Salma Badr
July 7, 2024

Spotify 52k Songs Analysis

Dimension Reduction: I began by z-scoring the data in each column to assure that my principal components were not influenced by the varying scales of the data. I then performed PCA using the Sklearn package and extracted 3 principal components by the Kaiser criterion.

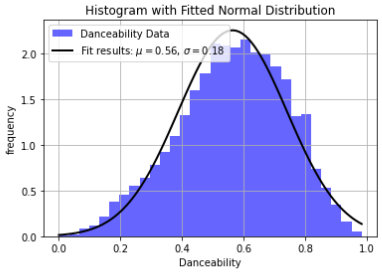
Data Cleaning: I cleaned the data by creating subsets of the data frame and using the drop na function in the Pandas library. I assured that each column in the subset data frames was relevant to the operation I was performing so as to not drop any unnecessary data.

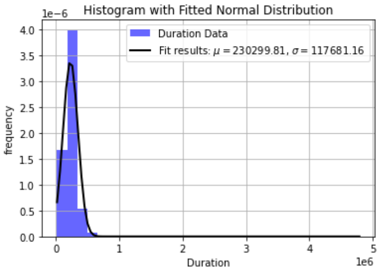
Data Transformations: I z-scored the data before performing PCA. I cast the values for track_genre onto the binary values, 0 and 1, before performing logistic regression.

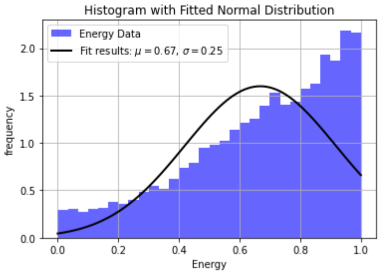
Alpha Threshold: For each significance test I performed in my analysis, I implemented an alpha threshold of 0.005 instead of the conventional threshold of 0.05.

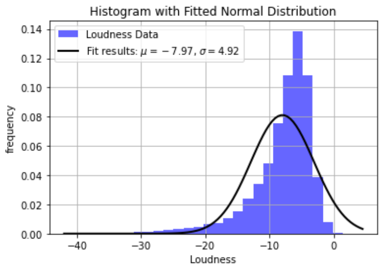
1. Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Are any of these features reasonably distributed normally? If so, which one?

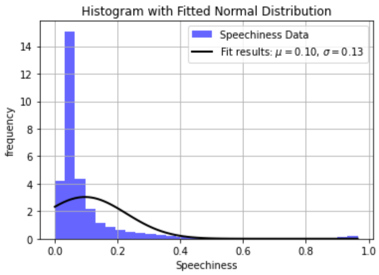
For each feature listed above, I plotted a histogram of the data points with a fit normal distribution line for contrast. None appeared to closely model a normal distribution. However, danceability showed only a slight positive skew – which was notable in comparison to the other plots. Below are the results:

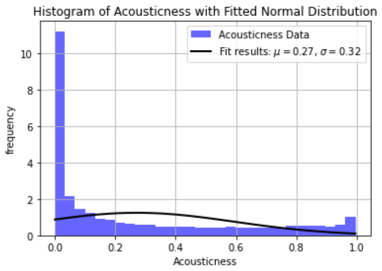


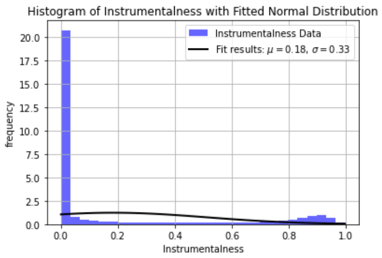


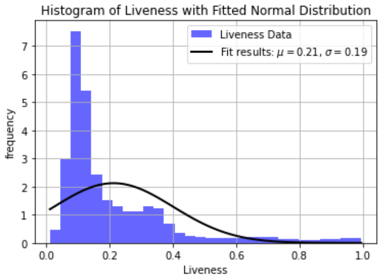


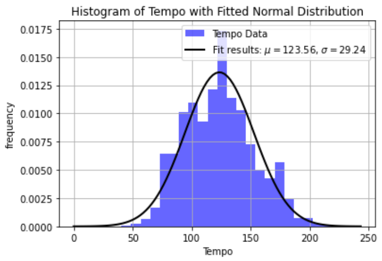

Histogram with Fitted Normal Distribution

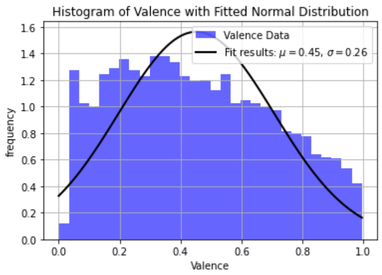

Histogram with Fitted Normal Distribution


Histogram of Acousticness with Fitted Normal Distribution


Histogram of Instrumentalness with Fitted Normal Distribution

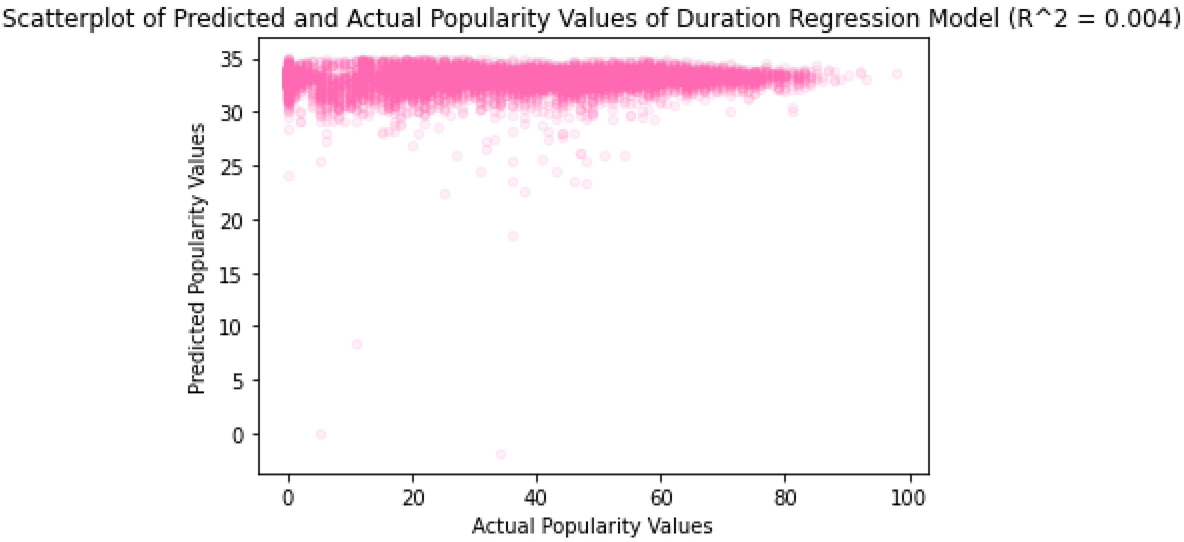

Histogram of Liveness with Fitted Normal Distribution


Histogram of Tempo with Fitted Normal Distribution


Histogram of Valence with Fitted Normal Distribution

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

I ran an Ordinary Least Squares model predicting popularity from duration. The model reported a 99.5% confidence interval for Beta with the bounds [-0.000012, -0.000008] which provided statistically significant evidence of a negative relationship between song length and popularity. The model explains 0.4% of the variance in popularity.



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

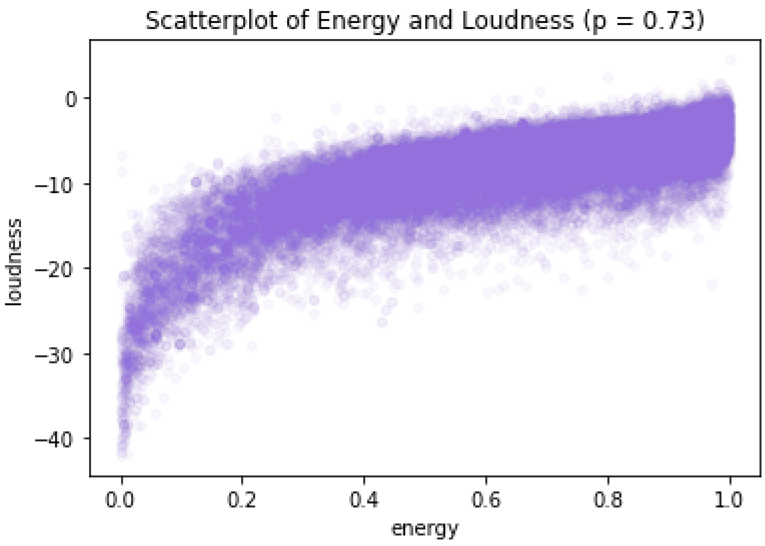
Because I could not assume that popularity is normally distributed, I decided to use a Mann-Whitney U test to evaluate if the difference in median popularity between explicit songs and non-explicit songs is statistically significant. The significance test reported a U-Statistic of 120356317.5 and a corresponding p-value ≈ 0. The median popularity for explicitly rated songs is 34, while the median for non-explicitly rated songs is 33.

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

I used a Mann-Whitney U test again to evaluate if the difference in median popularity between major and minor key songs is statistically significant. The significance test reported a U-Statistic of 309702373 and a corresponding p-value ≈ 0. The median popularity for major key songs is 32, while the median for minor key songs is 34.

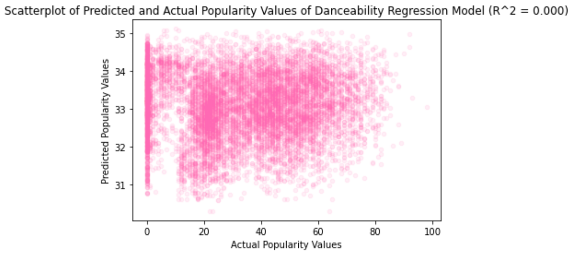
5. Energy is believed to largely reflect the “loudness” of a song. Can you substantiate (or refute) that this is the case?

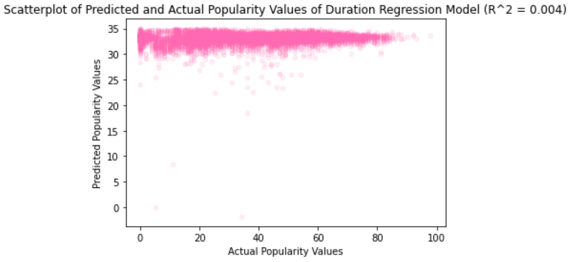
I first performed EDA and observed that the relationship between energy and loudness appeared to be monotonic and non-linear. So I used Spearman’s correlation coefficient to quantify the strength of the relationship. P = .73

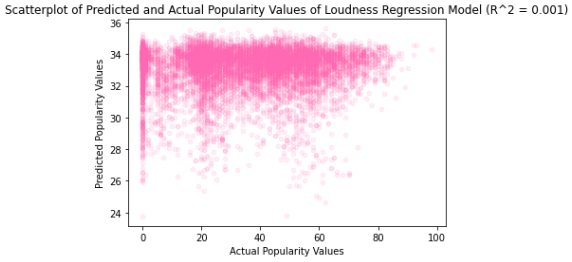


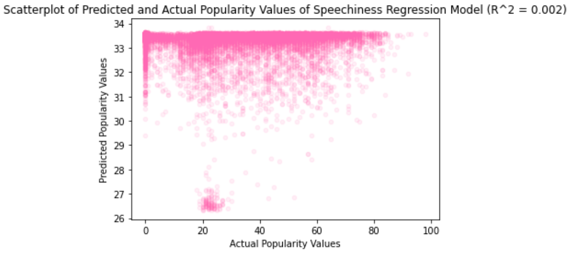
6. Which of the 10 individual (single) song features from question 1 predicts popularity best? How good is this “best” model?

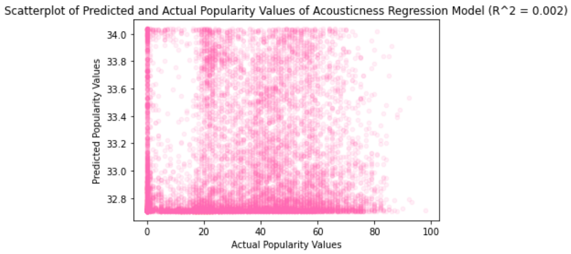
For each song feature, I trained a linear regression model to predict popularity and calculated the model’s R^2. The model with Instrumentalness as a predictor reported the highest R^2 value out of the features, explaining 2% of the variance in popularity. Below are scatterplots of the predicted vs. real popularity values from each model and their respective R^2 values.

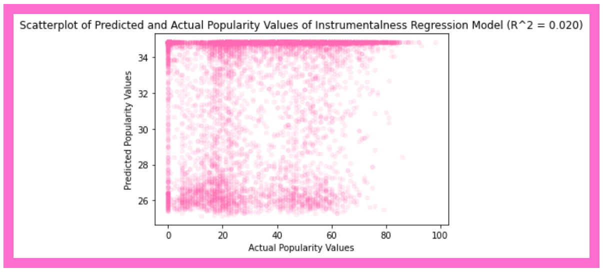


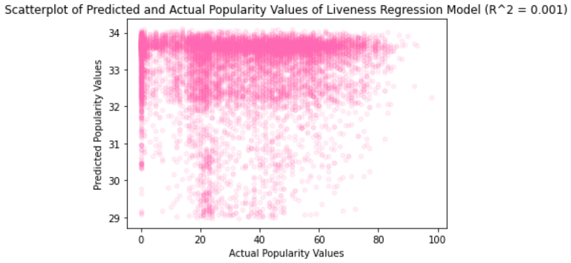


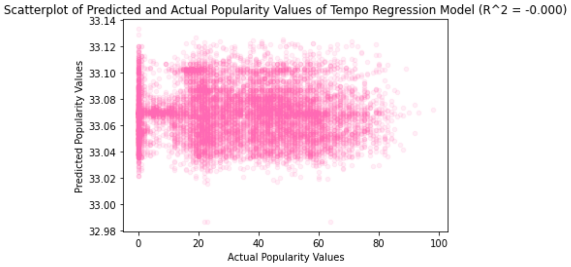


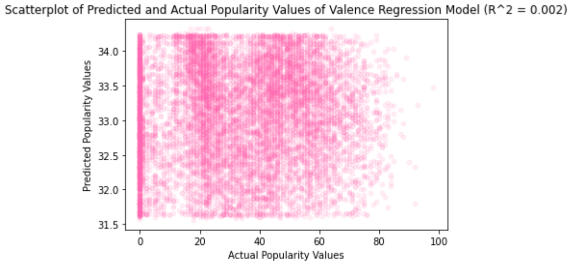


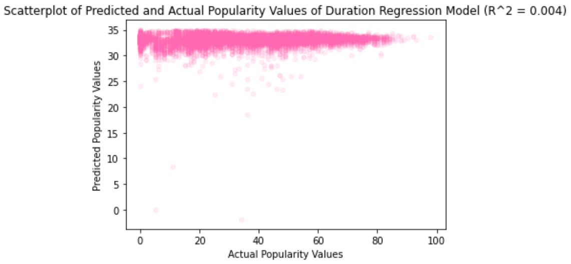






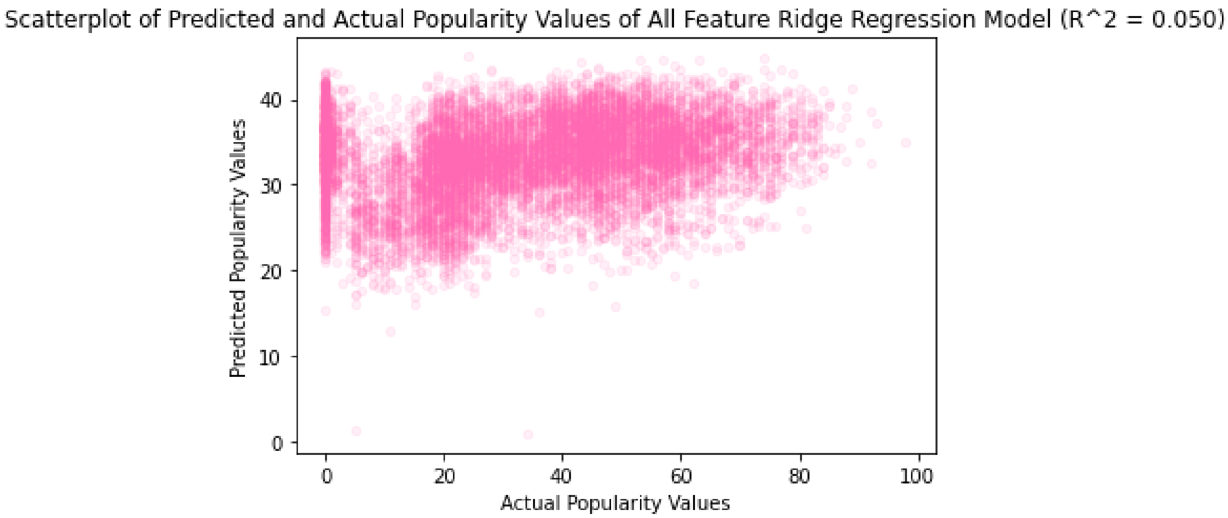


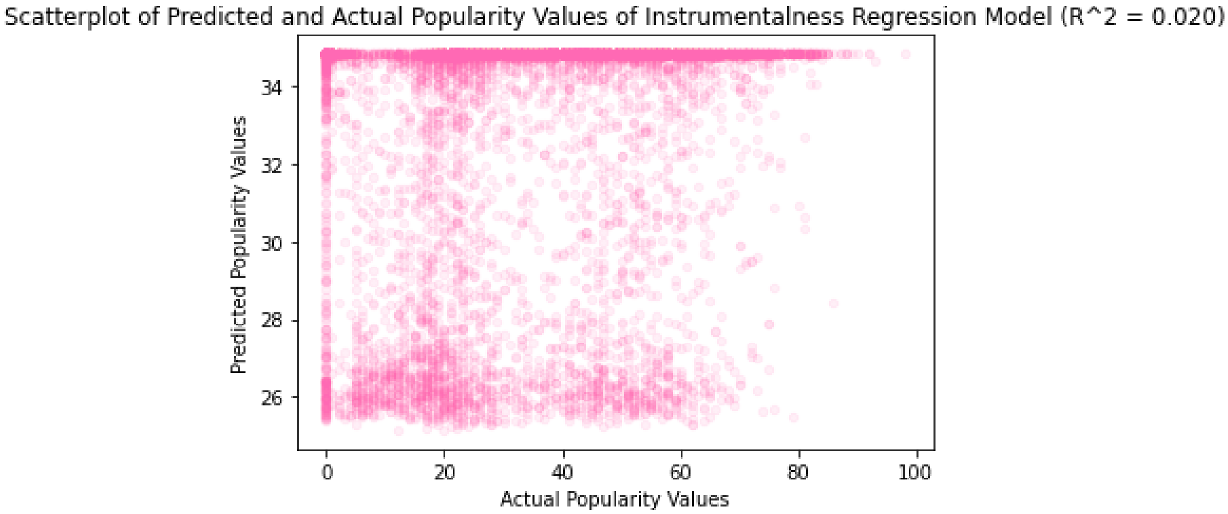




7. Building a model that uses *all* of the song features from question 1, how well can you predict popularity now? How much (if at all) is this model improved compared to the best model in question 6). How do you account for this?

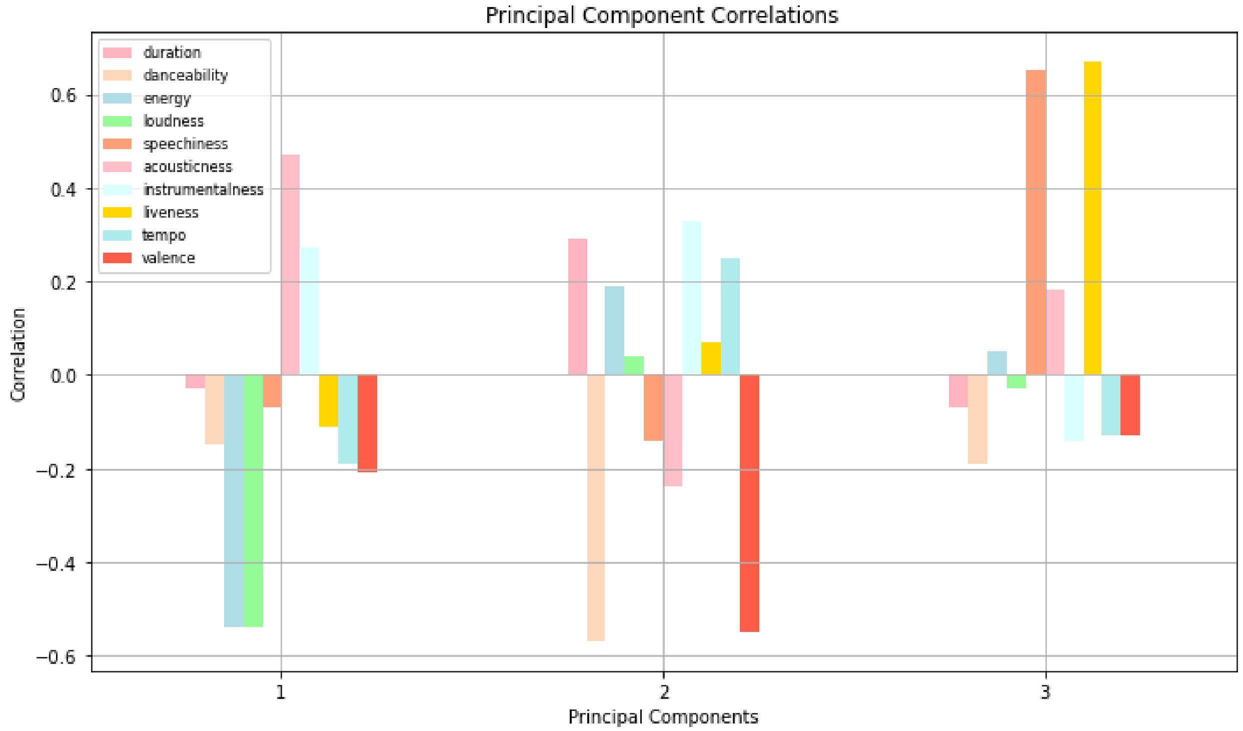
To account for multicollinearity, I trained a Ridge regression model to answer this question rather than a multiple linear regression. The model explains 5% of the variance in popularity, 3% more than the instrumentalness model. This difference is likely due to the higher count of predictors, as a model with more predictors will always be better fit to the data and hence report a higher R^2 even if the predictors themselves are not highly correlated with the outcomes.





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?

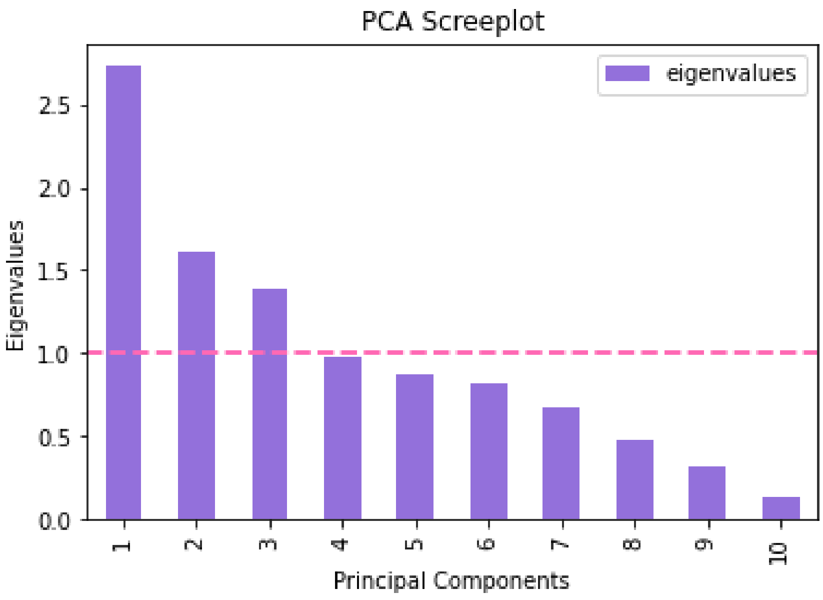
I performed a principal component analysis on the standardized data and extracted 3 principal components by the Kaiser criterion. The 3 principal components explained 6.72% of the variance in the data. Below is a bar plot of each principal component’s respective correlations with the song features.



PC1: PC1 shows a strong positive correlation with acousticness and a strong negative correlation with energy and loudness. My conclusion is that PC1 represents the *mellowness* of a song.

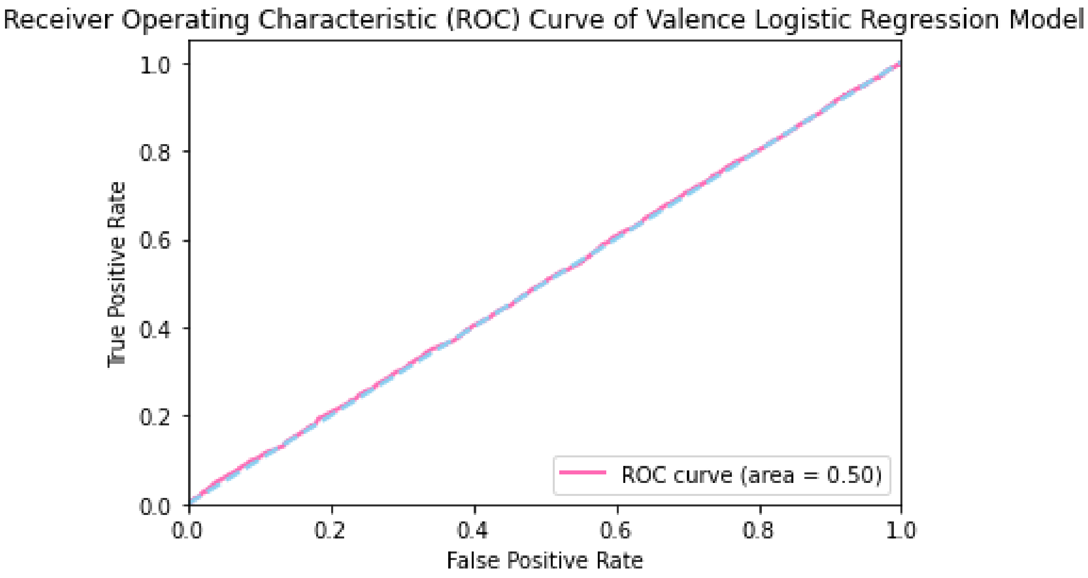
PC2: PC2 shows strong negative correlations with valence and danceability and a positive correlation with acousticness. My conclusion is that PC2 represents the somberness of a song.

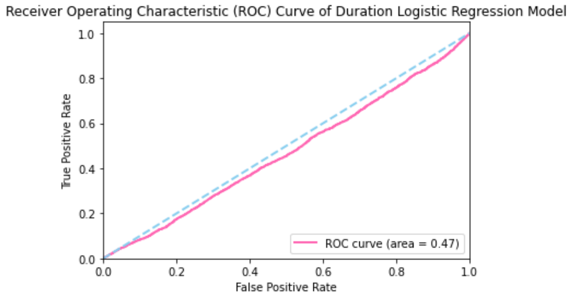
PC3: PC3 shows strong positive correlations with speechiness and liveness. My conclusion is that PC3 represents the oratoriness of a song.

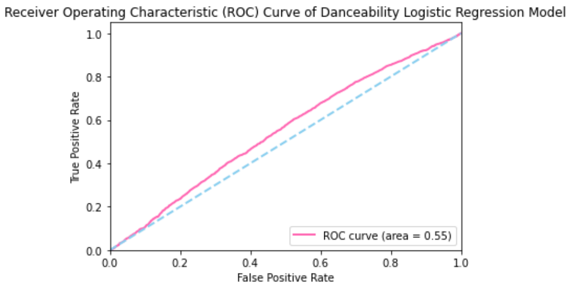


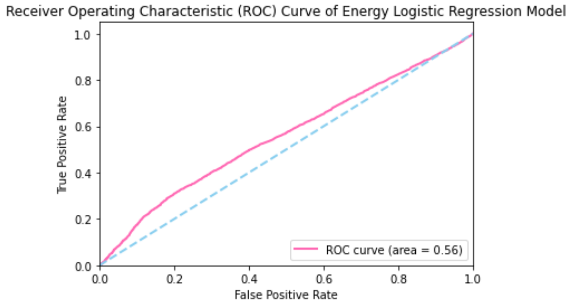
9. Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?

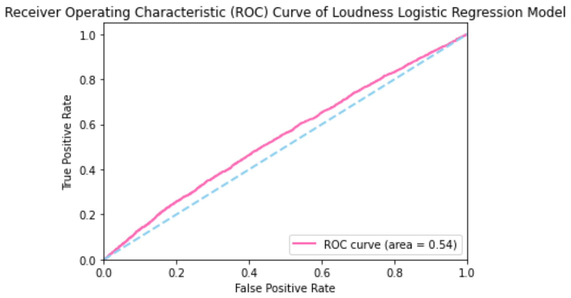
I trained 10 separate logistic regression models with each respective song feature as a predictor and mode as the outcome. I initially analyzed the confusion matrices of each model to measure accuracy, but found that each model could perform no better than predicting the majority reflected in the training set, so I opted to instead use area under the ROC curve as an accuracy metric. The valence model reported a relatively low AUC score of 50.2%. The acousticness model reported the highest AUC score of 57%.

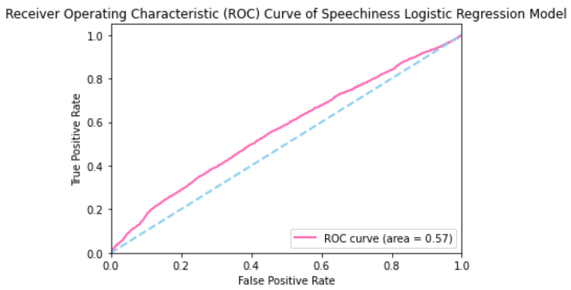

Receiver Operating Characteristic (ROC) Curve of Valence Logistic Regression Model

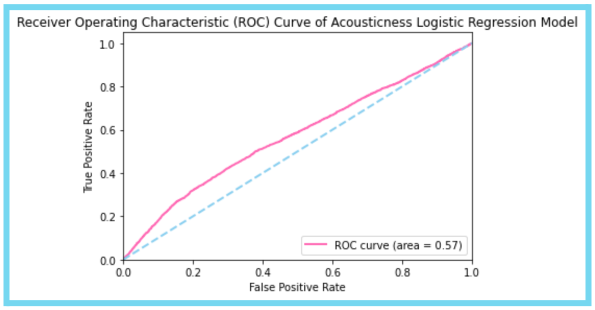

Receiver Operating Characteristic (ROC) Curve of Duration Logistic Regression Model

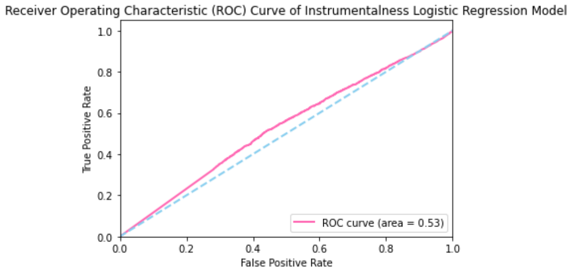

Receiver Operating Characteristic (ROC) Curve of Danceability Logistic Regression Model

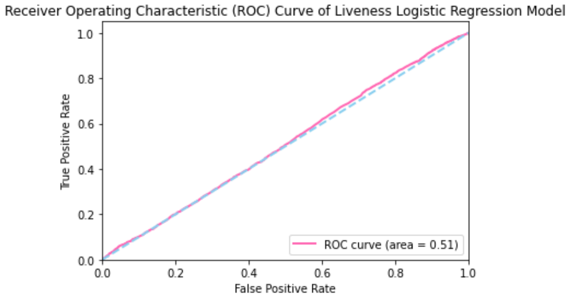

Receiver Operating Characteristic (ROC) Curve of Energy Logistic Regression Model

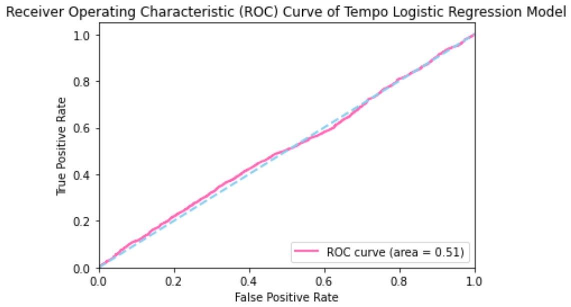

Receiver Operating Characteristic (ROC) Curve of Loudness Logistic Regression Model





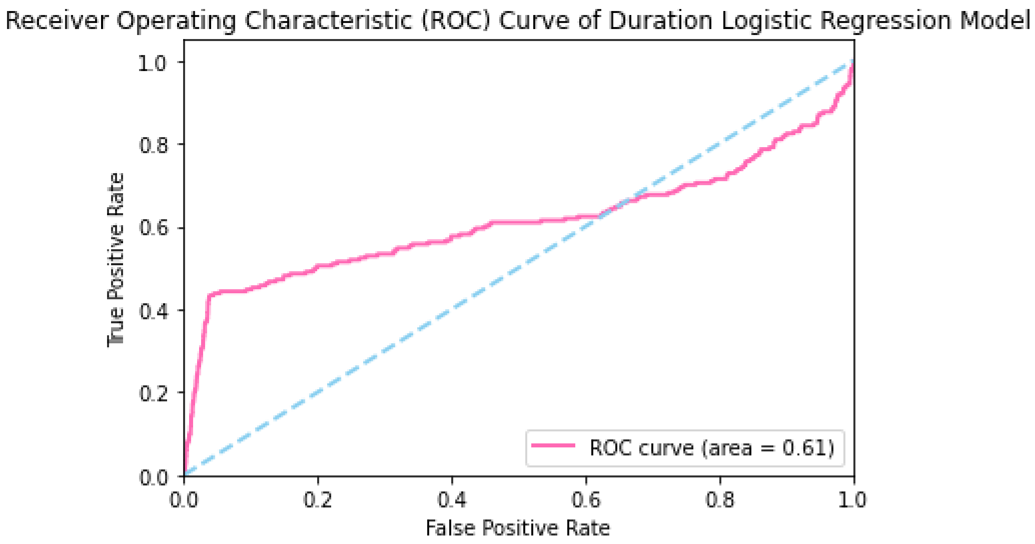


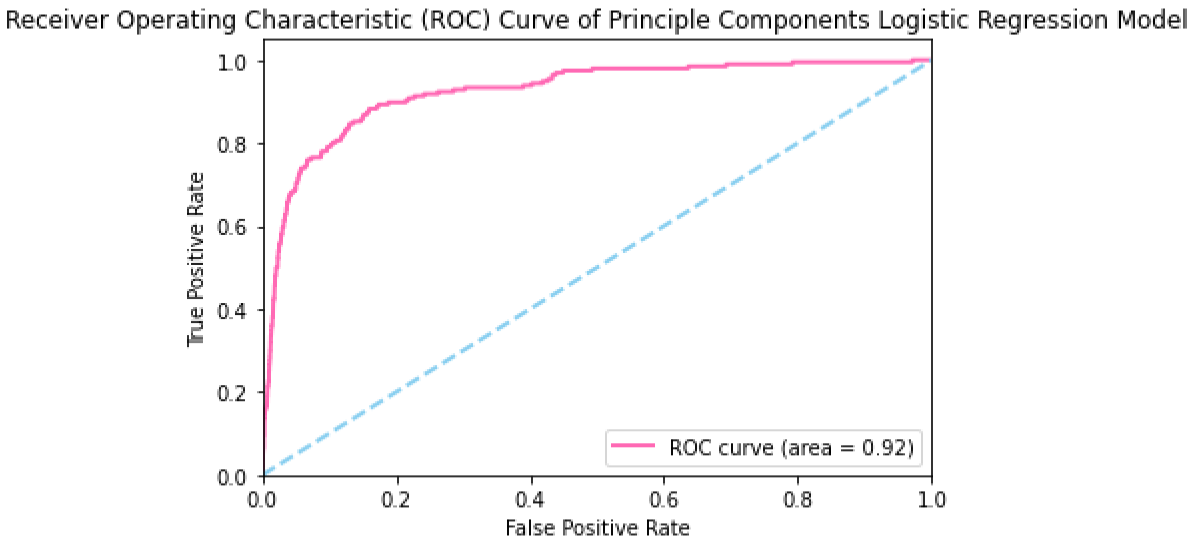




10. Which is a better predictor of whether a song is classical music –duration or the principal components you extracted in question 8?

I first transformed the variable 'track_genre' into binary values of 0 and 1 indicating if the song is classical or not. I then trained 2 logistic regression models, one using duration and the other using the 3 principal components extracted in question 8. Using AUC as an accuracy metric, I determined that the principal components model is a better predictor of whether a song is classical. The duration model reported an AUC of 61% while the model trained by the principal components reported an AUC of 92.3%.

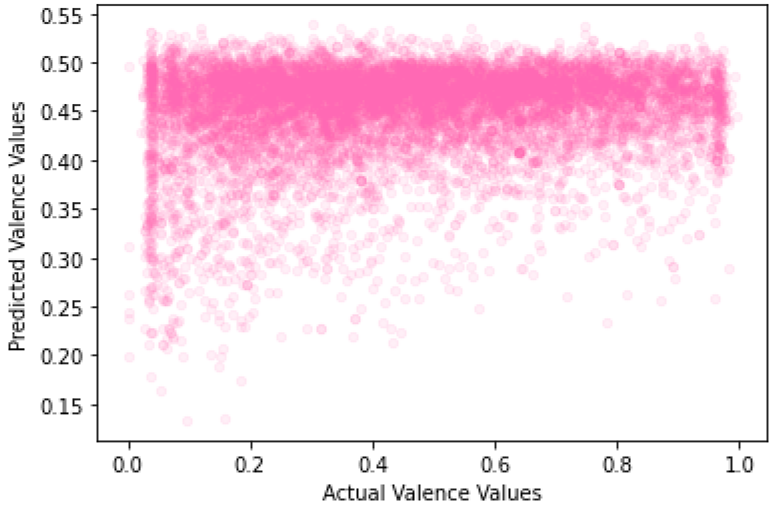




Extra Credit:

I ran an Ordinary Least Squares model predicting valence from loudness which reported a 99.5% confidence interval with the bounds [0.008823 0.010079]. The model explains 3.3% of the variance in valence, providing statistically significant evidence that loudness is positively associated with a song's ability to uplift people.

Scatterplot of Predicted and Actual Valence Values of Loudness Regression Model (R^2 = 0.026)


Predicted Valence Values
Actual Valence Values