Air Quality Index Prediction using Linear Regression

Paper replication

Summary

01 - Context and problematic

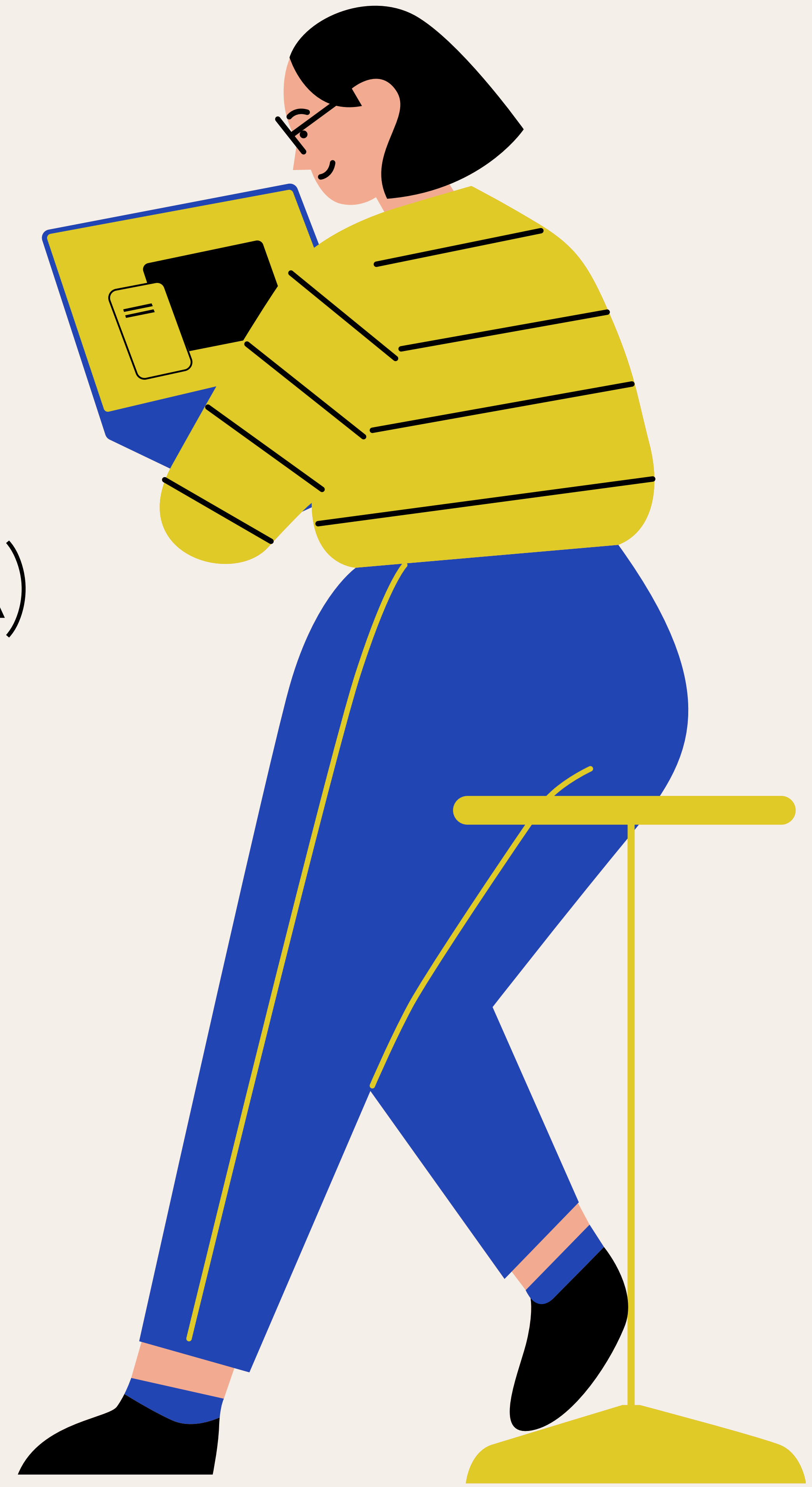
02 - Objectives

03 - Exploratory Data Analysis (EDA)

04 - Linear Regression Models & Evaluation

05 - Tableau

06 - Conclusion



01 - Context

Air pollution threatens public health, the environment, and the economy, causing diseases, ecosystem damage, and economic losses. Monitoring and predicting **AQI** are vital for identifying risks and guiding effective mitigation. Data analytics and predictive models enable proactive strategies to address this global challenge.

The AQI is a numerical scale used to communicate how polluted the air is.

01 - problematic

- How can we use pollutant and vehicle data to accurately predict AQI?
- What are the key contributing factors to AQI variations?

02 - Objectives

Primary Objectives:

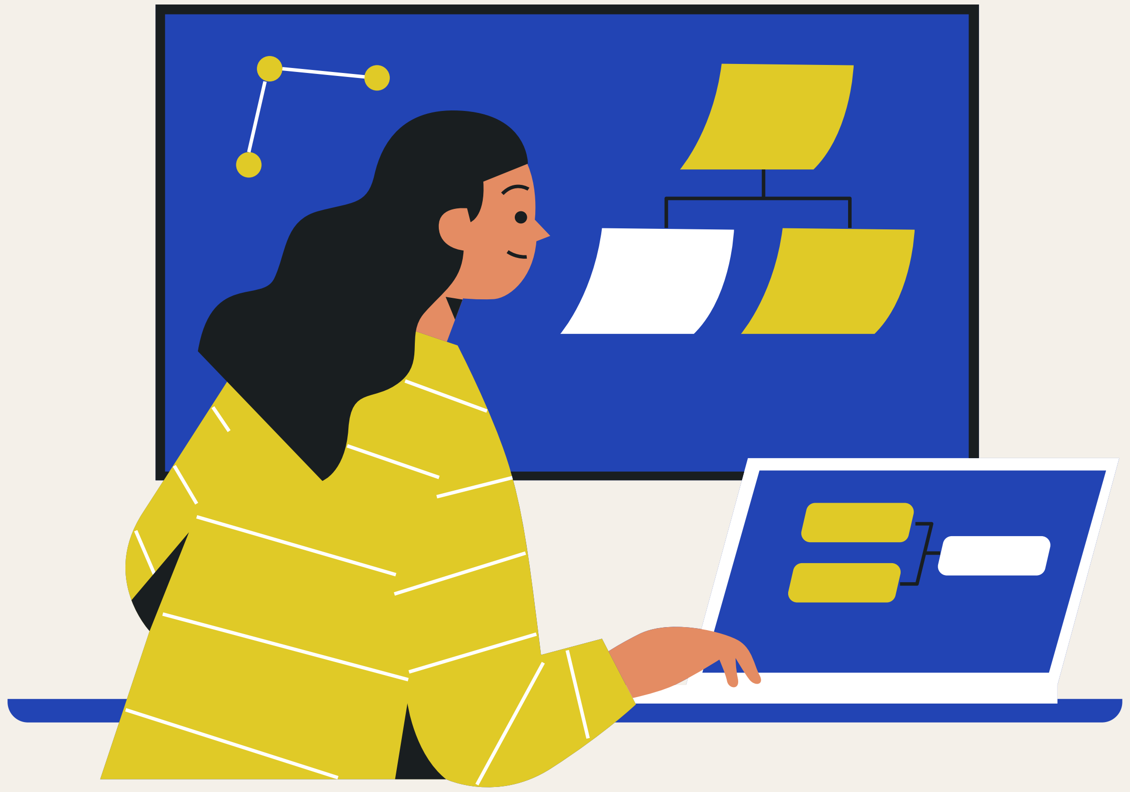
- Analyze pollutant and vehicle data to identify significant trends.
- Build regression models (simple and multiple) to predict AQI.
- Evaluate model performance using standard metrics.

Secondary Objectives:

- Visualize AQI variations across regions.
- Provide actionable insights for reducing pollution levels.

03 - Exploratory Data Analysis (EDA)

Data Overview



Air Pollution Dataset

Description:

- Contains air quality data collected from various countries and cities, including pollutant-specific AQI values and overall AQI categories.
- Number of Observations: 23,463 rows.
- Number of Features: 12 columns.

Key Features:

- **Country:** Name of the country where the data was collected.
- **City:** City within the country.
- **AQI.Value:** The Air Quality Index value, representing overall air quality.
- **AQI.Category:** Describes the AQI (e.g., "Good", "Moderate", "Unhealthy").
- **Pollutant-Specific Values:** Includes AQI values for CO, Ozone, NO2, and PM2.5, along with their corresponding AQI categories.

03 - Exploratory Data Analysis (EDA)

Data Overview

Registered Vehicles Dataset

Description:

- Provides information on vehicle registration density per 1,000 people for various countries.
- Number of Observations: 161 rows.
- Number of Features: 4 columns.

Key Features:

- **Entity:** The country of the dataset.
- **Year:** The year the data was recorded.
- **Registered.vehicles.per.1.000.people:** The number of vehicles registered per 1,000 people in that country.
- **Code:** The country code for identification.

Registered Vehicles Per 1000 people Dataset

	Entity	Code	Year	Registered vehicles per 1,000 people
1	Afghanistan	AFG	2013	20.724253
2	Albania	ALB	2016	194.31734
3	Antigua and...	ATG	2016	400.41788
4	Argentina	ARG	2016	492.7889
5	Australia	AUS	2016	[illegible]
6	Austria	AUT	2016	
7	Azerbaijan	AZE	2016	
8	Bangladesh	BGD	2016	
9	Barbados	BRB	2015	

```
> print("Vehicles Data Overview")
[1] "Vehicles Data Overview"
> print(summary(vehicles_data))
    Entity               Code                Year      Registered.vehicles.per.1.000.people
 Length:161         Length:161         Min.   :2007   Min.   :   4.457
 Class :character   Class :character   1st Qu.:2016   1st Qu.:  84.071
 Mode  :character   Mode  :character   Median :2016   Median : 258.296
                                       Mean   :2016   Mean   : 319.263
                                       3rd Qu.:2016   3rd Qu.: 500.625
                                       Max.   :2017   Max.   :1607.512
```

Global Air Pollution Dataset

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
1	Russian Fed...	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
2	Brazil	Presidente D...	41	Good	1	Good	5	Good	1	Good	41	Good
3	Italy	Priolo Gargal...	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
4	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
5	France	Punaauia	22	Good								
6	United State...	Punta Gorda	54	Moderate								
7	Germany	Puttlingen	62	Moderate								
8	Belgium	Puurs	64	Moderate								
9	Russian Fed...	Pyatigorsk	54	Moderate								
10	Egypt	Qalyub	142	Unhealthy fo...								
11	China	Qinzhou	68	Moderate								
12	Netherlands	Raalte	41	Good								

```
> print("Pollution Data Overview")
[1] "Pollution Data Overview"
> print(summary(pollution_data))
   Country              City             AQI.Value      AQI.Category        CO.AQI.Value     CO.AQI.Category    Ozone.AQI.Value  Ozone.AQI.Category
 Length:23463       Length:23463       Min.   :  6.00   Length:23463       Min.   :  0.000   Length:23463       Min.   :  0.00   Length:23463
 Class :character   Class :character   1st Qu.: 39.00   Class :character   1st Qu.:  1.000   Class :character   1st Qu.: 21.00   Class :character
 Mode  :character   Mode  :character   Median : 55.00   Mode  :character   Median :  1.000   Mode  :character   Median : 31.00   Mode  :character
                                       Mean   : 72.01                      Mean   :  1.368                      Mean   : 35.19
                                       3rd Qu.: 79.00                      3rd Qu.:  1.000                      3rd Qu.: 40.00
                                       Max.   :500.00                      Max.   :133.000                      Max.   :235.00
 NO2.AQI.Value    NO2.AQI.Category   PM2.5.AQI.Value  PM2.5.AQI.Category
 Min.   : 0.000   Length:23463       Min.   :  0.00   Length:23463
 1st Qu.: 0.000   Class :character   1st Qu.: 35.00   Class :character
 Median : 1.000   Mode  :character   Median : 54.00   Mode  :character
 Mean   : 3.063                      Mean   : 68.52
 3rd Qu.: 4.000                      3rd Qu.: 79.00
 Max.   :91.000                      Max.   :500.00
```

03 - Exploratory Data Analysis (EDA)

Data Overview

Merged Dataset

Description:

- Combines the air pollution and vehicle registration datasets by country to analyze the relationship between air quality and vehicle density.
- Number of Observations: 17,195 rows.
- Number of Features: 15 columns.

Key Features:

- Includes all features from the pollution dataset (e.g., AQI, pollutants) and vehicle density data.
- Added Registered.vehicles.per.1.000.people from the vehicle dataset.
- Dropped Code from the combined Dataset.
- Renamed Registered.vehicles.per.1.000.people to Vehicles

Necessity of Merge:

- Objective: To explore how vehicle density influences air pollution levels and AQI trends.
- By merging these datasets, we can link vehicle data to specific AQI observations, enabling comprehensive regression analysis.

Merged Dataset

	Country	City	AQI.Value	AQI.Category	CO.AQI.Value	CO.AQI.Cate...	Ozone.AQI.V...	Ozone.AQI.Category	NO2.AQI.Value	NO2.AQI.Category	PM2.5.AQI.Value	PM2.5.AQI.Category	Year	Vehicles
1	Afghanistan	Zaranj	133	Unhealthy fo...	1	Good	46	Good	0	Good	133	Unhealthy for Sensitive...	2013	20.724253
2	Afghanistan	Asmar	151	Unhealthy	2	Good	48	Good	1	Good	151	Unhealthy	2013	20.724253
3	Afghanistan	Panjab	83	Moderate	0	Good	28	Good	0	Good	83	Moderate	2013	20.724253
4	Afghanistan	Mehtar Lam	63	Moderate	1	Good	45	Good	0	Good	63	Moderate	2013	20.724253
5	Afghanistan	Mirabad	165	Unhealthy	1	Good	45	Good	0	Good	165	Unhealthy	2013	20.724253
6	Afghanistan	Rostaq	113	Unhealthy fo...	1	Good	42	Good	0	Good	113	Unhealthy for Sensitive...	2013	20.724253
7	Afghanistan	Kholm	123	Unhealthy fo...	1	Good	35	Good	0	Good	123	Unhealthy for Sensitive...	2013	20.724253
8	Afghanistan	Lar Gerd	70	Moderate	0	Good	34	Good	0	Good	70	Moderate	2013	20.724253
9	Afghanistan	Kuhestan	151	Unhealthy	1	Good	41	Good	0	Good	151	Unhealthy	2013	20.724253
10	Afghanistan	Gazni	83	Moderate	0	Good	40	Good	0	Good	83	Moderate	2013	20.724253
11	Afghanistan	Taloqan	127	Unhealthy fo...	1	Good	29	Good	0	Good	127	Unhealthy for Sensitive...	2013	20.724253
12	Afghanistan	Baglan	72	Moderate	1	Good	44	Good	0	Good	72	Moderate	2013	20.724253

```
> print("Combined Data Overview")
[1] "Combined Data Overview"
> print(summary(merged_data))
   Country               City             AQI.Value      AQI.Category        CO.AQI.Value     CO.AQI.Category     Ozone.AQI.Value  Ozone.AQI.Category
 Length:17195       Length:17195       Min.   :  7.00   Length:17195       Min.   : 0.000   Length:17195       Min.   :  0.00   Length:17195
 Class :character   Class :character   1st Qu.: 40.00   Class :character   1st Qu.: 1.000   Class :character   1st Qu.: 22.00   Class :character
 Mode  :character   Mode  :character   Median : 57.00   Mode  :character   Median : 1.000   Mode  :character   Median : 31.00   Mode  :character
                                       Mean   : 76.95                      Mean   : 1.361                      Mean   : 37.19
                                       3rd Qu.: 88.00                      3rd Qu.: 1.000                      3rd Qu.: 41.00
                                       Max.   :500.00                      Max.   :67.000                      Max.   :210.00
 NO2.AQI.Value    NO2.AQI.Category   PM2.5.AQI.Value  PM2.5.AQI.Category      Year         Vehicles
 Min.   : 0.000   Length:17195       Min.   :  0.00   Length:17195       Min.   :2007   Min.   :  4.457
 1st Qu.: 0.000   Class :character   1st Qu.: 36.00   Class :character   1st Qu.:2015   1st Qu.:158.147
 Median : 1.000   Mode  :character   Median : 57.00   Mode  :character   Median :2016   Median :461.903
 Mean   : 2.396                      Mean   : 73.35                      Mean   :2016   Mean   :427.863
 3rd Qu.: 3.000                      3rd Qu.: 87.00                      3rd Qu.:2016   3rd Qu.:652.578
 Max.   :64.000                      Max.   :500.00                      Max.   :2017   Max.   :949.482
```

Missing Values

```
> print(merged_data_missing)
          Country              City         AQI.Value      AQI.Category      CO.AQI.Value   CO.AQI.Category   Ozone.AQI.Value Ozone.AQI.Category
                0                 0                 0                 0                 0                 0                 0                  0
     NO2.AQI.Value  NO2.AQI.Category   PM2.5.AQI.Value PM2.5.AQI.Category          Year          Vehicles
                0                 0                 0                 0                 0                 0
```

Our Merged Dataset Doesn't contain any missing values

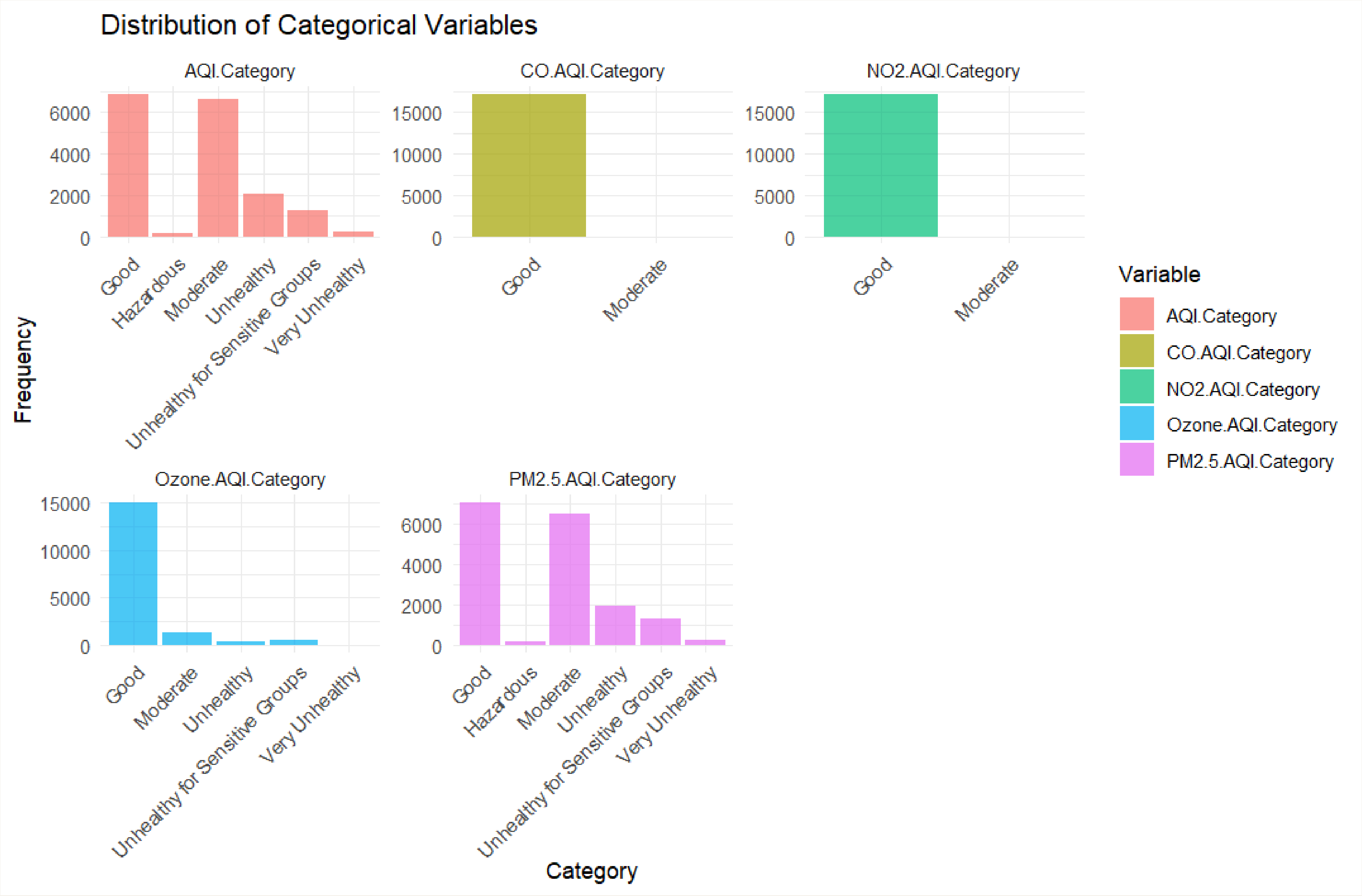
Duplicated Rows

```
> print(merged_data_missing)
          Country              City         AQI.Value      AQI.Category      CO.AQI.Value   CO.AQI.Category   Ozone.AQI.Value Ozone.AQI.Category
                0                 0                 0                 0                 0                 0                 0                  0
     NO2.AQI.Value  NO2.AQI.Category   PM2.5.AQI.Value PM2.5.AQI.Category          Year          Vehicles
                0                 0                 0                 0                 0                 0
```

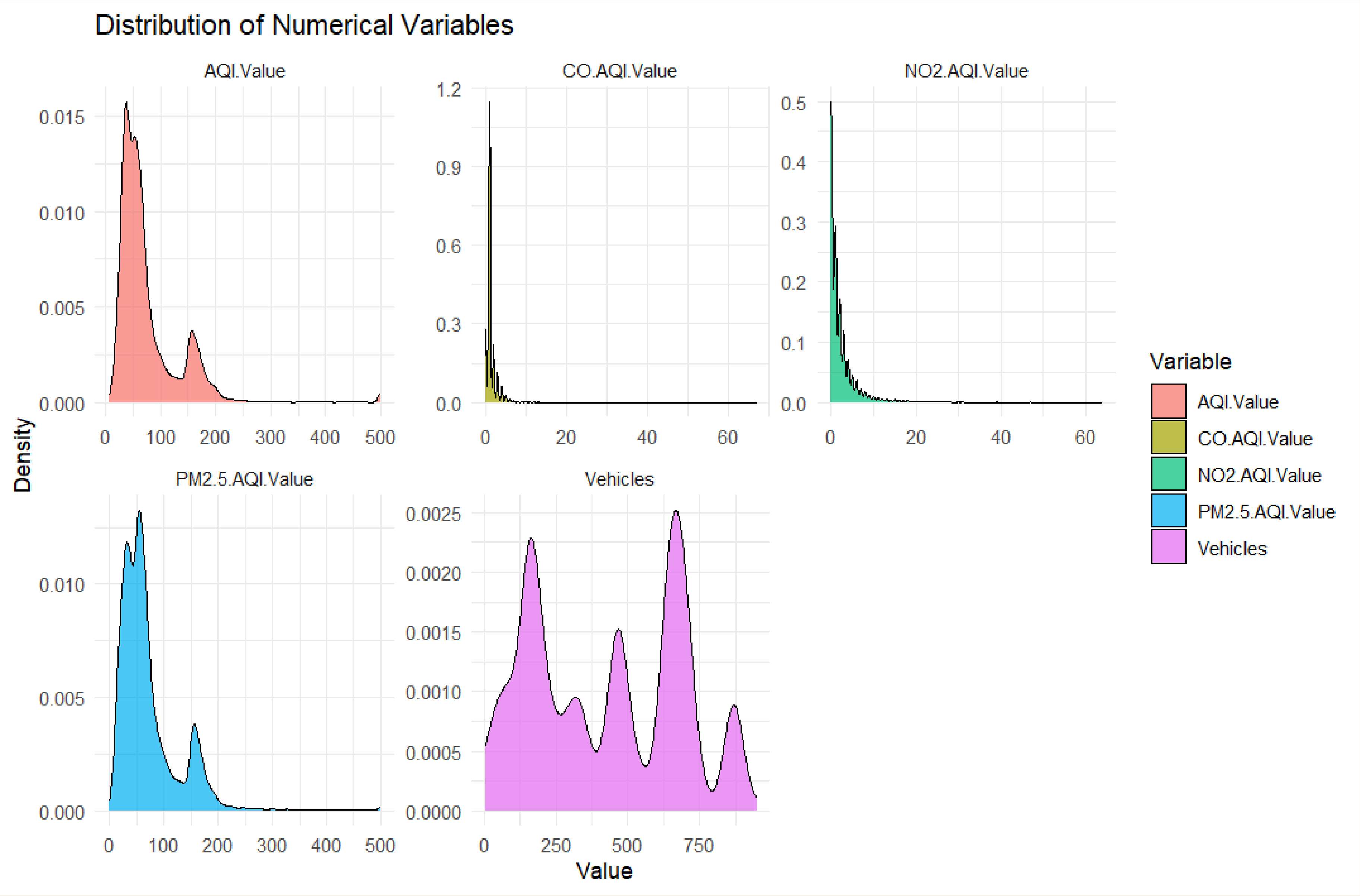
Our Merged Dataset Doesn't contain any duplicated rows

Numerical Vs Categorical

Categorical Features



Numerical Features



Label Encoding

```r
# label encoding
merged_data <- merged_data %>%
  mutate(
    AQI.Category = as.numeric(factor(AQI.Category)),
    CO.AQI.Category = as.numeric(factor(CO.AQI.Category)),
    Ozone.AQI.Category = as.numeric(factor(Ozone.AQI.Category)),
    NO2.AQI.Category = as.numeric(factor(NO2.AQI.Category)),
    PM2.5.AQI.Category = as.numeric(factor(PM2.5.AQI.Category))
  )
```

AQI.Category

All | Possible | Excluded | None

Value	Count
Good	6848
Hazardous	181
Moderate	6598
Unhealthy	2053
Unhealthy for Sensit...	1257
Very Unhealthy	258

Values: 6

CO.AQI.Category

All | Possible | Excluded | None

Value	Count
Good	17194
Moderate	1

Values: 2

Ozone.AQI.Category

All | Possible | Excluded | None

Value	Count
Good	15009
Moderate	1280
Unhealthy	393
Unhealthy for Sensit...	465
Very Unhealthy	48

Values: 5

NO2.AQI.Category

All | Possible | Excluded | None

Value	Count
Good	17183
Moderate	12

Values: 2

PM2.5.AQI.Category

All | Possible | Excluded | None

Value	Count
Good	7054
Hazardous	163
Moderate	6502
Unhealthy	1959
Unhealthy for Sensit...	1287
Very Unhealthy	230

Values: 6

	Country	City	AQI.Value	AQI.Category	CO.AQI.Value	CO.AQI.Cate...	Ozone.AQI.V...	Ozone.AQI.Category	NO2.AQI.Value	NO2.AQI.Category	PM2.5.AQI.Value	PM2.5.AQI.Category	Year	Vehicles
1	Afghanistan	Zaranj	133	Unhealthy fo...	1	Good	46	Good	0	Good	133	Unhealthy for Sensitive...	2013	20.724253
2	Afghanistan	Asmar	151	Unhealthy	2	Good	48	Good	1	Good	151	Unhealthy	2013	20.724253
3	Afghanistan	Panjab	83	Moderate	0	Good	28	Good	0	Good	83	Moderate	2013	20.724253
4	Afghanistan	Mehtar Lam	63	Moderate	1	Good	45	Good	0	Good	63	Moderate	2013	20.724253
5	Afghanistan	Mirabad	165	Unhealthy	1	Good	45	Good	0	Good	165	Unhealthy	2013	20.724253
6	Afghanistan	Rostaq	113	Unhealthy fo...	1	Good	42	Good	0	Good	113	Unhealthy for Sensitive...	2013	20.724253
7	Afghanistan	Kholm	123	Unhealthy fo...	1	Good	35	Good	0	Good	123	Unhealthy for Sensitive...	2013	20.724253
8	Afghanistan	Lar Gerd	70	Moderate	0	Good	34	Good	0	Good	70	Moderate	2013	20.724253

We used label encoding to to convert categorical columns into numerical ones

AQI.Category

All | Possible | Excluded | None

Value	Count
1	6848
2	181
3	6598
4	2053
5	1257
6	258

Values: 6

CO.AQI.Category

All | Possible | Excluded | None

Value	Count
1	17194
2	1

Values: 2

Ozone.AQI.Category

All | Possible | Excluded | None

Value	Count
1	15009
2	1280
3	393
4	465
5	48

Values: 5

NO2.AQI.Category

All | Possible | Excluded | None

Value	Count
1	17183
2	12

Values: 2

PM2.5.AQI.Category

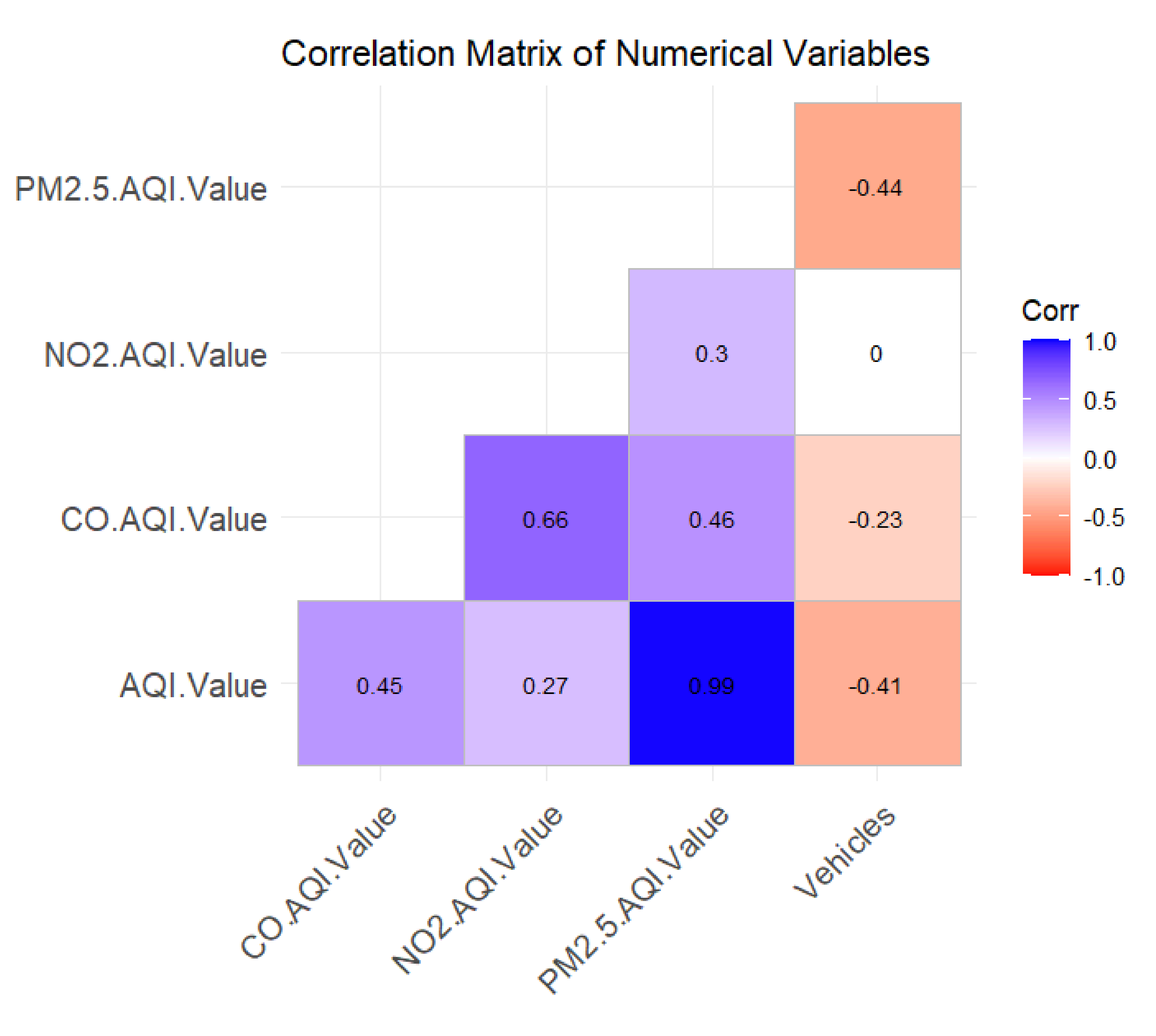
All | Possible | Excluded | None

Value	Count
1	7054
2	163
3	6502
4	1959
5	1287
6	230

Values: 6

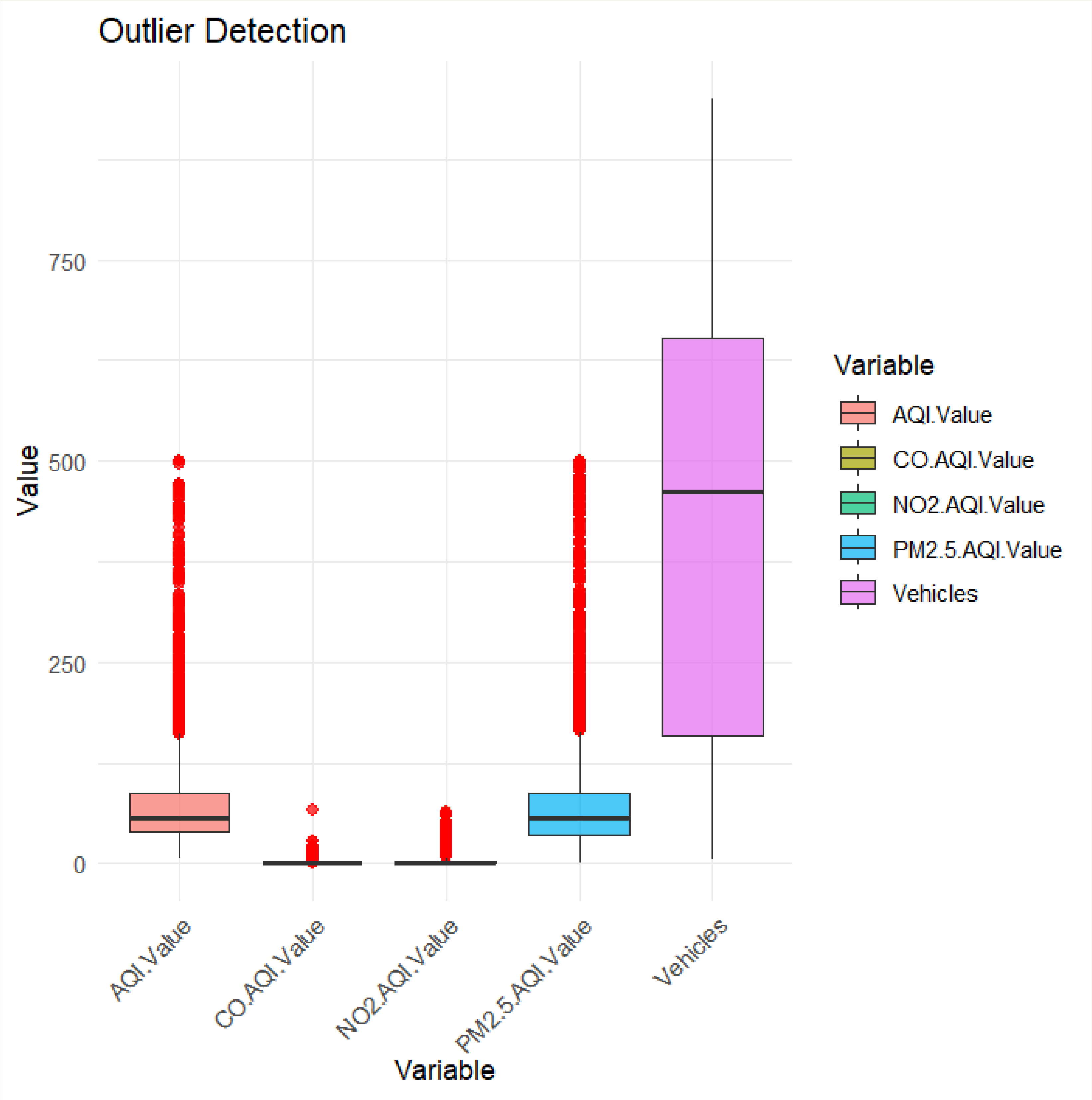
	Country	City	AQI.Value	AQI.Category	CO.AQI.Value	CO.AQI.Cate...	Ozone.AQI.V...	Ozone.AQI.C...	NO2.AQI.Value	NO2.AQI.Cat...	PM2.5.AQI.V...	PM2.5.AQI.C...	Year	Vehicles
1	Afghanistan	Zaranj	133	5	1	1	46	1	0	1	133	5	2013	20.724253
2	Afghanistan	Asmar	151	4	2	1	48	1	1	1	151	4	2013	20.724253
3	Afghanistan	Panjab	83	3	0	1	28	1	0	1	83	3	2013	20.724253
4	Afghanistan	Mehtar Lam	63	3	1	1	45	1	0	1	63	3	2013	20.724253
5	Afghanistan	Mirabad	165	4	1	1	45	1	0	1	165	4	2013	20.724253
6	Afghanistan	Rostaq	113	5	1	1	42	1	0	1	113	5	2013	20.724253
7	Afghanistan	Kholm	123	5	1	1	35	1	0	1	123	5	2013	20.724253

Correlation Matrix



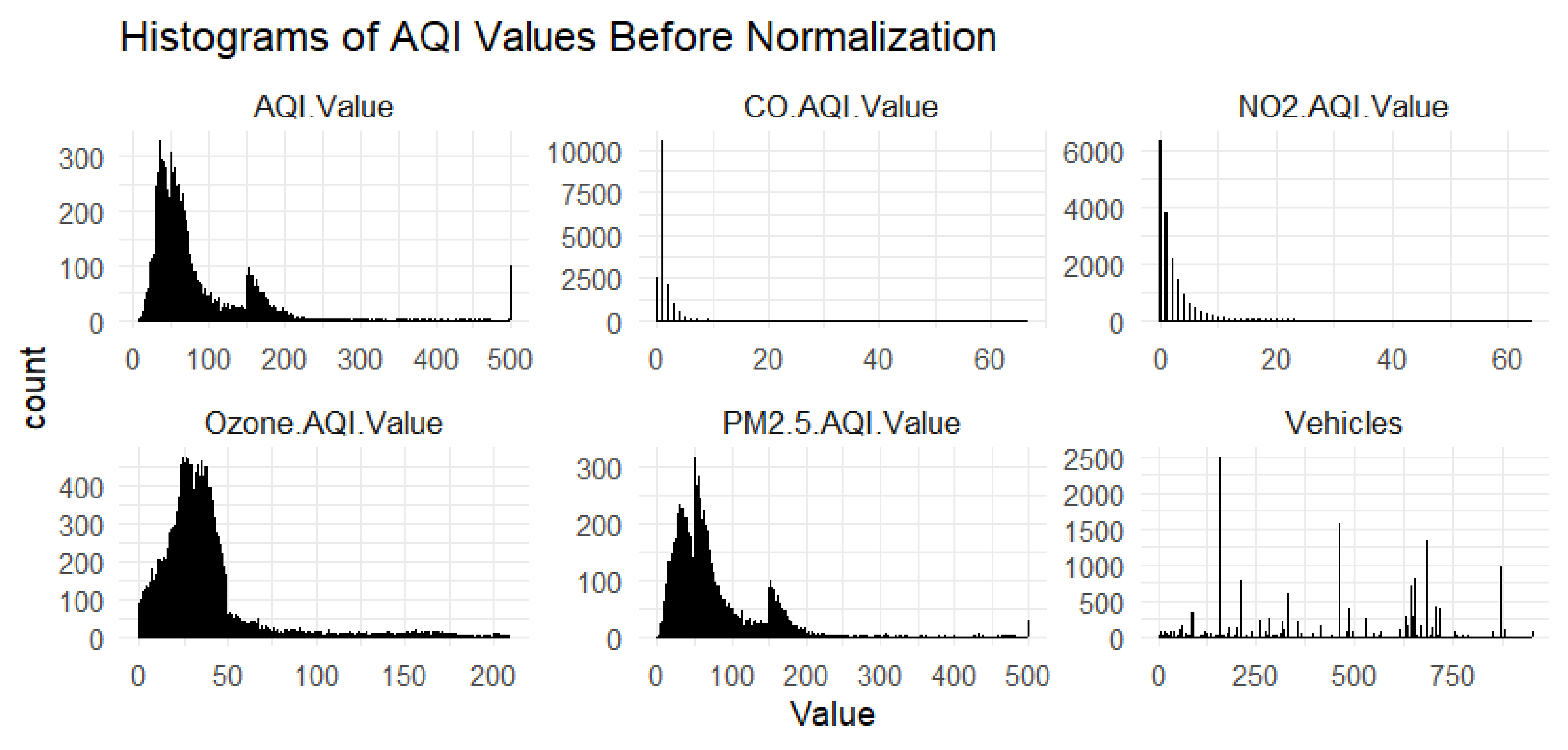
The correlation matrix shows strong multicollinearity between PM2.5.AQI.Value and AQI.Value (0.99) and a moderate correlation between CO.AQI.Value and other variables (e.g., 0.66).

Outliers



The boxplot highlights significant outliers in the AQI.Value, PM2.5.AQI.Value, and Vehicles variables. These outliers could skew results for parametric tests.

Histogram



The histograms reveal highly skewed distributions for most numerical variables (e.g., AQI.Value, CO.AQI.Value, PM2.5.AQI.Value), deviating significantly from normality.

Given the high skewness, presence of significant outliers, and multicollinearity, it is appropriate to proceed with normalization to standardize the scale of numerical features, reduce skewness, and prepare the dataset for further analysis.

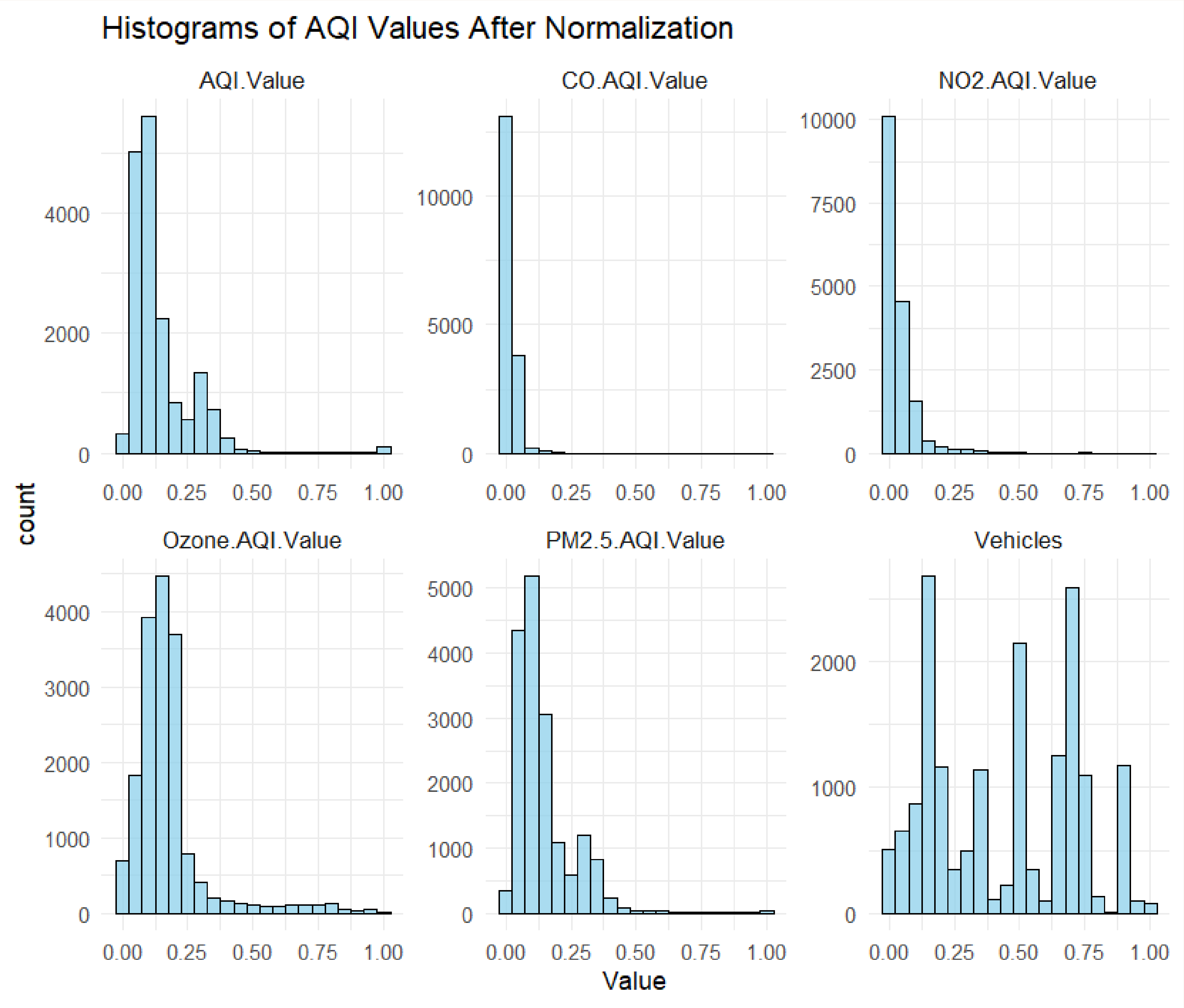
Normalization

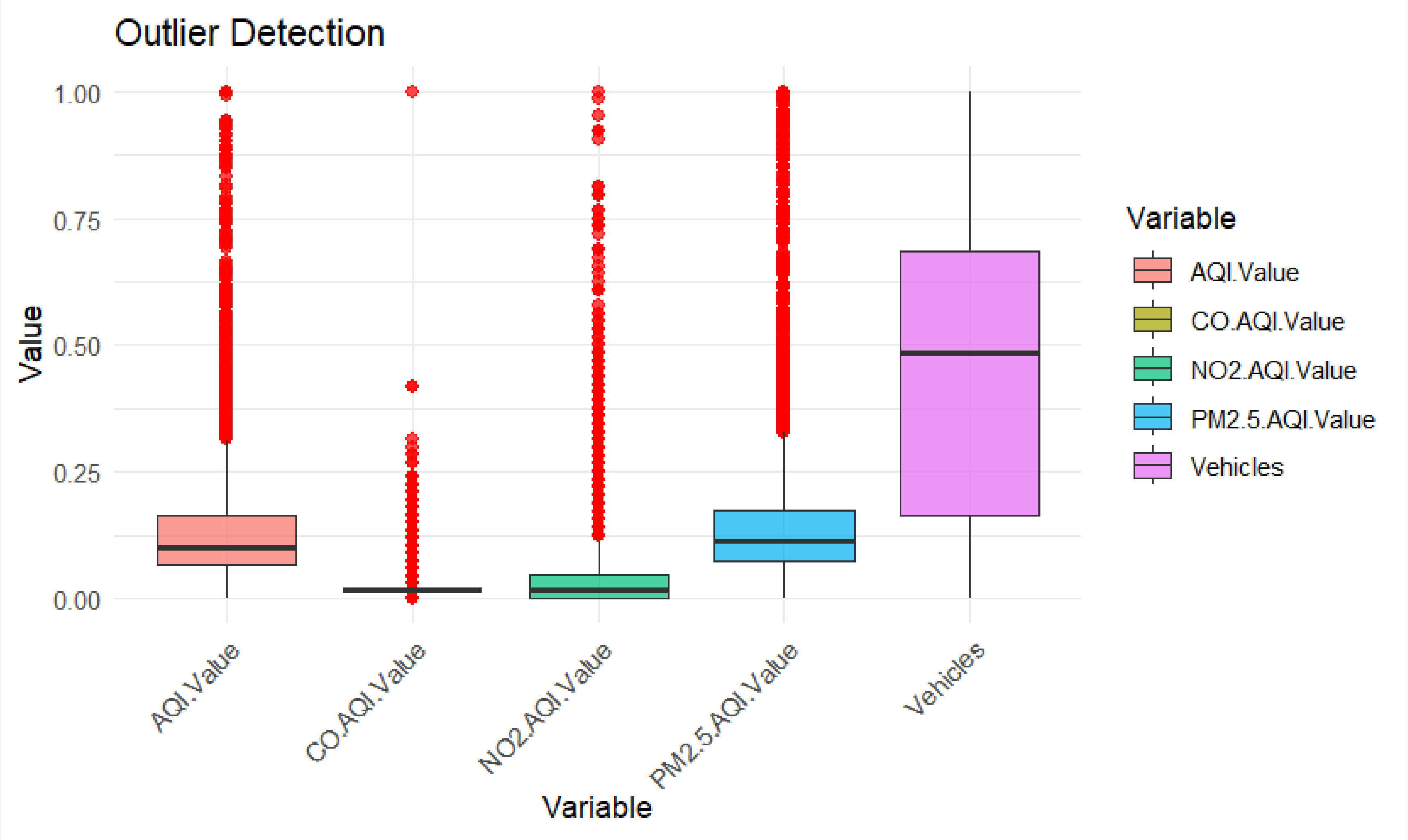
The method we used here is min-max normalization, which rescales the data into a range of [0, 1].

```r
merged_data <- merged_data %>%
  mutate(
    AQI.Value = rescale(AQI.Value, to = c(0, 1)),
    CO.AQI.Value = rescale(CO.AQI.Value, to = c(0, 1)),
    Ozone.AQI.Value = rescale(Ozone.AQI.Value, to = c(0, 1)),
    NO2.AQI.Value = rescale(NO2.AQI.Value, to = c(0, 1)),
    PM2.5.AQI.Value = rescale(PM2.5.AQI.Value, to = c(0, 1)),
    Vehicles = rescale(Vehicles, to = c(0, 1))
  )
```

This technique preserves the relative relationships of the original data while ensuring all numerical features are on the same scale, making it suitable for models sensitive to varying ranges.

```
> shapiro_test_results <- merged_data_cleaned %>%
+   select(AQI.Value, CO.AQI.Value, Ozone.AQI.Value, NO2.AQI.Value, PM2.5.AQI.Value, Vehicles) %>%
+   summarise(across(everything(), ~ shapiro.test(.)$p.value)) %>%
+   gather(key = "Variable", value = "Shapiro-Wilk p-value")
Error in `summarise()`:
i In argument: `across(everything(), ~shapiro.test(.)$p.value)`.
Caused by error in `across()`:
! Can't compute column `AQI.Value`.
Caused by error in `shapiro.test()`:
! sample size must be between 3 and 5000
Run `rlang::last_trace()` to see where the error occurred.
> # Print the results of the Shapiro-Wilk test
> print(shapiro_test_results)
Error: object 'shapiro_test_results' not found
>
```





The normalized data is still skewed for most variables (AQI.Value, CO, NO2, PM2.5, etc.), although the values are now constrained between 0 and 1.

The presence of outliers and non-normality suggests that assumptions for parametric tests like t-tests or ANOVA may not hold.

```r
# Compute skewness and kurtosis for numerical variables
skewness_kurtosis <- data.frame(
  Variable = c("AQI.Value", "CO.AQI.Value", "Ozone.AQI.Value", "NO2.AQI.Value", "PM2.5.AQI.Value", "Vehicles"),
  Skewness = c(
    skewness(merged_data$AQI.Value, na.rm = TRUE),
    skewness(merged_data$CO.AQI.Value, na.rm = TRUE),
    skewness(merged_data$Ozone.AQI.Value, na.rm = TRUE),
    skewness(merged_data$NO2.AQI.Value, na.rm = TRUE),
    skewness(merged_data$PM2.5.AQI.Value, na.rm = TRUE),
    skewness(merged_data$Vehicles, na.rm = TRUE)
  ),
  Kurtosis = c(
    kurtosis(merged_data$AQI.Value, na.rm = TRUE),
    kurtosis(merged_data$CO.AQI.Value, na.rm = TRUE),
    kurtosis(merged_data$Ozone.AQI.Value, na.rm = TRUE),
    kurtosis(merged_data$NO2.AQI.Value, na.rm = TRUE),
    kurtosis(merged_data$PM2.5.AQI.Value, na.rm = TRUE),
    kurtosis(merged_data$Vehicles, na.rm = TRUE)
  )
)
```

```
> print(skewness_kurtosis)
         Variable   Skewness   Kurtosis
1       AQI.Value 3.03206853  17.699996
2    CO.AQI.Value 8.75102520 255.434551
3 Ozone.AQI.Value 2.79921156  12.244145
4   NO2.AQI.Value 4.87096042  41.409858
5 PM2.5.AQI.Value 2.61983127  14.508415
6        Vehicles 0.08333046   1.707021
```

AQI.Value:

Skewness: 3.03, indicating a right-skewed distribution (long tail on the right).

Kurtosis: 17.7, indicating a highly leptokurtic distribution (heavy tails with more outliers than normal).

CO.AQI.Value:

Skewness: 8.75, suggesting high positive skew (significant right skew).

Kurtosis: 255.43, suggesting extremely heavy tails (very high presence of outliers).

Ozone.AQI.Value:

Skewness: 2.80, showing moderate right skew.

Kurtosis: 12.24, indicating moderately high kurtosis, meaning some outliers.

NO2.AQI.Value:

Skewness: 4.87, indicating strong right skew.

Kurtosis: 41.41, showing extremely high kurtosis, pointing to many outliers.

PM2.5.AQI.Value:

Skewness: 2.62, indicating moderate right skew.

Kurtosis: 14.51, showing high kurtosis (more outliers).

Vehicles:

Skewness: 0.08, indicating very little skew, close to a normal distribution.

Kurtosis: 1.71, indicating a platykurtic distribution (light tails, fewer outliers than normal).

Analyzing Air Quality Across Vehicle Density Groups

Test if AQI significantly differs across four vehicle density groups.

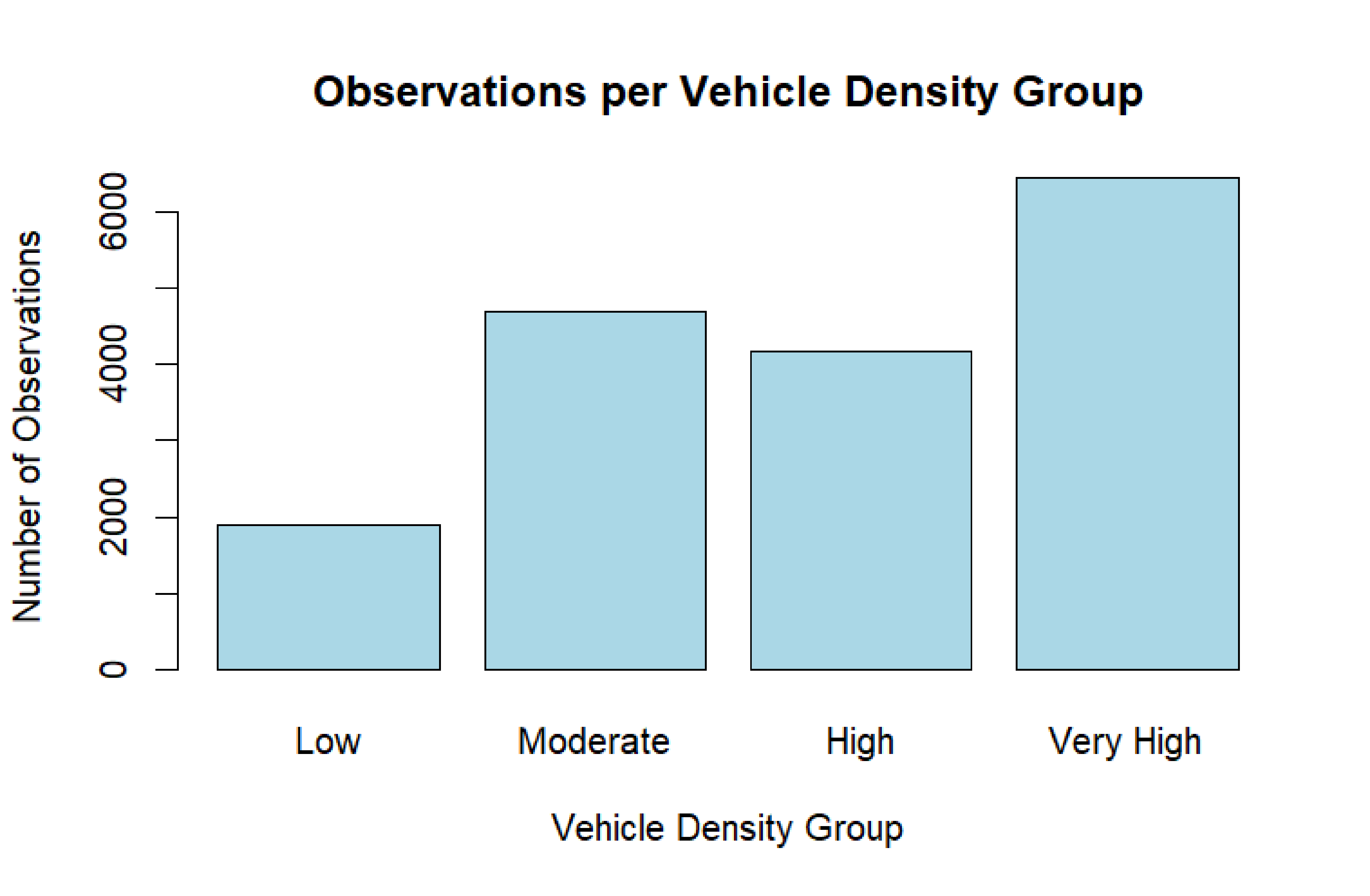
We will Investigate the relationship between vehicle density and air quality index (AQI).

Key Variables:

AQI.Value: *Air Quality Index (normalized).*

Vehicle Density Groups: *Created by binning vehicle density into "Low," "Moderate," "High," and "Very High."*

```
# Step 1: Binning the vehicle density data based on 'Vehicles'
merged_data_cleaned <- merged_data_cleaned %>%
  mutate(Vehicle_Density_Group = cut(Vehicles,
                                     breaks = c(-Inf, 0.1, 0.3, 0.6, Inf),  # Adjust bin thresholds as needed
                                     labels = c("Low", "Moderate", "High", "Very High")))
```



Analyzing Air Quality Across Vehicle Density Groups

```
kruskal_test <- kruskal.test(AQI.Value ~ Vehicle_Density_Group, data = merged_data_cleaned)
```

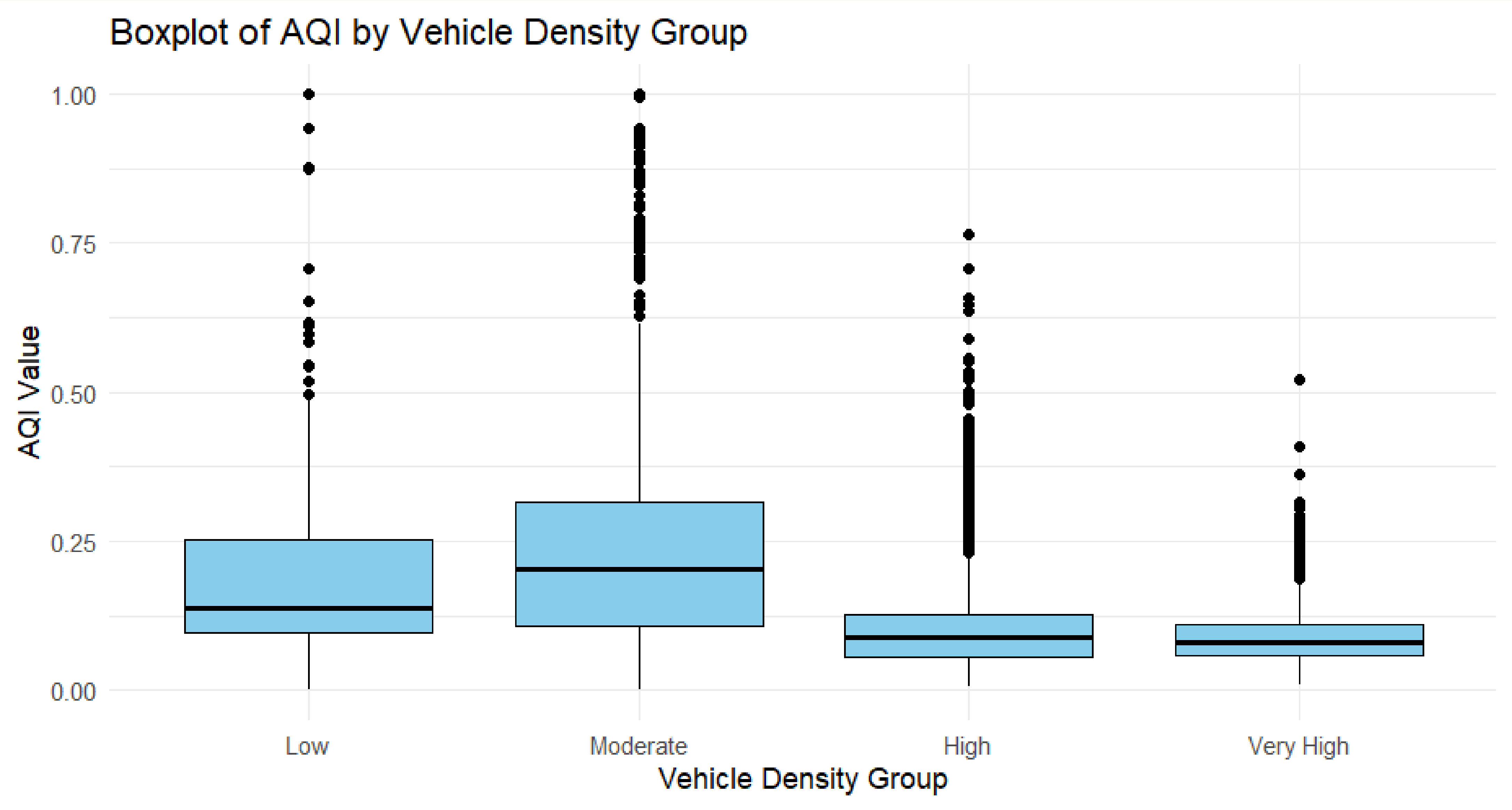
Kruskal-Wallis Test for Group Differences
Step 1: Hypotheses
Null Hypothesis (H0):
AQI values are the same across vehicle density groups.
Alternative Hypothesis (H1):
At least one group has significantly different AQI values.
Step 2: Kruskal-Wallis Test Results
Chi-squared value: 4150.9
Degrees of freedom: 3
p-value: < 2.2e-16

There is strong evidence to suggest that the AQI values differ significantly among the four vehicle density groups ("Low," "Moderate," "High," and "Very High"). This result implies that vehicle density likely impacts air quality (as represented by AQI).

```
> cat("Kruskal-Wallis Test Results:\n")
Kruskal-Wallis Test Results:
> print(kruskal_test)

	Kruskal-Wallis rank sum test

data:  AQI.Value by Vehicle_Density_Group
Kruskal-Wallis chi-squared = 4150.9, df = 3, p-value < 2.2e-16
```

```
if (kruskal_test$p.value < 0.05) {
  cat("The Kruskal-Wallis test indicates significant differences in AQI values across vehicle density groups.\n")
  cat("Review the pairwise Wilcoxon test results for detailed group comparisons.\n")
} else {
  cat("No significant difference in AQI values across vehicle density groups.\n")
}
```

The boxplot visualizes the distribution of AQI values across the different vehicle density groups. It also includes significance levels, showing which comparisons have statistically significant differences in AQI values.



04 - Linear Regression Models & Evaluation

Simple Linear Regression

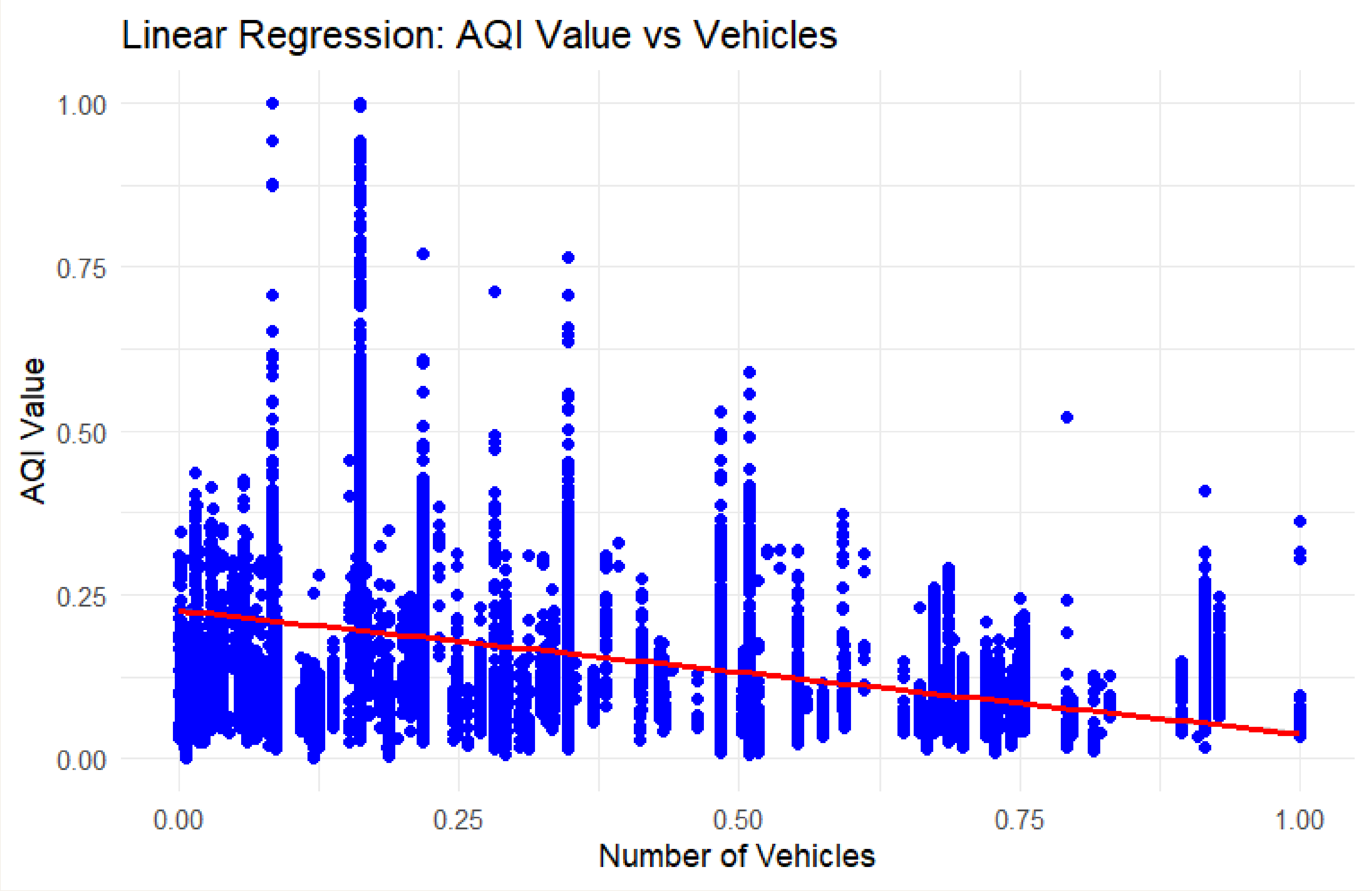
```r
# Linear Regression: AQI vs Vehicles_Per_1000_People
linear_model <- lm(AQI.Value ~ Vehicles, data = merged_data_cleaned)
```

```
> cat("Linear Regression Summary:\n")
Linear Regression Summary:
> summary(linear_model)

Call:
lm(formula = AQI.Value ~ Vehicles, data = merged_data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.22283 -0.06519 -0.02361  0.04138  0.80439 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.226219   0.001653  136.87   <2e-16 ***
Vehicles    -0.188206   0.003147  -59.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1131 on 17193 degrees of freedom
Multiple R-squared:  0.1722,	Adjusted R-squared:  0.1722 
F-statistic:  3577 on 1 and 17193 DF,  p-value: < 2.2e-16
```

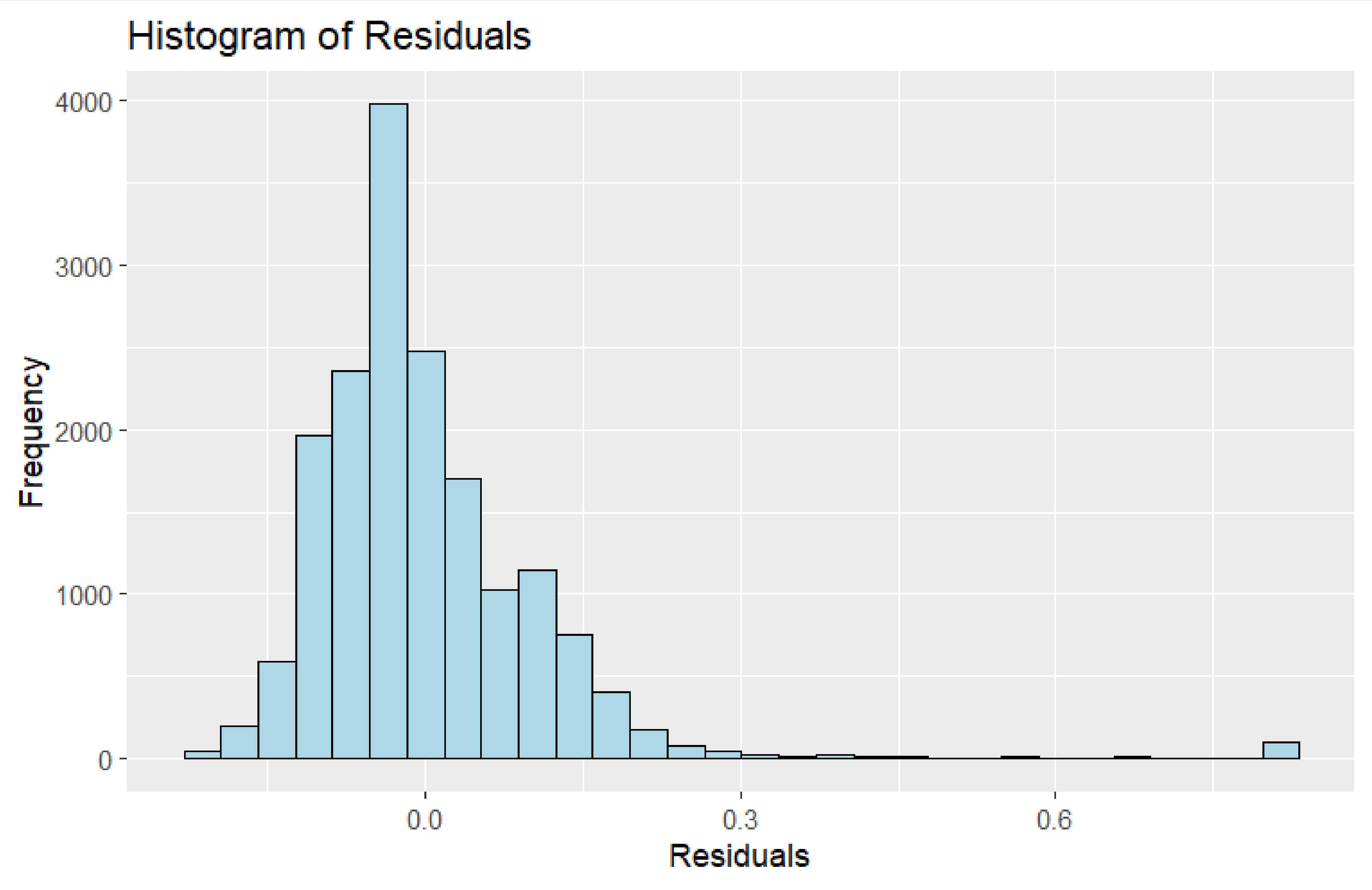
Simple Linear Regression

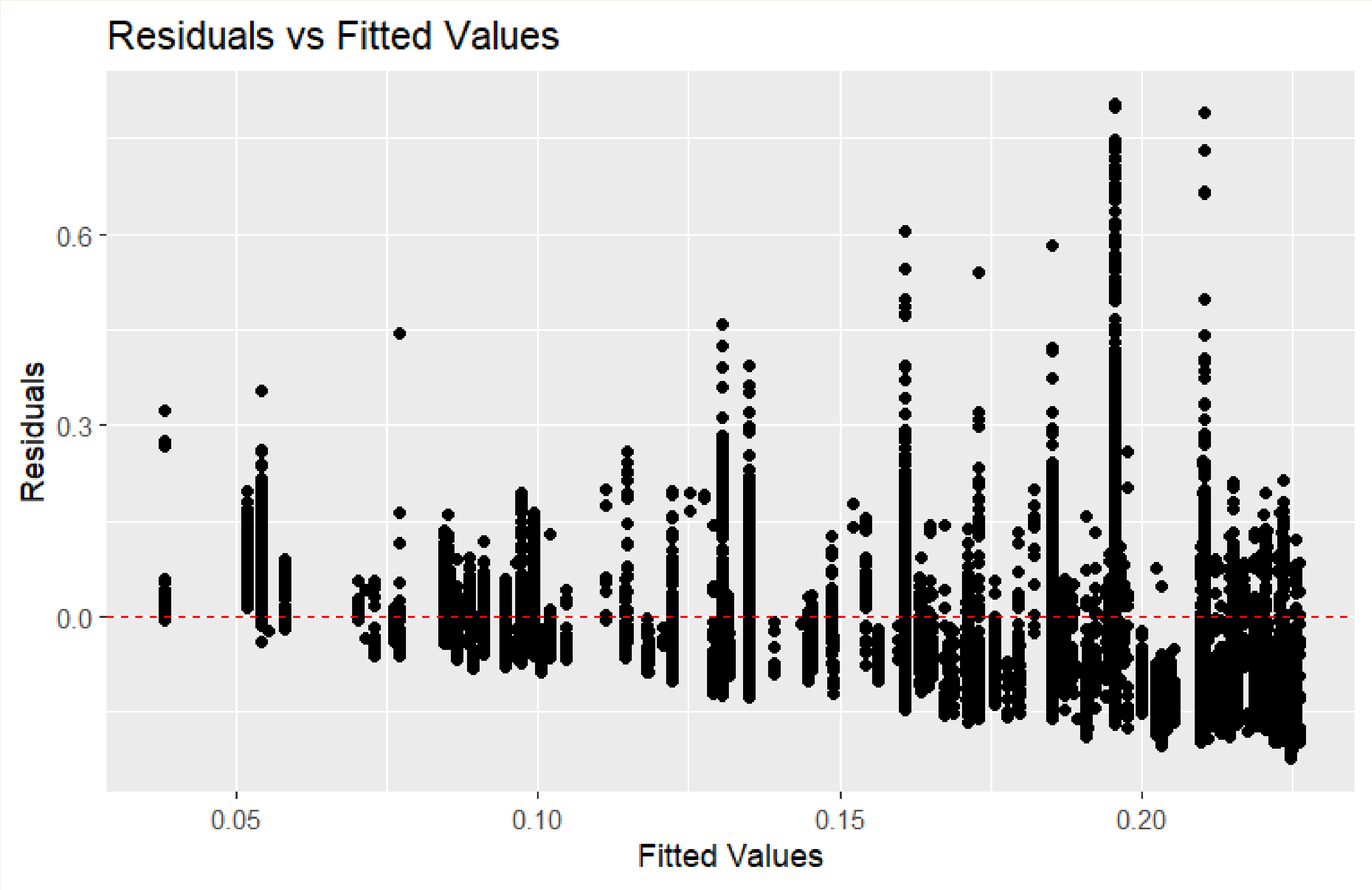
The linear regression analysis between AQI.Value and Vehicles (number of vehicles) shows the following:

1. Negative Trend:
   - The regression line (red) shows a negative slope. This indicates a weak negative relationship between the number of vehicles and AQI value.
   - As the number of vehicles increases, the AQI value tends to decrease slightly.
2. Scattered Data:
   - The data points (blue) are widely scattered, suggesting high variance in AQI values for any given number of vehicles.
   - This weakens the strength of the linear relationship.
3. Outliers:
   - The presence of many points far from the regression line shows potential outliers or noise in the data, which could affect the model's accuracy.



Simple Linear Regression Evaluation






```
> cat("Linear Regression RMSE:", linear_rmse, "\n")
Linear Regression RMSE: 0.1130997
> cat("Linear Regression MAE:", linear_mae, "\n")
Linear Regression MAE: 0.07592861
> cat("Linear Regression MAPE:", linear_mape, "%\n")
Linear Regression MAPE: Inf %
```

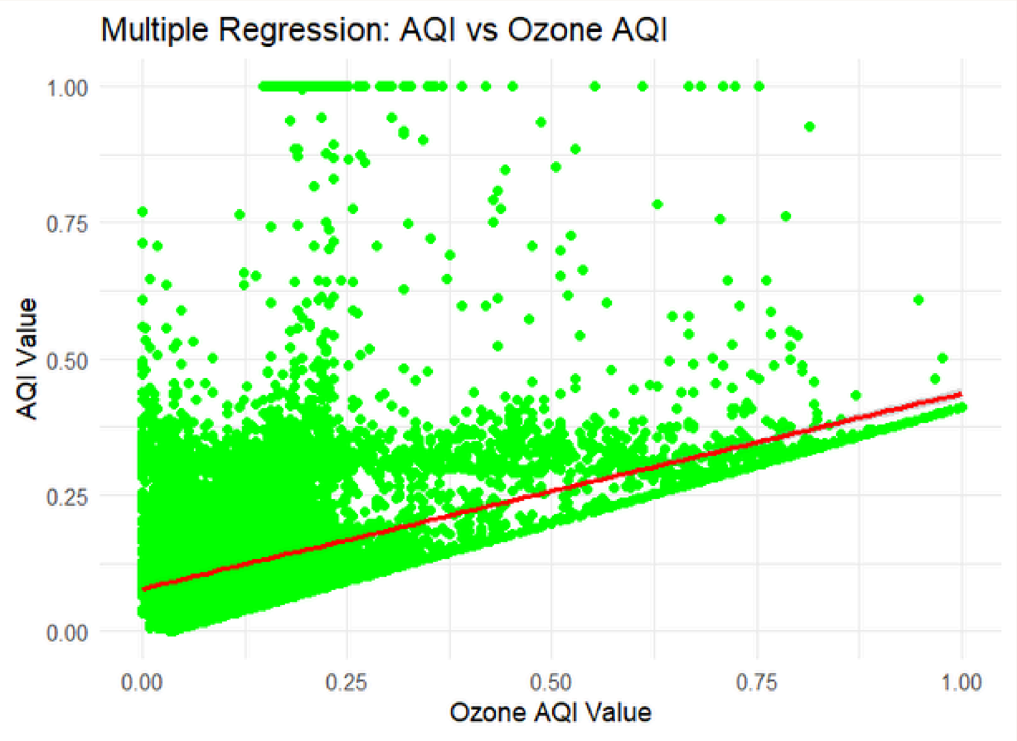
04 - Linear Regression Models

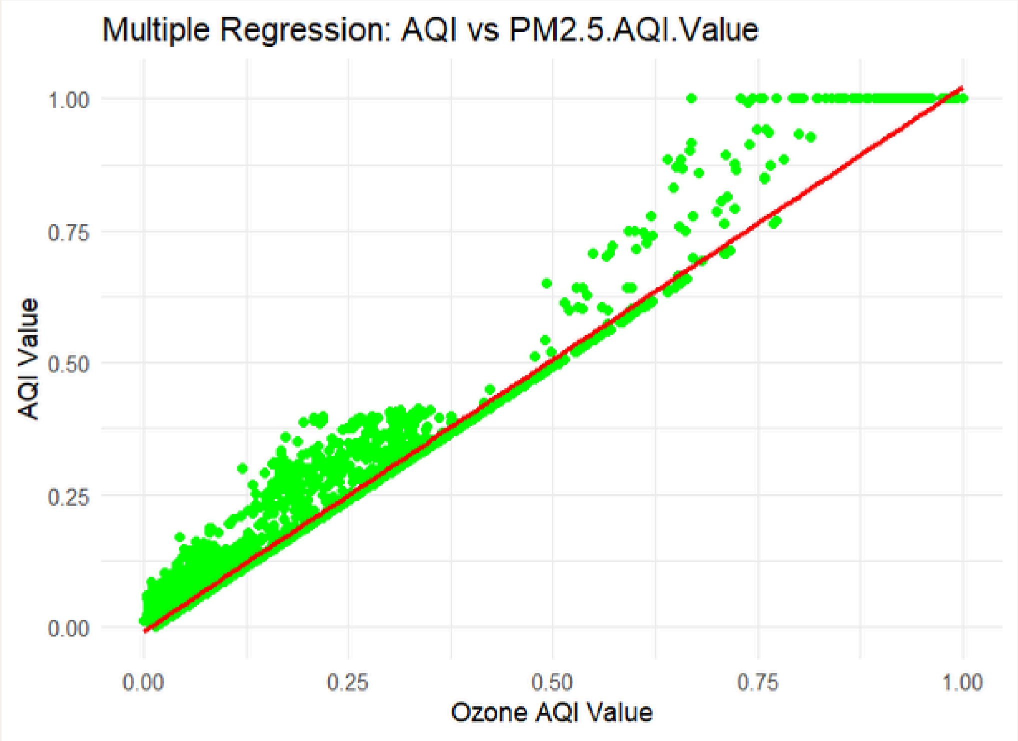
Multiple Linear Regression

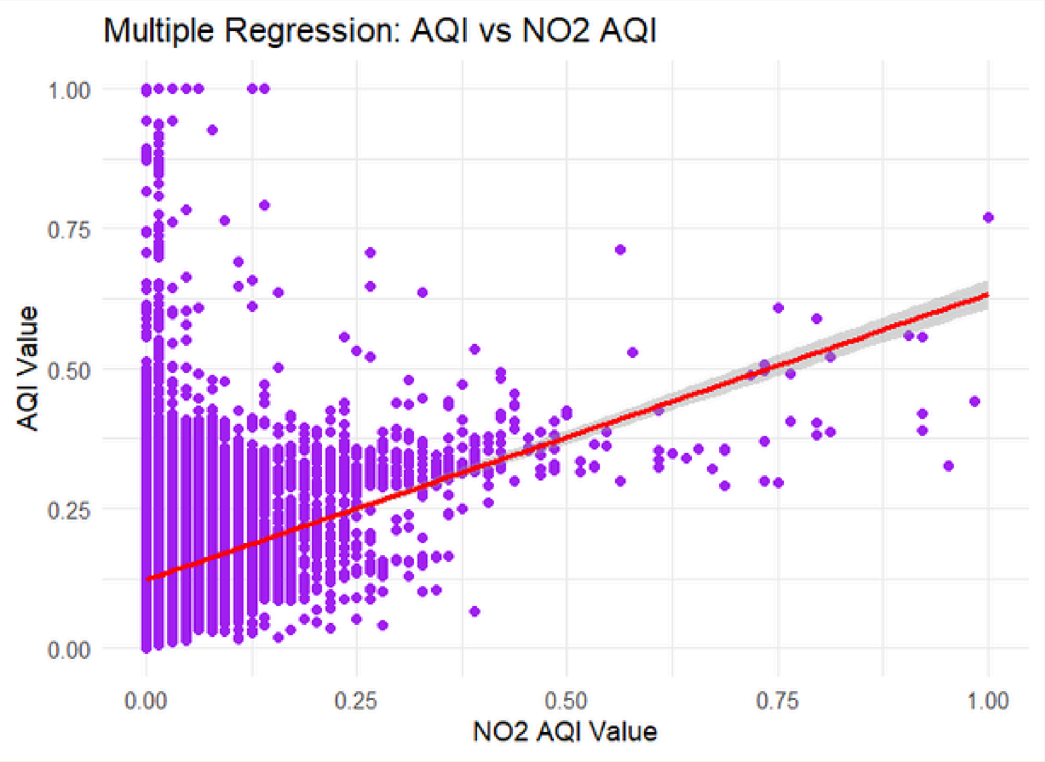
```
# Multiple Regression
multiple_model <- lm(AQI.Value ~ CO.AQI.Value + Ozone.AQI.Value + NO2.AQI.Value + Vehicles, data = merged_data_cleaned)
```

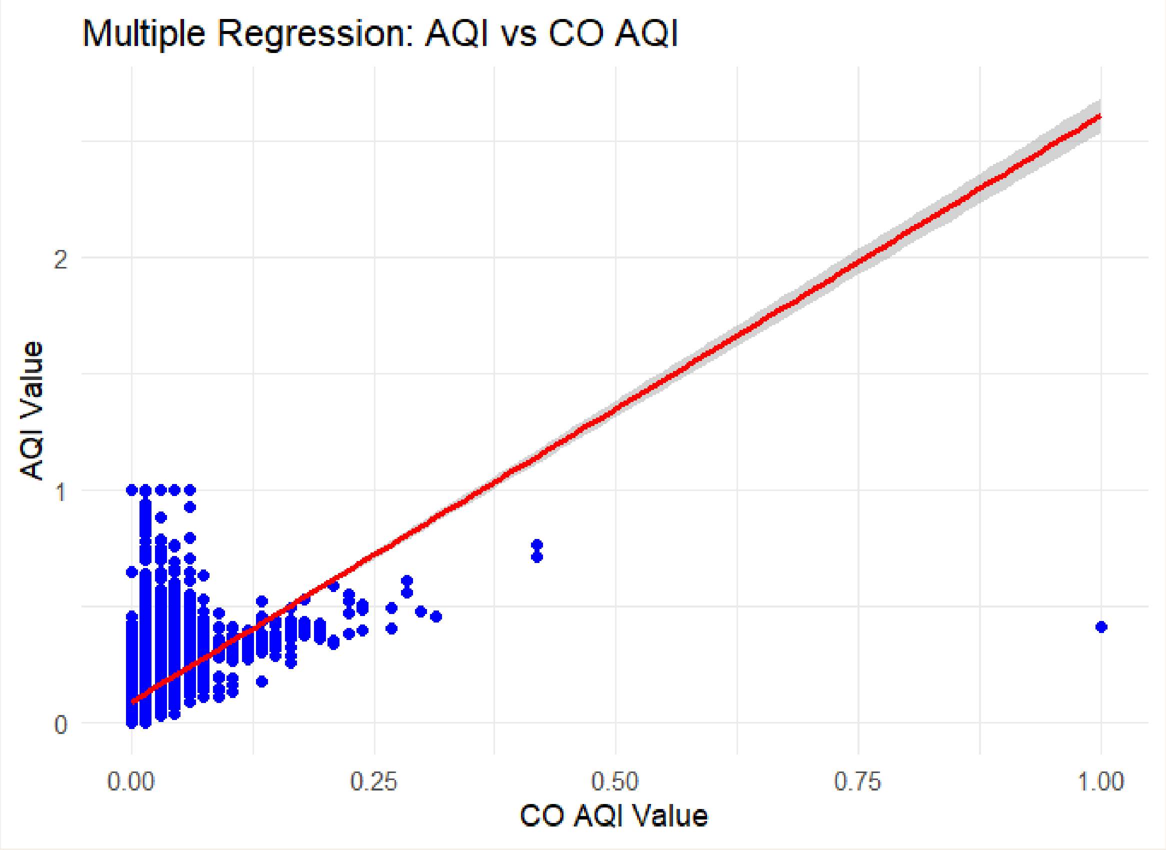
```
> cat("Multiple Regression Summary:\n")
Multiple Regression Summary:
> summary(multiple_model)

Call:
lm(formula = AQI.Value ~ CO.AQI.Value + Ozone.AQI.Value + NO2.AQI.Value +
    Vehicles, data = merged_data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58603 -0.04357 -0.01718  0.01797  0.85405

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.123828   0.001872   66.14   <2e-16 ***
CO.AQI.Value     0.681102   0.049821   13.67   <2e-16 ***
Ozone.AQI.Value  0.318839   0.005445   58.56   <2e-16 ***
NO2.AQI.Value    0.440345   0.015828   27.82   <2e-16 ***
Vehicles        -0.153368   0.002761  -55.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09458 on 17190 degrees of freedom
Multiple R-squared:  0.4213,    Adjusted R-squared:  0.4211
F-statistic:  3128 on 4 and 17190 DF,  p-value: < 2.2e-16
```

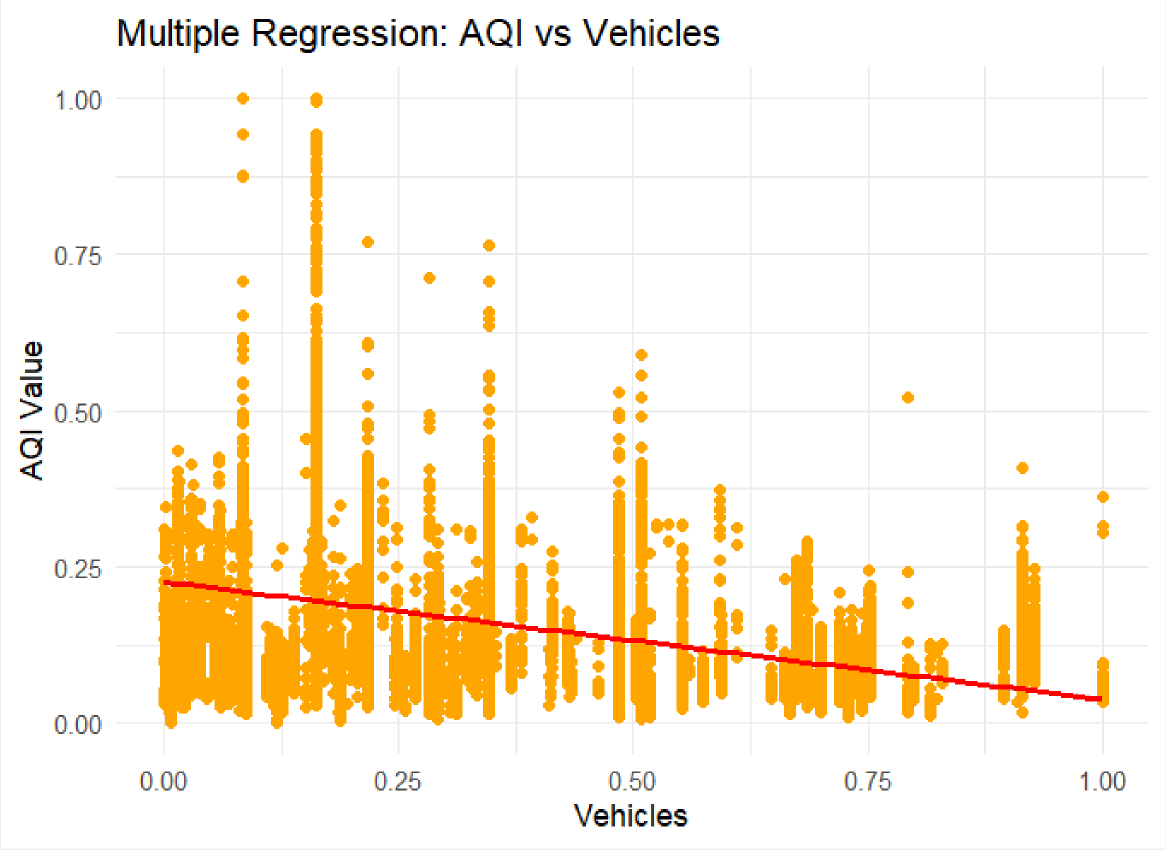
Multiple Linear Regression











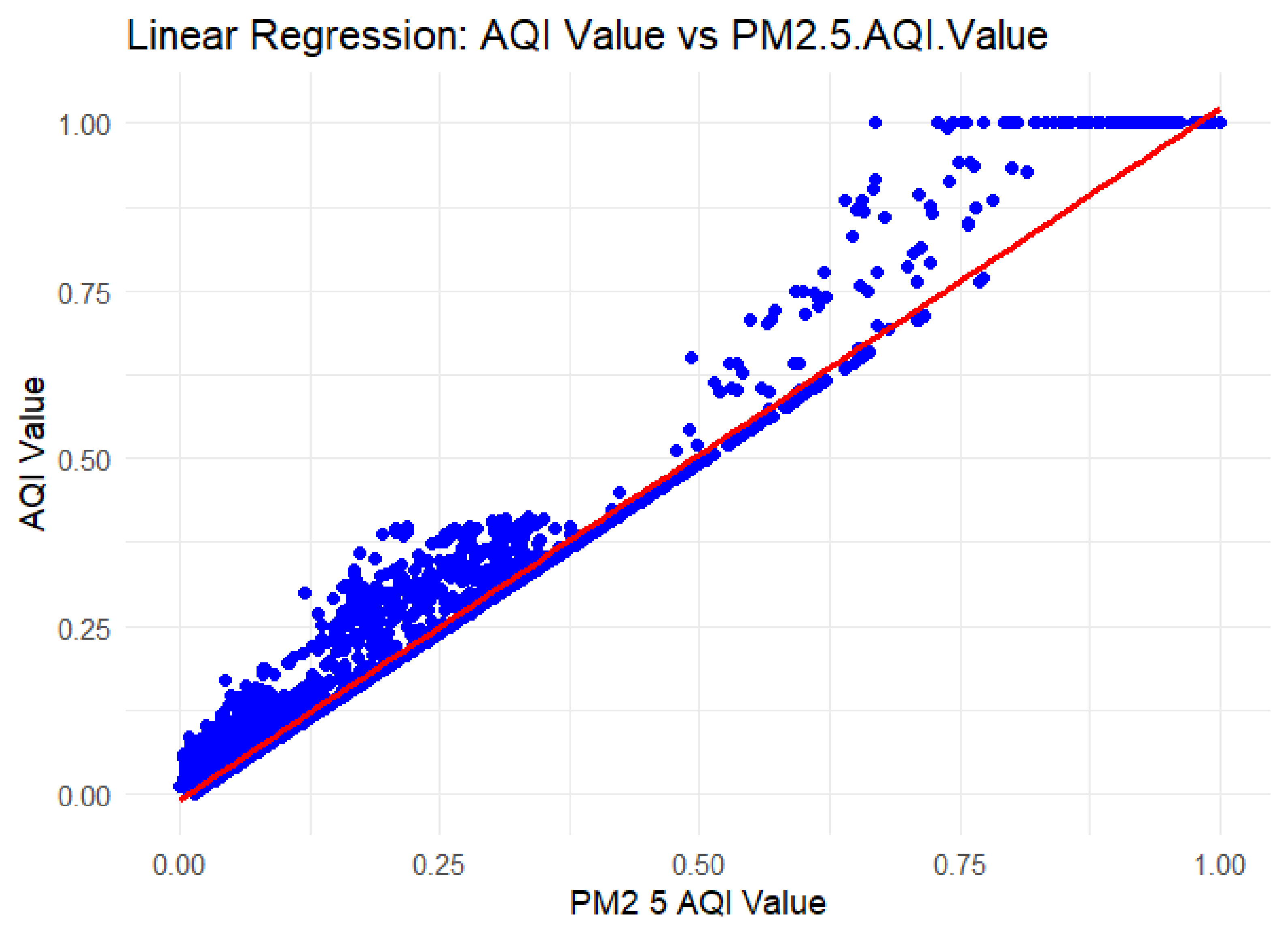
Further Analysis

Simple Linear Regression: AQI vs PM2.5.AQI.Value

```
# Linear Regression: AQI vs PM2.5.AQI.Value
linear_model <- lm(AQI.Value ~ PM2.5.AQI.Value, data = merged_data_cleaned)
```

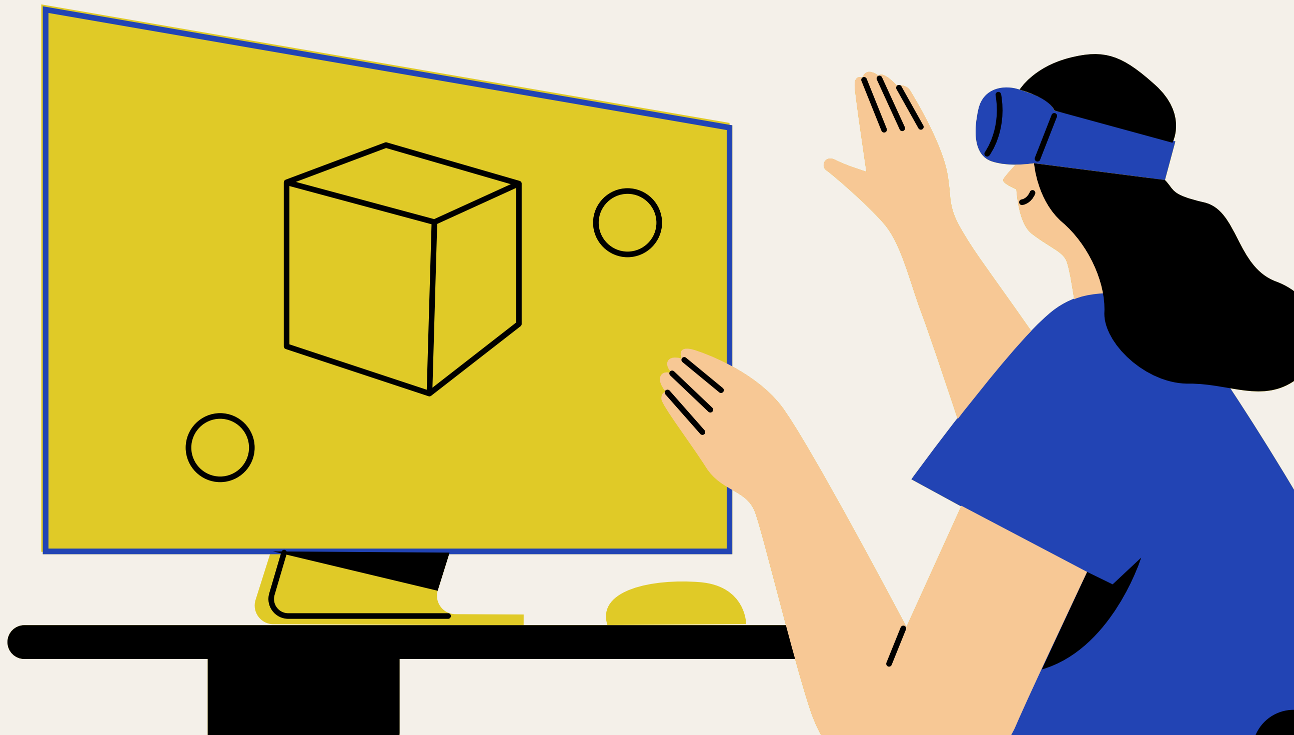
```
Linear Regression Summary:
> summary(linear_model)

Call:
lm(formula = AQI.Value ~ PM2.5.AQI.Value, data = merged_data_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.02165 -0.00754 -0.00667 -0.00576  0.32064 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -0.0093419  0.0002554  -36.58   <2e-16 ***
PM2.5.AQI.Value  1.0309961  0.0013530  762.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02108 on 17193 degrees of freedom
Multiple R-squared:  0.9712,	Adjusted R-squared:  0.9712 
F-statistic: 5.807e+05 on 1 and 17193 DF,  p-value: < 2.2e-16
```

```
> cat("Linear Regression RMSE:", linear_rmse, "\n")
Linear Regression RMSE: 0.02108051
> cat("Linear Regression MAE:", linear_mae, "\n")
Linear Regression MAE: 0.01196818
```



05 - Tableau

Access Tableau



06 - Conclusion

Our analysis reveals that while vehicle density does not directly impact AQI, PM2.5 is a significant contributor to poor air quality. Addressing PM2.5 emissions is vital and can be achieved by promoting renewable energy, adopting electric vehicles, enhancing urban planning with green spaces and efficient public transport, enforcing stricter industrial regulations, and raising public awareness on reducing pollution. These steps collectively pave the way for a healthier and more sustainable environment.