Regression analysis for establishing a relation between weather parameters

Group – 13

Jaswanth Krishna Eaga – S20200020257

Kavya Sree Kaitepalli – S20200020270

C.V. Bhanu Prakash – S20200020255

Sayee Sreenivas G B – S20200020259

Kavya Sai Isheka – S20200020314

Date of submission : 23 Nov 2022

Understanding the theory to solve the project problem:

In daily life, weather forecasting is a significant factor of analysis. Given that many businesses, including agriculture, rely heavily on the weather, forecasting the weather is one of the most important aspects. Future planning in agriculture and industry, as well as many other professions including defence, mountaineering, shipping, and aerospace navigation, all require the ability to predict weather conditions. Natural catastrophes are frequently utilised as a warning when sudden climate circumstances change.

Various methods for statistical analysis and weather forecasting have been developed during the past few decades. Regression models are still frequently utilised in these models to estimate future events or values. In order to better handle flaws in numerical forecasts, they are frequently utilised as post-processing techniques (such as model output statistics and perfect programming).

The models are based on observational relationships of the predictand variable with various predictors in a form of a fitting function with unknown parameters to be determined by regression or other optimization methods with the data.

The given climate time series data has 4 parameters mean temperature, humidity, wind speed, mean pressure and independent variable date of 4 years. We need to find the relation between each weather parameters.

We need to create a model that, given a date as an input, can predict the weather parameters temperature, pressure, wind speed, and humidity for that specific day.

We use one of the variables (mean Temp) as input and predict the other three parameters (Humidity, mean Pressure, Wind speed) because we are unfamiliar with working with time series data.

Doing Relational Analysis with Simple Linear and Simple non-linear Regression Analysis approaches, calculate R square values in each case and conclude the results precisely.

Simple Linear Regression

Simple linear regression is used to model the relationship between two continuous variables. Often, the objective is to predict the value of an output variable (or response) based on the value of an input (or predictor) variable.

Formula for finding the slope of the regression line:

$$\mathrm{m} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Formula to calculate the R2 value(Coefficient of determination):

$$R^2 = \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

The R2 value varies from 0(very poor fit) to 1(very good fit).

Simple Non-Linear Regression

It is a method to model a non-linear relationship between the dependent and independent variables. It is used in place when the data shows a curvy trend, and linear regression would not produce very accurate results when compared to non-linear regression.

RESULTS

Data Pre-Processing:

1.DATA CLEANING

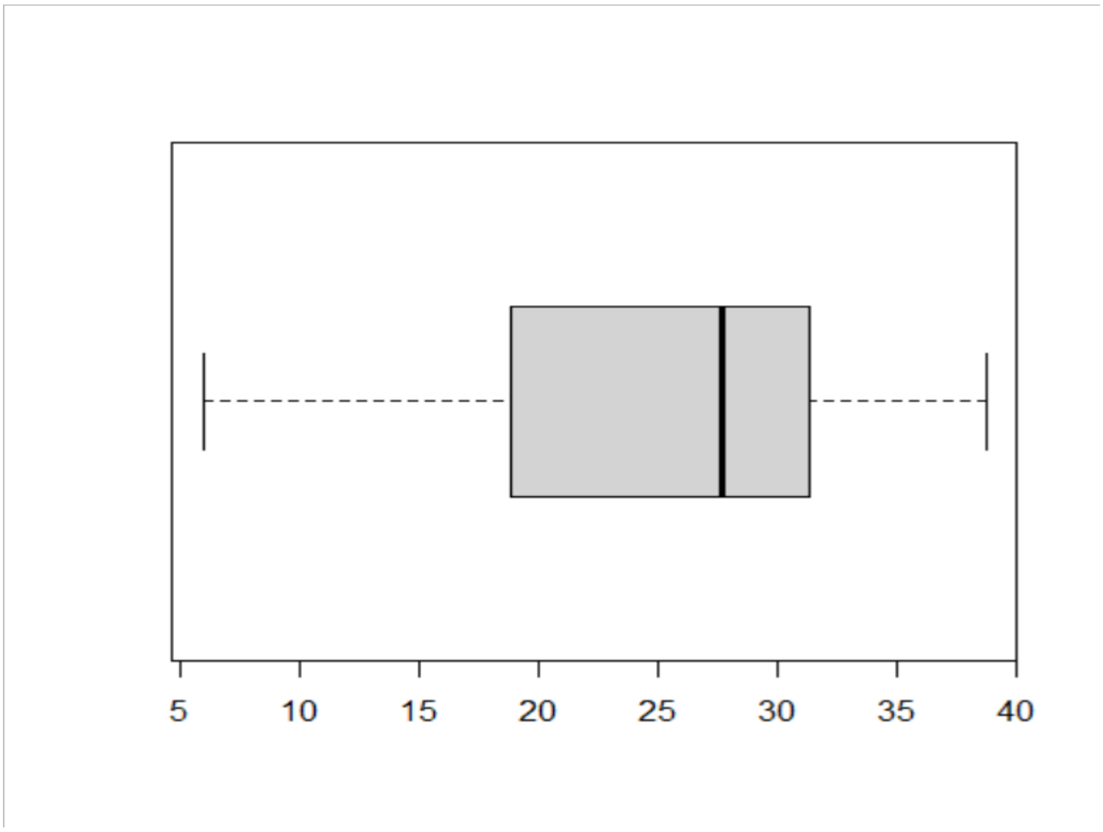
I]Filling Missing Values

On the provided climatic data, we need to do data pre-processing. The null values should be detected, and they will be replaced out with the matching column mean values.

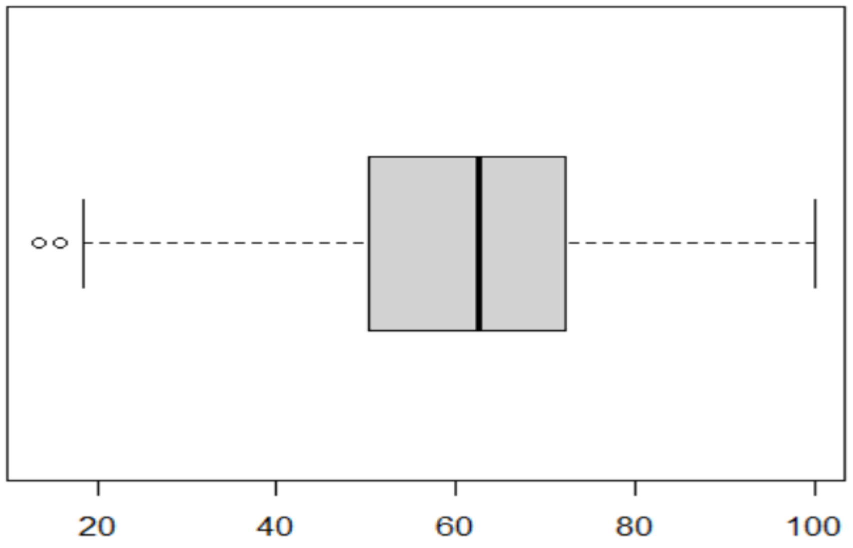
II]Correct Inconsistencies in the Data(outlier detection)

The Box-Plots of weather parameters before removing outliers:

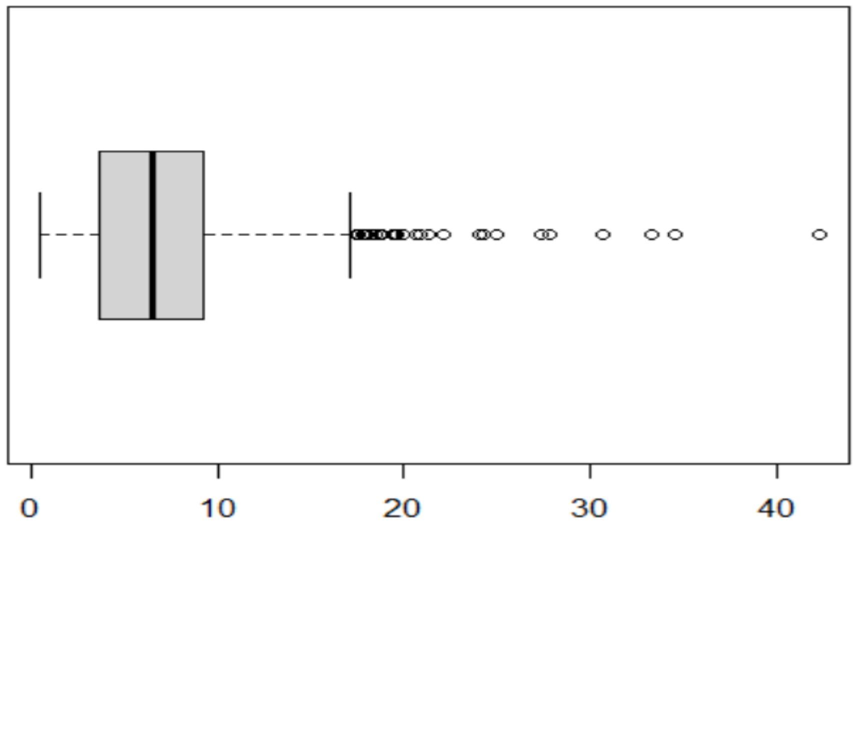
1.Mean Temperature



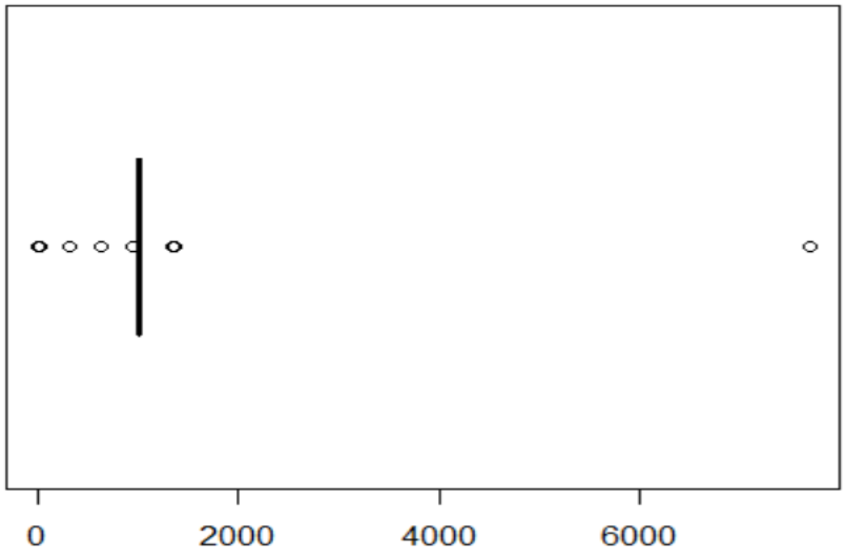
2.Humidity



3.Wind Speed

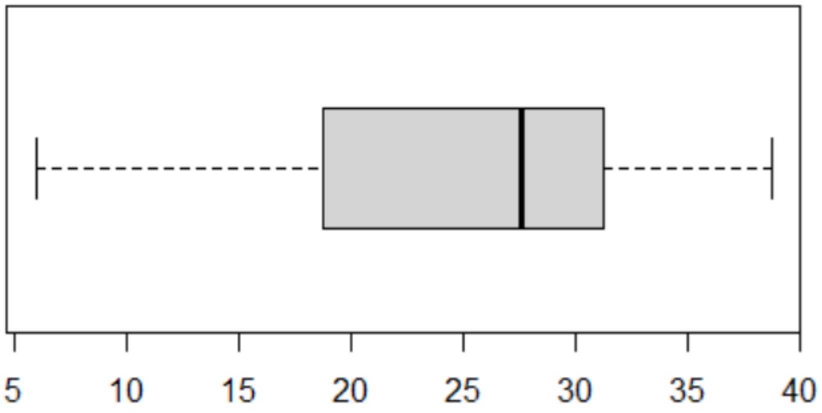


4.Mean Pressure

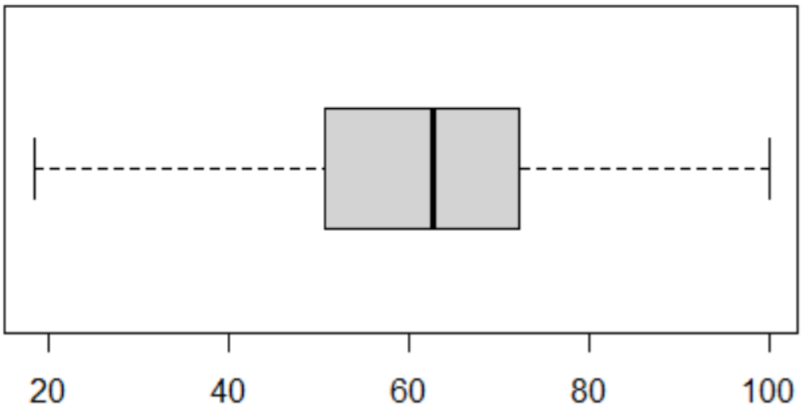


The Box-Plots of weather parameters after removing outliers:

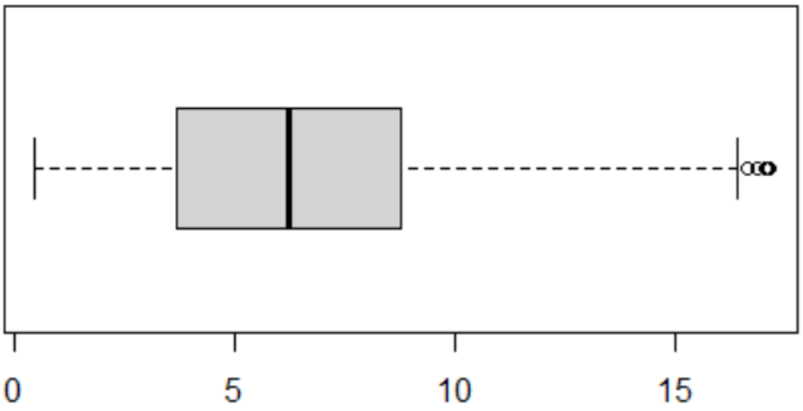
1.Mean Temperature



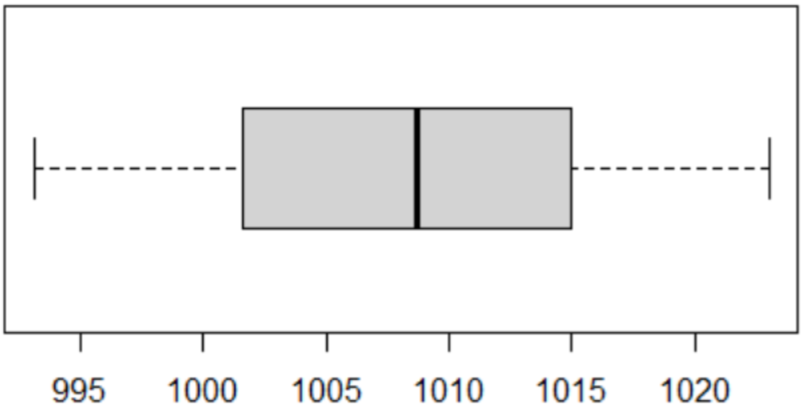
2.Humidity



3.Wind Speed



4.Mean Pressure



II]DATA INTEGRATION

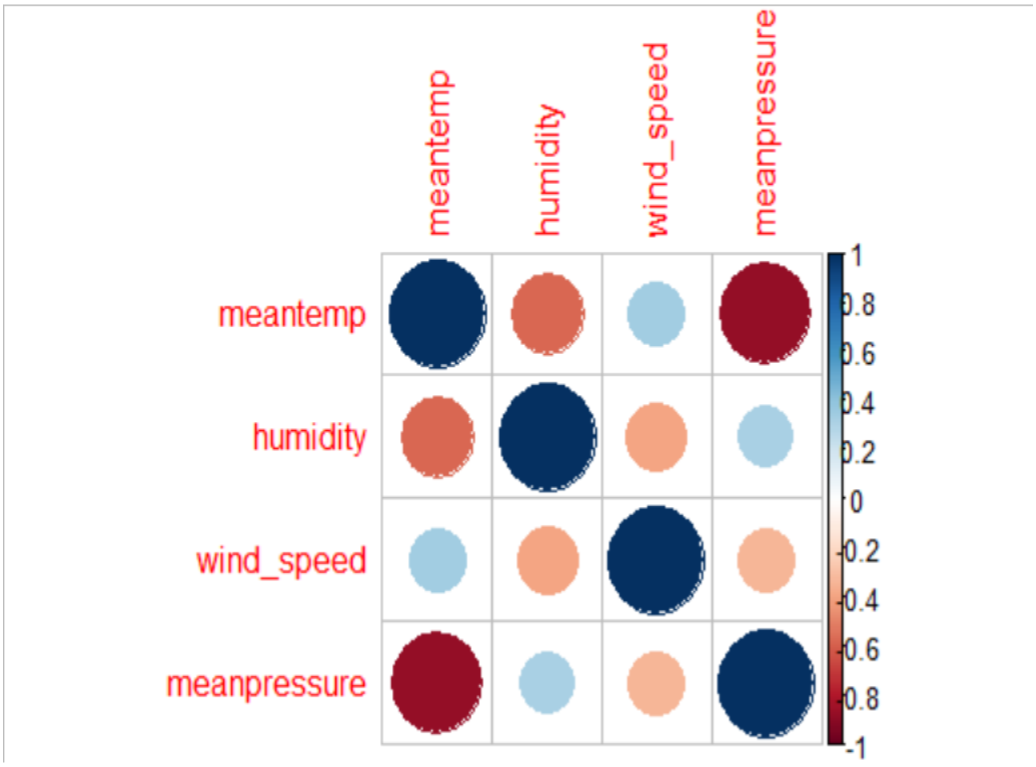
Correlation Analysis

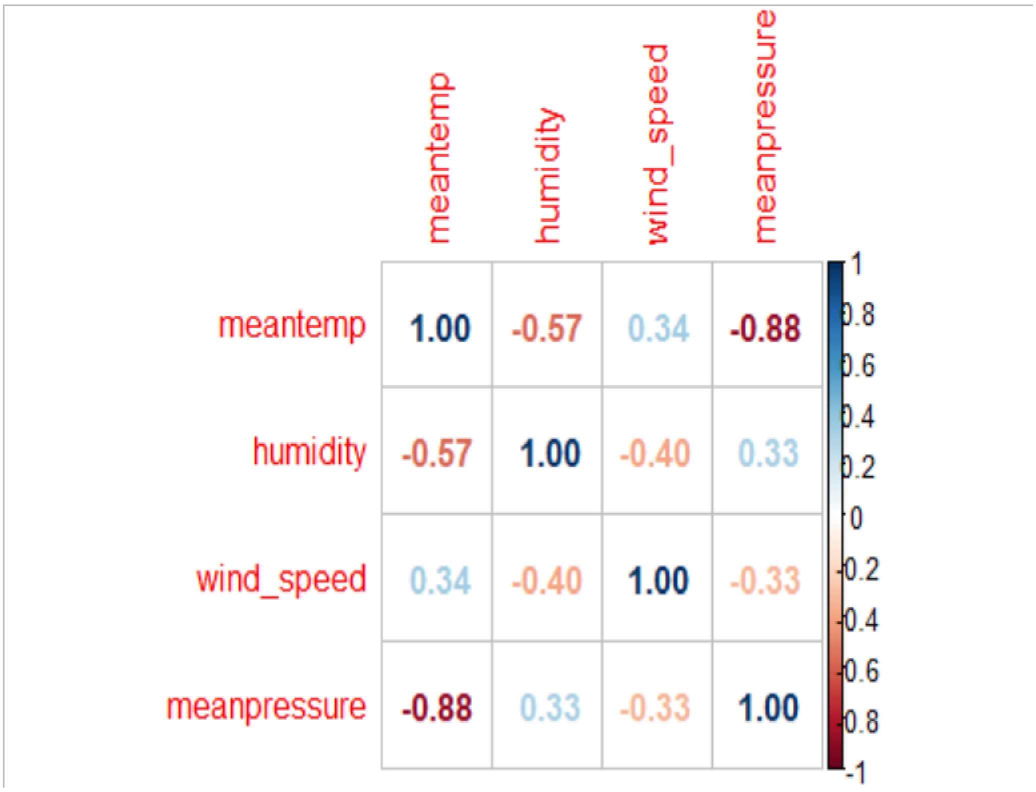
Karl Pearson's correlation coefficient is used to determine the correlation between attributes since these are numerical attributes.

$$r^* = \frac{\sum_{1=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N.\sigma_X.\sigma_Y}$$

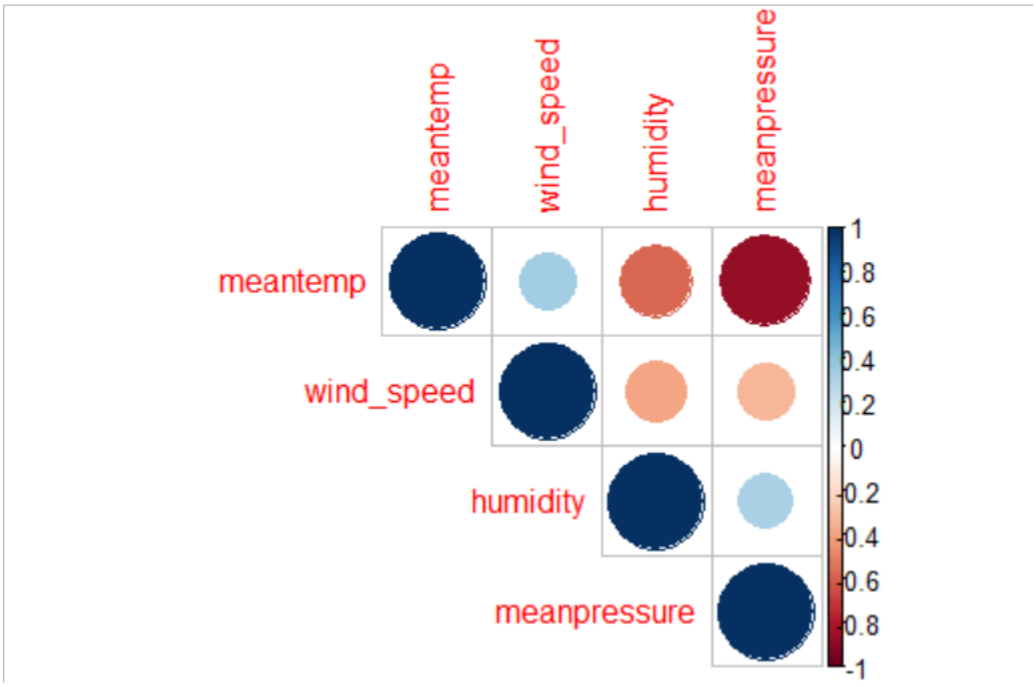
There is a stronger relationship between the attributes if the correlation(r*) value is positive and high, and we can eliminate any one of them because of redundancy.

Correlation Plots:





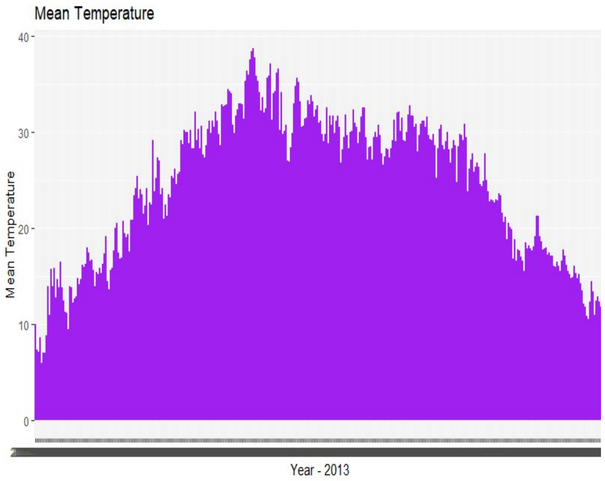
	meantemp	humidity	wind_speed	meanpressure
meantemp	1.00	-0.57	0.34	-0.88
humidity	-0.57	1.00	-0.40	0.33
wind_speed	0.34	-0.40	1.00	-0.33
meanpressure	-0.88	0.33	-0.33	1.00

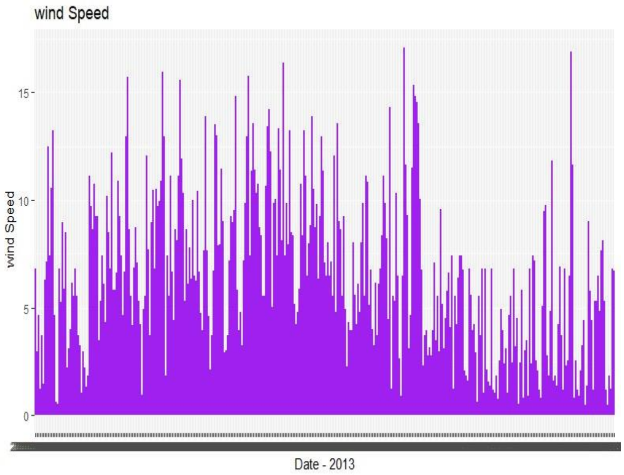


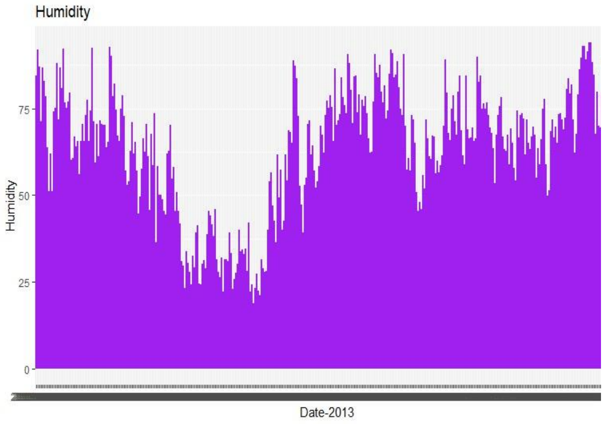
The correlation between mean temperature and mean pressure is -0.88 it means they are negatively correlated. So we have dropped the column.

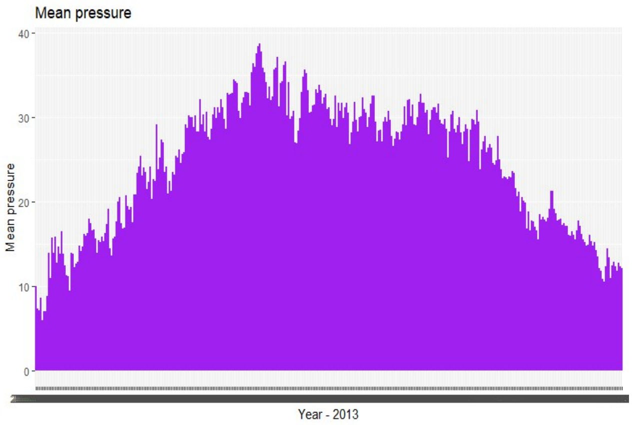
We considered mean temperature as independent variable and humidity as dependent variable and created a Simple Linear Regression model.

2013 Plots of Weather parameters

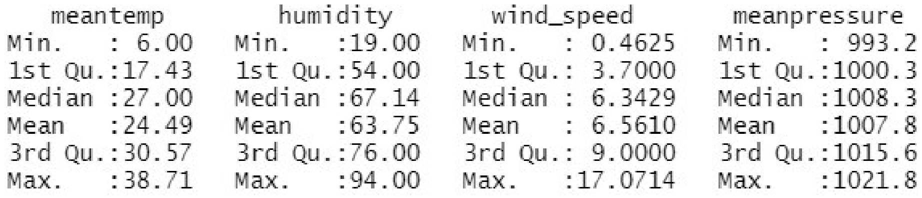






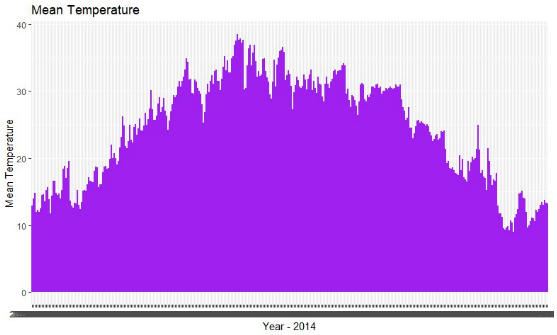


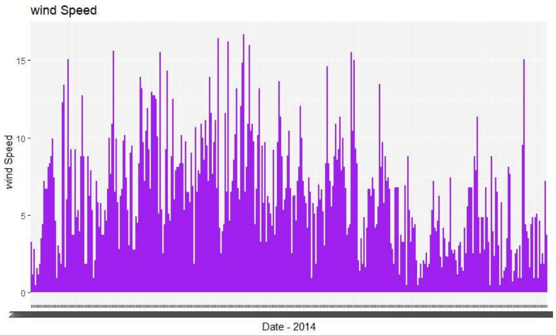
Summary of the plot:

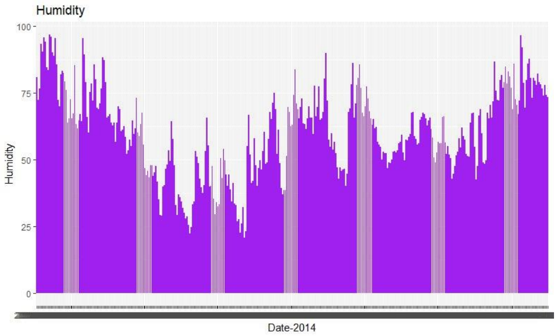


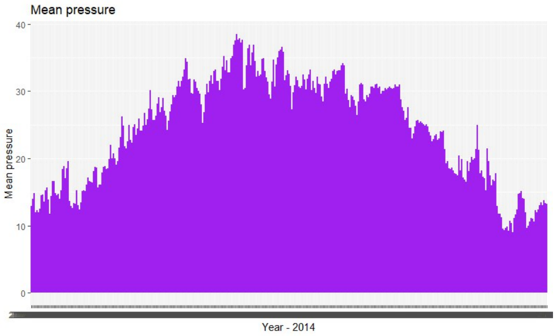
```
    meantemp          humidity         wind_speed         meanpressure
 Min.   : 6.00   Min.   :19.00   Min.   : 0.4625   Min.   : 993.2
 1st Qu.:17.43   1st Qu.:54.00   1st Qu.: 3.7000   1st Qu.:1000.3
 Median :27.00   Median :67.14   Median : 6.3429   Median :1008.3
 Mean   :24.49   Mean   :63.75   Mean   : 6.5610   Mean   :1007.8
 3rd Qu.:30.57   3rd Qu.:76.00   3rd Qu.: 9.0000   3rd Qu.:1015.6
 Max.   :38.71   Max.   :94.00   Max.   :17.0714   Max.   :1021.8
```

2014 Plots of Weather parameters

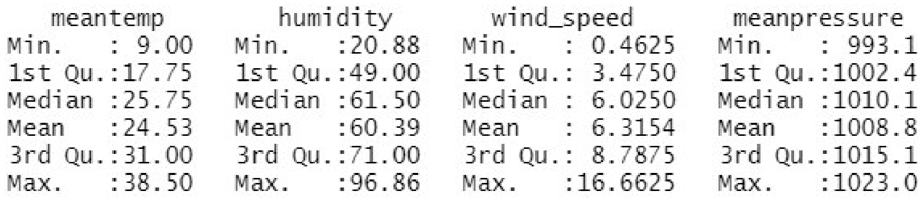






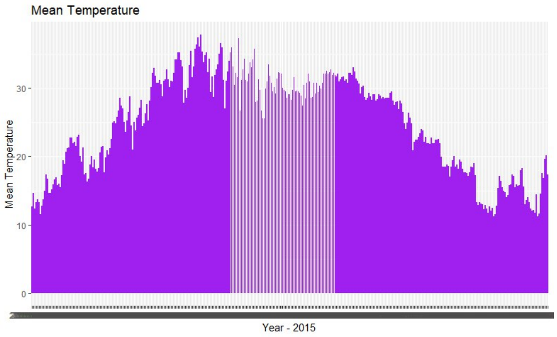


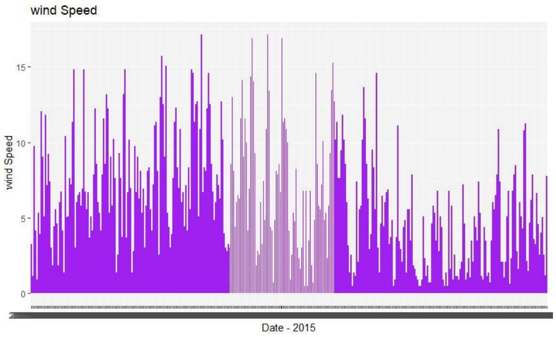
Summary of the plot:

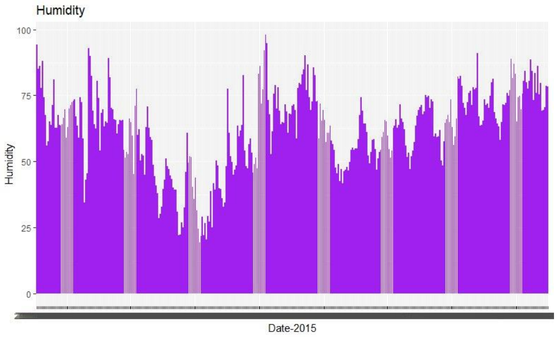


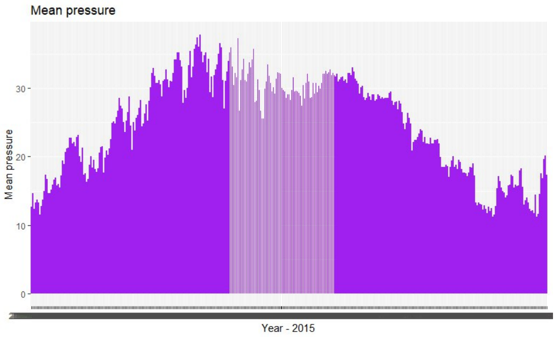
```
    meantemp         humidity         wind_speed         meanpressure
 Min.   : 9.00   Min.   :20.88   Min.   : 0.4625   Min.   : 993.1
 1st Qu.:17.75   1st Qu.:49.00   1st Qu.: 3.4750   1st Qu.:1002.4
 Median :25.75   Median :61.50   Median : 6.0250   Median :1010.1
 Mean   :24.53   Mean   :60.39   Mean   : 6.3154   Mean   :1008.8
 3rd Qu.:31.00   3rd Qu.:71.00   3rd Qu.: 8.7875   3rd Qu.:1015.1
 Max.   :38.50   Max.   :96.86   Max.   :16.6625   Max.   :1023.0
```

2015 Plots of Weather parameters

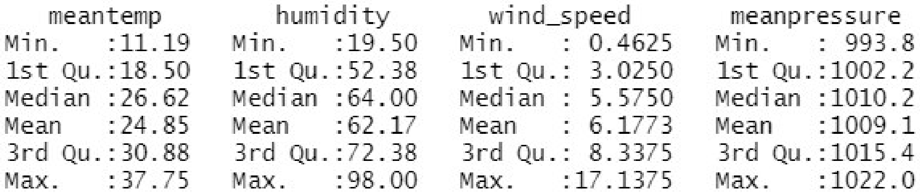






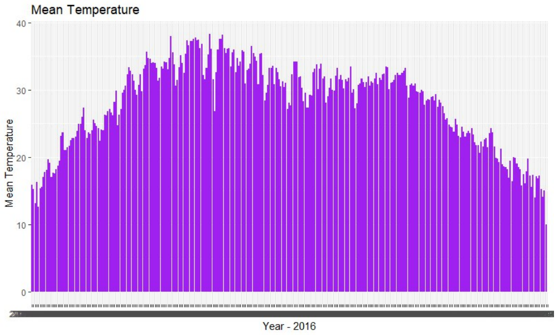


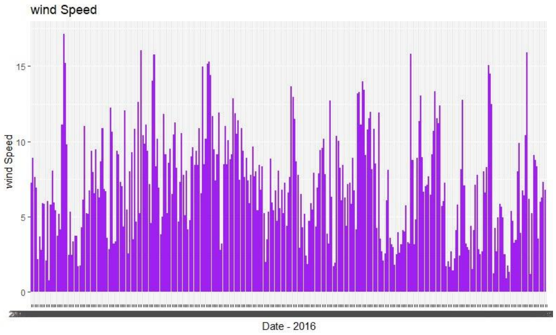
Summary of the plot:

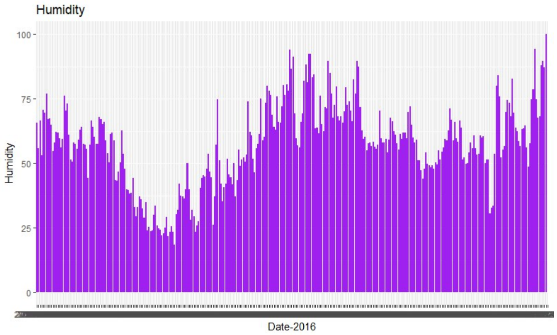


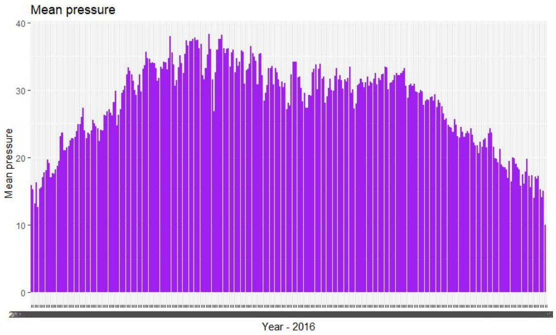
```
    meantemp          humidity         wind_speed         meanpressure
 Min.    :11.19   Min.    :19.50   Min.    : 0.4625   Min.    : 993.8
 1st Qu.:18.50   1st Qu.:52.38   1st Qu.: 3.0250   1st Qu.:1002.2
 Median :26.62   Median :64.00   Median : 5.5750   Median :1010.2
 Mean    :24.85   Mean    :62.17   Mean    : 6.1773   Mean    :1009.1
 3rd Qu.:30.88   3rd Qu.:72.38   3rd Qu.: 8.3375   3rd Qu.:1015.4
 Max.    :37.75   Max.    :98.00   Max.    :17.1375   Max.    :1022.0
```

2016 Plots of Weather parameters

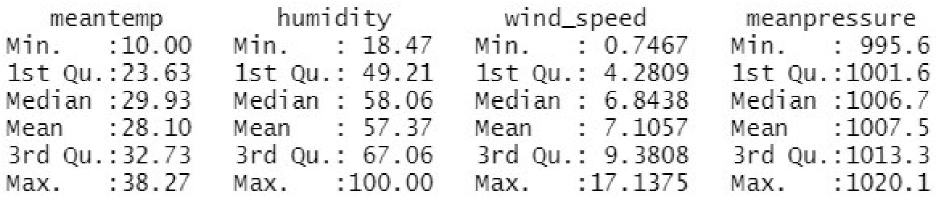








Summary of the plot:

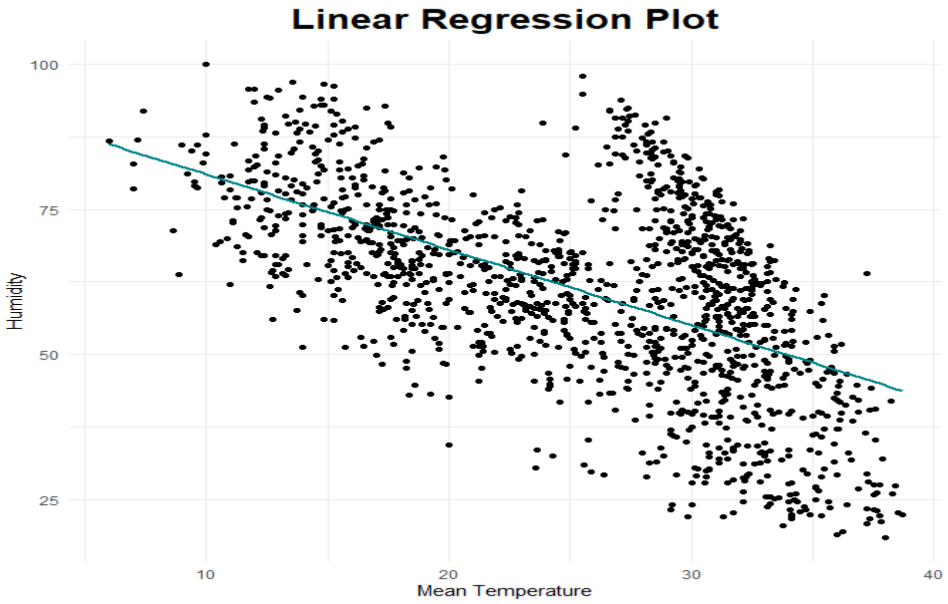


```
    meantemp         humidity          wind_speed         meanpressure
 Min.   :10.00   Min.   : 18.47   Min.   : 0.7467   Min.   : 995.6
 1st Qu.:23.63   1st Qu.: 49.21   1st Qu.: 4.2809   1st Qu.:1001.6
 Median :29.93   Median : 58.06   Median : 6.8438   Median :1006.7
 Mean   :28.10   Mean   : 57.37   Mean   : 7.1057   Mean   :1007.5
 3rd Qu.:32.73   3rd Qu.: 67.06   3rd Qu.: 9.3808   3rd Qu.:1013.3
 Max.   :38.27   Max.   :100.00   Max.   :17.1375   Max.   :1020.1
```

Weather Parameters	Minimum Recorded in	Maximum Recorded in
Mean Temperature	2013	2013
Humidity	2016	2016
Wind Speed	2015	2016
Mean Pressure	2014	2014

III]REGRESSION ANALYSIS

1]Linear Regression Plot of Mean Temperature and Humidity



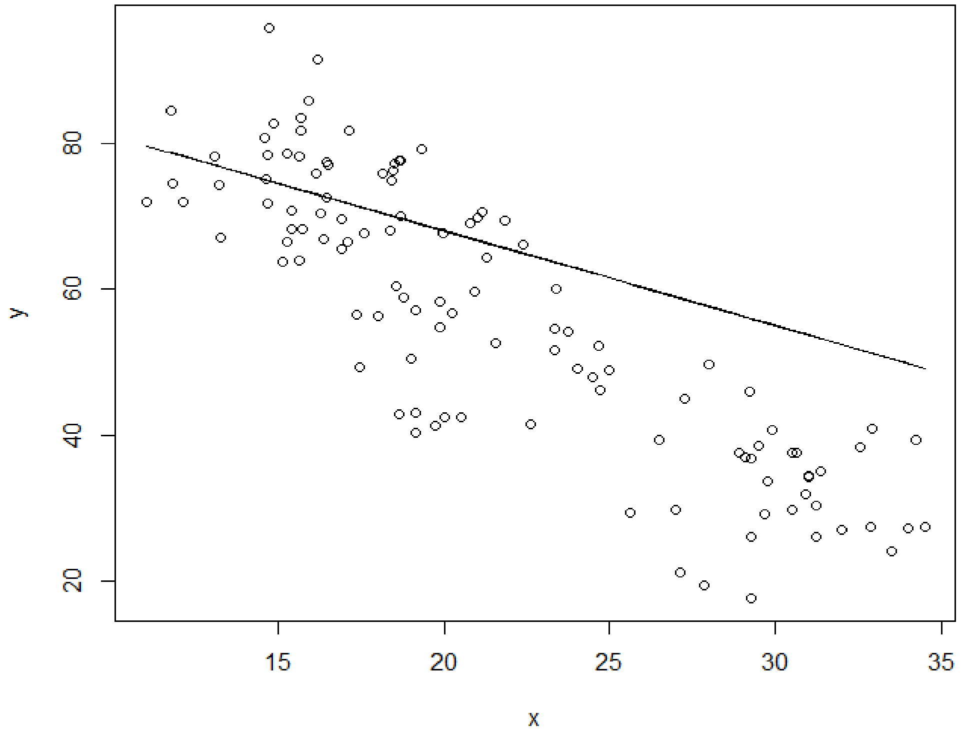
For this plot we used ggplot2 library for the visualization of the regression line, and the colour is turquoise.

1(a)]Simple Linear Regression [using inbuilt lm()]

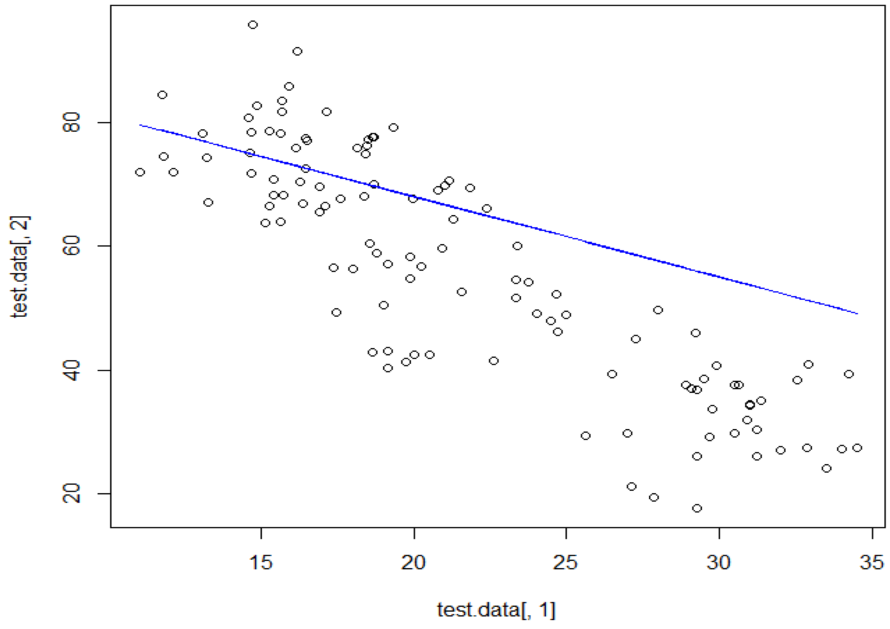
The formula we used for simple linear regression is Humidity ~ Mean temperature. Which means X=Mean temperature and Y=Humidity of training data. With X and Y as inputs for linear regression, we obtained the intercept(c) as 93.992 and slope(m) as -1.297 for the equation Y = m*X + c.

We took X as input from the testing data(mean Temperature) and predicted the Humidity by using the model which we have created with the training data.

The function we used for this is **predict()**. For visualization we have created scattered plot for mean temperature and humidity by using the function **plot()**. For the Regression line we have plotted mean Temperature with the predicted humidity values. The function used for this is **lines().** We got **R square value as 0.2359407**.

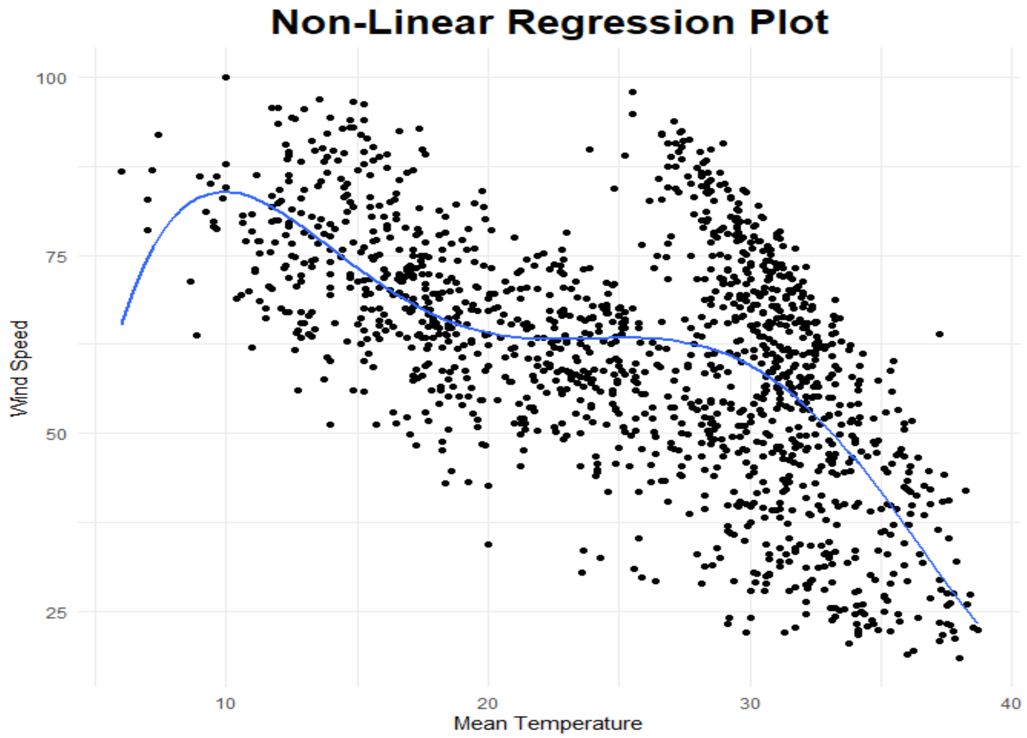


1(b)]Simple Regression Analysis without using any inbuilt functions:



Non-Linear Regression of Mean Temperature and Humidity

For this plot we used ggplot2 library for the visualization of the regression line, with 5 degree polynomial and the colour is turquoise.

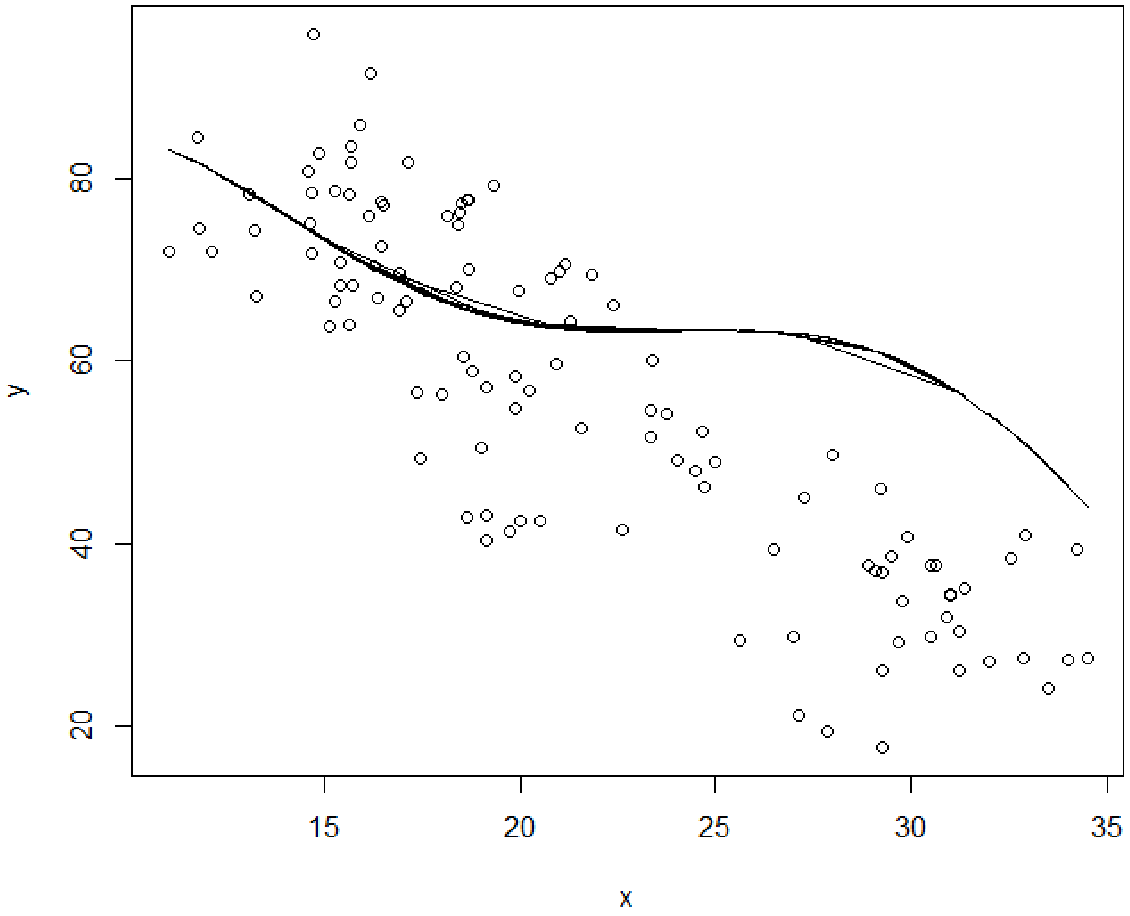


Simple Non-Linear Regression [using inbuilt nls()]

The formula we used for simple linear regression is y~(a*x^5 + b*x^4 + c*x^3 +d*x^2 +e*x +f), a = 5.095e-05,b =-6.327e-03,c= 2.929e-01,d =-6.267e+00,e =6.011e+01,f = -1.252e+02), with X=Mean temperature and Y=Humidity of training data. With X and Y as inputs for non-linear regression.

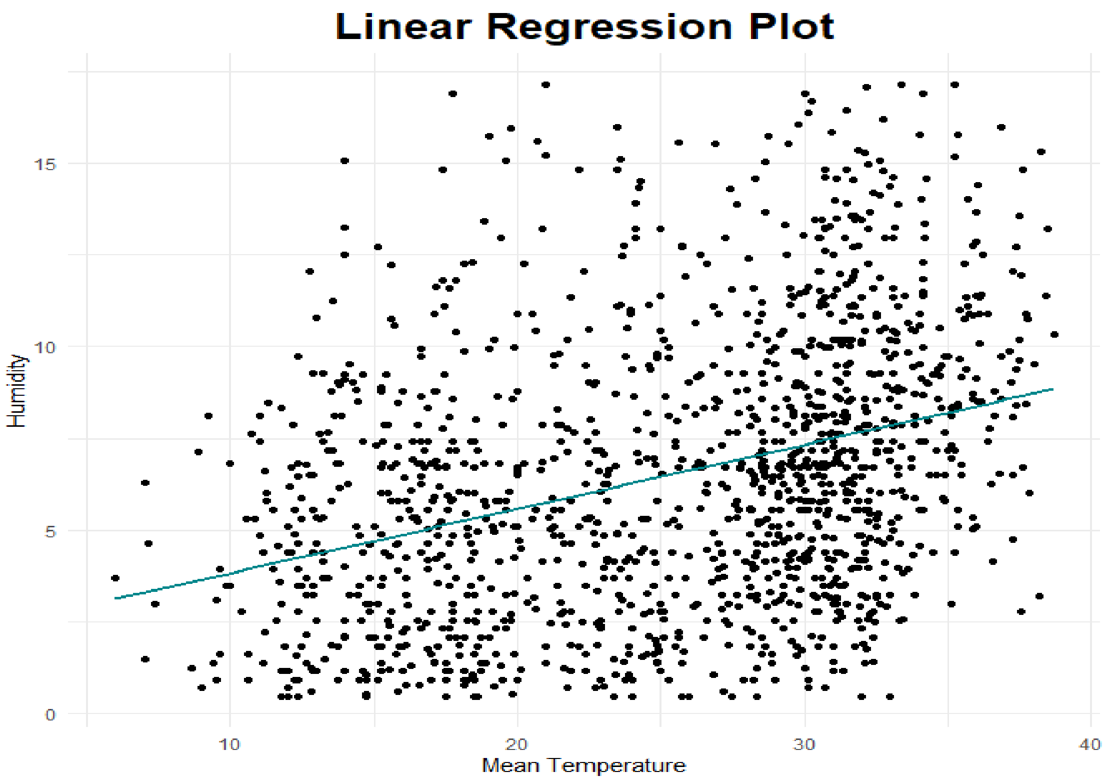
We took X as input from the testing data(mean Temperature) and predicted the Humidity by using the model which we have created with the training data.

The function we used for this is **predict()**. For visualization we have created scattered plot for mean temperature and humidity using function **plot().** For the Regression line we have plotted mean Temperature with the predicted humidity values. The function used for this is **lines().** We got **R square value as 0.4010732**.



2]Linear Regression Plot of Mean Temperature and Wind Speed

For this plot we used ggplot2 library for the visualization of the regression line, and the colour is turquoise.

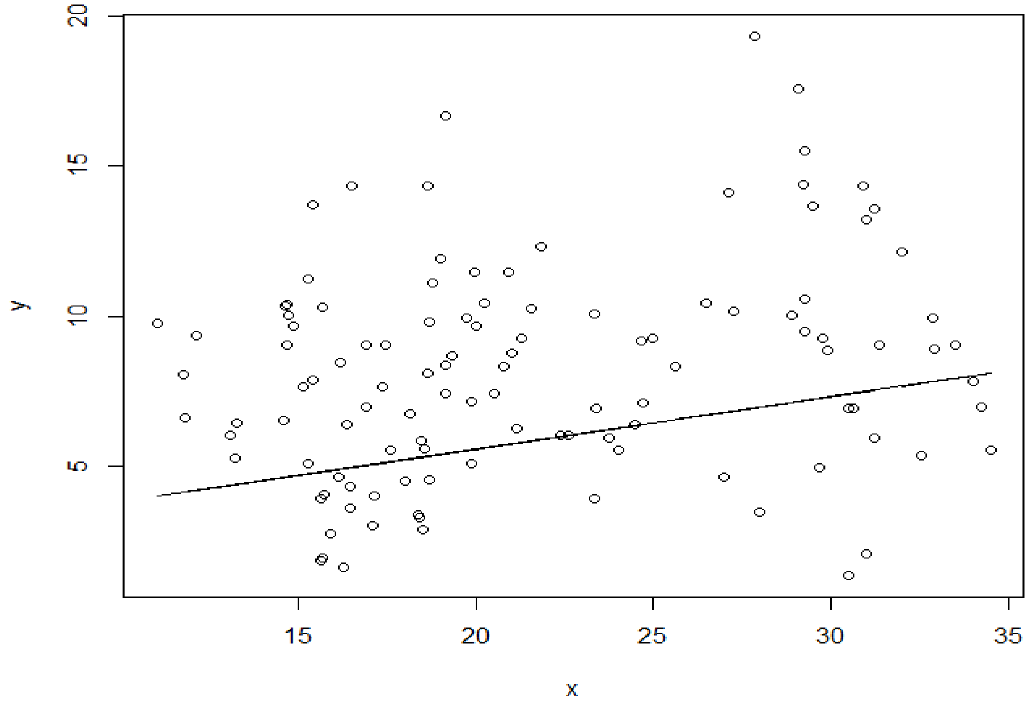


2(a)]Simple Linear Regression [using inbuilt lm()]

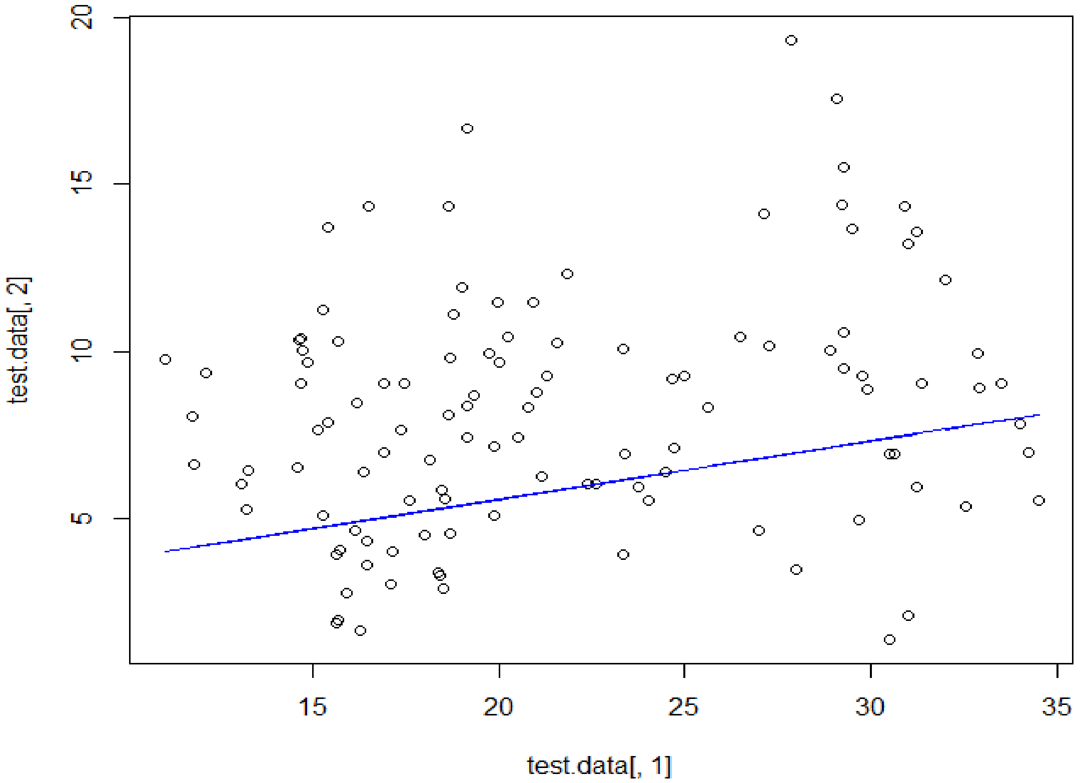
The formula we used for simple linear regression is Wind Speed~Mean temperature. Which means X=Mean temperature and Y=Wind Speed of training data. With X and Y as inputs for linear regression, we obtained the intercept(c) as 2.0819 and slope(m) as 0.1747 for the equation Y=m*X+c.

We took X as input from the testing data(mean Temperature) and predicted the Wind Speed by using the model which we have created with the training data.

The function we used for this is **predict()**. For visualization we have created scattered plot for mean temperature and Wind Speed by using function **plot()**. For the Regression line we have plotted mean Temperature with the predicted Wind Speed values. The function used for this is **lines().** We got **R square value as 0.1187509**.

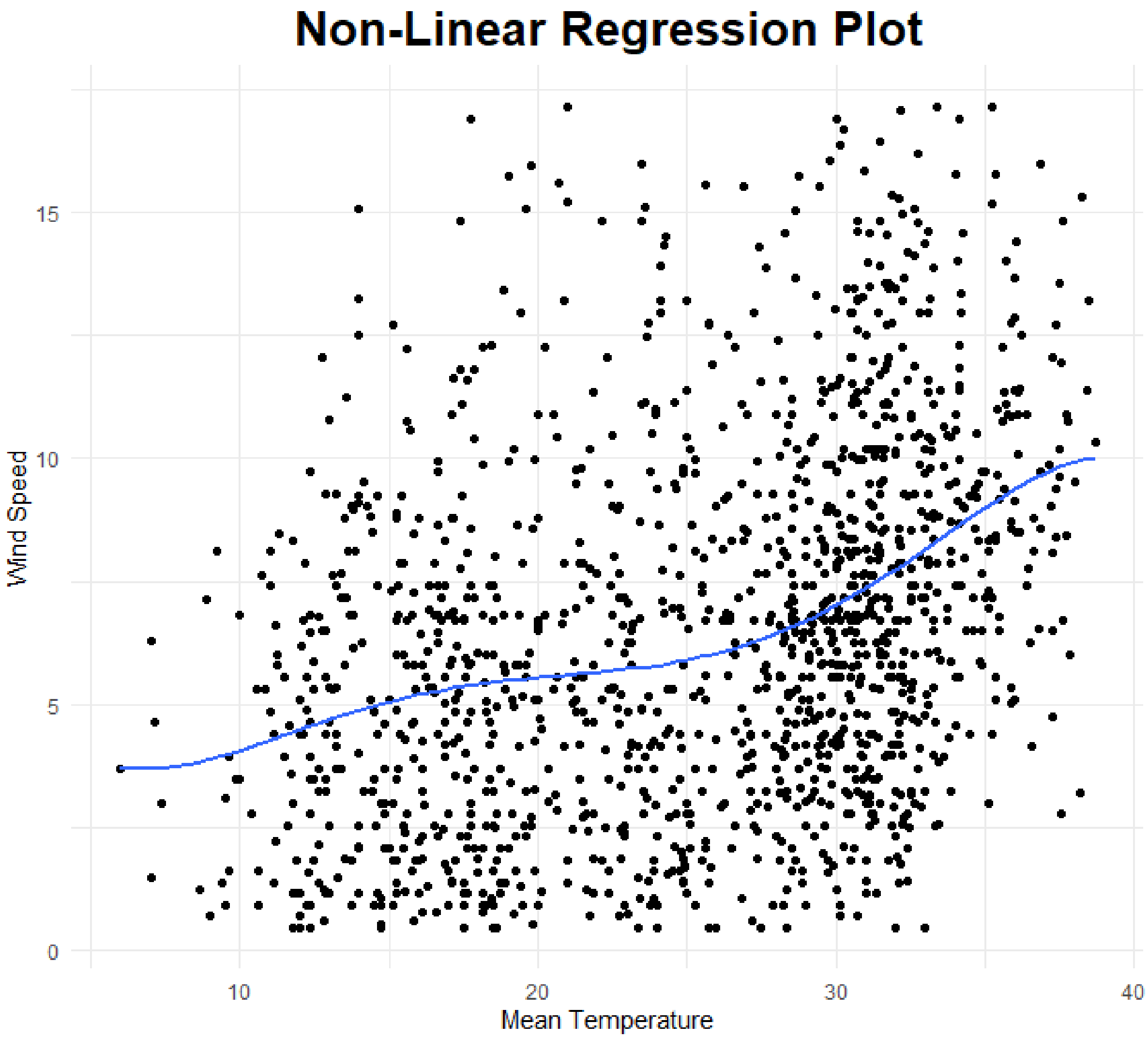


2(b)]Simple Regression Analysis without using any inbuilt functions:



Non-Linear Regression of Mean Temperature and Wind Speed

For this plot we used ggplot2 library for the visualization of the regression line, with 5 degree polynomial and the colour is turquoise.



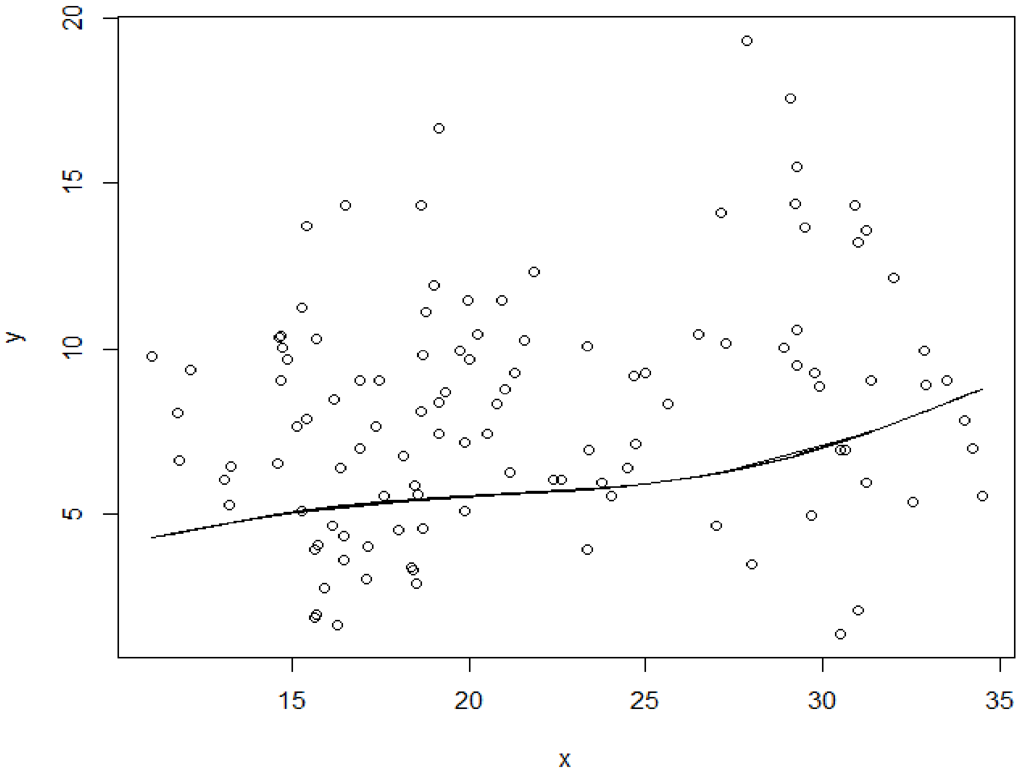
Simple Non-Linear Regression [using inbuilt nls()]

The formula we used for simple linear regression is y~(a*x^5 + b*x^4 + c*x^3 +d*x^2 +e*x +f), a =-3.351e-06 ,b =3.689e-04,c=-1.486e-02,d =2.707e-01 ,e =-2.066e+00,f =9.156e+00), with X=Mean temperature and Y= Wind Speed of training data. With X and Y as inputs for non-linear regression.

We took X as input from the testing data(mean Temperature) and predicted the Wind Speed by using the model which we have created with the training data.

The function we used for this is **predict()**. For visualization we have created scattered plot for mean temperature and Wind Speed using function **plot().** For the Regression line we have plotted mean Temperature with the predicted Wind

Speed values. The function used for this is **lines().** We got **R square value as 0.132098**.



While the regression coefficients and predicted values focus on the mean, R-squared measures the scatter of the data around the regression lines. That's why the four R-squared values are so different. For a given dataset, **higher variability around the regression line** produces a **lower R-squared value.**

A low R-squared value indicates that your independent variable is not explaining much in the variation of your dependent variable - regardless of the variable significance, this is letting us know that the identified independent variable, even though significant, is not accounting for much of the mean of the dependent variable. We may want to look into adding more non-correlated independent variables to your model - variables that some how relate to your dependent variable.

CONCLUSION:

This project was about doing proper data pre-processing and find the relation between each weather parameters in the Daily Climate time series data. The 4 parameters here are mean temperature, humidity, wind speed, mean pressure of 4 years. Doing Relational Analysis with Simple Linear and Simple non-linear Regression Analysis approaches, calculate R square values in each case and conclude the results precisely. Non-linear regression gives better prediction curve than in linear regression.