Name: Yassen Ayman // Track: Data Science and Bussiness Analytics

task 3 for sparks

In [21]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
data=pd.read_csv('SampleSuperstore.csv')
```

In [3]:
```
data.head()
```

Out[3]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sa
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.96
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.94
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.62
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.57
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.36

In [4]:
```
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 [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

In [7]:
```
data.duplicated().sum()
```

Out[7]:
```
17
```

In [8]:
```
# drop duplicated
data.drop_duplicates()
```

Out[8]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage
...	...	...	...	...	...	...	...	...	...
9989	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings
9990	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper
9993	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances

9977 rows × 13 columns

In [10]:
```
data.isnull().sum()
```

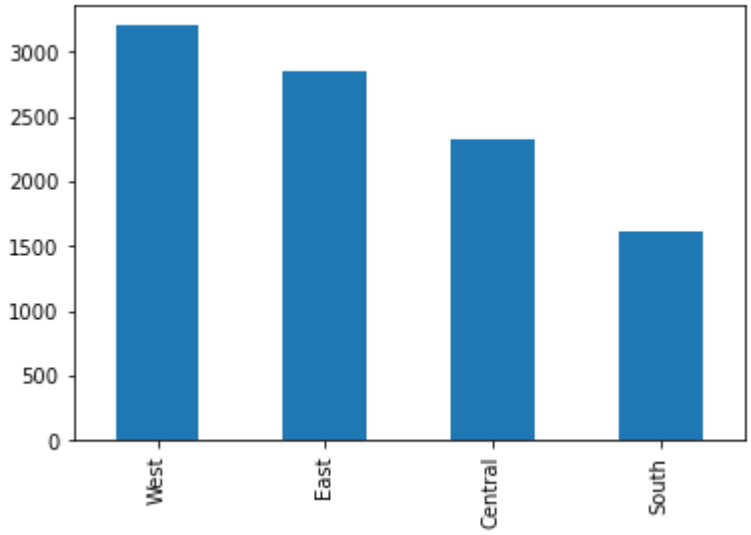
Out[10]:
```
Ship Mode       0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Category        0
Sub-Category    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

In [28]:
```
sns.heatmap(data.corr(),annot=True)
```

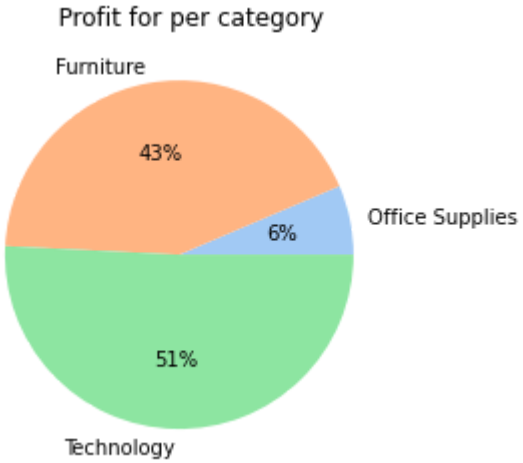
Out[28]:
```
<AxesSubplot:>
```

In [15]:
```
data['Region'].value_counts().plot.bar()
```

Out[15]: <AxesSubplot:>



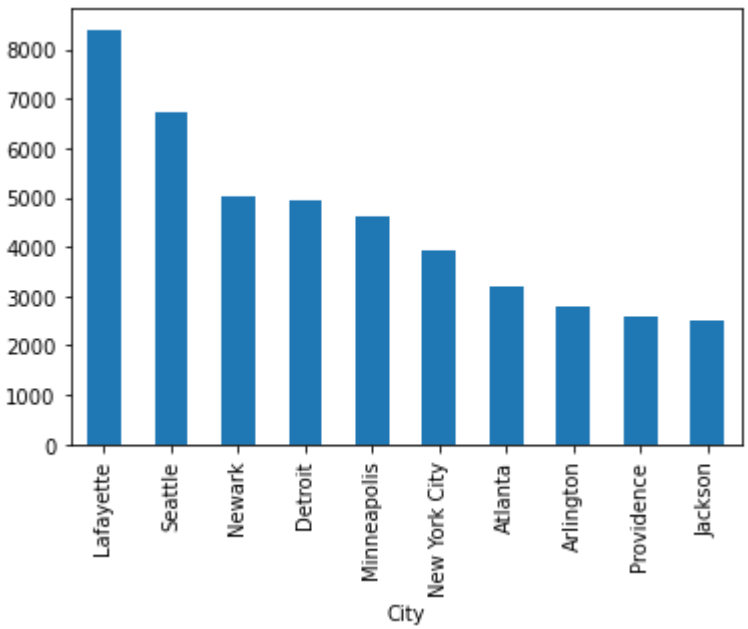
In [27]:
```
exp=data.groupby('Category')['Profit'].sum()
colors = sns.color_palette('pastel')[0:5]
labels=['Office Supplies','Furniture','Technology']
#create pie chart
plt.pie(exp ,labels=labels,colors = colors, autopct='%.0f%%')
plt.title('   Profit for per category')
plt.show()
```



Calculate Max profit per each city

In [35]:
```
data.groupby('City')['Profit'].max().nlargest(10).plot.bar()
```

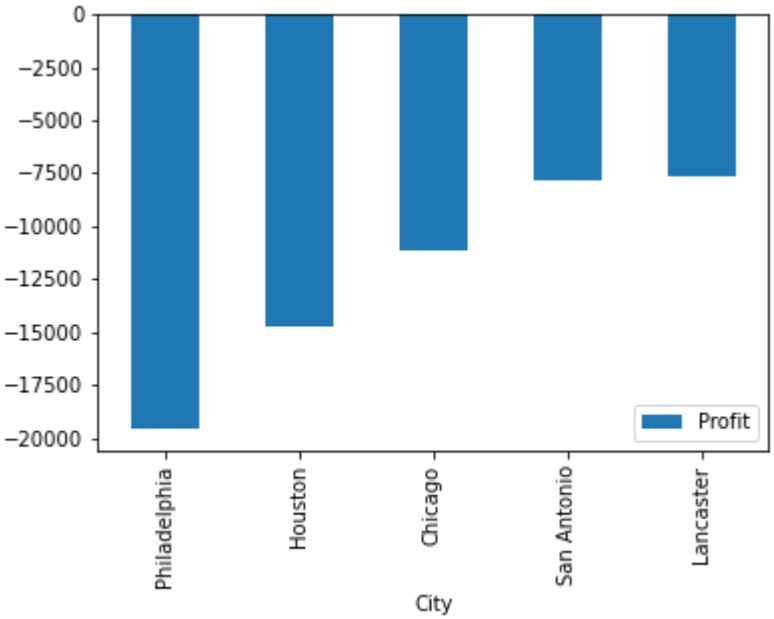
Out[35]: <AxesSubplot:xlabel='City'>



Calculate lost money per each city

In [71]:
```
JustSubCatProf = data[["Sub-Category", "Profit","City"]]
NegProfFilter = data["Profit"] < 0.0
JustNegSubCatProf = JustSubCatProf[NegProfFilter].groupby(by = "City").sum().sort
print(JustNegSubCatProf)
```

```
AxesSubplot(0.125,0.125;0.775x0.755)
```



In [74]:
```
JustSubCatProf = data[["Sub-Category", "Profit","City","Discount"]]
NegProfFilter = data["Profit"] < 0.0
JustNegSubCatProf = JustSubCatProf[NegProfFilter].groupby(by = "City").sum().sort
print(JustNegSubCatProf)
```

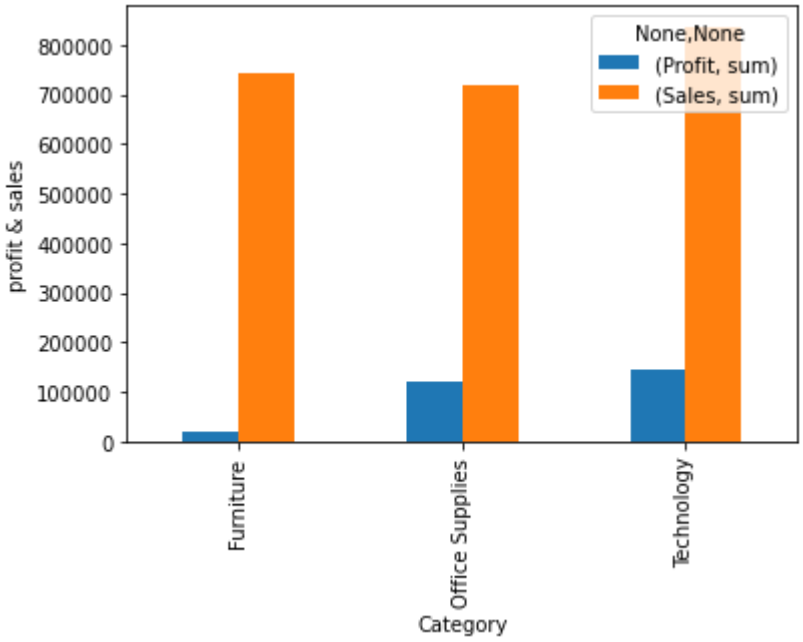
```
                Profit  Discount
City
Philadelphia -19590.7411    115.30
Houston      -14785.3668    104.14
Chicago      -11120.6271     88.20
San Antonio   -7831.0254     17.10
Lancaster     -7632.4946      9.40
...                  ...       ...
Loveland         -1.5948      0.90
Pensacola        -1.4760      0.70
Elyria           -1.3984      0.70
Homestead        -0.6624      0.20
Coppell          -0.2098      0.20

[229 rows x 2 columns]
```

Visualize Profit and Sales

In [76]:
```
plt.figure(figsize=(10,16))
data.groupby('Category')['Profit','Sales'].agg(['sum']).plot.bar()
plt.ylabel('profit & sales')
plt.show()
```

```
<ipython-input-76-986a0b826ae7>:2: FutureWarning: Indexing with multiple keys
(implicitly converted to a tuple of keys) will be deprecated, use a list instea
d.
  data.groupby('Category')['Profit','Sales'].agg(['sum']).plot.bar()
```

```
<Figure size 720x1152 with 0 Axes>
```



In [90]:
```
JustSubCatProf = data[["Category", "Profit","Region"]]
NegProfFilter = data["Profit"] < 0.0
JustNegSubCatProf = JustSubCatProf[NegProfFilter].groupby(["Category","Region"]).
print(JustNegSubCatProf)
```

```
AxesSubplot(0.125,0.125;0.775x0.755)
```