EDS Activity:-

Name: Kamakhya Kanzade

Roll no.:CS7-82

PRN:202401110016

Y elp Reviews

Q1 Find the average star rating.

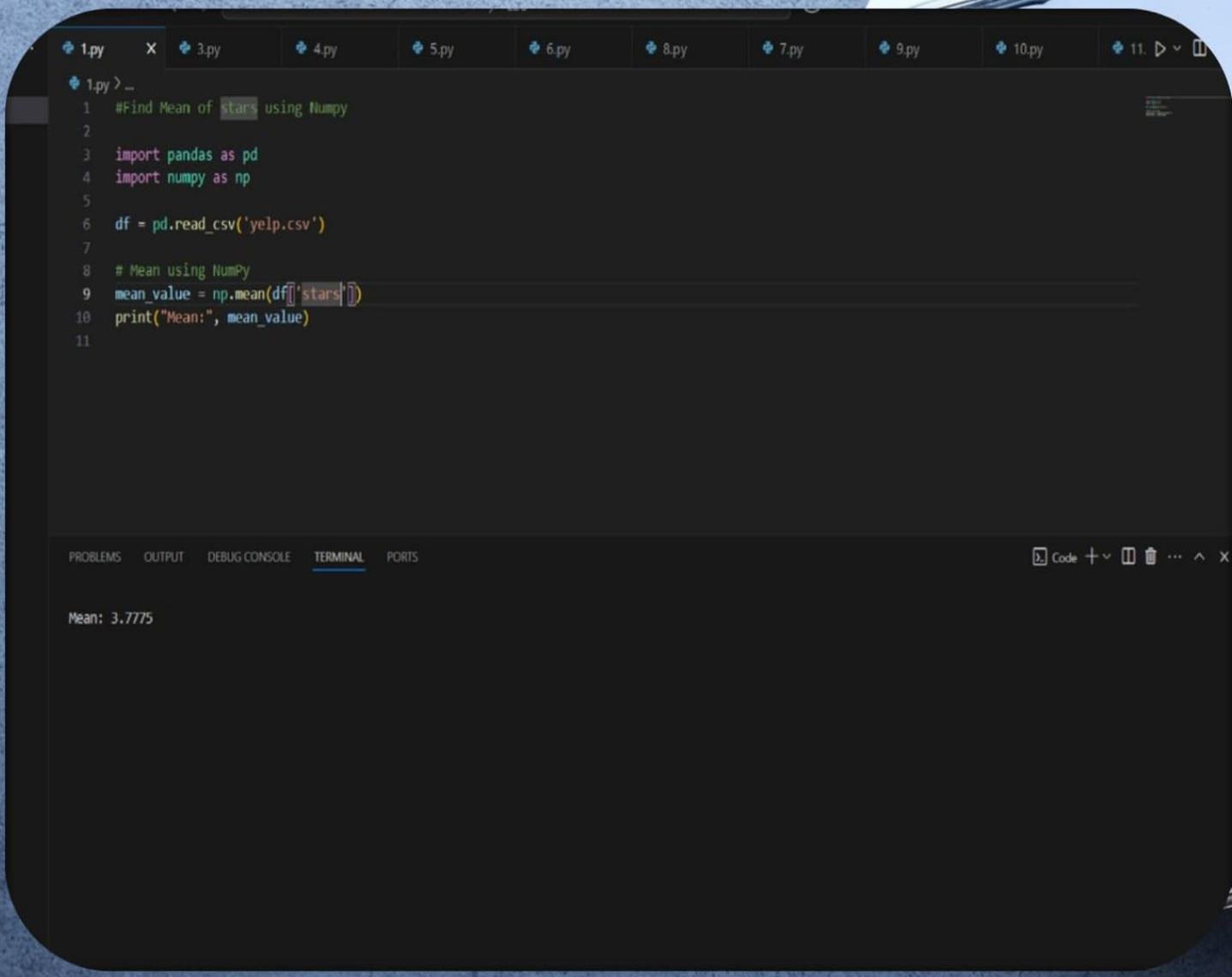
3.py > ...

```python
# Find the average star rating.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
average_stars = np.mean(df['stars'])
print("Average star rating:", average_stars)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Average star rating: 3.7775
```

Q.2] Find Mean of stars using Numpy



Q3) Find the total number of reviews.

2.py > ...

```python
#Find the total number of review.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
total_reviews = np.unique(df['business_id']).size
print("Total Reviews:", total_reviews)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Total Reviews: 4174
```

Q.4List the top 5 businesses with the highest number of stars.

5.py > ...

```python
#List the top 5 businesses with the highest number of stars.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
top_5_indices = np.argsort(-df['stars'].values)[:5]
top_reviewed = df.iloc[top_5_indices]
print(top_reviewed[['business_id', 'stars']])

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
9981  AX8lx9wHNYT45lyd7pxaYw      5
30    V1nEpIRmEa1768oj_tuxeQ      5
9969  qhIlkXgcC4j34lNTIqu9WA      5
9970  R6aazv8FB-6BeanY3ag8kw      5
9971  JOZqBKIOB8WEBAWm7v1JFA      5
```

Q.5] Count how many businesses have a 5-star rating.

```python
#Count how many businesses have a 5-star rating.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
five_star_count = np.sum(df['stars'] == 5.0)
print("5-star businesses:", five_star_count)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
5-star businesses: 3337
```

Q.6] Find the review with maximum 'cool' votes.

e review with maximum 'cool' votes.

4.py > ...

```python
# Find the review with maximum 'cool' votes.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
coolest_idx = np.argmax(df['cool'])
coolest_review = df.iloc[coolest_idx]
print(coolest_review[['review_id', 'text']])
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
review_id                          1kc50TqtMsIHDGR8yQgF8g
text         Love this place! Amazing Happy Hour Specials!!
Name: 4957, dtype: object
```

Q.7] Find the review with minimum 'funny' votes.

8.py > ...

```python
# Find the review with minimum 'funny' votes.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
least_funny_idx = np.argmin(df['funny'])
least_funny_review = df.iloc[least_funny_idx]
print(least_funny_review[['review_id', 'funny', 'text']])

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
review_id
funny
text        My wife took me here on my birthday for breakf...
Name: 0, dtype: object
```

Q.8] Find the total number of unique users.

7.py > ...

```python
# Find the total number of unique users.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
unique_users = np.unique(df['user_id']).size
print("Unique users:", unique_users)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Unique users: 6403
```

Q.9] Find how many reviews have more than 5 useful votes.

```python
# Find how many reviews have more than 5 useful votes.



import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
useful_reviews = np.sum(df['useful'] > 5)
print("Reviews with >5 useful votes:", useful_reviews)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Reviews with >5 useful votes: 431
```

Q.10] Find the standard deviation of star ratings.

10.py > ...

```python
# Find standard deviation of star ratings.


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
stars_std = np.std(df['stars'])
print("Standard deviation of stars:", stars_std)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Standard deviation of stars: 1.2145755431425416
```

Q.11 Find the average number of characters per review text


```python
# Find the average number of characters per review text


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
average_text_length = np.mean(df['text'].str.len())
print("Average number of characters per review:", average_text_length)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Average number of characters per review: 710.7387
```

Q.12 Find the earliest review date.


```python
#  Find the earliest review date.


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
df['date'] = pd.to_datetime(df['date'])
earliest_date = np.min(df['date'])
print("Earliest review date:", earliest_date)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Earliest review date: 2005-04-18 00:00:00
```

Q.13] Find the percentage of reviews with a 5-star rating.


```python
# Find the percentage of reviews with a 5-star rating.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
percentage_5_stars = np.mean(df['stars'] == 5) * 100
print("Percentage of 5-star reviews:", percentage_5_stars)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Percentage of 5-star reviews: 33.37
```

Q.14 FFind average 'cool' votes per review.

12.py > ...

```python
#  FFind average 'cool' votes per review.


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
average_cool = np.mean(df['cool'])
print("Average cool votes per review:", average_cool)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Average cool votes per review: 0.8768
```

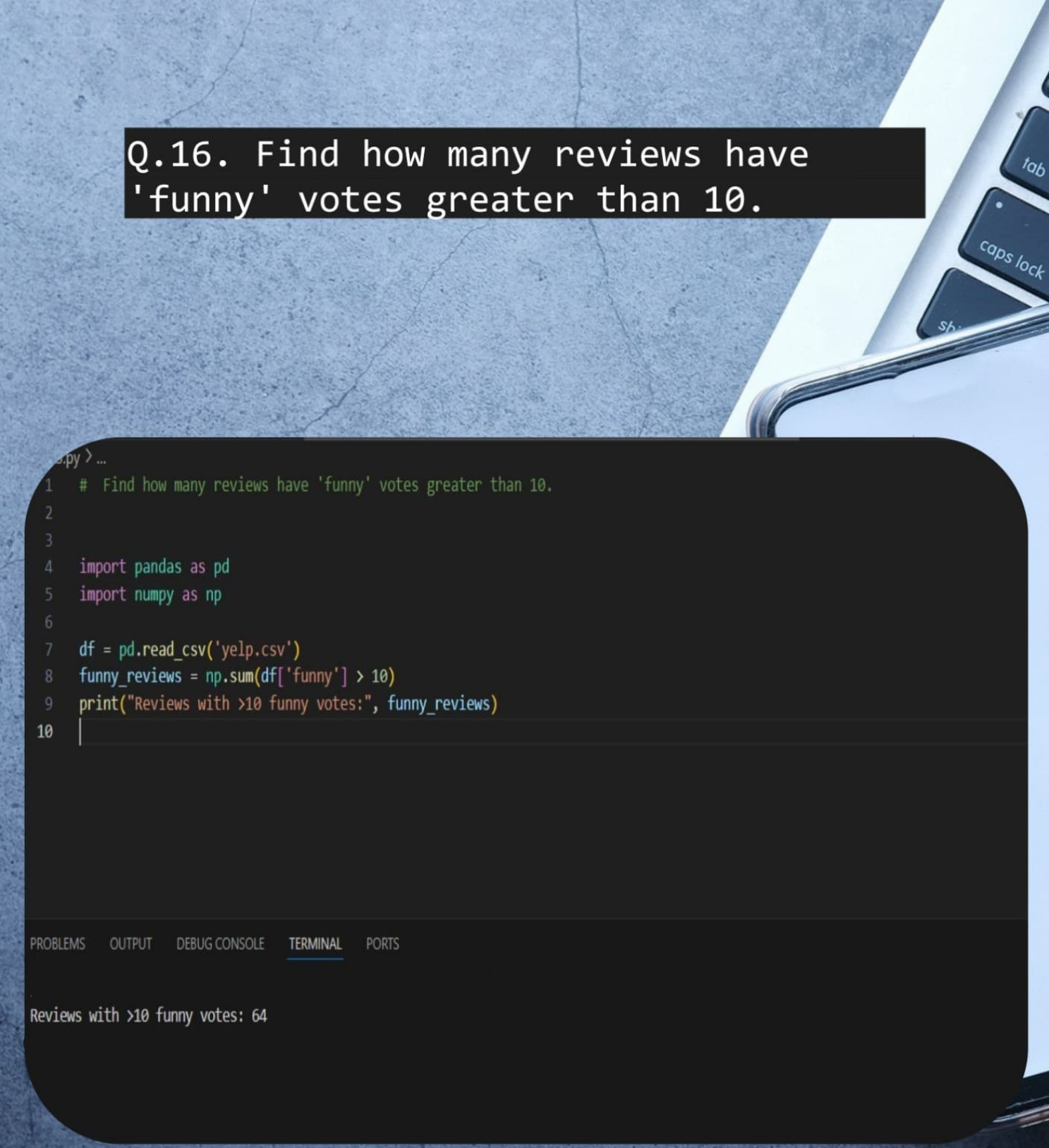
Q.15 Find the user who wrote the most reviews.

```python
#  Find the user who wrote the most reviews.

import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
top_user = df['user_id'].value_counts().idxmax()
print("User with most reviews:", top_user)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
User with most reviews: fczQCSmaWF78toLEmb0Zsw
```

Q.16. Find how many reviews have 'funny' votes greater than 10.



```python
# Find how many reviews have 'funny' votes greater than 10.


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
funny_reviews = np.sum(df['funny'] > 10)
print("Reviews with >10 funny votes:", funny_reviews)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Reviews with >10 funny votes: 64
```

Q.17] Find the review posted most recently.

```python
#  Find the review posted most recently.


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
df['date'] = pd.to_datetime(df['date'])
latest_idx = np.argmax(df['date'])
latest_review = df.iloc[latest_idx]
print(latest_review[['review_id', 'date']])

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
review_id    lI8Bo4AMQT7C-zNLgvRasw
date           2013-01-05 00:00:00
Name: 633, dtype: object
```

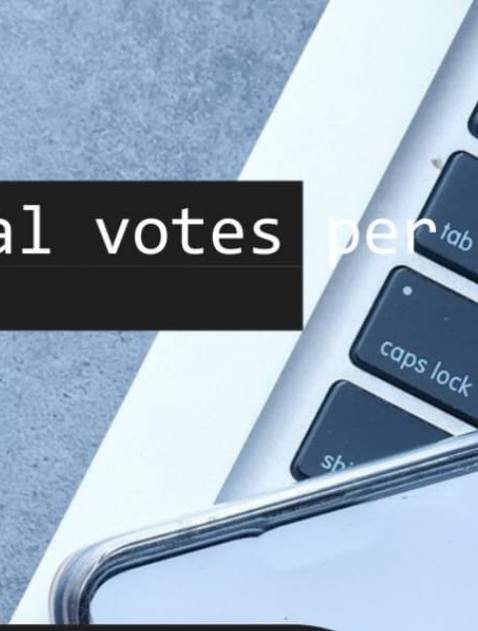
Q.18] Create a new column 'total votes' (cool + useful + funny).

```
.py > ...
1   #  Create a new column 'total_votes' (cool + useful + funny).
2
3
4   import pandas as pd
5   import numpy as np
6
7   df = pd.read_csv('yelp.csv')
8   df['total_votes'] = np.add(np.add(df['cool'], df['useful']), df['funny'])
9   print(df[['review_id', 'total_votes']].head())
10
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
              review_id  total_votes
0  fWKvX83p0-ka4JS3dc6E5A           7
1  IjZ33sJrzXqU-0X6U8NwyA           0
2  IESLBzqUCLdSzSqm0eCSxQ           1
3  G-WvGaISbqqaMHlNnByodA           3
4  1uJFq2r5QfJG_6ExMRCaGw           0
```

Q.19 Find average total votes per review.




```
18.py > ...
1   #  Find  total votes per review.
2
3
4   import pandas as pd
5   import numpy as np
6
7   df = pd.read_csv('yelp.csv')
8   df['total_votes'] = np.add(np.add(df['cool'], df['useful']), df['funny'])
9   average_total_votes = np.mean(df['total_votes'])
10  print("Average total votes per review:", average_total_votes)
11
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

Average total votes per review: 2.9874
```

Q.20] Find the review with the longest text (most characters).

```python
#  Find the review with the longest text (most characters).


import pandas as pd
import numpy as np

df = pd.read_csv('yelp.csv')
text_lengths = df['text'].str.len()
longest_idx = np.argmax(text_lengths)
longest_review = df.iloc[longest_idx]
print(longest_review[['review_id', 'text']])

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
review_id                              6jRs2P6zTYMn36fVnCu1Zw
text         In our continuing quest to identify cool, loca...
Name: 55, dtype: object
```