Name:- Rudra Pratap Singh
Task 5:- Exploratory Data Analysis - Retail
Iot & Computer Vision Intern
The Sparks Foundation

In [1]:

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

In [2]:

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

Out[2]:

	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

In [59]:

```
data_set.shape
```

Out[59]:

```
(9994, 13)
```

In [60]:

```
data_set.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 [61]:

```
data_set.describe()
```

Out[61]:

	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 [62]:

```
data_set.nunique()
```

Out[62]:

```
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
```

checking for duplicate values

In [63]:

```
data_set.duplicated().sum()
```

Out[63]:

```
17
```

In [65]:

```
#dropping duplicate values from the dataset
data_set.drop_duplicates()
```

Out[65]:

	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

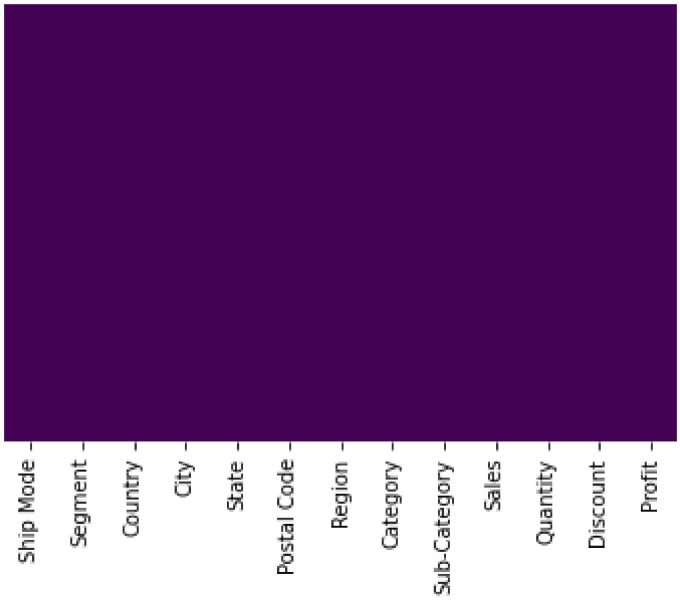
9977 rows × 13 columns

In [66]:

```
sns.heatmap(data_set.isnull(), cbar=False, yticklabels=False, cmap='viridis')
```

Out[66]:

<AxesSubplot:>



Analysis using Pairplot of each column

In [67]:

```
sns.pairplot(data_set, hue = 'Category')
```

Out[67]:

```
<seaborn.axisgrid.PairGrid at 0x261365904f0>
```

In [12]:

```
#Based on Region
sns.pairplot(data_set, hue = 'Region')
```

Out[12]:

```
<seaborn.axisgrid.PairGrid at 0x261252630a0>
```

In [68]:

```
#Based on Segment
sns.pairplot(data_set, hue = 'Segment')
```

Out[68]:

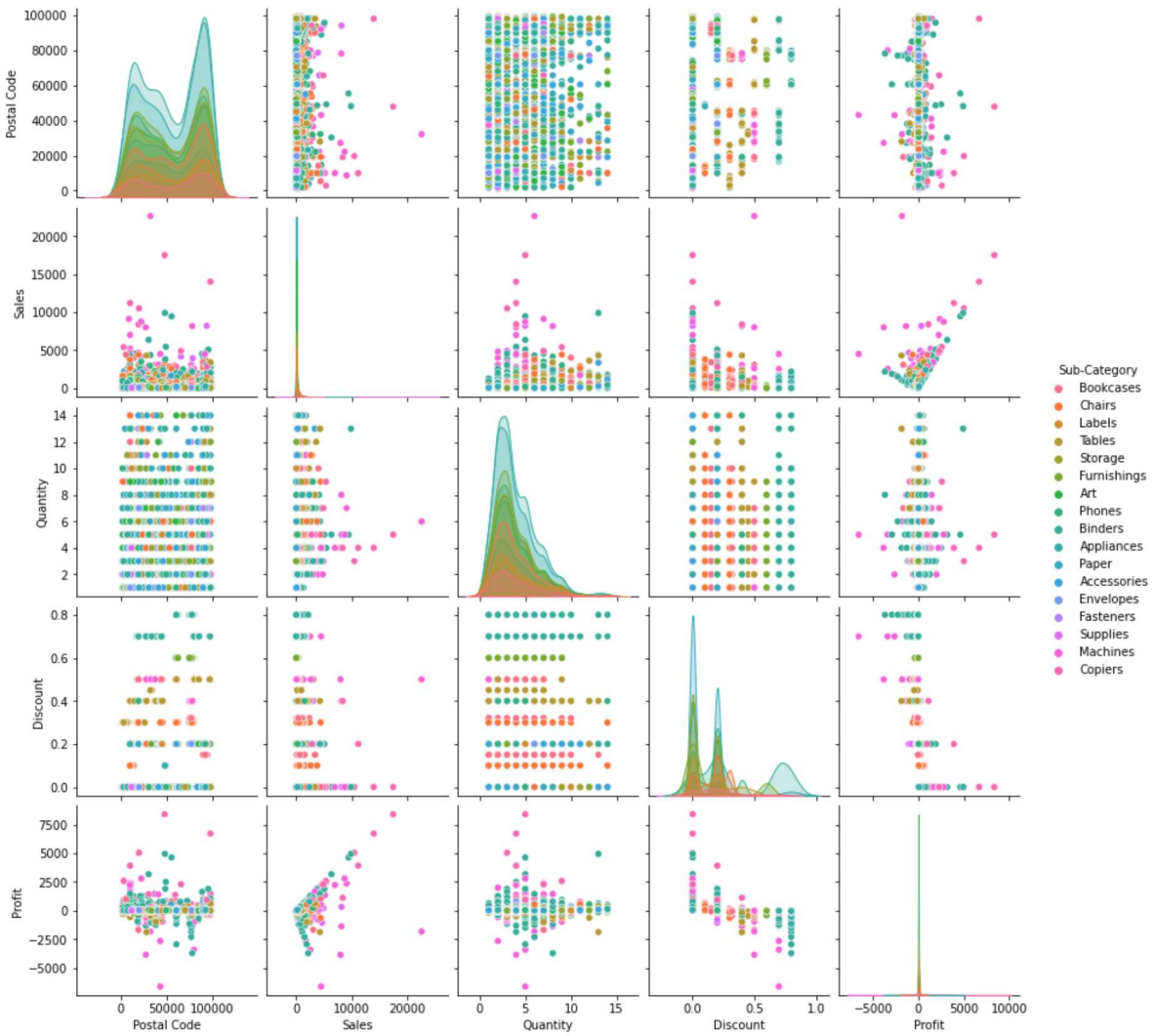
```
<seaborn.axisgrid.PairGrid at 0x26139fd1f40>
```

In [16]:

```
#Based on Sub-Category
sns.pairplot(data_set, hue = 'Sub-Category')
```

Out[16]:

```
<seaborn.axisgrid.PairGrid at 0x2612fdba4c0>
```

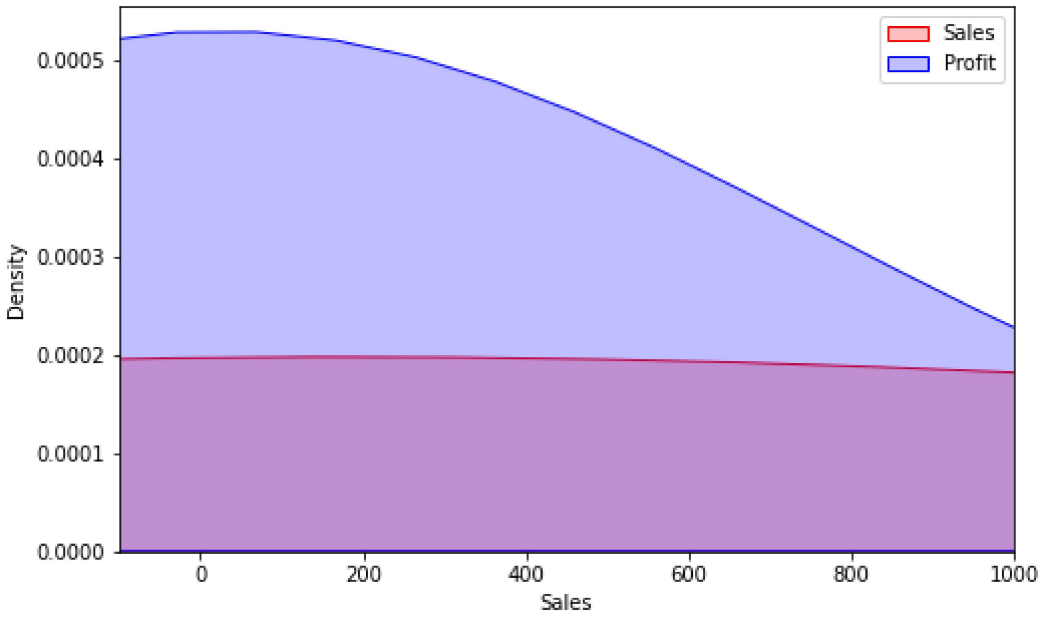


In [69]:

```
#Exploratory Data Analysis
plt.figure(figsize=(8,5))
sns.kdeplot(data_set['Sales'],color='red',label='Sales',shade=True,bw_adjust=20)
sns.kdeplot(data_set['Profit'],color='Blue',label='Profit',shade=True,bw_adjust=20)
plt.xlim([-100,1000])
plt.legend()
```

Out[69]:

```
<matplotlib.legend.Legend at 0x2613667f550>
```

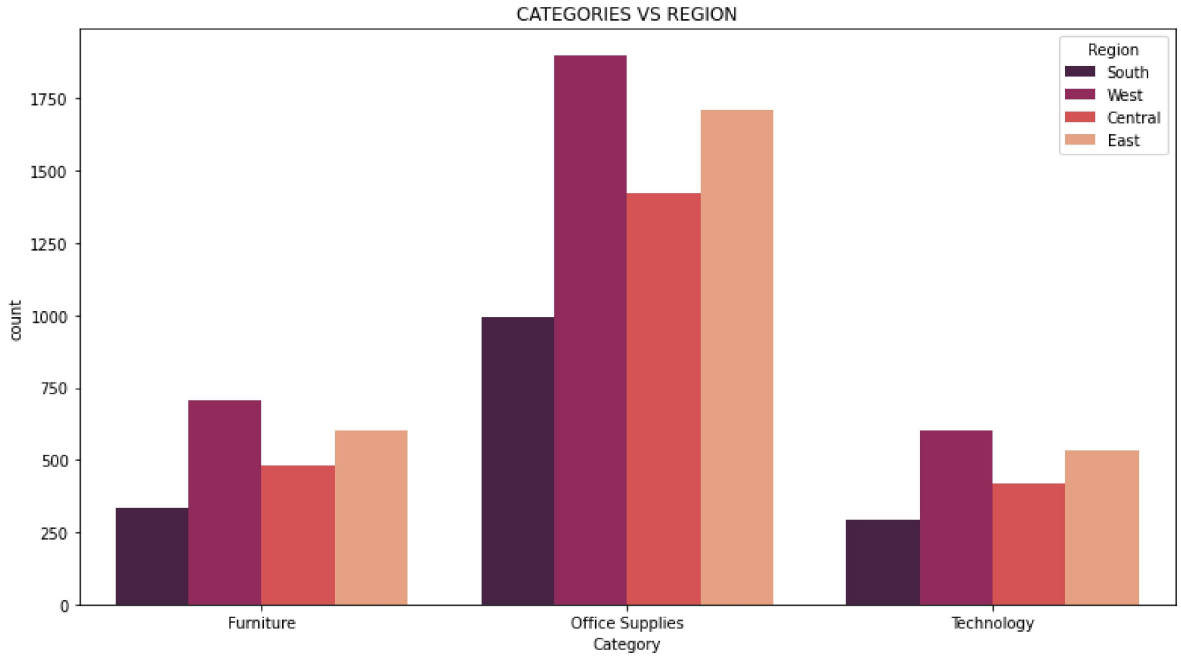


In [70]:

```
#CATEGORIES VS REGION
plt.figure(figsize=(13,7))
plt.title('CATEGORIES VS REGION')
sns.countplot(x=data_set['Category'],hue=data_set['Region'],palette='rocket')
plt.xticks()
```

Out[70]:

```
(array([0, 1, 2]),
 [Text(0, 0, 'Furniture'),
  Text(1, 0, 'Office Supplies'),
  Text(2, 0, 'Technology')])
```

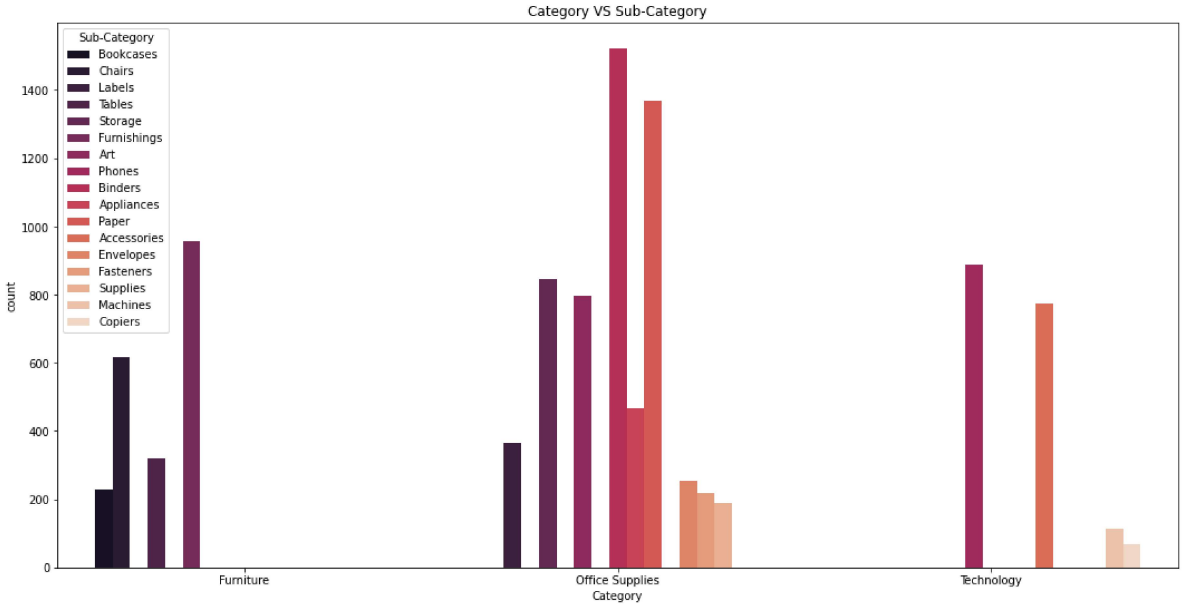


In [71]:

```
#Category VS Sub-Category
plt.figure(figsize=(18,9))
plt.title('Category VS Sub-Category')
sns.countplot(x=data_set['Category'],hue=data_set['Sub-Category'],palette='rocket')
plt.xticks()
```

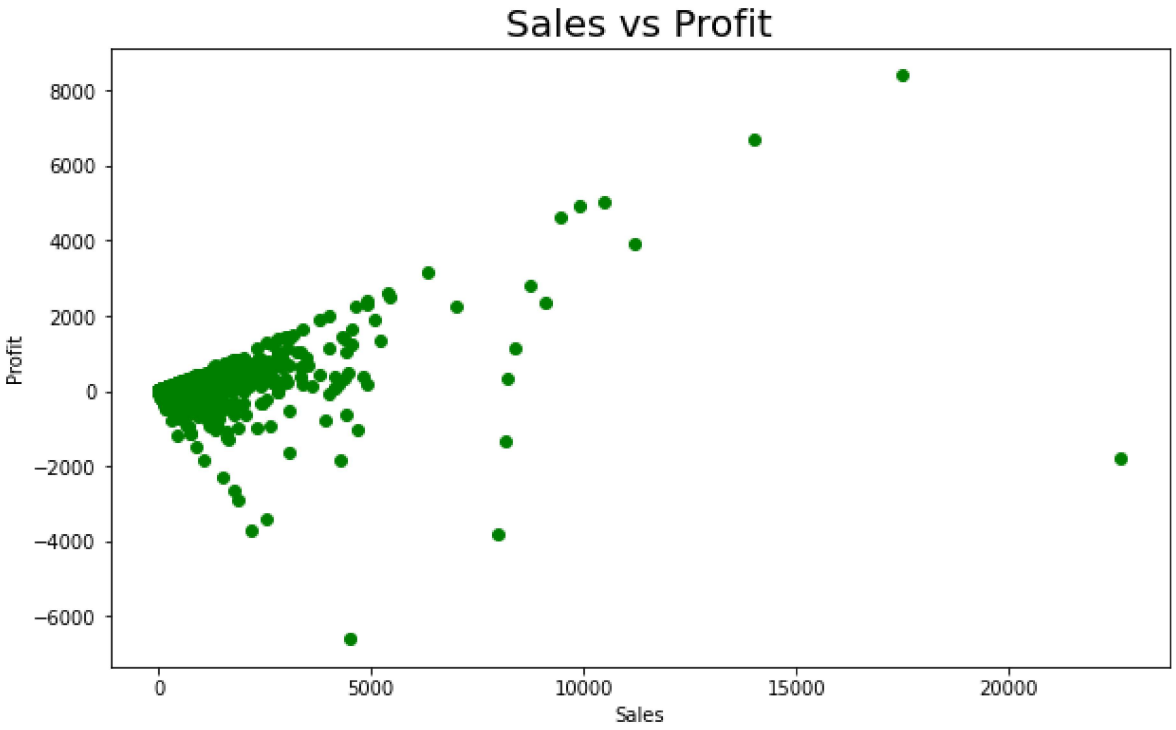
Out[71]:

```
(array([0, 1, 2]),
 [Text(0, 0, 'Furniture'),
  Text(1, 0, 'Office Supplies'),
  Text(2, 0, 'Technology')])
```



In [72]:

```
#Sales vs Profit
plt.subplots(figsize=(10,6))
plt.scatter(data_set['Sales'],data_set['Profit'],color='green')
plt.xlabel('Sales')
plt.ylabel("Profit")
plt.title('Sales vs Profit',fontsize=20)
plt.show()
```



In [73]:

```
# correlation matrix of the data
data_set.corr()
```

Out[73]:

	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

In [74]:

```
# covariance matrix of data
data_set.cov()
```

Out[74]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.028080e+09	-476682.766590	910.415885	386.870404	-225045.849445
Sales	-4.766828e+05	388434.455308	278.459923	-3.627228	69944.096586
Quantity	9.104159e+02	278.459923	4.951113	0.003961	34.534769
Discount	3.868704e+02	-3.627228	0.003961	0.042622	-10.615173
Profit	-2.250458e+05	69944.096586	34.534769	-10.615173	54877.798055

In [75]:

```
#Heatmap for Correlation
sns.heatmap(data_set.corr(),annot= True,cmap = 'rocket')
plt.figure(figsize=(14,7))
```

Out[75]:

```
<Figure size 1008x504 with 0 Axes>
```


```
<Figure size 1008x504 with 0 Axes>
```

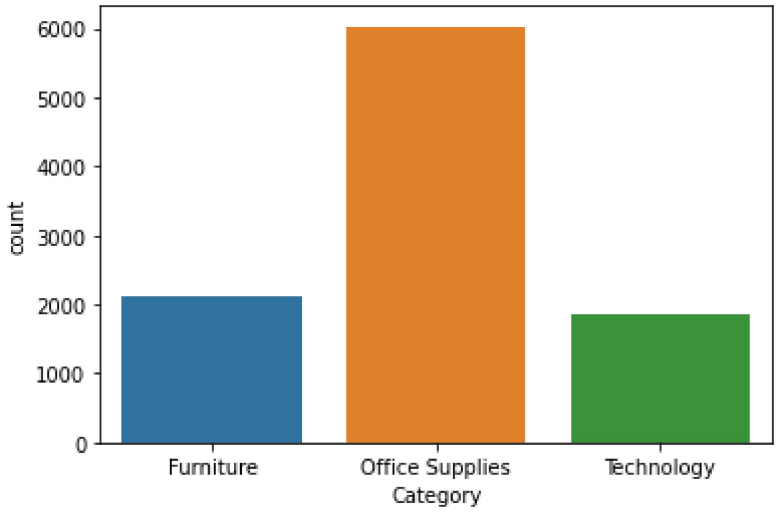
In [76]:

```
sns.countplot(data_set['Category'])
plt.figure(figsize=(15,7))
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```

Out[76]:

```
<Figure size 1080x504 with 0 Axes>
```




```
<Figure size 1080x504 with 0 Axes>
```

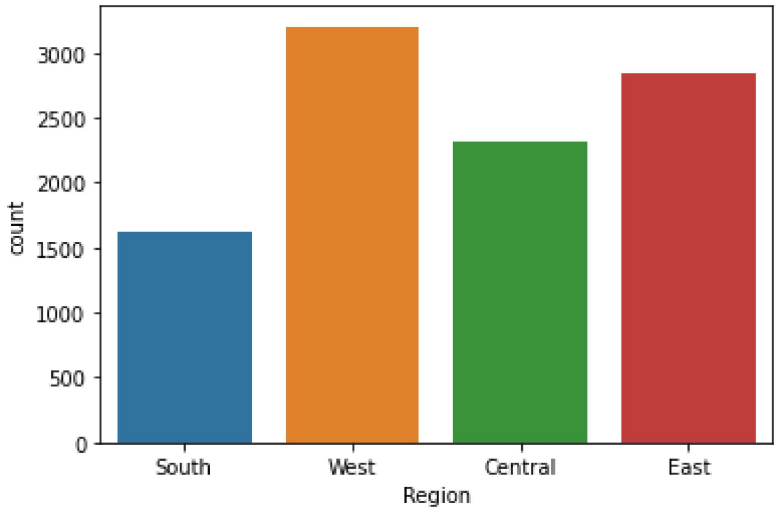
In [77]:

```
sns.countplot(data_set['Region'])
plt.figure(figsize=(12,6))
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```

Out[77]:

```
<Figure size 864x432 with 0 Axes>
```

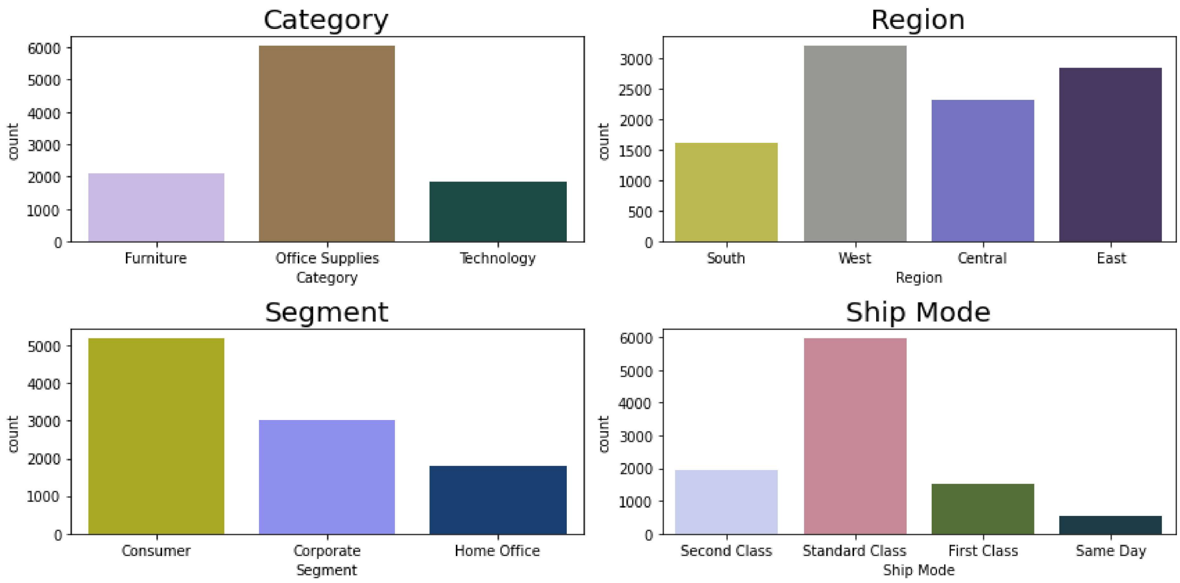



```
<Figure size 864x432 with 0 Axes>
```

In [30]:

```
fig, axs = plt.subplots(nrows = 2,ncols = 2,figsize = (12,6))
sns.countplot(data_set['Category'],ax = axs[0][0],palette='cubehelix_r')
sns.countplot(data_set['Region'],ax = axs[0][1],palette='gist_stern_r')
sns.countplot(data_set['Segment'],ax = axs[1][0],palette='gist_stern_r')
sns.countplot(data_set['Ship Mode'],ax = axs[1][1],palette='cubehelix_r')
axs[0][0].set_title('Category',fontsize=20)
axs[0][1].set_title('Region',fontsize=20)
axs[1][0].set_title('Segment',fontsize=20)
axs[1][1].set_title('Ship Mode',fontsize=20)
plt.tight_layout()
```

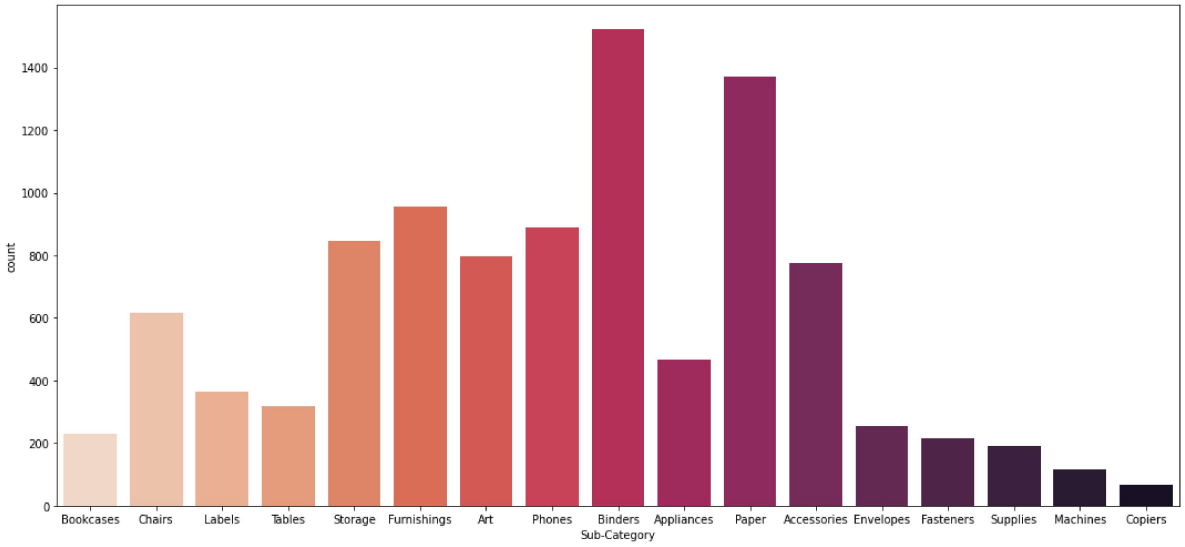
```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



In [78]:

```
plt.figure(figsize=(15,7))
sns.countplot(data_set['Sub-Category'], palette='rocket_r')
plt.tight_layout()
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



In [79]:

```
data_set['Quantity'].value_counts()
```

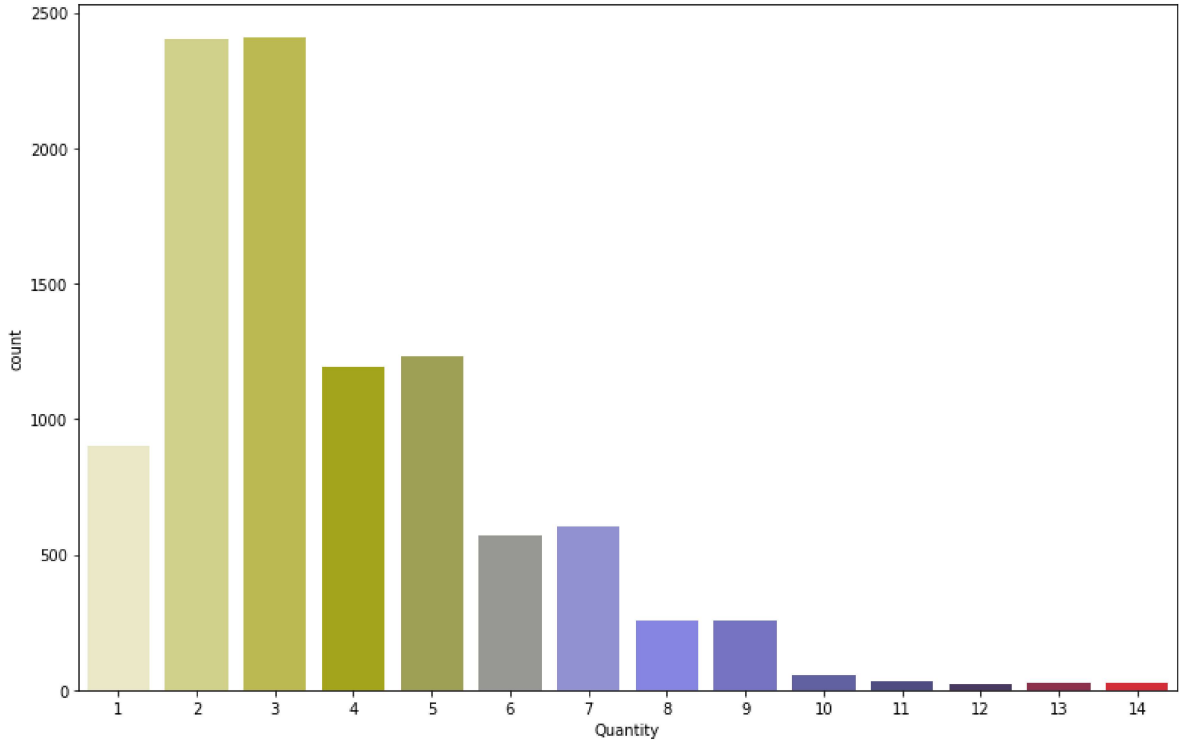
Out[79]:

```
3     2409
2     2402
5     1230
4     1191
1      899
7      606
6      572
9      258
8      257
10      57
11      34
14      29
13      27
12      23
Name: Quantity, dtype: int64
```

In [80]:

```
plt.figure(figsize=(11,7))
sns.countplot(data_set['Quantity'], palette='gist_stern_r')
plt.tight_layout()
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



In [81]:

```
data_set['State'].value_counts()
```

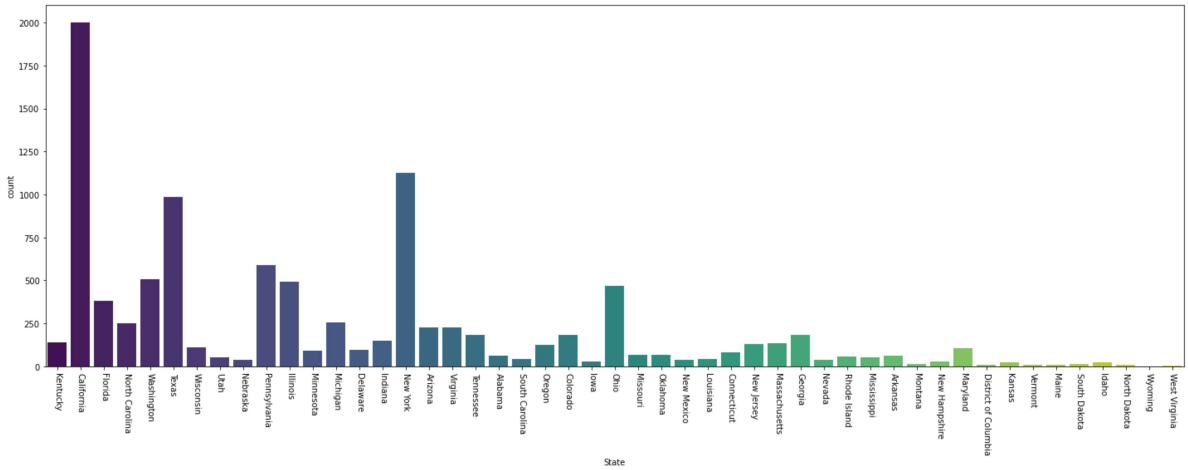
Out[81]:

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

In [82]:

```
plt.figure(figsize=(20,8))
sns.countplot(data_set['State'], palette='viridis')
plt.xticks(rotation=270)
plt.tight_layout()
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



In [83]:

```
data_set['Discount'].value_counts()
```

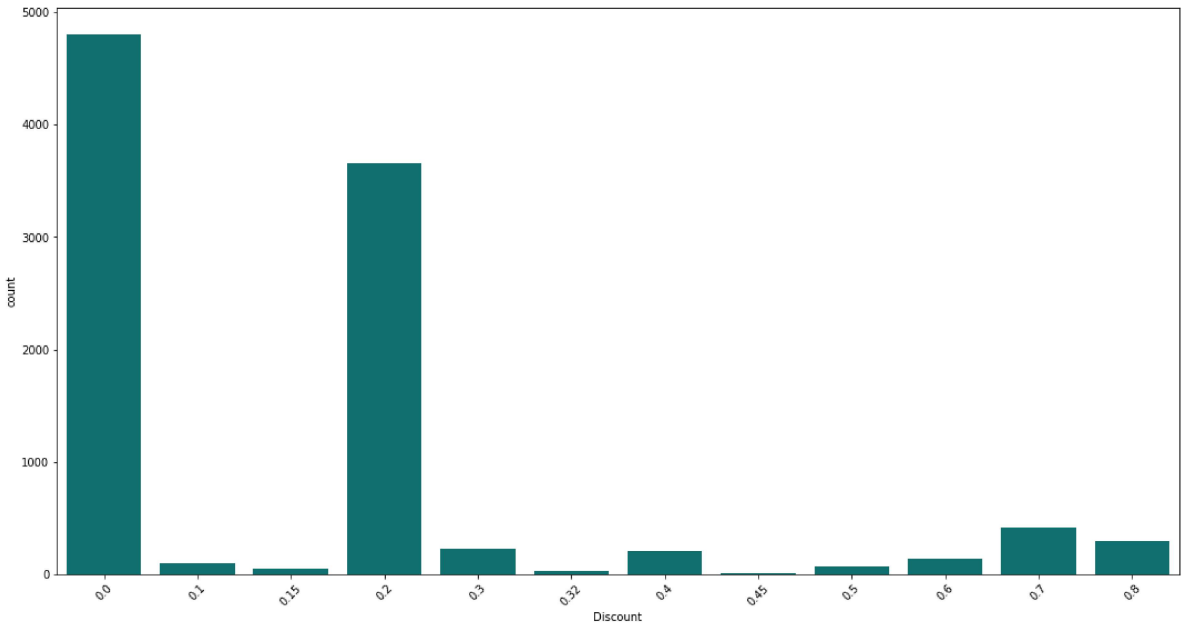
Out[83]:

```
0.00    4798
0.20    3657
0.70     418
0.80     300
0.30     227
0.40     206
0.60     138
0.10      94
0.50      66
0.15      52
0.32      27
0.45      11
Name: Discount, dtype: int64
```

In [84]:

```
plt.figure(figsize=(15,8))
sns.countplot(data_set['Discount'], color='Teal')
plt.xticks(rotation=45)
plt.tight_layout()
```

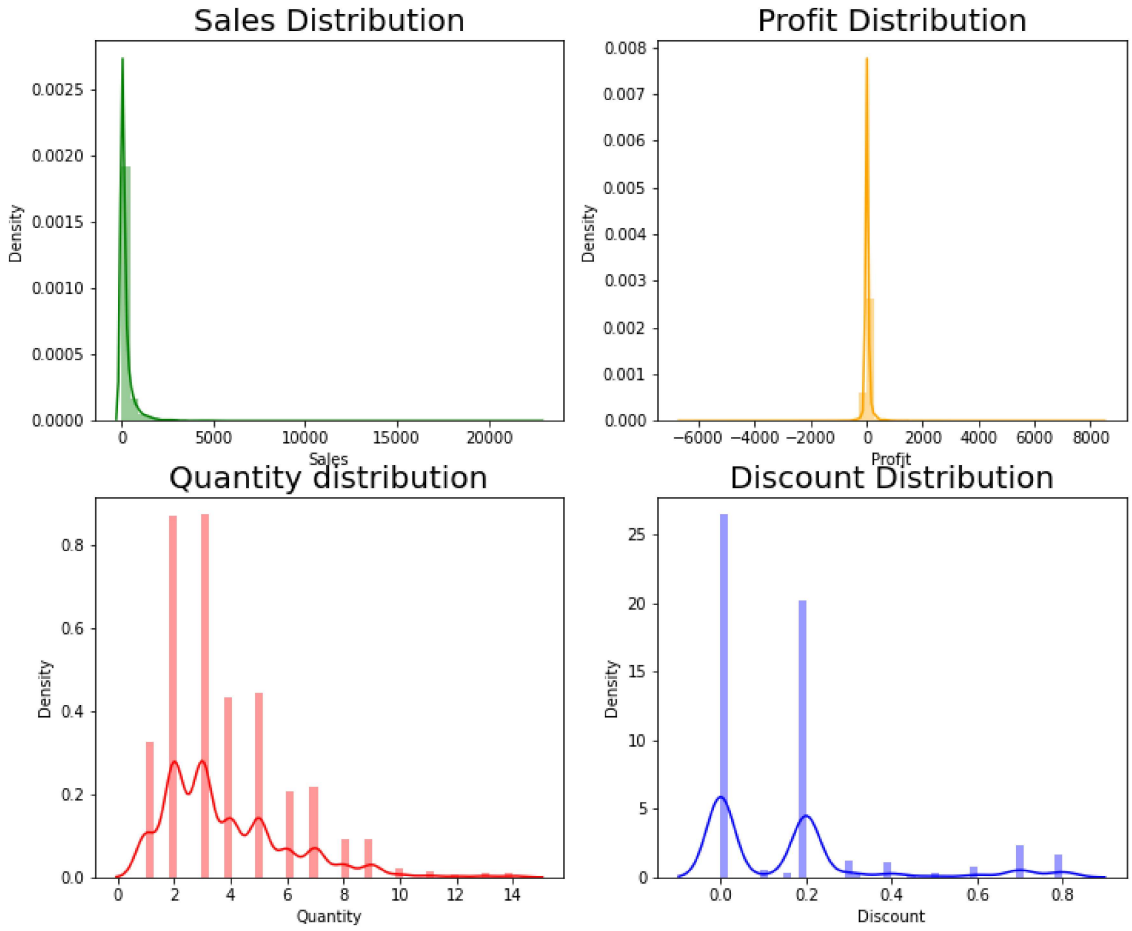
```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\_decorators.py:36: FutureWarning: Pass the following variable as a keywo
rd arg: x. From version 0.12, the only valid positional argument will be `da
ta`, and passing other arguments without an explicit keyword will result in
an error or misinterpretation.
  warnings.warn(
```



In [85]:

```
fig, axs = plt.subplots(ncols=2, nrows = 2, figsize = (12,10))
sns.distplot(data_set['Sales'], color = 'green',  ax = axs[0][0])
sns.distplot(data_set['Profit'], color = 'orange',  ax = axs[0][1])
sns.distplot(data_set['Quantity'], color = 'red',  ax = axs[1][0])
sns.distplot(data_set['Discount'], color = 'blue',  ax = axs[1][1])
axs[0][0].set_title('Sales Distribution', fontsize = 20)
axs[0][1].set_title('Profit Distribution', fontsize = 20)
axs[1][0].set_title('Quantity distribution', fontsize = 20)
axs[1][1].set_title('Discount Distribution', fontsize = 20)
plt.show()
```

```
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\distributions.py:2557: FutureWarning: `distplot` is a deprecated functio
n and will be removed in a future version. Please adapt your code to use eit
her `displot` (a figure-level function with similar flexibility) or `histplo
t` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\distributions.py:2557: FutureWarning: `distplot` is a deprecated functio
n and will be removed in a future version. Please adapt your code to use eit
her `displot` (a figure-level function with similar flexibility) or `histplo
t` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\distributions.py:2557: FutureWarning: `distplot` is a deprecated functio
n and will be removed in a future version. Please adapt your code to use eit
her `displot` (a figure-level function with similar flexibility) or `histplo
t` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
c:\users\shray\appdata\local\programs\python\python39\lib\site-packages\seab
orn\distributions.py:2557: FutureWarning: `distplot` is a deprecated functio
n and will be removed in a future version. Please adapt your code to use eit
her `displot` (a figure-level function with similar flexibility) or `histplo
t` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



In [86]:

```
df=data_set['State'].value_counts()
df.head(10)
```

Out[86]:

```
California        2001
New York          1128
Texas              985
Pennsylvania       587
Washington         506
Illinois           492
Ohio               469
Florida            383
Michigan           255
North Carolina     249
Name: State, dtype: int64
```

In [87]:

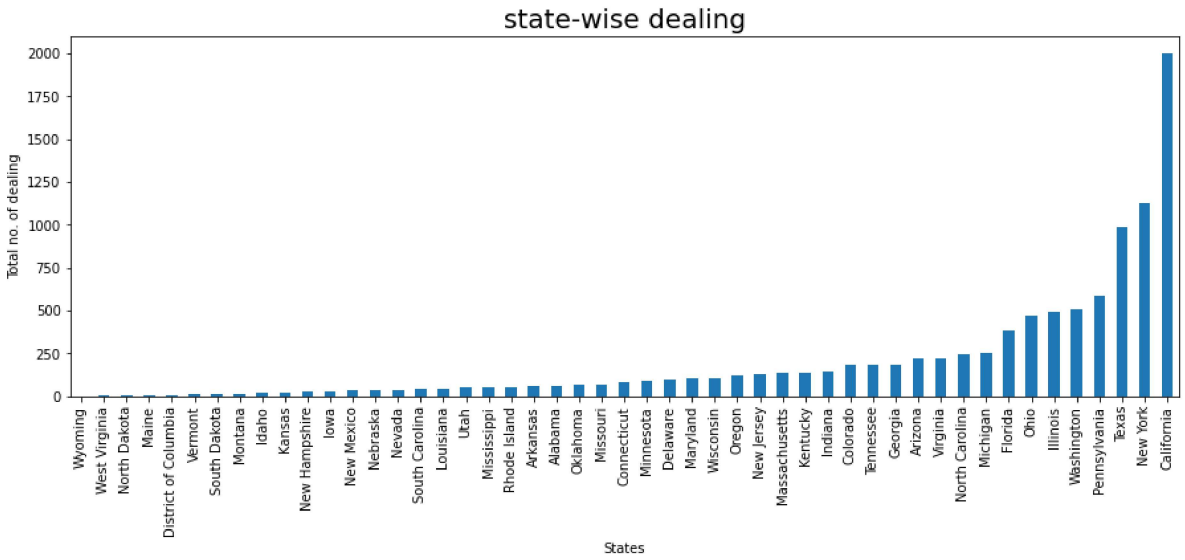
```
df_states=data_set.groupby(['State'])[['Sales','Discount','Profit']].mean()
df_states.head(10)
```

Out[87]:

State	Sales	Discount	Profit
Alabama	319.846557	0.000000	94.865989
Arizona	157.508933	0.303571	-15.303235
Arkansas	194.635500	0.000000	66.811452
California	228.729451	0.072764	38.171608
Colorado	176.418231	0.316484	-35.867351
Connecticut	163.223866	0.007317	42.823071
Delaware	285.948635	0.006250	103.930988
District of Columbia	286.502000	0.000000	105.958930
Florida	233.612815	0.299347	-8.875461
Georgia	266.825217	0.000000	88.315453

In [88]:

```
df_state = data_set.groupby('State')['Quantity'].count().sort_values(ascending=True).plot.b
plt.title('state-wise dealing',fontsize=20)
plt.xlabel('States')
plt.ylabel('Total no. of dealing')
plt.show()
```

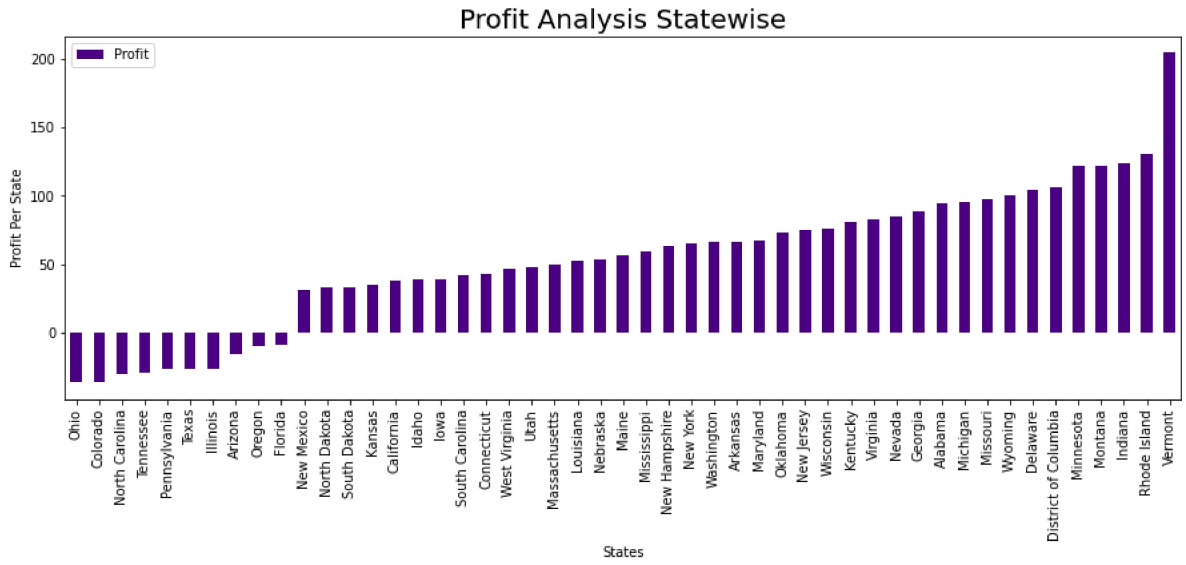


In [89]:

```
#Profit Analysis Statewise
df1=df_states.sort_values('Profit')
df1['Profit'].plot(kind= 'bar',figsize=(15,5),color='indigo')
plt.xlabel('States')
plt.ylabel('Profit Per State')
plt.title('Profit Analysis Statewise', fontsize=20)
plt.legend()
```

Out[89]:

```
<matplotlib.legend.Legend at 0x2613c641e80>
```

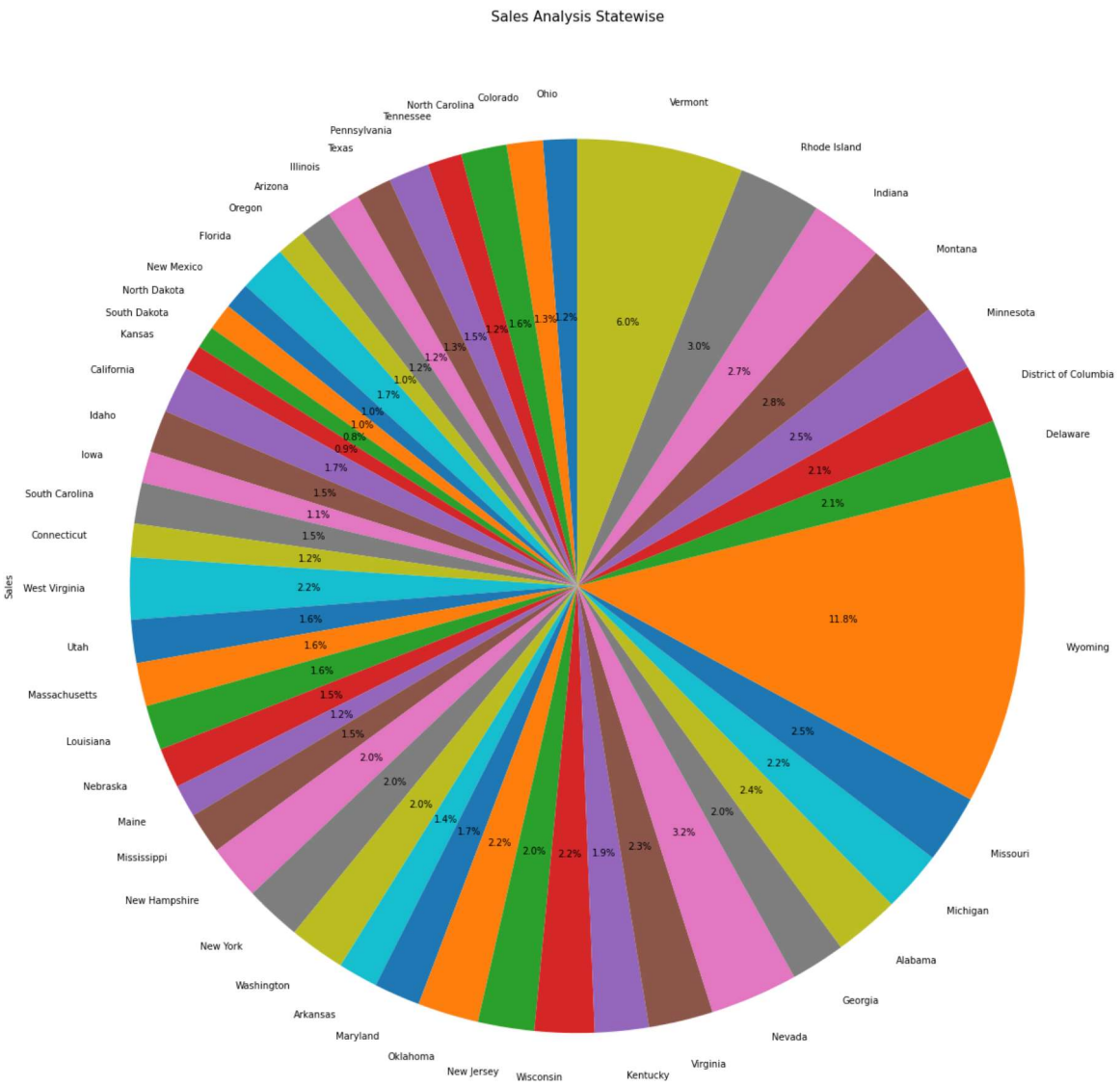


In [90]:

```
df1['Sales'].plot(kind= 'pie',figsize=(22,22),autopct='%1.1f%%',startangle=90)
plt.title('Sales Analysis Statewise', fontsize=15)
```

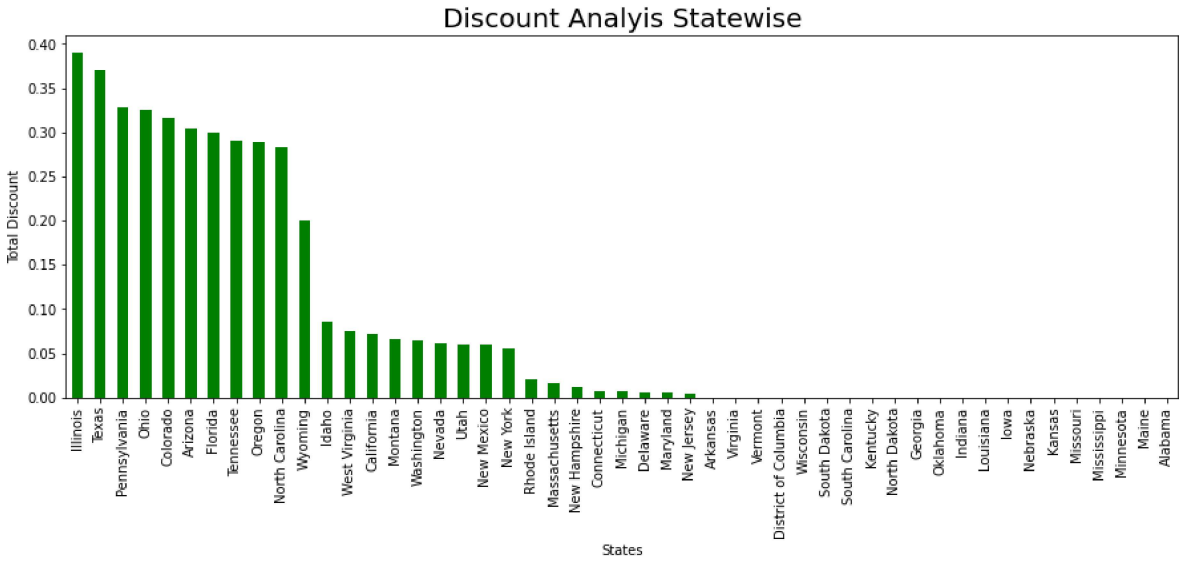
Out[90]:

```
Text(0.5, 1.0, 'Sales Analysis Statewise')
```



In [91]:

```
df_discount = data_set.groupby('State')['Discount'].mean().sort_values(ascending=False).plo
plt.title('Discount Analyis Statewise',fontsize=20)
plt.xlabel('States')
plt.ylabel('Total Discount')
plt.show()
```



In [92]:

```
segment=data_set.Segment.value_counts().reset_index()
segment.columns=("Segment","Count")
segment
```

Out[92]:

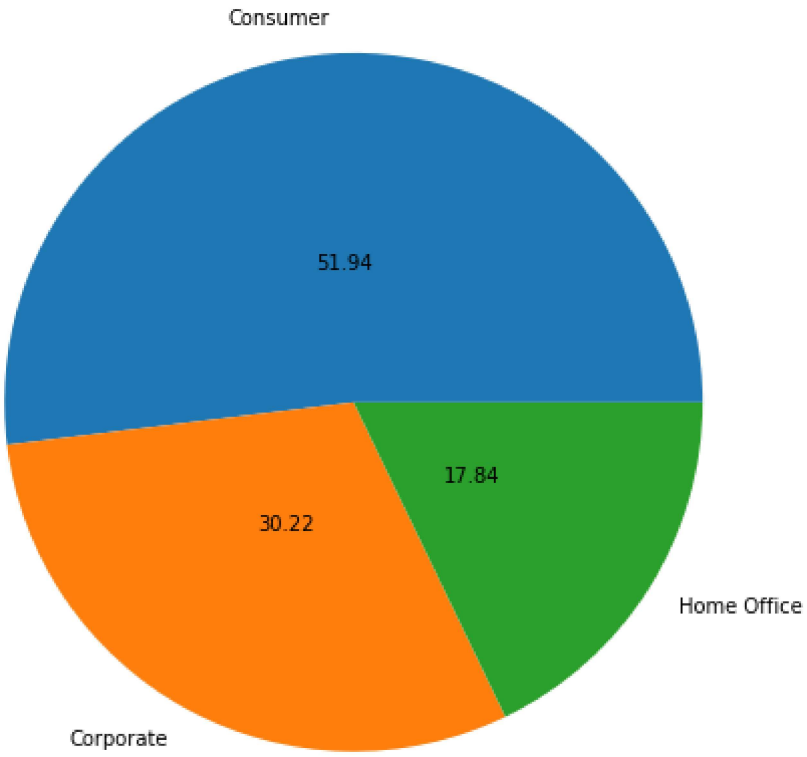
	Segment	Count
0	Consumer	5191
1	Corporate	3020
2	Home Office	1783

In [93]:

```
plt.pie(x="Count",labels="Segment",data=segment,radius=2,autopct="%.2f",pctdistance=0.4)
```

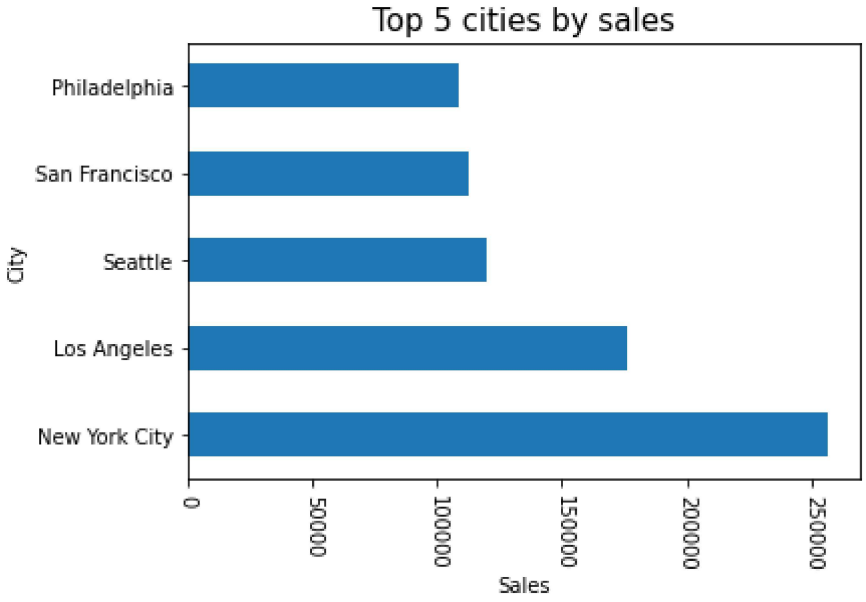
Out[93]:

```
([<matplotlib.patches.Wedge at 0x2613ef43460>,
  <matplotlib.patches.Wedge at 0x2613ef43be0>,
  <matplotlib.patches.Wedge at 0x2613ef4f340>],
 [Text(-0.13408036713364035, 2.1959103932422446, 'Consumer'),
  Text(-1.0537686579632295, -1.9312098838537397, 'Corporate'),
  Text(1.8633972391176687, -1.1695087555245793, 'Home Office')],
 [Text(-0.048756497139505584, 0.798512870269907, '51.94'),
  Text(-0.3831886028957198, -0.7022581395831781, '30.22'),
  Text(0.6775989960427886, -0.425275911099847, '17.84')])
```



In [94]:

```
#Top 5 cities by Sales,Profit,Discount
total_sales = data_set.groupby('City')['Sales'].sum()
top_5_cities = total_sales.sort_values(ascending = False).iloc[0:5]
top_5_cities.plot(kind = 'barh')
plt.title('Top 5 cities by sales',fontsize=15)
plt.xticks(rotation=270)
plt.xlabel('Sales')
plt.show()
```

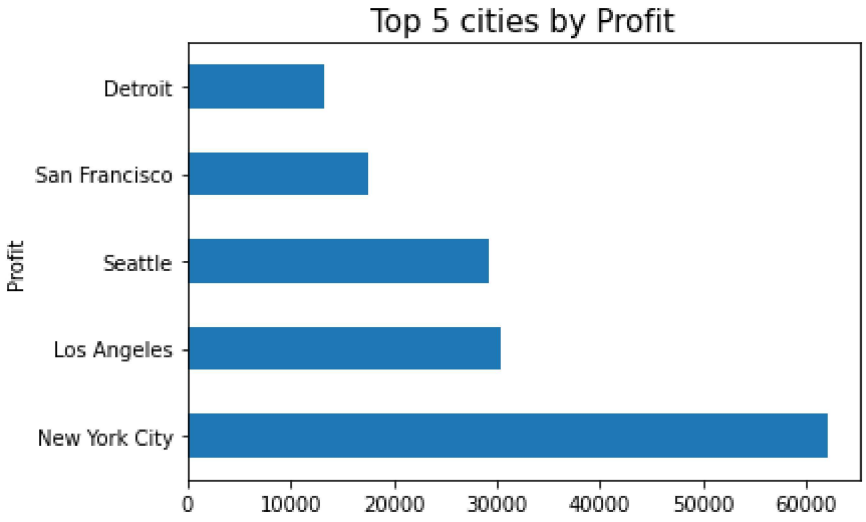


In [95]:

```
total_profit = data_set.groupby('City')['Profit'].sum()
top_5_cities = total_profit.sort_values(ascending= False).iloc[0:5]
top_5_cities.plot(kind = 'barh')
plt.title('Top 5 cities by Profit',fontsize=15)
plt.ylabel('Profit')
```

Out[95]:

```
Text(0, 0.5, 'Profit')
```

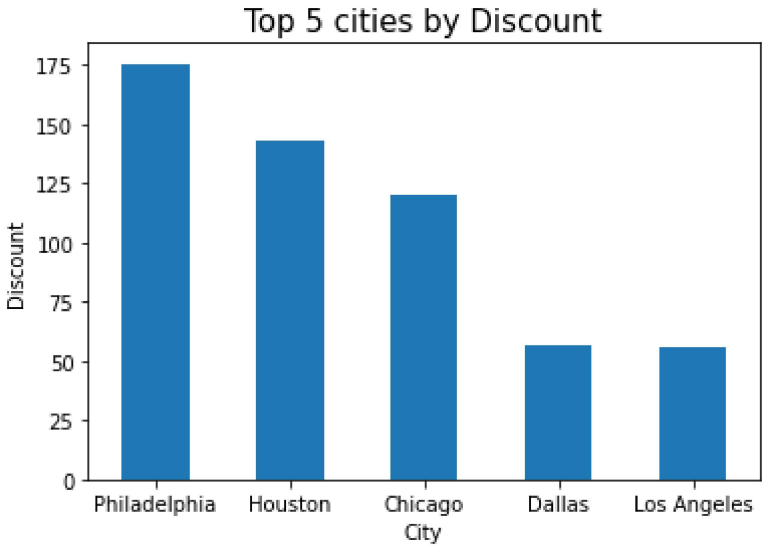


In [96]:

```
total_profit = data_set.groupby('City')['Discount'].sum()
top_5_cities = total_profit.sort_values(ascending= False).iloc[0:5]
top_5_cities.plot(kind = 'bar')
plt.title('Top 5 cities by Discount',fontsize=15)
plt.xticks(rotation=0)
plt.ylabel('Discount')
```

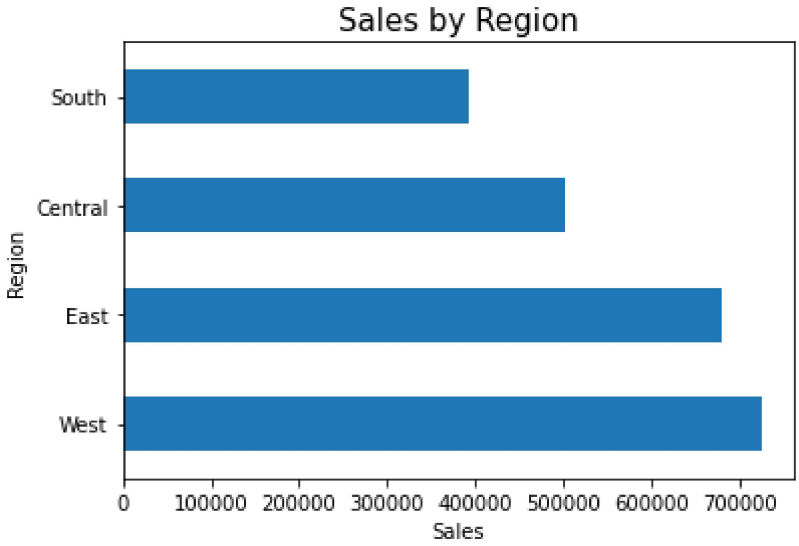
Out[96]:

```
Text(0, 0.5, 'Discount')
```



In [97]:

```
total_sales = data_set.groupby('Region')['Sales'].sum()
region_sales = total_sales.sort_values(ascending = False).iloc[0:4]
region_sales.plot(kind ='barh')
plt.title('Sales by Region',fontsize=15)
plt.xlabel('Sales')
plt.show()
```

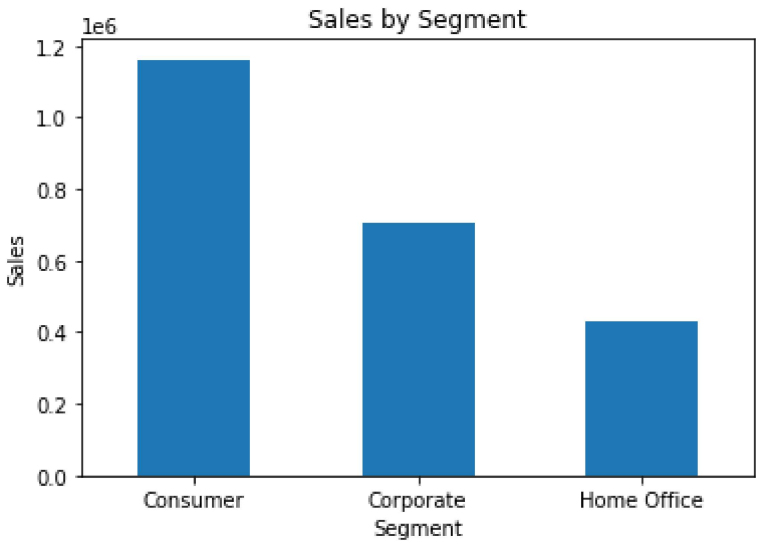


In [98]:

```
total_profit = data_set.groupby('Segment')['Sales'].sum()
segment_sales = total_profit.sort_values(ascending= False).iloc[0:4]
segment_sales.plot(kind = 'bar',fontsize=10)
plt.title('Sales by Segment')
plt.xticks(rotation=0)
plt.ylabel('Sales')
```

Out[98]:

```
Text(0, 0.5, 'Sales')
```



Thank you so much. Please give your valuable feedback in the comment box below.

Thank you.....