PYTHON

DATA CLEANING

```
In [21]: import pandas as pd

In [22]: data = pd.read_csv("airquality.csv")

In [24]: data.head()
Out[24]:
   Unnamed: 0  Ozone  Solar.R  Wind  Temp  Month  Day
0           1   41.0    190.0   7.4    67      5    1
1           2   36.0    118.0   8.0    72      5    2
2           3   12.0    149.0  12.6    74      5    3
3           4   18.0    313.0  11.5    62      5    4
4           5    NaN      NaN  14.3    56      5    5

In [25]: data.isnull().sum()
Out[25]:
Unnamed: 0     0
Ozone         37
Solar.R        7
Wind           0
Temp           0
Month          0
Day            0
dtype: int64

In [26]: data["Ozone"].fillna(data['Ozone'].median(),inplace = True)

In [27]: data.isnull().sum()
Out[27]:
Unnamed: 0    0
Ozone         0
Solar.R       7
Wind          0
Temp          0
Month         0
Day           0
dtype: int64

In [28]: data["Solar.R"].fillna(data['Solar.R'].median(),inplace = True)

In [29]: data.isnull().sum()
Out[29]:
Unnamed: 0    0
Ozone         0
Solar.R       0
Wind          0
Temp          0
Month         0
Day           0
dtype: int64
```

```
In [30]: data.head()
Out[30]:
   Unnamed: 0  Ozone  Solar.R  Wind  Temp  Month  Day
0           1   41.0    190.0   7.4    67      5    1
1           2   36.0    118.0   8.0    72      5    2
2           3   12.0    149.0  12.6    74      5    3
3           4   18.0    313.0  11.5    62      5    4
4           5   31.5    205.0  14.3    56      5    5
[153 rows x 7 columns]
```

DATA TRANSFORMATION

```
In [31]: data['newsolar'] = [True if x >100 else False for x in
data['Solar.R']]

In [32]: print(data[12:27])
    Unnamed: 0  Ozone  Solar.R  Wind  Temp  Month  Day  newsolar
12          13   11.0    290.0   9.2    66      5   13      True
13          14   14.0    274.0  10.9    68      5   14      True
14          15   18.0     65.0  13.2    58      5   15     False
15          16   14.0    334.0  11.5    64      5   16      True
16          17   34.0    307.0  12.0    66      5   17      True
17          18    6.0     78.0  18.4    57      5   18     False
18          19   30.0    322.0  11.5    68      5   19      True
19          20   11.0     44.0   9.7    62      5   20     False
20          21    1.0      8.0   9.7    59      5   21     False
21          22   11.0    320.0  16.6    73      5   22      True
22          23    4.0     25.0   9.7    61      5   23     False
23          24   32.0     92.0  12.0    61      5   24     False
24          25   31.5     66.0  16.6    57      5   25     False
25          26   31.5    266.0  14.9    58      5   26      True
26          27   31.5    205.0   8.0    57      5   27      True


In [33]: vector = [0,50,100,150,200,250,300,350,400]

In [35]:

In [35]: data.head()
Out[35]:
   Unnamed: 0  Ozone  Solar.R  Wind  Temp  Month  Day  newsolar
0           1   41.0    190.0   7.4    67      5    1      True
1           2   36.0    118.0   8.0    72      5    2      True
2           3   12.0    149.0  12.6    74      5    3      True
3           4   18.0    313.0  11.5    62      5    4      True
4           5   31.5    205.0  14.3    56      5    5      True

In [36]: data['Solar.R']=pd.cut(data['Solar.R'],bins=vector)

In [37]: data
Out[37]:
     Unnamed: 0  Ozone     Solar.R  Wind  Temp  Month  Day  newsolar
0             1   41.0  (150, 200]   7.4    67      5    1      True
1             2   36.0  (100, 150]   8.0    72      5    2      True
```

```
2             3   12.0  (100, 150]  12.6    74      5    3      True
3             4   18.0  (300, 350]  11.5    62      5    4      True
4             5   31.5  (200, 250]  14.3    56      5    5      True
..          ...    ...         ...   ...   ...    ...  ...       ...
148         149   30.0  (150, 200]   6.9    70      9   26      True
149         150   31.5  (100, 150]  13.2    77      9   27      True
150         151   14.0  (150, 200]  14.3    75      9   28      True
151         152   18.0  (100, 150]   8.0    76      9   29      True
152         153   20.0  (200, 250]  11.5    68      9   30      True

In [39]: data.loc[data['Month']==5,'Month']="May"

In [40]: data.loc[data['Month']==6,'Month']="June"

In [41]: data.loc[data['Month']==7,'Month']="July"

In [42]: data.loc[data['Month']==8,'Month']="August"

In [43]: data.loc[data['Month']==9,'Month']="September"

In [44]: data
Out[44]:
     Unnamed: 0  Ozone     Solar.R  Wind  Temp      Month  Day  newsolar
0             1   41.0  (150, 200]   7.4    67        May    1      True
1             2   36.0  (100, 150]   8.0    72        May    2      True
2             3   12.0  (100, 150]  12.6    74        May    3      True
3             4   18.0  (300, 350]  11.5    62        May    4      True
4             5   31.5  (200, 250]  14.3    56        May    5      True
5             6   28.0  (200, 250]  14.9    66        May    6      True
6             7   23.0  (250, 300]   8.6    65        May    7      True
7             8   19.0   (50, 100]  13.8    59        May    8     False
8             9    8.0     (0, 50]  20.1    61        May    9     False
9            10   31.5  (150, 200]   8.6    69        May   10      True
10           11    7.0  (200, 250]   6.9    74        May   11      True
11           12   16.0  (250, 300]   9.7    69        May   12      True
12           13   11.0  (250, 300]   9.2    66        May   13      True
13           14   14.0  (250, 300]  10.9    68        May   14      True
14           15   18.0   (50, 100]  13.2    58        May   15     False
15           16   14.0  (300, 350]  11.5    64        May   16      True
16           17   34.0  (300, 350]  12.0    66        May   17      True
17           18    6.0   (50, 100]  18.4    57        May   18     False
18           19   30.0  (300, 350]  11.5    68        May   19      True
19           20   11.0     (0, 50]   9.7    62        May   20     False
20           21    1.0     (0, 50]   9.7    59        May   21     False
21           22   11.0  (300, 350]  16.6    73        May   22      True
22           23    4.0     (0, 50]   9.7    61        May   23     False
23           24   32.0   (50, 100]  12.0    61        May   24     False
24           25   31.5   (50, 100]  16.6    57        May   25     False
25           26   31.5  (250, 300]  14.9    58        May   26      True
26           27   31.5  (200, 250]   8.0    57        May   27      True
27           28   23.0     (0, 50]  12.0    67        May   28     False
28           29   45.0  (250, 300]  14.9    81        May   29      True
29           30  115.0  (200, 250]   5.7    79        May   30      True
..          ...    ...         ...   ...   ...        ...  ...       ...
```

123	124	96.0	(150, 200]	6.9	91	September	1	True
124	125	78.0	(150, 200]	5.1	92	September	2	True
125	126	73.0	(150, 200]	2.8	93	September	3	True
126	127	91.0	(150, 200]	4.6	93	September	4	True
127	128	47.0	(50, 100]	7.4	87	September	5	False
128	129	32.0	(50, 100]	15.5	84	September	6	False
129	130	20.0	(250, 300]	10.9	80	September	7	True
130	131	23.0	(200, 250]	10.3	78	September	8	True
131	132	21.0	(200, 250]	10.9	75	September	9	True
132	133	24.0	(250, 300]	9.7	73	September	10	True
133	134	44.0	(200, 250]	14.9	81	September	11	True
134	135	21.0	(250, 300]	15.5	76	September	12	True
135	136	28.0	(200, 250]	6.3	77	September	13	True
136	137	9.0	(0, 50]	10.9	71	September	14	False
137	138	13.0	(100, 150]	11.5	71	September	15	True
138	139	46.0	(200, 250]	6.9	78	September	16	True
139	140	18.0	(200, 250]	13.8	67	September	17	True
140	141	13.0	(0, 50]	10.3	76	September	18	False
141	142	24.0	(200, 250]	10.3	68	September	19	True
142	143	16.0	(200, 250]	8.0	82	September	20	True
143	144	13.0	(200, 250]	12.6	64	September	21	True
144	145	23.0	(0, 50]	9.2	71	September	22	False
145	146	36.0	(100, 150]	10.3	81	September	23	True
146	147	7.0	(0, 50]	10.3	69	September	24	False
147	148	14.0	(0, 50]	16.6	63	September	25	False
148	149	30.0	(150, 200]	6.9	70	September	26	True
149	150	31.5	(100, 150]	13.2	77	September	27	True
150	151	14.0	(150, 200]	14.3	75	September	28	True
151	152	18.0	(100, 150]	8.0	76	September	29	True
152	153	20.0	(200, 250]	11.5	68	September	30	True

[153 rows x 8 columns]

DATA INTREGATION

```
In [53]: rollno = {'rollno':[1,2,3,4,5]}

In [54]: rollno
Out[54]: {'rollno': [1, 2, 3, 4, 5]}

In [55]: name = {'name':['A','B','C','D','E']}

In [56]: name
Out[56]: {'name': ['A', 'B', 'C', 'D', 'E']}

In [57]: marks = {'marks':[20,25,14,18,19]}

In [59]:

In [59]: data1 = pd.DataFrame(rollno)

In [60]: data2 = pd.DataFrame(name)

In [61]: pd.concat([data1,data2],axis=1,ignore_index=True)
```

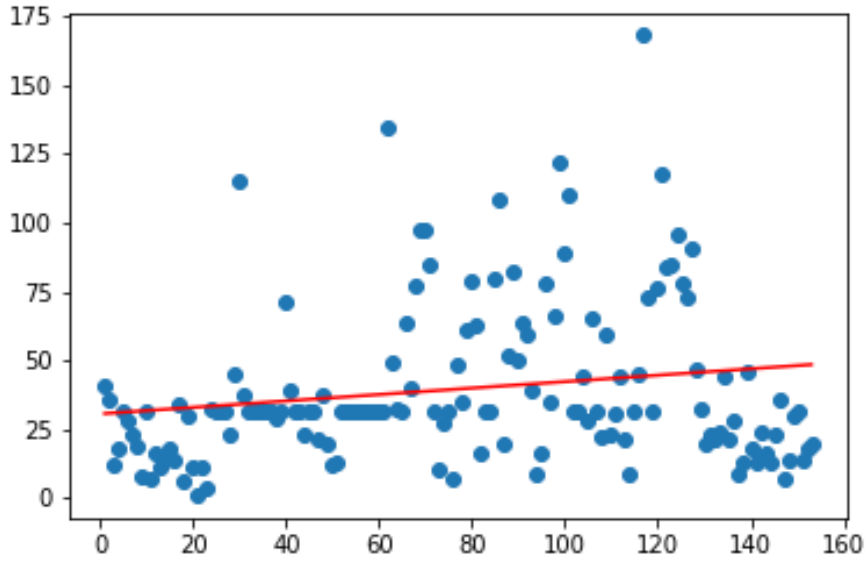
```
Out[61]:
   0  1
0  1  A
1  2  B
2  3  C
3  4  D
4  5  E

In [62]: data3 = pd.DataFrame(marks)

In [63]: pd.concat([data1,data2,data3],axis=1,ignore_index=True)
Out[63]:
   0  1   2
0  1  A  20
1  2  B  25
2  3  C  14
3  4  D  18
4  5  E  19
```

DATA MODEL BUILDING

```
import numpy as np
import matplotlib.pyplot as plt  # To visualize
import pandas as pd  # To read data
from sklearn.linear_model import LinearRegression
data = pd.read_csv('airquality.csv')  # load data set
data["Ozone"].fillna(data['Ozone'].median(),inplace = True)
data["Solar.R"].fillna(data['Solar.R'].median(),inplace = True)
X = data.iloc[:, 0].values.reshape(-1, 1)  # values converts it
into a numpy array
Y = data.iloc[:, 1].values.reshape(-1, 1)  # -1 means that
calculate the dimension of rows, but have 1 column
linear_regressor = LinearRegression()  # create object for the
class
linear_regressor.fit(X, Y)  # perform linear regression
Y_pred = linear_regressor.predict(X)  # make predictions
plt.scatter(X, Y)
plt.plot(X, Y_pred, color='red')
plt.savefig("plot10.png")
plt.show()
```


175
150
125
100
75
50
25
0
0
20
40
60
80
100
120
140
160