TASK 3 ::EXPLORATORY DATA ANALYSIS

dataset:: https://bit.ly/3i4rbWl

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

In [2]:
```
df=pd.read_csv(r"C:\Users\akank\Downloads\SampleSuperstore.csv")
```

In [4]:
```
df.sample(5)
```

Out[4]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
9288	First Class	Consumer	United States	Philadelphia	Pennsylvania	19134	East	Furniture	Chairs	1079.316	6	0.3	-15.4188
3250	Standard Class	Corporate	United States	Grand Rapids	Michigan	49505	Central	Office Supplies	Fasteners	24.850	7	0.0	11.6795
2517	Standard Class	Corporate	United States	Westfield	New Jersey	7090	East	Furniture	Furnishings	129.930	3	0.0	12.9930
9025	Standard Class	Corporate	United States	New York City	New York	10035	East	Technology	Accessories	139.960	4	0.0	9.7972
1596	First Class	Corporate	United States	New York City	New York	10011	East	Furniture	Furnishings	547.300	13	0.0	175.1360

In [5]:
```
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 [8]:
```
df.describe()
```

Out[8]:

	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]:
```
for i in df.columns:
    print(i,len(df[i].unique()))
```

```
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
```

In [12]:
```
df.isnull().sum()
```

Out[12]:
```
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 [16]:
```
df.duplicated().sum()
```

Out[16]:
```
17
```

In [19]:
```
df.shape
```

Out[19]:
```
(9994, 13)
```

In [20]:
```
df.drop_duplicates()
```

Out[20]:

	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

9977 rows × 13 columns

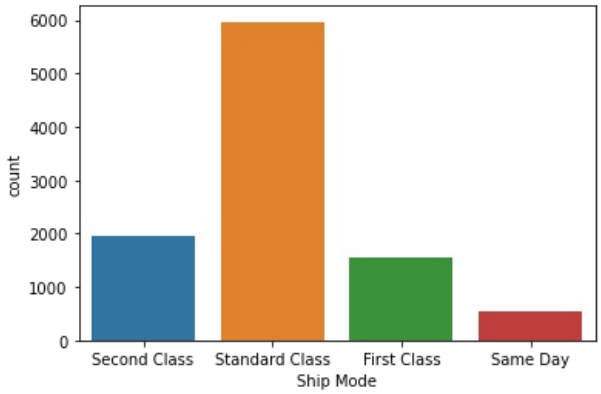
In [21]:
```
df.shape
```

Out[21]: (9994, 13)

countplot to know most preferred shipment

In [79]:
```
sns.countplot(x=df["Ship Mode"])
```

Out[79]: <AxesSubplot:xlabel='Ship Mode', ylabel='count'>

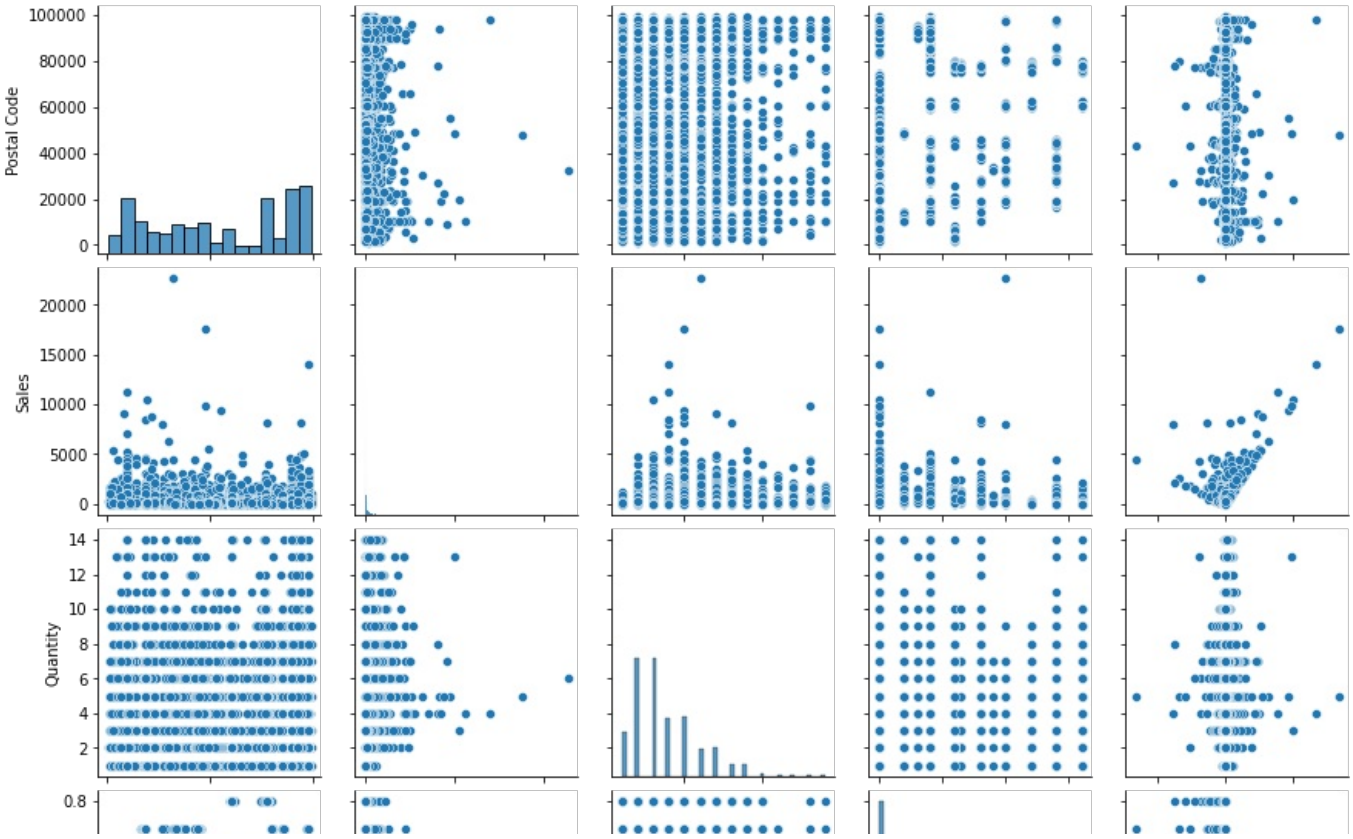


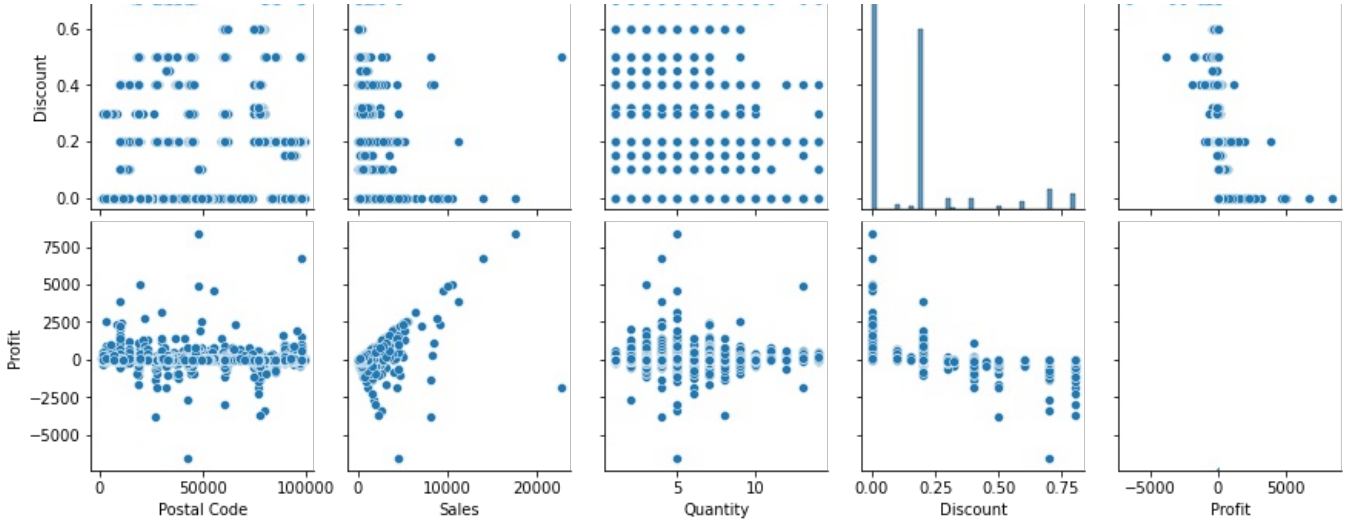
Standard class Is the most preferred shipment mode

PAIRWISE RELATIONSHIP BETWEEN COLUMNS

In [22]:
```
sns.pairplot(df)
```

Out[22]: <seaborn.axisgrid.PairGrid at 0x13977ceec40>

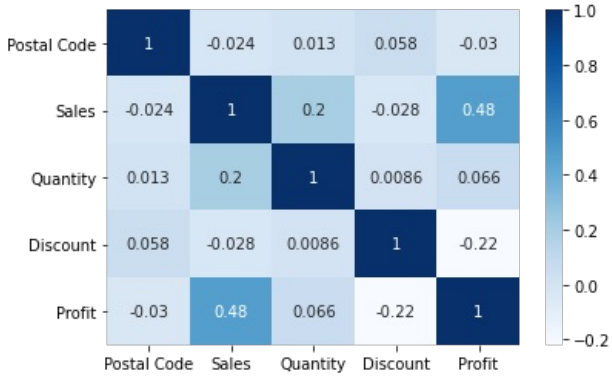




HEATMAP

In [25]:
```
samp_heat=df.corr()
sns.heatmap(samp_heat,annot=True,cmap="Blues")
```

Out[25]: <AxesSubplot:>



most correlated:: sales and profit (0.48)

least correlated:: discount and quantity (0.0086)

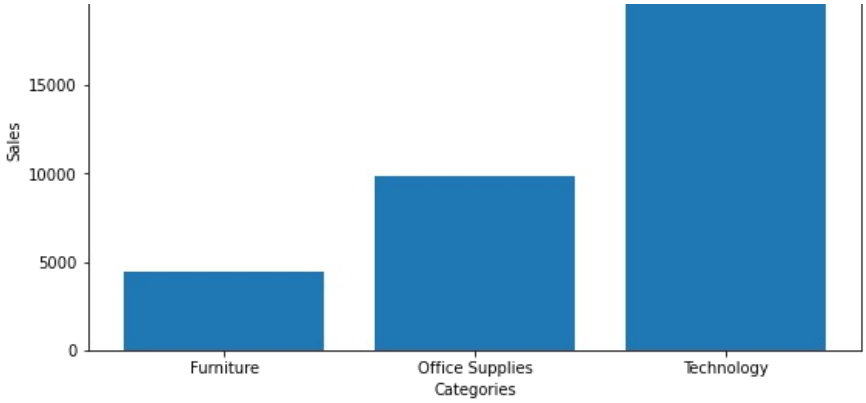
TOTAL SALES AND PROFIT

In [30]:
```
print("Total sales in the dataset is {}".format(df["Sales"].sum()))
print("Total profit in the dataset is {}".format(df["Profit"].sum()))
```

```
Total sales in the dataset is 2297200.8603000003
Total profit in the dataset is 286397.0217
```

Graph showing sales across different categories

In [34]:
```
plt.figure(figsize=(9,5))
plt.bar(df['Category'],df['Sales'])
plt.xlabel("Categories")
plt.ylabel("Sales")
plt.show()
```

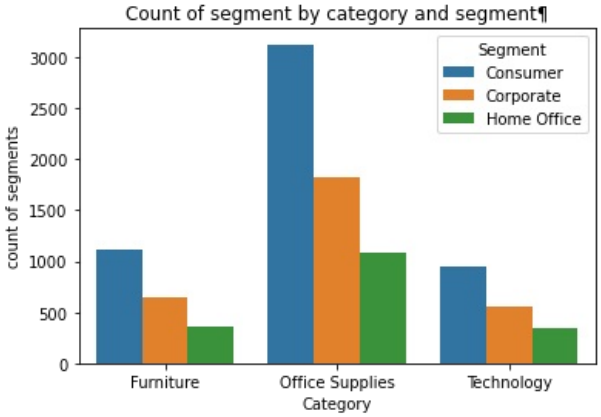


from the barplot we can understand that maximum sales is of technology and minimum sales is of furniture

Count of segment by category and segment

In [36]:
```
sns.countplot(x="Category",hue="Segment",data=df)
plt.ylabel("count of segments")
plt.title("Count of segment by category and segment¶")
```

Out[36]:
```
Text(0.5, 1.0, 'Count of segment by category and segment¶')
```

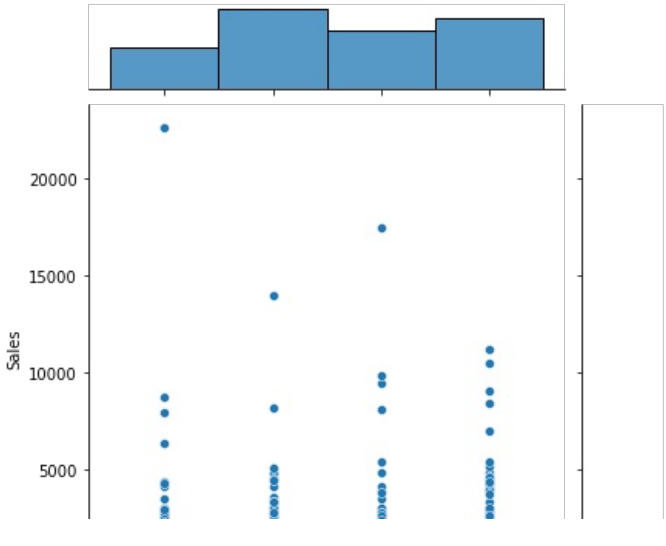


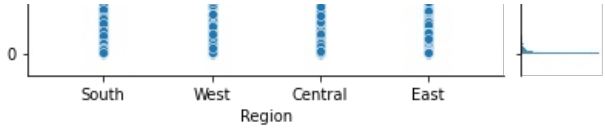
from the above plot we can understand that around 1100 consumers have ordered furnitute,650 coperates have ordered furniture and 400 home offices.

JOINTPLOT DEPICTING SALES ACROSS REGIONS

In [38]:
```
sns.jointplot(x="Region",y="Sales",data=df)
```

Out[38]:
```
<seaborn.axisgrid.JointGrid at 0x13905d690d0>
```



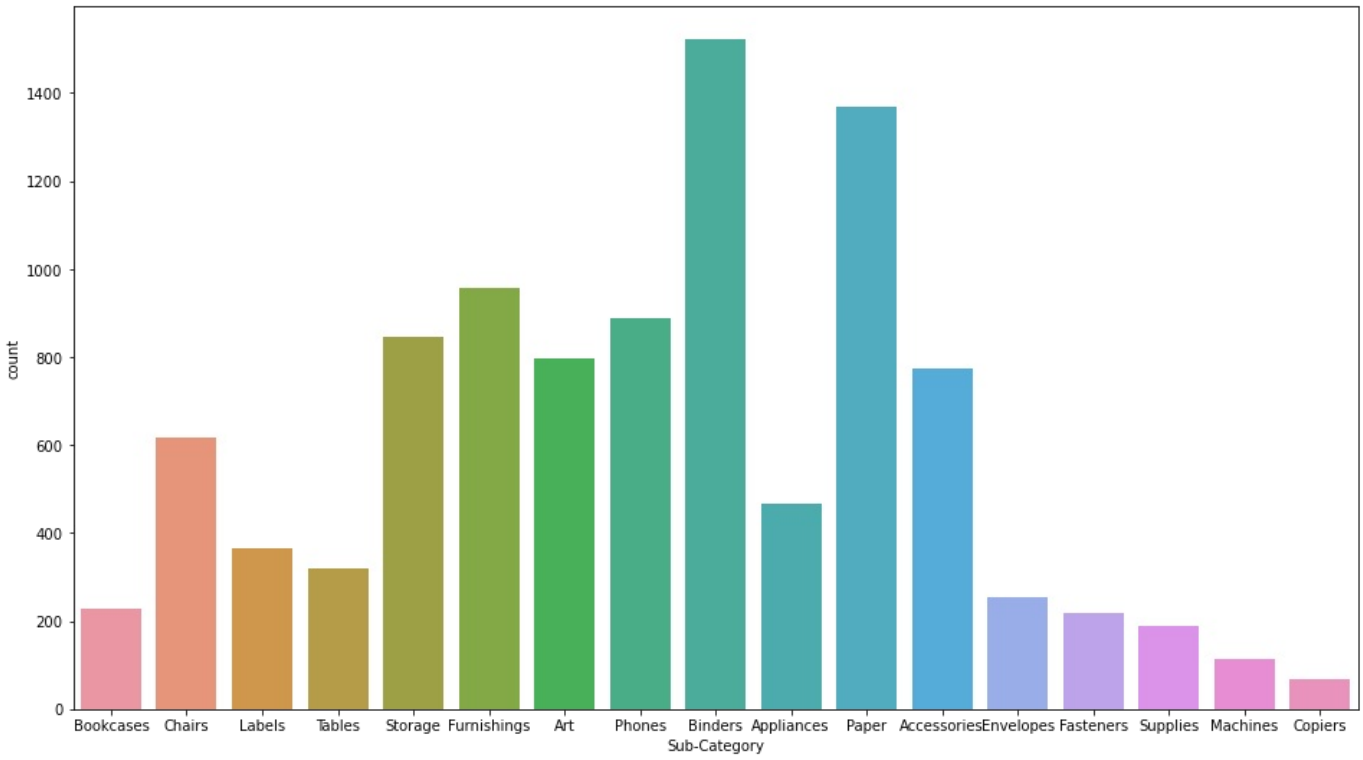


Countplot depicting nunber of subplots

In [40]:
```
plt.figure(figsize=(16,9))
sns.countplot(df['Sub-Category'])
```

```
C:\Users\akank\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable
as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other argu
ments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

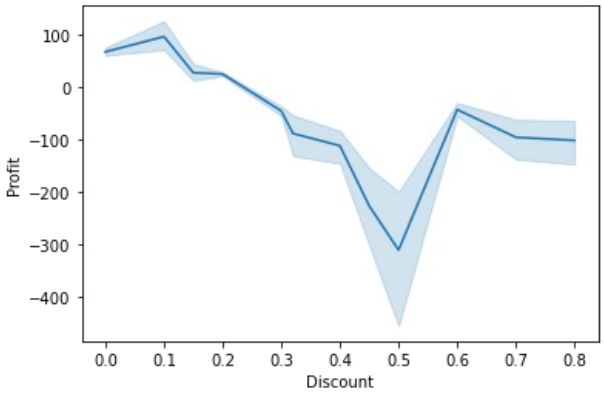
Out[40]: <AxesSubplot:xlabel='Sub-Category', ylabel='count'>



Line plot depicting relation between discount and profit

In [46]:
```
sns.lineplot(x="Discount",y="Profit",data=df)
```

Out[46]: <AxesSubplot:xlabel='Discount', ylabel='Profit'>



In [52]:
```
sample_subcategory=df.groupby(['Sub-Category'])[['Sales','Profit','Discount']].sum()
sample_subcategory
```

Out[52]:

Sub-Category	Sales	Profit	Discount
Accessories	167380.3180	41936.6357	60.80
Appliances	107532.1610	18138.0054	77.60
Art	27118.7920	6527.7870	59.60
Binders	203412.7330	30221.7633	567.00
Bookcases	114879.9963	-3472.5560	48.14
Chairs	328449.1030	26590.1663	105.00
Copiers	149528.0300	55617.8249	11.00
Envelopes	16476.4020	6964.1767	20.40
Fasteners	3024.2800	949.5182	17.80
Furnishings	91705.1640	13059.1436	132.40
Labels	12486.3120	5546.2540	25.00
Machines	189238.6310	3384.7569	35.20
Paper	78479.2060	34053.5693	102.60
Phones	330007.0540	44515.7306	137.40
Storage	223843.6080	21278.8264	63.20
Supplies	46673.5380	-1189.0995	14.60
Tables	206965.5320	-17725.4811	83.35

phones has the made the highest sales .

copiers made the highest profit.

bookcases made least profit but a good sale, so its price can be increased for a increse profit

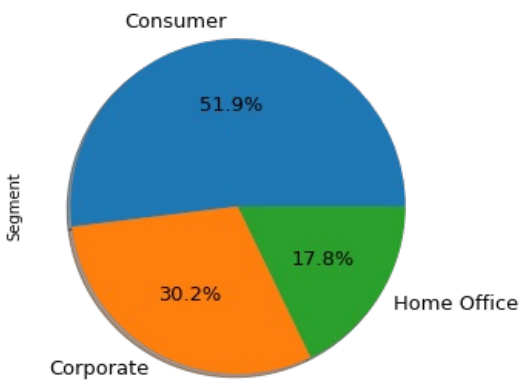
Distribution of segment using piechart.

In [56]:
```
seg=df["Segment"].value_counts()
seg
```

Out[56]:
```
Consumer       5191
Corporate      3020
Home Office    1783
Name: Segment, dtype: int64
```

In [61]:
```
seg.plot(kind='pie',figsize=(5,5),fontsize=13,shadow=True,autopct='%1.1f%%')
```

Out[61]: <AxesSubplot:ylabel='Segment'>

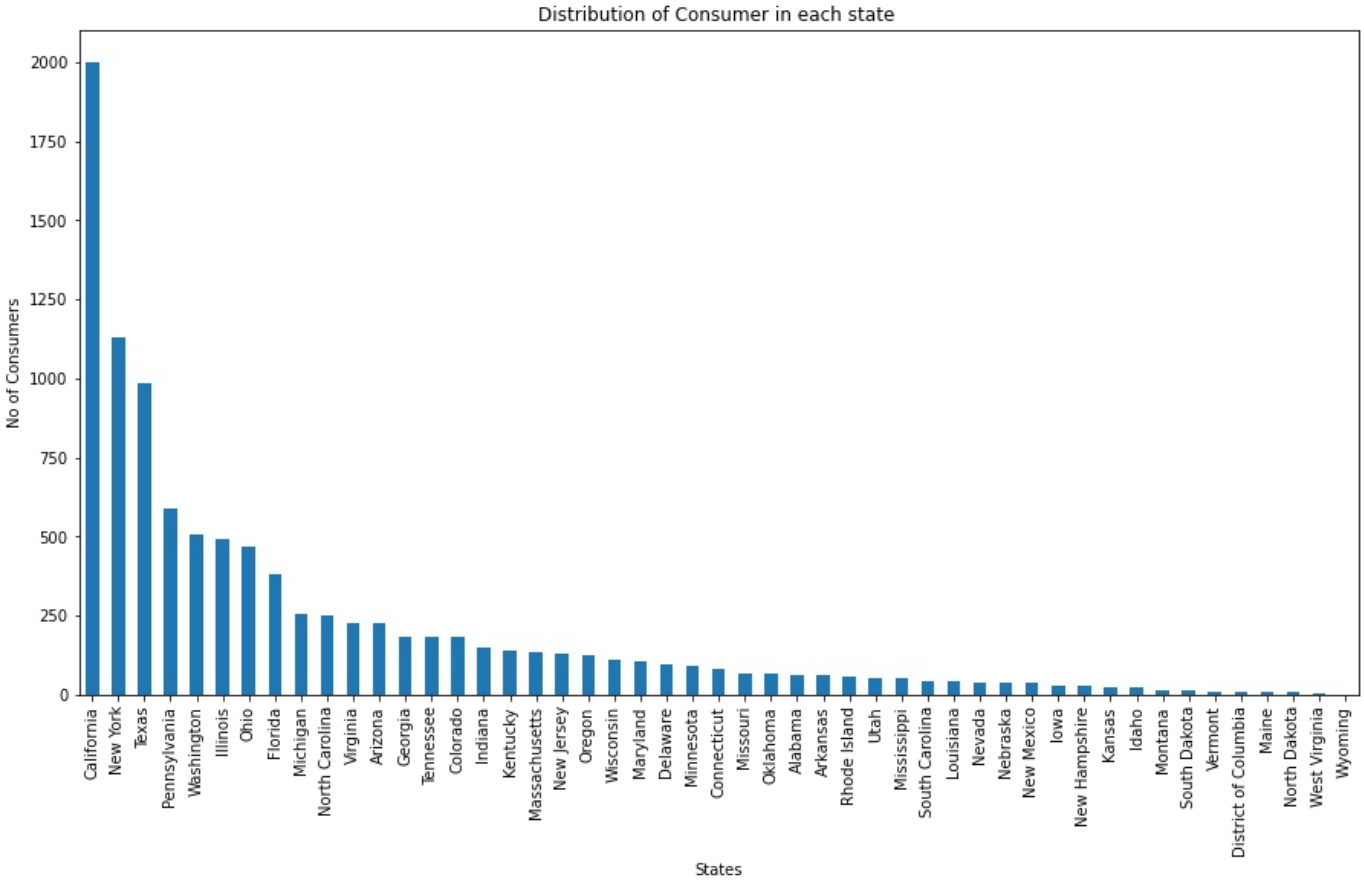


consumers segment has the highest share followed by Corporate and home offices

Statewise distribution of consumers

In [71]:
```
count=df['State'].value_counts()
count.plot(kind='bar',figsize=(15,8))
plt.xlabel("States")
plt.ylabel("No of Consumers")
plt.title('Distribution of Consumer in each state ')
```

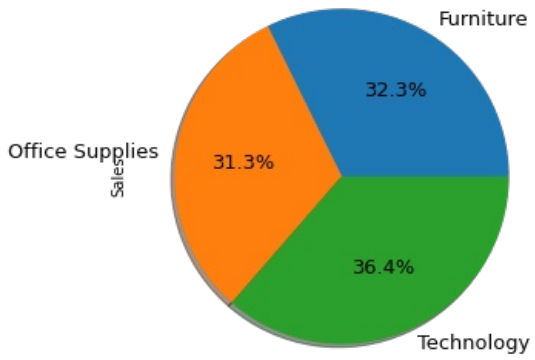
Out[71]: Text(0.5, 1.0, 'Distribution of Consumer in each state ')



Summing up the sales across various categories

In [73]:
```
cs_sum=df.groupby('Category').Sales.sum()
cs_sum.plot(kind='pie',figsize=(5,5),fontsize=13,shadow=True,autopct='%1.1f%%')
```

Out[73]: <AxesSubplot:ylabel='Sales'>



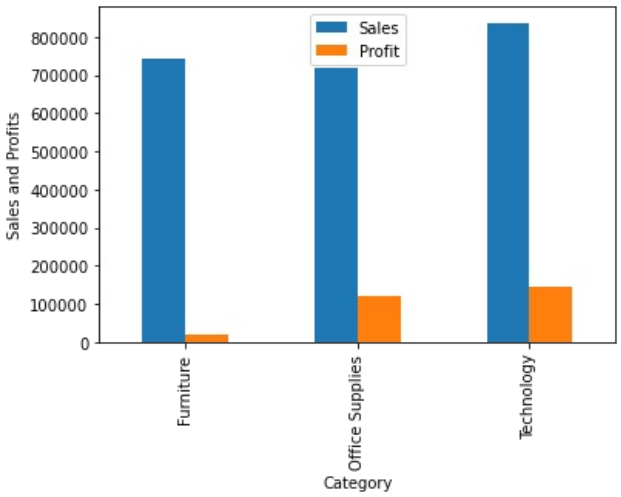
Technology has the higest sales followed by office supply and furniture

Profit and sales across various categories.

In [76]:
```
df.groupby('Category')['Sales','Profit'].agg(sum).plot(kind='bar')
plt.ylabel("Sales and Profits")
```

```
<ipython-input-76-e65d92d7863b>:1: FutureWarning: Indexing with multiple keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  df.groupby('Category')['Sales','Profit'].agg(sum).plot(kind='bar')
```

Out[76]:
```
Text(0, 0.5, 'Sales and Profits')
```

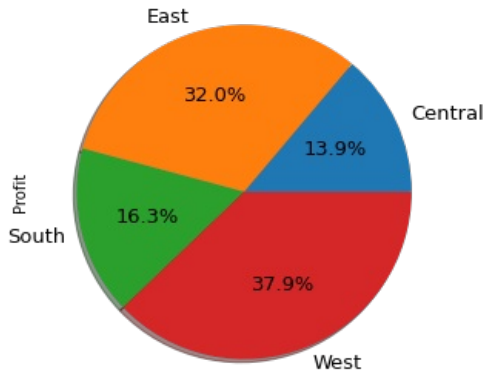


furniture is making the least profit

Summing up profits in various regions

In [78]:
```
rp_sum=df.groupby("Region").Profit.sum()
rp_sum.plot(kind='pie',figsize=(5,5),fontsize=13,shadow=True,autopct='%1.1f%%')
```

Out[78]:
```
<AxesSubplot:ylabel='Profit'>
```



West made the highest profit

CONCLUSION

1.Standard class Is the most preferred shipment mode.

2.Highest Sales and Profits:::::

a)Region-:West

b)Segment-:Consumer

c)Category-:Technology

3.Lowest Sales and Profits::::

a)Region-:Central

b)Segment-:Home Offices

c)Category-:Furniture

4.California and New York have the highest No of consumers.

5.Among the subcategories, Phones had the highest sales followed by Chairs.Copiers made the highest profit.Bookcases made the least profit but a good sale,Prices of bookcases can be incresed for a increase in Profit