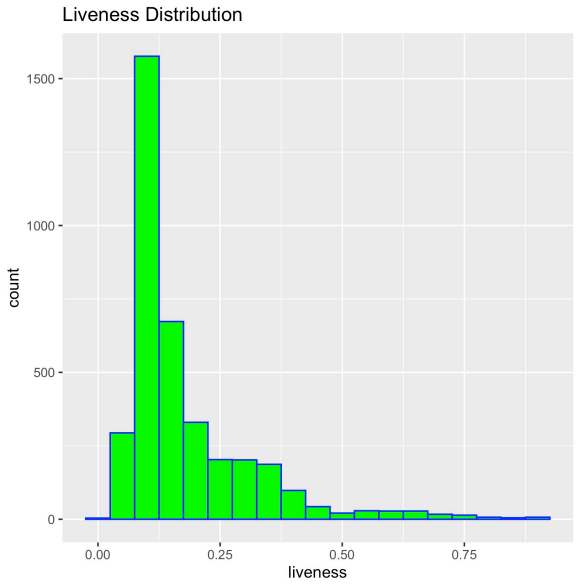
Abstract

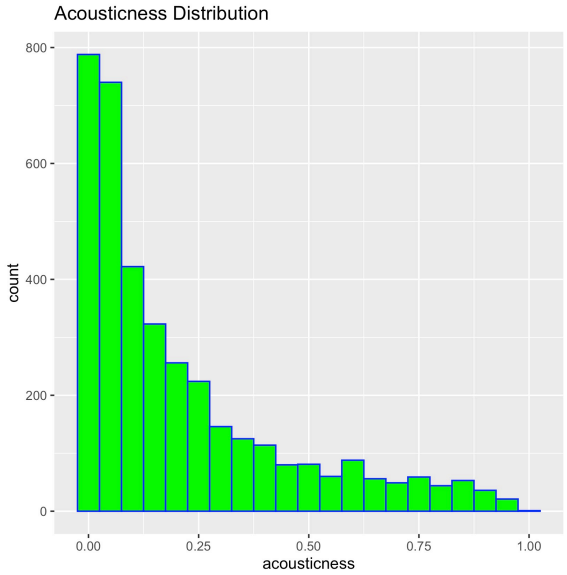
For our project, we examined the Billboard Top 100 Songs for every day from January 2016 to 2022 and aggregated statistics for each individual song over this time period, where there were 3611 unique songs that entered the top 100. At the outset, our objective was to predict the amount of time that a given song would stay in the Billboard Top 100, regardless of rank. This way, we could better understand why certain songs may be more popular than others, and we could observe what characteristics of a song would most affect its popularity. After trying various models, we decided to use an ordinal logistic regression model, where we input variables obtained from backwards stepwise regression. We found that speechiness and valence were two of the most impactful variables in a song's popularity, and the model itself had an accuracy of 65.3% on the test set of songs, which was 20% of the data.

Methodology

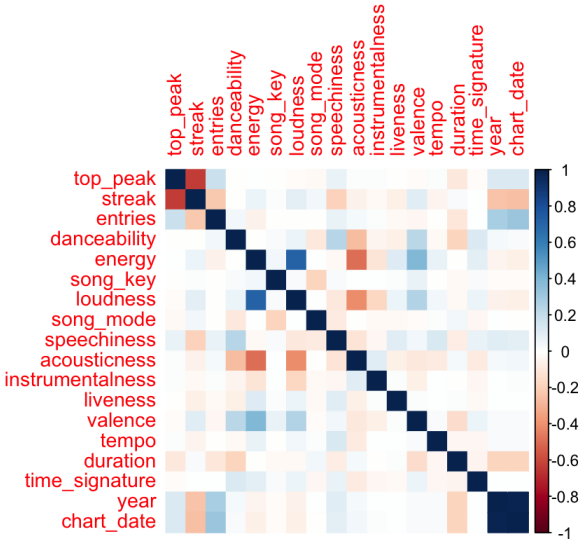
Coming into the project, we knew we wanted to somehow predict the length of time that a given song would stay in the Billboard Top 100. However, there are many possible ways to do this, so part of the challenge of this project was clearly defining what we wanted to predict. We considered looking at all of the times a given song would enter and exit the top 100, and then take the average time spent in the top 100 over the entire period. However, this would prove to be too tedious, so to simplify the operation and maintain the original goal of analyzing time on the list, we initially decided to look at only the longest streak (in weeks) for each song and use that as our response variable. In addition, we used the number of re-entries into the top 100 for each song as a predictor, and we created variables like the average ranking, standard deviation of ranking, and top ranking based on the daily ranking data.

Since we started with two different datasets, one with the Billboard ranking data and the other with Spotify song metric data, we joined the two datasets by joining on the song title and artist name variables; this gave us 3611 unique songs. This allowed us to perform some intriguing EDA, as we examined the distributions of the predictors. While most variables were skewed in some form, there were a few variables that stood out in terms of needing alterations. Specifically, the song_mode and time_signature variables were integer variables concentrated on only a few values, which made us change these variables to factors. Additionally, variables like acousticness and liveness were quite right skewed, so we were inclined to apply a log transformation to these variables.



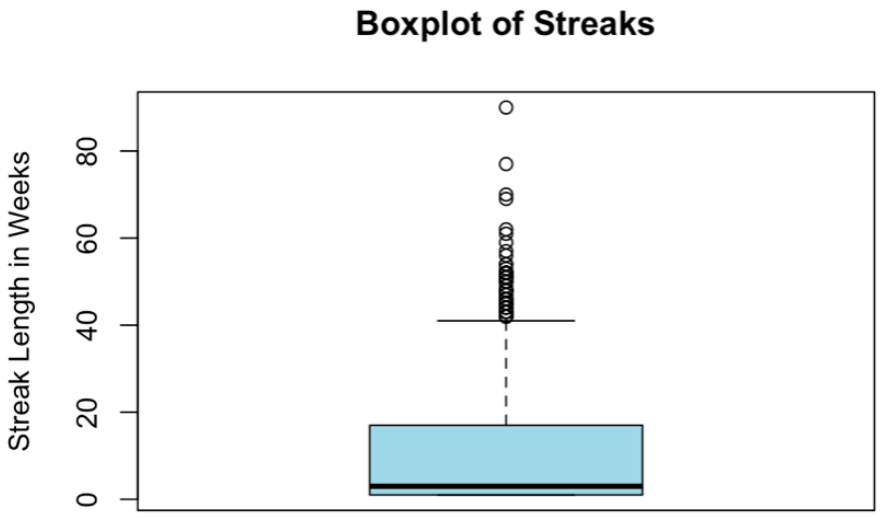


In terms of the correlation plot, we found that variables like top peak and average position were most negatively correlated with the response, as high rankings bode well for a song staying in the top 100 for a longer period of time. On the other hand, the loudness and standard deviation of position were a couple of the most positively correlated variables with the response.

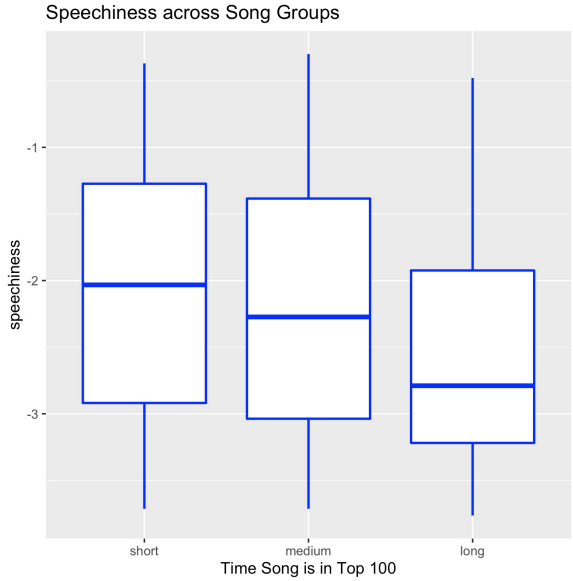


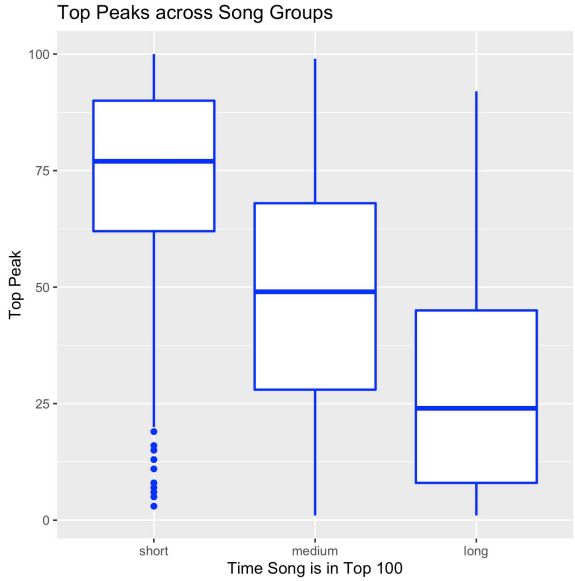
Results

The first model we attempted was the "full" model, which included all numerical predictors (including the log transformed and factor variables), and we used an 80-20 train-test split on the data so that we could train our model before evaluating it against the "new" test data. When we evaluated the model using the test set, we found an R-squared of 0.4536 and an RMSE of 8.754, which was quite high, considering the mean of the response variable was 9.667 weeks. Due to the heavily right skewed distribution of the response variable, we decided to run the same model, with the only change being a log transformation of the response. After using this new model on the test set, we found an R-squared of 0.3687 and an RMSE of 9.948, which is even higher than the previous model. So, we gave the linear regression one more try by reducing the model to 7 terms through the use of backwards stepwise regression. This model, which included valence, log(liveness), log(speechiness), top peak, entries, standard deviation of ranking, and loudness, gave the best results of the three linear regression models, with an R-squared of 0.4599 and an RMSE of 8.707. However, this RMSE was still too high to be satisfied with linear regression, which motivated us to change the scope of our question. We concluded that the heavily skewed distribution of the response variable and the few outliers with streaks above 40 weeks hindered the success of these models, as seen in the boxplot below.



As a result, we moved on to a new method of analysis. For one, we redefined the response variable as a three-level factor variable. The levels of this factor included short, medium, and long; to determine the labels of the songs, we took a look at the summary statistics of the streak variable from the first approach and used percentiles to form three groups. The short group included all streaks of at most 1 week, which was roughly the 33rd percentile. Then, the medium group included streaks greater than 1 week but at most 12 weeks, which was the 66th percentile. Finally, the long groups included streaks greater than 12 weeks. After labeling each song, we found that the groups were of similar size; the short group had 1269 songs, the medium group had 1170 songs, and the long group had 1172 songs. Thankfully, we can see through some EDA that variables like speechiness and top peak have some separation between each of the three groups, which should help us develop a more accurate model.





We proceeded to use ordinal logistic regression by using variables obtained from the stepwise regression from the linear regression attempts. The exponentiated coefficients of the ordinal logistic regression model are seen below in Table 1, along with the bounds of the 95 percent confidence interval for each predictor's exponentiated coefficient.

\begin{table}[htp]
\caption{Ordinal Logistic Regression Coefficients}
\begin{center}
\begin{tabular}{lclclcl}
\hline
Variable & Estimate & Lower Bound & Upper Bound\\
\hline
valence & 0.294 & 0.209 & 0.413\\
\hline
log(liveness) & 1.332 & 1.183 & 1.501\\
\hline
log(speechiness) & 1.672 & 1.541 & 1.815\\
\hline
peak & 1.041 & 1.037 & 1.045\\
\hline
sdposition & 0.916 & 0.906 & 0.925\\
\hline
loudness & 0.913 & 0.884 & 0.943\\
\hline

\end{tabular}
\end{center}
\label{default}
\end{table}%

With regard to the model's accuracy, we split the data into training and test sets, with 80 percent of the data going in the training set and 20 percent of the data going in the test set. After training the model on the training set of 2890 songs, we made predictions on the test set for what each song's category would be. In the end, we found this model performed reasonably well on the test set, as 65.3 percent of songs were classified correctly, which is even more impressive than usual since there are three possible categories as opposed to the usual two categories in standard logistic regression. Below, Table 2 is the three-way contingency table showcasing the results of

the model, where the rows represent the predicted values and the columns represent the actual values.

```
\begin{table}[htp]
\caption{Contingency Table}
\begin{center}
\begin{tabular}{|c|c|c|c|}
\hline
Prediction & Long & Medium & Short\\
\hline
Long & 154 & 95 & 0\\
\hline
Medium & 74 & 94 & 30\\
\hline
Short & 6 & 45 & 223\\
\hline

\end{tabular}
\end{center}
\label{default}
\end{table}%
```

Conclusion

In the ordinal regression model, two variables stood out. The log(speechiness) variable had an estimated coefficient of 1.672, which means that for every two-fold increase in speechiness, the predicted odds of the song being in the top 100 for a longer period of time increases by a factor of roughly 3.18. We get this value by choosing k = 2 to represent a two-fold increase in speechiness, and we raise k to the power of the estimated coefficient, 1.672. This gives us 2^1.672 = 3.18. On the other hand, the valence variable had an estimated coefficient of 0.294, which signifies that for every unit increase in this variable, the predicted odds of the song being in the top 100 for a longer period of time decreases by approximately 70.6 percent.

As for the contingency table, we see that the model did a great job of predicting songs in the short category correctly, with an 81.4 percent success rate. The model also did quite well predicting songs in the long category correctly, with a 61.8 percent success rate and zero incorrect predictions that were short songs. However, the model struggled the most when predicting songs in the medium category correctly; the success rate here was only 47.5 percent. This result makes sense though, as this is the middle group of the three and can lead to some borderline cases that are difficult to differentiate between the three categories. A potential remedy for this issue in the future could be to split the medium group into two, but even this strategy will not be able to remove all ambiguity.