TASK - 3: Exploratory Data Analysis on dataset 'SampleSuperstore

Our Objective:

1. Perform 'Exploratory Data Analysis' on dataset 'SampleSuperstore'
2. As a business manager, try to find out the weak areas where you can work to make more profit.
3. What all business problems you can derive by exploring the data?

In [1]:

```
### Import all necessary Libraries
import numpy as np
import pandas as pd
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
```

In [2]:

```
sample = pd.read_csv("SampleSuperstore.csv")
sample.head()
```

Out[2]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	26
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	73
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	1
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	95
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	2

In [3]:

```
sample.shape
```

Out[3]:

```
(9994, 13)
```

In [4]:

```
sample.info
```

Out[4]:

```
<bound method DataFrame.info of           Ship Mode    Segment        Count
ry             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

[9994 rows x 13 columns]>
```

In [5]:

```
sample.describe()
```

Out[5]:

	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

In [6]:

```
#Checking Missing Values
sample.isnull().sum()
```

Out[6]:

```
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
```

In [7]:

```
# Chaecking the dupilication in data
sample.duplicated().sum()
```

Out[7]:

```
17
```

In [8]:

```
sample.drop_duplicates()
```

Out[8]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Su Categ
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcas
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Cha
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Lab
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tab
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Stora
...	...	...	...	...	...	...	...	...	
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishin
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishin
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phon
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Pap
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Applianc

9977 rows × 13 columns

In [9]:

```
sample.nunique()
```

Out[9]:

```
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
```

In [10]:

```
#Deleting the Variable.
col=['Postal Code']
sample1=sample.drop(columns=col,axis=1)
```

In [11]:

```
#Correlation Between Variables.
sample1.corr()
```

Out[11]:

	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

In [12]:

```
# Covariance of columns
sample1.cov()
```

Out[12]:

	Sales	Quantity	Discount	Profit
Sales	388434.455308	278.459923	-3.627228	69944.096586
Quantity	278.459923	4.951113	0.003961	34.534769
Discount	-3.627228	0.003961	0.042622	-10.615173
Profit	69944.096586	34.534769	-10.615173	54877.798055

In [13]:

```
sample1.head()
```

Out[13]:

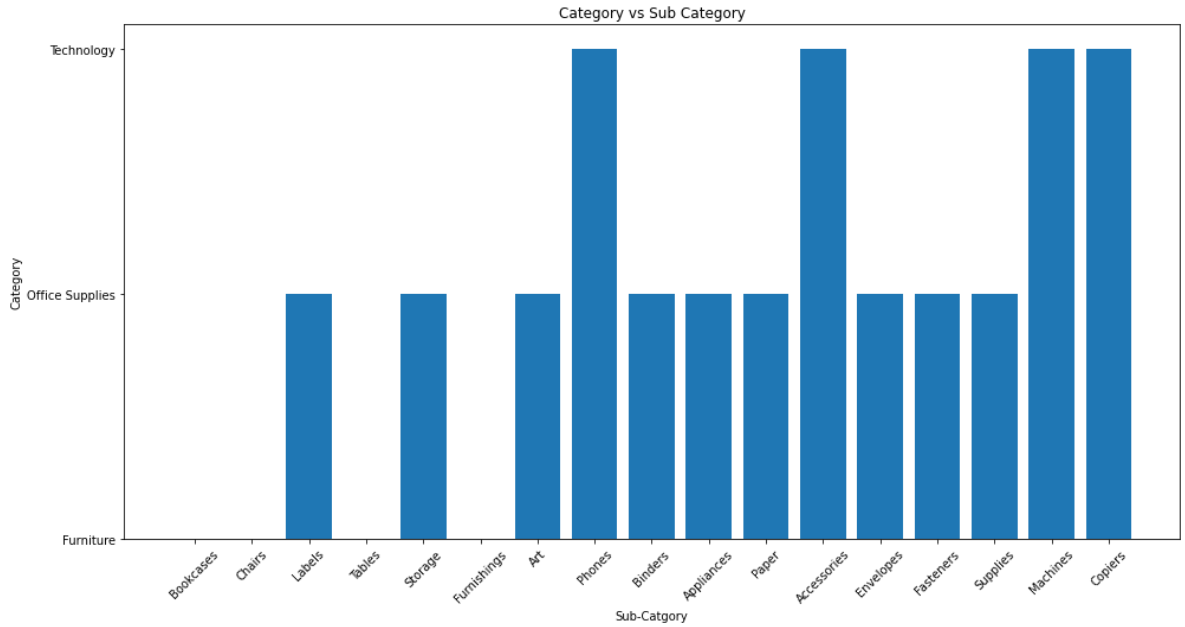
	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400
2	Second Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680

Exploratory Data Analysis

Data Visualization

In [14]:

```
plt.figure(figsize=(16,8))
plt.bar('Sub-Category','Category', data=sample1)
plt.title('Category vs Sub Category')
plt.xlabel('Sub-Catgory')
plt.ylabel('Category')
plt.xticks(rotation=45)
plt.show()
```



In [15]:

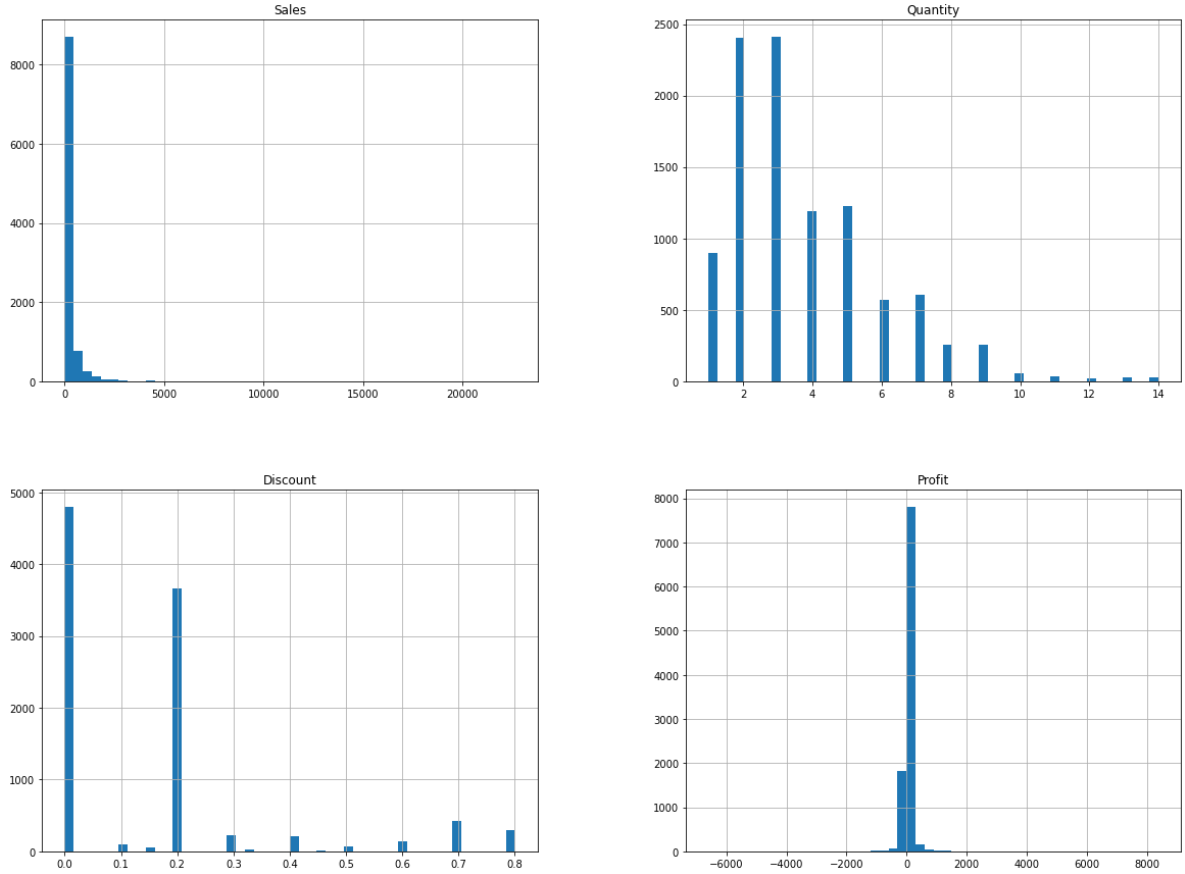
```
sample1.corr()
```

Out[15]:

	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

In [16]:

```
sample1.hist(bins=50 ,figsize=(20,15))
plt.show();
```



In [17]:

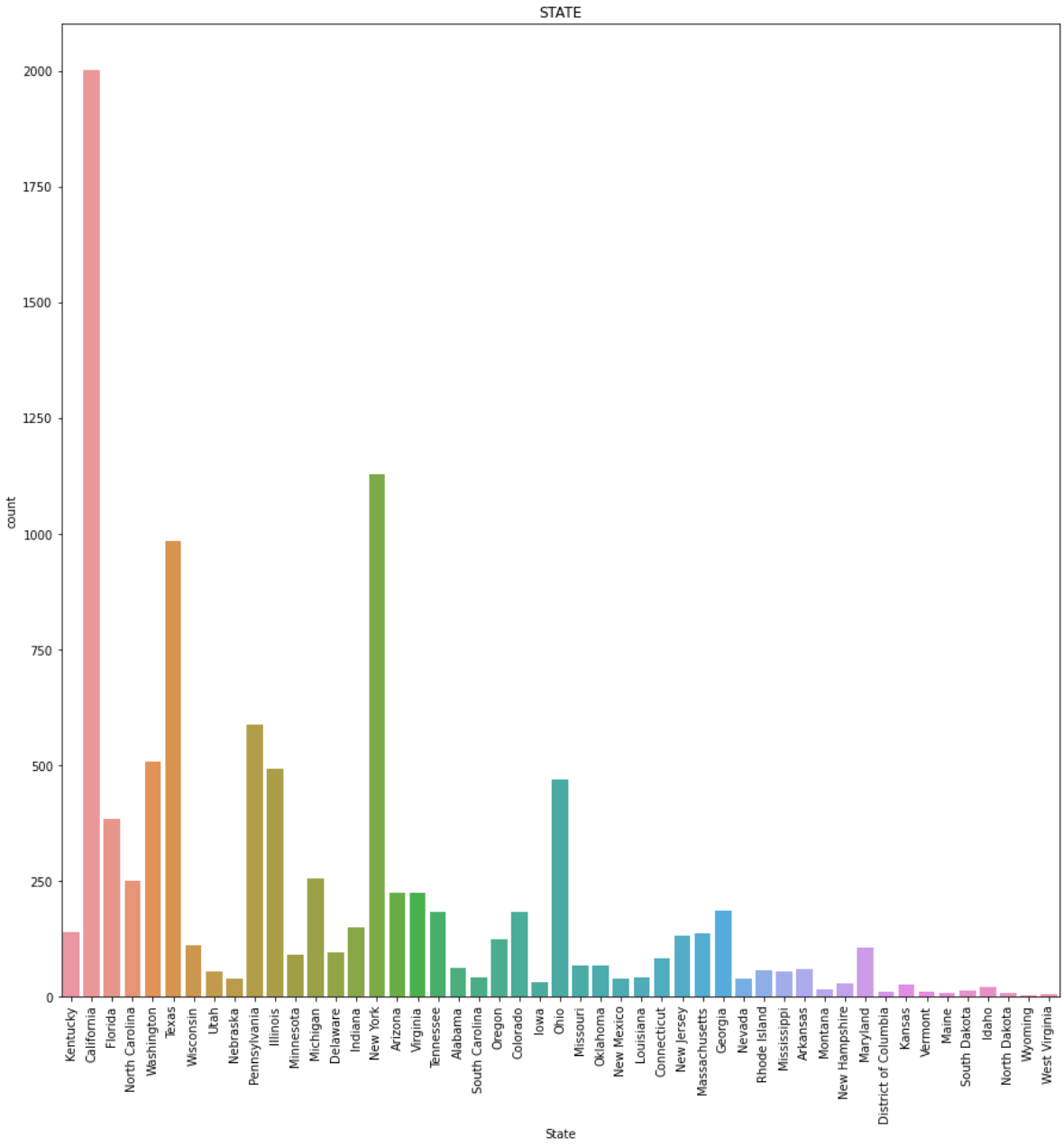
```
# Count the total repeatable states
sample1['State'].value_counts()
```

Out[17]:

```
California              2001
New York                1128
Texas                    985
Pennsylvania             587
Washington               506
Illinois                 492
Ohio                     469
Florida                  383
Michigan                 255
North Carolina           249
Arizona                  224
Virginia                 224
Georgia                  184
Tennessee                183
Colorado                 182
Indiana                  149
Kentucky                 139
Massachusetts            135
New Jersey               130
Oregon                   124
Wisconsin                110
Maryland                 105
Delaware                  96
Minnesota                 89
Connecticut               82
Oklahoma                  66
Missouri                  66
Alabama                   61
Arkansas                  60
Rhode Island              56
Utah                      53
Mississippi               53
Louisiana                 42
South Carolina            42
Nevada                    39
Nebraska                  38
New Mexico                37
Iowa                      30
New Hampshire             27
Kansas                    24
Idaho                     21
Montana                   15
South Dakota              12
Vermont                   11
District of Columbia      10
Maine                      8
North Dakota               7
West Virginia              4
Wyoming                    1
Name: State, dtype: int64
```

In [18]:

```
plt.figure(figsize=(15,15))
sns.countplot(x=sample1['State'])
plt.xticks(rotation=90)
plt.title("STATE")
plt.show()
```



In [19]:

```
!pip install git+https://github.com/has2k1/plotnine.git
```

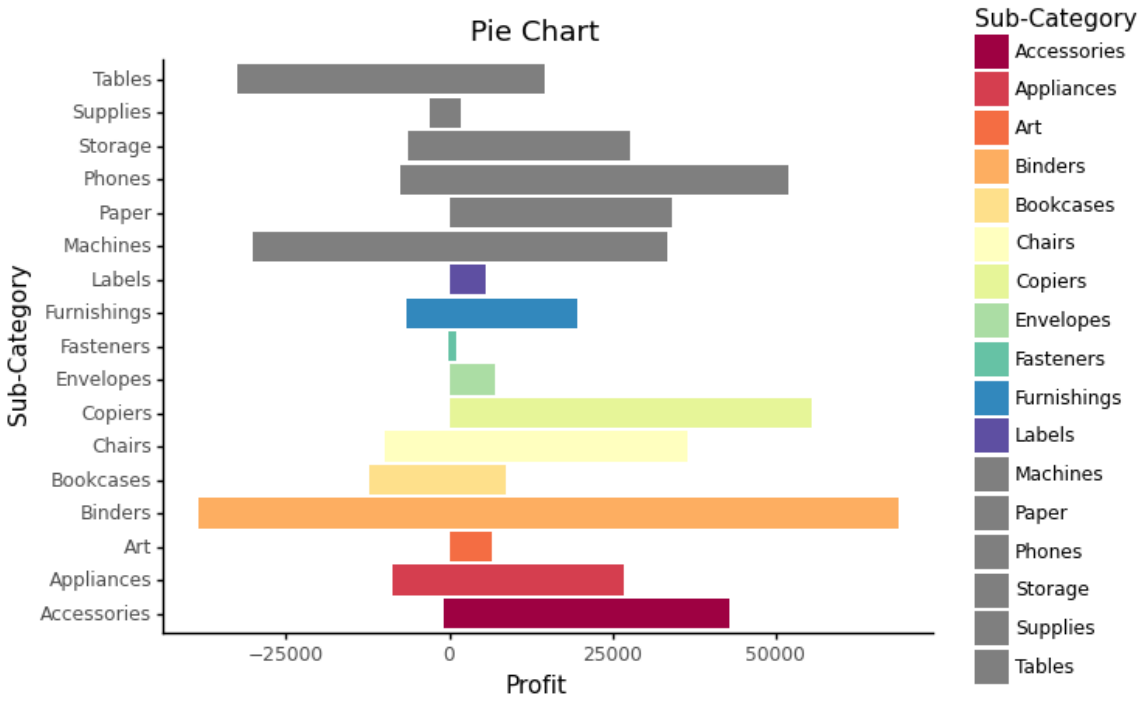
```
Collecting git+https://github.com/has2k1/plotnine.git

WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
  Running command git clone -q https://github.com/has2k1/plotnine.git (http
s://github.com/has2k1/plotnine.git) 'C:\Users\vanda\AppData\Local\Temp\pip-r
eq-build-hqr9iu5u'
  ERROR: Error [WinError 2] The system cannot find the file specified while
 executing command git clone -q https://github.com/has2k1/plotnine.git (http
s://github.com/has2k1/plotnine.git) 'C:\Users\vanda\AppData\Local\Temp\pip-r
eq-build-hqr9iu5u'
ERROR: Cannot find command 'git' - do you have 'git' installed and in your P
ATH?
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)



  Cloning https://github.com/has2k1/plotnine.git (https://github.com/has2k1/
plotnine.git) to c:\users\vanda\appdata\local\temp\pip-req-build-hqr9iu5u
```

In [20]:

```
from plotnine import *
from plotnine.data import mtcars
```

In [21]:

```
Profit_plot = (ggplot(sample, aes(x='Sub-Category', y='Profit', fill='Sub-Category')) + geo
+ scale_fill_brewer(type='div', palette="Spectral") + theme_classic() + ggtitle('Pie Chart'

display(Profit_plot)
```



<ggplot: (125242968311)>

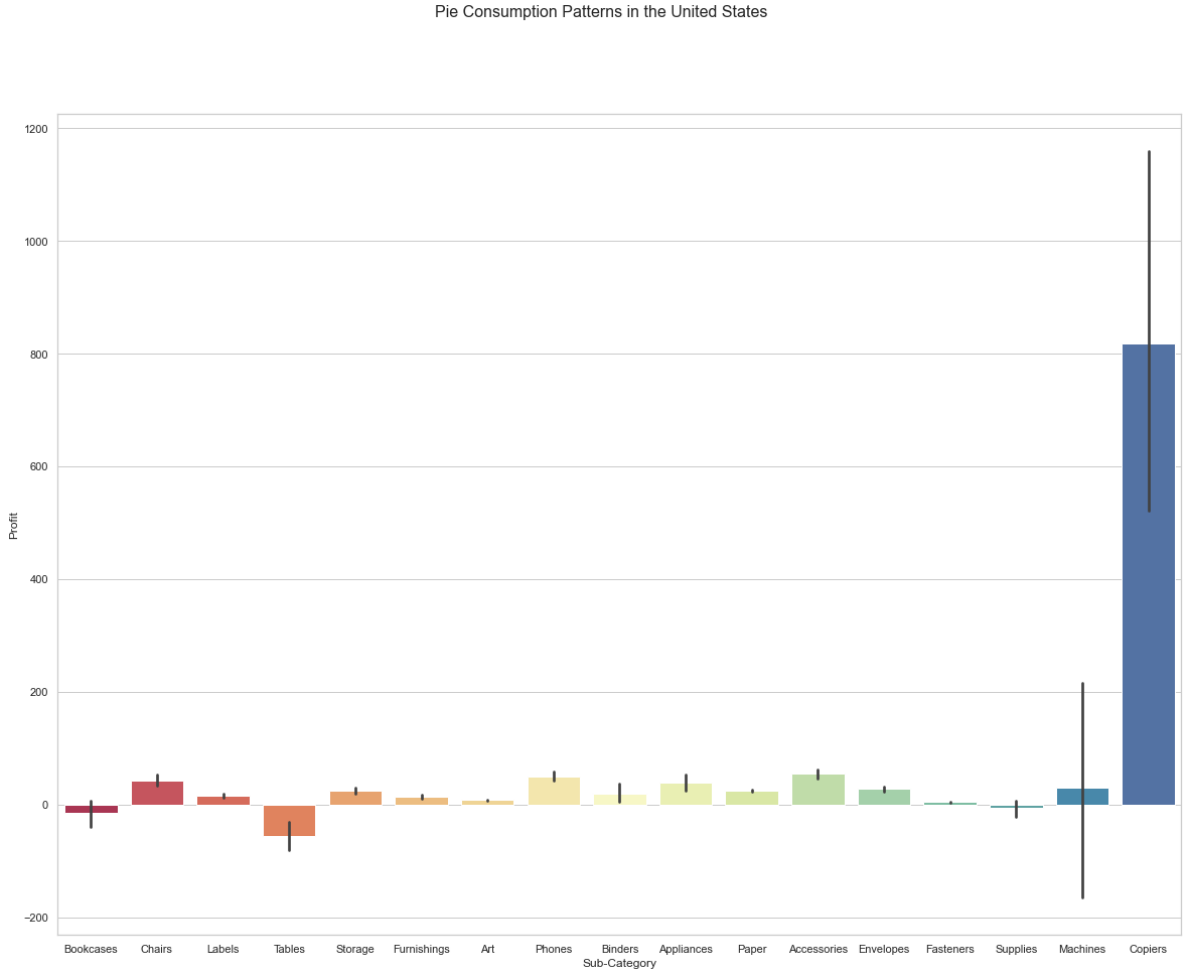
Above Pie chart Shows the profit and loss of each and every subcategories.Here from graph we can visualize that "binders" sub-category has suffered the highest amount of loss and also profit amongst all other sub-Categories (For now we can't say that what is the reason it may be because of discounts given on binders subcategory)

Next,"Copiers" Sub-category has gain highest amount of profit with no loss.There are other sub-categories too who are not faced any kind of losses but their profit margins are also low.

Next,Suffering from highest loss is machines

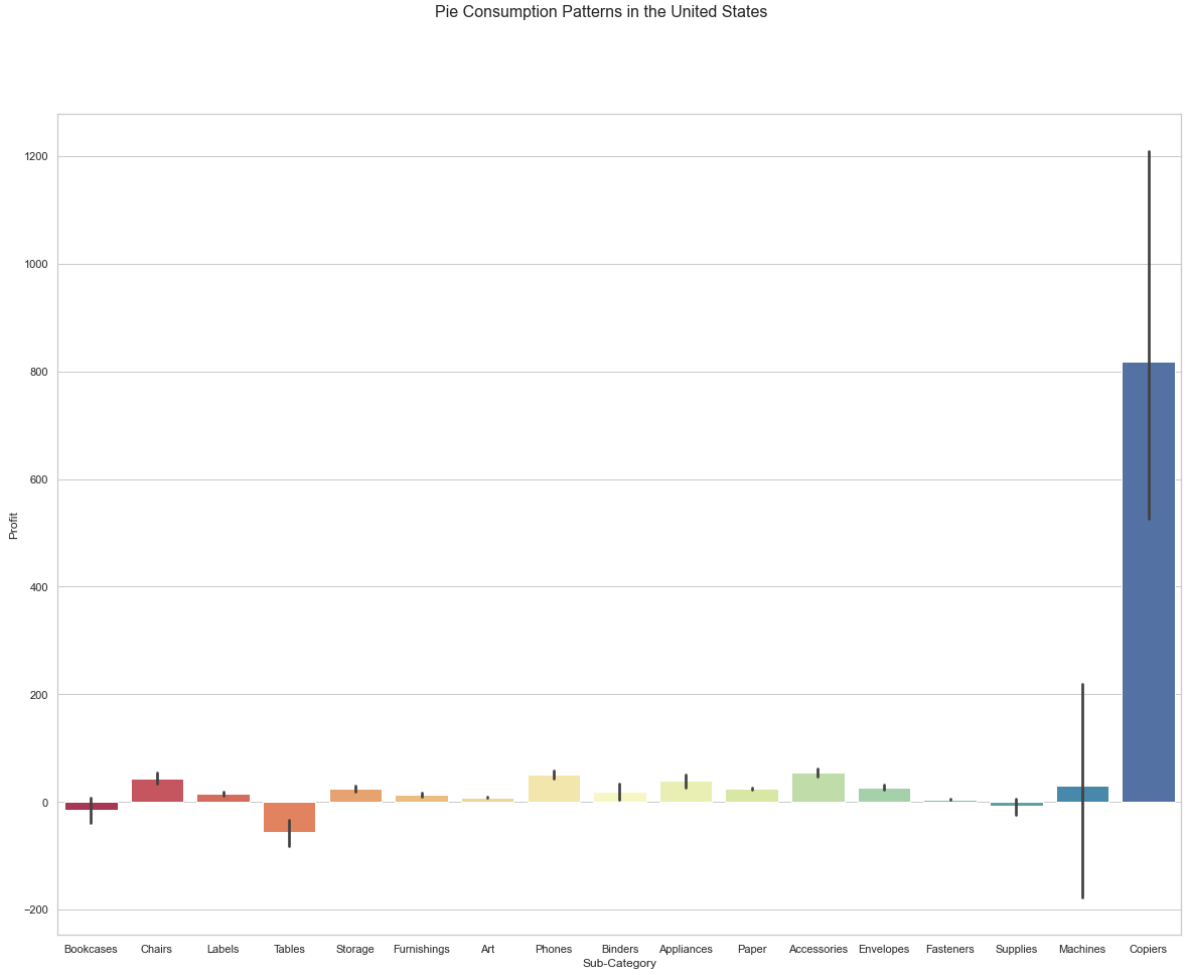
In [22]:

```
sns.set(style="whitegrid")
plt.figure(2, figsize=(20,15))
sns.barplot(x='Sub-Category',y='Profit', data=sample, palette='Spectral')
plt.suptitle('Pie Consumption Patterns in the United States', fontsize=16)
plt.show()
```



In [23]:

```
sns.set(style="whitegrid")
plt.figure(2, figsize=(20,15))
sns.barplot(x='Sub-Category',y='Profit', data=sample, palette='Spectral')
plt.suptitle('Pie Consumption Patterns in the United States', fontsize=16)
plt.show()
```



In [24]:

```
figsize=(15,10)
sns.pairplot(sample1,hue='Sub-Category')
plt.show
```

Out[24]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

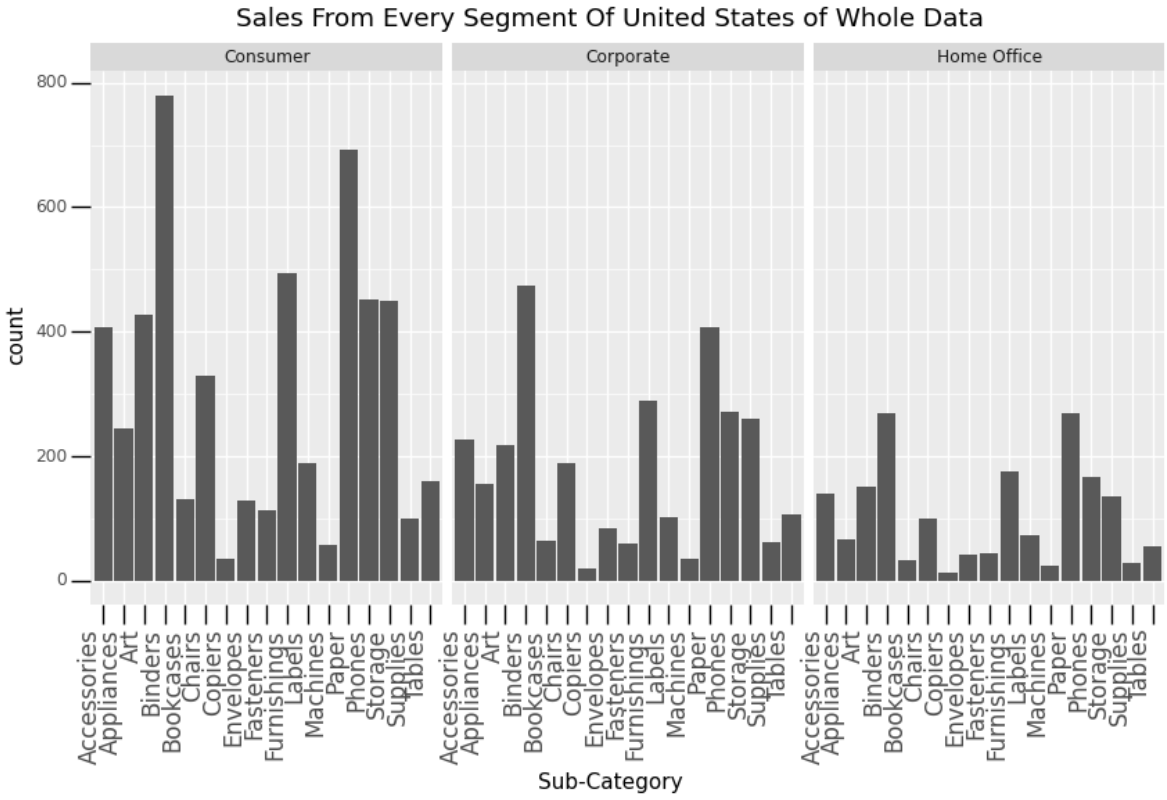
From the above plot we can say that Our Data is not Normal and it has some amount of outliers too.

Let's explore more about these outliers by using boxplots.

Ist we'll check Sales from Every Segments of Whole Data.

In [25]:

```
flip_xlabels = theme(axis_text_x = element_text(angle=90, hjust=1),figure_size=(10,5),
                     axis_ticks_length_major=10,axis_ticks_length_minor=5)
(ggplot(sample, aes(x='Sub-Category', fill='Sales')) + geom_bar() + facet_wrap(['Segment'])
+ flip_xlabels +theme(axis_text_x = element_text(size=12))+ggtitle("Sales From Every Segmen
```



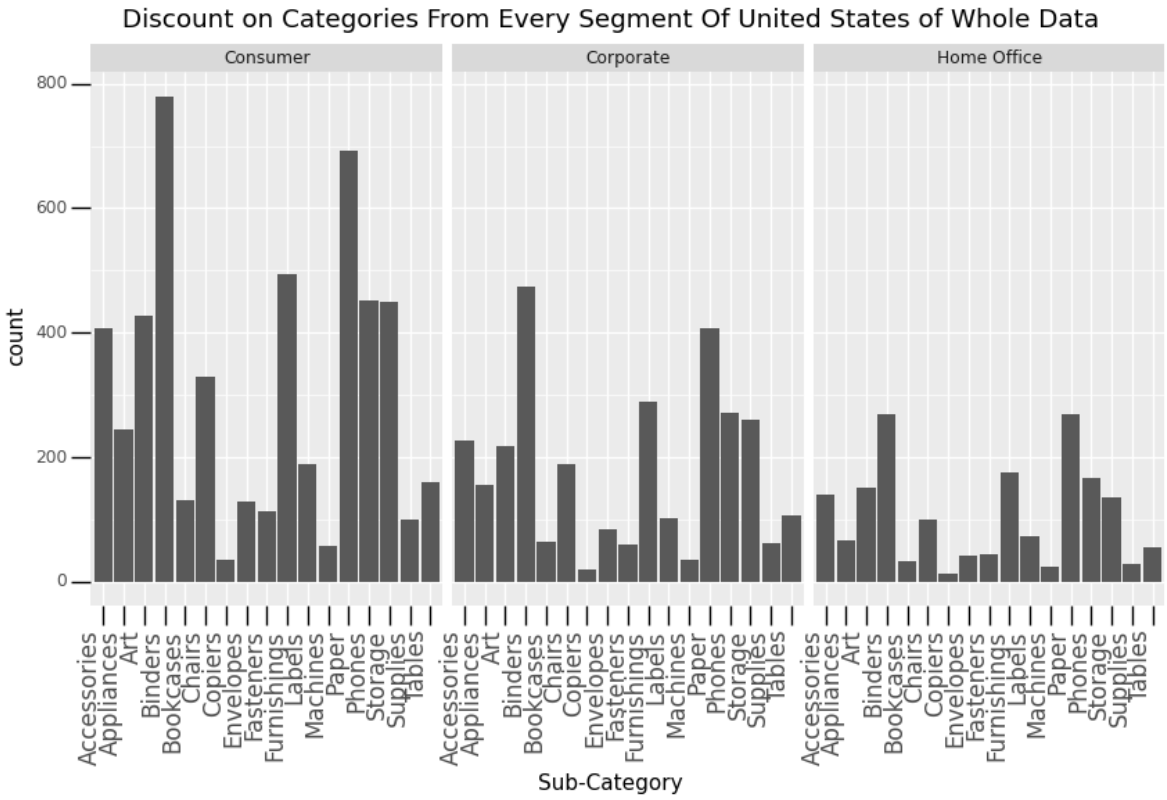
Out[25]:

```
<ggplot: (125243028405)>
```

From above Graph we can say that "Home Office" segment has less purchased Sub-Categories and in that "Tables","Supplies","Machines","Copiers","Bookcases" has the lowest Sales. "Consumer" has purchased more sub-categories as compared to other segments.

In [26]:

```
flip_xlabels = theme(axis_text_x = element_text(angle=90, hjust=1),figure_size=(10,5),
                     axis_ticks_length_major=10,axis_ticks_length_minor=5)
(ggplot(sample, aes(x='Sub-Category', fill='Discount')) + geom_bar() + facet_wrap(['Segment
+ flip_xlabels +theme(axis_text_x = element_text(size=12))+ggtitle("Discount on Categories
```

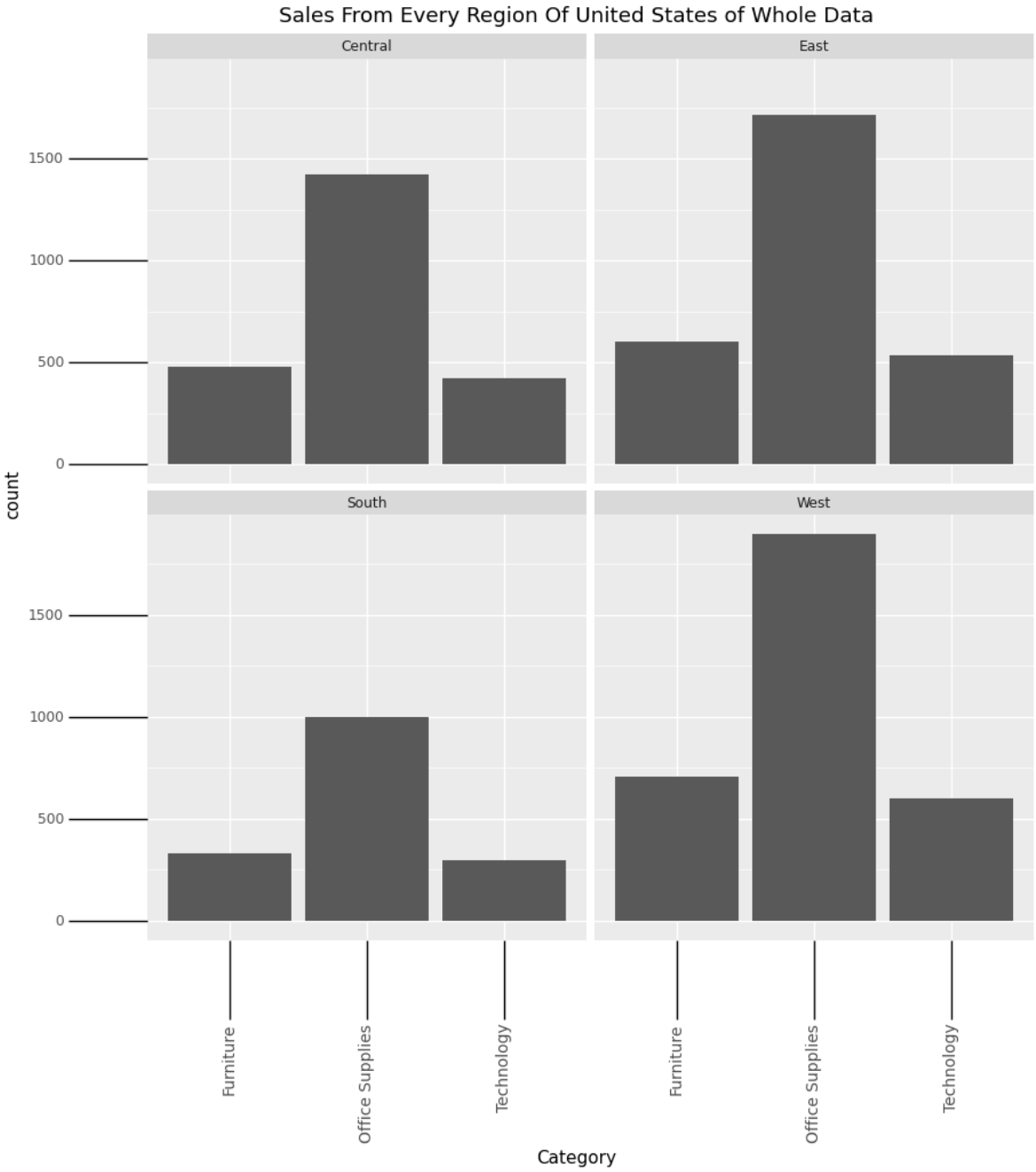


Out[26]:

<ggplot: (125242948136)>

In [27]:

```
flip_xlabels = theme(axis_text_x = element_text(angle=90, hjust=10),figure_size=(10,10),
                     axis_ticks_length_major=50,axis_ticks_length_minor=50)
(ggplot(sample1, aes(x='Category', fill='Sales')) + geom_bar() + theme(axis_text_x = elemen
+ facet_wrap(['Region']) + flip_xlabels+ ggtitle("Sales From Every Region Of United States
```

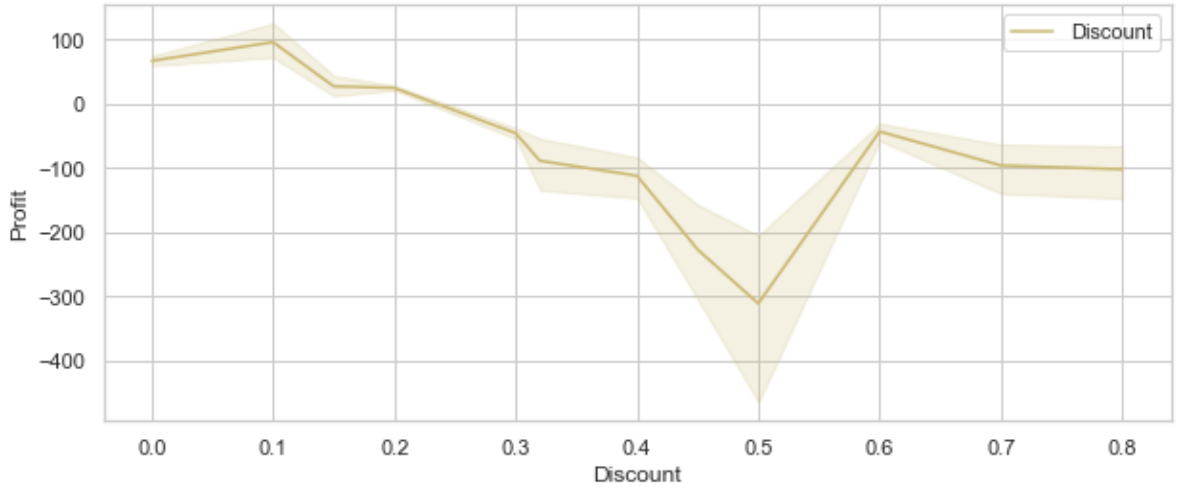


Out[27]:

```
<ggplot: (125243076340)>
```

In [28]:

```
plt.figure(figsize=(10,4))
sns.lineplot('Discount','Profit', data=sample1 , color='y',label='Discount')
plt.legend()
plt.show()
```



In [29]:

```
!pip install plotly==5.7.0
```

```
Requirement already satisfied: plotly==5.7.0 in c:\users\vanda\anaconda3\lib
\site-packages (5.7.0)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\vanda\anaconda3\l
ib\site-packages (from plotly==5.7.0) (8.0.1)
Requirement already satisfied: six in c:\users\vanda\anaconda3\lib\site-pack
ages (from plotly==5.7.0) (1.16.0)

WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
WARNING: Ignoring invalid distribution -atplotlib (c:\users\vanda\anaconda3
\lib\site-packages)
```

In [30]:

```
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
```

In [31]:

```
state_code = {'Alabama': 'AL','Alaska': 'AK','Arizona': 'AZ','Arkansas': 'AR','California':
sample1['state_code'] = sample1.State.apply(lambda x: state_code[x])
```

In [32]:

```
state_data = sample1[['Sales', 'Profit', 'state_code']].groupby(['state_code']).sum()


fig = go.Figure(data=go.Choropleth(
    locations=state_data.index,
    z = state_data.Sales,
    locationmode = 'USA-states',
    colorscale = 'Reds',
    colorbar_title = 'Sales in USD',
))

fig.update_layout(
    title_text = 'Total State-Wise Sales',
    geo_scope='usa',
    height=800,
)

fig.show()
```

> Now, let us analyze the sales of a few random states from each profit bracket (high profit, medium profit, low profit, low loss and high loss) and try to observe some crucial trends which might help us in increasing the sales.

We have a few questions to answer here.

1 What products do the most profit making states buy?

2 What products do the loss bearing states buy?

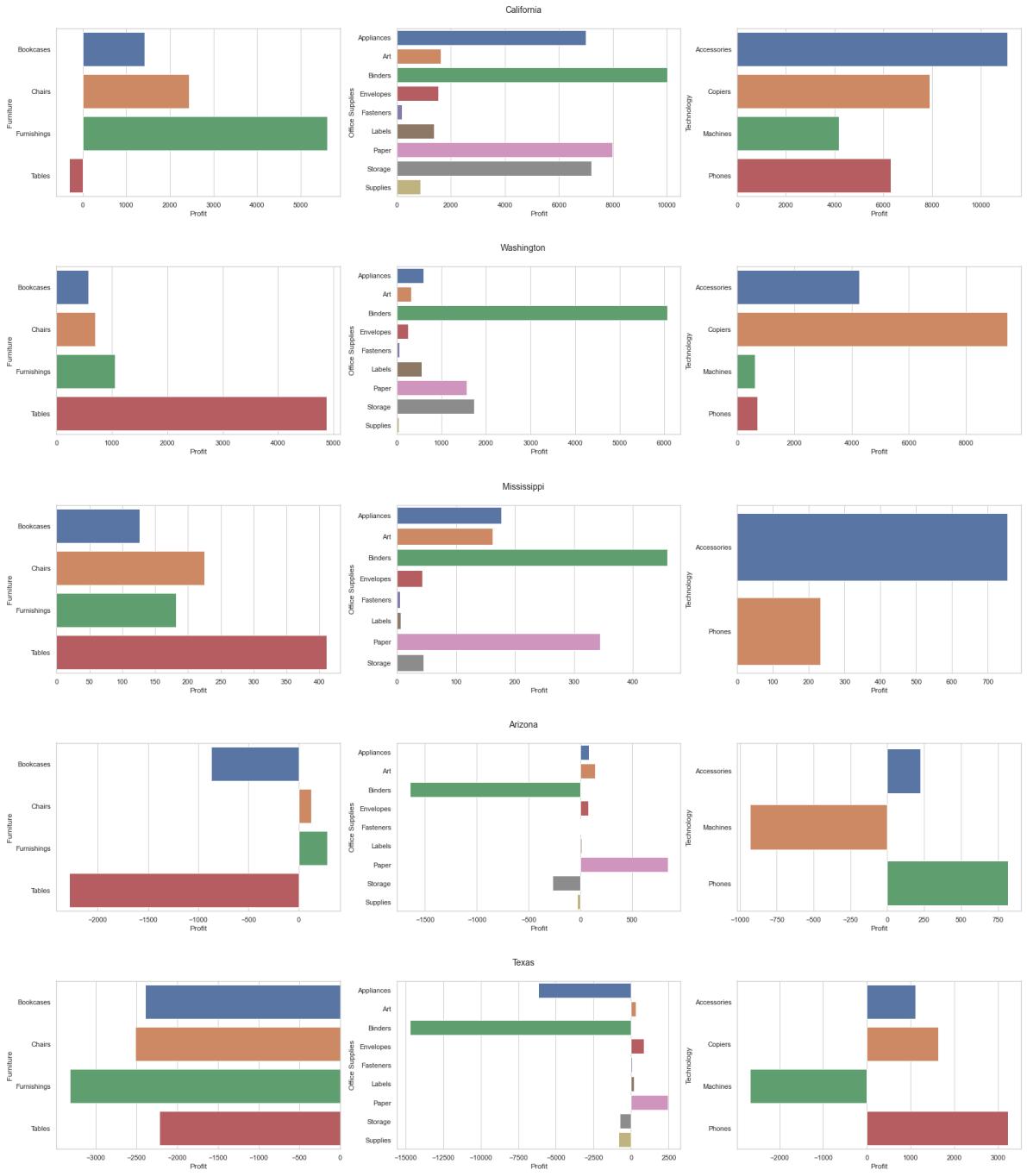
3 What product segment needs to be improved in order to drive the profits higher?

In [33]:

```
def state_data_viewer(states):
    """Plots the turnover generated by different product categories and sub-categories for
    Args:
        states- List of all the states you want the plots for
    Returns:
        None
    """
    product_data = sample1.groupby(['State'])
    for state in states:
        data = product_data.get_group(state).groupby(['Category'])
        fig, ax = plt.subplots(1, 3, figsize = (28,5))
        fig.suptitle(state, fontsize=14)
        ax_index = 0
        for cat in ['Furniture', 'Office Supplies', 'Technology']:
            cat_data = data.get_group(cat).groupby(['Sub-Category']).sum()
            sns.barplot(x = cat_data.Profit, y = cat_data.index, ax = ax[ax_index])
            ax[ax_index].set_ylabel(cat)
            ax_index +=1
        fig.show()
```

In [34]:

```
states = ['California', 'Washington', 'Mississippi', 'Arizona', 'Texas']
state_data_viewer(states)
```



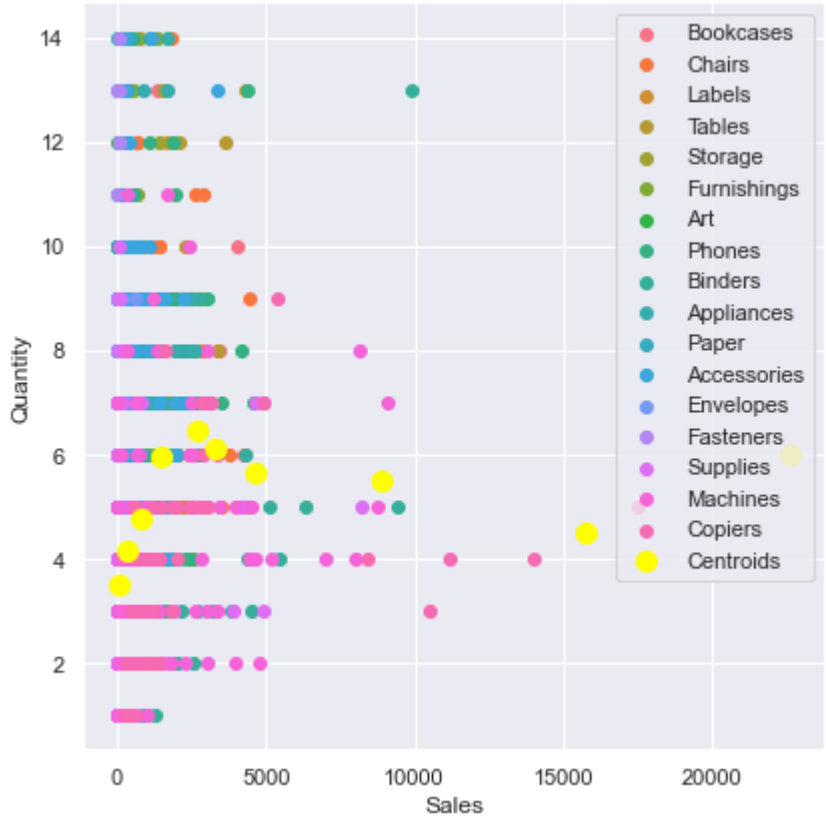
Using Cluster Analysis(K-Mean Clustering)

In [35]:

```
x = sample.iloc[:, [9, 10, 11, 12]].values

from sklearn.cluster import KMeans
wcss = []

for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++',
                    max_iter = 300, n_init = 10, random_state = 0).fit(x)
    wcss.append(kmeans.inertia_)
```

In [36]:

```
sns.set_style("darkgrid")
sns.FacetGrid(sample, hue ="Sub-Category",height = 6).map(plt.scatter,'Sales','Quantity')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1],
            s = 100, c = 'yellow', label = 'Centroids')

plt.legend()
plt.show()
```



In [37]:

```
sns.set_style("darkgrid")
sns.FacetGrid(sample, hue ="Sub-Category",height = 6).map(plt.scatter,'Sales','Profit')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:,1],
            s = 100, c = 'yellow', label = 'Centroids')

plt.legend()
plt.show()
```

In [38]:

```
fig, ax = plt.subplots(figsize = (10 , 6))
ax.scatter(sample1["Sales"] , sample1["Profit"])
ax.set_xlabel('Sales')
ax.set_ylabel('Profit')
ax.set_title('Sales vs Profit')
plt.show()
```

From the Above data Visualization and Clustering we can see that in Which states and in which Category Sales and profits are High or less,We can improve in that States By Providing Discounts in prefered Range so that Company and cosumer both will be in profit.So For Deciding that Range we have to do some Technical Analysis.One can Do it through Factor Analysis,or also can Do it throgh neural networks.

One thing to be noted is that while the superstore is incurring losses due to giving discounts on its products, they can't stop giving discounts of their products. Most of the heavy discounts are during festivals, end-of-season and clearance sales which are necessary so that the store can make space in their warehouses for fresh stock. Also, by incurring small losses, the company gains in the future by attracting more long term customers. Therefore, the small losses from discounts are an essential part of company's business.

In [ ]: