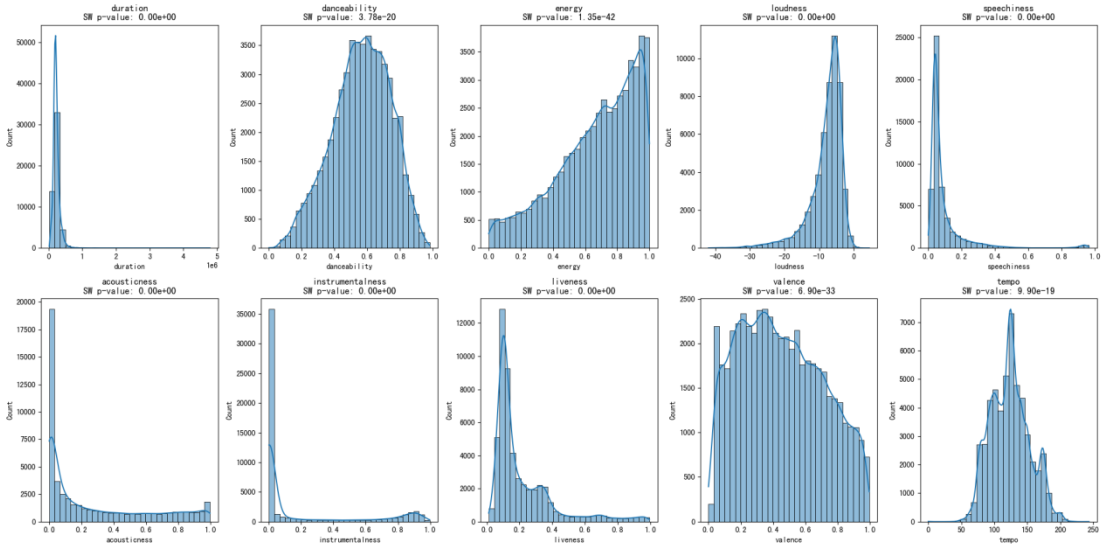
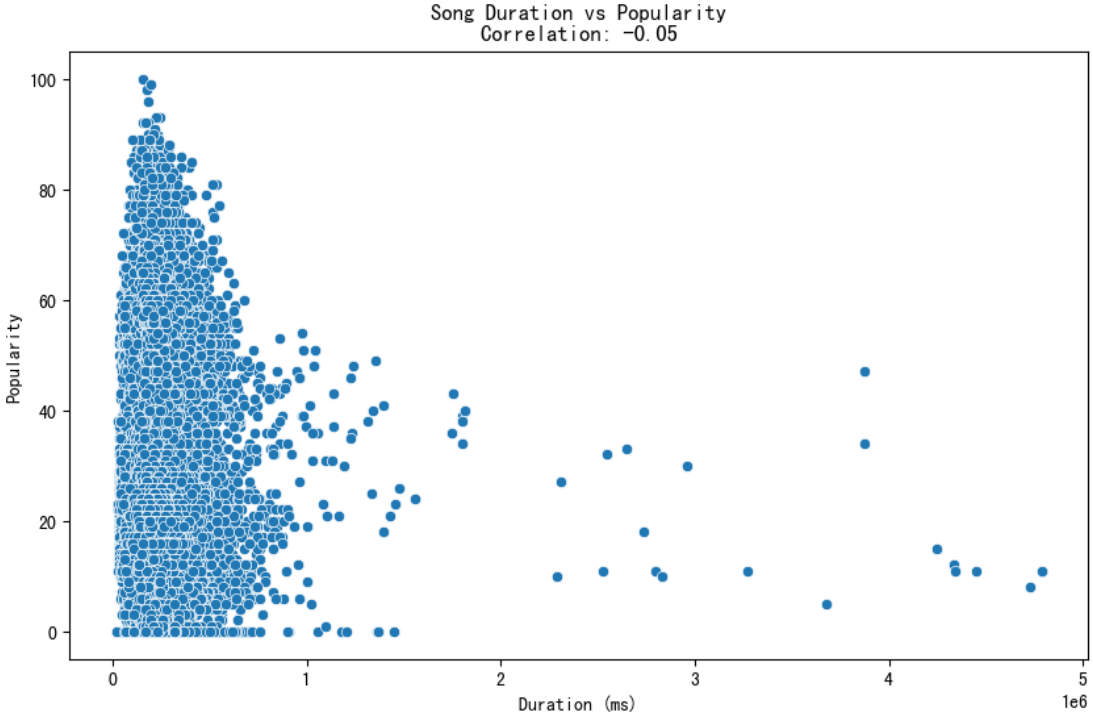
1.



The histogram above shows the distribution of these ten features. Intuitively, the distribution of the feature "danceability" and the feature "rhythm" may be approximately a normal distribution. But I performed a Shapiro-Wilk test on each feature to statistically determine normality. Due to the large size of the data set, the test was performed on a random sample of 5000 observations for each characteristic. All characteristics were not normally distributed according to the Shapiro-Wilk test (critical p value to determine normality was 0.05). This shows that all analyzed features deviate significantly from the normal distribution.

2.



I have calculated the correlation coefficient song length and the popularity. The correlation coefficient is approximately -0.055. This indicates a very weak negative relationship between

song duration and popularity. The negative value suggests that as the duration increases, popularity slightly decreases. However, the strength of this relationship is very weak.

3.

A Mann-Whitney U test, a non-parametric test, was performed to compare the popularity of explicitly rated songs and songs that are not explicit. This test is chosen as it does not assume normal distribution of the data.

Null Hypothesis: There is no difference in popularity between explicit and non-explicit songs.

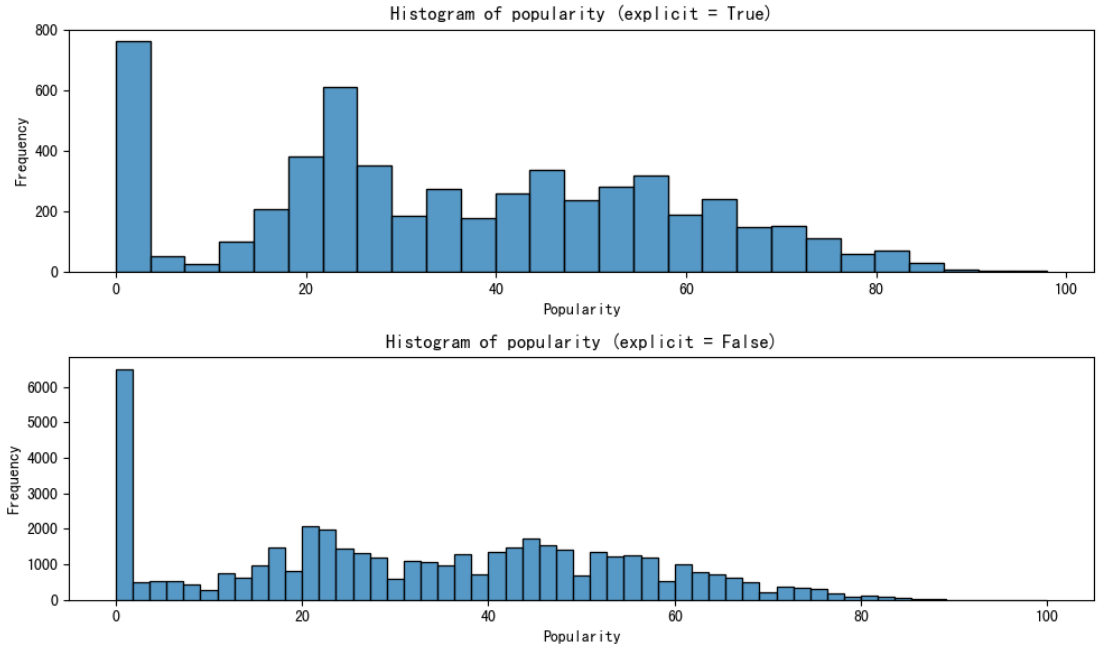
Alternative Hypothesis: There is a difference in popularity between explicit and non-explicit songs.

Significance Level: 0.05

Test Statistic: 139361273.5

p-Value: Approximately 3.07e-19

The p-value is significantly less than the significance level of 0.05. This result allows us to reject the null hypothesis. Therefore, there is a statistically significant difference in popularity between explicitly rated songs and songs that are not explicit.



4.

I still use the Mann-Whitney U test. This test is appropriate as it does not require the assumption of normal distribution.

Null Hypothesis: There is no difference in popularity between songs in major key and songs in minor key.

Alternative Hypothesis: There is a difference in popularity between songs in major key and songs in minor key.

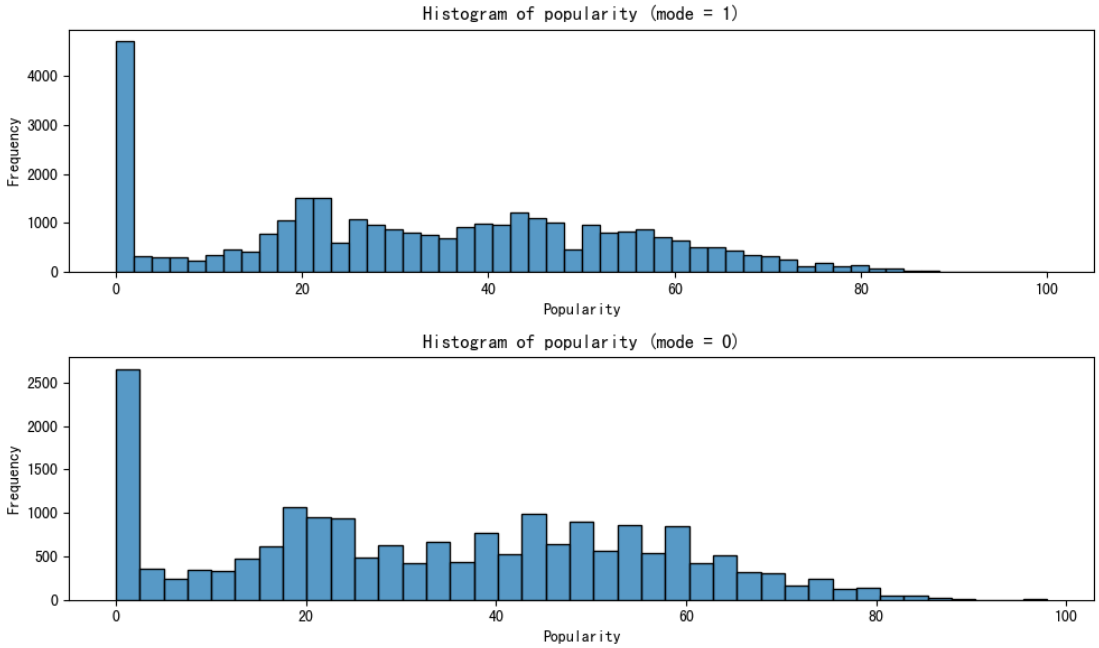
Significance Level: 0.05

Test Statistic: 309702373.0

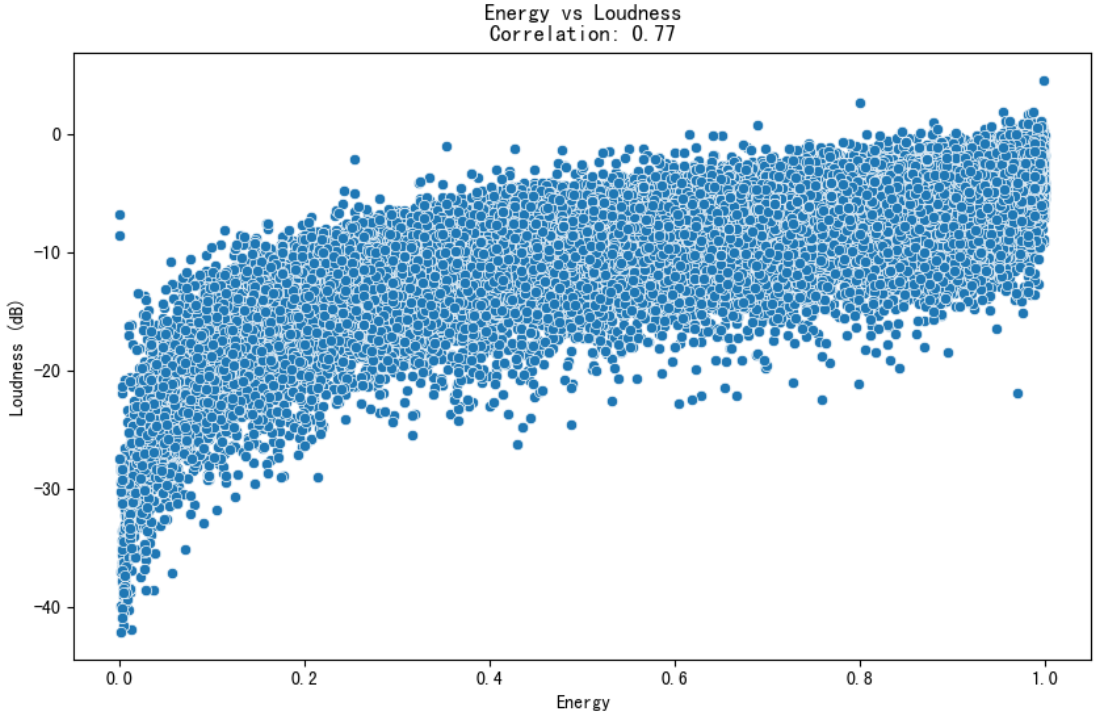
p-Value: Approximately 2.02e-06

The p-value is significantly less than the significance level of 0.05. This result allows us to

reject the null hypothesis. Therefore, there is a statistically significant difference in popularity between songs in major key and songs in minor key.



5.



I have made a scatter plot between energy and loudness, calculating the correlation coefficient. The scatterplot shows a general trend where higher energy is associated with higher loudness. The correlation coefficient is approximately 0.77. This indicates a strong positive relationship between energy and loudness. The positive value suggests that as the energy of a song

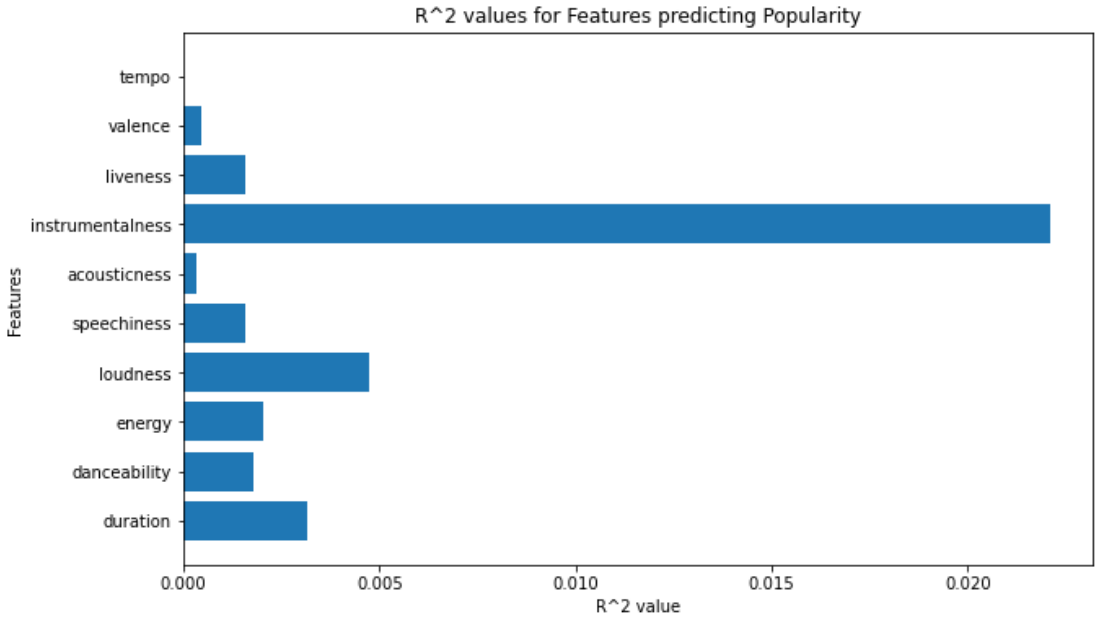
increases, its loudness also tends to increase.

6.

A linear regression model was built to predict the popularity of songs based on the 10 features respectively. The dataset was split into training and testing sets to evaluate the model. The R^2 score was calculated to assess the performance of each model.

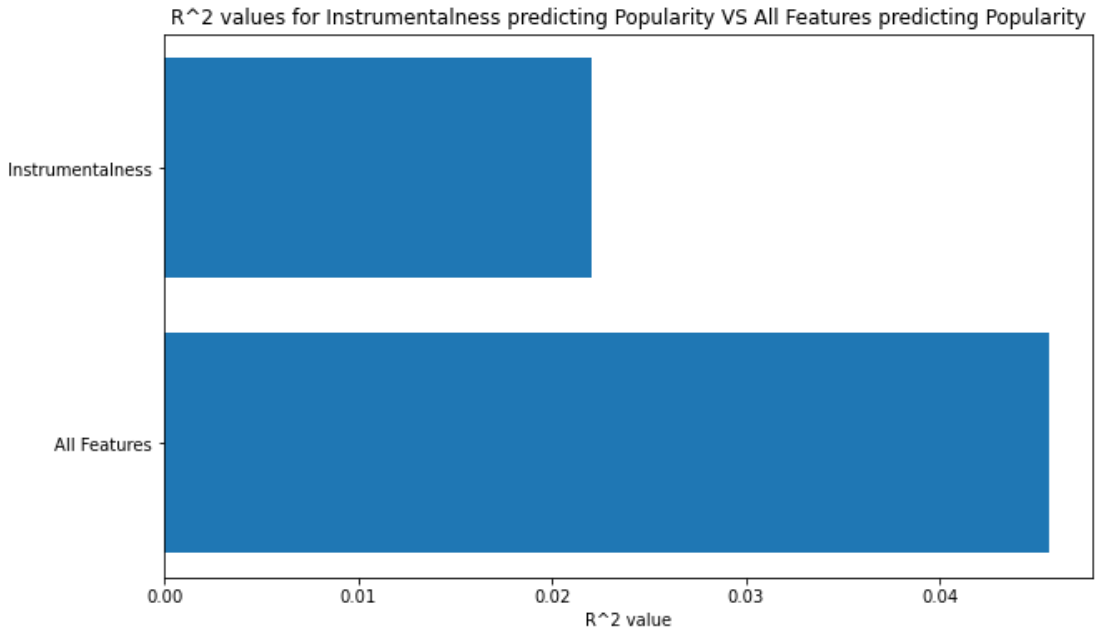
Best Predictor: 'instrumental' emerged as the feature with the highest R^2 value, suggesting it is the best predictor of popularity among the 10 features.

Model Performance: The R^2 score of the model is approximately 0.0221. This means that about 2.21% of the variability in the popularity can be explained by the model. Though it is the highest R^2 value among all 10 features, the value is still quite low, which indicates the model does not have a quite high predictive power.

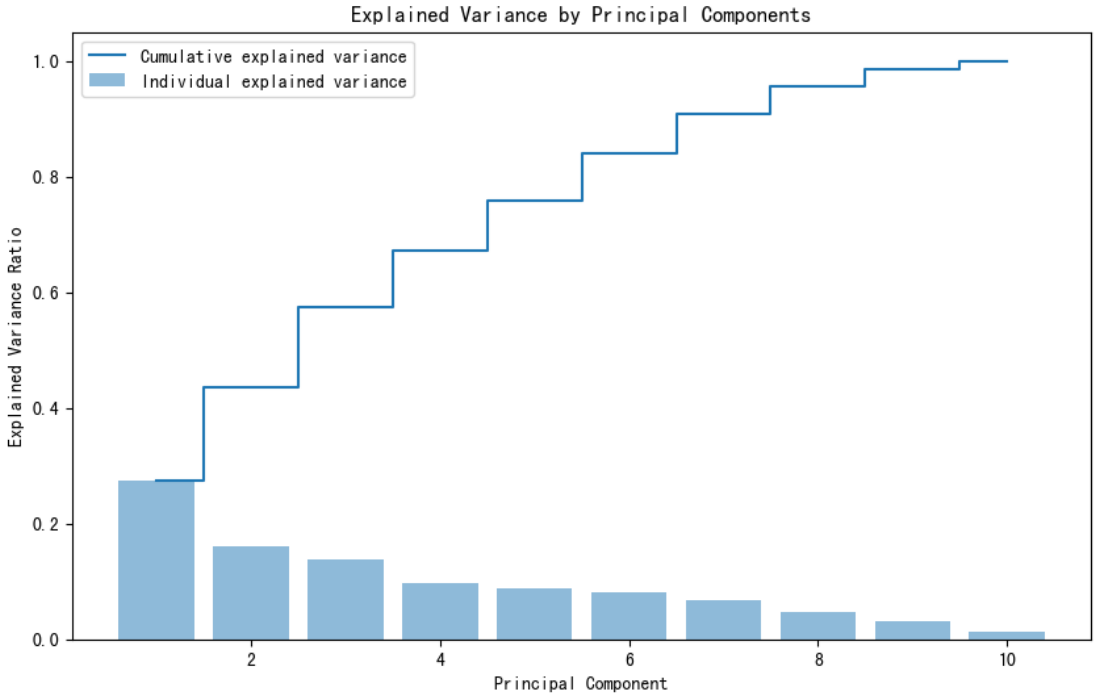


7.

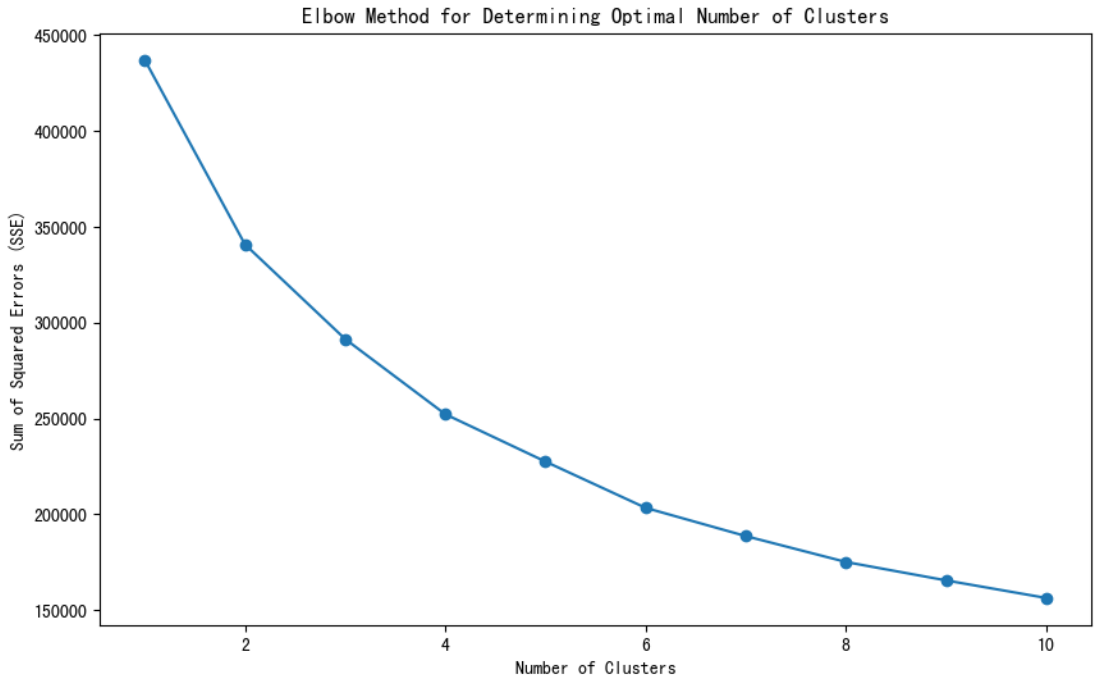
A linear regression model was built, which includes all the features as the predictors. The dataset was split into training and testing sets to evaluate the model. The R^2 score was calculated to assess the performance of the model. The model has a low R^2 score of about 0.046, suggesting that even combining all the features the model still has a limited predictive power for song popularity. But the predictive power of this model is better than the former one, which is only based on the instrumentalness alone.



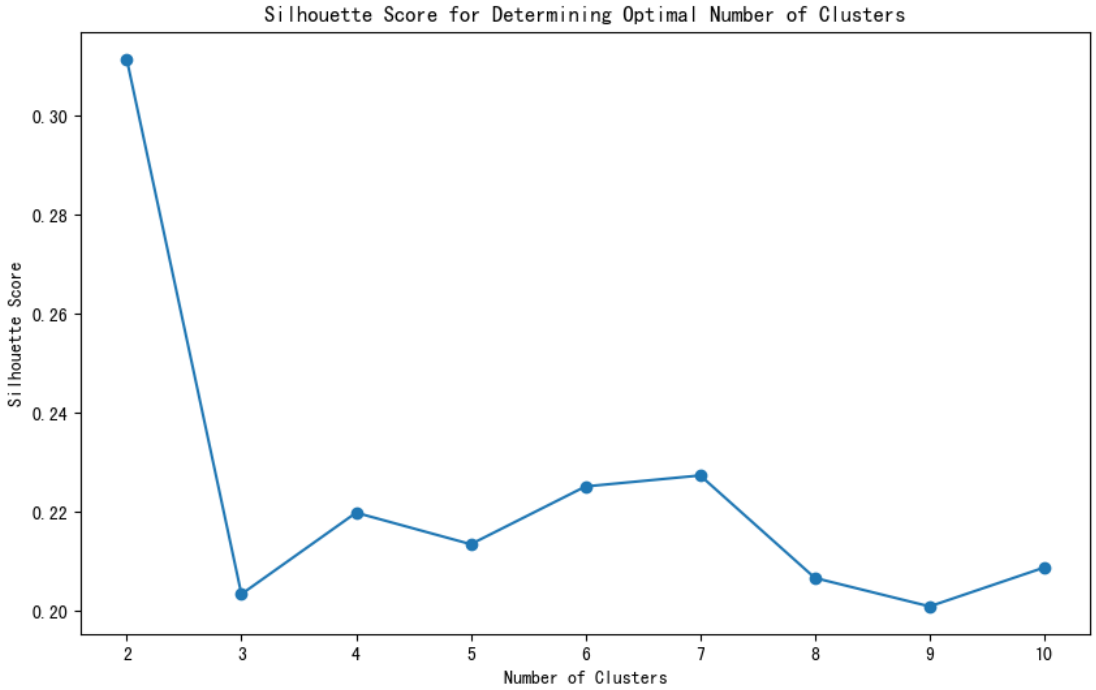
8.



I have done the PCA for this one. According to the principal component analysis results, when 6 principal components are selected, the cumulative contribution of the variance can reach about 84%, so I chose 6 meaningful principal component digits.



From the figure, it is hard to eyeball what is the optimal number of clusters. So I made a silhouette score figure to determine this number.



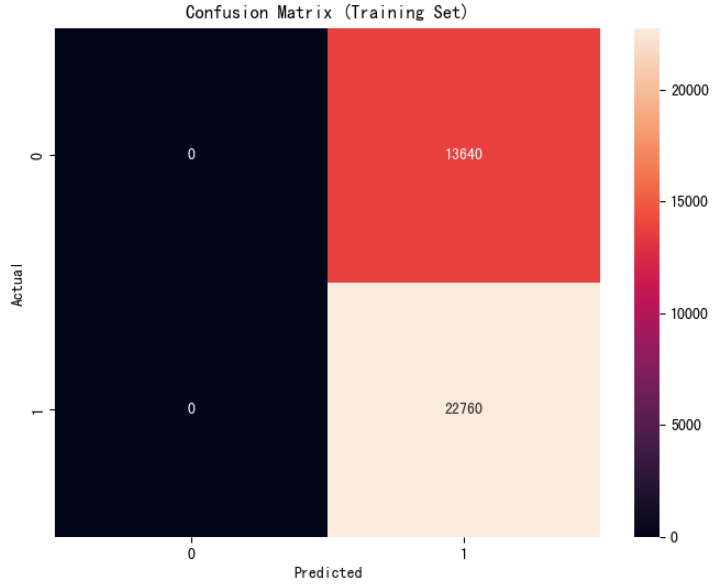
From the figure, we can see that the Silhouette Score is highest when 2 clusters are selected. Therefore, we can choose the 2 cluster in the process of k-means clustering.

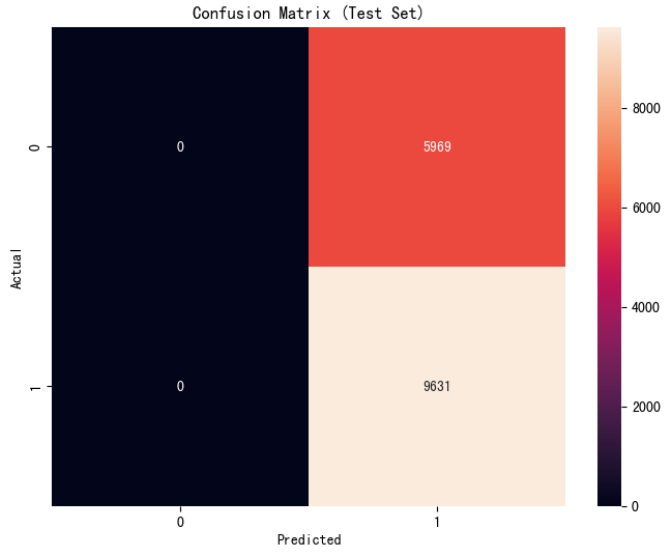
9.

Use "valence" as the predictor variable and "mode" (major or minor) as the target variable. Split the data into training and test sets. Train a logistic regression model on the training data

and get the results of the model on the test set. I also select other predictor variables to perform logistic regression on the target variable.

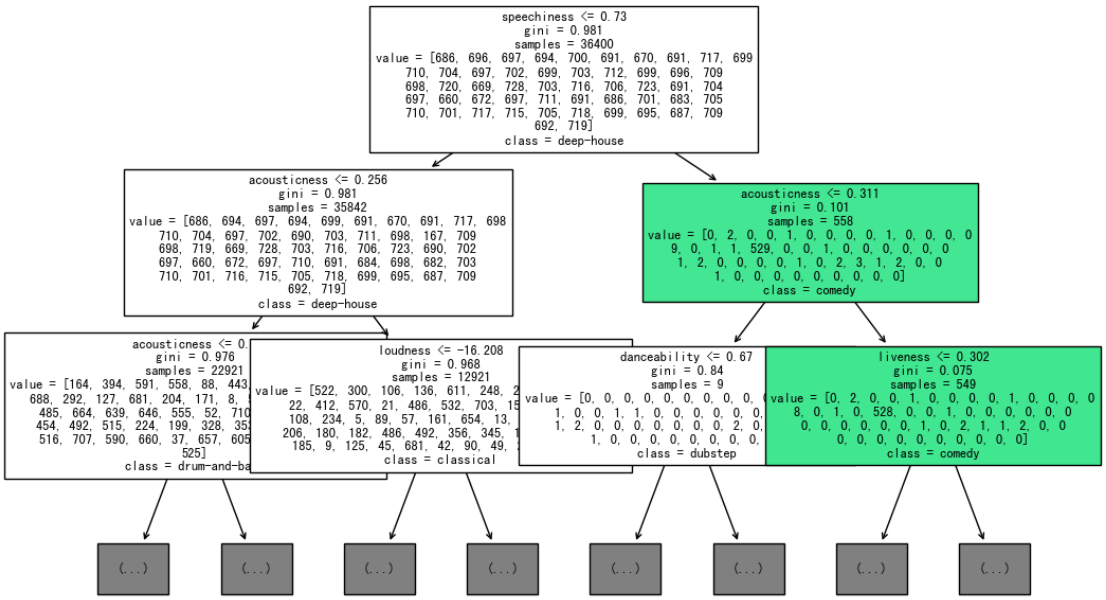
The results show that among the variables "valence" and the variables I selected, the accuracy of the logistic regression model based on the variable "valence" is the highest, with an accuracy of approximately 62%.





In addition, I used AUC as an evaluation index to select features for logistic regression. The final feature acousticness had the highest AUC value, which was 0.566.

10.



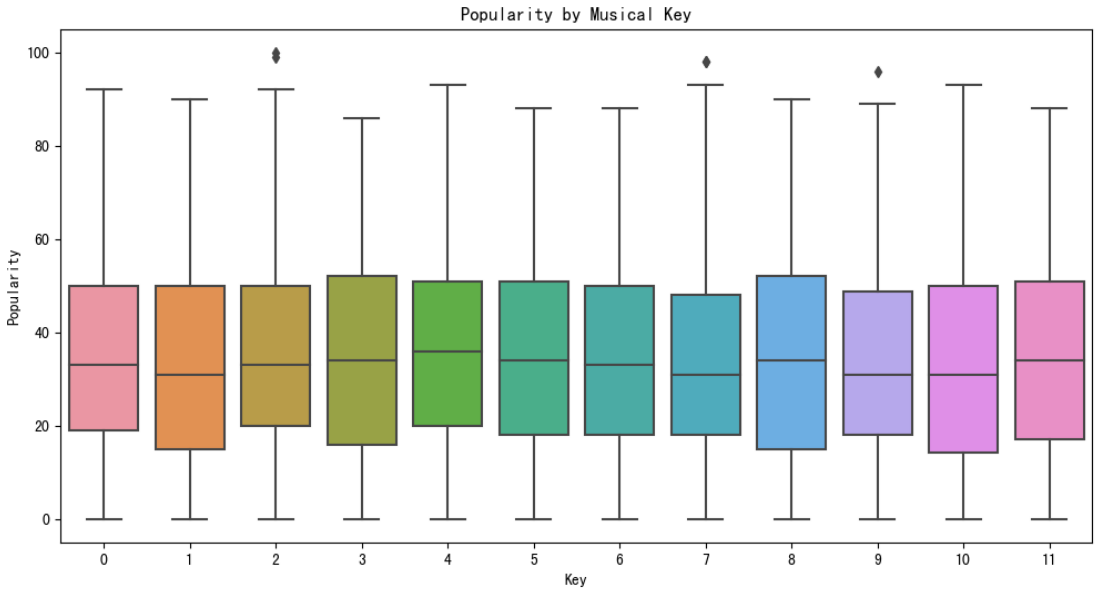
Using the 10 features from the first question to build a classification decision tree, the tree has an accuracy of 0.2437 on the test set, which is not a very high accuracy. This means that we cannot use these 10 features to predict which genre a song belongs to. We need more features to make more accurate predictions.

Extra credit

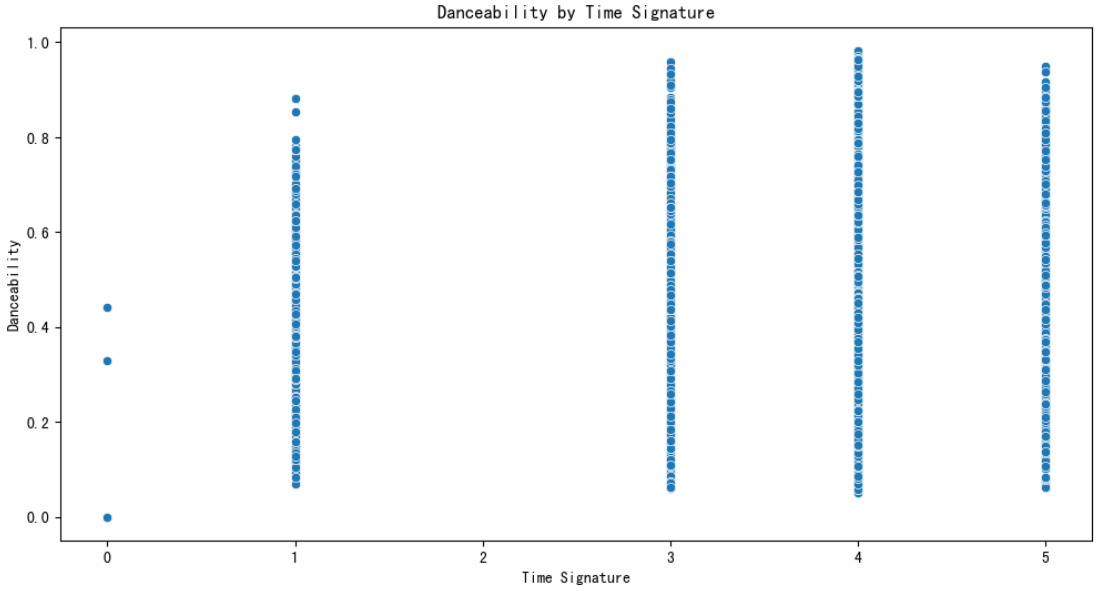
RQ1: Is there a relationship between a song's key and its popularity?

RQ2: How does the time signature of a song relate to its danceability?

For RQ1, I used Analysis of Variance (ANOVA) to determine if there are statistically significant differences in the popularity scores across different keys. For RQ2, I used linear regression to explore the relationship between the time signature and danceability of songs.



The boxplot displays the distribution of popularity scores across different musical keys. There's variability in the median popularity scores for each key, suggesting that certain keys might be associated with higher popularity. However, the overlap in the interquartile ranges indicates that the key alone may not be a strong predictor of popularity.



The scatter plot shows the relationship between time signature and danceability. It appears there's a concentration of songs with higher danceability in specific time signatures, particularly in 4/4 time. However, the spread also suggests other factors might influence danceability.

The ANOVA results show a statistically significant difference in popularity scores across different keys ($F=5.59$, $p<0.01$). This indicates that the key of a song could have an association with its popularity. However, the overlap in popularity ranges for different keys suggests that while the key is a factor, it's not the sole determinant of a song's popularity.

The linear regression analysis reveals a positive relationship between time signature and danceability, with an R-squared of 0.027. This means that about 2.7% of the variation in

danceability can be explained by the time signature. The coefficient for time signature is 0.0727, indicating that as the time signature increases, the danceability of a song also tends to increase. However, the low R-squared value suggests that other factors also play a significant role in determining danceability.