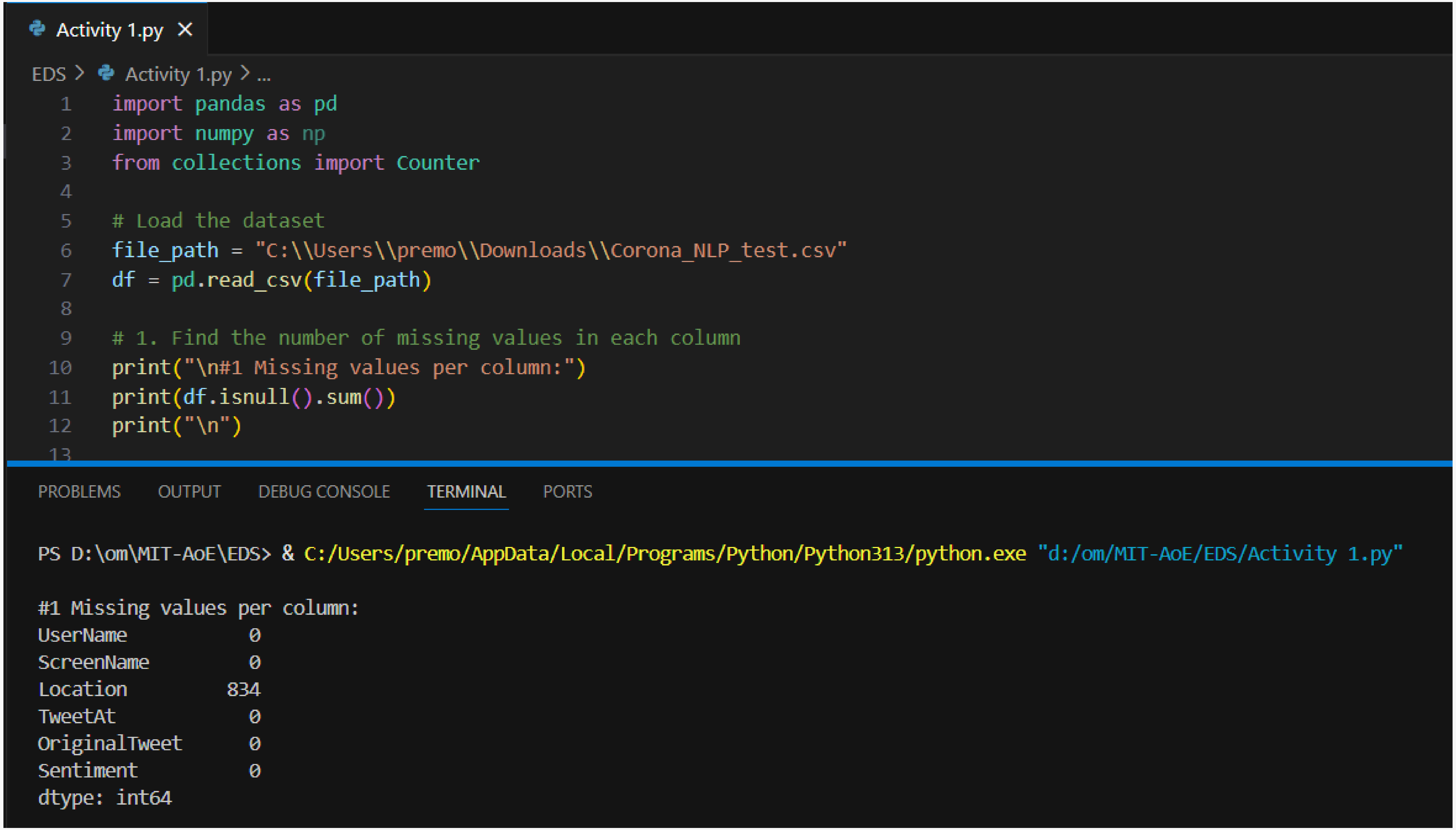
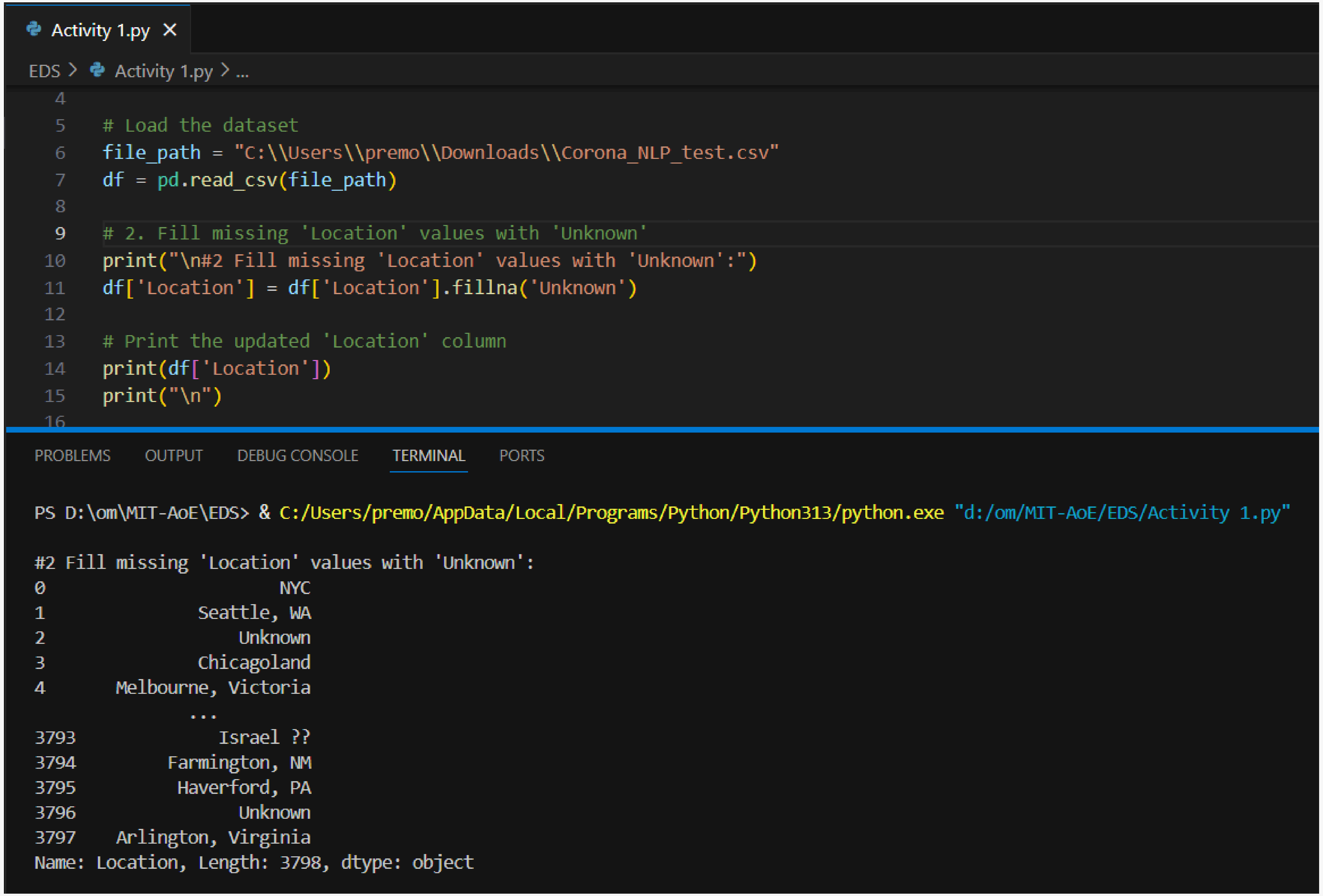
EDS THEORY ACTIVITY NO. 1

Name:- Om Channawar
Division:- CS7
Roll No.:- CS7-54
PRN:- 202401110044



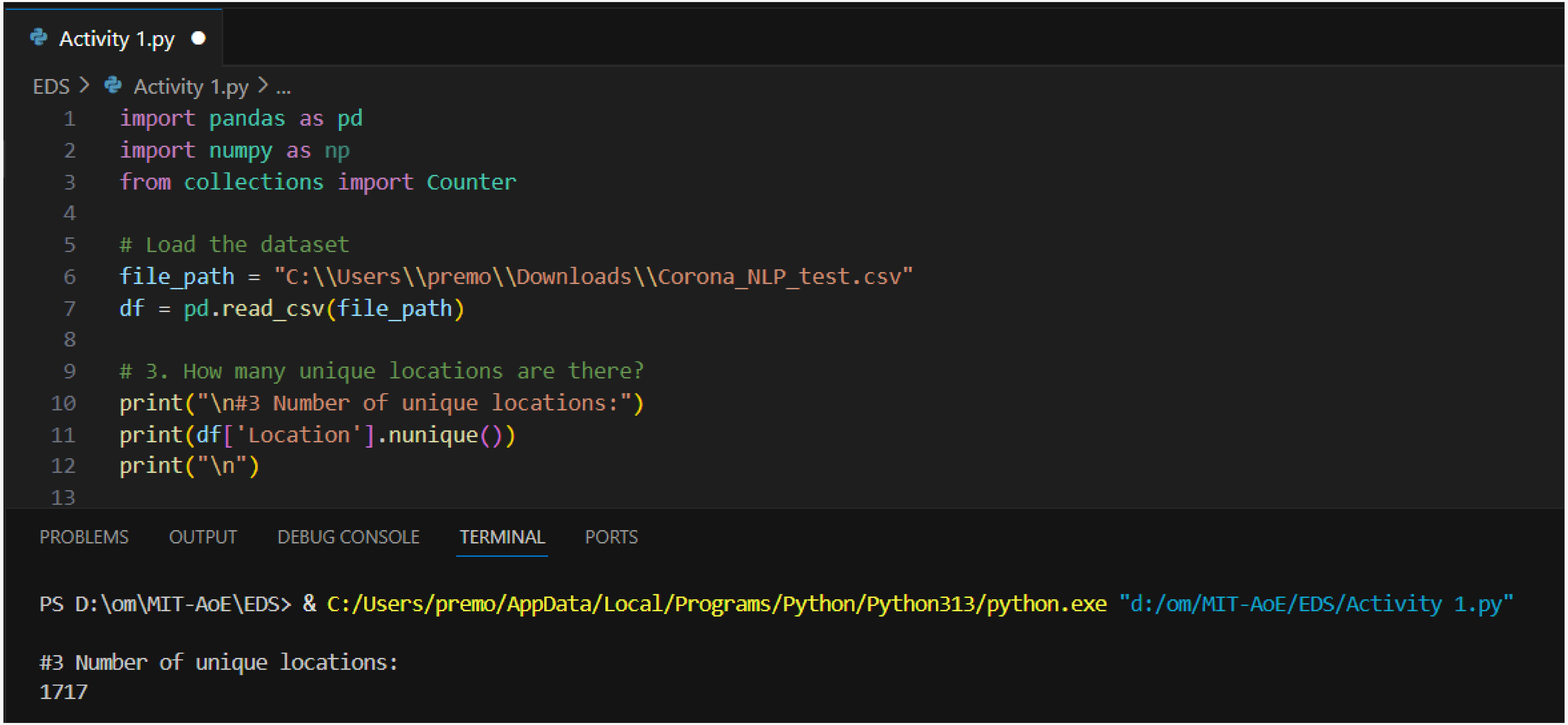
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 1. Find the number of missing values in each column
print("\n#1 Missing values per column:")
print(df.isnull().sum())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#1 Missing values per column:
UserName           0
ScreenName         0
Location         834
TweetAt            0
OriginalTweet      0
Sentiment          0
dtype: int64
```



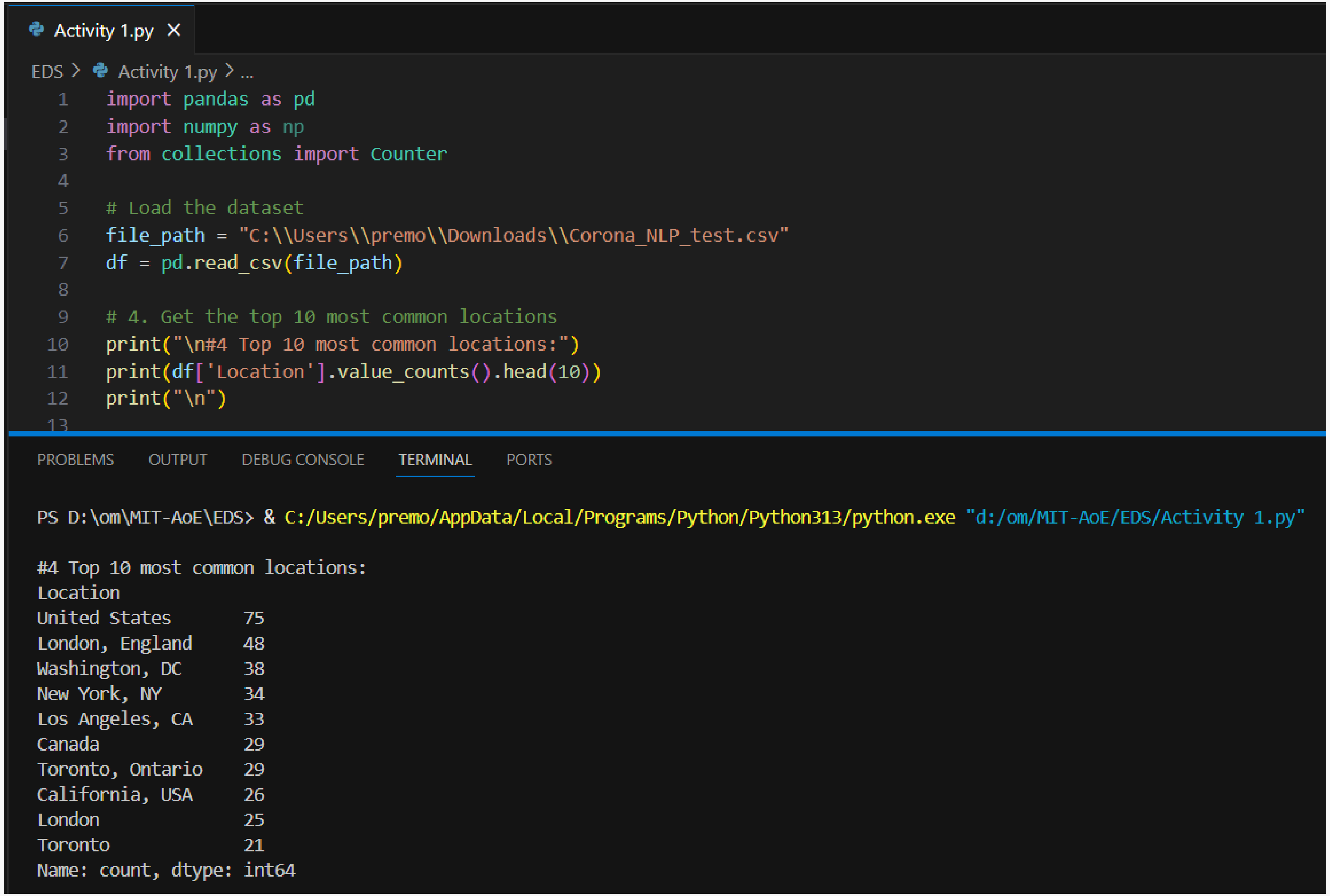
```python
# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 2. Fill missing 'Location' values with 'Unknown'
print("\n#2 Fill missing 'Location' values with 'Unknown':")
df['Location'] = df['Location'].fillna('Unknown')

# Print the updated 'Location' column
print(df['Location'])
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#2 Fill missing 'Location' values with 'Unknown':
0                        NYC
1                Seattle, WA
2                    Unknown
3                Chicagoland
4        Melbourne, Victoria
               ...
3793               Israel ??
3794          Farmington, NM
3795           Haverford, PA
3796                 Unknown
3797     Arlington, Virginia
Name: Location, Length: 3798, dtype: object
```



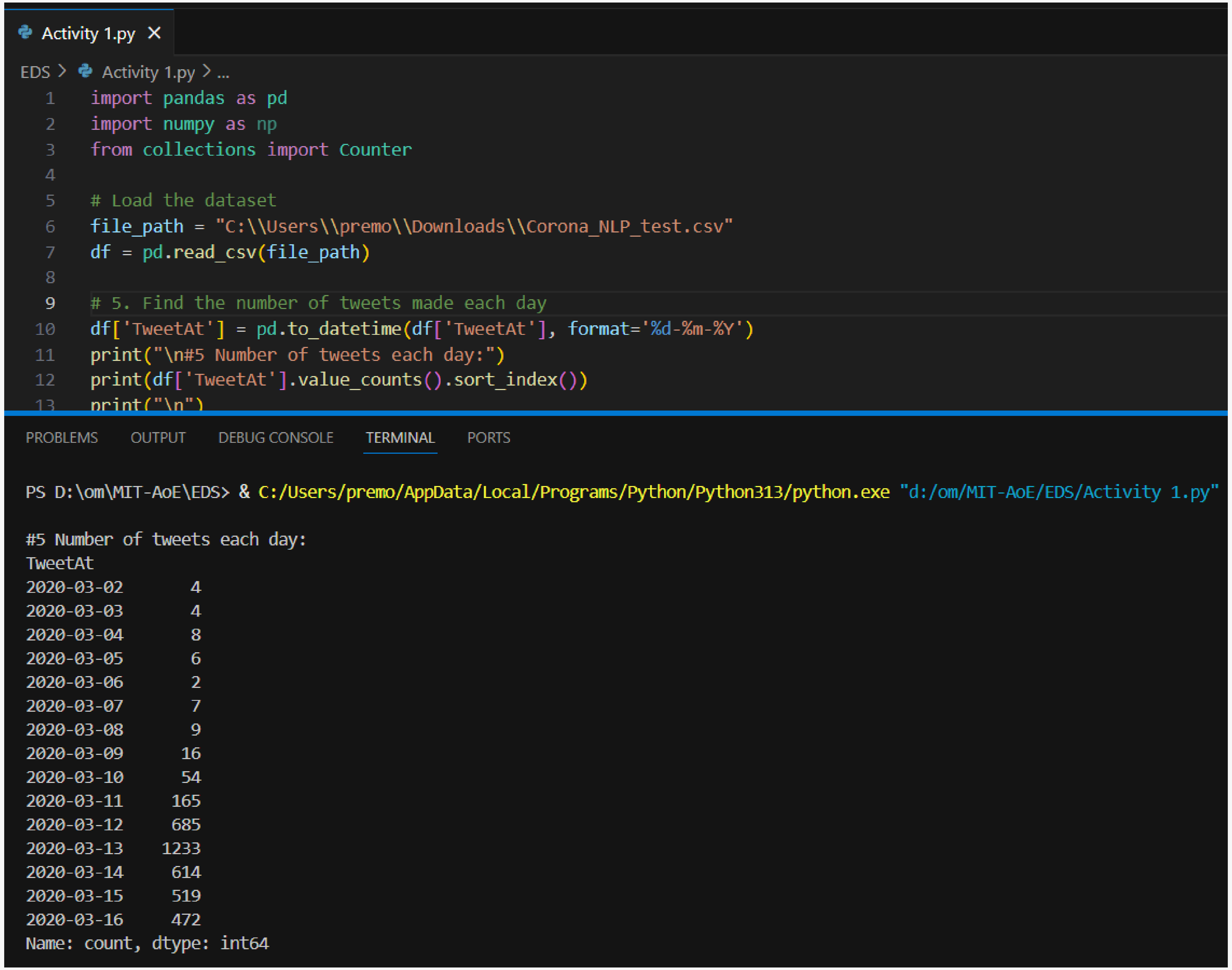
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 3. How many unique locations are there?
print("\n#3 Number of unique locations:")
print(df['Location'].nunique())
print("\n")

```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#3 Number of unique locations:
1717
```



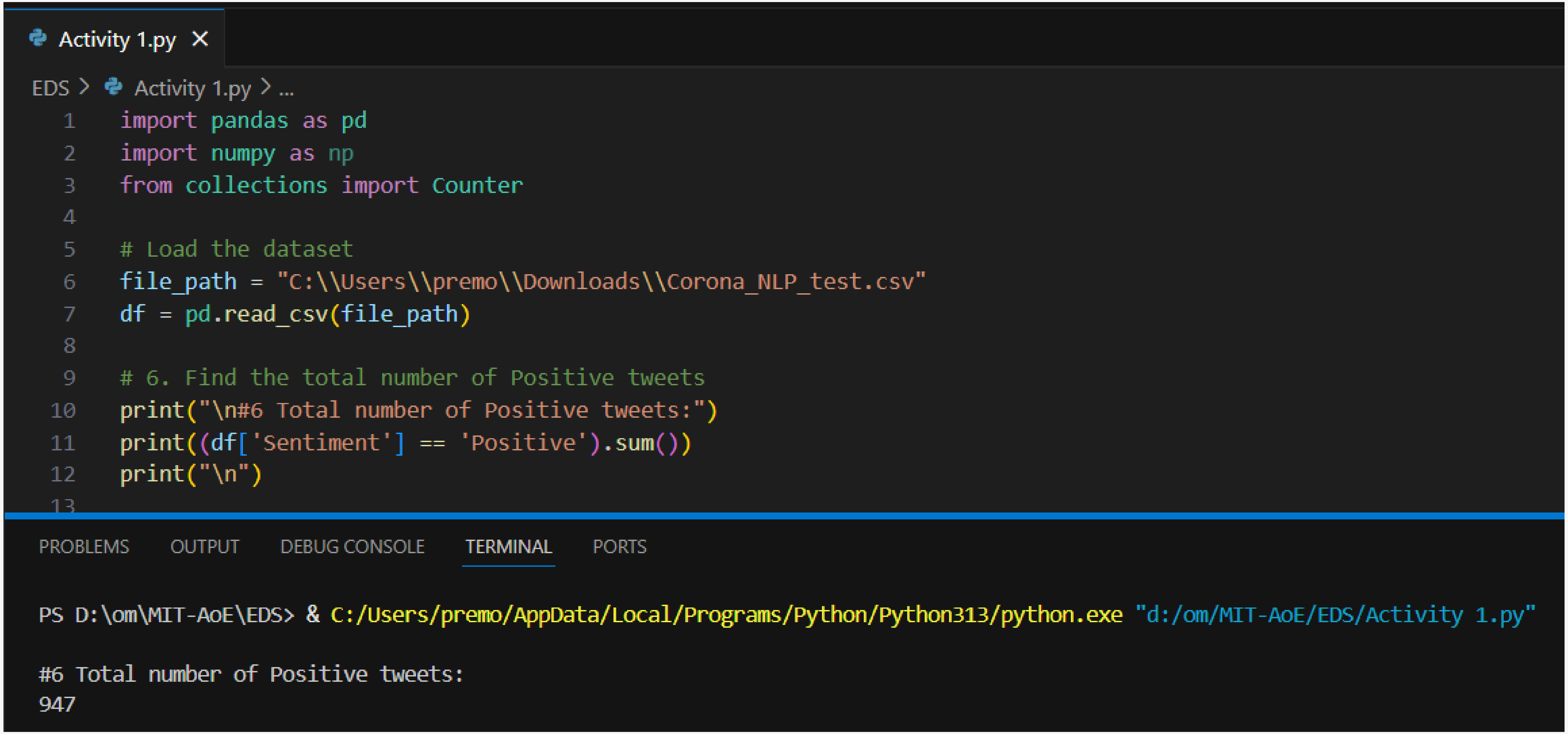
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 4. Get the top 10 most common locations
print("\n#4 Top 10 most common locations:")
print(df['Location'].value_counts().head(10))
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#4 Top 10 most common locations:
Location
United States       75
London, England     48
Washington, DC      38
New York, NY        34
Los Angeles, CA     33
Canada              29
Toronto, Ontario    29
California, USA     26
London              25
Toronto             21
Name: count, dtype: int64
```



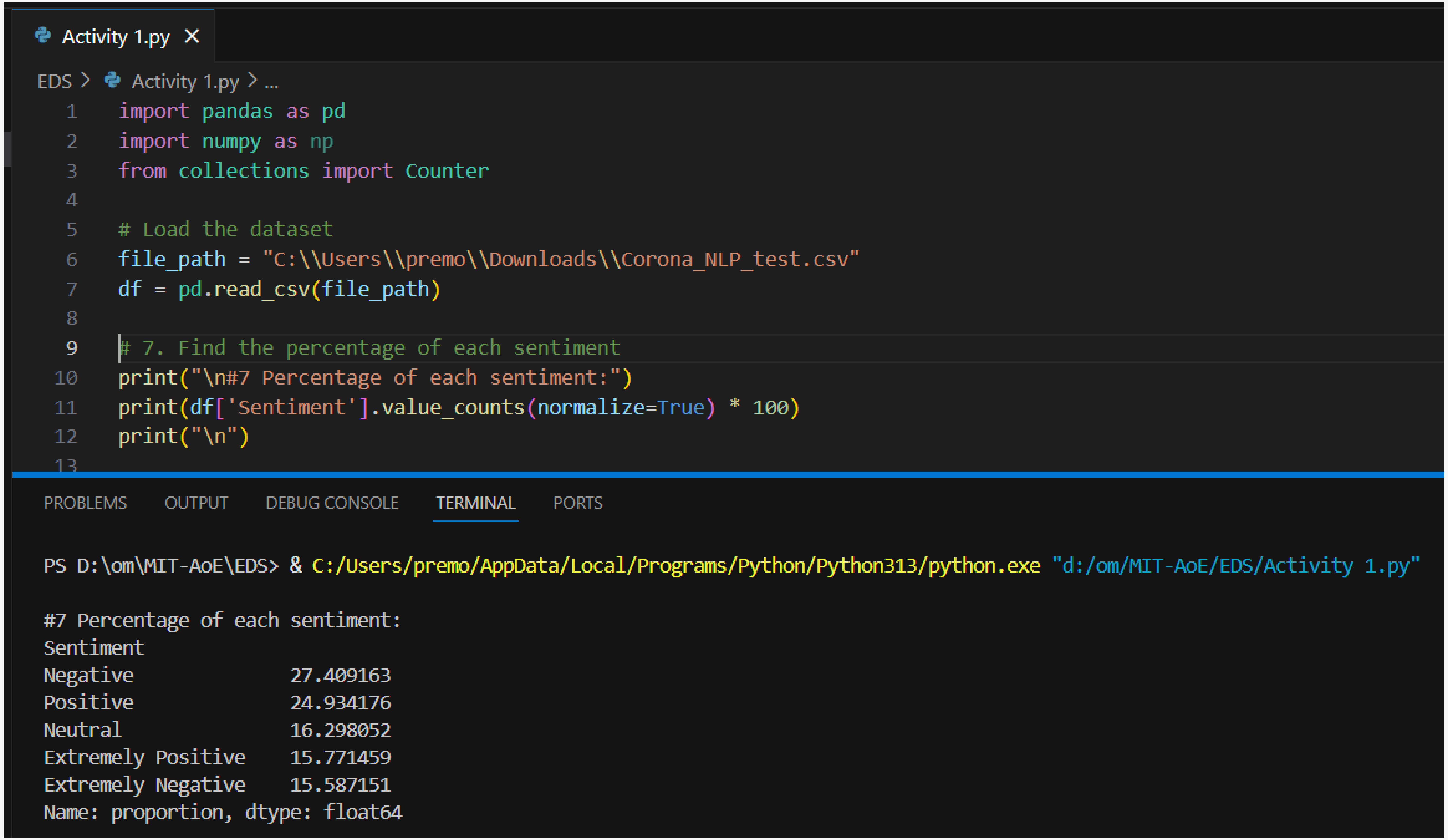
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 5. Find the number of tweets made each day
df['TweetAt'] = pd.to_datetime(df['TweetAt'], format='%d-%m-%Y')
print("\n#5 Number of tweets each day:")
print(df['TweetAt'].value_counts().sort_index())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#5 Number of tweets each day:
TweetAt
2020-03-02       4
2020-03-03       4
2020-03-04       8
2020-03-05       6
2020-03-06       2
2020-03-07       7
2020-03-08       9
2020-03-09      16
2020-03-10      54
2020-03-11     165
2020-03-12     685
2020-03-13    1233
2020-03-14     614
2020-03-15     519
2020-03-16     472
Name: count, dtype: int64
```



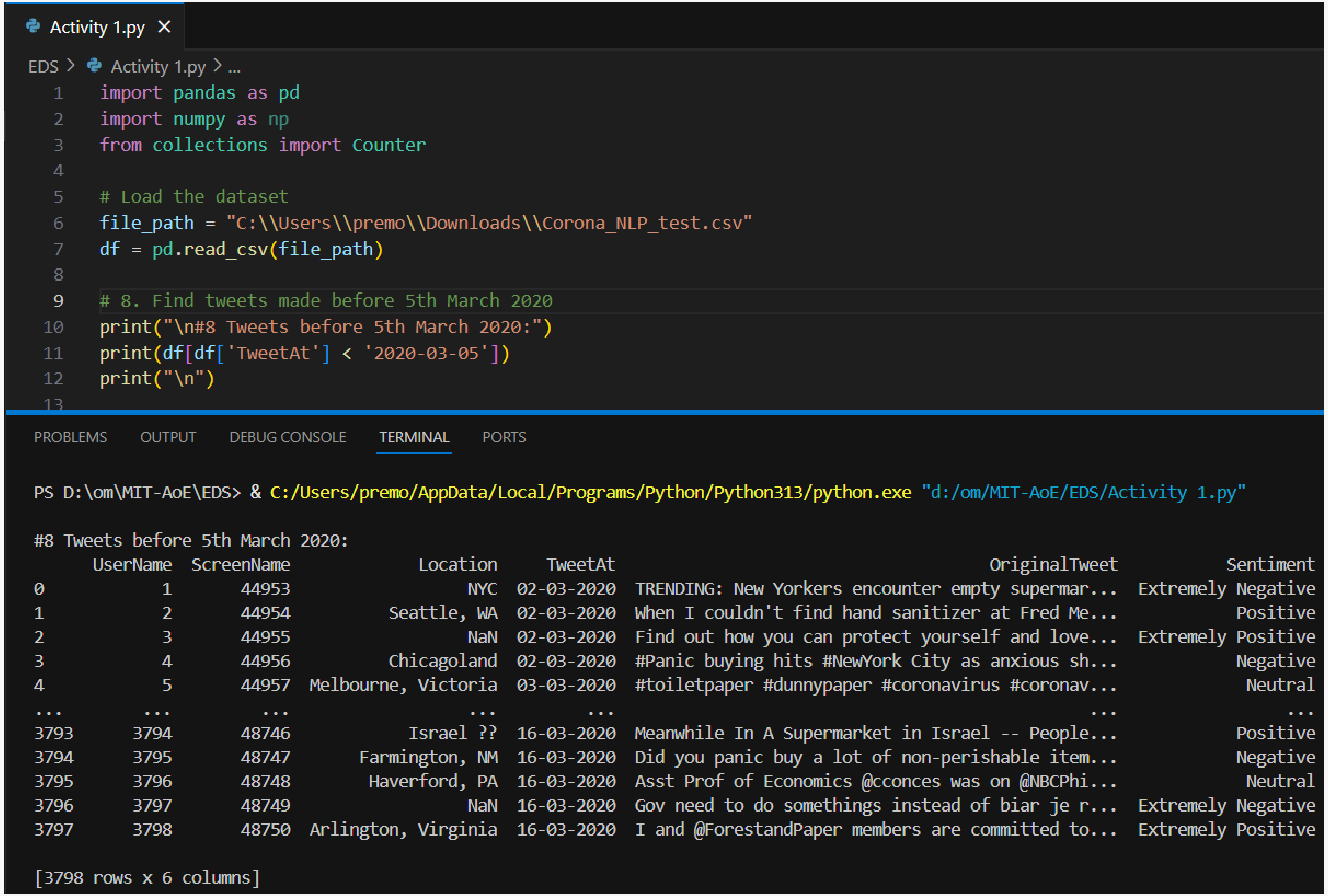
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 6. Find the total number of Positive tweets
print("\n#6 Total number of Positive tweets:")
print((df['Sentiment'] == 'Positive').sum())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#6 Total number of Positive tweets:
947
```



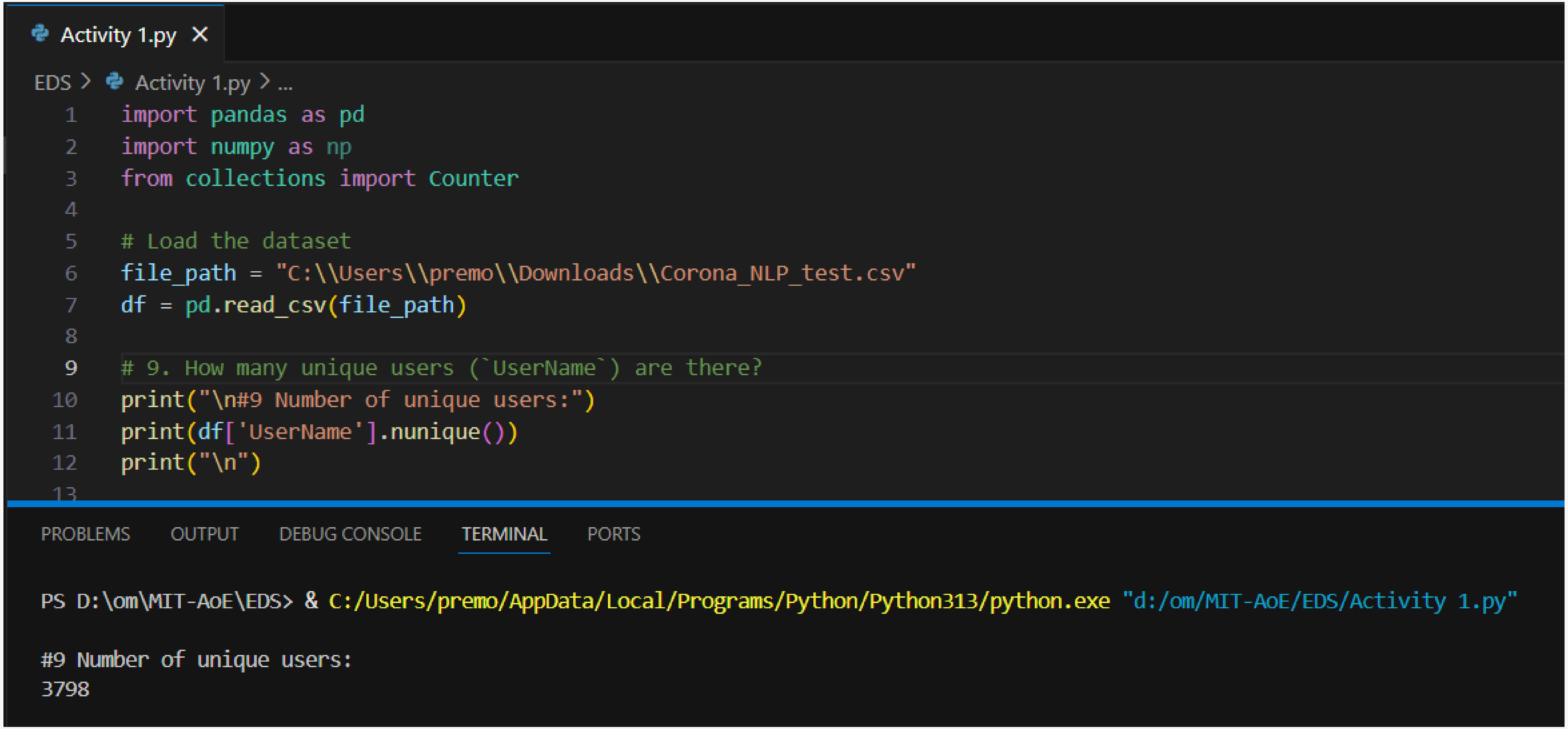
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 7. Find the percentage of each sentiment
print("\n#7 Percentage of each sentiment:")
print(df['Sentiment'].value_counts(normalize=True) * 100)
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#7 Percentage of each sentiment:
Sentiment
Negative              27.409163
Positive              24.934176
Neutral               16.298052
Extremely Positive    15.771459
Extremely Negative    15.587151
Name: proportion, dtype: float64
```



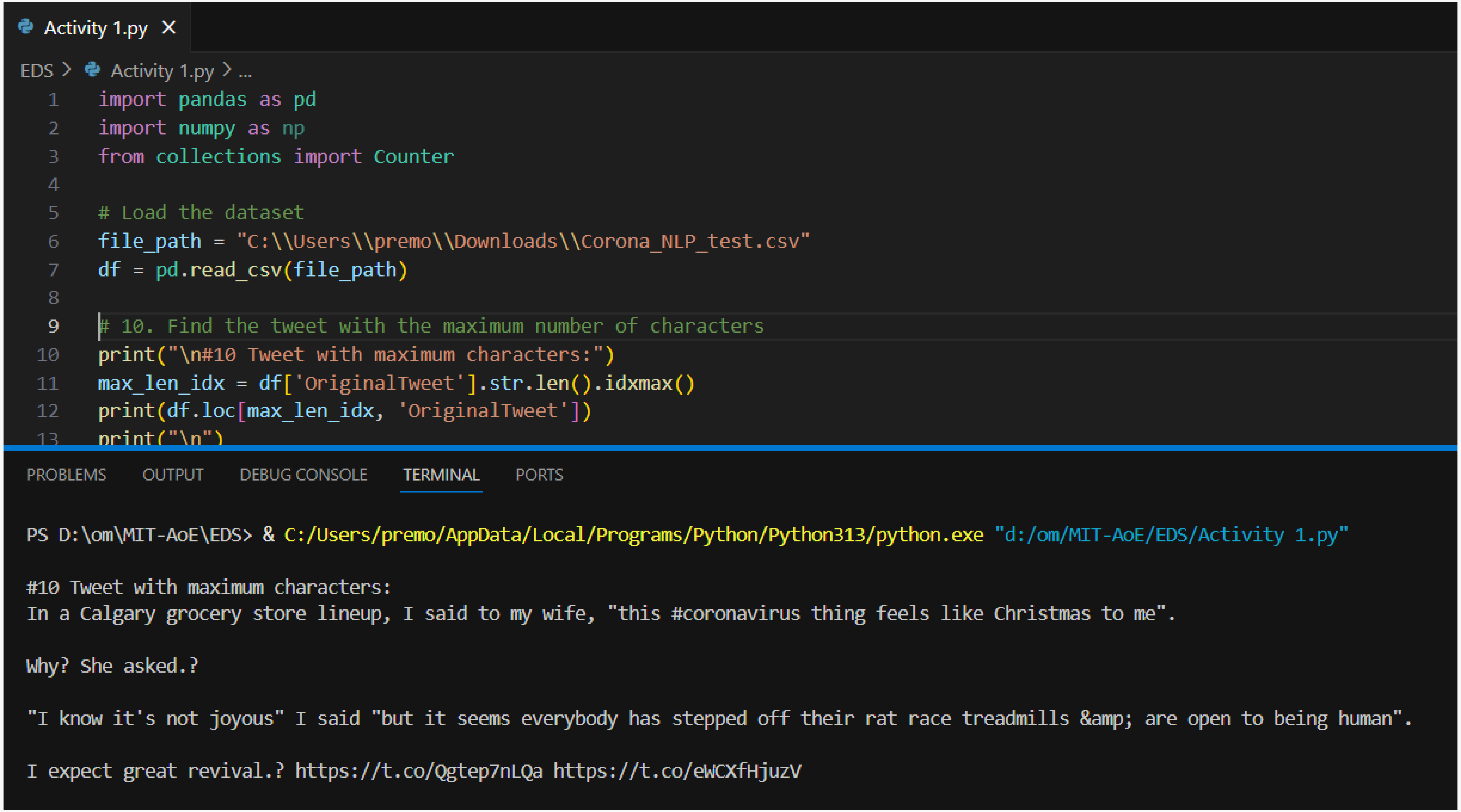
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 8. Find tweets made before 5th March 2020
print("\n#8 Tweets before 5th March 2020:")
print(df[df['TweetAt'] < '2020-03-05'])
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#8 Tweets before 5th March 2020:
      UserName  ScreenName             Location     TweetAt                                       OriginalTweet           Sentiment
0            1       44953                  NYC  02-03-2020  TRENDING: New Yorkers encounter empty supermar...  Extremely Negative
1            2       44954          Seattle, WA  02-03-2020  When I couldn't find hand sanitizer at Fred Me...            Positive
2            3       44955                  NaN  02-03-2020  Find out how you can protect yourself and love...  Extremely Positive
3            4       44956          Chicagoland  02-03-2020  #Panic buying hits #NewYork City as anxious sh...            Negative
4            5       44957  Melbourne, Victoria  03-03-2020  #toiletpaper #dunnypaper #coronavirus #coronav...             Neutral
...        ...         ...                  ...         ...                                                ...                 ...
3793      3794       48746            Israel ??  16-03-2020  Meanwhile In A Supermarket in Israel -- People...            Positive
3794      3795       48747       Farmington, NM  16-03-2020  Did you panic buy a lot of non-perishable item...            Negative
3795      3796       48748        Haverford, PA  16-03-2020  Asst Prof of Economics @cconces was on @NBCPhi...             Neutral
3796      3797       48749                  NaN  16-03-2020  Gov need to do somethings instead of biar je r...  Extremely Negative
3797      3798       48750  Arlington, Virginia  16-03-2020  I and @ForestandPaper members are committed to...  Extremely Positive

[3798 rows x 6 columns]
```



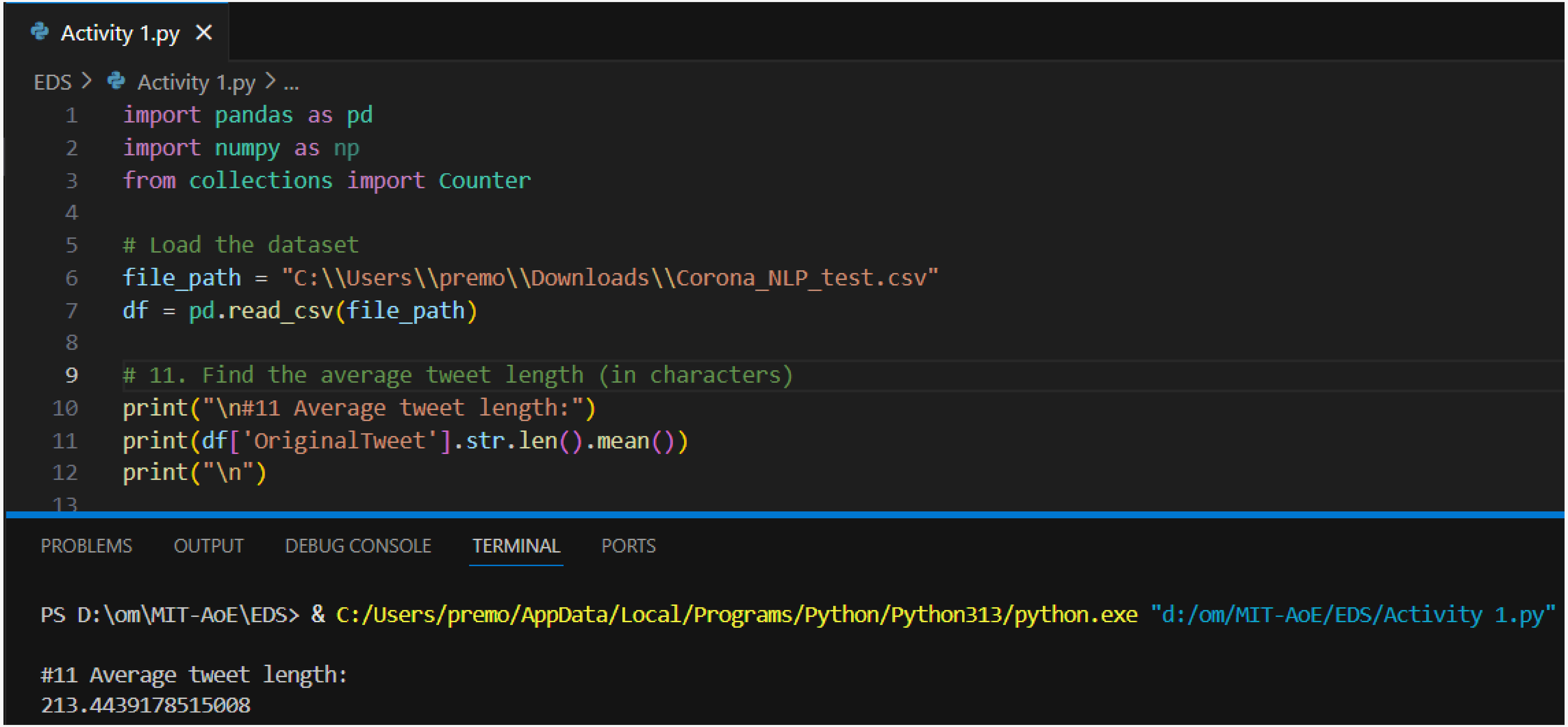
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 9. How many unique users (`UserName`) are there?
print("\n#9 Number of unique users:")
print(df['UserName'].nunique())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#9 Number of unique users:
3798
```



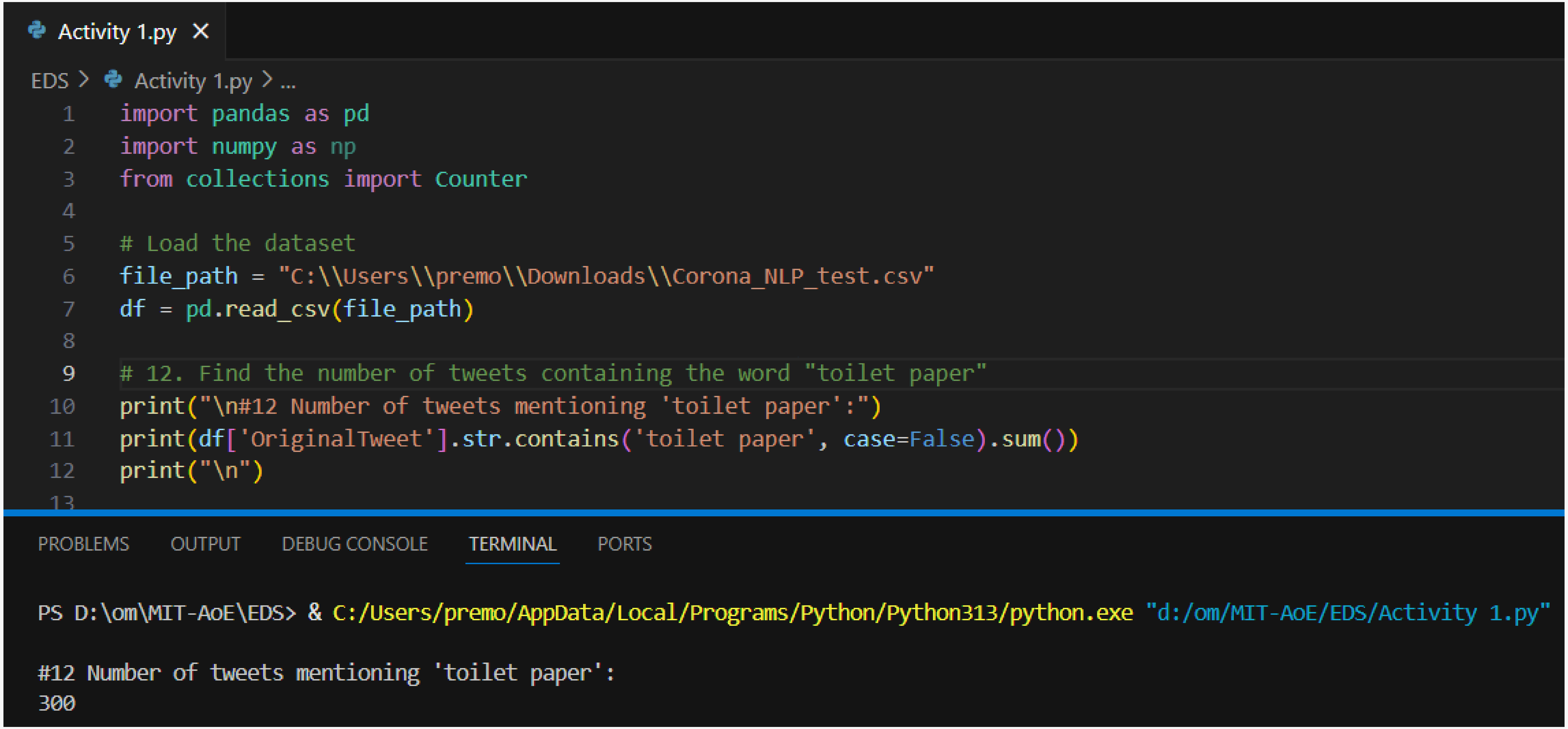
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 10. Find the tweet with the maximum number of characters
print("\n#10 Tweet with maximum characters:")
max_len_idx = df['OriginalTweet'].str.len().idxmax()
print(df.loc[max_len_idx, 'OriginalTweet'])
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#10 Tweet with maximum characters:
In a Calgary grocery store lineup, I said to my wife, "this #coronavirus thing feels like Christmas to me".

Why? She asked.?

"I know it's not joyous" I said "but it seems everybody has stepped off their rat race treadmills &amp; are open to being human".

I expect great revival.? https://t.co/Qgtep7nLQa https://t.co/eWCXfHjuzV
```



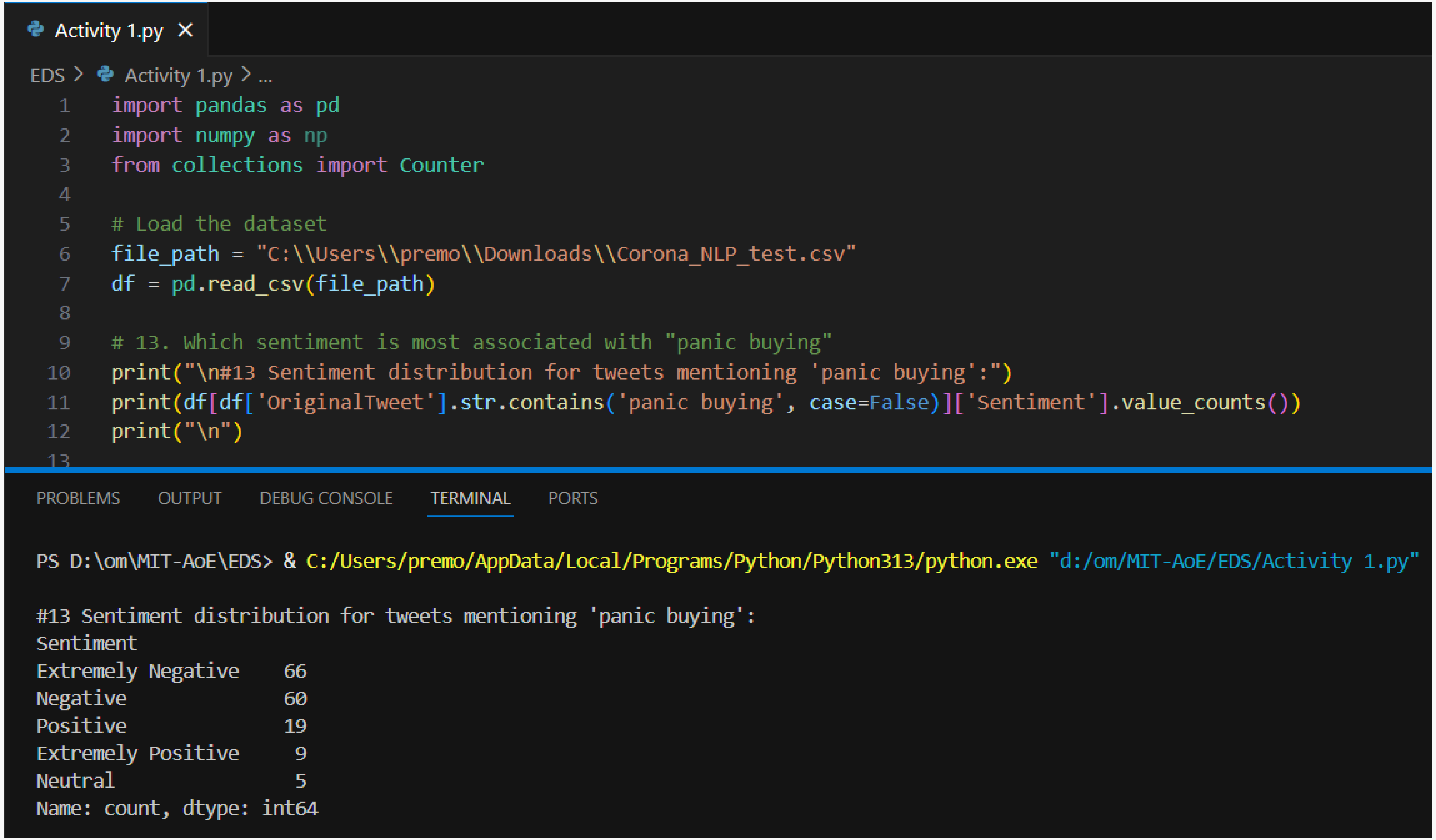
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 11. Find the average tweet length (in characters)
print("\n#11 Average tweet length:")
print(df['OriginalTweet'].str.len().mean())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#11 Average tweet length:
213.4439178515008
```



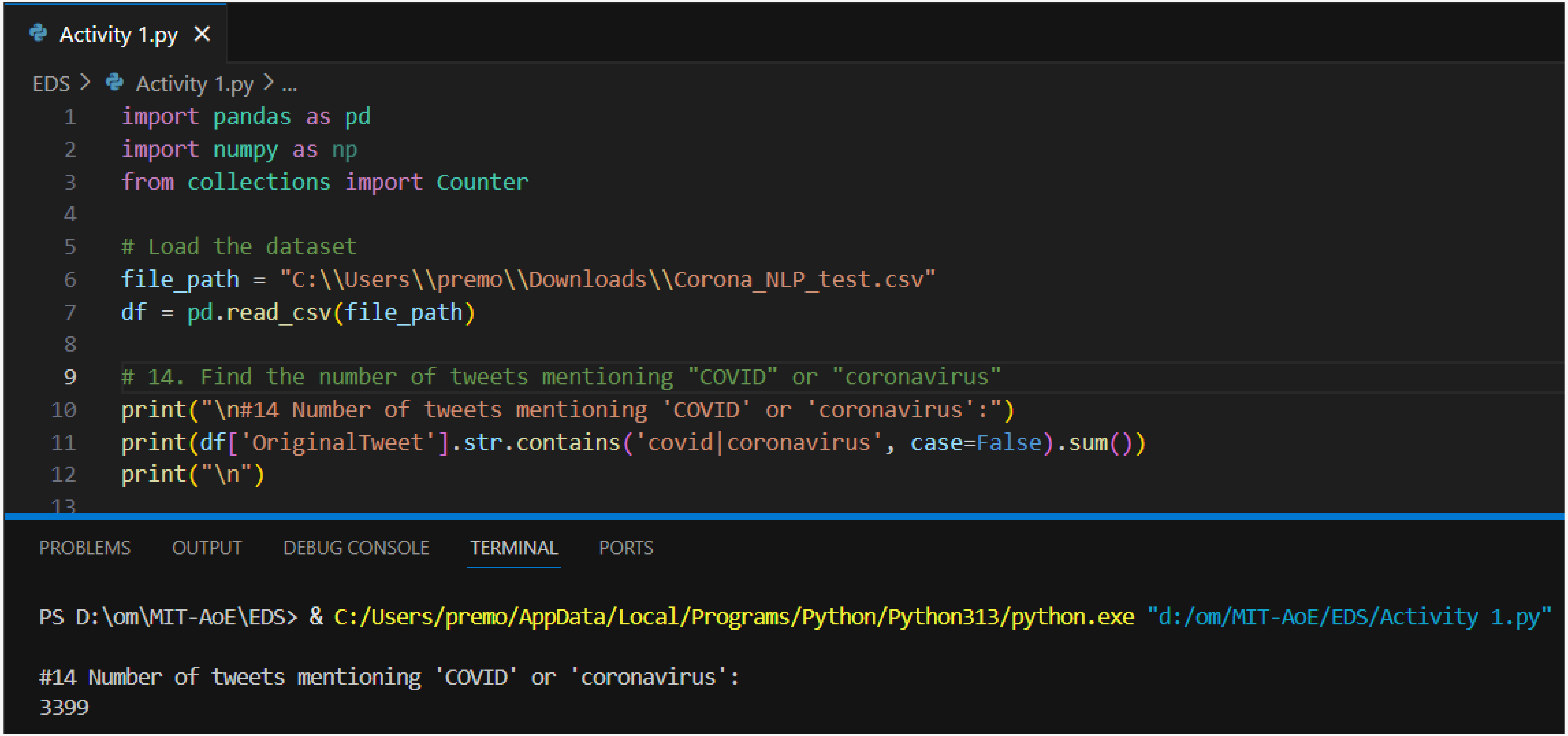
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 12. Find the number of tweets containing the word "toilet paper"
print("\n#12 Number of tweets mentioning 'toilet paper':")
print(df['OriginalTweet'].str.contains('toilet paper', case=False).sum())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#12 Number of tweets mentioning 'toilet paper':
300
```



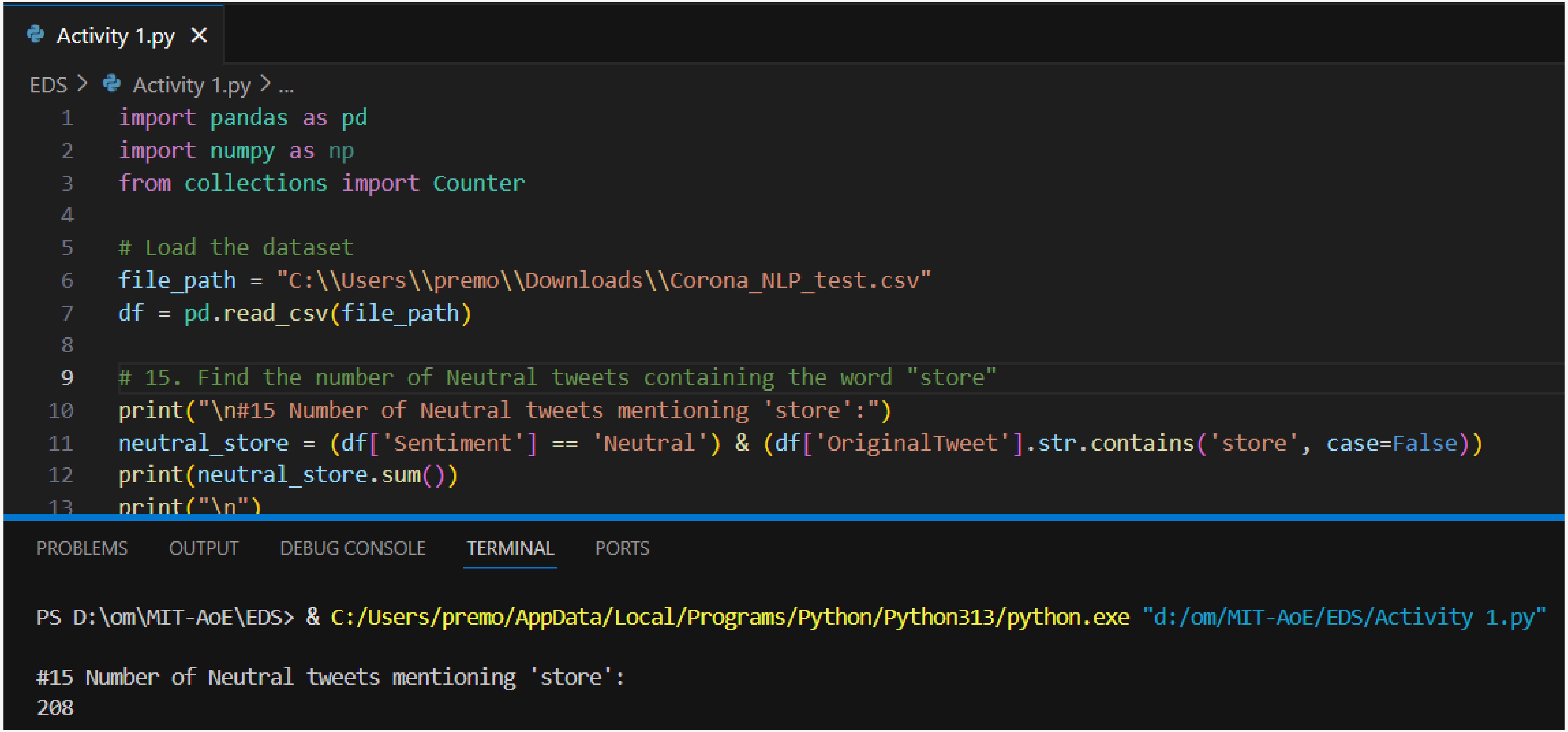
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 13. Which sentiment is most associated with "panic buying"
print("\n#13 Sentiment distribution for tweets mentioning 'panic buying':")
print(df[df['OriginalTweet'].str.contains('panic buying', case=False)]['Sentiment'].value_counts())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#13 Sentiment distribution for tweets mentioning 'panic buying':
Sentiment
Extremely Negative    66
Negative              60
Positive              19
Extremely Positive     9
Neutral                5
Name: count, dtype: int64
```



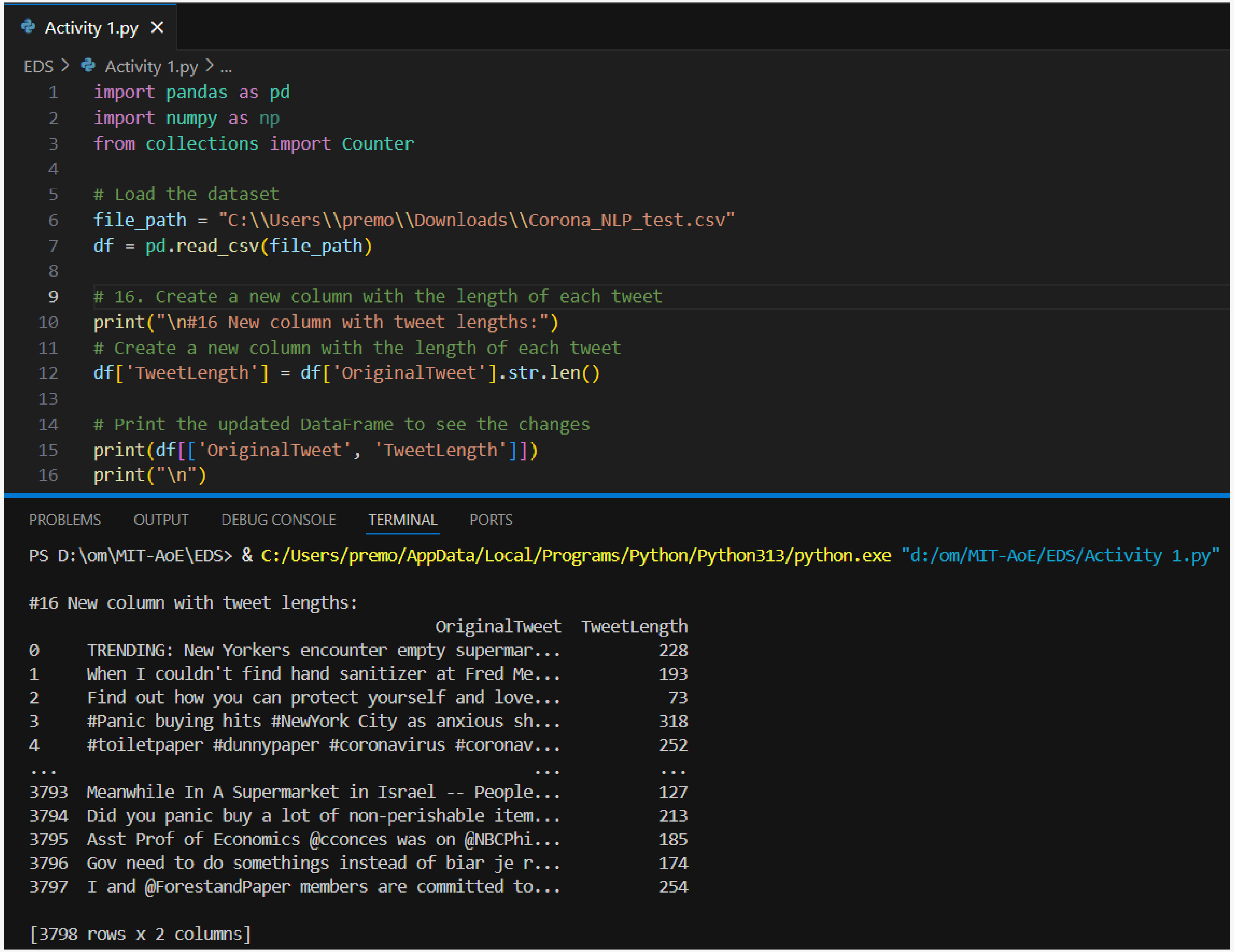
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 14. Find the number of tweets mentioning "COVID" or "coronavirus"
print("\n#14 Number of tweets mentioning 'COVID' or 'coronavirus':")
print(df['OriginalTweet'].str.contains('covid|coronavirus', case=False).sum())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#14 Number of tweets mentioning 'COVID' or 'coronavirus':
3399
```



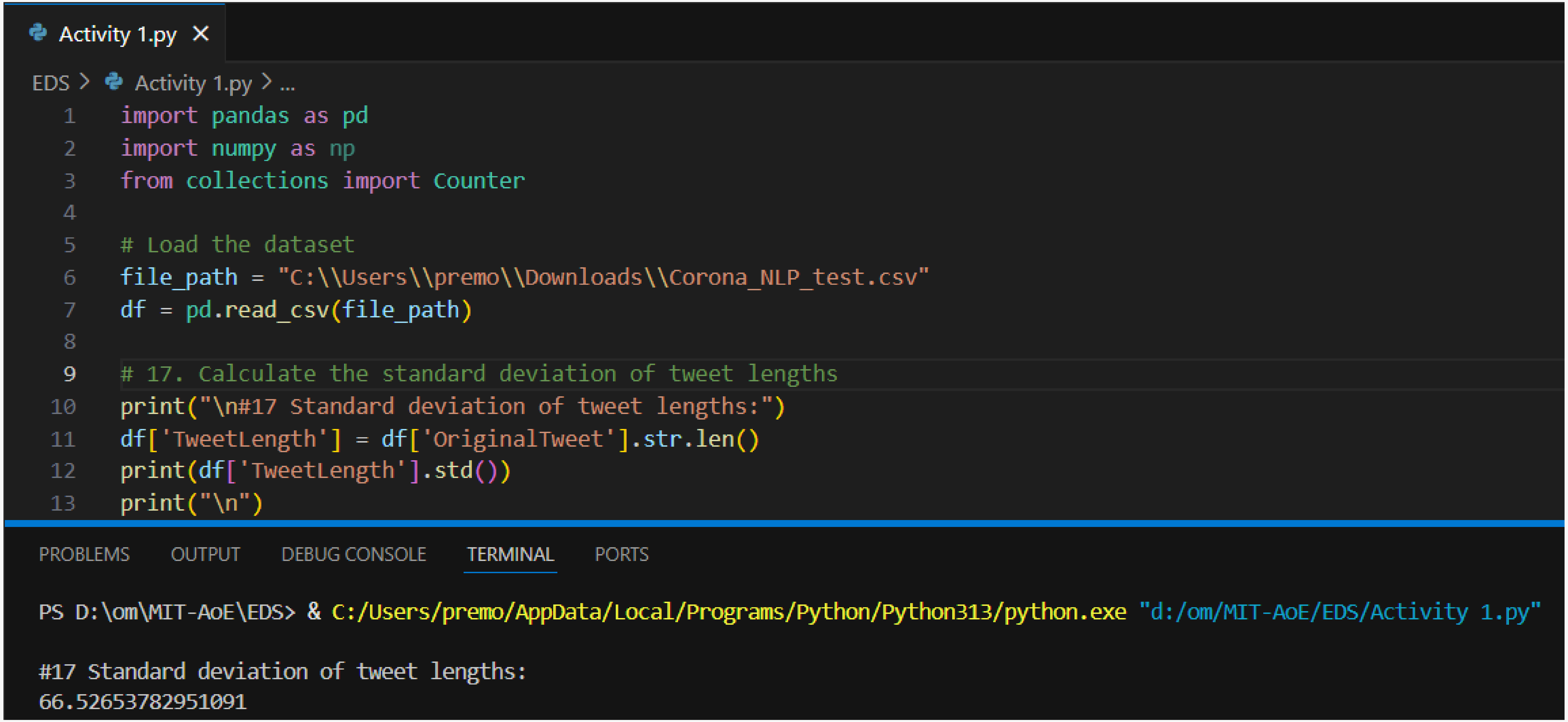
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 15. Find the number of Neutral tweets containing the word "store"
print("\n#15 Number of Neutral tweets mentioning 'store':")
neutral_store = (df['Sentiment'] == 'Neutral') & (df['OriginalTweet'].str.contains('store', case=False))
print(neutral_store.sum())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#15 Number of Neutral tweets mentioning 'store':
208
```



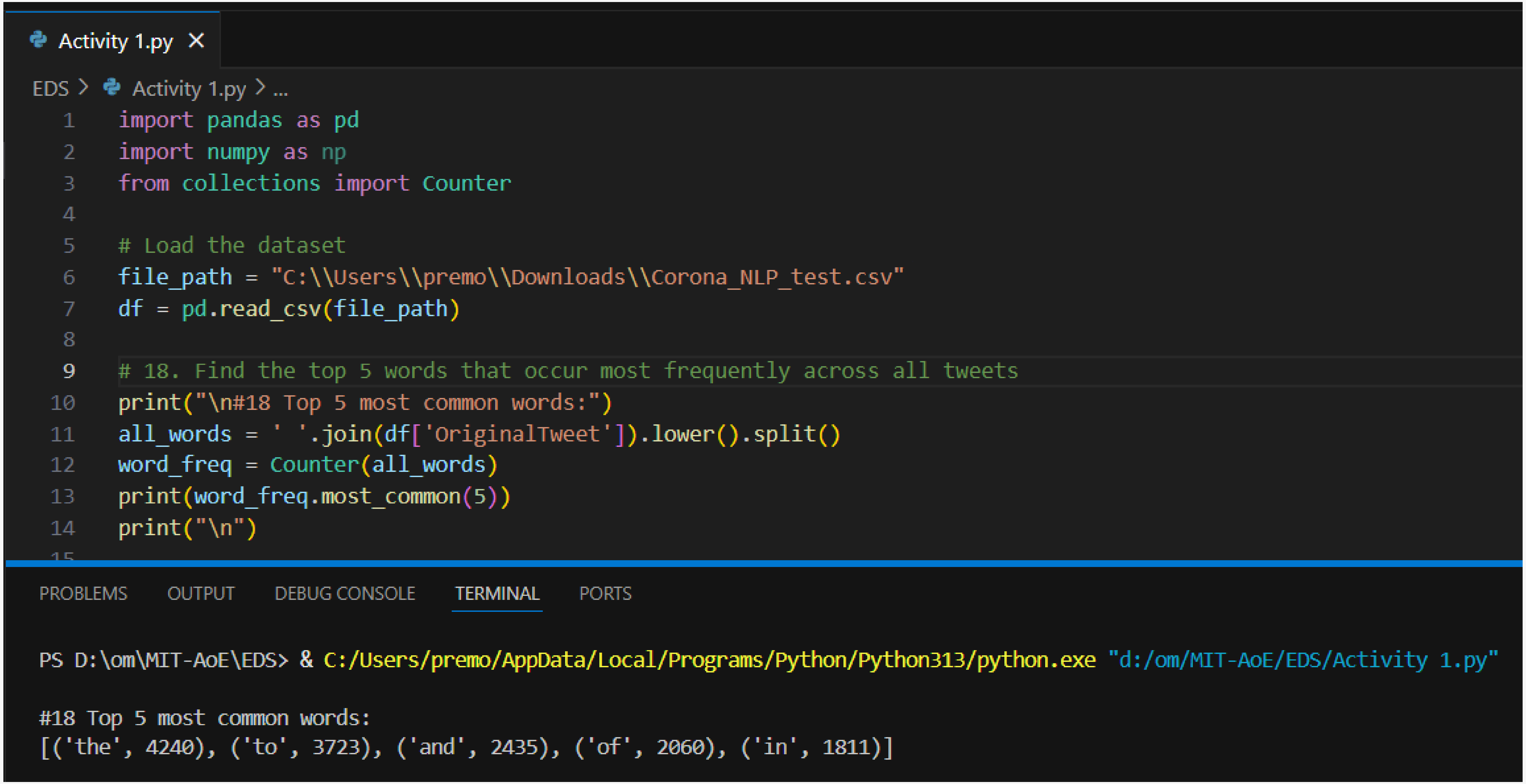
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 16. Create a new column with the length of each tweet
print("\n#16 New column with tweet lengths:")
# Create a new column with the length of each tweet
df['TweetLength'] = df['OriginalTweet'].str.len()

# Print the updated DataFrame to see the changes
print(df[['OriginalTweet', 'TweetLength']])
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#16 New column with tweet lengths:
                                          OriginalTweet  TweetLength
0     TRENDING: New Yorkers encounter empty supermar...          228
1     When I couldn't find hand sanitizer at Fred Me...          193
2     Find out how you can protect yourself and love...           73
3     #Panic buying hits #NewYork City as anxious sh...          318
4     #toiletpaper #dunnypaper #coronavirus #coronav...          252
...                                                 ...          ...
3793  Meanwhile In A Supermarket in Israel -- People...          127
3794  Did you panic buy a lot of non-perishable item...          213
3795  Asst Prof of Economics @cconces was on @NBCPhi...          185
3796  Gov need to do somethings instead of biar je r...          174
3797  I and @ForestandPaper members are committed to...          254

[3798 rows x 2 columns]
```



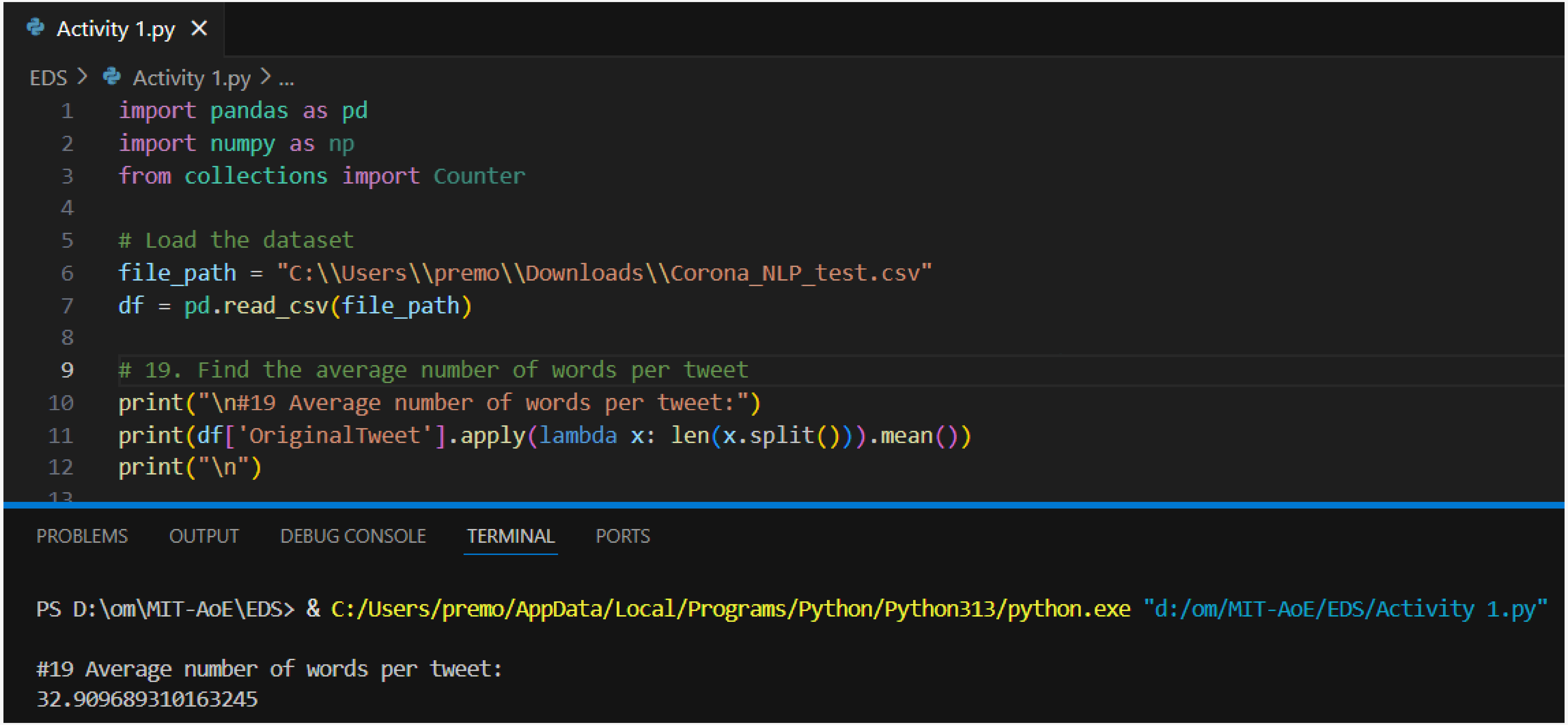
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 17. Calculate the standard deviation of tweet lengths
print("\n#17 Standard deviation of tweet lengths:")
df['TweetLength'] = df['OriginalTweet'].str.len()
print(df['TweetLength'].std())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#17 Standard deviation of tweet lengths:
66.52653782951091
```



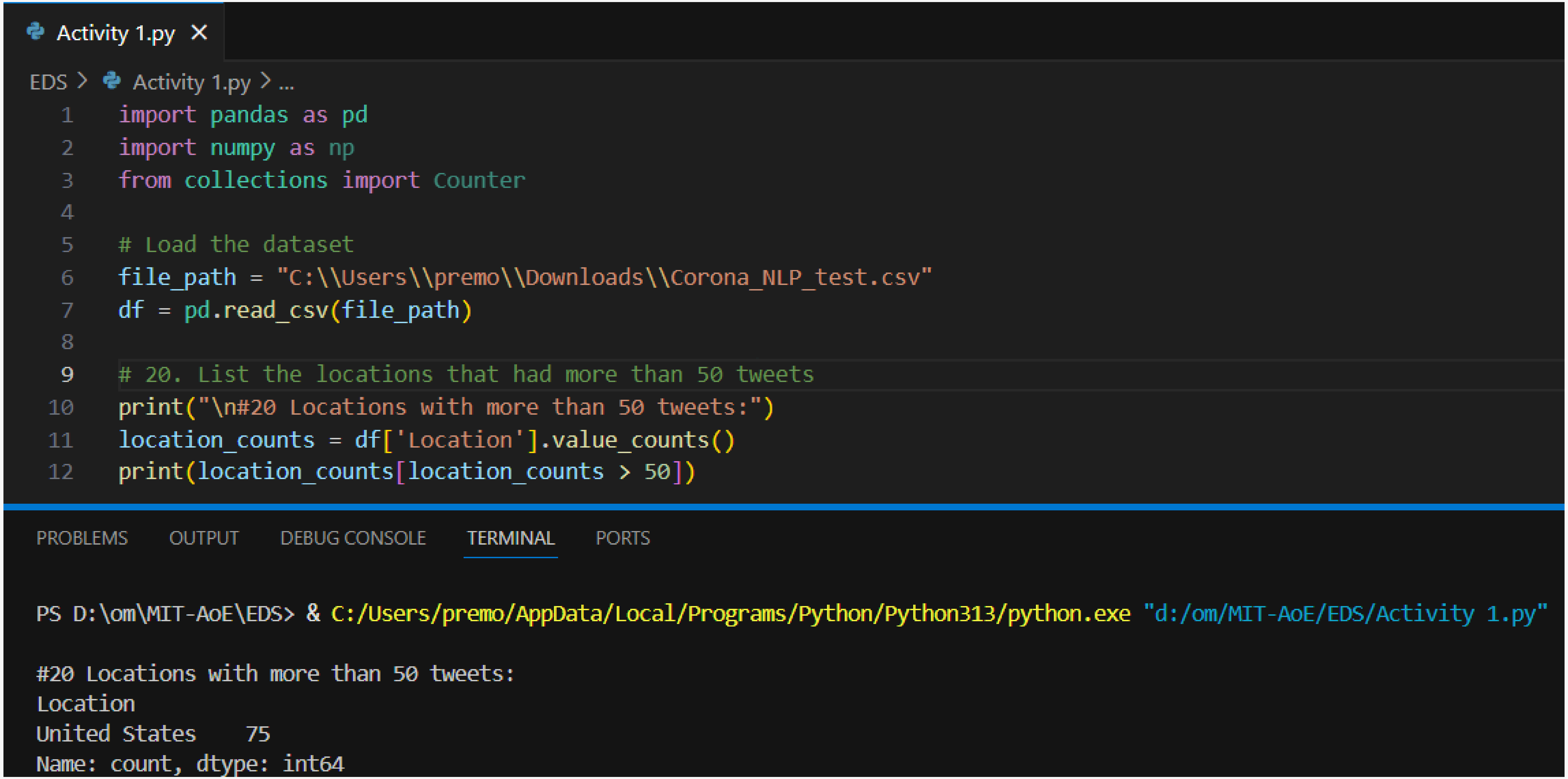
```python
import pandas as pd
import numpy as np
from collections import Counter

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 18. Find the top 5 words that occur most frequently across all tweets
print("\n#18 Top 5 most common words:")
all_words = ' '.join(df['OriginalTweet']).lower().split()
word_freq = Counter(all_words)
print(word_freq.most_common(5))
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#18 Top 5 most common words:
[('the', 4240), ('to', 3723), ('and', 2435), ('of', 2060), ('in', 1811)]
```



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

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 19. Find the average number of words per tweet
print("\n#19 Average number of words per tweet:")
print(df['OriginalTweet'].apply(lambda x: len(x.split())).mean())
print("\n")
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#19 Average number of words per tweet:
32.909689310163245
```



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

# Load the dataset
file_path = "C:\\Users\\premo\\Downloads\\Corona_NLP_test.csv"
df = pd.read_csv(file_path)

# 20. List the locations that had more than 50 tweets
print("\n#20 Locations with more than 50 tweets:")
location_counts = df['Location'].value_counts()
print(location_counts[location_counts > 50])
```

```
PS D:\om\MIT-AoE\EDS> & C:/Users/premo/AppData/Local/Programs/Python/Python313/python.exe "d:/om/MIT-AoE/EDS/Activity 1.py"

#20 Locations with more than 50 tweets:
Location
United States    75
Name: count, dtype: int64
```

THANK YOU