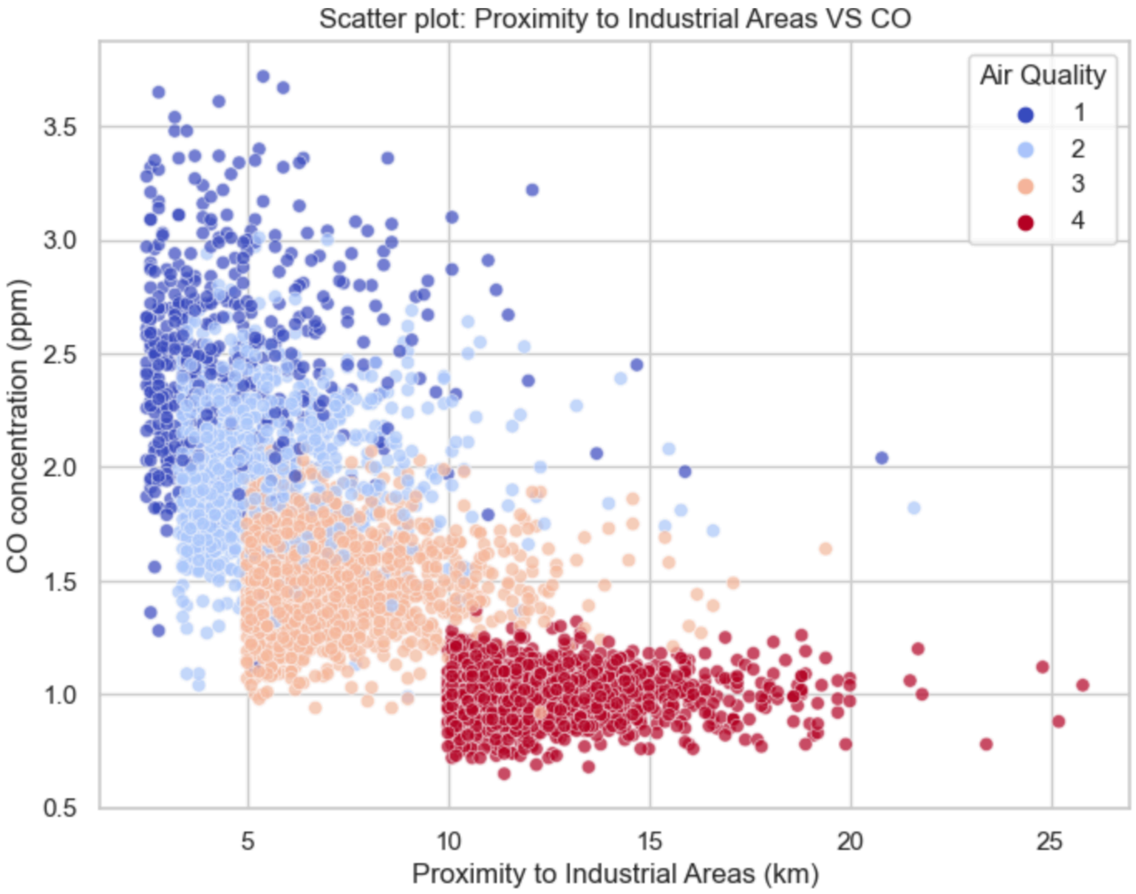
1.

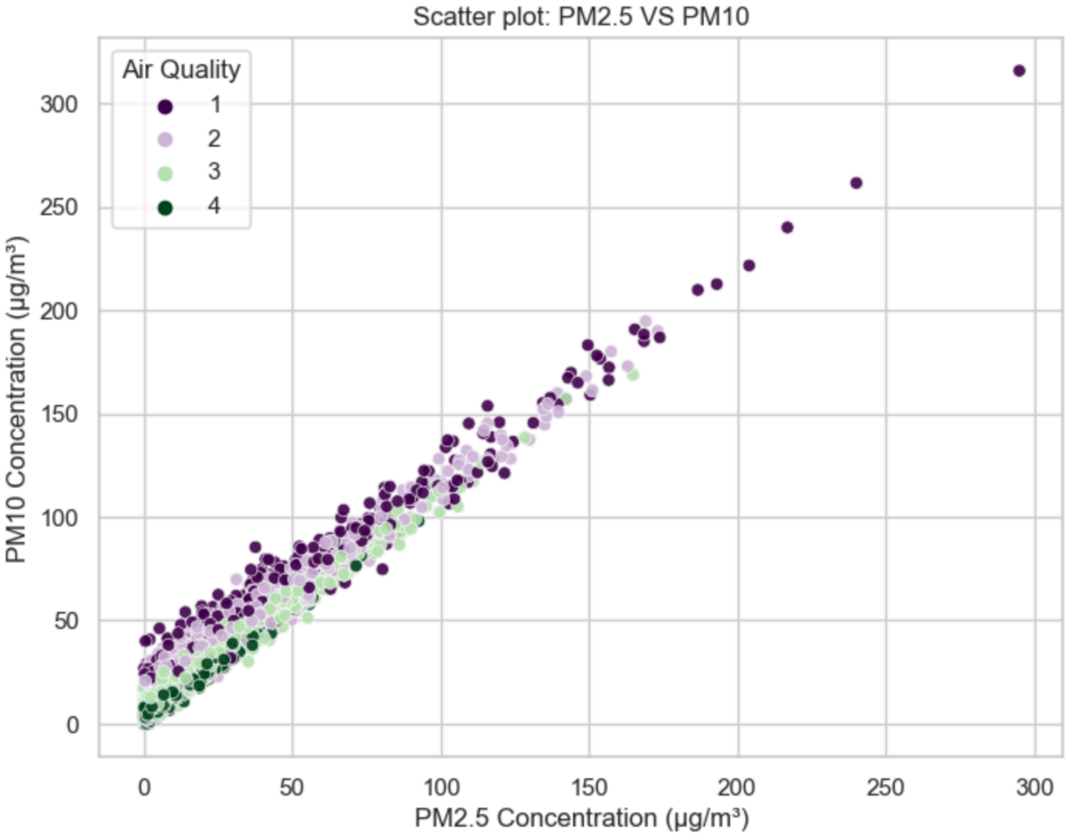


Interpretation:

- The scatter plot shows a negative correlation between Proximity to Industrial Areas and CO concentration.
- As the distance from industrial areas increases, the CO concentration decreases.
- Areas closer to industrial zones (<= 5 km) have **hazardous** and **poor** air quality.
- Areas within the proximity of 5 km to 10 km have **moderate** air quality.
- Areas farther from industrial zones (10+ km) have significantly lower CO concentrations and **good** air quality.

Industrial areas contribute significantly to CO pollution. The further a location is from an industrial area, the lower the CO concentration, improving air quality.This suggests that proximity to industrial zones is a key factor affecting air pollution.

2.

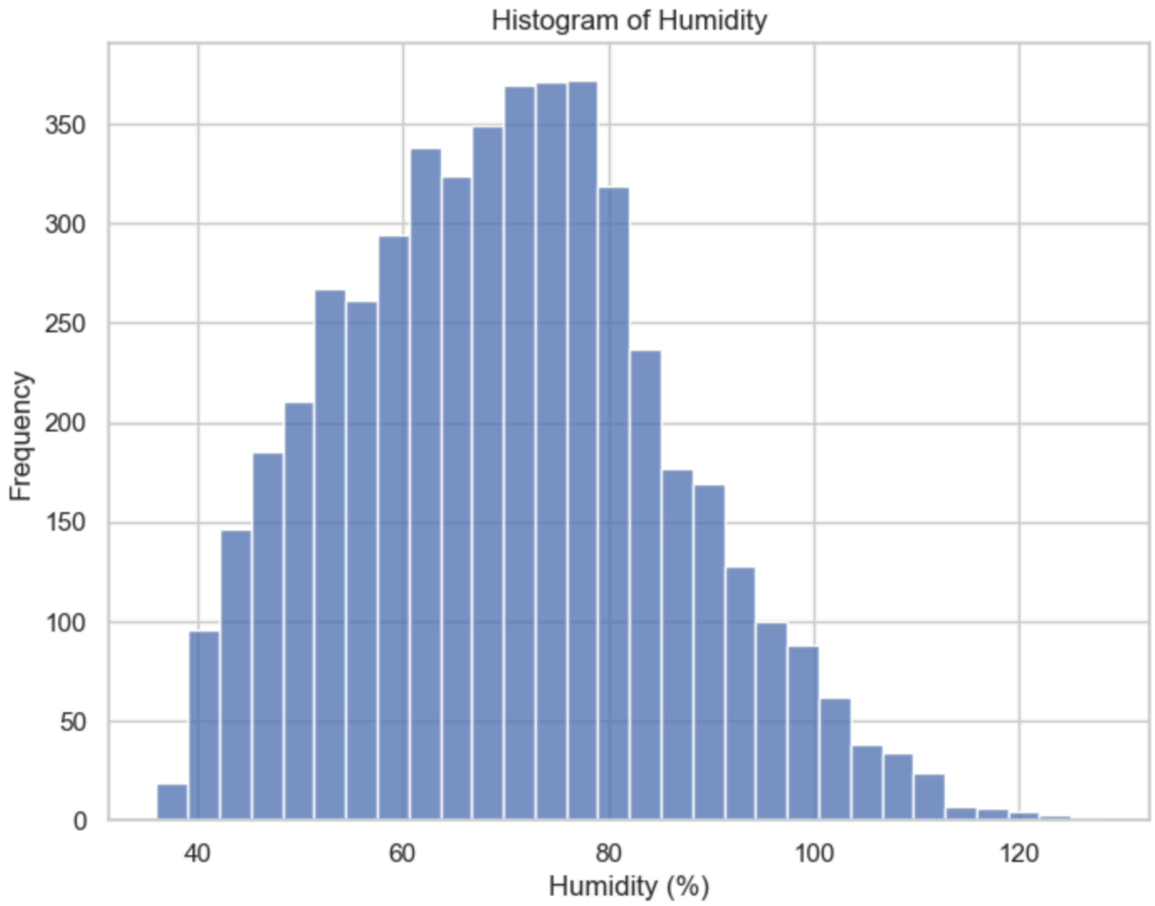


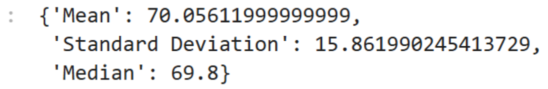
Interpretation

- The scatter plot shows a **strong positive correlation** between PM2.5 and PM10 concentrations.
- As PM2.5 concentration increases, PM10 concentration also increases almost linearly.
- Data points with higher PM2.5 and PM10 concentrations are more dark purple, indicating hazardous air quality.
- Lower concentrations of PM2.5 and PM10 are dark green, signifying good air quality.

PM2.5 and PM10 are closely related pollutants. Higher PM2.5 and PM10 levels strongly correlate with poorer air quality.

3.



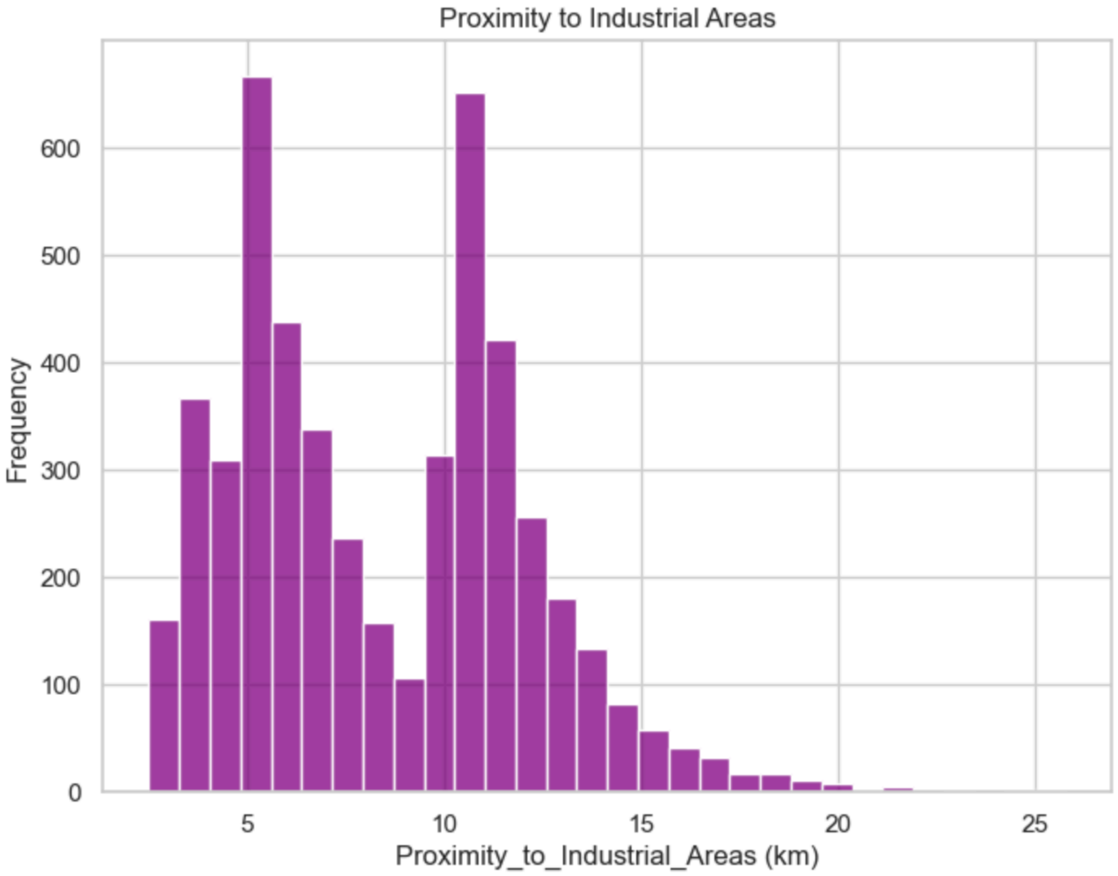


```
:  {'Mean': 70.05611999999999,
    'Standard Deviation': 15.861990245413729,
    'Median': 69.8}
```

Interpretation:

- Moderate spread with a standard deviation of 15.86, indicating that humidity values are somewhat dispersed around the mean.
- Unimodal (one peak around 75%), meaning that most observations are concentrated around a single central value.
- Slight right skew (mean = 70.06, median = 69.8), meaning a few higher humidity values extend the tail of the distribution.
- Outliers: Values above 120% can be considered outliers, as they are far from the main distribution.
- Bell Curve (approximately normal but slightly skewed right).

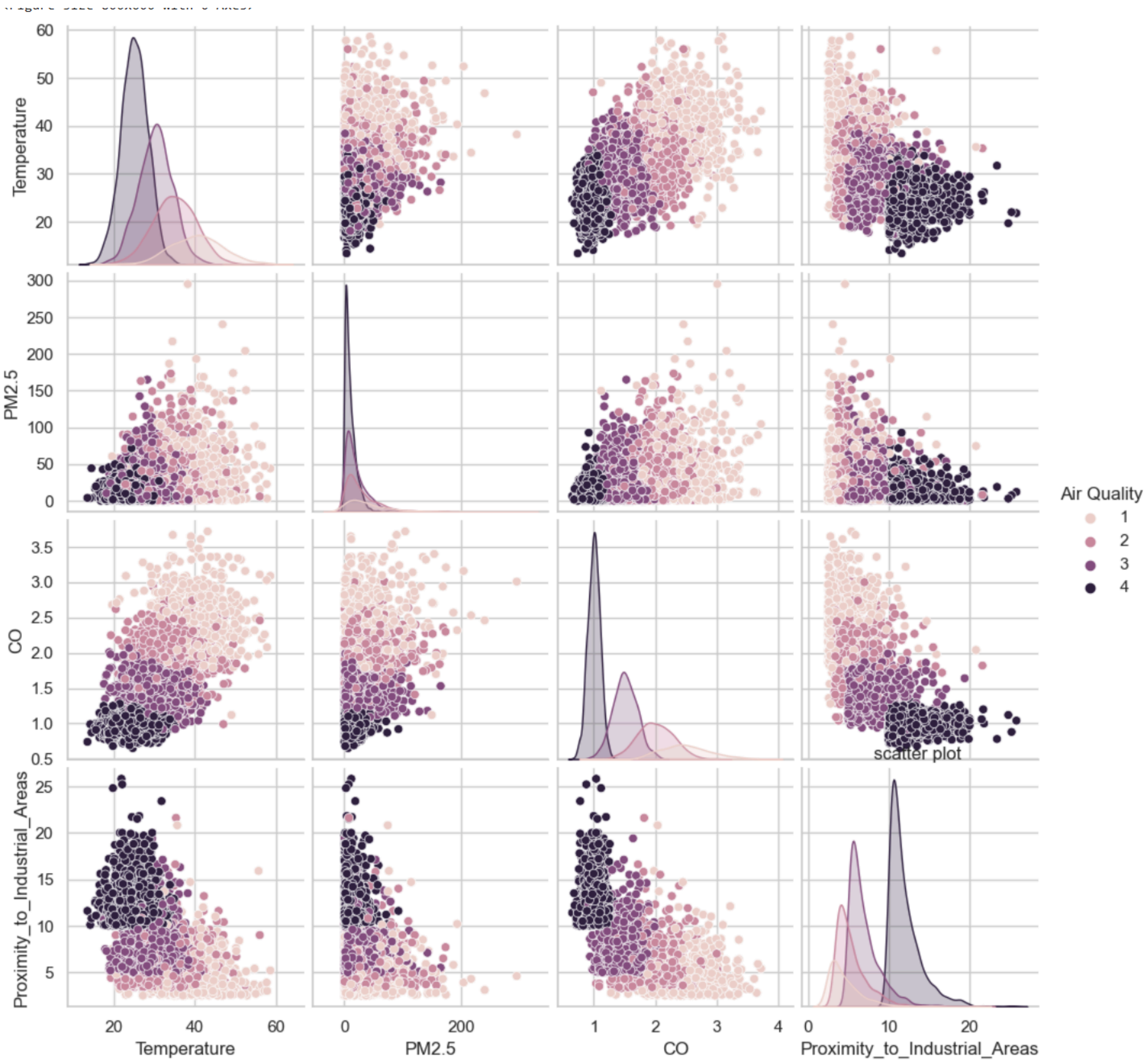
4.



```
{'Mean': 8.4254, 'Standard Deviation': 3.6105826178055773, 'Median': 7.9}
```

- Moderate spread, as indicated by the standard deviation.
- The histogram appears bimodal (two peaks), suggesting two major groups in the dataset. There are two peaks around 5 km and 10 km, indicating two common distances to industrial areas.
- Right-skewed (positively skewed) – The tail extends towards higher distances, meaning most observations are near industrial areas, but some are farther away.
- Outliers: Values beyond ~20 km can be considered outliers.
- The frequency drops at certain distance ranges, suggesting gaps between clusters of data points.there are two dominant groupings of industrial proximity distances rather than a smooth bell curve.

5.



1. Correlation in the Scatter Plot Matrix

- **CO vs Temperature**: There is a strong positive correlation (0.685), indicating that as temperature increases, CO levels also tend to rise.
- **CO vs PM2.5**: There is a moderate positive correlation (0.395), meaning that higher PM2.5 levels are often associated with increased CO levels.
- **CO vs Proximity to Industrial Areas**: There is a strong negative correlation (-0.708), which implies that locations closer to industrial areas have higher CO concentrations.

2. Separation of Air Quality Classes

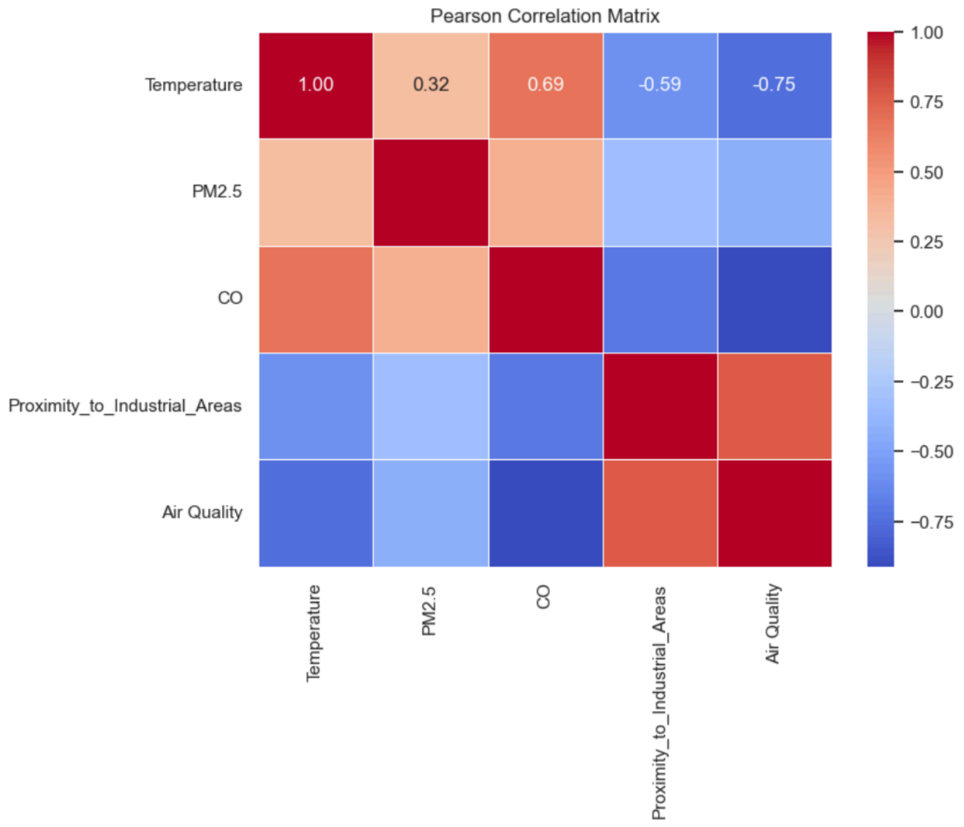
- The scatter plot matrix shows clear separation of air quality classes:
  - High CO and PM2.5 levels are strongly linked to poor air quality.
  - Lower CO levels tend to be associated with better air quality classes (darker colors in the plot).

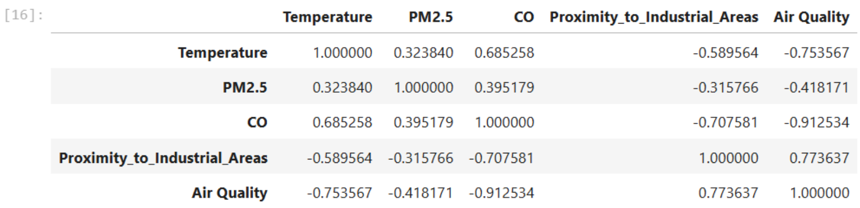
3. Decision Boundaries for Classification

- **CO & Proximity to Industrial Areas provide a strong separation for air quality classification.**
  - The decision boundary can be drawn where CO is above 1.5 ppm and proximity is below 10 km for poor air quality.
  - Temperature and PM2.5 alone may not provide as strong a classification boundary.

4. Difficulty of Classification

- **Easy classification:**
  - Since CO has a high negative correlation with proximity to industrial areas (-0.708) and a high positive correlation with temperature (0.685), we can infer clear trends for classifying air quality.
  - The Pearson correlation matrix confirms these relationships.




[16]:

	Temperature	PM2.5	CO	Proximity_to_Industrial_Areas	Air Quality
Temperature	1.000000	0.323840	0.685258	-0.589564	-0.753567
PM2.5	0.323840	1.000000	0.395179	-0.315766	-0.418171
CO	0.685258	0.395179	1.000000	-0.707581	-0.912534
Proximity_to_Industrial_Areas	-0.589564	-0.315766	-0.707581	1.000000	0.773637
Air Quality	-0.753567	-0.418171	-0.912534	0.773637	1.000000

1. CO vs Temperature (0.685) – Strong Positive Correlation

- Higher temperatures are associated with **increased CO levels.**

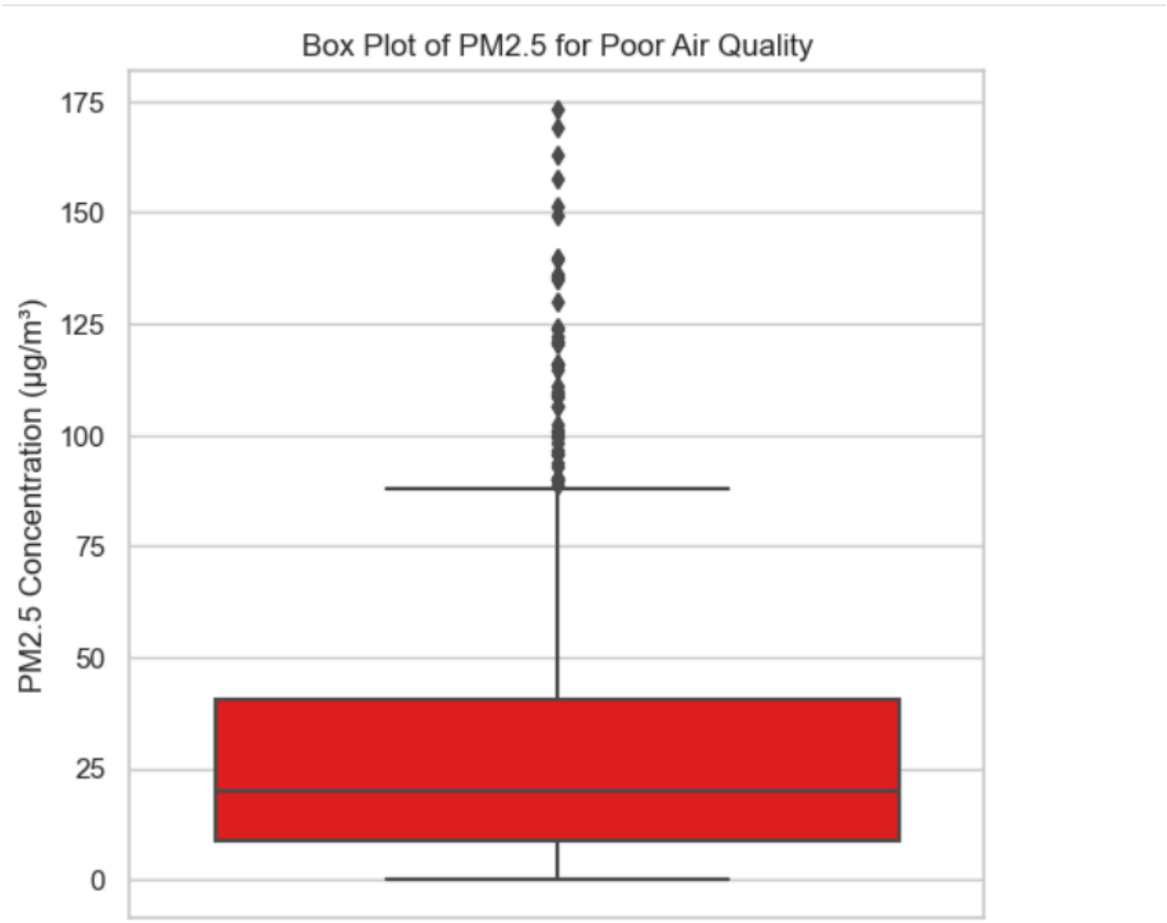
2. CO vs PM2.5 (0.395) – Moderate Positive Correlation

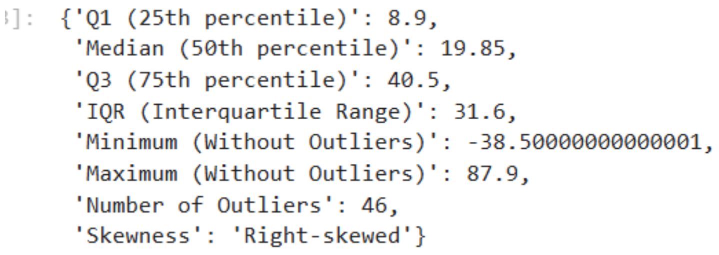
- CO and PM2.5 tend to increase together..

3. CO vs Proximity to Industrial Areas (-0.708) – Strong Negative Correlation

- Closer proximity to industrial areas results in higher CO levels.
- As distance from industrial areas increases, CO levels drop significantly.

6.

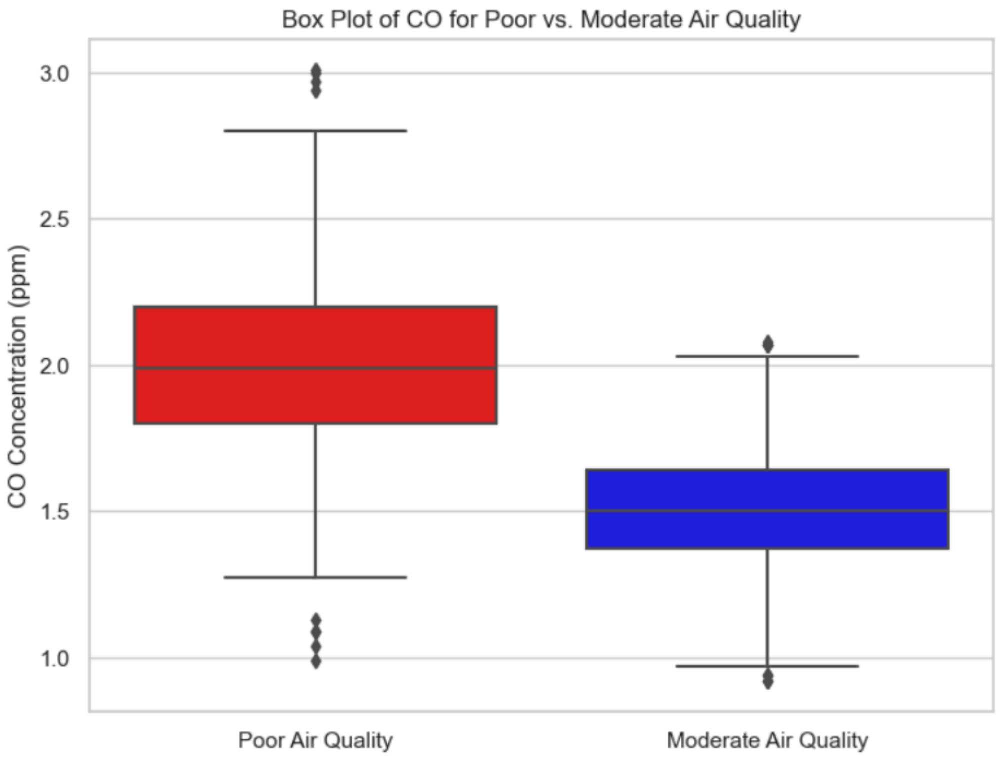


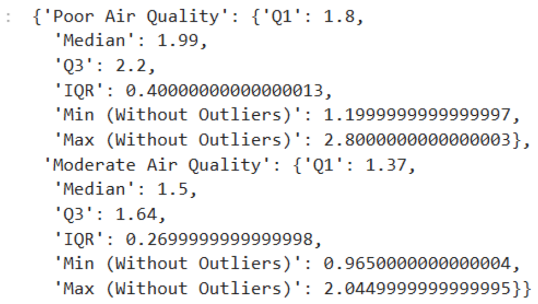


```
]:  {'Q1 (25th percentile)': 8.9,
     'Median (50th percentile)': 19.85,
     'Q3 (75th percentile)': 40.5,
     'IQR (Interquartile Range)': 31.6,
     'Minimum (Without Outliers)': -38.50000000000001,
     'Maximum (Without Outliers)': 87.9,
     'Number of Outliers': 46,
     'Skewness': 'Right-skewed'}
```

- Q1 (25th percentile): 8.9
- Median (50th percentile): 19.85
- Q3 (75th percentile): 40.5
- IQR (Interquartile Range): 31.6
- Minimum (Without Outliers): -38.5
- Maximum (Without Outliers): 87.9
- Number of outliers: 46
- Outliers are located above the upper whisker, indicating extreme high PM2.5 values.
- **Skewness of the Box Plot:**The distribution is right-skewed (positively skewed)**.**
    - The median is closer to Q1 (lower quartile) rather than centered.
    - The upper whisker is longer than the lower whisker, indicating a spread of higher PM2.5 values.
    - There are multiple outliers on the higher end (above Q3 + 1.5 * IQR), suggesting extreme pollution events.

7.





```
: {'Poor Air Quality': {'Q1': 1.8,
    'Median': 1.99,
    'Q3': 2.2,
    'IQR': 0.40000000000000013,
    'Min (Without Outliers)': 1.1999999999999997,
    'Max (Without Outliers)': 2.8000000000000003},
   'Moderate Air Quality': {'Q1': 1.37,
    'Median': 1.5,
    'Q3': 1.64,
    'IQR': 0.2699999999999998,
    'Min (Without Outliers)': 0.9650000000000004,
    'Max (Without Outliers)': 2.0449999999999995}}
```

1. Medians:

- **Poor Air Quality Median:** 1.99
- **Moderate Air Quality Median:** 1.50
- **Interpretation:** The medians are somewhat similar but still show a notable difference. Since they are about 0.5 ppm apart, this suggests that CO concentration is consistently higher for Poor Air Quality compared to Moderate Air Quality.

2. IQRs (Interquartile Ranges):

- **Poor Air Quality IQR:** 0.40
- **Moderate Air Quality IQR:** 0.27
- **Interpretation:** The IQRs are somewhat overlapping, meaning the spread of CO values within the middle 50% of both classes shares some common values, but Poor Air Quality generally has a higher range.

3. Whiskers (Overall Range of Values):

- **Poor Air Quality Range (Without Outliers):** 1.20 to 2.80 ppm
- **Moderate Air Quality Range (Without Outliers):** 0.97 to 2.04 ppm
- **Interpretation:** The whiskers indicate that Poor Air Quality has a much wider range of CO values, extending to higher values compared to Moderate Air Quality.

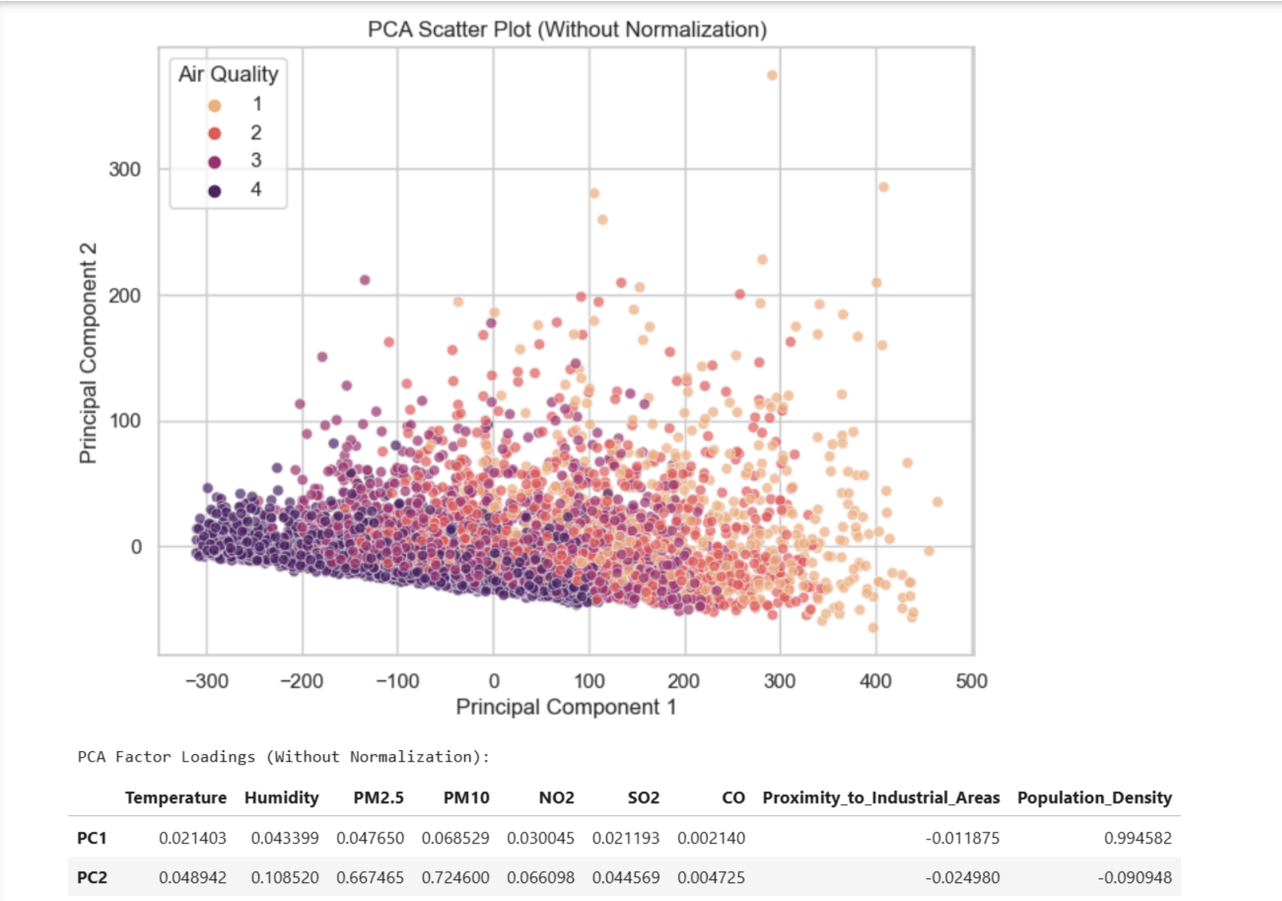
4. Skewness:

- Both box plots show slight right skewness, meaning there are a few instances where CO values are significantly higher than the median. However, Poor Air Quality has more extreme high values, which is evident from the number of outliers.

5. Outliers Comparison:

- **Poor Air Quality:** More outliers (values exceeding 2.80 ppm).
- **Moderate Air Quality:** Fewer outliers, with a maximum value around 2.04 ppm.
- **Interpretation:** This indicates that Poor Air Quality has extreme pollution events where CO spikes significantly, while Moderate Air Quality is more stable with fewer extreme cases.

8.

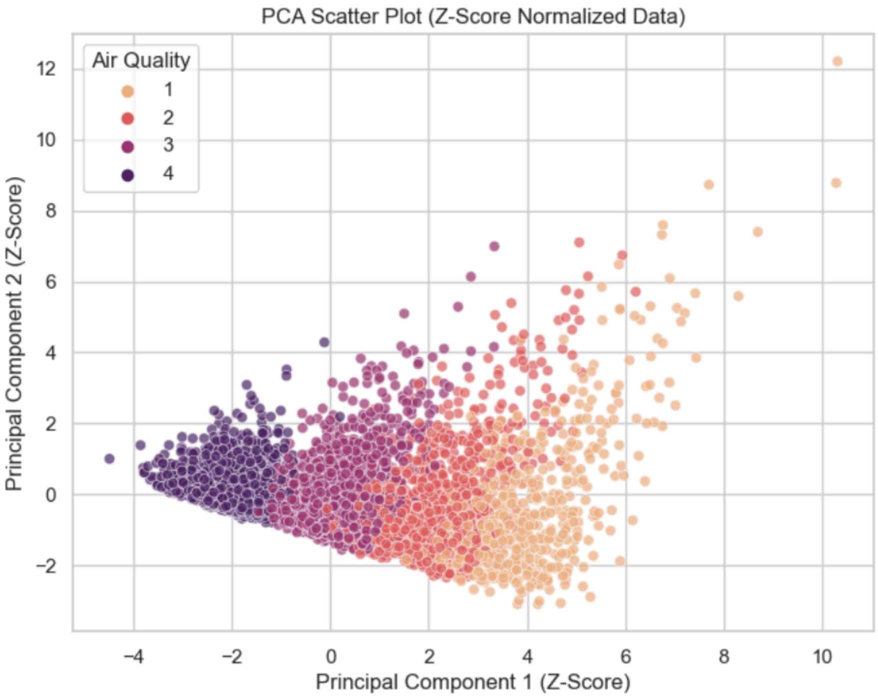


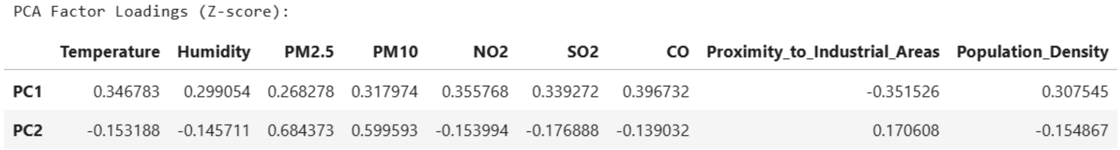
```
PCA Factor Loadings (Without Normalization):
```

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density
PC1	0.021403	0.043399	0.047650	0.068529	0.030045	0.021193	0.002140	-0.011875	0.994582
PC2	0.048942	0.108520	0.667465	0.724600	0.066098	0.044569	0.004725	-0.024980	-0.090948

PCA scatter plot (Without Normalization):

- The scatter plot shows wide-spread data, and the variance among attributes is dominated by their original scales.
- Population Density has the highest loading on PC1 (0.994582), meaning it contributes most to the first principal component.
- PM10 and PM2.5 dominate PC2 (0.724600 and 0.667465, respectively), indicating that fine particulate matter is a key differentiator in the dataset.



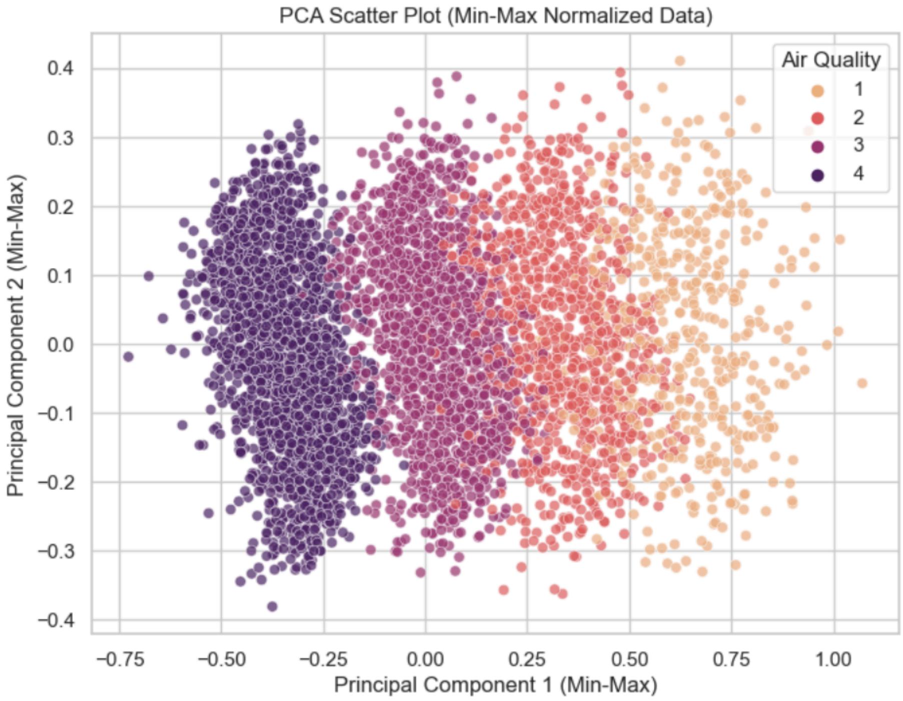


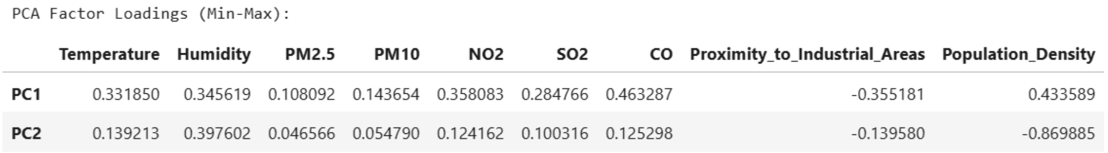
PCA Factor Loadings (Z-score):

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density
PC1	0.346783	0.299054	0.268278	0.317974	0.355768	0.339272	0.396732	-0.351526	0.307545
PC2	-0.153188	-0.145711	0.684373	0.599593	-0.153994	-0.176888	-0.139032	0.170608	-0.154867

PCA scatter plot(Z-score Normalized Data):

- The spread of points in the scatter plot appears more balanced.
- PC1 is strongly influenced by CO (0.396732), SO2 (0.396732), NO2 (0.355768), and Temperature (0.346783). This suggests that gaseous pollutants and temperature are major factors in air quality.
- PC2 is still highly dominated by PM2.5 (0.684373) and PM10 (0.599593), reinforcing their importance in air quality analysis.
- Population Density is less influential (-0.154867 in PC2), suggesting that its dominance in the unnormalized PCA was due to scale differences rather than true importance.





```
PCA Factor Loadings (Min-Max):
```

	Temperature	Humidity	PM2.5	PM10	NO2	SO2	CO	Proximity_to_Industrial_Areas	Population_Density
PC1	0.331850	0.345619	0.108092	0.143654	0.358083	0.284766	0.463287	-0.355181	0.433589
PC2	0.139213	0.397602	0.046566	0.054790	0.124162	0.100316	0.125298	-0.139580	-0.869885

PCA scatter plot(Max-Min Normalized Data):

- The scatter plot now shows a more evenly distributed pattern, suggesting better separation between different air quality classes.
- PC1 is influenced by SO2 (0.463287), NO2 (0.358083), and PM10 (0.143654), while PC2 is largely driven by Humidity (0.397602) and Temperature (0.139213).
- Population Density (-0.869885 in PC2) has an inverse impact, meaning areas with high density correlate with lower principal component scores.

Key Differences Before and After Normalization

- **Without normalization:** Population Density was the dominant factor, likely due to its large numerical scale.
- **With Z-Score normalization**: CO, NO2, and PM pollutants emerge as the key contributors, revealing true influential factors.
- **With Min-Max normalization:** The influence of temperature and humidity increases, making pollutant-based factors clearer.

Benefits of PCA:

- PCA compresses 9 features into 2 principal components, retaining the most important variance.
- Identifies which attributes contribute most to variations in air quality.
- Helps in better clustering of air quality levels based on key contributing factors.
- Reduce redundancy. Strongly correlated attributes like PM2.5 and PM10 are effectively combined into fewer dimensions.

Conclusion: Key Attributes for Predicting Air Quality

Based on the PCA analysis, correlation matrix, and visualizations, the most important attributes for predicting air quality are:

1. **PM2.5 & PM10** : These pollutants have a strong contribution to principal components (PC1 & PC2) and are highly correlated with poor air quality. Their increased concentrations are a clear indicator of air pollution.
2. **CO (Carbon Monoxide)** : CO has strong correlations with air quality and is consistently important in PCA factor loadings. Higher CO levels are often found in poor air quality zones.
3. **NO2 (Nitrogen Dioxide) & SO2 (Sulfur Dioxide)** : These gaseous pollutants are significant in PC1, meaning they play a major role in distinguishing air quality levels..
4. **Proximity to Industrial Areas** : This factor is negatively correlated with air quality, meaning closer proximity to industrial zones leads to worse air quality due to emissions from factories.
5. **Temperature & Humidity** : While not the primary pollutants, temperature affects pollutant dispersion, and humidity can impact particle accumulation. They contribute moderately to air quality variations.