MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Department Of Computer Engineering

B.TECH SEM-II

NAME: Sujal Ashok Pimparkar

DATA SET: Yelp Reviews

DIVISION: CS5

BATCH: C54

ROLL NO: 81

PRN NO: 202401100048

Under the Guidance of Course In-charge,

Prof. Priyanka Mane

1.Find the total number of reviews in the dataset.

```
[1] import pandas as pd
    import numpy as np

df = pd.read_csv('/content/sample_data/yelp.csv.zip')
num_reviews = df.shape[0]
print("Total Reviews:", num_reviews)

Total Reviews: 10000
```

2.Find the average star rating of all reviews.

```
# 1. Total number of reviews
num_reviews = df.shape[0]
print(f"\n1. Total Reviews: {num_reviews}")

1. Total Reviews: 10000

[7] # 2. Average star rating
    average_rating = df['stars'].mean()
    print(f"2. Average Rating: {average_rating:.2f}")

2. Average Rating: 3.78
```

3.Count the number of unique users.
4.find the most common rating
5.find maximum and minimum stars.
6.find the percentage of reviews with 1-star

```
[8] # 3. Count of 5-star reviews
    five_star_reviews = (df['stars'] == 5).sum()
    print(f"3. 5-Star Reviews: {five_star_reviews}")

3. 5-Star Reviews: 3337

# 4. Most common rating
most_common_rating = df['stars'].mode()[0]
print(f"4. Most Common Rating: {most_common_rating}")

4. Most Common Rating: 4

[10] # 5. Max and Min stars
     max_stars = df['stars'].max()
     min_stars = df['stars'].min()
     print(f"5. Max Stars: {max_stars}, Min Stars: {min_stars}")

5. Max Stars: 5, Min Stars: 1

# 6. Percentage of reviews with 1-star
one_star_reviews = (df['stars'] == 1).sum()
percentage_one_star = (one_star_reviews / num_reviews) * 100
print(f"6. Percentage of 1-Star Reviews: {percentage_one_star:.2f}%")
```

7. Find the average length of review
8. Find the longest review
9. Find the shortest review
10. Findthe standard deviation of text length

```
[12] # 7. Average review text length
     df['text_length'] = df['text'].apply(len)
     avg_text_length = df['text_length'].mean()
     print(f"7. Average Review Text Length: {avg_text_length:.2f} characters")

     7. Average Review Text Length: 710.74 characters

[13] # 8. Longest review (by text length)
     longest_review = df.loc[df['text_length'].idxmax()]
     print(f"8. Longest Review Stars: {longest_review['stars']}")

     8. Longest Review Stars: 4

[14] # 9. Shortest review (by text length)
     shortest_review = df.loc[df['text_length'].idxmin()]
     print(f"9. Shortest Review Stars: {shortest_review['stars']}")

     9. Shortest Review Stars: 3

     # 10. Standard deviation of text length
     std_text_length = df['text_length'].std()
     print(f"10. Std Deviation of Text Length: {std_text_length:.2f}")

     10. Std Deviation of Text Length: 617.40
```

11. Find number of review with text length>1000
12. Find number of review containing word 'great'
13. Find number of reviews containing word 'bad'
14. Find correlation between stars and text length

```
[16] # 11. Number of reviews with text length > 1000
     long_reviews = (df['text_length'] > 1000).sum()
     print(f"11. Reviews longer than 1000 characters: {long_reviews}")

     11. Reviews longer than 1000 characters: 2230

[17] # 12. Reviews containing the word 'great' (case insensitive)
     great_reviews = df['text'].str.contains('great', case=False, na=False).sum()
     print(f"12. Reviews containing 'great': {great_reviews}")

     12. Reviews containing 'great': 3601

[18] # 13. Reviews containing the word 'bad'
     bad_reviews = df['text'].str.contains('bad', case=False, na=False).sum()
     print(f"13. Reviews containing 'bad': {bad_reviews}")

     13. Reviews containing 'bad': 912

[19] # 14. Correlation between stars and text length
     correlation_stars_textlength = df['stars'].corr(df['text_length'])
     print(f"14. Correlation between Stars and Text Length: {correlation_stars_textlength:.4f}")

     14. Correlation between Stars and Text Length: -0.1147
```

15. Find average text length per star rating

16 .find count of reviews per star rating

```
[20] # 15. Average text length per star rating
     avg_text_length_per_star = df.groupby('stars')['text_length'].mean()
     print(f"15. Average Text Length per Star Rating:\n{avg_text_length_per_star}")

15. Average Text Length per Star Rating:
stars
1    826.515354
2    842.256742
3    758.498289
4    712.923142
5    624.999101
Name: text_length, dtype: float64

     # 16. Count of reviews per star rating
     reviews_per_star = df['stars'].value_counts().sort_index()
     print(f"16. Reviews per Star Rating:\n{reviews_per_star}")

16. Reviews per Star Rating:
stars
1     749
2     927
3    1461
4    3526
5    3337
Name: count, dtype: int64
```

17. Find top 5 longest review
18. Find number of unique text lengths
19. Find median star rating

```
[22] # 17. Top 5 longest reviews (text)
     top5_long_reviews = df.nlargest(5, 'text_length')[['stars', 'text_length']]
     print(f"17. Top 5 Longest Reviews:\n{top5_long_reviews}")

17. Top 5 Longest Reviews:
      stars  text_length
55        4         4997
2622      5         4986
4033      3         4975
3686      2         4972
1870      4         4968

[23] # 18. Number of unique text lengths
     unique_text_lengths = df['text_length'].nunique()
     print(f"18. Number of Unique Text Lengths: {unique_text_lengths}")

18. Number of Unique Text Lengths: 2134

[24] # 19. Median stars rating
     median_stars = df['stars'].median()
     print(f"19. Median Stars: {median_stars}")

19. Median Stars: 4.0
```

20. Find quantile value of text length

```
[23] # 18. Number of unique text lengths
     unique_text_lengths = df['text_length'].nunique()
     print(f"18. Number of Unique Text Lengths: {unique_text_lengths}")
```

```
18. Number of Unique Text Lengths: 2134
```

```
[24] # 19. Median stars rating
     median_stars = df['stars'].median()
     print(f"19. Median Stars: {median_stars}")
```

```
19. Median Stars: 4.0
```

```
[25] # 20. Quantile values (25%, 50%, 75%) of text length
     quantiles_text_length = df['text_length'].quantile([0.25, 0.5, 0.75])
     print(f"20. Text Length Quantiles:\n{quantiles_text_length}")
```

THANK YOU