Spotify Music Analysis

Project 2 Part 3: Report - STAT 6021

Sirish Desai, Kevin Kuc, Suraj Kunthu, Matt Scheffel

Executive Summary

Music preference is a subjective topic. Everyone has their own opinion as to what they consider to be good music. Pop music tends to have certain qualities with the "aim of appealing to a general audience"[1] . What are these qualities and how can we produce a song to be more popular? This report will go in detail analyzing the Top 50 Spotify Songs of 2019 with these questions in mind. We primarily ask two questions related to what factors affect a song's popularity and compare the characteristic differences between Pop and non-Pop songs.

The data set includes information regarding the Track Name, Artist Name, Genre, a Song's Beats Per Minute, Energy level, Danceability measure, Loudness volume, Liveness measure, Valence rating, Length in seconds, Acousticness, Speechiness, and a Song's popularity score. After some minor data wrangling, we proceeded to analyze the data.

Our first question delves into what predictor variable(s) has the greatest impact on a song's popularity. We were interested in what qualities of a song engage listeners in a positive manner and figure out what was special about the popular songs. After performing some exploratory data analysis, we utilized the Automatic Model Selection search procedure to help identify only the necessary predictors. We ultimately found a linear model that would estimate popularity as a function of a song's Speechiness (presence of words in a track) and Valence (musical positiveness conveyed by the track). What we discovered was that we couldn't use more variables than Speechiness and Valence without the result we received being muddled. So, using a few methods to test outliers, we tried to see if there were any outliers in our dataset, to which we could attribute some unaccounted change in being able to understand our data. However, what's interesting is that our data showed that the more positive the track was, there was a negative effect on how Popular the song was, which could be due to more popular songs being more negative, or vice versa of having a very upbeat song being seen as cheesy and therefore people don't want to have to listen to it all the time. We also can see that the more words a song has, the more popular the song becomes. However, what is interesting is that we can't tell if the song has unique words or just repeats the chorus a few times.

Our Second question compares the different qualities of a song to determine whether the song is classified as a Pop or not. After some exploratory data analysis, we use the same wrangled data from the first question of interest. However, to do a Logistic Regression, we needed to make a variable that has only two outcomes. So, the PopCheck classifier was created to read the text string in the Genre column. If the word "Pop" appeared in that song's genre, it was classified as "Pop." If "Pop" did not appear, the song was classified as "NotPop." Then we utilized the Automatic Model Selection search procedure to help identify only the necessary

[1] https://en.wikipedia.org/wiki/Pop_music

predictors. What we found was that the best thing you could do to try and figure out if a song is Pop or not was to look at how many words the song has, which could be attributed to Pop songs having repeated lyrics. We could also guess that if you knew the song's Popularity level, you can tell if the song is going to be classified as Pop, which makes sense because if the song is going to try and be in the Top 50 songs. What we found was similar to the behavior seen in the linear regression model. This may be a factor of the song's that were released in 2019, the more popular songs tended to be non-Pop. However, if we were to look at this data again, we would try to delve more into the different Genres that appear.

Data Description

The data set was pulled by Leonardo Henrique from the Spotify music database API and is hosted on Kaggle. It is the "Top 50 Spotify Songs of 2019" [2]. The variables included in this data set are as follows:

Variable	Definition
Track Name	The name of the song.
Artist Name	The name of the artist who performs the song.
Beats Per Minute (BPM)	The tempo of a song, measured in beats per minute.
Energy	A measure from 0 to 100 that represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
Danceability	Measure of how "danceable" a song is based on tempo, rhythm stability, beat strength, and overall regularity.
Loudness	The volume of a song, measured in decibels. 0 dB indicates the normal human hearing threshold. A negative Loudness indicates a softer volume while a positive Loudness indicates a harsher volume.
Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live.
Valence	Musical positiveness conveyed by the track. Tracks with high valence sound more positive while tracks with low valence sound more negative.
Length	The duration of a song, measured in seconds.
Acousticness	A measure of how likely the song uses acoustic instruments versus synthesizers.
Speechiness	The presence of spoken words in a track. The higher the value, the more likely the song is "speech-like."
Genre	The classification of what Genre the songs would fit into. We modified this variable to make it Binary, because we noticed many songs had the classification of "pop" in some form, so we decided to create a new column from this data called Popcheck which is binary: "Pop" or "NotPop."
PopCheck	The new column we formed by seeing if a song had Pop or Not in the Genre column. This new column is Binary with the songs classified as "Pop" or "NotPop."

[2] https://www.kaggle.com/datasets/leonardopena/top50spotify2019

Certain variables such as Energy, Danceability, Liveness, Valence, Acousticness, and Speechiness are measured by Spotify's internal algorithm and given a value. We assume these variable values to be accurate for our analysis purposes.

Linear Regression

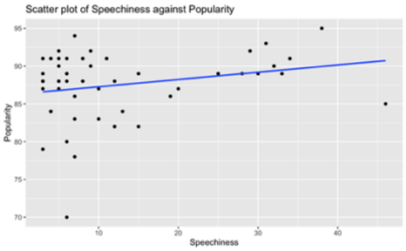
Our first question explores the quantitative relationship between variables in the "Top 50 Spotify Songs - 2019" dataset, specifically what variables affect how popular (represented by the "Popular" variable) a song is going to be. Our question delves into which variable included in the data set has the greatest impact on a song's popularity. A linear regression analysis would be of interest to artists and producers who are seeking to produce music that engages listeners in a positive manner and who are seeking to create or sustain a successful music career. It may also be of interest to see if certain variables reach a "peak" in their impact on popularity - i.e., if a song reaches a point of becoming *too* long or *too* wordy, would this cause a drop off in the song's popularity? This study may help artists find the ideal characteristics required to produce a hit song.

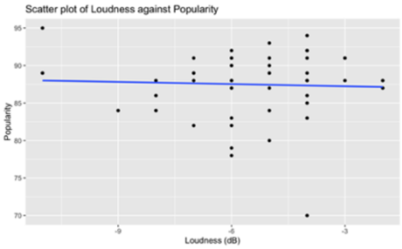
Linear Regression – Data Visualization

The data set provided is already in a format we can digest for analysis. We only drop the very first column as it is a repeat of the index column given in R. In **Figure 1**, a scatter plot of Speechiness against Popularity was generated. Although the data does not appear to be linear, we can see somewhat of a positive relationship between Speechiness and Popularity. However, based on the data points plotted, more songs with less Speechiness appear with relatively high Popularity but the most popular song in this data set has the second highest Speechiness. A transformation may be necessary to help linearize the data.

In **Figure 2**, a scatter plot of Loudness against Popularity was generated. There appears to be little to no relationship between the two variables. This indicates that a song's Loudness appears to have no effect on its popularity.

In **Figure 3**, a scatter plot of Length against Popularity was created. There appears to be a slight negative linear relationship between the two variables. This indicates that the longer a song is, the less popular it is. However, from the data set, there appears to be an optimized spot around the 200 second (3 minutes, 20 seconds) mark with most of the data hovering around there.





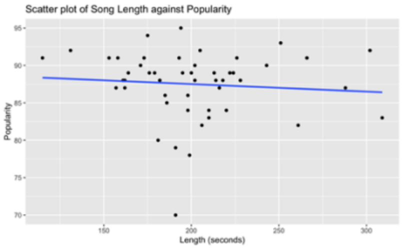
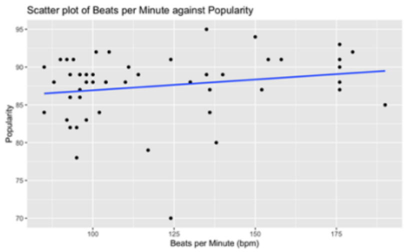


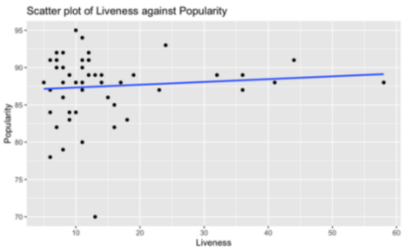
Figure 1 (left): Scatterplot of Speechiness against Popularity
Figure 2 (center): Scatterplot of Song Length against Popularity
Figure 3 (left): Scatterplot of Length against Popularity

In **Figure 4**, a scatter plot of Beats per Minute (BPM) against Popularity was created. There appears to be a positive relationship between Beats per Minute and Popularity. We can infer from the regression model that a song with a higher BPM would result in a more popular song. However, from this dataset, that does not appear to be true for all cases. There are more songs in the lower end of BPM that have relatively high popularity and the most popular song in this dataset is in the middle in terms of Beats per minute.

In **Figure 5**, a scatter plot of Liveness against Popularity was created. There appears to be a slight positive relationship between Liveness and Popularity which indicates that a song that is closer to a live recorded setting will be more popular. Looking at the dataset, this does not appear to be true. Majority of the songs on this data set are not recorded live and are mostly likely recorded in a studio. The most popular song in this dataset also displays this phenomenon.

In **Figure 6**, a scatter plot of Valence against Popularity was created. There appears to be a negative relationship between Valence and Popularity which indicates that a song that is too happy sounding maybe reach a point where the listener may think it sounds too "cheesy". This may be an interesting question in the future to examine what popularity, valence, and age demographic distribution.





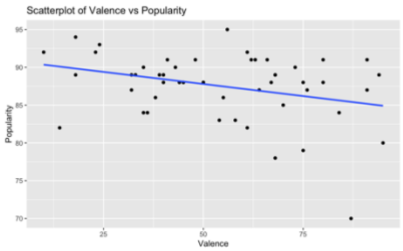
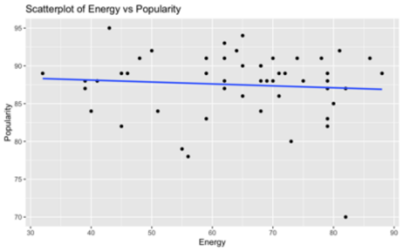


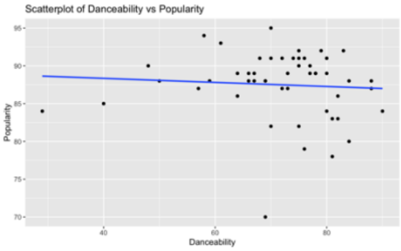
Figure 4 (left): Scatterplot of Beats per Minute Liveness against Popularity
Figure 5 (center): Scatterplot of Liveness against Popularity
Figure 6 (right): Scatterplot of Valence against Popularity

In **Figure 7**, a scatter plot of Energy against Popularity was created. There appears to be a slight negative relationship between Energy and Popularity. This may indicate that as a song's "energy" level increases, it will eventually be considered as less popular.

In **Figure 8,** a scatter plot of Danceability against Popularity was created. There appears to be a slight negative relationship between Danceability and Popularity. This may indicate that as a song's "Danceability" rating increases, it will eventually be considered as less popular.

In **Figure 9,** a scatter plot of Acousticness against Popularity was created. There appears to be a very slight negative relationship between Acousticness and Popularity. This may indicate that as a song is more acoustic sounding, it less popular. However, the regression line is essentially flat so this may not be a good predictor in determining popularity.





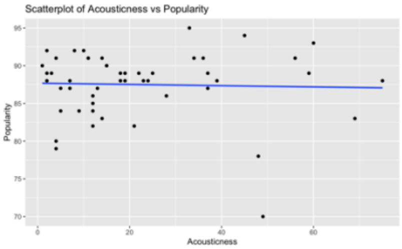


Figure 7 (left): Scatterplot of Energy against Popularity
Figure 8 (center): Scatterplot of Danceability against Popularity
Figure 9 (right): Scatterplot of Acousticness against Popularity

Linear Regression – Model Building

After researching how Spotify collects data on the variables above, we were concerned that multicollinearity would affect our linear regression model. To check this, a correlation matrix was created with all the variables on interest and found that most variables ranged from –0.3 to 0.3. The variance inflation factor (VIF) was then calculated, and we found that the variables in this dataset ranged between 1-2. We were able to conclude that multicollinearity was not a factor that affected our data.

After multicollinearity between predictors was ruled out, we began establishing a base model with popularity as the response variable. We first looked at the Adjusted R^2, Mallow's C$_p$, and BIC. The model that was best for Adjusted R^2, and Mallow's C$_p$ used the predictors Valence and Speechiness:

$$y_{Popularity} = 89.74257 - 0.06152x_{Valence} + 0.08944x_{Speechiness}$$

The model that was best for BIC used only the predictor Valence:

$$y_{Popularity} = 90.9887 - 0.06389x_{Valence}$$

We utilized the forward, backward, and stepwise automated search procedures to give us the lowest AIC value and aid us in our ability to choose a model. We found that the forward selection, backward elimination, and stepwise regression search procedures all resulted in Valence and Speechiness as the variables that would produce the most optimal model. We set up those predictors in the linear regression but found the p-value for Speechiness was insignificant, greater than 0.05.

$$H_0: \beta_{Speechiness} = 0 \text{ and } H_A: \beta_{Speechiness} \neq 0$$

We conducted a hypothesis test between our two-predictor model and a one predictor model with only Speechiness and ran an ANOVA test. We then obtained a p-value of 0.028 which is less than our significance level of 0.05 and concluded that the two-predictor model (Valence and Speechiness) is the best model for our dataset.

With our model selection process complete, we proceeded to check the residual plot assumptions. The residuals had a mean of zero, but they did not have a constant variance as the plot was right skewed. We started off by performing a log transformation on the response variable but that did not have the intended effect we were looking for, so we tested out a variety of other transformations. To correct the residual plot, we ended up taking the log of the predictor variable Speechiness.

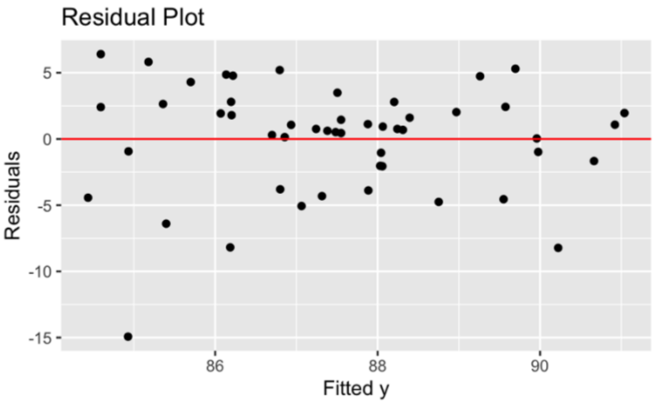


Figure 10: Residual plot of Speechiness and Valence on Popularity

In **Figure 10**, a residual plot was created to observe the mean of the residuals was 0 along with the residuals having constant variance. We noticed a bowtie shape in the data and concluded that although the residuals have a mean of zero, they do not have constant variance. Also, we wanted to display the partial residual plot as it will illustrate the marginal effect of adding a predictor when the other predictors are already in the model.

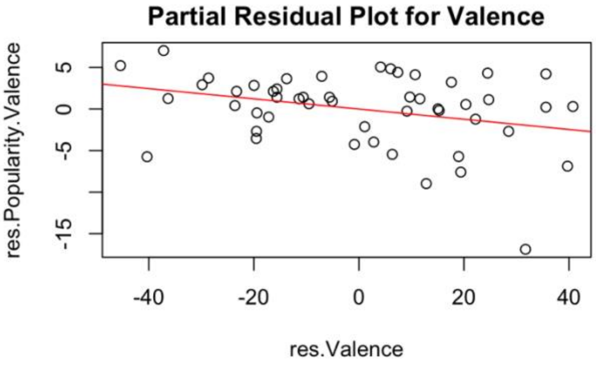


Figure 11: Partial Residual plot of Valence on Popularity

In **Figure 11,** we concluded that we would keep Valence as there is a linear increase in the plot along with the slope being equal to the coefficient for Valence in the regression model. Also, the residuals are evenly scattered across the regression line. The partial residual plot for Valence informs us that a linear term for Valence will be appropriate when Speechiness is already in the model, and that the estimated coefficient for Valence would be negative in the MLR model with Valence and Speechiness as predictors.

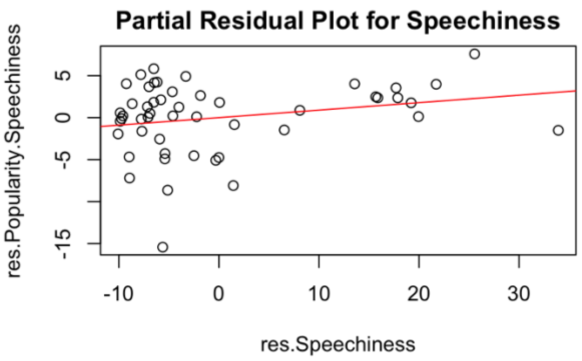


Figure 12: Partial Residual plot of Speechiness on Popularity

In **Figure 12,** the residuals are not evenly scattered across the regression line. The partial residual plot for Speechiness informs us that a linear term for Speechiness is not appropriate when Valence is already in the model. A transformation may be necessary. We will try a log transformation.

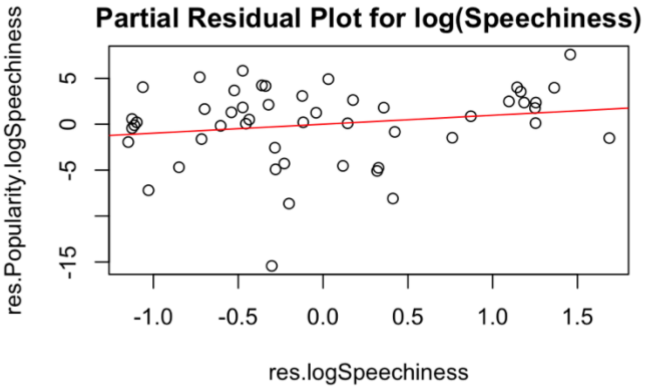


Figure 13: Partial Residual plot of log(Speechiness) on Popularity

In **Figure 13,** we log transformed Speechiness, now the residuals are evenly scattered across the regression line. The partial residual plot for Speechiness informs us that a linear term for Speechiness will be appropriate when Valence is already in the model, and that the estimated coefficient for Speechiness would be positive in the MLR model with Valence and log(Speechiness) as predictors.

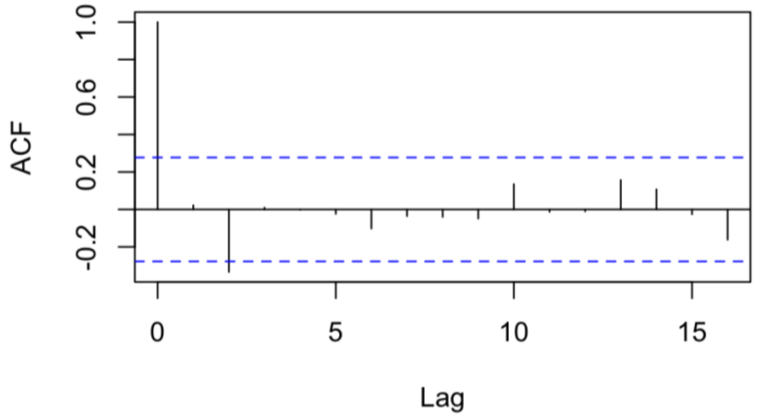


Figure 14: ACF (autocorrelation function) plot of the residuals

In **Figure 14**, we can see the ACF plot of the residuals. A significant ACF at lag 0 is always present 1. However, we also notice a significant ACF at Lag 2 which could show that the residuals are correlated. This may be the case as there may be some inherent structure in the observations. It is unknown whether the data is pre-ranked in descending order for the Top 50 or if the top 50 were randomly selected based on some predictor. There may also be a highly influential outlier in this data set that may be skewing the residuals, especially since this is a small data set. We ran a few model diagnostics to find if there truly was an outlier influencing the data. We found the critical value using the Bonferroni procedure and applied that into the externally studentized formula to identify if any outliers existed. We found that data point 26 ("If I Can't Have You" by Shawn Mendes) was the outlier. We also tested for leverage points and found that data point 30 ("QUE PRETENDES" by J. Balvin had high leverage. Then we tested DFFITS and found that data points 16 ("No Guidance (feat. Drake)" by Chris Brown) and 26 ("If I Can't Have You" by Shawn Mendes) were influential in terms of DFFITS. We then tested for DFBETAS and found:

Data Point	Influential in terms of:
1: "Senorita: by Shawn Mendes	$\beta_{Speechiness}$
10: "bad guy" by Billie Eilish	$\beta_{Speechiness}$
16: "No Guidance (feat. Drake)" by Chris Brown	Intercept, $\beta_{Valence}$
18: "Sunflower - Spider-Man: Into the Spider-Verse" by Post Malone	β_{Valence}
26: "If I Can't Have You" by Shawn Mendes	Intercept, $\beta_{Valence}, \beta_{Speechiness}$
39: "Sucker" by Jonas Brothers	$\beta_{Valence}$

We finally performed the last model diagnostic test, Cook's distance and found that no songs are influential in terms of Cook's distance. It is safe to assume that there exists data in this data set that has some level of influence.

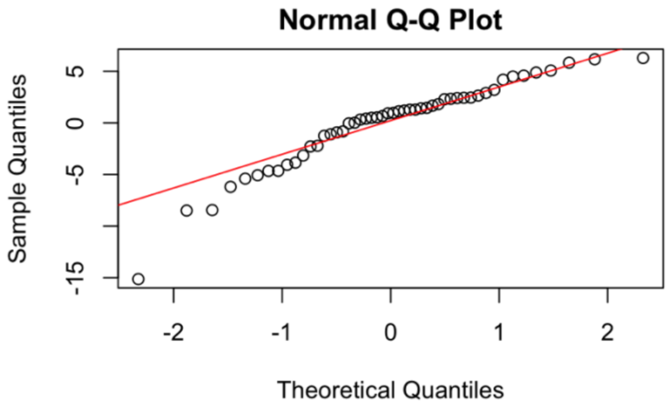


Figure 15: QQplot of Quantiles

In **Figure 15**, we can see a QQplot of the Quantiles. The quantiles fall closely to the line, so we can assume the residuals with the transformed predictor follows a normal distribution. With Linear Regression assumptions 1, 2, and 4 met, we only rely on assumption 3 to decide if our model is good for use. Since Our final Multiple Linear Regression model:

$$y = 89.74257 - 0.06152x_{Valence} + 0.08944x^*_{Speechiness} \text{ where } x^*_{Speechiness} = \log(x_{Speechiness})$$

Linear Regression – Conclusion

Our question was tailored around which predictors would have the greatest impact on our response variable, Popularity. We found that Speechiness and Valence were the highest contributing predictors to affect Popularity. Holding Speechiness constant, an increase in one unit of Valence equates to a 0.08944 decrease in Popularity. When Valence is constant, an increase in one unit of Speechiness produces an increase of 1.0935 of Popularity.

When we originally chose to analyze this data, we selected a subset of data's predictors thinking those would be sufficient: Beats Per Minute, Loudness, Liveness, Length, and Speechiness. The model ended up resorting to the intercept only model which does not give a clear picture of the response. From there, we decided to expand our predictor variables to create a more beneficial model.

Logistic Regression

Our second question will explore the logistic relationship between variables in the dataset what affects whether the song is categorized in the genre of "Pop" or Not ("NotPop"). Pop music has an ever-evolving definition, but often features a consistent rhythm, simple beats and melodies, and catchy hooks/choruses. An analysis of this dataset can help further clarify what factors lead to a song being considered to belong to the pop genre. In the public opinion, pop music is likely considered as the go-to for listening at clubs, parties, or other social gatherings. Similar to our first question, artists and producers (as well as others in the music industry) may be interested in results that show which variables help earn a song the desired “Pop” genre label. According to a 2018 study from Statista[3]. Pop music still dominates the worldwide music market, with 64% of consumers reporting that they listen to pop music. Due to the genre’s prevailing popularity in our culture, many people could benefit from learning more about which factors lead to songs being labeled as “Pop” or “Other”.

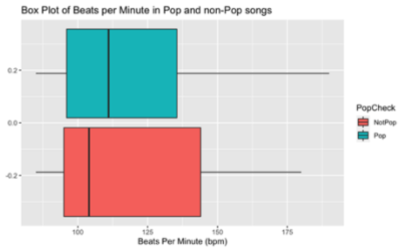
Logistic Regression – Data Visualization

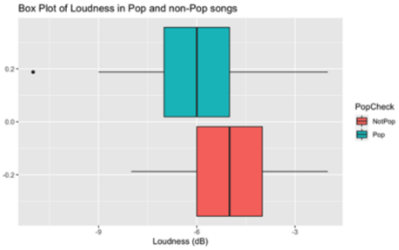
For the second question of interest, we analyze the top 50 Spotify songs in 2019 between Pop and non-Pop songs and compare their spreads across a song’s characteristics. We had to create a binary variable in the data set in order to be able to perform the logistic regression analysis. The PopCheck classifier was created to read the text string value in the Genre column in the dataset. If the word “Pop” appeared in that song’s genre, it was classified as “Pop.” If “Pop” did not appear, the song was classified as “NotPop.”

In **Figure 16**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Beats per Minute. The medians between the two categories are about the same as well as their respective spreads. We can infer, from this dataset, there is no significant differences in a song’s Beats per Minute between Pop and non-Pop songs.

In **Figure 17**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Loudness. The medians between the two categories are not the same. The Loudness value is a measure of how far away it is from the threshold of human hearing. A negative Loudness value indicates a softer Loudness rating. We can infer, from this dataset, Pop songs tend to have lower Loudness ratings than non-Pop songs.

In **Figure 18**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Length. Although the medians between Pop and non-Pop songs appear to be the same, their spreads are different. From this dataset, Pop songs tend to be within 180 to 210 seconds while non-Pop songs are anywhere from 165 seconds to 220 seconds. The medians hovering around the 200 second mark does align with a comment made about the scatter plot in **Figure 3**.





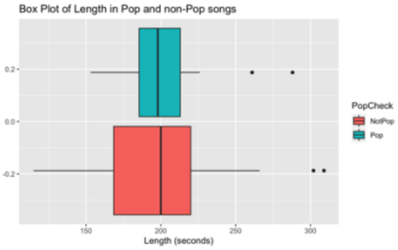


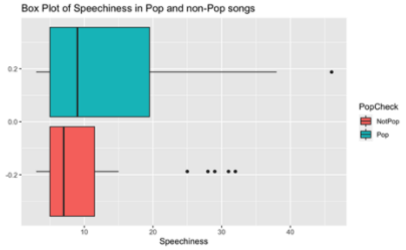
Figure 16 (left): Box Plot of Beats per Minute in Pop and non-Pop Songs
Figure 17 (center): Box Plot of Loudness in Pop and non-Pop Songs
Figure 18 (right): Box Plot of Length in Pop and non-Pop Songs

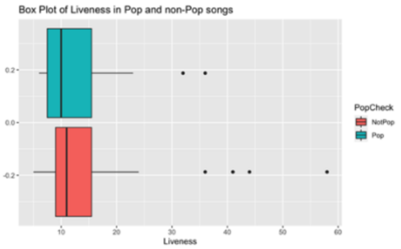
[3] https://www.statista.com/chart/15763/most-popular-music-genres-worldwide/

In **Figure 19,** a box plot was created to observe the spread of data between Pop and non-Pop Songs on Speechiness. The medians and spreads between the two categories are not the same. Since non-Pop can include any song that does not have “Pop” in its genre, this spread makes sense as some songs may be instrumentals or part of a soundtrack that do not have any words in it. From this dataset, we can infer that pop songs tend to have a higher Speechiness than non-Pop songs.

In **Figure 20**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Liveness. The medians and the spreads between the two categories are about the same. From this dataset, these visualizations indicate that the Pop and non-Pop songs in this list tend not to be live recordings but rather songs recorded in a recording studio.

In **Figure 21**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Valence. The medians between Pop and non-Pop songs are not the same, however, the spreads appear to be the same. This is an interesting observation as Valence is a measure of Happiness. How does Spotify measure Happiness? This may be calculated based on if a listener favorites the song or not.





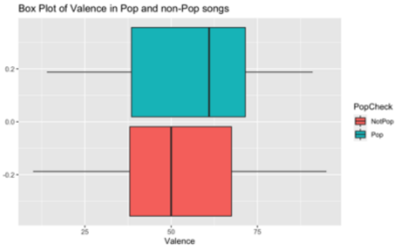
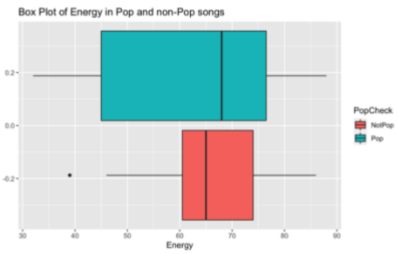


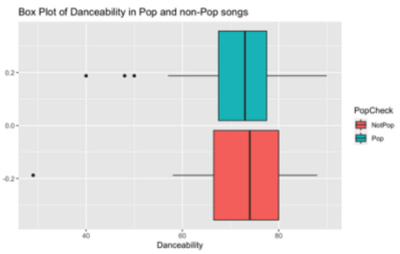
Figure 19 (left): Box Plot of Speechiness in Pop and non-Pop Songs
Figure 20 (center): Box Plot of Liveness in Pop and non-Pop Songs
Figure 21 (right): Box Plot of Valence in Pop and non-Pop Songs

In **Figure 22**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Energy. Both the medians and spreads of Pop and non-Pop songs are not the same. It makes sense that Pop songs have such a large spread of Energy levels as Pop music can be songs of varying levels of energy.

In **Figure 23**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Danceability. It appears that both the median and spreads of data are the same between the two categories. We can infer, from this dataset, both Pop and non-Pop songs have equal levels of Danceability.

In **Figure 24**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Acousticness. The medians between Pop and non-Pop songs appear to be the same, while the spread of Pop songs seems to skew more acoustic than non-Pop songs. We can infer that in 2019, the top 50 songs tended to sound more Acoustic than non-Pop songs.





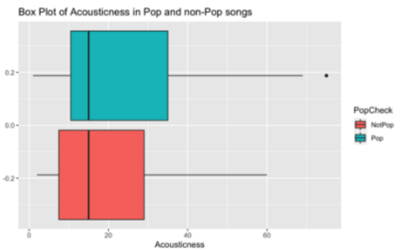


Figure 22 (left): Box Plot of Energy in Pop and non-Pop Songs
Figure 23 (center): Box Plot of Danceability in Pop and non-Pop Songs
Figure 24 (right): Box Plot of Acousticness in Pop and non-Pop Songs

In **Figure 25**, a box plot was created to observe the spread of data between Pop and non-Pop Songs on Popularity. Both the spreads and medians of popularity between Pop and non-Pop songs are not the same. This is interesting as one would think that pop songs would consist of all of the higher popularity values. Pop songs have the highest Popularity and lowest Popularity.

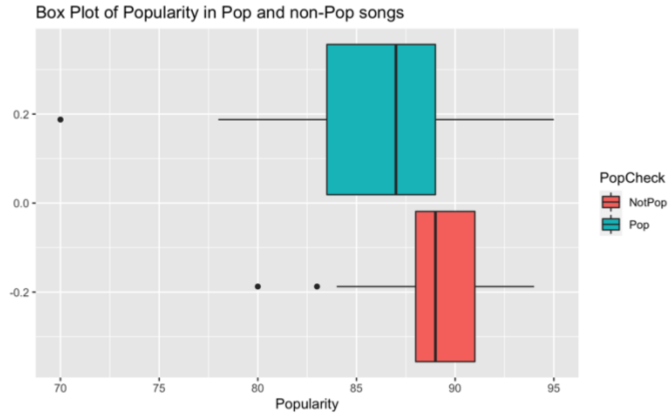


Figure 25: Box Plot of Popularity in Pop and non-Pop Songs

Logistic Regression – Model Building

To create the logistic regression model, first we created a data frame that dropped the columns that hold unique values such as: Track Name, Artist Name, and Genre. Since there are only 50 data points in the data set, we could not split the data 50-50. We needed at least 30 samples in the training data set, so the data was split 80/20 train and test. Next, we needed to find which predictors in the data set were useful in the model, so we utilized the Automatic Model Selection search procedures: Forward Selection, Backward Elimination, and Stepwise Regression. Across the three search procedures, the model with the lowest AIC dropped all the predictors except for only Speechiness and Popularity. When creating the logistic regression model with the two predictors against the user created PopCheck, we noticed that Speechiness had an insignificant p-value of 0.069 which is greater than our significance level of 0.05. So, we conducted a Wald hypothesis test:

$$H_0\colon \beta_{Speechiness} = 0 \text{ and } H_A\colon \beta_{Speechiness} \neq 0$$

We then compared the p-values by running a Chi-Square test. The reduced model without Speechiness had an insignificant p-value of 0.051, which is greater than our significance level of 0.05. So, we reject the null hypothesis and continue with the 1-predictor model that only uses Popularity. Our final logistic regression model is as follows:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 20.3225399 - 0.2361914x_{Popularity}$$

We still needed to evaluate the performance of the model. So, with the test set of the data, we now create a ROC curve to evaluate its classification performance, seen in **Figure 26**.

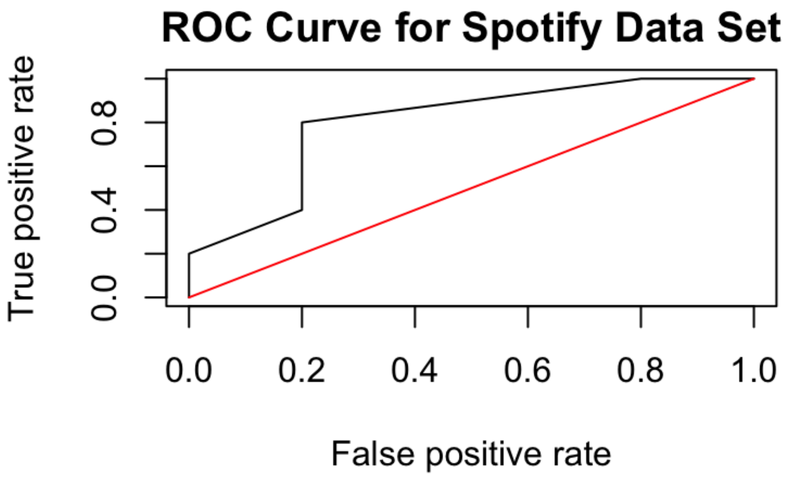


Figure 26: ROC Curve of Logistic Regression Model

Since this ROC curve is above the diagonal line, the logistic regression performs better than random guessing. We also obtained an Area Under the ROC-Curve (AUC) value of 0.8, which confirms that our logistic regression model performs better than random guessing. Now we can analyze the accuracy of the model on the test data. With a threshold of 0.5, we can see the confusion matrix result of the test data in **Table 1.**

Cutoff > 0.5	**False**	**True**
NotPop	5	-
Pop	5	-

Table 1: Confusion matrix with a cutoff of 0.5

This yields us an accuracy of 0.6, however, we also obtain an error rate of 0.5. These values are not ideal as we are trying to minimize the error rate. We can develop a density plot of the predicted probabilities to identify a better threshold value. Based on **Figure 27**, it appears that a cutoff value of 0.25 may be optimal.

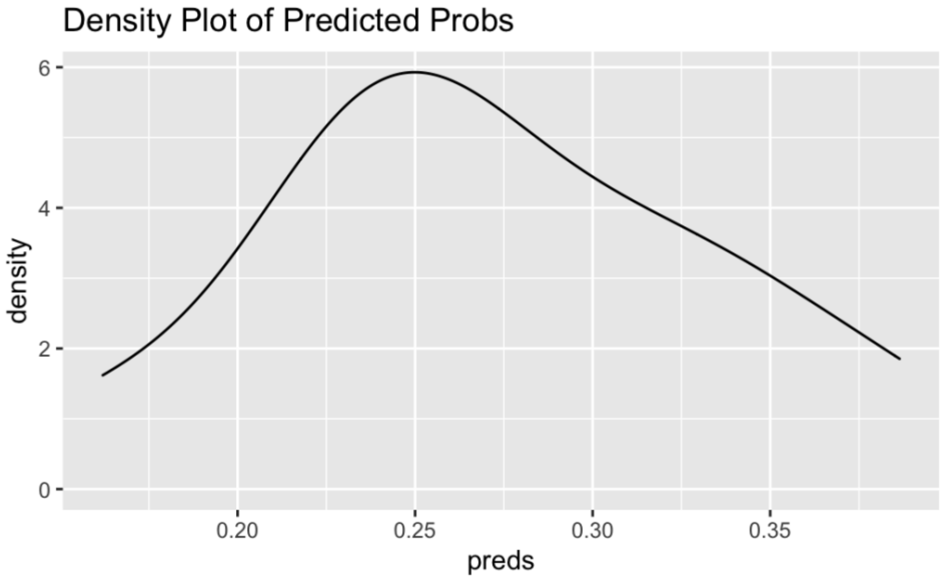


Figure 27: Density Plot of Predicted Probabilities

This lower threshold resulted in a confusion matrix seen in **Table 2**. The accuracy of this threshold is 0.8 with an error rate of 0.2.

Cutoff > 0.25	**False**	**True**
NotPop	4	1
Pop	1	4

Table 2: Confusion matrix with a cutoff of 0.25

Logistic Regression – Conclusion

With the final logistic regression model found, we can now figure out whether a song will be classified as Pop or non-Pop based solely on its Popularity. For a 1% increase in Popularity, the log odds of a song being non-Pop decreases by 0.2361914. Rather the odds of a song being non-Pop is divided by a factor of 1.266417. This conclusion is similar to the behavior seen in the linear regression model. This may be a factor of the song's that were released in 2019, the more popular songs tended to be non-Pop.

One challenge faced with creating the binary variable with PopCheck was the fact that some songs do not have "Pop" in their Genre column even though they are considered as Pop songs to the public. This causes the PopCheck classifier not to see it and thus it gets categorized incorrectly. A future study may need to be conducted analyzing the Genre column further to help better parse through the text so that a song that is considered as Pop even when the word "Pop" may not appear in its Genre column. Our test data was also on 10 samples. If we had a large data, we believe our model would have been more accurate in classifying songs.