Sales Insight Analysis Using Retail Superstore Dataset

Name: Vedant P. Darvekar

Tools Used: Excel, SQL(postgresql), Python(Pandas,Seaborn,Matplotlib), Power BI

Date: October 2025

Project Summary:

This project, “Sales Insight Analysis,” focuses on analyzing retail sales data to uncover key trends in sales, profit, customer behaviour, and product performance across different regions and categories. The analysis follows a complete workflow — data cleaning in Excel, data validation and aggregation in SQL, exploratory data analysis (EDA) in Python, and visualization in Power BI. The goal is to transform raw data into actionable insights that can support informed decision-making in a retail environment.

Problem Statement & Objective:

The Superstore is experiencing fluctuations in sales and profit across various product categories, regions, and customer segments. Despite having a large amount of transactional data, the management lacks a clear understanding of which factors drive profitability and how discounts, shipping modes, or regional demand impact overall performance. Without proper analysis, it becomes difficult to identify underperforming areas, forecast future sales, and make informed business decisions. Therefore, a detailed data

analysis is needed to uncover sales patterns, customer preferences, and operational inefficiencies that affect business growth.

Objectives

- To clean and preprocess the Superstore dataset to ensure accuracy and reliability of analysis.
- To explore and analyz sales and profit trends across regions, categories, and customer segments.
- To identify key performance indicators (KPIs) influencing overall sales and profitability.
- To visualize insights through an interactive Power BI dashboard for better decision-making.
- To provide actionable recommendations for optimizing sales strategy, discount distribution, and inventory management.

Dataset Overview:

The dataset used for this project is the Superstore Sales Dataset, which provides detailed transactional data related to product sales, customers, and shipping across different regions. It serves as the foundation for performing sales, profit, and customer behaviour analysis.

Total Records: 9,994

Total Columns: 21

Missing Values: None

Data Structure:

Numerical Features: Sales, Quantity, Discount, Profit, Postal_Code, Row_ID.

Categorical Features: Order_ID, Order_Date, Ship_Date, Ship_Mode, Customer_ID, Customer_Name, Segment, Country, City, State, Region, Product_ID, Category, SubCategory, Product_Name.

Key Statistical Insights:

- Average Sales: $229.86
- Average Quantity Sold: 3.79 units
- Average Discount: 15.6%
- Average Profit: $28.66
- Highest Sales Value: $22,638.48
- Highest Profit: $8,399.98
- Lowest Profit: -$6,599.98

Data Processing and Visualization Workflow:

- **Data Cleaning and Preprocessing With Excel**: The dataset was first imported and reviewed to ensure its quality and readiness for analysis.

  Steps performed:

- Checked for missing values — none were found, ensuring a complete dataset.
- Verified data types for each column and converted date fields (Order_Date, Ship_Date) into proper datetime format.

- Removed unnecessary columns like Row_ID which didn't contribute to analysis.
- Standardized categorical fields (e.g., ensuring consistent casing in Category, Region, Segment).
- Validated that all numerical columns (Sales, Profit, Discount, Quantity) contained only valid numeric entries.

Data Validation and Aggregation With SQL:

To ensure the accuracy and consistency of the Superstore dataset, several validation and aggregation steps were performed using SQL queries. This stage verified data integrity before further analysis and visualization.

Data Validation Steps:

- **Table Creation:**
  A structured table named superstore_data was created with appropriate data types for each column.
- **Row Count Validation & Missing Value Checks:**
  Verified total record & checked for null values iin key numeric fields.

```sql
-- Total Row Count --
select count(*) as total_rows from superstore_data;

-- Missing Sales --
select sales from superstore_data as missing_sales
where sales is null;

-- Missing profit --
select profit from superstore_data as missing_profit
where profit is null;

-- Missing Quantity --
select quantity from superstore_data as missing_quantity
where quantity is null;

-- Missing order_date --
select order_date from superstore_data as missing_order_date
where order_date is null;
```

- **Duplicate ,Inconsistent & Negative Value Check:** Ensured uniqueness of each order entry by grouping on order_id.

Detected and removed inconsistent data. Such as where profit exceeded sales or when sales /quantity were negative or zero.

```
-- Checking Duplicates --
select order_id, count(*) as duplicate_count from superstore_data
group by order_id having count(*) > 1;

-- Inconssistent Data --
select * from superstore_data
where profit > sales;

-- Negative or Zero Sales/Quantity --
select * from superstore_data
where sales <= 0 or quantity <= 0;
```

Data Aggregation Steps:

After validation, data was aggregated to summarize sales performance across multiple dimensions.

- **Total Sales, Profit & Discount:**

```
select sum(sales) as total_sale from superstore_data;
select sum(discount) as total_discount from superstore_data;
select sum(profit) as total_profit from superstore_data;
```

- **Maximun and Minimum Sale & Profit:**

```
-- Min & Max sales & Profit --
select  max(sales) as maximum_sale,
min(sales) as minimum_sale
from superstore_data;
select  max(profit) as maximum_profit,
min(profit) as minimum_profit
from superstore_data;
```

- **Total Sales and Profit by Region:**

```sql
--Total Sales And Profit By Region --
select region,
       sum(sales) as total_sales,
       sum(profit) as total_profit
from superstore_data
group by region
order by total_sales desc;
```

- **Average Profit By Category:**

```sql
-- Average Profit By Category --
select category,
       avg(profit) as avg_profit
from superstore_data
group by category
order by avg_profit;
```

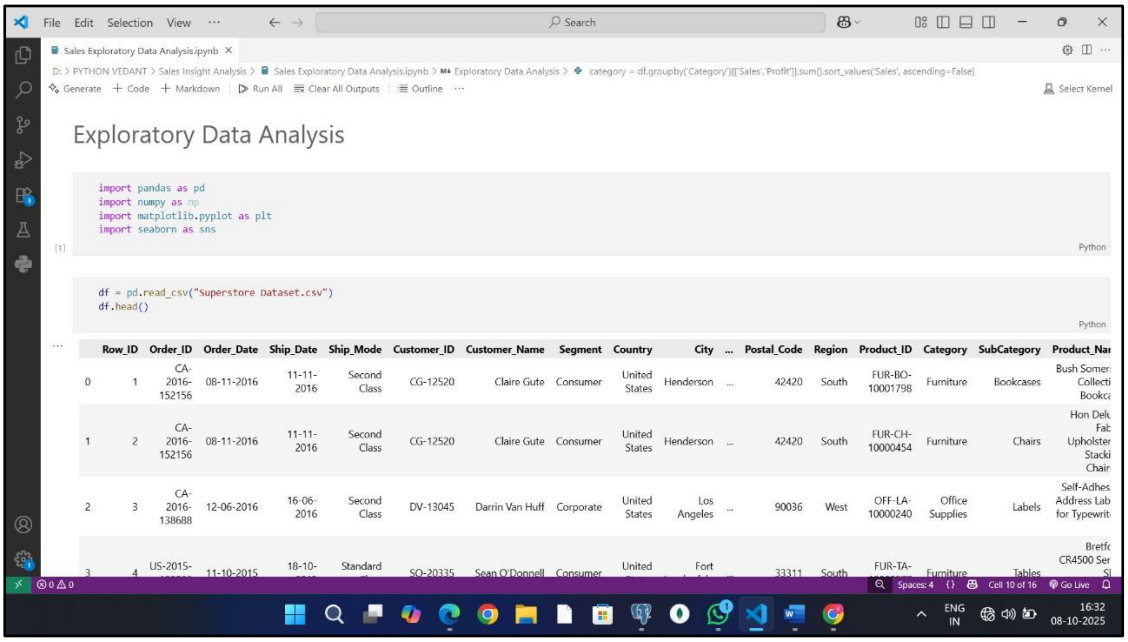
- **Monthly Sales Trend:**

```sql
-- Monthly Sales Trend --
select extract(year from order_date) as year,
       extract(month from order_date) as month,
       sum(sales) as monthly_sales
from superstore_data
group by year,month
order by year,month;
```

- **Top 5 product by Sales:**

```sql
-- Top 5 product by Sales--
select product_id,
       product_name,
       sum(sales) as Total_sale
from superstore_data
group by product_id,product_name
order by total_sale desc limit 5
```

Exploratory Data Analysis With Python:



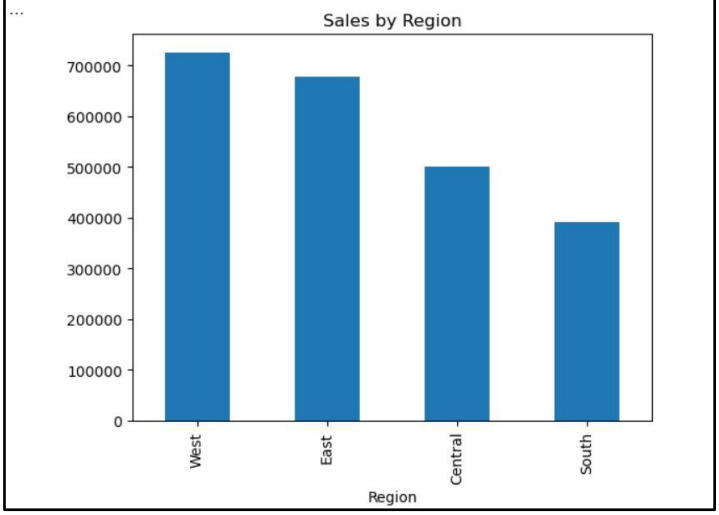
Installation of the essential liberaries and reading csv.

- **Sales & Profit By Category**:

Technology category has highest profit even with moderate sales

- **Sales By Region:**

```
region = df.groupby('Region')['Sales'].sum().sort_values(ascending=False)
region.plot(kind='bar', title='Sales by Region')
plt.show()

## West region generates the most revenue
```



West region generates the most revenue.

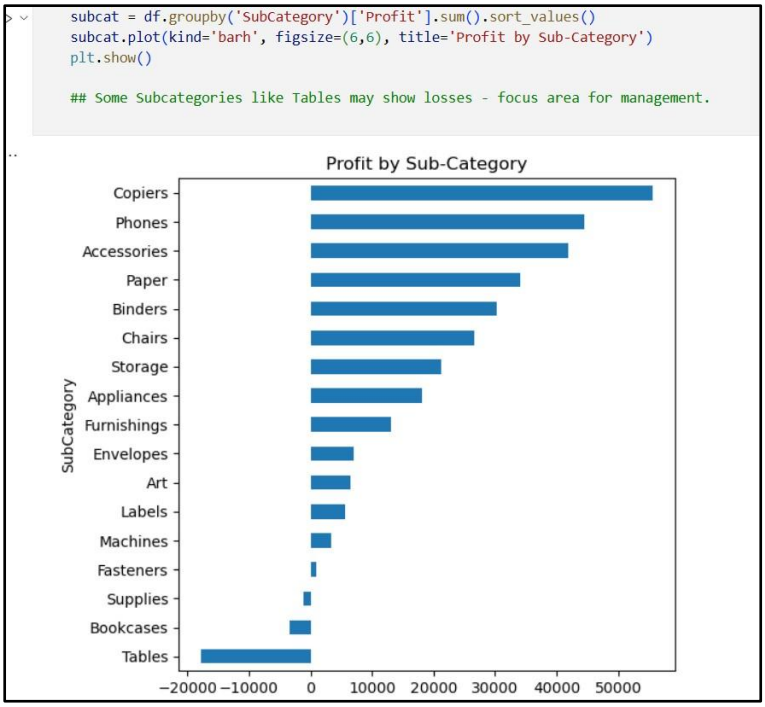
- **Top 10 Customers By Sales:**

```
top_customers = df.groupby('Customer_Name')['Sales'].sum().sort_values(ascending=False).head(10)
top_customers.plot(kind='bar', title='Top 10 Customers by Sales')
plt.show()

## A few customer contribute large share of revenue.
```

A few customers contribute large share of revenue.

- **Profit By Subcategory:**

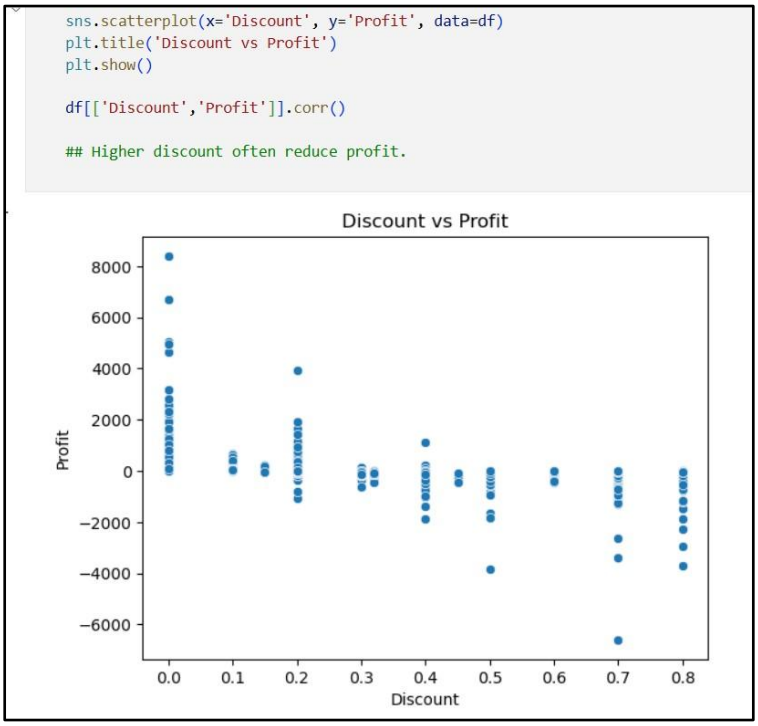
```
subcat = df.groupby('SubCategory')['Profit'].sum().sort_values()
subcat.plot(kind='barh', figsize=(6,6), title='Profit by Sub-Category')
plt.show()

## Some Subcategories like Tables may show losses - focus area for management.
```



Some Subcategories like Tables may show losses - focus area for management**.**

- **Discount vs Profit:**

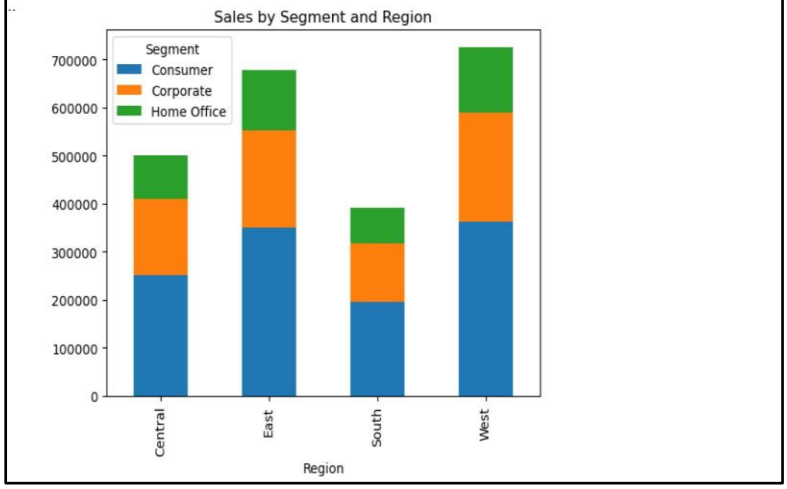
```
sns.scatterplot(x='Discount', y='Profit', data=df)
plt.title('Discount vs Profit')
plt.show()

df[['Discount','Profit']].corr()

## Higher discount often reduce profit.
```



Higher discount often reduce profit.

- **Sales by Segment & Region:**

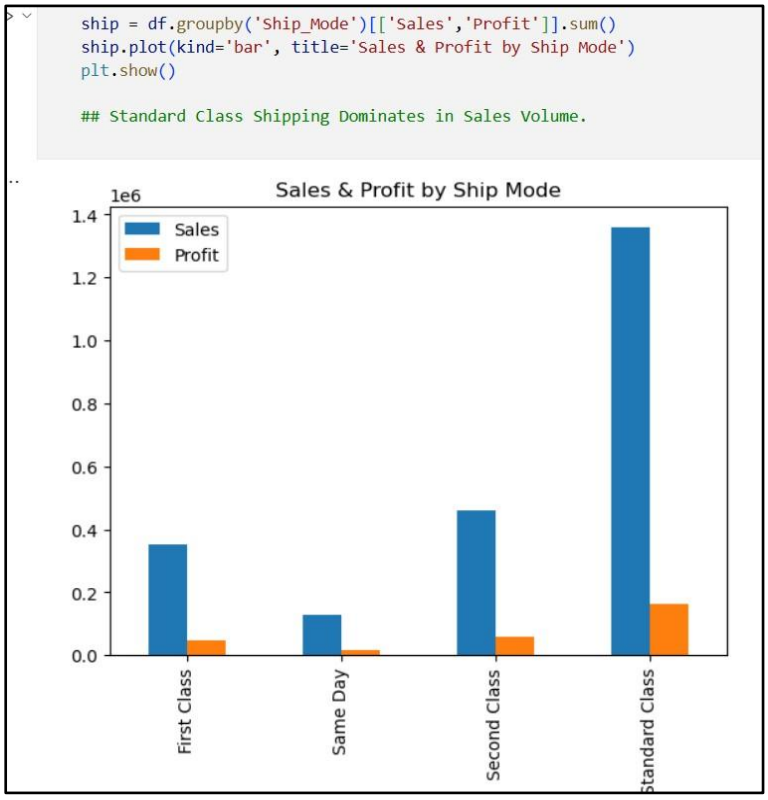
```
pd.crosstab(df['Region'], df['Segment'], values=df['Sales'], aggfunc='sum').plot(kind='bar', stacked=True)
plt.title('Sales by Segment and Region')
plt.show()

## Corporate segment perform best in the west and east regions.
```



Corporate segment perform best in the west and east regions.

- **Sales & Profit By Ship Mode:**

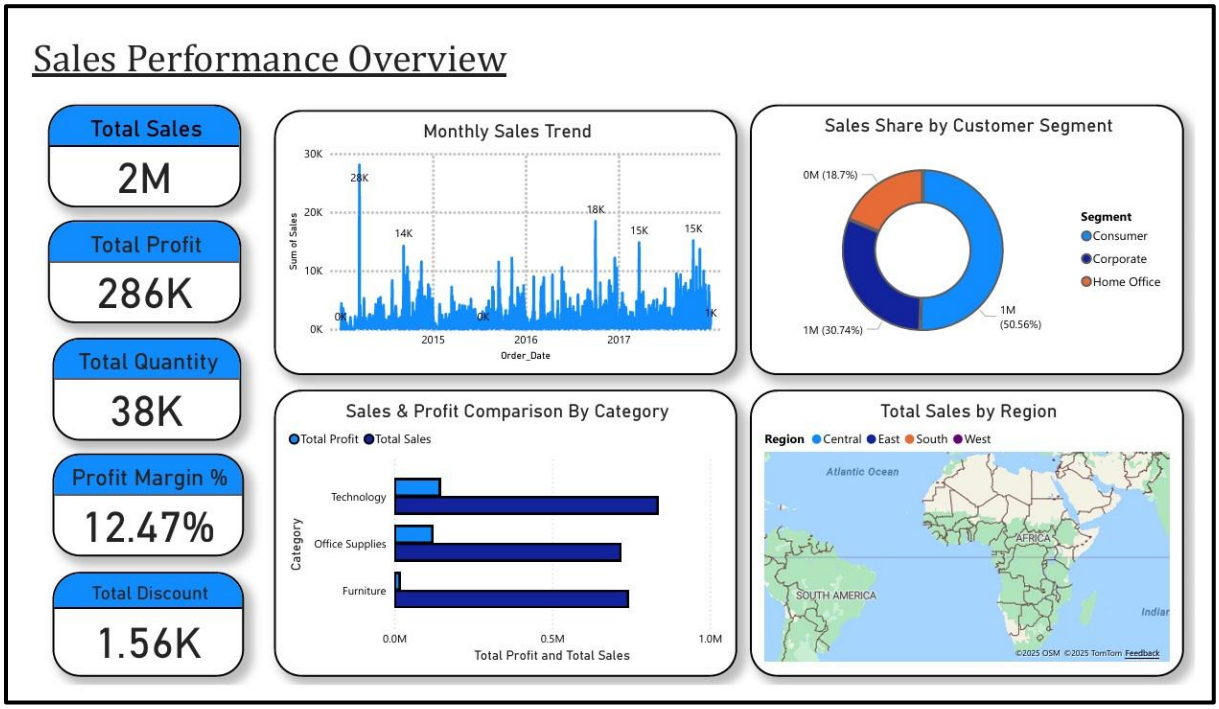
```
ship = df.groupby('Ship_Mode')[['Sales','Profit']].sum()
ship.plot(kind='bar', title='Sales & Profit by Ship Mode')
plt.show()

## Standard Class Shipping Dominates in Sales Volume.
```



Standard Class Shipping Dominates in Sales Volume.

Data Visualization With Power BI:

- **Sales Performance Overview :**

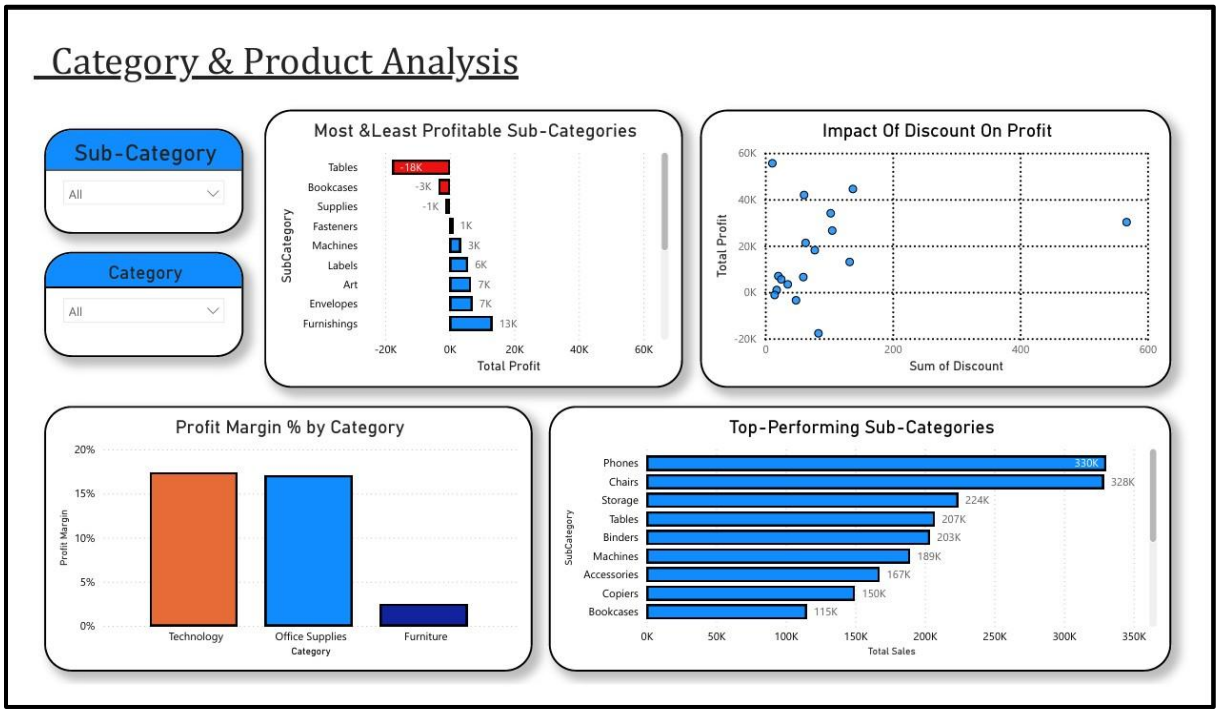


Insights:

1. KPI Cards:
   Total Sales: 2M | Total Profit: 286K| Total Quantity: 38K
   Profit Margin: 12.47%| Total Discount: 1.56K.
2. Monthly Sales Trend:
   Sales show a steady upward trend from 2015 to 2017, with spikes during year-end months, indicating strong seasonal demand likely driven by holiday shopping and promotions.
3. Sales & Profit Comparison by Category:
   The Technology category leads in both sales and profit, followed by Office Supplies.
   Although Furniture has substantial sales, its profit contribution is lower, suggesting higher costs or discounting.

4. Sales Share by Customer Segment:
   The Consumer segment contributes about 50% of total sales, followed by Corporate (31%) and Home Office (19%).
   Targeting the Consumer group with loyalty programs could enhance revenue further.
5. Total Sales by Region:
   The West region dominates in both sales and profitability, followed by the East.
   The Central region shows moderate sales but lower profit, signaling potential issues in pricing or logistics.

- **Category & Product Analysis:**



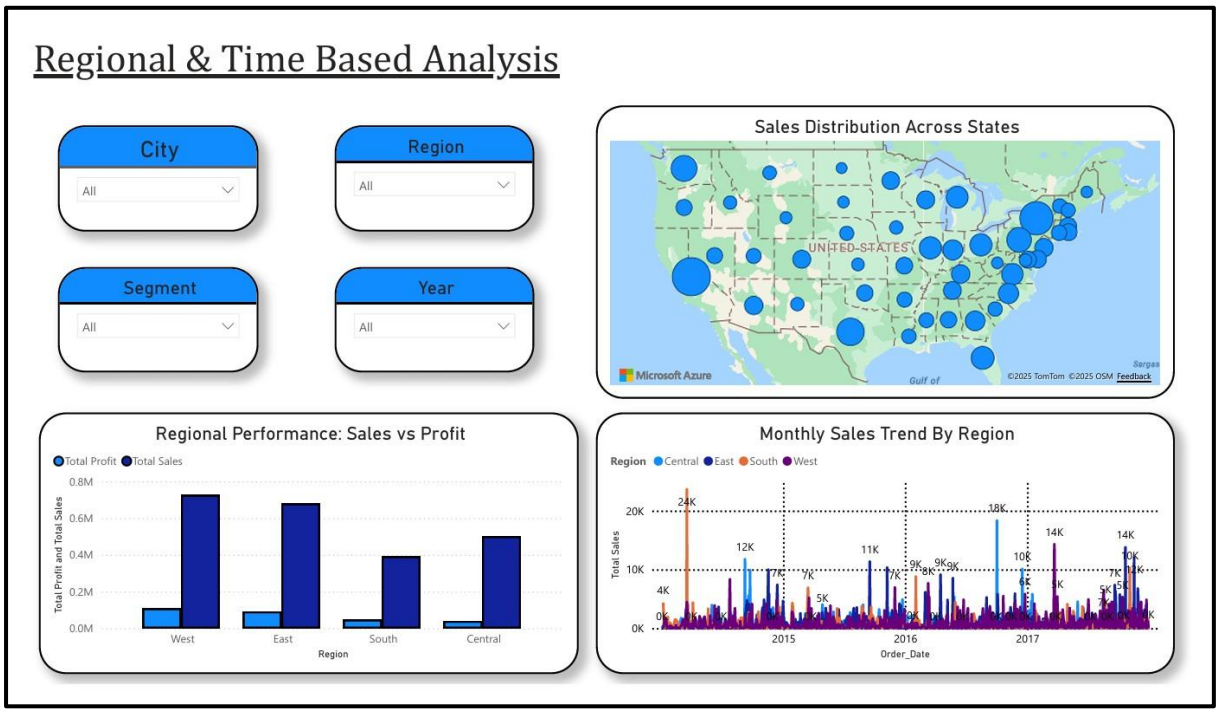
Insights:

1. Top-Performing Sub-Categories:
   The Phones and Chairs sub-categories top the list, generating over $300K each in sales.

Other strong performers include Storage, Tables, and Binders.

2. Most & Least Profitable Sub-Categories:
   Copiers and Phones deliver the highest profits, while Tables and Bookcases show losses.
   Some product lines within Furniture need pricing or cost optimization to improve profitability.
3. Impact of Discount on Profit:
   Higher discounts correlate with reduced profit — confirming that over-discounting negatively affects profitability.
4. Profit Margin % by Category:
   Technology has the highest profit margin (~18%), followed by Office Supplies (13%) and Furniture (7%).
   Furniture, while driving sales volume, contributes the least profit margin.

- **Regional & Time Based Analysis:**



Insights:

1. Sales Distribution Across States:
   States like California, New York, and Texas record the highest sales volumes.
   Western and Eastern states outperform Southern and Central ones overall.
2. Monthly Sales Trend By Region:
   The West region consistently leads sales across all months, showing strong, sustained performance. The East region follows similar trends but with slightly lower volumes.
   The Central and South regions remain comparatively stable with minimal growth over time**.**
3. Regional Performance — Sales vs Profit:
   The West region shows the highest total sales and profit, indicating it's the best-performing market.
   The Central region has lower profit despite decent sales, suggesting high operational costs or lower pricing efficiency.

Recommendation:

1. Refine Discount Strategy:
   High discounts, especially in Furniture and Office Supplies, hurt profit margins. Implement a tiered discount policy to balance sales growth and profitability.

2. Leverage Strong Segments and Regions:
   The Consumer segment (50%+ of sales) and the West region drive top performance. Focus on retention here, while improving Corporate/Home Office sales and strengthening presence in weaker regions like Central.

3. Optimize Product Mix:
   Sub-categories such as Phones and Chairs perform well, while Tables and Bookcases underperform. Reassess pricing, promotions, and inventory for lowprofit items.

4. Use Data for Inventory Planning:
   Stable monthly sales trends can guide better demand forecasting and inventory control, reducing overstock costs.

Conclusion:

This Sales Insight Analysis project applied an end-toend data analytics process—from Data Processing and Cleaning in Excel, SQL-based validation and aggregation, through EDA in Python, to Power BI visualization.

The project uncovered key business insights:

1. Total sales of approximately $2 million with a 12.47% profit margin.
2. The Technology category emerged as the most profitable, while Furniture faced margin challenges.

3. The Consumer segment dominated overall sales contribution.
4. The West region consistently outperformed others in both sales and profit metrics.
   The analysis underscores how data-driven insights enable smarter pricing, marketing, and inventory strategies, ultimately improving overall business performance.

Project Repository:

You can explore the complete project files, including datasets, SQL queries, EDA notebooks, and Power BI dashboard, at the following link:
https://github.com/vedantdarvekar/Sales-Performance-Analysis.git