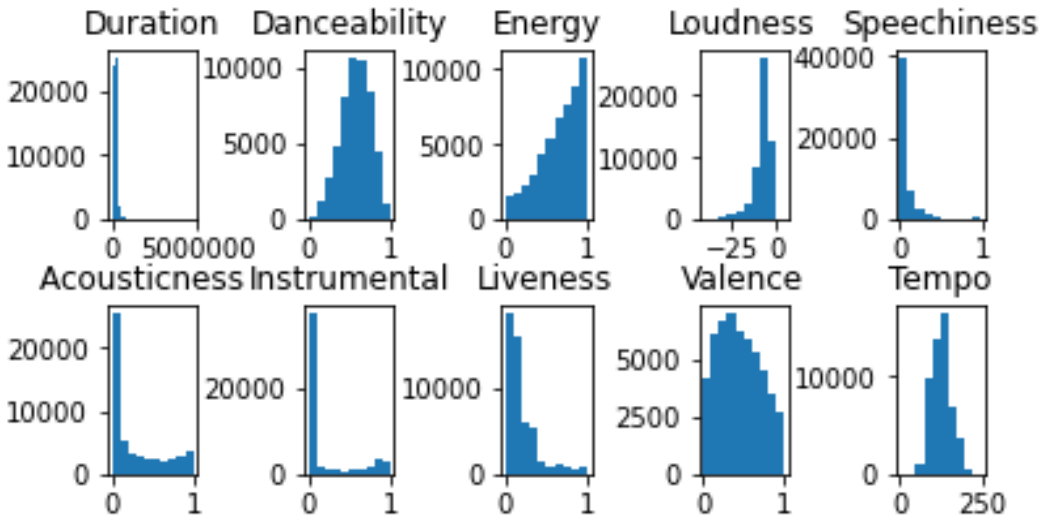
Preprocessing

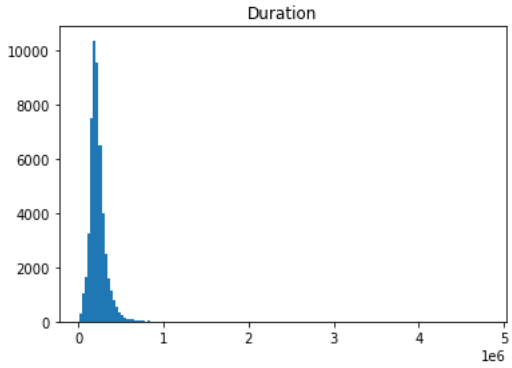
I started off by importing the data as a pandas dataframe to maintain all the column labels. I then converted the data into numpy and created arrays for certain song features so that it will be easier to do calculations and tests for those individual features. In addition, when I created those arrays, I converted the data types for the song features with numerical values to either float for decimals or int for integer data because numpy originally converted the data as type object and converting it to a numerical type makes it easier to use in calculations and other functions.

Question 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?

After plotting a histogram for these 10 categories, the set of histograms look like this:



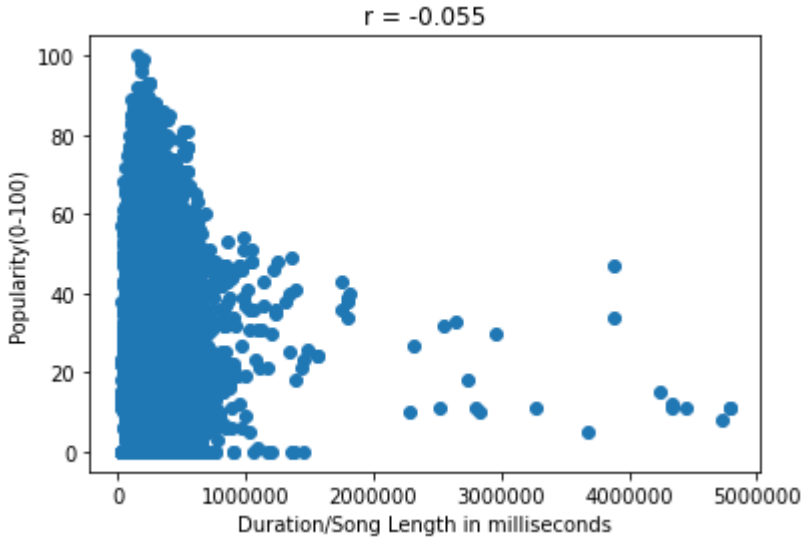
Upon inspection of the distribution for each feature, it can be seen that most of the distributions are not normal. However, 2 interesting distributions to look at are Danceability and Duration. For Danceability, the shape is roughly close to the shape of a normal distribution but since it is still slightly left skewed, I would not consider it reasonably normally distributed. In addition, I also noticed that the Duration histogram looked like it has a few outliers since most of the data is visibly concentrated on the left. I decided to plot the histogram for duration individually to get a better idea of the shape of the distribution and this is the histogram:



After seeing the histogram for duration, I noticed that the histogram is too skewed to the right for the shape to be considered reasonably normally distributed. Therefore, none of the distributions can be considered reasonably normally distributed.

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

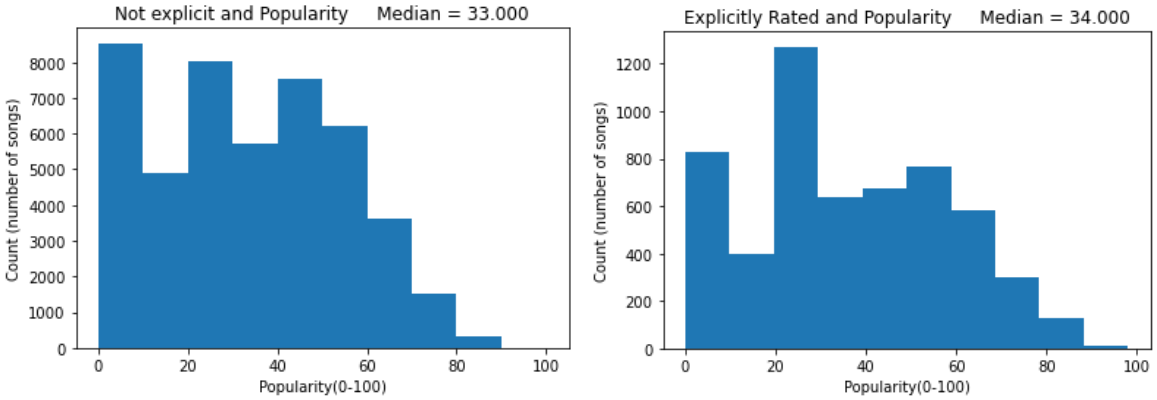
After creating a scatterplot with the duration of the song in milliseconds on the x-axis and the popularity of the song on the y-axis, the final result looks like this:



From looking at the scatterplot, songs that have similar durations/lengths can have a variety of popularity ratings so these 2 variables will likely have a weak or no relationship(since there is a big block of data points for songs all under 1 million milliseconds that spans almost the entire popularity range). This is also supported by the Pearson correlation between duration and popularity, which is -0.055. This illustrates a very weak negative linear relationship between the 2 variables overall. In addition, an interesting trend to note is that looking at the songs with durations over 1 million milliseconds, all of these songs have popularity under 60.

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

First, I plotted a histogram of the popularity scores for both songs that are explicitly rated and not explicitly rated. The histograms look like this:

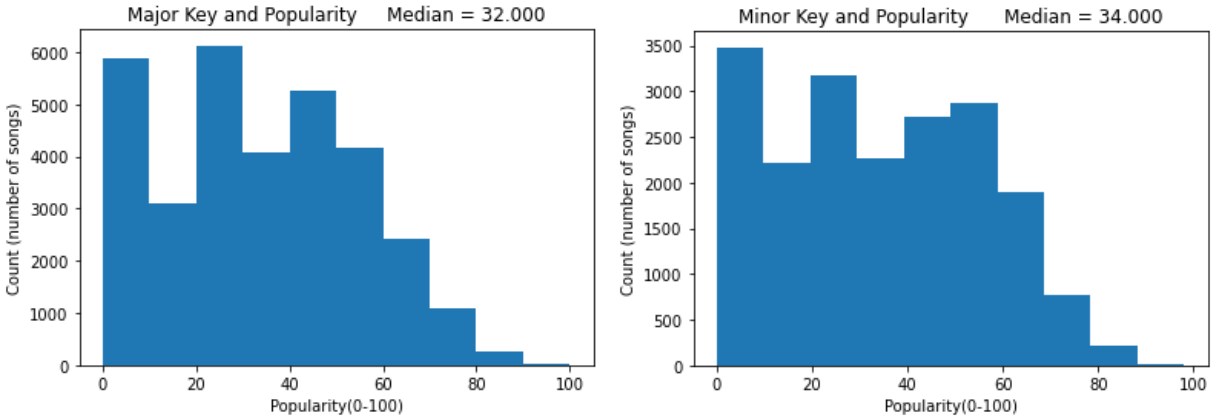


Looking at the histograms, it is evident that the distributions are not normally distributed, so a nonparametric test is needed. Therefore, I chose to do the Mann-Whitney U test to compare the medians of the 2 samples. The median popularity for non explicit songs and explicit songs are 33 and 34 respectively. From this, I can see that the explicitly rated songs do have a slightly higher median popularity than the not explicitly rated songs for this specific sample.

After doing the Mann-Whitney U test, the p-value is 3.0679199339114678e-19, which is very close to 0. Therefore, since the p-value is less than the predetermined alpha of 0.05, we can conclude that the difference in the medians is very unlikely due to chance. Therefore, the null hypothesis that there is no relationship between popularity and if a song is explicitly rated or not can be dropped. It can be concluded that songs that are explicitly rated are more popular than songs that are not explicitly rated.

Question 4) Are songs in major key more popular than songs in minor key?

First, I started off by plotting a histogram of popularity scores for songs in major and minor keys:

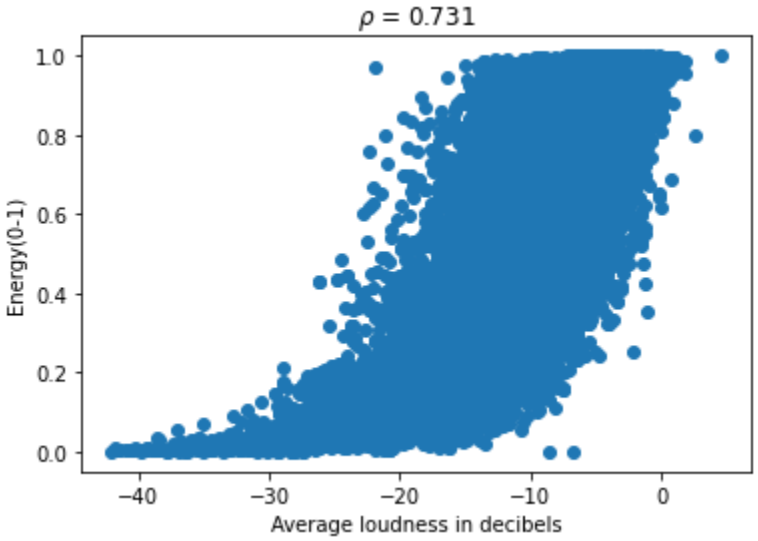


Looking at the shape of the histogram, it is evident that the distribution of popularity scores are not normally distributed. Therefore, a Mann-Whitney U test will be needed to determine if songs in major keys are more popular than songs in minor keys.

In addition, the median popularity for songs in major and minor keys are 32 and 34, respectively. From this, I can see that songs in minor key have a slightly higher median than songs in major key in this sample. After doing the Mann-Whitney test, the resulting p-value is 2.0175287554899416e-06. This is smaller than the predetermined alpha value/threshold of 0.05, which means that the probability of getting this difference in median is unlikely to be due to chance. Therefore, the null hypothesis that there is no relationship between the popularity of a song and if the song is in major or minor key can be dropped. Therefore, I conclude that songs in minor key are more popular than songs in major key.

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

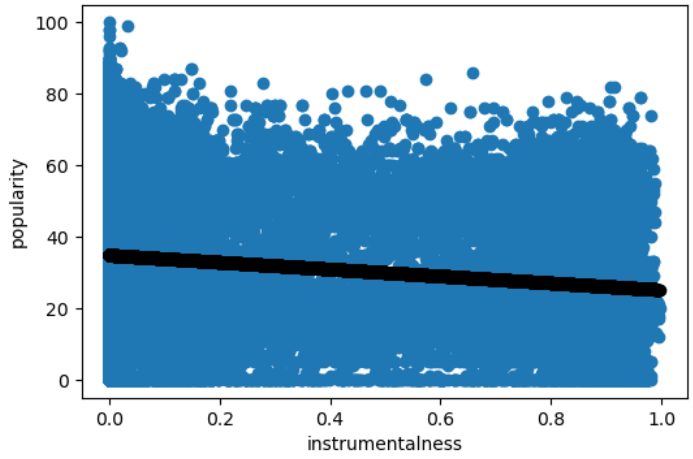
After plotting a scatter plot of the loudness in decibels on the x-axis and the energy level on the y-axis, the scatterplot looked like this:



I noticed that there is a general upward trend formed by the data points and there is a monotonic relationship. Therefore, the correlation coefficient that is most suitable for this relationship is Spearman's rank coefficient. After calculating the Spearman's rank coefficient, rho is 0.731, which suggests that there is a moderately strong positive monotonic relationship between the loudness of a song and energy. Therefore, I would say that this supports the belief that energy largely reflects the loudness of a song.

Question 6) Which of the 10 song features in question 1 predicts popularity best? How good is this model?

For this question, I decided to use linear regression for using a song feature to predict the popularity score. I created a loop that iterates through all 10 features and finds the slope, intercept, R^2, and predicted popularity for the regression model. In addition, I also created a loop to plot a scatterplot for each song feature and popularity.
The highest R^2 score out of the ten features was 0.021017(which meant that this song feature predicted popularity best). This score corresponds to the feature “instrumentalness”. This R^2 was very low and suggests that instrumentalness is not a good predictor of popularity since it only explains about 2% of the variation in the outcomes. Because of the low R^2 score, I decided to look at the plot and this is the resulting scatterplot:



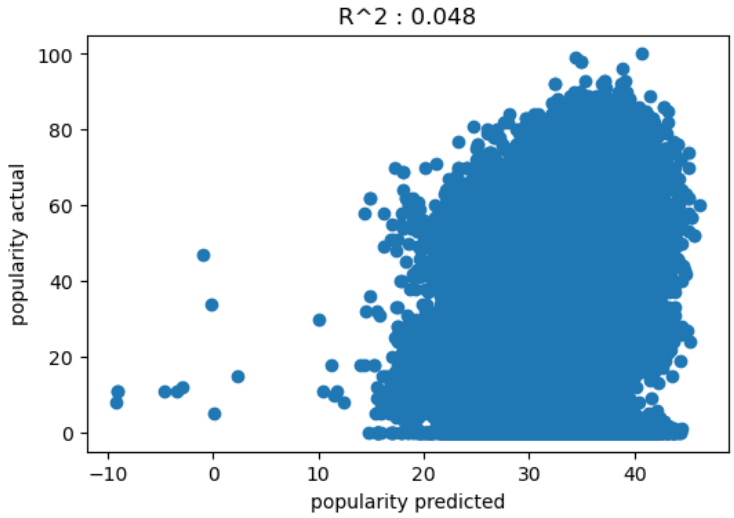
The blue data points are the instrumentalness score and the actual popularity score and the black data points are the instrumentalness score and predicted popularity score from the model. This graph explains the low R^2 because since the data is so scattered, there is no clear relationship between instrumentalness score and popularity. Looking at the scatterplots for the other 9 features, a similar phenomenon is noticed: similar values for the song features can correspond to a large range of popularity values for different songs. Thus, it would be hard to create any model that is able to accurately predict popularity using one song feature. The reason for this might be because multiple song features contribute to the overall popularity and no one single song feature has a large influence on the song’s popularity.

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

Using all 10 song features from question 1, I used multiple regression to predict popularity from the song features and the R^2 score is 0.048(rounded to the nearest thousandth). This suggests that this model that uses all the features is able to predict popularity a little better(it explains about 2% more of the variance in outcomes) than the model that uses only the instrumentalness feature. However, this model only accounts for 4.8% of the variance in the

outcomes, which is still a very low percentage, meaning that although this model is better than the previous model that only uses the instrumentalness feature, it is not a good predictor of popularity. A possible explanation for this small improvement in R^2 may be because of multicollinearity among the different song features.

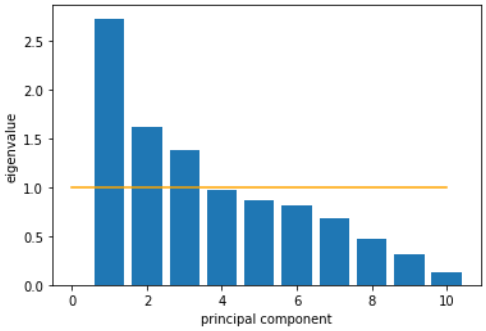
Plotting a scatter plot of predicted popularity scores versus the actual popularity scores, this is the scatter plot:



This cluster of points on the right suggests that for a popularity score predicted, the actual popularity score is a wide range of values instead of having one correct value. This may be because people like songs for different reasons and there might not be a specific song feature or set of song features that is associated with a high or low popularity score.

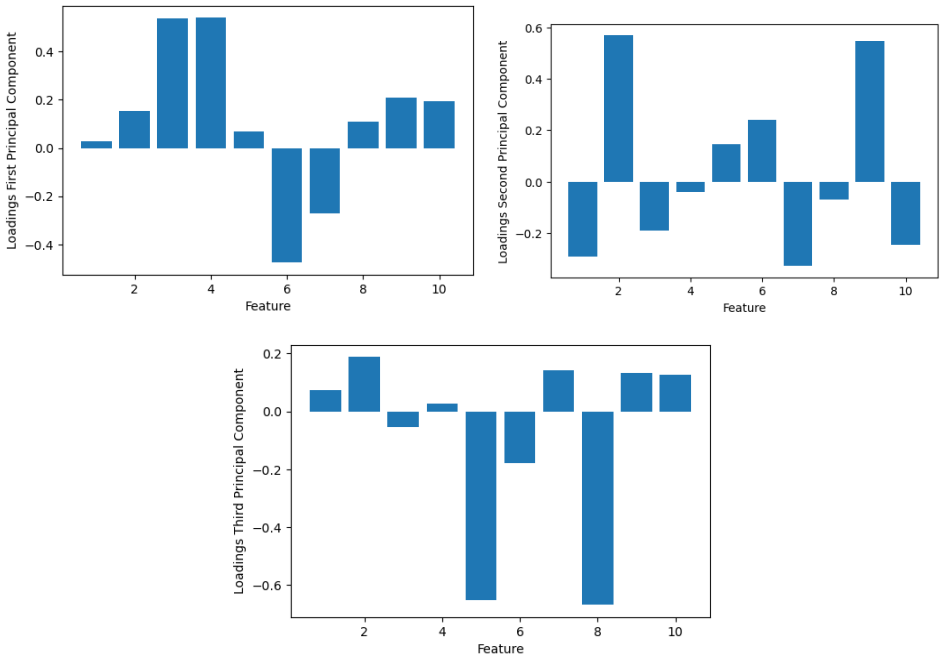
Question 8) When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using the principal components, how many clusters can you identify?

After running a PCA on the 10 song features, the scree plot looks like this:



Using the Kaiser Criterion(shown by the orange line), there are only 3 meaningful principal components.

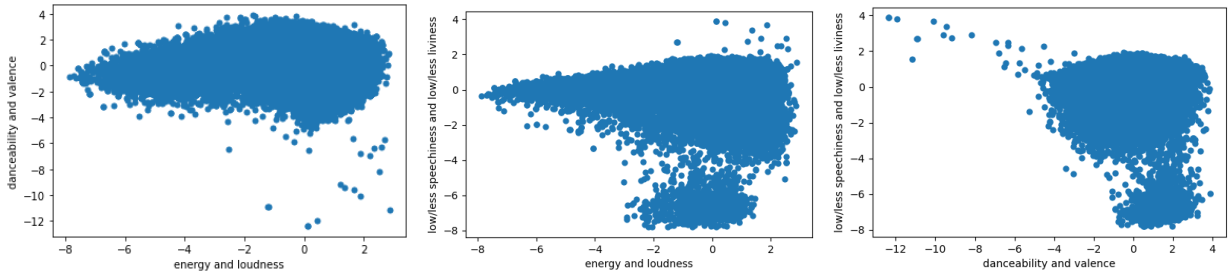
Looking at the loadings for the three principal components, the loadings matrices look like this:



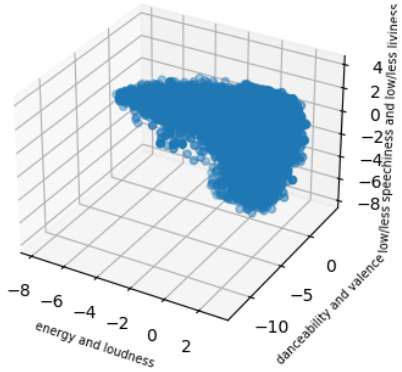
The features that contribute most to the first principal component (in the positive direction) are features 3 and 4, which correspond to energy and loudness, respectively.
The features that contribute the most to the second principal component(in the positive direction) are features 2 and 9, which correspond to danceability and valence, respectively.
The features that contribute the most to the third principal component (in the negative direction) are features 5 and 8, which correspond to low/less speechiness and low/less liveness.

I calculated the variance that each principal component accounts for by dividing the eigenvalue for each principal component by the total sum of eigenvalues. To calculate the total variance that the 3 meaningful principal components account for, I just summed the variance accounted for for the 3 principal components. The proportion of the variance that the 3 meaningful principal components account for is about 57.358% (rounded to the nearest thousandth).

When plotting the rotated data, I first plotted 2D scatter plots comparing 2 of the principal components at once:



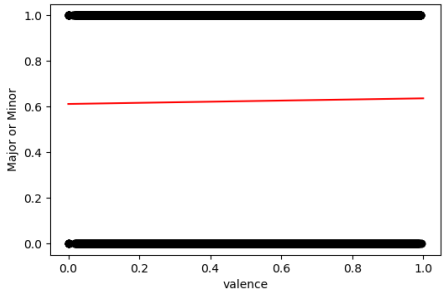
Then, I plotted a 3D graph comparing all 3 features:



Therefore, given the scatterplots, it is evident that there are 2 clusters formed by the principal components(separated where low/less speechiness and low/less liviness is about -5, as seen in the second and third 2D plots).

Question 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?

After using a logistic regression to predict if a song is in major or minor key from valence, the resulting model looks like this:



Looking at the graph, it is evident that valence is not a good predictor. For both songs that are in major key and songs in minor key, the songs have valence scores that span the entire range from 0 to 1, so valence will not be a good feature to use to differentiate between songs in major and minor key. This can also be seen through the slope of the logistic regression curve, which in the diagram is almost flat, shows how there is no clear division between songs that are in major and minor keys given only valence.

In addition, by calculating the AUC score using the roc_auc_score() function from sklearn, the AUC score for this model is 0.507(rounded to the nearest thousandth). Given that the worst AUC score is 0.5, this score is very close to that value which means that this model has an accuracy similar to randomly guessing if the song is in major or minor key.

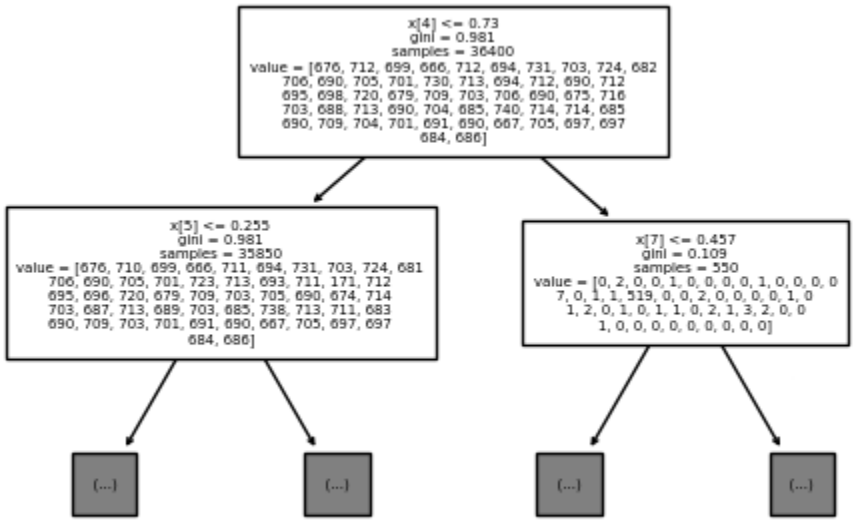
In addition, there is no better predictor for if a song is in major/minor key. For each of the 10 song features from question 1, plotting a scatter plot of the major/minor key versus the score for that feature yields a similar scatter plot as using valence as a predictor: there is no clear

difference in the values for each song feature for songs that are in major keys versus songs that are in minor keys .

Question 10) Can you predict the genre, either from the 10 song features from question 1 directly or the principal components you extracted in question 8?

Using the 10 song features from question 1, it is not possible to predict the genre of the song. I used a decision tree to determine song genre from the song features by using 30% of the data as the test set and 70% as the train set. Since this model is predicting multiclass data, an AUC score may not be the best option so for this model, I am assessing how well the model predicts the data with its accuracy score. The accuracy of the model is 24.5%, which is a low accuracy score.

The decision tree will have a lot of nodes and leaves because there are 52 different genres and 10 features used. Showing the entire tree is not possible but the first 2 levels of the decision tree look like this:

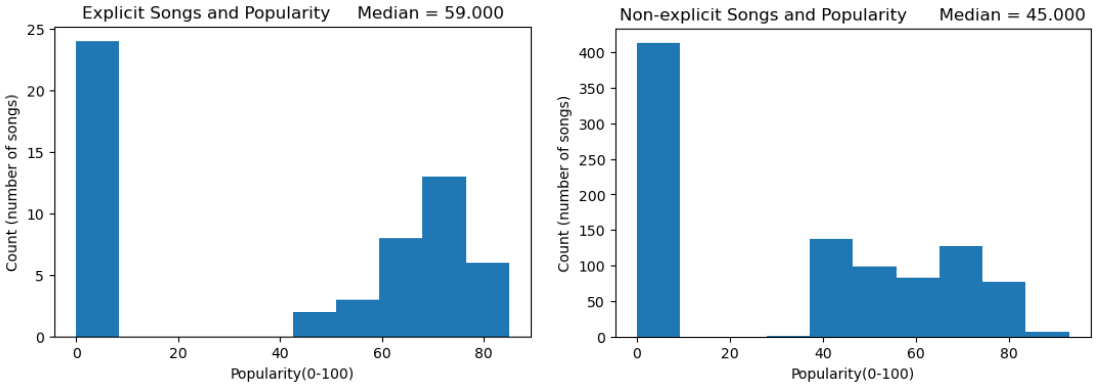


Extra Credit) Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions

Are explicitly rated songs more popular than songs that are not explicit for songs that are alt-rock?

This essentially solves question 3 again but this time, I am controlling the genre variable and only using data from songs that are classified as alt-rock. Using this, I want to compare the results with the results from question 3 and see if there are any interesting observations.

First, I plotted a histogram of the distribution of popularity scores for both songs that are explicitly rated and songs that are not explicit and compared the median popularity scores.



After using a Mann-Whitney U test, I got a p-value of 0.04099008249033649, which is less than the predetermined alpha value of 0.05. Something that I noticed is that even though this is still significant, the p-value is much higher than the p-value in question 3, which was only 3.0679199339114678e-19. This may be because of the large difference in median popularity between explicitly rated songs and not explicitly rated songs and that the sample sizes are smaller when only looking at songs that are alt-rock. Nevertheless, since p-value is less than alpha, it can be concluded that explicitly rated songs are more popular than not explicitly rated songs (songs that are alt-rock). However, even though the conclusion is the same, there are some interesting differences.

By looking at the median popularity ratings, I noticed that the median popularity for explicit songs is much higher than non explicit songs(14 higher) for alt-rocks songs compared to question 3 where the median popularity scores only has a difference of 1. In addition, the median popularity is also higher when only considering songs that are alt-rock: 59 for explicitly rated songs and 45 for not explicitly rated songs. In comparison, the median popularity is 34 for explicitly rated songs and 33 for not explicitly rated songs in question 3.