Spotify Top Songs Analysis

Arnob Chanda, Kai Mei, Hanying Chen

1. Introduction

Since the digitization of the music label industry in the past decade, online music streaming has become the main medium for people to listen to and interact with music in their daily life. Due to this new form of music consumption, vast amounts of music related data can be utilized to help companies better understand the preferences of music listeners.

Spotify is a Swedish audio streaming and media services provider founded on 23 April 2006 by Daniel Ek and Martin Lorentzon. It is the world's largest music streaming service provider, with over 381 million monthly active users, including 172 million paying subscribers, as of September 2021.

One of the company's key product features is the service of providing customized music recommendations based on the user's favorite singers, music genres and listening history. The current music recommendation analytics are more focused on the user specific metrics and we believe adding the song specific metrics such as song duration, danceability, loudness, liveliness etc. can help Spotify improve its recommendation model.

Our study is to provide a music specific metrics focused complementary popularity prediction model to mitigate the limitations of user specific metrics focused model.

In this study, we gathered each song's popularity score and its corresponding metrics including beats per minute, energy, danceability, loudness, length, valence etc. from the top songs database of Spotify. We built multiple variable regression models to forecast song's popularity based on its song specific metrics.

2. Related Work

The main idea of a project like this is to guess the popularity of something based on some known metrics. This is used everywhere where there is some kind of product or idea targeted towards a specific person or groups of people. For example, in the case of Netflix, by using the known variables of a movie or a series, the popularity of it can be estimated. Those movies or series would then be recommended and if the person watches the recommended movie or series, the model is successful. Also in cases of advertising companies or agencies, models like these can be used to target a specific ad or a piece of information towards a person or a group of people. Companies like Facebook can (and have) use known viewing patterns of a user correlate it with the known metrics of an ad and then using that can target a specific ad or piece of information towards them.

In general, a basic recommendation system is the basis for maintaining user retention which allows a company to make their platform more user-friendly and in turn generate more revenue.

3. Data

3.1 Dataset Description

For our dataset we are using the top spotify songs from 2010 to 2019 from kaggle with each song having 13 variables that can be explored. [1].

The dataset we have consists of 15 columns and 603 observations. Each of these observations have the below information.

- Title
- Artist
- Genre
- Year
- Bpm
- Energy
- Danceability
- Db
- Liveness
- Valence
- Duration
- Acousticness
- Speechiness
- Popularity

Most of the observation points are self explanatory except a few which are explained here. Energy value shows how energetic a song is. It's a value that ranges from 0 to 100. Danceability is a value that shows how easy it is to dance to a certain song. This also ranges from 0 to 100. Liveness is a value that shows how likely a song is a live recording. Valence value shows how positive a song is. Acousticness shows how acoustic the song is. Speechiness value increases the spoken word a song contains. All the above explained values can range from 0 to 100.

3.2 Data Screening

By taking a first look at the data, we found that the first column was a row number and was not needed in the analysis so it was removed.

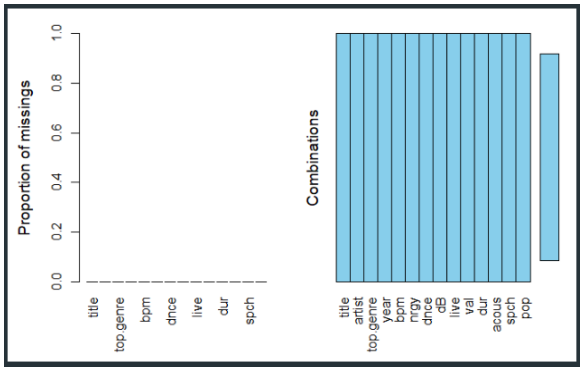
Next, a check for duplicate songs was done. We found that there are some instances where the song name is the same but the artist is different and in other cases that have similar song names only the year was different. In the latter scenario we decided to go with the first occurrence of the song and remove the second occurrence of it, and for cases that had the same song name but different artist we kept them all. This reduced our data set to 587 observations.

3.2.1 Accuracy

For accuracy check, all the variables in the dataset were checked if they are all in range. As mentioned before, many variable values in this dataset range between 0 to 100. By taking a look at the summary of the dataset, it was determined that all the values were in fact in range and accurate.

3.2.2 Missing Values

Next, a test for missing values was done. We used VIM to check the proportion of missing values and found none.



3.2.3 Outliers

Data was then tested for outliers. Outliers were calculated based on Leverage, Mahal scores and Cook's distance.

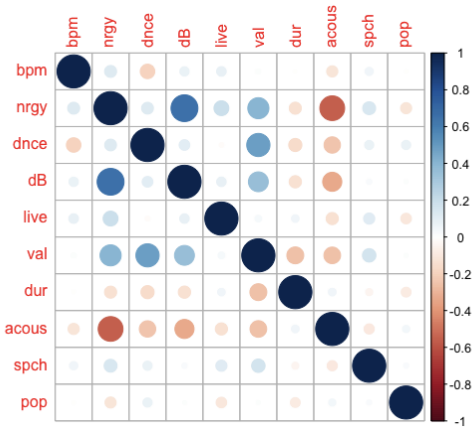
Mahalanobis scores, leverage values and cook's distances were calculated and for each data point more than these scores were marked as outliers. In total with these, we found 25 outliers, out of which 17 data points were part of 2 of the outlier group and 8 data points were part of all 3 outlier groups. We proceeded to remove all these outlier values and that reduced our data set to 562 observations.

3.3 Check Assumptions

3.3.1 Additivity

To assess the additivity assumption, we computed the correlation matrix and visualized it using the `corrplot` package (see figure 1). From the graph, we can tell that some pairs of variables are correlated to some extent. For example, `dB` and `nrgy` are positively correlated with Pearson correlation=0.6434, which makes intuitive sense since the louder the music, the more energy it might bring.

Figure 1. Correlation plot

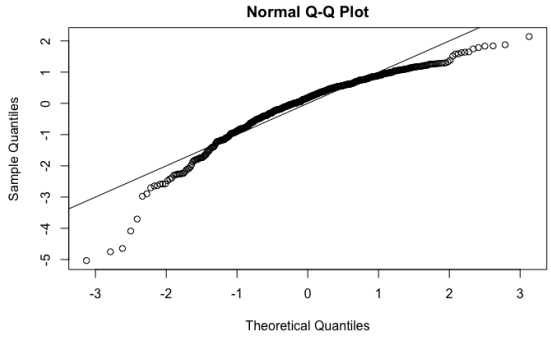


There is no concerning multicollinearity according to the graph, therefore the additivity assumption is met.

3.3.2 Linearity

We validated the linearity assumption by fitting a dummy linear model with `pop` being the dependent variable, and plotting the normal QQ plot (figure 2). We observed light tails at both ends, indicating the potential non-linearity in the dataset.

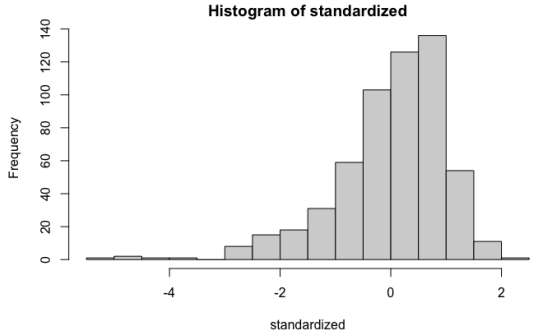
Figure 2. Q-Q plot



3.3.3 Normality

In order to test normality, we applied histogram on the standardized residuals (figure 3). The distribution is left skewed based on the graph, so the normality assumption does not hold well.

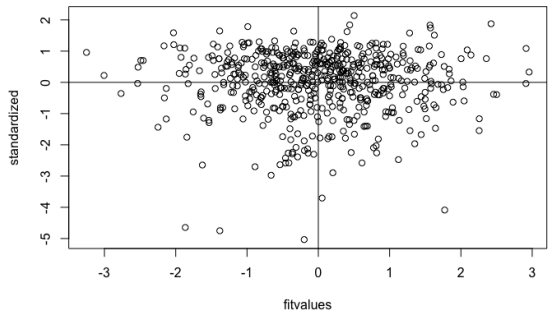
Figure 3. Histogram of standardized residuals



3.3.4 Homogeneity and Homoscedasticity

Homogeneity and homoscedasticity assumptions were assessed based on the scatterplot of standardized residuals (figure 4). Since the spread of data points across x- and y-axis are relatively even, both assumptions are met.

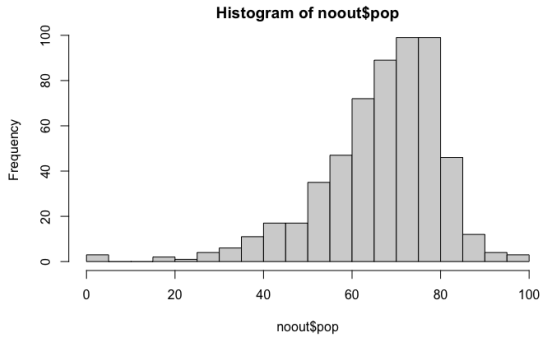
Figure 4. Fitted values vs. standardized residuals



3.4 Data Visualization

3.4.1 Popularity

We firstly visualized the distribution of the dependent variable `pop`. Based on the histogram below, we can see it is left skewed. There are three songs with zero popularity score which seems like data error, given that the median is 69 and 25th quantile is 60. However, without understanding how `pop` was determined, we decided to keep them in this analysis.



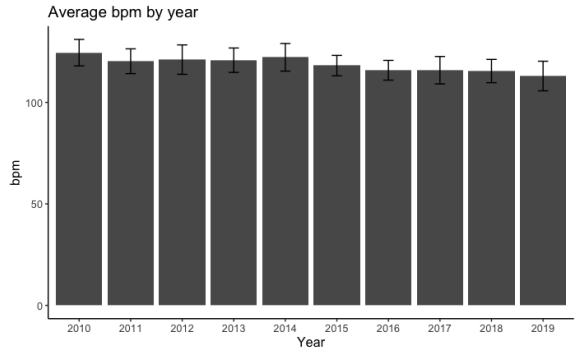
Songs with `pop=0`

	title <chr>	artist <chr>	top.genre <chr>	year <int>	pop <int>
51	Hello	Martin Solveig	big room	2010	0
139	Blow Me (One Last Kiss)	P!nk	dance pop	2012	0
268	Not a Bad Thing	Justin Timberlake	dance pop	2014	0

3.4.2 Trend of `bpm`

From figure 5, we can tell that the average bpm is decreasing over time, indicating a trend that popular songs tend to have slower bpm nowadays.

Figure 5. Average `bpm` over years



3.4.3 Artist

We counted the number of songs by artists and sorted them in descending order. The top 10 popular artists are shown in table 1.

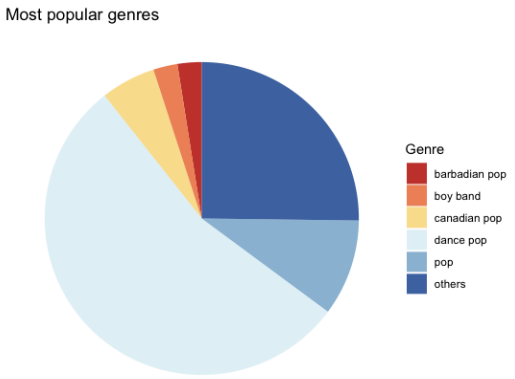
Table 1. Top 10 popular artists

Artist <fctr>	Count <int>
Katy Perry	17
Lady Gaga	14
Maroon 5	14
Rihanna	14
Justin Bieber	13
Bruno Mars	11
Ed Sheeran	11
Pitbull	11
The Chainsmokers	11
Calvin Harris	10

3.4.4 Genres

We have a total 50 distinct genres in the dataset. We only took the top five genres and group the rest into 'others' category, then visualized using the pie chart. Figure 6 shows that the most popular genres are dance pop, others, pop, canadian pop, barbadian pop and boy band respectively.

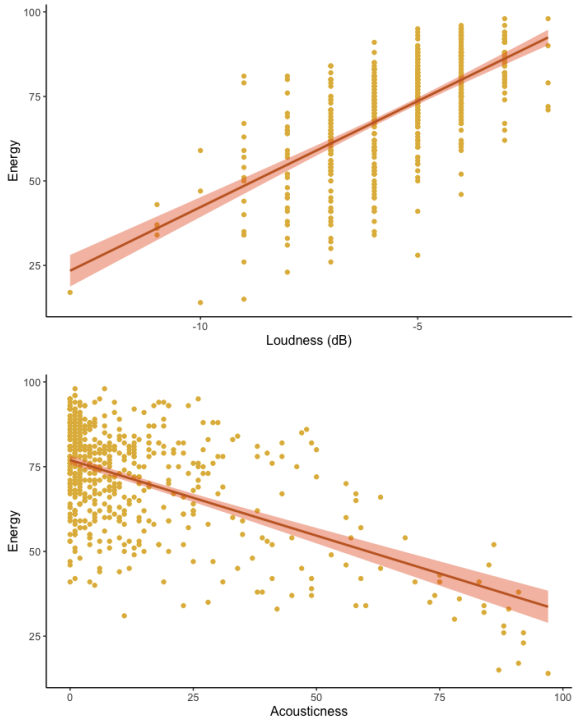
Figure 6. Top 5 popular genres



3.4.5 Loudness, acousticness and energy

We used scatterplot to analyze the relationship between loudness, acousticness and energy. Figure 7 and 8 confirmed that there is a positive correlation between loudness and energy and a negative correlation between acousticness and energy.

Figure 7. `dB`, `acous` vs. `nrgy`



4. Technical Approach

4.1 Model Building Process

We chose multiple linear regression to build the forecast model and pop score is our dependent variable. The model was built in three stages. We first run the pop score against all other numerical variables to build a full model. Then we run pop score against nrgy, live, dur, dnce, and acous, which were shown as highly correlated with pop score in our exploratory data analysis. Finally, we built a stepwise regression model based on both backward and forward selection. Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on the form of a forward, backward, or

combined sequence of F-tests or t-tests. In our analysis, we used both directions.
The ‘lm’ function was used in the first two steps. Then we installed the ‘leap’ and ‘MASS’ packages and used the ‘stepAIC’ function from the packages to build a stepwise regression model. The full model we built in the first step was an input in the ‘stepAIC’ function and we chose ‘both’ and ‘FALSE’ for direction and trace arguments respectively.
After building the three models, we used root mean square error (RMSE) as our model evaluation metric. RMSE is the standard deviation of the residuals, and it tells how concentrated the data is around the line of best fit. RMSE is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

4.2 Model Results

4.2.1 Full Model

We used pop as a dependent variable and all other numerical variables including bpm, nrgy, dnce, dB, live, val, dur, acous and spch as independent variables. The modeling result is shown below:

```
Call:
lm(formula = pop ~ ., data = noout[-c(1, 2, 3, 4)])

Residuals:
    Min      1Q  Median      3Q     Max
-66.326  -6.331   2.356   9.286  28.610

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.627563   9.142370   9.694  < 2e-16 ***
bpm          0.015860   0.025587   0.620 0.535612
nrgy        -0.191369   0.057732  -3.315 0.000977 ***
dnce         0.088435   0.054260   1.630 0.103700
dB           1.166329   0.474544   2.458 0.014283 *
live        -0.077270   0.046247  -1.671 0.095322 .
val         -0.012441   0.032975  -0.377 0.706114
dur         -0.036647   0.019220  -1.907 0.057069 .
acous       -0.009102   0.036643  -0.248 0.803923
spch         0.112403   0.097596   1.152 0.249928

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.57 on 557 degrees of freedom
Multiple R-squared:  0.0457, Adjusted R-squared:  0.03028
F-statistic: 2.964 on 9 and 557 DF,  p-value: 0.00188
```

The model is statistically significant as p-value of F-statistic is below 0.01.

4.2.2 Model based on High Correlation Variables

Then, We used pop as a dependent variable and ran the regression against nrgy, live, dur, dnce, and acous, which were shown as highly correlated with pop score in our exploratory data analysis. The modeling result is shown below:

```
Call:
lm(formula = pop ~ nrgy + live + dur + dnce + acous, data = noout[-c(1,
    2, 3, 4)])

Residuals:
    Min      1Q  Median      3Q     Max
-66.067  -5.792   2.657   8.927  27.122

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 79.604522   6.930839  11.486   <2e-16 ***
nrgy        -0.108247   0.045129  -2.399   0.0168 *
live        -0.077677   0.046087  -1.685   0.0925 .
dur         -0.038831   0.018954  -2.049   0.0410 *
dnce         0.078679   0.046694   1.685   0.0926 .
acous       -0.007393   0.036433  -0.203   0.8393

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.61 on 561 degrees of freedom
Multiple R-squared:  0.03326,    Adjusted R-squared:  0.02464
F-statistic:  3.86 on 5 and 561 DF,  p-value: 0.001902
```

The model is statistically significant as p-value of F-statistic is below 0.01.

4.2.3 Stepwise Model

Finally, we used ‘stepAIC’ to build a stepwise regression model based on the full model we built in the first step. Five variables including nrgy, dnce, dB, live and dur were selected by the function. The result of stepwise regression model is shown below:

```
Call:
lm(formula = pop ~ nrgy + dnce + dB + live + dur, data = noout[-c(1,
    2, 3, 4)])

Residuals:
    Min      1Q  Median      3Q     Max
-66.597  -6.086   2.677   8.945  28.156

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.50962    7.64763  11.704  < 2e-16 ***
nrgy        -0.17710    0.04947  -3.580 0.000373 ***
dnce         0.07891    0.04536   1.740 0.082472 .
dB           1.08411    0.46626   2.325 0.020420 *
live        -0.07086    0.04593  -1.543 0.123421
dur         -0.03603    0.01887  -1.910 0.056689 .


Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 561 degrees of freedom
Multiple R-squared:  0.04241,    Adjusted R-squared:  0.03388
F-statistic:  4.97 on 5 and 561 DF,  p-value: 0.0001837
```

The model is statistically significant as p-value of F-statistic is below 0.01.

With either model, we can see nrgy and dB are significant variables. To interpret the estimate of coefficient, take the stepwise model for example, b(nrgy) is -0.17710, meaning holding everything else unchanged, increasing one unit of nrgy would decrease the pop score by 0.177 on average.

5. Test and Evaluation

5.1 Model Evaluation

5.1.1 Predicted vs. Actual

We plotted the predicted popularity versus actual popularity (see figure 9 and 10) to visualize the model performance.

Figure 9. Predicted popularity (full model) vs. actual popularity

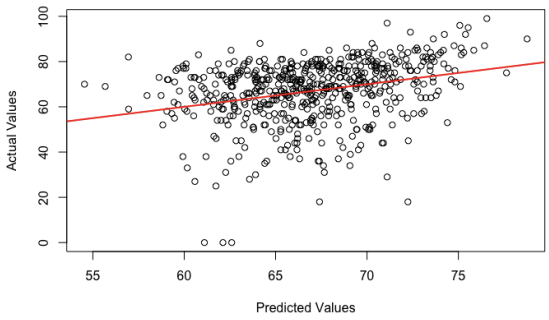
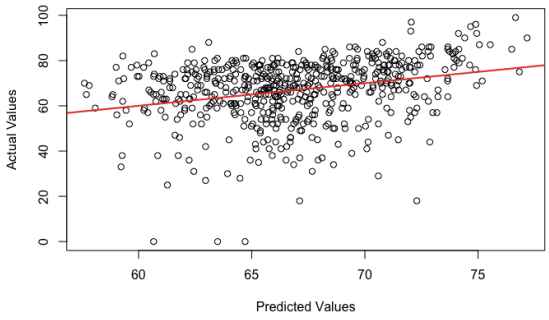


Figure 10. Predicted popularity (stepwise model) vs. actual popularity



5.1.2 Evaluation Metric

We mainly used adjusted R^2 and root mean square error (RMSE) as our model evaluation metrics. Compared to regular R^2, adjusted R^2 takes the model's degree freedom (i.e number of independent variables) into account which avoids overfitting. RMSE is commonly used for measuring regression model accuracy.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Table2. RMSE and adjusted R^2 results

	RMSE	Adjusted R^2
Full model	13.16876	0.03028
High correlation model	13.37352	0.0.02464
Stepwise model	13.21132	0.03388

Based on table 2, we can see that the full model has the lowest RMSE while the stepwise model has the highest adjusted R^2.

5.2 Future Work

We identified the following potential directions for the future work.

- Currently, we observed some degrees of skewness in variables including `pop` and `dB`. This can be remediated using non-linear transformations like log transformation.
- To better evaluate the model performance, it is recommended to split the dataset into training and testing sets, so that the model can be assessed using the unseen data. This would help avoid overfitting.
- Collect more data. So far, we only have less than 600 data points to train the model. ThIf we can gather more data from Spotify, the model

would be more significant and predictive.

- Add categorical variable genre into the model. This can be done by applying clustering techniques (e.g KNN) to group the minor genres into major ones.
- More advanced machine learning models can be applied to achieve better accuracy. Models such as polynomial regression, generalized linear model, random forest and xgboost are all good candidates.
- Analyze Spotify user comments for each song using NLP techniques. If we have access to scrape the user reviews or reactions for those songs, we can potentially create word embeddings and train the model with it.