Author: Kai-Hsuan Chan

PODS Capstone Project

**Data Processing**

1. **Dimension Reduction:**

For dimension reduction, I first extracted the variables that are needed, then z-scored the data. After that, I performed PCA on the data. To determine the representative principal components, I used the Kaiser Criterion, and looked for components with eigenvalue>1.

1. **Data Cleaning:**

For data cleaning, I decided to do a row-wise removal for rows that contain NaN values. However, there doesn’t seem to be any rows in the data that contains NaN values, which suggests that the data has already been cleaned before I received it.

1. **Data Transformation:**

Data transformation was used when z-scoring the data for PCA. In addition, the loadings matrix and the data in the new coordination system (rotatedData) from the PCA was multiplied by -1 to address polarity issues. The genre column was also transformed to numerical values in question 10 to quantify the data.

1. **RNG Seeding:**

All the random number generators in this analysis were seeded with my NYU N-Number.

**Question 1)**

This is straightforward, I extracted the ten song features, then I plotted a histogram for each of them. As shown below:

A group of blue and black graphs

Description automatically generated

Looking at the histograms, danceability and tempo seems to be reasonably normally distributed. To confirm this, I looked at the median and mean of both. The mean of danceability is 0.5639 and the median is 0.574. The mean of tempo is 123.558 and the median is 123.366. Although mean and median alone do not characterize a normal distribution, this is supportive of our assumption as a similar mean and median follows the behavior of a normal distribution.

**Question 2)**

Before determining the relationship between song length and popularity, I performed an EDA by plotting a scatterplot between the two variables. As shown below:

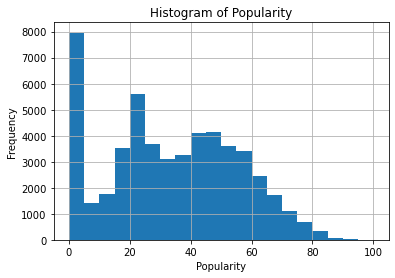
A graph of blue dots

Description automatically generated

There doesn’t seem to be a relationship between the two variables by observing the scatterplot. For further confirmation, the Pearson correlation between the two variables is -0.0546, with a p-value of 1.06916e-35. Therefore, we can conclude that there isn’t really a relationship between the variables (at least linearly speaking).

**Question 3)**

Before conducting a significance test, I did an EDA by plotting a histogram of popularity:

There is a spike for songs where the popularity is zero. This makes sense, as there are a lot of songs that people aren’t listening to.

Due to the spike in the distribution, I decided to do a Mann-Whitney U test and picked an alpha value of 0.05. This is a reasonable decision as the Mann-Whitney U test relies on the median, which is more robust in this case.

I then spliced the dataset into explicitly rated songs and non-explicitly rated songs. The median popularity is 34 for explicitly rated songs and 33 for non-explicitly rated songs. Therefore, I hypothesized that explicitly rated songs are more popular than non-explicitly rated songs. I decided to do a one-tailed test since directionality is specified in my hypothesis. After running the Mann Whitney U test, I got a p-value of 1.533959e-19, which is less than 0.05. Therefore, I concluded that explicitly rated songs are more popular than non-explicitly rated songs at the 5 percent significance level.

**Question 4)**

Following the reasoning in question 3, I decided to use a one-tailed Mann-Whitney U test again, with an alpha value of 0.05. The median popularity is 32 for major Songs and 34 for minor songs. Therefore, I hypothesized that minor songs are more popular than major songs. I decided to do a one-sided test as directionality is specified in my hypothesis. After running the test, I got a p-value of 1.00876e-06, which is less than 0.05. Therefore, I concluded that minor songs are more popular than major songs at the 5 percent significance level.

**Question 5)**

I first visualized the data in a scatterplot. Shown below:

A graph of blue dots

Description automatically generated

I noticed that the relation doesn’t seem to be entirely linear. However, it does display a monotonic trend. Therefore, I decided to find the spearman correlation between them. The calculations yielded a correlation coefficient of 0.73 between the two variables. This is a strong correlation, which confirms that the energy of a song does largely reflect its loudness.

**Question 6)**

To see which song feature predicts popularity the most, I built a linear regression model for each of the features. The models were trained and evaluated with an 80-20 train-test split for cross validation. I then compared the models based on their respective R-squared scores:

A graph with blue squares

Description automatically generated

The result shows that instrumentalness has the highest R-squared score among the ten features. However, this does not mean that instrumentalness is a good predictor for popularity, as its R-squared score is only 0.02, meaning that it only accounts for 2% of the variance in popularity. This suggests that popularity cannot be determined by a single variable and requires multiple factors to be accurately predicted.

**Question 7)**

To incorporate all the song features, I built a multiple linear regression model. To assess how good the model performs, I did an 80-20 train-test split for cross validation. After cross validation, the model has an R-squared score of 0.0419, which is double the R-squared score for using instrumentalness as a predictor alone. However, even though we used all the features to predict popularity, the model isn’t performing well, since it only accounts for 4% of the variance in popularity. This implies that the dataset we have lacks important information needed to predict popularity and that popularity might require more complex models to accurately predict.

**Question 8)**

A graph with a red line

Description automatically generatedBefore running the PCA on the song features, I first z-scored the data. This is a crucial step, since the PCA is extremely sensitive to units of measurement. I decided to use the kaiser criterion and looked at the eigenvalues of the principal components to determine the relevant ones:

The histogram shows that there are 3 principal components with eigenvalue>1. Therefore, by the kaiser criterion, there are 3 relevant components.

Next, I calculated the variance accounted for by the 3 components, which is 57.359. This can be done by dividing the sum of their eigenvalues by the sum of the eigenvalues of all the principal components. Then, I looked at the loadings matrix of the 3 components to try and understand what they represent:

A graph of different sizes and colors

Description automatically generated with medium confidence  
I interpreted the first component being “Vitality”, the second being “Positivity”, and the third being “How emotional?”.

Now that I have determined the relevant components, I ran a k-means clustering algorithm to classify the datapoints. Before doing so, I used the silhouette method to determine the optimal K:

A graph with a line

Description automatically generatedThe silhouette method indicated that 2 is the optimal K.

A graph with a blue and orange color

Description automatically generatedThen, I ran the k-means clustering algorithm with k=2. Here is the visualized data with two clusters in 3d space (since we have 3 principal components):

I interpreted the orange cluster as being songs that are more emotional, and the blue cluster as being songs that are less emotional. However, this is just my interpretation, and it is difficult to fully understand the meaning behind the clusters without further analysis.

**Question 9)**

Since predicting if a song is major or minor is a binary outcome, I decided to build a logistic regression model to see if valence accurately predicts the mode of a song. I performed a simple EDA by plotting the datapoints to see if there is difference in valence between songs that are major and songs that are minor:

A rectangular object with black circles

Description automatically generated with medium confidenceInspecting the scatterplot, I don’t see a discrepancy between songs that are minor and songs that are major. Therefore, I wouldn’t expect the logistic regression model to perform well.

Then, I built the logistic regression model. To assess the model, I plotted the ROC curve:  
A graph of a line

Description automatically generatedThe curve is nearly identical to a diagonal line with a slope of 0.45, and the area under it is only 0.5. This means that the model is basically randomly guessing the outcome, and confirms my assumption that valence isn’t a good predictor for the mode of a song.

In addition, I tried using the other features as predictors. However, none of them accurately predicts the mode of the song.

A blue bar graph with white text

Description automatically generatedThe AUC score for all of the song features are close to 0.5, indicating that these aren’t good predictors for the mode.

**Question 10)**

Before building the model, I used a label encoder to convert the genre column into numerical values, ranging from 1-52. Then, I used the 10 song features to build a classification tree using the random forest classifier in sklearn. To assess the performance, an 80-20 train-test split was conducted. This resulted in an accuracy score of 0.36875, meaning that the model correctly predicts the genre about 1/3 of the time. Considering that there are 52 genres to predict, the performance of the model is quite good. However, an accuracy of 0.36 is far from perfect, and we might need more relevant data to improve the prediction.