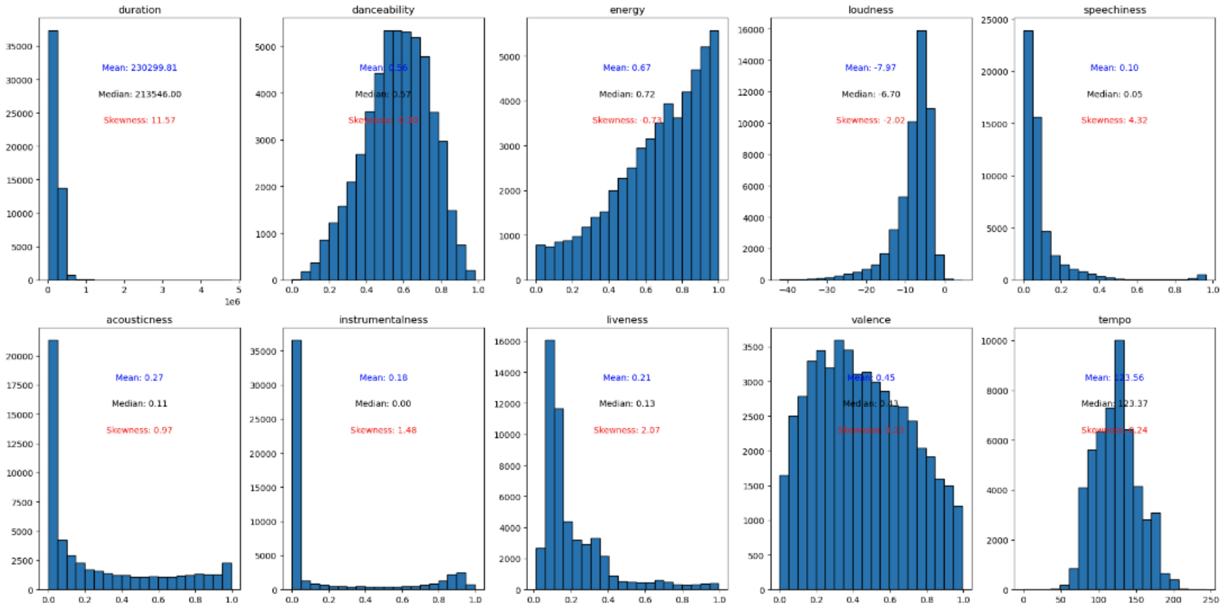
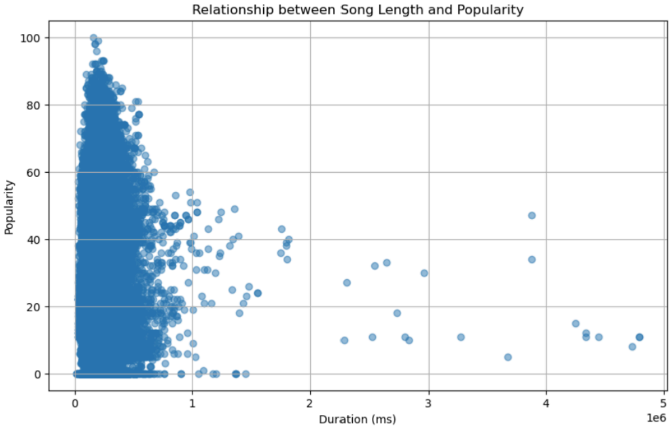
Introduction:
I first investigated the dataset’s shape, values, and datatypes as well as searched for null/nan values or invalid values. I found none, so there was no missing data and no need for imputing or removing any of the data. I did not find the need to apply any dimensionality reduction methods either for any of the questions.

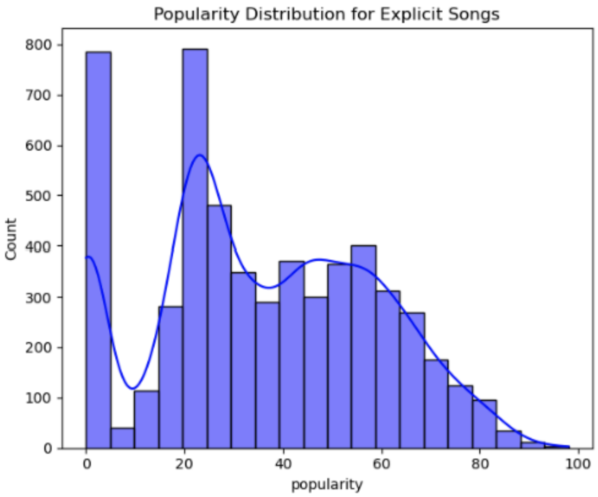
1. From observing the scatterplots, we can see that tempo and danceability are the only features that are reasonably normally distributed. Tempo has a mean of 123.56 and median of 123.37 with a skewness of 0.24. Danceability has a mean of 0.56, median of 0.57, and skewness -0.30. A mean and median close to each other with a low skew makes the distribution reasonably/approximately normal and these are the only two features are distributed as such.

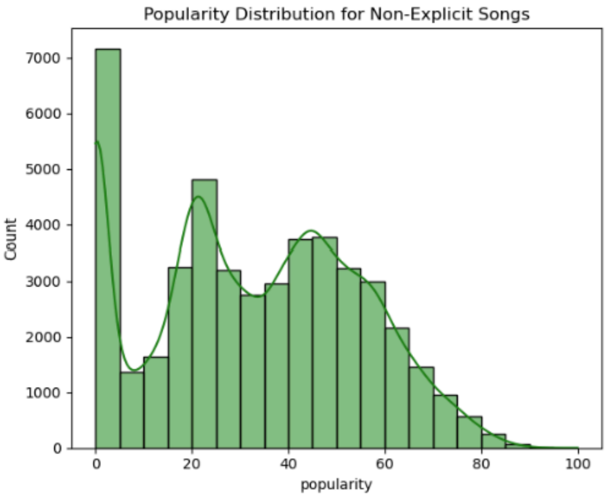


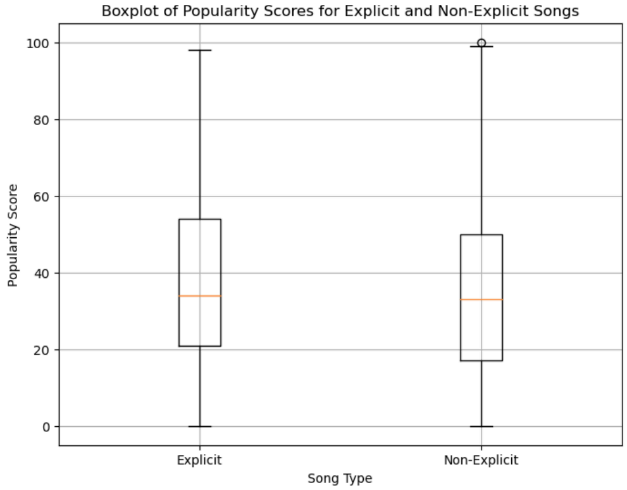
2. There is not a linear or monotonic relationship between song length and popularity as indicated by the near-zero pearson correlation coefficient of -0.0547 and spearman correlation coefficient of -0.0373. Additionally, inspecting the scatterplot also does not reveal any kind of strong relationship.

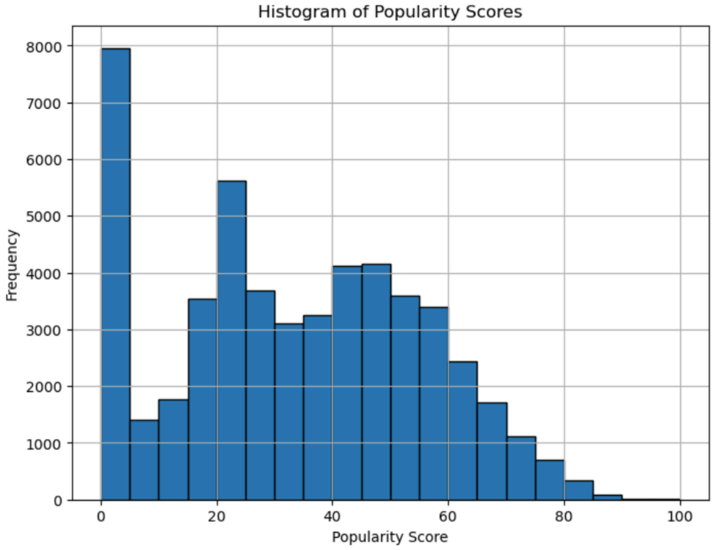


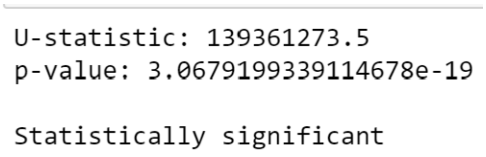
3. The null hypothesis for this test is that explicit songs are not more popular than non-explicit songs. First I looked the how popularity scores were distributed and as seen by the histogram below it is clearly not normal, and the same goes for the distribution of explicit and non-explicit songs. Therefore, I used a Mann-Whitney U test that does not have any assumptions about populations being distributed normally. According to it, the p-value was 3.07e-19 which is much lower than 0.05, making this result statistically significant and indicating explicitly rated songs are more popular than non-explicit songs. I additionally created a boxplot to visualize these results.


Popularity Distribution for Explicit Songs

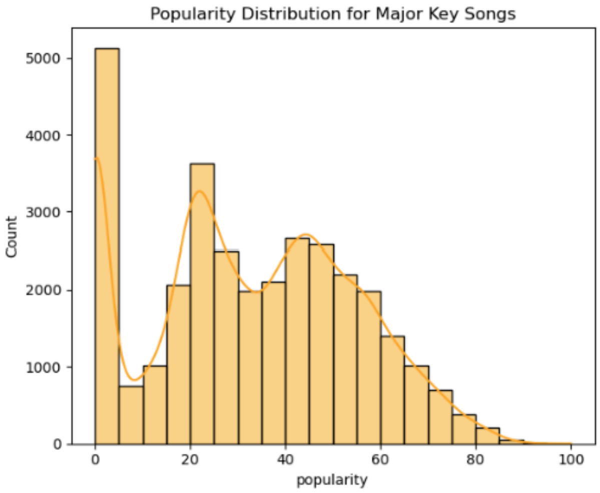

Popularity Distribution for Non-Explicit Songs

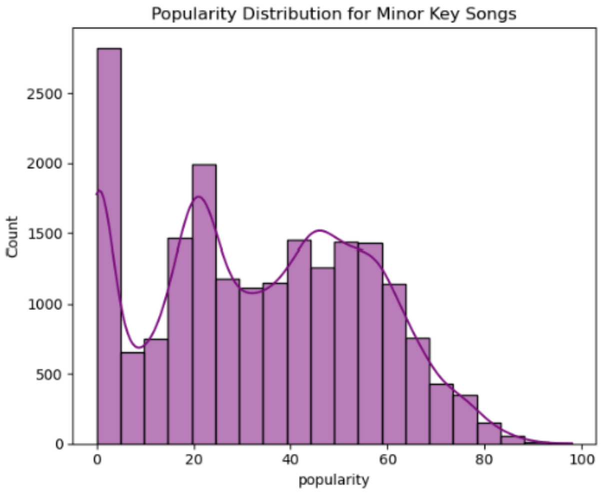

Boxplot of Popularity Scores for Explicit and Non-Explicit Songs

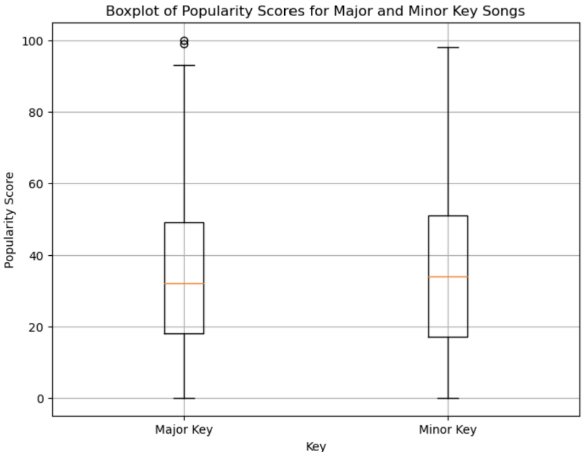

Histogram of Popularity Scores

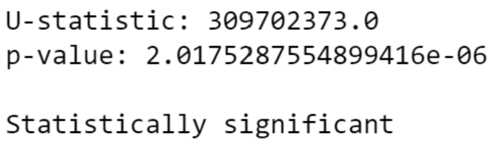

```
U-statistic: 139361273.5
p-value: 3.0679199339114678e-19

Statistically significant
```

4. The null hypothesis for this test is that songs in the major key are not more popular than songs in the minor key. For the same reasons as above, I utilized another Mann-Whitney U test, obtaining a p-value of 2.02e-06, again making this result statistically significant. However, to interpret the “direction” we observe the boxplot and see that minor key songs are more popular than major key songs, and this is the statistically significant result. Songs in the major key are not more popular than songs in the minor key.



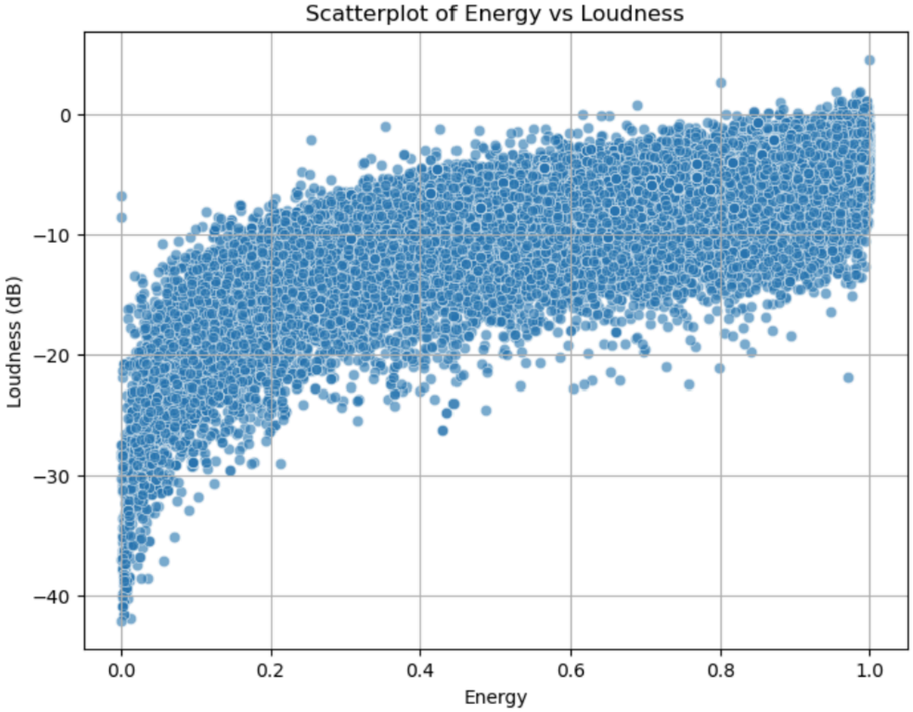






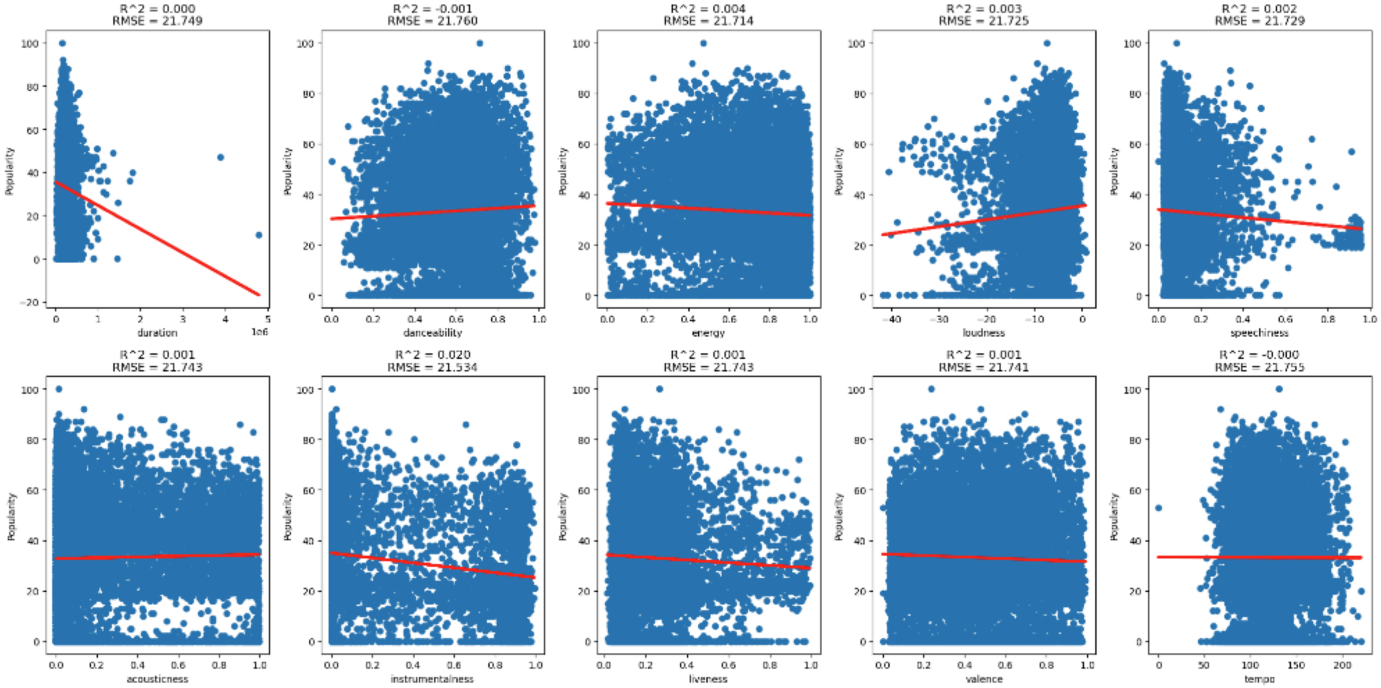
```
U-statistic: 309702373.0
p-value: 2.0175287554899416e-06

Statistically significant
```

5. The below scatterplot shows a substantial positive linear relationship with a calculated correlation coefficient of approximately 0.775, substantiating the claim that energy largely reflects the “loudness” of the song. As the energy increases, the loudness generally also increases.



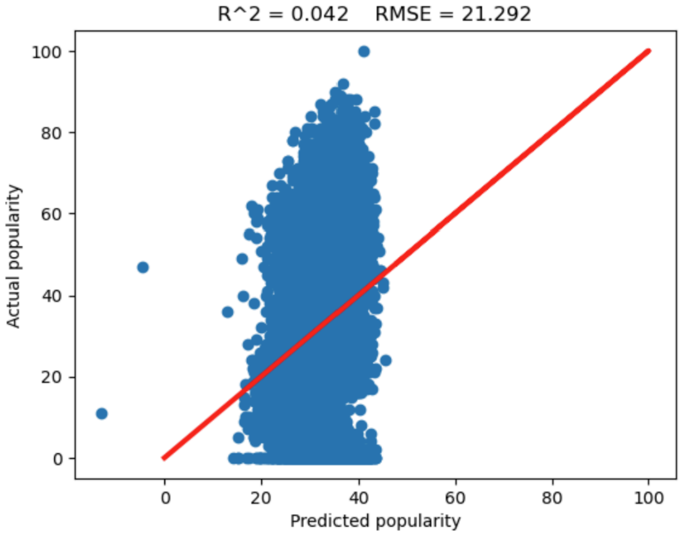
```
0.774880829185019
```

6. I utilized linear regression because based on one predictor we are attempting to predict a continuous value (popularity scores). Below are scatterplots of each feature vs. popularity scores and the line indicating the linear regression. None of the scatterplots feature a linear relationship or really any clear/strong relationship at all just upon visually observing it. The best model utilizes instrumentalness which is the best predictive feature as it has the highest R^2 of 0.02 and lowest RMSE of 21.5. However this indicates that despite being the “best”, the model is still not very good and instrumentalness alone is not a great predictor of a song’s popularity as the model only explains about 2% of the variance.

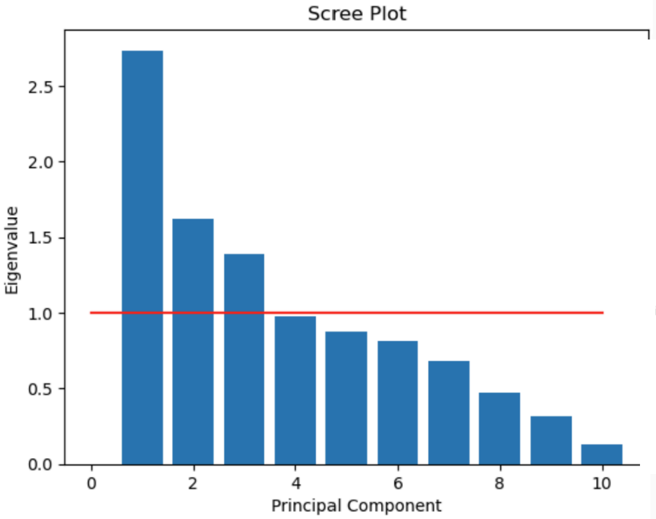


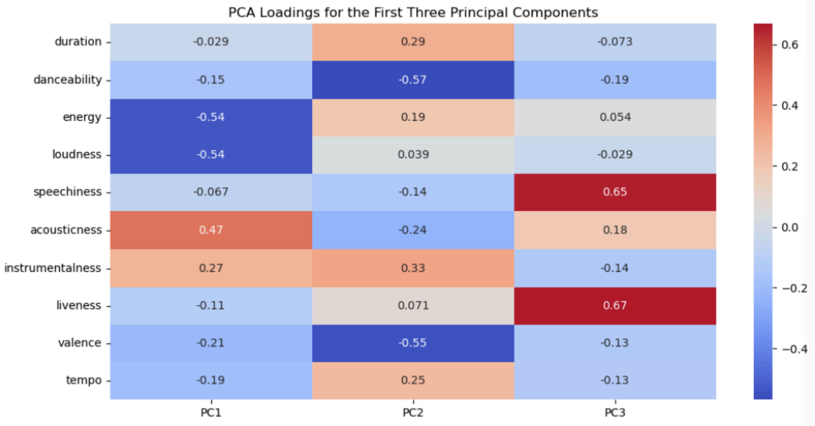
7. Training a Linear Regression model on all of the features improved model performance slightly, resulting in an R^2 of 0.042 and RMSE of 21.292. However, this is still very poor performance as the model is only accounting for 4.2% of the variance and is not very good at all at predicting popularity scores based on a song features. This indicates the model is underfitting, possibly because none of the relationships are very linear. I think it would be beneficial to use a more complex model that captures more nonlinear relationships. I tried to standardize the scaling of the dating using StandardScaler() to see if it would help but it made no difference to the model’s performance.

```
R-squared using all features: 0.041946946217162444
RMSE using all features: 21.29241819088791
```

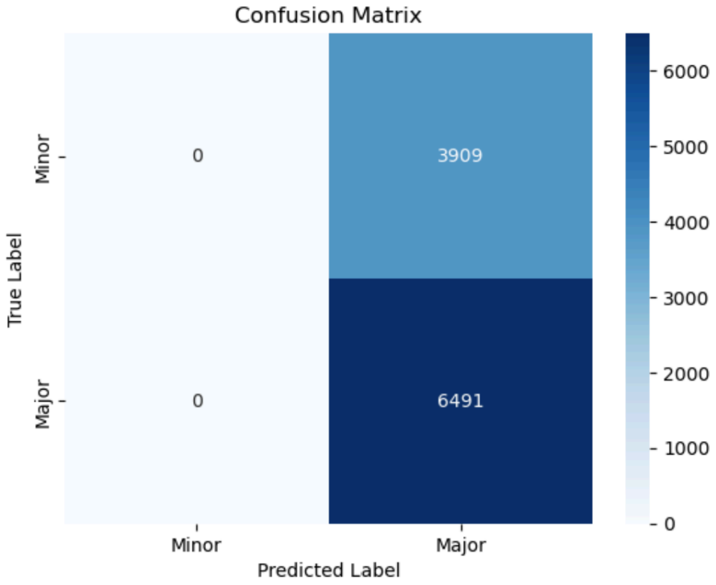


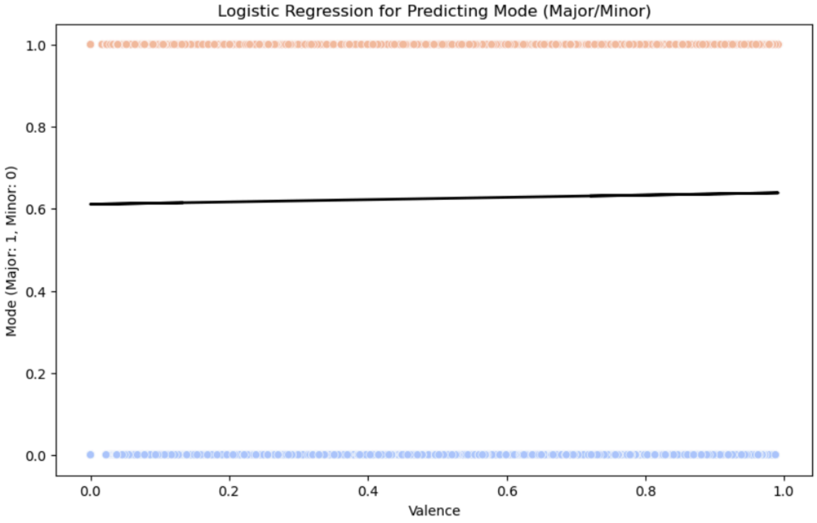
8. I was able to extract 3 meaningful principal components utilizing the Kaiser criterion that cumulatively account for about 57% of the variance with component 1 accounting for about 27%, component 2 about 16% and component 3 about 13.9%. I inspected the loadings for each component and determined the component 1 seems to represent quieter more instrumental songs as they have high acoustic ness and instrumentalness and low energy and loudness. Component 2 seems to represent calmer, average-feeling signs as they are slightly instrumental with a slightly higher tempo and energy but have very low danceability and valence. Finally component 3 seems to represent concert recordings/live albums as they have high liveness and speechiness. More figures can be found in my code.

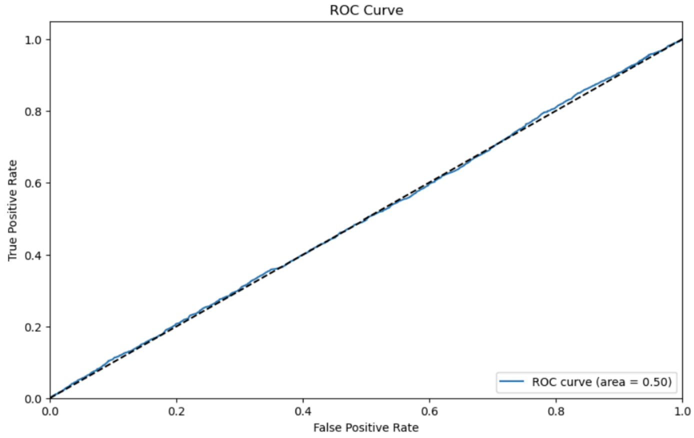




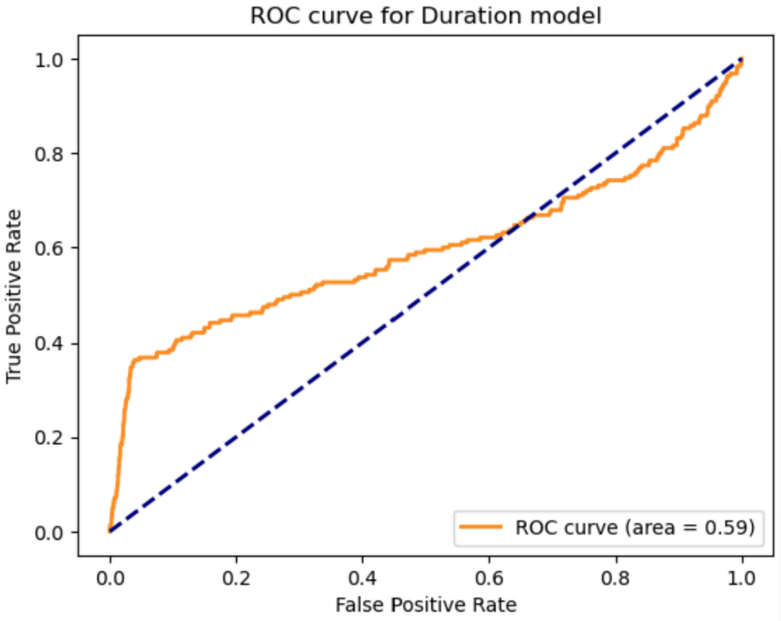
9. You cannot predict whether a song is in major or minor key from valence. The model had an accuracy of 62.4% and looking at the confusion matrix the logistic regression is always predicting that the song is in major key. The plot below also shows that there are songs in major key and minor key for almost every possible value of valence. Additionally, plotting the ROC curve shows that it is a straight line and the AUC score is 0.50, indicating the model is just randomly guessing. Specifically it is guessing the majority class for every prediction as checking the distribution of modes reveals about 32.4k songs are in major key while only 19.6k are in minor key in this dataset, showing a class imbalance that must be tackled for any model to be a good predictor. That being said, building a logistic regression for all of the other song features with which to predict mode revealed a similar distribution, where for a given value for that song feature it had songs both in major and minor key. None of the features are good predictors of mode.

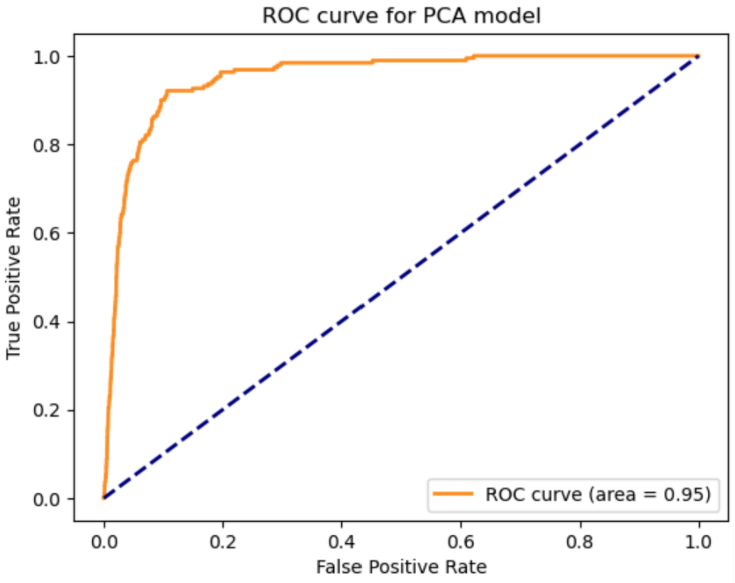






10. The principal components are a better predictor of if a song is classical or not. I made two logistic regression models and initially it appears duration and pca are equally good predictors as they have an accuracy of 98.2% and 97.9% respectively but observing the confusion matrices, ROC curves, and AUC scores reveals pca is much better than duration. Especially the AUC scores as for duration it is 0.59 indicating the model is randomly guessing the majority class (in this case, that is not classical which has 51k counts as opposed to only 1000 classical songs) but the AUC score for the pca model is 0.95, indicating that it is correctly prediction both the majority and minority classes.

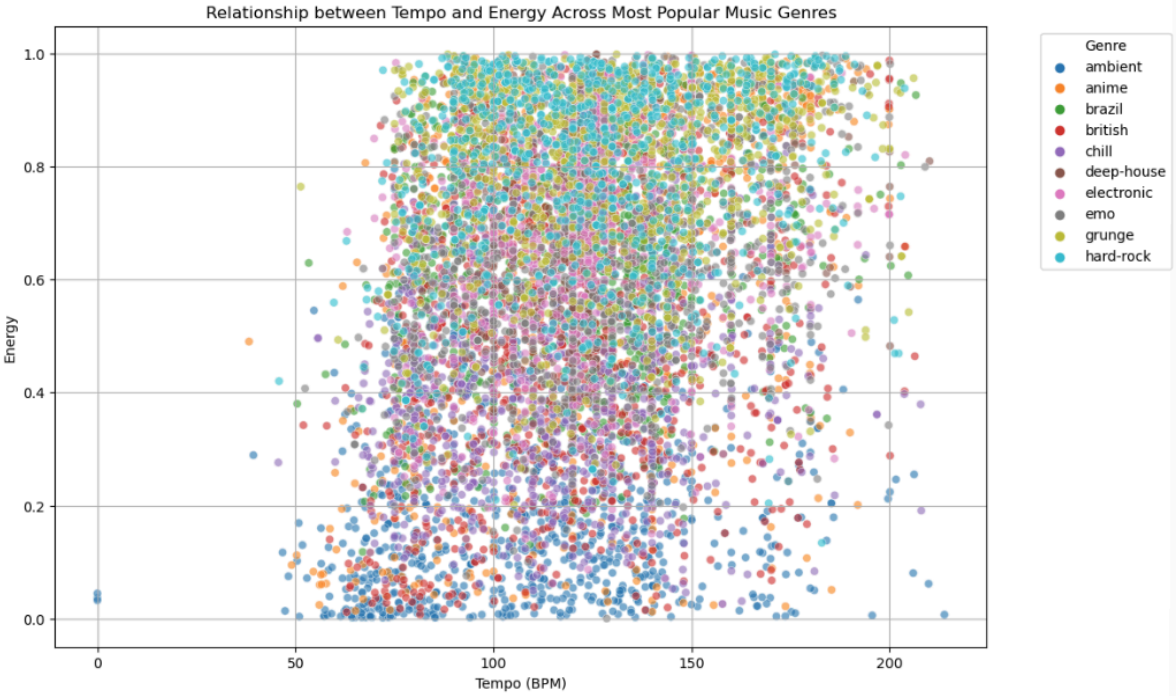




(EXTRA CREDIT ON NEXT PAGE; PLEASE READ)
[My apologies, it was difficult to format and fit it onto this same page]

Extra Credit:

For this I wanted to explore the relationship between tempo and energy for the 10 most popular genres. I found the most popular genres by taking the average popularity score of each genre and taking the top ten, then making a plot of their energy and tempo. On spotify, the 10 genres which the highest average popularity scores are: ambient, anime, brazil, british, chill, deep-house, electronic, emo, grunge, and hard-rock. Ambient had the lowest energy but variable tempo. Hard-rock and grunge tended to be high energy with electronic mostly concentrated around 100 BPM. Very few songs from these genres had a tempo around below 50 or above 200. There were songs from these genres in all ranges of energy, however.


Relationship between Tempo and Energy Across Most Popular Music Genres
Energy
Tempo (BPM)
Genre
● ambient
● anime
● brazil
● british
● chill
● deep-house
● electronic
● emo
● grunge
● hard-rock