GRIP - The Sparks Foundation Data Science & Business Analytics Intern(Aug21)

Deepti Taram

Task3: Exploratory Data Analysis - Retail

To perform 'Exploratory Data Analysis' on dataset 'SampleSuperstore'

In [1]:
```
!pip install plotly --quiet
```

Importing the required libraries

In [3]:
```
#Importing all the required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
```

In [4]:
```
sample_df = pd.read_csv('SampleSuperstore.csv')
sample_df
```

Out[4]:

	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
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

9994 rows × 13 columns

In [5]:
```
sample_df.head()
```

Out[5]:

	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

In [6]:
```
sample_df.tail()
```

Out[6]:

	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 [7]:
```
sample_df.describe()
```

Out[7]:

	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 [8]:
```
sample_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 [9]:
```
sample_df.isnull().sum()
```

Out[9]:
```
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 [10]:
```
sample_df.columns
```

Out[10]:
```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity', 'Discount',
       'Profit'],
      dtype='object')
```

In [11]:
```
sample_df.duplicated().sum()
```

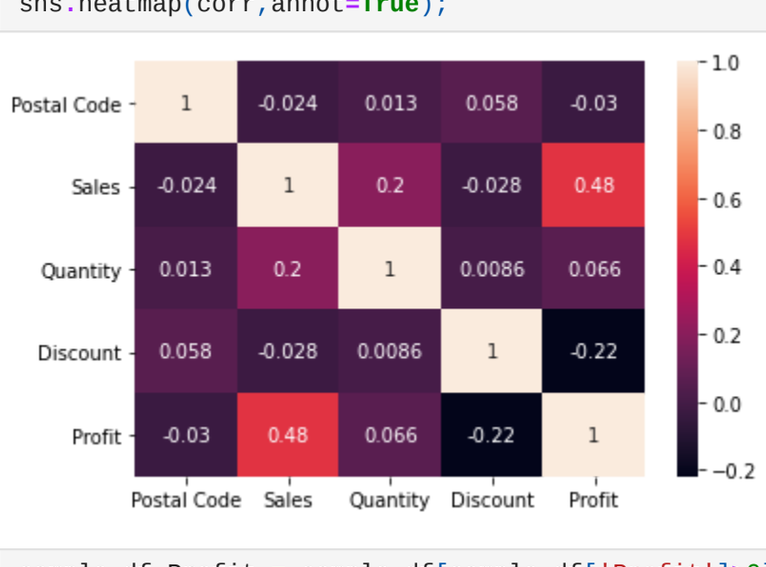
Out[11]: 17

In [12]:
```
sample_df.nunique()
```

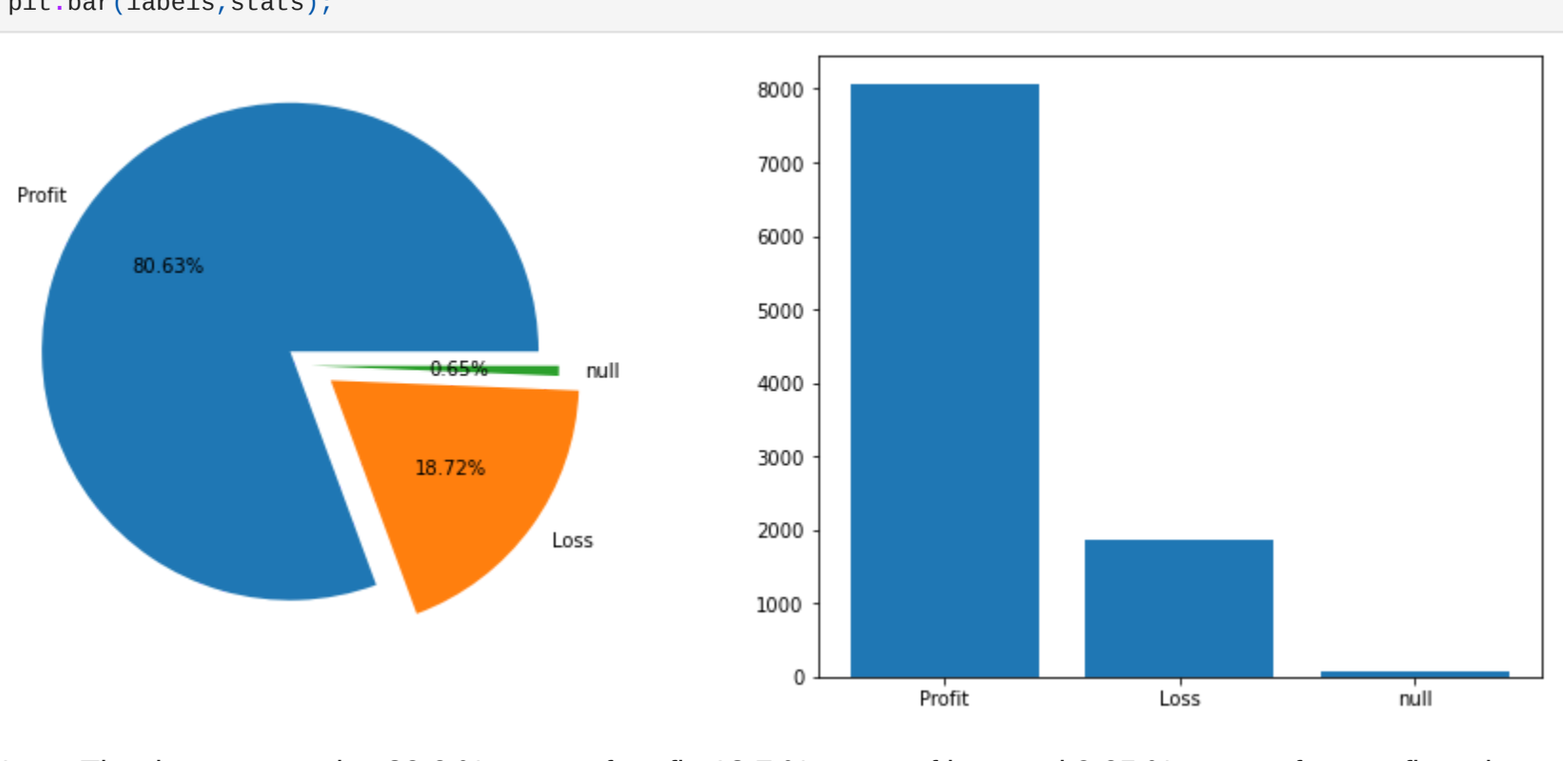
Out[12]:
```
Ship Mode          4
Segment            3
Country            1
City             531
State             49
Postal Code      631
Region             4
Category           3
Sub-Category      17
Sales           5825
Quantity          14
Discount          12
Profit          7287
dtype: int64
```

Exploratory Data Analysis

In [13]:
```
corr = sample_df.corr()
sns.heatmap(corr,annot=True);
```



In [14]:
```
sample_df_Profit = sample_df[sample_df['Profit']>0]
sample_df_loss = sample_df[sample_df['Profit']<0]
sample_df_null = sample_df[sample_df['Profit']==0]

stats = [sample_df_Profit.shape[0],sample_df_loss.shape[0],sample_df_null.shape[0]]
labels = ['Profit','Loss','null']

plt.figure(figsize=(15,6))
plt.subplot(121)
plt.pie(stats,labels=labels,explode=[0.1,0.1,0],autopct='%1.2f%%')
plt.subplot(122)
plt.bar(labels,stats);
```



Note : The dataset contains 80.6 % cases of profit, 18.7 % cases of loss and 0.65 % cases of no-profit-no-loss . From this, we can probably say that the Retail store is working fine

Let's add a new column which signifies if a particular row incurs profit or gain

In [15]:
```
x = []
for i in sample_df['Profit']:
  if (i>0):
   x.append('Gain')
  elif (i<0):
   x.append('Loss')
  else :
   x.append('null')

sample_df['Gain/Loss'] = x
sample_df.sample(2)
```

Out[15]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Gain/Loss
1581	First Class	Consumer	United States	New York City	New York	10024	East	Office Supplies	Storage	41.96	2	0.0	2.9372	Gain
3747	Second Class	Corporate	United States	Inglewood	California	90301	West	Furniture	Furnishings	198.46	2	0.0	99.2300	Gain

Examining features 'Ship Mode' and 'Segment'

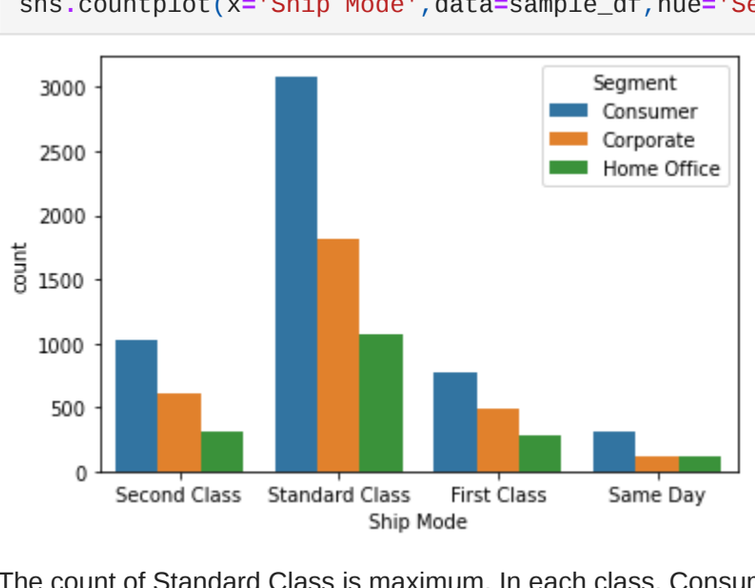
In [16]:
```
print('Number of classes : ',sample_df['Ship Mode'].nunique())
print('Different classes ', sample_df['Ship Mode'].unique())
```

```
Number of classes :  4
Different classes  ['Second Class' 'Standard Class' 'First Class' 'Same Day']
```

In [17]:
```
print('Number of classes ',sample_df['Segment'].nunique())
print('Different classes ',sample_df['Segment'].unique())
```

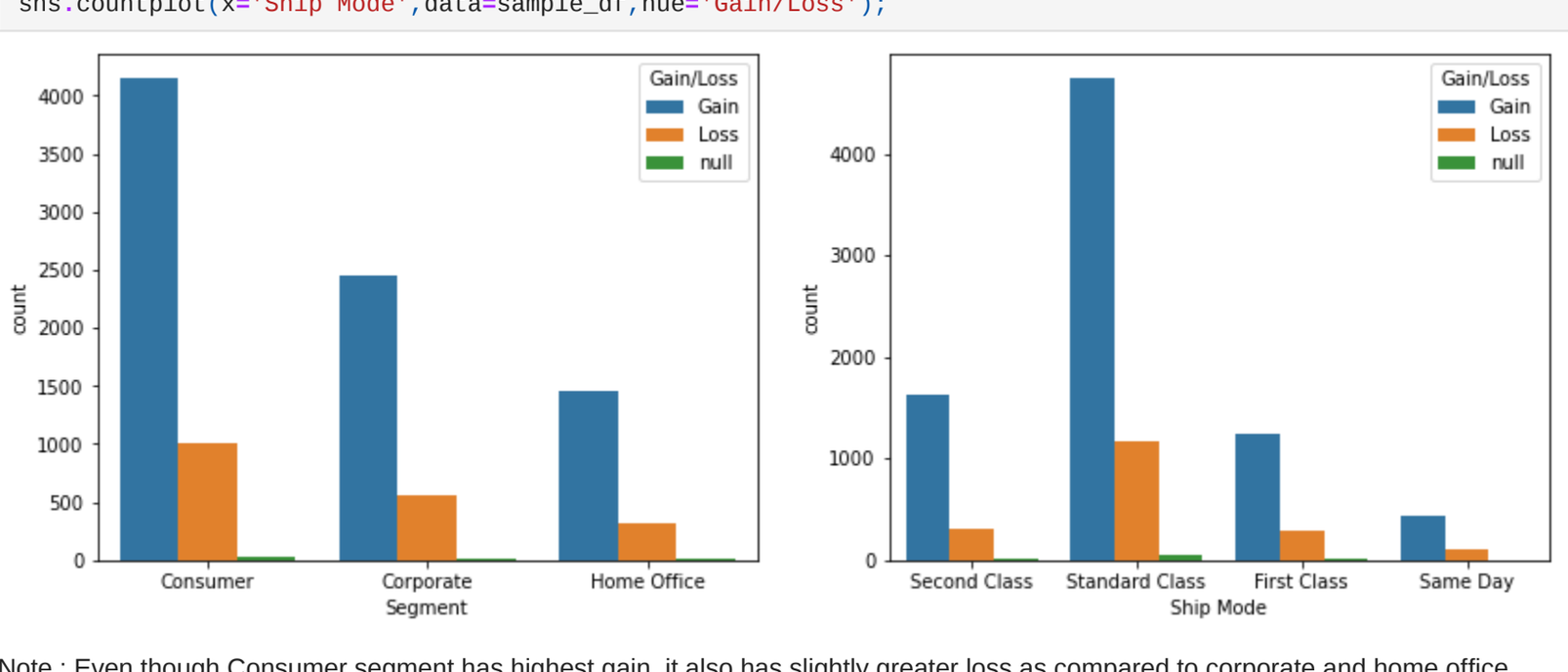
```
Number of classes  3
Different classes  ['Consumer' 'Corporate' 'Home Office']
```

In [18]:
```
sns.countplot(x='Ship Mode',data=sample_df,hue='Segment');
```



The count of Standard Class is maximum. In each class, Consumer segment holds the majority. We may conclude that the stores prefer Consumer Segment as it may draw more profit

In [19]:
```
plt.figure(figsize=(14,5))
plt.subplot(121)
sns.countplot(x='Segment',data=sample_df,hue='Gain/Loss')
plt.subplot(122)
sns.countplot(x='Ship Mode',data=sample_df,hue='Gain/Loss');
```

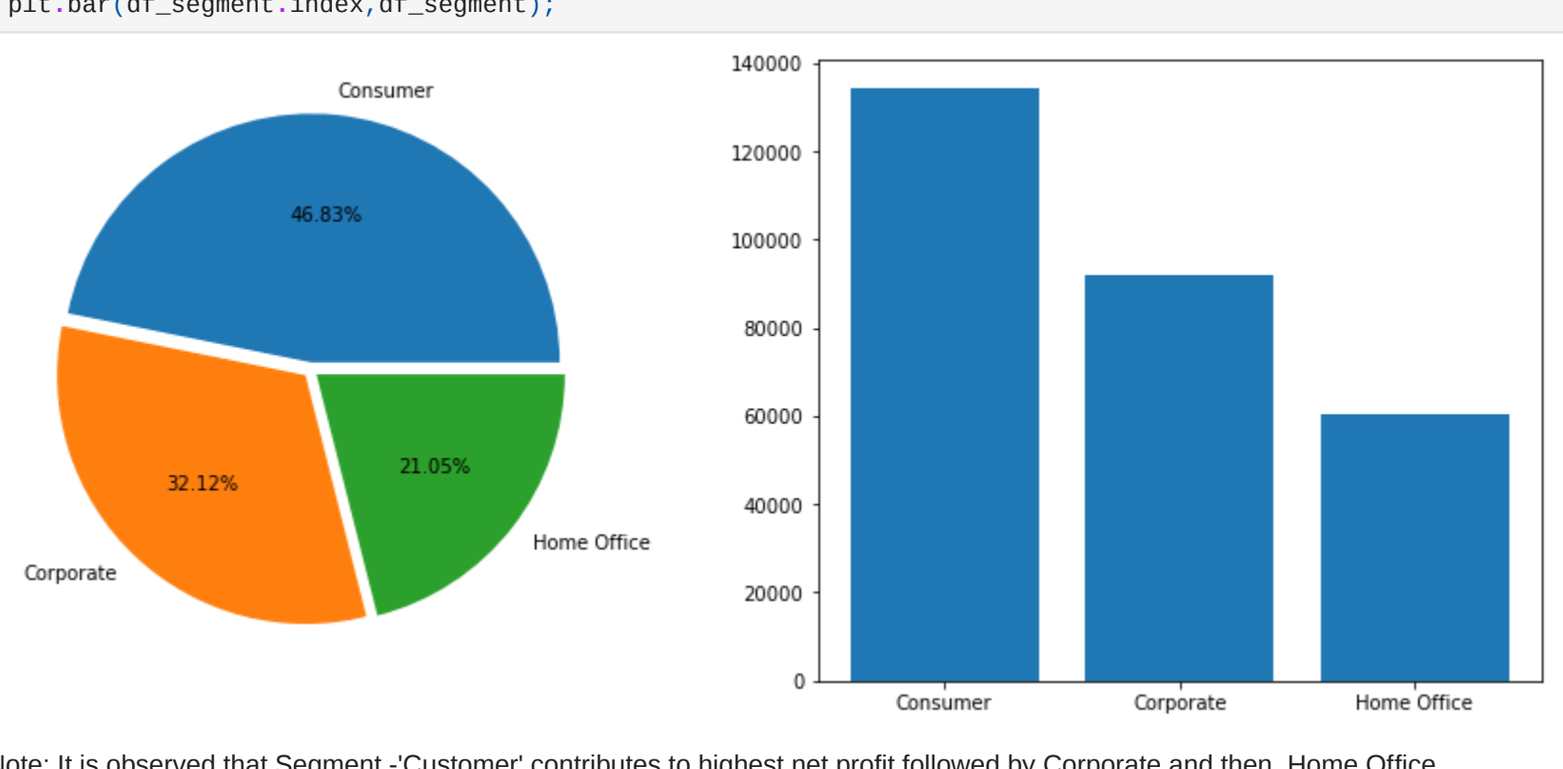


Note : Even though Consumer segment has highest gain, it also has slightly greater loss as compared to corporate and home office.

In [20]:
```
# Analyzing net Profits based on feature 'Segment'.

df_segment = sample_df.groupby('Segment').Profit.sum()

print('NET PROFIT IN EACH SEGMENT')
print(df_segment)
```

```
NET PROFIT IN EACH SEGMENT
Segment
Consumer       134119.2092
Corporate       91979.1340
Home Office     60298.6785
Name: Profit, dtype: float64
```

In [21]:
```
# Visualizing net Profits based on feature 'Segment'

plt.figure(figsize=(15,6))
plt.subplot(121)
plt.pie(x=df_segment,labels=df_segment.index,autopct='%1.2f%%',explode=[0.03,0.03,0.03])
plt.subplot(122)
plt.bar(df_segment.index,df_segment);
```

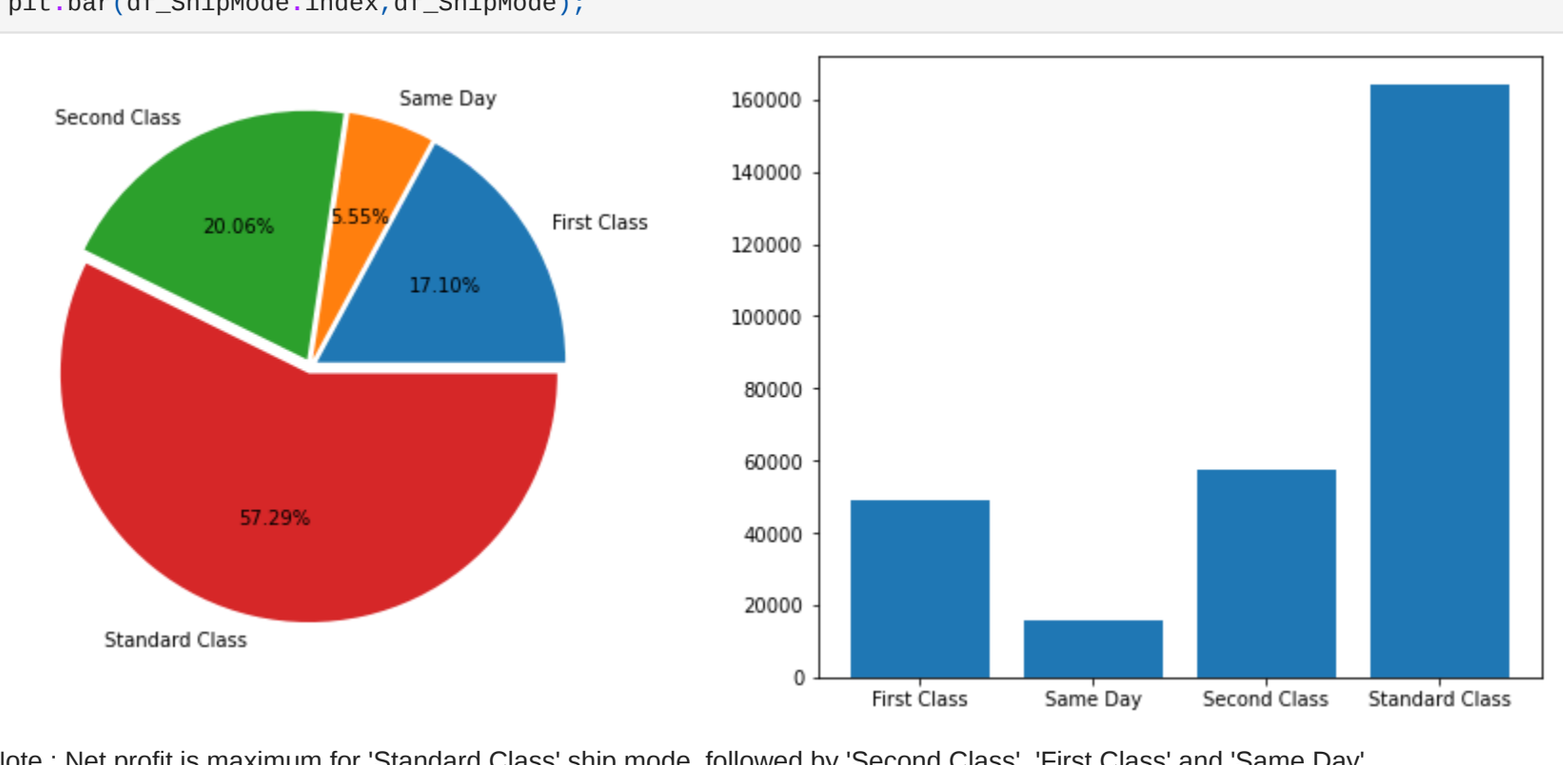


Note: It is observed that Segment -'Customer' contributes to highest net profit followed by Corporate and then, Home Office

In [22]:
```
# Analyzing net Profits based on feature 'hip ModSe'.

df_ShipMode = sample_df.groupby('Ship Mode').Profit.sum()

print('NET PROFIT IN EACH SHIP MODE')
print(df_ShipMode)
```

```
NET PROFIT IN EACH SHIP MODE
Ship Mode
First Class        48969.8399
Same Day           15891.7589
Second Class       57446.6354
Standard Class    164088.7875
Name: Profit, dtype: float64
```

In [23]:
```
# Visualizing net Profits based on feature 'Ship Mode'

plt.figure(figsize=(15,6))
plt.subplot(121)
plt.pie(x=df_ShipMode,labels=df_ShipMode.index,autopct='%1.2f%%',explode=[0.03,0.03,0.03,0.03])
plt.subplot(122)
plt.bar(df_ShipMode.index,df_ShipMode);
```



Note : Net profit is maximum for 'Standard Class' ship mode, followed by 'Second Class', 'First Class' and 'Same Day',

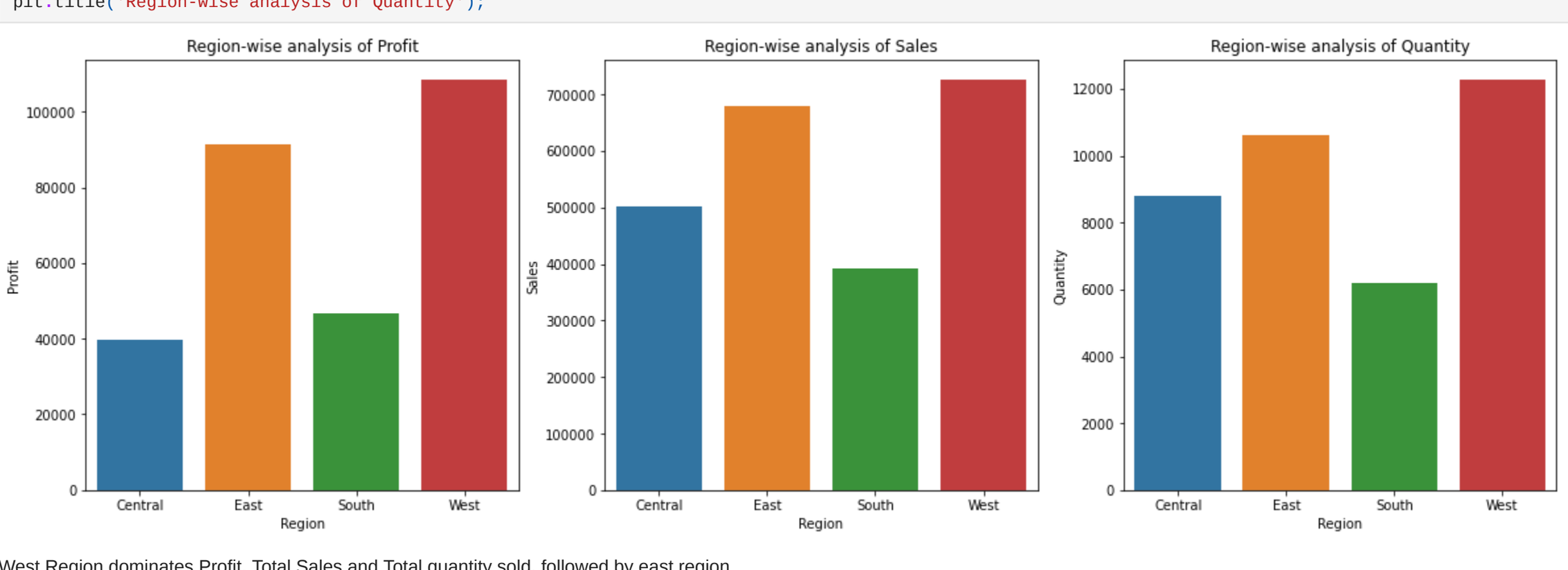
Analyzing feature 'Regions'

In [24]:
```
df_Region = pd.DataFrame(sample_df.groupby('Region').sum())
df_Region
```

Out[24]:

	Postal Code	Sales	Quantity	Discount	Profit
Region					
Central	151786150	501239.8908	8780	558.34	39706.3625
East	50171698	678781.2400	10618	414.00	91522.7800
South	55875052	391721.9050	6209	238.55	46749.4303
West	293739752	725457.8245	12266	350.20	108418.4489

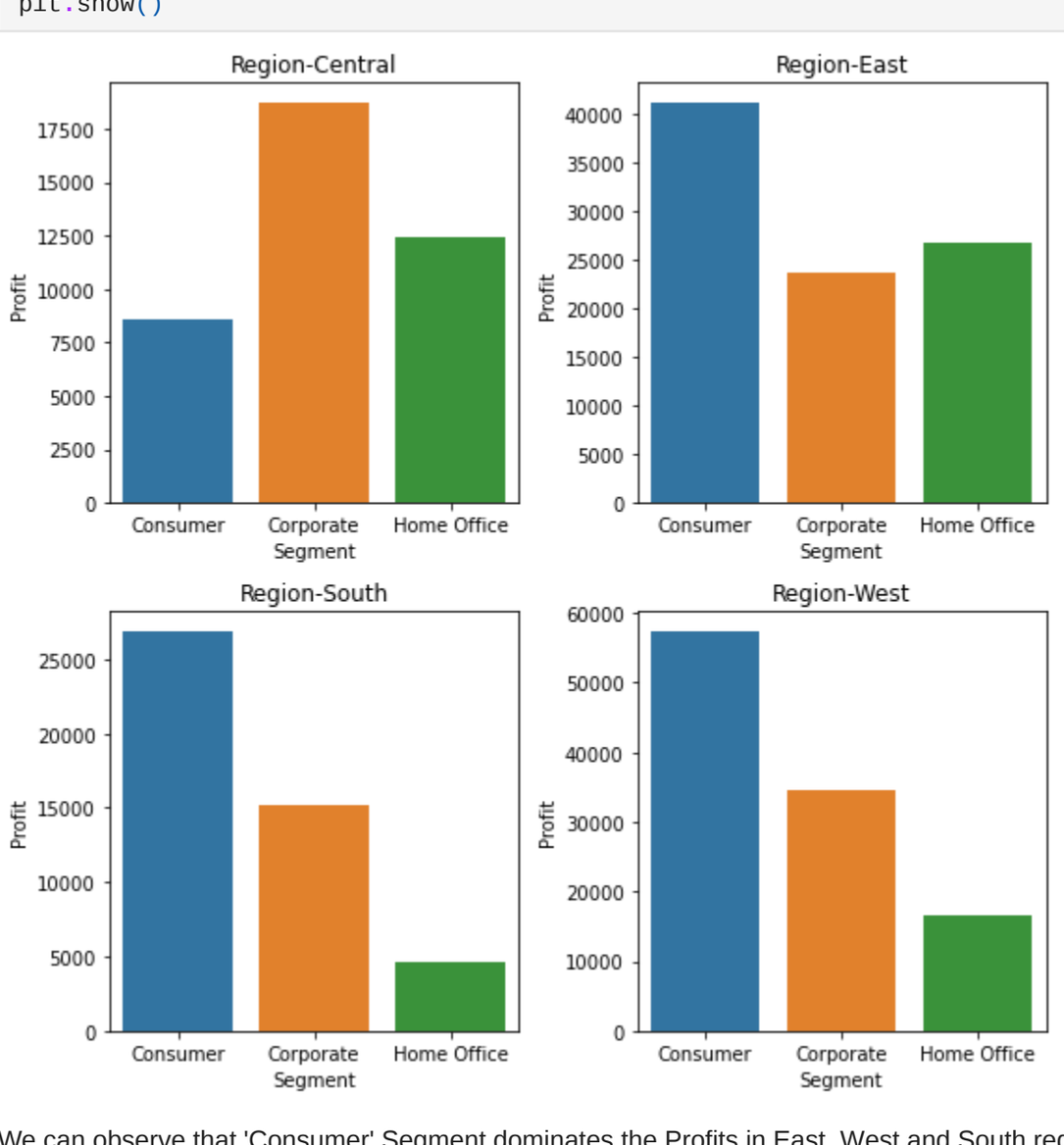
In [25]:
```
plt.figure(figsize=(20,6))
plt.subplot(131)
sns.barplot(x=df_Region.index,y='Profit',data=df_Region)
plt.title('Region-wise analysis of Profit')
plt.subplot(132)
sns.barplot(x=df_Region.index,y='Sales',data=df_Region)
plt.title('Region-wise analysis of Sales')
plt.subplot(133)
sns.barplot(x=df_Region.index,y='Quantity',data=df_Region)
plt.title('Region-wise analysis of Quantity');
```



West Region dominates Profit, Total Sales and Total quantity sold, followed by east region

In [26]:
```
central_profit = sample_df[sample_df['Region']=='Central'].groupby('Segment').sum()
east_profit = sample_df[sample_df['Region']=='East'].groupby('Segment').sum()
south_profit = sample_df[sample_df['Region']=='South'].groupby('Segment').sum()
west_profit = sample_df[sample_df['Region']=='West'].groupby('Segment').sum()
```

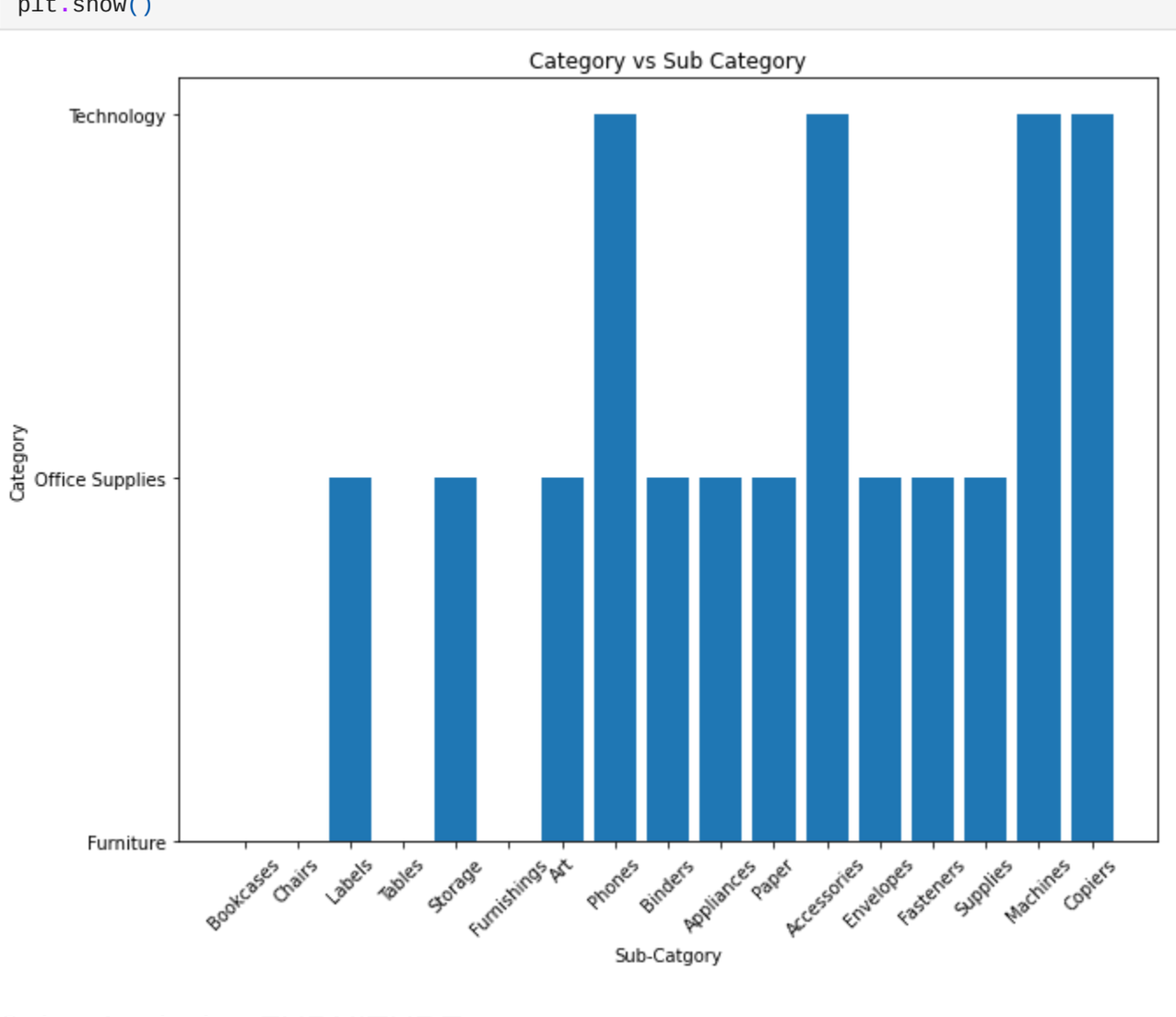
In [27]:
```
# Region-wise contribution to profit by Segment feature

plt.figure(figsize=(8,8))

plt.subplot(2,2,1)
sns.barplot(x=central_profit.index,y=central_profit['Profit'],data=central_profit)
plt.title('Region-Central')

plt.subplot(2,2,2)
sns.barplot(x=east_profit.index,y=east_profit['Profit'],data=east_profit)
plt.title('Region-East')

plt.subplot(2,2,3)
sns.barplot(x=south_profit.index,y=south_profit['Profit'],data=south_profit)
plt.title('Region-South')

plt.subplot(2,2,4)
sns.barplot(x=west_profit.index,y=west_profit['Profit'],data=west_profit)
plt.title('Region-West')

plt.tight_layout()
plt.show()
```



We can observe that 'Consumer' Segment dominates the Profits in East, West and South regions whereas 'Corporate' Segment dominates the Profits in Central region.

Analyzing features 'Categories' and 'Sub-Categories'

In [28]:
```
plt.figure(figsize=(10,8))
plt.bar('Sub-Category','Category', data=sample_df)
plt.title('Category vs Sub Category')
plt.xlabel('Sub-Catgory')
plt.ylabel('Category')
plt.xticks(rotation=45)
plt.show()
```



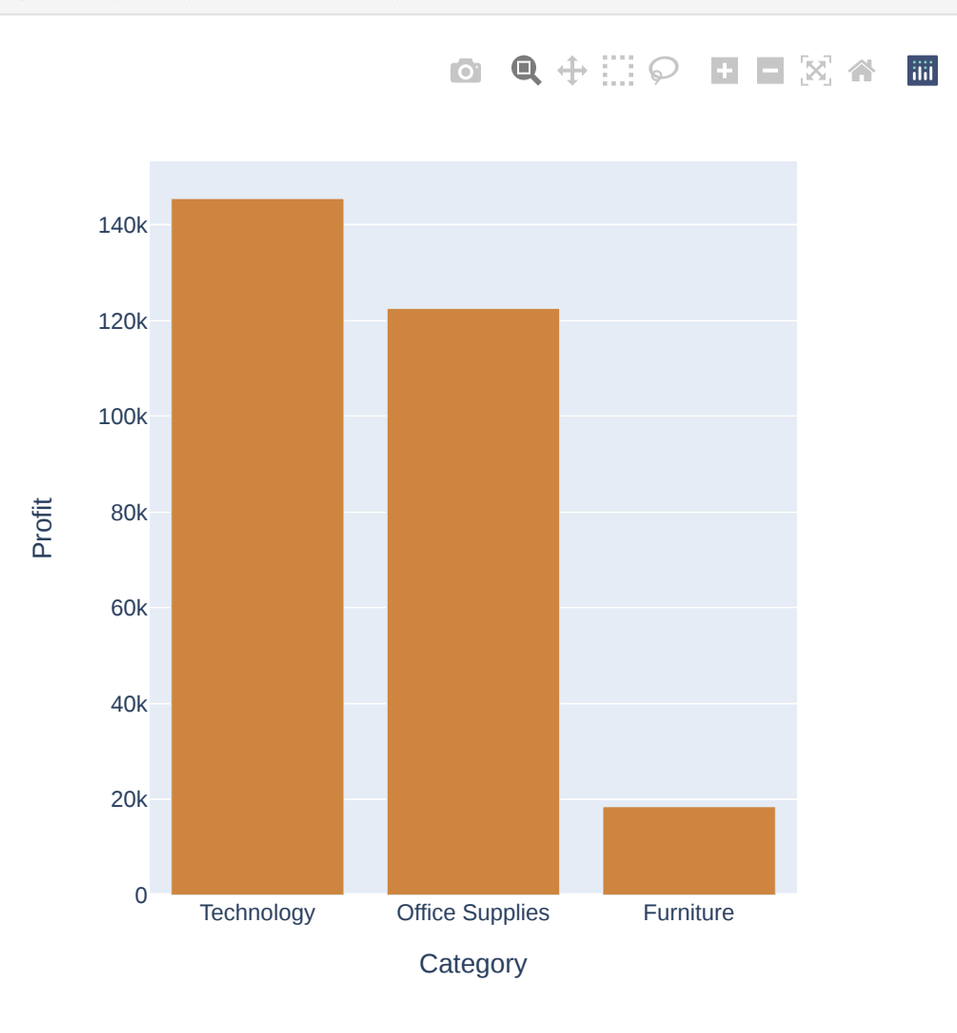
Category vs Sub Category
Technology
Office Supplies
Furniture
Category
Bookcases
Chairs
Labels
Tables
Storage
Furnishings
Art
Phones
Binders
Appliances
Paper
Accessories
Envelopes
Fasteners
Supplies
Machines
Copiers
Sub-Catgory

Sales Analysis - FURNITURE

In [29]:
```
furn = sample_df[sample_df.Category == 'Furniture'].groupby('Sub-Category').sum().sort_values('Profit', ascending=False).iloc[:, [-1]]
plt.figure(figsize=(10,4))
plt.title('Profit - Furniture ')
sns.barplot(x=furn.Profit, y=furn.index,data=furn);
```

Profit of Different Categories

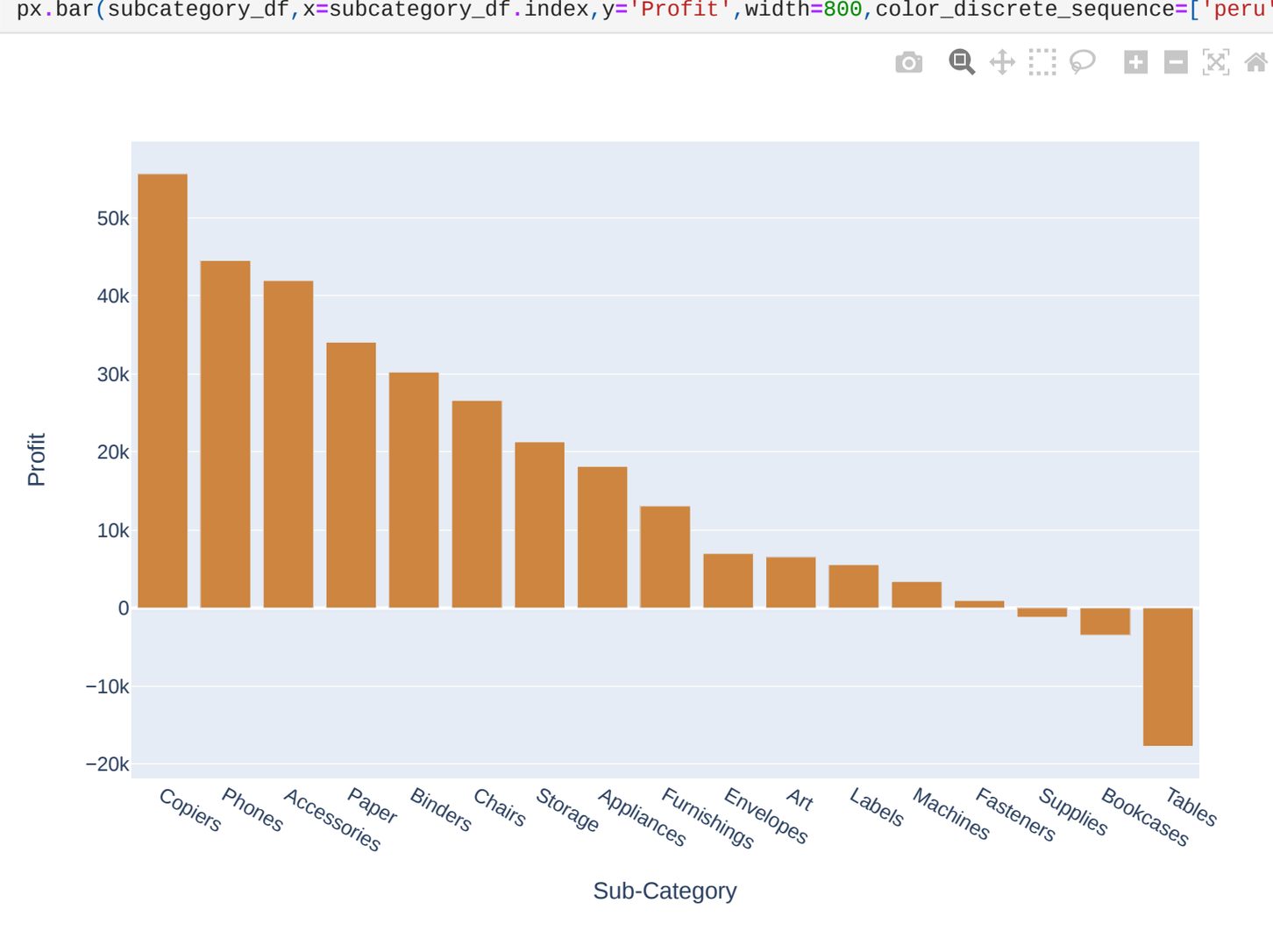
In [30]:
```
category_df = sample_df.groupby('Category').sum().sort_values('Profit',ascending=False)
px.bar(category_df,x=category_df.index,y='Profit',width=500,color_discrete_sequence=['peru'],hover_data=['Sales'])
```



Note: Maximum net profit is incurred in category of Technology followed by Office Supplies and Furniture.

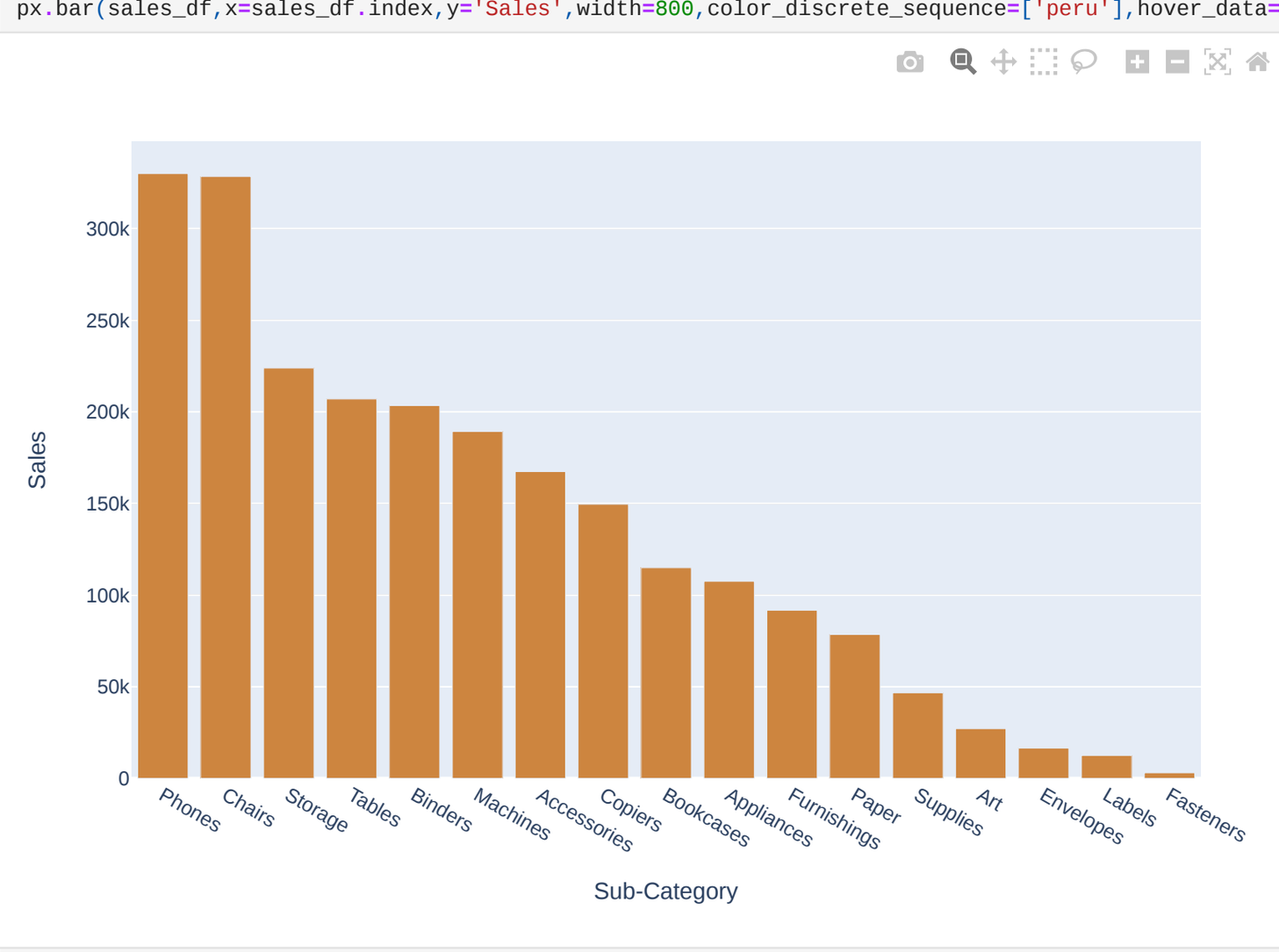
For the category of Furniture, sub-categories - Tables and Bookcases need to be taken care of where losses are incurred

In [31]:
```
subcategory_df = sample_df.groupby('Sub-Category').sum().sort_values('Profit',ascending=False)
px.bar(subcategory_df,x=subcategory_df.index,y='Profit',width=800,color_discrete_sequence=['peru'])
```

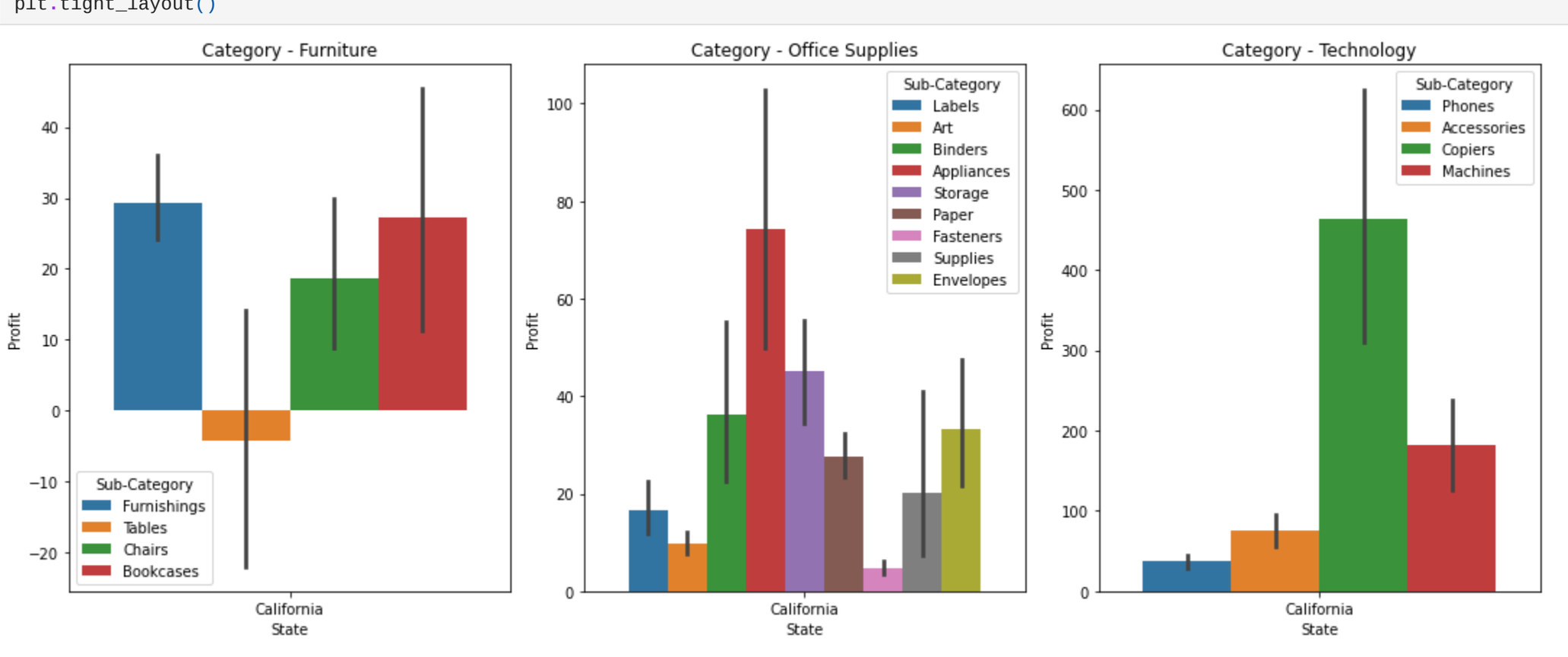


Sales of Different Sub-Categories

In [32]:
```
sales_df = sample_df.groupby('Sub-Category').sum().sort_values('Sales',ascending=False)
px.bar(sales_df,x=sales_df.index,y='Sales',width=800,color_discrete_sequence=['peru'],hover_data=['Profit','Quantity'])
```



In [34]:
```
# Analyzing sales by category for state 'California'

df_California = sample_df[sample_df['State']=='California']
df_furn = df_California[df_California['Category']=='Furniture']
df_offc = df_California[df_California['Category']=='Office Supplies']
df_tech = df_California[df_California['Category']=='Technology']
plt.subplots(figsize=(15,6))


plt.subplot(1,3,1)
sns.barplot(x=df_furn.State, y=df_furn.Profit, hue=df_furn['Sub-Category'])
plt.title('Category - Furniture')

plt.subplot(1,3,2)
sns.barplot(x=df_offc.State, y=df_offc.Profit, hue=df_offc['Sub-Category'])
plt.title('Category - Office Supplies')

plt.subplot(1,3,3)
sns.barplot(x=df_tech.State, y=df_tech.Profit, hue=df_tech['Sub-Category'])
plt.title('Category - Technology')


plt.tight_layout()
```



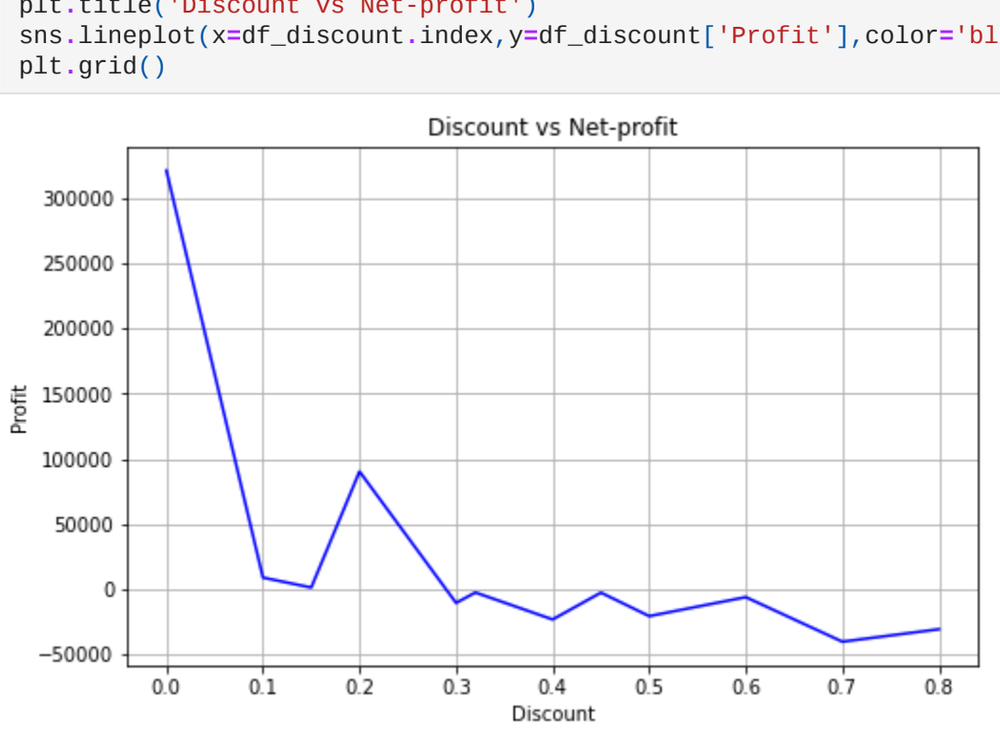
Category - Furniture
Category - Office Supplies
Category - Technology
Profit
Sub-Category
Furnishings
Tables
Chairs
Bookcases
Labels
Art
Binders
Appliances
Storage
Paper
Fasteners
Supplies
Envelopes
Phones
Accessories
Copiers
Machines
California
State

Analysis of Feature-Discount

In [38]:
```
print('Available discounts : ',sample_df.Discount.unique())
```

```
Available discounts :  [0.   0.45 0.2  0.8  0.3  0.5  0.7  0.6  0.32 0.1  0.4  0.15]
```

In [45]:
```
df_discount = sample_df.groupby('Discount').sum()
plt.figure(figsize=(8,5))
plt.title('Discount vs Net-profit')
sns.lineplot(x=df_discount.index,y=df_discount['Profit'],color='blue');
plt.grid()
```



As evident from the plot, net profit decreases when discount is increased.

Business problems that can be derived by looking into the data

- How much is the sales, profit and quantity sold varies region-wise, state-wise and segment-wise, category-wise?
- Which category of items gives the more profit and sold more?
- Which type of mode is suitable for more profit?
- Which state has the highest profit?
- Which region has the highest sales and profit?

Conclusion

- The dataset contains 80.6 % cases of profit, 18.7 % cases of loss and 0.65 % cases of no-profit-no-loss . From this, we can probably say that the overall Retail store in US is working fine.
- The count of Standard Class is maximum. In each class of Ship Mode, Consumer segment holds the majority. We may conclude that the stores prefer Consumer Segment as it may draw more profit.
- It is observed that Segment -'Customer' contributes to highest net profit followed by Corporate and then, Home Office.
- Net profit is maximum for 'Standard Class' ship mode, followed by 'Second Class', 'First Class' and 'Same Day'.
- California, New York, Washington, Michigan, Virginia are the five states which incur maximum net profit
- For the category of Furniture, sub-categories - Tables and Bookcases need to be taken care of where losses are incurred.
- Net profit decreases when discount is increased

In [ ]: