NAME : NARAYAN SANJAYRAO TERKHEDKAR

CLASS : CS3

ROLL NO : CS3-32

PRN : 202401040049

```
import pandas as pd
import numpy as np
import re
from collections import Counter

df = pd.read_csv('/content/drive/MyDrive/Dataset/IMDB Dataset.csv')
```

1. Total reviews

```
print("Total reviews:", len(df))
```

```
Total reviews: 50000
```

2. Positive and negative reviews

```
print(df['sentiment'].value_counts())
```

```
sentiment
positive    25000
negative    25000
Name: count, dtype: int64
```

3. Percentage of positive reviews

```
print("Positive %:", (df['sentiment'].value_counts()['positive'] / len(df)) * 100)
```

```
Positive %: 50.0
```

4. Average characters per review

```
print("Avg characters per review:", df['review'].apply(len).mean())
```

```
Avg characters per review: 1309.43102
```

5. Average words per review

```
print("Avg words per review:", df['review'].apply(lambda x: len(x.split())).mean())
```

6. Review with max words

```
print(df.loc[df['review'].apply(lambda x: len(x.split())).idxmax()])
```

```
review       Match 1: Tag Team Table Match Bubba Ray and Sp...
sentiment                                             positive
Name: 31481, dtype: object
```

7. Review with min characters

```
print(df.loc[df['review'].apply(len).idxmin()])
```

```
review       Read the book, forget the movie!
sentiment                            negative
Name: 27521, dtype: object
```

8. Reviews containing 'excellent'

```
print("Reviews with 'excellent':", df['review'].str.contains('excellent', case=False).sum())
```

```
Reviews with 'excellent': 3625
```

9. First 5 reviews mentioning 'boring'

```
print(df[df['review'].str.contains('boring', case=False)].head(5))
```

```
                                               review sentiment
5   Probably my all-time favorite movie, a story o...  positive
8   Encouraged by the positive comments about this...  negative
23  First of all, let's get a few things straight ...  negative
34  I watched this film not really expecting much,...  negative
63  Besides being boring, the scenes were oppressi...  negative
```

10. Median words per review

```
print("Median words:", df['review'].apply(lambda x: len(x.split())).median())
```

```
Median words: 173.0
```

11. New column review_length

```
df['review_length'] = df['review'].apply(lambda x: len(x.split()))
print(df[['review', 'review_length']].head())
```

```
                                              review  review_length
0  One of the other reviewers has mentioned that ...            307
1  A wonderful little production. <br /><br />The...            162
2  I thought this was a wonderful way to spend ti...            166
3  Basically there's a family where a little boy ...            138
4  Petter Mattei's "Love in the Time of Money" is...            230
```

12. Avg length of positive reviews

```
print("Positive avg length:", df[df['sentiment']=='positive']['review_length'].mean())
```

```
Positive avg length: 232.84932
```

13. Avg length of negative reviews

```
print("Negative avg length:", df[df['sentiment']=='negative']['review_length'].mean())
```

14. Most common word in positive reviews

```
positive_words = ' '.join(df[df['sentiment']=='positive']['review']).lower()
positive_words = re.findall(r'\b\w+\b', positive_words)
print("Most common in positive:", Counter(positive_words).most_common(1))
```

```
Most common in positive: [('the', 341281)]
```

15. Most common word in negative reviews

```
negative_words = ' '.join(df[df['sentiment']=='negative']['review']).lower()
negative_words = re.findall(r'\b\w+\b', negative_words)
print("Most common in negative:", Counter(negative_words).most_common(1))
```

```
Most common in negative: [('the', 326712)]
```

16. Reviews mentioning 'awful'

```
print("Reviews with 'awful':", df['review'].str.contains('awful', case=False).sum())
```

```
Reviews with 'awful': 3119
```

17. Top 10 longest reviews

```
print(df.nlargest(10, 'review_length')[['review', 'review_length']])
```

```
                                                  review  review_length
31481  Match 1: Tag Team Table Match Bubba Ray and Sp...           2470
40521  There's a sign on The Lost Highway that says:<...           2278
31436  Back in the mid/late 80s, an OAV anime by titl...           2125
31240  (Some spoilers included:)<br /><br />Although,...           2108
12647  Titanic directed by James Cameron presents a f...           1839
5708   **Attention Spoilers**<br /><br />First of all...           1830
3024   If anyone ever assembles a compendium on moder...           1737
42946  By now you've probably heard a bit about the n...           1723
3654   *!!- SPOILERS - !!*<br /><br />Before I begin ...           1601
1531   Warning: Does contain spoilers.<br /><br />Ope...           1527
```

18. Top 10 shortest reviews

```
print(df.nsmallest(10, 'review_length')[['review', 'review_length']])
```

```
                                                  review  review_length
28920  Primary plot!Primary direction!Poor interpreta...              4
27521                   Read the book, forget the movie!              6
13109  More suspenseful, more subtle, much, much more...              8
40817   I hope this group of film-makers never re-unites.              8
31072          What a script, what a story, what a mess!              9
11926  I wouldn't rent this one even on dollar rental...             10
18400  Brilliant and moving performances by Tom Court...             10
19874  This movie is terrible but it has some good ef...             10
20274  You'd better choose Paul Verhoeven's even if y...             11
31761  Ming The Merciless does a little Bardwork and ...             12
```

19. Remove HTML tags

```
df['clean_review'] = df['review'].apply(lambda x: re.sub('<.*?>', ' ', x))
print(df[['review', 'clean_review']].head())
```

```
                                              review  \
0  One of the other reviewers has mentioned that ...   
1  A wonderful little production. <br /><br />The...   
2  I thought this was a wonderful way to spend ti...   
3  Basically there's a family where a little boy ...   
4  Petter Mattei's "Love in the Time of Money" is...   

                                        clean_review  
0  One of the other reviewers has mentioned that ...  
1  A wonderful little production.   The filming t...  
2  I thought this was a wonderful way to spend ti...  
3  Basically there's a family where a little boy ...  
4  Petter Mattei's "Love in the Time of Money" is...  
```

20. Correlation between review_length and sentiment

```
df['sentiment_encoded'] = df['sentiment'].map({'positive': 1, 'negative': 0})
print("Correlation:", np.corrcoef(df['review_length'], df['sentiment_encoded'])[0,1])
```

```
Correlation: 0.009877188293045663
```

21. Reviews containing 'amazing'

```
print("Reviews with 'amazing':", df['review'].str.contains('amazing', case=False).sum())
```

```
Reviews with 'amazing': 2479
```

22. Avg number of exclamation marks (!)

```
print("Avg ! marks per review:", df['review'].apply(lambda x: x.count('!')).mean())
```

```
Avg ! marks per review: 0.98328
```