PROJECT - REPORT

Beat Analytics : Spotify Data Analysis and Song Popularity Prediction

Data Preparation & Analysis (CSP-571)

Project Group

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

Table of Contents

1. Abstract

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.

2. Overview

The "Beat Analytics" project aims to leverage the extensive data provided by Spotify to explore the various factors that influence a song's success. By using data analysis techniques and machine learning techniques, this project seeks to pinpoint the important elements that affect song popularity. The ultimate goal is to provide musicians, record labels, and industry stakeholders with actionable insights that can enhance a song's performance and reception in today's music industry. This project highlights the key factors that could be optimized to boost a song's impact on platforms (digital) within the modern music landscape.

2.1. 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.

2.2. Relevant Literature

The "Beat Analytics: Spotify Data Analysis and Song Popularity Prediction" project's literature study examines previous research and studies on the subjects of machine learning applications in the music industry, song popularity prediction, and music data Analysis.

Where several key studies and resources have informed this project’s approach like The article "Are Hit Songs Becoming Less Musically Diverse?" [1] explores the evolution of musical diversity in hit songs, suggesting a potential narrowing in the variety of popular music. "Song Popularity Predictor" by Mohamed Nasreldin [2] discusses various methodologies for predicting song popularity using machine learning techniques. Scholarly articles such as "Predicting Music Popularity with Machine Learning" [3] and research on the evolution of Western popular music [4] provide foundational insights into the application of analytical techniques in understanding music trends. Resources like "R for Data Science" and various Spotify engineering blogs offer methodological and technical guidance, supporting the project's use of the R programming language for data analysis and modeling.

2.3. Proposed Methodology

1. Data Collection: We are utilizing two primary sources for our data:

- Top Spotify Songs from 2010-2019 by Year: This dataset includes about 600 top songs for each year, encapsulating various features like artist, genre, and music properties.

- Ultimate Spotify Tracks Database: Provides comprehensive information on Spotify tracks, including features such as acousticness, danceability, and overall popularity.

2. Data Preprocessing: Essential steps to ensure the data is ready for analysis:

- Handling Missing Data: We checked for missing data or anomalies in the datasets and decided how to address them—either by imputation or removal and summarized them, which is crucial for clean modeling
- Removing Duplicates: Removing duplicated entries to ensure the dataset's uniqueness.
- Data Conversion and Cleaning: Modifying data types for better analysis, such as converting ‘year’ to a factor and Removing unnecessary columns.

3. Feature Engineering: Wel derived features to better understand the underlying music patterns:

- Aggregation by Artist or Genre: Summarizing data at the artist or genre level could reveal trends over time or across categories.

- Extracting and Transforming Features:
  - Year Extraction and Factor Transformation: Convert the year column from numeric to a factor, which could help in treating each year as a categorical variable, potentially capturing different trends across years.

- Aggregating Data:
  - Artist Appearances: Summarizing and Arranging the number of appearances each artist has in the dataset, help understand influence by artist frequency.

4. Exploratory Data Analysis (EDA): An in-depth EDA helps to uncover relationships and trends:

- Visualizing Feature Distributions: Using histograms, box plots, and scatter plots to understand the distribution of each feature.
- Correlation Analysis: Investigating how various features like energy and danceability correlate with song popularity.
- Impact Analysis: Assessing how different features influence the popularity of the songs.

5. Model Development: Developing a predictive model for song popularity:

- Data Splitting: Dividing the data into training and testing sets to ensure model validity.
- Algorithm Selection: Choosing suitable algorithms like logistic regression for baseline modeling, followed by more complex models like decision trees, naive bayes and KNNs.
- Model Training: Using the training data to fit the models and train them and then test them for their prediction accuracy.

6. Model Evaluation: Assessing the model using various metrics:

- R-squared: To measure the amount of variance in song popularity explained by the model.
- Mean Absolute Error (MAE): To evaluate the average error magnitude.
- Root Mean Squared Error (RMSE): To measure the average magnitude of error squared, providing a sense of error volume.
- Using Confusion Matrix.

*With KNN, precision, recall, F1-score, and AUC are more informative about model performance.

7. Insights and Recommendations:

- Understanding Song Popularity: Using the model to understand what drives song popularity.

- Recommendations for Stakeholders: Probable to offer data-driven insights to artists, producers, and labels on how to enhance song popularity.

3. Data Processing

3.1. Data Set

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.

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

3.2. Data 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)).
    - 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.
    - 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.

3.3. 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.
  - 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.
  - 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.

3.4. Data Issues & Assumptions/ Adjustments

- **Data Issues**:
  - Quality Concerns: Potential issues with data completeness, consistency, and accuracy, such as incorrect tagging of genres or mismatches in artist names.
  - Duplication: Redundant entries for the same songs could skew the analysis, necessitating de-duplication.

- **Assumptions/Adjustments**:
  - Popularity Metric Reliability: Assuming that Spotify's popularity index accurately reflects listener preferences and trends.
  - Stable Data Sources: Assuming the data sources remain stable and consistent over time, without significant shifts in data collection methods.

4. Data Analysis

4.1. Summary Statistics

Top Spotify Songs from 2010-2019 by Year (top10s.csv):

Total Records and Features: The dataset includes approximately 600 songs with 13 features for exploration such as BPM (beats per minute), energy, danceability, loudness (dB), and popularity.

Correlations: Using correlation plots, we explored the relationships between features such as energy, danceability, and popularity to identify the most influential factors on song popularity.

Artist and Genre Analysis: The dataset allows analysis of top artists and genres over the years, including counts of appearances and distinct years present in the top charts.

Ultimate Spotify Tracks Database (SpotifyFeatures.csv):

Total Records and Features: This dataset is more extensive, with over 232,725 tracks spanning multiple genres and 18 features including acousticness, danceability, duration, and popularity.

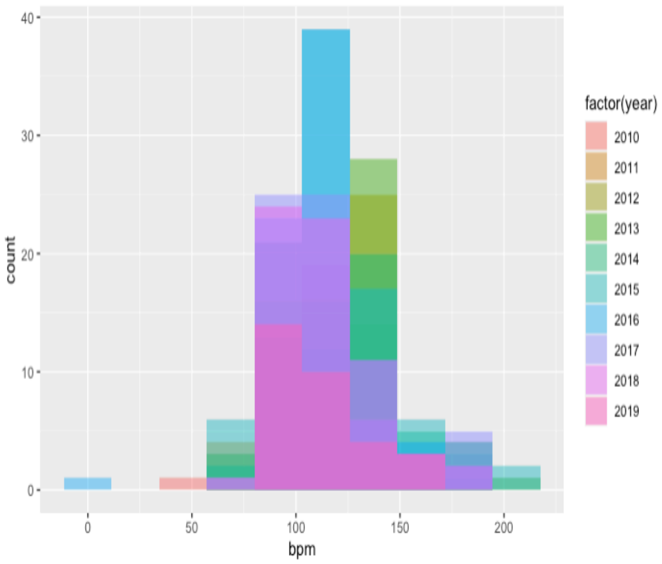
Popularity: Similar to the first dataset, popularity is analyzed, but with a broader scope given the dataset's size and diversity. Histograms and density plots are used to visualize the distribution of popularity scores across all tracks.

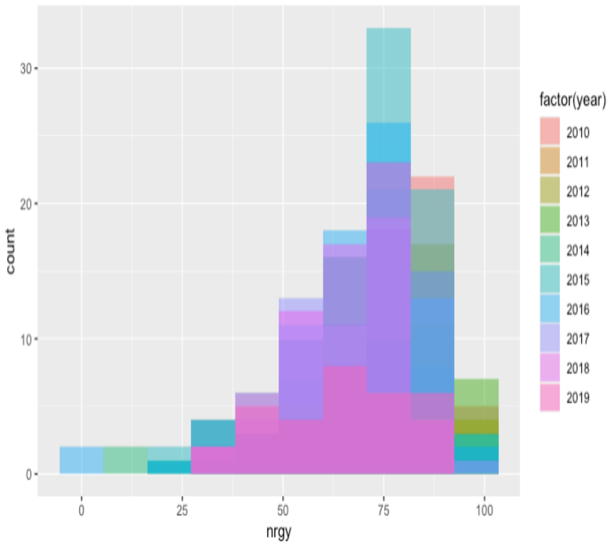
Key and Mode: The distribution of musical keys and modes (major vs. minor) provides insights into their correlation with popularity, which we visualized using bar plots and density plots.

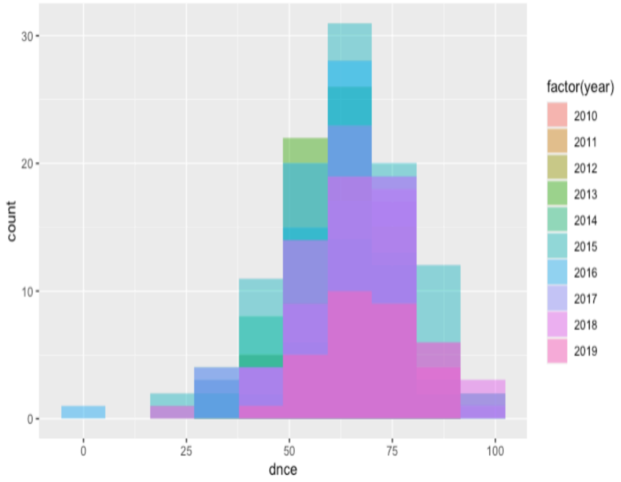
Correlations: We made a detailed correlation matrix to explore the relationships between numerical features, highlighting how different musical characteristics interact with each other and influence song popularity.

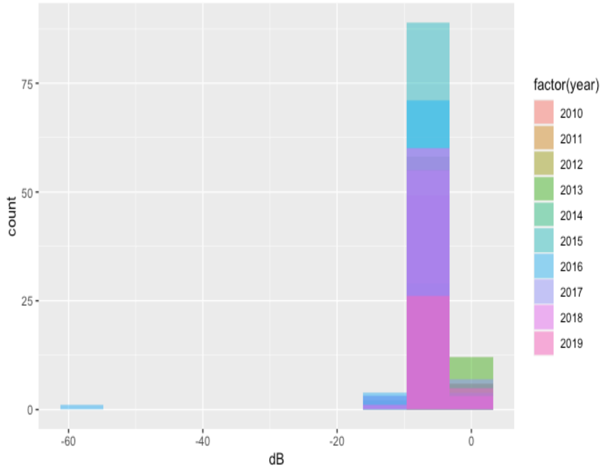
4.2. Visualization

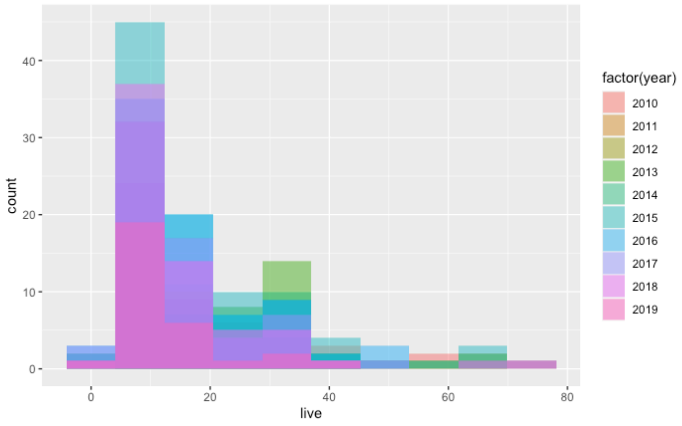
In this section, we delve into the analysis of song and artists characteristics through a series of histograms, each providing a visual summary of the distribution of a specific musical attribute across a decade's worth of data. These features include Beats Per Minute (BPM), energy (nrgy), danceability (dnce), and loudness (dB), key elements that contribute to the auditory and emotional qualities of music.

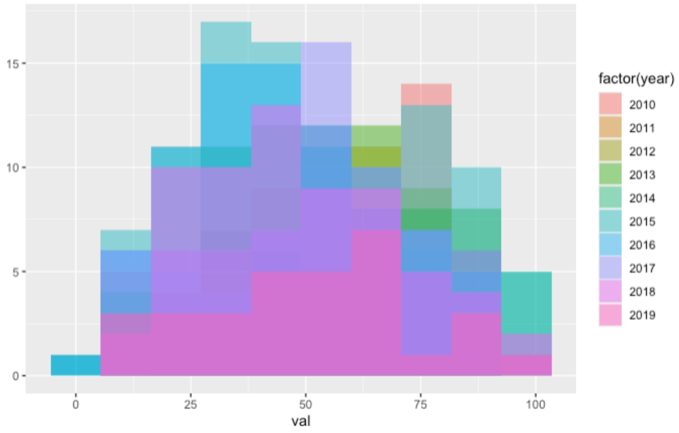


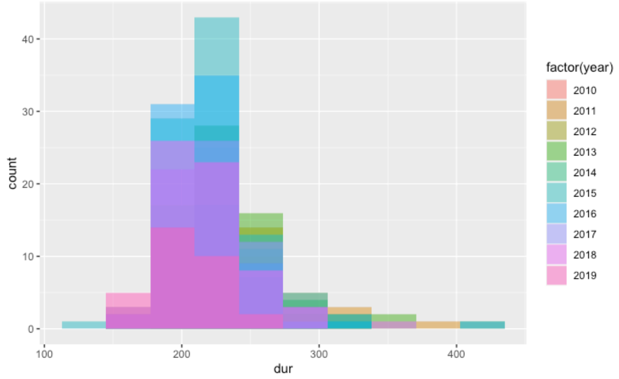


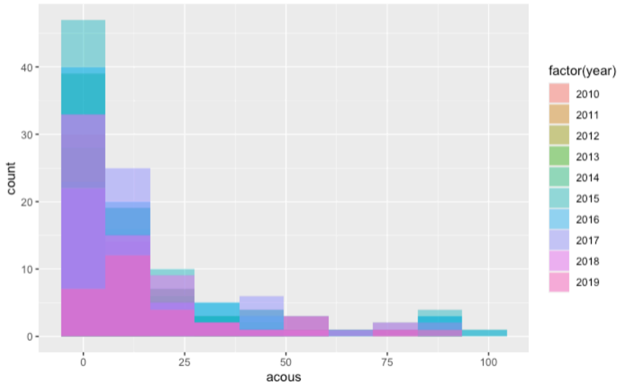


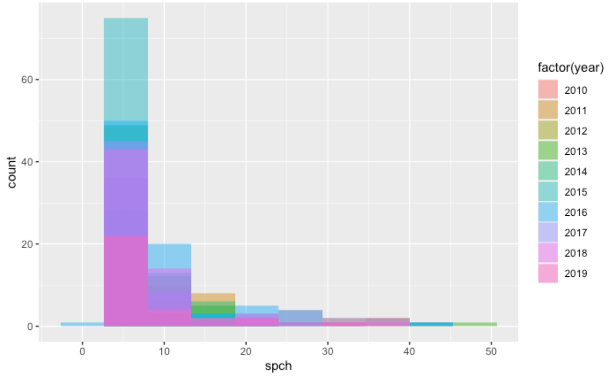


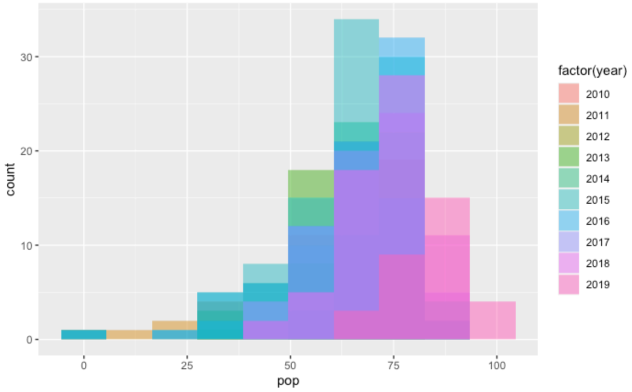

count
40
30
20
10
0
0
20
40
60
80
live
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


count
15
10
5
0
0
25
50
75
100
val
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

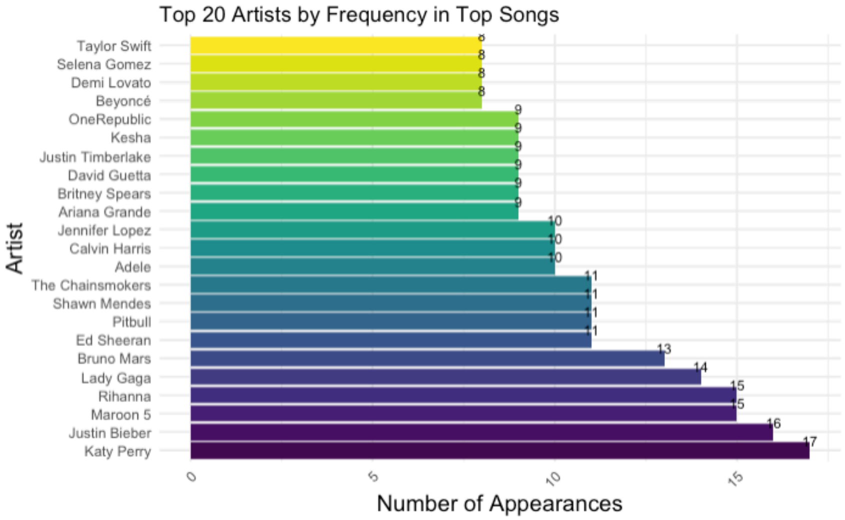

count
40
30
20
10
0
100
200
300
400
dur
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


count
40
30
20
10
0
0
25
50
75
100
acous
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

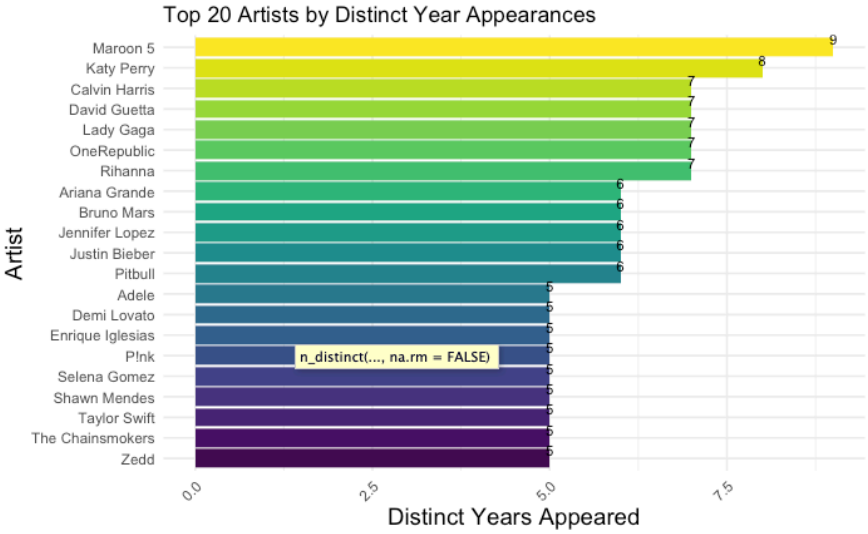

count
60
40
20
0
0
10
20
30
40
50
spch
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019


count
30
20
10
0
0
25
50
75
100
pop
factor(year)
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

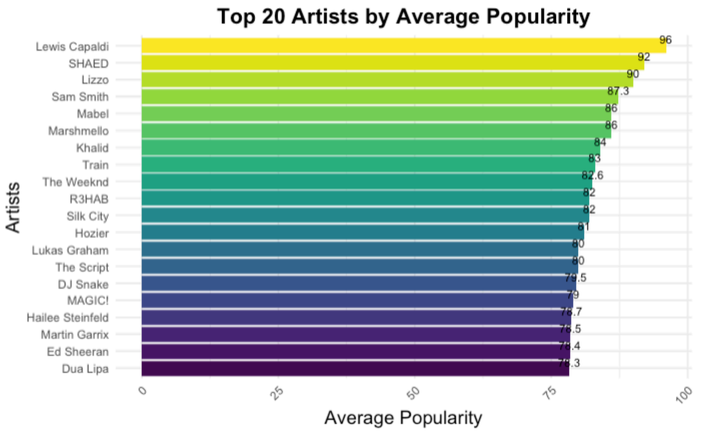
The series of histograms above visually represents the distribution of various song attributes, including beats per minute (BPM), energy, danceability, loudness, liveness, valence, duration, acousticness, speechiness, and popularity, across different years, highlighting trends and shifts in musical characteristics over time.



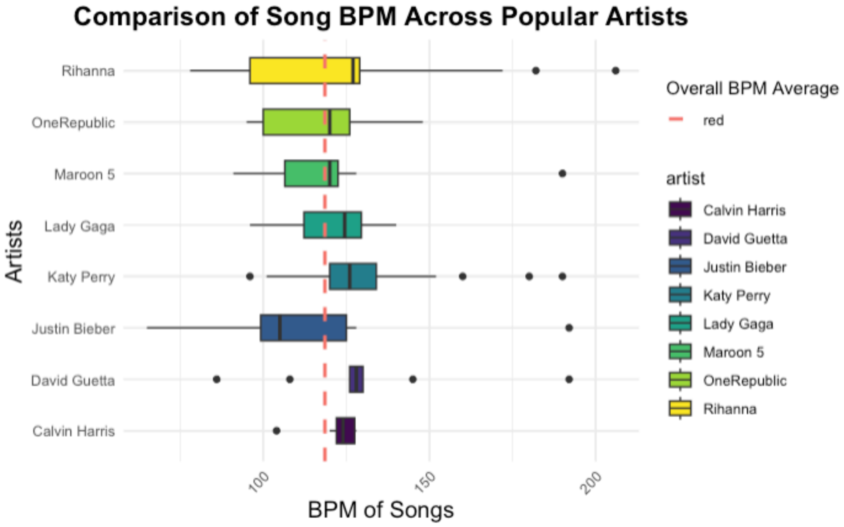
This bar chart ranks the top 20 artists by their frequency of appearances in the dataset, with each bar representing the total number of times an artist appears, visually highlighting the most prevalent artists in terms of song counts over the examined period.



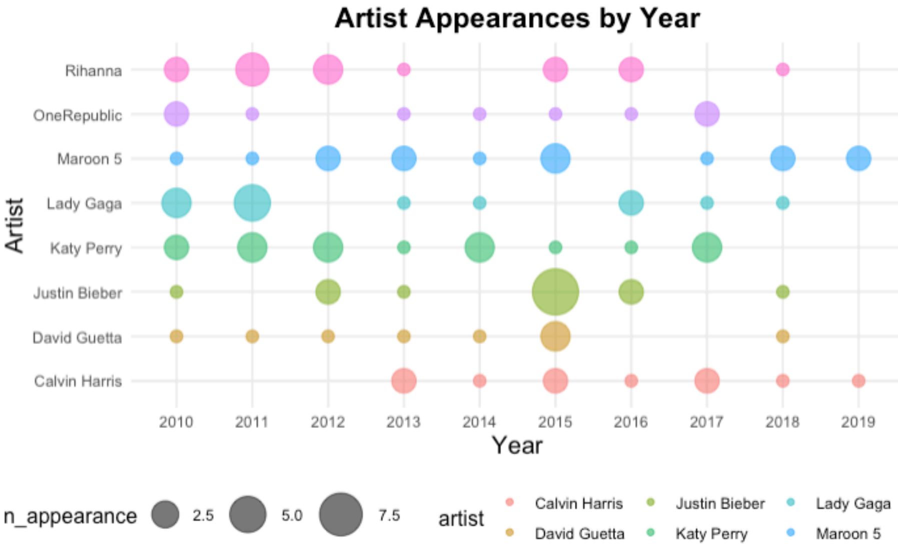
This bar chart displays the top 20 artists ranked by the diversity of years in which they have appeared, showing each artist's presence across different years to highlight their enduring popularity or resurgence in the music industry. We can see that “Maroon 5” is at the top with 9 distinct year appearances.



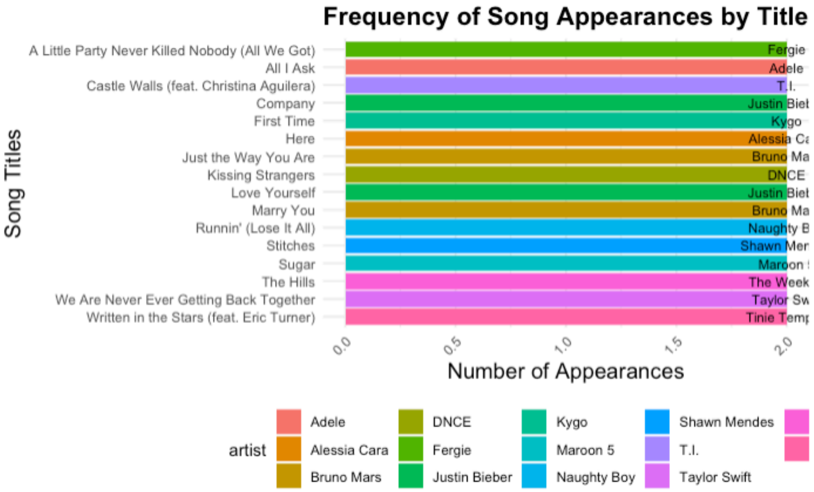
This bar chart showcases the top 20 artists sorted by their average popularity, quantifying each artist's typical appeal and highlighting those who consistently resonate the most with listeners over the surveyed period.



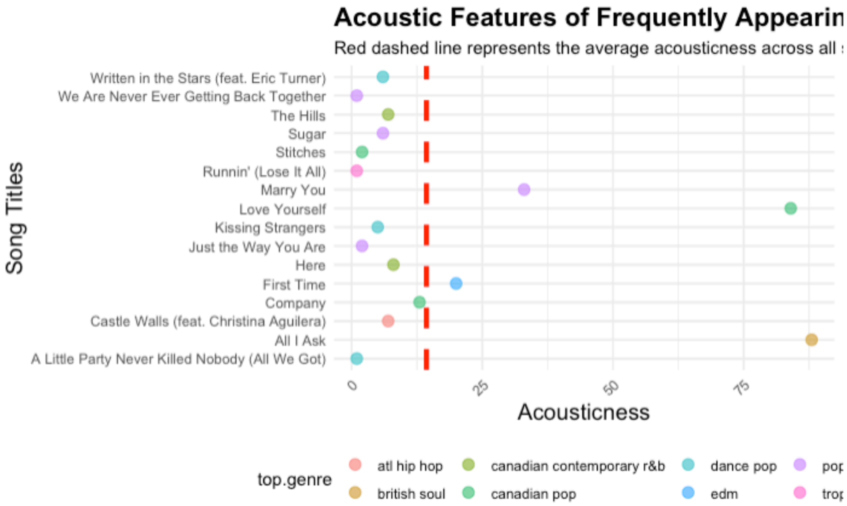
This boxplot visualizes the range and distribution of BPM for songs by selected top artists, highlighting variability and typical tempo characteristics within each artist's music catalog, with a dashed red line indicating the average BPM across all songs in the dataset.



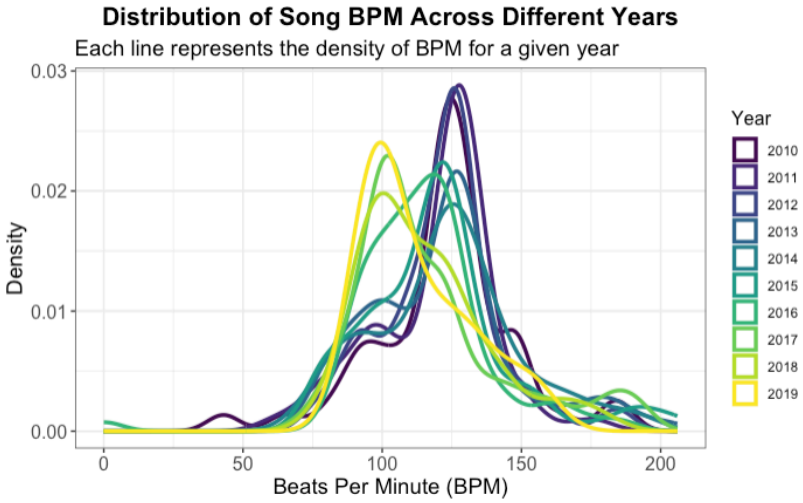
This scatter plot displays the number of appearances for each selected artist by year, using point size to represent the frequency of appearances, thereby illustrating trends and the prominence of each artist in the dataset over time.

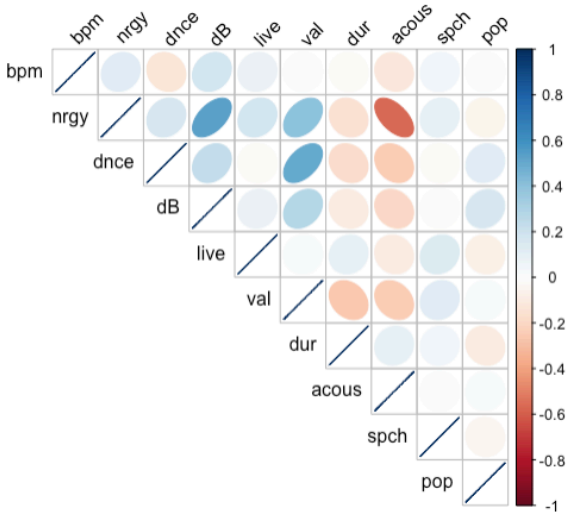


This bar chart displays the frequency of song appearances sorted by title, with each bar representing the number of times a particular song appeared in the dataset, colored by artist, thus highlighting which songs and artists have been recurrently popular over the observed period.

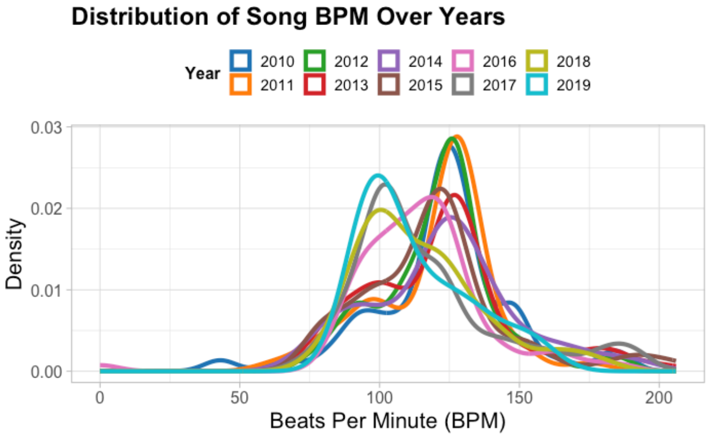


This scatter plot visualizes the acousticness of frequently appearing songs, highlighting individual song characteristics against the average acousticness (indicated by a red dashed line), offering insights into how these songs compare to the overall musical texture observed across the dataset.

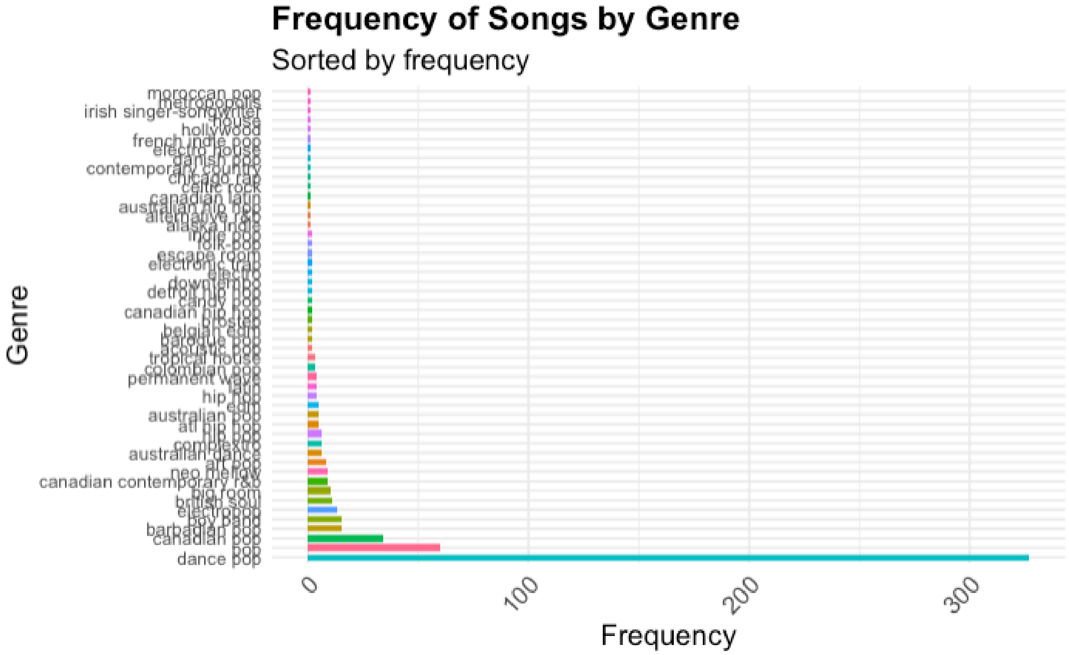




This correlation plot uses elliptical visual representations to illustrate the strength and direction of correlations between various musical attributes in the dataset. For example, we can see that “nrgy” and “dB” have higher correlations.

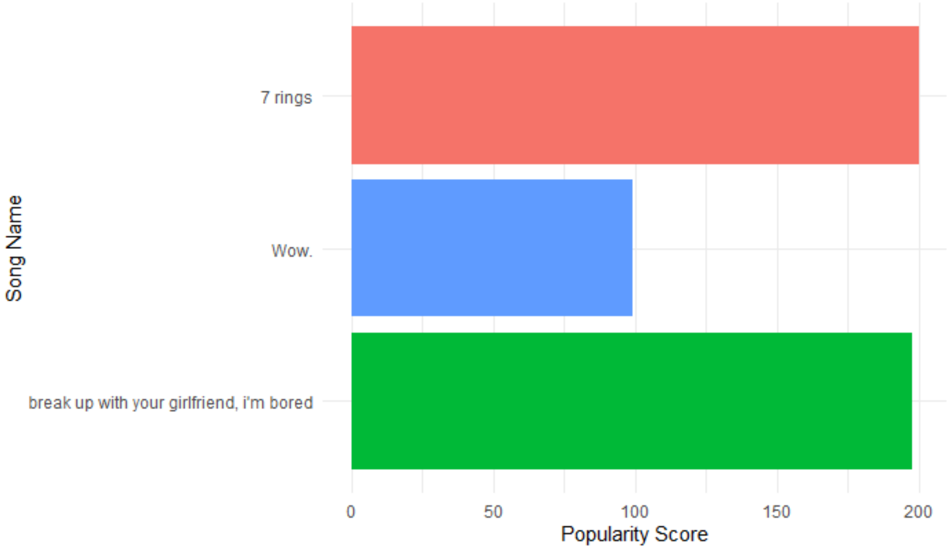


This density plot showcases the distribution of Beats Per Minute (BPM) for songs from different years.. The plot helps identify temporal trends in song BPM, providing insights into how musical tempo preferences have evolved over the decade.

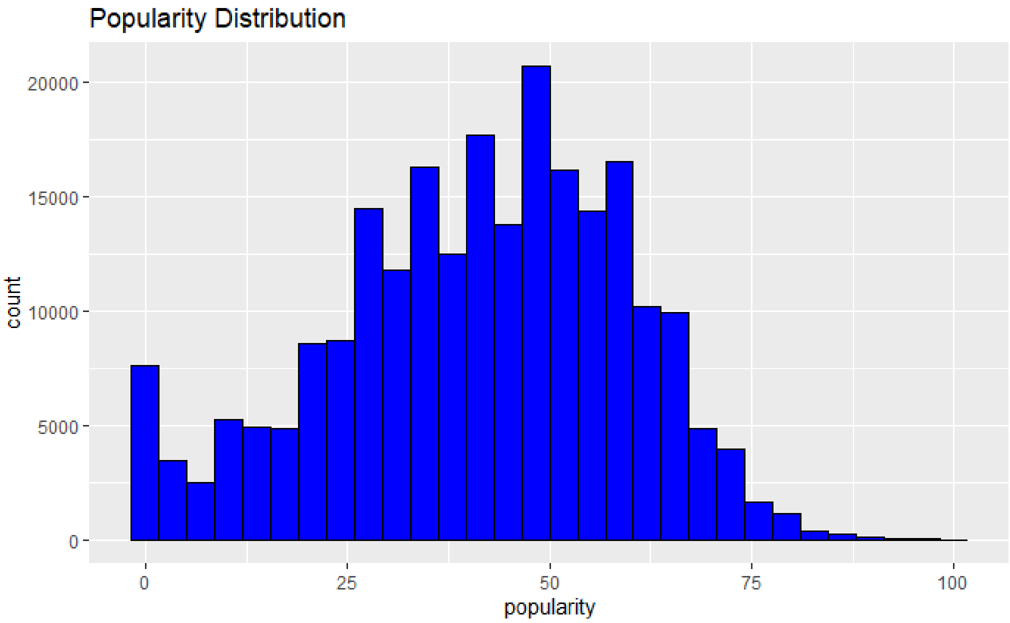


This bar chart quantifies the frequency of songs across various genres, displaying each genre's prevalence in the dataset. By arranging genres in descending order of frequency, the chart highlights the most popular music styles and provides a clear visual comparison of genre popularity, with enhanced readability through a flipped coordinate system for easy genre identification. Where we can see that “dance pop” is the most frequent genre of songs made.

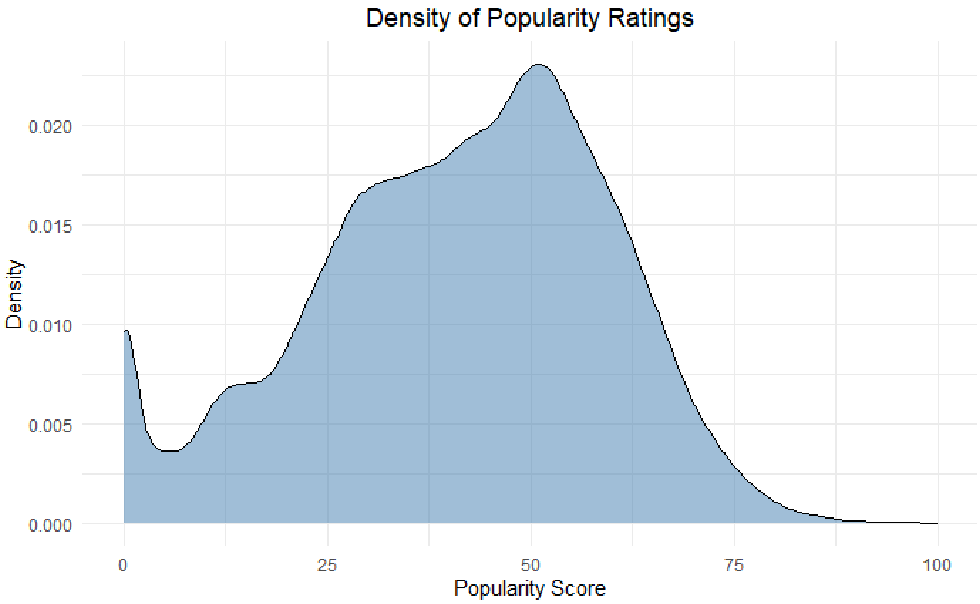
Top Songs



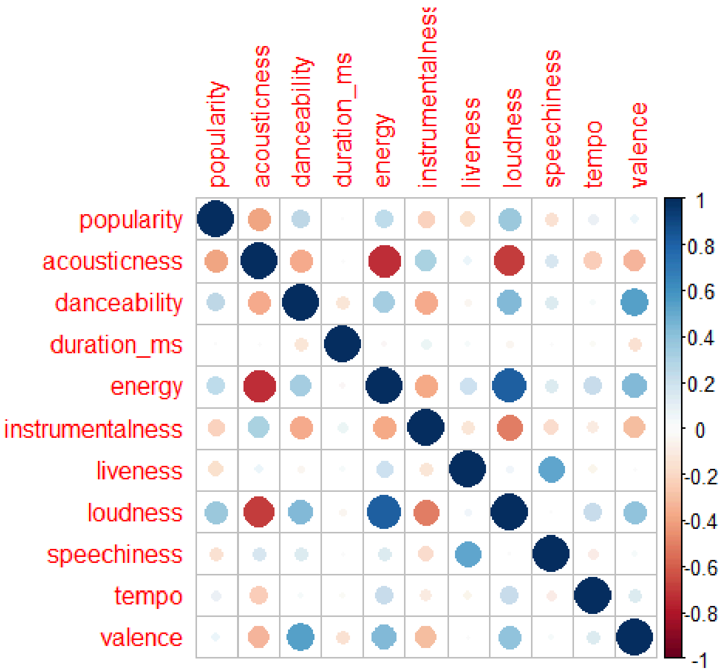
This bar chart shows the top songs by popularity score, with the song "7 rings" having the highest score, followed by "Wow." and "break up with your girlfriend, i'm bored".



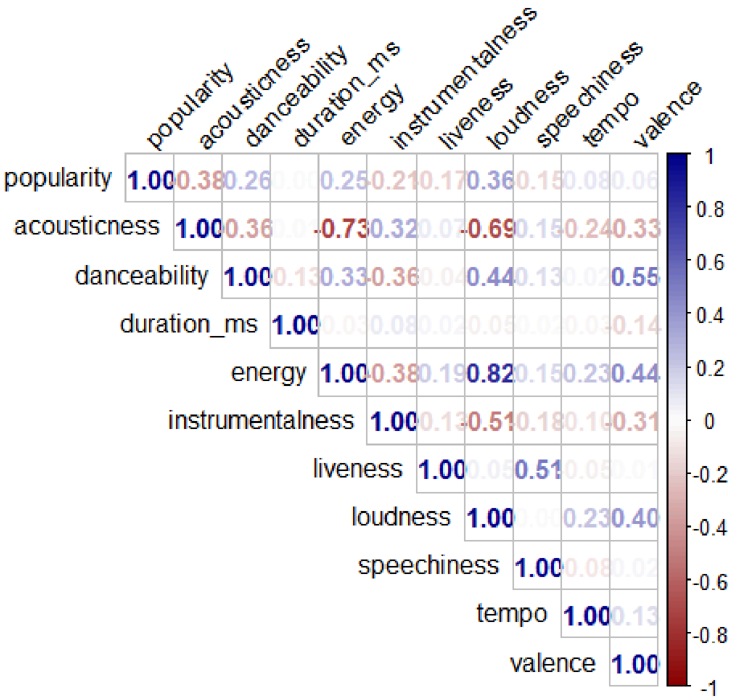
A histogram that shows the distribution of popularity scores for a set of songs, indicating that most songs have a popularity score around the 40 to 60 range.



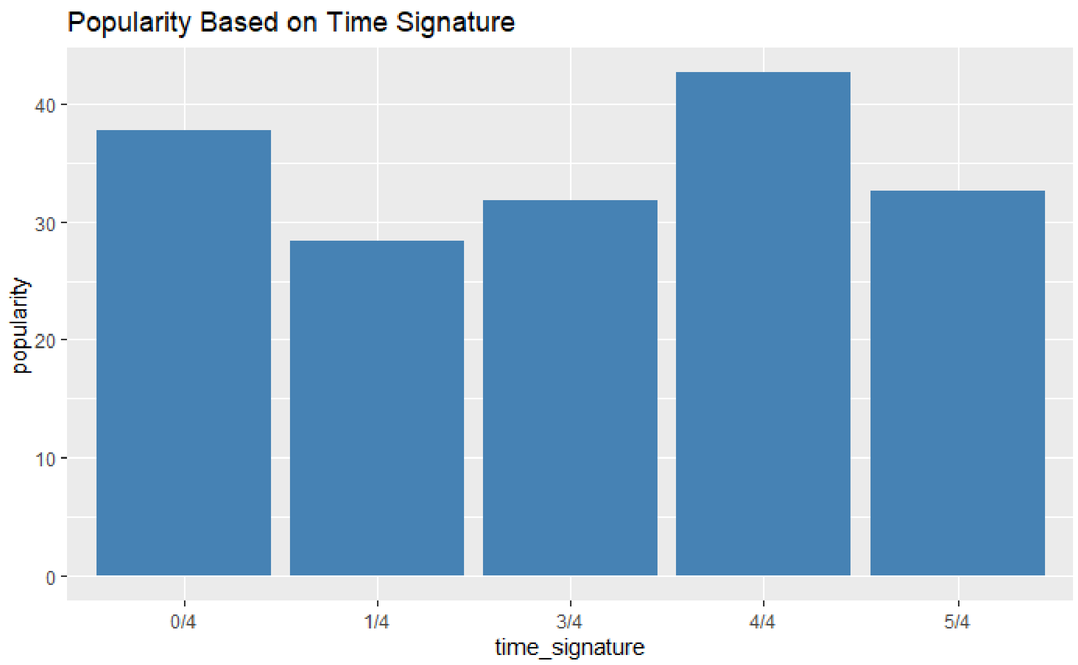
This density plot represents the distribution of popularity scores, with a clear peak around 50, suggesting that is the most common popularity score among the songs analyzed.



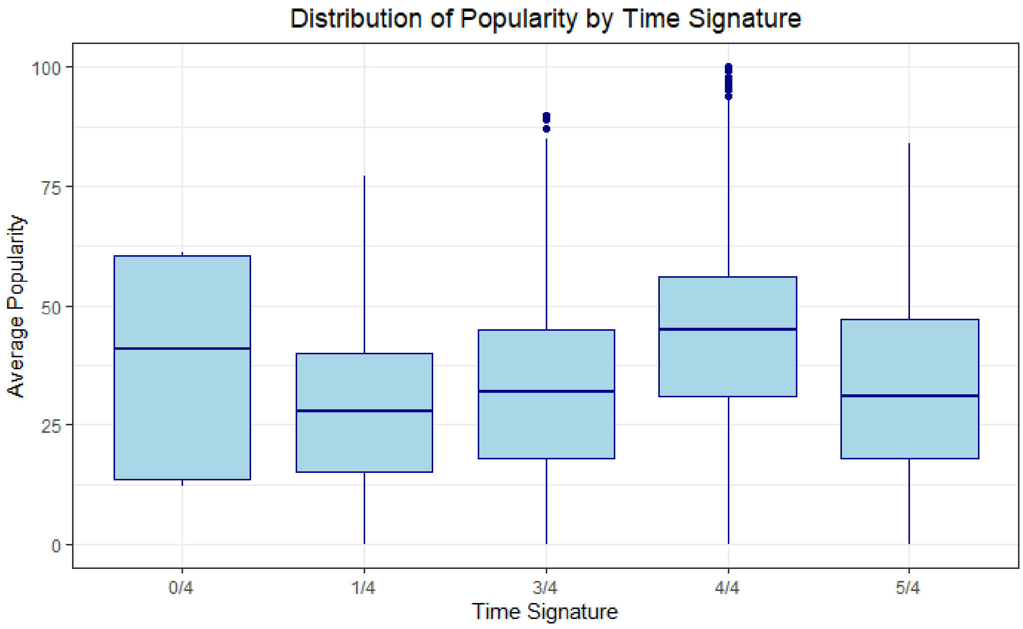
A correlation matrix plot presented as a grid of circles, where the size and color of the circles represent the strength and direction of correlation between different song characteristics such as popularity, danceability, energy, etc.



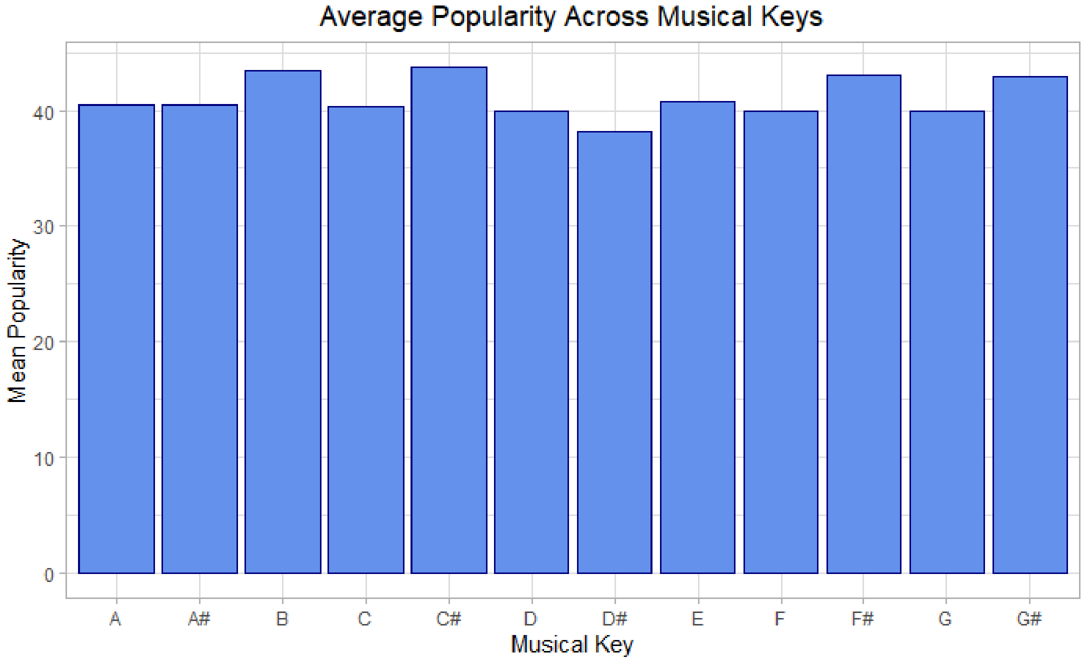
A correlation matrix with numerical values, highlighting the strength of linear relationship between song attributes, with 1 indicating a perfect positive correlation and -1 indicating a perfect negative correlation.



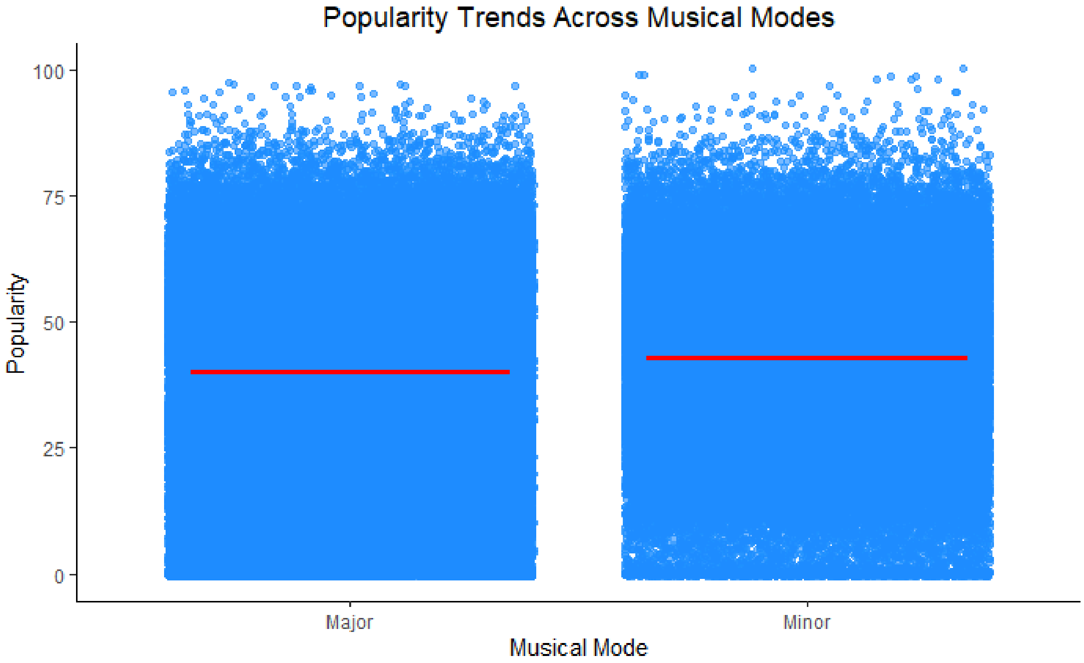
A bar chart that illustrates the popularity of songs based on their time signature, with songs in 4/4 time being the most popular, followed closely by 5/4 and then 0/4.



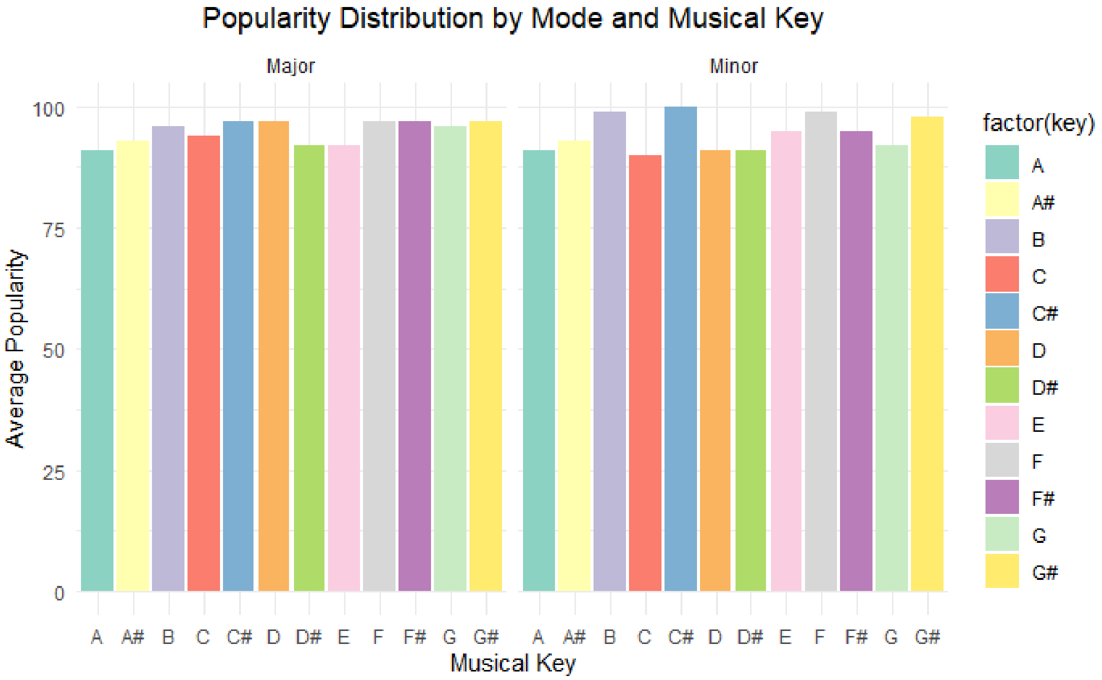
Box plots representing the distribution of song popularity for different time signatures, showing a wide range of popularity for 4/4, and relatively less variation for other time signatures.



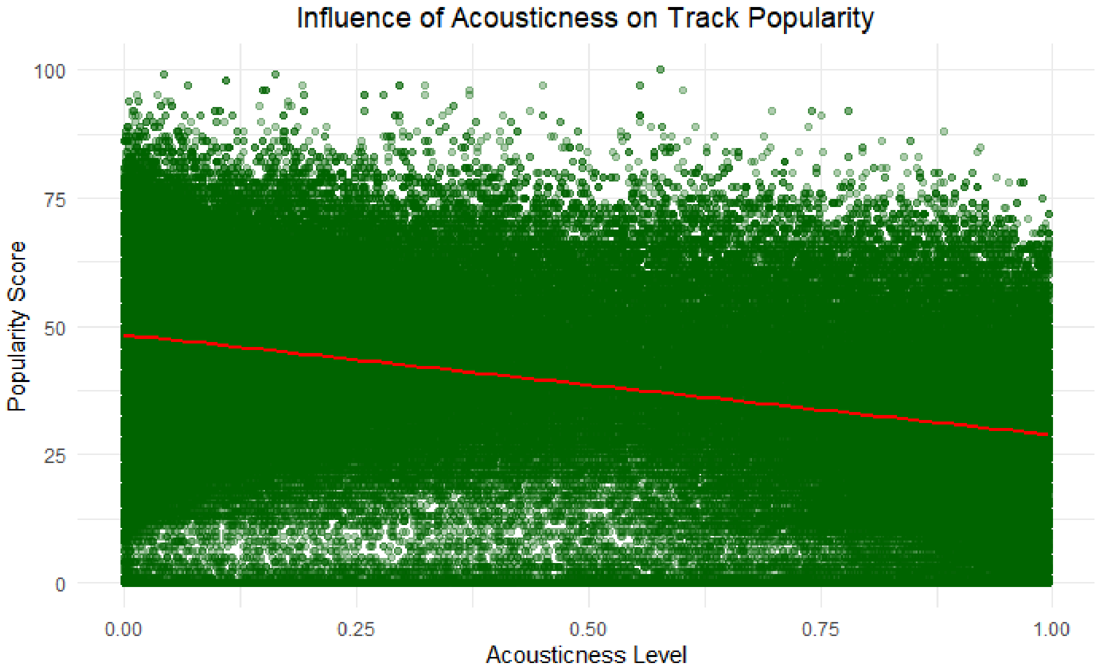
.A bar chart that shows the popularity of songs based on their musical key, with all keys appearing to have a similar level of popularity, though B, C#, F# and G# are slightly higher.



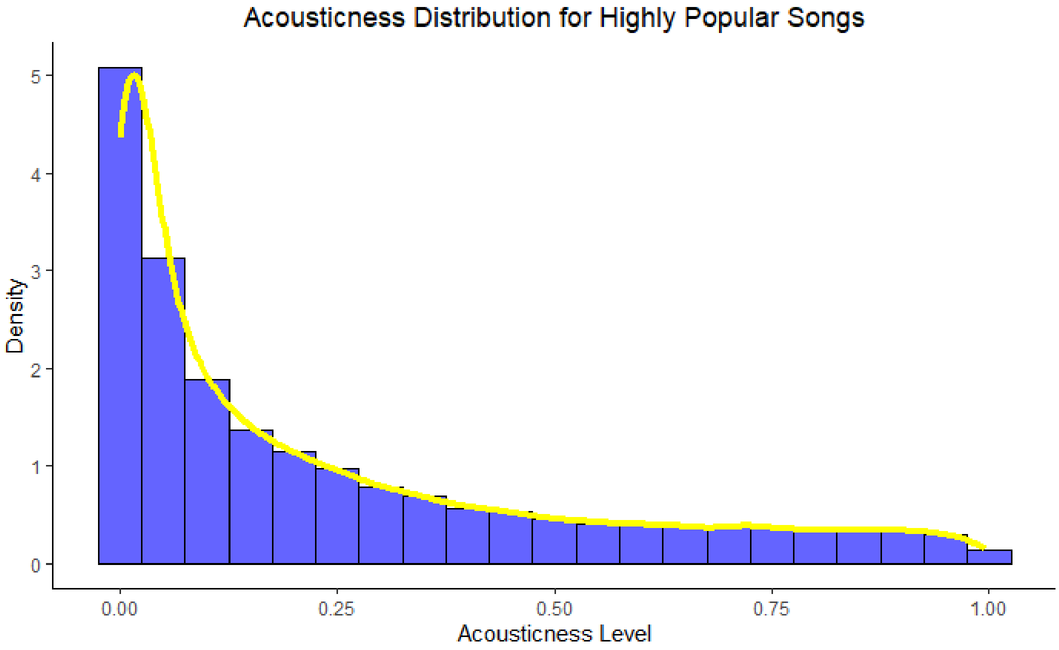
A scatter plot comparing the popularity of songs in major vs. minor modes, with a horizontal line indicating the median popularity level, showing a broad distribution in both modes.



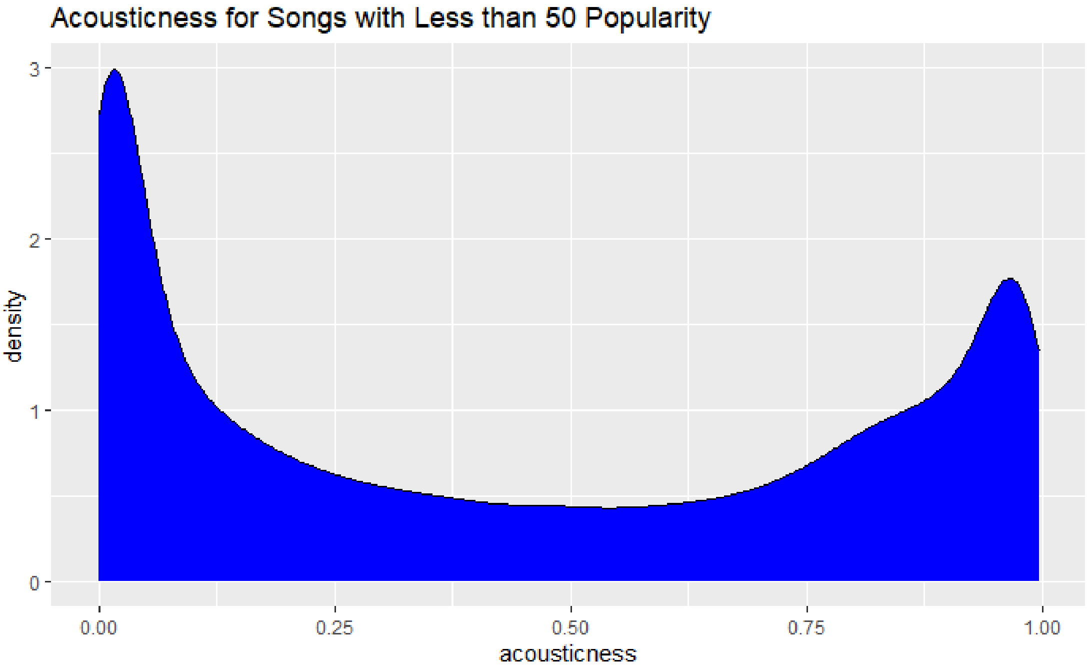
A stacked bar chart representing the cumulative popularity of songs categorized by musical mode (major or minor) and key, with different colors indicating the contribution of each key to the total popularity.



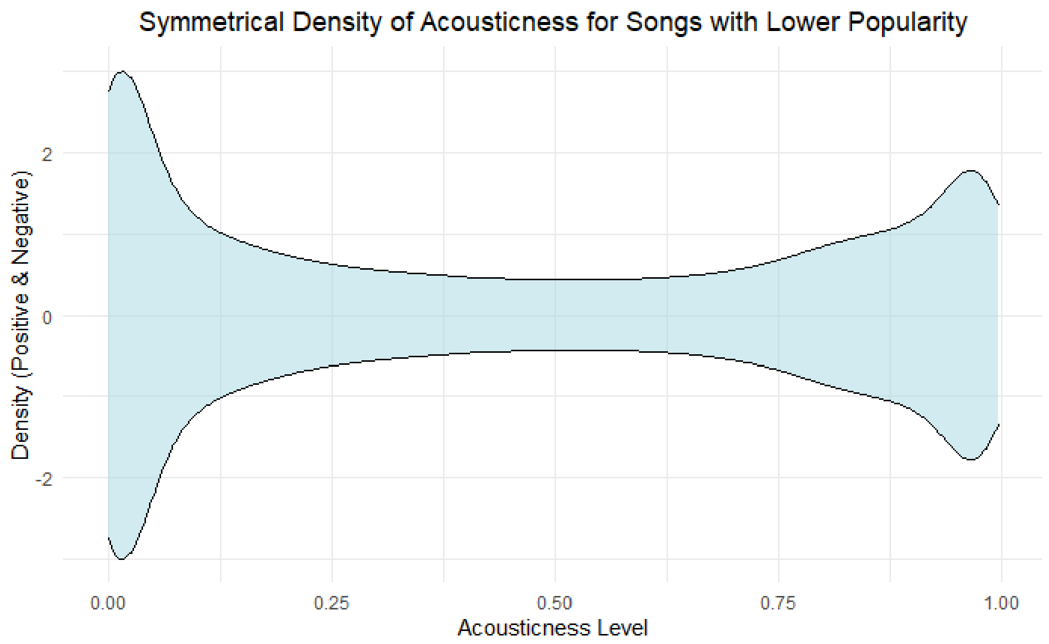
A scatter plot with a trend line that examines the influence of acousticness on track popularity, suggesting a slight negative trend as acousticness increases.



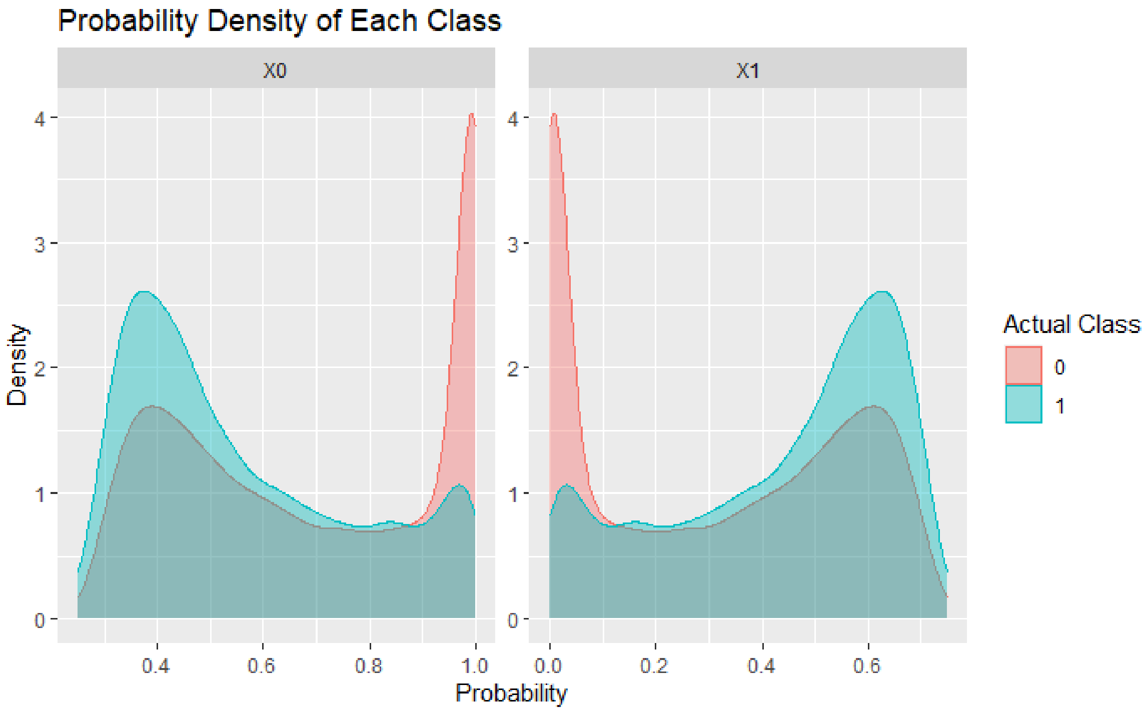
A histogram overlaid with a density plot showing the distribution of acousticness levels for highly popular songs, indicating a higher frequency of songs with low acousticness.



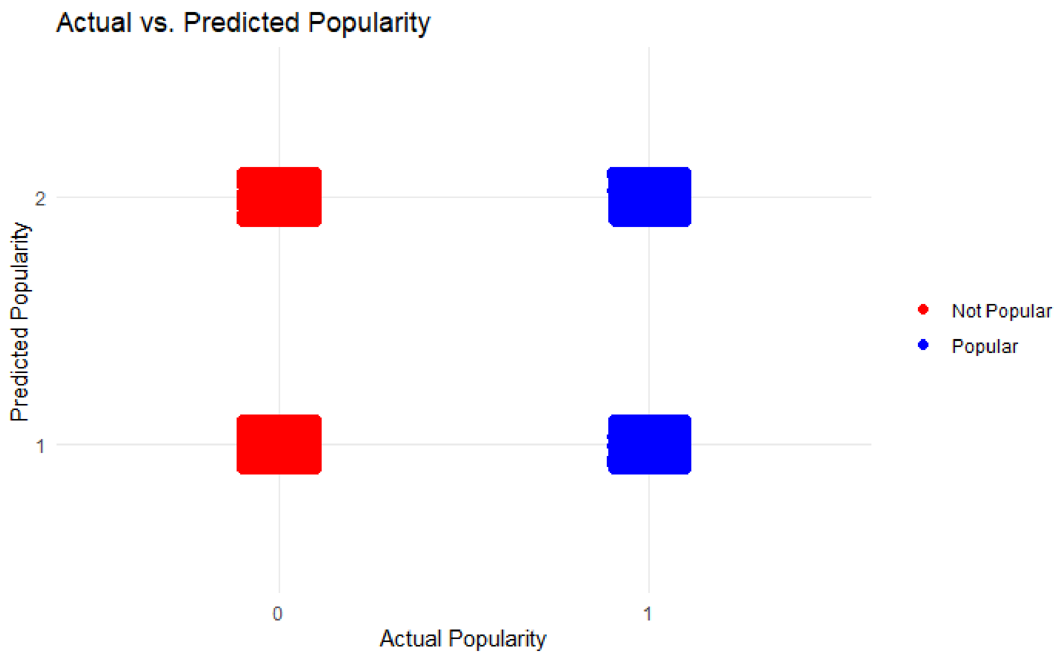
A density plot representing the acousticness levels of songs with popularity scores less than 50, showing peaks at both the lower and higher ends of the acousticness scale.



A mirrored density plot illustrates the distribution of acousticness for songs with lower popularity, highlighting a symmetric pattern with two peaks at the extremes.



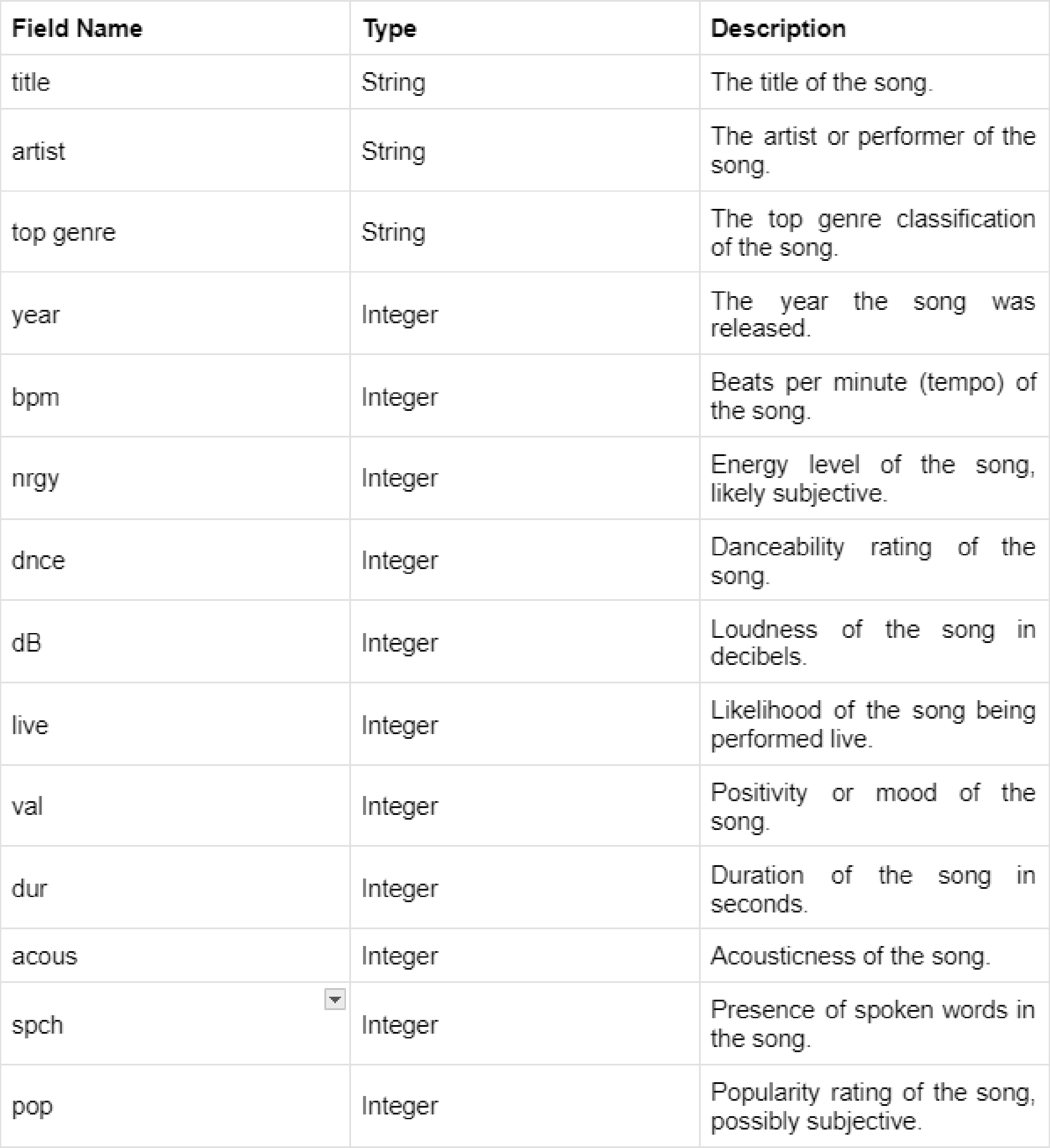
A pair of density plots showing the probability distribution for two classes, likely representing the predicted probabilities for a binary classification, with clear peaks at probability values near 0 and 1.



A scatter plot with color-coded points representing the comparison between actual and predicted popularity, divided into four quadrants to show the distribution of predictions against actual values.

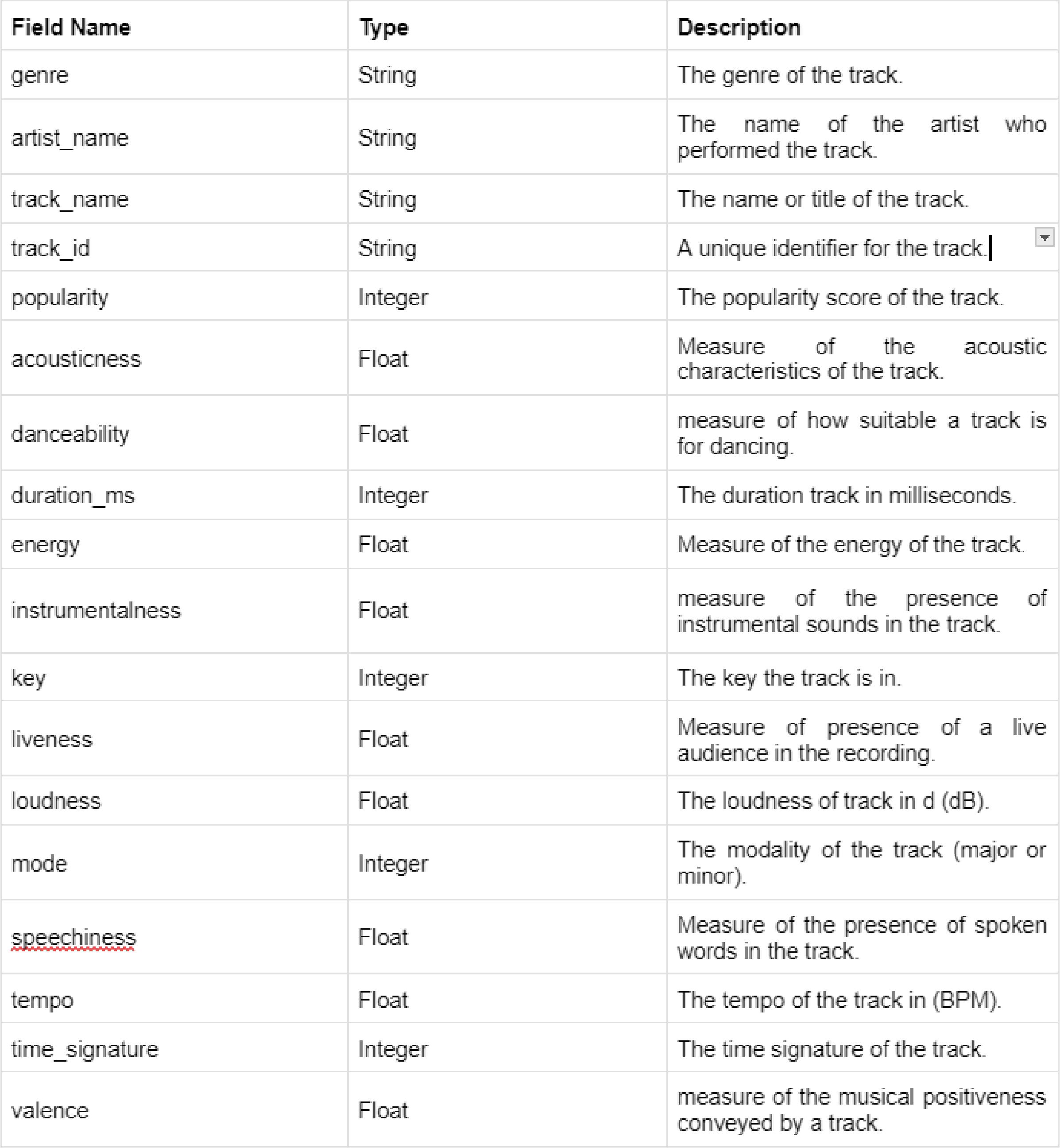
4.3. Feature Extraction

Extracted Features from Top Spotify Songs from 2010-2019 by Year (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.

Extracted Features from Ultimate Spotify Tracks Database (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.

5. Model Training

5.1. Feature Engineering

- **Modification of Year Data:** The year attribute in ‘top_by_year’ data is converted to a categorical variable using as.factor. This is crucial for models that handle categorical data differently from numerical data and can help in identifying specific year-based trends in song popularity.

- **Temporal Features**: Extracting and utilizing the 'year' from release dates to analyze trends over time.

- **Creating Artist Appearance Features**: New variables are created to count the number of appearances of each artist and the distinct years they appear in the dataset. This can be used to measure the consistency and longevity of an artist's popularity.

- **Engineering Popularity Features**: For each artist, the average popularity score is computed. This engineered feature provides a condensed view of an artist’s overall popularity across the dataset.

- **Handling Key and Mode**: Converting the key and mode from categorical/string types to numeric/binary formats. This is vital for including these features in predictive modeling since many machine learning algorithms require numerical input.

- **Binarization of Popularity**: Transforming the continuous popularity score into a binary format (popular/not popular) based on a threshold, which is particularly useful for classification models.

5.2. 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).

5.3. 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).

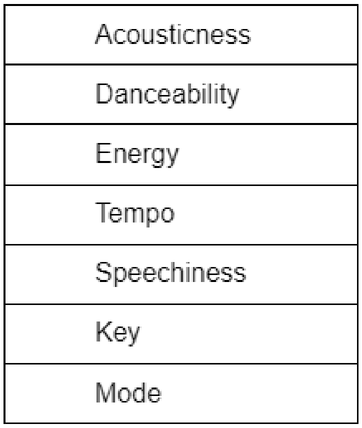
- **Logistic Regression**: For predicting whether a song is popular or not based on a binarized popularity score.

- **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

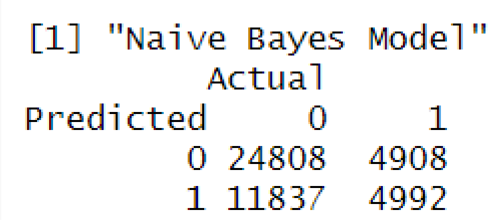
6. Model Validation

6.1. Testing Results

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

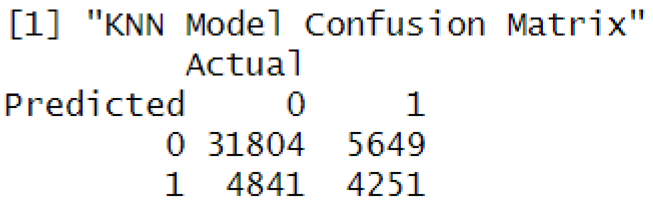
Model Type	Accuracy	MSE	RMSE	F1 Score
Decision Tree	78.73%	0.1675	0.4092	88.10%

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
```

6.2. 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.

6.3. Biases/ Risks

- **Sampling Bias**

  Given that our data comes from two specific datasets:

  - ➔ **Top Spotify Songs from 2010-2019**: This dataset may limit the generalization of our findings to only top songs per year and may not be representative of all types of songs on Spotify.
  - ➔ **Ultimate Spotify Tracks Database**: While its comprehensive, the selection and inclusion criteria for these tracks aren't specified in depth, which could lead to the dataset that is not entirely representative of the entire Spotify music catalog.

These biases could make our model to learn patterns which may not be applicable to the broader updated Spotify track database or newer music trends post-2019.

7. 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.

7.1 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.

8. Data Sources

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

9. Source Code

Our source code along with the datasets & documentation can be found at this GitHub link - https://github.com/AkshatBehera/CSP571-DPA-Project-BeatAnalytics-Spotify

10. 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.