**Project Tite**: What Attributes of Music Impact the Current Popularity?

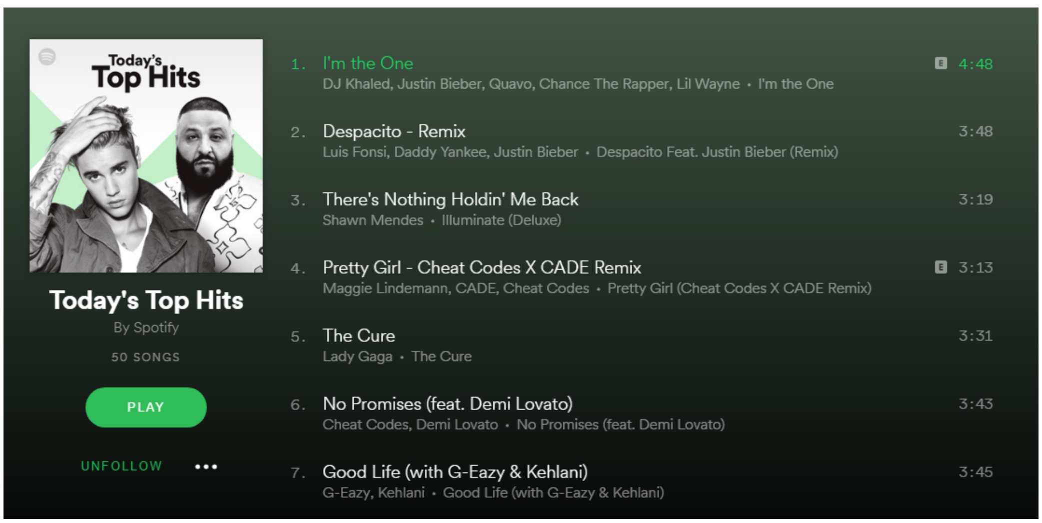
**Course and Section Number**: STAT 3220 Section 002

**Team Name**: Section 002 - Gooch

**Group Members**: Elit Dogu, Sam Switz, Brian Wimmer

**Due Date**: Sunday October 18, 2020

**Photo Representation**:




**Introduction and Research Question:**

Our main research question is: What attributes of music impact the current popularity, and how has the importance of different attributes changed over time? The music industry has been rapidly growing in the past decade with the introduction of streaming services, leading to more money to be made by labels and artists alike. Although music has been a huge part of the lives of everyone for decades, streaming services have made data more easily accessible to be utilized by those in the music industry. By examining this research question, we can look at how labels and artists may be changing their sound to increase their popularity. Similarly, there have been a lot of new platforms, like Tik Tok, that have increased the popularity of certain songs drastically. By analyzing our data by decade, we will be able to see if there were any sudden peaks of certain music trends/characteristics. By doing this analysis, we will be able to come up with a model that predicts the factors of a song that allow it to become popular. Music companies, artists, and others involved in the music industry may find these analyses helpful in determining how music tastes are changing over time and how they can stay ahead of the trends.

Specifically, we are interested in if music with a greater valence (positivity/happiness) has a higher current popularity, if there are certain trends that can influence music's current popularity (and does this change depending on the year(s)), and does the duration of a song have an impact on its current popularity? We likely will be observing more sub questions than these four to get a better understanding of the data and a better understanding of what impacts music's current popularity.

We found an article that helped us in deciding what we wanted our research question to be: *How Has the Internet Changed the Music Industry in the 2010s?* The article supports our claim that music trends have changed throughout the last decade; for example, the article goes into detail about how hip hop has become the new version of "pop" again, and how rock is dead. The article also looks at platforms like Tik Tok to see its impact on music trends. It discusses certain factors of music that can lead to its popularity increasing, which is a critical question of our research for project 1. We could use this article in addition to the analysis we conduct on the Spotify.csv file to have a more comprehensive understanding of music popularity trends.

Link to the article: https://medium.com/swlh/pop-music-in-2019-year-in-review-1a0c7fe89f31

In Stage 1 of our analysis, we examined the effect of three of our "most important" quantitative variables on the popularity of a song. Acousticness, energy, and danceability were chosen. In Stage 2, we continued with our original model from Stage 1, but added an interaction variable. Through our process, we found that our interaction variable (between key and explicit) was insignificant and should be removed. Finally, in Stage 3, we examined the interactions between explicit and both danceability and energy through a global f test and individual t tests. Both of these were significant and added into the model. This gave us our finally model for determining the current popularity of a song:

y(*popularity*) = $\beta_0$ + $\beta_1$(*acousticness*) + $\beta_2$(*danceability*) + $\beta_3$(*energy*) + $\beta_4$(*duration_ms*) + $\beta_5$(*instrumentalness*) + $\beta_6$(*explicit*) + $\beta_7$(*key*) + $\beta_7$(*exp_dance*) + $\beta_7$(*exp_eng*) + epsilon

## Data Summary

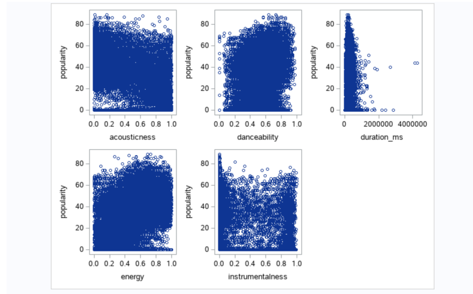
The spotify.csv dataset was originally created as "independent research" into Spotify's Web API by Yamaç Eren Ay, and was last updated as early as three months ago. The dataset includes a staggering 160,000 songs, spanning from 1921 to 2020. Each observation is a different song ID, which includes information about the name of the song, artist, release date, and many other musical identifiers - ranging from danceability, to tempo, to energy, and more.

We did not have any NA's in our dataset. We verified this by using a Google Colab Notebook in Python to look at the dataframe. We ran the "code df.isnull().any()", which showed "False" for all of our variables, meaning that we did not have any NA in our dataset. We will be subsetting the data by decade to look at how trends change over time. We found that the earliest year in our dataset was 1921, and the latest was 2020. We ended up having a total of 10 subsetted data frames once we broke down the original dataset into decades. This data manipulation was all done using Python. The subsetted data (by decade) can be found in the link below.

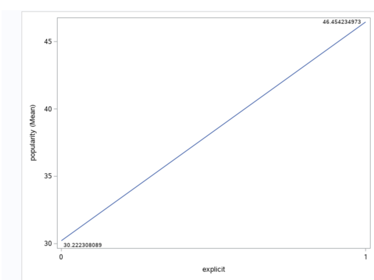
The original spotify.csv dataset can be found at: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks
The subsetted dataset for the breakdown of each decade can be found here: https://www.kaggle.com/elitdogu/spotify-data-broken-down-by-decade

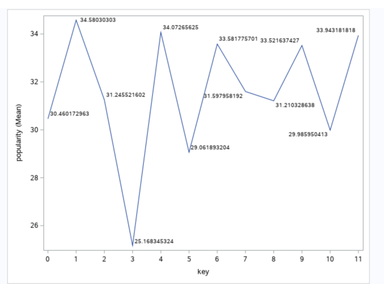
| Variable | Description | Quantitative/Qualitative | Units/Levels |
|---|---|---|---|
| acousticness | a measure of how acoustic the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 |
| danceability | a measure of how danceable the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 |
| energy | a measure of how energetic the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 |
| duration_ms | a measure of how long the song is in milliseconds | Quantitative | milliseconds, integer number that usually ranges from 200k to 300k |
| instrumentalness | a measure of how instrumental the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 |
| | | | |
| **Dummy Variables:** | | | |
| mode | measures if the song begins with a major chord progression | Qualitative | 0 = minor, 1 = major |
| explicit | measures if the song contains explicit content | Qualitative | 0 = non-explicit, 1 = explicit |
| key | measures the primary key of the track | Qualitative | measurement ranging from 0-11 |

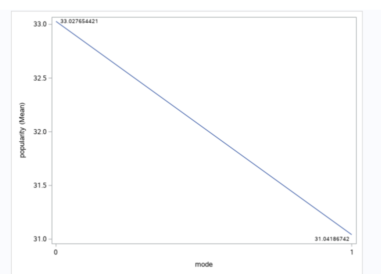
## Exploratory Data Analysis




Acousticness had a weak negative correlation between acousticness and popularity. Danceability and Energy had a weak positive correlation. Duration was hard to tell because of the outliers, we might see more of a relationship if they're taken out. Similarly, instrumentalness did not show much of a relationship; overall none of the predictors seem to have a strong correlation with popularity, although the sheer size of the data set make the graphs hard to decipher




On the left graph, we can see the popularity for songs based on if they are explicit (shown as 1) or non-explicit (shown as 0). We can see that explicit songs have a significantly higher mean popularity. As for the middle graph, we see the mean popularities for each key, with the main outlier being D# having a mean popularity of just 25.17, whereas all the other keys are around 31 for mean popularity. Lastly on the right graph, we have the mean popularities for the two mode dummy variables. 1 signals that the song starts in a major chord progression; 0 signals that the song starts in a minor chord progression. A mode of 1 has a lower mean popularity than a mode of 0 by about 2 ranking units.

**The CORR Procedure**

| | |
|---|---|
| **1 With Variables:** | popularity |
| **5 Variables:** | acousticness danceability duration_ms energy instrumentalness |

| Pearson Correlation Coefficients, N = 16992 | | | | | |
|---|---|---|---|---|---|
| | **acousticness** | **danceability** | **duration_ms** | **energy** | **instrumentalness** |
| **popularity** | -0.60069 | 0.21924 | 0.07759 | 0.50469 | -0.29353 |

Acousticness has the strongest correlation of the bunch, at -0.60069. Energy has the second strongest at 0.50469, then instrumentalness at -0.029353. The weakest two are danceability and duration, at 0.21924 and 0.07759, respectively.

## Analysis

For our analysis, we will be using an alpha level of 0.05 throughout. We determined through our EDA that the variables of interest that we believe to influence the current popularity of a song the most are acousticness, danceability, and energy. The analysis of these variables will be conducted through 3 stages, using Global F Tests, individual t tests, and nested F tests.

### Stage 1

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: popularity**

| | |
|---|---|
| Number of Observations Read | 169909 |
| Number of Observations Used | 169909 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 29443803 | 5888761 | 20130.8 | <.0001 |
| Error | 169903 | 49700910 | 292.52521 | | |
| Corrected Total | 169908 | 79144713 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 17.10337 | R-Square | 0.3720 |
| Dependent Mean | 31.55661 | Adj R-Sq | 0.3720 |
| Coeff Var | 54.19900 | | |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 36.37747 | 0.26459 | 137.49 | <.0001 |
| acousticness | 1 | -26.30961 | 0.17206 | -152.91 | <.0001 |
| danceability | 1 | 6.24966 | 0.25373 | 24.63 | <.0001 |
| duration_ms | 1 | 0.00000683 | 3.491317E-7 | 19.56 | <.0001 |
| energy | 1 | 8.94934 | 0.23524 | 38.04 | <.0001 |
| instrumentalness | 1 | -7.16681 | 0.14645 | -48.94 | <.0001 |

The three quantitative predictors that we deemed the most important through the EDA were acousticness, danceability, and energy. Danceability and energy both had positive linear correlations and acousticness had a negative linear correlation. We will be using danceability as our variable of interest, as it corresponds with our sub question regarding the characteristics of most popular songs by decade. Our initial model was stated as follows:

$$y(\textit{popularity}) = \beta_0 + \beta_1(\textit{acousticness}) + \beta_2(\textit{danceability}) + \beta_3(\textit{energy}) + \beta_4(\textit{duration\_ms}) + \beta_5(\textit{instrumentalness}) + \text{epsilon}$$

The global F-test for our model resulted in an F-value of 20130.8, which gives us a P-value of <0.0001. We then deem our model to be significant at our alpha level of 0.05, meaning that at least one of the betas is significantly different from zero. We also noted a MSE of 17.10337 and an R-Squared value of 0.372. We then conducted a Nested F test on danceability, in which we were able to reject the null hypothesis after finding an F-Value of 606.67 and a p-value of < 0.0001. This Nested F test also allowed us to determine that we did not need to remove any variables, as danceability was significant. Our original model will stay the same:

$$y(\textit{popularity}) = \beta_0 + \beta_1(\textit{acousticness}) + \beta_2(\textit{danceability}) + \beta_3(\textit{energy}) + \beta_4(\textit{duration\_ms}) + \beta_5(\textit{instrumentalness}) + \text{epsilon}$$

| Test 1 Results for Dependent Variable popularity | | | | |
|---|---|---|---|---|
| Source | DF | Mean Square | F Value | Pr > F |
| Numerator | 1 | 177467 | 606.67 | <.0001 |
| Denominator | 169903 | 292.52521 | | |

## Stage 2

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 8 | 29733483 | 3716685 | 12779.8 | <.0001 |
| Error | 169900 | 49411230 | 290.82537 | | |
| Corrected Total | 169908 | 79144713 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 17.05360 | R-Square | 0.3757 |
| Dependent Mean | 31.55661 | Adj R-Sq | 0.3757 |
| Coeff Var | 54.04130 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 35.96495 | 0.27140 | 132.52 | <.0001 |
| acousticness | 1 | -25.24396 | 0.17491 | -144.33 | <.0001 |
| danceability | 1 | 4.85361 | 0.25698 | 18.89 | <.0001 |
| duration_ms | 1 | 0.00000724 | 3.483738E-7 | 20.78 | <.0001 |
| energy | 1 | 9.58121 | 0.23546 | 40.69 | <.0001 |
| instrumentalness | 1 | -7.07408 | 0.14605 | -48.44 | <.0001 |
| explicit | 1 | 4.58132 | 0.26381 | 17.37 | <.0001 |
| key | 1 | -0.03817 | 0.01239 | -3.08 | 0.0021 |
| key_exp | 1 | 0.06538 | 0.03996 | 1.64 | 0.1018 |

For stage two, we will be adding both explicit and key as our qualitative predictors, which have been indicated as dummy variables within the data set. Of the three qualitative variables we had (including mode), explicit and key seemed to have the greatest possibility and difference when predicting the popularity of a song. We added an interaction between key and explicit to examine if the two qualitative variables are correlated. We will be using the explicit variable as our variable of interest, as it corresponds with [sub question]. We continued with our end stage 1 model, with the interactions added on:

$$y(popularity) = \beta_0 + \beta_1(acousticness) + \beta_2(danceability) + \beta_3(energy) + \beta_4(duration\_ms) + \beta_5(instrumentalness) + \beta_6(explicit) + \beta_7(key) + \beta_8(key\_exp) + \text{epsilon}$$

Our global F test for the overall model returns a F statistic of 12779.8 and a P-value of <0.0001, meaning that our model is significant. After finding that the model was in fact significant, we had to do individual t-tests to determine which variables should be included in the model and move into stage 3. We determined that the interaction variable between key and explicit was insignificant with a p-value of 0.1018 (T statistic of 1.64), which is above our cutoff of 0.05. The two qualitative variables, key and explicit, were significant on their own. Our end stage 2 model was equated as follows:

$$y(popularity) = \beta_0 + \beta_1(acousticness) + \beta_2(danceability) + \beta_3(energy) + \beta_4(duration\_ms) + \beta_5(instrumentalness) + \beta_6(explicit) + \beta_7(key) + \text{epsilon}$$

## Stage 3

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 9 | 30761166 | 3417907 | 12002.0 | <.0001 |
| **Error** | 169899 | 48383547 | 284.77829 | | |
| **Corrected Total** | 169908 | 79144713 | | | |

| | | | |
|---|---|---|---|
| **Root MSE** | 16.87538 | **R-Square** | 0.3887 |
| **Dependent Mean** | 31.55661 | **Adj R-Sq** | 0.3886 |
| **Coeff Var** | 53.47652 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | 1 | 39.55774 | 0.27501 | 143.84 | <.0001 |
| **acousticness** | 1 | -27.15835 | 0.17633 | -154.02 | <.0001 |
| **danceability** | 1 | 4.74702 | 0.26476 | 17.93 | <.0001 |
| **duration_ms** | 1 | 0.00000563 | 3.457996E-7 | 16.27 | <.0001 |
| **energy** | 1 | 5.13507 | 0.24679 | 20.81 | <.0001 |
| **instrumentalness** | 1 | -7.40115 | 0.14467 | -51.16 | <.0001 |
| **explicit** | 1 | -31.10468 | 0.84608 | -36.76 | <.0001 |
| **key** | 1 | -0.03502 | 0.01165 | -3.01 | 0.0027 |
| **exp_dance** | 1 | 17.80371 | 0.97107 | 18.33 | <.0001 |
| **exp_eng** | 1 | 38.86627 | 0.65007 | 59.79 | <.0001 |

In our final analysis stage, stage 3, we will be adding the interactions between explicit and danceability and explicit and energy to our model. Of the two qualitative variables we selected in stage 2, explicit seemed to have the more significant differences in average popularity for each value of the dummy variables than key, so we decided to test its interaction with the two quantitative variables we deemed to be most important: danceability and energy. We conducted another Global F test, which indicates that the model is significant as it returned an F value of 12002 (p-value = <0.0001). We then conducted individual T-tests for the interaction variables, which showed that both were significant to the model. From this, we were able to conclude our final model:

$$y(popularity) = \beta_0 + \beta_1(acousticness) + \beta_2(danceability) + \beta_3(energy) + \beta_4(duration\_ms) + \beta_5(instrumentalness) + \beta_6(explicit) + \beta_7(key) + \beta_7(exp\_dance) + \beta_7(exp\_eng) + \text{epsilon}$$

### Summary of the Model Building

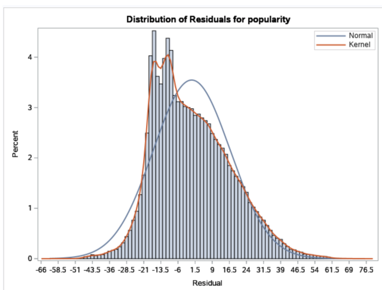
| | Predictor Count | Overall utility - global f test | adj-R Square | Root MSE |
|---|---|---|---|---|
| Stage 1 | 4 | F-value: 20130.8<br>P-value: <0.0001 | 0.3720 | 17.10337 |
| Stage 2 | 7 | F-value: 12779.8<br>P-value: <0.0001 | 0.3757 | 17.05360 |
| Stage 3 | 8 | F-value: 12002.0<br>P-value: <0.0001 | 0.3886 | 16.87538 |

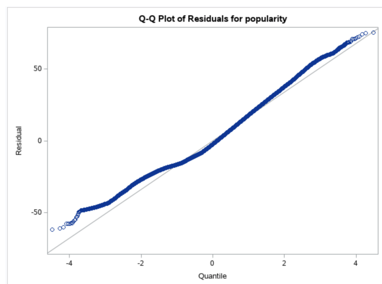
# Evaluating the Model

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: popularity**

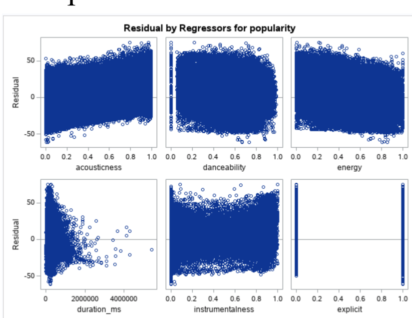
| | |
|---|---|
| **Durbin-Watson D** | 0.794 |
| **Pr < DW** | <.0001 |
| **Pr > DW** | 1.0000 |
| **Number of Observations** | 169909 |
| **1st Order Autocorrelation** | 0.603 |

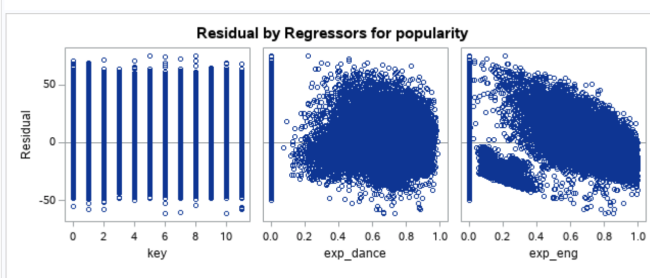
**Autocorrelation:** We detected autocorrelation in our model through the Durbin Watson test. In order to fix this we would fit an autocorrelation model by adding extra parameters to the model. This also gave us information about the error correlation - noting that the errors in our model are uncorrelated.




**Multivariate Normality:** When examining the normal distribution compared to the actual distribution of the graph, we can note that the graph seems to follow the normal distribution curve fairly well. The histogram and the QQ plot both seem to follow the curve and line well, therefore we would not make any changes based solely off of this and would need to look at the other assumptions.

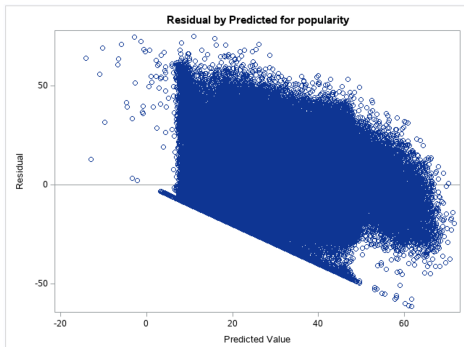


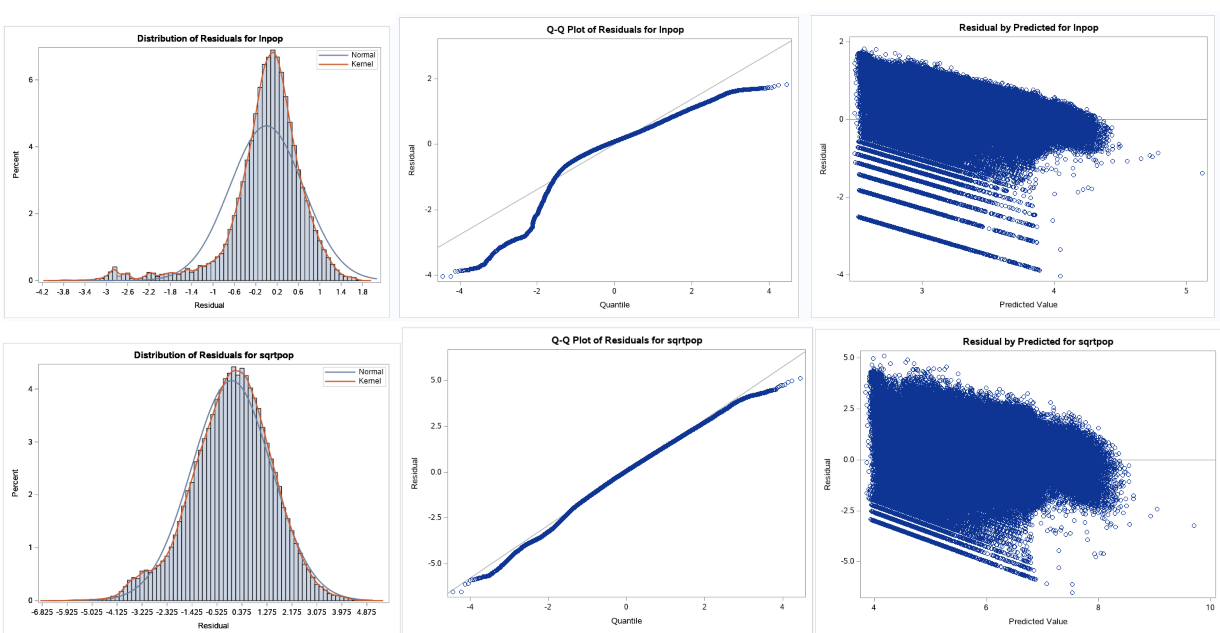

No assumptions were violated.

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | 1 | 35.92897 | 0.27051 | 132.82 | <.0001 | 0 |
| acousticness | 1 | -25.24039 | 0.17489 | -144.32 | <.0001 | 2.53484 |
| danceability | 1 | 4.85090 | 0.25698 | 18.88 | <.0001 | 1.18618 |
| energy | 1 | 9.58545 | 0.23544 | 40.71 | <.0001 | 2.31545 |
| duration_ms | 1 | 0.00000724 | 3.483678E-7 | 20.79 | <.0001 | 1.04359 |
| instrumentalness | 1 | -7.07386 | 0.14605 | -48.43 | <.0001 | 1.19244 |
| explicit | 1 | 4.92830 | 0.15693 | 31.41 | <.0001 | 1.11731 |
| key | 1 | -0.03189 | 0.01178 | -2.71 | 0.0068 | 1.00119 |

**Multicollinearity:** According to our Variance Inflation analysis, no multicollinearity was detected. All VIF values were between 1 and 3, which is far less than the 10 than indicates extreme multicollinearity.




**Homoscedasticity:** There are patterns in some of the residual graphs that indicate non-constant variance, so we transformed the model in an attempt to correct this. We performed log and square root transformations on our model, lnpop and sqrtpop, respectively. The transformations did not seem to have a noticeable effect on the variance of the residuals.



**Outliers/Influential Observations:** We were unable to completely analyze our model for any outliers or influential observations as the data set is incredibly large and sas could not handle it. We were able to observe about 20,000 outliers initially with respect leverage, but we couldn't remove them and test the model without them.

## Conclusion

Our final prediction equation after doing the analyses is:
y(*popularity*) = $\beta_0$ + $\beta_1$(*acousticness*) + $\beta_2$(*danceability*) + $\beta_3$(*energy*) + $\beta_4$(*duration_ms*) + $\beta_5$(*instrumentalness*) + $\beta_6$(*explicit*) + $\beta_7$(*key*) + $\beta_7$(*exp_dance*) + $\beta_7$(*exp_eng*) + epsilon

Our final model with beta estimates included is:
y(*popularity*) = 39.55774 - 27.15835(*acousticness*) + *4.74702*(*danceability*) + 5.13507(*energy*) + 0.00000563(*duration_ms*) - 7.40115(*instrumentalness*) - 31.10468(*explicit*) - 0.03502(*key*) + 17.80371(*exp_dance*) + 38.86627(*exp_eng*) + epsilon

Each of these beta estimates that correspond with their individual variables show us how the popularity of a song can change, depending on its certain factors. For example, for every unit that the popularity of a song increases, the key of that song would decrease by 0.03502. Overall, we cannot say that our model is "good" at predicting the current popularity of a song. From a statistical standpoint, the Root MSE is fairly high at 17.05 considering that popularity is rated on a scale of 0 to 100. Additionally, the adjusted R-squared is only 0.3756, which means that just 38% of the variation in song popularity is explained by our model. Our model only attempts to predict current popularity for a given song, but we don't expect it to predict with great accuracy. We tried to correct these problems throughout the model-building process by adding interaction and higher-order terms; however, they did not improve the model. The key limitation of the data we worked on was that it was only focused on the songs themselves. That posed two problems: one, the data were likely to be correlated with each other and two, we weren't able to account for outside cultural influences that can greatly affect how many streams a song gets. For a prediction, we will use the audio features for the song "Laugh Now Cry Later" by Drake. We calculated the predicted value by and and got a popularity of 8 for the song, which seems very inaccurate given that the song has been near the top of the U.S. charts for the past two months. One reason for the large error could be that we got the data for this particular song directly from Spotify as opposed to the data set we built the model with because the song wasn't released when the data set was created.

We believe that our model could definitely be improved using other metrics - or that other questions can be answered using our model and other variables. Something that would be interesting to look at in the future is the impact of economic and social media metrics within the music industry. Musical metrics - such as energy and danceability - are not the only things that could result in the popularity of a song. We believe that the label, budget, amount spent on promotion, number of TikTok shares, and number of Instagram story adds are just a few of the economic/social media metrics that could influence the popularity of a song. It would be highly interesting and beneficial to do an analysis on these metrics in the future.

## SAS Code

```
*Importing the csv file with the sampled song data, sampling done in python;
proc import datafile='/folders/myfolders/data/spot_samps.csv'
out=work.spotsamps
dbms=csv
;
run;

*Creating scatter plots to show the relationship of all the quantitative variables;
*with popularity;
proc sgscatter data=mydata.spotsamps;
    plot popularity*(acousticness danceability duration_ms energy instrumentalness);
run;

proc corr data=mydata.spotsamps pearson nosimple noprob plots=none;
    var acousticness danceability duration_ms energy instrumentalness;
    with popularity;
run;

*Mean popularity for each explicit dummy variable;
proc sgplot data=mydata.spotsamps;
    vline explicit / response=popularity datalabel stat=mean;
run;

*Mean popularity for each key dummy variable;
proc sgplot data=mydata.spotsamps;
    vline key / response=popularity datalabel stat=mean;
run;

*Mean popularity for each mode dummy variable;
proc sgplot data=mydata.spotsamps;
    vline mode / response=popularity datalabel stat=mean;
run;
```

```
*### Stage 1;
proc reg data=mydata.spotify plots=none;
model popularity = acousticness danceability duration_ms energy instrumentalness;
run;
*nested F test;
proc reg data=mydata.spotify plots=none;
model popularity = acousticness danceability duration_ms energy instrumentalness;
test danceability;
run;

*### Stage 2;
*Adding interaction term;
data mydata.spotify2;
set mydata.spotify;
key_exp = key*explicit;
run;
*Testing model with new predictor;
proc reg data=mydata.spotify2 plots=none;
model popularity = acousticness danceability duration_ms energy instrumentalness explicit key key_exp;
run;

*### Stage 3;
*Adding two new interaction terms;
data mydata.spotify3;
set mydata.spotify;
exp_dance = explicit*danceability;
exp_eng = explicit*energy;
run;
*Testing model with two new predictors;
proc reg data=mydata.spotify3 plots=none;
model popularity = acousticness danceability duration_ms energy instrumentalness explicit key exp_dance exp_eng;
run;
```

```
*### Evaluation;
proc reg data=mydata.spotify3 plots(maxpoints=170000)
plots(only)=(residualbypredicted residualplot qqplot residualhistogram);
model popularity = acousticness danceability duration_ms energy instrumentalness explicit key exp_dance exp_eng / dwprob;
run;
*Testing Variance Inflation Factors;
proc reg data=mydata.spotify3 plots=none;
model popularity = acousticness danceability duration_ms energy instrumentalness explicit key / vif;
run;

*Transforming response to solve homoscedasticity problem;
data mydata.spotify4;
set mydata.spotify3;
lnpop=log(popularity);
sqrtpop=sqrt(popularity);
invpop=1/popularity;
run;

*Removing 0 popularity songs so transformed models can work;
data mydata.spotify5;
set mydata.spotify4;
if popularity = 0 then delete;
run;

*Testing log transformation model;
proc reg data=mydata.spotify5 plots(maxpoints=170000)=(residualbypredicted residualplot qqplot residualhistogram);
model lnpop = acousticness danceability duration_ms energy instrumentalness explicit key exp_dance exp_eng;
run;
*Testing square root transformation model;
proc reg data=mydata.spotify5 plots(maxpoints=170000)=(residualbypredicted residualplot qqplot residualhistogram);
model sqrtpop = acousticness danceability duration_ms energy instrumentalness explicit key exp_dance exp_eng;
run;
```

## Appendix

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | **Variable** | **Description** | **Quantitative/Qualitative** | **Units/Levels** | |
| 2 | acousticness | a measure of how acoustic the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 3 | danceability | a measure of how danceable the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 4 | energy | a measure of how energetic the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 5 | duration_ms | a measure of how long the song is in milliseconds | Quantitative | milliseconds, integer number that usually ranges from 200k to 300k | |
| 6 | instrumentalness | a measure of how instrumental the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 7 | | | | | |
| 8 | **Dummy Variables:** | | | | |
| 9 | mode | measures if the song begins with a major chord progression | Qualitative | 0 = minor, 1 = major | |
| 10 | explicit | measures if the song contains explicit content | Qualitative | 0 = non-explicit, 1 = explicit | |
| 11 | key | measures the primary key of the track | Qualitative | measurement ranging from 0-11 | |
| 12 | | | | | |
| 13 | | | | | |
| 14 | | | | | |
| 15 | valence | a measure of the valence of the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 16 | popularity | a mesure of how popular the song is today from 0 to 100 | Quantitative | measurement ranging from 0-100 | |
| 17 | tempo | the tempo of the song in bpm | Quantitative | float, typically ranges from 50-150 | |
| 18 | liveness | a measure of how lively the song is from 0 to 1 | Quantitative | measurement ranging from 0-1 | |
| 19 | loudness | Relative loudness of the track in the typical range [-60, 0] in decib | Quantitative | decibels | |
| 20 | speechiness | a measure of the speechiness of the song is from 0 to 1 | Quantitative | measurement from 0 to 1 | |
| 21 | year | year the song was released | Quantitative | year | |
| 22 | | | | | |
| 23 | artists | all the contributing artists | Qualitative | artists names | |
| 24 | release_date | specific date the song was released | Qualitative | typically listed in yyyy-mm-dd format | |
| 25 | name | name of the song | Qualitative | name | |
| 26 | id | unique id for the song | Qualitative | unique id code generated by Spotify | |
| 27 | | | | | |

Full data dictionary, including variables that we may have to use for the final project submission.

The Data Dictionary can be found here: https://docs.google.com/spreadsheets/d/1wgJS-aqyJgZW5PgSI6tBxaiE29cMQDYpUWmrGc2PYDE/edit?usp=sharing

**Works Cited**

Bloom, M. (2020, August 14). Drake and Lil Durk Share Video for New Song "Laugh Now Cry Later". Retrieved September 19, 2020, from https://pitchfork.com/news/drake-and-lil-durk-share-video-for-new-song-laugh-now-cry-later-watch/?mbid=synd_yahoo_rss

Sounds Click: Spotify Today's Top Hits 320kbps (May 1, 2017) Dir...: Top hits, Spotify, Today. (n.d.). Retrieved September 19, 2020, from https://www.pinterest.com/pin/858287641455818943/

Tauberg, M. (2020, April 18). How Has the Internet Changed Pop Music in the 2010s? Retrieved September 19, 2020, from https://medium.com/swlh/pop-music-in-2019-year-in-review- 1a0c7fe89f31