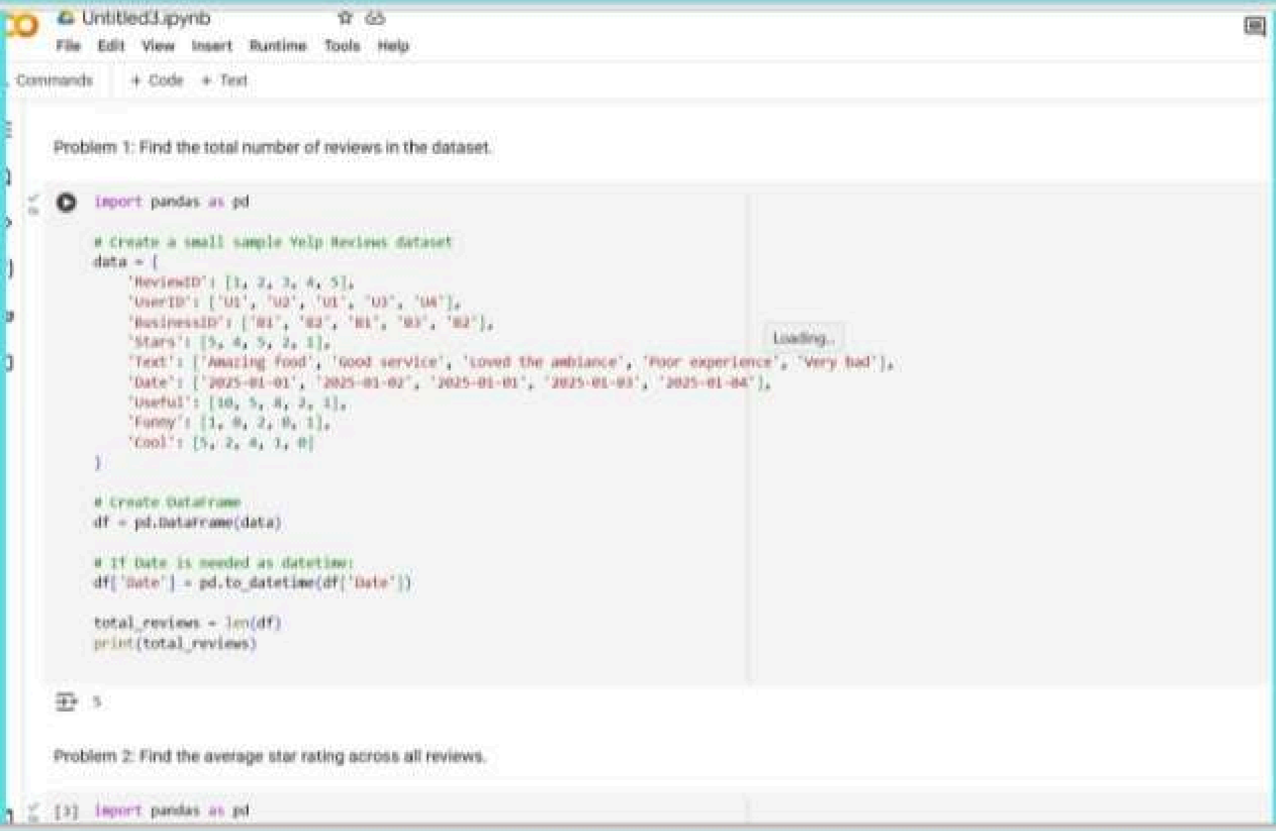
Name : Varad Moholkar

Roll no : ET2-81 DIV : ET-2 Batch : E-24 PRN : 202401070210



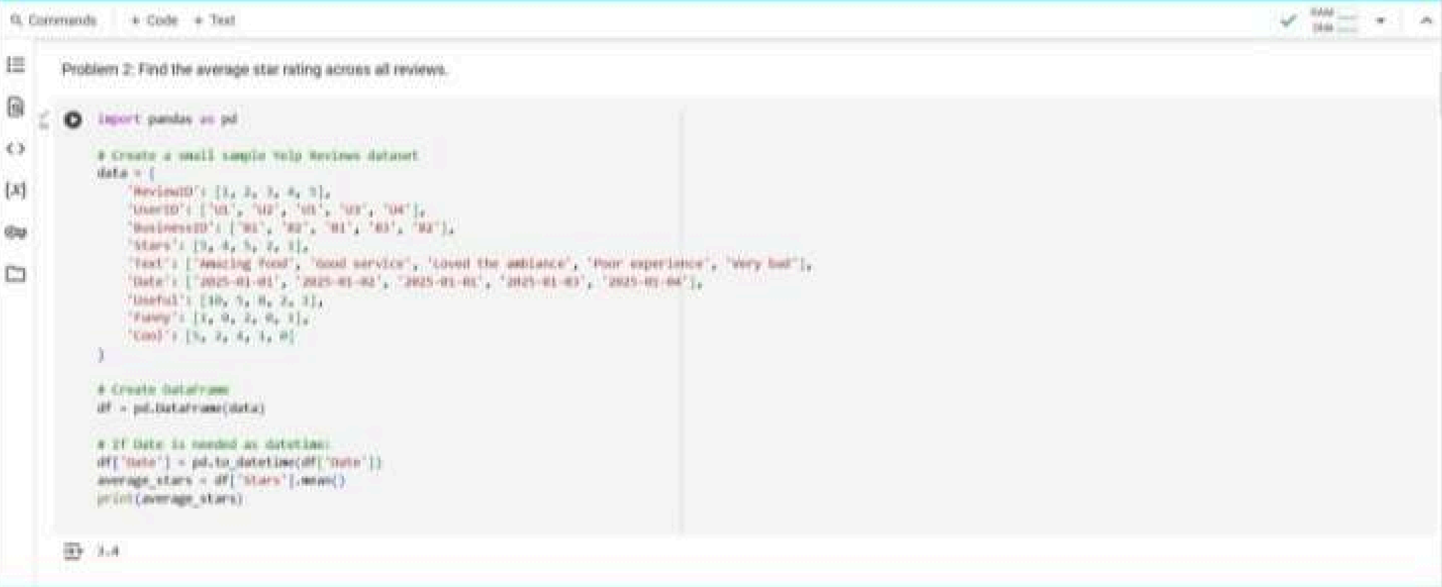
Problem 1: Find the total number of reviews in the dataset.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)

# If Date is needed as datetime:
df['Date'] = pd.to_datetime(df['Date'])

total_reviews = len(df)
print(total_reviews)
```

```
5
```

Problem 2: Find the average star rating across all reviews.

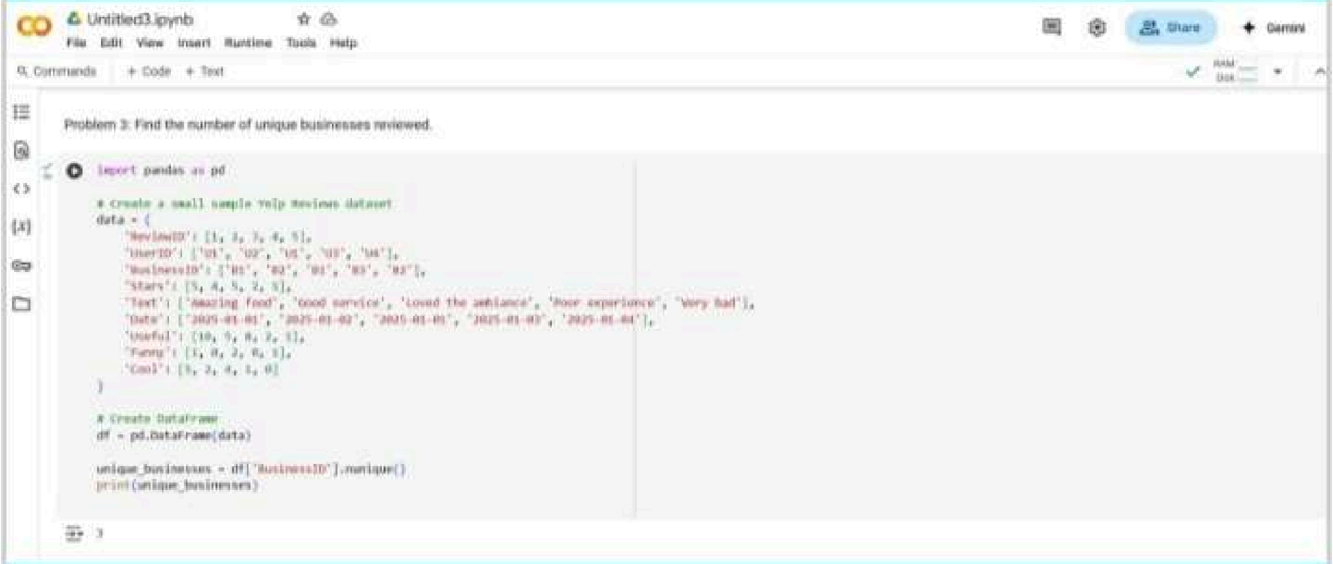
```
[3] import pandas as pd
```



Problem 2: Find the average star rating across all reviews.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)

# If Date is needed as datetime:
df['Date'] = pd.to_datetime(df['Date'])
average_stars = df['Stars'].mean()
print(average_stars)
```

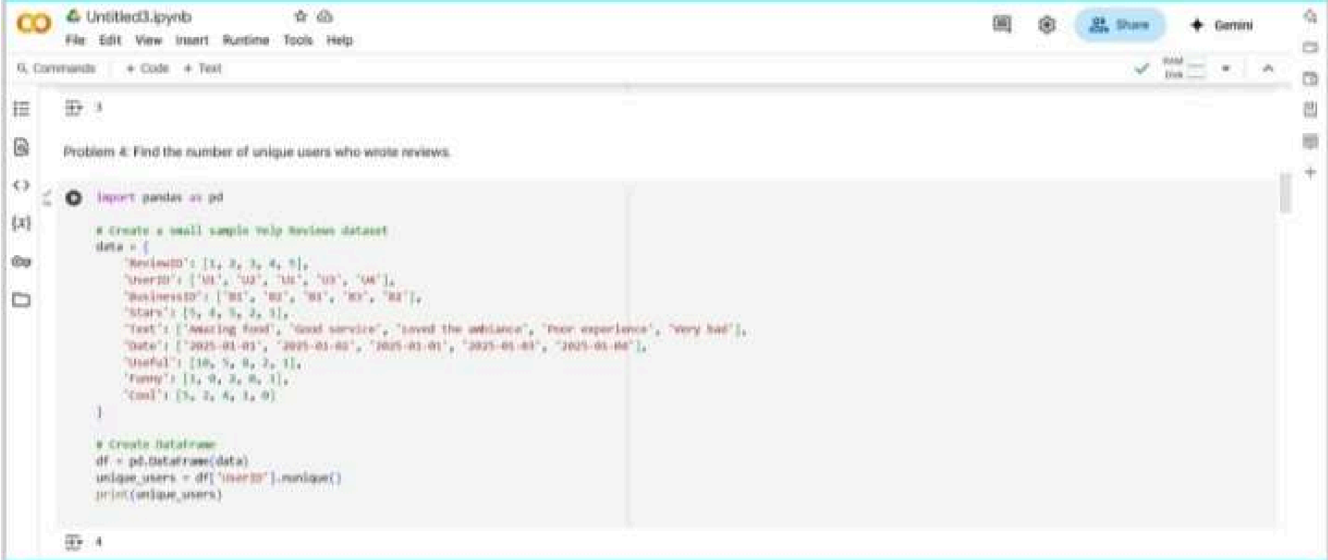
```
3.4
```



Problem 3: Find the number of unique businesses reviewed.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)

unique_businesses = df['BusinessID'].nunique()
print(unique_businesses)
```

```
3
```

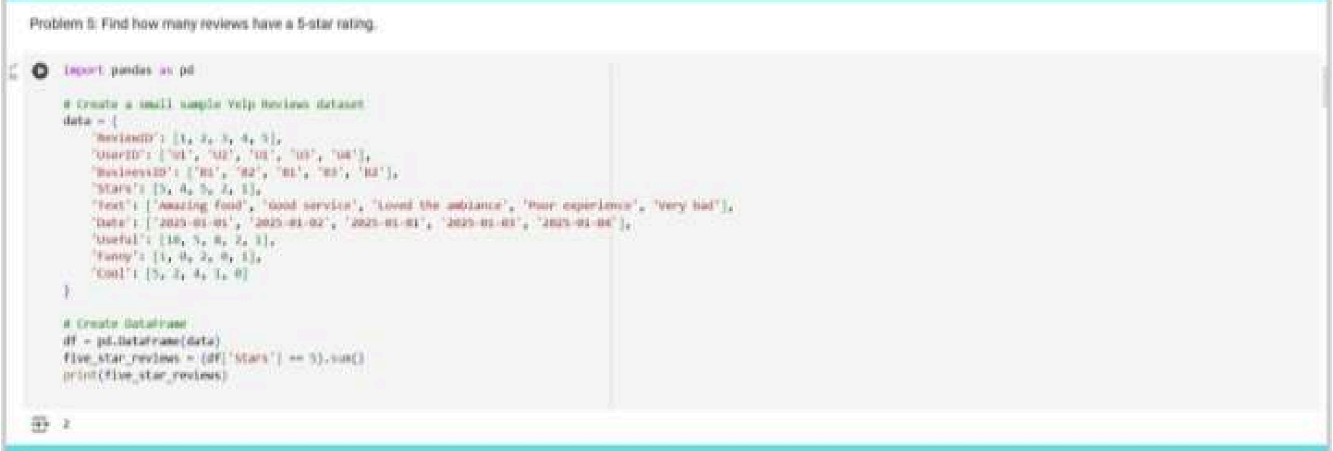


```
3
```

Problem 4: Find the number of unique users who wrote reviews.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
unique_users = df['UserID'].nunique()
print(unique_users)
```

```
4
```

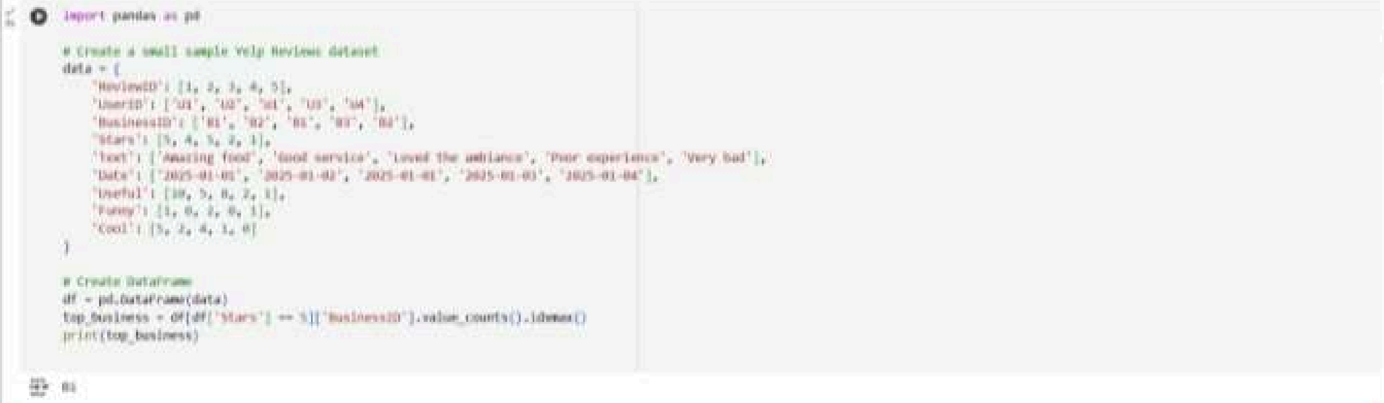


Problem 5: Find how many reviews have a 5-star rating.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
five_star_reviews = (df['Stars'] == 5).sum()
print(five_star_reviews)
```

```
2
```

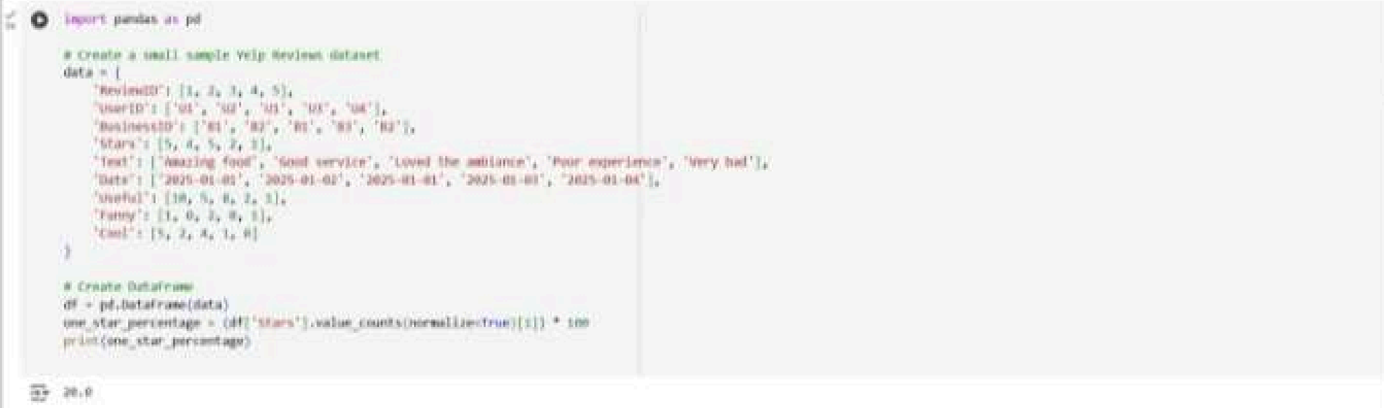
Problem 6: Find the business with the highest number of 5-star reviews.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
top_business = df[df['Stars'] == 5]['BusinessID'].value_counts().idxmax()
print(top_business)
```

```
B1
```

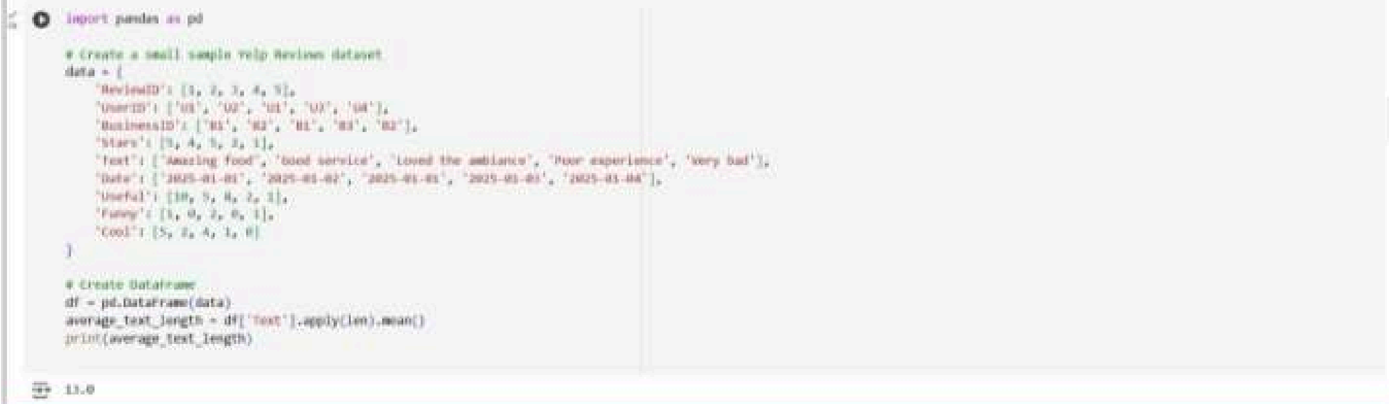
Problem 7: Find the percentage of reviews rated as 1-star.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
one_star_percentage = (df['Stars'].value_counts(normalize=True)[1]) * 100
print(one_star_percentage)
```

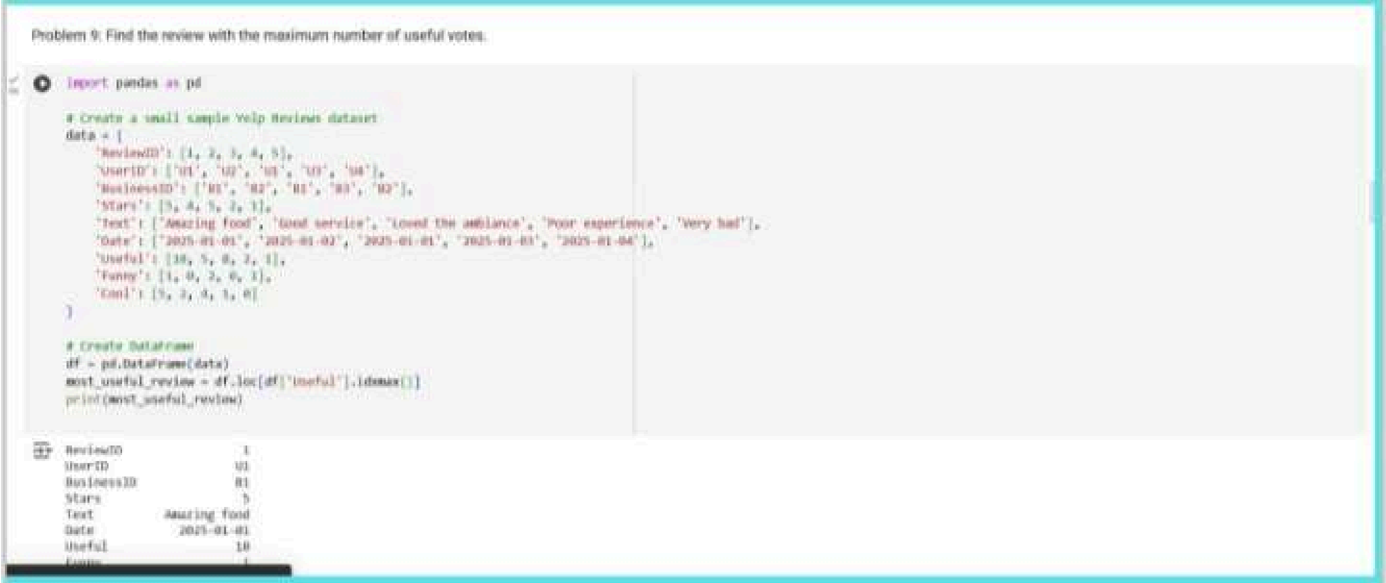
```
20.0
```

Problem 8: Find the average length of review text.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
average_text_length = df['Text'].apply(len).mean()
print(average_text_length)
```

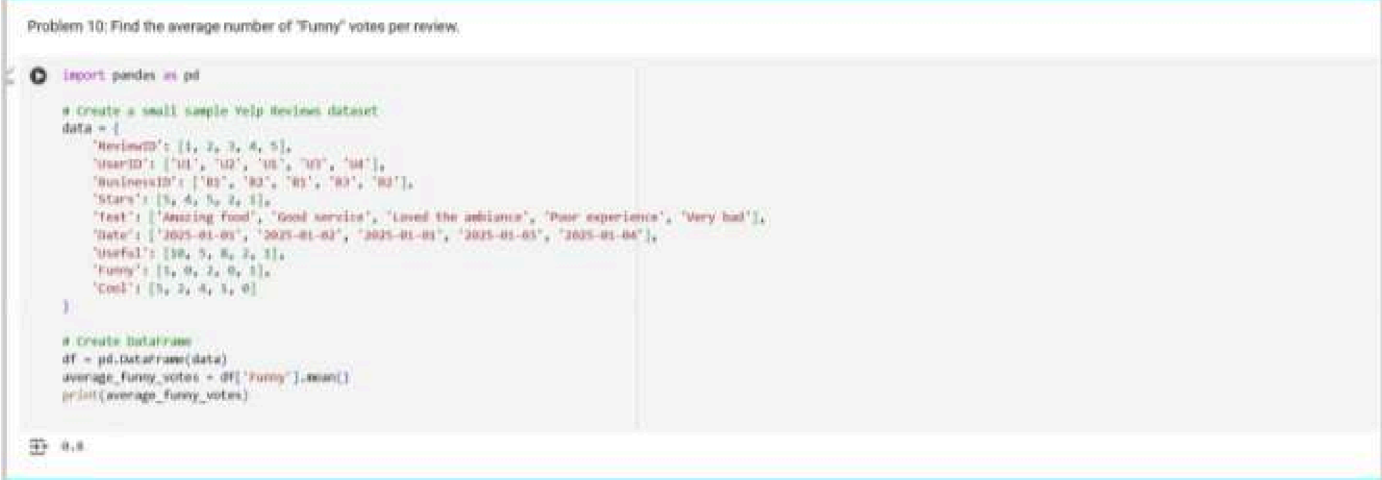
```
13.0
```



Problem 9: Find the review with the maximum number of useful votes.

```python
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
most_useful_review = df.loc[df['Useful'].idxmax()]
print(most_useful_review)
```

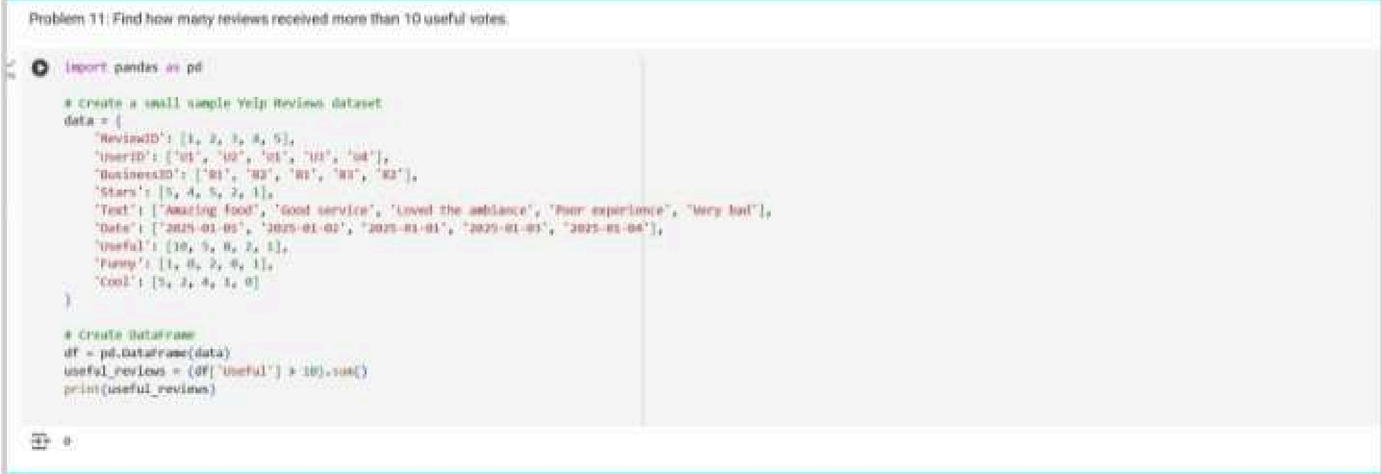
```
ReviewID                 1
UserID                  U1
BusinessID              B1
Stars                    5
Text          Amazing food
Date            2025-01-01
Useful                  10
Funny                    1
```



Problem 10: Find the average number of "Funny" votes per review.

```python
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
average_funny_votes = df['Funny'].mean()
print(average_funny_votes)
```

```
0.8
```

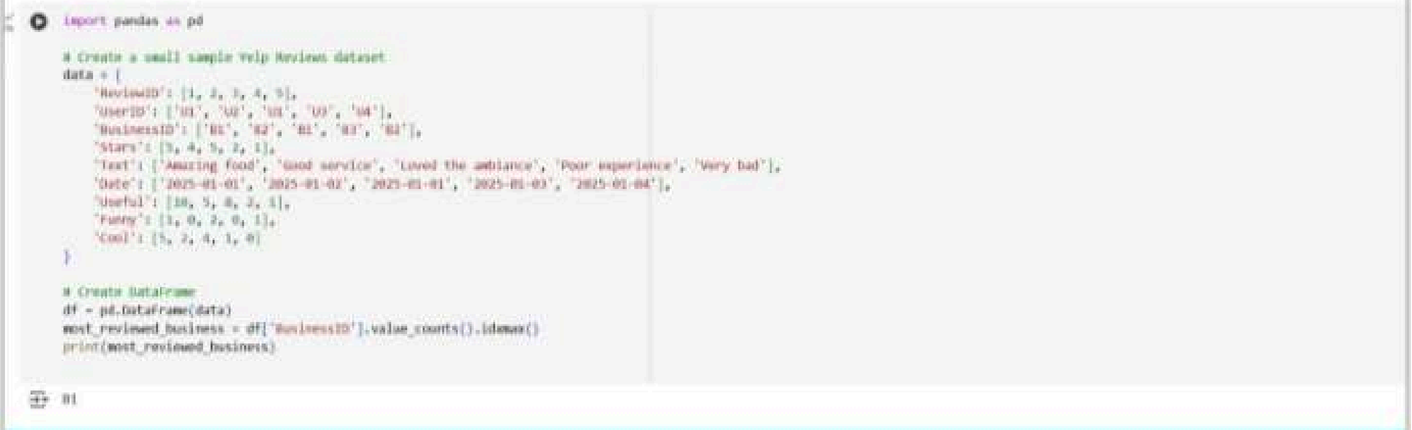


Problem 11: Find how many reviews received more than 10 useful votes.

```python
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
useful_reviews = (df['Useful'] > 10).sum()
print(useful_reviews)
```

```
0
```

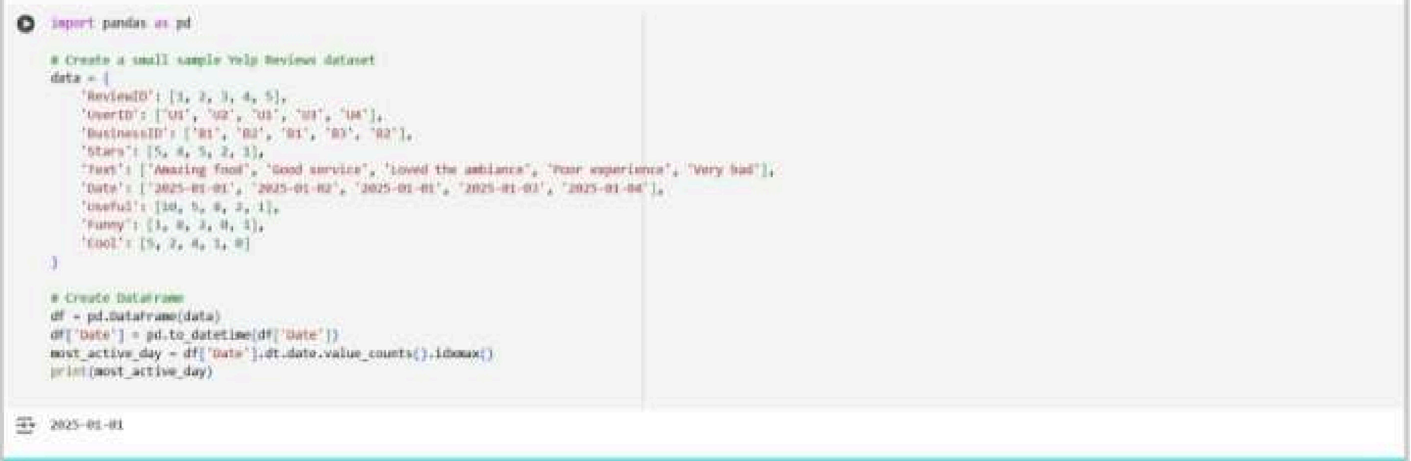
Problem 12: Find the most reviewed business.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
most_reviewed_business = df['BusinessID'].value_counts().idxmax()
print(most_reviewed_business)
```

```
B1
```

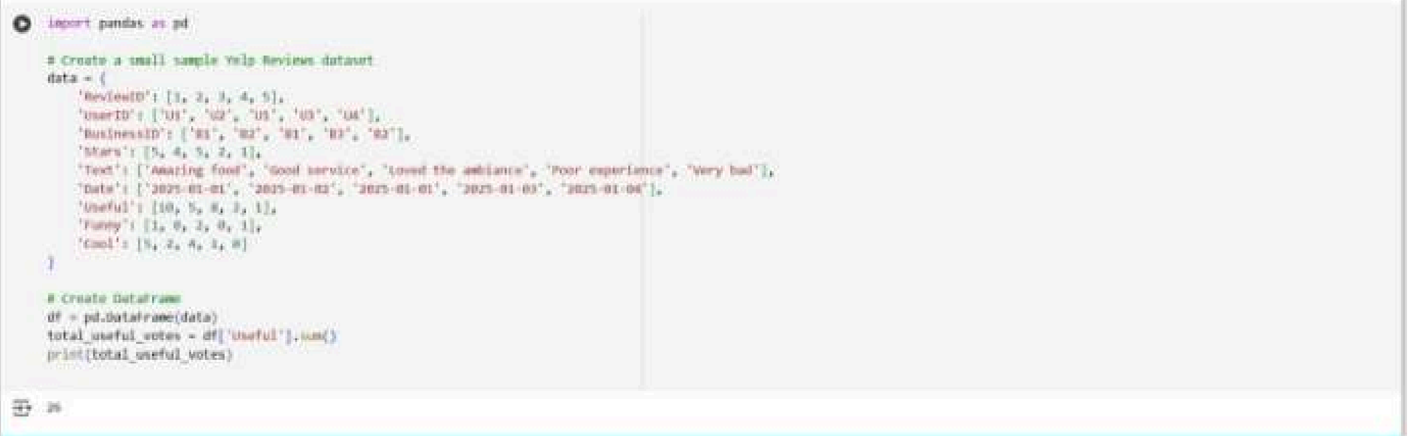
Problem 13: Find the day with the most reviews.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
df['Date'] = pd.to_datetime(df['Date'])
most_active_day = df['Date'].dt.date.value_counts().idxmax()
print(most_active_day)
```

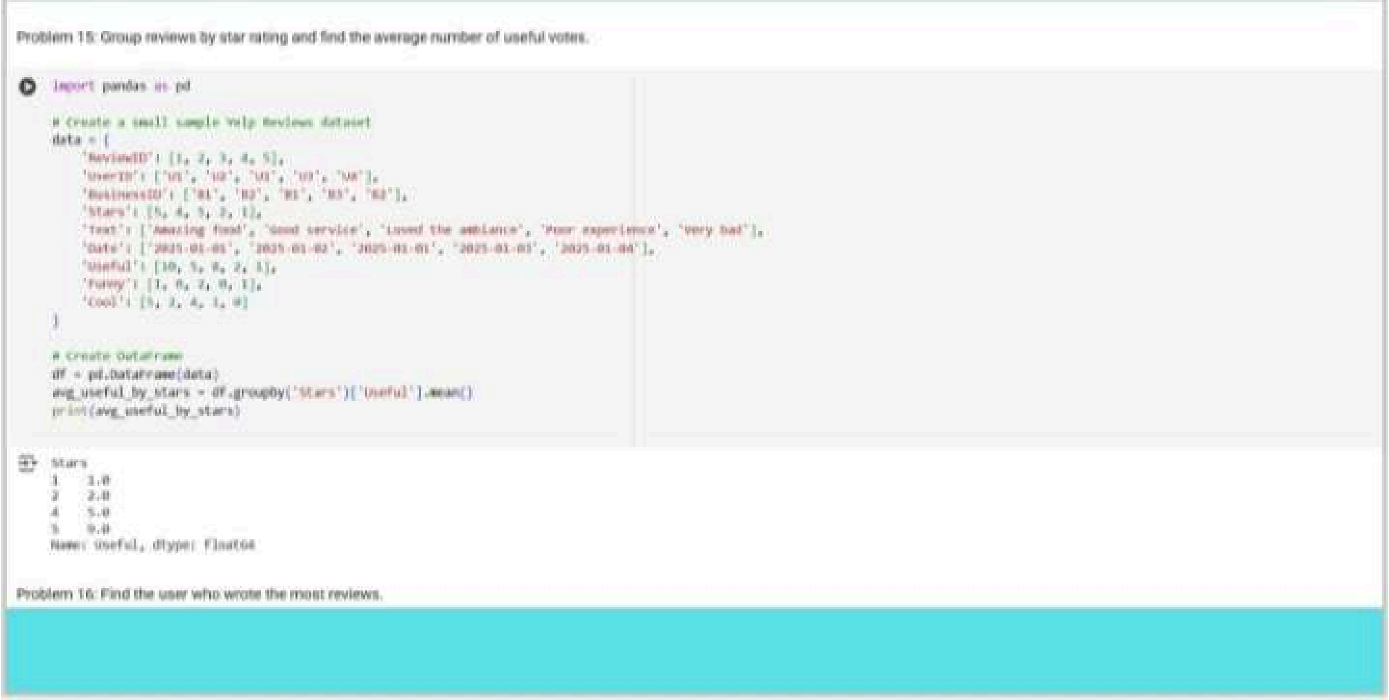
```
2025-01-01
```

Problem 14: Calculate total useful votes across all reviews.



```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
total_useful_votes = df['Useful'].sum()
print(total_useful_votes)
```

```
26
```

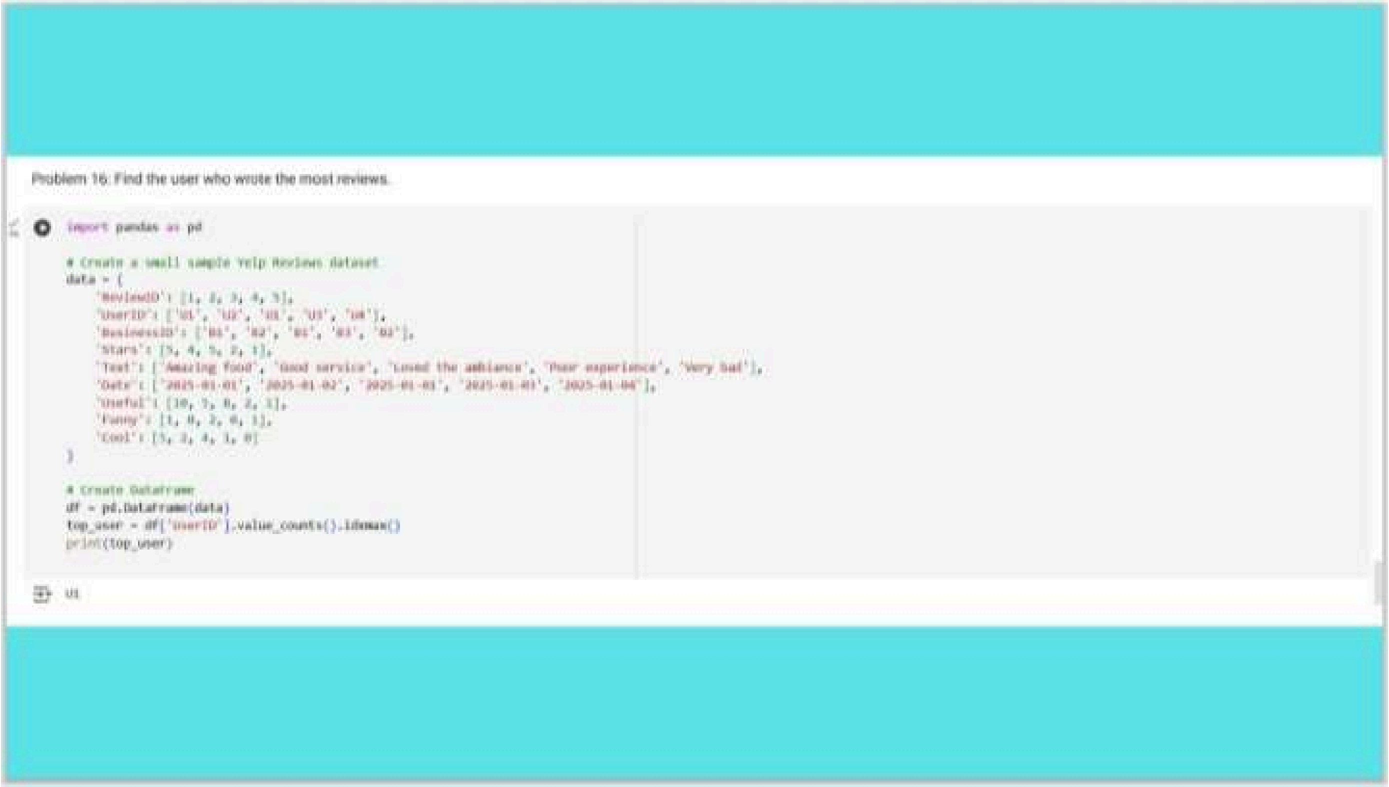


Problem 15: Group reviews by star rating and find the average number of useful votes.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
avg_useful_by_stars = df.groupby('Stars')['Useful'].mean()
print(avg_useful_by_stars)
```

```
Stars
1    1.0
2    2.0
4    5.0
5    9.0
Name: Useful, dtype: float64
```

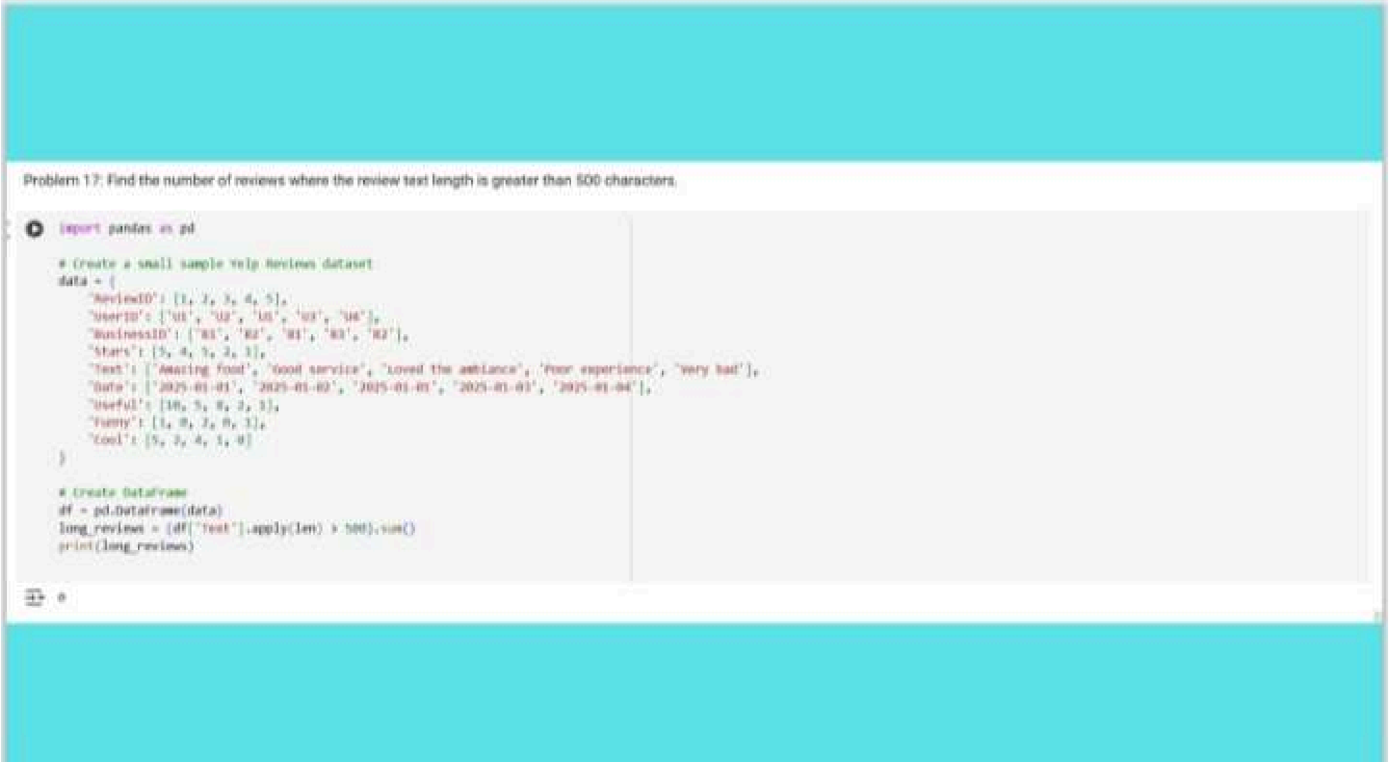
Problem 16: Find the user who wrote the most reviews.



Problem 16: Find the user who wrote the most reviews.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
top_user = df['UserID'].value_counts().idxmax()
print(top_user)
```

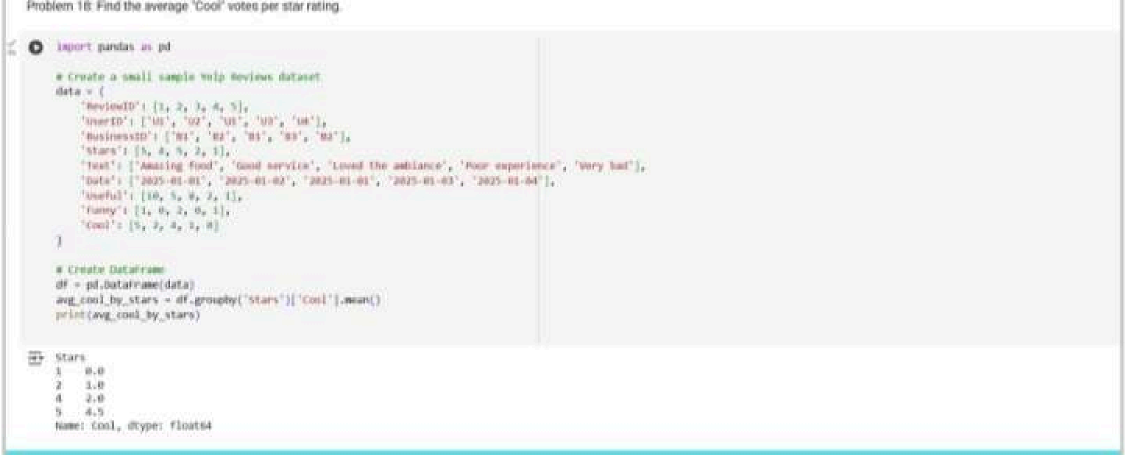
```
U1
```



Problem 17: Find the number of reviews where the review text length is greater than 500 characters.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
long_reviews = (df['Text'].apply(len) > 500).sum()
print(long_reviews)
```

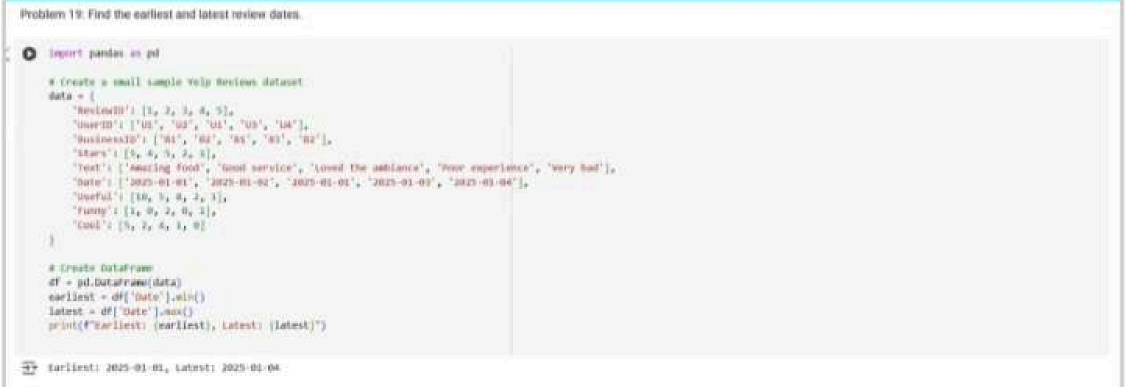
```
0
```



Problem 18: Find the average 'Cool' votes per star rating.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
avg_cool_by_stars = df.groupby('Stars')['Cool'].mean()
print(avg_cool_by_stars)
```

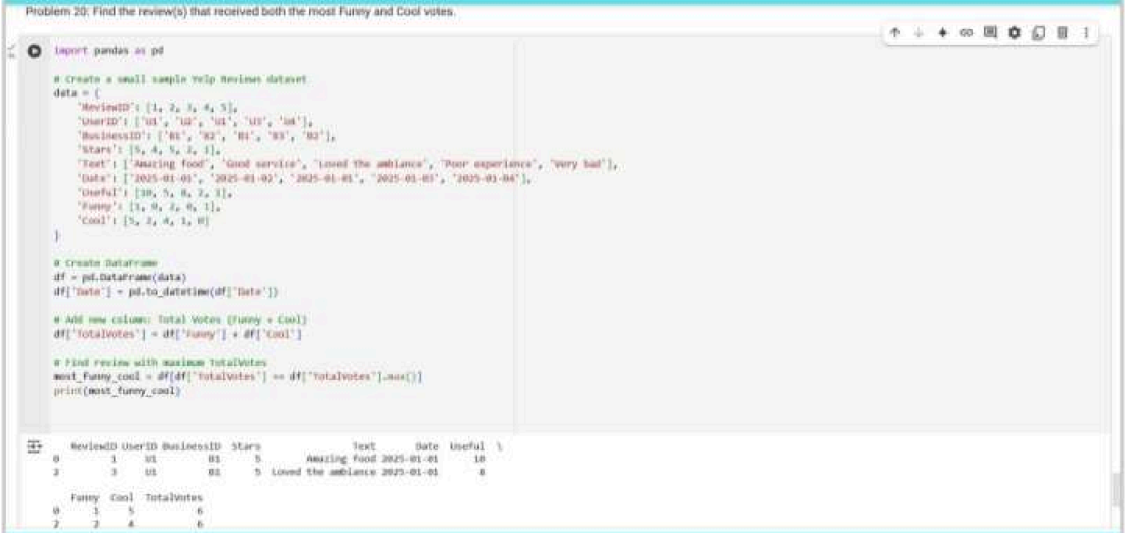
```
Stars
1    0.0
2    1.0
4    2.0
5    4.5
Name: Cool, dtype: float64
```



Problem 19: Find the earliest and latest review dates.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
earliest = df['Date'].min()
latest = df['Date'].max()
print(f"Earliest: {earliest}, Latest: {latest}")
```

```
Earliest: 2025-01-01, Latest: 2025-01-04
```



Problem 20: Find the review(s) that received both the most Funny and Cool votes.

```
import pandas as pd

# Create a small sample Yelp Reviews dataset
data = {
    'ReviewID': [1, 2, 3, 4, 5],
    'UserID': ['U1', 'U2', 'U1', 'U3', 'U4'],
    'BusinessID': ['B1', 'B2', 'B1', 'B3', 'B2'],
    'Stars': [5, 4, 5, 2, 1],
    'Text': ['Amazing food', 'Good service', 'Loved the ambiance', 'Poor experience', 'Very bad'],
    'Date': ['2025-01-01', '2025-01-02', '2025-01-01', '2025-01-03', '2025-01-04'],
    'Useful': [10, 5, 8, 2, 1],
    'Funny': [1, 0, 2, 0, 1],
    'Cool': [5, 2, 4, 1, 0]
}

# Create DataFrame
df = pd.DataFrame(data)
df['Date'] = pd.to_datetime(df['Date'])

# Add new column: Total Votes (Funny + Cool)
df['TotalVotes'] = df['Funny'] + df['Cool']

# Find review with maximum TotalVotes
most_funny_cool = df[df['TotalVotes'] == df['TotalVotes'].max()]
print(most_funny_cool)
```

```
   ReviewID UserID BusinessID  Stars                Text       Date  Useful  \
0         1     U1         B1      5        Amazing food 2025-01-01      10
2         3     U1         B1      5  Loved the ambiance 2025-01-01       8

   Funny  Cool  TotalVotes
0      1     5           6
2      2     4           6
```

Thank You!!!