Task:1 Exploratory Data Analysis (EDA) of Retail Domain.

we have downloaded the dataset of sample superstore from the link:https://bit.ly/3i4rbWl

Business Problem:

we have to perform exploratory data analysis on the dataset "SampleSuperStore and as a business manager, we have to find out the weak areas where we can work to make more profit and what are the business problem we can derive from the data.

importing all the important libraries

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

reading the dataset using pandas

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

Checking some values of the dataset using head and tell function

```
sample_superstore.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
```

sample_superstore.tail()

```
           Ship Mode   Segment        Country         City       State
\
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  \
9989        33180  South        Furniture  Furnishings   25.248
3
9990        92627   West        Furniture  Furnishings   91.960
2
9991        92627   West       Technology       Phones  258.576
2
9992        92627   West  Office Supplies        Paper   29.600
4
9993        92683   West  Office Supplies   Appliances  243.160
2

      Discount   Profit
9989       0.2   4.1028
9990       0.0  15.6332
9991       0.2  19.3932
9992       0.0  13.3200
9993       0.0  72.9480
```

checking the shape of the dataset

```
sample_superstore.shape
```

```
(9994, 13)
```

To see all the general information of the dataset like how many columns are there, what are the data type of those columns

```
sample_superstore.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
Ship Mode       9994 non-null object
Segment         9994 non-null object
Country         9994 non-null object
City            9994 non-null object
State           9994 non-null object
Postal Code     9994 non-null int64
Region          9994 non-null object
Category        9994 non-null object
Sub-Category    9994 non-null object
Sales           9994 non-null float64
Quantity        9994 non-null int64
Discount        9994 non-null float64
Profit          9994 non-null float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

check the number of uniques values in our dataset.

```
sample_superstore.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
```

check the correlation between the features

```
sample_superstore.corr()
```

```
             Postal Code     Sales  Quantity  Discount    Profit
Postal Code     1.000000 -0.023854  0.012761  0.058443 -0.029961
Sales          -0.023854  1.000000  0.200795 -0.028190  0.479064
Quantity        0.012761  0.200795  1.000000  0.008623  0.066253
Discount        0.058443 -0.028190  0.008623  1.000000 -0.219487
Profit         -0.029961  0.479064  0.066253 -0.219487  1.000000
```

```
corr = sample_superstore.corr()
sns.heatmap(corr)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x262a42c7b38>
```

from this above we can see that variable with itself has the correlation but with others, with someone it has positive correlation and with some others it has negative correlation.

finding the covariance of the dataset

```
sample_superstore.cov()
```

```
              Postal Code          Sales    Quantity    Discount  \
Postal Code  1.028080e+09 -476682.766590  910.415885  386.870404
Sales       -4.766828e+05  388434.455308  278.459923   -3.627228
Quantity     9.104159e+02     278.459923    4.951113    0.003961
Discount     3.868704e+02      -3.627228    0.003961    0.042622
Profit      -2.250458e+05   69944.096586   34.534769  -10.615173
```

```
                    Profit
Postal Code -225045.849445
Sales         69944.096586
Quantity         34.534769
Discount        -10.615173
Profit        54877.798055
```

checking for the null or NAN values

```
sample_superstore.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
```

From the above data, we can see that there is no null values in our dataset.

```
print("total number of null values =
",sample_superstore.isnull().sum().sum())
```

```
total number of null values =  0
```

Check for the duplicate values

```
sample_superstore[sample_superstore.duplicated(keep='last')]
```

```
           Ship Mode      Segment        Country           City
State  \
568   Standard Class    Corporate  United States        Seattle
Washington
591   Standard Class     Consumer  United States          Salem
Oregon
935   Standard Class  Home Office  United States   Philadelphia
Pennsylvania
1186  Standard Class    Corporate  United States        Seattle
Washington
1479  Standard Class     Consumer  United States  San Francisco
California
2803  Standard Class     Consumer  United States  San Francisco
```

```
California
2807    Second Class     Consumer  United States        Seattle
Washington
2836  Standard Class     Consumer  United States    Los Angeles
California
3127  Standard Class     Consumer  United States  New York City
New York
3405  Standard Class  Home Office  United States       Columbus
Ohio
3412  Standard Class    Corporate  United States  San Francisco
California
5372  Standard Class    Corporate  United States        Houston
Texas
5493        Same Day  Home Office  United States  San Francisco
California
6245  Standard Class  Home Office  United States        Seattle
Washington
6409     First Class     Consumer  United States        Houston
Texas
8457    Second Class    Corporate  United States        Chicago
Illinois
8533  Standard Class     Consumer  United States        Detroit
Michigan

      Postal Code   Region         Category Sub-Category    Sales
Quantity  \
568         98105     West  Office Supplies        Paper   19.440
3
591         97301     West  Office Supplies        Paper   10.368
2
935         19120     East  Office Supplies        Paper   15.552
3
1186        98103     West  Office Supplies        Paper   25.920
4
1479        94122     West  Office Supplies        Paper   25.920
4
2803        94122     West  Office Supplies        Paper   12.840
3
2807        98115     West  Office Supplies        Paper   12.960
2
2836        90036     West  Office Supplies        Paper   19.440
3
3127        10011     East  Office Supplies        Paper   49.120
4
3405        43229     East        Furniture       Chairs  281.372
2
3412        94122     West  Office Supplies          Art   11.760
4
5372        77041  Central  Office Supplies        Paper   15.552
3
```

```
5493         94122     West  Office Supplies       Labels   41.400
4
6245         98105     West        Furniture  Furnishings   22.140
3
6409         77041  Central  Office Supplies        Paper   47.952
3
8457         60653  Central  Office Supplies      Binders    3.564
3
8533         48227  Central        Furniture       Chairs  389.970
3

      Discount   Profit
568        0.0   9.3312
591        0.2   3.6288
935        0.2   5.4432
1186       0.0  12.4416
1479       0.0  12.4416
2803       0.0   5.7780
2807       0.0   6.2208
2836       0.0   9.3312
3127       0.0  23.0864
3405       0.3 -12.0588
3412       0.0   3.1752
5372       0.2   5.4432
5493       0.0  19.8720
6245       0.0   6.4206
6409       0.2  16.1838
8457       0.8  -6.2370
8533       0.0  35.0973
```

These above rows are the duplicate rows in our dataset and these duplicates values should be removed so we will drop these values.

checking the shape of the duplicated values dataframe

for this we can save the values in a variable and check the shape of that variable as you can see here we have given the name duplicate and then we are checking its shape.

```
duplicate = 
sample_superstore[sample_superstore.duplicated(keep='last')]
duplicate.shape
```

```
(17, 13)
```

dropping the duplicate values so that it can not affect our data.

```
sample_superstore.drop_duplicates(keep='last',inplace=True)
```

after dropping the duplicate values checking the shape of our remaining data.

```
sample_superstore.shape
```

```
(9977, 13)
```

check all the statistical features

```
sample_superstore.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
```

from this above we can see that till 75% it is acceptable but after the 75%, profit is increased repidly so we have to check after 75% that what is happening there and where is the sudden change in the data.

To see all the percentiles from 90 to 100 in the gap of 1%, to see where is the sudden change.

```
for i in range(90,101,1):
    print(np.percentile(sample_superstore.Profit,i))
```

```
89.3142
99.23
111.59100000000001
126.34005600000005
146.39486399999998
168.61271999999923
210.47357599999955
260.4091519999995
342.94862399999914
```

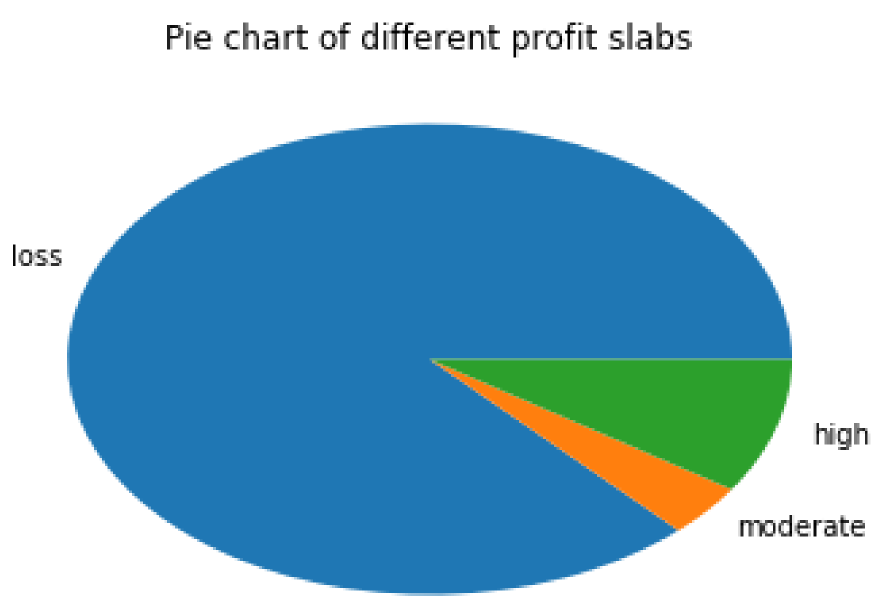
```
580.9456239999996
8399.976
```

so from this above we can see that profit is suddenly changing (increased with large profit) after 98 percentile and it becomes more.

Here we have categories our profit in three parts: LOSS, worst, high profit

```
high = 343
worst = 0
loss,moderate,high_profit = [],[],[]
for i in sample_superstore.Profit:
    if i in range(0,high):
        moderate.append(i)
    elif i<worst:
        loss.append(i)
    elif i>=high:
        high_profit.append(i)

#mask_shape=list(dict(df["mask_image_shapes"].value_counts()).keys())
#mask_shape_count=list(dict(df["mask_image_shapes"].value_counts()).va
lues())
plt.pie(x=[len(loss),len(moderate),len(high_profit)], labels=
['loss','moderate','high'])
plt.title("Pie chart of different profit slabs")
plt.show()
```



The above pie chart is of different profit slabs and from the above pie chart,we can see that loss length is more than the high and moderate because in this loss category the data comes which has profit less than the 0 and any profit that is less than the zero is not profit, it is loss actually.so these comes under loss category. and moderate category whose profit is between 0 to 343 means this is average and considerable profit and high comes under very large profit and in this the data comes it has large profit. so we can see why we are in more loss except getting profit. we can work upon those places where we are getting loss.

checking all the value counts for different-2 categorical features

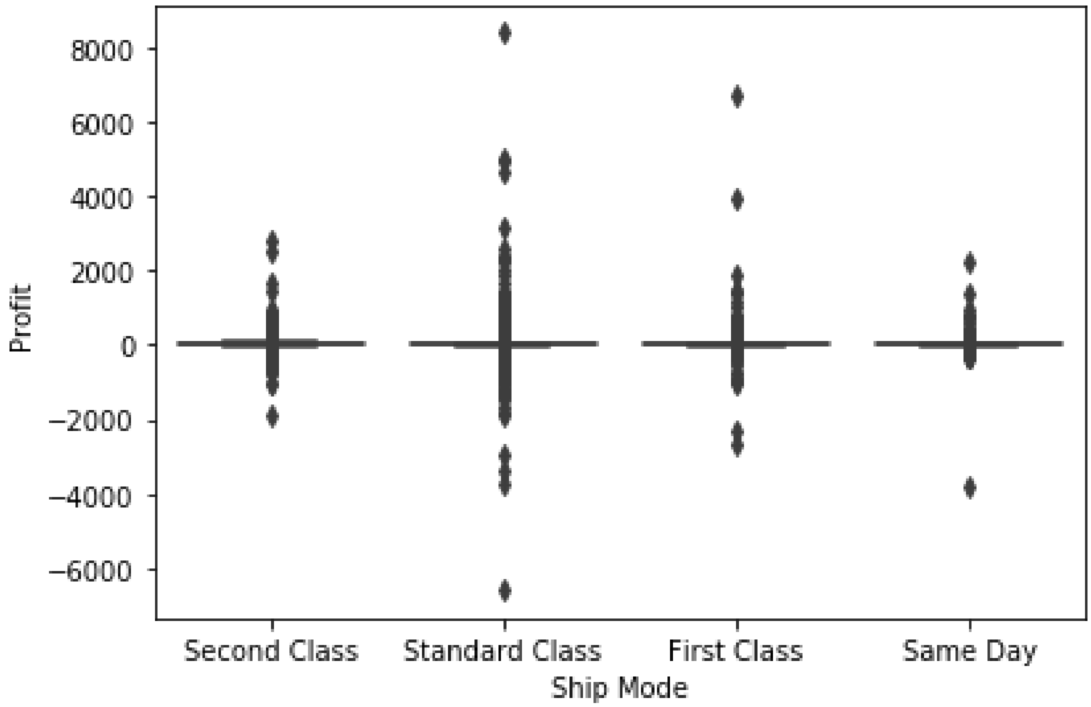
```
sample_superstore['Ship Mode'].value_counts()
```

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

from the above data we can see that ship mode by standard class is maximum and ship mode by same day is least so we see the same day shipment mode.

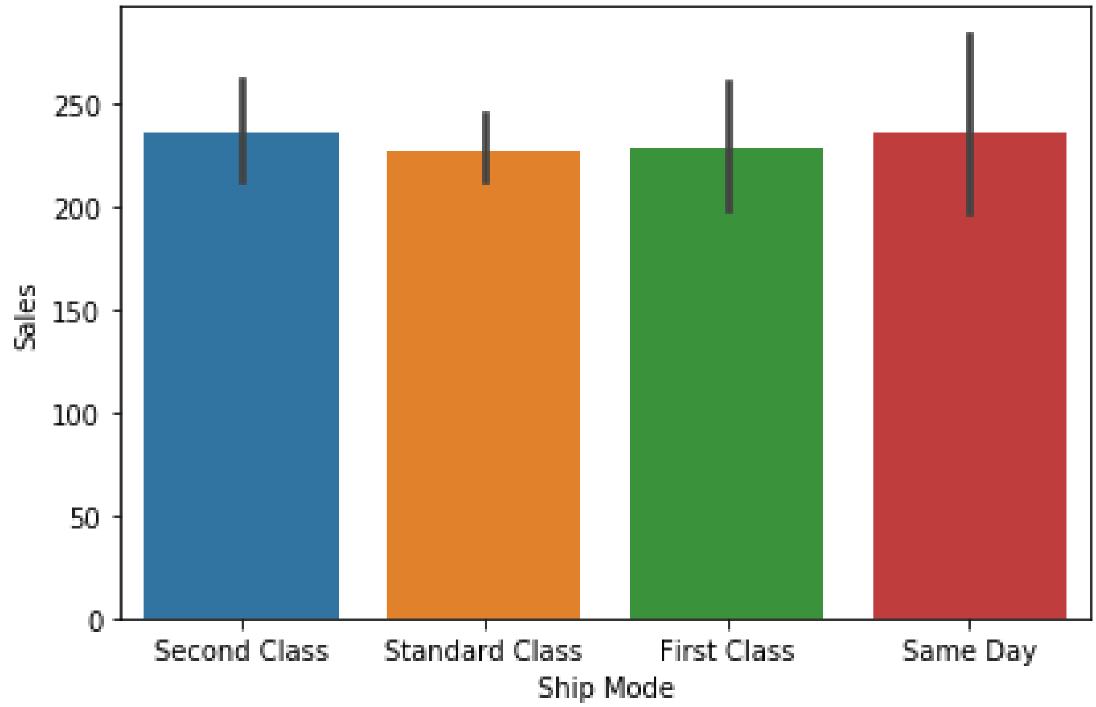
box plot between ship mode and profit

```
sns.boxplot(sample_superstore['Ship
Mode'],sample_superstore['Profit']);
plt.tight_layout()
plt.show()
```



looking at the above boxplot, we can say that the profit is maximum by first class ship mode and least profit is by standard class ship mode but sip mode by standard class is maximum till then profit is less so we can focus on this thing why the profit is less there and we can also check ship mode by same day and first class is less, even same day ship mode is least till then these are giving maximum profit so we can check the things there why it is happening.

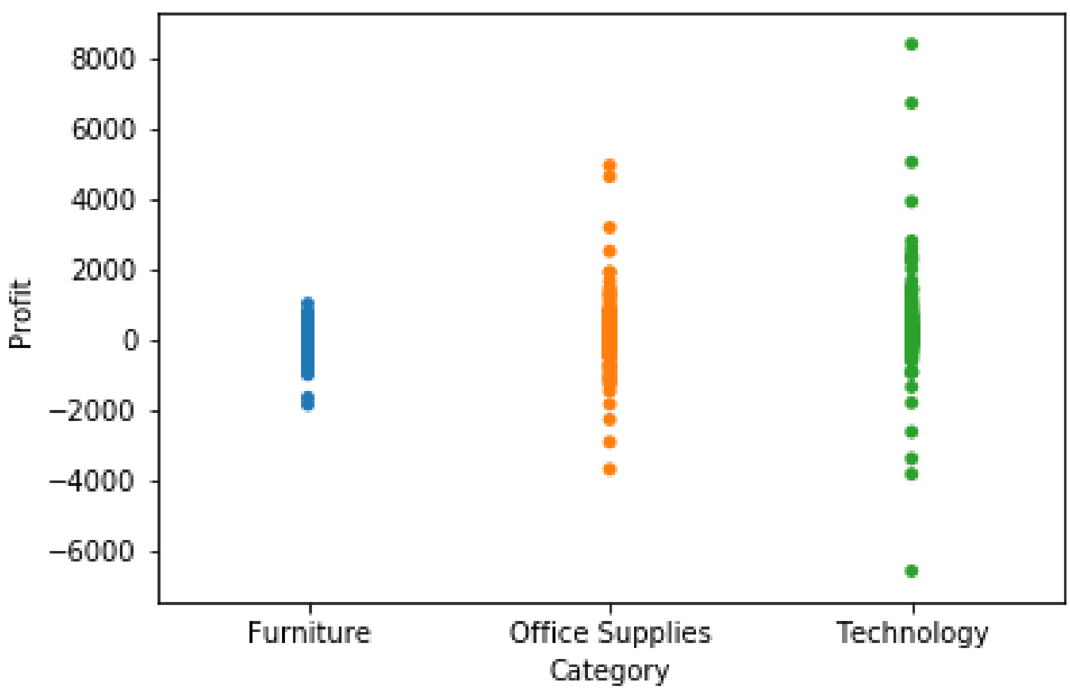
```
sns.barplot(sample_superstore['Ship Mode'],sample_superstore['Sales']);
plt.tight_layout()
plt.show()
```



looking above barplot, we can see that maximum sales is by the ship mode on the same day.

```
sns.stripplot(x="Category", y="Profit", data=sample_superstore)
```

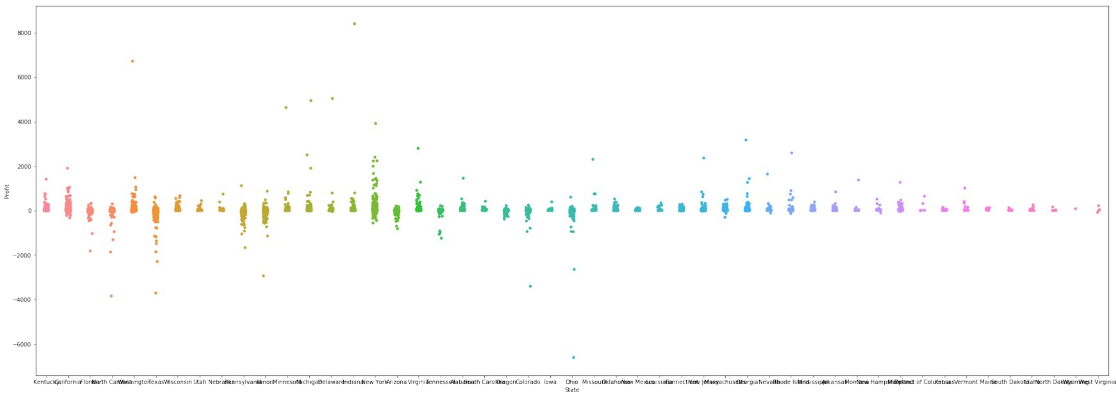
```
<matplotlib.axes._subplots.AxesSubplot at 0x1c3eefea128>
```



from this above plot we can see that we are getting more loss in comparision of Furniture and office Supplies category in the Technology category , we should work upon that part and for all the categories profit is very less it is almost constant for furniture category but for the officed supplies category it is increasing and going till 6000 but for the technology part it is sometime going till 8000 which is maximum.

```
plt.figure(figsize=(34,12))
sns.stripplot(x="State", y="Profit",jitter= True,
data=sample_superstore,orient = 'v')
#gfg.legend(fontsize=5)
#plt.setp(gfg.get_legend().get_title(), fontsize='20')
#plt.show()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x262a379c0b8>
```

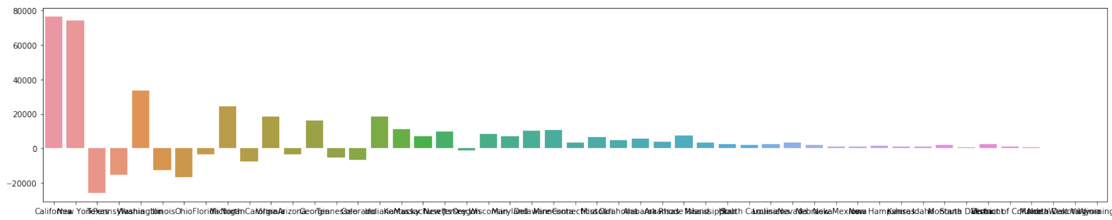


here from the above strip plot we can see that some states have loss so we should see there.

```
State = list(dict(sample_superstore['State'].value_counts()).keys())
#State_wise_profit=list(map(lambda s:
sum(list(sample_superstore[sample_superstore.State==s].Profit)),State)
)
state_wise_profit = []
for s in State :

state_wise_profit.append(sum(list(sample_superstore[sample_superstore.
State==s].Profit)))
```

```
plt.figure(figsize=(20,4))
sns.barplot(State,state_wise_profit)
plt.tight_layout()
plt.show()
```



check for value counts

```
sample_superstore['Region'].value_counts()
```

```
West       3203
East       2848
Central    2323
South      1620
Name: Region, dtype: int64
```

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

```
Office Supplies    6026
Furniture          2121
Technology         1847
Name: Category, dtype: int64
```

```
sample_superstore['Segment'].value_counts()
```

```
Consumer       5191
Corporate      3020
Home Office    1783
Name: Segment, dtype: int64
```

```
print(sample_superstore['State'].value_counts())
```

```
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
Missouri                 66
Oklahoma                 66
Alabama                  61
Arkansas                 60
Rhode Island             56
Mississippi              53
Utah                     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
```

```
sample_superstore['Country'].value_counts()
```

```
United States    9994
Name: Country, dtype: int64
```

here, from the above data we can see that our dataset has only data of one country so this feature is not more useful even we can remove this feature and it will not affect our data analysis so much.

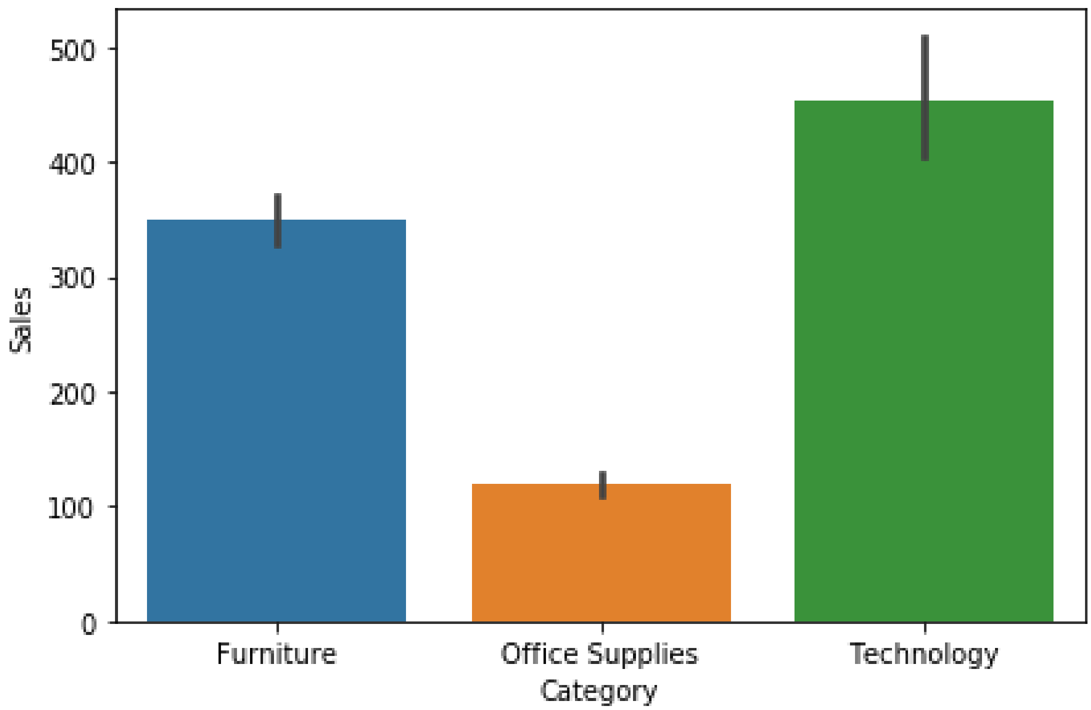
```
print(sample_superstore['City'].value_counts())
```

```
New York City    915
Los Angeles      747
Philadelphia     537
San Francisco    510
Seattle          428
Houston          377
Chicago          314
```

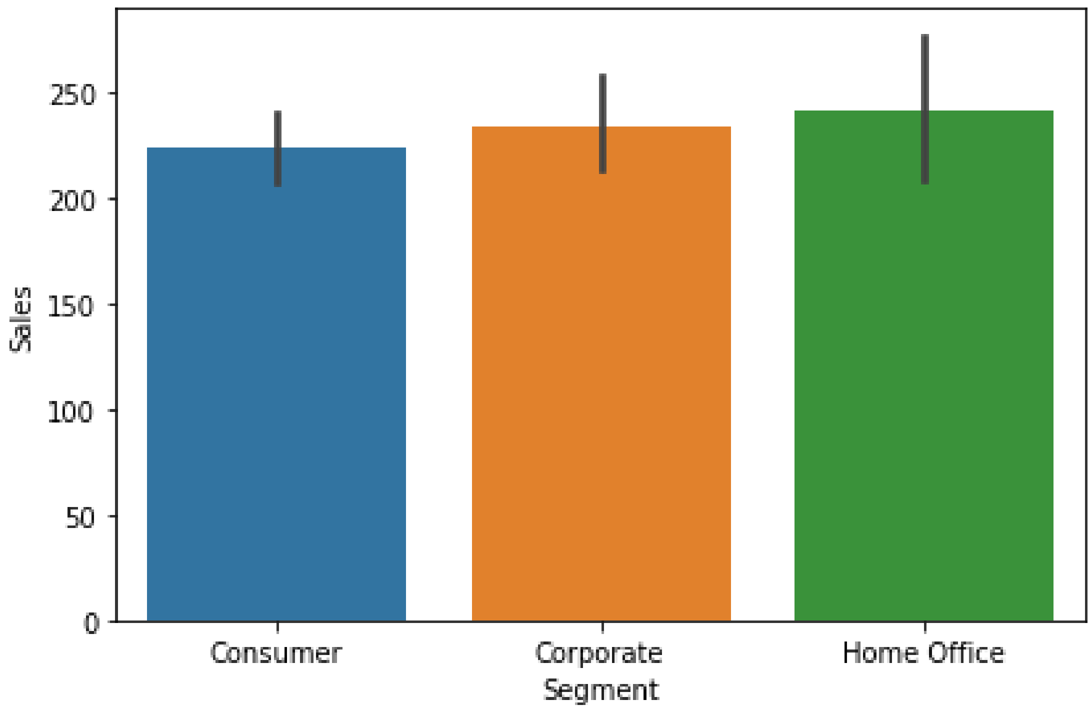
```
Columbus         222
San Diego        170
Springfield      163
Dallas           157
Jacksonville     125
Detroit          115
Newark            95
Richmond          90
Jackson           82
Columbia          81
Aurora            68
Phoenix           63
Long Beach        61
Arlington         60
San Antonio       59
Louisville        57
Miami             57
Rochester         53
Charlotte         52
Henderson         51
Lakewood          49
Lancaster         46
Fairfield         45
                ...
Vacaville          1
Rogers             1
Keller             1
Margate            1
Bartlett           1
Chapel Hill        1
Davis              1
Redding            1
Antioch            1
Cheyenne           1
Norfolk            1
Yucaipa            1
Atlantic City      1
Jupiter            1
La Quinta          1
Palatine           1
Normal             1
Port Orange        1
Littleton          1
Lake Elsinore      1
Lindenhurst        1
Melbourne          1
Commerce City      1
Whittier           1
Linden             1
Deer Park          1
```

```
Romeoville         1
Kissimmee          1
Abilene            1
Conroe             1
Name: City, Length: 531, dtype: int64
```

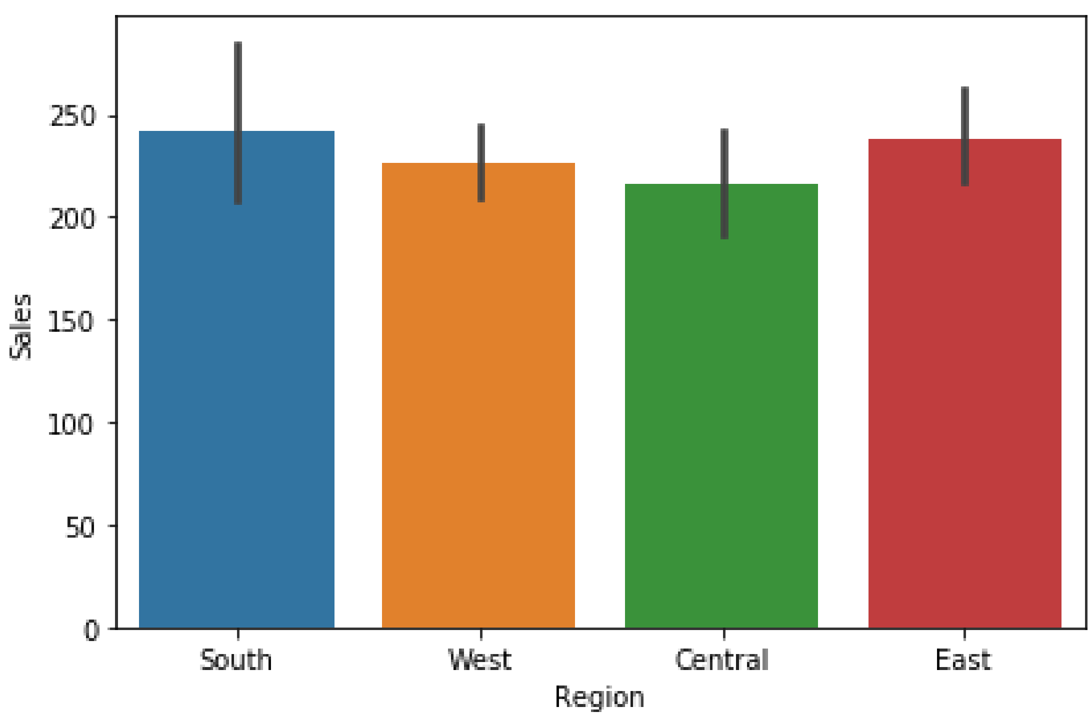
```
sns.barplot(sample_superstore['Category'],sample_superstore['Sales']);
plt.tight_layout()
plt.show()
```



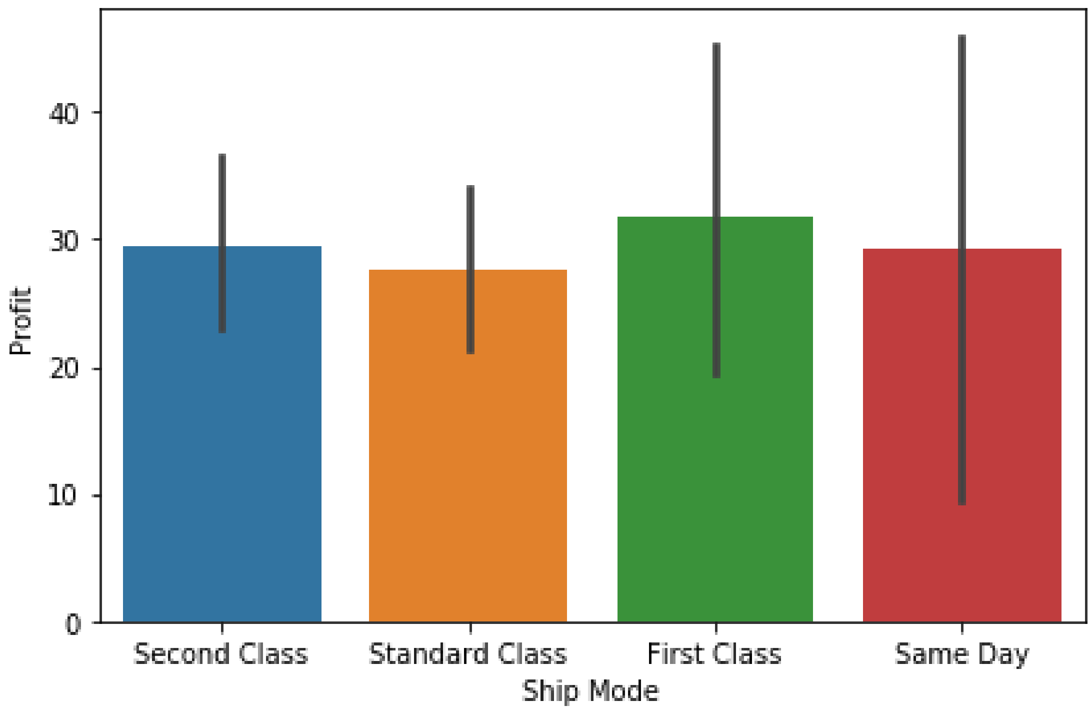

```
sns.barplot(sample_superstore['Segment'],sample_superstore['Sales']);
plt.tight_layout()
plt.show()
```



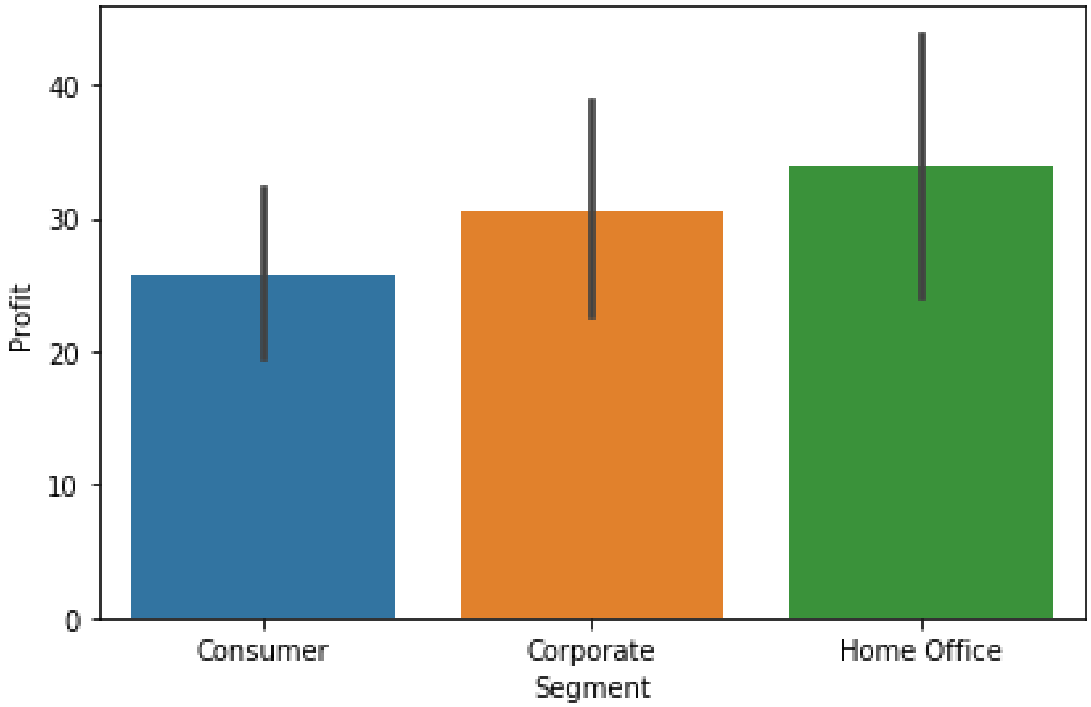

```
A = sample_superstore['Region']
sns.barplot(A,sample_superstore['Sales']);
plt.tight_layout()
plt.show()
```



```
sns.barplot(sample_superstore['Ship
Mode'],sample_superstore['Profit']);
plt.tight_layout()
plt.show()
```



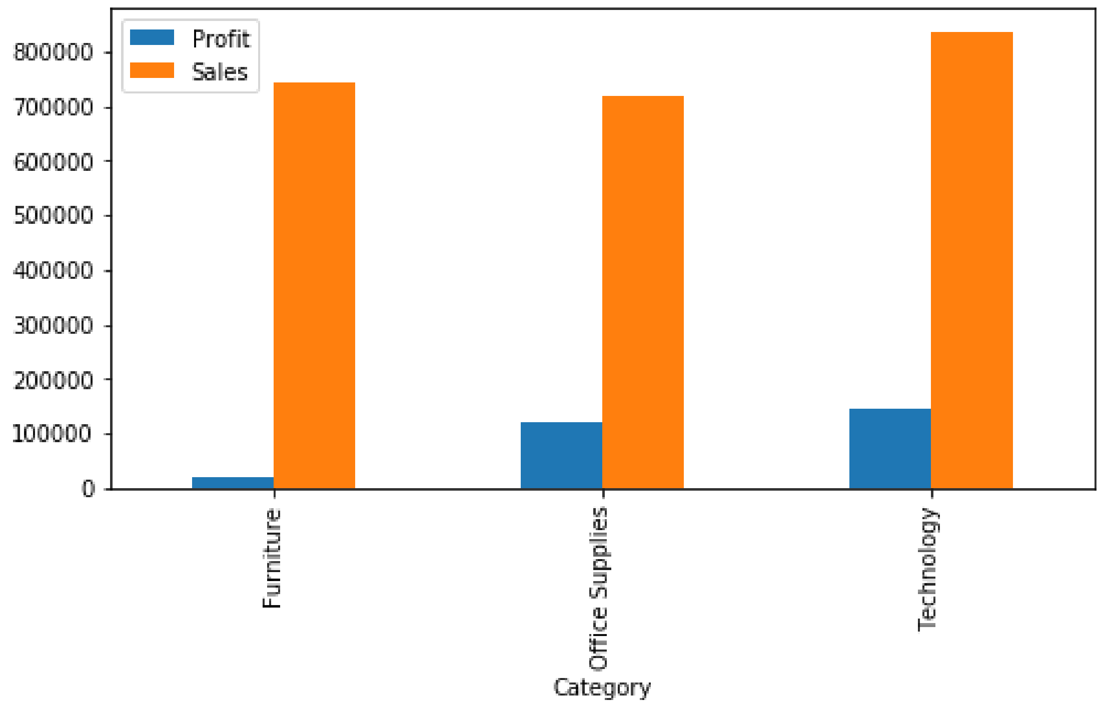

```
sns.barplot(sample_superstore['Segment'],sample_superstore['Profit']);
plt.tight_layout()
plt.show()
```



from this, we can see that profit is more from home office and less from consumer segment.so we should work on the consumer segment.

```
sample_superstore.groupby('Category')
['Profit','Sales'].agg(sum).plot(kind='bar',figsize=(8,4))
```

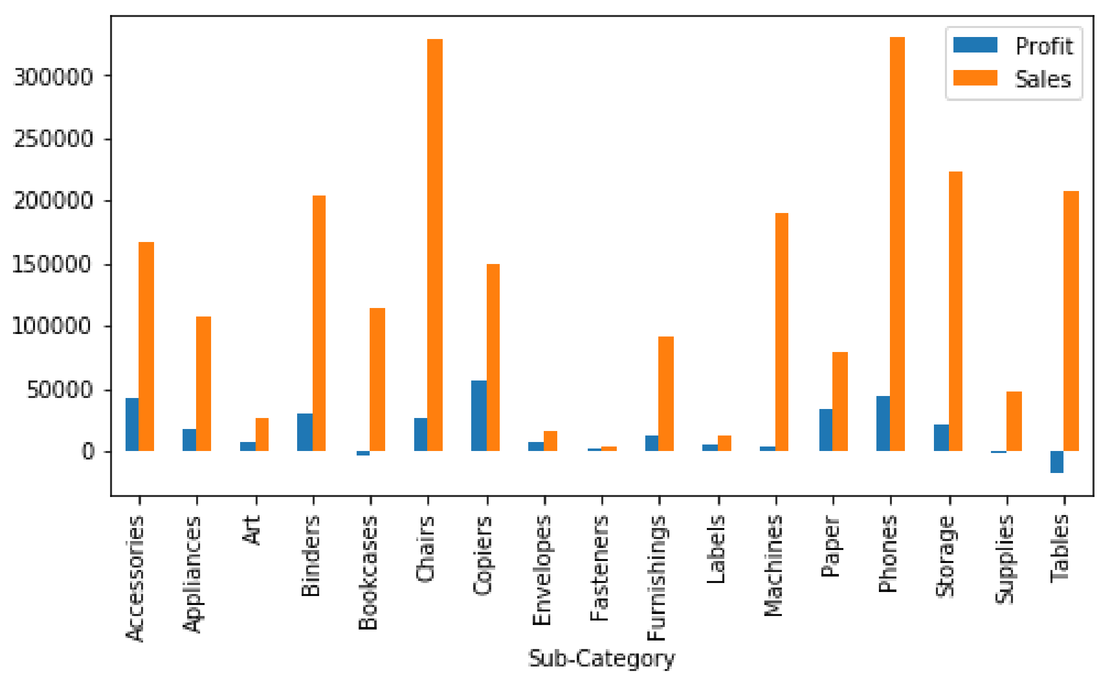
```
<matplotlib.axes._subplots.AxesSubplot at 0x2629f03b208>
```



from this above, we can see that profit is less in comprison of sales for furniture category.

```
sample_superstore.groupby('Sub-Category')
['Profit','Sales'].agg(sum).plot(kind='bar',figsize=(8,4))
```

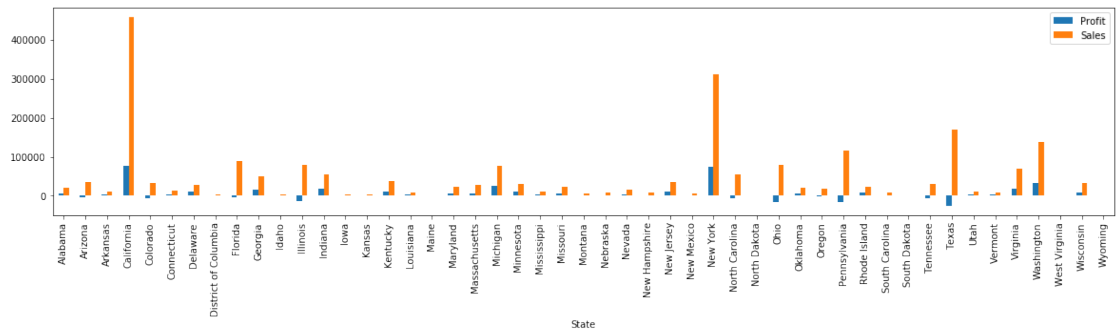
```
<matplotlib.axes._subplots.AxesSubplot at 0x2629eab9908>
```



from the above bar plot we can conclude that the sales of chairs and phones are maximum even they are giving very less profit so we can reduce the sales of those and there re some sub-categories like art,envelopes,fasteners,lables which has less sales even they are giving comparatevly good profit so we should increase the sales of these items and work upon them. also from this we can see that there is an item tables that has very good sales till then we are in loss, so we should check with this item why we are in loss or we should not sold that item.

```
sample_superstore.groupby('State')
['Profit','Sales'].agg(sum).plot(kind='bar',figsize=(20,4))
```

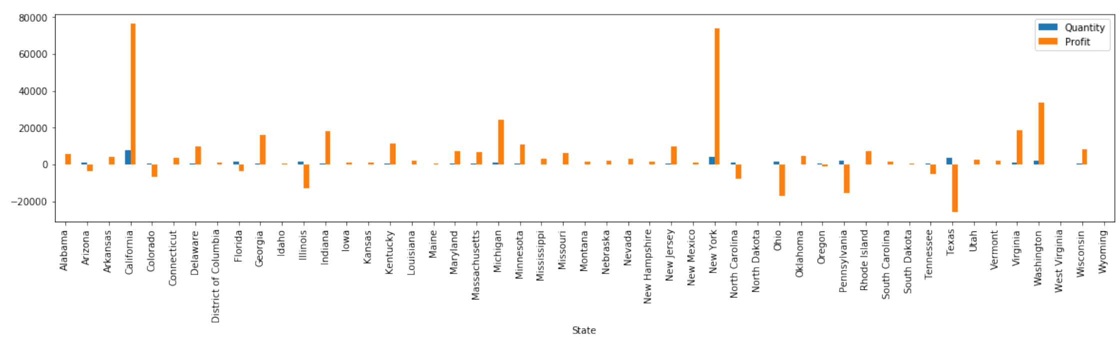
```
<matplotlib.axes._subplots.AxesSubplot at 0x262a0301e80>
```



from the above, we can see that there are some states where amount of sales is good but we are in loss so we should check there, what could be the problem and in the california sales is maximum even we re not getting so much profit in comparison of sales.

```
sample_superstore.groupby('State')
['Quantity','Profit'].agg(sum).plot(kind='bar',figsize=(20,4))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x262a4f52470>
```



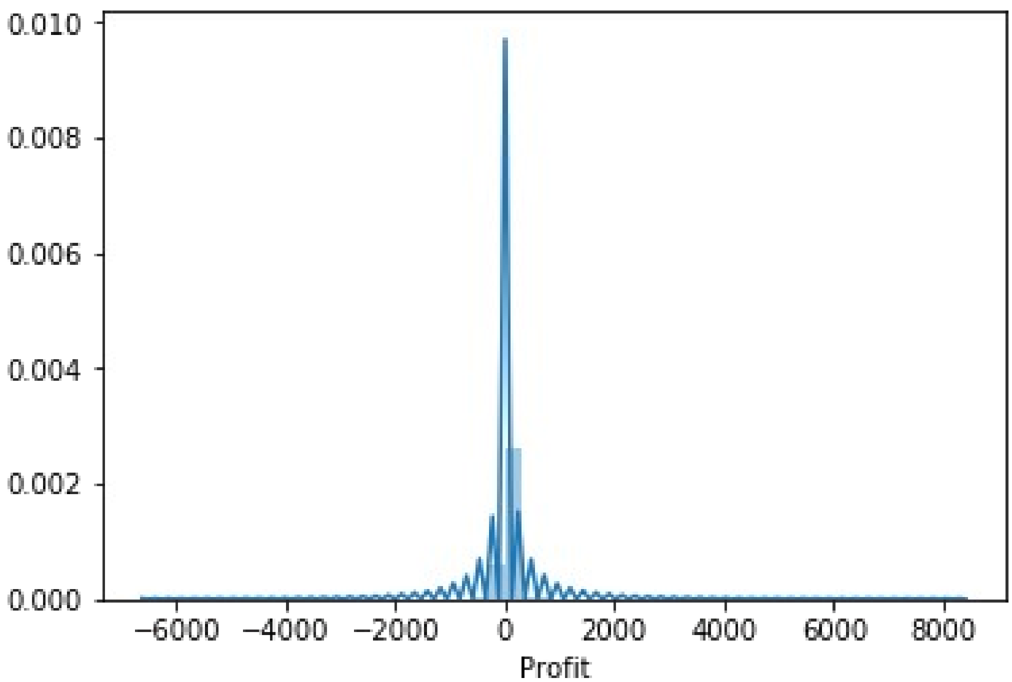
from the above bar plot, we can say that some states has sold some quantity but we are in loss in that states so we can work there. also we can see that some states has good profit in comparison of quantity sold so we can see what is happening there.

```
sns.distplot(sample_superstore.Profit)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\matplotlib\axes\
_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has
```

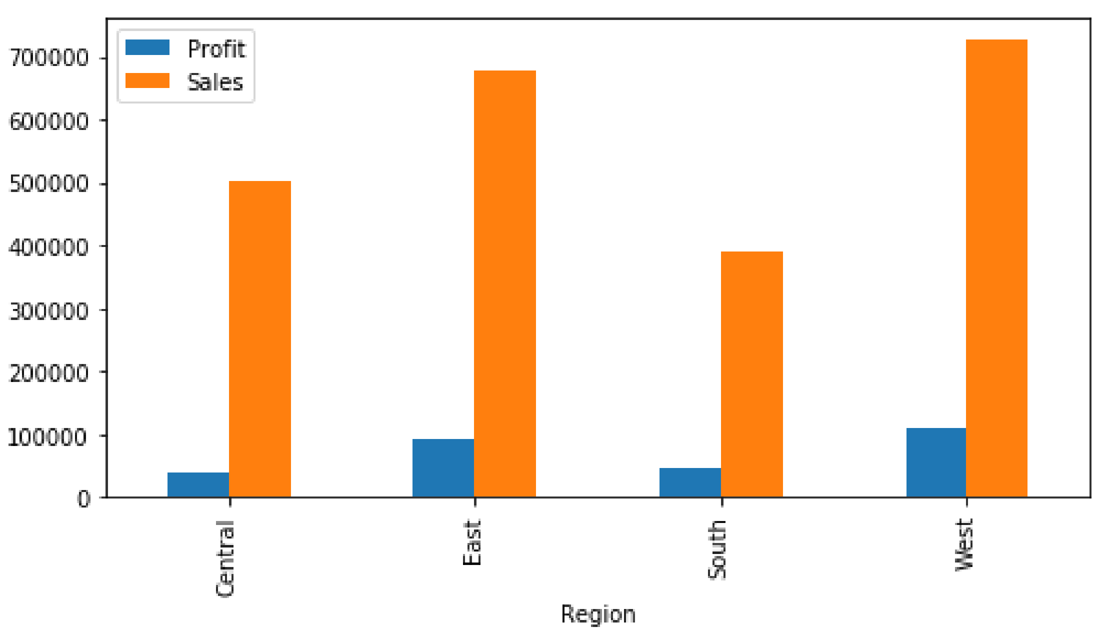
```
been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x262a4f32208>
```




```
sample_superstore.groupby('Region')
['Profit','Sales'].agg(sum).plot(kind='bar',figsize=(8,4))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x262a671ada0>
```



from the above, we cn conclude that profit is very less in comparison of sales, specially in the south and central region so we should work there to increase the profit.

Conclusions:-

1. profit is very less in comparison of sales, specially in the south and central region so we should work there to increase the profit.
1. there are some states where amount of sales is good but we are in loss so we should check there, what could be the problem and in the california sales is maximum even we re not getting so much profit in comparison of sales.
1. some states has sold some quantity but we are in loss in that states so we can work there. also we can see that some states has good profit in comparison of quantity sold so we can see what is happening there.
1. we can conclude that the sales of chairs and phones are maximum even they are giving very less profit so we can reduce the sales of those and there re some sub-categories like art,envelopes,fasteners,lables which has less sales even they are giving comparatevly good profit so we should increase the sales of these items and work upon them. also from this we can see that there is an item tables that has very good sales till then we are in loss, so we should check with this item why we are in loss or we should not sold that item.
1. we can see that profit is less in comprison of sales for furniture category.
1. profit is more from home office and less from consumer segment.so we should work on the consumer segment.
1. we are getting more loss in comparision of Furniture and office Supplies category in the Technology category , we should work upon that part and for all the categories profit is very less it is almost constant for furniture category but for the officed supplies category it is increasing and going till 6000 but for the technology part it is sometime going till 8000 which is maximum.
1. we can see that maximum sales is by the ship mode on the same day.
1. from this above we can see that variable with itself has the correlation but with others, with someone it has positive correlation and with some others it has negative correlation.