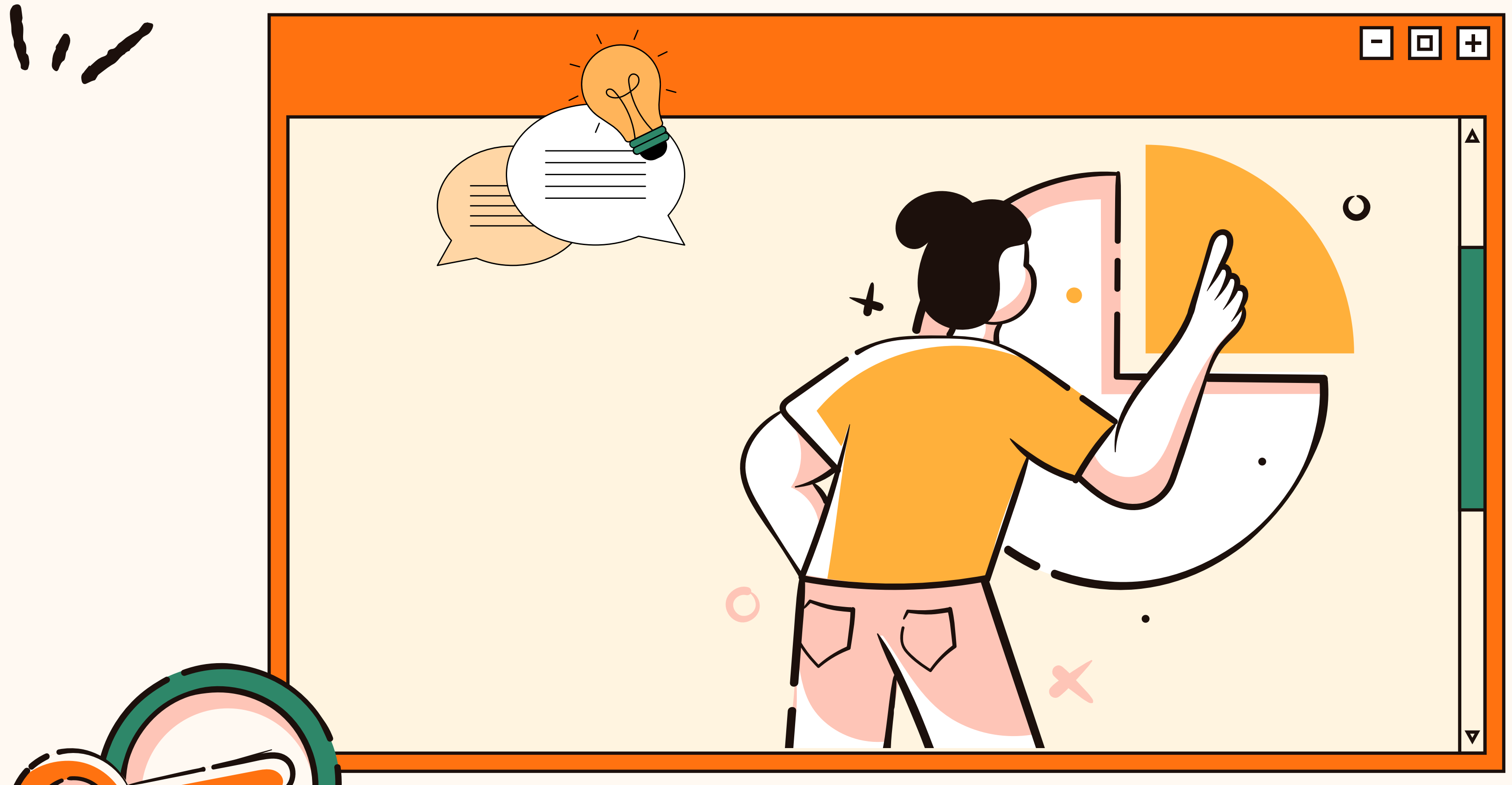
Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics

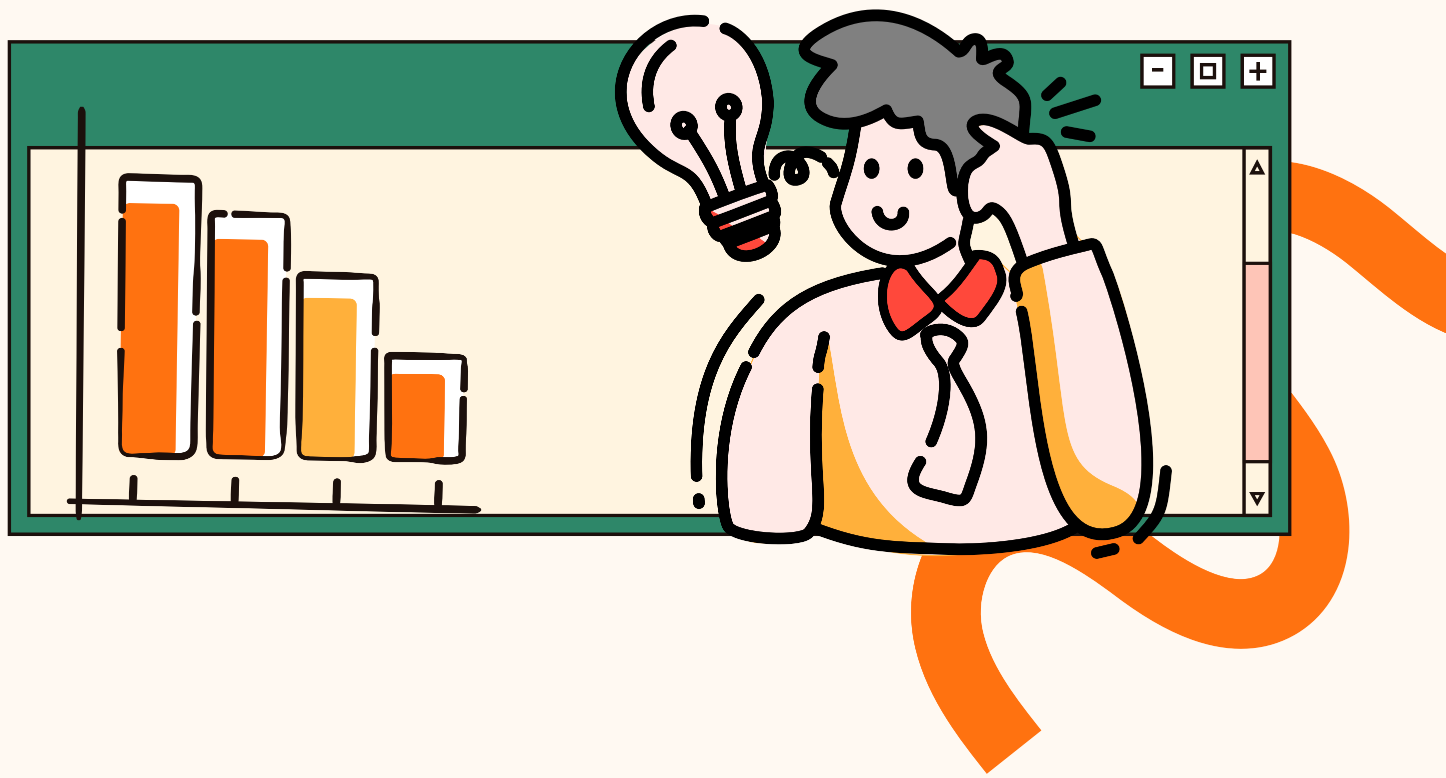
Quotes Analysis



Lecturer:

Course: **Mr. OL Say**

Instructor: **Mr. Min Sothearith**



Group members:

KHUN Limchheang	e20230393
CHHAY Lyveng	e20230135

Group members

KHUN Limchheang
e20230393

Chhay Lyveng
e20230135

Contents

1 Project Overview

2 Web Scraping Process

3 Data Description

4 Feature Engineering and Data Cleaning

5 Descriptive Statistics

6 Exploratory Data Analysis & Visualizations

7 Conclusion

8 Limitations & Future Work

1. Project Overview

Objective: Web scraping and data analysis of quotes from quotes.toscrape.com

Tools Used: Python (BeautifulSoup, Pandas, Matplotlib, Seaborn)

Process: Scraping → Feature Engineering → Cleaning → Visualization

2. Web Scraping Process

Source: http://quotes.toscrape.com

Method: Automated pagination scraping using BeautifulSoup

Data **Extracted** per Quote:

- Quote text
- Author name
- List of tags

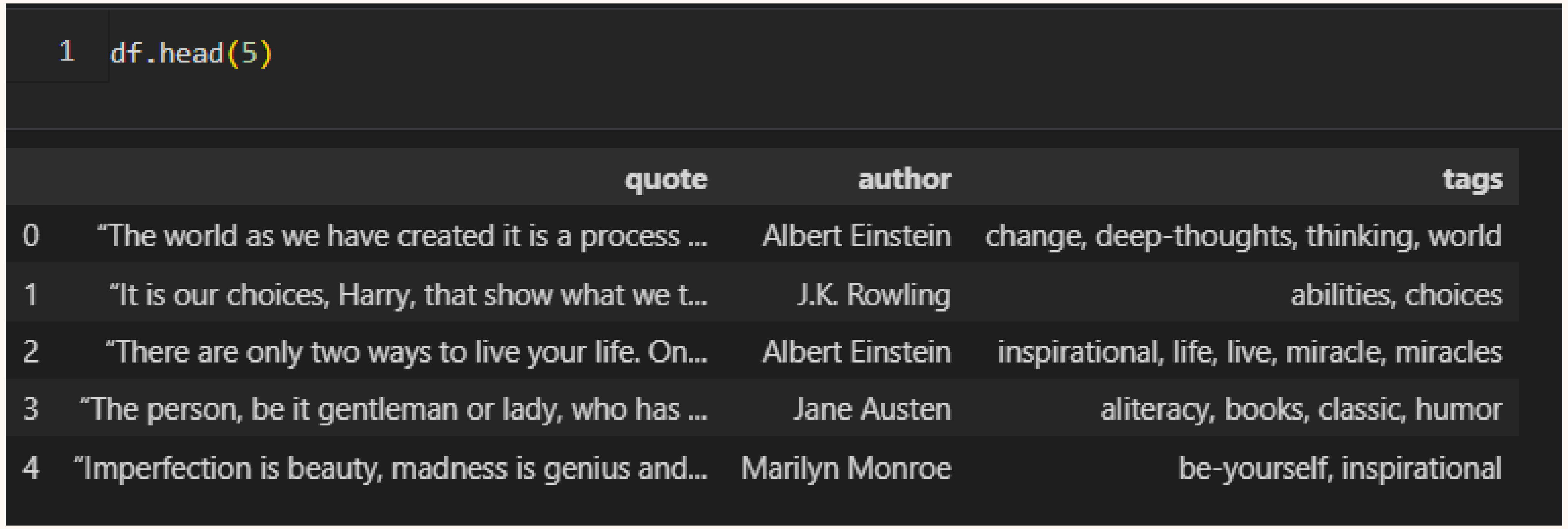
Output: CSV file containing 100 quotes

3. Data Description

Dataset: quotes_raw.csv

Columns:

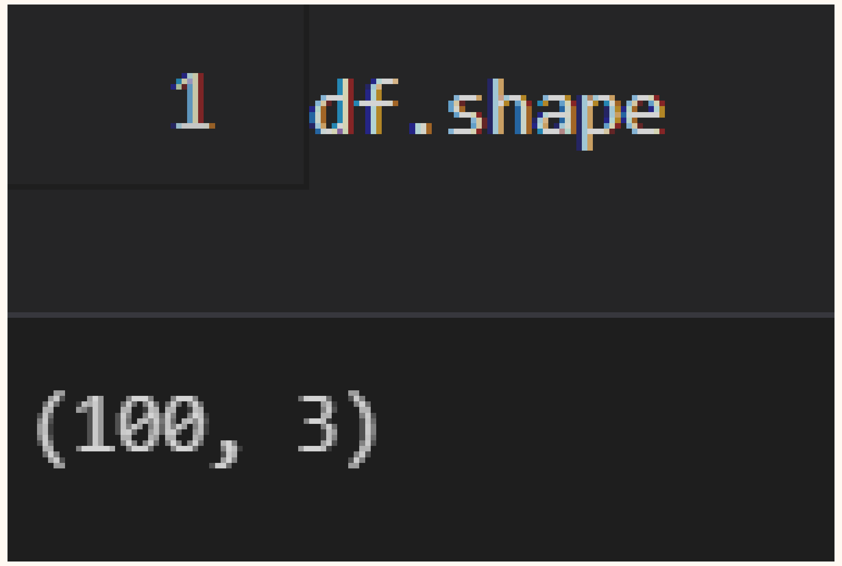
- quote: Text of the quote
- author: Name of the author
- tags: Comma-separated list of tags



```
df.head(5)
```

	quote	author	tags
0	“The world as we have created it is a process ...	Albert Einstein	change, deep-thoughts, thinking, world
1	“It is our choices, Harry, that show what we t...	J.K. Rowling	abilities, choices
2	“There are only two ways to live your life. On...	Albert Einstein	inspirational, life, live, miracle, miracles
3	“The person, be it gentleman or lady, who has ...	Jane Austen	aliteracy, books, classic, humor
4	“Imperfection is beauty, madness is genius and...	Marilyn Monroe	be-yourself, inspirational

Initial Size: 100 rows x 3 columns

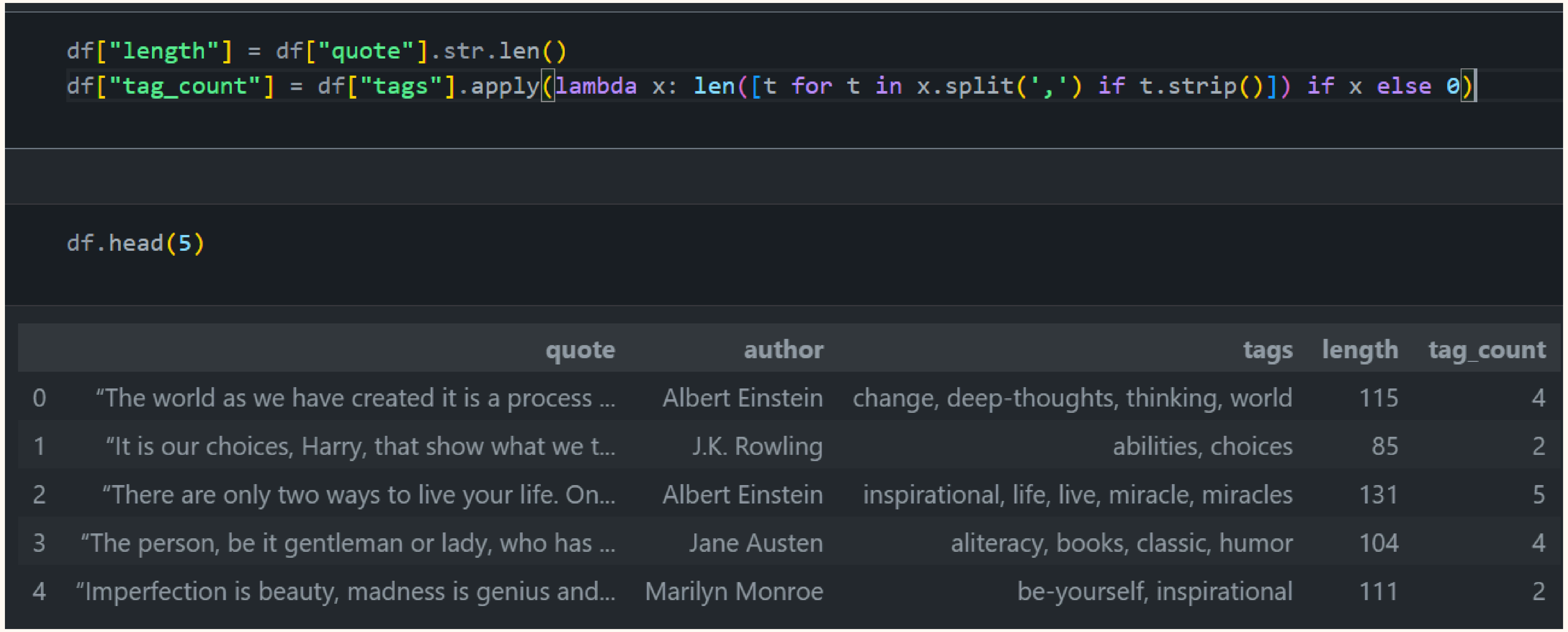


```
df.shape
```

```
(100, 3)
```

4. Feature Engineering

- length: Character count of each quote
- tag_count: Number of tags per quote



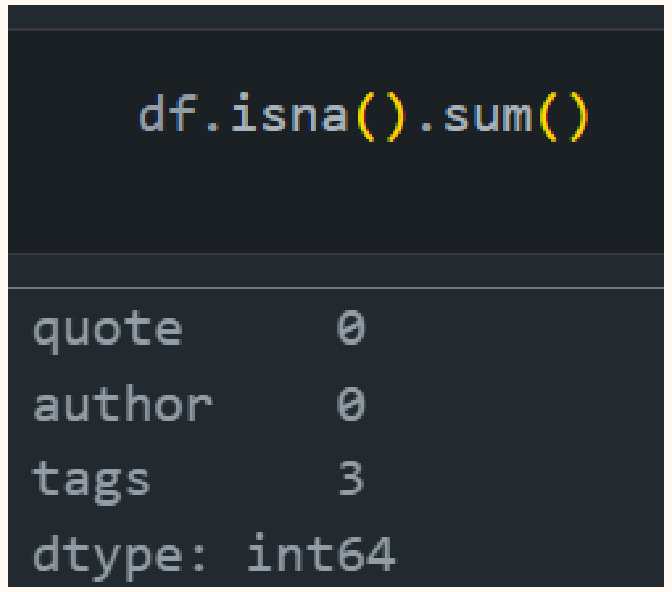
```
df["length"] = df["quote"].str.len()
df["tag_count"] = df["tags"].apply(lambda x: len([t for t in x.split(',') if t.strip()]) if x else 0)
```

```
df.head(5)
```

	quote	author	tags	length	tag_count
0	“The world as we have created it is a process ...	Albert Einstein	change, deep-thoughts, thinking, world	115	4
1	“It is our choices, Harry, that show what we t...	J.K. Rowling	abilities, choices	85	2
2	“There are only two ways to live your life. On...	Albert Einstein	inspirational, life, live, miracle, miracles	131	5
3	“The person, be it gentleman or lady, who has ...	Jane Austen	aliteracy, books, classic, humor	104	4
4	“Imperfection is beauty, madness is genius and...	Marilyn Monroe	be-yourself, inspirational	111	2

4. Data Cleaning

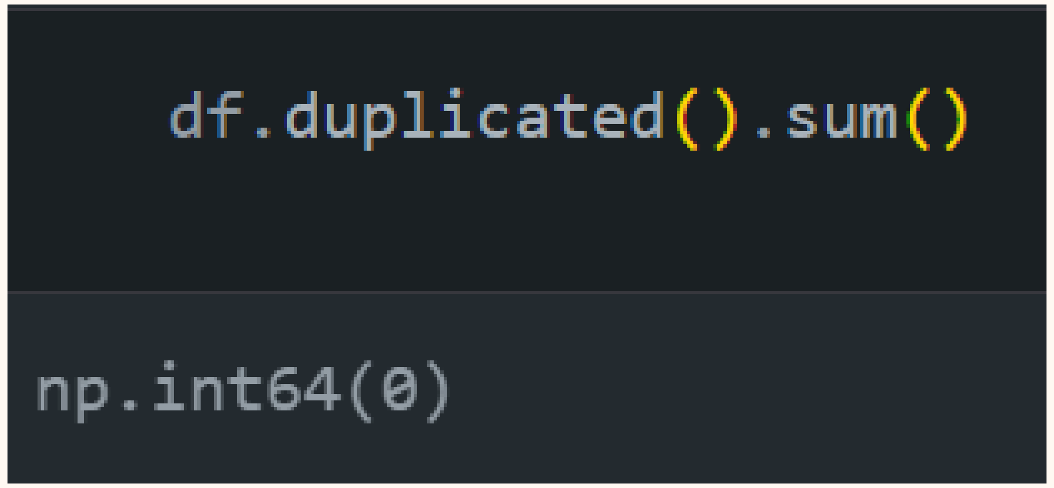
Missing Values Check: 3 missing values found **(Drop)**



```
df.isna().sum()
```

```
quote      0
author     0
tags       3
dtype: int64
```

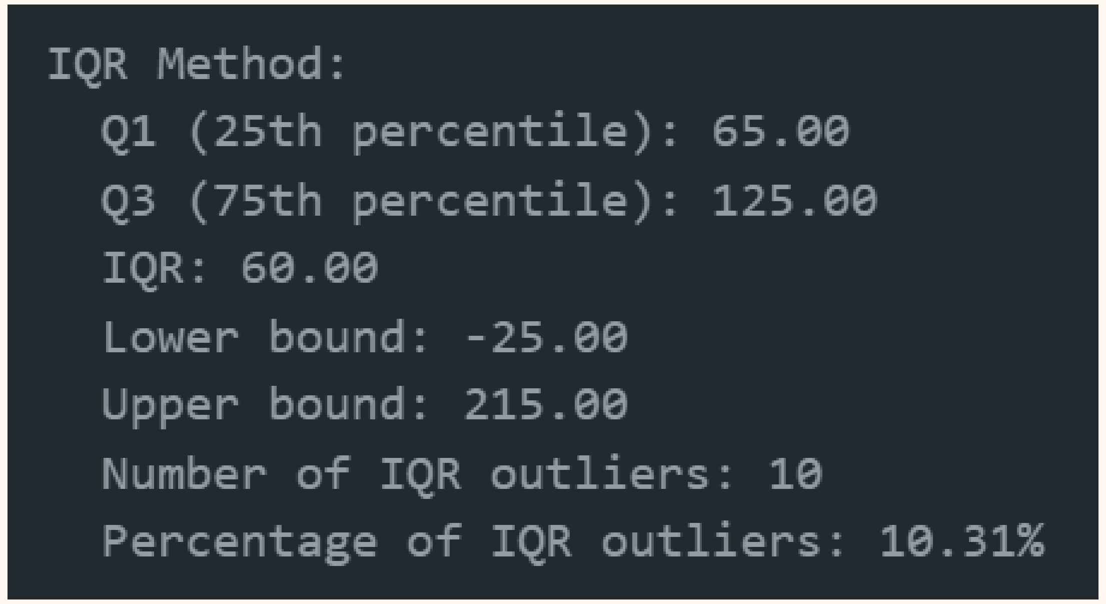
Duplicate Removal: No duplicates detected



```
df.duplicated().sum()
```

```
np.int64(0)
```

Outlier Detection: **IQR method** applied to quote lengths



```
IQR Method:
  Q1 (25th percentile): 65.00
  Q3 (75th percentile): 125.00
  IQR: 60.00
  Lower bound: -25.00
  Upper bound: 215.00
  Number of IQR outliers: 10
  Percentage of IQR outliers: 10.31%
```

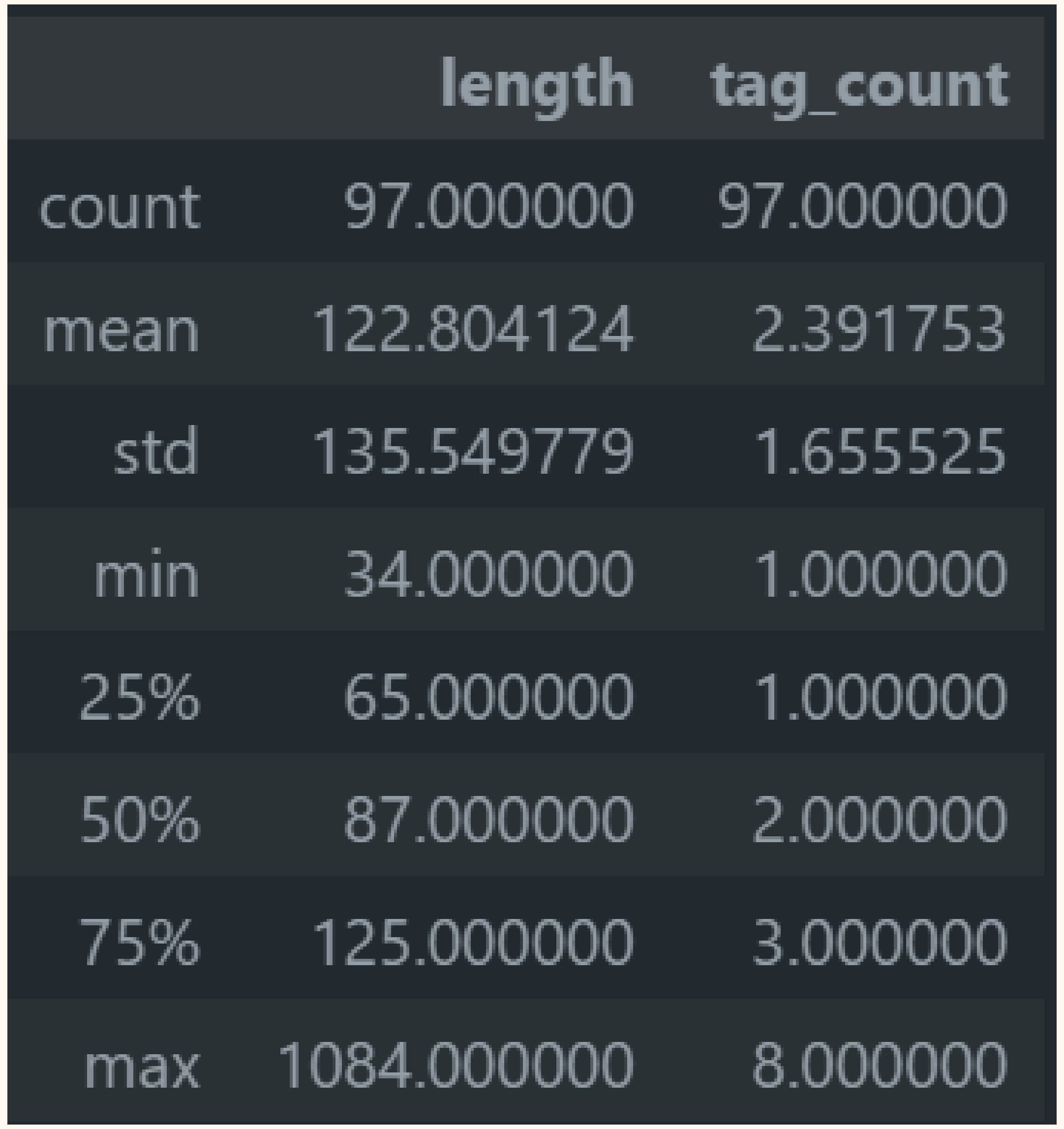
5. Descriptive Statistics

Quote Length (characters):

- Mean: ~122 chars
- Min/Max: Varies widely

Tags per Quote:

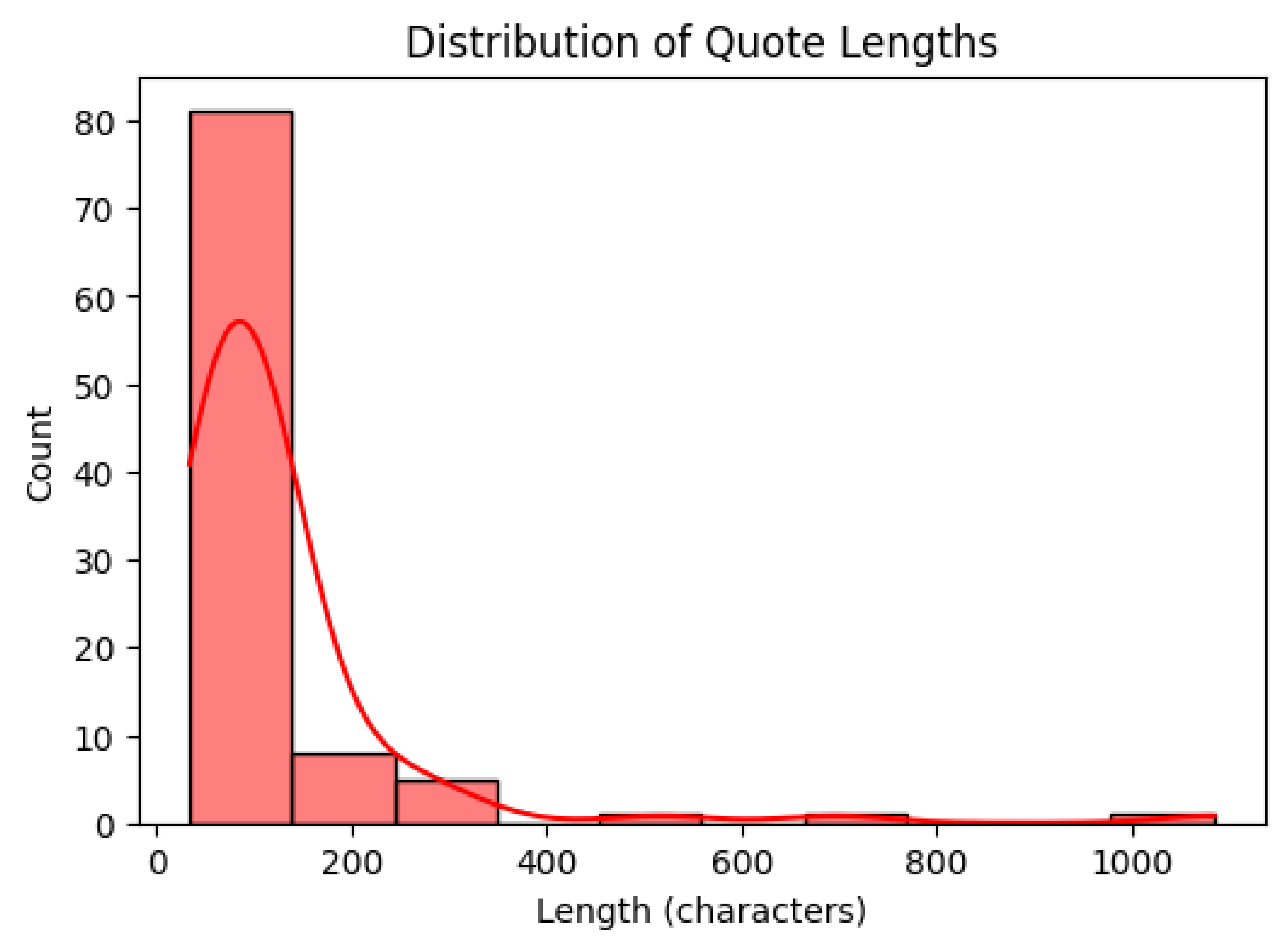
- Average: ~2 tags per quote
- Range: 1 to 8 tags



	length	tag_count
count	97.000000	97.000000
mean	122.804124	2.391753
std	135.549779	1.655525
min	34.000000	1.000000
25%	65.000000	1.000000
50%	87.000000	2.000000
75%	125.000000	3.000000
max	1084.000000	8.000000

6. Exploratory Data Analysis & Visualizations

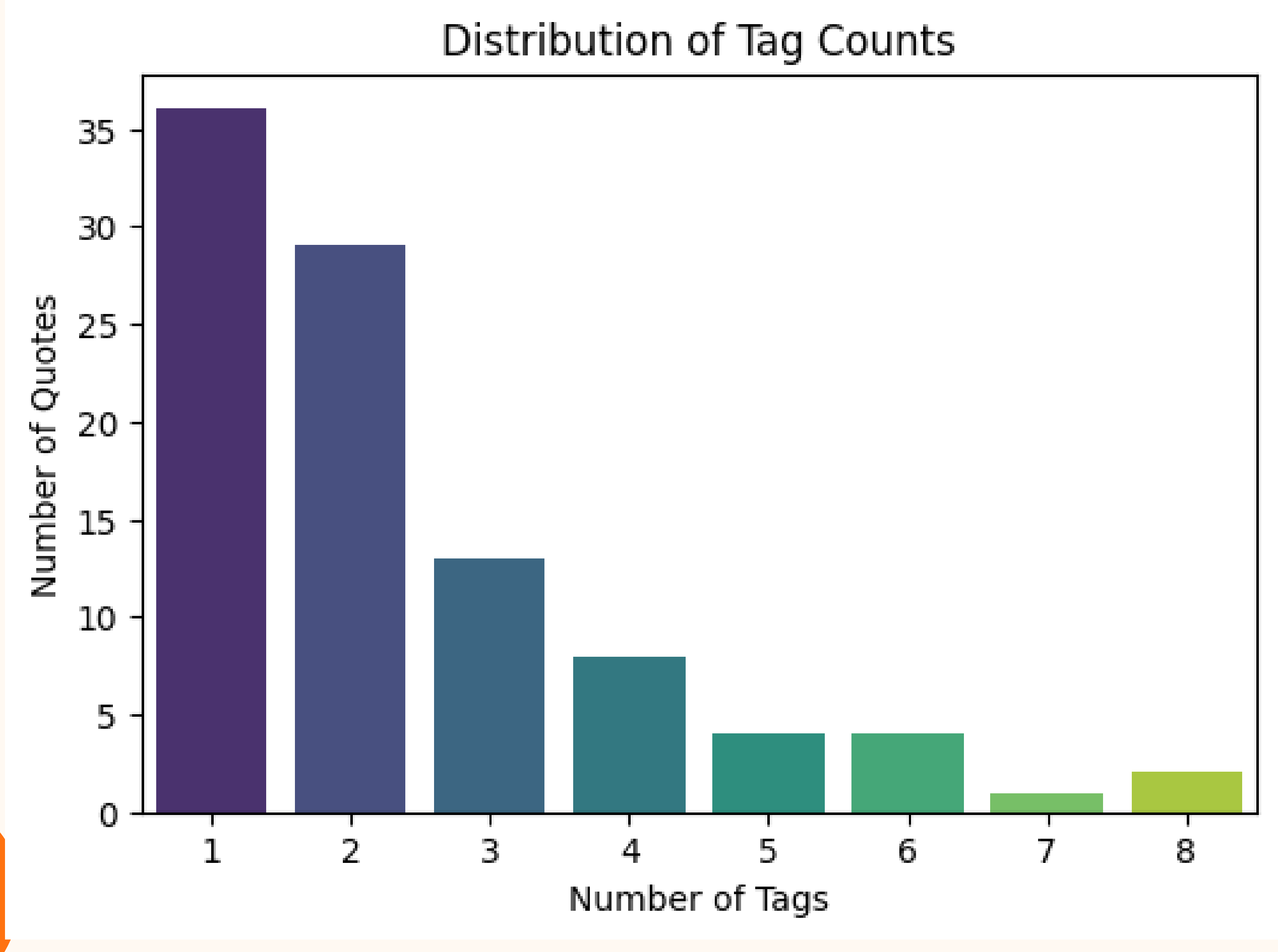
Quote Length Distribution:

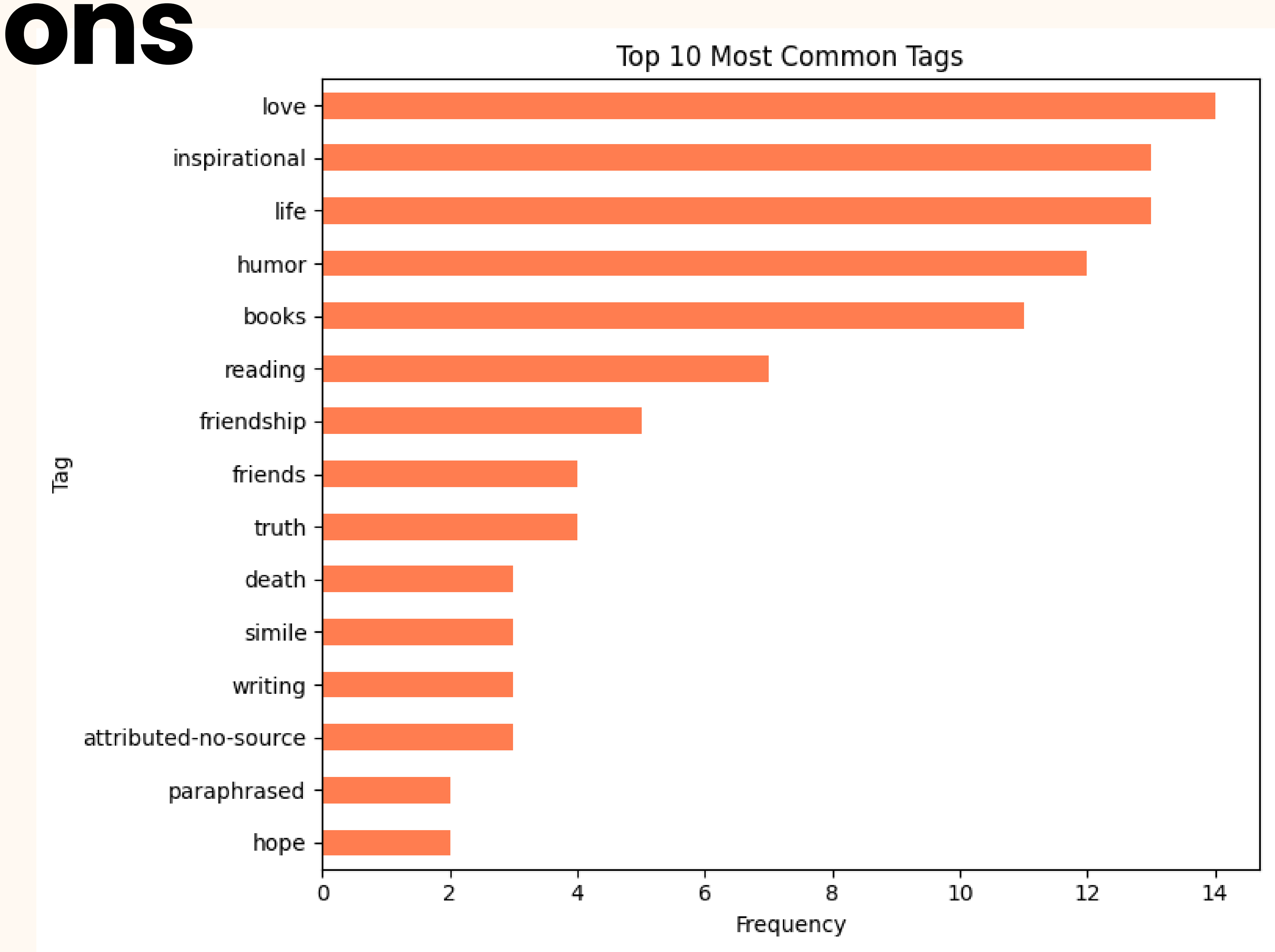


- Most quotes are 70-100 characters
- Right skewed

6. Exploratory Data Analysis & Visualizations

Tag Frequency:

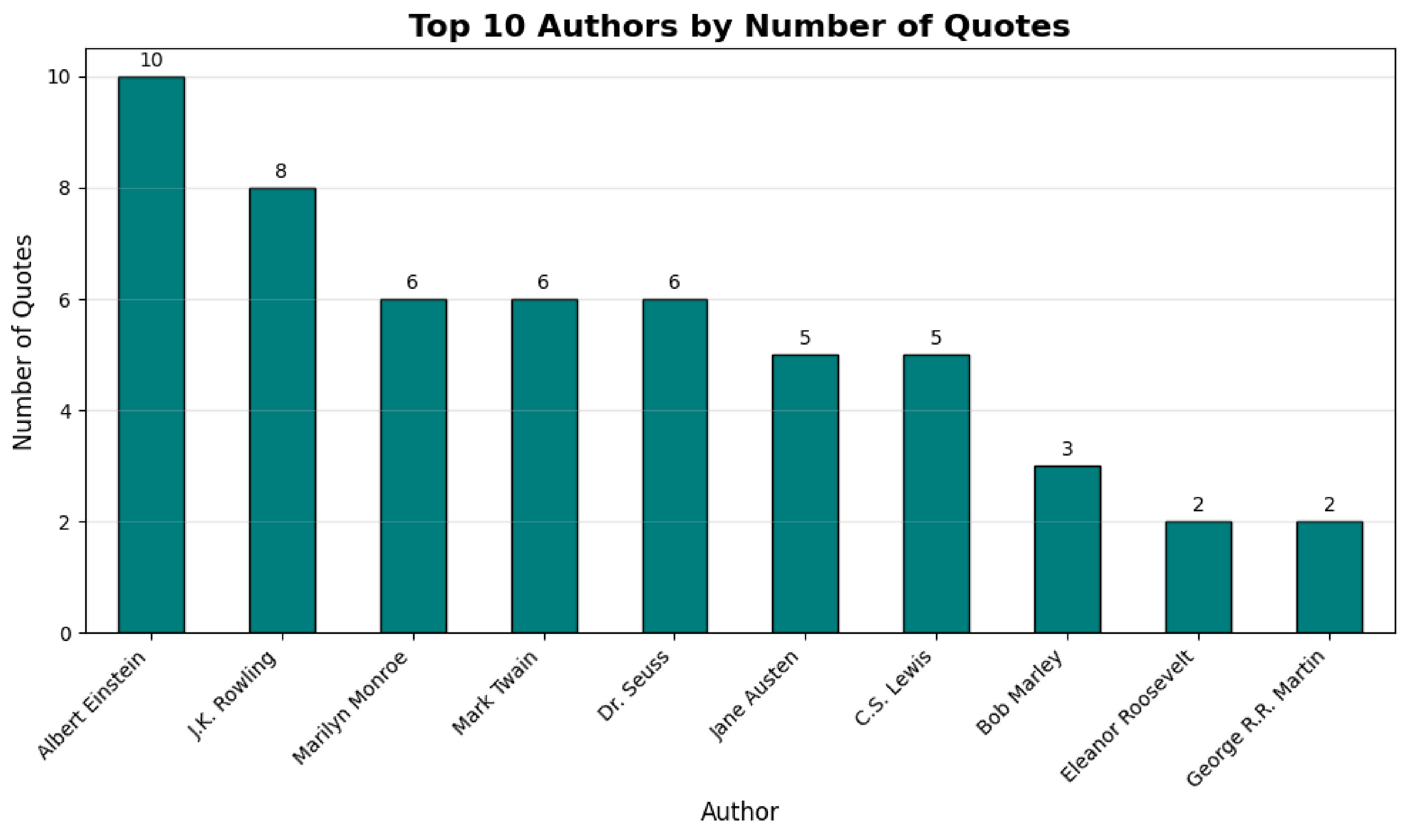


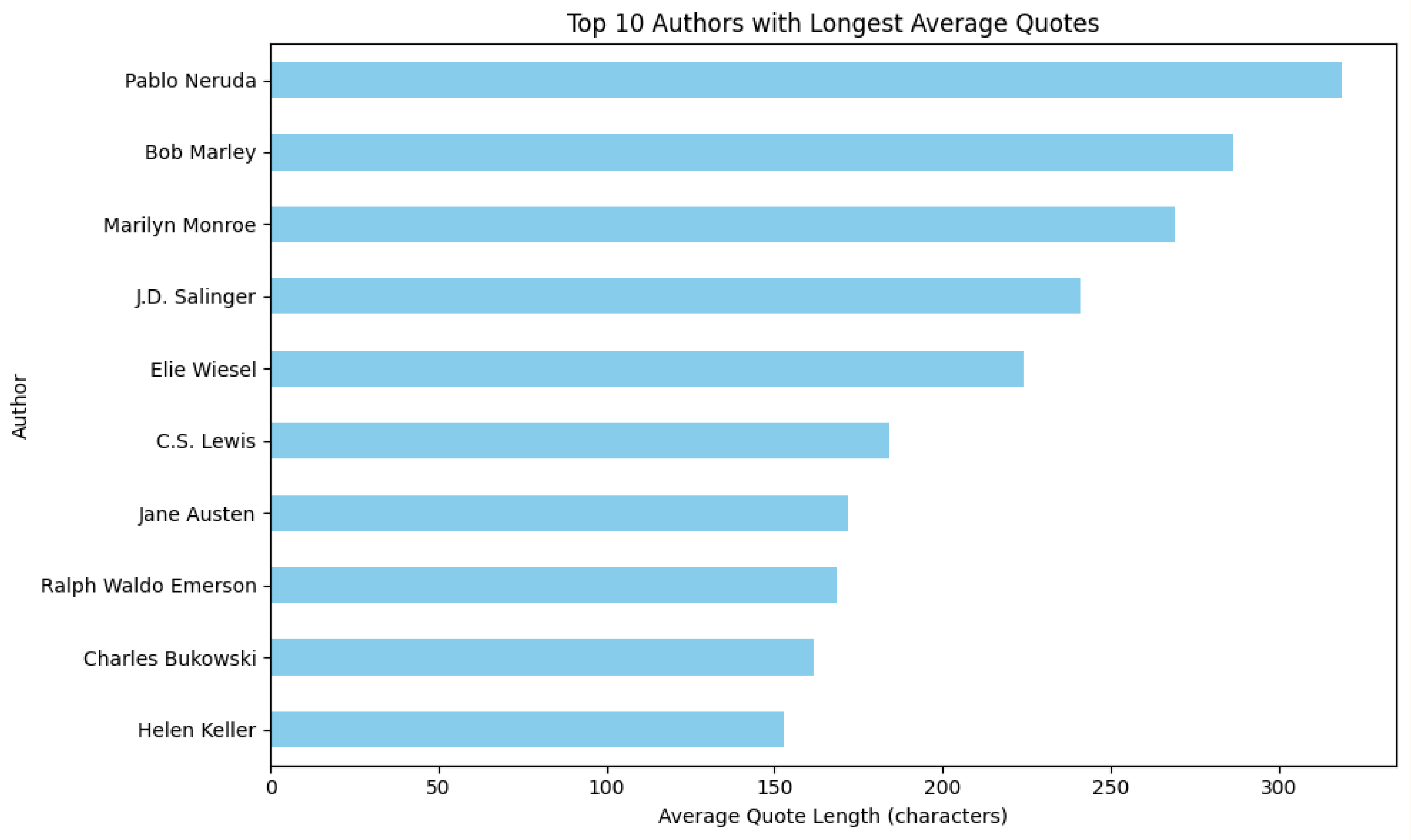


- Majority of quotes have 1-2 tags
- "love", "Inspirational" and "life" are most common tags

6. Exploratory Data Analysis & Visualizations

Author Analysis:

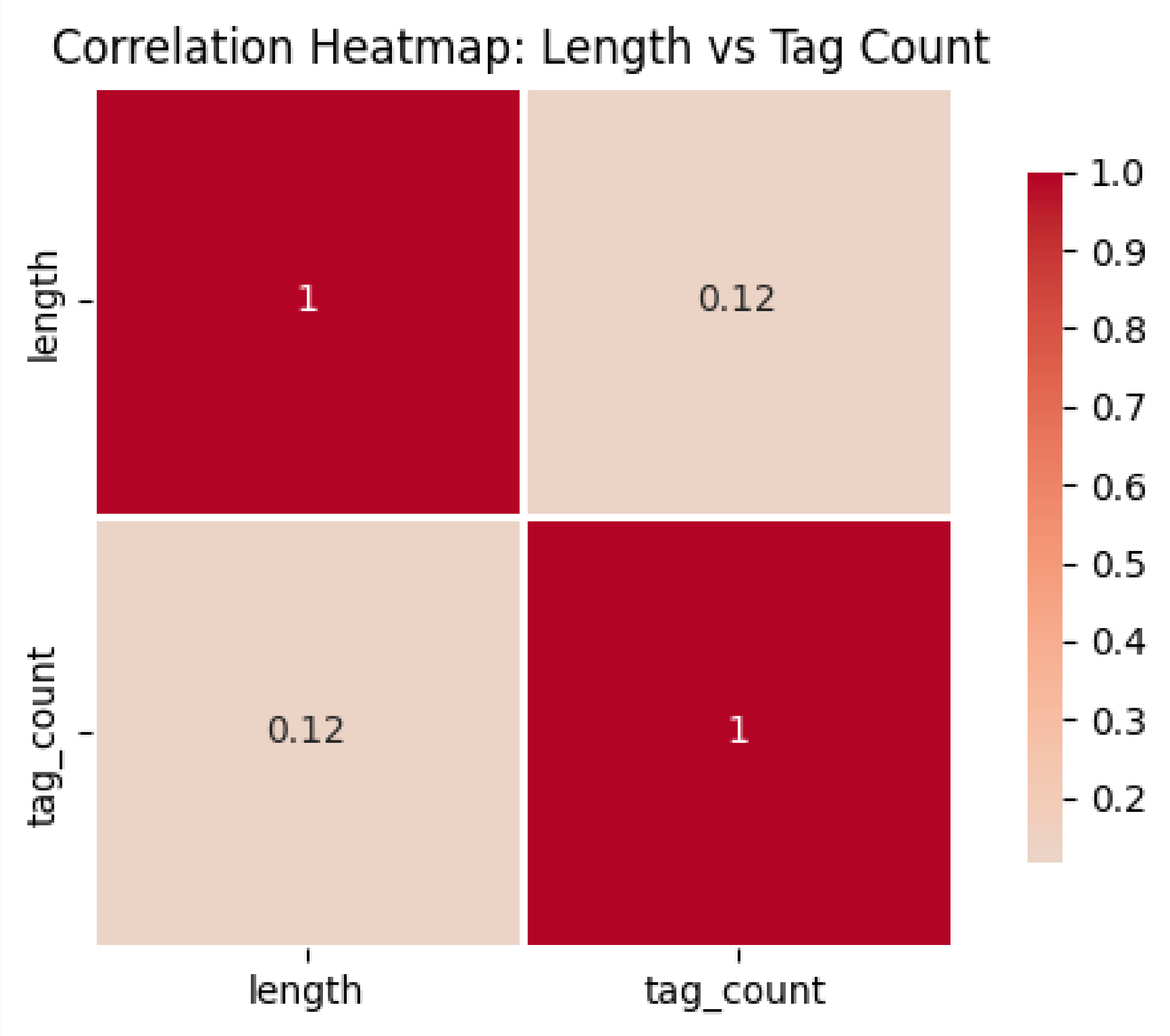




- Albert Einstein has the most quotes with 10 quotes
- Some authors consistently write longer quotes

6. Exploratory Data Analysis & Visualizations

Correlation:



- **Weak positive** correlation between **quote length** and **tag count**
- **Longer** quotes tend to have **higher** tags

7. Conclusion

- **Successfully** web scraping → Feature Engineering→ cleaning → EDA workflow
- **Quotes** are generally concise, authored by well-known figures, and tagged with universal themes
- **Feature engineering** (quote length, tag count) enabled meaningful pattern discovery
- **Highlights the value** of combining web scraping with EDA for text data insights

8. Limitations & Future Work

Limitations:

Single website source, small dataset, subjective tagging, no deep semantic analysis

Future Work:

- Collect more data from multiple sources
- Apply NLP (sentiment analysis, topic modeling)
- Use machine learning for theme/author classification
- Conduct advanced statistical analysis

Thank You!