ProblemStatement:

Perform the following operations using Python on the Air quality and Heart Diseases data sets

a. Data cleaning

b. Data integration

c. Data transformation

d. Error correcting

e. Data model building

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
df=pd.read_csv('airquality3.csv')
```

In [3]:

```
df
```

Out[3]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp	Month	Day	humidity
0	1	41.0	190.0	7.4	67	5	1	high
1	2	36.0	118.0	8.0	72	5	2	high
2	3	12.0	149.0	12.6	74	5	3	high
3	4	18.0	313.0	11.5	62	5	4	high
4	5	NaN	NaN	14.3	56	5	5	high
...	...	...	...	...	...	...	...	...
148	149	30.0	193.0	6.9	70	9	26	high
149	150	NaN	145.0	13.2	77	9	27	high
150	151	14.0	191.0	14.3	75	9	28	high
151	152	18.0	131.0	8.0	76	9	29	high
152	153	20.0	223.0	11.5	68	9	30	high

153 rows × 8 columns

A)Data Cleaning

Checking Missing Values in Dataframe

In [4]:

```
df.isnull().sum()
```

Out[4]:

```
Unnamed: 0     0
Ozone         37
Solar.R        7
Wind           0
Temp           0
Month          0
Day            0
humidity       4
dtype: int64
```

In [5]:

```
df.drop('Unnamed: 0',axis=1,inplace=True)
```

In [6]:
```
df
```

Out[6]:

	Ozone	Solar.R	Wind	Temp	Month	Day	humidity
0	41.0	190.0	7.4	67	5	1	high
1	36.0	118.0	8.0	72	5	2	high
2	12.0	149.0	12.6	74	5	3	high
3	18.0	313.0	11.5	62	5	4	high
4	NaN	NaN	14.3	56	5	5	high
...	...	...	...	...	...	...	...
148	30.0	193.0	6.9	70	9	26	high
149	NaN	145.0	13.2	77	9	27	high
150	14.0	191.0	14.3	75	9	28	high
151	18.0	131.0	8.0	76	9	29	high
152	20.0	223.0	11.5	68	9	30	high

153 rows × 7 columns

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

Out[7]:

	Ozone	Solar.R	Wind	Temp	Month	Day
count	116.000000	146.000000	153.000000	153.000000	153.000000	153.000000
mean	42.129310	185.931507	9.957516	77.882353	6.993464	15.803922
std	32.987885	90.058422	3.523001	9.465270	1.416522	8.864520
min	1.000000	7.000000	1.700000	56.000000	5.000000	1.000000
25%	18.000000	115.750000	7.400000	72.000000	6.000000	8.000000
50%	31.500000	205.000000	9.700000	79.000000	7.000000	16.000000
75%	63.250000	258.750000	11.500000	85.000000	8.000000	23.000000
max	168.000000	334.000000	20.700000	97.000000	9.000000	31.000000

In [8]:
```
df.shape
```

Out[8]:
```
(153, 7)
```

In [9]:
```
df["Ozone"].fillna(df["Ozone"].mean(),inplace=True)
```

In [10]:
```
df
```

Out[10]:

	Ozone	Solar.R	Wind	Temp	Month	Day	humidity
0	41.00000	190.0	7.4	67	5	1	high
1	36.00000	118.0	8.0	72	5	2	high
2	12.00000	149.0	12.6	74	5	3	high
3	18.00000	313.0	11.5	62	5	4	high
4	42.12931	NaN	14.3	56	5	5	high
...	...	...	...	...	...	...	...
148	30.00000	193.0	6.9	70	9	26	high
149	42.12931	145.0	13.2	77	9	27	high
150	14.00000	191.0	14.3	75	9	28	high
151	18.00000	131.0	8.0	76	9	29	high
152	20.00000	223.0	11.5	68	9	30	high

153 rows × 7 columns

In [11]:
```
df["Solar.R"].fillna(df["Solar.R"].mean(),inplace=True)
```

In [12]:
```
df
```

Out[12]:

	Ozone	Solar.R	Wind	Temp	Month	Day	humidity
0	41.00000	190.000000	7.4	67	5	1	high
1	36.00000	118.000000	8.0	72	5	2	high
2	12.00000	149.000000	12.6	74	5	3	high
3	18.00000	313.000000	11.5	62	5	4	high
4	42.12931	185.931507	14.3	56	5	5	high
...	...	...	...	...	...	...	...
148	30.00000	193.000000	6.9	70	9	26	high
149	42.12931	145.000000	13.2	77	9	27	high
150	14.00000	191.000000	14.3	75	9	28	high
151	18.00000	131.000000	8.0	76	9	29	high
152	20.00000	223.000000	11.5	68	9	30	high

153 rows × 7 columns

In [13]:
```
df["humidity"].mode()
```

Out[13]:
```
0    high
Name: humidity, dtype: object
```

In [14]:
```
df["humidity"]=df["humidity"].fillna('high')
```

In [15]:

```
df
```

Out[15]:

	Ozone	Solar.R	Wind	Temp	Month	Day	humidity
0	41.00000	190.000000	7.4	67	5	1	high
1	36.00000	118.000000	8.0	72	5	2	high
2	12.00000	149.000000	12.6	74	5	3	high
3	18.00000	313.000000	11.5	62	5	4	high
4	42.12931	185.931507	14.3	56	5	5	high
...	...	...	...	...	...	...	...
148	30.00000	193.000000	6.9	70	9	26	high
149	42.12931	145.000000	13.2	77	9	27	high
150	14.00000	191.000000	14.3	75	9	28	high
151	18.00000	131.000000	8.0	76	9	29	high
152	20.00000	223.000000	11.5	68	9	30	high

153 rows × 7 columns

In [16]:

```
df.isnull().sum()
```

Out[16]:

```
Ozone       0
Solar.R     0
Wind        0
Temp        0
Month       0
Day         0
humidity    0
dtype: int64
```

B) Data Integration

In [46]:

```
subset1=df[['Ozone','Solar.R','Wind','Temp']].loc[0:15]
```

In [47]:

```
subset1
```

Out[47]:

	Ozone	Solar.R	Wind	Temp
0	41.00000	190.000000	7.4	67
1	36.00000	118.000000	8.0	72
2	12.00000	149.000000	12.6	74
3	18.00000	313.000000	11.5	62
4	42.12931	185.931507	14.3	56
5	28.00000	185.931507	14.9	66
6	23.00000	299.000000	8.6	65
7	19.00000	99.000000	13.8	59
8	8.00000	19.000000	20.1	61
9	42.12931	194.000000	8.6	69
10	7.00000	185.931507	6.9	74
11	16.00000	256.000000	9.7	69
12	11.00000	290.000000	9.2	66
13	14.00000	274.000000	10.9	68
14	18.00000	65.000000	13.2	58
15	14.00000	334.000000	11.5	64

In [48]:
```
subset2=df[['Ozone','Solar.R','Wind','Temp']].loc[16:30]
```

In [49]:
```
subset2
```

Out[49]:

	Ozone	Solar.R	Wind	Temp
16	34.00000	307.000000	12.0	66
17	6.00000	78.000000	18.4	57
18	30.00000	322.000000	11.5	68
19	11.00000	44.000000	9.7	62
20	1.00000	8.000000	9.7	59
21	11.00000	320.000000	16.6	73
22	4.00000	25.000000	9.7	61
23	32.00000	92.000000	12.0	61
24	42.12931	66.000000	16.6	57
25	42.12931	266.000000	14.9	58
26	42.12931	185.931507	8.0	57
27	23.00000	13.000000	12.0	67
28	45.00000	252.000000	14.9	81
29	115.00000	223.000000	5.7	79
30	37.00000	279.000000	7.4	76

In [50]:
```
merge=pd.concat([subset1,subset2])
```

In [51]:

```
merge
```

Out[51]:

	Ozone	Solar.R	Wind	Temp
0	41.00000	190.000000	7.4	67
1	36.00000	118.000000	8.0	72
2	12.00000	149.000000	12.6	74
3	18.00000	313.000000	11.5	62
4	42.12931	185.931507	14.3	56
5	28.00000	185.931507	14.9	66
6	23.00000	299.000000	8.6	65
7	19.00000	99.000000	13.8	59
8	8.00000	19.000000	20.1	61
9	42.12931	194.000000	8.6	69
10	7.00000	185.931507	6.9	74
11	16.00000	256.000000	9.7	69
12	11.00000	290.000000	9.2	66
13	14.00000	274.000000	10.9	68
14	18.00000	65.000000	13.2	58
15	14.00000	334.000000	11.5	64
16	34.00000	307.000000	12.0	66
17	6.00000	78.000000	18.4	57
18	30.00000	322.000000	11.5	68
19	11.00000	44.000000	9.7	62
20	1.00000	8.000000	9.7	59
21	11.00000	320.000000	16.6	73
22	4.00000	25.000000	9.7	61
23	32.00000	92.000000	12.0	61
24	42.12931	66.000000	16.6	57
25	42.12931	266.000000	14.9	58
26	42.12931	185.931507	8.0	57
27	23.00000	13.000000	12.0	67
28	45.00000	252.000000	14.9	81
29	115.00000	223.000000	5.7	79
30	37.00000	279.000000	7.4	76

c) Data Transformation

In [52]:

```
from sklearn import preprocessing
```

In [53]:

```
Label_Encoder=preprocessing.LabelEncoder()
```

In [54]:

```
df["humidity"] = Label_Encoder.fit_transform(df["humidity"])
```

In [55]:

```
df['humidity']
```

Out[55]:

```
0      0
1      0
2      0
3      0
4      0
      ..
148    0
149    0
150    0
151    0
152    0
Name: humidity, Length: 153, dtype: int64
```

In [56]:

```
df['humidity'].values
```

Out[56]:

```
array([0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2,
       2, 2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      dtype=int64)
```

In [61]:

```
df["humidity"].count()
```

Out[61]:

```
153
```

D)Data model building

In [62]:

```
from sklearn.model_selection import train_test_split
```

In [63]:

```
X=df[['Ozone']]
Y=df[['Temp']]
```

In [64]:

```
Xtrain,Xtest,Ytrain,Ytest=train_test_split(X,Y,test_size=0.3,random_state=42)
```

In [65]:

```
Xtrain.shape
```

Out[65]:

```
(107, 1)
```

In [66]:

```
Xtest.shape
```

Out[66]:

```
(46, 1)
```

In [67]:

```
Ytrain.shape
```

Out[67]:

```
(107, 1)
```

In [68]:

```
Ytest.shape
```

Out[68]:

```
(46, 1)
```

In [69]:

```
from sklearn import linear_model
```

In [70]:

```
reg=linear_model.LinearRegression()
```

```
In [71]: model=reg.fit(Xtrain,Ytrain)
```

```
In [72]: model
```

```
Out[72]: LinearRegression()
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
In [73]: print(model.intercept_)
         print(model.coef_)
```

```
[69.42232301]
[[0.20312469]]
```

```
In [74]: Ypred=model.predict(Xtest)
```

```
In [76]: print(Ypred)
```

```
[[85.67229785]
 [73.48481672]
 [82.82855225]
 [78.56293386]
 [92.78166184]
 [77.97982592]
 [81.81292882]
 [72.67231798]
 [75.51606357]
 [72.26606861]
 [71.65669455]
 [71.65669455]
 [77.97982592]
 [77.97982592]
 [77.97982592]
 [89.1254175 ]
 [74.09419078]
 [71.25044518]
 [82.42230288]
 [76.531687  ]
 [70.84419581]
 [88.92229281]
 [77.97982592]
 [77.54731043]
 [77.97982592]
 [76.93793637]
 [77.97982592]
 [73.6879414 ]
 [75.1098142 ]
 [77.97982592]
 [71.25044518]
 [72.67231798]
 [89.1254175 ]
 [70.23482175]
 [77.97982592]
 [79.17230791]
 [72.67231798]
 [73.6879414 ]
 [74.29731546]
 [75.92231294]
 [74.09419078]
 [77.97982592]
 [72.67231798]
 [78.96918323]
 [77.97982592]
 [77.97982592]]
```

In [43]:

```
from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score
mse = mean_squared_error(Ytest,Ypred)
rmse = np.sqrt(mse)
mae = mean_absolute_error(Ytest,Ypred)
r2 = r2_score(Ytest,Ypred)
ar2 = 1-(1-r2) * (len(Y)-1) / (len(Y)-1-1)
```

In [44]:

```
print('Mean squared error: ',mse)
print('Root Mean Square error: ',rmse)
print('Mean Absolute error: ',mae)
print('R2: ',r2)
print('Adjusted  R2: ',ar2)
```

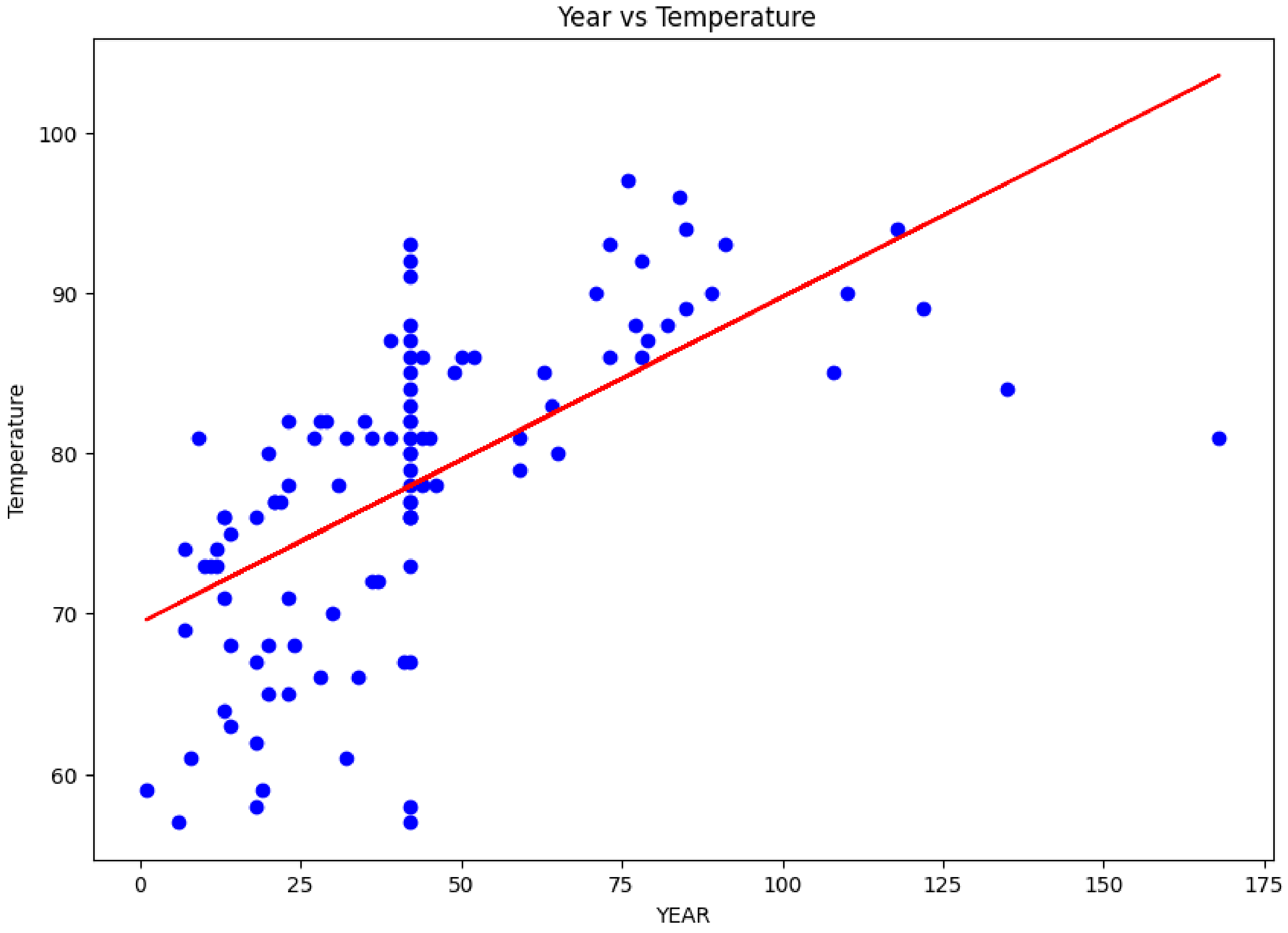
```
Mean squared error:  54.13493693361354
Root Mean Square error:  7.357644795286976
Mean Absolute error:  5.337665610314505
R2:  0.3049896154430294
Adjusted  R2:  0.30038689766450644
```

In [45]:

```
import matplotlib.pyplot as plt
plt.figure(figsize=(10,7))
plt.title(' Year vs Temperature')
plt.xlabel('YEAR')
plt.ylabel('Temperature')
plt.plot(Xtrain,reg.predict(Xtrain),color='red')
plt.scatter(Xtrain,Ytrain,color='blue')
```

Out[45]:

```
<matplotlib.collections.PathCollection at 0x1ee295fdab0>
```


Year vs Temperature
Temperature
YEAR