CAPSTONE PROJECT

Name : Vedant Desai

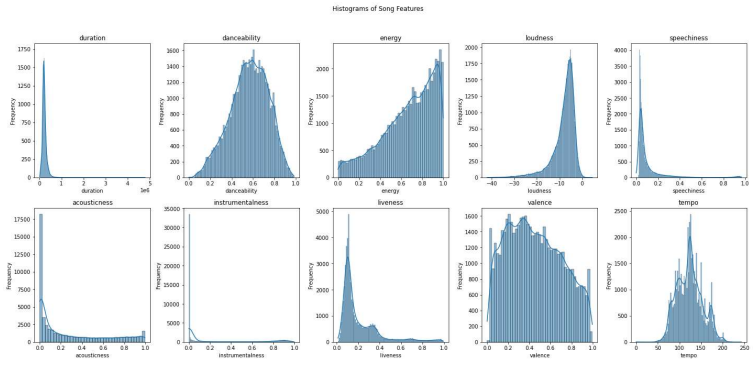
Preprocessing of data:

In the preprocessing step of the analysis, I did not encounter any missing values in the dataset, which simplified the preprocessing. I had to handle skewed data in answering some of the questions. In question 5, I used z-score normalization to account for the skewed data in loudness. For some of the latter questions I performed dimensionality reduction. I also cross validated my models wherever I deemed necessary to guard against overfitting. In the beginning of my code file, I seeded the random number generator with my N-number.

ANSWERS

Question 1 : Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Is any of these features reasonably distributed normally? If so, which one? [Suggestion: Include a 2x5 figure with histograms for each feature]

Answer 1: I plotted a histogram and based on visual interpretation of the histograms shown below. I concluded that two of the features that are danceability and tempo are normally distributed.

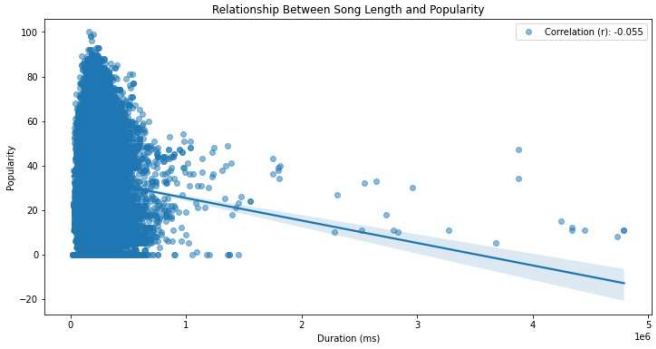


Histogram 1

Question 2 : Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative? [Suggestion: Include a scatterplot]

Answer: I plotted a scatterplot and used correlation coefficient to determine the relationship. Based on the scatterplot and the correlation coefficient (-0.05), there seems to be a very slight negative relationship between song length and popularity. However, this relationship is not strong enough to suggest that longer or shorter songs are consistently more popular.

R value: -0.05



Scatter Plot 1

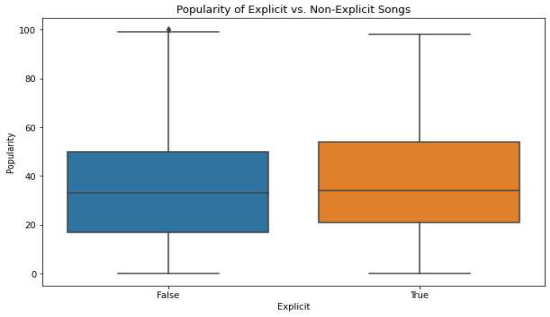
Question 3 : Are explicitly rated songs more popular than songs that are not explicit? [Suggestion: Do a suitable significance test, be it parametric, non-parametric or permutation]

Answer - Based on the Mann-Whitney U test and the boxplot, there is a statistically significant difference in the popularity of explicit and non-explicit songs. This result suggests that the explicitness of a song has a notable impact on its popularity, although the direction of this impact (whether explicit songs are popular) is better understood by examining the box plot which I have included below. I conclude that explicitly rated songs are more popular than songs that are not explicit Given the low p-value, we can confidently say that the difference in popularity between the two groups is not due to random chance.

Mann-Whitney U Test Results:

- **U-Statistic**: Approximately 139,361,273.5139,361,273.5
- **P-Value**: Approximately 3.07×10^{-19}

I chose the Mann-Whitney U test because it is suitable for comparing two independent samples. This test was chosen as it does not assume a normal distribution of data.



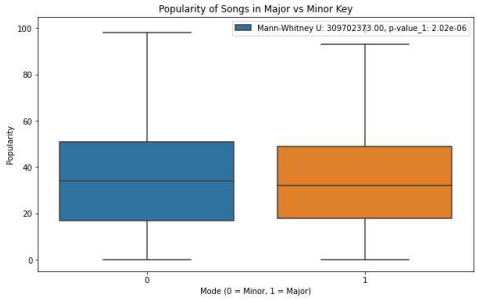
Box plot 1

Question 4 - Are songs in major key more popular than songs in minor key? [Suggestion: Do a suitable significance test, be it parametric, non-parametric or permutation]

Answer: I found based on the Mann-Whitney U test and the visual evidence from the boxplot, there is a statistically significant difference in popularity between songs in major and minor keys. This suggests that the key to a song (major or minor) may have a notable impact on its popularity. I conclude that songs in minor key are more popular than the songs in major key.

Mann-Whitney U Test Results:

- **U-Statistic**: Approximately 309702373
- **P-Value**: Approximately 2.01×10^{-19}

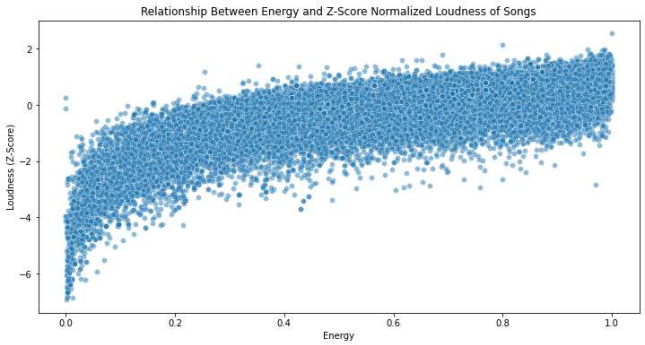


Box Plot 2

Question 5: Energy is believed to largely reflect the "loudness" of a song. Can you substantiate (or refute) that the is the case? [Suggestion: Include a scatterplot]

Answer : I plotted the data on scatterplot and accounted for skewed data in loudness by z-score normalization. The analysis I did substantiates the belief that energy largely reflects the loudness of a song, with a strong positive correlation between these two attributes.

The correlation coefficient I found between energy and z-score normalized loudness is approximately 0.775. The scatterplot illustrates this relationship.



Scatterplot 1

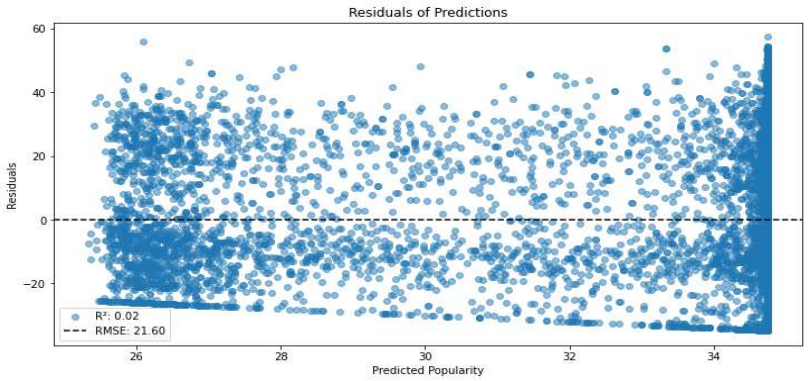
Question 6: Which of the 10 song features in question 1 predicts popularity best? How good is this model?

Answer : Based on the correlation and the magnitude of coefficients from the model's output, the most influential feature in predicting song popularity is instrumentalness

r of instrumentanlness : -0.144

R^2 (Coefficient of Determination): Approximately 0.02

RMSE (Root Mean Square Error): Approximately 21.6



Scatterplot 2

The R^2 value indicates that the model explains about 2% of the variance in the song popularity. This is relatively low, suggesting that these features collectively have limited predictive power for song popularity in this dataset. The RMSE value of 21.60 suggests that the average deviation of the predicted popularity from the actual popularity is about 21.60 points.

While the model includes a comprehensive set of features, its ability to predict song popularity accurately is quite limited. This outcome could imply that song popularity is influenced by factors not entirely captured by these features.

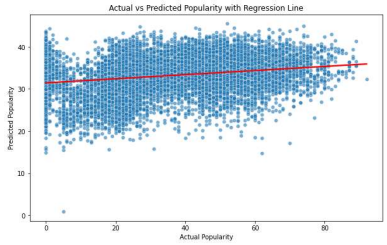
Question 7: Building a model that uses *all* of the song features in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question 6). How do you account for this?

Answer: This model accounts for approximately 5% of the variance, which is slightly better than the model in question 6. This feature accounts for 3% more variance than the previous model.

Multivariate Model R^2: Approximately 0.05

Multivariate Model RMSE: Approximately 21.6

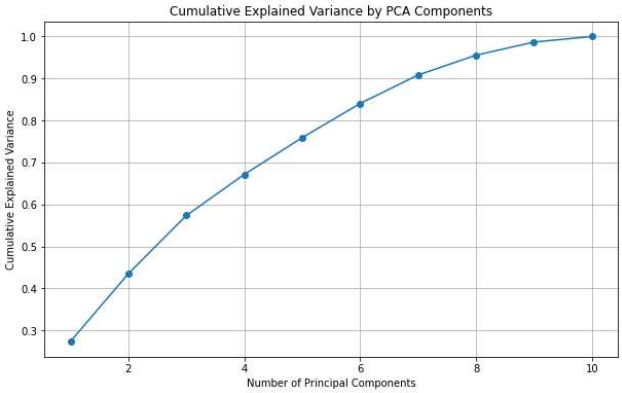
However, an increase from 2% to 5% is still modest, indicating that while the model has improved, there are likely other unaccounted factors that influence song popularity. The increase from the previous model to the new one in terms of the percentage of variance explained (R-squared) suggests that the additional features included in the new model might contribute to a better understanding of the factors influencing song popularity.



Scatterplot 3

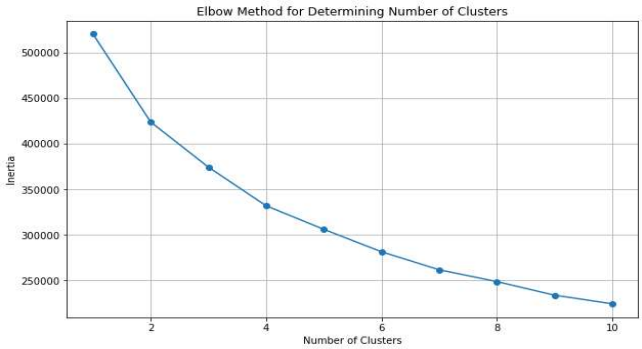
Question 8: When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using the principal components, how many clusters can you identify?

Answer : I have considered the 10 song features and identified 7 principal features. I have included the graph to illustrate it better. With 7 principle features we can account for about 90 percent of the variance. I have applied the elbow method to identify the number of clusters. I have identified 2 clusters and included the graphs below.



Line Graph 1

The elbow method is used to identify the number of clusters. From the graph below we can see that at 2 clusters are where we see an elbow like break.



Line Graph 2

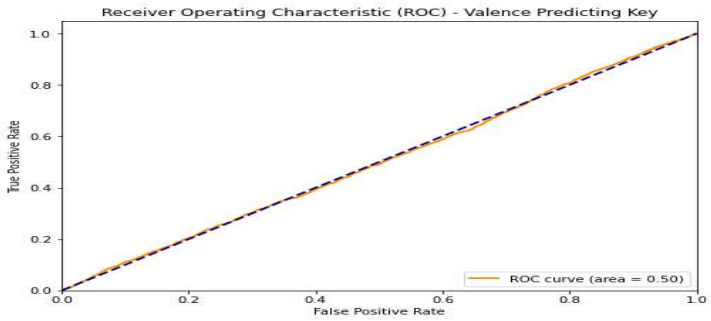
Question 9: Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor? [Suggestion: It might be nice to show the logistic regression once you are done building the mode.]

Answer: No, you cannot predict a song is in major or minor key from valence. I have built a logistic regression model to answer this question, and these were the results:

Accuracy: Approximately 61.94%

ROC AUC: Approximately 0.50

I have also included a visualization for better understanding.



Line Graph 3

The AUC (Area Under the Curve) value of about 0.50 suggests that the model's ability to distinguish between major and minor key songs is no better than random guessing. Ideally, a well-performing model would have an ROC curve that bows towards the top left corner of the plot, with an AUC closer to 1.0. The model's accuracy of about 61.94% might seem moderate, but the ROC AUC value indicates that this might be misleading, primarily due to the imbalanced nature of the classes (major and minor keys).

In all, I conclude that the model does not effectively use valence to predict the key to a song.

Better Predictor: To find a better predictor, other features can be examined, or combinations of features could be used. Feature selection techniques or exploratory data analysis might reveal more informative predictors.

Question 10: Can you predict the genre, either from the 10 song features from question 1 directly or the principal components you extracted in question 8? [Suggestion: Use a classification tree, but you might have to map the qualitative genre labels to numerical labels first]

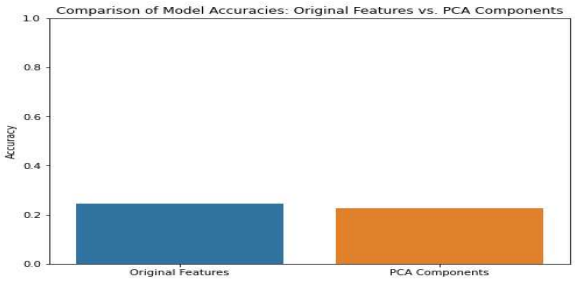
Answer: The decision tree classifier was applied to predict the genre of a song using two different approaches: directly using the 10 song features and using the principal components derived from these features.

Using Original 10 Song Features:

Accuracy: Approximately 24.36%

Using PCA Components:

Accuracy: Approximately 22.66%



Bar graph 1

Comparison shown above shows that accuracy of the original features is better than the PCA components. Both models show relatively low accuracy, indicating the challenge of predicting song genre based on these features alone. While using the original features directly offers a slight improvement over PCA components, the overall effectiveness of both approaches is limited for the task of genre classification.

Extra Credit:

I t analyzed the data based on time signature and found that the songs with time signature 4 have the highest averages in popularity, danceability, energy, and valence. This suggests that songs with a 4-beat measure are generally more popular, danceable, energetic, and have a more positive mood.

Index	ne_signatu	popularity	danceability	energy	valence
0	0	35.8889	0.0856667	0.0859856	0.117778
1	1	29.8322	0.435418	0.531248	0.353889
2	3	30.4144	0.438026	0.497322	0.339087
3	4	33.3992	0.576745	0.685882	0.462879
4	5	31.2497	0.490864	0.538301	0.401323

Figure 1