Isabella Wrobleski

Pascal Wallisch

Principles of Data Science

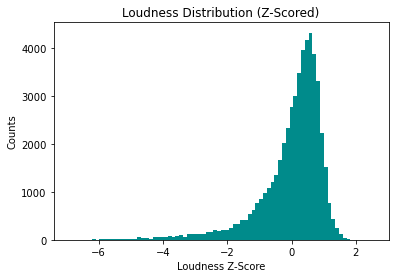
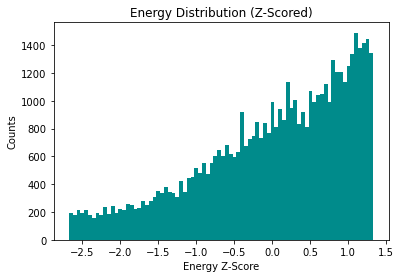
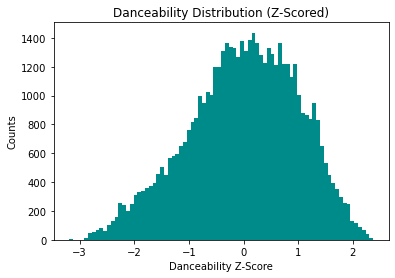
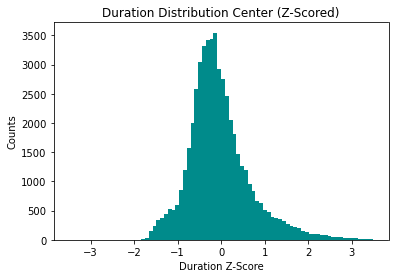
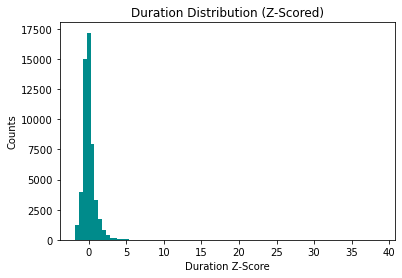
December 19, 2023

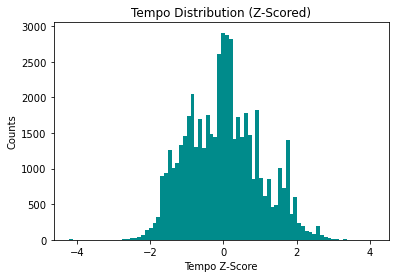
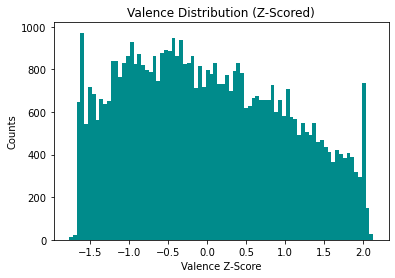
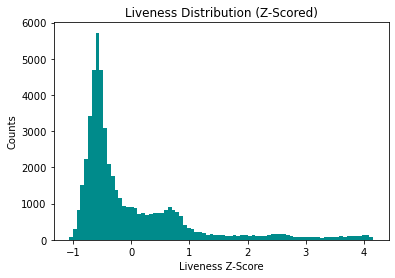
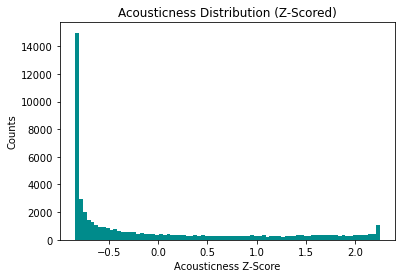
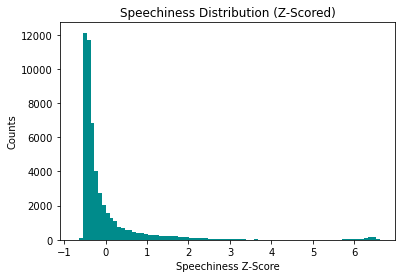
Data Loading and Pre-Processing

To analyze this dataset, I used both numpy and pandas. I was unable to load the dataset directly into numpy, so I loaded it as a pandas dataframe and then converted it to a numpy array. I also kept the pandas version of the dataframe to use for EDA as it had labeled columns. I then created separate columns for features that I would need to use later, as well as z-scored versions of those columns, and since the numpy array was an object, I converted them to their appropriate data types. Before beginning the project, I also did some exploratory data analysis. First, I computed means for the numerical data and modes for the categorical data to get a better understanding of it. I also used the pandas dataframe to get value counts for certain columns such as the genres. I also looked through the data to try to find songs that I knew, so I could get an idea of what certain values for a feature meant. For example, I found that most of the songs I knew in the dataset had a speechiness score of around 0.1, which I expected to be higher. I also looked at the popularity distributions to get an idea of how popular songs I knew were.

Question 1

To determine if any of the ten features were normally distributed, I created a histogram for each one, using the z-scored data columns so that it would be easier to compare the distributions to one another. I used the entire data column for each histogram, but I created two histograms for the duration data, since it included some extreme values and I wanted to look more closely at the distribution of the bulk of the data.

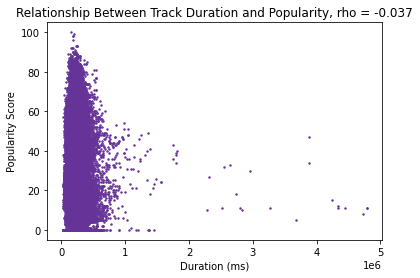




Based on the histograms above, duration, danceability, and loudness seem to be reasonably normally distributed. They are not perfectly normally distributed, as they are all a bit skewed, but in comparison to the other features, they are the most normal distributions. The distribution for danceability is also a bit more dispersed than those for loudness and duration, which makes sense as most songs are around the same length and volume, while danceability varies greatly between songs.

Question 2

In order to determine whether there is a relationship between song length and popularity, I calculated the Spearman correlation coefficient between duration and popularity. I used the Spearman coefficient as it can indicate any monotonic relationship, as opposed to the Pearson coefficient, which only indicates linear relationships.

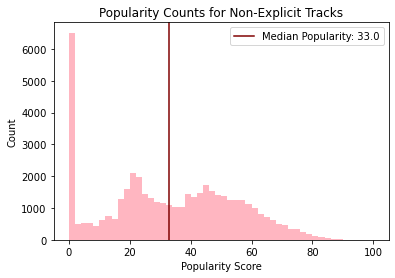
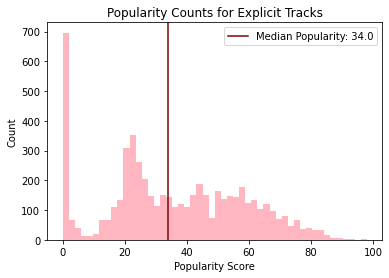


The scatterplot above shows the relationship between track duration and popularity score. The value of ρ was -0.037, indicating that there is an extremely weak negative monotonic relationship between duration and popularity. This is likely due to the fact that the majority of the tracks are in the range of 0 to about 500,000 ms (about 8 minutes). However, it does look like there is a more deterministic relationship after the tracks reach a duration of around 1,000,000 ms. So, perhaps duration does not have any relationship with popularity until a certain duration is reached, after which the duration and the popularity become correlated.

Question 3

My hull hypothesis for this test was that there is no difference in popularity between songs that are explicit and songs that are non-explicit. The alternative hypothesis was that songs that are explicit are more popular than songs that are clean. I ran a one-tailed Mann-Whitney U test, since my EDA indicated that a large proportion of tracks have a popularity score of 0, so I felt that it would not be reasonable to reduce the popularity of each group to a mean. They are also rankings, so the “distance” between each popularity score may not be uniform, depending on the range of plays Spotify corresponds to each score.

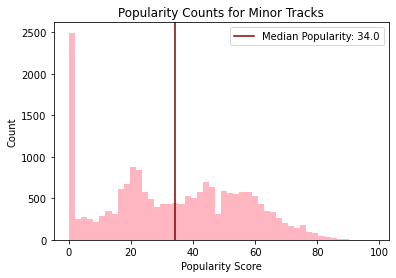
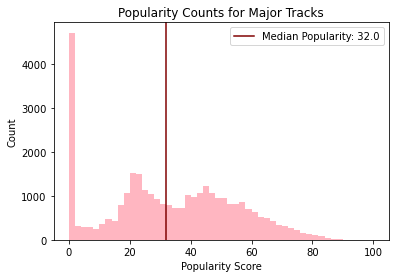
After running the test, the value of the empirical test statistic, u, was 139361273.5 and the p value was 1.53 × 10-19 which is less than the chosen alpha of 0.05, so the test was statistically significant. Therefore, I was able to drop the assumption that the null hypothesis is true and conclude that explicit songs are most likely more popular than non-explicit songs. After running the U test, I wanted to know how large the difference in median rating was. I plotted histograms for the popularities of the explicit and non-explicit tracks and computed the effect size. The effect size was about 0.14, and the histograms indicate that the difference in median is only 1 popularity point. So, the effect is not very large, and there is not a large difference in median popularity between explicit and non-explicit tracks even though the difference is statistically significant.



Question 4

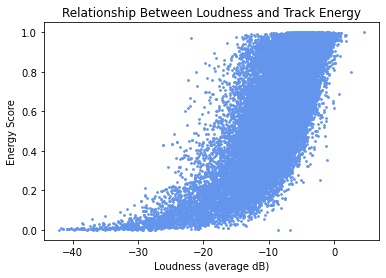
For this test, my null hypothesis was that there is no difference in popularity between songs that are in a major key and songs that are in a minor key. The alternative hypothesis was that songs that are in a major key are more popular than songs that are in a minor key. I ran a one-tailed Mann-Whitney U test, for the same reasons as the previous question since the outcome variable is the same. The value of U for this test was 309702373.0 and the p value was 0.99 which is greater than the chosen alpha of 0.05, so the test was not significant, and I was unable to drop the assumption that the null hypothesis is true.

Because the p value was so high in this test, I wanted to run a one-tailed U test in the other direction. In this case, the alternative hypothesis was that songs in a minor key are more popular than songs in a major key. The U value for this test was 325452746.0 and the p value was 1.009 × 10-6  which is less than 0.05, so this test was significant, and I was able to drop the assumption that this null hypothesis is true and conclude that songs in a minor key are actually more likely to be more popular than songs in a major key. I also plotted histograms for the popularities of the major and minor tracks and computed the effect size. The histograms indicate that the difference in median is 2 popularity points, and the effect size is about 0.04. So, the effect is very small, and the median difference in popularity between major and minor tracks is not large even though it is statistically significant.



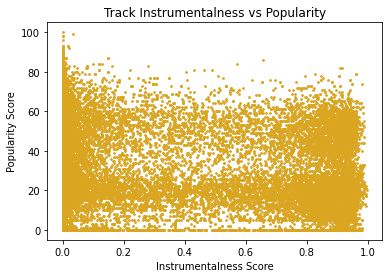
Question 5

To determine whether energy reflects the loudness of a song, I first calculated correlation coefficients to see if there is a relationship between them. I calculated the Spearman coefficient first, and the value of ρ was 0.73, indicating a positive monotonic relationship between loudness and energy. I wanted to determine whether that relationship is linear, so I then calculated the Pearson coefficient, and the value of r was 0.77, indicating that the relationship is positive and linear. Based on the Pearson correlation coefficient, I calculated R2 to be 0.60, indicating that loudness accounts for 60% of the variability in energy, which is the majority of the variance. So, the energy of a track does largely reflect the loudness of the track. I also created a scatter plot of loudness and energy in order to visualize the relationship, and found that while it looks more exponential than linear, there is a clear positive correlation between loudness and energy.



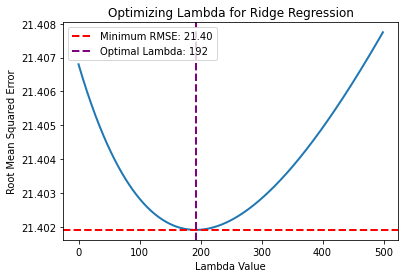
Question 6

In order to determine which of the ten song features best predicts popularity, I fit a linear regression model to each feature. In order to avoid overfitting, I split the data into training and test sets; 80% of the data was used to create the model, and 20% was used to evaluate it. 20% of the data points was about 10,000 points, which I felt would be sufficient to evaluate the models. For each model, I computed the R2 value, which is the proportion of popularity variance accounted for by the predictor. The highest R2 value was 0.019, which was the R2 value for instrumentalness, indicating that instrumentalness is the best predictor of popularity out of these ten, even though it only accounts for about 2% of the variance in popularity. I also created a scatter plot to visualize the relationship, and found that there is no clear visible correlation between instrumentalness and popularity, though there look to be two bands of points.



Question 7

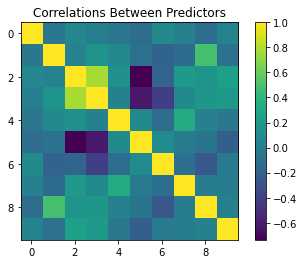
To create a model that used all ten features to predict popularity, I used a multiple regression model. To avoid overfitting, I used a ridge regression model, since a LASSO model did not seem necessary based on the low R2 values for each predictor. I also used the same 80/20 train/test split as above. I ran a loop that fit ridge regression models with a range of lambda values from 0 to 500, and used cross validation to evaluate each model by computing the root mean squared error. I then plotted each lambda value and its corresponding root mean squared error to determine the optimal lambda.



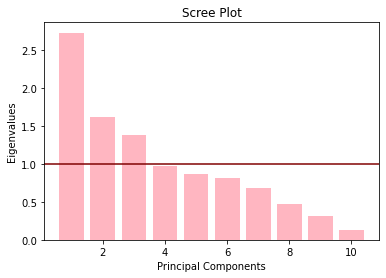
As shown in the plot above, the optimal lambda value for this model was 192. I then evaluated that model using the test set, and found that the R2 value was 0.039, which was about twice the value of R2 for the model in question 6. So, although this model is an improvement on the model in question 6, it is still not very good, especially for having ten predictors. This is likely due to the fact that each of the ten predictors account for less than 2% of the variance in popularity, as well as the fact that a penalty term was added with the ridge regression model.

Question 8

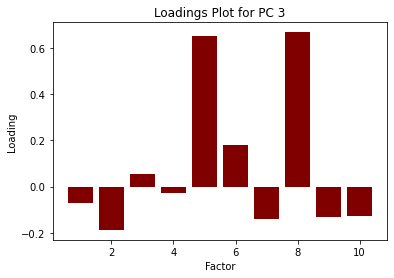
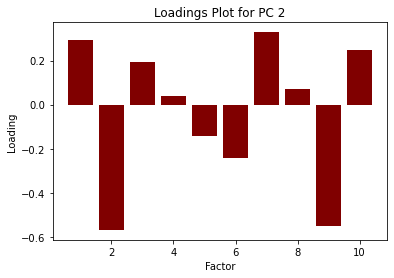
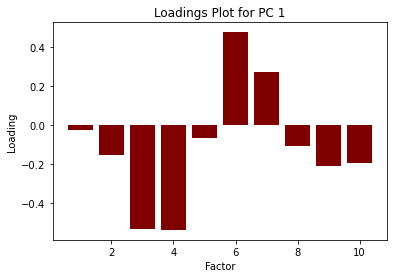
Before performing PCA on the predictors, I created an array of all of their z-scored columns. I then computed the correlation matrix and made the heatmap of the correlations to get an idea of how many principal components there might be.



Based on this heatmap, it looked like there may be 2 or 3 meaningful principal components, since there are a few squares that are green, indicating a strong correlation. I then fitted a PCA model to the data, and created the scree plot to determine the number of principal components I should interpret.

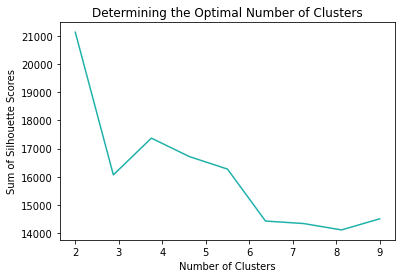
****

Using different methods to analyze the scree plot indicated different numbers of meaningful principal components. Using the elbow method, it looks like there is one meaningful principal component. The kaiser criterion indicates three, as there are three principal components with eigenvalues above the kaiser criterion line on the scree plot. Using the 90% variance method indicated six principal components. Based on the loadings plots for all of these principal components, I felt that the first three were the most distinct and meaningful.

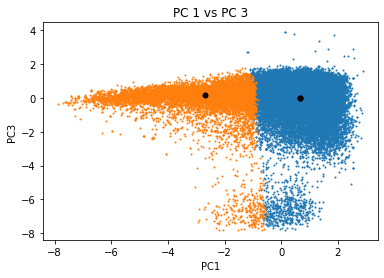
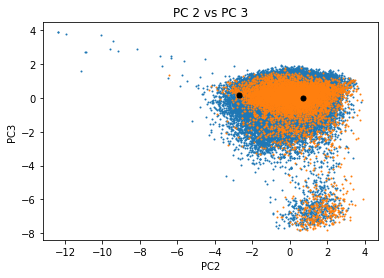
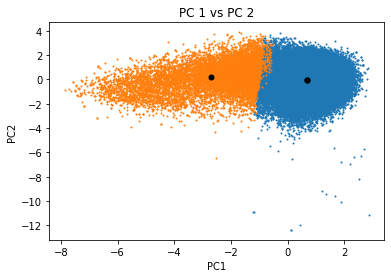
****

The loadings of the first principal component indicate high acousticness and instrumentalness, but low energy and loudness, which I interpreted as “softness” or “calmness”. The second principal component is composed of low danceability and low valence, as well as moderate instrumentalness, which I interpreted as “melancholy”. The third principal component is composed of high liveness and speechiness, which I feel indicates “word-based performance.” I wasn’t sure what to call this principal component, but it indicates that a track is likely live and is mostly spoken word, which based on my EDA could be something like a podcast, a comedy show, or an audiobook. These principal components account for 57.36% of variance.

I then ran a loop to determine the number of clusters that can be identified from the principal components. I created a k-means model with the rotated data for the first three principal components, with values of k ranging from 2 to 10. For each k, I computed the silhouette score, and I found that the optimal number of clusters was 2.

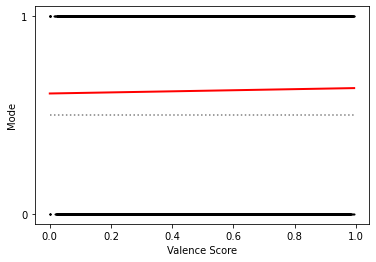


I then fit a k-means model to two clusters. I was unable to plot the data in three dimensions, so I created three two-dimensional plots with each combination of principal components.



Question 9

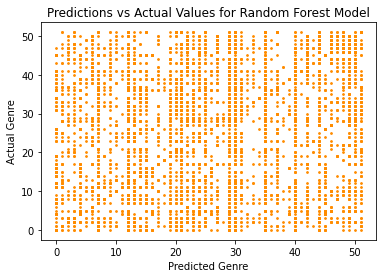
Before attempting to fit a logistic regression model to predict mode from valence, I looked at the sample statistics of valence for the major tracks and the minor tracks. They were pretty similar, and I also created the scatter plot of valence and mode, which looked like it would not cooperate with a logistic regression model. However, I still fit the logistic regression model to the data, and fit the logit curve.



Based on this plot, it does not look like mode can be predicted from valence. I also computed the confusion matrix of the model using the test set, and found that the model predicted 1 every time (major). So, although the model’s mean accuracy score was about 60%, this was only because the majority of the points in the test set were in a major key. This model clearly does not work for this predictor. I also fit the model for every other numerical column in the dataset, to see if any of them were a better predictor for mode. However, all of the models had the same mean accuracy score when predicting the mode of the test data, and they all predicted 1 every time, so none of these predictors work to predict the mode of a track.

Question 10

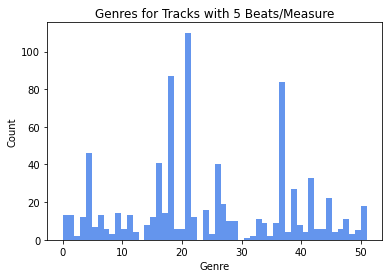
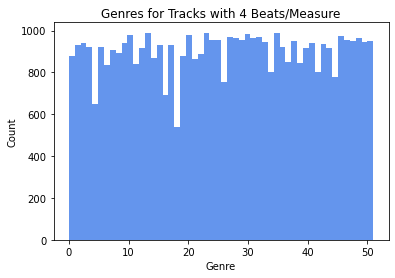
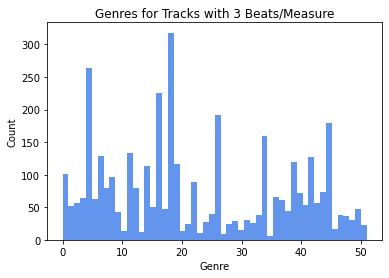
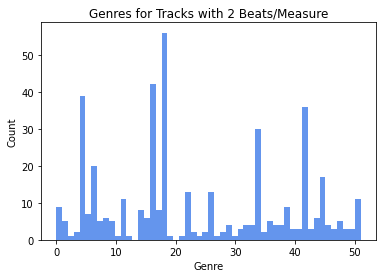
I chose to use my three principal components to try to predict genre. In order to fit the random forest model, I first mapped each of the genres to a corresponding number. Then, I fit the random forest model using the 80/20 train/test split to avoid overfitting, and computed its accuracy, which was 0.198. I also plotted the model’s predictions for the test set vs their actual values.



Based on this plot as well as the accuracy score, this is not a very good model for predicting genre. From my EDA as well as experience with the data over the rest of the project, I feel that this is due to the fact that there are a lot of factors not present in this dataset that determine genre. For example, some of the genres are things like “British,” “French,” and “German,” so the genre is based on the nationality of the artist. This data does not include that information, so the model would have no way of knowing and would therefore try to predict the genre based on the features of the song, which may differ greatly within the genre. Some of the data is also meaningless for certain genres. For example, there is a “comedy” genre with comedy shows, but they still have keys, time signatures, and tempos, although that information does not really mean anything as the track is just a spoken show. So, the random forest model could not be expected to be completely accurate, because the determination of genre was likely based on information that is not included in the dataset.

Extra Exploration

For additional exploration, I wanted to determine if there was any sort of relationship between the time signature of a track and its genre. Most of the tracks in this dataset had 4-based time signatures, so I was mainly wondering whether there was a certain genre that held the majority of tracks with 5-based and 1-based time signatures. I created subsets of the data with tracks for each number of beats per measure, and plotted histograms for the genres within each subset.



I found that while the distribution of the genres for tracks with 4 beats per measure looked pretty uniform, the rest did not. The distribution of the genres for tracks with 4 beats per measure did have some gaps, which were the more popular genres for the other time signatures. For example, genre 18 is comedy, which is the majority genre for 2, 3, and 5 beats per measure, but is the genre for which there are the least number of tracks with 4 beats per measure. Other genres like this are 16, which is classical, and 4, which is ambient. I also noticed that the most popular genre for tracks with 5 beats per minute is genre 21, which is dancehall. I feel that these distributions are very interesting and somewhat surprising, but I also think that the “time signature” category would have been better represented by actual time signatures, like 6/8 or 4/4, since songs with 4 beats per measure can still have many different actual time signatures. This may have resulted in a more specific understanding of how genres are distributed within time signatures.