Spotify Music Popularity Analysis & Predictive Modeling
Hairong (Rona) Zhang

Preprocessing:

1) Dimension reduction:

   Before analyzing data for some steps, we need to generate the dimension reduction of the dataset. Here I used **Principal Components Analysis (PCA)** to accomplish the dimension reduction. In order to do a PCA successfully, the dataset should be **standardized through z-score normalization**. Then the PCA was applied to the dataset, which transformed the original variables into a new set of variables, the principal components. They were ordered by the amount of variance they capture from the data, or the eigenvalues. Based on the graph of new set of variables, for deciding the number of factors or principal components to retain, I chose to use Kaiser criterion, which means only those components with eigenvalues greater than 1 should be chosen.

2) Data cleaning:

   To get the dataset from the excel, we used *read_csv* in pandas. Then we got the dataset of all songs' information, which was named *spotify_data*. And to handle missing data, here I cleaned the dataset by row-wise removal. To be more specific, I used *dropna()* to remove rows that have any missing values. After the removal, the cleaned dataset was named *spotify_data_cleaned.*

3) Data transformations:

   Since PCA for dimension reduction is influenced by the scale of the features, it is important to standardize the dataset first. So, the dataset was transformed by z-score normalization firstly. After applying PCA, the dataset is rotated into a new coordinate system.

Seeded RNG:

To protect my work and doing analysis, I seeded the random number generator with my unique N-number. To do that, I used *random.seed(n)* from random. Then there will always be a unique random number for my answers. And when I use my N-number for the *random_state* when splitting dataset, the results of my work will also be stable and unique.

In order to analyze the distributions of the 10 song features, including duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo, I used matplotlib and seaborn to generate histograms for each of the 10 song features in a 2x5 figure as below (Figure.1). Then we can use these graphs to determine if any of them appear to be reasonably normally distributed. Let's analyze each of them one by one.

i. **Duration:** The distribution seems skewed, so it does not appear to be normally distributed. The histogram shows that most songs having shorter durations.
ii. **Danceability**: The distribution seems somewhat bell-shaped with a very slight skew. So, it is not perfectly normal distribution, but it is relatively close to it.
iii. **Energy**: The distribution is skewed and does not show a normal distribution.
iv. **Loudness**: The distribution is somewhat bell-shaped but is still slightly skewed.
v. **Speechiness**: The distribution is significantly skewed towards lower values, indicating it is not normally distributed.
vi. **Acousticness**: The distribution also shows a skewed distribution, not normal.
vii. **Instrumentalness**: The distribution is skewed with most values close to zero, indicating a non-normal distribution.
viii. **Liveness**: Skewed towards lower values and is not normally distributed.
ix. **Valence**: The distribution seems somewhat to be normal but is not perfect since some skews.
x. **Tempo**: The distribution seems to be normal since it shows bell-shaped and only very slight skewness.

Overall, danceability and tempo seem to be reasonably distributed normally among all of these features.

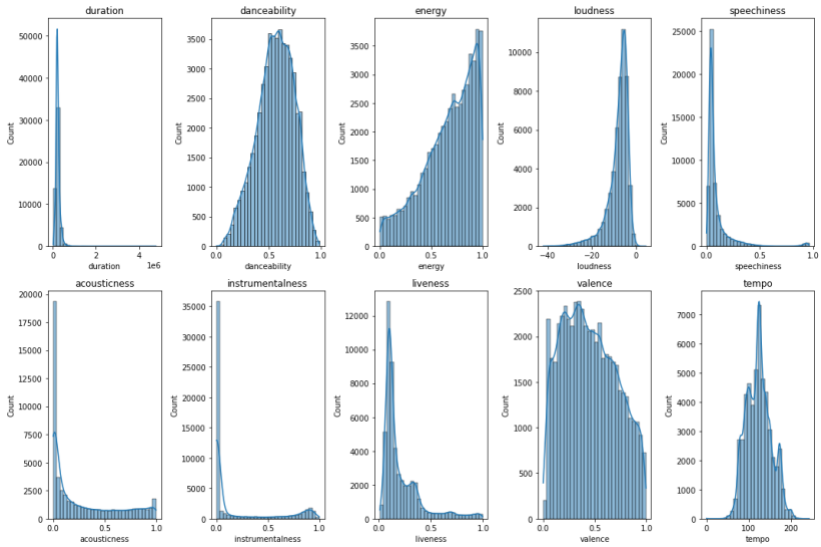


Figure.1

To examine the relationship between song length (duration) and popularity of a song, I generated a scatterplot to show the overall trend of the relationship, where the duration of each song is plotted on the x-axis and the popularity on the y-axis. Based on the figure below (Figure.2) and the **Pearson correlation coefficient** which is about -0.0547, pretty close to 0, there **does not appear to be a clear and strong positive or negative relationship** between song length and popularity. The distribution of these points did not show a consistent trend that songs with longer or shorter duration are more popular.

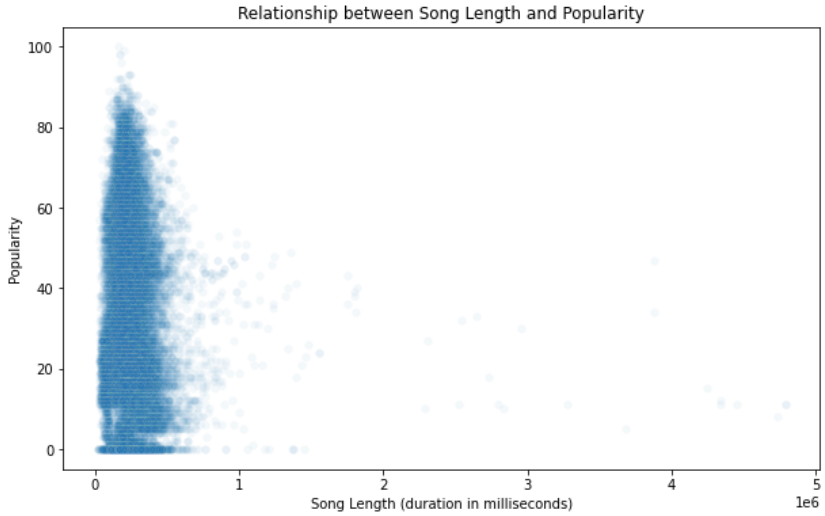
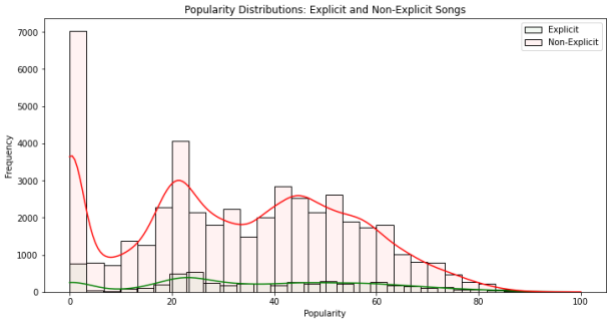


Figure.2

In order to choose the suitable significance test, I found the distributions of popularity for both explicit and non-explicit songs as below (Figure.3). From the figure, we can see that both distributions are skewed and not normal. So, I chose a non-parametric test, **Mann-Whitney U test**, to compare differences between two groups. Then I conducted the test to determine if there is a statistically significant difference in popularity between explicit and non-explicit songs. The null hypothesis is that there is no difference in the distribution of popularity between the two groups. Based on the results (U statistic and p-value) of the test (Figure.4), since p-value of 3.068*10^-19 is much smaller than significance level of 0.05, the null hypothesis has been rejected. So, there is a statistically significant difference in the popularity between explicit and non-explicit songs. And to compare them, I found medians of each distributions. Since the median of popularity for explicit songs is higher than that for non-explicit songs, it suggests that **explicitly rated songs are statistical significantly more popular than songs that are not explicit**.

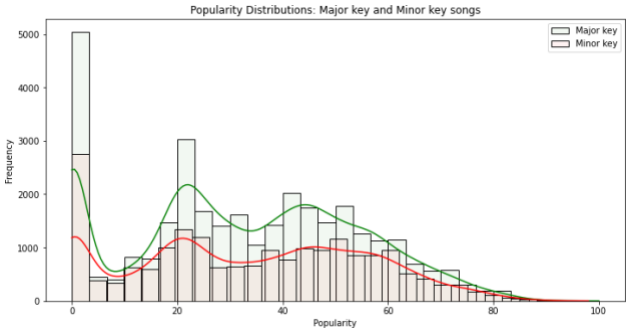



```
U statistic: 139361273.5
p-value: 3.0679199339114678e-19
Median of popularity for explicitly rated songs: 34.0
Median of popularity for non-explicitly rated songs: 33.0
```

Figure.4

Figure.3

Again, I found the distributions of popularity for both songs with major key (mode) and with minor key as below (Figure.5) to choose the suitable significance test. And since both distributions are not normal, I used non-parametric test, **Mann-Whitney U test**, again. The null hypothesis for this test is that there is no difference in the distribution of popularity between songs in major key and in minor key. Based on the results (U statistic and p-value) of the test (Figure.6), since p-value of 2.0175*10^-6 is much smaller than significance level of 0.05, the null hypothesis has been rejected. Then we can conclude that there is a statistically significant difference in the popularity between songs in major key and in minor key. And since the median of popularity for songs in major key is lower than that in minor key, **songs in minor key are more popular than songs in major key**.

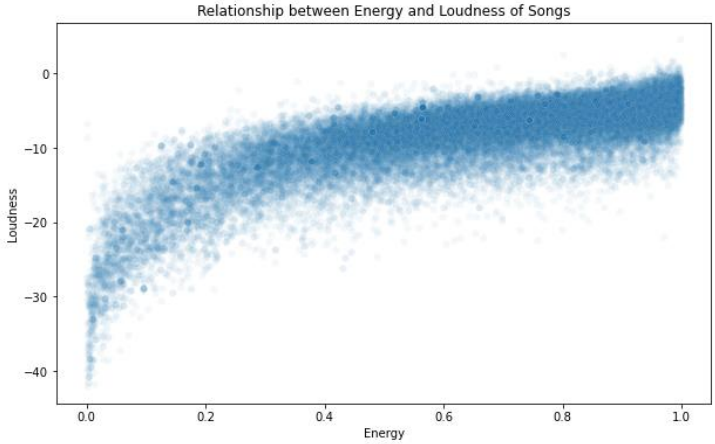



```
U statistic: 309702373.0
p-value: 2.0175287554899416e-06
Median of popularity for songs in major key: 32.0
Median of popularity for songs in minor key: 34.0
```

Figure.6

Figure.5

To find whether energy is believed to largely reflect the “loudness” of a song or not, I tried to find the relationship between energy and loudness by drawing the scatterplot as below (Figure.7). Based on the figure and the **Pearson correlation coefficient** of 0.775 which is close to 1, there shows a positive relationship between energy and loudness. Although the relationship is not perfect positive, we can still conclude that **energy is believed to largely reflect the “loudness” of a song**.




```
Correlation coefficient:0.7748808291850185
```

Figure.7

Which of the 10 song features predicts popularity best, I used **Simple Linear Regression model** for each feature and then evaluate their performance using **Root Mean Square Error (RMSE) and R-squared (R^2) values**. Firstly, I divided the dataset into training and testing sets with *random_state* to be my N-number so that my answers could be unique. Then for each feature, I generated a simple linear regression model and train it by the training sets. And I also used the testing sets to predict popularity. For the evaluation of the prediction performance, I calculated RMSE and R-Squared for each model and stored them in a dictionary named *modelPerformance* for their corresponding features. Finally, I tried to find the feature with the lowest RMSE and the highest R-Squared among all of these 10 features in the dictionary and the results are shown below (Figure.8) that **instrumentalness predicts popularity best** with RMSE of 21.3954 and R-Squared of 0.0245.

```
{'duration': {'RMSE': 21.62770780163697, 'R-Squared': 0.0032224931550222102},
'danceability': {'RMSE': 21.64588164749077, 'R-Squared': 0.0015465973924084775},
'energy': {'RMSE': 21.626583017384846, 'R-Squared': 0.0033261685420752363},
'loudness': {'RMSE': 21.627156819145167, 'R-Squared': 0.0032732798554711007},
'speechiness': {'RMSE': 21.64054948529217, 'R-Squared': 0.002038447051464831},
'acousticness': {'RMSE': 21.651872033025843, 'R-Squared': 0.0009938873166602802},
'instrumentalness': {'RMSE': 21.395424040261236, 'R-Squared': 0.024518496856643646}
'liveness': {'RMSE': 21.64911278818979, 'R-Squared': 0.0012484913355187421},
'valence': {'RMSE': 21.647551206851258, 'R-Squared': 0.0013925688765975552},
'tempo': {'RMSE': 21.66331581902167, 'R-Squared': -6.241242160309746e-05}}

Best feature for prediction:  instrumentalness
RMSE: 21.395424040261236
R-Squared: 0.024518496856643646
```

Figure.8

To predict popularity using all of the 10 features, I used **Multiple Regression model** to do so. Just like previous steps, I divided dataset into training and testing sets with my N-number again. Then I used these datasets to conduct and train the multiple regression model. After generating the predictions, I calculated RMSE and R-Squared. The results are as below (Figure.9) with **RMSE of 21.097 and R-Squared of 0.0515**. Compared with the linear regression model above, this multiple regression model is **improved by decrease of RMSE for 0.2984 and increase of R-Squared for 0.02702**. So, this multiple regression model performed a little bit better for predicting popularity. The improvement of the model can be attributed to the ability of dealing with more complex relationships of the multiple regression model and the combination of more effects from all features, instead of just relying on only one feature.

```
RMSE: 21.09702301294336
R-Squared: 0.051538742827321515
The multiple regression model is improved compared to the model in question 6) with
decrease of RMSE for 0.2984010273178761 and increase of R-Squared for 0.02702024597067787
```

Figure.9

To find the number of meaningful principal components to extract from the 10 features, I used **PCA** to conduct the **dimension reduction**. I firstly standardized the dataset through z-score normalization to avoid the sensitivity of PCA to the variances of the initial data. Next, I applied PCA to the dataset, which transformed the original variables into the principal components in the order of decreasing amount of variance, or eigenvalues. Based on **Kaiser criterion** and the Screeplot (Figure.10), **3 components** with eigenvalues greater than 1 should be extracted, and the total proportion of the variance these principal components account for is about **57.36%**. Finally, for these 3 principal components, I used both **Elbow Method** and **Silhouette Method** to identify the number of clusters in K-means clustering. As in Figure.11, since for Elbow Method, at the point of K equals 2, the distance starts to decrease at a slower rate, this point suggests the optimal number of clusters. And for Silhouette Method, since the highest silhouette score is also at 2 clusters, the **number of clusters should be 2**.

```
Number of meaningful principal components to extract: 3
Proportion of the variance these principal components account for: 0.5735819422797209
```

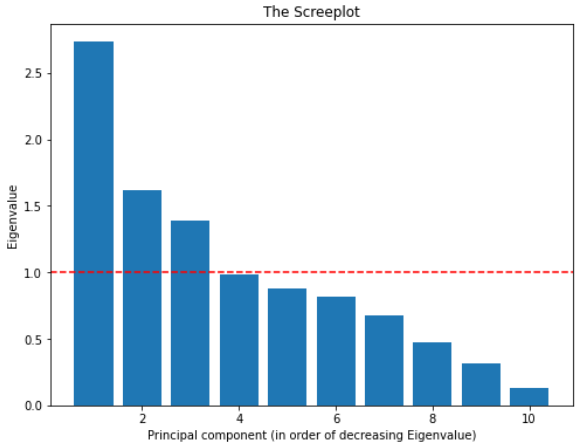


Figure.10

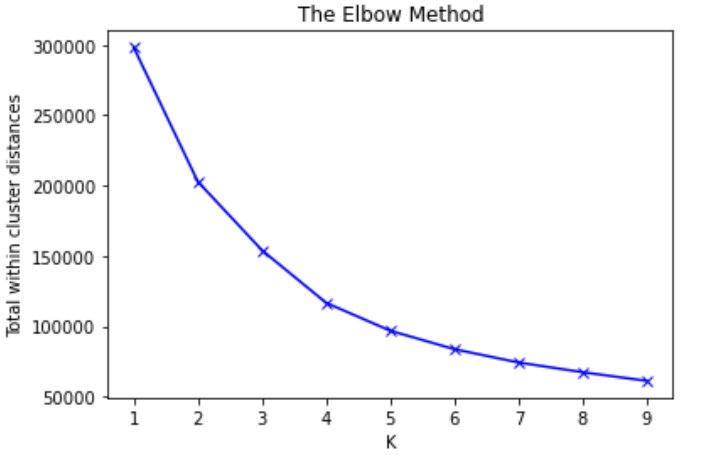


Figure.11

Since we needed to predict the major or minor key (mode) which is a binary outcome, I used **Logistic Regression model** to do it. I still divided the dataset of valence into training and testing sets firstly and then conducted the logistic regression model. Then I plotted the graph of the results of the logistic regression (Figure.12). To evaluate the performance of the model, I found the confusion matrix and **AUC** (Figure.13). Based on the graph and the values, and since AUC for valence is only about 0.505, **the prediction is not so good**. Then I tried to find a better predictor for that. I repeated the above process of logistic regression model for all of the rest features and compared their confusion matrix and AUC (Figure.14). Finally, I found the feature with highest AUC among all features, speechiness with AUC of 0.5656, and the scatter plot of speechiness is shown below (Figure.15). So, in this way, **speechiness can be a little better predictor for the major or minor key (mode) than valence**.

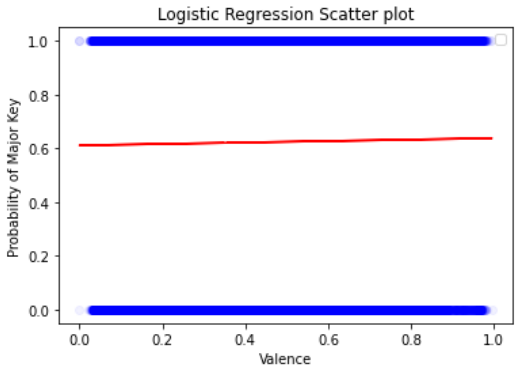


Figure.12

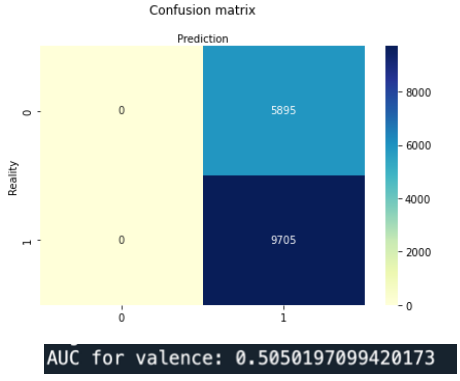


Figure.13

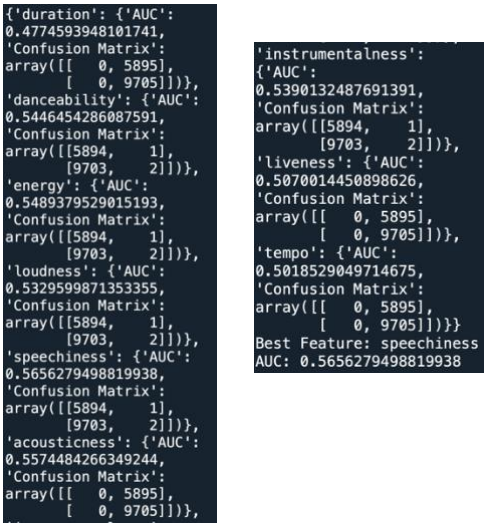


Figure.14.

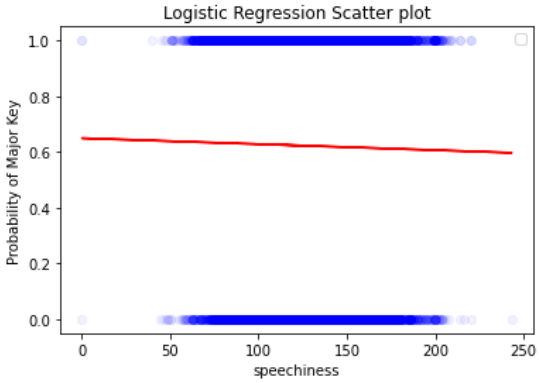


Figure.15

I firstly performed the prediction of genre from the 10 song features directly by the **Classification Tree**. I mapped the qualitative genre labels to numerical labels in the beginning and divided the dataset into training and testing sets as usual. Then conducted the classification tree and found AUC of 0.6316 (Figure.16). Next, I performed the prediction of genre from the principal components extracted before. With the same process, I found AUC of 0.6209 in this case (Figure.16). Since a higher AUC indicates a better model, the prediction of genre from the 10 song features is relatively better than that from the extracted principal components.

```
AUC of using 10 song features in question 1 directly: 0.6316429365844601
AUC of using principal components extracted in question 8: 0.6209020839110586
```

Figure.16

For the key of songs, if we label 0-5 to be low key and 6-11 to be high key, I wonder whether songs with low key are more popular than songs with high key or not. To do it, I found the distributions of popularity for both songs with high key and low key as below (Figure.17) firstly. Since both distributions are not normal, I used non-parametric test, **Mann-Whitney U test**. The null hypothesis is that there is no difference in the distribution of popularity between songs with high and low key. Then the results of the test, including U statistic and p-value of 0.000866 (Figure.18), indicates that the null hypothesis has been rejected since p-value is smaller than significance level of 0.05. And since the median of popularity for songs with low key is higher than that with high key, **songs with low key are more popular than songs with high key**.

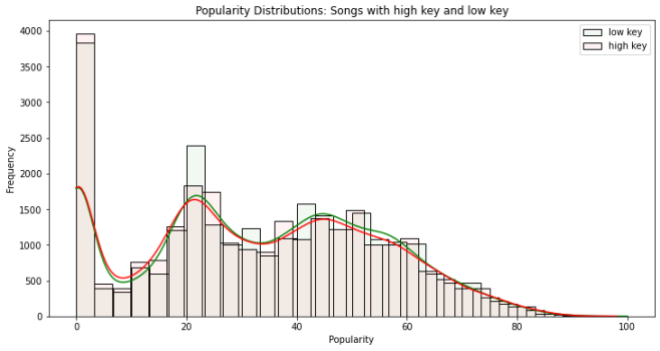


Figure.17

```
U statistic: 343690333.0
p-value: 0.0008655569902940956
Median of popularity for songs with low key: 33.0
Median of popularity for songs with high key: 32.0
```

Figure.18