Rithvik Kottapalli

PODS Final Project

Preprocessing for each problem was done through pandas or raw arrays. I searched for null inputs and made sure the data was in the correct form (str, float, etc.). Filtering was done through Pandas lambda functions or basic for-loops. Each section of the code is separated, labeled, and commented out except for when it needs to be used.

1. On this problem, looking through the description of the data I determined the only columns worth checking for the normal distribution were columns [5,6,8,9,11,13,14,15,16,17,18], as these were the only columns that could have any sort of numerical distribution. After performing EDA on all columns, I determined none were distributed normally. This makes sense as there is a very low chance any of these variables were made by completely random and independent preferences by the user or choices by the artists. I did however find that column 8 (danceability) closely resembled a Beta distribution (Figure 1), and column 6 (duration) closely resembled a Poisson distribution (Figure 2).

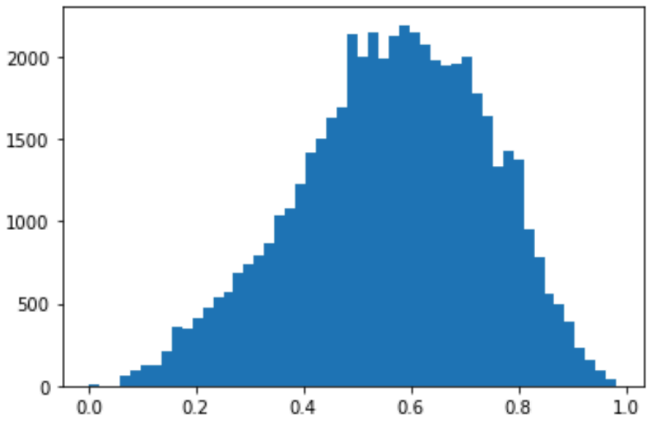

Figure 1

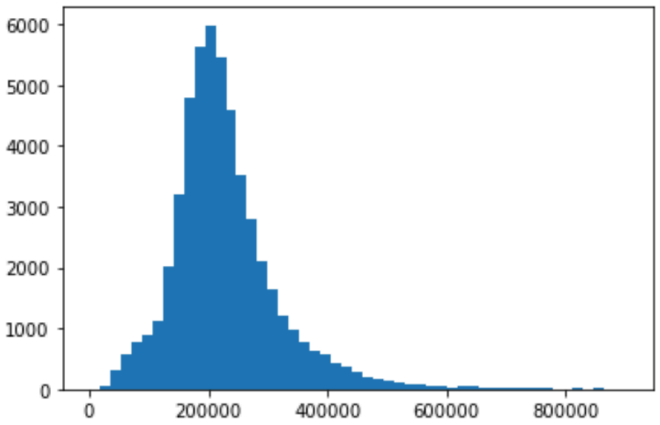

Figure 2

2. I can confidently say there is very little correlation if any between the duration of a song and the popularity. Our correlation value is -0.055, indicating a very slight negative

correlation, which would mean that shorter songs tend to be more popular, however, there are many more short songs (see Figure 2) so this assumption may be flawed. Duration and popularity are not strongly correlated. The scatterplot is represented in Figure 4, and the same scatterplot is represented in Figure 3 with songs longer than 1 million ms cut off for ease of viewing. It is worth noting that no song longer than 800,000 ms has a popularity rating above 60.

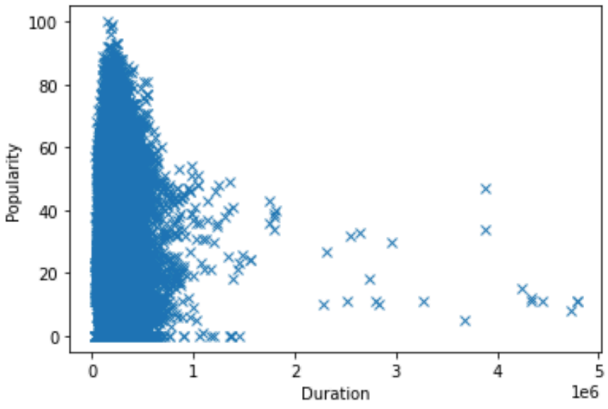

Figure 3

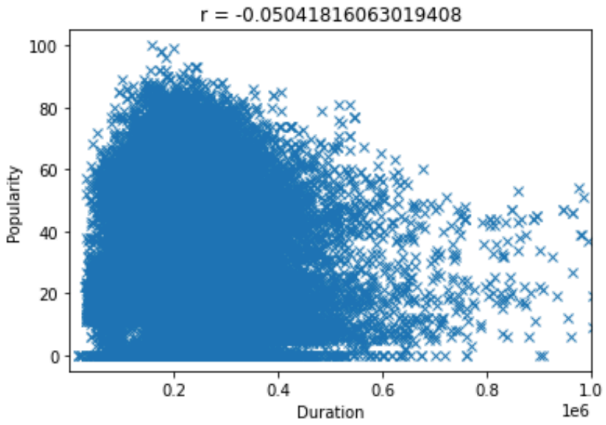

Figure 4

3. To determine if one group is more popular than another, we need to do a significance test. To pick the significance test, I determined that for popularity, it was not reasonable to reduce the population to a sample mean as the underlying distribution for probability is more heavily influenced by very unpopular (near 0 rated) songs. The data is not categorical, there are 2 groups to compare (explicit and non-explicit), and it would make sense to compare their medians as the underlying distribution is not heavily skewed. Therefore, I used the Mann-Whitney U test to test whether the populations were significantly different. There were 46403 non-explicit songs and 5597 explicit songs. The significance test showed an extremely low p-value of $3.07 * 10^{-19}$ which is much lower than our alpha of 0.05. This shows that we must reject the null hypothesis that explicit and non-explicit songs are rated the same. After confirming this, EDA and central

tendency measures show that explicit songs are rated more highly than non-explicit songs.

4. Using the same logic as in the previous problem, I also decided to use the Mann-Whitney U test for this problem. There are 32391 major key songs and 19609 minor key songs. The p-value was $2.02 * 10^{-6}$, which is much lower than our alpha. We can reject the null hypothesis that there is no difference between their popularities. Using central tendencies and EDA we determine that songs in the minor key are more popular than songs in the major key.
5. Testing the correlation between energy and loudness, we see that they are extremely correlated, as seen in Figure 5 with an r of 0.77. This makes our $R^2 = 0.60$, meaning that energy accounts for 60% of the variance in the loudness level. The higher the energy of the song, the higher the loudness will likely be.

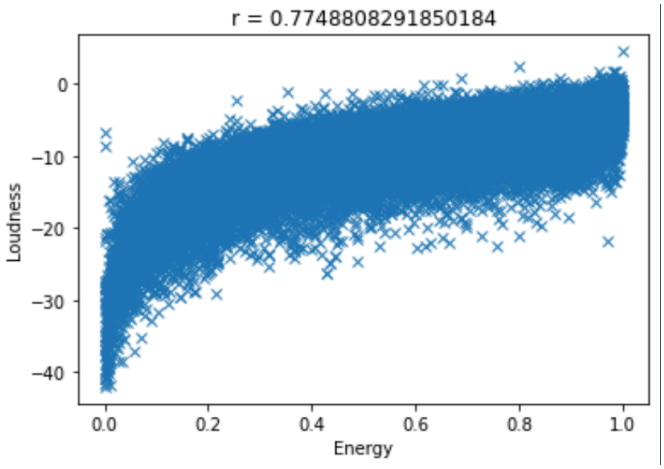

Figure 5

6. To solve this problem, I created a linear regression model to predict popularity given each of the 10 variables. I made sure to create a suitable train test split to avoid predicting with training data. After testing all 10 variables with a for-loop in the linear regression model, the single column that is the best predictor of popularity is instrumentalness, however, even this alone is a very poor predictor, with an r^2 of only

0.02 as shown in Figure 14, meaning it alone can only account of 2 percent of the variance of popularity. No single variable is a good predictor of popularity.

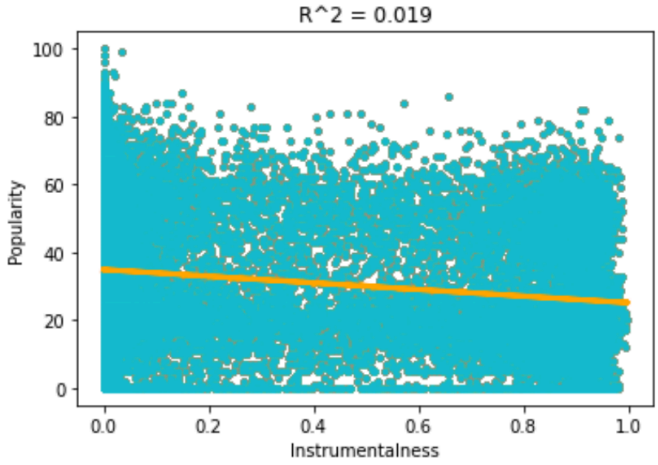

Figure 14

7. Creating another linear regression model using the same steps as outlined above, however, including all 10 predictors gives us an R^2 of 0.048. This is still extremely poor, considering how much more data we processed for an improvement of only 3 percent variance over our single best predictor. Even using all 10 variables, 95% of the variance is unaccounted for.
8. To start PCA, first I plot all correlations between the 10 variables on a heatmap. Right off the bat, I see some important correlations to keep an eye out for, such as a strong positive correlation between energy and loudness as discovered before, a strong negative correlation between energy and acousticness, and a strong positive correlation between danceability and acousticness. There are a few more slightly weaker but still important correlations also visible as seen in the heatmap in Figure 6, which can be understood by the key in Figure 7. The next step is to z-score our data to get it ready for PCA. Stats.zscore() was throwing an error so I used a pandas function to do it. Finally, we can run PCA. After running PCA, there was no clear elbow, so I used the Kaiser criterion to choose my principal components, of which there were 3 that exceeded an eigenvalue of 1, as seen in Figure 8.

Inspecting the Loadings matrices for each of these 3 principle components to identify them, we find that factor 1 is mostly Energy, Loudness, and lack of acousticness, as seen in Figure 9. This means it quantifies how hard a song goes, the average sound in dB, and how synthesized it is. I will interpret this factor as “Hardstyle”, as this factor would be maximized in the hardstyle genre. Factor 2 is determined mostly by Danceability and Valence, as seen in Figure 10. I will characterize this factor as “Excitement” as it encapsulates positivity and danceability. Factor 3 is determined mostly by lack of speechiness and lack of liveness, as seen in Figure 11. I will characterize this factor as “Ambience”, as this encapsulates more instrumental/non-lyrical music in a quiet setting. These 3 factors, “Hardstyle”, “Excitement”, and “Ambience” account for 57.36% of the variance of songs.

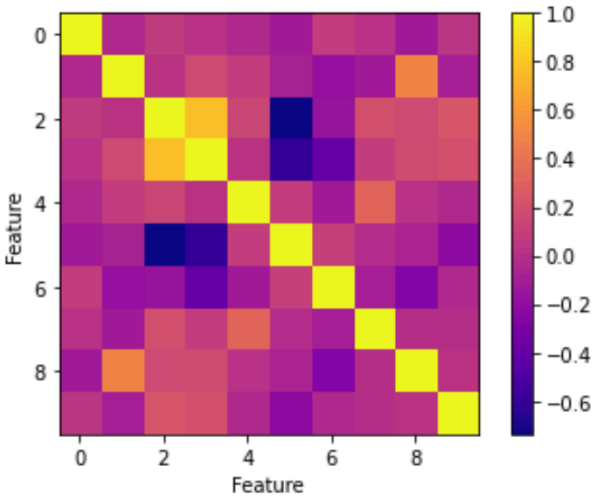

Figure 6

```
'''
0  Duration
1  Danceability
2  Energy
3  Loudness
4  Speechiness
5  Acousticness
6  Instrumentalness
7  Liveness
8  Valence
9  Tempo
'''
```

Figure 7

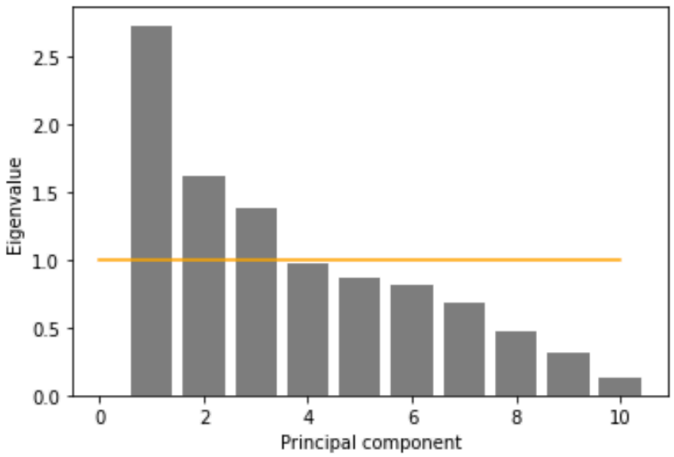

Figure 8

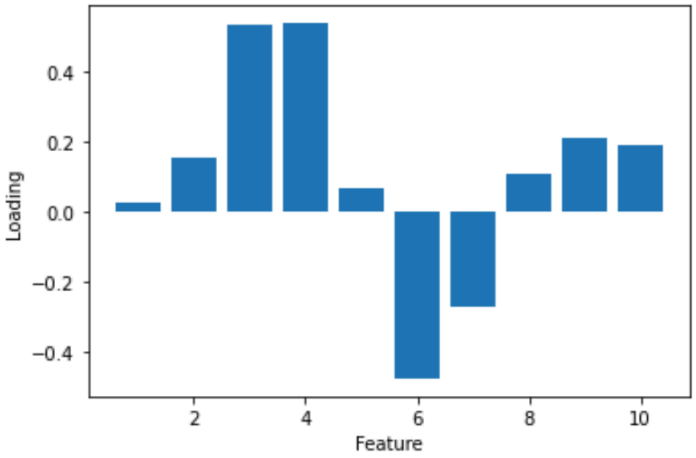

Figure 9

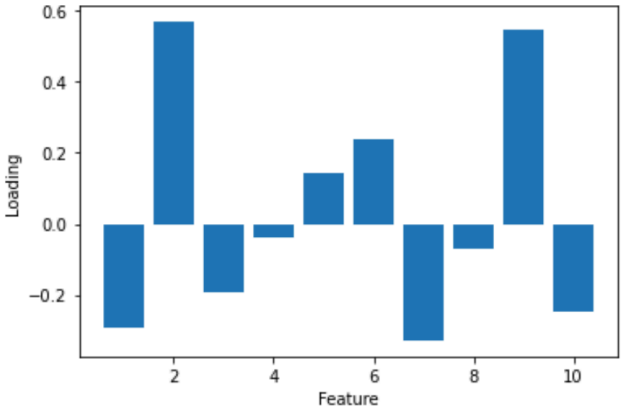

Figure 10

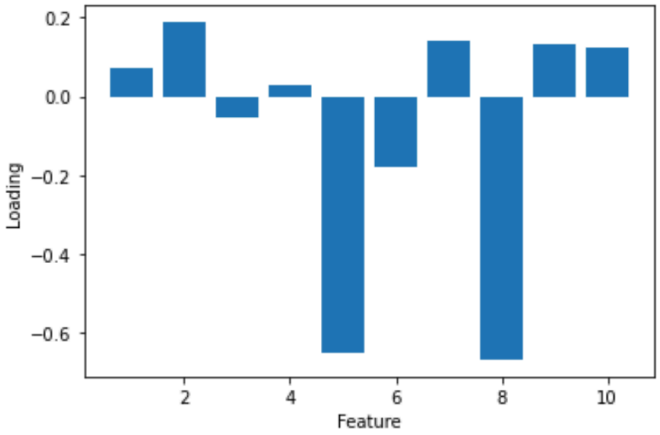

Figure 11

9. First, I'll plot the data to see if there is a clear cutoff. Since this is a binary classification, I will use logistic regression. From the plot, we can see that there is no clear cutoff/criteria that relates valence to mode as seen in Figure 12. We can predict that the logistic regression will not be very good. After running the Logistic Regression as seen in Figure 13 (0 is minor and 1 is major) and testing it with a proper train/test split, we achieved an AUC score of 0.5056. This indicates that our model is exceptionally bad, being only 0.56% better than randomly guessing. We can conclude that valence is not a good

predictor of mode. Speechiness and Acousticness were both better predictors than Valence testing at an AUC-ROC of 0.56 for both. This is still a relatively poor performance but 6% better than random guessing. Their sigmoid curves were not more interesting than the one below so I have omitted them.

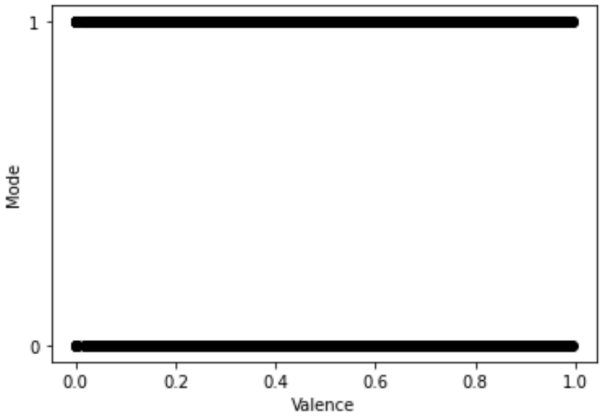

Figure 12

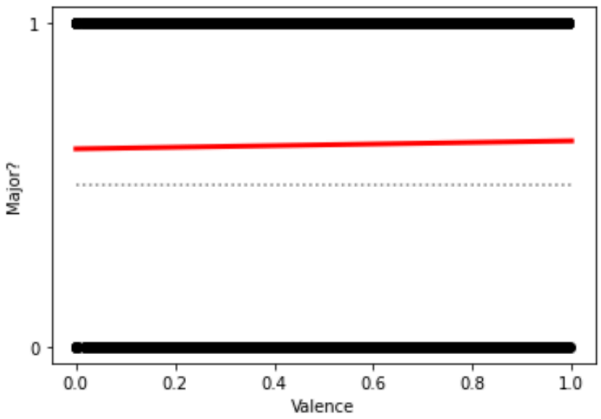

Figure 13

10. First I write a function to transform the genre column from a column containing many strings to a column containing a 1 if it was classical, and 0 otherwise. Then I use the steps from the previous question to run logistic regression with my predictor as duration. I achieved an AUC-ROC Score of 0.63, meaning duration can be used to predict if music is classical or not at a demonstrably better rate than randomly guessing. Next, I run logistic regression using the 3 principal components I extracted in question 8. I make sure to zscore the data again, then I use the first 3 principal components as my predictors in Logistic Regression. The AUC-ROC is exceptional in this model, as it achieved a 0.953. Using the 3 principal components leads to an almost perfect prediction of whether or not a song is classical. Thus we can conclude that using the PCA components leads to a better prediction than duration.
11. For the extra credit, I wanted to see if the key of a song was related to its danceability, and if certain keys were more danceable than others. To do this I employed a Kruskal-Wallis Test, as I had multiple groups with non-categorical data and wanted to

test the underlying distributions. First I performed EDA, and the median danceability for each key varied from 0.541 to 0.609, which was a slight difference but enough to warrant a Kruskal Wallis test to test the underlying distributions. After grouping each key into its own array, I ran a Kruskal-Wallis test on all 12 groups and found an exceptionally low p-value (on the order of 10^{-81}). I ran the test multiple times on multiple different smaller combinations of groups as well, and each time I ended up with exceptionally small P-values. There is virtually no chance that the danceability is independent of the key, and using EDA and central tendency measures, I can see that the most danceable key is most likely G# and the least danceable key is most likely B.