Aesthetic Preference in Music

based on a dataset from

Prepared for

Data Science 2: Statistics for Data Science

by

Group 11

Benhamida, Jamil

Laudrum, Eric

Shah, Akash

Treshan, Waas

August 2, 2021

Content

Content 2
Objectives 3
Data Preparation 3
- Dataset 3
- Dataset pre-processing 3
- Dataset Cleaning 4
- Exploration 4

Analysis 5
- Hypothesis testing 5
- Predictive Model 7
  - Predicting energy level (OLS) 9
  - Predicting energy level (Random Forest) 9

Conclusions 10
References 10
Appendix 11

Objectives

Although it was previously believed that musical preference is rooted in the brain, a 2016 study suggests that “musical tastes are cultural in origin, not hardwired in the brain” (Trafton Anne, 2016). Given that aesthetic preferences in music is influenced by one’s culture and that every country has its own culture, is there any way to predict hit songs by country? Given all the financial scale of the music industry, being able to predict musical top 200 could be beneficial for any artist or music production company. Which parameters of a song (i.e. tempo, danceability, duration, words related to joy, etc.) would provide enough insights to determine hit songs per country and how do these parameters change from country to country (i.e. from culture to culture)?

Data Preparation

To answer these questions, we will investigate predictive modelling and correlation in the dataset through all the songs in Spotify’s Daily Top 200 Charts in 35+1 countries in the world for a period of over 3 years (2017 to 2020) retrieved from Kaggle. Raw metadata from Spotify was pre-processed through different steps such as feature engineering and natural language processing to generate this dataset. This preprocessing allowed us to focus our investigation on a thorough statistical analysis of the data.

Dataset

The dataset was retrieved from Kaggle. It includes some of the basic information of the songs from the top 200 charts from 35 countries in addition to data specific to global top charts. Some key identifiers captured included: the country, title, artist, album, genre, etc. Additionally, the dataset included different musical variables such as: the measure of energy (perceptual measure of intensity and activity), key (overall key of the track), loudness (in decibels), liveliness (presence of an audience in the recording), duration of the track, etc. This dataset had a total of 150 features and 26,444 entries.

Dataset pre-processing

The Kaggle users that published the dataset engineered a couple of fields. One of the fields that can be the most useful is the engineered field ‘popularity’. It includes the number of days a song stayed in the charts and at which position (adjusted with modificator to give more weight to top positions). Additionally, approximately 70 dummy fields were created with the country in which the song was evaluated for, the song genre, and if the song was in the top 50 and top 10 in the associated country.

The Kaggle users also processed the english songs only through Natural Language Processing (NLP) and Latent Dirichlet Allocation (LDA) to assign tone, emotion and topic. Then, approximately 50 dummy fields were created with the tone, emotion and topic. Given that NLP and LDA were only performed on English songs, these musical variables contain a lot of null values. Based on the countries where Spotify operates, a parallel analysis was performed to identify countries that had a majority English-speaking population by using a supplementary dataset from the United Nations.

We decided to use the engineered field “Popularity” as our benchmark to determine the level of popularity of an entry. As explained by the Kaggle user that uploaded the dataset, this field was already engineered in the dataset and represents a score from 1 to 200 for each song, where the #1 ranked song was assigned 200, #2 ranked song was assigned 199, #200 ranked song assigned 1 and so on. Then, this rank was multiplied by a modifier: 3 for #1, 2.2 for the #2, 1.7 for the #3, 1.3 for #4-10, 1 for # 11-50, 0.85 for #51-100, 0.8 for #101-200. This rank and modificator multiplication was performed for every day for the 3 years of the dataset (2017 to 2020) to provide a final score that is normalized for each song in the top 200 for all 35 countries and the global top charts.

Dataset Cleaning

Although most of the dataset was already cleaned with dummy fields, the main features we decided to use for our analysis needed to be transformed from string to float. All the null values were assigned to NLP and LDA entries for songs that were not in English. These were outside our scope.

Exploration

Given the size of the dataset and the number of features to consider, we quickly realized that this dataset would require a more advanced methodology such as machine learning or clustering to accurately predict the popularity of a song based on its indicators and a given nation's taste preferences. Therefore, we decided to reduce the scope of our analysis to a handful of key indicators, such as: danceability, energy, loudness, speechiness, acoustics, instrumentalness, liveliness, valence, tempo, and artist followers.

Upon visual exploration of the basic features of interest compared to the popularity for the global subset, no trend was observed at first through scatter plotting. A copy of this image is provided in the appendix. Given the high number of points, the scatter plot didn't provide enough insight to help direct our investigation. Through a joint plot of histogram and kernel density estimate (KDE), the visual inspection of potential correlation between the features of interest and popularity was not successful.

A Spearman correlation analysis was performed between the popularity and each individual dummy field. The Spearman correlation was selected for this exploration as we didn't observe any linear relationships in the plots generated comparing these factors to the measured popularity index. This correlation evaluation allows a monotonic relationship assessment. A table highlighting the factors with the highest Spearman correlation are shown below:

```
Popularity          1.000000
Popu_max            0.866468
Top50_dummy         0.653970
Top10_dummy         0.416660
single              0.122047
Artist_followers    0.118070
album               0.114534
latin               0.108534
else                0.084328
metal               0.083274
rock                0.070909
pop                 0.056520
Taiwan              0.054980
Ecuador             0.037712
Argentina           0.036799
Name: Popularity, dtype: float64
```

The strongest correlation with the popularity for the global entries were as expected. The factors which possessed the highest Spearman's correlation with the engineered popularity variable included: popularity maximum (maximum popularity a song reached during the time period of the whole dataset), top 50 (if a song ever made it into the top 50), and top 10 (if a song ever made it into the top 10). As these indicators were described as key factors used to quantify the popularity of each song the high level of correlation (relative to the bulk of additional factors available) was to be expected.

This posed an issue in generating predictive models to estimate the popularity of a song based on certain key musical factors. This will be discussed in detail in the section titled 'Predictive Model'.

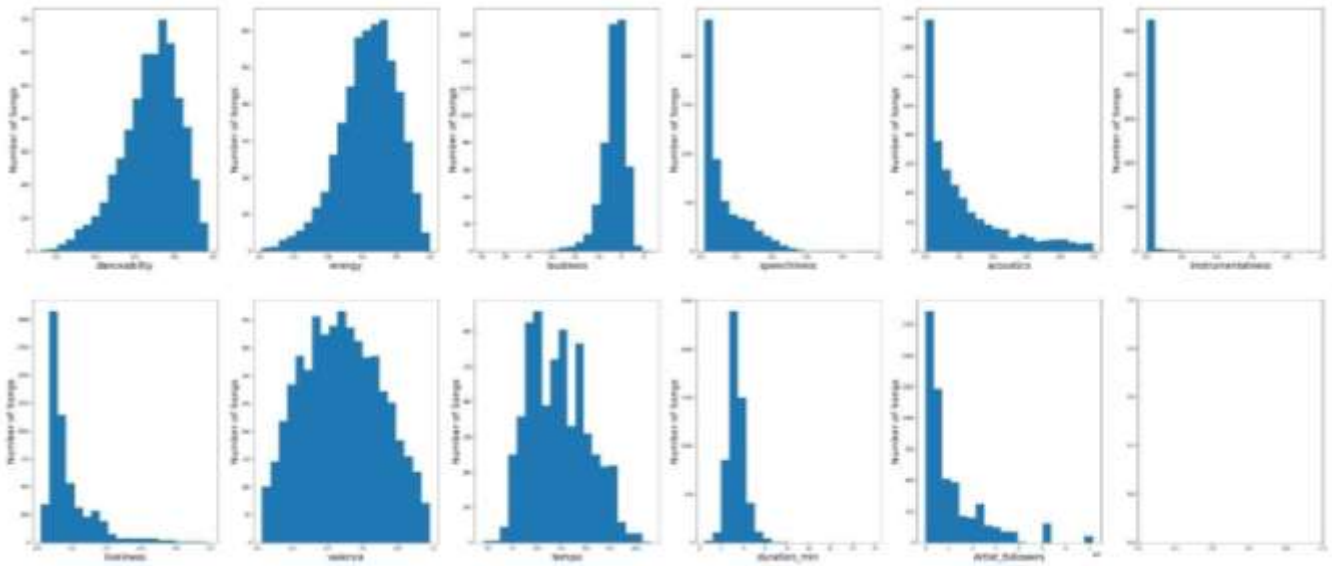
Analysis

Hypothesis testing

The dataset provided a plethora of information that we used to test various scenarios; however, it was a challenge to conduct hypothesis testing on all key indicators due to non-normal/skewed distributions in the data. The histograms below show the distribution of a few key musical indicators available in the dataset and were used for some preliminary hypothesis testing.

The main features investigated include:

- Popularity - A popularity grade generated by the team that pre-processed raw Spotify data through NLP and longevity of specific titles remaining on the Top 200 lists of their respective countries.
- Danceability - A combined variable generated by the dataset creators that incorporated tempo, rhythmic stability, and beat strength to assess how suitable a track is to dancing. This is measured between 0 and 1.
- Energy - A perceptual measure of the intensity of a track based on dynamic range, perceived loudness, timbre, onset rate and general entropy. This is measured between 0 and 1.
- Loudness - Overall loudness of the track measured in decibels.
- ‘Speechiness’ - The presence of spoken words in a track. This is limited between 0.0 to 1.
- Acoustics - A subjective measure of the ‘acousticness’ of a track.
- ‘Instrumentalness’ - A measure of a track’s lack of vocals, this is limited between 0.0 and 1.
- Liveliness - A measure of audience sound predominantly used to discern live recordings.
- Valence - A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)
- Tempo - Overall estimated paces of the song measured in BPM (beats per minute)
- Duration_min - The song length in minutes.



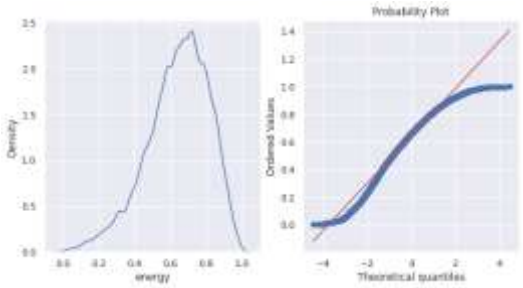
A couple of scenarios were tested to check the variability in specific key musical traits compared to global standards. Additionally, a country to country benchmark was considered to determine whether the selected features are significantly different at a 95% confidence level.

As discussed previously, plenty of basic indicators had skewed/non-normal distributions in data, as such, we decided to further investigate whether energy, loudness and valence were significantly different in English-speaking and Nordic countries, Southern European and Portuguese heritage countries and Spanish-speaking countries when comparing with global music hits.

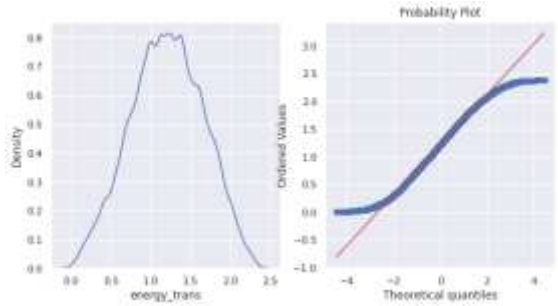
To improve the normality of these factors, we transformed the data columns using a Yeo-Johnson power transformation. This transformation is similar to the Box-Cox transformation; however, it is applicable for all values of a given variable. Meanwhile, the Box-Cox transformation is only applicable for strictly positive values; as such, this was not a valid method to apply to the loudness variable especially.

A summary of the transformed distribution for ‘energy’ is provided below. All other transformed distributions may be found in the appendix.

`Before conversion:`



`After conversion:`



After transformation we summarized the data by grouping according to clusters based on language and key demographics available in the dataset. We obtained the mean, standard deviation and sample numbers for each cluster and completed hypothesis testing (z-tests) to assess if any ethno-linguistic group was statistically different from the global dataset benchmark.

A summary of the p-values obtained are shown below:

	cluster	feature	p_value
0	southern europe and portuguese heritage	energy_trans	9.967419e-01
1	english speaking and nordic	energy_trans	5.342931e-01
2	spanish speaking	energy_trans	1.000000e+00
3	southern europe and portuguese heritage	danceability_trans	9.190091e-16
4	english speaking and nordic	danceability_trans	7.542977e-28
5	spanish speaking	danceability_trans	9.922601e-01
6	southern europe and portuguese heritage	valence_trans	9.921777e-01
7	english speaking and nordic	valence_trans	2.979797e-02
8	spanish speaking	valence_trans	1.000000e+00

Compared to the global average, the following results were statistically significant at the 95% confidence level:

	cluster	feature	p_value
0	southern europe and portuguese heritage	danceability_trans	9.190091e-16
1	english speaking and nordic	danceability_trans	7.542977e-28
2	english speaking and nordic	valence_trans	2.979797e-02

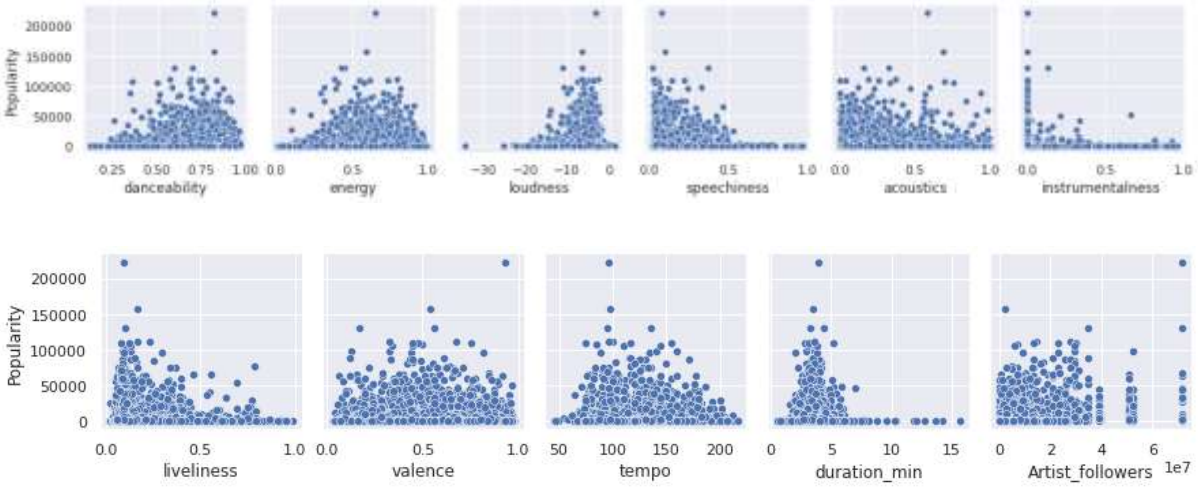
This analysis indicates that danceability is significantly different in ‘Southern Europe and Portuguese heritage’ and ‘English-speaking and Nordic countries’ compared to the global average. Valence is significantly different in English speaking and Nordic countries.

Cluster	energy_trans	danceability_trans	valence_trans
english speaking and nordic	1.188345	1.361032	0.426630
global	1.187829	1.429564	0.431404
southern europe and portuguese heritage	1.205969	1.379456	0.437529
spanish speaking	1.293777	1.444826	0.511371

As such, the group of ‘Southern Europe and Portuguese heritage’ and ‘English-speaking and Nordic countries’ appear to favour danceability less than the global average. ‘English-speaking and Nordic countries’ appears to favour less positive music based on its lower valence compared to the global mean.

Predictive Model

The key challenge of generating a predictive model was identifying meaningful relationships in the data as there were numerous features in the dataset. The main ideology was to check whether any of the features had an evident relationship with music popularity. To check relationships we used a scatter plot. The scatter plots below highlight the global dataset.



Per the scatter plots, we did not see any relationship between the key musical features and calculated popularity.

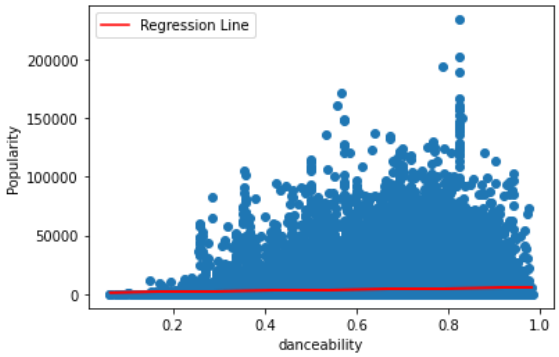
To further investigate these weak relationships, we obtained OLS models for each continent and the Global dataset as a whole. We noticed extremely weak R-Squared values in all models generated irrespective of geographic region.

Please see below an example of OLS regression of all basic musical features at the global level.

```
Grouping:  Continent  -  Global
Number of variables:  12
Rows:  5460
                                 OLS Regression Results
==============================================================================
Dep. Variable:             Popularity   R-squared (uncentered):                   0.173
Model:                            OLS   Adj. R-squared (uncentered):              0.171
Method:                 Least Squares   F-statistic:                              103.3
Date:                Thu, 29 Jul 2021   Prob (F-statistic):                   1.32e-214
Time:                        02:19:23   Log-Likelihood:                         -58998.
No. Observations:                5460   AIC:                                  1.180e+05
Df Residuals:                    5449   BIC:                                  1.181e+05
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
danceability      7857.7598   1041.570      7.544      0.000    5815.866    9899.654
energy           -1291.3339   1167.093     -1.106      0.269   -3579.303     996.635
loudness           289.5554     72.255      4.007      0.000     147.906     431.204
speechiness      -5993.1420   1381.370     -4.339      0.000   -8701.179   -3285.105
acoustics         1850.5152    790.017      2.342      0.019     301.766    3399.265
instrumentalness -1940.1170   2173.457     -0.893      0.372   -6200.960    2320.726
liveliness       -2571.7957   1183.068     -2.174      0.030   -4891.081    -252.510
valence           1517.7631    835.486      1.817      0.069    -120.123    3155.649
tempo               -4.7940      5.212     -0.920      0.358     -15.011       5.423
duration_min       336.9856    179.907      1.873      0.061     -15.704     689.675
Artist_followers     0.0002   1.24e-05     12.313      0.000       0.000       0.000
==============================================================================
Omnibus:                     5659.632   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           418555.420
Skew:                           5.126   Prob(JB):                         0.00
Kurtosis:                      44.650   Cond. No.                     2.31e+08
==============================================================================
```

To further confirm this we ran OLS models for each feature to identify relationships with music popularity individually. As expected, the models obtained did not adequately describe the variability in data.

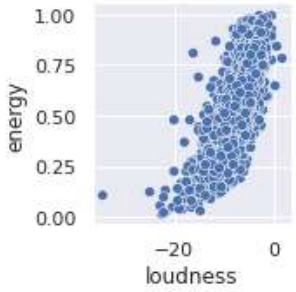
Grouping: Continent - Europe
Number of variables: 2
n: 91134



Therefore, we concluded that we are unable to create a model to predict popularity based on various musical features.

Predicting energy level (OLS)

To determine if any potential relationships were prevalent between different factors a scatter matrix was generated. This would then be used to obtain a predictive model. From this, we were able surmise that energy levels and loudness had a positive relationship. Therefore, we created an OLS model and also included the song type as dummy variables in different model iterations.



Scatter Plot of Energy vs. Loudness

Model 1

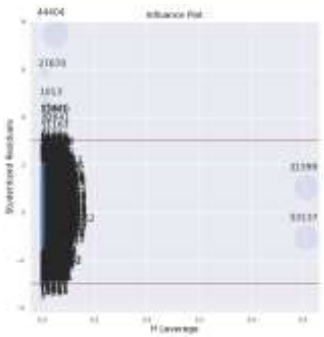
The model obtained had a R-squared value of 0.54 which suggested that the variability was slightly explained by this model; however, this was still not adequate as this is only considered a moderate effect size. To potentially improve the explained variability (R-squared by extension) additional factors were considered.

Model 2

To further improve the model we included the genre types as supporting variables. These were binary variables.The explained variability increased slightly suggesting further iteration was required.

Model 3

To further improve the model we wanted to analyse points that are influencing the model as outliers. We used the 'Get_Influence" function to get a table with all parameters. We calculated `dffits score (`square root of (number of parameters/number of observations).
The influence plot below provides a visualization of some of the key outliers found.



Model 4

To iterate further, we removed the features/parameters that possessed a p-value that was greater than 0.05 as they were not statistically significant. A final model was obtained with the trimmed list of factors. This approach did not appreciably impact the model's performance relative to model 3. This said, the reduced number of factors required to obtain the same level of explained error is favourable.

Predicting energy level (Random Forest)

The random forest model was also used to predict energy levels to determine if any improvement could be obtained. We used the dataset that was cleaned with outliers and executed the random forest model. The accuracy score function was not possible to execute as it had continuous variables. Therefore, the R-squared model was used to evaluate the model. Overall the random forest accuracy was much lower than the OLS model.

Conclusions

In our analysis, we were not able to observe and predict trends pertaining to the popularity of a given song and its distinct musical features using this dataset.

The dataset allowed us to explore and test a few features such as energy, danceability and valence to justify whether there are significant differences on the levels compared to the global average. We determined that danceability is significantly different in 'Southern Europe and Portuguese heritage' and 'English-speaking and Nordic countries' compared to the global average. Valence is significantly different in English speaking and Nordic countries compared to the global average as well. Using more complex methods involving machine learning it may be feasible to utilize this information to generate models to predict popularity.

Although this dataset didn't allow us to predict popularity based on key musical indicators through OLS, it did prove its value to identify a trend between song energy. We were able to predict energy levels with loudness and the song genre. As discussed, a more powerful model through complex machine learning methods such as deep learning or employing various sorting algorithms.

Different hypotheses could explain why it is not possible to predict what parameters will create a hit song. Additionally, since music is art, there are so many factors, globally and socially, on a micro or a macro-scale, that could influence musical preference of different demographics in the same country. Such information was not available for this analysis.

References

Herremans, Dorien. May 2019. Towards Data Science: Data Science for Hit Song Prediction. Retrieved from: https://towardsdatascience.com/data-science-for-hit-song-prediction-32370f0759c1 [Accessed: July 4, 2021].

Kaggle. Updated July, 2021. Spotify HUGE Database - Daily Charts Over 3 Years. Retrieved from: https://www.kaggle.com/pepepython/spotify-huge-database-daily-charts-over-3-years [Accessed: July 4, 2021].

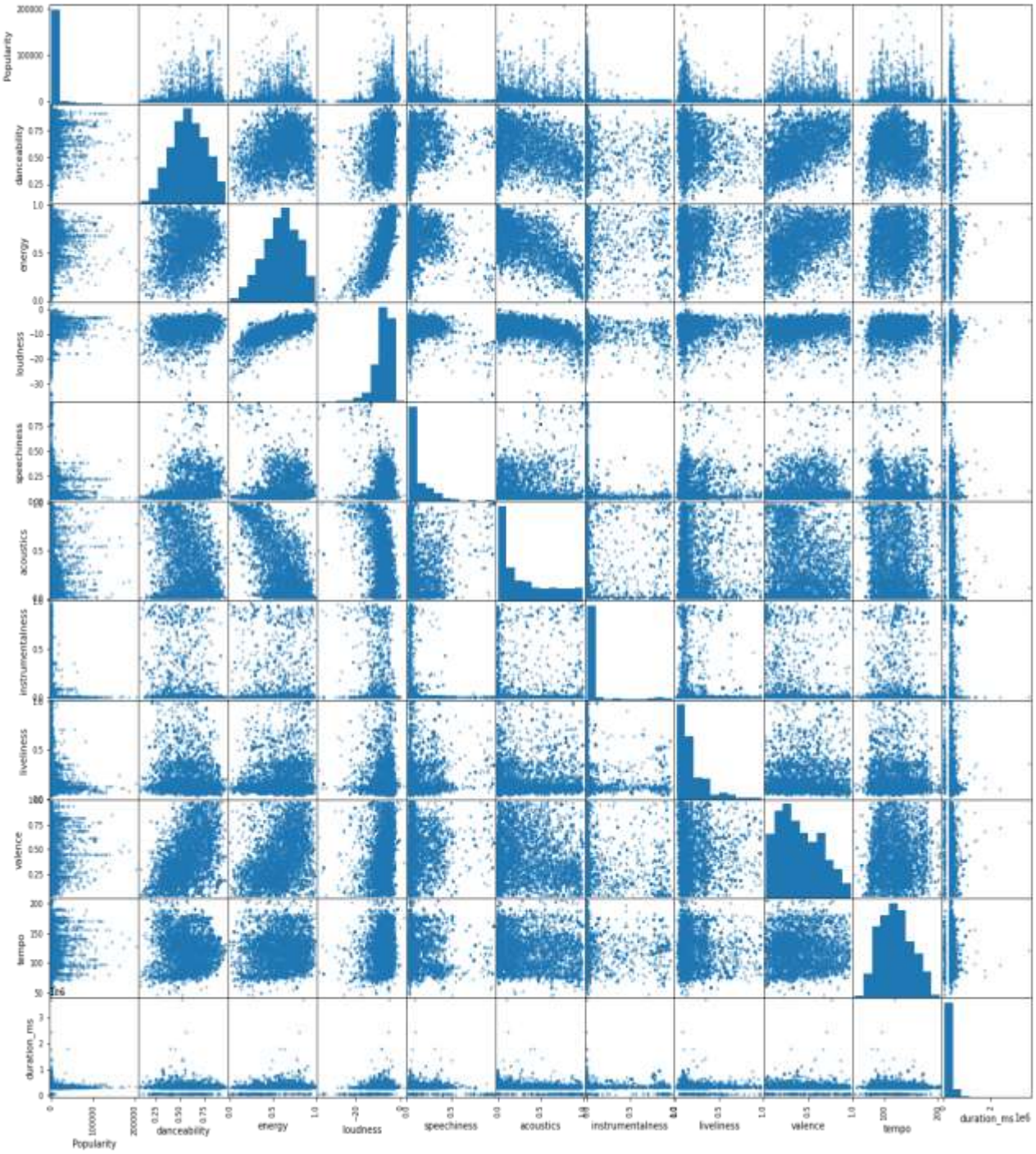
Spotify. Company Info. Retrieved from: https://newsroom.spotify.com/company-info/ [Accessed: July 28, 2021].

Trafton, Anne. July 2016. MIT News: Why We Like The Music We Do. Retrieved from: https://news.mit.edu/2016/music-tastes-cultural-not-hardwired-brain-0713 [Accessed: July 4, 2021].

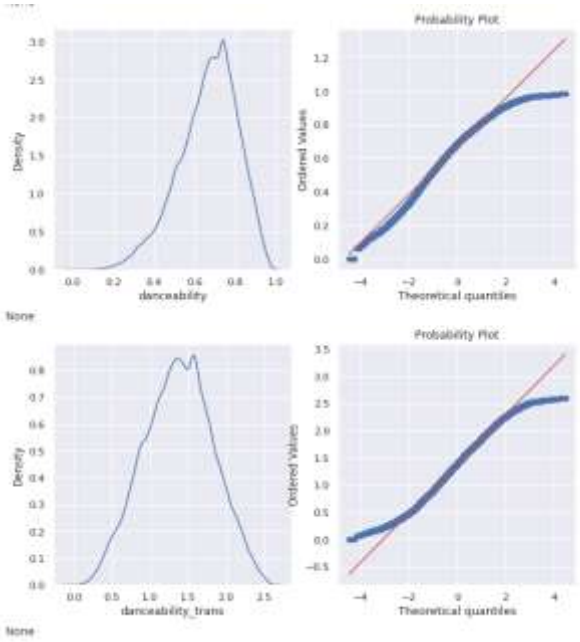
United Nation. United Nation Data. Retrieved from: https://data.un.org/Data.aspx?q=language&d=POP&f=tableCode%3a27 [Accessed: July 9, 2021].

Appendix

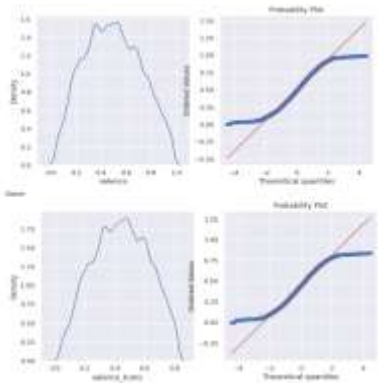
The following scatter matrix was utilized to identify any relationships in the features available.



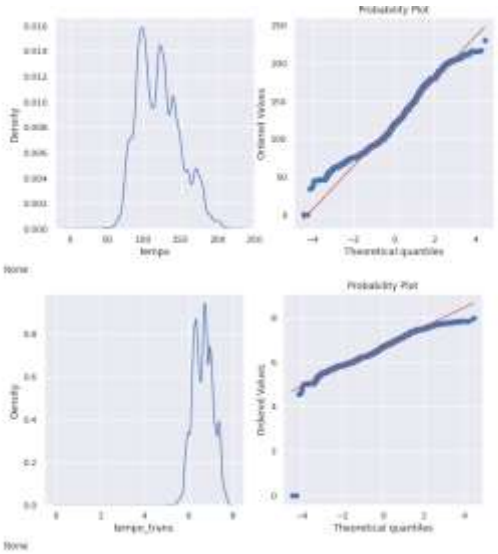
The following plots show Yeo-Johnson transformed data for loudness, valence, and danceability:



Top: Non-Transformed Danceability Data
Bottom: Yeo-Johnson Transformed Danceability Data



Top: Non-Transformed Valence Data
Bottom: Yeo-Johnson Transformed Valence Data



Top: Non-Transformed Tempo Data
Bottom: Yeo-Johnson Transformed Tempo Data