In [1]:
```
import pandas as pd
import os
import matplotlib.pyplot as plt
```

In [2]:
```
files=[file for file in os.listdir('SalesAnalysis/Sales_Data')]
allmonths_data=pd.DataFrame()
for file in files:
    df=pd.read_csv('SalesAnalysis/Sales_Data/'+ file)
    allmonths_data=pd.concat([allmonths_data,df])
allmonths_data.to_csv('full_data.csv',index=False)
```

In [3]:
```
full_data=pd.read_csv('full_data.csv')
full_data.head()
```

Out[3]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
1	NaN	NaN	NaN	NaN	NaN	NaN
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001

In [4]:
```
full_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 186850 entries, 0 to 186849
Data columns (total 6 columns):
```

```
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Order ID          186305 non-null  object
 1   Product           186305 non-null  object
 2   Quantity Ordered  186305 non-null  object
 3   Price Each        186305 non-null  object
 4   Order Date        186305 non-null  object
 5   Purchase Address  186305 non-null  object
dtypes: object(6)
memory usage: 8.6+ MB
```

In [5]:
```
full_data.shape
```

Out[5]: (186850, 6)

In [6]:
```
full_data.tail()
```

Out[6]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
186845	259353	AAA Batteries (4-pack)	3	2.99	09/17/19 20:56	840 Highland St, Los Angeles, CA 90001
186846	259354	iPhone	1	700	09/01/19 16:00	216 Dogwood St, San Francisco, CA 94016
186847	259355	iPhone	1	700	09/23/19 07:39	220 12th St, San Francisco, CA 94016
186848	259356	34in Ultrawide Monitor	1	379.99	09/19/19 17:30	511 Forest St, San Francisco, CA 94016
186849	259357	USB-C Charging Cable	1	11.95	09/30/19 00:18	250 Meadow St, San Francisco, CA 94016

In [7]:
```
full_data.describe()
```

Out[7]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
count	186305	186305	186305	186305	186305	186305
unique	178438	20	10	24	142396	140788
top	Order ID	USB-C Charging Cable	1	11.95	Order Date	Purchase Address
freq	355	21903	168552	21903	355	355

In [8]:
```
nan_df=full_data[full_data.isna().any(axis=1)]
nan_df.head()
```

Out[8]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
1	NaN	NaN	NaN	NaN	NaN	NaN
356	NaN	NaN	NaN	NaN	NaN	NaN
735	NaN	NaN	NaN	NaN	NaN	NaN
1433	NaN	NaN	NaN	NaN	NaN	NaN
1553	NaN	NaN	NaN	NaN	NaN	NaN

In [9]:
```
full_data=full_data.dropna(how='all')
full_data.head()
```

Out[9]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001

In [10]:
```
full_data['Month']=full_data['Order Date'].str[0:2]
full_data.head()
```

Out[10]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	04
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	04
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	04

In [11]:
```
temp_df=full_data[full_data['Order Date'].str[0:2] =='Or']
temp_df.head()
```

Out[11]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
519	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
1149	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
1155	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
2878	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
2893	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or

In [12]:
```
full_data=full_data[full_data['Order Date'].str[0:2] != 'Or']
```

In [13]:
```
full_data['Quantity Ordered']= pd.to_numeric(full_data['Quantity Ordered'])
full_data['Price Each']=pd.to_numeric(full_data['Price Each'])
full_data.head()
```

Out[13]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	04
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	04
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	04

In [ ]:
```
```

In [14]:
```
full_data['Month']=full_data['Month'].astype('int32')
full_data.head()
```

Out[14]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4

In [15]:
```
full_data['Sales']=full_data['Quantity Ordered']*full_data['Price Each']
]
full_data.head()
```

Out[15]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99

In [16]:
```
def get_city(address):
    return address.split(',')[1]
def get_state(address):
    return address.split(',')[2].split(' ')[1]
full_data['City']=full_data['Purchase Address'].apply(lambda x:get_city
```

```
(x) +' '+get_state(x))
full_data.head()
```

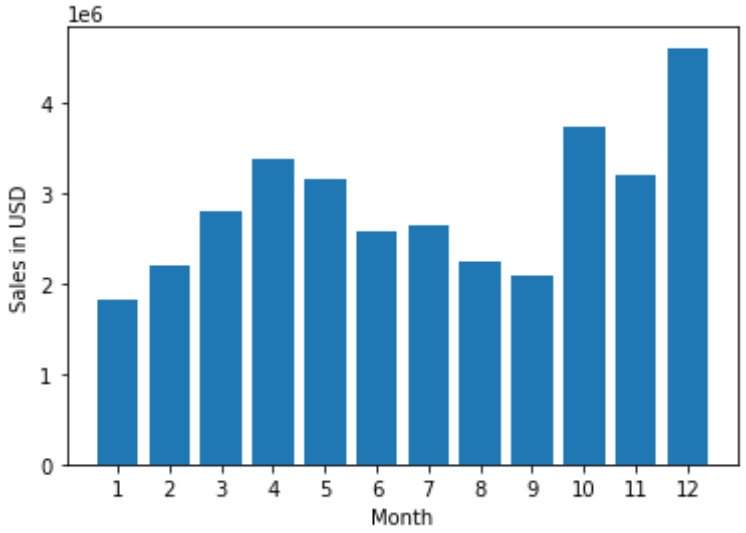
Out[16]:

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	4	23.90	Dallas TX
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	4	99.99	Boston MA
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles CA
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles CA
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles CA

What was the best month for sales? How much was earned that month?

In [17]:
```
full_data.groupby('Month').sum()
months=range(1,13)
results=full_data.groupby('Month').sum()
plt.bar(months,results['Sales'])
plt.xticks(months)
plt.ylabel('Sales in USD')
plt.xlabel('Month')
plt.show
```

Out[17]:
```
<function matplotlib.pyplot.show(*args, **kw)>
```



By Bar-Chart it is clear that December has the most Sale.The revenue generated this month is around 4.5million dollars.The Sales were less in the month of January but It increased the following months till April,then again there was constant drop in Sales.

What city has the highest number of sales?

In [20]:
```
results=full_data.groupby('City').sum()
results
```

Out[20]:

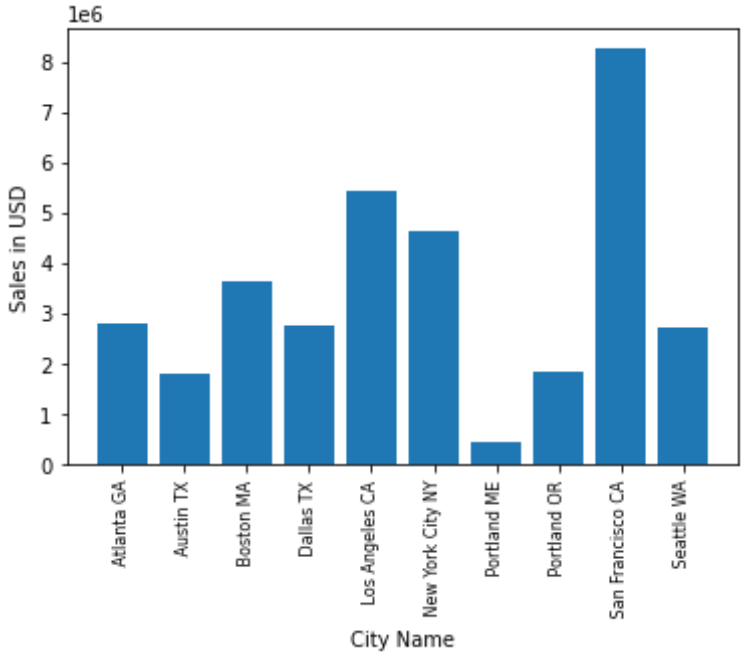
City	Quantity Ordered	Price Each	Month	Sales
Atlanta GA	16602	2.779908e+06	104794	2.795499e+06
Austin TX	11153	1.809874e+06	69829	1.819582e+06
Boston MA	22528	3.637410e+06	141112	3.661642e+06
Dallas TX	16730	2.752628e+06	104620	2.767975e+06

City	Quantity Ordered	Price Each	Month	Sales
Los Angeles CA	33289	5.421435e+06	208325	5.452571e+06
New York City NY	27932	4.635371e+06	175741	4.664317e+06
Portland ME	2750	4.471893e+05	17144	4.497583e+05
Portland OR	11303	1.860558e+06	70621	1.870732e+06
San Francisco CA	50239	8.211462e+06	315520	8.262204e+06
Seattle WA	16553	2.733296e+06	104941	2.747755e+06

In [25]:
```
cities=[city for city, df in full_data.groupby('City')]

plt.bar(cities,results['Sales'])
plt.xticks(cities,rotation='vertical',size=8)
plt.ylabel('Sales in USD')
plt.xlabel('City Name')
```

Out[25]:
```
Text(0.5, 0, 'City Name')
```



It is clear that Sab Francisco bagged the top position in Sales which corresponds to 8.26Million USD whereas Portland is at the bottom position.The two cities Dallas and Seattle are on the same position.

In [ ]: