Statistics Report

An Analytical Overview

Prepared By:

- **Sliti Maram**
- **Mojaat Meriem**
- **Charrada Yosr**
- **Amorri Nour**
- **Etteib Tessnim**
- **Bahri Yassmine**

Table Of Contents

1. Introduction

This statistical report leverages advanced analytical techniques to explore relationships and patterns within datasets from diverse domains. Using methods like ANOVA, regression, correlation, and t-tests, the analysis aims to uncover insights that support data-driven decision-making. The study involves a dataset for analyzing gas turbine emissions (CO and NOx). By applying these statistical tools, the report provides actionable insights, supporting predictive modeling and performance evaluation.

2. Business Understanding

Objective:
The primary goal is to identify significant factors influencing outcomes in industrial and environmental settings to optimize processes and predict failures.
Dataset: Gas Turbine Emissions
This dataset, sourced from a turbine in Turkey, investigates CO and NOx emissions under varying operational conditions. Understanding these emissions is critical for improving energy efficiency and compliance with environmental regulations.

Key Question:

- What are the major contributors to emissions in gas turbines?

3. Data Understanding

3.1 Dataset Overview

Source: UCI Machine Learning Repository.
Size: Data collected from 2011 to 2015, which we merged to create a unified dataset for analysis.
Features: 11 features

- Ambient Temperature (AT)
- Ambient Pressure (AP)
- Turbine Inlet Temperatures (TIT)
- Carbon Monoxide (CO)
- Nitrogen Oxide (NOx)
- Turbine Energy Yield (TEY)
- Ambient Humidity (AH)
- Air Filter Differential Pressure (AFDP)
- Gas Turbine Exhaust Pressure (GTEP)
- Turbine After Temperature (TAT)
- Compressor Discharge Pressure (CDP)

Purpose: Predict CO and NOx emissions, which are critical indicators of environmental impact and operational efficiency.

3.2 Data Merged

The merged dataset consolidates data from 2011 to 2015 into a single comprehensive file, enabling a holistic analysis of trends and patterns. This integration facilitates a more robust statistical approach, eliminating the need to handle fragmented yearly datasets while retaining the original data granularity. It simplifies comparisons across years, enhancing the accuracy and reliability of insights derived from the analysis.

Necessary steps were performed, like data exploration, data cleaning, etc., to ensure consistency and remove redundancy across the years.

Size: 36,733 instances with 11 sensor measurements.

Step 1: Data import & Data merged

```
#import data

data2011=read.table(file=file.choose(),header=T,sep=",",dec=".")
data2012=read.table(file=file.choose(),header=T,sep=",",dec=".")
data2013=read.table(file=file.choose(),header=T,sep=",",dec=".")
data2014=read.table(file=file.choose(),header=T,sep=",",dec=".")
data2015=read.table(file=file.choose(),header=T,sep=",",dec=".")
```

```
#merge data
fulldata=rbind(data2011, data2012, data2013, data2014, data2015)
```

The data from 2011 to 2015 was imported by reading each yearly CSV file and consolidating them into a unified dataset called “fulldata.” This process ensured consistency in variable formatting and prepared the data for comprehensive analysis.

Step 2: Data Exploration

- Data types

```
> str(fulldata)
'data.frame':   36733 obs. of  11 variables:
 $ AT   : num  4.59 4.29 3.9 3.74 3.75 ...
 $ AP   : num  1019 1018 1018 1018 1018 ...
 $ AH   : num  83.7 84.2 84.9 85.4 85.2 ...
 $ AFDP: num  3.58 3.57 3.58 3.58 3.58 ...
 $ GTEP: num  24 24 24 23.9 23.9 ...
 $ TIT  : num  1086 1086 1086 1086 1086 ...
 $ TAT  : num  550 550 550 550 550 ...
 $ TEY  : num  135 135 135 135 135 ...
 $ CDP  : num  11.9 11.9 12 12 11.9 ...
 $ CO   : num  0.327 0.448 0.451 0.231 0.267 ...
 $ NOX  : num  82 82.4 83.8 82.5 82 ...
>
```

- The analysis of data types showed that all variables in the dataset are quantitative.This implies that the dataset consists solely of measurable values, which is ideal for applying statistical methods like regression, correlation, and ANOVA to uncover meaningful patterns and relationships.

- Missing Values

```
> sum(is.na(fulldata))        # Total missing values
[1] 0
>
```

- After checking for missing values, we found that there were no missing data points in the dataset. This ensures that the analysis can proceed without the need for imputation or handling of incomplete records, maintaining the integrity and completeness of the data throughout the statistical processes.

- Count of unique values

```
> print(unique_counts_all_columns)
   AT    AP    AH  AFDP  GTEP   TIT   TAT   TEY   CDP    CO   NOX
22523   791 25708 20495 12967   799  2769  6236  4447 26185 23637
>
```

- The count of unique values for each column indicates the level of variability in the dataset. **AT (Ambient Temperature)**, **CO (Carbon Monoxide)**, and **NOX (Nitrogen Oxides)** have the highest counts of unique values, suggesting significant variability in these measurements across the dataset.Columns like **AP (Ambient Pressure)** and **TIT (Turbine Inlet Temperature)** have relatively fewer unique values, indicating that these measurements are more consistent or clustered compared to others.This variability in unique values helps highlight which variables exhibit more diverse data and which are more stable, providing insights into the range and spread of the data.

-Duplicate Rows

```
> duplicate_count <- sum(duplicated(fulldata))
> print(paste("Number of duplicate rows:", duplicate_count))
[1] "Number of duplicate rows: 7"
>
```

- The dataset contains **7 duplicate rows** out of a total of **36,733 instances** with 11 sensor measurements. This means that the duplicates represent a very small fraction of the total dataset (less than 0.02%).

- Outliers

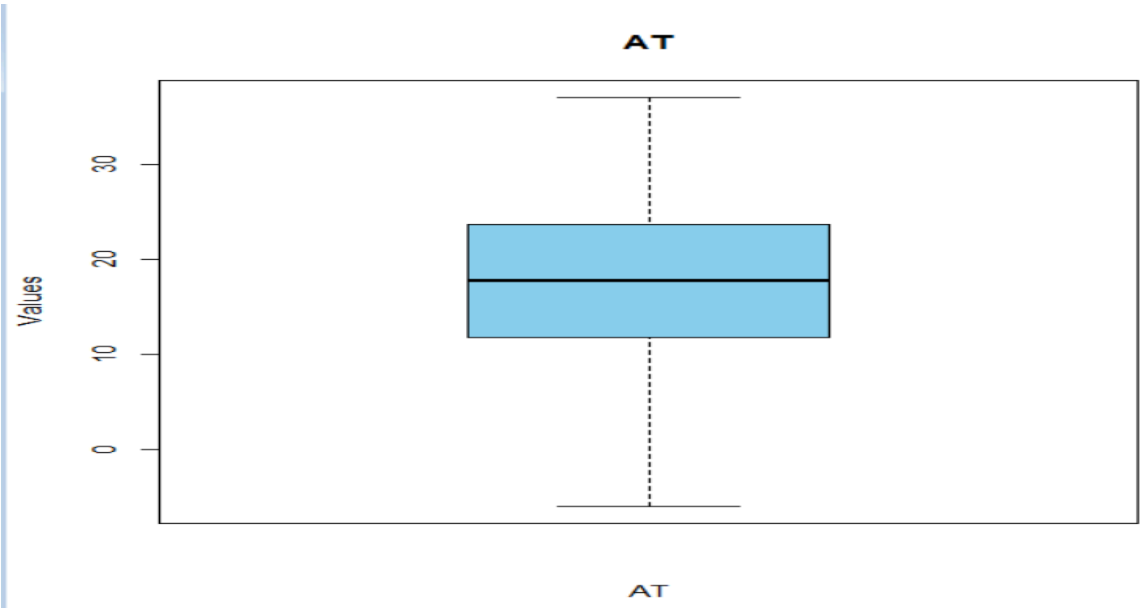


Figure 1: This boxplot for the AT variable shows a symmetric distribution with no clear outliers. The data appears to be concentrated within a relatively narrow range.

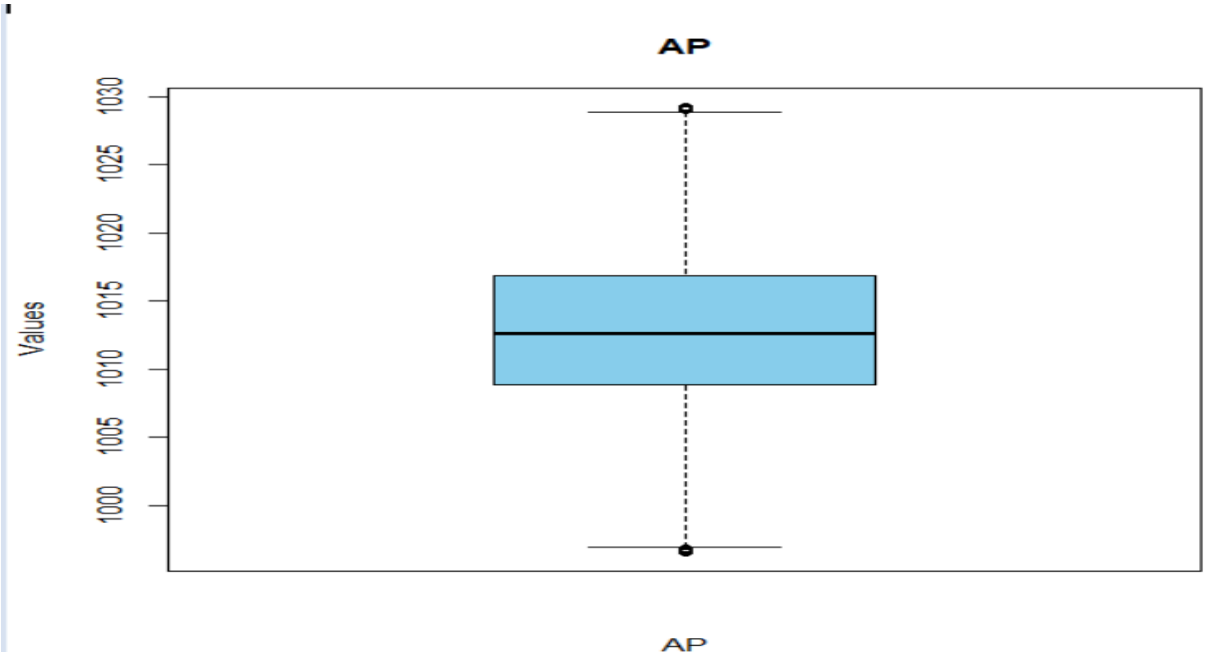


Figure 2: The AP boxplot shows a similar pattern, with some outlier values that are considerably higher than the other data points and some outlier values that are considerably lower than the other data points.

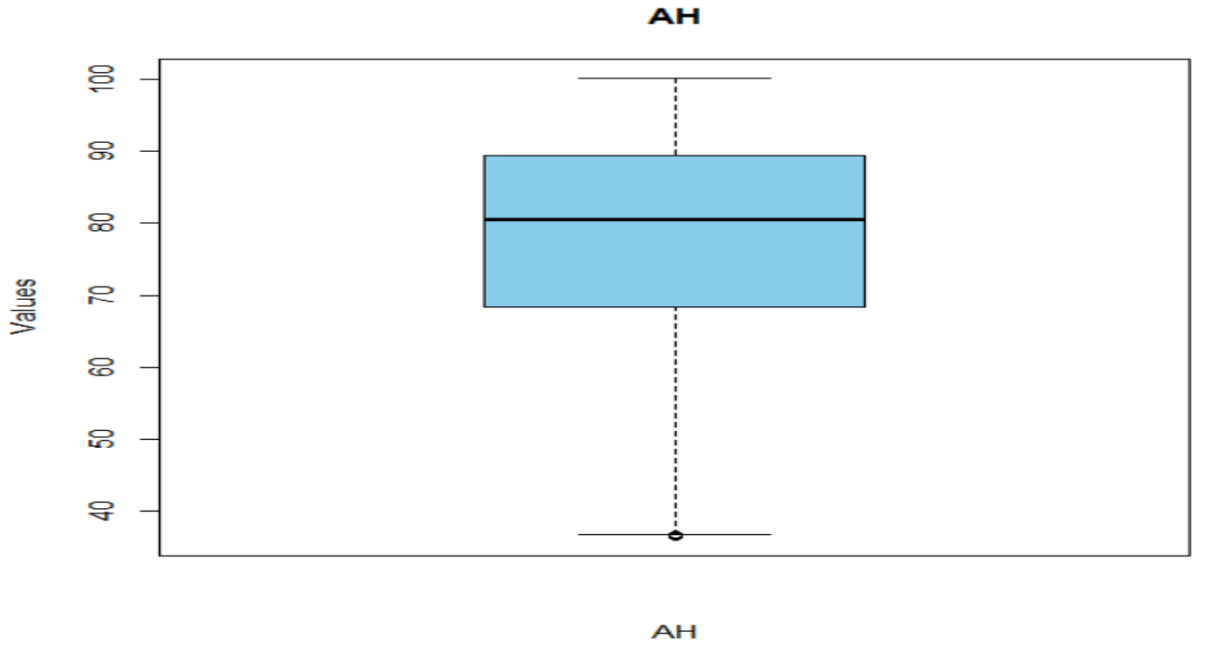


Figure 3: The AH boxplot reveals only one outlier that is considerably lower than the main group of data points.

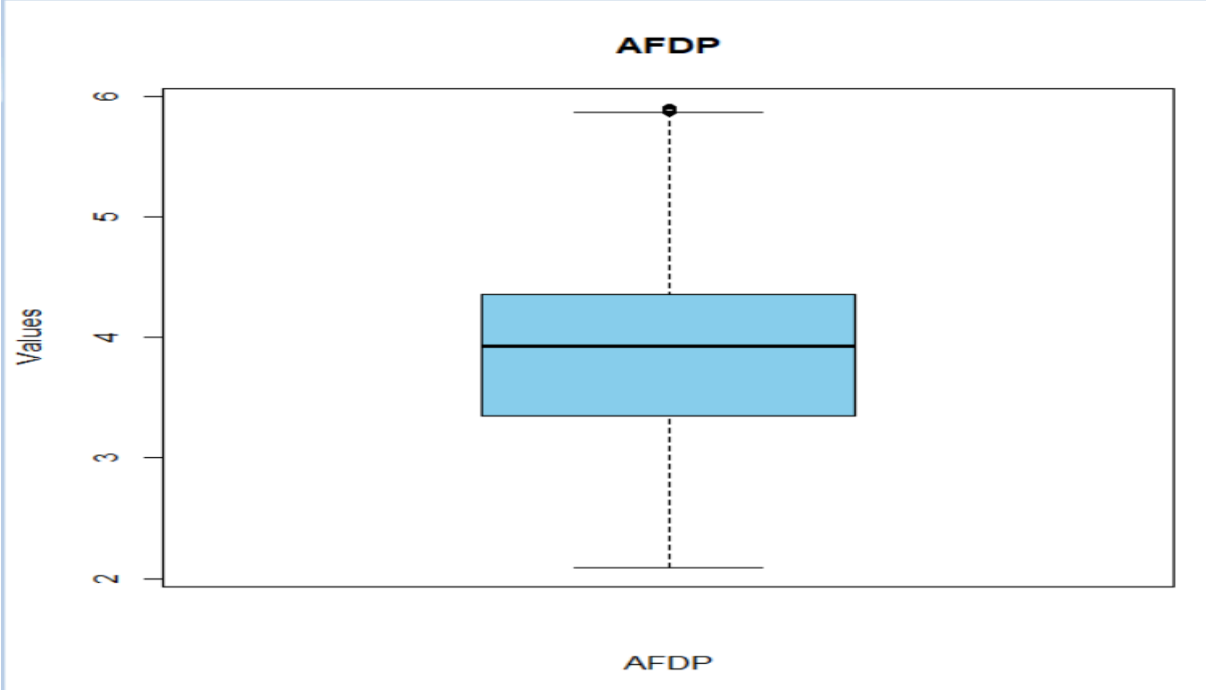


Figure 4: The image shows a boxplot for the AFDP variable. The plot indicates that there is one clear outlier value that is significantly higher than the rest of the data points.

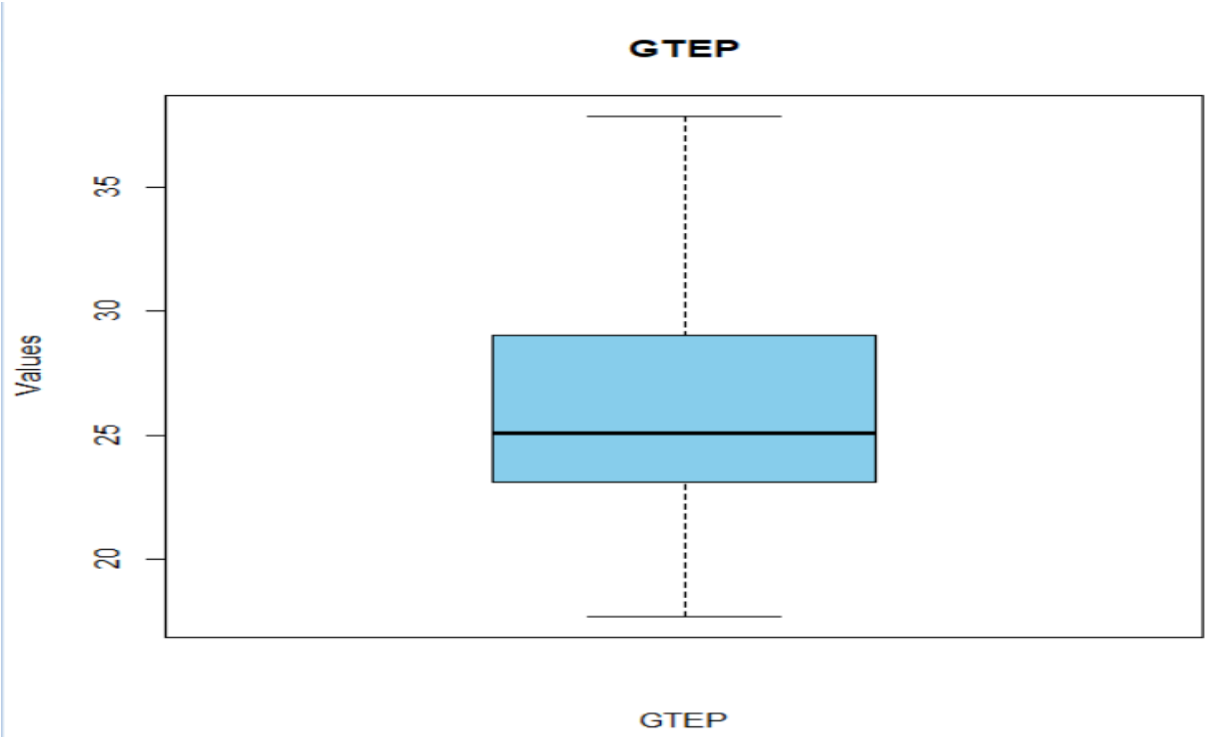


Figure 5: The GTEP boxplot shows no clear outlier value that is significantly higher or lower than the rest of the data points.

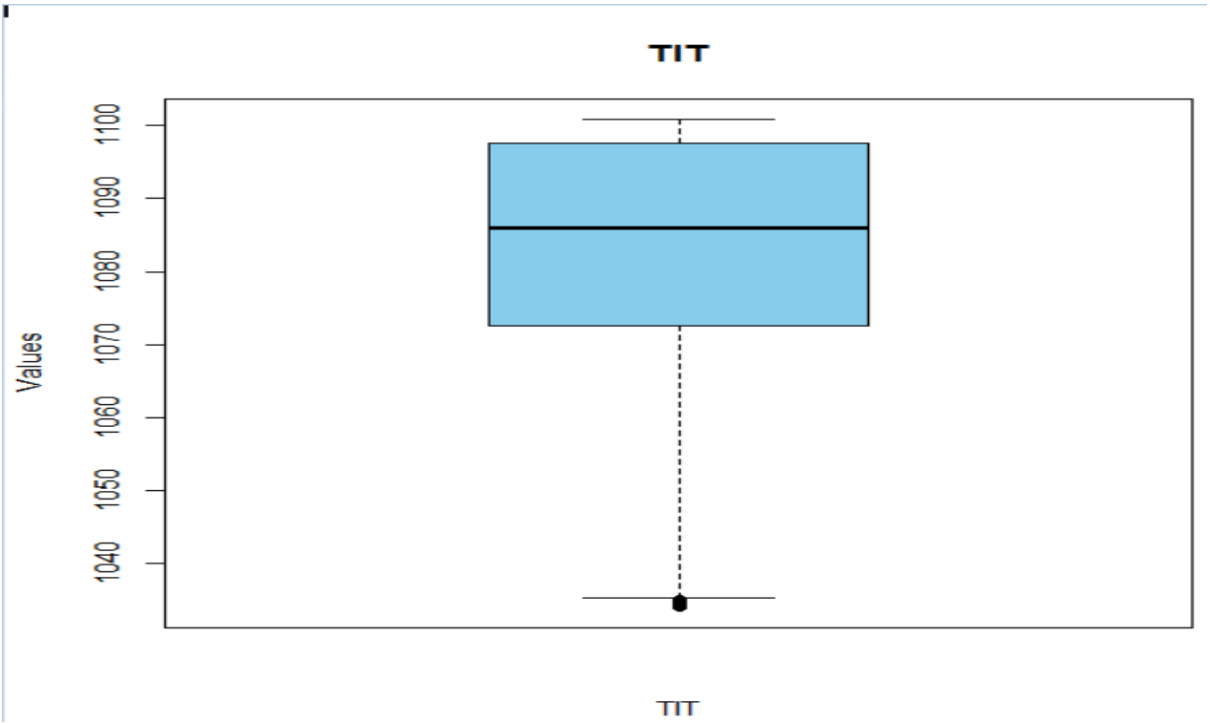


Figure 6: The TIT boxplot reveals some outliers that are considerably lower than the main group of data points.

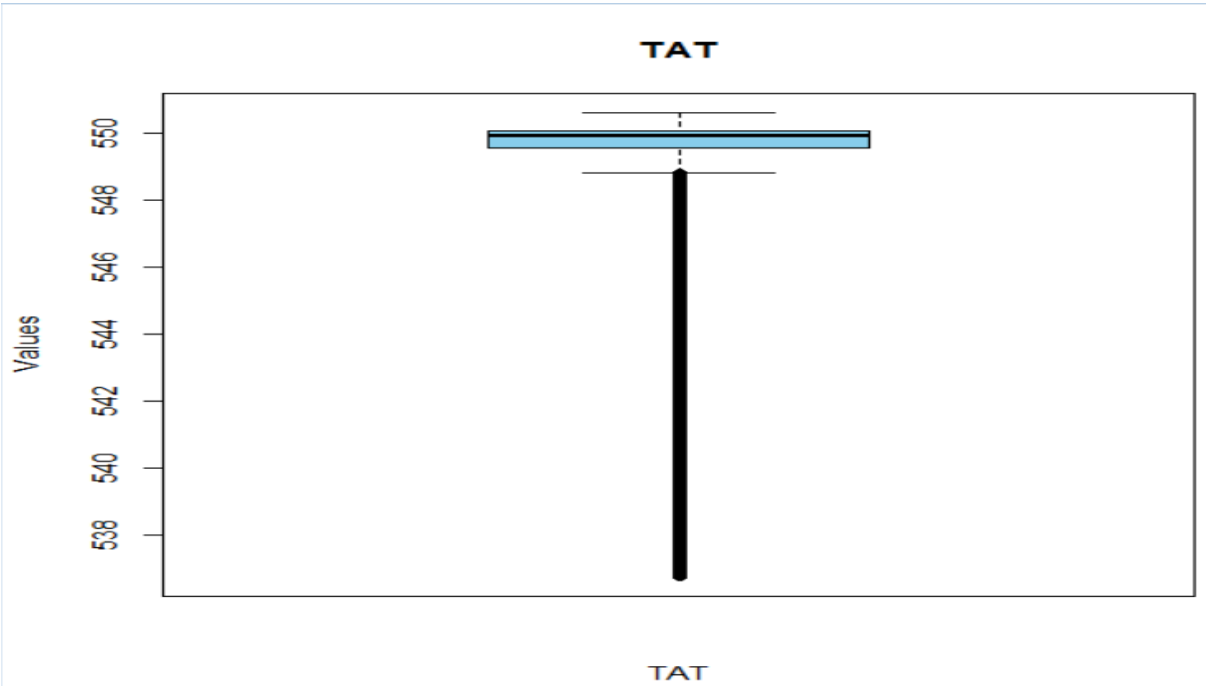


Figure 7: The TAT boxplot reveals multiple outliers that are much lower than the rest of the data.The percentage of outliers is higher than the data.

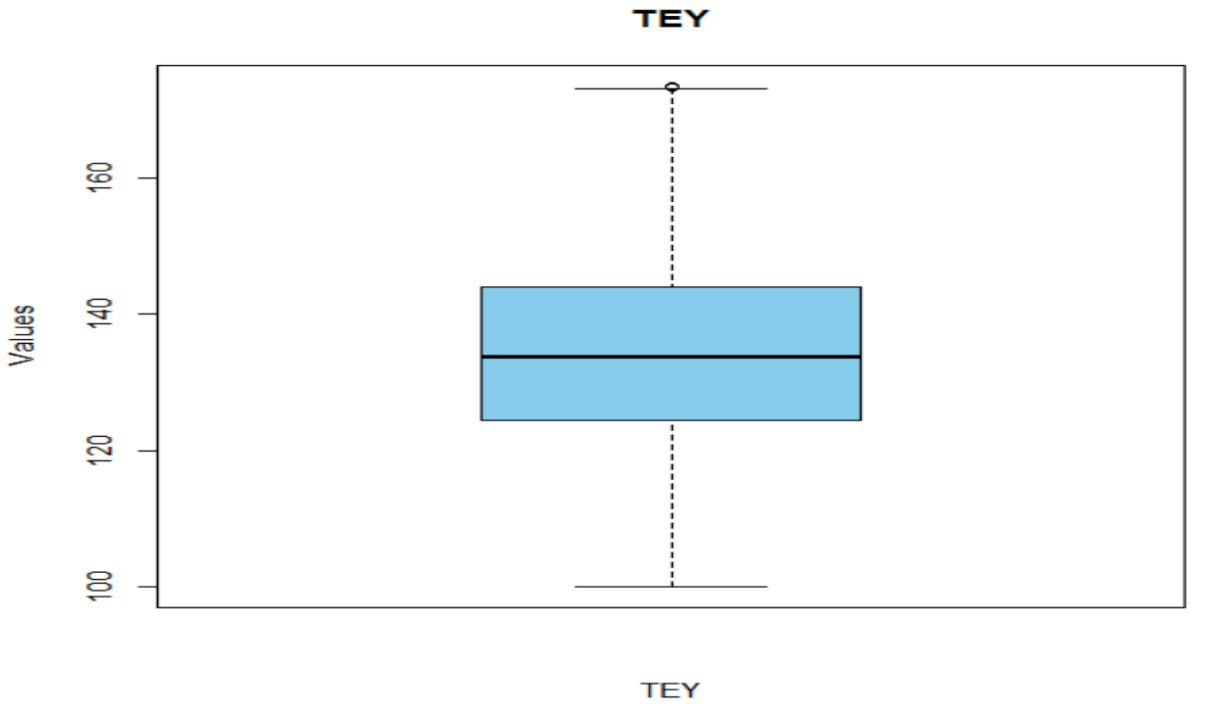


Figure 8: For the TEY variable, the boxplot indicates the presence of one outlier value that is noticeably higher than the rest of the data points. This outlier stands out as being quite different from the main cluster of TEY values.

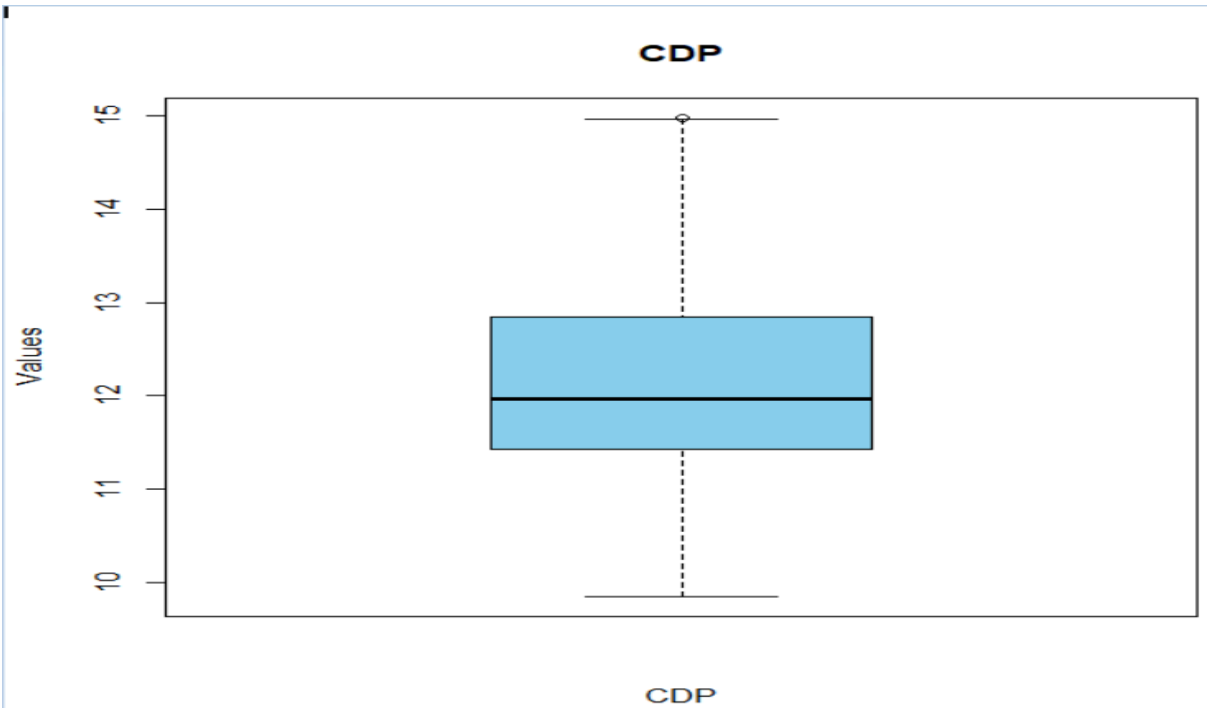


Figure 9: The CDP boxplot displays only one outlier that is considerably higher than the rest of the data points.

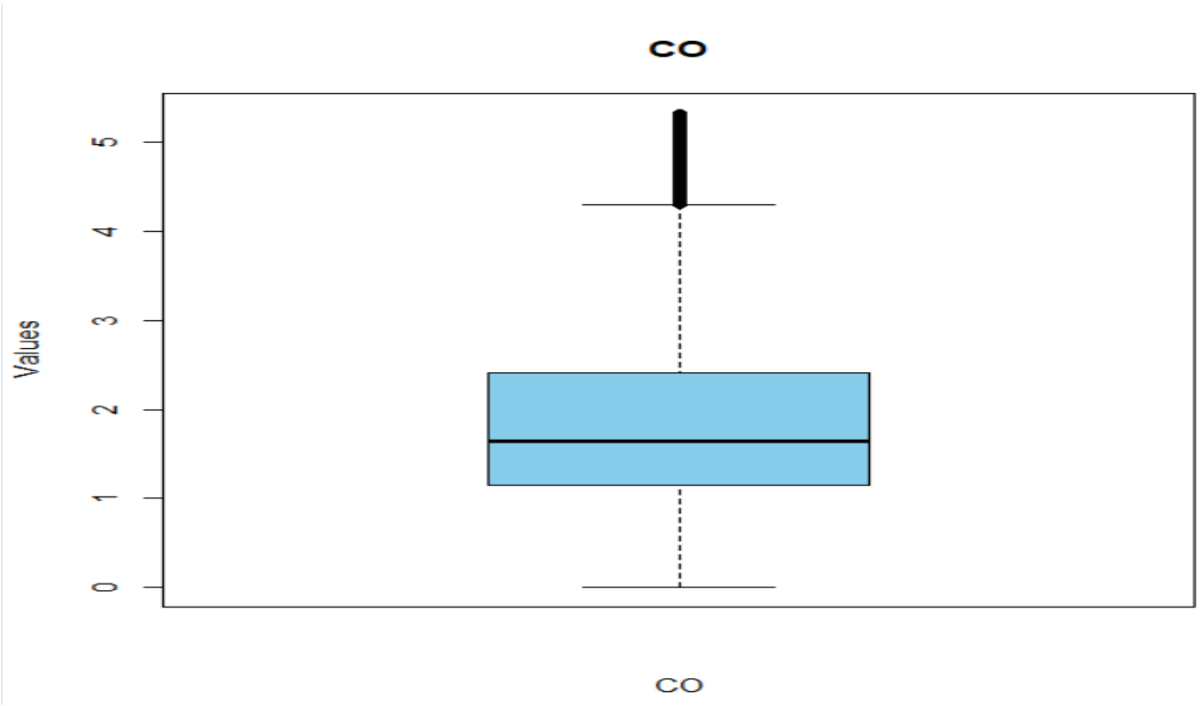


Figure 10: The CO boxplot shows outliers that are substantially higher than the other data points.

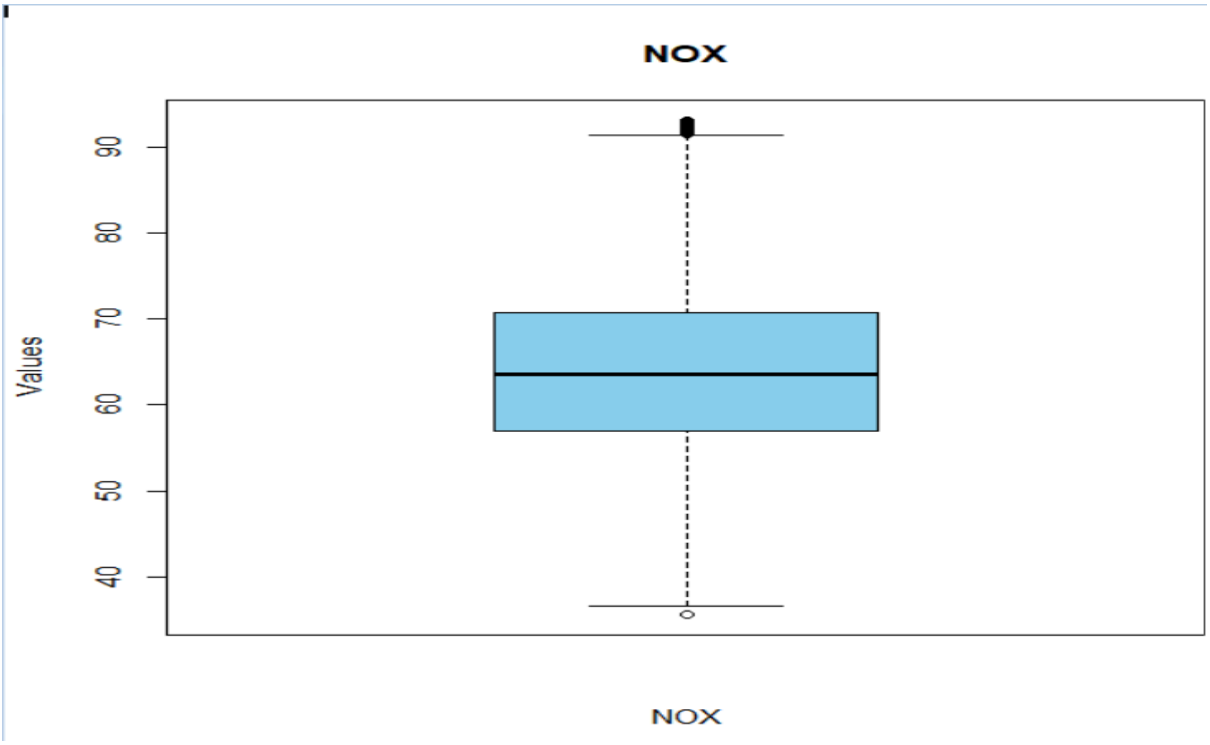


Figure 11: The NOX boxplot indicates the presence of two outlier values—one that is quite high and another that is relatively low compared to the main group of data.

- The boxplots reveal a mix of symmetric distributions and notable outliers across several variables. While **AT** and **GTEP** show stable distributions with no significant outliers, variables like **AP**, **AH**, **AFDP**, **TIT**, **TAT**, **TEY**, **CO**, and **NOX** have outliers, with some values much higher or lower than the main data cluster. **TAT** has a higher percentage of outliers. These outliers may indicate rare events or errors and should be further investigated for their impact on the analysis.

-Correlation Table

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1	-0.378042951584244	-0.525620056676072	0.436409093855256	0.295748467323948	0.263205773231759	-0.155406133809274	0.0789134546189298	0.255514651596901	-0.162421573654029	-0.608095992914672
AP	-0.378042951584244	1	0.00724676560552747	-0.144977416911883	-0.0973543103663175	-0.0461398625458531	0.0366630670781794	0.00139641956220428	-0.0427592564088693	0.0413219745622171	0.230819849445985
AH	-0.525620056676072	0.00724676560552747	1	-0.172888458400922	-0.324936818240692	-0.255648295229442	0.179794038608391	-0.188136377504028	-0.278328077872772	0.10097427849433	0.124632086187408
AFDP	0.436409093855256	-0.144977416911883	-0.172888458400922	1	0.593760171508362	0.642916582482503	-0.410555439141225	0.561846357010944	0.629295629632322	-0.448741570767774	-0.0979757554347423
GTEP	0.295748467323948	-0.0973543103663175	-0.324936818240692	0.593760171508362	1	0.883589842254337	-0.662437806667143	0.935238520035373	0.963297285793129	-0.576011159194319	-0.126494657260114
TIT	0.263205773231759	-0.0461398625458531	-0.255648295229442	0.642916582482503	0.883589842254337	1	-0.448987381952725	0.938547654261918	0.943741696322779	-0.729859332010346	0.0669942490418065
TAT	-0.155406133809274	0.0366630670781794	0.179794038608391	-0.410555439141225	-0.662437806667143	-0.448987381952725	1	-0.610357387195859	-0.656667733671014	0.212218895769592	0.136354204333411
TEY	0.0789134546189298	0.00139641956220428	-0.188136377504028	0.561846357010944	0.935238520035373	0.938547654261918	-0.610357387195859	1	0.97838065788942 9	-0.661106593430792	0.0637445735272839
CDP	0.255514651596901	-0.0427592564088693	-0.278328077872772	0.629295629632322	0.963297285793129	0.943741696322779	-0.656667733671014	0.97838065788942 9	1	-0.649758846980672	-0.0611409592082919
CO	-0.162421573654029	0.0413219745622171	0.10097427849433	-0.448741570767774	-0.576011159194319	-0.729859332010346	0.212218895769592	-0.661106593430792	-0.649758846980672	1	-0.0208679253827068
NOX	-0.608095992914672	0.230819849445985	0.124632086187408	-0.0979757554347423	-0.126494657260114	0.0669942490418065	0.136354204333411	0.0637445735272839	-0.0611409592082919	-0.0208679253827068	1

- The diagonal elements (where the row and column headers match) all show a value of 1.0, which indicates a perfect positive correlation between each variable and itself; this is expected.
- Looking at the off-diagonal elements, we can see some strong positive correlations, such as between TEY and CDP (0.978). These suggest these pairs of variables tend to move together in a linear fashion.
- There are also some strong negative correlations, such as between TIT and CO (-0.72). These indicate an inverse relationship between those variable pairs.

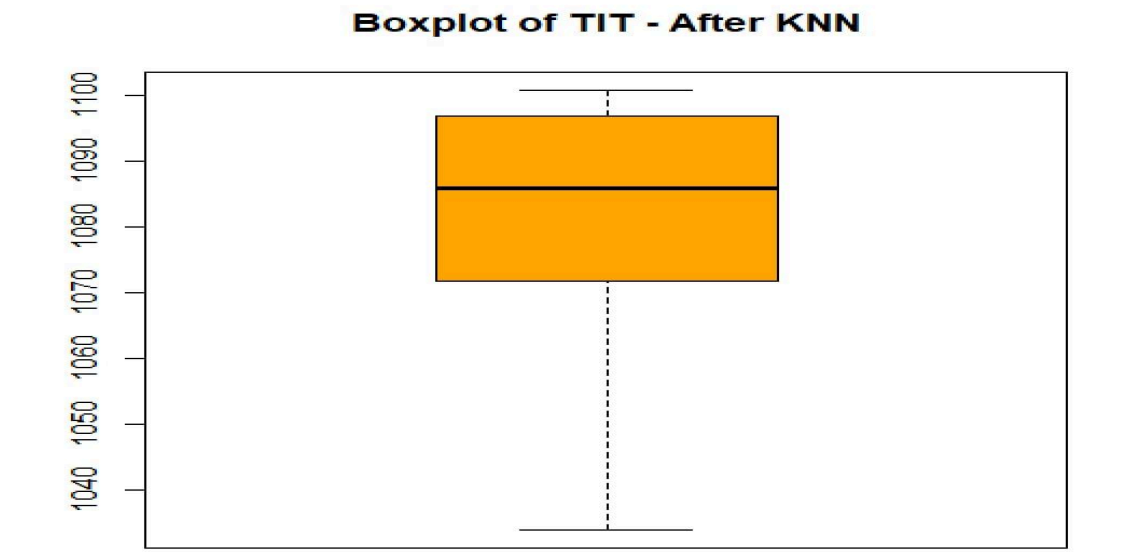
4. Data Cleaning

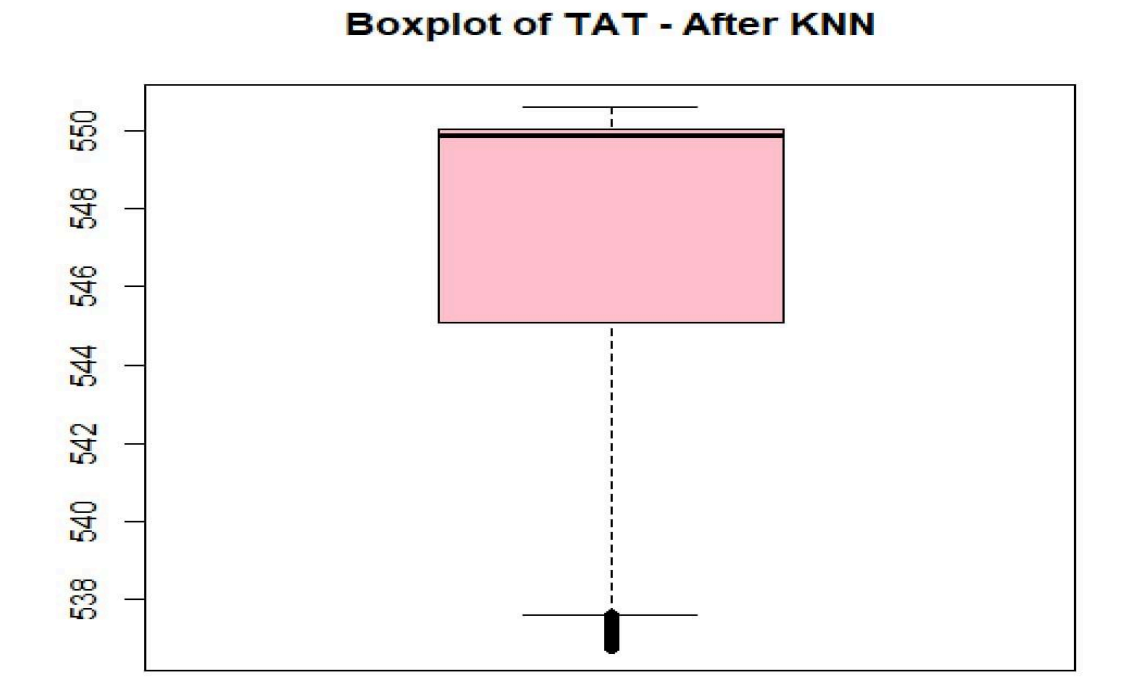
In the data cleaning phase, we focused on ensuring the dataset's quality by addressing potential issues such as missing values, duplicates, and outliers. This step is essential for preparing the data for further analysis and ensuring that the results are reliable and accurate.

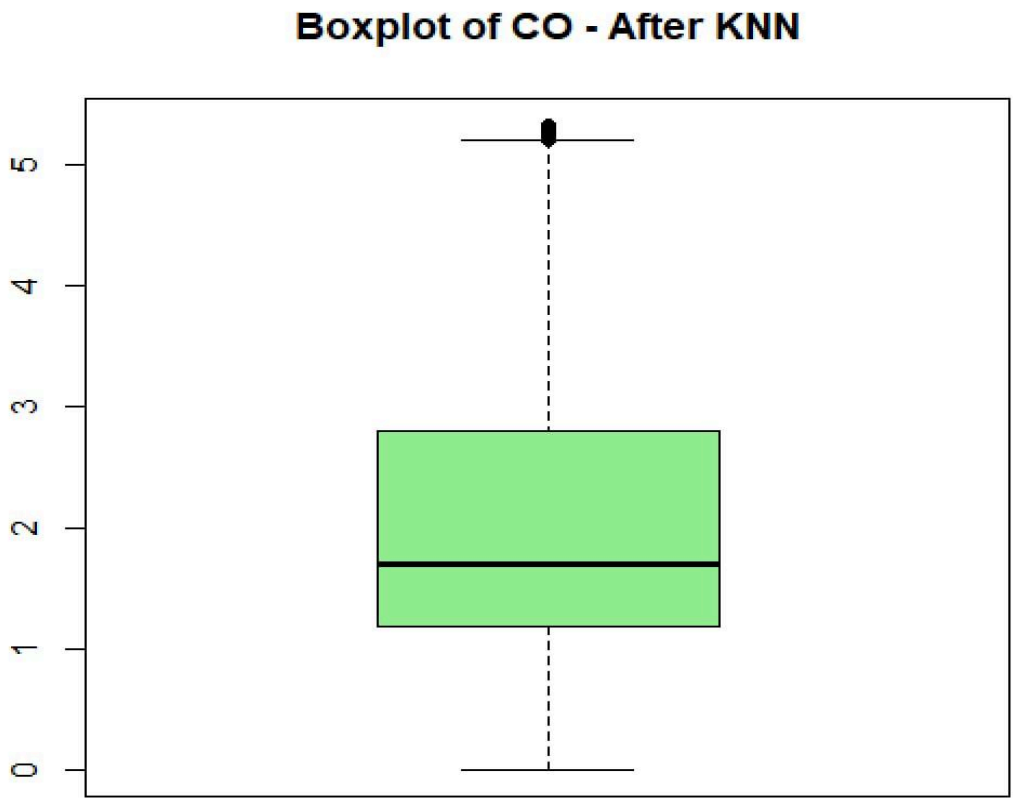
During the data understanding process, we found no missing values in the dataset. However, we identified **7 duplicate rows**, which is a small number compared to the overall size of the dataset (36,733 instances and less than 0.02%). Since we are working with quantitative values only, we decided to keep these duplicates for now as they represent a minimal portion of the data. Additionally, we observed several outliers across different variables. To handle these outliers, we plan to use **K-Nearest Neighbors (KNN)** and

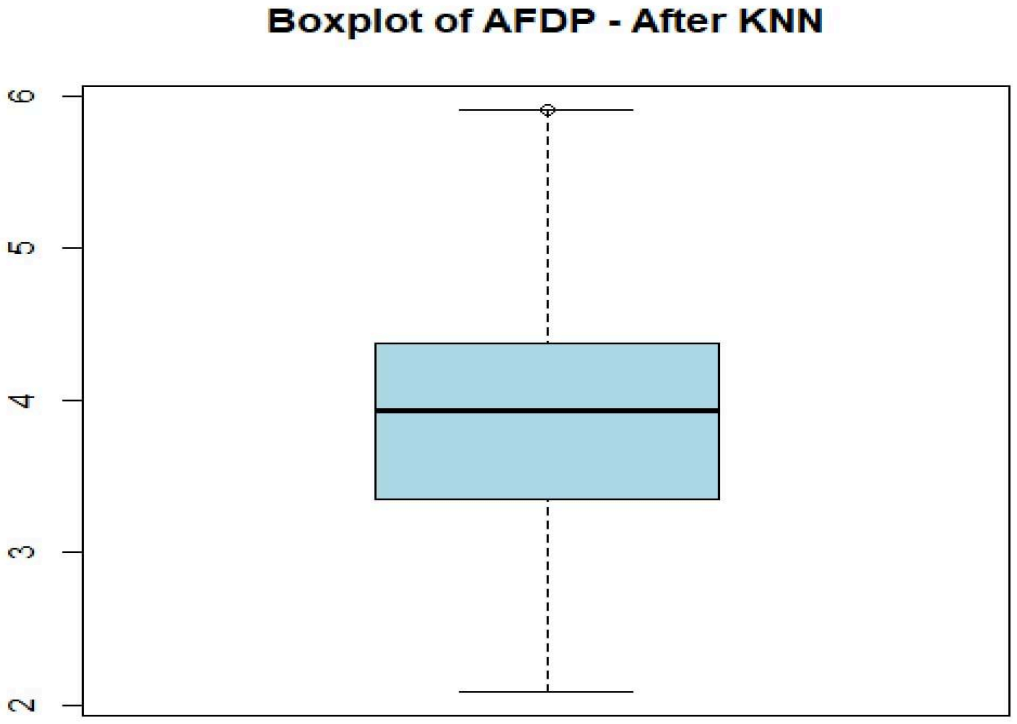
median imputation to treat them, ensuring a more consistent dataset for analysis. After applying both techniques, we will evaluate and choose the best method based on their effectiveness.

-Treat outliers using K-NN

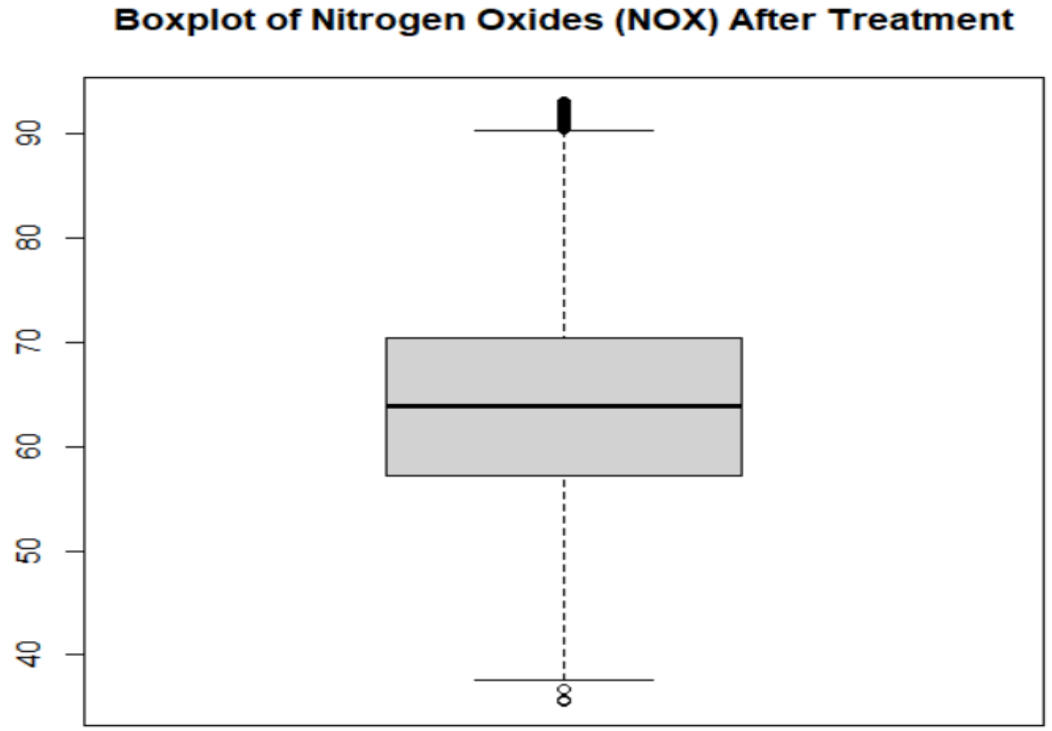


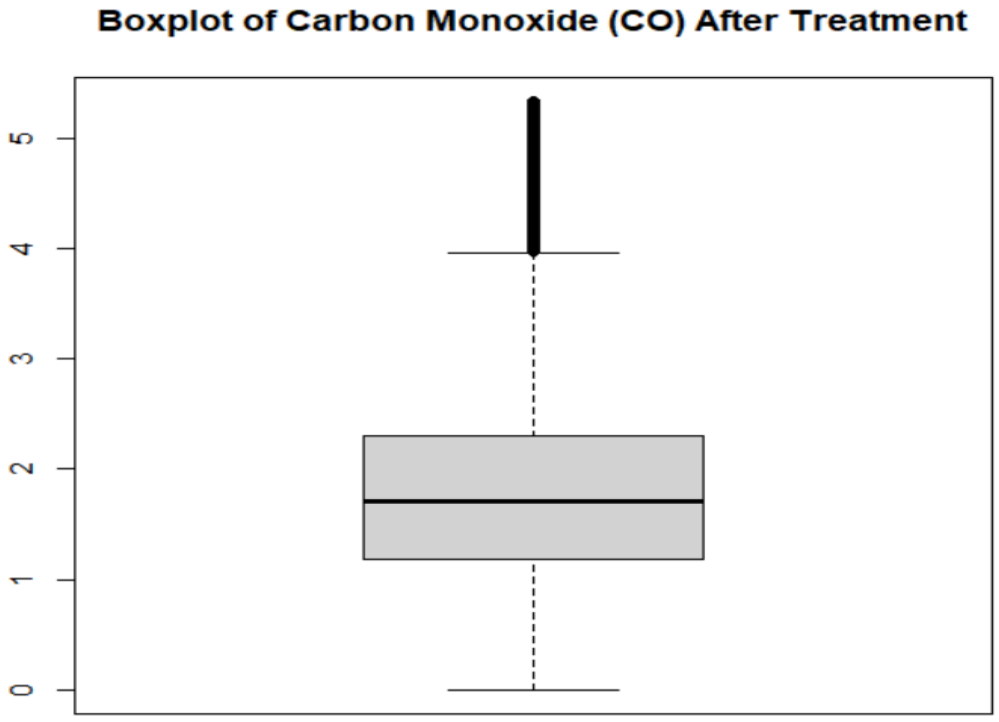


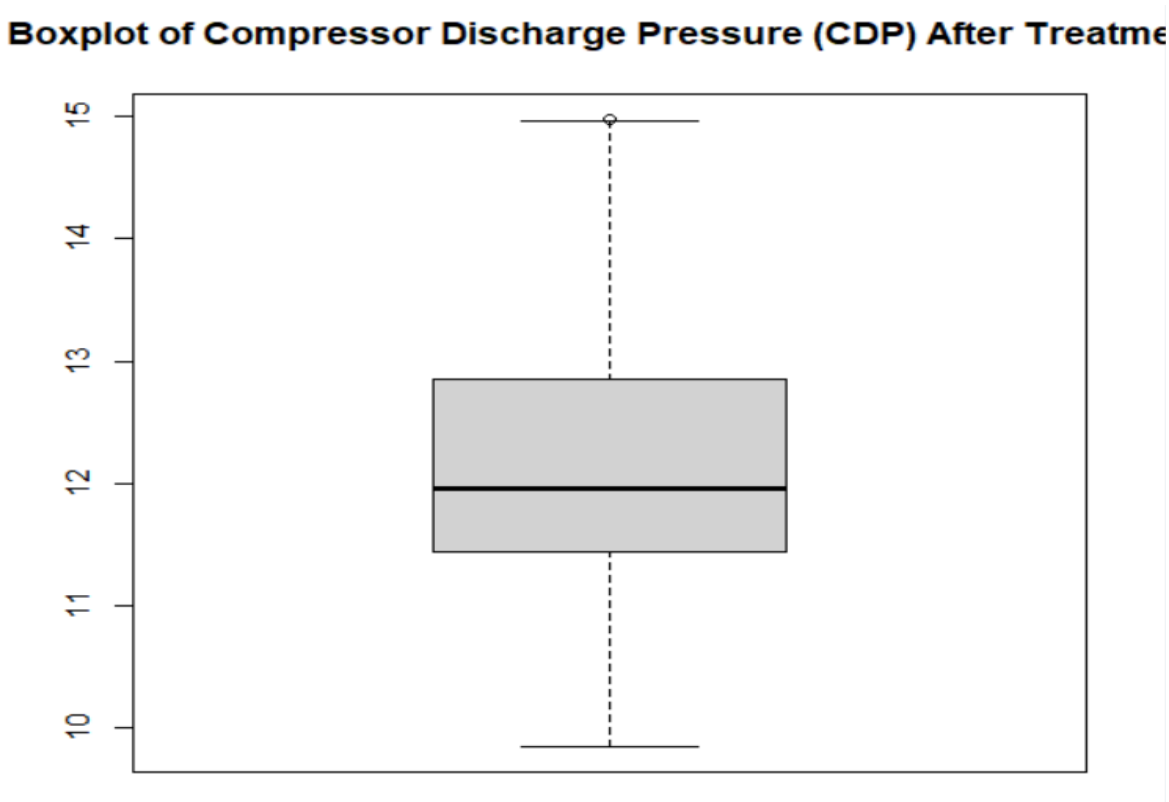

Boxplot of CO - After KNN
0
1
2
3
4
5

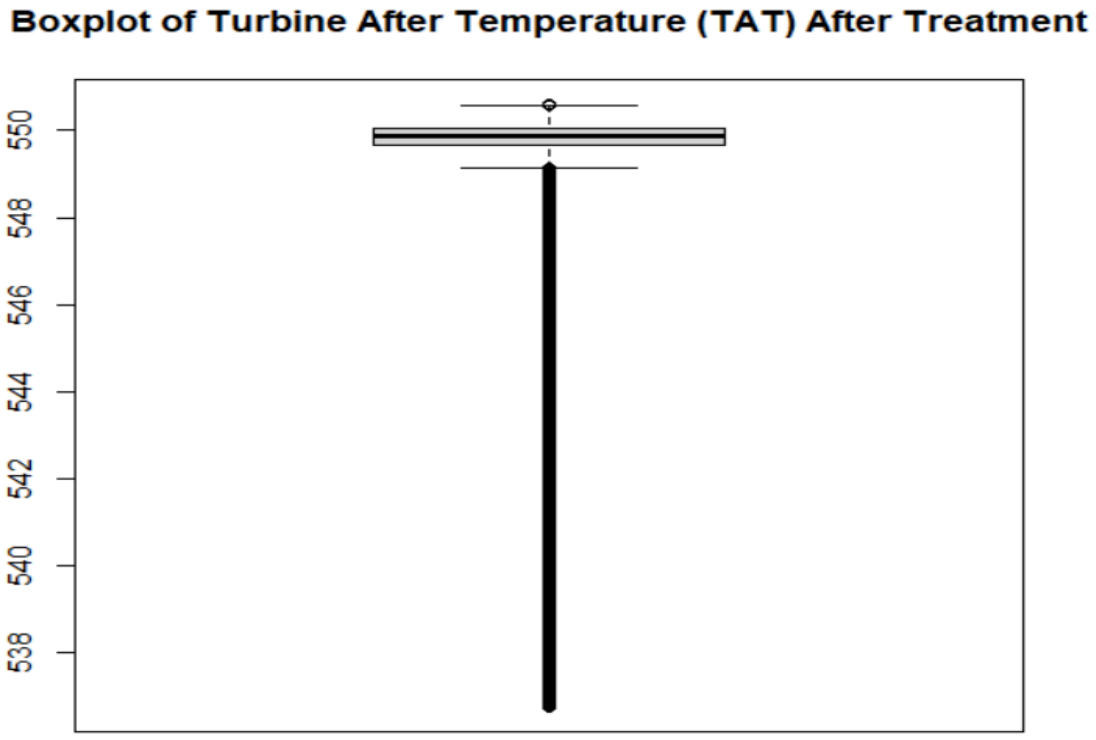

Boxplot of AFDP - After KNN
2
3
4
5
6

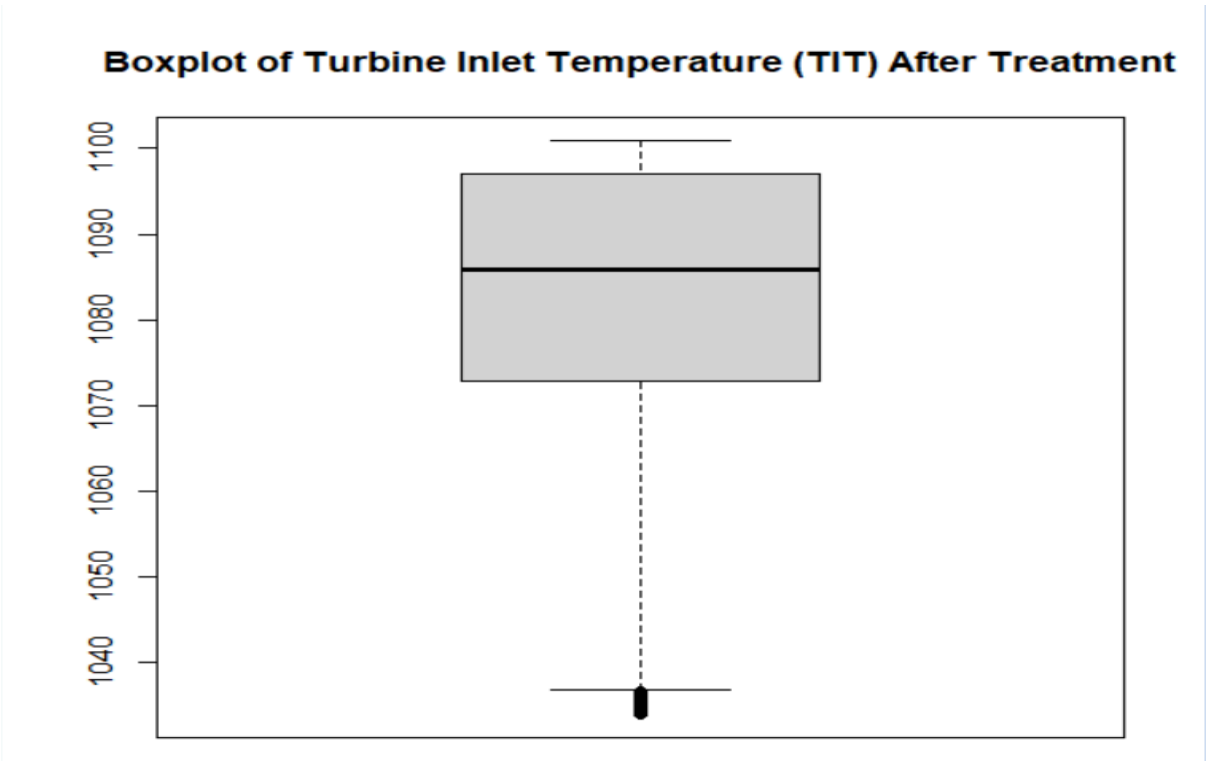
- Treat outliers using median imputation

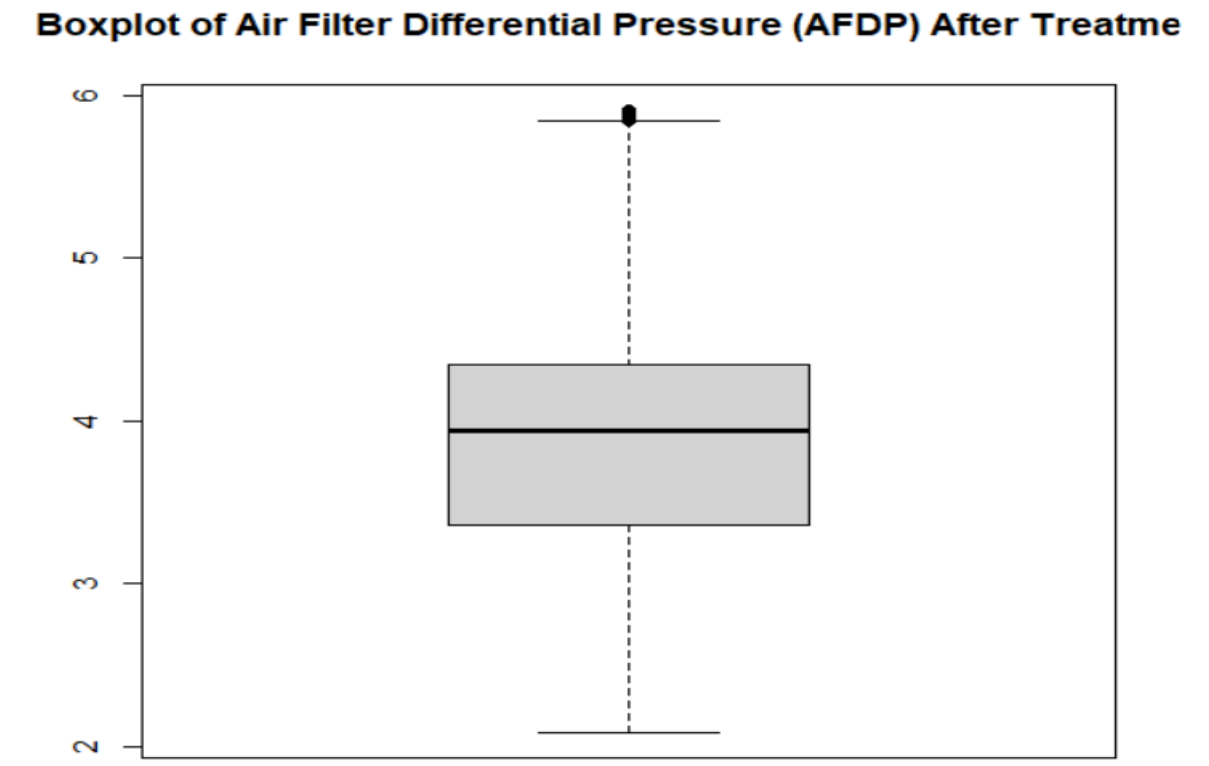


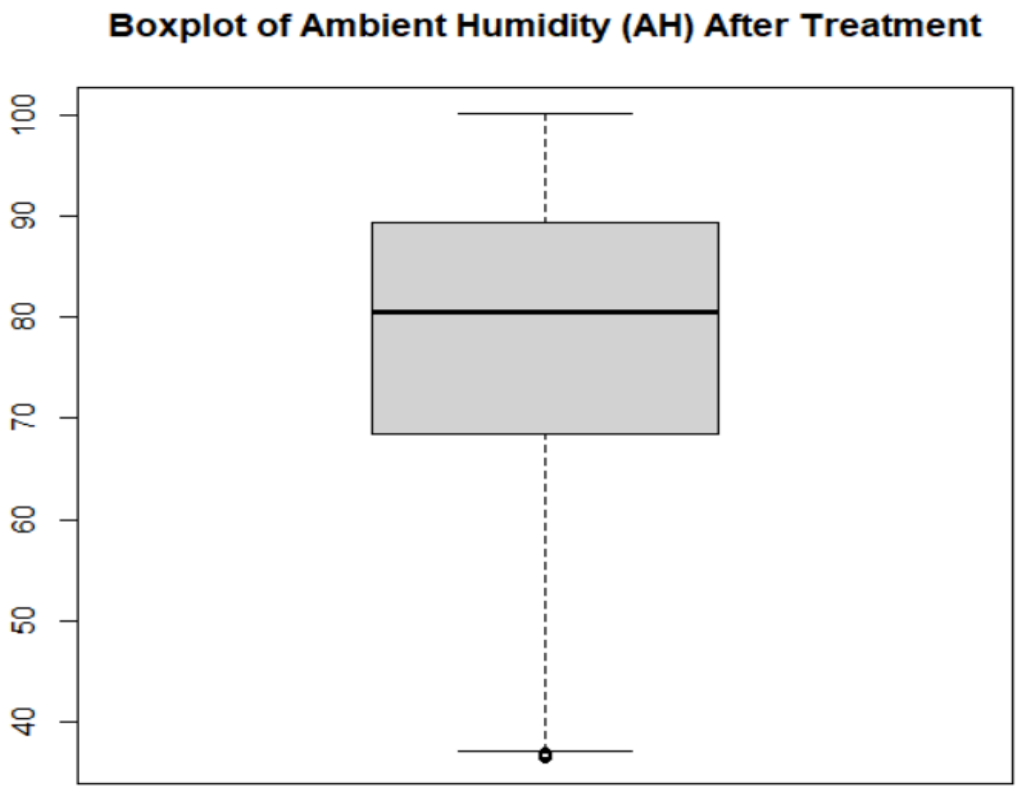


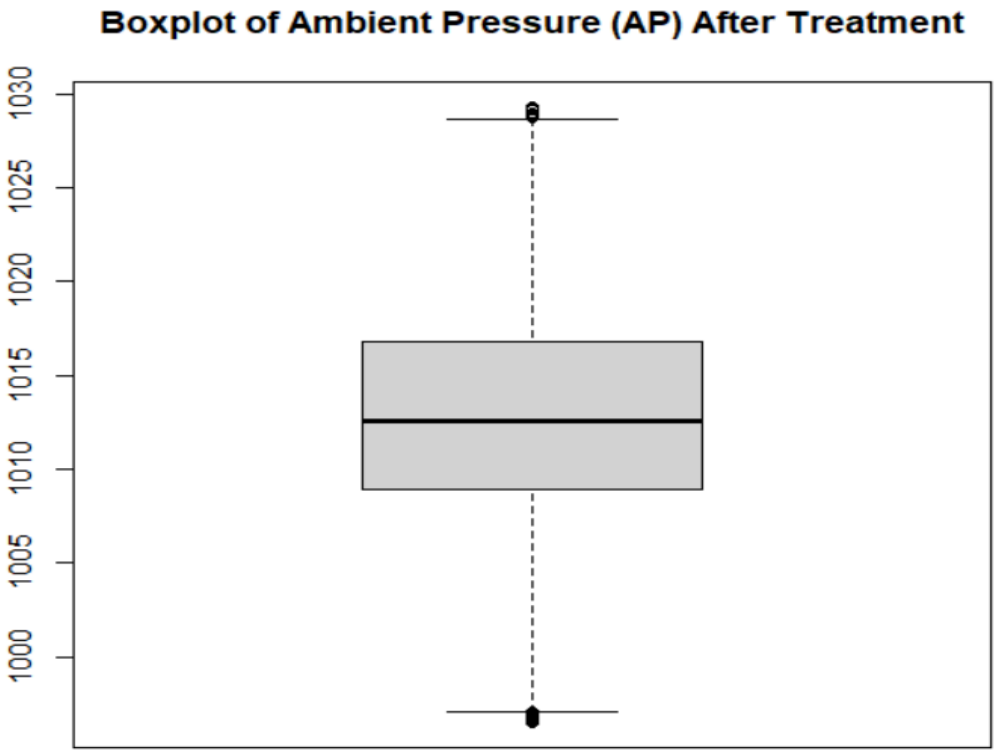

Boxplot of Compressor Discharge Pressure (CDP) After Treatme
15
14
13
12
11
10

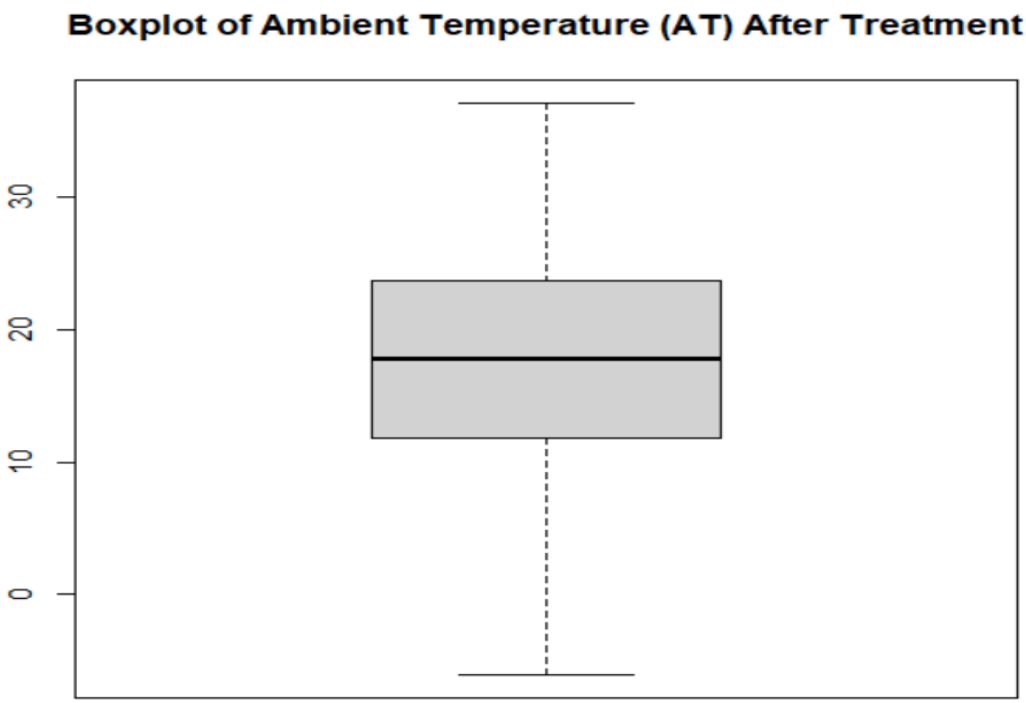

Boxplot of Turbine After Temperature (TAT) After Treatment
550
548
546
544
542
540
538


Boxplot of Turbine Inlet Temperature (TIT) After Treatment
1040
1050
1060
1070
1080
1090
1100


Boxplot of Air Filter Differential Pressure (AFDP) After Treatme
2
3
4
5
6


Boxplot of Ambient Humidity (AH) After Treatment
100
90
80
70
60
50
40


Boxplot of Ambient Pressure (AP) After Treatment
1030
1025
1020
1015
1010
1005
1000



- ❖ After handling the outliers using both K-Nearest Neighbors (KNN) and median imputation, we found that KNN provided better results. The KNN method more effectively maintained the integrity of the data by considering the relationships between nearby data points, while median imputation, though useful, did not account for the data's underlying structure as well. Therefore, we concluded that KNN is the more suitable approach for treating the outliers in this dataset.

5. Scaling

Scaling is an important preprocessing step to ensure that all variables in the dataset contribute equally to the analysis. In this case, we applied normalization to the data, which scales the quantitative values to a consistent range, typically between 0 and 1. This step helps prevent any variable with a larger range from disproportionately influencing the statistical models, ensuring more accurate and reliable results in subsequent analyses.

-Normalization

```
> data <- read.table(file = file.choose(), header = TRUE, sep = ",")
> # Normalize the numeric columns (Min-Max scaling)
> normalize <- function(x) {
+   return ((x - min(x)) / (max(x) - min(x)))
+ }
>
> # Apply normalization to each numeric column
> normalized_data <- as.data.frame(lapply(data, function(x) if(is.numeric(x))
>
> # Check the summary of the normalized data
> summary(normalized_data)
```

```
> data <- read.table(file = file.choose(), header = TRUE, sep = ",")
> # Normalize the numeric columns (Min-Max scaling)
> normalize <- function(x) {
+   return ((x - min(x)) / (max(x) - min(x)))
+ }
>
> # Apply normalization to each numeric column
> normalized_data <- as.data.frame(lapply(data, function(x) if(is.numeric(x)) n$
>
> # Check the summary of the normalized data
> summary(normalized_data)
       AT                AP                AH                AFDP
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.4157   1st Qu.:0.3777   1st Qu.:0.4992   1st Qu.:0.3318
 Median :0.5546   Median :0.4905   Median :0.6906   Median :0.4840
 Mean   :0.5526   Mean   :0.5036   Mean   :0.6514   Mean   :0.4754
 3rd Qu.:0.6899   3rd Qu.:0.6217   3rd Qu.:0.8303   3rd Qu.:0.5984
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      GTEP              TIT               TAT               TEY
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.2691   1st Qu.:0.5650   1st Qu.:0.6003   1st Qu.:0.3326
 Median :0.3670   Median :0.7758   Median :0.9473   Median :0.4590
 Mean   :0.3897   Mean   :0.7115   Mean   :0.7562   Mean   :0.4557
 3rd Qu.:0.5631   3rd Qu.:0.9417   3rd Qu.:0.9589   3rd Qu.:0.5999
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      CDP               CO               NOX
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.3090   1st Qu.:0.2212   1st Qu.:0.3741
 Median :0.4124   Median :0.3203   Median :0.4899
 Mean   :0.4310   Mean   :0.3916   Mean   :0.5034
 3rd Qu.:0.5861   3rd Qu.:0.5236   3rd Qu.:0.6189
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
> # Check the minimum and maximum values of each column
> mins <- sapply(normalized_data, min)
```

```
> # Check the minimum and maximum values of each column
> mins <- sapply(normalized_data, min)
> maxs <- sapply(normalized_data, max)
>
> cat("Column Minimums (Should be 0):\n", round(mins, 3), "\n")
Column Minimums (Should be 0):
 0 0 0 0 0 0 0 0 0 0 0
> cat("Column Maximums (Should be 1):\n", round(maxs, 3), "\n")
Column Maximums (Should be 1):
 1 1 1 1 1 1 1 1 1 1 1
```

- The normalization was successfully applied, with the data now scaled between 0 and 1 for all variables. The summary statistics show a balanced distribution with most variables having means and medians near the middle of the range, indicating relatively symmetrical data. This normalization prepares the data for further statistical analysis and modeling.

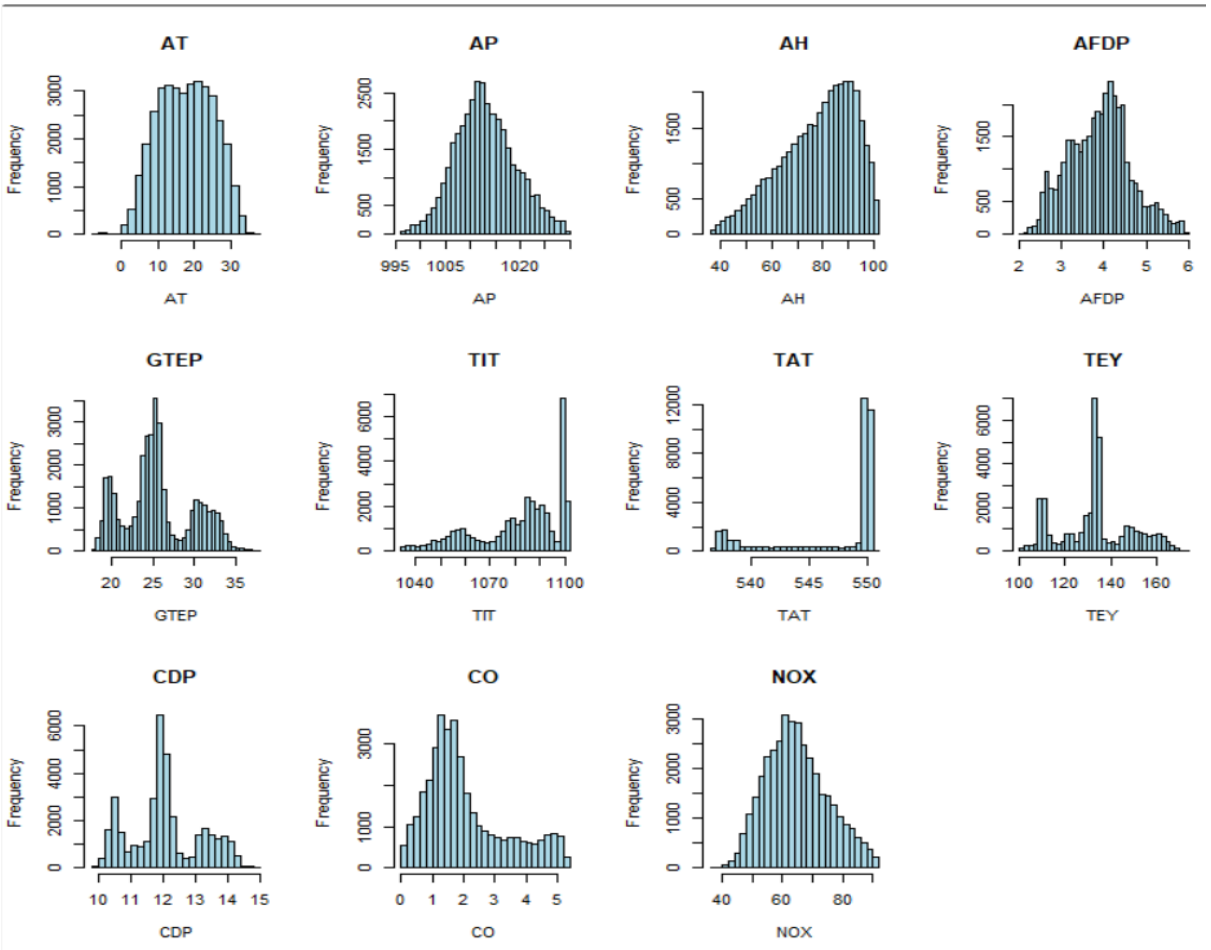
6. Test Normality

Testing normality is an essential step in data analysis to assess whether the data follows a normal distribution. In this process, we used **kurtosis**, **skewness**, and **histograms** to evaluate the distribution of the variables. Kurtosis helps measure the tail heaviness, skewness indicates the asymmetry of the distribution, and histograms provide a visual representation, allowing us to determine if any transformations are needed for accurate analysis.

- Histogram displayed

```
# Set plotting layout to display multiple histograms
par(mfrow = c(3, 4)) # Adjust to fit all variables in multiple rows and columns

# Generate histograms for each variable in fulldata_knn_cleaned
for (col in colnames(fulldata_knn_cleaned)) {
  hist(fulldata_knn_cleaned[[col]], main = col, col = "lightblue",
xlab = col, breaks = 30)
}
```



```
> print(summary_stats)
   Feature     Skewness Kurtosis
1       AT -0.04354494 2.173349
2       AP  0.19303636 2.813087
3       AH -0.57914652 2.577764
4     AFDP  0.12759644 2.689074
5     GTEP  0.32532700 2.334110
6      TIT -0.78275173 2.541754
7      TAT -1.25431930 2.870321
8      TEY  0.10589464 2.479998
9      CDP  0.23350257 2.361297
10      CO  0.85879792 2.752433
11     NOX  0.32172461 2.619673
> 
```

- ❖ Based on the histograms and the results, we observed that AT, AP, and NOX follow approximately normal distributions, with AP and NOX showing symmetry and a kurtosis close to 3. However, variables like AH, GTEP, TIT, TAT, AFDP, TEY, CDP, and CO exhibit non-normal distributions, with skewness and kurtosis values. indicating either right or left skew and slightly lower than normal kurtosis. These findings suggest that while some variables can be analyzed with standard parametric methods, others may require transformation or special handling due to their non-normal distributions.

7. Bivariate Analysis

Bivariate analysis involves examining the relationship between two variables to understand how they are associated. In this analysis, we used the **t-test** to compare the means of two groups and the **variance test (-test)** to assess if the variances between two groups are significantly different. These tests help identify whether there are any significant differences between the groups and provide insights into the strength and nature of their relationship.

- Variance Homogeneity Test

```
# Create two separate data frames
normal_data <- normalized_data[, normal_vars]
non_normal_data <- normalized_data[, non_normal_vars]

#two by two testing
var.test(normal_data$NOX, normal_data$AT)
var.test(normal_data$NOX, normal_data$AP)
var.test(normal_data$AT, normal_data$AP)
```

```
> var.test(normal_data$NOX, normal_data$AT)

        F test to compare two variances

data:  normal_data$NOX and normal_data$AT
F = 1.1722, num df = 36732, denom df = 36732, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.148512 1.196468
sample estimates:
ratio of variances
          1.172245
```

```
> var.test(normal_data$NOX, normal_data$AP)

        F test to compare two variances

data:  normal_data$NOX and normal_data$AP
F = 1.0137, num df = 36732, denom df = 36732, p-value = 0.1909
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9932192 1.0346908
sample estimates:
ratio of variances
          1.013743
```

```
> var.test(normal_data$AT, normal_data$AP)

        F test to compare two variances

data:  normal_data$AT and normal_data$AP
F = 0.86479, num df = 36732, denom df = 36732, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8472794 0.8826573
sample estimates:
ratio of variances
         0.8647874
```

- The results from the variance tests (F-tests) comparing the variances of different pairs of variables are as follows:

1. **NOX vs. AT**: The F-test yielded an F-statistic of 1.1722 with a very small p-value (< 2.2e-16), indicating a significant difference between the variances of NOX and AT. The 95% confidence interval for the ratio of variances is between 1.1485 and 1.1965, suggesting that the variance of NOX is slightly higher than that of AT.
2. **NOX vs. AP**: The F-test returned an F-statistic of 1.0137 with a p-value of 0.1909, which is greater than the common significance level of 0.05. This suggests there is no significant difference between the variances of NOX and AP, as the null hypothesis (that the variances are equal) cannot be rejected. The ratio of variances is 1.0137, indicating very similar variances.

3. **AT vs. AP**: The F-test produced an F-statistic of 0.8648 with a very small p-value (< 2.2e-16), indicating a significant difference between the variances of AT and AP. The 95% confidence interval for the ratio of variances is between 0.8428 and 0.8827, suggesting that the variance of AT is smaller than that of AP.

Conclusion:

- NOX and AT have significantly different variances.
- NOX and AP do not have significantly different variances.
- AT and AP have significantly different variances.

- Independent t-test for normally distributed data

```
        Welch Two Sample t-test

data:  normalized_data$NOX and normalized_data$AT
t = -38.44, df = 73443, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05166603 -0.04665292
sample estimates:
mean of x mean of y
0.5034188 0.5525783
```

```
 # Perform Welch's t-test between NOX and AT
> t_test_nox_at <- t.test(normalized_data$NOX, normalized_data$AT, var.equal = FALSE)
> cat("Welch's t-test between NOX and AT:\n")
Welch's t-test between NOX and AT:
> print(t_test_nox_at)


> # Perform Welch's t-test between NOX and AP
> t_test_nox_ap <- t.test(normalized_data$NOX, normalized_data$AP, var.equal = FALSE)
> cat("\nWelch's t-test between NOX and AP:\n")

Welch's t-test between NOX and AP:
> print(t_test_nox_ap)
```

```
        Welch Two Sample t-test

data:  normalized_data$NOX and normalized_data$AP
t = -0.11548, df = 73270, p-value = 0.9081
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.002748124  0.002442313
sample estimates:
mean of x mean of y
0.5034188 0.5035717
```

```
> # Perform Welch's t-test between AT and AP
> t_test_at_ap <- t.test(normalized_data$AT, normalized_data$AP, var.equal = FALSE)
> cat("\nWelch's t-test between AT and AP:\n")

Welch's t-test between AT and AP:
> print(t_test_at_ap)
```

```
        Welch Two Sample t-test

data:  normalized_data$AT and normalized_data$AP
t = 37.306, df = 73124, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04643185 0.05158129
sample estimates:
mean of x mean of y
0.5525783 0.5035717
```

- The results from the Welch Two Sample t-test are as follows:

1. NOX vs. AT:
    - The t-test statistic is -38.44, with a p-value < 2.2e-16, which is highly significant, indicating that the means of NOX and AT are significantly different.
    - The 95% confidence interval for the difference in means is [-0.0517, -0.0467], confirming that the mean of NOX is significantly lower than that of AT.
    - Mean of NOX: 0.5034, Mean of AT: 0.5526.
2. NOX vs. AP:
    - The t-test statistic is -0.1155, with a p-value of 0.9081, which is not significant. This suggests there is no significant difference between the means of NOX and AP.
    - The 95% confidence interval for the difference in means is [-0.00275, 0.00244], which includes 0, further supporting the conclusion that the means are equal.
    - Mean of NOX: 0.5034, Mean of AP: 0.5036.
3. AT vs. AP:

- The t-test statistic is 37.306, with a p-value < 2.2e-16, which is highly significant, indicating a significant difference between the means of AT and AP.
- The 95% confidence interval for the difference in means is [0.0464, 0.0516], showing that AT has a significantly higher mean than AP.
- Mean of AT: 0.5526, Mean of AP: 0.5036.

Conclusion:

- NOX and AT: There is a significant difference in means. AT has a higher mean (0.5526) than NOX (0.5034). p-value = < 2.2e-16, indicating strong evidence against the null hypothesis.
- NOX and AP: There is no significant difference in means. The means of NOX (0.5034) and AP (0.5036) are nearly identical. p-value = 0.9081, failing to reject the null hypothesis.
- AT and AP: There is a significant difference in means. AT has a higher mean (0.5526) than AP (0.5036). p-value = < 2.2e-16, indicating strong evidence against the null hypothesis.

- Kruskal-Wallis Test

TIT groups:

```
> kruskal.test(AH ~ TIT)    # Compare AH across TIT groups

        Kruskal-Wallis rank sum test

data:  AH by TIT
Kruskal-Wallis chi-squared = 5723.6, df = 984, p-value < 2.2e-16

> kruskal.test(AFDP ~ TIT) # Compare AFDP across TIT groups

        Kruskal-Wallis rank sum test

data:  AFDP by TIT
Kruskal-Wallis chi-squared = 21100, df = 984, p-value < 2.2e-16

> kruskal.test(GTEP ~ TIT) # Compare GTEP across TIT groups

        Kruskal-Wallis rank sum test

data:  GTEP by TIT
Kruskal-Wallis chi-squared = 33064, df = 984, p-value < 2.2e-16
```

```
> kruskal.test(TAT ~ TIT)  # Compare TAT across TIT groups

        Kruskal-Wallis rank sum test

data:  TAT by TIT
Kruskal-Wallis chi-squared = 21573, df = 984, p-value < 2.2e-16

> kruskal.test(TEY ~ TIT)  # Compare TEY across TIT groups

        Kruskal-Wallis rank sum test

data:  TEY by TIT
Kruskal-Wallis chi-squared = 32957, df = 984, p-value < 2.2e-16

> kruskal.test(CDP ~ TIT)  # Compare CDP across TIT groups

        Kruskal-Wallis rank sum test

data:  CDP by TIT
Kruskal-Wallis chi-squared = 34743, df = 984, p-value < 2.2e-16
```

- Significant differences are observed in GTEP, TAT, TEY, and CDP across TIT groups, with p-values < 0.05.

Conclusion:

For each Kruskal-Wallis test result, since the p-value is very small, we reject H0 in favor of H1, indicating a significant difference between the distributions of the variables.

AFDP groups:

```
> kruskal.test(CDP ~ AFDP)  # Compare CDP across AFDP groups

        Kruskal-Wallis rank sum test

data:  CDP by AFDP
Kruskal-Wallis chi-squared = 28488, df = 20536, p-value < 2.2e-16
```

```
> kruskal.test(GTEP ~ AFDP) # Compare GTEP across AFDP groups

        Kruskal-Wallis rank sum test

data:  GTEP by AFDP
Kruskal-Wallis chi-squared = 27803, df = 20536, p-value < 2.2e-16

> kruskal.test(TAT ~ AFDP)  # Compare TAT across AFDP groups

        Kruskal-Wallis rank sum test

data:  TAT by AFDP
Kruskal-Wallis chi-squared = 23125, df = 20536, p-value < 2.2e-16

> kruskal.test(TEY ~ AFDP)  # Compare TEY across AFDP groups

        Kruskal-Wallis rank sum test

data:  TEY by AFDP
Kruskal-Wallis chi-squared = 27035, df = 20536, p-value < 2.2e-16
```

- ❖ Significant differences are observed in GTEP, TAT, TEY, and CDP across AFDP groups, with p-values < 0.05.

Conclusion:

For each Kruskal-Wallis test result, since the p-value is very small, we reject H0 in favor of H1 indicating a significant difference between the distributions of the variables (GTEP, TAT, TEY, and CDP) across the AFDP groups.

TAT groups:

```
> kruskal.test(TEY ~ TAT)

        Kruskal-Wallis rank sum test

data:  TEY by TAT
Kruskal-Wallis chi-squared = 18438, df = 6337, p-value < 2.2e-16

> kruskal.test(CDP ~ TAT)

        Kruskal-Wallis rank sum test

data:  CDP by TAT
Kruskal-Wallis chi-squared = 18741, df = 6337, p-value < 2.2e-16
```

```
> kruskal.test(TAT ~ GTEP)

        Kruskal-Wallis rank sum test

data:  TAT by GTEP
Kruskal-Wallis chi-squared = 26197, df = 12966, p-value < 2.2e-16

> kruskal.test(TEY ~ GTEP)

        Kruskal-Wallis rank sum test

data:  TEY by GTEP
Kruskal-Wallis chi-squared = 33597, df = 12966, p-value < 2.2e-16

> kruskal.test(CDP ~ GTEP)

        Kruskal-Wallis rank sum test

data:  CDP by GTEP
Kruskal-Wallis chi-squared = 35321, df = 12966, p-value < 2.2e-16
```

- ❖ Significant differences are observed in TEY and CDP across TAT groups (with p-values < 0.05 for all comparisons), indicating that these variables differ significantly across the groups.

Conclusion:

For each Kruskal-Wallis test result, since the p-value is very small, we reject H0 in favor of H1, indicating a significant difference between the distributions of the variables (TEY and CDP) across the TAT groups.

GTEP groups:

- Significant differences are observed in TAT, TEY, and CDP across GTEP groups, with p-values < 0.05.

Conclusion:

For each Kruskal-Wallis test result, since the p-value is very small, we reject H0 in favor of H1, indicating a significant difference between the distributions of the variables (TAT, TEY, and CDP) across the GTEP groups.

```
> kruskal.test(TAT ~ GTEP)

        Kruskal-Wallis rank sum test

data:  TAT by GTEP
Kruskal-Wallis chi-squared = 26197, df = 12966, p-value < 2.2e-16

> kruskal.test(TEY ~ GTEP)

        Kruskal-Wallis rank sum test

data:  TEY by GTEP
Kruskal-Wallis chi-squared = 33597, df = 12966, p-value < 2.2e-16

> kruskal.test(CDP ~ GTEP)

        Kruskal-Wallis rank sum test

data:  CDP by GTEP
Kruskal-Wallis chi-squared = 35321, df = 12966, p-value < 2.2e-16
```

AH groups:

```
> kruskal.test(TEY ~ AH)  # Compare TEY across AH groups

        Kruskal-Wallis rank sum test

data:  TEY by AH
Kruskal-Wallis chi-squared = 25652, df = 25709, p-value = 0.5984

> kruskal.test(CDP ~ AH)  # Compare CDP across AH groups

        Kruskal-Wallis rank sum test

data:  CDP by AH
Kruskal-Wallis chi-squared = 26074, df = 25709, p-value = 0.05447
```

```
> kruskal.test(AFDP ~ AH) # Compare AFDP across AH groups

        Kruskal-Wallis rank sum test

data:  AFDP by AH
Kruskal-Wallis chi-squared = 25810, df = 25709, p-value = 0.3277

> kruskal.test(GTEP ~ AH) # Compare GTEP across AH groups

        Kruskal-Wallis rank sum test

data:  GTEP by AH
Kruskal-Wallis chi-squared = 26199, df = 25709, p-value = 0.01569

> kruskal.test(TAT ~ AH)  # Compare TAT across AH groups

        Kruskal-Wallis rank sum test

data:  TAT by AH
Kruskal-Wallis chi-squared = 26102, df = 25709, p-value = 0.0421
```

- Significant differences are observed in GTEP and TAT across AH groups (both with p-values < 0.05).No significant differences are found for AFDP, TEY, and CDP across AH groups, as their p-values are greater than 0.05.

Conclusion:

For GTEP and TAT, since the p-value is small, we reject H0 in favor of H1, indicating a significant difference between the distributions of these variables across the AH groups. For AFDP, TEY, and CDP, since the p-value is not significant, we fail to reject H0, indicating no significant difference between the distributions of these variables across the AH groups.

TEY groups:

```
> kruskal.test(CDP ~ TEY)

        Kruskal-Wallis rank sum test

data:  CDP by TEY
Kruskal-Wallis chi-squared = 34458, df = 6235, p-value < 2.2e-16
```

- The p-value is significantly less than 0.05, indicating a highly significant difference in the distribution of CDP across TEY groups.

Conclusion:

For the Kruskal-Wallis test result, since the p-value is very small, we reject H0 in favor of H1, indicating a significant difference between the distributions of CDP across the TEY groups.

8. Linear Regression

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. By fitting a linear equation to the data, it helps quantify the strength and direction of these relationships, allowing for predictions and insights into how changes in independent variables influence the dependent variable.

- Regression model

```
# Define the dependent and independent variables
target_var <- "NOX"  # Target variable
independent_vars <- c("AP", "AT")  # Independent variables

# Fit the regression model
model <- lm(as.formula(paste(target_var, "~", paste(independent_vars,
collapse = " + "))),
            data = normal_data)

# Summary of the model
summary(model)
```

```
> summary(model)

Call:
lm(formula = as.formula(paste(target_var, "~", paste(independent_vars,
    collapse = " + "))), data = normal_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62298 -0.10969 -0.01083  0.10790  0.72195

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.883515   0.004431  199.400   <2e-16 ***
AP          -0.044142   0.004730   -9.331   <2e-16 ***
AT          -0.634267   0.005087 -124.687   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1536 on 36730 degrees of freedom
Multiple R-squared:  0.3242,    Adjusted R-squared:  0.3241
F-statistic:  8808 on 2 and 36730 DF,  p-value: < 2.2e-16
```

- ❖ The linear regression model examines the relationship between NOX (dependent variable) and AP and AT (independent variables).
- Intercept: The baseline value of NOX is 0.8835 when AP and AT are 0.
- AP Coefficient (0.0414): For every unit increase in AP, NOX increases by 0.0414, holding AT constant (statistically significant with p-value < 2e-16).

- AT Coefficient (-0.8363): For every unit increase in AT, NOX decreases by 0.8363, holding AP constant (statistically significant with p-value < 2e-16).

❖ Model Fit:

- R-squared (0.324): About 32.4% of the variability in NOX is explained by AP and AT.
- F-statistic (8808, p-value < 2.2e-16): The model as a whole is highly significant.

Conclusion:

While the regression model explains a moderate proportion of the variability in NOX, the predictors AP and AT are highly significant. AT has a much stronger negative impact on NOX compared to AP. Further improvements, such as adding more variables or testing interactions, could potentially enhance the model's explanatory power.

- Strategy to Improve the Model

```
#check if there's loss of information
summary(resid(model))
```

```
> summary(resid(model))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.62298 -0.10969 -0.01083  0.00000  0.10790  0.72195
```

❖ The model is effective as the residuals summary indicates a mean of 0, confirming that no information is lost.

9. Dimensional reduction

Dimensionality reduction is a crucial step in statistical analysis, used to simplify datasets with many variables while preserving their essential information. By reducing the number of dimensions, we can improve the efficiency of the analysis, reduce computational complexity, and mitigate potential issues like multicollinearity. This process helps to focus on the most impactful variables, making the data easier to interpret and visualize.

-Correlation test for variables with the same informational content

```
> cor.test(NOX ,CDP,method="spearman")

        Spearman's rank correlation rho

data:  NOX and CDP
S = 8.9799e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.08705732

Message d'avis :
Dans cor.test.default(NOX, CDP, method = "spearman") :
  Impossible de calculer la p-value exacte avec des ex-aequos
```

```
> cor(NOX ,CDP, method = "spearman")
[1] -0.08705732
```

- A rho of -0.087 is very close to 0, indicating a very weak negative monotonic relationship (almost negligible).
  2. p-value: < 2.2e-16

```
> cor(NOX ,TEY, method = "spearman")
[1] 0.01979976
> cor.test(NOX ,TEY,method="spearman")

        Spearman's rank correlation rho

data:  NOX and TEY
S = 8.0972e+12, p-value = 0.0001476
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01979976

Message d'avis :
Dans cor.test.default(NOX, TEY, method = "spearman") :
  Impossible de calculer la p-value exacte avec des ex-aequos
```

```
> cor(NOX ,AH, method = "spearman")
[1] 0.1438031
> cor.test(NOX ,AH,method="spearman")

        Spearman's rank correlation rho

data:  NOX and AH
S = 7.0728e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1438031

Message d'avis :
Dans cor.test.default(NOX, AH, method = "spearman") :
  Impossible de calculer la p-value exacte avec des ex-aequos
```

- NOX and TEY: Spearman's rho = 0.0198 (very weak correlation) p-value = 0.0001476 (statistically significant)Action: Since the correlation is weak (ϱ close to 0), you may drop TEY despite the significance.

- NOX and CDP: Spearman's rho = -0.087 (weak negative correlation); p-value 2.2e-16 (statistically significant) Action: The weak correlation suggests dropping CDP, even though it's statistically significant.

- NOX and AH Spearman's rho = 0.1438 (weak but stronger correlation) p-value < 2.2e-16 (statistically significant)Action: You can keep AH, as it has a relatively stronger correlation compared to the others.

Conclusion:

Drop: Variables with weak correlations (TEY, CDP)

```
> normalized_data_cleaned <- normalized_data[, !(colnames(normalized_data) %in% c("TEY", "CDP"))]
> head(normalized_data_cleaned)
        AT        AP        AH      AFDP      GTEP       TIT       TAT
1 0.2497266 0.6766321 0.7408575 0.3895010 0.3112642 0.7802691 0.9437229
2 0.2429288 0.6644295 0.7496393 0.3882187 0.3098766 0.7787743 0.9595960
3 0.2339597 0.6674802 0.7594091 0.3913328 0.3118093 0.7847534 0.9696970
4 0.2302470 0.6644295 0.7684419 0.3908094 0.3078943 0.7847534 0.9682540
5 0.2304316 0.6491763 0.7644900 0.3901028 0.3081917 0.7757848 0.9559885
6 0.2335282 0.6461257 0.7451073 0.3912281 0.3074979 0.7772795 0.9545455
          CO       NOX
1 0.06117300 0.8059603
2 0.08390082 0.8133498
3 0.08457584 0.8376743
4 0.04325476 0.8155753
5 0.05008004 0.8072817
6 0.04394104 0.8024133
```

```
> dim(normalized_data_cleaned)
[1] 36733     9
```

- ❖ After applying dimensionality reduction, the dataset's dimensions were reduced to 36,733 rows and 9 columns. This indicates that the number of features was streamlined to focus on the most relevant variables, simplifying the dataset while retaining its key information for analysis.

10. Conclusion

This statistical analysis provided valuable insights into the dataset through a comprehensive exploration of its characteristics and relationships. The data cleaning process ensured the quality of the dataset by addressing duplicates and outliers while confirming the absence of missing values. Through normalization, variables were scaled for consistency, and normality tests highlighted the need for non-parametric methods in some cases.

Bivariate analyses, including t-tests, variance tests, and the Kruskal-Wallis test, revealed significant relationships and differences between variables. Dimensionality reduction effectively simplified the dataset, retaining key information for further analysis. Finally, linear regression identified the significant influence of variables such as **AP** and **AT** on **NOX**, explaining a moderate proportion of its variability.

In conclusion, this report highlights the effective application of statistical techniques to understand, clean, and analyze the dataset, providing meaningful results to support data-driven decision-making.