Alissa Caltagirone, Yaxi Lei, and Sam Webster

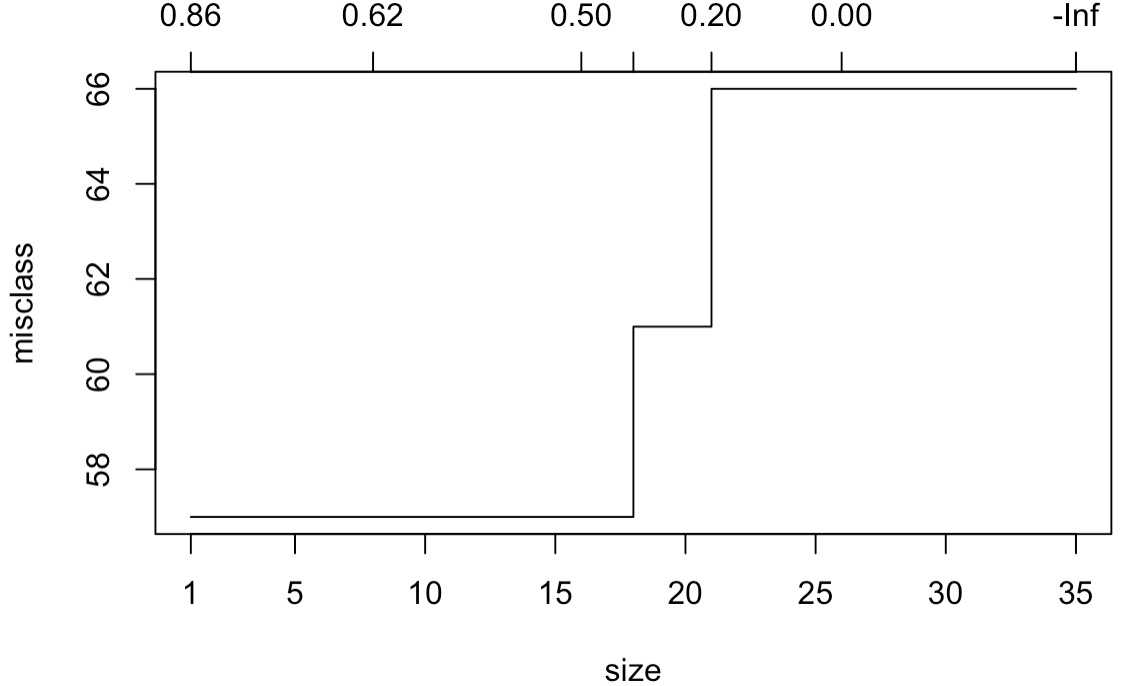
MGSC 310 Final Project Essay: Spotify Group

December 14, 2018

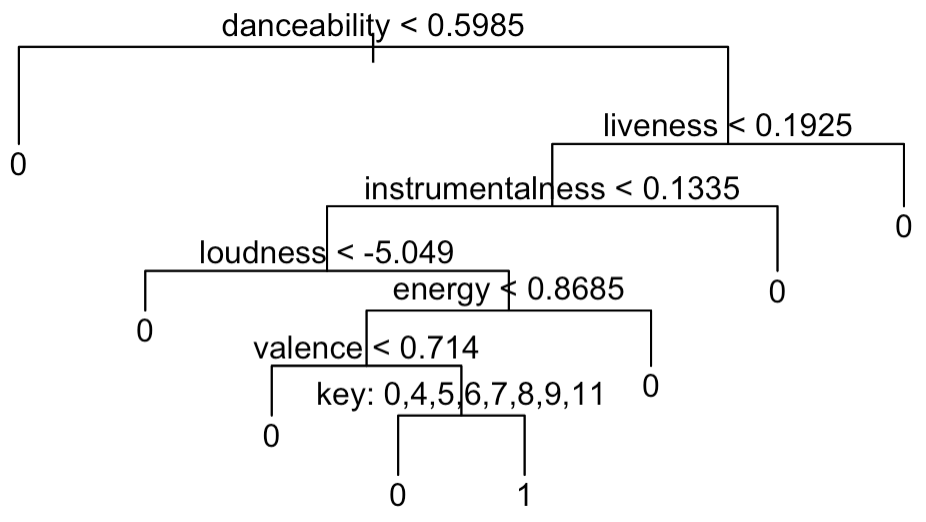
We initially analyzed a dataset provided by Kaggle named “Top Spotify Tracks of 2017,” which included 100 observations and 13 variables. The original variables were as follows: id, name, artists, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration\_ms, and time\_signature. Some of the variables had measures from 0 to 1 such as danceability and energy. For example, energy represents a “perpetual measure of intensity and activity.” As the project continued, Professor Hersh provided us with a second data set named “Spotify Song Attributes,” which included 2,017 observations and 19 variables.

We found that the larger dataset contained all of the original variables and clearly more observations about songs not included in the Top 100 list. Since that was the case, we did not actually combine the two datasets, instead, we created a list that contains all of the Top 100 song names and loop through the larger dataset to see if the song name is an element of the list. We added a new list called “top100songs” in the larger dataset and marked all songs that also appear on the Top 100 list as 1 and the rest as 0. We ended up with 43 songs marked as 1 in this column. To make our further analysis easier, we also did a pre-data cleaning that deleted the attributes which are clearly not related to what we want to predict such as ID. After the datasets were ready to be analyzed, we began with some simple summary analysis.

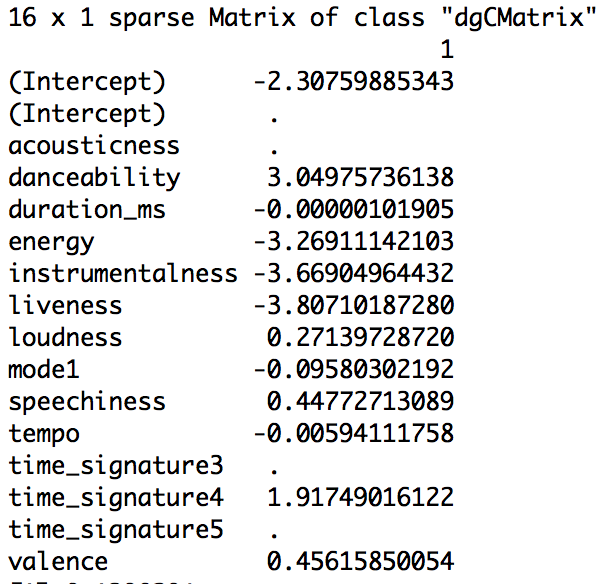
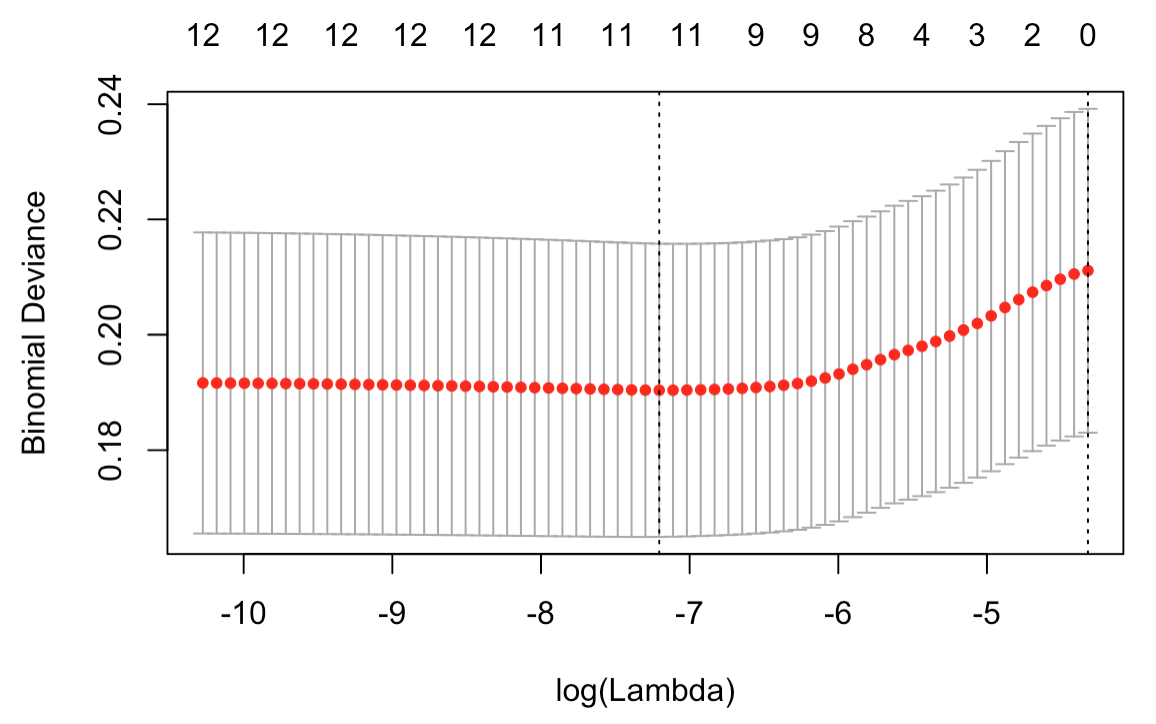
First of all, we want to know which songs are going to be in the Top 100. In order to determine this we trained a tree model on the larger dataset to try to get the answer. Before we built the tree, we cross validated the tree to see how many layers our tree has to avoid overfitting. It gives us the following plot:



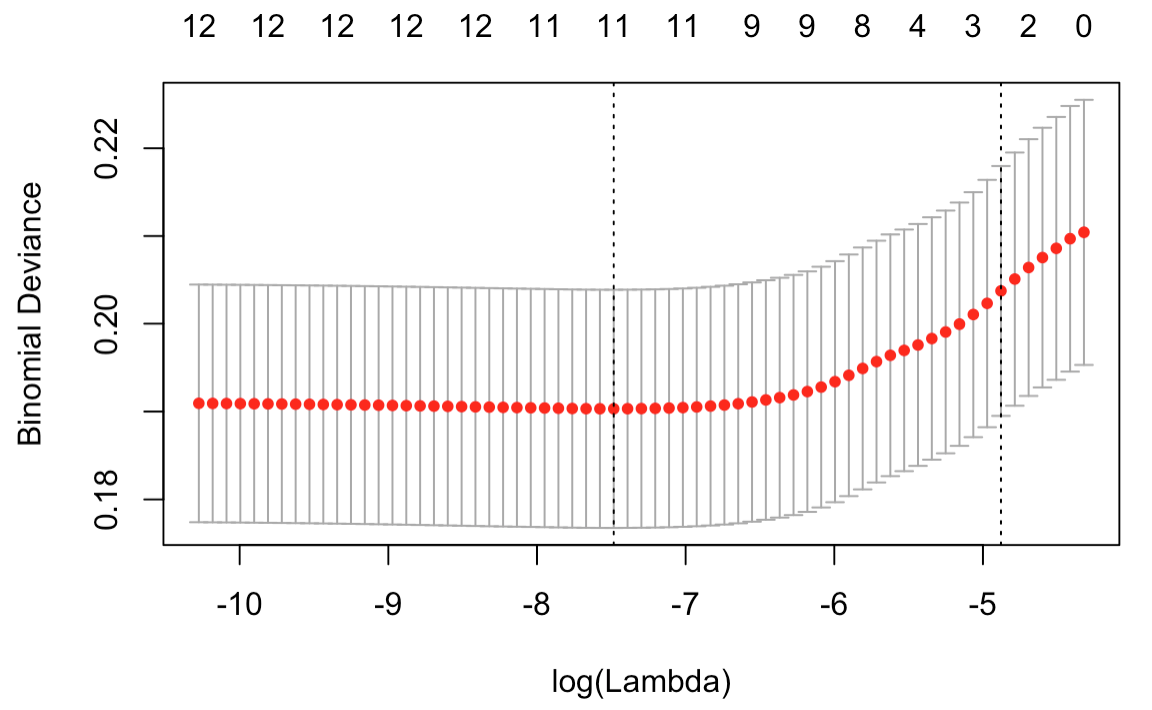
As we can see, this plot basically tells us that if our tree has layers from 1 to 18, it should be fine. As a result, we chose the 3 as the parameter and we got this tree model:

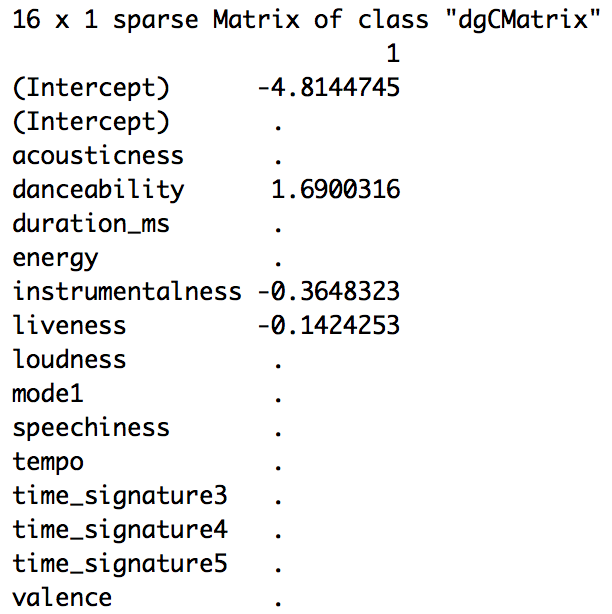
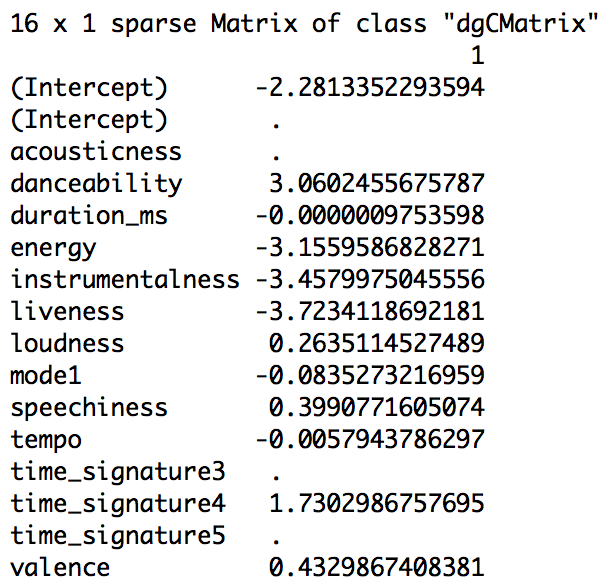


This tree tells us is that the danceability is the most important attribute for deciding if a song will successfully goes into Top 100. If the danceability is smaller than 0.6, it is not possible for the song to goes into Top 100. The second most important attribute is the liveness. We want the liveness to be smaller than 0.19 to get the song into Top 100. The instrumentalness, loudness, energy, valence and key are the other variables that has effects on the result. To convince our result, we also built a Lasso model on the same attribute in the same dataset.

After building our Lasso model, we found that our Tree model was extremely accurate. When we made our first Lasso model, the lambda minimum predicted that the model with the least amount of error would use eleven variables. This result had an R-squared error of 0.13.  ****

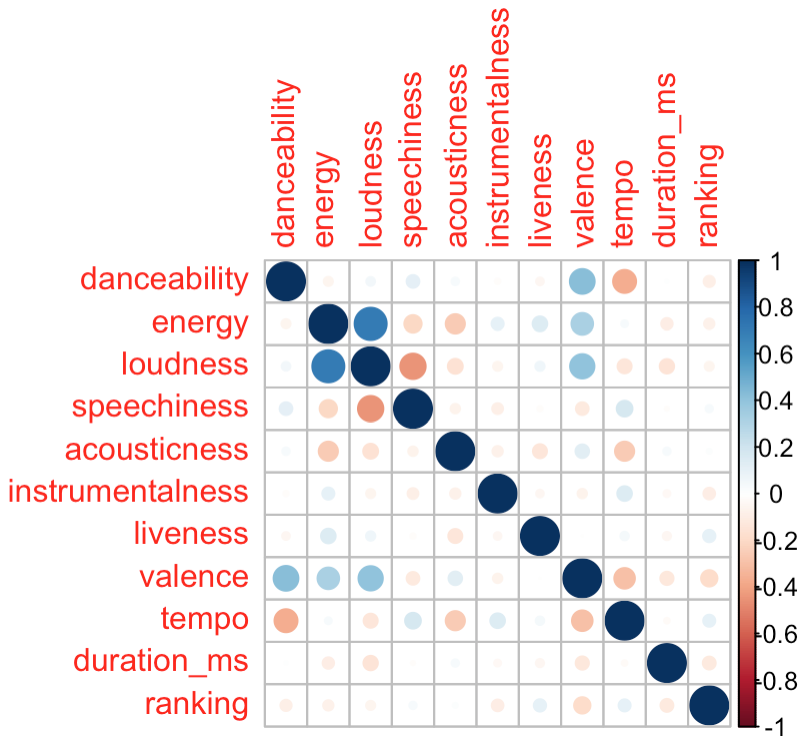
This outcome gave us great insight, but we wanted to look to the one standard error in order to narrow down the results and be more cost efficient with our variables. Unfortunately, we found that the 1se gave us zero variables. To receive better results, we set the seed to twenty-three.

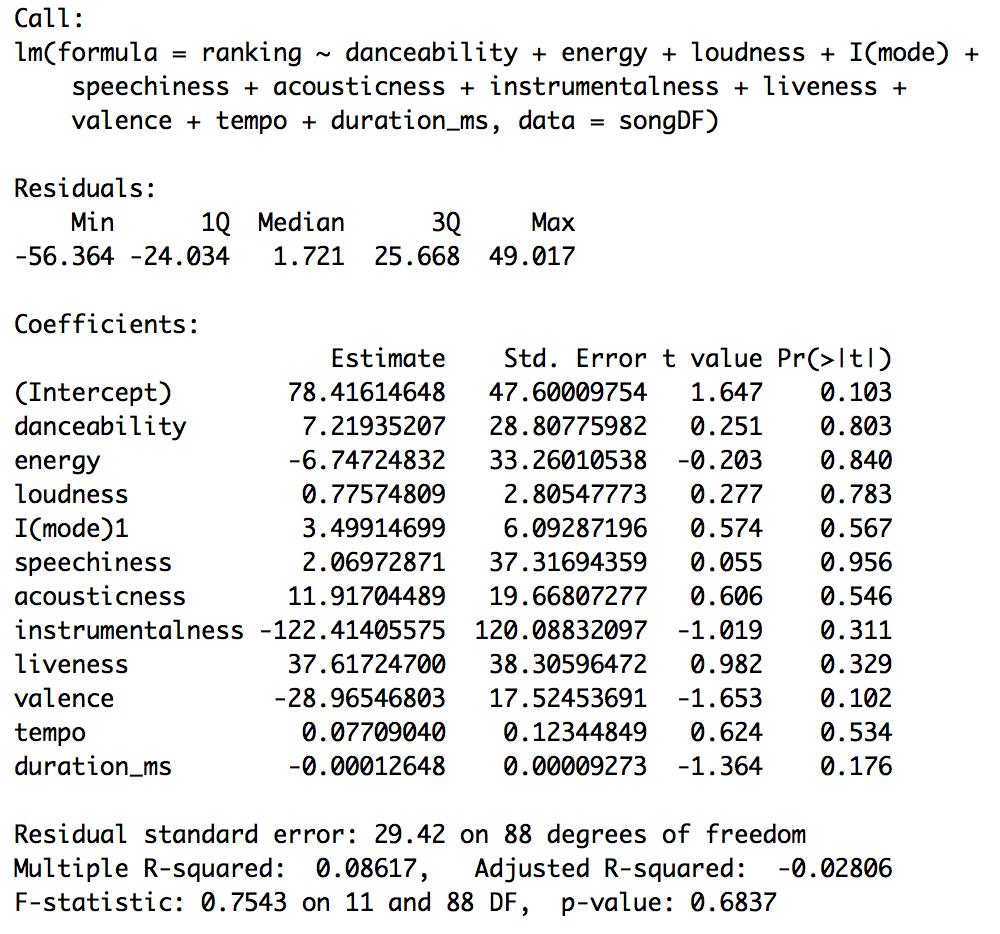
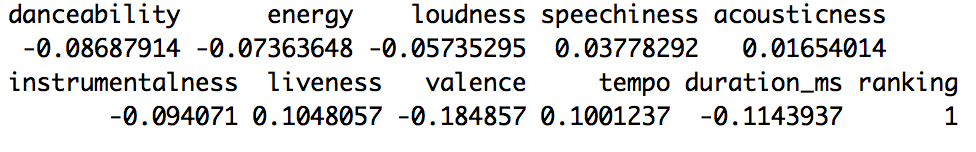
****

****

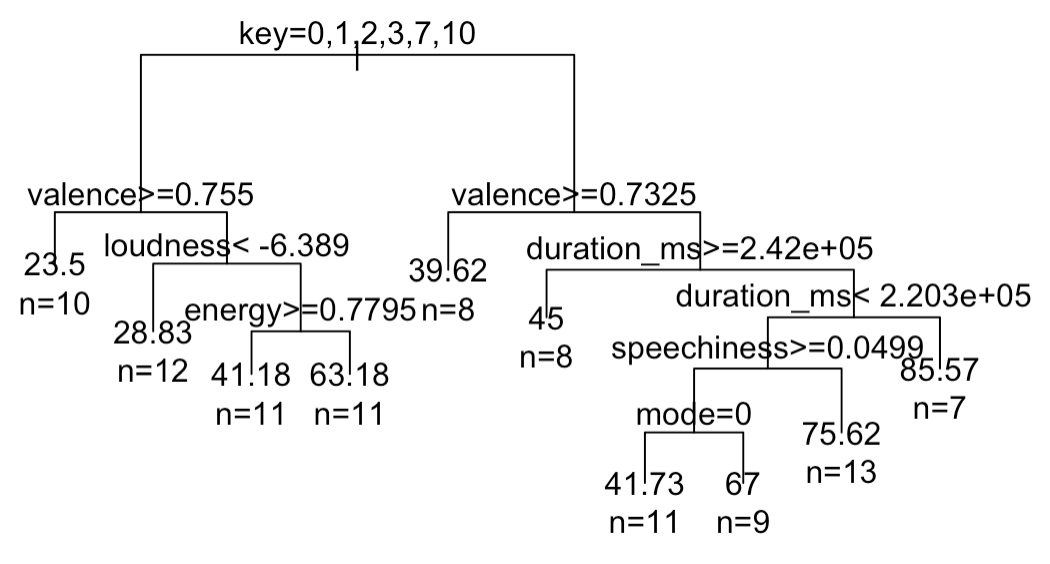
The lambda minimum returned eleven variables with an increase in the R-squared value by only 0.01, from 0.13 to 0.14. The lambda 1SE gave us the three strongest variables that influence a song reaching the Top 100. These three variables are danceability, instrumentalness and liveness. The top three variables that our Tree model predicted were these exact same variables, which proved that our results were accurate.

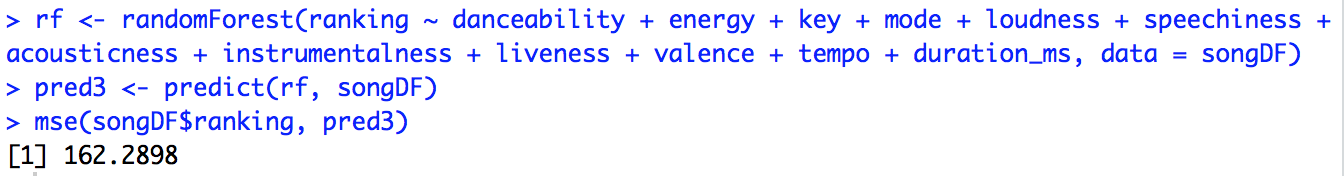
After we found out that which attributes will make a song goes into the Top 100, we still want to know if the song is already in the Top 100 list, is there any factors that have influences on the ranking? It should still have a difference for a song to be on the top 1 comparing to the Top 100. As a consequence, we did a regression model only on the small dataset to predict ranking using the other attributes. Before we built the regression model, we did a corrplot and a correlation table first to see if we need to choose just a subset of the variables because some variables may have a very weak relationship with the ranking that we do not need to test on. The corrplot and the coefficient table is shown below:



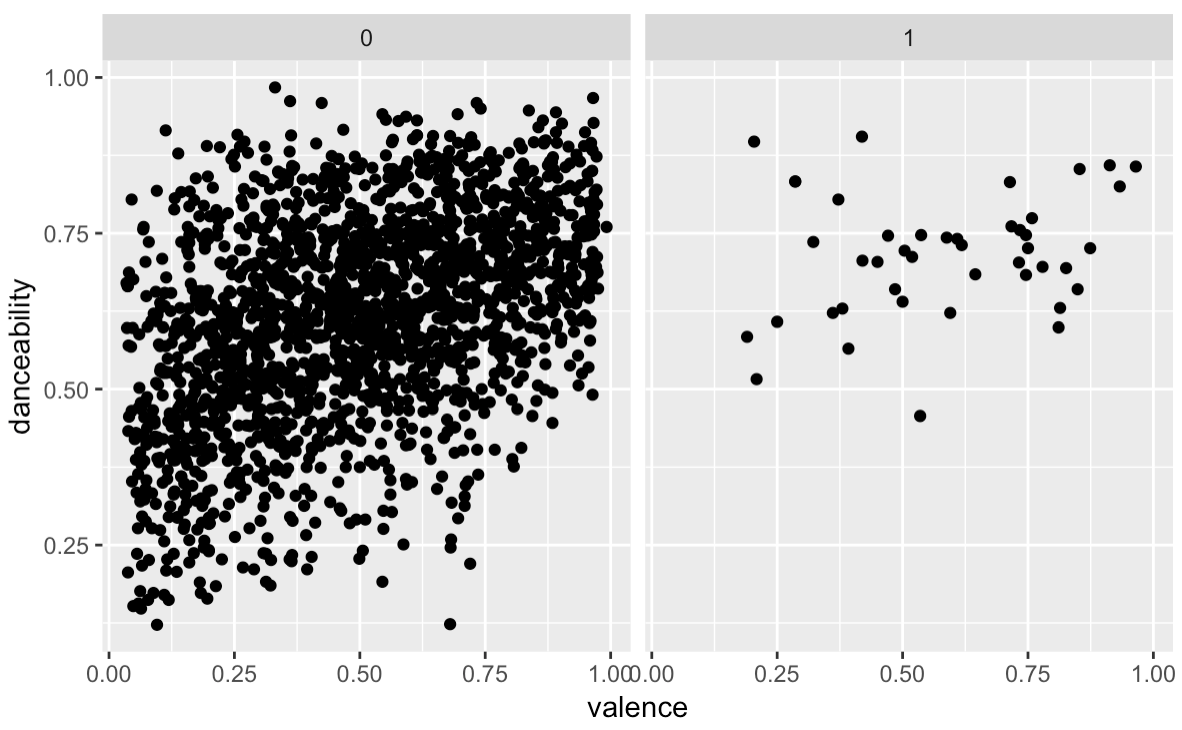
From that we thought that there is no significant difference among different variables with its relation to the ranking. The coefficients are similar and the plot confirms that. As a result, we used all of these variables as well as a factor level variable called “mode” in the regression model. The result is shown here:

This model obviously does not work well. The R-squared is 0.08 and all of the p-values are high. We calculated the MSE and got a large number “761.45” as imaged. Because of that, we decided to use a different model to predict the ranking, which is another tree model shown here:

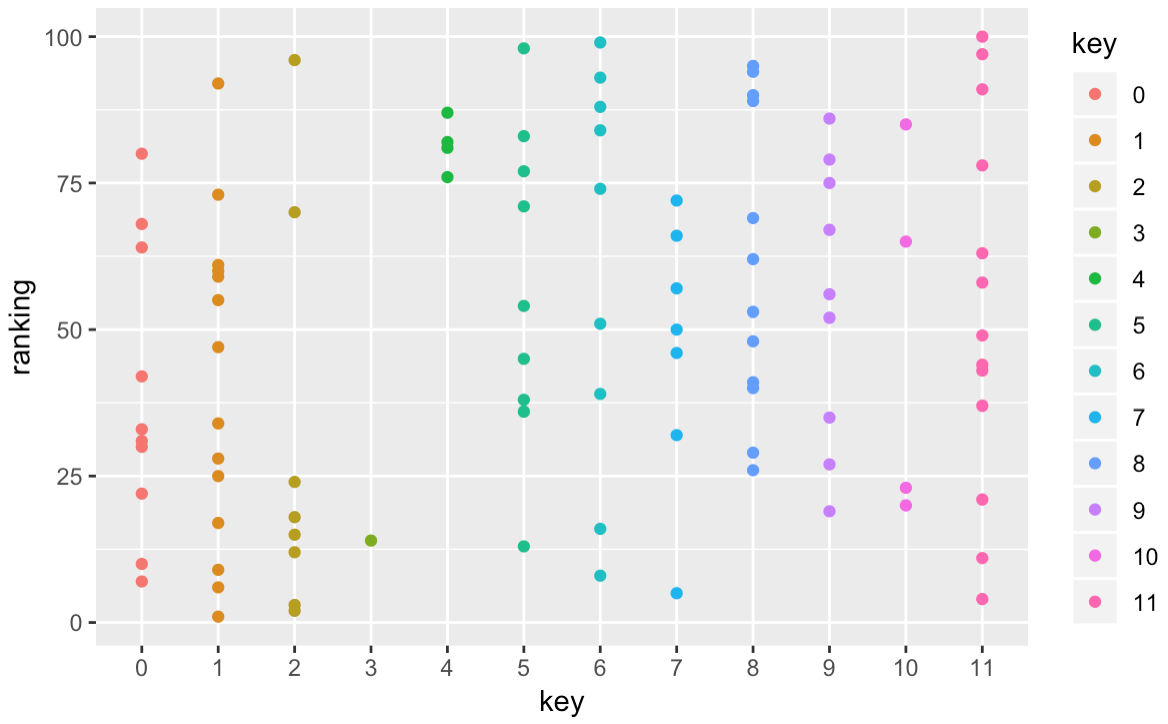


Our second tree model predicted the ranking within the Top 100 list using just the data from the original dataset. It concluded that the key and the valence replaces danceability as the most important variable once the song enters the Top 100 List. Valence means the intrinsic attractiveness (positive valence) versus unattractiveness (negative valence) of a song while the key represents the overall key of the track (0 as C, 1 as C♯/D♭, 2 as D, etc.). This model has an MSE 464, which is clearly not great. However, this tree still performed better than the regression model, which had an MSE of about 761. As an example to interpret the tree, there is a left node of 23.5, which in simple terms means that if the song has a key in 0,1,2,3,7,10 and a valence level greater than or equal to 0.755 then the song will rank around number 23.5. This tree also emphasizes that variables such as loudness and duration are still very important for a song once they are in the Top 100 list. After that, we did a random forest model as well. We got an MSE of 162.29 on the random forest one, which was much better than the other two. (Code is shown as below)

After the two analysis, we found that the danceability is the most important thing in predicting whether a song will go into Top 100, and key and valance is the most important thing in affecting the ranking. To understand it a little bit better, we created some data visualization based on our results:

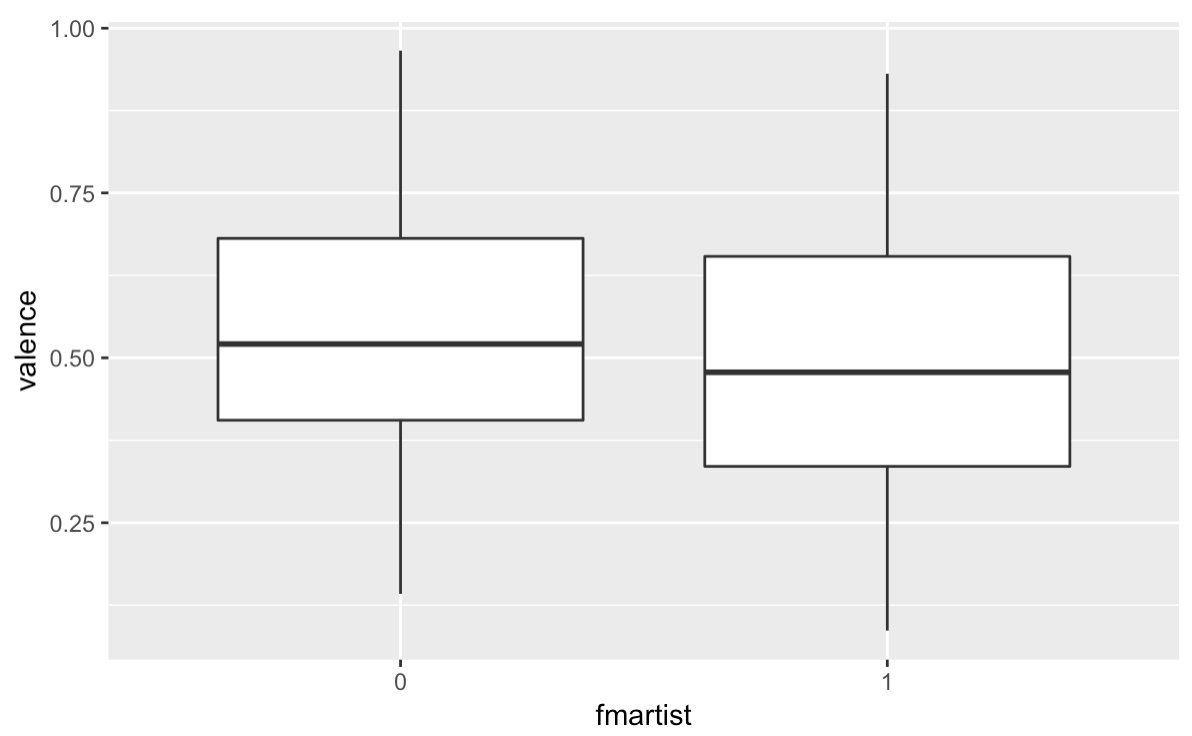


For example, the plot above is a scatterplot about danceability and valence. The right one with label 1 are the songs that are on the Top 100 while the left are not. We can clearly see that both the danceability and the valence for the songs that are on the Top 100 are higher than those are not, which convince our results again. We did another visualization for key:



It does show that when the key is 0,1,2,3, and 7, more points are shown at a lower position, which indicates a higher ranking. It is consistent with what we found before.

However, our analysis still have some limitations.

For our analysis, we predicted that the artist of a song would have an extreme effect on the outcome of a song’s ranking. We decided to take out the artist variable. This is because we wanted to analyze the characteristics of each song to find out what made people want to stream, buy, and ultimately put them in the Top 100. After finding what characteristics influence a song reaching the Top 100, we decided to test how top artists’ songs compared to the average song on the list. We created a variable that held all of the songs in which the artist had two or more songs in the Top 100. We called this mini dataset “fmartist,” which stands for “famous artists.” We compared their characteristics to those of all the other songs in the Top 100. We were quite surprised with our results. 

We found that for valance, and a few other top correlated variables, the average song written by famous artists had inferior statistics to songs that were sung by artists who only had one song in the Top 100. This was interesting because this meant that famous artist’s songs were reaching the Top 100 list not because of the characteristics of their songs, but because of their name.

In conclusion, we learned a lot from our initial summary statistics and were able to grow using our chosen models. We concluded that the most important characteristic of a song before they enter the Top 100 list is danceability. After more modeling and analysis, we found that once a song enters the Top 100 list, then valence and key become the most important variables. The subject overall was very interesting and relevant to the time. We would have desired to do some more analysis including the artists’ impact and other variables if given more time and knowledge of other models to go deeper into researching the trends to fully understand an algorithm as to why a song reaches the Top 100 list.