NCPL SALES

Power BI Project: Sales Data Analysis with Predictive Modeling

Table Of Content		
1	Introduction	2
2	Business Questions	2
3	Dataset Overview	3
4	Data Preprocessing	6
5	Data Cleaning and Transformation	8
7	Business Analysis	9
8	Predictive analysis using Python	18
9	Dashboard	20
10	Final Analysis and Conclusion	22

Introduction

Introduction

A comprehensive **Power BI project for sales analysis and forecasting**, aimed at uncovering key business insights and providing actionable recommendations to support decision-making. The analysis focused on understanding overall sales performance, identifying trends and fluctuations over time, and examining the contributions of different product categories to highlight top-performing and underperforming areas. It also explored geographic insights to determine which regions show strong sales potential and which may require targeted marketing efforts. Additionally, the project assessed profitability, analyzing which products were most profitable, factors affecting overall profit, and patterns in the company's financial performance during the period. Using **Python (Pandas and Statsmodel)**, predictive models were developed to forecast future sales by leveraging historical trends, seasonality, promotional impacts, and other derived features, supporting inventory planning, resource allocation, and strategic decisions. By combining **interactive Power BI dashboards** with predictive analytics, stakeholders can explore historical data, track future forecasts, and gain a clear understanding of sales dynamics, enabling data-driven actions that enhance overall business performance and operational efficiency.

Business Questions

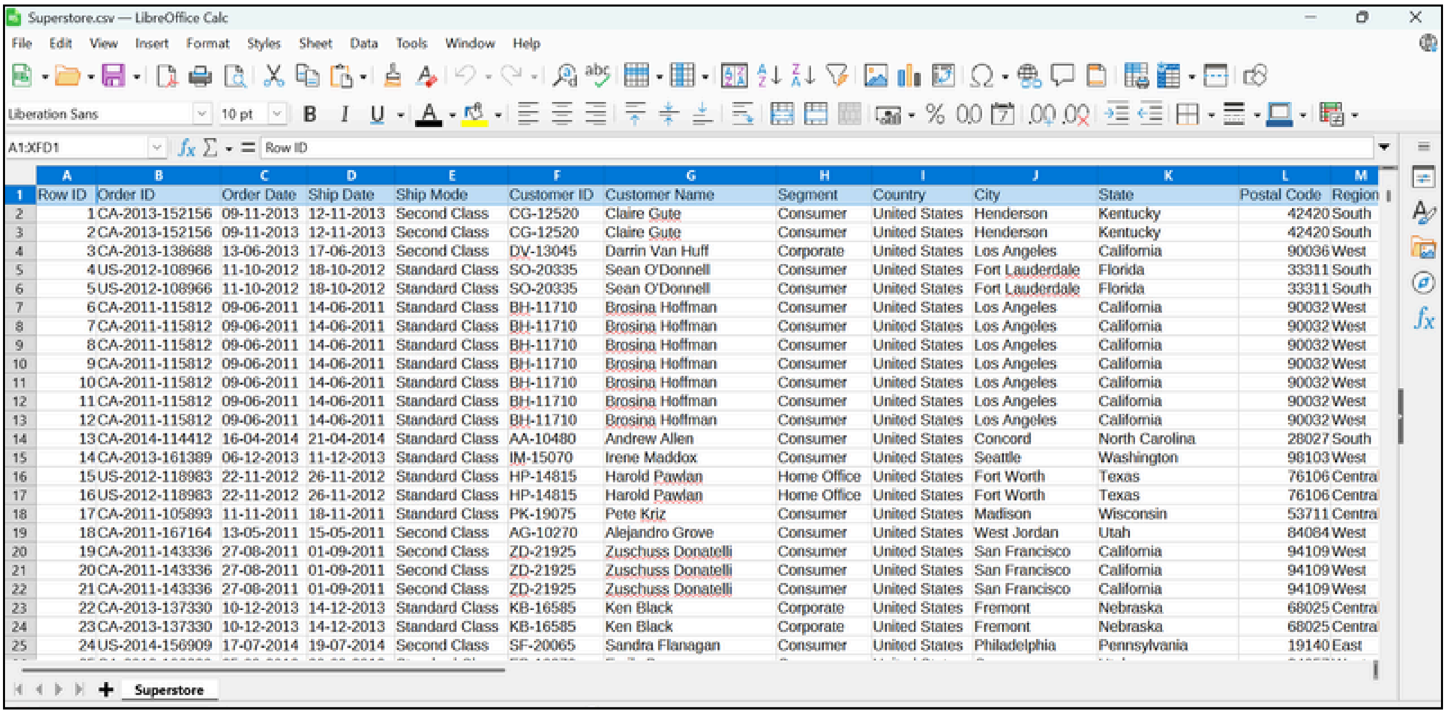
This analysis aims to address the following key business questions:

1. **Sales Performance**: What are the overall sales trends, and how have they evolved over time? Are there any significant fluctuations that need to be addressed?
2. **Product Categories**: Which product categories contributed the most to the company's sales? Which categories are underperforming, if any?
3. **Geographic Insights**: How does sales performance vary across the regions? Are there promising geographical regions or areas requiring improved marketing?
4. **Profitability**: Which products are more profitable and which were not? With the available data, what factors affected the company's profit? How is the company's profitability during the period?

This analysis also aims to discover other valuable insights about the dataset. Ultimately, this analysis intends to provide actionable insights to guide decision-making and enhance overall business performance.

Dataset Overview (Metadata)

Dataset has 9994 **rows × 21 columns**



Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region
1	CA-2013-152156	09-11-2013	12-11-2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South
2	CA-2013-152156	09-11-2013	12-11-2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South
3	CA-2013-138688	13-06-2013	17-06-2013	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West
4	US-2012-108966	11-10-2012	18-10-2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
5	US-2012-108966	11-10-2012	18-10-2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
6	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
7	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
8	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
9	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
10	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
11	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
12	CA-2011-115812	09-06-2011	14-06-2011	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West
13	CA-2014-114412	16-04-2014	21-04-2014	Standard Class	AA-10480	Andrew Allen	Consumer	United States	Concord	North Carolina	28027	South
14	CA-2013-161389	06-12-2013	11-12-2013	Standard Class	IM-15070	Irene Maddox	Consumer	United States	Seattle	Washington	98103	West
15	US-2012-118983	22-11-2012	26-11-2012	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texas	76106	Centra
16	US-2012-118983	22-11-2012	26-11-2012	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texas	76106	Centra
17	CA-2011-105893	11-11-2011	18-11-2011	Standard Class	PK-19075	Pete Kriz	Consumer	United States	Madison	Wisconsin	53711	Centra
18	CA-2011-167164	13-05-2011	15-05-2011	Second Class	AG-10270	Alejandro Grove	Consumer	United States	West Jordan	Utah	84084	West
19	CA-2011-143336	27-08-2011	01-09-2011	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West
20	CA-2011-143336	27-08-2011	01-09-2011	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West
21	CA-2011-143336	27-08-2011	01-09-2011	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West
22	CA-2013-137330	10-12-2013	14-12-2013	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Centra
23	CA-2013-137330	10-12-2013	14-12-2013	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Centra
24	US-2014-156909	17-07-2014	19-07-2014	Second Class	SF-20065	Sandra Flanagan	Consumer	United States	Philadelphia	Pennsylvania	19140	East

Column Details:

Column Name	Description
Row ID	Unique identifier for each record/row.
Order ID	Unique identifier for each order (multiple rows may belong to one order).
Order Date	The date when the order was placed.
Ship Date	The date when the order was shipped.
Ship Mode	Shipping method chosen (e.g., Standard, Second Class).
Customer ID	Unique identifier for each customer.
Customer Name	Name of the customer.
Segment	Customer segment (Consumer, Corporate, Home Office).
Country	Country of the customer (mainly USA).
City	Customer’s city.
State	Customer’s state.
Postal Code	Zip/postal code of the customer’s address.
Region	Geographic region (Central, East, South, West).
Product ID	Unique identifier for each product.
Category	High-level product category (Furniture, Office Supplies, Technology).
Sub-Category	More specific product classification (Chairs, Phones, Binders, etc.).
Product Name	Full product description/name.

Sales	Total sales revenue for the product/order line.
Quantity	Number of units sold.
Discount	Discount applied on the product.
Profit	Profit earned from the product/order line.

Important Columns by Analysis Area

1. **Sales Performance**

- **Order Date** → Trend analysis over time.
- **Sales** → Measuring revenue growth/fluctuations.
- **Quantity** → Supporting sales volume analysis.

2. **Product Categories**

- **Category** → Major sales contribution.
- **Sub-Category** → Detailed product-level performance.
- **Sales, Profit** → Identify top-performing vs underperforming products.

3. **Geographic Insights**

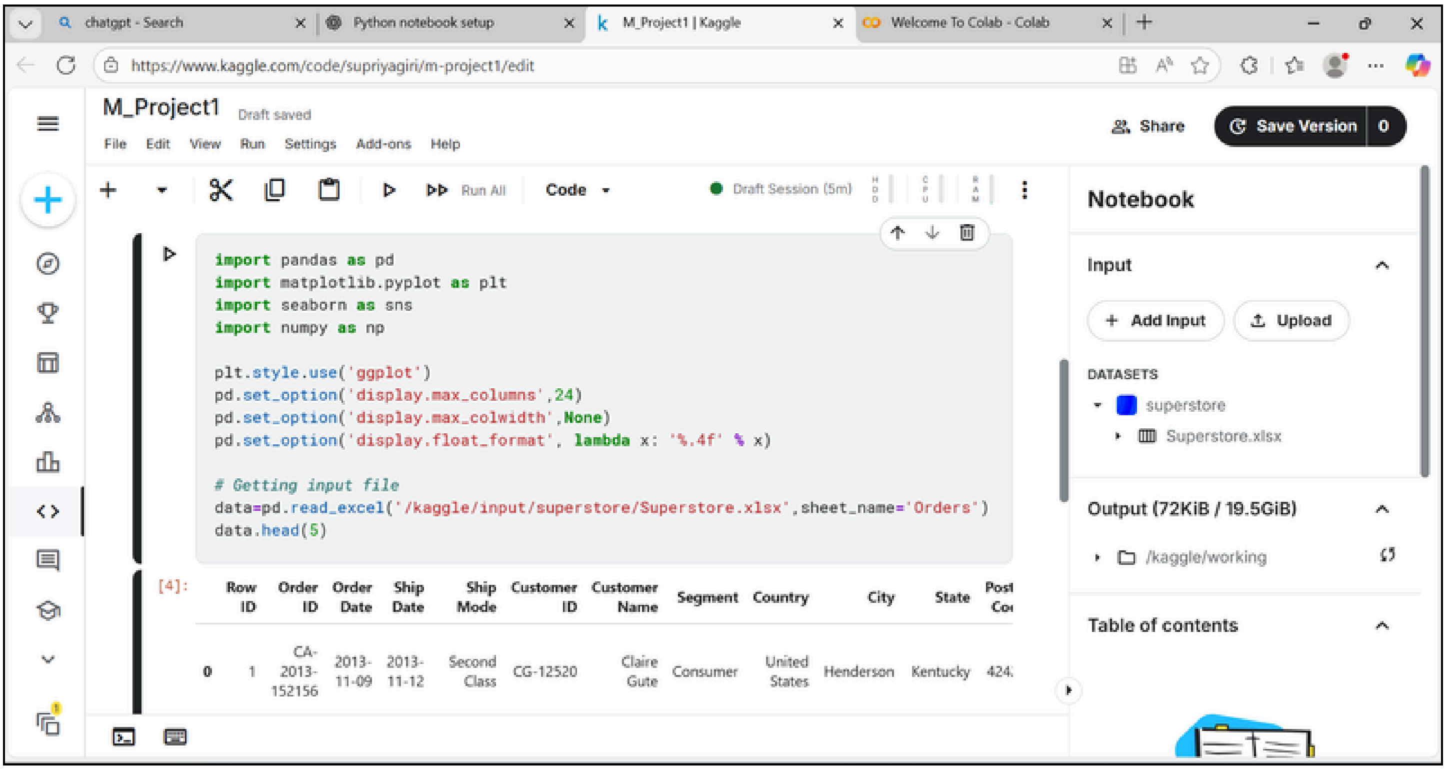
- **Region, State, City** → Regional sales & profitability analysis.
- **Sales, Profit** → Geographic contribution and performance variations.

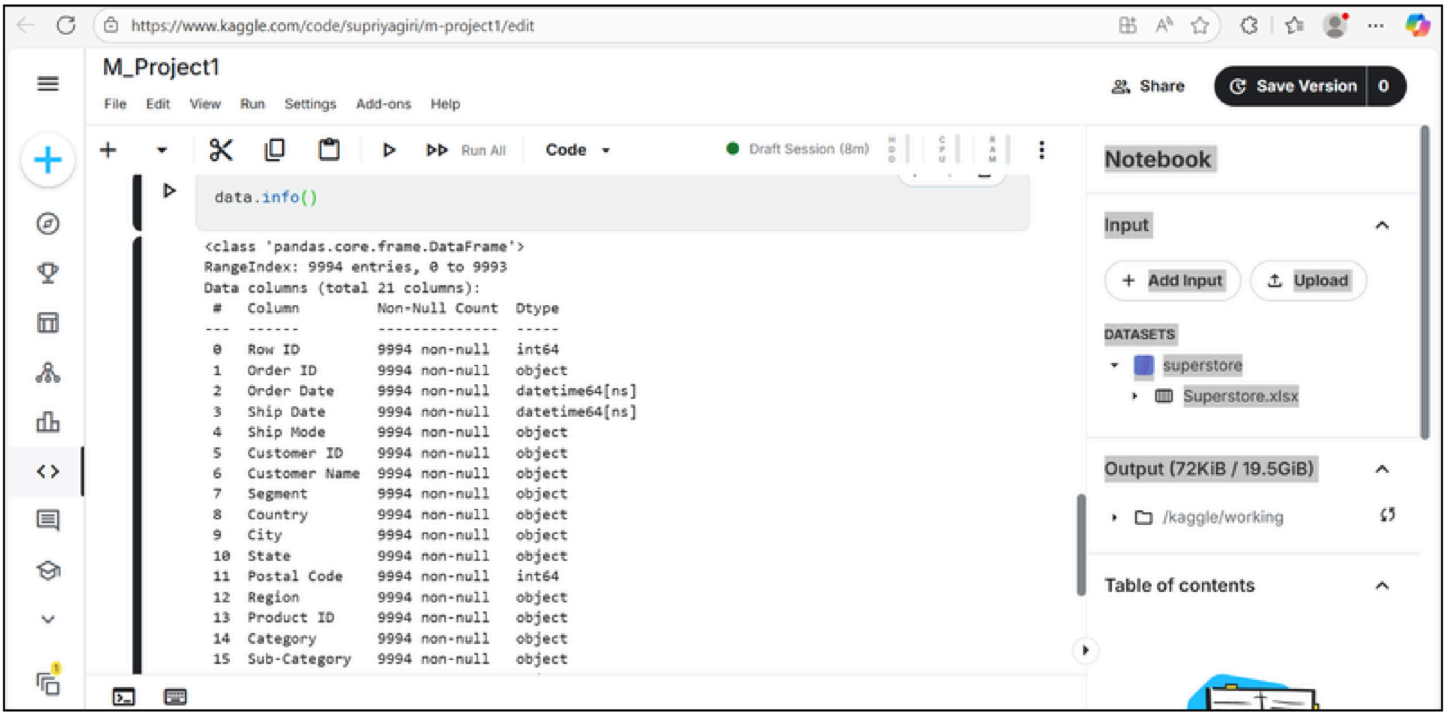
4. **Profitability**

- **Profit** → Direct profitability measure.
- **Discount** → Understand discount impact on profits.
- **Sales, Quantity** → Identify profit margins relative to sales volume.
- **Category, Sub-Category, Product Name** → Pinpoint profitable vs loss-making products.

Step 1: Data Preprocessing

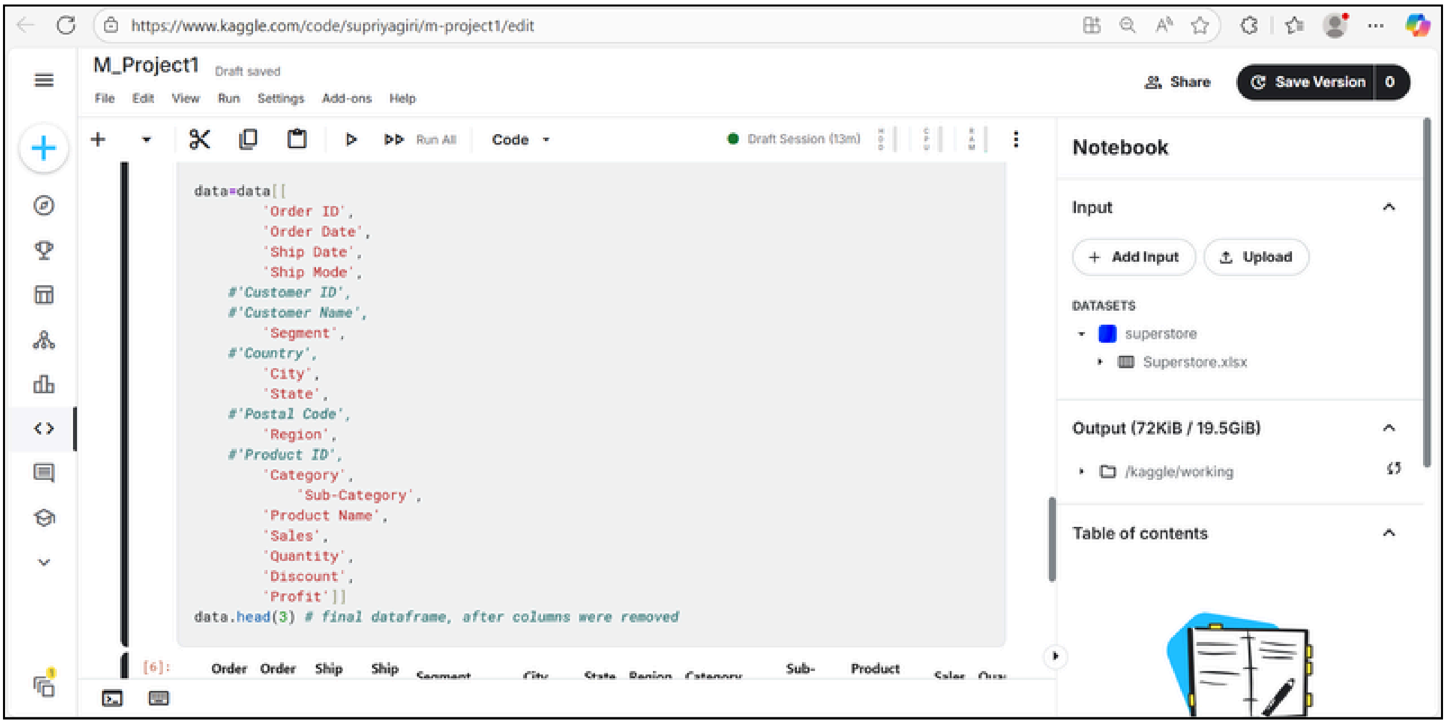
- With the overview, the dataset will not need a lot of data cleaning. However, there are certain transformations that needs to be done to ready the data for the analysis. Specifically, the Row ID column and some others are not necessary for this particular analysis and will be removed. Feature engineering will also be done.
- The Data preprocessing is done in Python
- Importing relevant libraries:

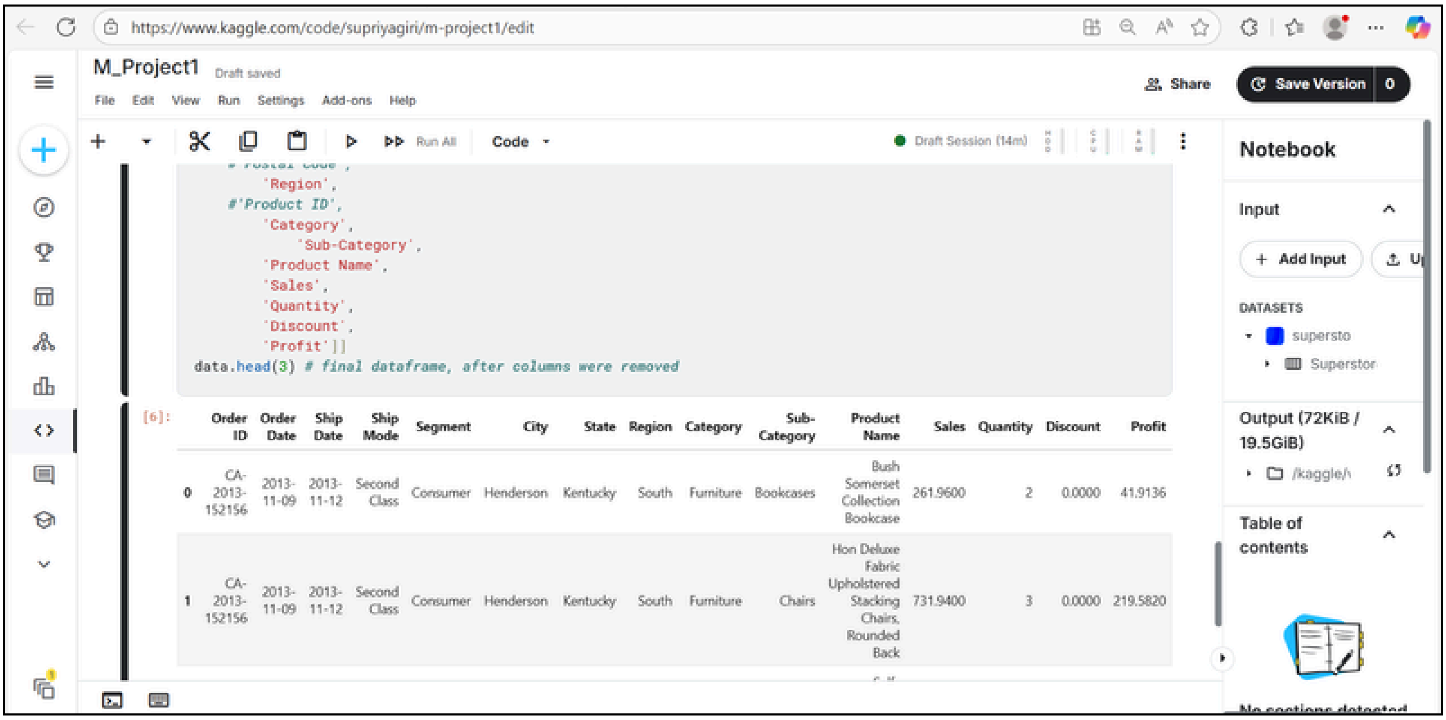




The dataset also has no missing values. (see Non-Null Count)

We can drop Some of the Columns which are not needed in our analysis.





Now in feature Engineering, new columns will be created with the help of existing columns. e.g. year and month are extracted from date column. Saling Price calculated by dividing total quantity by Total Sales

Typewriters
by Universal

```
# feature engineering, extracts specific date values from the `Order Date` column, creates new features from existing
data['month']=data['Order Date'].dt.month
data['year']=data['Order Date'].dt.year
data['year_month']=data['Order Date'].dt.to_period('M')
data['total_discount_in_dollars']=data['Sales'] * data['Discount'] # discount's equivalent to dollars
data['selling_price']=data['Sales'] / data['Quantity'] # calculates selling price for the each product
data['(net)_profit_before_discount']=data['Sales'] * data['Discount'] + data['Profit'] # net profit before deducting
data['order_fulfillment_time']=data['Ship Date'] - data['Order Date'] # interval between order placed and order ship
data['net_profit_per_unit_sold']=data['Profit'] / data['Quantity'] # net profit generated per unit sold
data=data.rename(columns={'Profit':'net_profit'}) # renames Profit column with net_profit, a more specific name
data['profit_margin']=data['net_profit'] / data['Sales'] * 100 # for a 25% profit margin, the company makes .25 doll
data['discounted_sales']=data['Sales'] - (data['Discount']*data['Sales']) # extracts sales accounted for discount
```

+ Code + Markdown

Output:

```
[10]: print('Output dataframe:')
      data.head(5)
```

Output dataframe:

[10]:

	Order ID	Order Date	Ship Date	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Product Name	Sales	...	Discount	net_profit	mo
0	CA-2013-152156	2013-11-09	2013-11-12	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	...	0.0000	41.9136	
1	CA-2013-152156	2013-11-09	2013-11-12	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.9400	...	0.0000	219.5820	
2	CA-2013-138688	2013-06-13	2013-06-17	Second Class	Corporate	Los Angeles	California	West	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.6200	...	0.0000	6.8714	

The transformed dataset now contains 9994 rows and 22 columns.

With this, the data is now ready for analysis. Data cleaning and transformations are always done to almost all real-world datasets. This includes handling for missing values, casting data to appropriate data types, standardizing or normalizing values, feature engineering, and date time and string types transformations, among others.

https://www.kaggle.com/code/supriyagiri/m-project1/edit

M_Project1 Draft saved

```
[12]: print('Annual total sales: ')
      data.groupby('year')['Sales'].sum()
```

```
Annual total sales:
[12]: year
2011    484247.4981
2012    470532.5090
2013    608473.8300
2014    733947.0232
Name: Sales, dtype: float64
```

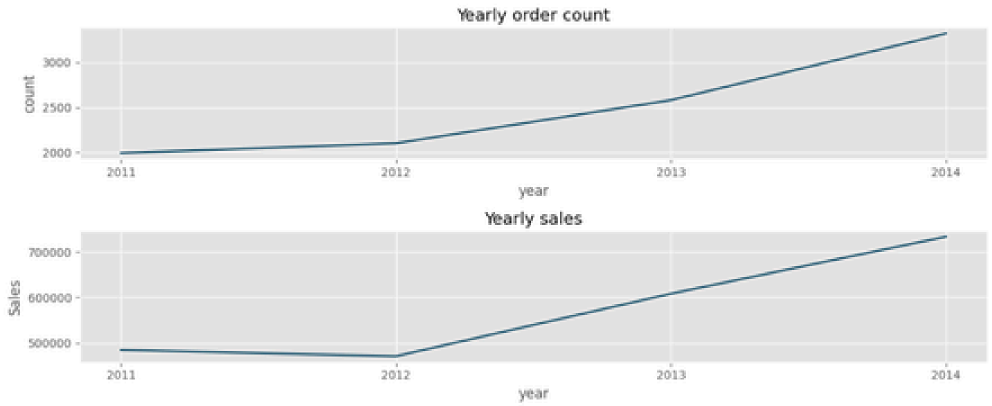
```
# Save as CSV
data.to_csv("/kaggle/working/Superstore_cleaned.csv", index=False)
```

```
plt.figure(figsize=(12,5))   # (width = 12, height = 5).

plt.subplot(211)             # make a grid of 2 rows × 1 column, and activate the 1st plot
data.groupby(['year'])['Order Date'].count().plot(c='#003f5c')
plt.ylabel('count')          # showing number of orders.
plt.xticks(data.groupby(['year'])['Order Date'].count().index)
plt.title('Yearly order count')

plt.subplot(212)
data.groupby('year')['Sales'].sum().plot(c='#003f5c')
plt.ylabel('Sales')
plt.xticks(data.groupby('year')['Sales'].sum().index)
plt.title('Yearly sales')

plt.tight_layout()
plt.show()
```

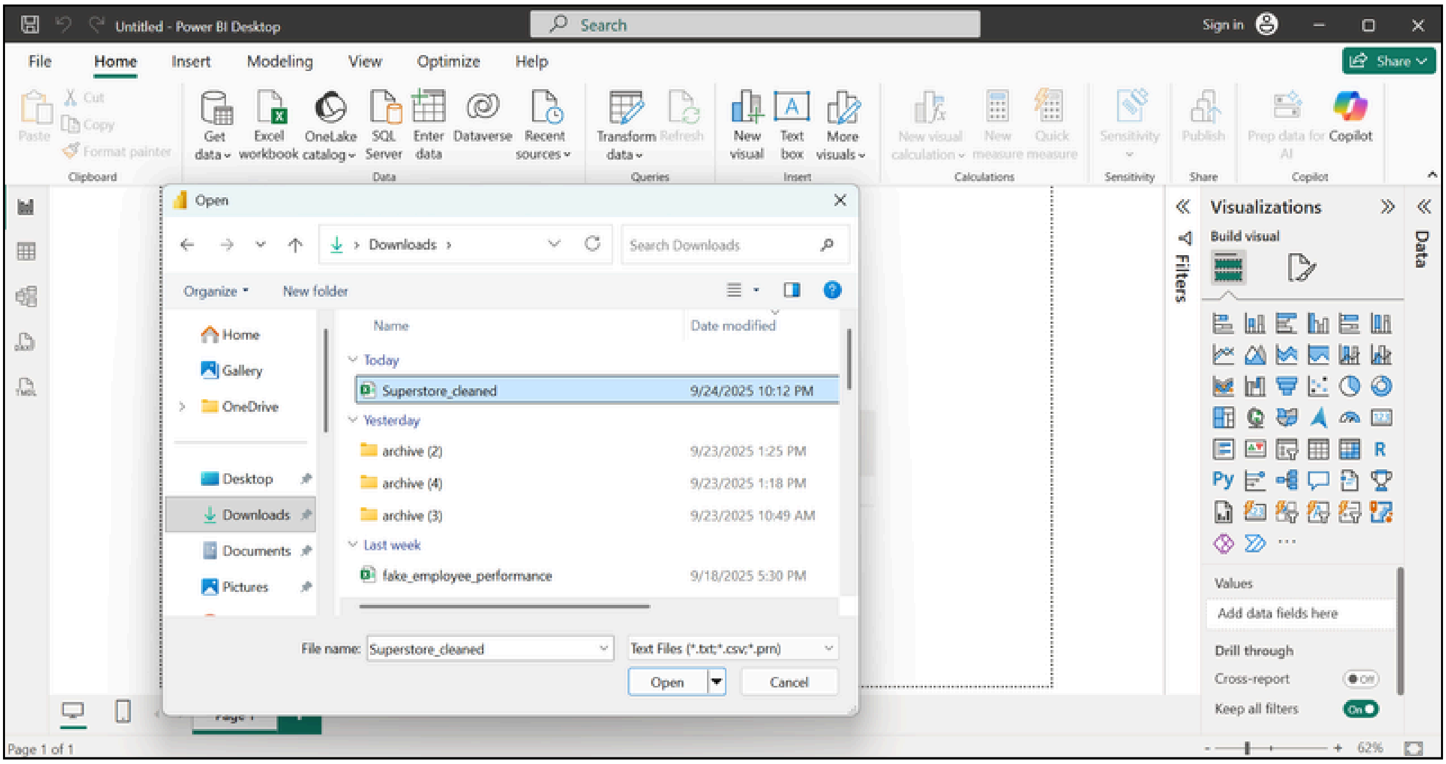


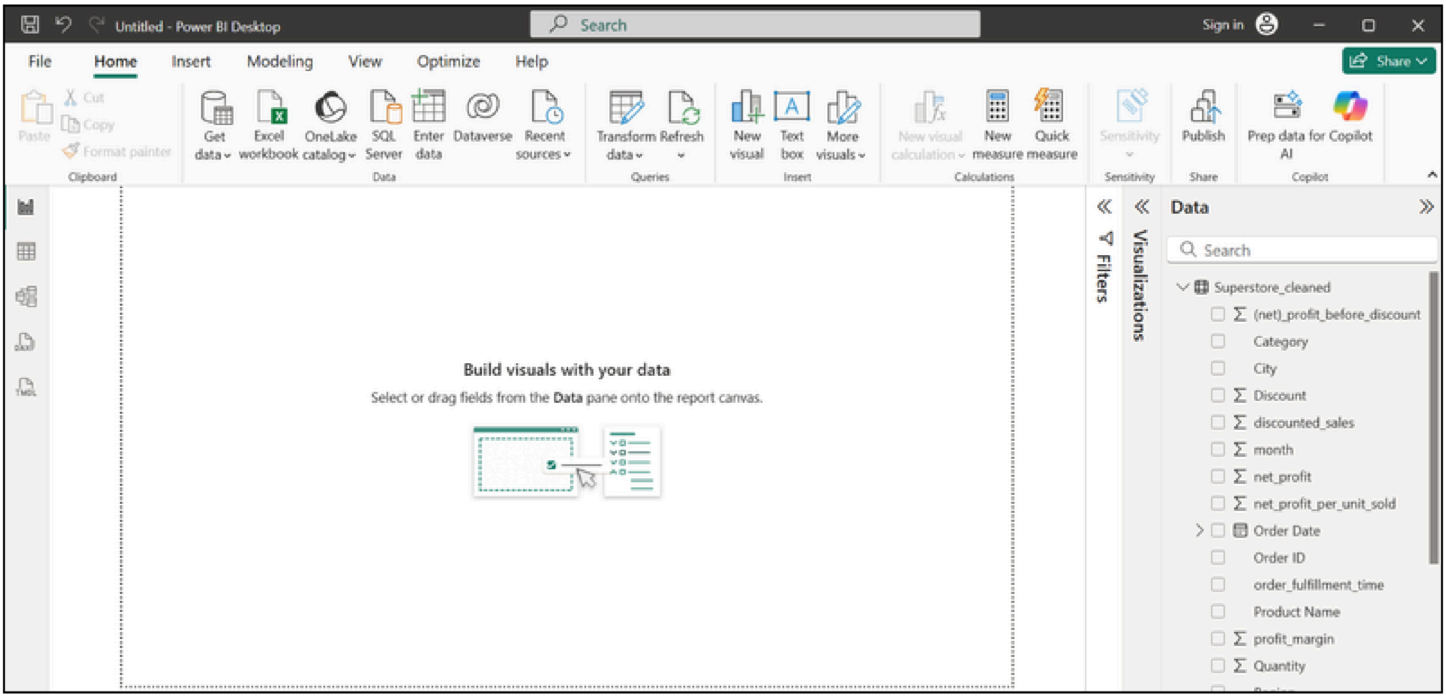

```
print('Annual total sales: ')
data.groupby('year')['Sales'].sum()
```

```
Annual total sales:
[12]: year
2011    484247.4981
2012    470532.5090
2013    608473.8300
2014    733947.0232
Name: Sales, dtype: float64
```

Business Analysis

For visualizations we need to import the cleaned and transformed file into PowerBI

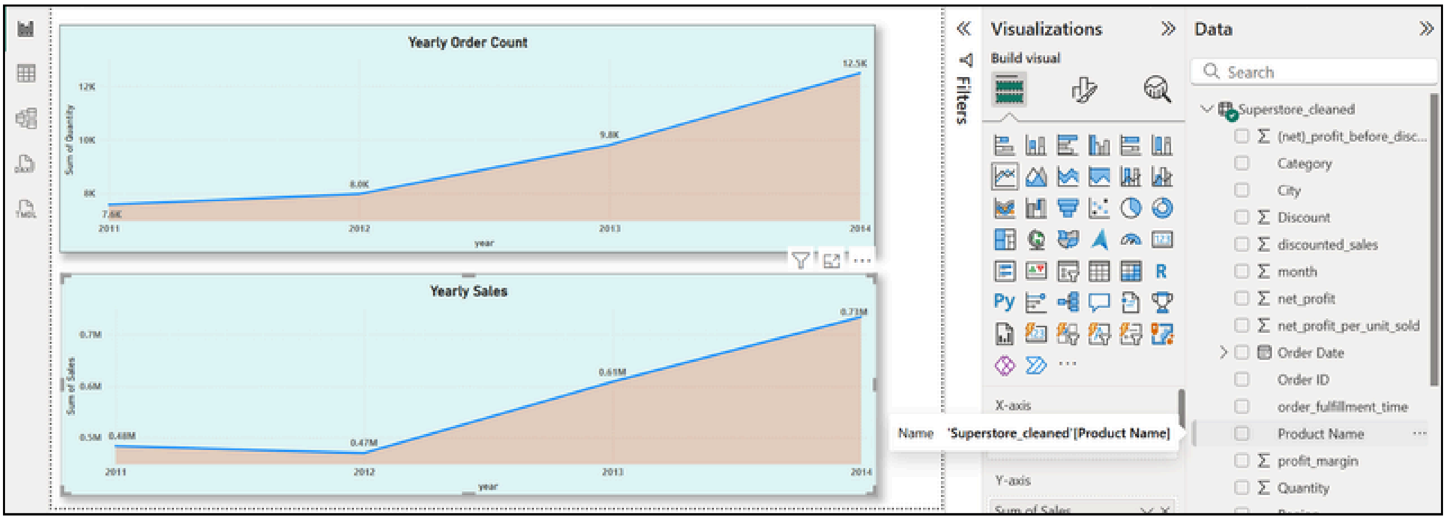




Business Question 1: **Sales Performance**:

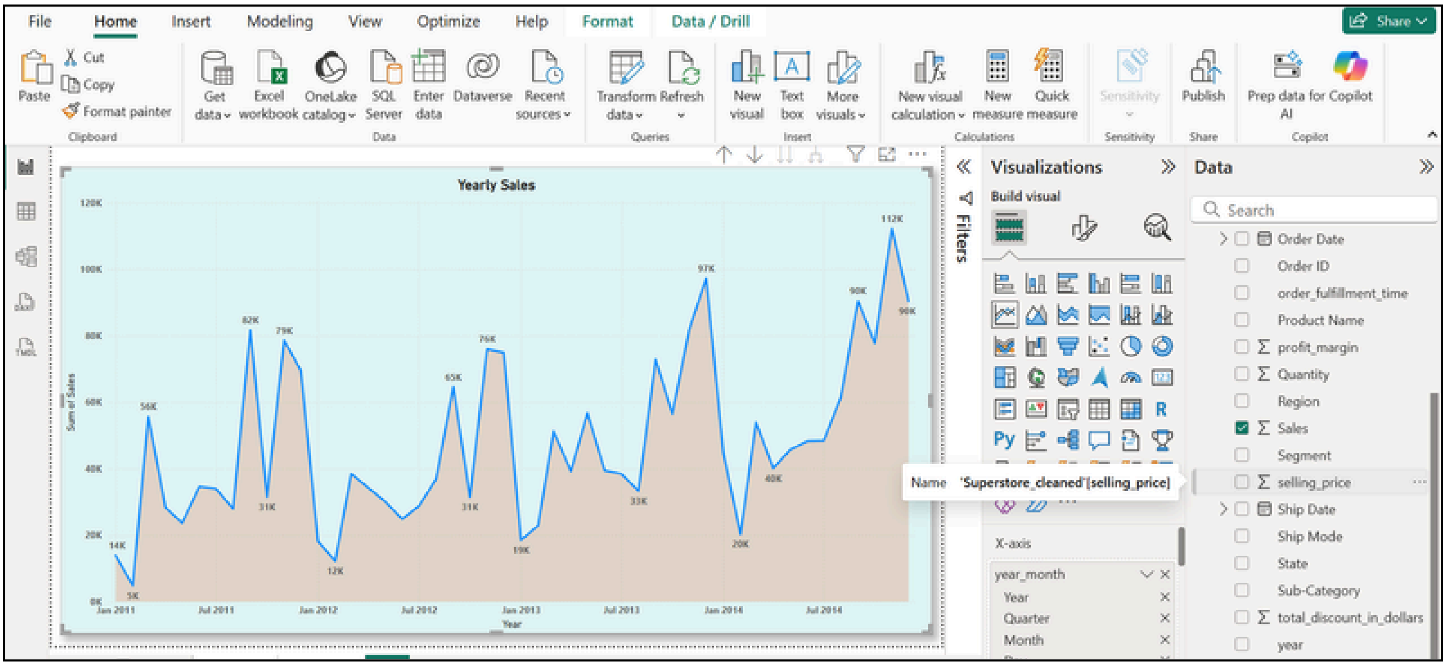
Here we want to see yearly Sales, Monthly Sales, Sales Trend.

Yearly Sales Pattern:



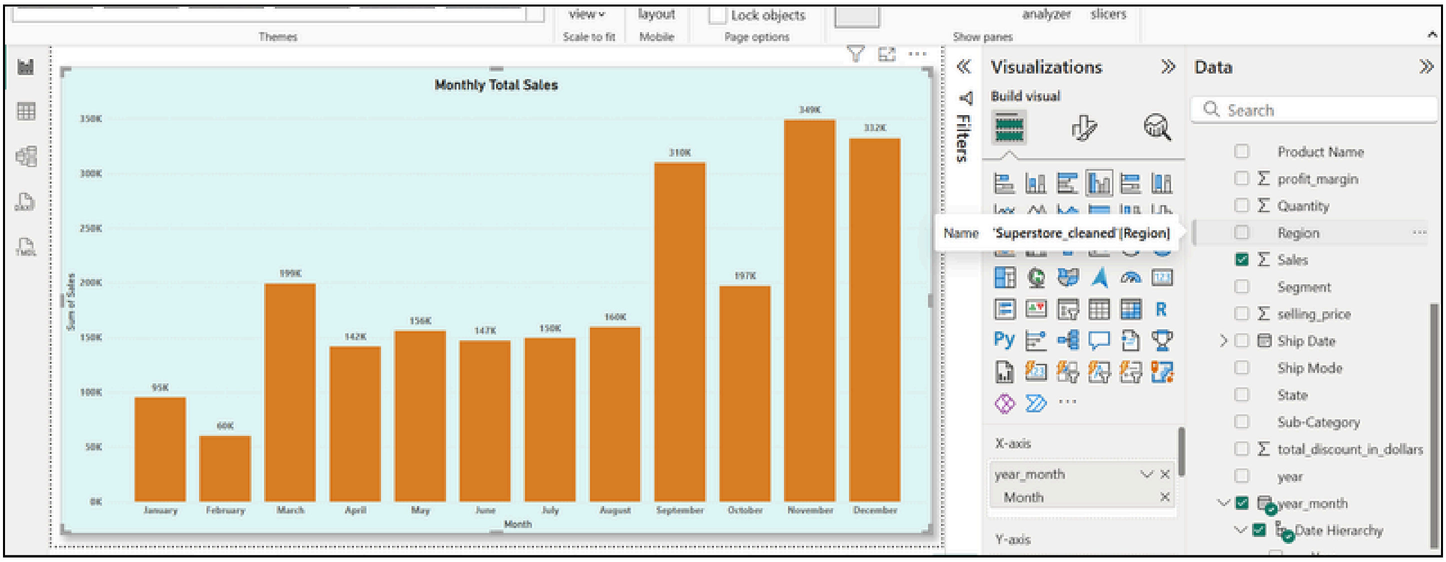
Insight: Over time, orders had increased and so are sales. However, a slight dip in sales can be observed in 2012. From 484,247 dollars total sales in 2011, Superstore sales slightly dipped to 470,532 dollars in the following year, which is a 2.83% difference or 13,715 dollars.

Monthly Sales Pattern throughout all years:



Insight: Seasonal trends occurred. Superstore sales increase towards the end of the year starting in November and is sustained until December and then drops in January. Between February and March each year, sales rise again. From April to August, a generally stable trend is evident every year. Furthermore, a sharp downward trend is observed during October.

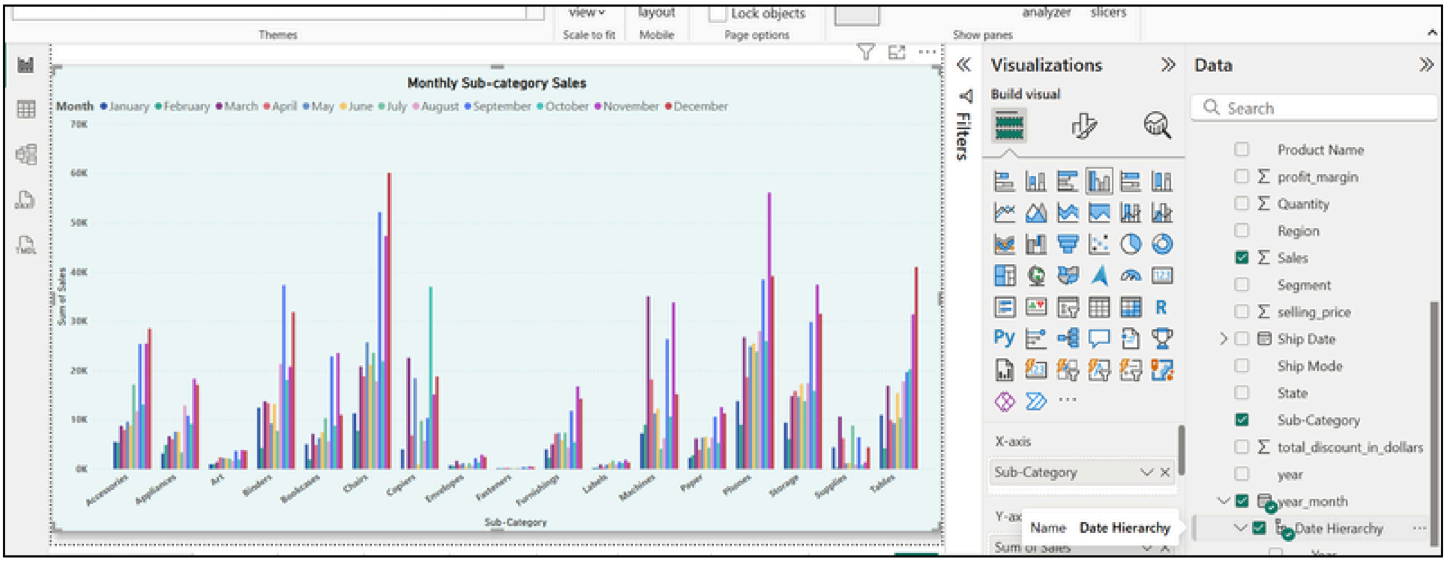
Monthly Sales Pattern:



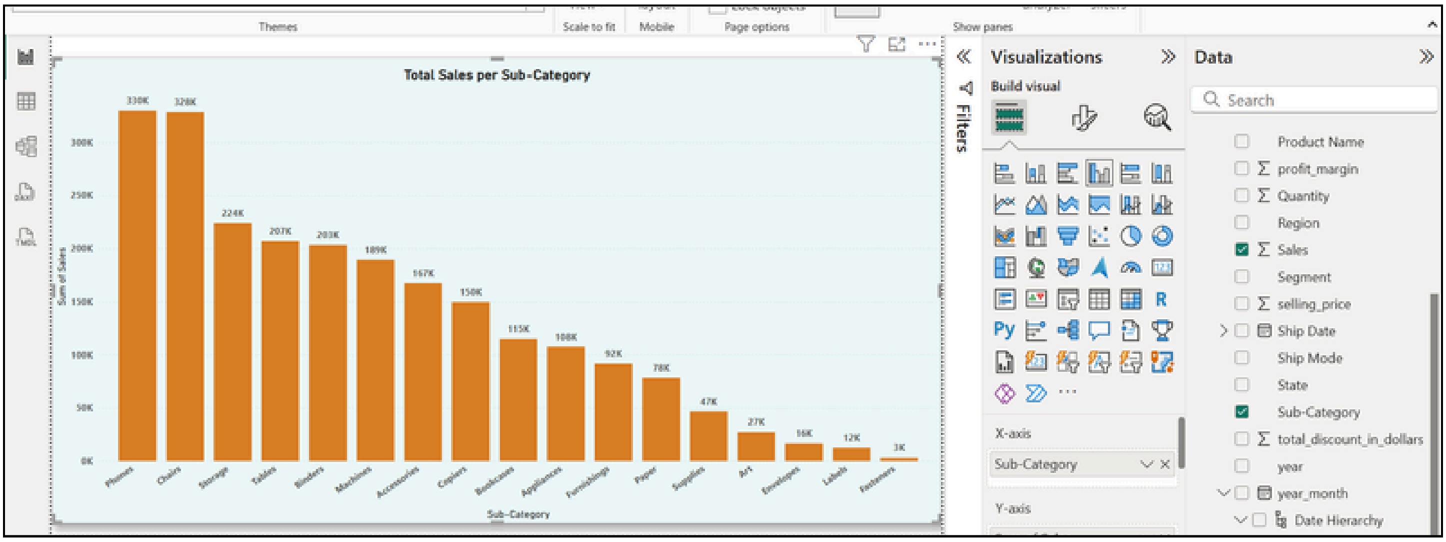
Insight: The visualization above shows total sales for each month over the course of 4 years. By magnitude, sales are higher towards the holiday seasons. Also, the academic year (opening of schools) in America usually starts in late August or early September, which can possibly explain higher sales in September of school-related products such as binders, home and office supplies, papers, bookcases, and accessories, among others (see graph below).

Business Question 2: **Sales Product Categories**:

Sub-Category Sales across the Months Pattern:



Total Sales of Subcategories over all 4 years:

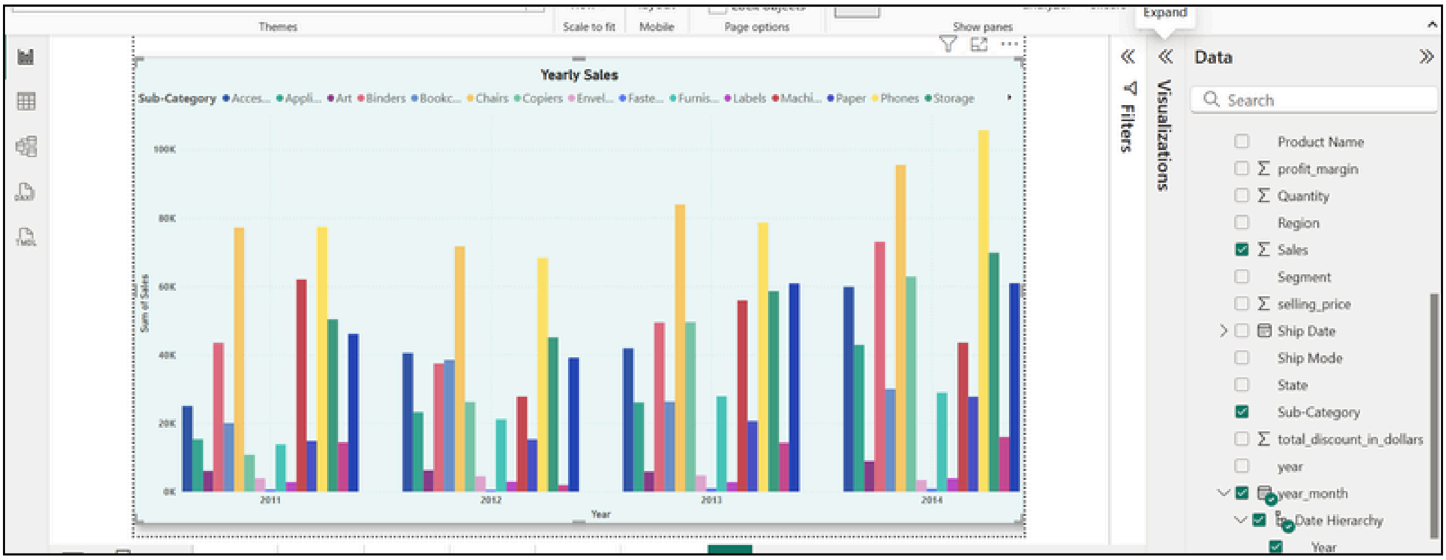


Insight: The visualization above shows a general overview of the magnitude of sales for each product sub-category. For technology category, phones are the top sales-generating products. Chairs products for Furnitures category, and storage products for office supplies category. Throughout the 4-year period from 2011-2014, phones, chairs, and storage products are the three most sales-generating products. Along with them are tables, binders, and machine products.

Under the technology category, copier products are the least performing. For the furniture and office supplies category, furnishings and fasteners are the least performing.

It is worth noting that phones and chairs products sales, which are significantly higher than the rest of the sub-categories, belong to Technology and Furniture product categories, products that are generally expensive.

Subcategories Performance over 4 years:

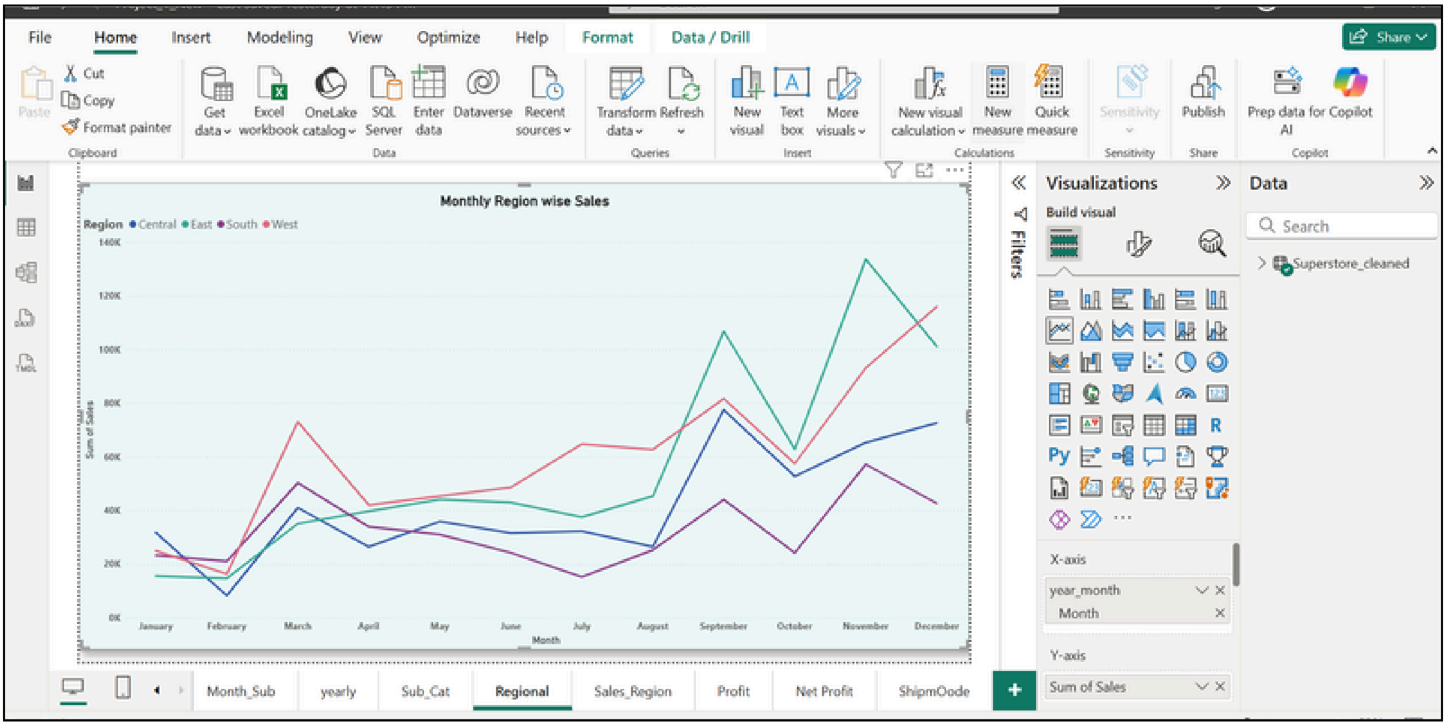


Insight: Shown is how sales on different products had changed over the 4-year period. For some product categories, sales had been fastest growing in 2014. This was not the case for bookcases, machines, supplies, and tables, which all saw a slow growth in sales in the same year. In 2012, products under binders, phones, storages, supplies, and tables experienced negative growth in sales, especially machine products.

By average sales annual growth rate, envelope products had been the slowest while supplies products had been the fastest followed by copier and appliances products.

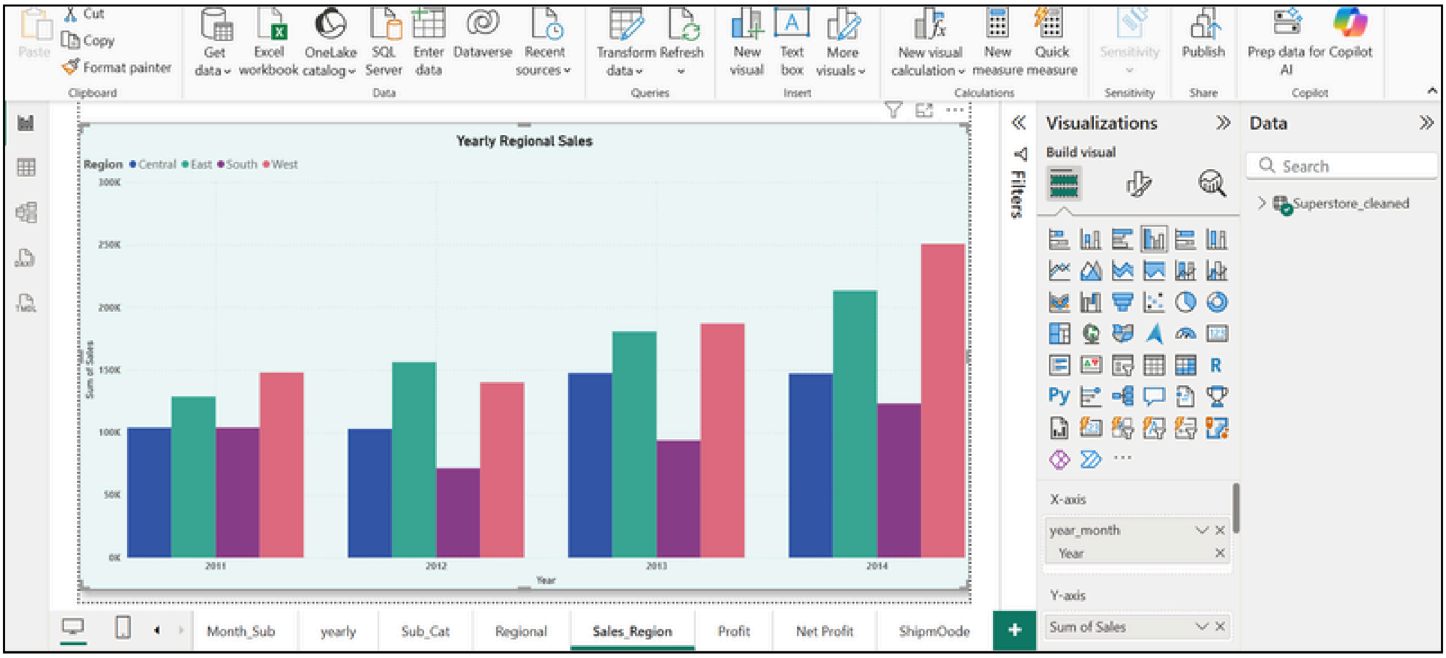
Business Question 3: **Sales Geographic**

Monthly Region wise Sales Trends:



Insight: As shown before, seasonal trend occurred, with sales increasing in holidays (November and December), opening of classes (September), and possibly Easter (March). This is also the case with regional sales data per year. Total sales have been generally higher most of the year in the West, followed by the East compared to the remaining two regions. Sales in the South has been lowering each year compared to other regions with the exception in March 2011 when sales in the South were more than thrice the next best performer.

Yearly Region wise Sales:

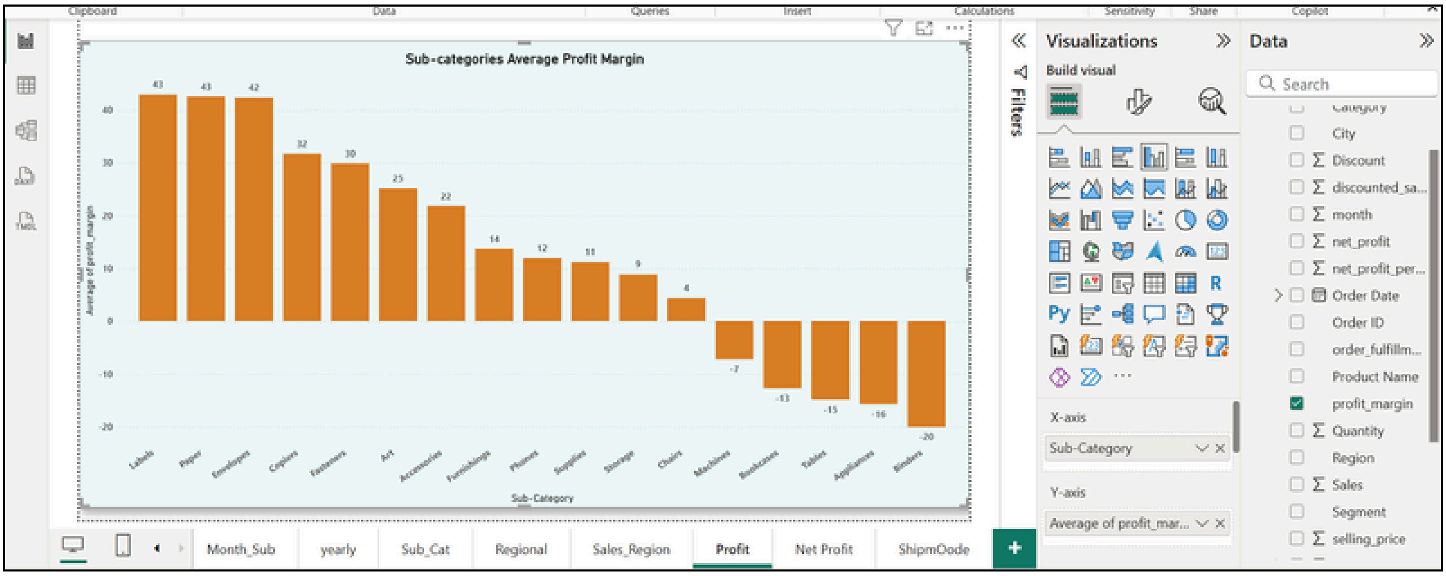


Insight: Shown is how sales for each region changed over time. In the Central region, a sharp positive growth was observed from 2013, while yearly positive growth was consistent in the East. For the South region, a negative growth was observed in 2012, but the region rebounded thereafter. Similarly for the West, negative growth was observed in 2012, but rebounded and sustained positive growth thereafter.

It was observed that there was a slight dip in total sales in 2012. From the visualization above, it can be inferred that South had contributed the most to that dip, followed by the West and then the Central. Interestingly, the East region still grew positively in 2012.

Business Question 4: **Profitability**

We can see profit margin of Categories:

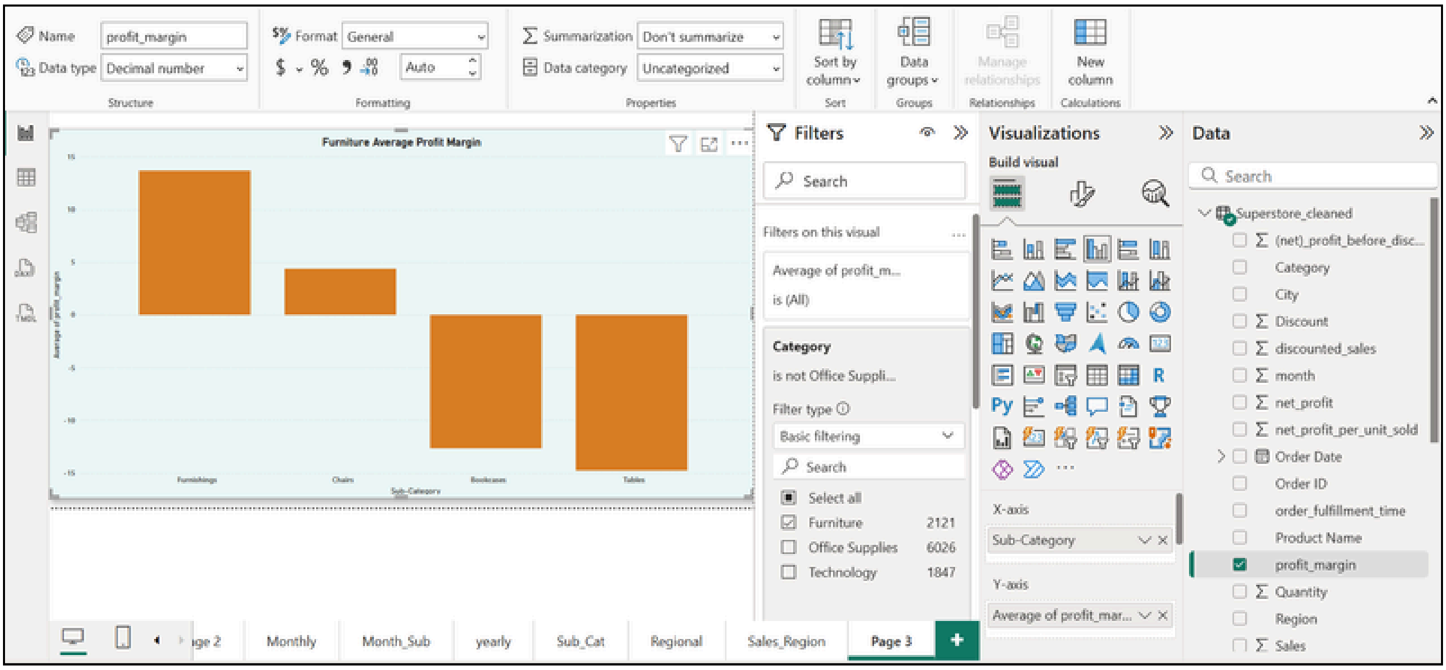


Insight: The profit margin of sub-categories like Labels, Paper, Copiers, Art, Suppliers etc is positive. Label, Paper and Envelopes have high Profit margin.

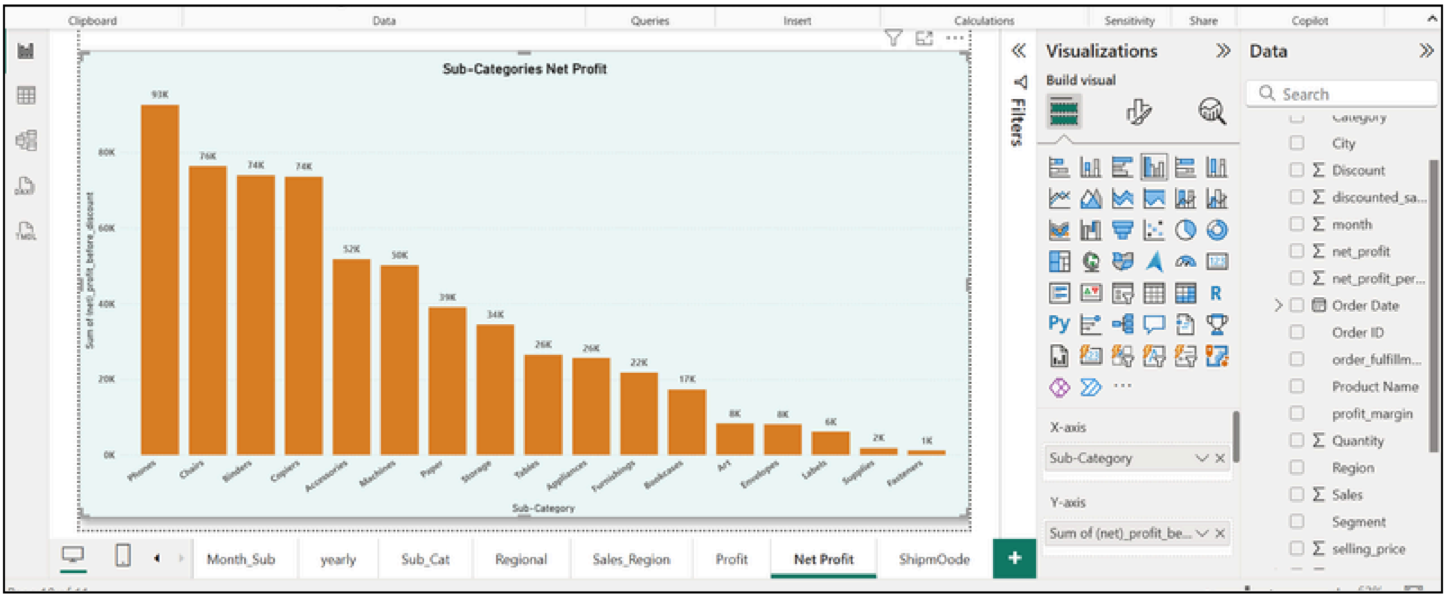
Sub-categories Tables, Binders, Appliances, Machines, Bookcases have negative profit margin.

Profit Margin of Individual Category and its subcategories.

Profit Margin of Category Furniture:

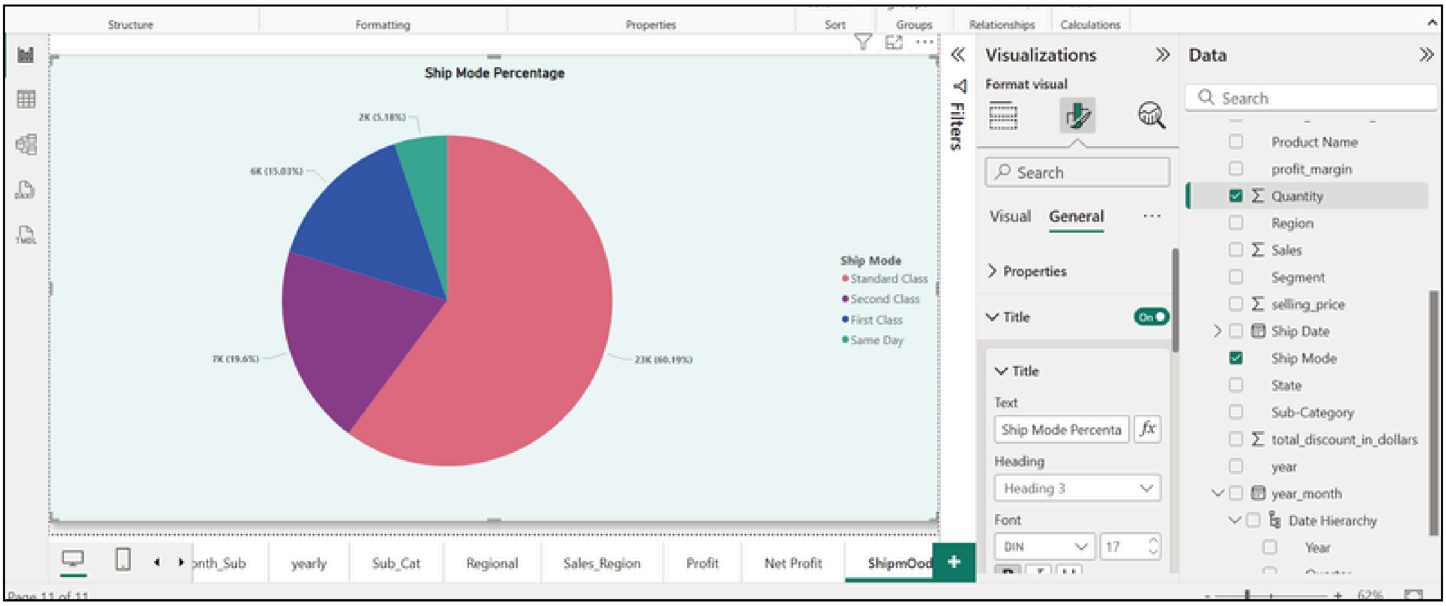


Insight: For Furnitures products, furnishings products, on average, are the most profitable followed by chairs products. On the other hand, the company was operating at a loss on tables and bookcases products. Assuming similar sales, loss on tables are higher than gains on furnishings.



Insight: By product sub-category, phones, chairs, and binders products had generated the highest **total** profit from the technology, furniture, and office supplies category, respectively. They are also the ones who generated the highest profit, overall.

Ship Mode Percentages:



Insight: 60% of all orders had been shipped through standard class. 20% was shipped through second class, and 15% was shipped through first class. 5% of all orders was shipped through same day mode.

Predictive Modeling

Business Question: To build a model that predicts future sales for a business, aiding inventory management, resource allocation, and planning.

Predictive Modeling using Python in Kaggle Notebook:

Holt Winter Forecasting Method: Holt-Winters is a **time series forecasting method** that extends exponential smoothing to handle both **trend** and **seasonality** in your data. It's widely used for sales forecasting because most sales data have patterns that repeat over time (like seasonal spikes during holidays) and may have an upward or downward trend.

Core Idea: Holt-Winters has **three components**:

1. **Level (L)** – the baseline value of the series.
2. **Trend (T)** – the direction the series is moving (up or down).
3. **Seasonality (S)** – recurring patterns within a fixed period (like weekly, monthly, or yearly cycles).

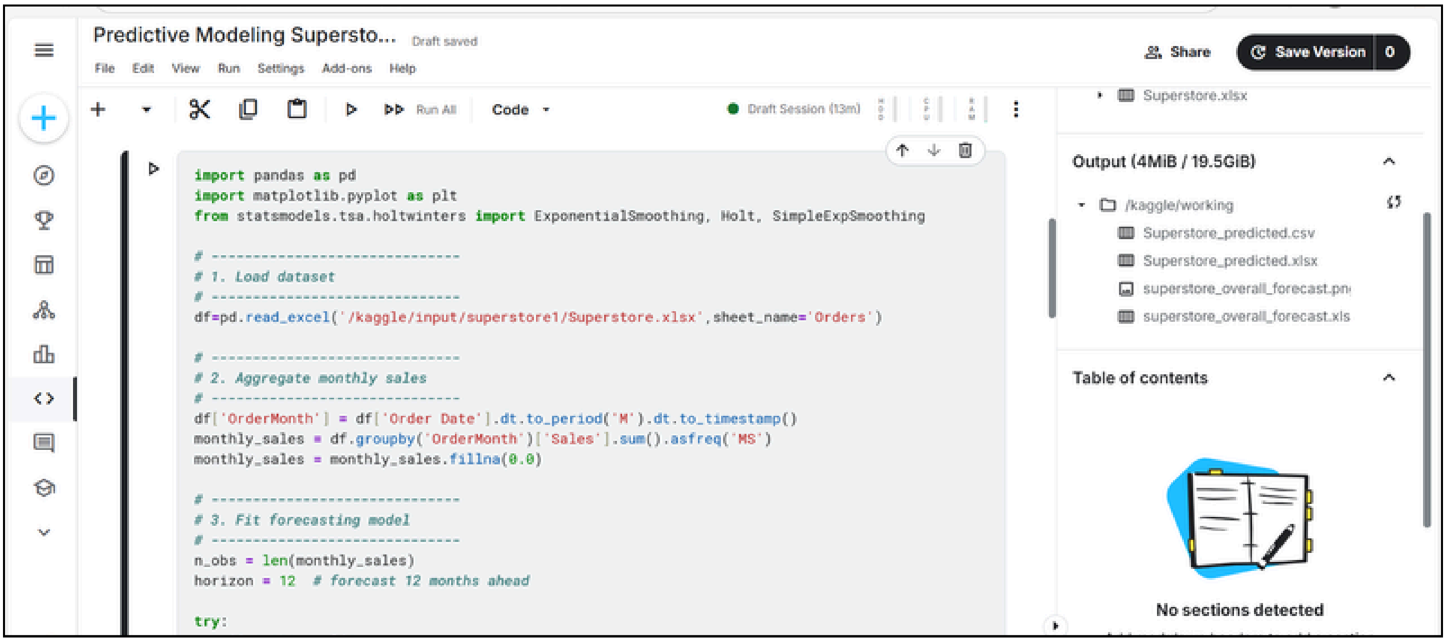
The method updates these components at each time step to make forecasts.

Exponential smoothing is a **time series forecasting technique** that predicts future values based on **weighted averages of past observations**, giving **more weight to recent data** and less to older data. It's simple, yet surprisingly effective for many forecasting tasks, especially when data has no complex seasonality or trends.

Types of Exponential Smoothing:

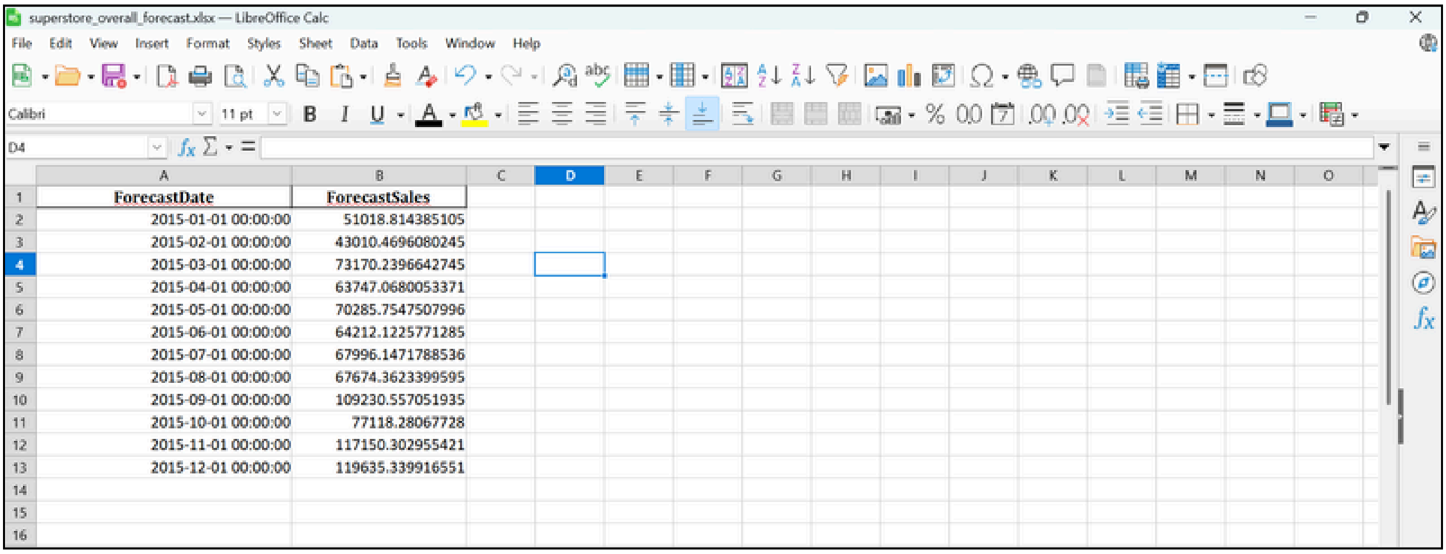
1. **Simple Exponential Smoothing (SES)**
   - Suitable for **data without trend or seasonality.**
   - Only smooths the **level.**
2. **Holt's Linear Trend**
   - Extends SES to include **trend.**
   - Good for data with **upward or downward trends.**
3. **Holt-Winters (Triple Exponential Smoothing)**
   - Adds **seasonality** on top of level and trend.
   - Used for **sales data with seasonal patterns**.

So Holt-Winters is really just a **more advanced version of exponential smoothing**.



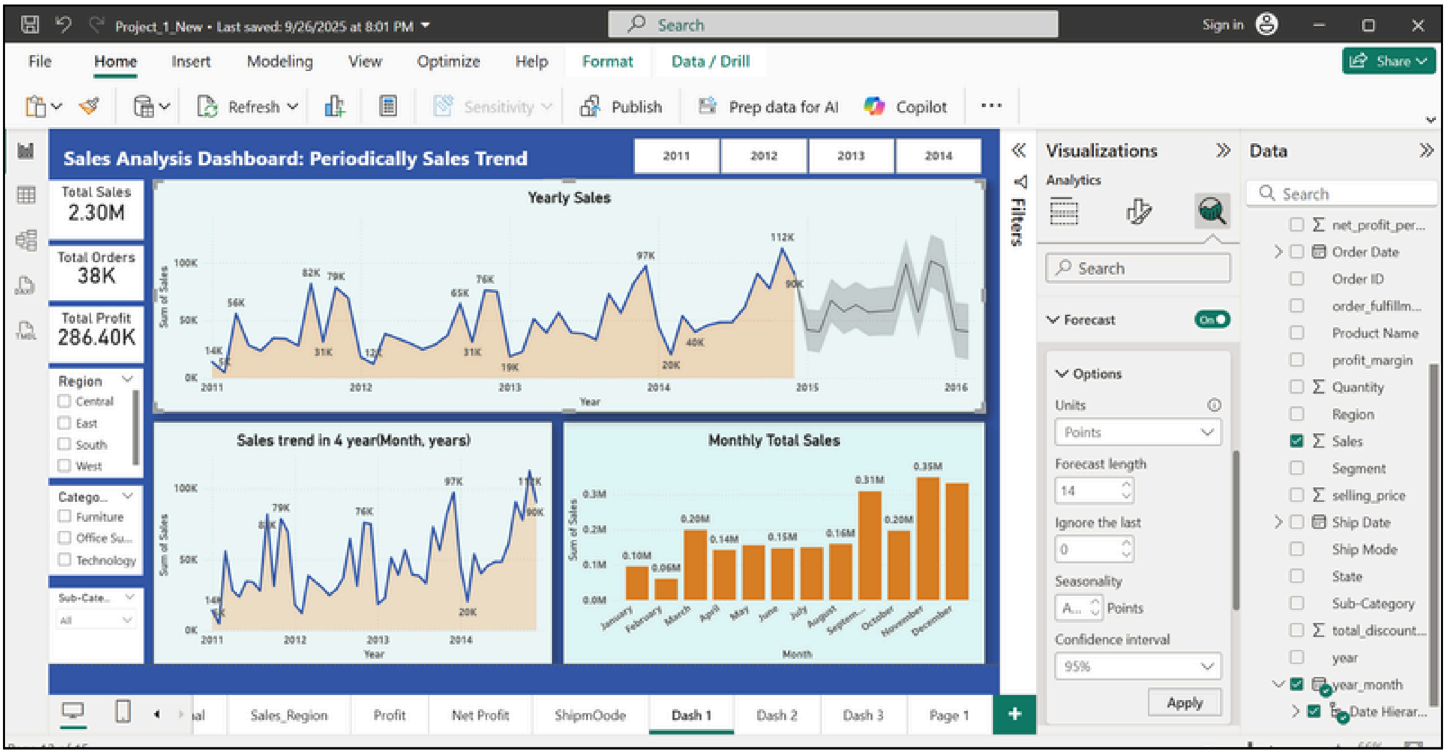
Output:

Output File showing forecast for each month in 2015:



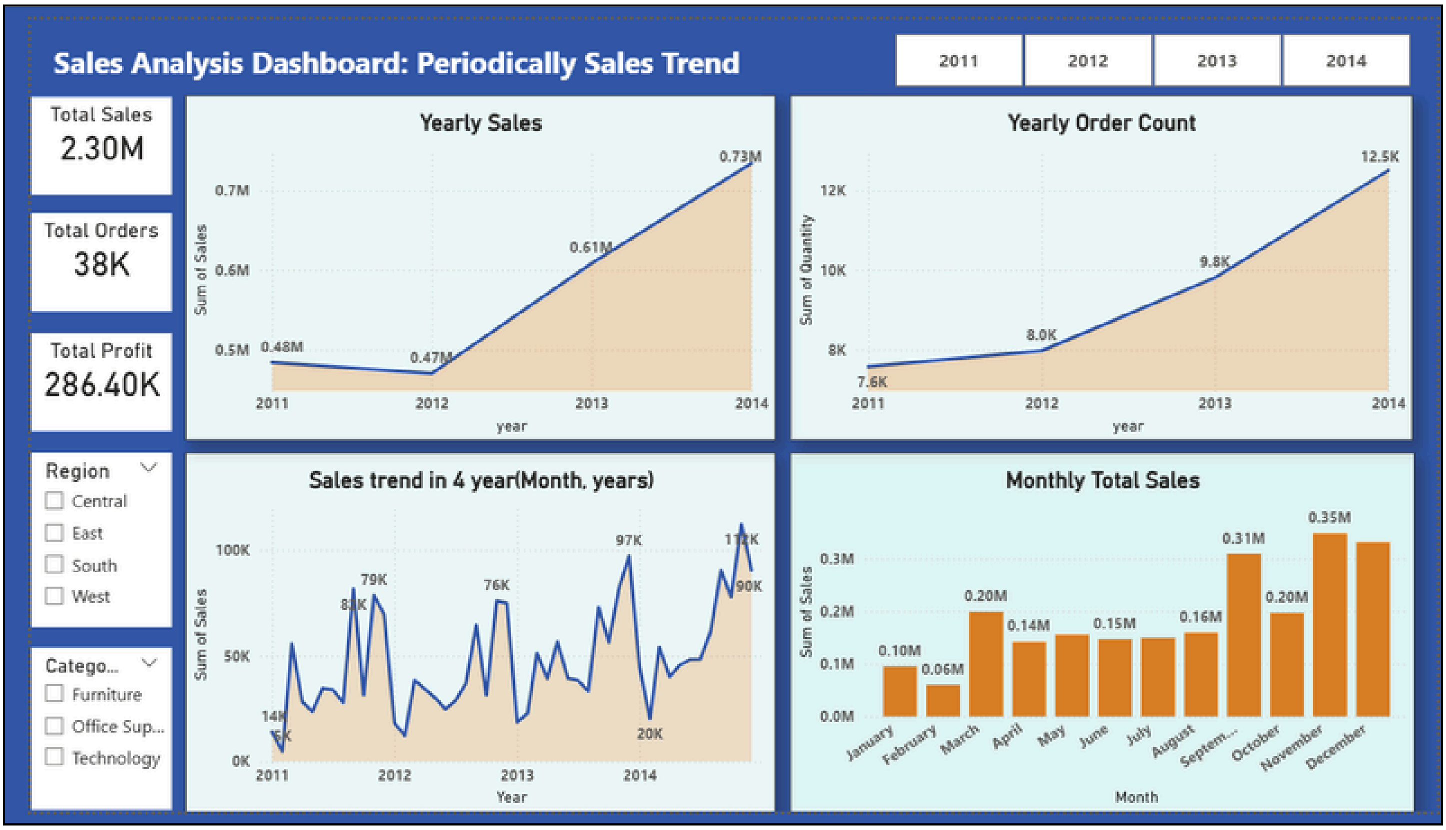
ForecastDate	ForecastSales
2015-01-01 00:00:00	51018.814385105
2015-02-01 00:00:00	43010.4696080245
2015-03-01 00:00:00	73170.2396642745
2015-04-01 00:00:00	63747.0680053371
2015-05-01 00:00:00	70285.7547507996
2015-06-01 00:00:00	64212.1225771285
2015-07-01 00:00:00	67996.1471788536
2015-08-01 00:00:00	67674.3623399595
2015-09-01 00:00:00	109230.557051935
2015-10-01 00:00:00	77118.28067728
2015-11-01 00:00:00	117150.302955421
2015-12-01 00:00:00	119635.339916551

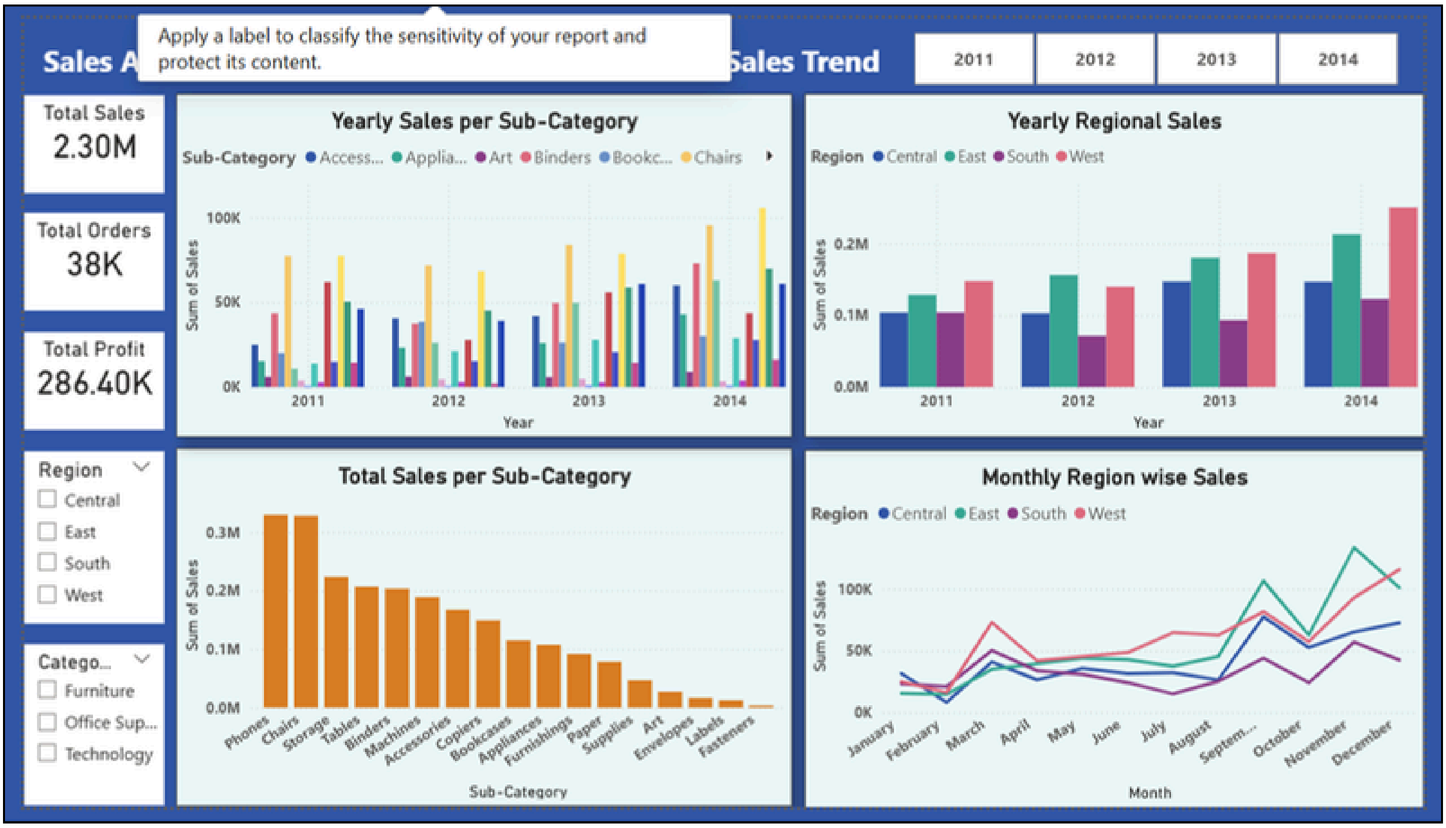
Comparing Forecast calculated by Python with Forecast generated with Power BI

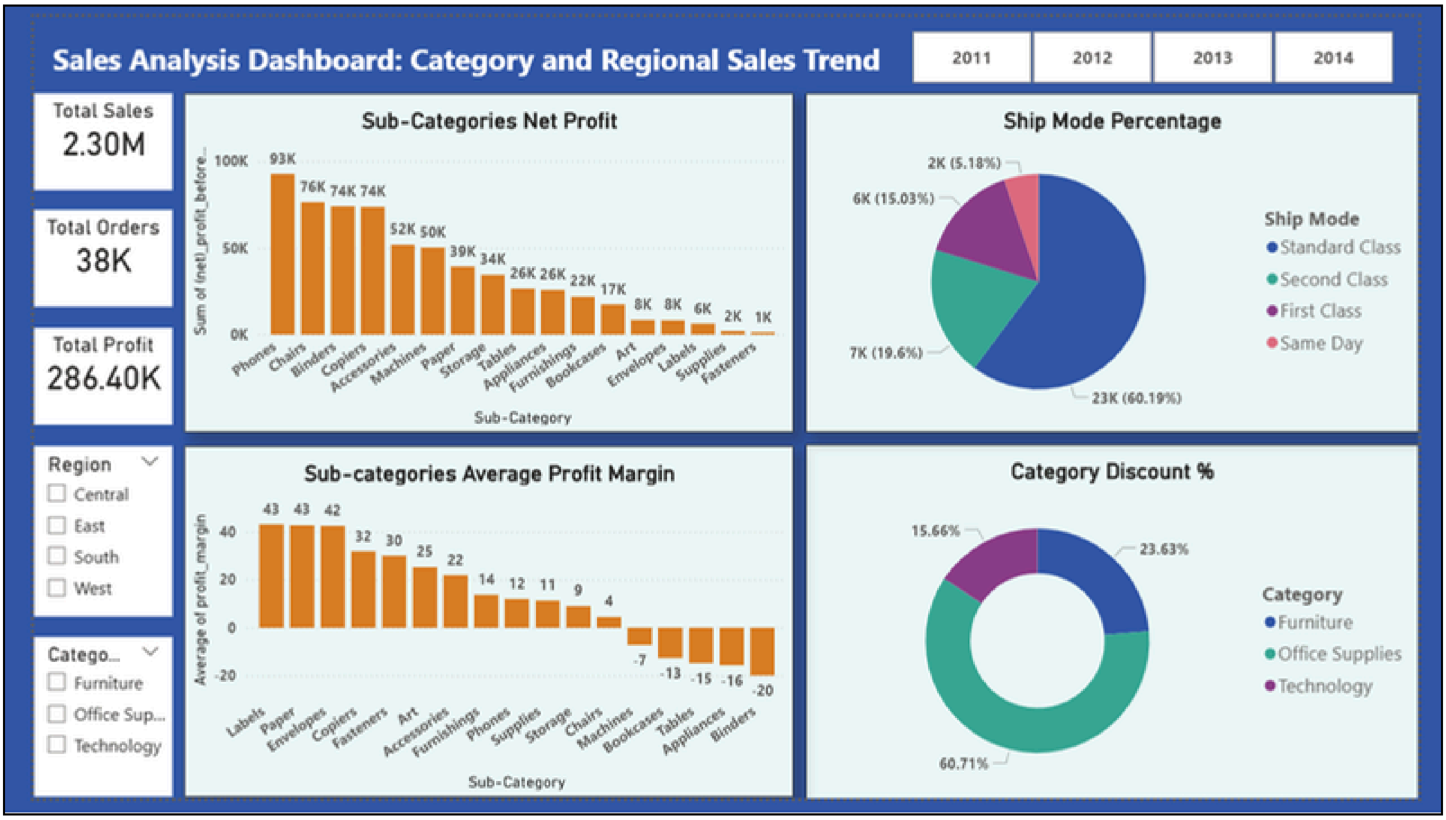


Dashboard

Dashboard Design:




Apply a label to classify the sensitivity of your report and protect its content.
Sales A
Sales Trend
2011
2012
2013
2014
Total Sales
2.30M
Total Orders
38K
Total Profit
286.40K
Yearly Sales per Sub-Category
Sub-Category Access... Applia... Art Binders Bookc... Chairs
100K
50K
0K
Sum of Sales
Year
Yearly Regional Sales
Region Central East South West
0.2M
0.1M
0.0M
Region
Central
East
South
West
Catego...
Furniture
Office Sup...
Technology
Total Sales per Sub-Category
0.3M
0.2M
0.1M
0.0M
Phones
Chairs
Storage
Tables
Binders
Machines
Accessories
Copiers
Bookcases
Appliances
Furnishings
Paper
Supplies
Art
Envelopes
Labels
Fasteners
Sub-Category
Monthly Region wise Sales
100K
50K
January
February
March
April
May
June
July
August
Septem...
October
November
December
Month


Sales Analysis Dashboard: Category and Regional Sales Trend
2011
2012
2013
2014
Total Sales
2.30M
Total Orders
38K
Total Profit
286.40K
Sub-Categories Net Profit
Sum of (net)_profit_before...
100K
50K
0K
93K
76K
74K
74K
52K
50K
39K
34K
26K
26K
22K
17K
8K
8K
6K
2K
1K
Phones
Chairs
Binders
Copiers
Accessories
Machines
Paper
Storage
Tables
Appliances
Furnishings
Bookcases
Art
Envelopes
Labels
Supplies
Fasteners
Sub-Category
Ship Mode Percentage
2K (5.18%)
6K (15.03%)
7K (19.6%)
23K (60.19%)
Ship Mode
Standard Class
Second Class
First Class
Same Day
Region
Central
East
South
West
Catego...
Furniture
Office Sup...
Technology
Sub-categories Average Profit Margin
Average of profit_margin
43
43
42
32
30
25
22
14
12
11
9
4
-7
-13
-15
-16
-20
Labels
Paper
Envelopes
Copiers
Fasteners
Art
Accessories
Furnishings
Phones
Supplies
Storage
Chairs
Machines
Bookcases
Tables
Appliances
Binders
Sub-Category
Category Discount %
15.66%
23.63%
60.71%
Category
Furniture
Office Supplies
Technology

Final Analysis and Insights

Final Analysis and Insights:

The company experienced year-over-year sales growth, with 2013 showing the highest growth and 2012 the slowest. Seasonal sales trends were also evident, with spikes in November, December, and September. Sales were highly variable in March, and around September and October. Phones, chairs, and storage products led sales within their respective categories, while copiers, furnishings, and fasteners lagged.

By region, the West and East had higher sales across most categories, while the South had consistently lower sales. Central, on the other hand, showed positive growth in 2012, in contrast to all other regions in the same year. Furthermore, the West had the fastest average annual sales growth.

Despite fluctuating profitability, the company maintained a 10%+ profit margin from 2011 to 2014. Chairs, phones, and storage products were the least profitable, while furnishings, copiers, and labels were the most profitable sub-categories. Sales discounts significantly impacted profits, with tables and office supplies experiencing the largest drops.

In conclusion, while the company saw sales growth and maintained profitability, seasonal and regional variations played a significant role in its performance.