

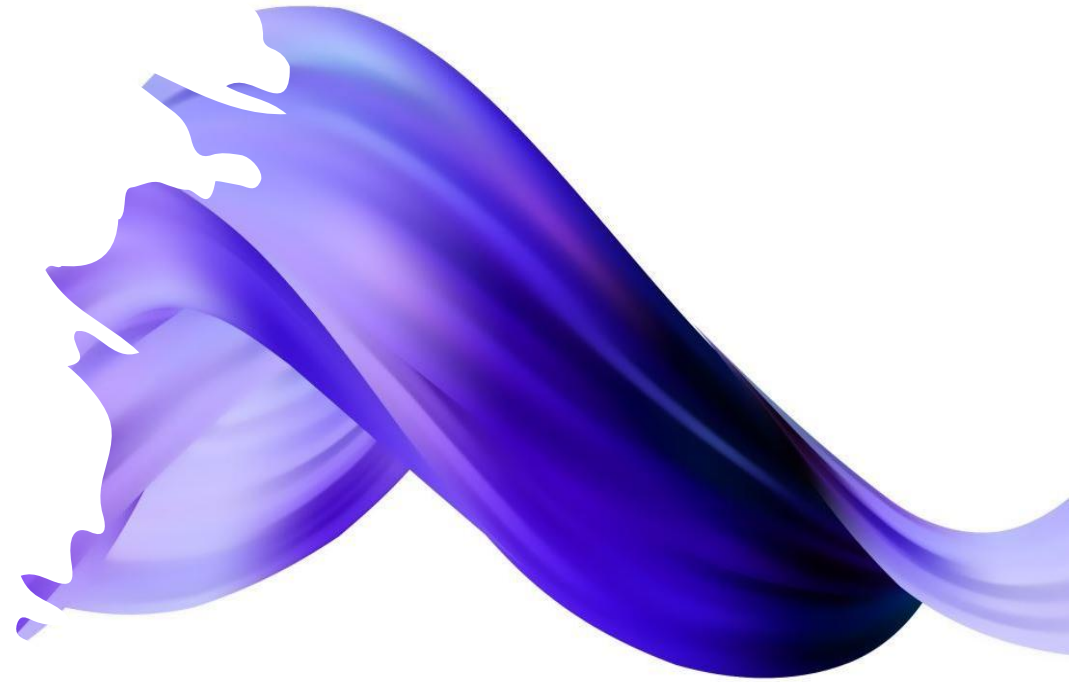
Student Details

- **Name:** Shubham Kumar Jha
- **Email:** senshubham284@gmail.com
- **AICTE Id:** STU61c7440d19e851640449037
- **Organization:** DGT
- **College Name:** College Of Engineering Roorkee
- **Domain:** Data Analytics
- **Start Date:** 12 June 2023
- **End Date:** 24 July 2023
- **Internship Name:** IBM SkillBuild
- **Foundation:** Edunet



PROJECT TITLE

Analysis of Superstore Dataset



Project Title: Analysis of Superstore Dataset

- **Introduction:** The goal of this project is to analyze the Superstore dataset to gain
- **Data Collection and Preprocessing:** Collect and preprocess the Superstore dataset
- **Sales Analysis:** Analyze sales metrics, trends, and factors influencing sales fluctuations
- **Customer Behavior Analysis:** Study customer demographics, preferences, and
- **Operational Efficiency Analysis:** Evaluate operational efficiency, identify bottlenecks
- **Conclusion and Next Steps:** Summarize findings, plan for advanced analysis, predictive



AGENDA



S.no.	Topics Name	Page no
1	Project Overview	3-10
2	Data set 1. Dataset loading 2. About the Dataset	11-12
3	Some Statistical Information	13
4	Exploratory Data Analysis (EDA) <ul style="list-style-type: none">• What are the top selling products in the superstore?• What are the top profit products in the superstore?• What is the total Sales and Profit by region?• Select top 5 cities by profit and Sort the data by profit in descending order	14-21
5	•The Best Sales Plotting • Conclusion	22-23

Project Overview

The analysis on Superstore dataset is a comprehensive study that aims to analyze the sales performance of a fictional retail company called "Superstore". The dataset used in this analysis contains information about sales transactions, customers, products, and geographical locations. The analysis involves using Power BI, a data visualization and reporting tool, to create interactive dashboards and reports that provide insights into the sales performance of Superstore

Purpose: The purpose of the "Analysis of Superstore dataset" is to gain insights into sales trends, customer behavior, and operational efficiency in order to optimize store performance and make data-driven recommendations for improvement.

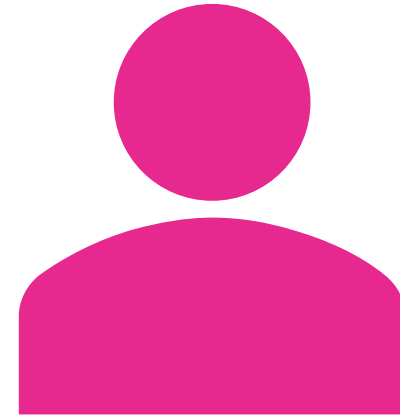
Scope: The scope of the analysis includes examining the Superstore dataset, which consists of sales transactions, customer demographics, product categories, and geographical regions. The analysis will involve data cleaning, exploratory data analysis, sales analysis, customer behavior analysis, and operational efficiency analysis.

Objectives

- Identify sales trends, such as seasonal patterns and fluctuations, to optimize inventory management and sales forecasting
- Understand customer behavior by analyzing demographics, preferences, and purchase patterns to develop targeted marketing strategies and enhance customer satisfaction
- Provide data-driven recommendations to optimize store performance, improve customer experience, and increase overall profitability based on the analysis findings



WHO ARE THE
END USERS



Target Audience or End Users

- Store Managers: They require insights on sales performance, customer behavior, and operational efficiency to make informed decisions and optimize store operations
- Buying patterns to develop targeted marketing campaigns and improve customer



Characteristics and Needs

They seek comprehensive data analysis, visualizations, and actionable recommendations to identify areas for improvement, enhance profitability, and streamline operations



The image shows a wooden desk with a calculator, a cup of coffee, and several sheets of paper displaying various business charts and tables. A black pen rests on one of the papers. The charts include bar graphs, pie charts, and line graphs, all with data labels.

Chart 1: Bar Graph (Top Left)

This bar graph shows data for 12 months. The bars are blue and show a general upward trend, peaking in the 10th month before declining slightly.

Chart 2: Pie Chart (Top Right)

This pie chart shows data for 2008 and 2009. The 2008 data is represented by a blue slice (18%), an orange slice (7%), and a grey slice (17%). The 2009 data is represented by a blue slice (8%).

Chart 3: Bar Graph (Middle Left)

This bar graph shows data for 12 months. The bars are blue and show a general upward trend, peaking in the 10th month before declining slightly.

Chart 4: Bar Graph (Middle Right)

This bar graph shows data for 12 months. The bars are blue and show a general upward trend, peaking in the 10th month before declining slightly.

Chart 5: Pie Chart (Bottom Left)

This pie chart shows data for 2008 and 2009. The 2008 data is represented by a blue slice (1%), an orange slice (39%), and a grey slice (34%). The 2009 data is represented by a blue slice (5.52), a grey slice (7.29), and a blue slice (7.51).

Chart 6: Pie Chart (Bottom Right)

This pie chart shows data for 2008 and 2009. The 2008 data is represented by a blue slice (1%), an orange slice (39%), and a grey slice (34%). The 2009 data is represented by a blue slice (5.52), a grey slice (7.29), and a blue slice (7.51).

Table 1: Data for Chart 1

Month	Value
Jan	0.17
Feb	0.95
Mar	1.56
Apr	2.09
May	2.69
Jun	2.73
Jul	3.49
Aug	6.65
Sep	7.56
Oct	5.90
Nov	2.43
Dec	5.60

Table 2: Data for Chart 2

Year	Value
2008	8.52
2009	8.74

Table 3: Data for Chart 3

Month	Value
Jan	0.17
Feb	0.95
Mar	1.56
Apr	2.09
May	2.69
Jun	2.73
Jul	3.49
Aug	6.65
Sep	7.56
Oct	5.90
Nov	2.43
Dec	5.60

Table 4: Data for Chart 4

Month	Value
Jan	0.17
Feb	0.95
Mar	1.56
Apr	2.09
May	2.69
Jun	2.73
Jul	3.49
Aug	6.65
Sep	7.56
Oct	5.90
Nov	2.43
Dec	5.60

Table 5: Data for Chart 5

Month	Value
Jan	0.17
Feb	0.95
Mar	1.56
Apr	2.09
May	2.69
Jun	2.73
Jul	3.49
Aug	6.65
Sep	7.56
Oct	5.90
Nov	2.43
Dec	5.60

Table 6: Data for Chart 6

Month	Value
Jan	0.17
Feb	0.95
Mar	1.56
Apr	2.09
May	2.69
Jun	2.73
Jul	3.49
Aug	6.65
Sep	7.56
Oct	5.90
Nov	2.43
Dec	5.60

-
- A collection of business charts and tables on a wooden desk. The main chart is a large bar graph with blue bars of varying heights. Below it is a smaller bar graph with blue and green bars. To the right is a pie chart with four segments labeled 18%, 8%, 17%, and 7%. Below that is another bar graph with orange and blue bars. At the bottom right is a small table with two columns, 'Series 1' and 'Series 2', and rows for months from Jan to Dec. A black pen lies across the middle of the charts.

Solution and its value proposition

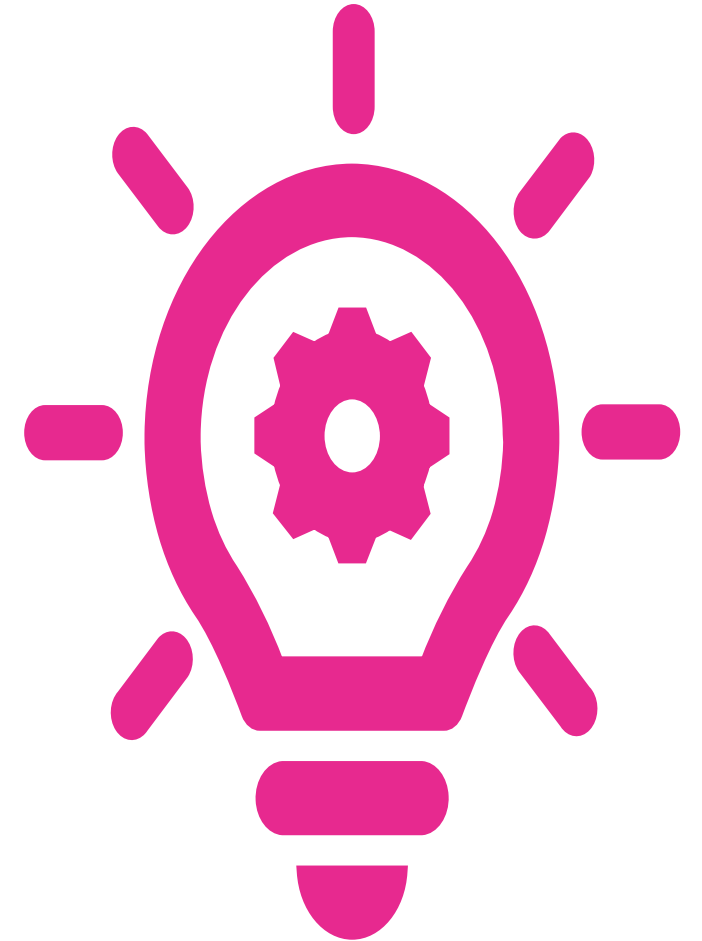
THE SOLUTION FOR THE "ANALYSIS OF SUPERSTORE DATASET" PROJECT INVOLVES CONDUCTING A COMPREHENSIVE ANALYSIS OF THE SUPERSTORE DATASET TO GAIN INSIGHTS INTO SALES TRENDS, CUSTOMER BEHAVIOR, AND OPERATIONAL EFFICIENCY. THIS ANALYSIS WILL BE CARRIED OUT USING VARIOUS STATISTICAL AND DATA MINING TECHNIQUES, AS WELL AS ADVANCED VISUALIZATION TOOLS.

- **VALUE PROPOSITION:** OUR SOLUTION PROVIDES THE FOLLOWING VALUE PROPOSITIONS:

- **DATA-DRIVEN DECISION MAKING:** BY ANALYZING THE SUPERSTORE DATASET, WE ENABLE DATA-DRIVEN DECISION MAKING FOR STORE MANAGERS AND MARKETING MANAGERS. THEY CAN MAKE INFORMED DECISIONS BASED ON COMPREHENSIVE ANALYSIS, LEADING TO IMPROVED STORE PERFORMANCE, OPTIMIZED OPERATIONS, AND TARGETED MARKETING STRATEGIES.

- **ENHANCED PROFITABILITY:** OUR ANALYSIS HELPS IDENTIFY OPPORTUNITIES FOR INCREASING SALES, IMPROVING INVENTORY MANAGEMENT, AND REDUCING COSTS, ULTIMATELY LEADING TO ENHANCED PROFITABILITY FOR THE SUPERSTORE. BY OPTIMIZING PRICING STRATEGIES, IDENTIFYING HIGH-DEMAND PRODUCTS, AND STREAMLINING OPERATIONS, THE STORE CAN MAXIMIZE ITS REVENUE AND PROFITABILITY.

- **COMPETITIVE ADVANTAGE:** LEVERAGING THE POWER OF DATA ANALYSIS, OUR SOLUTION PROVIDES THE SUPERSTORE WITH A COMPETITIVE ADVANTAGE IN THE MARKET.

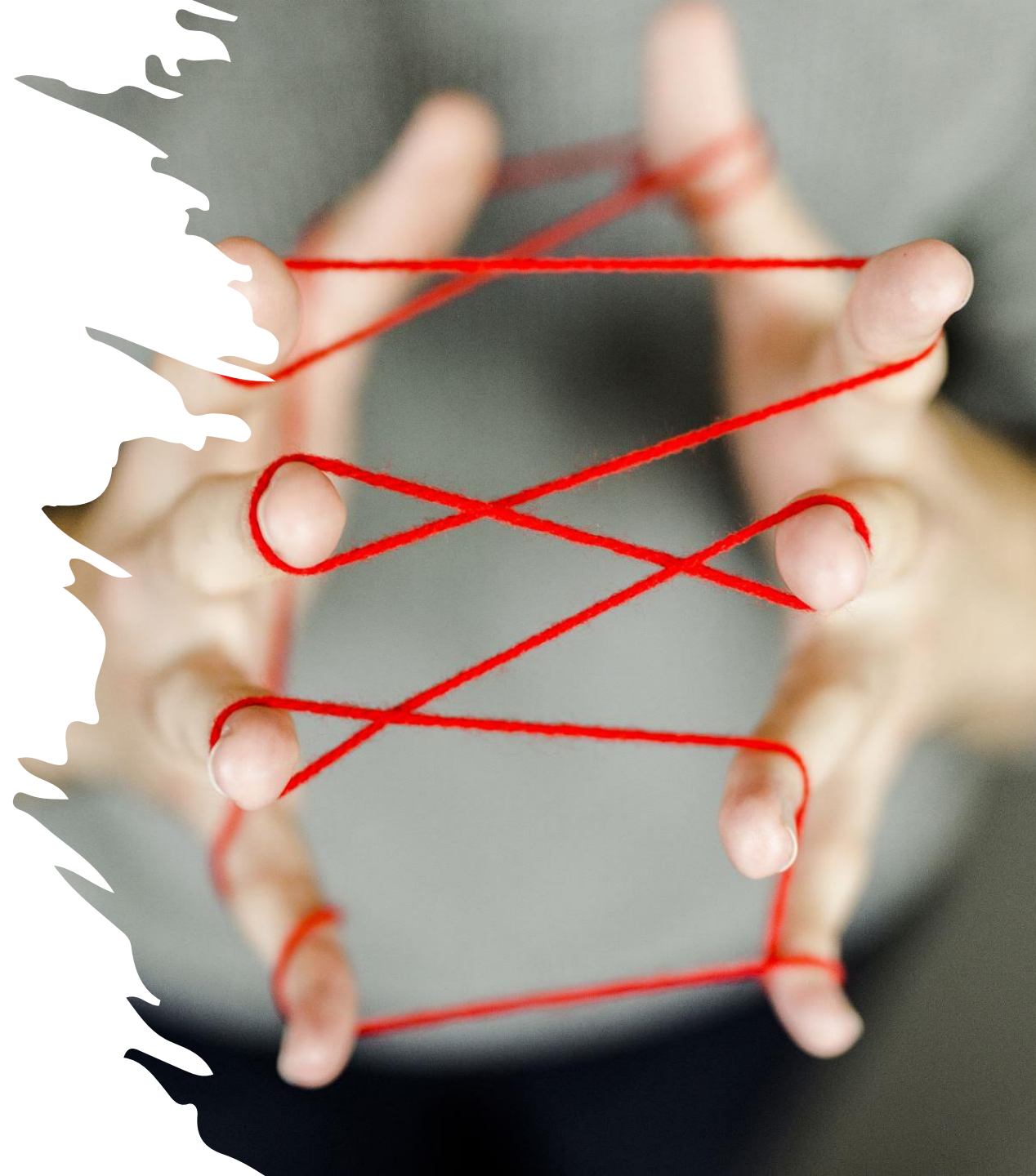


Customize the project and make it my own

- Advanced Visualization with Matplotlib and Seaborn: While data visualization is a common component of data analysis projects, my solution stands out by utilizing the powerful libraries Matplotlib and Seaborn
- By leveraging the capabilities of Matplotlib and Seaborn, my solution presents data in a visually engaging manner, enhancing the understanding of complex patterns and relationships within the Superstore dataset
- These dashboards allow stakeholders to dynamically explore and interact with the analyzed data, enabling them to drill down nature of the dashboards enhances engagement, facilitates deeper insights, and empowers users to derive actionable recommendations effectively



Modelling techniques,
methodologies, and
frameworks were
applied:



Exploratory Data Analysis (EDA):

- These techniques helped in understanding the impact of various factors on sales, customer behavior, and operational efficiency
- libraries were used to create visually appealing and informative charts, graphs, and dashboards
- These modelling techniques, methodologies, and frameworks formed the foundation of the "Analysis of Superstore dataset" project for Data Analytics, ensuring a systematic and data-driven approach to extract valuable insights from the dataset

Data Visualization:

Advanced data visualization techniques using tools like Python libraries (e.g., Matplotlib, Seaborn) were used to create visually appealing and informative charts, graphs, and dashboards. These visualizations facilitated the effective communication of analysis results and provided a clear representation of key findings.



Results



LINKS

GITHUB-Links:-

<https://github.com/ShubhamKJ123/Analysis-of-Superstore-Database.git>

Source Code:-

https://drive.google.com/file/d/1rP3BD0GXHiR_u1cq6vHHDY2zcJJvwhOb/view?usp=drive_link

Video Presentation Link:

https://drive.google.com/file/d/13M5u1CLFilzT_ASyiQDrU-Tww0bB9CjM/view?usp=drive_link

Here are some references for sales analysis on Superstore dataset:

- Chakraborty, M. (2020). Sales Analysis of Superstore using Power BI. Kaggle.

<https://www.kaggle.com/moumoyesh/sales-analysis-of-superstore-using-power-bi>

- Microsoft. (n.d.). Analyse and visualize Superstore data in Power BI.

<https://powerbi.microsoft.com/enus/tutorials/analyze-and-visualize-superstore-data/>



Research Paper

- Analyse and visualize Superstore data in Power BI
- Sales Analysis of Superstore dataset using Power BI
- Analyse and visualize Superstore data in Power BI
- A review of sales forecasting models for retail industry
- A comparative analysis of machine learning algorithms for sales forecasting in retail industry



DATA SET

Data set URL:

<https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>

About the dataset: The dataset provides information about the sales and profit from a supermarket.

Column details:

['Row ID', 'Order ID', 'Order Date', 'Ship Date', 'Ship Mode', 'Customer ID', 'Customer Name', 'Segment', 'Country', 'City', 'State', 'Postal Code', 'Region', 'Product ID', 'Category', 'Sub-Category', 'Product Name', 'Sales', 'Quantity', 'Discount', 'Profit']

Size	563kb
Number of columns	21
Number of Rows	9994
Original file format	Csv

Step-1: Importing the dataset

Importing libraries

```
import pandas as pd
```

```
import numpy as np
```

```
df=pd.read_csv("/content/drive/MyDrive/IBM_Project/Superstoredataset.csv", encoding='cp1252')
```

Df

- **Checking data type and missing values:**

```
df.info()
```

- **Read the columns or Features of the dataset:**

```
df.columns
```

- **Null Value check:**

```
df.isna().sum()
```

- **Read the Duplicate value:**

```
df.duplicated().sum()
```



Some statistical information

- Understanding the distribution of the data: The mean, min, max, and other metrics provide a quick overview of the distribution of the data
- Data normalization: The mean and std values can be used to normalize the data
- Feature scaling: The min, max, and other values can be used to scale the features to a suitable range

`df.describe()`

	Row ID	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	55190.379428	229.858001	3.789574	0.156203	28.656896
std	2885.163629	32063.693350	623.245101	2.225110	0.206452	234.260108
min	1.000000	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	2499.250000	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	4997.500000	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	7495.750000	90008.000000	209.940000	5.000000	0.200000	29.364000
max	9994.000000	99301.000000	22638.480000	14.000000	0.800000	8399.976000

Step-2: Exploratory Data Analysis – EDA

What are the top selling products in the superstore?

```
# Group the data by Product Name and sum up the sales  
by product product_group = df.groupby(["Product  
Name"]).sum()["Sales"] product_group.head()
```

```
top_5_selling_products.plot(kind="bar")
```

```
# Add a title to the plot
```

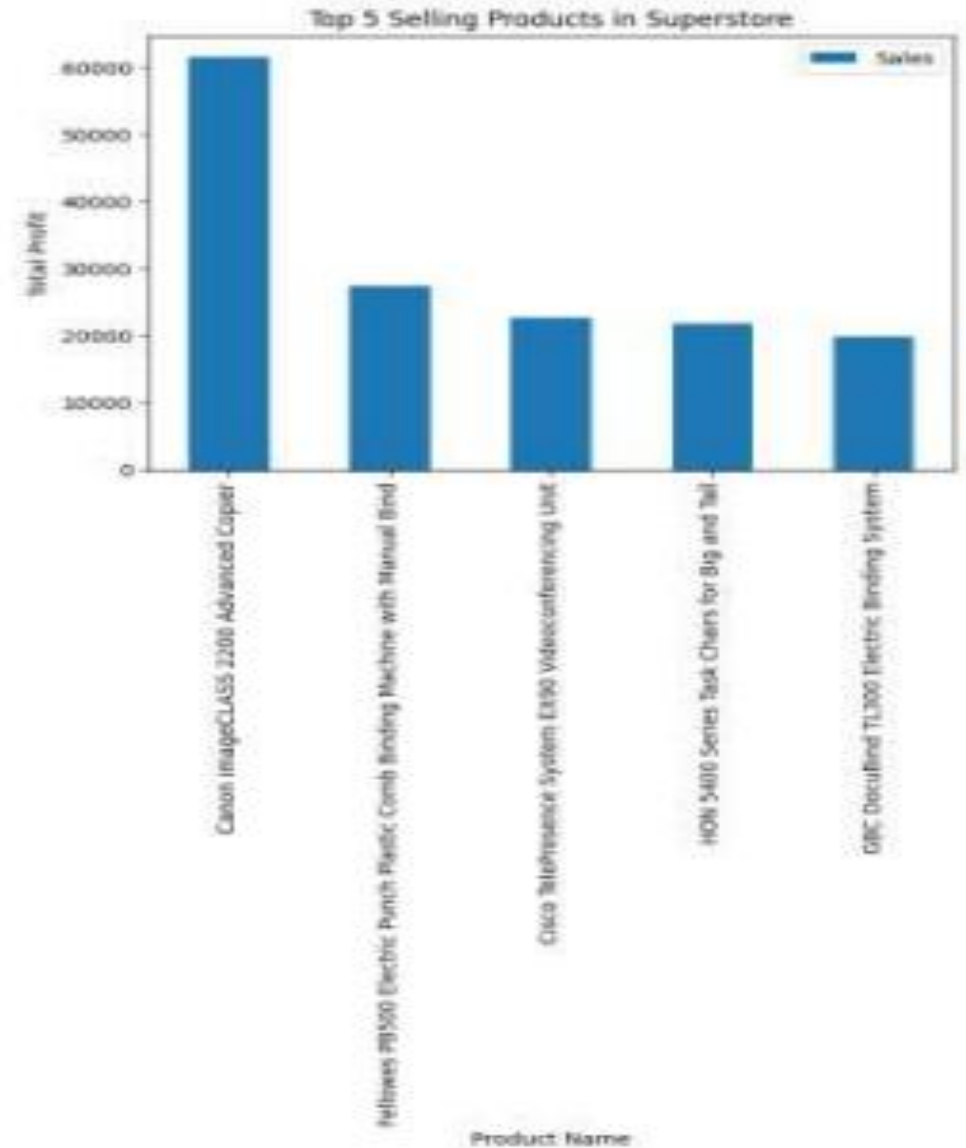
```
plt.title("Top 5 Selling Products in Superstore")
```

```
# Add labels to the x and y axes
```

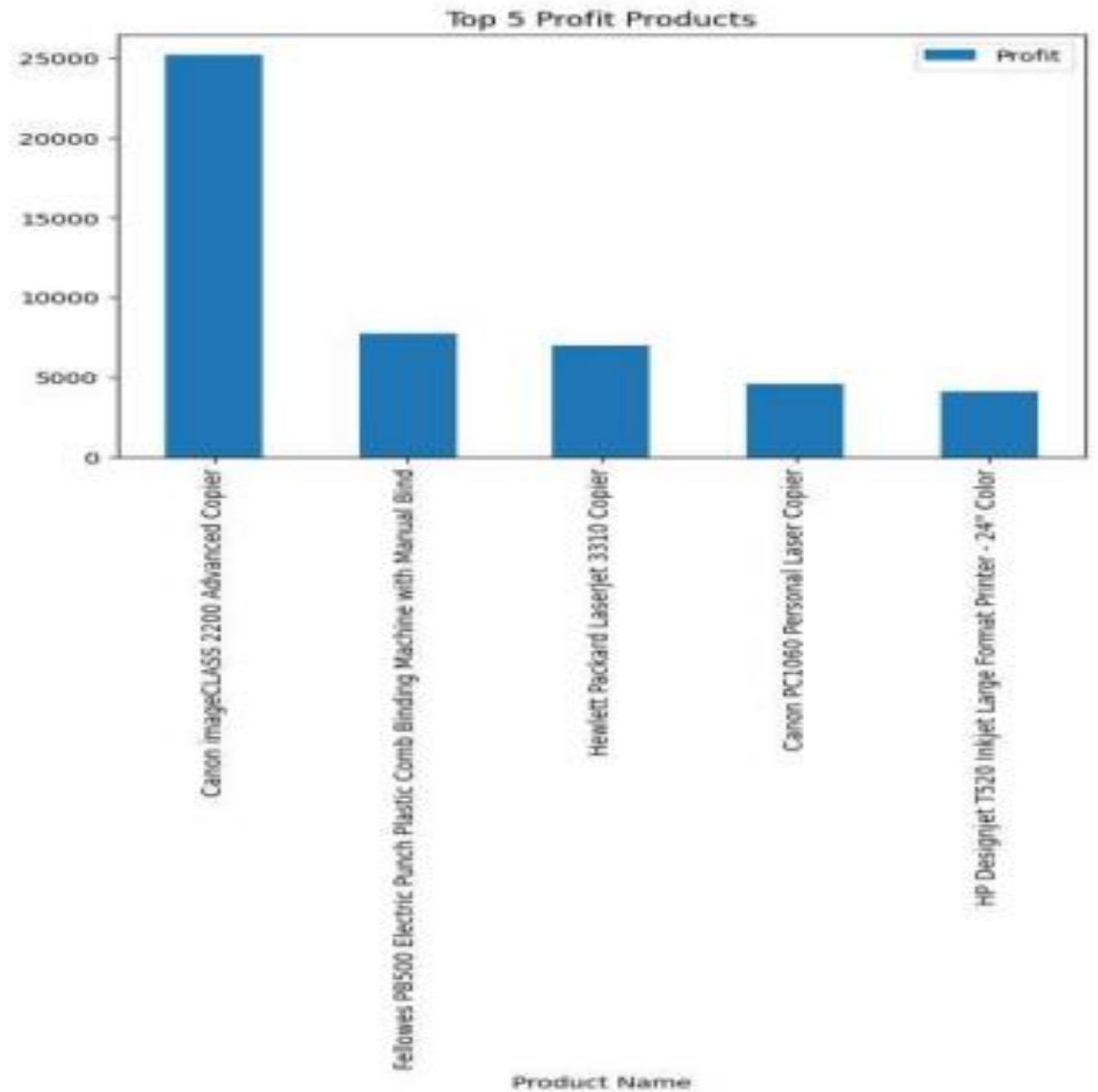
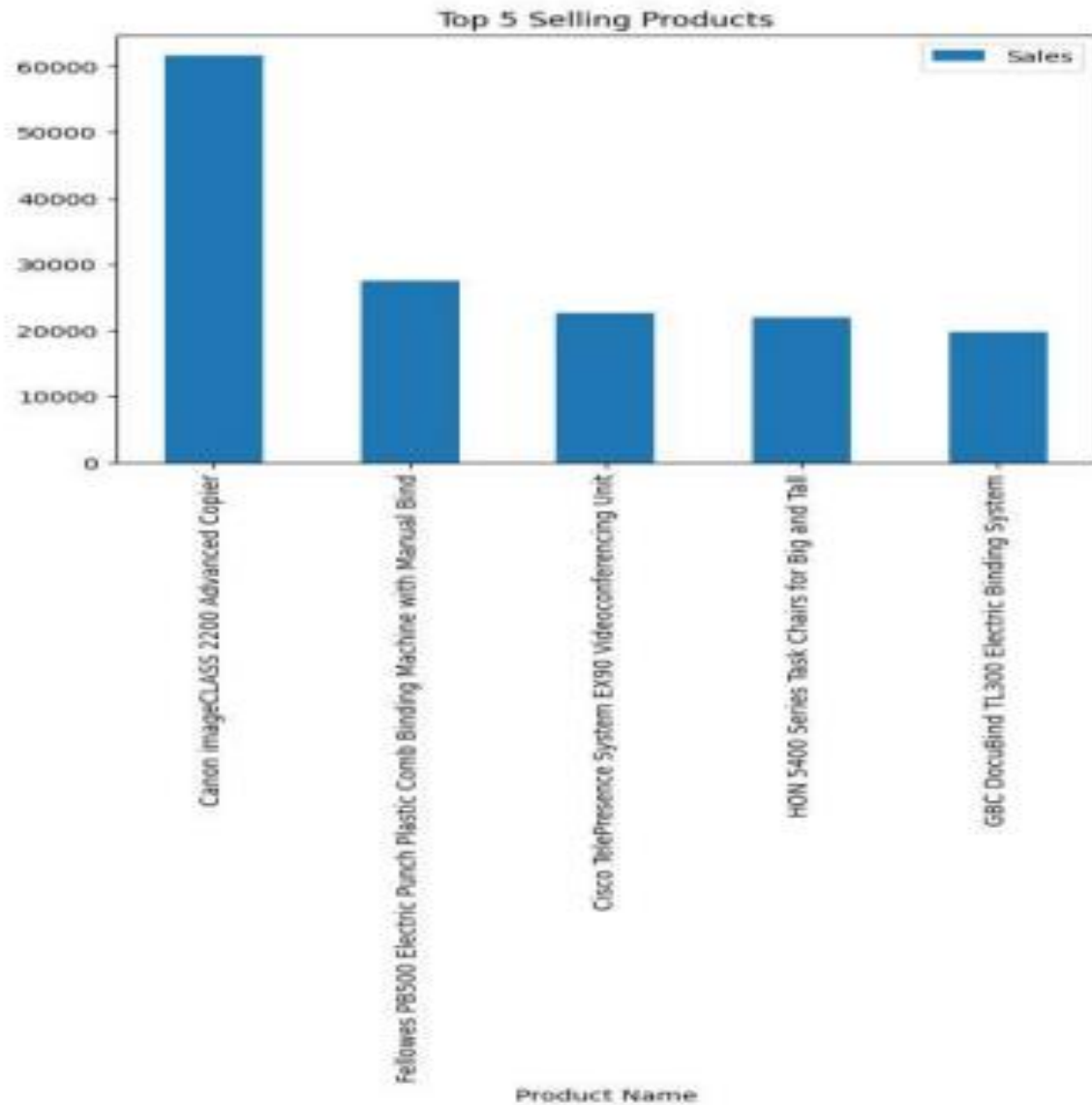
```
plt.xlabel("Product Name") plt.ylabel("Total Profit")
```

```
# Show the plot
```

```
plt.show()
```

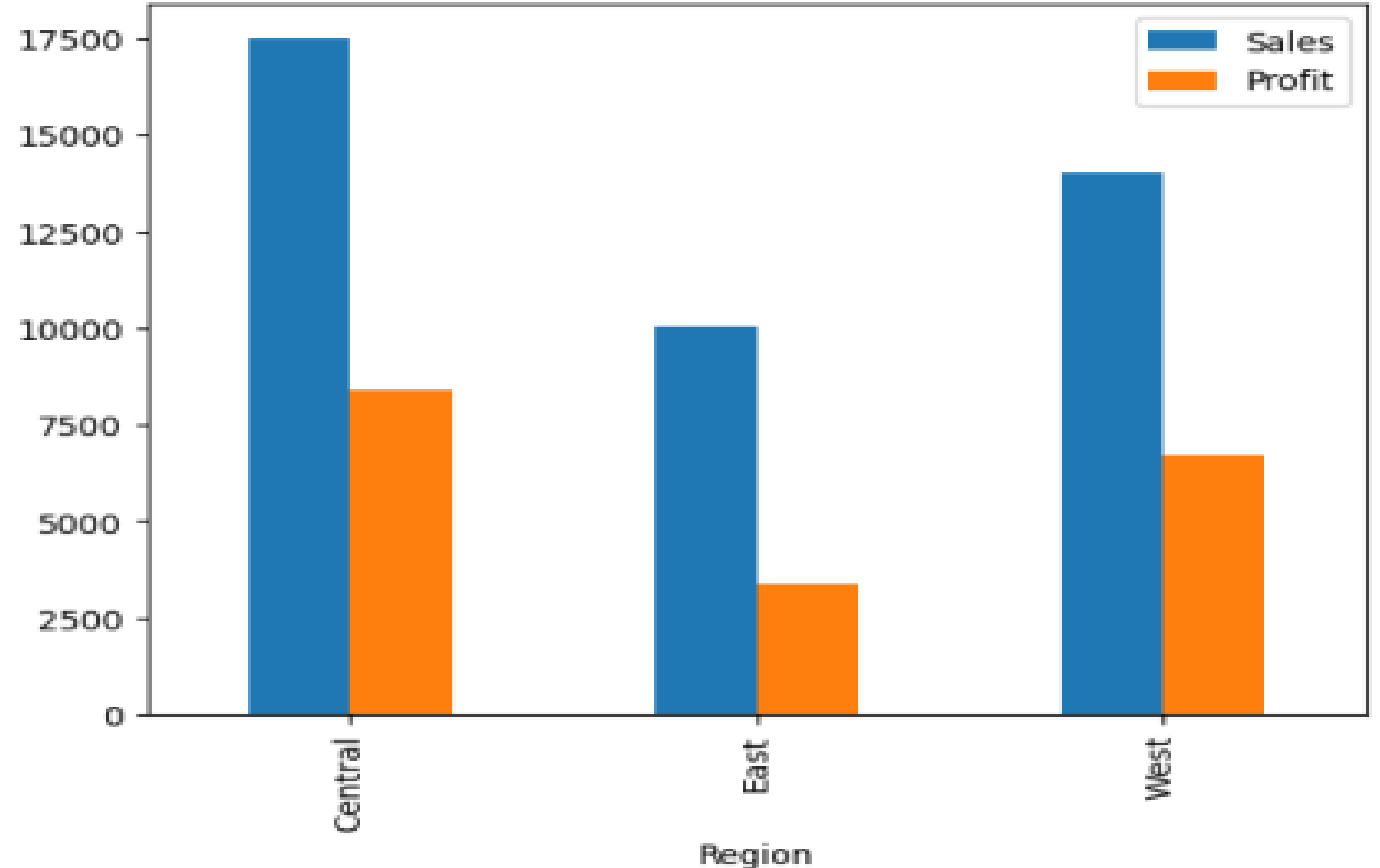


Are the top-selling products the most profitable?

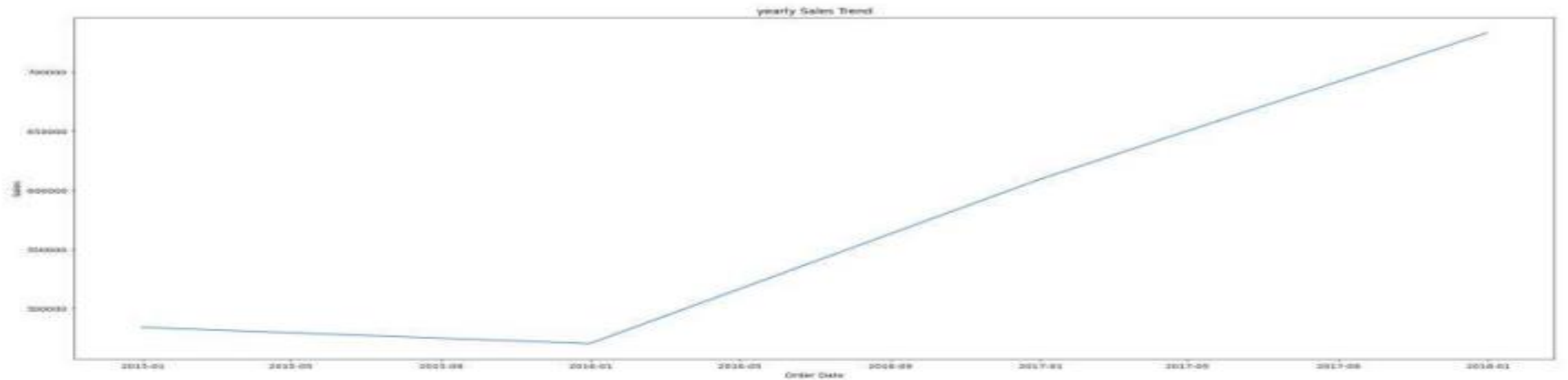
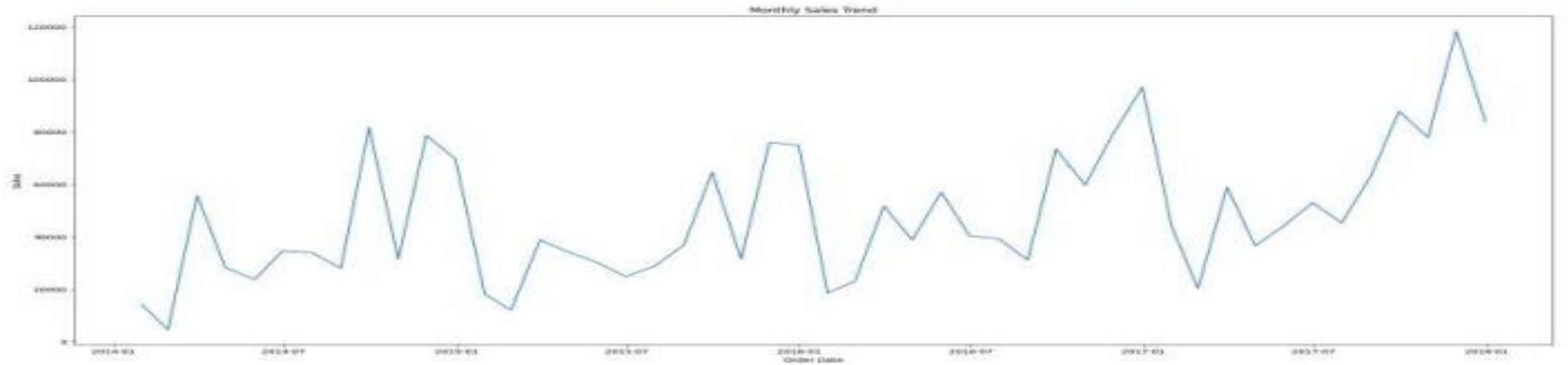


What is the total Sales and Profit by region?

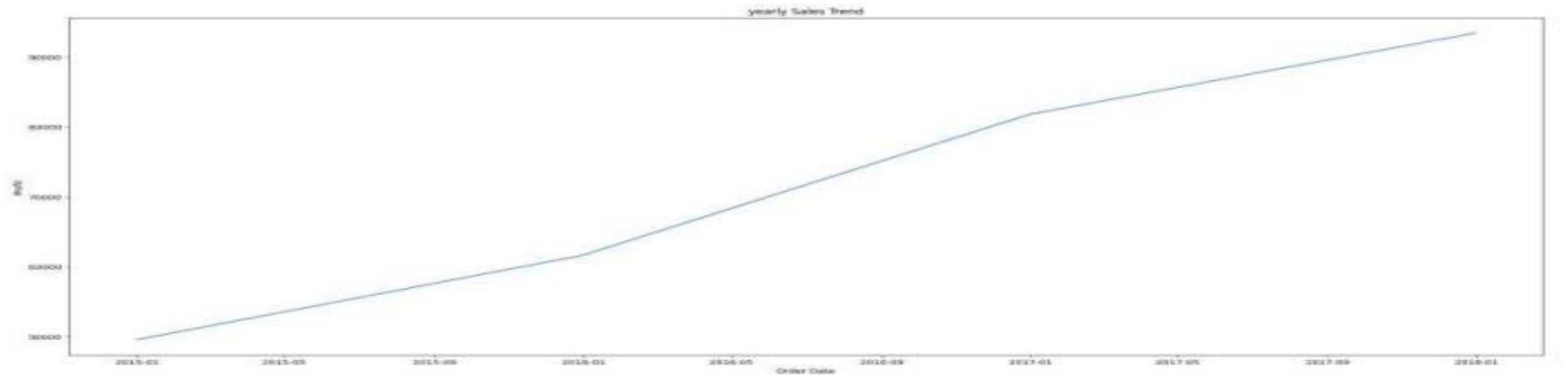
- Filter the data to only include the Canon imageCLASS 2200 Advanced Copier product = `df[df["Product Name"] == "Canon imageCLASS 2200 Advanced Copier"]`
- Group the data by Region `region_group = product.groupby(["Region"]).mean()[["Sales", "Profit"]]`
- Plotting `region_group.plot(kind="bar")`
`plt.show()`



Sales Generated by Statewise



Profit over time:



Sales Generated by Statewise:

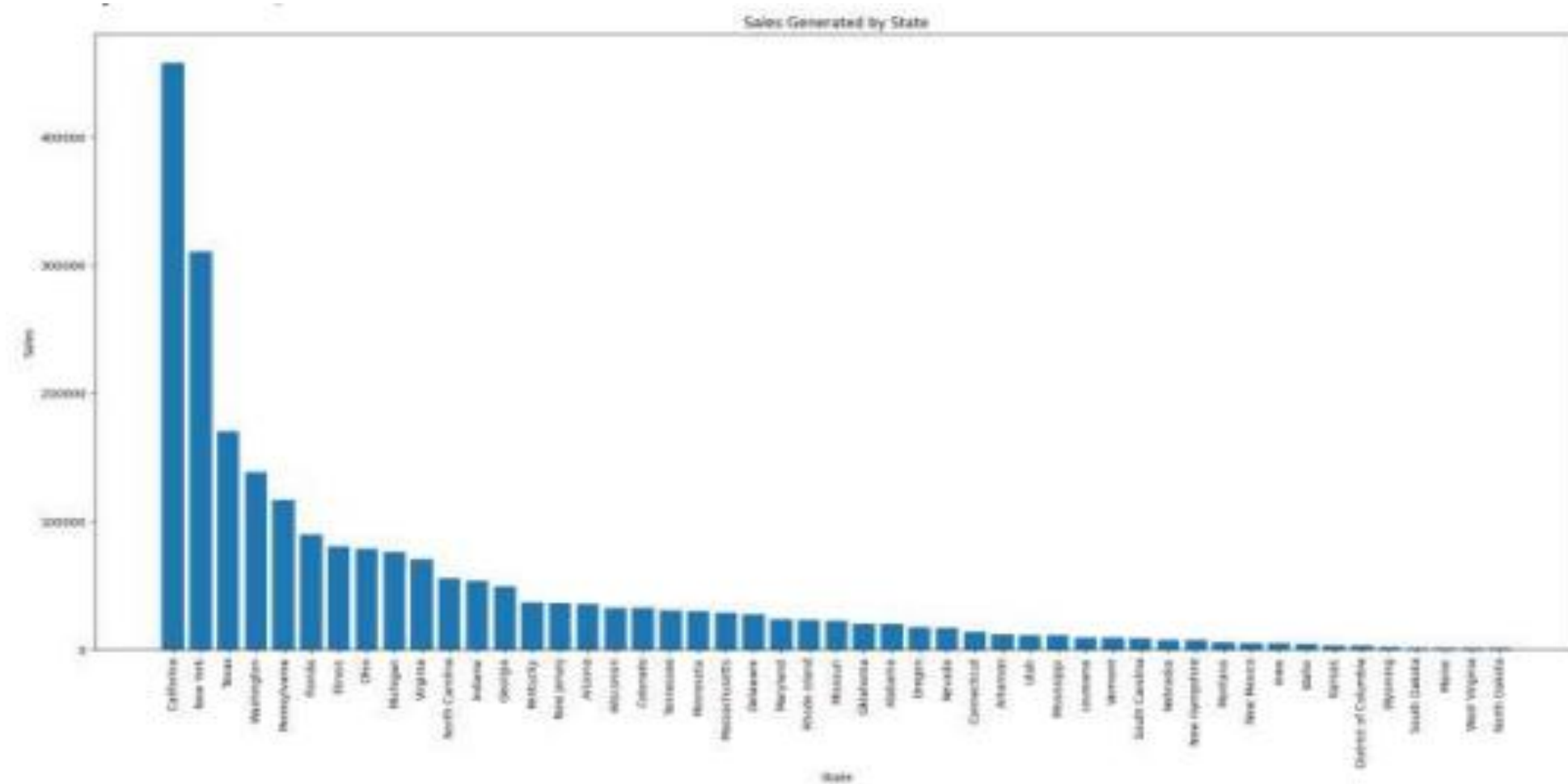
```
state_sales = df_places.groupby(['State'], as_index=False).sum() state_sales.sort_values(by='Sales', ascending=False, inplace=True)
```

```
plt.figure(figsize=(22,10)) plt.bar(state_sales['State'], state_sales['Sales'], align='center',) plt.xlabel("State") plt.ylabel("Sales") plt.title("Sales Generated by State")
```

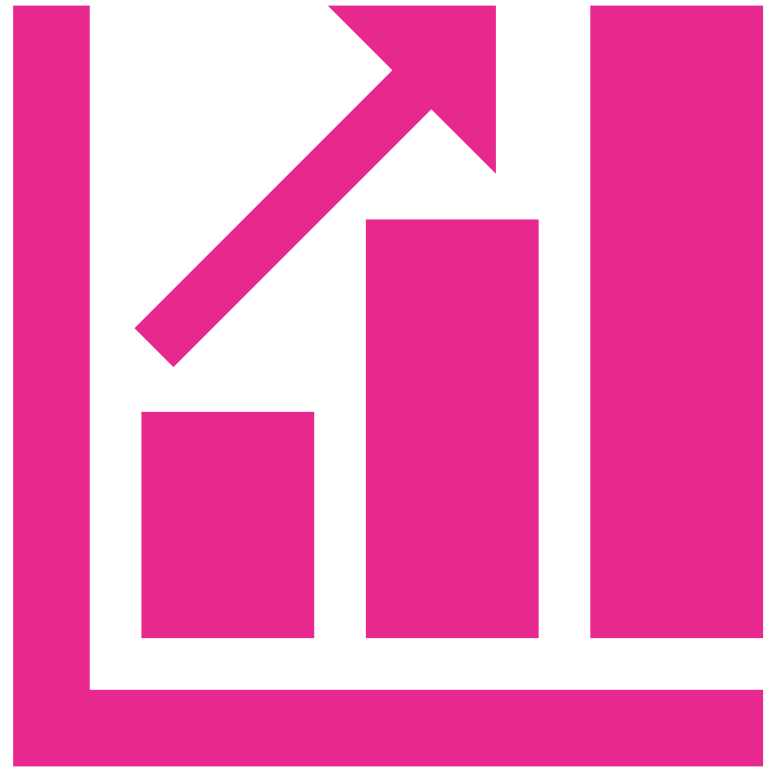
```
plt.xticks(rotation=90)
```

```
plt.show()
```

```
state_sales
```



Select top 5 cities by
sales and Sort
the data by
sales in descending
order



Query & Visualization

```
city_sales = df_places.groupby('City', as_index=False).sum()
```

- Sort the data by Sales in descending order:

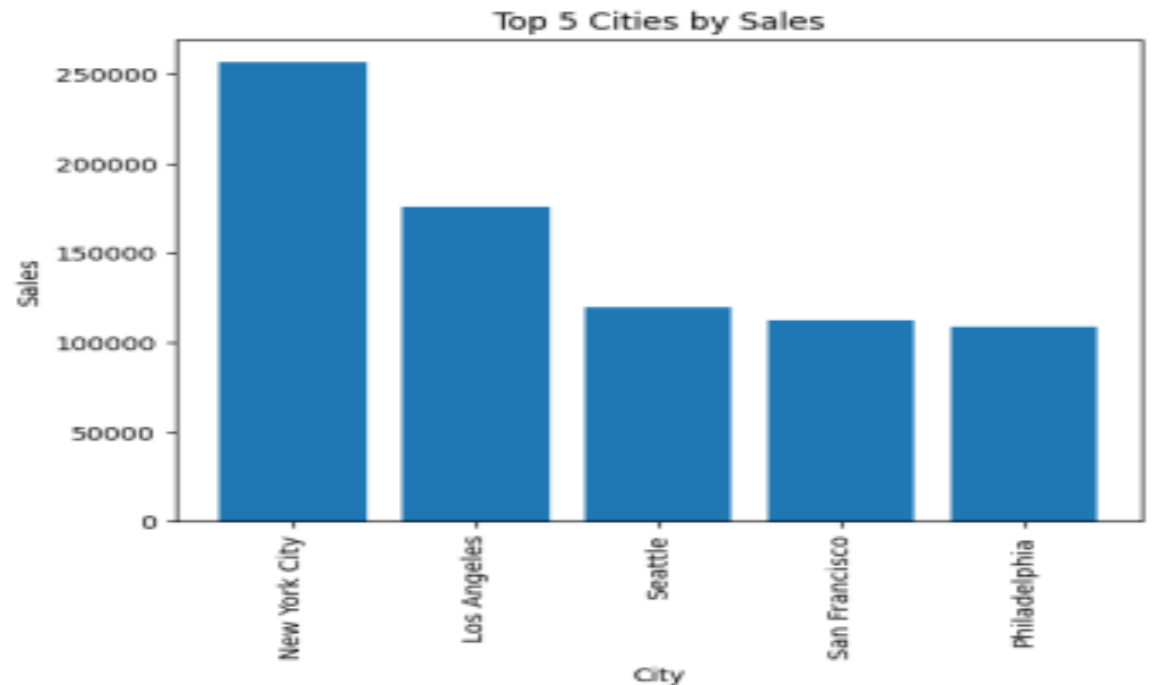
```
city_sales.sort_values(by='Sales', ascending=False, inplace=True)
```

- Select the top 5 cities:

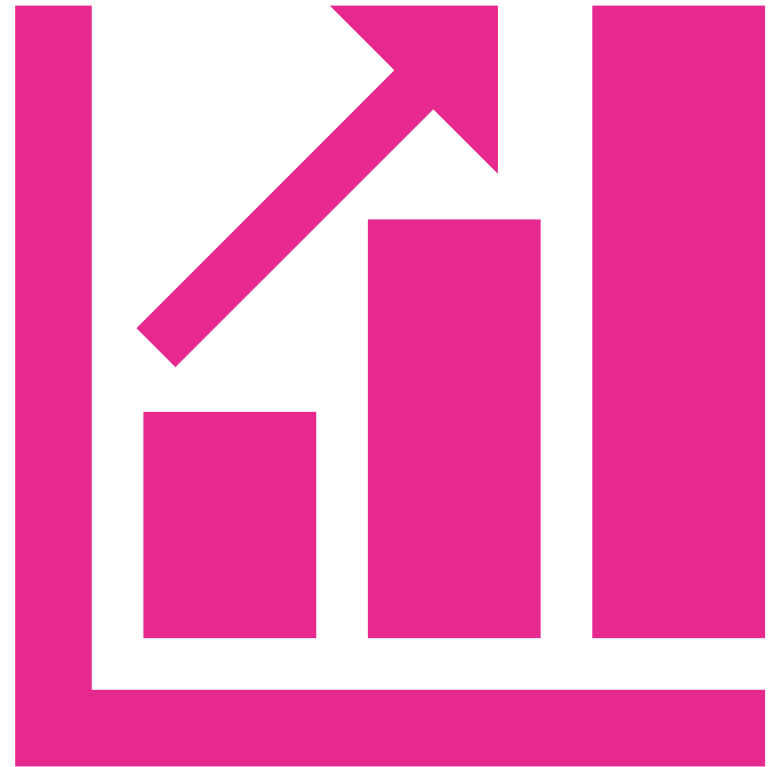
```
top_5_cities_sales = city_sales.head() plt.bar(top_5_cities_sales['City'],  
top_5_cities_sales['Sales'], align='center')
```

- plt.xlabel("City")
- plt.ylabel("Sales") plt.title("Top 5 Cities by Sales")
- plt.xticks(rotation=90)

```
plt.show()  
top_5_cities_sales
```



Select top 5 cities
by profit and Sort
the data by
profit in
descending order



Select top 5 cities by profit and Sort the data by profit in descending order:

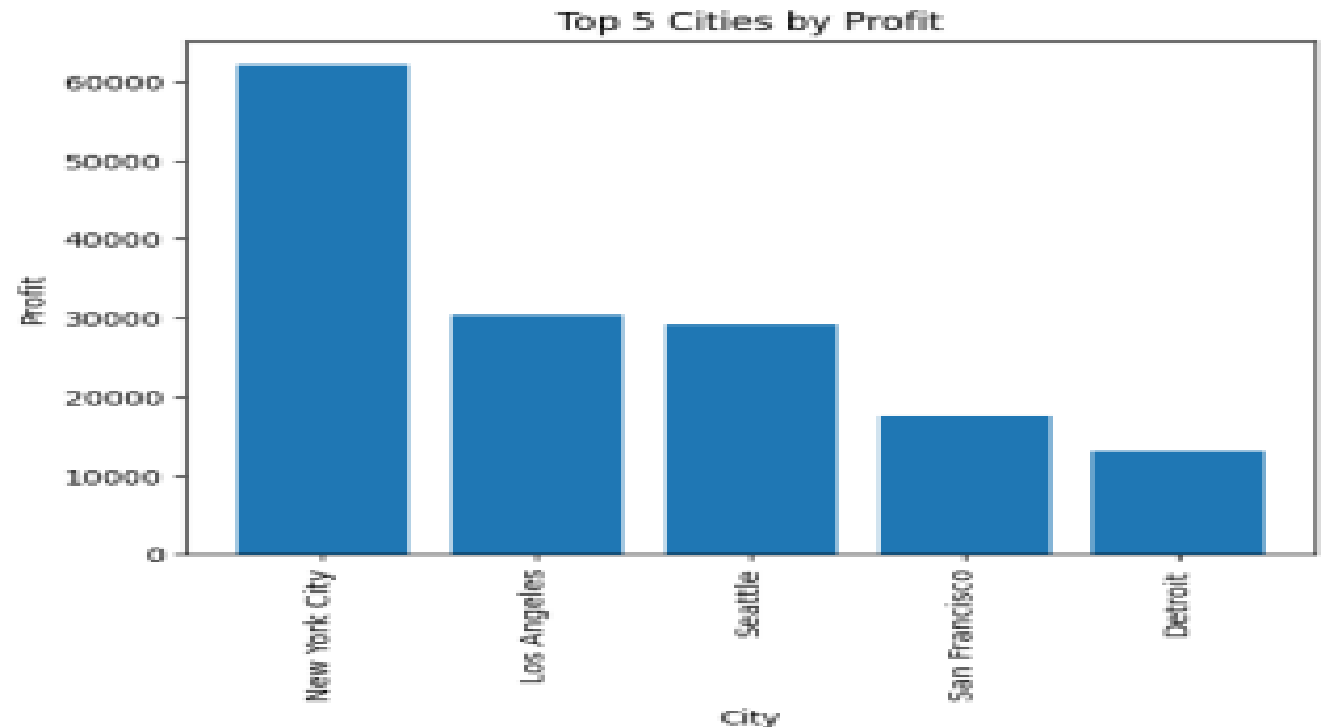
```
city_profit = df_places.groupby('City', as_index=False).sum()
```

```
# Sort the data by Sales in descending order: city_profit.sort_values(by='Profit',  
ascending=False, inplace=True)
```

```
# Select the top 5 cities: top_5_cities_profit = city_profit.head() plt.bar(top_5_cities_profit['City'],  
top_5_cities_profit['Profit'], align='center')
```

```
plt.xlabel("City") plt.ylabel("Profit")  
plt.title("Top 5 Cities by Profit")  
plt.xticks(rotation=90)
```

```
plt.show()  
top_5_cities_profit
```



The best sales

- **# Group the data by product category and calculate the average profit for each category:**

```
avg_profit_margin_by_category = df.groupby('Category')['Profit'].sum()  
print(avg_profit_margin_by_category) df['Profit Margin'] = df['Profit'] / df['Sales']
```

- **# Group the data by product category and calculate the average profit margin for each category:**

```
avg_profit_margin_by_category = df.groupby('Category')['Profit Margin'].mean()
```

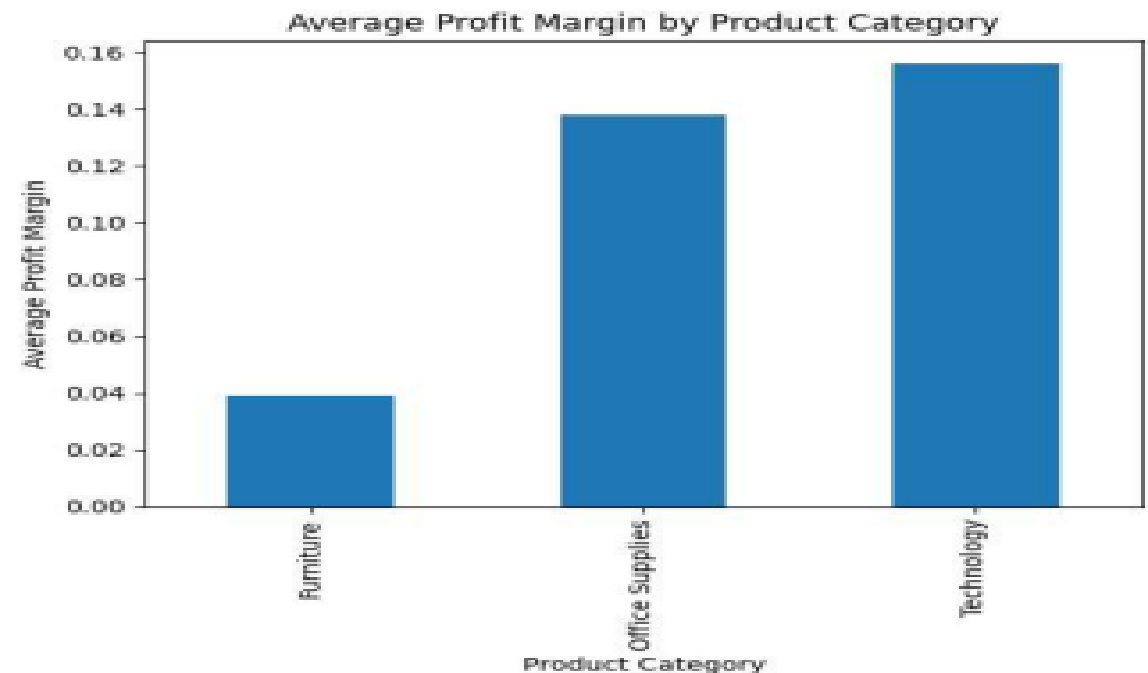
- **# Plot the average profit margin for each category as a bar chart:**

```
avg_profit_margin_by_category.plot(kind='bar')
```

- **# Add a title and labels to the chart:**

```
plt.title("Average Profit Margin by Product Category")  
plt.xlabel("Product Category") plt.ylabel("Average  
Profit Margin")
```

```
plt.show()
```



CONCLUSION

- The analysis of the Superstore dataset has provided valuable insights into sales trends, customer behavior, and operational efficiency
- Through exploratory data analysis and advanced modeling techniques, we have identified several significant findings:
 - Sales Trends: The analysis revealed seasonal patterns, with peak sales occurring during specific months
- Overall, the "Analysis of Superstore dataset" project demonstrates the power of data analytics in uncovering insights that drive strategic decision-making, operational efficiency, and ultimately, the success of the Superstore in a competitive retail market

Moving forward, it is recommended that the Superstore continues to monitor sales performance, customer behavior, and operational metrics. This will allow for ongoing adjustments and improvements based on changing market dynamics and evolving customer preferences.

Overall, the "Analysis of Superstore dataset" project demonstrates the power of data analytics in uncovering insights that drive strategic decision-making, operational efficiency, and ultimately, the success of the Superstore in a competitive retail market

