Beat Analytics : Spotify Data Analysis and Song Popularity Prediction

Data Preparation & Analysis (CSP-571)

Project Group Members:

Student Name	A# Number
Kasargod Kailash Chandra Shenoy	A20526053
Aditya Nayak	A20528097
Akshat Behera	A20516439
Veerendra Gopichand Karuturi	A20529571

Project Group Leader:

Student Name	A# Number
Kasargod Kailash Chandra Shenoy	A20526053

Introduction

Our project, "Beat Analytics: Spotify Data Analysis and Song Popularity Prediction," dives deep into Spotify's extensive datasets to analyze musical trends, uncover patterns, and identify what drives song popularity. By harnessing the power of two significant datasets—the "Top Spotify Songs from 2010-2019 by Year" and the "Ultimate Spotify Tracks Database"—our team employs a range of data science methodologies including machine learning, data visualization, and exploratory data analysis.

Using the R programming language, we use its comprehensive and the required libraries for statistical analysis and data manipulation, we develop predictive models that can estimate song popularity based on various musical features such as tempo, energy, and danceability.

The objective of our project is to pinpoint the crucial features that impact a song's popularity on Spotify, to forecast future song popularity, and to detect emerging trends in popular music over the past decade. Through detailed data preparation, feature engineering, and model development & evaluation, we wish to provide valuable insights to musicians, record labels, and industry stakeholders.

Our efforts are directed towards guiding decision-making and uncovering the formula behind that contribute to successful music hits. Ultimately, the project seeks to enhance the application of data science in the music industry, providing a blueprint for how quantitative analysis can influence creative industries.

Problem Statement

The predictability of song popularity is a persistent challenge in our modern digital music landscape due to its rapid evolution and complexity. Even with the abundance of data available, it is still difficult to pinpoint the precise elements that lead to a song's success because they are frequently hidden behind flimsy metrics. The important task of methodically identifying, evaluating, and forecasting these variables is addressed by this research. Our objective is to create a predictive framework that can identify the traits of hit songs and forecast future releases, giving music industry participants a competitive advantage when creating hit songs.

The project seeks to answer the following questions:

What are the key features that influence the popularity of a song on Spotify?

How can we predict the future popularity of a song based on its features?

What trends and patterns can be observed in the top songs from 2010 to 2019?

Project Plan & Detail

Resources

Name	Initials
Kasargod Kailash Chandra Shenoy	KKCS
Aditya Nayak	AN
Akshat Behera	AB
Veerendra Gopichand Karuturi	VGK

Project Plan & Detail

WBS	Activity	Dependency	Beginning Date	Conclusion Date	Assigned Resources	Progress
1	DPA Project		11th Feb	21st Apr		Done
1.1	**Project Group & Topic (PGT)**		**11th Feb**	**18th Feb**		**Done**
1.1.1	Literature Survey		11th Feb	12th Feb	KKCS, AN, AB, VGK	Done
1.1.2	Idea Generation & Finalization	1.1.1	12th Feb	13th Feb	KKCS, AN, AB, VGK	Done
1.1.3	PGT Document	1.1.2	13th Feb	16th Feb		Done
1.1.3.1	Initial Report	1.1.2	13th Feb	14th Feb	KKCS, AN, AB, VGK	Done
1.1.3.2	PGT Report Evaluation Session	1.1.3.1	15th Feb	15th Feb	KKCS, AN, AB, VGK	Done
1.1.3.3	PGT Report Revision	1.1.3.2	16th Feb	17th Feb	KKCS, AN, AB, VGK	Done
1.1.4	Submission on Blackboard	1.1.3.3	18th Feb	18th Feb	KKCS, AN, AB, VGK	Done
1.2	**Data**	**1.1.4**	**16th Feb**	**1st Mar**		**Done**
1.2.1	Data Gathering and Research	1.1.4	16th Feb	19th Feb		Done
1.2.1.1	Assembling all Relevant Datasets	1.1.4	16th Feb	17th Feb	AB	Done
1.2.1.2	Review Meetings of Dataset	1.2.1.1	17th Feb	18th Feb	KKCS, AN, AB,VGK	Done
1.2.1.3	Documentation of Datasets	1.2.1.2	18th Feb	19th Feb	AN	Done
1.2.2	Analysis of Datasets	1.2.1.3	20th Feb	24th Feb		Done
1.2.2.1	Style of Datasets	1.2.1.3	20th Feb	21st Feb	VGK	Done
1.2.2.2	Processing of Dataset	1.2.2.1	22nd Feb	22nd Feb	VGK	Done

1.2.2.3	Meeting Review	1.2.2.2	23rd Feb	23rd Feb	KKCS, AN, AB,VGK	Done
1.2.2.4	Dataset Documents	1.2.2.3	24th Feb	24th Feb	VGK	Done
1.2.3	Dataset Applications of the Project	1.2.2.4	25th Feb	29th Feb		Done
1.2.3.1	Response and Key Predictions	1.2.2.4	25th Feb	26th Feb	KCS	Done
1.2.3.2	Pipelines and Processing of Data	1.2.3.1	27th Feb	27th Feb	AN	Done
1.2.3.3	Review Meeting	1.2.3.2	28th Feb	28th Feb	KKCS, AN, AB,VGK	Done
1.2.3.4	Documentation of Dataset	1.2.3.3	29th Feb	29th Feb	KKCS	Done
1.2.4	Rework	1.2.3.4	1st Mar	1st Mar	AN	Done
1.3	**Analysis Overview**	**1.2.4**	**17th Feb**	**22th Feb**		**Done**
1.3.1	Review of Related Literature	1.2.4	17th Feb	17th Feb	AN, AB	Done
1.3.2	Gathering Research Materials	1.3.1	17th Feb	17th Feb	AB	Done
1.3.3	Synthesis of Key Insights	1.3.2	18th Feb	18th Feb	KKCS, VGK	Done
1.3.4	Pinpointing Crucial Variables	1.3.3	19th Feb	19th Feb	VGK	Done
1.3.5	Justification for Model Choices	1.3.4	20th Feb	21st Feb	KKCS, AN	Done
1.3.6	Overall Review	1.3.1 - 1.3.5	21th Feb	22th Feb	KKCS, AN, AB, VGK	Done
1.4	**Project Software - RStudio Setup**	**1.3.6**	**13th Feb**	**14th Feb**		**Done**
1.4.1	Install Latest RStudio	1.3.6	13th Feb	13th Feb	KKCS, AN, AB, VGK	Done
1.4.1.1	Download RStudio	1.4.1	13th Feb	13th Feb	KKCS, AN, AB, VGK	Done
1.4.1.2	Installation Process	1.4.1.2	13th Feb	13th Feb	KKCS, AN, AB, VGK	Done
1.4.1.3	Installation Verification	1.4.1. - 1.4.1.2	13th Feb	13th Feb	KKCS, AN, AB, VGK	Done
1.4.2	Install Necessary Libraries	1.4.1.3	14th Feb	14th Feb	KKCS, AN, AB, VGK	Done
1.4.2.1	Identify Required Libraries	1.4.2	14th Feb	14th Feb	KKCS, AN, AB, VGK	Done
1.4.2.2	Verify Installation	1.4.2 - 1.4.2.1	14th Feb	14th Feb	KKCS, AN, AB, VGK	Done
1.5	**Project Proposal & Outline (PPO)**	**1.3.6 - 1.4.2.2**	**18th Feb**	**3rd Mar**		**Done**
1.5.1	Proposal Documentation	1.3.6	18th Feb	24th Feb	KKCS, AN, AB, VGK	Done
1.5.1.1	Project Description	1.3.6	18th Feb	19th Feb	AN	Done
1.5.1.2	Questions	1.5.1.1	20th Feb	20th Feb	AB	Done
1.5.1.3	Proposed Methodology	1.5.1.2	21st Feb	21st Feb	KKCS	Done
1.5.1.4	Outlining Metrics	1.5.1.3	22nd Feb	22nd Feb	VGK	Done
1.5.2	Outline Documentation	1.5.1.4	25th Feb	29th Feb	KKCS, AN, AB, VGK	Done
1.5.2.1	Literature Review - References - Supplemental Resources	1.3.1 - 1.3.2	25th Feb	25th Feb	VGK, AB	Done
1.5.2.2	Dataset and Sources - Overview and Feature Description	1.2 - 1.5.2.1	26th Feb	26th Feb	AB	Done
1.5.2.3	Data Processing and Pipeline & Data Stylized Facts	1.5.2.2	27th Feb	27th Feb	KKCS, AN	Done
1.5.2.4	Model Selection	1.3.5 - 1.5.2.3	28th Feb	29th Feb	KKCS, AB	Done
1.5.2.5	Software and Libraries	1.5.2.4	29th Feb	29th Feb	AN, VGK	Done
1.5.3	Refining the PPO	1.5.2.5	1st Mar	12th Mar	AB, AN	Done
1.5.4	Final Review	1.5.3	12th Mar	3rd Mar	KKCS, AN, AB, VGK	Done

1.5.5	Submission Preparation	1.5.4	12th Mar	3rd Mar	KKCS, AN, AB, VGK	Done
1.6	**Coding**	**1.5.5**	**6th Mar**	**9th Apr**		Done
1.6.1	Data Handling	1.5.5	6th Mar	10th Mar		Done
1.6.1.1	Scripting	1.5.5	10th Mar	16th Mar	KKCS	Done
1.6.1.2	Review Session	1.6.1.1	16th Mar	17th Mar	KKCS, AN, AB, VGK	Done
1.6.1.3	Modifications	1.6.1.2	17th Mar	19th Mar	AN	Done
1.6.2	Charting	1.6.1.3	19th Mar	20th Mar		Done
1.6.2.1	Scripting	1.6.1.3	15th Mar	26th Mar	AB, VGK	Done
1.6.2.2	Review Session	1.6.2.1	22nd Mar	27th Mar	KKCS, AN, AB, VGK	Done
1.6.2.3	Modifications	1.6.2.2	23rd Mar	24th Mar	AB	Done
1.6.3	Linear Models	1.6.1.3	25th Mar	26th Mar		Done
1.6.3.1	Scripting	1.6.1.3	26th Mar	27th Mar	AN, AB	Done
1.6.3.2	Review Session	1.6.3.1	27th Mar	28th Mar	KKCS, AN, AB, VGK	Done
1.6.3.3	Modifications	1.6.3.2	28th Mar	29th Mar	KKCS	Done
1.6.4	Combine Versions 1.6.1-1.6.3	1.6.2.3 1.6.3.3	29th Mar	1st Apr		Done
1.6.4.1	Scripting	1.6.2.3 1.6.3.3	1st Apr	2nd Apr	KKCS, VGK	Done
1.6.4.2	Review Session	1.6.4.1	2nd Apr	4th Apr	KKCS, AN, AB, VGK	Done
1.6.4.3	Modifications	1.6.4.3	4th Mar	9th Apr	KKCS, AN, AB, VGK	Done
1.7	**Project Plan & Detail**	**1.1.4 - 1.6**	**12th Mar**	**24th Mar**		**Done**
1.7.1	Documentation Creation	1.1.4 - 1.6	12th Mar	14th Mar	KKCS, AN, AB, VGK	Done
1.7.2	Documentation Review	1.7.1	14th Mar	16th Mar	KKCS, AN, AB, VGK	Done
1.7.3	Documentation Revision	1.7.2	16th Mar	20th Mar	KKCS, AN, AB, VGK	Done
1.7.4	Submit on Blackboard	1.7.3	20th Mar	24th Mar	KKCS, AN, AB, VGK	Done
1.8	**Project Demo & Report**	**1.7.4**	**8th Apr**	**17th Apr**		Done
1.8.1	Report Creation	1.6.4.3	8th Apr	14th Apr		Done
1.8.1.1	Report Drafting	1.6.4.3	14th Apr	15th Apr	KKCS, AN, AB, VGK	Done
1.8.1.2	Report Review	1.8.1.1	15th Apr	16th Apr	KKCS, AN, AB, VGK	Done
1.8.1.3	Report Revision	1.8.1.2	16th Apr	17th Apr	KKCS, AN, AB, VGK	Done
1.8.2	Presentation Preparation	1.6.4.3	17th Apr	18th Apr		Done
1.8.2.1	Presentation Draft	1.6.4.3	18th Apr	19th Apr	KKCS, AN, AB, VGK	Done
1.8.2.2	Presentation Review	1.8.2.1	19th Apr	20th Apr	KKCS, AN, AB, VGK	Done
1.8.2.3	Presentation Revision	1.8.2.2	20st Apr	20th Apr	KKCS, AN, AB, VGK	Done
1.8.3	Final Adjustments	1.8.2.3	20th Apr	21st Apr	KKCS, AN, AB, VGK	Done
1.8.4	Report and Presentation Submission	1.8.3	21st Apr	21st Apr	KKCS, AN, AB, VGK	Done

Dataset

a) Top Spotify Songs from 2010-2019 by Year

Description: This dataset comprises approximately 600 songs that were among the top songs of the year from 2010 to 2019, as measured by Billboard. It includes 13 features for exploration.

Source: Kaggle/DataCamp Dataset – Top Spotify Songs 10’-19

Data Origin: Extracted from organizeyourmusic.playlistmachinery.com

b) Ultimate Spotify Tracks Database

Description: This dataset provides comprehensive information on Spotify tracks, including various features such as acousticness, danceability, energy, etc., along with the popularity of the songs.

It has 232,725 tracks (i.e. Observations) spread across 26 genres of music. Where each genre approximately has 10,000 songs belonging to a particular genre. along with that 18 columns (i.e., Features) which can be used for exploration and analysis.

Source: Kaggle Dataset – Ultimate Spotify Dataset

Spotify Web API documentation on getting audio features: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

Methodology

1. Data Sources

Top Songs (2010-2019) & Ultimate Tracks Database: Features like genre, danceability, popularity.

2. Data Preprocessing

Clean-Up: Handle missing data, remove duplicates, standardize data formats.

3. Feature Engineering

Aggregations & Transformations: By artist, genre, and year.

4. Exploratory Data Analysis (EDA)

Visuals & Correlations: Analyze feature distributions and relationships.

5. Model Development

Train & Test: Use algorithms like logistic regression, decision trees, NB, and KNN.

6. Model Evaluation

Metrics: R-squared, MAE, RMSE, precision, recall, F1-score, AUC.

7. Insights

Drivers of Popularity: Key findings and recommendations for stakeholders.

Data Processing & Pipeline

Data Loading: Loaded the datasets using R's read.csv() function for "top10s.csv" and read_csv() for "SpotifyFeatures.csv".

Data Cleaning: Checked for missing values (sum(is.na(data))) and duplicate entries (sum(duplicated(data)==TRUE)).

1. Adjustments to the data structure, such as removing an unspecified column from "top10s.csv" and converting the 'year' column to a factor to better handle categorical analysis.

Exploratory Data Analysis (EDA): Visualization of data distributions and relationships using ggplot2, with plots like histograms, box plots, and density plots to explore various features such as bpm, energy, danceability, and popularity.

1. Using the 'ggpairs' to display pairwise relationships and distributions among several musical features.

Feature Manipulation: The modification of song attributes, such as converting year into a categorical factor, which is typical in preparing data for analysis that involves trends over time.

Pre-processing for Modeling: Features like 'key', 'mode', and 'time_signature' are converted to numeric or binary formats to prepare for modeling.

The data is split into training and testing sets to evaluate the model's performance accurately.

Data Stylization

Consistency in Categorical Data: The transformation of year to a factor indicates standardization of categorical data for consistency in visual and statistical analysis.

1. Factor adjustment for plotting to ensure that categorical levels (e.g., artist) are consistent and meaningful when visualized, particularly in functions like geom_bar() and coord_flip().

Visualization Enhancements: Using scale_fill_viridis_d() to maintain a consistent and accessible color palette across different plots.

1. Enhancement of plot aesthetics such as titles, axis labels, and legends to improve readability and interpretation.

Consolidation: Aggregation based on artist, genre, or year to facilitate group-wise analysis and comparisons.

Feature Extraction

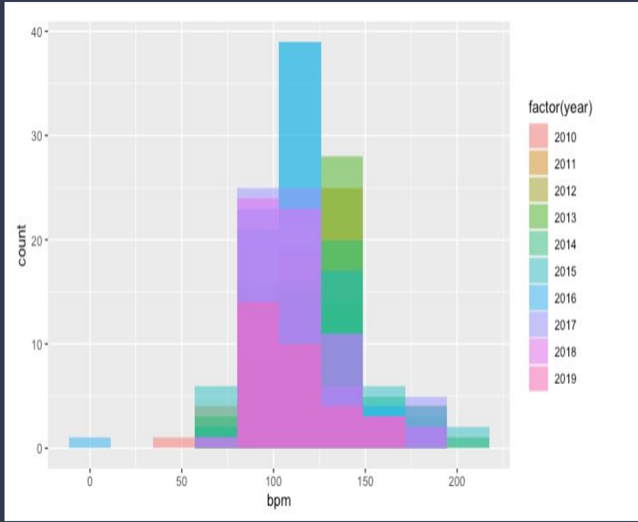
Top10s.csv: It has 600 songs (i.e., Observations), along with 14 columns (i.e., Features) from which 13 can be used for the exploration and analysis.

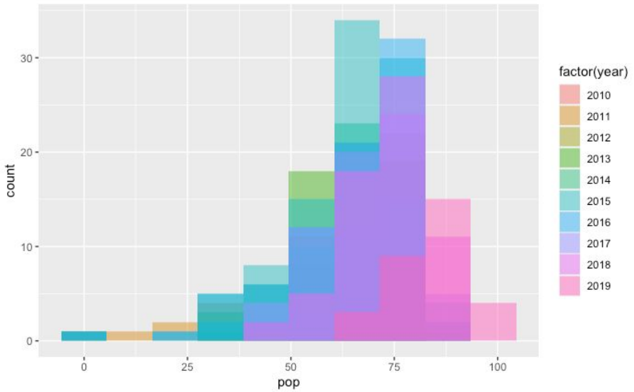
Field Name	Type	Description
title	String	The title of the song.
artist	String	The artist or performer of the song.
top genre	String	The top genre classification of the song.
year	Integer	The year the song was released.
bpm	Integer	Beats per minute (tempo) of the song.
nrgy	Integer	Energy level of the song, likely subjective.
dnce	Integer	Danceability rating of the song.
dB	Integer	Loudness of the song in decibels.
live	Integer	Likelihood of the song being performed live.
val	Integer	Positivity or mood of the song.
dur	Integer	Duration of the song in seconds.
acous	Integer	Acousticness of the song.
spch	Integer	Presence of spoken words in the song.
pop	Integer	Popularity rating of the song, possibly subjective.

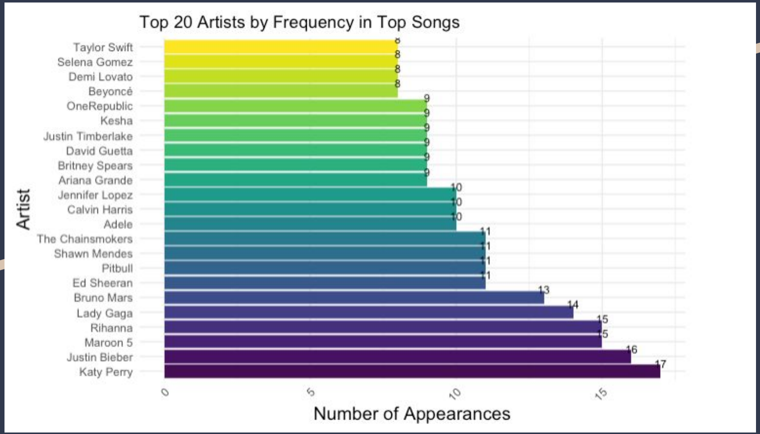
SpotifyFeatures.csv: It has approximately 232,725 tracks (i.e. Observations) spread across 26 genres of music. Where each genre approximately has 10,000 songs belonging to a particular genre. along with that 18 columns (i.e., Features) which can be used for exploration and analysis.

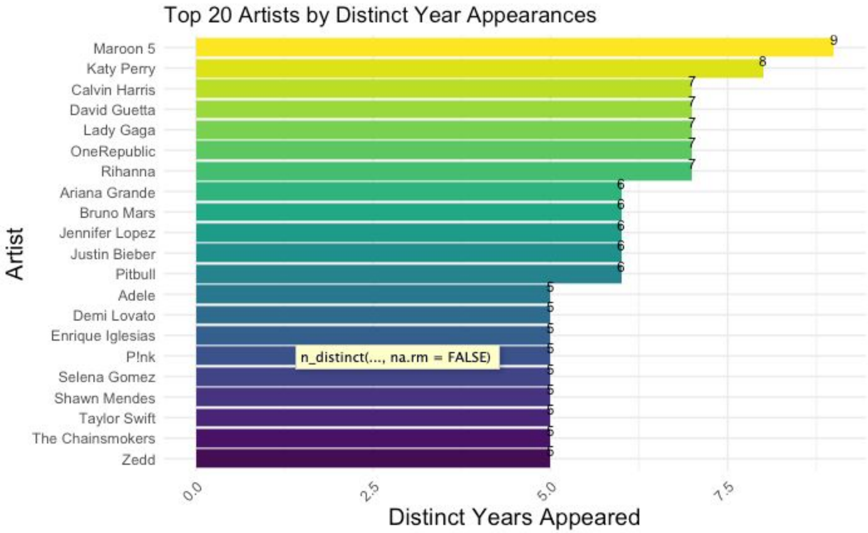
Field Name	Type	Description
genre	String	The genre of the track.
artist_name	String	The name of the artist who performed the track.
track_name	String	The name or title of the track.
track_id	String	A unique identifier for the track.
popularity	Integer	The popularity score of the track.
acousticness	Float	Measure of the acoustic characteristics of the track.
danceability	Float	measure of how suitable a track is for dancing.
duration_ms	Integer	The duration track in milliseconds.
energy	Float	Measure of the energy of the track.
instrumentalness	Float	measure of the presence of instrumental sounds in the track.
key	Integer	The key the track is in.
liveness	Float	Measure of presence of a live audience in the recording.
loudness	Float	The loudness of track in d (dB).
mode	Integer	The modality of the track (major or minor).
speechiness	Float	Measure of the presence of spoken words in the track.
tempo	Float	The tempo of the track in (BPM).
time_signature	Integer	The time signature of the track.
valence	Float	measure of the musical positiveness conveyed by a track.

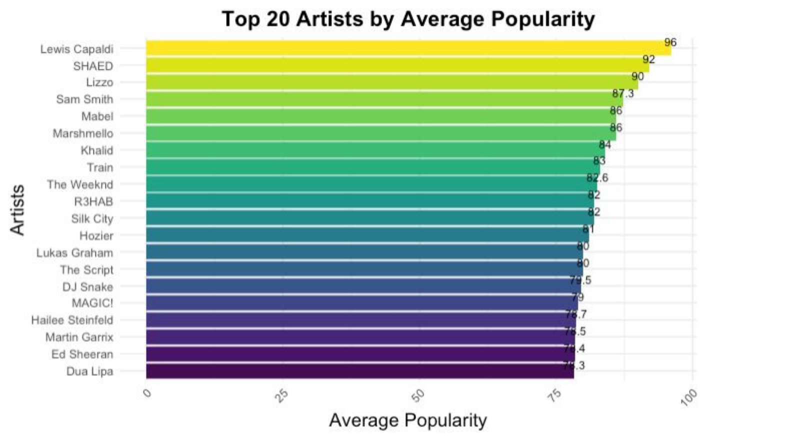
Data Visualization - top10s.csv

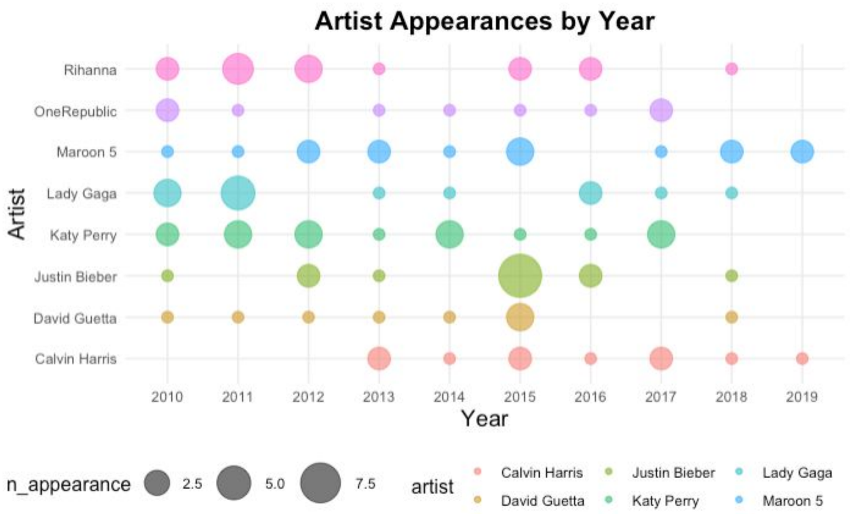


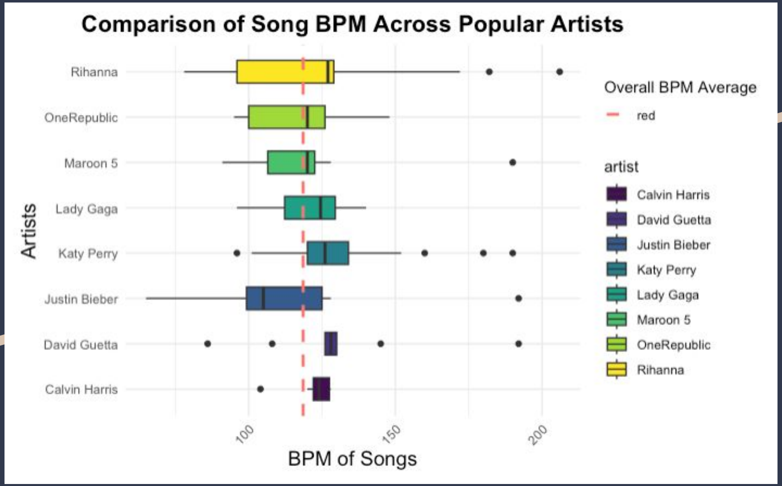


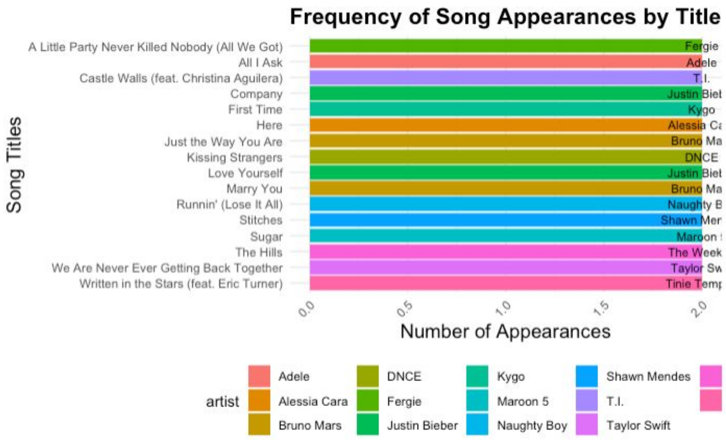


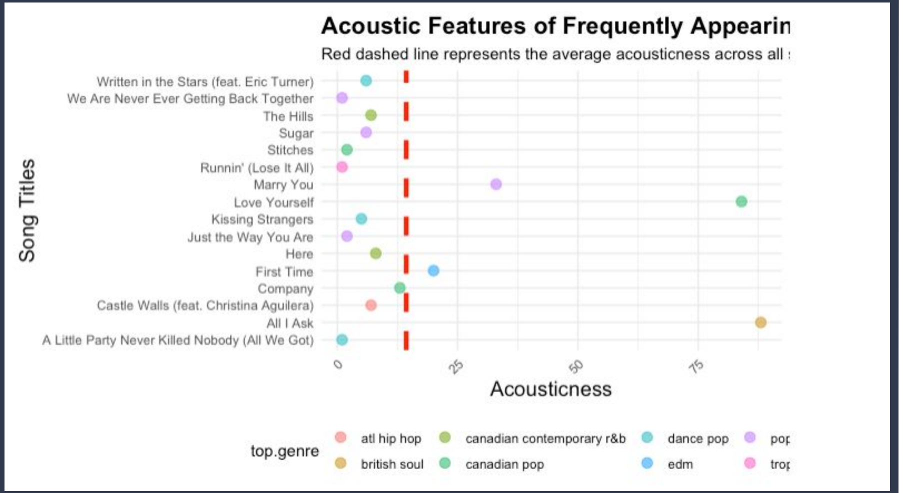


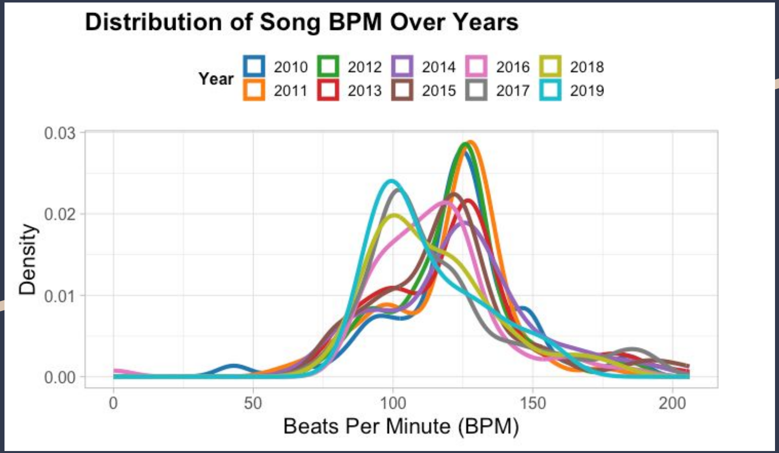

Top 20 Artists by Average Popularity
Lewis Capaldi
SHAED
Lizzo
Sam Smith
Mabel
Marshmello
Khalid
Train
The Weeknd
R3HAB
Silk City
Hozier
Lukas Graham
The Script
DJ Snake
MAGIC!
Hailee Steinfeld
Martin Garrix
Ed Sheeran
Dua Lipa
96
92
90
87.3
86
86
84
83
82.6
82
82
81
80
80
79.5
79
Artists
0
25
50
75
100
Average Popularity

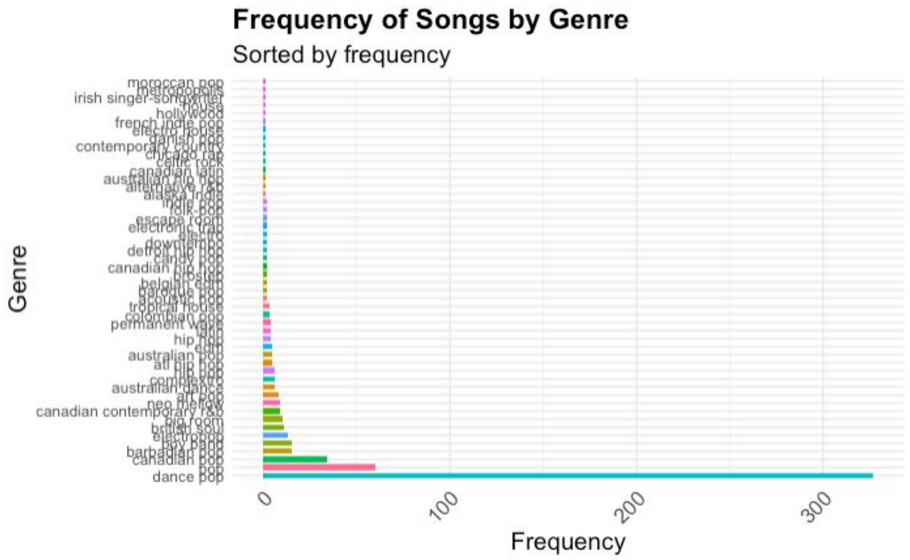

Artist Appearances by Year
Rihanna
OneRepublic
Maroon 5
Lady Gaga
Katy Perry
Justin Bieber
David Guetta
Calvin Harris
Artist
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year
n_appearance
2.5
5.0
7.5
artist
Calvin Harris
Justin Bieber
Lady Gaga
David Guetta
Katy Perry
Maroon 5


Comparison of Song BPM Across Popular Artists
Rihanna
OneRepublic
Maroon 5
Lady Gaga
Katy Perry
Justin Bieber
David Guetta
Calvin Harris
Artists
100
150
200
BPM of Songs
Overall BPM Average
red
artist
Calvin Harris
David Guetta
Justin Bieber
Katy Perry
Lady Gaga
Maroon 5
OneRepublic
Rihanna

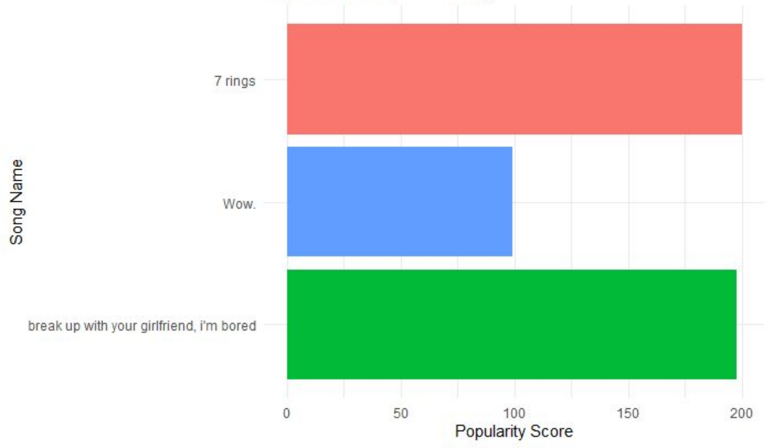

Frequency of Song Appearances by Title
A Little Party Never Killed Nobody (All We Got)
All I Ask
Castle Walls (feat. Christina Aguilera)
Company
First Time
Here
Just the Way You Are
Kissing Strangers
Love Yourself
Marry You
Runnin' (Lose It All)
Stitches
Sugar
The Hills
We Are Never Ever Getting Back Together
Written in the Stars (feat. Eric Turner)
Song Titles
Fergie
Adele
T.I.
Kygo
DNCE
0.0
0.5
1.0
1.5
2.0
Number of Appearances
artist
Adele
Alessia Cara
Bruno Mars
DNCE
Fergie
Justin Bieber
Kygo
Maroon 5
Naughty Boy
Shawn Mendes
T.I.
Taylor Swift

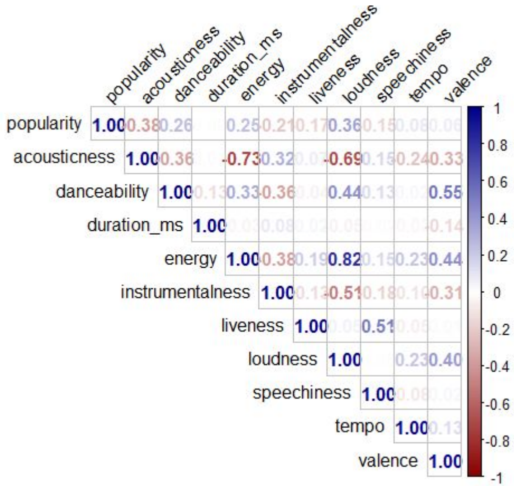

Acoustic Features of Frequently Appearin
Red dashed line represents the average acousticness across all
Song Titles
Written in the Stars (feat. Eric Turner)
We Are Never Ever Getting Back Together
The Hills
Sugar
Stitches
Runnin' (Lose It All)
Marry You
Love Yourself
Kissing Strangers
Just the Way You Are
Here
First Time
Company
Castle Walls (feat. Christina Aguilera)
All I Ask
A Little Party Never Killed Nobody (All We Got)
0
25
50
75
Acousticness
top.genre
atl hip hop
british soul
canadian contemporary r&b
canadian pop
dance pop
edm
pop
trop

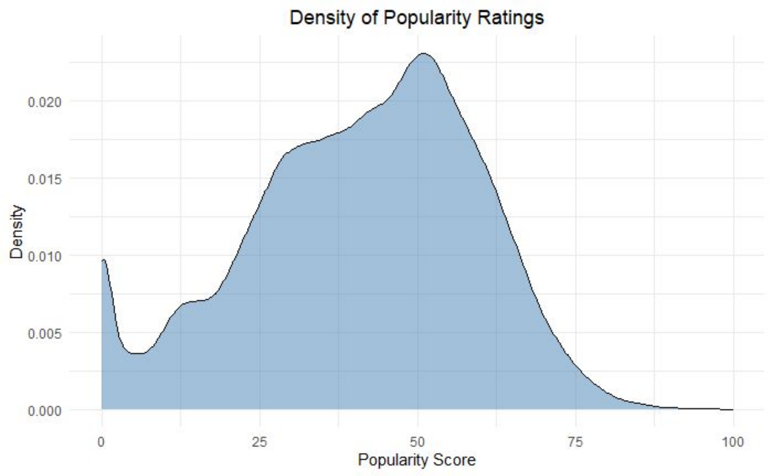

Distribution of Song BPM Over Years
Year
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Density
0.03
0.02
0.01
0.00
0
50
100
150
200
Beats Per Minute (BPM)

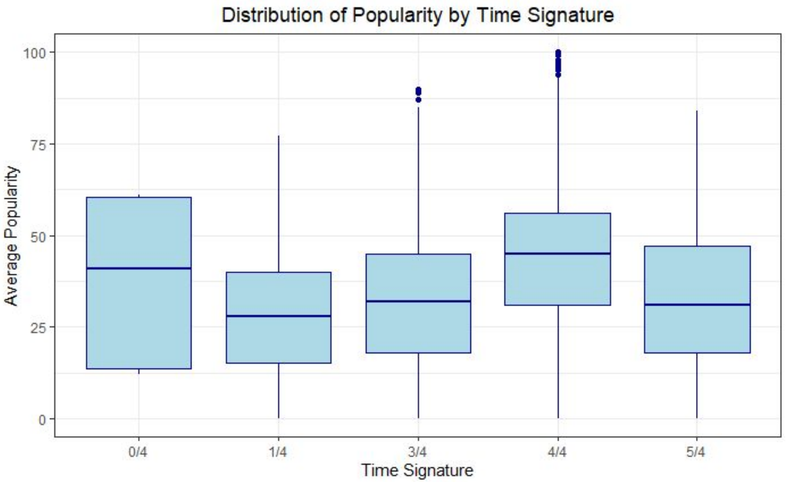

Frequency of Songs by Genre
Sorted by frequency
Genre
moroccan pop
irish singer-songwriter
contemporary country
canadian contemporary r&b
dance pop
0
100
200
300
Frequency

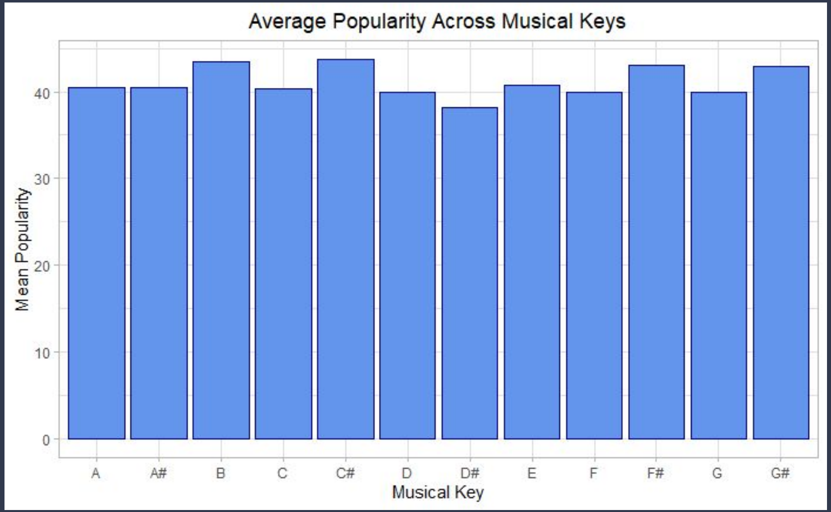
Data Visualization - SpotifyFeatures.csv

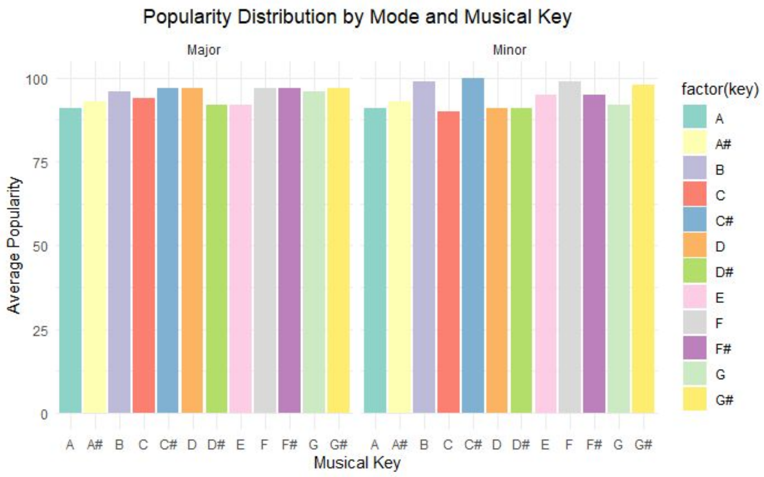


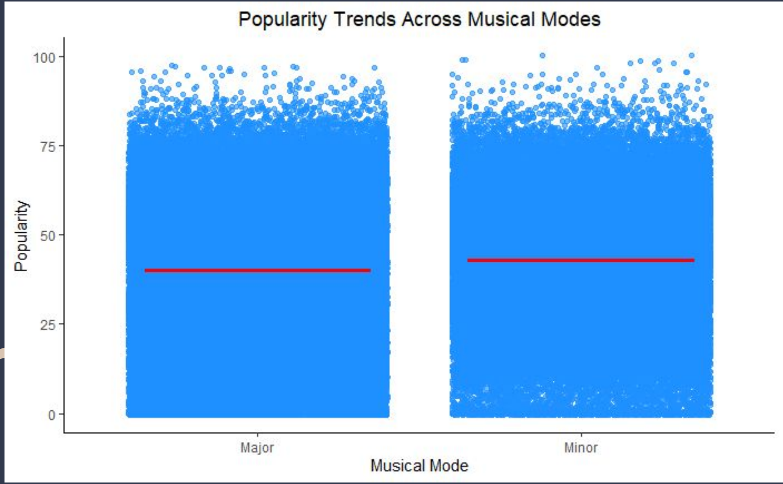


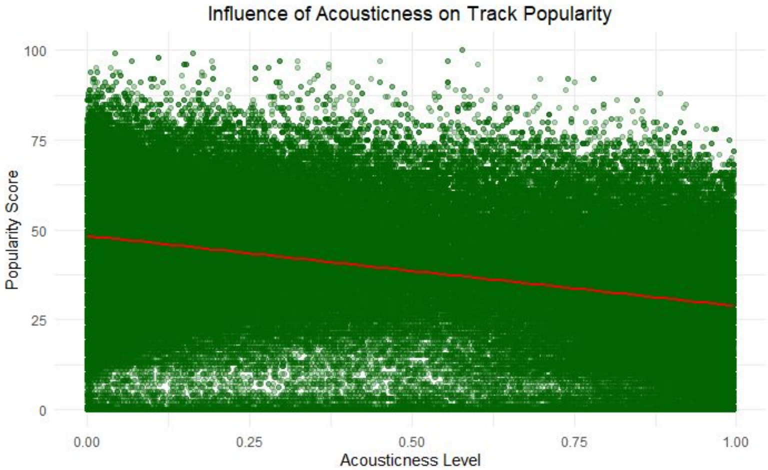


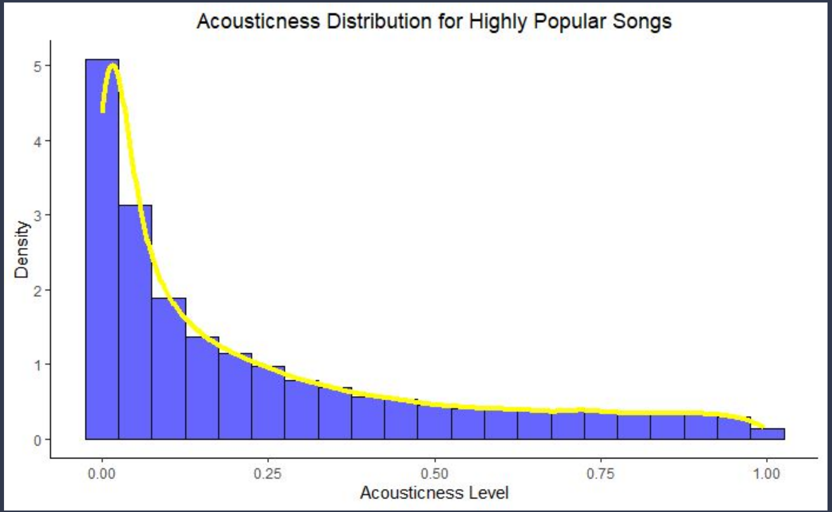


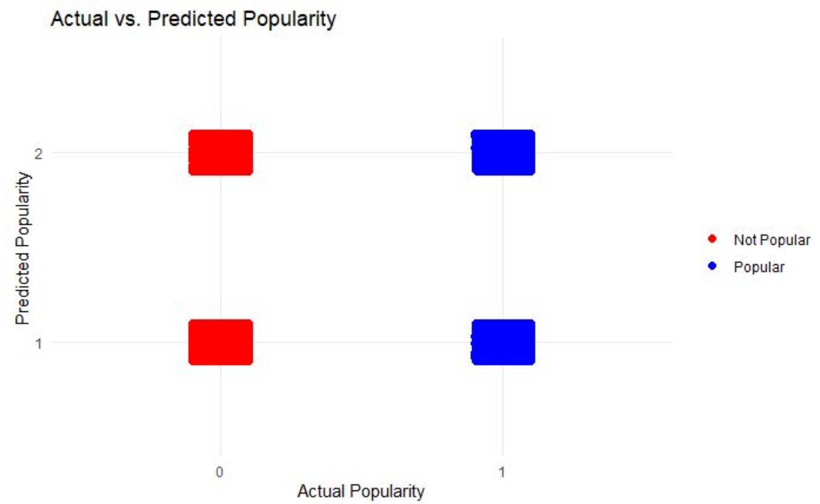

Average Popularity Across Musical Keys
Mean Popularity
0
10
20
30
40
A
A#
B
C
C#
D
D#
E
F
F#
G
G#
Musical Key

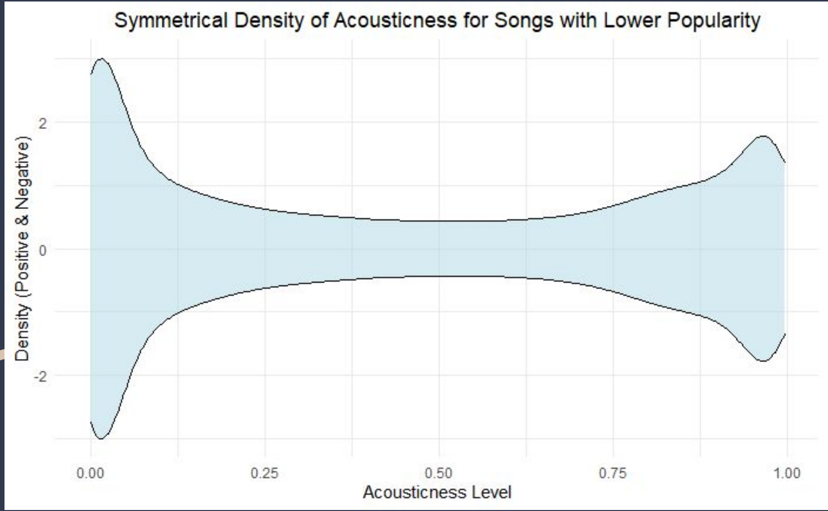

Popularity Distribution by Mode and Musical Key
Major
Minor
Average Popularity
0
25
50
75
100
A A# B C C# D D# E F F# G G#
A A# B C C# D D# E F F# G G#
Musical Key
factor(key)
A
A#
B
C
C#
D
D#
E
F
F#
G
G#

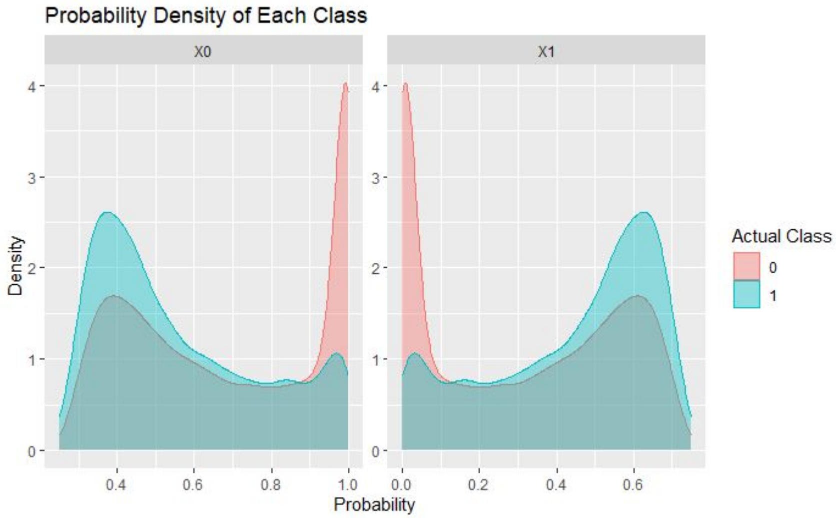

Popularity Trends Across Musical Modes
Popularity
0
25
50
75
100
Major
Minor
Musical Mode


Influence of Acousticness on Track Popularity
Popularity Score
0
25
50
75
100
0.00
0.25
0.50
0.75
1.00
Acousticness Level


Acousticness Distribution for Highly Popular Songs
Density
Acousticness Level
0.00
0.25
0.50
0.75
1.00


Actual vs. Predicted Popularity
Predicted Popularity
Actual Popularity
Not Popular
Popular


Symmetrical Density of Acousticness for Songs with Lower Popularity
Density (Positive & Negative)
Acousticness Level


Probability Density of Each Class
X0
X1
Density
Probability
Actual Class
0
1

Model Selection

The models were selected in a way that addresses both the continuous nature of the popularity scores and the categorical outcomes of song classification (popular/not popular).

- **Logistic Regression**: For predicting whether a song is popular or not based on a binarized popularity score.
- **Decision Tree**: It was selected for its ability to handle nonlinear relationships and their robustness against overfitting.
- **Naive Bayes**: A simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features.
- **K-Nearest Neighbors (KNN)**: Its selection was based on its efficacy in capturing complex patterns by considering the proximity of similar data points.
- **Predictors**:
  These are the predictors used to input into the models to predict the response variable:

Acousticness
Danceability
Energy
Tempo
Speechiness
Key
Mode

- **Response**:
  The response variable is what the models aim to predict based on the predictors. In our project, the response variable is:
  - Popularity Score: This metric tells how often the song is played and how recent those plays are. It is used to gauge the success and reach of the tracks on Spotify, making it a key outcome variable for our analysis

Evaluation Metrics

In predictive modeling and data analysis, evaluation metrics are crucial for assessing the performance of models. The Metrics Used in Our Project:

- **R-squared (R^2)**: For predicting song popularity from features like acousticness, danceability, and energy, It quantifies how well the variations in song popularity.
- **Mean Absolute Error (MAE)**: To assess the accuracy of continuous variables such as song popularity, providing a clear measure of prediction accuracy on a comprehensible scale.
- **Root Mean Squared Error (RMSE)**: To evaluate regression models where minimizing large errors is critical, ensuring the model's predictive accuracy across various song features.
- **Accuracy**: To predict whether a song is popular, accuracy measures the overall correctness of the model in classifying songs as hits or non-hits.
- **Confusion Matrix**: For predicting categorical outcomes (popular/not popular), the confusion matrix helps in visualizing the model's performance.
- **Area Under the Curve (AUC):** It helps in judging how well the model can distinguish between the two classes (popular vs. not popular).
- **Precision, Recall, and F1-Score**: In music popularity predictions, ensuring a song identified as a potential hit truly has the characteristics of a hit (high precision), or that the model captures as many actual hits as possible (high recall).

Model Validation

Testing Results

Model Type	Accuracy	MAE	RMSE	F1 Score	AUC
Logistic Regression	78.73%	0.3138	0.3976	88.10%	66.55%

Model Type	Accuracy	MSE	RMSE	F1 Score
Decision Tree	78.73%	0.1675	0.4092	88.10%

Model Type	Accuracy	MSE	RMSE	F1 Score
Naive Bayes	64.02%	0.2020	0.4494	74.77%

```
[1] "Naive Bayes Model"
         Actual
Predicted     0     1
        0 24808  4908
        1 11837  4992
```

Model Type	Accuracy	Precision	Recall	F1 Score	AUC
KNN	77.46%	84.92%	86.79%	85.84%	64.86%

```
[1] "KNN Model Confusion Matrix"
         Actual
Predicted     0     1
        0 31804  5649
        1  4841  4251
```

Performance Criteria

The performance criteria we used for our model evaluation and comparison are as follows:

1. **Accuracy**: High accuracy is crucial for our case where correct overall predictions are paramount, affecting every decision based on the model’s output.

2. **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)**: MSE measures the average of the squares of the errors—i.e., the average squared difference between estimated values and the actual value. RMSE is the square root of MSE.

3. **Precision**: In our scenario, it is critical as high precision ensures that the model's positive predictions are reliable, reducing unnecessary expenses or actions based on incorrect data.

4. **Recall**: It ensures that most positive cases are caught, even if some false positives occur.

5. **F1 Score**: The F1 score becomes a key metric as It helps optimize models to find an effective balance between recall and precision, which is important for maintaining a stable performance.

6. **AUC**: It provides us a measure of how well the model can discriminate between the classes across different thresholds. Higher AUC indicates a better performing model in terms of its capability to differentiate between the +ve and -ve classes.

Conclusion

Analysis of the top songs from 2010 to 2019 revealed that certain features like danceability, energy, and acousticness consistently correlate with higher popularity scores.

Visual and statistical analyses indicated that trends in music preference might shift over time, reflecting changes in listener demographics and technological advancements in music consumption.

Upon evaluating our models we see that Logistic Regression emerged as the most balanced model, offering a robust combination of accuracy, F1 score, and a reasonable AUC. The Decision Tree model matched Logistic Regression in terms of accuracy and F1 score but was notably inferior in its ability to discriminate between classes. Naive Bayes, while the fastest to implement, lagged behind in accuracy and F1 score. KNN demonstrated commendable precision and recall, making it suitable for project objectives.

Overall, these models were able to predict song popularity with reasonable accuracy giving us a positive result w.r.t our project objectives, highlighting the importance of certain features in influencing a song's success on Spotify.

Future Work / Recommendations

Future work upon this project could involve expanding the dataset to include newer data and additional variables or metrics like social media influence and artist popularity. More could be explored in the direction of more complex models such as neural networks and ensemble methods to improve prediction accuracy.

Source Links:

1. **Top Spotify Songs from 2010-2019 by Year**

   Description: This dataset comprises approximately 600 songs that were among the top songs of the year from 2010 to 2019, as measured by Billboard. It includes 13 features for exploration.

   Source: https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year

   Data Origin: Extracted from http://organizeyourmusic.playlistmachinery.com/

2. **Ultimate Spotify Tracks Database**

   Description: This dataset provides comprehensive information on Spotify tracks, including various features such as acousticness, danceability, energy, etc., along with the popularity of the songs.

   Source: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db#SpotifyFeatures.csv

   Additional Resources: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

Source Code GitHub Link:
https://github.com/AkshatBehera/CSP571-DPA-Project-BeatAnalytics-Spotify

Bibliography

1. Thompson, Andrew, Matt Daniels, and Damián Gaume. "Are Hit Songs Becoming Less Musically Diverse?" The Pudding. Accessed. https://www.the-pudding.com/

2. Nasreldin, Mohamed. "Song Popularity Predictor." Towards Data Science, 2020. Accessed. https://towardsdatascience.com/song-popularity-predictor-1ef69735e380

3. Agarwal, S., R. Cherakkara, and M. Jhonnalagadda. "Predicting Music Popularity with Machine Learning." In Proceedings of the IEEE Conference on Machine Learning, 1-6. IEEE Xplore, 2018.

4. Mauch, M., R. M. MacCallum, M. Levy, and A. M. Leroi. "Measuring the Evolution of Contemporary Western Popular Music." PLOS ONE 10, no. 4 (2015): e0131732. doi:10.1371/journal.pone.0131732.

5. Lau, Cher. "5 Steps of a Data Science Project Lifecycle." Towards Data Science, 2019. Accessed. https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

6. Wickham, Hadley, and Garrett Grolemund. R for Data Science. Accessed. https://r4ds.had.co.nz/

7. Spotify Engineering. "Insights into Data Analysis and Engineering Practices." Spotify Engineering Blog. Accessed. https://engineering.atspotify.com/

8. Harvey, Alan. Music, Evolution, and the Harmony of Souls. Cambridge: Cambridge University Press, 2017.

THANK YOU

Project Group Members:

Student Name	A# Number
Kasargod Kailash Chandra Shenoy	A20526053
Aditya Nayak	A20528097
Akshat Behera	A20516439
Veerendra Gopichand Karuturi	A20529571

Project Group Leader:

Student Name	A# Number
Kasargod Kailash Chandra Shenoy	A20526053