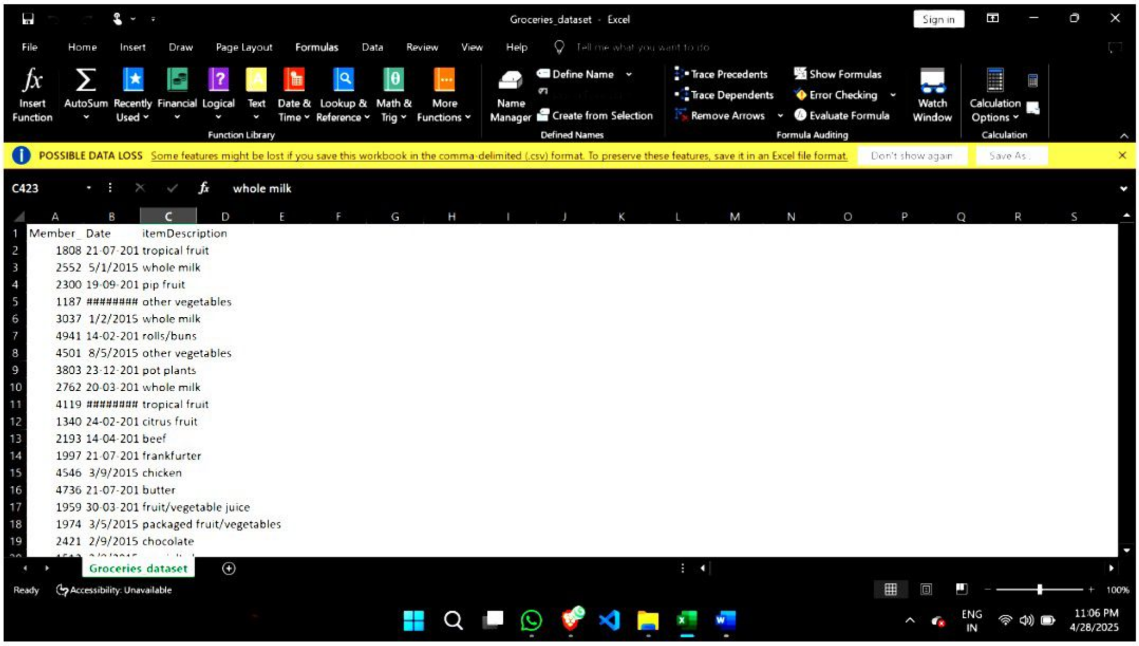
EDS theory assignment no.1

Name:- Aniket Nilesh Rajput. Div:- CS3 PRN:- 202401040138

Roll no:- 23 Google Colab Notebook Link:- https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?usp=sharing

Kaggle dataset link:- https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

DATASET: Groceries



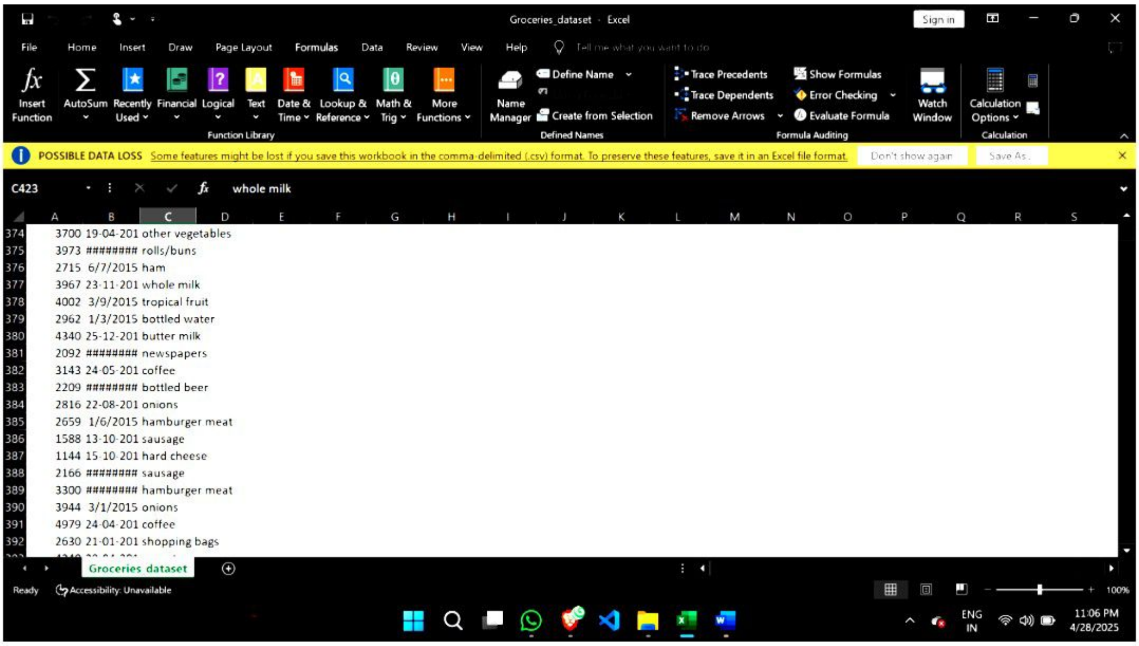
Groceries_dataset - Excel

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

C423 whole milk

	A	B	C
1	Member_	Date	itemDescription
2	1808	21-07-201	tropical fruit
3	2552	5/1/2015	whole milk
4	2300	19-09-201	pip fruit
5	1187	########	other vegetables
6	3037	1/2/2015	whole milk
7	4941	14-02-201	rolls/buns
8	4501	8/5/2015	other vegetables
9	3803	23-12-201	pot plants
10	2762	20-03-201	whole milk
11	4119	########	tropical fruit
12	1340	24-02-201	citrus fruit
13	2193	14-04-201	beef
14	1997	21-07-201	frankfurter
15	4546	3/9/2015	chicken
16	4736	21-07-201	butter
17	1959	30-03-201	fruit/vegetable juice
18	1974	3/5/2015	packaged fruit/vegetables
19	2421	2/9/2015	chocolate

Groceries dataset



Groceries_dataset - Excel

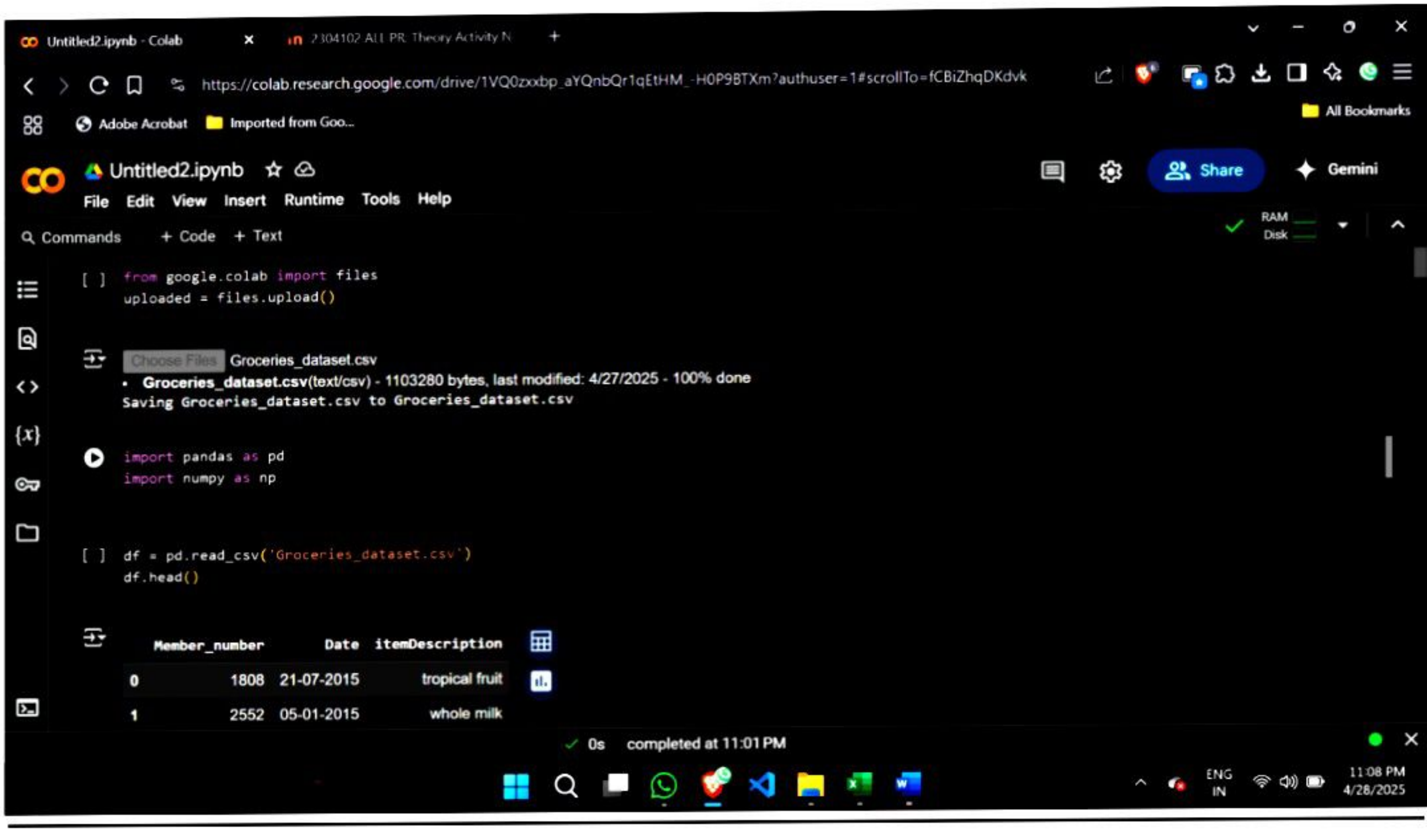
POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

C423 whole milk

	A	B	C
374	3700	19-04-201	other vegetables
375	3973	########	rolls/buns
376	2715	6/7/2015	ham
377	3967	23-11-201	whole milk
378	4002	3/9/2015	tropical fruit
379	2962	1/3/2015	bottled water
380	4340	25-12-201	butter milk
381	2092	########	newspapers
382	3143	24-05-201	coffee
383	2209	########	bottled beer
384	2816	22-08-201	onions
385	2659	1/6/2015	hamburger meat
386	1588	13-10-201	sausage
387	1144	15-10-201	hard cheese
388	2166	########	sausage
389	3300	########	hamburger meat
390	3944	3/1/2015	onions
391	4979	24-04-201	coffee
392	2630	21-01-201	shopping bags

Groceries dataset

Screenshots of assignment done on the google colab:


Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
Choose Files

```
from google.colab import files
uploaded = files.upload()
```

Groceries_dataset.csv

- **Groceries_dataset.csv**(text/csv) - 1103280 bytes, last modified: 4/27/2025 - 100% done

```
Saving Groceries_dataset.csv to Groceries_dataset.csv
```

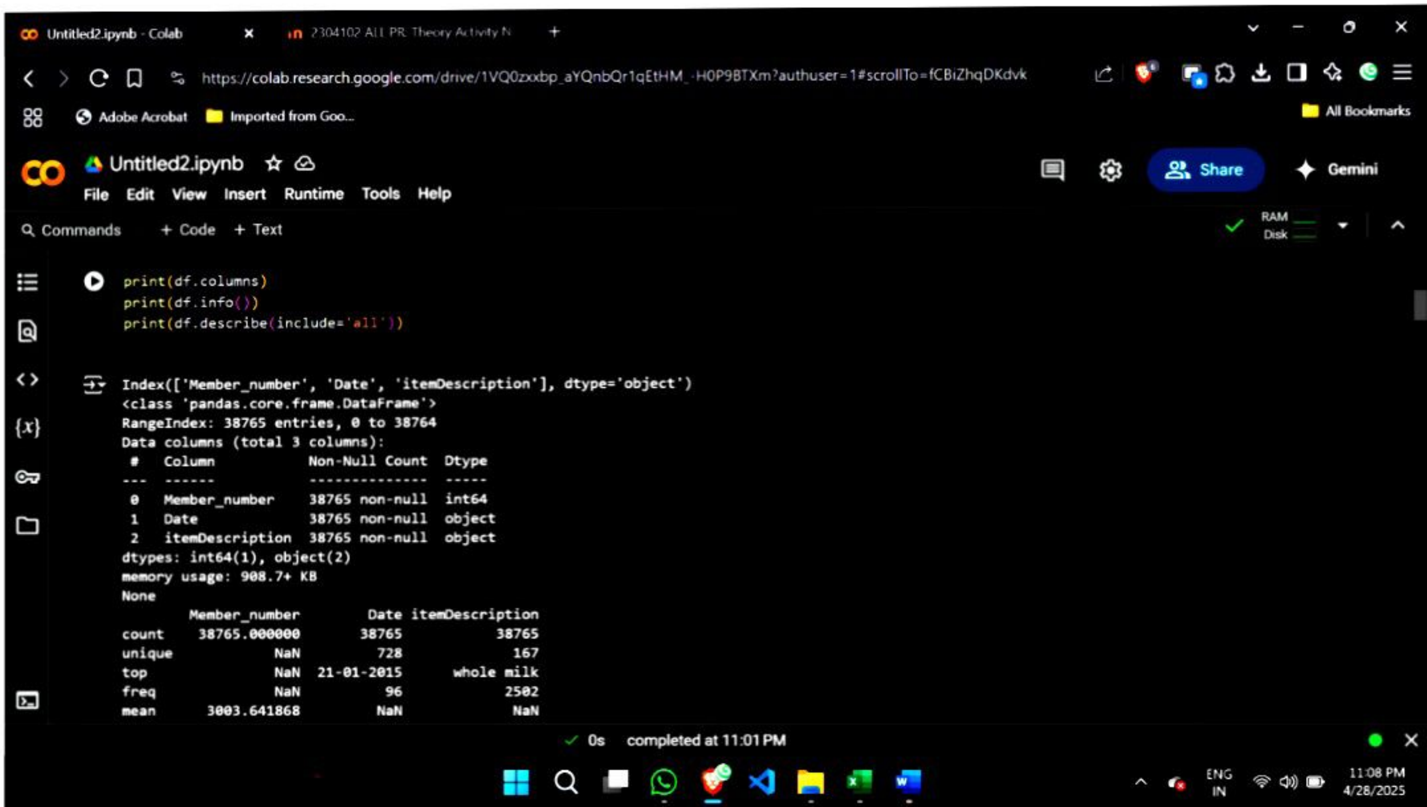
```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('Groceries_dataset.csv')
df.head()
```

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk

✓ 0s completed at 11:01 PM

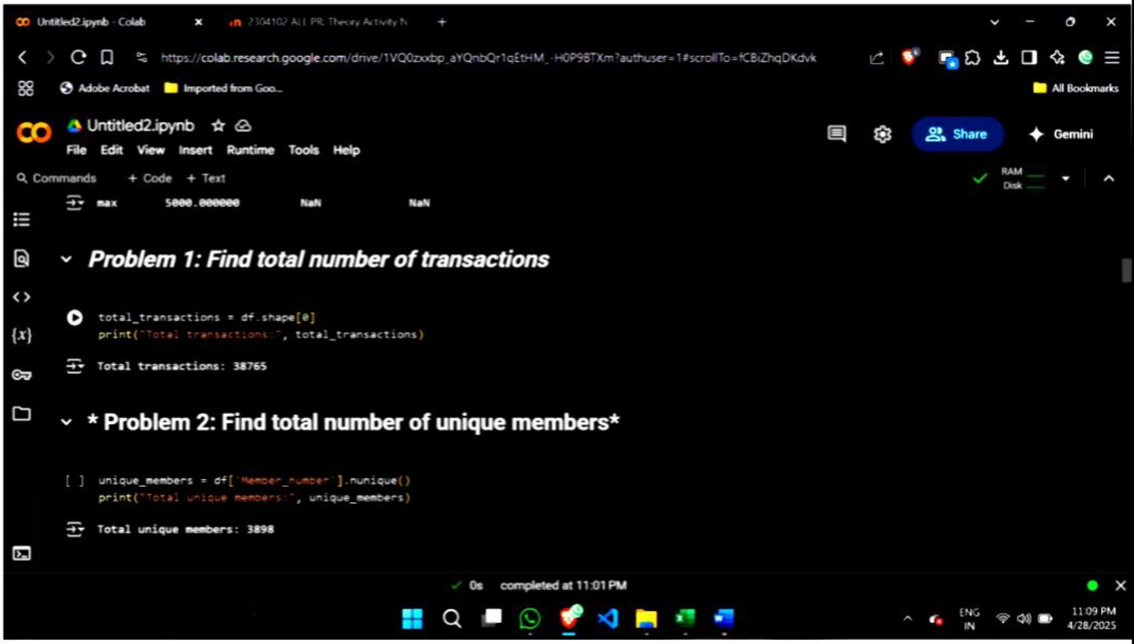
ENG
IN
11:08 PM
4/28/2025


Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk

```
print(df.columns)
print(df.info())
print(df.describe(include='all'))
```

```
Index(['Member_number', 'Date', 'itemDescription'], dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38765 entries, 0 to 38764
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Member_number    38765 non-null  int64 
 1   Date             38765 non-null  object
 2   itemDescription  38765 non-null  object
dtypes: int64(1), object(2)
memory usage: 908.7+ KB
None
       Member_number        Date itemDescription
count   38765.000000       38765           38765
unique           NaN         728             167
top              NaN  21-01-2015      whole milk
freq             NaN          96            2502
mean     3003.641868         NaN             NaN
```

✓ 0s completed at 11:01 PM


Untitled2.ipynb - Colab
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text
Share
Gemini
✓ 0s completed at 11:01 PM
11:09 PM
4/28/2025

```
max       5000.000000          NaN              NaN
```

Problem 1: Find total number of transactions

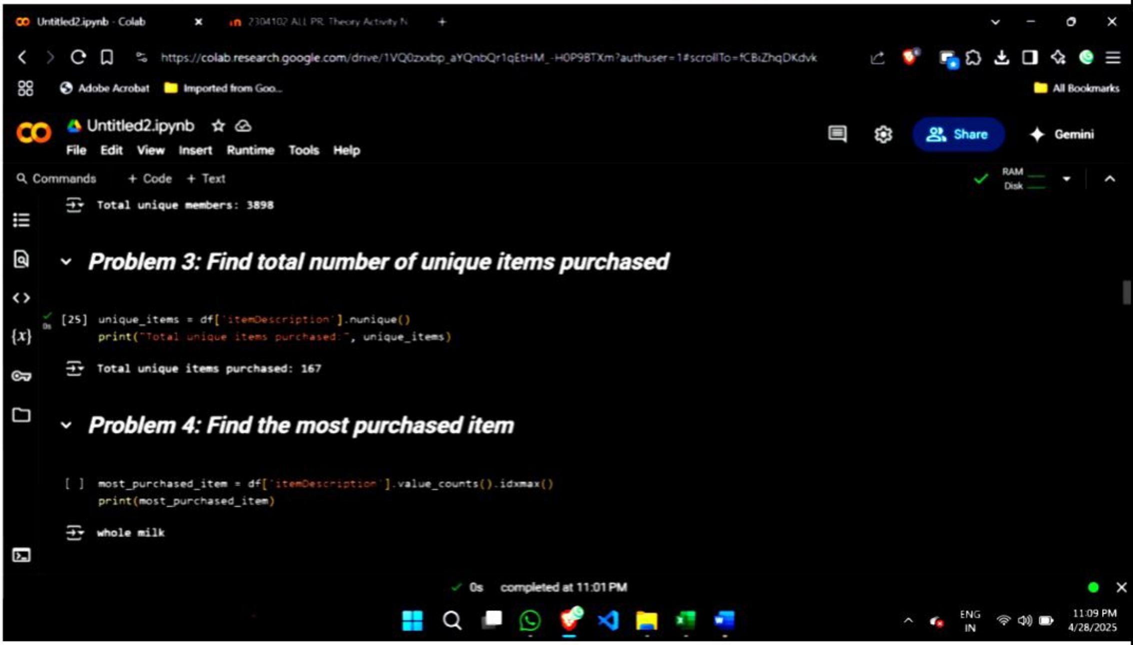
```
total_transactions = df.shape[0]
print("Total transactions:", total_transactions)
```

```
Total transactions: 38765
```

* Problem 2: Find total number of unique members*

```
unique_members = df['Member_number'].nunique()
print("Total unique members:", unique_members)
```

```
Total unique members: 3898
```


Untitled2.ipynb - Colab
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text
Share
Gemini
✓ 0s completed at 11:01 PM
11:09 PM
4/28/2025

```
Total unique members: 3898
```

Problem 3: Find total number of unique items purchased

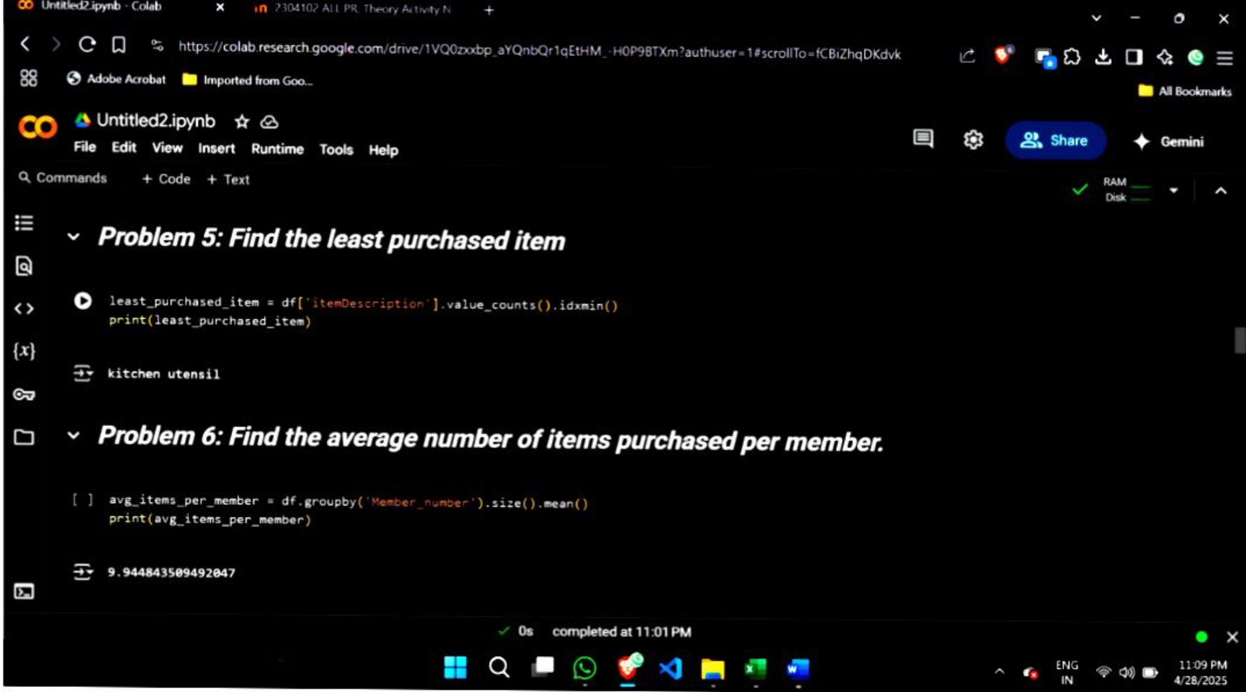
```
[25] unique_items = df['itemDescription'].nunique()
     print("Total unique items purchased:", unique_items)
```

```
Total unique items purchased: 167
```

Problem 4: Find the most purchased item

```
most_purchased_item = df['itemDescription'].value_counts().idxmax()
print(most_purchased_item)
```

```
whole milk
```



Problem 5: Find the least purchased item

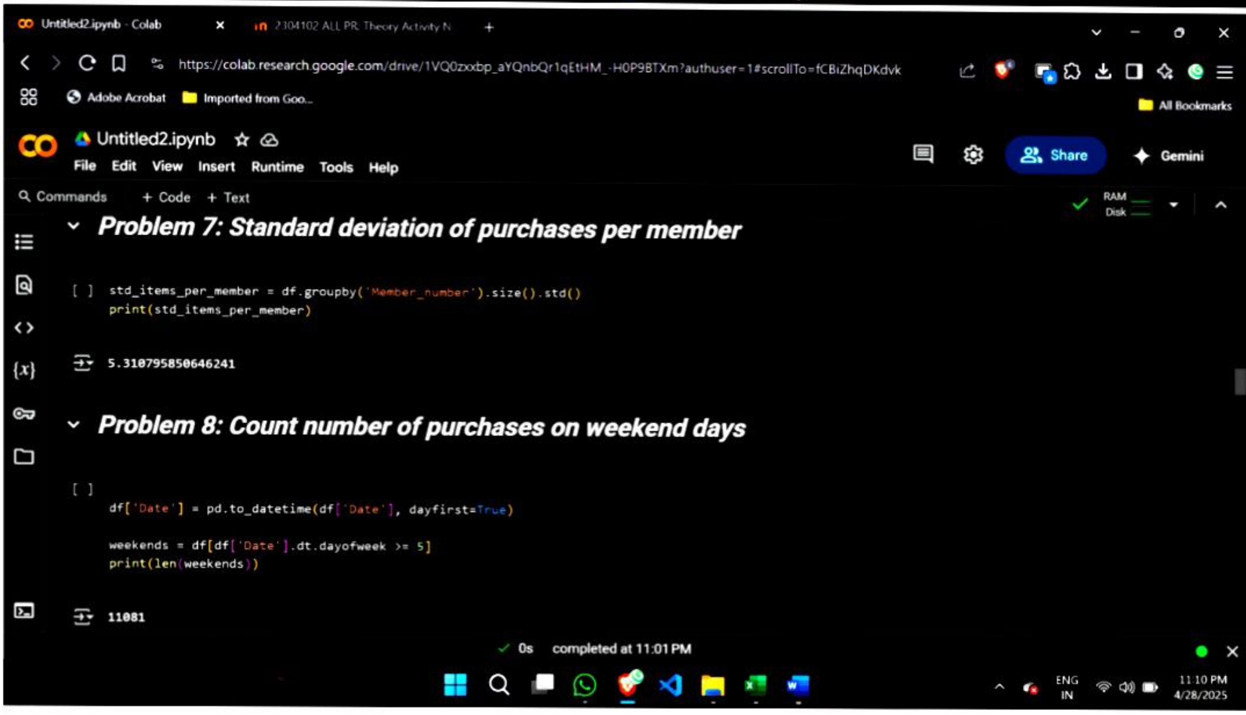
```
least_purchased_item = df['itemDescription'].value_counts().idxmin()
print(least_purchased_item)
```

```
kitchen utensil
```

Problem 6: Find the average number of items purchased per member.

```
avg_items_per_member = df.groupby('Member_number').size().mean()
print(avg_items_per_member)
```

```
9.944843509492047
```



Problem 7: Standard deviation of purchases per member

```
std_items_per_member = df.groupby('Member_number').size().std()
print(std_items_per_member)
```

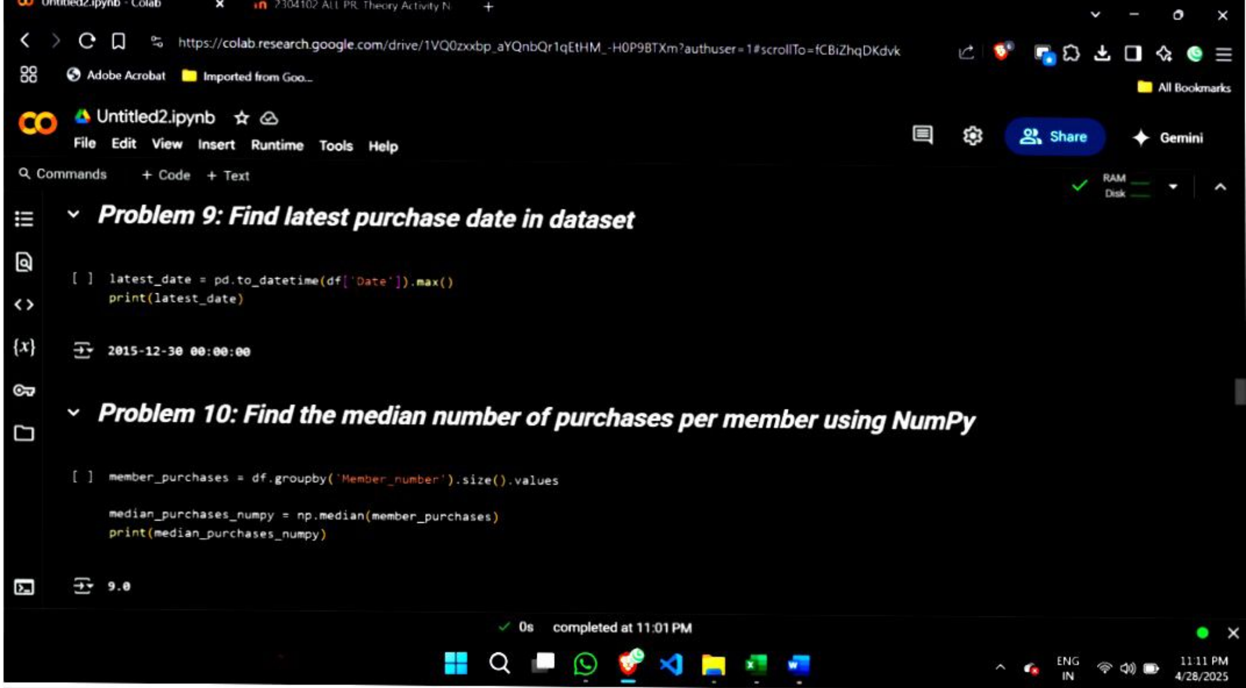
```
5.310795850646241
```

Problem 8: Count number of purchases on weekend days

```
df['Date'] = pd.to_datetime(df['Date'], dayfirst=True)

weekends = df[df['Date'].dt.dayofweek >= 5]
print(len(weekends))
```

```
11081
```


Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
0s completed at 11:01 PM
ENG
IN
11:11 PM
4/28/2025

Problem 9: Find latest purchase date in dataset

```
latest_date = pd.to_datetime(df['Date']).max()
print(latest_date)
```

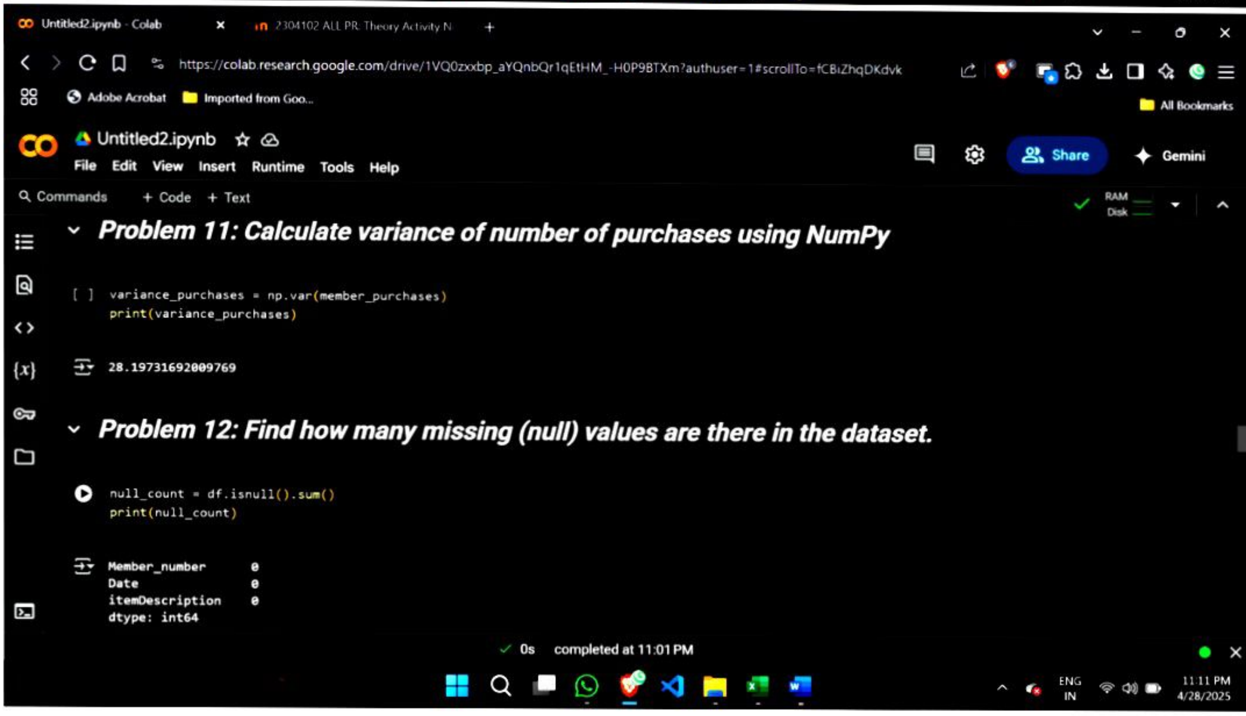
```
2015-12-30 00:00:00
```

Problem 10: Find the median number of purchases per member using NumPy

```
member_purchases = df.groupby('Member_number').size().values

median_purchases_numpy = np.median(member_purchases)
print(median_purchases_numpy)
```

```
9.0
```


Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
0s completed at 11:01 PM
ENG
IN
11:11 PM
4/28/2025

Problem 11: Calculate variance of number of purchases using NumPy

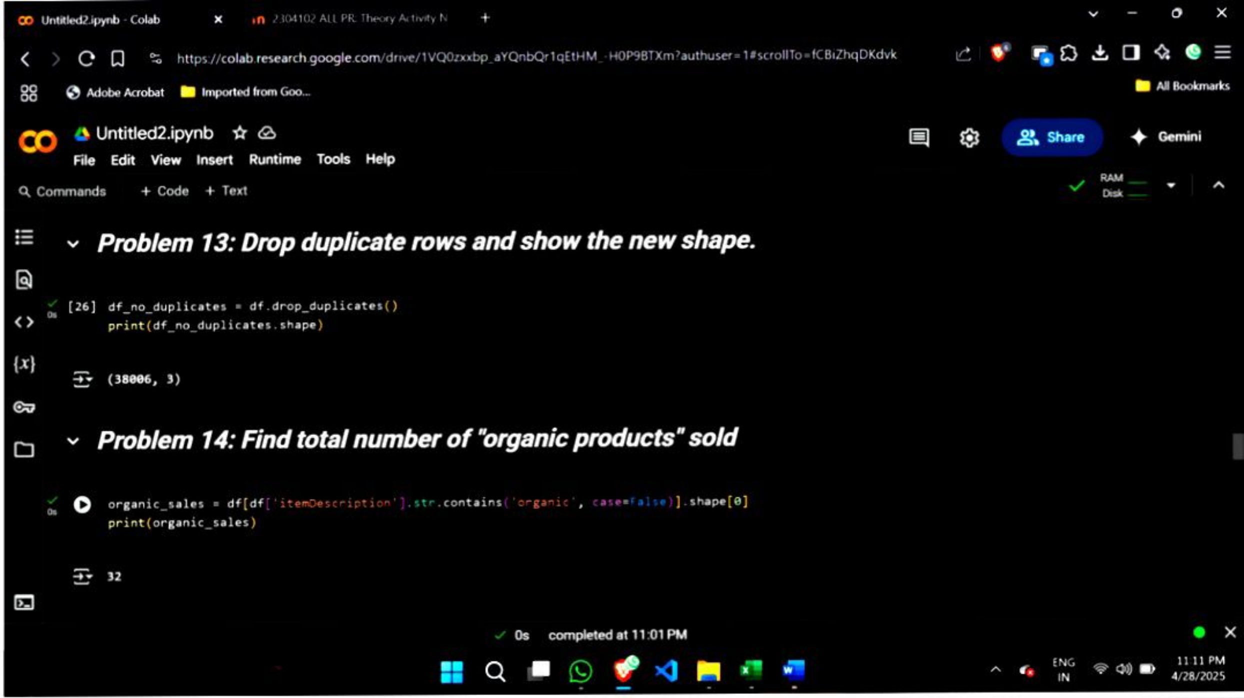
```
variance_purchases = np.var(member_purchases)
print(variance_purchases)
```

```
28.19731692009769
```

Problem 12: Find how many missing (null) values are there in the dataset.

```
null_count = df.isnull().sum()
print(null_count)
```

```
Member_number      0
Date               0
itemDescription    0
dtype: int64
```



Problem 13: Drop duplicate rows and show the new shape.

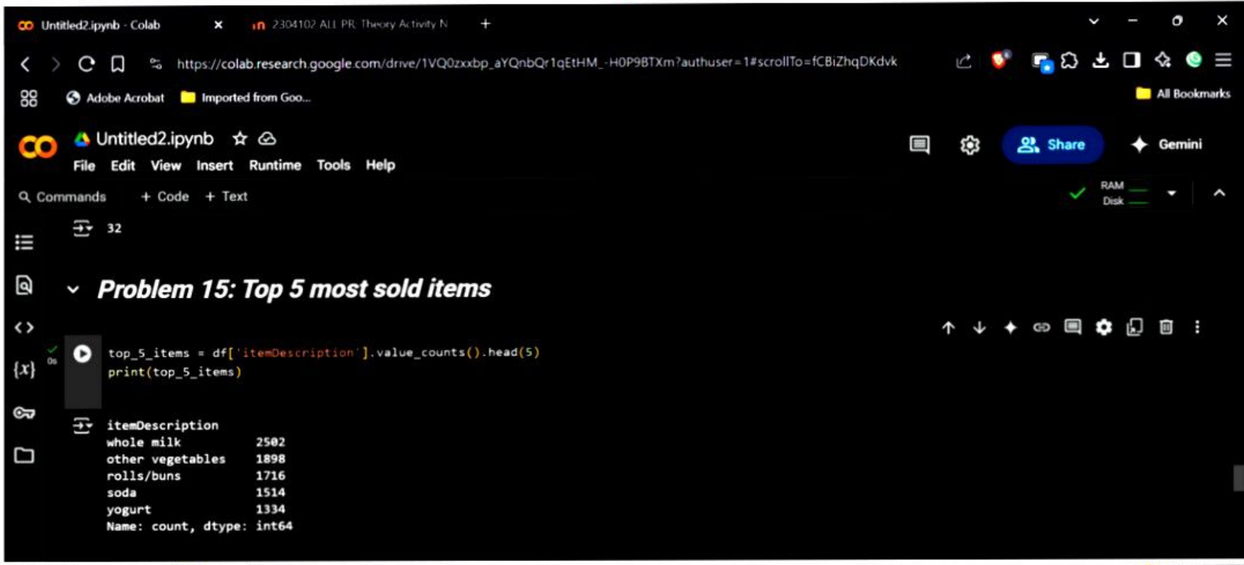
```
[26] df_no_duplicates = df.drop_duplicates()
     print(df_no_duplicates.shape)
```

```
(38006, 3)
```

Problem 14: Find total number of "organic products" sold

```
organic_sales = df[df['itemDescription'].str.contains('organic', case=False)].shape[0]
print(organic_sales)
```

```
32
```

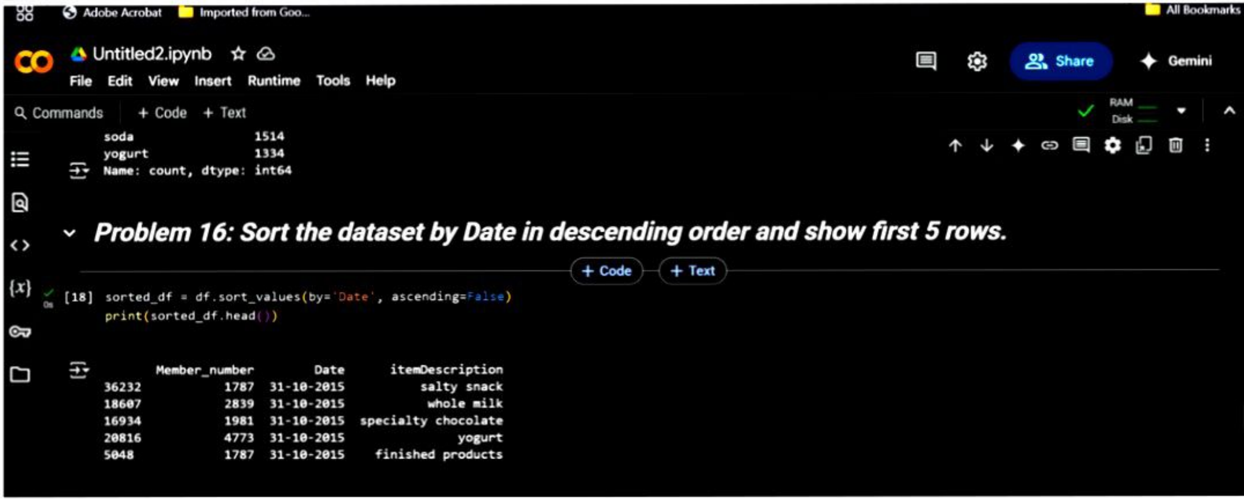


```
32
```

Problem 15: Top 5 most sold items

```
top_5_items = df['itemDescription'].value_counts().head(5)
print(top_5_items)
```

```
itemDescription
whole milk          2502
other vegetables    1898
rolls/buns          1716
soda                1514
yogurt              1334
Name: count, dtype: int64
```

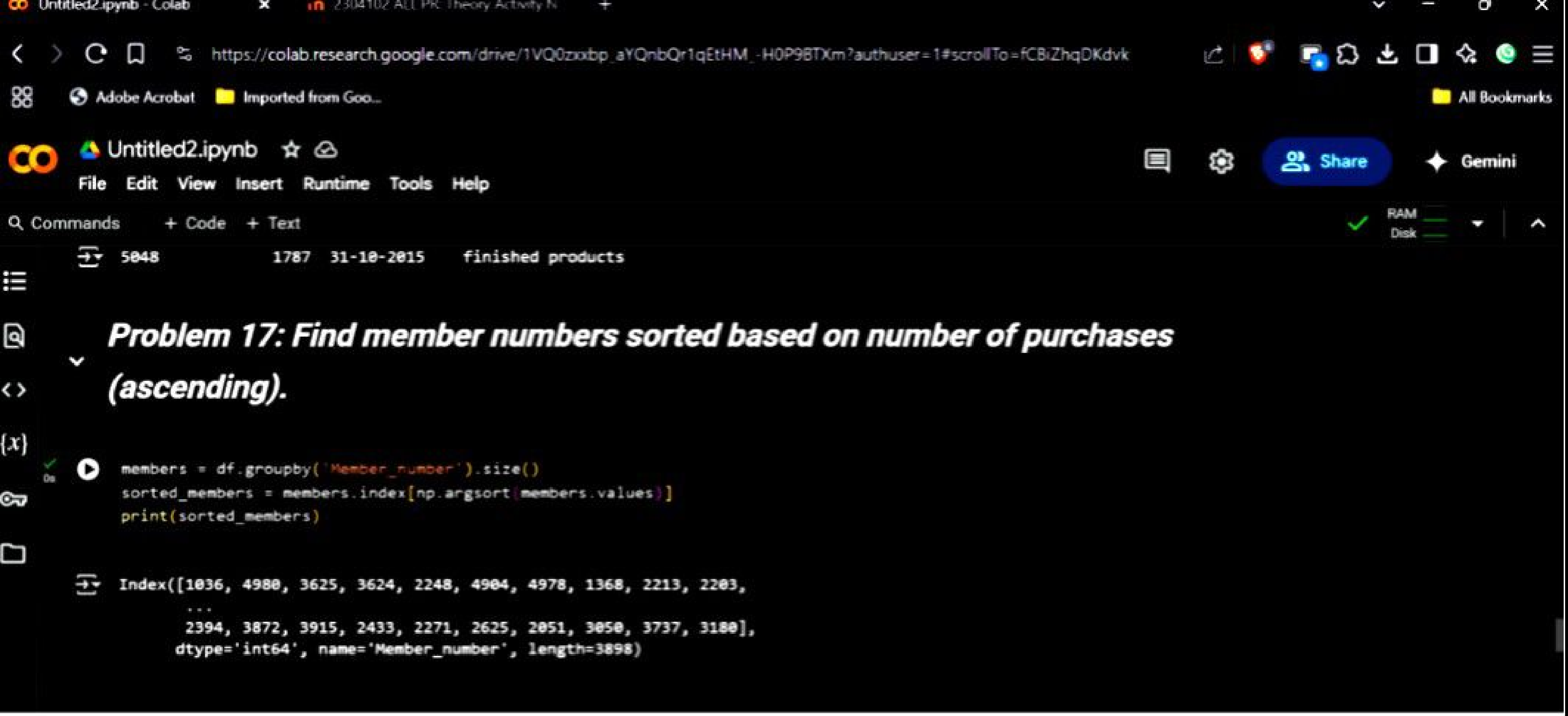


```
soda                1514
yogurt              1334
Name: count, dtype: int64
```

Problem 16: Sort the dataset by Date in descending order and show first 5 rows.

```
[18] sorted_df = df.sort_values(by='Date', ascending=False)
     print(sorted_df.head())
```

```
       Member_number        Date       itemDescription
36232           1787  31-10-2015           salty snack
18607           2839  31-10-2015            whole milk
16934           1981  31-10-2015  specialty chocolate
20816           4773  31-10-2015                yogurt
5048            1787  31-10-2015    finished products
```

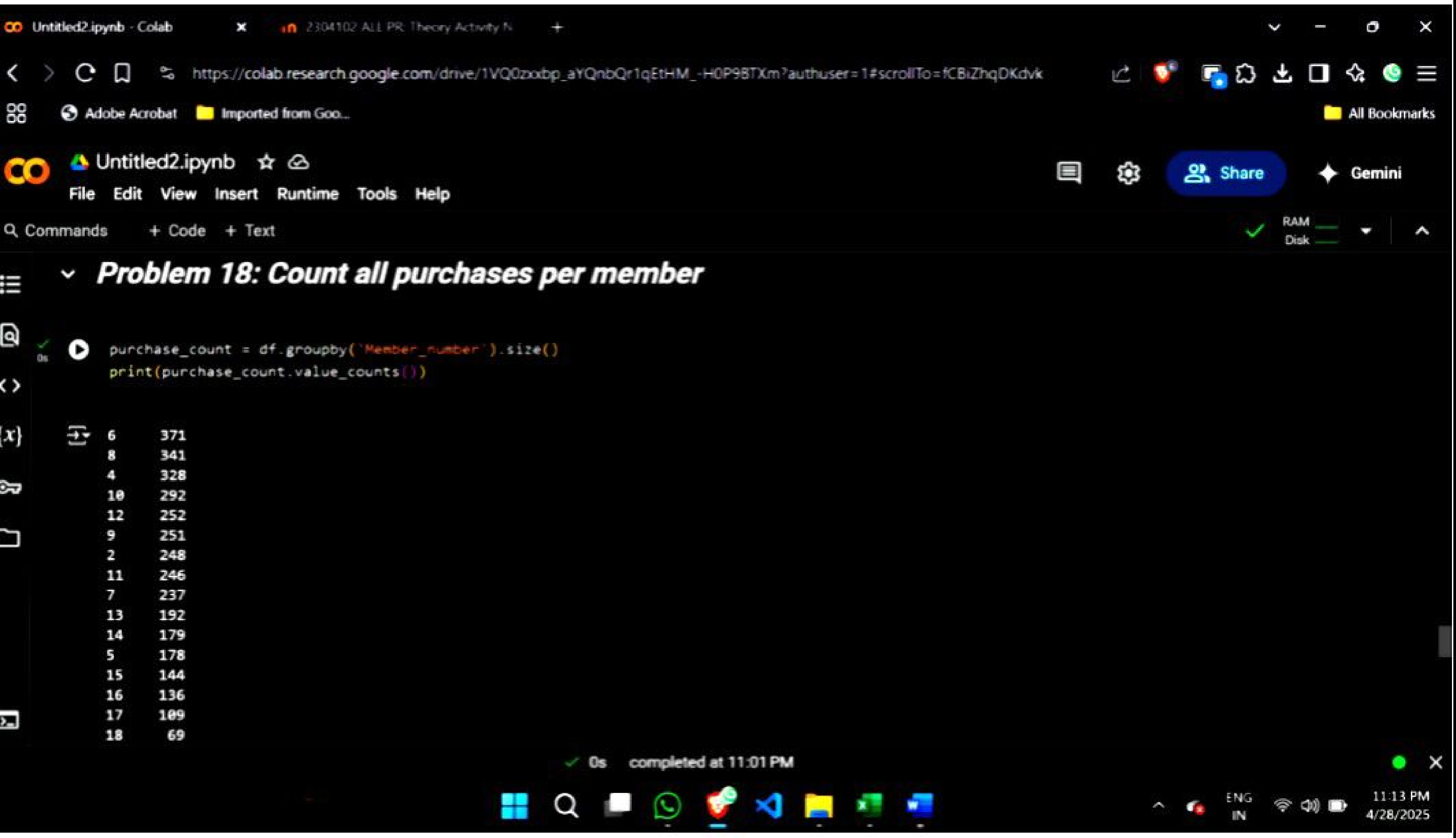

Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text
Share
Gemini
RAM
Disk

```
5048          1787  31-10-2015     finished products
```

Problem 17: Find member numbers sorted based on number of purchases (ascending).

```python
members = df.groupby('Member_number').size()
sorted_members = members.index[np.argsort(members.values)]
print(sorted_members)
```

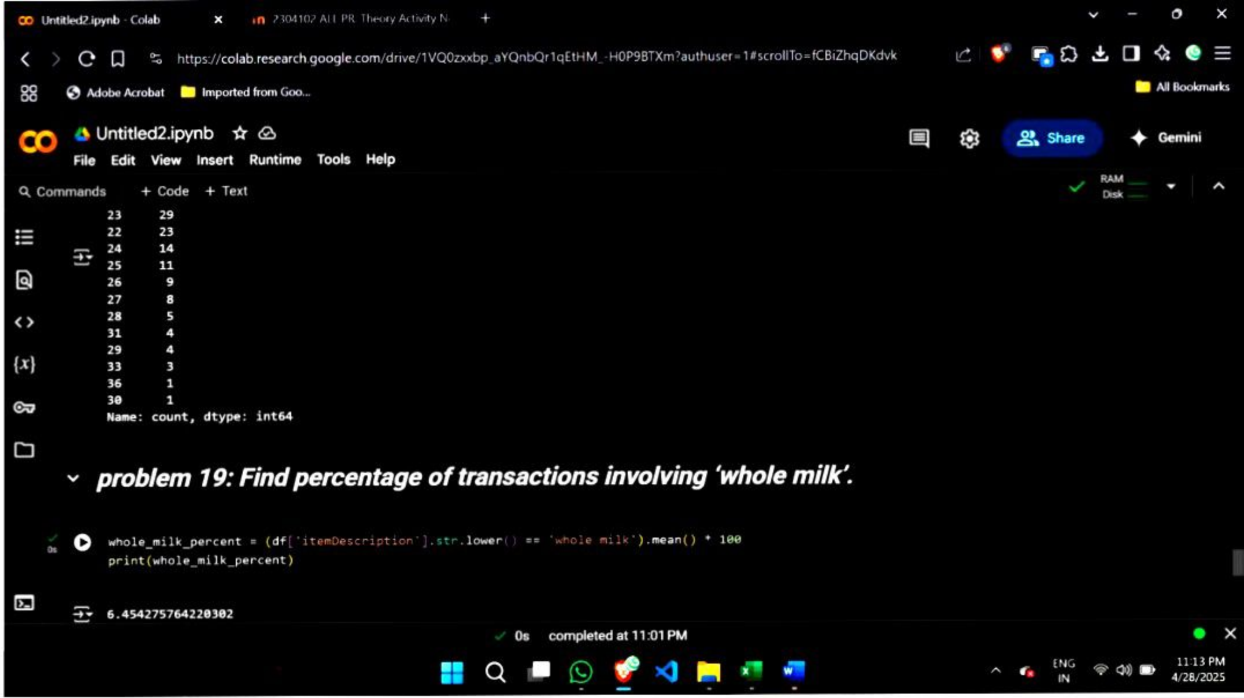
```
Index([1036, 4980, 3625, 3624, 2248, 4904, 4978, 1368, 2213, 2203,
       ...
       2394, 3872, 3915, 2433, 2271, 2625, 2051, 3050, 3737, 3180],
      dtype='int64', name='Member_number', length=3898)
```


Untitled2.ipynb - Colab
2304102 ALL PR Theory Activity N
https://colab.research.google.com/drive/1VQ0zxxbp_aYQnbQr1qEtHM_-H0P9BTXm?authuser=1#scrollTo=fCBiZhqDKdvk
Adobe Acrobat
Imported from Goo...
All Bookmarks
Untitled2.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text
Share
Gemini
RAM
Disk
0s completed at 11:01 PM
ENG IN
11:13 PM
4/28/2025

Problem 18: Count all purchases per member

```python
purchase_count = df.groupby('Member_number').size()
print(purchase_count.value_counts())
```

```
6     371
8     341
4     328
10    292
12    252
9     251
2     248
11    246
7     237
13    192
14    179
5     178
15    144
16    136
17    109
18     69
```

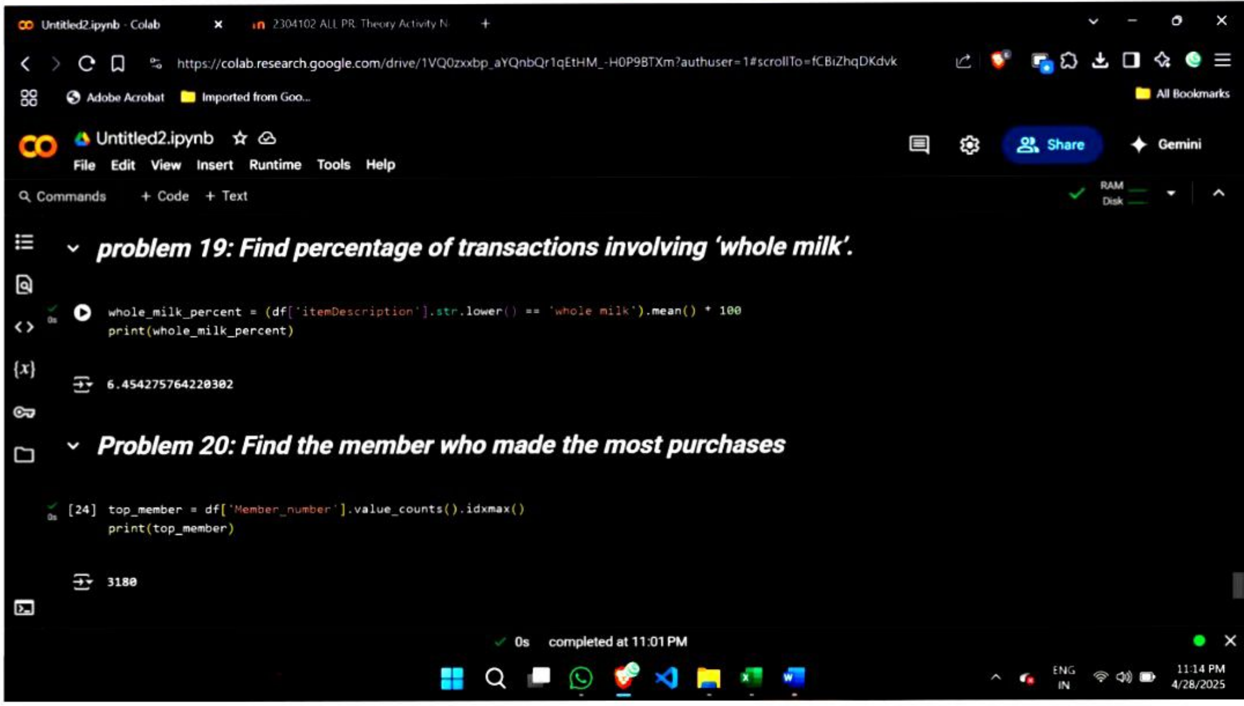


```
23     29
22     23
24     14
25     11
26      9
27      8
28      5
31      4
29      4
33      3
36      1
30      1
Name: count, dtype: int64
```

problem 19: Find percentage of transactions involving 'whole milk'.

```
whole_milk_percent = (df['itemDescription'].str.lower() == 'whole milk').mean() * 100
print(whole_milk_percent)
```

```
6.454275764220302
```



problem 19: Find percentage of transactions involving 'whole milk'.

whole_milk_percent = (df['itemDescription'].str.lower() == 'whole milk').mean() * 100
print(whole_milk_percent)

6.454275764220302

Problem 20: Find the member who made the most purchases

```
[24] top_member = df['Member_number'].value_counts().idxmax()
     print(top_member)
```

```
3180
```