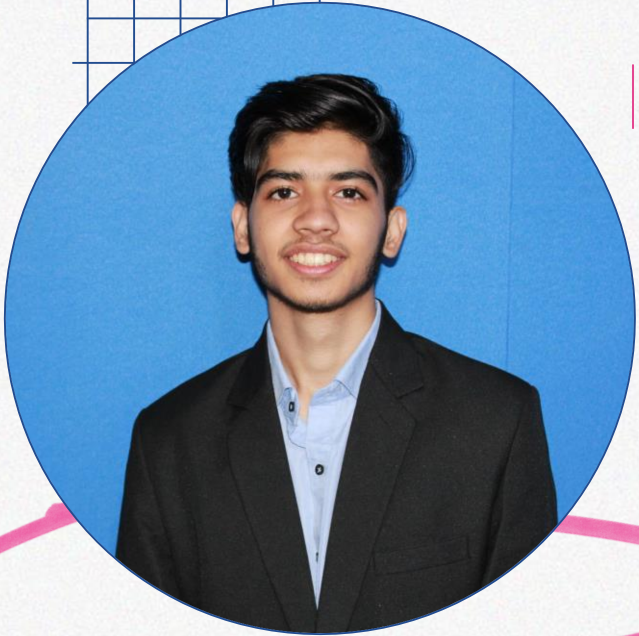
Data Analytics Project-IBM INTERNSHIP

Sahil Thakur

sahilthakurco@gmail.com

Thakur College of Engineering and Technology

Organization - DGT



Date – 12/06/2023 - 24/07/2023

Topic

Analysis of Superstore Dataset

This project centers around conducting a thorough analysis of the SuperStore dataset, which comprises sales data from a fictitious retail store. The primary objective is to extract valuable insights regarding the store's performance and to know about specific areas that offer potential for enhancement and growth.

We are provided with various information in data set such as product type, customer demographics, regional infographics

INDEX

Agenda

The goal of the "Analysis of Superstore Dataset" project is to investigate and analyze a dataset from a superstore in order to learn important details about its sales, clients, merchandise, and general performance. The project attempts to evaluate the store's strengths and weaknesses using data-driven approaches and processes, identify possible development areas, and provide data-backed recommendations for improving business operations and increasing profitability.

Project Overview

The project entails performing an extensive study of the Superstore dataset, which includes historical data on sales transactions, customer data, and product details. The dataset includes data on a variety of characteristics, including sales revenue, profit margins, consumer demographics, product categories, and the regions where the business is located. The project aims to identify patterns, trends, and correlations in the data by utilizing data analysis methods.

END USERS

1. Store Management: By using the analysis's insights, store managers will be better able to manage their inventories, set reasonable prices, and spot areas where they may cut costs.

2. Marketing Team: The study can be used by the marketing team to identify target consumer categories, understand client preferences, and create focused marketing efforts.

3. Sales Team: By recognizing top-performing products, analyzing sales patterns, and adjusting sales tactics for various geographies, the sales team can benefit.

4. Executives and Stakeholders: The results of the project will be helpful to executives and stakeholders as they can aid in formulating strategic plans, establishing long-term objectives, and assessing overall performance.

Solution

- Utilized the SuperStore dataset to conduct an extensive analysis of sales data, providing a deep understanding of the business's performance.
- Explored the dataset comprehensively, gaining insights into its structure, variables, and data quality, ensuring the reliability of subsequent analysis.
- Ensured accurate and reliable analysis results by performing meticulous data cleaning and preprocessing techniques on the SuperStore dataset.
- Conducted in-depth exploratory data analysis (EDA) to unveil hidden patterns, trends, and relationships within the sales data, revealing valuable insights.
- Investigated key performance metrics, including sales revenue, profit, and customer segments, to identify areas for improvement and growth opportunities.
- Identified potential target markets by analyzing geographical sales distribution, providing actionable information for strategic expansion.
- Examined top-selling products and popular categories, evaluating their impact on overall store performance and informing future inventory management decisions.
- Utilized advanced techniques to analyze customer behavior, including buying patterns and loyalty, enabling the optimization of marketing strategies for increased customer satisfaction and retention.

Value Proposition

The following value propositions are offered by our solution:

•Data-Driven Decision Making: By examining the Superstore dataset, we help marketing and store managers make data-driven decisions. Because they can base their judgments on thorough analysis, they may improve store performance, streamline operations, and develop more precise marketing plans.

•Enhanced Profitability: Our analysis aids in locating chances for raising sales, enhancing inventory control, and cutting costs, all of which contribute to the Superstore's enhanced profitability. The store can increase its income and profitability by streamlining processes, discovering high-demand products, and improving pricing methods.

•consumer Insights and Personalized Marketing: Our system offers marketing managers insightful data on consumer behavior, demographics, and preferences. Due to their ability to create targeted marketing campaigns, customize promotions, and increase client engagement, they are able to generate more revenue.

How did I customize the project

Data visualization is a regular part of projects involving data analysis, but my method stands out by making use of the potent tools Matplotlib and Seaborn. These libraries include a wide range of customisation possibilities, enabling the development of aesthetically pleasing and illuminating charts, graphs, and plots. My approach improves understanding of complicated patterns and relationships within the Superstore dataset by presenting data in a visually appealing manner and utilizing the capabilities of Matplotlib and Seaborn.

Interactive Dashboards: My solution includes interactive dashboards to deliver an amazing user experience. These dashboards give users the ability to interactively explore and interact with the data that has been evaluated, allowing them to dig down into particulars, use filters, and visualize various dimensions. The dashboards' interactive features increase engagement.

MODELLING Techniques, Frameworks ,methods used

Exploratory Data Analysis (EDA)

1. Data Understanding: Exploratory Data Analysis (EDA) helps in gaining a deep understanding of the dataset, including its structure, variables, and content.

2. Pattern Identification: EDA allows the identification of patterns, trends, and relationships within the data, enabling insights into sales trends, customer behavior, and product performance.

3. Data Visualization: EDA involves creating various visualizations, making it easier to communicate complex information and identify trends that may not be apparent from raw data.

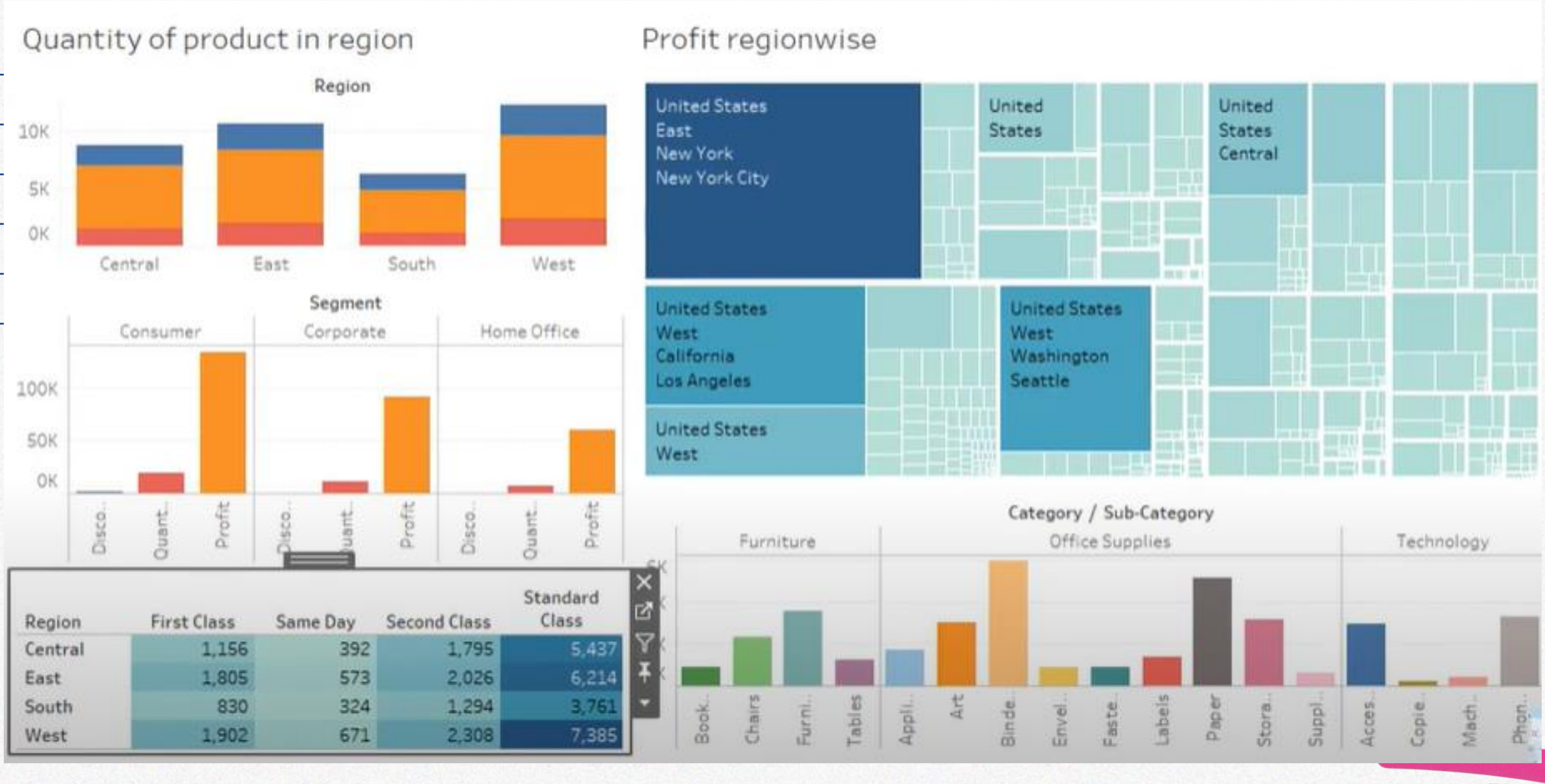
Market Segmentation

Market segmentation is used to categorize customers based on purchasing behavior, enabling targeted marketing strategies and personalized offerings to optimize sales and customer satisfaction in the superstore.

Data Visualization

Python libraries such as Matplotlib,Seaborn were used to create informative graphs, charts to properly display the findings of the analysis

Data Visualization result



Quantity of product in region

Profit regionwise

Region	First Class	Same Day	Second Class	Standard Class
Central	1,156	392	1,795	5,437
East	1,805	573	2,026	6,214
South	830	324	1,294	3,761
West	1,902	671	2,308	7,385

LINKS

Project Link

https://github.com/SahilRT/Analysis_ of_ SuperStore_ Dataset-Data-Analytics-Project-IBM-INTERNSHIP

Research Papers

- SALES ANALYSIS ON SUPERSTORE DATASET https://www.irjmets.com/uploadedfiles/paper//issue_ 4_ april_ 2023/36572/final/fin_ irjmets1682186035.pdf
- 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/en-us/tutorials/analyze-and-visualize-superstore-data/
- Pranav, B. (2021). Sales Analysis of Superstore Data using Power BI. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/sales-analysis-of-superstore-data-using-power-bi/

Other

Super Store Sales Analysis
https://medium.com/analytics-vidhya/exploratory-data-analysis-super-store-cb91c37bcb06

Dataset

Dataset Url

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

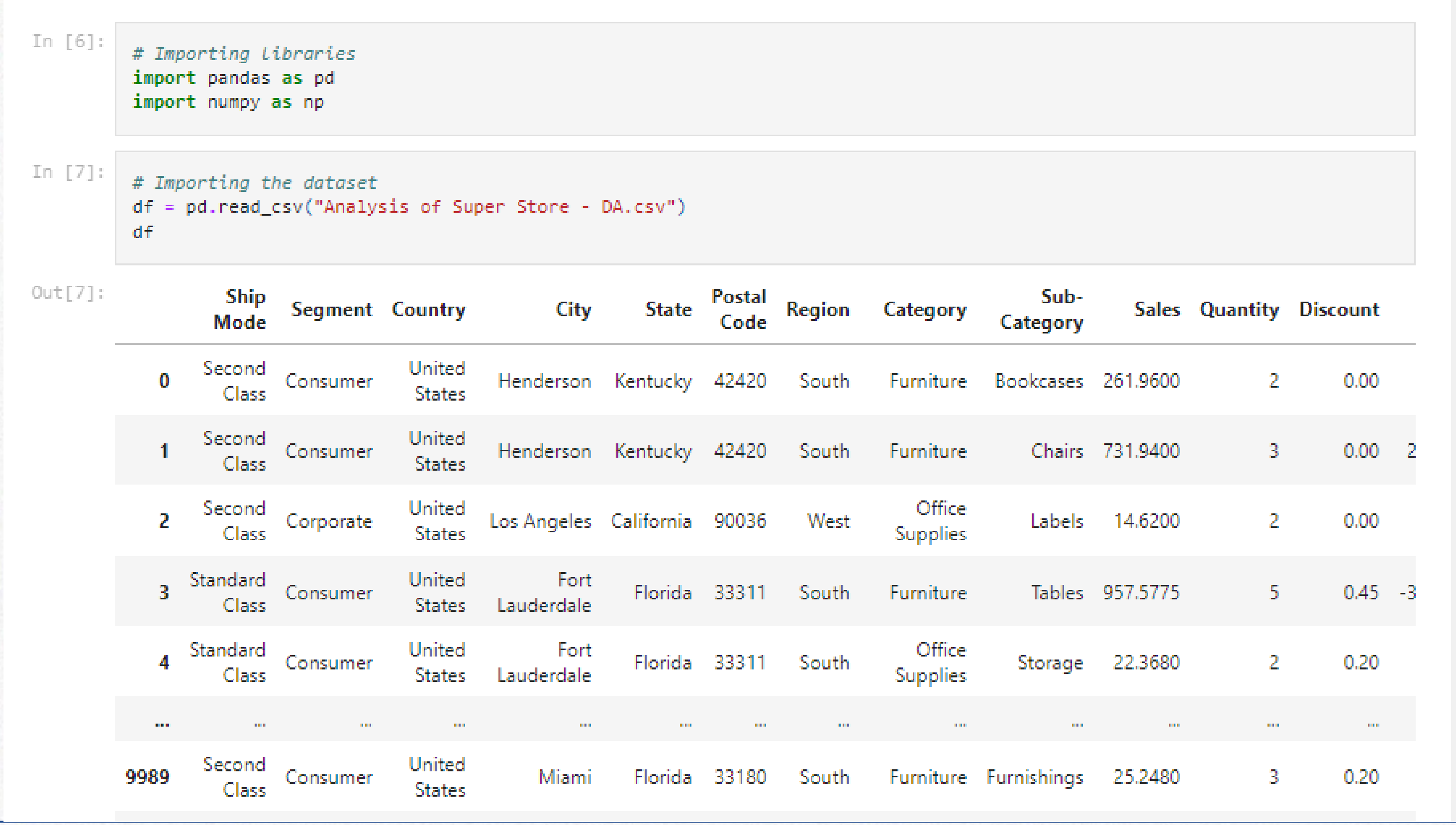
About Dataset

This is a sample superstore dataset, a kind of a simulation where you perform extensive data analysis to deliver insights on how the company can increase its profits while minimizing the losses.

Details

- Size – 1.11 mb (.csv)
- Rows - 9994
- Columns - 13

Import Dataset



In [6]:

```
# Importing libraries
import pandas as pd
import numpy as np
```

In [7]:

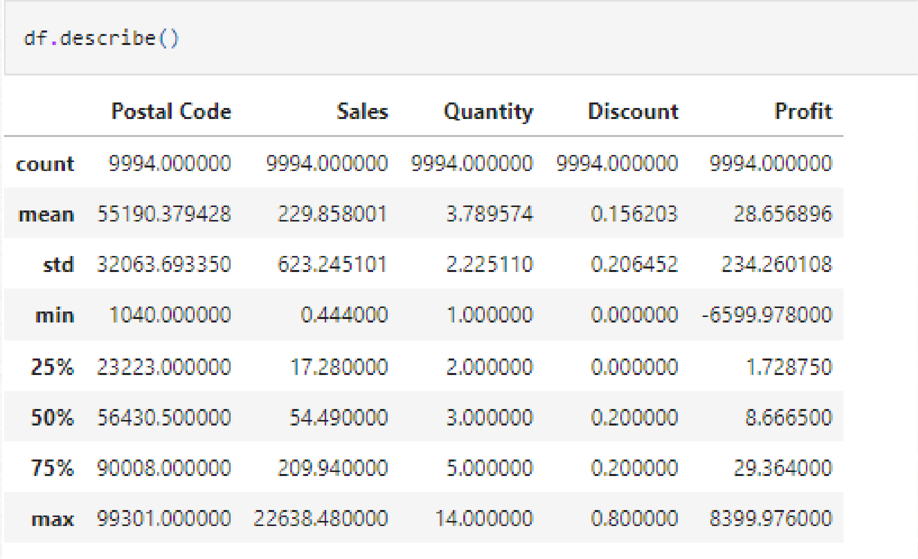
```
# Importing the dataset
df = pd.read_csv("Analysis of Super Store - DA.csv")
df
```

Out[7]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	2
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-3
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	
...	...	...	...	...	...	...	...	...	...	...	...	...	
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.2480	3	0.20	

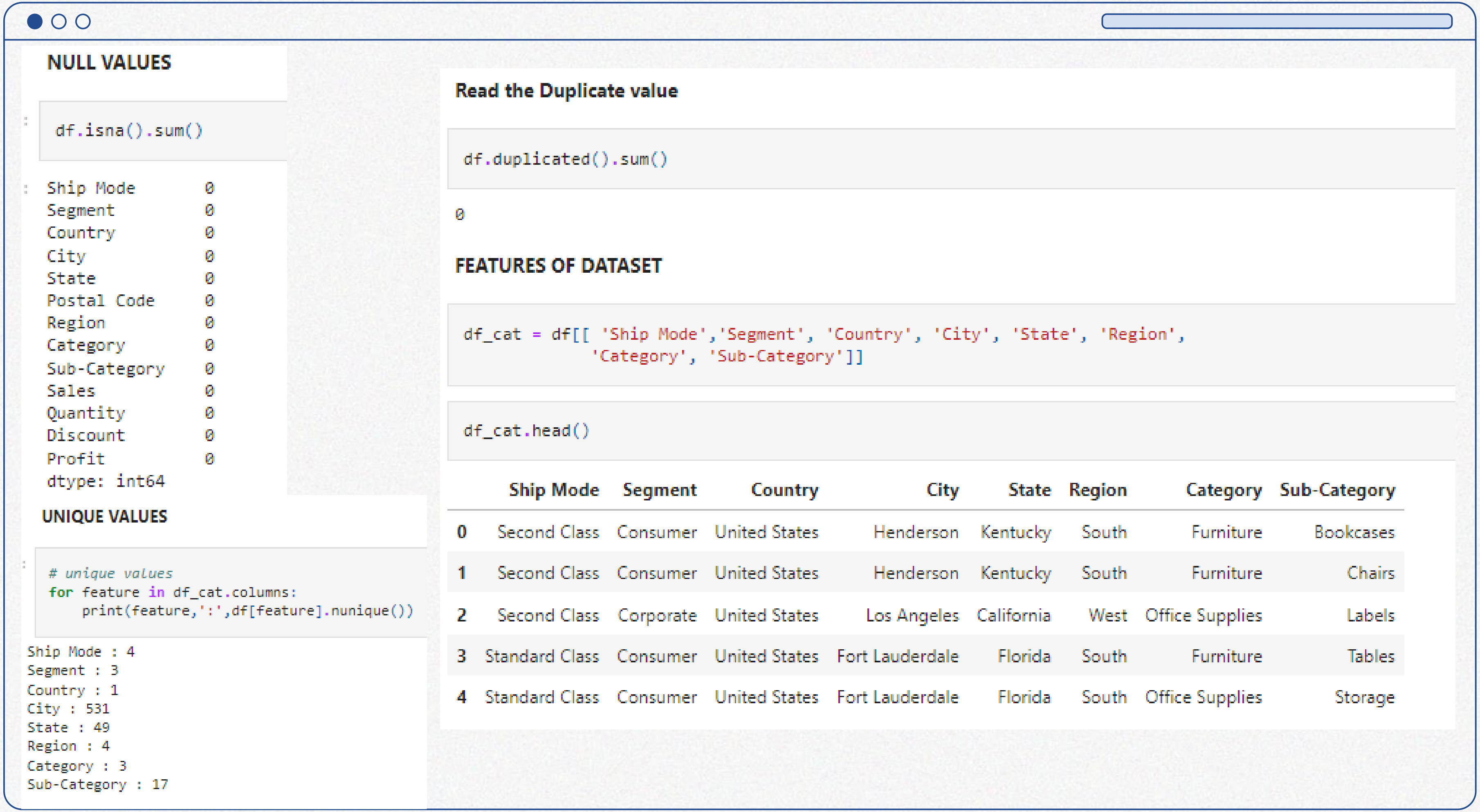
DATASET INFO

DataFrame.count : Count number of non-NA/null observations.
DataFrame.max : Maximum of the values in the object.
DataFrame.min : Minimum of the values in the object.
DataFrame.mean : Mean of the values.
DataFrame.std : Standard deviation of the observations.
DataFrame.select-dtypes : Subset of a DataFrame including/excluding columns based on their dtype.



```
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



NULL VALUES

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

```
Ship Mode       0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Category        0
Sub-Category    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

UNIQUE VALUES

```
# unique values
for feature in df_cat.columns:
    print(feature,':',df[feature].nunique())
```

```
Ship Mode : 4
Segment : 3
Country : 1
City : 531
State : 49
Region : 4
Category : 3
Sub-Category : 17
```

Read the Duplicate value

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

```
0
```

FEATURES OF DATASET

```
df_cat = df[[ 'Ship Mode','Segment', 'Country', 'City', 'State', 'Region',
             'Category', 'Sub-Category']]
```

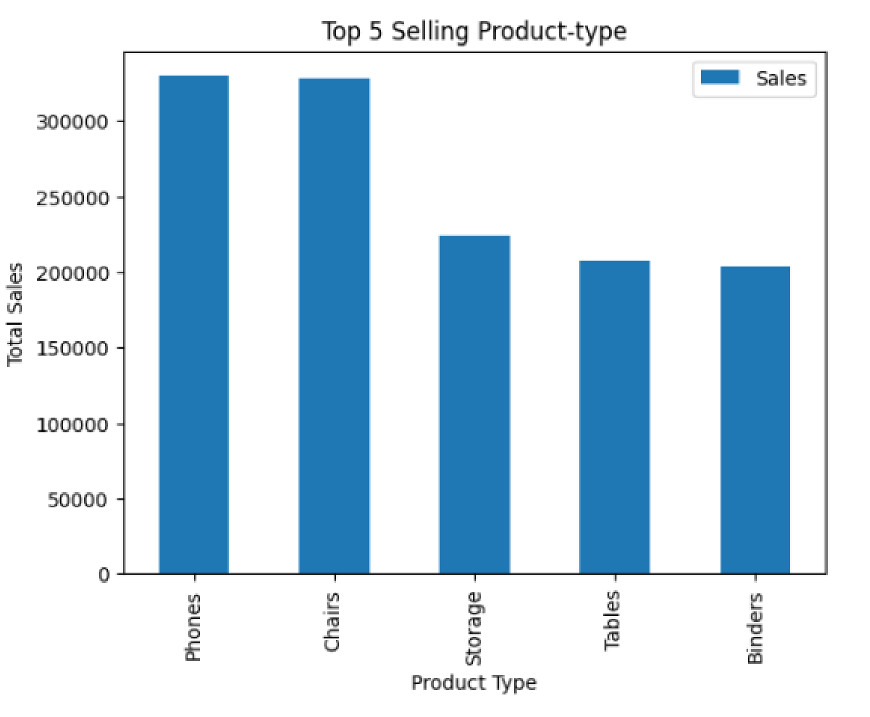
```
df_cat.head()
```

	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs
2	Second Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Furniture	Tables
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage

Exploratory Data Analysis

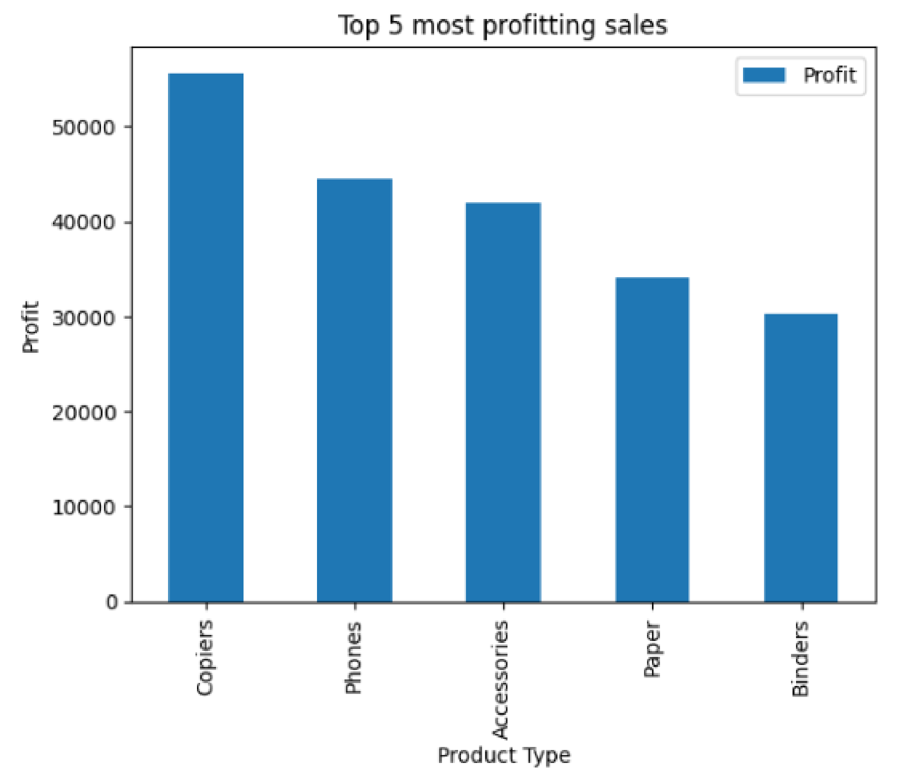
Top 5 Selling Products

```
# Group the data by Subcategory and sum up the sales
subcategory_group = df.groupby(["Sub-Category"]).sum()["Sales"]
 # Sort the data by sales in descending order
top_subcategory_sales =
subcategory_group.sort_values(ascending=False)
top5_subcategory_sales =
pd.DataFrame(top_subcategory_sales.head())


top5_subcategory_sales.plot(kind="bar")
plt.title("Top 5 Selling Product-type")
# Add labels to the x and y axes
plt.xlabel("Product Type")
plt.ylabel("Total Sales")
# Show the plot
plt.show()
```



Top 5 Profitable Products

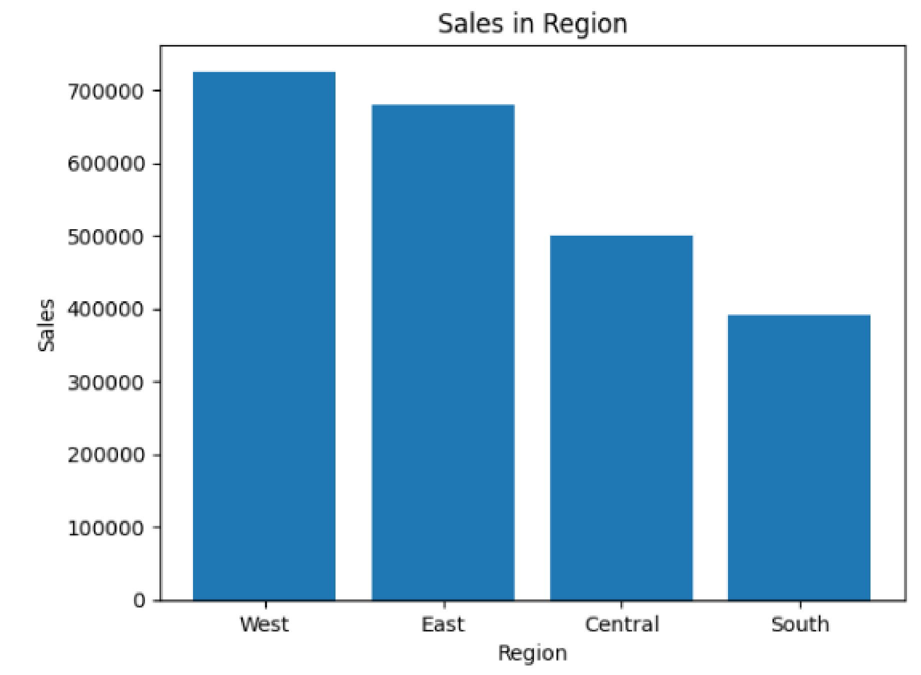
```
product_profit = df.groupby(["Sub-Category"]).sum()["Profit"]

top_profit = product_profit.sort_values(ascending=False)

top5_profit =pd.DataFrame(top_profit.head())

#Top 5 Profitting products
top5_profit.plot(kind="bar")

plt.title("Top 5 most profitting sales")

plt.xlabel("Product Type")
plt.ylabel("Profit")

plt.show()
```



Top Sales and Profit by

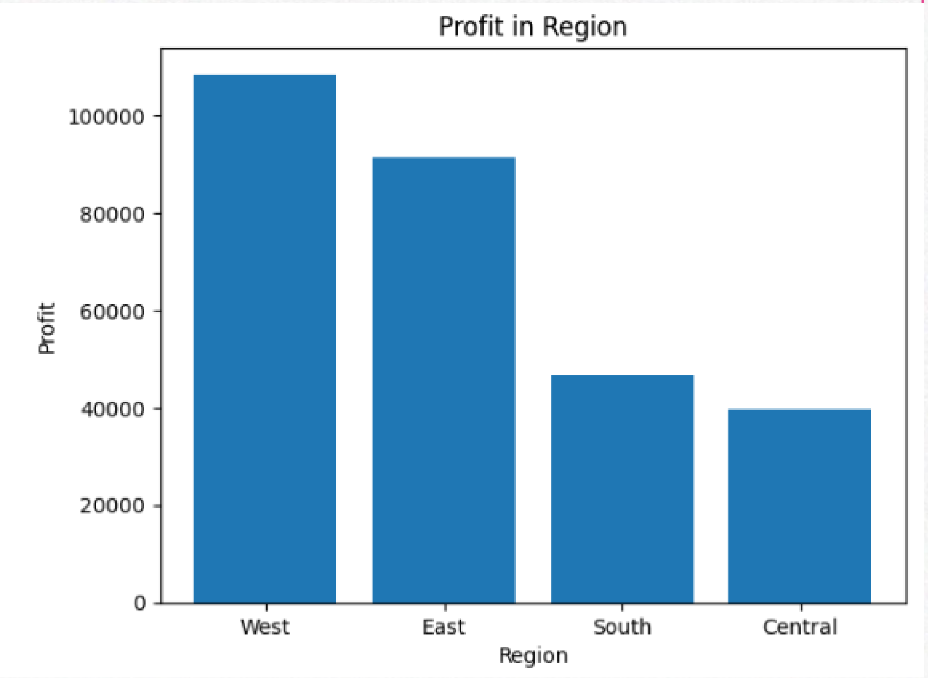
Top Regions by Sales

```
# Group the data by Region and calculate the total sales for each
region_sales = df_places.groupby(['Region'], as_index=False).sum()
region_sales.sort_values(by='Sales', ascending=False, inplace=True)

# Total sales by region
plt.bar(region_sales['Region'], region_sales['Sales'])
plt.xlabel("Region")
plt.ylabel("Sales")
plt.title("Sales in Region")
plt.show()
```



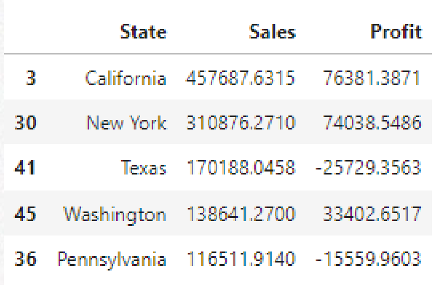
Top Regions by Profit

```
# Group the data by Region and calculate the
total profit for each
region_profit = df_places.groupby(['Region'],
as_index=False).sum()
region_profit.sort_values(by='Profit',
ascending=False, inplace=True)

# Profit in each region
plt.bar(region_profit['Region'],
region_profit['Profit'])
plt.xlabel("Region")
plt.ylabel("Profit")
plt.title("Profit in Region")
plt.show()
```

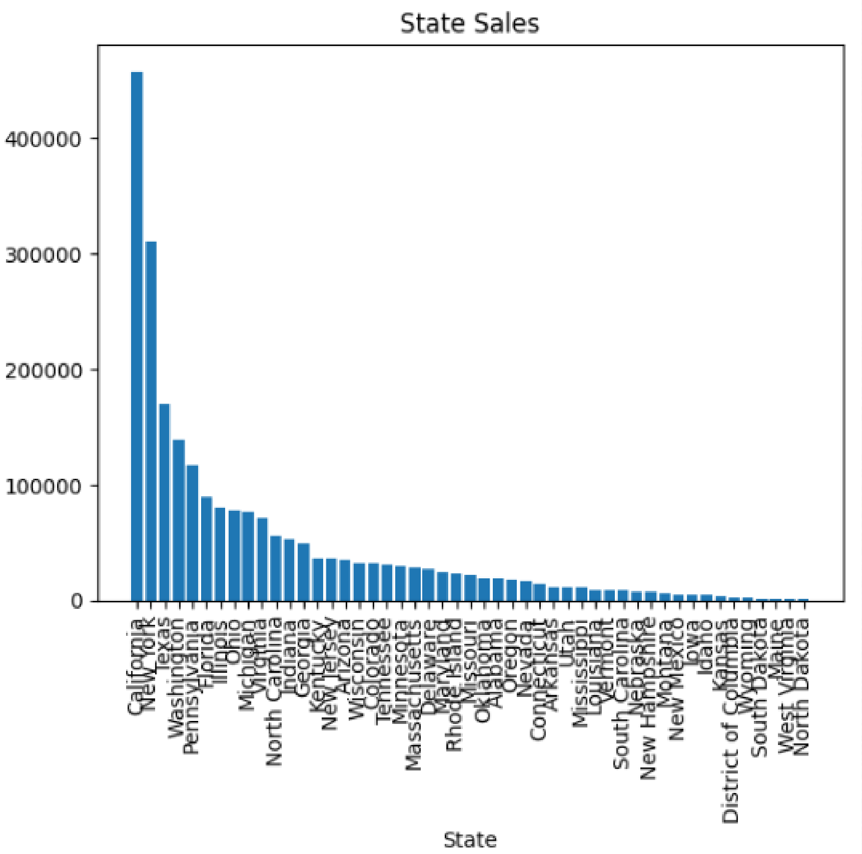


Top States by Sales

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

plt.bar(state_sales['State'],
state_sales['Sales'])
plt.xlabel("State")
plt.ylabel("Sales")
plt.title("State Sales")
plt.xticks(rotation=90)

plt.show()
state_sales.head()
```

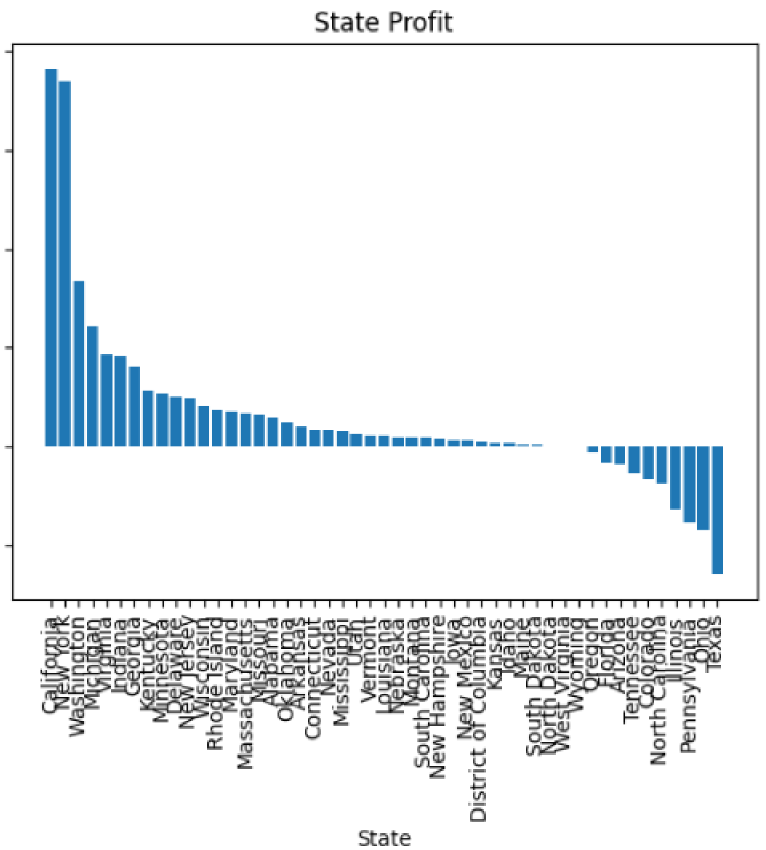


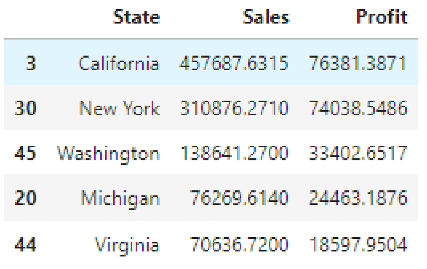
	State	Sales	Profit
3	California	457687.6315	76381.3871
30	New York	310876.2710	74038.5486
41	Texas	170188.0458	-25729.3563
45	Washington	138641.2700	33402.6517
36	Pennsylvania	116511.9140	-15559.9603



Top States by Profit

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


plt.bar(state_profit['State'], state_profit['Profit'])
plt.xlabel("State")
plt.ylabel("Profit")
plt.title("State Profit")
plt.xticks(rotation=90)

plt.show()
state_profit.head()
```

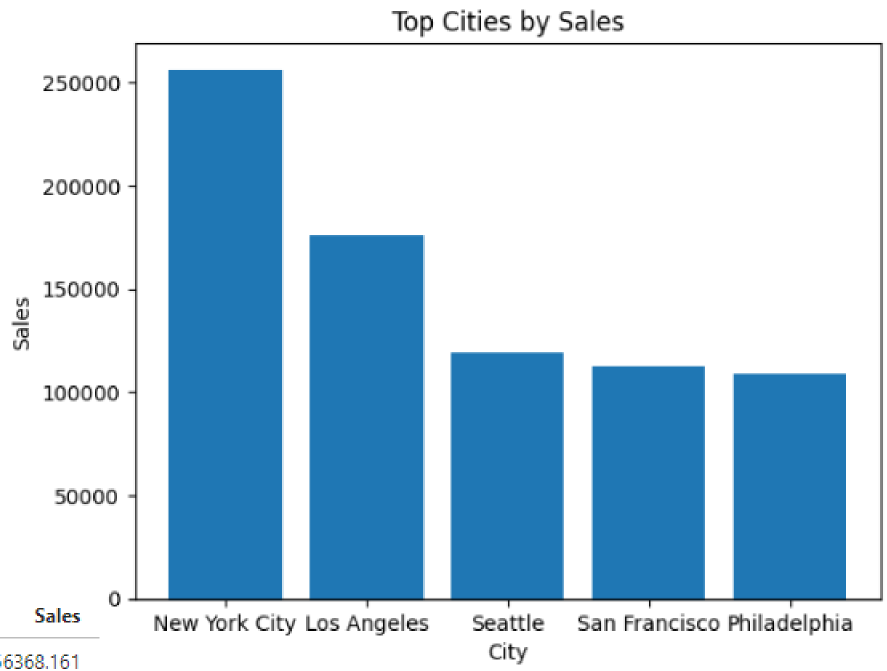


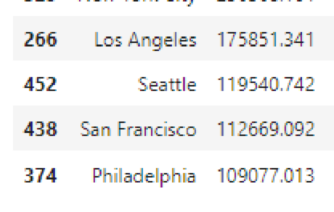


	State	Sales	Profit
3	California	457687.6315	76381.3871
30	New York	310876.2710	74038.5486
45	Washington	138641.2700	33402.6517
20	Michigan	76269.6140	24463.1876
44	Virginia	70636.7200	18597.9504

Top Cities by Sales

```
city_sales = df_places.groupby('City',
as_index=False).sum()
city_sales.sort_values(by='Sales',
ascending=False, inplace=True)

# Select the top 5 cities
top5_cities_sales = city_sales.head()

plt.bar(top5_cities_sales['City'],
top5_cities_sales['Sales'])
plt.xlabel("City")
plt.ylabel("Sales")
plt.title("Top Cities by Sales")

plt.show()
top5_cities_sales
```

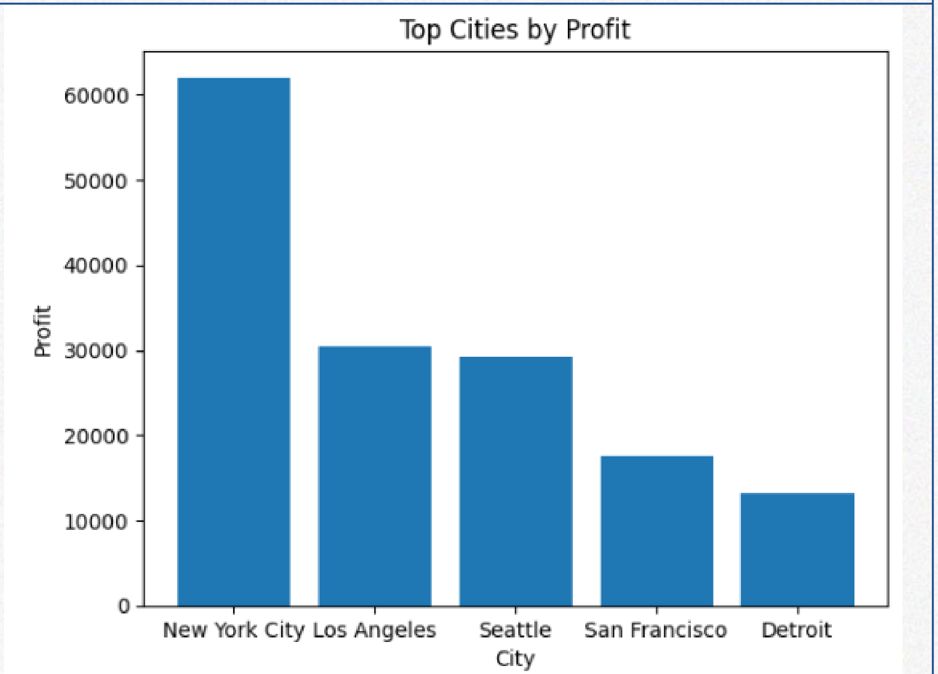


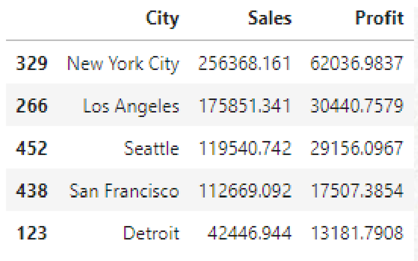


	City	Sales
329	New York City	256368.161
266	Los Angeles	175851.341
452	Seattle	119540.742
438	San Francisco	112669.092
374	Philadelphia	109077.013

Top Cities by Profit

```
city_profit = df_places.groupby('City',
as_index=False).sum()
city_profit.sort_values(by='Profit',
ascending=False, inplace=True)

# Select the top 5 cities
top5_cities_profit =city_profit.head()

plt.bar(top5_cities_profit['City'],
top5_cities_profit['Profit'])
plt.xlabel("City")
plt.ylabel("Profit")
plt.title("Top Cities by Profit")

plt.show()
top5_cities_profit
```

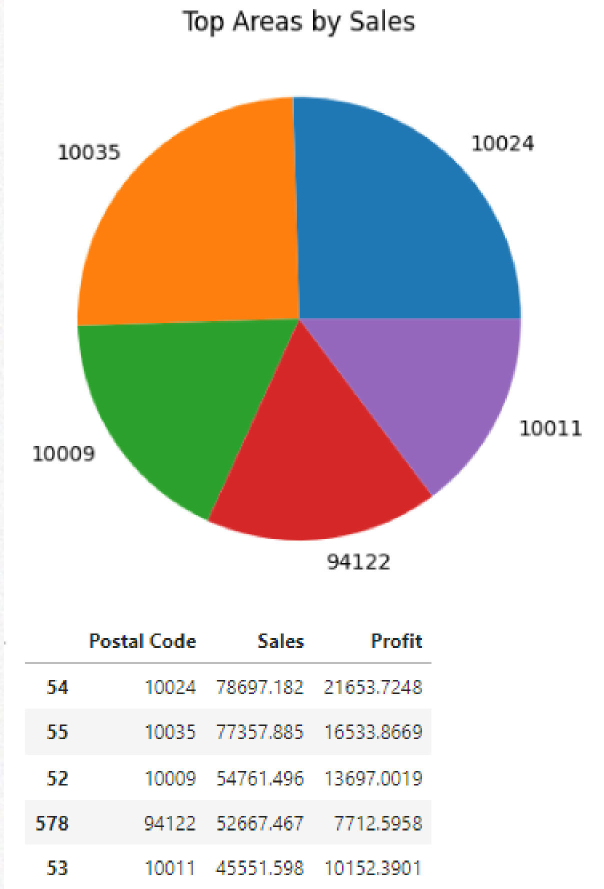




	City	Sales	Profit
329	New York City	256368.161	62036.9837
266	Los Angeles	175851.341	30440.7579
452	Seattle	119540.742	29156.0967
438	San Francisco	112669.092	17507.3854
123	Detroit	42446.944	13181.7908

Top Areas by Sales

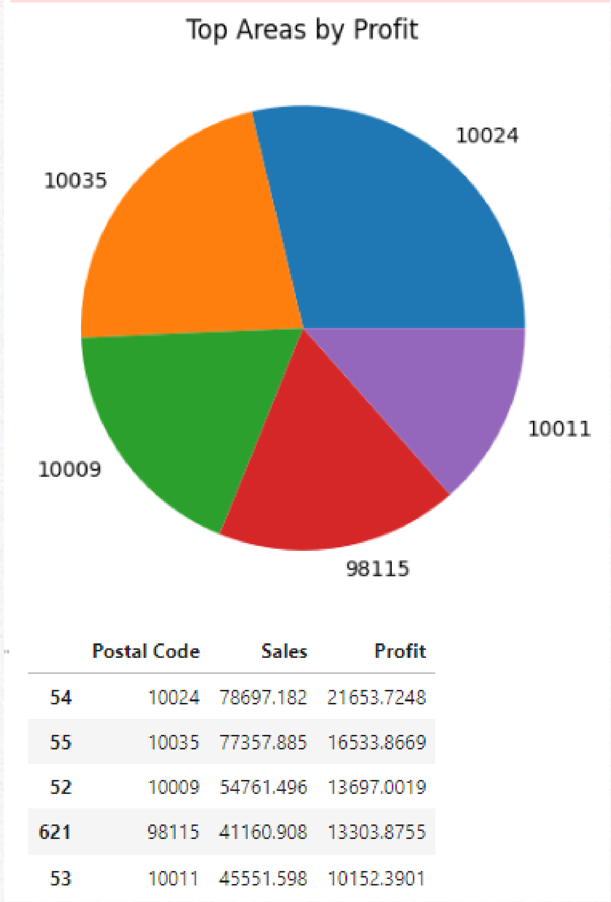
```
area_sales = df_places.groupby('Postal Code',
as_index=False).sum()
area_sales.sort_values(by='Sales',
ascending=False, inplace=True)

# Select the top 5 areas
top5_areas_sales = area_sales.head()
mylabels=(top5_areas_sales['Postal Code'])
y=np.array(top5_areas_sales['Sales'])
plt.pie(y, labels = mylabels)

plt.title("Top Areas by Sales")

plt.show()
top5_areas_sales
```



	Postal Code	Sales	Profit
54	10024	78697.182	21653.7248
55	10035	77357.885	16533.8669
52	10009	54761.496	13697.0019
578	94122	52667.467	7712.5958
53	10011	45551.598	10152.3901

Top Areas by Profit

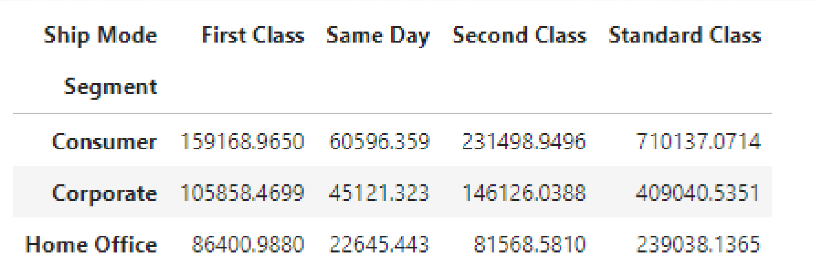
```
area_profit = df_places.groupby('Postal Code',
as_index=False).sum()
area_profit.sort_values(by='Profit',
ascending=False, inplace=True)

# Select the top 5 areas
top5_areas_profit = area_profit.head()
mylabels=(top5_areas_profit['Postal Code'])
y=np.array(top5_areas_profit['Profit'])
plt.pie(y, labels = mylabels)

plt.title("Top Areas by Profit")

plt.show()
top5_areas_profit
```



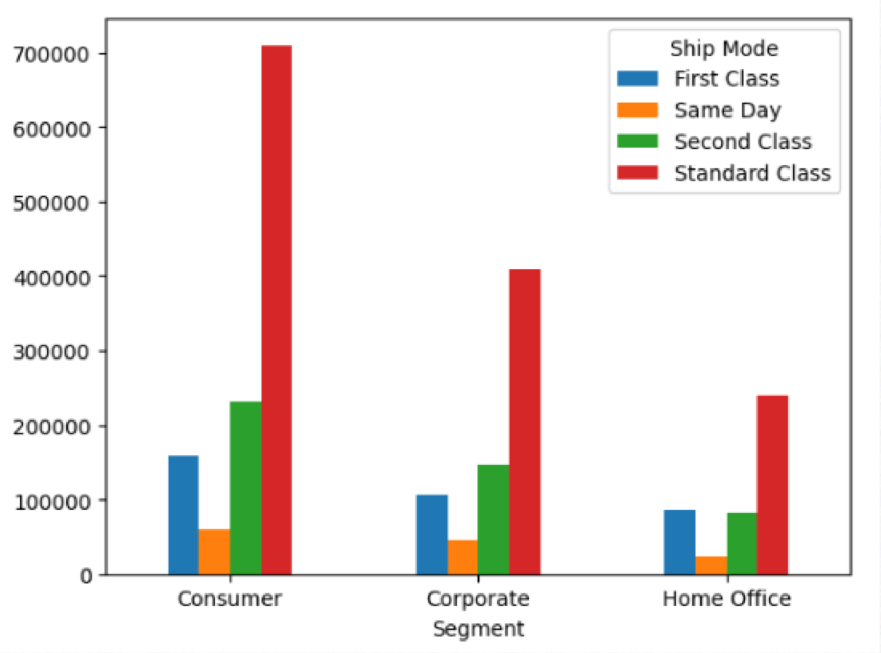
	Postal Code	Sales	Profit
54	10024	78697.182	21653.7248
55	10035	77357.885	16533.8669
52	10009	54761.496	13697.0019
621	98115	41160.908	13303.8755
53	10011	45551.598	10152.3901

Most Active Segment and Mode

```
#Related Sales
table= df.pivot_table(index='Segment',
columns='Ship Mode', values='Sales',
aggfunc='sum')
table.plot(kind='bar')
plt.xticks(rotation=0)
plt.show()

table
```



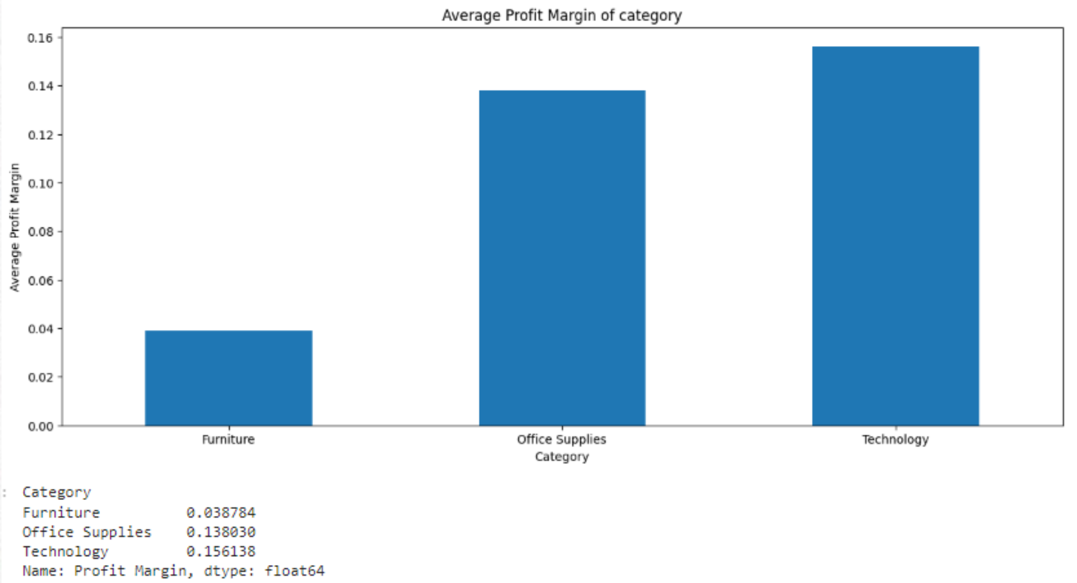
Ship Mode	First Class	Same Day	Second Class	Standard Class
Segment				
Consumer	159168.9650	60596.359	231498.9496	710137.0714
Corporate	105858.4699	45121.323	146126.0388	409040.5351
Home Office	86400.9880	22645.443	81568.5810	239038.1365



Results

BEST SALES

```
df['Profit Margin'] = df['Profit'] / df['Sales']
# Group category and data and calculate the average profit margin for each
avg_profit_margin = df.groupby('Category')['Profit Margin'].mean()
plt.figure(figsize=(15,6))
avg_profit_margin.plot(kind='bar')

plt.title("Average Profit Margin of category")
plt.xlabel("Category")
plt.ylabel("Average Profit Margin")
plt.xticks(rotation=0)
plt.show()

avg_profit_margin
```



```
:  Category
   Furniture          0.038784
   Office Supplies    0.138030
   Technology         0.156138
   Name: Profit Margin, dtype: float64
```

Conclusion

The study of the superstore dataset revealed useful insights into sales trends, customer behavior, and product performance, allowing data-driven recommendations to optimize business operations and increase overall profitability. The project's findings provide a strategic roadmap for decision-making and enhancing the competitiveness of the superstore in the market for store management, marketing teams, and executives.

Best Region : [West]
Best State : [california, New York]
Best Cities: [New York City, Los Angeles, Seattle, San Francisco, Detroit]
Best Areas: [10024,10035,1009]

Category with highest avg profit margin - **Technology** (*0.156*)
Most active sales segment - **Consumer**
Most used Ship mode - **Standard Class**

Thank you!

Credits - https://github.com/SahilRT