The Sparks Foundation

DataScienceAndBusinessAnalytics

DoneBy : GajalS

TASK 3

- Perform ‘Exploratory Data Analysis’ on dataset ‘SampleSuperstore’
- As a business manager, try to find out the weak areas where you can work to make more profit.
- What all business problems you can derive by exploring the data?

Importing libraries based on the requirements

In [1]:

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

Reading the csv dataset using the command pd.read_csv()

In [2]:

```
data = pd.read_csv("SampleSuperstore.csv")
data.head()
```

Out[2]:

	Ship Mode	Segment	Country	City	State	Postal_Code	Region	Category	Su Catego
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	Labe
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tabl
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Stora

In [3]:

```
data.tail()
```

Out[3]:

	Ship Mode	Segment	Country	City	State	Postal_Code	Region	Category	(
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Fu
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Fu
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	A

Performing Exploratory Data Analysis

In [4]:

```
data.shape
```

Out[4]:

```
(9994, 13)
```

To Generate descriptive statistics

In [5]:

```
data.describe()
```

Out[5]:

	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

To Print a concise summary of a DataFrame

In [6]:

```
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
```

Detecting missing values

In [7]:

```
data.isna().sum()
```

Out[7]:

```
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
```

Count distinct observations over each columns

In [8]:

```
data.nunique()
```

Out[8]:

```
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
```

To Compute pairwise correlation of columns

In [9]:

```
data.corr()
```

Out[9]:

	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

Duplicated values in the data

In [10]:

```
data.duplicated().sum()
```

Out[10]:

```
17
```

Dropping the duplicate values

In [11]:

```
data.drop_duplicates(inplace = True)
```

In [12]:

```
data.shape
```

Out[12]:

```
(9977, 13)
```

Data Visualization

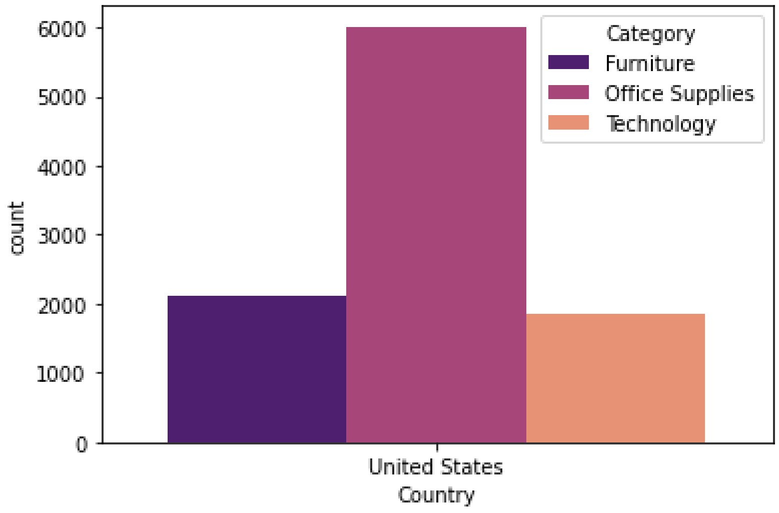
Which Category has the highest supply in United States?

In [13]:

```
sns.countplot(x = "Country", hue = "Category" , data = data, palette = "magma")
```

Out[13]:

```
<AxesSubplot:xlabel='Country', ylabel='count'>
```



- From the above plot, we can see that Office Supplies have better sales than Furniture and technology in United States

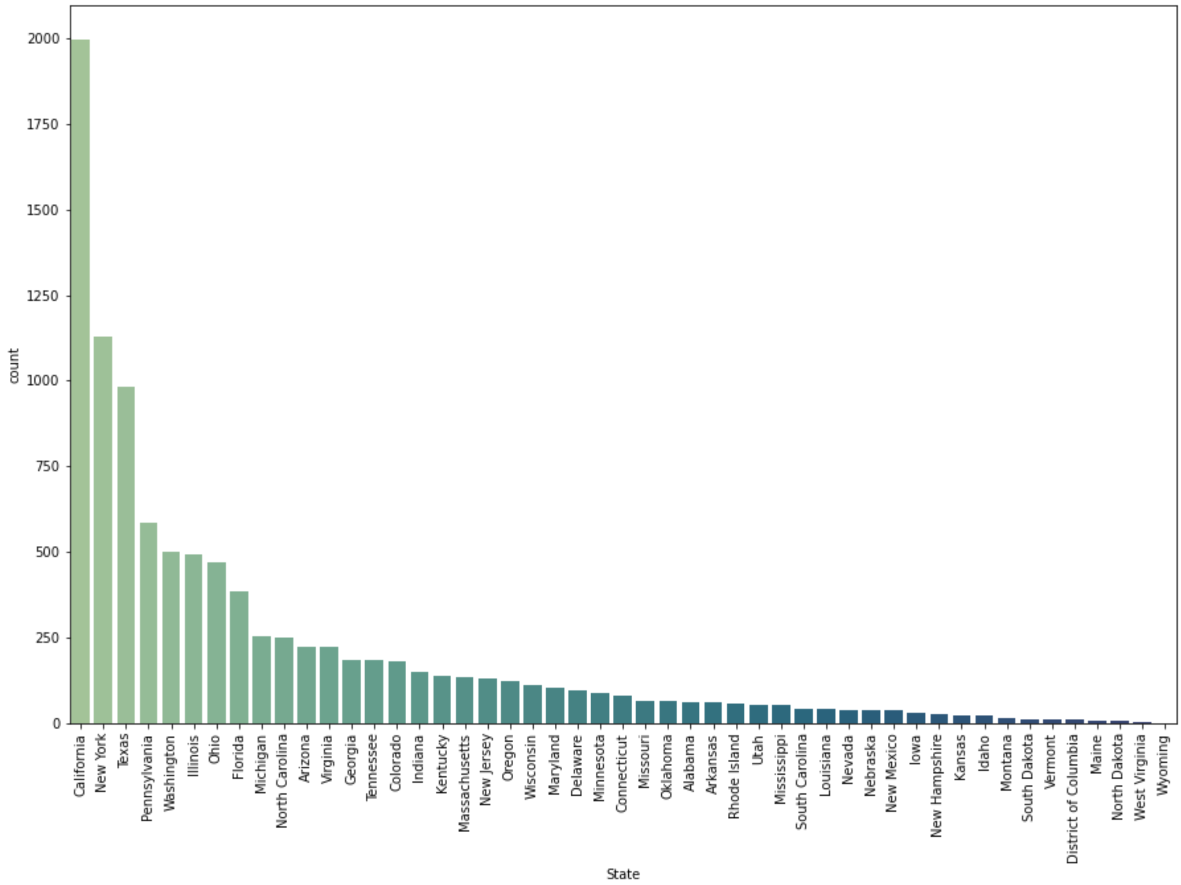
Which city has the most and the least sales in United States?

In [14]:

```
plt.figure(figsize=(15,10))
sns.countplot(x = "State", data = data,palette="crest",order = data["State"].value_counts()
plt.xticks(rotation = 90)
```

Out[14]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
        34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]),
 [Text(0, 0, 'California'),
  Text(1, 0, 'New York'),
  Text(2, 0, 'Texas'),
  Text(3, 0, 'Pennsylvania'),
  Text(4, 0, 'Washington'),
  Text(5, 0, 'Illinois'),
  Text(6, 0, 'Ohio'),
  Text(7, 0, 'Florida'),
  Text(8, 0, 'Michigan'),
  Text(9, 0, 'North Carolina'),
  Text(10, 0, 'Arizona'),
  Text(11, 0, 'Virginia'),
  Text(12, 0, 'Georgia'),
  Text(13, 0, 'Tennessee'),
  Text(14, 0, 'Colorado'),
  Text(15, 0, 'Indiana'),
  Text(16, 0, 'Kentucky'),
  Text(17, 0, 'Massachusetts'),
  Text(18, 0, 'New Jersey'),
  Text(19, 0, 'Oregon'),
  Text(20, 0, 'Wisconsin'),
  Text(21, 0, 'Maryland'),
  Text(22, 0, 'Delaware'),
  Text(23, 0, 'Minnesota'),
  Text(24, 0, 'Connecticut'),
  Text(25, 0, 'Missouri'),
  Text(26, 0, 'Oklahoma'),
  Text(27, 0, 'Alabama'),
  Text(28, 0, 'Arkansas'),
  Text(29, 0, 'Rhode Island'),
  Text(30, 0, 'Utah'),
  Text(31, 0, 'Mississippi'),
  Text(32, 0, 'South Carolina'),
  Text(33, 0, 'Louisiana'),
  Text(34, 0, 'Nevada'),
  Text(35, 0, 'Nebraska'),
  Text(36, 0, 'New Mexico'),
  Text(37, 0, 'Iowa'),
  Text(38, 0, 'New Hampshire'),
  Text(39, 0, 'Kansas'),
  Text(40, 0, 'Idaho'),
  Text(41, 0, 'Montana'),
  Text(42, 0, 'South Dakota'),
  Text(43, 0, 'Vermont'),
  Text(44, 0, 'District of Columbia'),
  Text(45, 0, 'Maine'),
  Text(46, 0, 'North Dakota'),
  Text(47, 0, 'West Virginia'),
  Text(48, 0, 'Wyoming')])
```



From the above cell, it is clear that:

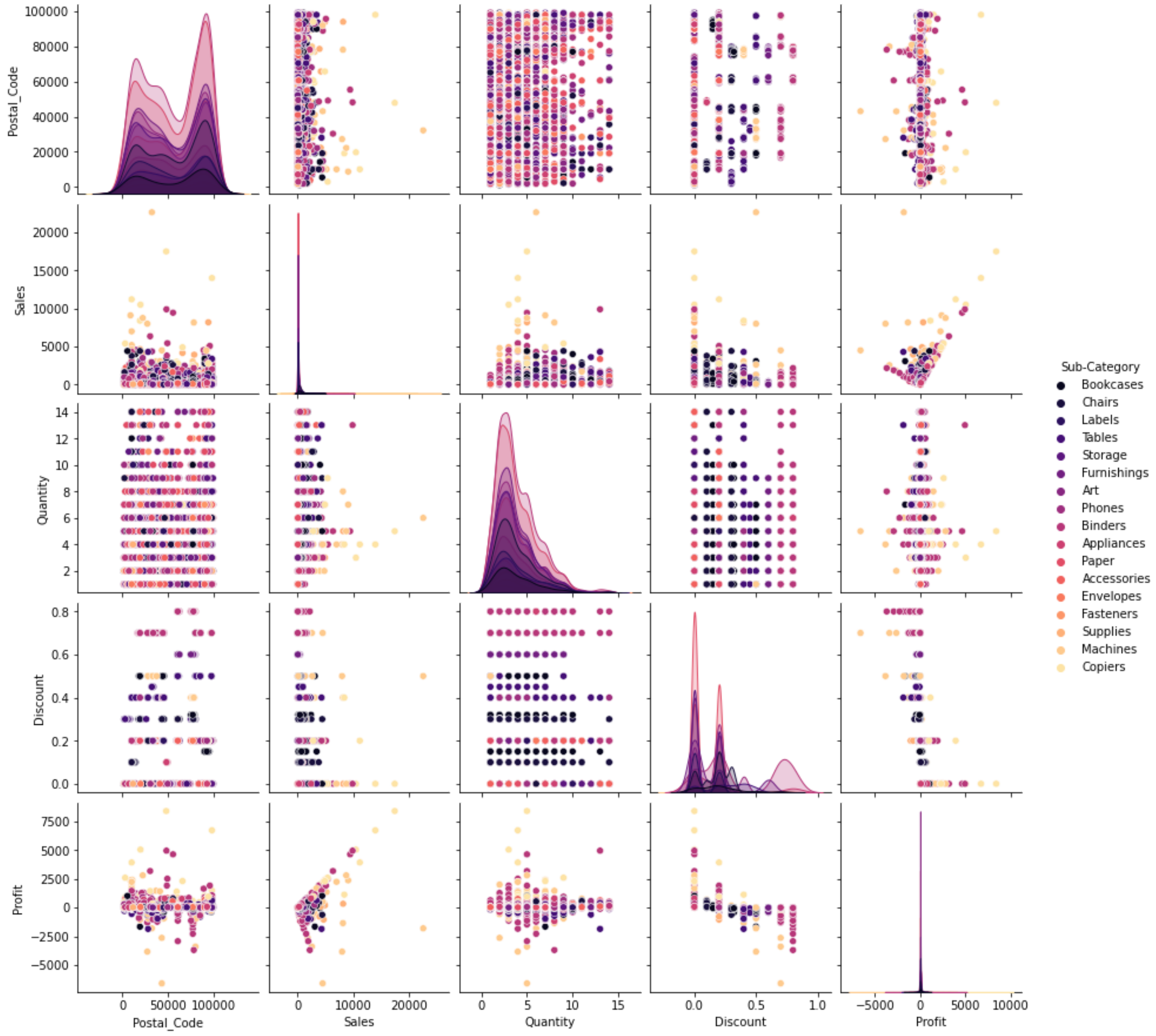
- Sales are in its peak in California and New York
- The least buyers are from Wyoming

In [15]:

```
sns.pairplot(data, hue ='Sub-Category', palette = "magma")
```

Out[15]:

```
<seaborn.axisgrid.PairGrid at 0x296cd63b5b0>
```



In [16]:

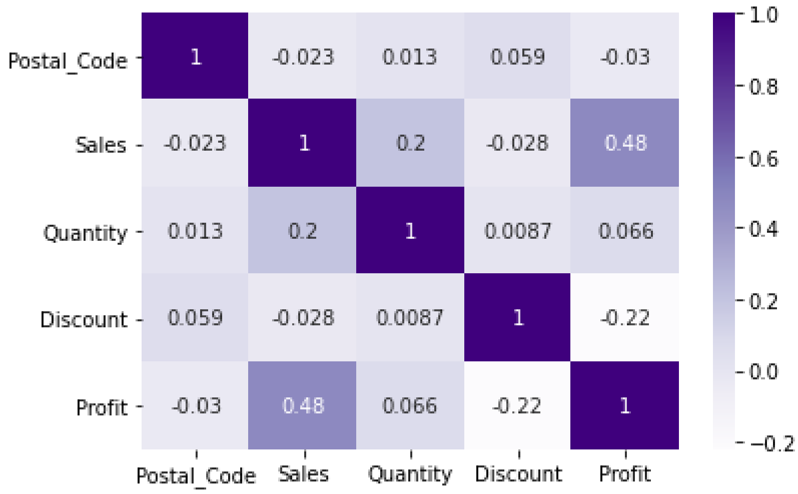
```
corr_data = data.corr()
```

In [17]:

```
sns.heatmap(data = corr_data, annot = True, cmap = "Purples")
```

Out[17]:

```
<AxesSubplot:>
```



The above heatmap shows the pairwise correlation of columns

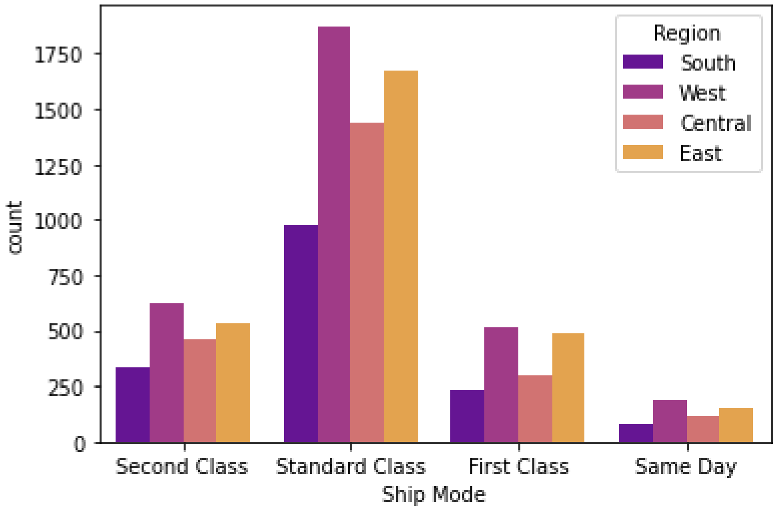
Which mode of Shipping is the most preferred?

In [18]:

```
sns.countplot(x = "Ship Mode", hue = "Region", data = data, palette = "plasma")
```

Out[18]:

```
<AxesSubplot:xlabel='Ship Mode', ylabel='count'>
```



Most preferred mode of Shipping and highest sales in region:

- The sales in the west are the highest according to the above countplot.
- Standard Class is the most preferred mode of shipping.

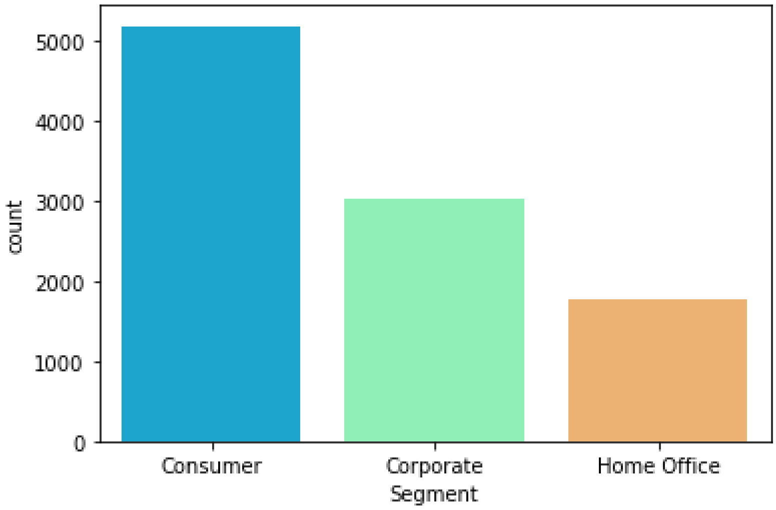
Which Segment is the most preferred by customers?

In [19]:

```
sns.countplot(x = "Segment", palette = "rainbow", data = data)
```

Out[19]:

```
<AxesSubplot:xlabel='Segment', ylabel='count'>
```



- The above plot show that the segment "Consumer" is the most preferrable, while Home Office and corporate are comparitvely very less.

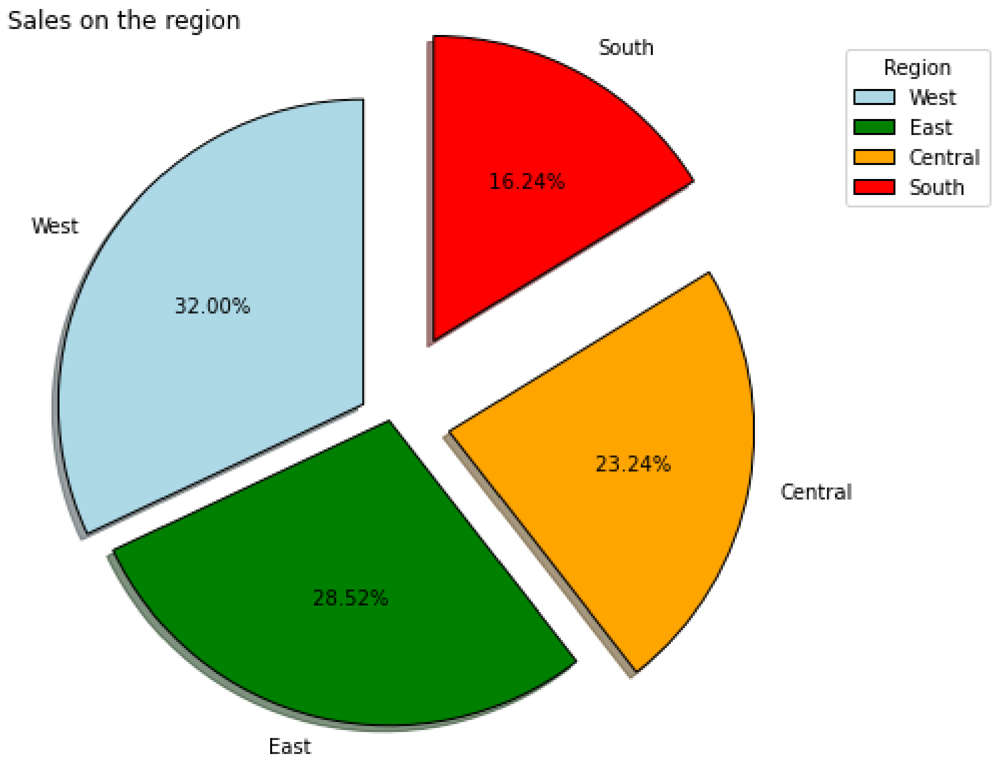
Which Region has the highest sale?

In [20]:

```
plt.figure(figsize =(12, 7))
explode = (0.1, 0.0, 0.2, 0.3)

wp = { 'linewidth' : 1, 'edgecolor' : "black" }
regions = ['West', 'East', 'Central', 'South']
colors = ("lightblue","green", "orange", "red")
region_data = data["Region"].value_counts()

plt.pie(region_data, explode = explode, autopct = "% .2f%%",labels = regions, colors = colo
plt.title("Sales on the region", loc = "left")
plt.legend(regions, title ="Region", bbox_to_anchor =(0.8, 0.5, 0.5, 0.5))
plt.show()
```



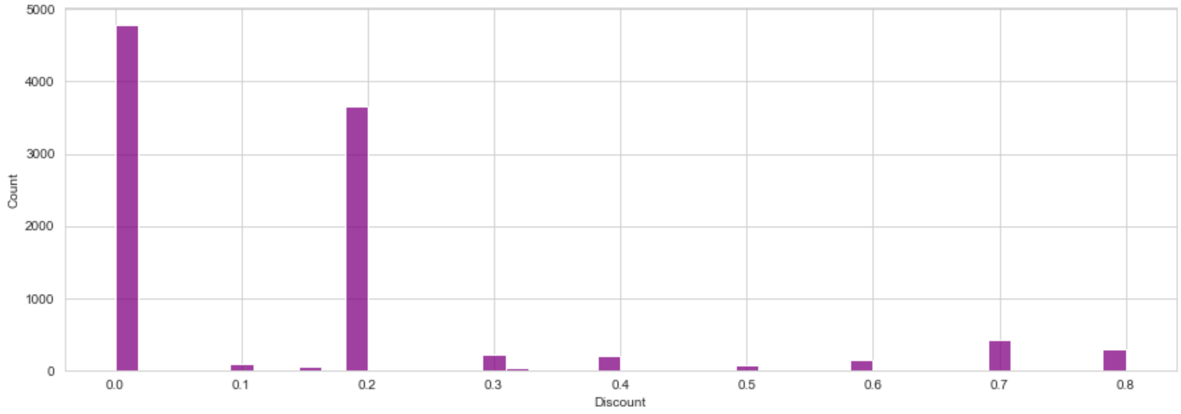
What percentage of the discount is the most given?

In [21]:

```
plt.figure(figsize = (15,5))
sns.set_style("whitegrid")
sns.histplot(data["Discount"], color ='purple')
```

Out[21]:

```
<AxesSubplot:xlabel='Discount', ylabel='Count'>
```



The above plot shows that 0 to 20% is the highest given discount.

In [22]:

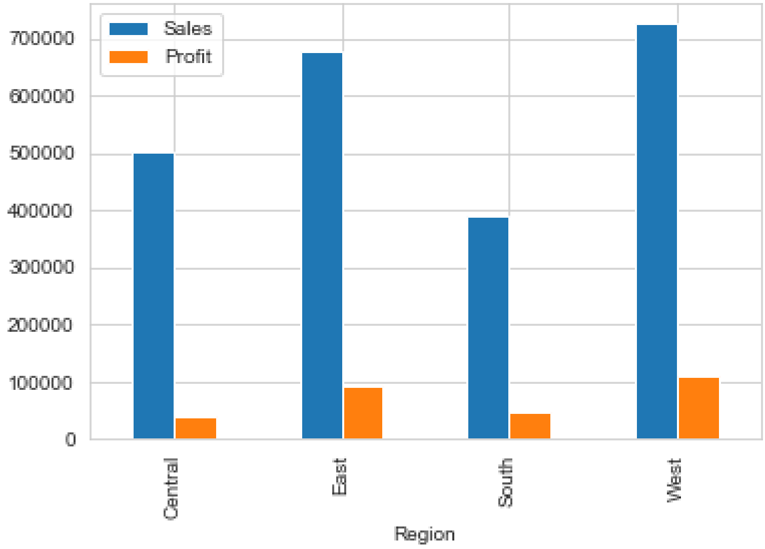
```
sales_profit_data = data.groupby("Region")[["Sales","Profit"]].sum()
```

In [23]:

```
sales_profit_data.plot.bar()
```

Out[23]:

```
<AxesSubplot:xlabel='Region'>
```



In [24]:

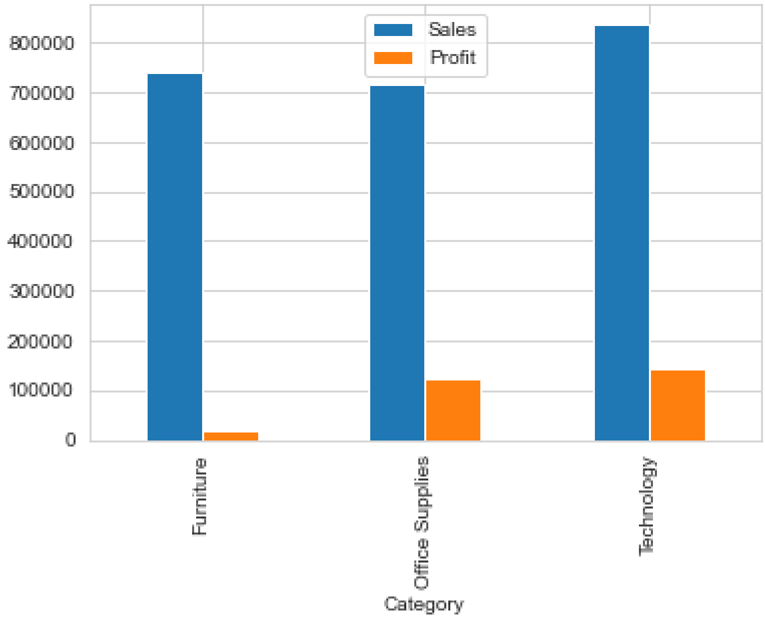
```
Category_profit = data.groupby("Category")[["Sales","Profit"]].sum()
```

In [25]:

```
Category_profit.plot.bar()
```

Out[25]:

```
<AxesSubplot:xlabel='Category'>
```



In [26]:

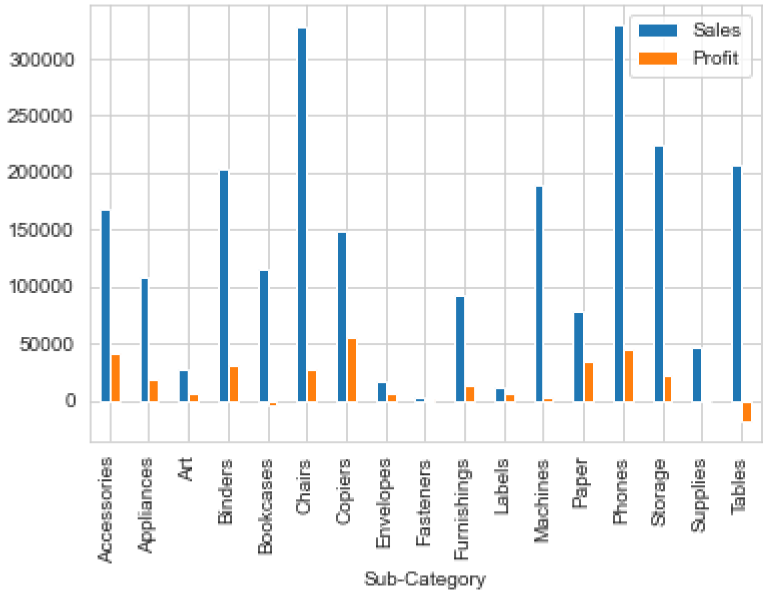
```
Sub_Category_profit = data.groupby("Sub-Category")[["Sales","Profit"]].sum()
```

In [27]:

```
Sub_Category_profit.plot.bar()
```

Out[27]:

```
<AxesSubplot:xlabel='Sub-Category'>
```

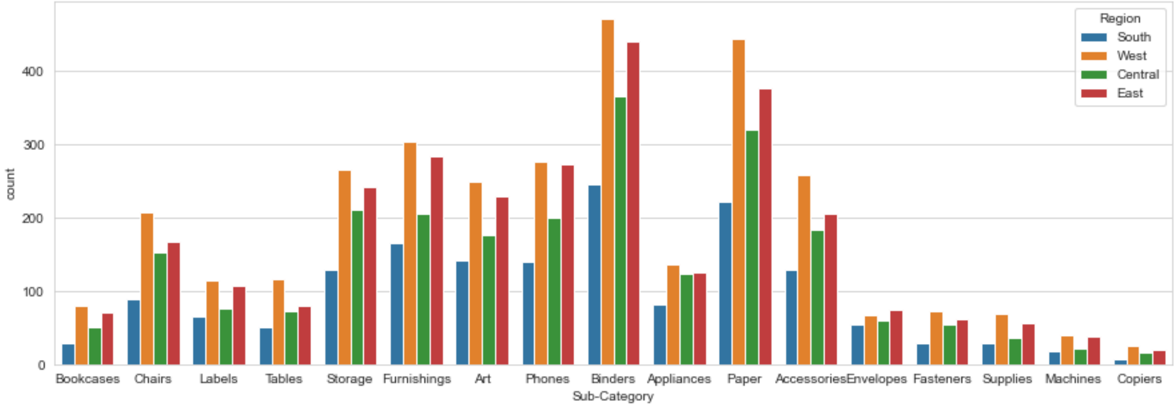


In [28]:

```
plt.figure(figsize = (15,5))
sns.countplot(x = "Sub-Category", hue = "Region", data = data)
```

Out[28]:

```
<AxesSubplot:xlabel='Sub-Category', ylabel='count'>
```

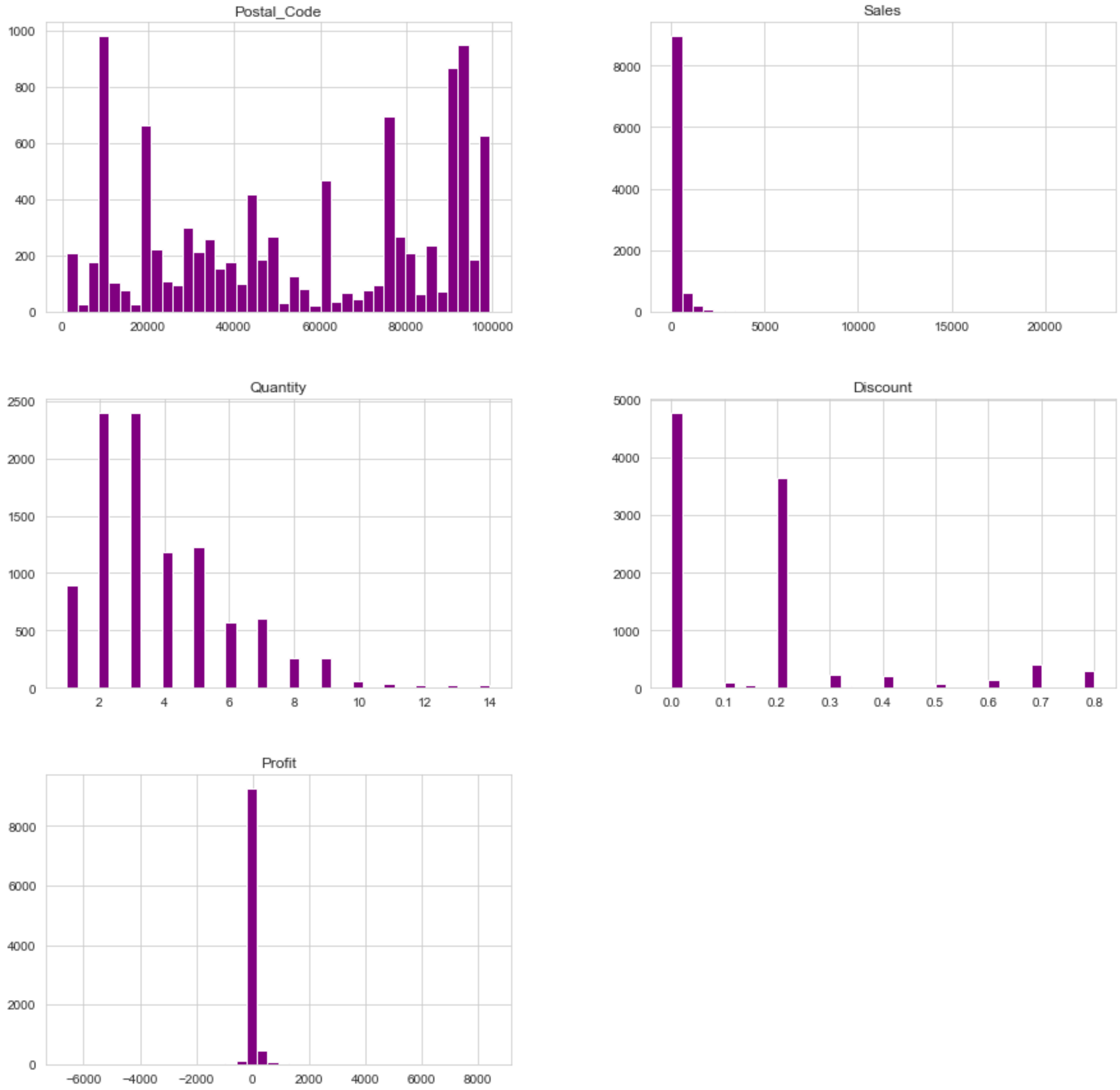


In [29]:

```
data.hist(figsize=(15,15), bins = 40, color = "purple")
```

Out[29]:

```
array([[<AxesSubplot:title={'center':'Postal_Code'}>,
        <AxesSubplot:title={'center':'Sales'}>],
       [<AxesSubplot:title={'center':'Quantity'}>,
        <AxesSubplot:title={'center':'Discount'}>],
       [<AxesSubplot:title={'center':'Profit'}>, <AxesSubplot:>]],
      dtype=object)
```



Observations/Buisness Problems

- Office Supplies are most bought than Furniture and technology
- Sates California and New York has the highest sales
- Wyoming has the least sales
- Customer's most bought quantity is 2 and 3
- Standard Class is the most preferred mode of shipping
- Consumer is the most preferrable Segment
- West region has the highest sale
- 0 - 20% is the most given discount
- Papers and Binders are the most bought in Sub-category(mainly in the west region)
- There is no to very less profit in furnitures
- All in all, the sales are high in all the region while the profit is comparitively less

Conclusion

- South region has the least sales
- Tables are Bookcases are in loss, no profit
- Office supplies are the most bought by the customers
- There are more loss than profit, to resolve this we must focus on the city/product/segment/mode of shipping and procced the sales.
- In this case Office supplies are most bought, west region has the highest sales, counsumer is the most preferred segment, Standand class is the most preferred mode of shipping

Thank you!