SEEM2460 Final Report

Identifying Trends and Indicators of Popular Music on Spotify

Group Members
Jason, Wang Cheuk Yeung 1155127081
Hans, Nathanael Junoes 1155147304
Felix, Tsui Fan Yau 1155143241

Table of Contents

1. Introduction

1.1 Context (Jason)

In the music industry, the objective is clearly to craft a song that resonates with the general public – to capture their emotions and rise in popularity as a result. However, art is inherently that of a subjective matter, and it can be difficult to pinpoint with utmost certainty the many factors that determine a popular song in a quantitative manner. Music producers and artists alike often find it challenging to cater to the ever-shifting musical taste.

In the era of big data, there is an influx of valuable data available ever since the digitization of entertainment. Companies are able to capture data and analyze the latest trend with the latest technology to incredible accuracy. Despite this, analyzing a subjective and ever-changing topic as with musical tastes remains a problematic topic to handle. This research aims to give a general insight into how one may approach such a problem.

1.2 Motivation (Hans)

The world's largest streaming platform is, without a doubt, Spotify; boasting an impressive 422 millions of users and 182 million subscribers across 183 markets. In consequence to the global, disastrous impact the pandemic has brought, Spotify has seen a significant rise in users over the last few years as many found themselves spending more time in the comfort of their homes. As much as a 27% increase in monthly active users was recorded in early 2021, and premium subscribers, which account for the majority of the streaming service giant's revenue, were up 24% in the late fourth quarter of 2020 [7].

Yet, despite the immense user and subscriber growth, audio music consumption decreased by 12.5% after the declaration of a pandemic by the World Health Organization in March of 2020. Owing to this, Spotify suffered a major loss of $838 million in revenue in the first three quarters of 2020 [1]. In a time when live concerts and tours are put to an end, it has been brought to our attention that the current situation is detrimental to artists seeking recognition. Evidently, there is a need in the market for better models to predict the trend and popularity of songs. It is in our best interests, therefore, that this paper may prove useful for the reader.

2. Data Modeling (Jason)

2.1 Data Description

Due to the considerable size of the user base on Spotify, there is a vast opportunity to analyze the constantly-evolving trends and preferences of songs worldwide. The source of the data comes directly from Spotify, which has compiled a wealth of information about the songs on the platform and publishes the numbers via the Spotify Web API. Owing to

Spotify hitting 1 million users in March 2011 [2], the top 100 songs per year between 2011 and 2019 are used to have a consistently larger sample size.

There are numerous online resources available that have compiled the Spotify numbers into useful datasets. For the purpose of this research, the dataset titled 'Top 100 tracks of Spotify from 2001-2019' from Kaggle [3] is used, with an additional self-explanatory parameter named "popularity" ranging from 1 to 100, where 100 denotes the most popular song.

2.2 Methodology

This research makes use of Tableau to visualize the overall trends and distribution of different parameters across the years, as explained in section 3. A general overview of which parameters have the greater effects can be shown, while section 4 makes use of more rigorous statistics to build a comprehensive regression model using only the most significant parameters. This is used to understand the inner mechanisms of what determines a song's popularity. In section 5, these parameters are used to do a clustering model – providing a means to predict the popularity of new songs to be released.

3. Trends & Visualization (Hans)

3.1 General Trend

The first step in the visualization process is to make sense of the general trend in the different metrics concerning the top popular songs in Spotify, across 2011 to 2019 inclusive. Here, there are 100 tracks inputted in each year, totaling to 900 such tracks[3]. All graphs are done with the visualization tool Tableau.

Brief description of each metric provided by the dataset along with our findings across all 900 tracks are as follows :

1. **Acousticness** - determines whether or not a song is acoustic. A score of 1 would mean an entirely acoustic track with no electronic modifications.
   -> Nearly 60% of popular songs fall between 0.00 to 0.12 in acousticness, resulting in an extremely right-skewed graph.
2. **Danceability** - a score of 1 would indicate a song that is extremely easy to dance to. Danceability is the culmination of tempo, beat, and rhythm.
   -> 47% of songs have values ranging from 0.61 to 0.76.
3. **Duration** - a duration, in milliseconds, of the song in question.
   -> 67% of songs last between 3m10s to 3m48s.
4. **Energy** - measures the intensity of a song. Soft, melodic tracks have low energy closer to 0 while loud, dynamic songs have high energy closer to 1.

-> The graph is left-skewed with an overwhelming number of songs exuding more than 0.5 energy(mid-value). 43% of all songs have values between 0.70 to 0.84.

5. **Key** - the key the song is in. Each key from A to G# is described as integers from 0 to 11.

   -> Approximately 10% of songs are composed in the key of A and Bb each. The lowest favored keys are C and G with 3% and 5% of songs respectively.
6. **Loudness** - measured in decibels and averaged across the track.

   -> 53% of songs have values ranging from -6.3 to -3.8dB.
7. **Mode** - a binary variable that indicates if the song is in major key (integer 1) or minor key (integer 0).

   -> 40% of songs are composed in minor while 60% are in major.
8. **Speechiness** - the portion of the song that consists of spoken words. Rap or other music that is highly composed of speech have scores closer to 1.

   -> The graph is significantly right-skewed with 63% of songs falling between 0.02 to 0.07 in speechiness.
9. **Valence** - describes how uplifting a song is. A song with a positive mood would have a score closer to 1.

   -> Roughly approximating a bell-shaped curve with 57% of songs with valence less than 0.5 (mid-value).
10. **Tempo** - the overall speed of the song.

   -> A multimodal graph. The three highest peaks are of ranges 117 to 130, 94 to 108, and 135 to 148 with percentages of 30%, 20% and 11% respectively.

3.2 Acousticness, Speechiness, and Tempo : Visualization

Arguably, some of these discoveries could be explained by common sense even without rigorous analysis, such as those of popular songs being mostly danceable ones. Other metrics, intuitively, may not have as much weight in determining the popularity of a song, such as those of duration and mode(refer to section 4 in regression analysis). In this section we have decided to focus on three metrics that, perhaps, are not immediately apparent and appear to suggest more bearing on the matter of a track's popularity.

The second step involves the comparison of three metrics, namely acousticness, speechiness, and tempo, across the years 2011 to 2019 in a specific range that exhibits the highest values in the step prior. By doing comparisons in these specific ranges, we hope to find an overall difference in trend across these 9 years. Afterwards, a direct comparison is made in the first and last years in the dataset, 2011 and 2019, to demonstrate the movement across the entire range.

Fig. 1, Fig. 2, and Fig. 3 show the metrics acousticness, speechiness, and tempo respectively compared across different years.

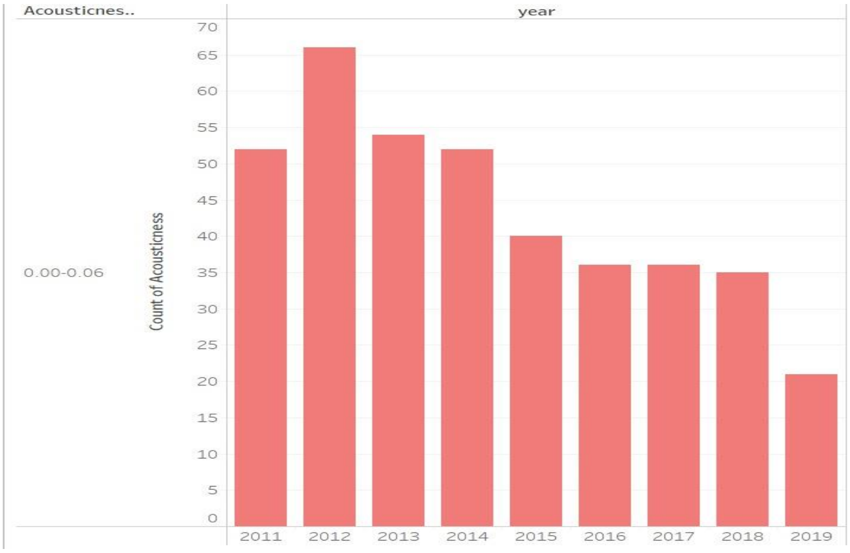


Figure 1.1: Count of Acousticness of range 0.00-0.06 for 2011-2019

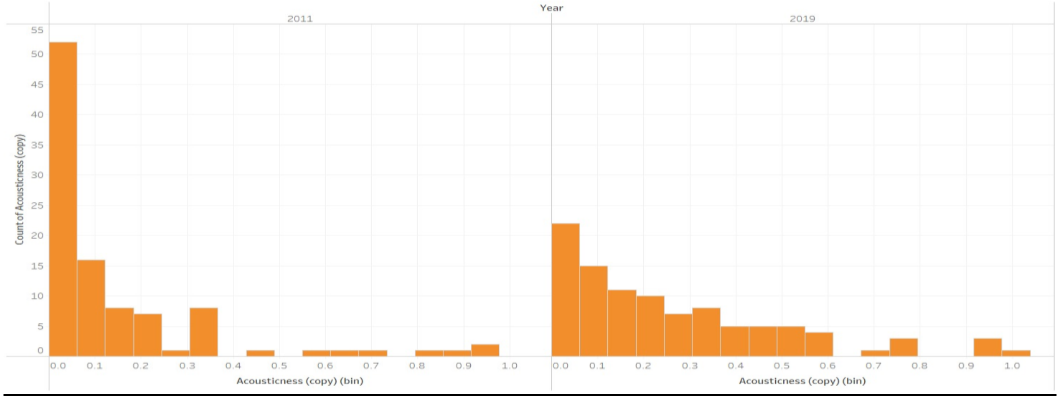


Figure 1.2: Acousticness for 2011 VS 2019

As shown in Figure 1.1, there is gradual decline in the lowest and most favored range of acousticness. Since low values of acousticness mean non-acoustic songs, there has been a general increase in the popularity of tracks that exhibit more acoustic or less electronic features. Figure 1.2 depicts in direct comparison the graphs of 2011 as opposed to 2019. Observe that in 2019, the skewness has decreased substantially.

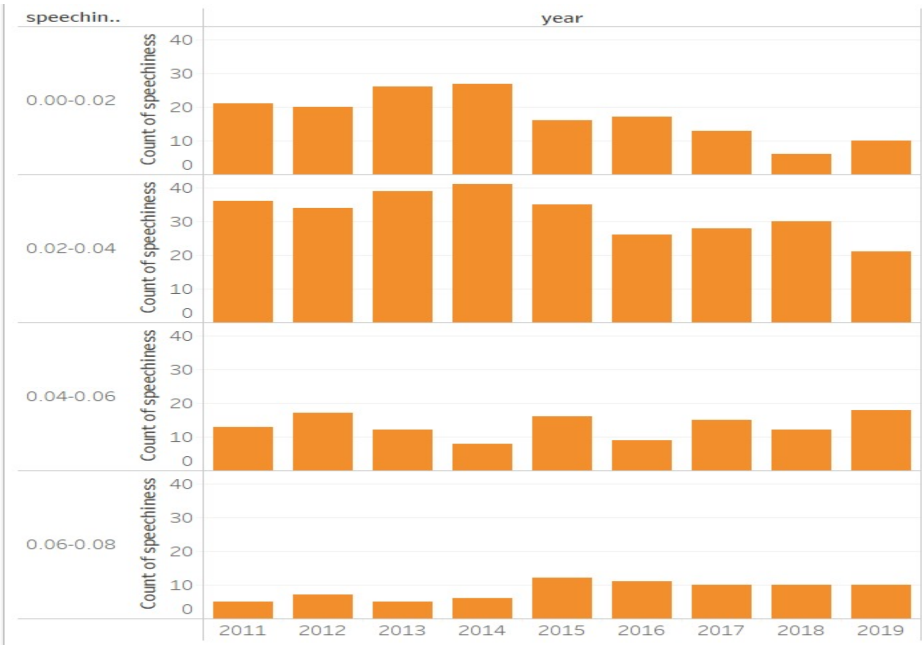


Figure 2.1: Count of Speechiness of range 0.00-0.08 for 2011-2019

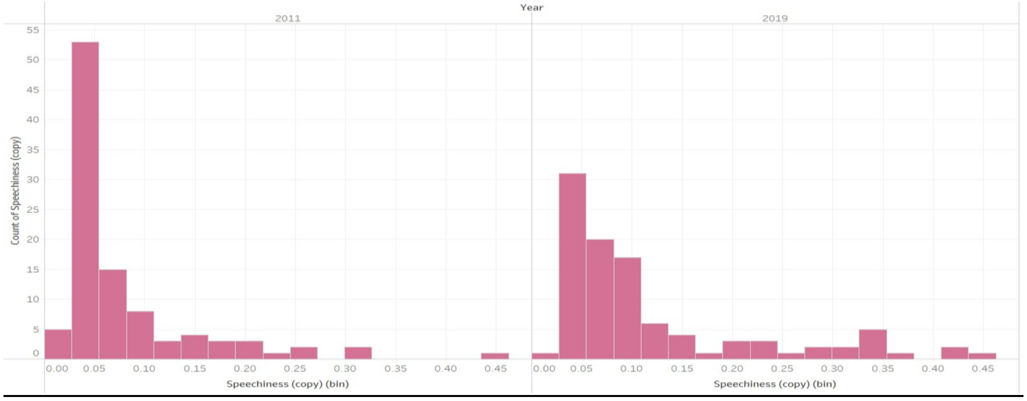


Figure 2.2: Speechiness for 2011 VS 2019

Figure 2 demonstrates a much similar trend to that of Figure 1. Overall, ranges of values that were once extremely common in earlier years were less so in the later ones. Again, Figure 2.2 depicts the decreasing skewness of the graph, allowing for greater diversity in the speechiness values of popular songs in 2019.

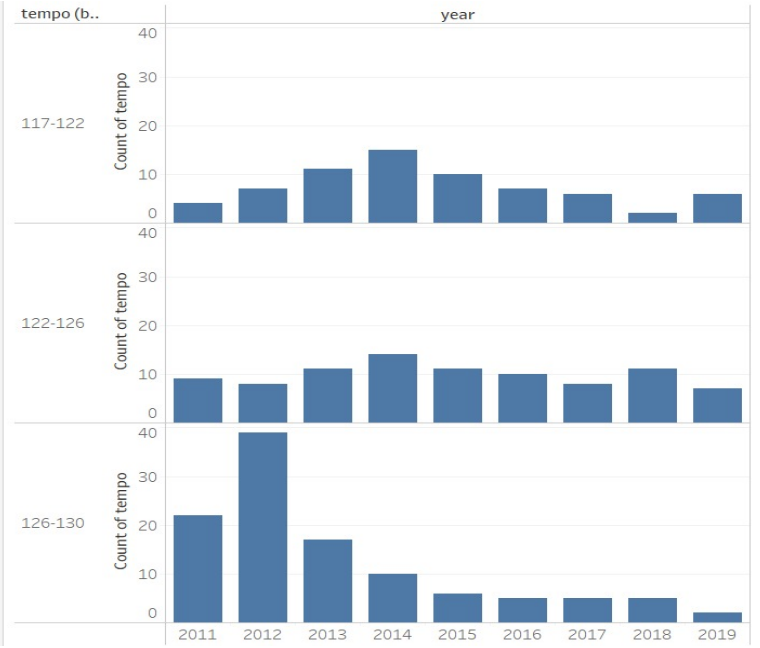


Figure 3.1: Count of Tempo of range 117-130 for 2011-2019

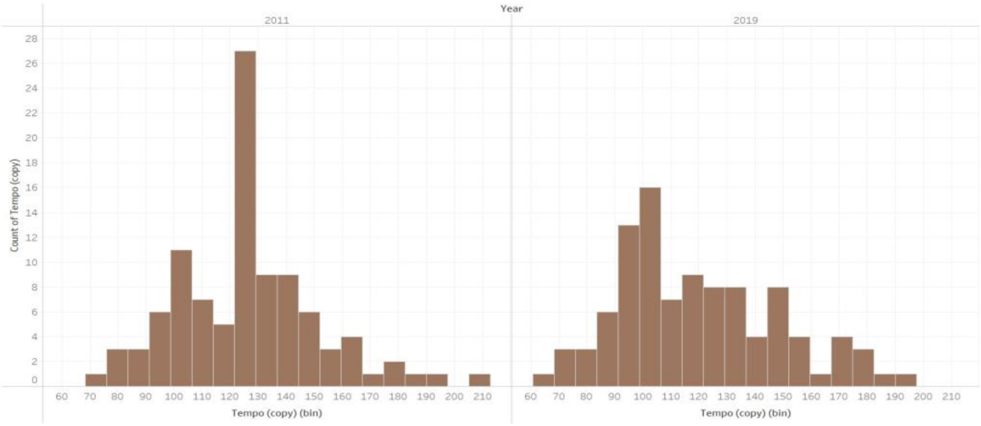


Figure 3.2: Tempo for 2011 VS 2019

The trend repeats itself onto tempo, the average speed of the track, whereby 2019 sees a flatter curve with less skewness intensity. Overall, these different observations of three distinct metrics seems to suggest that in later years, there are a greater variety of songs that have made it in Spotify's top tracks.

4. Regression Analysis (Jason)

There are many parameters that potentially contribute to the popularity of a song. However, not all parameters are strongly correlated with popularity, i.e. not all of them exert a strong influence on whether a song will become a hit among Spotify listeners, as alluded to in the previous part. The aim of this section is to eliminate any inactive parameters, in order to fit different weighting values using only the most significant parameters for the best regression model.

To provide a general idea of the overall trend, the datasets from 2011-2019 are combined as one, and a correlation coefficient heatmap is plotted as shown in Figure 4. Note that the parameters used are identical to that of the section prior. The Pearson correlation coefficient between an explanatory variable x and the response variable y is defined as

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right), \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are the means, while s_x and s_y are the sample standard deviations [4]. In our research, x corresponds to all the parameters in the dataset, while y corresponds to popularity. The closer $|r_{xy}|$ is to 1, the higher the correlation between x and y. Likewise, variables are more closely-related for darker colors in the heatmap (excluding the diagonal elements), where red and blue imply negative and positive correlations, respectively.

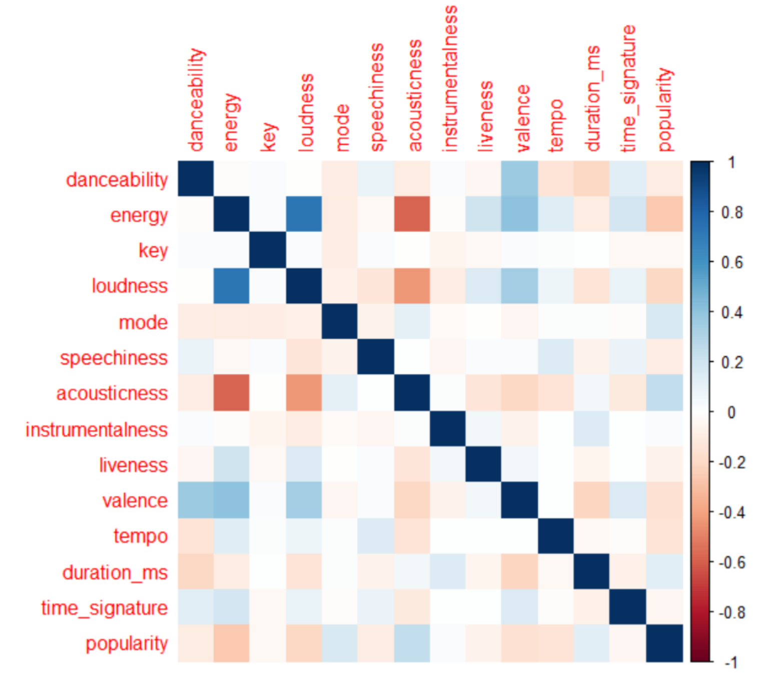


Figure 4: Correlation coefficient heatmap of popularity vs other variables

By observing the last row of the heatmap, there are a few parameters that show varying degrees of correlation with popularity. However, from the rows above, there are also clear correlations between some of the parameters themselves - for example, intuitively, energy has a positive correlation with valence (higher value for happier songs), as shown in figure 5. These collinearities between explanatory variables may result in redundancy in the final regression model, and may even make the weighting variance blow up. It is important for them to be removed.

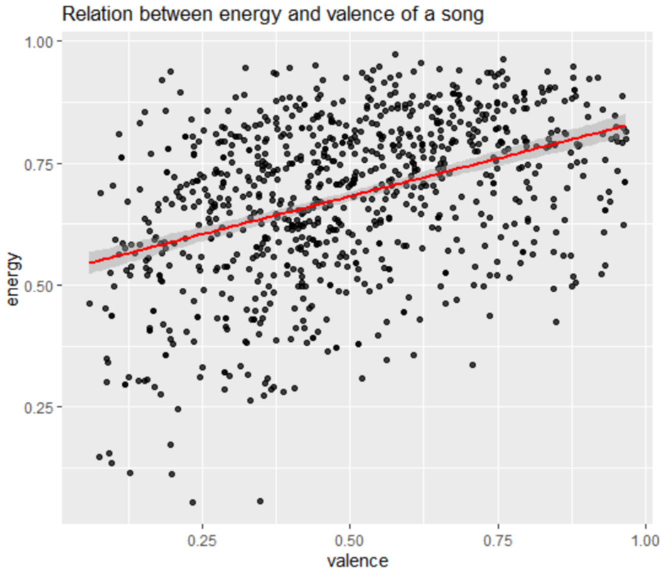


Figure 5: A positive correlation between the energy and the valence of a song

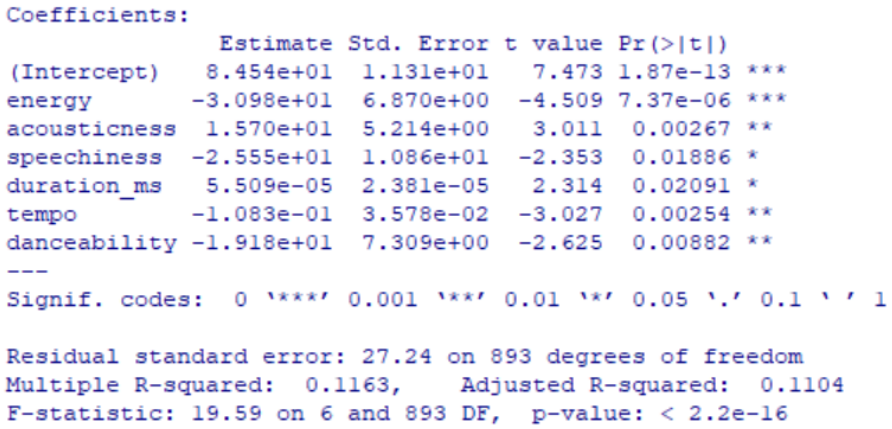
In statistics, there are many different ways to extract the simplest regression model such that the least amount of parameters can be used to do a good fit of the data. The Akaike Information Criterion (AIC) is a popular choice [5]:

$$AIC = n \log \left(\frac{RSS}{n} \right) + 2p_c, \tag{2}$$

where n is the total number of data points, p_c is the number of explanatory variables and RSS is the residual sum of squares (i.e. related to variance of the error of the regression fit). The lower the AIC, usually the better the model. We start from the null model (intercept only) and add one parameter to the regression and determine the addition of which parameter provides the most decrease to the AIC value. Likewise, we proceed to add a second parameter to the model and again, use the one that decreases the AIC the most, and so on and so forth until the AIC no longer decreases. Using this forward selection model, we arrive at the following regression - Model 1:

$$\text{popularity} = \beta_0 + \beta_1 \times \text{energy} + \beta_2 \times \text{acousticness} + \beta_3 \times \text{speechiness} + \beta_4 \times \text{duration\_ms} + \beta_5 \times \text{tempo} + \beta_6 \times \text{danceability} + e, \tag{3}$$

where e is the error and β_i are the corresponding weightings for i = 1 - 6. The values of these weightings are calculated using ordinary least squares (OLS) fitting and are given in the "estimate" column in figure 6.



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.454e+01  1.131e+01   7.473 1.87e-13 ***
energy       -3.098e+01  6.870e+00  -4.509 7.37e-06 ***
acousticness  1.570e+01  5.214e+00   3.011  0.00267 ** 
speechiness  -2.555e+01  1.086e+01  -2.353  0.01886 *  
duration_ms   5.509e-05  2.381e-05   2.314  0.02091 *  
tempo        -1.083e-01  3.578e-02  -3.027  0.00254 ** 
danceability -1.918e+01  7.309e+00  -2.625  0.00882 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.24 on 893 degrees of freedom
Multiple R-squared:  0.1163,	Adjusted R-squared:  0.1104 
F-statistic: 19.59 on 6 and 893 DF,  p-value: < 2.2e-16
```

Figure 6: OLS estimates for the weightings of the regression model

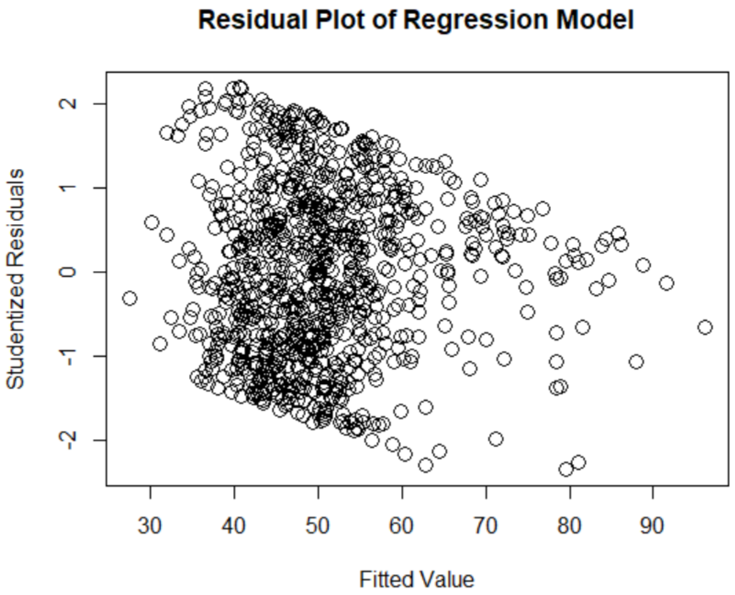


Figure 7: Residuals of the linear regression model

From figure 6, the rightmost column shows the p-value for the weighting of each parameter in Model 1. Each p-value provides a result for the following hypothesis test:

$$H_0\colon\ \beta_i = 0 \text{ in Model 1} \qquad H_1\colon\ \text{Model 1.} \qquad (4)$$

Since all the p-values are smaller than the common alpha level of 0.05, the 95% confidence interval does not contain $\beta_i = 0$ [6]. Hence, the inclusion of those parameters are significant and cannot be ignored. From an intuitive point of view, we can see that parameters like energy, tempo and danceability all have negative correlations with popularity, which imply that Spotify listeners seem to have a preference for slower, less upbeat songs. On the other hand, lower speechiness with higher acousticness seem to increase the popularity in general, putting emphasis on the music part of the song other than the vocal portions. Hence, it appears that more melodic songs that are rich in acoustic instruments are favored among the audience. The duration of the song also has a tiny positive correlation, where longer songs are slightly preferred, but the effect is small.

Since the model is a multi-variable regression, it is not ideal to plot several 2D graphs with each showing the popularity against one of the parameters. To show the combined effects of all the parameters, figure 7 shows the studentized residuals t_i (dividing residual by the standard deviation estimate) against the fitted values. Outliers are denoted by $|t_i| > 2$, and from the graph, most points fit within [-2, 2]. It provides a simple picture to show Model 1 as a good regression fit.

Refer to Appendix 1 for the code used in this section.

5. Clustering (Felix)

Based on the results of the regression analysis, it is evident that some factors are more correlated to top songs, which are energy, speechiness and danceability. In this section, these parameters will be used to do clustering.

K-means clustering enables one to consider all the factors in a single 2D graph when doing comparison, which are suitable for our purpose. The aim of this section is to predict the possible extent of popularity of the song.

The general idea of the prediction is to use the data of energy, speechiness and danceability of song to create a database. Afterwards the most suitable K value by the Elbow Method is calculated. Once the clustering result is found, it will be ranked according to the number of the song in each cluster. Finally, compute the distance between the new object and the center of each cluster by the data of a new song and hence decide the class of the new object.

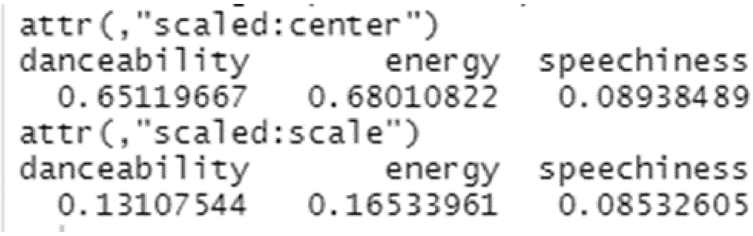
Firstly, the clustering model built by the database of the top 100 songs from 2011-2019. We select the energy, speechiness and danceability of 900 songs. To prevent attributes with large ranges outweighing ones with small ranges and the function of machine learning algorithms work well, Standard Scaling will be used to normalize data. The function name is “scale()”. The principle of Standard scaling is to change the mean value of all data to be zero.

By using equation (4):

$$x\_scaled = (x - \bar{x}) / s , \quad (4)$$

where x is the real x-value, x̄ is the sample mean and s is the sample SD.

Figure 8 shows the normalized data of the 900-song list.



```
attr(,"scaled:center")
danceability       energy  speechiness
  0.65119667   0.68010822   0.08938489
attr(,"scaled:scale")
danceability       energy  speechiness
  0.13107544   0.16533961   0.08532605
```

Figure 8: Normalized data of all 900 songs

After the data are normalized, the K value should be chosen. The function name is called “fviz_nbclust ()”. The actual method that we are using is called “Elbow method”. The principle of “Elbow method” is based on the sum of the squared errors (SSE) as shown in equation (5):

$$SSE = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{p \in C_i} \left| p - m_i \right|^2 \qquad (5)$$

where:

Ci: i-th cluster p: sample point in Ci

mi: centroid of Ci k: no. of cluster

When k increases, the sample division will be more refined. The aggregation degree of each cluster will gradually increase, so the SSE will gradually become smaller, as shown in figure 9.

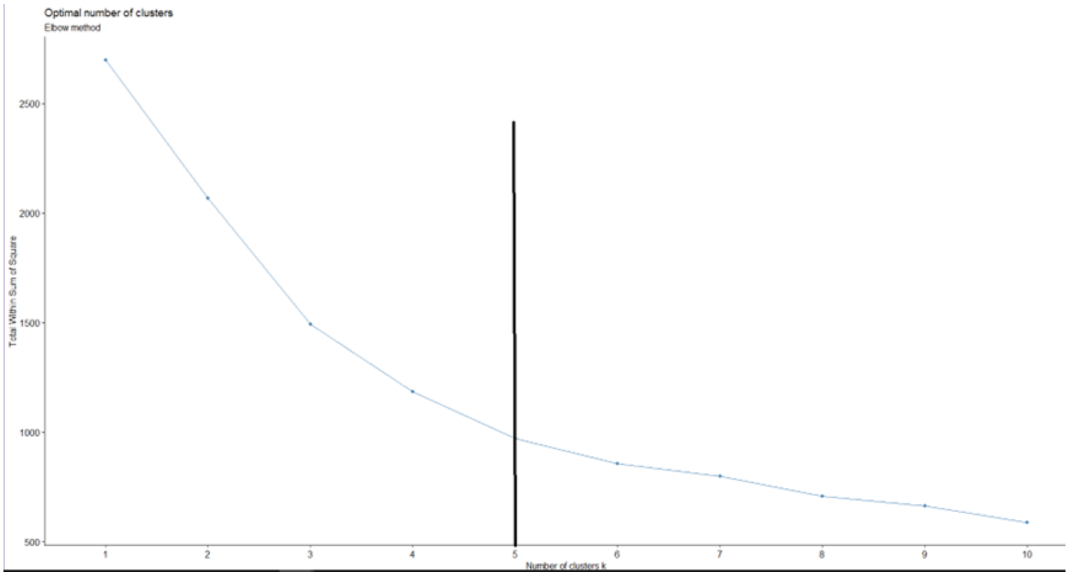


Figure 9: The Elbow Method

We choose K=5. Since the data can be clearly seen at K=5 and it has a great change of slope at K=5. After we choose the K value, we can use the data to form a clustering system. We use a function called “fviz_cluster ()”. It can generate the clustering system and plot a graph for visualization, as shown in figure 10.

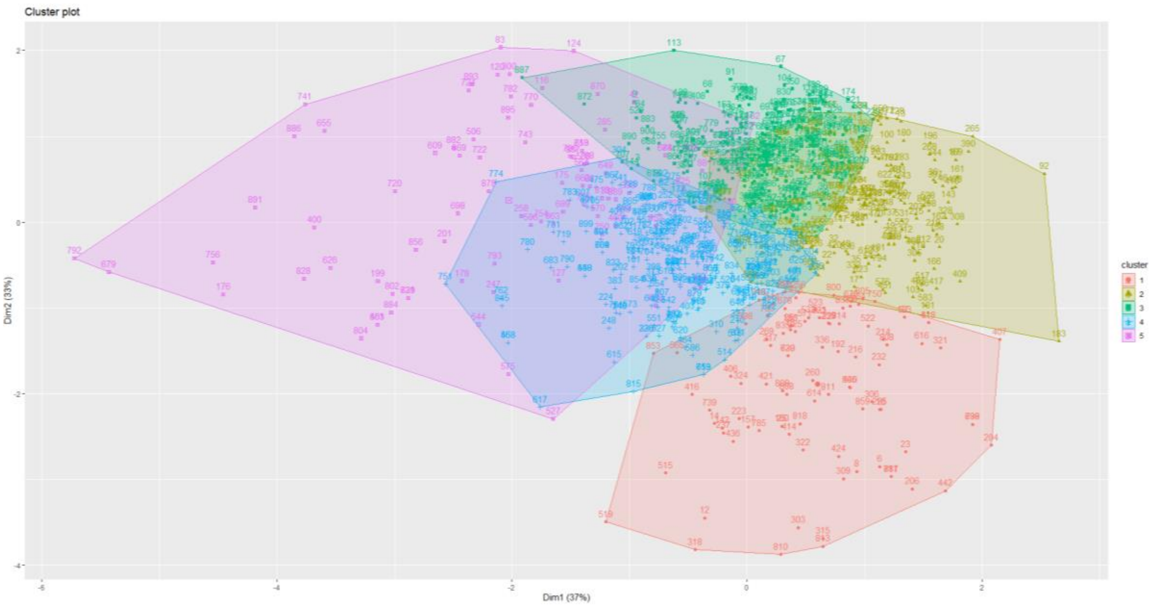
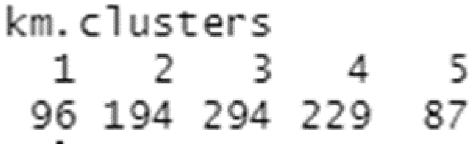


Figure 10: The 5-cluster graph



```
km.clusters
  1   2   3   4   5
 96 194 294 229  87
```

Figure 11: The summation of the song of each cluster

Since the song here is the top 100 most popular each year, we can assume that the five clusters all represent popular songs. The difference of each cluster is the relation of three parameters. There are a total of 900 songs in 5 clusters. The sum of each cluster is shown in figure 11. We will rank the clusters as 3>4>2>1>5.

The system of clustering is ready, and we can classify new objects now. The procedure is nearly the same as the aforementioned. Assume we choose a song from Kaggle, which is not on our 900-song list.

Data of song: Danceability: 0.78 Energy: 0.33 Speechiness: 0.06

By using equation (4):

Attribute: Danceability: 0.983 Energy: -2.119 speechiness: -0.345

Then, assume K=5 should be used.

According to the result of the program, the new song is in cluster 4, which is the second rank on the list. Based on the result of different rankings in each cluster, a higher popularity song has a higher probability to gather in a higher cluster rank. The classification of the clustering model is pretty accurate.

Refer to Appendix 2 for the code used in this section.

6. Conclusion (Jason)

In conclusion, this research has focused on the effects of different numerical parameters on the overall popularity of a song on the Spotify platform, based on numbers provided directly from Spotify. Through visualization in section 3, it is found that the distributions of some parameters, i.e. people's music tastes, have shifted gradually over the years, but mostly retaining the same overall preference. For instance, people generally prefer songs with lower acousticness, with 60% of songs lying in the range of 0-0.12, though that preference has slightly evened out from 2010-2019.

The results from visualization is a neat segue into the regression analyses, which indicate similar correlations between popularity and each of the parameters, e.g. popularity having a negative correlation with acousticness. Using criterion models and hypothesis testings, it is found that the most significant parameters are energy, acousticness, speechiness and danceability, while the duration and tempo are also useful parameters to build the best regression model.

(Felix)One of the purposes of the project is to classify the different songs into subgenres by K-means clustering. The clustering system mainly focuses on three parameters to do grouping. The 900 songs are clustered into 5 groups, which are ranked according to the number of songs. It is shown that a higher ranking cluster represents a greater probability of the song having a higher popularity.

References

[1] Philiptrapp. (2022, February). Why streaming on Spotify actually declined during the pandemic. Loudwire. Retrieved April 5, 2022, from *https://loudwire.com/spotify-streaming-declined-during-pandemic-study/*

[2] BBC News. (2011, March). Spotify hits milestone with 1 million subscribers.Retrieved April 3, 2022, from *https://www.bbc.com/news/business-12676327.*

[3] Delaney. (2021, April). Spotify Top 100 Tracks (2001-2019), Version 1. Retrieved April 3, 2022 from *https://www.kaggle.com/datasets/delaneyisabella/spotify-top-100-tracks-20012019.*

[4] Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product–moment correlation coefficient. *Animal behavior*, *93*, 183-189.

[5] Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853.

[6] Long, J. (2022, April 25). Tools for summarizing and visualizing Regression Models. Retrieved May 10, 2022, from *https://cran.r-project.org/web/packages/jtools/vignettes/summ.html*

[7] Mukherjee, S. (2021, February). Spotify outlook weakens as pandemic uncertainty persists. Reuters. Retrieved April 5, 2022, from *https://www.reuters.com/article/us-spotify-tech-results-idUSKBN2A31JB*

Appendix

Appendix 1 (Jason)

```
library(corrplot)
M<-cor(data0[5:18])
corrplot(M, method ="color")

library("ggplot2")
p <- ggplot(data0, aes(x = valence, y = energy)) + geom_point(alpha = 0.75) + geom_smooth(method="lm", color = 'red')
p + ggtitle("Relation between energy and valence of a song")

fit0<-lm(popularity~1, data=data0)
fit1<-lm(popularity~danceability+energy+key+loudness+speechiness+acousticness+instrumentalness+liveness+valence+tempo+duration_ms,data=data0)

library(MASS)
stepAIC(fit0,scope=list(lower=fit0, upper=fit1),direction="forward",trace=1)
stepAIC(fit1,scope=list(lower=fit0, upper=fit1),direction="backward",trace=1)
```

```
fit<-lm(popularity~energy+acousticness+speechiness+duration_ms+tempo+danceability,data=data0)
summary(fit)

plot(fit$fitted,rstudent(fit),main="Residual Plot of Regression Model",xlab="Fitted Value",ylab="Studentized Residuals",cex=1.5)
```

Appendix 2 (Felix)

```
> library(readxl)
> dance_energy_speech <- read_excel("C:/Users/p/Desktop/dance_energy_speech.xlsx",
+     range = "A1:E902")

> library(factoextra)

> dance_energy_speech_data <- dance_energy_speech[2:4]
> dance_energy_speech_scale <- scale(dance_energy_speech_data)
> dance_energy_speech_data <- dist(dance_energy_speech_scale)

> fviz_nbclust(dance_energy_speech_scale, kmeans, method = "wss") + labs(subtitle = "Elbow method")

> km.out <- kmeans(dance_energy_speech_scale, centers = 5, nstart = 100)
> km.clusters <- km.out$cluster

> paste(dance_energy_speech$popularity, 1:dim(dance_energy_speech)[1], sep="_")
> fviz_cluster(list(data=dance_energy_speech_scale, cluster = km.clusters))
> km.out$cluster
```