NAME - Shravani Birajdar CLASS -

ET2

ROLL NO - ET2-24

BATCH - ET22

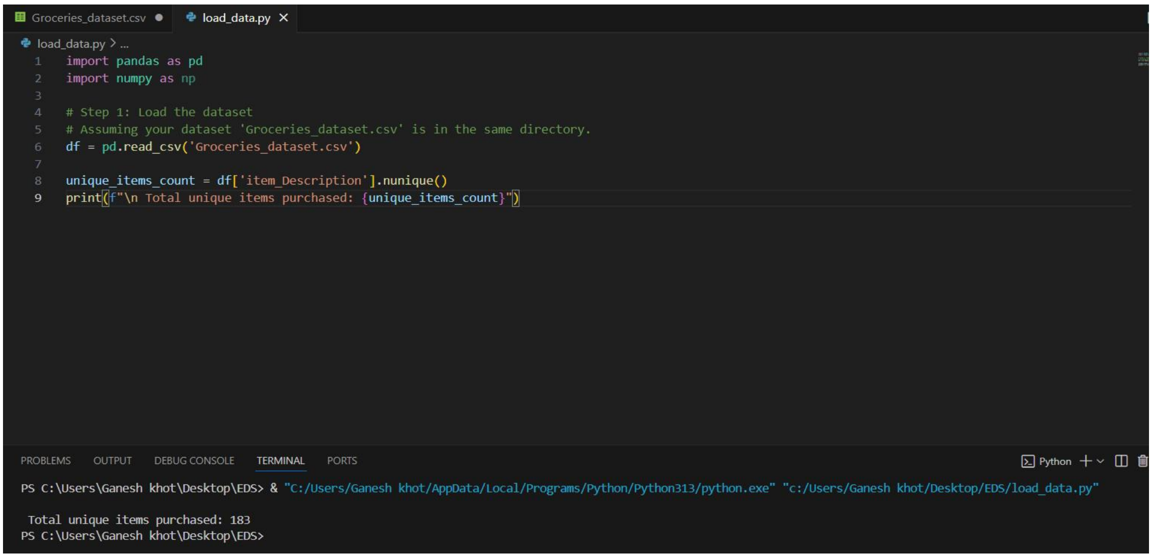
PRN-202401070065

DATA SET LINK:

https://www.kaggle.com/heeraldedhia/groceries-dataset

---

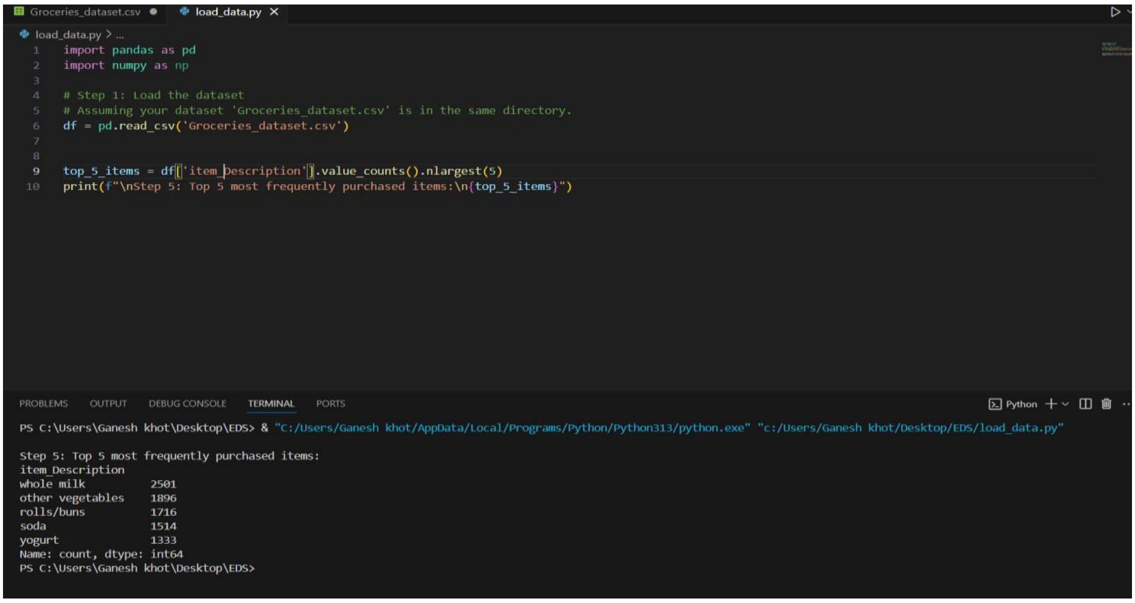
1) How many unique items were purchased in total?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')

unique_items_count = df['item_Description'].nunique()
print(f"\n Total unique items purchased: {unique_items_count}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

 Total unique items purchased: 183
PS C:\Users\Ganesh khot\Desktop\EDS>
```

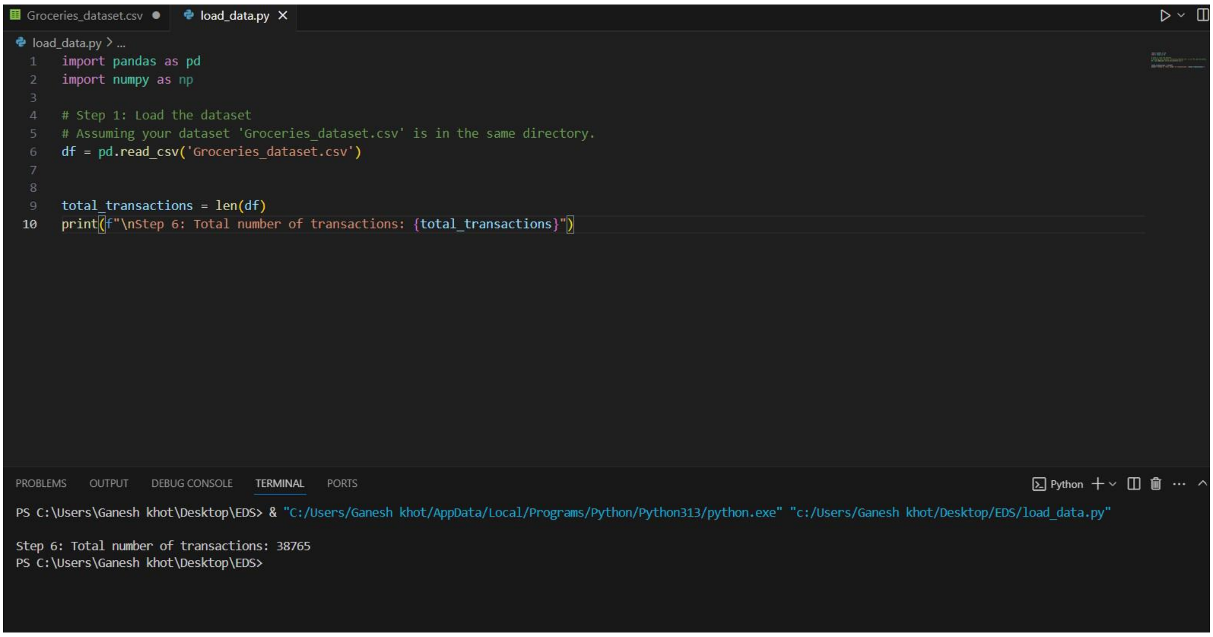
2) What are the top 5 most frequently purchased items?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')


top_5_items = df['item_Description'].value_counts().nlargest(5)
print(f"\nStep 5: Top 5 most frequently purchased items:\n{top_5_items}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 5: Top 5 most frequently purchased items:
itemDescription
whole milk          2501
other vegetables    1896
rolls/buns          1716
soda                1514
yogurt              1333
Name: count, dtype: int64
PS C:\Users\Ganesh khot\Desktop\EDS>
```

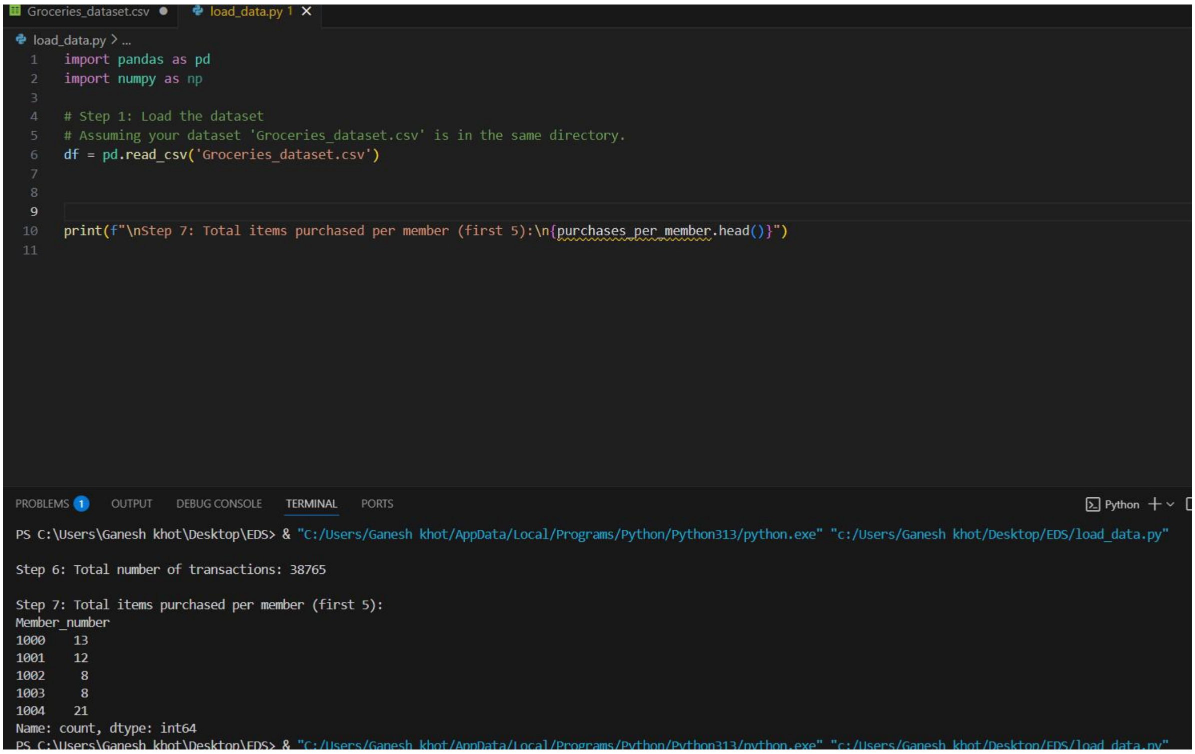
3) What is the total number of transactions in the dataset?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')


total_transactions = len(df)
print(f"\nStep 6: Total number of transactions: {total_transactions}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 6: Total number of transactions: 38765
PS C:\Users\Ganesh khot\Desktop\EDS>
```

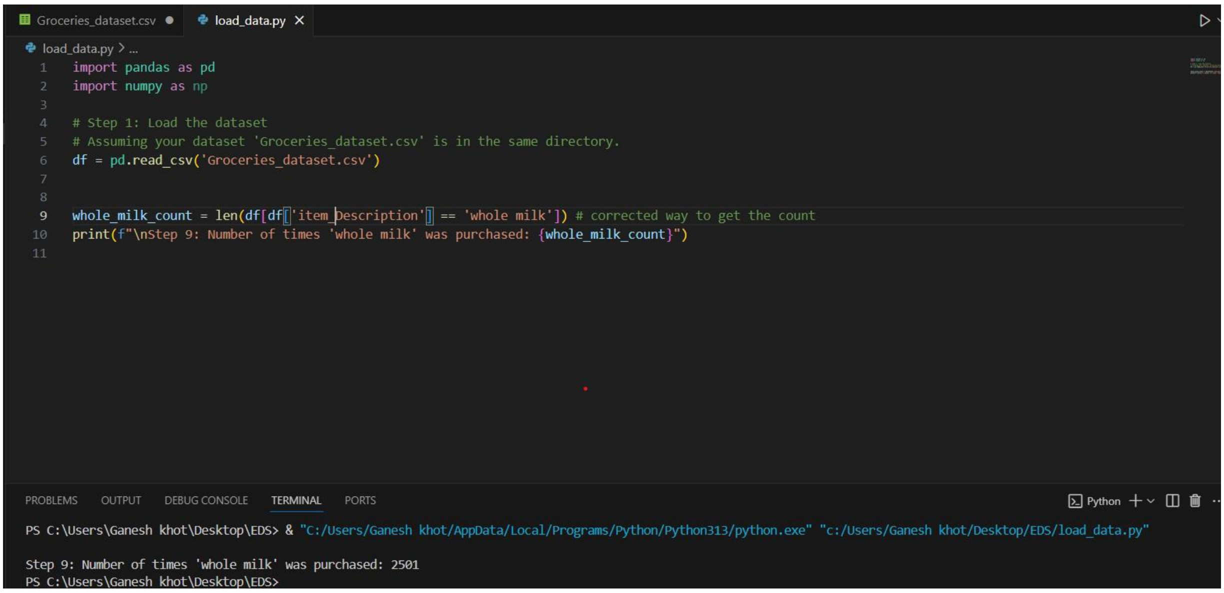
4) For each member, how many items did they purchase in total?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')



print(f"\nStep 7: Total items purchased per member (first 5):\n{purchases_per_member.head()}")

```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 6: Total number of transactions: 38765

Step 7: Total items purchased per member (first 5):
Member_number
1000    13
1001    12
1002     8
1003     8
1004    21
Name: count, dtype: int64
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"
```

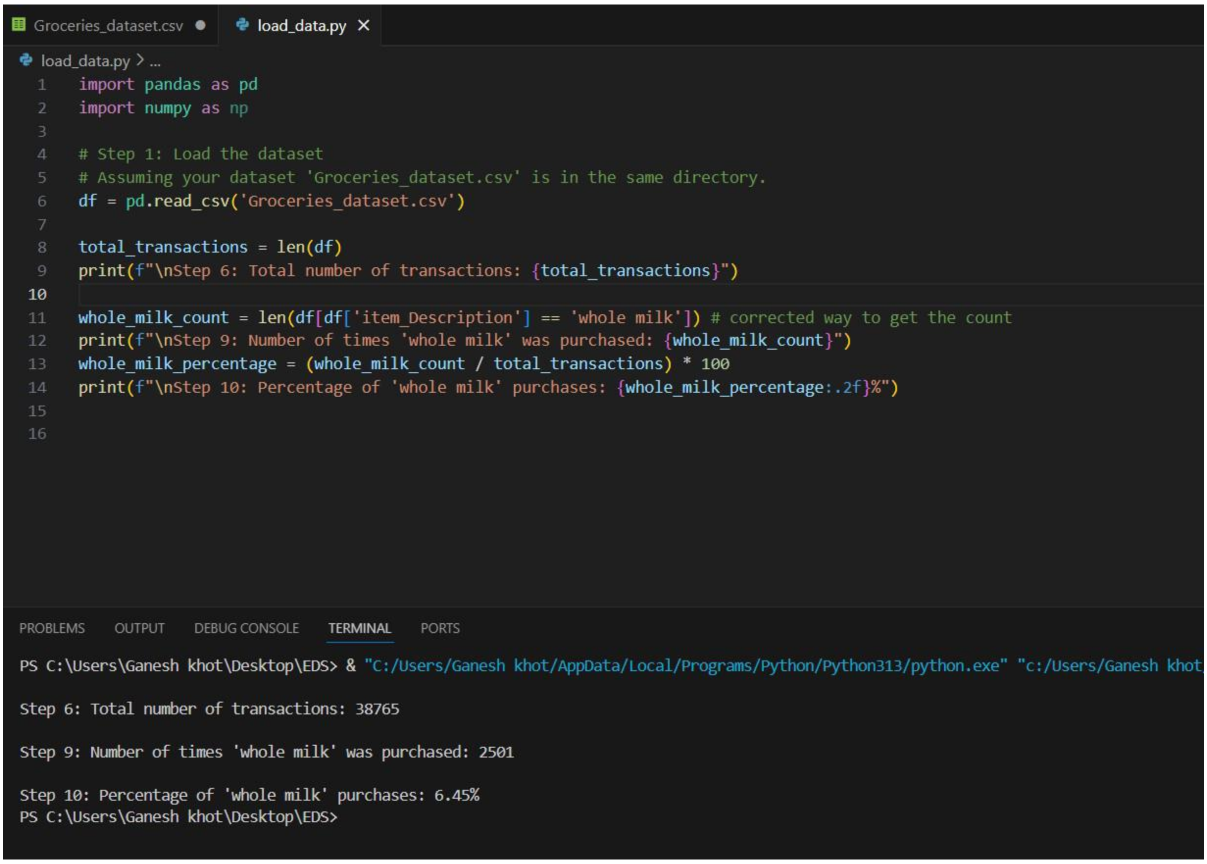
5) How many times was 'whole milk' purchased?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')


whole_milk_count = len(df[df['item_Description'] == 'whole milk']) # corrected way to get the count
print(f"\nStep 9: Number of times 'whole milk' was purchased: {whole_milk_count}")

```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 9: Number of times 'whole milk' was purchased: 2501
PS C:\Users\Ganesh khot\Desktop\EDS>
```

6) What percentage of total purchases does 'whole milk' represent?

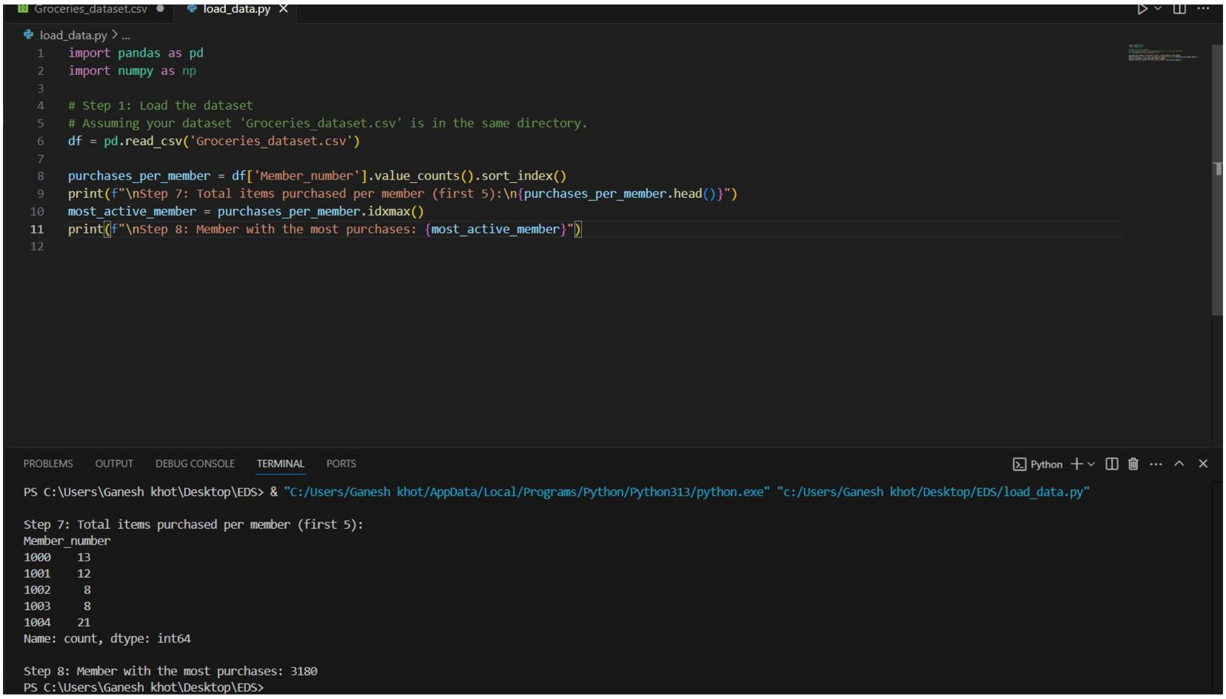


```python
import pandas as pd
import numpy as np

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')

total_transactions = len(df)
print(f"\nStep 6: Total number of transactions: {total_transactions}")

whole_milk_count = len(df[df['item_Description'] == 'whole milk']) # corrected way to get the count
print(f"\nStep 9: Number of times 'whole milk' was purchased: {whole_milk_count}")
whole_milk_percentage = (whole_milk_count / total_transactions) * 100
print(f"\nStep 10: Percentage of 'whole milk' purchases: {whole_milk_percentage:.2f}%")


```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot

Step 6: Total number of transactions: 38765

Step 9: Number of times 'whole milk' was purchased: 2501

Step 10: Percentage of 'whole milk' purchases: 6.45%
PS C:\Users\Ganesh khot\Desktop\EDS>
```

8)purchase per member

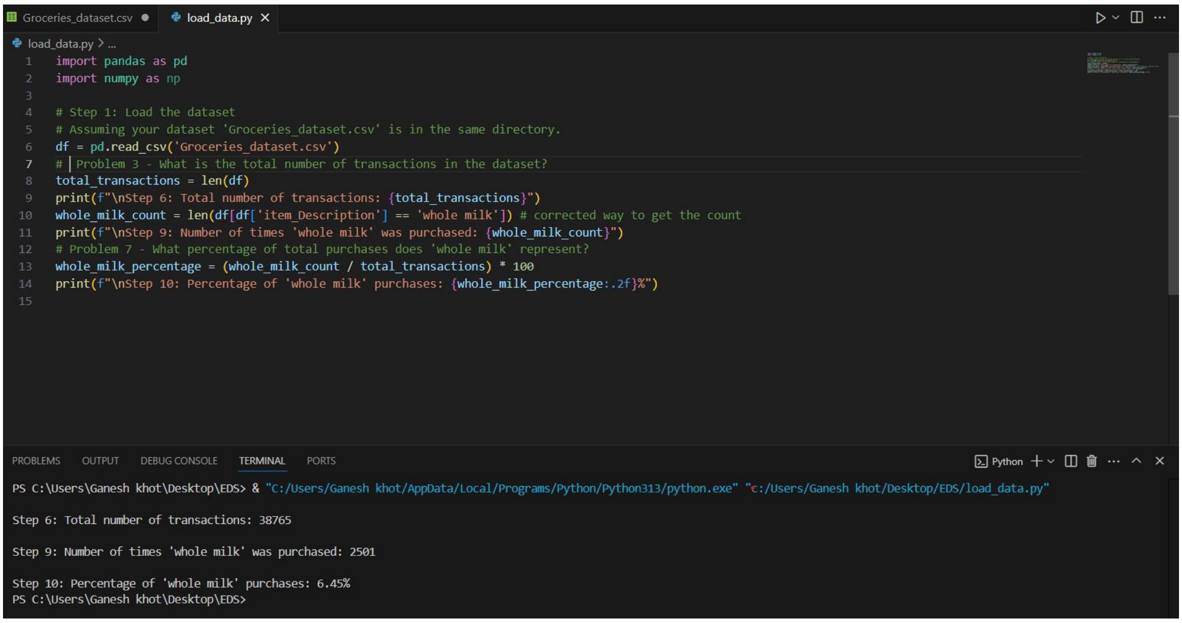
8) Which member made the most purchases?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')

purchases_per_member = df['Member_number'].value_counts().sort_index()
print(f"\nStep 7: Total items purchased per member (first 5):\n{purchases_per_member.head()}")
most_active_member = purchases_per_member.idxmax()
print(f"\nStep 8: Member with the most purchases: {most_active_member}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 7: Total items purchased per member (first 5):
Member_number
1000    13
1001    12
1002     8
1003     8
1004    21
Name: count, dtype: int64

Step 8: Member with the most purchases: 3180
PS C:\Users\Ganesh khot\Desktop\EDS>
```

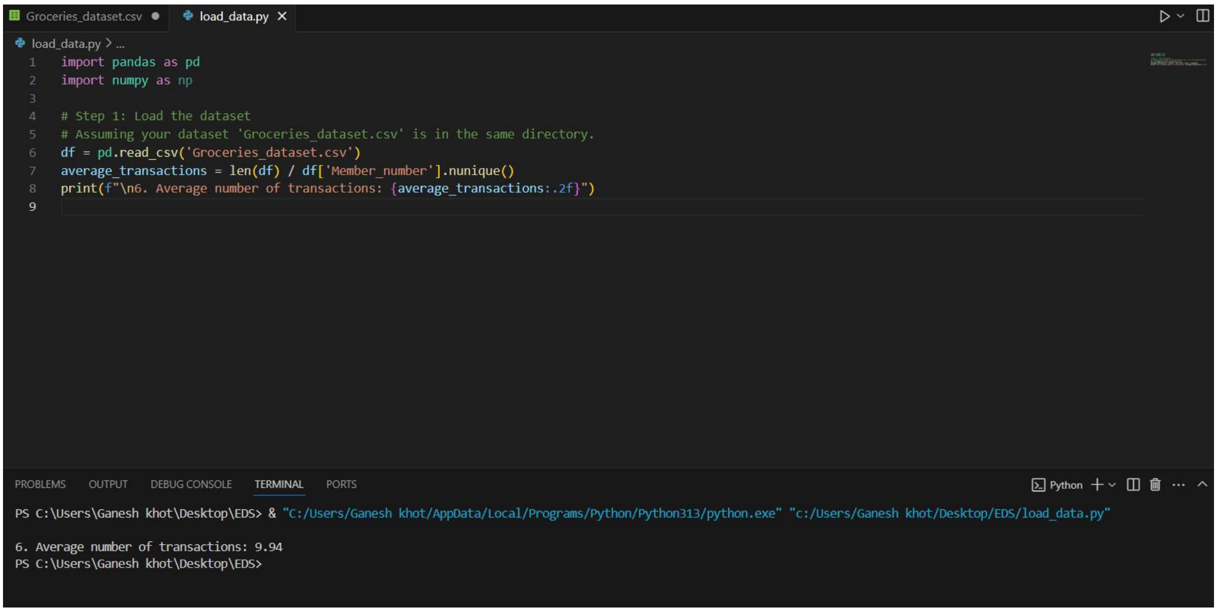
9) What percentage of total purchases does 'whole milk' represent?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')
# Problem 3 - What is the total number of transactions in the dataset?
total_transactions = len(df)
print(f"\nStep 6: Total number of transactions: {total_transactions}")
whole_milk_count = len(df[df['item_Description'] == 'whole milk']) # corrected way to get the count
print(f"\nStep 9: Number of times 'whole milk' was purchased: {whole_milk_count}")
# Problem 7 - What percentage of total purchases does 'whole milk' represent?
whole_milk_percentage = (whole_milk_count / total_transactions) * 100
print(f"\nStep 10: Percentage of 'whole milk' purchases: {whole_milk_percentage:.2f}%")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

Step 6: Total number of transactions: 38765

Step 9: Number of times 'whole milk' was purchased: 2501

Step 10: Percentage of 'whole milk' purchases: 6.45%
PS C:\Users\Ganesh khot\Desktop\EDS>
```

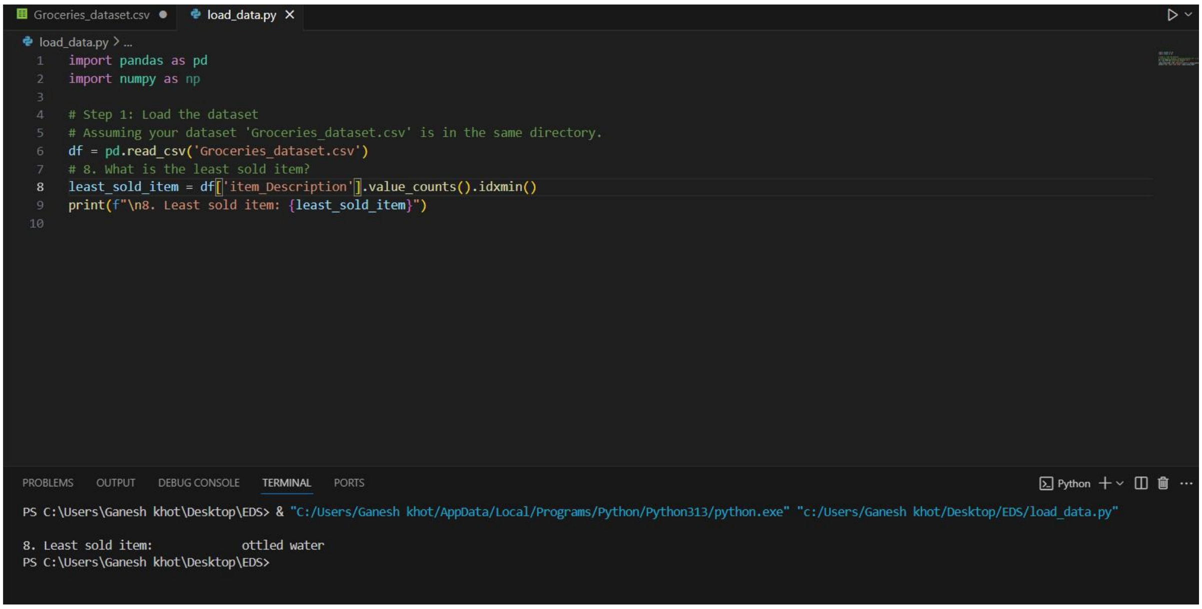
10) What is the average number of transactions in the dataset?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')
average_transactions = len(df) / df['Member_number'].nunique()
print(f"\n6. Average number of transactions: {average_transactions:.2f}")

```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

6. Average number of transactions: 9.94
PS C:\Users\Ganesh khot\Desktop\EDS>
```

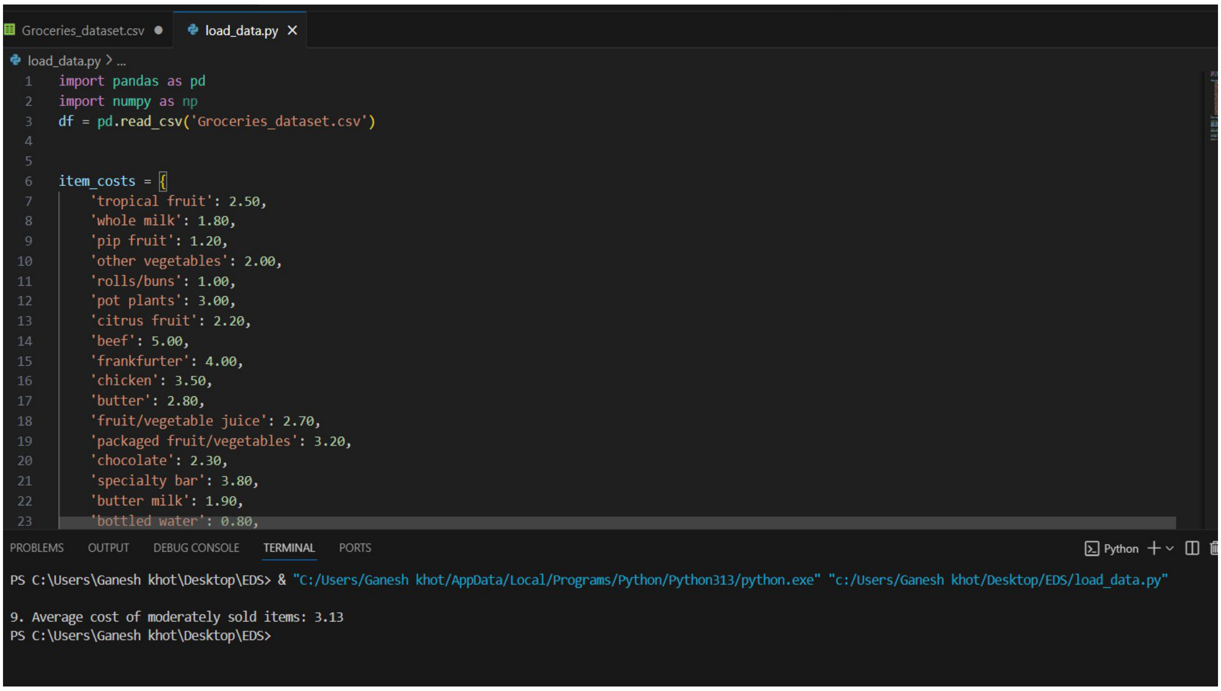
11) What is the least sold item?



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

# Step 1: Load the dataset
# Assuming your dataset 'Groceries_dataset.csv' is in the same directory.
df = pd.read_csv('Groceries_dataset.csv')
# 8. What is the least sold item?
least_sold_item = df['itemDescription'].value_counts().idxmin()
print(f"\n8. Least sold item: {least_sold_item}")

```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

8. Least sold item:             ottled water
PS C:\Users\Ganesh khot\Desktop\EDS>
```

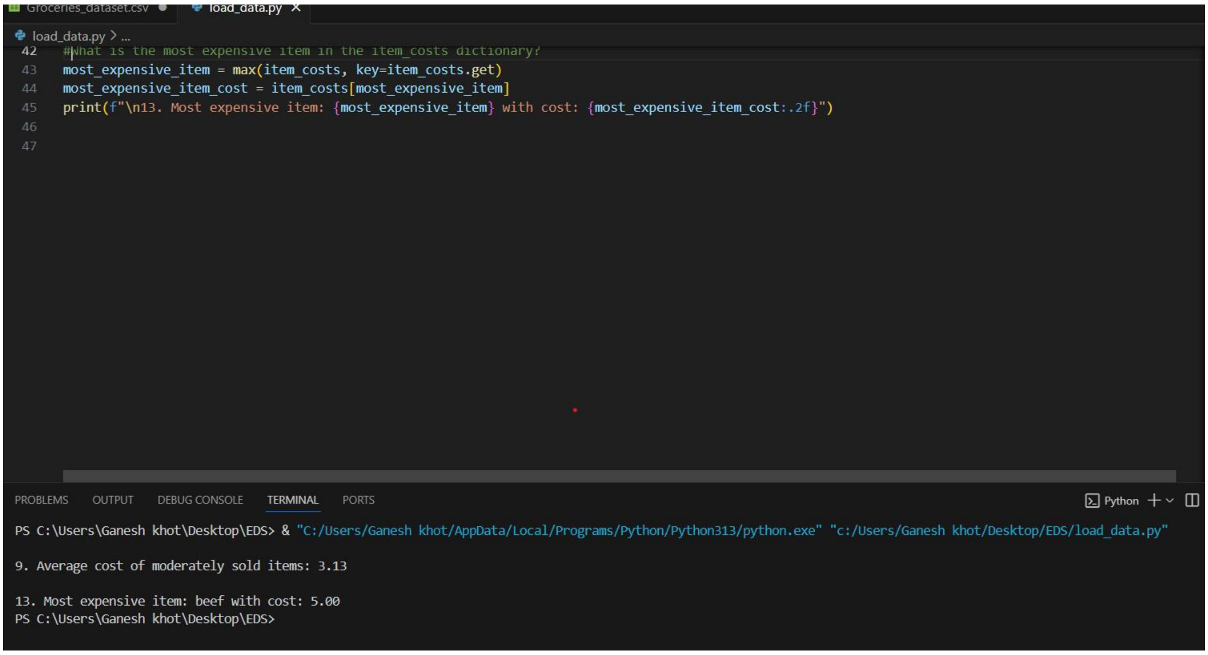
12) Average cost of moderalty sold itmes.



```python
import pandas as pd
import numpy as np
df = pd.read_csv('Groceries_dataset.csv')


item_costs = {
    'tropical fruit': 2.50,
    'whole milk': 1.80,
    'pip fruit': 1.20,
    'other vegetables': 2.00,
    'rolls/buns': 1.00,
    'pot plants': 3.00,
    'citrus fruit': 2.20,
    'beef': 5.00,
    'frankfurter': 4.00,
    'chicken': 3.50,
    'butter': 2.80,
    'fruit/vegetable juice': 2.70,
    'packaged fruit/vegetables': 3.20,
    'chocolate': 2.30,
    'specialty bar': 3.80,
    'butter milk': 1.90,
    'bottled water': 0.80,
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

9. Average cost of moderately sold items: 3.13
PS C:\Users\Ganesh khot\Desktop\EDS>
```

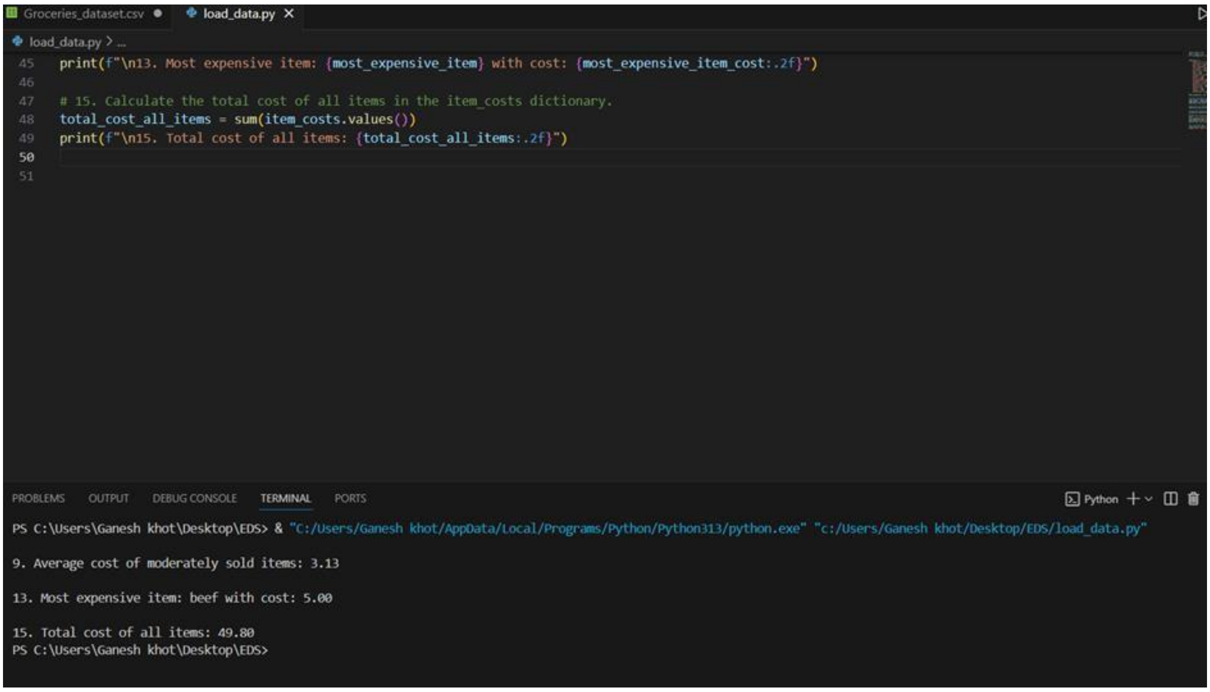
13) Most expensive item



```python
#What is the most expensive item in the item_costs dictionary?
most_expensive_item = max(item_costs, key=item_costs.get)
most_expensive_item_cost = item_costs[most_expensive_item]
print(f"\n13. Most expensive item: {most_expensive_item} with cost: {most_expensive_item_cost:.2f}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

9. Average cost of moderately sold items: 3.13

13. Most expensive item: beef with cost: 5.00
PS C:\Users\Ganesh khot\Desktop\EDS>
```

14) Calculate the total cost of all items in the item_costs dictionary.



```python
print(f"\n13. Most expensive item: {most_expensive_item} with cost: {most_expensive_item_cost:.2f}")

# 15. Calculate the total cost of all items in the item_costs dictionary.
total_cost_all_items = sum(item_costs.values())
print(f"\n15. Total cost of all items: {total_cost_all_items:.2f}")
```

```
PS C:\Users\Ganesh khot\Desktop\EDS> & "C:/Users/Ganesh khot/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Ganesh khot/Desktop/EDS/load_data.py"

9. Average cost of moderately sold items: 3.13

13. Most expensive item: beef with cost: 5.00

15. Total cost of all items: 49.80
PS C:\Users\Ganesh khot\Desktop\EDS>
```

15) What are the top 5 most frequently purchased items, and how do their costs compare to the average cost of all items?

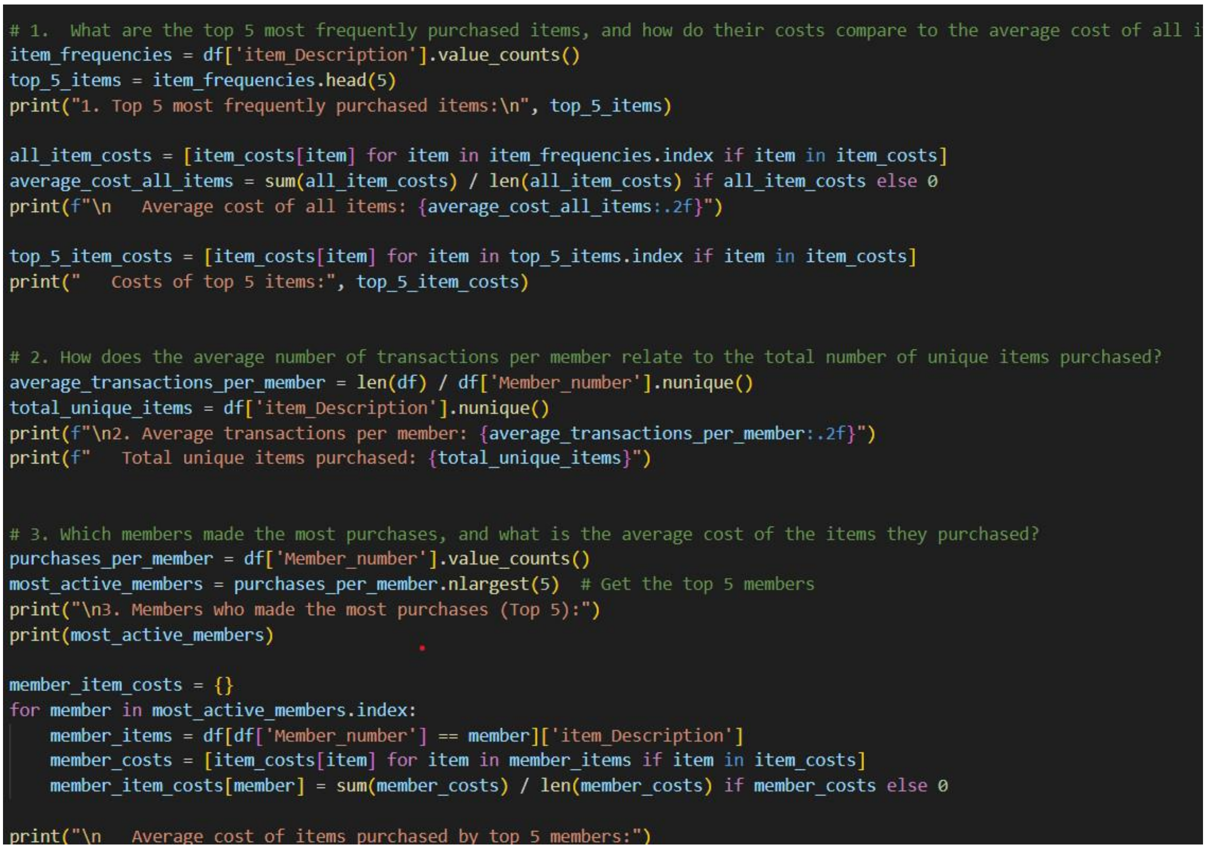
16). How does the average number of transactions per member relate to the total number of unique items purchased?

17) Which members made the most purchases, and what is the average cost of the items they purchased?

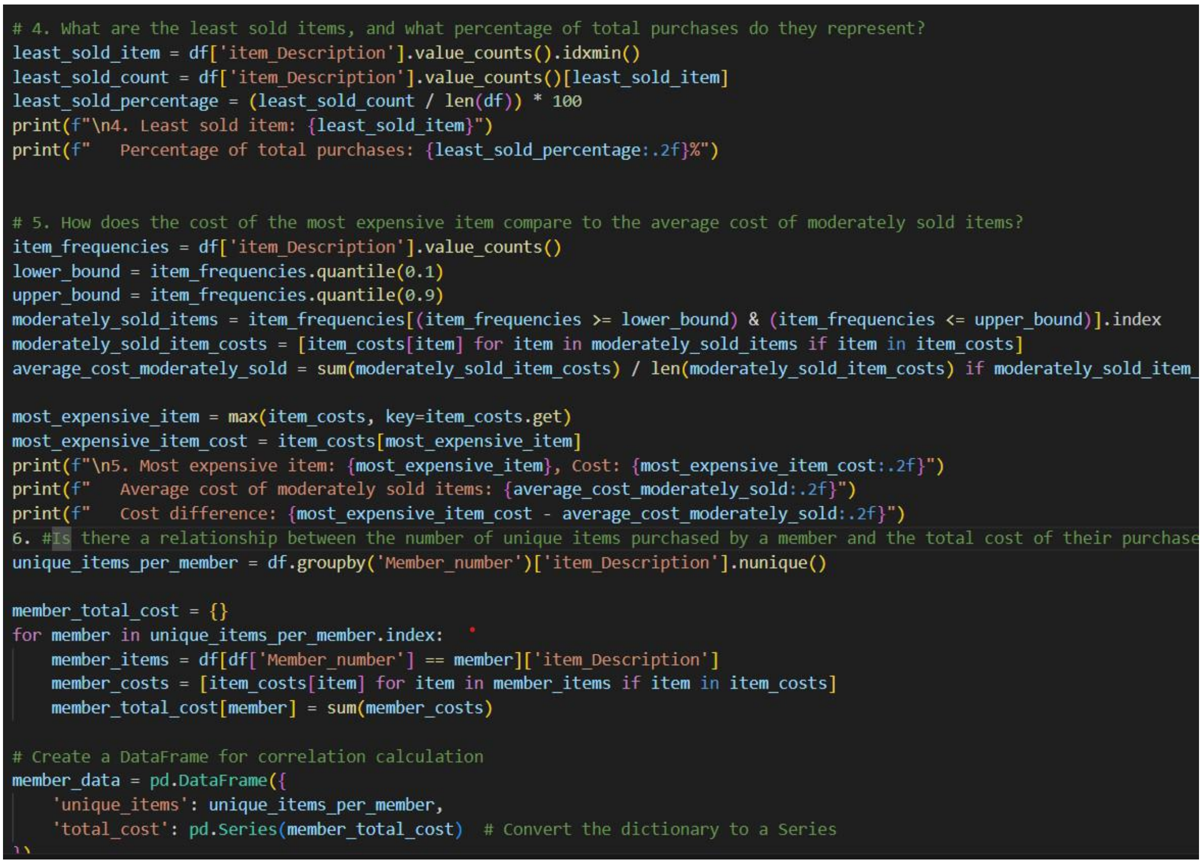
18) What are the least sold items, and what percentage of total purchases do they represent?

19)How does the cost of the most expensive item compare to the average cost of moderately sold items?

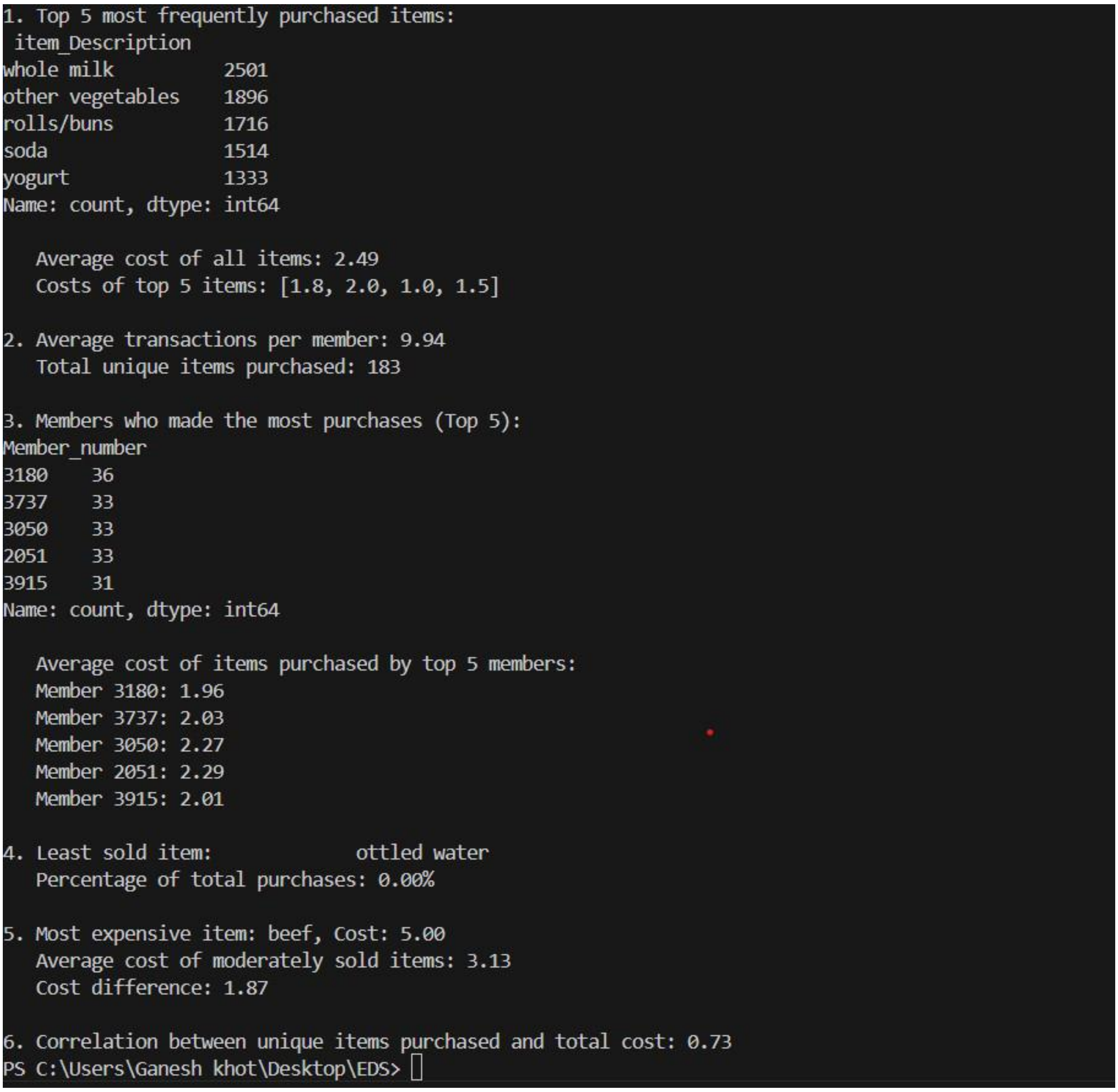
20) Is there a relationship between the number of unique items purchased by a member and the total cost of their purchases?



```
# 1.  What are the top 5 most frequently purchased items, and how do their costs compare to the average cost of all i
item_frequencies = df['item_Description'].value_counts()
top_5_items = item_frequencies.head(5)
print("1. Top 5 most frequently purchased items:\n", top_5_items)

all_item_costs = [item_costs[item] for item in item_frequencies.index if item in item_costs]
average_cost_all_items = sum(all_item_costs) / len(all_item_costs) if all_item_costs else 0
print(f"\n   Average cost of all items: {average_cost_all_items:.2f}")

top_5_item_costs = [item_costs[item] for item in top_5_items.index if item in item_costs]
print("   Costs of top 5 items:", top_5_item_costs)


# 2. How does the average number of transactions per member relate to the total number of unique items purchased?
average_transactions_per_member = len(df) / df['Member_number'].nunique()
total_unique_items = df['item_Description'].nunique()
print(f"\n2. Average transactions per member: {average_transactions_per_member:.2f}")
print(f"   Total unique items purchased: {total_unique_items}")


# 3. Which members made the most purchases, and what is the average cost of the items they purchased?
purchases_per_member = df['Member_number'].value_counts()
most_active_members = purchases_per_member.nlargest(5)  # Get the top 5 members
print("\n3. Members who made the most purchases (Top 5):")
print(most_active_members)

member_item_costs = {}
for member in most_active_members.index:
    member_items = df[df['Member_number'] == member]['item_Description']
    member_costs = [item_costs[item] for item in member_items if item in item_costs]
    member_item_costs[member] = sum(member_costs) / len(member_costs) if member_costs else 0

print("\n   Average cost of items purchased by top 5 members:")
```



```
# 4. What are the least sold items, and what percentage of total purchases do they represent?
least_sold_item = df['item_Description'].value_counts().idxmin()
least_sold_count = df['item_Description'].value_counts()[least_sold_item]
least_sold_percentage = (least_sold_count / len(df)) * 100
print(f"\n4. Least sold item: {least_sold_item}")
print(f"   Percentage of total purchases: {least_sold_percentage:.2f}%")


# 5. How does the cost of the most expensive item compare to the average cost of moderately sold items?
item_frequencies = df['item_Description'].value_counts()
lower_bound = item_frequencies.quantile(0.1)
upper_bound = item_frequencies.quantile(0.9)
moderately_sold_items = item_frequencies[(item_frequencies >= lower_bound) & (item_frequencies <= upper_bound)].index
moderately_sold_item_costs = [item_costs[item] for item in moderately_sold_items if item in item_costs]
average_cost_moderately_sold = sum(moderately_sold_item_costs) / len(moderately_sold_item_costs) if moderately_sold_item_

most_expensive_item = max(item_costs, key=item_costs.get)
most_expensive_item_cost = item_costs[most_expensive_item]
print(f"\n5. Most expensive item: {most_expensive_item}, Cost: {most_expensive_item_cost:.2f}")
print(f"   Average cost of moderately sold items: {average_cost_moderately_sold:.2f}")
print(f"   Cost difference: {most_expensive_item_cost - average_cost_moderately_sold:.2f}")
6. #Is there a relationship between the number of unique items purchased by a member and the total cost of their purchase
unique_items_per_member = df.groupby('Member_number')['item_Description'].nunique()

member_total_cost = {}
for member in unique_items_per_member.index:
    member_items = df[df['Member_number'] == member]['item_Description']
    member_costs = [item_costs[item] for item in member_items if item in item_costs]
    member_total_cost[member] = sum(member_costs)

# Create a DataFrame for correlation calculation
member_data = pd.DataFrame({
    'unique_items': unique_items_per_member,
    'total_cost': pd.Series(member_total_cost)  # Convert the dictionary to a Series
})
```



```
1. Top 5 most frequently purchased items:
 item_Description
whole milk          2501
other vegetables    1896
rolls/buns          1716
soda                1514
yogurt              1333
Name: count, dtype: int64

   Average cost of all items: 2.49
   Costs of top 5 items: [1.8, 2.0, 1.0, 1.5]

2. Average transactions per member: 9.94
   Total unique items purchased: 183

3. Members who made the most purchases (Top 5):
Member_number
3180    36
3737    33
3050    33
2051    33
3915    31
Name: count, dtype: int64

   Average cost of items purchased by top 5 members:
   Member 3180: 1.96
   Member 3737: 2.03
   Member 3050: 2.27
   Member 2051: 2.29
   Member 3915: 2.01

4. Least sold item:             ottled water
   Percentage of total purchases: 0.00%

5. Most expensive item: beef, Cost: 5.00
   Average cost of moderately sold items: 3.13
   Cost difference: 1.87

6. Correlation between unique items purchased and total cost: 0.73
PS C:\Users\Ganesh khot\Desktop\EDS> 
```