Author Yatam Sumalatha

Task 3 Exploratory Data Analysis (EDA)

GRIP @ The Sparks Foundation

Importing Labraies

In [1]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
```

Dataset loading

In [3]:
```
df=pd.read_csv(r'C:\Users\ADMIN\Desktop\Sparks Foundation Internship\Datasets\SampleSuperstore.csv')
```

In [4]:
```
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]:
```
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]:
```
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]:
```
df.shape
```

Out[7]:
```
(9994, 13)
```

In [8]:
```
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]:
```
df.describe()
```

Out[9]:

	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 [10]:
```
df.duplicated().sum()
```

Out[10]:
```
17
```

In [11]:
```
superstore=df.drop_duplicates()
superstore.shape
```

Out[11]: (9977, 13)

Total superstore sales

In [12]:
```
print('Total Sales:')
print(superstore['Sales'].sum())
```

```
Total Sales:
2296195.5903
```

Superstore profit

In [13]:
```
print('Total Profit:')
print(superstore['Profit'].sum())
```

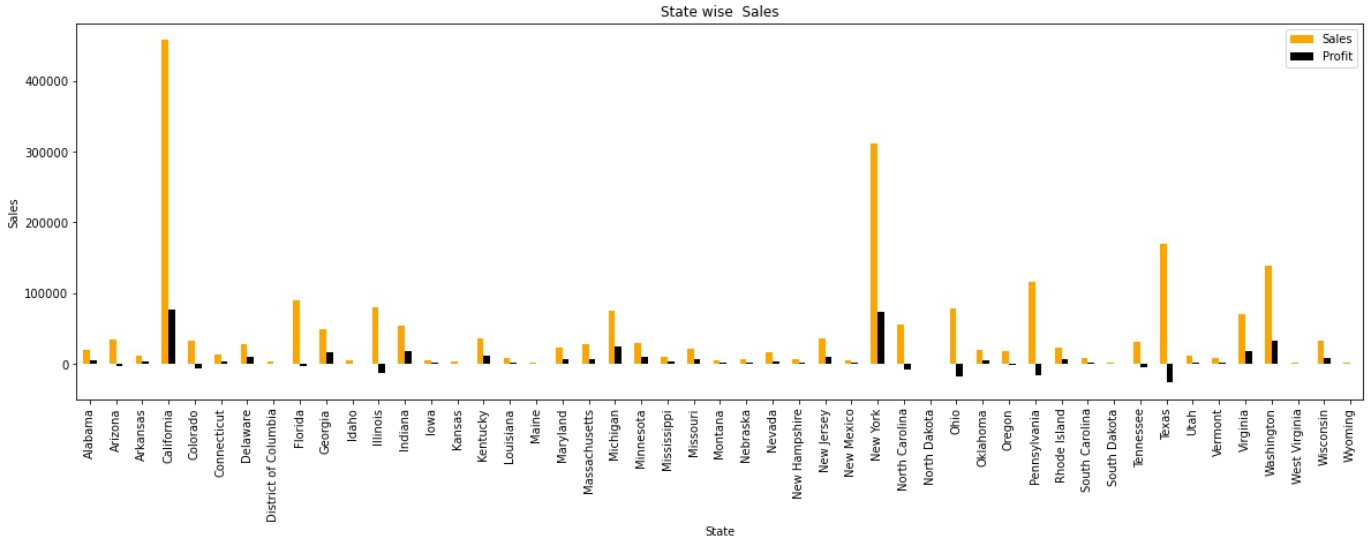
```
Total Profit:
286241.4226
```

Exploratory Data Analysis

Data Visualization

In [14]:
```
superstore.groupby(['State'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(20,6))
plt.title('State wise  Sales')
plt.ylabel('Sales')
plt.xlabel('State')
plt.show()
```

```
C:\Users\ADMIN\AppData\Local\Temp/ipykernel_42384/2083496388.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  superstore.groupby(['State'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(20,6))
```

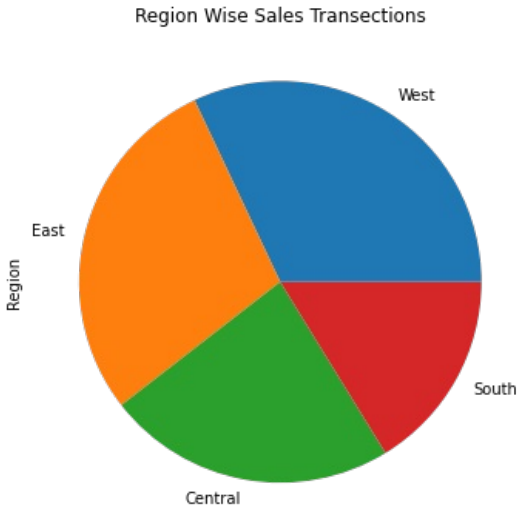


Sales and profit (State Wise)

Highest -- California Profit --- california & Newyork

Region Wise Sales Transactions

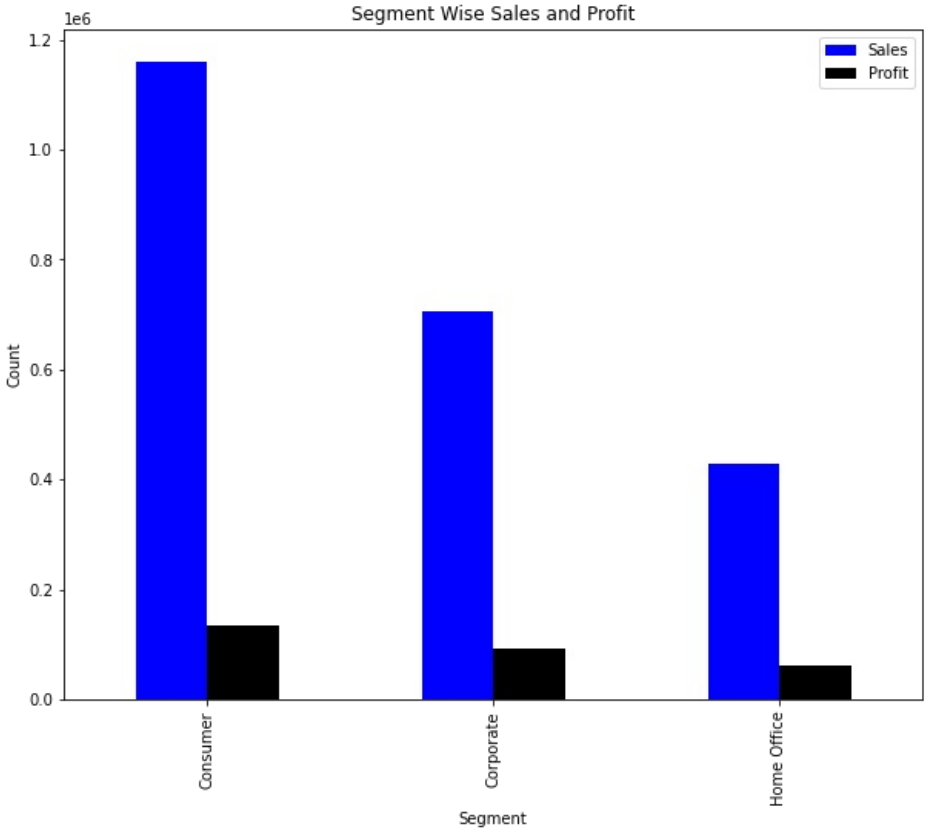
In [15]:
```
plt.figure(figsize=(10,6))
superstore['Region'].value_counts().plot.pie()
plt.title('Region Wise Sales Transections')
plt.show()
```



Bar chart (Sales and profit)

In [19]:
```
superstore.groupby(['Segment'])['Sales','Profit'].sum().plot.bar(color=['blue','black'],figsize=(10,8))
plt.title('Segment Wise Sales and Profit')
plt.ylabel('Count')
plt.xlabel('Segment')
plt.show()
```

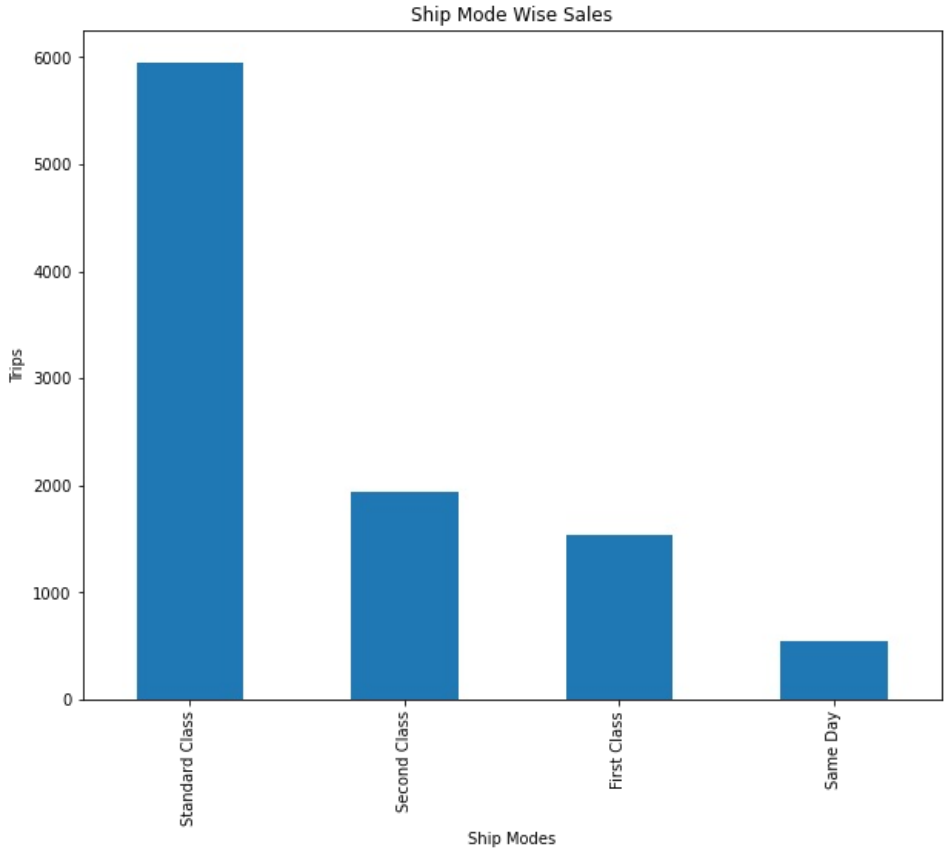
```
C:\Users\ADMIN\AppData\Local\Temp/ipykernel_42384/1735703115.py:1: FutureWarning: Indexing with multiple keys (im
plicitly converted to a tuple of keys) will be deprecated, use a list instead.
  superstore.groupby(['Segment'])['Sales','Profit'].sum().plot.bar(color=['blue','black'],figsize=(10,8))
```



Consumers has more sales and profit as compare to Corporate and Home Ofiice

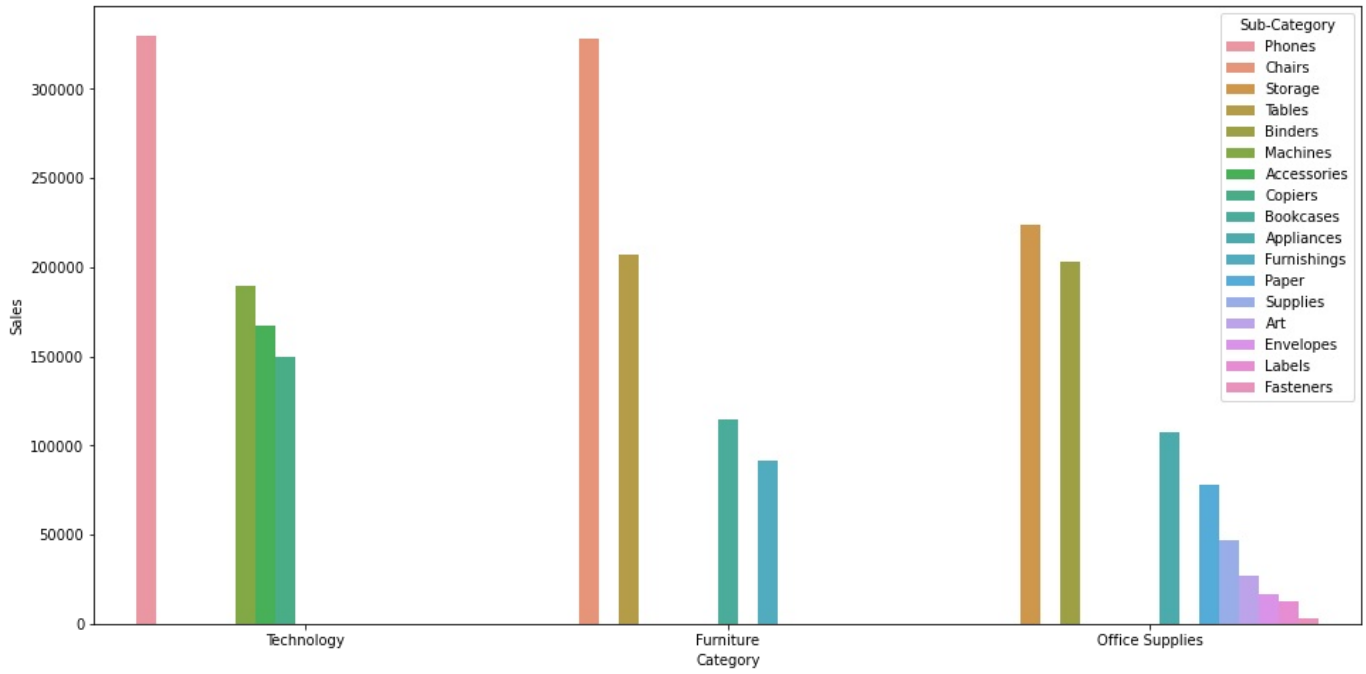
Ship Mode Wise Sales

In [21]:
```
plt.figure(figsize=(10,8))
superstore['Ship Mode'].value_counts().plot.bar()
plt.title('Ship Mode Wise Sales')
plt.ylabel('Trips')
plt.xlabel('Ship Modes')
plt.show()
```



Sales by product Category, Sub-category

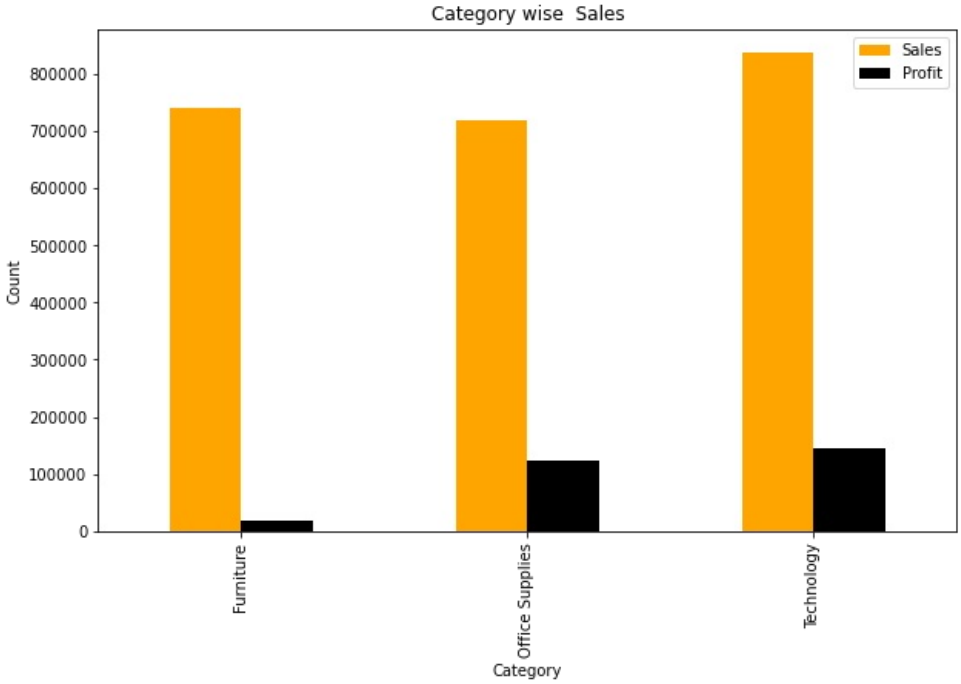
In [22]:
```
plt.figure(figsize=(16,8))
sale_category = superstore.groupby(["Category","Sub-Category"])['Sales'].aggregate(np.sum).reset_index().sort_val
sns.barplot(x = "Category", hue="Sub-Category", y= "Sales", data=sale_category)
plt.show()
```



Categories Wise Sales and Profit

In [23]:
```
superstore.groupby(['Category'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(10,6))
plt.title('Category wise  Sales')
plt.ylabel('Count')
plt.xlabel('Category')
plt.show()
```

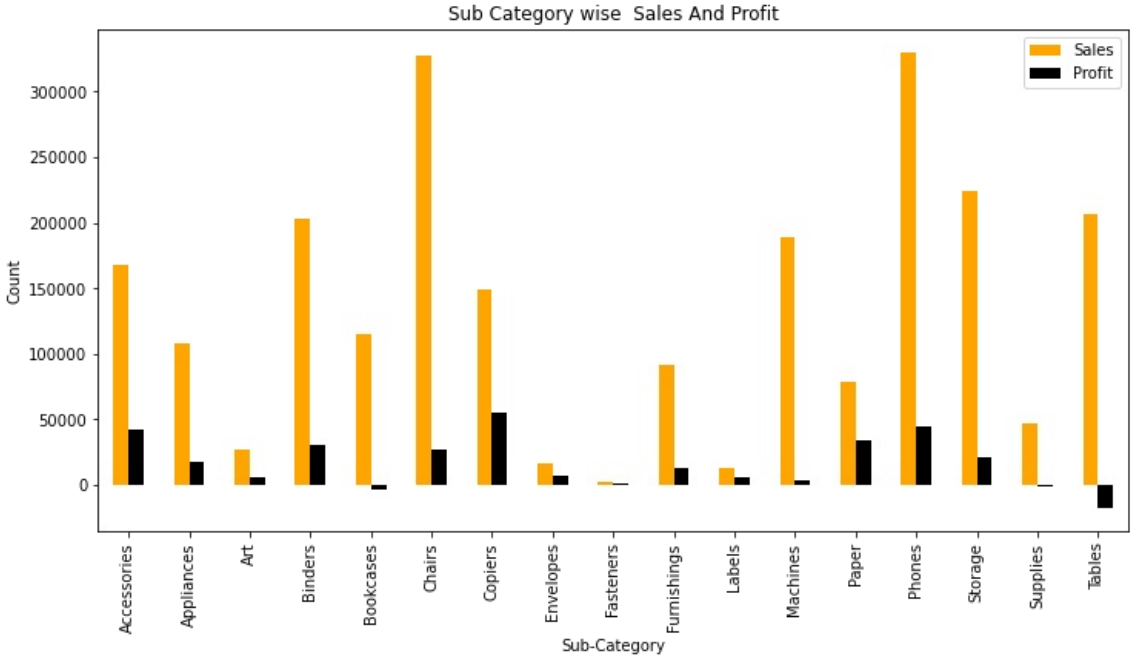
```
C:\Users\ADMIN\AppData\Local\Temp/ipykernel_42384/1445596966.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  superstore.groupby(['Category'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(10,6))
```



SUB-Categories Wise Sales and Profit

In [24]:
```
superstore.groupby(['Sub-Category'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(12,6))
plt.title('Sub Category wise  Sales And Profit')
plt.ylabel('Count')
plt.xlabel('Sub-Category')
plt.show()
```

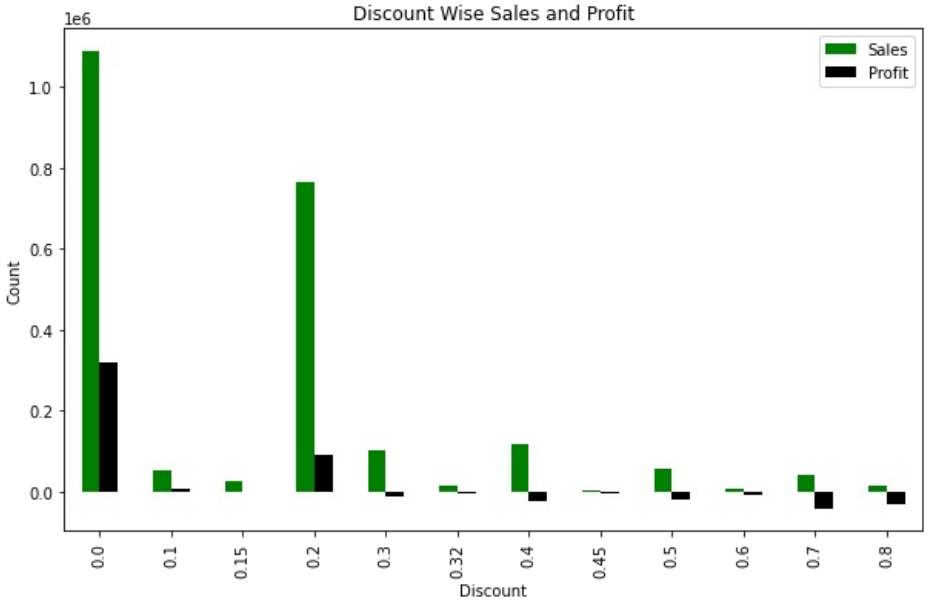
```
C:\Users\ADMIN\AppData\Local\Temp/ipykernel_42384/1932217870.py:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  superstore.groupby(['Sub-Category'])['Sales','Profit'].sum().plot.bar(color=['Orange','Black'],figsize=(12,6))
```



Discounts wise Sales and Profit

In [25]:
```
superstore.groupby(['Discount'])['Sales','Profit'].sum().plot.bar(color=['Green','black'],figsize=(10,6))
plt.title('Discount Wise Sales and Profit')
plt.ylabel('Count')
plt.xlabel('Discount')
plt.show()
```

```
C:\Users\ADMIN\AppData\Local\Temp/ipykernel_42384/2888861062.py:1: FutureWarning: Indexing with multiple keys (im
plicitly converted to a tuple of keys) will be deprecated, use a list instead.
  superstore.groupby(['Discount'])['Sales','Profit'].sum().plot.bar(color=['Green','black'],figsize=(10,6))
```



Conclusion

I was successfully able to carry-out Exploratory data analysis task and was able to evaluate the model's performance on various parameters.

ThankYou

In [ ]:
```
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js