

Spotify Music Popularity Analysis

Jillian Hay, 20200484, 19jfh@queensu.ca

Ganfu Fan, 20163296, 18gf7@queensu.ca

Liguang(Lee) Zhang, 101798204, 14lz22@queensu.ca

Yuren Xia, 20148705, 18yx61@queensu.ca

Apr. 24, 2022

Brian Ling

List of Contributions

Jillian Hay - presentation - results, report - results, formatting into APA

Ganfu Fan - coding & graphing in R, the method in presentation and report, graphs in result.

Liguang(Lee) Zhang - Presentation& Report: introduction, 1/3 of discussion & conclusion, references, presentation template and slides aesthetics & formatting.

Yuren Xia - Dataset descriptions, 1/3 conclusion in the presentation and report.

1 Introduction

The rise of various music streaming services have provided easy access to high-quality music using subscription-based models such as Spotify and Apple music. In the case of Spotify, a \$9.99 monthly fee allows its users to download high-quality music, listen to any music advertisement-free, and be able to stream music on the Spotify mobile application. These music streaming services have gained widespread success—as of July 2019, Spotify had 232 million monthly active users, with 108 million paying a subscription fee.

In recent years, Spotify has made extensive use of recommender systems to assemble and recommend playlists to its subscribers. Its motivation behind this is to keep its subscribers entertained with its never-ending stream of fresh and chart-topping music in order to keep them from switching to another music streaming platform.

There also exists a loudness war in the music industry—loudness war is a trend of increasing audio levels in recorded music.

Our project is centred around discovering the key song features that are associated with popularity. We will use the Spotify dataset found on Kaggle to perform data analysis. This dataset was originally obtained by accessing the Spotify Web API and consists of 586673 tracks' features such as popularity and tempo. We believe that certain features of songs, for example, high tempo and volume are positively correlated with popularity. And that by having a deeper understanding into what makes a song chart-topping, music streaming services like Spotify could provide better recommendations—recommending songs that have more features positively correlated with popularity. Artists could also use this gained understanding to produce songs more aligned with the “definition” of popular music. This would in turn result in more trendy music for the public to consume and result in a more vibrant music industry. In addition, if high volume is proven to be positively correlated with popularity, it would justify the loudness war in the industry.

Our analysis could also reveal the changing trend in the music industry, for example, the changes in average loudness from 1922 to 2021 (as our dataset contains records in that timespan), or the changes in average song duration from 1922 to 2021.

Al-Beitawi, Salehan & Zhang (2020) performed cluster analysis on Spotify dataset containing the top 100 songs of 2017 and 2018 using 9 features including danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. The dataset was compiled by Spotify. The authors used k-means clustering for the cluster analysis and confirmed the resulting clusters using agglomerative clustering. The biggest cluster out of the four clusters for the year 2017 contained features including high danceability, high loudness, low instrumentalness, high valence, and low tempo. They found clusters that vary greatly in size and contain varying features in each cluster.

The most popular songs were found to be the more exciting and radio friendly songs which generally have a formulaic, pop-friendly, danceable nature that put the listeners in a good mood afterwards.

2 Description of the dataset

- Describe each variable in the dataset. Are they categorical? Are they numeric? Are they binary?

#paraphrase this

<https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features>

#use a table to summarize could be a good idea

#binary is categorical

#keys and time_signature also are categorical

- May include data visualization and summary statistics here.

#use the summary function for the table

- Always label the figures and tables properly.

3 Methods

3.1 overview of methods

The original dataset contains 20 variables. It includes unique variables, numeric variables, and categorical variables. Our goal is to find out what influences music popularity. We use the following method and progress to obtain our result. First, the unique variables and abnormal data were removed from the dataset for cleaning. Second, linear regression was applied to the numeric variables. Third, ANOVA analysis was applied to the categorical variable. Forth, a stepwise regression was applied to find the best model. Lastly, we did quality control on the models by drawing the residuals and calculating MSE.

3.2 clean the data

The data contains unique values (eg: id, artist), numeric variables (eg: danceability), and categorical variables (eg: mode). So, the first thing is converting these variables into the correct type. For example, the “key” is converted from numeric to categorical. Then, we removed the abnormal data. The data is removed if they have popularity equal to 0, the tempo equal to 0, or the release date is “1900-01-01”.

3.3 Regressions of numeric variables

There are 12 numeric variables in total. Popularity is the response variable and the others are predictor variables. In order to understand how one numeric variable influences popularity, we used linear regression. For example, our hypothesis is there is a relationship between loudness and popularity. The null hypothesis is there is no relationship between them. The coefficient shows the direction of the relationship. If the coefficient equals 0, then there is no relationship between them.

The accuracy of this model is measured in 3 ways. The first method is using the *rcompanion* package to calculate the MSE and the Max-Min accuracy of the model directly.

The second method is creating a training set that contains 10% data from the entire dataset. Then, apply the training model to the testing model which contains the rest of 90% of the data. Last, the accuracy of the model equals the correlation between the observed value and the predicted value in the testing set.

The third method is logistic regression. We define a binary variable to measure popularity. If the popularity of the pieces is greater than the average (29.84), it is popular (denoted as 1). Otherwise is not popular (denoted as 0). Then, using a glm regression between the predicted variable and binary popularity in the training set provides a training model. Then, apply the training model to the testing dataset to get the predicted results. Last, the accuracy can be found by comparing the table of observed results and predicted results.

3.4 ANOVA

Not every variable is numeric in this project. To find out if 4 categorical variables influence popularity we did an ANOVA test. The hypothesis of the ANOVA test is there is a difference in the mean among the different groups. For example, we found there is a significant difference in the mean between major mode and minor mode by ANOVA test which means the mode can affect the popularity.

3.5 Regressions of categorical variables

We used linear regression to understand how categorical variables influence popularity. First, the categorical variables are converted to dummy code by using `as.factor` function. Then, we used linear regression to measure the direction of the relationship. For example, there is positive relationship between positive mode and popularity. In contrast, there is negative relationship between minor mode and popularity.

Similar to the linear regressions of numeric variables accuracy, the accuracy of categorical variables models also calculated in two ways. First is done by *rcompanion* package. The second is done by logistic regression as same as numeric variables `glm`.

3.6 Stepwise Regression

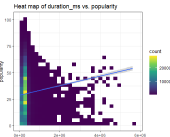
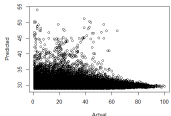
The stepwise regression allows us to find the best model in the big dataset. Comparing the AIC value among the possible models, it can reduce the complexity of the model without reducing accuracy too much. Since the release year and categorical variable (some categorical variables may consider as random variables) may cause unnecessary bias, we tried the stepwise regression 4 times for the following group separately:

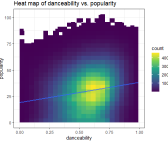
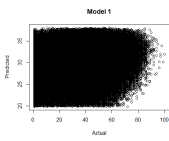
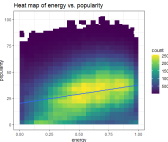
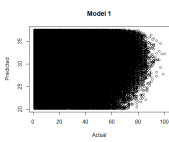
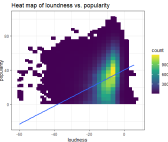
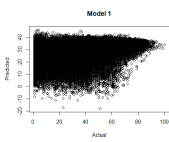
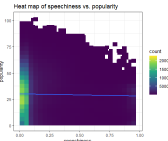
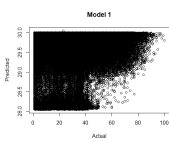
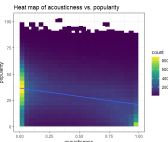
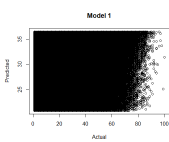
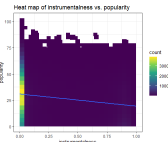
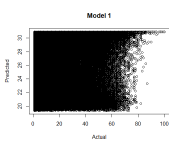
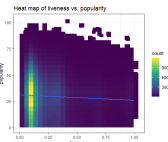
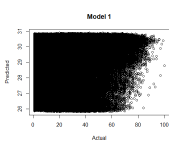
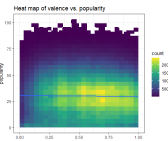
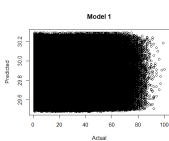
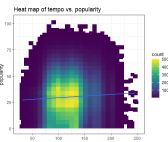
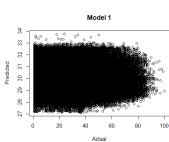
1. The full dataset contains release years and categorical variables (dummy code)
2. The dataset does not include release years but includes categorical variables (dummy code)
3. The dataset contains release years but does not include categorical variables
4. The dataset that dose does not contain release year and categorical variables

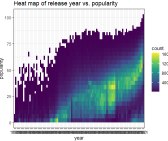
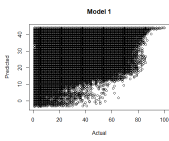
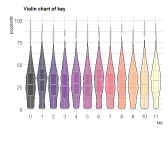
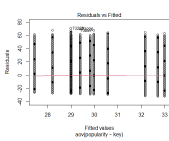
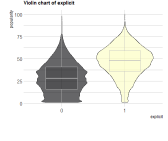
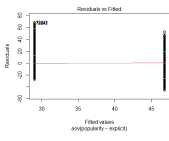
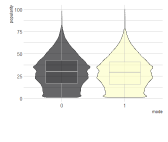
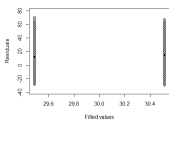
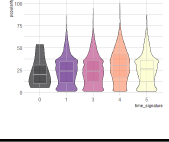
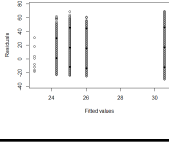
As same categorical variables regression, the accuracy is measured by the *rcompanion* package and the logistic regression with a binary response variable.

4 Results

The results of Spotify's music's popularity analysis are displayed in the table below. The table is complete with title, method, graph, accuracy and residual or quality control. These results consist of heat maps, ANOVA tests and residuals.

Title	Method	Graph	Accuracy (Min.max.accuracy, MSE)	Residual/quality control
Duration_ms vs. Popularity	Linear regression (LM)		0.62 ,297	

Danceability vs. Popularity	LM		0.624,287	
Energy vs. Popularity	LM		0.628,279	
Loudness vs. Popularity	LM		0.631,270	
Speechiness vs. Popularity	LM		0.62,297	
Acousticness vs. Popularity	LM		0.633,269	
Instrumentalness vs. Popularity	LM		0.622,289	
Liveness vs. Popularity	LM		0.62,296	
Valence vs. Popularity	LM		0.62,297	
Tempo vs. Popularity	LM		0.62,297	

Release Year vs. Popularity	LM		0.668, 207	
Violin Chart of Key	ANOVA		NA	
Violin Chart of Explicit	ANOVA		NA	
Violin Chart of Mode	ANOVA		NA	
Violin Chart of Time Signature	ANOVA		NA	

Each heat map is reflective of a specific variable evaluated in our report compared to popularity. This is exemplified by the heat map titled Tempo vs. Popularity which assesses the track's tempo relative to its popularity on Spotify. This heat map demonstrates that the majority of tracks have a tempo of 100 with popularity at 30. The heat map also establishes that popularity in terms of tempo is very dispersed and any tempo can range in popularity. Moreover, the regression line in this heat map which fits the data is almost straight across the heat map. This is indicative of the heat map's generally dispersed display.

There are also several Analysis of Variance (ANOVA) tests represented in the table which calculate whether a relationship exists between more than two groups. The violin charts in the table also depict the density of each variable and summary statistics. This is illustrated with the ANOVA test titled Violin Chart of Time Signature which displays the analysis of variance related to popularity and time signature. This violin chart represents time signatures from 1 to 5 and popularity ranging from 0 to 100. As seen at time signature 4, the median popularity is the highest of the time signatures at around 31.

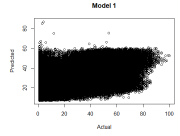
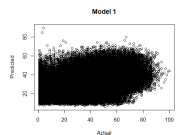
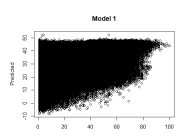
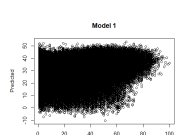
The results from the heat maps, ANOVA tests and residuals all lead to implications and insights that can be made for Spotify. As aforementioned, the heat map relating to Tempo vs. Popularity revealed that any tempo can range in popularity. This gives insight to artists as well as employees of Spotify that tempo does not weigh too heavily on a track's popularity.

Implications can be more helpful in other heat maps such as Loudness vs. Popularity and Release Year vs. Popularity. The Loudness vs. Popularity heat map detected that the optimal

loudness to encourage popularity in a track was located at around -5. This is useful information to Spotify and its artists as these statistics reveal the loudness level that maximizes the popularity of a track. The Release Year vs. Popularity heat map detected a pattern of popularity emerging in the 1930s and escalating into the early 2000s. From this information, it can be inferred that Spotify's demographics are more focused on younger users. Furthermore, if hypothetically a Spotify employee thought to invest money into adding older tracks to Spotify this may differ that employee from this decision as they would most likely not perform well.

Additionally, the ANOVA tests present in the table can give insights as well. The Violin Chart of Time Signature concluded that time signature 4 had the highest median popularity. This could lead artists and Spotify executives to aim for songs in the 4th time signature used in the violin chart.

To conclude, these results can be further evaluated and each graph has beneficial results which can lead to more insight and knowledge on Spotify's music popularity analysis. The graphs represented in the table all provide practical insight and implications surrounding track popularity within Spotify.

Model	AIC	Accuracy (Min.max.accuracy, MSE)	Residual/ Quality Control
With release year and categoricla variables	4406246	0.672 200	
Without Year	4512243	0.645 239	
Without Categorical variables	4408409	0.670 202	
Without release year and categorical variables	4525159	0.642 247	

5 Discussion & Conclusion

Conclusion. What are the limitations of your study? Any possible improvements? Any problems discovered?

I separate the discussion and conclusion because we might have a lot of things that can discuss.

#a possible structure for this part: describe the problem we found+discuss the possible solution (eg: problem: low model fitness, solution: add more variables/only select short period of time) A problem may have multiple solutions

#limitation: no genres, no artists, limited sample (not every region use Spotify)

#issue: low model fitness, the stepwise model unable to include all variables, bias due to the sample only from one platform

Problems discovered:

Our stepwise regression models suffered from large AIC values, and included all the variables which resulted in a bulky model. In addition, we think there exists data-based multicollinearity in our model. Multicollinearity exists where two or more of the predictor variables in a regression model are moderately or highly correlated. The consequence of having multicollinearity in a regression model is that the estimated regression coefficient of a variable depends on which other variables are present in the model. In addition, when the predictor variables are correlated, the precision of the coefficients decrease as more predictor variables are included in the model. Lastly, our best performing model only had an accuracy of 0.672. In the future, using more advanced machine learning approaches such as deep-learning-based or random-forest models might help improve the accuracy of the model.

6 Reference

Al-Beitawi, Z., Salehan, M., & Zhang, S. (2020). What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs. *Journal of Marketing Development and Competitiveness*, 14(3), 79-91.

Ay, Y. (2021). *Spotify Dataset 1921-2020, 600k+ Tracks*[Dataset]. Kaggle.
<https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks?select=tracks.csv>