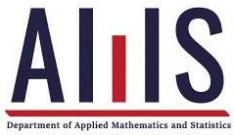
Kingdom of Cambodia

Nation Religion King



Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics

Report of Final Project

Subject : Programming For Data Science

Name	ID	Score
1. KHUN Limchheang	e20230393	…….
2. CHHAY Lyveng	e20230135	…….

Lecturer: Mr. OL Say (Course)

Mr. MIN Sothearith (TP)

Academic year 2025 - 2026

Contents

Group 9: Quote Analysis

Group Members: KHUN Limchheang and CHHAY Lyveng

---

1 Introduction

1.1 Background and Motivation

Quotes are widely used to express ideas, inspiration, and wisdom across different themes such as life, love, and knowledge. With the increasing availability of online textual data, analyzing quotes provides an opportunity to explore patterns in authorship, text structure, and thematic categorization. This project applies web scraping and data analysis techniques to study a real-world quote dataset.

1.2 Objectives

The main objectives of this project are:

- To collect quote data from a publicly available website using web scraping techniques
- To clean and preprocess the scraped data
- To engineer new features such as quote length and tag count
- To perform exploratory data analysis to identify patterns related to authors and themes

1.3 Dataset Description

- Website: `http://quotes.toscrape.com`
- Dataset size: Approximately 100 quotes
- Variables collected: Quote, Author, Tags
- Derived variables: Quote Length, Tag Count

2 Data Collection

Data was collected from the website *quotes.toscrape.com* using Python libraries such as `requests` and `BeautifulSoup`. The website contains multiple pages of quotes, and pagination was handled automatically by detecting the "Next" button.

To ensure ethical data collection, only publicly available data was scraped and a delay was added between requests to avoid overloading the server. The collected data was stored as raw data in CSV format before any cleaning or analysis was performed.

3 Data Cleaning and Feature Engineering

3.1 Data Cleaning

The raw dataset was examined for inconsistencies and missing values. The following cleaning steps were applied:

- Removal of unnecessary quotation marks from quote text
- Trimming whitespace from author names
- Removing records with missing tag information

These steps ensured data quality by removing incomplete records while preserving consistency across the dataset.

3.2 Feature Engineering

Some variables required for analysis were not explicitly available on the website. Therefore, feature engineering was applied:

- Quote Length: calculated as the number of characters in each quote
- Tag Count: calculated as the number of tags associated with each quote

These derived features enable deeper analysis of text structure and thematic categorization.

4 Descriptive Statistics

4.1 Quote Length

The average quote length in the dataset is approximately 120 characters.

Interpretation: This indicates that most quotes are relatively concise, which aligns with the purpose of quotes as short expressions of ideas or wisdom.

4.2 Authors with the Most Quotes

The analysis shows that a small number of authors contribute multiple quotes. Authors such as Albert Einstein and J.K. Rowling appear more frequently than others.

Interpretation: This suggests that the dataset emphasizes well-known authors whose quotes are commonly shared.

4.3 Tag Distribution

Tags such as *life*, *love*, and *inspiration* appear most frequently.

Interpretation: This reflects the thematic focus of the website on universal and relatable topics.

5 Data Visualization and Exploratory Data Analysis (EDA)

Data visualization is used in this project to better understand patterns and relationships within the quote dataset. While descriptive statistics provide numerical summaries, visualizations help reveal distributions, comparisons, and potential relationships more clearly.

5.1 Distribution of Quote Length

Objective: This visualization examines how quote lengths are distributed across the dataset.

Reason for visualization choice: A histogram is appropriate for understanding the distribution and spread of numerical data such as quote length.

Key observations: Most quotes are relatively short, with the majority clustered around lower character counts. A small number of longer quotes appear as outliers, indicating that lengthy quotes are less common.

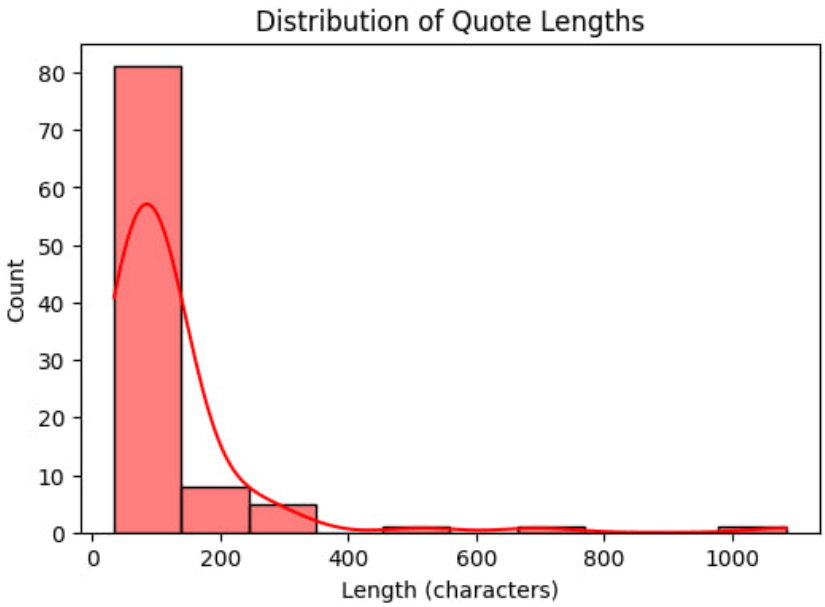


Figure 1: Distribution of Quote Length

5.2 Distribution of Tag Count

Objective: This figure shows how many tags are typically assigned to each quote.

Reason for visualization choice: A bar chart effectively displays the frequency of discrete values such as the number of tags.

Key observations: Most quotes contain between one and three tags. Very few quotes have a large number of tags, indicating that tags are applied selectively rather than extensively.

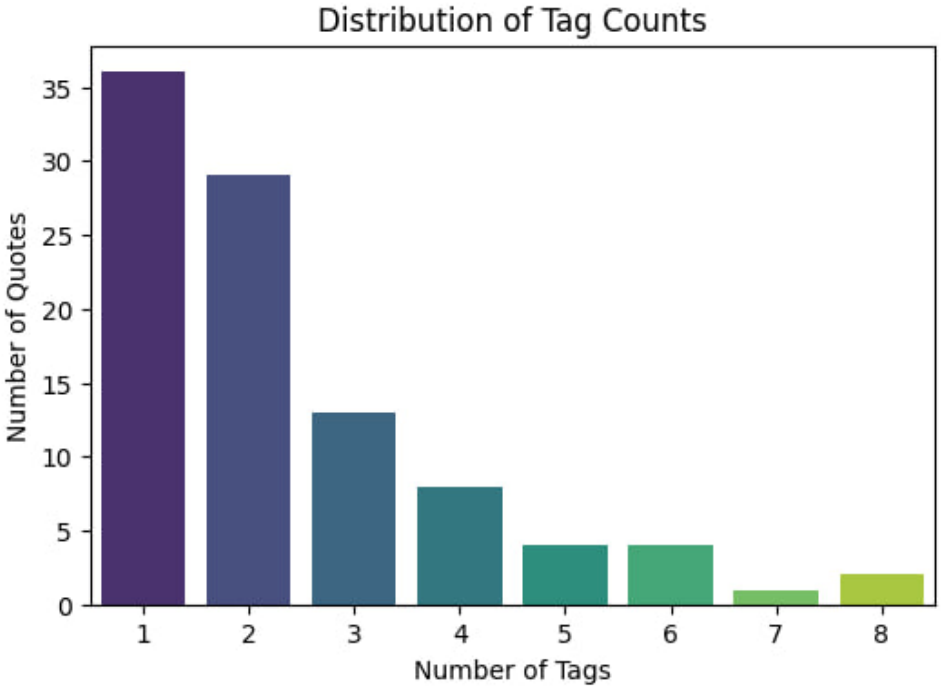


Figure 2: Distribution of Tag Count per Quote

5.3 Authors with the Most Quotes

Objective: This visualization identifies authors who appear most frequently in the dataset.

Reason for visualization choice: A bar chart is suitable for comparing the frequency of categorical variables such as author names.

Key observations: A small number of authors contribute a disproportionately large number of quotes. Authors such as Albert Einstein and J.K. Rowling appear most frequently, while many authors appear only once or twice.

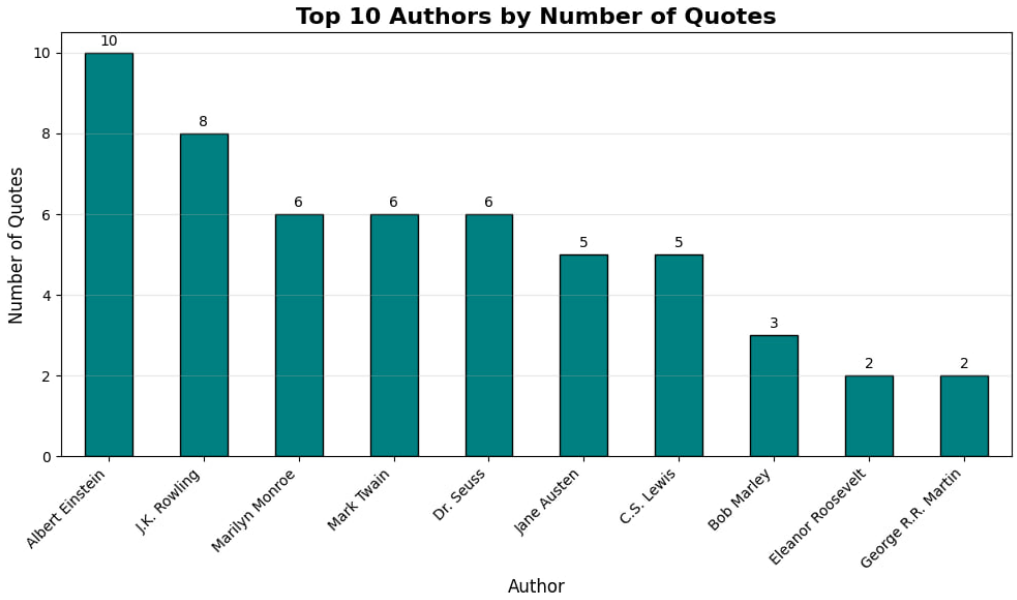


Figure 3: Top Authors by Number of Quotes

5.4 Authors with the Longest Average Quote Length

Objective: This visualization identifies authors whose quotes are longest on average, based on the mean quote length per author.

Reason for visualization choice: A horizontal bar chart is suitable for comparing average values across multiple authors and allows author names to be displayed clearly.

Key observations: Authors such as Pablo Neruda and Bob Marley tend to have longer quotes on average, while other authors show shorter average quote lengths. This suggests differences in writing style, with some authors favoring more elaborate expressions.

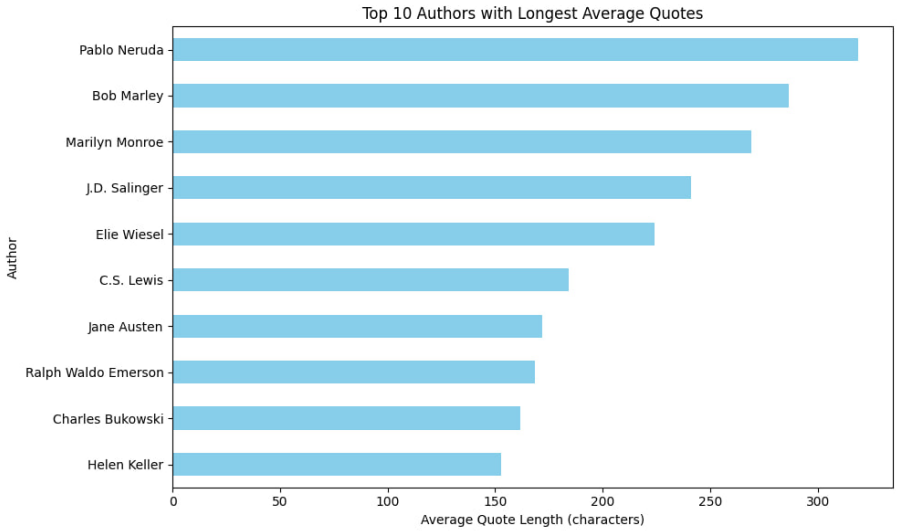


Figure 4: Top 10 Authors with the Longest Average Quote Length

5.5 Most Common Tags

Objective: This visualization highlights the most frequently occurring tags in the dataset.

Reason for visualization choice: A bar chart allows easy comparison of tag frequencies and helps identify dominant themes.

Key observations: Tags such as *life*, *love*, and *inspiration* appear most frequently. This suggests that the dataset focuses on universal and broadly relatable themes.

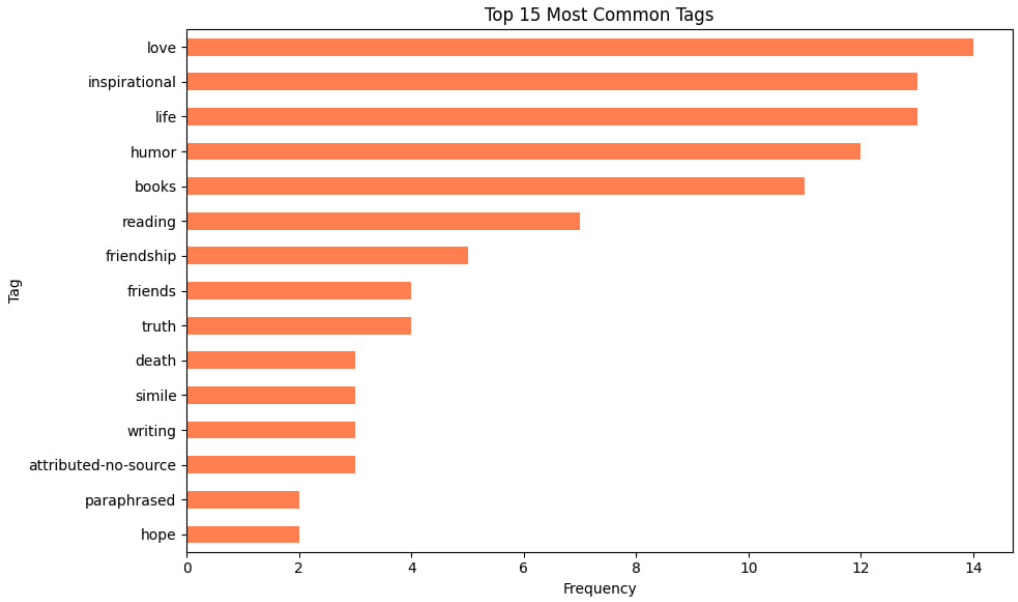


Figure 5: Most Common Tags in the Dataset

5.6 Cumulative Distribution of Quote Lengths

Objective: This visualization shows how quote lengths increase when quotes are sorted from shortest to longest. It helps understand how quickly the dataset moves from short quotes to very long quotes.

Reason for visualization choice: A cumulative (sorted) line plot is useful for observing the overall growth pattern and identifying whether the dataset contains extreme outliers. The median reference line provides a clear midpoint comparison.

Key observations: The curve increases gradually for most quotes and rises sharply near the end, indicating that a small number of quotes are much longer than the rest. The median length (shown by the dashed line) suggests that at least half of the quotes are below the median value.

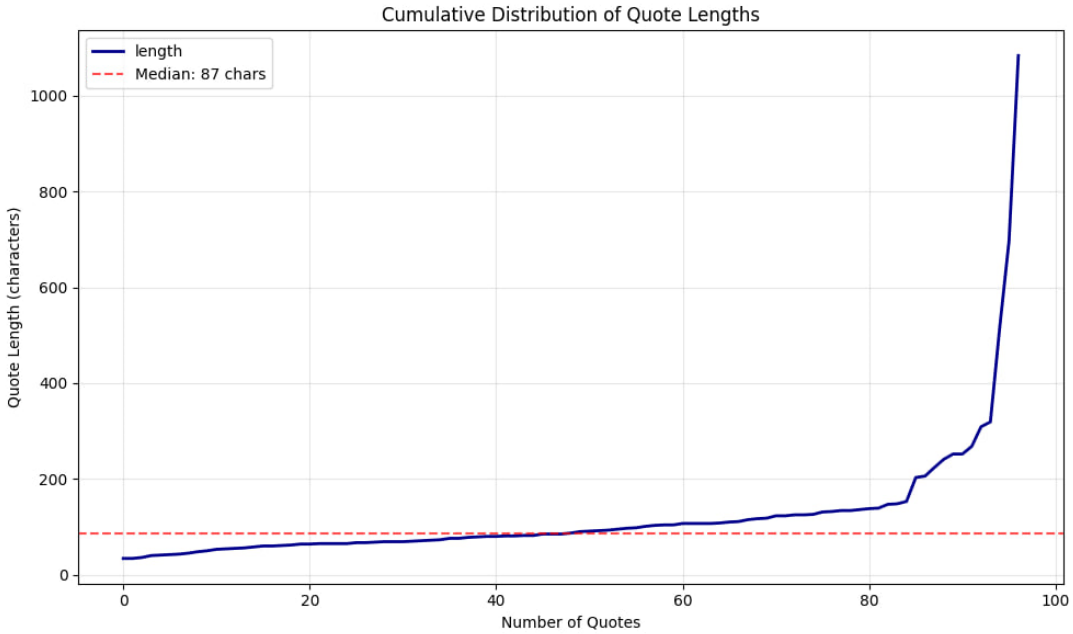


Figure 6: Cumulative Distribution of Quote Lengths with Median Reference

5.7 Correlation Between Quote Length and Tag Count

Objective: This visualization evaluates whether quote length is correlated with the number of tags assigned to a quote.

Reason for visualization choice: A correlation heatmap provides a clear and interpretable summary of linear relationships between numerical variables. It highlights both the direction and strength of correlation.

Key observations: The correlation between quote length and tag count is low (approximately 0.12), indicating a weak positive relationship. This suggests that longer quotes do not strongly imply more tags, and tag assignment may depend more on content theme than quote length.

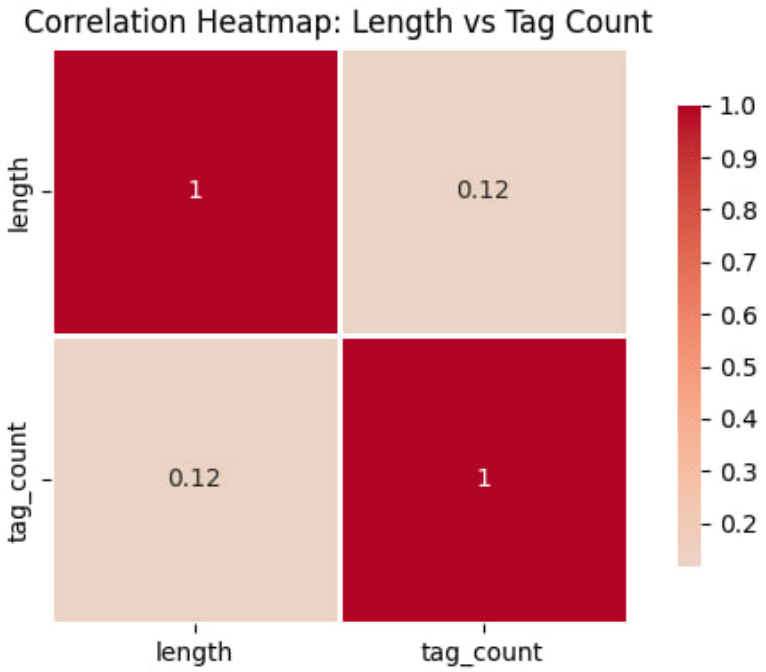


Figure 7: Correlation Heatmap: Quote Length vs Tag Count

6 Conclusion

This project successfully demonstrated the complete workflow of web scraping and data analysis using a real-world textual dataset. Quotes were collected from a publicly available website, cleaned, and transformed through feature engineering to enable meaningful analysis. Key features such as quote length and tag count were derived to support exploratory data analysis.

The analysis revealed that most quotes are relatively concise, with only a small number of long quotes acting as outliers. A limited number of authors contributed multiple quotes, indicating that the dataset is dominated by well-known figures. Additionally, thematic tags were found to focus mainly on universal topics such as life and love, and they were applied selectively rather than extensively.

Overall, the project met its objectives and demonstrated how web scraping combined with exploratory data analysis can provide valuable insights into unstructured text data. The results highlight patterns in authorship, text structure, and thematic categorization while reinforcing the importance of feature engineering in data analysis.

7 Limitations

Despite achieving the project objectives, several limitations should be acknowledged. First, the dataset is limited to a single website, which restricts the diversity and generalizability of the findings. The number of available quotes is relatively small, which may limit the strength of statistical conclusions.

Second, the tags associated with each quote are subjective and manually assigned by humans. As a result, tag usage may be inconsistent and influenced by personal interpretation rather than objective criteria. Finally, the analysis focuses primarily on descriptive and exploratory methods and does not include deeper linguistic or semantic analysis of the quote content.

8 Future Work

Future work could expand this project in several ways. Additional data could be collected from multiple quote websites to increase dataset size and improve representativeness. Natural language processing techniques, such as sentiment analysis or topic modeling, could be applied to extract deeper insights from the quote text.

Furthermore, machine learning methods could be explored to automatically classify quotes by theme or author style. More advanced statistical analysis could also be conducted to examine relationships between textual features and thematic tags. These extensions would enhance the analytical depth and practical applications of the project.

9 References

- Quotes to Scrape Website
- Project GitHub Repository