Innobyte Services Internship

Project : Data Analysis of Superstore Retail Sales dataset

Introduction :

Welcome to the Sales Store Analysis notebook! In this notebook, we will be delving into the intricate details of sales data from our store. The objective of this analysis is to gain valuable insights into our sales performance, understand customer behavior, identify trends, and ultimately make data-driven decisions to improve our business operations.

Throughout this analysis, we will explore various aspects of our sales data, including but not limited to:

1. Sales Trends: Examining overall sales trends over time to identify any seasonal patterns or fluctuations.
2. Product Performance: Analyzing the performance of individual products or product categories to identify top-selling items and areas for improvement.
3. Customer Segmentation: Understanding our customer base by segmenting them based on demographics, purchasing behavior, or other relevant factors.
4. Geographical Analysis: Investigating sales performance across different regions to identify geographical trends and opportunities.
5. The columns available in dataset

Importing the Libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
from scipy import stats
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from xgboost import XGBRegressor
from sklearn.pipeline import Pipeline
from sklearn.metrics import r2_score, mean_squared_error
import time
```

Data Exploration

Loading the dataset

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

df.head()
```

```
        Ship Mode    Segment        Country             City
State  \
0    Second Class   Consumer  United States        Henderson
```

```
Kentucky
1    Second Class   Consumer  United States        Henderson
Kentucky
2    Second Class  Corporate  United States      Los Angeles
California
3  Standard Class   Consumer  United States  Fort Lauderdale
Florida
4  Standard Class   Consumer  United States  Fort Lauderdale
Florida

   Postal Code Region         Category Sub-Category     Sales
Quantity  \
0        42420  South        Furniture    Bookcases  261.9600
2
1        42420  South        Furniture       Chairs  731.9400
3
2        90036   West  Office Supplies       Labels   14.6200
2
3        33311  South        Furniture       Tables  957.5775
5
4        33311  South  Office Supplies      Storage   22.3680
2

   Discount    Profit
0      0.00   41.9136
1      0.00  219.5820
2      0.00    6.8714
3      0.45 -383.0310
4      0.20    2.5164
```

```
df.shape
```

```
(9994, 13)
```

Columns in Dataset

```
df.columns
```

```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal
Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity',
'Discount',
       'Profit'],
      dtype='object')
```

```
df.dtypes
```

```
Ship Mode        object
Segment          object
Country          object
City             object
```

```
State             object
Postal Code        int64
Region            object
Category          object
Sub-Category      object
Sales            float64
Quantity           int64
Discount         float64
Profit           float64
dtype: object
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Ship Mode     9994 non-null   object
 1   Segment       9994 non-null   object
 2   Country       9994 non-null   object
 3   City          9994 non-null   object
 4   State         9994 non-null   object
 5   Postal Code   9994 non-null   int64
 6   Region        9994 non-null   object
 7   Category      9994 non-null   object
 8   Sub-Category  9994 non-null   object
 9   Sales         9994 non-null   float64
 10  Quantity      9994 non-null   int64
 11  Discount      9994 non-null   float64
 12  Profit        9994 non-null   float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

Data cleaning and preprocessing

Handling Null Values and Duplicates.

```
# Checking for null values
df.isnull().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
```

Observation : No Missing Values find

```
#checking for duplicates value

df.duplicated().sum()

17

df.drop_duplicates(inplace=True)

# Removing duplicate values

df

           Ship Mode    Segment        Country             City
State  \
0       Second Class   Consumer  United States        Henderson
Kentucky
1       Second Class   Consumer  United States        Henderson
Kentucky
2       Second Class  Corporate  United States      Los Angeles
California
3     Standard Class   Consumer  United States  Fort Lauderdale
Florida
4     Standard Class   Consumer  United States  Fort Lauderdale
Florida
...              ...        ...            ...              ...
...
9989    Second Class   Consumer  United States            Miami
Florida
9990  Standard Class   Consumer  United States       Costa Mesa
California
9991  Standard Class   Consumer  United States       Costa Mesa
California
9992  Standard Class   Consumer  United States       Costa Mesa
California
9993    Second Class   Consumer  United States      Westminster
California

      Postal Code Region         Category Sub-Category     Sales
Quantity  \
0           42420  South        Furniture    Bookcases  261.9600
```

```
2
1           42420  South        Furniture       Chairs  731.9400
3
2           90036   West  Office Supplies       Labels   14.6200
2
3           33311  South        Furniture       Tables  957.5775
5
4           33311  South  Office Supplies      Storage   22.3680
2
...           ...    ...              ...          ...       ...
...
9989        33180  South        Furniture  Furnishings   25.2480
3
9990        92627   West        Furniture  Furnishings   91.9600
2
9991        92627   West       Technology       Phones  258.5760
2
9992        92627   West  Office Supplies        Paper   29.6000
4
9993        92683   West  Office Supplies   Appliances  243.1600
2

      Discount    Profit
0         0.00   41.9136
1         0.00  219.5820
2         0.00    6.8714
3         0.45 -383.0310
4         0.20    2.5164
...        ...       ...
9989      0.20    4.1028
9990      0.00   15.6332
9991      0.20   19.3932
9992      0.00   13.3200
9993      0.00   72.9480

[9977 rows x 13 columns]
```

```
df.shape
```

```
(9977, 13)
```

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

```
0
```

Statistical Summary of data

```
df.describe()
```

```
        Postal Code         Sales     Quantity     Discount
Profit
count   9977.000000   9977.000000  9977.000000  9977.000000
9977.00000
mean   55154.964117    230.148902     3.790719     0.156278
28.69013
std    32058.266816    623.721409     2.226657     0.206455
234.45784
min     1040.000000      0.444000     1.000000     0.000000 -
6599.97800
25%    23223.000000     17.300000     2.000000     0.000000
1.72620
50%    55901.000000     54.816000     3.000000     0.200000
8.67100
75%    90008.000000    209.970000     5.000000     0.200000
29.37200
max    99301.000000  22638.480000    14.000000     0.800000
8399.97600
```

Describe method shows:

There are 9977 records (sales)

Sales values are in the range of 0.444000 to 22,638.48 with average 230.77 and standard deviation of 623.72

```
df.describe(include="object")
```

```
             Ship Mode   Segment        Country           City
State  \
count             9977      9977           9977           9977
9977
unique               4         3              1            531
49
top     Standard Class  Consumer  United States  New York City
California
freq              5955      5183           9977            914
1996

       Region         Category Sub-Category
count    9977             9977         9977
unique      4                3           17
top      West  Office Supplies      Binders
freq     3193             6012         1522
```

Exploring Unique Values.

```
df.nunique()
```

```
Ship Mode          4
Segment            3
Country            1
City             531
State             49
Postal Code      631
Region             4
Category           3
Sub-Category      17
Sales           5825
Quantity          14
Discount          12
Profit          7287
dtype: int64
```

Let's see how many unique values in each of State, Category, Sub-Category, and Ship Mode

```
print('* There are stores in {}
states'.format(len(df['State'].unique())))

print('* There are {} different
categories'.format(len(df['Category'].unique())))

print('* There are {} different sub categories'.format(len(df['Sub-
Category'].unique())))

print('* There are {} different ship mode'.format(len(df['Ship
Mode'].unique())))
```

```
* There are stores in 49 states
* There are 3 different categories
* There are 17 different sub categories
* There are 4 different ship mode
```

```
# Total sales and Profit

print('Total profit of the superstore:',df['Profit'].sum())
```

```
Total profit of the superstore: 286241.4226
```

```
print('Total sales of the superstore:',df['Sales'].sum())
```

```
Total sales of the superstore: 2296195.5903
```

Customer segmentation

```
# Types of unique values in segment
df['Segment'].unique()
```

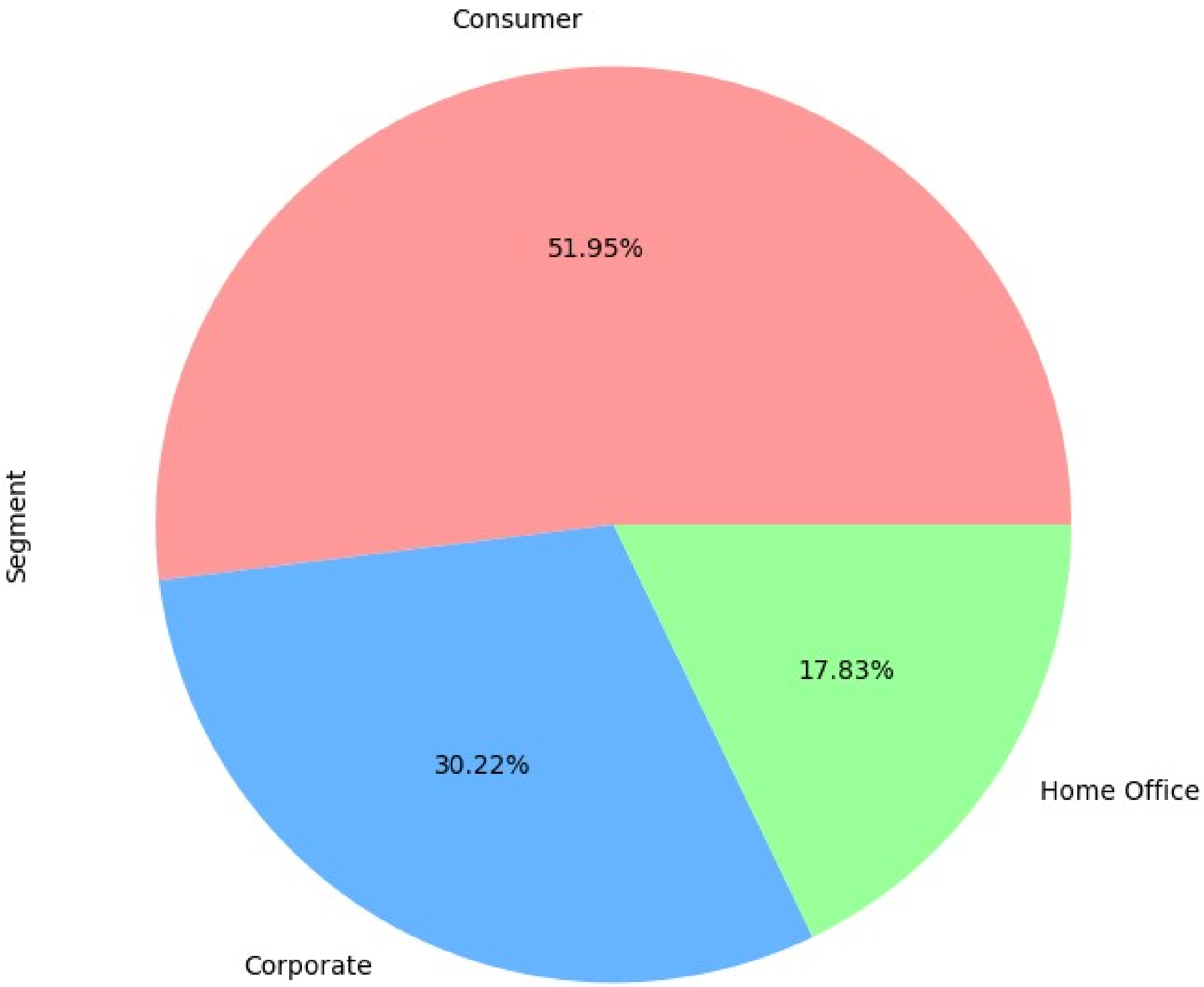
```
array(['Consumer', 'Corporate', 'Home Office'], dtype=object)
```

```
# No. of unique values in each segment
df['Segment'].value_counts()
```

```
Consumer       5183
Corporate      3015
Home Office    1779
Name: Segment, dtype: int64
```

```
df['Segment'].value_counts().plot(kind='pie', autopct = '%1.2f%%',
colors = ['#ff9999','#66b3ff','#99ff99'])
```

```
<Axes: ylabel='Segment'>
```

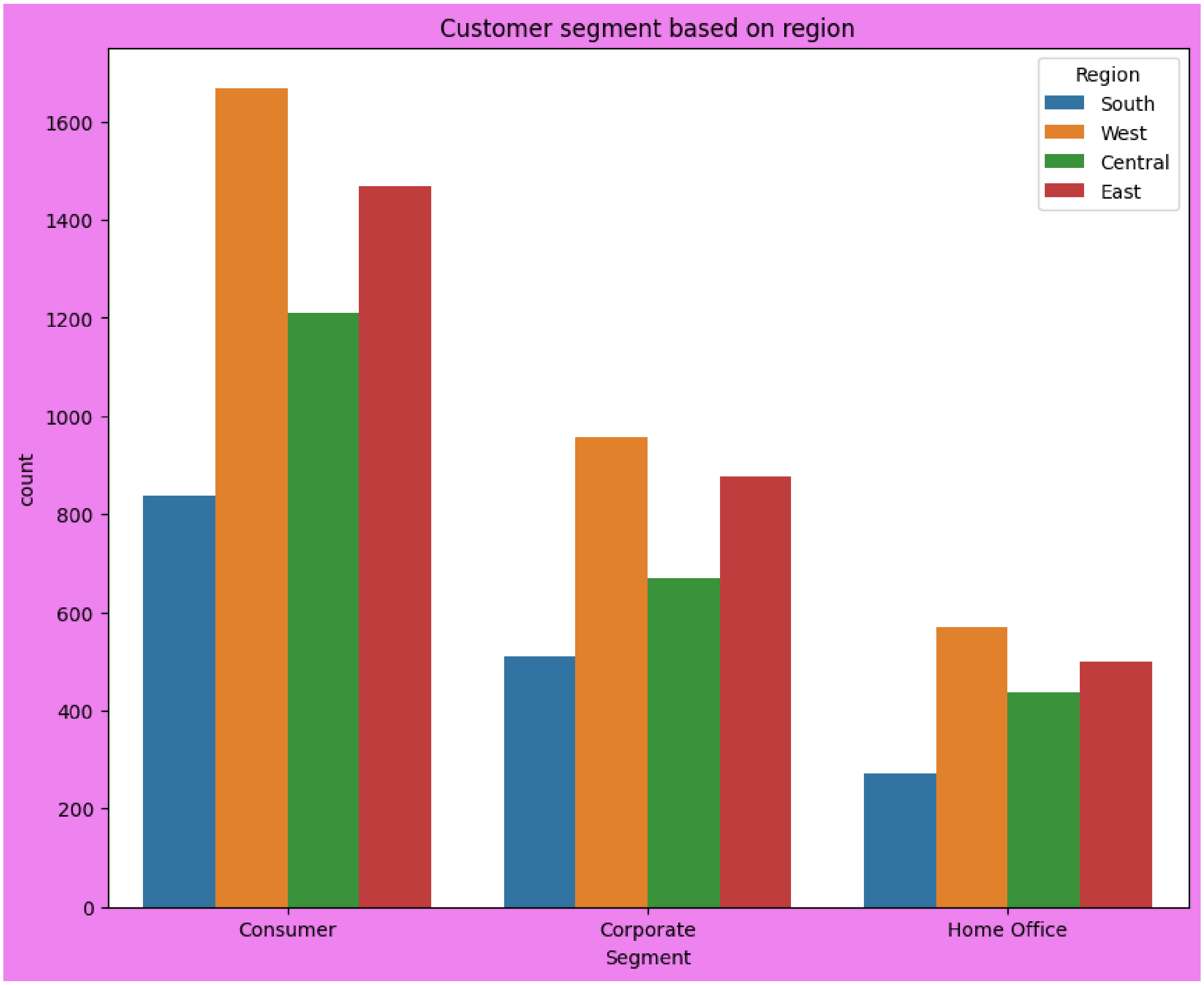


Observation :

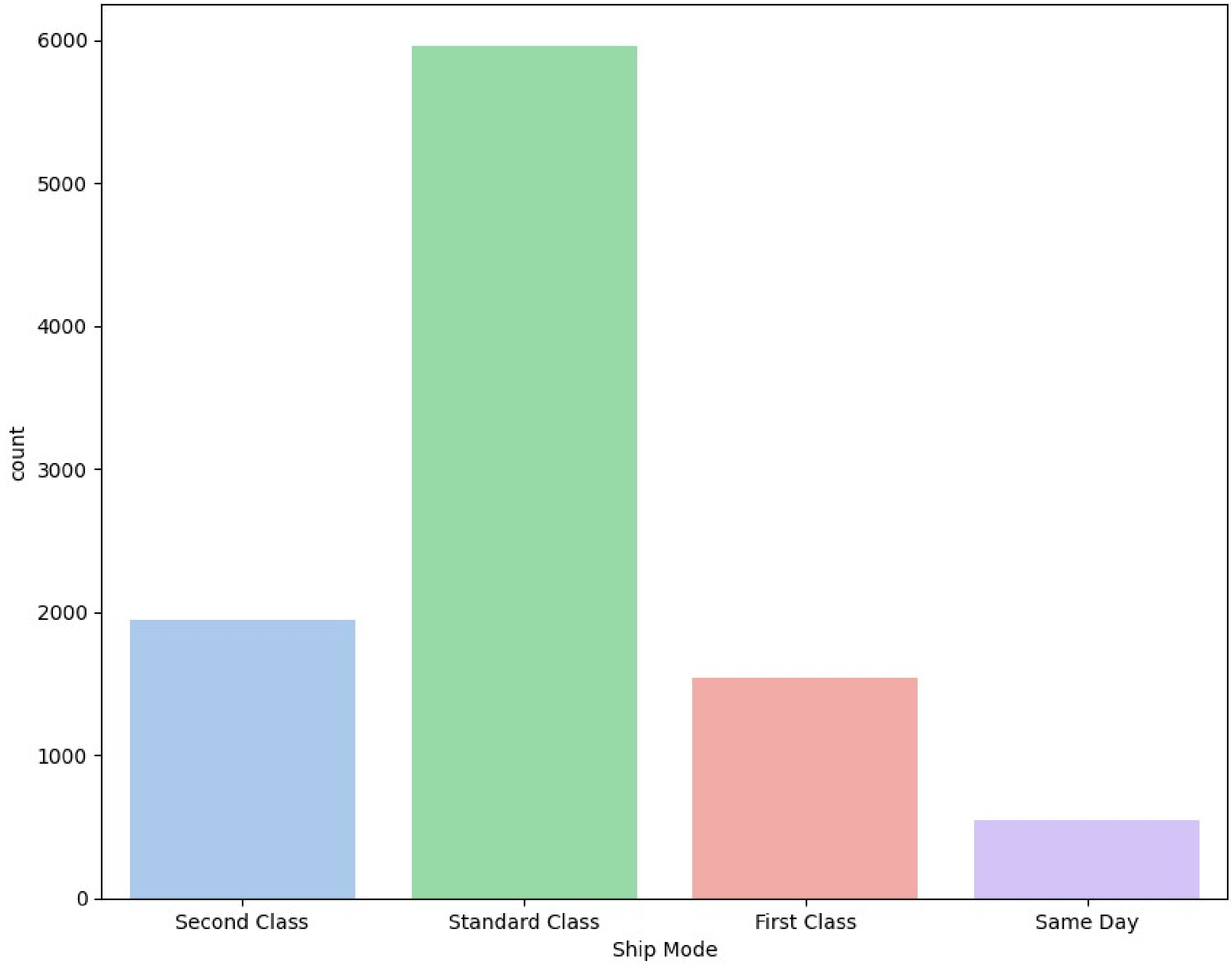
Around 50% of customers in the dataset are classified as consumers, indicating a significant portion of individual buyers among the customer base.

```
plt.figure(facecolor='violet')
plt.title('Customer segment based on region')
sns.countplot(x='Segment',data=df,hue='Region')
```

```
<Axes: title={'center': 'Customer segment based on region'},
xlabel='Segment', ylabel='count'>
```



```
df['Ship Mode'].unique()

array(['Second Class', 'Standard Class', 'First Class', 'Same Day'],
      dtype=object)

df['Ship Mode'].value_counts()

Standard Class    5955
Second Class      1943
First Class       1537
Same Day           542
Name: Ship Mode, dtype: int64

sns.countplot(data=df,x='Ship Mode',palette=['#a1c9f4', '#8de5a1',
'#ff9f9b', '#d0bbff'])
plt.show()
```



Observation :

The majority of customers prefer the standard class ship mode compared to other options like first class, second class, or same day.

Product Analysis

Analysis of Order Count Distribution Across Various Categories.

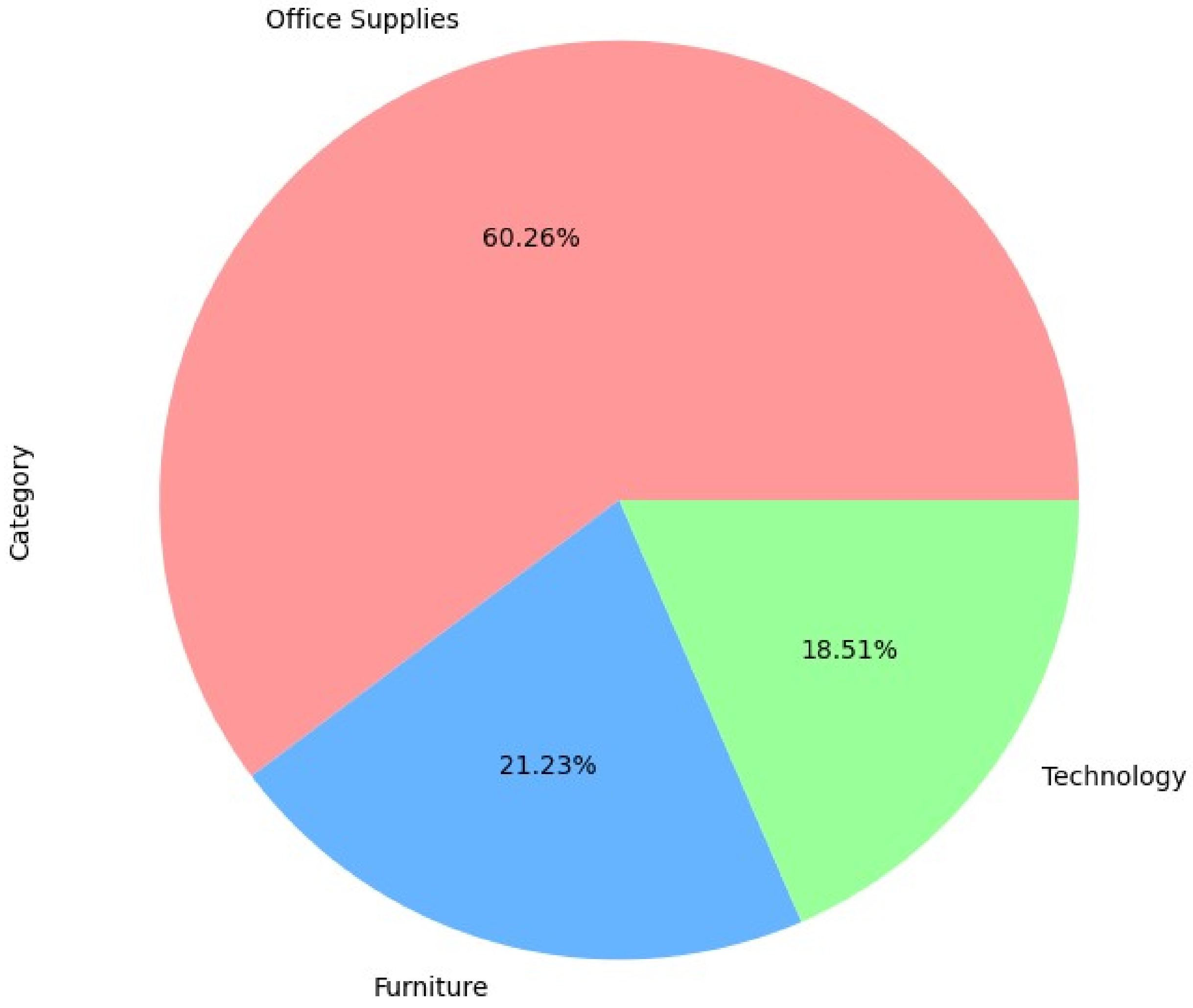
```
df['Category'].unique()
```

```
array(['Furniture', 'Office Supplies', 'Technology'], dtype=object)
```

```
df['Category'].value_counts()
```

```
Office Supplies    6012
Furniture          2118
```

```
Technology         1847
Name: Category, dtype: int64

df['Category'].value_counts().plot(kind='pie', autopct = '%1.2f%%',
colors = ['#ff9999','#66b3ff','#99ff99'])

<Axes: ylabel='Category'>
```

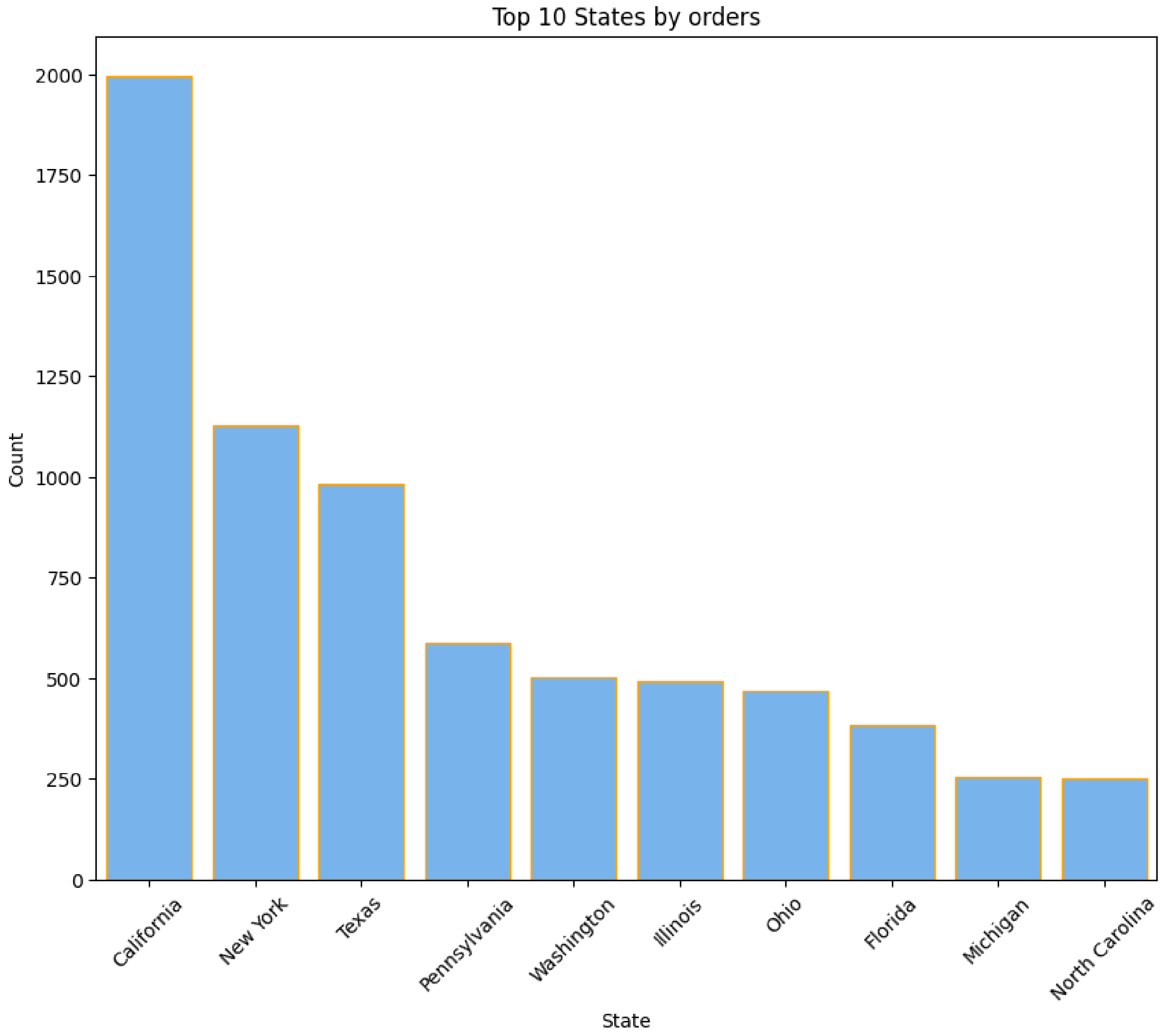


Observation :

Above 60% of customers in the dataset place orders for office supplies.

Distribution of orders count across top 10 states

```
state=df['State'].value_counts().index[:10]
count=df['State'].value_counts().values[:10]
sns.barplot(x=state,y=count,data=df,color=
'#66b3ff',edgecolor='orange')
plt.xticks(rotation=45)
plt.xlabel('State')
plt.ylabel('Count')
plt.title('Top 10 States by orders')
plt.show()
```



Observation :

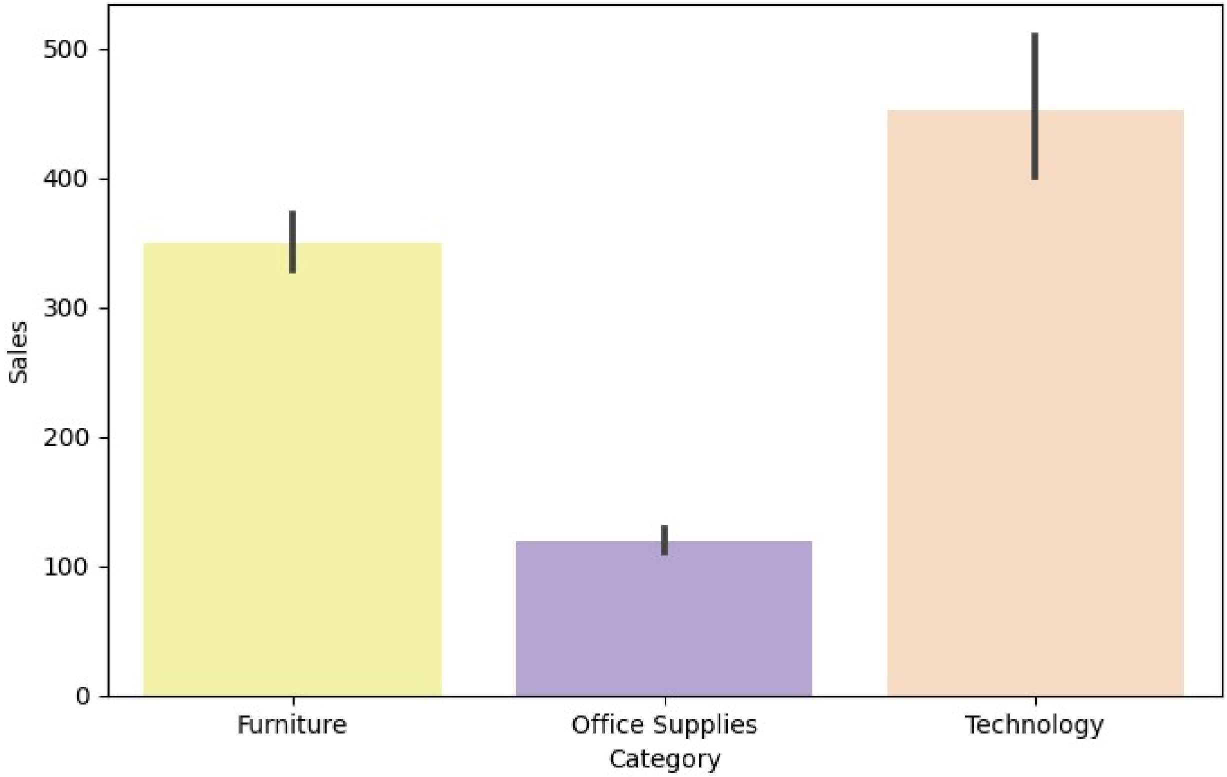
The plot above displays the top 10 cities by some metric, where California standing out as having the highest number of order counts.

Analysis of Sales Distribution.

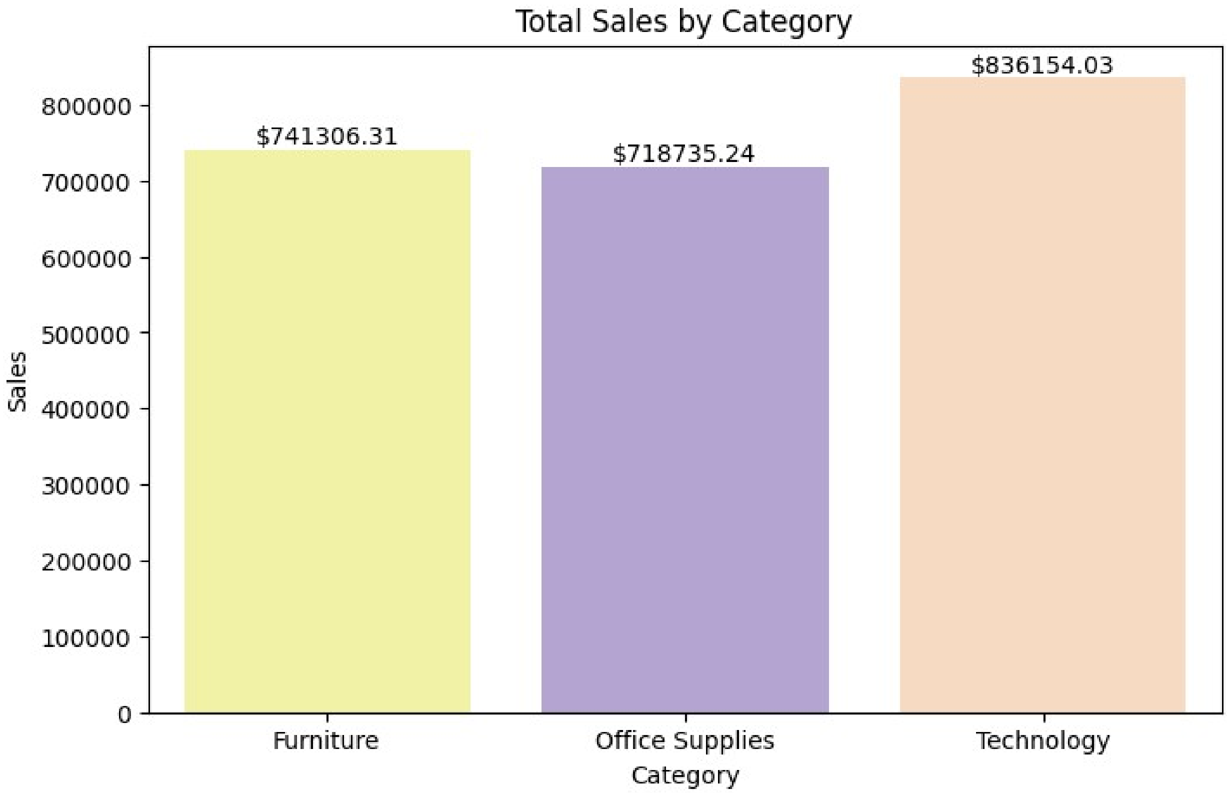
This title conveys that you have conducted an analysis based on the sales column, comparing it with different categories.

```
plt.figure(figsize=(8,5))
sns.barplot(data=df,x="Category",y="Sales",palette=["#FFFF99",
"#B19CD9", "#FFDAB9"])
```

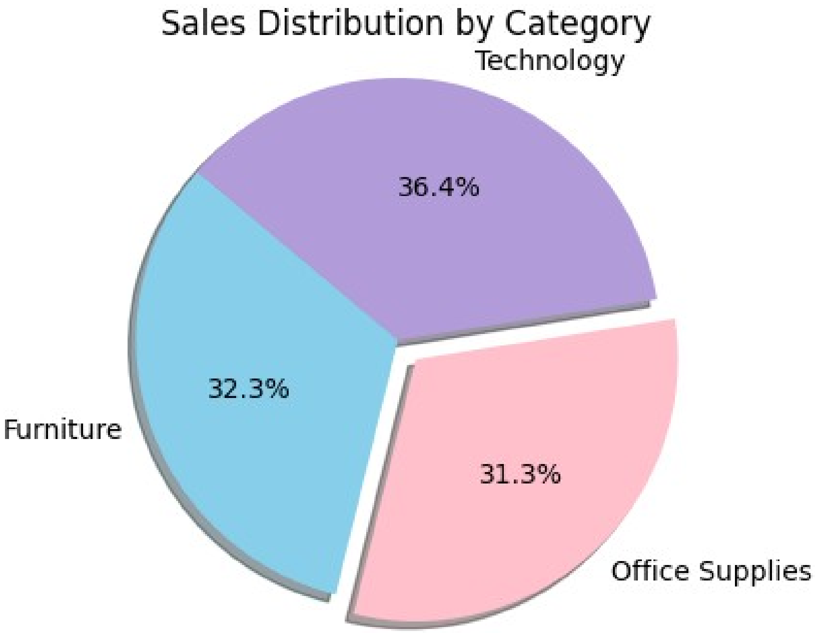
```
<Axes: xlabel='Category', ylabel='Sales'>
```



```
cat_s=df.groupby("Category")["Sales"].sum().reset_index()
plt.figure(figsize=(8,5))
sns.barplot(data=cat_s,x="Category",y="Sales",palette=["#FFFF99",
"#B19CD9", "#FFDAB9"])

for index, row in cat_s.iterrows():
    plt.annotate(f"${row['Sales']:.2f}", (index, row['Sales']),
ha='center', va='bottom')

plt.xlabel('Category')
plt.ylabel('Sales')
plt.title('Total Sales by Category')
plt.show()
```



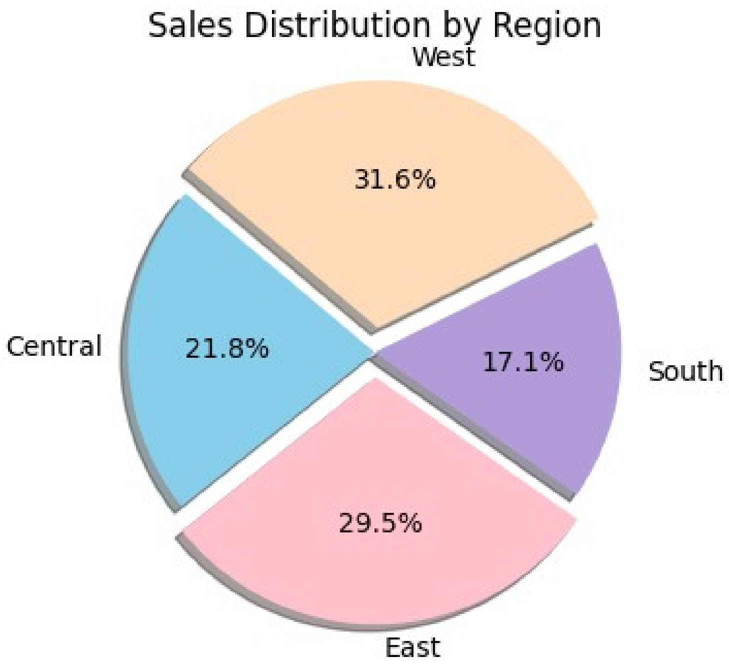
```
reg_s=df.groupby("Category")["Sales"].sum().reset_index()
sales=reg_s["Sales"].tolist()
category=reg_s["Category"].tolist()


plt.figure(figsize=(8,4))
plt.pie(sales, labels=category,
        autopct='%1.1f%%',
        colors=["#87CEEB", "#FFC0CB", "#B19CD9"],
        shadow=True,
        explode = [0, 0.1, 0],
        startangle=140)
plt.title('Sales Distribution by Category')
plt.axis('equal')
plt.show()
```



Observation :

The plot above depicts the sales distribution across different categories, highlighting the Technology as the leader in terms of sales.

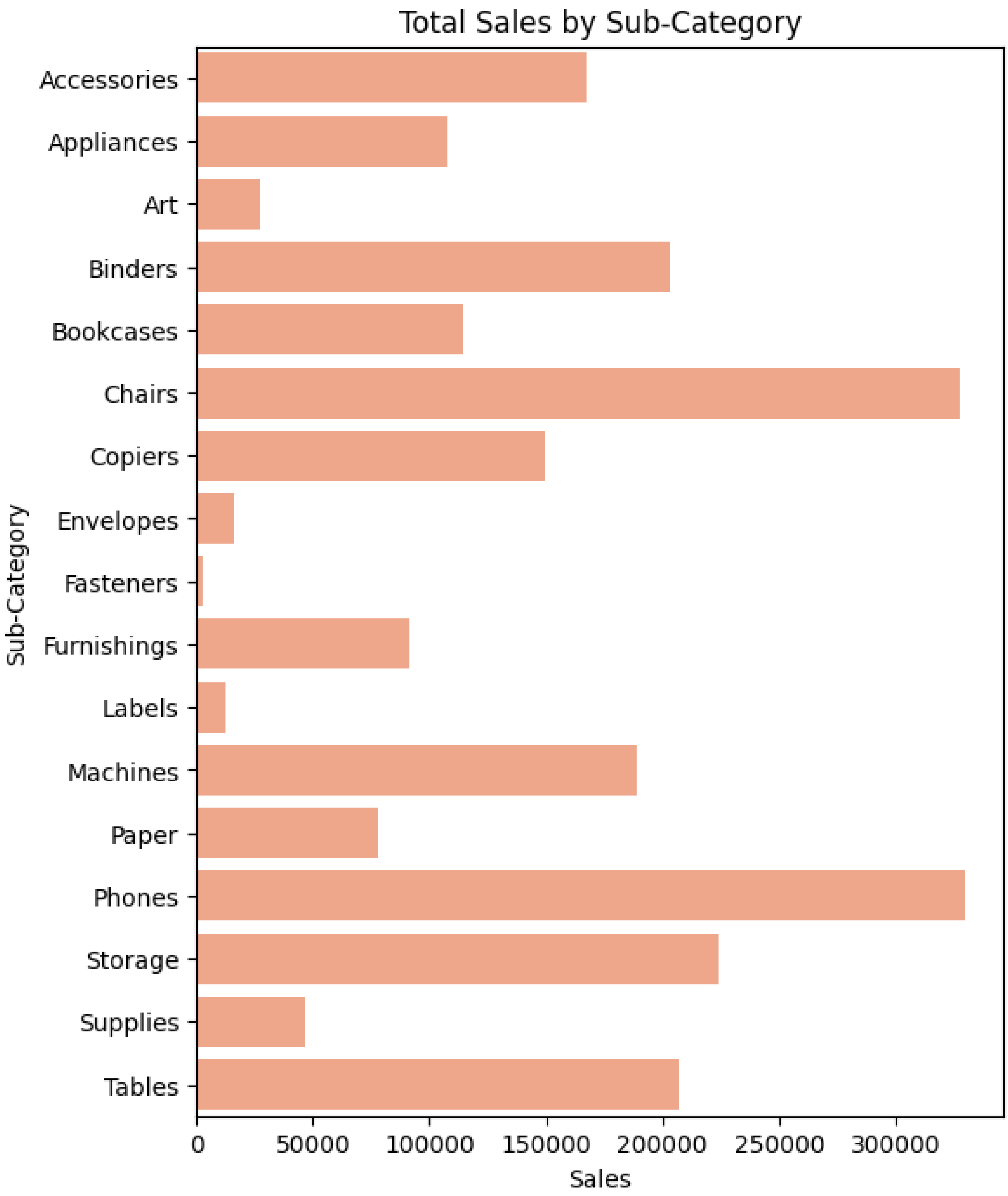
```
reg_s=df.groupby("Region")["Sales"].sum().reset_index()
sales=reg_s["Sales"].tolist()
regions=reg_s["Region"].tolist()

plt.figure(figsize=(8,4))
plt.pie(sales, labels=regions,
        autopct='%1.1f%%',
        colors=["#87CEEB", "#FFC0CB", "#B19CD9", "#FFDAB9"],
        shadow=True,
        explode = [0, 0.1, 0, 0.1],
        startangle=140)
plt.title('Sales Distribution by Region')
plt.axis('equal')
plt.show()
```



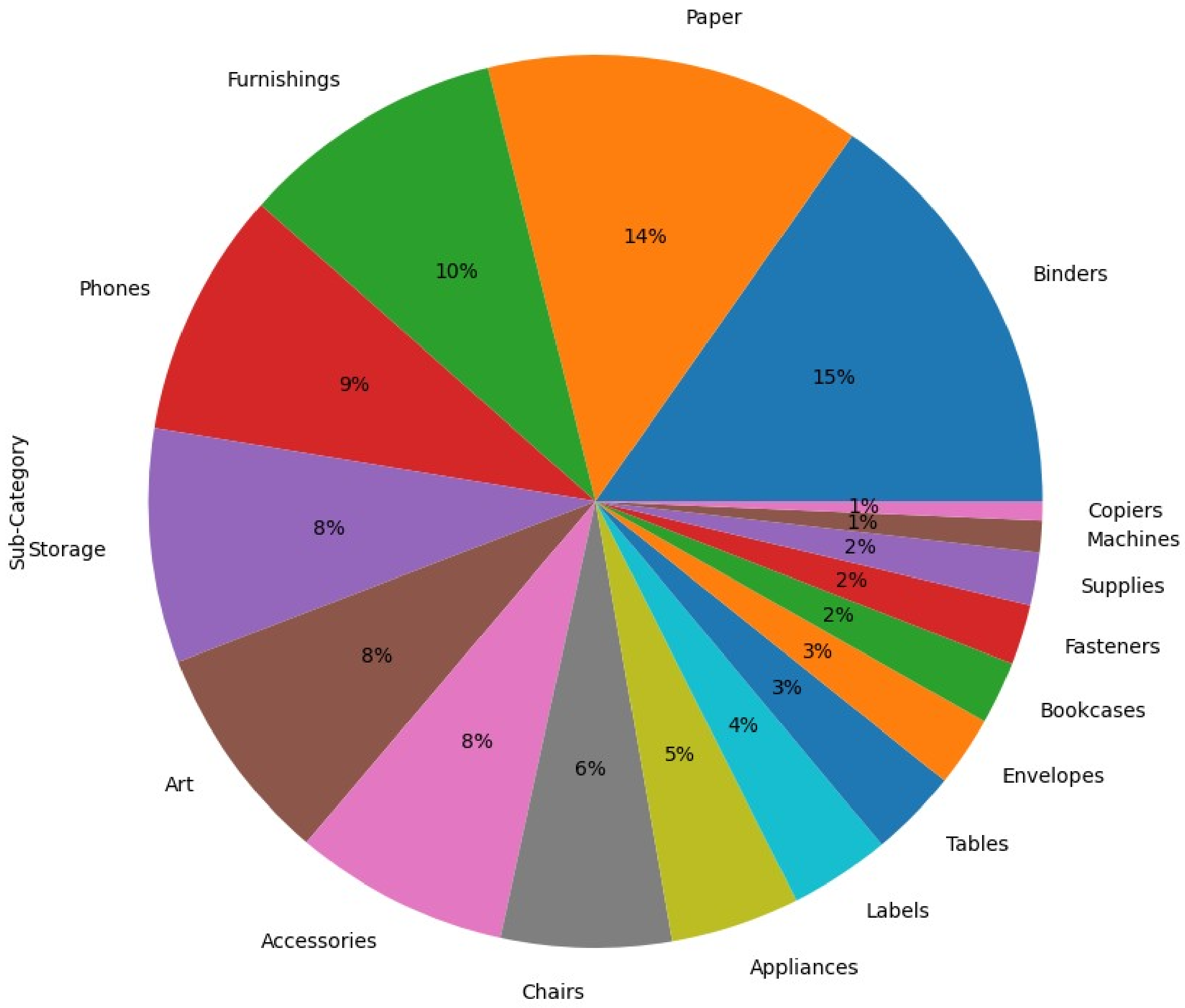
Observation :

The plot above illustrates the distribution of sales by region. The West region stands out with the highest sales.

```
scat_s=df.groupby("Sub-Category")["Sales"].sum().reset_index()

plt.figure(figsize=(6,8))
sns.barplot(data=scat_s,y="Sub-Category",x="Sales",color="#FFA07A")
plt.xlabel('Sales')
plt.ylabel('Sub-Category')
plt.title('Total Sales by Sub-Category')
plt.show()
```



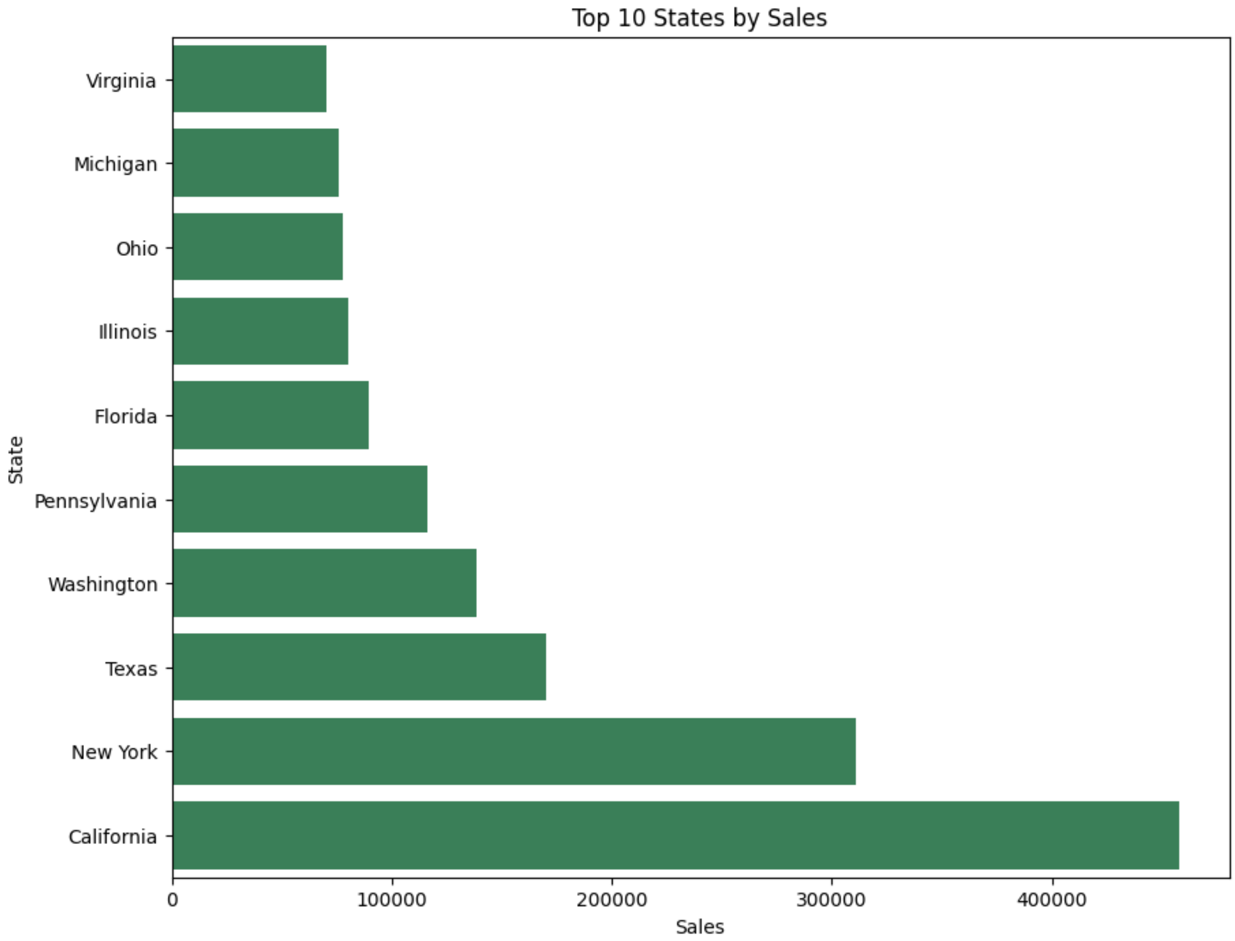
```
plt.figure(figsize=(12,10))
df["Sub-Category"].value_counts().plot.pie(autopct="%1.0f%%")
plt.show()
```



Observation

Chairs and Tables have high sales, both around $300,000.

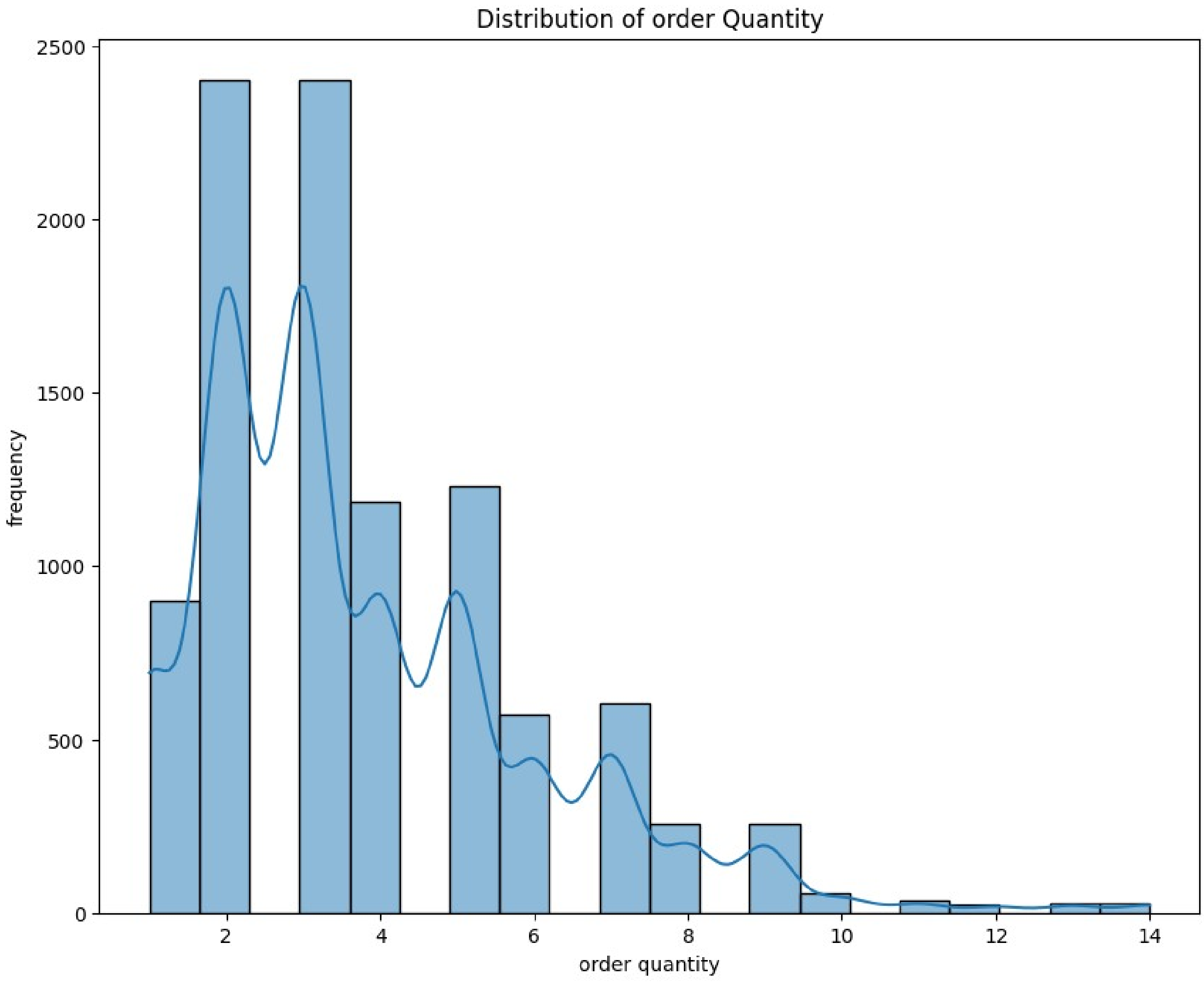
```
sta_s=df.groupby("State")["Sales"].sum().reset_index()
sta_s=sta_s.sort_values(by="Sales")
sta_s=sta_s.tail(10)

sns.barplot(data=sta_s,y="State",x="Sales",color="seagreen")
plt.xlabel('Sales')
plt.ylabel('State')
plt.title('Top 10 States by Sales')
plt.show()
```



Observation :

The distribution of sales across states reveals a notable disparity, with California and New York leading in sales volume, suggesting strong market presence and economic activity

```
sns.histplot(df['Quantity'], bins=20, kde=True)
plt.title('Distribution of order Quantity')
plt.xlabel('order quantity')
plt.ylabel('frequency')
plt.show()
```



```python
count_sub=df.groupby(["Category","Sub-Category"]).size().reset_index(name='Count')
fig = px.bar(count_sub, x='Category', y='Count', color='Sub-Category',
             title='Counts of Sub-Categories within Main Categories',
             labels={'Count': 'Number of Items Sold'},
             barmode='group')
fig.show()
```

```
{"config":{"plotlyServerURL":"https://plot.ly"},"data":[{"alignmentgroup":"True","hovertemplate":"Sub-Category=Bookcases<br>Category=%{x}<br>Number of Items Sold=%{y}<extra></extra>","legendgroup":"Bookcases","marker":{"color":"#636efa","pattern":{"shape":""}},"name":"Bookcases","offsetgroup":"Bookcases","orientation":"v","showlegend":true,"textposition":"auto","type":"bar","x":["Furniture"],"xaxis":"x","y":[228],"yaxis":"y"},{"alignmentgroup":"True","hovertemplate":"Sub-Category=Chairs<br>Category=%{x}<br>Number of Items Sold=%{y}<extra></extra>","legendgroup":"Chairs","marker":
```

```
{"color":"#EF553B","pattern":
{"shape":""}},"name":"Chairs","offsetgroup":"Chairs","orientation":"v"
,"showlegend":true,"textposition":"auto","type":"bar","x":
["Furniture"],"xaxis":"x","y":[615],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Furnishings<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Furnishings","marker":
{"color":"#00cc96","pattern":
{"shape":""}},"name":"Furnishings","offsetgroup":"Furnishings","orient
ation":"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Furniture"],"xaxis":"x","y":[956],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Tables<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Tables","marker":
{"color":"#ab63fa","pattern":
{"shape":""}},"name":"Tables","offsetgroup":"Tables","orientation":"v"
,"showlegend":true,"textposition":"auto","type":"bar","x":
["Furniture"],"xaxis":"x","y":[319],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Appliances<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Appliances","marker":
{"color":"#FFA15A","pattern":
{"shape":""}},"name":"Appliances","offsetgroup":"Appliances","orientat
ion":"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Office Supplies"],"xaxis":"x","y":[466],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Art<br>Category=%{x}<br>Number of Items
Sold=%{y}<extra></extra>","legendgroup":"Art","marker":
{"color":"#19d3f3","pattern":
{"shape":""}},"name":"Art","offsetgroup":"Art","orientation":"v","show
legend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[795],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Binders<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Binders","marker":
{"color":"#FF6692","pattern":
{"shape":""}},"name":"Binders","offsetgroup":"Binders","orientation":"
v","showlegend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[1522],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Envelopes<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Envelopes","marker":
{"color":"#B6E880","pattern":
{"shape":""}},"name":"Envelopes","offsetgroup":"Envelopes","orientatio
n":"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Office Supplies"],"xaxis":"x","y":[254],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Fasteners<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Fasteners","marker":
```

```
{"color":"#FF97FF","pattern":
{"shape":""}},"name":"Fasteners","offsetgroup":"Fasteners","orientatio
n":"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Office Supplies"],"xaxis":"x","y":[217],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Labels<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Labels","marker":
{"color":"#FECB52","pattern":
{"shape":""}},"name":"Labels","offsetgroup":"Labels","orientation":"v"
,"showlegend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[363],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Paper<br>Category=%{x}<br>Number of Items
Sold=%{y}<extra></extra>","legendgroup":"Paper","marker":
{"color":"#636efa","pattern":
{"shape":""}},"name":"Paper","offsetgroup":"Paper","orientation":"v","
showlegend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[1359],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Storage<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Storage","marker":
{"color":"#EF553B","pattern":
{"shape":""}},"name":"Storage","offsetgroup":"Storage","orientation":"
v","showlegend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[846],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Supplies<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Supplies","marker":
{"color":"#00cc96","pattern":
{"shape":""}},"name":"Supplies","offsetgroup":"Supplies","orientation"
:"v","showlegend":true,"textposition":"auto","type":"bar","x":["Office
Supplies"],"xaxis":"x","y":[190],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Accessories<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Accessories","marker":
{"color":"#ab63fa","pattern":
{"shape":""}},"name":"Accessories","offsetgroup":"Accessories","orient
ation":"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Technology"],"xaxis":"x","y":[775],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Copiers<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Copiers","marker":
{"color":"#FFA15A","pattern":
{"shape":""}},"name":"Copiers","offsetgroup":"Copiers","orientation":"
v","showlegend":true,"textposition":"auto","type":"bar","x":
["Technology"],"xaxis":"x","y":[68],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Machines<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Machines","marker":
```

```
{"color":"#19d3f3","pattern":
{"shape":""}},"name":"Machines","offsetgroup":"Machines","orientation"
:"v","showlegend":true,"textposition":"auto","type":"bar","x":
["Technology"],"xaxis":"x","y":[115],"yaxis":"y"},
{"alignmentgroup":"True","hovertemplate":"Sub-
Category=Phones<br>Category=%{x}<br>Number of Items Sold=%
{y}<extra></extra>","legendgroup":"Phones","marker":
{"color":"#FF6692","pattern":
{"shape":""}},"name":"Phones","offsetgroup":"Phones","orientation":"v"
,"showlegend":true,"textposition":"auto","type":"bar","x":
["Technology"],"xaxis":"x","y":[889],"yaxis":"y"}],"layout":
{"barmode":"group","legend":{"title":{"text":"Sub-
Category"},"tracegroupgap":0},"template":{"data":{"bar":[{"error_x":
{"color":"#2a3f5f"},"error_y":{"color":"#2a3f5f"},"marker":{"line":
{"color":"#E5ECF6","width":0.5},"pattern":
{"fillmode":"overlay","size":10,"solidity":0.2}},"type":"bar"}],"barpo
lar":[{"marker":{"line":{"color":"#E5ECF6","width":0.5},"pattern":
{"fillmode":"overlay","size":10,"solidity":0.2}},"type":"barpolar"}],"
carpet":[{"aaxis":
{"endlinecolor":"#2a3f5f","gridcolor":"white","linecolor":"white","min
orgridcolor":"white","startlinecolor":"#2a3f5f"},"baxis":
{"endlinecolor":"#2a3f5f","gridcolor":"white","linecolor":"white","min
orgridcolor":"white","startlinecolor":"#2a3f5f"},"type":"carpet"}],"ch
oropleth":[{"colorbar":
{"outlinewidth":0,"ticks":""},"type":"choropleth"}],"contour":
[{"colorbar":{"outlinewidth":0,"ticks":""},"colorscale":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"contour"}],"contourcarpet":[{"colorbar":
{"outlinewidth":0,"ticks":""},"type":"contourcarpet"}],"heatmap":
[{"colorbar":{"outlinewidth":0,"ticks":""},"colorscale":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"heatmap"}],"heatmapgl":[{"colorbar":
{"outlinewidth":0,"ticks":""},"colorscale":[[0,"#0d0887"],
[0.1111111111111111,"#46039f"],[0.2222222222222222,"#7201a8"],
[0.3333333333333333,"#9c179e"],[0.4444444444444444,"#bd3786"],
[0.5555555555555556,"#d8576b"],[0.6666666666666666,"#ed7953"],
[0.7777777777777778,"#fb9f3a"],[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"heatmapgl"}],"histogram":[{"marker":{"pattern":
{"fillmode":"overlay","size":10,"solidity":0.2}},"type":"histogram"}],
"histogram2d":[{"colorbar":{"outlinewidth":0,"ticks":""},"colorscale":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"histogram2d"}],"histogram2dcontour":
[{"colorbar":{"outlinewidth":0,"ticks":""},"colorscale":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"histogram2dcontour"}],"mesh3d":[{"colorbar":
{"outlinewidth":0,"ticks":""},"type":"mesh3d"}],"parcoords":[{"line":
{"colorbar":{"outlinewidth":0,"ticks":""}},"type":"parcoords"}],"pie":
[{"automargin":true,"type":"pie"}],"scatter":[{"fillpattern":
{"fillmode":"overlay","size":10,"solidity":0.2},"type":"scatter"}],"sc
atter3d":[{"line":{"colorbar":{"outlinewidth":0,"ticks":""}},"marker":
{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scatter3d"}],"scattercarpet":
[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scattercarpet"}],"scattergeo":
[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scattergeo"}],"scattergl":
[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scattergl"}],"scattermapbox":
[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scattermapbox"}],"scatterpolar"
:[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scatterpolar"}],"scatterpolargl
":[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scatterpolargl"}],"scatterterna
ry":[{"marker":{"colorbar":
{"outlinewidth":0,"ticks":""}},"type":"scatterternary"}],"surface":
[{"colorbar":{"outlinewidth":0,"ticks":""},"colorscale":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],
[1,"#f0f921"]],"type":"surface"}],"table":[{"cells":{"fill":
{"color":"#EBF0F8"},"line":{"color":"white"}},"header":{"fill":
{"color":"#C8D4E3"},"line":
{"color":"white"}},"type":"table"}]},"layout":{"annotationdefaults":
{"arrowcolor":"#2a3f5f","arrowhead":0,"arrowwidth":1},"autotypenumbers
":"strict","coloraxis":{"colorbar":
{"outlinewidth":0,"ticks":""}},"colorscale":{"diverging":
[[0,"#8e0152"],[0.1,"#c51b7d"],[0.2,"#de77ae"],[0.3,"#f1b6da"],
[0.4,"#fde0ef"],[0.5,"#f7f7f7"],[0.6,"#e6f5d0"],[0.7,"#b8e186"],
[0.8,"#7fbc41"],[0.9,"#4d9221"],[1,"#276419"]],"sequential":
```

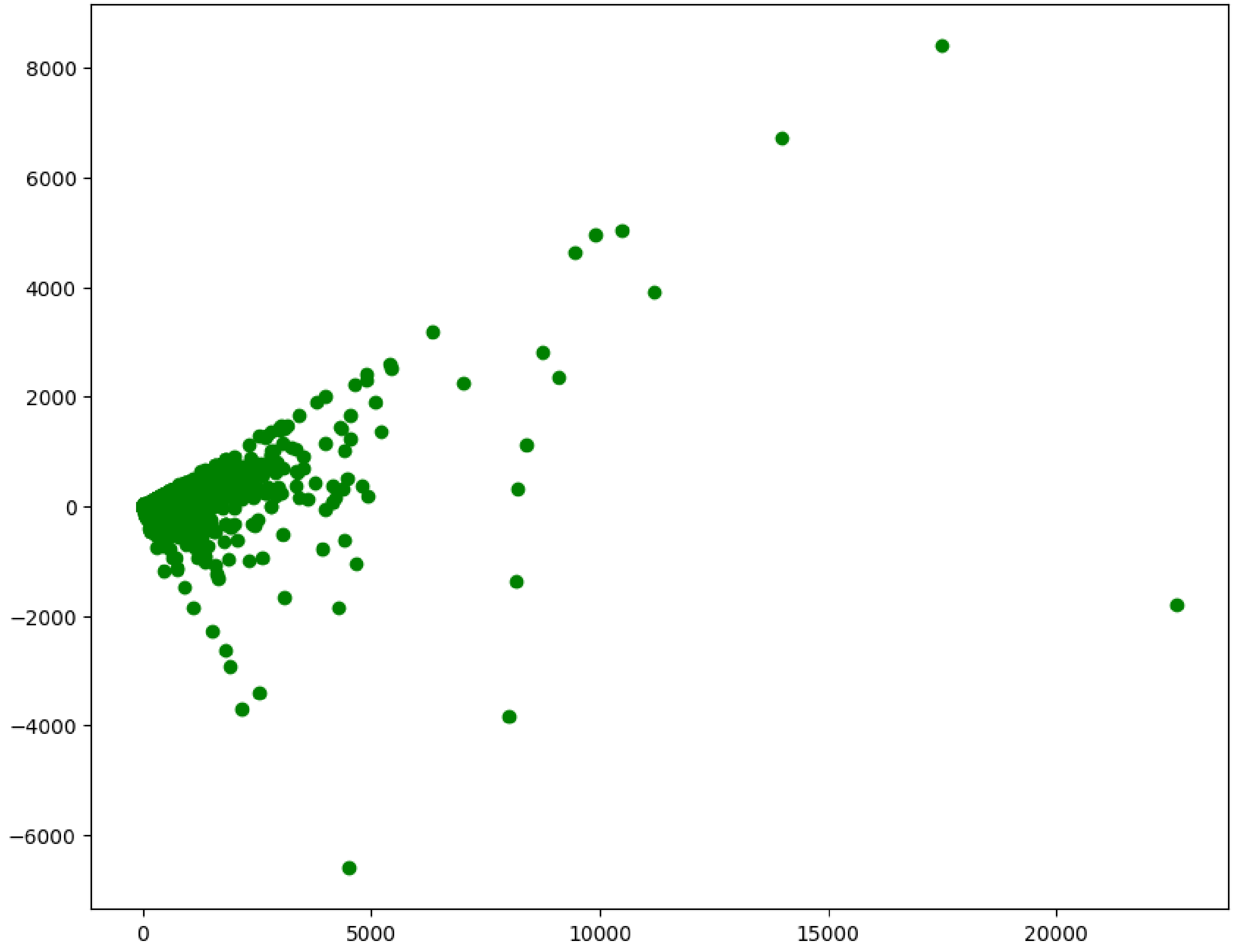
```
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],[1,"#f0f921"]],"sequentialminus":
[[0,"#0d0887"],[0.1111111111111111,"#46039f"],
[0.2222222222222222,"#7201a8"],[0.3333333333333333,"#9c179e"],
[0.4444444444444444,"#bd3786"],[0.5555555555555556,"#d8576b"],
[0.6666666666666666,"#ed7953"],[0.7777777777777778,"#fb9f3a"],
[0.8888888888888888,"#fdca26"],[1,"#f0f921"]]},"colorway":
["#636efa","#EF553B","#00cc96","#ab63fa","#FFA15A","#19d3f3","#FF6692"
,"#B6E880","#FF97FF","#FECB52"],"font":{"color":"#2a3f5f"},"geo":
{"bgcolor":"white","lakecolor":"white","landcolor":"#E5ECF6","showlake
s":true,"showland":true,"subunitcolor":"white"},"hoverlabel":
{"align":"left"},"hovermode":"closest","mapbox":
{"style":"light"},"paper_bgcolor":"white","plot_bgcolor":"#E5ECF6","po
lar":{"angularaxis":
{"gridcolor":"white","linecolor":"white","ticks":""},"bgcolor":"#E5ECF
6","radialaxis":
{"gridcolor":"white","linecolor":"white","ticks":""}},"scene":
{"xaxis":
{"backgroundcolor":"#E5ECF6","gridcolor":"white","gridwidth":2,"lineco
lor":"white","showbackground":true,"ticks":"","zerolinecolor":"white"}
,"yaxis":
{"backgroundcolor":"#E5ECF6","gridcolor":"white","gridwidth":2,"lineco
lor":"white","showbackground":true,"ticks":"","zerolinecolor":"white"}
,"zaxis":
{"backgroundcolor":"#E5ECF6","gridcolor":"white","gridwidth":2,"lineco
lor":"white","showbackground":true,"ticks":"","zerolinecolor":"white"}
},"shapedefaults":{"line":{"color":"#2a3f5f"}},"ternary":{"aaxis":
{"gridcolor":"white","linecolor":"white","ticks":""},"baxis":
{"gridcolor":"white","linecolor":"white","ticks":""},"bgcolor":"#E5ECF
6","caxis":
{"gridcolor":"white","linecolor":"white","ticks":""}},"title":
{"x":5.0e-2},"xaxis":
{"automargin":true,"gridcolor":"white","linecolor":"white","ticks":"",
"title":
{"standoff":15},"zerolinecolor":"white","zerolinewidth":2},"yaxis":
{"automargin":true,"gridcolor":"white","linecolor":"white","ticks":"",
"title":
{"standoff":15},"zerolinecolor":"white","zerolinewidth":2}}},"title":
{"text":"Counts of Sub-Categories within Main Categories"},"xaxis":
{"anchor":"y","domain":[0,1],"title":{"text":"Category"}},"yaxis":
{"anchor":"x","domain":[0,1],"title":{"text":"Number of Items
Sold"}}}}
```

Observation :

Here, we can see that throughout the sub-categories the main category of Office Supplies having highest no. of sales distribution

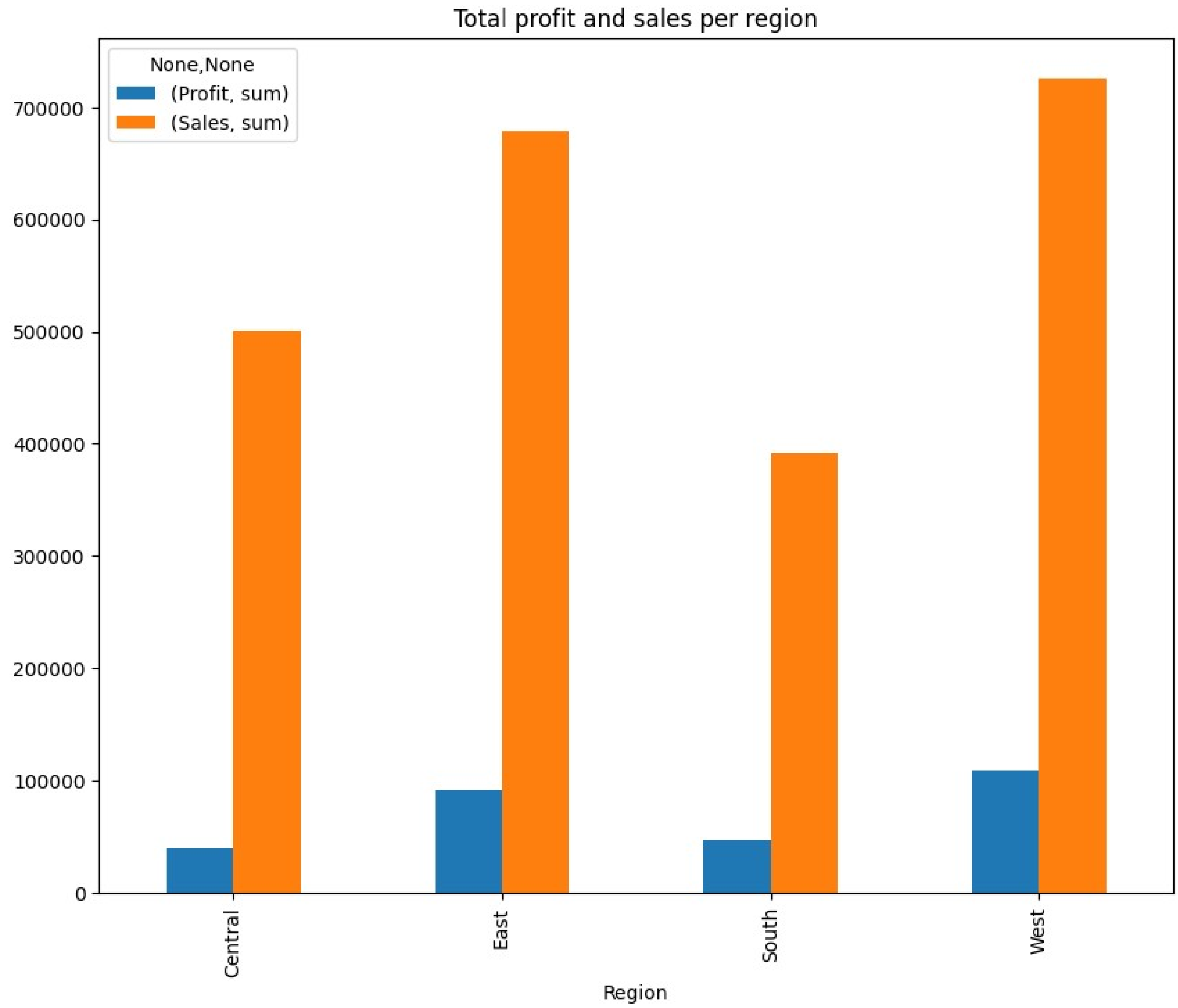
```
plt.scatter(df['Sales'],df['Profit'],color='green')
```

```
<matplotlib.collections.PathCollection at 0x2260e896350>
```



Profit of sales based on region

```
df.groupby('Region')['Profit','Sales'].agg(['sum']).plot.bar()
plt.title('Total profit and sales per region')
plt.rcParams['figure.figsize']=[10,8]
plt.show()
```

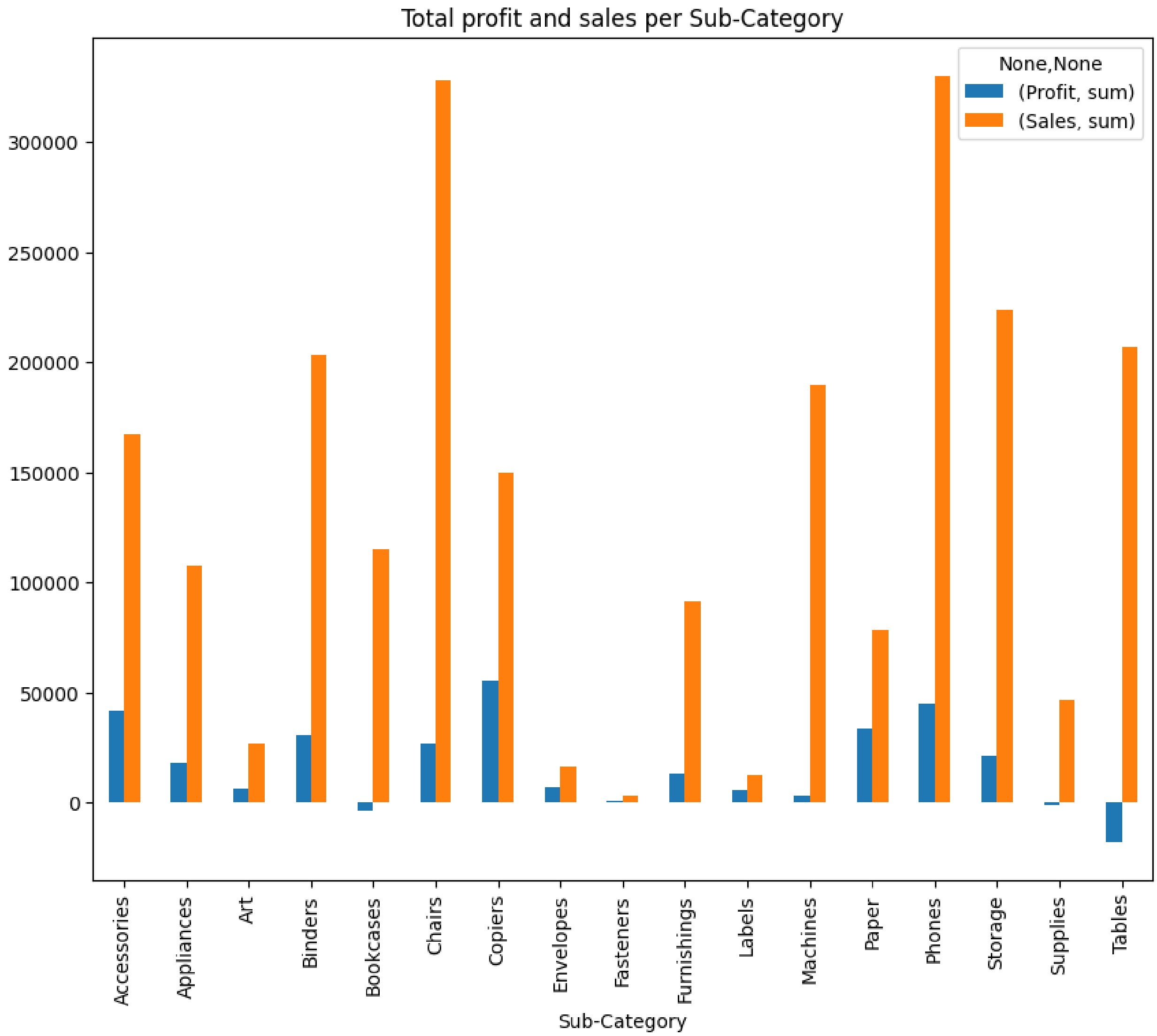


Observation :

The highest profit earn in East and west region and also sales are high no. of sales are belongs to the same region.

Profit of sales based on Sub-Category

```
df.groupby('Sub-Category')['Profit','Sales'].agg(['sum']).plot.bar()
plt.title('Total profit and sales per Sub-Category')
plt.rcParams['figure.figsize']=[10,8]
plt.show()
```

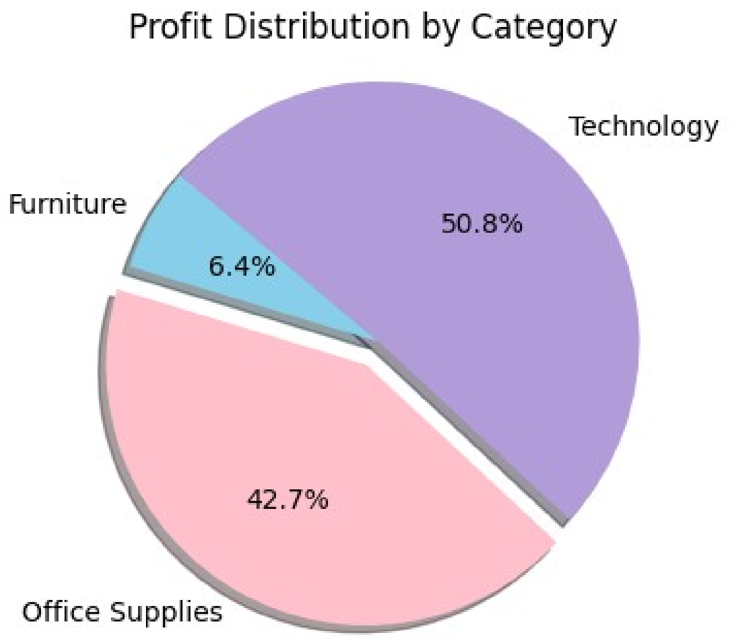


Observation :

The Highest profit is earned in copiers while, the selling of phones and chairs are extremely high compared to other products.

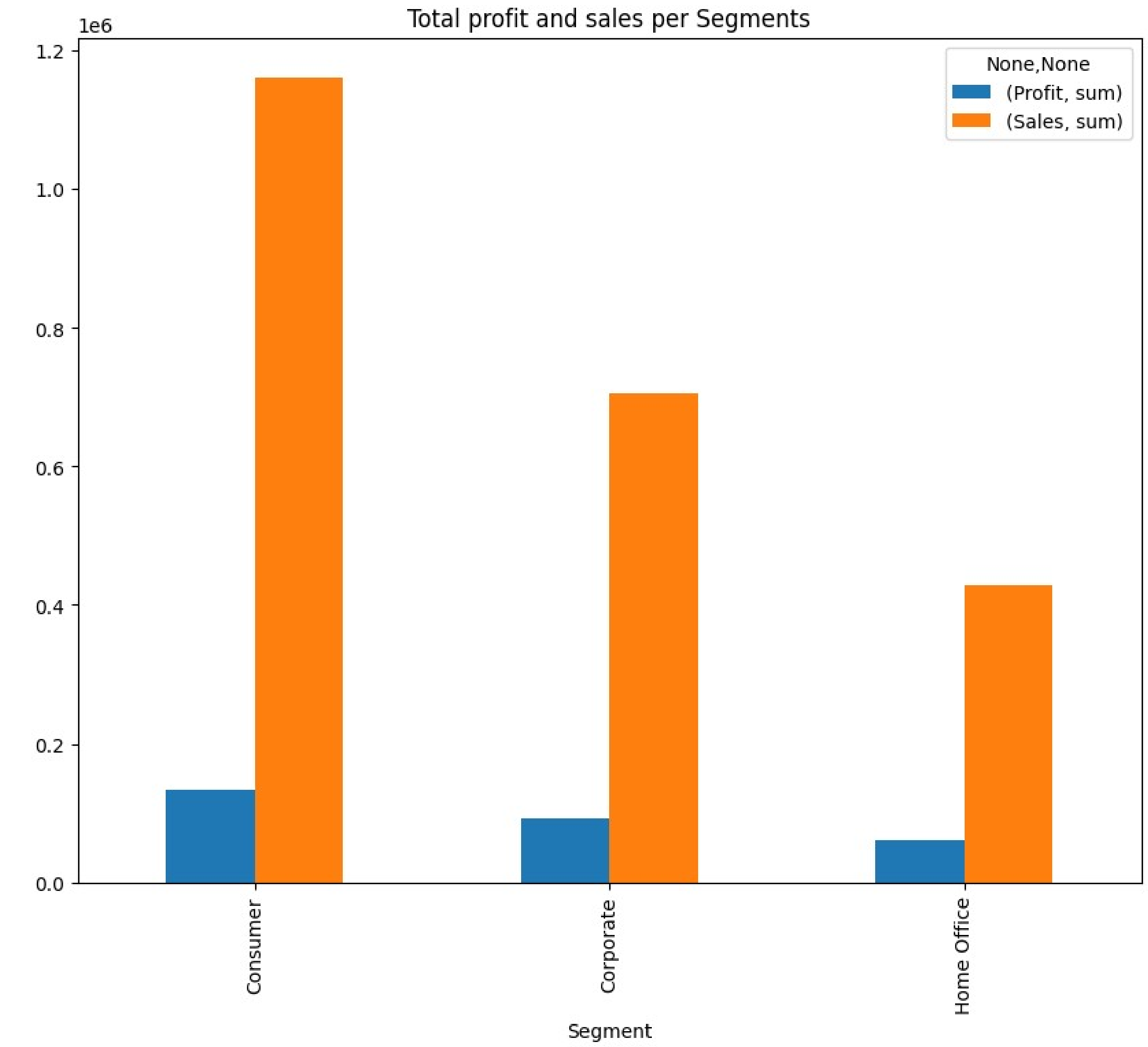
Another interesting fact-peoples don't prefer to buy tables and Bookcases from superstore as sales is medium but they are facing loss

```
cat_s=df.groupby("Category")["Profit"].sum().reset_index()
profit=cat_s["Profit"].tolist()
category=cat_s["Category"].tolist()


plt.figure(figsize=(8,4))
plt.pie(profit, labels=category,
        autopct='%1.1f%%',
        colors=["#87CEEB", "#FFC0CB", "#B19CD9"],
        shadow=True,
        explode = [0, 0.1, 0],
        startangle=140)
plt.title('Profit Distribution by Category')
plt.axis('equal')
plt.show()
```



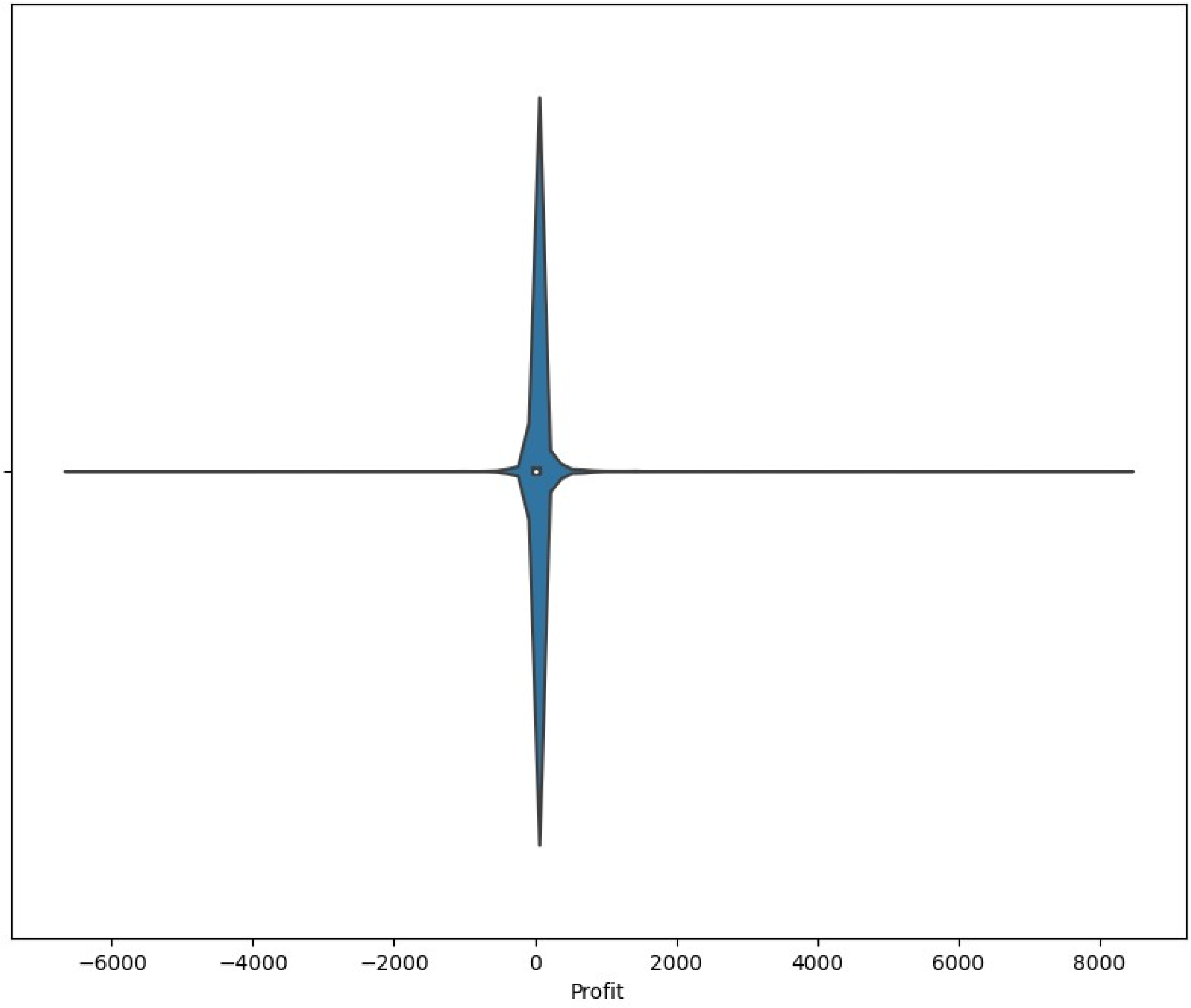
Profit of sales based on Segments

```
df.groupby('Segment')['Profit','Sales'].agg(['sum']).plot.bar()
plt.title('Total profit and sales per Segments')
plt.rcParams['figure.figsize']=[10,8]
plt.show()
```



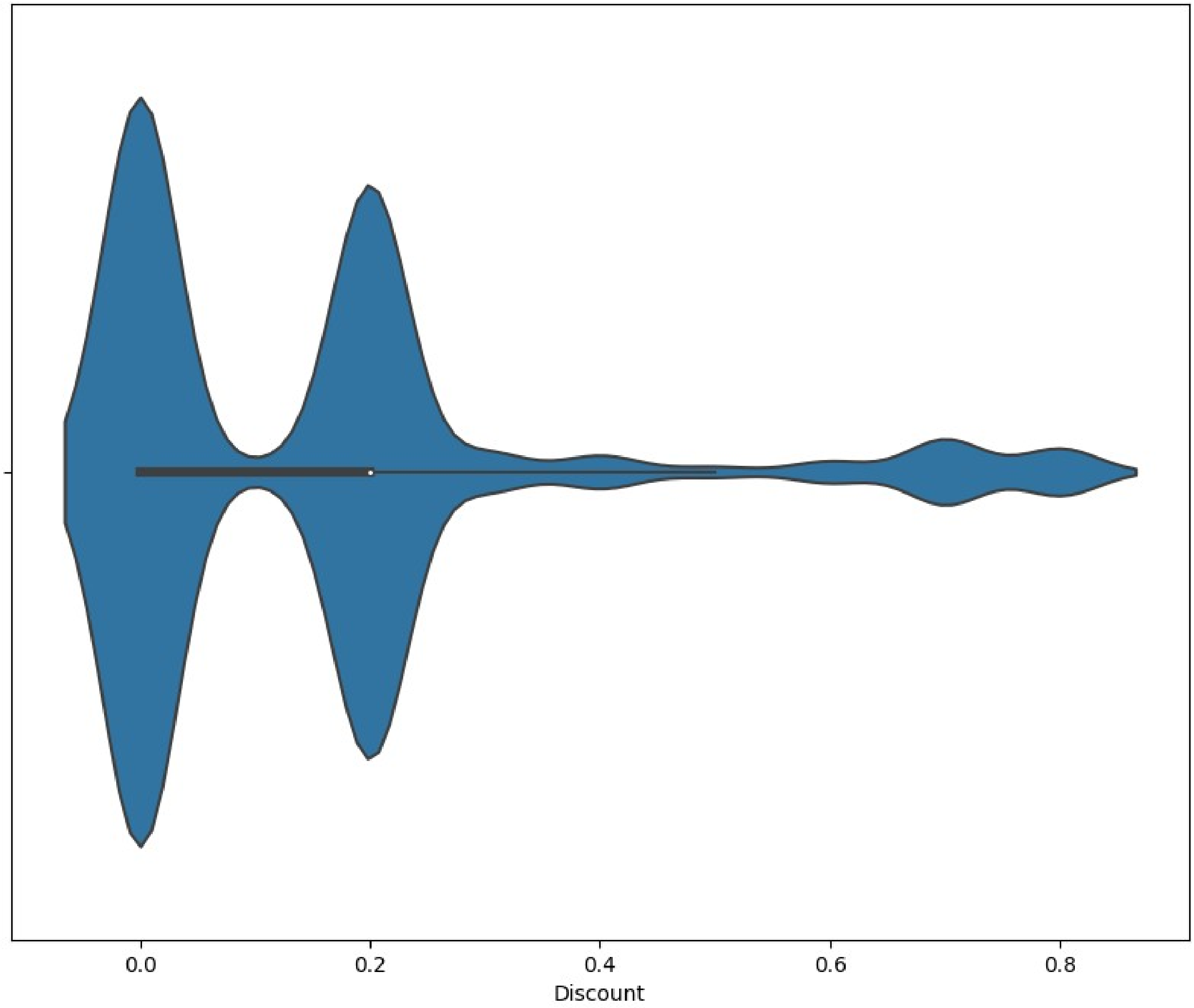
```
sns.violinplot(x='Profit',data=df)
```

```
<Axes: xlabel='Profit'>
```



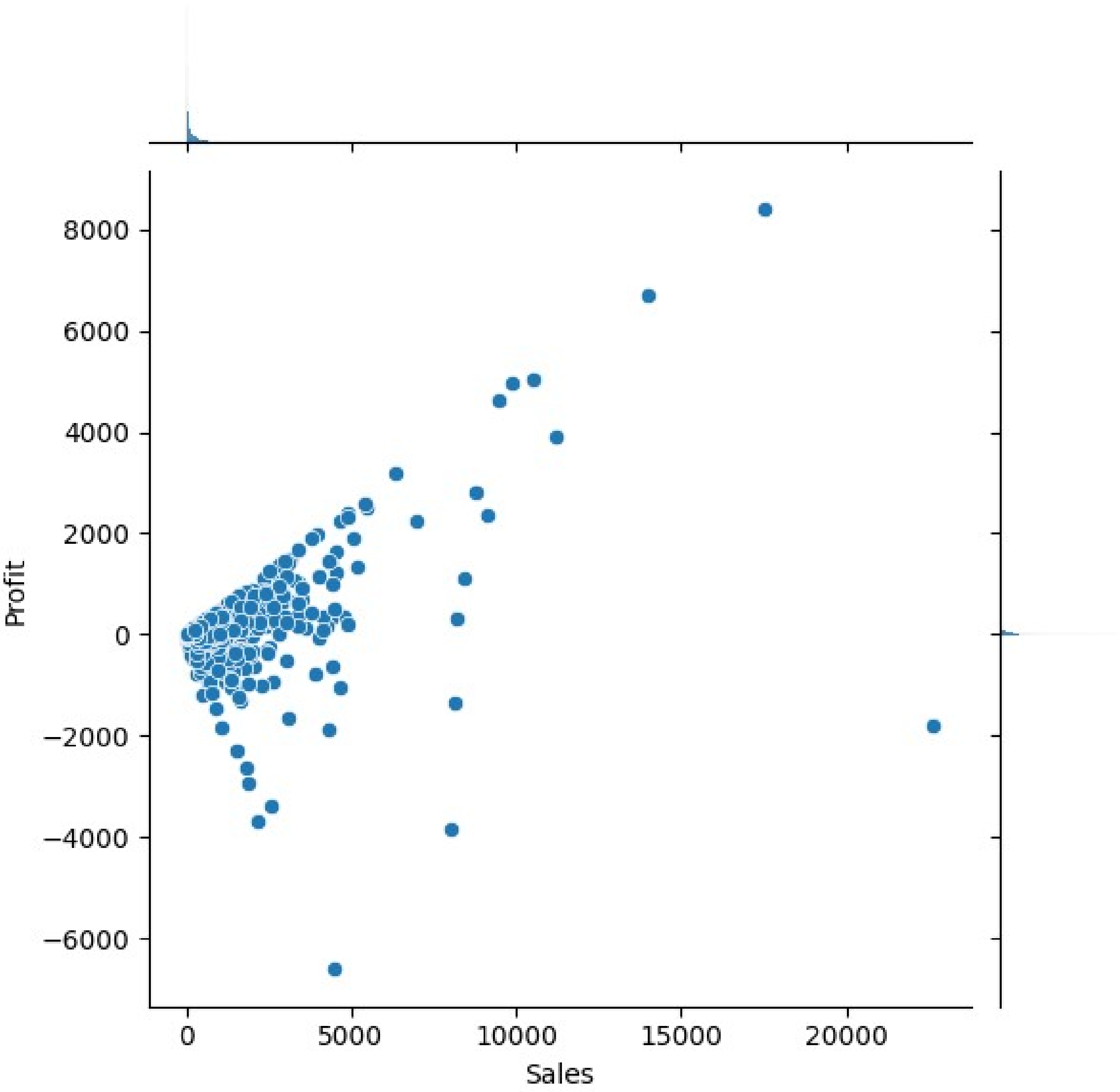
```
sns.violinplot(x='Discount',data=df)
```

```
<Axes: xlabel='Discount'>
```



```
sns.jointplot(data=df,x='Sales',y='Profit')
```

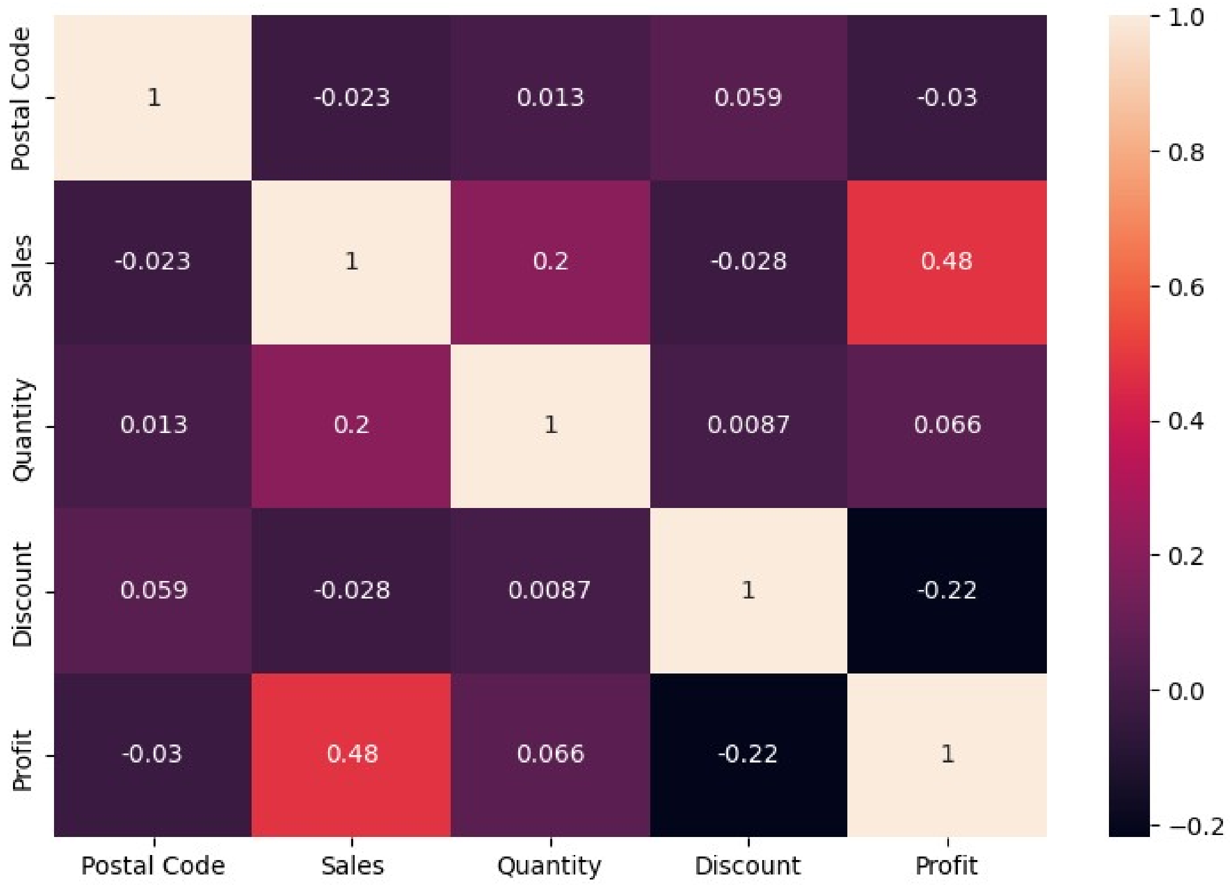
```
<seaborn.axisgrid.JointGrid at 0x22611b43ed0>
```



```
df_corr=df.corr()
df_corr
```

```
             Postal Code     Sales  Quantity  Discount    Profit
Postal Code     1.000000 -0.023476  0.013110  0.059225 -0.029892
Sales          -0.023476  1.000000  0.200722 -0.028311  0.479067
Quantity        0.013110  0.200722  1.000000  0.008678  0.066211
Discount        0.059225 -0.028311  0.008678  1.000000 -0.219662
Profit         -0.029892  0.479067  0.066211 -0.219662  1.000000
```

```
fig,axes=plt.subplots(1,1,figsize=(9,6))
sns.heatmap(df_corr,annot=True)
plt.show()
```



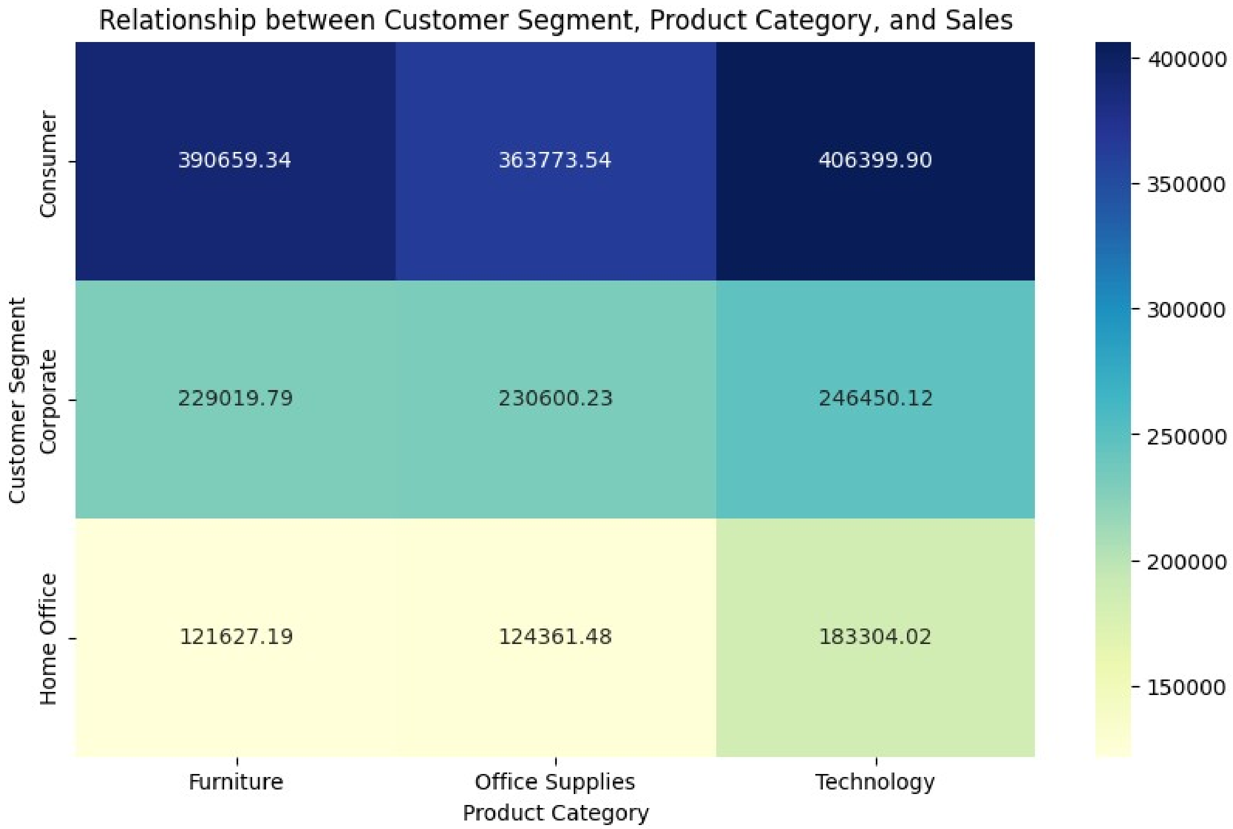
Observation :

From above heatmap we can observe that there is negative corerelation between discount and profit

Relation between the customer segment,product category with the sales

```
grouped_data = df.groupby(['Segment', 'Category'])
['Sales'].sum().reset_index()
grouped_data

       Segment         Category        Sales
0     Consumer        Furniture  390659.3420
1     Consumer  Office Supplies  363773.5360
2     Consumer       Technology  406399.8970
3    Corporate        Furniture  229019.7858
4    Corporate  Office Supplies  230600.2260
5    Corporate       Technology  246450.1190
6  Home Office        Furniture  121627.1855
```

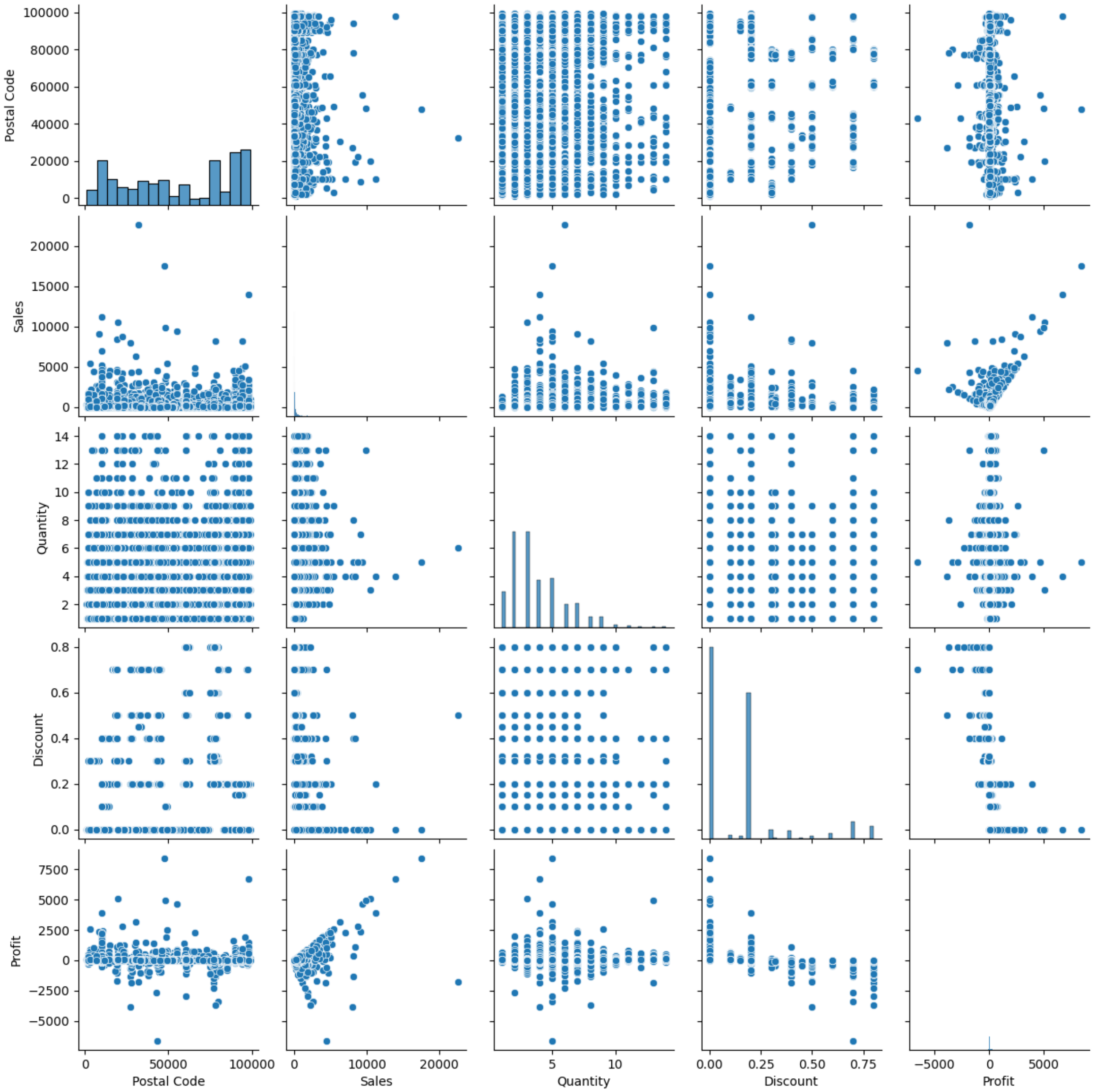
```
7  Home Office  Office Supplies  124361.4820
8  Home Office       Technology  183304.0170

pivot_df = grouped_data.pivot(index='Segment', columns='Category',
values='Sales')
pivot_df

Category       Furniture  Office Supplies  Technology
Segment
Consumer     390659.3420       363773.536  406399.897
Corporate    229019.7858       230600.226  246450.119
Home Office  121627.1855       124361.482  183304.017

plt.figure(figsize=(10, 6))
sns.heatmap(pivot_df, annot=True, fmt=".2f", cmap="YlGnBu")
plt.title('Relationship between Customer Segment, Product Category,
and Sales')
plt.xlabel('Product Category')
plt.ylabel('Customer Segment')
plt.show()
```



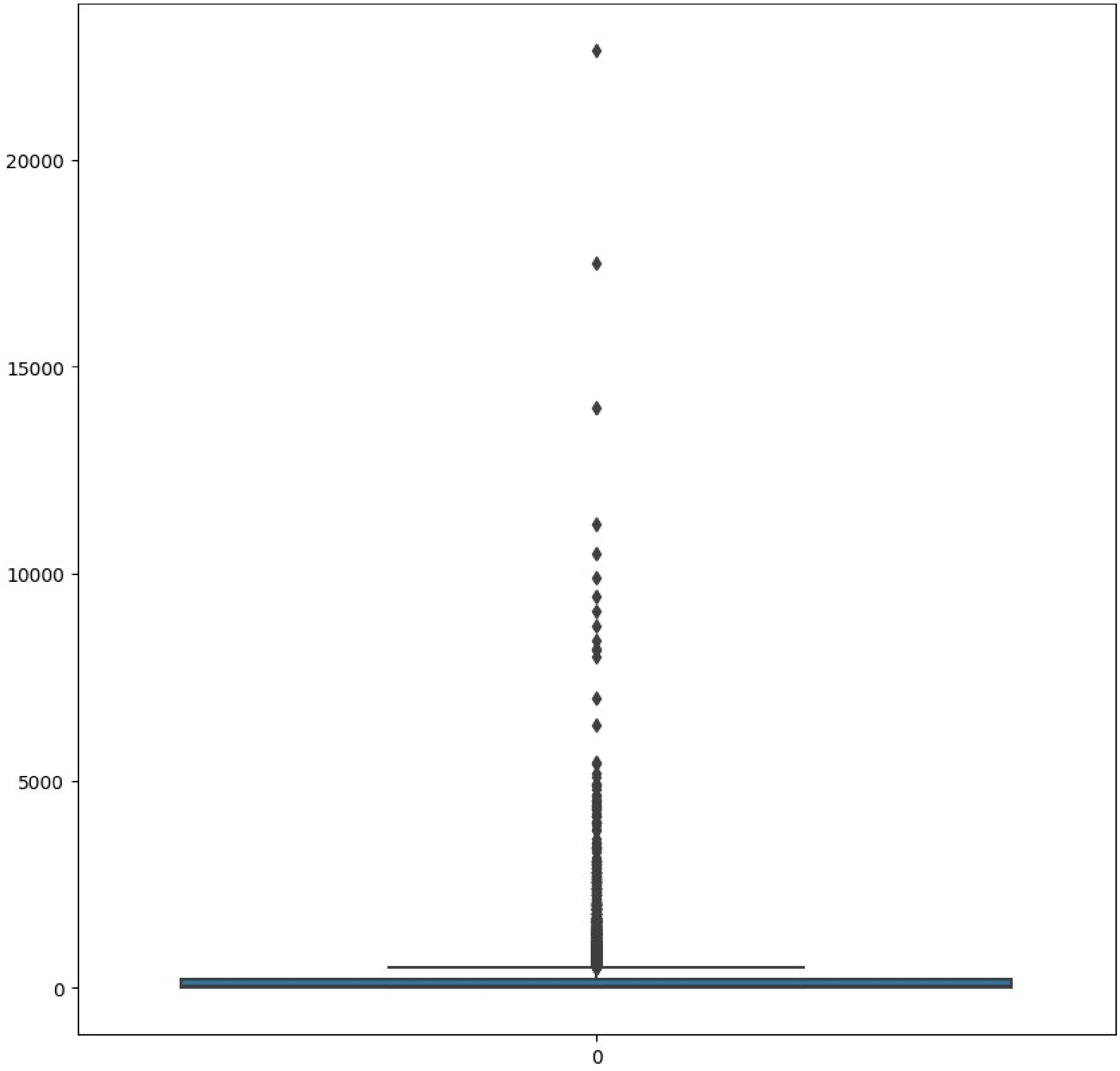
Observation :

consumers who byes Technology have the highest sales

```
sns.pairplot(df)

<seaborn.axisgrid.PairGrid at 0x22612912790>
```

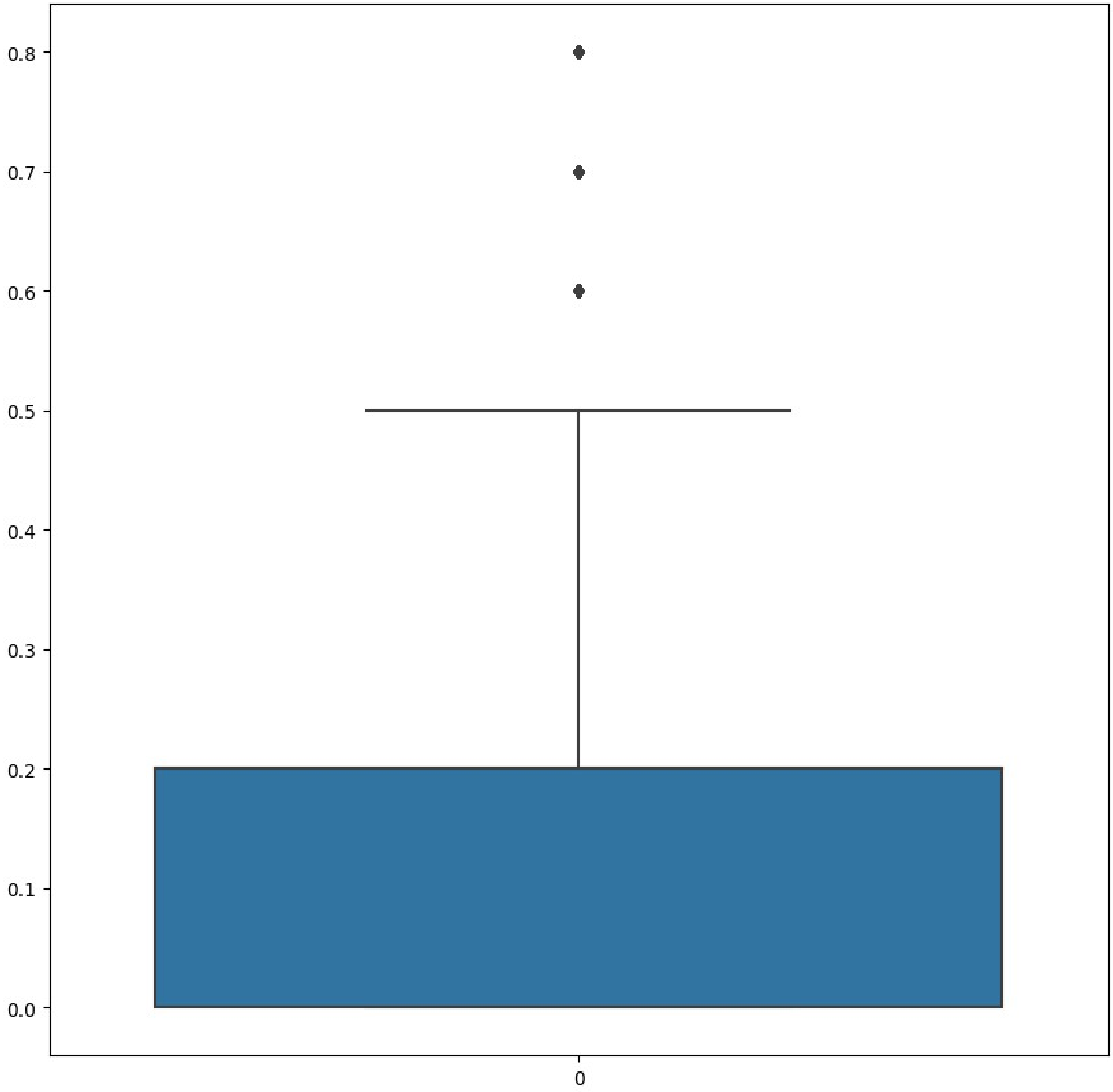


```
fig,axes=plt.subplots(figsize=(10,10))
sns.boxplot(df['Sales'])

<Axes: >
```



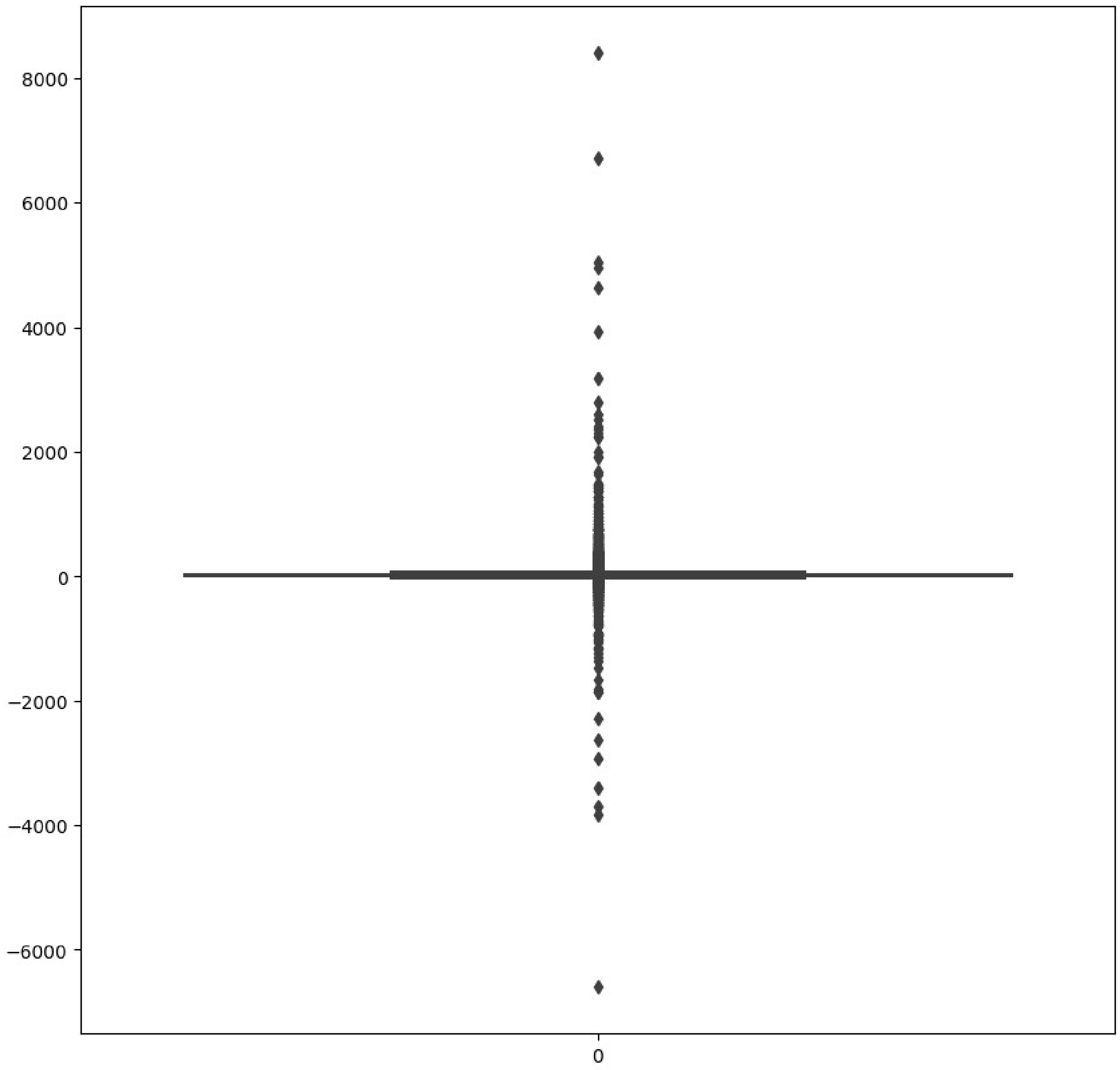
```
fig,axes=plt.subplots(figsize=(10,10))
sns.boxplot(df['Discount'])
```

```
<Axes: >
```



```
fig,axes=plt.subplots(figsize=(10,10))
sns.boxplot(df['Profit'])
```

```
<Axes: >
```

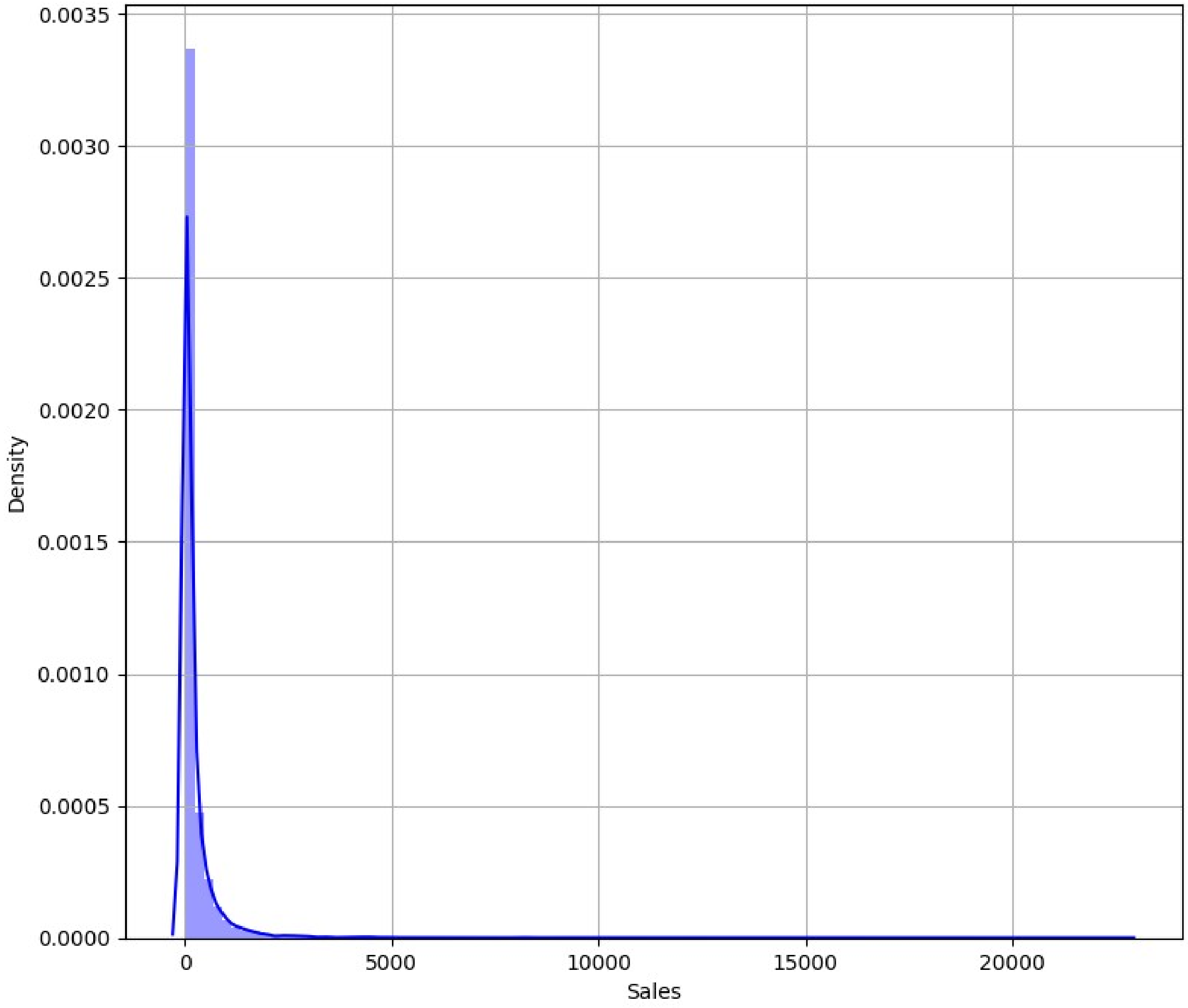


Sales Statistical data

```
print(df['Sales'].describe())
plt.figure(figsize=(9,8))
plt.grid()
sns.distplot(df['Sales'],color='b',bins=100,hist_kws={'alpha':0.4})
```

```
count     9977.000000
mean       230.148902
std        623.721409
min          0.444000
25%         17.300000
50%         54.816000
```

```
75%           209.970000
max         22638.480000
Name: Sales, dtype: float64

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

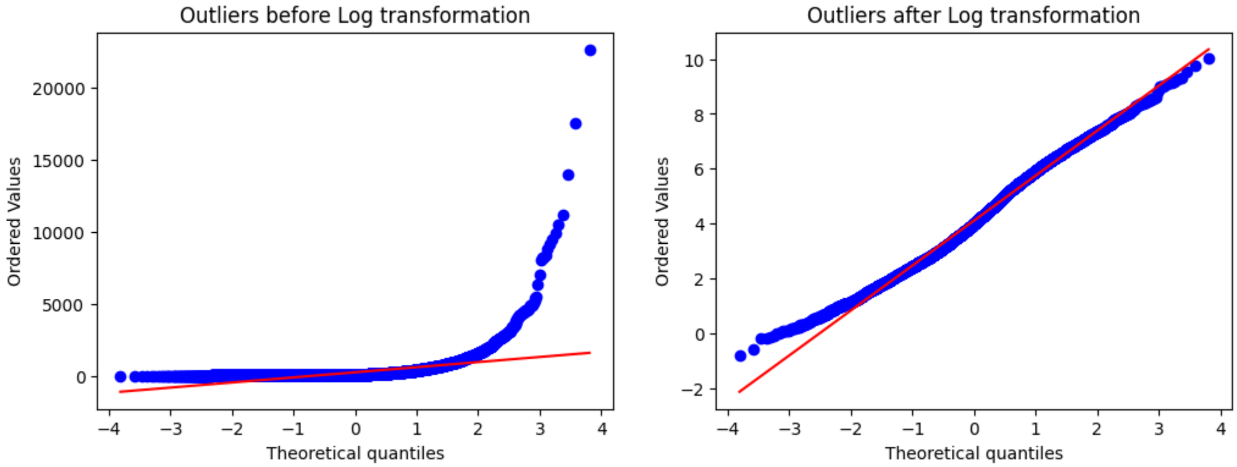


Handling Outliers

As we already see in the Data Visualization part, the Sales column having some outliers so it is important to handle this

```
df['Sales_log'] = np.log(df['Sales'])

fig = plt.figure(figsize=(12,4))
ax1 = fig.add_subplot(121)
stats.probplot(df['Sales'], dist="norm", plot=ax1)
```

```
ax1.set_title('Outliers before Log transformation')
ax2 = fig.add_subplot(122)
stats.probplot(df['Sales_log'],dist="norm", plot=ax2)
ax2.set_title('Outliers after Log transformation')
plt.show()
```



Conclusion

1. We can say that more profitable region is West and East whereas New york and California having highest profitable states.
2. and in terms of the Product Category Technology is highest but Furniture and Office Supplier are also good there are so many demand in all these product category.
3. Also the profit in South and Central is less,The Highest profit is earned in copiers while the selling of phones and chairs are extremely high compared to other products.
4. No or very less profit in sales of supplies.
5. Profit is more in sale of copiers.
6. Total sum of profit in sale of tables is negative.
7. Negative correlation between profit and Discount.

Thank you