In [5]:

```
import pandas as pd
data=pd.read_excel("C:/Users/ayush/Desktop/TE Practical Exam/DSBDA/Q2/AirQualityUCI.xls
```

In [6]:

```
data
```

Out[6]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	pr
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.87
1	2004-03-10	0.791667	2.0	1292.25	112	9.397165	954.75	103.0	4619.06
2	2004-03-10	0.833333	2.2	1402.00	88	8.997817	939.25	131.0	5862.09
3	2004-03-10	0.875000	2.2	1375.50	80	9.228796	948.25	172.0	7682.24
4	2004-03-10	0.916667	1.6	1272.25	51	6.518224	835.50	131.0	5862.09
...	...	...	...	...	...	...	...	...	
9352	2005-04-04	0.416667	3.1	1314.25	-200	13.529605	1101.25	471.7	
9353	2005-04-04	0.458333	2.4	1162.50	-200	11.355157	1027.00	353.3	
9354	2005-04-04	0.500000	2.4	1142.00	-200	12.374538	1062.50	293.0	
9355	2005-04-04	0.541667	2.1	1002.50	-200	9.547187	960.50	234.5	
9356	2005-04-04	0.583333	2.2	1070.75	-200	11.932060	1047.25	265.2	

9357 rows × 17 columns

In [7]:

```
data.dropna()
```

Out[7]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	predic
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.87878
1	2004-03-10	0.791667	2.0	1292.25	112	9.397165	954.75	103.0	4619.06060
2	2004-03-10	0.833333	2.2	1402.00	88	8.997817	939.25	131.0	5862.09090
3	2004-03-10	0.875000	2.2	1375.50	80	9.228796	948.25	172.0	7682.24242

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	predic
4	2004-03-10	0.916667	1.6	1272.25	51	6.518224	835.50	131.0	5862.09090

In [8]:

```
data.fillna(1)
```

Out[8]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	pr
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.87
1	2004-03-10	0.791667	2.0	1292.25	112	9.397165	954.75	103.0	4619.06
2	2004-03-10	0.833333	2.2	1402.00	88	8.997817	939.25	131.0	5862.09
3	2004-03-10	0.875000	2.2	1375.50	80	9.228796	948.25	172.0	7682.24
4	2004-03-10	0.916667	1.6	1272.25	51	6.518224	835.50	131.0	5862.09
...	...	...	...	...	...	...	...	...	
9352	2005-04-04	0.416667	3.1	1314.25	-200	13.529605	1101.25	471.7	1.00
9353	2005-04-04	0.458333	2.4	1162.50	-200	11.355157	1027.00	353.3	1.00
9354	2005-04-04	0.500000	2.4	1142.00	-200	12.374538	1062.50	293.0	1.00
9355	2005-04-04	0.541667	2.1	1002.50	-200	9.547187	960.50	234.5	1.00
9356	2005-04-04	0.583333	2.2	1070.75	-200	11.932060	1047.25	265.2	1.00

9357 rows × 17 columns

In [9]:

```
mean=data['CO(GT)'].mean()
```

In [10]:

```
mean
```

Out[10]:

```
-34.20752377898902
```

In [11]:

```
median=data['CO(GT)'].median()
```

In [12]:

```
median
```

Out[12]:

```
1.5
```

In [13]:

```
mode=data['CO(GT)'].mode()
```

In [14]:

```
mode
```

Out[14]:

```
0   -200.0
dtype: float64
```

In [15]:

```
mean_data=data.groupby('CO(GT)')['NO2(GT)'].mean()
```

In [16]:

```
mean_data
```

Out[16]:

```
CO(GT)
-200.0   -122.407427
 0.1       66.345455
 0.2       33.653333
 0.3        5.003061
 0.4        2.786875
              ...
 9.9      269.000000
 10.1     255.000000
 10.2     209.500000
 11.5     190.000000
 11.9     220.000000
Name: NO2(GT), Length: 97, dtype: float64
```

In [17]:

```
mean_data=data.groupby('CO(GT)')['NO2(GT)'].mean().rename("user_mean").reset_index()
```

In [18]:

```
mean_data
```

Out[18]:

	CO(GT)	user_mean
0	-200.0	-122.407427
1	0.1	66.345455
2	0.2	33.653333
3	0.3	5.003061
4	0.4	2.786875
...	...	...
92	9.9	269.000000
93	10.1	255.000000
94	10.2	209.500000
95	11.5	190.000000
96	11.9	220.000000

97 rows × 2 columns

In [19]:

```
final_data=data.merge(mean_data)
```

In [20]:

```
final_data
```

Out[20]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.8
1	2004-03-13	0.958333	2.6	1418.00	116	10.873367	1009.75	172.0	
2	2004-03-16	0.916667	2.6	1379.25	183	13.529605	1101.25	184.0	
3	2004-03-17	0.583333	2.6	1389.25	152	13.735290	1108.00	161.0	
4	2004-03-17	0.958333	2.6	1488.00	216	15.710274	1170.75	178.0	
...	...	...	...	...	...	...	...	...	
9352	2004-12-16	0.833333	9.1	-200.00	-200	-200.000000	-200.00	1253.0	
9353	2004-12-23	0.833333	9.1	1701.00	-200	36.263240	1691.75	1220.0	
9354	2004-12-23	0.750000	8.5	1629.50	-200	33.088098	1621.75	1089.0	
9355	2005-02-11	0.666667	7.1	-200.00	-200	-200.000000	-200.00	1218.0	
9356	2005-02-11	0.791667	7.1	-200.00	-200	-200.000000	-200.00	1074.8	

9357 rows × 18 columns

In [21]:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model as lm
```

In [22]:

```
data
```

Out[22]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	pr
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.87
1	2004-03-10	0.791667	2.0	1292.25	112	9.397165	954.75	103.0	4619.06

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	pr
2	2004-03-10	0.833333	2.2	1402.00	88	8.997817	939.25	131.0	5862.09
3	2004-03-10	0.875000	2.2	1375.50	80	9.228796	948.25	172.0	7682.24
4	2004-03-10	0.916667	1.6	1272.25	51	6.518224	835.50	131.0	5862.09
...	...	...	...	...	...	...	...	...	
9352	2005-04-04	0.416667	3.1	1314.25	-200	13.529605	1101.25	471.7	
9353	2005-04-04	0.458333	2.4	1162.50	-200	11.355157	1027.00	353.3	
9354	2005-04-04	0.500000	2.4	1142.00	-200	12.374538	1062.50	293.0	
9355	2005-04-04	0.541667	2.1	1002.50	-200	9.547187	960.50	234.5	
9356	2005-04-04	0.583333	2.2	1070.75	-200	11.932060	1047.25	265.2	

9357 rows × 17 columns

In [23]:

```
data1=data.head(5)
```

In [24]:

```
data1
```

Out[24]:

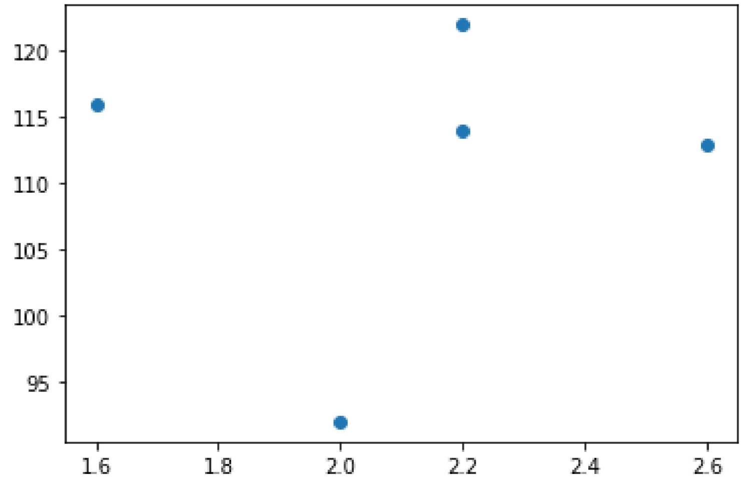
	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	predic
0	2004-03-10	0.750000	2.6	1360.00	150	11.881723	1045.50	166.0	7415.87878
1	2004-03-10	0.791667	2.0	1292.25	112	9.397165	954.75	103.0	4619.06060
2	2004-03-10	0.833333	2.2	1402.00	88	8.997817	939.25	131.0	5862.09090
3	2004-03-10	0.875000	2.2	1375.50	80	9.228796	948.25	172.0	7682.24242
4	2004-03-10	0.916667	1.6	1272.25	51	6.518224	835.50	131.0	5862.09090

In [25]:

```
plt.scatter(data1[['CO(GT)']],data1[['NO2(GT)']])
```

Out[25]:

```
<matplotlib.collections.PathCollection at 0x1db40306eb0>
```



```
In [27]: reg=lm.LinearRegression()
```

```
In [28]: reg.fit(data1[['CO(GT)']],data1[['NO2(GT)']])
```

```
Out[28]: LinearRegression()
```

```
In [29]: reg.coef_
```

```
Out[29]: array([[3.33333333]])
```

```
In [30]: reg.intercept_
```

```
Out[30]: array([104.33333333])
```

```
In [31]: y_predict=reg.predict(data1[['NO2(GT)']])
```

```
In [32]: y_predict
```

```
Out[32]: array([[481.        ],
                [411.        ],
                [484.33333333],
                [511.        ],
                [491.        ]])
```

```
In [ ]:
```