GRIP - The Sparks Foundation

Task - Exploratory Data Analysis - Retail

Dataset- Sample Superstore

This dataset is about the sale and the profit earned by a Sample store in the US.

Lets first import all the essentials library and the instances.

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Reading the Data

In [4]:

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

In [5]:

```
1 df
```

Out[5]:

	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

9994 rows × 13 columns

Following set of code is about the general information about the data:

In [4]:

```
1 df.shape
```

Out[4]:

```
(9994, 13)
```

In [5]:

```
1 df.columns
```

Out[5]:

```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity', 'Discoun
t',
       'Profit'],
      dtype='object')
```

In [6]:

```
1 df.dtypes
```

Out[6]:

```
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
```

In [7]:

```
1 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
```

In [8]:

```
1 df.describe()
```

Out[8]:

	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

The above output gives good information about the dataset, in this analysis I mainly foccused on the Sales and Profit and thier trend with various aspects.

In [9]:

```
1 df.loc[df.Profit==8399.976000]
```

Out[9]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	
6826	Standard Class	Corporate	United States	Lafayette	Indiana	47905	Central	Technology	Copiers	17

In [10]:

```
1 df.loc[df.Profit==-6599.978]
```

Out[10]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	
7772	Standard Class	Consumer	United States	Lancaster	Ohio	43130	East	Technology	Machines	449

The above outputs clearly tell that the Lafayette city of US had the maximum profit and a sale of 17499.95 and the maximum loss with the sale of 4499.985 was of Lancaster city of the US.

Lets sort the dataset by the Profit gained, and take a look at the first few rows of

the sorted dataset.

In [11]:

```
df.sort_values('Profit', ascending=False, inplace = True)
```

In [12]:

```
df.head(15)
```

Out[12]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Su Categ
6826	Standard Class	Corporate	United States	Lafayette	Indiana	47905	Central	Technology	Copie
8153	First Class	Consumer	United States	Seattle	Washington	98115	West	Technology	Copie
4190	Standard Class	Consumer	United States	Newark	Delaware	19711	East	Technology	Copie
9039	Standard Class	Consumer	United States	Detroit	Michigan	48205	Central	Office Supplies	Binde
4098	Standard Class	Consumer	United States	Minneapolis	Minnesota	55407	Central	Office Supplies	Binde
2623	First Class	Home Office	United States	New York City	New York	10024	East	Technology	Copie
509	Standard Class	Consumer	United States	Atlanta	Georgia	30318	South	Office Supplies	Binde
8488	Second Class	Consumer	United States	Arlington	Virginia	22204	South	Technology	Machin
7666	Standard Class	Home Office	United States	Providence	Rhode Island	2908	East	Technology	Copie
6520	Second Class	Consumer	United States	Jackson	Michigan	49201	Central	Office Supplies	Binde
1085	Standard Class	Consumer	United States	Yonkers	New York	10701	East	Technology	Machin
4277	Standard Class	Corporate	United States	Lakewood	New Jersey	8701	East	Technology	Machin
8990	Standard Class	Corporate	United States	Springfield	Missouri	65807	Central	Technology	Copie
6626	Standard Class	Consumer	United States	New York City	New York	10024	East	Technology	Machin
8204	Same Day	Corporate	United States	New York City	New York	10024	East	Technology	Machin

Cleaning the Data, i.e, removing the columns which are not required in this analysis.

The Country column has only one value(US), so we can drop that column. Also, the analysis was done on the regions and state, thus, City and Postal Column is also of no use to us.

In [13]:

```
df_state =df.drop(['Country','City','Postal Code'], axis=1)
```

In [14]:

```
df_state
```

Out[14]:

	Ship Mode	Segment	State	Region	Category	Sub-Category	Sales	Quantity	Disc
6826	Standard Class	Corporate	Indiana	Central	Technology	Copiers	17499.950	5	
8153	First Class	Consumer	Washington	West	Technology	Copiers	13999.960	4	
4190	Standard Class	Consumer	Delaware	East	Technology	Copiers	10499.970	3	
9039	Standard Class	Consumer	Michigan	Central	Office Supplies	Binders	9892.740	13	
4098	Standard Class	Consumer	Minnesota	Central	Office Supplies	Binders	9449.950	5	
...	...	...	...	...	...	...	...	...	
4991	Standard Class	Corporate	Illinois	Central	Office Supplies	Binders	1889.990	5	
3011	Standard Class	Home Office	Colorado	West	Technology	Machines	2549.985	5	
9774	Standard Class	Consumer	Texas	Central	Office Supplies	Binders	2177.584	8	
683	Same Day	Corporate	North Carolina	South	Technology	Machines	7999.980	4	
7772	Standard Class	Consumer	Ohio	East	Technology	Machines	4499.985	5	

9994 rows × 10 columns

Checking if there is any null value present in the dataset.

In [15]:

```
1 df_state.isnull().sum()
```

Out[15]:

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

ANALYSIS of the Data

This can be done in major three ways:

1. Product Level Analysis(Category and Sub-Category wise)
2. Region Level Analysis
3. Customer Level Analysis

Lets start with the Product level Analysis.

Broadly, it is divide into two major domains- Category and Sub-Category.

Lets, first check what are the unique Categories in the dataset, plotting the same with Number of such products by counting the respective values.

In [16]:

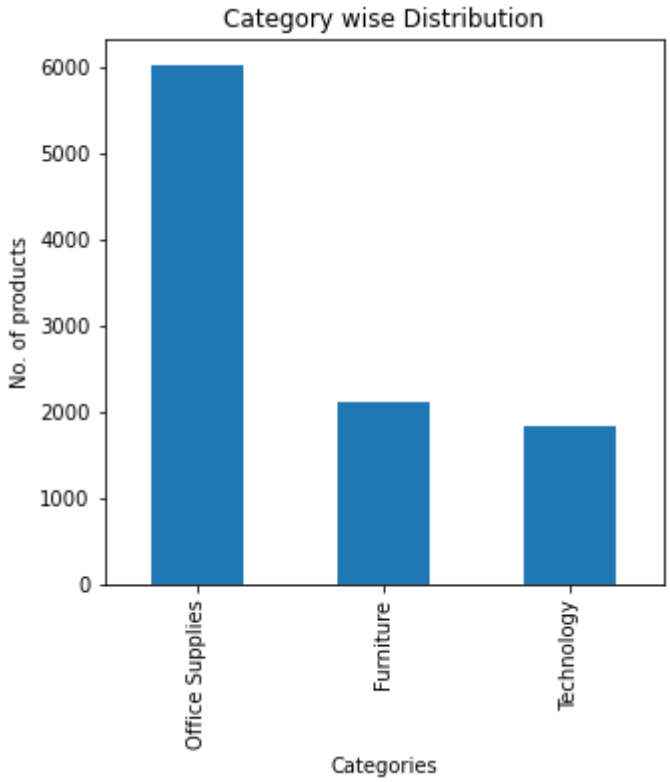
```
1 df_state['Category'].unique()
```

Out[16]:

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

In [17]:

```
df_state['Category'].value_counts().plot(kind='bar', figsize=(5,5))
plt.title('Category wise Distribution')
plt.xlabel('Categories')
plt.ylabel('No. of products')
plt.show()
```



With the above plot, we can easily conclude the highest number of products sold from the supermarket was of the Office Suppliers Category. Let's now check which sub category had the highest sale.

For that we need to know what are the unique Sub-Categories present in the Dataset.

In [18]:

```
1 df_state['Sub-Category'].unique()
```

Out[18]:

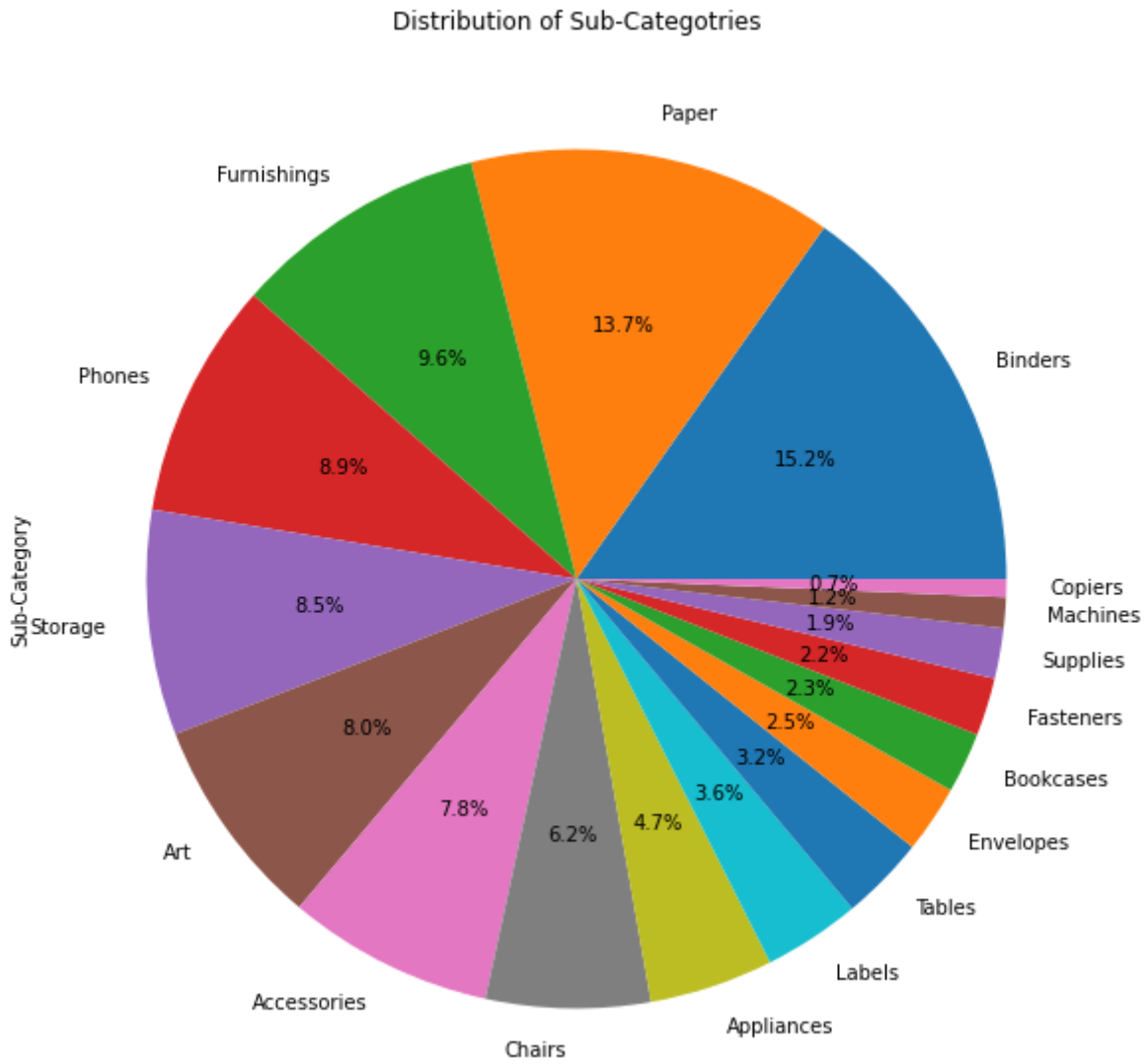
```
array(['Copiers', 'Binders', 'Machines', 'Phones', 'Bookcases',
       'Accessories', 'Appliances', 'Storage', 'Chairs', 'Tables',
       'Furnishings', 'Labels', 'Paper', 'Supplies', 'Envelopes', 'Art',
       'Fasteners'], dtype=object)
```

In [19]:

```
1 a=df_state['Sub-Category'].value_counts()
```

In [20]:

```
1 a.plot.pie(autopct="%1.1f%%",figsize=(10,10))
2 plt.title('Distribution of Sub-Categotries')
3 plt.show()
```



This can be clearly seen, the Copiers and Machines have the least number of products sold where as Papers and Binders had the highest Number of products sold.

Now, to have even more clear picture, lets now plot the Sub-Categories inside each Category with the Sales and Profit Gained.

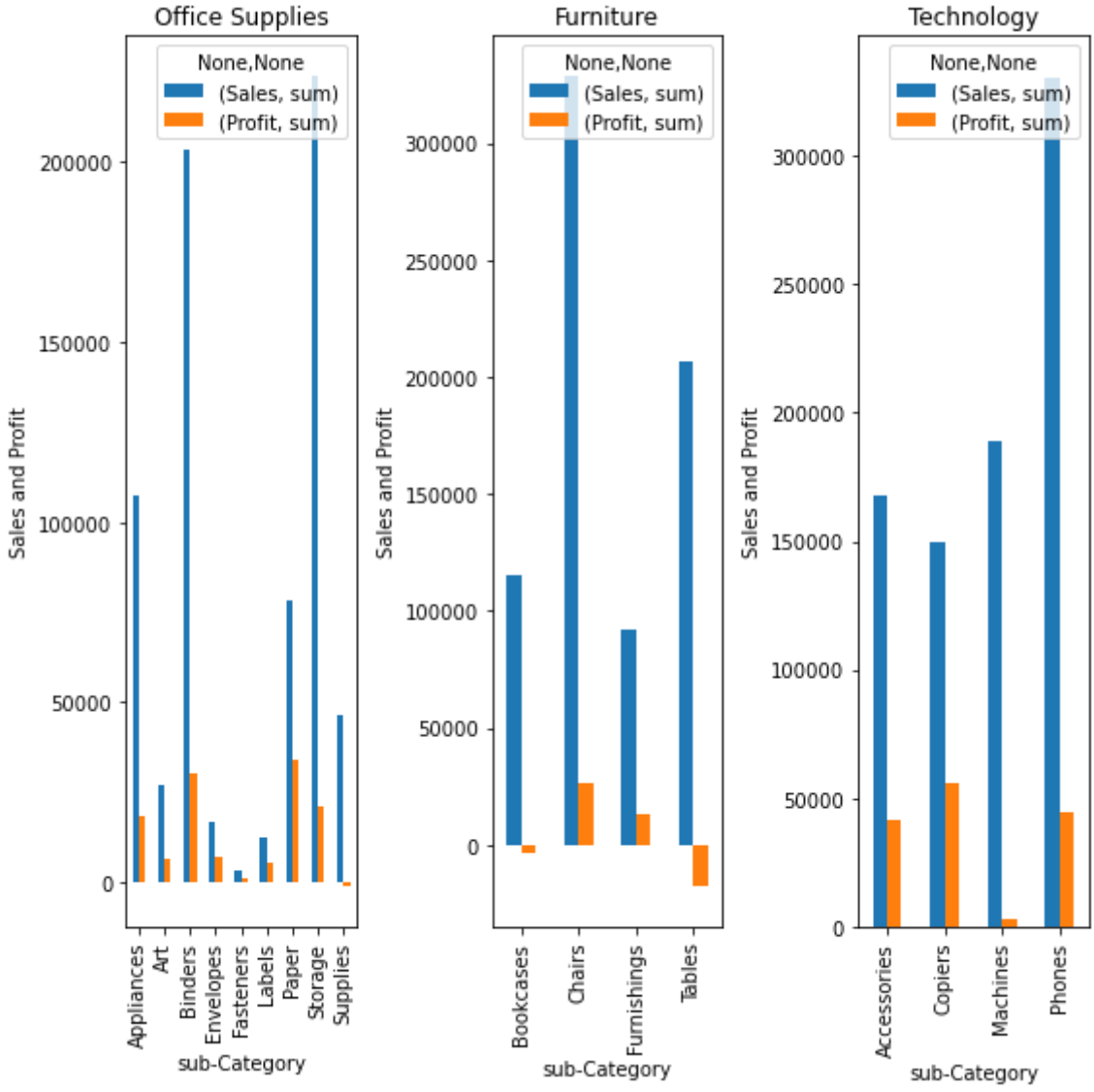
In [21]:

```
df_c1= df_state.loc[df_state['Category']=='Office Supplies']
df_c2=df_state.loc[df_state['Category']=='Furniture']
df_c3=df_state.loc[df_state['Category']=='Technology']
a=df_c1.groupby('Sub-Category')[['Sales','Profit']].agg(['sum'])
b=df_c2.groupby('Sub-Category')[['Sales','Profit']].agg(['sum'])
c=df_c3.groupby('Sub-Category')[['Sales','Profit']].agg(['sum'])
```

In [22]:

```
fig=plt.figure()
ax0=fig.add_subplot(1,3,1)
ax1=fig.add_subplot(1,3,2)
ax2=fig.add_subplot(1,3,3)

a.plot(kind='bar', ax=ax0,figsize=(8,8))
ax0.set_title('Office Supplies')
ax0.set_xlabel('sub-Category')
ax0.set_ylabel('Sales and Profit')
b.plot(kind='bar', ax=ax1,figsize=(8,8))
ax1.set_title('Furniture')
ax1.set_xlabel('sub-Category')
ax1.set_ylabel('Sales and Profit')
c.plot(kind='bar', ax=ax2,figsize=(8,8))
ax2.set_title('Technology')
ax2.set_xlabel('sub-Category')
ax2.set_ylabel('Sales and Profit')
fig.tight_layout()
plt.show()
```



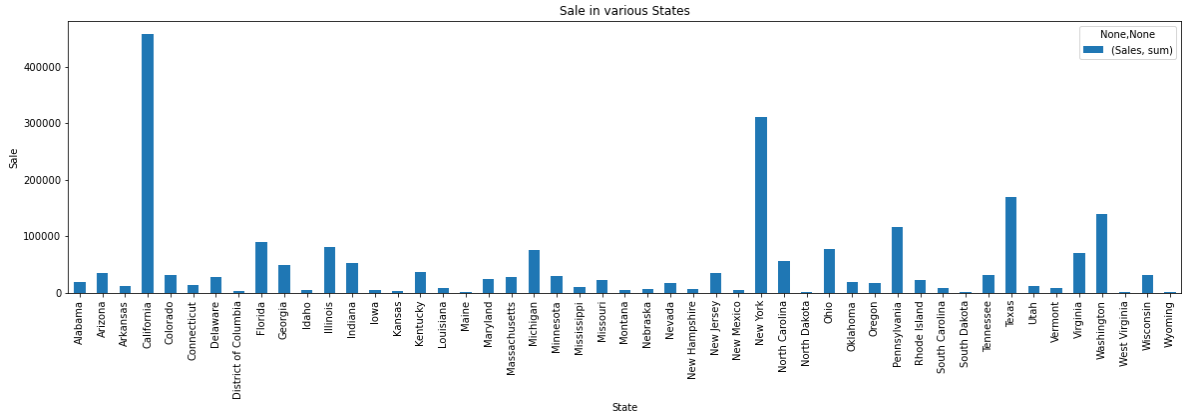
As per the above plot, it is well understood that the sale of Fasteners was least among all the three Sub-Categories, and that of Chairs and Phones were highest and the profit from Copiers is highest among all, where as the loss by Tables is highest.

Lets now work on the State and Region Level Analysis.

First of all, check how the Sale Trend goes with respect to the State.

In [23]:

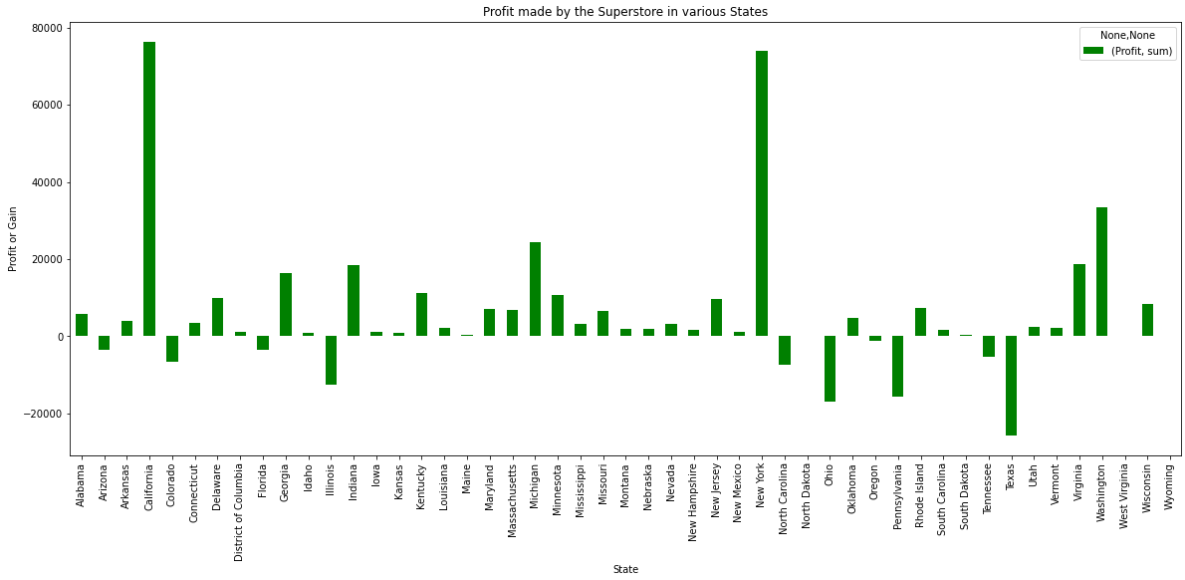
```
df_state.groupby('State')[['State','Sales']].agg(['sum']).plot(kind='bar',figsize=(20,5
plt.ylabel('Sale')
plt.title('Sale in various States')
plt.show()
```



The highest sale was in California and New york, the Texas and Washington had the moderate Sale and the other state had comparably low Sale.

In [24]:

```
df_state.groupby('State')[['State','Profit']].agg(['sum']).plot(kind='bar',figsize=(20,
plt.ylabel('Profit or Gain')
plt.title('Profit made by the Superstore in various States')
plt.show()
```

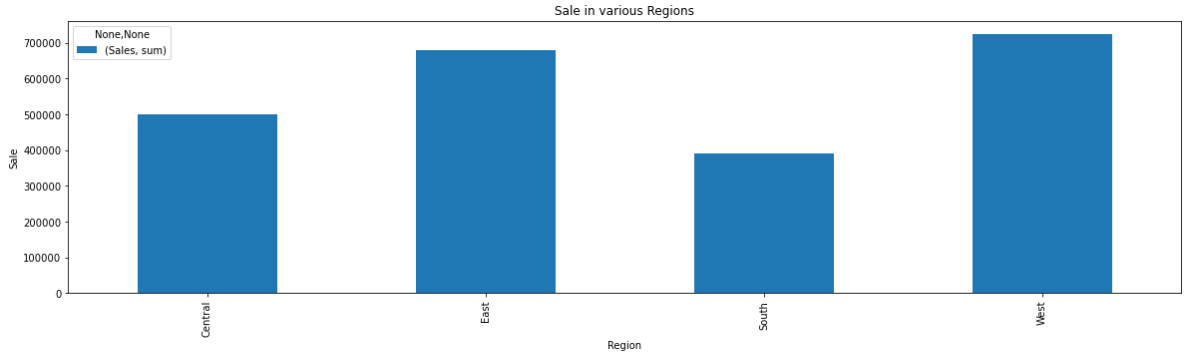


The Sale and Profit are directly propotional to each other, the same can be seem here, the States with the highest sale i.e., California and New york shows the cities where the Supermarket earned the highest Profit. The Texas being an exception case shows a higher loss and Washington also had a good profit when compared to the other states.

Lets now see the Sale and Profit Trend with respect to the different regions in the Country.

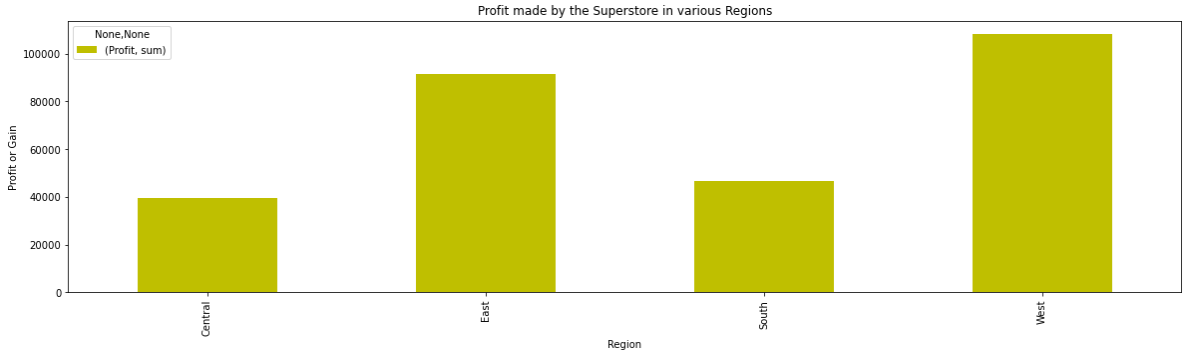
In [25]:

```
df_state.groupby('Region')[['Region','Sales']].agg(['sum']).plot(kind='bar',figsize=(20
plt.ylabel('Sale')
plt.title('Sale in various Regions')
plt.show()
```



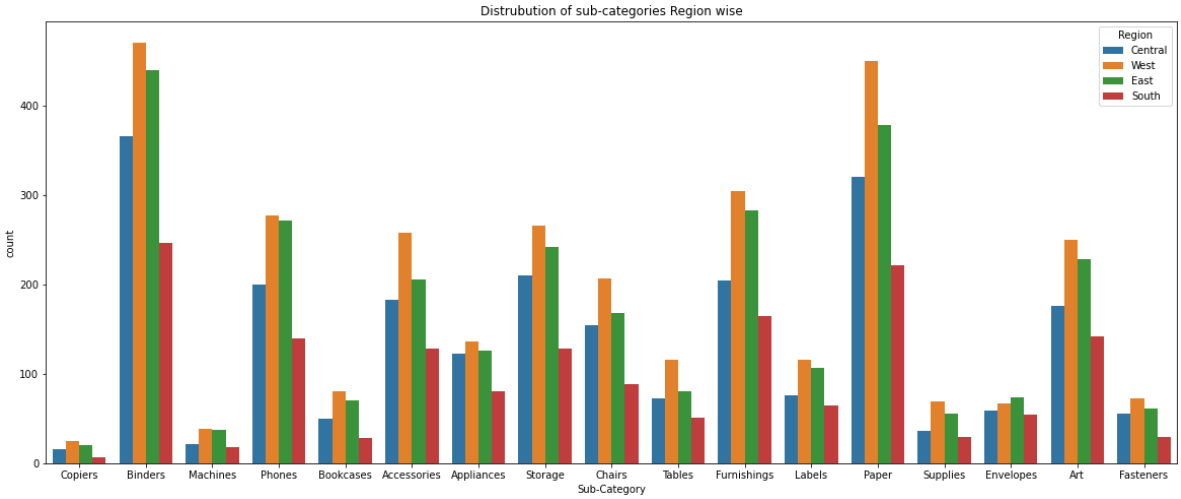
In [26]:

```
df_state.groupby('Region')[['Region','Profit']].agg(['sum']).plot(kind='bar',figsize=(2
plt.ylabel('Profit or Gain')
plt.title('Profit made by the Superstore in various Regions')
plt.show()
```



In [27]:

```
1 plt.figure(figsize=(20,8))
2 sns.countplot(x="Sub-Category", hue="Region", data=df_state)
3 plt.title('Distrubution of sub-categories Region wise')
4 plt.show()
```



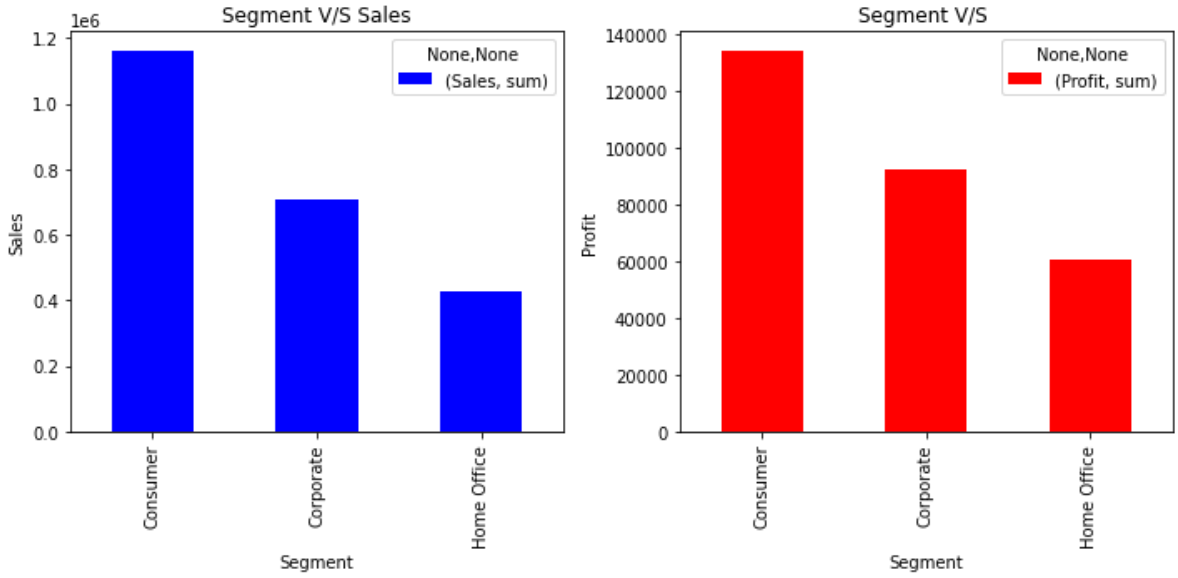
With above three plots, the inference be taken out the West Region of the Country is the region where supermarket had the highest as well as profit, the central region had more sale then Souhtern region but had lesser Profit from it.

Lets now work on the Segment or Consumer Level Analysis.

We have to check how the Sale and Profit trend goes with Segment.

In [28]:

```
fig=plt.figure()
ax0=fig.add_subplot(1,2,1)
ax1=fig.add_subplot(1,2,2)
df_state.groupby('Segment')[['Segment','Sales']].agg(['sum']).plot(kind='bar',ax=ax0,fi
ax0.set_title('Segment V/S Sales')
ax0.set_xlabel('Segment')
ax0.set_ylabel('Sales')
df_state.groupby('Segment')[['Segment','Profit']].agg(['sum']).plot(kind='bar',ax=ax1,f
ax1.set_title('Segment V/S ')
ax1.set_xlabel('Segment')
ax1.set_ylabel('Profit')
fig.tight_layout()
plt.show()
```



The Consumer domain of the Segment had the highest Sale as well as the profit and home office had the least.

Conclusion

We have seen the analysis at various level, and at every level we found that there are various areas where we can work to make more profit.

By - Vedangi Sharma

In [ ]:

```
1
```