Assignment (Section 2)

Name : Priyanshu Kumar

Mail id : priyanshu_24a12res1193@iitp.ac.in

XInsight : See below 👇🏻

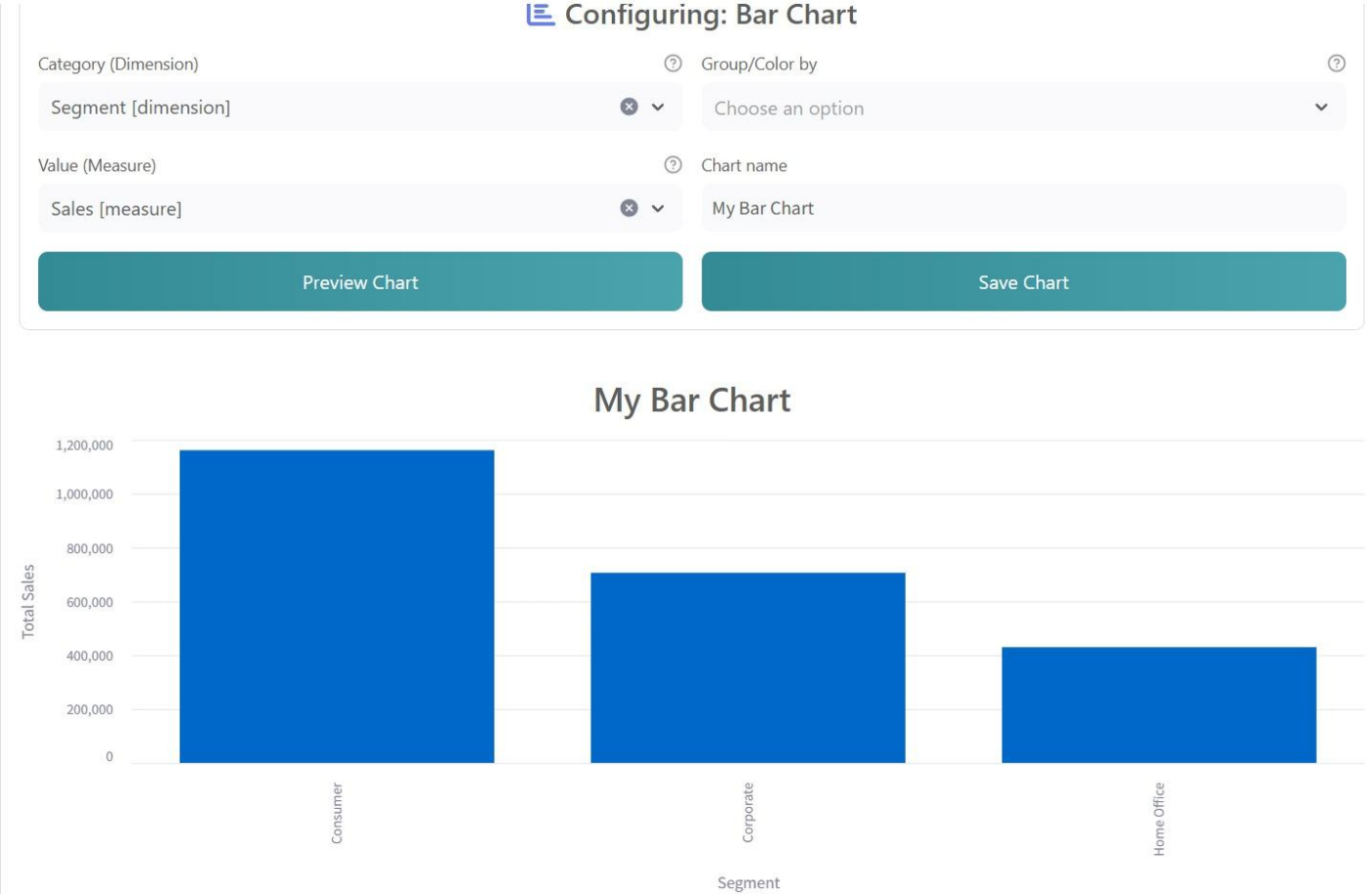
GitHub Repo (Python) : Click Here

Dataset Link

ASSIGNMENT

1. Perform a market share analysis to visualize the contribution of different Customer Segments to the total revenue. Identify which segment dominates the portfolio and discuss if the business is over-reliant on a single group.

Ans:

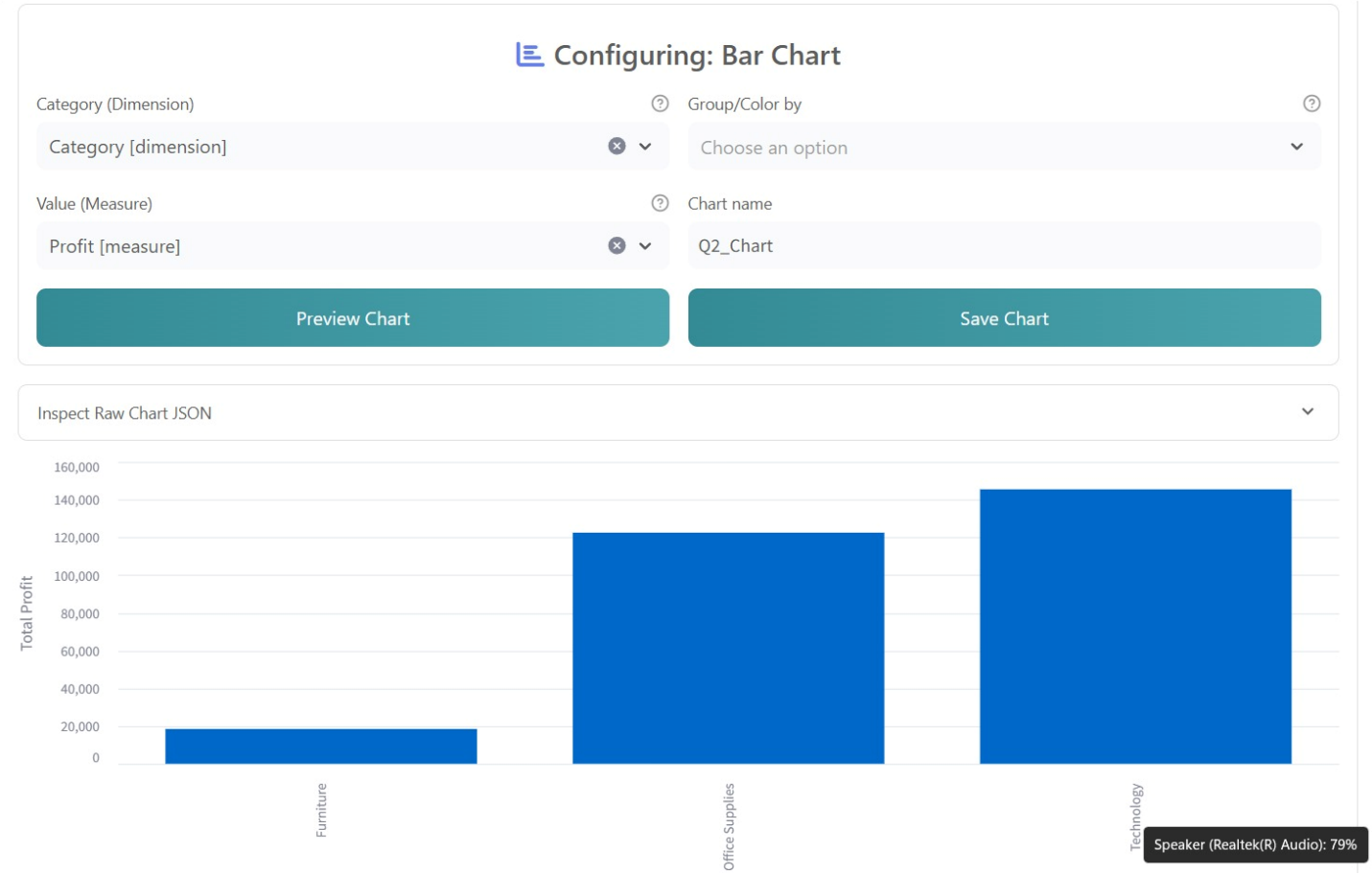


A market share analysis was performed using a bar chart to visualize the contribution of different customer segments to total revenue. The analysis shows that the Consumer segment dominates the revenue portfolio, contributing the highest share of total sales, followed by the corporate segment, while the Home Office segment contributes the least.

This indicates that the business is largely consumer-driven, suggesting a certain level of reliance on the Consumer segment for revenue generation. However, since Corporate and Home Office segments also contribute a significant portion, the business is not entirely over-reliant on a single customer group, though diversification across segments could further reduce revenue risk.

2. Rank the Product Categories from highest to lowest profit and identify if there is a significant 'performance gap' between the top-performing and bottom-performing categories

Ans:



Product categories were ranked based on total profit using a bar chart. The analysis shows that Technology is the most profitable category, followed by Office Supplies, while Furniture generates the lowest profit. The substantial gap between Technology and Furniture indicates a significant performance difference, suggesting that Furniture requires strategic improvements in pricing or cost management.

3. Evaluate the Sales-Profit Correlation to determine the efficiency of revenue generation. Use the scatter plot to identify 'High-Volume, Low-Profit' anomalies- orders that generate significant sales but fail to contribute to the bottom line.

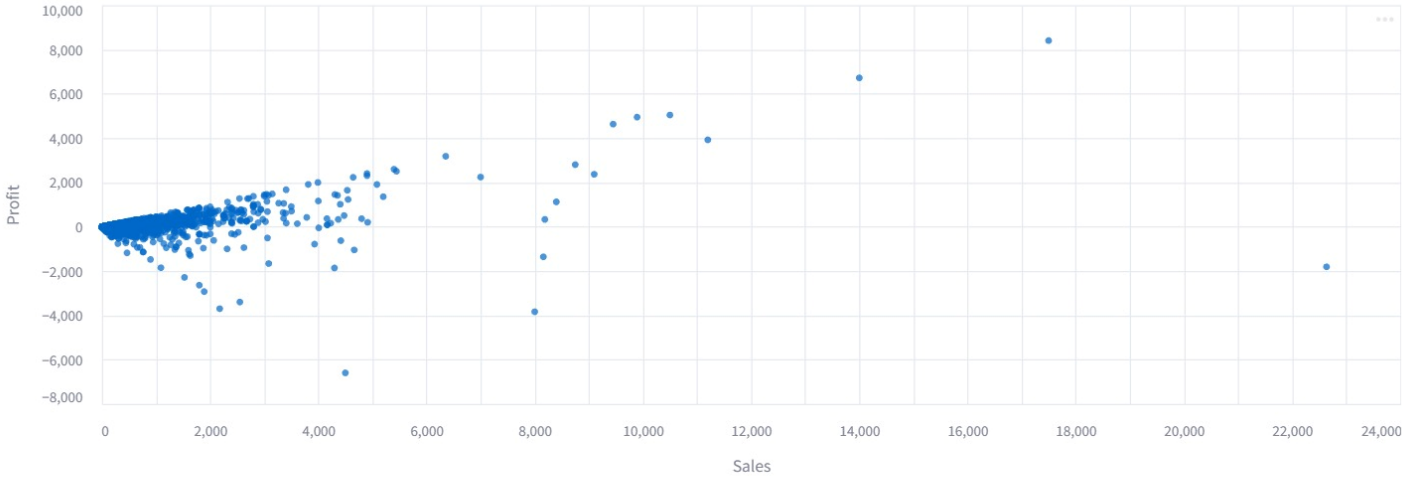
Ans:

The Pearson's correlation between x and y is 0.479

The Kendall's correlation between x and y is 0.452

The Spearman's correlation between x and y is 0.518

Two columns have weak or no correlation

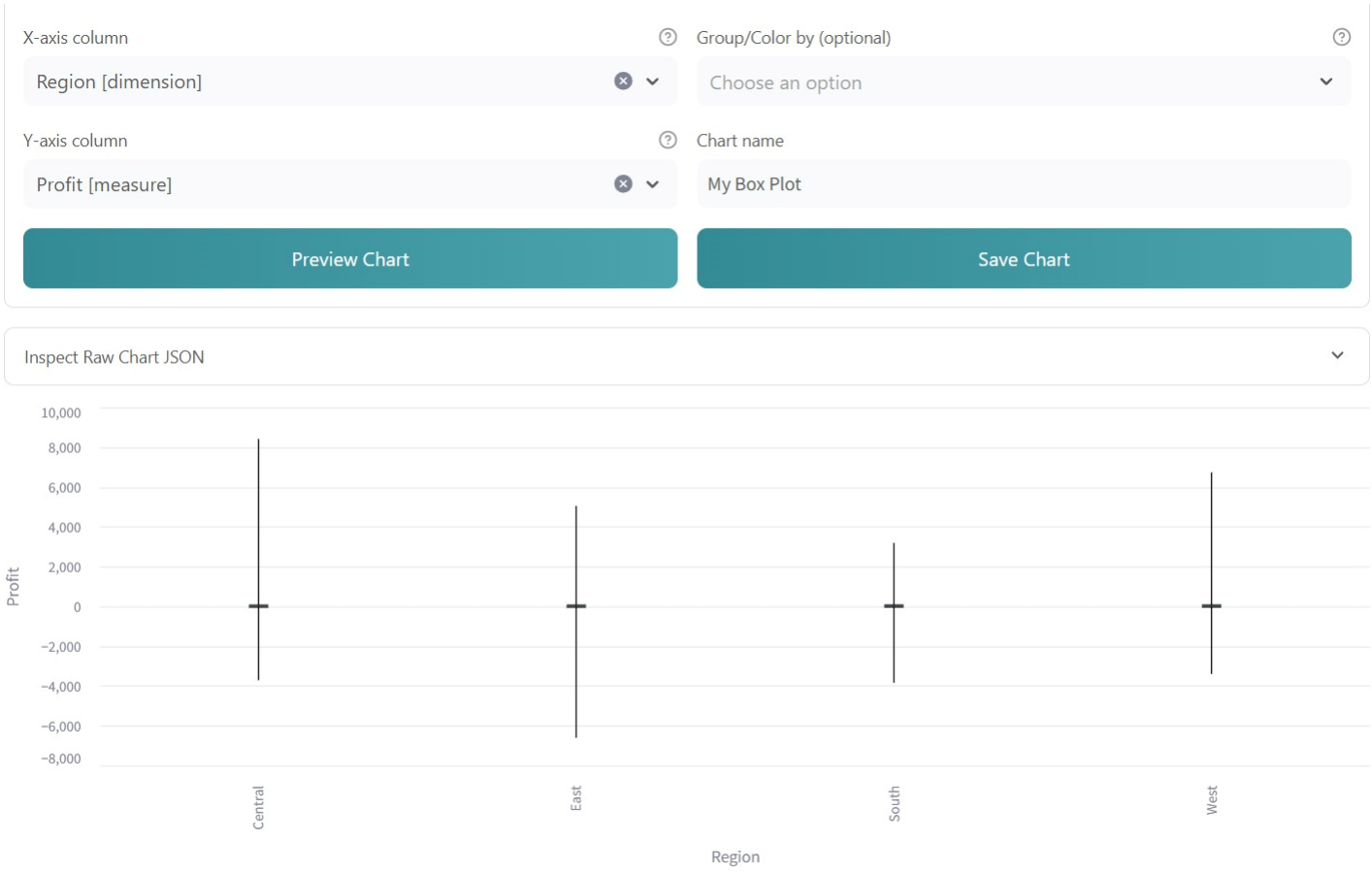


Download PDF

The Sales–Profit scatter plot shows a weak to moderate positive correlation between sales and profit. While higher sales generally lead to higher profits, several high-sales orders exhibit low or even negative profit. These high-volume, low-profit anomalies indicate inefficiencies in revenue generation, likely due to heavy discounting or high operational costs.

4. Using boxplots, identify which geographic region has the most 'volatile' profit margins and pinpoint extreme outliers (extreme losses or gains).

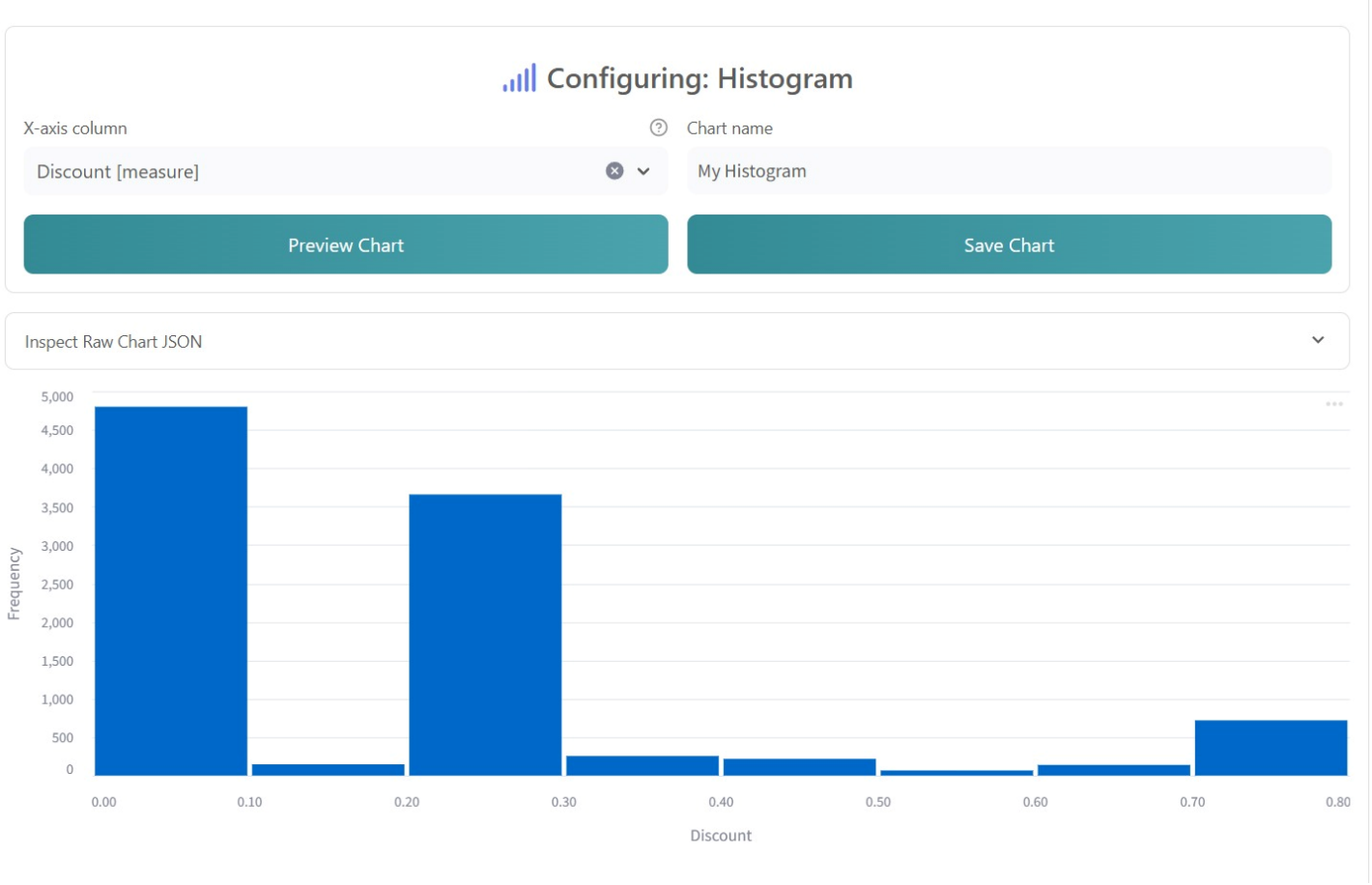
Ans:



Boxplot analysis of profit margins across regions shows that the Central region has the most volatile profit distribution, with a wide interquartile range and multiple extreme outliers, including significant losses and gains. The East region also exhibits high variability with extreme negative outliers. In contrast, the South and West regions display relatively stable profit margins with fewer extreme fluctuations. This indicates higher financial risk in the Central region.

5. Use a histogram to visualize the distribution of discount rates. Is the company mostly using 'shallow' discounts (10%) or 'aggressive' discounts (50%)? How does this impact the overall transaction volume?

Ans:

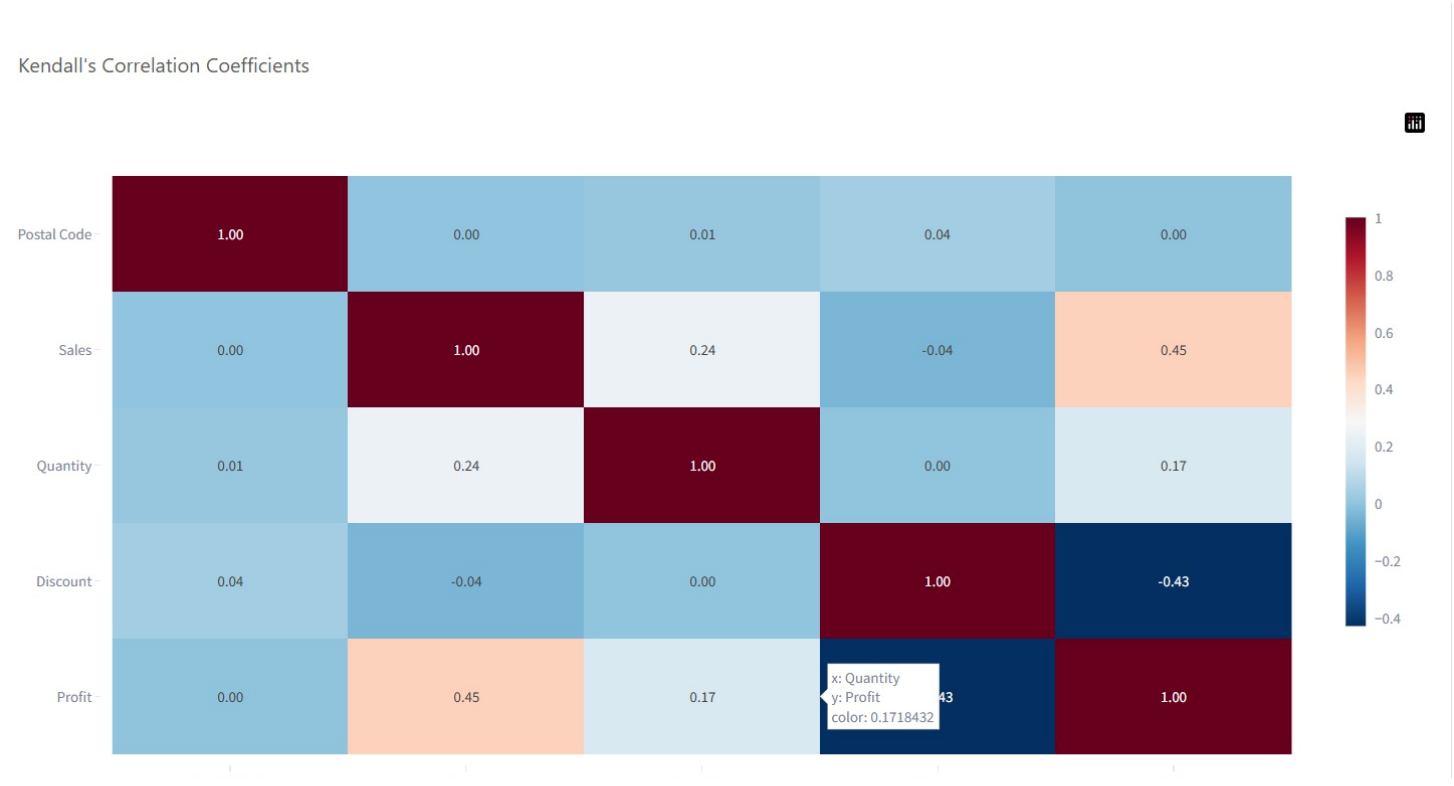


The histogram of discount rates shows that the company primarily uses shallow discounts (around 10%), with the highest frequency of transactions occurring at low discount levels. Moderate discounts (20–30%) are also common, while high discounts (50% or more) are used infrequently. This indicates that the company focuses on maintaining profit margins, using higher discounts selectively to boost transaction volume when required.

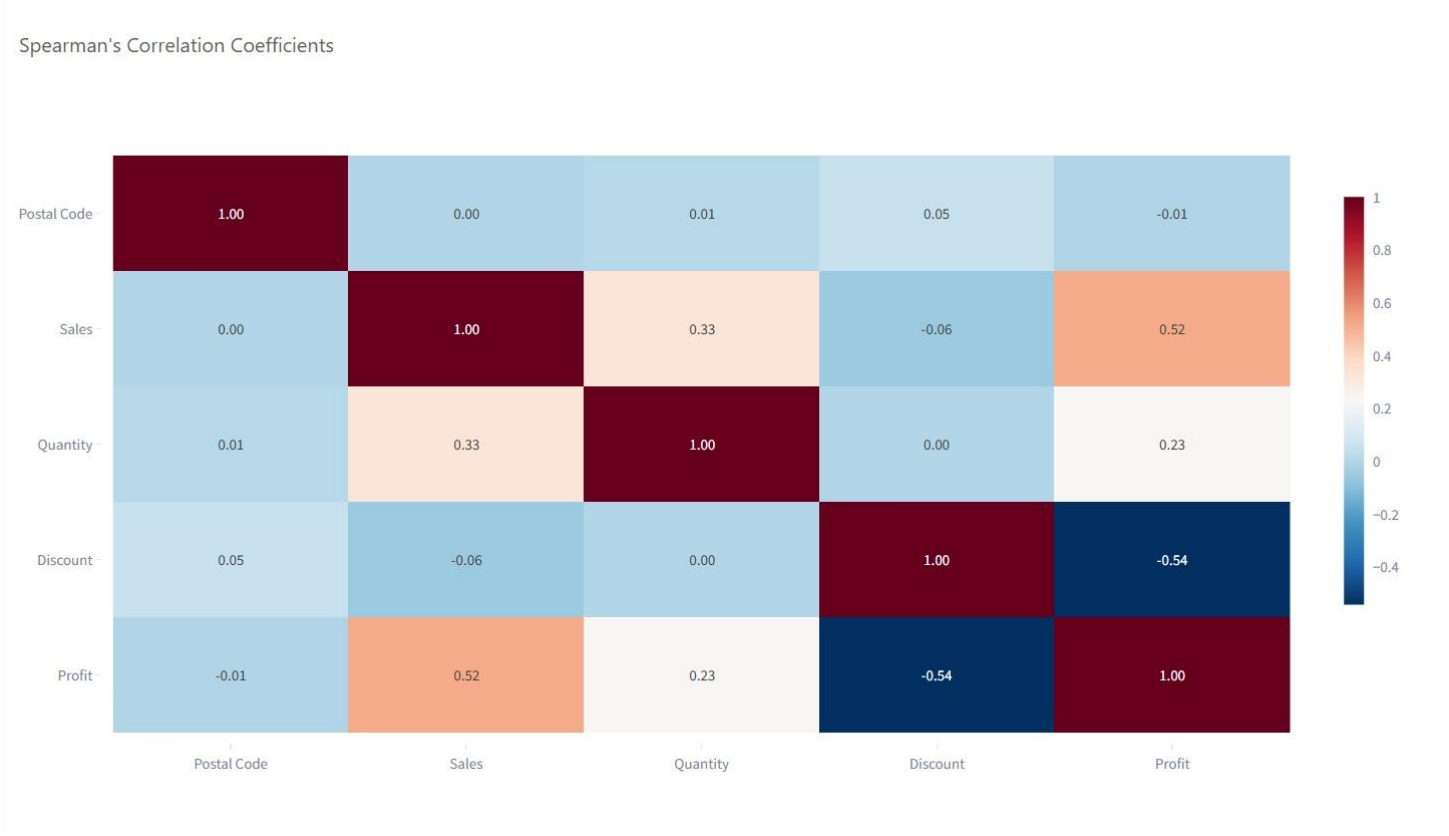
6. Perform a Multivariate Dependency Study between Sales, Quantity, and Discount. Quantify the impact of aggressive discounting on net profit using the correlation matrix.

Ans: Pearson

Kendall's



Spearman’s



DataFrame

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1	-0.0239	0.0128	0.0584	-0.03
Sales	-0.0239	1	0.2008	-0.0282	0.4791
Quantity	0.0128	0.2008	1	0.0086	0.0663
Discount	0.0584	-0.0282	0.0086	1	-0.2195
Profit	-0.03	0.4791	0.0663	-0.2195	1

A multivariate dependency analysis using Pearson, Kendall, and Spearman correlations reveals important relationships among Sales, Quantity, Discount, and Profit. Sales show a moderate positive correlation with Profit, indicating that higher revenue generally improves profitability. Quantity is positively correlated with Sales but has a weak relationship with Profit, suggesting that higher volumes alone do not ensure profit growth. In contrast, Discount exhibits a strong negative correlation with Profit across all correlation measures, confirming that aggressive discounting significantly reduces net profitability. Overall, the analysis highlights discounting as the most critical factor impacting profit.

Python – Implementation Google Collab	GitHub Repo (Python) : Click here

ASSIGNMENT

SampleSuperstore dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

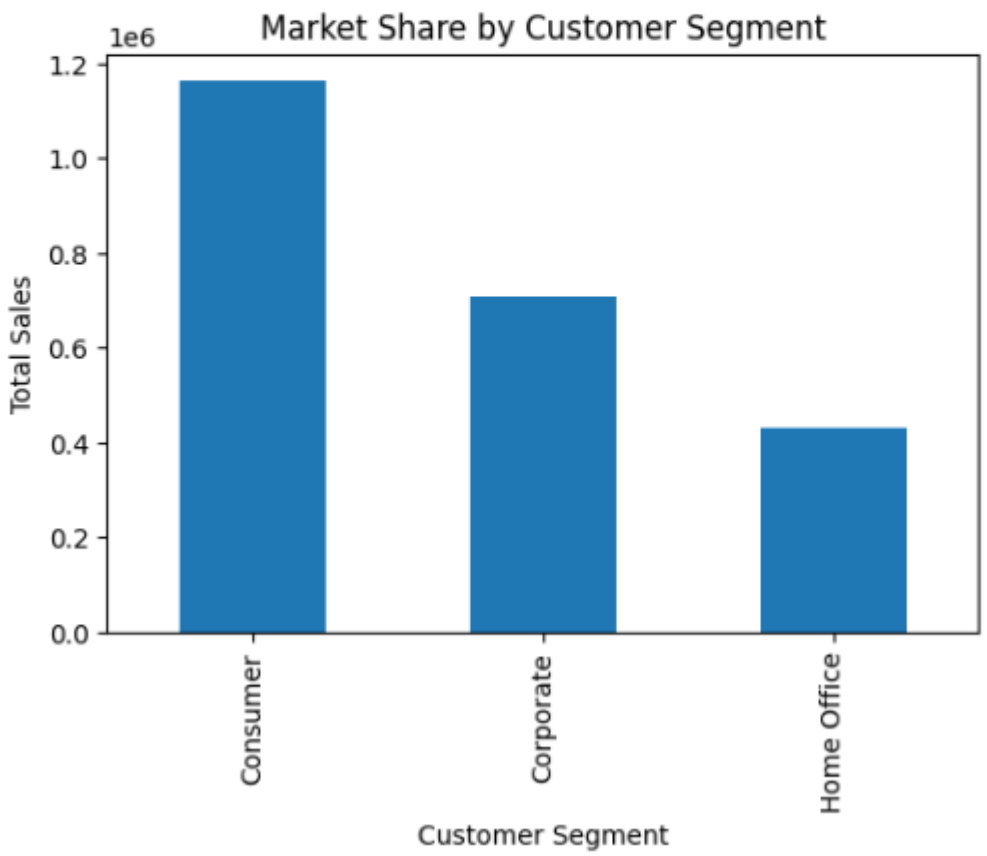
```
df=pd.read_csv("SampleSuperstore.csv")
```

```
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

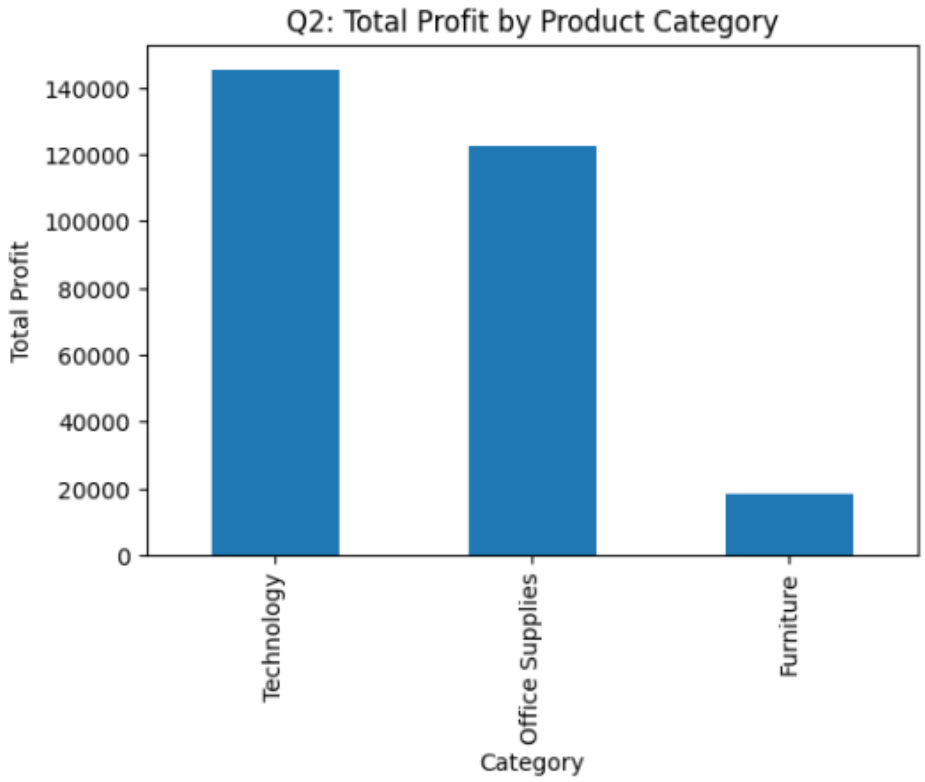
Q1 : Market Share (Segment vs Sales)

```
segment_sales=df.groupby("Segment")["Sales"].sum()

plt.figure(figsize=(6,4))
segment_sales.plot(kind="bar")
plt.title("Market Share by Customer Segment")
plt.xlabel("Customer Segment")
plt.ylabel("Total Sales")
plt.show()
```



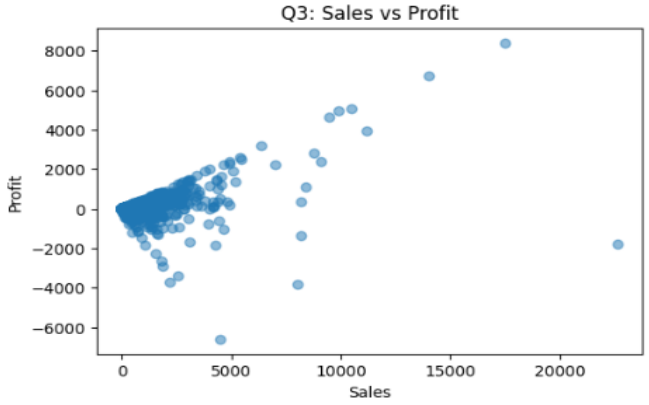
Q2: Category-wise Profit

```
category_profit = df.groupby("Category")["Profit"].sum().sort_values(ascending=False)

plt.figure(figsize=(6,4))
category_profit.plot(kind="bar")
plt.title("Q2: Total Profit by Product Category")
plt.ylabel("Total Profit")
plt.xlabel("Category")
plt.show()
```



Q3: Sales Vs Profit

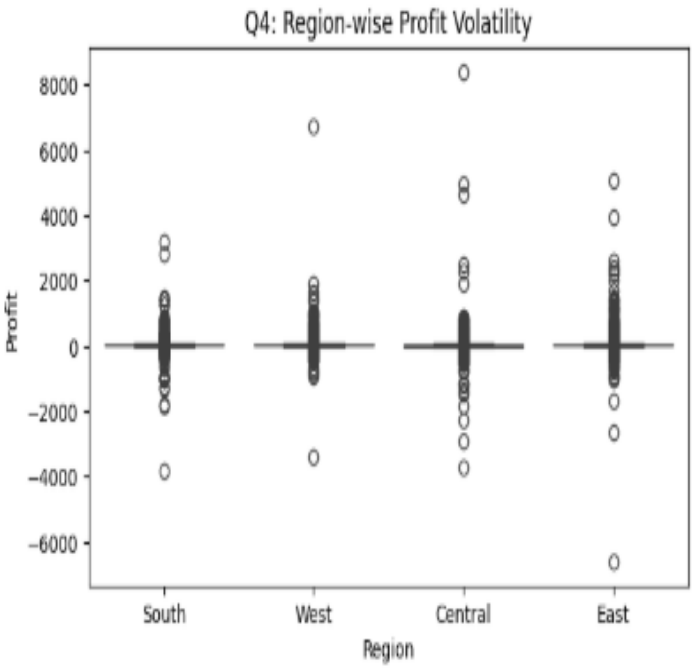
```
plt.figure(figsize=(6,4))
plt.scatter(df["Sales"], df["Profit"], alpha=0.5)
plt.title("Q3: Sales vs Profit")
plt.xlabel("Sales")
plt.ylabel("Profit")
plt.show()

print("Pearson:", df["Sales"].corr(df["Profit"], method="pearson"))
print("Spearman:", df["Sales"].corr(df["Profit"], method="spearman"))
print("Kendall:", df["Sales"].corr(df["Profit"], method="kendall"))
```




```
Pearson: 0.4790643497377062
Spearman: 0.5184066611400607
Kendall: 0.4521182435817151
```

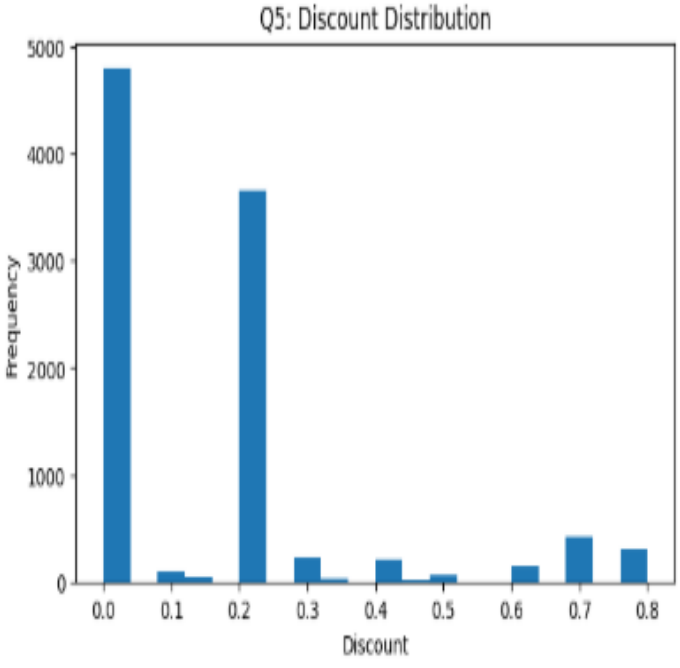
Q4: Region-wise Boxplot

```
plt.figure(figsize=(7,4))
sns.boxplot(x="Region", y="Profit", data=df)
plt.title("Q4: Region-wise Profit Volatility")
plt.show()
```



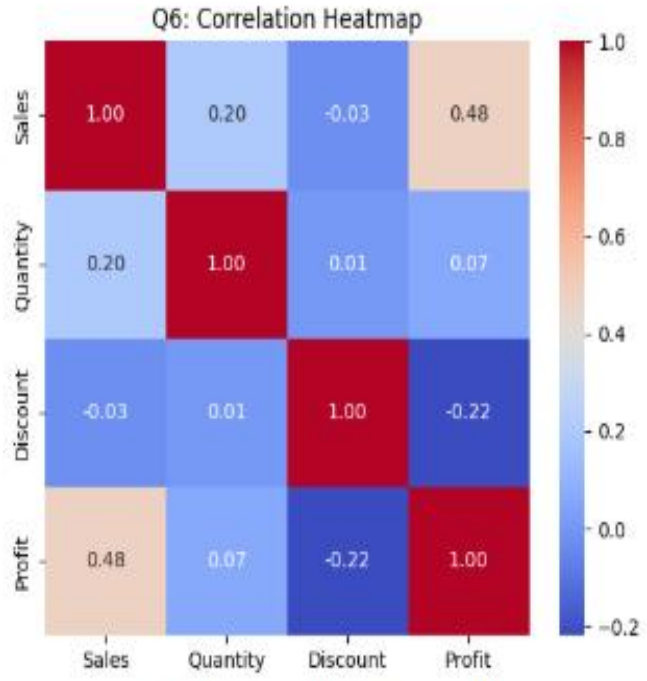
Q5: Discount Histogram

```
plt.figure(figsize=(7,4))
plt.hist(df["Discount"], bins=20)
plt.title("Q5: Discount Distribution")
plt.xlabel("Discount")
plt.ylabel("Frequency")
plt.show()
```



Q6: Correlation Heatmap

```
cols = ["Sales", "Quantity", "Discount", "Profit"]
corr = df[cols].corr()

plt.figure(figsize=(6,5))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Q6: Correlation Heatmap")
plt.show()

corr
```



	Sales	Quantity	Discount	Profit
Sales	1.000000	0.200795	-0.028190	0.479064
Quantity	0.200795	1.000000	0.008623	0.066253
Discount	-0.028190	0.008623	1.000000	-0.219487
Profit	0.479064	0.066253	-0.219487	1.000000

Name : Priyanshu Kumar

Mail id : priyanshu_24a12res1193@iitp.ac.in

GitHub Repo (Python) : Click Here