ANALYSIS OF FACTORS EFFECTING THE RELATIVE HUMIDITY IN AN ITALIAN CITY

Group K

192002 - R. M. N. B. Abeyrathna

192010 - A. H. M. G. Bakmeedeniya

192016 - K. K. C. Chandrathilaka

192020 - M. D. A. Darshika

192021 - B. G. I. H. Dayasiri

STAT 3232 - Data Analysis and Preparation of Statistical Reports

Department of Mathematical Sciences

Faculty of Applies Sciences

Wayamba University of Sri Lanka

Data of Submission: October 2023

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to Mrs. P.M.O.P. Panahetipola for enriching us with the knowledge in course module Data Analysis and Preparation of Statistical Reports.

We are also grateful to Ms. Samara Maduwanthi and all the other demonstrators from the junior academic staff for their constant support throughout the course module and in the preparation of this report.

And we would like to thank each and every person who guided us, supported us, and encourage us directly and indirectly to create this report. Finally, we appreciate the encouragement given by our fellow batch mates. Those comments and advises helped us to improve our Data Analysis and Preparation of Statistical Reports project.

ABSTRACT

This study is conducted to determine the factors effecting the Daily average Relative Humidity of Italian city. The regression analysis was conducted using responses of a gas multisensory device deployed on the field in an Italian city. This data set includes four independent variables, daily average Carbon monoxide (CO) concentration in atmosphere, daily average Benzene (C_6H_6) concentration in atmosphere, daily average Nitrogen dioxide (NO_2) concentration in atmosphere and temperature of the Italian city.

This analysis is based on multiple linear regression theories and the necessary assumptions made accordingly. It showed that the relationship between Relative Humidity and Carbon monoxide, Benzene, Nitrogen dioxide and Temperature. The result of this study suggests that only the daily temperature is significantly associated with the Relative Humidity. But the residuals were linearly auto correlated in this study. Therefore, further study or research is needed to confirm this results.

TABLE OF CONTENT

Page No.

LIST OF FIGURES

Page No.

LIST OF TABLES

Page No.

1. INTRODUCTION

This project is based on the data gathered from daily data collection from several factors affecting the daily average Relative Humidity of an Italian City. The data has been gathered from the website which contains the independent researches [1]. The collection includes 200 instances of daily averaged responses from an Air Quality Chemical Multisensory Device that includes an array of five metal oxide chemical sensors. The device was situated at street level on a field in a heavily polluted part of a city in Italy. The data, which includes a year from March 2004 to September 2004, are the open-source records of the reactions of field-deployed air quality chemical sensor devices. A co-located reference certified analyzer supplied the ground truth daily averaged readings for CO concentration, daily averaged Benzene (C_6H_6) concentration, daily averaged NO_2 concentration Temperature, and Relative Humidity [2].

Climatic science and air quality research, one often encounters a critical yet sometimes overlooked parameter: Relative Humidity. Relative humidity, expressed as a percentage, quantifies the amount of moisture in the atmosphere concerning the maximum amount it could hold when saturated. This seemingly straightforward metric plays a pivotal role in understanding and managing air quality, as well as its far-reaching implications for human health and the environment.

Our study delves into the intricate relationship between relative humidity and the concentration of three significant air pollutants—carbon monoxide (CO), Benzene, and nitrogen dioxide (NO_2)—alongside the influence of temperature. CO, due to its reactivity with water vapor under sunlight, initiates a transformative process that can deplete atmospheric moisture, thereby reducing relative humidity. Benzene and NO2 can directly absorb water vapor from the air, causing a decrease in relative humidity. Temperature increases air's moisture-holding capacity, causing relative humidity to decrease even when moisture levels remain constant.

In the below, we can get some idea about the variables that we get for analysis.

1. Carbon monoxide - CO

In general, carbon monoxide (CO) in the Earth's atmosphere is less than 0.001%, or parts per million, or ppm. CO concentrations can change based on some variables, including location and pollution levels. The levels of CO may be somewhat greater in cities with heavy traffic and manufacturing activities, but they still make up a very minor portion of the atmosphere overall. It's crucial to remember that carbon monoxide is a pollutant and that large amounts of it may be dangerous to human health.

2. Benzene - C_6H_6

The quantity of benzene, an air contaminant and volatile organic compound (VOC), in the Earth's atmosphere, is normally quite low. Even though the quantity of benzene could be slightly higher in metropolitan regions and places with more traffic and manufacturing activities, it still makes up a relatively minor portion of the atmosphere overall. The location, local sources of emissions, and weather can all have a significant impact on the precise concentration of benzene in the air. Benzene may be hazardous to human health even at low doses, it is crucial to monitor and regulate the amount of this chemical in the air.

3. Nitrogen dioxide - NO_2

Nitrogen dioxide (NO_2) concentrations in the Earth's atmosphere are generally very low, frequently falling between 200 and 300 parts per billion, or ppb, or 0.00002% to 0.00003%. NO_2 is a result of combustion activities, such as those in cars and industrial settings, and is regarded as a trace gas in the atmosphere. Location, nearby pollution sources, and other variables might affect its concentration. Comparing urban to rural regions, there may be higher NO_2 levels in places with high traffic and industrial emissions.

4. Temperature

Depending on the location, time of day, season, and weather, the air's temperature can fluctuate significantly. Generally, temperature is expressed in either Fahrenheit (°F) or Celsius (°C). The temperature of the Earth's surface varies by place and time, although

on average it is about 15°C (59°F). However, depending on variables like location, temperature, and time of day, this might be significantly higher or lower. Temperature influences many elements of the environment.

5. Relative Humidity

The quantity of moisture or water vapor in the air relative to the greatest amount the air could store at a specific temperature is known as relative humidity or RH. A percentage is used to express RH. It is essential to many natural and artificial processes as well as the weather, climate, and human comfort.

Italy, like many other countries, experiences variations in air quality depending on factors such as location, weather, industrial activity, and transportation. Our objectives are analysis of the factors affecting the Quality of Air in the country incurred by each independent variable and develop a multiple regression model to identify the impact of those variables.

2. STATISTICAL THEORY

2. 1. Descriptive Analysis

2. 1. 1. Mean

Mean is the average of all the data values and sample mean is the point estimator of the population mean μ.

$$\bar{X} = \frac{\sum_0^n X_i}{n}$$

$$\mu = \frac{\sum_0^n X_i}{n}$$

Where;

- $\sum_{i=0}^{n} x_i$ - Sum of the values of the n observations.
- n - Number of observations in the sample.
- $\bar{x}$ - Sample mean
- μ - Population mean

2. 1. 2. Median

It is the value in the middle of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is median and if it is even then the median would be the average of two central elements.

2. 1. 3. Standard Derivation

The standard deviation is a summary measure of the differences of each observation from the mean. If the differences themselves were added up, the positive would exactly balance the negative and so their sum would be zero. Consequently, the squares of the differences are added. The sum of the squares is then divided by the number of observations minus onto give the mean of the squares, and the square root is taken to bring the measurements back to the units the calculations started with.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)}{N}}$$

Where;

- σ = Population standard deviation
- N = Size of the population
- x_i = Each value from the population
- μ = The population mean

2. 1. 4. Maximum Value

This number is the data value that is less than or equal to all other values in the set of data.

2. 1. 5. Minimum Value

This number is the data value that is greater than or equal to all other values in our set of data

2. 2. Pearson Correlation Coefficient

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

r_{xy} can be defined as;

$$\mathbf{r} = \frac{\sum(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{Y_i} - \bar{\mathbf{Y}})}{\sqrt{\sum(\mathbf{X_i} - \bar{\mathbf{X}})^2 \sum(\mathbf{Y_i} - \bar{\mathbf{Y}})^2}}$$

Where,

- r - correlation coefficient
- X_i - values of the x-variable in a sample
- $\bar{X}$ - mean of the values of the x-variable
- Y_i - values of the y-variable in a sample
- $\bar{Y}$- mean of the values of the y-variable

2. 2. 1. Positive correlation

When values of the correlation coefficient are greater than zero there is a positive correlation. The values close to +1 defines a perfect positive correlation. The variables move into same direction; therefore, one variable increases as other one increases or decreases as the other decreases.

2. 2. 2. Negative correlation

When values of the correlation coefficient are less than zero there is a negative correlation. The values close to -1 defines a perfect negative correlation. The variables move into same direction, therefore increase in one variable as other one decreases.

2. 2. 3. No correlation

When the correlation is zero there is no correlation hence there is no relationship between two variables. In general, correlation coefficient value greater than 0.8 defined as strong, where the value less than 0.5 defined as weak relationship

2. 3. Coefficient of Determinations (R Square)

The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. In other words, this coefficient, which is more commonly known as R-squared (or R^2), assesses how strong the linear

relationship is between two variables, and is heavily relied on by researchers when conducting trend analysis.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where,

- R^2 - coefficient of determination
- RSS - sum of square of residuals
- TSS - total sum of squares

2. 4. Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as a key into a more advanced analysis and as a diagnostic for advanced analyses.

2. 5. Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable. Formula and Calculation of Multiple Linear Regression,

$$\mathbf{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon}$$

Where,

for i=n observations;

- y_i = dependent variable
- x_i = explanatory variables

- β_0 = y-intercept (constant term)
- β_p = slope coefficients for each explanatory variable
- ϵ = the model's error term (also known as the residuals)

2. 6. Stepwise Regression

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model that is a model that lowers prediction error.

There are three strategies of stepwise regression (James et al. 2014,P. Bruce and Bruce (2017)):

1. Forward selection, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.

2. Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.

3. Stepwise selection (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

2. 7. Assumptions

Linear regression is the core process for various prediction analytics. By definition, linear regression refers to fitting of two continuous variables of interest. Not all datasets can be fitted into a linear fashion. There are few assumptions that must be fulfilled before fit the adequate regression model.

Assumptions are;

1. Linearity
2. Normality
3. Homoscedasticity
4. Multicollinearity
5. Autocorrelation

Linearity

The first assumption is very obvious and straightforward. There variable that we are trying to fit should maintain a linear relationship. If there is no linear relationship, the data can be transformed to make it linear. These type of transformation include taking logs on the response data or square rooting the response data. Checking scatterplot is the best and easiest way to check the linearity.

Normality

This assumption states that the residuals from the model is normally distributed. After determining the model parameters, it is good to check the distribution of the residuals. Apart from the visual of the distribution, one should check the Q-Q plot for better understanding of the distribution.

Homoscedasticity

Homoscedasticity is another assumption for multiple linear regression modeling. It requires equal variance among the data points on both side of the linear fit. If it is not the case, the data is heteroscedastic. Typically the quality of the data gives rise to this

heteroscedastic behavior. Instead of linear increase of decrease, if the response variable exhibits cone shaped distribution, we can say that variance cannot be equal at every point of the model.

Multi-collinearity

Multi-collinearity is observed when two or more independent variables are correlated to one another. If that is the case, the model's estimation of the coefficients will be systematically wrong. One can check Variance Inflation Factor (VIF) to determine the variables which are highly correlated and potentially drop those variables from the model. R^2 is a measure of how correlated the variables are and VIF is determined from this R^2 value. If the variables have high correlation, VIF value shoots up. Typically VIF value >5 indicates the presence of multi-collinearity. The minimum value of VIF is 1 which is evident for the equation and it indicates that there is no multi-collinearity out there.

Autocorrelation

Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. Using Durbin Watson Test to get the autocorrelation between variables.

3. METHODOLOGY

First, Independent and Dependent variables of the data set should be identified.

Independent Variables

- Daily average Temperature
- Daily average carbon monoxide concentration (CO)
- Daily average Benzene concentration (Benzene)
- Daily average nitrogen dioxide concentration (NO2)

Dependent Variable

- Relative Humidity (RH)

3. 1. Method of Analysis

- In this statistical analysis study, data was analyzed using R console.
- Descriptive analysis and graphs were used to explain the distribution of the independent and dependent variables.
- Then scatter plot and Correlations matrix were obtain to discuss the linear relationship between independent variables and dependent variables.
- Multiple regression models were fitted to the independent variables. Stepwise regression method was followed to get the adequate model.
- If the linear model is not significant, Stepwise regression method should be revised again and again to get the adequate model.
- Then check the assumption of linearity, normality and homoscedasticity to check whether the identified model is fitted or not.
- Finally check the correlation among residuals and multi-collinearity.

4. RESULTS

4. 1. Descriptive Analysis

4. 1. 1. Summary of Dependent and Independent variables

This table gives summary of distribution of each and every dependent and independent variables numerically. According to this we can get an idea about the range of each and every variables.

Table 1 - Summary of the data set

	CO	**Benzene**	**NO2**	**Temperature**	**RH**
Minimum	0.4	1	28	6.3	14.9
1st Quartile	1.675	6.9	96	11.38	36.42
Median	2.5	11.9	119.5	14.8	50.95
Mean	2.791	12.92	119.2	15.61	48.61
3rd Quartile	3.5	16.65	143	18.85	60.05
Maximum	8.1	39.2	230	29.3	81.1

4. 1. 2. Basic boxplots of Dependent and Independent variables

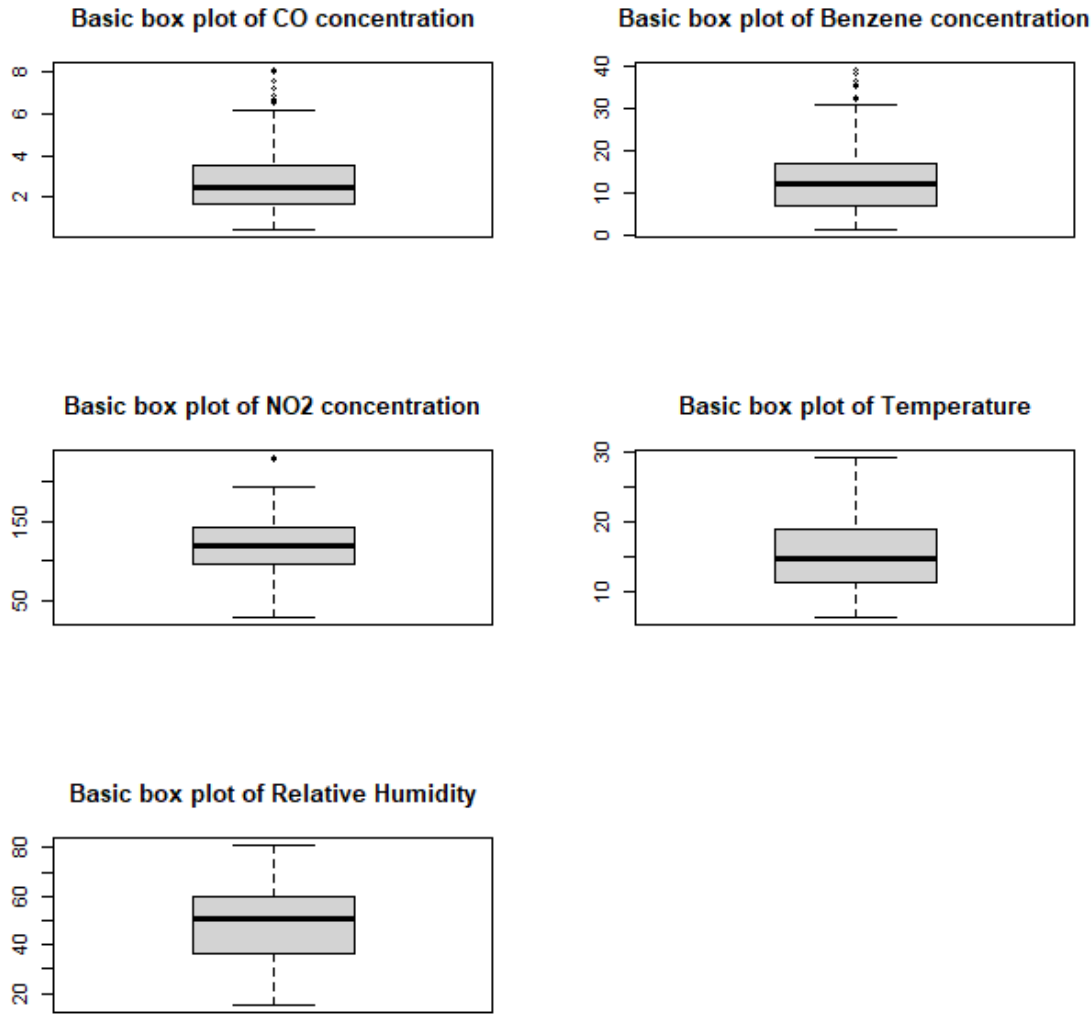

Figure 1 - Basic Boxplots of Independent and Dependent variables

Basic box plot of CO concentration indicates that mean of its lies in between 2 and 3. Also box plot of Benzene indicates that its mean is close to 10 while mean of NO2 concentration is close to 150. There are two other variables named Temperature and Relative Humidity and there means are close to 15 and 50 respectively.

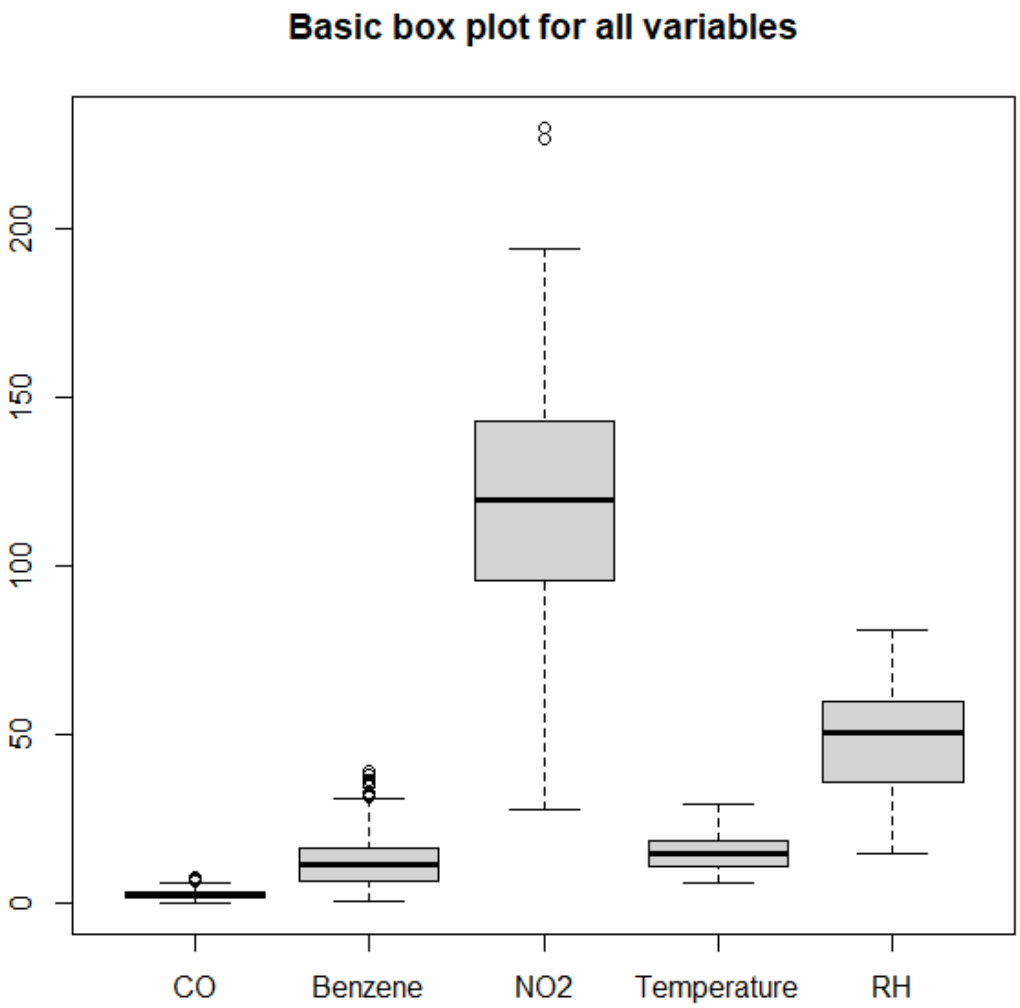

Figure 2 - Basic Boxplot of all variables

This box plot shows distribution of all variables. According to this figure, box plot of NO2 concentration comparatively taller than all other box plots while box plot of CO concentration comparatively shorter than other variables. Means of both Benzene concentration and Temperature lie in between 10 and 15.

4. 2. Correlation Matrix

Table 2 - Correlation matrix

	CO	**Benzene**	**NO2**	**Temperature**	**RH**
CO	1	0.9745872	0.8213264	0.3035964	-0.2988077
Benzene	0.9745872	1	0.8283961	0.4194497	-0.4000033
NO2	0.8213264	0.8283961	1	0.501423	-0.5084791
Temperature	0.3035964	0.4194497	0.501423	1	-0.9301893
RH	-0.2988077	-0.4000033	-0.5084791	-0.9301893	1

The obtained Correlation matrix demonstrates the relationship of Dependent variable "Relative Humidity" and Independent variables "Daily Average CO concentration", "Daily Average Benzene concentration", "Daily Average NO2 concentration" and "Temperature".

According to this matrix, CO and Benzene, CO and NO2, NO2 and Benzene have the strong positive correlation with each other. This means both variables are proportional to each other.

Temperature has weak but positive correlation with all three gases, CO, Benzene and NO2 while Relative Humidity weak negatively correlated with the all three gasses.

Also Relative Humidity is highly but negatively correlated with the Temperature.

4. 3. Scatter Plot of all pair-wise combinations

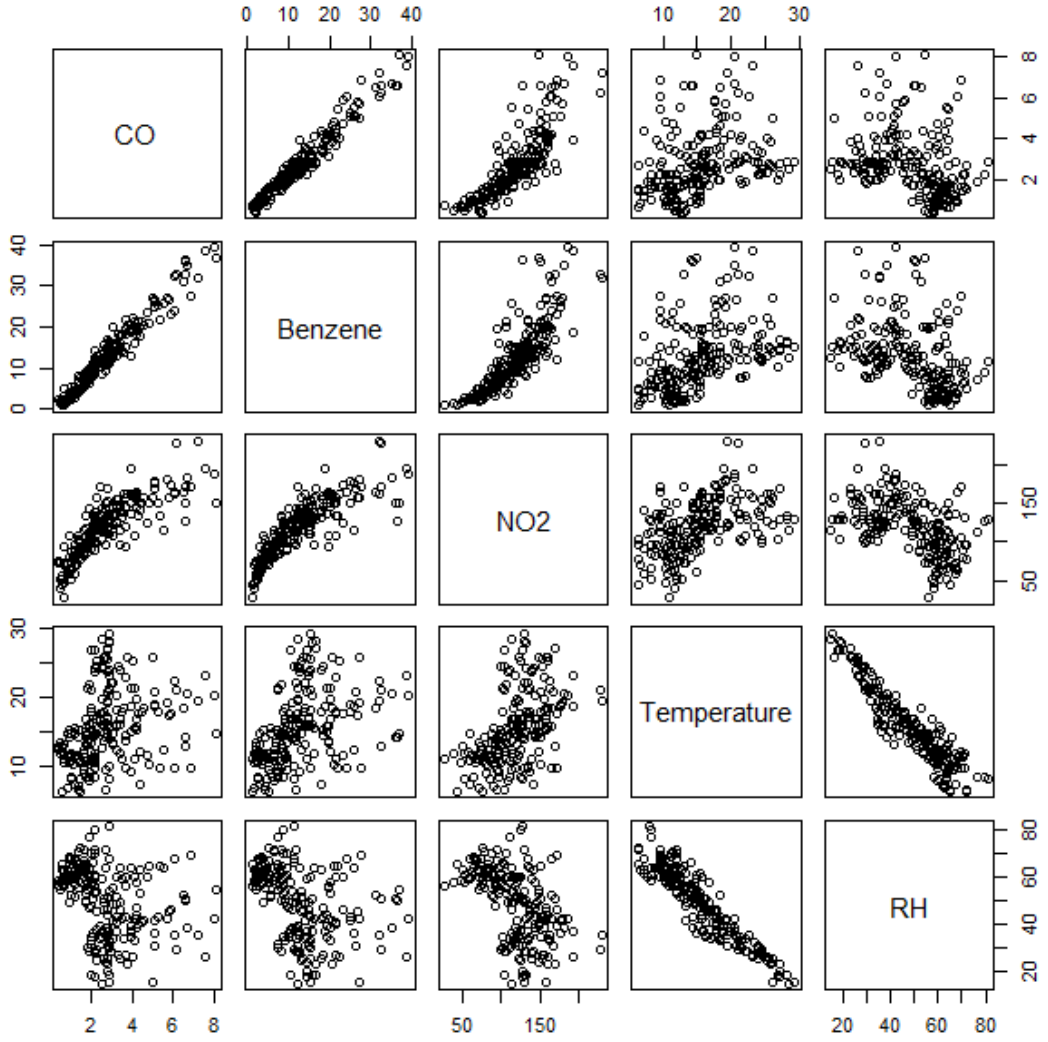

Figure 3 - Scatter plot of Dependent and Independent variables

Scatter plot also gives the brief and basic idea of the relationship in between the all variables. According to this scatter plot CO-Benzene, CO-NO2, Benzene-NO2 variables show positive correlation. Also Temperature – Relative Humidity indicates the negative correlation. The plot is not provided clear clarification between the rest of the independent and dependent variables.

4. 4. Multiple Linear Regression Model

The below table gives the summary of regression model using all variables of the data set and Hypothesis testing should be done for this.

Table 3 - Summary of the Regression model of all Independent variables

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	91.92672	1.80581	50.906	<2e-16 ***
CO	-0.6154	1.30899	-0.47	0.6388
Benzene	0.25586	0.25486	1.004	0.3167
NO2	-0.04841	0.02181	-2.22	0.0276 *
Temperature	-2.50656	0.1009	-24.843	<2e-16 ***

For Intercept;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected at 5% level of significance. Hence we can conclude Intercept is significant.

For Daily average CO concentration;

Hypothesis;

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

Since p value = 0.6388 greater than $\alpha = 0.05$; Therefor H_0 is not rejected at 5% level of significance. Hence we can conclude CO variable is not significant at 5% level of significance.

For Daily average Benzene concentration;

Hypothesis;

$H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$

Since p value = 0.3167 greater than $\alpha = 0.05$; Therefor H_0 is not rejected at 5% level of significance. Hence we can conclude Benzene variable is not significant at 5% level of significance.

For Daily average NO2 concentration;

Hypothesis;

$H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0$

Since p value = 0.0276 less than $\alpha = 0.05$; Therefor H_0 is rejected at 5% level of significance. Hence we can conclude NO2 variable is significant at 5% level of significance.

For Temperature;

Hypothesis;

$H_0: \beta_4 = 0$ vs $H_1: \beta_4 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected at 5% level of significance. Hence we can conclude Temperature variable is significant at 5% level of significance.

Residual standard error: 5.428 on 195 degrees of freedom
Multiple R-squared: 0.8697, Adjusted R-squared: 0.8671
F-statistic: 325.5 on 4 and 195 DF, p-value: < 2.2e-16

For the Regression model;

Hypothesis;

H_0: Model is not statistically significant

H_1: Model is statistically significant

Since the p-value of whole set of data is < 2.2e-16 and it is less than $\alpha = 0.05$, Therefore, H_0 is rejected at 5% level of significance. Hence we can conclude that obtained model is statistically significant at 5% level of significance.

R^2 Suggest that 86.97 percent of variability of "Relative Humidity" is explained by the regression model

But since all the variables are not statistically significant, we can't accept this model as the best model for this data set. Therefore, Step-Wise Regression must be done.

4. 2. Stepwise Regression

4. 2. 1. Checking Step Wise Regression and Hypothesis Testing for Model 1

Step-wise regression is used to build statistically significant model for regression analysis by adding or removing independent variables one at a time.

Start: AIC=1081.18

RH ~ 1

Table 4 - Stepwise Regression - Model 1 - Step 1

	Df	**Sum of Sq**	**RSS**	**AIC**
+ Temperature	1	38158	5942	682.31
+ NO2	1	11402	32698	1023.35
<none>			44100	1081.18

Step: AIC=682.31

RH ~ Temperature

Table 5 - Stepwise Regression - Model 1 - Step 2

	Df	**Sum of Sq**	**RSS**	**AIC**
+ NO2	1	104	5838	680.77
<none>			5942	682.31
- Temperature	1	38158	44100	1081.18

Step: AIC=680.77

RH ~ Temperature + NO2

Table 6 - Stepwise Regression - Model 1 - Step 3

	Df	**Sum of Sq**	**RSS**	**AIC**
<none>			5838	680.77
- NO2	1	104.2	5942	682.31
- Temperature	1	26859.7	32698	1023.35

Call:

lm(formula = RH ~ Temperature + NO2)

Coefficients:

Table 7 - Coefficients of the Model 1

(Intercept)	**Temperature**	**NO2**
90.14407	-2.47778	-0.02391

After three steps of stepwise regression we can get the regression model as;

RH = 90.14407 - 2.47778 (Temperature) - 0.02391 (NO2)

After the Step Wise Regression we should check the significance of model that we build and the variables in that model

Table 8 - Summary of the Regression Model 1

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	90.14407	1.4663	61.477	<2e-16 ***
Temperature	-2.47778	0.0823	-30.106	<2e-16 ***
NO2	-0.02391	0.01275	-1.875	0.0622 .

For Intercept;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected. Hence we can conclude Intercept is statistically significant at 5% level of significance.

For Temperature;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected. Hence we can conclude Temperature variable is statistically significant at 5% level of significance.

For Daily average NO2 concentration;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value = 0.0622 greater than $\alpha = 0.05$; Therefor H_0 is not rejected. Hence we can conclude NO2 variable is not statistically significant at 5% level of significance

Since one variable is not statistically significant, Step-wise regression is should perform again

4. 2. 2. Checking Step Wise Regression and Hypothesis Testing for Model 2

Start: AIC=1081.18

RH ~ 1

Table 9 - Stepwise Regression - Model 2 - Step 1

	Df	**Sum of Sq**	**RSS**	**AIC**
+Temperature	1	38158	5942	682.31
<none>			44100	1081.18

Step: AIC=682.31

RH ~ Temperature

Table 10 - Stepwise Regression - Model 2 - Step 2

	Df	**Sum of Sq**	**RSS**	**AIC**
<none>			5942	682.31
-Temperature	1	38158	44100	1081.18

Call:

lm(formula = RH ~ Temperature)

Coefficients:

Table 11 - Coefficients of Model 2

(Intercept)	**Temperature**
88.503	-2.555

After two steps of stepwise regression we can get the regression model as;

RH = 88.503 - 2.555 (Temperature)

After the Step Wise Regression we should check the significance of the model 2 that we build and the variables in that model

Table 12 - Summary of Regression Model 2

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	88.50326	1.18409	74.74	<2e-16 ***
Temperature	-2.55517	0.07166	-35.66	<2e-16 ***

For Intercept;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected. Hence we can conclude Intercept is statistically significant at 5% level of significance.

For Temperature;

Hypothesis;

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Since p value < 2e-16 less than $\alpha = 0.05$; Therefor H_0 is rejected. Hence we can conclude Temperature variable is statistically significant at 5% level of significance.

Residual standard error: 5.478 on 198 degrees of freedom

Multiple R-squared: 0.8653, Adjusted R-squared: 0.8646

F-statistic: 1271 on 1 and 198 DF, p-value: < 2.2e-16

For the Regression model;

Hypothesis;

H_0: Model is not statistically significant

H_1: Model is statistically significant

Since the p-value of whole set of data is < 2.2e-16 and it is less than $\alpha = 0.05$, Therefore, H_0 is rejected at 5% level of significance. Hence we can conclude that obtained model is statistically significant at 5% level of significance.

Obtained Multiple R-squared: 0.8653

Therefore, **86.53%** of the total variation in the Relative Humidity can be explained by this model.

Since all the variables and whole model is statistically significant, This model can be accepted as the best model for the data set. This model has the high multiple R^2 value also

Hence the best Regression Model is;

Relative Humidity = 88.50326 - 2.55517*(Temperature)

4. 6. Assumption Checking

4. 6. 1. Linearity

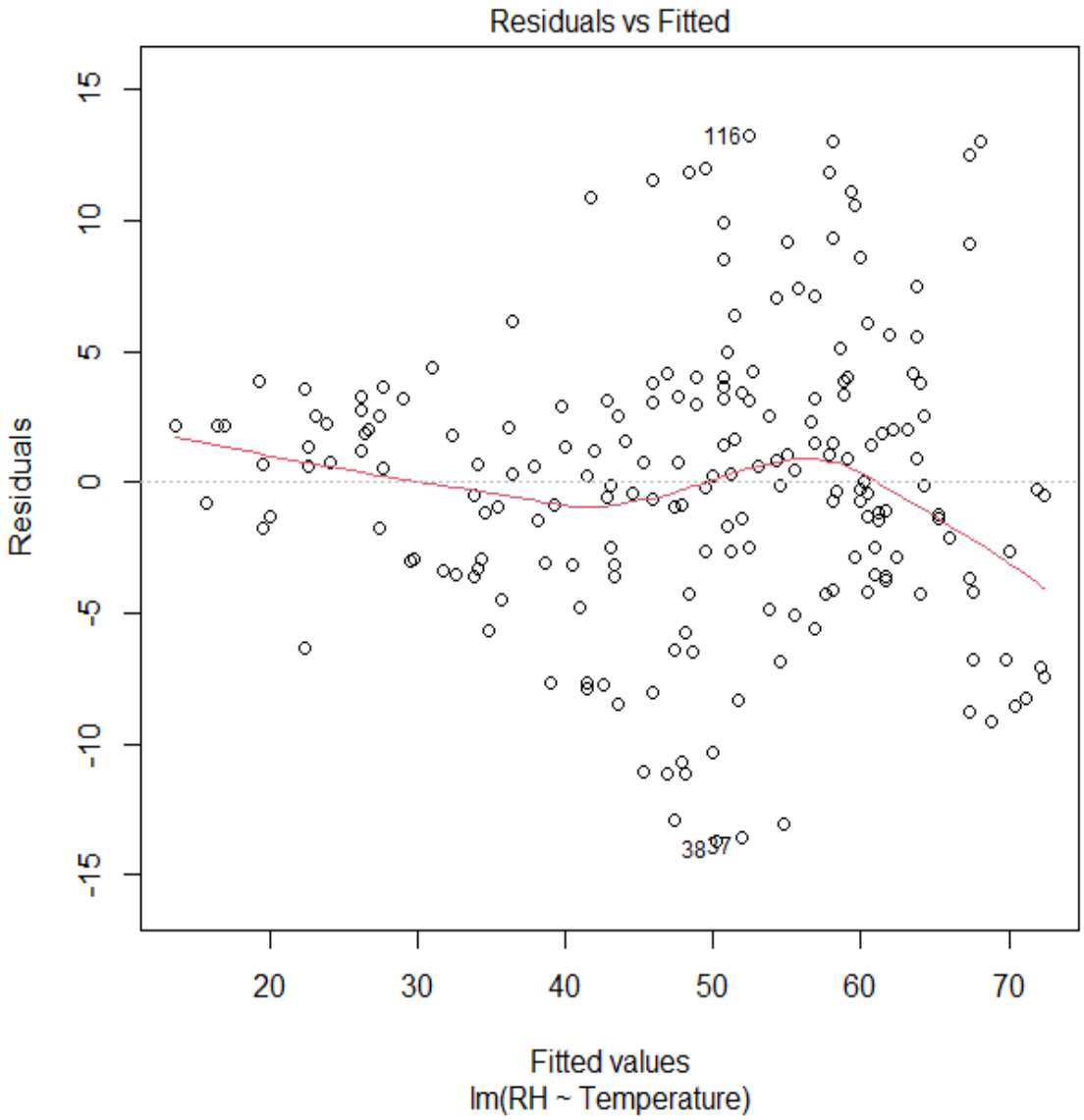

Figure 4 - Diagnostic plot of Linearity

Residuals are approximately randomly scattered in this above plot of Residuals vs Fitted values. It indicating a constant variance of residuals. Therefore, we can conclude that the residuals are random

4. 6. 2. Normality of Residuals

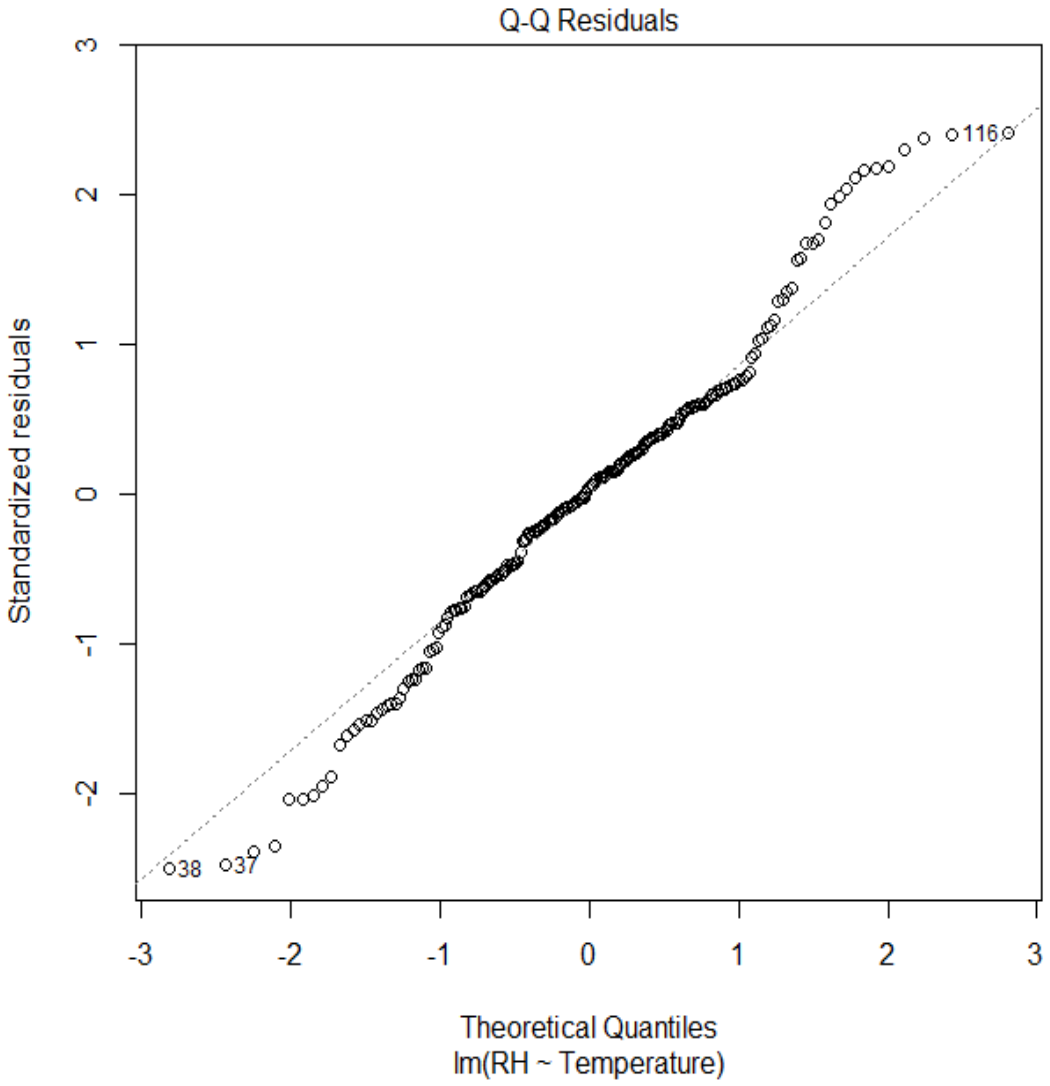

Figure 5 - Diagnostic plot of Normality

The residuals are approximately scattered in the straight line. Therefore, residuals are normality distributed.

4. 6. 3. Homoscedasticity

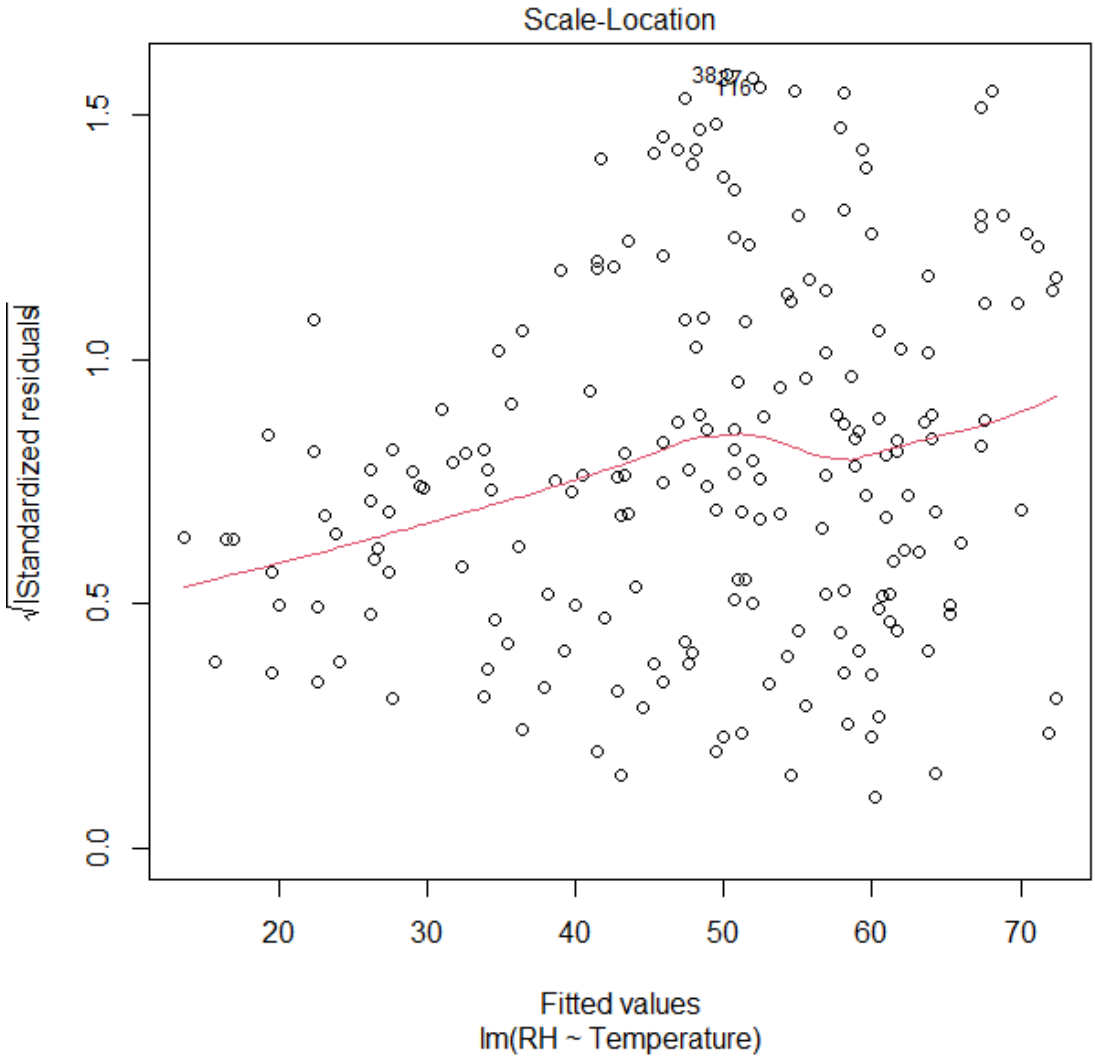

Figure 6 - Diagnostic plot of Homoscedasticity

According to the plot, residuals are scattered in approximately horizontal line with equally spread points. Therefore we can conclude that the residuals are assumed to have a constant variance. (Homoscedasticity)

4. 6. 4. Autocorrelation of the Residuals

Autocorrelation of the residuals were calculated by using Durbin Watson Test. It suggests the independency of the residuals.

Hypothesis:

H_0: Residuals are independent ($\varrho = 0$)

H1: Residuals are dependent ($\varrho \neq 0$)

Table 13 - Durbin Watson Test Value

lag	Autocorrelation	D-W Statistic	p-value
1	0.8890269	0.2177403	0

Since the P-value = 0 is less than 0.05, H_0 is rejected at 5% level of significance. Therefore, the residuals from the regression model is linearly autocorrelated. It can be assume that there is no any independency of the errors

4. 6. 4. Multi-collinearity

Since there is only one variable in the regression model, Variance Inflation Factor (VIF) cannot be calculated.

5. CONCLUSION

This study was carried out with the variables, daily CO concentration, daily Benzene concentration, daily NO_2 concentration in the atmosphere, Temperature and the Relative Humidity of the an Italian city to find out the factors which are effecting to the Relative Humidity in that city. After the applying step-wise regression process and statistical theories, we finally got the model,

Relative Humidity = 88.50326 - 2.55517*(Temperature)

But after the checking assumptions on residuals, we got the result that, Residuals are linearly correlated according to the Durbin Watson test. That means error term also can calculated using an equation. Although this model explained the 86.53% of the total variation in the Relative Humidity, we can't get this model as the best fitted model because of the error term. Hence we suggest the alternative method to continue this study to find the best fitted model.

6. DISCUSSION

Relative Humidity plays the major role in this study. Relative Humidity is a ratio, expressed in percent of the amount of atmospheric moisture present relative to the amount that would be present if the air was saturated. In this study, we checked the relationship of CO, Benzene and NO_2 concentration in the air and Temperature to relative humidity.

Usually;

- CO reacts with water vapor in the presence of sunlight to form carbon dioxide and hydrogen gas and because of that reaction can reduce the amount of water vapor in the air, leading to a decrease in relative humidity.
- Benzene and NO_2 can also absorb water vapor from the air, reducing relative humidity.
- Temperature has a direct inverse relationship on relative humidity. As temperature increases, the air can hold more water vapor. This means that relative humidity will decrease as temperature increases, even if the amount of water vapor in the air remains constant.

Although according to this study and this data set, Relative Humidity only depends on Temperature. But according to our final statistically significant model, error term is linearly auto-correlated and can be expressed using a formula also. That may be because of the effect of air particles and pollutants and some other variable which we were not included for the study. Because of those causes, we can't fit the best statistically significant model for the data set.

Relative Humidity can effect on Air Quality by both direct and indirect methods.

- Relative Humidity is inversely proportional to the Air Circulation. Which means high relative humidity can trap the pollutants in the air.
- High relative humidity can make it difficult for the body to cool itself, which can lead to heat stress and other health problems. It can also worsen respiratory problems such as asthma and bronchitis [2].

- Relative Humidity also promote the growth the mold and mildew in buildings and materials such as wood, metal and fabric [3].
- Low relative humidity can also irritate the skin and eyes [2].

These are some few real life effects of the Relative Humidity and it indicates that relative humidity can effect vast area of human life. Hence it is good to fit a model to find the variation of the relative humidity.

Reference

[1] UC Irvine School of Information and Computer Science, “Air Quality - Dataset by UCI,” *data.world*, Aug. 16, 2017. Available: https://data.world/uci/air-quality. [Accessed: Oct. 25, 2023]

[2] G. Guarnieri, B. Olivieri, G. Senna, and A. Vianello, “Relative Humidity and Its Impact on the Immune System and Infections,” *International Journal of Molecular Sciences*, vol. 24, no. 11, p. 9456, Jan. 2023, doi: https://doi.org/10.3390/ijms24119456. Available: https://www.mdpi.com/1422-0067/24/11/9456#:~:text=Higherorlowerlevelsof. [Accessed: Oct. 27, 2023]

[3] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005, .( )

Appendix

This appendix contains the daily data collection of several factors affecting the daily average Relative Humidity of an Italian City. The data has been gathered from the Air Quality - Dataset by UCI by *data.world* website [1], and the research paper of the S. De Vito and others [3].

Date	Daily Averaged CO Concentration (mg/m^3)	Daily Averaged Benzene (C6H6) Concentration (mg/m^3)	Daily Averaged NO2 Concentration (µg/m^3)	Temperature (°C)	Relative Humidity
3/10/2004	2.6	11.9	113	13.6	48.9
3/11/2004	2	9.4	92	13.3	47.7
3/12/2004	2.2	9.0	114	11.9	54.0
3/13/2004	2.2	9.2	122	11.0	60.0
3/14/2004	1.6	6.5	116	11.2	59.6
3/15/2004	1.2	4.7	96	11.2	59.2
3/16/2004	1.2	3.6	77	11.3	56.8
3/17/2004	1	3.3	76	10.7	60.0
3/18/2004	0.9	2.3	60	10.7	59.7
3/19/2004	0.7	1.1	28	11.0	56.2
3/20/2004	0.7	1.6	48	10.5	58.1
3/21/2004	1.1	3.2	82	10.2	59.6
3/22/2004	2	8.0	112	10.8	57.4
3/23/2004	2.2	9.5	101	10.5	60.6
3/24/2004	1.7	6.3	98	10.8	58.4
3/25/2004	1.5	5.0	92	10.5	57.9
3/26/2004	1.6	5.2	95	9.5	66.8
3/27/2004	1.9	7.3	112	8.3	76.4
3/28/2004	2.9	11.5	128	8.0	81.1
3/29/2004	2.2	8.8	126	8.3	79.8
3/30/2004	2.2	8.3	131	9.7	71.2
3/31/2004	2.9	11.2	135	9.8	67.6
4/1/2004	4.8	20.8	151	10.3	64.2
4/2/2004	6.9	27.4	172	9.7	69.3
4/3/2004	6.1	24.0	165	9.6	67.8
4/4/2004	3.9	12.8	136	9.1	64.0
4/5/2004	1.5	4.7	85	8.2	63.4
4/6/2004	1	2.6	53	8.2	60.8
4/7/2004	1.7	5.9	97	8.3	58.5

4/8/2004	1.9	6.4	110	7.7	59.7
4/9/2004	1.4	4.1	91	7.1	61.8
4/10/2004	0.6	1.0	44	6.3	65.0
4/11/2004	0.8	1.8	71	6.8	62.9
4/12/2004	1.4	4.4	104	6.4	65.1
4/13/2004	4.4	17.9	141	7.3	63.1
4/14/2004	3.1	14.0	122	13.2	41.7
4/15/2004	2.7	11.6	143	14.3	38.4
4/16/2004	2.1	10.2	113	15.0	36.5
4/17/2004	2.5	11.0	116	16.1	34.5
4/18/2004	2.7	12.8	123	16.3	35.7
4/19/2004	2.9	14.2	126	15.8	37.0
4/20/2004	2.8	12.7	120	15.9	37.2
4/21/2004	2.4	11.7	119	16.9	34.3
4/22/2004	3.9	19.3	149	15.1	39.6
4/23/2004	3.7	18.2	145	14.4	43.4
4/24/2004	6.6	32.6	170	12.9	50.5
4/25/2004	4.4	20.1	149	12.1	53.3
4/26/2004	3.5	14.3	139	11.0	59.1
4/27/2004	5.4	21.8	134	9.7	64.6
4/28/2004	2.7	9.6	113	9.5	64.1
4/29/2004	1.9	7.4	97	9.1	63.9
4/30/2004	1.6	5.4	82	8.8	63.9
5/1/2004	1	2.6	65	8.3	63.6
5/2/2004	1.2	2.9	60	7.2	67.5
5/3/2004	1.5	5.1	77	6.3	71.9
5/4/2004	2.7	11.8	96	6.5	71.6
5/5/2004	3.7	15.1	119	9.6	59.7
5/6/2004	3.2	12.9	126	12.4	51.2
5/7/2004	4.1	16.1	158	15.6	42.2
5/8/2004	3.6	14.0	161	18.4	33.8
5/9/2004	2.8	12.3	124	19.4	31.3
5/10/2004	2	8.6	102	18.0	34.8
5/11/2004	2	9.2	116	18.4	33.6
5/12/2004	2.5	10.2	124	17.6	35.1
5/13/2004	2.3	10.6	125	16.7	37.8
5/14/2004	3.2	15.5	148	16.1	41.0
5/15/2004	4.2	19.6	165	15.8	42.4
5/16/2004	4.2	19.2	161	15.7	44.1
5/17/2004	4.2	18.3	159	15.3	46.8
5/18/2004	3.1	13.1	143	14.6	48.6
5/19/2004	2.6	10.9	130	14.7	49.3
5/20/2004	2.9	11.0	129	13.9	53.6

5/21/2004	2.8	11.9	119	14.6	51.5
5/22/2004	2.5	8.6	104	12.5	58.9
5/23/2004	1.2	3.7	70	11.5	63.1
5/24/2004	1	2.5	63	11.6	62.2
5/25/2004	0.9	2.4	67	10.4	67.6
5/26/2004	1.4	4.2	84	11.6	62.7
5/27/2004	1.6	6.4	83	12.4	60.0
5/28/2004	2.2	8.6	98	14.5	53.1
5/29/2004	2.8	10.9	114	16.9	46.1
5/30/2004	2.8	10.7	119	19.3	38.3
5/31/2004	2	7.5	104	21.2	31.4
6/1/2004	1.8	7.5	102	21.4	30.2
6/2/2004	1.9	8.2	107	21.9	29.0
6/3/2004	3	11.9	129	22.2	28.4
6/4/2004	2.9	12.0	128	21.3	30.8
6/5/2004	2.5	12.2	121	19.7	36.7
6/6/2004	4.6	20.6	157	18.4	41.7
6/7/2004	5.9	23.1	173	17.6	46.1
6/8/2004	3.4	14.7	146	16.7	49.6
6/9/2004	2.1	9.0	121	16.3	51.0
6/10/2004	2.2	8.8	119	14.7	55.9
6/11/2004	1.8	7.4	99	14.8	54.7
6/12/2004	1.8	6.9	93	14.0	57.0
6/13/2004	1.8	7.0	88	13.4	61.3
6/14/2004	1	3.9	74	11.9	67.4
6/15/2004	1.4	6.4	80	11.4	70.5
6/16/2004	2.2	9.7	89	11.3	70.2
6/17/2004	5.5	25.9	114	12.4	63.9
6/18/2004	8.1	36.7	149	14.8	54.3
6/19/2004	5.8	26.6	157	17.4	45.6
6/20/2004	4.2	20.1	155	19.8	38.5
6/21/2004	3.1	14.1	134	22.0	34.1
6/22/2004	2.9	14.9	119	23.3	32.2
6/23/2004	2.9	15.4	111	23.9	30.0
6/24/2004	2.5	12.1	104	24.4	28.9
6/25/2004	2.3	11.5	99	24.4	29.4
6/26/2004	2.8	14.8	110	23.8	31.3
6/27/2004	6.1	32.1	162	22.5	35.4
6/28/2004	8	39.2	187	20.4	42.5
6/29/2004	6.5	31.0	165	18.3	52.6
6/30/2004	4.2	19.9	145	16.7	57.4
7/1/2004	3.2	15.3	125	15.7	60.2
7/2/2004	1.4	6.9	101	15.3	61.4

7/3/2004	2.1	11.1	103	14.1	65.7
7/4/2004	1.2	5.4	88	14.8	60.6
7/5/2004	0.8	2.8	61	14.8	59.2
7/6/2004	0.6	2.0	52	12.8	63.2
7/7/2004	0.9	3.5	64	11.2	68.5
7/8/2004	1.3	5.1	70	11.0	66.5
7/9/2004	3.4	16.2	97	11.7	63.7
7/10/2004	3.7	19.7	95	13.6	56.3
7/11/2004	5.3	25.1	150	17.8	42.9
7/12/2004	4.1	20.0	162	21.4	33.3
7/13/2004	3.3	18.3	154	24.4	27.4
7/14/2004	4	22.3	161	25.3	26.1
7/15/2004	3.8	20.4	161	25.8	23.2
7/16/2004	2.8	14.6	128	27.0	20.2
7/17/2004	2.9	16.6	129	28.2	18.6
7/18/2004	2.9	15.8	133	28.0	19.1
7/19/2004	3.4	17.8	139	23.9	25.7
7/20/2004	3.9	19.1	137	21.3	34.8
7/21/2004	3.2	15.8	143	20.4	36.7
7/22/2004	5.1	24.9	177	19.0	41.3
7/23/2004	2.6	13.5	138	17.9	45.9
7/24/2004	1.7	9.1	117	16.7	48.9
7/25/2004	1.7	8.6	107	15.5	52.9
7/26/2004	1.2	5.4	90	15.5	51.9
7/27/2004	0.9	4.1	79	14.1	55.6
7/28/2004	0.5	1.6	40	11.8	58.0
7/29/2004	0.5	1.9	53	11.9	57.4
7/30/2004	1.6	7.5	84	9.9	65.2
7/31/2004	4.1	21.4	108	11.1	60.2
8/1/2004	6.6	36.4	127	14.1	50.0
8/2/2004	4.3	21.3	134	17.7	40.1
8/3/2004	2.9	15.4	135	21.1	33.4
8/4/2004	2.5	12.5	142	24.3	28.3
8/5/2004	2.8	15.1	153	25.6	25.6
8/6/2004	2.6	13.7	123	25.9	25.9
8/7/2004	2	10.4	104	26.8	18.7
8/8/2004	2.9	15.2	129	29.3	15.8
8/9/2004	2.5	12.3	114	28.5	14.9
8/10/2004	5	27.0	158	25.9	16.0
8/11/2004	7.6	38.4	194	23.1	26.5
8/12/2004	6.7	35.1	182	20.5	38.2
8/13/2004	5.7	27.2	180	19.1	42.6
8/14/2004	2.8	15.0	136	17.2	44.1

8/15/2004	2.6	15.7	127	16.0	50.9
8/16/2004	2.3	13.0	116	14.8	53.9
8/17/2004	1.4	8.1	107	14.3	55.4
8/18/2004	1	5.5	88	14.8	52.1
8/19/2004	0.6	2.5	57	12.0	58.9
8/20/2004	0.7	3.0	71	10.9	62.1
8/21/2004	1.5	7.7	85	10.6	63.3
8/22/2004	4.7	23.3	124	11.5	60.0
8/23/2004	6.6	35.8	151	14.3	50.6
8/24/2004	4.5	21.3	150	17.8	40.5
8/25/2004	2.8	14.3	152	20.8	34.4
8/26/2004	2.2	12.5	139	23.8	28.2
8/27/2004	2.2	12.2	133	24.2	28.7
8/28/2004	2.3	13.1	126	25.2	24.9
8/29/2004	2.2	14.4	128	27.0	17.8
8/30/2004	2.8	16.8	169	27.1	23.1
8/31/2004	2.7	14.5	149	25.8	23.9
9/1/2004	3.7	21.5	156	23.0	26.8
9/2/2004	5.1	26.4	168	20.7	31.1
9/3/2004	5.1	26.0	176	18.6	36.2
9/4/2004	3.2	14.1	135	16.0	48.4
9/5/2004	2.1	10.3	121	14.5	57.8
9/6/2004	1.7	8.3	99	13.1	64.2
9/7/2004	2	8.9	106	12.0	69.7
9/8/2004	1.6	6.6	96	11.9	71.1
9/9/2004	6.2	32.6	227	21.0	29.2
9/10/2004	7.2	32.0	230	19.5	35.6
9/11/2004	3.9	18.8	194	18.8	37.3
9/12/2004	2.4	12.1	172	17.7	39.7
9/13/2004	2.3	13.0	158	16.7	45.2
9/14/2004	1.9	9.9	144	15.9	47.0
9/15/2004	1.7	8.1	133	15.1	50.2
9/16/2004	0.9	3.8	109	15.3	49.2
9/17/2004	0.7	3.1	94	13.1	56.1
9/18/2004	0.5	2.0	73	13.3	54.4
9/19/2004	0.4	2.2	76	12.4	58.3
9/20/2004	0.4	2.3	74	12.9	56.0
9/21/2004	0.7	4.9	73	11.9	59.6
9/22/2004	3.1	19.8	107	13.4	55.1
9/23/2004	3.4	15.0	115	16.1	46.4
9/24/2004	3.5	18.1	132	17.9	42.2
9/25/2004	3.2	15.3	155	18.2	43.2