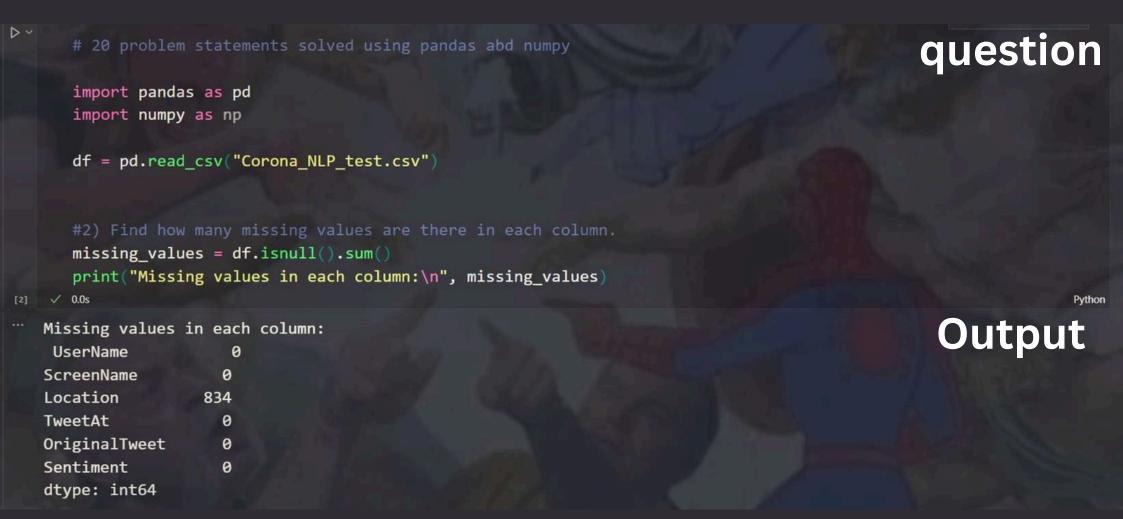
## Theory Activity No. 1

Formulate 20 problem statements for a given dataset using Numpy and Pandas and Apply Numpy and pandas methods to find the solution for the formulated problem statements.

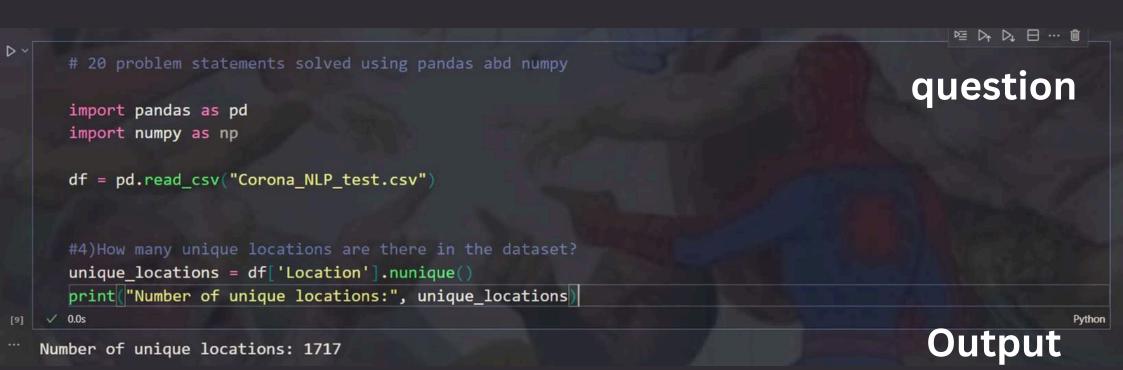
**Dataset: Kaggle Text Classification Dataset** Name: Mahir Mulani Roll No: ET1-54 PRN: 202401070130 Batch: ET13

**Dataset link** 





```
# 20 problem statements solved using pandas abd numpy
                                                                                            question
   import pandas as pd
   import numpy as np
   df = pd.read_csv("Corona_NLP_test.csv")
   #3) Fill missing Location with "Unknown"
   df['Location'] = df['Location'].fillna('Unknown')
   print(df['Location'])
✓ 0.0s
                                                                                                            Python
0
                       NYC
                                                                                             Output
                Seattle, WA
                    Unknown
                Chicagoland
       Melbourne, Victoria
                  Israel ??
3793
             Farmington, NM
3794
3795
             Haverford, PA
3796
                   Unknown
        Arlington, Virginia
3797
Name: Location, Length: 3798, dtype: object
```

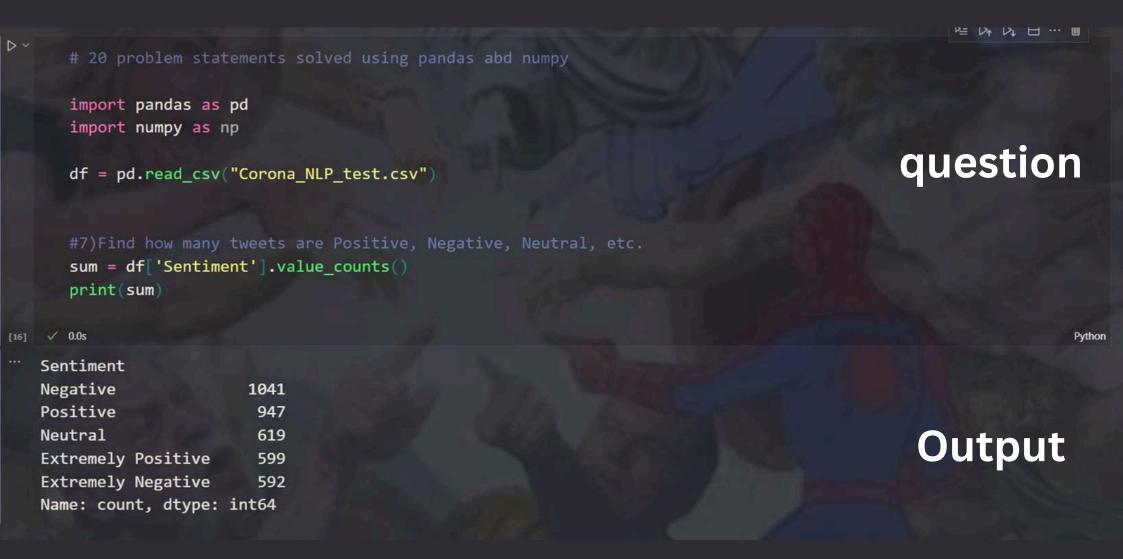


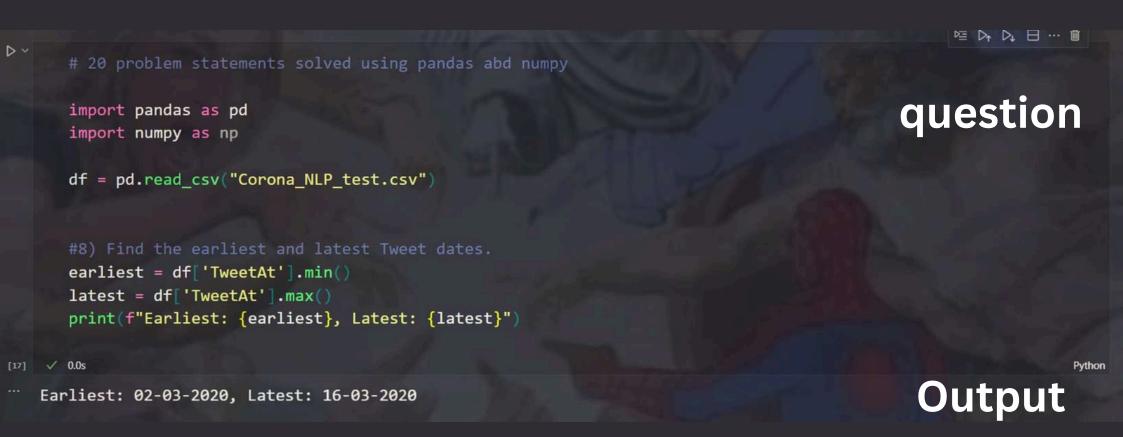


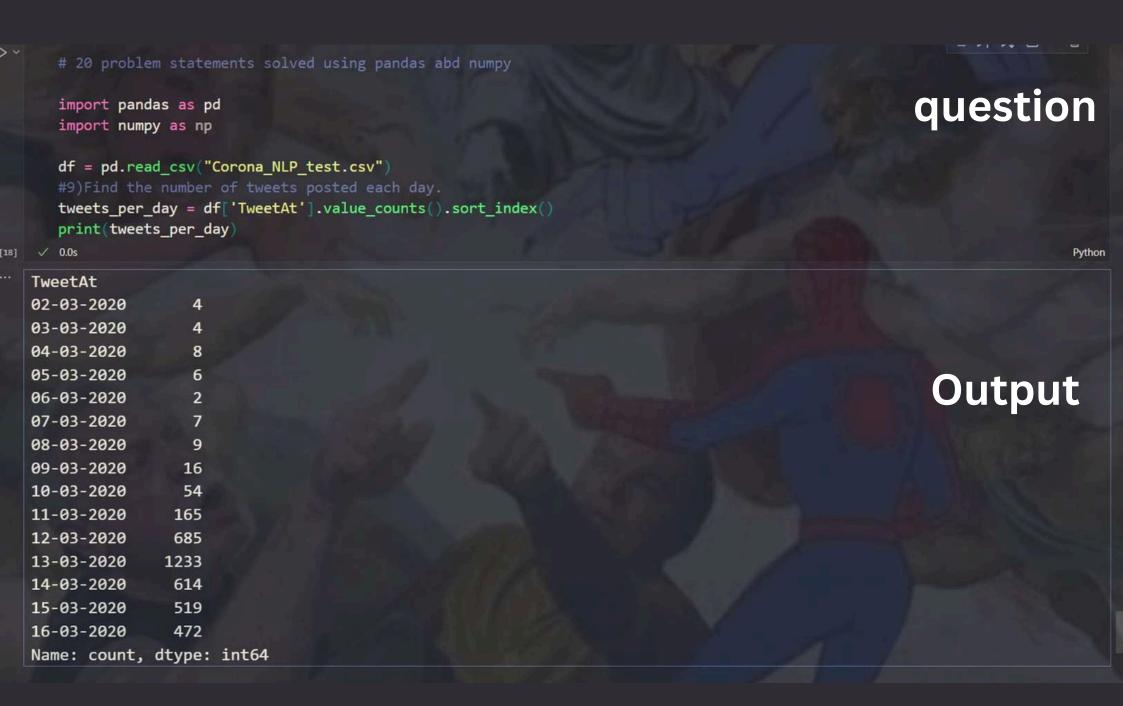
```
DV
      # 20 problem statements solved using pandas abd numpy
      import pandas as pd
      import numpy as np
                                                                                            question
      df = pd.read_csv("Corona_NLP_test.csv")
      #6) Convert TweetAt to datetime, extract month and year, and create a new column
      df['TweetAt'] = pd.to datetime(df['TweetAt'], dayfirst=True)
      print(df['TweetAt'])

√ 0.0s

                                                                                                             Python
          2020-03-02
          2020-03-02
          2020-03-02
                                                                                              Output
          2020-03-02
          2020-03-03
   3793
          2020-03-16
   3794
          2020-03-16
   3795
          2020-03-16
   3796
          2020-03-16
          2020-03-16
   3797
   Name: TweetAt, Length: 3798, dtype: datetime64[ns]
```









```
# 20 problem statements solved using pandas abd numpy

import pandas as pd
import numpy as np

df = pd.read_csv("Corona_NLP_test.csv")

#11) Find the tweet with the maximum number of characters.

max_length_tweet = df.loc[df['OriginalTweet'].str.len().idxmax(), 'OriginalTweet']

print(max_length_tweet)

#20 problem statements solved using pandas abd numpy

question

#20 problem statements solved using pandas abd numpy

question

#20 problem statements solved using pandas abd numpy

question

#20 problem statements solved using pandas abd numpy

question

#20 problem statements solved using pandas abd numpy

question

#21 print(max_length_tweet)

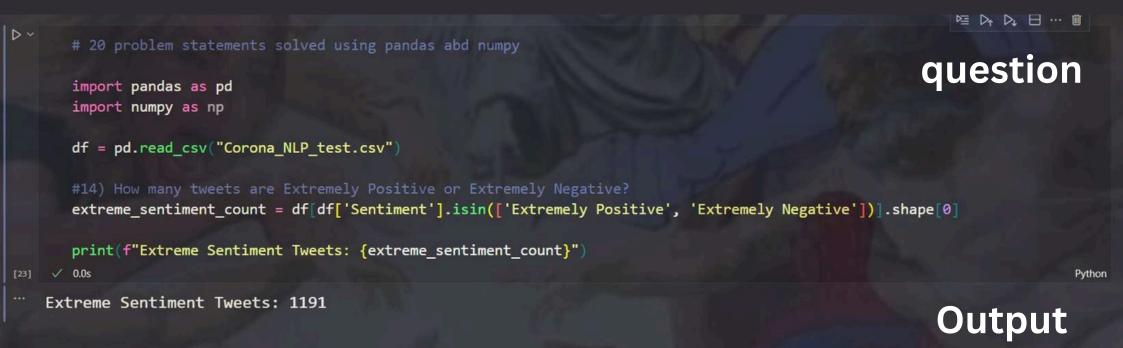
Python

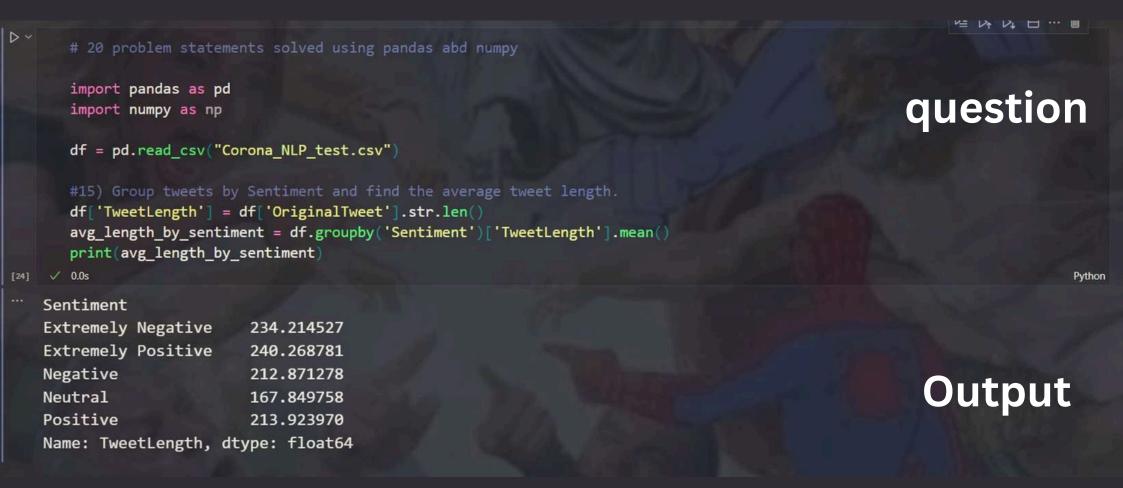
Why? She asked.?

Output
```



```
喧 际 及 日 … 前
   # 20 problem statements solved using pandas abd numpy
   import pandas as pd
                                                                                              question
   import numpy as np
  df = pd.read csv("Corona NLP test.csv")
   #13) List all tweets containing the word 'coronavirus'
   coronavirus tweets = df[df['OriginalTweet'].str.contains('coronavirus', case=False)]
  print(coronavirus_tweets[['OriginalTweet']])
✓ 0.0s
                                                                                                                 Python
                                           OriginalTweet
      TRENDING: New Yorkers encounter empty supermar...
0
      When I couldn't find hand sanitizer at Fred Me...
                                                                                                Output
      Find out how you can protect yourself and love...
      #Panic buying hits #NewYork City as anxious sh...
4
      #toiletpaper #dunnypaper #coronavirus #coronav...
      Stuck inside? How about getting some reading ...
      We've noticed a shift in consumer #research in...
3786
3787
      Its funny seeing all these people fight and pa...
      Meanwhile In A Supermarket in Israel -- People...
3793
      Asst Prof of Economics @cconces was on @NBCPhi...
3795
[1698 rows x 1 columns]
```





```
四下以出… 画
# 20 problem statements solved using pandas abd numpy
import pandas as pd
                                                                                          question
import numpy as np
df = pd.read csv("Corona NLP test.csv")
#16) Create a new column indicating if the tweet contains the word 'toiletpaper'.
df['ContainsToiletPaper'] = df['OriginalTweet'].str.contains('toiletpaper', case=False)
print(df[['OriginalTweet', 'ContainsToiletPaper']].head())
                                                                                                              Python
                                     OriginalTweet ContainsToiletPaper
                                                                   False
TRENDING: New Yorkers encounter empty supermar...
When I couldn't find hand sanitizer at Fred Me...
                                                                   False
Find out how you can protect yourself and love...
                                                                   False
                                                                                            Output
#Panic buying hits #NewYork City as anxious sh...
                                                                   False
#toiletpaper #dunnypaper #coronavirus #coronav...
                                                                   True
```



```
# 20 problem statements solved using pandas abd numpy

import pandas as pd
import numpy as np

df = pd.read_csv("Corona_NLP_test.csv")

#18) Find the 5 most frequent words in all tweets.

from collections import Counter
all_words = ' '.join(df['OriginalTweet']).lower().split()
common_words = Counter(all_words).most_common(5)
print(common_words)

Python

('the', 4240), ('to', 3723), ('and', 2435), ('of', 2060), ('in', 1811)]

Output
```



