**“A Time Series Analysis”**

Weather analysis of Delhi over a decade starting from mid 90s till 2016.

**Submitted by**

Mehul Patil (T184099)

Niranjan Patil (T184104)

Pranay Patil (T184115)

Prajwal Warad (T18423)

**Under Supervision of**

Prof. Jayvant H. Devare

Miss. Shobha Mourya



**School of Computer Engineering and Technology**

MIT ACADEMY OF ENGINEERING

December 2018

**Acknowledgments**

Time Analysis of weather is the project we chose for the first semester mini project in the third year of our engineering studies. I think this a very great opportunity to showcase our knowledge in terms of practical knowledge

I would like to show a very great gratitude towards our project guide i.e. Miss. Shobha Mourya Ma’am and Prof. Jayvant H. Devare Sir. They both encouraged and helped us in our very first project. Shobha Ma’am provided us the backend information and the structure part of our project and Devare Sir helped us with the front end of the project.

At the academics level I would also like to thank our Principal Sir and the whole management team which is working in our campus

Finally, I would like to give a special gratitude to our team.

December 2018 Mehul Patil (T184099)

Niranjan Patil (T184104)

Pranay Patil (T184115)

Prajwal Warad (T18423)

**Contents**

**ACKNOWLEDGEMENTS** **1**

**CONTENTS** **2**

**1. INTRODUCTION** **4**

OVERVIEW 4

BACKGROUND AND MOTIVATION 5

OBJECTIVE 6

METHODOLOGY 6

**2. DATA SET** **7**

DESCRIPTION 7

NUMBER OF ATTRIBUTES USED 7

**3. SYSTEM WORKING** **8**

BLOCK DIAGRAM……………………………………………………………………………………………………….......**8**

SAMPLE CODE……………………………………………………………………………………………8

SCREENSHOT RESULT…………………………………………………………………………………….10

**4. CONCLUSION** **15**

**5. FUTURE WORK**  **16**

**6. REFERENCES** **17**

**1. INTRODUCTION**

**OVERVIEW**

A time series is a sequence of observations collected at some time intervals. Time plays an important role here. The observations collected are dependent on the time at which it is collected.

Or we can say that:

Any metric that is measured over regular time intervals forms a time series. Analysis of time series is commercially importance because of industrial need and relevance especially w.r.t forecasting (demand, sales, supply etc).

The time series model can be done by:

* The understanding of the underlying forces and structures that produced the observed data is done.
* Start to fit a model and also start to forecasting, monitoring or even feedback and feed forward control is done.

Several data mining techniques have been employed in diversified applications such as predicting rainfall, weather, storms and flood. Weather forecasting falls under predictive mining which focuses on the data analysis, formulates the database, and forecasts the features of anonymous data. This research work focuses on solving the weather prediction anomalies and in-efficiency based on linear regression algorithms through normal equation model. The major contribution of this research work is to formulate an efficient weather prediction model based on the linear regression algorithms.

Weather forecasting means predicting the weather conditions (conditions of atmosphere) of a particular given area or location. More importantly, accurate weather prediction is very important to pursue day-to-day activities. Living and non-living things are dependent on weather predictions. Even after decades of weather forecasting, the weather industry in India is still in its initial stage, facing many obstacles. One of the major obstacles that weather forecasting faces is the arbitrary & ill-suited expectations from the nature and global warming.

**BACKGROUND AND MOTIVATION**

Currently the entire world’s population is believed to be suffering from harsh climatic conditions. Climatic changes have the ability to influence nature and thus threaten humans in different aspects of life economically, socially and politically. Despite the fact that the climatic changes pose a threat to the entire globe, many people believe that developing countries (like India) in the tropical regions of the world will be impacted more severely than developed ones. The most influential factors in the climate are temperature and moisture.

The gradual rise in the mean temperature of the Earth’s atmosphere and its oceans is referred to as Global warming. It is widely believed that the changing temperature due to global warming is permanently changing the entire Earth’s climate. For a long time the biggest debate in a number of local and international forums worldwide has been whether global warming is real.

Some people think that global warming is not real. However several climate scientists have carried out researches and have come to a conclusion that the globe is gradually warming. People perceive the impacts of global warming differently with some taking the necessary precautions to help reduce the rates of the rising temperatures. In the past century alone, studies have shown that the globe’s mean temperature has risen by between 0.4°C and 0.8 °C.

The prediction of weather conditions can have significant impacts on various sectors of society in different parts of the country. Forecasts are used by government and industry to protect life and property and to improve the efficiency of operations, and by individuals to plan a wide range of daily activities. The notable improvement in forecast accuracy has been achieved since the 1950s, that is, a direct outgrowth of technological developments, basic and applied research, and the application of new knowledge and methods by weather forecasters. The advance knowledge of weather parameters in a particular region is advantageous in effective planning.

Several studies on forecasting weather variables based on time series data in reference to a particular region have been carried out at national and international level both in the farm and nonfarm sectors. It was observed that the combination of two or more computational models/ hybrid models decomposes a time series into linear and non-linear form and prove to be better approach in comparison to single models for the reason that hybrid model produces small forecasting error in terms of accuracy. In contrary some of the studies also mentioned that hybrid approaches are not always better. Such uncertainty in weather forecasting models open up new opportunities for the selection of precise forecasting model. These aspects motivate this thesis, to explore the existing opportunities to identify the precise weather forecasting model. Predictions of weather parameters provide by such identify models based on time series data will be of particular interest to the weather forecasters.

**OBJECTIVE**

The advance knowledge of weather parameters in a particular region is very helpful in sound planning. A reliable prediction of Indian monsoon in a region on seasonal and inter seasonal time series is not only scientifically challenging but also important for the future planning. The role of statistical techniques for predicting the weather parameters at a particular place and time depends on an understanding of the past time series data. The transient behavior of weather parameters over a particular period of time makes difficult to predict correctly and consistently.

Keeping in view a comparative study of weather forecasting models and to propose hybrid model for seasonal and inter-seasonal time series data is planned with the following objectives:

1. To study the distribution pattern of weather parameters.
2. Development of weather forecasting models.
3. To compare the predictive ability of the developed model.
4. To identify the precise and reliable weather forecasting model.

**METHODOLOGY**

In a developing country and an economy like India where major population is dependent on agriculture, weather conditions play an important and vital role in economic growth of the overall nation. So, weather prediction should be more precise and accurate. The data used in this research is for the years 1996 to 2016. This research uses software called ‘R-studio’ for programming. The programming language used is ‘R-language’. The following diagram visualizes the system in the form of a block diagram.

Three types of weather parameters are predicted: Temperature, Humidity and Dew-point, etc. Temperature is the measure of hotness or coldness, generally measured using thermometer. Units of temperature most frequently used are Celsius and Fahrenheit. Humidity is the quantity of water vapor present in the atmosphere. It is a relative quantity. Dew point is the temperature of the atmosphere (which varies according to pressure and humidity) below which water droplets begin to condense and dew is formed.

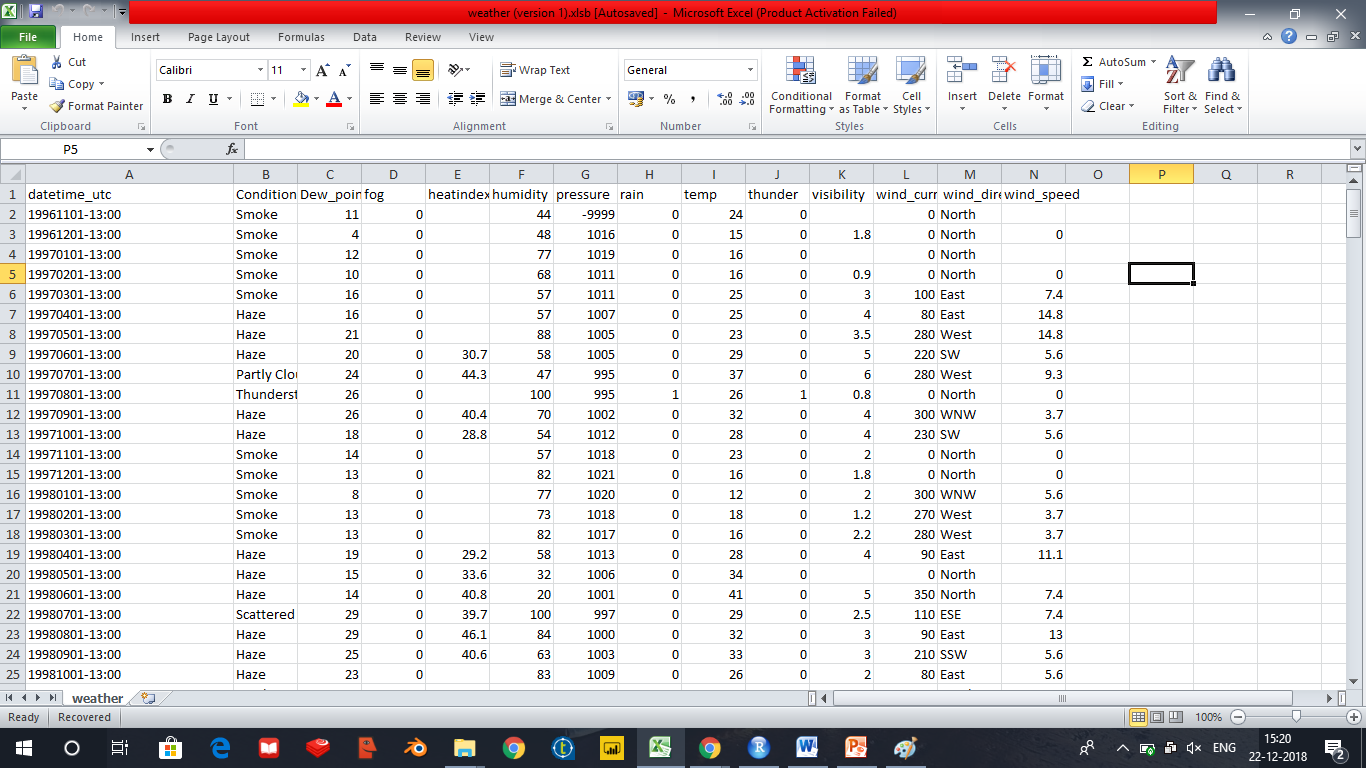
**2. DATA SET**

**DESCRIPTION**

The data set used here is a CSV file containing the weather observations, such as conditions of the day, heat index, humidity etc, recorded every first day of the months starting from November 1996 to December 2016 in the Union Territory of Delhi.

The observations are all recorded at 13:00 hrs in the afternoon.

Following is a sample from the dataset we have used :



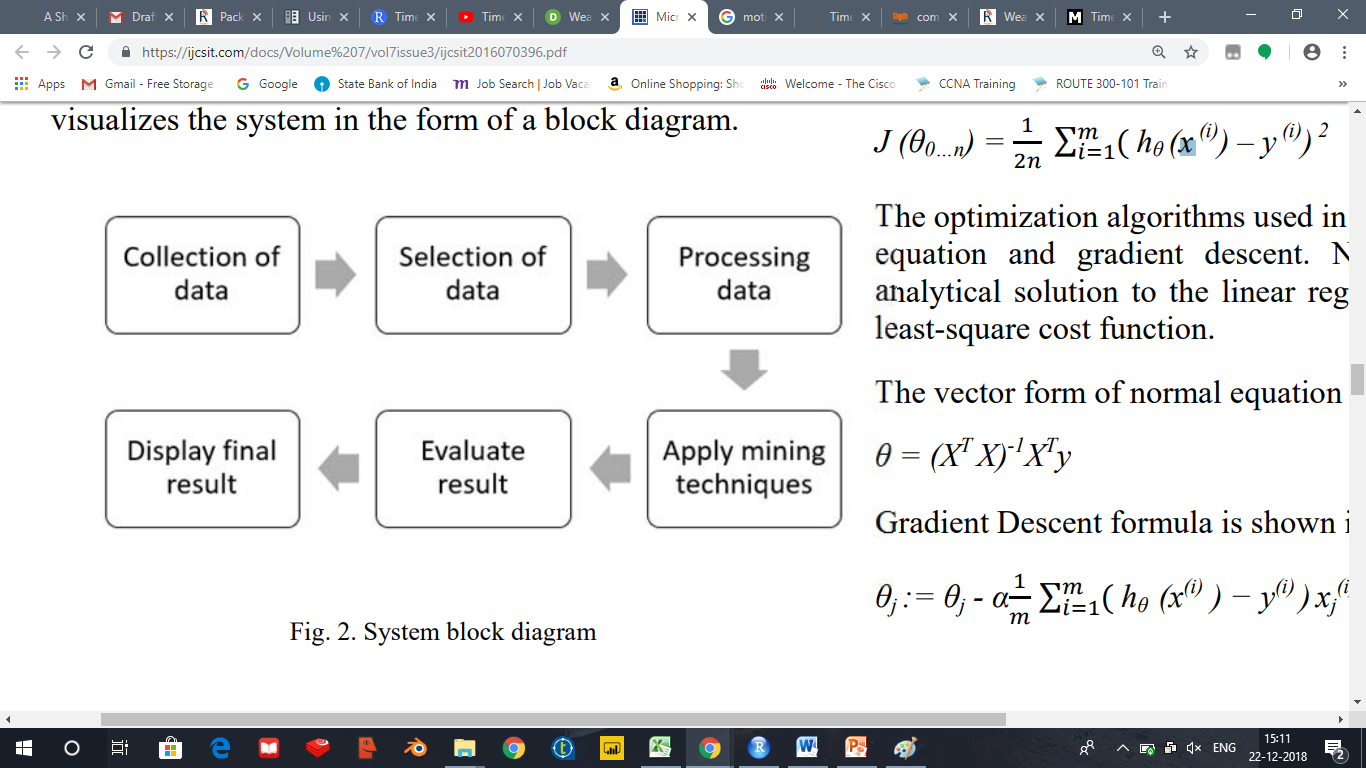
**NUMBER OF ATTRIBUTES USED**

Following are the attributes used:

Condition, Dew Point, Fog, Humidity, Heat Index, Pressure, Rain, Temperature, Thunder, Visibility, Wind Current, Wind Direction and Wind Speed.

**3.SYSTEM WORKNG**

**BLOCK DIAGRAM**



**CODE:**

getwd()

options(repos = c(CRAN ="http://cran.rstudio.com"))

install.packages("dplyr")

install.packages("ggplot2")

install.packages("stringr")

install.packages("imputeTS")

install.packages("fpp")

install.packages("lubridate")

install.packages("knitr")

install.packages("caret")

install.packages("gmodels")

install.packages("lattice")

install.packages("gridExtra")

install.packages("kmisc")

install.packages("ROCR")

install.packages("corrplot")

library(dplyr)

library(ggplot2)

library(stringr)

library(imputeTS)

library(fpp)

library(lubridate)

library(knitr)

library(caret)

library(gmodels)

library(lattice)

library(gridExtra)

library(kmisc)

library(ROCR)

library(corrplot)

weather\_org = read.csv("F:/desktop/weather.csv", header = TRUE, sep = "," ,stringsAsFactors = TRUE)

kable(head(weather\_org))

class(weather\_org)

colnames(weather\_org)

str(weather\_org)

c(as.character(weather\_org$datetime\_utc[1]))

as.character(weater\_org$datetime\_utc)

all.equal(weather\_org$thunder > 1,

weather\_org$rain == "Yes")

weather\_org2 = subset(weater\_org, select = -c(datetime\_utc,rain,pressure))

colnames(weather\_org2)

(cols\_withNa <- apply(weather\_org2, 2, function(x)

sum(is.na(x))))

weather\_org3 = weather\_org2[complete.cases(weather\_org2),]

factor\_vars = names(which(sapply(weather\_org3, class) == "factor"))

factor\_vars = setdiff(factor\_vars, "heatindex")

chisq\_test\_res = lapply(factor\_vars, function(x) {

chisq.test(weather\_org3[,x], weather\_org3[, "heatindex"], simulate.p.value = TRUE)

})

names(chisq\_test\_res) = factor\_vars

chisq\_test\_res

weather\_org5 = weather\_org2[complete.cases(weather\_org2),]

colnames(weather\_org5)

factor\_vars = names(which(sapply(weather\_org5, class) == "factor"))

numeric\_vars = setdiff(colnames(weather\_org5), factor\_vars)

numeric\_vars = setdiff(numeric\_vars, "thunder")

numeric\_vars

numeric\_vars\_mat = as.matrix(weather\_org5[, numeric\_vars, drop=FALSE])

numeric\_vars\_cor = cor(numeric\_vars\_mat)

corrplot(numeric\_vars\_cor)

pairs(weather\_org5[,numeric\_vars], col=weather\_org5$thunder)

args(ts)

#converting into time series

weather.timeseries <- ts(weather\_org,start = c(1997),frequency = 12)

weather.timeseries

colnames(weather.timeseries)

class(weather.timeseries)

#checking for na values

is.na(weather.timeseries)

#impute the missing values with na.mean using option median

na.mean(weather.timeseries,option = "median")

#ploting histogram or high-density vertical lines

plot(x = pressure,type = "h")

#ploting overplotted

plot(x = pressure,type = "o", col = "red")

#plot(weather.timeseries[10:20],main = 'temp')

plot(weather.timeseries[100:200],ylab = "heatindex",xlab = "Dew\_point",main = "Conditions",type = 'h')

plot(weather.timeseries[100:300],type = "o",col = "red", lty = "dashed")

#prints cycles across the years

cycle(weather.timeseries)

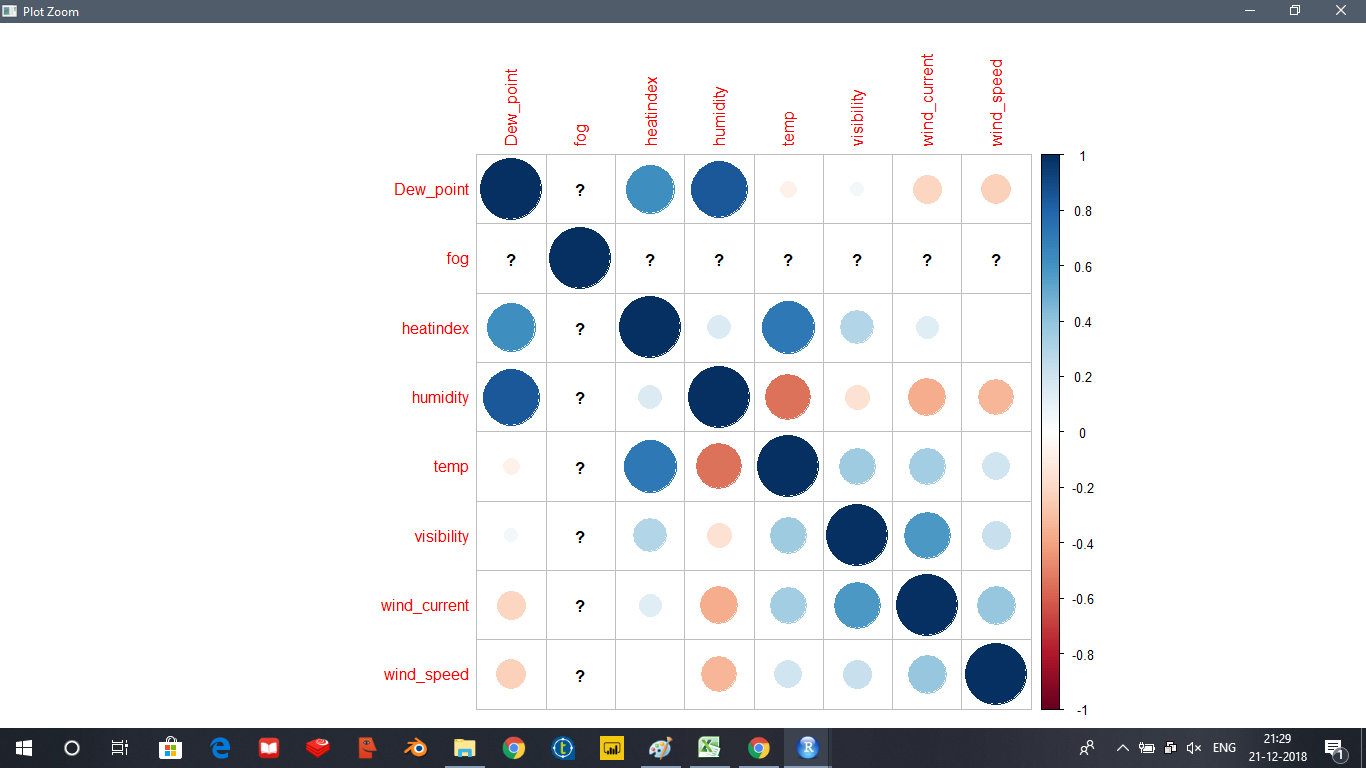
#Boxplot

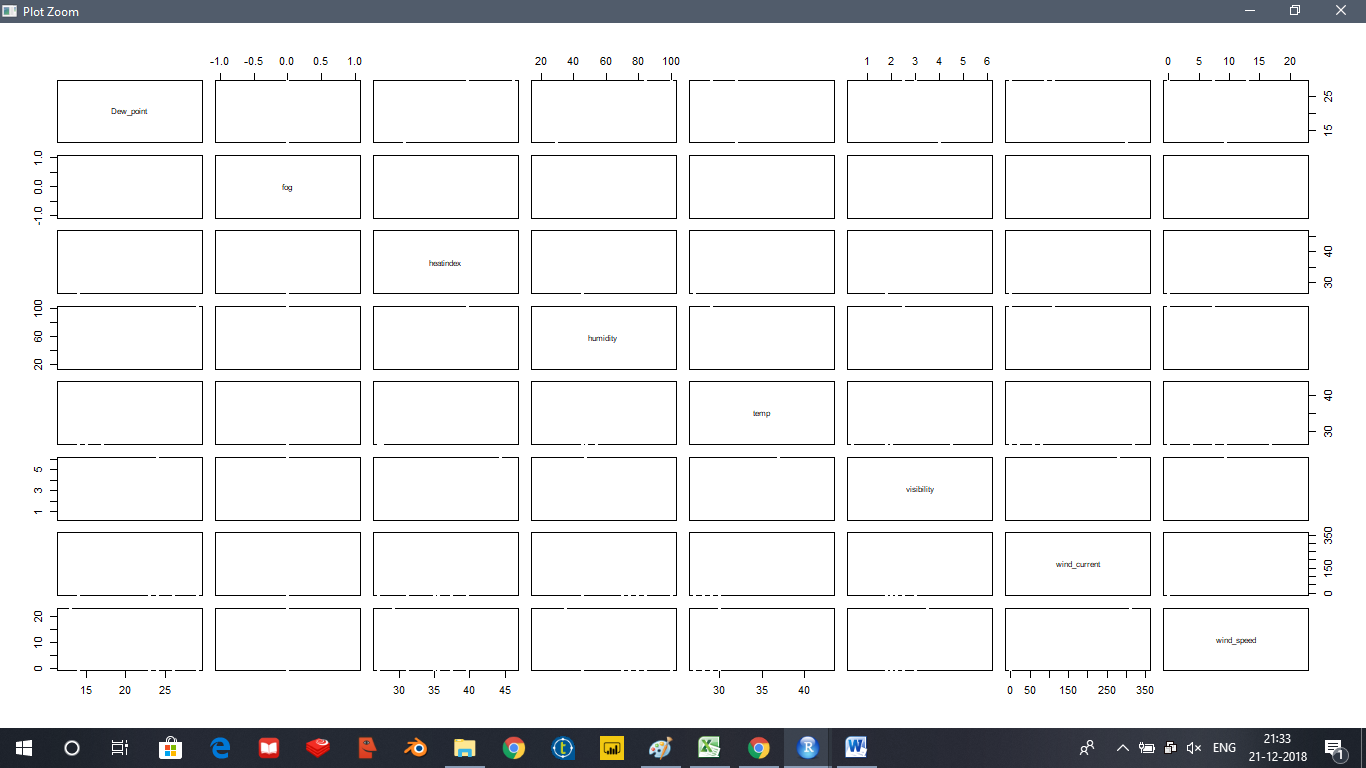
boxplot(weather.timeseries~cycle(weather.timeseries))

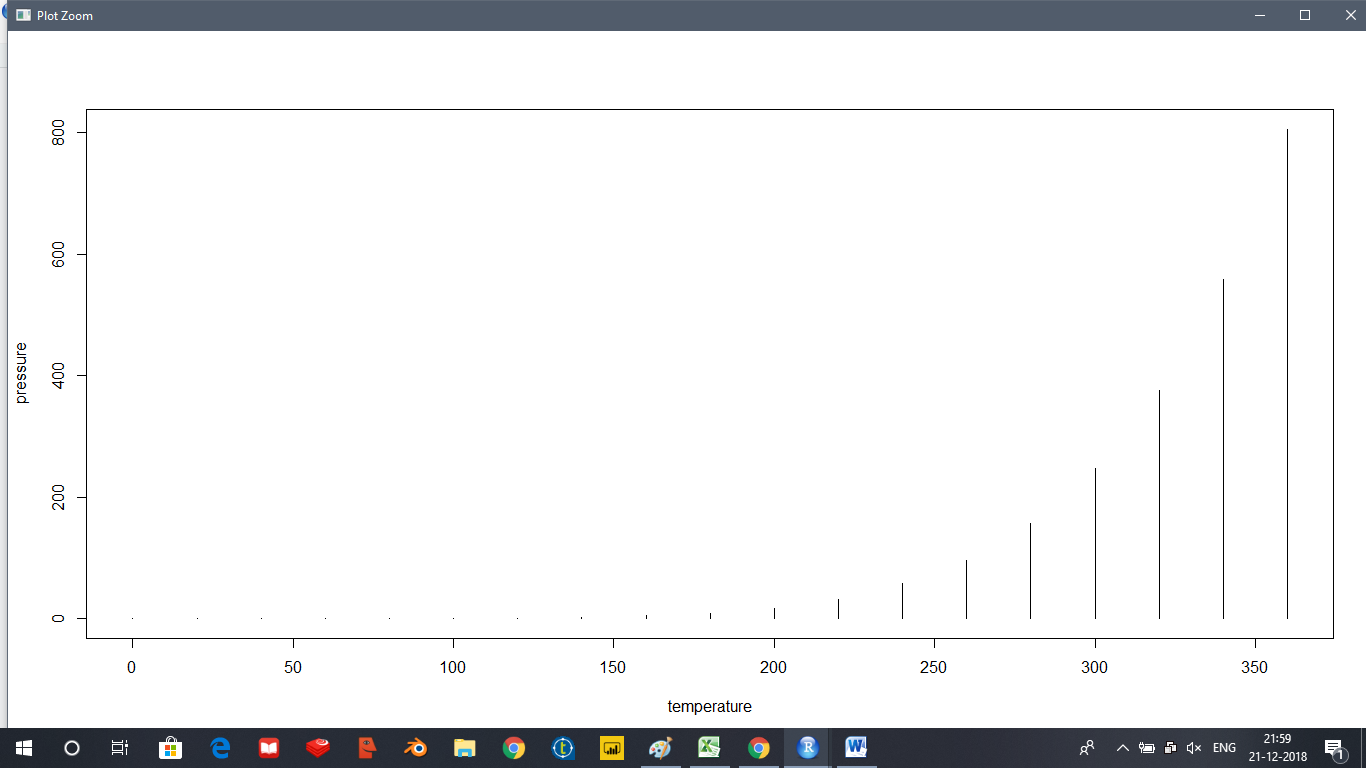
#forecasting

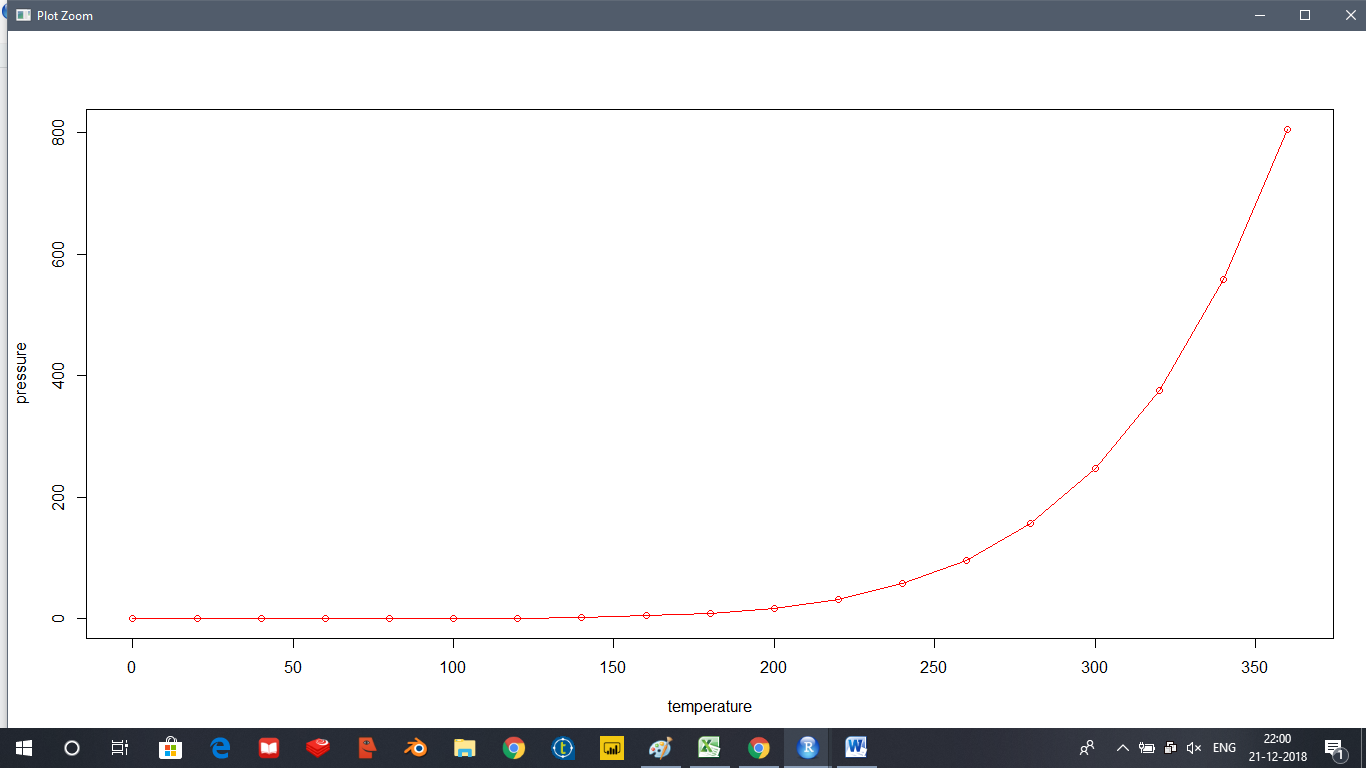
plot(forecast(weather.timeseries))

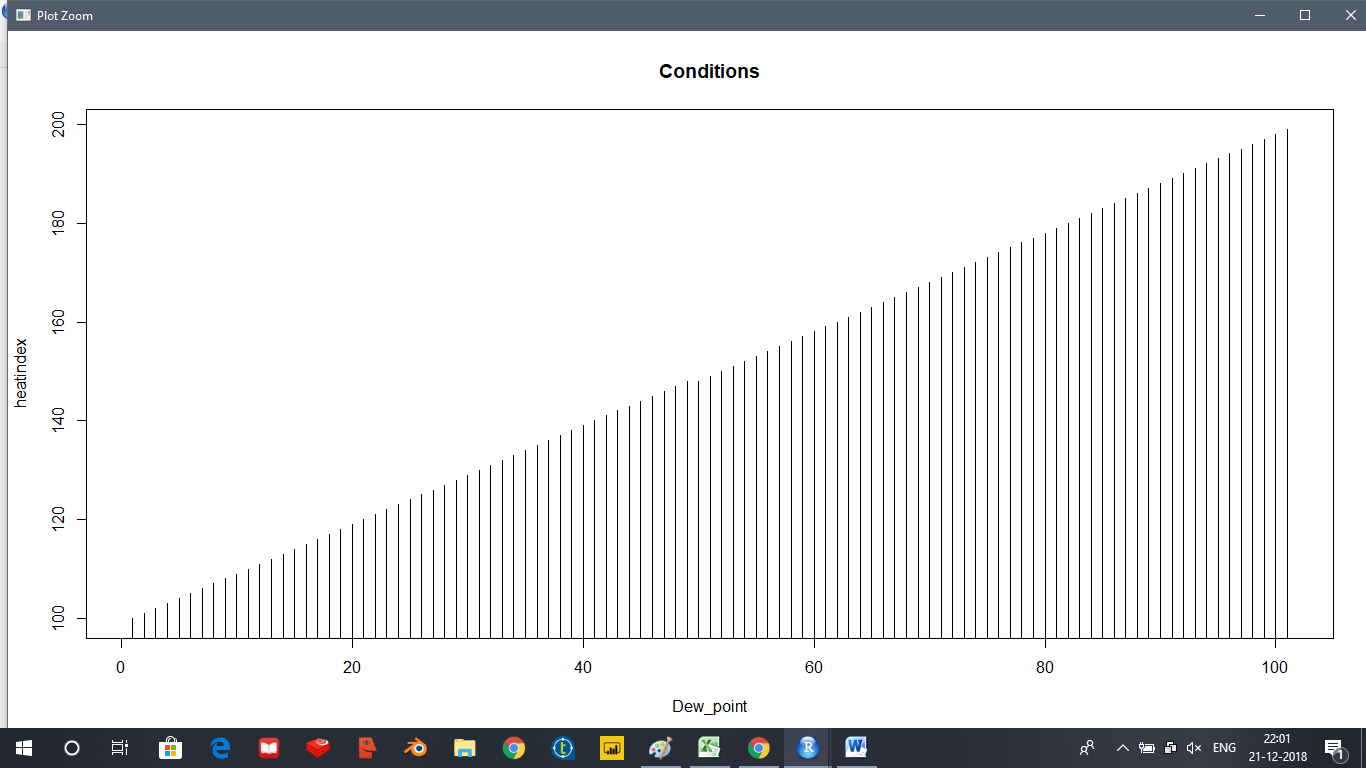
**SCREENSHOT** **RESULT**

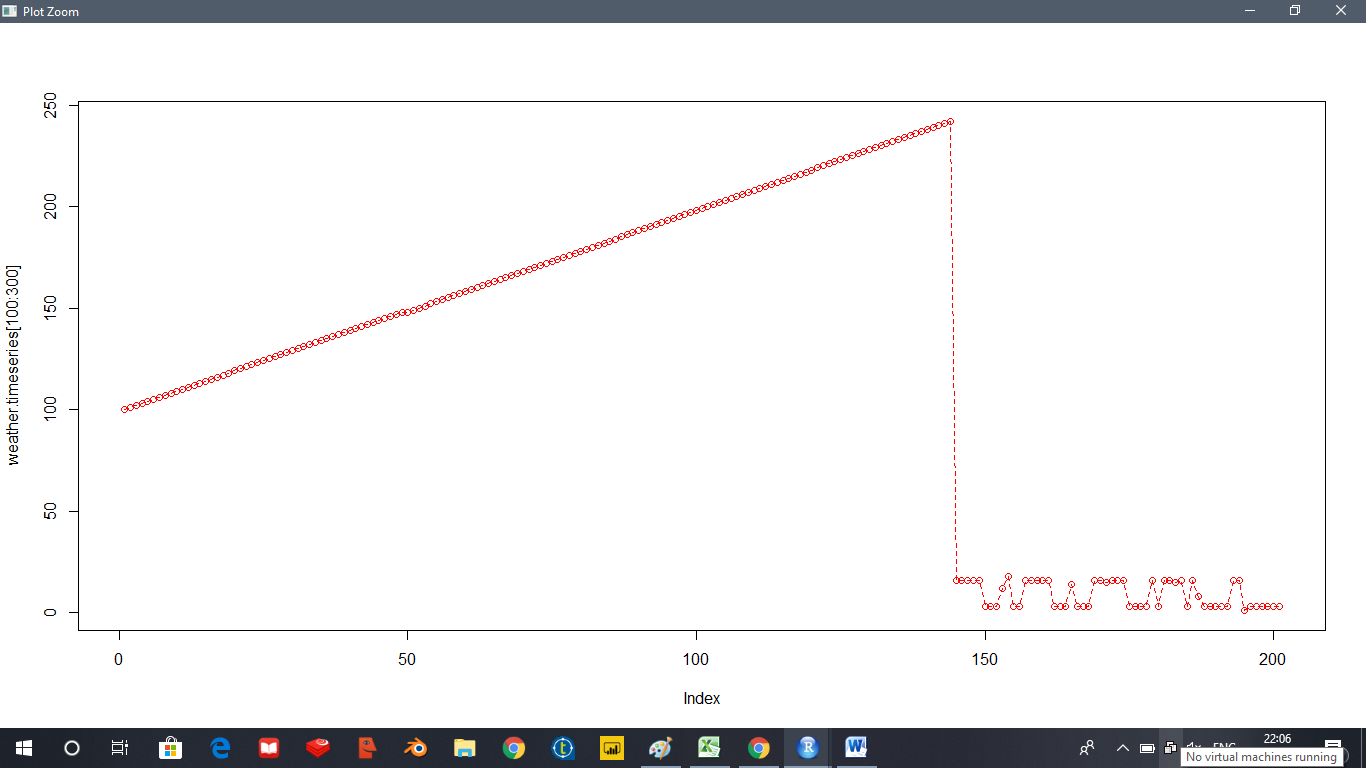


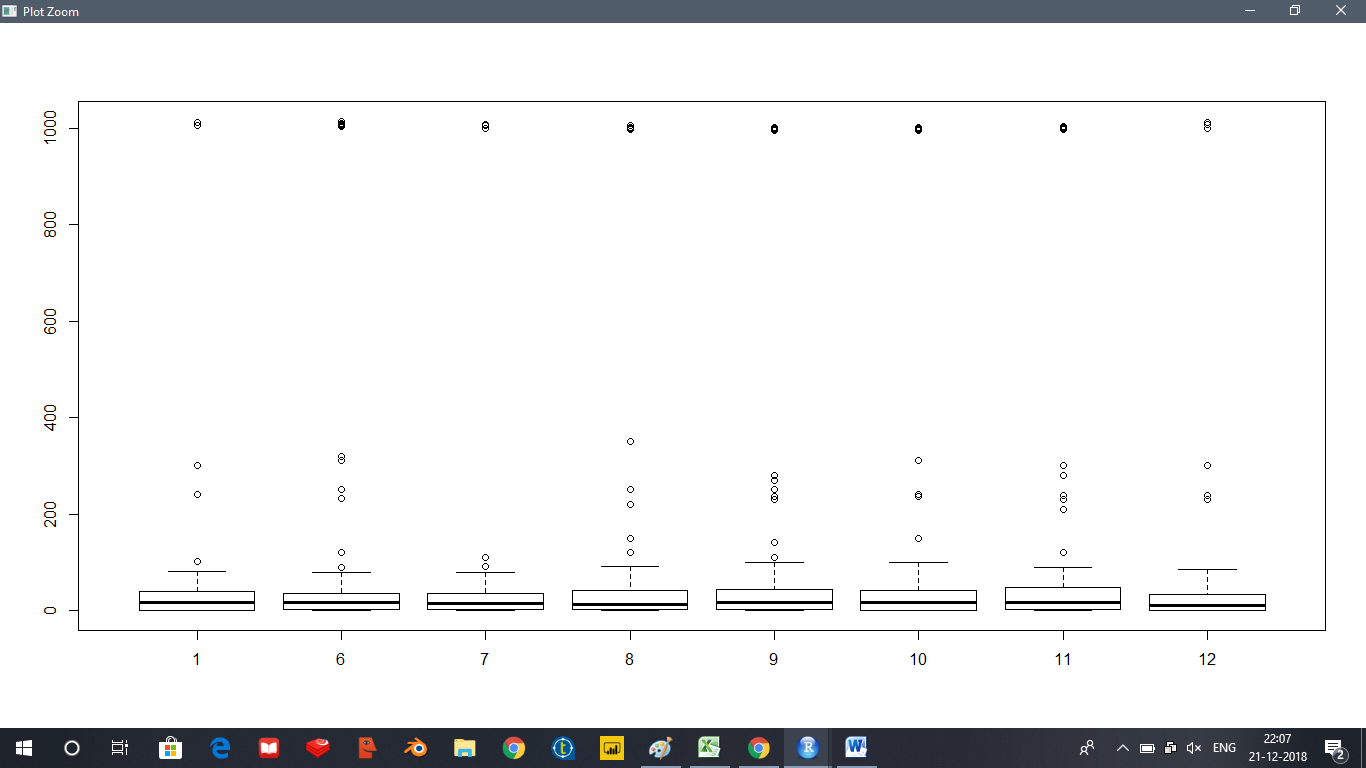














**4. CONCLUSION**

This research suggests and proposes an efficient and accurate weather prediction and forecasting model using linear regression concepts and normal equation model. All these concepts are a part of machine learning. The normal equation is a very efficient weather prediction model and using the entities temperature, humidity and dew-point, it can be used to make reliable weather predictions. This model also facilitates decision making in day to day life. It can yield better results when applied to cleaner and larger datasets. Pre-processing of the datasets can be effective in the prediction as unprocessed data can also affect the efficiency of the model.

**5. FUTURE WORK**

In future the data processing and analysis will result in more accurate prediction of future weather events due to more powerful and precise algorithms for forecasting .

Due to continuous development in this field and accurate weather prediction with the help of these powerful algorithms will result in early detection of any weather calamity plus more viable lifestyle choices for people living in Delhi.

**6. REFERENCES**

1. Wikipedia
2. R Programming page
3. Kaggle for the dataset.
4. <http://manishbarnwal.com/blog/2017/05/03/time_series_and_forecasting_using_R/>
5. http://r-statistics.co/Time-Series-Analysis-With-R.html