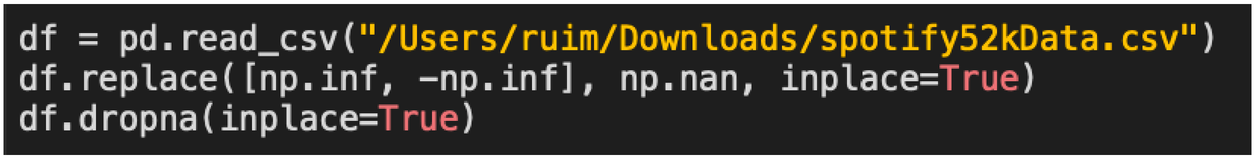
Merry Ma
15 May, 2024

Capstone Project

I used NumPy and pandas for the project's data analysis and manipulation. To read the .csv file, I used pandas and worked with the data frame. When necessary, I conducted Principal Component Analysis (PCA) to decrease the dataset's dimensionality. Before any manipulation, I first performed data cleansing. The second line of the code shown in Figure 1 replaces infinite values with NaN and the third row drops rows with missing values (if appropriate).



```
df = pd.read_csv("/Users/ruim/Downloads/spotify52kData.csv")
df.replace([np.inf, -np.inf], np.nan, inplace=True)
df.dropna(inplace=True)
```

Figure 1: cleaning infinite and NaN data

Questions

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? [Suggestion: Include a 2x5 figure with histograms for each feature)

Two features are reasonably distributed normally based on the shapes of the graph from Figure 2, which are danceability and tempo.

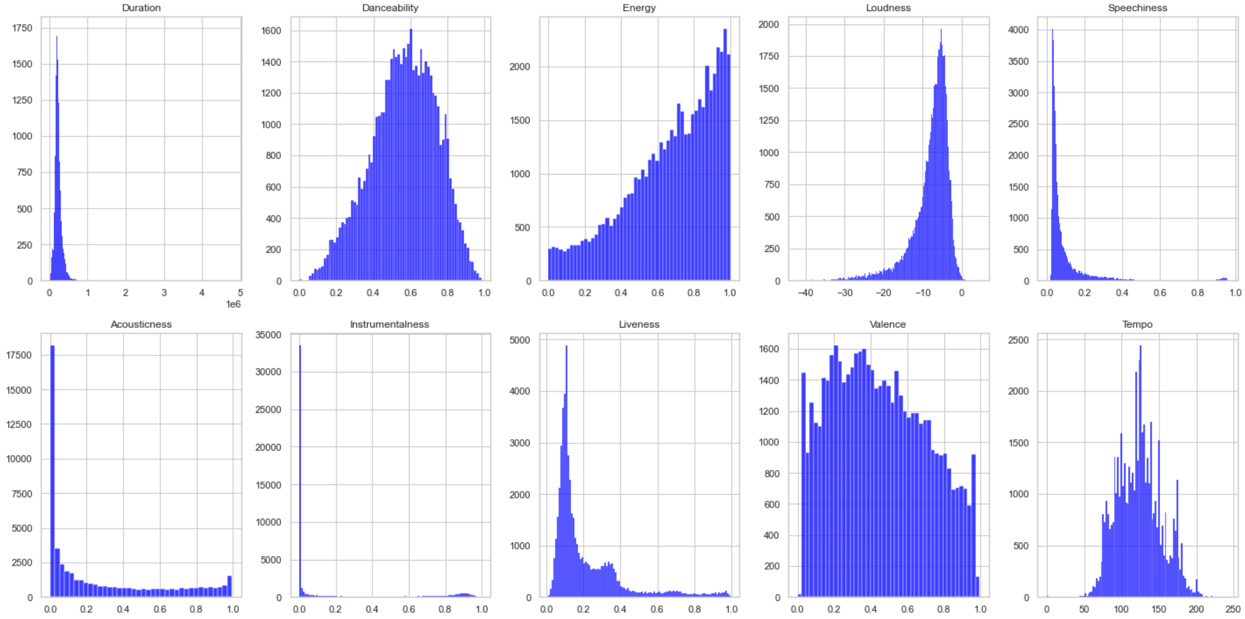


Figure 2: 2x5 histograms for each feature

The ten graphs provided depict the distribution of various song features. Danceability and tempo are reasonably distributed normally because the majority of songs exhibit normal levels of danceability, while tempo tends to hover around a standard pace. This coherence suggests that these features align with common expectations for typical songs.

2) Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative? [Suggestion: Include a scatterplot]

There is no clear relationship between song length and popularity of a song, as seen in Figure 3. However, by calculating the correlation coefficient, the linear relationship is negative with a correlation coefficient of Pearson's R of -0.05. In the case of a non-linear relationship, I calculated Spearman's r on the data again and yielded a correlation of -0.04, which is also negative and not very significant. Thus, I can conclude that there is no relationship between these two features.

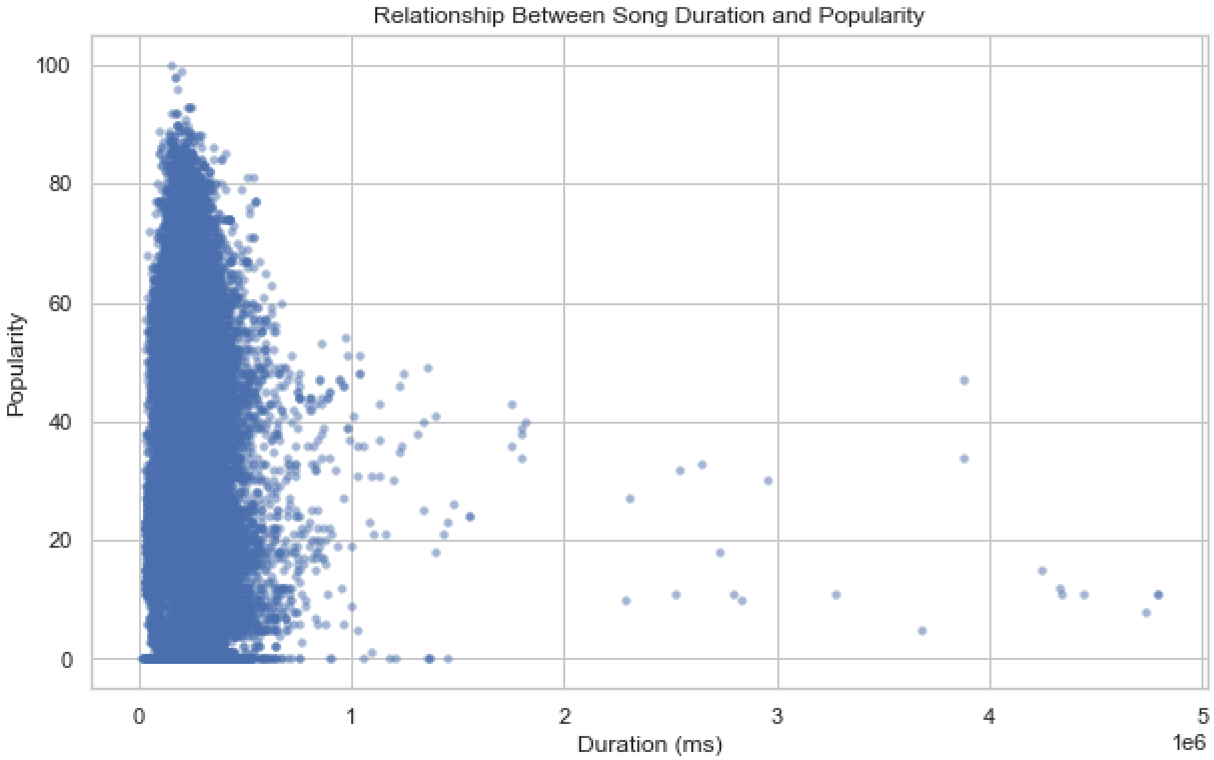


Figure 3: scatterplot of popularity vs. Duration

3) Are explicitly rated songs more popular than songs that are not explicit? [Suggestion: Do a suitable significance test, be it parametric, non-parametric, or permutation]

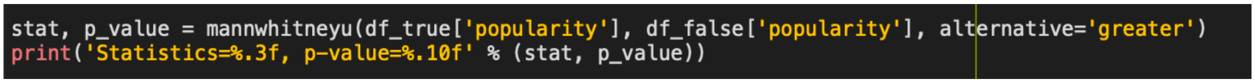
I conducted the Mann-Whitney U test due to the non-normal distribution of the data and the data being ordinal, such that it is not reasonable to reduce the data into sample means. This test was employed to examine potential differences between two independent groups: explicit and non-explicit songs.

The Mann-Whitney U Statistics is 139361273.5, resulting in a p-value that is almost 0. The resulting p-value was found to be smaller than the predetermined alpha level of 0.05 for a one-tailed test, indicating a statistically significant distinction between the groups.

The median popularity of explicit songs is 34.0, while that for non-explicit songs is 33.0. Upon comparing the median values, it became evident that explicitly rated songs were associated with higher levels of popularity.

4) Are songs in the major key more popular than songs in the minor key? [Suggestion: Do a suitable significance test, be it parametric, non-parametric or permutation]

Again, I used the Mann-Whitney U test for ordinal data. Since we are wondering if the major key is more popular, this is a one-tailed test, and I added the “alternative = greater” in the code for this as seen in Figure 4. The Mann-Whitney U statistic is 309702373.0, and the p-value is 0.9999989912. This means that the two groups have different underlying distributions, and since the alpha level is 0.05, the difference is not statistically significant. The median popularity for the songs in a major key is 32.0, and that for the songs in a minor key is 34.0. From these, we can conclude that the songs in the major key are not more popular than songs in the minor key.



```
stat, p_value = mannwhitneyu(df_true['popularity'], df_false['popularity'], alternative='greater')
print('Statistics=%.3f, p-value=%.10f' % (stat, p_value))
```

Figure 4: Mann Whitney U test for comparing major key and minor key

5) Energy is believed to largely reflect the “loudness” of a song. Can you substantiate (or refute) that this is the case? [Suggestion: Include a scatter plot]

The Pearson correlation coefficient between loudness and energy is 0.77. This means that they are highly correlated. From Figure 5 it is observed that they may not have a linear relationship, so I performed a Spearman’s r correlation calculation as well, yielding 0.73. Although it is lower than that for linear relationships, it is still a very high correlation.

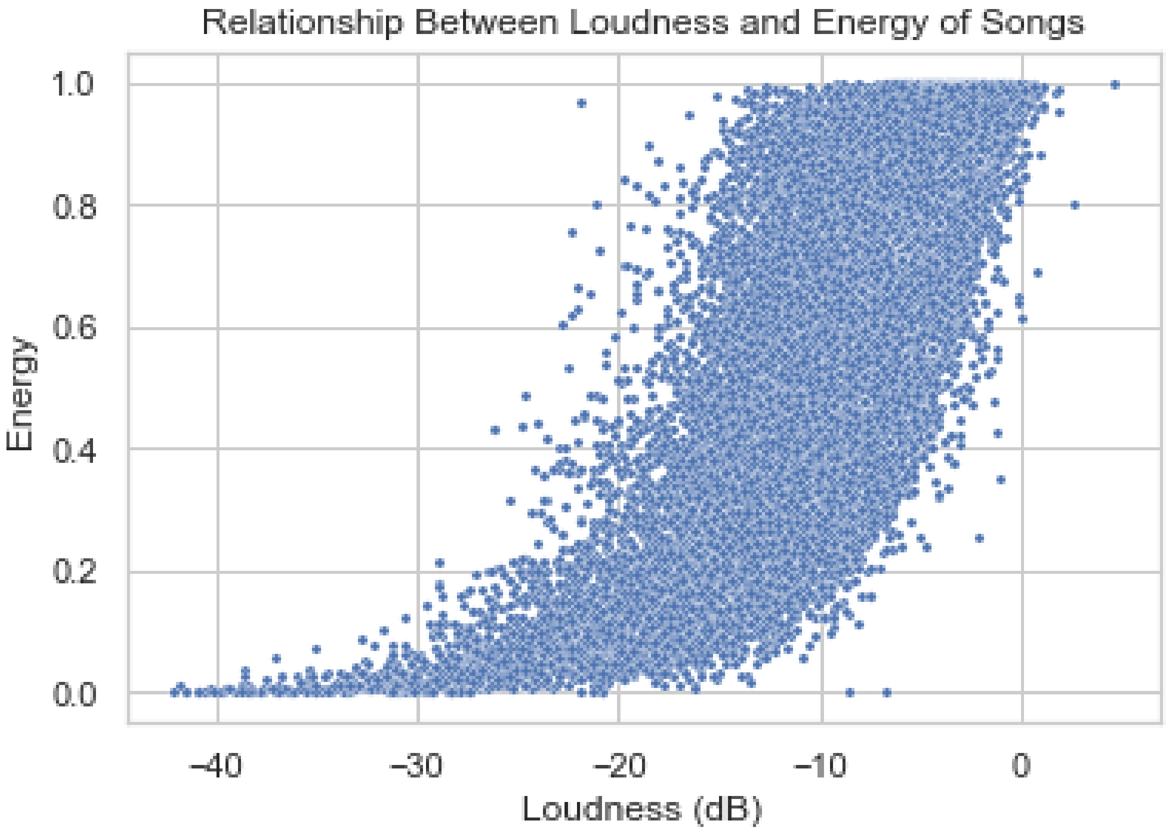
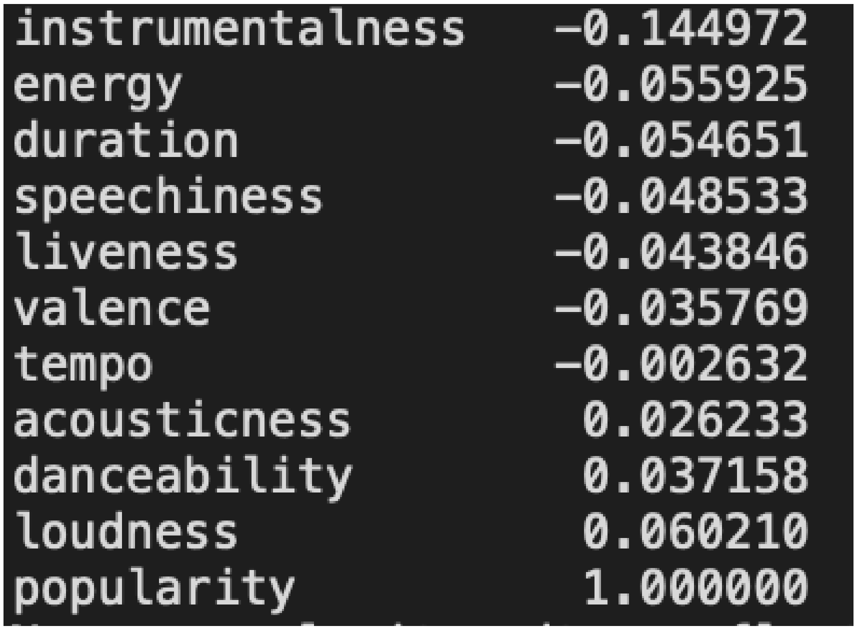


Figure 5: scatter plot of Energy vs. Loudness

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

After calculating the Pearson’s correlation coefficient for each of the 10 song features, with results in Figure 6, it appears that Instrumentalness predicts popularity the best, with a correlation coefficient of -0.144972, as seen from the image below. To evaluate how good this best model is, I created a linear regression model for predicting popularity from instrumentalness. This “best” model has a r-squared value of 0.0211, which is small, so this “best” model is still not good.



```
instrumentalness   -0.144972
energy             -0.055925
duration           -0.054651
speechiness        -0.048533
liveness           -0.043846
valence            -0.035769
tempo              -0.002632
acousticness        0.026233
danceability        0.037158
loudness            0.060210
popularity          1.000000
```

Figure 6: correlation of the 10 individual song features

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?

A multiple linear regression is suitable for this case. After building and evaluating the model, the mean squared error turned out to be 445.689, and the r-squared value was 0.05235. Compared to the r-squared value from the model in question 6, this multiple linear regression has improved, but not by a lot. This improvement may be because more parameters have been used for prediction.

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 10 song features and created the scree plot in Figure 7. By the elbow criterion, there is 1 meaningful principal component, accounting for 27.34% of the explained variance. By the Kaiser criterion, there are 3 meaningful components, accounting for 57.36%. By the third criterion for the number of factors that account for 90% of the variance, I have 7 meaningful components for 90.84% of variance accounted for. I choose to use the Kaiser criterion with 3 meaningful components.

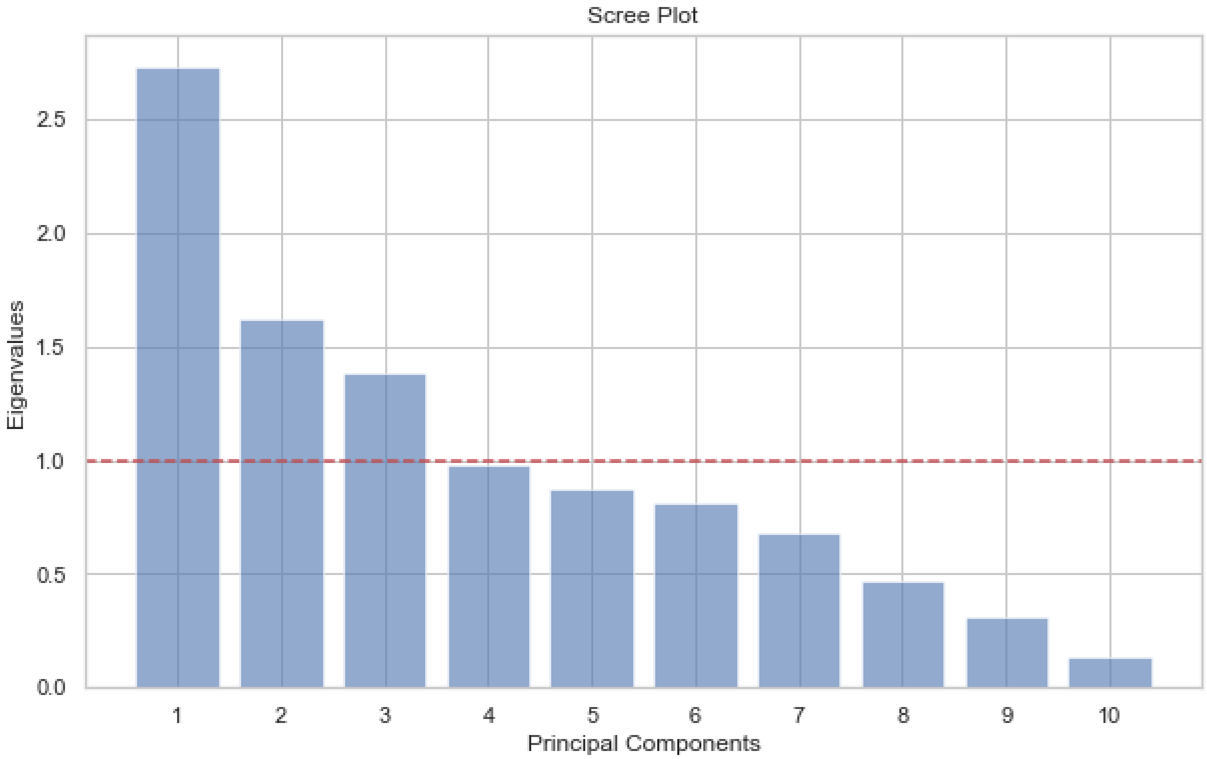


Figure 7: Scree plot of the principal components

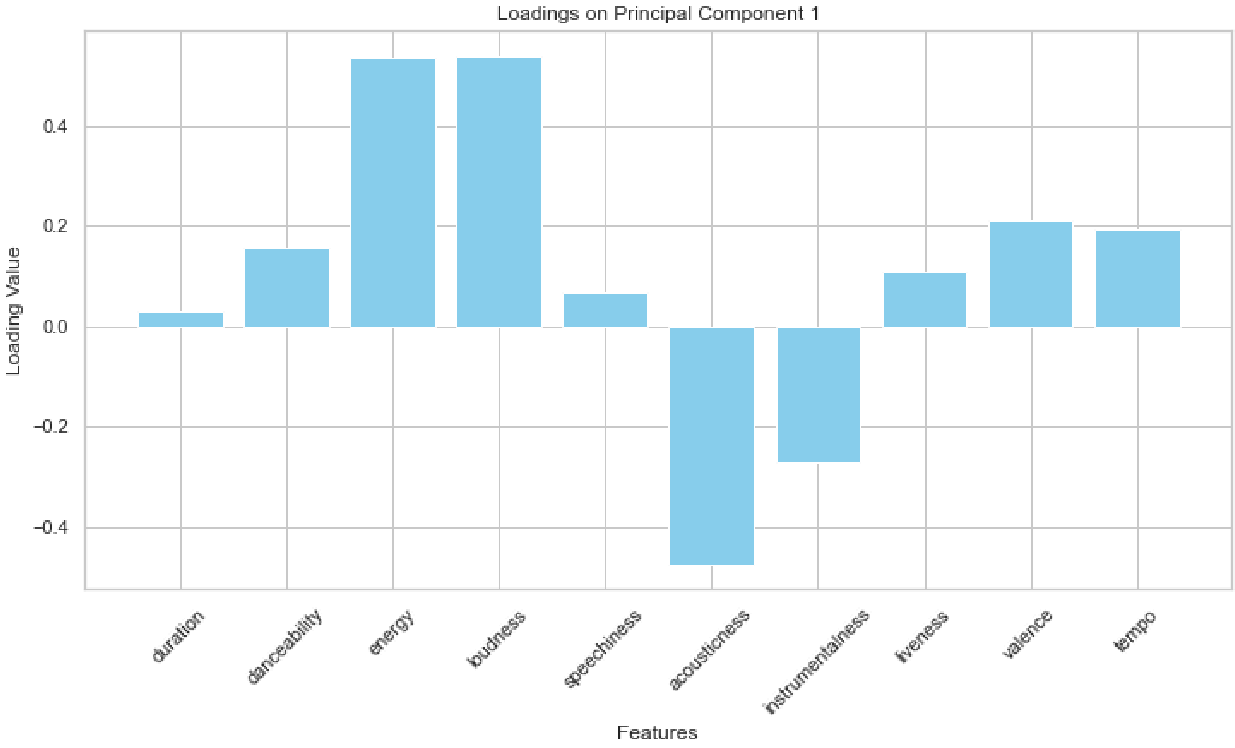


Figure 8: Loadings on Principal Component 1

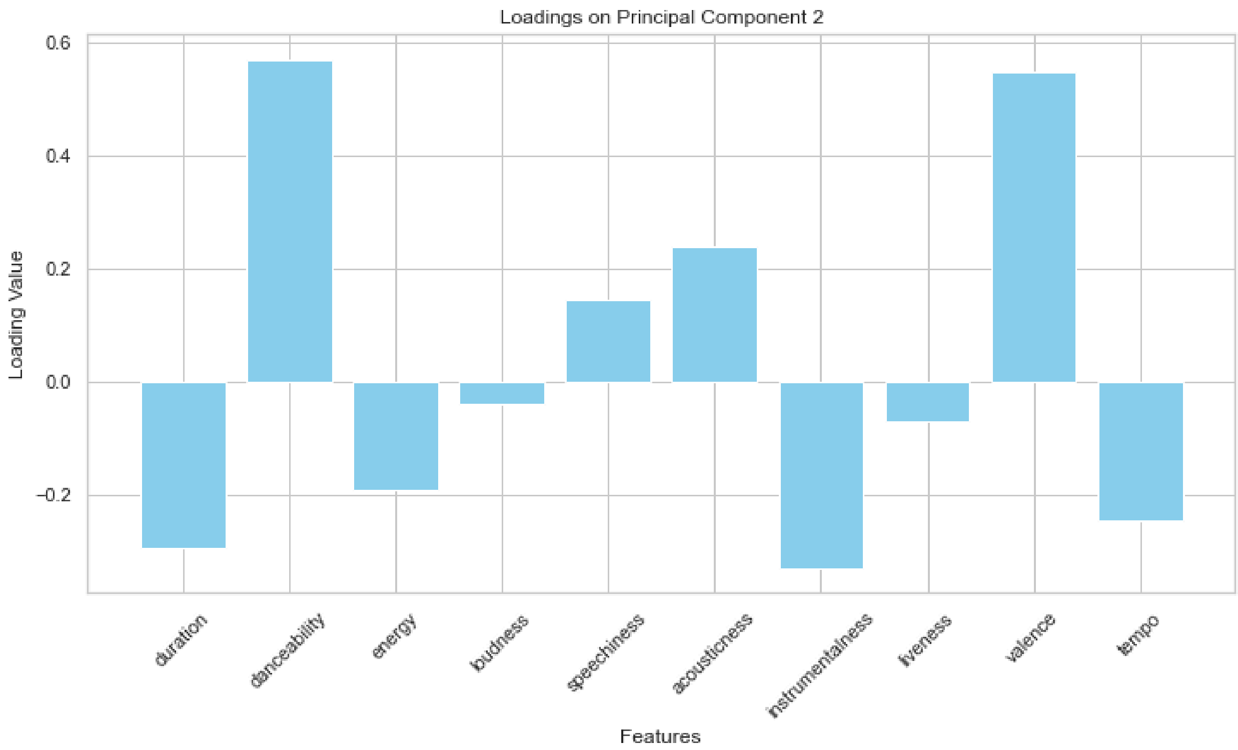


Figure 9: Loadings on Principal Component 2

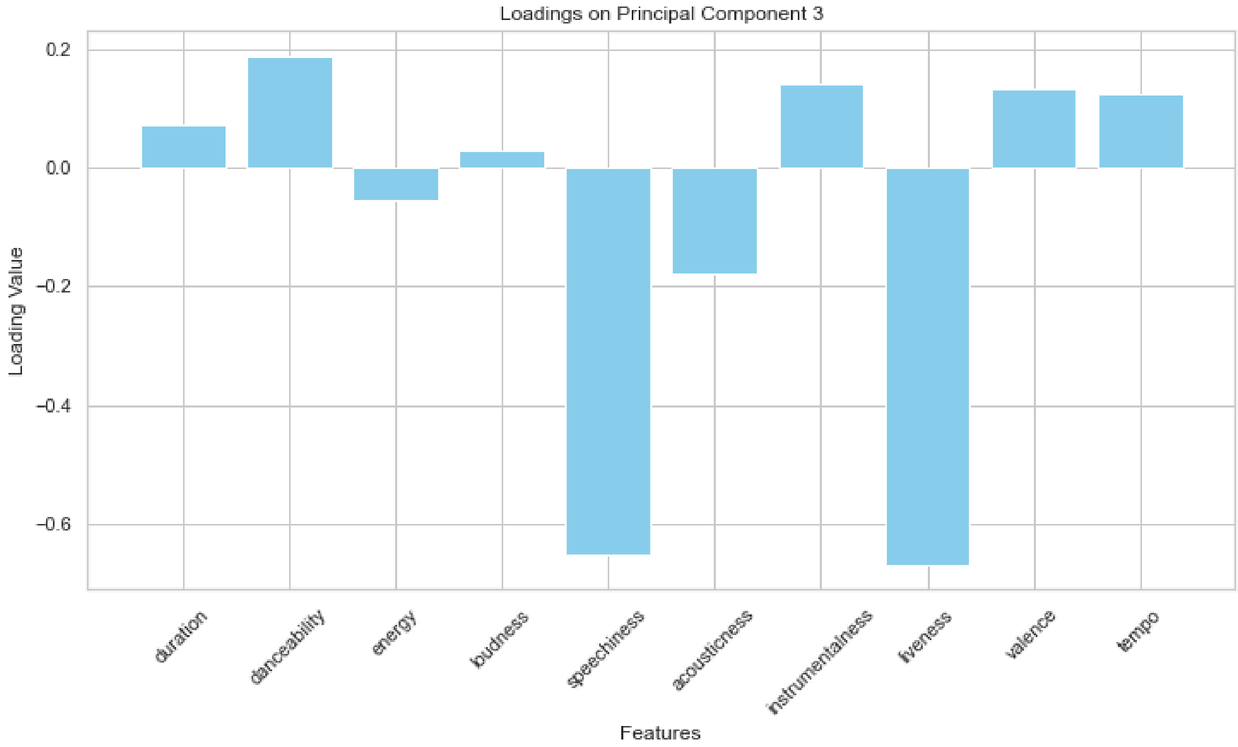
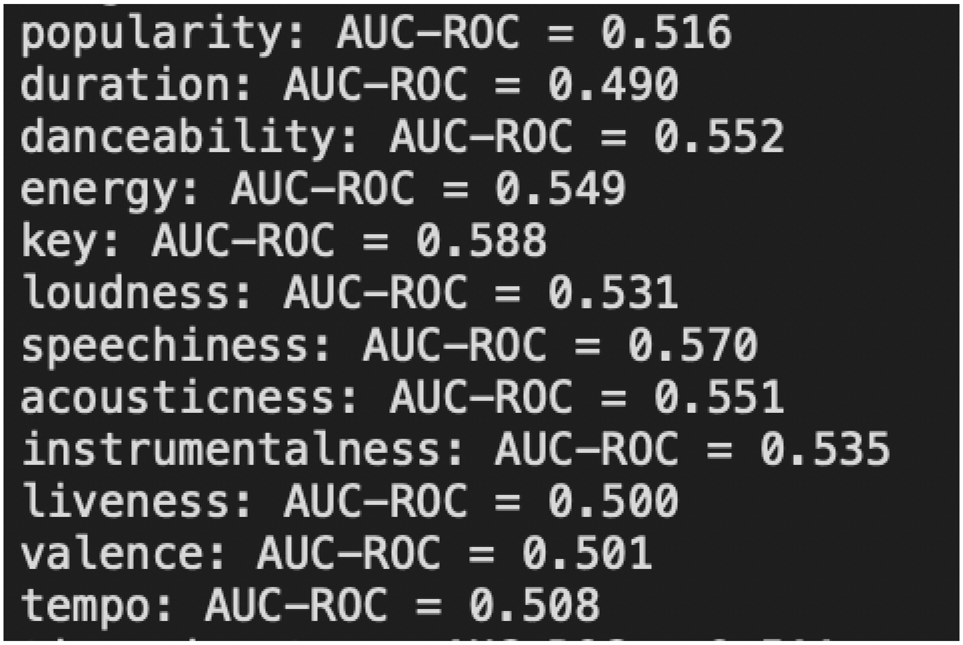


Figure 10: Loadings on Principal Component 3

For principal component 1, it is clear that energy and loudness are highly correlated, as discovered in question 5. With accousticness and instrumentalness having negative loadings. Therefore, this may be accounting for the **intensity** of the music. For principal component 2, the loadings show that danceability and valance are highly positive, while duration, energy, instrumentalness, and tempo are all notably negative. This may account for the **emotion** of the music, for danceability and valence are highly correlated in contrast to duration, energy, tempo, and instrumentalness. The third principal component results in speechiness and liveness highly correlated in contrast to other features. This may be accounting for the **vibrancy** of the song, encapsulating the energetic and dynamic quality while also incorporating the presence of spoken words.

9) Can you predict whether a song is in a 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]

Yes. The AUC-ROC score of valence predicting the key of a song is 0.501. To find a better predictor, we do an exploratory analysis. Selecting from the 10 features, speechiness has an AUC-ROC score of 0.570, the highest among all. Therefore, speechiness is a better predictor. So are other features that have AUC-ROC scores that are higher than 0.501.



```
popularity: AUC-ROC = 0.516
duration: AUC-ROC = 0.490
danceability: AUC-ROC = 0.552
energy: AUC-ROC = 0.549
key: AUC-ROC = 0.588
loudness: AUC-ROC = 0.531
speechiness: AUC-ROC = 0.570
acousticness: AUC-ROC = 0.551
instrumentalness: AUC-ROC = 0.535
liveness: AUC-ROC = 0.500
valence: AUC-ROC = 0.501
tempo: AUC-ROC = 0.508
```

Figure 11: AUC-ROC scores of other columns, including the 10 features.

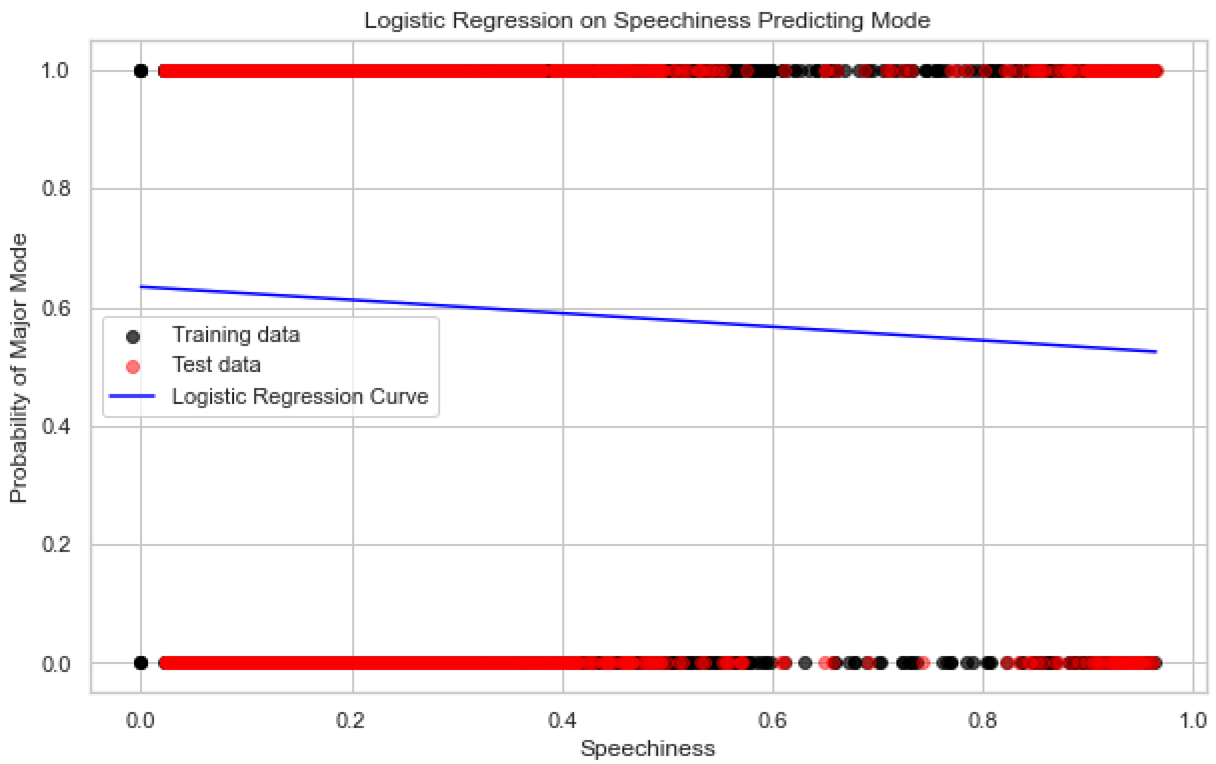
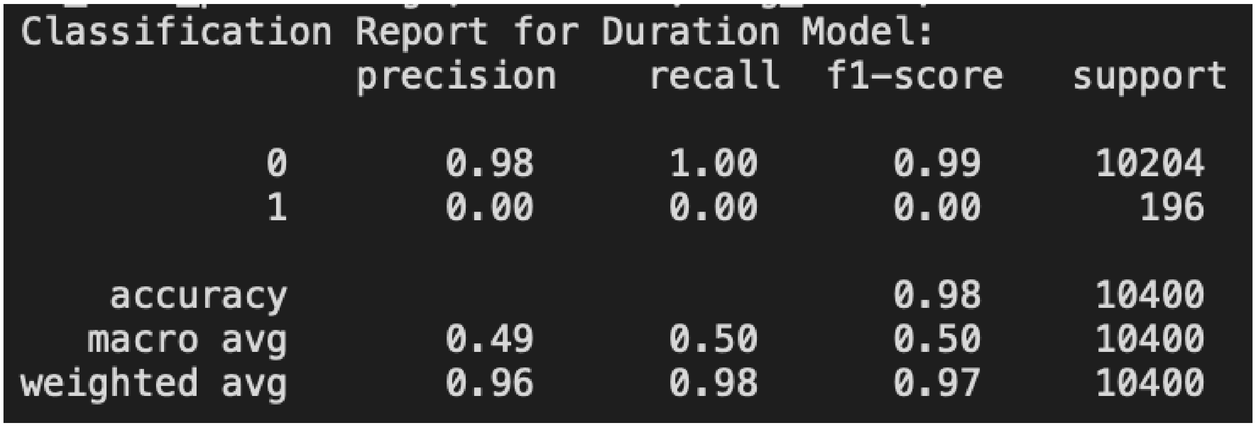


Figure 12: logistic regression model

10) Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8? [Suggestion: You might have to convert the qualitative genre label to a binary numerical label (classical or not)]

The Principal component from question 8 is a better predictor of classical music. After calculating the AUC-ROC scores of the two models, I found that duration has an AUC-ROC score of 0.5582, while PCA with 3 features has an AUC-ROC score of 0.9404. Since the score for PCA is higher, the principal components are a better predictor. The classification reports are Figures 13 and 14.

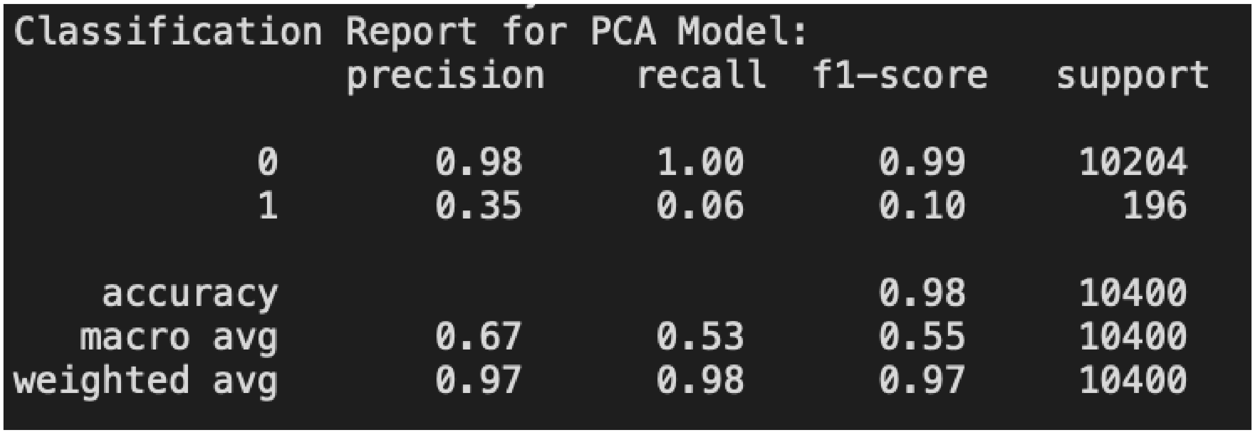


```
Classification Report for Duration Model:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10204
           1       0.00      0.00      0.00       196

    accuracy                           0.98     10400
   macro avg       0.49      0.50      0.50     10400
weighted avg       0.96      0.98      0.97     10400
```

Figure 13: classification report for duration model



```
Classification Report for PCA Model:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10204
           1       0.35      0.06      0.10       196

    accuracy                           0.98     10400
   macro avg       0.67      0.53      0.55     10400
weighted avg       0.97      0.98      0.97     10400
```

Figure 14: classification report for PCA model

Extra credit:

After graphing the distribution of beats per measure in a bar graph, I found that the most common time signature in the data set is 4, as seen in Figure 15.

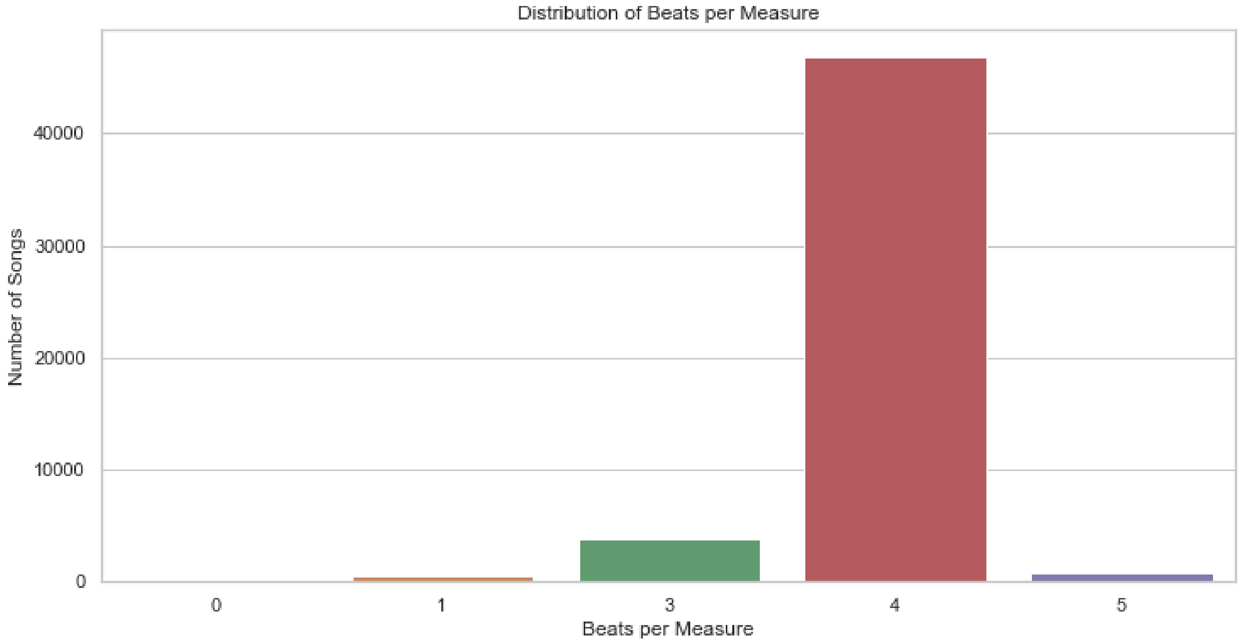


Figure 15: Distribution of Beats per Measure

From my understanding of beats per measure, I wonder if it is correlated with tempo. So I performed a Linear Regression and found the following results. I yielded a mean squared error of 844.47, and the R-squared score was 0.000283, which is very small. Thus I can conclude that there is no notable linear relationship between beats per measure and tempo.