Huyen Truong - BDS22 - BS22DSY032

AIR QUALITY

Statistical Data Analysis Project

INTRODUCTION

Air pollution is a global problem that affects the health and well-being of millions of people. Understanding the factors that contribute to air pollution is essential for developing effective policies and interventions to improve air quality. In this report, we will be analyzing a dataset on air quality collected from a device located in a significantly polluted area, at road level, within an Italian city. The dataset contains information on different air pollutants, as well as weather variables such as temperature and humidity. We will be performing a thorough statistical analysis of the dataset, including exploratory data analysis and multiple regression analysis. Our goal is to identify the variables that are most important in predicting air quality and to provide insights into the factors that contribute to air pollution in urban areas. The findings from this analysis can be used to inform policy decisions and interventions aimed at improving air quality in Italian cities and other urban areas around the world.

METHODOLOGY

We followed a structured approach to analyze the air quality dataset, consisting of the following steps:

1. Raw data collection
2. Data Cleaning and Preprocessing
3. Exploratory Data Analysis
4. Multiple Regression Analysis

1. Raw data collection:

https://archive.ics.uci.edu/ml/datasets/air+quality

2. Data Cleaning and Preprocessing

Overview:

The dataset used in this study contains 9358 instances of hourly averaged responses. It consists of 15 attributes including Date, Time, and measurements for various air pollutants such as CO, NMHC, C6H6, NOx, NO2, and O3, and weather variables such as temperature and humidity.

To ensure the accuracy and reliability of our analysis, we performed the following data cleaning and processing steps:

- Dealing with missing values:

The missing values in the dataset were identified as -200 and were replaced with NaN values to facilitate data inspection. The columns NMHC(GT), CO(GT), NOx(GT), and NO2(GT) were dropped due to the high percentage of null values they contained, as compared to other sensors.

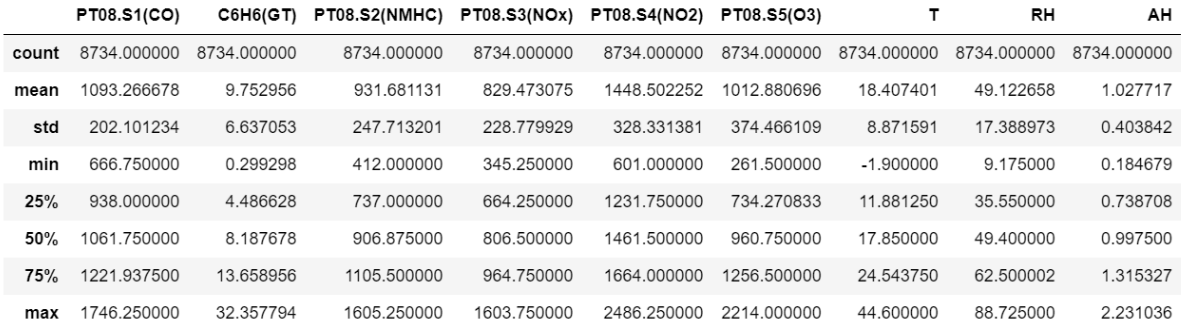
- Dealing with outliers:
  Extreme outliers were identified using Z-scores and were removed from the dataset as they might have a negative impact on the analysis. For each column, the Z-score was calculated relative to the column mean and standard deviation. The absolute value of the Z-score was taken, and any value beyond 3 standard deviations from the mean was considered an extreme outlier and removed from the dataset.

After cleaning and processing the data, there are 8734 rows remaining with 11 columns.

3. Exploratory Data Analysis

To gain insights into the air quality dataset, we performed the following exploratory data analysis techniques:

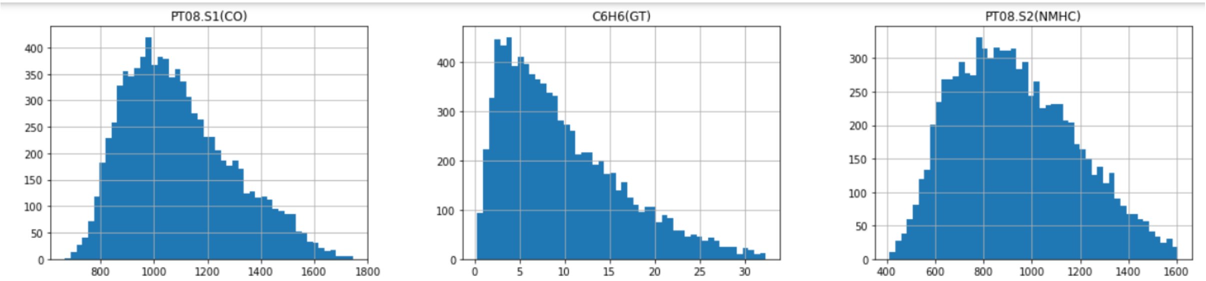
The descriptive statistics for numerical features:

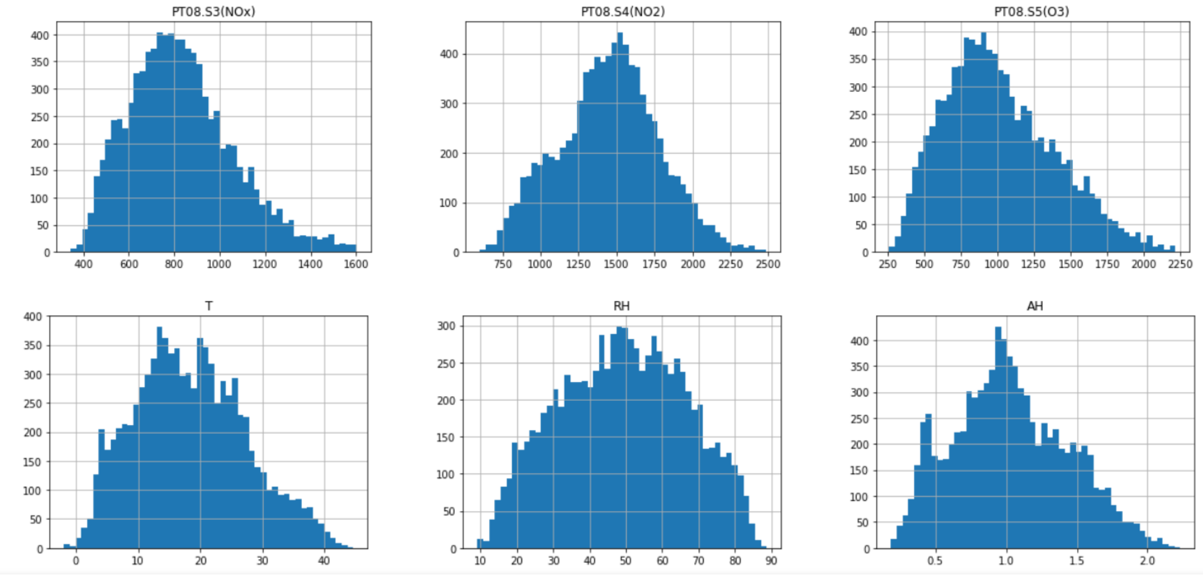


	PT08.S1(CO)	C6H6(GT)	PT08.S2(NMHC)	PT08.S3(NOx)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
count	8734.000000	8734.000000	8734.000000	8734.000000	8734.000000	8734.000000	8734.000000	8734.000000	8734.000000
mean	1093.266678	9.752956	931.681131	829.473075	1448.502252	1012.880696	18.407401	49.122658	1.027717
std	202.101234	6.637053	247.713201	228.779929	328.331381	374.466109	8.871591	17.388973	0.403842
min	666.750000	0.299298	412.000000	345.250000	601.000000	261.500000	-1.900000	9.175000	0.184679
25%	938.000000	4.486628	737.000000	664.250000	1231.750000	734.270833	11.881250	35.550000	0.738708
50%	1061.750000	8.187678	906.875000	806.500000	1461.500000	960.750000	17.850000	49.400000	0.997500
75%	1221.937500	13.658956	1105.500000	964.750000	1664.000000	1256.500000	24.543750	62.500002	1.315327
max	1746.250000	32.357794	1605.250000	1603.750000	2486.250000	2214.000000	44.600000	88.725000	2.231036

As we can see from the table, the number of attribute vectors, their mean, standard deviation, minimum/maximum, 1st-2nd-3rd quartiles.

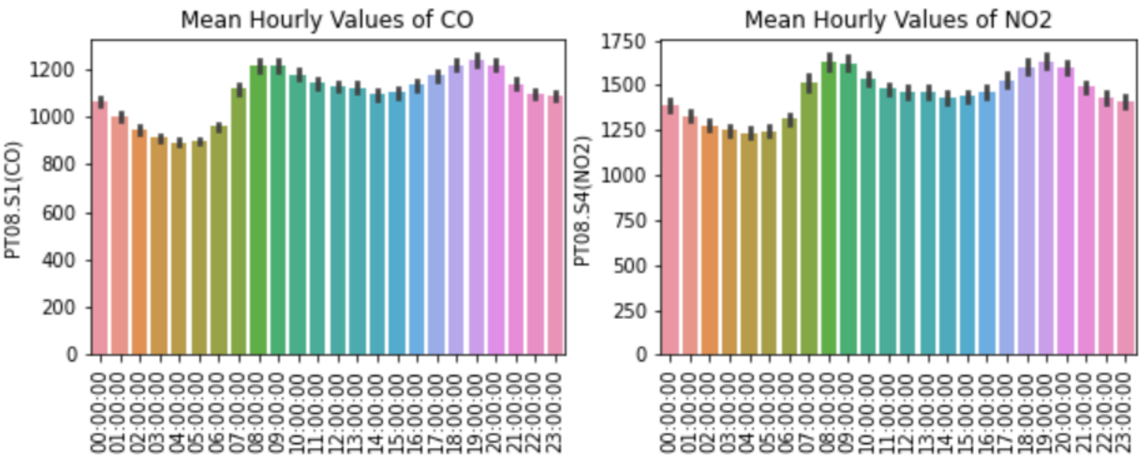
Plot 1:

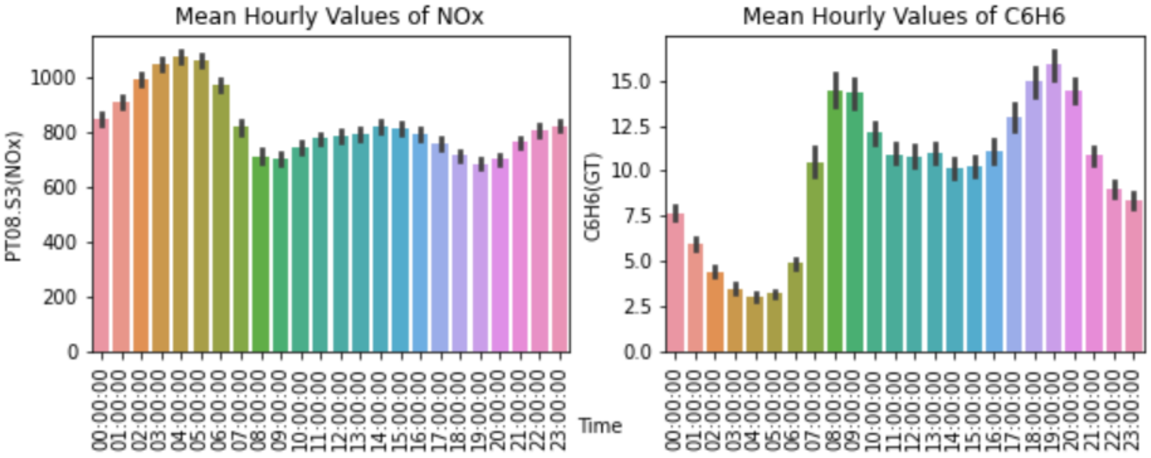




- The distributions of PT08.S1(CO), C6H6(GT), PT08.S3(NOx), PT08.S4(NO2) are positively skewed.
- The approximately normal distributions of T and RH suggest that the temperature and humidity in the area are relatively stable.

Plot 2:

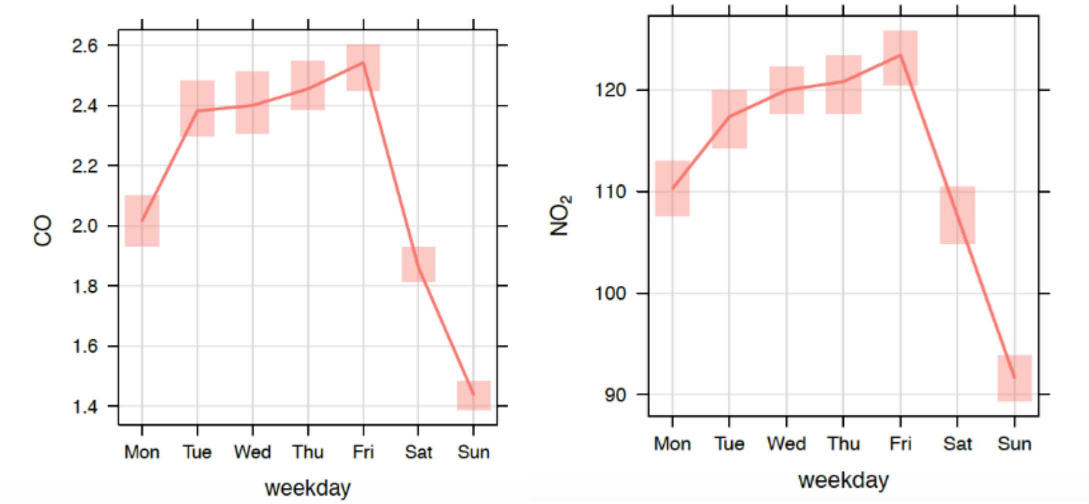


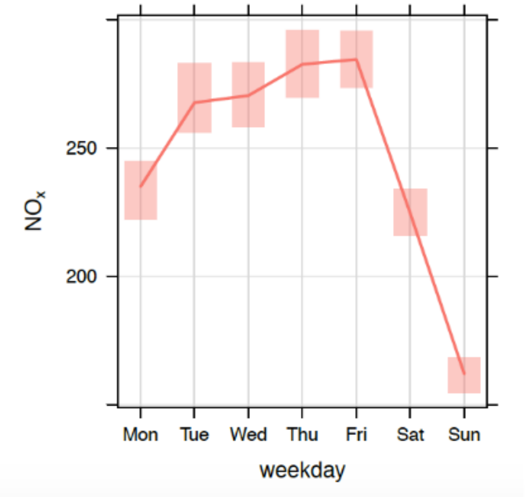


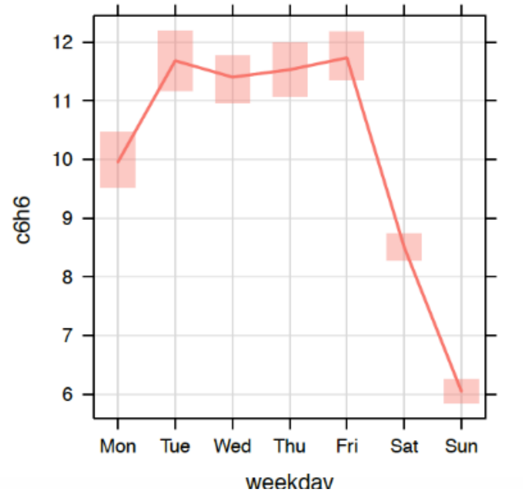
From these above bar charts, we can see that:

- The peak concentrations of CO, NO2, and C6H6 in the city are between 8 AM and 9 AM and between 6 PM and 8 PM, which coincide with the beginning and ending of office hours, respectively.
- The concentration level of NOx is high in the evening between 9 PM and 11 PM.

Plot 3:

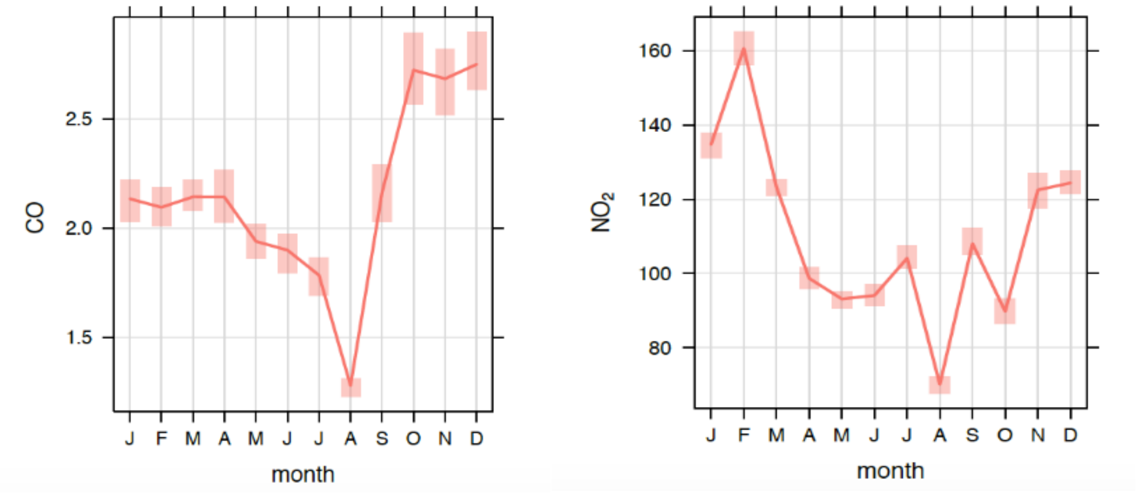


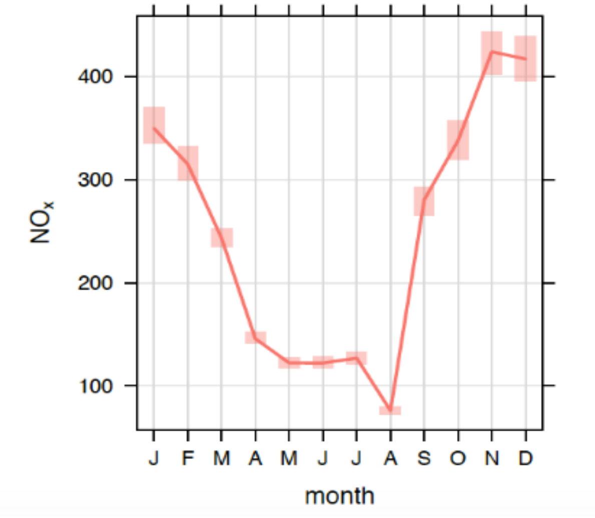


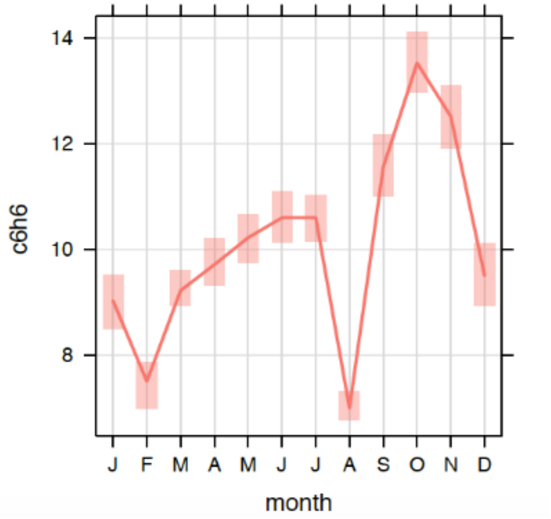


Based on the line charts for the pollutants CO, NO2, C6H6, and NOx, it appears that the levels of these pollutants are generally higher on weekdays compared to weekends.

Plot 4:





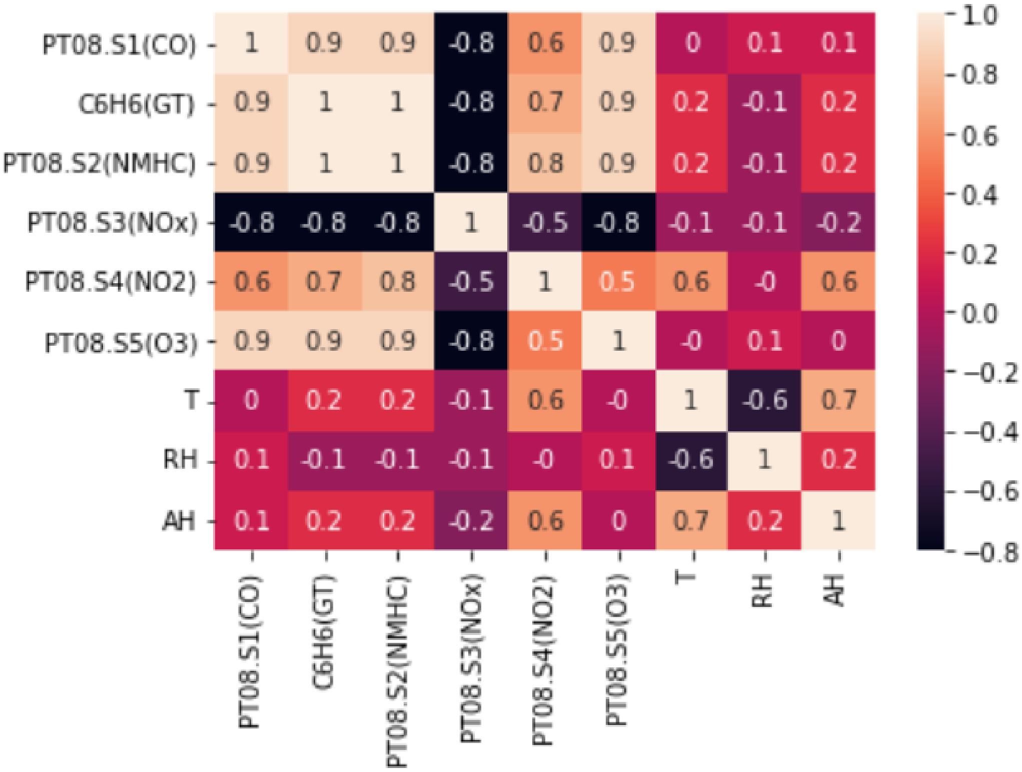


From these line charts, we can see that:

- The month of August has the lowest level of contaminants in the air.
- These pollutant concentrations (CO, NO2, NOx and C6H6) are higher in autumn and winter than other seasons.

(It is important to identify such seasons in the data analysis as it is measured in an Italian city, where Autumn in Italy usually begins in September and lasts until about mid-December, the winter season typically starts in December and ends by the end of February or early March.)

Plot 5:



From the heat map, the variables PT08.S1(CO), C6H6(GT), PT08.S2(NMHC), PT08.S4(NO2), and PT08.S5(O3) have a strong positive correlation with each other, indicating that they may be good predictors of air quality.

4. Multiple Regression Analysis

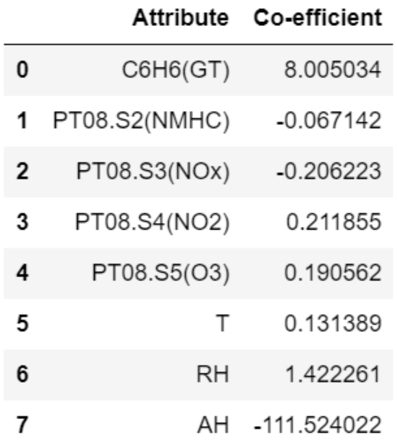
The objective of the multiple regression analysis is to examine the relationship between carbon monoxide (CO) concentration and other air quality and meteorological variables, and to identify the factors that contribute to CO concentration in the air.

CO concentration is defined as the dependent variable, and other air quality and meteorological variables such as NOx, NO2, temperature, and humidity as independent variables. The analysis was performed using Jupyter Notebook code.

To maximize the accuracy of our model's predictions, we use the data file that was cleaned. The dataset was then split into a training set and a test set, with a ratio of 80:20. The model was trained to predict the actual value of CO.

<u>OUTPUTS:</u>

Multiple Linear Regression Model:



	Attribute	Co-efficient
0	C6H6(GT)	8.005034
1	PT08.S2(NMHC)	-0.067142
2	PT08.S3(NOx)	-0.206223
3	PT08.S4(NO2)	0.211855
4	PT08.S5(O3)	0.190562
5	T	0.131389
6	RH	1.422261
7	AH	-111.524022

Based on the coefficients obtained from the multiple regression analysis, we can conclude that **C6H6(GT) and RH are the most important variables to predict the value of PT08.S1(CO).**

- The coefficient of C6H6(GT) is 8.005034, indicating that for every unit increase in C6H6(GT), the value of PT08.S1(CO) increases by 8.005034 units, holding all other variables constant. This suggests a *positive relationship* between C6H6(GT) and PT08.S1(CO).

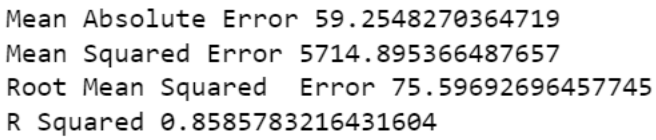
- The coefficient of RH (relative humidity) is 1.422261, indicating that for every unit increase in RH, the value of PT08.S1(CO) increases by 1.422261 units, holding all other variables constant. This suggests a *positive relationship* between RH and PT08.S1(CO).

Other variables such as PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3) and T also influence PT08.S1(CO) but may be a lesser extent compared to C6H6(GT) and RH.

***Note:**

It is important to note that the coefficient of AH (absolute humidity) in the multiple regression analysis is -111.524022, which suggests a *strong negative relationship* between AH and PT08.S1(CO). BUT the coefficient for AH is much larger in absolute terms compared to the other variables, which could indicate a possible issue and needs to be considered in other statistical diagnostics before drawing conclusions about the effect of AH on PT08.S1(CO).

Evaluate the model:

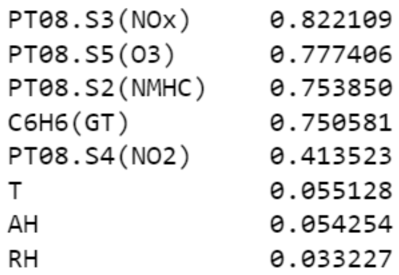


```
Mean Absolute Error 59.2548270364719
Mean Squared Error 5714.895366487657
Root Mean Squared  Error 75.59692696457745
R Squared 0.8585783216431604
```

Based on the evaluation, the regression model appears to perform reasonably well with an R-squared value of 0.86, indicating that the model explains 86% of the variation in the target variable.

Although the Mean Absolute Error (MAE) value of 59.25 is relatively large and the Mean Squared Error (MSE) value of 5714.89 indicates a significant error variance, these errors may be acceptable depending on the context of the problem being solved and additional analysis is needed.

Mutual Information value:



```
PT08.S3(NOx)      0.822109
PT08.S5(O3)       0.777406
PT08.S2(NMHC)     0.753850
C6H6(GT)          0.750581
PT08.S4(NO2)      0.413523
T                 0.055128
AH                0.054254
RH                0.033227
```

Based on the mutual information scores, these findings suggest that **the levels of PT08.S3(NOx), PT08.S5(O3), PT08.S2(NMHC), and C6H6(GT) have a significant impact on the Y variable PT08.S1(CO).**

NOTE:

From the coefficients from the linear regression above and the mutual information value, it can be seen that the results seem to be different. This is not uncommon as these are two different methods that provide distinct information about the relationship between variables.

Mutual information measures the dependence between variables and is based on the information shared between them. It is a non-parametric measure that does not assume any specific form of the relationship between variables. On the other hand, regression coefficients represent the strength and direction of the linear relationship between variables in a parametric model.

However, in the report, we want a more general measure of dependence, we decide to use the results from mutual information that '***The levels of PT08.S3(NOx), PT08.S5(O3), PT08.S2(NMHC), and C6H6(GT) have a significant impact on the Y variable PT08.S1(CO)***'. Based on the coefficients, it appears that PT08.S3(NOx), PT08.S5(O3), PT08.S2(NMHC), and C6H6(GT) are still important variables for predicting the Y variable. In addition, from the graphs in the EDA section above, we also see the relationship between PT08.S1(CO) and C6H6(GT), PT08.S3(NOx).

Finding:

PT08.S3(NOx), PT08.S5(O3), PT08.S2(NMHC), and C6H6(GT) are all air pollutants that are commonly monitored by regulatory agencies because of their negative impact on human health and the environment. The fact that they have a significant impact on PT08.S1(CO) suggests that controlling these pollutants can help to reduce levels of CO in the air, which is important for maintaining good air quality and protecting public health.

CONCLUSION

In conclusion, our statistical analysis of the air quality dataset has shown that various factors contribute to air pollution, including weather variables such as temperature and humidity, as well as different air pollutants such as CO, NOx, NO2, and C6H6. Our exploratory data analysis revealed that the concentrations of these pollutants vary according to the season, time of day, and day of the week. Furthermore, the analysis indicated a strong correlation between PT08.S1(CO) and other factors: C6H6(GT), PT08.S2(NMHC), PT08.S4(NO2), and PT08.S5(O3). Our multiple regression analysis found that C6H6 and relative humidity are also good predictors of CO concentration in the air. Overall, our findings can inform policymakers and stakeholders in developing effective policies and interventions to improve air quality in Italian cities and other urban areas around the world. It should be noted that this report has certain limitations that may impact on the accuracy and comprehensiveness of the findings. Therefore, further research and analysis may be necessary to fully understand the topic.