Zipf's Law Analysis on Selena Gomez Song Lyrics

A Data-Driven Analysis of Selena Gomez's Discography

Team Name: Visioners

Team Members:

Parth Tandalwade

Ansh Sharma

Yatin Bisht

Ved Vadnere

Individual contributions of each member

Parth Tandalwade → Data Cleaning , Data Handling, Data Visulisation , Understanding and Implementing Zipf's Law

Ansh Sharma → Making of PPT , Data Visulisation , Understanding and Implementing Zipf's Law

Ved Vadnere → Data Research , Understanding Zipf's Law

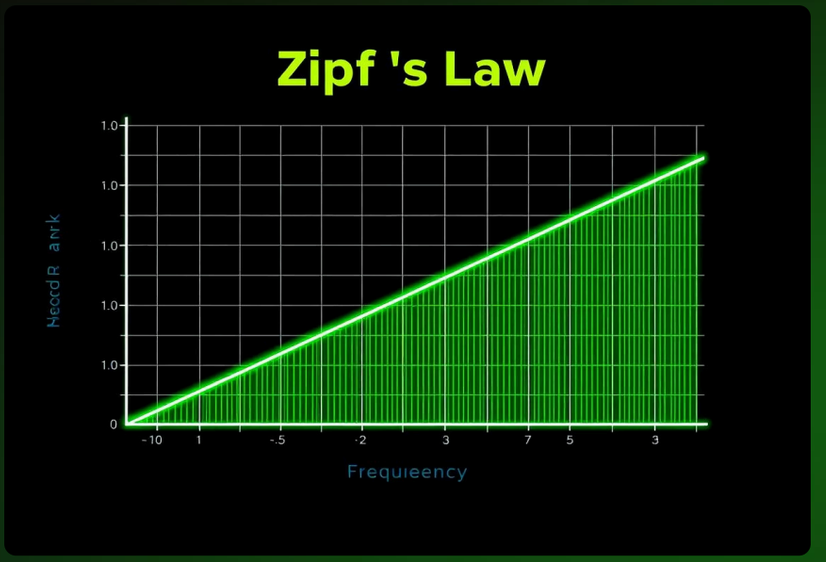
Yatin → Data Analysis , Understanding Zipf's Law , Reviewer

📘 What is Zipf's Law?

- A principle from linguistics and statistics.
- States that **the frequency of a word is inversely proportional to its rank** in the frequency table.
  - Example: 2nd most frequent word appears ≈ ½ as often as the most frequent.
- Commonly observed in:
  - Natural languages
  - City populations
  - Income distributions
  - Web traffic and more

🎯 Project Goal:

- **Analyze Selena Gomez's song lyrics**
- Check if the word frequency in her lyrics follows **Zipf's Law**
- Use visualization and frequency-rank plots to validate



Understanding Zipf's Law

Zipf's Law: Word frequency is inversely proportional to its rank.

Common in Language: Few words are very frequent; most are rare.

Log-Log Plot: Word frequency vs. rank on log-log scale is a linear plot with slope ≈ -1.

Dataset Overview

Dataset Name: SelenaGomez.csv

Total Songs: 81

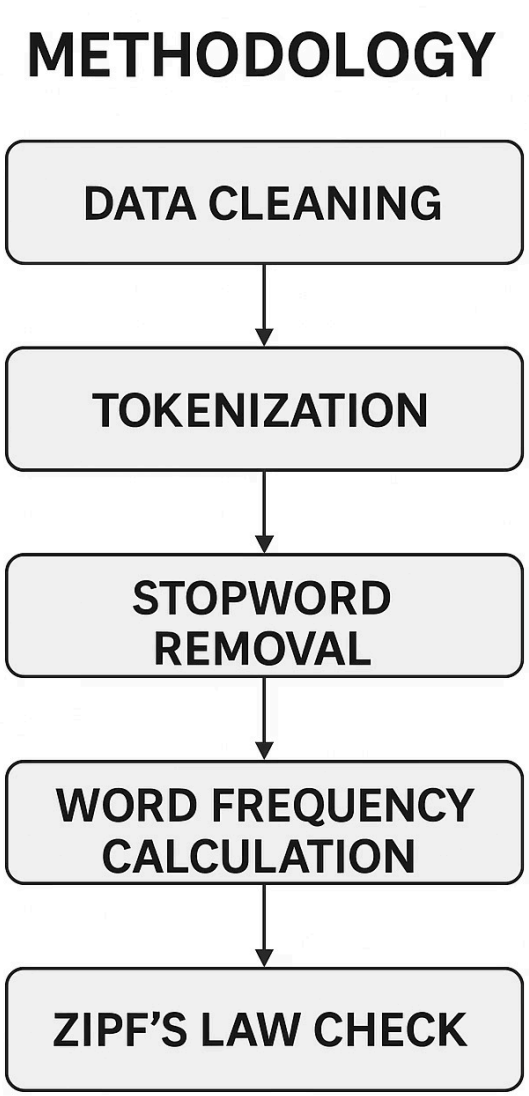
Columns: Title(Name of the song), Album(Album it belongs to), Year(Release year), Lyric(Full song lyrics)

Focus: Lyrics column will be the primary focus for analyzing word frequencies.

🧪 Methodology

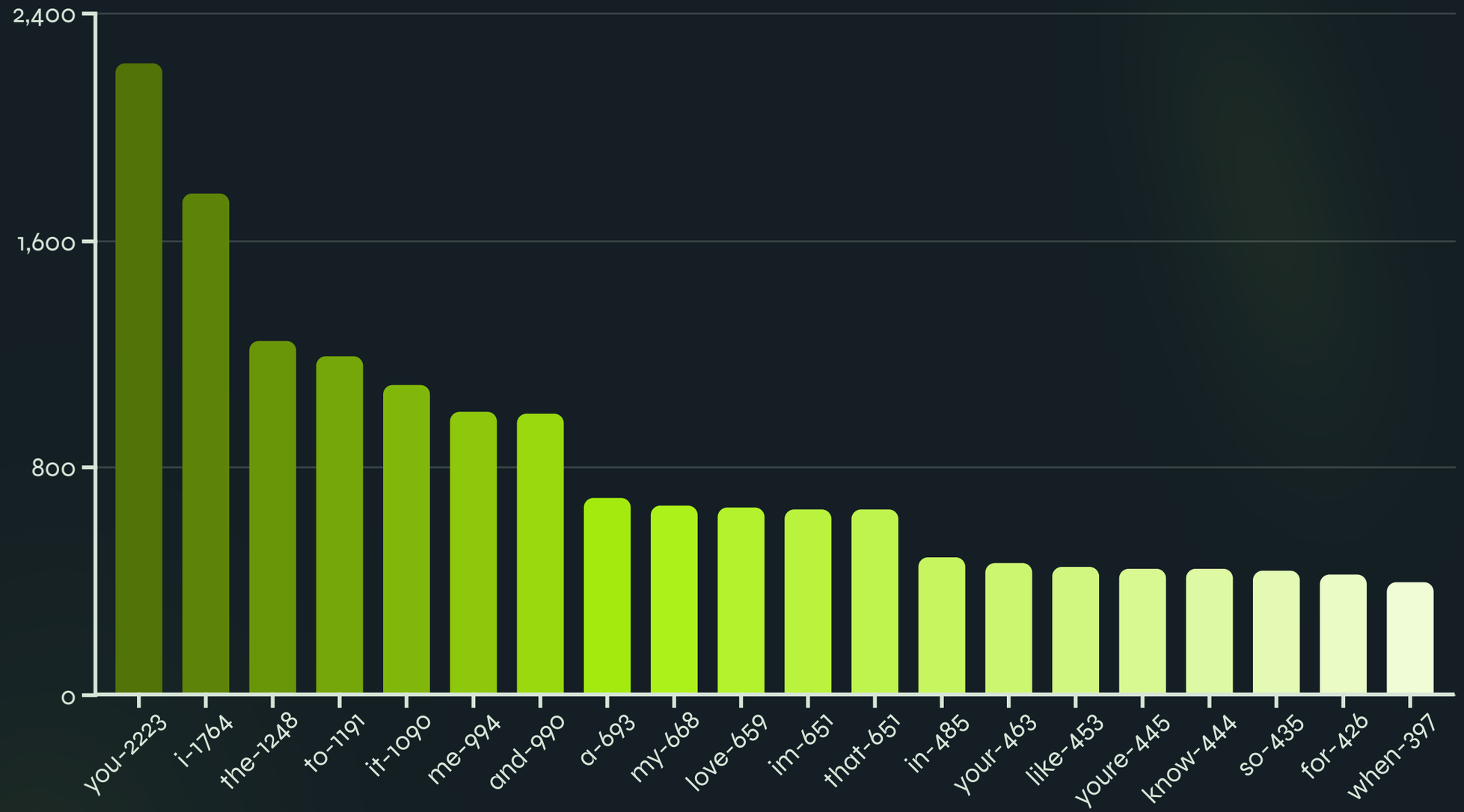
🔄 Steps Followed:

1. **Data Cleaning**
   - Converted lyrics to lowercase
   - Removed punctuation, numbers, and special characters
2. **Tokenization**
   - Split lyrics into individual words (tokens)
3. **Stopword Removal**
   - Removed common words (e.g., "the", "is", "and") that carry little meaning
4. **Word Frequency Calculation**
   - Counted the occurrences of each remaining word
5. **Zipf's Law Check**
   - Plotted word rank vs. frequency on a log-log scale
   - Compared with ideal Zipfian distribution



Word Frequency Analysis

Top Words:



Insights:

- Dominance of emotional words such as "love" and "need".
- Repetition in choruses causes words like "yeah" to appear frequently.

Zipf's Law Verification

Log-Log Plot: Observed near-linear trend but with a flatter slope (e.g., -0.8). Artistic repetition disrupts natural distribution.

Rank	Word	Frequency	Expected (Zipf)
1	you	2223	2223.00
2	i	1764	1111.50

Plot Rank vs Frequency

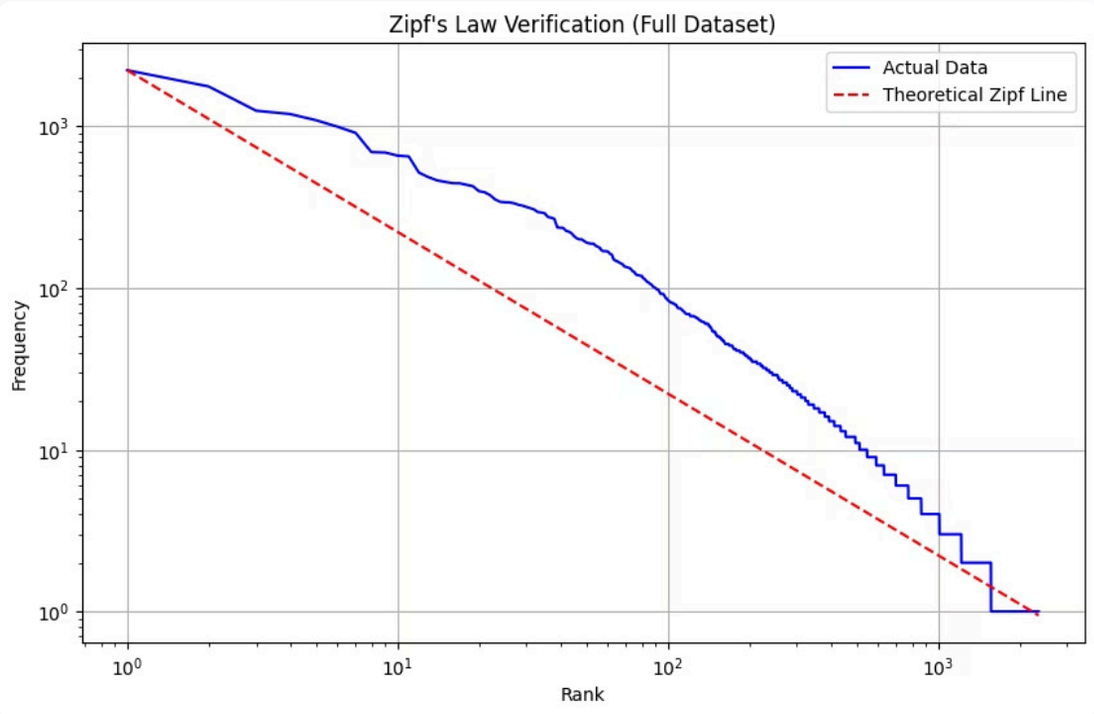
Log-log scale visualization

Theoretical Line

Compare actual data to Zipf's prediction

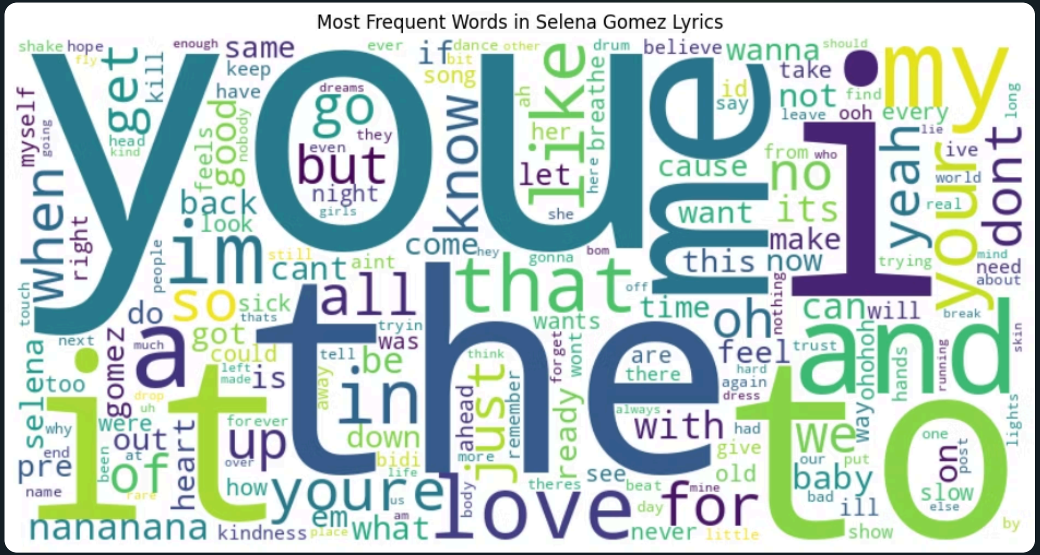
Result

Data closely follows Zipf's law



☁️ Word Cloud Visualization

- A **word cloud** is a visual representation of text data.
- **Larger words** appear more frequently in the lyrics.
- Helps quickly identify **common themes**, emotions, or repeated terms in Selena Gomez's songs.
- Used as a **visual summary** before deeper statistical analysis like Zipf's Law.



Key Insights

- Choruses/refrains inflate specific word frequencies.
- Slope ≈ -0.8 (weaker than natural language).
- Intentional repetition for artistic effect.
- Genre conventions influence word choice.

Conclusion

- **Zipf's Law Applicability:** The law is partially valid in Selena Gomez's song lyrics, showing general trends with some deviations.
- **Limitations:** The analysis is constrained by a relatively small corpus size and the influence of genre-specific lyrical patterns that affect word frequencies.
- **Future Work:** Expanding the study by comparing word frequency distributions across different artists and musical genres to validate and generalize findings.

Thank You

EDA Dashboard

Data Analytics

Data Visualization