In [ ]:

Part A: Superstore Sales Analysis

Aim and Objectives

Aim:

To use the Superstore dataset to investigate sales and profitability that will inform business development for the firm.

Objectives:

For the purpose of identifying patterns of changes of the sales, profit and customer segmentation. To establish correlation between discounts and profitability which exists in different product types. For the purposes of defining high-performing geographical markets and customers, on which high-level strategic prescriptions should be applied.

Tool(s) and Techniques

Tools:

1. Pandas: For data manipulation and analysis.
2. Seaborn: For creating high-level visualizations.
3. Matplotlib: For plotting customized visualizations.

Techniques:

Data cleaning and preprocessing. Exploratory Data Analysis (EDA) using descriptive statistics and visualizations. Aggregation and grouping to derive actionable insights.

Data Collection

Dataset:

Name: Superstore Dataset

Source: Kaggle

Justification: The dataset contains comprehensive transactional data, including sales, profit, discount, and customer segments, making it ideal for profitability and performance analysis.

Data Processing

Descriptive Statistics: Summary statistics of sales and profit were derived using .describe(). New Features: Added a Cost column (Sales - Profit) and extracted Year, Month, and Day from the Order Date column for trend analysis. Date Conversion: Converted Order Date and Ship Date to datetime format using pd.to_datetime.

Import Libraries

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading Data: The dataset was loaded using Pandas.

In [ ]:
```
df = pd.read_csv("Superstore.csv");
```

Provide Dataframe Information

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10194 entries, 0 to 10193
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Row ID          10194 non-null  int64
 1   Order ID        10194 non-null  object
 2   Order Date      10194 non-null  object
 3   Ship Date       10194 non-null  object
 4   Ship Mode       10194 non-null  object
 5   Customer ID     10194 non-null  object
 6   Customer Name   10194 non-null  object
 7   Segment         10194 non-null  object
 8   Country/Region  10194 non-null  object
 9   City            10194 non-null  object
 10  State/Province  10194 non-null  object
 11  Postal Code     10194 non-null  object
 12  Region          10194 non-null  object
 13  Product ID      10194 non-null  object
 14  Category        10194 non-null  object
 15  Sub-Category    10194 non-null  object
 16  Product Name    10194 non-null  object
 17  Sales           10194 non-null  float64
 18  Quantity        10194 non-null  int64
 19  Discount        10194 non-null  float64
 20  Profit          10194 non-null  float64
dtypes: float64(3), int64(2), object(16)
memory usage: 1.6+ MB
```

Dataframe Heading

In [ ]:
```
df.head()
```

Out[ ]:

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/I
0	1	US-2021-103800	03-01-2021	07-01-2021	Standard Class	DP-13000	Darren Powers	Consumer	United
1	2	US-2021-112326	04-01-2021	08-01-2021	Standard Class	PO-19195	Phillina Ober	Home Office	United
2	3	US-2021-112326	04-01-2021	08-01-2021	Standard Class	PO-19195	Phillina Ober	Home Office	United
3	4	US-2021-112326	04-01-2021	08-01-2021	Standard Class	PO-19195	Phillina Ober	Home Office	United
4	5	US-2021-141817	05-01-2021	12-01-2021	Standard Class	MB-18085	Mick Brown	Consumer	United

5 rows × 21 columns

Checking for Missing Values: Missing values were identified using .isnull().sum().

In [ ]:
```
df.isnull().sum()
```

Out[ ]:

	0
Row ID	0
Order ID	0
Order Date	0
Ship Date	0
Ship Mode	0
Customer ID	0
Customer Name	0
Segment	0
Country/Region	0
City	0
State/Province	0
Postal Code	0
Region	0
Product ID	0
Category	0
Sub-Category	0
Product Name	0
Sales	0
Quantity	0
Discount	0
Profit	0

dtype: int64

Descriptive Statistics: Summary statistics of sales and profit were derived using .describe().

In [ ]:
```
df[["Sales","Profit"]].describe()
```

Out[ ]:

	Sales	Profit
count	10194.000000	10194.000000
mean	228.225854	28.673417
std	619.906839	232.465115
min	0.444000	-6599.978000
25%	17.220000	1.760800
50%	53.910000	8.690000
75%	209.500000	29.297925
max	22638.480000	8399.976000

Calculate Cost

In [ ]:
```
df["Cost"] = df["Sales"] - df["Profit"]
```

Date Conversion: Converted Order Date and Ship Date to datetime format using pd.to_datetime.

In [ ]:
```
df["Order Date"] = pd.to_datetime(df["Order Date"], dayfirst=True, errors
df["Ship Date"] = pd.to_datetime(df["Ship Date"], dayfirst=True)
```

New Features: Extracted Year, Month, and Day from the Order Date column for trend analysis.

In [ ]:
```
df["Year"] = df["Order Date"].dt.year
df["Month"] = df["Order Date"].dt.month
df["Day"] = df["Order Date"].dt.day
```

Visual Data Exploration

Heatmap: Profit by Region and Category

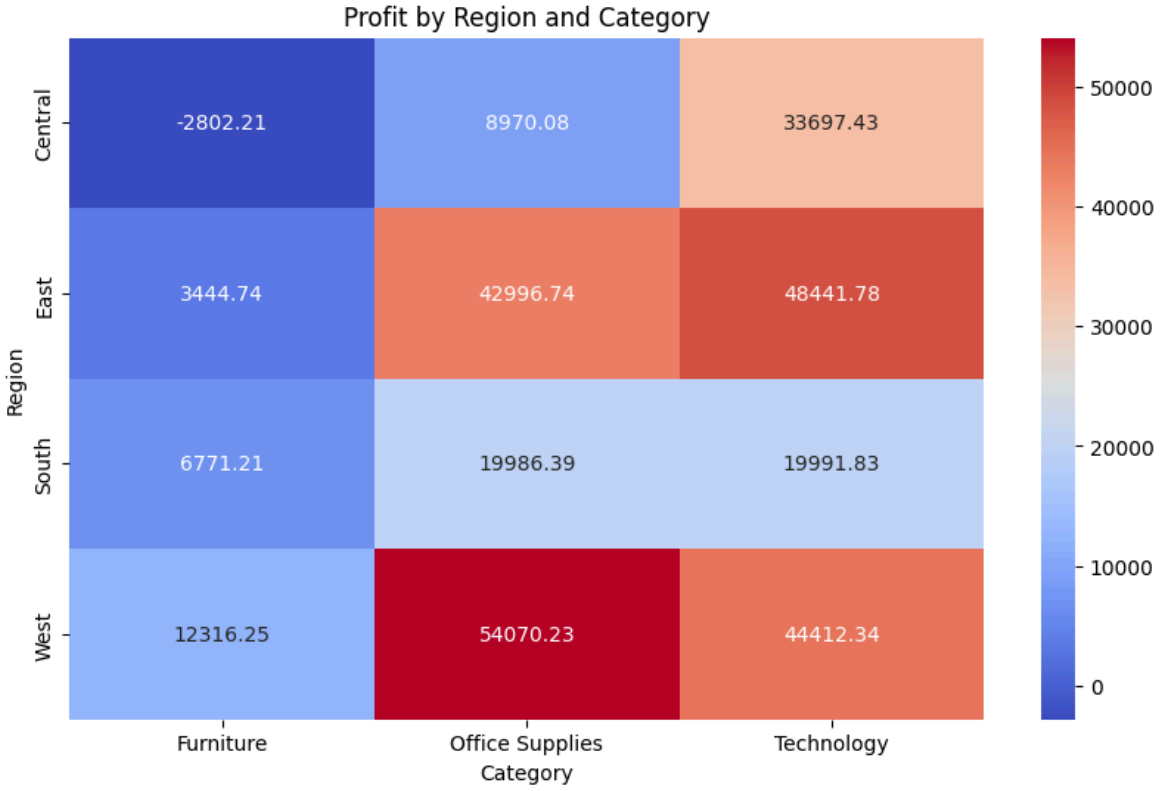
Description: A heatmap showing profit distribution across regions and product categories.

Method: Used pd.pivot_table for aggregation and sns.heatmap for visualization.

Insight: High profits observed in the "Technology" category, especially in the "West" and "Central" regions.

In [ ]:
```
# Visualization 1: Heatmap of Profit by Region and Category
pivot_table = pd.pivot_table(df, values="Profit", index="Region", columns
plt.figure(figsize=(10, 6))
```

```
sns.heatmap(pivot_table, annot=True, fmt=".2f", cmap="coolwarm")
plt.title("Profit by Region and Category")
plt.show()
```



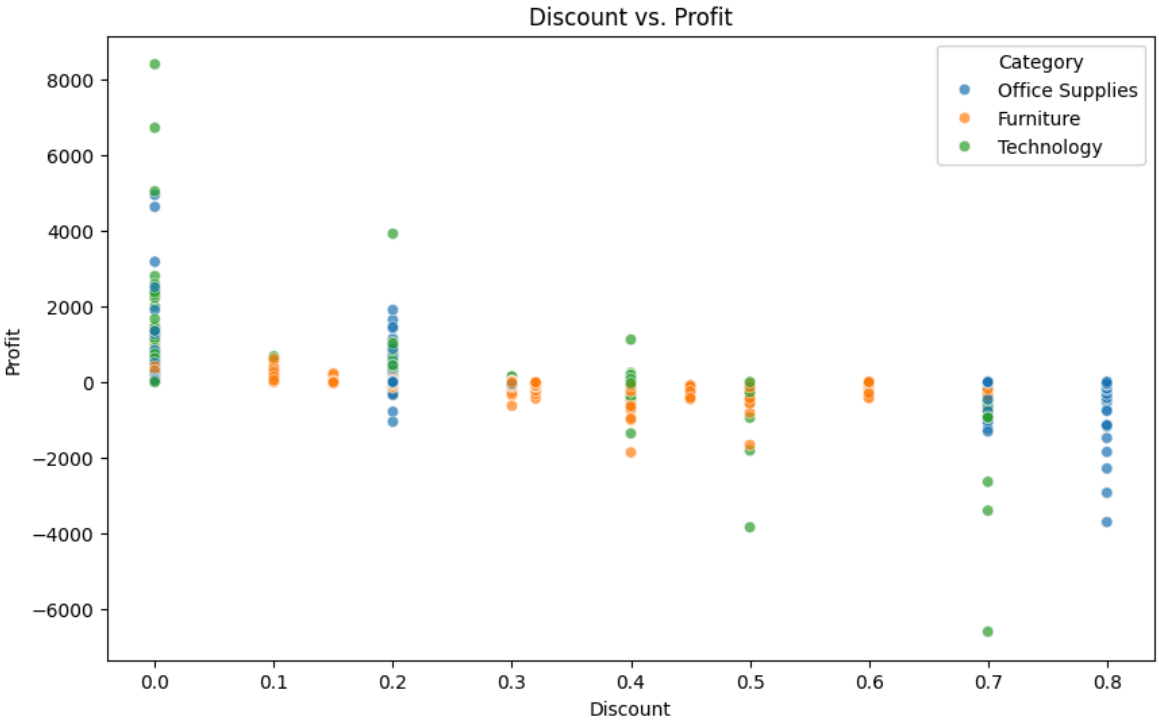
Scatter Plot: Discount vs. Profit

Description: A scatter plot examining the relationship between discount levels and profit.

Method: Used sns.scatterplot with hue to distinguish between categories.

Insight: Higher discounts correlate with lower profits, particularly in the "Furniture" category.

In [ ]:
```
# Visualization 2: Scatter plot of Discount vs. Profit
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x="Discount", y="Profit", hue="Category", alpha=
plt.title("Discount vs. Profit")
plt.xlabel("Discount")
plt.ylabel("Profit")
plt.show()
```



Bar Chart: Sales and Profit by Segment

Description: A bar chart comparing sales and profit across customer segments.

Method: Grouped data by segment and plotted a bar chart using sns.barplot.

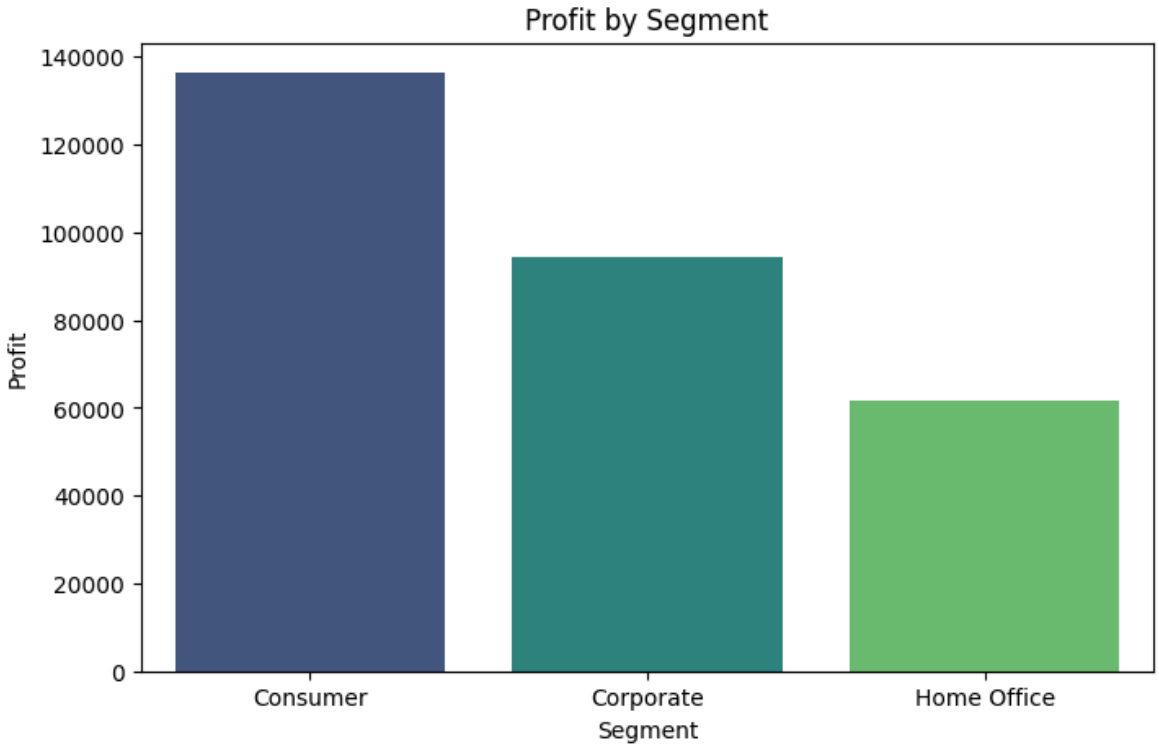
Insight: "Corporate" and "Consumer" segments are the most profitable.

In [ ]:
```
# Visualization 3: Bar chart of Sales and Profit by Segment
segment_data = df.groupby("Segment")[["Sales", "Profit"]].sum().reset_ind
segment_data = segment_data.sort_values(by="Profit", ascending=False)
plt.figure(figsize=(8, 5))
sns.barplot(data=segment_data, x="Segment", y="Profit", palette="viridis"
plt.title("Profit by Segment")
plt.ylabel("Profit")
plt.show()
```

```
<ipython-input-14-a88acf1b0efc>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be remove
d in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for
the same effect.

  sns.barplot(data=segment_data, x="Segment", y="Profit", palette="viridi
s")
```



Results

Heatmap: Revealed key profitable regions and categories.

Scatter Plot: Highlighted the negative impact of high discounts on profitability.

Bar Chart: Showed customer segments that drive profits, providing focus areas for targeted strategies.

In [ ]:
```
# Summary of Insights
print("\nInsights:")
print("1. Heatmap: Certain regions and categories have higher profits. Fo
print("2. Scatter Plot: High discounts often correlate with lower profits
print("3. Bar Chart: Some customer segments are significantly more profit
```

```
Insights:
1. Heatmap: Certain regions and categories have higher profits. Focus on t
hese areas.
2. Scatter Plot: High discounts often correlate with lower profits.
3. Bar Chart: Some customer segments are significantly more profitable.
```

Applying the data analysis on superstore some emerging insights were highlighted; areas with high performance, product categories and customers group that can generate increased sales and profit through optimization of the discount offers. These suggestions form a premise for extra contemplation and planning.

Cleaned Dataset:

The cleaned dataset has been saved as Superstore_cleaned.csv.

In [ ]:
```
# Save cleaned dataset to a new CSV file
df.to_csv("Superstore_cleaned.csv", index=False)
print("\nCleaned dataset saved as 'Superstore_cleaned.csv'.")
```

```
Cleaned dataset saved as 'Superstore_cleaned.csv'.
```

(open my notebook)
(http://localhost:8889/lab/tree/MN5813/SuperStore_Jalawan%20Khan.ipynb)

Conclusion

This supplementary report is aimed at providing the Interpretations and Recommendations in addition to the data analysis and data visualizations, presented in the Jupyter Notebook. In pursuing the most significant rise in profitability level, studies discussions point towards a new agenda with special reference to regional, product and customer profit ability analysis. For instance, the “Technology” sub-segment is always a winner in areas such as the West; however, the “Corporate” and “Consumer” segments have higher revenues. On the other hand, high discounts have a direct effect of reducing the level of profitability more so in the Furniture category, this show that discounts have to be well thought out. In this context, the analysis presented in this report can be viewed as a set of guidelines for identifying potential vectors for improving the quality of decision-making, focusing the marketing message, and fine-tuning operating activities. To the same effect, the annotated bibliography presents the main findings and develops more references for future reference as well as further research. Combined, these resources serve as a framework to support the development of evidentiased approaches for driving business results and recognizing opportunities for future enhancement.

Bibliography

Primary Sources

1. Superstore Dataset (Kaggle)

o Relevance: The dataset was the primary source for all analysis. It contains transactional data on sales, profits, discounts, and customer segmentation. o Link: https://www.kaggle.com/datasets/aditirai2607/super-market-dataset

2. Seaborn Documentation

o Relevance: Used as a reference for creating visualizations like scatter plots, heatmaps, and bar charts. o Link: https://seaborn.pydata.org/

3. Pandas Documentation

o Relevance: Provided guidance on data cleaning, processing, and manipulation techniques. o Link: https://pandas.pydata.org/docs/ Supplementary Sources

4. Matplotlib Documentation

o Relevance: Assisted in creating advanced visualizations and fine-tuning chart aesthetics.

5. Business Analytics Techniques

o Relevance: Academic references for understanding profit analysis, customer segmentation, and discount strategies. o Black, K. (2023). Business statistics: for contemporary decision making. John Wiley & Sons.