Problem Statement:- 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

Importing libraries and reading dataset

In [78]:
```
import pandas as pd
import numpy as np
```

In [79]:
```
data=pd.read_csv("airquality.csv")
data
```

Out[79]:

	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	Medium
3	4	18.0	313.0	11.5	62	5	4	Medium
4	5	NaN	NaN	14.3	56	5	5	NaN
...	...	...	...	...	...	...	...	...
148	149	30.0	193.0	6.9	70	9	26	Low
149	150	NaN	145.0	13.2	77	9	27	Low
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	Medium

153 rows × 8 columns

Removing the Unnamed column

In [80]:
```
data.drop("Unnamed: 0",axis=1)
```

Out[80]:

	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	Medium
3	18.0	313.0	11.5	62	5	4	Medium
4	NaN	NaN	14.3	56	5	5	NaN
...	...	...	...	...	...	...	...
148	30.0	193.0	6.9	70	9	26	Low
149	NaN	145.0	13.2	77	9	27	Low
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	Medium

153 rows × 7 columns

Sum of null values in each column

In [81]:
```
data.isnull().sum()
```

Out[81]:
```
Unnamed: 0     0
Ozone         37
Solar.R        7
Wind           0
Temp           0
Month          0
Day            0
Humidity       6
dtype: int64
```

Replacing null values

In [82]:
```
data["Humidity"].fillna("Medium",inplace=True)
data["Ozone"].fillna(data["Ozone"].mean(),inplace=True)
data["Solar.R"].fillna(data["Solar.R"].mean(),inplace=True)
data
```

Out[82]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp	Month	Day	Humidity
0	1	41.00000	190.000000	7.4	67	5	1	High
1	2	36.00000	118.000000	8.0	72	5	2	High
2	3	12.00000	149.000000	12.6	74	5	3	Medium
3	4	18.00000	313.000000	11.5	62	5	4	Medium
4	5	42.12931	185.931507	14.3	56	5	5	Medium
...	...	...	...	...	...	...	...	...
148	149	30.00000	193.000000	6.9	70	9	26	Low
149	150	42.12931	145.000000	13.2	77	9	27	Low
150	151	14.00000	191.000000	14.3	75	9	28	High
151	152	18.00000	131.000000	8.0	76	9	29	High
152	153	20.00000	223.000000	11.5	68	9	30	Medium

153 rows × 8 columns

In [83]:
```
data.isnull().sum()
```

Out[83]:
```
Unnamed: 0    0
Ozone         0
Solar.R       0
Wind          0
Temp          0
Month         0
Day           0
Humidity      0
dtype: int64
```

Label Encoding on column Humidity

In [84]:
```
from sklearn.preprocessing import LabelEncoder
label_Encoder=LabelEncoder()
data["Humidity"]=label_Encoder.fit_transform(data["Humidity"])
data
```

Out[84]:

	Unnamed: 0	Ozone	Solar.R	Wind	Temp	Month	Day	Humidity
0	1	41.00000	190.000000	7.4	67	5	1	0
1	2	36.00000	118.000000	8.0	72	5	2	0
2	3	12.00000	149.000000	12.6	74	5	3	2
3	4	18.00000	313.000000	11.5	62	5	4	2
4	5	42.12931	185.931507	14.3	56	5	5	2
...	...	...	...	...	...	...	...	...
148	149	30.00000	193.000000	6.9	70	9	26	1
149	150	42.12931	145.000000	13.2	77	9	27	1
150	151	14.00000	191.000000	14.3	75	9	28	0
151	152	18.00000	131.000000	8.0	76	9	29	0
152	153	20.00000	223.000000	11.5	68	9	30	2

153 rows × 8 columns

Assigning Variables

In [85]:
```
x=data[["Day"]]
y=data[["Temp"]]
```

Splitting dataset into Training and Testing part

In [86]:
```
from sklearn.model_selection import train_test_split
Xtrain,Xtest,Ytrain,Ytest=train_test_split(x,y,test_size=0.2)
```

Creating and Training Linear Regression Model

In [87]:
```
from sklearn.linear_model import LinearRegression
model=LinearRegression()
model.fit(Xtrain,Ytrain)
```

Out[87]:
```
▾ LinearRegression

LinearRegression()
```

Predicting Value

In [88]:
```
predict=model.predict(Xtest)
```

Mean Squared Error

In [89]:
```
from sklearn.metrics import mean_squared_error
mse=mean_squared_error(predict,Ytest)
mse
```

Out[89]:
```
96.80759948604907
```

Root Mean Square Error

In [90]:
```
rmse=np.sqrt(mse)
rmse
```

Out[90]:
```
9.83908529722398
```

Visualization of Model

In [91]:
```
import matplotlib.pyplot as plt

plt.scatter(Xtrain,Ytrain,color="blue")
plt.title("Temperature vs Day Graph")
plt.xlabel("Day")
plt.ylabel("Temperature")
plt.plot(Xtrain,model.predict(Xtrain),color="red")
```

Out[91]:
```
[<matplotlib.lines.Line2D at 0x284da2f2d10>]
```