MSIS 510: Introduction to Data Mining and Analytics

Group Project

OSWALD
RECORDS

The secret formula for the next hit artist on Spotify

By: MSIS Purple 2

Bahnny Das, Laura Bettinger, Malavika Nandakumar, Nick Griffin, Partha Ray, Yen Phung

Table of Contents

Project Goal

Oswald Records is an established record label with a strong history of artist management and content creation. In addition to the traditional processes, such as using Artists and Repertoires (A&Rs) that scout new musical talent and undertaking market research, we recently expanded our capabilities to include a business intelligence team focused on analyzing data to better understand demographic and social trends correlated with hit music at the top of the billboards. With increased competition today, the market is saturated, especially with the increased popularity of streaming services. Our analysis has determined the influx of tracks and artists into the market threatens our current business model. Moreover, any artists' ability to upload songs on a streaming service – regardless of popularity or name recognition – indicates a necessary change in Oswald Records' strategy.

This project aims to identify the characteristics that make a track popular on a streaming service such as Spotify, empowering our company to advise our artists and produce music that earns us the maximum return on investment through increased popularity and a larger number of streams.

More importantly, we are seeking to provide the answer to these questions:

1. *What are the characteristics that an artist's tracks must have to be popular on Spotify?*
2. *What time of year is best suited to release particular types of music?*
3. *What is the optimal genre of music to be released in 2021?*

Dataset description

For this exercise, our team selected the Spotify Dataset on Kaggle[1] which includes data from over 160 thousand songs / tracks released between 1921 to 2020. This can be referenced here. This data was extracted via the Spotify Web API and presented in multiple CSV files. Our team focused on the data.csv, data_by_artist.csv, and data_by_genres.csv files. There was a total of 202308 observations, and we summarize the variables we have focused on below. Our target variable for this exercise is **popularity**. As per Oswald Records' business requirements, and after analyzing the popularity of numerous top 1000 songs, we are designating a popularity of 70 or above as the score to identify a track as popular, 60 or above for an artist.

Field Name	Description	Type	Range / Sample values
Id	Id of track – assigned by Spotify	Numerical	
acousticness	Measure of how acoustic a track is.	Numerical	0 to 1 1 is the most acoustic
danceability	Suitability of a track for dancing, based on tempo, rhythm stability, beat strength and regularity.	Numerical	0 to 1 0 is least danceable.
energy	Perceptual measure of intensity and activity	Numerical	0 to 1 1 is highest energy.
duration_ms / Avg Song Duration	Duration of track in milliseconds	Numerical	200K to 300K
instrumentalness	Measure of whether a track contains vocal content	Numerical	0 to 1 1 is the most instrumental, with no vocal content
valence	Measure of how positive the music is (happy / cheerful)	Numerical	0 to 1 1 is the most positive

[1] Spotify Dataset 1921 – 2020, 160K+ Tracks Version 9 by Yamac Eren Ay

Field Name	Description	Type	Range / Sample values
tempo	Pace of the music measured in beats per minute	Numerical	50 to 150
liveness	Presence of an audience during the recording	Numerical	0 to 1 .8 or above indicates the track was recorded live
loudness	Loudness measured in decibels, averaged across the entire track.	Numerical	-60 to 0
speechiness	Presence of spoken words in a track.	Numerical	0 to 1 .66 or above are made entirely of spoken words
year	The release year of the track	Categorical	1921 to 2020
mode	Modality of the track/ type of scale for melodic content	Dummy	0 = Minor, 1 = Major
explicit	Indicator whether a track contains explicit content	Dummy	0 = No Explicit content, 1 = Explicit content
key	Overall key of a track	Categorical	0 to 11, 0 = C, 1 = C#/Db
artists	The artist for the track	Categorical	list of artists E.g., Foo Fighters
release_date	The exact date the track was released	Categorical	YYYY-mm-dd
release_month	The exact month that the track was released in. This was created during our pre-processing steps	Categorical	1 – 12 1 is January
name	The name of the track.	Categorical	
genres	The genre the track belongs to.	Categorical	Genre name E.g., Alternative Rock
popularity	Indication of the popularity of the Track.	Numerical	0 – 100

Data preparation details

In order to prepare the data for our mining efforts, we undertook the following pre-processing steps:

1. All Datasets:
   a. Removed NULL values.
   b. Data transformation applied:
      1. Normalization, discretization, Concept Hierarchy Generation.
      2. Changed duration to seconds and changed column name to "Avg Song Duration".
2. data.csv: (Observations- Before processing: 170654, After processing: 118188)
   a. Removed all rows with incomplete release dates.
   b. Extracted month from release_date field and added column release_month.
   c. Removed all non-alphanumeric characters from track name and artist name.
   d. Removed rows with empty track names.
   e. Removed the identifier column (id).
3. data_by_genre.csv: (Observations - 2973)
   a. Removed rows with empty genre values.
   b. Removed all non-alphanumeric characters from genre.
   c. Combined terms within a genre.
4. data_by_artist.csv: (Observations- Before processing: 28681, After processing: 27262)
   a. Removed non-English terms from columns.
   b. Removed rows with missing data

Data Visualization and Exploratory analysis

We structured our exploratory analysis efforts first by validating the playing field, i.e., the competition. The histograms below show that the distribution of popular tracks/artists gets much lower around the 70 + mark. This indicates the industry's competitive nature, as very few tracks/artists are rising to that level.

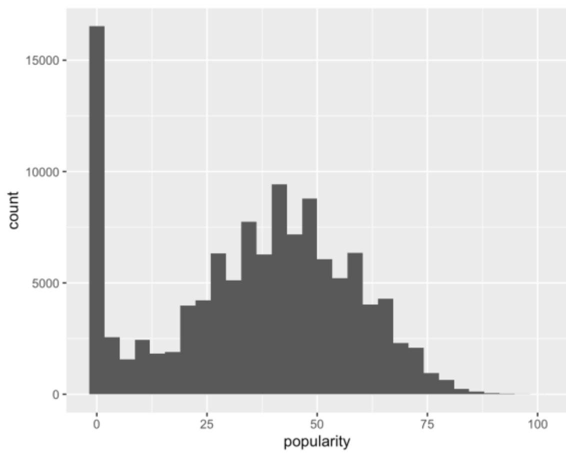


Figure 1: Distribution of Popular tracks

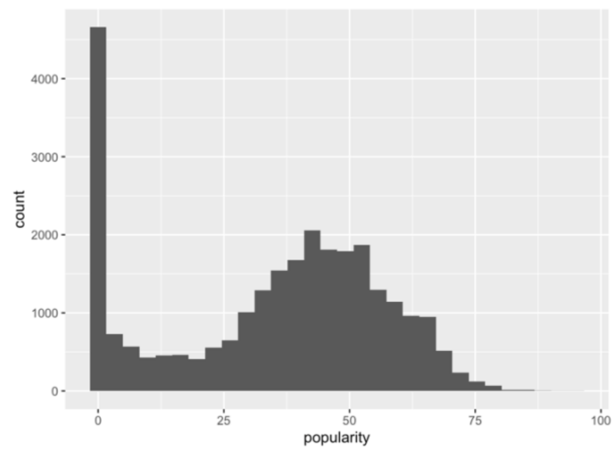


Figure 2: Distribution of Popular artists

Next, we have dived into a better understanding of how the various features of tracks have evolved in music released from 1921 till the present day, and then comparing them with the popularity of tracks listed by Spotify. We have done this by selecting all the features associated with tracks, and then used line charts to identify trends in the change of these features over time and the impact on popularity.

To understand how the average valence, acousticness, danceability, energy, liveness, and loudness for all songs on Spotify changed over time relative to average song popularity, the team plotted those data elements in a single line chart, as seen in Figure 3, before diving into each feature individually.

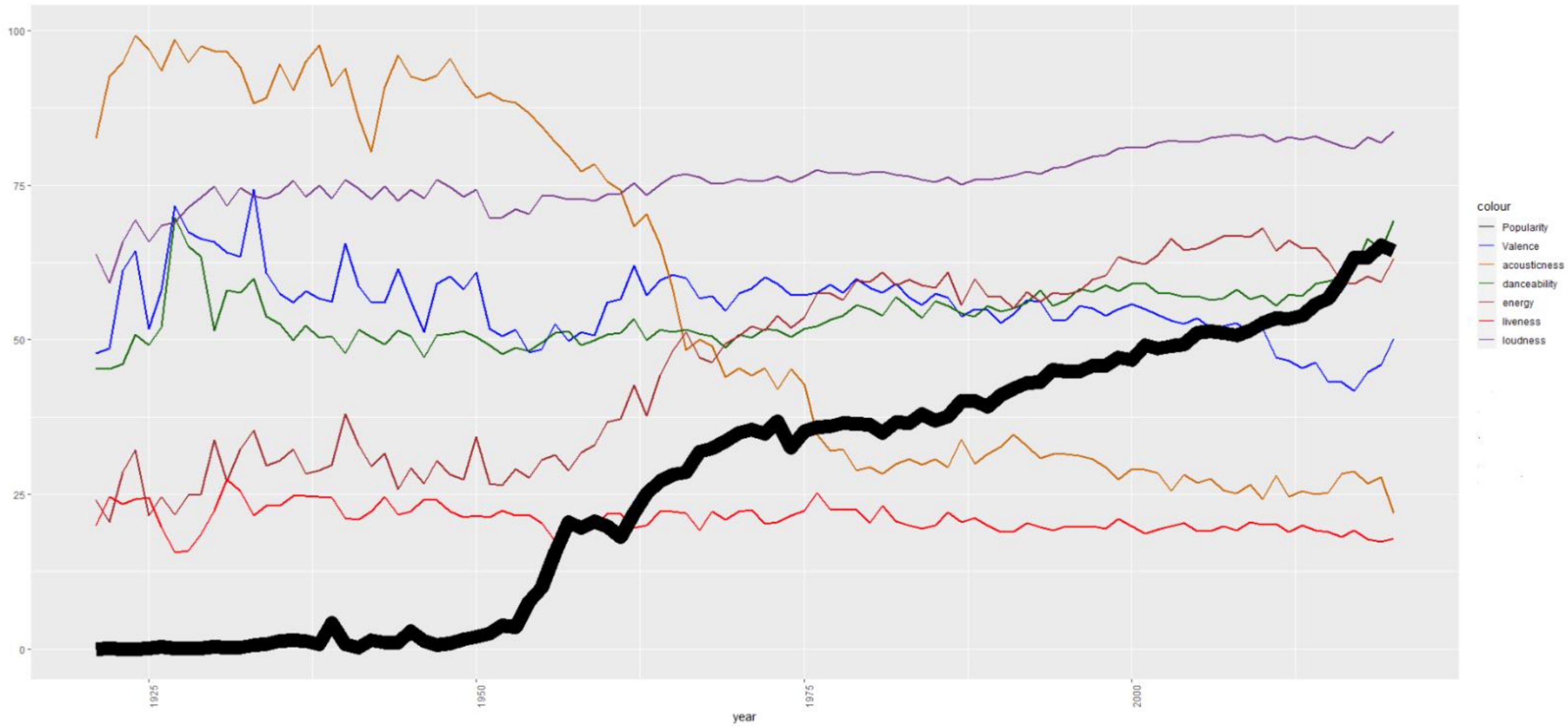


Figure 3: Average popularity and plotting valence, acousticness, danceability, energy, liveness and loudness for all songs on Spotify based on release date

Figure 3 shows that popularity of songs is much more for music released recently, compared to older release dates. Our team believes the increase is due to more music being created, especially due to the advent of technology that enabled and simplified music production and the easy availability of streaming services such as Spotify. This observation is reinforced by Figure 4, which illustrates the number of songs on Spotify by year created.

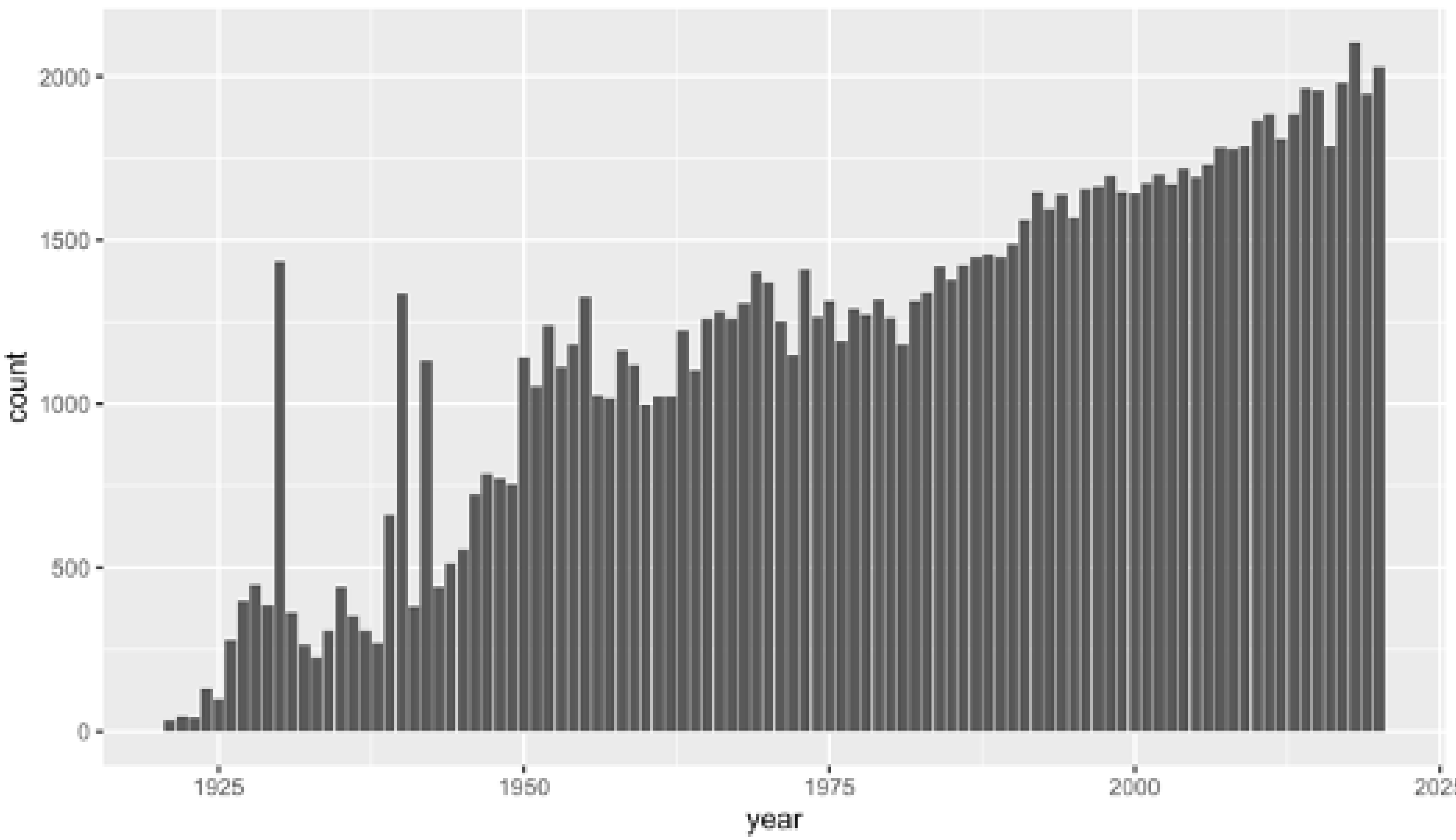


Figure 4: number of songs on Spotify by year created

As per figure 2, we can see that the average valence, acousticness, danceability, and energy of Spotify songs have changed significantly as average popularity has increased. Average loudness and liveness have not changed significantly. To explore these song features further, they are each broken out in separate line charts in Figure 5 below.

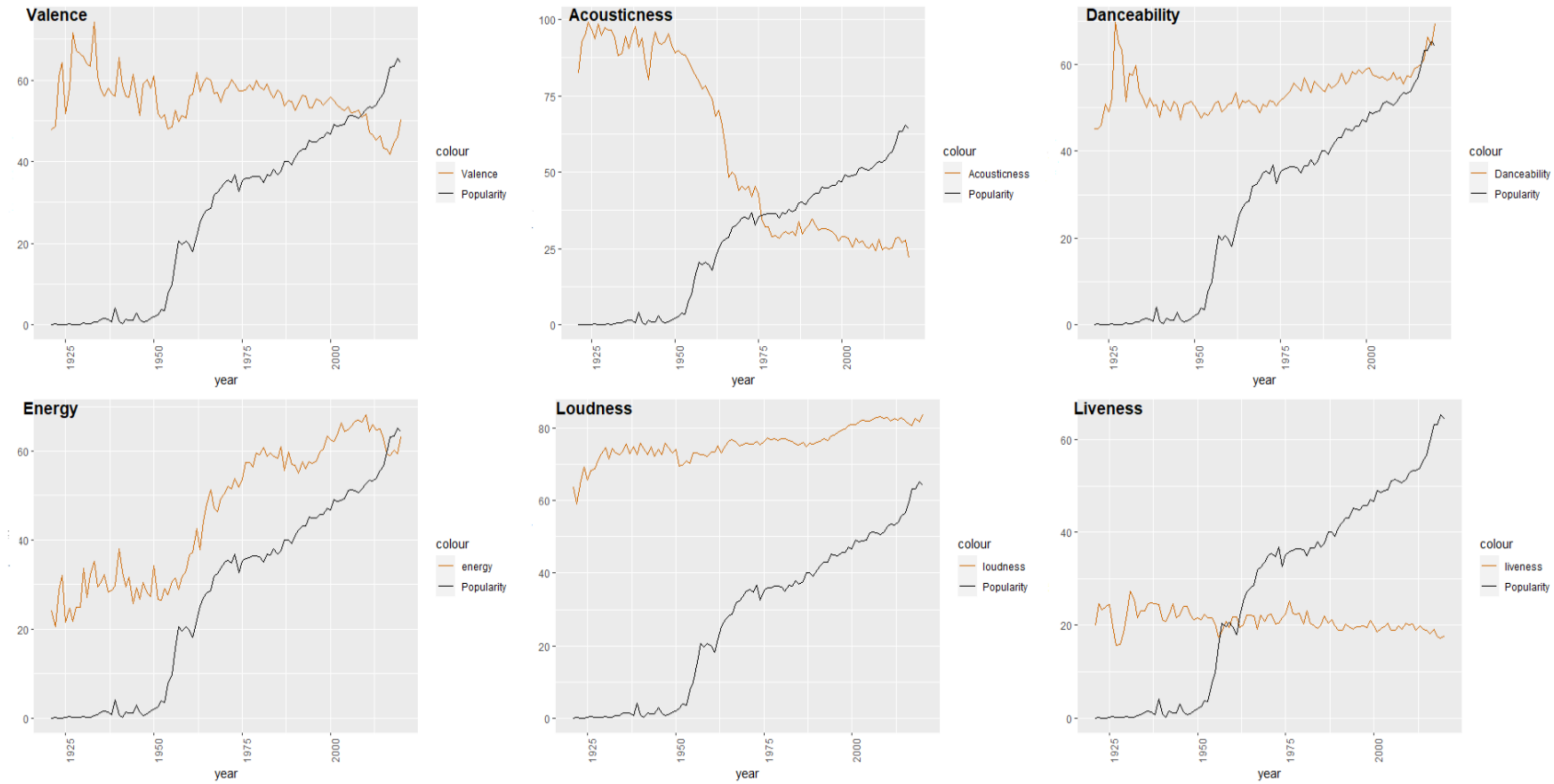


Figure 5: Deep dive of average popularity valence, acousticness, danceability, energy, liveness and loudness for all songs on Spotify changed over time

Figure 5 shows that average song valence and acousticness have dropped as song popularity has increased. There isn't a significant change for average song loudness and liveness.

Average song danceability and energy, on the other hand, have increased as popularity has increased. Notably, from this analysis, we can see spikes in danceability for songs released during challenging times, such as the Great Depression in the 1930s and the 2020 COVID pandemic.

Additional song features in the Spotify dataset were analyzed and found to not significantly correlated with popularity. Data visualizations for these features can be found in Appendix A. We also looked at the popularity of songs by month released. The findings from this analysis are captured in Figure 6. From our analysis, we were able to determine that tracks released in January and December are, on average, less popular than songs released in other months.

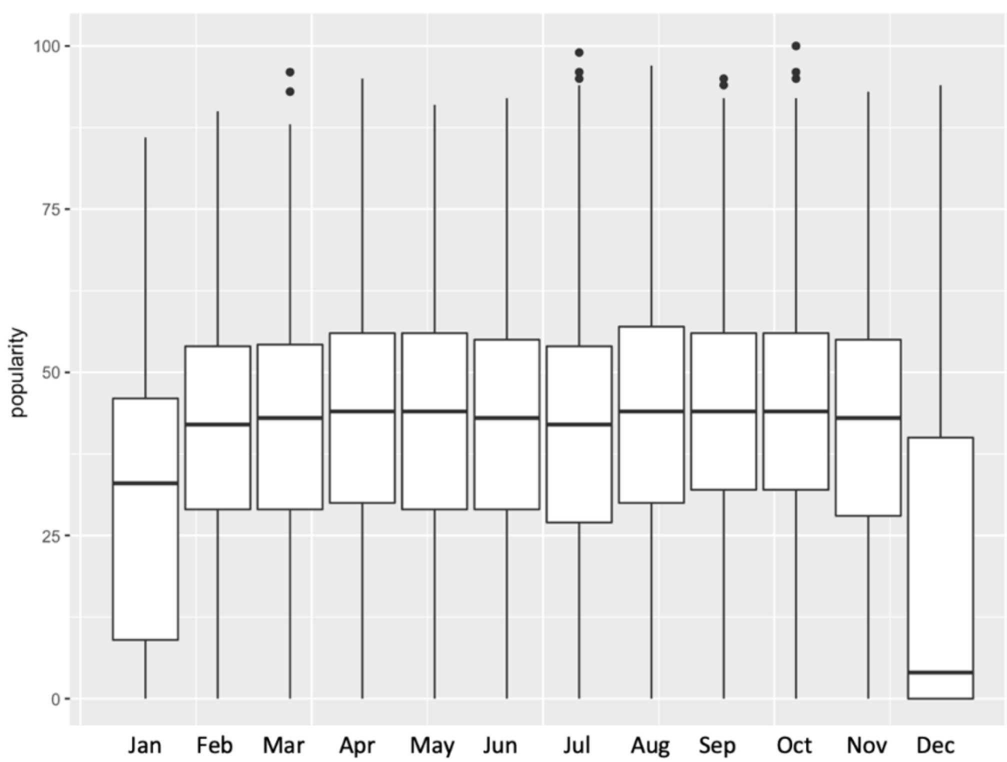


Figure 6: Song popularity by release month

The team was also curious what the most popular genres and song titles are on Spotify. As structured data was lacking, we used text analysis to identify trends, as seen in Figures 7 and 8. From our analysis in Figure 7, we learned that love is the most popular word for song titles.

Figure 7: Popular words in song titles

From the analysis in Figure 8, we learned the movie-tunes, pop and show-tunes are the most popular genres in Spotify.

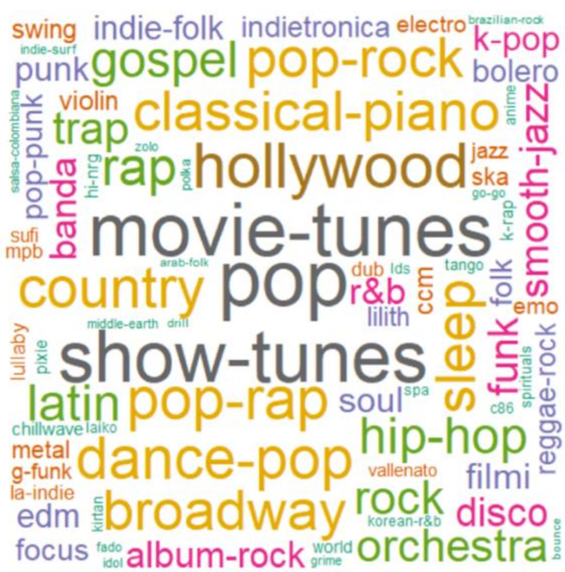


Figure 8: Popular genres

Data Modeling and Methodology

At this point in our analysis, we will delve into setting up Data models needed to answer the key questions we outlined at the introduction of this report.

Q1: What are the characteristics that an artist's tracks must have to be popular on Spotify?

To answer this question, we set up a logistic regression model using the artists datasets. Also, we converted popularity into a categorical feature with a value of 1 if it is > 60 for artist data sets. The training dataset was set to 60 percent of the pre-processed data, which amounted to 16357 rows. We also ran a correlation matrix and a CORR plot (See Appendix C).

The correlation matrix indicated a strong relationship between energy and loudness and a negative correlation with acousticness. However, acousticness and loudness themselves are not correlated. This information guided our decision to remove energy from our analysis, as it is a redundant variable. The CORR plot below confirms our suspicions.

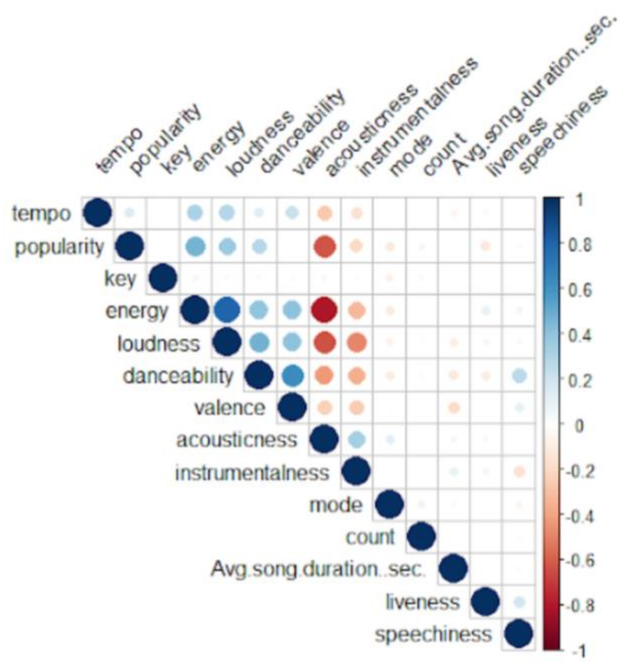


Figure 9: Corrplot to show correlation of variables

After this, we ran a logistic regression to come up with the following (Appendix C):

Feature	Coefficient	Interpretation
Valence	-4.034	This analysis further solidifies the negative correlation between positivity and popularity
Acousticness	-1.402	Confirms that more acoustic tracks are less popular
Danceability	4.009	Further confirms the importance of danceability
Liveness	-0.711	Liveness has an impact on popularity for artists, but we could not find correlation within the songs dataset. Hence, analysis for Liveness is inconclusive
Instrumentalness	0.507	This feature had a negative impact for songs, but a positive impact for artists. Hence analysis for Instrumentalness is also inconclusive
Loudness	0.045	Minimal impact, but confirms that loudness plays a role in popularity
Mode (Major)	-0.320	Further confirms that Mode 1 (Major modality) tends to be less popular
Speechiness	-0.133	Further confirmation that speechiness has a negative impact on popularity
Tempo	0.0003	Further confirms that tempo is not a relevant variable for our models
Keys (1 to 11)	Range -0.39 to 0.106	The wide range indicates that further analysis will be required
Average Track Duration	-0.009	Similar to before, longer tracks can have a minimal impact on popularity
Count	-0.003	Number of songs that an artist has created, but with such a minimal impact to popularity, we have concluded that this variable can be excluded

Running a prediction based on this model and evaluating the confusion matrix, we get an accuracy of 75.5%. Setting an optimal threshold of 0.12 gives us a sensitivity of .732 and a specific of 0.758 (See Appendix C). After this step, we created a decision tree for predicting whether a song/artist would be a hit or not. The results are shown below.

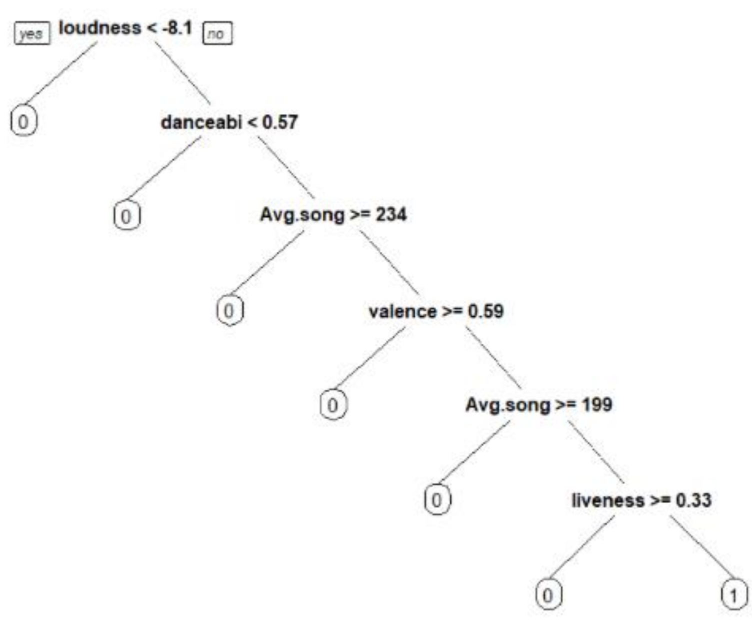


Figure 10: CART for hit song prediction

Based on our analysis, we came to the following conclusions for an optimal track an artist must produce to be popular on Spotify.

Feature	Impact on Track popularity	Conclusion for popularity
Valence	The models suggest a negative impact on popularity, to a much more significant extent when artists are generating	There is no need for overt positivity in a track. The CART suggests less than .59.
Acousticness	Acousticness has a negative impact overall	Avoid acoustic tracks.
Danceability	Has a significant positive impact on popularity.	Ensure tracks can have a high danceability score. The CART suggests more than .57.
Explicit	Positive impact, but not to a significant extent	Having explicit lyrics does not detract from popularity.
Loudness	Has a positive impact on popularity, but not to a significant extent	Ensure that a song is not below a certain level of loudness. The CART suggests not to go below -8.1.
Mode (Major)	Has a negative impact on popularity, but not to a significant extent	Avoiding Major modality and focusing on Minor modality is an option.
Speechiness	Negative impact on popularity	Ensure there is less spoken words in a track.
Instrumentalness	Positive impact on popularity	Having songs with instrumentalness increased popularity here, but as per Appendix A, purely from the songs dataset, there was negative correlation. Hence the analysis for this attribute is inconclusive.
Average Track duration	Has a small negative impact on popularity.	Since the CART shows Average duration as a node, there are sufficient records to consider this as having an impact. Hence the CART suggests below 234 seconds.
Keys	Negative or Positive impact on popularity – inconclusive from logit models.	To be further analyzed below.

Now on to our second question.

Q2: What time of year is best suited to release particular types of music?

Since music is a creative field, sometimes it may not always be possible to stick to prescriptive features to generate tracks. This creates an opportunity for Oswald Records to maximize our investment return by ensuring we release music at an optimum time during the year.

Additionally, we wanted to see if the inconclusive attribute from our analysis above has further impact. Hence, we created a classification tree after running the following preprocessing steps:

1. Excluded year, artist name, energy, and release_date as these do not have any analytical value or are redundant.
2. Converted month, explicit, key and mode as categorical values.
3. Set release month as the outcome variable.
4. Taking a subset of only those songs that have a popularity value > 70.

The resulting tree (shown in the next page), based on a track's attributes, provides guidance on when best to release a track during a year to ensure a popularity over 70. We also see that in this instance, the key of a track place a significant role.

For example, it suggests that if a song has a key value of 3 (D# or Eb), the best time to release it is in November. Similarly, if a song has a key value that is 8 (G# or Ab), additional decision points around acousticness, danceability, and song duration come into play. To further explain this, if the song with the key value of 8 has a song duration less than 252 seconds and a danceability less than .41, then the ideal time to release it is in February.

To further validate the CART, we ran a confusion matrix (See Appendix D) which reported an accuracy of 83.62%, with comparatively high Sensitivity scores throughout, except for April (0.491), and comparatively high Specificity scores except in September (0.268). Hence, we are confident that this CART can be used to make decisions on when to release a track on Spotify.

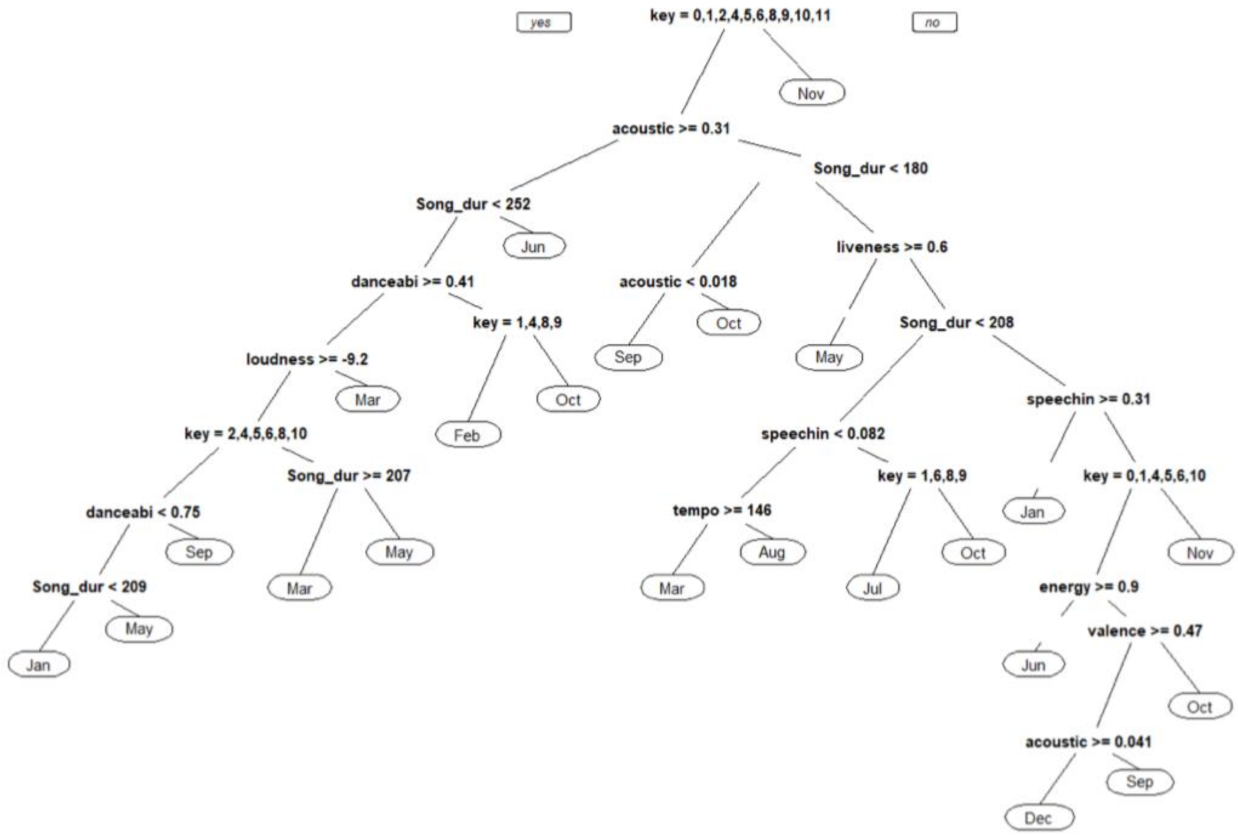


Figure 11: CART depicting model to determine ideal month to add to Spotify

Now on to our third question.

Q3: What is the optimal genre of music to release in 2021?

Genre plays an essential role in Spotify's classification of songs; hence selecting the right genre with the features mentioned above increases a song's chances of being added to the right playlist giving it more visibility. This allows for more streaming opportunities and hence more opportunities to ramp up a popularity score. To help identify the genres that have a high popularity and exhibit features similar to the ones discussed above, we ran a K-means cluster analysis. First, we ran the following pre-processing steps on the data_by_genre dataset.

1. Selected Genre as the target and all the other columns as predictors.
2. Normalizing the predictor values by subtracting mean and dividing by standard deviation.

Our next objective was to determine the best k, after running the following plots with multiple options.

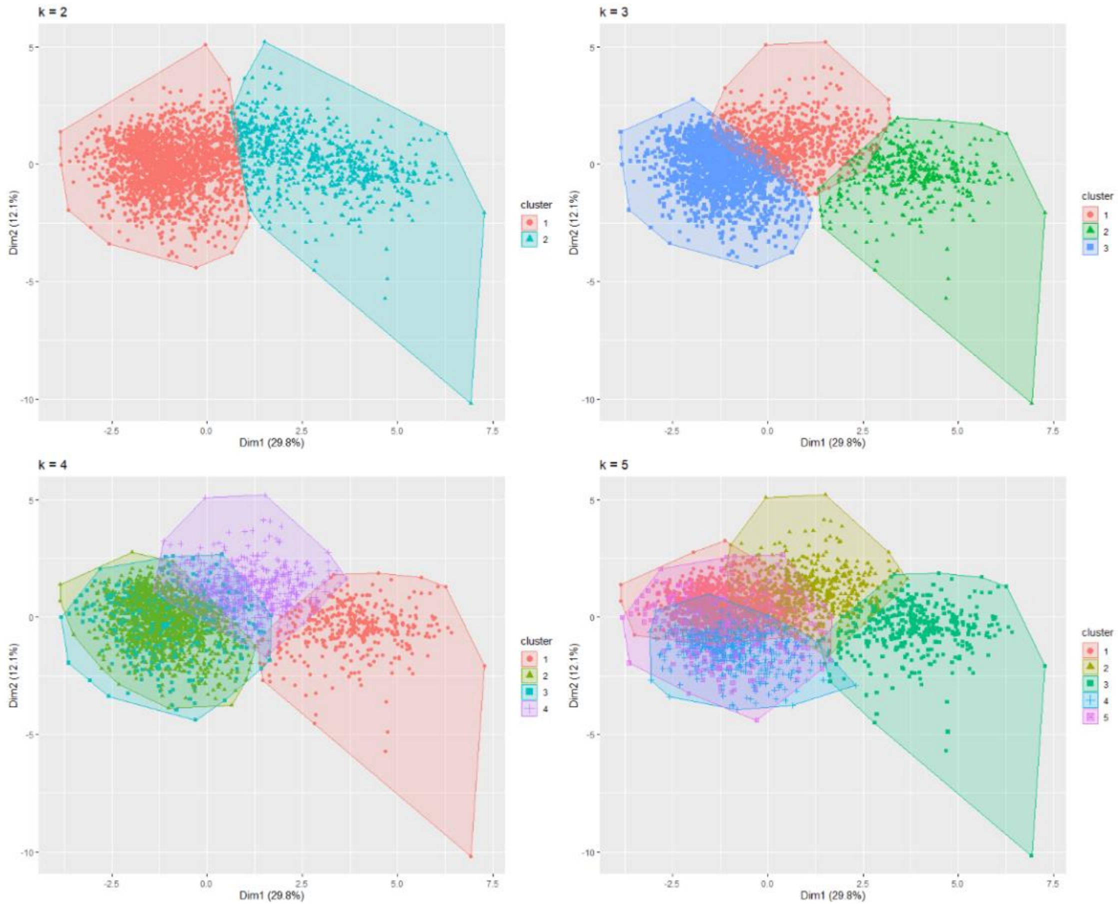


Figure 12: K-Means Clustering with multiple K values

Based on the findings above, we can see that as the value of K increases, the overlap between clusters also increased. Hence, we choose k = 2.

Running K-means clustering with k = 2 and plotting each cluster's characteristics by referring to their km centers as per below (Also see Appendix E), we can see that Cluster 2 has a higher popularity, and the features of this cluster corresponds to low acousticness, instrumentalness, and, high danceability, valence, loudness, and tempo. We were then able to review this with the word cloud for popular genres identified and found alignment.

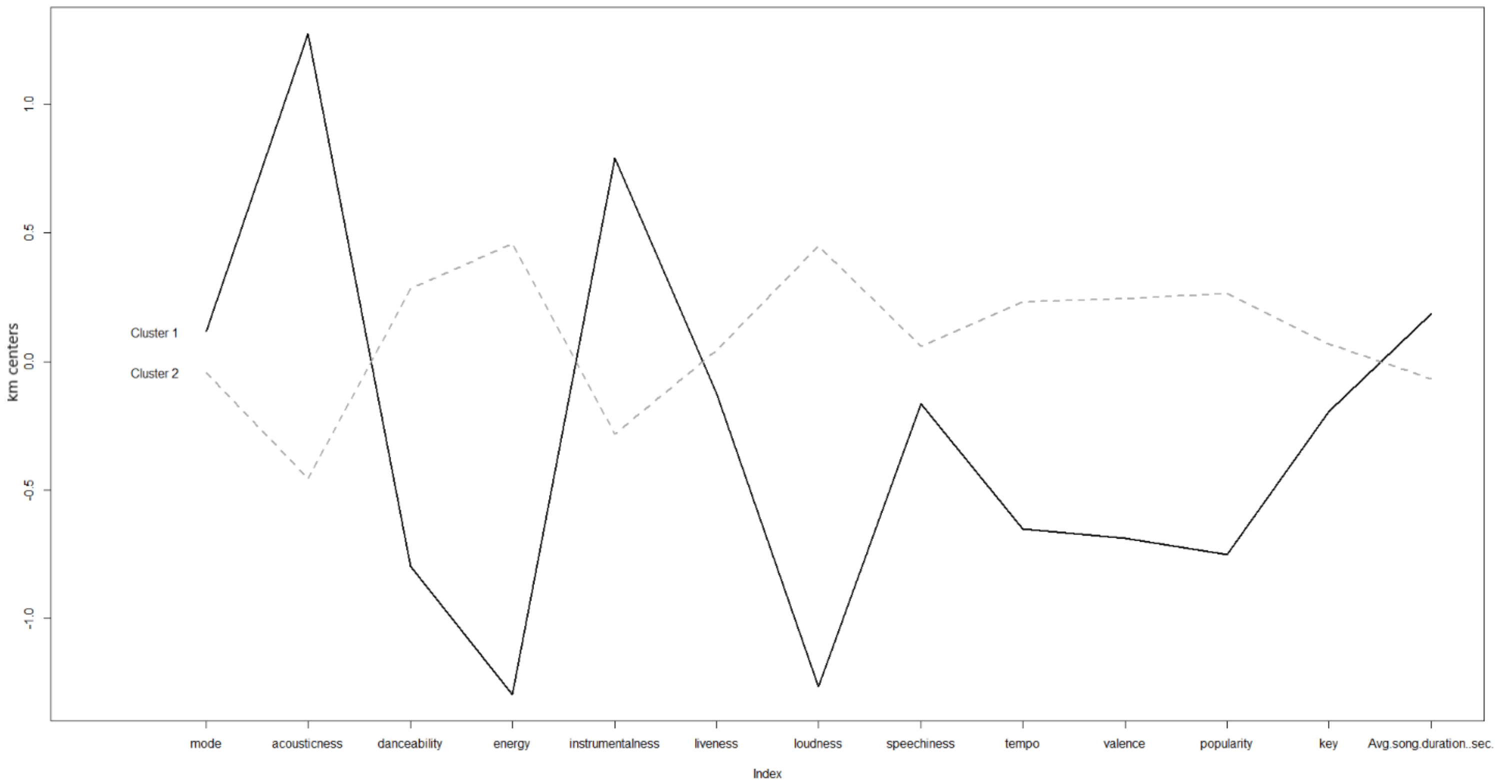


Figure 13: Plot to show features for each cluster

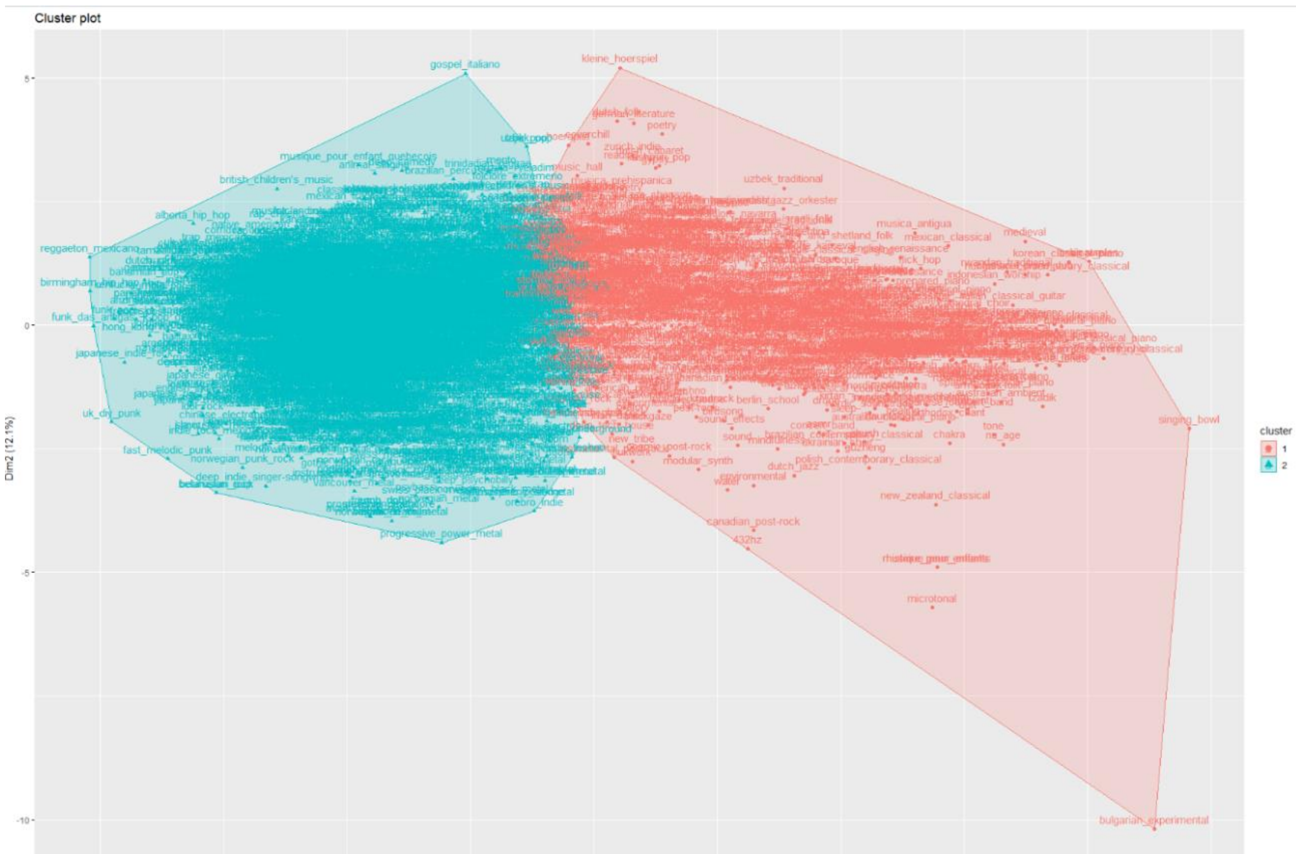


Figure 14: Clusters showing actual Genre names

Figure 15: Word cloud to indicate popular Genres

Based on our analysis, the most popular genres we would expect in 2021 would be the ones depicted in the word cloud, and some key ones to highlight are - Pop, Movie Tunes, Show Tunes, Rap, Trap, Latin, Dance-pop, Hip-hop, etc. A point of interest, the genre of "Sleep" has also appeared here, which provides unique insight as to how Spotify is used by people for assistance with sleep!

Conclusion

In summary, our team has found that there are a couple of factors that increase the probability of producing a hit song: the song has to have key characteristics such as high danceability and low valence, be released on an optimal date, contain "love" in the song name, and be in the pop, show tunes or movie tunes genre. The rules for determining the key characteristics for a hit song are shown in Figure 10, and the rules for determining the optimal release schedule are shown in Figure 11.

The biggest challenge we faced was handling this massive dataset with 5 different files. Understanding each dataset's significance was crucial to identifying the project's limitations and narrowing down the scope to suit our requirements. Exploring the raw format data, reading the descriptions, and understanding the meaning of each terminology used in the column names helped us clear the doubts around the initial hindrance.

We can also suggest the following scope for improvement when it comes to the datasets themselves:

1. A column for genre in the song dataset would have provided more definitive conclusions about our analysis. It would have helped validate our assumptions about the relationship between a song and its release month.
2. A dataset containing lyrics of the top songs was not readily available and had to be collated from multiple sources. A solid dataset about the same could extend the scope of text mining and help us gain more insights about a song's contents and its popularity. However, we could not include lyrics in our analysis due to this issue.

3. Information about customer ratings would help us validate the significance of the popularity score, whether the score has been influenced by too few ratings or not.

Lastly, the team tested the framework above with a team favorite song: Dakiti by Bad Bunny.

Key characteristics of this song are:

Valence	Genre	Danceability	Energy	Explicit	Key	Liveness	Loudness	Release Date	Speechiness	Duration
0.145	Pop	0.731	0.573	Yes (1)	4	0.113	10.059	10/30/2020	0.0544	205.09

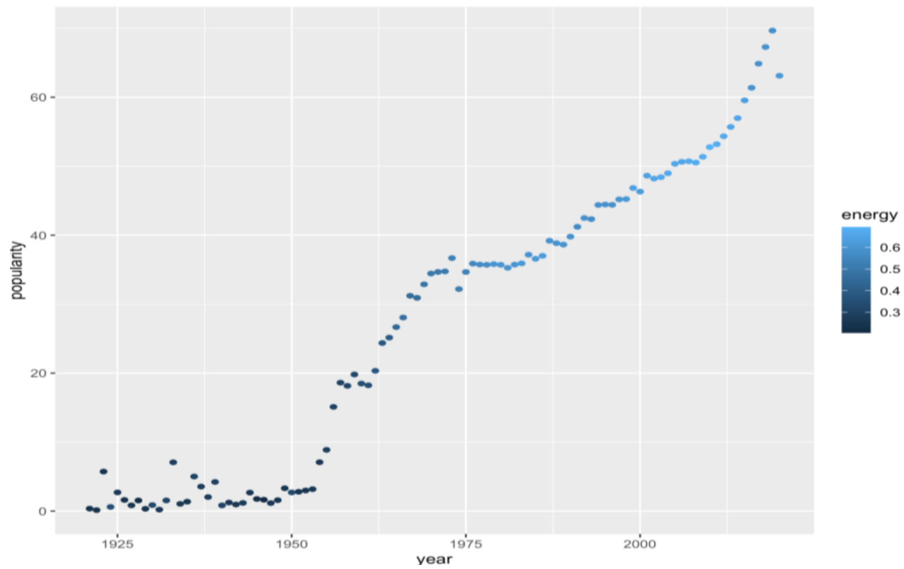
Based on the rules we've developed for identifying a hit song, this song does quality as a "hit" - which is accurate as it has a 100-popularity score according to Spotify. Based on our model, the optimal release schedule is in August, which is extremely close to the actual release date in October. This song also is in the pop genre. It doesn't contain "love" in the song title, but it meets many of the other criteria in our model or a "hit" song.

After this test, the team feels confident that our framework for identifying the next hit artist and song is accurate. We look forward to using this framework to inform investment decisions at Oswald Records – ROCK ON!

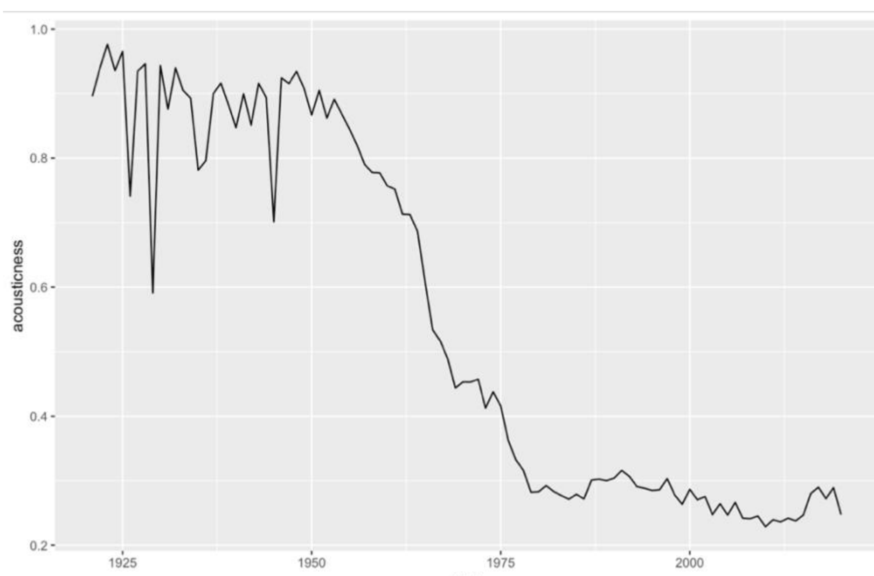
Appendix

Appendix A – Additional data visualizations

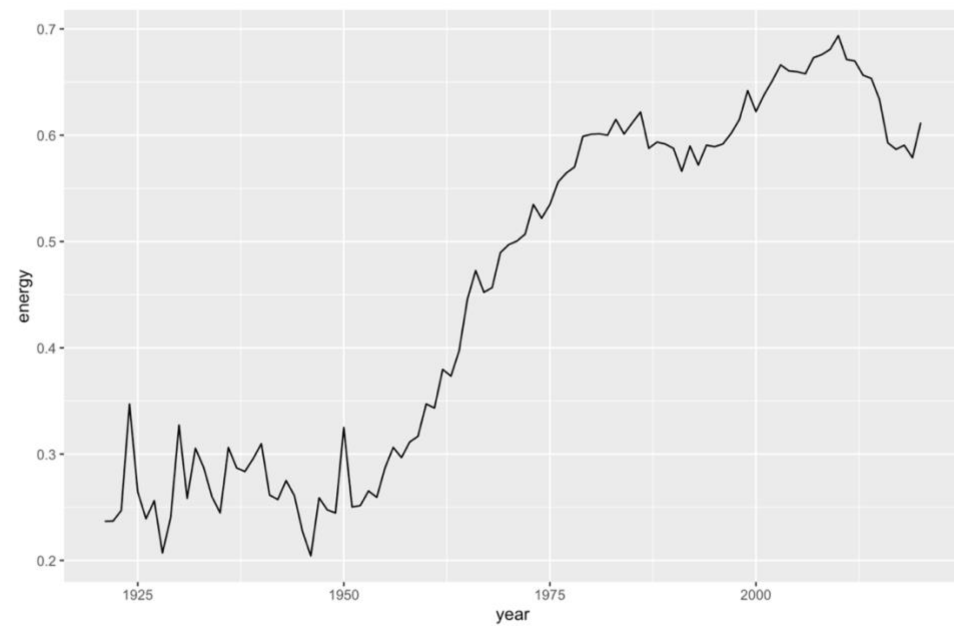
Average song popularity based on creation date and corresponding increase in energy.



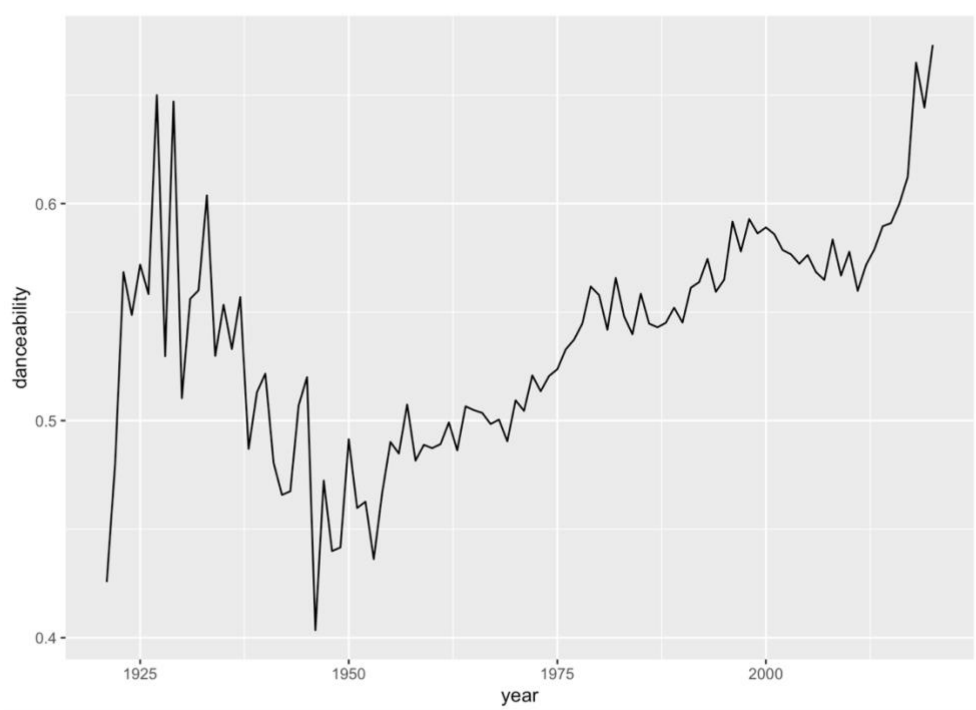
Average acousticness of songs over the years



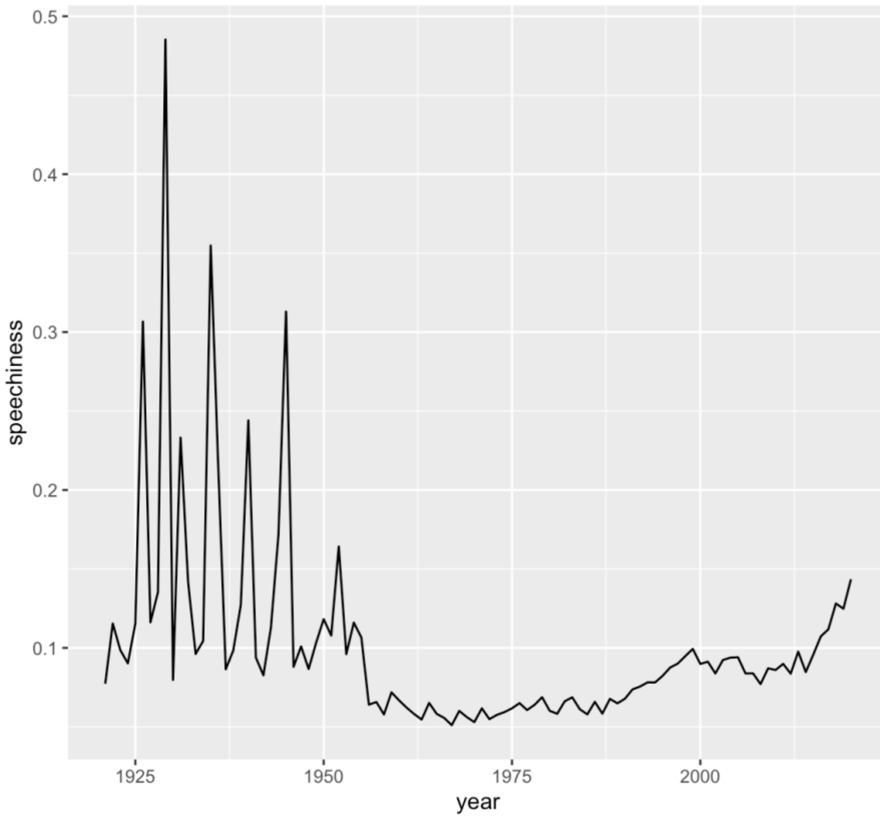
Energy of songs over the years



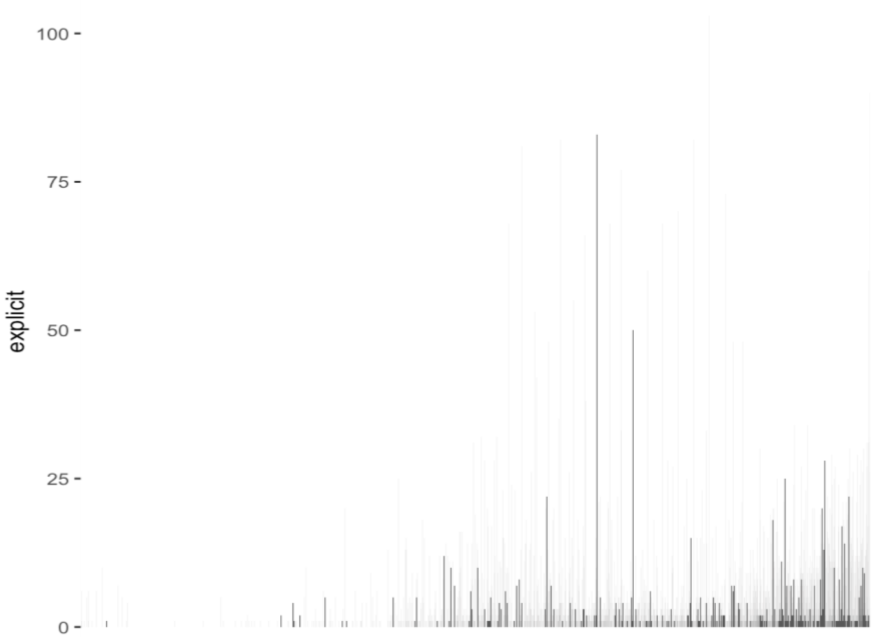
Average danceability of songs over the years



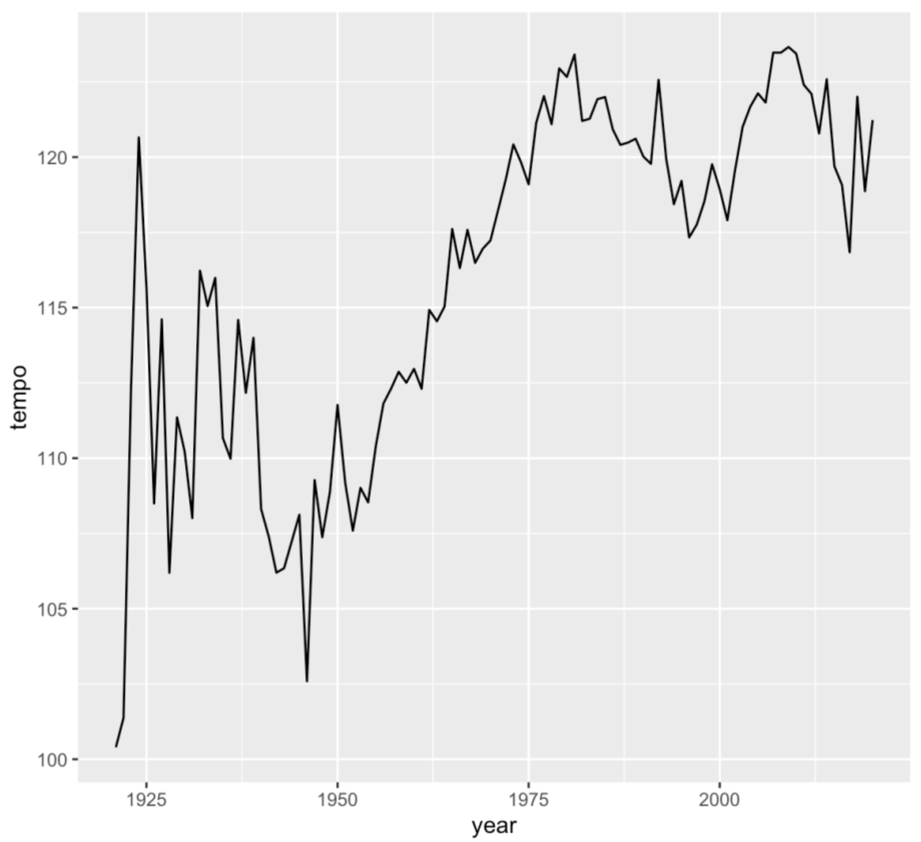
Average speechiness of songs over the years



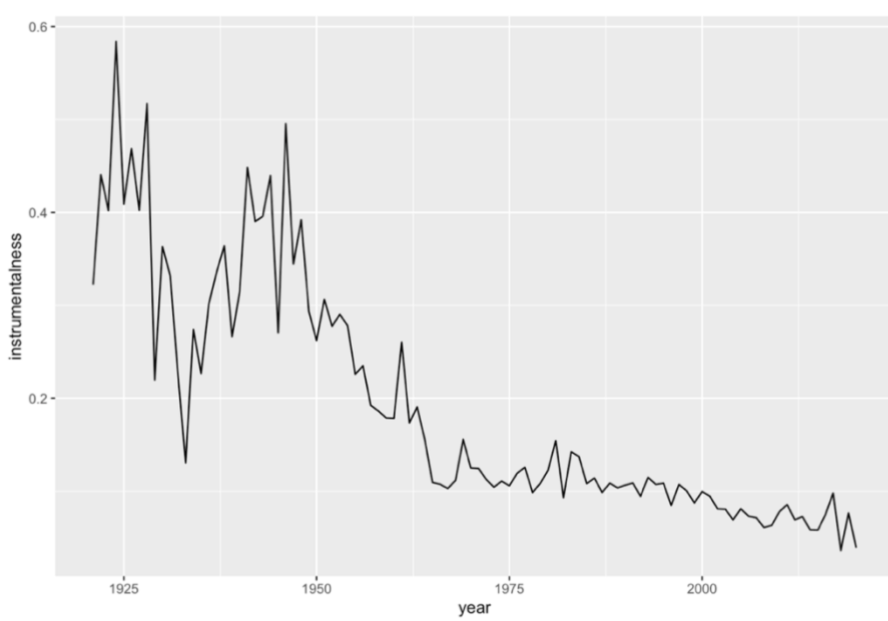
Count of explicit tracks over the years.



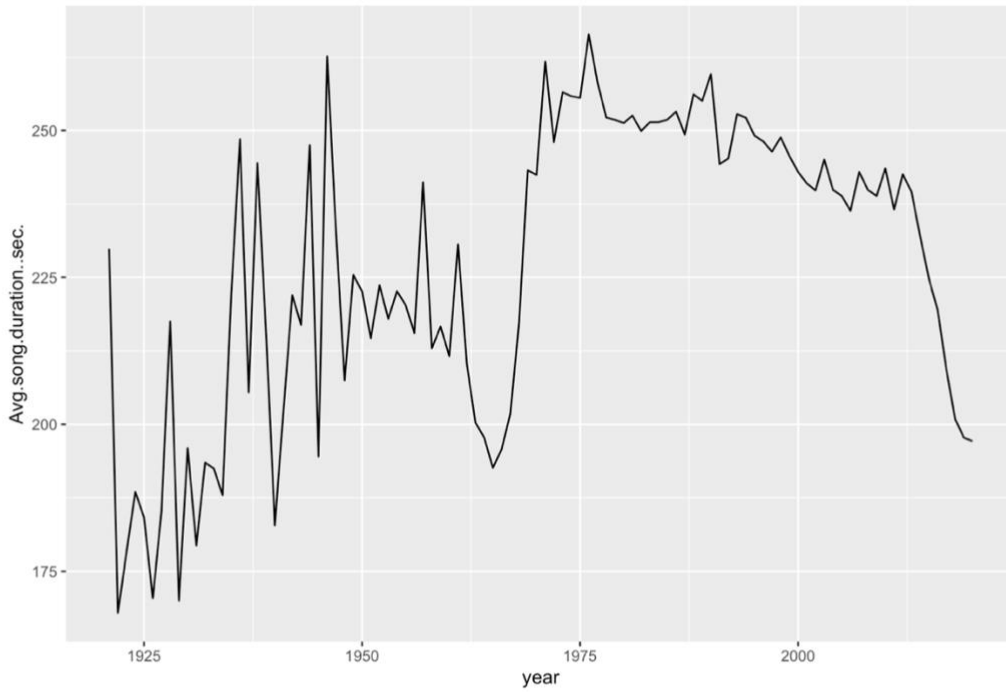
Average tempo of songs over the years



Average instrumentalness of songs over the years



Average track duration over the years



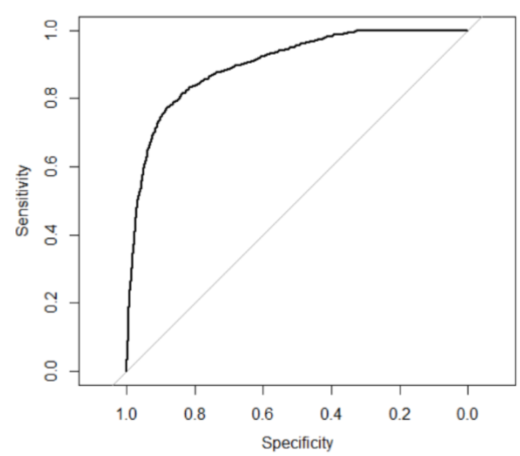
Appendix B

Logit

```
Call:
glm(formula = popularity ~ ., family = "binomial", data = train.df1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6366  -0.2756  -0.1643  -0.0855   5.0061

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.5698238  0.2526167  -2.256 0.024090 *
valence                 -1.8336069  0.1001633 -18.306  < 2e-16 ***
acousticness            -0.7273138  0.0981031  -7.414 1.23e-13 ***
danceability             3.1228028  0.1570777  19.881  < 2e-16 ***
energy                  -1.9580267  0.1725773 -11.346  < 2e-16 ***
explicit1                0.5267146  0.0545355   9.658  < 2e-16 ***
instrumentalness        -1.9856695  0.1933446 -10.270  < 2e-16 ***
loudness                 0.2468038  0.0097256  25.377  < 2e-16 ***
mode1                   -0.2556610  0.0414455  -6.169 6.89e-10 ***
speechiness             -1.2135539  0.2293594  -5.291 1.22e-07 ***
tempo                    0.0024149  0.0007245   3.333 0.000858 ***
release_monthFeb         1.1420501  0.1032867  11.057  < 2e-16 ***
release_monthMar         1.1522648  0.0935964  12.311  < 2e-16 ***
release_monthApr         0.9150206  0.1011086   9.050  < 2e-16 ***
release_monthMay         1.1592046  0.0904226  12.820  < 2e-16 ***
release_monthJun         0.9356483  0.0934750  10.010  < 2e-16 ***
release_monthJul         1.2219140  0.0951559  12.841  < 2e-16 ***
release_monthAug         1.2455185  0.0901793  13.812  < 2e-16 ***
release_monthSep         1.0492356  0.0879158  11.935  < 2e-16 ***
release_monthOct         1.0891124  0.0853017  12.768  < 2e-16 ***
release_monthNov         1.0195728  0.0845896  12.053  < 2e-16 ***
release_monthDec         0.8074246  0.0999391   8.079 6.52e-16 ***
Avg.song.duration..sec. -0.0048322  0.0003590 -13.460  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25079  on 82730  degrees of freedom
Residual deviance: 20366  on 82708  degrees of freedom
AIC: 20412

Number of Fisher Scoring iterations: 8
```

Confusion Matrix and ROC Curve

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27591   308
         1  6650   908

               Accuracy : 0.8038
                 95% CI : (0.7996, 0.8079)
    No Information Rate : 0.9657
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1572

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.74671
            Specificity : 0.80579
         Pos Pred Value : 0.12014
         Neg Pred Value : 0.98896
             Prevalence : 0.03430
         Detection Rate : 0.02561
   Detection Prevalence : 0.21316
      Balanced Accuracy : 0.77625

       'Positive' Class : 1
```



Appendix C

Correlation Matrix:

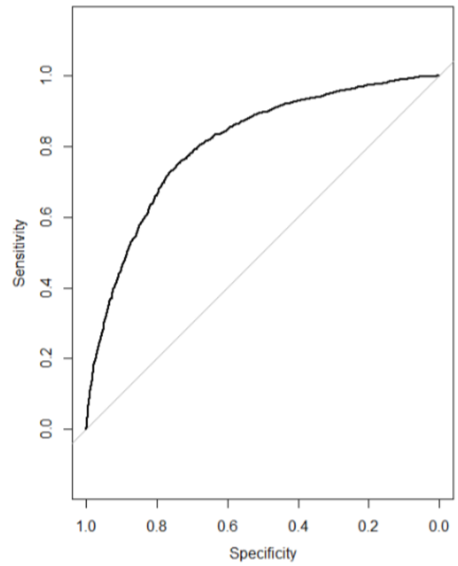
	acousticness	danceability	Avg.song.duration..sec.	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity	key	mode	count
acousticness	1.00	-0.42	-0.04	-0.80	0.33	0.04	-0.63	-0.02	-0.26	-0.23	-0.62	-0.04	0.11	0.02
danceability	-0.42	1.00	-0.12	0.40	-0.36	-0.10	0.48	0.27	0.12	0.62	0.28	0.04	-0.11	-0.03
Avg.song.duration..sec.	-0.04	-0.12	1.00	-0.02	0.10	-0.02	-0.09	0.03	-0.05	-0.19	0.02	-0.01	-0.03	0.00
energy	-0.80	0.40	-0.02	1.00	-0.33	0.10	0.80	0.06	0.31	0.40	0.47	0.04	-0.10	-0.02
instrumentalness	0.33	-0.36	0.10	-0.33	1.00	-0.06	-0.48	-0.16	-0.15	-0.25	-0.20	-0.03	-0.02	0.01
liveness	0.04	-0.10	-0.02	0.10	-0.06	1.00	0.05	0.19	-0.03	0.01	-0.13	-0.01	0.02	0.01
loudness	-0.63	0.48	-0.09	0.80	-0.48	0.05	1.00	0.03	0.28	0.41	0.37	0.03	-0.07	-0.02
speechiness	-0.02	0.27	0.03	0.06	-0.16	0.19	0.03	1.00	-0.02	0.10	-0.04	0.01	-0.05	-0.02
tempo	-0.26	0.12	-0.05	0.31	-0.15	-0.03	0.28	-0.02	1.00	0.22	0.15	0.01	-0.01	0.01
valence	-0.23	0.62	-0.19	0.40	-0.25	0.01	0.41	0.10	0.22	1.00	-0.02	0.04	-0.02	0.00
popularity	-0.62	0.28	0.02	0.47	-0.20	-0.13	0.37	-0.04	0.15	-0.02	1.00	0.01	-0.11	-0.05
key	-0.04	0.04	-0.01	0.04	-0.03	-0.01	0.03	0.01	0.01	0.04	0.01	1.00	-0.08	-0.04
mode	0.11	-0.11	-0.03	-0.10	-0.02	0.02	-0.07	-0.05	-0.01	-0.02	-0.11	-0.08	1.00	0.08
count	0.02	-0.03	0.00	-0.02	0.01	0.01	-0.02	-0.02	0.01	0.00	-0.05	-0.04	0.08	1.00

Logit

```
Call:
glm(formula = popularity ~ ., family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1974  -0.4733  -0.3226  -0.2042   5.3012

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.2705411  0.2086502   6.089 1.13e-09 ***
acousticness            -1.4024883  0.0847893 -16.541  < 2e-16 ***
danceability             4.0091667  0.1684275  23.804  < 2e-16 ***
Avg.song.duration..sec. -0.0098199  0.0003792 -25.897  < 2e-16 ***
instrumentalness         0.5077576  0.0896148   5.666 1.46e-08 ***
liveness                -0.7111140  0.1727432  -4.117 3.85e-05 ***
loudness                 0.0455846  0.0059704   7.635 2.26e-14 ***
speechiness             -0.1332391  0.2028233  -0.657  0.51123
tempo                    0.0003489  0.0009036   0.386  0.69946
valence                 -4.0345916  0.1170448 -34.470  < 2e-16 ***
key1                     0.0718793  0.0909804   0.790  0.42950
key2                    -0.2827032  0.0947176  -2.985  0.00284 **
key3                    -0.2303447  0.1337538  -1.722  0.08504 .
key4                    -0.3456761  0.1089077  -3.174  0.00150 **
key5                    -0.1515866  0.0975749  -1.554  0.12029
key6                     0.1066496  0.0983434   1.084  0.27816
key7                    -0.2645807  0.0877907  -3.014  0.00258 **
key8                    -0.0599071  0.1080070  -0.555  0.57913
key9                    -0.3927379  0.0994671  -3.948 7.87e-05 ***
key10                   -0.1455275  0.1024867  -1.420  0.15562
key11                   -0.2228574  0.0969252  -2.299  0.02149 *
mode1                   -0.3288723  0.0475465  -6.917 4.62e-12 ***
count                   -0.0038426  0.0009461  -4.062 4.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18814  on 27260  degrees of freedom
Residual deviance: 15467  on 27238  degrees of freedom
AIC: 15513

Number of Fisher Scoring iterations: 6
```

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7382  314
         1 2349  860

               Accuracy : 0.7558
                 95% CI : (0.7476, 0.7638)
    No Information Rate : 0.8923
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2787

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.73254
            Specificity : 0.75861
         Pos Pred Value : 0.26800
         Neg Pred Value : 0.95920
             Prevalence : 0.10766
         Detection Rate : 0.07886
   Detection Prevalence : 0.29427
      Balanced Accuracy : 0.74557

       'Positive' Class : 1
```



Confusion Matrix for CART

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9635 1040
         1   96  134

               Accuracy : 0.8958
                 95% CI : (0.8899, 0.9015)
    No Information Rate : 0.8923
    P-Value [Acc > NIR] : 0.123

                  Kappa : 0.1613

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.11414
            Specificity : 0.99013
         Pos Pred Value : 0.58261
         Neg Pred Value : 0.90258
             Prevalence : 0.10766
         Detection Rate : 0.01229
   Detection Prevalence : 0.02109
      Balanced Accuracy : 0.55214

       'Positive' Class : 1
```

Appendix D

Confusion Matrix for Month prediction CART

```
Confusion Matrix and Statistics
          Reference
Prediction Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
       Jan  69   0   0   0   0   0   0   0   0   0   0   8
       Feb   0  37   0   0   0   0   0   0   0   0   0   2
       Mar   0   0  29   0   0  11   0   0   0   1   0   0
       Apr   0   0   0  20   0   0   0   0   0   0   0   0
       May   0   0   0   0  72   0   1   0   0   7   0   7
       Jun   0   0   0   0   2  56   2   0   0   0   0   0
       Jul   9   2  10   0   1   2  29   0   1   0   2   0
       Aug   0   0   0   9   0   0   0  56   0   0   0   0
       Sep   1   0   1   0   5   0   7   5  76   3   5   0
       Oct   0   0   4   6   0  12   9   9   1  78   0   3
       Nov   0   3   2   9   0   7   0   0   2   0  90   0
       Dec   0   0   0   0   0   0   0   0   0   0   0  29

Overall Statistics

               Accuracy : 0.8362
                 95% CI : (0.8028, 0.8395)
    No Information Rate : 0.9304
    P-Value [Acc > NIR] : 0.9961

                  Kappa : 0.1512

 Mcnemar's Test P-Value : <2e-14

Statistics by Class:

                     Class: Jan Class: Feb Class: Mar Class: Apr Class: May Class: Jun Class: Jul Class: Aug Class: Sep
Sensitivity             0.93234    0.91286    0.60734    0.49174     0.9590     0.6283    0.73360     0.7837    0.95253
Specificity             0.84620    0.89371    0.96939    0.98904     0.8384     0.9384    0.78135     1.9772    0.26830
Pos Pred Value          0.87672    0.92672    0.54854    0.25745     0.9791     0.7901    0.01725     0.0047    0.10345
Neg Pred Value          0.92495    0.95413    0.89623    0.95264     0.1102     0.9191    0.89158     0.8807    0.01250
Prevalence              0.05505    0.04587    0.00092    0.04626     0.0367     0.1009    0.18257     0.1193    0.09174
Detection Rate          0.71375    0.56876    0.67101    0.20726     0.7830     0.4374    0.56428     0.0263    0.79687
Detection Prevalence    0.28254    0.06627    0.18752    0.00372     0.2743     0.0917    0.02018     0.0735    0.30606
Balanced Accuracy       0.77386    0.87946    0.54545    0.42834     0.8803     0.6827    0.69264     0.6602    0.89869
                     Class: Oct Class: Nov Class: Dec
Sensitivity             0.88250    0.93953    0.81269
Specificity             0.83871    0.87234    0.88273
Pos Pred Value          0.85000    0.74286    0.04286
Neg Pred Value          0.17640    0.26316    0.17248
Prevalence              0.14679    0.13761    0.12752
Detection Rate          0.64587    0.81835    0.77869
Detection Prevalence    0.12349    0.39844    0.25573
Balanced Accuracy       0.87560    0.90284    0.79572
```

Appendix E

```
> km$centers
         mode acousticness danceability     energy instrumentalness    liveness   loudness
1  0.11705883    1.2749979   -0.7949947 -1.2938465        0.7913502 -0.12457543 -1.2632998
2 -0.04158177   -0.4529062    0.2823990  0.4596017       -0.2811043  0.04425183  0.4487508
  speechiness      tempo    valence popularity         key Avg.song.duration..sec.
1 -0.16461681 -0.6503090 -0.6866921 -0.7493119 -0.19397987              0.18789526
2  0.05847537  0.2310035  0.2439276  0.2661714  0.06890575             -0.06674437
> km$withinss
[1] 10646.92 19805.87
```