Data Analytics Research on Spotify Music Ranking: An In-Depth Exploration

Table of Contents

Introduction

Understanding the elements that affect song rankings is crucial for both music platforms and artists in the fast-paced world of music streaming. We explore the fascinating world of Spotify song ranking in this paper, using advanced data analytics approaches to reveal underlying trends and insights. Our research is divided into a number of components, each of which adds to our knowledge of the underlying dynamics at play. We begin by delving into an exploratory data analysis to understand the nuances of the multitude of audio features that might have an influence on music rankings. Following this, we create prediction models utilising random forest and logistic regression to identify the major determinants of song popularity. Subsequently, we employ the technique of Principal Component Analysis (PCA) to pinpoint the key elements influencing the development of the music industry. Finally, in order to reveal hidden patterns and groups throughout the enormous musical repertoire, we embrace the field of clustering algorithms to identify commonalities within each unique cluster of audio tracks.

Data Preparation

The Spotify dataset consists of over 1 million rows with 19 variables. However, due to device capabilities limitations, the most recent completed year (2022) will be utilised in this report. As the number of observations in 2022 consists of over 53,000 rows, a further subset of 10% of the filtered selection will be utilized for efficient processing and modelling purposes. It should be noted that although most of the data are reasonably well distributed, there are instances of skewed or extreme values. These anomalies are not removed due to random subsetting, which might not be outliers as a whole. Furthermore, logistics transformation is not performed as the data are primarily between 0 to 1. Those that exist outside of the 0 to 1 range have values that have meaningful contextual representation. Therefore, no further work to be performed was suggested.

In order to facilitate classification for high or low popularity grouping, a new dependent variable, “verdict”, was generated, This new variable is based on the raw popularity score, where tracks with a popularity score of 50 or above popularity score are classed “popular”, and the remaining as “low”.

Exploratory Data Analysis

Correlation Plot

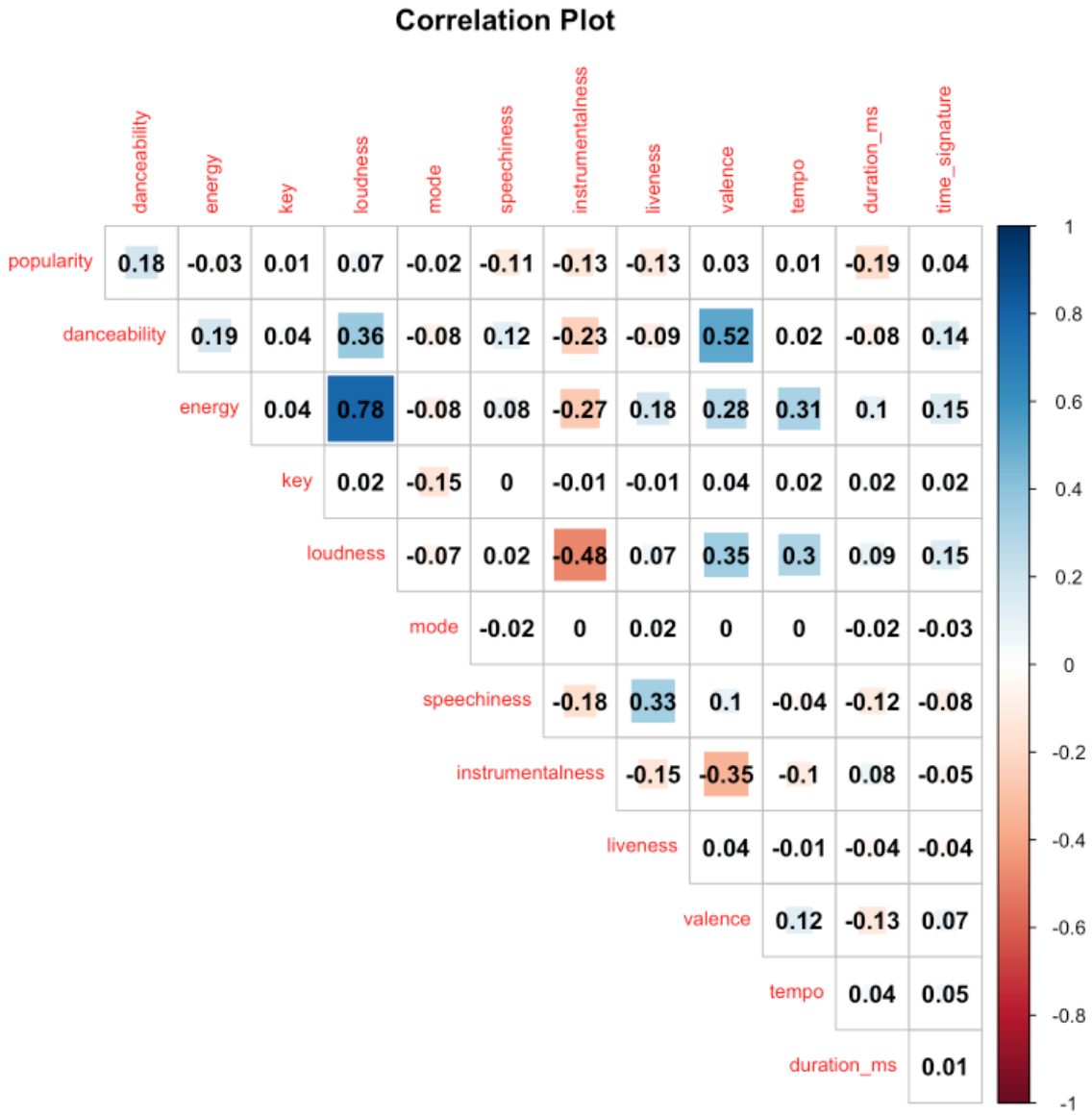

Figure 1: Correlation Plot

An indirect classifier (popularity) would be utilised to assess the relationship with the other relevant parameters in order to rank music on Spotify. Based on Figure 1, all of the variables are only slightly associated with popularity on both positive and negative scales, according to the correlation matrix. Duration_ms has the strongest correlation with a negatively weak correlation of -0.19, while danceability comes in second with a positively weak correlation of 0.18. Tempo is noted to be the least correlated variable, with a 0.01 correlation score. This suggests that there may be complicated relationships since a song's popularity cannot be well explained by any one factor.

```
> print("Highly Correlated Pairs: \n")
[1] "Highly Correlated Pairs: \n"
> print(highly_correlated_positive_pairs)
     Variable1    Variable2 Correlation
1      valence danceability    0.516591
2     loudness       energy    0.783326
3       energy     loudness    0.783326
4 danceability      valence    0.516591
> print("\nHighly Correlated Negative Pairs: \n")
[1] "\nHighly Correlated Negative Pairs: \n"
> print(highly_correlated_negative_pairs)
[1] Variable1   Variable2   Correlation
<0 rows> (or 0-length row.names)
```

Figure 2: List of Highly Correlated Pairs

From the above Figure 2, we can observe that while there isn't any strong association between features and the popularity of a track, there are strong associations among the independent features. Loudness and energy are highly correlated, which would be a logical deduction. However, an interesting find from this review is danceability of a track is associated with the track's emotional positiveness. On the contrary, there are no strongly negatively correlated features pair.

Patterns and Trend Analysis

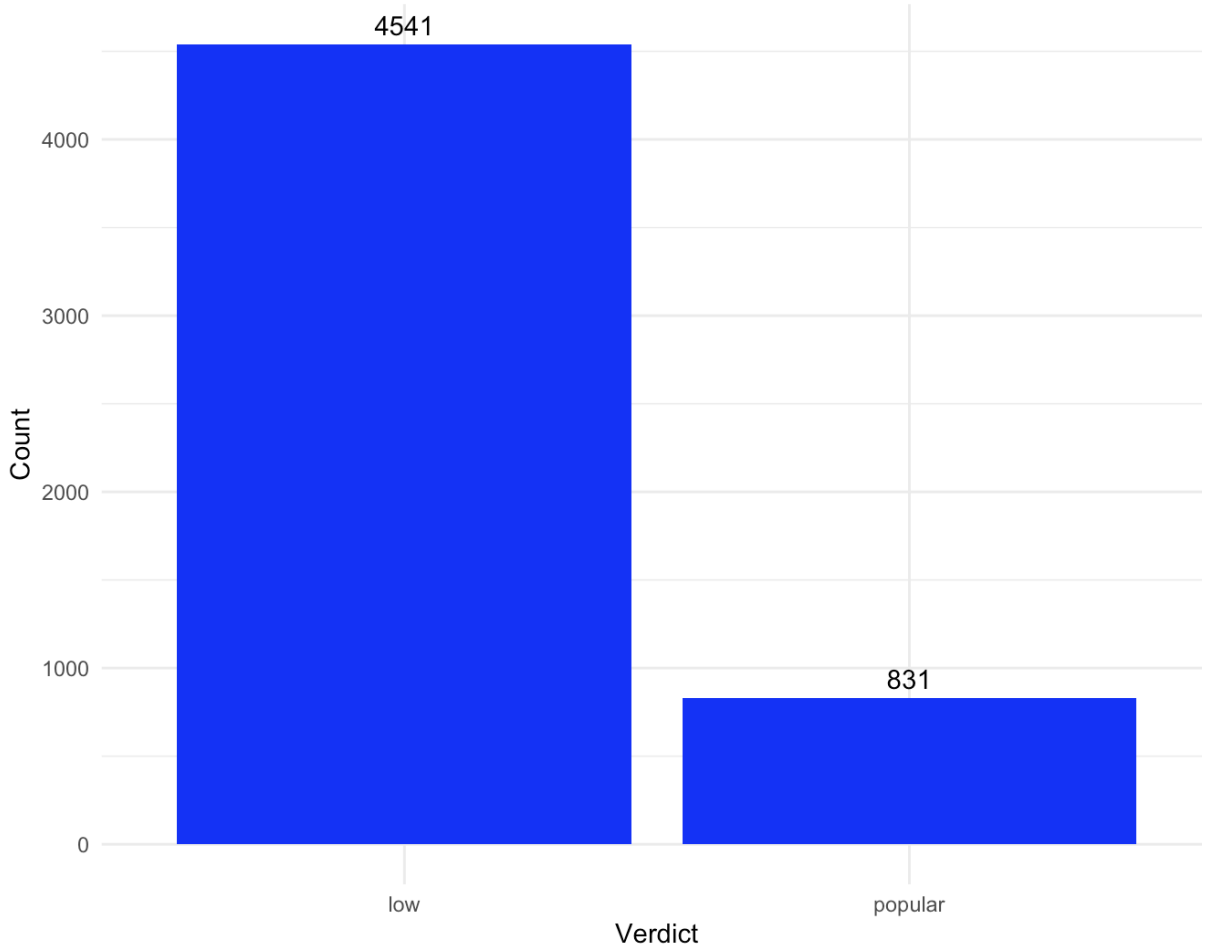

Figure 3: Graph of Verdict's Distribution

It can be seen from the verdict distribution (Figure 3) above that around 20 percent of the published music is well-received by Spotify platform users. This shows that the Creative Audio industry experiences a high failure rate, with just one out of every five songs being

successful. From the dataset, we can conclude that the dataset is unbalanced since the dependent variable is biased towards songs with low popularity which could result in complications during the modelling process.

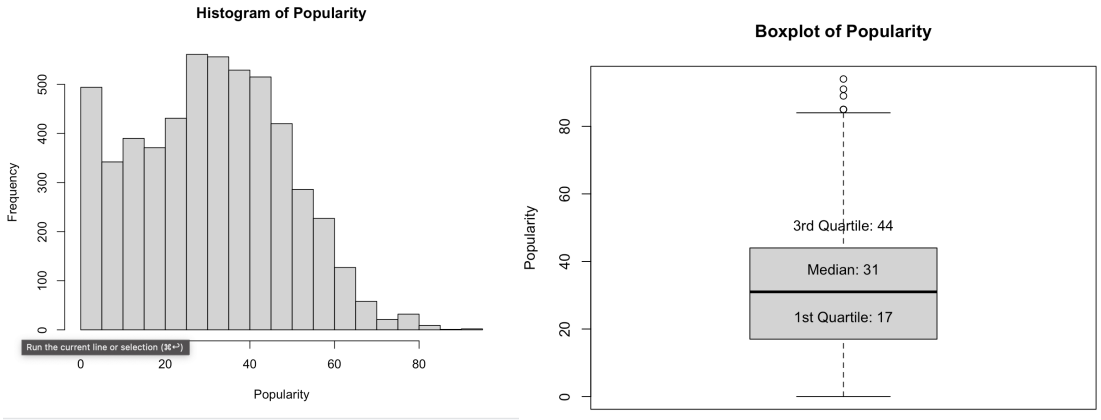

Figure 4: Histogram and Boxplot of Popularity

According to the popularity histogram, the majority of the songs had popularity scores between 30 and 40, which is corroborated by the boxplot, which showed a median score of 31. It is intriguing that many in the third quarter, scoring 44, got close to 50 before falling short. Investigating this region could help identify potential causes and provide insights to this group of artists.

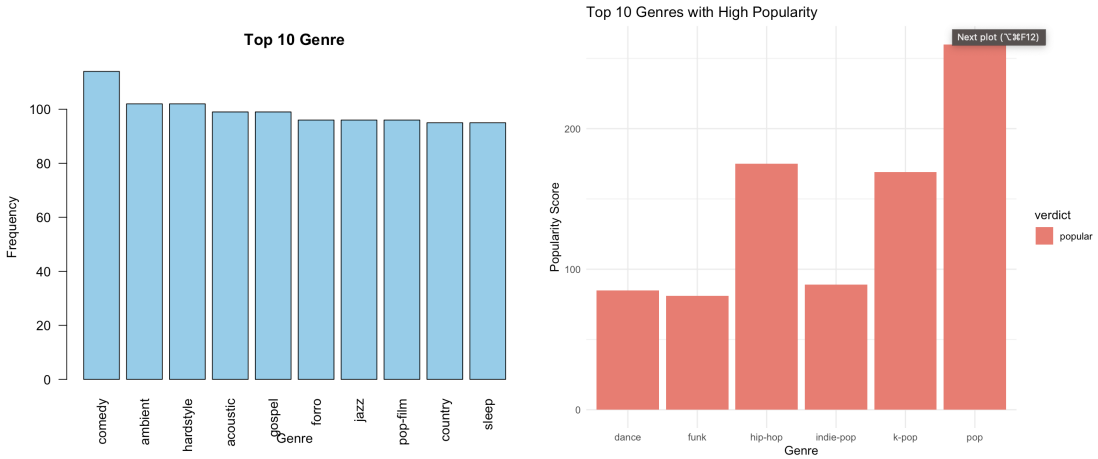

Figure 5: Graph for Top 10 Genre for the Whole Tracks and High Popularity of 2022

Cross-examining the above charts from Figure 5, it is evident that the creative audio industry is not a numbers game where quantity releases would result in a hit track of high popularity score. None of the genres from the largest 10 are in the top 10 in terms of popularity. This is definitely interesting with the old adage saying of “quality over quantity”.

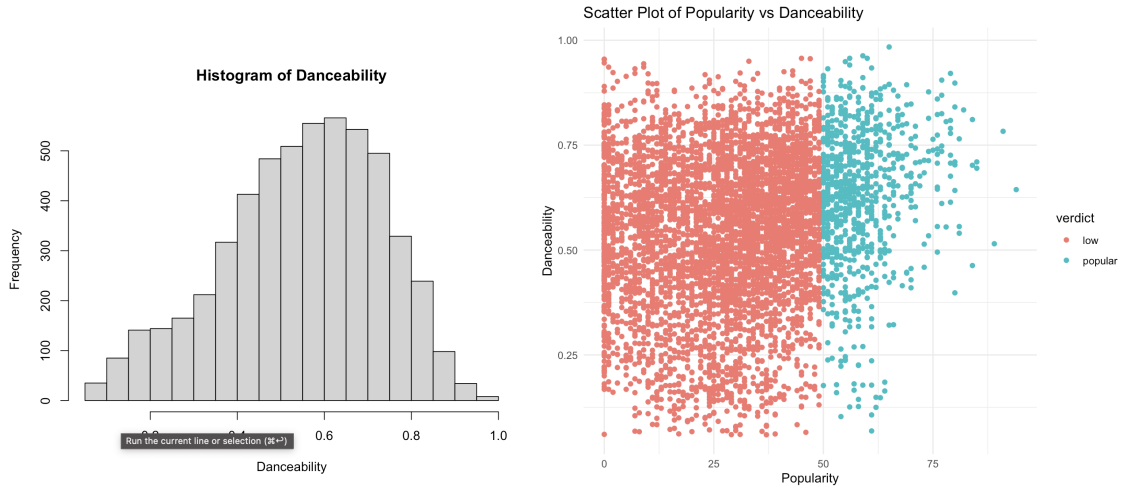

Figure 6: Histogram of Danceability (Left) and Popularity Distribution for Danceability Graph (Right)

Tracks are ever present in many aspects of our lives. However, popular tracks selected for dance scenes appear to have common characteristics. Based on Figure 6, these popular tracks have a danceability score of between 0.4 to 0.8, as evident in the tight clustering. As the popularity score improves, the danceability begins to congregate around 0.6 to 0.8. Therefore we are able to conclude that tracks that are danceable tend to be popular hits which are in sync with the number of tracks released that possess a particular dance score of 0.3 to 0.9.

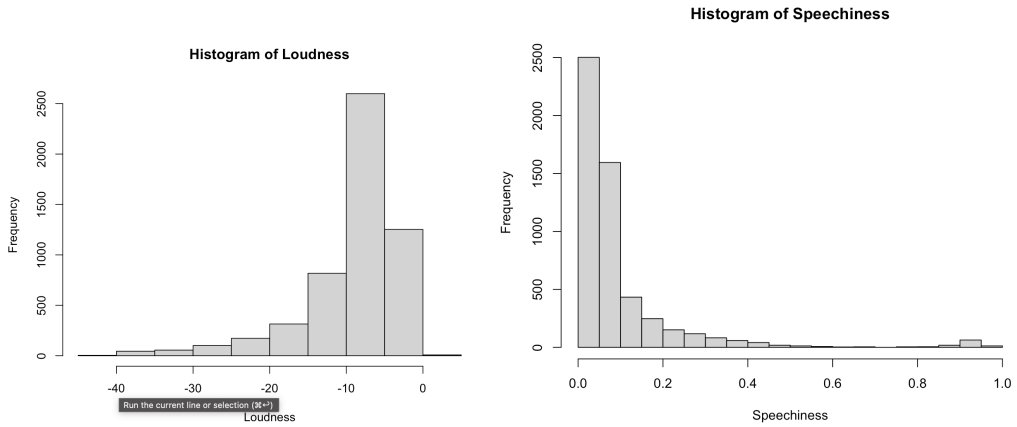

Figure 7: Histogram of Loudness (Left) and Histogram of Speechniness (Right)

Comparing loudness to the presence of lyrics (speechiness), it appears that most tracks have a higher tone with lesser lyrics.

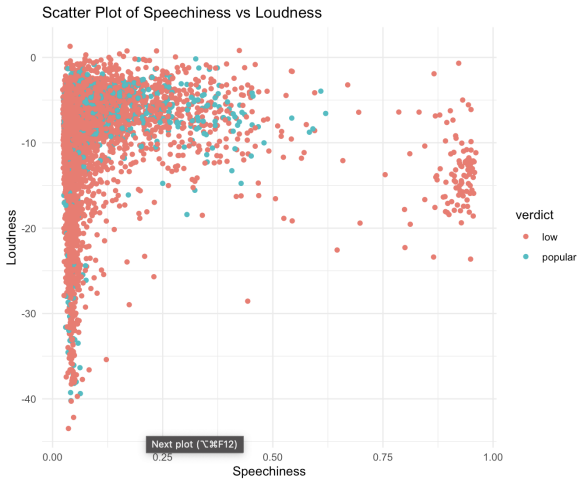

Figure 8: Speechiness and Loudness Scatter Plot based on Popularity

The characteristics of the tracks released are supported by their popularity. The high-popularity tracks possess similar traits to most audio tracks released, and those that deviate are less likely to be popular. Most popular songs have a speechiness score of less than 0.5. This is interesting as it reveals the fact that listeners might actually enjoy the background audio to a certain extent or are unable to keep up with the barrage of word lyrics.

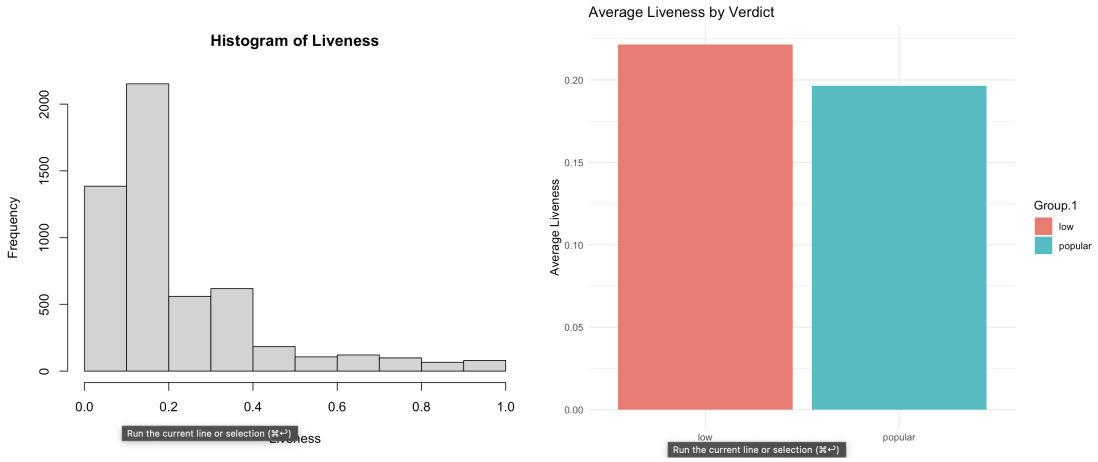

Figure 9: Histogram of Liveness (Left) and Average Liveness by Verdict (Right)

Another interesting tidbit found from Figure 9 was that Spotify users do prefer the audio of the audience in the background, similar to a concert, albeit to a limited extent of 0.18. This is depicted by the average liveness, and anything greater would be rejected. This would be against the logical deduction of having a "release album" containing only the singer's voice.

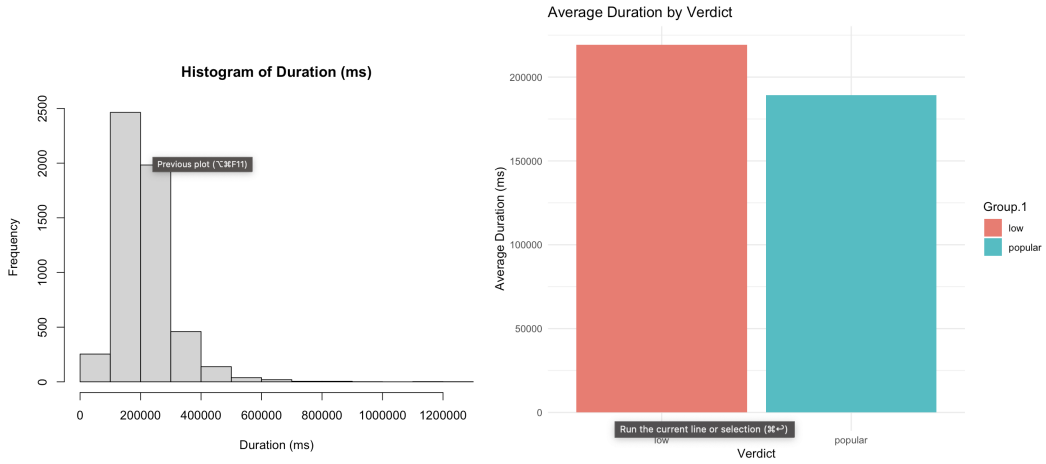

Figure 10: Histogram of Duration (Left) and Average Duration by Verdict (Right)

The length of a song's duration can range from a minute to more than an hour, making it a contentious issue. The preferred duration, based on the average duration calculated on Figure 10, is around 186,000 milliseconds (3.1 minutes). This shows the typical amount of focus a Spotify user will give to a song before losing interest.

Research Questions

1. Predicting Song Popularity

Using classification techniques, we would build a model, evaluate, and predict the popularity of a song. In this section, we will use the audio features in the dataset as the independent features and verdict as the classification target. After which, evaluation of the model performance will include specifically the accuracy and F1-Score.

Training & Model Building

As we are using a subset of an imbalanced dataset, a 5 cross-fold validation was implemented to supplement the training phase and tackle the data issue. Furthermore, with a large number of independent variables, complications could arise due to complex interactions. Therefore, to aid in our model building in both logistic regression and random forest, we have selected the positive/negative variables with a score > 0.5 from the correlation matrix as follows:

1. Danceability
2. Loudness
3. Speechiness
4. Instrumentalness
5. Liveness
6. Duration_ms

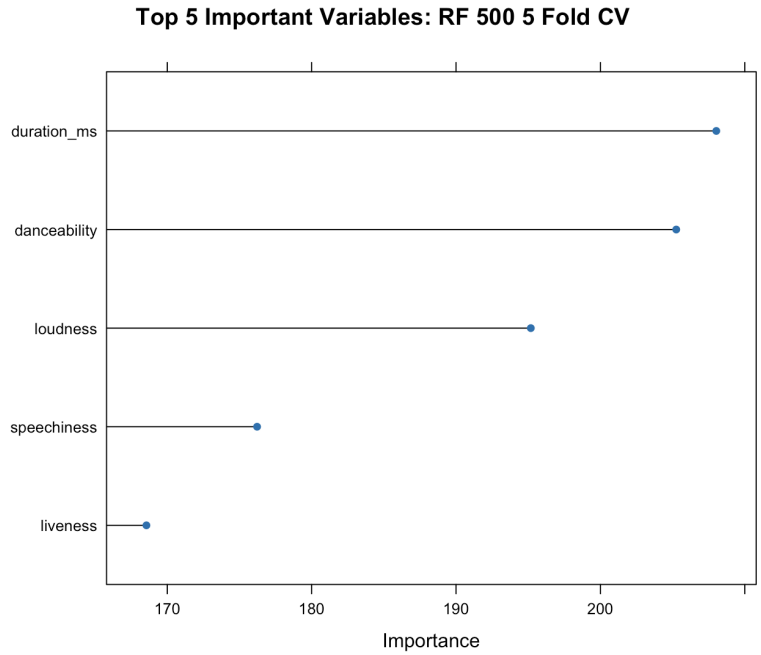

Figure 11: Top 5 Important Variables of Random Forest Model

From the random forest model, we are able to identify the top features as observed above. This indicates that the first criterion is the song duration, followed by danceability and loudness. This would conclude that popular songs should have the following characteristics:

1. Duration_ms: average duration of 186,000ms (3.1 minutes)

2. Danceability: 0.4 to 0.8
3. Loudness: -20 to -0db

```
> finalPerformance
     Performance                    LR                  RF500
1       Accuracy                0.8344                 0.8233
2         95% CI ( 0.8108 ,  0.8562 ) ( 0.7991 ,  0.8456 )
3          Kappa                     0                 0.0472
4             F1                0.9097                 0.9021
5            MSE                  <NA>                   <NA>
6            MAE                  <NA>                   <NA>
7 Precision Rate                     1                 0.9755
8    Recall Rate                0.8344                 0.8389
```

Figure 12: Final Performance Comparison Matrix Between LR and RF500

For our evaluation, we will be placing emphasis on Accuracy and F1-score, as mentioned prior. This is crucial because the model maximises overall prediction through accuracy. High accuracy, however, could be deceptive and ineffective for predicting minority classes. Here, the F1-Score begins to show its effectiveness. As it takes precision and recall into consideration, the F-score is more informative and recognises class imbalance when there is a significant difference in class frequencies.

From the performance evaluation above, Logistic Regression is observed to perform better in both accuracy 83.44% and F1- score 90.97% as compared to random forest of 82.33% and 90.21%. Therefore, Logistic Regression will be our model of choice.

2. PCA for Dimensionality Reduction

PCA (Principal Component Analysis) was employed in this data analytic project to reduce dimensionality in the dataset. The dataset consists of numerous audio features, and using all of them in our analysis could lead to the curse of dimensionality, resulting in increased computational complexity and potential overfitting. PCA helps us overcome this challenge by identifying the most critical components that capture the majority of the variance in the audio features.

Following the application of PCA, we obtained the eigenvalues for each of the 13 principal components. Here are the resulting eigenvalues:

- PC1 → 3.25
- PC2 → 1.59
- PC3 → 1.34
- PC4 → 1.15
- PC5 → 0.95
- PC6 → 0.91
- PC7 → 0.86
- PC8 → 0.77
- PC9 → 0.73
- PC10 → 0.64

- PC11 → 0.43
- PC12 → 0.26
- PC13 → 0.13

The eigenvalues resulting from Principal Component Analysis (PCA) offer valuable information about the individual contributions of each principal component in representing the dataset's overall variance. The eigenvalues range from 3.25 for PC1, indicating its prominent role in explaining patterns and structure within the data and capturing a substantial portion of the total variance, to 0.13 for PC13, signifying its minimal impact on the overall variance. These eigenvalues are crucial for dimensionality reduction, feature selection, and better comprehension of the data through PCA.

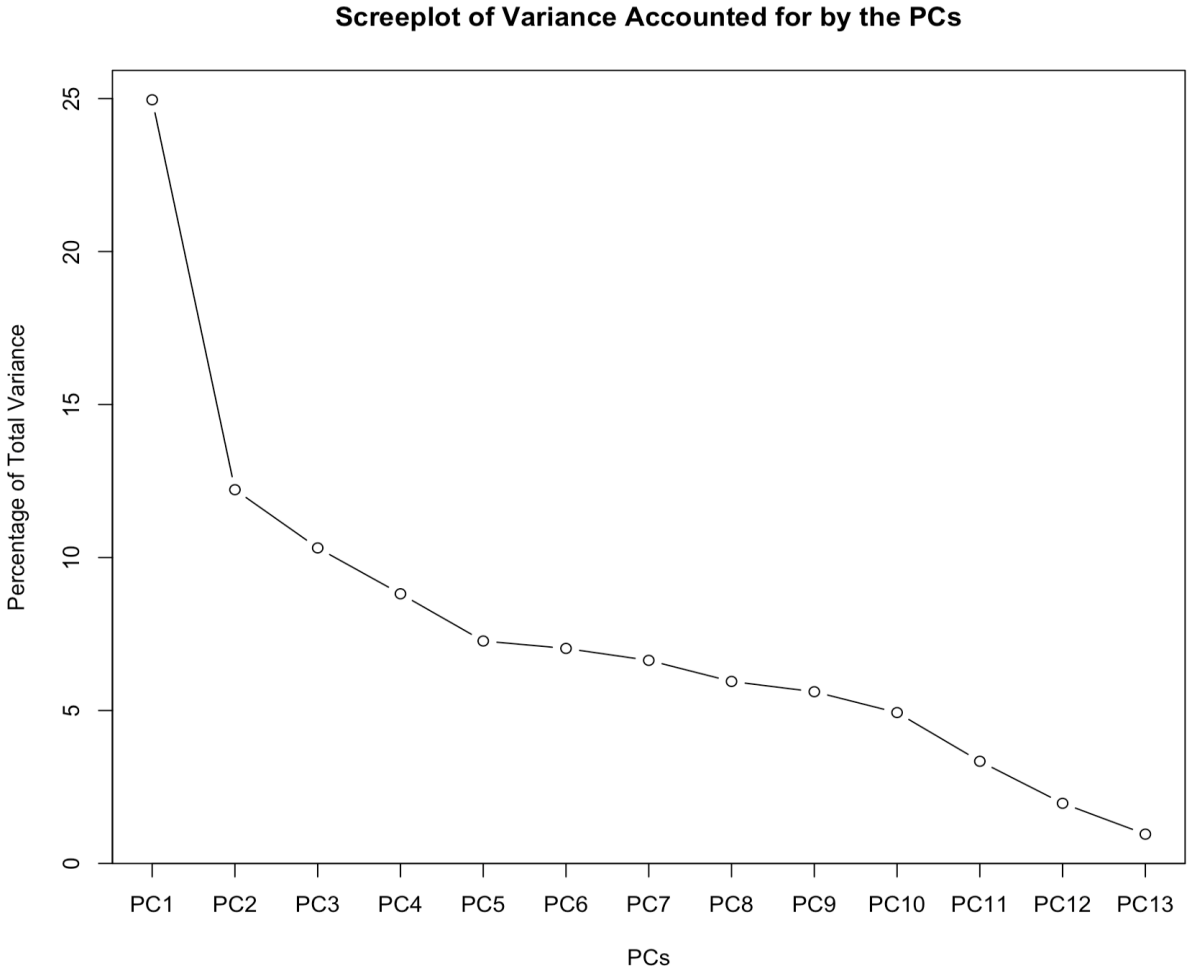

Figure 13: Variance Explained by Principal Components (Scree Plot)

Utilizing the elbow method for principal component selection, we examined Figure X to determine the number of PCs to retain. The eigenvalues plot displays a gradual decrease, with a clear inflection or "elbow" point visible. After analyzing the graph, it becomes evident that the elbow point occurs following PC5. As a result, we will opt to retain PC1, PC2, PC3, PC4, and PC5 as the principal components that effectively capture a significant portion of the overall variance in the dataset, as indicated by the elbow method. These selected principal components will serve as essential representatives of the original features, facilitating dimensionality reduction while preserving the crucial patterns in the data.

In contrast, if our aim is to account for 70% of the total variance, a modified selection of principal components would involve retaining only PC1, PC2, and PC3, while excluding PC4 and PC5. These three principal components individually account for 32.5%, 15.9%, and 13.4% of the variance, respectively, summing up to a cumulative variance of 61.8%. By choosing PC1, PC2, and PC3, we achieve a substantial reduction in dimensionality while still capturing a significant portion of the dataset's variability, precisely aligning with the specific objective of accounting for 70% of the total variation. This approach allows us to focus on the most informative components that collectively represent a substantial portion of the original data, facilitating efficient data analysis while maintaining essential patterns and structures.

Based on our analysis, we have decided to proceed with PC1, PC2, and PC3, as they collectively account for 61.8% of the total variance in the dataset. This selection aligns perfectly with our goal of maximizing information retention and maintaining the key characteristics of the dataset for further analysis and interpretation.

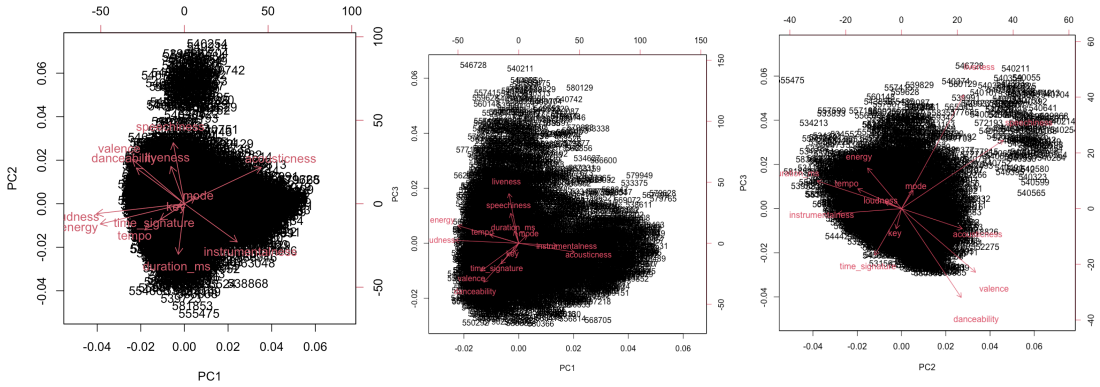

Figure 14: Biplots for PC1 <> PC2 <> PC3

The biplot analysis has provided valuable insights into the relationships between variables and the principal components (PC1, PC2, and PC3). In all three biplots, interesting groupings of variables with similar loading patterns have been observed. Notably, the variables 'Valence' and 'Danceability' consistently appear together on all three biplots, indicating that they share similar patterns of variation and contribute significantly to PC1, PC2, and PC3. Similarly, 'Speechiness' and 'Liveness' also exhibit consistent grouping across the biplots, suggesting their strong association with each other and their joint contribution to the principal components.

These consistent groupings of variables demonstrate that certain audio features tend to co-vary and are closely related in their impact on the underlying structure of the data captured by PC1, PC2, and PC3. Such findings can be instrumental in understanding the inherent structure of the dataset and the underlying factors influencing the audio features.

Another noteworthy observation from all three biplots is that no audio feature stands independently on its own. In other words, every audio feature is influenced by multiple principal components to varying degrees. This interconnectedness among the variables highlights the complex nature of the dataset and reinforces the importance of considering multiple principal components to comprehensively capture the variance and relationships within the data.

```
> loadings_pc1
    danceability           energy              key         loudness             mode      speechiness     acousticness instrumentalness
      -0.27268855      -0.47625773      -0.03870972      -0.49691278       0.06244002      -0.06332371       0.43560841       0.29883654
        liveness          valence            tempo      duration_ms   time_signature
      -0.07619032      -0.29482228      -0.22517077      -0.03473915      -0.13710707
> loadings_pc2
    danceability           energy              key         loudness             mode      speechiness     acousticness instrumentalness
       0.28961667      -0.16485018      -0.02559364      -0.08072574       0.05634368       0.49439565       0.29554128      -0.30976953
        liveness          valence            tempo      duration_ms   time_signature
       0.30243430       0.35922343      -0.21247536      -0.41024311      -0.13007030
> loadings_pc3
    danceability           energy              key         loudness             mode      speechiness     acousticness instrumentalness
      -0.47026975       0.21415436      -0.10705352       0.03534052       0.09592231       0.36225218      -0.10689818      -0.02641908
        liveness          valence            tempo      duration_ms   time_signature
       0.60090839      -0.33658974       0.10482945       0.14654610      -0.24672541
```

Figure 15: List of Loading Variables For Each PCs

In Figure X, the top loading features for each of the three principal components (PC1, PC2, and PC3) have been identified. Notably, loudness emerges as the most influential feature for PC1, indicating that variations in loudness contribute significantly to the overall variability captured by this component. For PC2, speechiness stands out as the dominant factor, implying that changes in speech-like elements play a crucial role in shaping the patterns represented by PC2. Moving on to PC3, liveness emerges as the top loading feature, suggesting that variations in the perception of live recordings strongly influence the variability observed in PC3. These findings complement the insights obtained from the biplot analysis, reinforcing the importance of these three audio features as key determinants in our dataset. By understanding and leveraging the impact of these dominant factors, we can gain deeper insights into the underlying patterns and characteristics of the audio data.

3. Clustering Song Based on Audio Features

In order to identify groups of songs that share similar audio characteristics, clustering techniques like K-means or hierarchical clustering can be used to group the songs based on their audio features. We can then analyze the various clusters to understand the distinct musical patterns or genres that emerge.

In this section, we will be employing the hierarchical clustering technique as it can handle both categorical and numeric data types and is adaptable to a wide range of datasets. Furthermore, it is based on similarity measurements between data points, it is often less susceptible to outliers than other clustering algorithms like k-means, and most importantly, it is able to handle relatively large datasets.

```
> table(clusters, data$verdict)

clusters  low popular
       1 3994     777
       2  449      30
       3   98      24
```

Figure 16: Confusion matrix of popularity's verdict based on the first three splits for hierarchical clustering of audio features using complete method

After clustering the dataset into 3 groups, the clusters are observed to be ordered by their group size, with Cluster 1 having over 3994 low popularity and 777 popular tracks.

```
> cluster_profiles
  cluster        X artist_name track_name track_id popularity year genre danceability    energy
1       1 556795.5          NA         NA       NA   31.29763 2022    NA    0.5733991 0.6970635
2       2 556450.2          NA         NA       NA   27.15658 2022    NA    0.3937601 0.1835437
3       3 566708.5          NA         NA       NA   38.14754 2022    NA    0.2433836 0.0313582
       key   loudness      mode speechiness acousticness instrumentalness  liveness   valence
1 5.317124  -7.200476 0.6120310  0.10298231    0.2471842        0.1954734 0.2208359 0.4539654
2 5.098121 -21.084672 0.6680585  0.10379165    0.8832520        0.6574395 0.2125294 0.2162759
3 4.811475 -33.966451 0.7786885  0.05039344    0.9541393        0.8433171 0.1160172 0.1048180
      tempo duration_ms time_signature verdict popularity_score
1 123.88879    217981.2       3.917208      NA         31.29763
2 104.32210    199792.6       3.705637      NA         27.15658
3  93.80618    143836.1       3.754098      NA         38.14754
```

Figure 17: Cluster Profile for All Three Clusters

Analysing the Spotify subset data of 2022 as a whole, it is observed that the following characteristics are observed:

Cluster 1

- Highest danceabilty of 0.57
- Highest energy of 0.69
- Highest loudness of -7.2db
- Lowest acousticness of 0.25
- Lowest instrumentalness of 0.20
- Highest valence of 0.45
- Highest tempo of 123.9
- Longest duration_ms of 217,981.2

Cluster 2

- Lowest popularity of 27.2

Cluster 3

- Highest popularity of 31.3
- Lowest danceability of 0.24
- Lowest energy of 0.03
- Lowest key of 4.81
- Lowest tone of -34.0
- Lowest Speechiness of 0.05

- Highest acousticness of 0.95
- Highest instrumentalness of 0.84
- Lowest liveness of 0.12
- Lowest Valence of 0.10
- Slowest tempo of 93.81
- Shortest duration of 143,836.1

Conclusion

In summary, this project analysed a 10% subset of Spotify 2022 audio tracks to acquire valuable insights regarding the qualities and popularity of music tracks on the Spotify platform. From the analysis, we are able to identify trends in the popularity of a track and learn how certain characteristics affect Spotify listeners' reactions.

The association between numerous features and track popularity was another intriguing conclusion of our investigation. While no single characteristic can fully explain a track's appeal, we found trends that suggested a track's **duration**, **danceability**, and **loudness** played significant roles in a track's reception. Spotify may concentrate on promoting high-potential music and increase those tunes' chances of economic success by precisely identifying tracks.

Future revisions and implementation of enhancements are probable. The most immediate improvement would be to use greater amounts of data, as only approximately 0.5 percent of the available data was used due to the limitations of the device's processing power. This may result in the deletion of some features or the failure to recognise underlying trends. Time series analytics may also be used to identify seasonal data trends, which may be necessary for making decisions.

The other revision would be to perform clustering analysis on a larger dataset of "popular" tracks. This is useful as the characteristics of this particular subset could allow Spotify to further unravel the mystery of successful tracks and provide an addition avenue of consultancy service or feedback to production studios and/or artists.

APPENDIX A – Details of Dataset

Link: https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks

Music is part of many people everyday life, and the musical industry has been growing consistently. With the rise of technology and streaming platforms, the amount of data generated and collected grows exponentially. These raw data harness untapped potential, which has the ability to identify trends and predict a song's popularity before it is released. The Kaggle dataset extracted from Spotify Music via APIs comprises of approximately 1 million song tracks from the period of 2000 to 2023, with 19 different audio features. Using machine learning/deep learning methods, unique insights could be unravelled, and prediction models could be built.

The dataset contains the following variables:

Audio (Features)	**Description**
Popularity	How popular a track is from 0 to 100
Year	Year of track release from 2000 to 2023
Danceability	Suitabilty to be used as a dance track
Energy	Intensity and activity level of track
Key	Central note of track (-1 to -11)
Loudness	Volume of track in decibles (-60 to 0 db)
Mode	Modality of track (1: Major; 0: Minor)
Speechiness	Existence of lyrics (words) in track
Acousticness	Measure of acousticness of track (0 to 1)
Instrumentalness	Degree of vocals in track (0 to 1.0)
Liveness	Presence of audience in track (0 to 1.0)
Valence	Measure of positiveness conveyed by track
Tempo	The rhythm of the track in beats per minute (BPM)
Time_signature	Estimated time signature (rhythmic structure) from 3 to 7
Duration_ms	Duration of the track in milliseconds