Thiago Pari

tp2330

Principles of Data Science

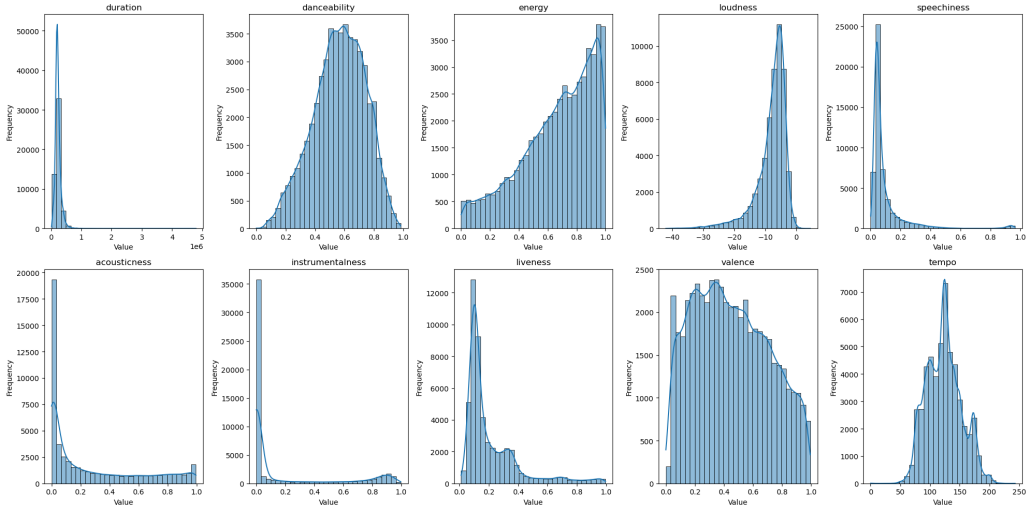
Final Project

Introduction

For this project, I used both NumPy and pandas to handle and analyze the Spotify dataset. The dataset was loaded using pandas, and I performed various preprocessing steps including cleaning the data and addressing missing values. To reduce the dimensionality of the dataset, I applied Principal Component Analysis (PCA). Before conducting any calculations or statistical tests, I standardized the features when necessary to ensure consistent scaling.

In addition, the random number generator seed was set to 19259759

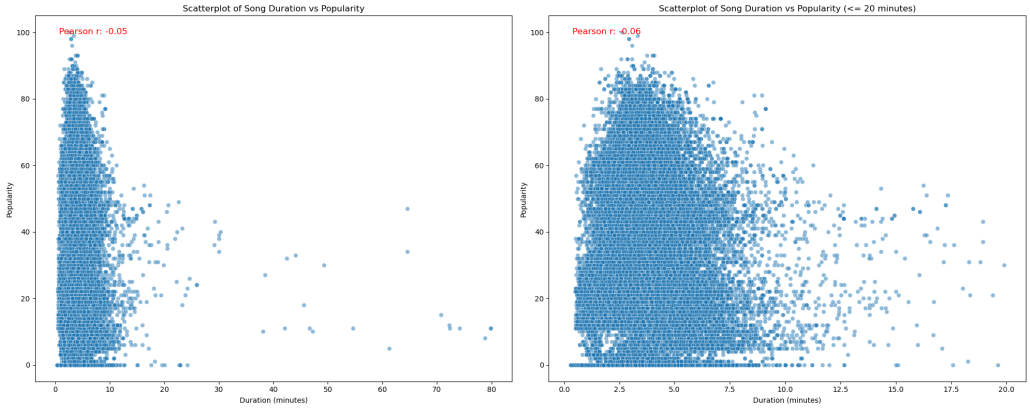
1) Consider the 10 song features: Duration, Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo. Are any of these features reasonably distributed normally? If so, which one? Suggestion: Include a 2x5 figure with histograms for each feature



To determine if any of the song features are normally distributed, I conducted both the Shapiro-Wilk test and D'Agostino's K squared test for each feature. The results indicated

that none of the features are normally distributed as all p-values were below the significance level. It can also be seen in the histograms that, although danceability and valence may look like a normal distribution, they have skewness, which is reflected on the tests.

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



The scatterplot illustrating the relationship between song length and popularity showed a very weak negative correlation, with a Pearson correlation coefficient of approximately -0.055. We can see that even by removing outliers and limiting it to songs with a duration of 20 minutes or less, the correlation does not change drastically. This suggests that there is almost no linear relationship between the duration of a song and its popularity.

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]

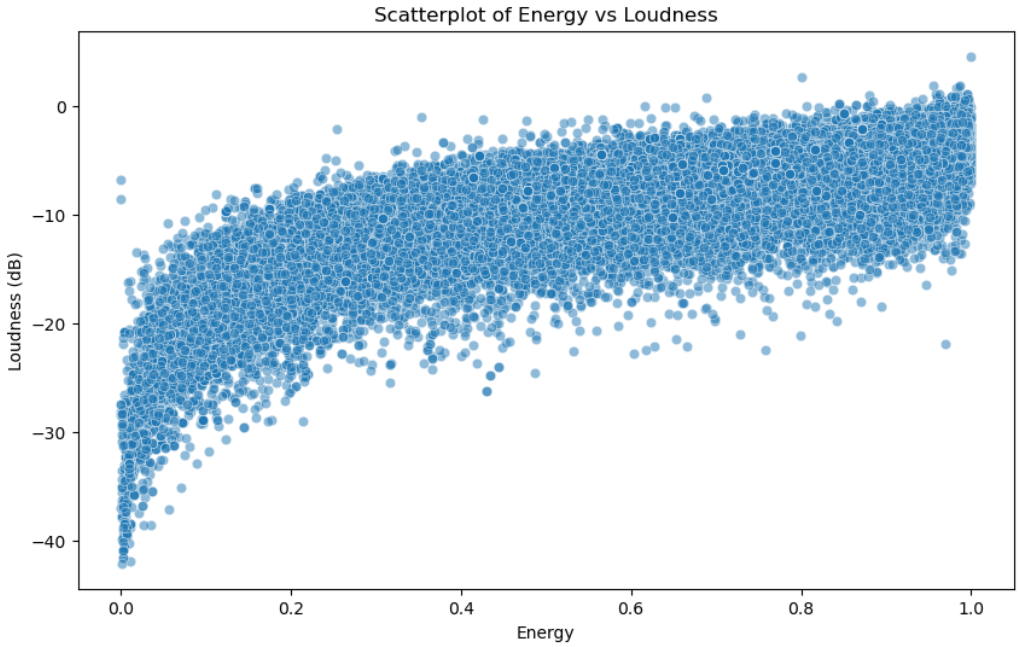
To investigate whether explicitly rated songs are more popular than non-explicit songs, I used the Mann-Whitney U test, a non-parametric test suitable for comparing two independent samples without assuming a normal distribution. The dataset was divided into two groups: explicit songs and non-explicit songs. The test revealed a significant difference in popularity between these groups, with explicit songs having a higher mean popularity (35.81) compared to non-explicit songs (32.79). The U-Statistic was 139,361,273.5, and the p-value was extremely low (3.07e-19), indicating strong evidence against the null hypothesis of no difference in popularity. This suggests that explicit songs are generally more popular, possibly due to audience preferences or the nature of explicit content. However, further analysis could explore specific genres or artists to understand this trend better.

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

To determine if there is a significant difference in popularity between songs in major and minor keys, the Mann-Whitney U test was employed again. The dataset was divided into

major key songs (mode = 1) and minor key songs (mode = 0). The test results indicated a significant difference, with minor key songs having a slightly higher mean popularity (33.71) compared to major key songs (32.76). The U-Statistic was 309,702,373.0, and the p-value was 2.02e-06, suggesting strong evidence against the null hypothesis of no difference in popularity. This finding challenges the common assumption that major key songs, often associated with positive emotions, are more popular. The preference for minor key songs could be due to their emotional depth or musical complexity. Further analysis could investigate the influence of genres, lyrical content, or cultural trends on this preference.

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 scatterplot]



The scatterplot and correlation analysis between energy and loudness revealed a strong positive correlation (Pearson correlation coefficient of approximately 0.775). This suggests that songs with higher energy tend to be louder.

6) Which of the 10 song features in question 1 predicts popularity best? How good is this model?

For this question, I built 10 individual OLS models with each feature as an input variable to predict popularity.

	Feature	RMSE	R^2
0	duration	21.713905	0.002999

1	danceability	21.731276	0.001403
2	energy	21.715194	0.002880
3	loudness	21.693340	0.004886
4	speechiness	21.720691	0.002375
5	acousticness	21.742401	0.000380
6	instrumentalness	21.474318	0.024879
7	liveness	21.728270	0.001679
8	valence	21.745656	0.000081
9	tempo	21.746961	-0.000039

Among these individual models, instrumentalness had the highest R^2 value of 0.025, making it the best single predictor of popularity among the features.

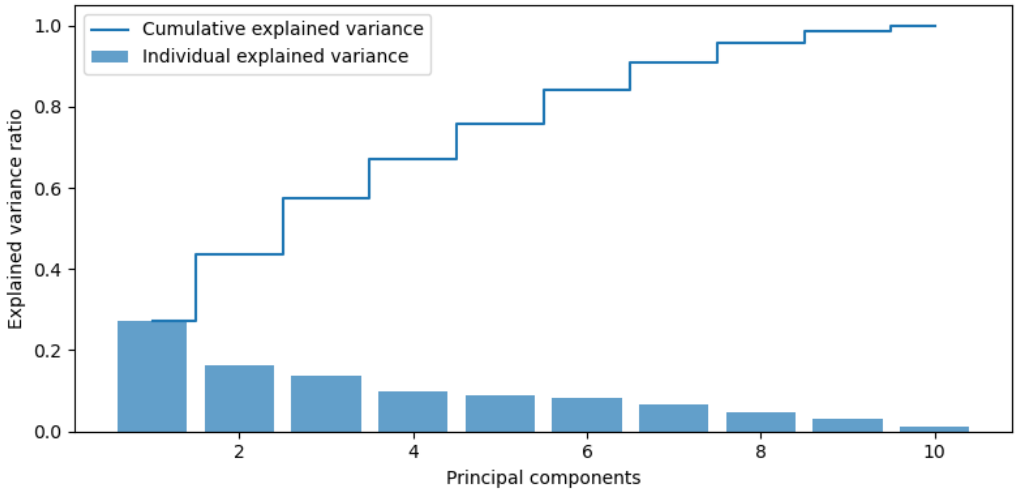
7) Building a model that uses *all* of the song features in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question 6). How do you account for this?

R^2 value using all features: 0.051535237581818105

R^2 value using 'instrumentalness' only: 0.02487861316214779

The R^2 value for the model using all features is higher than the R^2 value for the model using only 'instrumentalness'. This indicates that the model with all features explains a greater proportion of the variance in the target variable (popularity). One possible reason for this improvement is collinearity among the features. When multiple features are correlated, they can collectively contribute to a better prediction even if individually they may not be strong predictors. The model with all features can capture complex relationships that a single feature model cannot.

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?



Applying PCA to the song features, I determined that 8 principal components are needed to explain at least 95% of the variance. The first principal component alone explained 27.34% of the variance, with the subsequent components contributing less. The exact proportion of variance explained by the selected PCs is as follows:

```
PC1: 27.34%
PC2: 14.05%
PC3: 12.47%
PC4: 10.58%
PC5: 10.14%
PC6: 9.39%
PC7: 7.21%
PC8: 3.83%
```

9) Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?

To predict whether a song is in a major or minor key based on the feature *valence*, a logistic regression model was used. Additionally, class imbalance was handled using SMOTE (Synthetic Minority Over-sampling Technique).

The logistic regression model using *valence* as a predictor yielded the following results:

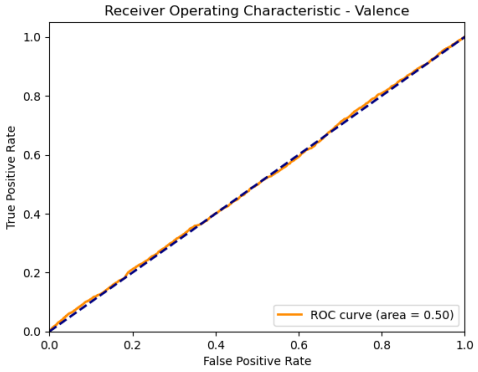
- **Accuracy:** 0.4986
- **AUC:** 0.5037

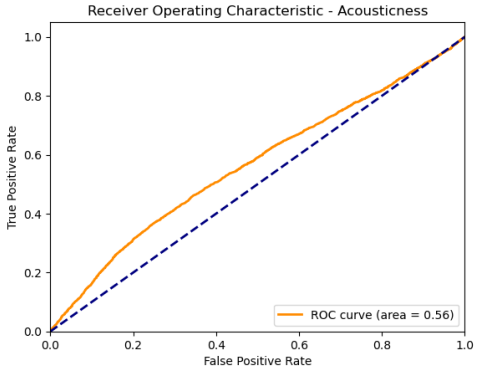
Upon examining other features, *acousticness* was identified as a better predictor:

- **Best Predictor:** acousticness
- **Accuracy:** 0.5572
- **AUC:** 0.5641

Justification: The R^2 value for the model using *acousticness* as a predictor is higher than that for *valence*, indicating that *acousticness* explains a greater proportion of the variance in the target variable (mode). The logistic regression model with *acousticness* shows improved performance in both accuracy and AUC. This improvement can be attributed to the fact that *acousticness* captures more relevant information for predicting the mode of a song. Additionally, handling class imbalance using SMOTE contributed to the improved performance of the model.

Feature	Accuracy	AUC
valence	0.4986	0.5037
acousticness	0.5572	0.5641





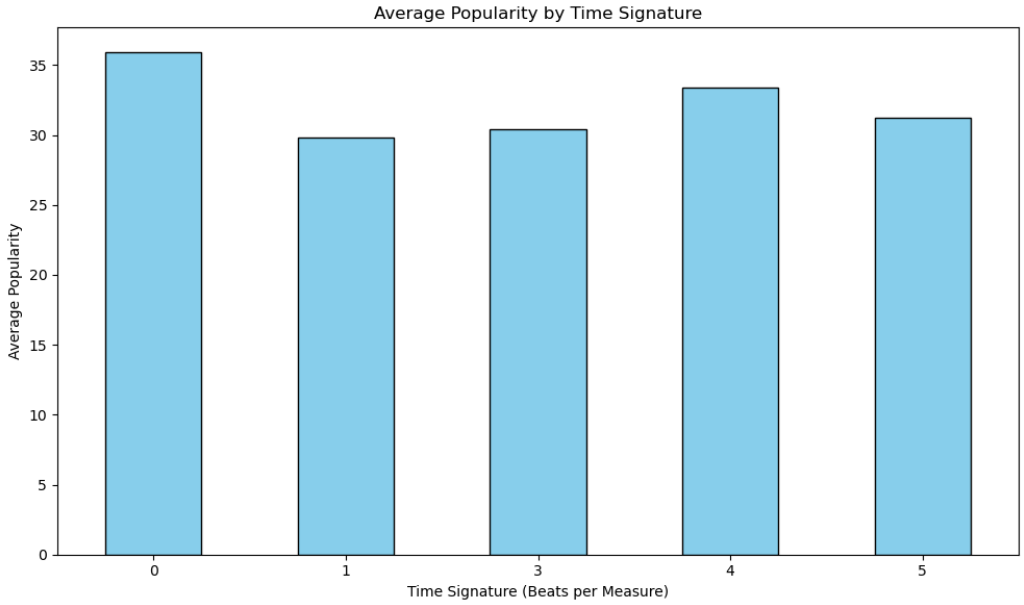
10) Which is a better predictor of whether a song is classical music – duration or the principal components you extracted in question 8?

```
Duration Model Accuracy:
0.9801923076923077

Duration Model Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10194
           1       0.00      0.00      0.00       206

    accuracy                           0.98     10400
```

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

PCA Model Accuracy:
0.9806730769230769

PCA Model Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     10194
           1       0.53      0.19      0.28       206

    accuracy                           0.98     10400
   macro avg       0.76      0.59      0.63     10400
weighted avg       0.97      0.98      0.98     10400
```

Comparing logistic regression models using duration and PCA components to predict whether a song is classical, I found that the model using PCA components had a higher AUC (0.959) compared to the model using duration alone (0.558). Both models had similar high accuracy, but the PCA model performed significantly better in distinguishing classical tracks.

Extra Credit Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the number of beats per measure, something with the key, or something with the song or album titles]



Analyzing the relationship between time signatures and song popularity, I found that songs with a 4/4 time signature are the most common and moderately popular, while less common time signatures like 1/4 and 3/4 do not significantly outperform 4/4 in popularity. Songs with 0 beats per measure had the highest average popularity, but this may be due to the small sample size. This analysis provides an interesting perspective on how the structure of a song might influence its popularity.