Exploratory Data Analysis

Dheeman Kumar **Aryaman Sharma** **Raj Devanshu**

1 Introduction

In this report, we delve into the exploratory data analysis (EDA) conducted on the provided dataset. Our aim is to meticulously examine the dataset's structure, characteristics, and content to derive meaningful insights and inform subsequent phases of analysis. Through a systematic approach, we scrutinize each processing and cleaning step, ensuring data integrity and reliability. Additionally, various statistical analyses and visualization techniques are employed to uncover hidden patterns, relationships, and trends within the dataset. By elucidating the dataset's nuances and intricacies, this EDA lays a solid foundation for informed decision-making and further analysis in subsequent phases.

2 Dataset Overview

The dataset comprises columns including *date*, *likes*, *content*, *username*, *media*, and *inferred company*. An initial inspection of the dataset's structure provides insights into its organization and contents. Understanding the composition and distribution of these columns is crucial for conducting effective exploratory data analysis (EDA). This section serves as a foundation for comprehensively analyzing the dataset and deriving meaningful insights.

3 Data Cleaning and Processing

3.1 Initial Steps

- Removed the *id* column as it was deemed unnecessary for Phase 2 and 3.
- Removed text like `<mention>` and `<hyperlink>` from the *content* column to reduce noise.
- Checked for duplicate rows, but found none.

3.2 Additional Processing

- Separated hashtags from the content and stored them in a new column to highlight tweet topics.
- Extracted the media type from the *media* column and saved it in a separate column.
- Checked for missing values, but none were found.

This comprehensive cleaning and processing workflow ensures data integrity and prepares the dataset for in-depth analysis in subsequent phases.

4 Exploratory Data Analysis

4.1 Univariate Analysis

In this section, we conduct univariate analysis to explore individual variables in the dataset. This includes examining the central tendency, spread, and distribution characteristics of each variable. Visualizations such as histograms and box plots are used to gain insights into the data's distribution and identify potential outliers.

4.1.1 Summary Statistics

- Summary statistics for 'likes' column

Table 1: Summary Statistics for likes

Statistic	Value
Count	300000.000
Mean	773.364
Standard Deviation	4931.463
Minimum	0.000
25% Quantile	3.000
50% Quantile (Median)	76.000
75% Quantile	364.000
Maximum	560193.000

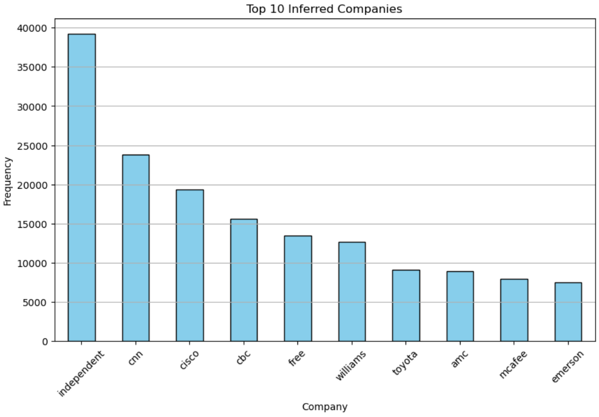

Figure 1: Most frequent Inferred Companies

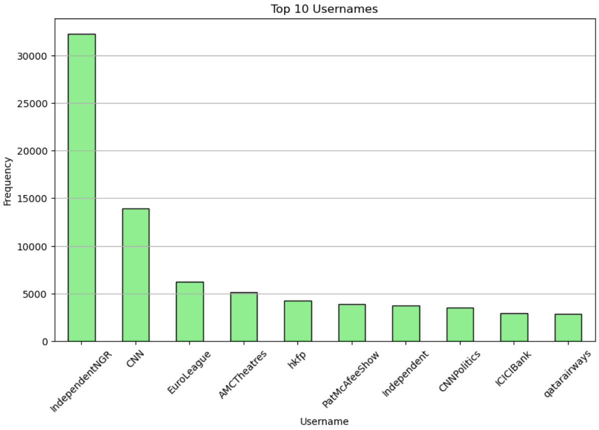

Figure 2: Most frequent Usernames

4.1.2 Frequency Counts

- Frequency counts for top 10 ’inferred company’
- Frequency counts for top 10 ’usernames’
- Frequency counts for ’media type’
- Frequency counts for top 10 ’hashtags’

4.1.3 Temporal Analysis

- Change in number of likes with respect to months in each year

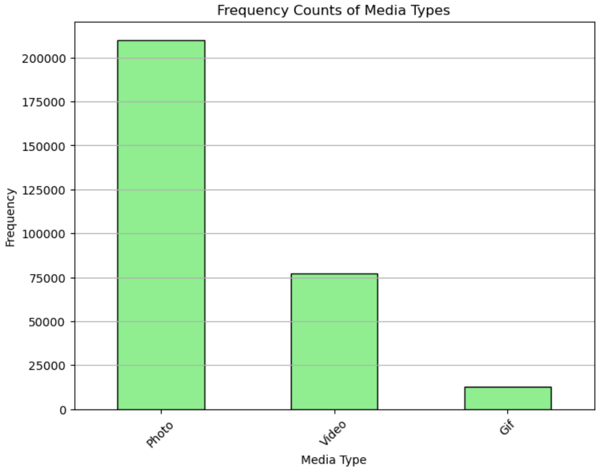

Figure 3: Most frequent Media Types

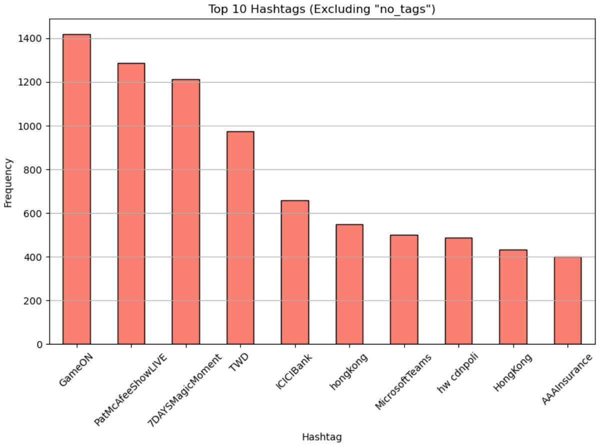

Figure 4: Most frequent Hashtags

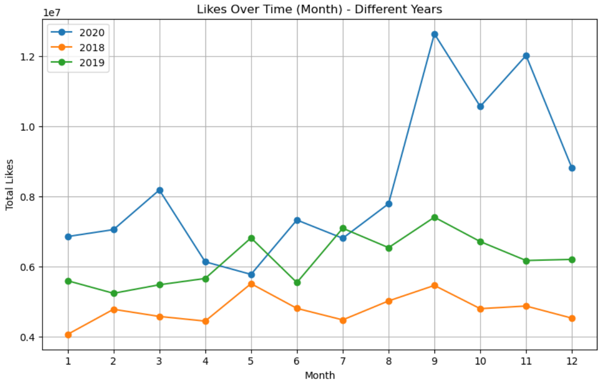

Figure 5: Likes Over Time

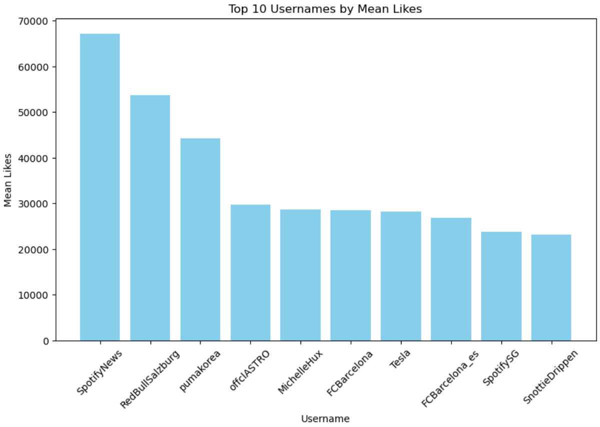

Figure 6: Most Liked Usernames

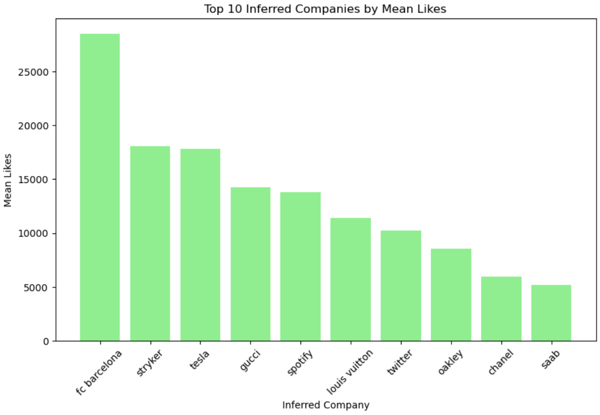

Figure 7: Most Liked Companies

4.2 Multivariate Analysis

This section explores relationships between variables. We examine correlations and patterns to uncover insights into how variables interact within the dataset.

4.2.1 Top 10 Most Liked Tweets

- Provided a table of the top 10 most liked tweets in the dataset

4.2.2 Top 10 Most Liked Usernames

- Provided a table of the top 10 usernames by mean likes

4.2.3 Top 10 Most Liked Companies

- Provided a table of the top 10 inferred companies by mean likes

4.2.4 Content Length Analysis

- Analyzed the relationship between word count and likes

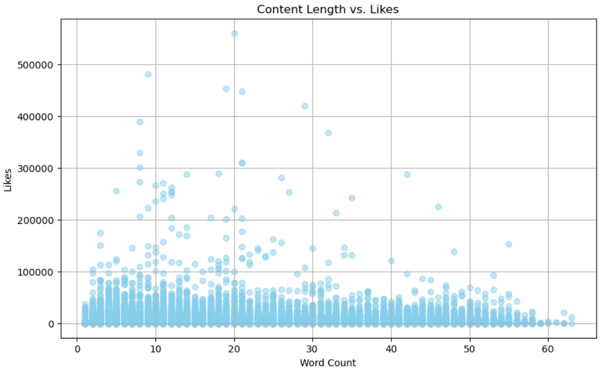

Figure 8: Content Length vs. Likes

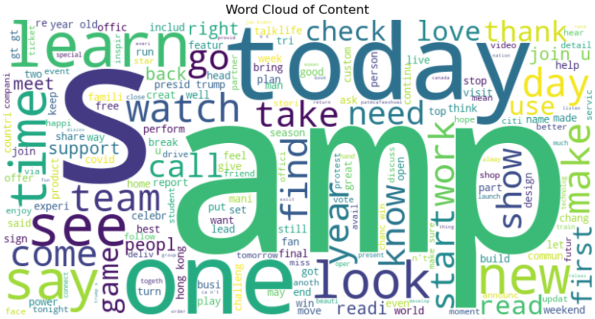

Figure 9: Text Analysis

4.2.5 Text Analysis using NLTK

- Conducted text analysis using the NLTK library

5 Conclusion

In conclusion, our exploratory data analysis (EDA) has provided valuable insights into the dataset. By meticulously cleaning and processing the data, we ensured its integrity and prepared it for in-depth analysis. Through univariate and multivariate analyses, we gained a comprehensive understanding of the dataset's characteristics, including distributions, trends, and relationships between variables. These insights will inform feature selection, preprocessing, and modeling in subsequent phases. By leveraging the knowledge gained from this EDA, we are better equipped to derive meaningful insights and make informed decisions in future stages of analysis and modeling.