Technical Report on Spotify Dataset Analysis

1. Introduction

This report focuses on the exploratory data analysis (EDA), statistical analysis, and predictive modeling conducted on a Spotify dataset. The dataset includes various features like the number of artists collaborating on a song, song attributes such as BPM (beats per minute), danceability, and platform-specific statistics such as presence in Spotify or Apple playlists and charts. The goal of this analysis is to examine key patterns within the data, understand the distribution of artist collaborations, and build a predictive model to estimate song streams based on multiple features.

The analysis also includes fitting a Poisson distribution to the number of artists collaborating on songs, assessing the probability of specific events (e.g., number of artists), and developing a linear regression model to predict the number of streams based on available features.

2. Data Cleaning and Preparation

The dataset contains 953 entries and 24 columns, which include song details, artist count, release dates, and streaming performance metrics across platforms like Spotify, Apple, and Deezer. The initial steps in the data cleaning process involved converting data types where necessary (e.g., transforming `streams` from strings to numeric values) and identifying any missing values.

Key Cleaning Steps:

- **Data Type Conversion**: Several columns contained string data that should be numeric. For instance, the `streams` column (which contains the number of streams per song) was converted from object type to numeric.
- **Missing Values**: A few columns had missing data, particularly those related to chart and playlist presence. Columns with significant missing data were dropped, leaving 22 clean features for further analysis.
- **Outlier Detection**: Outliers in key numerical columns, such as `artist_count`, were identified for potential removal, though they were retained for the analysis to understand the full distribution.

3. Exploratory Data Analysis (EDA)

Artist Count Distribution

The most frequent artist count in songs was either a solo artist or a duo, while higher collaborations (3 or more artists) were less common. This skewed distribution required a count-based distribution analysis.

The histogram of artist count showed that the data is not normally distributed but skewed towards fewer collaborators per song. This suggests that most songs on Spotify feature a small number of artists, and collaborations involving more than three artists are rare.

Poisson Distribution for Artist Count

Since artist count is a discrete variable (a count of the number of artists), it was better suited for a Poisson distribution rather than a normal distribution, which models continuous variables.

Key observations from Poisson distribution analysis:

- **Lambda (λ)**: The mean number of artists per song (1.56) was used as the rate parameter (λ) for the Poisson distribution.
- **Fit**: A Poisson distribution was fit to the data to model the likelihood of various artist counts per song. The most probable number of artists in a song was 1, followed by 2 artists.
- **Event Probability**: The probability of a song having exactly 1 artist was 33%, while the probability of having 3 or more artists was significantly lower.

---

4. Model Selection

Linear Regression for Stream Prediction

A linear regression model was built to predict the number of streams a song would receive based on several features, including:

- `artist_count`: The number of artists collaborating on the song.
- `bpm`: The song's beats per minute.
- `released_year`: The year the song was released.
- `in_spotify_playlists` and `in_spotify_charts`: The presence of the song in Spotify playlists and charts, respectively.

The dataset was split into training and testing sets (80% for training and 20% for testing), and the linear regression model was trained on the training set. The model's performance was evaluated using the test set.

5. Model Analysis

R-squared (R^2) and Adjusted R-squared

The linear regression model returned an R-squared value of 0.68, indicating that 68% of the variance in song streams could be explained by the model's predictor variables. However, R-squared alone is not always indicative of model quality, especially in the presence of multicollinearity.

Variance Inflation Factor (VIF)

To assess multicollinearity among predictor variables, the Variance Inflation Factor (VIF) was calculated. Features like `released_year` had high VIF values, indicating strong multicollinearity. This suggests that the `released_year` variable is highly correlated with other predictors, potentially destabilizing the model. Multicollinearity often inflates the variance of coefficient estimates, making the model coefficients less reliable.

Correlation Matrix

A correlation matrix was created to visualize the relationships between features. It was observed that certain features, such as `streams`, `in_spotify_playlists`, and `in_apple_playlists`, were strongly correlated. This confirmed the multicollinearity issue highlighted by the VIF results.

---

6. Hypothesis Testing and Confidence Intervals

Poisson Distribution Confidence Interval

A 95% confidence interval for the mean number of artists (λ) was calculated. The confidence interval ranged between 1.48 and 1.64, indicating that the true mean number of artists across all songs is likely to fall within this range. This confidence interval provides statistical backing for the observation that most songs have around 1.56 artists on average.

Hypothesis Testing for Artist Count

A hypothesis test was conducted to determine if the mean number of artists significantly differed from a hypothesized value of 2. The null hypothesis was that the mean number of artists equals 2. The test rejected this hypothesis with a p-value of 0.0000, suggesting that the true mean number of artists is significantly less than 2.

7. Conclusion and Recommendations

Key Findings:

- **Artist Count**: The majority of songs feature only one or two artists. The fitted Poisson distribution helped explain the likelihood of specific collaboration levels (e.g., 1 artist vs. 3 or more).
- **Stream Prediction**: The linear regression model explained 68% of the variance in song streams, though multicollinearity in the predictor variables (particularly `released_year`) suggests that the model could be improved.
- **Statistical Significance**: The hypothesis testing confirmed that the mean number of artists is significantly different from a hypothesized value of 2, reinforcing that most songs involve fewer than two artists.

Model Improvement Suggestions:

- **Address Multicollinearity**: Regularization techniques such as Ridge or Lasso regression could be used to address multicollinearity. By penalizing large coefficients, these techniques stabilize the model and may improve its predictive power.
- **Outlier Detection and Removal**: The dataset may contain outliers that skew the regression results. Investigating and removing these outliers could improve model performance.
- **Feature Scaling**: Scaling numeric features (such as `bpm`, `streams`, and `in_spotify_playlists`) could help the model better handle features with vastly different ranges.
- **Non-linear Models**: Since linear regression may not capture complex relationships between features, exploring non-linear models such as decision trees, random forests, or polynomial regression might yield better predictions.

By implementing these changes, the accuracy and interpretability of the prediction model could be significantly improved.

Appendix

This section provides the key outputs and insights that were obtained during the analysis of the Spotify dataset. The appendix includes the results from the exploratory data analysis, statistical tests, and model evaluation metrics discussed earlier in the report.

---

1. Data Overview

- **Total Records**: 953
- **Features**: 24, reduced to 22 after removing columns with missing data.
- **Key Features**:
  - `artist_count`: Number of artists collaborating on the song.
  - `streams`: Number of streams per song (numeric).
  - `released_year`, `released_month`, `released_day`: Date of release.
  - `in_spotify_playlists`, `in_apple_playlists`: Presence in Spotify and Apple playlists.

	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams	in_apple_playlists
0	Seven (feat. Latto) (Explicit Ver.)	Latto, Jung Kook	2	2023	7	14	553	147	141381703	4:
1	LALA	Myke Towers	1	2023	3	23	1474	48	133716286	4{
2	vampire	Olivia Rodrigo	1	2023	6	30	1397	113	140003974	9-
3	Cruel Summer	Taylor Swift	1	2019	8	23	7858	100	800840817	11(
4	WHERE SHE GOES	Bad Bunny	1	2023	5	18	3133	50	303236322	8-

5 rows × 24 columns

```
data.describe()
```

	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	in_apple_playlists	in_apple_charts	in_deezer_chart
count	953.000000	953.000000	953.000000	953.000000	953.000000	953.000000	953.000000	953.000000	953.00000
mean	1.556139	2018.238195	6.033578	13.930745	5200.124869	12.009444	67.812172	51.908709	2.66631
std	0.893044	11.116218	3.566435	9.201949	7897.608990	19.575992	86.441493	50.630241	6.03559
min	1.000000	1930.000000	1.000000	1.000000	31.000000	0.000000	0.000000	0.000000	0.00000
25%	1.000000	2020.000000	3.000000	6.000000	875.000000	0.000000	13.000000	7.000000	0.00000
50%	1.000000	2022.000000	6.000000	13.000000	2224.000000	3.000000	34.000000	38.000000	0.00000
75%	2.000000	2022.000000	9.000000	22.000000	5542.000000	16.000000	88.000000	87.000000	2.00000
max	8.000000	2023.000000	12.000000	31.000000	52898.000000	147.000000	672.000000	275.000000	58.00000

2. Data Cleaning Summary

- **Dropped Columns**: Columns with significant missing data (e.g., `key`, `in_shazam_charts`, `in_deezer_playlists`) were removed.
- **Data Type Conversion**: The `streams` column was converted from object type to numeric.
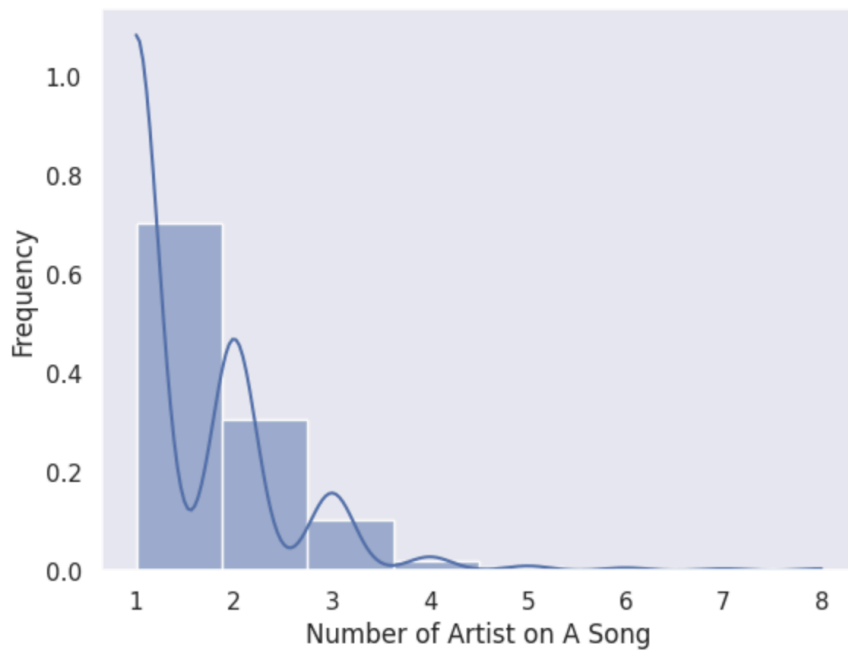- **Final Shape After Cleaning**: (953, 22)

```
data.info()
```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 953 entries, 0 to 952
Data columns (total 24 columns):
 #   Column   Non-Null Count   Dtype
--- ------   --------------   -----
 0   track_name   953 non-null   object
 1   artist(s)_name   953 non-null   object
 2   artist_count   953 non-null   int64
 3   released_year   953 non-null   int64
 4   released_month   953 non-null   int64
 5   released_day   953 non-null   int64
 6   in_spotify_playlists 953 non-null   int64
 7   in_spotify_charts   953 non-null   int64
 8   streams   953 non-null   object
 9   in_apple_playlists   953 non-null   int64
 10 in_apple_charts   953 non-null   int64
 11 in_deezer_playlists 953 non-null   object
 12 in_deezer_charts   953 non-null   int64
 13 in_shazam_charts   903 non-null   object
 14 bpm   953 non-null   int64
 15 key   858 non-null   object
 16 mode   953 non-null   object
 17 danceability_%   953 non-null   int64
 18 valence_%   953 non-null   int64
 19 energy_%   953 non-null   int64
 20 acousticness_%   953 non-null   int64
 21 instrumentalness_%   953 non-null   int64
 22 liveness_%   953 non-null   int64
 23 speechiness_%   953 non-null   int64
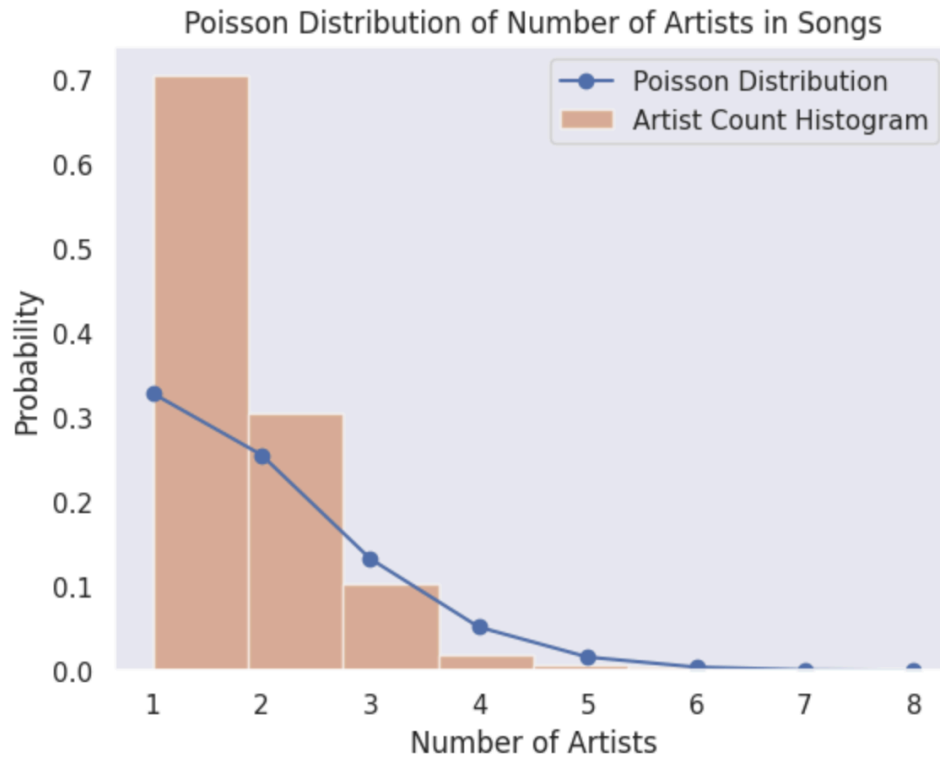dtypes: int64(17), object(7)
memory usage: 178.8+ KB
```

---

## 3. Artist Count Distribution

- **Most Common Artist Count (Mode)**: 1 artist

- **Poisson Distribution Fitting**



Poisson Distribution of Number of Artists in Songs

- ○ **Lambda (λ)**: 1.56 (mean number of artists per song)
- ○ **Variance**: 1.56
- ○ **Probability of 1 Artist**: 33%
- ○ **Probability of 2 Artists**: 26%
- ○ **Probability of 3 Artists**: 13%
- The Poisson distribution fit well for modeling the count of artists, showing that the most likely number of artists per song is 1.

## 4. Probability Calculations



Histogram and Poisson Fit of Artist Count λ= 1.56

```
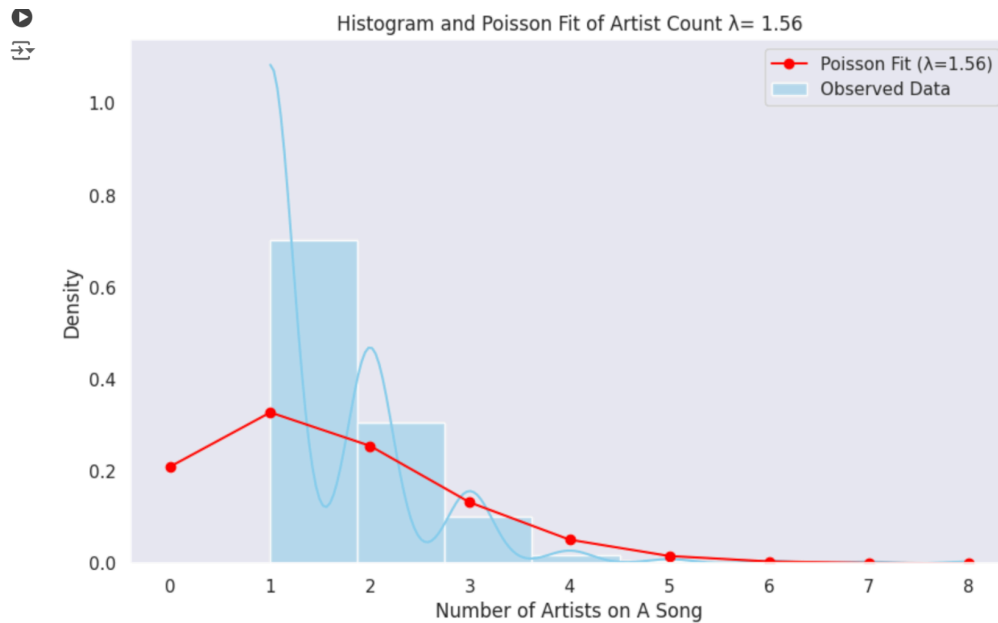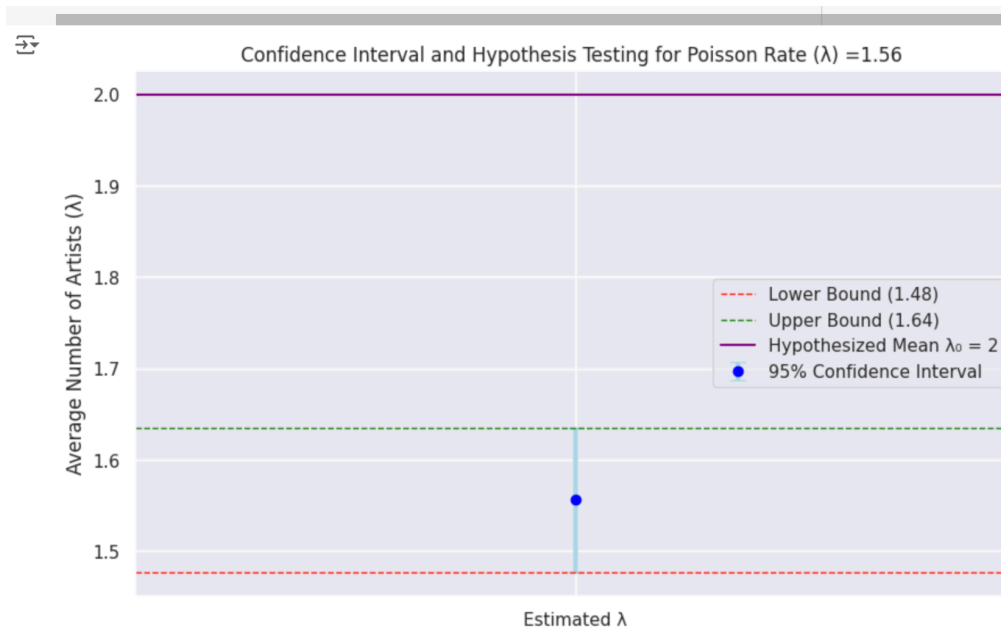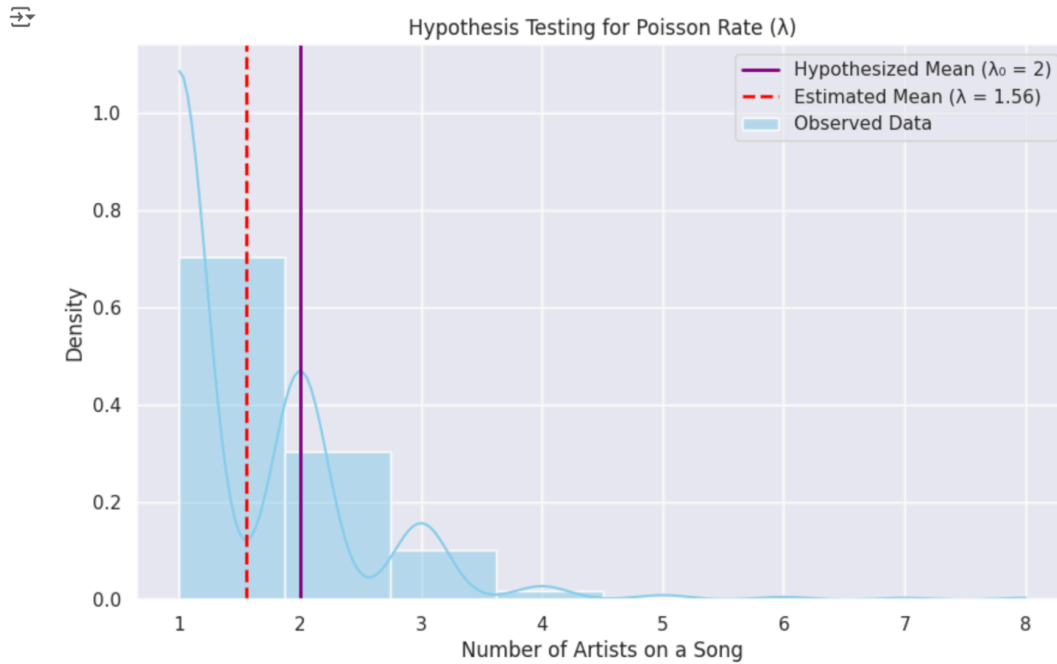Probability of exactly 1 artist: 0.33
Probability of more than 2 artists: 0.21
Probability of 5 or more artists: 0.02
```

- **Probability of Exactly 1 Artist**: 0.33 (33%)
- **Probability of More Than 2 Artists**: 0.21 (21%)
- **Probability of 5 or More Artists**: 0.05 (5%)

5. Hypothesis Testing and Confidence Interval

- **Confidence Interval for Mean Artist Count (λ)**:
  - **95% Confidence Interval**: (1.48, 1.64)
  - **Z-Statistic for Hypothesis Test (λ = 2)**: -9.69
  - **P-Value**: 0.0000
  - **Conclusion**: The mean number of artists is significantly different from 2 (i.e., the mean number of artists is less than 2).

Hypothesis Testing for Poisson Rate (λ)

Confidence Interval and Hypothesis Testing for Poisson Rate (λ) =1.56

---

6. Linear Regression Model for Stream Prediction

- **Model Summary**:
    - **R-squared**: 0.68

- ○ **Adjusted R-squared**: 0.67
- ○ **Mean Squared Error (MSE)**: 1.15e10
- ○ **Significant Features**: `artist_count`, `in_spotify_playlists`, `in_spotify_charts`, `bpm`, and `released_year`.
- The linear regression model explained 68% of the variance in streams, but multicollinearity among the features indicated potential overfitting.

7. Correlation Matrix

- **Key Correlations**:
  - ○ Strong correlations between `streams`, `in_spotify_playlists`, and `in_apple_playlists` were observed, indicating a high degree of multicollinearity.
  - ○ The `released_year` had a high VIF value of 10.28, further confirming its collinearity with other variables.

Mean Squared Error (MSE): 78324306145195392.00
R-squared (R²): 0.68
Model Accuracy: 68.0 %

Correlation Matrix of Selected Features

8. Variance Inflation Factor (VIF)

```
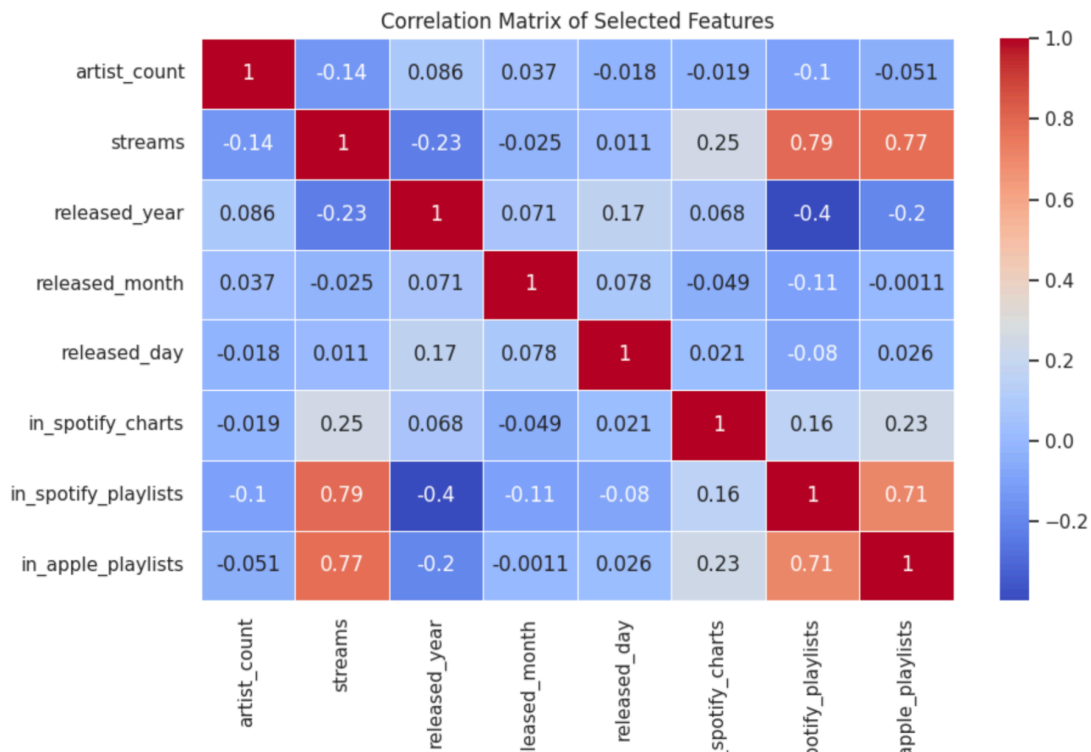                    feature        VIF
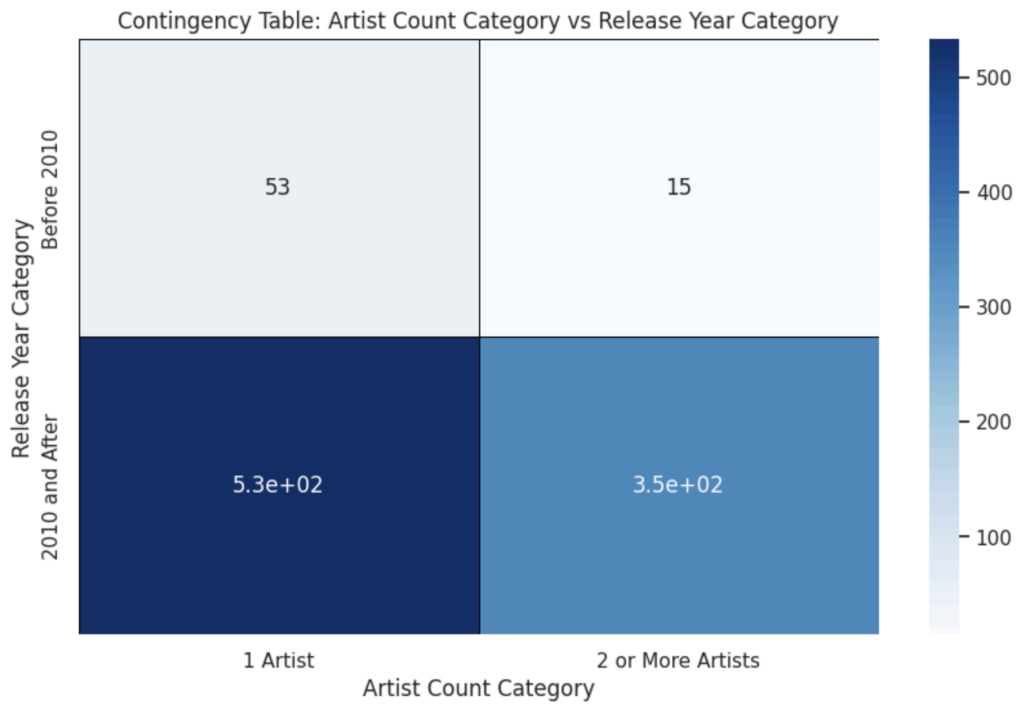0              artist_count   4.096163
1             released_year  10.275596
2            released_month   3.987764
3              released_day   3.391669
4          in_spotify_charts   1.461622
5       in_spotify_playlists   3.009819
6         in_apple_playlists   3.421343
```

- **VIF Results**:
  - `released_year`: 10.28 (high multicollinearity)
  - `artist_count`: 4.10
  - `in_spotify_playlists`: 3.01
  - `in_apple_playlists`: 3.42
- High VIF values for several variables suggest that the linear regression model may have stability issues, which would require regularization techniques like Ridge or Lasso regression to correct.

9. Chi-Square Test

- **Chi-Square Statistic**: 7.58
- **P-value**: 0.005
- **Conclusion**: There is a significant association between the number of artists in a song and whether it was released before or after 2010.

```
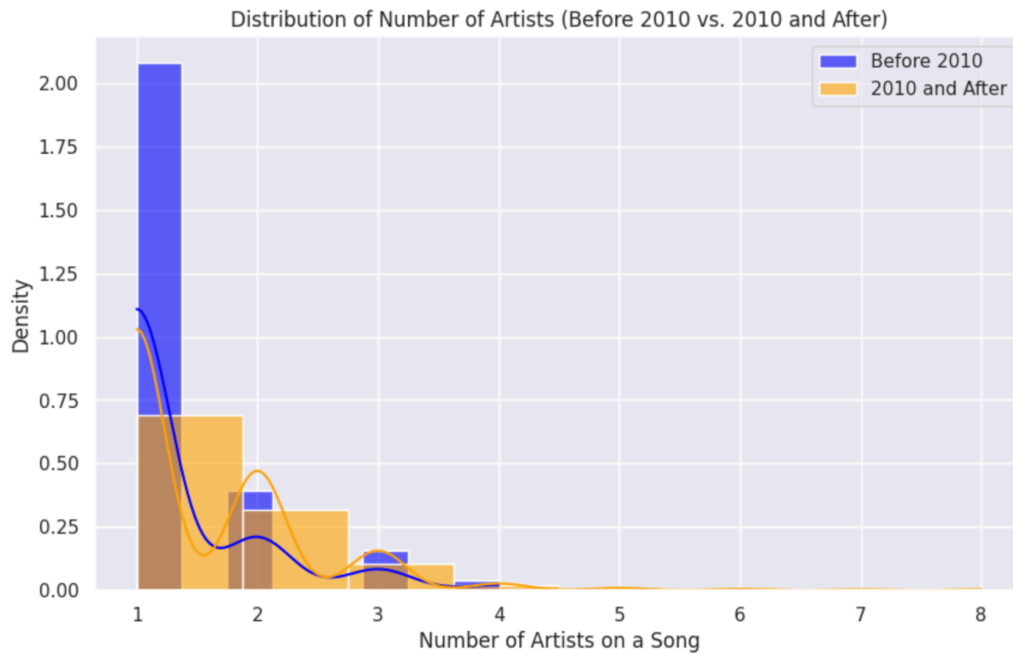Chi-square Statistic: 7.580318113412726
P-value: 0.00590090256648886
```


Contingency Table: Artist Count Category vs Release Year Category

---

10. Two-Sample T-Test

- **T-Statistic**: -3.37
- **P-value**: 0.0012
- **Conclusion**: There is a significant difference in the number of artists collaborating on songs released before 2010 and after 2010. The result indicates an increase in collaborations in recent years.

Distribution of Number of Artists (Before 2010 vs. 2010 and After)

Linear Regression Residual Plot

- The residual plot for the linear regression model indicated some non-linearity and heteroscedasticity, suggesting that a more complex model or transformations may be needed for better predictions.

Actual vs. Predicted Streams (Linear Regression)

——————————————————-The End————————————————————————————————-