**Final Report**

**Factors contributing to a top song**

*Blanche Chung*

**Part 1: Introduction to the problem**

Executives in the music industry want to know the formula for a top ranking song. The question is, why do some songs make it to the top while others fall behind? We want to uncover what are music listeners true preferences and what gravitates them towards a song.

Producers and music labels want to know what to invest in: is it more important to have a top producer or a well known artist? How long should a song be, and what tempo do listeners prefer?

Our team was hired to see what music executives should consider when producing albums and contracting artists. We will look at the top 200 songs of the year, and discover why top songs are at the top, while other songs fall to the bottom 200. Is a top ranking song more likely to remain on the charts for several weeks, or be a short lived hit? After our analysis, we will be able to present music executives with the formula for a perfect song, so that they can see a high return on their investments.

**Part 2: Exploratory data analysis**

There were three types of plots which we utilized in this analysis:

1. Correlation Plot
2. Histogram
3. Box plot

We are using the dataset “Spotify top chart songs 2022” dataset from Kaggle. In this dataset, there are 646 observations in the dataset and 17 columns. We dropped the column “URI”, as it is not useful to the analysis. We noted our target variable “peak rank” isn't continuous, and therefore that led us to decide that we need to do a categorical analysis, since there are 200 possible values for this variable. We decided to “peak rank” it into 3 groups: “Low” (127 observations), “Medium” (233 observations) and “High” (286 observations).

We started the exploratory process by creating a correlation matrix. This matrix reveals that many of our variables are closely related, such as loudness, energy, danceability and acousticness, however there does not seem to be a noticeable correlation between peak rank and any of the target variables. Based on the multicollinearity, we are anticipating the statistical value of a regression model to be weak.

We also observed the median values for factors such as “danceability”, “energy”, “speechiness” etc., and wondered if top songs are above or below average values. We created exploratory histograms looking at the distribution of each variable based on “high”, “low” and “mid” songs to see if there is any clear difference in skew of data. It does look as if songs in the top 200 skew towards larger values for loudness, liveness and tempo. Based on this exploratory analysis, we predict that these variables will influence song ranking and top songs will have slightly higher average values than the average for the total data.

**Part 3: Choose data mining methods, reach final model, compare model performances**

Since we are working with categorical variables, we have to look at classification models including classification trees/ random forest bagging, clusters and logistic regression. We explored all of these methods to determine which tells the best story.

Logistic Regression: We want to build a model that will predict whether a song will be a top rated song. For this model, insteading of dividing the data into 3 possible outcomes, we reduced it to 2, “High” and “Low”. “High” will use data where peak\_ranking is between 1 and 40, and “Low” will use peak\_rankings between 41 and 200. We will also use 12 inputs as variables (Danceability, Energy, Key, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Tempo, Time\_signature, Duration\_ms, and Loudness. We performed a logistic regression model using 5-fold Cross-Validation and 10NN. The 5-fold Cross Validation logistic regression model had an accuracy of 65.43%. The 10NN had an accuracy of 56.24%.

Clustering: We used clustering to see if there were visible relationships between a “top song” and each variable. We used both K-means clustering and hierarchical clustering.

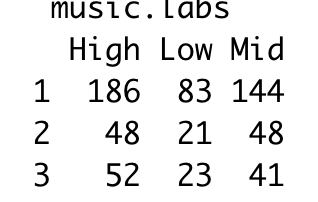
1. **Hierarchical clustering:** We explored Hclustering to see dendrograms and explore values of k. We explored using “average” “complete” and “single” methods. It looks like average and single divide the observations the best *(Figure 1).*

We then plotted the clusters to see if there was a clear distinction between them. There don't seem to be any clear distinctions between the clusters and 2 variables *(Figure 2).*

We did a boxplot and saw that the most evident differences in variables are loudness, ms\_duration and speechiness, running the clusters again, there are no changes *(Figure 3).*

1. **K-means clustering**: we sought to partition the observations into a pre-specified number of clusters. To perform k-means clustering, we first specified the desired number of clusters K, then used the K means algorithm to assign each observation to one of the clusters (3 clusters) .

* *Results of the cluster*
  + We see a high distribution of songs ranked in the third cluster; a large distribution of “high songs” in the second cluster and “low and mid” songs in the third cluster. The table (*Figure 4)* shows initial k means clustering results using all of the variables.
  + We looked at the sum of squares, we know if they are closer to zero, the model has a more accurate fit: km.out1$betweenss: 1.456579e+12. km.out1$totss: 1.948764e+12 . These results aren't too close to zero. In clustering, we see certain variables help to differentiate among the song rankings; we are going to isolate the variables “liveness”, “tempo” and “loudness”.
  + We scaled this version of the model, and saw that the third and second clusters now have fewer observations and the first now have the most. The sum of squares for this model are huge, and therefore this model is not a better fit.



> km.out4$tot.withinss [1] 1103.417

> km.out4$betweenss [1] 831.5827

* We then looked at an unscaled version of this same model with the 3 variables. We see that the sum of squares are basically identical to the original model
* Lastly, we wanted to see if the optimal amount of clusters is correct (Figure 5).

Bagging and Random Forest: After dropping the categorical variables such as artist\_name, track\_name etc. we created training and testing datasets. We proceeded to perform random forest using all predictors or bagging on the dataset and evaluated the performance of bagging. We ended up with a mean value of 2266.

Next, we decided to proceed with the same methodology but this time, performing random forest and considering only 5 variables instead of all. In this case, the mean value was still very high: 2249. When looking at the measures of the variable importance, it appears that besides weeks on chart, the duration, loudness and instrumentalness are significant when it comes to peak rank as shown in Figure 5.

Being that the peak rank takes values between 1-200, we thought the prediction is not as accurate as we would like and decided we would also try random forest but this time considering values only 1-100. By following the same procedure the mean value resulted to be much smaller 937. In addition, as shown in Figure 6, this time energy, danceability and loudness are more significant.

**Part 4: Model Performance and Final Model**

We created a logistic regression model using 2 different sampling techniques. We found the 5 fold cross-validation method gave us a more accurate method. The accuracy was approx 65% compared to the 56% of the 10NN logistic regression model. Although our 5 fold model was better than our 10NN, we ultimately decided it was not the best model to predict whether a song will be ranked in Top 40. Our cluster analysis did not show a strong relationship between song ranking and any of the variables. Both KNN and Hierarchical clustering did not change too much when including our predictors into the model. In our H CLUST, we do see that there is a slight improvement in the way that observations are classified, however this is disproved by the KNN clusters. In our KNN cluster, when we isolated our predictor variables, the means increased significantly, revealing that it was a less reliable model. Overall we are not confident that this model should be used as the final model for the analysis.

**Final Model - Random Forest:** the performance of the model on the original dataset was not as good as we expected which led us to retry by using a smaller range of values as mentioned in the section above.

**Part 6: Findings, conclusion and recommendations**

**Is there a formula for a perfect song?**

Our Random Forest model supports the claim that loudness is the contributing factor to a song being highly ranked. We also see that energy and danceability contribute to high ranking.

In order to create a hit, make sure you are prioritizing **loudness, high energy and danceability.**

When producing a song aim for the below ranges..

Loudness >= -6.307 decibels

Energy >= .63 points

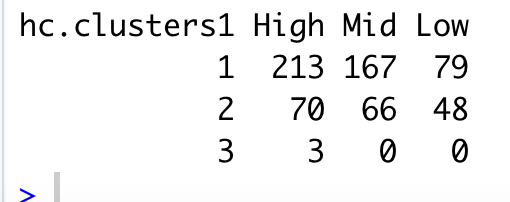
Danceability >= .68 points

These ranges are based on the median and sd values for these variables in top ranking songs.

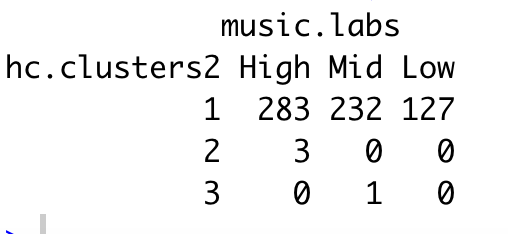
**What to consider in the next iteration of analysis**: We would like to factor in the artist to the next version. Since this was a categorical variable, it was very complicated to classify these values and include them in our analysis; however, we do think it would be a worthwhile move because in our exploratory analysis, we see that certain artists have multiple top ranking songs.

**Appendix**

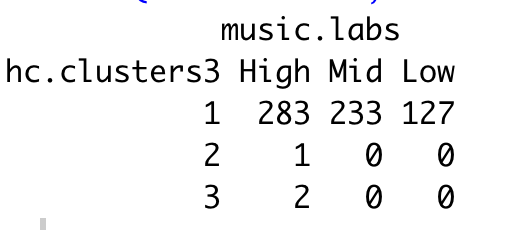
*Figure 1:*



*Complete linkage analysis*

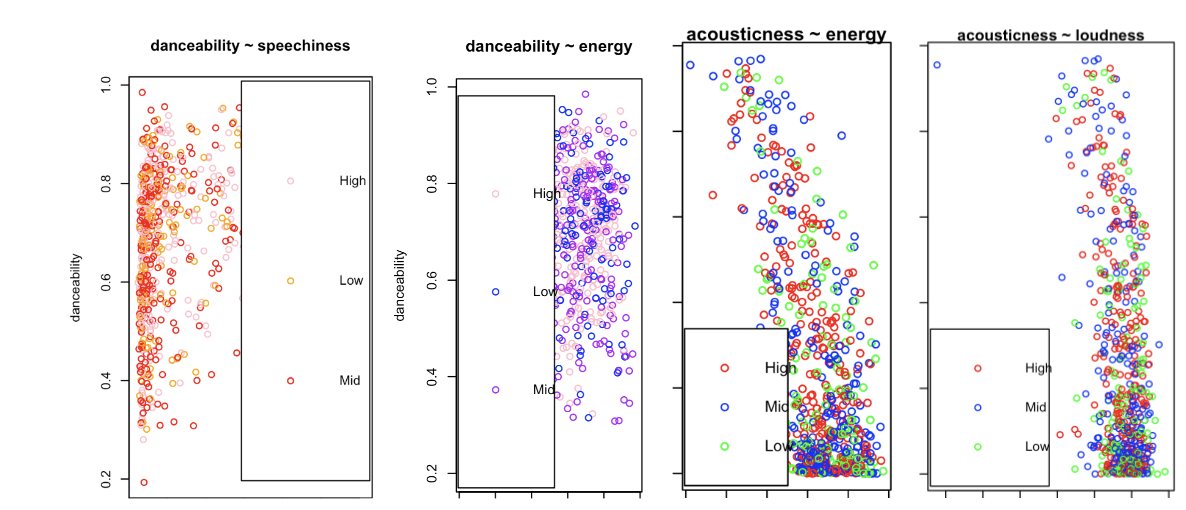


*Average Linkage Analysis*

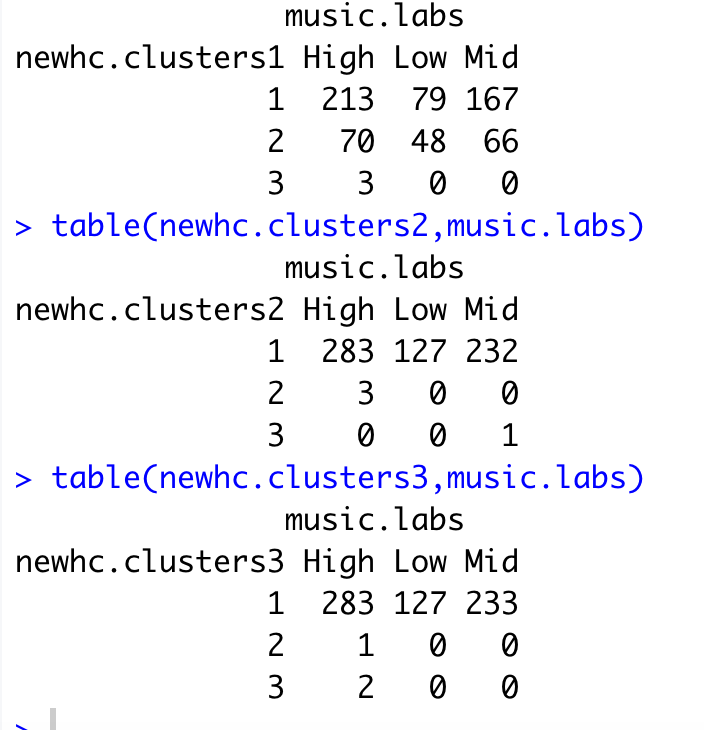


*Single linkage analysis*

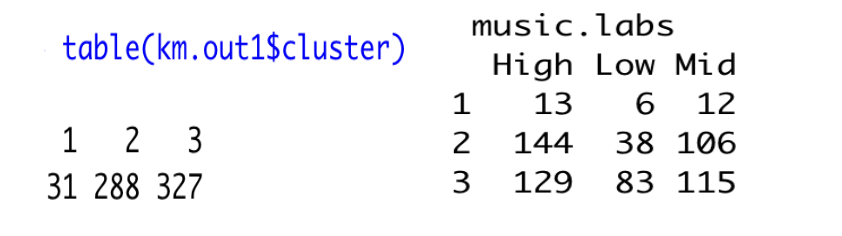
*Figure 2:*



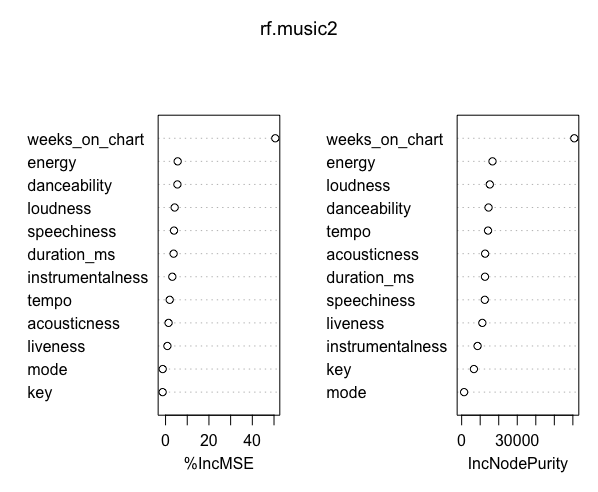
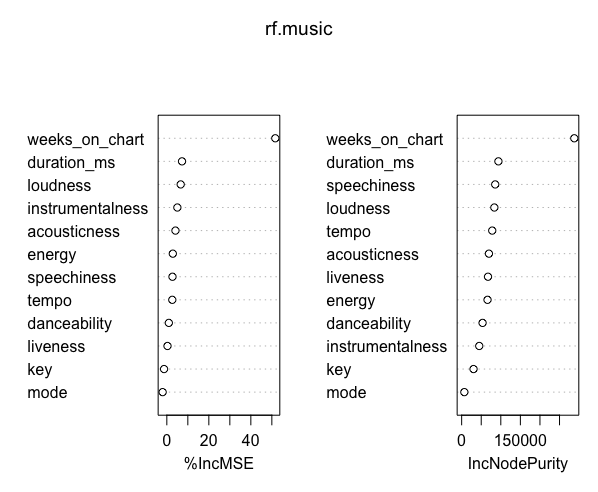
*Figure 3:*

****

*Figure 4:*



*Figure 5 Figure 6*



*Figure 7*

