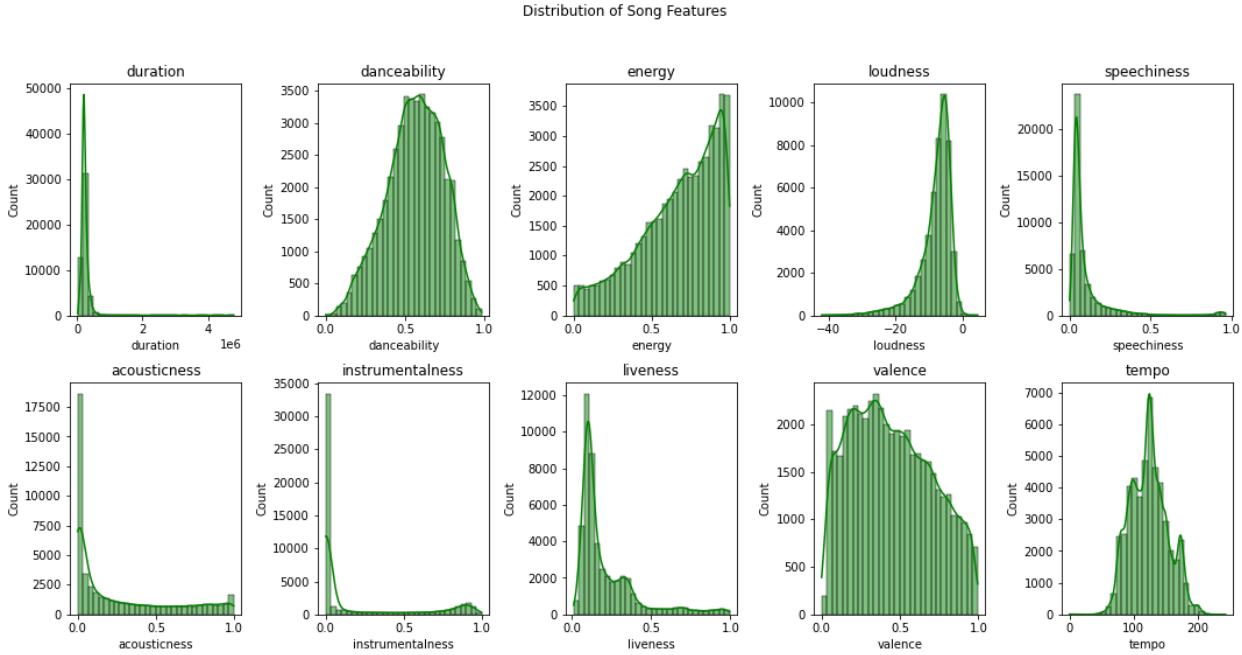

Figure 1. Distribution of Song Features

The features danceability and tempo seem to have reasonably normal distributions.

Question 2

A scatter plot with duration on the x axis and popularity on the y axis was created in order to visualize the relationship between song length and popularity of a song.

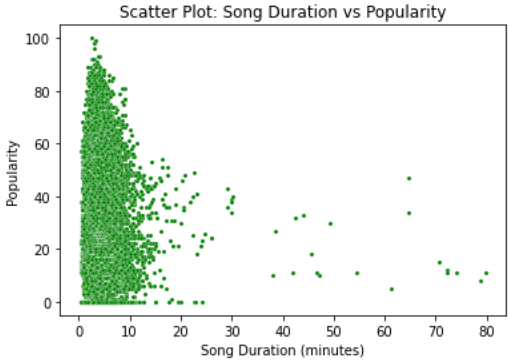

Figure 2. Scatter Plot: Song Duration vs Popularity

Visually, there does not seem to be a strong correlation between song length and popularity. To further assess any potential relationship, the Pearson correlation coefficient and the Spearman correlation coefficient were calculated, yielding ~ -0.08 and ~0.07 respectively. These coefficients indicate a weak negative relationship, whether linear or monotonic, between song length and popularity.

Question 3

When choosing an appropriate test for assessing the question of "Are explicitly rated songs more popular than songs that are not explicit?", the counts, means, and medians were printed of both explicit and non-explicit songs.

```
Number of Explicit Songs: 5334
Number of Non-Explicit Songs: 43887
Mean (Explicit):  37.526621672290965
Mean (Non-Explicit): 34.583999817713675
Median (Explicit):  36.0
Median (Non-Explicit): 35.0
```

Figure 3. Analyzing explicit vs non-explicit songs

The distribution of the data was also shown.

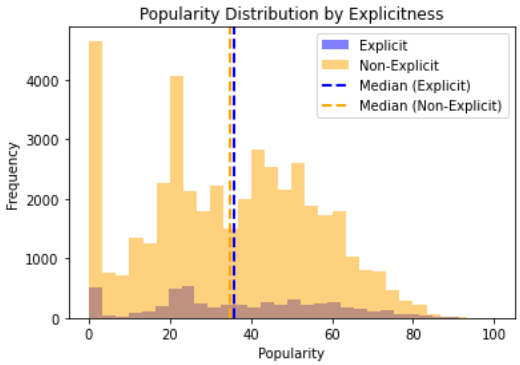

Figure 4. Popularity Distribution by Explicitness

Given that the data is not normally distributed, as well as the vast difference in number of explicit vs non-explicit songs, a parametric test was not the most viable option. However, since the explicit and non-explicit songs have similar distributions, a Mann Whitney U test was performed. The p-value resulting from this test was 4.983748064755883e-19, which is less than the alpha value of 0.05. Therefore, the assumption that no relationship exists between the explicitness of a song and song popularity was dropped. The different in popularity of explicit and non-explicit songs is statistically significant.

Question 4

When choosing an appropriate test for assessing the question of "Are songs in major key more popular than songs in minor key?", the counts, means, and medians were printed of both major and minor key songs.

```
Number of Major Key Songs: 30514
Number of Minor Key Songs: 18707
Mean (Major Key): 34.68460378842499
Median (Major Key): 35.0

Mean (Minor Key): 35.25894050355482
Median (Minor Key): 36.0
```

Figure 5. Analyzing songs in major key vs songs in minor key

The distribution of the data was also shown.

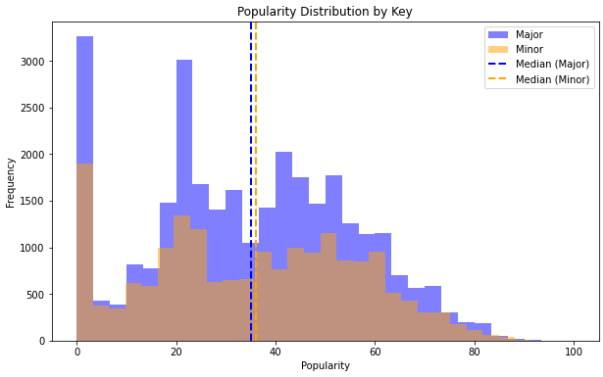

Figure 6. Popularity Distribution by Key

Given that the data is not normally distributed, as well as the difference in number of songs in minor key vs songs in major key, a parametric test was not the most viable option. However, since the major key and minor key songs have similar distributions, a Mann Whitney U test was performed. The p-value resulting from this test was 0.9978254724696891, which is greater than the alpha value of 0.05. Therefore, the assumption that no relationship exists between the mode of a song and song popularity was not dropped. The different in popularity of major key vs minor key songs is not statistically significant.

Question 5

A scatter plot with duration on the x axis and popularity on the y axis was created in order to visualize the relationship between song length and popularity of a song.

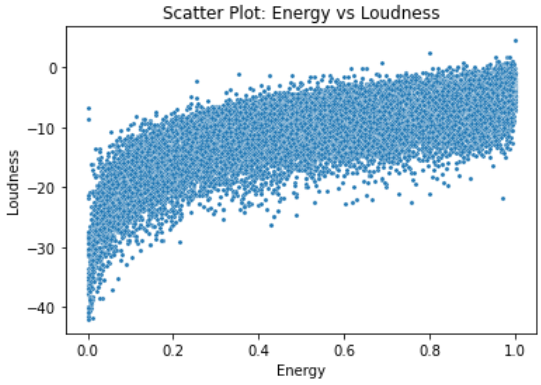

Figure 7. Scatter Plot: Energy vs Loudness

Visually, there seems to be a positive, possibly linear, correlation. To further assess any potential relationship, the Pearson correlation coefficient and the Spearman correlation coefficient were calculated, yielding ~ -0.78 and ~0.73 respectively. These coefficients indicate a somewhat strong positive relationship between energy and loudness that is more linear than monotonic in nature. Any assumption that energy does not largely reflect the loudness of a song can be dropped.

Question 6

A simple linear regression model for each of the 10 features was built to predict popularity. Since the data is skewed, the data was transformed by z-scoring beforehand. After completing this model on all 10 features and printing RMSE and R^2 values for all 10 features, the feature with the lowest RMSE as well as the highest R^2 was instrumentalness. With an RMSE of ~20.45 and an R^2 of ~0.03, instrumentalness predicted popularity best. However, when assessing the quality of this model, the RMSE value indicates that the model's predictions have significant error. The R^2 value also indicates that instrumentalness only explains ~3% of the variance in the model. Therefore, instrumentalness is not necessarily a good predictor of popularity. The scatter plot below further illustrates the lack of a strong relationship between instrumentalness and popularity.

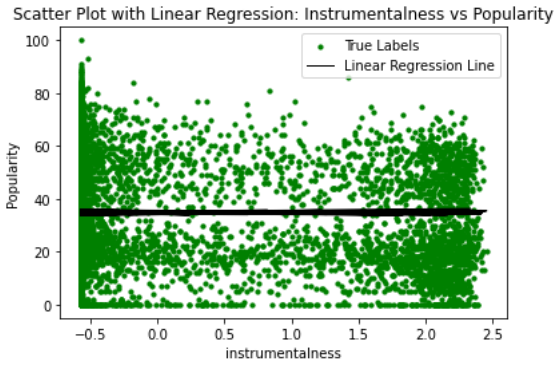

Figure 8. Linear Regression

Question 7

A multiple linear regression model was built to create a model that uses all of the song features in question 1. Since the data is skewed, the data was transformed by z-scoring beforehand. This model had an RMSE of ~20.06 and an R^2 value of ~0.07. This model was an improvement from the model in question 6, with an improvement of ~0.4 in RMSE and an improvement of ~0.04 in R^2. However, the improvements were not large, indicating that the 10 features do not predict popularity well.

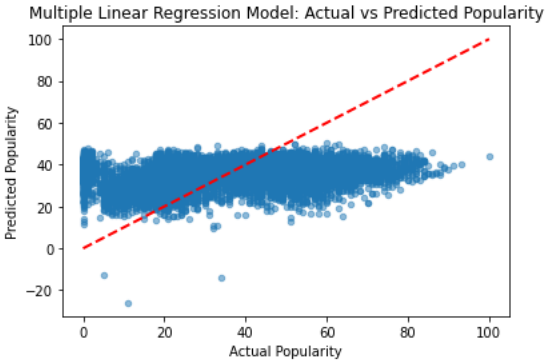

Figure 9. Actual vs Predicted Popularity

In this plot, the red line serves a reference line for a perfect prediction scenario in which the model's predictions are identical to the actual values. Given this reference, the low power of prediction held by this model is further apparent.

Question 8

To first visualize potential meaningful factors, a correlation matrix was created.

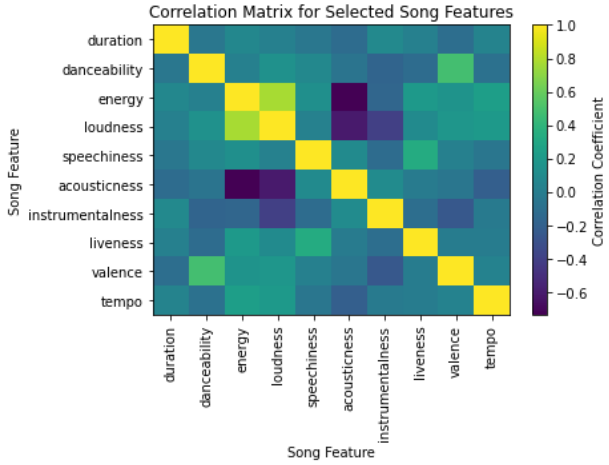

Figure 10. Correlation matrix

There are several noticeable clusters in the correlation matrix, highlighted by the groupings in dark purple (noticeable correlations between acousticness and energy, acousticness and loudness, and loudness and instrumentalness) as well as light green (noticeable correlations between loudness and energy).

The explained variance for every factor was printed.

```
27.13
16.192
13.948
9.836
8.772
8.199
6.738
4.721
3.15
1.312
```

Figure 11. Explained variance for 10 features

In order to discover the number of meaningful principal components, a scree plot was created and the Kaiser criterion was used to extract 3 components.

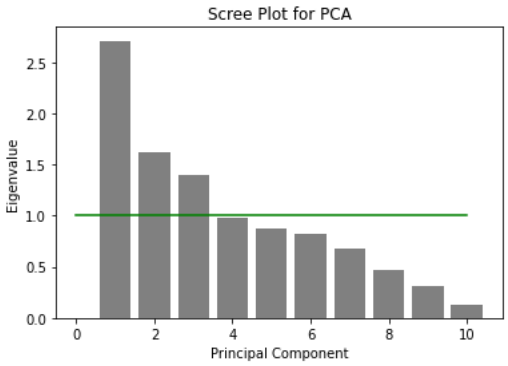

Figure 12. Scree plot

The three principal components account for 57.27% of the variance. These components were described as intensity, expressiveness, and quietness, respectively, as shown by these loadings plots which depict how much each feature contributes to a given component.

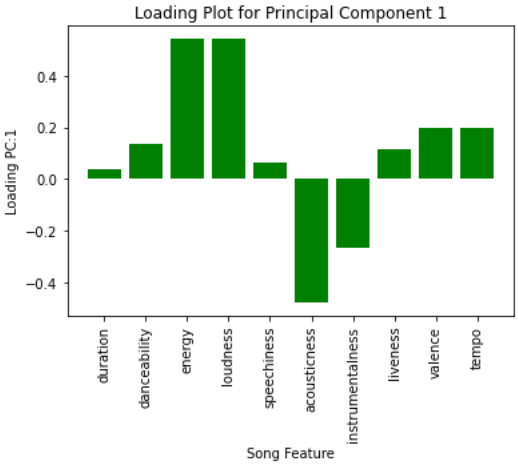

Figure 13. Component 1: Intensity

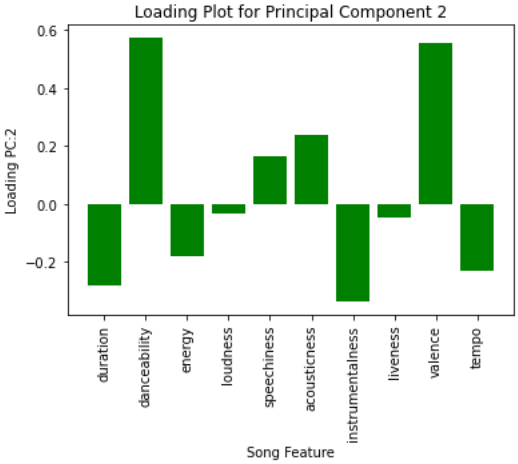

Figure 14. Component 2: Expressiveness

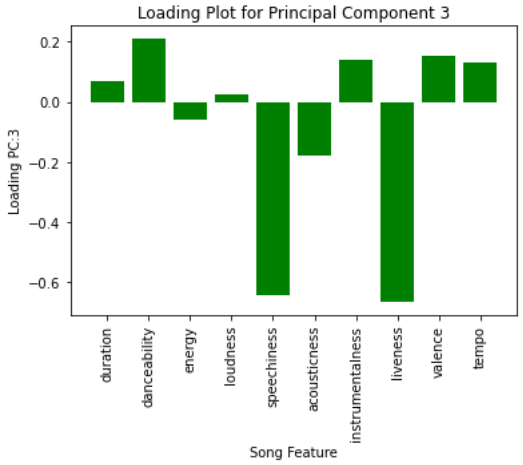

Figure 15. Component 3: Quietness

K-means clustering using the silhouette analysis method was then employed to identify 2 clusters.

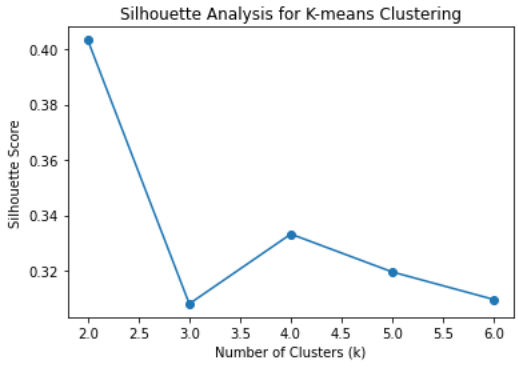

Figure 16. Silhouette Analysis to identify 2 clusters

Question 9

A logistic regression model was built to predict a song's key from valence. Since the data is skewed, the data was transformed by z-scoring beforehand. The accuracy of this model was 61.30%, and its AUC score was ~0.5011. A confusion matrix was also printed for this model.

Figure 17. Confusion Matrix

This matrix indicates that the model correctly predicted 6035 instances of major key songs, incorrectly predicted 3810 instances of major key songs when they were, in reality, minor key. The model also did not correctly predict minor key songs that truly are in minor key, nor minor key when they were truly major key. The model was also plotted as showcased below.

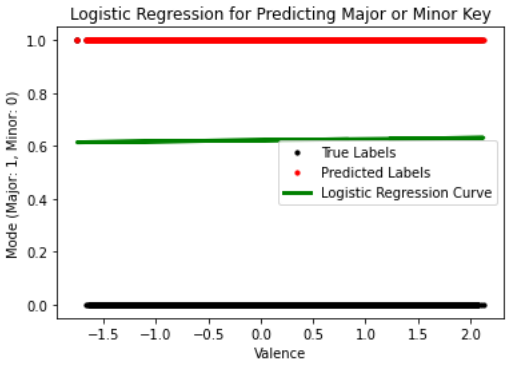

Figure 18. Logistic Regression for Predicting Major or Minor Key

Therefore, the model does not predict well. The question of a better predictor was then tested by creating a logistic regression model for the 10 features in question 1, yielding AUC scores that were higher than the original model for all 10 features. However, the AUC scores fell in the similar range of ~-0.5-0.6, showcasing that none of the features were necessarily better predictors.

Question 10

To predict the genre from the 10 song features in question 1, a random forest classifier was used with a cross-validation method involving splitting the dataset into five folds and training and evaluating the model five times, with each instance using a different fold as the test set and the remaining data as the training set. Since the data is skewed, the data was transformed by z-scoring beforehand. The average accuracy of this model among the five folds was ~0.32, not yielding a very high accuracy.

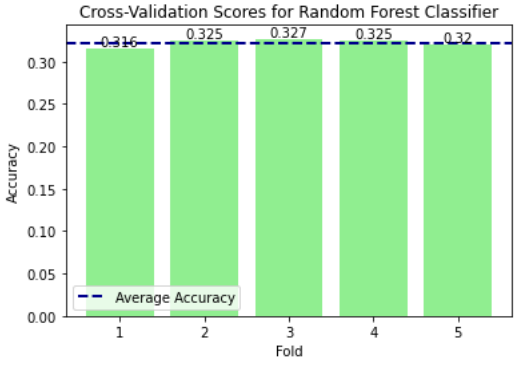

Figure 19. Accuracy scores for each fold

The feature importance plot indicates the importance of each feature when predicting the genre using the random forest classifier. Although the model as a whole struggles to reach a high accuracy, the feature importance plot provides a guide to understanding important features considered by the model.

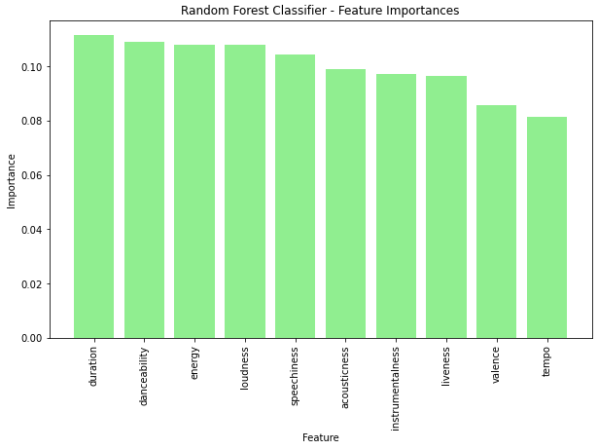

Figure 20. Feature importance