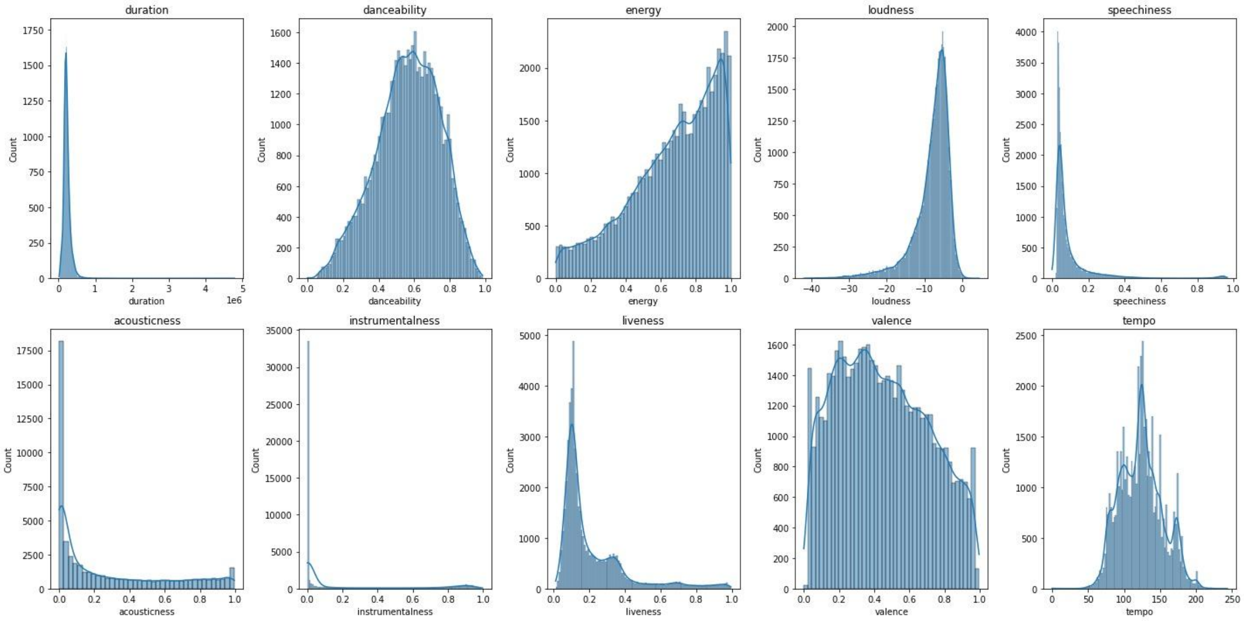
Tiffany He
Pascal Wallisch
Principles of Data Science (DS-UA-112)
20 Dec 2023

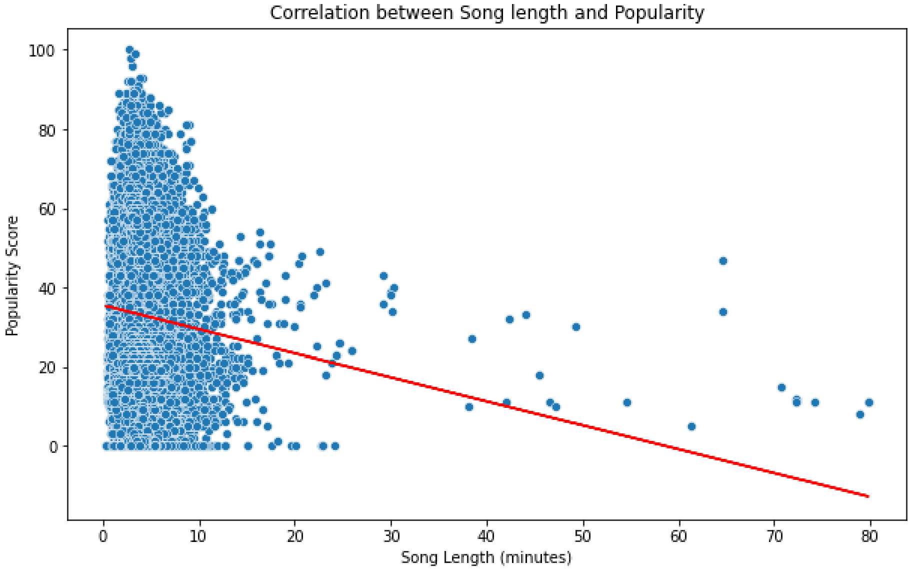
Analysis of Spotify Music

Before diving into the analysis, preprocessing the data is essential due to its imperfect nature. Once I loaded the Spotify dataset with pandas, a preliminary exploratory data analysis (EDA) was conducted to check for missing values. Fortunately, the dataset was complete, lacking any missing data. This allowed me to proceed with plotting histograms for each numerical feature and creating a correlation heatmap to observe the relationship among all features. The data's squared matrix format was noted, which could be advantageous for potential dimensional reduction in further analysis. Additionally, I set the seed for the random number generator to my N-number (17754923) wherever it's used in the code.

1) Upon examining the histograms, it's evident that danceability and tempo roughly follow normal distributions. Danceability aligns most closely with a typical bell curve. Tempo, despite having some irregularities and multiple peaks, gravitates around a steady BPM, suggesting a general trend towards a normal distribution. These two features, although not perfectly normal in their distribution, are comparatively more normal than other features, which show a range of skewness and irregular patterns.

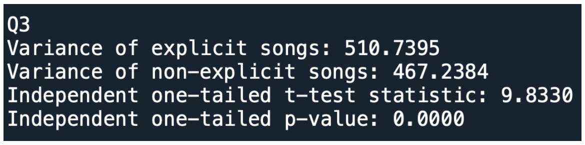


2) Before analyzing the precise relationship between song duration and its popularity, I created a scatter plot for an initial data visualization. The scatter plot shows a broad dispersion of data points, with a notable concentration at shorter song lengths. Although there are songs of varying popularity at these lengths, there isn't a clear upward or downward pattern indicating a strong correlation between song length and popularity.



To delve deeper into the relationship, I calculated both the Pearson and Spearman correlation coefficients. The Pearson coefficient is around -0.0547, and the Spearman coefficient is approximately -0.0373, both indicating a very weak negative relationship between song length and popularity. This suggests that as song length increases, there's a marginal decrease in popularity. Consequently, the relationship is so minimal that it lacks practical significance. It's more likely that other factors play a much more crucial role in determining a song's popularity than its length.

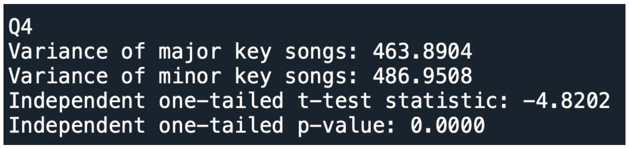
3) To evaluate this statement, I've chosen the Independent samples t-test. This test is appropriate for comparing the popularity of explicitly rated songs against songs not explicitly rated, since it is reasonable to reduce the data to sample mean. Given that we have two samples with unknown population parameters and slight variability among individuals, the similar variances calculated for each group validate this choice. My null hypothesis states that there is no significant difference in popularity between explicit and non-explicit songs, while the alternative hypothesis posits that explicit songs are more popular than non-explicit songs. With the goal of determining if explicit songs are indeed more popular, I employed a one-tailed test and divided the p-value by 2.



```
Q3
Variance of explicit songs: 510.7395
Variance of non-explicit songs: 467.2384
Independent one-tailed t-test statistic: 9.8330
Independent one-tailed p-value: 0.0000
```

The t-statistic is approximately 9.8330, and the p-value is extremely small, about 0.0000, which is well below the standard significance level of 0.05. Consequently, we can confidently reject the null hypothesis, inferring that on Spotify, explicit songs tend to be more popular than their non-explicit counterparts.

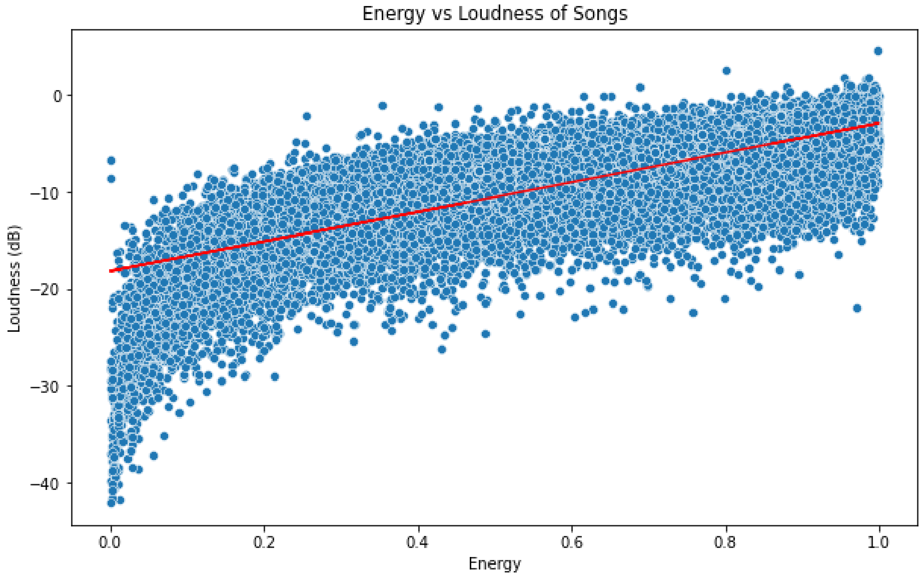
4) Similar to the earlier question, I employ a one-tailed Independent samples t-test, now aiming to test if songs in a major key are more popular than those in a minor key. My null hypothesis is that there is no significant difference in popularity between songs in major and minor keys, while the alternative hypothesis suggests that songs in a major key are more popular than those in a minor key.



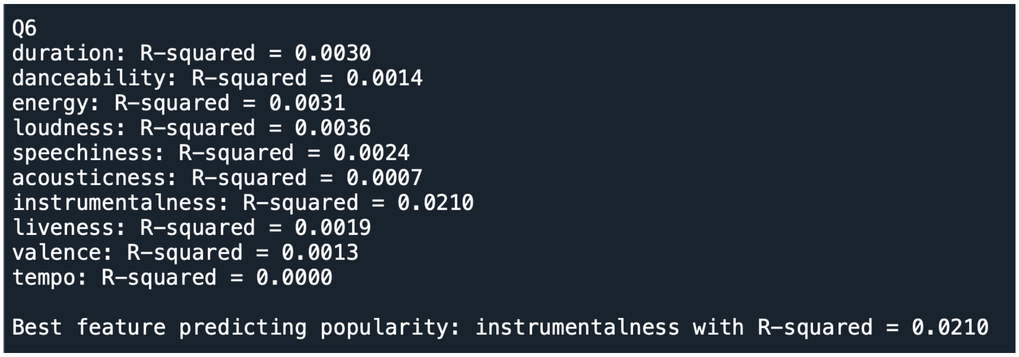
```
Q4
Variance of major key songs: 463.8904
Variance of minor key songs: 486.9508
Independent one-tailed t-test statistic: -4.8202
Independent one-tailed p-value: 0.0000
```

The t-statistic is approximately -4.8202, and the p-value is extremely small, about 0.0000, significantly below the standard significance level of 0.05. Given that the t-statistic is negative, we can reject the null hypothesis and conclude that there is a statistically significant difference in popularity, with songs in minor keys being more popular than those in major keys.

5) This question involves examining the relationship between two audio features of songs: energy and loudness, to determine if they are correlated. The scatter plot reveals a clear positive linear relationship: as energy increases, loudness generally increases as well. Therefore, to further investigate this linear relationship, I calculated the Pearson correlation coefficient, which is approximately 0.7749. This indicates a strong positive linear relationship between the two variables. Hence, songs with higher energy levels tend to be louder, supporting the statement that energy is largely indicative of a song's loudness.

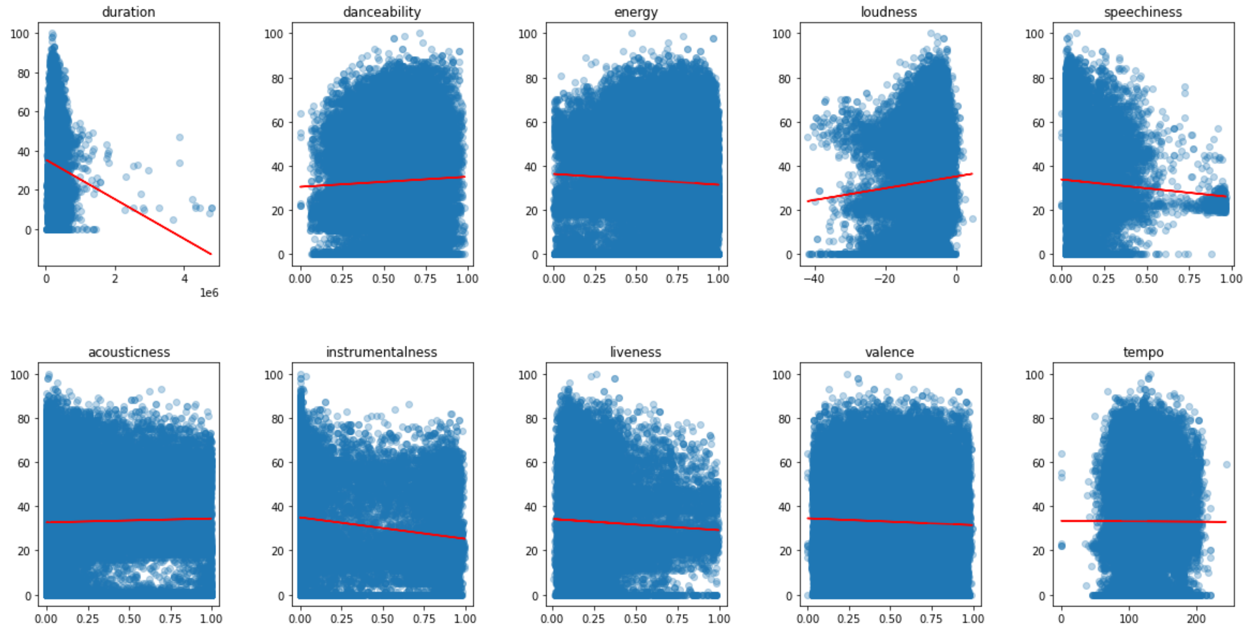


6) To evaluate how good the model of using song feature is to predict the popularity, I initialized a linear regression model for each of the 10 features. Each model was then fitted to the data and assessed using the R-squared value. R-squared represents the proportion of the variance in the dependent variable (popularity) that can be predicted from the independent variable (each feature), where a higher value indicates a better model fit.



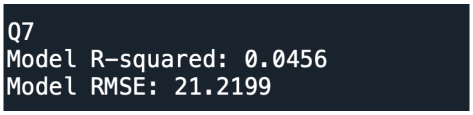
```
Q6
duration: R-squared = 0.0030
danceability: R-squared = 0.0014
energy: R-squared = 0.0031
loudness: R-squared = 0.0036
speechiness: R-squared = 0.0024
acousticness: R-squared = 0.0007
instrumentalness: R-squared = 0.0210
liveness: R-squared = 0.0019
valence: R-squared = 0.0013
tempo: R-squared = 0.0000

Best feature predicting popularity: instrumentalness with R-squared = 0.0210
```

The R-squared values for all features are close to zero, suggesting that none are strong predictors of song popularity individually. The feature with the highest R-squared value is instrumentalness, at 0.0210. This indicates that it accounts for 2.1% of the variance in popularity scores. Although this value is relatively low, among the 10 song features in question 1, instrumentalness emerges as the best individual predictor of song popularity. I also created scatterplots with regression lines for each feature, which further support my conclusion that using these 10 features to predict popularity is not an effective approach.



7) When constructing a model with multiple features, the risk of overfitting increases. To mitigate this and to obtain an unbiased assessment of the model's predictive performance, I divided the data into training and testing sets. The multiple linear regression model was then

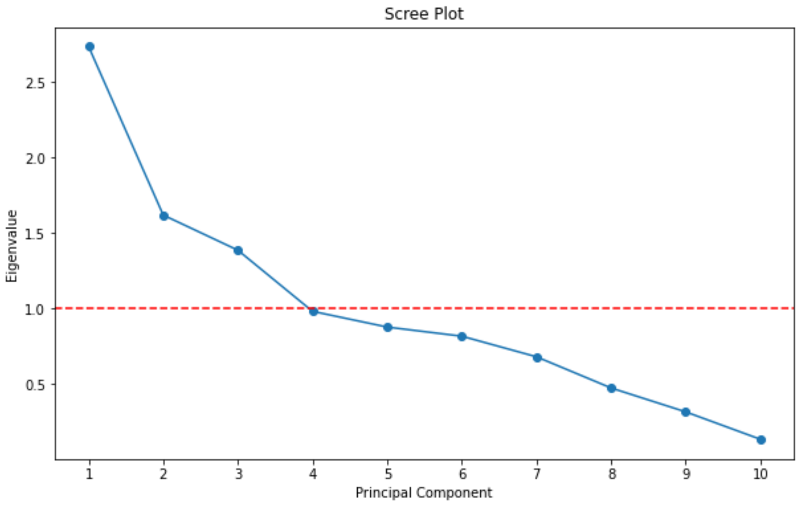
applied to the training data. Following this, it made predictions on the test set, which I evaluated using the R-squared value and Root Mean Square Error (RMSE).



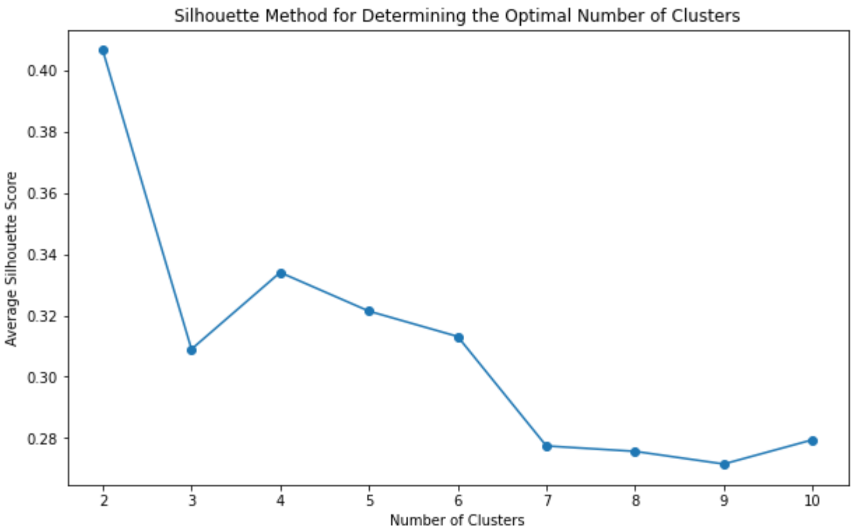
```
Q7
Model R-squared: 0.0456
Model RMSE: 21.2199
```

The R-squared value is approximately 0.0456. This indicates that the model, incorporating all features, explains about 4.56% of the variance in popularity. This is a slight improvement over the R-squared for the best individual predictor, instrumentalness, which was 0.0210. It demonstrates that the model using all features collectively explains more variability in song popularity than instrumentalness alone. The RMSE is roughly 21.2199, suggesting that the model's predictions deviate from actual values by an average of 21.11 popularity points. Considering popularity scores range from 0 to 100, an error margin of 21.11 is relatively significant. Therefore, even though the model shows some improvement from the previous one, it still indicates that a substantial portion of the variance in popularity is not accounted for by these features.

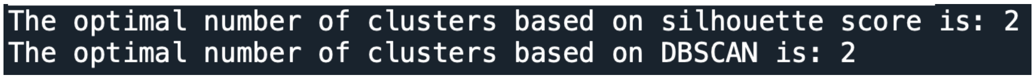
8) To assess the optimal number of principal components, I implemented PCA for dimensionality reduction. I observed that the features in our dataset varied in units, requiring standardization to achieve a mean of 0 and a standard deviation of 1. The correlation matrix indicated that the data is structured as a symmetric square matrix, making it suitable for Eigen-decomposition.



After applying PCA, I examined the Scree plot through the Kaiser criterion, retaining principal components with eigenvalues exceeding 1. This approach led me to identify three significant principal components. Although the fourth eigenvalue was marginally below 1, it was excluded per the Kaiser criterion. These three components cumulatively explain approximately 57.36% of the dataset's variability.



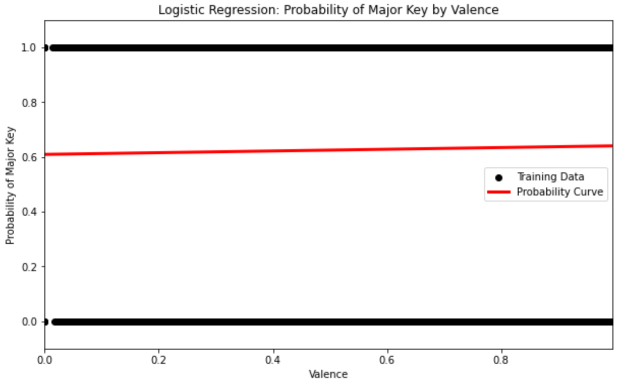
For cluster determination, I employed K-means and DBSCAN clustering methods. In K-means, the optimal cluster count was determined using silhouette scores, which range from -1 to +1, with higher values signifying better fit within a cluster and distinct separation from neighboring clusters. According to the graph above, the highest average silhouette score occurs at 2, which suggests that two as the optimal number of clusters. To validate this, I also used DBSCAN, a method that forms clusters based on data point density. This approach also identified two as the optimal number of clusters, aligning with the K-means findings. Thus, both methods indicated the presence of two distinct clusters in the data.

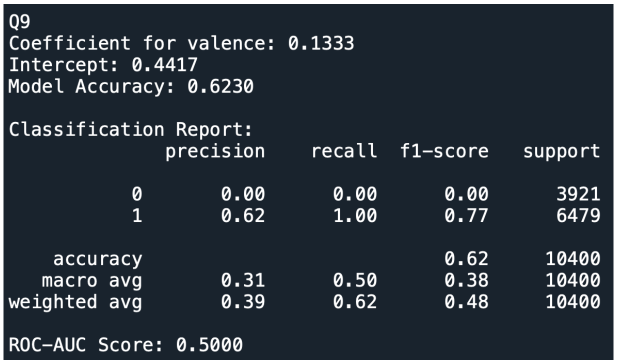


```
The optimal number of clusters based on silhouette score is: 2
The optimal number of clusters based on DBSCAN is: 2
```

9) For this question, I have chosen to perform logistic regression, as it is a supervised machine learning classification model specifically designed to predict binary or categorical dependent variables. In this case, the dependent variable is the song mode (whether in major or minor key) is a binary categorical data. I then split the dataset into a training set and a testing set. After training the logistic regression model on the training set, where 'valence' is the independent variable, and 'mode' is the dependent variable, the coefficient for valence is approximately 0.133 and the intercept is approximately 0.442, suggesting a positive relationship between valence and the likelihood of a song being in a major key.

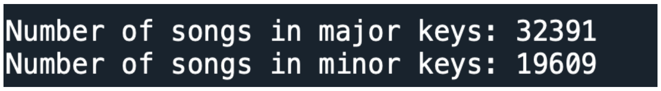
For the performance of the model, the accuracy is roughly 62.30%, indicating that the model correctly predicted the key mode for about 62.30% of the songs in the test set. The classification report, however, indicates a significant issue. The precision, recall, and F1-score for the class predicting minor keys (class 0) are all zero, indicating the model did not correctly predict any minor key songs. The model only predicts major keys, as shown by the recall of 1.00 for the major key class. Besides, the ROC-AUC score is 0.5000, which is the score that a random classifier would typically achieve.



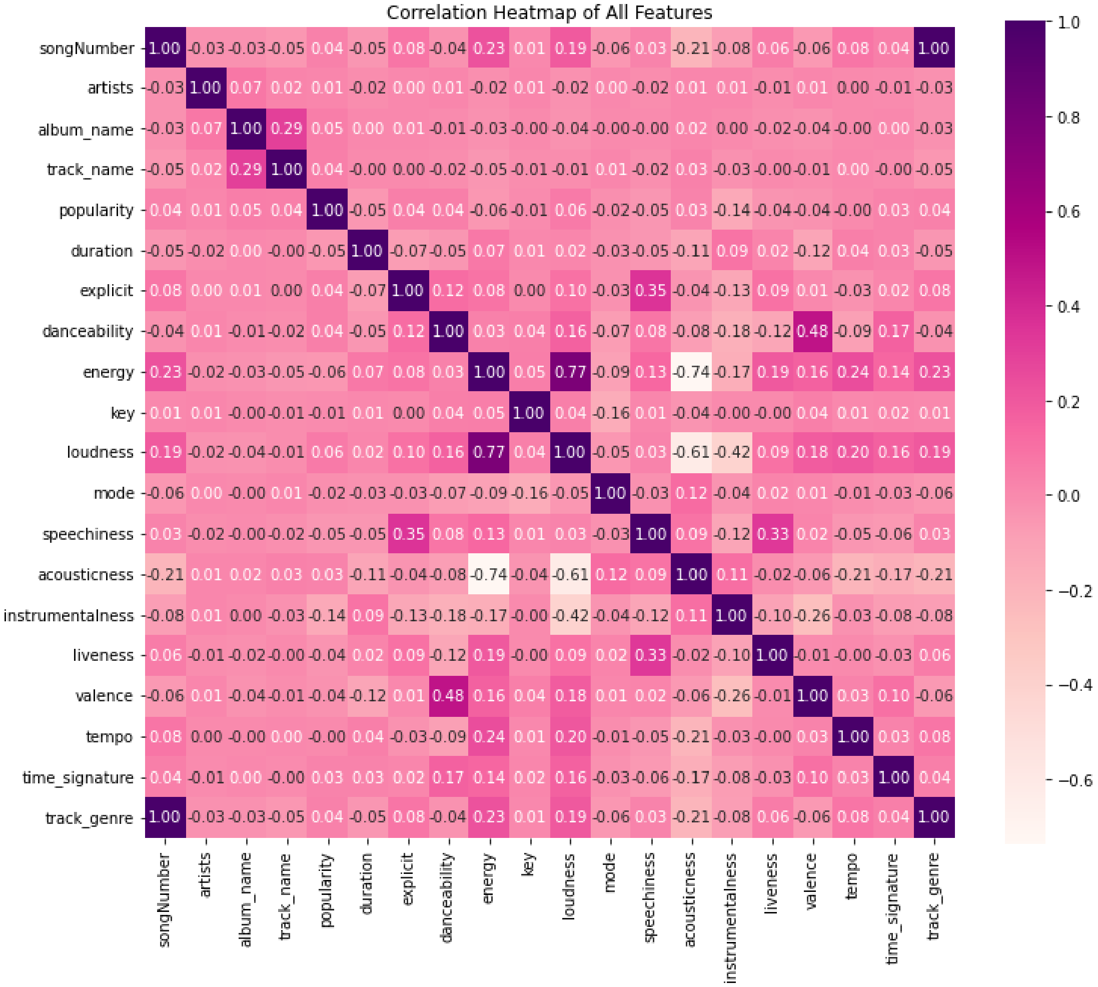


```
Q9
Coefficient for valence: 0.1333
Intercept: 0.4417
Model Accuracy: 0.6230

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      3921
           1       0.62      1.00      0.77      6479

    accuracy                           0.62     10400
   macro avg       0.31      0.50      0.38     10400
weighted avg       0.39      0.62      0.48     10400

ROC-AUC Score: 0.5000
```

The potential issue might be that the number of songs in minor keys (19609) is roughly two thirds of the songs in major keys (32391), so a potential way to improve the model is to increase the number of instances in the minority class by randomly replicating them until you achieve a balance using resampling methods.

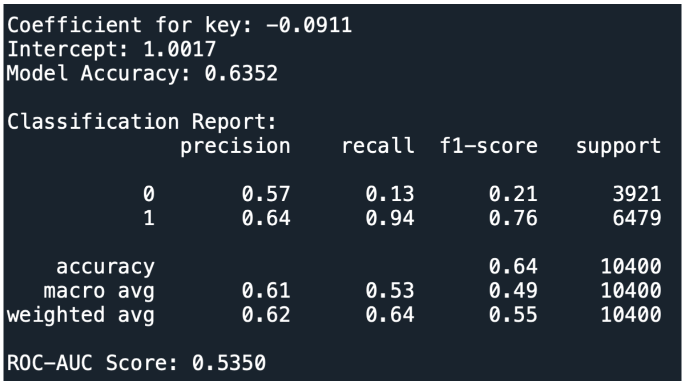


```
Number of songs in major keys: 32391
Number of songs in minor keys: 19609
```



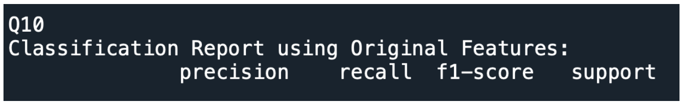
	songNumber	artists	album_name	track_name	popularity	duration	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	track_genre
songNumber	1.00	-0.03	-0.03	-0.05	0.04	-0.05	0.08	-0.04	0.23	0.01	0.19	-0.06	0.03	-0.21	-0.08	0.06	-0.06	0.08	0.04	1.00
artists	-0.03	1.00	0.07	0.02	0.01	-0.02	0.00	0.01	-0.02	0.01	-0.02	0.00	-0.02	0.01	0.01	-0.01	0.01	0.00	-0.01	-0.03
album_name	-0.03	0.07	1.00	0.29	0.05	0.00	0.01	-0.01	-0.03	-0.00	-0.04	-0.00	-0.00	0.02	0.00	-0.02	-0.04	-0.00	0.00	-0.03
track_name	-0.05	0.02	0.29	1.00	0.04	-0.00	0.00	-0.02	-0.05	-0.01	-0.01	0.01	-0.02	0.03	-0.03	-0.00	-0.01	0.00	-0.00	-0.05
popularity	0.04	0.01	0.05	0.04	1.00	-0.05	0.04	0.04	-0.06	-0.01	0.06	-0.02	-0.05	0.03	-0.14	-0.04	-0.04	-0.00	0.03	0.04
duration	-0.05	-0.02	0.00	-0.00	-0.05	1.00	-0.07	-0.05	0.07	0.01	0.02	-0.03	-0.05	-0.11	0.09	0.02	-0.12	0.04	0.03	-0.05
explicit	0.08	0.00	0.01	0.00	0.04	-0.07	1.00	0.12	0.08	0.00	0.10	-0.03	0.35	-0.04	-0.13	0.09	0.01	-0.03	0.02	0.08
danceability	-0.04	0.01	-0.01	-0.02	0.04	-0.05	0.12	1.00	0.03	0.04	0.16	-0.07	0.08	-0.08	-0.18	-0.12	0.48	-0.09	0.17	-0.04
energy	0.23	-0.02	-0.03	-0.05	-0.06	0.07	0.08	0.03	1.00	0.05	0.77	-0.09	0.13	-0.74	-0.17	0.19	0.16	0.24	0.14	0.23
key	0.01	0.01	-0.00	-0.01	-0.01	0.01	0.00	0.04	0.05	1.00	0.04	-0.16	0.01	-0.04	-0.00	-0.00	0.04	0.01	0.02	0.01
loudness	0.19	-0.02	-0.04	-0.01	0.06	0.02	0.10	0.16	0.77	0.04	1.00	-0.05	0.03	-0.61	-0.42	0.09	0.18	0.20	0.16	0.19
mode	-0.06	0.00	-0.00	0.01	-0.02	-0.03	-0.03	-0.07	-0.09	-0.16	-0.05	1.00	-0.03	0.12	-0.04	0.02	0.01	-0.01	-0.03	-0.06
speechiness	0.03	-0.02	-0.00	-0.02	-0.05	-0.05	0.35	0.08	0.13	0.01	0.03	-0.03	1.00	0.09	-0.12	0.33	0.02	-0.05	-0.06	0.03
acousticness	-0.21	0.01	0.02	0.03	0.03	-0.11	-0.04	-0.08	-0.74	-0.04	-0.61	0.12	0.09	1.00	0.11	-0.02	-0.06	-0.21	-0.17	-0.21
instrumentalness	-0.08	0.01	0.00	-0.03	-0.14	0.09	-0.13	-0.18	-0.17	-0.00	-0.42	-0.04	-0.12	0.11	1.00	-0.10	-0.26	-0.03	-0.08	-0.08
liveness	0.06	-0.01	-0.02	-0.00	-0.04	0.02	0.09	-0.12	0.19	-0.00	0.09	0.02	0.33	-0.02	-0.10	1.00	-0.01	-0.00	-0.03	0.06
valence	-0.06	0.01	-0.04	-0.01	-0.04	-0.12	0.01	0.48	0.16	0.04	0.18	0.01	0.02	-0.06	-0.26	-0.01	1.00	0.03	0.10	-0.06
tempo	0.08	0.00	-0.00	0.00	-0.00	0.04	-0.03	-0.09	0.24	0.01	0.20	-0.01	-0.05	-0.21	-0.03	-0.00	0.03	1.00	0.03	0.08
time_signature	0.04	-0.01	0.00	-0.00	0.03	0.03	0.02	0.17	0.14	0.02	0.16	-0.03	-0.06	-0.17	-0.08	-0.03	0.10	0.03	1.00	0.04
track_genre	1.00	-0.03	-0.03	-0.05	0.04	-0.05	0.08	-0.04	0.23	0.01	0.19	-0.06	0.03	-0.21	-0.08	0.06	-0.06	0.08	0.04	1.00

According to the correlation matrix, I identified that the feature 'key' (the key of the song) has the highest correlation coefficient with 'mode' at -0.16, while the 'valence' used in the previous question only has a correlation of 0.01. Consequently, I performed another logistic regression using 'key' as the predictor for 'mode'. The results show an accuracy of 0.6352 and an ROC-AUC score of 0.5330, both of which are slightly higher than our last model. This suggests that the new model's ability to distinguish between classes is still weak, but it represents a slight improvement over random guessing and shows an advancement from using 'valence' to predict the song's mode.

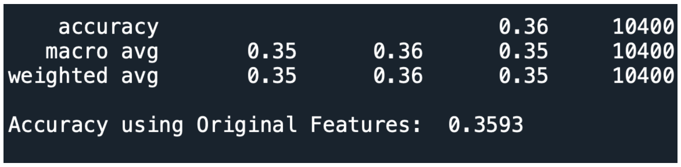


```
Coefficient for key: -0.0911
Intercept: 1.0017
Model Accuracy: 0.6352

Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.13      0.21      3921
           1       0.64      0.94      0.76      6479

    accuracy                           0.64     10400
   macro avg       0.61      0.53      0.49     10400
weighted avg       0.62      0.64      0.55     10400

ROC-AUC Score: 0.5350
```

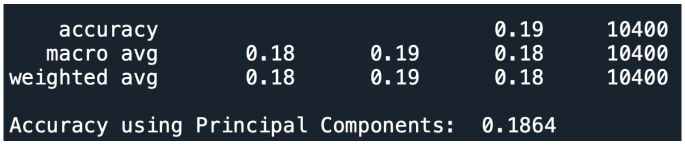
10) The question essentially seeks to determine if supervised or unsupervised learning methods are better suited for genre prediction tasks using data from Spotify. Given the ample labels and predictive nature of this data, supervised learning is the preferable approach. Unsupervised learning, such as PCA, reduces data dimensionality by creating new variables. However, this can lead to data loss and lower accuracy in predictions. To support this, I first map the qualitative genre labels to numerical labels and then divide the data into training and test sets. For the classification model, I chose the Random Forest classifier for its effectiveness with numerical and categorical data, resistance to overfitting, and high accuracy in classification tasks. Consequently, I conduct a comparative analysis using two approaches: one using Random Forest classification with 10 song features and another applying PCA to three principal components. The results indeed show higher accuracy with the original features (0.3593) compared to the principal components (0.1864).



```
Q10
Classification Report using Original Features:
              precision    recall  f1-score   support
```



```
    accuracy                           0.36     10400
   macro avg       0.35      0.36      0.35     10400
weighted avg       0.35      0.36      0.35     10400

Accuracy using Original Features:  0.3593
```



```
    accuracy                           0.19     10400
   macro avg       0.18      0.19      0.18     10400
weighted avg       0.18      0.19      0.18     10400

Accuracy using Principal Components:  0.1864
```