

Jerry Huang

Pascal Wallisch

Principles of Data Science

May 14, 2024

### Data Analysis of Spotify songs

Data Processing: I began the analysis by addressing missing values and processing the data as needed. Initially, I reviewed all the data labels and considered the significance of each. I decided to retain the zeros for all features, as they hold meaningful information within these labels. For instance, “popularity” includes integer from 0 to 100 and I interpret it as “the least listened song”. Also, I checked if there is any Nah value, and it turned out to be 0, suggesting that we have a full matrix. Eventually, I turned duration into minutes for me to better understand how long the songs are.

```
#%% Data Processing
# Loading files
df = pd.read_csv("spotify52kData.csv")

# Try to see if there is any missing value
missing_values = df.isnull().sum()
print(missing_values)

# Turning duration into minutes for me to better understand the song
df['duration_in_minutes'] = df['duration'] / 60000
```

songNumber	0
artists	0
album_name	0
track_name	0
popularity	0
duration	0
explicit	0
danceability	0
energy	0
key	0
loudness	0
mode	0
speechiness	0
acousticness	0
instrumentalness	0
liveness	0
valence	0
tempo	0
time_signature	0
track_genre	0
dtype:	int64

*Figure 1: Data Processing and Nah value checking.*

Additionally, I created a heatmap to visualize the relationships between features, driven by curiosity. The heatmap revealed that certain features, like energy and loudness, are strongly correlated.

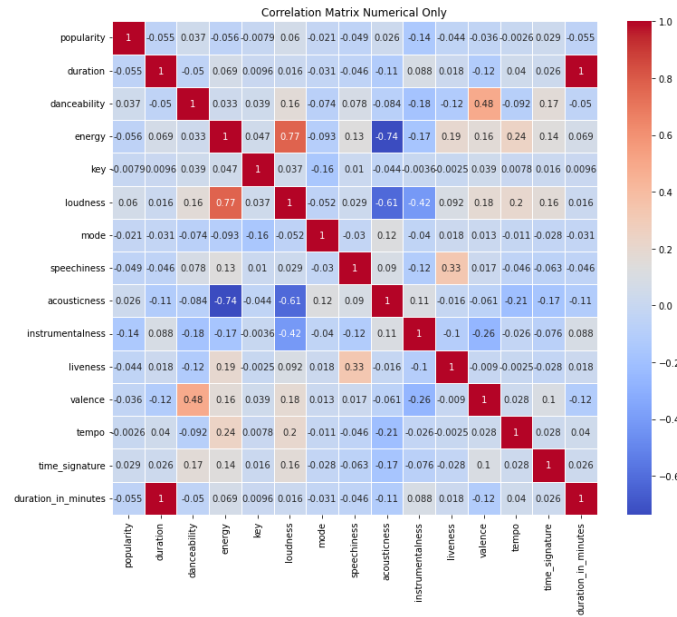


Figure 2: Heatmap visualizing relationship within features

1) **Question:** Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Is any of these features reasonably distributed normally? If so, which one?

I plotted a 2x5 figure histograms for each feature below. It has shown that there aren't any reasonably normally distributed features among those 10 columns.

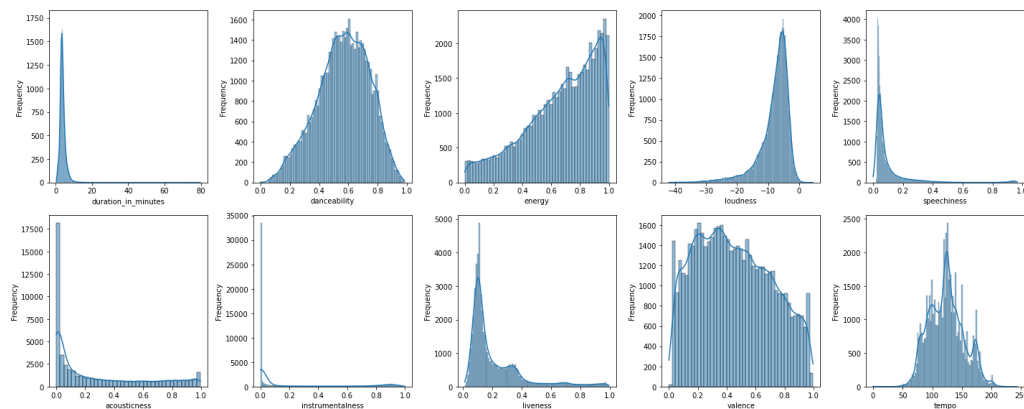
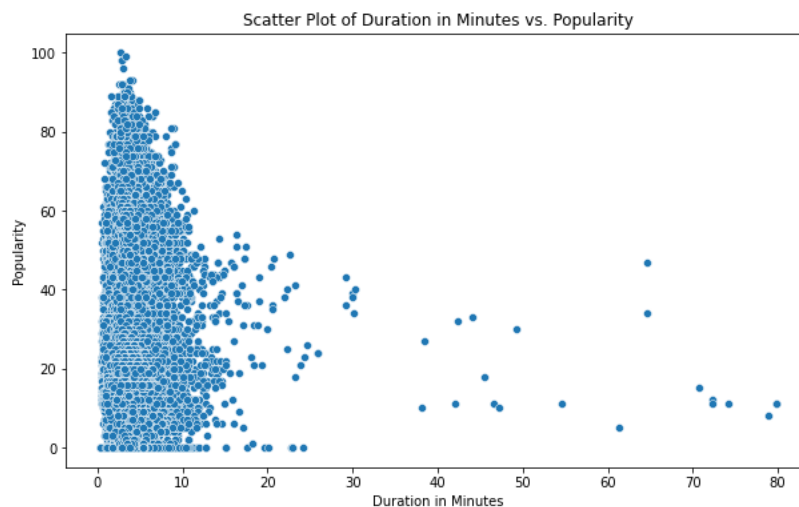


Figure 3: Histograms for each feature.

2) **Question:** Is there a relationship between song length and popularity of a song? If so, if the relationship positive or negative?

For question 2 I plotted a scatterplot for popularity and duration in minutes. The correlation coefficient between song duration and popularity is **-0.0547**, and the Spearman's rank correlation coefficient **-0.037**. Most songs are concentrated from 0 to 10 minutes, but the correlation coefficient may not be trustworthy here, as we have some extreme outliers. However, I don't think we can remove those outliers because we are finding the relationship between song length and popularity of a song. We will have long songs in real life such as classical concert. Thus, I sliced the data into half, marking short songs as duration smaller or equal to 20 minutes, and long songs as duration larger than 20 minutes. Eventually, the Correlation in Short Songs is **-0.0578**, suggesting a weak negative relationship within this range; the Correlation in Long Songs is **-0.3262**, suggesting a moderate negative relationship within this range. Therefore, I think the relationship between song length and popularity of a song is negative, which means the longer the song, the less popular that song is.



*Figure 4: Scatterplot of Duration vs. Popularity*

3) Are explicitly rated songs more popular than songs that are not explicit?

For this analysis, I employed the One-tailed Mann-Whitney U test, which is appropriate given that popularity is a numerical label, and we are comparing two distinct groups: explicit and non-

explicit songs. The Mann-Whitney U test statistic is **139361273.5** and the P-value is **1.5339599669557339e-19**. The result of p-value is statistically significant given that **alpha = 0.05**. We reject the null hypothesis that there is no difference in median of popularity scores between explicit and implicit songs. We have evidence suggesting that explicitly rated songs are more popular than songs that are not explicit, as the median popularity score of explicitly rated songs is higher than that of non-explicit songs.

```
# We split the songs according to their explicitness feature. Our two different sample groups
# Popularity is ordered data, so I decide to use Mann-Whitney U test
# We can't do a permutation test for reasons that : 1. sample size is small, but we have a large dataset

explicit = df[df['explicit'] == True]['popularity']
implicit = df[df['explicit'] == False]['popularity']

u_statistic, p_value = stats.mannwhitneyu(explicit, implicit, alternative='greater')
print("Mann-Whitney U test statistic:", u_statistic)
print("P-value:", p_value)
```

*Figure 5: Codes of U-Test*

4) Are songs in major key more popular than songs in minor key?

For this analysis, I used the Mann-Whitney U test with the alternative hypothesis that songs in a major key are more popular than those in a minor key. The Mann-Whitney U test statistic is **309702373.0**, and the P-value is **0.9999989912386331**. Given that the P-value is much higher than the significance level of **0.05**, we fail to reject the null hypothesis that there is no difference in popularity between songs in major and minor keys.

This high P-value indicates that we do not have significant evidence to support the hypothesis that songs in a major key are more popular than those in a minor key. The direction of the test (testing if major key songs are more popular) is not supported by the data.

```
### 4) Are songs in major key more popular than songs in minor key?
major = df[df['mode'] == 1]['popularity']
minor = df[df['mode'] == 0]['popularity']

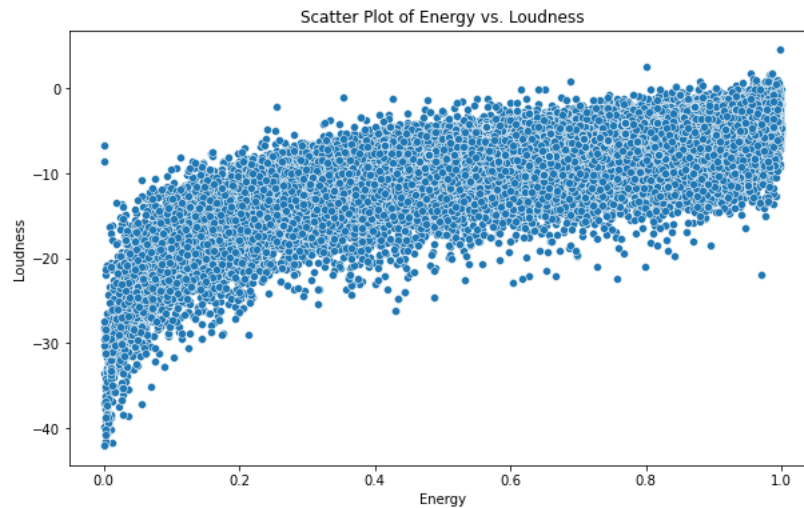
u, p = stats.mannwhitneyu(major, minor, alternative='greater')
print("Mann-Whitney U test statistic:", u)
print("P-value:", p)
```

*Figure 6: Codes of U-Test*

5) Energy is believed to largely reflect the “loudness” of a song. Can you substantiate (or refute)

that this is the case?

To assess whether I can substantiate the idea that energy is believed to largely reflect the “loudness” of a song, I decided to use correlation to investigate the relationship between those 2 variables. I found the correlation coefficient between song energy and popularity is **0.7749**, which shows a strong positive relationship between them. Therefore, I can substantiate the sentence that as the energy of a song increases, its popularity tends to increase as well.

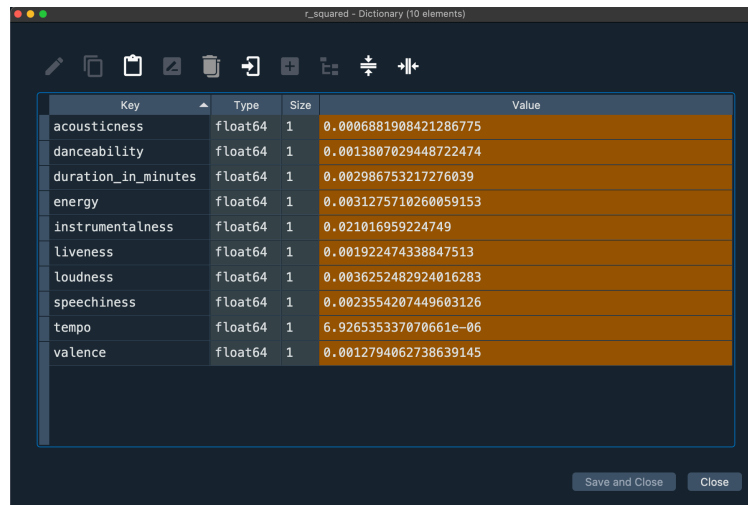


**Figure 7:** Energy vs. Loudness ( $r = 0.7749$ )

6) Which of the 10 individual (single) song features from question 1 predicts popularity best?

How good is this “best” model?

The model utilizes individual song features to predict popularity scores. Thus, I applied a linear regression model to forecast popularity using each feature individually, comparing the R-squared values to assess each model's effectiveness. Eventually, I found the best predictor of popularity is “instrumentalness” with an  $R^2$  of **0.0210**. We are using R-squared because it measures the proportion of the variance explained by the model. Here, we found a small R-squared even for the “best” feature, meaning that explained variance by the model is only **2.1%**.



Key	Type	Size	Value
acoustictness	float64	1	0.0006881908421286775
danceability	float64	1	0.0013807029448722474
duration_in_minutes	float64	1	0.002986753217276039
energy	float64	1	0.0031275710260059153
instrumentalness	float64	1	0.021016959224749
liveness	float64	1	0.001922474338847513
loudness	float64	1	0.0036252482924016283
speechiness	float64	1	0.0023554207449603126
tempo	float64	1	6.926535337070661e-06
valence	float64	1	0.0012794062738639145

*Figure 8: Lists of R-squared.*

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

I started with multiple regression using the 10 features, and I found the R-Square for the multiple regression model is 0.0477. The R-squared has improved approximately **0.0267** compared to the best model in question 6. This indicates that the model explains **4.77%** of the variance in the outcome, which is an improvement over the **2.1%** explained by the single-feature model using “instrumentalness.”

The small improvement in R-squared could be due to **collinearity** among the features.

Collinearity occurs when some of the predictor variables are correlated with each other, which can reduce the overall effectiveness of each predictor in explaining the variance in the dependent variable. Despite the increase in  $R^2$ , the overall explanatory power of the model remains low, suggesting that other factors not included in the model might play a significant role in determining the popularity of songs.

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?

Initially, I **plotted a correlation matrix** using a heatmap (Figure 7). We can see that some features have a strong relationship with each other. Therefore, it is important to use PCA to reduce dimensionality and potential noise. Then, I **normalized** the data so that we can compare everything within the same scale and started with PCA. Next, I **plotted the components** (Figure 8) and used Kaiser criterion to extract the components. I also plotted **the loading matrix** to see how each components are weighted by the features (Figure 9). For instance, PC1 is largely weighted by energy (-0.54), loudness(-0.54), and Acousticness(0.47). Eventually, I **calculated the explained variance** (eigensum) of the 3 components extracted, which is **57.36%**. It means that those 3 components explain **57.36%** of the variance.

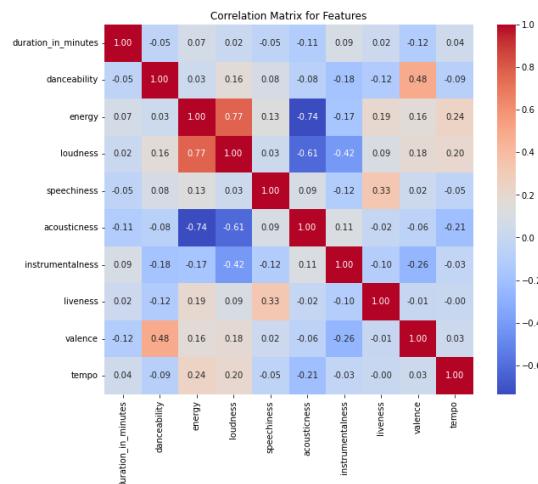
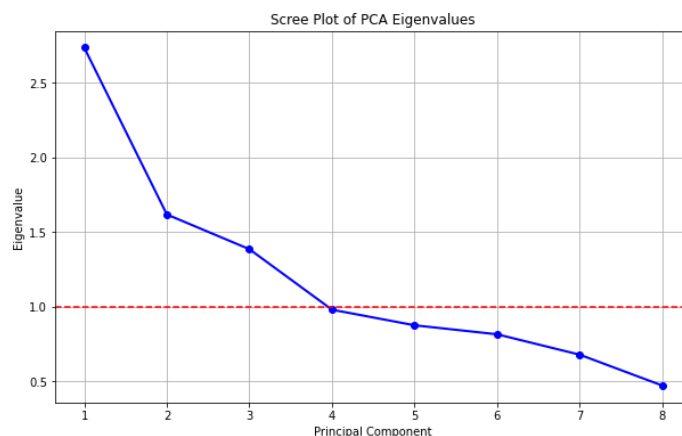
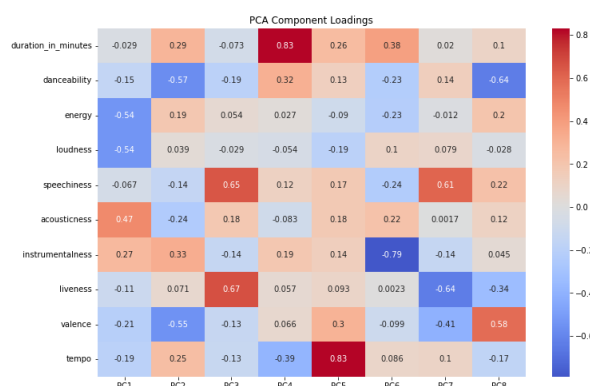


Figure 9: Correlation Matrix for features.



**Figure 10: Eigenvalues (Kaiser Criterion).**



**Figure 11: Loading Matrix**

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?

No, I can't predict a song's mode from valence. I have used the logistic regression model to see if I can predict the mode from valence. I separated the mode into major and minor, then set up the mode using cross-validation to train and test the model. I assessed the model using ROC curve, but the area under the ROC curve has a value of **0.50**, meaning that the classifier performs no better than the random chance, so the predictor is bad. To find a better predictor, I did the logistic regression for each features again, and plotted a new graph (Figure 11). The AUC values pointed out that the better predictor can be acousticness (AUC = 0.56).



Confusion Matrix:

```
[[ 0 3980]
 [ 0 6420]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3980
1	0.62	1.00	0.76	6420
accuracy			0.62	10400
macro avg	0.31	0.50	0.38	10400
weighted avg	0.38	0.62	0.47	10400

*Figure 12: Confusion Matrix and Classification Report*

### **True Negatives (TN): 0**

The model did not classify any instances as True Negatives (class 0 correctly).

### **False Positives (FP): 3980**

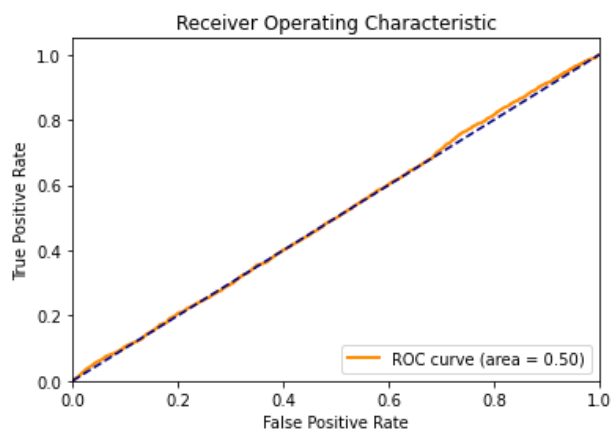
The model classified 3980 instances as False Positives (class 1 incorrectly when they were actually class 0).

### **False Negatives (FN): 0**

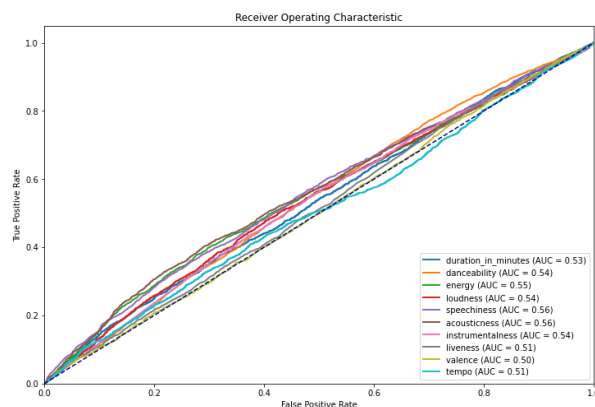
The model did not classify any instances as False Negatives (class 0 incorrectly when they were actually class 1).

### **True Positives (TP): 6420**

The model classified 6420 instances as True Positives (class 1 correctly).



*Figure 13: ROC curve of valence and mode.*

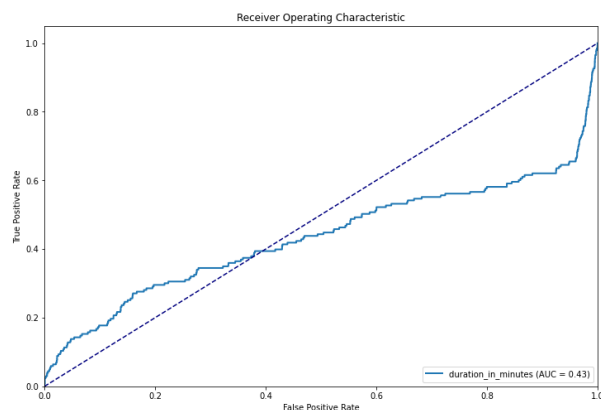


**Figure 14:** The ROC curves.

The logistic regression model provided a robust evaluation. However, the results indicate that valence alone is not a strong predictor of whether a song is in a major or minor key. The ROC AUC score of the model suggests that there is limited predictive power in using valence alone for this classification task.

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

Initially, I started with turning genre into binary numerical label according to if it is classical or not. Then, I set up the model using cross-validation for both duration and Principal components. Eventually, I drew the graphs to compare. We can see that principal components (AUC = 0.94) is a better predictor than duration (AUC = 0.43) because of having a high AUC value.



**Figure 15:** ROC of duration in minutes.

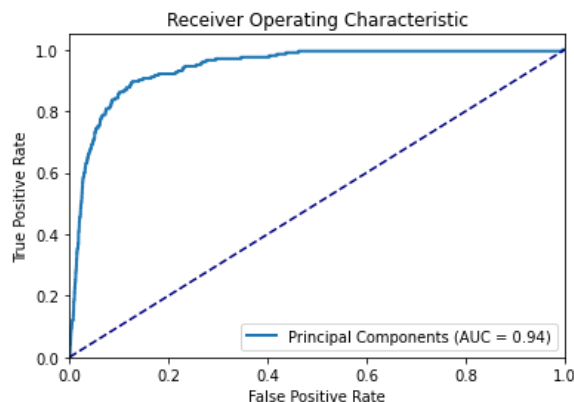


Figure 16: ROC of Principal Components.

11) Extra:

I am interested in analyzing whether certain genres are more popular or have distinct musical features such as danceability, energy, valence, and loudness. To do this, I initially create new data frame grouping the dataset by genre and include the mean of danceability, popularity scores and energy scores in terms of genres. There are 52 genres shown in this dataset, and I successfully get a data with genres and their musical features. After that, I selected the top 10 genre with the highest popularity scores and created a graph of boxplots to show their energy and danceability.

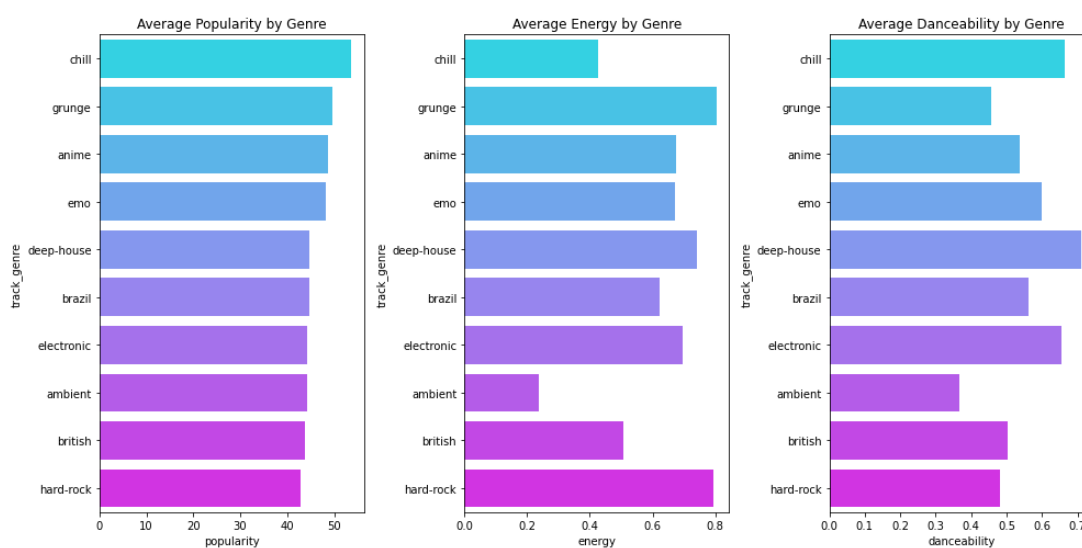


Figure 17: Boxplots for energy and danceability of the top 10 genres (sorted by popularity)

It is interesting to see from the graph above that Chill genre has the highest popularity and a high danceability scores but a low energy level.