The Sparks Foundation-Data Science and Business Analytics Internship

TASK 3:Exploratory Data Analysis-Retail

Done by:Samyuktha Rajkumaran

Understanding our data

In [157…
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [158…
```
data=pd.read_csv(r"C:\Users\Rajkumar\Downloads\SampleSuperstore.csv")
data.shape
```

Out[158…
```
(9994, 13)
```

In [159…
```
data.head()
```

Out[159…

	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 [160…
```
data.tail()
```

Out[160…

	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 [161…
```
data.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 [162…
```
data.notnull()
```

Out[162…

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	True	True	True	True	True	True	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True	True	True	True	True	True	True
2	True	True	True	True	True	True	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True	True	True	True	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	True	True	True	True	True	True	True	True	True	True	True	True	True
9990	True	True	True	True	True	True	True	True	True	True	True	True	True
9991	True	True	True	True	True	True	True	True	True	True	True	True	True
9992	True	True	True	True	True	True	True	True	True	True	True	True	True
9993	True	True	True	True	True	True	True	True	True	True	True	True	True

9994 rows × 13 columns

In [163…
```
data.isnull()
```

Out[163…

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	False	False	False	False	False	False	False	False	False	False	False	False	False
9990	False	False	False	False	False	False	False	False	False	False	False	False	False
9991	False	False	False	False	False	False	False	False	False	False	False	False	False
9992	False	False	False	False	False	False	False	False	False	False	False	False	False
9993	False	False	False	False	False	False	False	False	False	False	False	False	False

9994 rows × 13 columns

In [164…
```
data.columns
```

Out[164…
```
Index(['Ship Mode', 'Segment', 'Country', 'City', 'State', 'Postal Code',
       'Region', 'Category', 'Sub-Category', 'Sales', 'Quantity', 'Discount',
       'Profit'],
      dtype='object')
```

In [165…
```
data.nunique()
```

Out[165…
```
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 order to clean our data we should drop redundant and irrelevant columns.Here,we should drop Country and Postal Code

In [166…
```
newdata = data.drop(['Country','Postal Code'],axis=1)
```

In [167…
```
newdata['Cost']= newdata['Sales'] - newdata['Profit']
newdata['Profit%']= (newdata['Profit']/newdata['Cost'])*100
```

In [168…
```
newdata.head()
```

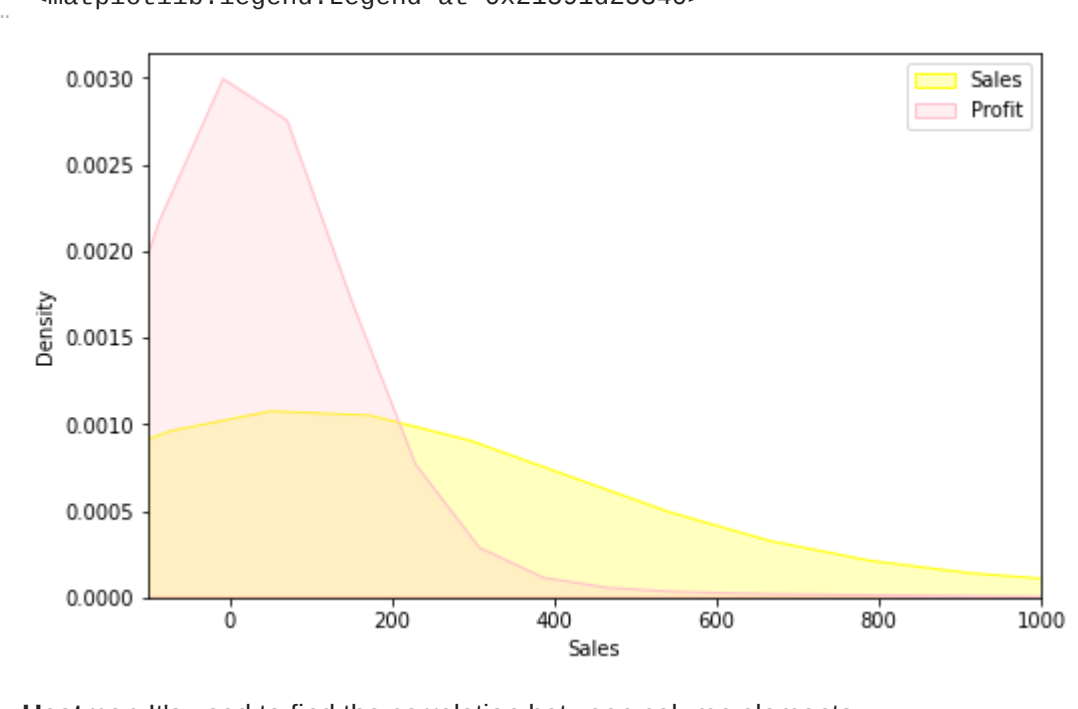
Out[168…

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Cost	Profit%
0	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136	220.0464	19.047619
1	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820	512.3580	42.857143
2	Second Class	Corporate	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714	7.7486	88.679245
3	Standard Class	Consumer	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310	1340.6085	-28.571429
4	Standard Class	Consumer	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164	19.8516	12.676056

Analyzing the relationship

In [169…
```
plt.figure(figsize=(8,5))
sns.kdeplot(data['Sales'],color='yellow',label='Sales',shade=True,bw=0.5)
sns.kdeplot(data['Profit'],color='Pink',label='Profit',shade=True,bw=0.5)
plt.xlim([-100,1000])
plt.legend()
```

Out[169…
```
<matplotlib.legend.Legend at 0x21591d25340>
```



Heatmap It's used to find the correlation between column elements

In [170…
```
correlation = newdata.corr()
```

In [171…
```
sns.heatmap(newdata.corr(),cmap='viridis',annot=True)
```

Out[171…
```
<AxesSubplot:>
```

Heatmap Inference
Sales and profit are moderately correlated.
Quantity and profit are moderately correlated.
Discount and profit are negatively correlated.

Dividing the dataset further
into:prof(profit) and los(loss)

In [172…
```
prof=newdata.loc[newdata['Profit%']>0]
```

In [173…
```
prof.head()
```

Out[173…

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Cost	Profit%
0	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Bookcases	261.960	2	0.0	41.9136	220.0464	19.047619
1	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Chairs	731.940	3	0.0	219.5820	512.3580	42.857143
2	Second Class	Corporate	Los Angeles	California	West	Office Supplies	Labels	14.620	2	0.0	6.8714	7.7486	88.679245
4	Standard Class	Consumer	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.368	2	0.2	2.5164	19.8516	12.676056
5	Standard Class	Consumer	Los Angeles	California	West	Furniture	Furnishings	48.860	7	0.0	14.1694	34.6906	40.845070

In [174…
```
los=newdata.loc[newdata['Profit%']<0]
```

In [175…
```
los.head()
```

Out[175…

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Cost	Profit%
3	Standard Class	Consumer	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310	1340.6085	-28.571429
14	Standard Class	Home Office	Fort Worth	Texas	Central	Office Supplies	Appliances	68.8100	5	0.80	-123.8580	192.6680	-64.285714
15	Standard Class	Home Office	Fort Worth	Texas	Central	Office Supplies	Binders	2.5440	3	0.80	-3.8160	6.3600	-60.000000
23	Second Class	Consumer	Philadelphia	Pennsylvania	East	Furniture	Chairs	71.3720	2	0.30	-1.0196	72.3916	-1.408451
27	Standard Class	Consumer	Philadelphia	Pennsylvania	East	Furniture	Bookcases	3083.4300	7	0.50	-1665.0522	4748.4822	-35.064935

In [176…
```
prof.reset_index(drop=True)
```

Out[176…

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Cost	Profit%
0	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Bookcases	261.960	2	0.0	41.9136	220.0464	19.047619
1	Second Class	Consumer	Henderson	Kentucky	South	Furniture	Chairs	731.940	3	0.0	219.5820	512.3580	42.857143
2	Second Class	Corporate	Los Angeles	California	West	Office Supplies	Labels	14.620	2	0.0	6.8714	7.7486	88.679245
3	Standard Class	Consumer	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.368	2	0.2	2.5164	19.8516	12.676056
4	Standard Class	Consumer	Los Angeles	California	West	Furniture	Furnishings	48.860	7	0.0	14.1694	34.6906	40.845070
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8053	Second Class	Consumer	Miami	Florida	South	Furniture	Furnishings	25.248	3	0.2	4.1028	21.1452	19.402985
8054	Standard Class	Consumer	Costa Mesa	California	West	Furniture	Furnishings	91.960	2	0.0	15.6332	76.3268	20.481928
8055	Standard Class	Consumer	Costa Mesa	California	West	Technology	Phones	258.576	2	0.2	19.3932	239.1828	8.108108
8056	Standard Class	Consumer	Costa Mesa	California	West	Office Supplies	Paper	29.600	4	0.0	13.3200	16.2800	81.818182
8057	Second Class	Consumer	Westminster	California	West	Office Supplies	Appliances	243.160	2	0.0	72.9480	170.2120	42.857143

8058 rows × 13 columns

In [177…
```
los.reset_index(drop=True)
```

Out[177…

	Ship Mode	Segment	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit	Cost	Profit%
0	Standard Class	Consumer	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310	1340.6085	-28.571429
1	Standard Class	Home Office	Fort Worth	Texas	Central	Office Supplies	Appliances	68.8100	5	0.80	-123.8580	192.6680	-64.285714
2	Standard Class	Home Office	Fort Worth	Texas	Central	Office Supplies	Binders	2.5440	3	0.80	-3.8160	6.3600	-60.000000
3	Second Class	Consumer	Philadelphia	Pennsylvania	East	Furniture	Chairs	71.3720	2	0.30	-1.0196	72.3916	-1.408451
4	Standard Class	Consumer	Philadelphia	Pennsylvania	East	Furniture	Bookcases	3083.4300	7	0.50	-1665.0522	4748.4822	-35.064935
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1866	Standard Class	Corporate	Bryan	Texas	Central	Office Supplies	Binders	22.3860	7	0.80	-35.8176	58.2036	-61.538462
1867	Standard Class	Home Office	Akron	Ohio	East	Office Supplies	Binders	5.7420	3	0.70	-4.5936	10.3356	-44.444444
1868	Standard Class	Consumer	San Bernardino	California	West	Furniture	Bookcases	683.3320	4	0.15	-40.1960	723.5280	-5.555556
1869	Second Class	Corporate	Los Angeles	California	West	Furniture	Tables	71.0880	2	0.20	-1.7772	72.8652	-2.439024
1870	First Class	Home Office	Houston	Texas	Central	Furniture	Bookcases	383.4656	4	0.32	-67.6704	451.1360	-15.000000

1871 rows × 13 columns

In [178…
```
prof.shape
```

Out[178…
```
(8058, 13)
```

In [179…
```
los.shape
```

Out[179…
```
(1871, 13)
```

In [180…
```
prof.value_counts()
```

Out[180…
```
Ship Mode       Segment      City          State       Region   Category         Sub-Category  Sales    Quantity  Discount  Profit   Cost     Profit%  
Standard Class  Consumer     Los Angeles   California  West     Office Supplies  Paper         19.440   3         0.0       9.3312   10.1088  92.307692
4
                Corporate    Houston       Texas       Central  Office Supplies  Paper         15.552   3         0.2       5.4432   10.1088  53.846154
3
First Class     Corporate    Los Angeles   California  West     Office Supplies  Binders       7.184    2         0.2       2.2450   4.9390   45.454545
2
Standard Class  Consumer     Dallas        Texas       Central  Office Supplies  Labels        4.928    2         0.2       1.7248   3.2032   53.846154
2
                             Salem         Oregon      West     Office Supplies  Paper         10.368   2         0.2       3.6288   6.7392   53.846154
2
                                                                                                                                                           
..
Second Class    Corporate    Jacksonville  Florida     South    Office Supplies  Labels        6.264    3         0.2       2.0358   4.2282   48.148148
1
                                                                                 Art           5.248    2         0.2       0.5904   4.6576   12.676056
1
                                                                                 Appliances    20.808   3         0.2       1.8207   18.9873  9.589041 
1
                                                                Furniture        Furnishings   12.320   5         0.2       1.8480   10.4720  17.647059
1
Standard Class  Home Office  Yonkers       New York    East     Technology       Machines      52.440   4         0.0       24.1224  28.3176  85.185185
1
Length: 8014, dtype: int64
```

In [181…
```
los.value_counts()
```

Out[181…
```
Ship Mode       Segment      City           State       Region   Category         Sub-Category  Sales     Quantity  Discount  Profit     Cost       Profi
t%
Standard Class  Consumer     San Francisco  California  West     Furniture        Bookcases     359.499   3         0.15      -29.6058   389.1048   -7.60
8696     2
Second Class    Corporate    Chicago        Illinois    Central  Office Supplies  Binders       3.564     3         0.80      -6.2370    9.8010     -63.6
36364    2
Standard Class  Home Office  Columbus       Ohio        East     Furniture        Chairs        281.372   2         0.30      -12.0588   293.4308   -4.10
9589     2
                Consumer     Chicago        Illinois    Central  Office Supplies  Storage       35.168    2         0.20      -8.3524    43.5204    -19.1
91919    2
                             New York City  New York    East     Furniture        Bookcases     353.568   2         0.20      -44.1960   397.7640   -11.1
11111    2
                                                                                                                                                            
..
Second Class    Corporate    Mcallen        Texas       Central  Furniture        Chairs        56.686    1         0.30      -20.2450   76.9310    -26.3
15789    1
                             Loveland       Colorado    West     Office Supplies  Binders       1.188     2         0.70      -0.9900    2.1780     -45.4
54545    1
                             Los Angeles    California  West     Furniture        Tables        1322.352  3         0.20      -99.1764   1421.5284  -6.97
6744     1
                                                                                                71.088    2         0.20      -1.7772    72.8652    -2.43
9024     1
Standard Class  Home Office  Yuma           Arizona     West     Technology       Machines      599.985   5         0.70      -479.9880  1079.9730  -44.4
44444    1
Length: 1865, dtype: int64
```

It's noteworthy that the business is overall yielding profits as the company is making profits in 8058 entries and losses in 1871 entries.Let's analyse this further.

In [182…
```
prof['City'].value_counts()
```

Out[182…
```
New York City    868
Los Angeles      705
San Francisco    482
Seattle          406
Philadelphia     283
                ... 
Orland Park        1
Whittier           1
Holyoke            1
Richardson         1
Mason              1
Name: City, Length: 513, dtype: int64
```

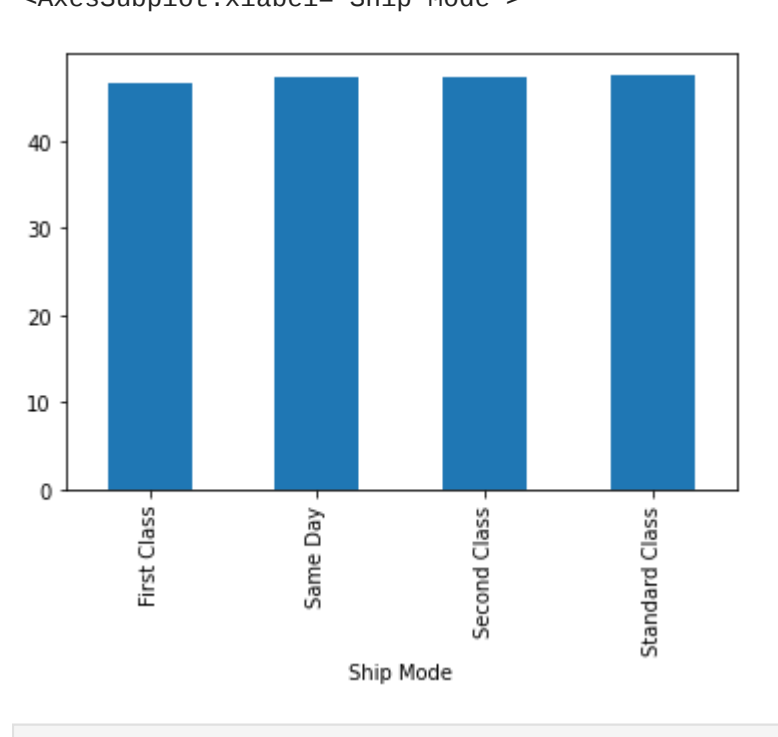
In [183…
```
los['City'].value_counts()
```

Out[183…
```
Philadelphia       250
Houston            185
Chicago            155
Dallas              73
Columbus            51
                  ... 
Harlingen            1
Champaign            1
North Las Vegas      1
Altoona              1
San Bernardino       1
Name: City, Length: 229, dtype: int64
```

We can hypothesize that the most profit is generated in New York City while the most loss is generated in Philadelphia.

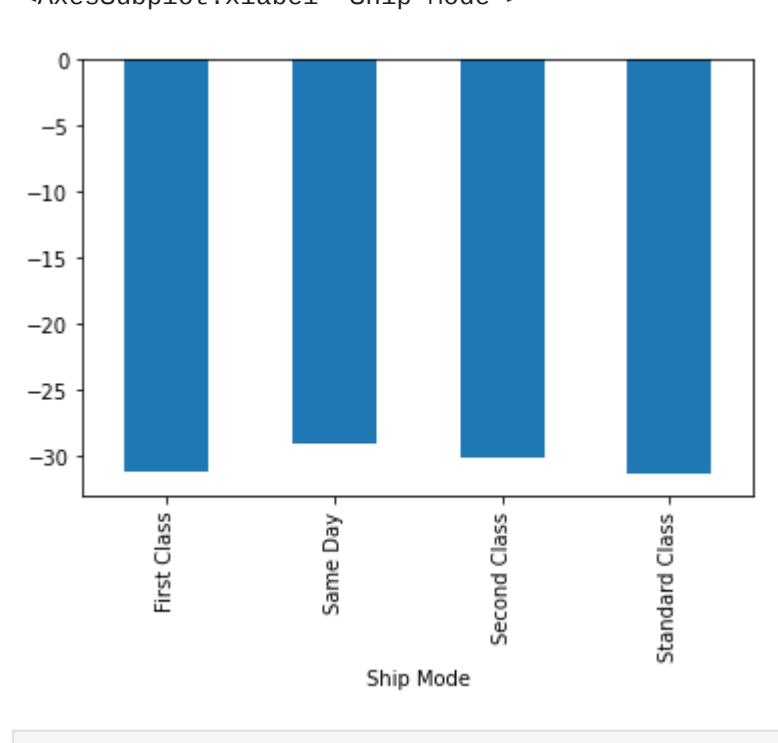
In [185…
```
prof.groupby('Ship Mode')['Profit%'].mean().plot.bar()
```

Out[185…
```
<AxesSubplot:xlabel='Ship Mode'>
```



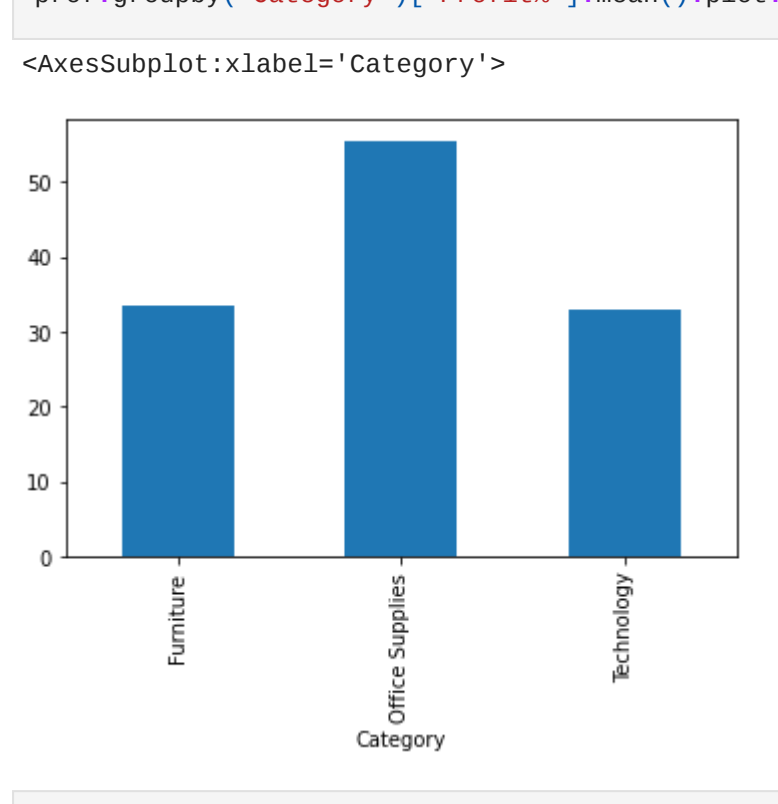
In [186…
```
los.groupby('Ship Mode')['Profit%'].mean().plot.bar()
```

Out[186…
```
<AxesSubplot:xlabel='Ship Mode'>
```



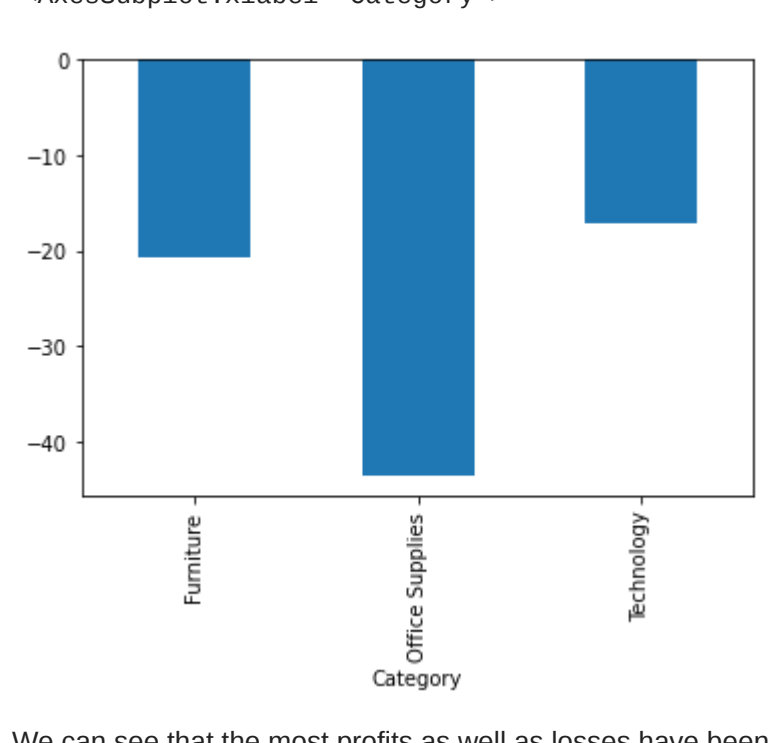
In [187…
```
prof.groupby('Category')['Profit%'].mean().plot.bar()
```

Out[187…
```
<AxesSubplot:xlabel='Category'>
```



In [188…
```
los.groupby('Category')['Profit%'].mean().plot.bar()
```

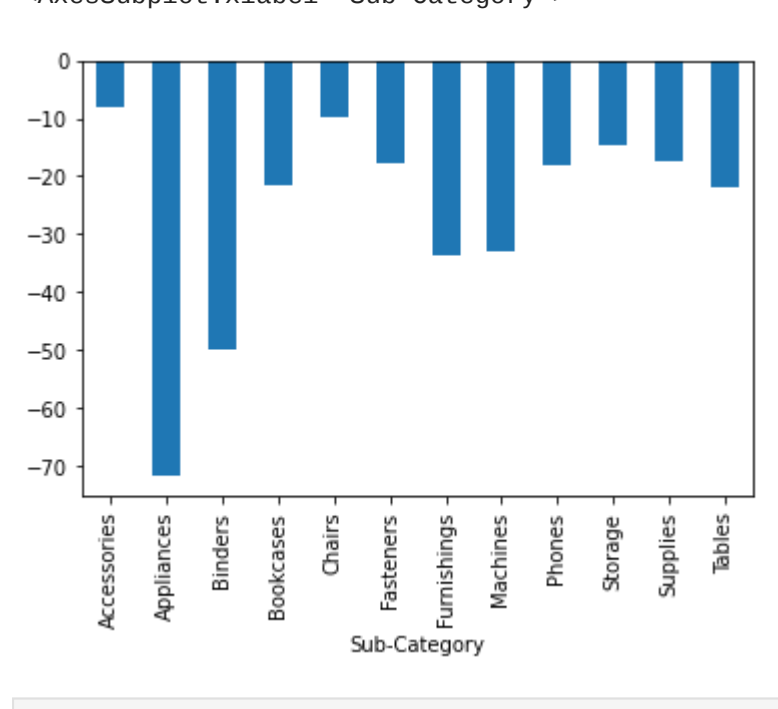
Out[188…
```
<AxesSubplot:xlabel='Category'>
```



We can see that the most profits as well as losses have been incurred in the Office Supplies Category.We need to analyse the sub-categories as well.

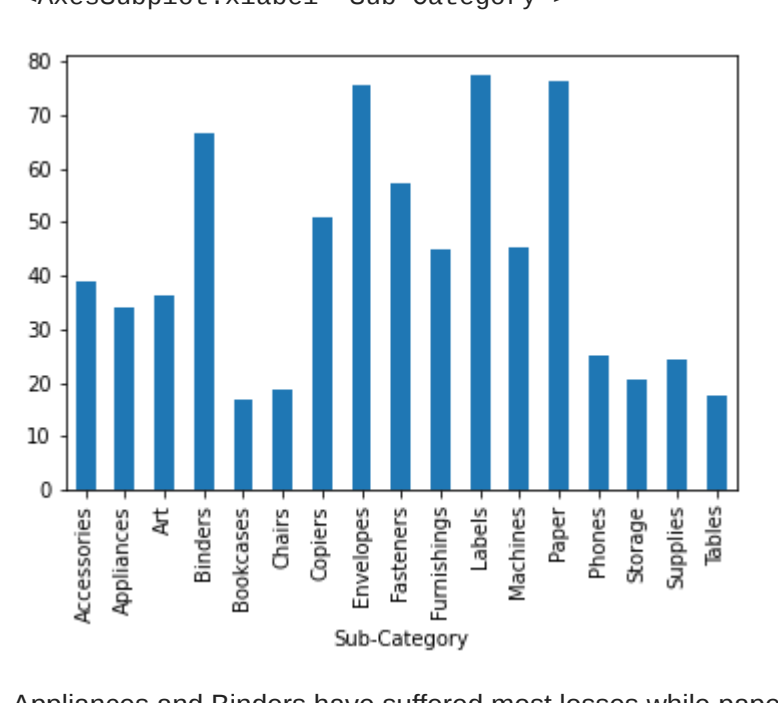
In [189…
```
los.groupby('Sub-Category')['Profit%'].mean().plot.bar()
```

Out[189…
```
<AxesSubplot:xlabel='Sub-Category'>
```



In [190…
```
prof.groupby('Sub-Category')['Profit%'].mean().plot.bar()
```

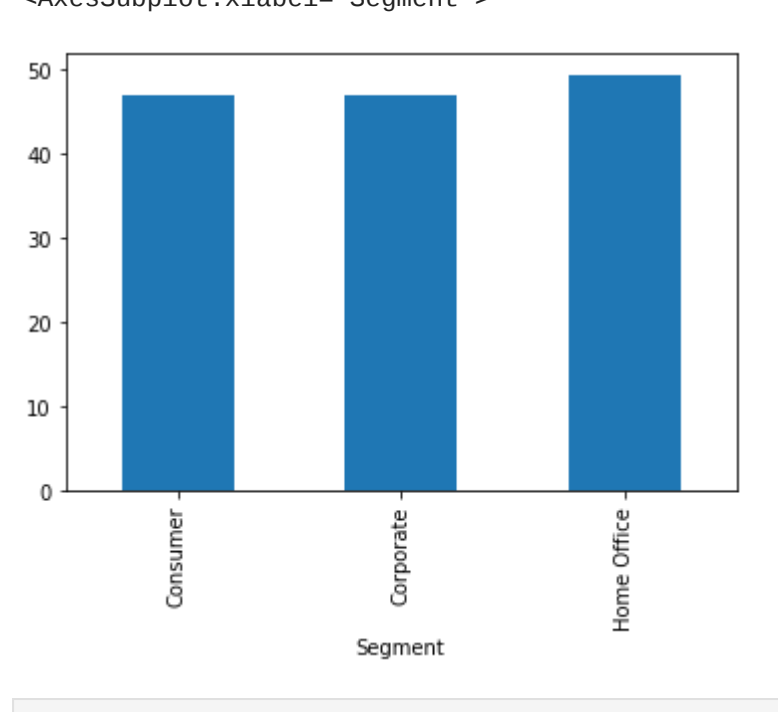
Out[190…
```
<AxesSubplot:xlabel='Sub-Category'>
```



Appliances and Binders have suffered most losses while paper and labels have brought out the most profit.

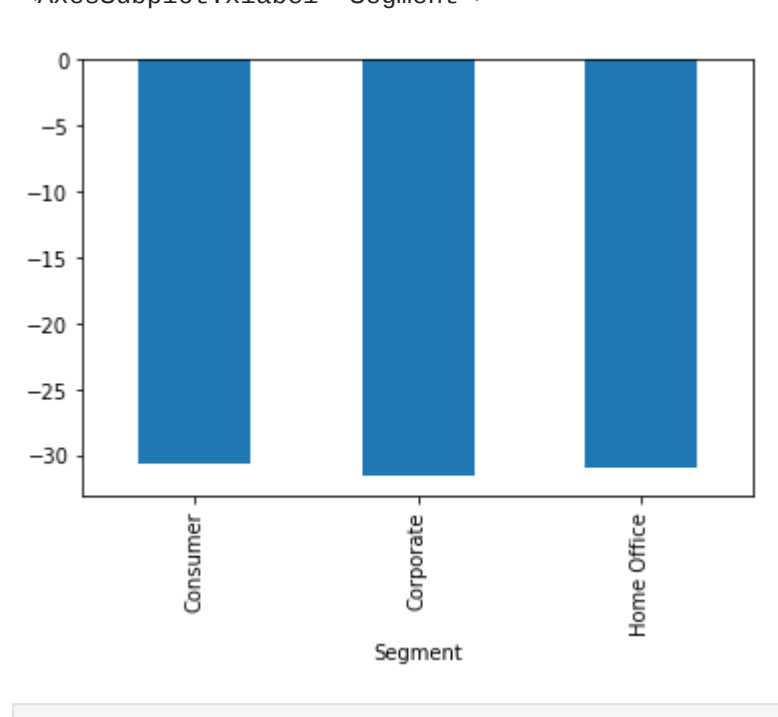
In [191…
```
prof.groupby('Segment')['Profit%'].mean().plot.bar()
```

Out[191…
```
<AxesSubplot:xlabel='Segment'>
```



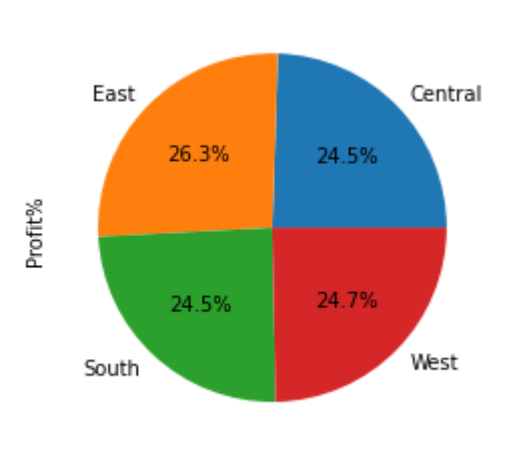
In [192…
```
los.groupby('Segment')['Profit%'].mean().plot.bar()
```

Out[192…
```
<AxesSubplot:xlabel='Segment'>
```



In [194…
```
prof.groupby('Region')['Profit%'].mean().plot.pie(autopct="%.1f%%")
```

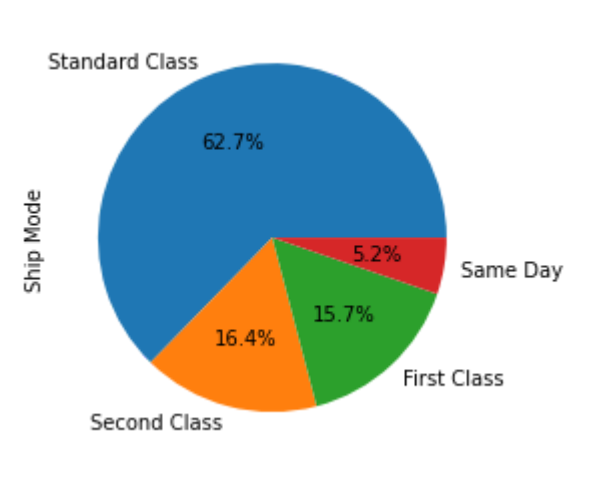
Out[194…
```
<AxesSubplot:ylabel='Profit%'>
```



In [198…
```
graph=los['Ship Mode'].value_counts()
```

In [199…
```
graph.plot.pie(autopct="%.1f%%")
```

Out[199…
```
<AxesSubplot:ylabel='Ship Mode'>
```



In [200…
```
prof.groupby('State')['Profit%'].mean()
```

Out[200…
```
State
Alabama                 57.035981
Arizona                 29.596523
Arkansas                66.037987
California              49.296384
Colorado                26.723697
Connecticut             59.535868
Delaware                61.867211
District of Columbia    74.969477
Florida                 29.573372
Georgia                 58.758111
Idaho                   45.027213
Illinois                31.171962
Indiana                 60.042150
Iowa                    71.072453
Kansas                  55.323261
Kentucky                56.021568
Louisiana               57.003986
Maine                   60.659094
Maryland                60.162794
Massachusetts           60.291566
Michigan                56.013903
Minnesota               62.775354
Mississippi             60.653238
Missouri                59.800716
Montana                 50.835528
Nebraska                55.741688
Nevada                  53.004812
New Hampshire           61.055753
New Jersey              58.659983
New Mexico              50.973856
New York                52.961044
North Carolina          27.811624
North Dakota            57.396117
Ohio                    31.453295
Oklahoma                55.598073
Oregon                  30.513179
Pennsylvania            31.261299
Rhode Island            58.975207
South Carolina          55.013330
South Dakota            65.246039
Tennessee               30.783915
Texas                   33.678226
Utah                    50.474091
Vermont                 55.873440
Virginia                56.222189
Washington              47.318674
West Virginia           96.078431
Wisconsin               51.819017
Wyoming                  6.666667
Name: Profit%, dtype: float64
```

In [201…
```
los.groupby('State')['Profit%'].mean()
```

Out[201…
```
State
Arizona          -30.265361
California        -9.274100
Colorado         -34.971742
Connecticut       -6.650071
Delaware         -14.621449
Florida          -29.034221
Illinois         -40.464033
Maryland          -8.492823
Massachusetts    -13.895860
Nevada           -13.978495
New Hampshire     -9.090909
New Jersey       -10.824408
New Mexico       -19.191919
New York         -17.066147
North Carolina   -30.815065
Ohio             -27.484667
Oregon           -30.983764
Pennsylvania     -26.750433
Rhode Island     -10.264431
Tennessee        -27.695211
Texas            -38.113559
Washington       -10.468661
West Virginia    -10.256410
Name: Profit%, dtype: float64
```

Most profits have come from West Virginia while the most losses have come from Illinois.

Conclusion
By dividing the data and analysing it from various perspectives using different fields,we can conclude that though our business is doing well,improvement is necessary in the Home Office Segment.Furthermore, we need to improve our reach in the southern region.