NAME: AVIK MONDAL

COURSE NAME: M.Tech (Computer Science and Engineering)

COLLEGE NAME: JADAVPUR UNIVERSITY

INTERNSHIP DOMAIN: DATA ANALYTICS (DA) / [12.06.2023-23.07.2023]

SKILLSBUILD EMAIL ID: aviksecondmail@gmail.com

PROJECT TITLE:

CASE STUDY: ANALYSIS OF SUPERSTORE DATASET

Introduction:

The goal of this project is to conduct a comprehensive analysis of the Sample Superstore dataset to gain valuable insights into Sales trends and Profitability of the store.

- **Sales Analysis:** Analyze sales metrics, trends, and factors influencing sales fluctuations.
- **Profit Analysis:** Analyze the profit and factors affecting profit of various items in the store.

The Superstore dataset encompasses a wide range of information, including sales data, customer demographics, product categories, and geographical regions. By leveraging this dataset, our objective is to identify areas for improvement and provide data-driven recommendations to optimize the performance of the store. The purpose of this report is to present our findings and recommendations based on the analysis of this dataset.

AGENDA

- Introduction to the Project- Analysis of Superstore Dataset
- Project Overview- Purpose, Scope and Objective
- End Users
- Solution and Value Proposition
- Customization of the Project & Code Snippets
- Modelling & Insights
- Results of Analysis – Key Insights, Actionable Insights, Customized Solutions
- Relevant Links

PROJECT OVERVIEW

PURPOSE:

The purpose of the project is to perform **Descriptive Data Analysis on Superstore data** to gather relevant insights regarding sales and profit of the superstore-

- Extract, Transform and Load (ETL) the data
- Perform Exploratory Data Analysis (EDA) on the dataset

SCOPE:

The project focused on understanding sales trends and profitability across regions, cities, categories, and sub-categories within the Superstore dataset.

- **Goal:**
  - Perform Descriptive Analysis of the 'SampleSuperstore' Dataset using Python
  - Highlight the insights regarding Sales and Profit of the Superstore as gathered from analysing the data.
- **Tools Used:** Python and libraries- Numpy, Pandas, Matplotlib and Seaborn

PROJECT OVERVIEW

OBJECTIVES:

- Understand, clean and visualize the dataset to gain insights from the data
- Analyze the Sales and Profit based on different regions, categories and other parameters, that is-
  - Analyze sales patterns in different regions and cities to identify the highest-selling areas.
  - Determine the top-selling categories and sub-categories within the Superstore dataset.
  - Assess the profitability of different products and identify the most profitable areas.
  - Find the Cities, States and Regions having maximum Profit and maximum Sales.
- Provide data-driven insights and recommendations for optimizing sales and improving profitability.

END USERS

- **Store managers and executives:** The insights from the analysis can guide data-driven decision-making and strategic planning to optimize store operations.

- **Sales and marketing teams:** The analysis provides valuable information for designing effective sales and marketing strategies based on customer demographics, preferences and buying patterns to develop targeted marketing campaigns and improve customer engagement.

- **Financial analysts and stakeholders:** The analysis offers insights into profit margins and can enhance financial decision-making for stakeholders of the superstore.

SOLUTION AND ITS VALUE PROPOSITION

Solution: The project involved a comprehensive analysis of the Superstore dataset to gain insights into sales trends and profitability of the Superstore. This analysis utilized various Python libraries like NumPy, Pandas, statistical and data mining techniques and visualization tools such as Matplotlib and Seaborn to efficiently analyze the Superstore dataset.

Value Proposition: The analysis provides data-driven recommendations for decision making to -

- optimize sales,
- improve profitability, and
- inform business strategies for personalized marketing.

The use of Python libraries and visualization tools enhances the efficiency and effectiveness of the analysis by making it easily interpretable.

CUSTOMIZATION OF PROJECT & CODE SNIPPETS

- The analysis was tailored to the specific requirements of the Superstore dataset.

- **Use of advanced Visualization Tools:**
  In this project, I have used advanced visualization libraries like Matplotlib and Seaborn to emphasize on the uniqueness of my solution by presenting it in a visually appealing manner such that it provides clear understanding of the insights to the end users.

- **Descriptive Analysis & EDA:**
  This project also utilizes the various techniques of Descriptive Analytics and Exploratory Data Analysis to summarize and present the key insights regarding Sales and Profit of the Superstore.

CUSTOMIZATION OF PROJECT & CODE SNIPPETS

Code Snippets: Here are a few code snippets from the project to demonstrate the data loading, cleaning, and transformation processes. These snippets showcases the use of Python libraries such as Pandas, NumPy, and Matplotlib for data manipulation and visualization in this project.

UNDERSTANDING DATA:

```
# Viewing the file headers to derive a primary meaning of the data
df.head()
```

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

DATA CLEANING:

Since we need to analyze the Sales of the Superstore, the column named Postal Code is redundant. Hence, we drop that column.

```
# Drop redundant columns
df.drop(columns="Postal Code", inplace=True)
df.head()
```

	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

```
[9] # Checking unique values for the column "Country"
    df['Country'].unique()
```

```
array(['United States'], dtype=object)
```

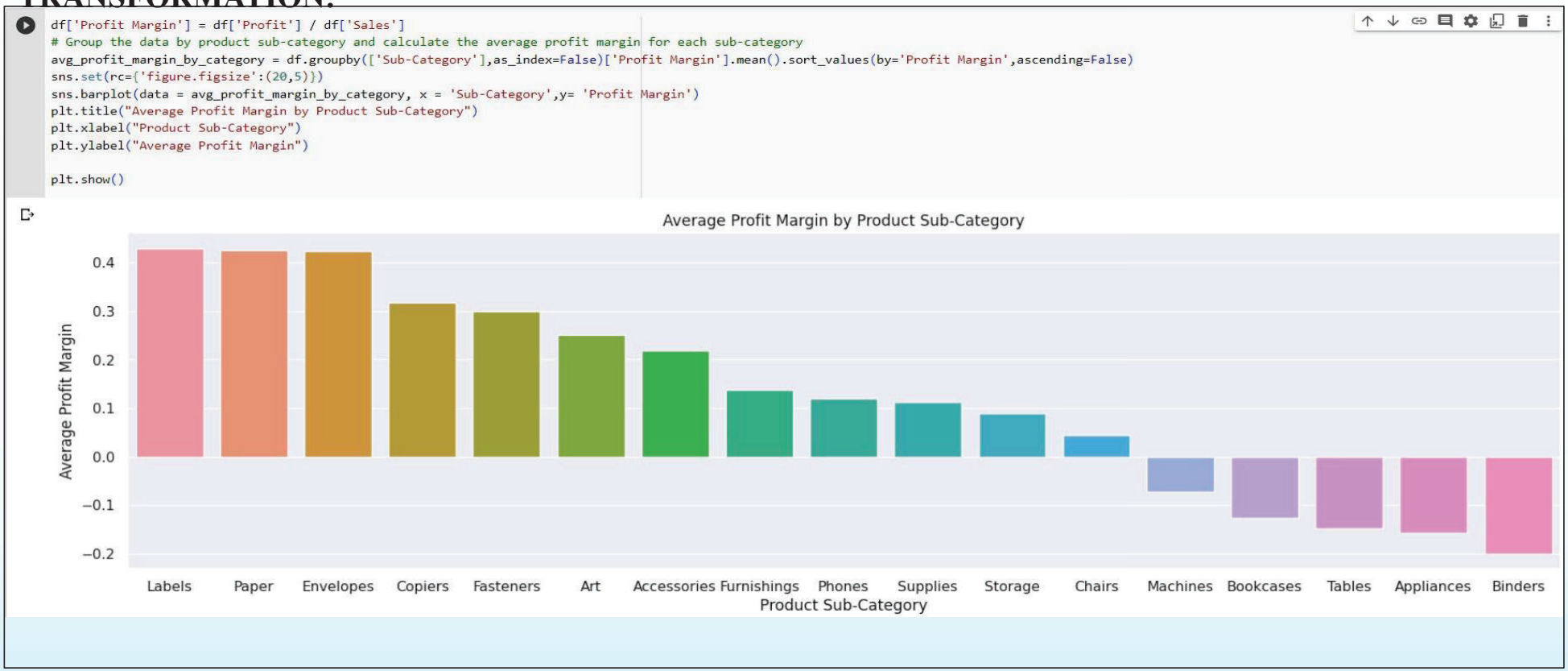
Since the entire dataset contains Superstore data of United States alone, thus including the 'Country' column is redundant while analyzing the dataset.

CUSTOMIZATION OF PROJECT & CODE SNIPPETS

DATA TRANSFORMATION:

```
df['Profit Margin'] = df['Profit'] / df['Sales']
# Group the data by product sub-category and calculate the average profit margin for each sub-category
avg_profit_margin_by_category = df.groupby(['Sub-Category'],as_index=False)['Profit Margin'].mean().sort_values(by='Profit Margin',ascending=False)
sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = avg_profit_margin_by_category, x = 'Sub-Category',y= 'Profit Margin')
plt.title("Average Profit Margin by Product Sub-Category")
plt.xlabel("Product Sub-Category")
plt.ylabel("Average Profit Margin")

plt.show()
```



MODELLING AND INSIGHTS

The modelling techniques used for data analysis in this project are –

- **Statistical Analysis:**

Used to discover correlations, trends, and patterns within the Superstore dataset.

These techniques helped in understanding the impact of various factors on Sales and Profitability.

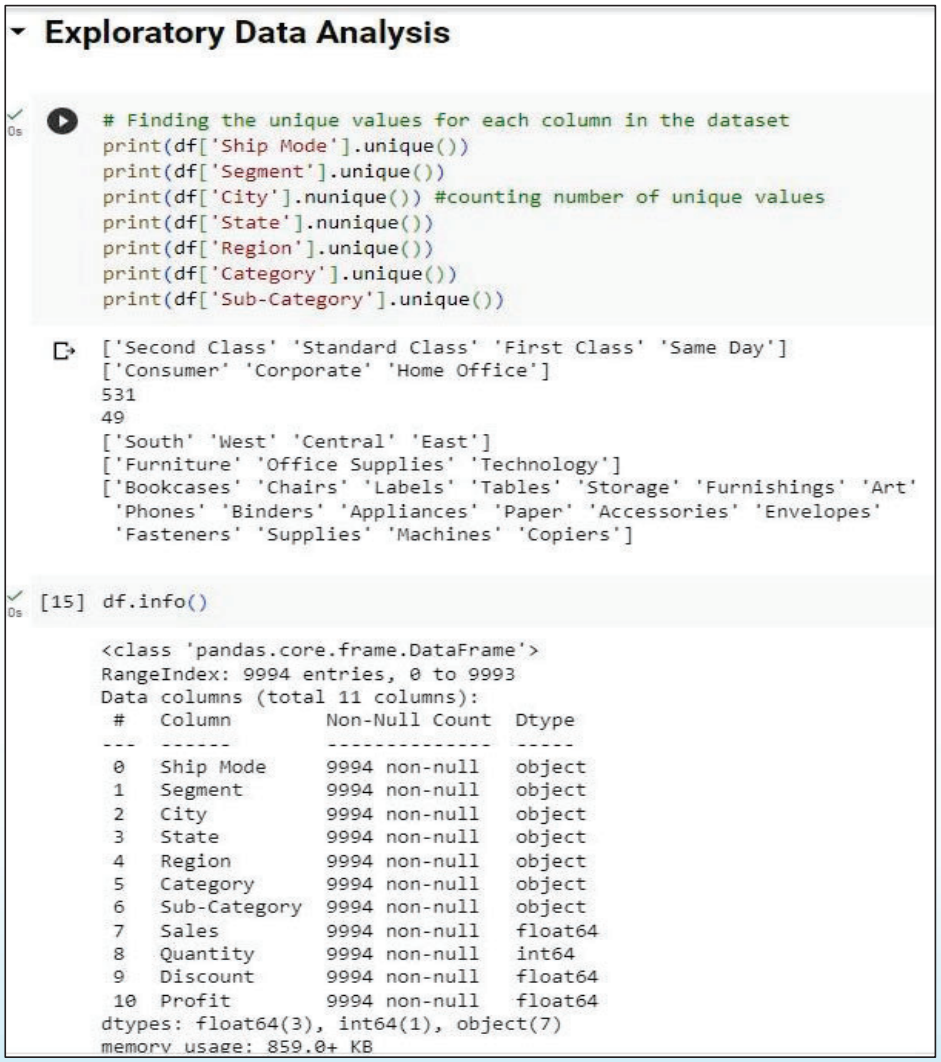
Statistical Analysis:

```
[7] df.describe()
```

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

MODELLING AND INSIGHTS

- **Exploratory Data Analysis (EDA):** EDA techniques were employed to gain initial insights into the dataset. This included data visualization through charts, graphs, and plots to understand the distribution of variables, identify outliers, and detect patterns or relationships between different variables.



Exploratory Data Analysis

```
# Finding the unique values for each column in the dataset
print(df['Ship Mode'].unique())
print(df['Segment'].unique())
print(df['City'].nunique()) #counting number of unique values
print(df['State'].nunique())
print(df['Region'].unique())
print(df['Category'].unique())
print(df['Sub-Category'].unique())
```

```
['Second Class' 'Standard Class' 'First Class' 'Same Day']
['Consumer' 'Corporate' 'Home Office']
531
49
['South' 'West' 'Central' 'East']
['Furniture' 'Office Supplies' 'Technology']
['Bookcases' 'Chairs' 'Labels' 'Tables' 'Storage' 'Furnishings' 'Art'
 'Phones' 'Binders' 'Appliances' 'Paper' 'Accessories' 'Envelopes'
 'Fasteners' 'Supplies' 'Machines' 'Copiers']
```

```
[15] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Ship Mode     9994 non-null   object
 1   Segment       9994 non-null   object
 2   City          9994 non-null   object
 3   State         9994 non-null   object
 4   Region        9994 non-null   object
 5   Category      9994 non-null   object
 6   Sub-Category  9994 non-null   object
 7   Sales         9994 non-null   float64
 8   Quantity      9994 non-null   int64
 9   Discount      9994 non-null   float64
 10  Profit        9994 non-null   float64
dtypes: float64(3), int64(1), object(7)
memory usage: 859.0+ KB
```

MODELLING AND INSIGHTS

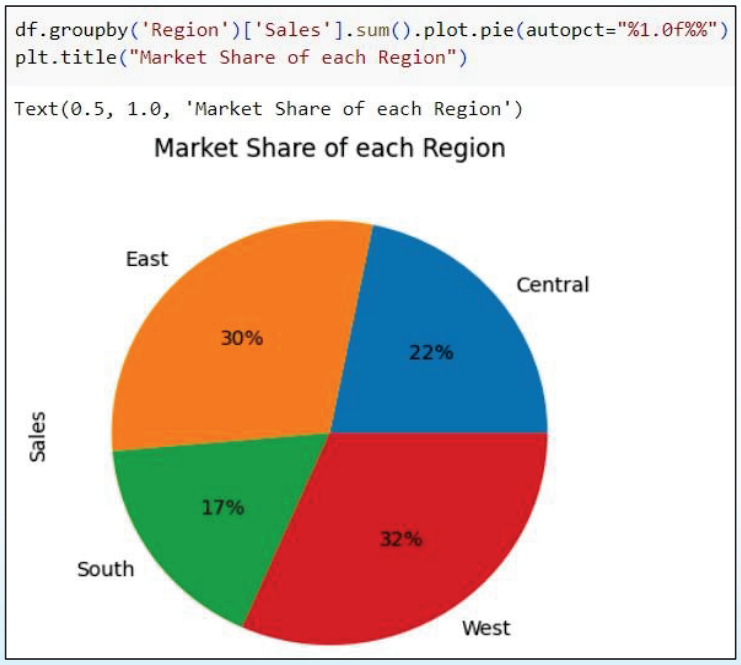
- **Exploratory Data Analysis (EDA):**

This comprises mainly two basis of analysis –

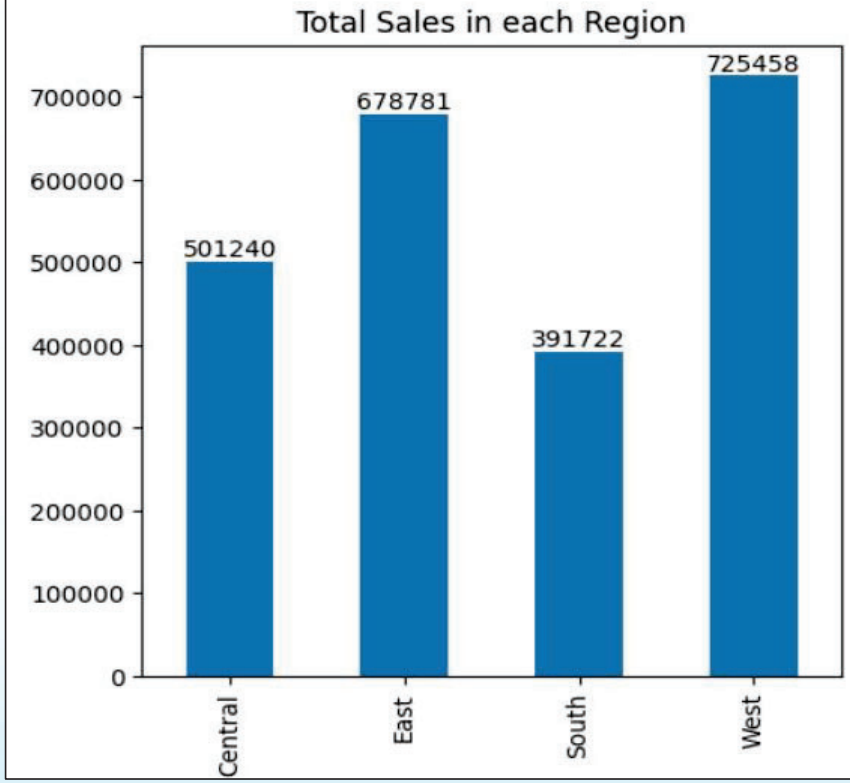
- **Sales Analysis:**

```
df.groupby('Region')['Sales'].sum().plot.pie(autopct="%1.0f%%")
plt.title("Market Share of each Region")
```

```
Text(0.5, 1.0, 'Market Share of each Region')
```

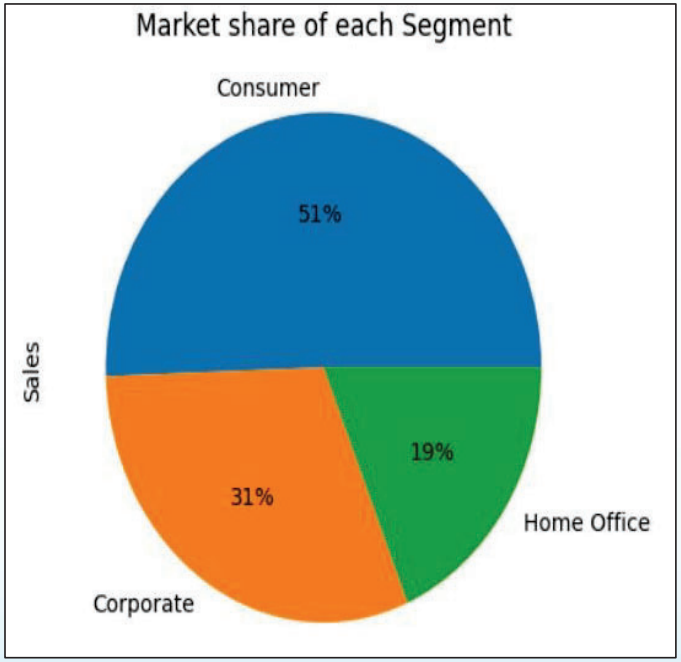



```
# Sales Analysis based on region
sales_region = df.groupby('Region')['Sales'].sum().plot.bar(figsize=(5,5))
plt.title("Total Sales in each Region")
for container in sales_region.containers:
    sales_region.bar_label(container)
```

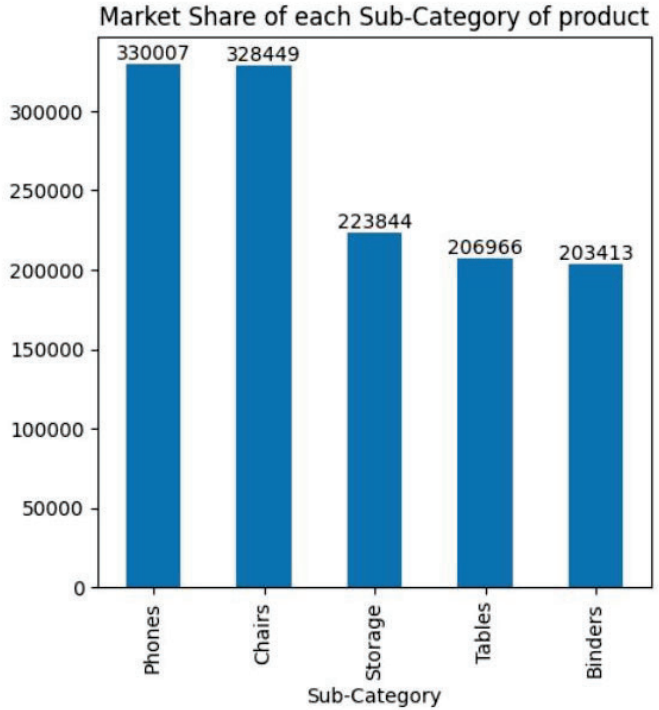


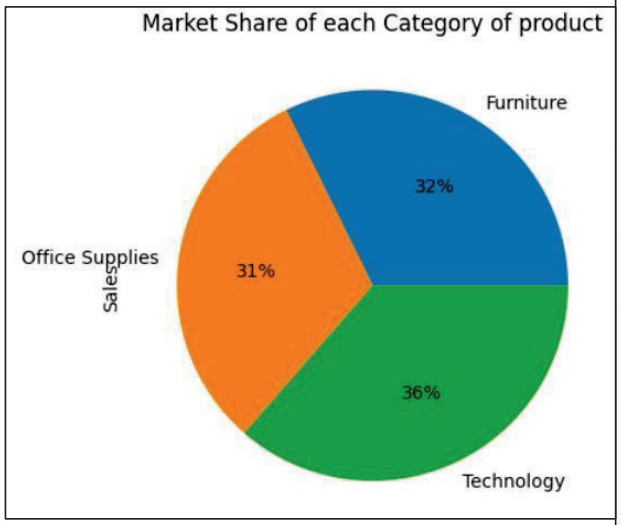
MODELLING AND INSIGHTS

- **Exploratory Data Analysis (EDA):**
  - **Sales Analysis:**



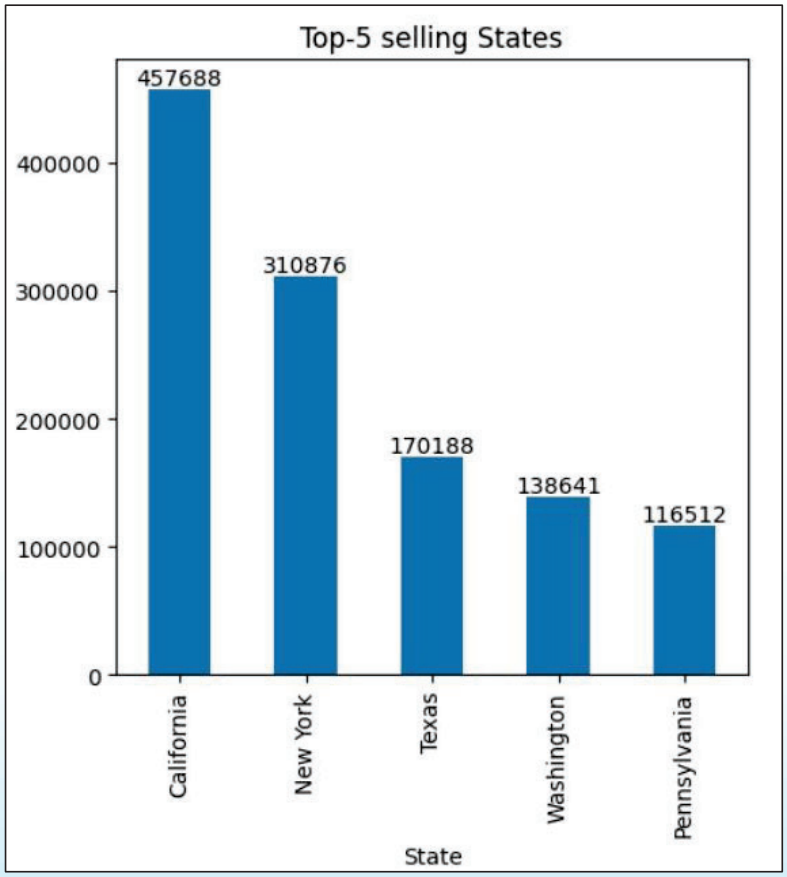

```
# Sales Analysis based on sub-category
sales_subc = df.groupby('Sub-Category')['Sales'].sum().sort_values(ascending=False).head(5).plot.bar(figsize=(5,5))
plt.title("Market Share of each Sub-Category of product")
for container in sales_subc.containers:
    sales_subc.bar_label(container)
```

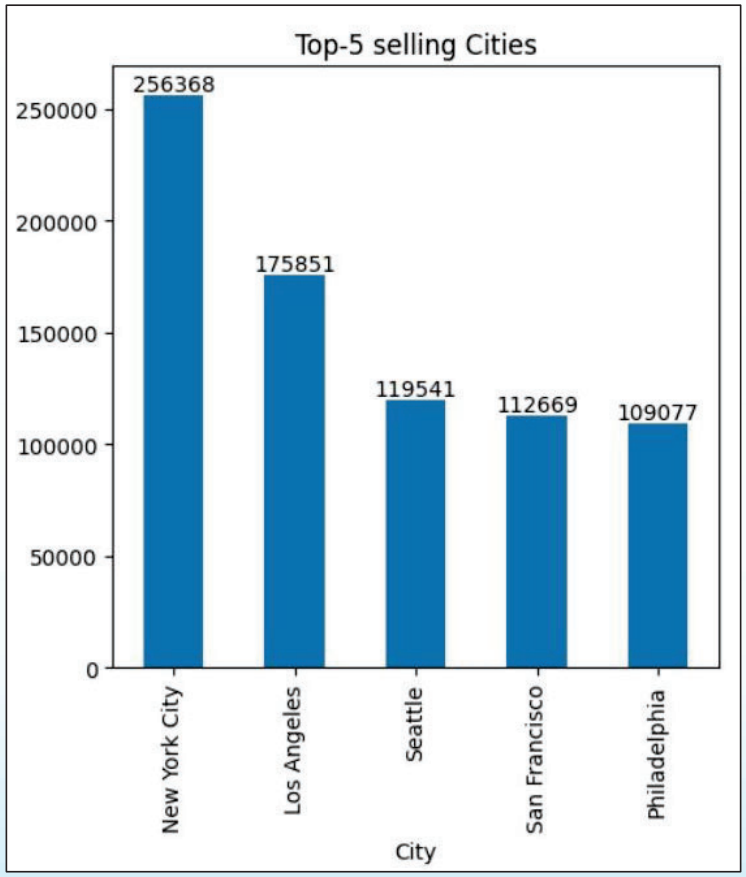




MODELLING AND INSIGHTS

- **Exploratory Data Analysis (EDA):**
  - **Sales Analysis:**



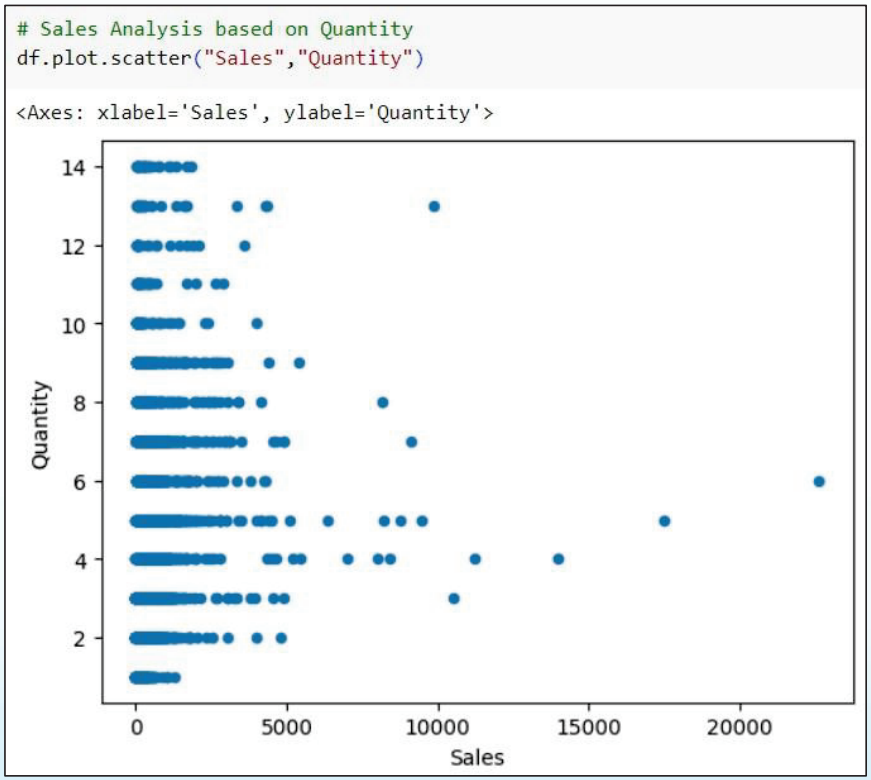


MODELLING AND INSIGHTS

- **Exploratory Data Analysis (EDA):**
  - **Sales Analysis:**

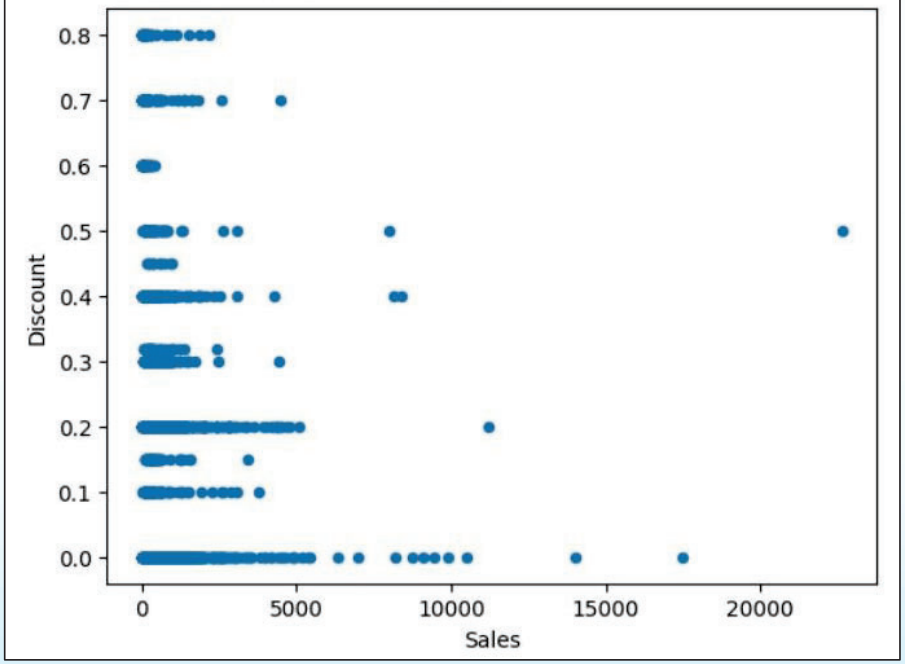
```
# Sales Analysis based on Quantity
df.plot.scatter("Sales","Quantity")
```

```
<Axes: xlabel='Sales', ylabel='Quantity'>
```




```
# Sales Analysis based on Discount
df.plot.scatter("Sales","Discount")
```

```
<Axes: xlabel='Sales', ylabel='Discount'>
```



MODELLING AND INSIGHTS

Exploratory Data Analysis (EDA):

❖ Sales Analysis Insights:

Sales Insights

- The **West** region contributes to the maximum amount (32%) of sales, followed by the **East** (30%). **South** has the **minimum** sales.
- The majority of the Sales opts for the shipping mode **Standard Class**
- The maximum sales occurs in the **Consumer** Segment (51%), followed by **Corporate** (31%)
- Highest Sales occurs in the **New York City**, followed by **Los Angeles**
- The State having the highest sales is **California**, followed by **New York**
- The maximum sales occurs in the **Technology** category (36%) in the Sub-Category of **Phones**, followed by **Chairs** of the category **Furniture**. Furniture and Office supplies have nearly equal share of sales.
- Sales does not depend upon Discount and Quantity significantly, since the coefficient of correlation between them is negligible (~0). However, Sales is directly dependent on Profit, which is quite obvious. Higher the Sales, more is the Profit generated.

Conclusion:

1. The **West** Region has the maximum amount of Sales followed by the **East**
2. The city of **Los Angeles** in the State of **California** in the **West** Region of the United States and **New York City** in **New York** in the **East** contributes to the maximum amount of Sales.
3. The maximum Sales is encountered in the Sub-Category of **Phones** in the Category **Technology**, followed by **Chairs** in the Category **Furniture**. However, Furniture and Office supplies have nearly equal share of Sales.

MODELLING AND INSIGHTS

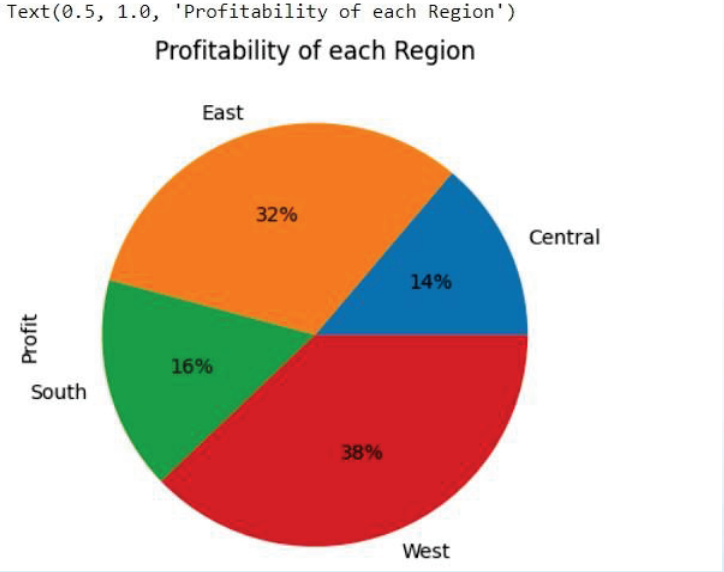
- **Exploratory Data Analysis (EDA):**

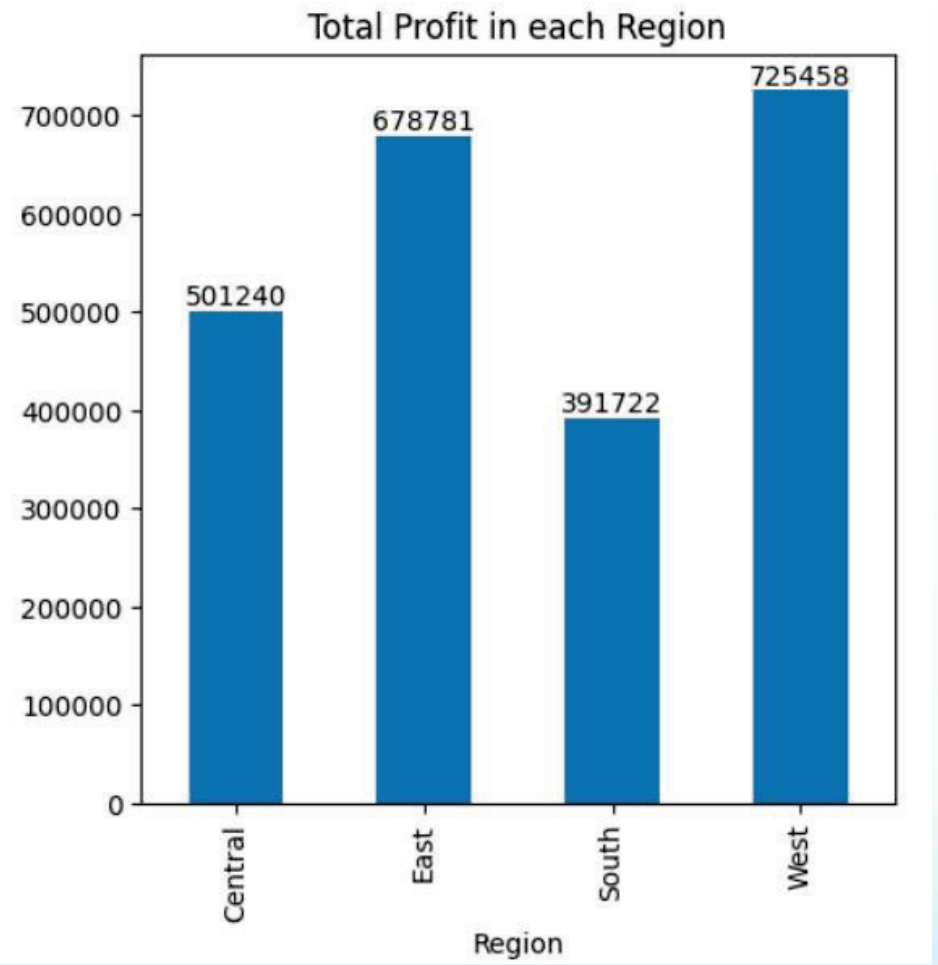
This comprises of mainly two basis of analysis –

- **Profit Analysis:**

```
df.groupby('Region')['Profit'].sum().plot.pie(autopct="%1.0f%%")
plt.title("Profitability of each Region")
```

```
Text(0.5, 1.0, 'Profitability of each Region')
```



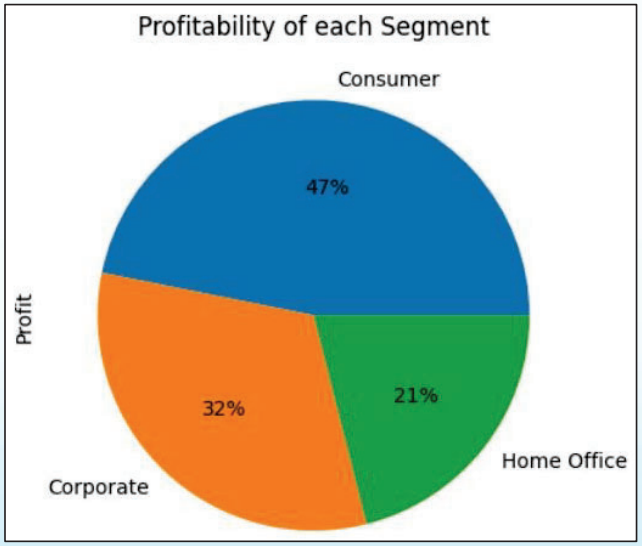


MODELLING AND INSIGHTS

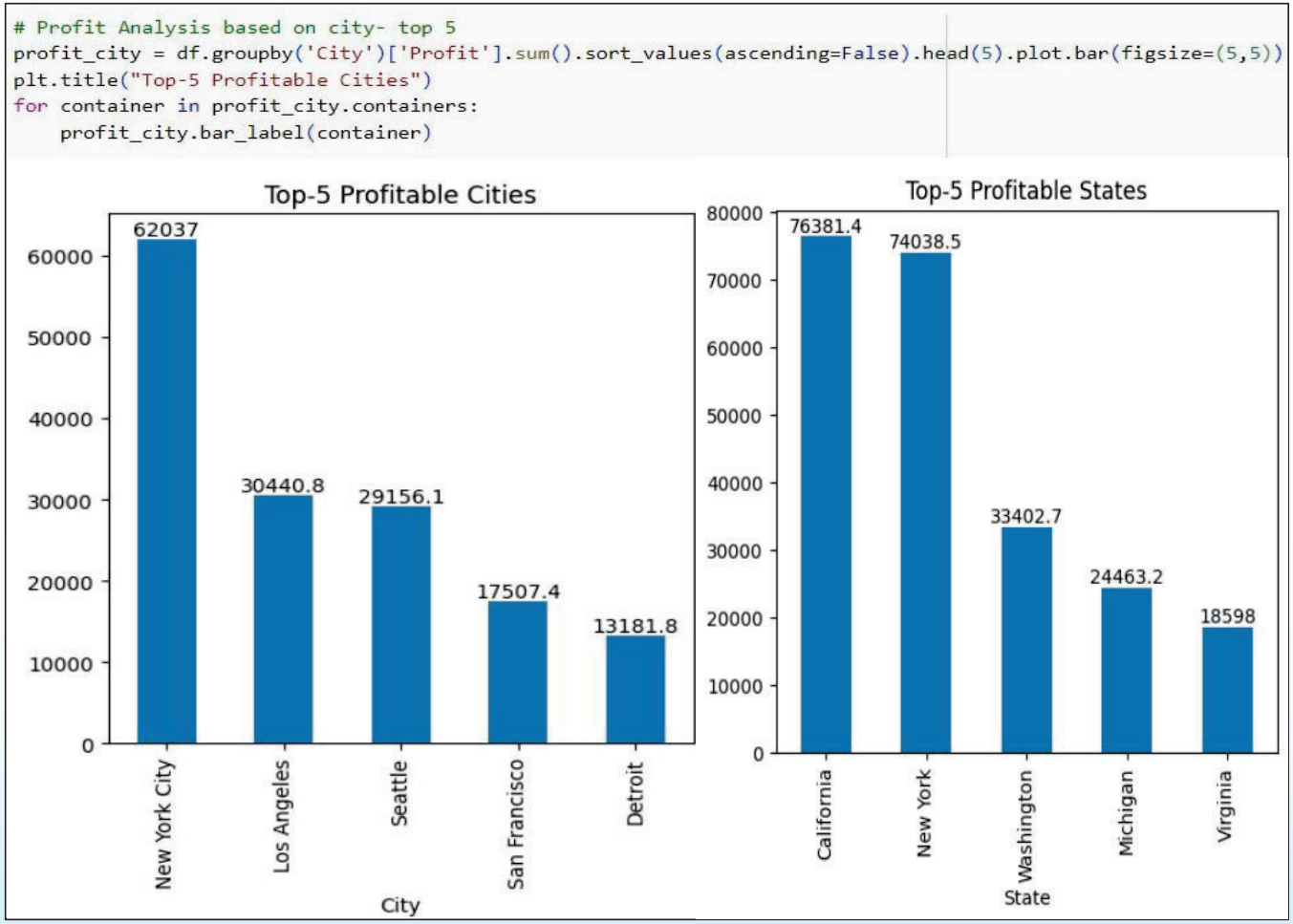
- **Exploratory Data Analysis (EDA):**

This comprises of mainly two basis of analysis –

- **Profit Analysis:**




```
# Profit Analysis based on city- top 5
profit_city = df.groupby('City')['Profit'].sum().sort_values(ascending=False).head(5).plot.bar(figsize=(5,5))
plt.title("Top-5 Profitable Cities")
for container in profit_city.containers:
    profit_city.bar_label(container)
```

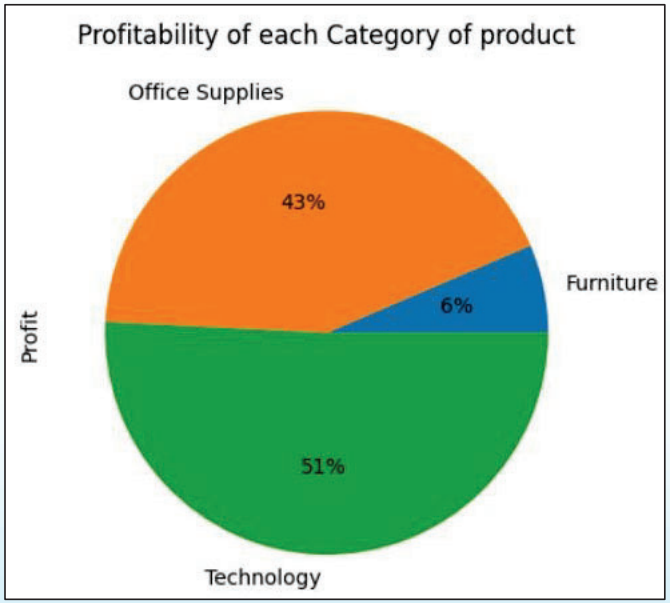


MODELLING AND INSIGHTS

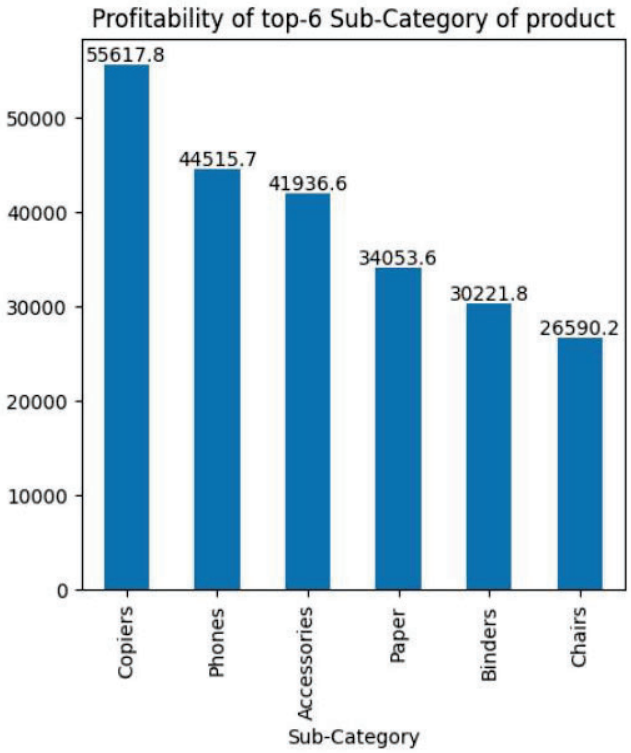
Exploratory Data Analysis (EDA):

This comprises of mainly two basis of analysis –

- **Profit Analysis:**




```
# Profit Analysis based on sub-category
profit_subc = df.groupby('Sub-Category')['Profit'].sum().sort_values(ascending=False).head(6).plot.bar(figsize=(5,5))
plt.title("Profitability of top-6 Sub-Category of product")
for container in profit_subc.containers:
    profit_subc.bar_label(container)
```



MODELLING AND INSIGHTS

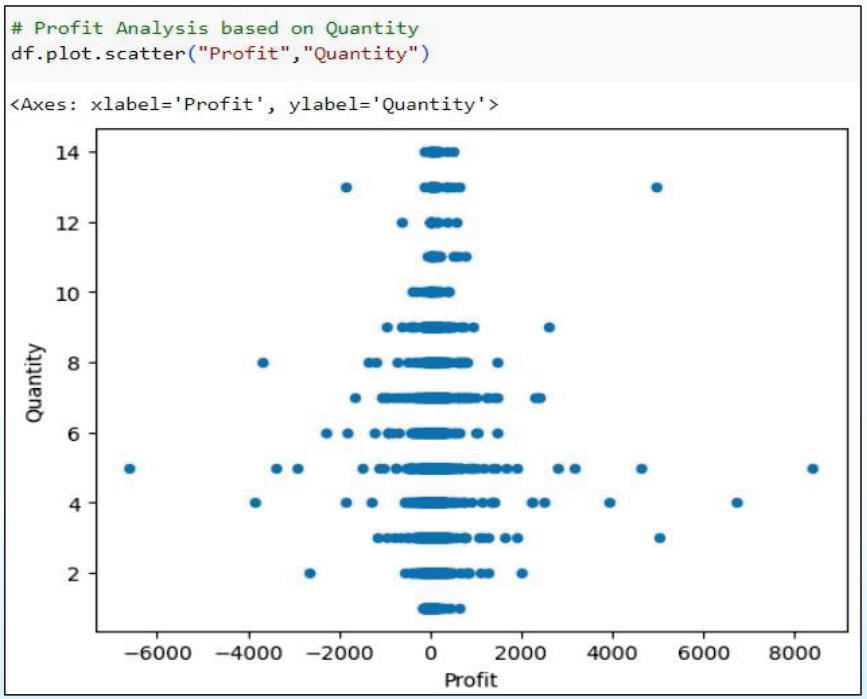
- **Exploratory Data Analysis (EDA):**

This comprises of mainly two basis of analysis –

- **Profit Analysis:**

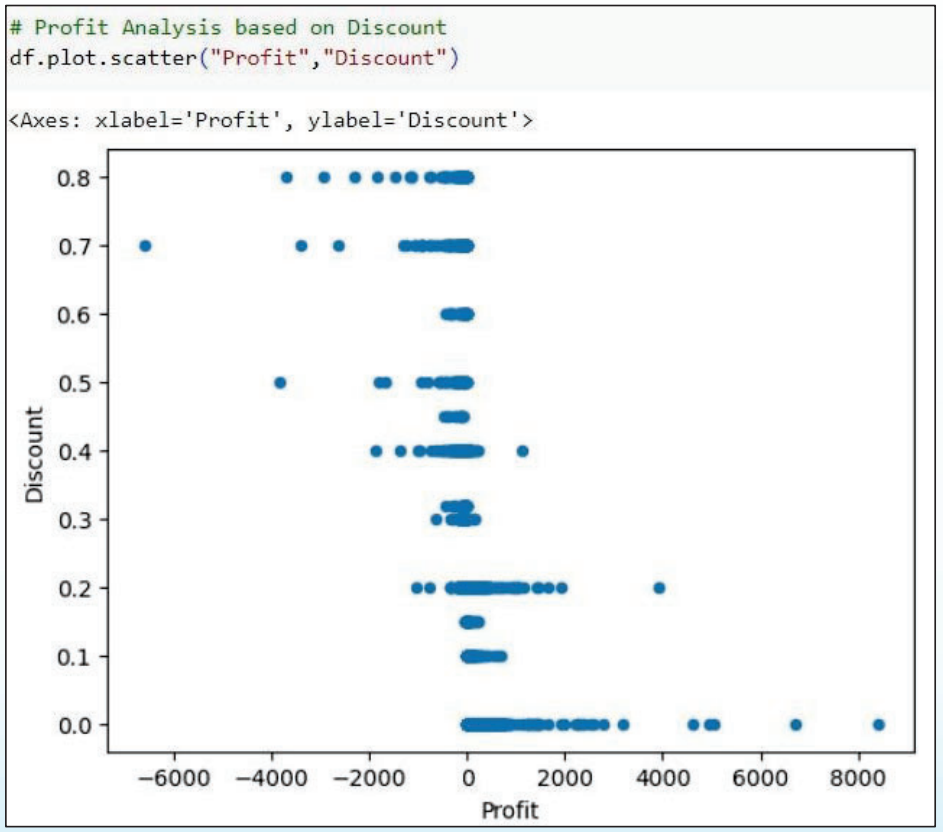
```
# Profit Analysis based on Quantity
df.plot.scatter("Profit","Quantity")
```

```
<Axes: xlabel='Profit', ylabel='Quantity'>
```




```
# Profit Analysis based on Discount
df.plot.scatter("Profit","Discount")
```

```
<Axes: xlabel='Profit', ylabel='Discount'>
```



MODELLING AND INSIGHTS

Exploratory Data Analysis (EDA):

❖ Profit Analysis Insights:

Profit Insights

- The **West** region contributes to the maximum amount (38%) of profit, followed by the **East** (32%)
- The majority of the Profit is earned from sales which opts for the shipping mode **Standard Class**
- The maximum profit is obtained from the **Consumer** Segment (47%), followed by **Corporate** (32%)
- Highest Profit is earned in the **New York City**, followed by **Los Angeles** and **Seattle**, which have nearly equal total profit earned.
- The State having the highest profit is **California** and **New York**, having nearly equal share of Profit
- The maximum Profit is incurred from the **Technology** category (51%) in the Sub-Category of **Copiers** and **Phones**, followed by the category **Office Supplies** (43%).**Furniture** yields the **least profit** (6%)
- Profit does not depend upon Discount(related inversely) and Quantity, since the coefficient of correlation between them is negligible (~0). However, Profit is directly dependent on Sales, which is quite obvious. Higher the Sales, more is the Profit generated.

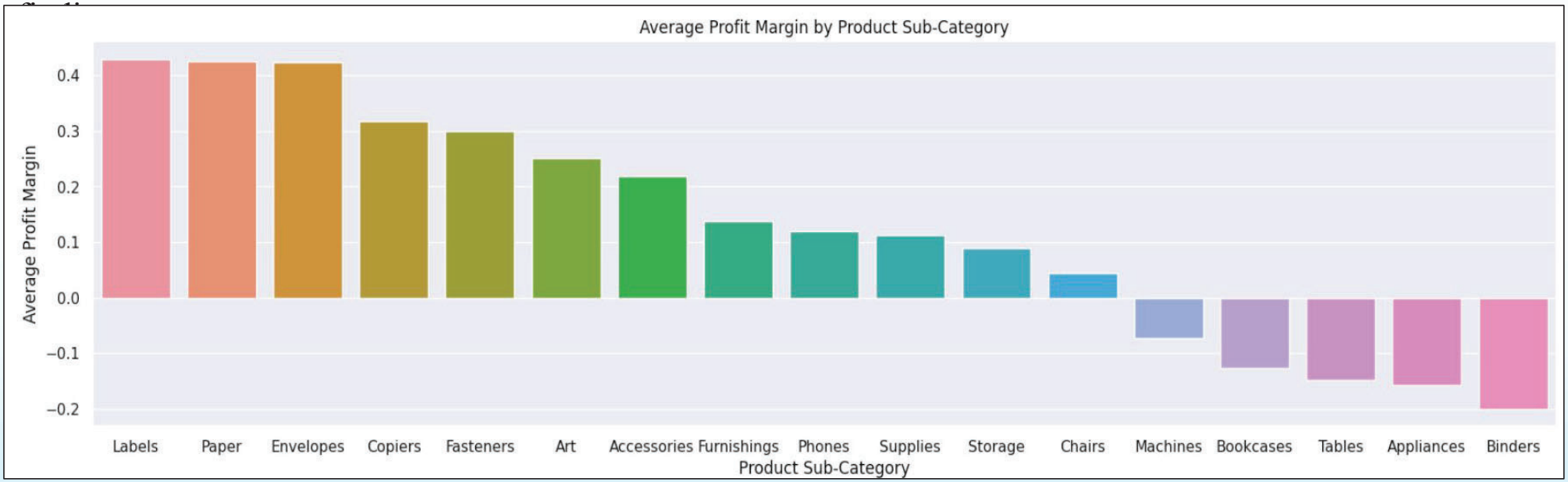
Conclusion:

1. The **West** Region contributes to the maximum amount of Profit followed by the **East**
2. The city of **Los Angeles** in the State of **California**, followed by **Seattle** in **Washington** in the **West** Region of the United States and **New York City** in **New York** in the **East** contributes to the maximum amount of Profit.
3. The maximum Profit is earned in the Sub-Category of **Copiers**, followed by **Phones** in the Category **Technology**, followed by the Category of **Office Supplies**. Furniture yields the least amount of Profit, although its sales is high.

MODELLING AND INSIGHTS

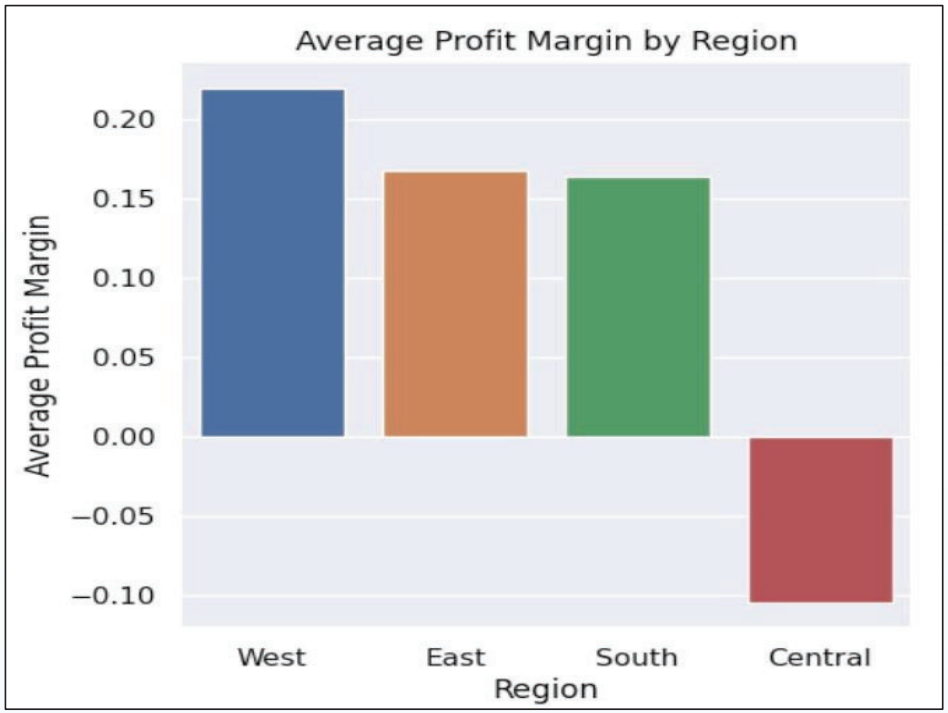
- **Data Visualization:**

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



MODELLING AND INSIGHTS

These modelling techniques formed the pillars of the Project – **Analysis of Superstore Dataset** for Data Analytics, ensuring a systematic and data-driven approach to extract valuable insights from the dataset.




```
df['Profit Margin'] = df['Profit'] / df['Sales']
# Group the data by region and calculate the average profit margin for each region
avg_profit_margin_by_category = df.groupby(['Region'],as_index=False)['Profit Margin'].mean().sort_values(by='Profit Margin',ascending=False).head(50)
sns.set(rc={'figure.figsize':(5,5)})
sns.barplot(data = avg_profit_margin_by_category, x = 'Region',y= 'Profit Margin')
plt.title("Average Profit Margin by Region")
plt.xlabel("Region")
plt.ylabel("Average Profit Margin")
plt.show()
```

RESULTS OF ANALYSIS

- **Key Insights:** The key findings of the Sales and Profit Analysis are summarized, including the highest-selling regions, cities, and sub-categories, as well as the most profitable areas –

- Sale of products in the **Technology** Category results in **Maximum Profit Margin**, more specifically **Copiers, Fasteners, Accessories and Phones**, followed by **Office Supplies** including **Labels, Papers and Envelopes**. **Chairs** are the only product in the **Furniture** Category which is **profitable**, other products of this category results in **Loss**
- The **Segment- Home Office** is the most profitable, followed by **Corporate**
- Sales in the **West** Region has the Highest **Profitability**, followed by the **East** while that in the **Central** Region suffers the Highest **Loss**

RESULTS OF ANALYSIS

- **Actionable Insights:** Actionable insights derived from the analysis are presented, highlighting the areas where optimization opportunities exists for sales improvement and profitability enhancement –

- To ensure **Maximum Profit,** the production in the **Technology** Sector must be upgraded so that the products mentioned above are available to the customers in required quantity so as to always remain ahead of the Market Demand.
- Suitable actions must be taken **ensure Profit** in the **Furniture** Category for products other than Chairs or else the cost of production of those commodities must be reduced.
- **More products** of the Segment- **Home Office** must be produced so as to ensure abundant supply and **maximum profit** of the Superstore.
- The Sales in the **Central** Region of the United States must be **inspected for the cause of overall Loss** and suitable steps must be taken to rectify the same to ensure Profit.

RESULTS OF ANALYSIS

- **Customized Solutions:** The presentation emphasizes on the significance of customized solutions based on data-driven analysis to achieve better business outcomes –

- The city of **Los Angeles** in the State of **California** in the **West** Region of the United States and **New York City** in **New York** in the **East** contributes to the **maximum amount of Sales**. Thus, proper infrastructure, machinery and production must be especially ensured in the Superstore in these cities to facilitate higher Sales of Profit-earning commodities.

- The city of **Los Angeles** in the State of **California**, followed by **Seattle** in **Washington** in the **West** Region of the United States and **New York City** in **New York** in the **East** contributes to the **maximum amount of Profit**. Thus, higher production must be ensured for the Superstores in these areas so as to ensure Maximum Profit from the sale of products here.

RELEVANT LINKS

- **GitHub Repository Link:**

  https://github.com/JU-Avik/IBM-Internship-Project-for-Data-Analytics

- **Dataset Link:**

  https://www.kaggle.com/datasets/bravehart101/sample-supermarket-dataset

- **Source Code (Google Colab):**

  https://colab.research.google.com/drive/1LjXR1NJv9cwph61yWCGNNdaAZE0iXeaR?usp=sharing

…THANK YOU