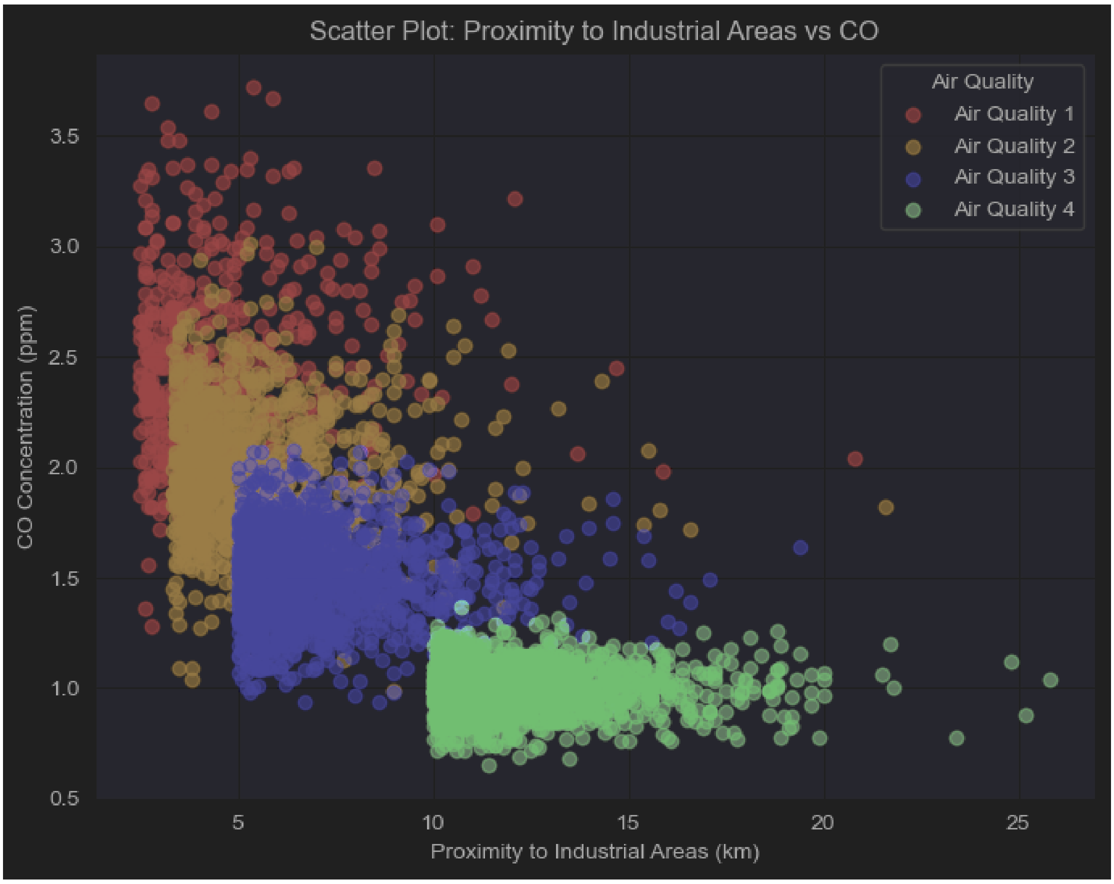
Assignment 1 Report



1. Proximity to Industrial Areas vs. CO

Correlation: Quite strong negative overall relationship between 2 attributes.

Separation between classes: Green is clearly sperate from the three others; red, yellow, and blue has significant overlap.

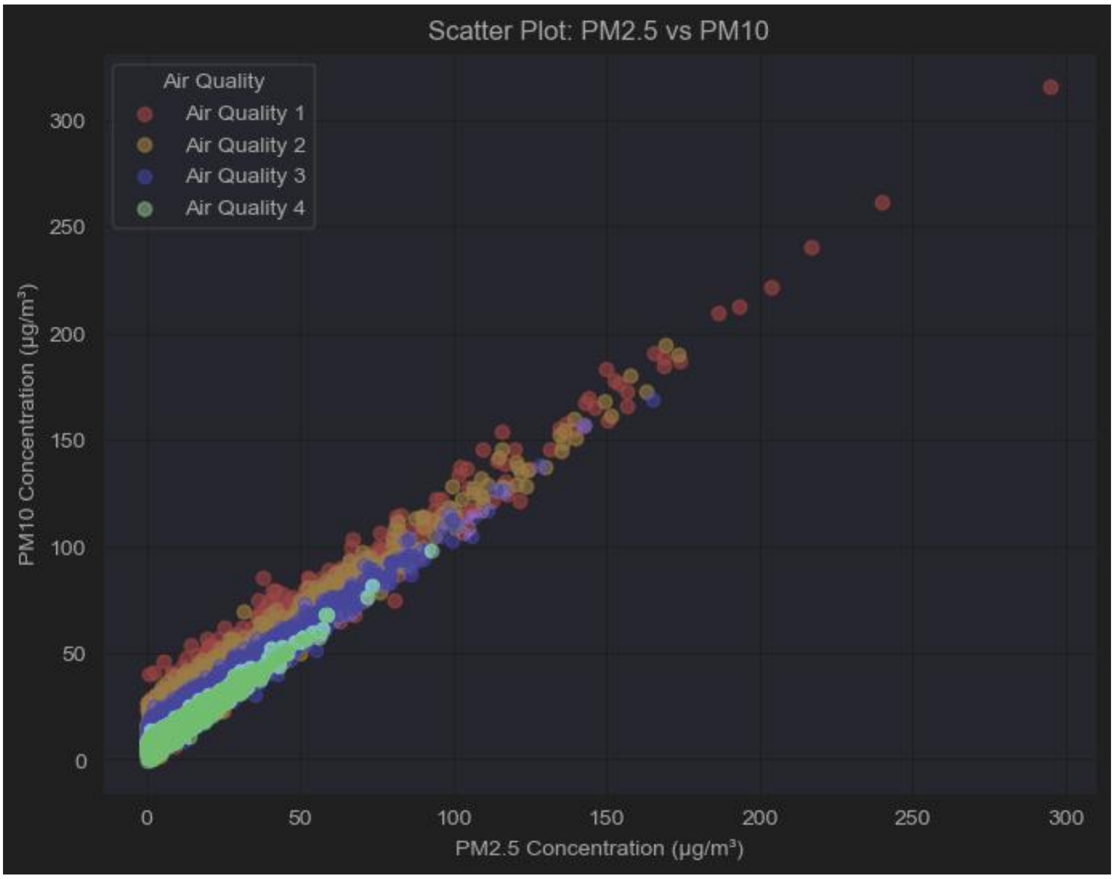
Decision Boundary: Because of the overlap between the three classes AQ1, AQ2 and AQ3, it is more appropriate to use nonlinear decision boundaries. However, for the linear boundary works for AQ4 to separate between others 3 classes.

Difficulty in classification: Hard

Location of each class:

- AQ1 (Red): [1, 11] * [1.3, 3.7]
- AQ2 (Yellow): [2, 13] * [1.0, 3.0]
- AQ3 (Blue): [5, 17] * [1.0, 5.0]
- AQ4 (Green): [10, 27] * [0.7, 1.3]

Attribute Proximity to Industrial Areas works well to separate AQ1 with AQ2, AQ3 and AQ4. Attribute CO works well to separate AQ1 with AQ2, AQ3 and AQ4 AQ1 with AQ2, AQ3 and AQ4



2. PM2.5 vs PM10

Correlation: Quite a strong positive overall correlation between two attributes.

Separation between classes: There is not a clear separation between classes.

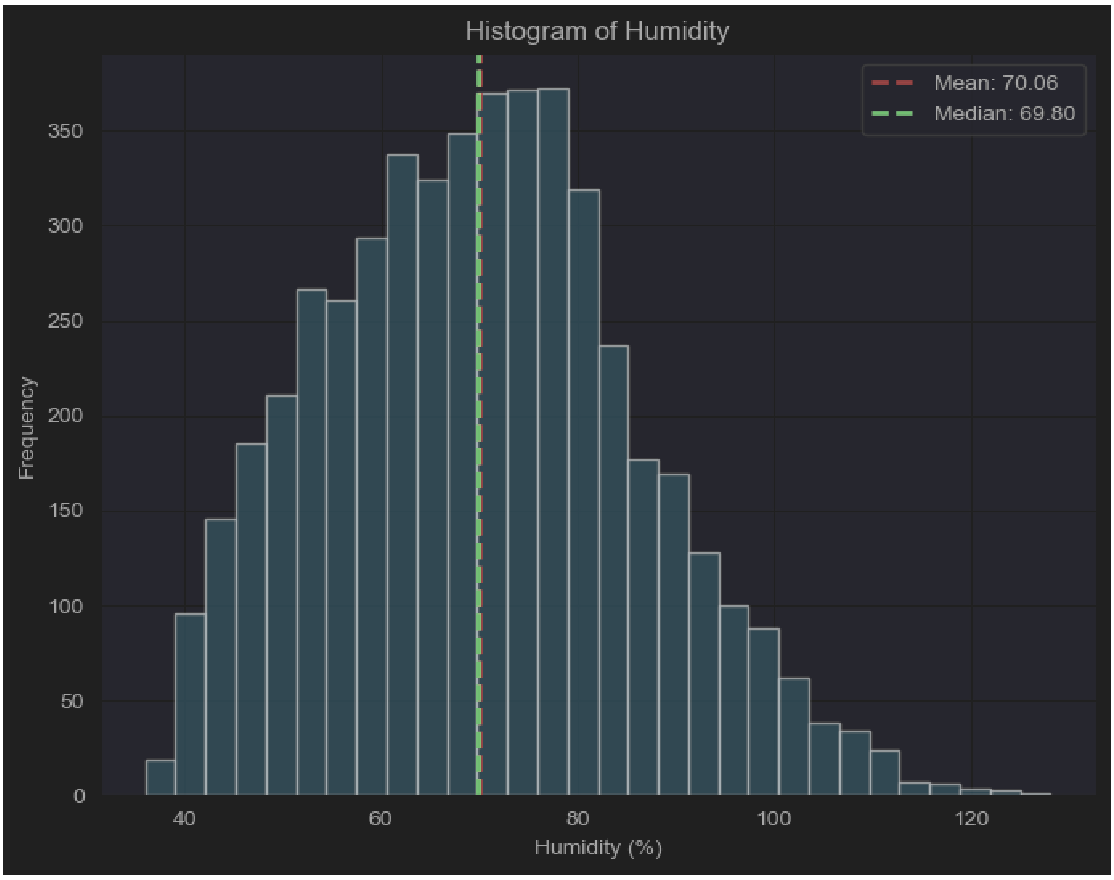
Decision Boundary: Because of the overlap between the four classes, it is more appropriate to use nonlinear decision boundaries

Difficulty in classification: Hard

Location of each class:

- AQ1 (Red): [0, 240] * [0, 260]
- AQ2 (Yellow): [0, 170] * [0, 199]
- AQ3 (Blue): [0, 160] * [0, 160]
- AQ4 (Green): [0,99] * [0,100]

Attribute PM10 and PM2.5 works good to separate AQ1 with AQ2, AQ3 and AQ4



Mean Humidity: 70.06

Standard Deviation: 15.86

Median Humidity: 69.80

3. Histogram of Humidity

Type of Attribute: Positive numbers.

Spread: Medium spread.

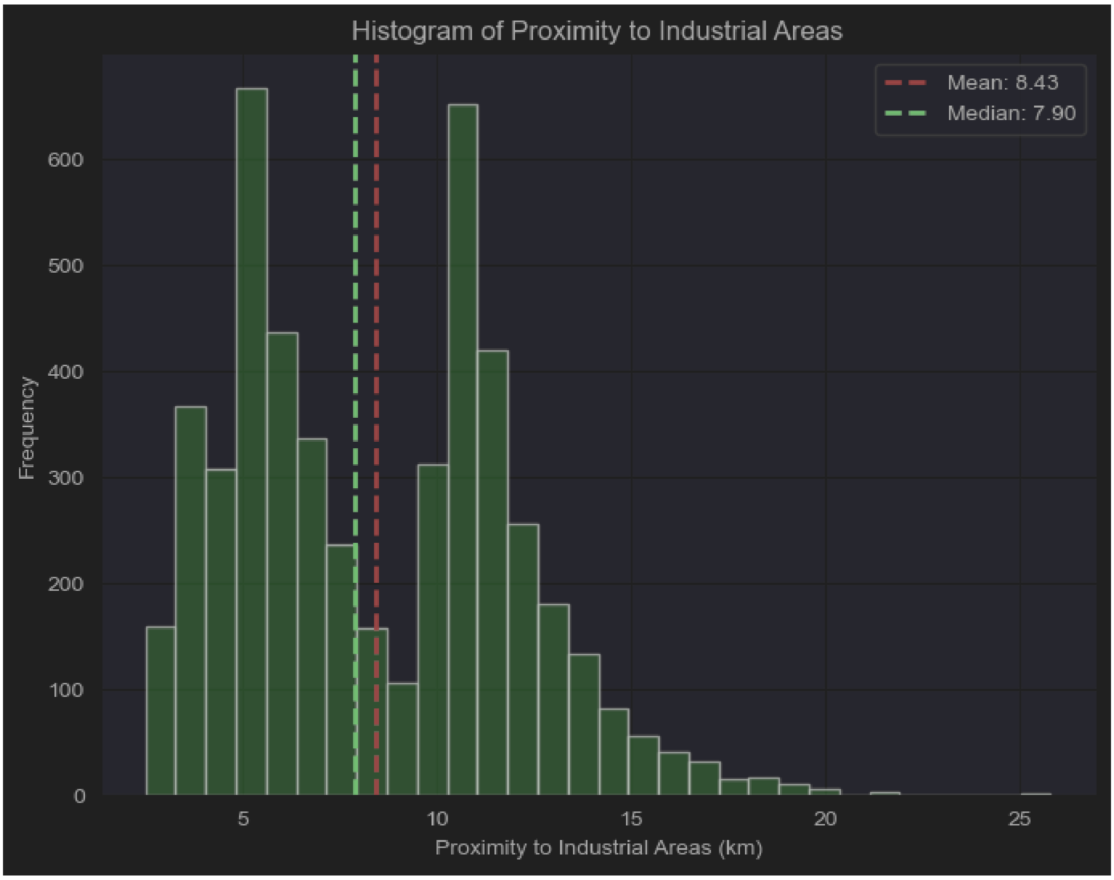
Pattern: Unimodal (One hill)

Skewed: Mildly skewed right (Mean > Median)

Outliers: None

Gap: None

Shape: Bell curve



Mean Proximity: 8.43 km

Standard Deviation: 3.61 km

Median Proximity: 7.90 km

4. Histogram of Proximity to Industrial Areas

Type of Attribute: Positive numbers.

Spread: High

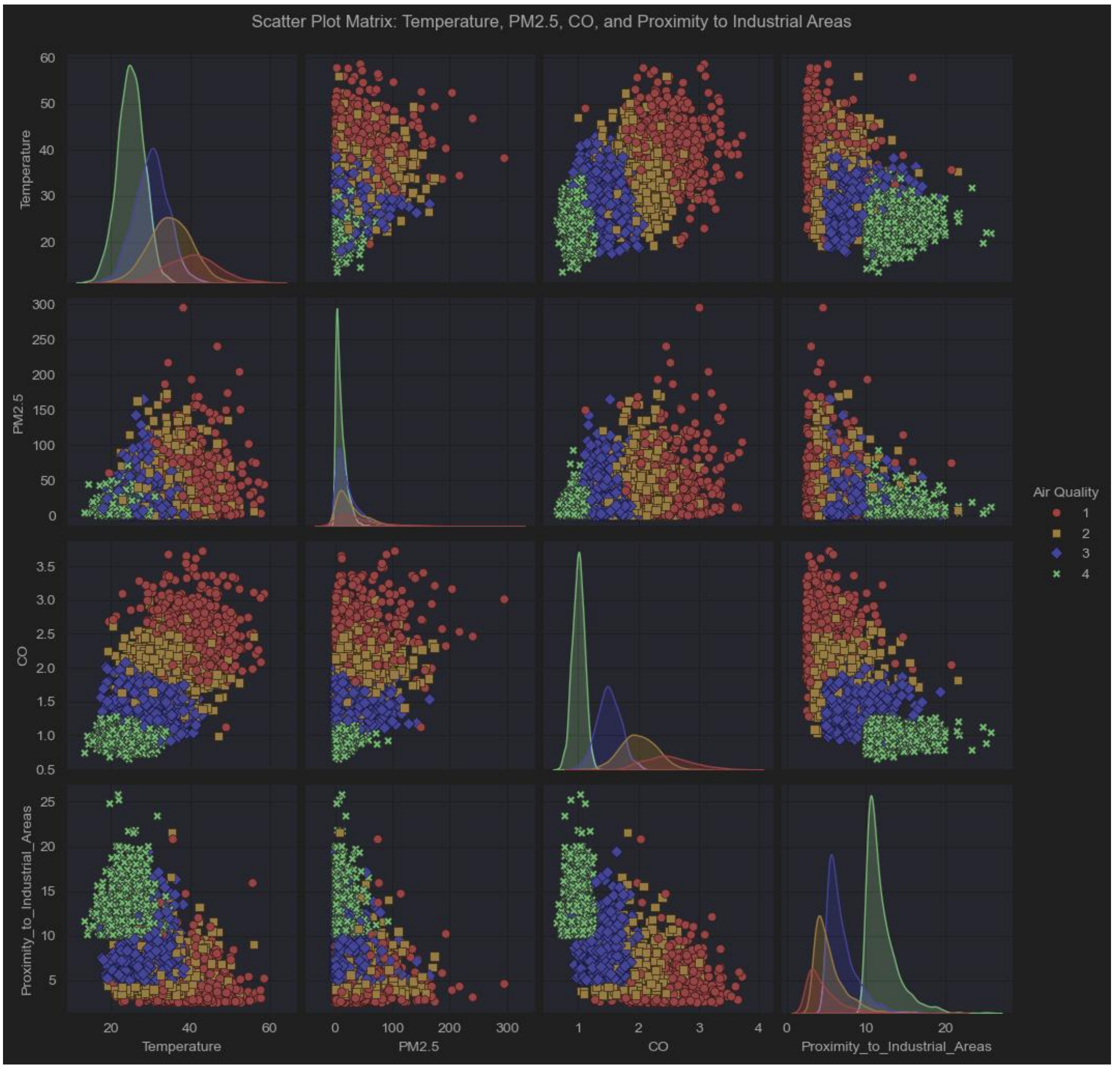
Pattern: Multi-modal

Skewed: Skewed right (Mean > Median)

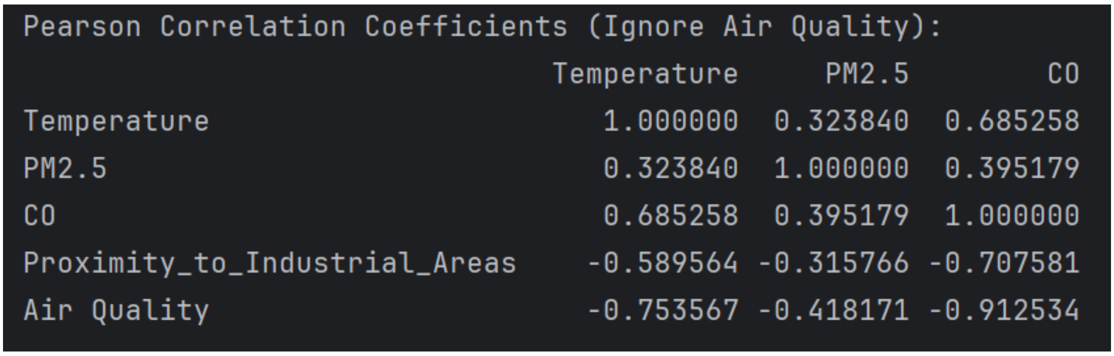
Outliers: Yes, value above 25

Gap: Yes, two gaps, 1) Value above 21 2) Value above 25

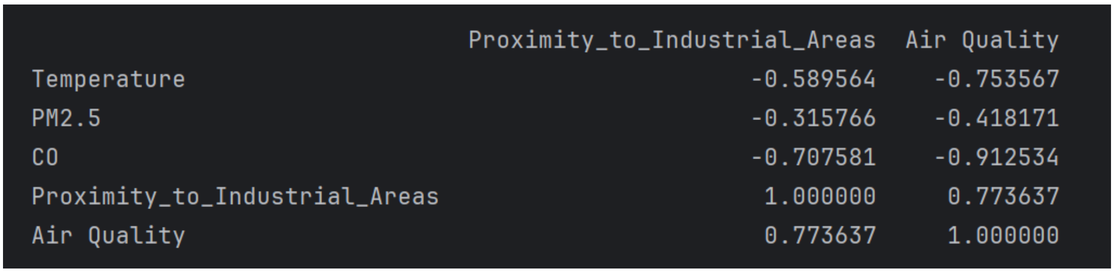
Shape: U-shape



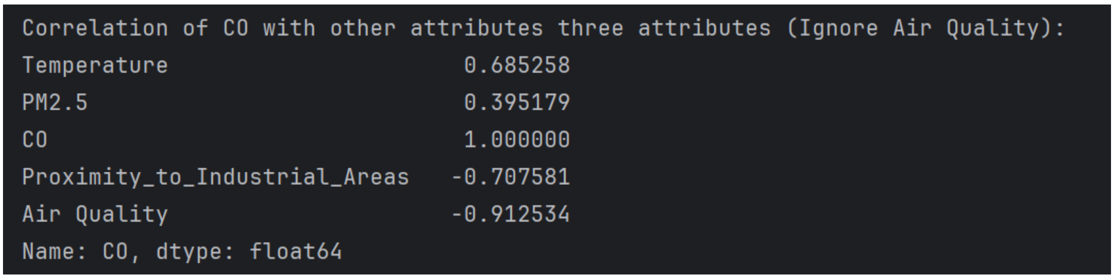
(Air quality attribute in the generated text is just to meet the requirement of Seaborn Pairplot function for color rendering the datapoints, also for further analysis in the conclusion part.)



```
Pearson Correlation Coefficients (Ignore Air Quality):
                               Temperature     PM2.5        CO
Temperature                       1.000000  0.323840  0.685258
PM2.5                             0.323840  1.000000  0.395179
CO                                0.685258  0.395179  1.000000
Proximity_to_Industrial_Areas    -0.589564 -0.315766 -0.707581
Air Quality                      -0.753567 -0.418171 -0.912534
```



```
                               Proximity_to_Industrial_Areas  Air Quality
Temperature                                        -0.589564    -0.753567
PM2.5                                              -0.315766    -0.418171
CO                                                 -0.707581    -0.912534
Proximity_to_Industrial_Areas                       1.000000     0.773637
Air Quality                                         0.773637     1.000000
```

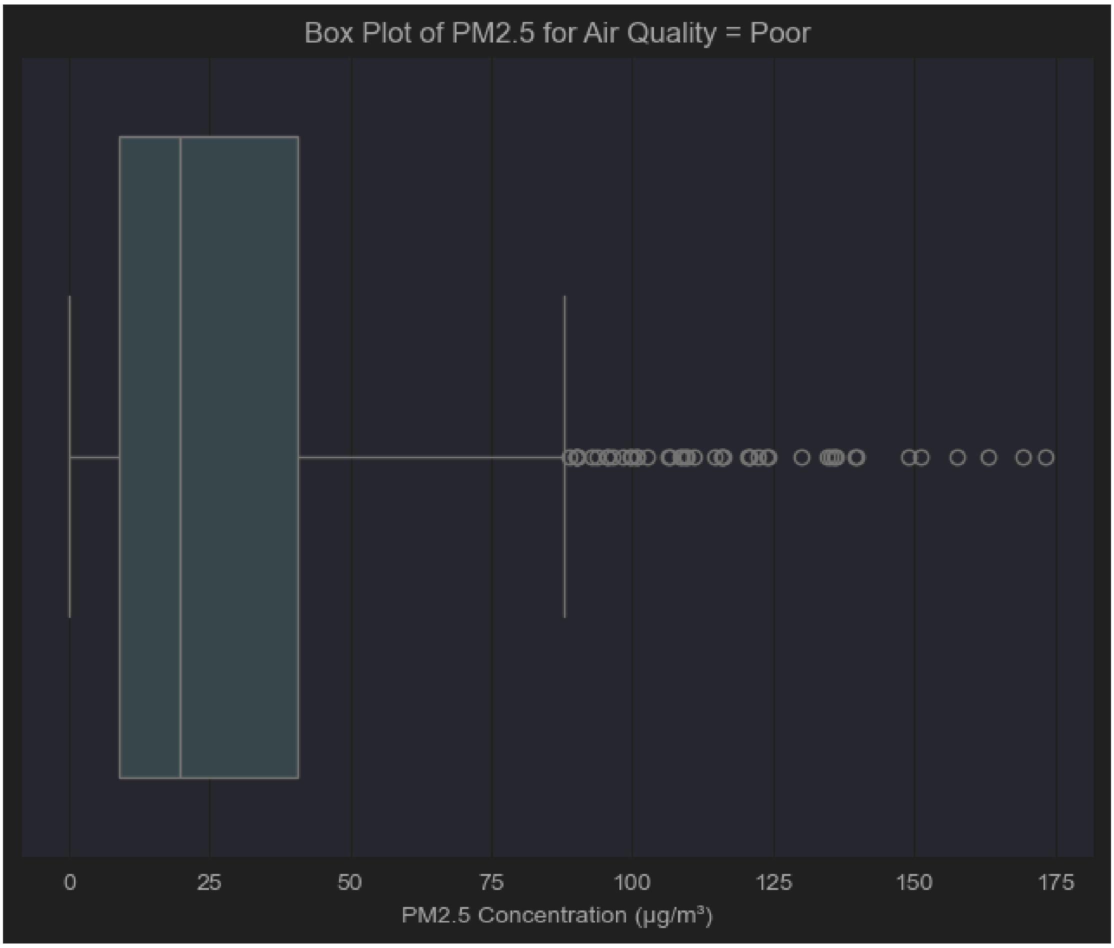


```
Correlation of CO with other attributes three attributes (Ignore Air Quality):
Temperature                      0.685258
PM2.5                            0.395179
CO                               1.000000
Proximity_to_Industrial_Areas   -0.707581
Air Quality                     -0.912534
Name: CO, dtype: float64
```

5. CO with other attributes:

- Temperature: Moderate strong, positive
- PM 2.5: Weak, negative
- Proximity to Industrial Areas: Strong, negative

.



6. Boxplot of PM 2.5 to Poor Air Quality

1) Q1: 8.9

Median (Q2): 19.85

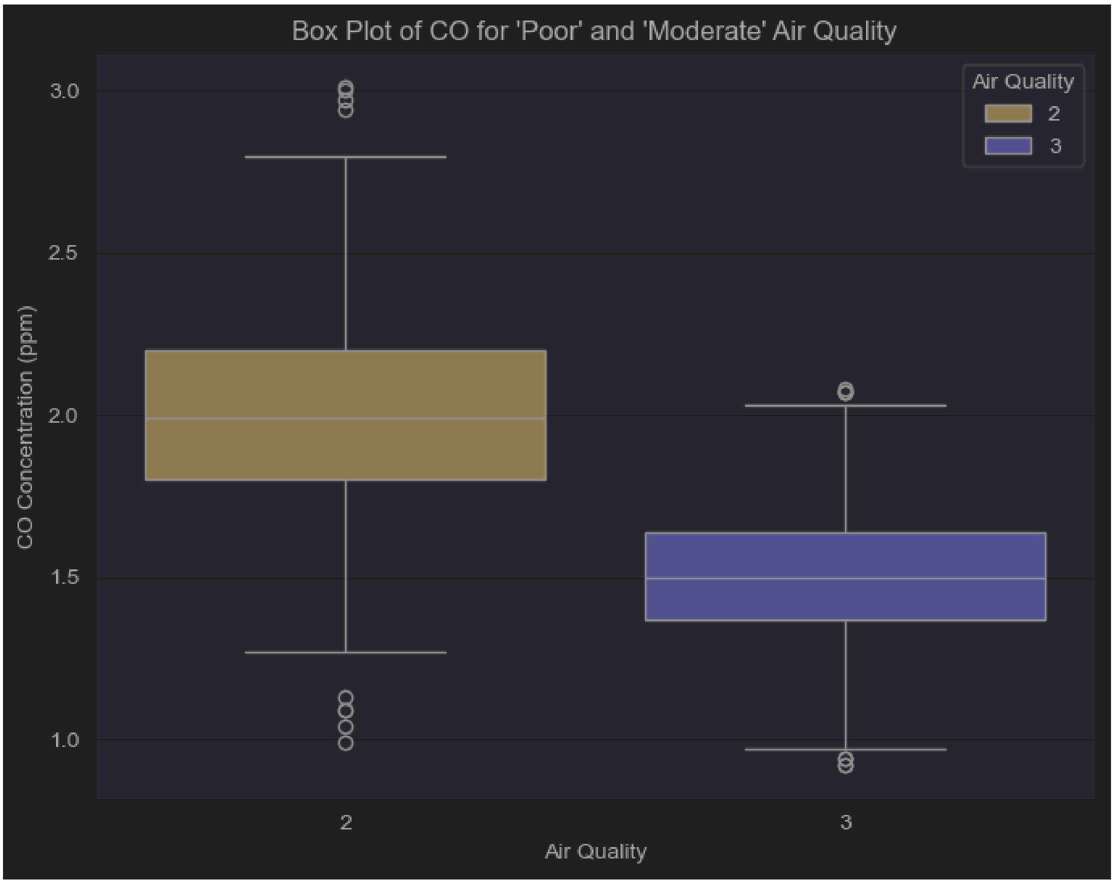
Q3: 40.5

IQR: 31.6

Min: 0.1

Max: 173.2

2) Number of outliers: 46
3) Skewed right, the median is closer to Min



7. Box plot of CO for ‘Poor’ and ‘Moderate’ Air Quality

Medians: Different (0.25 to 1 box size apart)

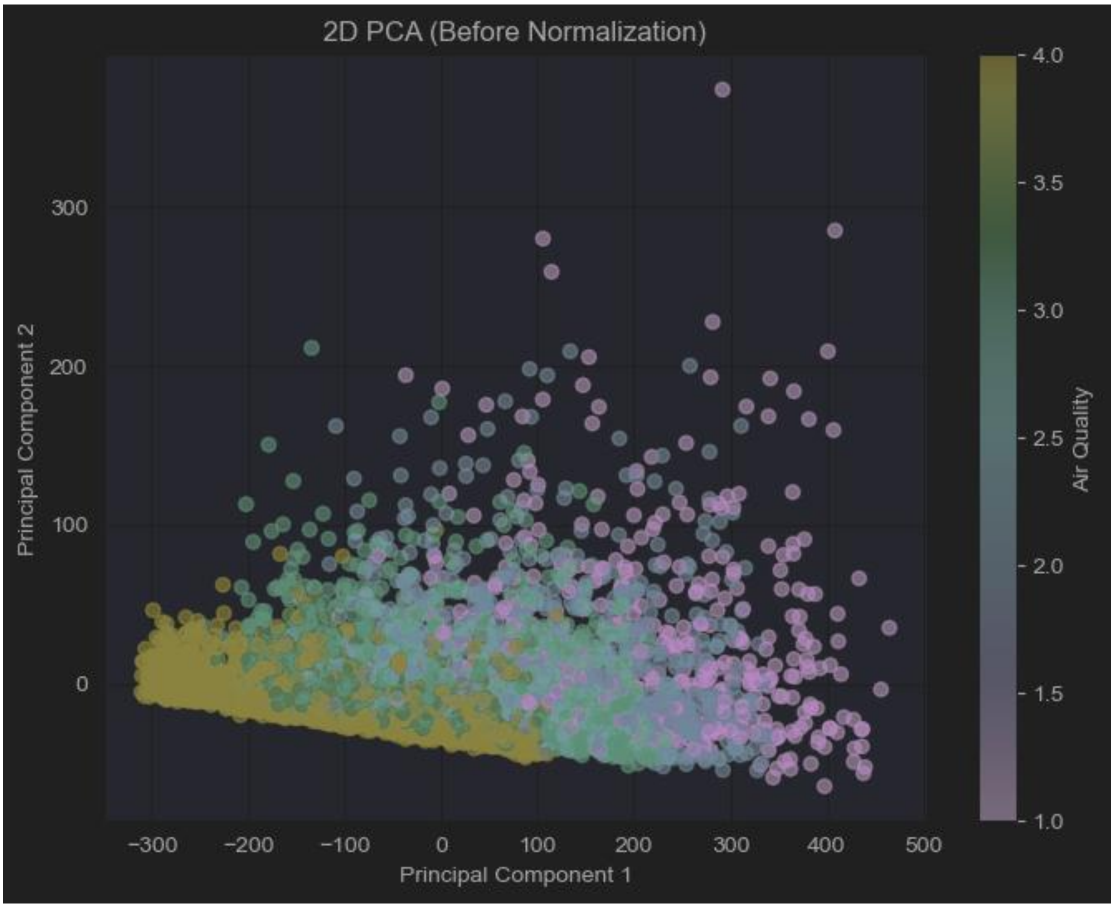
IQRs: very dissimilar (Intervals of IQR 10% or less overlap)

Intervals defined by the two whiskers: 25%-75% Medium overlap

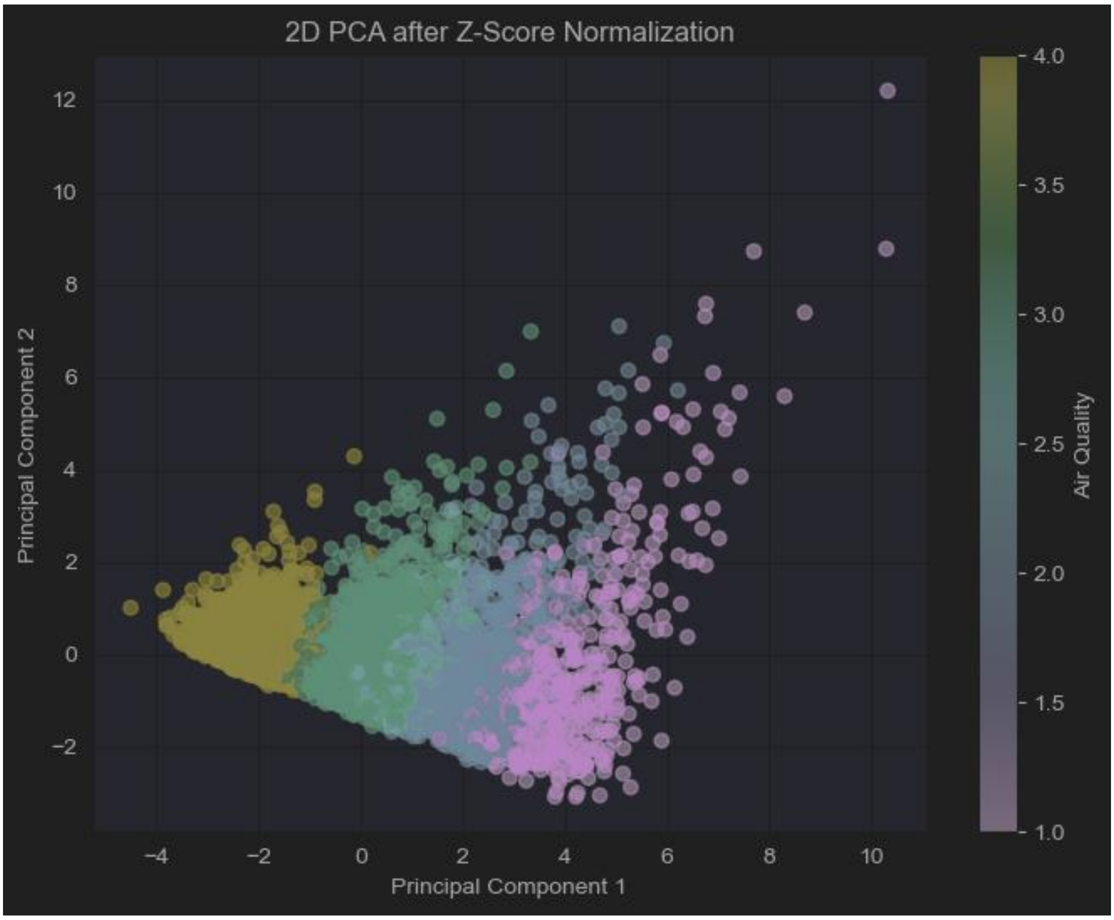
Skewed: Both of the air quality box plots are symmetric distribution.

Outliers: 'Poor' air quality (2) has outliers above 2.5 ppm. 'Moderate' air quality (3) has outliers below 1.0 ppm and above 2.0 ppm. There is no agreement in the specific outliers between the two box plots.

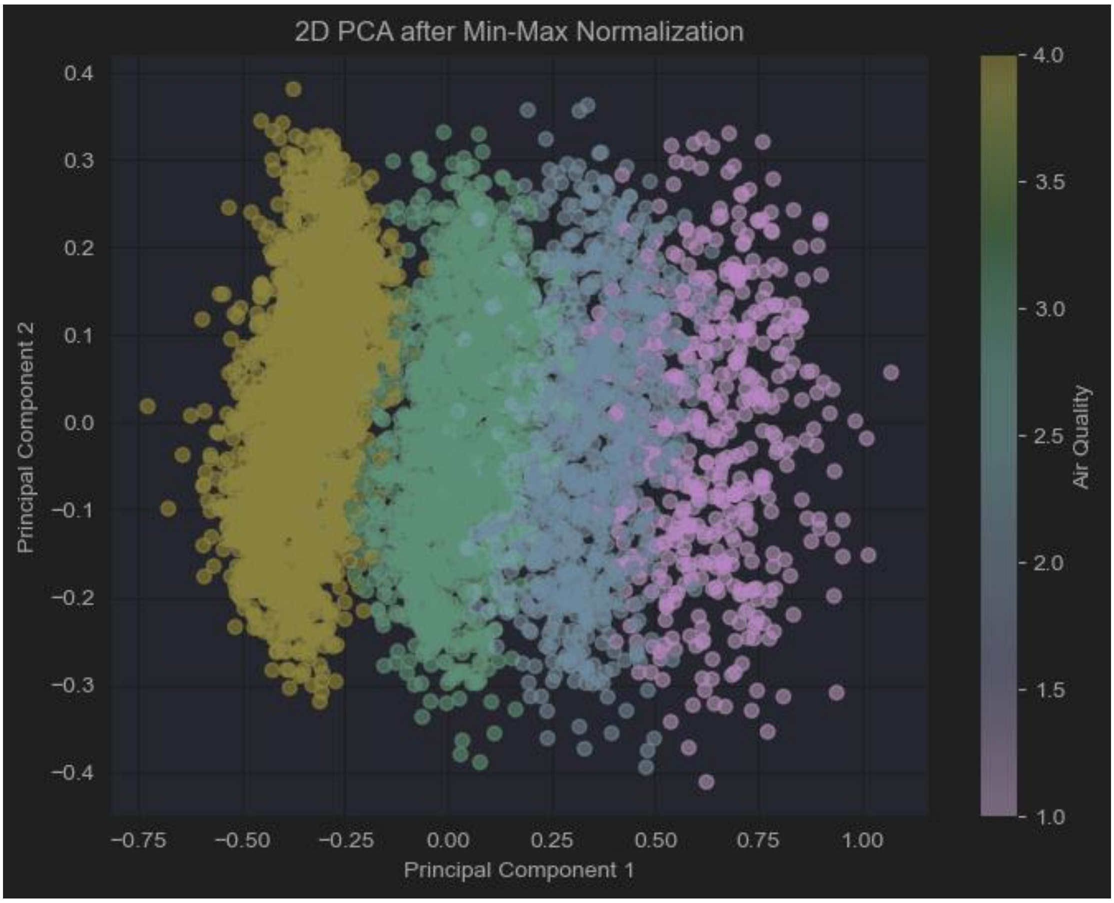
Summary: Overall, the two box plots show different distributions of CO concentration for 'Poor' and 'Moderate' air quality levels, with differences in medians, IQRs, and similar skewness, no agreement in outliers.



Explained Variance Ratio: [0.9386382 0.04787276]



```
Explained Variance Ratio after Z-Score Normalization: [0.55721851 0.14969909]
```



Explained Variance Ratio after Min-Max Normalization: [0.59660027 0.10332071]

8. Differences Before and After Normalization:

Before Normalization:

- The first principal component captures the vast majority of the variance (93.86%).
- There is a significant drop-off in the variance explained by the second component (4.79%).

After Z-Score Normalization:

The variance explained by the first principal component is reduced to 55.72%.

The second principal component explains a significantly larger portion of the variance (14.97%).

The total variance explained by the first two components is lower (70.69%).

After Min-Max Normalization:

The variance explained by the first principal component is slightly higher (59.66%) than that after Z-Score normalization.

The second principal component explains a bit less of the variance (10.33%) compared to Z-Score normalization.

The total variance explained by the first two components is similar to that of Z-Score normalization (69.99%).

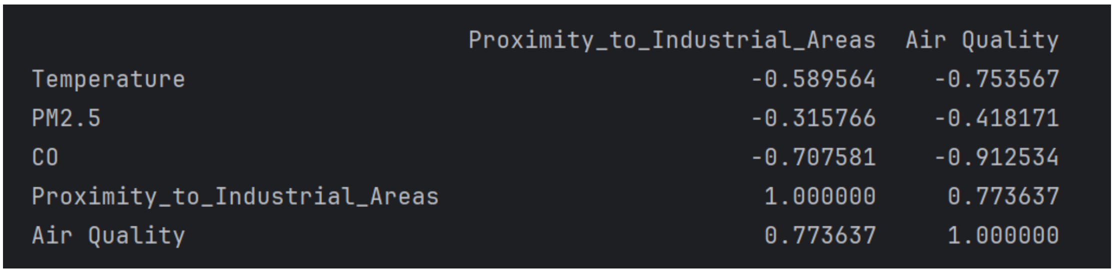
Reason of difference:

Normalization Impact: Normalizing the data affects the variance distribution among the principal components. Without normalization, attributes with larger scales can dominate the principal components. Normalization mitigates this issue by scaling the data, leading to a more balanced variance distribution among components.

Z-Score vs. Min-Max: Z-Score normalization centers the data around the mean with a standard deviation of 1, while Min-Max normalization scales the data to a specific range (e.g., 0 to 1). These different normalization methods influence the variance distribution in the PCA results.

Benefit and usefulness: Performing PCA on a dataset after Z-Score or Min-Max normalization achieves a more accurate capture of the main information, enhancing the quality and effectiveness of data analysis. Normalization ensures that PCA analysis is fair and balanced, allowing for a better understanding and utilization of the underlying patterns and structures in the data. This makes PCA particularly valuable and beneficial in the specific data analysis case.

Based on the dataset analysis, Temperature, Distance to industrial areas, CO Concentration are the most useful attributes for predicting air quality. PM2.5 and PM10 have strong correlations between but has weaker correlations to Air Quality, therefore is not the most useful attribute for prediction air quality.



```
                               Proximity_to_Industrial_Areas  Air Quality
Temperature                                        -0.589564    -0.753567
PM2.5                                              -0.315766    -0.418171
CO                                                 -0.707581    -0.912534
Proximity_to_Industrial_Areas                       1.000000     0.773637
Air Quality                                         0.773637     1.000000
```