Data Science And Bussiness Analytics Intern At TheSparksFoundation

GRIPJAN21

Author- Rajat Kumar

Task 3:-Exploratory Data Analysis - Retail

Problem Statement:- From the given SampleSuperstore dataset my work is to find Business Problem and also find weak areas where I can work to make more profit.

Step 1: Importing Libraries

In [82]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Step 2: Reading The Dataset

In [83]:
```
sample=pd.read_csv("E:\TSF\Task3\samplesuperstore.csv")
sample.head() # This command is used to load first five row of dataset
```

Out[83]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

In [84]:
```
sample.tail() # This command is used to load last five row of dataset.
```

Out[84]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.248	3	0.2	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.960	2	0.0	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	258.576	2	0.2	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	29.600	4	0.0	13.3200
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	243.160	2	0.0	72.9480

In [85]:
```
sample.shape # This command will give total numbers of row and total number of columns in a array.
```

Out[85]:
```
(9994, 13)
```

Step3: In this step I am checking type of data

In [86]:
```
sample.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 [87]:
```
sample.describe() #  This command will give some statistics Information such as mean, std,meadian etc.
```

Out[87]:

	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 [88]:
```
sample.isnull().sum() # This command will give us any missiing value.
```

Out[88]:
```
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
```

Step 4: Checking for the duplicate data, if yes then drop those dat.

In [89]:
```
sample.duplicated().sum() # This command will give number of duplicate data.
```

Out[89]: 17

In [90]:
```
sample.drop_duplicates() # Here I are trying to drop duplicate data.
```

Out[90]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	25.2480	3	0.20	4.1028

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	91.9600	2	0.00	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	258.5760	2	0.20	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	29.6000	4	0.00	13.3200
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	243.1600	2	0.00	72.9480

9977 rows × 13 columns

In [91]:
```
sample.nunique() # Here I am checking for unique data
```

Out[91]:
```
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
```

Step 5:Dropping Unwanted Columns

In [92]:
```
col=['Postal Code'] # Here we do not requried Postal code so we will drop this column.
sample1=sample.drop(columns=col,axis=1)
```

Step 6: Finding the statical relation between various rows and columns.

In [93]:
```
sample1.corr() # It will give corelation of variables
```

Out[93]:

	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 [94]:
```
sample1.cov() # it will give covariance of variable
```

Out[94]:

	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 [95]:
```
sample1.head() # It will first five row.
```

Out[95]:

	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

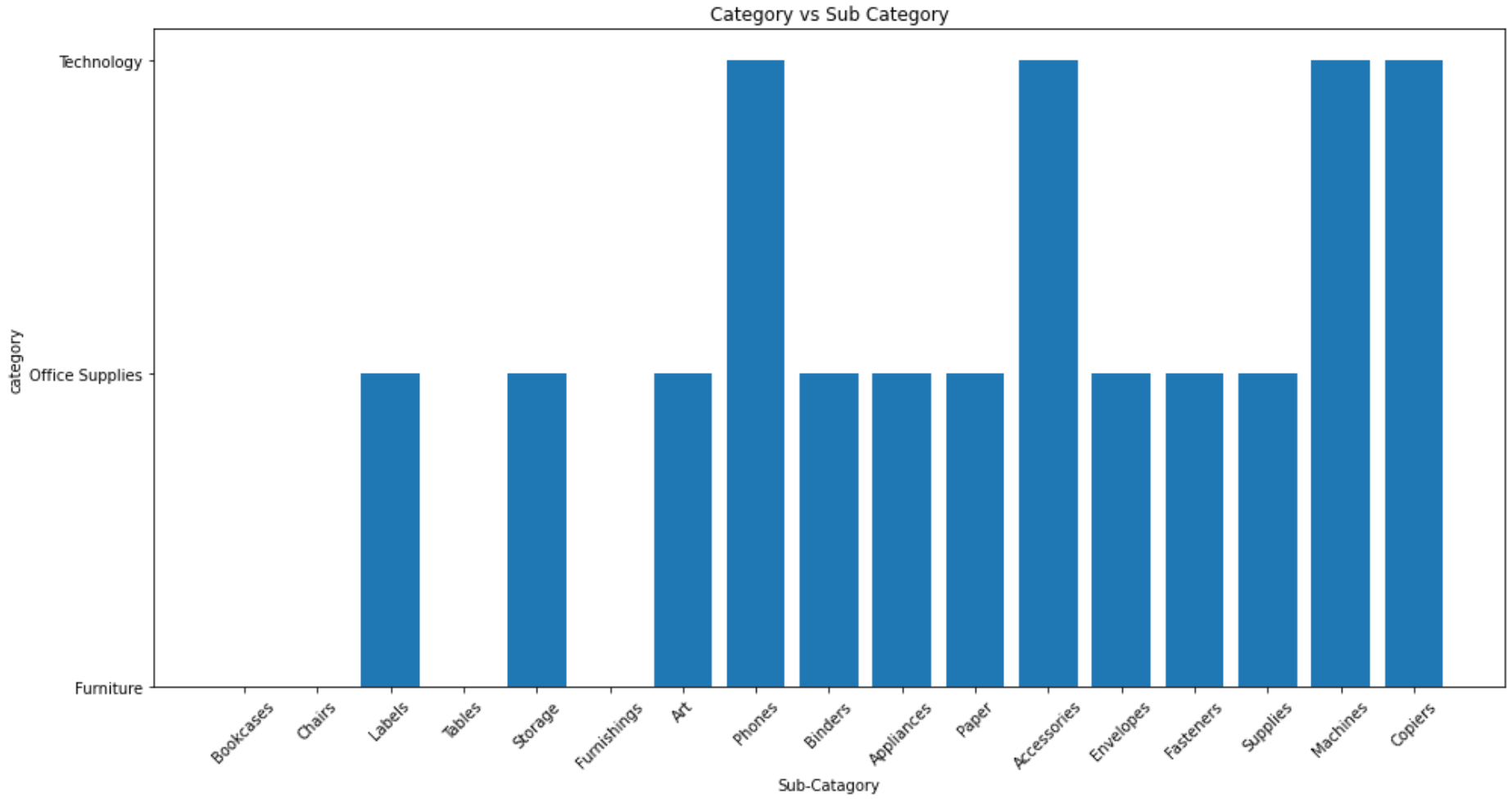
In [96]:
```
sample1.tail() # it will give last five row.
```

Out[96]:

	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9989	Second Class	Consumer	United States	Miami	Florida	South	Furniture	Furnishings	25.248	3	0.2	4.1028
9990	Standard Class	Consumer	United States	Costa Mesa	California	West	Furniture	Furnishings	91.960	2	0.0	15.6332
9991	Standard Class	Consumer	United States	Costa Mesa	California	West	Technology	Phones	258.576	2	0.2	19.3932
9992	Standard Class	Consumer	United States	Costa Mesa	California	West	Office Supplies	Paper	29.600	4	0.0	13.3200
9993	Second Class	Consumer	United States	Westminster	California	West	Office Supplies	Appliances	243.160	2	0.0	72.9480

Step 7: Data Visualization

In [97]:
```
plt.figure(figsize=(16,8))
plt.bar("Sub-Category","Category",data=sample)
plt.title("Category vs Sub Category")
plt.xlabel("Sub-Catagory")
plt.ylabel("category")
plt.xticks(rotation=45)
plt.show()
```



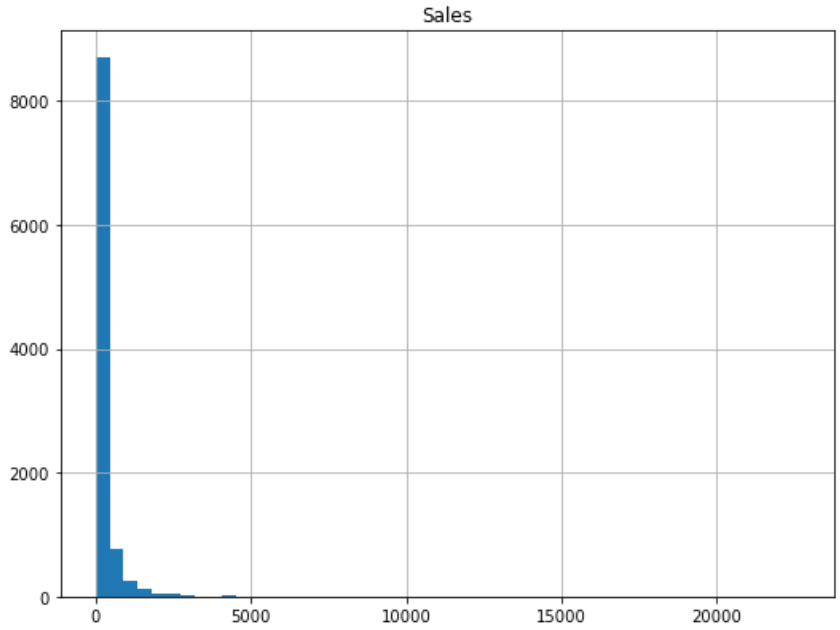
In [98]:
```
sample1.corr()
```

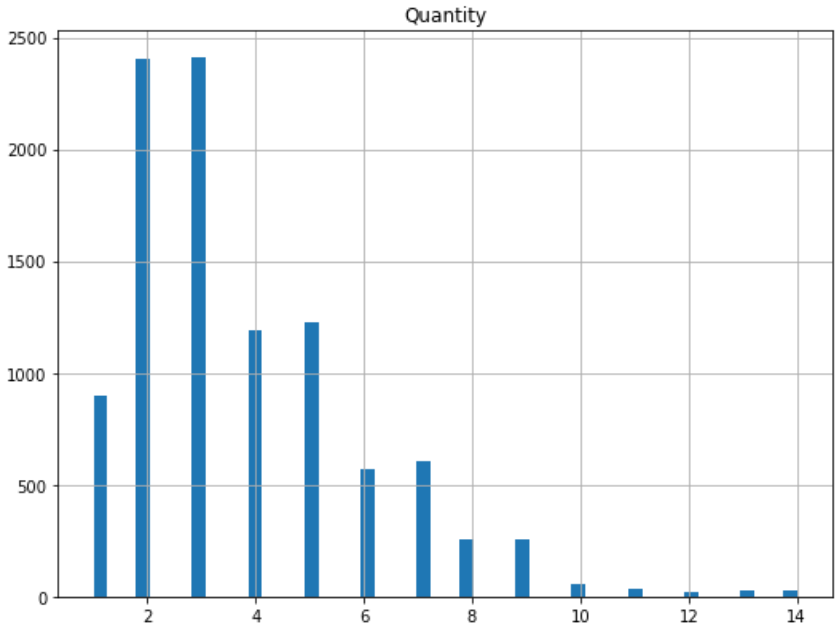
Out[98]:

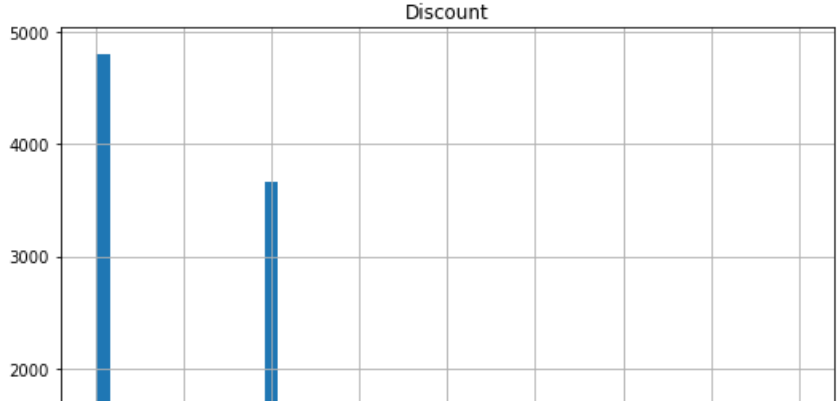
	Sales	Quantity	Discount	Profit
Sales	1.000000	0.200795	-0.028190	0.479064
Quantity	0.200795	1.000000	0.008623	0.066253
Discount	-0.028190	0.008623	1.000000	-0.219487
Profit	0.479064	0.066253	-0.219487	1.000000

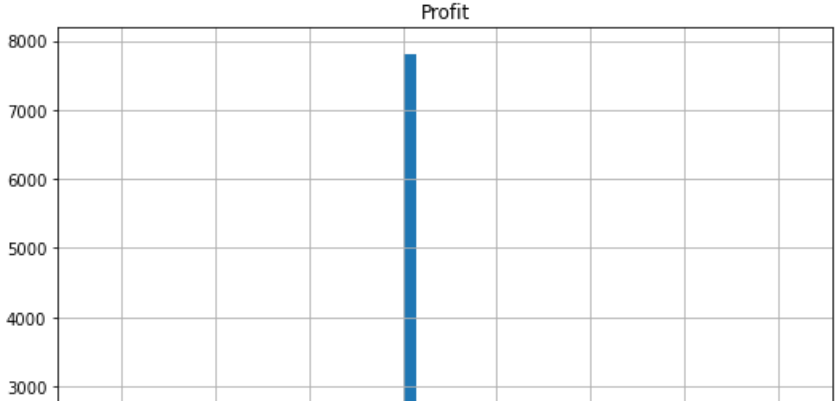
Step 8: With the help of histogram I have represented four variables.

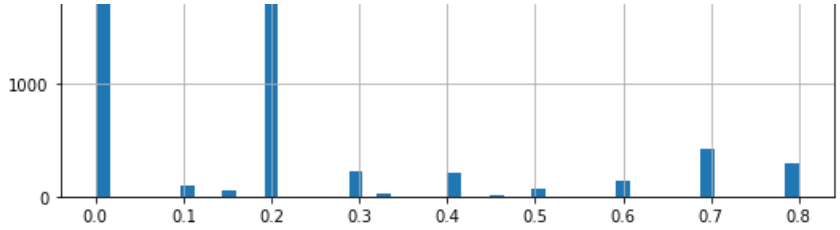
In [99]:
```
sample1.hist(bins=50,figsize=(20,15))
plt.show();
```

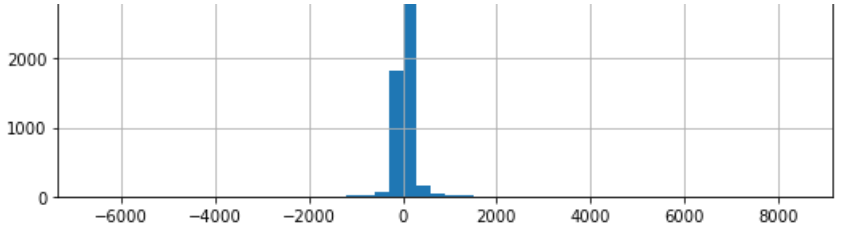












In [100…

```
# In this step we want to count total repaeatable state in our dataset.
sample1["State"].value_counts()
```

Out[100…

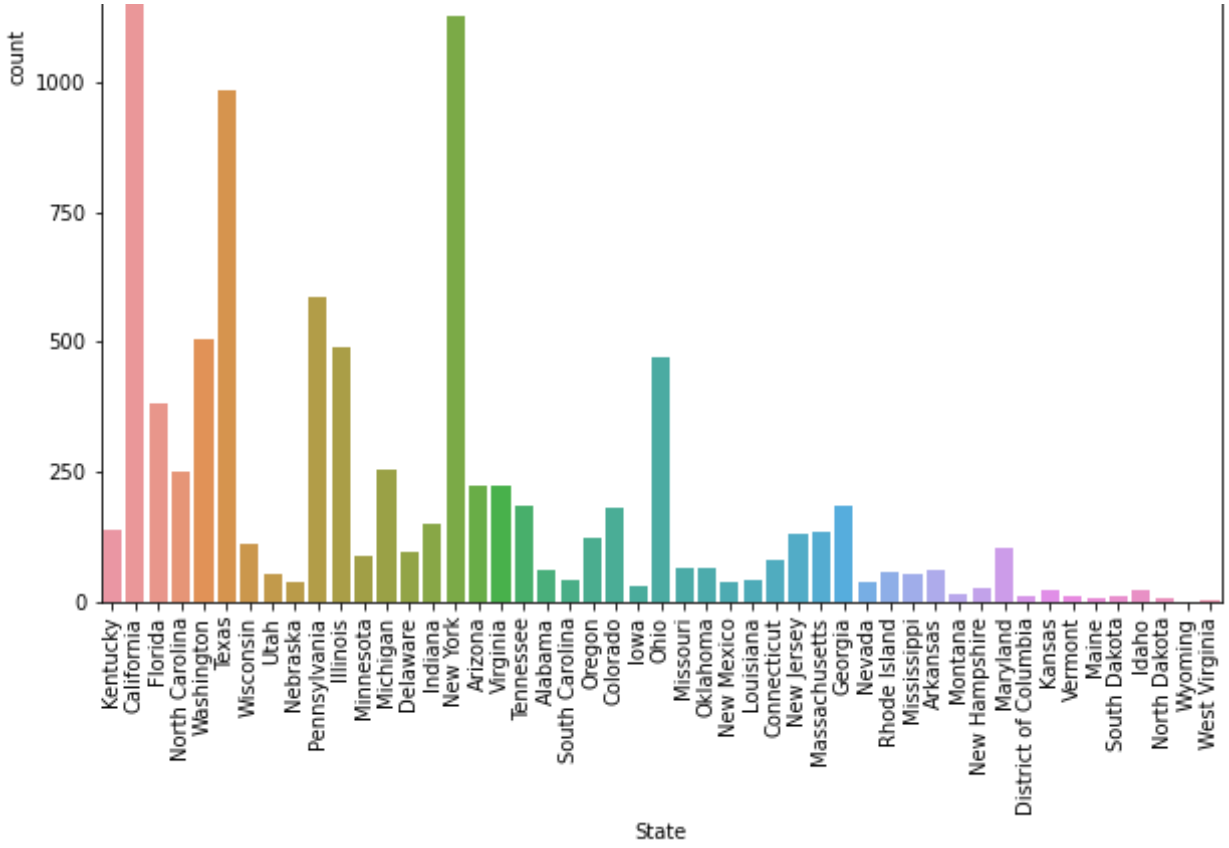
```
California              2001
New York                1128
Texas                    985
Pennsylvania             587
Washington               506
Illinois                 492
Ohio                     469
Florida                  383
Michigan                 255
North Carolina           249
Virginia                 224
Arizona                  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
South Carolina            42
```

```
Louisiana                  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
```

Step 9: In this step we will plot above state in the form graphical representation.

In [101…

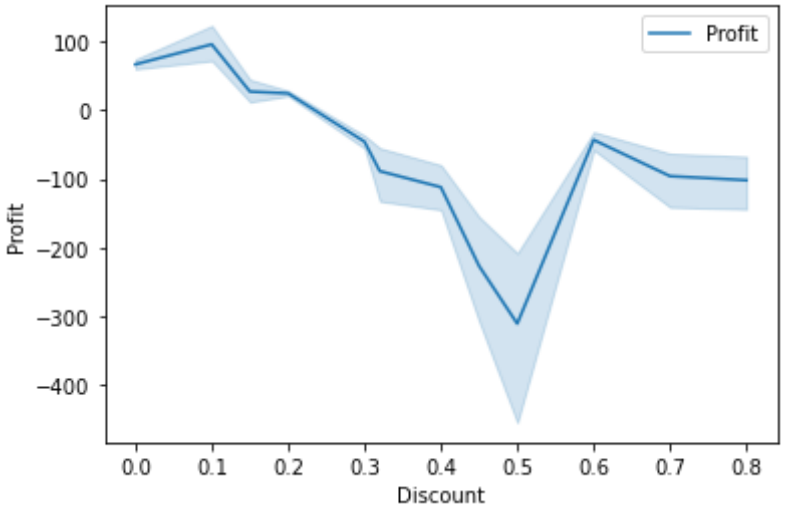
```
plt.figure(figsize=(10,10))
sns.countplot(x=sample1["State"])
plt.xticks(rotation=90)
plt.title("STATE")
plt.show()
```



Step 10: Profit and Discount Relation

In [102…

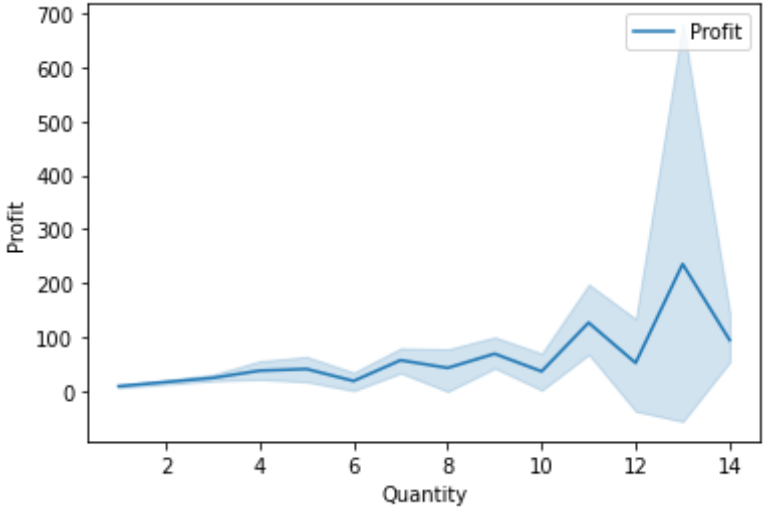
```
sns.lineplot(x='Discount',y='Profit',label='Profit',data=sample)
plt.legend()
plt.show()
```



Step 11: Profit and Quantity Relation

In [103…

```
sns.lineplot(x='Quantity',y='Profit',label='Profit',data=sample)
plt.legend()
plt.show()
```

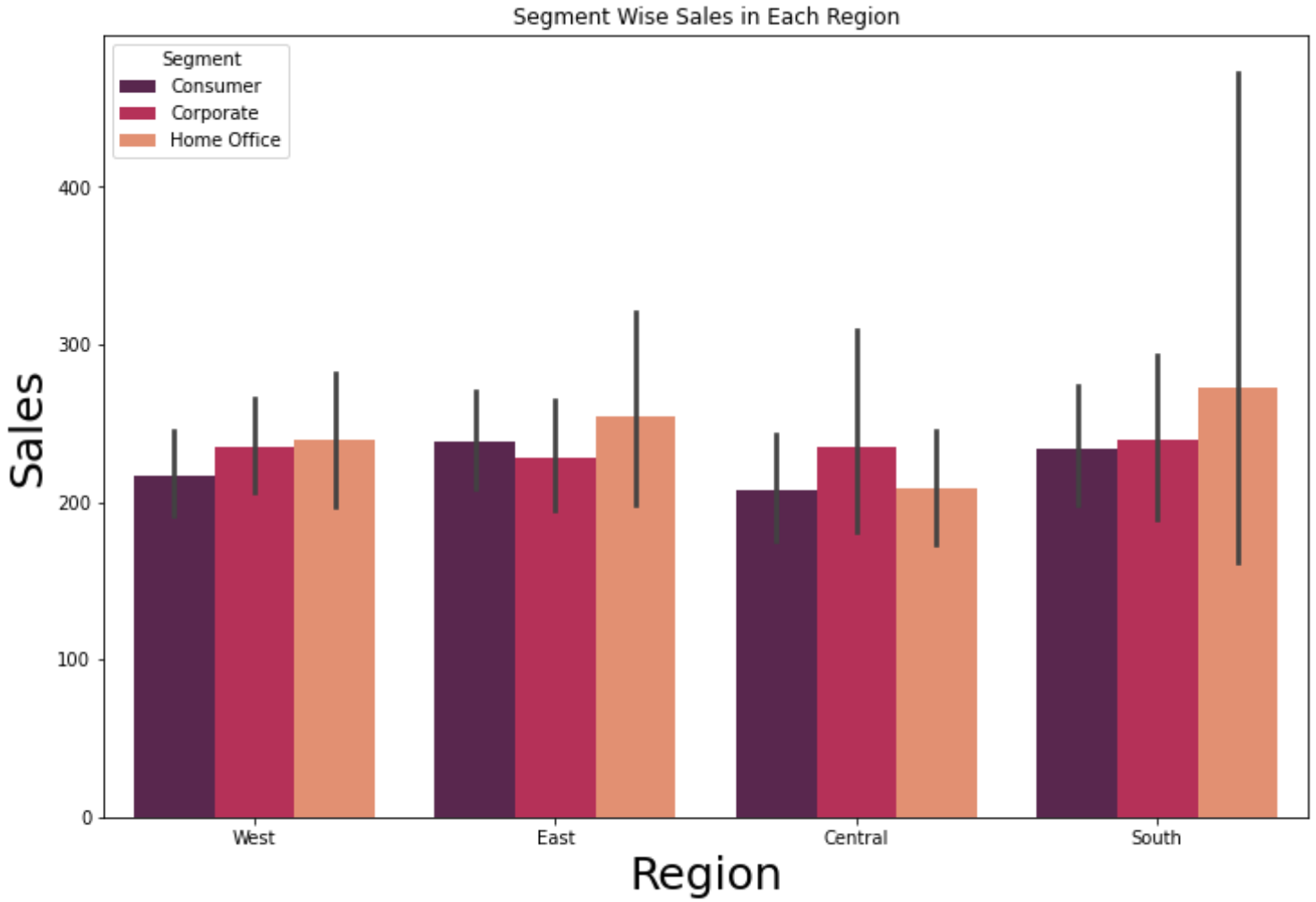


Higher the quantity higher the profit

Step 12:Segment Wise Sales in each Region

In [104…

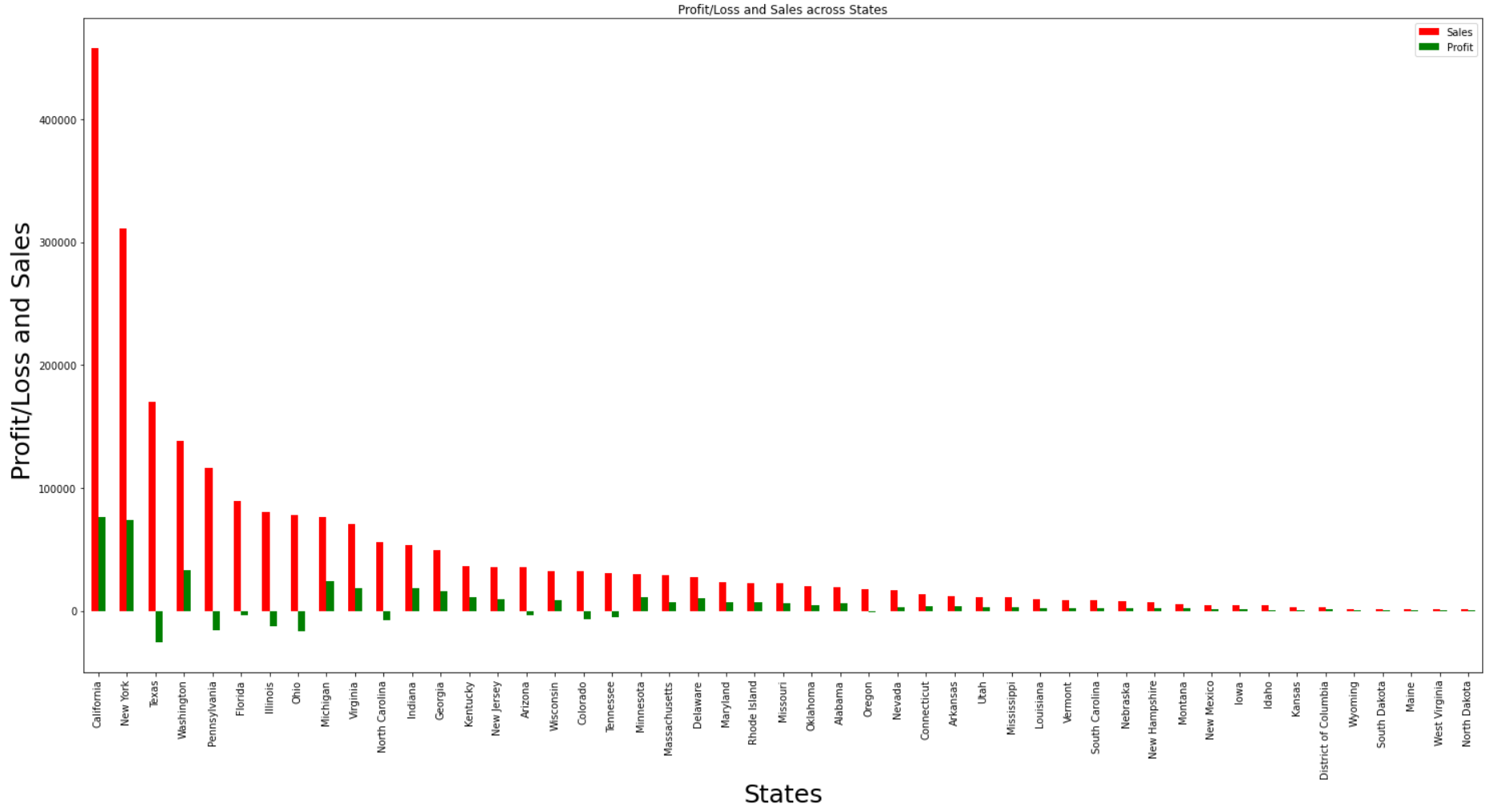
```
plt.figure(figsize=(12,8))
plt.title(" Segment Wise Sales in Each Region ")
sns.barplot(x='Region',y='Sales',data=sample,hue='Segment',order=sample['Region'].value_counts().index,palette='rocke
plt.xlabel('Region',fontsize=25)
plt.ylabel('Sales',fontsize=25)
plt.show()
```



Step 13:Profit/Loss and Sales by States

In [105…

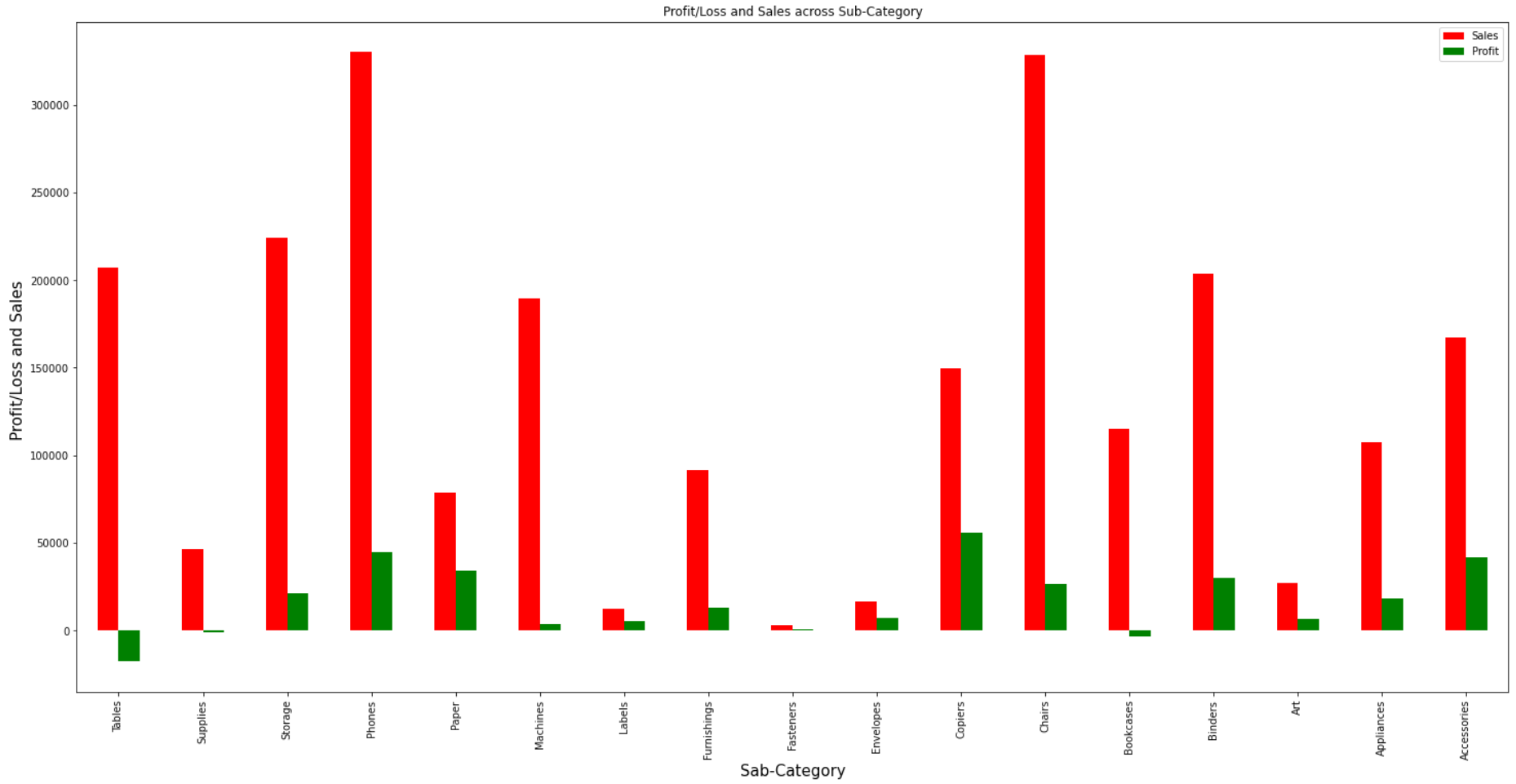
```
ps=sample.groupby('State')[['Sales','Profit']].sum().sort_values(by='Sales',ascending=False)
ps[:].plot.bar(color=['red','green'],figsize=(25,12))
plt.title('Profit/Loss and Sales across States')
plt.xlabel('States',fontsize=25)
plt.ylabel('Profit/Loss and Sales ',fontsize=25)
plt.show()
```



Step 14: Profit/Loss and Sales by Sub Category

In [106…

```
ps=sample.groupby('Sub-Category')[['Sales','Profit']].sum().sort_values(by='Sub-Category',ascending=False)
ps[:].plot.bar(color=['red','green'],figsize=(25,12))
plt.title('Profit/Loss and Sales across Sub-Category')
plt.xlabel('Sab-Category',fontsize=15)
plt.ylabel('Profit/Loss and Sales ',fontsize=15)
plt.show()
```



Conclusion Based on above plots:

1: Based on Category:- From Plot we can say that technology products are having higher sales and hgher profit compare to other category. for furniture products improvement is needed also try towards margin for the profit should increase.

2: Based upon segment:- sales and profit both are high in consumer segment so major focus sould be there to maintain it. and in home and office segment sales are low and due to that profit is also low so it should be improvewise with good streteg.

3: Based upon product:- Tables are facing loss as product so discount and offer shouls be optimised copies are good profit with excellent number of sales so it should be continue where sales is very low that produts are of office category so there is need of good stretegy.

Thnak You