Research on Top Spotify Songs in 2010-2019

Principal Investigators:

Luyang Xu (lxu256@wisc.edu), Yuanru Gao (ygao277@wisc.edu)

Group 9

Introduction

As the COVID-19 pandemic presents challenges unimaginable, we have stayed at home for a few months and many of us choose listening to music in our spare time. In this project, we will use data from the music application Spotify to study the characteristics of popular music, such as energy, loudness, valence and the relationship between them. At the same time, we try to explore the relationship between these characteristics and song popularity, which will help us discover music with popular potential and explore its commercial value, such as advertising.

In the first part of our project we explore the quantitative features of the top songs. Most of them have the theme of love. Dance pop takes a large proportion , and the top singers often have consistent high quality works in a long time. In the second part, we try to explore the qualitative features. The variables that have a significant impact on popularity are energy, loudness, length. Loudness has a positive impact, while energy and length have negative impact.

Analysis

Data

We are using the data of the top songs from 2010 to 2019 in the world by Spotify based on Billboard. This dataset has 602 observations and 14 variables about the songs, 12 of the variables describe the features of the top songs.

Part I

In part I, we will describe qualitative variables including song, genre and artist. By discovering the characteristics of popular songs over the years to make preparation for the modeling part. To better understand how the data distributes and getting a clear picture. We use matplotlib and seaborn to depict the data. We also use several tables to show the results clearly.

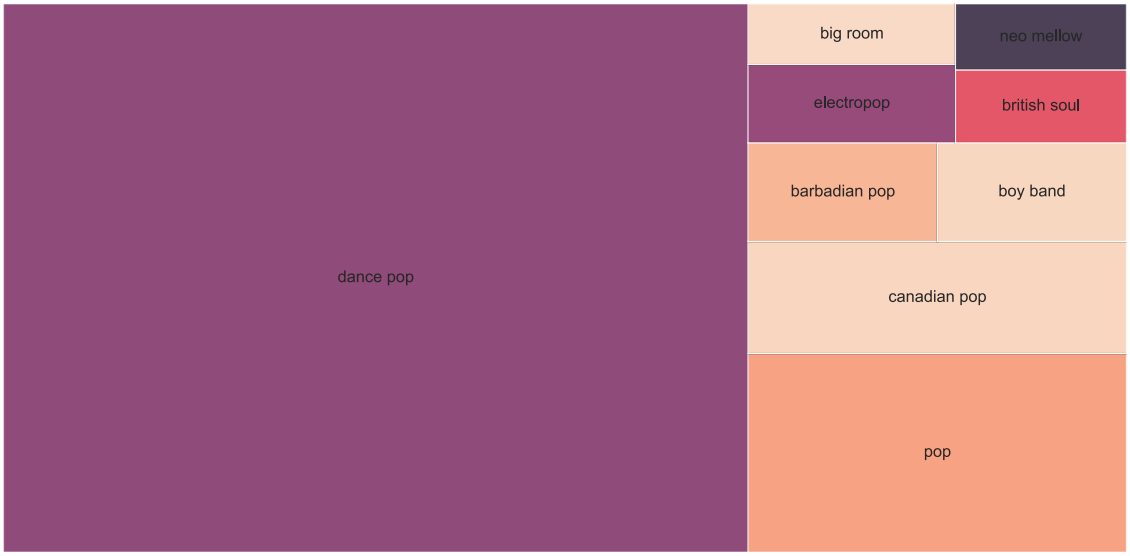
The first point of interest is the keyword that appear most frequently in top song titles. Intuitively we think it is “Love”. In order to check our thought is correct or not, we generate a word-cloud to describe the words in the titles, the larger the font, the more often it appears. We delete the words “feat” and “Remix” ahead. Because they can not reflect the artist's preference when naming the song. Figure 1 shows the result.

Figure 1: Words in the Titles of Top Spotify Songs in 2010-2019

Figure 1 proves our guess that the keyword appears most frequently in top song titles is “Love”. Besides, the word “Girl”, “One”, “Never”, “Time”, “Heart” also appear frequently. We concluded that many top songs in this decade have the theme of love.

The second point of interest is the genre of the top songs. There are 50 unique genres, we pick the top 9 genres and make a square plot to see their approximate proportion among the group of top 9 genres. Figure 2 shows the result.

Figure 2: Top 9 common genres of Top Spotify Songs in 2010-2019



During the period from 2010 to 2019, most of the top songs belongs to the pop genre including dance pop, pop, Canadian pop. Dance pop accounts for a large proportion among the top 9 genres.

Since we would like to construct models to explore the data in the next part and our data has the year variable, we need to check if features of the top songs change over time. Table 1 shows the top 5 genres from 2010 to 2019.

Table 1: The top 5 genres each year of Top Spotify Songs in 2010-2019

Year	Top1	Top2	Top3	Top4	Top5
2010	dance pop	pop	atl hip hop	hip pop	barbadian pop
2011	dance pop	barbadian pop	pop	british soul	acoustic pop
2012	dance pop	pop	canadian pop	barbadian pop	baroque pop
2013	dance pop	boy band	pop	electro	canadian pop
2014	dance pop	pop	neo mellow	colombian pop	electropop
2015	dance pop	canadian pop	pop	canadian contemporary r&b	boy band
2016	dance pop	canadian pop	electropop	british soul	canadian contemporary r&b
2017	dance pop	pop	canadian pop	electropop	canadian contemporary r&b
2018	dance pop	pop	canadian pop	hip hop	edm
2019	pop	dance pop	electropop	escape room	boy band

Table 1 shows that the top genre each year is dance pop which number is several times of the number of the second genre, except for 2019. In 2019 the most common genre of top songs is pop which has 9 , the second one is dance pop which has 7. Thus we conclude that the popular genre does not change by year. So we can remove the year indicator when constructing the model.

The third point of interest is the artists of the top songs. We count the number of songs by all the artists who has top songs in these 10 years and pick the top 30 artists by the number of their songs. Figure 3 shows the result.

Figure 3: Top 30 artists of Top Spotify Songs in 2010-2019

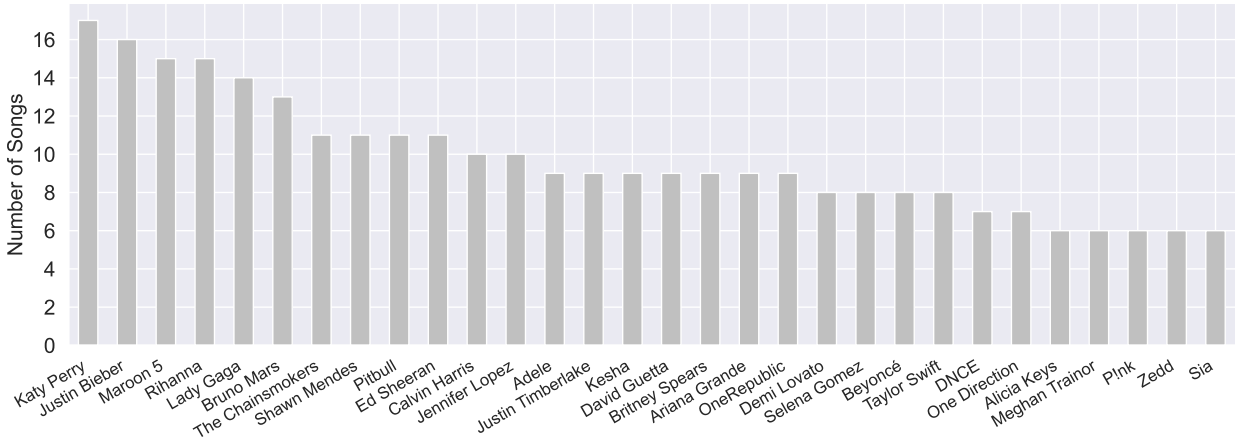


Figure 3 shows that there are 10 artists have more than 10 songs in the top songs list from 2010-2019, and 15 artists' number of songs are in the rank from 5 to 10. Then we can conclude that in these 10 years the top singers often have consistent high quality works in a long time.

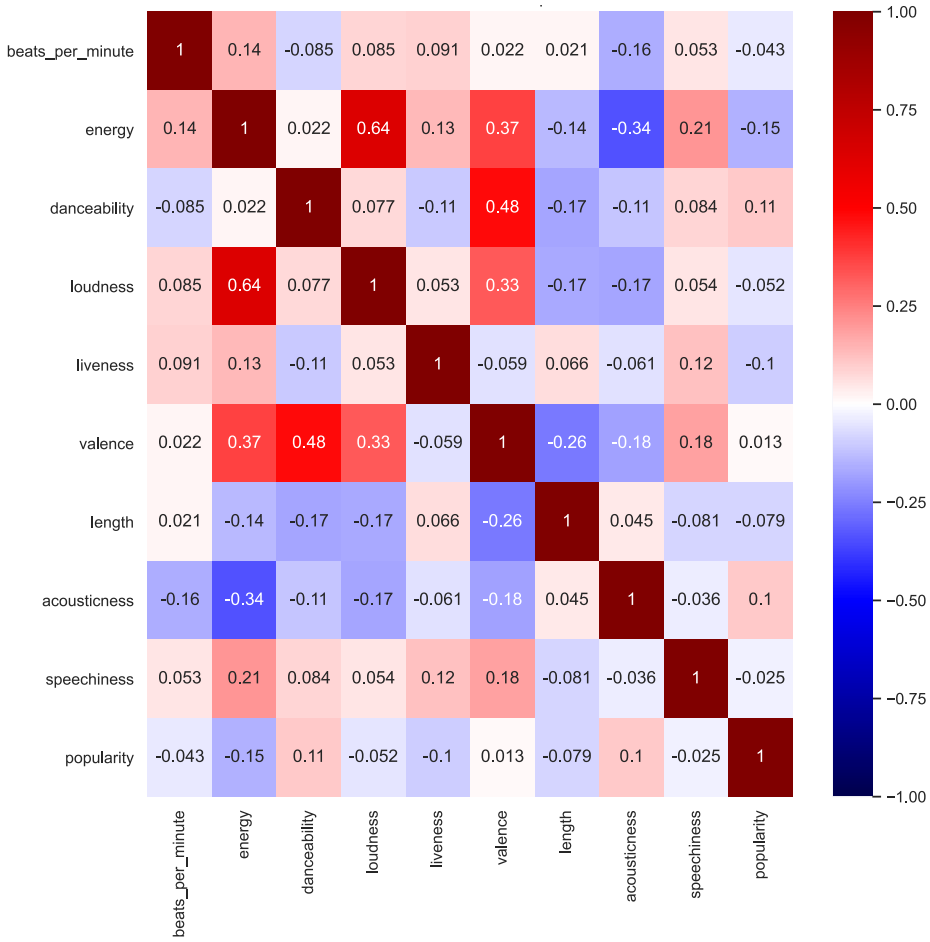
To further prove our conclusion, we sort the artists by the popularity of their songs and pick the top 5 artists each year. Table 2 shows the result.

Table 2: The top 5 artists each year of Top Spotify Songs in 2010-2019

Year	Top1	Top2	Top3	Top4	Top5
2010	Christina Aguilera	The Black Eyed Peas	Kesha	Alicia Keys	Lady Gaga
2011	Lady Gaga	Beyoncé	Jennifer Lopez	Rihanna	Bruno Mars
2012	Rihanna	Katy Perry	Maroon 5	P!nk	Taylor Swift
2013	Justin Timberlake	One Direction	Miley Cyrus	Demi Lovato	Daft Punk
2014	Birdy	Katy Perry	Beyoncé	Bruno Mars	Pharrell Williams
2015	Justin Bieber	Ed Sheeran	Maroon 5	Ariana Grande	Meghan Trainor
2016	Britney Spears	DNCE	Shawn Mendes	The Chainsmokers	Meghan Trainor
2017	Katy Perry	DNCE	The Chainsmokers	Lana Del Rey	OneRepublic
2018	Shawn Mendes	Justin Timberlake	Taylor Swift	Dua Lipa	Liam Payne
2019	Ed Sheeran	The Chainsmokers	Lizzo	Jonas Brothers	Mark Ronson

We find that artists like Lady Gaga, Rihanna, Justin Bieber, Bruno Mars and so on have appear more than once in Table 2 which consistent with our last conclusion.

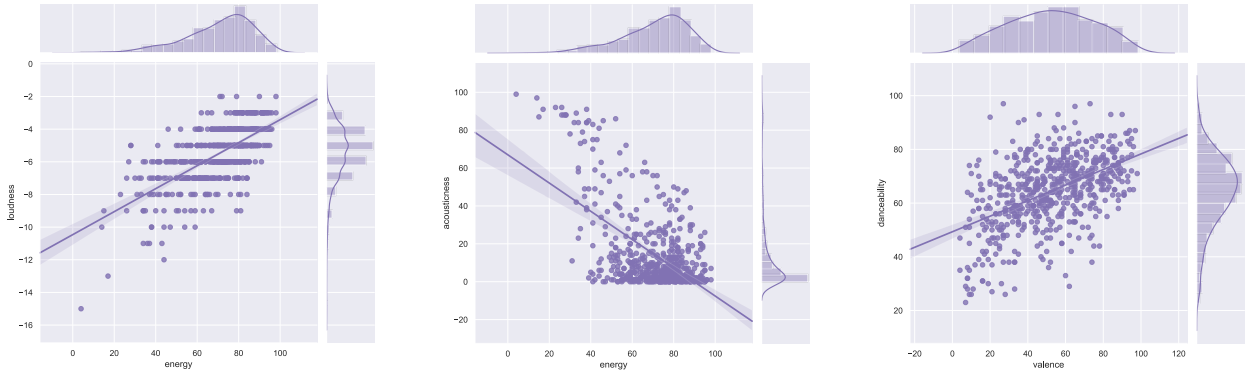
Figure 4: Correlation Heatmap of all Quantitative Variables



The heat map describes the correlation between variables. We can find that there is a relatively strong positive correlation between energy and loudness, and between valence and danceability, and a relatively strong negative correlation between energy and acousticness.

According to the conclusions we got above, we use jointplot to generate fitted lines of the data points and the 95% confidence interval. Figure 5 shows the result.

Figure 5: Jointplots of High Correlated Variables



(left: energy and loudness; middle: energy and acousticness; right: valence and danceability)

The plot on the left is the jointplot of energy and loudness. The data points are evenly distributed, they have relatively strong positive correlation, and the 95% confidence interval is narrow. The plot in the middle is the jointplot of energy and acousticness. they have relatively strong negative correlation, and the 95% confidence interval is relatively wide. Since the data points are concentrated in the 0-20 interval of acousticness, and there are not many data points at the top left, we consider that their relationship is not very robust. The plot on the right is the jointplot of valence and danceability. The data points are evenly distributed, they have relatively strong positive correlation, and the 95% confidence interval is narrow as well.

Part II

In part II, we will try to explore the relationship between qualitative characteristics and popularity of the songs.

First, we use all variables as independent variables to perform OLS regression on popularity. We do this for two purposes, one is finding out the variables which play an important role in deciding the popularity of a song, another is checking if OLS can fit the data well. Table 3 shows the results.

Table 3: OLS Regression on Popularity (Data for 2010-2019)

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.040
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	2.763
Date:	Thu, 10 Dec 2020	Prob (F-statistic):	0.00357
Time:	10:19:08	Log-Likelihood:	-2441.7
No. Observations:	602	AIC:	4903.
Df Residuals:	592	BIC:	4947.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	89.8857	8.958	10.035	0.000	72.293	107.478
beats_per_minute	0.0029	0.024	0.117	0.907	-0.045	0.051
energy	-0.1800	0.058	-3.092	0.002	-0.294	-0.066
danceability	0.0758	0.053	1.418	0.157	-0.029	0.181
loudness	0.9659	0.464	2.082	0.038	0.055	1.877
liveness	-0.0531	0.045	-1.167	0.244	-0.142	0.036
valence	-0.0019	0.033	-0.056	0.955	-0.067	0.063
length	-0.0399	0.018	-2.254	0.025	-0.075	-0.005
acousticness	-0.0229	0.035	-0.651	0.515	-0.092	0.046
speechiness	-0.0162	0.080	-0.202	0.840	-0.174	0.141

Omnibus:	179.571	Durbin-Watson:	0.506
Prob(Omnibus):	0.000	Jarque-Bera (JB):	523.881
Skew:	-1.449	Prob(JB):	1.74e-114
Kurtosis:	6.534	Cond. No.	4.36e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.36e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

From the regression results, the variables that have a significant impact on popularity are energy, loudness, length when the significance level is 0.95. The adjusted R-squared is only 0.026, which reflects that the OLS of all variables can not fit the data well.

Next, we select the data for 2019 and add the indicator "Followers". "Followers" refers to the number of followers of each singer on Spotify. It can largely reflect the singer's reputation and influence. We will use this data to study the influence of a singer's reputation on the popularity of his work. We have renamed the names in data at the beginning to make them easy to understand.

Table 4: OLS Regression on Popularity (Data for 2019)

OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.208
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	1.025
Date:	Wed, 09 Dec 2020	Prob (F-statistic):	0.441
Time:	17:59:24	Log-Likelihood:	-139.72
No. Observations:	50	AIC:	301.4
Df Residuals:	39	BIC:	322.5
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	89.6255	9.719	9.222	0.000	69.967	109.284
beats_per_minute	0.0121	0.027	0.445	0.659	-0.043	0.067
energy	0.0158	0.072	0.220	0.827	-0.130	0.161
danceability	0.0206	0.057	0.360	0.721	-0.095	0.136
loudness	0.1807	0.472	0.383	0.704	-0.774	1.136
liveness	0.0555	0.062	0.899	0.374	-0.069	0.180
valence	-0.0756	0.034	-2.230	0.032	-0.144	-0.007
length	-0.0153	0.018	-0.863	0.393	-0.051	0.021
acousticness	-0.0103	0.037	-0.279	0.782	-0.085	0.064
speechiness	0.0783	0.079	0.988	0.329	-0.082	0.239
Followers	0.0064	0.009	0.710	0.482	-0.012	0.024

Omnibus:	15.775	Durbin-Watson:	1.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.245
Skew:	-1.068	Prob(JB):	2.44e-05
Kurtosis:	5.373	Cond. No.	4.31e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.31e+03. This might indicate that there are strong multicollinearity or other numerical problems.

From the regression results, it can be seen that the variable “Followers” has no significant effect on popularity, that is to say, the reputation of the singer does not have much influence on the popularity of his songs.

Then we go back to the previous step, where we have the conclusion that the variables that have a significant impact on popularity are energy, loudness, length. To verify this conclusion, we use other three machine learning methods which can do variables selection.

Method 1: LASSO Regression

LASSO regression (least absolute shrinkage and selection operator) considers a L_1-penalty

$$h(\beta) = \sum_{j=0}^{p} |\beta_j| = ||\beta||_1$$

The lasso estimator in the β that minimize the loss function:

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)\right)^2 + \lambda \sum_{j=0}^{p} |\beta_j|$$

By reducing the coefficients (β) of some variables to 0, LASSO regression can be used to select variables.

Method 2: Regression Tree

Regression Tree divides the predictor space—that is, the set of possible values for $X_1, X_2, \ldots X_p$—into J distinct and non-overlapping regions, $R_1, R_2, \ldots R_J$. The goal is to find boxes , $R_1, R_2, \ldots R_J$ that minimize the RSS, given by

$$\sum_{j=1}^{J} \sum_{j \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

where $\hat{y}_{R_j}$ is the mean response for the observations within the jth box.
Through the pruning of the regression tree, the effect of screening variables can also be achieved.

Method 3: Principal Components Analysis (PCA)

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables.

Given a set of features $X_1, X_2, \ldots X_p$, principal components are independent and linear combinations of features.

The first component is $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$, which has the largest variance.

The second component is $Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p$, which is independent of Z_1, and has the second largest variance.

Since it is often the case that the signal in a data set is concentrated in its first few principal components, PCA can lead to less noisy results. Thus we can use PCA to select some representative variables.

We first use the above three models to select variables, and then use these variables to perform OLS regression on popularity, and find variables that have a significant impact on popularity by the results of OLS regression.

Table 5: Variables selected by LASSO Regression, Regression Tree and PCA

Models	Variables	Variables that are significant in the OLS regression results
LASSO Regression	energy, danceability, loudness, liveness, length	energy (negaitve) loudness (positive) length (negative)
Regression Tree	beats_per_minute, energy, loudness, valence, length, acousticness	energy (negaitve) loudness (positive) length (negative)
PCA	beats_per_minute, energy, danceability, loudness, valence, acousticness	energy (negaitve) loudness (positive)

In the end, depending on the methods we have used, we conclude that the variables that have a significant impact on popularity are energy, loudness, length. Loudness has a positive impact, while energy and length have negative impact. These impacts make sense, since intuitively listeners usually prefer louder music, and too long songs will make the listeners lose their patience.

Conclusions and directions for future research

From the study above we conclude that in the time period from 2010 to 2019, many top songs have the theme of love. Most of the top songs belongs to the pop genre including dance pop, pop, Canadian pop. The top singers often have consistent high quality works in a long time. When doing regression on popularity, we are not able to establish an effective model of the factors affecting a song's popularity, however, we can conclude that the variables that have a significant impact on popularity are energy, loudness, length. Loudness has a positive impact, while energy and length have negative impact.

One difficulty we encounter is that we do not have a proper quantitative indicator of the popularity of artists each year. As we intuitively speculate that the popularity of the artist will affect the popularity of his/her songs. Even though the follower indicator is not significant in the OLS regression of 2019 data.

Another problem to be solved in the future is the limitation of the data set. Since all the songs in this data set are the most popular songs of that year, and there is no big difference in their popularity. In other words, we lack data on songs with real low popularity.

Regarding the inadequacy of the model, there is another possibility that there is no significant relationship between the popularity of a song and its singer or other indicators of the song (like energy, loudness, length). We cannot judge the popularity and commercial value of a song through these indicators, but our ability to appreciate music.