Music Popularity Analysis on Spotify Top Songs

Qingyin Ge qg2177

April 2020

1 Background

Nowadays, entertainment has been playing a more and more important role in people's daily life. People watch Youtube, listen to musics and play computer games to relieve anxiety and enjoy spare time, especially during this COVID-19 pandemic. Music, specifically, could accompany human beings and bring comfort, when they feel panic, lonely, helpless, distressed. We need music now. Then it comes to a question, what music would people like to listen? Many people listen to the songs which are popular and listened most by other people. With time going by, "popular music" has changed a lot in our mind, among different genres, by different year-released, and with different singers. What's the trend of popular music? Who sings most popular songs? Are there any popularity difference among different genres? What will be the main effect of being a top song? How popular would songs be in the future? In this project, we are aiming at solving these questions, and help people better get the vibe of popular musics.

2 Introduction

The data set comes from Kaggle, describing top songs by year in the world, through Spotify music platform. Nine numeric variables, including **BPM** which is the tempo of the music; **Energy** which describes the power and cheer level of the music; **Danceability** which indicates that whether the song is suitable for dance or not; **Loudness** which describes the decibels of the music; **liveliness** which represents the possibility of live recording; **Valence** which shows the positive mood of the song; **Duration** which is the length of the songs; **Acousticness** which represents the acoustic level of the song; **Speechiness** which shows the amount of spoken words the song contains. These variables are extracted from Spotify. Most importantly, the research interest is in music **Popularity**,

and their **Title**, **Artist**, **Year** and **Genres**, which come from Billboard. Totally I have collected 2476 observations in the data analysis with 14 variables. The main findings are based on group-mean comparisons methods, time series analysis, and multivariate regression models. It turns out that the mean popularity of songs among different genres are different, pop, rock, metal, soul, and adult standards are the top 5 genres, which have higher mean popularity than other genres; moreover, we could predict that the average popularity of top songs in 2020 should be 72.85 based on arima model, and it should be 73.76 based on sample acf/pacf plot. The final multivariate linear regression model selected by stepwise regression regresses popularity on Year which is a two-factor variable containing value centry20 and centry21, and Energy, Danceability, Loudness, liveliness, Acousticness, Speechiness, total 7 predictor variables. However, the result seems to be not satisfying, with adjusted multiple R-squared 11%. After fitting polynomial regression model the result is not improved significantly. So we may claim that the mean popularity is not quite associated with the numeric features. By loess method and kernel smoothing to fit non-parametric model, we figure out that the local regressions are almost horizontal with various value of predictors, therefore this validate our claim. Details are described below.

3 Data Overview

First let's glance over the data. Totally we have 9 numeric features describing different features of songs. We would use them to investigate whether there is strong association between each feature and the popularity of music, by regression analysis. Therefore we should first check each numeric features, including their distribution and correlation. Figure 1 shows that feature BPM, Danceability are almost normally distributed, while there are more music have higher energy, and are louder; more music are less likely to be recording lively, have lower acoustic level and have less spoken words. Valence of all music are similar instead. Their correlation are not obvious.

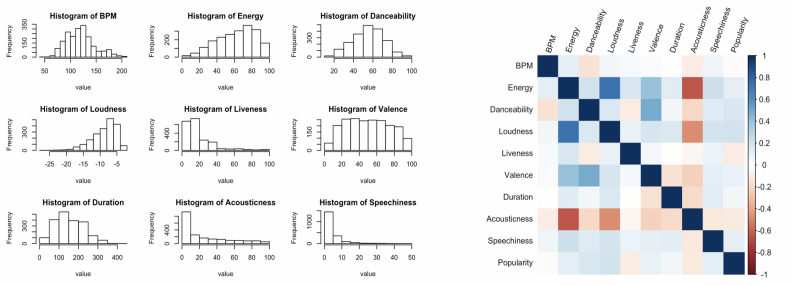

Figure 1: Numeric features value distribution histogram plot and their correlation plot

The box-plot for Popularity v.s. Year and Popularity v.s. Genre, both show that the popularity are normally distributed, but with different standard errors among each genre and each year. This is probably due to the unbalanced sample we get, we don't have similar amount of songs belong to each genre and are released in each year. Let's have a closer look later.

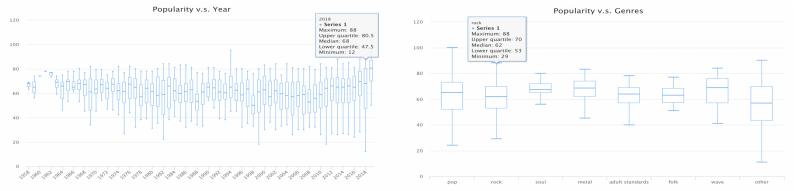

Figure 2: Popularity box-plot among each year and each genres

Second we step forward and dig into the data. By visualizing top 20 songs of each genre and each year, everyone can tell the most popular songs during recent 60 years, and people can also find popular songs in their favorite type of songs; moreover, I also filter out the best 20 songs and best 20 artist across different genres and eras. You can find the information below.

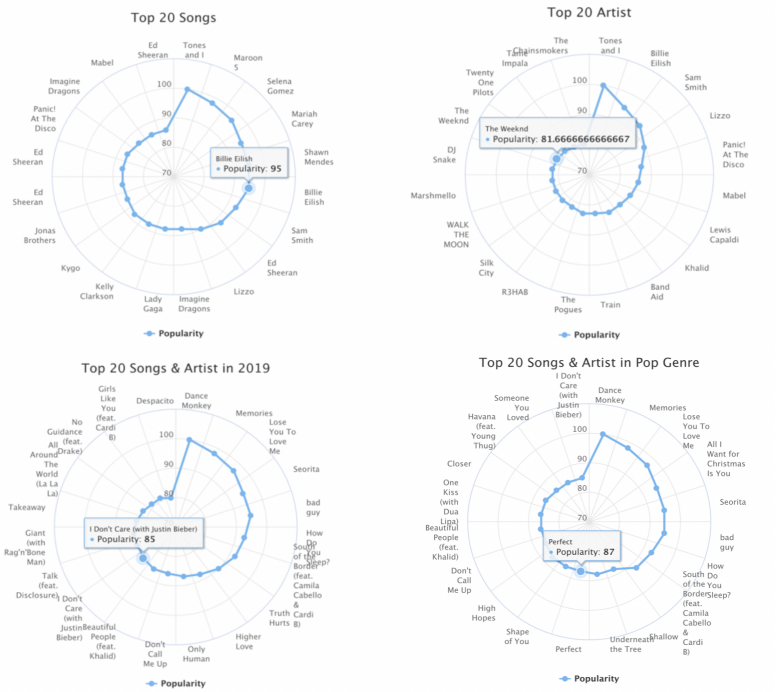

Figure 3: Polar representation of popular songs and artists

Finally we can glance at how does music popularity evolve during each year among each genre. No wonder that popularity in the 10s takes more part comparing with other eras and also, pop and rock are the most popular types across 60 years.

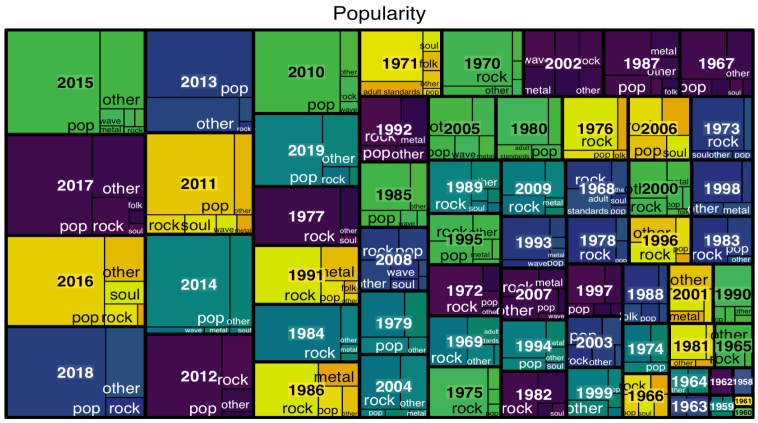

Figure 4: Treemap of popularity evolution

4 Data Analysis

4.1 Whether there exists difference among each genre?

In this section, we are focusing on group mean testing. We assume:

- The population have normal distributions
- The population standard deviations are all the same
- Observations within each sample are independent of each other
- Observations in each sample are independent of observations in others

Why can we assume those things? First, from the **box-plot** we can see that data are almost normally distributed; while by **Shapiro test** we end up with p-value $< 2.2e-16$ which suggests not normal. Why this is the case? By checking the **skewness** and **normal QQ plot**, we find out that the skewness $= -0.667$ and plot shows that it is strongly left skewed. Since data are from top songs so corresponding popularity is certainly high. We can believe that in population the popularity distribution is normal. The independence can be considered by

each song has nothing to do with other songs. Therefore by assuming these we are able to use t-test and F-test.

Genres	Number	Average	SD
Pop	667	62.09	15.59
Rock	674	60.70	12.71
Soul	32	67.28	7.20
Metal	72	66.79	10.09
Adult Standards	87	62.75	9.50
Folk	15	63.87	7.99
Wave	30	66.33	11.66
Other	423	56.36	16.84

Null hypothesis is *There is no difference in mean popularity between each genre, i.e.* $\mu_1 = ... = \mu_8 = \mu_0$ and the alternative is *At least one group is different, i.e. at least one* μ_i *is not equal.* By **ANOVA** we find that the p-value is $3.4e-11$ which is way smaller than 0.01, therefore we conclude that the mean of popularity is different between each genre. Moreover, we know the highest average popularity belongs to soul genre and lowest average popularity belongs to other genre. So we want to construct **Bonferroni confidence interval** for mean difference of these two groups. The result shows that the 95% simultaneous confidence interval is (3.7, 18.1), which doesn't include 0 inside. So we may conclude that the mean popularity of the 8 different genres are different. Moreover, we have 95% confidence to say that the mean popularity of soul genre is between 3.7 and 18.1 higher than other genre. Now the next step is to know which ones are different among these group means.

By **paired t-test with holm's adjustment**, which sequentially compares the lowest p-value with a Type I error rate that is reduced for each consecutive test, it is being shown that mean popularity of other genre is statistically different from that of pop, rock, soul, metal, adult standards, with p-value $5.0e-09, 3.3e-05, 8.9e-04, 3.9e-07, 4.0e-03$, respectively; in addition, there is also some evidence suggest that mean popularity of metal is different from rock genre, with p-value 0.014. By **Tukey's HSD Procedure**, we compare 28 pairwise mean popularity difference, seeing that mean popularity for other genre is 5.73 lower than that of pop genre, with 95 % confidence interval (3.02, 8.44); mean popularity for other genre is 4.33 lower than that of rock genre, with 95 % confidence interval (1.63, 7.04); mean popularity for other genre is 10.92 lower than that of soul genre, with 95 % confidence interval (2.92, 18.92); mean popularity for other genre is 10.43 lower than that of metal genre, with 95 % confidence interval (4.87, 15.99); mean popularity for other genre is 6.39 lower than that of adult standards genre, with 95 % confidence interval (1.25, 11.52).

4.2 How does popularity change with time going by?

In this section, we try to investigate the trend of songs' popularity since 1956. Can we know the approximate popularity of songs in 2020 and even later years? During which period music are more popular?

By fitting **auto.arima** function we can figure out the possible order for ARIMA time series, AR, Integrated, MA order are 1, 1, 2, respectively. In contrast to auto fitting, following by the series is almost stable via **Kwiatkowski-Phillips-Schmidt-Shin(KPSS) test** with p-value = 0.018, we plot the **sample ACF and PACF** as reference, finding that the mean popularity follows ARIMA(1,0,6) process.

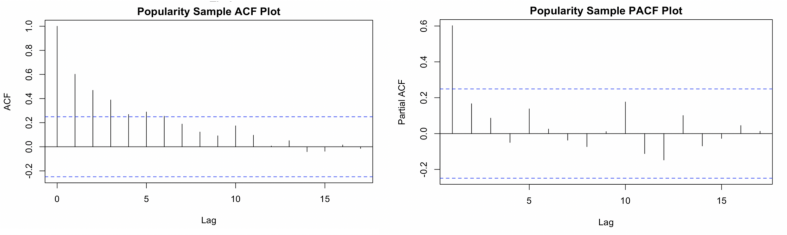

Figure 5: Sample ACF and PACF plot

Now we have two time series model, first one is **ARIMA(1,1,2)** by r function, another is **ARIMA(1,0,6)** by sample plot. We can see the predict popularity for the future 5 years in the figure below. We expect to have on average 72.85 popularity in year 2020 by ARIMA(1,1,2) and 73.76 popularity by ARIMA(1,0,6).

Figure 6: Future Prediction for popularity by time series model

4.3 Are the numeric features affect popularity of songs?

Now let's move on toward numeric features. Remember that we have 9 numeric features and 1 categorical variable Year. Here for the sake of convenience, I lump all the years to **"Centry20"** and **"Centry21"** where "Centry20" contains year

1956 to 1999, and "Centry21" contains year 2000 to 2019. After fitting null model which only contains intercept term, and full model which contains all the predictors, we do a model selection, which provides the result with 6 numeric features and the categorical variable "Year".

We could also do a model diagnostic aiming at detecting potential outliers and leverage points. By using **Cook's Distance** and **studentized residual plot** we find that case 1282, case 1099, and case 1548 might be high leverage points; case 1548, case 933, case 1099, case 346 are outliers, especially the first two. After deleting those influential points, we end up with our final model shows below:

Popularity = 77.42 - 5.60Centry21 - 0.13Energy + 0.14Danceability + 1.49Loudness - 0.09liveliness - 0.03Acousticness + 0.26Speechiness

with adjusted R-squared 0.11.

In order to consider comprehensively, we fit a **polynomial regression** model additionally, hoping to get better result. However, the result still shows that we can only explain around 0.12 variation of our data, so we decide not to talk about it here.

What would be possible explanations for such an unsatisfying result? By fitting local non-parametric regression on each feature,i.e. using **Loess Method** and **Kernel Smoothing** we find that the lines are almost horizontal. One possible reason is that the music popularity are not depend on those numeric features. People wouldn't choose to like one song by its loudness or liveliness. The characteristic of music that can attract people might be the rhythm, lyric or the singers, different people have different opinions on what is a mellifluous song.

5 Limitation and Future Improvement

Basically there are several limitations that are required further improvement.

- The sample we get is very unbalanced. We need about similar amount of cases in each genre and year. Also not only the top songs, but mediocre songs are needed as well.
- We are lacking of useful features, for example quantitative music style or rhythm, lyrics and flows.
- Sample size is insufficient for plenty of features. More data are necessary.