Katie Carlson

Professor Pascal Wallisch

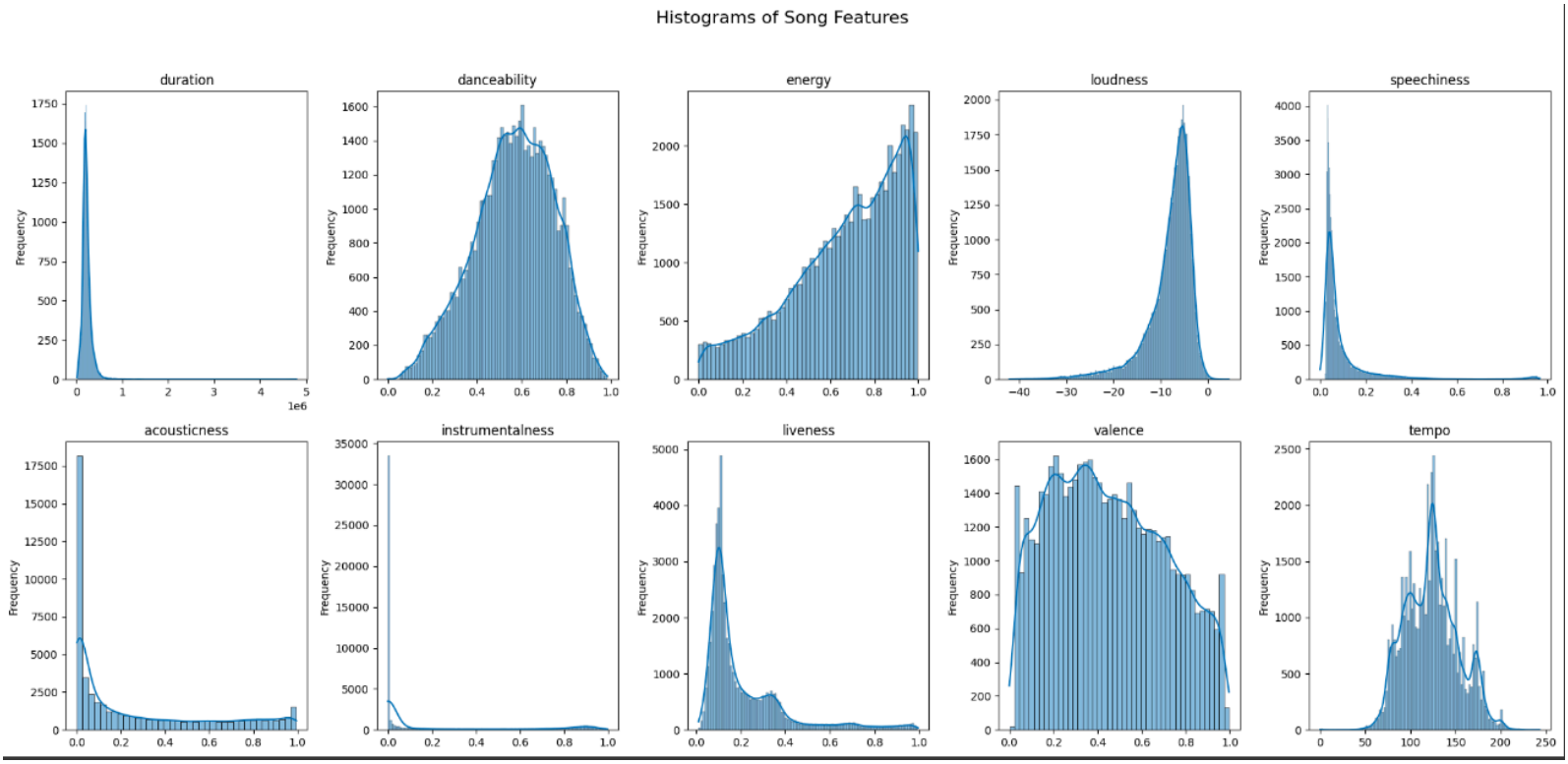
Principles of Data Science

14 May 2024

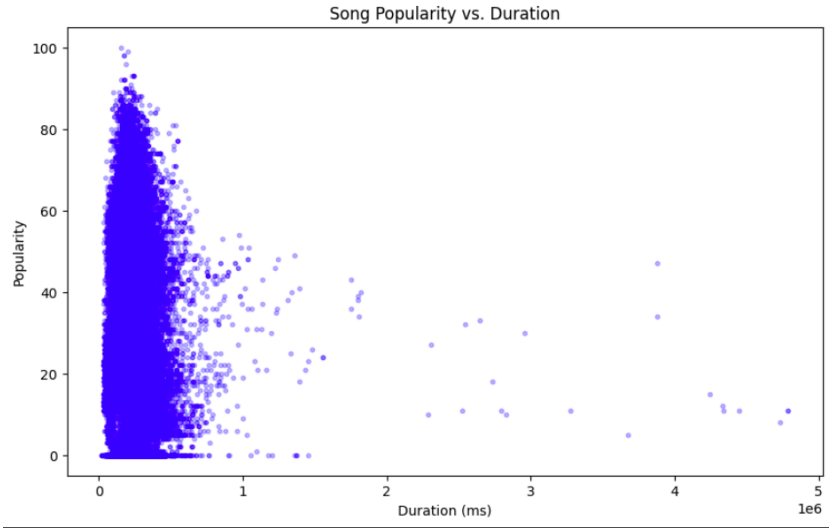
Capstone Project

For this project, I utilized the Pandas, Numpy, Seaborn, and Sci-Kit Learn libraries in Python. I loaded the data by reading the CSV file using Pandas. I removed any potential missing values using the dropna method and then proceeded. I also flattened the array inputs before making histograms. Before performing a PCA, I scaled the data with a standard scalar function. I seeded the RNG by setting the random-state argument of the sklearn model selection method to my N-number, N16594496.

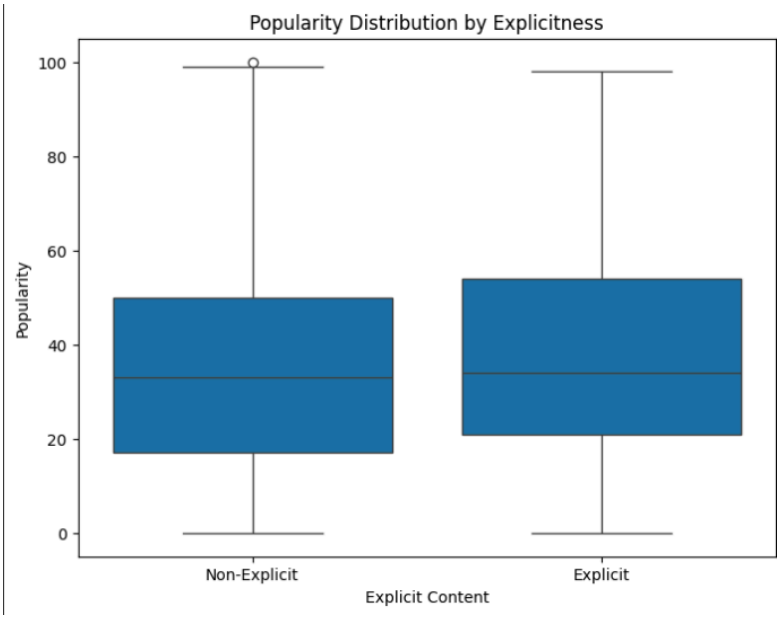
1. I plotted a histogram of the distribution across the data set for each song feature and visually analyzed the histograms for potential normality. Out of the 10 song features, only the "tempo" is reasonably normally distributed. The others appear to be skewed in various ways. The histograms are presented below.



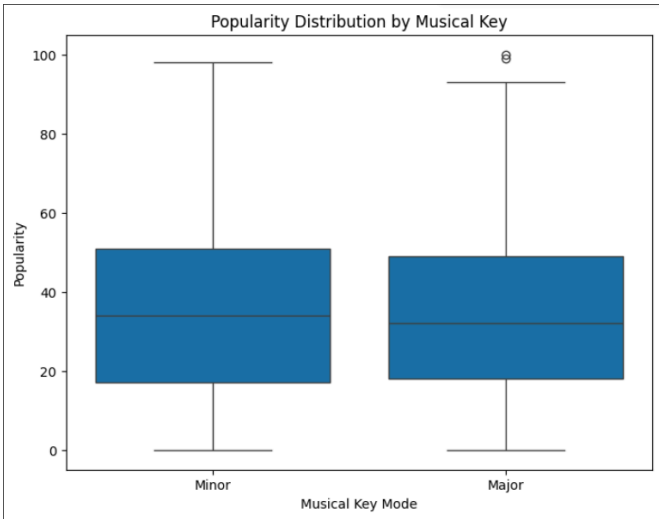
2. I plotted the popularity of each song against its duration as a scatterplot and visually inspected to see if there was a clear relationship between duration and popularity. As evident by the scatterplot, there is almost every possible value for popularity given the same duration. There appears to be no relationship between a song's length and its popularity. I deemed further analysis unnecessary.



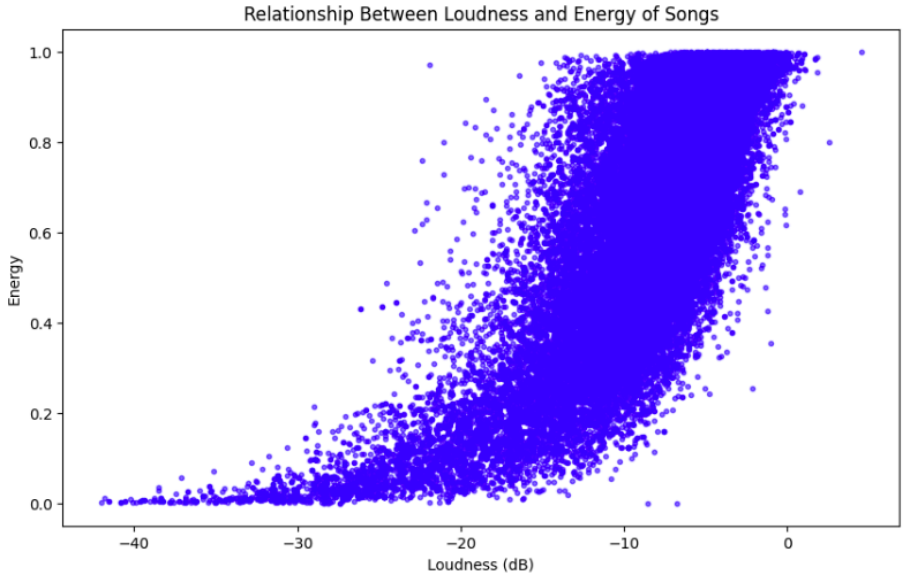
3. To determine whether explicitly rated songs are more popular than songs that are not explicitly–note: a visual inspection of the distribution of popularity indicates that the median for the explicit songs is *slightly* higher–I performed the Mann-Whitney U test because the data does not appear to be normally distributed. The test statistic = 139361273.5 The resultant p-value = 3.07e-19, which is very small. Setting the conventional alpha level to 0.05, this indicates that there is a statistically significant difference in the popularity of a song based on whether or not it is explicitly rated. I also plotted boxplots for the distributions of explicit and non-explicit songs side-by-side for visual inspection and comparison. The box plots are displayed below.



4. To determine whether major key songs are more popular than minor key songs, I performed the Mann-Whitney U test as well, a non-parametric test that does not assume normal distribution in the data. The test statistic = 309702373.0 and the resultant p-value = 2.018e-6, which is very small. Setting the conventional alpha level to 0.05, this indicates that there is a statistically significant difference in the popularity of a song based on the major or minor key. I also plotted boxplots for the distributions of major key and minor key songs side-by-side for visual inspection and comparison. The boxplots are displayed below.

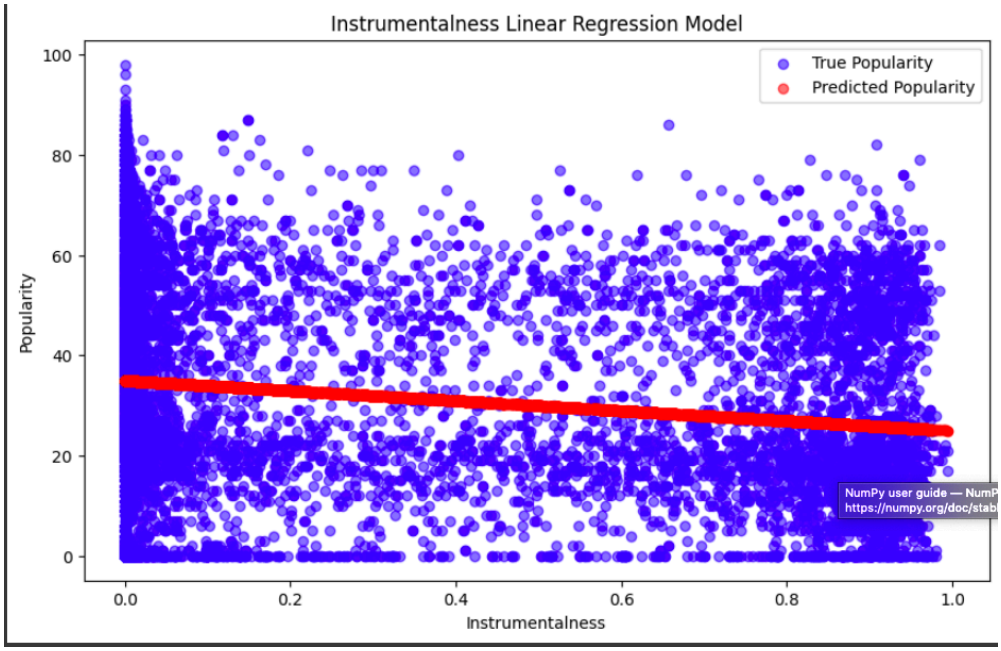


5. To substantiate or refute the claim that energy is believed to largely reflect the "loudness" of a song I created a scatterplot of loudness and the energy of each song in the data set. Upon visual inspection, the scatterplot suggests a positive correlation between the "loudness" and "energy" of a song. As the song gets louder, the energy tends to increase. The Pearson Correlation coefficient between "loudness" and "energy" = 0.775, indicating a strong, positive correlation. This substantiates the claim that the energy tends to reflect the "loudness" of a song. Since there does clearly exist a relationship between the two, the energy does "reflect" the loudness of a song.

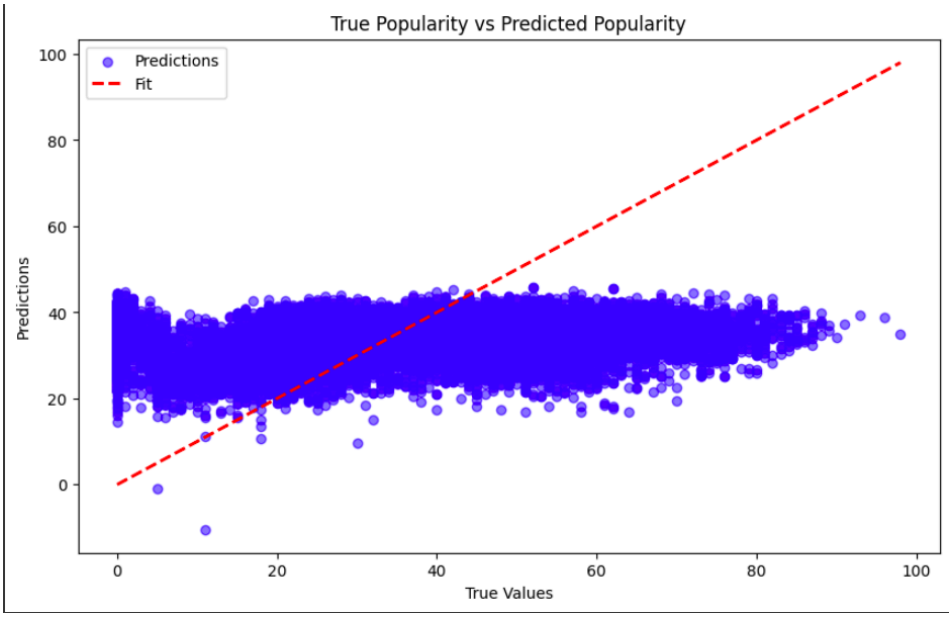


6. To determine which of the 10 song features best predicts the popularity of a song, I performed a simple linear regression for each song feature using the popularity as a response variable. Based on the R-squared value for each regression model, “Instrumentalness” appears to be the best predictor of popularity when compared against the other R-squared values; however, it still only explains about 1.87% of the variance in popularity. Overall, all of the models display a very weak predictive power. All R-squared values are displayed in the graphic below. Beneath the R-squared values is a plot of the “instrumentalness” model. I isolated the model for just “instrumentalness” and created a scatterplot of the true values in comparison to the predicted values for popularity based on the linear regression model.

```
{'duration': 0.0023361049392560673,
 'danceability': 0.0015540973100682809,
 'energy': 0.0034954514352990573,
 'loudness': 0.0016897087518026321,
 'speechiness': 0.0023211021272008248,
 'acousticness': 0.0006776756600284095,
 'instrumentalness': 0.017177996887789 98,
 'liveness': 0.0018870410280088512,
 'valence': 0.001513821091213341,
 'tempo': -7.585912446783638e-05}
```



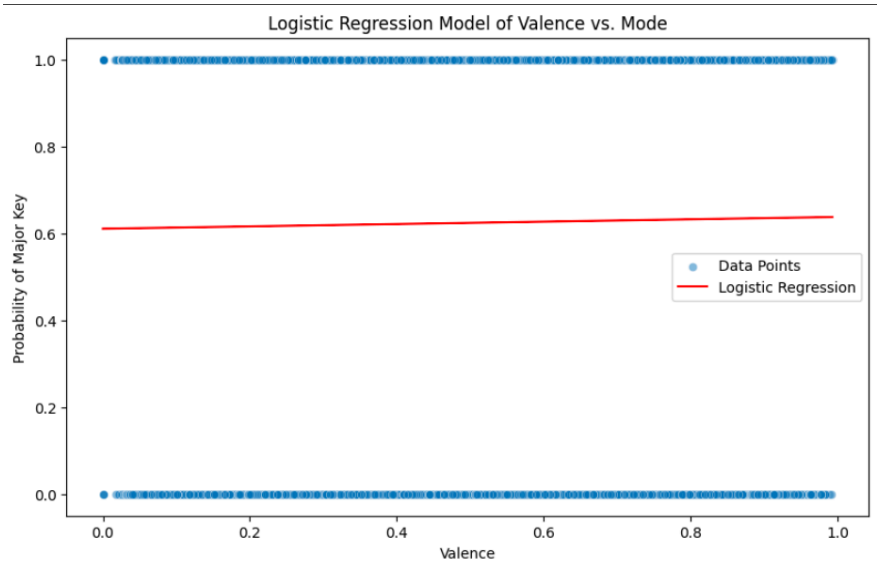
7. I used a multiple linear regression model to use all the song features to predict popularity. The R-squared value for this model, 0.0418 indicates that about 4.2% of the variance in song popularity can be explained by all of the features. This is still a weak predictive power; however, it is significantly greater than the best predictive model out of each of the individual features. Visually it also appears to have weak predicting power. In comparison to the model for instrumentalness, this model has improved by 225%. Since this is the case, it suggests that other features play a role in determining a song's popularity. Graphed below is the true value for song popularity compared against the predicted value for a song's popularity score. The red fit line displays the shape of the curve where the predicted values match the true values.



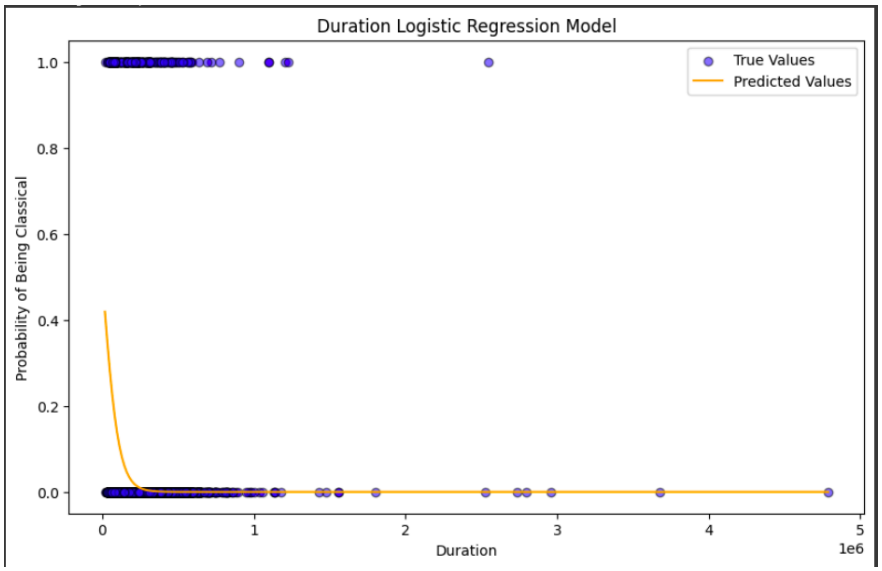
8. When considering the 10 song features, I was able to extract 7 meaningful principal components. I set my threshold so the components are able to explain at least 95% of the variance in the data. Component 1 explains roughly 36.7% of the variance in the data, the most out of the components. Based on the explained ratios for each song feature, Component 1 likely differentiates loud and energetic songs from instrumental and acoustic songs. These proportion of variance in the data explained through each principal component is described by the table below:

Component:	1	2	3	4	5	6	7
Explained Variance Ratio:	0.3670	0.2427	0.1422	0.0934	0.0504	0.0419	0.0283

9. Using a logistic regression model it is possible to predict whether a song is in a major or minor key (quantified by Spotify as the "mode") based on the valence; however, the model only accurately predicts the key about 62.31% of the time. Furthermore, the model's resultant ROC-AUC score = 0.505, which indicates that valence is not a strong predictor in determining the key. The graph below displays the predictions of this logistic regression model next to the true values of the valence and mode.



10. To prepare the data, I converted the genre label to a binary numerical of classical (1) or not classical/everything else (0). Then, I built two logistic regression models, one for the duration and one for the principal components to compare against each other's predictive capabilities. The duration model was highly accurate, able to successfully predict the classical or not classical music genre around 98.15% of the time. The principal components model was also quite accurate, able to predict the classical or not classical genre around 98.21% of the time. As these two accuracies are similar, I then produced the ROC-AUC scores for each model. Although the duration model was highly accurate, its ROC-AUC score= 0.5690, indicating a significantly weaker ability to distinguish between the classes compared to the principal components model, which had a score of 0.9677. This evidence suggests that the principal components are a better predictor of whether or not a song is classical music.


Duration Logistic Regression Model
True Values
Predicted Values
Probability of Being Classical
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
Duration
1e6