Han Le

Final Project: Spotify Analysis

Due: May 9th, 2021

**I. Introduction:**

Music is an essential necessity of life which brings happiness, motivation, and joy to the audiences. Music also connects everyone regardless of age, gender, and race. The music industry keeps growing with time, so people not only be able to enjoy the music, but also address many questions using its data. In this project, the provided Spotify data will be used to answer some interesting questions regarding tracks and artists.

Specifically, the project is divided into two parts. The first part addresses a general interest such as the most popular key, or Ed Sheeran, and Taylor Swift’s most popular songs. The second part will focus on the lyric analysis of the two selected singers (Ed and Taylor). R code will be used to extract the most used adjectives in their songs and compared their music vibe. The R program is also used to collect, clean, and process the data to support the corresponding problem statement/interest, and provide graphs which will visualize the data.

**II. Packages Required:**

* wordcloud: Plot a word cloud.
* dplyr: A grammar of Data Manipulation.
* genius: Easily access song lyrics from genius.com.
* lattice: Trellis Graphics for R.
* gridExtra: Miscellaneous Functions for “Grid” Graphics.

**III. Data Preparation:**

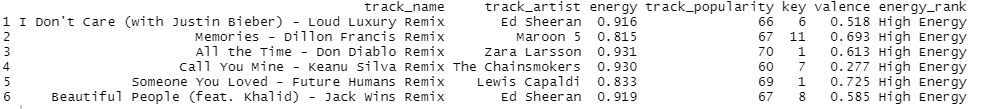
The Spotify data is provided in dataset 4 through MyCourses. This set of data is created by Kaylin Pavlik where she used the spotifyr package to collect about 5000 songs from 6 different categories such as EDM, Pop, R&B, Rock and Latin. To import the data to R, the data file must be moved to the same folder as R file, and the working directory must be set to that folder. There are 22 variables in this data set which is all not be used. The data is trimmed down to six variables to support the purpose of the project including: track\_name, energy, track\_popularity, key, and valence. The energy variable is then grouped into three different categories: Low Energy, Medium Energy, and High Energy based on the energy’s score. This categorical variable will later be used in plotting the distribution of the track\_popularity variable. The rank is distributed as below:

[0, 0.35] = Low Energy

(0.35, 0.7] = Medium Energy

(0.7, 1] = High Energy

It appears that there are some repeated rows in the dataset. Hence, distinct() function is used to eliminate those duplications. The printed head of trimmed data is below:



**IV. Exploratory Data Analysis:**

*A. General Interest:*

The first problem statement needed to be addressed is the most popular key used in all songs. Using tapply() function to create the table below:



The integers 0-11 represent the overall key of the track and map to pitches using standard Pitch Class notation. For example, 0 = C, 1 = C#/Db, 2 = D, and so on. Based on the table above, the key 1 = C#/Db appears to be used the most in this category.

Now, the data continue to be explored to find the most popular songs of Ed Sheeran and Taylor Swift. The grand data is extracted by track\_artist variable respected to the desired artists. Then the subset for each artist is extracted again by track\_popularity column with the maximum value. The results are shown below:



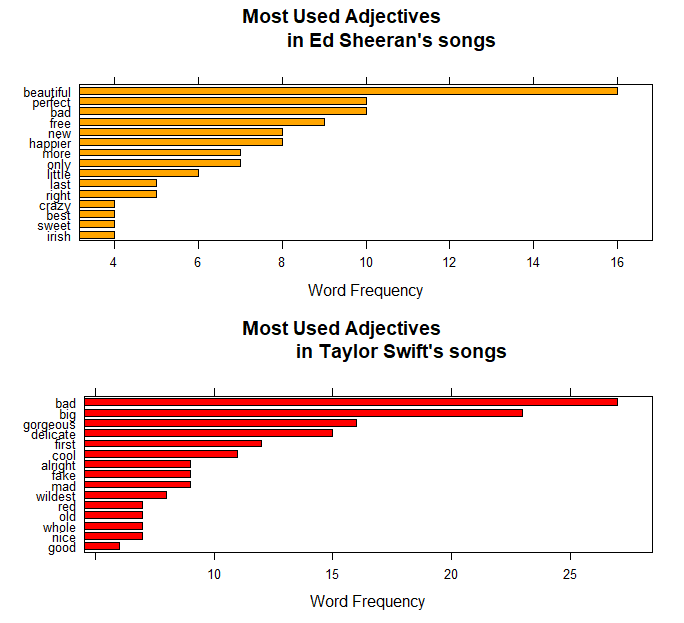


To have the condensed result, the data is further trimmed down to get rid of any other irrelevant variables, so that the result only consists of track\_name, track\_artist, and track\_popularity. Based on the outcome, the most popular song of Ed Sheeran is ‘South of the Border (feat. Camila Cabello & Cardi B), and ‘You Need to Calm Down’ for Taylor Swift.

*B. Lyrics Analysis:*

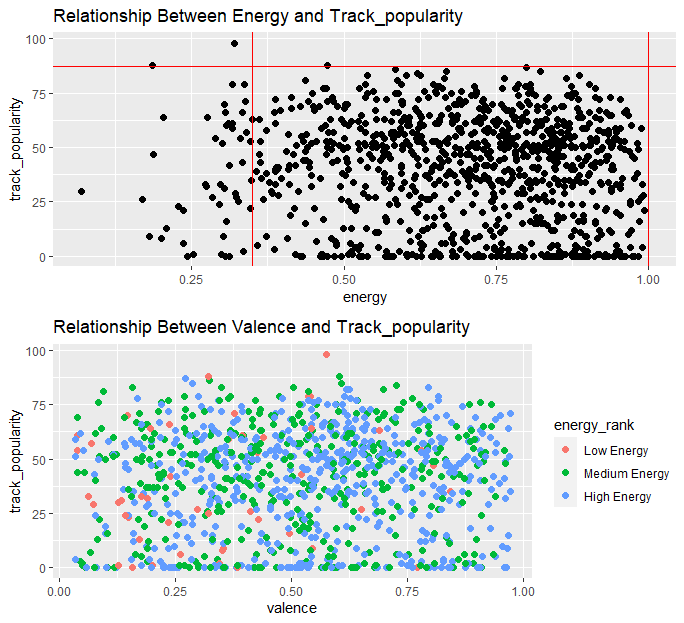
There are couple of interested inquiries that can be discovered using this dataset. Does Ed Sheeran’s or Taylor Swift’s songs have more negative vibe? Are there any relationship between energy and track\_popularity, or between valence and track\_popularity in all the songs? And how the track\_popularity is distributed respected to energy\_rank? Those questions will be answered in the second part of the project.

The lyric of the song is obtained using genius package. The required inputs are the name of artist and the name of the song. It is important to have a clean set of words for analysis, so each lyric has been through a cleaning process: eliminating punctuation, converting abbreviation to whole word, and converting words to lowercase for consistency. All the steps are written under a function called “cl”. After having the clean lyric, its adjectives will be extracted using UDPipe package. The UDPipe model provided by the UDPipe community for English language is downloaded and loaded in R. Then, each lyric is passed through annotate process to determine the parts of speech of each word. All the words tagged as adjective are extracted and store in a vector. The repeated process is applied for 12 different songs using “for loop”. Once having a combined lyric data set of Taylor Swift and Ed Sheeran, the problem statement can be addressed using graphs. Below is the bar chart of the most used adjectives of the two artists.

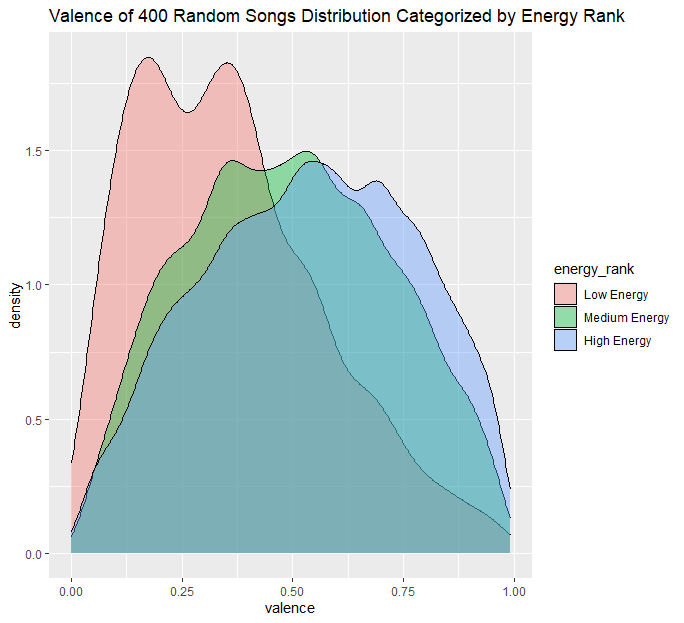


The plot is generated using barchart command in UDPipe package, and they combined using gridExtra package. The most used adjective across Ed Sheeran’s songs is “beautiful” while the word “bad” is used the most in Taylor’s Swift songs. The result makes sense because Ed Sheeran is well known for many positive songs which deliver the happy, cheerful, or romantic vibe to the audience. On the other hand, Taylor Swift’s songs tend to have a sad, depressed, or angry ambiance.

Next, we will explore whether there is any relationship between the variables in the data set. The graph below will describe whether energy or valence variable affect the track\_popularity.



It appears that there is no relationship between energy/valence and track\_popularity. The points are scattered randomly across the plot. However, both plots show that most of the songs have the medium-high energy. For the plot energy vs. track\_popularity, the songs are mostly distributed in the red rectangle, which displays the medium-high rank energy. For the valence vs. energy plot, the energy\_rank is differentiated by color. It can be concluded that the track\_popularity does not necessarily depend on the energy or the valence of the song. It can be a sad song, and people still love it. There are no boundaries or specific trends in the songs that can be analyzed using data because it mostly depends on the audience’s perspectives. If the vibe of the song fits with what people are currently feeling, they could give a high score in track\_popularity, and vice versa. Next, we will discuss the valence distribution respected to energy\_rank.



The density graph above is generated using ggplot2 package. Based on the result, the low valence song is spread mostly in the low energy rank region. From this finding, sad songs tend to have lower energy where the tracks feel slow, quiet, and chill. This way, the singer can fully express all types of emotion of the song and successfully deliver it to the listeners. The medium and high energy songs are likely to have higher valence score that the tracks display more positive atmosphere. They also have the normal distribution whereas the low energy category does not. It is because there are more songs in the medium and high rank than low rank. If more songs are collected and added to the dataset, the normal distribution will be better shown in those plots.

**V. Summary:**

There are couple of interesting findings after analyzing and visualizing the data. The most popular key that the composers use is C#/Db. Another finding is that the most popular song is “You Need to Calm Down” by Taylor Swift, and “South of the Border” by Ed Sheeran. For the analysis of the lyrics, there are no relationships between energy/valance and track\_popularity. The graph shows that the data points do not follow any specific pattern.

The analysis shows that Taylor Swift’s songs are more negative than Ed Sheeran’s. However, those songs are awhile ago, and it might not reflect truly the current vibe of the singer. Taylor Swift has many new songs with happier, more cheerful, and more romantic than her old ones. If those songs are added to the Spotify set, it would be interesting to see how her approaches change throughout time. The same logic could be applied to other singers as well. Having the up-to-date data would help a lot to do so.

Another issue is that the number of extracted adjectives from all the songs seem to change every time the code is run. It does not change the conclusion of the problem statements, but the “Most Used Adjectives” graph will be a little different. The error has not been found, so other people are welcome to improve this code.

Through the analysis, it is further proved that music is inspiration, creativity and emotional. There are no limits to what people can bring in the songs. There are also many different angles that the audiences can see, feel, and interpret to fit their current sentiment.