

MAY 8, 2022

FINAL PROJECT REPORT - SPOTIFY

SHWETA GUPTA
BIA-610 APPLIED ANALYTICS
Stevens Institute of Technology – Summer 2021

Revision: 1.0

Table of Contents

Section	Standard	Possible Points
Introduction	<p>1.1 Provide an introduction that explains the problem statement you are addressing. Why should I be interested in this?.....3</p> <p>1.2 Provide a short explanation of how you plan to address this problem statement (the data used and the methodology employed).....3</p> <p>1.3 Discuss your current proposed approach/analytic technique you think will address (fully or partially) this problem.....4</p> <p>1.4 Explain how your analysis will help the consumer of your analysis.....4</p>	<p>10</p> <p>10</p>
Data Preparation	<p>3.1 Original source where the data was obtained is cited and, if possible, hyperlinked.....4</p> <p>3.2 Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).....4</p> <p>3.3 Data importing and cleaning steps are explained in the text (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.....5</p> <p>3.4 Once your data is clean, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible...6</p> <p>3.5 Provide summary information about the variables of concern in your cleaned data set. Rather, provide me with a consolidated explanation, either with a table that provides summary info for each variable or a nicely written summary paragraph with inline code.....7</p>	.10
Exploratory Data Analysis	<p>4.1 Uncover new information in the data that is not self-evident (i.e. do not just plot the data as it is; rather, slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information)..9</p> <p>4.2 Provide findings in the form of plots and tables. Show me you can display findings in different ways.....9</p> <p>4.3 Graph(s) are carefully tuned for the desired purpose. One graph illustrates one primary point and is appropriately formatted (plot and axis titles, a legend if necessary, scales are</p>	10

Section	Standard	Possible Points
	appropriate, appropriate geoms used, etc.).....11	
	4.4 Table(s) carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features. The size of the table is appropriate.....12	
	4.5 Insights obtained from the analysis are thoroughly, yet succinctly, explained. Easy to see and understand the interesting findings that you uncovered.....14	
Summary	6.1 Summarize the problem statement you addressed.....16 6.2 Summarize how you addressed this problem statement (the data used and the methodology employed).....17 6.3 Summarize the interesting insights that your analysis provided.....18 6.4 Summarize the implications to the consumer of your analysis.....20 6.5 Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.....20	10
Formatting & Other Requirements	7.1 Achievement, mastery, cleverness, creativity: Tools and techniques from the course are applied very competently and, perhaps, somewhat creatively. Perhaps the student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, unusually sophisticated application of tools from the course.....21	

.....

1.1 PROVIDE AN INTRODUCTION THAT EXPLAINS THE PROBLEM STATEMENT YOU ARE ADDRESSING. WHY SHOULD I BE INTERESTED IN THIS?

Problem Statement:

We have songs from Spotify, which contains different kind of information about songs. In this exercise, we are trying to understand the factors driving popularity of the songs.

It is of interest to everyone who resonates with music. Often we like some songs which has high tempo and other sometimes songs with low tempo. So, if tempo alone is not a parameter, the question becomes crucial what does really require for a song to be famous?

1.2 PROVIDE A SHORT EXPLANATION OF HOW YOU PLAN TO ADDRESS THIS PROBLEM STATEMENT (THE DATA USED AND THE METHODOLOGY EMPLOYED)

Methodology:

- Drive insights from the data using descriptive statistics
- Build a predictive model to understand popularity measure against multiple song features using inferential statistics

1.3 DISCUSS YOUR CURRENT PROPOSED APPROACH/ANALYTIC TECHNIQUE YOU THINK WILL ADDRESS (FULLY OR PARTIALLY) THIS PROBLEM.

With the help of a predictive model, we will be able to create functional form of music likeliness, something which is entirely associated with experience. A

mathematical equation of musical features will help us understand the music from different perspective.

1.4 EXPLAIN HOW YOUR ANALYSIS WILL HELP THE CONSUMER OF YOUR ANALYSIS.

The analytical approach will immensely help music composers/ artists in identifying right set of music attributes e.g., acoustics, instruments, liveliness, danceability, energy, valence etc.

3.1 ORIGINAL SOURCE WHERE THE DATA WAS OBTAINED IS CITED AND, IF POSSIBLE, HYPERLINKED.

Data is pickup from Course Module provided by Professor.

Please find the link below:

<https://sit.instructure.com/courses/55988/files/8933638?wrap=1>

3.2 SOURCE DATA IS THOROUGHLY EXPLAINED (I.E. WHAT WAS THE ORIGINAL PURPOSE OF THE DATA, WHEN WAS IT COLLECTED, HOW MANY VARIABLES DID THE ORIGINAL HAVE, EXPLAIN ANY PECULIARITIES OF THE SOURCE DATA SUCH AS HOW MISSING VALUES ARE RECORDED, OR HOW DATA WAS IMPUTED, ETC.).

Source data contains 32853 rows and 23 columns.

Each row is presented for 1 song, however, there are few songs which are part of more than 1 playlist, gets repeated.

There are total 3 identifiers in data:

Track_id: Id of song

Artist_id: Id of artist

Album_id: Id of album

Playlist_id: Id of playlist song is part of

Popularity of songs is variable of interest and its value lies between 0 to 100, 0 being minimum and 100 maximum.

Data is collected in February 2020 as last released song in data is from January 2020.

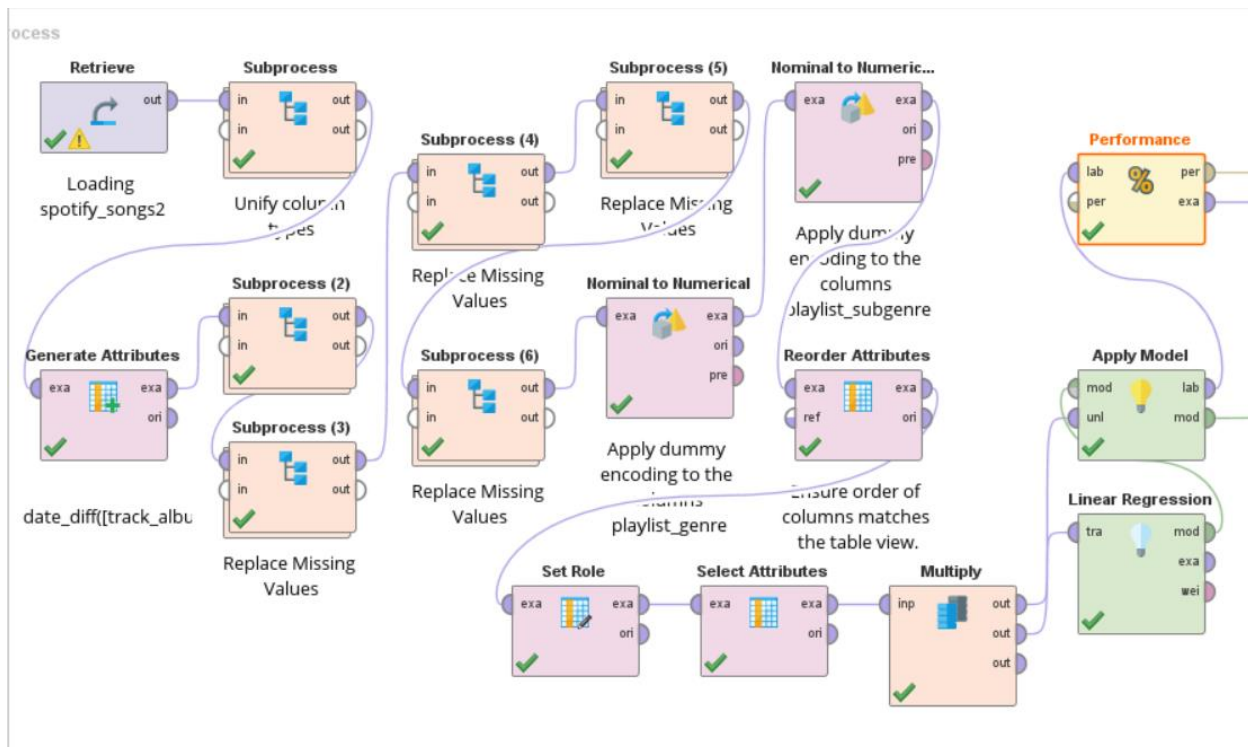
3.3 DATA IMPORTING AND CLEANING STEPS ARE EXPLAINED IN THE TEXT (TELL ME WHY YOU ARE DOING THE DATA CLEANING ACTIVITIES THAT YOU PERFORM) AND FOLLOW A LOGICAL PROCESS.

Missing value treatment:

Some of the data points are missing, but missing rate of columns are very low (< 1%), so it was safe to replace missing values by their mean.

Dummy encoding:

To build the predictive model, it was required to convert categorical variables to dummies. For this exercise, Genre and Sub-Genre are converted into dummy variables with the help of one-hot-encoding.



As represented in RapidMiner process, Replace Missing Values removed null values, while Nominal to Numerical function created one-hot-encoding.











3.4 ONCE YOUR DATA IS CLEAN, SHOW WHAT THE FINAL DATA SET LOOKS LIKE. HOWEVER, DO NOT PRINT OFF A DATA FRAME WITH 200+ ROWS; SHOW ME THE DATA IN THE MOST CONDENSED FORM POSSIBLE.











Once data cleaning was complete, data is saved and retrieved for modeling purposes. Below are first few observations:

playlist_sub...	playlist_sub...	playlist_sub...	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumenta...	liveness
0	0	0	0.748	0.916	6	-2.634	1	0.058	0.102	0	0.065
0	0	0	0.726	0.815	11	-4.969	1	0.037	0.072	0.004	0.357
0	0	0	0.675	0.931	1	-3.432	0	0.074	0.079	0.000	0.110
0	0	0	0.718	0.930	7	-3.778	1	0.102	0.029	0.000	0.204
0	0	0	0.650	0.833	1	-4.672	1	0.036	0.080	0	0.083
0	0	0	0.675	0.919	8	-5.385	1	0.127	0.080	0	0.143
0	0	0	0.449	0.856	5	-4.788	0	0.062	0.187	0	0.176
0	0	0	0.542	0.903	4	-2.419	0	0.043	0.034	0.000	0.111
0	0	0	0.594	0.935	8	-3.562	1	0.057	0.025	0.000	0.637
0	0	0	0.642	0.818	2	-4.552	1	0.032	0.057	0	0.092
0	0	0	0.679	0.923	6	-6.500	1	0.181	0.146	0.000	0.124
0	0	0	0.437	0.774	8	-4.918	1	0.055	0.148	0	0.133
0	0	0	0.744	0.726	1	-4.675	1	0.046	0.040	0	0.374
0	0	0	0.572	0.915	5	-4.451	0	0.062	0.011	0	0.339
0	0	0	0.690	0.780	5	-4.446	0	0.059	0.007	0.002	0.073
0	0	0	0.805	0.835	0	-4.603	1	0.090	0.130	0.000	0.365
0	0	0	0.694	0.901	1	-4.322	0	0.095	0.070	0	0.427

We can see in starting there are encoded variables based on playlist subgenre, later on there are music related attributes.

3.5 PROVIDE SUMMARY INFORMATION ABOUT THE VARIABLES OF CONCERN IN YOUR CLEANED DATA SET. RATHER, PROVIDE ME WITH A CONSOLIDATED EXPLANATION, EITHER WITH A TABLE THAT PROVIDES SUMMARY INFO FOR EACH VARIABLE OR A NICELY WRITTEN SUMMARY PARAGRAPH WITH INLINE CODE.

Name 	Type	Missing	Statistics			Filter (45 / 45 attributes)
 acousticness	Real	0	Min 0	Max 0.994	Average 0.175	
 danceability	Real	0	Min 0	Max 0.983	Average 0.655	
 days	Real	0	Min 3	Max 23041	Average 2698.176	
 duration_ms	Real	0	Min 4000	Max 517810	Average 225705.458	
 energy	Real	0	Min 0.000	Max 1	Average 0.699	
 instrumentalness	Real	0	Min 0	Max 0.994	Average 0.085	
 key	Real	0	Min 0	Max 11	Average 5.377	
 liveness	Real	0	Min 0	Max 0.996	Average 0.190	
 loudness	Real	0	Min -46.448	Max 1.275	Average -6.716	

Name 	Type	Missing	Statistics			Filter (45 / 45 attrit
 playlist_subgenre = southern h...	Integer	0	Min 0	Max 1	Average 0.051	
 playlist_subgenre = trap	Integer	0	Min 0	Max 1	Average 0.039	
 playlist_subgenre = tropical	Integer	0	Min 0	Max 1	Average 0.039	
 playlist_subgenre = urban cont...	Integer	0	Min 0	Max 1	Average 0.043	
 prediction(track_popularity)	Real	0	Min -11.130	Max 70.565	Average 42.496	
 speechiness	Real	0	Min 0	Max 0.918	Average 0.107	
 tempo	Real	0	Min 0	Max 239.440	Average 120.883	
 track_popularity	Real	0	Min 0	Max 100	Average 42.496	
 valence	Real	0	Min 0	Max 0.991	Average 0.510	

It can be seen the variables does not have any missing values. And their values appear to be in range.

Values of dummies variables are either 0 or 1 as expected.

4.1 UNCOVER NEW INFORMATION IN THE DATA THAT IS NOT SELF-EVIDENT (I.E. DO NOT JUST PLOT THE DATA AS IT IS; RATHER, SLICE AND DICE THE DATA IN DIFFERENT WAYS, CREATE NEW VARIABLES, OR JOIN SEPARATE DATA FRAMES TO CREATE NEW SUMMARY INFORMATION).

New information:

Song popularity is February 2020 can create bias in terms of when it was released, it was necessary to understand the number of days from February 2020 to song release date and use this variable in understanding song's popularity score.

Days since release = Days between Feb 1, 2020 and song release date

4.2 PROVIDE FINDINGS IN THE FORM OF PLOTS AND TABLES. SHOW ME YOU CAN DISPLAY FINDINGS IN DIFFERENT WAYS.

The exercise to drive insights is performed in Tableau.

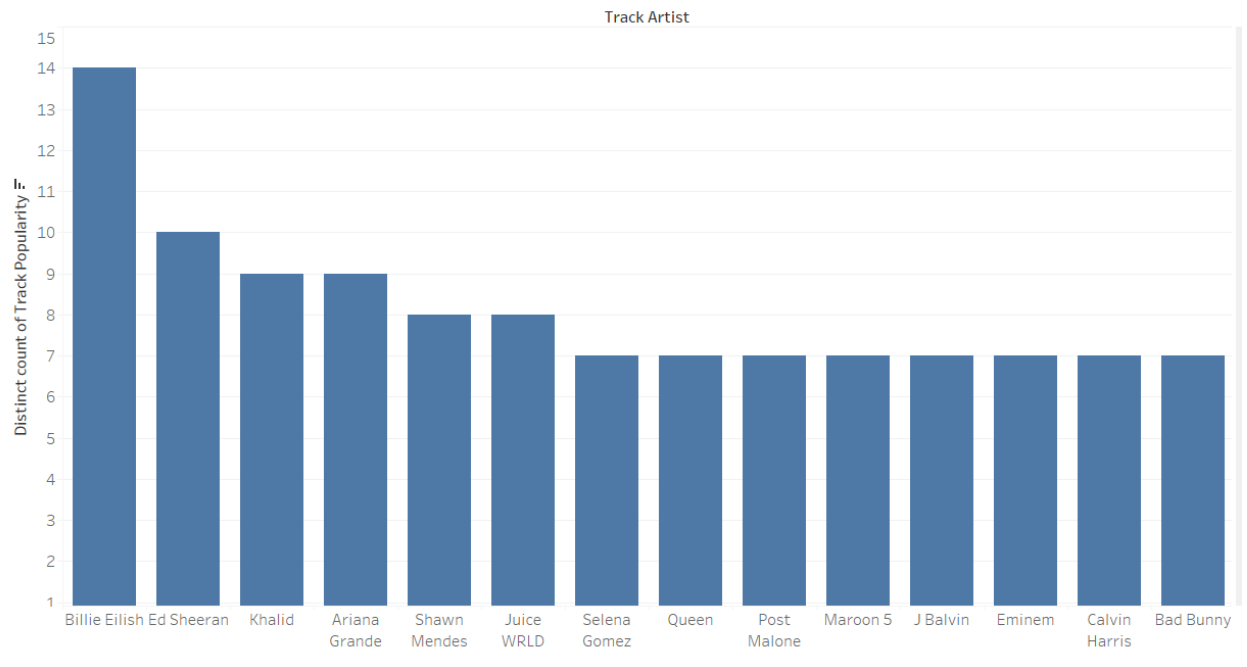
Top Artist:

Slice data for songs with popularity higher than 75 (more than 75%)

Group based on artist and take count of number of songs

Top artists are singers with the greatest number of popular songs

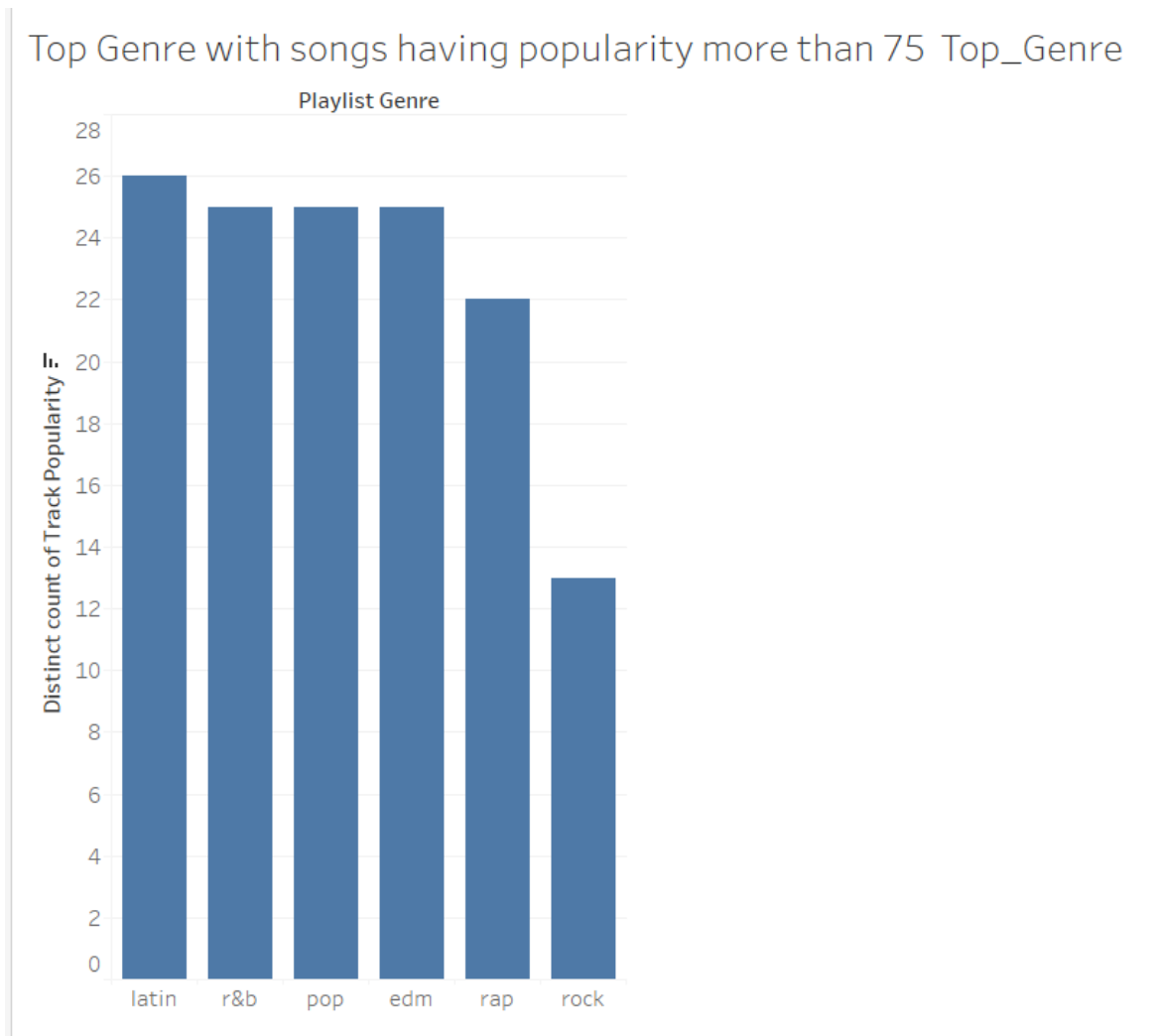
Top Artists with atleast 7 spogs having popularity more than 75 Top_Artist



Top Genre:

There are total 6 genres: RnB, Rock, EDM, Latin, Pop, Rap

Insight is derived to understand which genre drives most hits

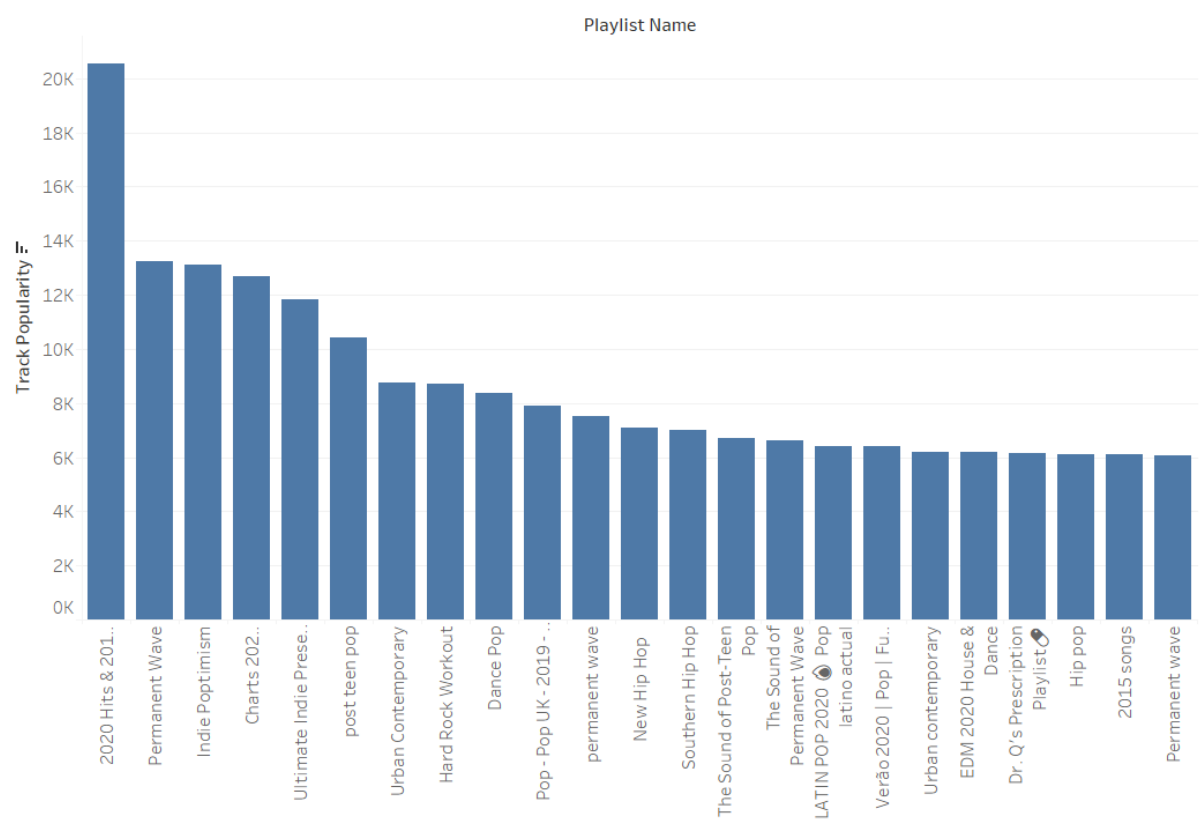


Latin is most popular Genre, followed by r&b, pop and edm. While Rap and Rack music are on lower end.

4.3 GRAPH(S) ARE CAREFULLY TUNED FOR THE DESIRED PURPOSE. ONE GRAPH ILLUSTRATES ONE PRIMARY POINT AND IS APPROPRIATELY FORMATTED (PLOT AND AXIS TITLES, A LEGEND IF NECESSARY, SCALES ARE APPROPRIATE, APPROPRIATE GEOMS USED, ETC.).

When finding information from data, it was vital to see top playlist going in February 2020, and match with our experience & knowledge. Purpose of this plot was more to verify data.

Top Playlists based on the total popularity of songs in it



We can clearly see top playlist have songs from 2020 and 2019, which validates our data and tell us about important of recency in music.

4.4 TABLE(S) CAREFULLY CONSTRUCTED TO MAKE IT EASY TO PERFORM IMPORTANT COMPARISONS. CAREFUL STYLING HIGHLIGHTS IMPORTANT FEATURES. THE SIZE OF THE TABLE IS APPROPRIATE.

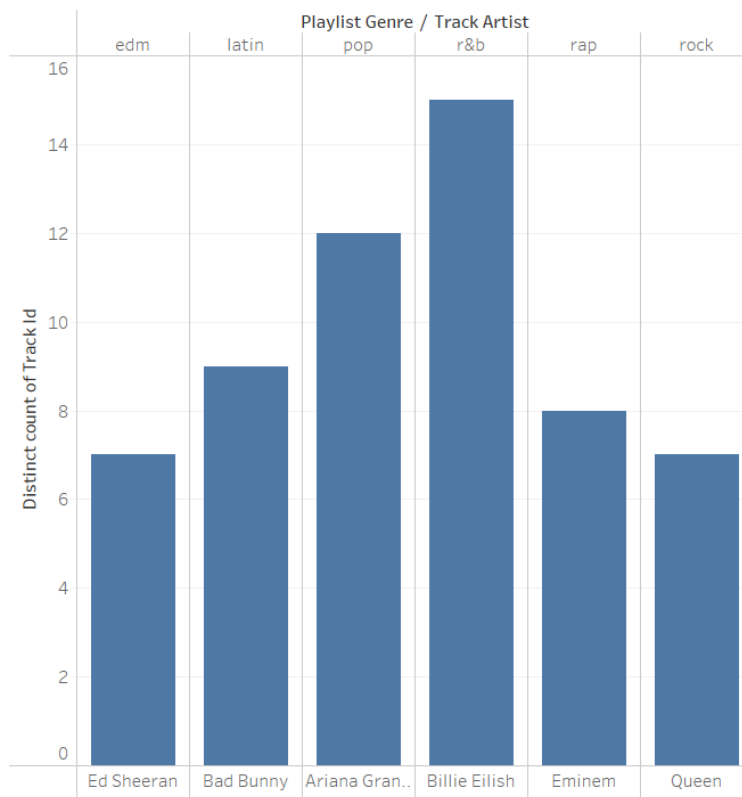
1. Best artist of Genre:
Within each 6 of the genres, who stands with maximum number of hits

Genre	Artist
EDM	Ed Sheeran
Latin	Bad Bunny
Pop	Ariana Grande
R&B	Billie Eilis

Rap	Eminem
Rock	Queen

Although in first 4 genres, current artists are dominating, Eminem is still most popular in Rap and while Freddy's Queen band is prevailing in rock music.

Top artist for each genre based on number of songs in genre with more than 75 popularity



2. Most famous artist of each era in current time

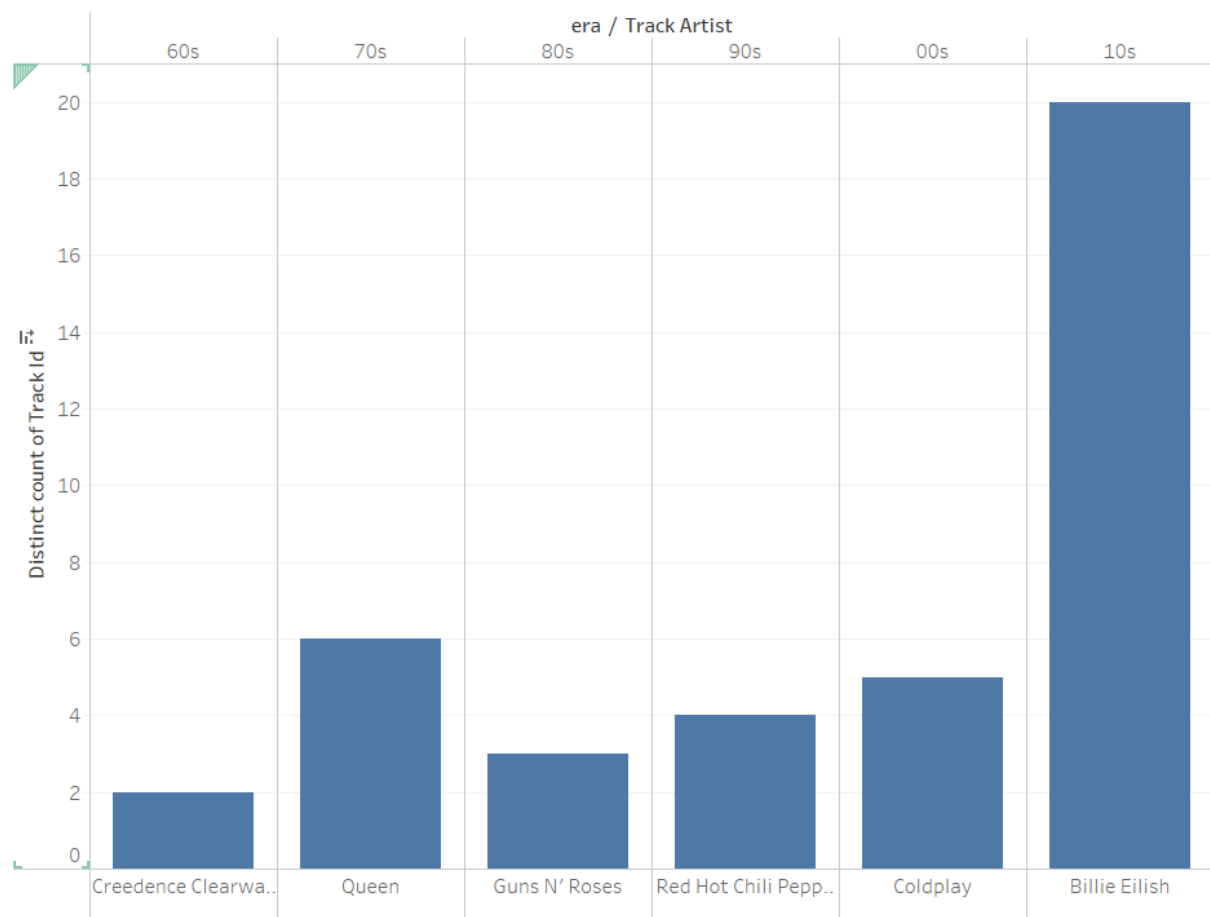
All songs of 60s, 70s, 80s, 90s, 2000s and 2010s are combined and based on the hits, most popular artist of the era is found that is still being heard in early 2020.

Era	Artist
60s	Credence
70s	Queen

80s	Guns n Roses
90s	Red hot chili pepper
2000s	Coldplay
2010s	Billie Eilis

So, out of all 70s bands, Queen has 6 songs in high popularity. Gun n Roses is most listened from 80s.

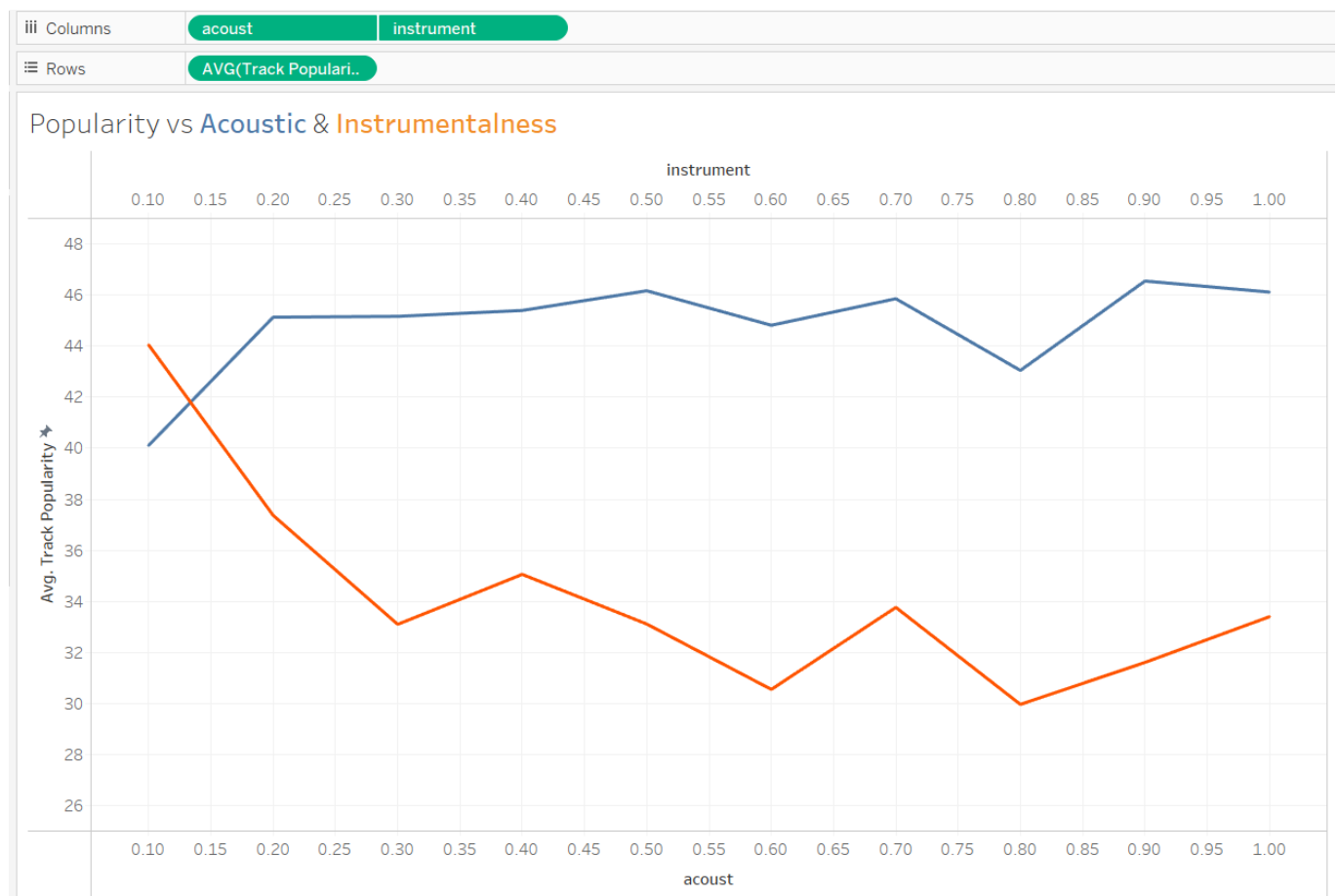
Top_Era_Artist



4.5 INSIGHTS OBTAINED FROM THE ANALYSIS ARE THOROUGHLY, YET SUCCINCTLY, EXPLAINED. EASY TO SEE AND UNDERSTAND THE INTERESTING FINDINGS THAT YOU UNCOVERED.

I wanted to see correlation of popularity of songs with different musical attributes. For this exercise, I created bins for variables.

Bin Number	Minimum	Maximum
1	0	0.1
2	0.1	0.2
3	0.2	0.3
4	0.3	0.4
5	0.4	0.5
6	0.5	0.6
7	0.6	0.7
8	0.7	0.8
9	0.8	0.9
10	0.9	1





For each of these bins, average value of popularity is plots. This exercise is done for below 4 variables:

Variable	Relationship with Popularity
Acoustics	Positive
Instrumentation	Negative
Danceability	Positive
Presence of Live audience in recording	Negative

6.1 SUMMARIZE THE PROBLEM STATEMENT YOU ADDRESSED.

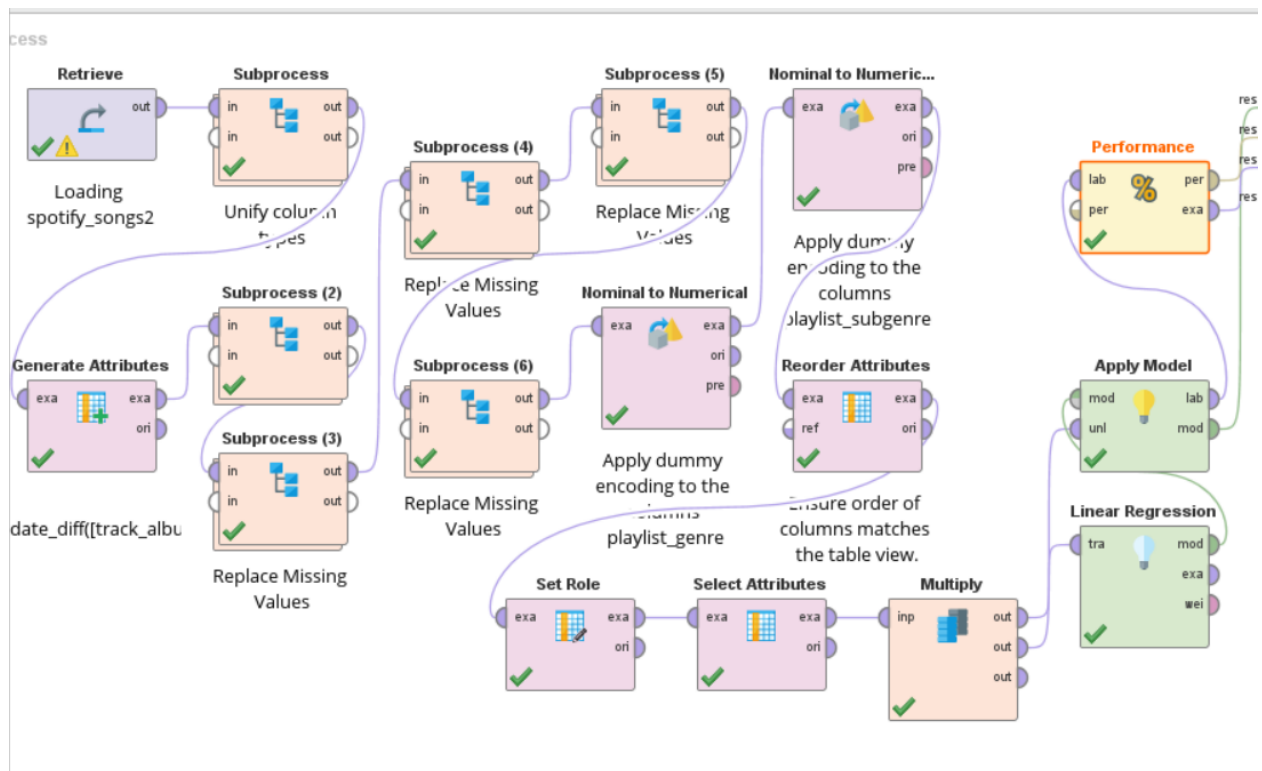
Objective:

Understand the popularity of different songs with the help of insights and model. To objective is to create a map with music attributes and other information of songs with its popularity.

6.2 SUMMARIZE HOW YOU ADDRESSED THIS PROBLEM STATEMENT (THE DATA USED AND THE METHODOLOGY EMPLOYED).

The predictive modeling is performed in DataMiner. Below steps are performed:

1. Load data and check columns data types
2. Check summary statistics and histogram of each column
3. Perform missing value treatment
4. Create dummy variables using one-hot-encoding
5. Set role of Popularity as Label
6. Select useful attributes for the model
7. Build linear regression model using features and label
8. Apply model and assess the model performance



The linear regression model will create the required map of attributes with song popularity.

Model has the Root mean square error of 22.935.

Criterion	
root mean squared error	root_mean_squared_error <code>root_mean_squared_error: 22.935 +/- 0.000</code>

6.3 SUMMARIZE THE INTERESTING INSIGHTS THAT YOUR ANALYSIS PROVIDED.

With the help of linear regression model, we can see the relationship between different variables.

The positive relationship of popularity with Danceability observed in Charts can be seen here with positive coefficient of 9.9. Similarly, negative correlation with instrumentalness can be verified with coefficient of -10.03.

Similarly, the relationship observed in charts can be verified with the equation of linear regression.

Please find the full equation below:

```

0.000 * days
+ 5.793 * playlist_genre = pop
+ 1.224 * playlist_genre = rap
+ 3.886 * playlist_genre = rock
+ 4.601 * playlist_genre = latin
- 0.602 * playlist_genre = r&b
- 2.642 * playlist_genre = edm
+ 4.963 * playlist_subgenre = dance pop
+ 8.331 * playlist_subgenre = post-teen pop
- 2.695 * playlist_subgenre = electropop
- 4.886 * playlist_subgenre = indie poptimism
+ 10.493 * playlist_subgenre = hip hop
- 6.285 * playlist_subgenre = southern hip hop
- 8.745 * playlist_subgenre = gangster rap
+ 5.865 * playlist_subgenre = trap
- 3.033 * playlist_subgenre = album rock
- 0.609 * playlist_subgenre = classic rock
+ 12.427 * playlist_subgenre = permanent wave
- 4.898 * playlist_subgenre = hard rock
- 2.293 * playlist_subgenre = tropical
+ 4.048 * playlist_subgenre = latin pop
+ 5.386 * playlist_subgenre = reggaeton
- 2.609 * playlist_subgenre = latin hip hop
+ 8.194 * playlist_subgenre = urban contemporary
+ 11.294 * playlist_subgenre = hip pop
- 11.124 * playlist_subgenre = new jack swing
- 8.966 * playlist_subgenre = neo soul
+ 0.784 * playlist_subgenre = electro house
- 2.936 * playlist_subgenre = big room
+ 6.558 * playlist_subgenre = pop edm
- 7.058 * playlist_subgenre = progressive electro house
+ 9.926 * danceability
- 25.187 * energy
+ 0.053 * key
+ 1.334 * loudness
- 10.083 * instrumentalness
- 2.154 * liveness
- 1.680 * valence
+ 0.010 * tempo
- 0.000 * duration_ms
+ 66.760

```

6.4 SUMMARIZE THE IMPLICATIONS TO THE CONSUMER OF YOUR ANALYSIS.

Implication to artist:

The predictive model can work as a benchmark to assess the song or track before the release. The track can be readjusted with fine-tuning these attributes. In this way, the model will help the artists in gaining more hit songs.

Implication to listeners:

The usage of model will also help the people to get to listen more likeable songs. And have better music experience.

Implications to music business:

Platforms like Spotify, Amazon Music, Apple music, YouTube music etc. can benefit from increased user time on platform and can generate incremental revenue.

6.5 DISCUSS THE LIMITATIONS OF YOUR ANALYSIS AND HOW YOU, OR SOMEONE ELSE, COULD IMPROVE OR BUILD ON IT.

Limitations:

There are two main limitations of this work:

1. Additional information:

The model here relies on the data it has. There are many factors which are not included in this analysis and will be crucial, e.g. Time from last album, New Technology introduced, Song used in movie etc. These factors can change popularity a lot apart from factors used in this exercise.

2. Artist Bias:

Lot of time, a good song from lesser-known artist may not be that popular compared to average song from star artist.

7.1 ACHIEVEMENT, MASTERY, CLEVERNESS, CREATIVITY: TOOLS AND TECHNIQUES FROM THE COURSE ARE APPLIED VERY COMPETENTLY AND, PERHAPS, SOMEWHAT CREATIVELY. PERHAPS THE STUDENT HAS GONE BEYOND WHAT WAS EXPECTED AND REQUIRED, E.G., EXTRAORDINARY EFFORT, ADDITIONAL TOOLS NOT ADDRESSED BY THIS COURSE, UNUSUALLY SOPHISTICATED APPLICATION OF TOOLS FROM THE COURSE.

Task	Tool
Understand data	Excel
Data visualizations & insights	Tableau
Data cleaning	RapidMiner
Predictive Modeling	RapidMiner