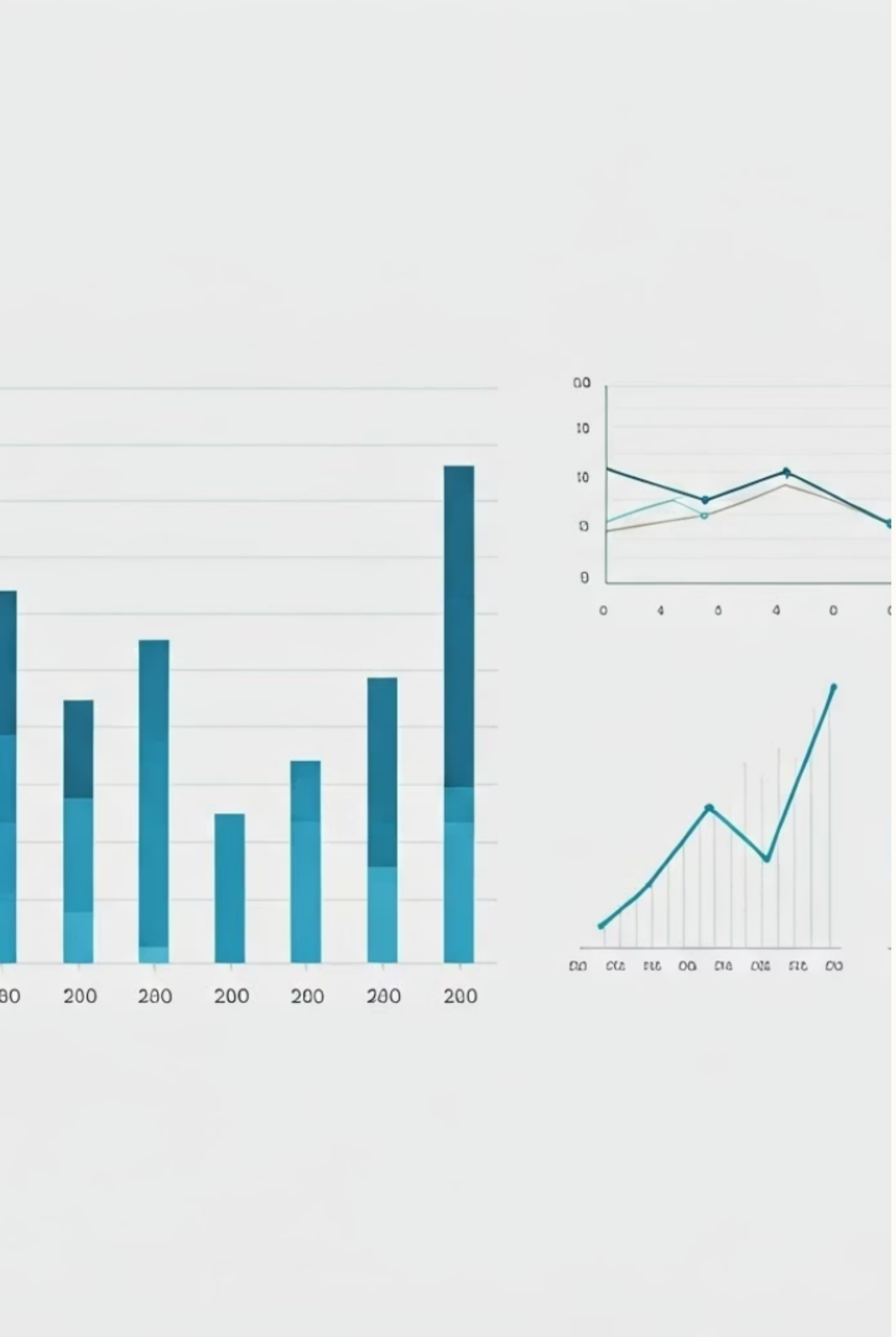


Exploratory Data Analysis on Dua Lipa Lyrics & Zipf's Law

SD **by Sleep Deprived**

Team Members: Aditya Singh, Pratik Kumar Pan, Shreyas Sarkar, Priyank Gaur.

Objective and Dataset Overview

The primary objective of this analysis is to investigate whether Dua Lipa's lyrics exhibit characteristics consistent with Zipf's Law, a well-known linguistic phenomenon. We will analyze a dataset containing the full lyrics of Dua Lipa's songs, exploring various aspects of the text data, including word frequencies, token distributions, and other statistical properties.

The dataset includes **246 song lyrics** from Dua Lipa's discography, spanning her debut self-titled album and subsequent releases. The key columns in the dataset are the song title and the full lyrics text.

Data Cleaning and Preprocessing

Before conducting the exploratory analysis, we performed several data cleaning and preprocessing steps to ensure the integrity of the lyrics data:

Missing Values

Checked for and handled any missing values in the dataset.

Text Normalization

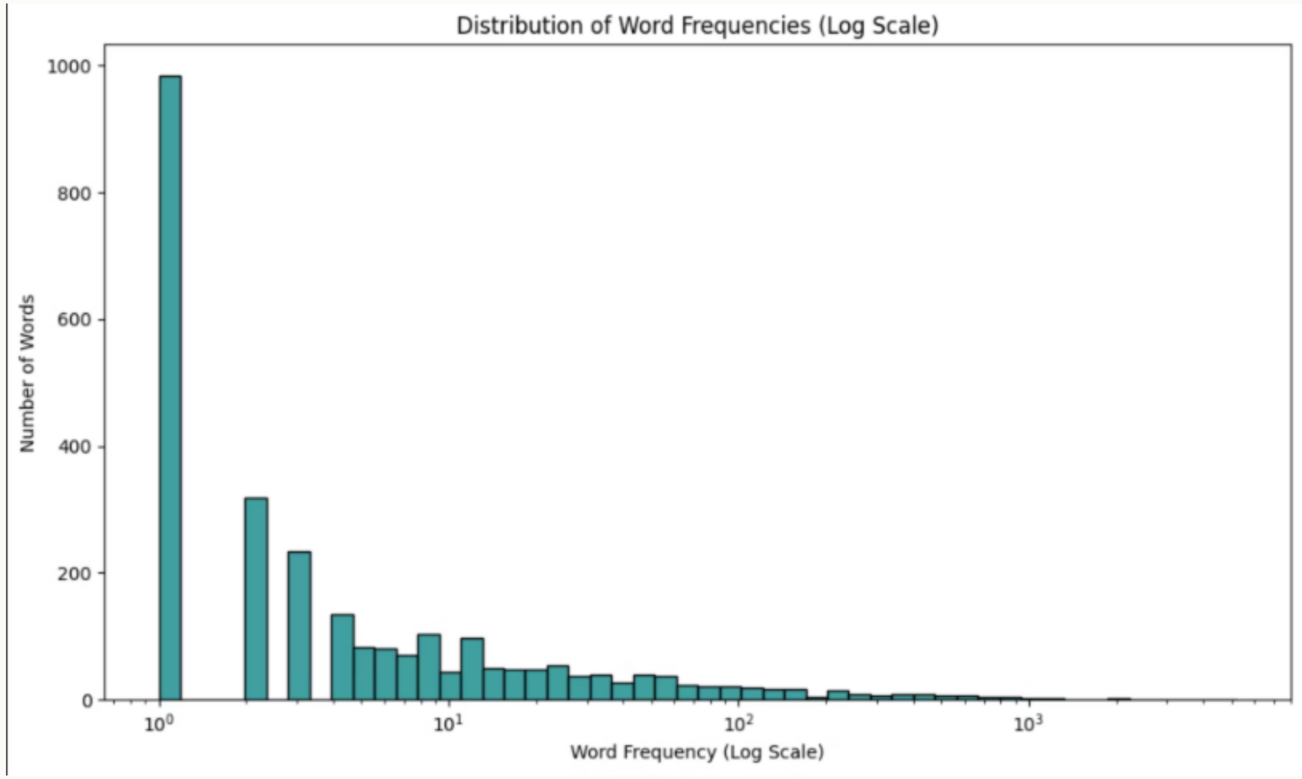
Converted all lyrics to lowercase and removed punctuation to standardize the text.

Tokenization

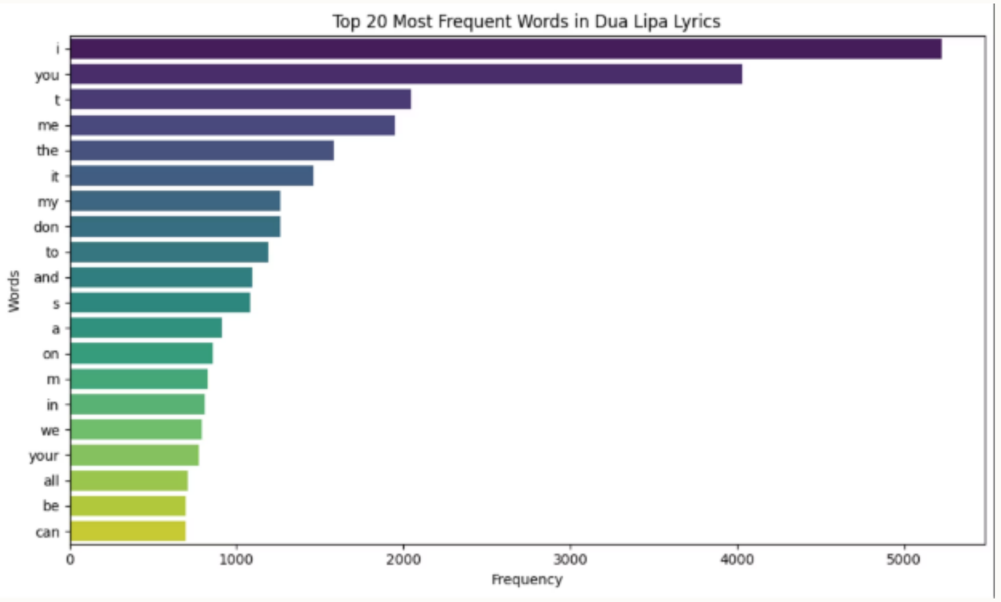
Split the lyrics into individual tokens (words) to facilitate word frequency analysis.

Univariate Analysis

We began our exploratory analysis by examining the distribution of word frequencies in Dua Lipa's lyrics. This univariate analysis provided valuable insights into the lexical richness and diversity of her songwriting.



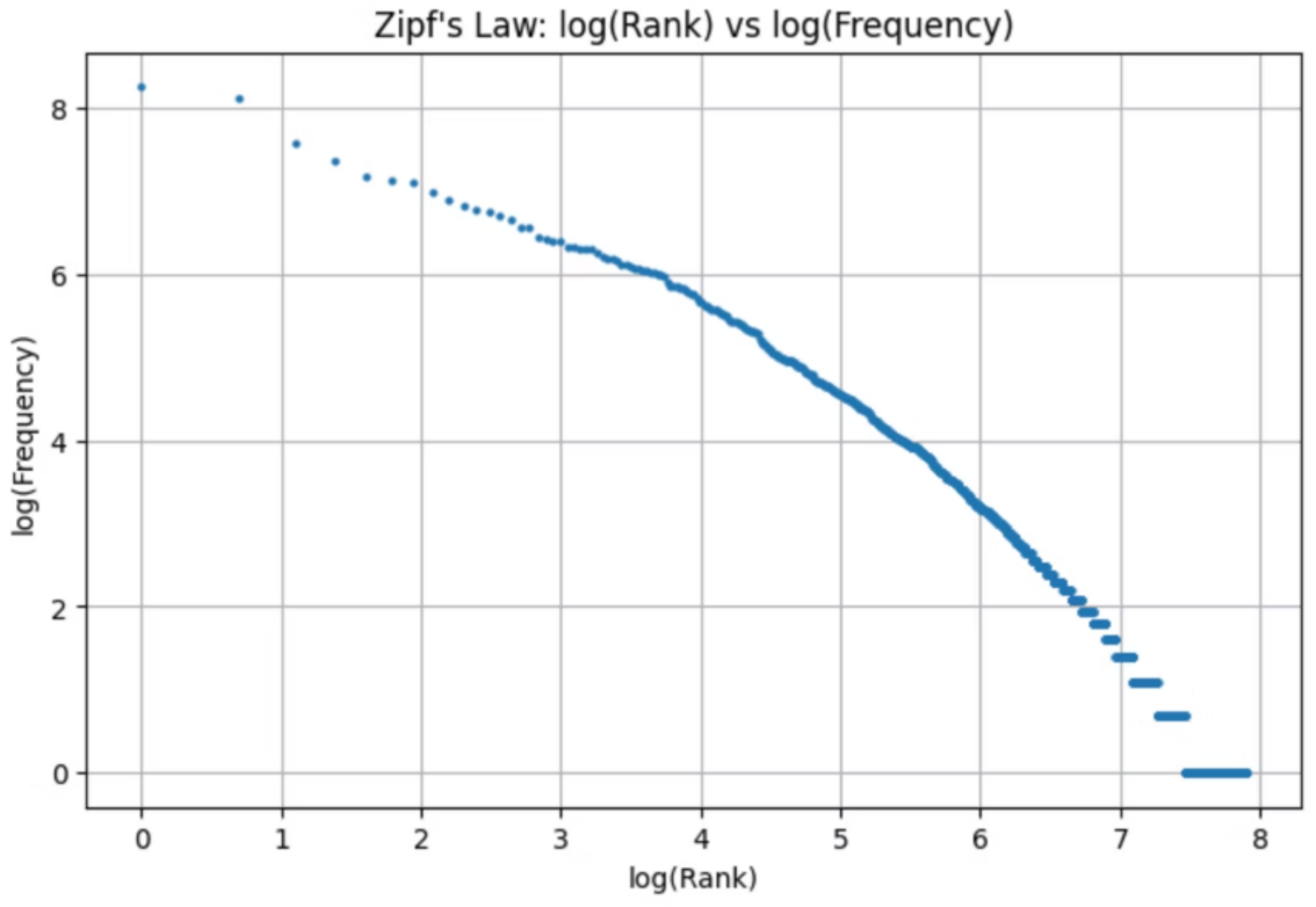
The histogram of word frequencies reveals a long-tailed distribution, with a few high-frequency words and a larger number of low-frequency words.



The bar chart of the top 20 most frequent words in the lyrics highlights common terms like "you", “t", and "i", which are typical of natural language.

Zipf's Law Analysis

Zipf's Law is a statistical principle that describes the relationship between the rank and frequency of words in a natural language. It states that the frequency of a word is inversely proportional to its rank, following a power-law distribution.

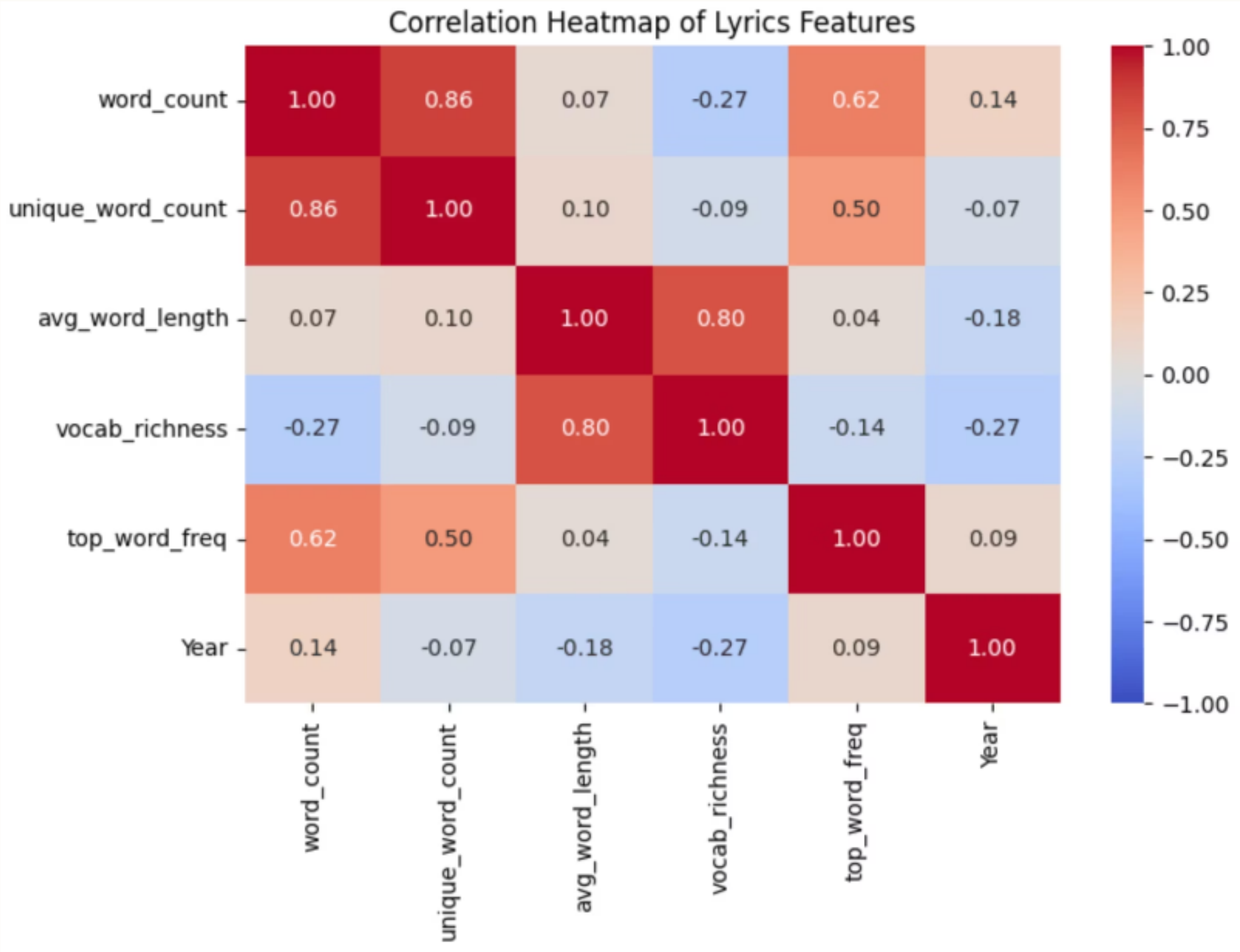


By plotting the logarithm of word rank against the logarithm of word frequency, we can examine whether Dua Lipa's lyrics conform to Zipf's Law. The resulting plot exhibits a linear relationship, suggesting that her songwriting aligns with this linguistic phenomenon.

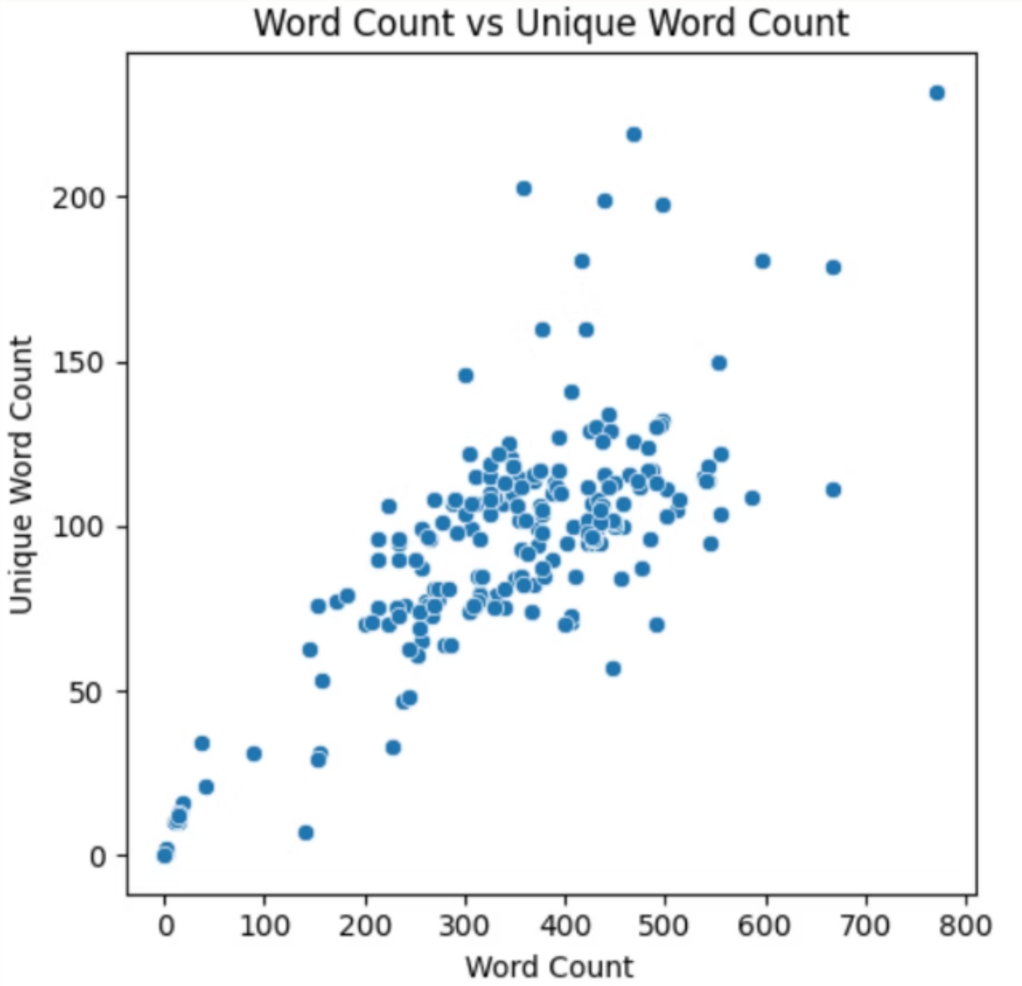
The regression line fitted to the Zipf's Law plot has a slope of approximately -1.57, which is close to the expected value of -1 for natural language. Additionally, the R-squared value is 0.96, indicating a strong linear fit and supporting the conclusion that Dua Lipa's lyrics follow Zipf's Law.

Bivariate and Multivariate Analysis

To further explore the relationships within Dua Lipa's lyrical data, we conducted bivariate and multivariate analyses. These techniques allowed us to uncover additional insights beyond the univariate patterns observed earlier.



The correlation heatmap reveals the strength of associations between different features, such as word counts, unique word counts, and average word length. These insights can inform our understanding of the overall structure and complexity of Dua Lipa's songwriting.



Scatterplots and pair plots provide a visual representation of the relationships between various attributes of the lyrics data. These plots help us identify patterns, outliers, and potential dependencies that may exist within Dua Lipa's creative process.

Findings and Conclusions

Based on the comprehensive exploratory data analysis performed on Dua Lipa's lyrical data, we can draw the following conclusions:

1

Zipf's Law Adherence

Dua Lipa's lyrics exhibit a strong alignment with Zipf's Law, a fundamental principle observed in natural language. The logarithmic plot of word rank versus frequency shows a linear relationship, with a slope close to the expected value of -1.

2

Lexical Diversity

The analysis of word frequencies and distributions reveals a rich and diverse vocabulary used in Dua Lipa's songwriting, with a small number of high-frequency words and a long tail of low-frequency terms.

3

Structural Insights

The correlation and scatterplot analyses provide insights into the relationships between various lyrical features, such as word counts, unique word counts, and average word length, highlighting the structural complexity of Dua Lipa's creative process.

Overall, the exploratory data analysis demonstrates that Dua Lipa's lyrics align with the statistical patterns observed in natural language, further validating the universality of Zipf's Law and the sophistication of her songwriting.

Team Member Contributions

Aditya Singh

Led data collection and cleaning to ensure dataset accuracy and reliability. Designed initial exploratory data analysis framework.

Pratik Kumar Pan

Performed bivariate and multivariate analyses, generating insightful visualizations and correlation heatmaps. Contributed to the presentation slides.

Shreyas Sarkar

Conducted Zipf's Law analysis, interpreting linguistic patterns within Dua Lipa's lyrics with statistical rigor.

Priyank Gaur

Coordinated documentation and presentation design, ensuring a cohesive narrative and visual consistency. Prepared the final PPT slides.

Thank You

We have explored the lyrical data of Dua Lipa through a comprehensive exploratory data analysis, uncovering insights about the structure and patterns within her songwriting. The findings suggest that Dua Lipa's lyrics adhere to Zipf's Law, a fundamental linguistic principle, and exhibit a rich and diverse vocabulary.

Thank you for your time and attention. We welcome any questions or further discussion around this exciting analysis of Dua Lipa's creative work.