Detailed EDA and Prediction of PM2.5 dataset(Beijing)

gautam patel

2022-10-04

# 1.Goal of the project

Throughout this project, I will be doing an Exploratory Data Analysis and prediction through Multiple Linear Regresssion(MLR) of the PM2.5 Dataset of Bejing, China.

# A Brief Intriduction to the PM2.5 Concentration

What is PM, and how does it get into the air?

PM stands for particulate matter (also called particle pollution): the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope.

Particle pollution includes: PM10 : inhalable particles, with diameters that are generally 10 micrometers and smaller; and PM2.5 : fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller.How small is 2.5 micrometers? Think about a single hair from your head. The average human hair is about 70 micrometers in diameter – making it 30 times larger than the largest fine particle.

Sources of PM:

These particles come in many sizes and shapes and can be made up of hundreds of different chemicals.

Some are emitted directly from a source, such as construction sites, unpaved roads, fields, smokestacks or fires.

Most particles form in the atmosphere as a result of complex reactions of chemicals such as sulfur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industries and automobiles.

# 2.Loading the Main packages

I will mainly be using the following packages:

library(dplyr) #To work with datasets

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(moments) #To calculate various values of descriptive sattistics

# 3.Importing And Cleaning Dataset

First, I need to import My dataset. The dataset I will be using can be found online at: [https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data#](https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data)

It was built from public data made available by Song Xi Chen, csx ‘@’ gsm.pku.edu.cn, Guanghua School of Management, Center for Statistical Science, Peking University and This hourly data set contains the PM2.5 data of US Embassy in Beijing. Meanwhile, meteorological data from Beijing Capital International Airport are also included,for each hour between Jan 1st, 2010 to Dec 31st, 2014. It contains 43824 observations.

#loading the dataset file  
library(readr)  
PM2.5\_data <- read.csv("C:/Users/gauta/Downloads/PRSA\_data\_2010.1.1-2014.12.31-1.csv", header=T)  
head(PM2.5\_data)

## No year month day hour pm2.5 DEWP TEMP PRES cbwd Iws Is Ir  
## 1 1 2010 1 1 0 NA -21 -11 1021 NW 1.79 0 0  
## 2 2 2010 1 1 1 NA -21 -12 1020 NW 4.92 0 0  
## 3 3 2010 1 1 2 NA -21 -11 1019 NW 6.71 0 0  
## 4 4 2010 1 1 3 NA -21 -14 1019 NW 9.84 0 0  
## 5 5 2010 1 1 4 NA -20 -12 1018 NW 12.97 0 0  
## 6 6 2010 1 1 5 NA -19 -10 1017 NW 16.10 0 0

#View(PM2.5\_data)

### Description of the dataset

No : row number  
year : year of data in this row  
month: month of data in this row  
day : day of data in this row  
hour : hour of data in this row  
PM2.5: PM2.5 concentration (ug/m^3)  
TEMP : temperature (degree Celsius)  
PRES : pressure (hPa)  
DEWP : dew point temperature (degree Celsius)  
Cbwd : Represents the Combined wind direction   
Iws : Represents the Cumulated wind speed in (m/s)  
Is : Represents the Cumulated hours of snow  
Ir : Represents the Cumulated hours of rain

### Duplicate and missing Values:

anyDuplicated(PM2.5\_data)

## [1] 0

sum(is.na(PM2.5\_data))

## [1] 2067

# Now since our dataset contains the NA value, therefore we have to impute those NA values with the mean of that variable, which can be done as follows:  
PM2.5\_data$pm2.5[is.na(PM2.5\_data$pm2.5)]<-mean(PM2.5\_data$pm2.5,na.rm=TRUE)  
head(PM2.5\_data)

## No year month day hour pm2.5 DEWP TEMP PRES cbwd Iws Is Ir  
## 1 1 2010 1 1 0 98.61321 -21 -11 1021 NW 1.79 0 0  
## 2 2 2010 1 1 1 98.61321 -21 -12 1020 NW 4.92 0 0  
## 3 3 2010 1 1 2 98.61321 -21 -11 1019 NW 6.71 0 0  
## 4 4 2010 1 1 3 98.61321 -21 -14 1019 NW 9.84 0 0  
## 5 5 2010 1 1 4 98.61321 -20 -12 1018 NW 12.97 0 0  
## 6 6 2010 1 1 5 98.61321 -19 -10 1017 NW 16.10 0 0

##Creating the date column  
  
PM2.5\_data$date<-as.Date(with(PM2.5\_data,paste(PM2.5\_data$year,PM2.5\_data$month,PM2.5\_data$day,sep="-")),"%Y-%m-%d")  
  
##  
##adding the season column  
season <- vector('character', nrow(PM2.5\_data))  
season[PM2.5\_data$month == 3 | PM2.5\_data$month == 4 | PM2.5\_data$month == 5] <- 'spring'  
season[PM2.5\_data$month == 6 | PM2.5\_data$month == 7 | PM2.5\_data$month == 8] <- 'summer'  
season[PM2.5\_data$month == 9 | PM2.5\_data$month == 10 | PM2.5\_data$month == 11] <- 'fall'  
season[PM2.5\_data$month == 12 | PM2.5\_data$month == 1 | PM2.5\_data$month == 2] <- 'winter'  
PM2.5\_data <- cbind(PM2.5\_data, season)

# 4.Exploring The Dataset

Before we start fitting a machine learning model on the data, we need to know much about the data by performing an Exploratory Data Analysis to gain insight from it. EDA is simply describing the data by means of visualization. It involves asking questions about the data and answering them with the help of charts/graphs (graphical representation of the data). In this process, we will try to study the behavoir of the amount of pollutant (PM2.5 concentration) in the air and the relationship between other features. Below are some of the questions we will try to answer by analyzing the data, to know more about our dependent and independent variables:

.what pattern does the amount of PM2.5 concentration in the air recorded in an hour follow for a daily time period ?  
.In which month does the amount of PM2.5 contained in the air rises ?  
.At what time of the day do we expect the amount of PM2.5 concentration in the air to be high ?  
.In which direction does polluted air/wind mostly move ?  
.How do the other environmental factors affect the amount of PM2.5 concentration in the air ?

We now have our questions so let’s just dive into our data and start finding and interpreting some results. But since we are going to take averages of the dependent variable, we should know the distribtution of the data before we do take averages. So lets start with the summary of the whole dataset as follows:

## 4.1 Summary Of the dataset PM2.5

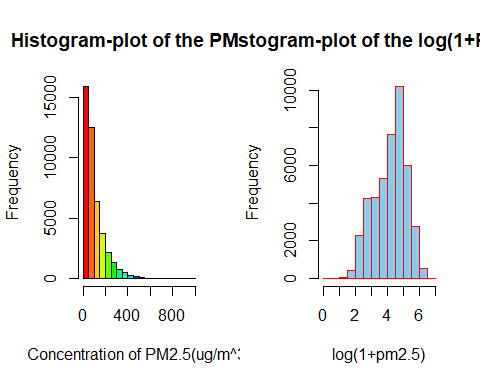
summary(PM2.5\_data)

## No year month day   
## Min. : 1 Min. :2010 Min. : 1.000 Min. : 1.00   
## 1st Qu.:10957 1st Qu.:2011 1st Qu.: 4.000 1st Qu.: 8.00   
## Median :21913 Median :2012 Median : 7.000 Median :16.00   
## Mean :21913 Mean :2012 Mean : 6.524 Mean :15.73   
## 3rd Qu.:32868 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.00   
## Max. :43824 Max. :2014 Max. :12.000 Max. :31.00   
## hour pm2.5 DEWP TEMP   
## Min. : 0.00 Min. : 0.00 Min. :-40.000 Min. :-19.00   
## 1st Qu.: 5.75 1st Qu.: 31.00 1st Qu.:-10.000 1st Qu.: 2.00   
## Median :11.50 Median : 77.00 Median : 2.000 Median : 14.00   
## Mean :11.50 Mean : 98.61 Mean : 1.817 Mean : 12.45   
## 3rd Qu.:17.25 3rd Qu.:132.00 3rd Qu.: 15.000 3rd Qu.: 23.00   
## Max. :23.00 Max. :994.00 Max. : 28.000 Max. : 42.00   
## PRES cbwd Iws Is   
## Min. : 991 Length:43824 Min. : 0.45 Min. : 0.00000   
## 1st Qu.:1008 Class :character 1st Qu.: 1.79 1st Qu.: 0.00000   
## Median :1016 Mode :character Median : 5.37 Median : 0.00000   
## Mean :1016 Mean : 23.89 Mean : 0.05273   
## 3rd Qu.:1025 3rd Qu.: 21.91 3rd Qu.: 0.00000   
## Max. :1046 Max. :585.60 Max. :27.00000   
## Ir date season   
## Min. : 0.0000 Min. :2010-01-01 Length:43824   
## 1st Qu.: 0.0000 1st Qu.:2011-04-02 Class :character   
## Median : 0.0000 Median :2012-07-01 Mode :character   
## Mean : 0.1949 Mean :2012-07-01   
## 3rd Qu.: 0.0000 3rd Qu.:2013-10-01   
## Max. :36.0000 Max. :2014-12-31

## 4.2 Plotting the histogram-plot of the continious variables

### 4.2.1.For Column 6(PM2.5)

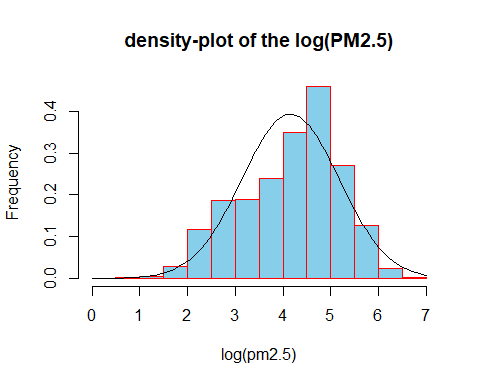
library(moments)  
par(mfrow=c(1,2))  
hist((PM2.5\_data$pm2.5),col=rainbow(15),xlab = "Concentration of PM2.5(ug/m^3)",ylab = "Frequency",main = "Histogram-plot of the PM2.5",breaks = 30)  
hist(log(1+(PM2.5\_data$pm2.5)),col="skyblue",xlab = "log(1+pm2.5)",ylab = "Frequency",main = "Histogram-plot of the log(1+PM2.5)",border = "red")



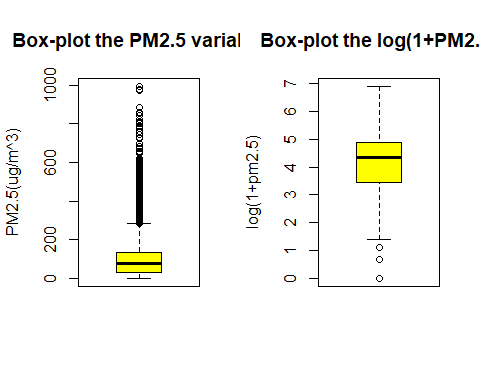
##  
par(mfrow=c(1,1))  
x=seq(2,8,by=0.1)  
which(PM2.5\_data$pm2.5==0)

## [1] 24035 24040

mn=mean(log(PM2.5\_data$pm2.5[-c(24035,24040)]))  
vr=var(log(PM2.5\_data$pm2.5[-c(24035,24040)]))  
hist(log(PM2.5\_data$pm2.5),col="skyblue",xlab = "log(pm2.5)",ylab = "Frequency",main = "density-plot of the log(PM2.5)",border = "red",probability = TRUE)  
curve(dnorm(x,mean=mn,sd=sqrt(vr)),add = TRUE)



##  
par(mfrow=c(1,2))  
boxplot(PM2.5\_data$pm2.5,col = "yellow",main="Box-plot the PM2.5 variable",ylab="PM2.5(ug/m^3)")  
boxplot(log(1+PM2.5\_data$pm2.5),col = "yellow",main="Box-plot the log(1+PM2.5)",ylab="log(1+pm2.5)")



#From the bar-plot of PM2.5 variable it can concluded that the distribution of PM2.5 is positively skewed, with a skewness of:  
skewness(PM2.5\_data$pm2.5)

## [1] 1.846314

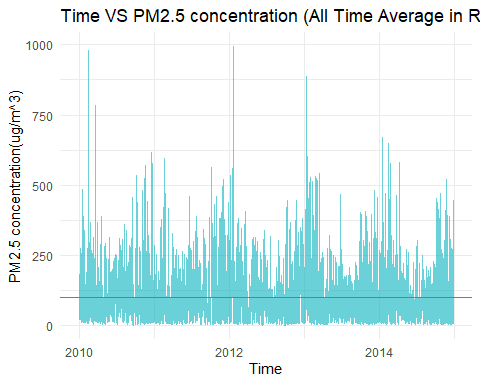
# Also from the box-plot it is clear that we ha too many out-liers, therefore i have transformed the variable PM2.5 to log(1+PM2.5) to reduce the number of outlier and from the box-plot of log(1+PM2.5) it can be seen that transformed PM2.5 variable has very less number of out-liers.  
# Kurtosis of the Pm2.5 Variable  
kurtosis(PM2.5\_data$pm2.5)

## [1] 8.152751

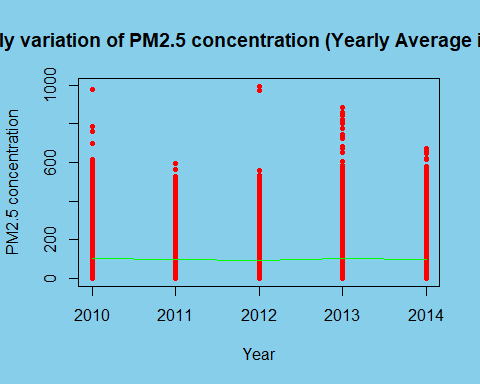
## Variation in Concentration of PM2.5 variable with repest to time:

### a.With repect to Year

library(ggplot2)  
ggplot(PM2.5\_data, aes(x = date, y = pm2.5)) +  
 geom\_line(alpha = 0.58, color = "#00AFBB") +   
 geom\_hline(yintercept=mean(PM2.5\_data$pm2.5),   
 color="#FC4E07",   
 linetype = 1) +  
 theme\_minimal() +  
 labs(title = "Time VS PM2.5 concentration (All Time Average in Red)", x = "Time", y = "PM2.5 concentration(ug/m^3)")

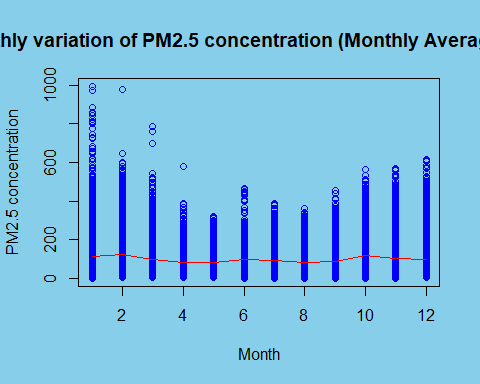


par(bg="skyblue")  
ty=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$year),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$year,PM2.5\_data$pm2.5,col="red",xlab = "Year",ylab = "PM2.5 concentration",main = "Yearly variation of PM2.5 concentration (Yearly Average in Green)",pch=20)  
lines(ty$Group.1,ty$x,col="green")



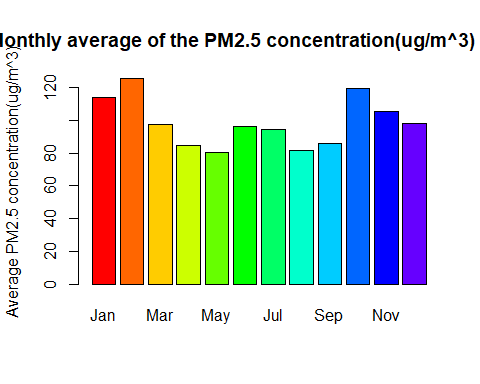
### b.With repect to Month

par(bg="skyblue")  
t=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$month),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$month,PM2.5\_data$pm2.5,col="blue",xlab = "Month",ylab = "PM2.5 concentration",main = "Monthly variation of PM2.5 concentration (Monthly Average in red)")  
lines(t$Group.1,t$x,col="red")



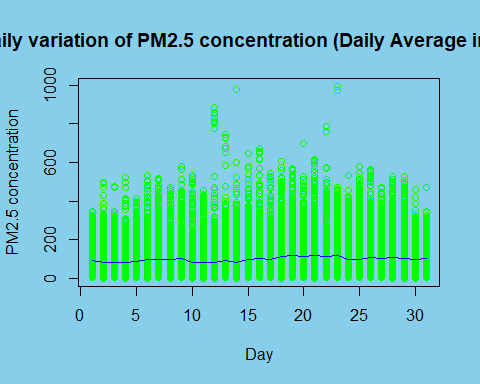
### In which month does the amount of PM2.5 contained in the air rises ?

barplot(t$x,names.arg = c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"),col = rainbow(15),main = "Monthly average of the PM2.5 concentration(ug/m^3) in air ",ylab = "Average PM2.5 concentration(ug/m^3)",horiz = FALSE)

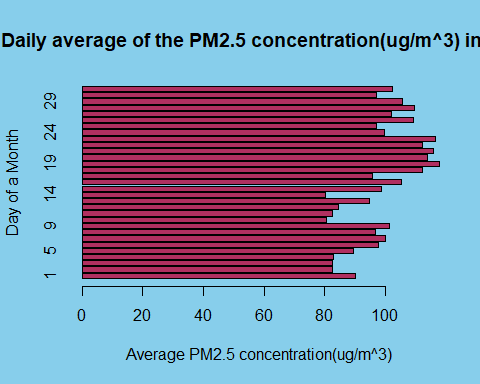


### c.With repect to daily

par(bg="skyblue")  
t=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$day),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$day,PM2.5\_data$pm2.5,col="green",xlab = "Day",ylab = "PM2.5 concentration",main = "Daily variation of PM2.5 concentration (Daily Average in blue)")  
lines(t$Group.1,t$x,col="blue")

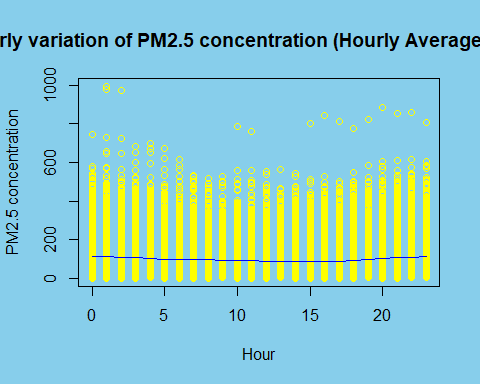


barplot(t$x,col = "maroon",main = "Daily average of the PM2.5 concentration(ug/m^3) in air ",xlab = "Average PM2.5 concentration(ug/m^3)",ylab = "Day of a Month",names.arg = c(1:31),horiz = TRUE)



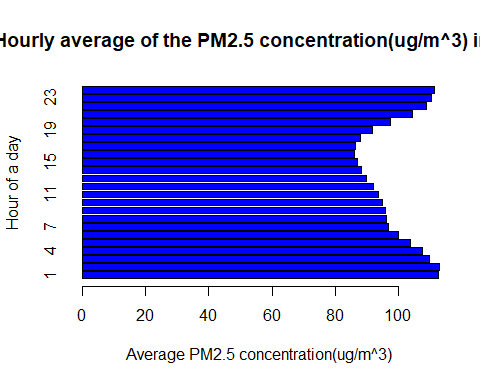
### d.With repect to hour

par(bg="skyblue")  
t=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$hour),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$hour,PM2.5\_data$pm2.5,col="yellow",xlab = "Hour",ylab = "PM2.5 concentration",main = "Hourly variation of PM2.5 concentration (Hourly Average in blue)")  
lines(t$Group.1,t$x,col="blue")



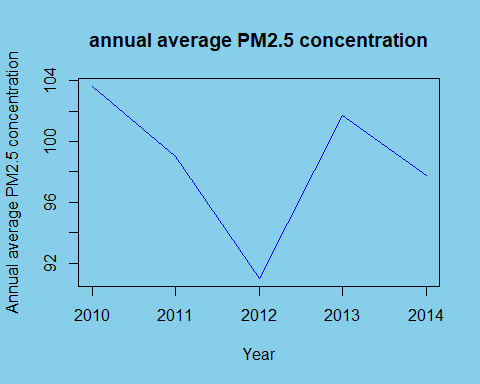
### At what time of the day do we expect the amount of PM2.5 concentration in the air to be high ?

barplot(t$x,col = "blue",main = "Hourly average of the PM2.5 concentration(ug/m^3) in air ",xlab = "Average PM2.5 concentration(ug/m^3)",ylab = "Hour of a day",names.arg = c(1:24),horiz = TRUE)



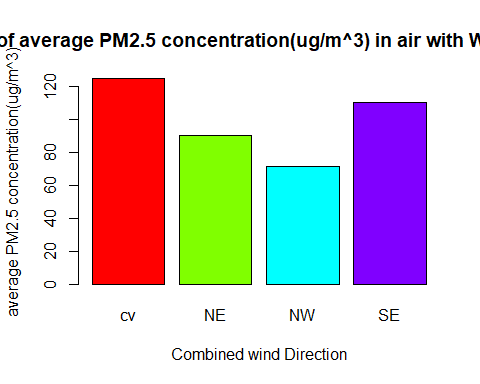
## e.Let’s take a look at annual average PM2.5 concentration:

par(bg="skyblue")  
plot(ty$Group.1,ty$x,col="blue","l",xlab = "Year",ylab = "Annual average PM2.5 concentration",main = "annual average PM2.5 concentration")



## f.In which direction does polluted air/wind mostly move ?

t=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$cbwd),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
barplot(t$x,col = rainbow(4),main = "Variation of average PM2.5 concentration(ug/m^3) in air with Wind-Direction ",xlab = "Combined wind Direction",ylab = "average PM2.5 concentration(ug/m^3)",names.arg = c("cv","NE","NW","SE"),horiz = FALSE)

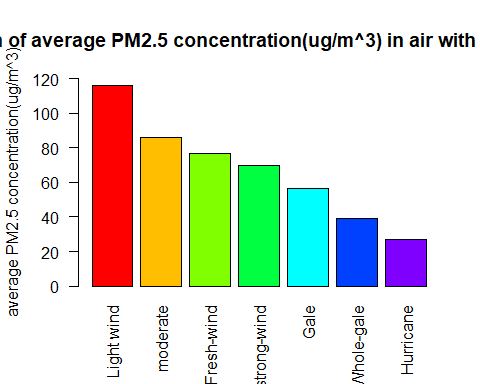


## g.Effect of wind speed on PM2.5 concentration

library(tidyverse)

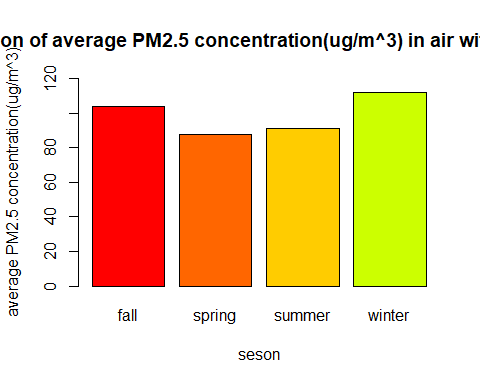
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ tibble 3.1.8 ✔ stringr 1.4.0  
## ✔ tidyr 1.2.0 ✔ forcats 0.5.1  
## ✔ purrr 0.3.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

t=PM2.5\_data %>%  
 pull(Iws) %>%  
 cut(breaks=c(-Inf,11,28,38,61,88,117,Inf),   
 labels=c("Light wind", "Gentle-moderate", "Fresh-wind","strong-wind","Gale","Whole-gale","Hurricane"))  
  
PM2.5\_data <- cbind(PM2.5\_data, t)  
ts=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$t),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
barplot(ts$x,col = rainbow(8),main = "Variation of average PM2.5 concentration(ug/m^3) in air with Wind-Speed ",ylab = "average PM2.5 concentration(ug/m^3)",names.arg = c("Light wind", "moderate", "Fresh-wind","strong-wind","Gale","Whole-gale","Hurricane"),horiz = FALSE,las=2,legend=T,ylim = c(0,120))



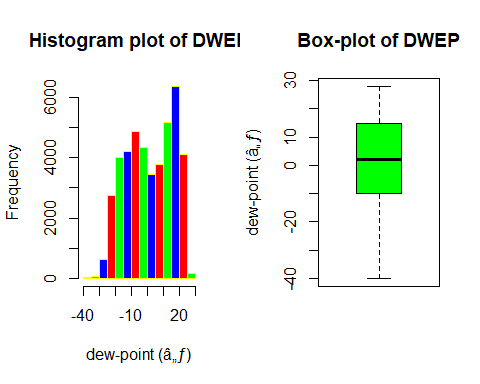
## h.In which seson does the PM2.5 concentration is high?

t=aggregate(x= PM2.5\_data$pm2.5,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$season),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
barplot(t$x,col = rainbow(15),main = "Variation of average PM2.5 concentration(ug/m^3) in air with Sesons ",xlab = "seson",ylab = "average PM2.5 concentration(ug/m^3)",names.arg = c("fall","spring","summer","winter"),horiz = FALSE,ylim = c(0,120))



## 4.2.2.for column 7(DWEP)

#since the variable DWEP(dew point in(â„ƒ)) is continuous variable therefore I am going to plot the histogram and box-plot:  
par(mfrow=c(1,2))  
hist(PM2.5\_data$DEWP,col=c("red","green","blue"),xlab = "dew-point (â„ƒ)",ylab = "Frequency",main = "Histogram plot of DWEP",border = "yellow")  
boxplot(PM2.5\_data$DEWP,ylab="dew-point (â„ƒ)",main="Box-plot of DWEP",col = "green")



#summary  
summary(PM2.5\_data$DEWP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -40.000 -10.000 2.000 1.817 15.000 28.000

skewness(PM2.5\_data$DEWP)

## [1] -0.1524422

kurtosis(PM2.5\_data$DEWP)

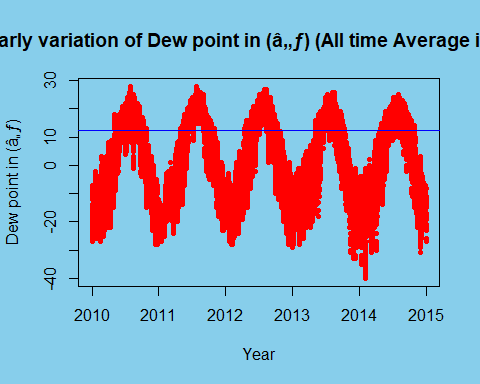
## [1] 1.80471

##From the above histogram plot and box-plot of the variable DWEP(dew-point) it can be concluded that the variable DWEP does not follows any fixed distribution(but slightely negatively skewed) and from the box-plot we can concluded that the variable has no outliers.

## Let us plot the DWEP variable with repest to time:

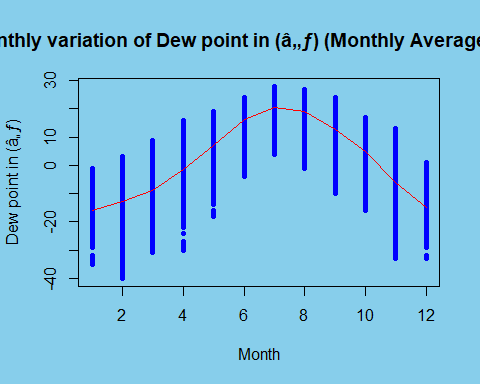
### a.With repect to Year

par(bg="skyblue")  
dy=aggregate(x= PM2.5\_data$DEWP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$year),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$date,PM2.5\_data$DEWP,col="red",xlab = "Year",ylab = "Dew point in (â„ƒ)",main = "Yearly variation of Dew point in (â„ƒ) (All time Average in blue)",pch=20)  
abline(h=mean(PM2.5\_data$TEMP),col="blue")

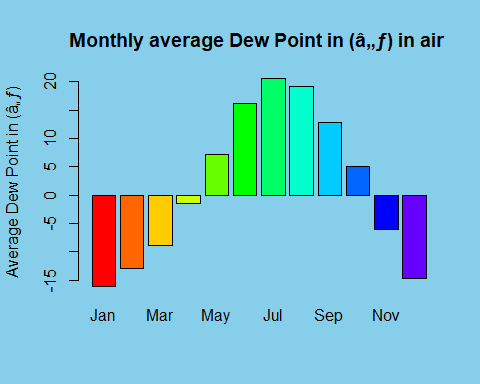


### b.With repect to Month

par(bg="skyblue")  
dm=aggregate(x= PM2.5\_data$DEWP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$month),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$month,PM2.5\_data$DEWP,col="blue",pch=20,xlab = "Month",ylab = "Dew point in (â„ƒ)",main = "Monthly variation of Dew point in (â„ƒ) (Monthly Average in red)")  
lines(dm$Group.1,dm$x,col="red")

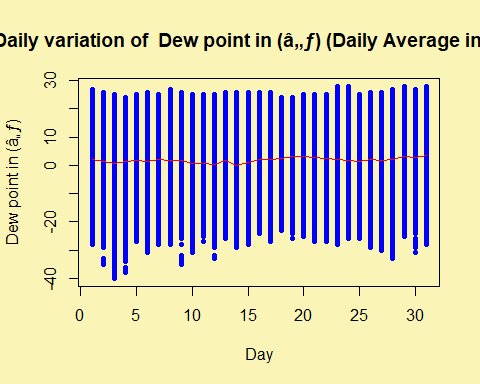


barplot(dm$x,names.arg = c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"),col = rainbow(15),main = "Monthly average Dew Point in (â„ƒ) in air ",ylab = "Average Dew Point in (â„ƒ)",horiz = FALSE)

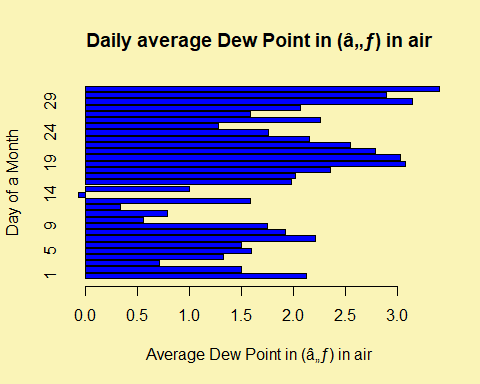


### c.With repect to daily

par(bg="#FAF4B7")  
dd=aggregate(x= PM2.5\_data$DEWP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$day),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$day,PM2.5\_data$DEWP,col="blue",xlab = "Day",ylab = "Dew point in (â„ƒ)",main = "Daily variation of Dew point in (â„ƒ) (Daily Average in red)",pch=20)  
lines(dd$Group.1,dd$x,col="red")

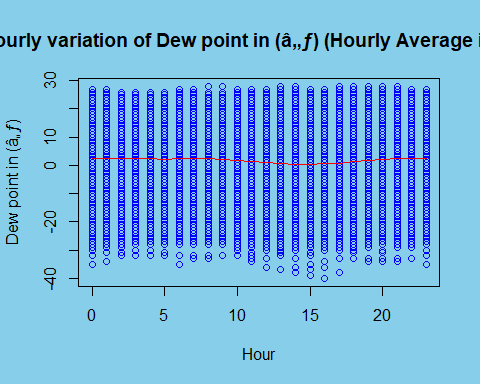


barplot(dd$x,col = "blue",main = "Daily average Dew Point in (â„ƒ) in air",xlab = "Average Dew Point in (â„ƒ) in air",ylab = "Day of a Month",names.arg = c(1:31),horiz = TRUE)

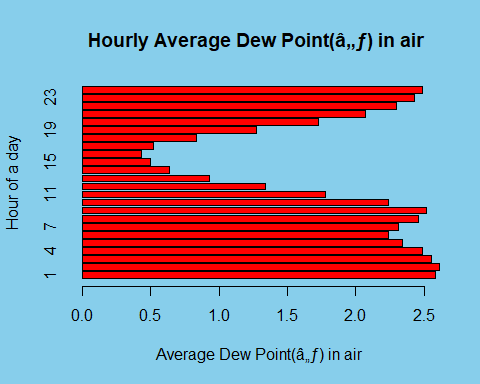


### d.With repect to hour

par(bg="skyblue")  
dh=aggregate(x= PM2.5\_data$DEWP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$hour),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$hour,PM2.5\_data$DEWP,col="blue",xlab = "Hour",ylab = "Dew point in (â„ƒ)",main = "Hourly variation of Dew point in (â„ƒ) (Hourly Average in red)")  
lines(dh$Group.1,dh$x,col="red")

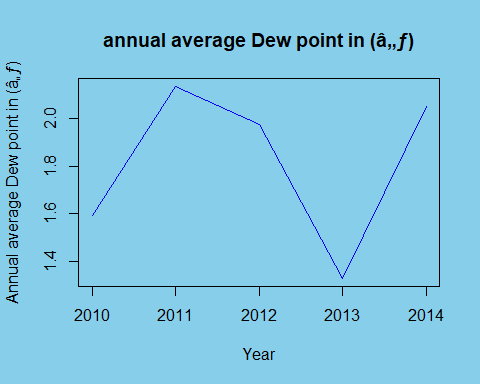


barplot(dh$x,col = "red",main = "Hourly Average Dew Point(â„ƒ) in air ",xlab = "Average Dew Point(â„ƒ) in air",ylab = "Hour of a day",names.arg = c(1:24),horiz = TRUE)



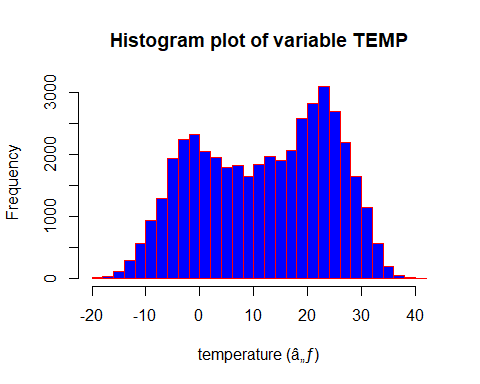
## e.Let’s take a look at annual average Dew point in (â„ƒ):

par(bg="skyblue")  
plot(dy$Group.1,dy$x,col="blue","l",xlab = "Year",ylab = "Annual average Dew point in (â„ƒ)",main = "annual average Dew point in (â„ƒ)")

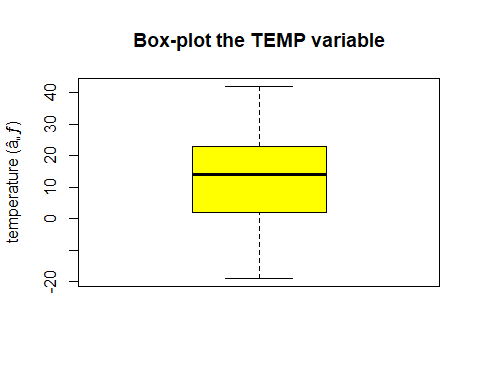


## 4.2.3.For Column 8(TEMP)

#par(mfrow=c(1,2))  
hist((PM2.5\_data$TEMP),col="blue",xlab = "temperature (â„ƒ)",ylab = "Frequency",main = "Histogram plot of variable TEMP",border = "red",breaks = 30)



##boxplot 0f TEMP  
boxplot(PM2.5\_data$TEMP,col = "yellow",main="Box-plot the TEMP variable",ylab="temperature (â„ƒ)")



#summary  
summary(PM2.5\_data$TEMP)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -19.00 2.00 14.00 12.45 23.00 42.00

skewness(PM2.5\_data$TEMP)

## [1] -0.163298

kurtosis(PM2.5\_data$TEMP)

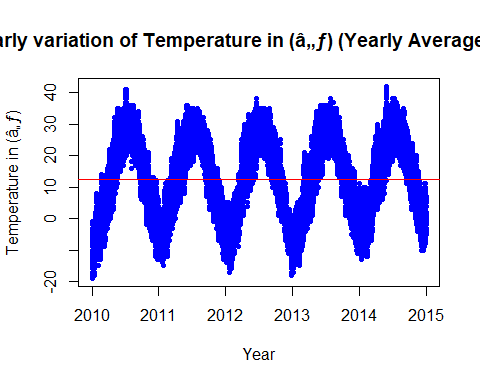
## [1] 1.889013

##From the above histogram plot and box-plot of the variable TEMP(Temperature in (â„ƒ)) it can be concluded that the variable TEMP does not follows any fixed distribution(but slightely negatively skewed)and slightly leptokurtic and from the box-plot we can concluded that the variable has no outliers.

## Let us plot the TEMP variable with repest to time:

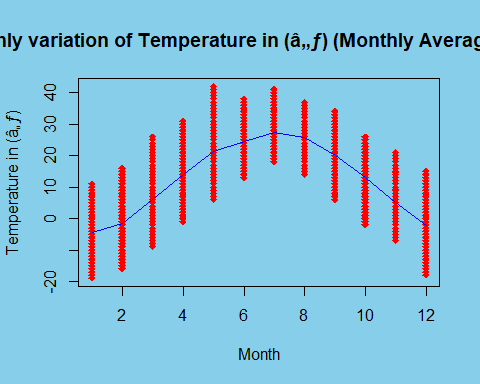
### a.With repect to Year

plot(PM2.5\_data$date,PM2.5\_data$TEMP,col="blue",pch=20,xlab = "Year",ylab = "Temperature in (â„ƒ)",main = "Yearly variation of Temperature in (â„ƒ) (Yearly Average in red)")  
abline(h=mean(PM2.5\_data$TEMP),col="red")



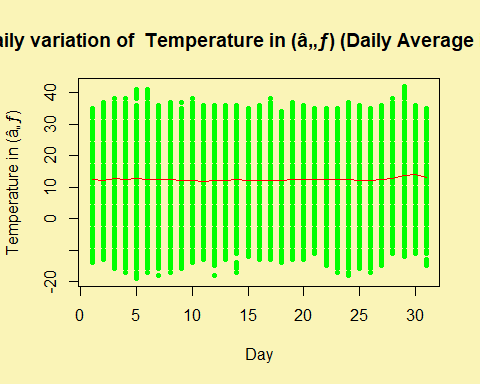
### b.With repect to Month

par(bg="skyblue")  
tm=aggregate(x= PM2.5\_data$TEMP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$month),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$month,PM2.5\_data$TEMP,col="red",pch=18,xlab = "Month",ylab = "Temperature in (â„ƒ)",main = "Monthly variation of Temperature in (â„ƒ) (Monthly Average in blue)")  
lines(tm$Group.1,tm$x,col="blue")



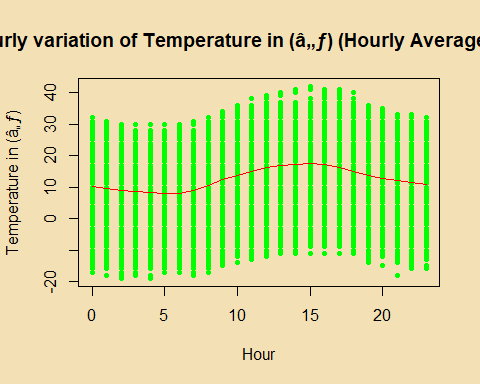
### c.With repect to day

par(bg="#FAF4B7")  
td=aggregate(x= PM2.5\_data$TEMP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$day),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$day,PM2.5\_data$TEMP,col="green",pch=20,xlab = "Day",ylab = "Temperature in (â„ƒ)",main = "Daily variation of Temperature in (â„ƒ) (Daily Average in red)")  
lines(td$Group.1,td$x,col="red")



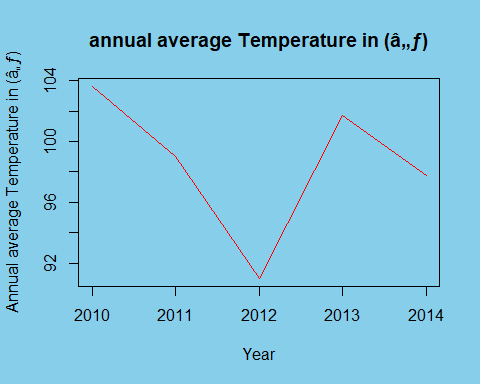
### d.With repect to hour

par(bg="#F3E0B5")  
th=aggregate(x= PM2.5\_data$TEMP,   
   
 # Specify group indicator  
 by = list(PM2.5\_data$hour),   
   
 # Specify function (i.e. mean)  
 FUN = mean)  
plot(PM2.5\_data$hour,PM2.5\_data$TEMP,col="green",pch=20,xlab = "Hour",ylab = "Temperature in (â„ƒ)",main = "Hourly variation of Temperature in (â„ƒ) (Hourly Average in red)")  
lines(th$Group.1,th$x,col="red")



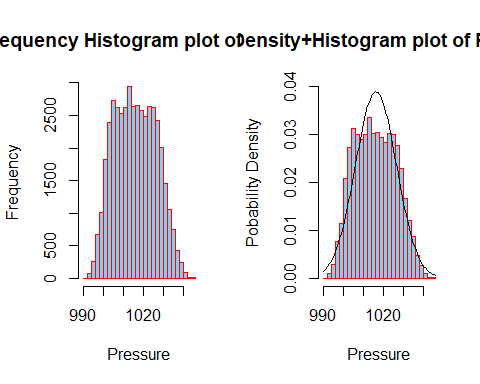
## e.Let’s take a look at annual average Temperature in (â„ƒ):

par(bg="skyblue")  
plot(ty$Group.1,ty$x,col="red","l",xlab = "Year",ylab = "Annual average Temperature in (â„ƒ)",main = "annual average Temperature in (â„ƒ)")

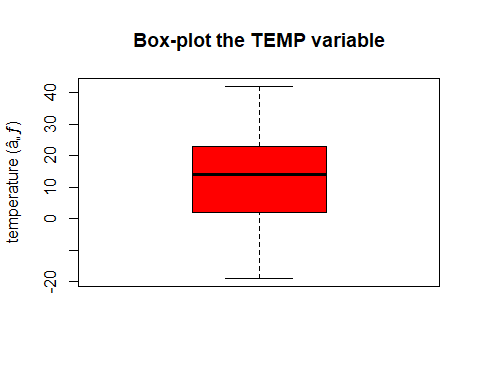


## 4.2.3.For Column 9(PRES)

par(mfrow=c(1,2))  
hist((PM2.5\_data$PRES),col="skyblue",xlab = "Pressure",ylab = "Frequency",main = "Frequency Histogram plot of PRES",border = "red",breaks = 20)  
hist((PM2.5\_data$PRES),col="skyblue",xlab = "Pressure",ylab = "Pobability Density",main = "Density+Histogram plot of PRES",border = "red",probability = TRUE,ylim = c(0,0.04),breaks = 20)  
x=seq(990,1050,by=1)  
curve(dnorm(x,mean = mean(PM2.5\_data$PRES),sd=sd(PM2.5\_data$PRES)),add = TRUE,col="black")



##boxplot 0f TEMP  
par(mfrow=c(1,1))  
boxplot(PM2.5\_data$TEMP,col = "red",main="Box-plot the TEMP variable",ylab="temperature (â„ƒ)")



#summary  
summary(PM2.5\_data$PRES)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 991 1008 1016 1016 1025 1046

skewness(PM2.5\_data$PRES)

## [1] 0.09820348

kurtosis(PM2.5\_data$PRES)

## [1] 2.153498

##From the above histogram plot and box-plot of the variable PRES(pressure) it can be concluded that the variable PRES follows approximately normal distribution with slightly Platykuratic and from the box-plot we can concluded that the variable has no outliers.

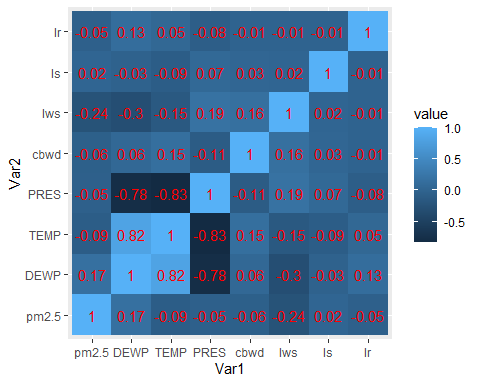
## 4.3 Relationship between PM2.5 variable and other Metrological factors (Heatmap of correlation matrix)

# creating correlation matrix  
PM2.5\_data$cbwd=factor(PM2.5\_data$cbwd)  
PM2.5\_data$cbwd=unclass(PM2.5\_data$cbwd)  
View(PM2.5\_data)  
PM2.5\_data$cbwd=as.numeric(PM2.5\_data$cbwd)  
df1=data.frame(PM2.5\_data[,c(6,7,8,9,10,11,12,13)])  
corr\_mat <- round(cor(df1),2)  
library(reshape2)

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

# reduce the size of correlation matrix  
melted\_corr\_mat <- melt(corr\_mat)  
   
# plotting the correlation heatmap  
library(ggplot2)  
ggplot(data = melted\_corr\_mat, aes(x=Var1, y=Var2,  
 fill=value)) +  
geom\_tile() +  
geom\_text(aes(Var2, Var1, label = value),  
 color = "red", size = 4)



## Modelling and Predicting PM2.5 concentration(Through Multiple Linear RegressionModel(MLRM)

Since our dataset contains the observation for four consutive years(2010,2011,2013,2014) but for the pridiction purpose one year data is enough, therefore I have considered the dataset for last year(year 2014), further since our dataset is hourely so first I have transformed the hourely data to daily data(by taking the daily average of all the variables) as follows:

library(readr)  
PM2.5\_data <- read.csv("C:/Users/gauta/Downloads/PRSA\_data\_2010.1.1-2014.12.31-1.csv", header=T)  
PM2.5\_data14=PM2.5\_data[35065:43824,]  
View(PM2.5\_data14)  
##  
anyDuplicated(PM2.5\_data14)

## [1] 0

sum(is.na(PM2.5\_data14))

## [1] 99

PM2.5\_data14$pm2.5[is.na(PM2.5\_data14$pm2.5)]<-mean(PM2.5\_data14$pm2.5,na.rm=TRUE)  
View(PM2.5\_data14)  
##  
PM2.5\_data14$cbwd=factor(PM2.5\_data14$cbwd)  
PM2.5\_data14$cbwd=unclass(PM2.5\_data14$cbwd)  
View(PM2.5\_data14)  
#creating the date column  
PM2.5\_data14$date<-as.Date(with(PM2.5\_data14,paste(PM2.5\_data14$year,PM2.5\_data14$month,PM2.5\_data14$day,sep="-")),"%Y-%m-%d")  
View(PM2.5\_data14)  
##  
library(dplyr)  
df1=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(pm2.5=mean(pm2.5)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df1)  
df2=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(DEWP=mean(DEWP)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df2)  
df3=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(TEMP=mean(TEMP)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df3)  
df4=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(PRES=mean(PRES)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df4)  
df5=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(Iws=mean(Iws)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df5)  
df6=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(Ir=mean(Ir)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df6)  
  
df7=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(cbwd=mean(cbwd)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df7)  
df8=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(Is=mean(Is)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df8)  
df9=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(month=mean(month)))

## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df9)  
df10=data.frame(PM2.5\_data14 %>% group\_by(date,day) %>%   
 summarise(day=mean(day)))

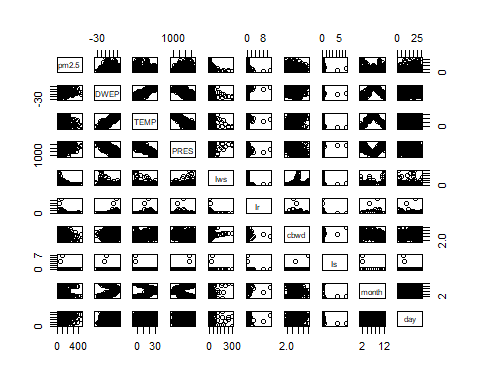
## `summarise()` has grouped output by 'date'. You can override using the  
## `.groups` argument.

#View(df10)  
data\_final=data.frame(pm2.5=df1$pm2.5,DWEP=df2$DEWP,TEMP=df3$TEMP,PRES=df4$PRES,Iws=df5$Iws,Ir=df6$Ir,cbwd=df7$cbwd,Is=df8$Is,month=df9$month,day=df10$day)  
View(data\_final)

### A Brief Description of the Data

We should already be familiar with the process of going through descriptive statistics in order better understand the data. However, there is one very useful plot for regression analysis that should be incorporated. This is the pairs plot.

pairs(data\_final[,c(1:10)])



### Checking for Multicollinearity(VIF test)

As we know from that for linear regression the dependent variable Y(pm2.5) should be normally or approximately normal distributed, and from the density-plot of log(pm2.5)  
 we can see that the distribution of log(pm2.5) is approximately normally distributed, therefore for regression purpose, I have transformed the pm2.5 variable to log(pm2.5).

library(faraway)  
s=225:365###making the training data  
# Construct our initial model   
model1 <- lm(log(pm2.5) ~ ., data = data\_final[s,])  
summary(model1)

##   
## Call:  
## lm(formula = log(pm2.5) ~ ., data = data\_final[s, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4915 -0.2536 0.0000 0.3589 1.3783   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.572540 11.599319 2.377 0.018897 \*   
## DWEP 0.134436 0.011086 12.126 < 2e-16 \*\*\*  
## TEMP -0.096062 0.024752 -3.881 0.000164 \*\*\*  
## PRES -0.029299 0.010954 -2.675 0.008435 \*\*   
## Iws -0.002682 0.001035 -2.592 0.010628 \*   
## Ir -0.116562 0.042798 -2.724 0.007340 \*\*   
## cbwd -0.071455 0.131152 -0.545 0.586802   
## Is 17.253408 12.378654 1.394 0.165737   
## month 0.699923 0.154114 4.542 1.25e-05 \*\*\*  
## day 0.028255 0.005634 5.015 1.69e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5088 on 131 degrees of freedom  
## Multiple R-squared: 0.7051, Adjusted R-squared: 0.6849   
## F-statistic: 34.8 on 9 and 131 DF, p-value: < 2.2e-16

##Computing the correlation matrix  
round(cor(data\_final,method = "pearson"),2)

## pm2.5 DWEP TEMP PRES Iws Ir cbwd Is month day  
## pm2.5 1.00 0.03 -0.22 0.12 -0.24 -0.07 -0.19 -0.01 -0.14 0.16  
## DWEP 0.03 1.00 0.90 -0.82 -0.30 0.15 0.13 -0.06 0.23 0.10  
## TEMP -0.22 0.90 1.00 -0.89 -0.20 0.07 0.22 -0.12 0.12 0.08  
## PRES 0.12 -0.82 -0.89 1.00 0.20 -0.08 -0.15 0.09 -0.02 -0.04  
## Iws -0.24 -0.30 -0.20 0.20 1.00 -0.05 0.23 0.02 0.13 -0.06  
## Ir -0.07 0.15 0.07 -0.08 -0.05 1.00 -0.01 -0.01 0.04 -0.01  
## cbwd -0.19 0.13 0.22 -0.15 0.23 -0.01 1.00 0.13 -0.21 -0.03  
## Is -0.01 -0.06 -0.12 0.09 0.02 -0.01 0.13 1.00 -0.10 -0.07  
## month -0.14 0.23 0.12 -0.02 0.13 0.04 -0.21 -0.10 1.00 0.01  
## day 0.16 0.10 0.08 -0.04 -0.06 -0.01 -0.03 -0.07 0.01 1.00

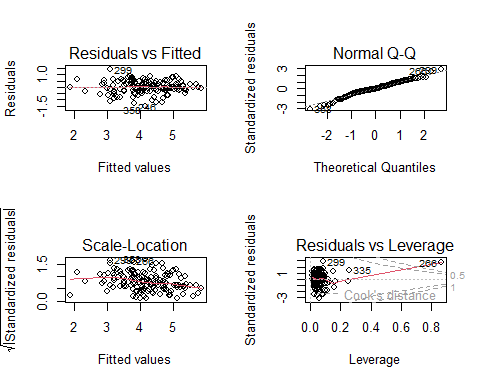
#Computes the VIF for every predictor  
vif(model1)

## DWEP TEMP PRES Iws Ir cbwd Is month   
## 11.276138 30.531262 3.583824 1.554716 1.090092 1.300653 1.020521 23.483785   
## day   
## 1.309213

### Normality of Residuals(Residual Analysis For initial model

par(mfrow = c(2, 2))  
plot(model1)

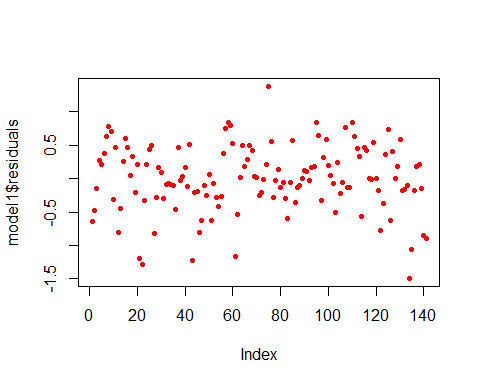
## Warning: not plotting observations with leverage one:  
## 120



mean(resid(model1))

## [1] -2.101353e-17

##the errors are un-correlated  
par(mfrow=c(1,1))  
plot(model1$residuals,col="red",pch=20)



Now From the above correlation analysis and VIF factor analysis it can concluded that the dependent variables have higher VIF value, and from the coreelation table it can be seen that some variables like day,Ir,Is are very weakly coreelated with the depentdent variable pm2.5 so in the next step we will discard these variables due to high multicollinarity between them,therefore the refigned model is like:

### Improvement of model:1(Model:2)

model2 <- lm(log(pm2.5) ~ ., data = data\_final[s,-c(4,6,8,10)])  
summary(model2)

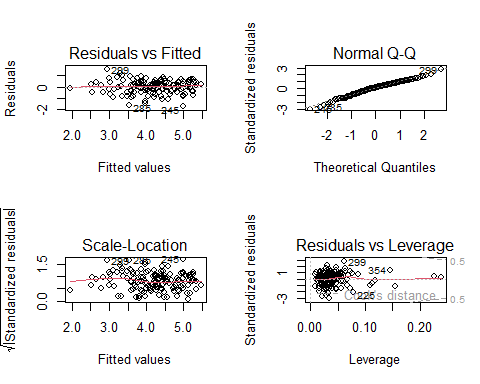
##   
## Call:  
## lm(formula = log(pm2.5) ~ ., data = data\_final[s, -c(4, 6, 8,   
## 10)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71039 -0.36574 0.07931 0.39279 1.53230   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.090464 1.784139 0.611 0.54210   
## DWEP 0.133307 0.012148 10.974 < 2e-16 \*\*\*  
## TEMP -0.114287 0.024431 -4.678 6.95e-06 \*\*\*  
## Iws -0.003272 0.001178 -2.778 0.00625 \*\*   
## cbwd -0.081918 0.149670 -0.547 0.58506   
## month 0.434992 0.153746 2.829 0.00538 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5821 on 135 degrees of freedom  
## Multiple R-squared: 0.6022, Adjusted R-squared: 0.5874   
## F-statistic: 40.87 on 5 and 135 DF, p-value: < 2.2e-16

#Computes the VIF for every predictor  
vif(model2)

## DWEP TEMP Iws cbwd month   
## 10.341982 22.719406 1.538947 1.293828 17.852222

### Residual Analysis

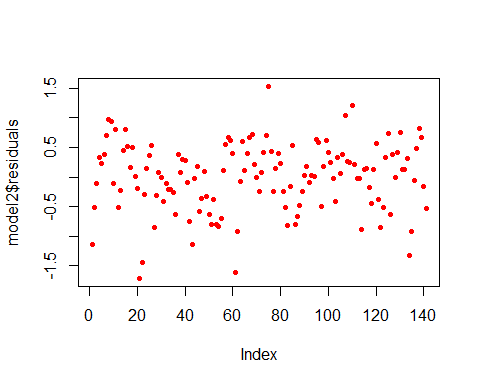
par(mfrow = c(2, 2))  
plot(model2)



mean(resid(model2))

## [1] -5.065954e-17

##the errors are un-correlated  
par(mfrow=c(1,1))  
plot(model2$residuals,col="red",pch=20)



### Final Multiple regression model is:

pm2.5=1.090464+0.133507\*DWEP-0.114287\*TEMP-0.003272\*Iws-0.081918\*cbwd+0.434992\*month

### The final value of R-square is: 0.6022

## Model Validation

Steps required:  
 1. Split the data into a training set and test set  
 2. Fit your model to the training set  
 3. Predict the responses in the test set  
 4. Evaluate the quality of the predictions

### step1: Split the data randomly into a training set and a test set

# Required for R2, RMSE and MAE commands  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':  
##   
## melanoma

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

set.seed(220)  
n\_train <- ceiling(0.6 \* length(data\_final$pm2.5))  
train\_sample <- sample(c(1:length(data\_final$pm2.5)), n\_train)  
train\_data <- data\_final[train\_sample,]  
test\_data <- data\_final[-train\_sample,]

### step2:Fit the model on the training data

model <- lm(pm2.5 ~ ., data = train\_data)  
predictions <- predict(model2, test\_data)# Measure performance by comparing the prediction with the data using multiple criterion  
R\_sq <- R2(predictions, test\_data$pm2.5)  
RMSE <- RMSE(predictions, test\_data$pm2.5)  
MAE <- MAE(predictions, test\_data$pm2.5)  
print(c(R\_sq, RMSE, MAE))

## [1] 0.02220384 121.80776620 93.71795456

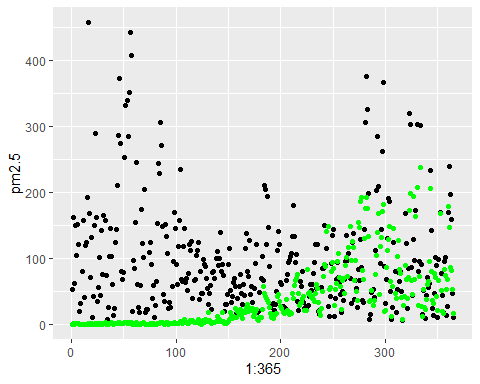
pred\_error\_rate <- RMSE / mean(test\_data$pm2.5)  
pred\_error\_rate

## [1] 1.267397

## Plotting the pridicted values of pm2.5 Vs actual values of pm2.5

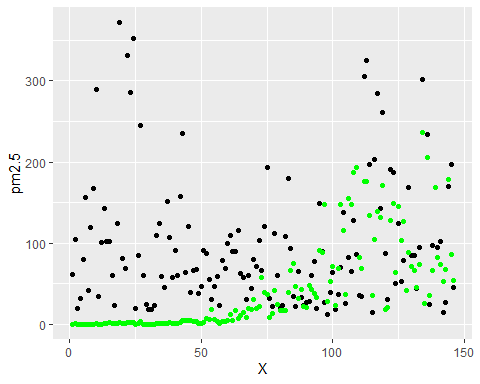
### To the whole DATA

par(mfrow=c(1,2))  
library(ggplot2)  
pr\_pm2.5=1.090464+0.133507\*data\_final$DWEP-0.114287\*data\_final$TEMP-0.003272\*data\_final$Iws-0.081918\*data\_final$cbwd+0.434992\*data\_final$month  
data\_final=data.frame(data\_final,pr\_pm2.5)  
ggplot(data = data\_final)+ geom\_point(aes(x=1:365,y=pm2.5))+geom\_point(aes(x=1:365,y=exp(pr\_pm2.5)),col="green")



### To the test DATA

par(mfrow=c(1,2))  
library(ggplot2)  
pr\_pm2.5=1.090464+0.133507\*test\_data$DWEP-0.114287\*test\_data$TEMP-0.003272\*test\_data$Iws-0.081918\*test\_data$cbwd+0.434992\*test\_data$month  
X=1:146  
test\_data=data.frame(test\_data,pr\_pm2.5,X)  
ggplot(data = test\_data)+ geom\_point(aes(x=X,y=pm2.5))+geom\_point(aes(x=X,y=exp(pr\_pm2.5)),col="green")



## Conclusion

In this project,I have Critically examined the variation of PM2.5 concentration with time (yearly, Monthly, Daily, hourly, seasonally).And as a result I got that the yearly variation of PM2.5 is approximately same in all the years,while taking about the monthly variation we can see that the concentration of PM2.5 is higher in the starting and last months of year while in the middle months it is quite less, and talking about on a particular day we can see that at the mid-nights the concentration is relatively high. We can also see that the PM2.5 concentration is higher in winter season followed by fall,summer and spring.And talking about the relation between wind speed and PM2.5 it can clearly observed that as the wind speed increases the concentration of PM2.5 decreases.I have also investigated the relationships between PM2.5 concentration and meteorological factors in terms of regional and seasonal variations. In this way, i have obtained a more comprehensive and precise understanding of how PM2.5 concentration is correlated with meteorological factors. This knowledge could provide a solid foundation for more accurate PM2.5 concentration retrieval, and for making more effective environmental protection policies for different regions. Although much work has been done in this area of research, there is still much room for improvement. The exploration of the cause of this regional and seasonal variation would help us to better understand the air pollution problem in China.  
 Lastely I have tried to predict the concentration of PM2.5 in the air by fitting a machine learning model(Multiple Linear Regression) as a result, I have obtained the value of R^2 as 0.6022, which reflects a fair amount of accuracy of the model.  
 Despite these limitations, my work could still provide a better chance for a more accurate prdiction of PM2.5 concentration.   
 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*