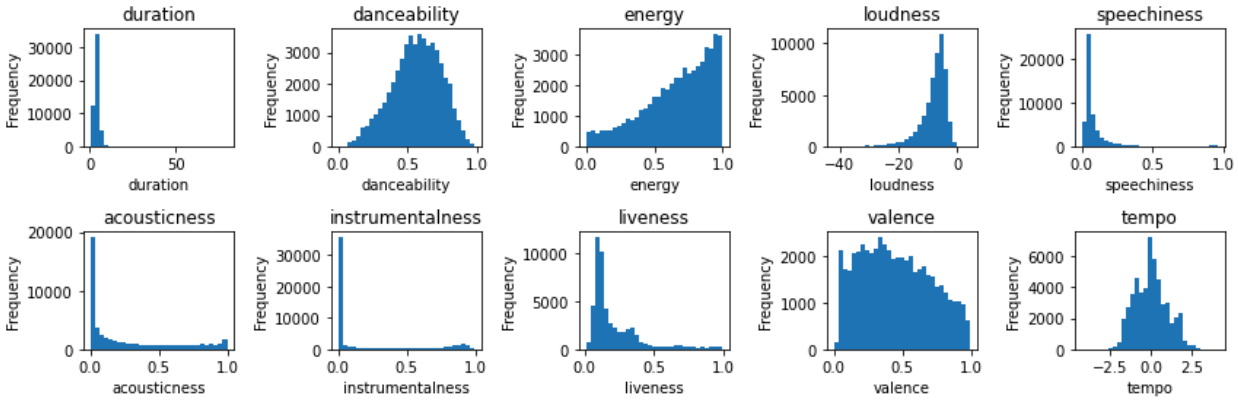
Spotify Analysis

Preprocessing Data:

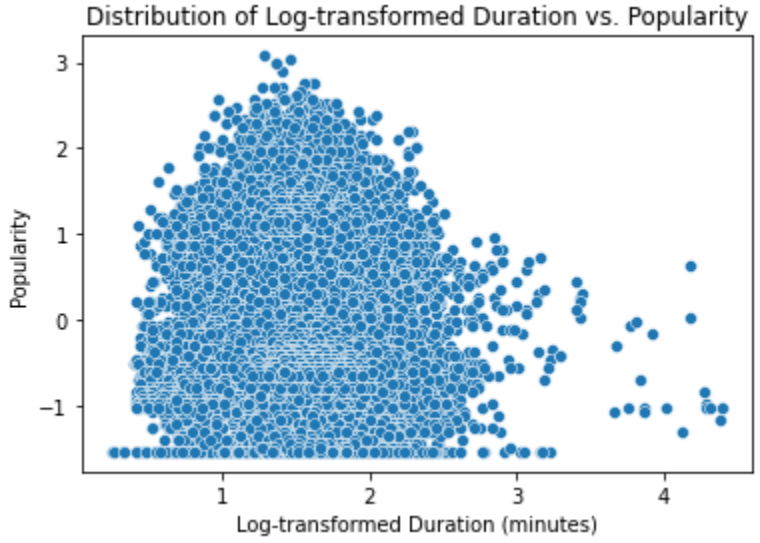
To preprocess my data, I began by data cleaning first checking for null values. Finding none, I confirmed this by iterating through each column to ensure no values were masked in the form of strings, such as “N/A”, and looked at their values. Data Transformations like log transformations and normalization were applied specifically to the analysis requirements of each question to ensure data integrity and suitability. With no data loss and under the assumption that each song is independent, the dataset is now ready for exploration and analysis. Dimension Reduction was achieved by transforming data to z-scores and applying PCA. A scree plot was made visualizing the Kaiser criterion and eigenvalues. It should be noted that duration was converted from milliseconds to minutes during the preprocessing. Whenever a RNG was required, I used my NYU ID as the seed. Regarding histograms, I used 31 bins throughout (odd number for data distribution).

1) Consider the 10 songs that feature duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Are any of these features reasonably distributed normally? If so, which one?
To observe the distribution of the 10 features, I chose to create a histogram. To do this, I created a list of the 10 features to create a data frame, then checked and dropped any rows with NA values in order to clean the data. To iterate through and create histograms, I used plt. subplots() and created a 2 x 5 figure for each feature. The ‘axes’ array was flattened to iterate through the loop, where each histogram is labeled by its feature, and every subplot displays the distribution of a feature. Out of these features, “danceability”, an audio feature provided by the Spotify API that quantifies how easy it is to dance to a song, was found to be reasonably distributed normally, with a slight left skew.



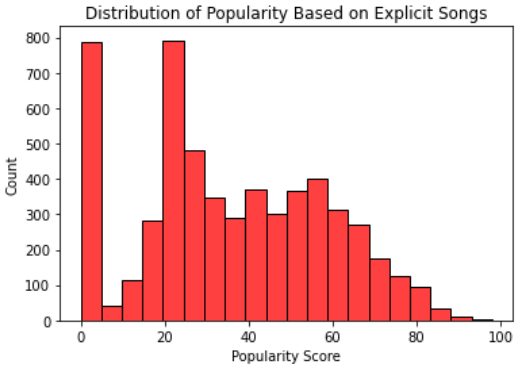
2. Is there a relationship between song length and popularity of a song? If so, is the relationship positive or negative?

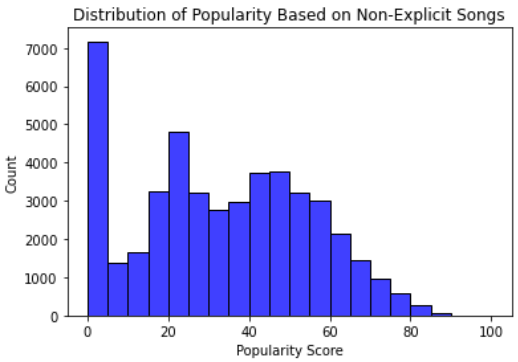
To observe any relationship between song length and popularity of a song, I used Matplotlib and the Seaborn library to visualize the data and found no strong relationship. To manage skewness and extreme values previously seen under the distribution for the duration, I applied a log transformation to the duration. Despite this adjustment, no strong relationship was found. I used Pearson's r, resulting in a correlation coefficient of -0.031 indicating a very weak linear relationship, Additionally, I conducted Spearman's rho in order to determine any monotonic correlation resulting in a correlation coefficient to be -0.037.



3. Are explicitly rated songs more popular than songs that are not explicit?

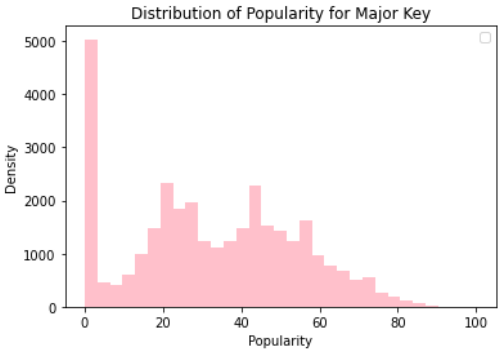
I conducted EDA by using histograms to observe the distribution of popularity scores between explicit songs and not explicit songs. The observed data from both distributions were not normally distributed. It is also important to take note of the type of variables being dealt with, which is categorical data (explicit or not explicit). Due to the data being categorical, I decided to use the Mann-Whitney U test, as this test does not assume a normal distribution and uses the median, as this data cannot be reduced to the mean to properly understand the relationship for nonparametric data. From the 'scipy. stats' library, I used the 'mannwhitneyu' function. Since I am conducting a null hypothesis statistical testing, the null hypothesis was that there is no statistically significant difference between non-explicit and explicit songs. The p-value was 3.06 x 10^-19, so if we hold that at a significance level of 0.05, this is statistically significant. Based on this analysis, explicit songs tend to be more popular than non-explicit songs.

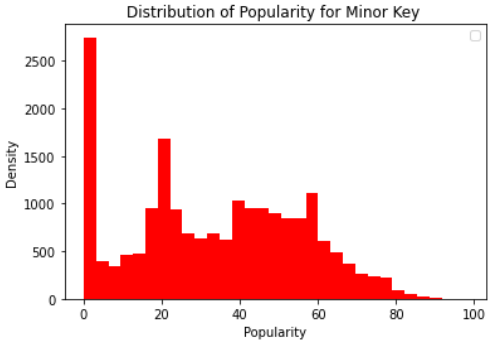




4. Are songs in major key more popular than songs in minor key?

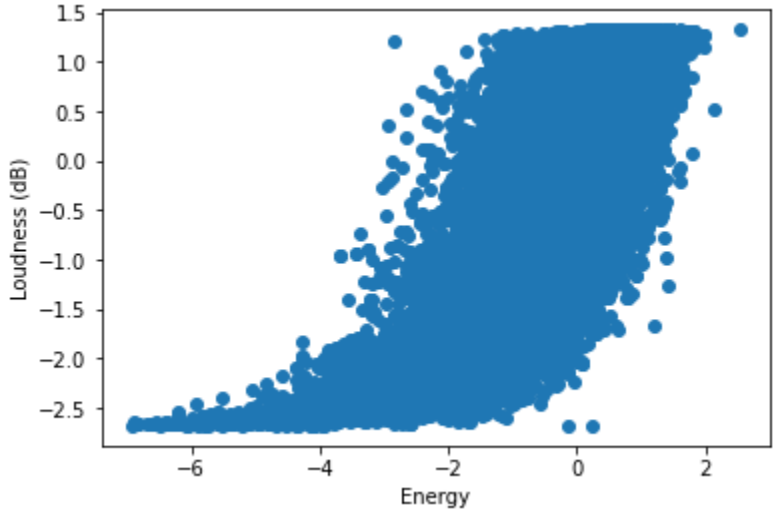
In order to determine if songs are more popular in major key than minor key. Before I began, I created a list for major key and minor key, each holding one value (1 for major, 0 for minor). I observed the distribution between popularity and frequency between both groups. Their distributions were not normal, so I decided to conduct a nonparametric test (U-test) which does not assume normality. Since I am conducting a null hypothesis statistical test, the setup of my null hypothesis was that there was no difference between popularity and the key of the song. To test this hypothesis. Then I conducted a U-test using the 'mannwhitneyu' function and found the resulting p-value to be 2.018 x 10 ^-6. Holding the p-value at the significance level of .05, the results are statistically significant and the null hypothesis can be rejected, indicating there is a difference between popularity in major key and minor key. Comparing their medians it was found that the major key had a median of 0.34 and the minor key had a median of .33. This combined analysis reveals that songs in major key tend to be more popular than songs in minor key.





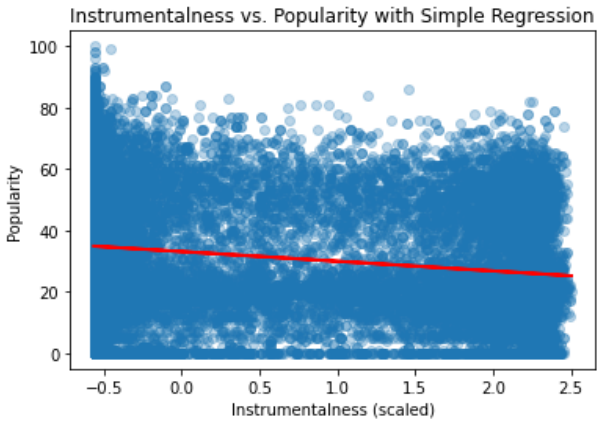
5. Energy is believed to largely reflect the “loudness” of a song. Can you substantiate (or refute) that this is the case?

To observe if energy largely reflects the “loudness” of a song, I first analyzed the data using a scatter plot, which revealed a slight monotonic relationship. Based on this observation, I decided to conduct a Spearman's rank correlation analysis. This resulted in Spearman's rho = .7306, reflecting the strong positive monotonic relationship between energy and “loudness”. The p-value was 0.0, which is statistically significant. Due to a high rho value and the lower p-value compared to the significance level of 0.05, this evidence substantiates that energy largely reflects the “loudness” of a song.



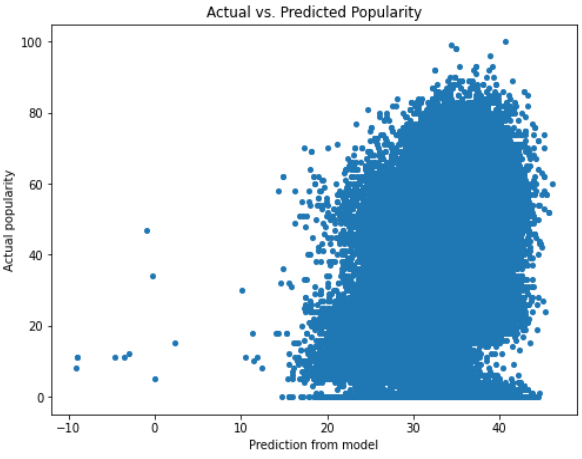
6. Which of the 10 individual (single) song features from question 1 predicts popularity best? How good is this “best” model?

To determine which of the 10 individual(single) song features from question 1 predicts popularity the best, I ran a for loop to iterate through each feature and fit it into a simple linear regression model. From this, all of the predictors showed low results, but instrumentalness seems to be the best predictor in comparison to the other features, with an R^2 squared value of 0.02101, indicating that 2.1% of the variance in popularity can be predicted by instrumentalness. The RMSE of 21.743 indicating the average error, suggests that the predictions the model makes are not accurate. The low R^2 Score and high RMSE suggest this “best” model is not a strong predictor of popularity.



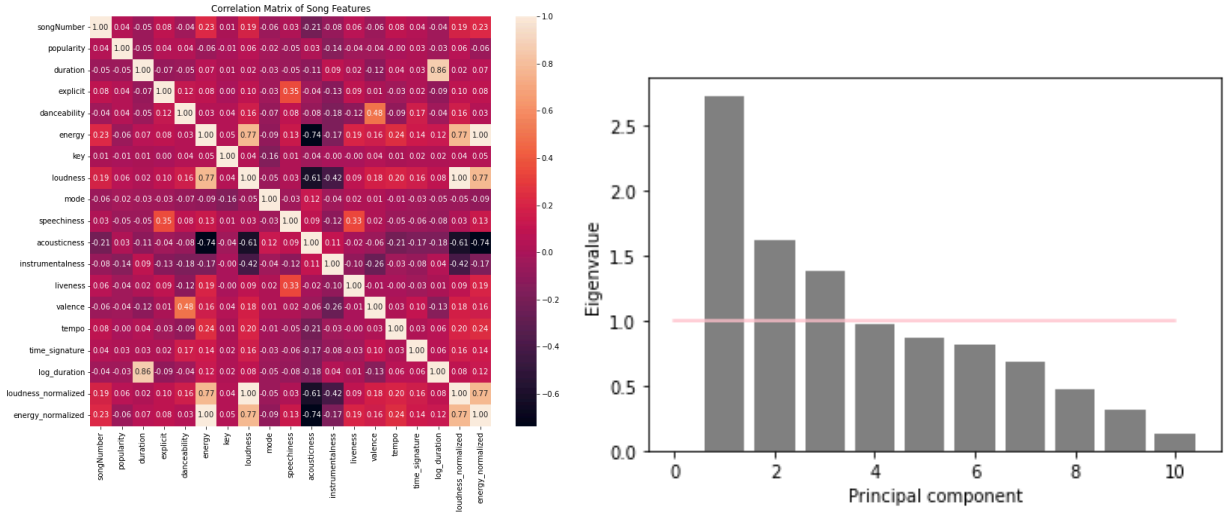
7. Building a model that uses *all* of the song features from question 1, how well can you predict popularity now? How much (if at all) is this model improved compared to the best model in question 6). How do you account for this?

To build a model that uses all of the song features from question 1, I built a multiple linear regression model. First, I began by first scaling the predictors in the model by using a scaler to normalize the data to make sure all features are on the same scale and fit the multiple regression model. Then I found the R^2 value, which resulted in 0.048. This means that 4.8% of the variance can be explained by the model. The RMSE of this model is 21.2187, indicating the model's predictions are not very accurate. This version of the model improved by 2.7% compared to the previous problem. This can be accounted for due to R^2 naturally increasing as more features of the model are added in comparison to a single feature. However, it should be noted that despite this improvement, the R^2 value suggests that this is not a strong model in terms of predicting popularity



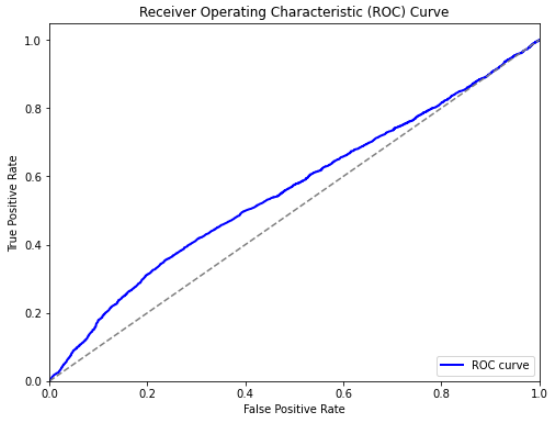
8. When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for?

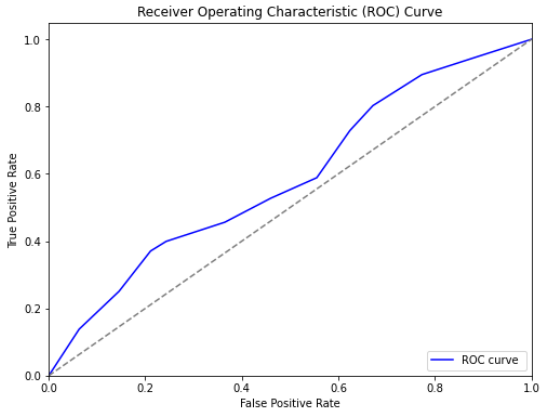
When considering the 10 song features above, I decided to create a heat plot using seaborn to visualize correlations between features. I observed a positive .77 correlation between energy and loudness, a negative 0.74 correlation between energy and acousticness
as well as a -0.61 correlation between loudness and acousticness. Then I decided to conduct a PCA by first standardizing the data using StandardScalar. Then, I fit the PCA with the standardized data and extracted the eigenvalues and loadings, where the eig_values variable contains the eigenvalues that represent the variance accounted by each principal component and loadings representing the direction of variance. I printed to find out how much variance is accounted for by each principal component, with the highest principal component being 27.3 %. I then created a scree plot to observe the principal components. Applying Kaiser's criterion(eigenvalues greater than 1), I identified three meaningful components. These components accounted for 57.359% of the variance.

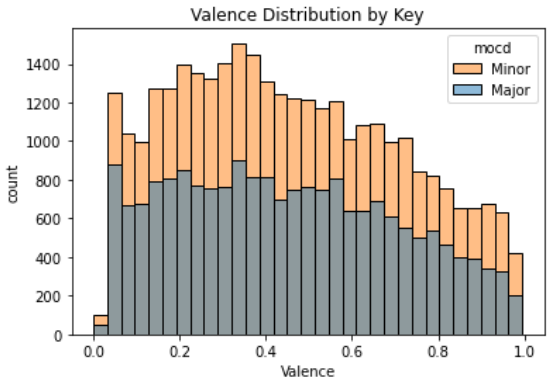


9.) Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?
To predict whether a song is in a major or minor key from valence, I first showed the correlation between valence and mode, which was 0.012, showing a very weak, slight correlation between the two variables. I then proceeded to plot the distribution of mode and major(figure on the bottom left), which revealed no pattern between variables. Then, I proceeded to build a logistic regression model. To avoid overfitting, I split my data. Once I built the model, I found the accuracy score to be 63% and the AUC to be 0.50, which means that the classifier performs no better than random chance (50%). To look for a better predictor, I conducted EDA again and produced a heat plot to see if there were any strong correlations between major or minor keys and another feature. I found a weak negative correlation between key and mode and a slight positive relationship between mode and acousticness of 0.12. I proceeded to conduct logistic regression for both these models and found AUC to be 0.56 (figure on top left) with an accuracy

of 62.7% between acousticness and major or minor key and AUC of 0.58 (figure on top right) with an accuracy of 63.5% between key and major or minor key, which means that the model has low discrimination ability, but may slightly be better than random guessing, but remains to be a better predictor in comparison to valence.

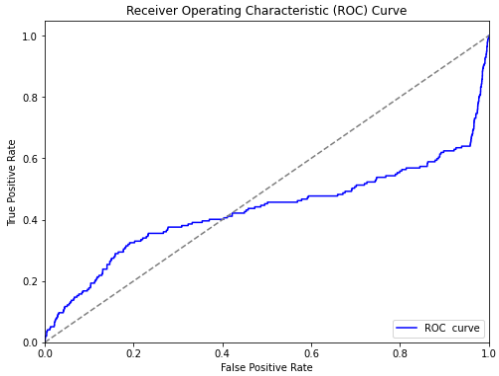


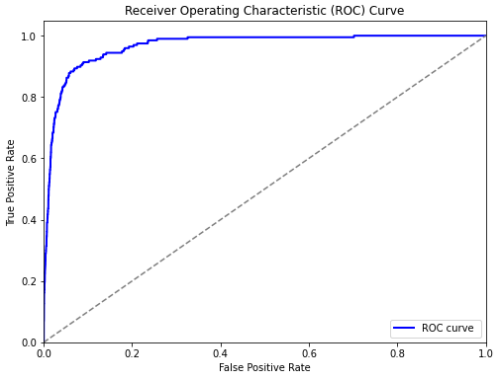




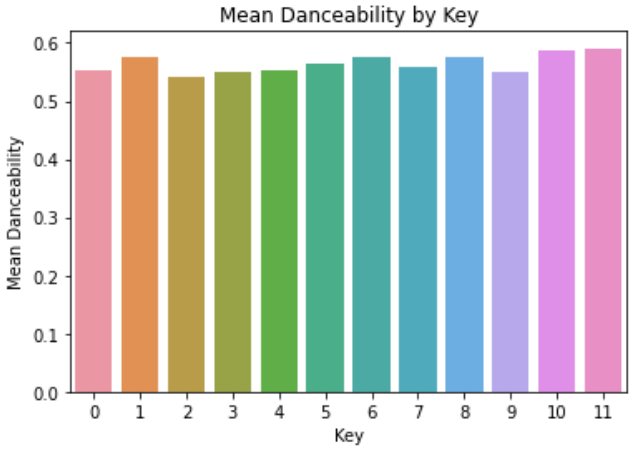
10 Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8?

To determine which is a better predictor of classical music, duration or principal components, I first began by creating another column of binary data, where 1 represented classical music and 0 represented the rest in the track_genre column. Then I used a logistic regression model to predict classical music based on duration, resulting in an AUC score of 0.58 (depicted below on the left), indicating low discrimination. Then, I created a classification model by using a logistic regression model with principal components and classical music, resulting in an AUC score of 0.97 (depicted on the bottom right), suggesting the model is highly accurate in distinguishing models. I then conducted a cross-validation, yielding consistently high accuracy results of around 98.17%. This indicates that the principal components extracted in question 8 are a better predictor in comparison to duration.





11. Is there a statistical difference between danceability and the type of key? **
To approach this question, I first determined the frequency of each key in my dataset. Then I calculated the mean danceability by key to understand how danceability varies between the keys. I visualized the mean of danceability through the usage of a bar plot, which was a uniform distribution. Then I used ANOVA to test my null hypothesis, that being that there was no statistically significant difference between danceability and type of key. I used the f_oneway() function to determine whether there are statistically significant differences in danceability among the different key categories, calculating the F statistic and p-value. With an F statistic of 34.888 and a p value of 3.33 10 ^-75 indicates there are statistically significant differences between danceability and type of key.




```
ANOVA F statistic: 34.888246891600545
ANOVA p value: 3.328454536225205e-75
```