Name:Vinayak V Thayil

Roll No:AM.EN.U4CSE21161

```
import pandas as pd
df=pd.read_csv('airquality.csv')
print("Shape of the dataset:",df.shape)
print("Structure of the dataset:",df.info())
```

```
Shape of the dataset: (29531, 16)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29531 entries, 0 to 29530
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   City        29531 non-null  object 
 1   Date        29531 non-null  object 
 2   PM2.5       24933 non-null  float64
 3   PM10        18391 non-null  float64
 4   NO          25949 non-null  float64
 5   NO2         25946 non-null  float64
 6   NOx         25346 non-null  float64
 7   NH3         19203 non-null  float64
 8   CO          27472 non-null  float64
 9   SO2         25677 non-null  float64
 10  O3          25509 non-null  float64
 11  Benzene     23908 non-null  float64
 12  Toluene     21490 non-null  float64
 13  Xylene      11422 non-null  float64
 14  AQI         24850 non-null  float64
 15  AQI_Bucket  24850 non-null  object 
dtypes: float64(13), object(3)
memory usage: 3.6+ MB
Structure of the dataset: None
```

```
print("Variables of the dataset:",df.columns)
```

```
Variables of the dataset: Index(['City', 'Date', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2',
       'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI', 'AQI_Bucket'],
      dtype='object')
```

```
print("Data type of AQI_Bucket:",df['AQI_Bucket'].dtype)
```

```
Data type of AQI_Bucket: object
```

```
print("Features Distributed:",df.describe())
```

```
Features Distributed:               PM2.5          PM10            NO           NO2           NOx  \
count  24933.000000  18391.000000  25949.000000  25946.000000  25346.000000
mean      67.450578    118.127103     17.574730     28.560659     32.309123
std       64.661449     90.605110     22.785846     24.474746     31.646011
min        0.040000      0.010000      0.020000      0.010000      0.000000
25%       28.820000     56.255000      5.630000     11.750000     12.820000
50%       48.570000     95.680000      9.890000     21.690000     23.520000
75%       80.590000    149.745000     19.950000     37.620000     40.127500
max      949.990000   1000.000000    390.680000    362.210000    467.630000

                NH3            CO           SO2            O3       Benzene  \
count  19203.000000  27472.000000  25677.000000  25509.000000  23908.000000
mean      23.483476      2.248598     14.531977     34.491430      3.280840
std       25.684275      6.962884     18.133775     21.694928     15.811136
min        0.010000      0.000000      0.010000      0.010000      0.000000
25%        8.580000      0.510000      5.670000     18.860000      0.120000
50%       15.850000      0.890000      9.160000     30.840000      1.070000
75%       30.020000      1.450000     15.220000     45.570000      3.080000
max      352.890000    175.810000    193.860000    257.730000    455.030000

             Toluene        Xylene           AQI
count  21490.000000  11422.000000  24850.000000
mean       8.700972      3.070128    166.463581
std       19.969164      6.323247    140.696585
min        0.000000      0.000000     13.000000
25%        0.600000      0.140000     81.000000
50%        2.970000      0.980000    118.000000
75%        9.150000      3.350000    208.000000
max      454.850000    170.370000   2049.000000
```

```
print("Missing Values:",df.isnull().sum())
```

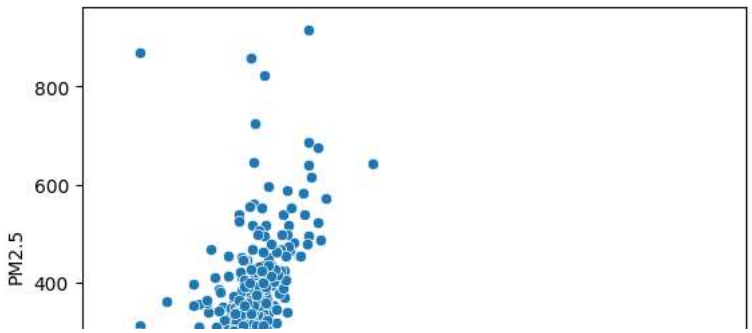
```
Missing Values: City              0
Date                0
PM2.5            4598
PM10            11140
NO               3582
NO2              3585
NOx              4185
NH3             10328
CO               2059
SO2              3854
O3               4022
Benzene          5623
Toluene          8041
Xylene          18109
AQI              4681
AQI_Bucket       4681
dtype: int64
```

```
df_imputed = df.fillna(df.mean())
df_cleaned = df.dropna()
df_cleaned_columns = df.dropna(axis=1)
print("\nShape of the imputed dataset:", df_imputed.shape)
print("Shape of the dataset after removing rows with missing values:", df_cleaned.shape)
print("Shape of the dataset after removing columns with missing values:", df_cleaned_columns.shape)
```

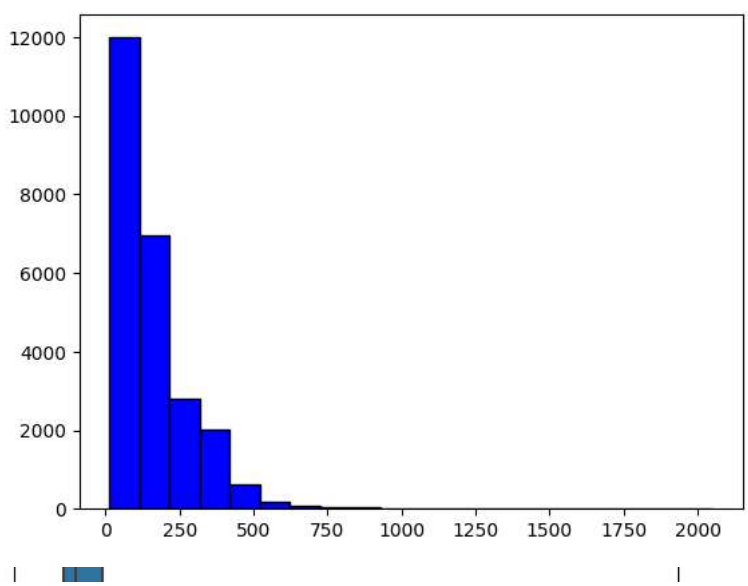
```
Shape of the imputed dataset: (29531, 16)
Shape of the dataset after removing rows with missing values: (6236, 16)
Shape of the dataset after removing columns with missing values: (29531, 2)
<ipython-input-9-df5c92da1ebc>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version,
  df_imputed = df.fillna(df.mean())
```

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.scatterplot(x='AQI', y='PM2.5', data=df)
plt.show()
sns.boxplot(x=df['AQI'])
plt.show()
```




```
plt.hist(df['AQI'], bins=20, color='blue', edgecolor='black')
plt.show()
```




```
correlation_matrix = df.corr()
print("\nCorrelation matrix:")
print(correlation_matrix)
```

```
Correlation matrix:
             PM2.5      PM10        NO       NO2       NOx       NH3        CO  \
PM2.5     1.000000  0.846498  0.433491  0.350709  0.436792  0.275086  0.089912
PM10      0.846498  1.000000  0.502349  0.464380  0.527768  0.376816  0.112588
NO        0.433491  0.502349  1.000000  0.478070  0.794890  0.185621  0.212607
NO2       0.350709  0.464380  0.478070  1.000000  0.627627  0.234938  0.356521
NOx       0.436792  0.527768  0.794890  0.627627  1.000000  0.166224  0.226992
NH3       0.275086  0.376816  0.185621  0.234938  0.166224  1.000000  0.104891
CO        0.089912  0.112588  0.212607  0.356521  0.226992  0.104891  1.000000
SO2       0.132325  0.256974  0.170322  0.392233  0.238397 -0.038998  0.489697
O3        0.161238  0.244919  0.014580  0.293349  0.093170  0.094972  0.041736
Benzene   0.023911  0.022265  0.035771  0.025260  0.039121 -0.015650  0.061861
Toluene   0.117080  0.169335  0.150857  0.273926  0.189386  0.013227  0.277904
Xylene    0.114579  0.081700  0.094237  0.171701  0.087398 -0.019813  0.154889
AQI       0.659181  0.803313  0.452191  0.537071  0.486450  0.252019  0.683346

              SO2        O3   Benzene   Toluene    Xylene       AQI
PM2.5     0.132325  0.161238  0.023911  0.117080  0.114579  0.659181
PM10      0.256974  0.244919  0.022265  0.169335  0.081700  0.803313
NO        0.170322  0.014580  0.035771  0.150857  0.094237  0.452191
NO2       0.392233  0.293349  0.025260  0.273926  0.171701  0.537071
NOx       0.238397  0.093170  0.039121  0.189386  0.087398  0.486450
NH3      -0.038998  0.094972 -0.015650  0.013227 -0.019813  0.252019
CO        0.489697  0.041736  0.061861  0.277904  0.154889  0.683346
SO2       1.000000  0.162142  0.036110  0.296139  0.251195  0.490586
O3        0.162142  1.000000  0.020255  0.130209  0.111410  0.198991
Benzene   0.036110  0.020255  1.000000  0.739286  0.415427  0.044407
Toluene   0.296139  0.130209  0.739286  1.000000  0.421432  0.279992
Xylene    0.251195  0.111410  0.415427  0.421432  1.000000  0.165532
AQI       0.490586  0.198991  0.044407  0.279992  0.165532  1.000000
<ipython-input-12-248659e80400>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future versior
  correlation_matrix = df.corr()
```