# Analysis of Spotify Music Popularity

*Spotify: An In-Depth Review*

# Content Page

# Data understanding and preparation

## Brief overview of data and business understanding

Spotify is a popular digital music streaming service that provides millions of users with access to a vast library of music and other audio content. Founded in 2006, it has since become one of the leading platforms for music streaming worldwide.

We have chosen the Spotify music dataset of 1 million songs, available from Kaggle (link). This dataset is a valuable resource for understanding music preferences, exploring trends in audio features, and gaining insights into what makes certain tracks more popular than others. Please refer to Appendix A for a detailed description of each variable in this dataset. However, due to the computational ability of our machines, running the full dataset is not feasible. Therefore, we have decided to limit our analysis to a 10% subset of data from 2022 of 5.3K data to showcase the most recent trends.

The primary objective of our 2022 analysis is to gain insights into the characteristics and trends of recent music tracks, discover patterns related to track popularity, and understand how different attributes influence a track's reception among users.

Spotify's platform allows users to discover, play, and share music. Furthermore, it is essential for the company to understand the platform users' preferences and track popularity to provide personalized recommendation algorithms, optimize their music library, and monetize the app more effectively. By conducting a thorough analysis of this dataset, we aim to unravel the underlying patterns and discover interesting relationships that can help enhance user experiences and refine the platform's music recommendation algorithms.

Through data mining techniques, we seek to identify the attributes and combinations of musical features that resonate most with users, leading to better-targeted recommendations. Improved recommendations contribute to increased user engagement and retention, fostering a stronger user-community relationship.

A detailed analysis of the dataset can assist Spotify in optimizing its music library curation and playlist creation. By understanding the most appealing musical characteristics, genres, and artist collaborations, Spotify can curate a diverse and engaging music collection that caters to the tastes of its vast user base.

Insights gained from the data mining process can empower Spotify to monetize the app more effectively. By understanding user preferences and track popularity, Spotify can strategically offer premium features, targeted advertisements, and curated content that align with users' tastes, thereby enhancing user satisfaction and generating more revenue opportunities.

By leveraging the wealth of information available in the Spotify 1 Million Tracks Dataset, this data mining project aims to contribute valuable insights to Spotify's continuous improvement, providing a win-win scenario for both the platform and its users.

## Descriptive Stats on Data


*Figure 1: Variables Statistics Table*

The Spotify 1 Million Tracks Dataset is a comprehensive collection of music tracks available on the Spotify music streaming platform. Based on the above figure 1, it consists of 14 predictors/features and 2 target variables after eliminating insignificant features. The dataset contains no missing values, ensuring that we have complete data for all the tracks.

## Overview of the Format and Types for Variables

1. Popularity (Target Variable)
   Format: Continuous numeric
   Type: Integer
   Description: Popularity has an integer data type with a minimum value of 0 and maximum value of 93, with an average of 31.237 and a deviation of 17.709.

2. Verdict (Target Variable)
   Format: Categorical
   Type: Text
   Description: The Verdict variable represents a track's popularity on Spotify, categorized as either 'Low' or 'High' based on verdict scores.

3. Danceability (Predictor/Feature)
   Format: Continuous numeric
   Type: Decimal
   Description: Danceability has a real data type with a minimum value of 0.054 and maximum value of 0.980, with an average of 0.548 and a deviation of 0.182.

4. Energy (Predictor/Feature)
   Format: Continuous numeric
   Type: Decimal
   Description: Energy has a real data type with a minimum value of 0 and a maximum value of 1, with an average of 0.649 and a deviation of 0.271.

5. Key (Predictor/Feature)
   Format: Categorical
   Type: Integer (values ranging from -1 to -11)
   Description: Key has an integer data type with a minimum value of 0 and a maximum value of 11, with an average of 5.303 and a deviation of 3.573.

6. Loudness (Predictor/Feature)
   Format: Continuous numeric
   Type: Decimal (in decibels, ranging from -60 to 0 dB)
   Description: Loudness has a real data type with a minimum value of -43.772 and a maximum value of 1.906, with an average of -8.862 and deviation of 6.163.

7. Mode (Predictor/Feature)
   Format: Categorical
   Type: Integer (values 0 or 1)
   Description: Mode has an integer data type with a minimum value of 0 and maximum value of 1, with an average of 0.615 and deviation of 0.487.

8. Speechiness (Predictor/Feature)
   Format: Continuous numeric
   Type: Decimal
   Description: Speechiness has a real data type with a minimum value of 0.023 and a maximum value of 0.960, with an average of 0.096 and deviation of 0.129.

9. Acousticness (Predictor/Feature)
   Format: Continuous numeric

Type: Decimal
Description: Acousticness has a real data type with a minimum value of 0 and maximum value of 0.996, with an average of 0.310 and deviation of 0.349.

10. Instrumentalness (Predictor/Feature)
    Format: Continuous numeric
    Type: Decimal
    Description: Instrumentalness has a real data type with a minimum value of 0 and a maximum value of 0.999, with an average of 0.246 and a deviation of 0.363.

11. Liveness (Predictor/Feature)
    Format: Continuous numeric
    Type: Decimal
    Description: Liveness has a real data type with a minimum value of 0.018 and a maximum value of 0.987, with an average of 0.223 and a deviation of 0.195.

12. Valence (Predictor/Feature)
    Format: Continuous numeric
    Type: Decimal
    Description: Valence has a real data type with a minimum value of 0 and a maximum value of 0.994, with an average of 0.428 and a deviation of 0.257.

13. Tempo (Predictor/Feature)
    Format: Continuous numeric
    Type: Decimal
    Description: Tempo has a real data type with a minimum value of 35.075 and maximum value of 216.317, with an average of 122.577 and deviation of 29.945.

14. Time_signature (Predictor/Feature)
    Format: Categorical
    Type: Integer (values ranging from 3 to 7)
    Description: Time signature has an integer data type with a minimum value of 0 and maximum value of 5, with an average of 3.882 and deviation of 0.473.

15. Duration_ms (Predictor/Feature)
    Format: Continuous numeric
    Type: Integer
    Description: Duration in ms has an integer data type with a minimum value of 24187 and maximum value of 2460440, with an average of 218681.264 and deviation of 108307.352.

16. Genre (Predictor/Feature)
    Format: Categorical
    Type: Text
    Description: Genre has a polynomial data type with values such as hardstyle, indian, alt-rock, club and more.

The dataset's distribution is quite well-distributed for the majority of the attributes, except for the duration_ms variable, which exhibits extreme values. This characteristic should be taken into account during data visualization and analysis, as it may impact the scaling and interpretation of the histogram.

## Data Preparation - Cleaning, Transformation, and Filtering

To prepare the dataset for thorough analysis, we meticulously executed essential data preprocessing steps using the powerful RapidMiner's operators. These preprocessing steps were vital to ensure the dataset's cleanliness, consistency, and suitability for our in-depth investigation. Here are the details steps explained based on Figure 2:

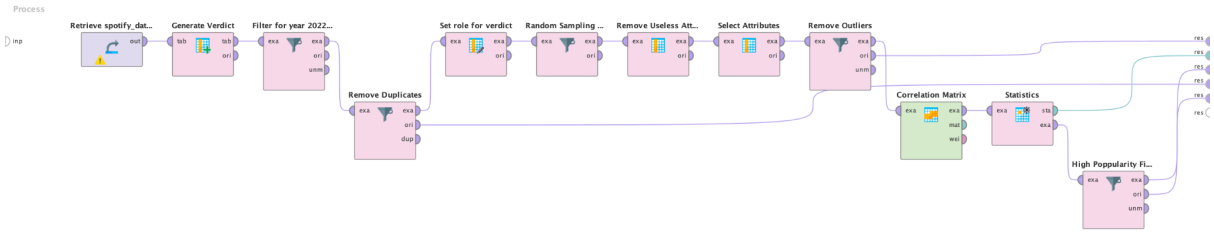


*Figure 2: Rapidminer Processes*

### 1. Data Retrieval

The initial step involved retrieving the Spotify data from the downloaded csv file from Kaggle.

### 2. Generate Verdict

Next, we created a new variable called verdict, the newly created Verdict variable will serve as our target variable for classification in the later stages of data mining.

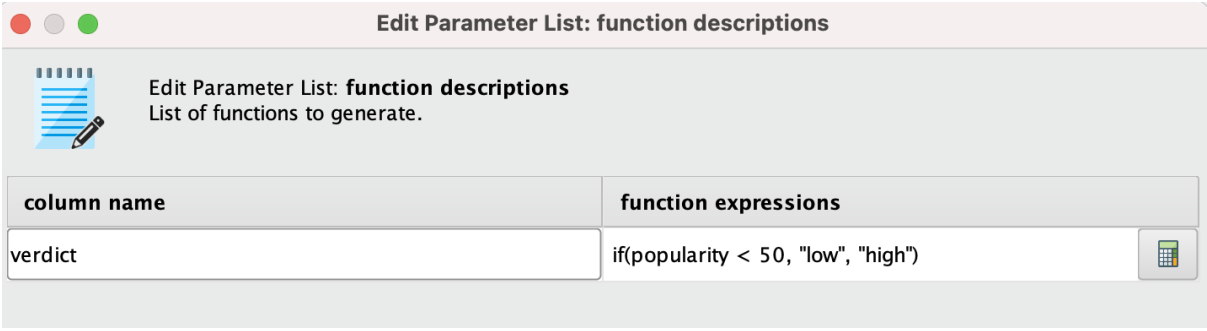


*Figure 3: Generate Verdict Function Descriptions*

In Figure 3, our verdict determination process is depicted, where we carefully establish the criteria for categorizing tracks as either low or high based on their verdict scores. Verdict scores below 50 were classified as low, while those equal to or greater than 50 were classified as high.

By setting this threshold, we aimed to create a clear demarcation between tracks that received relatively lower popularity scores and those that achieved higher levels of acclaim. This classification enables us to differentiate between less popular and more popular tracks, which will be instrumental in our subsequent data analysis and interpretation.

### 3. Filtering for Year 2022

To gain insights into recent music trends, we focused our analysis on tracks released in the year 2022. This filtering process allowed us to narrow down our dataset and concentrate on current music characteristics.

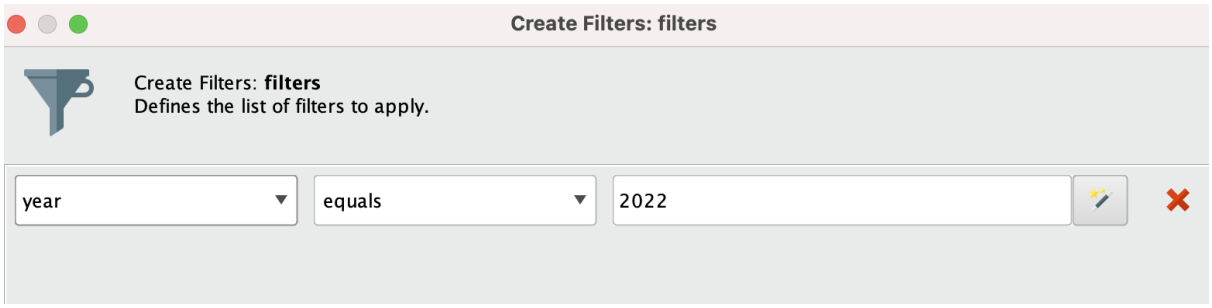


*Figure 4: Filters Tracks for the Year 2022 Release*

The filtering procedure is straightforward: as shown in Figure 4, we merely set the variable year to 2022.

### 4. Duplicate Removal

Ensuring data accuracy, we removed any duplicate entries from the dataset, eliminating any potential redundancy and maintaining data integrity. To eliminate duplicity, we employ the remove duplicate operator in RapidMiner, selecting the variables to be compared for redundancy removal.

### 5. Random Sampling

As part of our classification task, we set roles for the Verdict variable to aid in random sampling. This helped ensure the consistency of the target variable after random sampling, which was necessary to handle the immense size of the dataset.

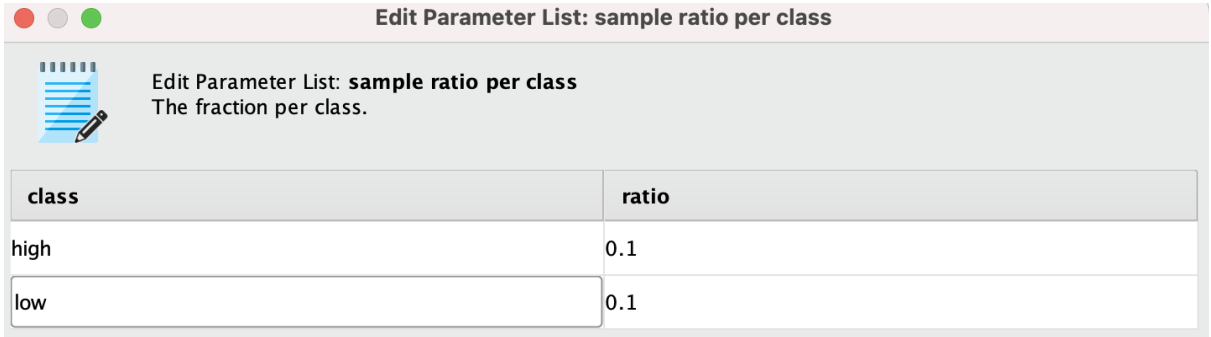


*Figure 5: Random Sampling Setting*

As depicted in Figure 5, we perform a random sampling of 10% from the entire dataset. This sampling approach allows us to work with a representative subset of the data, ensuring computational efficiency while preserving the integrity of the analysis.

## 6. Attribute Removal

Since our analysis is focused on the year 2022, we removed the Year attribute to maintain uniformity in the dataset. Additionally, we eliminated non-essential attributes, such as Track ID, Track Name, and Artist Name, that do not contribute directly to our analysis.

## 7. Handling Outliers

During our observation, we noticed that the Duration_ms attribute was affected by outliers. To mitigate this issue, we removed outliers from the dataset, enhancing the reliability of our analysis and interpretations.

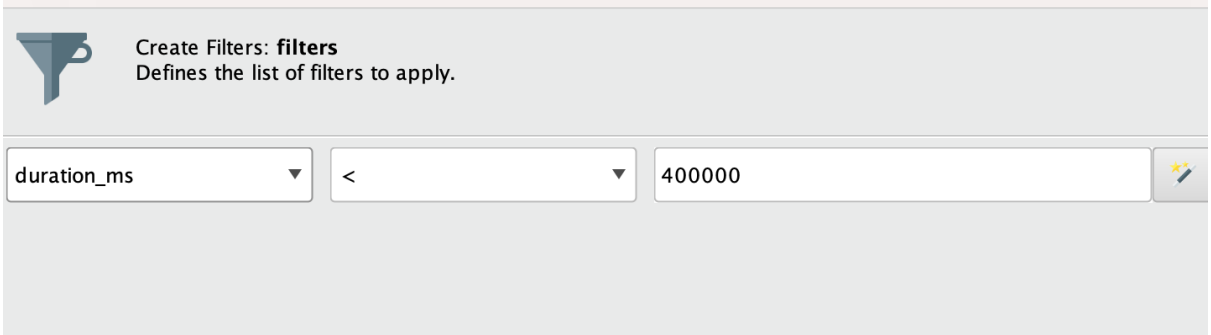


*Figure 6: Filter's Setting for Removing Outliers in duration_ms*

Figure 6 presents the configuration used to filter out data with a duration_ms greater than 400,000. As a result of this filtering process, approximately 230+ data points were removed from the dataset, refining our analysis and enhancing its accuracy.

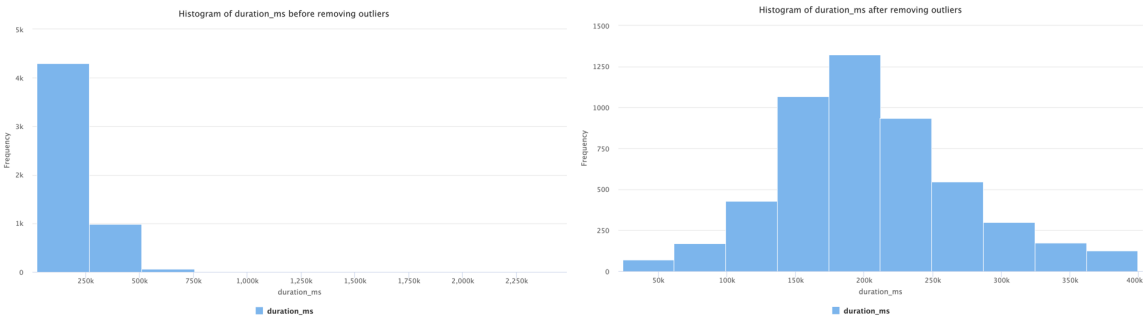


*Figure 7: Comparison of 'duration_ms' Distribution Before and After Outlier Removal*

Figure 7 showcases the comparison of the distribution of duration_ms'' before and after the removal of outliers. Following the outlier removal process, the distribution of the variable exhibits significant improvement. The data points now align more cohesively, resulting in a more refined and representative distribution. This enhancement reinforces the reliability of our analysis and ensures that the outlier-affect data points do not distort our conclusions.

## 8. Correlation Matrix and Statistical Table

To understand attribute relationships, we generated a correlation matrix, which revealed valuable insights into the interdependencies among attributes. Additionally, we created a statistical table to understand attribute distributions and descriptive statistics.

## 9. New Dataset with High Popularity Tracks

To facilitate further analysis, we derived a new dataset containing only high-popularity tracks. This subset of data enabled us to focus on tracks that are most influential on the Spotify platform.

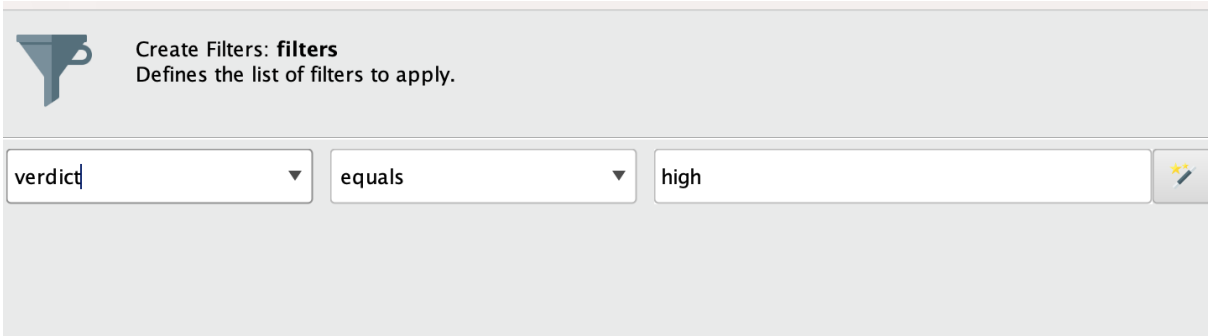


*Figure 8: Filter's Setting for Creating New Dataset with High Popularity Tracks*

Figure 8 displays the filter setting, where we selectively include only the verdicts that are labeled as 'high'.

# Data Exploratory & Analysis

## Correlation Matrix

| Attribu... | popula... | dancea... | energy | key | loudne... | mode | speech... | acousti... | instru... | liveness | valence | tempo | duratio... | time_si... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| popularity | 1 | 0.167 | -0.024 | 0.029 | 0.061 | -0.024 | -0.079 | -0.024 | -0.145 | -0.099 | 0.017 | 0.014 | -0.168 | 0.033 |
| dancea... | 0.167 | 1 | 0.155 | 0.042 | 0.349 | -0.066 | 0.093 | -0.192 | -0.222 | -0.126 | 0.510 | -0.018 | -0.018 | 0.182 |
| energy | -0.024 | 0.155 | 1 | 0.030 | 0.760 | -0.056 | 0.086 | -0.785 | -0.255 | 0.209 | 0.246 | 0.308 | 0.126 | 0.177 |
| key | 0.029 | 0.042 | 0.030 | 1 | 0.018 | -0.159 | -0.012 | -0.029 | -0.009 | -0.023 | 0.052 | 0.006 | 0.009 | 0.007 |
| loudness | 0.061 | 0.349 | 0.760 | 0.018 | 1 | -0.052 | 0.027 | -0.658 | -0.474 | 0.054 | 0.332 | 0.294 | 0.174 | 0.182 |
| mode | -0.024 | -0.066 | -0.056 | -0.159 | -0.052 | 1 | -0.006 | 0.058 | -0.026 | 0.017 | -0.012 | -0.021 | -0.048 | -0.035 |
| speechi... | -0.079 | 0.093 | 0.086 | -0.012 | 0.027 | -0.006 | 1 | 0.034 | -0.161 | 0.290 | 0.082 | -0.017 | -0.150 | -0.022 |
| acoustic... | -0.024 | -0.192 | -0.785 | -0.029 | -0.658 | 0.058 | 0.034 | 1 | 0.184 | -0.047 | -0.149 | -0.278 | -0.167 | -0.177 |
| instrum... | -0.145 | -0.222 | -0.255 | -0.009 | -0.474 | -0.026 | -0.161 | 0.184 | 1 | -0.105 | -0.344 | -0.104 | 0.034 | -0.081 |
| liveness | -0.099 | -0.126 | 0.209 | -0.023 | 0.054 | 0.017 | 0.290 | -0.047 | -0.105 | 1 | 0.028 | -0.008 | -0.077 | -0.016 |
| valence | 0.017 | 0.510 | 0.246 | 0.052 | 0.332 | -0.012 | 0.082 | -0.149 | -0.344 | 0.028 | 1 | 0.103 | -0.103 | 0.097 |
| tempo | 0.014 | -0.018 | 0.308 | 0.006 | 0.294 | -0.021 | -0.017 | -0.278 | -0.104 | -0.008 | 0.103 | 1 | 0.066 | 0.039 |
| duratio... | -0.168 | -0.018 | 0.126 | 0.009 | 0.174 | -0.048 | -0.150 | -0.167 | 0.034 | -0.077 | -0.103 | 0.066 | 1 | 0.041 |
| time_sig... | 0.033 | 0.182 | 0.177 | 0.007 | 0.182 | -0.035 | -0.022 | -0.177 | -0.081 | -0.016 | 0.097 | 0.039 | 0.041 | 1 |

*Figure 9: Correlation Matrix*

With the goal of ranking tracks on Spotify, popularity would be assigned as the dependent variable (DV). From the correlation matrix, all of the variables are weakly correlated to popularity on both the positive and negative scales. The highest correlation is duration_ms, with a negatively weak correlation of -0.168, followed by danceability, with a positively weak correlation of 0.167. While tempo is the least correlated variable, with a 0.014 correlation score. This indicates that a song's popularity will not be clearly explainable by any single variable, which signals the possibility of complex interactions.

## Attribute Weight

| attribute | weight |
|---|---|
| popularity | 0.958 |
| dancea... | 0.654 |
| energy | 0.055 |
| key | 1 |
| loudness | 0 |
| mode | 0.993 |
| speechi... | 0.918 |
| acoustic... | 0.215 |
| instrum... | 0.664 |
| liveness | 0.910 |
| valence | 0.637 |
| tempo | 0.841 |
| duratio... | 0.926 |
| time_sig... | 0.926 |

*Figure 10: Attribute Weight Table*

Popularity will be excluded in this analysis as it has been assigned as the DV. This indicates that the key, which has the highest weight of 1, is wholly included as it is highly relevant and contributes significantly to predicting the target variable. Following closely, mode ranks as the second most important feature with a weight of 0.993, while duration_ms and time_signature vie for third place with weight scores of 0.926. Notably, loudness has an attribute weight of 0, which indicates that it has been completely excluded.

# Exploratory Data Analysis

## Verdict

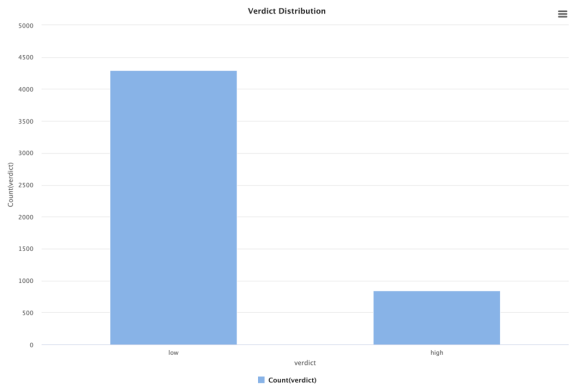


*Figure 11: Verdict Distribution Graph*

From the verdict distribution, it is observed that approximately 20% of the tracks released are popular. This indicates that only 1 out of 5 songs make it in the music industry, and the dataset is imbalanced as the dependent variable is skewed towards songs of low popularity.

## Popularity

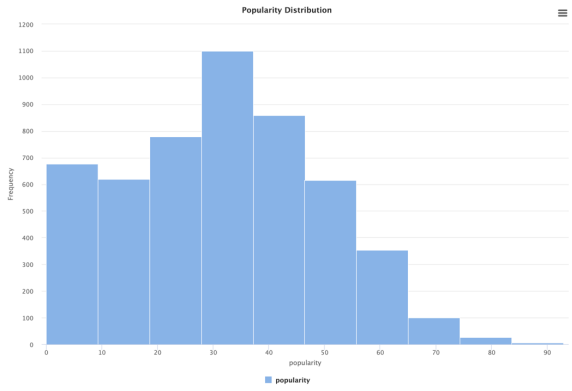


*Figure 12: Popularity Distribution Graph*

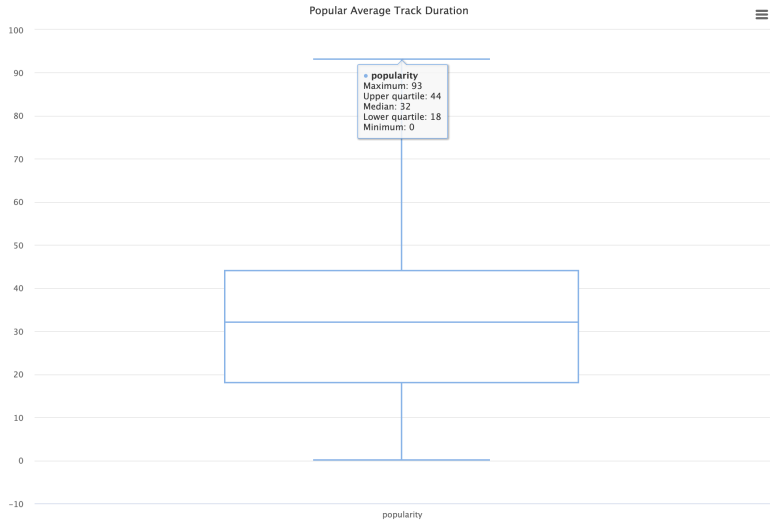


*Figure 13: Popularity Boxplot*

From the popularity histogram (Figure 12), it is depicted that most of the songs do not make the cut with a popularity score of approximately 30-40 which is supported by the boxplot (Figure 13) that indicated a median score of 32. However, it is interesting that many in the third quarter came close to 46 before falling short. This could be an area of investigation to determine possible causes.

## Genre

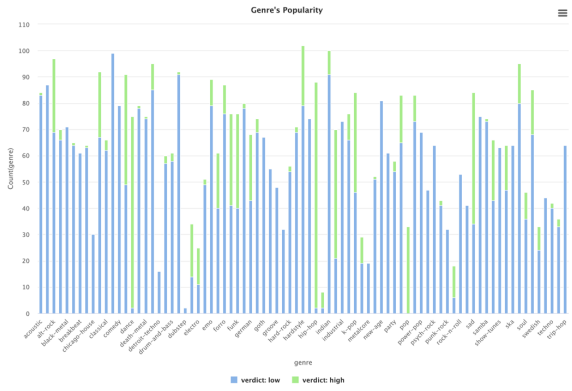


*Figure 14: Popularity Distribution Based on Genre*

Examining the genre popularity distribution (Figure 14), many exciting insights could be obtained. All of the tracks in the pop genre and most dance, hip-hop and house tracks are rated as highly popular. Similarly, over half of indie-pop, rock-n-roll, and sad tracks are popular. This indicates that these genres are well-received by the listeners. Furthermore, this is not an isolated case, as pop has 33 tracks in that particular genre.

## Danceability

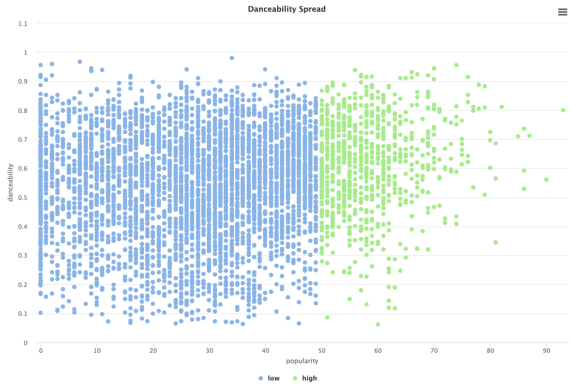


*Figure 15: Danceability Spread based on Popularity in Scatter Plot*

From Figure 15, most of the popular songs have a danceability score between 0.4 and 0.8 scores. The tracks in this range would receive a popularity score of 60 to 70. This indicates that while most tracks may not be popular, optimising the track's danceability score between this range would have a higher chance to succeed as evident in the tight congregations in the high popularity data points, with sporadic outliers being sprinkled around.

## Energy

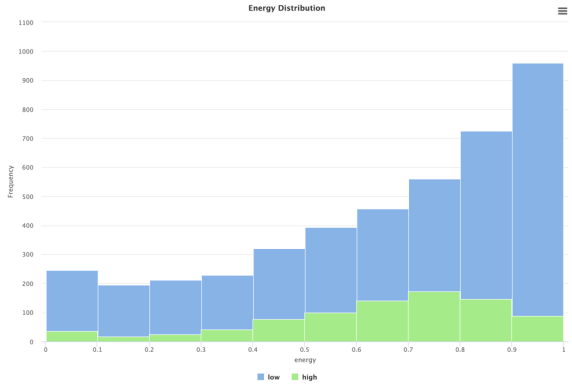


*Figure 16: Popularity Distribution based on Energy Graph*

Unsurprisingly based on the above Figure 16, low-energy songs do not appeal to listeners. Popular songs are observed to have energy scores between 0.6 to 0.8. However, too much energy (above 0.8) has been seen to experience a decline in popularity. This is heavily substantiated by the fact that while a significant portion of tracks have an energy score of 1, the popularity remains low.

## Speechiness

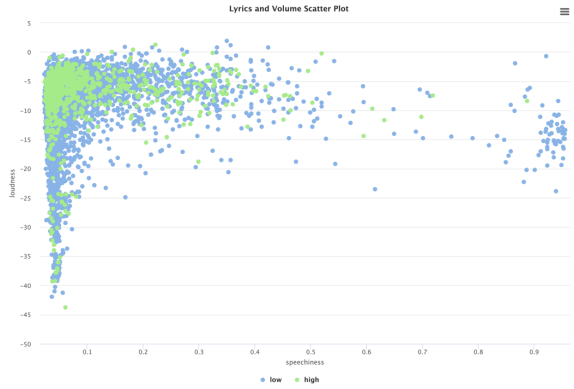


*Figure 17: Lyrics and Volume Scatter Plot based on Popularity*

From the Lyrics and Volume scatter plot (Figure 17), popular songs tend not to have more than 0.2 lyrics with a controlled volume between -20db to 0db. This trend is in line with most of the songs released. However, where it differs is the gap between -25db to -20db for popular songs. This means that tracks should be suggested not to have more than 0.2 lyrics and avoid going lower than -20db.

## Liveness

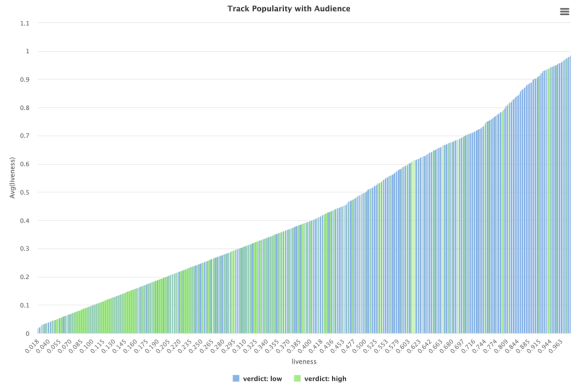


*Figure 18: Track Popularity with Audience Graph*

A surprising discovery in Figure 18 is that popular songs have at least a limited degree of audience in the track. This means that having audiences in the track background could popularise a track as evident in the chart above where most popular songs have a liveness score of 0.05 to 0.25.

## Valence

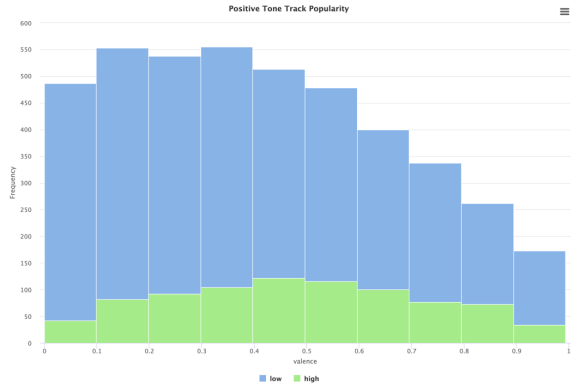


*Figure 19: Popularity Distribution based on Valence*

Figure 19 shows upbeat tone tracks are generally poorly received by listeners. Instead, tracks with a mid-emotional tone would be popular as contrary to the expectation of highly emotional tracks, which would be more relatable to the listener by logical deduction.

**Duration (ms)**

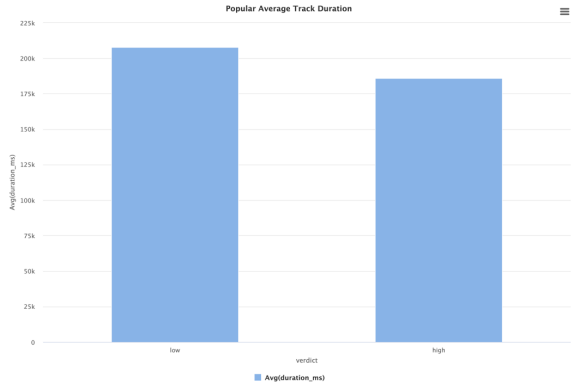


*Figure 20: Average Track Duration (in ms) Based on Popularity*

Songs' duration is a contesting topic, as the length of a song's duration could range from a minute to over an hour. From the average duration tabulated in Figure 20, the preferred duration is approximately 186,000 milliseconds (3.1 minutes). This indicates the average concentration of attention paid to a track for the Spotify user base before they lose interest.

## Modelling Approach

In this project, the area of analysis is focused on the popularity of the tracks available on Spotify. Therefore, we have phrased our approach as follows:

1. **Popularity – Raw score**

   The popularity rating of a track is the relative presentation which would be analysed via regression. This would allow us to predict a song's popularity score on a spectrum. As the score is a continuous variable with a known outcome (supervised), the models in consideration are generalised linear model (GLM), random forest (RF) decision tree, and support vector machine (SVM).

2. **Verdict – Classify if the track will be a hit**

   As raw scores are on a more granular level, artists and studios are focused on producing hit tracks. In this context, with a max rating ceiling of 100, 50 will be assigned as the separator benchmark. This would facilitate the classification of whether a song is classified by its popularity of being high or low (binary classification). Therefore, the models in consideration are GLM, logistic regression, RF, and SVM.

3. **Cluster – Identify group characteristics**

   In order to identify the next hit, features of certain groups of popular tracks have to be identified. However, in this context, we do not know the appropriate number of clusters. Thus, if K-means were to be relied upon, the assumption of clusters was to be made as it has to be defined prior. Therefore, the X-mean model was implemented as the optimal number of clusters is not known in advance.

# Model Selection & Evaluation

## Popularity - Regression

Root Mean Squared Error

| Model | Root Mean Squared... | Standard Deviation | Gains | Total Time | Training Time (1,0... | Scor |
|---|---|---|---|---|---|---|
| Generalized Linear Model | 9.718 | ± 0.529 | ? | 2 min 56 s | 6 ms | 19 n |
| Random Forest | 10.825 | ± 0.43 | ? | 9 min 23 s | 14 ms | 58 n |
| Support Vector Machine | 10.33 | ± 0.478 | ? | 2 min 9 s | 346 ms | 204 |


*Figure 21: Regression Result Table*

From the regression result table in Figure 21, the GLM performs best with the lowest MSE of 9.718. This is supported by the respective prediction charts, where most of the data points of the GLM cluster around the Line of best fit, as shown above. This is in contrast to the deviation at the starting point of RF and the deviation at the top area of SVM.

Generalized Linear Model – Weights

| Attribute | Weight |
|---|---|
| sqrt([acousticness]/[energy]) | 0.298 |
| energy | 0.152 |
| genre | 0.041 |
| mode | 0 |

*Figure 22: GLM - Weights Table*

From the above Figure 22, sqrt(acousticness/energy), a tabulated attribute has the heaviest weight of 0.298. After which, original attributes of energy, and genre rank second and third respectively with a weight of 0.152 and 0.041.

**Verdict - Classification**

Accuracy

| Model | Accuracy | Standard Deviation | Gains | Total Time | Training Time (1,0... | Scoring Ti |
|---|---|---|---|---|---|---|
| Generalized Linear Model | 84.9% | ± 0.4% | 200 | 8 min 22 s | 27 ms | 143 ms |
| Logistic Regression | 84.4% 84.9% | ± 0.3% | 172 | 15 min 54 s | 28 ms | 103 ms |
| Random Forest | 84.0% | ± 0.9% | 174 | 19 min 57 s | 35 ms | 760 ms |
| Support Vector Machine | 84.8% | ± 0.5% | 182 | 16 min 59 s | 199 ms | 286 ms |

F Measure

| Model | F Measure | Standard Deviation | Gains | Total Time | Training Time (1,0... | Scoring Ti |
|---|---|---|---|---|---|---|
| Generalized Linear Model | 91.0% | ± 0.3% | 200 | 8 min 22 s | 27 ms | 143 ms |
| Logistic Regression | 90.8% | ± 0.2% | 172 | 15 min 54 s | 28 ms | 103 ms |
| Random Forest | 90.7% | ± 0.6% | 174 | 19 min 57 s | 35 ms | 760 ms |
| Support Vector Machine | 90.9% | ± 0.3% | 182 | 16 min 59 s | 199 ms | 286 ms |

*Figure 23: Classification Models' Results*

In this context of classification, both accuracy and F- scores are taken into account. This is essential as accuracy maximises overall prediction in the imbalanced dataset as noted prior. However, high accuracy may be misleading and may not be useful for minority class prediction. This is where F-Measure takes effect. When there is a big variation in class frequencies, the F-score is more informative and takes into account class imbalance. The best accuracy and F-Measure results show that the GLM model is well-balanced (similar numbers of positive and negative cases), producing fewer errors.

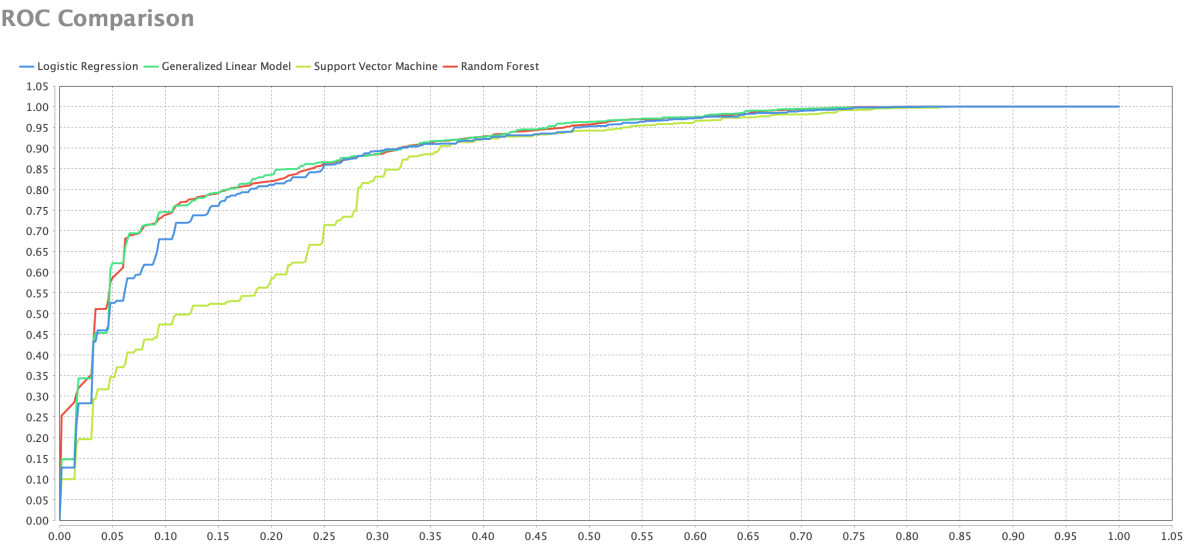


*Figure 24: ROC Comparison Graph*

From the ROC, GLM performs slightly better after 0.05, and tends to outperform the other models. However, when < 0.05, Random Forest may then be preferred as the model of choice.

Generalized Linear Model – Weights

| Attribute | Weight |
|---|---|
| genre | 0.072 |
| energy | 0.034 |
| loudness | 0.015 |
| danceability | 0.007 |

*Figure 25: GLM - Weights Table from Classification Model*

In contrast to regression, there is an absence of tabulated attributes. Genre, energy, loudness and danceability are the key features taken into account in this model.

## Clustering - High & Low

x-Means – Summary

Number of Clusters: 6

**Cluster 0** 67

**instrumentalness** is on average **86.41%** larger, **time_signature** is on average **74.63%** smaller, **energy** is on average **35.35%** smaller

**Cluster 1** 2,353

**mode** is on average **62.82%** larger, **instrumentalness** is on average **39.34%** smaller, **liveness** is on average **21.48%** smaller

**Cluster 2** 622

**instrumentalness** is on average **220.81%** larger, **energy** is on average **74.91%** smaller, **speechiness** is on average **61.55%** smaller

**Cluster 3** 395

**liveness** is on average **240.04%** larger, **energy** is on average **20.34%** larger, **instrumentalness** is on average **17.94%** smaller

**Cluster 4** 87

**speechiness** is on average **1,027.26%** larger, **liveness** is on average **217.69%** larger, **instrumentalness** is on average **99.99%** smaller

**Cluster 5** 1,613

**mode** is on average **100.00%** smaller, **liveness** is on average **21.90%** smaller, **instrumentalness** is on average **21.56%** smaller

*Figure 26: X-Means High & Low Clustering Summary*

x-Means - Centroid Table

| Cluster | energy | instrumentalness | key | liveness | loudness | mode | speechiness | time_signature |
|---|---|---|---|---|---|---|---|---|
| Cluster 0 | 0.419 | 0.435 | 5.194 | 0.210 | -15.216 | 0.657 | 0.085 | 0.985 |
| Cluster 1 | 0.714 | 0.141 | 4.875 | 0.178 | -6.777 | 1 | 0.083 | 3.939 |
| Cluster 2 | 0.163 | 0.748 | 5.198 | 0.131 | -20.816 | 0.696 | 0.051 | 3.791 |
| Cluster 3 | 0.780 | 0.191 | 4.947 | 0.711 | -8.014 | 0.668 | 0.103 | 3.929 |
| Cluster 4 | 0.565 | 0.000 | 4.805 | 0.665 | -13.367 | 0.701 | 0.848 | 3.644 |
| Cluster 5 | 0.721 | 0.183 | 6.102 | 0.177 | -6.826 | 0.000 | 0.091 | 3.958 |

*Figure 27: X-Means High & Low Centroid Table*

x-Means - Centroid Chart

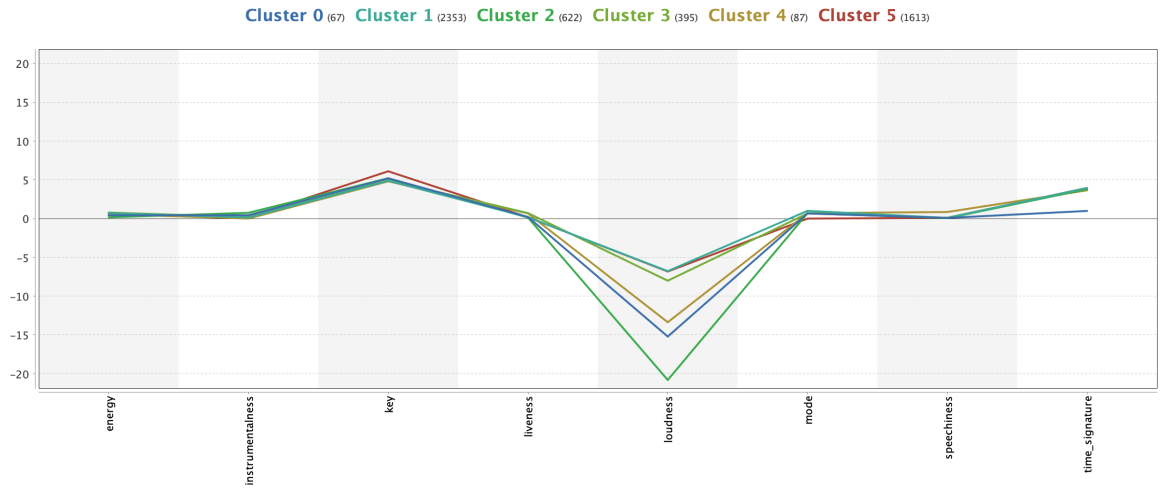


*Figure 28: X-Means High & Low Centroid Graph*

From the above chart (Figure 26), we can observe that Cluster 1 and 5 are the largest. It is noted that both groups have similar attributes of smaller instrumentalness and liveness. Where they differ is the mode, where Cluster 1 has a larger mode while Cluster 5 has a smaller mode on average. However, among all the clusters, the key differentiator is loudness, as seen in Figure 28 where the respective cluster deviates.

## Cluster - High

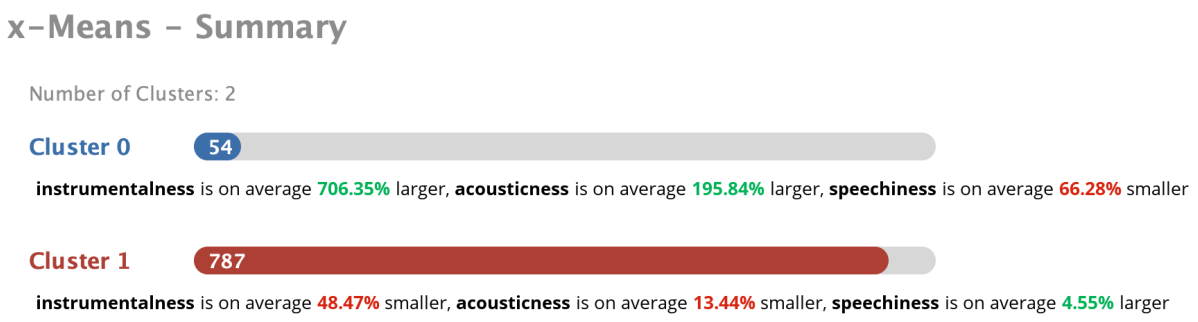


*Figure 29: X-Means High Clustering Summary*

x-Means - Centroid Table

| Cluster | acousticness | danceability | instrumentalness | loudness | speechiness | time_signature | valence |
|---|---|---|---|---|---|---|---|
| Cluster 0 | 0.898 | 0.369 | 0.851 | -26.416 | 0.050 | 3.630 | 0.178 |
| Cluster 1 | 0.263 | 0.637 | 0.054 | -6.471 | 0.104 | 3.929 | 0.504 |

*Figure 30: X-Means High Centroid Table*

x-Means - Cluster Tree

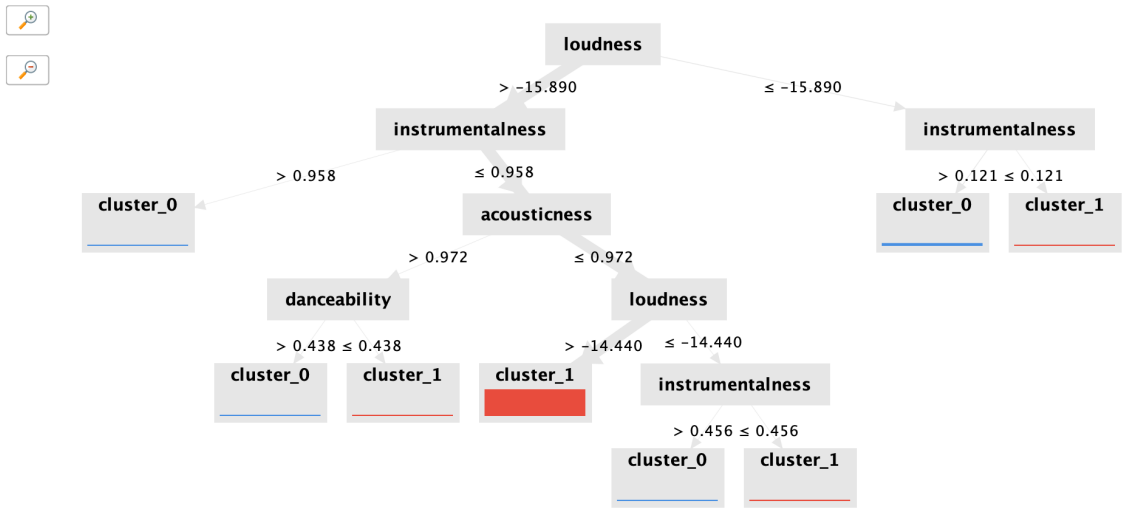


*Figure 31: X-Means High Cluster Tree*

Based on Figure 29, with an emphasis on unravelling the secrets to popular tracks, a further cluster analysis was performed on the high popularity subset. The popular tracks could be classified into two different clusters where instrumentalness, acousticness, and speechiness are critical. From the cluster tree above in Figure 31, we can observe that the majority of the tracks have the following characteristics:

1. Instrumentalness > -15.890
2. Acousticness < or = 0.958
3. Loudness > -14.440db

These characteristics could form the basis of determining whether a track could be popular. However, it should be noted that this differs from the other two analyses, which highlighted the common weighted attributes of energy and genre.

# Deployment Plan

The primary objective of the deployment plan is to leverage the insights obtained through data mining to enhance the Spotify music streaming platform. By implementing the models and recommendations, we aim to improve the personalized music recommendations provided to users, optimize the music library curation, and elevate the overall user experience.

## Deployment Steps

1. Model Integration

   The models developed during the data mining process, such as classification models for track popularity, will be integrated into the existing Spotify platform. These models will serve as the foundation for delivering personalized music recommendations to users based on their preferences and music history.

2. API Development

   To enable seamless integration with the Spotify platform, we will develop application programming interfaces (APIs) that facilitate data exchange between the data mining system and the platform. These APIs will ensure real-time data flow, enabling the platform to access the latest insights and recommendations.

3. Testing and Validation

   Before deployment, rigorous testing and validation of the models and algorithms will be conducted to ensure their accuracy and reliability. Validation against historical user data and A/B testing will be performed to assess the models' effectiveness and fine-tune recommendations.

4. User Interface Enhancements

   The user interface of the Spotify platform will be updated to incorporate the new personalized music recommendations. This will include features such as Recommended for You playlists and tailored song suggestions, providing users with a more engaging and satisfying music discovery experience.

5. Monitoring and Maintenance

   Post-deployment, continuous monitoring of the system's performance and user feedback will be carried out. Regular maintenance and updates will be implemented to ensure that the models stay up-to-date with evolving user preferences and music trends.

6. Performance Evaluation

   Performance evaluation metrics will be established to measure the impact of the deployed models on user engagement, customer satisfaction, and overall platform usage. The success of the deployment will be assessed against predefined key performance indicators (KPIs).

## Execution Timeline

The deployment plan will be executed in several phases, with specific milestones and deliverables defined for each stage. The timeline for deployment will be determined based on the complexity of integration, testing, and user interface enhancements. Here is the comprehensive timeline for the deployment plan, outlining key dates for the completion of each deployment step.

- **Phase 1** - Model Development and Integration
  Duration: 4 weeks
  Develop and fine-tune classification models for track popularity and integrate them into the Spotify platform. Conduct initial testing to ensure model accuracy.

- **Phase 2** - API Development and Data Flow
  Duration: 2 weeks
  Create application programming interfaces (APIs) to enable seamless data exchange between the data mining system and Spotify platform. Implement real-time data flow for up-to-date insights

- **Phase 3** - Testing and Validation
  Duration: 3 weeks
  Conduct comprehensive testing and validation of the deployed models. Perform A/B testing and validate against historical user data for model effectiveness.

- **Phase 4** - User Interface Enhancements
  Duration: 2 weeks
  Update the Spotify user interface to incorporate personalized music recommendations. Design and implement features such as Recommended for You playlists.

- **Phase 5** - Monitoring and Maintenance
  Duration: Indefinitely
  Implement continuous monitoring of the system's performance and user feedback. Regular maintenance and updates to keep models up-to-date with evolving user preferences and music trends.

- **Phase 6** - Performance Evaluation
  Duration: 4 weeks
  Evaluate the impact of deployed models on user engagement, customer satisfaction, and platform usage. Measure success against predefined key performance indicators (KPIs).

# Conclusion

In conclusion, our data mining project delved into the Spotify 1 Million Tracks Dataset to gain valuable insights into music track characteristics and popularity on the Spotify platform. By employing data preprocessing techniques using the RapidMiner operators, we ensured that the dataset was clean, consistent, and suitable for our analysis.

The primary objective of our analysis was to explore recent music trends in 2022, discover patterns related to track popularity, and understand how different attributes influence a track's reception among users. With this information, Spotify can enhance its user experiences, refine music recommendation algorithms, and optimize its music library curation.

Our analysis also revealed interesting findings related to the correlation between various attributes and track popularity. While no single attribute overwhelmingly determines a track's popularity, we identified patterns suggesting that elements like **danceability**, **energy**, **valence**, and **duration** play crucial roles in influencing a track's reception. By understanding these musical characteristics, Spotify can better tailor recommendations to users' preferences.

Furthermore, we established the **Verdict** variable as our target variable for classification models. This variable categorized tracks as Low or High based on their verdict scores, which served as a foundation for our classification analysis. By accurately classifying tracks, Spotify can focus on promoting high-potential tracks, thus maximizing their potential for commercial success.

However, it is worth noting that the dataset's vastness allowed us to analyze only a subset of approximately 0.5% due to limitations in device processing capabilities. To unlock even more valuable insights, we recommend scaling up the data size and performing more extensive analyses. Additionally, incorporating time series analytics could reveal seasonal trends in music preferences, providing critical information for decision-making and content promotion.

Our deployment plan outlines the strategic steps to integrate the developed models into the Spotify platform successfully. By leveraging APIs and continuous monitoring, Spotify can deliver personalized music recommendations and maintain an up-to-date music library to cater to its vast user base's diverse tastes.

Overall, this data mining project has provided valuable insights that can empower Spotify to optimize its music library curation, refine music recommendation algorithms, and enhance user satisfaction. By continuously improving its platform based on data-driven insights, Spotify can strengthen its position as a leading music streaming service, benefitting both the platform and its millions of users.

# Appendix

## Appendix A

The dataset contains a diverse set of attributes for each track, allowing us to examine various aspects of music and its reception among users. Some of the key variables included in the dataset are:

| **Audio (Features)** | **Description** |
|---|---|
| Popularity | This attribute represents the popularity of a track on Spotify and is measured on a scale from 0 to 100. Higher popularity scores indicate that the track is more frequently streamed and well-received by users. |
| Year | The Year variable denotes the year of the track's release, ranging from 2000 to 2023. This information is essential for studying music trends over time and understanding how musical preferences have evolved across different eras. |
| Danceability | The Danceability attribute quantifies the suitability of a track for dancing. Tracks with higher danceability scores are more likely to have a rhythmic and energetic composition, making them ideal for dance enthusiasts. |
| Energy | This variable measures the intensity and activity level of a track. High-energy tracks tend to be more powerful and dynamic, while low-energy tracks are typically more calming and mellow. |
| Key | The Key attribute represents the central note of the track, ranging from -1 to -11. Each value corresponds to a specific musical key, and this variable can help identify popular musical keys in different genres. |
| Loudness | Loudness indicates the volume of the track in decibels, with values ranging from -60 to 0 dB. It provides insights into the overall loudness or intensity of a track. |
| Mode | The Mode variable represents the modality of the track, with 1 indicating a major key and 0 indicating a minor key. This attribute can influence the emotional character of the music. |
| Speechiness | This attribute measures the presence of lyrics (words) in the track. Tracks with higher speechiness are likely to have more vocal content, while instrumental tracks have lower speechiness scores. |

| | |
|---|---|
| Acousticness | The Acousticness variable quantifies the extent to which a track is acoustic or electronic. Values closer to 1.0 indicate a more acoustic nature, while values closer to 0 represent electronic or synthesized music. |
| Instrumentalness | Instrumentalness represents the degree of vocals in a track. Tracks with higher instrumentalness are more likely to be instrumental compositions. |
| Liveness | The Liveness attribute indicates the presence of an audience during the recording of a track, ranging from 0 to 1.0. Higher values suggest live recordings with audience participation. |
| Valence | Valence measures the positiveness conveyed by a track. Higher valence scores indicate a more positive or happy emotional tone in the music. |
| Tempo | Tempo denotes the rhythm of a track in beats per minute (BPM). It offers insights into the track's pace and energy level. |
| Time_signature | The Time_signature attribute provides an estimated time signature (rhythmic structure) of the track, ranging from 3 to 7. It provides information about the track's rhythmic characteristics. |
| Duration_ms | Duration_ms represents the duration of a track in milliseconds. It provides insights into the length of the music piece. |