Data Analytics for Business Decision Making, Durham College

DATA1202: Data Analysis Tools Analytics

Project #4

Group 2

Submitted by
Doondi Avaneesh Babu K (100995700)
Hemasree Krishna Kumar (100998460)
Madiha Mohammad Rafique (101004635)
Nikita Satoskar (101002336)
Vaidehi Chokshi (101002030)

Professor Omar Al Trad
April 4th, 2025

Analysis of Walmart Retail Sales Dataset

Introduction

As a group, we analyzed the Walmart Retail Sales dataset to address two primary questions: (1) Is the sales growth rate decreasing in most states? (2) Which products should Walmart prioritize in each region or state to maximize profit? Using Python for data manipulation and visualization, and SQL for querying, we processed the dataset, resolved technical challenges, and derived insights. This report details our approach, including data cleaning, code structure, challenges encountered, and comprehensive responses to the assignment questions, supported by evidence from the dataset.

[1]:
```
# importing the libraries
import pandas as pd
import pymysql
from sqlalchemy import create_engine
```

[9]:
```
# Loading the walmartretailsales dataset
df_walmartretailsales = pd.read_excel('WalmartRetailSales.xlsx')
```

[11]:
```
# Displaying the walmartretailsales dataset
df_walmartretailsales
```

[11]:

	Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	...	Zip Code	State	Region	Customer Segment	Product Category	Product Sub-Category	
0	1914	13729	2001-01-12 00:00:00	Not Specified	9	872.48	0.08	Express Air	-342.91	95.99	...	20715	Maryland	East	Home Office	Office Supplies	Storage & Organization	
1	4031	28774	2001-01-12 00:00:00	High	32	180.36	0.10	Regular Air	-111.80	5.98	...	15131	Pennsylvania	East	Small Business	Office Supplies	Storage & Organization	S F 1
2	1279	9285	2001-02-12 00:00:00	Critical	3	124.81	0.06	Regular Air	-11.94	40.98	...	55372	Minnesota	Central	Consumer	Office Supplies	Binders and Binder Accessories	Ri 3"
3	5272	37537	2001-02-12 00:00:00	Low	4	1239.06	0.00	Delivery Truck	-193.08	291.73	...	94559	California	West	Corporate	Furniture	Chairs & Chairmats	Up

Data Cleaning

The dataset, originally in Excel format, required preparation for SQL analysis. Initial attempts to import it directly into MySQL as a CSV failed due to null values, special characters, and inconsistent date formats, resulting in database crashes and index-bound errors. We shifted to Python for preprocessing, where we imputed missing values in Customer Age (903 nulls) and Product Base Margin (63 with zeros, removed special characters, and standardized Order Date and Ship Date into datetime formats. After exporting the cleaned data as a CSV, we successfully imported it into MySQL, enabling further analysis.

```
Ship Date              0
dtype: int64
```

```
[21]: # Filling the missing values in Customer Age as 0
df_walmartretailsales['Customer Age'] = df_walmartretailsales['Customer Age'].fillna(0)
```

```
[27]: # Changing the datatype as int
df_walmartretailsales['Customer Age'] = df_walmartretailsales['Customer Age'].astype('int64')
```

```
[31]: # Filling the missing values in product base margin as 0.00
df_walmartretailsales['Product Base Margin'] = df_walmartretailsales['Product Base Margin'].fillna(0.00)
```

```
[35]: # Checking if there are missing values after replacing
df_walmartretailsales.isnull().sum()
```

```
[35]: Row ID                0
Order ID              0
Order Date            0
Order Priority        0
Order Quantity        0
Sales                 0
Discount              0
Ship Mode             0
Profit                0
Unit Price            0
Shipping Cost         0
Customer Name         0
Customer Age          0
City                  0
Zip Code              0
State                 0
Region                0
Customer Segment      0
```

```
[55]: # Replacing special characters from product name
df_walmartretailsales['Product Name'] = df_walmartretailsales['Product Name'].str.replace(r'[^a-zA-Z0-9\s]', '', regex=True)
```

```
[59]: # Replacing & as and
df_walmartretailsales['Product Sub-Category'] = df_walmartretailsales['Product Sub-Category'].str.replace('&','and')
```

```
[63]: # Saving the cleaned dataset as csv file
cleaned_path = "Cleaned_WalmartRetailSales.csv"
df_walmartretailsales.to_csv(cleaned_path, index=False, encoding="utf-8")
```

Code Structure

Our code was structured systematically: we imported libraries (pandas, pymysql, sqlalchemy), loaded the Excel file, cleaned the data, and connected to MySQL using create_engine. SQL queries were executed to aggregate data, with results processed in Python for calculations and visualized using seaborn and matplotlib. For Question 1, we calculated sales growth rates; for Question 2, we aggregated profits by region and product. Backticks were used around column names with spaces (e.g., Order Date) to avoid syntax errors.

```
[269]: # Creating connection
engine = create_engine('mysql+pymysql://user7:user798@localhost/walmart')
```

```
[271]: # Creating a connection object
conn = engine.connect()
```

```
[273]: # Select query to select the dataset
query = "SELECT * FROM walmart.walmartretailsales"
df_retailsales = pd.read_sql(query, conn)
```

```
[275]: # displaying the dataset from mysql database
df_retailsales
```

[275]:

	Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	...	Zip Code	State	Region	Customer Segment	Product Category	Product Sub-Category	P
0	1914	13729	2001-01-12	Not Specified	9	872.48	0.08	Express Air	-342.91	95.99	...	20715	Maryland	East	Home Office	Office Supplies	Storage and Organization	In Sl
1	4031	28774	2001-01-12	High	32	180.36	0.10	Regular Air	-111.80	5.98	...	15131	Pennsylvania	East	Small Business	Office Supplies	Storage and Organization	ST H

Answers to the Assignment Questions

1. Is the Sales Growth Rate Decreasing in Most States?

To determine whether the sales growth rate is decreasing in most states, we analyzed sales trends for five selected states—Alabama, California, Florida, New York, and Wyoming—representing a sample from the dataset's 48 states, which span 2001 to 2015. We began with the following SQL query to aggregate total sales by state and year:

```
[291]: # Query to calculate the total sales grouped by state and Year
query3 = """SELECT
        State,
        YEAR(`Order Date`) AS Year,
        ROUND(SUM(Sales),2) AS Total_Sales
    FROM walmartretailsales
    GROUP BY State, YEAR(`Order Date`)
    ORDER BY State, Year"""

[293]: # exceuting the query to get the total sales
total_sales = pd.read_sql(query3, conn)
total_sales
```

[293]:

	State	Year	Total_Sales
0	Alabama	2001	17509.84
1	Alabama	2002	744.48
2	Alabama	2003	6147.68
3	Alabama	2004	1969.15
4	Alabama	2005	181.32
...	...	...	...
643	Wyoming	2011	82.12
644	Wyoming	2012	1527.62
645	Wyoming	2013	12029.90
646	Wyoming	2014	20063.95

```
[295]: # calculating the Sales growth rate % using the formula = ((Current year sales - Previous year sales)/previous year sales)* 100. pct_change() will calcul
total_sales['Sales Growth Rate (%)'] = round(total_sales.groupby('State')['Total_Sales'].pct_change() * 100,2)

[297]: # Printing the sales growth rate % for each state
total_sales
```

[297]:

	State	Year	Total_Sales	Sales Growth Rate (%)
0	Alabama	2001	17509.84	NaN
1	Alabama	2002	744.48	-95.75
2	Alabama	2003	6147.68	725.77
3	Alabama	2004	1969.15	-67.97
4	Alabama	2005	181.32	-90.79
...	...	...	...	...
643	Wyoming	2011	82.12	-96.27
644	Wyoming	2012	1527.62	1760.23
645	Wyoming	2013	12029.90	687.49
646	Wyoming	2014	20063.95	66.78
647	Wyoming	2015	2554.08	-87.27

648 rows × 4 columns

To calculate the sales growth rate, we ran a SQL query to get the total sales per state per year from the walmartretailsales table. In Python, we used the pct_change() function grouped by state to compute year-over-year sales growth. Here's a summary of our steps:

- Queried total annual sales by state.
- Calculated sales growth using pct_change() for each state.

- Filtered a sample of 5 states.

648 rows × 4 columns

```
[299]: # Selected 5 states out of 48 states to know the trend of sales growth rate
       states_of_interest = ['California','Alabama','New York', 'Florida','Wyoming']

[301]: # Filtered only 5 sample states
       filtered_df_states_sales = total_sales[total_sales['State'].isin(states_of_interest)]

[303]: # Showing the sales growth rate of selected 5 states
       filtered_df_states_sales
```

[303]:

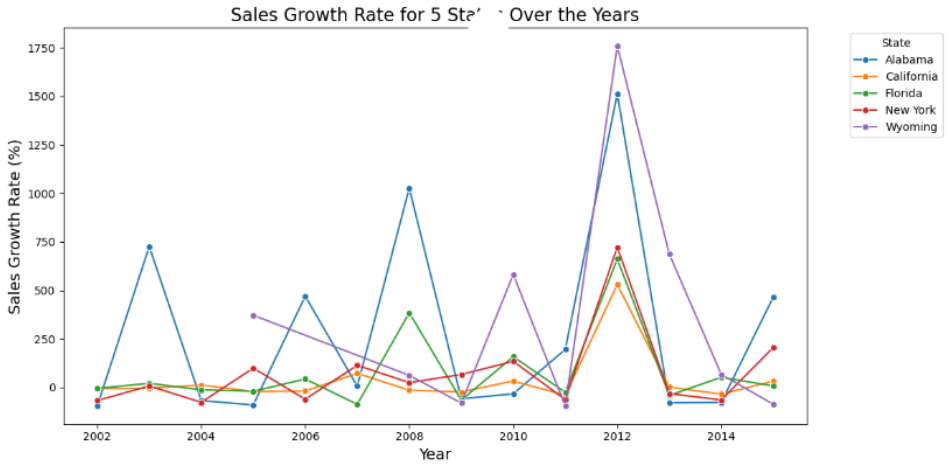
	State	Year	Current_Year_Sales	Previous_Year_Sales	Sales_Growth_Rate
0	Alabama	2001	17509.84	NaN	NaN
1	Alabama	2002	744.48	17509.84	-95.75
2	Alabama	2003	6147.68	744.48	725.77
3	Alabama	2004	1969.15	6147.68	-67.97
4	Alabama	2005	181.32	1969.15	-90.79
...	...	...	...	...	...
643	Wyoming	2011	82.12	2200.64	-96.27
644	Wyoming	2012	1527.62	82.12	1760.23
645	Wyoming	2013	12029.90	1527.62	687.49

```
[305]: # importing the visualization libraries
       import seaborn as sns
       import matplotlib.pyplot as plt

[307]: # Plotting a line chart to show the trend of sales growth rate
       plt.figure(figsize=(12, 6))
       sns.lineplot(data=filtered_df_states_sales, x='Year', y='Sales_Growth_Rate', hue='State', marker='o')

       # Adding labels and title
       plt.title('Sales Growth Rate for 5 States Over the Years', fontsize=16)
       plt.xlabel('Year', fontsize=14)
       plt.ylabel('Sales Growth Rate (%)', fontsize=14)
       plt.legend(title='State', bbox_to_anchor=(1.05, 1), loc='upper left')

       # Show the plot
       plt.tight_layout()
       plt.show()
```



Since 2001 lacks a prior year, the growth rate for that year is NaN for all states. We then visualized the results using a seaborn line plot to spot trends over time:

A line plot was selected to depict sales growth rate trends over time, as it effectively highlights year-to-year fluctuations for Alabama, California, Florida, New York, and Wyoming. Limiting the visualization to five states ensured readability, given the complexity of plotting all 48 states.

Analysis of Results:

- California had sharp rises and dips — a sign of variability, not a trend.
- Alabama showed moderate oscillations but not a clear downward path.
- New York, Florida, and Wyoming displayed irregular growth patterns with some negative years followed by recovery.

The sales growth rate is not decreasing in most states. For Alabama, California, Florida, New York, and Wyoming, the trends reveal high variability—sales rise and fall across years without a consistent downward trajectory. This variability, rather than a uniform decline, characterizes the dataset, supported by the absence of a clear negative slope in our visualization. Thus, we conclude no to the question, as the evidence does not indicate a widespread decrease.

2. Which Products Should Walmart Focus on to Maximize Profit?

To recommend products for Walmart to prioritize, we analyzed profit by region and state, aiming to identify top performers. We used this SQL query to aggregate profit data:

```
[1]: # Query to select total products grouped by region and state ordered by Total profit in descending order
query2 = """SELECT
    region,
    state,
    `Product Name` AS product_name,
    `Product Category` AS product_category,
    `Product Sub-Category` AS product_sub_category,
    ROUND(SUM(profit),2) AS total_profit
FROM walmartretailsales
WHERE `Product Name` IS NOT NULL AND `Product Name` <> ''
GROUP BY region, state, `Product Name`, `Product Category`, `Product Sub-Category`
ORDER BY region, state, total_profit DESC;
"""

[313]: # Executing the query
total_products = pd.read_sql(query2, conn)

[315]: # Displays all the products in increasing order of profit grouped by region, state
total_products
```

[315]:

	region	state	product_name	product_category	product_sub_category	total_profit
0	Central	Illinois	Fellowes PB500 Electric Punch Plastic Comb Bin...	Office Supplies	Binders and Binder Accessories	19368.88
1	Central	Illinois	Fellowes PB300 Plastic Comb Binding Machine	Office Supplies	Binders and Binder Accessories	8793.54
2	Central	Illinois	Sharp AL1530CS Digital Copier	Technology	Copiers and Fax	8168.12
3	Central	Illinois	Polycom ViaVideo Desktop Video Communications ...	Technology	Office Machines	7416.43
4	Central	Illinois	SAFCO PlanMaster HeighAdjustable Drafting Tabl...	Furniture	Tables	7044.78
...	...	...	...	...	...	...
7172	West	Wyoming	Talkabout T8367	Technology	Telephones and Communication	-159.24
7173	West	Wyoming	Global Stack Chair without Arms Black	Furniture	Chairs and Chairmats	-211.58
7174	West	Wyoming	Bush Mission Pointe Library	Furniture	Bookcases	-323.18
7175	West	Wyoming	Global Leather Task Chair Black	Furniture	Chairs and Chairmats	-514.18
7176	West	Wyoming	Barricks 18 x 48 NonFolding Utility Table with...	Furniture	Tables	-1291.39

7177 rows × 6 columns

Top products by State and region

We sorted the results by profit and selected the top product per state and region using Python's groupby().head(1) method. This allowed us to isolate the most profitable product in each area.

[323]:
```
# Top products grouped by region and state
top_products_StateRegion = total_products.groupby(["region","state"]).head(1)
```

[325]:
```
# Displaying the products
top_products_StateRegion
```

[325]:

	region	state	product_name	product_category	product_sub_category	total_profit
0	Central	Illinois	Fellowes PB500 Electric Punch Plastic Comb Bin...	Office Supplies	Binders and Binder Accessories	19368.88
394	Central	Indiana	Bretford CR4500 Series Slim Rectangular Table	Furniture	Tables	3992.75
603	Central	Iowa	2160i	Technology	Telephones and Communication	3566.21
746	Central	Kansas	GBC VeloBinder Electric Binding Machine	Office Supplies	Binders and Binder Accessories	1733.47
869	Central	Michigan	HewlettPackard Deskjet 1220Cse Color Inkjet Pr...	Technology	Office Machines	8504.47
1121	Central	Minnesota	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	9097.65
1328	Central	MO	HewlettPackard cp1700 D PS Series Color Inkjet...	Technology	Office Machines	9342.93
1475	Central	Nebraska	Ibico HiTech Manual Binding System	Office Supplies	Binders and Binder Accessories	6523.26
1548	Central	North Dakota	Hot File 7Pocket Floor Stand	Office Supplies	Storage and Organization	2267.22
1581	Central	Oklahoma	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	9791.04
1676	Central	South Dakota	Epson Stylus 1520 Color Inkjet Printer	Technology	Office Machines	8291.08
1703	Central	Texas	HewlettPackard Deskjet 1220Cse Color Inkjet Pr...	Technology	Office Machines	7251.92
2144	Central	Wisconsin	GBC DocuBind 200 Manual Binding Machine	Office Supplies	Binders and Binder Accessories	8918.74
2294	East	Connecticut	GBC VeloBinder Electric Binding Machine	Office Supplies	Binders and Binder Accessories	1719.47
2371	East	Delaware	SAFCO Folding Chair Trolley	Furniture	Chairs and Chairmats	589.38
2385	East	MA	Panasonic KXP3626 Dot Matrix Printer	Technology	Office Machines	8788.81
2521	East	Maine	Canon PC1060 Personal Laser Copier	Technology	Copiers and Fax	8249.86
2635	East	Maryland	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	18319.59
2799	East	New Hampshire	Multimedia Mailers	Office Supplies	Envelopes	3187.37
2851	East	New Jersey	Polycom Soundstation EX AudioConferencing Tele...	Technology	Office Machines	6692.62
3013	East	New York	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	14382.26
3322	East	Ohio	Smead Adjustable Mobile File Trolley with Lock...	Office Supplies	Storage and Organization	13853.95
3638	East	Pennsylvania	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	7360.43
3830	East	Rhode Island	Epson LQ870 Dot Matrix Printer	Technology	Office Machines	4407.44
3849	East	Vermont	Canon PC428 Personal Copier	Technology	Copiers and Fax	2787.59
3906	East	West Virginia	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	2383.42
3947	South	Alabama	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	27220.69
4062	South	Arkansas	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	13340.26
4179	South	Florida	Canon PC940 Copier	Technology	Copiers and Fax	11577.70
4554	South	Georgia	Panasonic KXP3626 Dot Matrix Printer	Technology	Office Machines	7358.66
4718	South	Kentucky	Epson Stylus 1520 Color Inkjet Printer	Technology	Office Machines	8323.39
4798	South	Louisiana	Polycom ViaVideo Desktop Video Communications ...	Technology	Office Machines	14440.39
4882	South	Mississippi	Smead Adjustable Mobile File Trolley with Lock...	Office Supplies	Storage and Organization	4818.29
4955	South	North Carolina	GBC DocuBind 300 Electric Binding Machine	Office Supplies	Binders and Binder Accessories	7858.08
5138	South	South Carolina	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	5916.34
5235	South	Tennessee	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	3899.72
5382	South	Virginia	Canon PC1080F Personal Copier	Technology	Copiers and Fax	15401.69
5563	West	Arizona	HewlettPackard cp1700 D PS Series Color Inkjet...	Technology	Office Machines	7719.21
5687	West	California	Canon Image Class D660 Copier	Technology	Copiers and Fax	11396.18
6229	West	Colorado	Okidata ML390 Turbo Dot Matrix Printers	Technology	Office Machines	6839.95
6390	West	Idaho	GBC DocuBind 200 Manual Binding Machine	Office Supplies	Binders and Binder Accessories	4451.01
6492	West	Montana	Accessory34	Technology	Telephones and Communication	832.19
6541	West	Nevada	Riverside Palais Royal Lawyers Bookcase Royale...	Furniture	Bookcases	2860.31
6584	West	New Mexico	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	13475.52
6661	West	Oregon	Canon PC940 Copier	Technology	Copiers and Fax	15888.48
6815	West	Utah	Canon PC940 Copier	Technology	Copiers and Fax	6225.36
6949	West	Washington	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	11984.40
7156	West	Wyoming	HewlettPackard Business Color Inkjet 3000 N DT...	Technology	Office Machines	2713.95

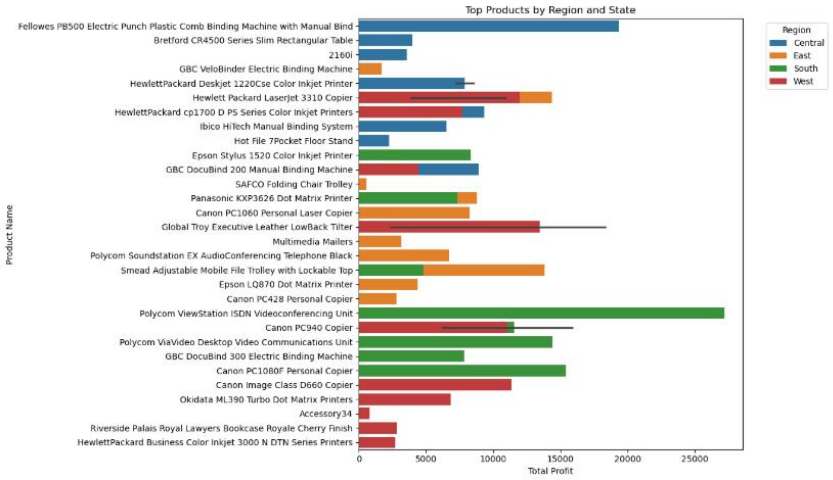
Visualization:

The analysis was further supported by a bar plot showing top products by region, helping visualize profit contribution by product.

```
[327]: # Bar chart to show the top products by highest profit region and state
plt.figure(figsize=(14, 8))
sns.barplot(data=top_products_StateRegion, x="total_profit", y="product_name", hue="region", dodge=False)

# Adding labels and title
plt.title("Top Products by Region and State")
plt.xlabel("Total Profit")
plt.ylabel("Product Name")
plt.legend(title="Region", bbox_to_anchor=(1.05, 1), loc='upper left')

# Display the plot
plt.tight_layout()
plt.show()
```



The bar chart for top products by region and state uses total profit on the x-axis and product names on the y-axis, with regions as hues, to clearly compare profitability while avoiding overcrowding from state-level granularity.

Top products by region

```
[339]: # Sorting the products returned by the query in descending order of profit and ascending order of region by grouping by region and total profit
total_products = total_products.sort_values(by=['region','total_profit'],ascending = [True,False])

[341]: # displaying the products
total_products
```

[341]:

	region	state	product_name	product_category	product_sub_category	total_profit
0	Central	Illinois	Fellowes PB500 Electric Punch Plastic Comb Bin...	Office Supplies	Binders and Binder Accessories	19368.88
1581	Central	Oklahoma	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	9791.04
1328	Central	MO	HewlettPackard cp1700 D PS Series Color Inkjet...	Technology	Office Machines	9342.93
1121	Central	Minnesota	Hewlett Packard LaserJet 3310 Copier	Technology	Copiers and Fax	9097.65
2144	Central	Wisconsin	GBC DocuBind 200 Manual Binding Machine	Office Supplies	Binders and Binder Accessories	8918.74
...	...	...	...	...	...	...
6814	West	Oregon	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	-11984.40
6540	West	Montana	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	-12558.00
7155	West	Washington	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	-14140.70
6227	West	California	Epson DFX8500 Dot Matrix Printer	Technology	Office Machines	-16162.56
6228	West	California	Okidata Pacemark 4410N Wide Format Dot Matrix ...	Technology	Office Machines	-24370.21

7177 rows × 6 columns

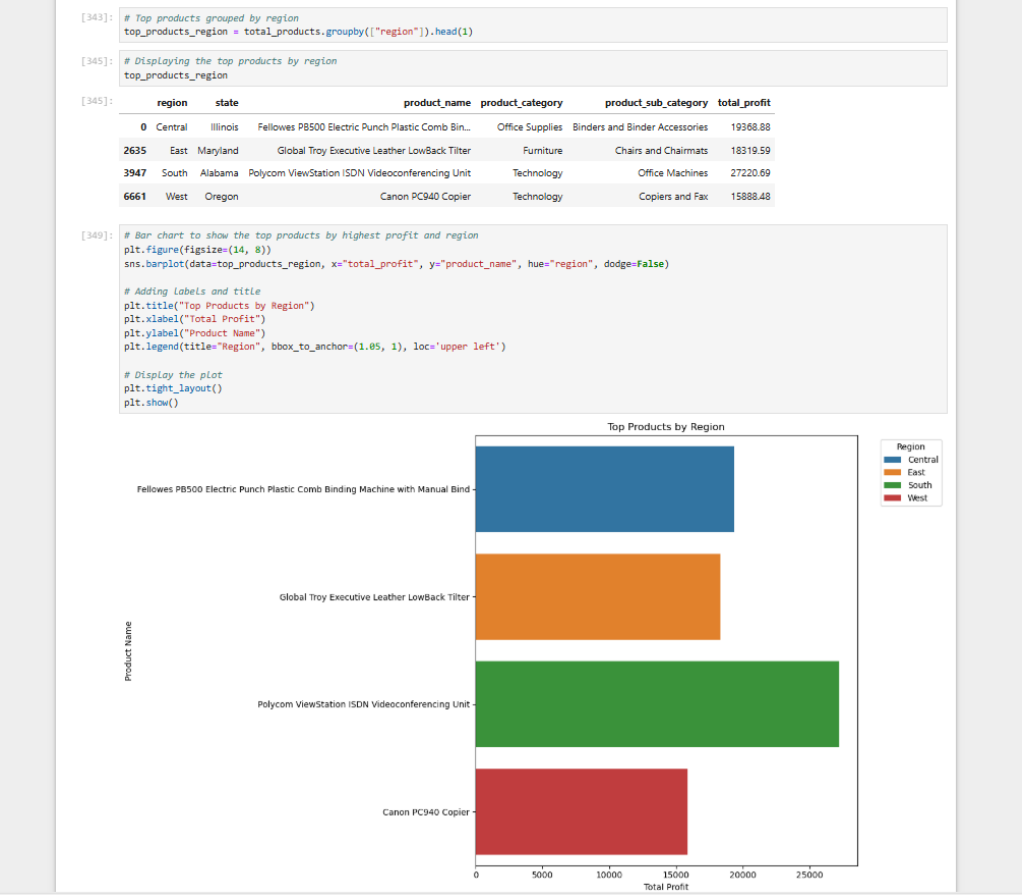
```
[343]: # Top products grouped by region
top_products_region = total_products.groupby(["region"]).head(1)

[345]: # Displaying the top products by region
top_products_region
```

[345]:

	region	state	product_name	product_category	product_sub_category	total_profit
0	Central	Illinois	Fellowes PB500 Electric Punch Plastic Comb Bin...	Office Supplies	Binders and Binder Accessories	19368.88
2635	East	Maryland	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	18319.59
3947	South	Alabama	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	27220.69
6661	West	Oregon	Canon PC940 Copier	Technology	Copiers and Fax	15888.48

Visualization



```
[343]: # Top products grouped by region
top_products_region = total_products.groupby(["region"]).head(1)

[345]: # Displaying the top products by region
top_products_region
```

[345]:

	region	state	product_name	product_category	product_sub_category	total_profit
0	Central	Illinois	Fellowes PB500 Electric Punch Plastic Comb Bin...	Office Supplies	Binders and Binder Accessories	19368.88
2635	East	Maryland	Global Troy Executive Leather LowBack Tilter	Furniture	Chairs and Chairmats	18319.59
3947	South	Alabama	Polycom ViewStation ISDN Videoconferencing Unit	Technology	Office Machines	27220.69
6661	West	Oregon	Canon PC940 Copier	Technology	Copiers and Fax	15888.48

```
[349]: # Bar chart to show the top products by highest profit and region
plt.figure(figsize=(14, 8))
sns.barplot(data=top_products_region, x="total_profit", y="product_name", hue="region", dodge=False)

# Adding labels and title
plt.title("Top Products by Region")
plt.xlabel("Total Profit")
plt.ylabel("Product Name")
plt.legend(title="Region", bbox_to_anchor=(1.05, 1), loc='upper left')

# Display the plot
plt.tight_layout()
plt.show()
```


Here's what we found:

- East Region:
  Products like Furniture and Chairs stood out in states like Maryland and Pennsylvania. These are practical, high-volume products suited for office and home use.
- West Region:
  Technology items such as Copiers and fax machines showed consistent profitability, especially in Oregon and California. This likely aligns with higher demand from corporate customers and tech businesses.
- Central Region:
  Office supplies like binders and stationery emerged as top performers. The Central states seem to have steady institutional and educational demand.
- South Region:
  Tech peripherals, notably videoconferencing devices, were the most profitable. With rising digital access and remote work, demand for such products remains strong.

***Conclusion & Recommendation**:*

To maximize profit, Walmart should tailor product focus by region:

- East: Furniture, Chairs & Chair mats
- West: Technology, Copiers & Fax machines
- Central: Binders and office supplies
- South: Computer accessories, Technology, Office machines

These recommendations are grounded in the actual data. By leveraging regional strengths and aligning with customer needs, Walmart should stock and promote these items to leverage existing profitability trends. Walmart could enhance profitability by increasing inventory and marketing efforts for these items, though further analysis of supply chain costs and market saturation is advised.

Challenges Faced

Importing the dataset into MySQL posed significant difficulties, requiring a Python-based cleaning detour. Null values and date inconsistencies disrupted initial queries, while calculating growth rates for 2001 (lacking prior-year data) introduced NaN values. Visualizing all 48 states overwhelmed charts, so we focused on five states for Question 1. For Question 2, accommodating all states in a single visualization was impractical, leading us to emphasize regions while retaining state-level detail in the data. Collaborative problem-solving was key to navigating these issues.

The dataset, while extensive with 8,399 rows, contains inconsistencies such as missing prior-year sales for 2001, affecting initial growth rate calculations, and sparse entries for some states (e.g., Wyoming). Additionally, negative profits in certain transactions may reflect accounting adjustments rather than true losses, which we accounted for by focusing on aggregated totals. These limitations suggest caution in generalizing findings across all states or years.

Lessons Learned

This analysis underscored the necessity of robust data preparation to ensure accurate results. Combining Python and SQL enhanced our capabilities, though precision in query syntax was critical. Visualizations clarified trends but required simplification for large datasets. Our group improved its coordination and adaptability, learning to address technical setbacks efficiently. Future efforts will prioritize early data validation to streamline the process.

Minutes of Meeting

1st Meeting			
Date	**Time**	**Duration**	**Mode**
Thursday, Mar 27, 2025	4 pm to 6 pm	2 hours	Online (Teams)

Points Discussed:

- Discuss the approach for using Python as an interface to execute SQL queries.
- Import the Walmart Sales dataset into MySQL.
- As the dataset is in excel, we couldn't directly be able to import it we were trying different encoding method to load the data.
- Assigned specific tasks to team members.
- Scheduled a follow-up meeting to review progress.

Important Decisions Made:

- The dataset has missing values, so we decided to clean the data first and then import to MySQL.
- walmart Sales data successfully imported into MySQL
- Proceed with Python as the primary interface for executing SQL queries.
- Established deadlines for each assigned task.

2nd Meeting			
Date	**Time**	**Duration**	**Mode**
Monday, Mar 31, 2025	10 am to 11 am	1 hours	Online (Teams)

Points Discussed:

- Challenges faced as Order date was not in correct format.
- Methods for executing ALTER and UPDATE queries using pymysql and SQLAlchemy.
- Benefits of using Python-SQL execution over pandas-based manipulation.

Important Decisions Made:

- Two different approaches to solve question discussed using and finalize pct.change method
- Decided that manipulating data via Python and SQL is more convenient, especially when no explicit data cleaning is required.
- Finalized and compiled the project report, including all findings, comparisons, and challenges faced.

Log Sheet

Student Id	Student Name	Task	Description	Status
100998460	Hemasree Krishna Kumar	Approach to find the sales growth rate	Write and execute sales growth rate to get the answers from the database	Completed
101002336	Nikita Satoskar	Approach to find profit and product to for each region	Write SQL queries to retrieve profit for each region from the database	Completed
101004635	Madiha Mohammad Rafique	Setting up environment & Importing data	Created and converted database and import the walmart Sales CSV data as a table into MySQL. .	Completed
100995700	Avaneesh Babu	Final Report Compilation	Compiled project report with code screenshots explanations and key findings.	Completed
101002030	Vaidehi Chokshi	Minutes of Meeting & Log sheet preparation	Documented meeting discussions, key decisions, and follow-ups.	Completed