Alexandra Przysucha
Pascal Wallisch
Principles of Data Science
19 December 2023

Capstone Project: Analysis of Spotify's 52,000 songs dataset (What makes music popular as well as the audio features that make up specific genres)

For starters, I used both NumPy and pandas to work on this project. To read the .csv file provided by the Spotify dataset of around 52,000 songs, I used pandas and converted the data frame to a NumPy array**.**

Data Cleaning: Prior to performing any tests or anything else with the data, I cleaned the data. It is important to clean and preprocess data prior to manipulating it and I did so with the following techniques: dropping duplicates of necessary rows (songs), in order to correct the repeated data in the project, and NaN removal of necessary rows (songs), in order to correct the missing data in the project. It is essential to note that removing duplicates and removing NaNs from **all** rows will not be done, because the goal is still to keep as much of the data as possible.

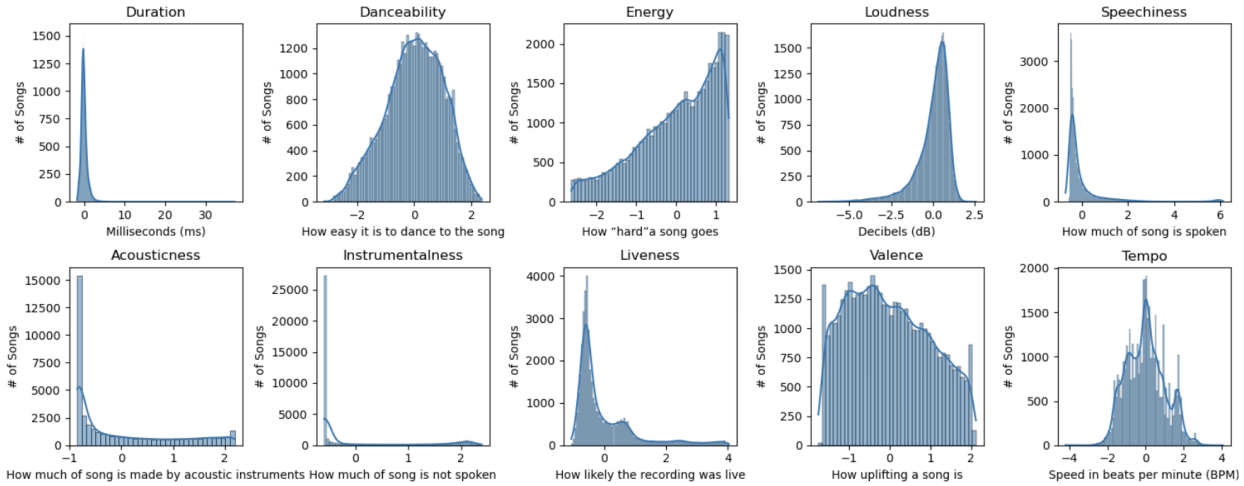
For dropping duplicates for songs, I noticed that there were many songs that had the same track/song name but were on different albums. I printed all of those duplicate songs, the songs with the same song title that were on different albums, to determine if they were different in any other way, such as in tempo, popularity, and any other song features. I came to the conclusion that some songs were on different albums because they were covers of the original song, and because of that they have different tempos, popularity ratings, and are different in all other aspects except for song title. On the other hand, songs that had the same track name that were on different albums only differed in the album name, with all other features being the same. I realized that the duration for these songs were the same, but for songs that were differing in other aspects had different durations than the original song, which logically makes sense because covers of songs are sung by different artists than in the original version of the song, and because of this some singers just naturally sing softer or louder and using more or less vocal runs than the original artist. Additionally, covers often use different instruments and from time to time, are done acoustically especially if the original was done with a heavy influence or reliance on electronic instruments. All of these changes can heavily impact the duration of a song as well as the tempo and other features, but through my investigation of the duplicate songs I notice that the duplicate songs that had the same duration, everything else was also the same excluding the differing album names. For this reason, I chose to only remove the duplicate songs that have the same duration, and only keep 1 occurrence of the song that has that duration, tempo, valence, etc. This removes a lot of duplicates but also doesn't just mindlessly remove all the duplicate songs, which does remove an enormous amount of data that does impact everything else done and all the correlations and tests we perform later on.

Also, by implementing the NaN removals for only the features and the songs that are being looked at for the specific question, so that we are dropping complete rows of data, a.k.a. Songs from the dataset that aren't necessary or needed to be dropped. Essentially, only the variables that are necessary for the question looked at specifically or the project would be taken from the data and would be cleaned at this step to minimize the amount of data lost.

Additionally, prior to any PCA, I made sure that the data was z-scored before it was used in the PCA, and then for the data transformation, I rotated to graph the old data in the new coordinate field. Furthermore,

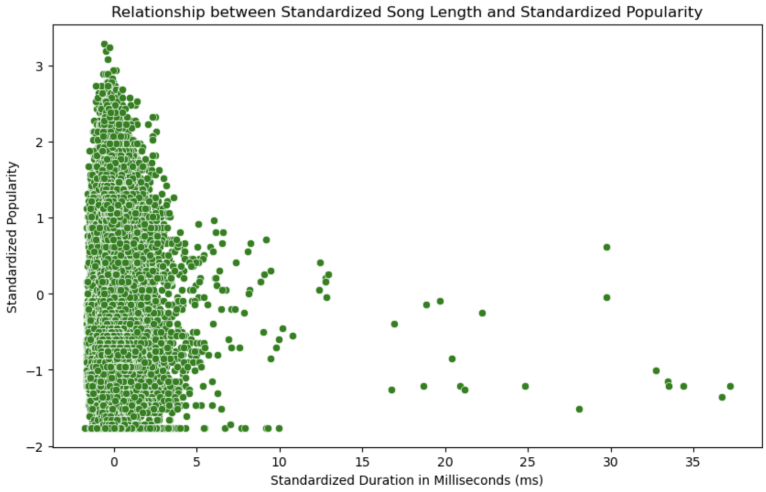
when the question required me to, in order to reduce the dimensionality of the dataset, I performed a Principal Component Analysis (PCA). Lastly, I seeded the RNG with my N number

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



Response to Question 1: Out of the 10 song features stated, I said that features that are the most normally distributed would be danceability and maybe tempo, but neither are perfectly normally distributed with danceability having a slight left skew. It is important to note that there is a possibility of the danceability and tempo feature to be more normally distributed if there was more data and more values, due to the combination of the ideas of the Law of Large Numbers and the Central Limit Theorem.

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

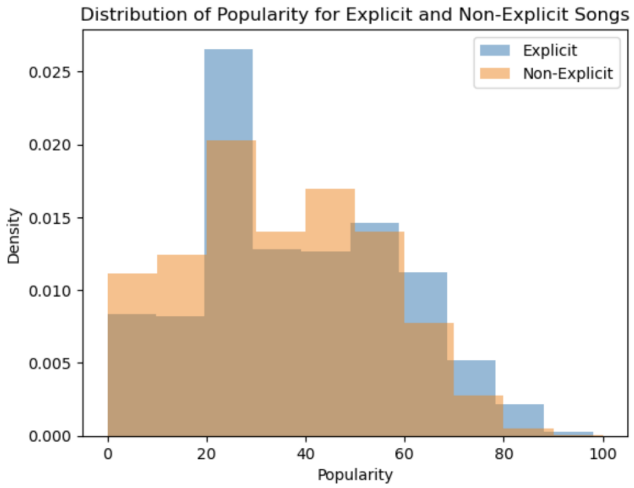



```
Spearman's Correlation Coefficient: -0.0948
Pearson's Correlation Coefficient: -0.0963
```

Response to Question 2: At the bottom of the plot, the correlation coefficient for both Spearman and Pearson is listed to be about -0.095. Spearman's correlation coefficient is used for when the data is not normally distributed or for ordinal or ranked data and Pearson's correlation coefficient is used when the data is normally distributed and to assess linear relationships. From question 1, we know that neither popularity nor duration are normally distributed, so we will use Spearman's Correlation coefficient to evaluate the relationship between duration and popularity. The correlation coefficient can range from -1 to 1 and a positive value indicates a positive correlation, and a negative value indicates a negative correlation. Additionally, the closer the value is to either the maximum value of 1 or the minimum value of -1, the stronger the correlation. Therefore, since our correlation coefficient is very close to and negative, we can state the correlation between song length and popularity is very weak and negative. In English, this means that there is practically no relationship between the duration of a song and the song's popularity.

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

```
Sample mean of the popularity of the explicit data: 37.648620181167054
Sample mean of the popularity of the non-explicit data:34.48352376660216

Sample size of the popularity of the explicit data: 4747
Sample size of the popularity of the non-explicit data:37721

Variance of the popularity of the explicit data: 431.7395487699498
Variance of the popularity of the non-explicit data: 385.1485156420776

Mann-Whitney U test p-value: 6.095345411719456e-21
Reject the null hypothesis. There is a significant difference in popularity.
```

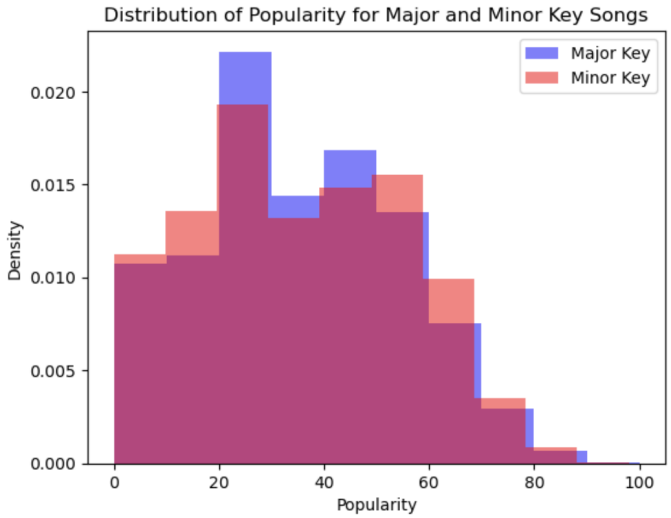


Response to Question 3: For this question, I performed a Mann Whitney U test to investigate whether explicitly rated songs more popular than songs that are not explicit, because the distributions are not normal, as seen in the Histograms that overlayed on top of each other, above, therefore a t-test could not be performed because it assumes the distributions are normal. Additionally, the variances were not the same and nor were the sample sizes and thus neither a t-test nor a Welch's t-test could be performed. Using the Mann Whitney U test, a p-value of practically 0 (6.095345411719456e-21) was produced, and because it was less than alpha, 0.05, I rejected the null hypothesis, with the null hypothesis being that there is no statistically significant difference between the popularity of explicit songs and non-explicit songs, and therefore I concluded that there is a significant difference in popularity. In English, this means that explicit songs are more popular than non-explicit songs and there is a statistically significant

difference in the two groups of songs: songs that are non-explicit and the songs that are explicit, that is unlikely due to chance.

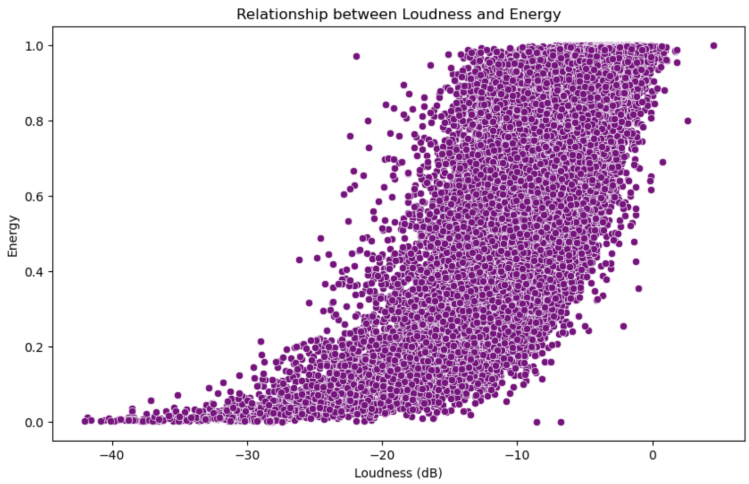
4. Are songs in major key more popular than songs in minor key?

```
Sample mean of the popularity of major key songs: 34.61602690040804
Sample mean of the popularity of minor key songs: 35.203375

Sample size of the popularity of major key songs: 26468
Sample size of the popularity of minor key songs: 16000

Variance of the popularity of major key songs: 381.516782460773
Variance of the popularity of minor key songs: 407.40247626414845

Mann-Whitney U test p-value: 0.9924090471044591
Fail to reject the null hypothesis. There is no significant difference in popularity.
```



Response to Question 4: For this question, I performed a Mann Whitney U test to investigate whether songs in major key more popular than songs in minor key, because the distributions are not normal, as seen in the histograms that overlayed on top of each other, above, therefore a t-test could not be performed because it assumes the distributions are normal. Additionally, the variances were not the same and nor were the sample sizes and thus neither a t-test nor a Welch's t-test could be performed. Using the Mann Whitney U test, a p-value of practically 1 (0.9924090471044591) was produced, and because it was more than alpha, 0.05, I failed to reject the null hypothesis, with the null hypothesis being that there is no statistically significant difference between the popularity of songs in major key and songs in minor key, and therefore I concluded that there is no significant difference in popularity. In English, this means that songs in major key are neither more nor less popular than songs in minor key and therefore the differences that exist between those two groups of songs is highly likely due to chance.

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

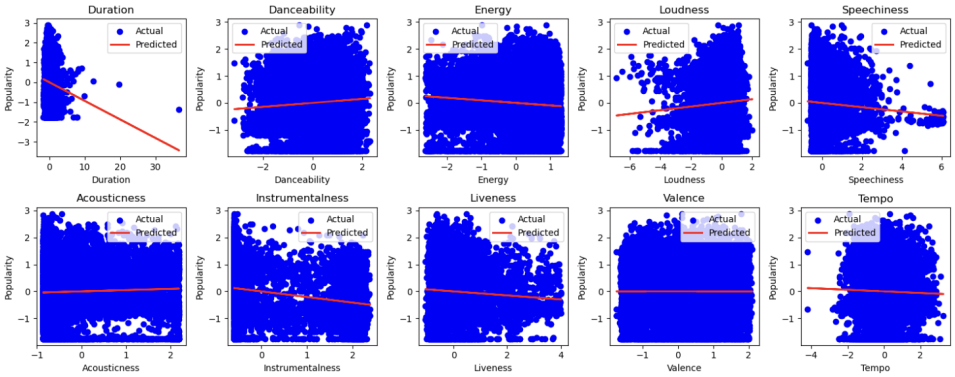



```
Spearman's Correlation Coefficient: 0.7390
Pearson's Correlation Coefficient: 0.7746
```

Response to Question 5: Based on the Pearson's correlation coefficient being 0.7746 and Spearman's correlation coefficient being 0.7390, it can be stated there is a relatively strong positive correlation between the energy and the loudness of a song, since 0.77 is closer to 1 than 0, meaning there is a relatively a strong correlation since the closer the correlation coefficient value is to 1 or -1 the stronger the correlation there is between the two variables, and since 0.7746 is positive, it is a positive correlation. This does support the idea or claim that the energy of a song does largely reflects the loudness of a song, meaning the "louder" a song is, the higher the energy of the song, and respectively, the "less loud" the song is, the lower the energy of the song.

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

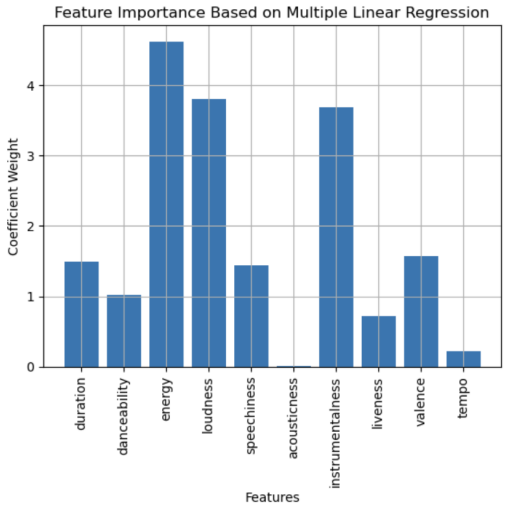
```
Feature: duration
Correlation to popularity: -0.1019
Mean RMSE: 3.96
RMSE: 1.0029
R-squared: 0.0101

Feature: danceability
Correlation to popularity: 0.0683
Mean RMSE: 3.96
RMSE: 1.0058
R-squared: 0.0045

Feature: energy
Correlation to popularity: -0.0936
Mean RMSE: 3.96
RMSE: 1.0037
R-squared: 0.0086

Feature: loudness
Correlation to popularity: 0.0696
Mean RMSE: 3.96
RMSE: 1.0057
R-squared: 0.0047

Feature: speechiness
Correlation to popularity: -0.0804
Mean RMSE: 3.96
RMSE: 1.0048
R-squared: 0.0064

Feature: acousticness
Correlation to popularity: 0.0458
Mean RMSE: 3.96
RMSE: 1.0071
R-squared: 0.0019

Feature: instrumentalness
Correlation to popularity: -0.2115
Mean RMSE: 3.96
RMSE: 0.9853
R-squared: 0.0446

Feature: liveness
Correlation to popularity: -0.0666
Mean RMSE: 3.96
RMSE: 1.0059
R-squared: 0.0042

Feature: valence
Correlation to popularity: 0.0068
Mean RMSE: 3.96
RMSE: 1.0081
R-squared: -0.0002

Feature: tempo
Correlation to popularity: -0.0255
Mean RMSE: 3.96
RMSE: 1.0078
R-squared: 0.0005
```



Response to Question 6: Based on the first image below the Question #6, which shows the correlation between each feature respectively and popularity. We can see the strongest correlation between one of the features listed and popularity out of all the features, is instrumentalness with a correlation coefficient of -0.2115, R^2 value of 0.0466, and RMSE value of 0.9853. Now, looking especially at instrumentalness, we can also see based on the linear regression models for each of the features and popularity, which is the second image, that none of them are very good models to predict popularity, and that makes sense because a song's popularity is not affected by only 1 characteristic or feature of a song, since music is complex and music's popularity is not just affected on how "danceable" it may be, there are many aspects that also affect a song's popularity that are outside of just a song's characteristics. Thus, the linear models perform poorly in predicting a song's popularity. Ultimately, a song's instrumentalness value is not a good predictor of a song's popularity.

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 7). How do you account for this?

```
Multiple Linear Regression:
RMSE: 18.991331268994223
R-squared: 0.09298249259621072
MAE using All Features: 15.407043367297948
```

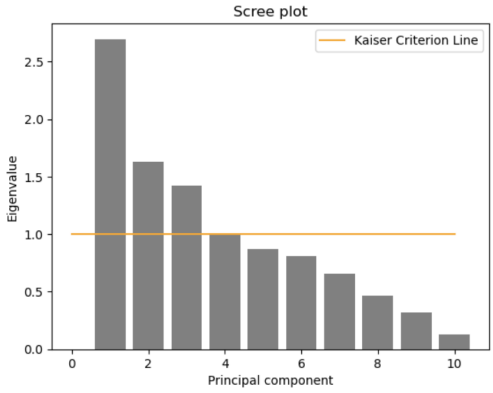


Response to Question 7: In the code, I built a multiple linear regression model to make this model use all the song features from question 1 to see how well I can predict popularity. Comparing the performance of

the model in question 6, which was a linear regression model for each of the features individually, specifically looking at the instrumentalness feature, and the multiple linear regression model for this question, I noticed the Multiple Linear Regression model with all features shows a relatively low R-squared value (0.09298), and the R-squared for the "instrumentalness" feature alone is also relatively low (0.0446), indicating that neither model explains a large portion of the variance in popularity. Although, even though the multiple linear regression model only explains 9% of the variation in popularity, it is still more than the variance explained by the model in Question 6, the linear regression model of the “instrumentalness” feature, which is 4%. The RMSE values for both models suggest that the model using all features has a larger prediction error (RMSE of 18.99) compared to the model in question 6, (which has an RMSE of 0.9853), which makes sense because none of the features really do well in predicting popularity so having a model with all the features would create a larger error in prediction than a model that uses only 1 feature. This same explanation and reasoning applies to why the MAE for the multiple linear regression model using all features is considerably higher (15.4070) compared to the model in question 6 (0.82). Ultimately, the low R-squared values for both models suggest that neither “single” linear nor multiple linear regression models are effective in capturing the complexity of factors influencing song popularity. As mentioned in Question 6’s response, a song's popularity is likely influenced by multiple complex factors beyond individual characteristics. Therefore, the multiple linear regression model in question 7 is only ever so slightly better than the model in question 6 just by the R-squared value, but not enough so to say that the model somehow improved a reasonable or considerable amount, which is realistic because a song’s popularity is influenced by not just song’ characteristics but also by people’s personal experiences, emotions, and many more outside factors that exist beyond this data.

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?

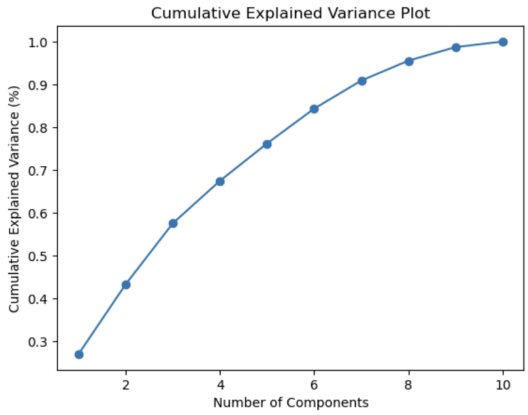
   Response to Question 8: In order to determine how many meaningful principal components there are, I have to perform the principal component analysis (PCA) to find the principal component(s) (PCs). After doing the PCA, I chose 3 principal components based on the Kaiser criterion and the Scree plot, which is below.



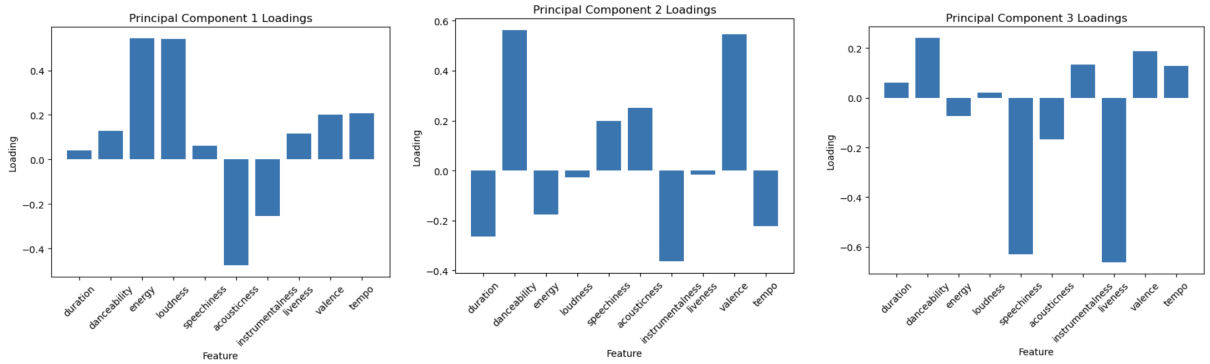

```
Number of factors selected by Kaiser criterion: 3
Number of factors selected by elbow criterion: 1
Number of factors to account for at least 90% variance: 7
```

Still, it is important to note that 3 principal components only account for 57.52% variance in the data, which is not a lot in comparison to where 7 principal components account for at least 90% of the variance of the data, but if we "reduced" our 10 components down to 7 principal components, it would be missing the whole idea surround PCA and dimension reduction of our data. For that reason, based on the Kaiser criterion, I chose to continue to reduce our 10 components to extract 3 principal components.

```
The variance of the data that's explained by our 3 PCs that we chose based on the Kaiser criterion, is 57.52%.
```

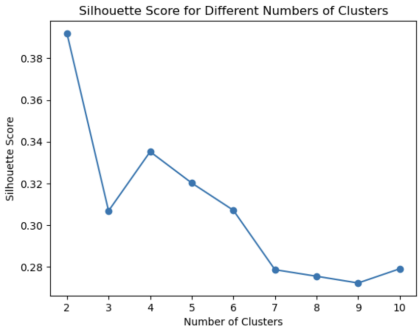


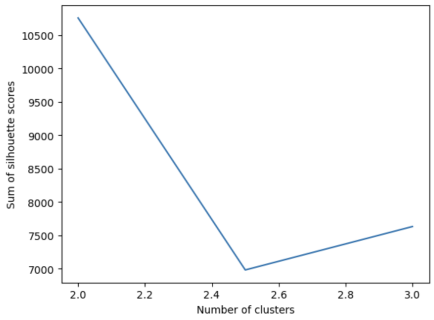
It is important to interpret what these principal components mean, which below are the 3 principal components loading to determine what's the meaning behind each of them.



Going from left to right, the Principal Component 1, I interpreted as extreme rave music or underground electronic dance music, because of the positive and high loading weight of the energy and loudness features as well as the negative and impactful loading weights of acousticness and instrumentalness. The Principal Component 2 I interpreted as mood booster electronic dance music, because of highly positive and strong the loading is for valence and the danceability, as well as the negative loading weight acousticness, I believe most of the dance music would be with synthesized sounds. The Principal Component 3, I interpreted as soft pop study lo-fi beats, since lo-fi music is not recorded live since it all is produced electronically and digitally and not sung by actual singers, as well as the strong negative loading weight of the speechiness feature, makes me think it is not sung vocally and with a weak instumentalness and energy, it makes me think it is music for studying or sleeping, such as lo-fi beats.

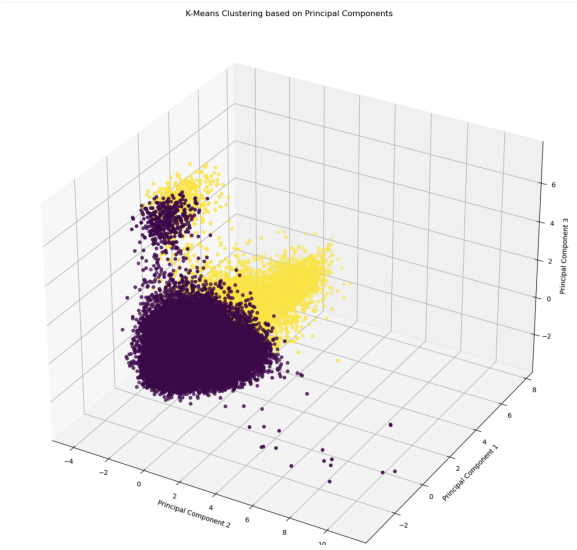
Also, in order to identify the clusters using the principal components, I did K-means clustering. But in order to this, I had to find the optimal number of clusters first, by plotting the sum of the silhouette scores on the y axis, and the range of number of clusters, the maximum of the sum of silhouette scores is found by the optimal number of the clusters, which is seen below is 2. From these 2 graphs and the sum of silhouette scores, it is evident that the highest sum of silhouette scores corresponds to 2 clusters.

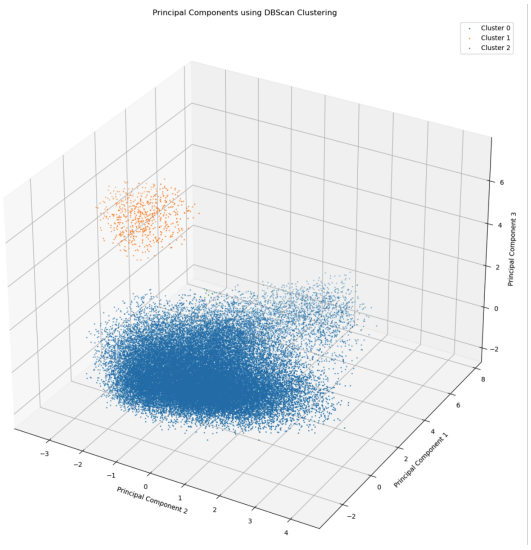




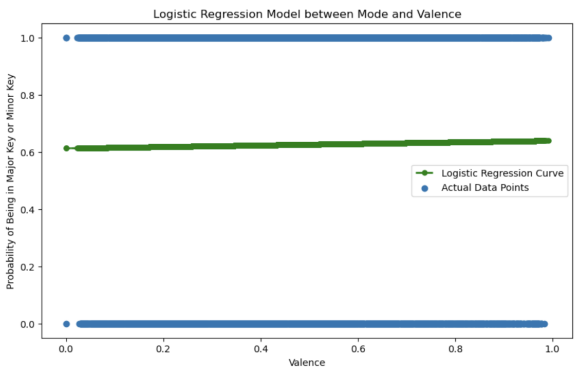

```
From the graph of number of clusters and sum of silhouette scores, we can observe that the highest sum of silhouett
e score corresponds to 2 clusters.
```

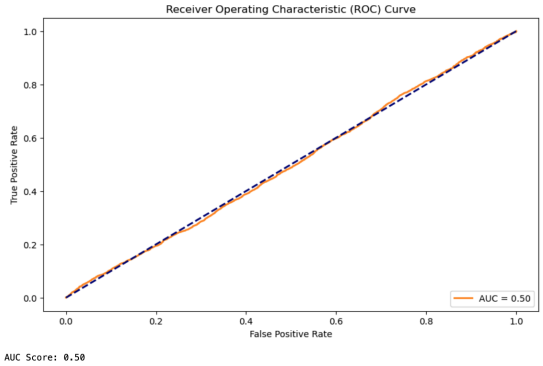
Based off of the results from the figures above, I used the 2 clusters in my K means clustering visualization and DBScan clustering visualization, which are shown below. The DB Scan has 3 clusters in the legend, which 2 of them are the clusters I recognized through the figures above, and the third cluster is noise that exists in the data.





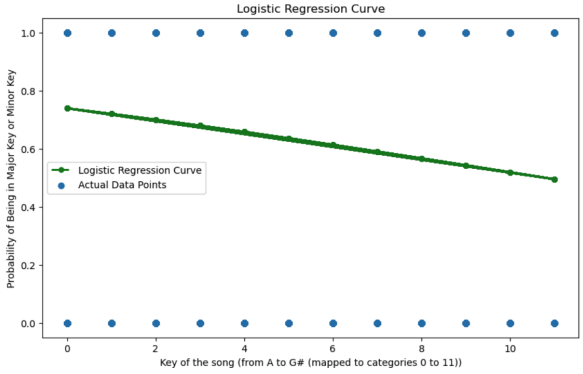
9. Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor? [Suggestion: It might be nice to show the logistic regression once you are done building the model]

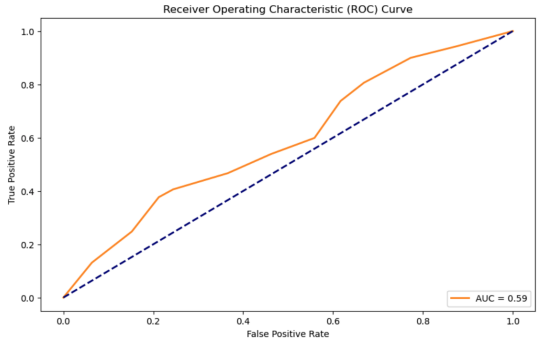




AUC Score: 0.50

Response to Question 9: I did a logistic regression model between valence and mode, where mode was a feature that was a binary categorical variable, where 1 represented the song being in major key, and 0 represented the song being in minor key. The model's accuracy was shown through the AUC (area under the ROC curve) that is displayed on the graph above on the right, and it performed poorly with an AUC score of 0.50, which meant the model was practically random guessing from the valence of song whether the song was in major or in minor key. The best predictor of "mode", or whether a song was in major or in minor key, was key, which makes logical sense because the key feature was representative of what key the song was from A to G# (mapped to categories 0 to 11). The model and ROC curve below of key predicting mode, shows an AUC score of 0.59, which is better than the logistic regression model of valence and mode, where valence was attempting to predict whether the song was in major or in minor key. Therefore, key is a better predictor than valence to predict whether the song is in major or in minor key.





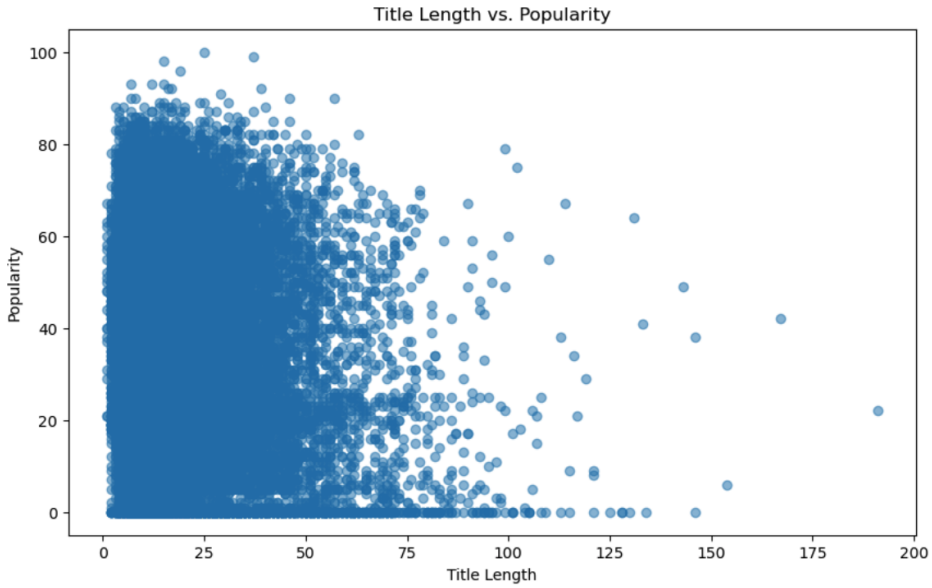
AUC Score: 0.59

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? [Suggestion: Use a classification tree, but you might have to map the qualitative genre labels to numerical labels first]

```
Using Original 10 Song Features:
For Classification Tree (Genre Prediction):
Accuracy: 0.23
Precision
 0.21036765773718297
Recall
 0.20757512357170585

Explained Variance Ratios: [0.26970697 0.16310306 0.14242392]
Using Principal Components:
For Classification Tree (Genre Prediction):
Accuracy: 0.11
Precision 0.10085136088248792
Recall: 0.09920216941737124
```

Response to Question 10: Using the 10 song features, the decision tree classification model had an accuracy of 23%, which was better than the accuracy produced from the decision tree classification model using the 3 principal components I extracted in question 8. In my opinion, I wouldn't say that either model is a good model for predicting the genre of a song, but as mentioned previously in prior parts and questions of this report, music and its aspects such as popularity and genre of song, is such a complex thing to model and predict just with song characteristics or just model using linear and logistic regression models, as well as many outside aspects that are not seen to be directly related to music have an effect on music and song features.

11. Extra Credit: Are songs with long song titles more popular than song with short titles?
I decided I wanted to see if there was a difference in popularity of a song depending on the length of the song's title. I visualized this relationship with a scatterplot, as seen below.



I determined if a song title was a long song title or track name by seeing if it was longer than the median song title length determined from all of the song title data, which was the “track_name” feature of the Spotify song dataset. For a song title to be considered to be a short song title, I did the same comparison as I did for the songs with long song titles but rather than being longer than the median, I checked to see if they were shorter than the median length of the song title. The median song title length was 15 characters. Following this, I performed a Mann Whitney U test to see if there was a significant difference in popularity between songs with short titles and long titles. The result from the Mann Whitney U test indicated that in fact there was a significant difference in popularity between songs with short song titles and long song titles, with the p-value being practically 0.

```
15.0
Mann-Whitney U test p-value: 6.2687136710903215e-65
Reject the null hypothesis.
There is a significant difference in popularity between the short track names and long track names.
```