Spotify Tracks Dataset Analysis

by

Vrushali Khatane, Achintya, Shravan Doda

FINAL REPORT

Course Professor: **Prof. Hong Do**

STEVENS INSTITUTE OF TECHNOLOGY
Castle Point on Hudson
Hoboken, NJ 07030

2024

Contents

1

Introduction

Music is a fundamental part of every human culture. Streaming platforms such as Spotify have become essential for accessing and enjoying music. For this project, we analyze a subset of the Tracks dataset from Spotify. By examining this data, we aim to uncover valuable insights that can enhance user experiences on the platform and provide a deeper understanding of current trends in the music industry.

1.1 Dataset Description

The dataset used in this project originates from the Spotify Tracks dataset available on Kaggle. The original dataset consists of around 90,000 unique tracks with over 100 different genres. However, to keep this analysis manageable, we pre-processed the data to select the 10 most popular genre. Additionally, we narrowed down the analysis to only six key features:

- Popularity (Numerical)
- Valence (Numerical)
- Loudness (Numerical)
- Acousticness (Numerical)
- Explicit (Categorical)
- Track Genre (Categorical)

1.2 Goal

The primary goal of this project is to predict the popularity of a track based on its musical characteristics and genre. Popularity serves as a key metric for gauging a song's success, and understanding its drivers can provide valuable insights for both listeners and industry stakeholders. Specifically, this study seeks to:

1. **Explore descriptive statistics:** Summarize and visualize the distribution of audio features associated with tracks, providing an overview of key characteristics within the dataset.
2. **Identify relationships:** Analyze the interplay between different audio features, such as how valence or acousticness relates to popularity.
3. **Develop predictive models:** Build and evaluate models to predict track popularity using the selected audio features.

2

Methods

Our analysis begins with **descriptive statistics**, providing a comprehensive overview of the dataset. This includes summary statistics for both numerical and categorical variables. We will then present data visualizations for univariate analysis, incorporating histograms and kernel density estimates (KDE) for numerical features to examine distributions, box plots to identify potential outliers, and QQ plots to assess normality. For categorical features such as Track Genre and Explicit, bar charts will be used to visualize frequency distributions. Additionally, we will compute sample means for all numerical features and verify their normality using the Central Limit Theorem (CLT).

Next, we proceed to **inferential statistics**, focusing on relationships between pairs of variables through bivariate analysis. Scatter plots will be employed to explore potential associations between numerical variables, while box plots will illustrate the distribution of Popularity (target variable) across categories of Explicit and Track Genre. A correlation heatmap will provide a broader view of the interdependence among numerical features. Building on these explorations, hypothesis tests will be conducted to assess the statistical significance of our findings.

Finally, we will implement regression models to predict Popularity using the other audio features as predictors. The performance of these models will be evaluated using appropriate metrics to ensure their reliability and accuracy.

3

Statistical Analysis

3.1 Descriptive Analysis

Statistic	Valence	Acousticness	Loudness	Popularity
Count	7279	7279	7279	7279
Mean	0.498	0.369	-8.772	26.820
Std	0.237	0.346	5.733	18.218
Min	0.000	0.000	-41.531	0.000
25%	0.315	0.045	-10.260	16.000
50%	0.497	0.255	-7.102	23.000
75%	0.683	0.685	-5.105	31.000
Max	1.000	1.000	1.023	100.000

Table 3.1: Summary Statistics for Valence, Acousticness, Loudness, and Popularity

The Table 3.1 above presents the summary statistics for the four numerical features in the dataset. A key observation is the variation in scales among these features. For instance, both Valence and Acousticness are bounded within the range $[0, 1]$, with mean values of 0.498 and 0.369, respectively. In contrast, Popularity spans a broader range of $[0, 100]$, with a mean of 26.82. Meanwhile, Loudness exhibits a mean value of -8.772. The variations in feature scales suggests that we may have to standardize the data before we implement our regression models.

Indian	Metal	Classical	Pop	Blues	Hip-hop	Reggae	Country	Rock	Jazz
933	866	814	776	764	745	634	630	608	509

Table 3.2: Count of Tracks by Genre (Genres as Columns)

Explicit Content	Count
False	6723
True	556

Table 3.3: Count of Tracks by Explicit Content Status

For the categorical features, Table 3.2 summarizes the distribution of tracks across different genres. The distribution appears relatively balanced, with no single genre dominating the dataset. The `indian` genre contains the highest number of tracks, while the `jazz` genre has the fewest. Similarly Table 3.3 illustrates that the dataset includes $6,723$ non-explicit tracks compared to only 553 explicit tracks, indicating skewnesses for this feature. Figure 3.1 illustrates the data distribution for the two categorical features.

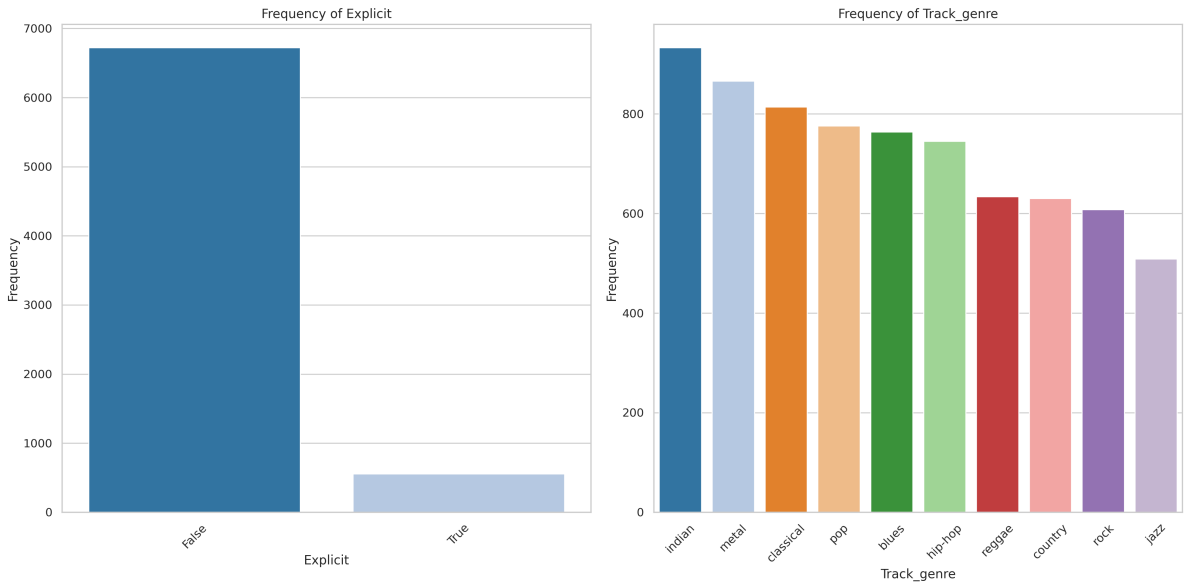


Figure 3.1: Data distribution - Explicit and Track Genre.

The distribution of numerical features will give us insights as well as help us decide the type of hypothesis test that we should choose for our inferential statistics part. Figure 3.2 illustrates these distributions using Histogram and KDE plots.

The distribution of Popularity is right-skewed with majority of the tracks having lower Popularity values clustered around 20–30. There are very few tracks at the higher end of the spectrum, near 100, indicating that most tracks in the dataset are not highly popular. The Valence feature, which measures the "happiness" of a track, exhibits a roughly uniform distribution. This suggests that tracks in the dataset are evenly spread across the entire valence range from 0 to 1, without any concentration to any specific range. The Acousticness is highly skewed to the left, with a peak near zero. This indicates that a large proportion of tracks have low acousticness, meaning they are less acoustic in nature. A small number of tracks also have high Acousticness values, nearing 1. This suggests that the values are concentrated at the ends. The Loudness feature shows a distribution that is shifted to the left. The peak is around -8 to -10 dB, and the values spread over a range between -42 dB and 1 dB.

These observations highlight the diverse characteristics of the dataset. There is skewness in some features, particularly Popularity and Acousticness, and overall none of the features appear to be normally distributed. Popularity and Loudness may be approximately normal but we'll have to test that further.

Figure 3.3 provides a visual summary and highlights the presence of potential outliers in the four numerical features.
The boxplot for Popularity highlights a large number of outliers on the right end of the distribution. These outliers represent tracks that are highly popular and exceed the upper whisker

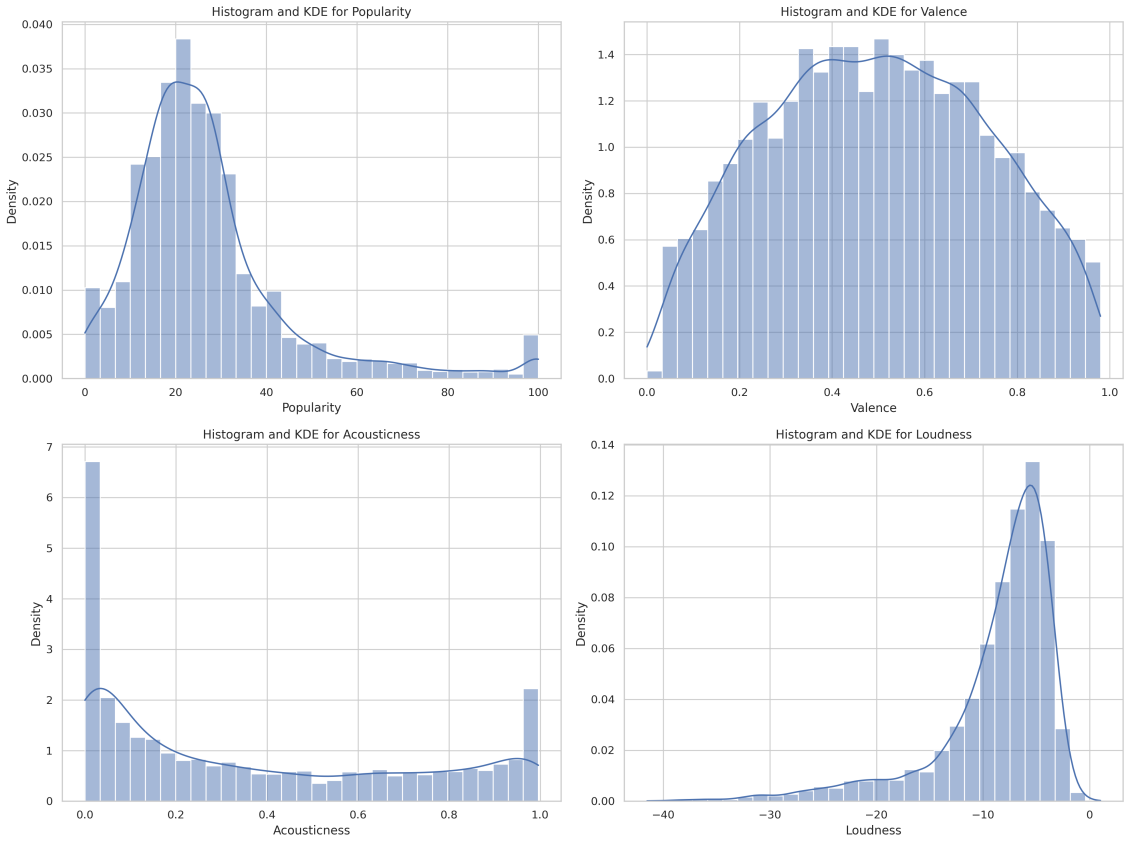


Figure 3.2: Histogram and KDE plots for numerical features

defined as 1.5 times the interquartile range (IQR) above the third quartile. This observation confirms the right-skewness noted in the histogram and KDE plot earlier, where most tracks have Popularity values concentrated between 20 and 30, with only a few achieving high Popularity. Both Valence and Acousticness exhibit no visible outliers, as all data points lie within the whiskers. This aligns with the earlier findings: a relatively uniform distribution for Valence and a broader, evenly spread range for Acousticness. The boxplot for Loudness reveals a substantial number of outliers on the lower end of the distribution. These correspond to tracks with extremely low loudness levels, consistent with the left-skewed distribution observed in the histograms and KDE plot. These outliers indicate the presence of a small subset of tracks that are significantly quieter than the majority.

The presence of outliers in Popularity and Loudness could influence subsequent analyses, particularly regression modeling. It may be necessary to perform transformations or deal with outliers to mitigate their impact on the results.

To examine the normality of the distribution of features we use the QQ plots illustrated in Figure 3.4. The QQ (Quantile-Quantile) plots evaluate the degree to which the data for the variables — Popularity, Valence, Acousticness, and Loudness — follows a normal distribution. The theoretical quantiles (x-axis) are compared against the sample quantiles (y-axis). Any significant deviation from the diagonal reference line indicates departures from normality. To statistically test the null hypothesis H_0 that the data is normally distributed, the Shapiro-Wilk test was performed alongside the visual analysis. A p-value less than 0.05 indicates a rejection of the null hypothesis H_0.

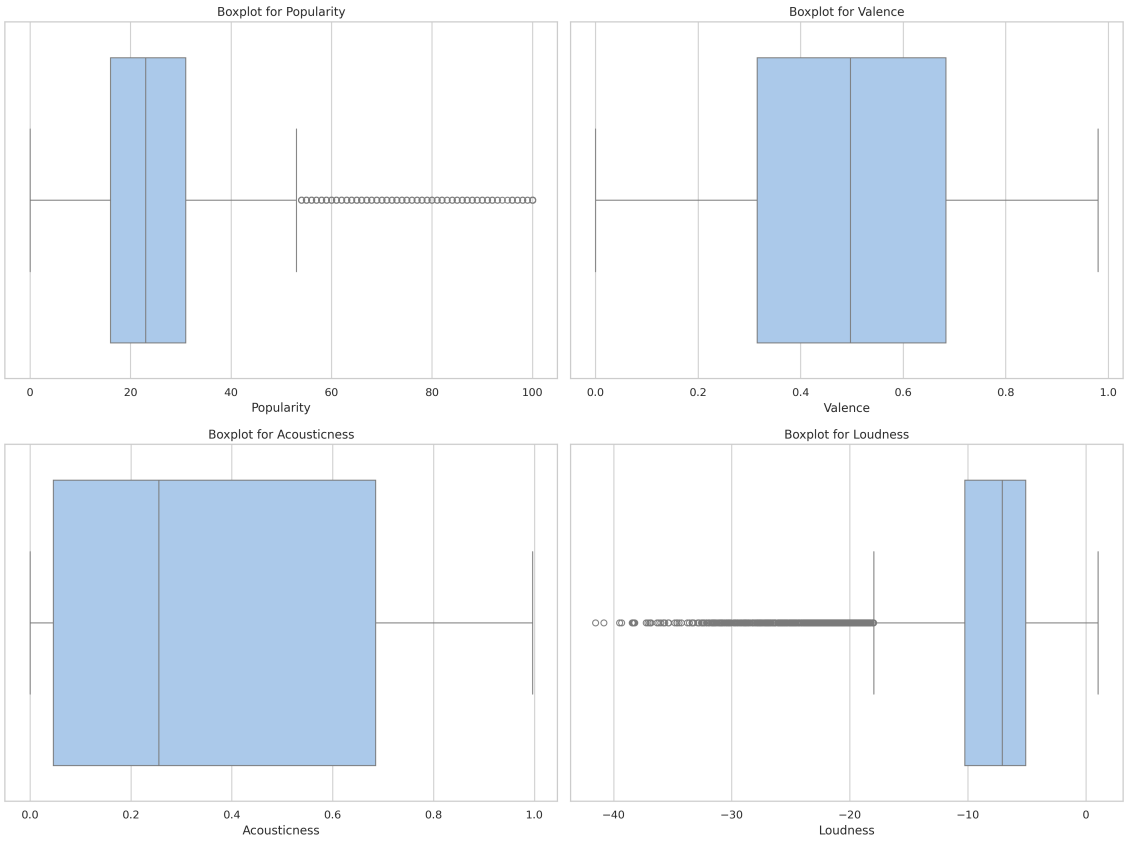


Figure 3.3: Box plots for numerical features

For Popularity, the data exhibits severe deviations from normality, with a divergence from the reference line at both ends. The pattern suggests heavy tails and the p-value of 0 indicates that the H_0 of normality is strongly rejected. For Valence, although the data is closer to the reference line in the central portion, significant departures occur at the tails. This suggests a departure from the normal distribution at the extreme values. The p-value is also effectively zero, providing strong evidence against H_0. For Acousticness, the data closely follows the reference line in the central region but diverges slightly at the extremes. This pattern suggests a slight departure from normality and with a p-value of 0, H_0 is rejected. For Loudness, there are strong deviations from the reference line, particularly in the tails. The heavy divergence at both extremes indicates a distribution far from normality, potentially with heavy tails or outliers. The p-value of 0 indicates H_0 is rejected.

All four variables show significant departure from normal distribution, as indicated by the QQ plots and the p-values. We'll have to adjust our approach for hypothesis testing and use non-parametric methods.

To get insights about the central tendencies and distributions of the features, we generated sampling distributions of the sample means for each feature and compared them against the theoretical normal distribution as illustrated in Figure 3.5a and Figure 3.5b.
The sampling distribution of the sample mean for Popularity appears approximately symmetric, closely following the theoretical normal distribution curve. This suggests that the sample means adhere to the Central Limit Theorem (CLT), as the distribution approaches normality, even if the underlying population may not be perfectly normal. Similarly, the sampling distribution of Valence demonstrates a good fit to the normal distribution. The histogram shows minimal

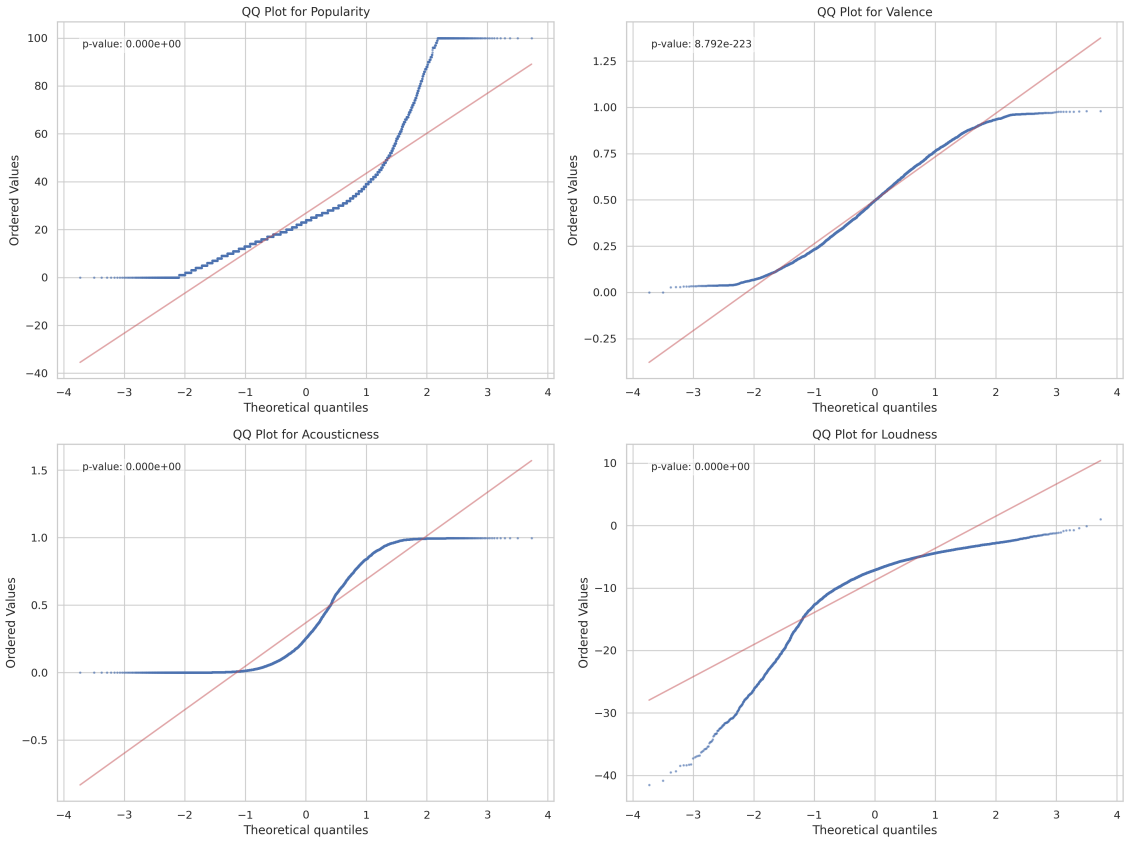
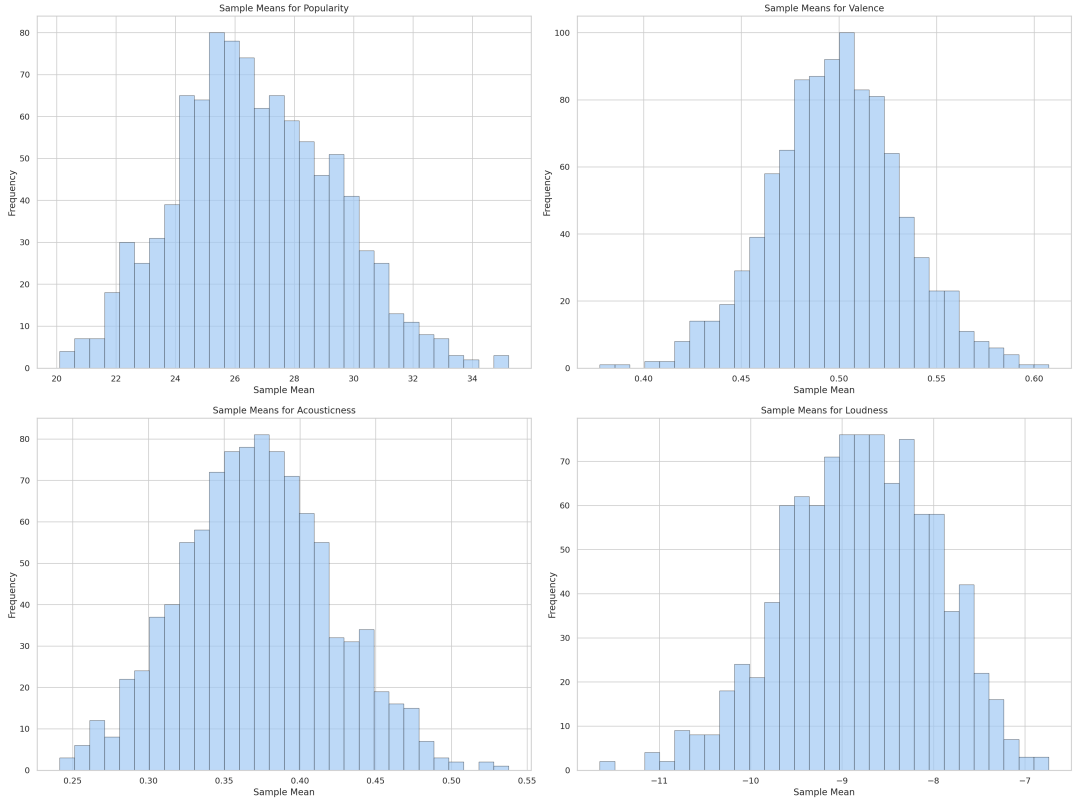
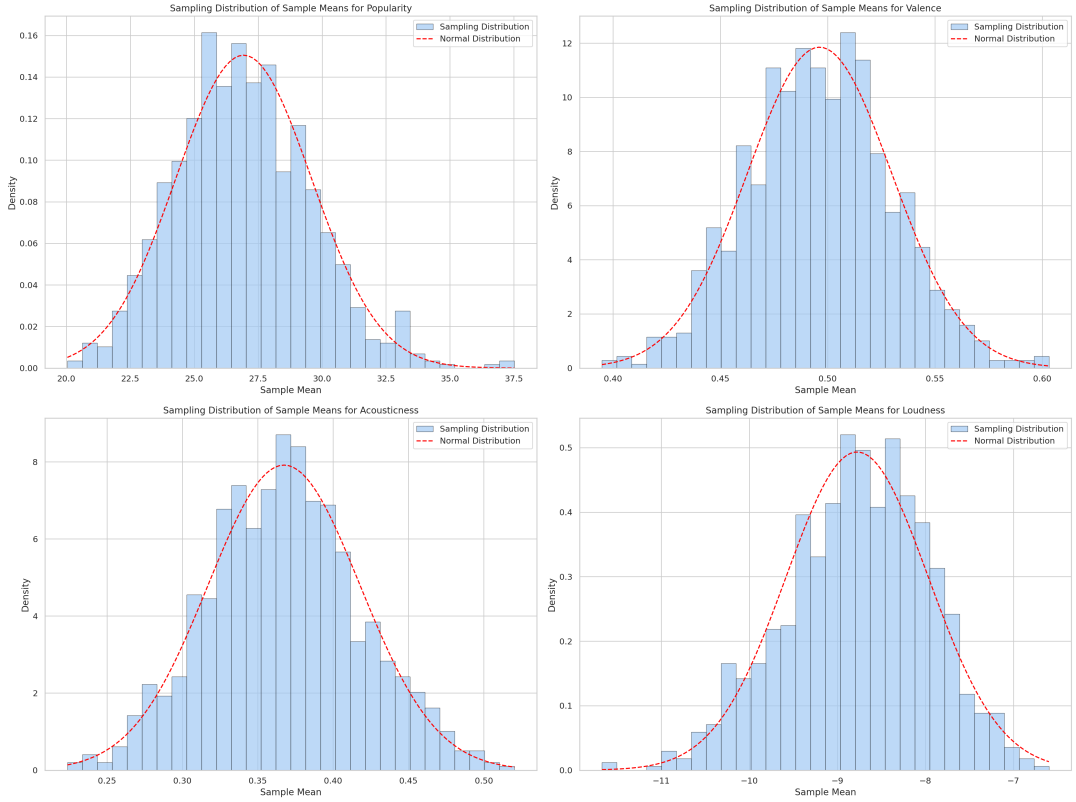


Figure 3.4: Quantile-Quantile plots for numerical features

skewness and the normal curve aligns well with the majority of the data points which is consistent with the CLT. The sampling distributions for Acousticness and Loudness show a slight asymmetry and mild deviations from the normal distribution at the tails. Despite this, the overall shape approximates a bell curve, suggesting that the sample means are largely normal but may be influenced by non-normality in the underlying data.
These histograms demonstrate a strong alignment with their respective normal distribution curves, which validates the theoretical expectation of the CLT that sample means will tend towards a normal distribution as sample sizes increase. The close fit suggests that the sample sizes are adequate to assume normality in the sampling distributions



(a) Sample means of numerical features (Histogram Plots)



(b) Sampling distribution for numerical features (Normal by CLT)

Figure 3.5: Sample means and Sampling distributions for numerical features.

3.2 Inferential Analysis

In this section we will focus on exploring and identifying relationships between variables and testing hypotheses. To start with, we explore the relationship between the numerical features.

3.2.1 Scatter Plots and Correlation Heatmap

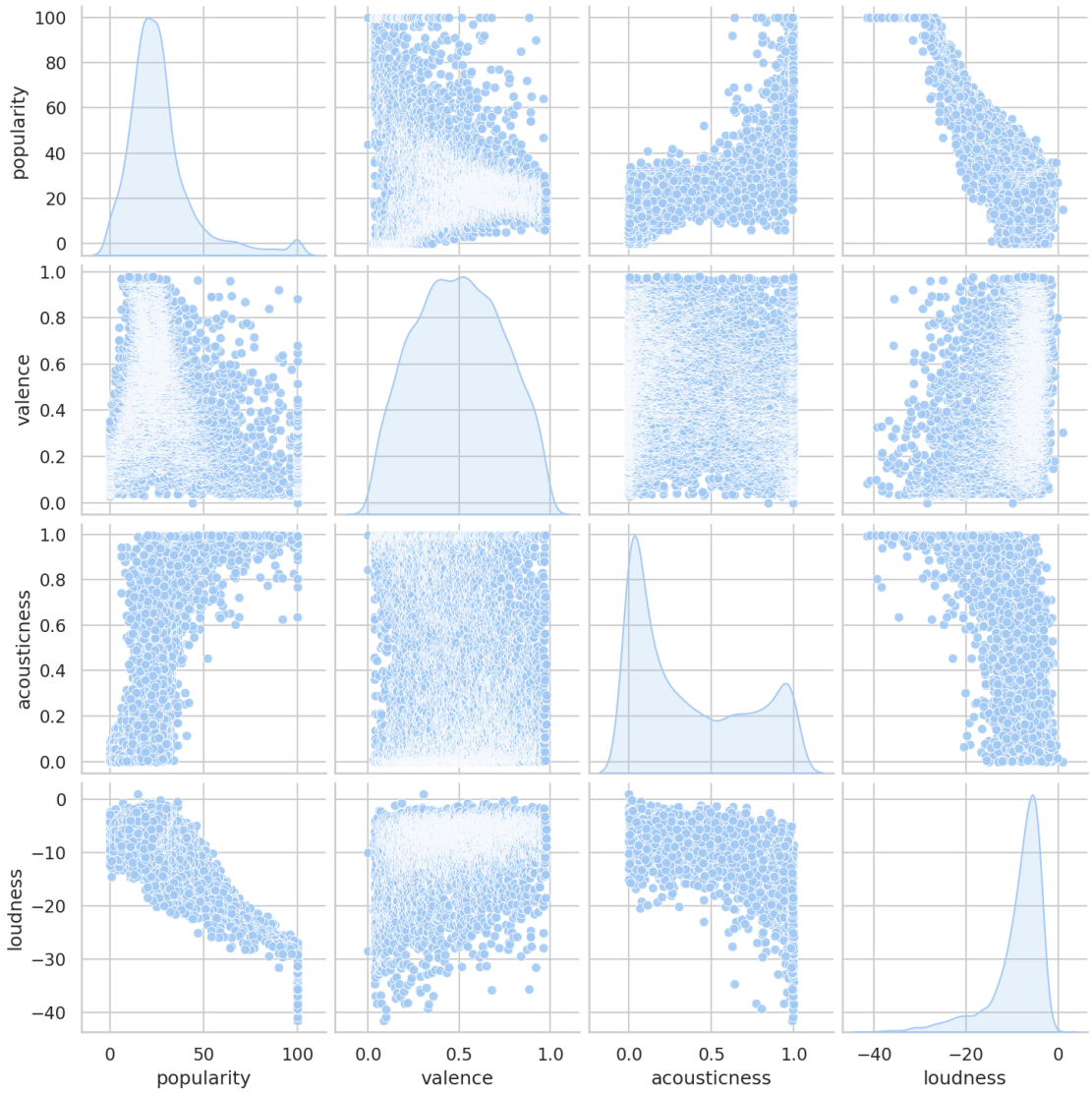


Figure 3.6: Pairwise scatter plots

Figure 3.6 presents pairwise scatter plots with density plots along the diagonal, highlighting the relationships between various numerical features. Each scatter plot compares two variables, illustrating the strength and direction of their relationship, while the density plots show the distribution of individual variables. This visualization allows for the identification of patterns and potential correlations between different numerical features. Scatter plots with upward trends suggest a positive correlation between the variables, indicating that as one variable increases, the other tends to increase as well. In contrast, downward trends suggest a negative correlation. From the plot, it can be observed that Popularity and Loudness appear to have a negative correlation, meaning that as Loudness increases, Popularity tends to decrease. On the other

hand, Popularity and Acousticness exhibit an upward trend, suggesting a Positive correlation. However, Popularity and Valence appear to have little to no correlation, as the scatter plot for these features shows points scattered randomly without a clear trend. These observations are further supported by the correlation heatmap shown in Figure 3.7.

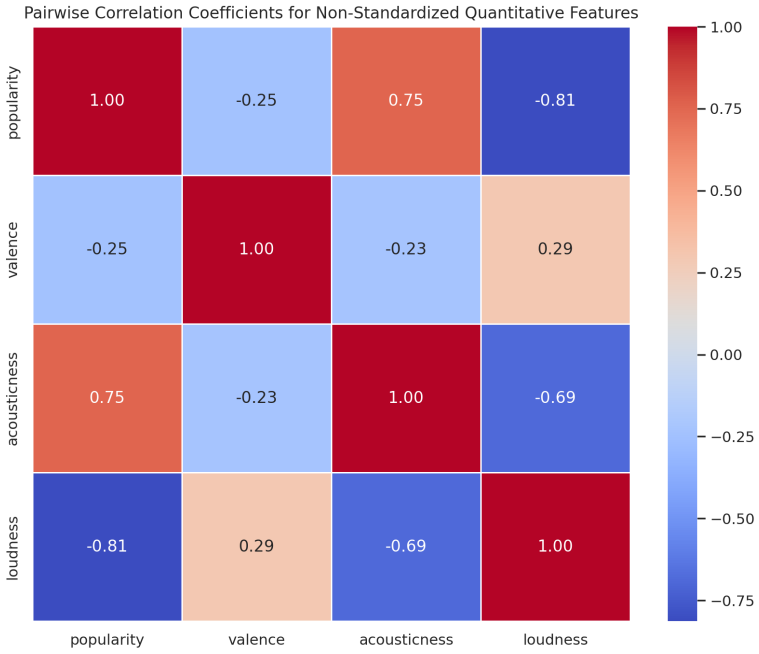


Figure 3.7: Correlation Heatmap for fumerical features

The heatmap in Figure 3.7 illustrates the correlation coefficients between numerical features. The color intensity reflects the strength of the relationship, with red indicating a higher correlation. The heatmap points out to a strong correlation coefficient of 0.75 between Popularity and Acousticness, indicating that more acoustic songs tend to be more popular. It also points out a strong negative correlation coefficient of -0.81 between Popularity and Loudness. This indicates that songs that are less loud and more acoustic tend to be more popular.

3.2.2 Hypothesis Testing

In order to analyse if there is a relationship between the Popularity (target variable) and the categorical features, we refer to the Figure 3.8 that illustrate line charts of Popularity means and standard deviations for explicit and non-explicit tracks across various genres. The charts hint at the fact that Popularity may not be similarly distributed across different genres. To get a statistically signification result, we perform a few hypothesis tests below.

Popularity across Genres: Given that our data does not follow a normal distribution, we will employ non-parametric methods to determine whether the distributions of Popularity vary across genres. Specifically, we will use the **Kruskal-Wallis** test. This test does not assume any particular distribution for the data, making it suitable for our analysis. For the hypothesis testing, we define the following terminology: let F_i represent the distribution of Popularity across the i^{th} genre. We can then state the null and alternative hypotheses as follows:

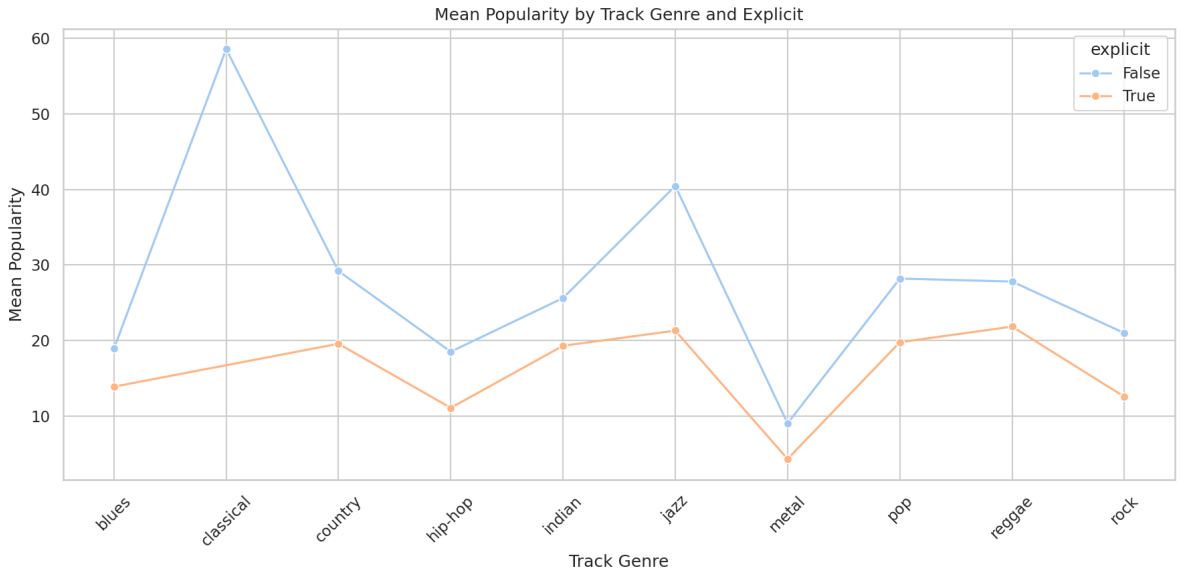


Figure 3.8: Line Chart for Popularity means across categorical variables

$$H_0 : F_1 = F_2 = \cdots = F_n$$

$$H_1 : \text{Not } H_0$$

The Kruskal-Wallis test statistic, H, is calculated using the following formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} J_i(\bar{R_i^2}) - 3(N+1)$$

where: N is the total number of observations across all groups, J_i is the number of observations in group i, R_i is the sum of ranks for group i, k is the number of groups (in this case, the number of genres). The value of H is then compared to a chi-square distribution with $k-1$ degrees of freedom to determine whether to reject the null hypothesis.

For our case, the calculated H-statistic is 4245.986, and the p-value is 0.0. With a significance level of $\alpha = 0.05$, we can **reject** H_0 since the **p-value < α**. This indicates that **Popularity varies with Genre**. Therefore, it is necessary to include Genre as a predictor when building a regression model for predicting Popularity.

Popularity vs Explicit/Not Explicit: Given that our data doesn't follow a normal distribution, we will use the non-parametric **Mann-Whitney** test to compare the distribution of Popularity between explicit and non-explicit songs. The Mann-Whitney test is suitable for this scenario because it does not assume any specific distribution for the data. For hypothesis testing, we define the following:
Let F represent the distribution of Popularity for explicit songs and G represent the distribution of Popularity for non-explicit songs. The null and alternative hypotheses are as follows:

$$H_0 : F = G$$

$$H_1 : F \neq G$$

The Mann-Whitney U test statistic is calculated using the following formula:

$$R^* = \min(R, R')$$

where $R' = n(m+n-1) - R$ is the U statistic for the second group (non-explicit), n and m are the sample sizes for the two groups, R is the sum of ranks for the first group (with smaller size). Once the value of U is computed, it is compared against a critical value from the Mann-Whitney distribution to determine whether to reject the H_0

For our case, the calculated R-statistic is 2980323.5, and the p-value is 0.0. With a significance level of $\alpha = 0.05$, we can **reject the** H_0 since the p**-value <** α. This indicates that **Popularity depends on whether the song is explicit**. Therefore, it is necessary to include Explicit/Non-Explicit as a predictor when building a regression model for predicting Popularity.

Genres vs Explicit/Non-Explicit: We suspect that Genre and Explicitness may be dependent as well. This implies that some genres may have higher proportions of explicit tracks compared to others. To test this statistically, we'll employ the **Chi-squared test for independence**. This will allow us to determine if there is a dependent relationship between Genre and Explicitness. Let the random variable G represent Genre, and E represent Explicitness. The null and alternative hypotheses are as follows:

$$H_0 : G \text{ and } E \text{ are independent.}$$

$$H_1 : G \text{ and } E \text{ are dependent.}$$

The Chi-squared statistic is calculated using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where: - O_{ij} is the observed frequency in the i^{th} row and j^{th} column of the contingency table, - E_{ij} is the expected frequency in the i^{th} row and j^{th} column of the contingency table, which is calculated as:

$$E_{ij} = \frac{(row\ total_i \times column\ total_j)}{total\ sample\ size}$$

The Chi-squared statistic is then compared to a critical value from the Chi-squared distribution with degrees of freedom $df = (r-1)(c-1)$, where r is the number of rows (explicit) and c is the number of columns (genres). If the calculated χ^2 value exceeds the critical value, we reject the H_0.

Explicit	Blues	Classical	Country	Hip-hop	Indian	Jazz	Metal	Pop	Reggae	Rock	Row Total
False	750	814	600	557	912	506	739	719	550	576	6563
True	14	0	30	188	21	3	127	57	84	32	566
Column Total	764	814	630	745	933	509	866	776	634	608	7129

Table 3.4: Contingency Table for Chi-Square test Between Explicitness and Track Genre

Table 3.4 shows the contingency table for Chi-squared test between Explicitness and Track Genre. The degree of freedom $df = 9$ and the χ^2 = 606.277 and the p-value comes out to be effectively 0. With a significance level of $\alpha = 0.05$, we can **reject** H_0 since the p**-value <** α. This indicates that **Explicitness and Track Genre may be dependent as well**, implying some Genres have higher proportions of explicit tracks compared to others.

4

Regression Analysis

In this section, we discuss and implement various regression techniques to predict a track's Popularity based on a number of predictor variables. We start with Multiple Linear Regression, then move on to Lasso Regression, and finally implement Support Vector Regression (SVR). To leverage categorical features for this analysis, they were label-encoded. We chose not to standardize the numerical features to keep the results interpretable.

4.1 Multiple Linear Regression

Table 4.1 shows the coefficients obtained from multiple regression table. The predictor variables were Valence, Acousticness, Explicit, Track Genre and Loudness.

Variable	Coefficient	Std. Error	p-value
intercept	23.992	0.235	0
valence	0.005	0.125	0.681
acousticness	6.845	0.166	0
explicit	-4.983	0.458	3.234e-27
track_genre	0.747	0.046	1.577e-57
loudness	-10.603	0.171	0

Table 4.1: Multiple Regression Coefficients

The multiple regression model reveals interesting insights about the relationship between different predictor variables and Popularity. The intercept is statistically significant and has a positive value. The Valence has a very small positive relationship with Popularity. A unit increase in Valence leads to a 0.005 increase in Popularity. A high p-value (0.681) indicates that the effect of valence is not statistically significant. Acousticness has a strong positive relationship with Popularity and is statistically significant as well (p-value > 0.05). A unit increase in Acousticness leads to a 6.845 increase in Popularity. Explicit has a strong negative relationship with popularity. Tracks marked as explicit reduce Popularity by -4.983 on average. Loudness has a strong negative relationship with Popularity. A unit increase in loudness reduces Popularity by -10.603. These observations align with with prior analysis of scatter plots and correlation matrix. Tracks that are more acoustic, less loud, and are not explicit tend to be more popular as compared to others.

4.2 Lasso Regression

We also implemented Lasso Regression. Lasso applies L1 regularization, which helps in feature selection by shrinking some coefficients to zero, effectively removing less relevant predictors.

The coefficients are shown in Table 4.2

Variable	Coefficient
const	24.041
valence	0
acousticness	6.835
explicit	-3.585
track_genre	0.708
loudness	-10.509

Table 4.2: Lasso Regression Coefficients

Lasso regression eliminates valence by setting its coefficient to zero, highlighting its irrelevance. This contrasts with the multiple regression model, where valence had a small coefficient but a non-significant p-value. For other predictors, Lasso regression coefficients are slightly reduced compared to the multiple regression model. This reflects the penalization imposed by Lasso, which shrinks coefficients to prevent overfitting.

4.3 Support Vector Regression

We also implemented a SVR with a linear kernel to predict Popularity as a function of the predictor variables. The coefficients from the SVR model are illustrated in Table 4.3

Variable	Coefficient
const	23.710
valence	0.676
acousticness	7.965
explicit	-4.845
track_genre	0.687
loudness	-7.724

Table 4.3: SVR Coefficients

The intercept is 23.7105, representing the baseline Popularity when all predictors are at their default (zero) values. The coefficient for Valence is 0.676, indicating a moderate positive relationship with Popularity. Unlike Lasso and multiple regression, SVR attributes more weight to this predictor. The coefficient for Acousticness is 7.965, suggesting a strong positive relationship with Popularity. The coefficient for explicit is -4.8452, showing a significant negative impact on Popularity. Tracks marked as explicit decrease the predicted Popularity by 4.8452 units, consistent with the findings from both multiple regression and Lasso. The coefficient for Loudness is -7.72494, reflecting a strong negative relationship with Popularity, although the magnitude is less than that in the Lasso and multiple regression models.

Table 4.4 compares the Mean Squared Error (MSE) and Coefficient of Determination (R^2) for the three regressions.

Model	R^2	Mean Squared Error
Multiple Regression	0.768	82.8318
Lasso Regression	0.767	83.238
Support Vector Regression (SVR)	0.742	92.3337

Table 4.4: Regression Model Evaluation Metrics

5

Summary and Conclusion

This project analyzed the Spotify Tracks Dataset to uncover insights about the relationships between audio features and track popularity. By examining numerical features like valence, acousticness, loudness, and categorical features such as genre and explicitness, the study identified key trends. Tracks with higher acousticness and lower loudness were found to be more popular, suggesting that quieter and more acoustic songs resonate better with listeners. Hypothesis testing further validated the significant influence of genre and explicitness on track popularity, emphasizing their importance as predictors in modeling.

Regression models, including multiple linear regression, Lasso regression, and SVR were built to predict track popularity. While these models provided foundational insights, their performance metrics suggested the need for more advanced techniques to capture the intricate relationships between features. Additionally, data distributions revealed outliers and skewness, underlining the importance of preprocessing steps such as scaling and transformation to improve model accuracy.

In conclusion, this study shed light on factors influencing the popularity of tracks on Spotify and provided a foundation for predictive modeling in this domain. The results emphasized the importance of considering both numerical and categorical features in analysis and highlighted areas where improvements, such as feature engineering, dataset expansion, and advanced modeling techniques, could enhance insights and predictions. These findings have implications for understanding music trends and optimizing user experiences on streaming platforms.

6

Future Work

Building on the findings of this study, several avenues for further research and analysis are suggested:

- **Feature Expansion:**
  - Incorporate additional audio features, such as tempo, energy, and instrumentalness, to capture more nuances in track characteristics.
  - Explore metadata features, such as artist popularity or release year, to enrich the analysis.
- **Data Preprocessing:**
  - Apply data transformation techniques, such as log transformations or scaling, to address skewness and outliers.
  - Explore advanced feature engineering, including polynomial features or interaction terms, to improve model expressiveness.
- **Advanced Modeling:**
  - Experiment with more complex algorithms, such as Random Forests, Gradient Boosting, or Neural Networks, for better predictive performance.
  - Implement cross-validation to ensure robust evaluation of model performance and reduce overfitting.
- **Exploration of Unsupervised Learning:**
  - Conduct clustering analysis to identify patterns or groupings within the dataset, potentially revealing latent structures such as genre-based clusters.
- **Dataset Enrichment:**
  - Integrate external datasets, such as listener demographics or streaming behavior, for more comprehensive analyses.
  - Utilize larger and more balanced datasets to improve generalizability.
- **Real-World Application:**
  - Develop a dashboard or recommendation system prototype that leverages these insights to suggest tracks based on user preferences.
  - Tailor the analysis to address business needs, such as optimizing playlist curation or marketing strategies.

These directions will help deepen the understanding of factors influencing track popularity and pave the way for more accurate and impactful models.