Spotify

Popularity of a Song

Gabriel Ferreira, Madeline Edmonds

Abstract

Spotify Web API is what is commonly known as a RESTful API. The web API is an interface that programs can use to retrieve and manage Spotify data over the internet. The Web API uses the same HTTP protocol that's used by every internet browser. In fact, you can access the API directly from your own browser. Understanding the different categories in music can help you as a consumer gain a grasp as to what is to come in the next year when it comes to musical entertainment. Over the course of this project, we will provide information regarding the different variables and observations and how they might impact the popularity of the song.

Contents

1. - Introduction
2. - Type of Project
3. - Data Description and Exploration
   3.1 - Removing variables not useful
   3.2 - Checking Missing Values
   3.3 - Understanding better each variable
      3.3.1 - Histogram of the Danceability Rating
      3.3.2 - Histogram of the Popularity Rating and Songs
      3.3.3 - Scatter Plot Between the Duration of a Song and Its Release Date
      3.3.4 - Boxplot graph containing variables with range from 0 to 1
      3.3.5 - Histogram providing a better visualization of the primary key of the track encoded as integers in between 0 and 11.
4. - Correlation
   4.1 - Correlation between variables
   4.2 - Top 10 variables most correlated to popularity
5. - The Most Popular Artists
6. - Modeling and Analyzing
   6.1 - Cluster 0
   6.2 - Cluster 1
   6.3 - Cluster 2
   6.4 - Cluster 3

7. - Final Predictions
8. - Conclusion

1 - Introduction

Spotify is one of the greatest online music platforms available on the market. There are all kinds of music available to listen, and the company has collected data from most if not all the available songs on the platform. In this project, we are going to analyze almost 170,000 songs to help the music industry to know what the best number for each variable to make a popular song. We are going to explore this data starting by explaining what each variable represents for the songs; then cleaning and modeling it to prepare the data to be analyzed and then to obtain good predictions of the best way to create a popular song in the Spotify.

2 - Type of Project

This project will be required data modeling and analysis techniques. The techniques that will be used to lead this modeling, analysis and predictions will be Clustering Analysis and Random Forest.

3 - Data Description and Exploration

The data that will be used is a music data and was collected from Spotify Web API. The data-frame is a matrix represented by 19 variables and 169,909 observations. Each observation represents a different song, and the following table contains the meaning of each column (variable):

Variable	**Meaning**	**Range**
	Categorical	
artists	First and last name of the artist responsible for the song. If they go by a stage name, their stage names are acknowledged.	-
Name	Name of the songs	-
key	The primary key of the track encoded as integers in between 0 and 11	-
release_date	Date of release mostly in yyyy-mm-dd format, however precision of date may vary	-
	Dummy	
Mode	0 = Minor, 1 = Major	-
explicit	0 = No explicit content, 1 = Explicit content	-
	Numeric	
popularity	The popularity of the song lately, default country = US	from 0 to 100
danceability	A float value from 0.0 to 1.0 to say how good the song is for dancing. If a song is perfect for dancing, the value given will be 1.0.	from 0 to 1

energy	A float value from 0.0 to 1.0 to display the energy of the song. For example, rock and rap will be considered high energy. If a song is very high energy, the value given will be 1.0.	from 0 to 1
count	The number of tracks from the original dataset, produced by the given artist	from 0 to 1
liveness	Measures if there is indication of and audience. For example, a live recording will have a measurement over 0.8.	from 0 to 1
speechiness	The presence of spoken words in a song. If a song is purely speech, like spoken word or poetry, its value will be close to 1.0	from 0 to 1
instrumentalness	Prediction values on how much of a song is instrumental.	from 0 to 1
acousticness	The relative metric of the track being acoustic.	from 0 to 1
valence	A value measuring how positive or negative the song is. An extremely positive song, meaning it has cheery, lively tones, will have a 1.0 rating.	from 0 to 1
loudness	Relative loudness of the track in the typical range.	from -60 to 0 (Float)
tempo	The beat or speed of the song. The tempo of the track in Beat Per Minute (BPM).	from 0 to 150 (Float)
duration_ms	The length of the track in milliseconds (ms)	200k to 300k
year	Year of the song was released.	1921 to 2020

Here a small sample of the data:

```
> head(spotify)
  acousticness                                 artists danceability duration_ms energy explicit                     id instrumentalness key
1        0.995                      ['Carl Woitschach']        0.708      158648 0.1950        0 6KbQ3uYMLKb5jDxLF7wYDD            0.563  10
2        0.994    ['Robert Schumann', 'Vladimir Horowitz']        0.379      282133 0.0135        0 6KuQTIu1KoTTkLXKrwlLPV            0.901   8
3        0.604                  ['Seweryn Goszczyński']        0.749      104300 0.2200        0 6L63VW0PibdM1HDSBoqnoM            0.000   5
4        0.995                     ['Francisco Canaro']        0.781      180760 0.1300        0 6M94FkXd15sOAOQYRnWPN8            0.887   1
5        0.990    ['Frédéric Chopin', 'Vladimir Horowitz']        0.210      687733 0.2040        0 6N6tiFZ9vLTSOIxkj8qKrd            0.908  11
6        0.995 ['Felix Mendelssohn', 'Vladimir Horowitz']        0.424      352600 0.1200        0 6NxAf7M8DNHOBTmEd3JSO5            0.911   6
  liveness loudness mode                                  name popularity release_date speechiness   tempo valence year
1   0.1510  -12.428    1             Singende Bataillone 1. Teil          0         1928      0.0506 118.469  0.7790 1928
2   0.0763  -28.454    1      Fantasiestücke, Op. 111: Più tosto lento          0         1928      0.0462  83.972  0.0767 1928
3   0.1190  -19.924    0            Chapter 1.18 - Zamek kaniowski          0         1928      0.9290 107.177  0.8800 1928
4   0.1110  -14.734    0 Bebamos Juntos - Instrumental (Remasterizado)          0   1928-09-25      0.0926 108.003  0.7200 1928
5   0.0980  -16.829    1   Polonaise-Fantaisie in A-Flat Major, Op. 61          1         1928      0.0424  62.149  0.0693 1928
6   0.0915  -19.242    0              Scherzo a capriccio: Presto          0         1928      0.0593  63.521  0.2660 1928
```

3.1 - Removing variables that are not useful

- "release_data" - This variable is the song's release data. The problem of these variables is because some songs have the day, month and year; but most of the songs have only the year. Also, we can just use the variable “year”, which is only the release year for all songs. Therefore, this variable is not needed in the model.

- “id” - This variable is also not needed in the model because it does not add any value to the model.
- “name” - This variable is also not needed in the model because it does not add any value to the model.

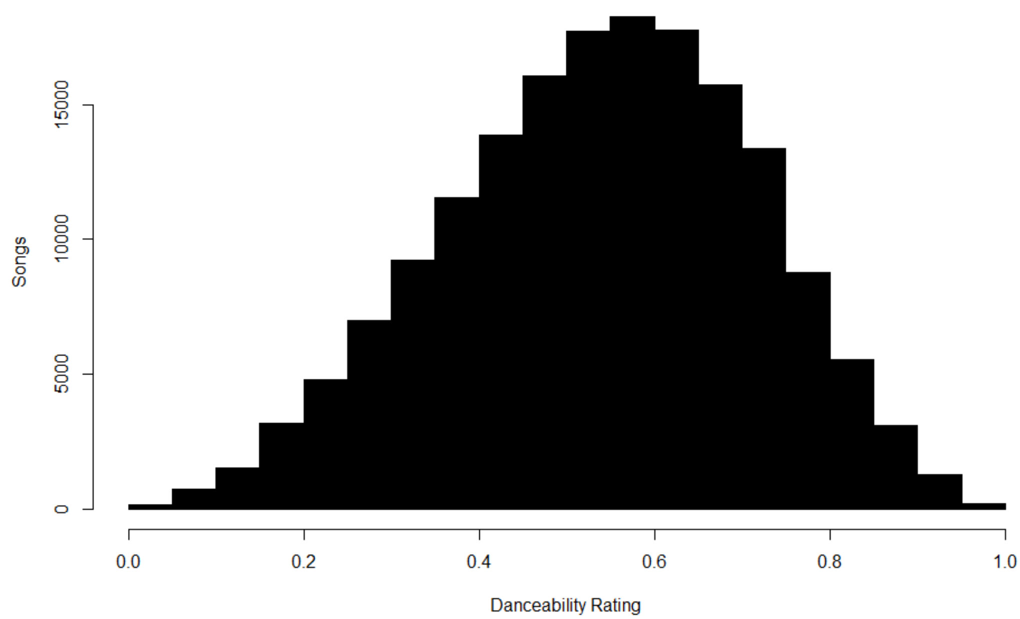
3.2 - Checking Missing Values

- There are no rows with missing values.
- There are no duplicated values

3.3 – Understanding each variable better

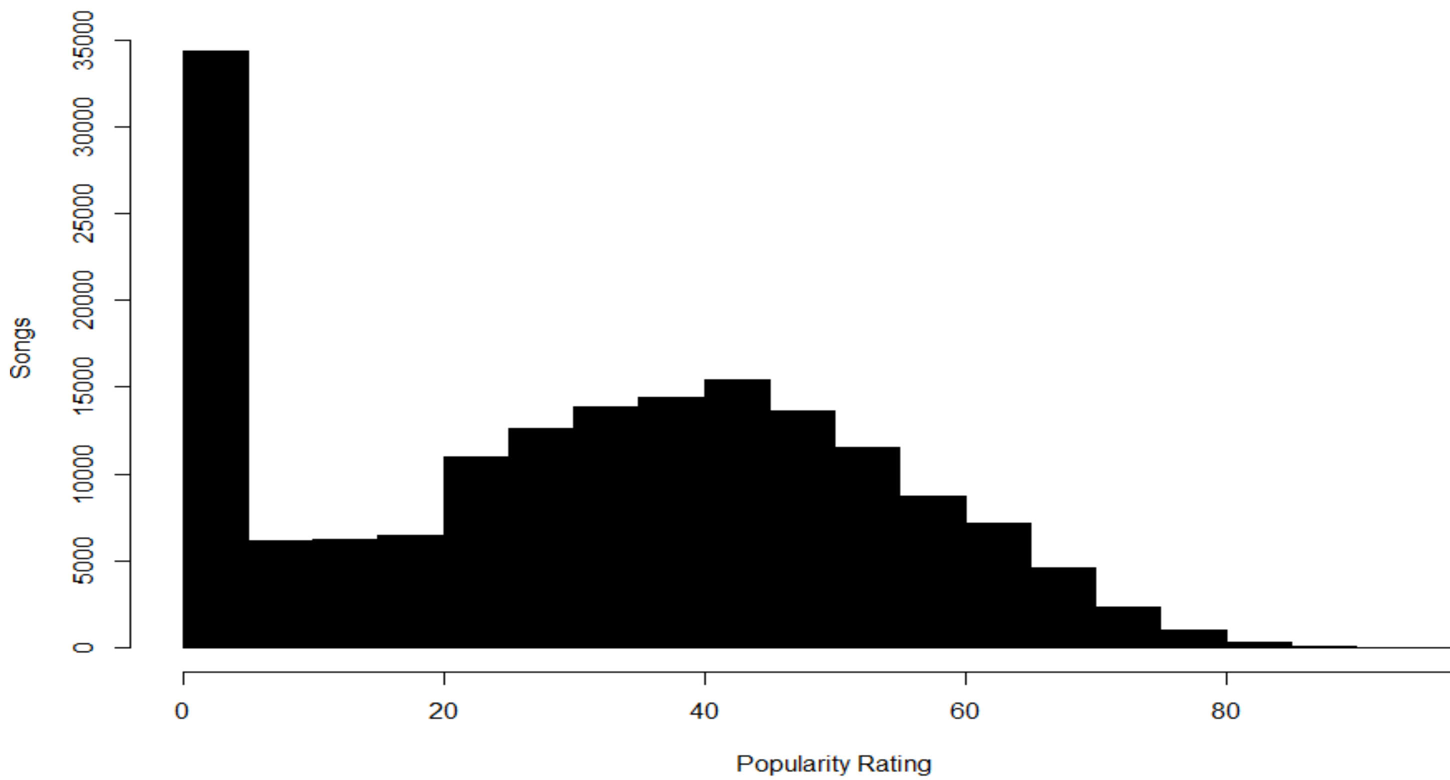
3.3.1 - Histogram of the Danceability Rating

- The graph below shows the relation of Danceability and Songs. Most songs are 60% danceable and the graph shows the data is very symmetrical around that point.



3.3.2 - Histogram of the Popularity Rating and Songs

- The graph below shows the relation of Popularity and Songs. Most songs are between a 20% and 80% popularity rate. The graph is also right skewed.



3.3.3 - Scatter Plot Between the Duration of a Song and Its Release Date

- This graph shows the relations between the duration of a song and it is release year to see if songs have gotten longer or shorter through the years.

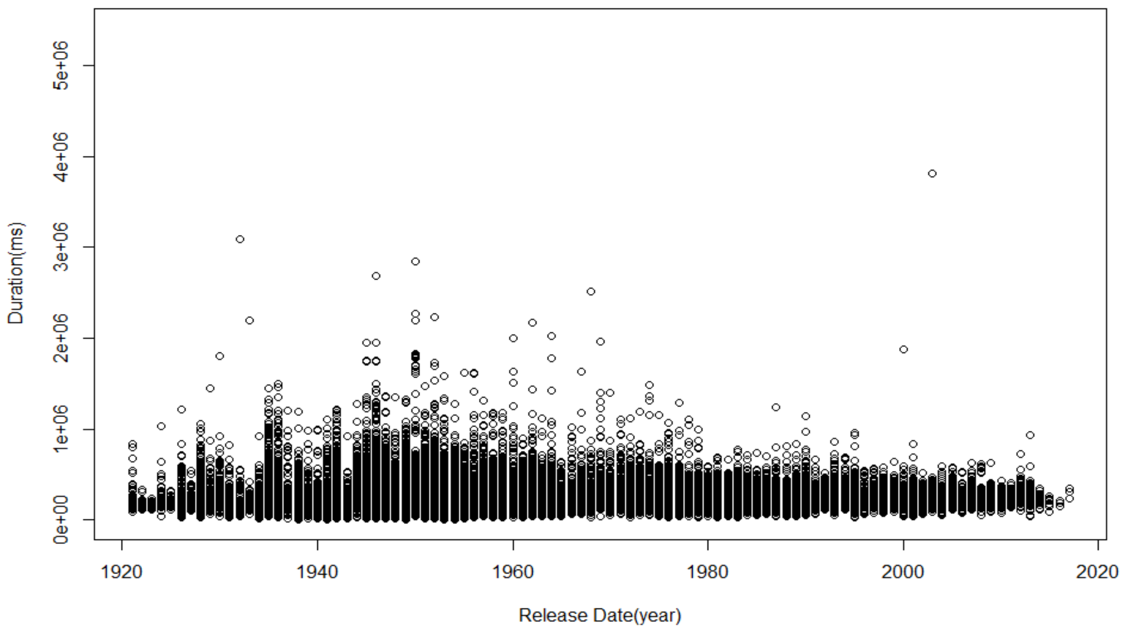


Figure 4.2.1: Scatter plot between Duration and Release Year

3.3.4 - Boxplot graph containing variables with range from 0 to 1:

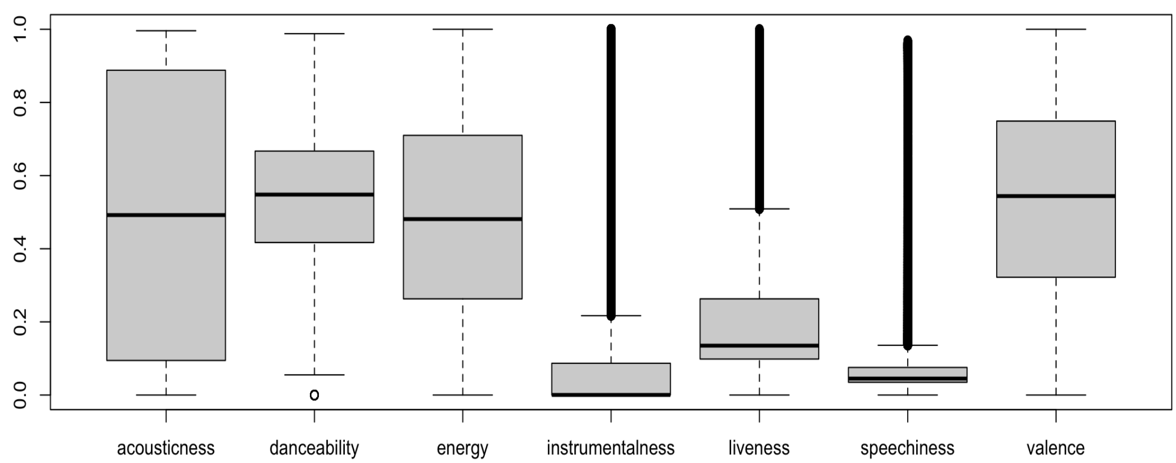
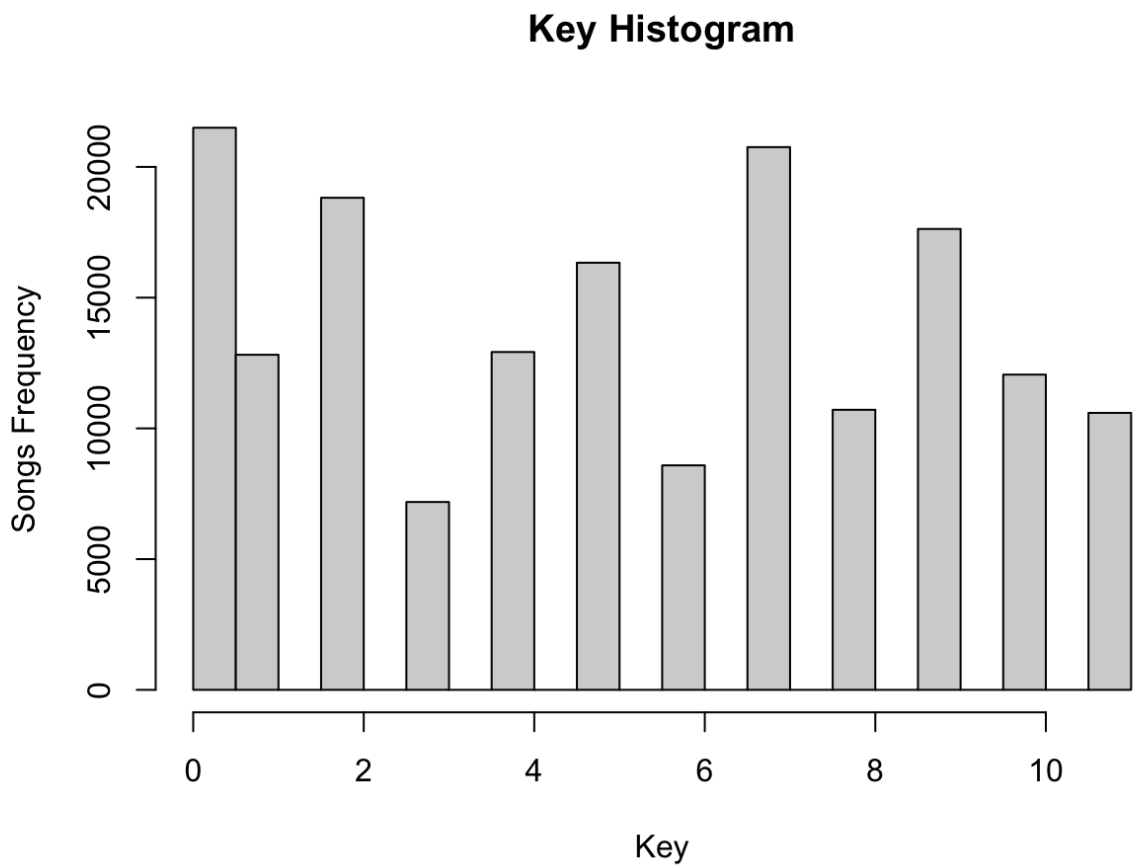


Figure 4.2.2: Boxplots of Acousticness, Danceability, Energy, Instrumentalness, Liveliness, Speechiness, and Valence

Observations from the above graph:

- Acousticness, danceability, energy, and valence variables do not have outliers.
- Instrumentalness, liveness, and speechiness variables contain a lot of outliers.
-

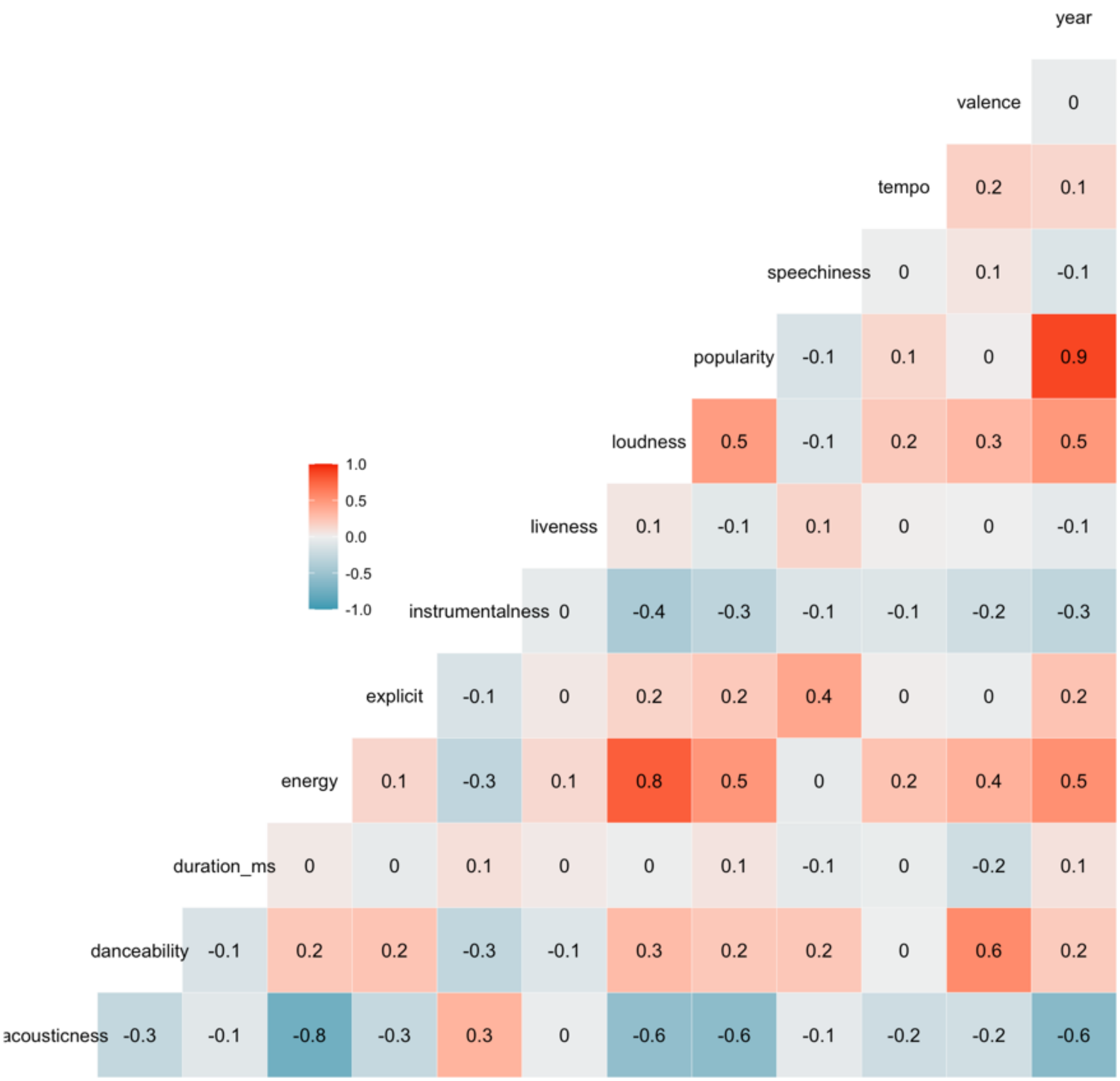
3.3.5 - Histogram provides a better visualization of the primary key of the track encoded as integers between 0 and 11.



4 - Correlation

Next, we are going to see how the variables correlate between them and which are the most correlated to popularity.

4.1 - Correlation between variables



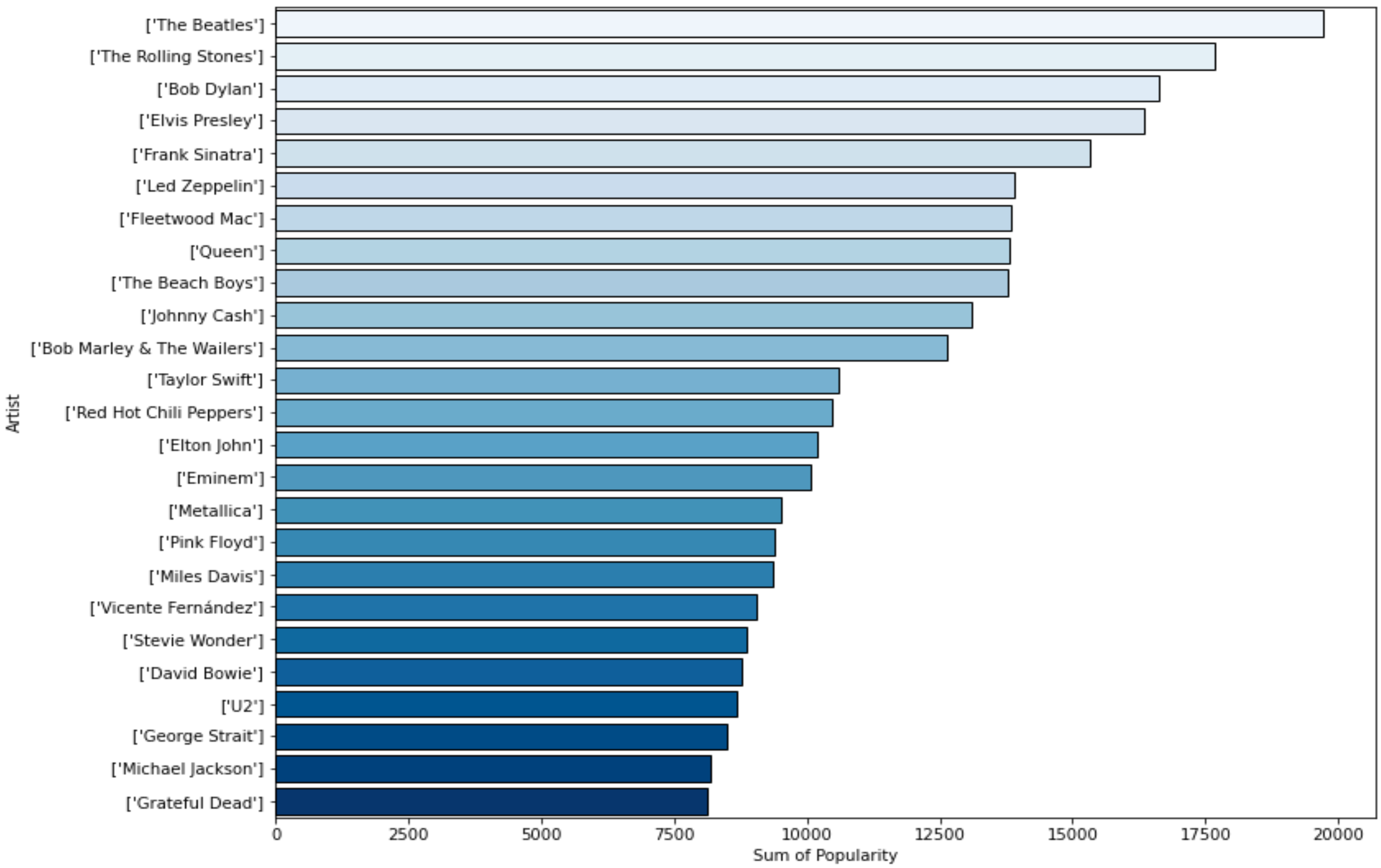
4.2 - Top 10 variables most correlated to popularity

Variables	Correalation (Absolute)
year	0.880
acousticness	0.593
energy	0.497
loudness	0.466
instrumentalness	0.299
danceability	0.221

explicit	0.214
speechiness	0.135
tempo	0.135
liveness	0.075

5 - The Most Popular Artists

Thinking logic, we all know that there is straight relationship between popularity of a song and the artist or singer who is performing it. From this pre knowledge, we made a barplot to show the top 25 artists who have the best songs by summing their songs' popularity in the Spotify.



6 - Modeling and Analyzing

With popularity as the target variable, the question became, what makes a song popular? Using the KMeans clustering and grouping the data into four clusters we were able to see how songs are analyzed and perceived.

6.1 - Cluster 0

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000	6064.000000
mean	0.835735	0.314999	249570.935588	0.265618	0.340864	0.216398	-16.653850	0.058815	105.594082	0.280499	23.039961
std	0.223058	0.102467	97527.413332	0.198608	0.370666	0.157511	6.560881	0.063152	29.366496	0.193448	22.293611
min	0.000000	0.000000	29080.000000	0.000000	0.000000	0.000000	-60.000000	0.000000	0.000000	0.000000	0.000000
25%	0.786717	0.246804	183500.000000	0.124150	0.000681	0.111814	-20.265015	0.037500	85.525500	0.129975	0.000000
50%	0.936477	0.330625	224367.000000	0.220414	0.149000	0.163417	-15.893000	0.043800	102.172125	0.246000	19.000000
75%	0.978000	0.394000	297013.333333	0.361100	0.744938	0.264331	-12.003312	0.055600	120.706583	0.394125	42.666667
max	0.996000	0.596500	571773.000000	1.000000	1.000000	0.983000	-1.532000	0.948000	217.743000	0.972000	85.000000

The cluster above contains about 6000 songs. It shows that with this cluster on average, songs that are acoustic and long have a 23% chance of being popular. This means form a company perspective they should be wary of having an acoustic musician releasing music.

6.2 - Cluster 1

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000	7956.000000
mean	0.806068	0.583101	189370.482379	0.367166	0.178857	0.212340	-12.529107	0.114621	113.966752	0.601604	22.245360
std	0.161921	0.095794	58150.807064	0.181046	0.295656	0.142748	4.539412	0.151701	23.303154	0.207324	21.734782
min	0.347000	0.377000	20293.000000	0.001590	0.000000	0.021500	-44.761000	0.023200	22.230500	0.017600	0.000000
25%	0.677000	0.508000	157726.785714	0.235150	0.000004	0.117000	-14.864000	0.040931	99.144875	0.459500	0.000000
50%	0.838000	0.572367	184884.500000	0.352000	0.002465	0.167914	-11.947565	0.057219	113.258915	0.610411	17.585714
75%	0.960569	0.645625	213871.450000	0.477000	0.252689	0.256242	-9.508714	0.113380	126.288000	0.755000	40.211364
max	0.996000	0.934000	572104.000000	0.997000	0.986000	0.977000	0.474000	0.960000	212.141000	0.989000	94.000000

The cluster above contains about 8000 songs. Even though the danceability and energy are higher for this cluster, the acoustics, duration, and popularity are lower than the cluster above. It still confirms to the company using this data that acoustic music still isn't popular. Maybe they should look at a pop song.

6.3 - Cluster 2

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	13237.000000	13237.000000	13237.000000	13237.000000	1.323700e+04	13237.000000	13237.000000	13237.000000	13237.000000	13237.000000	13237.000000
mean	0.163517	0.631813	242800.981284	0.680287	7.181471e-02	0.192255	-7.762991	0.100461	121.199184	0.580005	47.087392
std	0.145407	0.140357	65716.117858	0.159869	1.835346e-01	0.127983	3.144595	0.097965	22.248448	0.207878	13.493400
min	0.000001	0.230500	18795.500000	0.009270	0.000000e+00	0.012200	-31.004000	0.021350	50.113000	0.030800	0.000000
25%	0.034600	0.533000	202733.000000	0.569000	6.590909e-07	0.108000	-9.627000	0.040333	106.158250	0.435500	39.000000
50%	0.125346	0.638000	234377.666667	0.687000	1.790000e-04	0.159927	-7.211750	0.059000	120.567000	0.585000	47.466667
75%	0.267000	0.736000	272743.666667	0.798500	1.980633e-02	0.235860	-5.482000	0.118475	133.148250	0.736367	56.000000
max	0.659571	0.986000	657427.000000	0.999000	9.857500e-01	0.991000	1.342000	0.960000	210.654000	0.991000	95.000000

Cluster 2 contains over ten thousand songs. As big as this set is, it represents our pop, hip-hop, or techno songs. When looking at the averages, a company can fully confirm their assumption of not relying on an acoustic song. Low acoustics, high danceability and energy can give a song an average 47% popularity. This means that just looking at the data alone, a song in this cluster could appeal to large audiences and make a ton of money.

6.4 - Cluster 3

	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	364.000000	364.000000	3.640000e+02	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000
mean	0.806368	0.342199	8.947424e+05	0.274387	0.447113	0.212108	-17.985946	0.159879	102.174236	0.243081	19.472082
std	0.265905	0.182166	4.855405e+05	0.212966	0.371097	0.174963	5.755359	0.261811	26.019558	0.224351	19.940532
min	0.000023	0.000000	5.714930e+05	0.000075	0.000000	0.039900	-44.347000	0.000000	0.000000	0.000000	0.000000
25%	0.779000	0.206333	6.402361e+05	0.118586	0.005950	0.104000	-21.097500	0.040350	85.182000	0.067075	0.000000
50%	0.926750	0.292000	7.136270e+05	0.218000	0.491302	0.140500	-17.718500	0.046035	100.392181	0.157142	13.545455
75%	0.966219	0.473000	9.619522e+05	0.351813	0.833750	0.254728	-14.608750	0.084406	115.112250	0.348000	38.125000
max	0.993000	0.788000	5.403500e+06	0.966000	1.000000	0.958000	-4.724000	0.964000	193.041000	0.898000	67.000000

This cluster contains 364 songs with high acoustics, low danceability and very low energy. These could be classical songs where the energy is meant to be formal, subdued. It is not surprising to see the low popularity, and companies can learn that these more energy songs will not bring in the fans.

7 - Final Predictions

For the final predictions, we evaluated the performance of different numbers of trees by extracting the root mean square error (RMSE) and mean absolute error (MAE). After analyzing the output results, we decided the best performance is using 100 trees. So, according to the Random Forest Model, using 100 decisions trees to analyze all songs in the data, the number that provides the highest predicted popularity for each variable is the following:

Variables	Number
acousticness	0.3
danceability	0.9
duration_ms	679907ms
energy	1.0
explicit	>= 0.6
instrumentalness	0.1
liveness	0.1
loudness	-3.11
speechiness	0.3
tempo	90
valence	0

Using the same Random Forest Model with 100 decisions trees, we simulate a song containing the same data from the above table. The predicted popularity was 47.31.

8 – Conclusion

In this project, we examine the effect of every single variable on the popularity rate. We went through each variable meaning and learned why they are important for a song. However, some of those were not of much use; so, we removed them from the data to simplify the model. Next, we checked the correlation between all variables; then we look at which variables are the most correlated to popularity, ranking them in the top 10 variables where the most correlated variable comes on the top of the table. We also made a graph showing the most 25 popular artists in the data by summing their songs' popularity. After this data exploration, we started analyzing by clusters, where each cluster has different variables values. We interpreted it as different genres of music. Lastly, by going through this project we learned how difficult is to make a popular song on Spotify. In a range from 0 to 100, about only 25% of the songs had a rate greater than 47. For the overall model, we figured out that songs containing variables values close to the numbers from the above table, tend to have a popularity rate greater than 75% of the songs on Spotify.

Reference

[1] Ay, Y. (2020, November 25). Spotify Dataset 1921-2020, 160k+ Tracks. Retrieved December 04, 2020, from https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

[2] (n.d.). Retrieved December 04, 2020, from http://r-graph-gallery.com/index.html

[3] Bhalla, D. (n.d.). R : Keep / Drop Columns from Data Frame. Retrieved December 04, 2020, from https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html