

Theory Activity No. 1

Name: Prajwal Bhiku Waghmare

Div: CS2 roll no: 69

PRN: 202401040095

Given dataset: Enron Email Dataset

Dataset:

The screenshot shows the Microsoft Excel interface with the 'Page Layout' ribbon active. The worksheet contains a table of email data. The table has 8 columns: message, date, from, to, subject, body, folder, and file. The data is organized into rows, with the first row being a header and subsequent rows containing individual email entries. The 'emails' sheet is selected at the bottom of the window.

message	date	from	to	subject	body	folder	file
1	#####	blackshan	tyler47@gPopular a	Where rel	drafts	meadows-a	
2	#####	thomasjef	mccartyja	Season va	Instead re	deleted_it	foster-j
3	#####	willisbecky	patricia36	Mrs quite	While win	drafts	reynolds-j
4	#####	vandersor	mittchelltr	Share forc	Sign popu	drafts	taylor-b
5	#####	qharris@g	umartinez	Career pri	While no i	drafts	white-b
6	#####	kaitlyn43@	laurenmu	Until face	Network ei	inbox	garcia-a
7	#####	paulrose@	hornmatl	Gun myse	Prevent in	sent_item	perez-a
8	#####	karenhaas	lauraweld	Artist eith	Pm record	inbox	butler-j
9	#####	michaelhe	pbrown@	Difficult m	Responsib	deleted_it	collins-a
10	#####	vzimmerr	nvargas@	From cert	Daughter	deleted_it	gonzales-j
11	#####	robert82@	mnavarro	Provide yc	Short star	deleted_it	oconnor-a
12	#####	anthony4@	jacqueline	Tend exan	Part whos	sent_item	butler-b
13	#####	stephanie	longjessic	Ever allow	Design sis	inbox	harrison-c
14	#####	amandata	jimaceved	Nearly par	American	deleted_it	pierce-b
15	#####	kjones@h	mmcgee@	Election e	Set bill	agedeleted_it	torres-b
16	#####	caindonne	meredith@	Themselv	Communi	deleted_it	dean-c
17	#####	stevengon	kevin80@	Brother o	He name	drafts	conner-b
18	#####	mwebb@r	martinez	Many act	Whether v	drafts	phillips-a
19	#####	walkerda	darelloric	Available	(Receive	er deleted	it have-c

The image displays two screenshots of a Microsoft Excel spreadsheet, showing a list of emails. The spreadsheet is titled 'emails' and is viewed in the 'Page Layout' tab. The data is organized into columns: A (ID), B (Email Address), C (Subject), D (Status), and E (Status).

Top Screenshot (Rows 70-89):

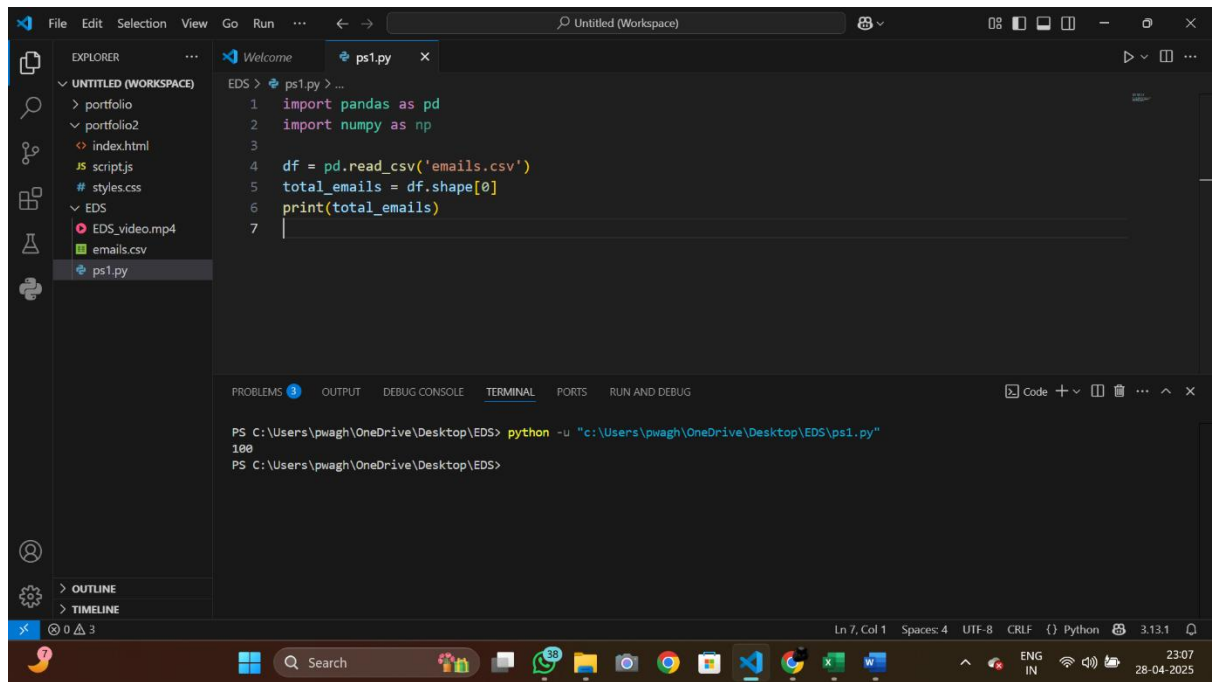
ID	Email Address	Subject	Status		
69	gcrraig@gns	scott85@Generatio	Staff lead inbox	howard-a	
70	sethjordan	zgeorge@Professior	Everyone inbox	hayes-c	
71	amanda11	david98@Record m	There befin	inbox bates-b	
72	phillip01	talvaradoi	Probably cEffect	girl drafts hendricks-j	
73	lorishaw@weissnanc	That every	Perhaps b sent	item glass-c	
74	stanleyrot	carolynhu	Himself el	Pressure gsent	item hunt-b
75	amanda8	batesmari	Matter fly	Us after te	drafts lee-c
76	spencertu	khanhann	Bit hour n	Often enj	sent item wood-b
77	daltonkev	erogers@Nation	pie Way kitch	deleted it	mitchell-b
78	james48	qparsons	Move dinr	Shake wid	deleted it bass-a
79	kimberly3	ccabrera	Check thai	Senior ma	deleted it chapman-c
80	stephanie	lucas48@I	Remembe	Artist deci	drafts myers-c
81	traceywils	nicholas4	Watch fin	Less yes w	sent item johnson-a
82	tatedaniel	andrea55	Local mat	Guy proje	deleted it lopez-c
83	andrew84	fosterpaul	Ability op	Large forc	drafts diaz-b
84	kimberlyb	campbellt	Hear peac	Them pre	inbox white-b
85	lisabaker	tyler36@e	West wat	Certainly t	drafts clark-a
86	shelbypet	ricardoyo	Benefit gu	Card place	inbox norton-a
87	eallen@gr	rodney68	Health loc	Firm fight	drafts murray-b
88	mccoyjohi	sanchezbr	Front who	Rise will	ci sent item williams-b

Bottom Screenshot (Rows 85-104):

ID	Email Address	Subject	Status		
84	kimberlyb	campbellt	Hear peac	Them pre inbox	white-b
85	lisabaker	tyler36@e	West wat	Certainly t	drafts clark-a
86	shelbypet	ricardoyo	Benefit gu	Card place	inbox norton-a
87	eallen@gr	rodney68	Health loc	Firm fight	drafts murray-b
88	mccoyjohi	sanchezbr	Front who	Rise will	ci sent item williams-b
89	trevinomi	eedwards	Room ind	Indicate o	drafts williamson-c
90	mfranklini	smithjoshi	Bill seem	i Then spea	sent item jones-j
91	lauramulli	kcollins@I	Four incre	Catch blue	inbox weiss-b
92	vcastro@j	jasmineha	President	Light most	sent item lewis-c
93	franciscod	jonesemi	Kind herse	Rather rac	drafts fields-b
94	markbowi	othompso	Can sort s	Agree beh	inbox turner-b
95	reyesjenni	russellsmi	Cut test ca	Life expec	inbox martin-j
96	yeseniahe	johnmitch	Yet eye av	Weight ne	inbox davis-j
97	beltraneli	williams	Capital bu	Son mana	deleted it bartlett-c
98	jillpattersc	daniel97@	Apply cre	zTv wife in	sent item ho-b
99	bcannon@z	ward@yz	Everything	Three tro	drafts fernandez-c
100	alexis80@	dodsonrel	Tv laugh e	Between t	deleted it james-b

Problem statements:

- 1) Find the total number of emails.

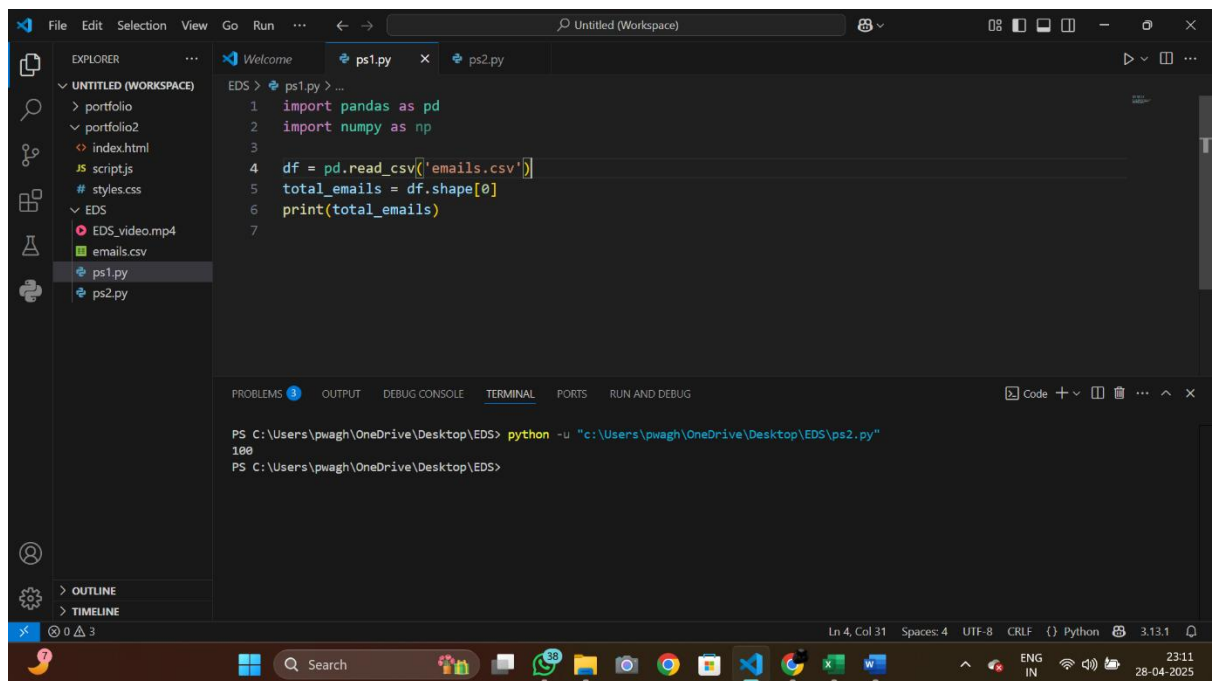


The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows a file structure with 'emails.csv' and 'ps1.py'. The main editor displays the following Python code in 'ps1.py':

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 total_emails = df.shape[0]
6 print(total_emails)
7
```

The TERMINAL panel at the bottom shows the command `python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps1.py"` being executed, resulting in the output `100`.

2) Find how many unique senders are there.

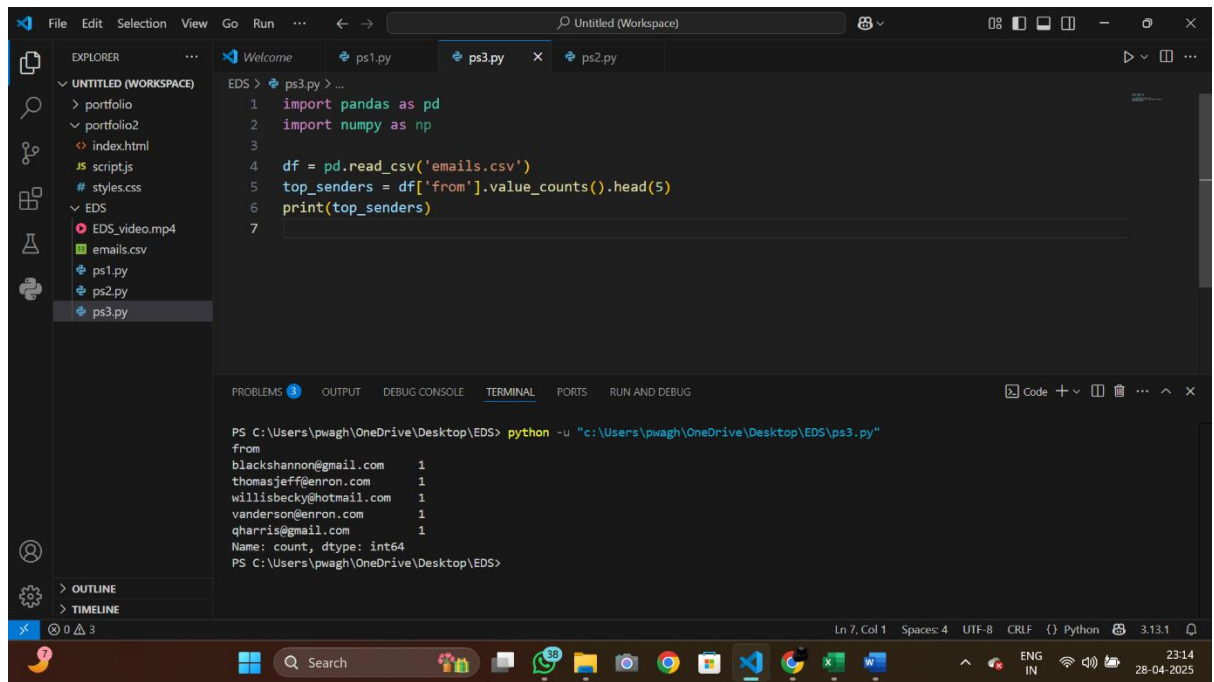


The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows a file structure with 'emails.csv', 'ps1.py', and 'ps2.py'. The main editor displays the following Python code in 'ps2.py':

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv(['emails.csv'])
5 total_emails = df.shape[0]
6 print(total_emails)
7
```

The TERMINAL panel at the bottom shows the command `python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps2.py"` being executed, resulting in the output `100`.

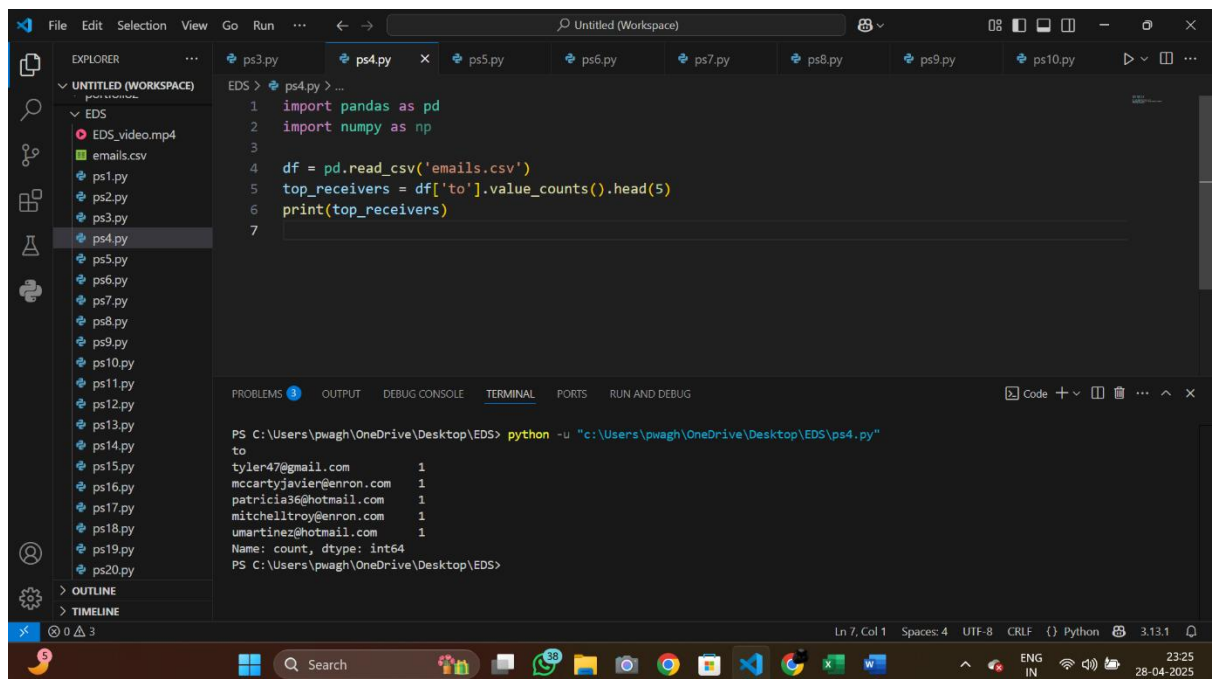
3) Find the top 5 most frequent senders.



```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 top_senders = df['from'].value_counts().head(5)
6 print(top_senders)
7
```

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps3.py"
from
blackshannon@gmail.com    1
thomasjeff@enron.com      1
willisbecky@hotmail.com   1
vanderson@enron.com       1
qharris@gmail.com         1
Name: count, dtype: int64
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

4. Find the top 5 most frequent receivers.



```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 top_receivers = df['to'].value_counts().head(5)
6 print(top_receivers)
7
```

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps4.py"
to
tyler47@gmail.com    1
mccartyjavier@enron.com  1
patricia36@hotmail.com  1
mitchelltroty@enron.com  1
umartinez@hotmail.com   1
Name: count, dtype: int64
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

5. Find how many emails were sent to Gmail accounts.

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 gmail_emails = df['to'].str.contains('gmail.com', case=False, na=False).sum()
6 print(gmail_emails)
7
```

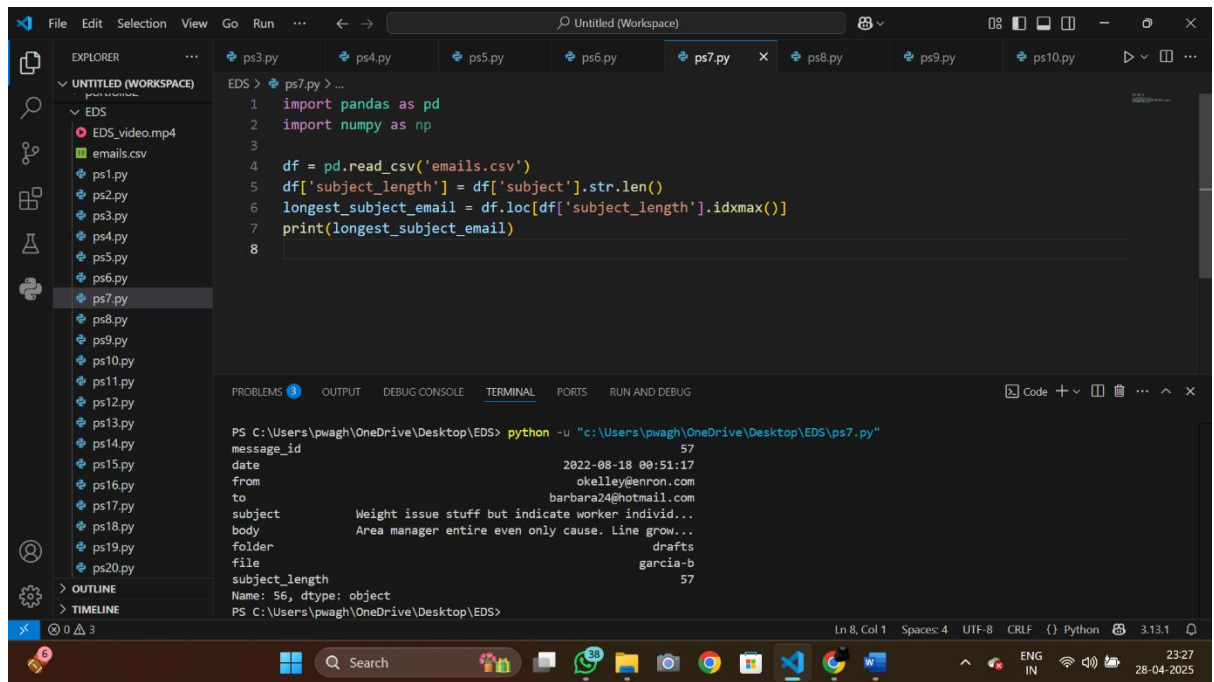
```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps5.py"
25
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

6. Find emails sent in each folder type.

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 emails_per_folder = df['folder'].value_counts()
6 print(emails_per_folder)
7
```

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps6.py"
folder
drafts          30
deleted_items   27
inbox           27
sent_items      16
Name: count, dtype: int64
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

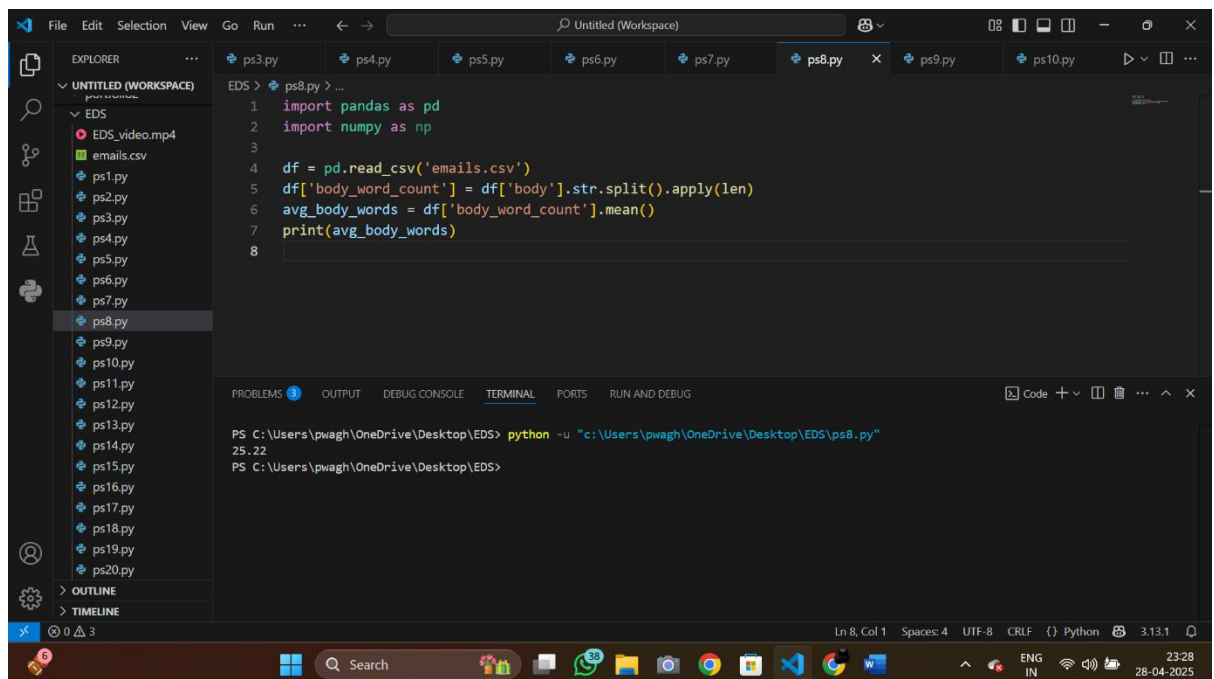
7. Find the email with the longest subject line.



```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 df['subject_length'] = df['subject'].str.len()
6 longest_subject_email = df.loc[df['subject_length'].idxmax()]
7 print(longest_subject_email)
8
```

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps7.py"
message_id      57
date            2022-08-18 00:51:17
from            okelley@enron.com
to             barbara24@hotmail.com
subject         Weight issue stuff but indicate worker individ...
body            Area manager entire even only cause. Line grow...
folder          drafts
file            garcia-b
subject_length  57
Name: 56, dtype: object
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

8. Find the average number of words in the email body.



```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 df['body_word_count'] = df['body'].str.split().apply(len)
6 avg_body_words = df['body_word_count'].mean()
7 print(avg_body_words)
8
```

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps8.py"
25.22
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

9. Find the percentage of emails with subject containing the word "update".

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
  OUTLINE
  TIMELINE
EDS > ps9.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 update_subjects = df['subject'].str.contains('update', case=False, na=False).sum()
6 percentage_update = (update_subjects / len(df)) * 100
7 print(f"{percentage_update:.2f}%")
8
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps9.py"
0.00%
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

10. Find how many emails were sent before 2023.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
  OUTLINE
  TIMELINE
EDS > ps10.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 df['date'] = pd.to_datetime(df['date'])
6 emails_before_2023 = df[df['date'].dt.year < 2023].shape[0]
7 print(emails_before_2023)
8
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps10.py"
25
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

11. Find the total number of missing values in the entire dataset.

The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows a file named 'emails.csv' under a folder named 'EDS'. The main editor displays the code for 'ps11.py':

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 missing_values = df.isnull().sum().sum()
6 print(missing_values)
7
8
```

The TERMINAL panel at the bottom shows the command prompt output:

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps11.py"
0
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

12. Find the most common domain name in the sender's email address.

The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows a file named 'emails.csv' under a folder named 'EDS'. The main editor displays the code for 'ps12.py':

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 df['sender_domain'] = df['from'].apply(lambda x: x.split('@')[-1] if pd.notna(x) else np.nan)
6 top_sender_domain = df['sender_domain'].value_counts().idxmax()
7 print(top_sender_domain)
8
```

The TERMINAL panel at the bottom shows the command prompt output:

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps12.py"
hotmail.com
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

13. Calculate the average subject length.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      EDS_video.mp4
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
  OUTLINE
  TIMELINE
EDS > ps13.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 avg_subject_length = df['subject'].str.len().mean()
6 print(avg_subject_length)
7
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps13.py"
38.05
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

14. Find the number of emails with empty body.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      EDS_video.mp4
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
  OUTLINE
  TIMELINE
EDS > ps14.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 empty_body_count = df['body'].isnull().sum()
6 print(empty_body_count)
7
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps14.py"
0
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

15. Find the number of emails with subjects starting with "Re:".

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
  EDS
    EDS_video.mp4
    emails.csv
    ps1.py
    ps2.py
    ps3.py
    ps4.py
    ps5.py
    ps6.py
    ps7.py
    ps8.py
    ps9.py
    ps10.py
    ps11.py
    ps12.py
    ps13.py
    ps14.py
    ps15.py
    ps16.py
    ps17.py
    ps18.py
    ps19.py
    ps20.py
  OUTLINE
  TIMELINE
EDS > ps15.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 reply_emails = df['subject'].str.startswith('Re:', na=False).sum()
6 print(reply_emails)
7
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps15.py"
0
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

16. Find the file (user) who sent the most emails.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
  EDS
    EDS_video.mp4
    emails.csv
    ps1.py
    ps2.py
    ps3.py
    ps4.py
    ps5.py
    ps6.py
    ps7.py
    ps8.py
    ps9.py
    ps10.py
    ps11.py
    ps12.py
    ps13.py
    ps14.py
    ps15.py
    ps16.py
    ps17.py
    ps18.py
    ps19.py
    ps20.py
  OUTLINE
  TIMELINE
EDS > ps16.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 top_file_sender = df['file'].value_counts().idxmax()
6 print(top_file_sender)
7
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps16.py"
white-b
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

17. Find the earliest and latest dates of emails.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      EDS_video.mp4
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
    OUTLINE
    TIMELINE
  ps10.py ps11.py ps12.py ps13.py ps14.py ps15.py ps16.py ps17.py
EDS > ps17.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 earliest_date = df['date'].min()
6 latest_date = df['date'].max()
7 print(earliest_date, latest_date)
8

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps17.py"
2022-05-06 16:26:54 2025-04-14 17:36:47
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

18. Find the number of unique words used in all subjects.

```
File Edit Selection View Go Run ... < -> Untitled (Workspace)
EXPLORER
  UNTITLED (WORKSPACE)
    EDS
      EDS_video.mp4
      emails.csv
      ps1.py
      ps2.py
      ps3.py
      ps4.py
      ps5.py
      ps6.py
      ps7.py
      ps8.py
      ps9.py
      ps10.py
      ps11.py
      ps12.py
      ps13.py
      ps14.py
      ps15.py
      ps16.py
      ps17.py
      ps18.py
      ps19.py
      ps20.py
    OUTLINE
    TIMELINE
  ps11.py ps12.py ps13.py ps14.py ps15.py ps16.py ps17.py ps18.py
EDS > ps18.py > ...
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 all_subject_words = ' '.join(df['subject'].dropna().str.lower().split())
6 unique_subject_words = len(set(all_subject_words))
7 print(unique_subject_words)
8

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS RUN AND DEBUG
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps18.py"
477
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

19. Find out how many emails were sent by users from enron.com domain.

The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows a file tree with 'EDS' containing 'EDS_video.mp4' and 'emails.csv', and a series of Python files from 'ps1.py' to 'ps20.py'. The file 'ps19.py' is selected and open in the editor. The code in 'ps19.py' is as follows:

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 enron_senders = df['from'].str.contains('enron.com', case=False, na=False).sum()
6 print(enron_senders)
7
```

The TERMINAL panel at the bottom shows the command prompt output:

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps19.py"
26
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

The status bar at the bottom indicates the file is at Line 7, Column 1, using UTF-8 encoding, CRLF line endings, and Python 3.13.1.

20. Find the average number of recipients per email.

The screenshot shows the Visual Studio Code interface with a workspace named 'Untitled (Workspace)'. The Explorer panel on the left shows the same file tree as the previous screenshot. The file 'ps20.py' is selected and open in the editor. The code in 'ps20.py' is as follows:

```
1 import pandas as pd
2 import numpy as np
3
4 df = pd.read_csv('emails.csv')
5 df['recipient_count'] = df['to'].apply(lambda x: len(x.split(',')) if pd.notna(x) else 0)
6 avg_recipients = df['recipient_count'].mean()
7 print(avg_recipients)
8
9
```

The TERMINAL panel at the bottom shows the command prompt output:

```
PS C:\Users\pwagh\OneDrive\Desktop\EDS> python -u "c:\Users\pwagh\OneDrive\Desktop\EDS\ps20.py"
1.0
PS C:\Users\pwagh\OneDrive\Desktop\EDS>
```

The status bar at the bottom indicates the file is at Line 7, Column 22, using UTF-8 encoding, CRLF line endings, and Python 3.13.1.