EDS

Atharv Madhav Darade

ET1-56

202401070046

E13

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

[1] 1s

```
import pandas as pd
import numpy as np

# Load the Excel file
df = pd.read_excel('IMDB_Large_Dataset.xlsx')

df.head()
```

	review	sentiment
0	Could have been better.	neutral
1	Had its moments.	neutral
2	A delightful experience.	positive
3	It was okay.	neutral
4	Could have been better.	neutral

Next steps: Generate code with df | View recommended plots | New interactive sheet

[2] 0s

```
total_reviews = df.shape[0]
print(f"Total Reviews: {total_reviews}")
```

```
Total Reviews: 500
```

0s completed at 8:15PM

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

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

```
sentiment
neutral     170
negative    168
positive    162
Name: count, dtype: int64
```

```
[4] sentiment_percentage = df['sentiment'].value_counts(normalize=True) * 100
print(sentiment_percentage)
```

```
sentiment
neutral     34.0
negative    33.6
positive    32.4
Name: proportion, dtype: float64
```

```
[5] longest_review = df['review'].apply(len).idxmax()
print(df.loc[longest_review])
```

```
review       Some parts were good, others not so much.
sentiment                                      neutral
Name: 9, dtype: object
```

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
[6] avg_words = df['review'].apply(lambda x: len(x.split())).mean()
    print(f"Average Words per Review: {avg_words}")
```

```
Average Words per Review: 3.312
```

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

```
review       It was okay.
sentiment         neutral
Name: 3, dtype: object
```

```
[8] amazing_reviews = df['review'].str.contains('amazing', case=False).sum()
    print(f"Reviews mentioning 'amazing': {amazing_reviews}")
```

```
Reviews mentioning 'amazing': 12
```

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
[9] df['review_length'] = df['review'].apply(len)
    df.head()
```

	review	sentiment	review_length
0	Could have been better.	neutral	23
1	Had its moments.	neutral	16
2	A delightful experience.	positive	24
3	It was okay.	neutral	12
4	Could have been better.	neutral	23

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
[10] sentiment_longest_review = df.loc[longest_review, 'sentiment']
     print(f"Sentiment of longest review: {sentiment_longest_review}")
```

```
Sentiment of longest review: neutral
```

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
[11] df['word_count'] = df['review'].apply(lambda x: len(x.split()))
     df.head()
```

	review	sentiment	review_length	word_count
0	Could have been better.	neutral	23	4
1	Had its moments.	neutral	16	3
2	A delightful experience.	positive	24	3
3	It was okay.	neutral	12	3
4	Could have been better.	neutral	23	4

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
[12] avg_word_count_by_sentiment = df.groupby('sentiment')['word_count'].mean()
     print(avg_word_count_by_sentiment)
```

```
sentiment
negative    3.327381
neutral     3.552941
positive    3.043210
Name: word_count, dtype: float64
```

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
[13] std_review_length = df['review_length'].std()
     print(f"Standard Deviation of Review Lengths: {std_review_length}")
```

```
Standard Deviation of Review Lengths: 6.638828819618149
```

```
long_reviews = df[df['word_count'] > 20]
print(long_reviews.shape[0])
```

```
0
```

```
[15] df['is_long_review'] = df['word_count'] > 20
     df.head()
```

	review	sentiment	review_length	word_count	is_long_review
0	Could have been better.	neutral	23	4	False
1	Had its moments.	neutral	16	3	False
2	A delightful experience.	positive	24	3	False
3	It was okay.	neutral	12	3	False
4	Could have been better.	neutral	23	4	False

Next steps: Generate code with df | View recommended plots | New interactive sheet

0s completed at 8:15PM

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
[16] long_review_percentage = df['is_long_review'].mean() * 100
     print(f"Percentage of Long Reviews: {long_review_percentage}%")
```

```
Percentage of Long Reviews: 0.0%
```

```
[17] bad_reviews = df['review'].str.contains('bad', case=False).sum()
     print(f"Reviews mentioning 'bad': {bad_reviews}")
```

```
Reviews mentioning 'bad': 38
```

```
[18] good_reviews = df[df['review'].str.contains('good', case=False)]
     avg_good_length = good_reviews['review_length'].mean()
     print(f"Average length of reviews mentioning 'good': {avg_good_length}")
```

```
Average length of reviews mentioning 'good': 28.096774193548388
```

Untitled6.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text

RAM Disk

```
avg_good_length = good_reviews['review_length'].mean()
print(f"Average length of reviews mentioning 'good': {avg_good_length}")
```

```
Average length of reviews mentioning 'good': 28.096774193548388
```

```
[19] df_sorted = df.sort_values('review_length', ascending=False)
     df_sorted.head()
```

	review	sentiment	review_length	word_count	is_long_review
160	Some parts were good, others not so much.	neutral	41	8	False
268	Some parts were good, others not so much.	neutral	41	8	False
280	Some parts were good, others not so much.	neutral	41	8	False
319	Some parts were good, others not so much.	neutral	41	8	False
86	Some parts were good, others not so much.	neutral	41	8	False

Next steps: Generate code with df_sorted | View recommended plots | New interactive sheet

```
df.to_excel('IMDB_Large_Dataset_Modified.xlsx', index=False)
print("Modified Excel file created successfully!")
```

```
Modified Excel file created successfully!
```

0s completed at 8:15PM