Introduction (Objective & Motivation):

We listen to music every day. Americans spend 32 hours listening to music every week.[1]
With this huge demand for music/music streaming, all music makers want to know: How to make a popular song? What is the characteristic of a popular song?

To answer these questions, we must obtain related data. We obtained the data from Kaggle based on a sample of Spotify song from past decades. Spotify is one of the largest song streaming companies, it has more than 248 million monthly active users and 30 million songs.[2] Its large number of users and songs provide workable and meaningful data for this topic.

Descriptive Analysis:

1) Brief description of Data

We found our dataset on Kaggle. The data includes the songs from 1950 to 2010 all divided into decades. The data includes various attributes from a selection of songs available in Spotify's playlist "All out [xx]'s". Attributions include categorical information such as release year, genre and quantitative variables such as beats per minute, the energy of the song and danceability score, etc.

2) Preliminary analysis on response variable and important predictors

In this part of analysis, we conducted preliminary analysis on our response variable- popularity and important predictors such as years and energy. The report will provide descriptive data of these variables through summary tables, histograms and boxplots.

---

[1] "Time with Tunes: How Technology Is Driving Music Consumption." Nielsen, February 11, 2017. https://www.nielsen.com/us/en/insights/article/2017/time-with-tunes-how-technology-is-driving-music-consumption/.

[2] Silva, Matthew De. "Spotify Is Still the King of Music Streaming-for Now." Quartz, October 28, 2019. https://qz.com/1736762/spotify-grows-monthly-active-users-and-turns-profit-shares-jump-15-percent/.

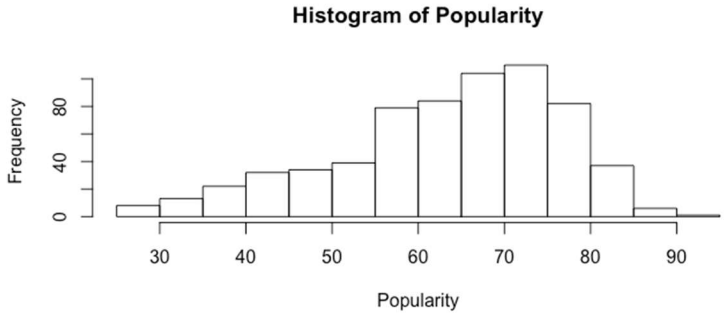
Levenson, Josh. "Apple Music vs. Spotify." Digital Trends, March 26, 2020. https://www.digitaltrends.com/music/apple-music-vs-spotify/.

Response variable: Popularity

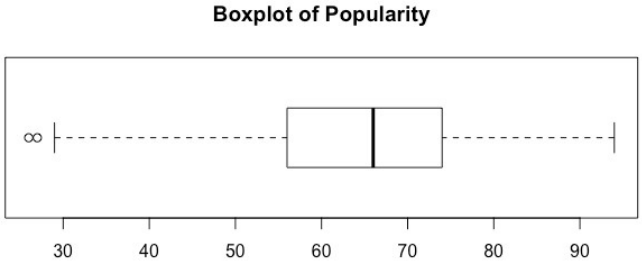
The response variable is calculated between the total number of plays and how recently the track was generally played. The song was played 100 times now scored higher than the song was played 100 times 10 years ago.[3] The range of popularity is [0,100].

Mean	Standard Deviation	Min	25th percentile	50th percentile	75th percentile	Max	IQR	Range
63.94777	13.16594	26	56	66	74	94	18	68

Summary Table of response variable: popularity



Histogram of popularity: the distribution of popularity is not normally distributed, and it is left-skewed.



Boxplot of popularity: we can see there are two outliers.

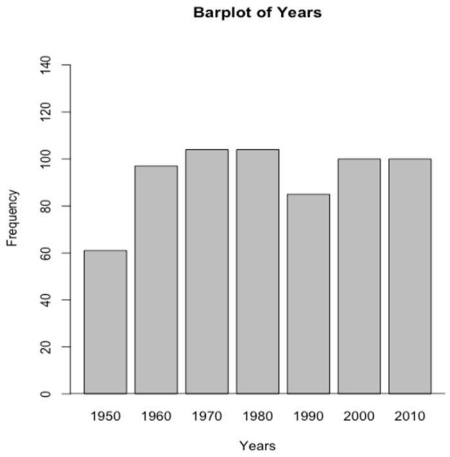
[3] "Get a Track." Spotify for Developers. Accessed March 31, 2020. https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/.

Years

This variable means the decade of this song released. It is qualitative data with 7 categories (1950s, 1960s, 1970s, 1980s, 1990s, 2000s, 2010s). We split it into 6 dummy variables with base-level = 1950s. Based on the article we found, we think that Years should be a significant variable, but the impact is not the strongest one.[4]

Year	1950	1960	1970	1980	1990	2000	2010
Relative Frequency	0.0937	0.149	0.1598	0.1598	0.1305	0.1536	0.1536

Relative Frequency table of variable Years



We can see that every level of the Year's categorical variable has a closed proportion around 13-15% except 1950.

Energy

"Energy represents a perceptual measure of intensity and activity."[5] The range of energy is [0,1]. The energy of sound is a critical criterion for people choosing the song, especially for sport or exercise.[6] People prefer songs with higher energy for sport.

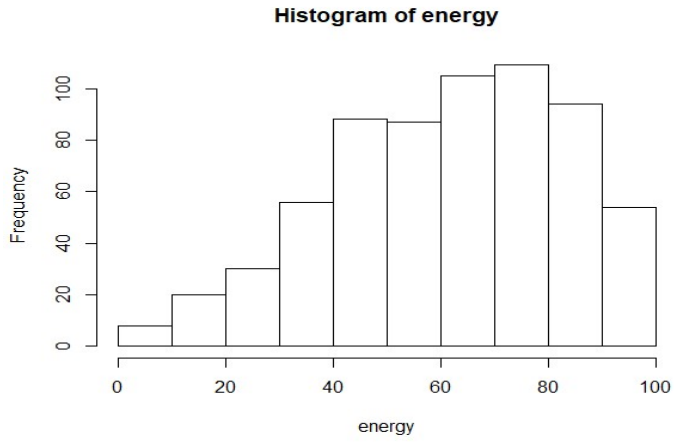
Mean	Standard Deviation	Min	25th percentile	50th percentile	75th percentile	Max	IQR	Range
61.74194	21.60123	6	45	65	79	100	34	94

Summary Table of response variable: popularity

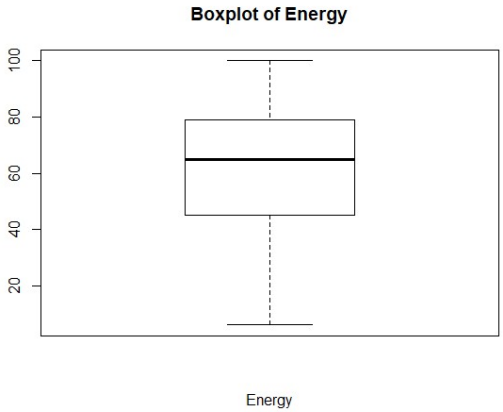
[4] Dodgson, Lindsay. "We Stop Discovering New Music at Age 30, a New Survey Suggests - Here Are the Scientific Reasons Why This Could Be." Business Insider, June 7, 2018. https://www.businessinsider.com/why-we-stop-discovering-new-music-around-age-30-2018-6.

[5] "Get Audio Features for a Track." Spotify for Developers. Accessed March 31, 2020. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/.

[6] Jabr, Ferris. "Let's Get Physical: The Psychology of Effective Workout Music." Scientific American, March 20, 2013. https://www.scientificamerican.com/article/psychology-workout-music/.



Histogram of energy: the distribution of energy is not normally distributed and it is left-skewed.



Boxplot of Energy: we can see there is no outlier.

Inferential Analysis:

1. Give the best linear regression model and fit with your data

- Model equation

$$y = \beta_0 + \beta_1 acous + \beta_2 dur + \beta_3 nrgy + \beta_4 2010s + \beta_5 1960s + \beta_6 2000s + \beta_7 1980s + \beta_9 1970s + \beta_{10} 1990s + \beta_{11} hiphop + \beta_{12} jazz + \beta_{13} rock + \beta_{14} folk + \beta_{15} dB + \beta_{16} rock * nrgy + \beta_{17} rock * 1990s + \beta_{18} rock * 2000s + \beta_{19} hiphop * 2010s + \beta_{20} dur * hiphop + \varepsilon$$

$$Where \quad \varepsilon \sim iid\ N(0, \sigma^2)$$

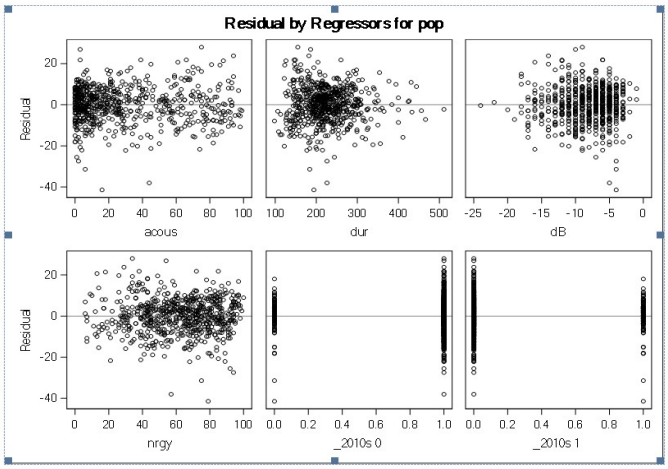
- Fitted equation

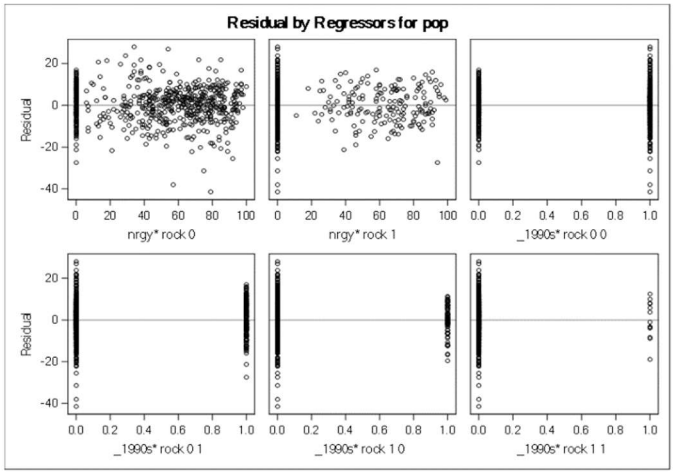
$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}acous + \widehat{\beta_2}dur + \widehat{\beta_3}nrgy + \widehat{\beta_4}2010s + \widehat{\beta_5}1960s + \widehat{\beta_6}2000s + \widehat{\beta_7}1980s + \widehat{\beta_8}1970s + \widehat{\beta_9}1990s + \widehat{\beta_{10}}hiphop + \widehat{\beta_{11}}jazz + \widehat{\beta_{12}}rock + \widehat{\beta_{13}}folk + \widehat{\beta_{14}}dB + \widehat{\beta_{15}}rock * nrgy + \widehat{\beta_{16}}rock * 1990s + \widehat{\beta_{17}}rock * 2000s + \widehat{\beta_{18}}hiphop * 2010s + \widehat{\beta_{19}}dur * hiphop$$

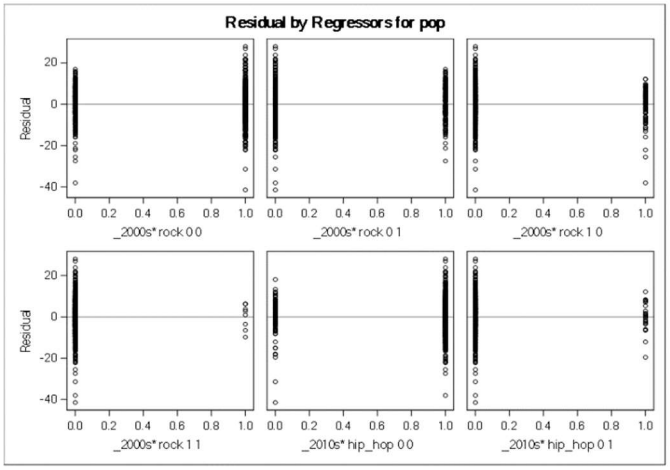
- Evaluate the model utility

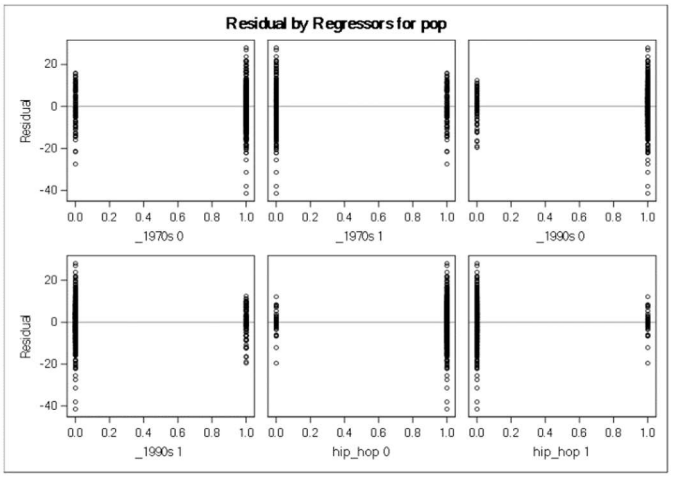
a) Goodness of fit

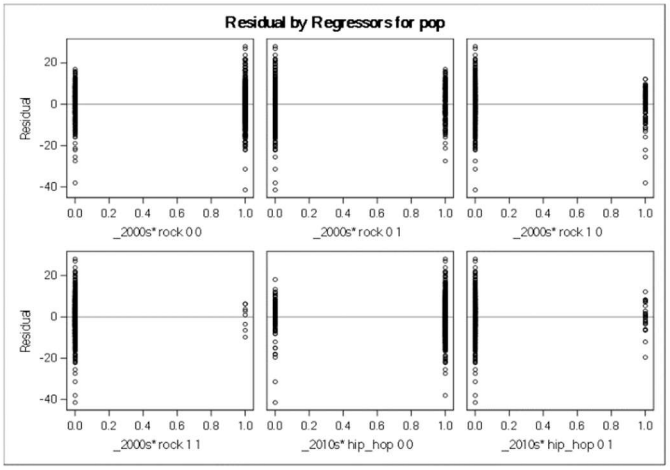
The first assumption is goodness of fit. To examine this, we plot the residuals against different variables to check is there any unusual pattern. If unusual pattern exists, that mean the variable is not good fit to the dependent variables. Here are residual plots against different variables.

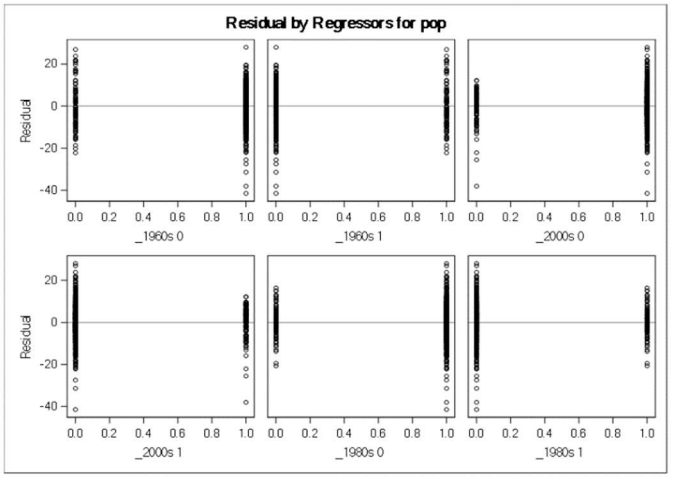










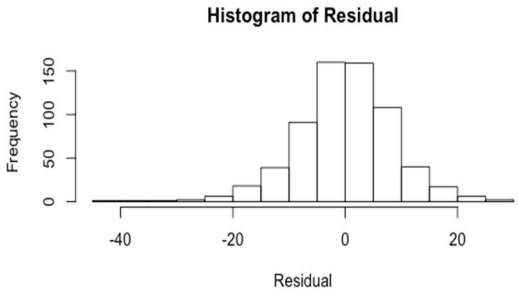


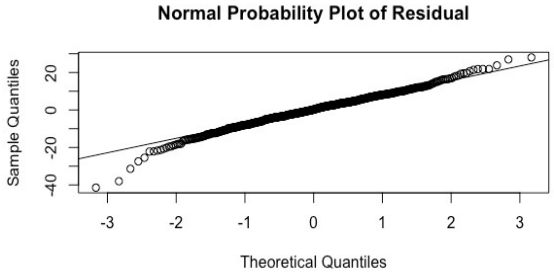
Residual plots against different variables

We can see that there is not any unusual pattern in the plots. No independent variables need to be transformed.

b) Normality

To check the normality, we can plot the histogram and the normal probability plot of residuals. Here is the result we concluded from histogram and probability plot.



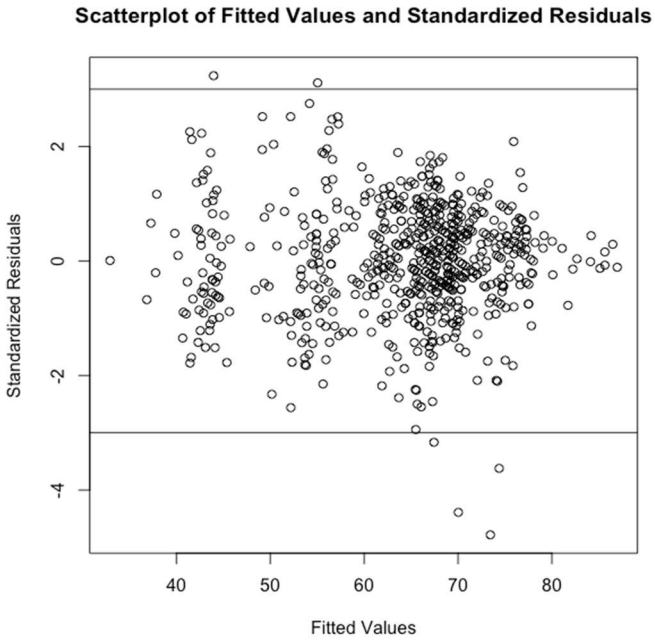


According to the histogram and the normal probability plot, the distribution of residuals is left-skewed and not normally distributed. There are some outliers on the left-hand side of the histogram.

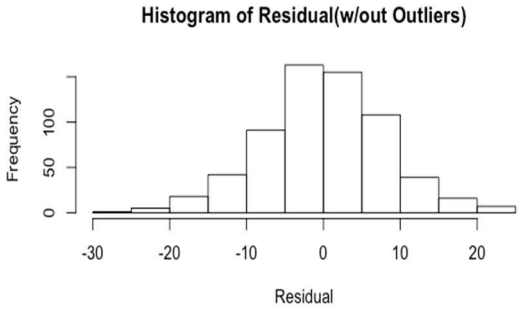
To detect the outliers, we should plot the scatter plot of standardized residual and fitted values to check the outliers. The data points with standardized residuals higher than 3 or smaller than -3 will be treated as outlier.

There are 4 data points with smaller than -3 standardized error, 2 data points with larger than 3 standardized error, total 6 outliers in this data set.

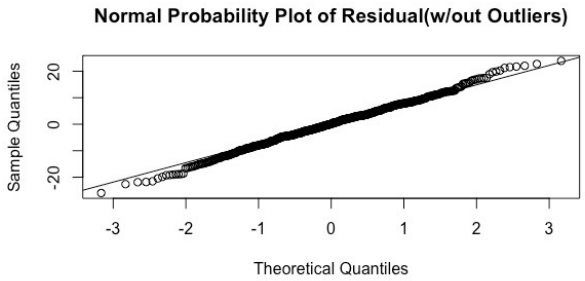
Those outliers are removed from the data set and we apply a new data set to our suggested method. The histogram and the normal probability plot of residuals without outliers are shown below:



scatter plot of standardized residual



Histogram of Residual without outliers

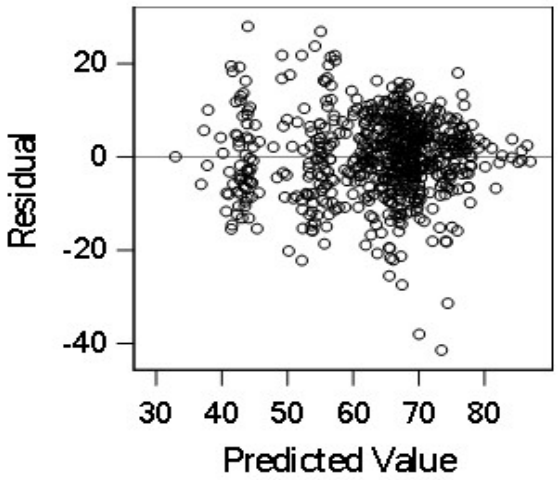


Normal probability plot of Residual without outliers

After removing the outlier, the distribution of the residual is nearly symmetric and normally distributed. The suggested model fulfills the assumption of normality.

c) Homoscedasticity

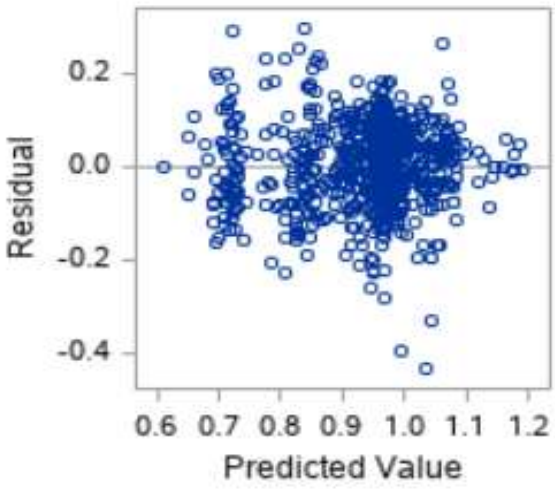
To check the homoscedasticity, we can plot the scatter plot of residuals against fitted value.



From residual plot of predicted model, we can detect a pattern which leading us to conclude that our response probably generated from binomial experiment. The graph shows a football-shaped pattern (smaller variances for small and large values of $\hat{y}$ and larger variance for intermediate values of $\hat{y}$). This indicates unequal variances. Therefore, we try to stabilize dependent variable pop as $sin^{-1}\sqrt{pop}$

After the y-transformation, the residual plot is shown as below:

There is no obvious change after transformation. For model simplicity, we do not implement the transformation.



Residual plot of Residual with variable transformation

d) Model Explanation

After we checked model utility, we tried to fit data into our model through R studio. Here is the result of R studio.

```
Call:
lm(formula = pop ~ acous + dur + nrgy + X2010s + X1960s + X2000s +
    X1980s + X1970s + X1960s + X1990s + hip.hop + jazz + rock +
    folk + dB + rock:nrgy + rock:X1990s + rock:X2000s + hip.hop:X2010s +
    dur:hip.hop, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9654  -4.7782   0.1289   5.1213  23.8985

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     53.554244   3.322191  16.120  < 2e-16 ***
acous           -0.027237   0.017550  -1.552  0.12116
dur              0.003809   0.006905   0.552  0.58143
nrgy            -0.151891   0.027540  -5.515 5.10e-08 ***
X2010s          34.042771   1.835501  18.547  < 2e-16 ***
X1960s          12.107991   1.492719   8.111 2.66e-15 ***
X2000s          28.127388   1.924334  14.617  < 2e-16 ***
X1980s          25.149591   1.757012  14.314  < 2e-16 ***
X1970s          22.799198   1.650437  13.814  < 2e-16 ***
X1990s          23.942075   1.895938  12.628  < 2e-16 ***
hip.hop         18.771608   8.310513   2.259  0.02424 *
jazz           -11.571588   3.523231  -3.284  0.00108 **
rock            -3.055362   2.437717  -1.253  0.21054
folk            -6.332787   2.663010  -2.378  0.01770 *
dB               0.377777   0.145505   2.596  0.00964 **
nrgy:rock        0.104747   0.036971   2.833  0.00476 **
X1990s:rock     -7.806130   2.669180  -2.925  0.00357 **
X2000s:rock      3.697777   3.228078   1.146  0.25244
X2010s:hip.hop   1.226471   3.716592   0.330  0.74151
dur:hip.hop     -0.063610   0.030227  -2.104  0.03574 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.221 on 625 degrees of freedom
Multiple R-squared:  0.6133,    Adjusted R-squared:  0.6015
F-statistic: 52.17 on 19 and 625 DF,  p-value: < 2.2e-16
```

R result of fitted model

We can conclude our result from R studio as following:

$$\hat{y} = 53.5542 - 0.0272x_1 + 0.0038x_2 - 0.1519x_3 + 12.108x_4 + 22.7992x_5 + 25.1496x_6 + 23.9421x_7 + 28.1274x_8 + 34.0428x_9 + 18.7716x_{10} - 11.5716x_{11} - 3.0554x_{12} - 6.3328x_{13} + 0.3778x_{14} + 0.1047x_3x_{12} - 7.8061x_7x_{12} + 3.6978x_8x_{12} + 1.2265x_9x_{10} - 0.0636x_2x_{10}$$

Where:

$x_1 = Score\ of\ Acousitc$

$x_2 = Duration\ of\ song\ (in\ second)$

$x_3 = Score\ of\ Energy$

$x_4 = \begin{cases} = 1 & from\ 1960 \\ = 0 & Otherwise \end{cases}$

$x_5 = \begin{cases} = 1 & from\ 1970 \\ = 0 & Otherwise \end{cases}$

$x_6 = \begin{cases} = 1 & from\ 1980 \\ = 0 & Otherwise \end{cases}$

$x_7 = \begin{cases} = 1 & from\ 1990 \\ = 0 & Otherwise \end{cases}$

$x_8 = \begin{cases} = 1 & from\ 2000 \\ = 0 & Otherwise \end{cases}$

```
Call:
lm(formula = pop ~ acous + dur + nrgy + X2010s + X1960s + X2000s +
    X1980s + X1970s + X1960s + X1990s + hip.hop + jazz + rock +
    folk + dB + rock:nrgy + rock:X1990s + rock:X2000s + hip.hop:X2010s +
    dur:hip.hop, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9654  -4.7782   0.1289   5.1213  23.8985

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    53.554244   3.322191  16.120  < 2e-16 ***
acous          -0.027237   0.017550  -1.552  0.12116
dur             0.003809   0.006905   0.552  0.58143
nrgy           -0.151891   0.027540  -5.515 5.10e-08 ***
X2010s         34.042771   1.835501  18.547  < 2e-16 ***
X1960s         12.107991   1.492719   8.111 2.66e-15 ***
X2000s         28.127388   1.924334  14.617  < 2e-16 ***
X1980s         25.149591   1.757012  14.314  < 2e-16 ***
X1970s         22.799198   1.650437  13.814  < 2e-16 ***
X1990s         23.942075   1.895938  12.628  < 2e-16 ***
hip.hop        18.771608   8.310513   2.259  0.02424 *
jazz          -11.571588   3.523231  -3.284  0.00108 **
rock           -3.055362   2.437717  -1.253  0.21054
folk           -6.332787   2.663010  -2.378  0.01770 *
dB              0.377777   0.145505   2.596  0.00964 **
nrgy:rock       0.104747   0.036971   2.833  0.00476 **
X1990s:rock    -7.806130   2.669180  -2.925  0.00357 **
X2000s:rock     3.697777   3.228078   1.146  0.25244
X2010s:hip.hop  1.226471   3.716592   0.330  0.74151
dur:hip.hop    -0.063610   0.030227  -2.104  0.03574 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.221 on 625 degrees of freedom
Multiple R-squared:  0.6133,    Adjusted R-squared:  0.6015
F-statistic: 52.17 on 19 and 625 DF,  p-value: < 2.2e-16
```

R result of fitted model

$$x_9 = \begin{cases} = 1 \ from\ 2010 \\ = 0 \ Otherwise \end{cases} \quad Baseline = 1950$$

$$x_{10} = \begin{cases} = 1 \ HipHop\ song \\ = 0 \ Otherwise \end{cases}$$

$$x_{11} = \begin{cases} = 1 \ Jazz\ song \\ = 0 \ Otherwise \end{cases}$$

$$x_{12} = \begin{cases} = 1 \ Rock\ song \\ = 0 \ Otherwise \end{cases}$$

$$x_{13} = \begin{cases} = 1 \ Folk\ song \\ = 0 \ Otherwise \end{cases} \quad Baseline = Pop\ Song$$

$$x_{14} = Loudness\ (in\ dB)$$

For every 1-unit increase in Score of Acoustic, we can expect the Popularity will decrease 0.0272 units.

For every 1-second increase in Duration, we can expect the Popularity will decrease 0.0598 for Hip Hop Song, increase 0.0038 unit for otherwise.

For every 1-unit increase Score of Energy, we can expect the Popularity will decrease 0.0472 units for rock song, decrease 0.1519 for otherwise.

We can expect the mean difference of Popularity between 1960s and 1950s is 12.108.

We can expect the mean difference of Popularity between 1970s and 1950s is 22.7992.

We can expect the mean difference of Popularity between 1980s and 1950s is 25.1496.

We can expect the mean difference of Popularity between 1990s and 1950s is 16.136 for Rock Song, 23.9421 for otherwise.

We can expect the mean difference of Popularity between 2000s and 1950s is 31.8225 for Rock Song, 28.1247 for otherwise

We can expect the mean difference of Popularity between 2010s and 1950s is 35.2693 for HipHop Song, 34.0428 for otherwise. We can expect the mean difference of Popularity between Hip Hop Song and Pop Song is 19.9981-0.0636x2 for 2010s, 18.7716-0.0636x2 for otherwise, holding x2 remains unchanged.

We can expect the mean difference of Popularity between Jazz Song and Pop Song is -11.5716.

We can expect the mean difference of Popularity between Rock Song and Pop Song is 10.8615+0.1074x3 for 1990s, 0.6424+0.1074x3 for 2000s, and-3.0554+0.1074x3 for otherwise, holding x3 remain unchanged.

We can expect the mean difference of Popularity between Folk Song and Pop Song is -6.3328.

For every 1 dB increase in Loudness, we can expect the Popularity will increase 0.3778 units.

e) T-test for Important Variables

i) Energy:

Setting Hypothesis as following:

$$H_0: \beta_3 \leq 0$$
$$H_1: \beta_3 > 0$$

Given that the significant level is 1%, the p-value of β_3 is 1-(5.10e-08)/2, which is larger than 0.01, we do not have enough evidence to reject the null hypothesis. We conclude that the effect of Level of Energy is not significantly larger than 0.
In other words, people are not significantly preferring a song with higher level of energy.

ii) Years:

As we split year in to 6 dummy variables, in order to avoid accumulating high type I error, we conduct the t-test for each dummy variables with the α =0.0085.

Setting Hypothesis as following:

$$H_0: \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0, \beta_9 = 0$$
$$H_1: \beta_4 \neq 0, \beta_5 \neq 0, \beta_6 \neq 0, \beta_7 \neq 0, \beta_8 \neq 0, \beta_9 \neq 0$$

The p-value of β_4 is 2.66e-15, $\beta_5, \beta_6, \beta_7, \beta_8, \beta_9$ is < 2e-16 respectively, which is closed to zero. We have enough evidence to reject the null hypothesis and conclude that the effect of each decades is significantly larger than 0.

2. Using model for estimation and prediction

- Estimation

We draw 10 samples from the datasets and fit it in the regression model. The result is:

Sample	1	2	3	4	5	6	7	8	9	10
Pop	72	54	71	67	56	54	77	65	63	77
Estimate Pop	70.133	52.0448	72.824	67.0233	53.4065	54.8885	75.4024	69.2887	52.349	70.6472
Absolute Residual	1.8702	1.9552	1.824	0.0233	2.5936	0.8885	1.5976	4.2885	10.651	6.3539

The residuals within these 10 sample are between 0.0233 and 10.651.

- Prediction

To test the prediction power of the model, we find additional 5 songs out of the original sample. It is the result of the prediction:

Sample	1	2	3	4	5
Pop	63	34	93	89	74
Predicted Pop	75.2154	77.4531	76.75028	77.7468	74.5026
95 %Prediction Interval (Upper)	91.462	94.0085	93.0462	94.0145	90.7495
95% Prediction Interval (Lower)	58.9687	60.8977	60.4544	61.4792	58.2556
Absolute Residual	12.2154	43.4531	16.2497	11.2532	0.5026

The absolute residual is 0.5026 – 43.531. 1 out of 5 prediction result is out of 95% prediction interval. The prediction power is not satisfied. The reason is the assumption of homoscedasticity is violated; it affects the power of the regression model.

3. Justification about model

- Did you use any variable selection method?

  Yes, we used the “Forward Selection Approach". This approach helped us reduce our model from 28 predictors down to 14. Out of the four approaches we learned this term, the forward regression approach gave us the model that made more sense to us. Although the Stepwise Regression approach had a very similar result, we decided to go with the Forward Regression model to diminish accumulation of p-values (Type 1 Error). Here is a table comparing the accuracy of the forward selection, backward elimination, and the stepwise approach:

	Backward Regression	Forward Regression	Stepwise Regression
Adj. R Squared*	55.7%	55%	54.7%
F Statistic*	41.827	55.63	64.23
MSE	79	79	80
C.V.*	13.9%	13.9%	13.9%
Additional Notes	This model gave us too many variables.	This model gave us the variables that we thought made more sense to predict the best popularity score while diminishing the accumulation of p-value.	This model was also good, but we decided to go with the forward selection model instead.

* For a referenced definition of these terms, please see the appendix.

- Have you conducted F-test, t-test, or other tests?

  Yes. After conducting a stepwise selection for all our predictors, we conducted an F-test to make sure our stepwise model was more significant:

  Global F-Test

  $$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$$
  $$H_a: At\ least\ one\ is\ non-zero$$

  α = 0.05 Critical Value = 1.57 F-value = 52.17

  Since the Critical Value < F-value, we reject Ho. At least one of the predictors is linearly associated with the response at α = 0.05. The overall model is statistically useful for predicting the response variable.

  After our global F-test passed, we conducted a partial F-test to eliminate unnecessary interaction terms; and we conducted t-tests (as seen in a previous section) to fine tune our model by not including those terms which did not proved to be significant.

  Partial F-Test:

  $$H_0: \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = 0$$
  $$H_a: At\ least\ one\ is\ non-zero$$

  α = 0.05 P-value = 0.0002566

  Since the P-value < 0.05, we reject Ho. We can conclude at least one of the interaction coefficients significantly contributes to the prediction of the popularity score α = 0.05. These terms can be included in the regression model.

- Have you done any variable transformation? Why?

  Variable transformation helps align data points closer to an ideal regression line. This makes linear regression calculations possible where each variable meets all assumptions. As seen above, our variables already fit the regression line as best as they can. We tried a number of transformations for our dependent and independent variables, but these did not make practical sense for our model.

- Did you check the multicollinearity?

  According to our textbook, “multicollinearity exists when two or more of the independent variables used in regression are moderately or highly correlated.” When wanting to know the relationships between the variables, multicollinearity can lead to incoherent and deceiving results in the summary output of the model.

  Our reduced model showed to have low VIF scores, and no correlation greater than |0.9|. Our significant coefficients do not have any multicollinearity among themselves, as the charts below will prove:

  Independent Variable Correlations:

	acous	dur	nrgy	dB	hip.hop	jazz	rock	folk
acous	1.00000000	-0.20724485	-0.66331087	-0.50870230	-0.11645584	0.14437536	-0.05058879	0.08366662
dur	-0.20724485	1.00000000	0.11701905	0.02784518	0.08315143	0.11733161	0.17981957	-0.03068391
nrgy	-0.66331087	0.11701905	1.00000000	0.70466234	0.07431151	-0.11725919	0.04772528	-0.07083964
dB	-0.50870230	0.02784518	0.70466234	1.00000000	0.19086255	-0.15469993	-0.08139348	-0.07848772
hip.hop	-0.11645584	0.08315143	0.07431151	0.19086255	1.00000000	-0.02401360	-0.14860691	-0.03109801
jazz	0.14437536	0.11733161	-0.11725919	-0.15469993	-0.02401360	1.00000000	-0.05756695	-0.01204666
rock	-0.05058879	0.17981957	0.04772528	-0.08139348	-0.14860691	-0.05756695	1.00000000	-0.07455013
folk	0.08366662	-0.03068391	-0.07083964	-0.07848772	-0.03109801	-0.01204666	-0.07455013	1.00000000

  VIF Scores for Independent variables:

acous	dur	nrgy	X2010s	X1960s	X2000s	X1980s
2.402335	1.594181	3.390453	4.158146	2.692609	4.567080	3.961099
X1970s	X1990s	hip.hop	jazz	rock	folk	dB
3.493333	3.899757	36.559395	1.091697	11.019817	1.032897	2.712485
nrgy:rock	X1990s:rock	X2000s:rock	X2010s:hip.hop	dur:hip.hop		
11.558641	1.342123	1.215382	2.599473	30.636128		

Note: a general rule of thumb when analyzing VIF scores, is that a score more than 5 means that there is a multicollinearity problem with that respective variable.

4. Conclusion & Limitations:

The analyses were done to find out the statistically significant prediction model of the Spotify songs from 1950s - 2010s and level of contribution of each predictors including the song style (Latin, Blues, Country, Jazz, Other styles, Folk, Soul, Hip-hop, Electronic, Mature, Rock, Pop) the overall tempo, the level of energy in a song, the dance score, the loudness score, the live score, the positive mood score, the song length, the acoustic score, the spoken word in a song, the count of words in a song’s name, and the year a song made to the popularity score.

For descriptive analysis, we found out that the distribution of popularity score is left-skewed and includes outliers, and the distribution of energy is also left-skewed but has no outlier. Meanwhile,

the year variable, which is divided into 6 dummy variables with the base level is 1950s, has the proportion around 13 -15% in all dummy variables except for 1950s.

For inferential analysis, we have the fitted model with predictors including the level of energy in a song, and the year a song made (from 1960s to 2010s), the acoustic score, the song length, the song style (Hip hop, Jazz, Rock, Folk), the loudness score, and other interaction terms) by using forward method of variable screening and our sense. By using histogram and normal probability plot of residuals, we found out a few outliers which make the distribution of residuals left-skewed and not normally distributed, then detect all outliers whose standardized residuals higher than 3 or smaller than -3 by using scatter plot of standardized residual. We conducted a t-test for the important variables including the level of energy in a song, and the year a song was made (from 1960s to 2010s). We also conducted an F-test and concluded that our stepwise model is more significant. Although we tried a number of transformations for both dependent and independent variables, we will not use transformation for our model because these transformations do not provide better results. Regarding multicollinearity, our reduced model showed low VIF scores between variables, leading to the conclusion that there is no multicollinearity among themselves.

Limitations:

1. Prediction power is limited as the assumption of homoscedasticity is violated, especially the confident interval and the prediction interval from this model.
2. This model is limited to explain the popularity of English Song, as the dataset is English song.
3. Selection bias is occurred, the users' behavior from other song streaming platform, such as Apple Music, may be totally different.
4. The genre of the songs is classified by manual misclassification may be existed. It makes the predication power of the model is limited.

Other References:

DANA 4810 Textbook: Mendenhall, William and Sincich, Terry, A second course in statistics: Regression Analysis – 7th ed., Prentice Hall (Pearson Publishing)

Appendix

	Definition
Adj. R Squared	"The adjusted R-squared is a modified version of R-squared, which accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. " Source: "Adjusted R-Squared - Overview, How It Works, Example". 2020. *Corporate Finance Institute*. Accessed April 8, 2020. https://corporatefinanceinstitute.com/resources/knowledge/other/adjusted-r-squared/.
F Statistic	"An F statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different [or] if a group of variables are jointly significant." Source: "F Statistic / F Value: Definition and How to Run An F-Test". 2020. Statistics How To. Accessed April 8, 2020. https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/.
C.V.	"The coefficient of variation (relative standard deviation) is a statistical measure of the dispersion of data points around the mean. The metric is commonly used to compare the data dispersion between distinct series of data." For a model to be considered good, according to the CV, it must be at less than 10%. Source: "Coefficient Of Variation - Definition, Formula, And Example". 2020. *Corporate Finance Institute*. Accessed April 8 2020. https://corporatefinanceinstitute.com/resources/knowledge/other/coefficient-of-variation/.

Appendix – R & SAS Code:

R code:

```
spotify <- read.delim("C:/Users/dvazq/OneDrive/School/Stats Data
Science/4810/Project/Dataset/spotify.txt")
View(spotify)

library(olsrr)
reg_spotify <- lm(pop ~ latin + blues + country + jazz + other + folk + soul + hip.hop + electronic +
                  mature + rock + bpm + nrgy + dnce + dB + live + val + dur + acous + spch +
                  feat + title_word + X1960s + X1970s + X1980s + X1990s + X2000s + X2010s, data = spotify)
#Variable Selection
stepwise_spotify <- ols_step_both_p(reg_spotify,pent = .1, prem = .3, details = TRUE)
stepwise_spotify

ols_step_backward_p(reg_spotify,details = TRUE, prem=.3)
ols_step_forward_p(reg_spotify,details = TRUE, pent=.1)



summary(reg_spotify)
stepwise_spotify

#recomended Model for 1st degree
spotify_recommended <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
                          X2000s + X1980s + X1970s + X1960s +
                          X1990s + hip.hop + jazz + rock + folk + dB, data = spotify)

summary(spotify_recommended)
plot(spotify_recommended)

abline(0,0)
plot(spotify_recommended$fitted.values,spotify_recommended$residuals)
hist(spotify_recommended$residuals)
boxplot(spotify_recommended$residuals)
barplot(spotify_recommended$residuals)
```

```
#Multiculliniarity Testing

only_num_var <- spotify[,c('acous' , 'dur' , 'nrgy', 'dB','hip.hop' , 'jazz' , 'rock' , 'folk')]
library(car)
summary(spotify_recommended)
vif(spotify_recommended)
cor(only_num_var)

#Model with Interactions that we think make sense due to year/music genre popularity
spotify_w_interaction <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
                        X2000s + X1980s + X1970s + X1960s +
                        X1990s + hip.hop + jazz + rock + folk + dB +
                        rock:nrgy + rock:X1960s + rock:X1970s + rock:+X1980s +
                        rock:X1990s +rock:X2000s + hip.hop:X2010s + hip.hop:X2000s +
                        jazz:X1960s + jazz:X1980s +  folk:X1960s + folk:X1970s +
                        folk:X1990s,data = spotify)
summary(spotify_w_interaction

#Final Improved Model (removed useless interaction terms)
spotify_recommended_w_interaction <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
                                X2000s + X1980s + X1970s +
                                X1990s + hip.hop + jazz + rock + folk + dB +
                                rock:nrgy + rock:X1990s + rock:X2000s +
                                hip.hop:X2010s + dur:hip.hop, data = spotify)
summary(spotify_recommended_w_interaction)
library(car)
summary(spotify_recommended)
vif(spotify_recommended_w_interaction)
cor(only_num_var)

#PARTIAL F-Tests:
spotify_reduced <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
                    X2000s + X1980s + X1970s +
                    X1990s + hip.hop + jazz + rock + folk + dB, data = spotify)
```

```
anova(spotify_recommended_w_interaction,spotify_reduced)

#Recommended model for second degree
spotify_int_stepwise<- ols_step_both_p(comp_sec_spotify,pent = .05, prem = .05, details = TRUE)
spotify_int_stepwise

spotify_int_recommended <- lm(pop ~ X1980s + X2000s + X1970s + X1990s + X1960s +
                        rock + log(dur) + acous:X2010s + X2010s:log(dur) + log(dur):nrgy +
                        nrgy:hip.hop + acous:log(dur) + X1970s:log(dur) +
                        X1990s:rock + nrgy:jazz +
                        rock:nrgy , data = df)

#Partial Residual Plots
termplot(spotify_recommended, partial.resid = TRUE)
termplot(spotify_int_recommended, partial.resid = TRUE)

#Residual Plots
abline(0,0)
#Pass:
plot(df$acous, resid(spotify_int_recommended), ylab = "Residuals", xlab = "Acous", main = "Residual
Plot")
plot(df$nrgy, resid(spotify_int_recommended), ylab = "Residuals", xlab = "Energy", main = "Residual
Plot")
plot(df$X2010s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X2010s", main = "Residual
Plot")
plot(df$X1960s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X1960s", main = "Residual
Plot")
plot(df$X2000s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X2000s", main = "Residual
Plot")
plot(df$X2010s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X2010s", main = "Residual
Plot")
plot(df$X1980s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X1980s", main = "Residual
Plot")
plot(df$X1970s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X1970s", main = "Residual
Plot")
plot(df$X1960s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X1960s", main = "Residual
Plot")
```

```
plot(df$X1990s, resid(spotify_int_recommended), ylab = "Residuals", xlab = "X1990s", main = "Residual Plot")
plot(df$jazz, resid(spotify_int_recommended), ylab = "Residuals", xlab = "jazz", main = "Residual Plot")
plot(df$hip.hop, resid(spotify_int_recommended), ylab = "Residuals", xlab = "hip.hop", main = "Residual Plot")
plot(df$rock, resid(spotify_int_recommended), ylab = "Residuals", xlab = "Rock", main = "Residual Plot")

####Normality
lm1 <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
    X2000s + X1980s + X1970s + X1960s + X1990s + hip.hop + jazz + rock + folk + dB +
    rock:nrgy + rock:X1990s + rock:X2000s + hip.hop:X2010s + dur:hip.hop, data = df)
par(mfrow=c(2,1))
hist(lm1$residuals,main = "Histogram of Residual", xlab = "Residual")
qqnorm(lm1$residuals, main = "Normal Probability Plot of Residual")
qqline(lm1$residuals)

par(mfrow=c(1,1))
plot(lm1$fitted.values,lm1$residuals/sd(lm1$residuals), main = "Scatterplot of Fitted Values and
Standardized Residuals",
    xlab = "Fitted Values", ylab = "Standardized Residuals")
abline(h=c(-3,3))

df2 <-df[abs(lm1$residuals/sd(lm1$residuals)) <= 3,]

lm2 <- lm(pop ~ acous + dur + nrgy + X2010s  + X1960s +
          X2000s + X1980s + X1970s + X1960s + X1990s + hip.hop + jazz + rock + folk + dB +
          rock:nrgy + rock:X1990s + rock:X2000s + hip.hop:X2010s + dur:hip.hop, data = df2)

par(mfrow=c(2,1))
hist(lm2$residuals,main = "Histogram of Residual(w/out Outliers)", xlab = "Residual")
qqnorm(lm2$residuals, main = "Normal Probability Plot of Residual(w/out Outliers)")
qqline(lm2$residuals)

# Condidence Interval
set.seed(1234)
est <- df2[sample(nrow(df2),10),]
result <- predict(lm2, newdata = est)
```

```
est['pop'] - result

# Prediction
predict(lm2, newdata = pred) - pred['pop']
predict(lm2, newdata = pred, interval = "prediction")
```

SAS Code: