Capstone: Analysis

Yuanzou Gao

Preprocessing

Before I begin the question, I'd like to do some cleaning of the data. First, I want to convert duration from ms to s. Then, I load the data and check the data to have a basic understanding. Then, I would like to know information about the missing values. And by the code, there is no missing value for the data. However, I still use "dropna" since I would like to make sure about this.

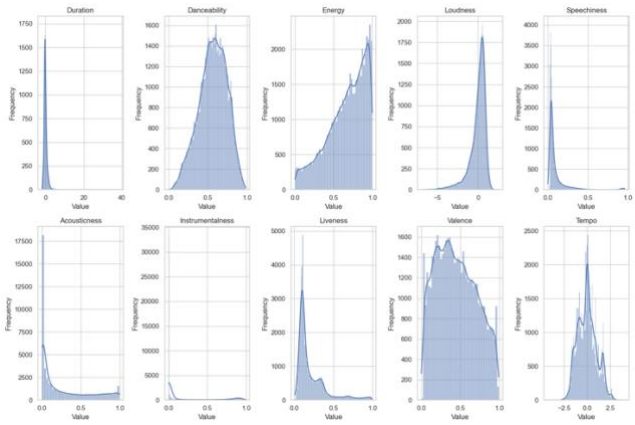
Then for the numerical variables, I would like to standardize 'popularity', 'duration', 'loudness' and 'tempo'. These four variables are not bounded between 0 and 1, and have a big range, so I would like to deal with them.

Question 1

What you did

For this question, we would like to figure out whether some features are distributed normally. I use 'histplot' to plot the histogram and let "kde" = True to visualizing the shape of the distribution. We have 10 plots in total, each subplot represents one feature. Then, I use Kolmogorov-Smirnov test to find whether the data is actually normal distribution. If the p-value is greater than 0.05, then the distribution is considered as normal.

What you found



```
K-S test for Duration:
Statistic=0.135, p-value=0.000

K-S test for Danceability:
Statistic=0.032, p-value=0.000

K-S test for Energy:
Statistic=0.092, p-value=0.000

K-S test for Loudness:
Statistic=0.134, p-value=0.000

K-S test for Speechiness:
Statistic=0.291, p-value=0.000

K-S test for Acousticness:
Statistic=0.200, p-value=0.000

K-S test for Instrumentalness:
Statistic=0.366, p-value=0.000

K-S test for Liveness:
Statistic=0.200, p-value=0.000

K-S test for Valence:
Statistic=0.053, p-value=0.000

K-S test for Tempo:
Statistic=0.042, p-value=0.000

Kolmogorov-Smirnov Test Results (p-values):
Duration: Not normally distributed (p-value = 0.000)
Danceability: Not normally distributed (p-value = 0.000)
Energy: Not normally distributed (p-value = 0.000)
Loudness: Not normally distributed (p-value = 0.000)
Speechiness: Not normally distributed (p-value = 0.000)
Acousticness: Not normally distributed (p-value = 0.000)
Instrumentalness: Not normally distributed (p-value = 0.000)
Liveness: Not normally distributed (p-value = 0.000)
Valence: Not normally distributed (p-value = 0.000)
Tempo: Not normally distributed (p-value = 0.000)
```

We first to pay attention to the ten plots. Actually, from the graphs, we can easily figure out that 'Energy', 'Loundness', 'Speechiness', 'Acousticness', 'Instrumentalness', 'Liveness', 'Valence', and 'Tempo' are not normal distribution, since the normal distribution is bell shape, and none of them is bell shape. 'Duration' and 'Danceability' may be the two that closest to the normal distribution. Then, we want to use Kolmogorov-Smirnov test to make me confidence about what I am thinking from the graph.

From the output, we can see that the p-values are all very small and, in that way, none of the distribution will be considered as normal. This is consistent with the graph.

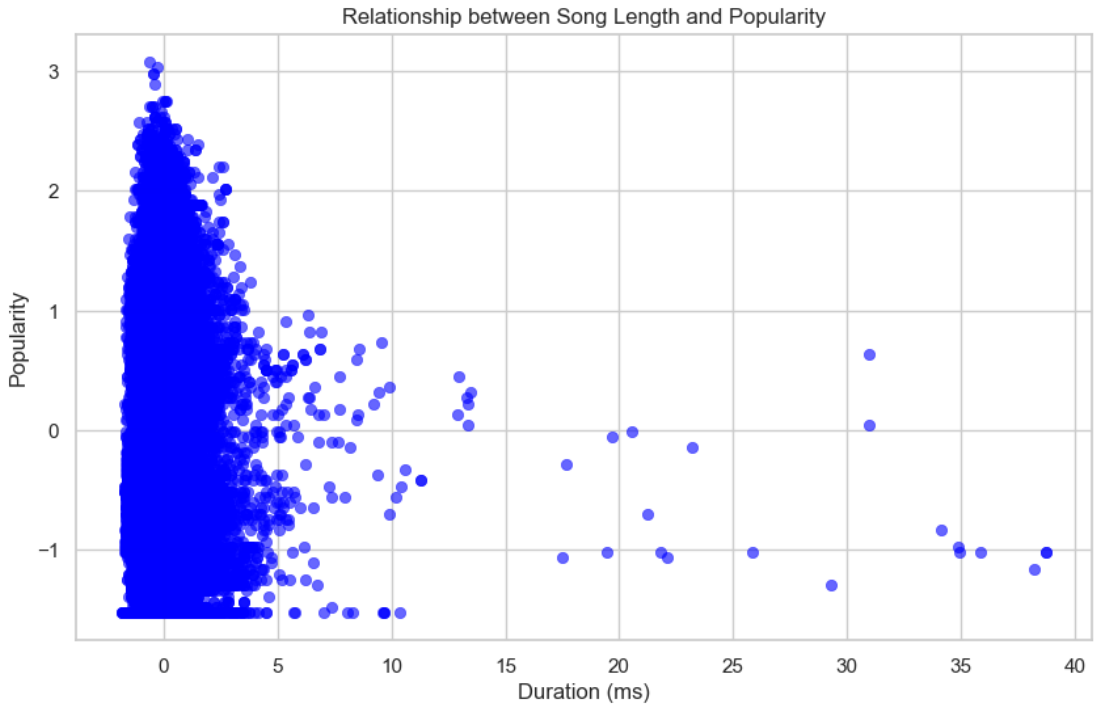
Question 2

What you did

For this question, we want to find the relationship between song length and popularity and create a scatterplot to visually the graph of the relationship. Then, we want to calculate the Pearson Correlation Coefficient to help us figure out the strength between

the relationship, and if there is relation, we can know it is a positive or negative.

What you found



From the plot, it will be very difficult for us to find out the relationship between song length and popularity. And I will consider there is no relationship between song length and popularity of a song. Then, I calculate the Pearson correlation and it equals to -0.055. In that way, the correlation coefficient is very close to 0, which means that there is no linear relationship. And it is consistent with what we figure out in the plot.

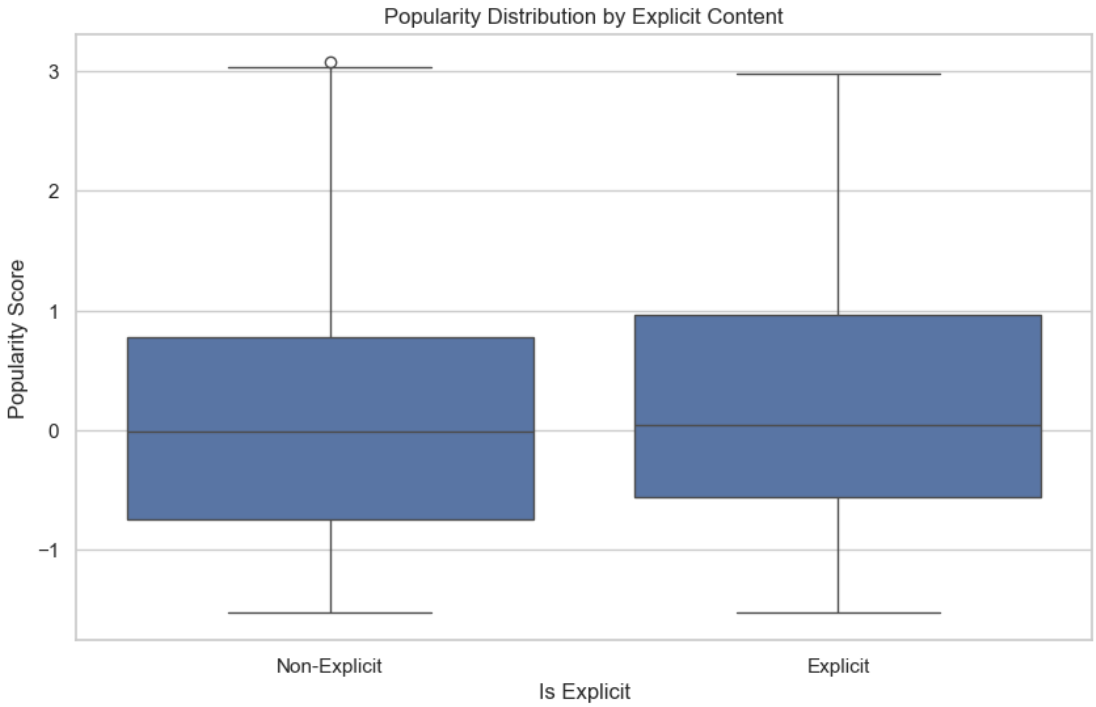
Question 3

What you did

For this question, we want to know explicitly rated songs are more popular than songs that are not explicit. Then, we want to use Mann-Whitney U test for this problem. The reason is that: explicit is a categorical variable. I would like to use a non-parametric test. Then, we divide into two groups with explicit = "FALSE" or "TRUE". Then, we use Mann-Whitney U Test. This could be useful when determine the difference in the popularity distributions for different explicit groups. The null hypothesis will be explicit songs are not more popular than non-explicit songs. And the alternative hypothesis is that explicit songs are more popular than non-explicit songs. I also make a histogram to visualize.

What you found

Here is the output I got.
U Statistic: 139361273.5
P-value: 1.5339599669557339e-19



We know that the p-value is approximately equal to zero, this suggests that there is a statistically significant difference in popularity between explicitly rated songs and those that are not explicit. From the boxplot, it seems like they have similar median value. This suggests that even if two samples have the same median, they can have a different distribution. And there are outliers for the 'Non-Explicit' group. The U statistics is very very large, this means that one groups tends to have rankings higher than the others. And the p-value is approximately equal to zero suggest that we should reject the null hypothesis, and this suggests that explicit songs are statistically more popular than non-explicit songs.

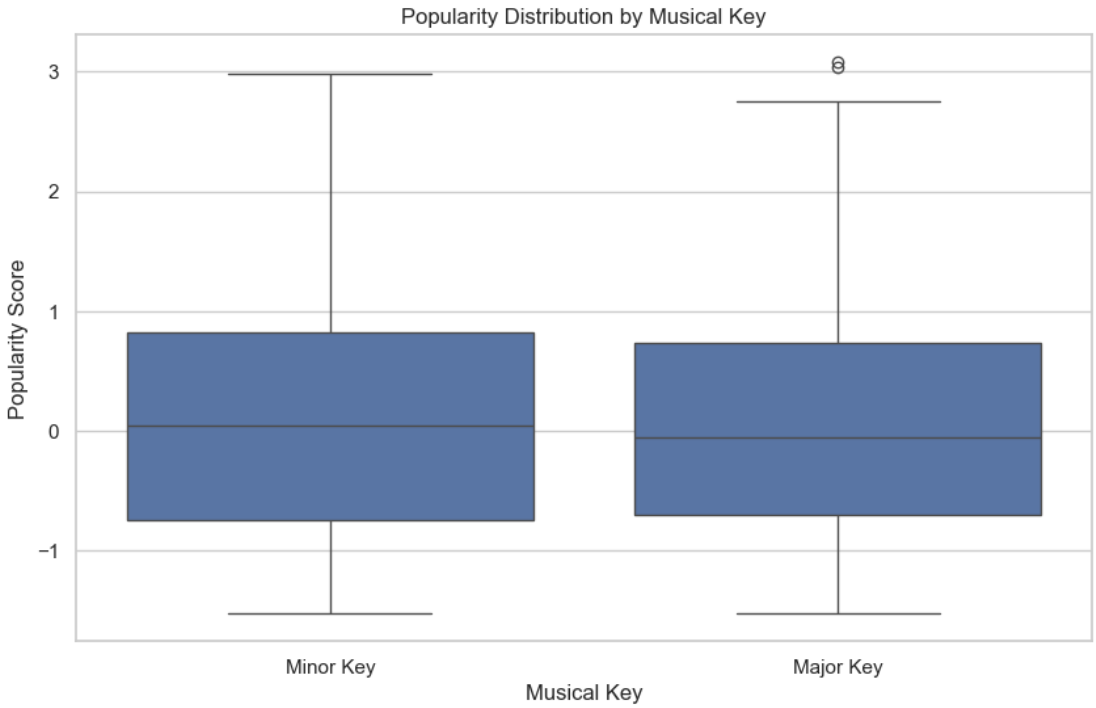
Question 4

What you did

For this question, we would definitely want to use a non-parametric test since the mode is categorical variable. And I choose the Mann-Whitney U test. First, we want to separate the data into two groups based on the mode. Then, we run the test to see if there is a statistically significant difference in popularity between songs in major and minor keys. For the null hypothesis, we have songs in a major key are not more popular than songs in a minor key. And the alternative hypothesis is defined as songs in a major key are more popular than songs in a minor key. Finally, I also create a graph to visualization.

What you found

Here is the output I got.
U Statistic: 309702373.0
P-value: 0.9999989912386331



For this problem, the p-value is approximately equals to 1 and greater than 0.05. This suggests that we cannot reject the null hypothesis that there is no difference between songs in major keys and those in minor keys. The U statistic is equals to 309702373.\
Then, we pay attention to the boxplot. We see that the major and the minor have a similar median popularity score. And there are some outliers for major key, where these points exceed the typical popularity scores. Given the information, we know that there is no significant difference in the popularity levels between major and minor keys for songs.

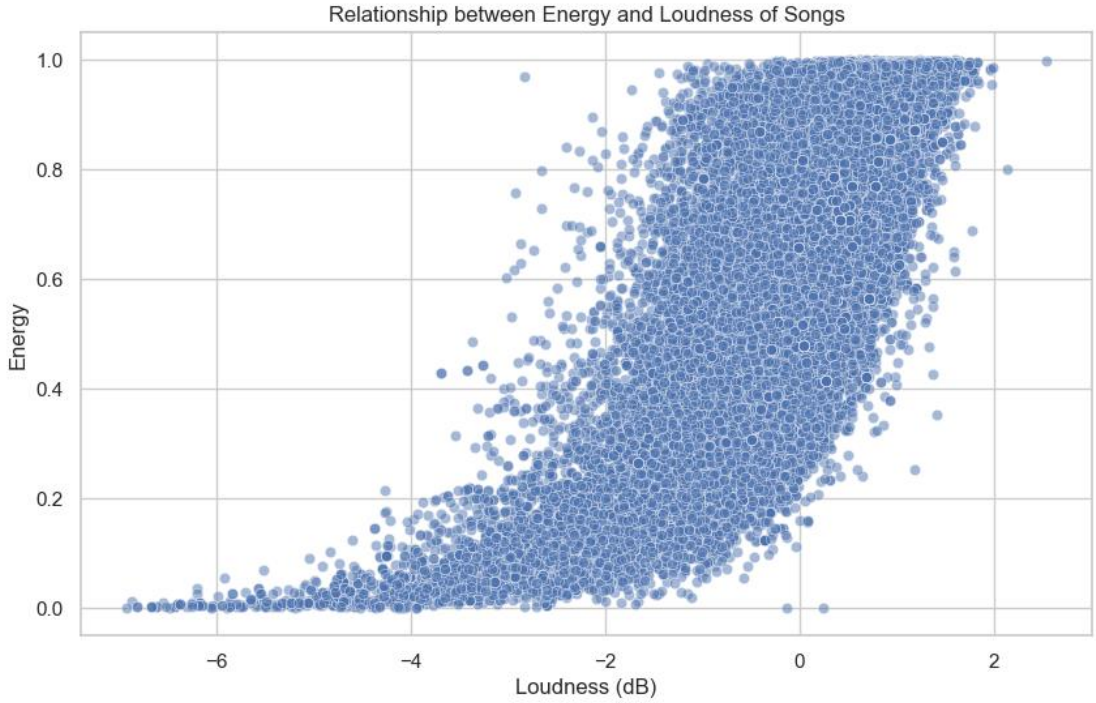
Question 5

What you did

For this question, we want to create a scatterplot to show the correlation between the two. First, we want to load the dataset, and create a scatterplot to visual the relationship between energy and loudness. We also want to calculate the Pearson correlation coefficient to learn about the strength and direction of linear relationship between energy and loudness.

What you found

We figure out that the Pearson correlation coefficient between energy and loudness is 0.775. This is a relative strong correlation, and with a positive number. This means that as loudness increases, there is also an increase in the energy.



The plot shows that as loudness increases, energy of the song also increases. There is a big cluster in the median part and right part, which is a positive relationship. In general, we have the evidence that louder sounds are often associated with higher energy.

Question 6

What you did

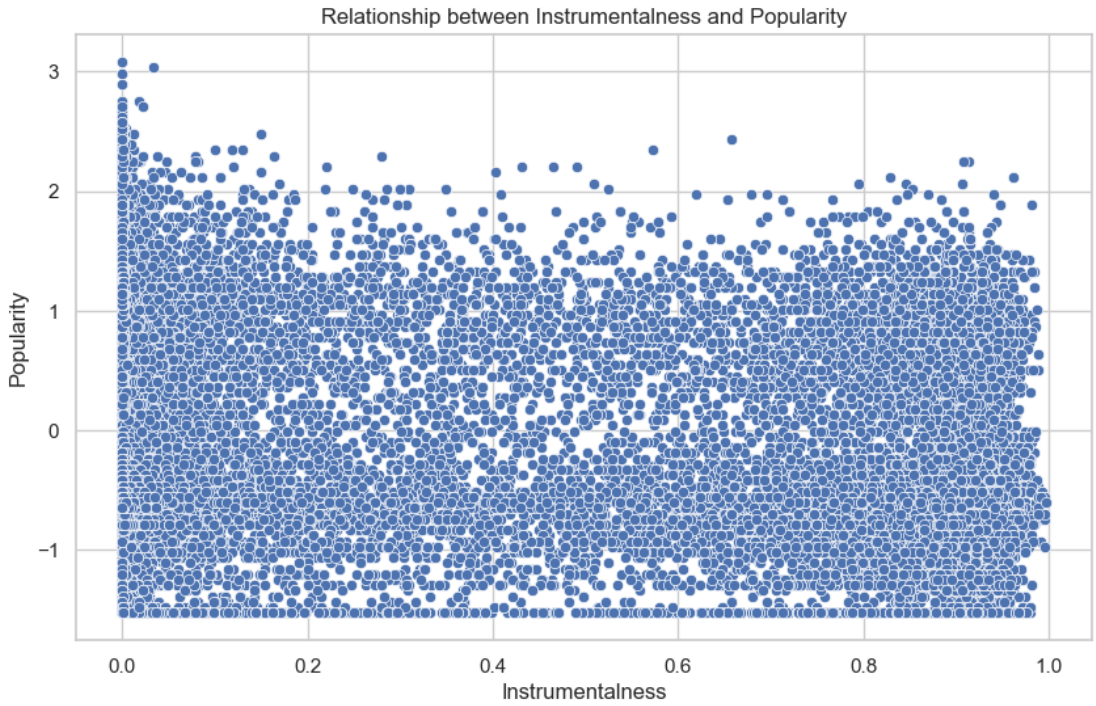
For this question, we want to conduct 10 regression analyses. And at each one, we want to find the relation between popularity and each feature. And then using R^2 to indicate the proportion of variance in the popularity that is predictable from the feature.

So, we first want to identify the features, and conduct an analysis for each feature against popularity, and calculate the R^2 for each model. Then we can compare the results and find which feature is the best predictor of the popularity and how good the model is. We also want to create a plot for the best feature, and visually see the relationship.

What you found

Here is the output I get:
Best predictor of popularity: instrumentalness

Coefficient of determination (R^2): 0.0210
Coefficient (Beta value) for instrumentalness: -0.4457
P-value for instrumentalness: 0.0000



We figure out that the best predictor of popularity is the instrumentalness. In that way, instrumentalness has the strongest relationship with song popularity in your model. And the coefficient of determination equals to 0.021, which is a very small number. This suggests that 2.1% of the variability in popularity can be explained by instrumentalness. While this value is relatively low, it highlights that instrumentalness still has a significant, albeit small, effect on popularity compared to other single features tested.

And the coefficient for the instrumentalness is -0.4457, which is a negative relationship. So as instrumentalness increases, popularity tends to decrease. And the p-value equals to 0 which means that the relationship between instrumentalness and popularity is statistically significant.

The graph could be a tool for us to better understand the numbers above, and actually, we can see from the graph that there will not have a very big coefficient of determination value.

Question 7

What you did

For this question, we want to build a model that uses all of the song features from question 1. And I would like to divide the dataset into training and testing sets since it

is a very big data and has a lot of features. It could possibly have overfitting, so we want to divide into two sets.

First, I load the data and define x and y. We want to split the data into training and testing sets. And then build the regression model and train it. We make predictions on the testing set and calculate the R^2 scores. Also, I evaluate the performance on the training set to compare with the testing set.

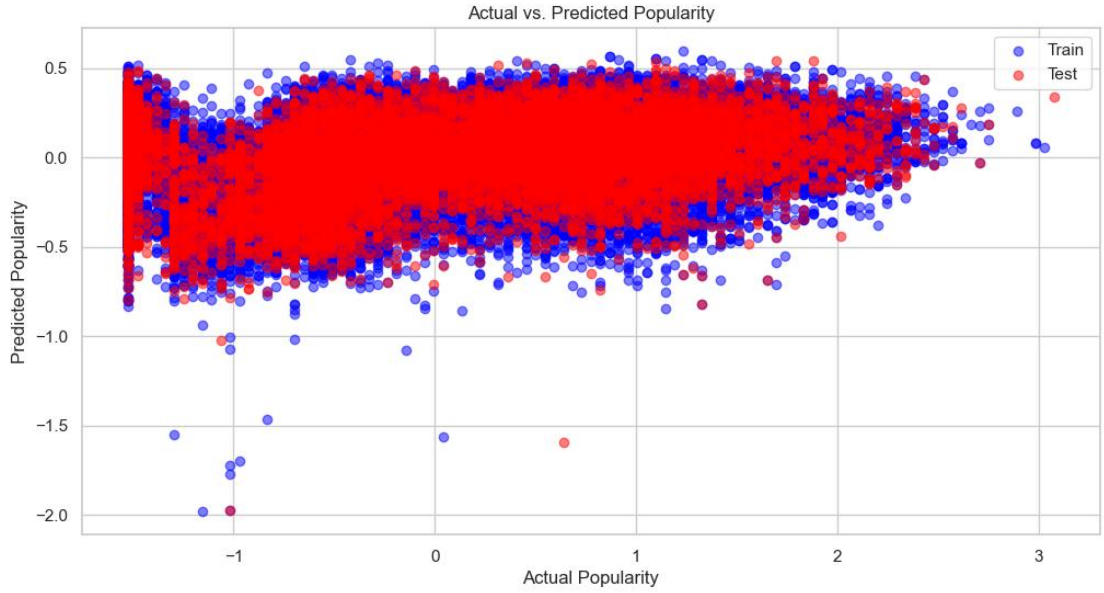
What you found

Here is my output:
R^2 score for the testing set: 0.0517
R^2 score for the training set: 0.0466

We figure out that the R^2 score for the testing set equals to 0.0517. This suggests that approximately 5.17% of the variance in popularity in the testing set is explained by the model. and although this indicates some predictive capability, the value is quite low, implying that the model explains only a small fraction of the total variability in the song popularity.

The R^2 score value for the training set equals to 0.0466. This means that approximately 4.66% of the variance in the training data is explained. And the value is very close to the testing model, this suggests a good performance of the model.

However, overall, this R^2 value is greater than the R^2 value for question 6, which means the model is getting better. However, the value is still very slow, which indicates that we still don't have a strong predictor.



We can see the plot shows a dense clustering of data points around the center. However, we could not see a clear trend about the correlation. Also, training and testing data

points are interspersed and show a similar pattern of distribution. This overlapping suggests that the model has a consistent performance across both training and testing sets, indicating good generalization without apparent overfitting.

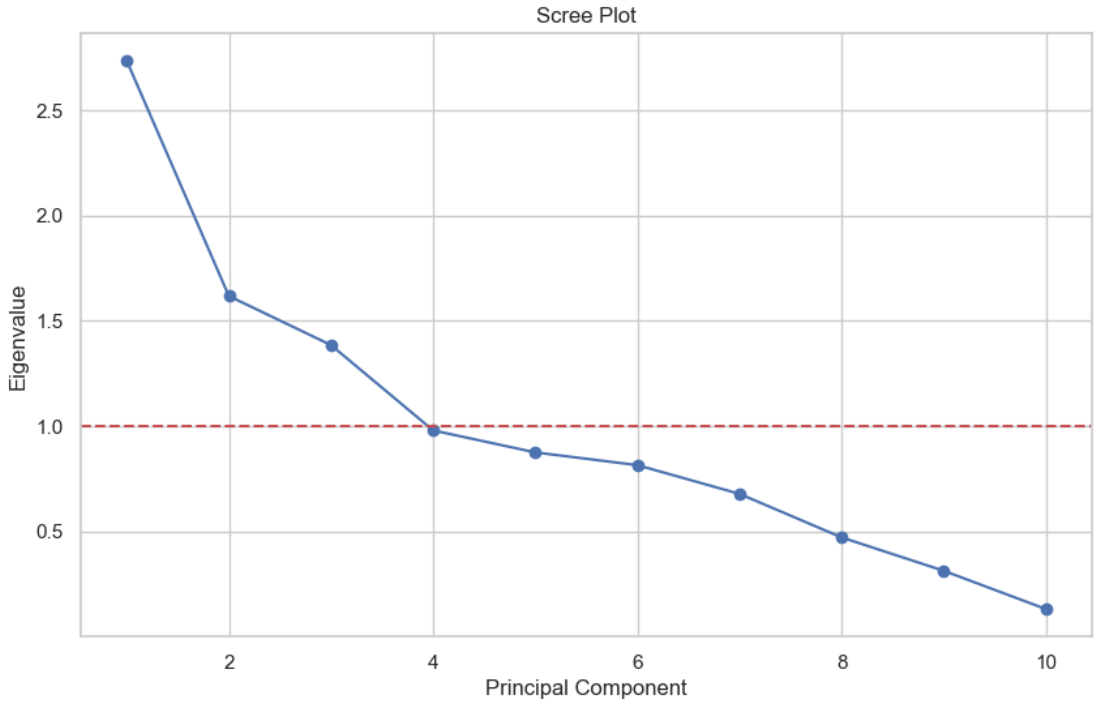
Question 8

What you did

For this problem, we first need to standardize the data since we want to do PCA. Then, we apply PCA to the scaled data. Then, we start fitting PCA to all components to see the variance expalined by each component. Then we can find the number of meaningful components by the eigenvalues.

What you found

Here is my output for this problem:
Eigenvalues: [2.73388097 1.61735976 1.3845787 0.97958798 0.8752094 0.81483073 0.67826859 0.47157203 0.31313367 0.13157818]
Number of components retained by Kaiser criterion: 3.



From the graph, we can find “Elbow”. By this criterion, there is 1 interpretable factor. And we know that three factors have Evas > 1 indicating that they are sufficient to summarize the majority of the information in the dataset without much loss of data integrity. 7 factors account for >90% (9.081=90.81%) of the eigensum.

Question 9

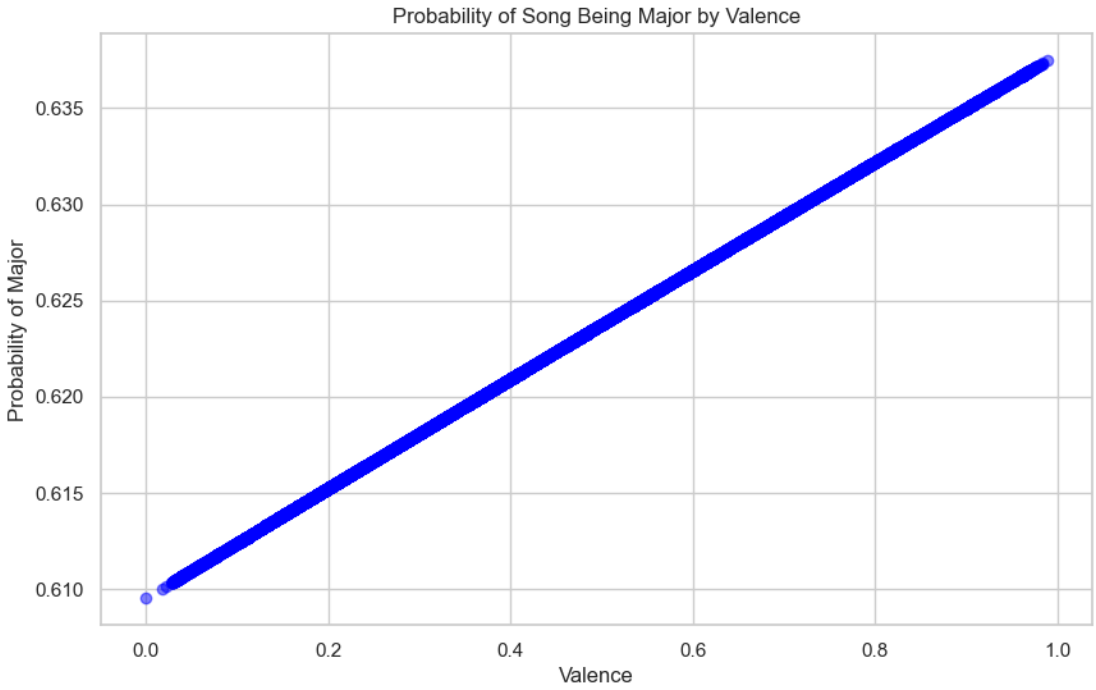
What you did

First, we want to build a model with valence as the predictor and the mode as the target. We then want to build and train the logistic regression model and then evaluate the model. We will use a classification report and a confusion matrix for this question. To better understand the model, I'd prefer to visualize the relationship between valence and the probability of a song being in a major key.

What you found

Here is the output I get:

```
              precision    recall  f1-score   support

           0       1.00      0.00      0.00      3899
           1       0.63      1.00      0.77      6501

    accuracy                           0.63     10400
   macro avg       0.81      0.50      0.38     10400
weighted avg       0.77      0.63      0.48     10400

[[   0 3899]
 [   0 6501]]
```



We can figure out that as valence increases, the probability of a song being in a major key also increases. This suggests that a positive relationship between valence and the likelihood of a song being in a major key. There's a perfect precision for the minor key,

where equals to 0, due to no predictions for this class, indicating poor performance. For the major key, the precision equals to 0.63, which is a moderate number. This results in a model that effectively ignores the minor key category, failing to identify any minor key songs accurately. In my opinion, I think that including additional features might improve model performance significantly.

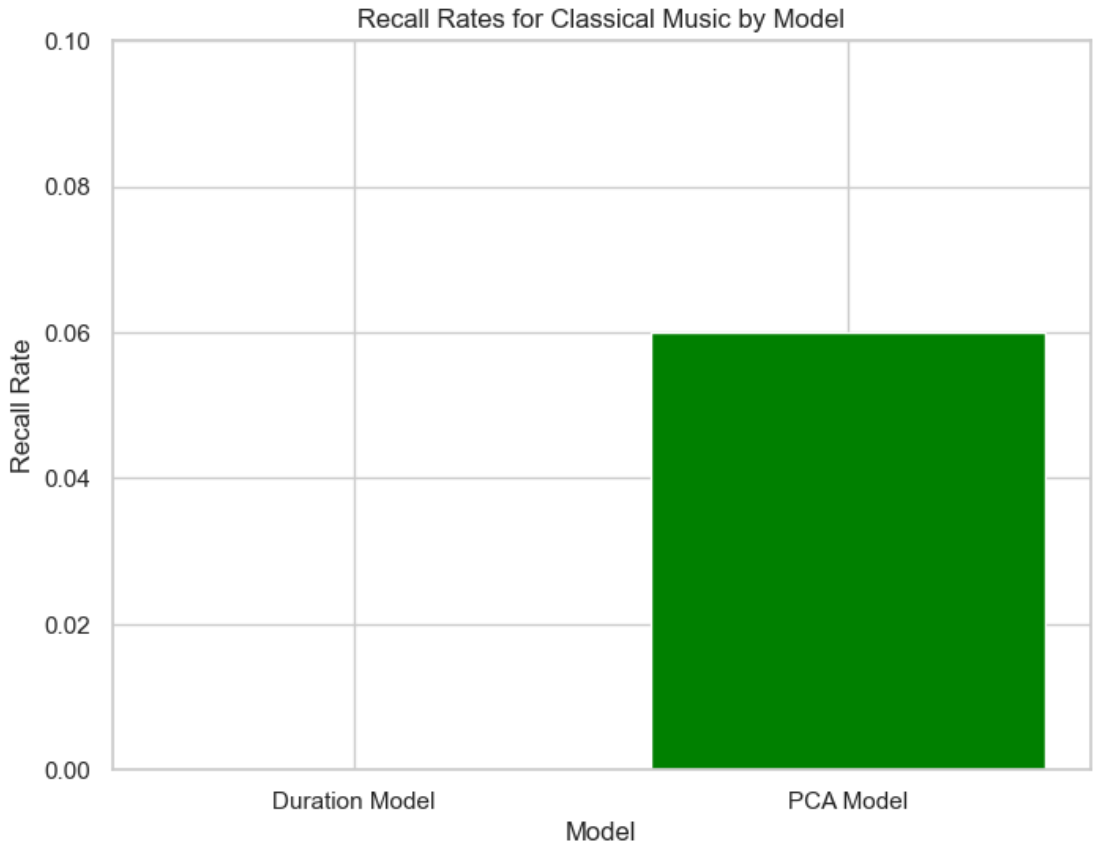
Question 10

What you did

The most important thing for this question is to convert genre into a binary numerical label with 'classical' is labeled as '1'. We want to compare 'duration' and 'principal components'. Since we know the number of components from problem 8, and we want to reapply PCA with this number. Then, we want to compare the predictive performance of duration and the principal components through logistic regression. Then, we want to evaluate the classification reports to determine which one is better predicting.

What you found

```
Duration Model Performance:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10189
           1       1.00      0.00      0.00       211

    accuracy                           0.98     10400
   macro avg       0.99      0.50      0.49     10400
weighted avg       0.98      0.98      0.97     10400

PCA Model Performance:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10189
           1       0.32      0.06      0.10       211

    accuracy                           0.98     10400
   macro avg       0.65      0.53      0.54     10400
weighted avg       0.97      0.98      0.97     10400
```

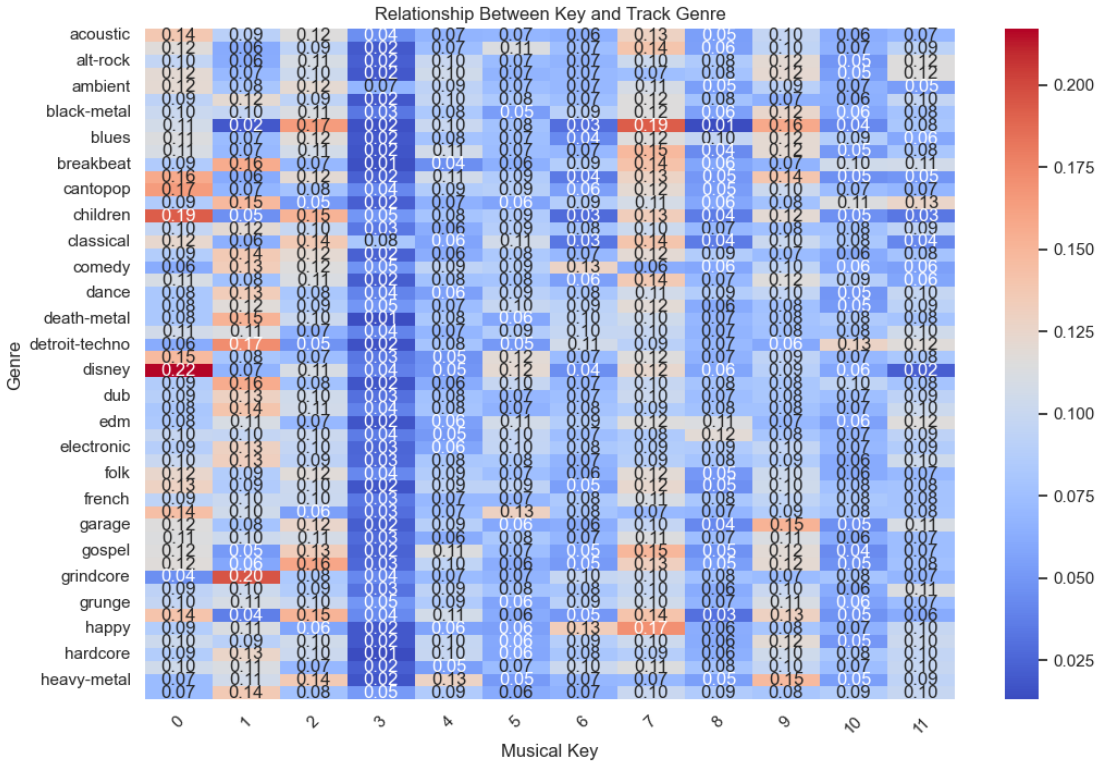


For the duration model, the precision is perfect at 1.00, indicating that the few predictions it made for the classical category were all correct. However, the recall is extremely low at 0.00, indicating that the model failed to identify almost all actual classical songs, resulting in an F1-score of 0.00. This suggests the model almost never predicts songs as classical. The accuracy for the model is high = 0.98, it may be due to the model is effective at predicting the non-classical songs.

For PCA model, precision = 0.32, which is much lower than in the duration model, indicating less reliability in the predictions made for classical songs. However, the recall = 0.06 is higher, although still very low, indicating that this model, while not very precise, does attempt to identify classical songs more often than the duration model. Also, the accuracy = 0.98 is very high, same as the duration model, it is because its goods at predicting the non-classical songs.

In general, both models show a strong bias towards predicting classical songs. The PCA model, which with low precision, has identifies some classical songs compared with the duration model, which almost ignore the whole model. So, although both model is bad, I'd prefer PCA.

Extra Credit



I was wondering whether there is a relationship between key and genre. To be more specifically, I was thinking maybe some genre will have their prefer key. So, I would like to build a plot to see whether there actually exists such a relationship.

Based on the graph, it confirms my guess. For example, Disney prefers key 0, and Grindcore has a strong prefer for key 2. Also, something very crazy is that key 3 seems to be abandoned by every genre. I think some reasons to cause certain keys might be preferred in specific genres due to their historical development, the instruments typically used and so on.