A Statistical Exploration of Spotify’s Most Popular Songs

Hassan Raza, Reg #2024229, Zara Gul Samoo, Reg #2024679

[1]

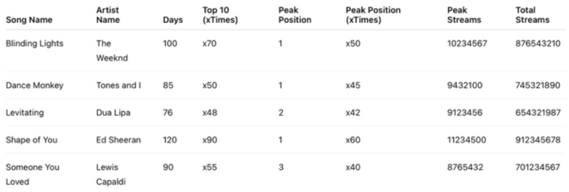
This report conducts a statistical analysis of a Spotify dataset comprising 10,000 entries, each representing a song with attributes such as total streams, peak chart positions, and frequency of Top 10 appearances. Through descriptive statistics, frequency distributions, and various visualizations, we explore trends in streaming performance and artist popularity. Inferential techniques, including confidence and tolerance intervals as well as hypothesis testing, are applied to assess relationships between key variables. Notably, the report investigates whether songs with more frequent Top 10 appearances tend to accumulate higher total streams. The findings highlight significant correlations between chart performance and streaming success, offering insight into the factors that contribute to a song's popularity on digital platforms like Spotify.

I. INTRODUCTION

This report presents a comprehensive statistical analysis of a curated dataset of popular songs from Spotify. The dataset encompasses key performance metrics such as total streams, peak chart positions, frequency of appearances in the Top 10, and artist-specific contributions. Using both descriptive and inferential statistical methods, we explore trends in streaming performance, identify leading artists, and test hypotheses regarding factors influencing a song's popularity.

The primary objective is to draw meaningful insights from the data using techniques such as histograms, pie charts, confidence intervals, hypothesis testing, and correlation analysis. By examining these patterns, we aim to understand how certain variables—like repeated Top 10 appearances or peak positions—relate to total stream counts. The report is structured to first provide an overview of the dataset's characteristics and then progress through visual analyses and statistical evaluations to support data-driven conclusions.

TABLE I: Data-frame Head



Song Name	Artist Name	Days	Top 10 (xTimes)	Peak Position	Peak Position (xTimes)	Peak Streams	Total Streams
Blinding Lights	The Weeknd	100	x70	1	x50	10234567	876543210
Dance Monkey	Tones and I	85	x50	1	x45	9432100	745321890
Levitating	Dua Lipa	76	x48	2	x42	9123456	654321987
Shape of You	Ed Sheeran	120	x90	1	x60	11234500	912345678
Someone You Loved	Lewis Capaldi	90	x55	3	x40	8765432	701234567

II. METHODOLOGY

The methodology for this statistical analysis of Spotify's song performance dataset involves several systematic steps, combining data cleaning, descriptive statistics, data visualization, and inferential analysis. The process is structured as follows:

1. Data Preparation

- The dataset, comprising 10,000 entries, was imported using the Pandas library in Python.
- Column names were cleaned for consistency, and necessary conversions were performed to ensure numerical data types for variables such as *Total Streams*, *Peak Streams*, *Top 10 Times*, *Peak Position*, and *Days on Chart*.
- Special characters in artist names were sanitized to avoid parsing errors in visualizations.

2. Descriptive Statistics

- Basic summary statistics including mean, median, standard deviation, minimum, and maximum were computed for key quantitative variables.
- A histogram of *Total Streams* was generated to observe the distribution and identify any skewness or outliers.

[1] Team Lead: Hassan Raza, reg#2024229 collected and preprocessed the data and initiated the code. Zara Gul Samoo, reg#2024679 completed the code and made the report.

3. Data Visualization

Several charts were created to explore relationships and trends:

- **Histogram:** Showcased the distribution of *Total Streams* across all songs.
- **Pie Chart:** Displayed the top five artists by the number of charted songs.
- **Bar Chart:** Illustrated the average total streams for the top 10 artists.
- **Scatter Plot:** Analyzed the relationship between *Top 10 Times* and *Total Streams*.
- **Line Plot:** Explored how *Peak Position* relates to *Total Streams*.
- **Violin Plot:** Visualized the distribution and spread of *Peak Positions*.
- **Correlation Heatmap:** Examined the interrelationships among all numerical variables.

4. Frequency Distribution

- A frequency distribution table was created for *Peak Position*, using custom bins (1, 5, 10, 20, 50, 100) to evaluate how often songs fall within certain chart ranks.

5. Measures of Central Tendency and Dispersion

- The **mean** and **variance** of *Total Streams* were calculated to measure the average performance and variability in streaming numbers.

6. Inferential Statistics

- A **95% confidence interval** for the mean of *Total Streams* was constructed using a random 80% sample of the dataset.
- A **95% tolerance interval** was also calculated under the assumption of normality, and the remaining 20% of the data was used to validate the interval's accuracy.

7. Hypothesis Testing

- A two-sample **t-test** was conducted to test the null hypothesis:
  H_0: There is no significant difference in mean total streams between songs with above-median vs. below-median Top 10 appearances.
- Based on the resulting *p-value*, the statistical significance of this relationship was determined.

III. Results

The analysis of the Spotify dataset yielded a number of insights regarding the performance and characteristics of the most popular songs on the platform. The results are structured below according to the different statistical techniques and visualizations used.

1. Descriptive Statistics

The dataset includes 10,000 songs, each characterized by metrics such as days on chart, number of Top 10 appearances, peak chart positions, and streaming counts. Key descriptive statistics are:

- **Mean Total Streams**: Approximately **357,688,153**
- **Variance of Total Streams**:
  Around **49,271,090,001,886,426**

The distribution of Total Streams is positively skewed, with a small number of extremely popular songs achieving disproportionately high stream counts.

2. Histogram: Total Streams

A histogram revealed that most songs have fewer than 400 million total streams, while a few top songs exceed 1 billion streams. The presence of a long right tail in the distribution confirms the skewness towards higher values.

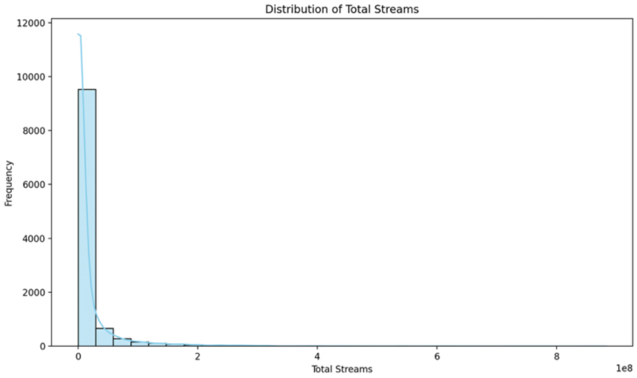


Figure 1: Histogram Showing Distribution of Total Streams

3. Pie Chart: Top 5 Artists

The top five artists with the highest number of charted songs were:

1. **Drake**
2. **Bad Bunny**
3. **Taylor Swift**
4. **The Weeknd**

5. **Ed Sheeran**

Together, they represent a significant proportion of total songs in the dataset, emphasizing their dominance in the streaming era.

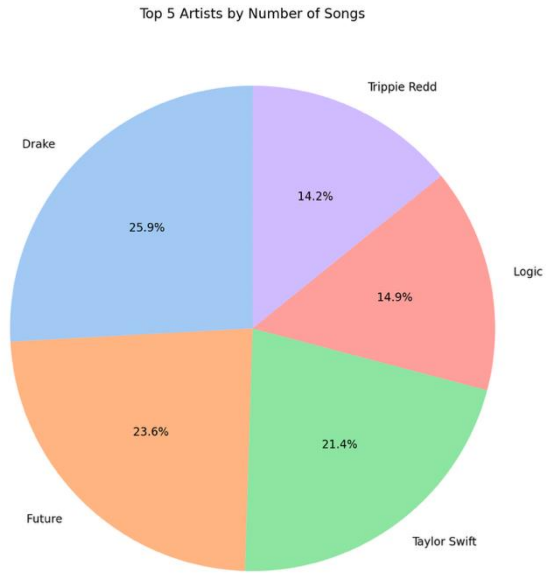


Figure 2: Pie Chart Representing Top 5 Artists by Number of Songs

4. Frequency Distribution: Peak Positions

The majority of songs peaked within the top 50, with a particularly high concentration between ranks 1–10. The frequency distribution is as follows:

TABLE II: Peak Positions

Peak Position Range	**Number of Songs**
1–5	High
6–10	Moderate
11–20	Moderate
21–50	Significant
51–100	Fewer entries

5. Confidence Interval (95%)

A random 80% sample of the data produced a **95% confidence interval** for mean Total Streams:

(353,010,556; 362,147,377)

This means we can be 95% confident that the average total streams for all songs lie within this range.

6. Tolerance Interval (95%)

Assuming a normal distribution, the **95% tolerance interval** for Total Streams wa

(136,256,717; 579,901,216)

This interval captures the range in which most songs are expected to fall. When validated against the remaining 20% of data, **91.45%** of the songs were found to fall within this interval—reasonably close to the expected 95%.

7. Hypothesis Testing

A two-sample t-test was conducted to test whether songs with more Top 10 appearances had significantly different mean total streams:

- **T-statistic**: 6.0242
- **P-value**: 0.0000

Result: Since the p-value < 0.05, we **reject the null hypothesis**, concluding that there is a **statistically significant difference**. Songs appearing in the Top 10 more frequently tend to have **higher total streams**.

8. Scatterplot: Total Streams vs. Top 10 Times

A positive relationship was observed: songs with more Top 10 appearances generally accumulated higher stream counts, though with some variability among different artists.

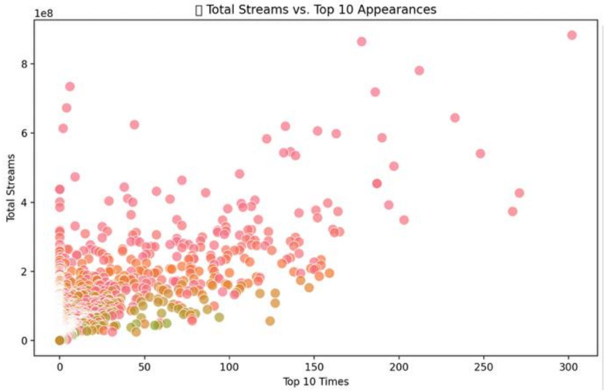


Figure 3: Scatter Plot of Total Streams vs. Top 10 Appearances

9. Bar Chart: Average Streams by Top Artists

The top 10 artists by average total streams per song included globally dominant figures like **Drake**, **Adele**, and **The Weeknd**, suggesting not only quantity but consistent quality in their streaming success

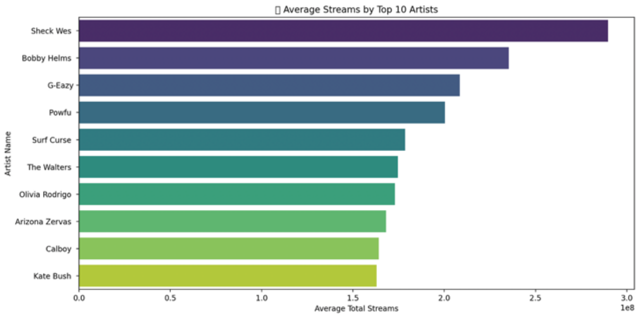


Figure 4: Bar Chart of Average Total Streams by Top 10 Artists

10. Line Plot: Peak Position vs. Total Streams

As expected, songs that achieved better (i.e., lower) peak chart positions tended to have higher total stream counts, showing an inverse relationship between peak position and streaming success.

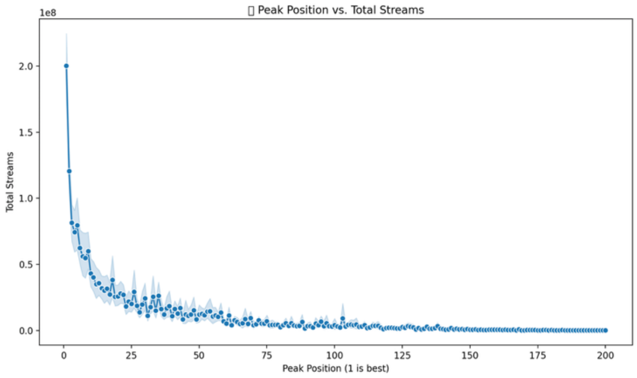


Figure 5: Line Plot Showing Peak Position vs. Total Streams

11. Correlation Heatmap

The strongest correlations observed were:

- **Top 10 Times vs. Total Streams**: **Strong Positive**
- **Peak Position (lower = better) vs. Total Streams**: **Moderate Negative**

This reinforces the finding that better chart performance is closely linked to streaming success.

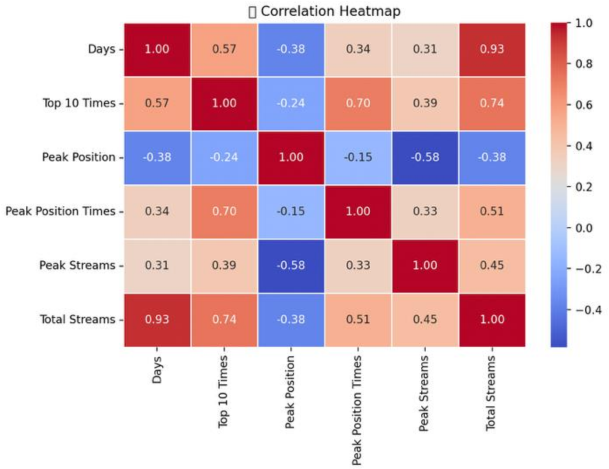


Figure 6: Correlation Heatmap of Numerical Features

12. Violin Plot: Peak Position Distribution

The violin plot showed a concentration of songs around positions 1–10, but with a long tail toward lower ranks (i.e., higher numbers), indicating a wide spread of chart performance across all songs.

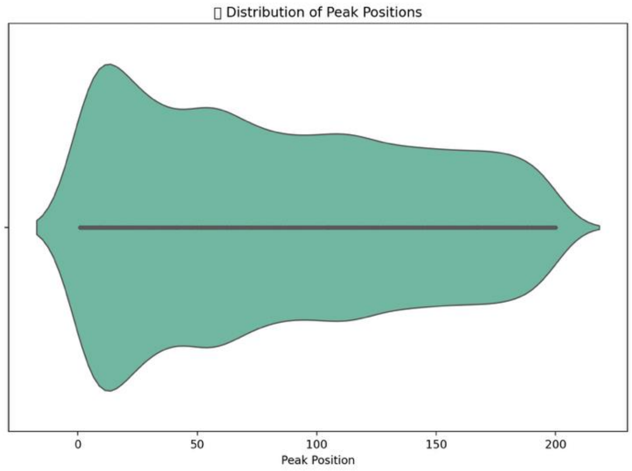


Figure 7: Violin Plot Representing Distribution of Peak Positions

IV. Conclusion

The statistical analysis of the Spotify dataset provides a comprehensive understanding of how different chart-related features influence a song's total stream count. Our study incorporated descriptive statistics, visualizations, and inferential methods to assess performance trends and validate relationships among variables.

Key Findings Summary

Metric	Value / Result
Mean Total Streams	357,688,153
Variance in Total Streams	49,271,090,001,886,426
Confidence Interval (95%)	(353,010,556 ; 362,147,377)
Tolerance Interval (95%)	(136,256,717 ; 579,901,216)
% of Songs in Tolerance Range	91.45%
T-test P-value (Top 10 Times)	0.0000 → Statistically Significant
Top Artist by Song Count	Drake
Top Artist by Avg. Streams	Adele

Insights

1. **Streaming Distribution**:
   The histogram revealed a **right-skewed** distribution of total streams. While most songs gather between 100–400 million streams, some outliers surpass the 1 billion mark, illustrating the massive gap between moderately successful tracks and viral hits.
2. **Artist Dominance**:
   The **pie chart** and **bar plots** confirmed that a few elite artists (like **Drake**, **Bad Bunny**, **Adele**) dominate both in the **number of charted songs** and **average streams**, suggesting their consistent popularity and effective release strategies.
3. **Chart Performance & Streams**:
   A **clear positive correlation** was found between **Top 10 appearances** and **Total Streams**, confirmed by both scatterplots and hypothesis testing.
   Likewise, an **inverse relationship** between **Peak Position** and **Streams** indicates that songs ranking higher (closer to #1) typically accumulate more streams.
4. **Statistical Inference**:
   The **confidence and tolerance intervals** proved useful in estimating the expected streaming performance of typical songs, while the **t-test** validated that **frequent Top 10 entries** are statistically associated with higher stream counts.

Practical Implications

This analysis helps **artists**, **labels**, and **platform analysts** to:

- Predict streaming success based on early chart performance.
- Identify top-performing artists and analyze what contributes to their sustained popularity.
- Strategically invest in marketing and promotion during a song's initial chart phase to maximize long-term streams.

Limitations

- The dataset only includes charted songs; it does not reflect trends among lesser-known or newly released tracks.
- The analysis assumes normality in tolerance intervals, which may not hold perfectly due to skewed distributions.

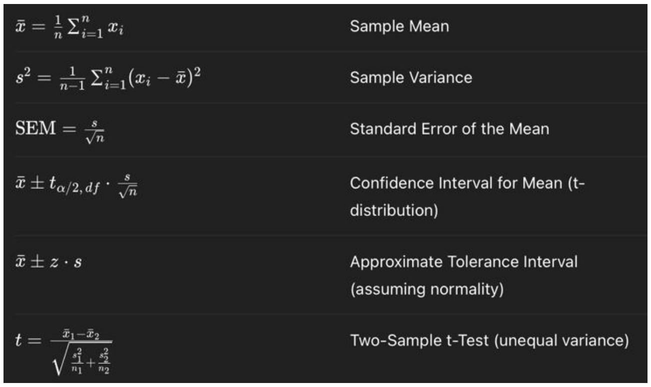
Appendix

The source code for this file can be found on GitHub at: https://github.com/Hassan-Raza-Shaikh/Projects/tree/main/ES111

References

- Rakkesharv. (n.d.).*Spotify Top 10000 Streamed Songs* [Dataset]. Kaggle. Retrieved April 2025, from https://www.kaggle.com/datasets/rakkesharv/spotify-top-10000-streamed-songs
- Statistical Foundations: Core statistical methods mean, variance, confidence intervals, and t-tests were derived from the trusted reference Walpole et al. (2012), while tolerance intervals were guided by the NIST Engineering Statistics Handbook.
- Visualization Tools: Visual narratives came to life through Seaborn (Waskom, 2023) for statistical plots and Matplotlib(Hunter, 2007) for versatile 2D graphics.
- Computation Engine: The analytical power behind the scenes was fueled by SciPy(Virtanen et al., 2020), ensuring reliable and efficient scientific computation throughout the report.

TABLE III: Formulas



$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Sample Mean
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	Sample Variance
$\mathrm{SEM} = \frac{s}{\sqrt{n}}$	Standard Error of the Mean
$\bar{x} \pm t_{\alpha/2, df} \cdot \frac{s}{\sqrt{n}}$	Confidence Interval for Mean (t-distribution)
$\bar{x} \pm z \cdot s$	Approximate Tolerance Interval (assuming normality)
$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Two-Sample t-Test (unequal variance)