Steps in Data Preprocessing Step 1: Import the necessary libraries

```
# importing libraries
import pandas as pd
import scipy
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
import matplotlib.pyplot as plt
```

Step 2: Load the dataset

```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
# Load the dataset
df = pd.read_csv('/content/gdrive/MyDrive/ADS_PHASE 3.csv')
print(df.head())
```

```
   Benzene  Eth-Benzene  MP-Xylene      BP  O Xylene    PM10  PM2.5     RH  \
0     1.08         0.04       0.00  754.05      2.50  140.23  90.62  42.51
1     0.83         0.03       0.00  754.28      1.74  124.91  61.11  28.34
2     1.38         0.27       0.06  754.49      2.39  114.27  70.89  36.48
3     1.97         0.47       0.12  754.28      3.51  128.15  78.52  43.68
4     1.80         0.75       0.20  754.00      4.00  122.36  70.48  51.57

       SR   Temp      WD    WS    CO    NH3     NO    NO2    NOx  Ozone    SO2
0  125.03  17.90  119.19  0.97  0.51  20.53   4.29  22.95  27.24  44.36   4.97
1  148.95  20.04   71.50  1.21  0.53  17.37   2.80  25.59  28.38  53.04   5.59
2  131.87  18.31  147.10  1.00  0.78  18.45   6.85  30.91  37.75  41.94  10.22
3  129.32  18.56  182.79  1.06  0.81  22.52   7.36  29.05  36.41  44.15  30.99
4  145.73  18.81  183.47  0.91  0.96  19.14  13.15  28.60  41.74  37.13  15.78
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 298 entries, 0 to 297
Data columns (total 19 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Benzene      298 non-null    float64
 1   Eth-Benzene  298 non-null    float64
 2   MP-Xylene    298 non-null    float64
 3   BP           298 non-null    float64
 4   O Xylene     298 non-null    float64
 5   PM10         298 non-null    float64
 6   PM2.5        298 non-null    float64
 7   RH           298 non-null    float64
 8   SR           298 non-null    float64
 9   Temp         298 non-null    float64
 10  WD           298 non-null    float64
 11  WS           298 non-null    float64
 12  CO           298 non-null    float64
 13  NH3          298 non-null    float64
 14  NO           298 non-null    float64
 15  NO2          298 non-null    float64
 16  NOx          298 non-null    float64
 17  Ozone        298 non-null    float64
 18  SO2          298 non-null    float64
dtypes: float64(19)
memory usage: 44.4 KB
```

```
df.head()
```

	Benzene	Eth-Benzene	MP-Xylene	BP	O Xylene	PM10	PM2.5	RH	SR	Temp	WD	WS	CO	NH3	NO	NO2	NOx
0	1.08	0.04	0.00	754.05	2.50	140.23	90.62	42.51	125.03	17.90	119.19	0.97	0.51	20.53	4.29	22.95	27.24
1	0.83	0.03	0.00	754.28	1.74	124.91	61.11	28.34	148.95	20.04	71.50	1.21	0.53	17.37	2.80	25.59	28.38
2	1.38	0.27	0.06	754.49	2.39	114.27	70.89	36.48	131.87	18.31	147.10	1.00	0.78	18.45	6.85	30.91	37.75
3	1.97	0.47	0.12	754.28	3.51	128.15	78.52	43.68	129.32	18.56	182.79	1.06	0.81	22.52	7.36	29.05	36.41
4	1.80	0.75	0.20	754.00	4.00	122.36	70.48	51.57	145.73	18.81	183.47	0.91	0.96	19.14	13.15	28.60	41.74

```
df.tail()
```

	Benzene	Eth-Benzene	MP-Xylene	BP	O Xylene	PM10	PM2.5	RH	SR	Temp	WD	WS	CO	NH3	NO	NO2	NOx
293	1.31	0.22	0.08	757.14	2.57	178.71	81.02	29.44	207.84	28.43	210.70	1.01	1.01	23.42	6.62	32.33	38.13
294	0.64	0.01	0.00	756.64	2.24	158.81	76.39	32.40	214.16	28.21	132.54	1.20	0.62	21.50	8.26	25.42	33.68
295	0.71	0.07	0.03	756.01	2.03	138.19	63.19	30.18	202.04	28.75	172.23	1.18	0.72	19.76	8.66	27.63	36.29
296	0.89	0.10	0.04	755.72	2.51	144.54	58.81	29.82	206.48	29.82	214.90	1.36	0.83	19.35	9.01	27.79	36.80
297	1.10	0.20	0.08	756.84	2.82	152.39	62.50	34.85	167.95	28.89	169.89	0.97	0.91	17.60	8.92	33.32	42.24

STEP 3: check the null values

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

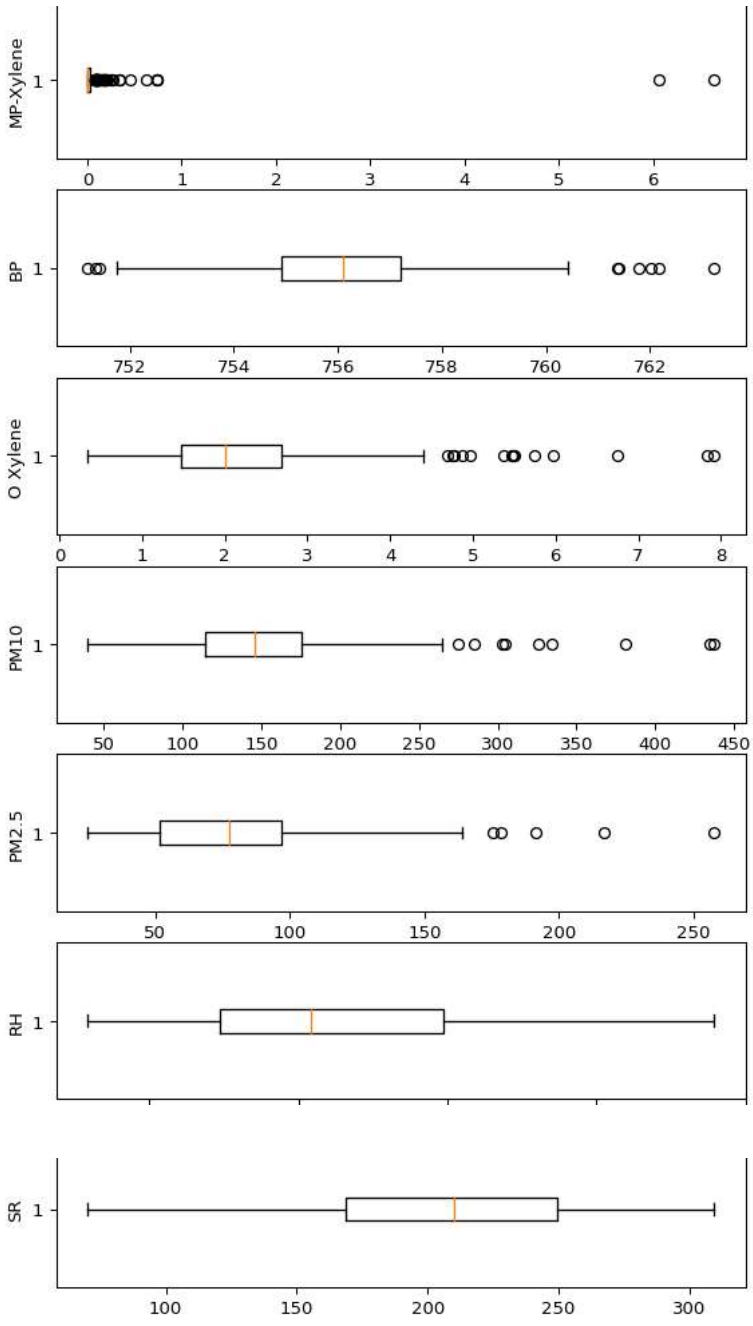
```
Benzene        0
Eth-Benzene    0
MP-Xylene      0
BP             0
O Xylene       0
PM10           0
PM2.5          0
RH             0
SR             0
Temp           0
WD             0
WS             0
CO             0
NH3            0
NO             0
NO2            0
NOx            0
Ozone          0
SO2            0
dtype: int64
```

```
df.describe()
```

	SR	Temp	
	298.000000	298.00000	2
	208.083020	30.42802	1
	50.891946	5.45345	
	69.960000	16.67000	
	168.785000	28.14250	1
	210.255000	30.86500	2
	249.402500	34.15250	2
	309.610000	40.07000	2

Step 5: Check the outliers

```
# Box Plots
fig, axs = plt.subplots(9,1,dpi=95, figsize=(7,17))
i = 0
for col in df.columns:
    axs[i].boxplot(df[col], vert=False)
    axs[i].set_ylabel(col)
    i+=1
plt.show()
```


MP-Xylene
1
0 1 2 3 4 5 6
BP
1
752 754 756 758 760 762
O Xylene
1
0 1 2 3 4 5 6 7 8
PM10
1
50 100 150 200 250 300 350 400 450
PM2.5
1
50 100 150 200 250
RH
1
SR
1
100 150 200 250 300