Data Analysis Course Work: FIREFOREST DATASET (Introduction to R)

Student name : Trinh Gia Huy
Student number: 290290
Email: giahuy.trinh@tuni.fi

Firstly, I included the csv file of forestfire and assign it to *steam variable*

```
>steam<- read.table(file = "forestfires.csv",header = T,sep = ",")
```

Then I want to get the brief information of the first 6 lines of dataset.

```
>attach(stream)
> head(steam)
  X Y month day FFMC  DMC    DC  ISI temp RH wind rain area
1 7 5   mar fri 86.2 26.2  94.3  5.1  8.2 51  6.7  0.0    0
2 7 4   oct tue 90.6 35.4 669.1  6.7 18.0 33  0.9  0.0    0
3 7 4   oct sat 90.6 43.7 686.9  6.7 14.6 33  1.3  0.0    0
4 8 6   mar fri 91.7 33.3  77.5  9.0  8.3 97  4.0  0.2    0
5 8 6   mar sun 89.3 51.3 102.2  9.6 11.4 99  1.8  0.0    0
6 8 6   aug sun 92.3 85.3 488.0 14.7 22.2 29  5.4  0.0    0
> str(steam)
'data.frame':   517 obs. of  13 variables:
 $ X    : int  7 7 7 8 8 8 8 8 8 7 ...
 $ Y    : int  5 4 4 6 6 6 6 6 6 5 ...
 $ month: chr  "mar" "oct" "oct" "mar" ...
 $ day  : chr  "fri" "tue" "sat" "fri" ...
 $ FFMC : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
 $ DMC  : num  26.2 35.4 43.7 33.3 51.3 ...
 $ DC   : num  94.3 669.1 686.9 77.5 102.2 ...
 $ ISI  : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
 $ temp : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
 $ RH   : int  51 33 33 97 99 29 27 86 63 40 ...
 $ wind : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
 $ rain : num  0 0 0 0.2 0 0 0 0 0 0 ...
 $ area : num  0 0 0 0 0 0 0 0 0 0 ...
```

I want to take some statistic calculations of each variable of the dataset

```
> mean(FFMC)
> sd(FFMC)
> mean(DC)
```

```
> sd(DC)
> mean(ISI)
> sd(ISI)
> mean(temp)
> sd(temp)
> mean(RH)
> sd(RH)
>mean(DMC)
>sd(DMC)
>mean(wind)
>sd(wind)
>mean(rain)
>sd(rain)
```

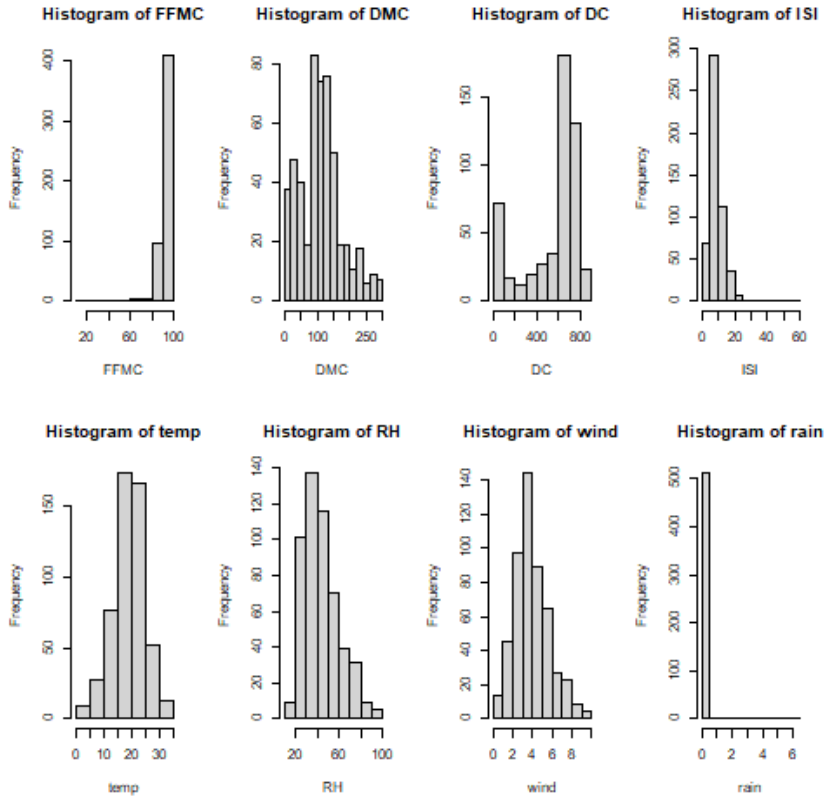
From mean and standard deviation we can calculate the coefficient of variation (CV). And the results are represented as table below:

	Mean	Standard Deviation(sd)	Coefficient of Variation (CV)
FFMC	90.64468	5.520111	0.06089835
DC	547.94	248.0662	0.4527251
ISI	9.021663	4.559477	0.5053921
Temp	18.88917	5.806625	0.307405
RH	44.2882	16.31747	0.3684383
DMC	110.8723	64.04648	0.5776599
Wind	4.017602	1.791653	0.4459508
Rain	0.02166344	0.2959591	13.66169

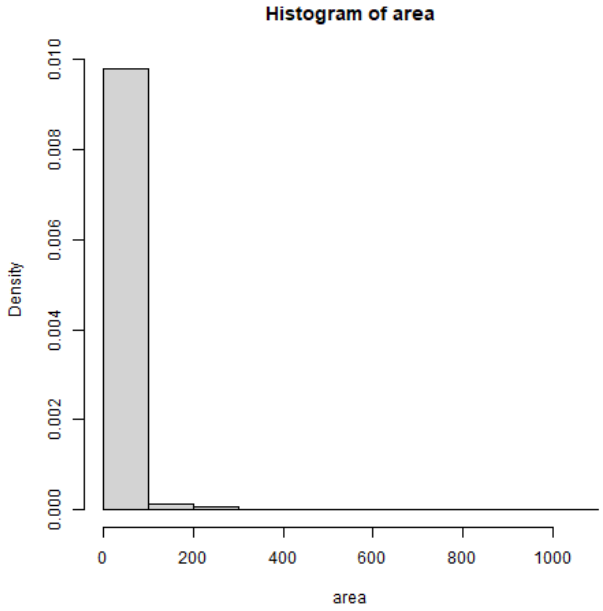
Table 1: Statistical calculation of dataset's variable

Then I would like to plot these variable

```
> par(mfrow = c(2,4))                    #save the result as a matrix 2x4
> hist(FFMC)
> hist(DMC)
> hist(DC)
> hist(ISI)
> hist(temp)
> hist(RH)
> hist(wind)
> hist(rain)
> hist(area,freq = F)
```



Picture 1: The histogram of variables in forest-fire's dataset.



Picture 2: The histogram of area variable in forest-fire's dataset.

As we can see there are 2 different parts of distribution. I will focus on the positive values.

```
>pos_steam = steam[which(area>0),]
>summary(pos_steam)
      X               Y              month              day
 Min.   :1.000   Min.   :2.000   Length:270         Length:270
 1st Qu.:3.000   1st Qu.:4.000   Class :character   Class :character
 Median :5.000   Median :4.000   Mode  :character   Mode  :character
 Mean   :4.807   Mean   :4.367
 3rd Qu.:7.000   3rd Qu.:5.000
 Max.   :9.000   Max.   :9.000
      FFMC            DMC              DC             ISI
 Min.   :63.50   Min.   :  3.2   Min.   : 15.3   Min.   : 0.800
 1st Qu.:90.33   1st Qu.: 82.9   1st Qu.:486.5   1st Qu.: 6.800
 Median :91.70   Median :111.7   Median :665.6   Median : 8.400
 Mean   :91.03   Mean   :114.7   Mean   :570.9   Mean   : 9.177
 3rd Qu.:92.97   3rd Qu.:141.3   3rd Qu.:721.3   3rd Qu.:11.375
 Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :22.700
      temp             RH             wind            rain
 Min.   : 2.20   Min.   :15.00   Min.   :0.400   Min.   :0.00000
 1st Qu.:16.12   1st Qu.:33.00   1st Qu.:2.700   1st Qu.:0.00000
 Median :20.10   Median :41.00   Median :4.000   Median :0.00000
 Mean   :19.31   Mean   :43.73   Mean   :4.113   Mean   :0.02889
 3rd Qu.:23.40   3rd Qu.:53.00   3rd Qu.:4.900   3rd Qu.:0.00000
 Max.   :33.30   Max.   :96.00   Max.   :9.400   Max.   :6.40000
      area
 Min.   :  0.09
 1st Qu.:  2.14
```

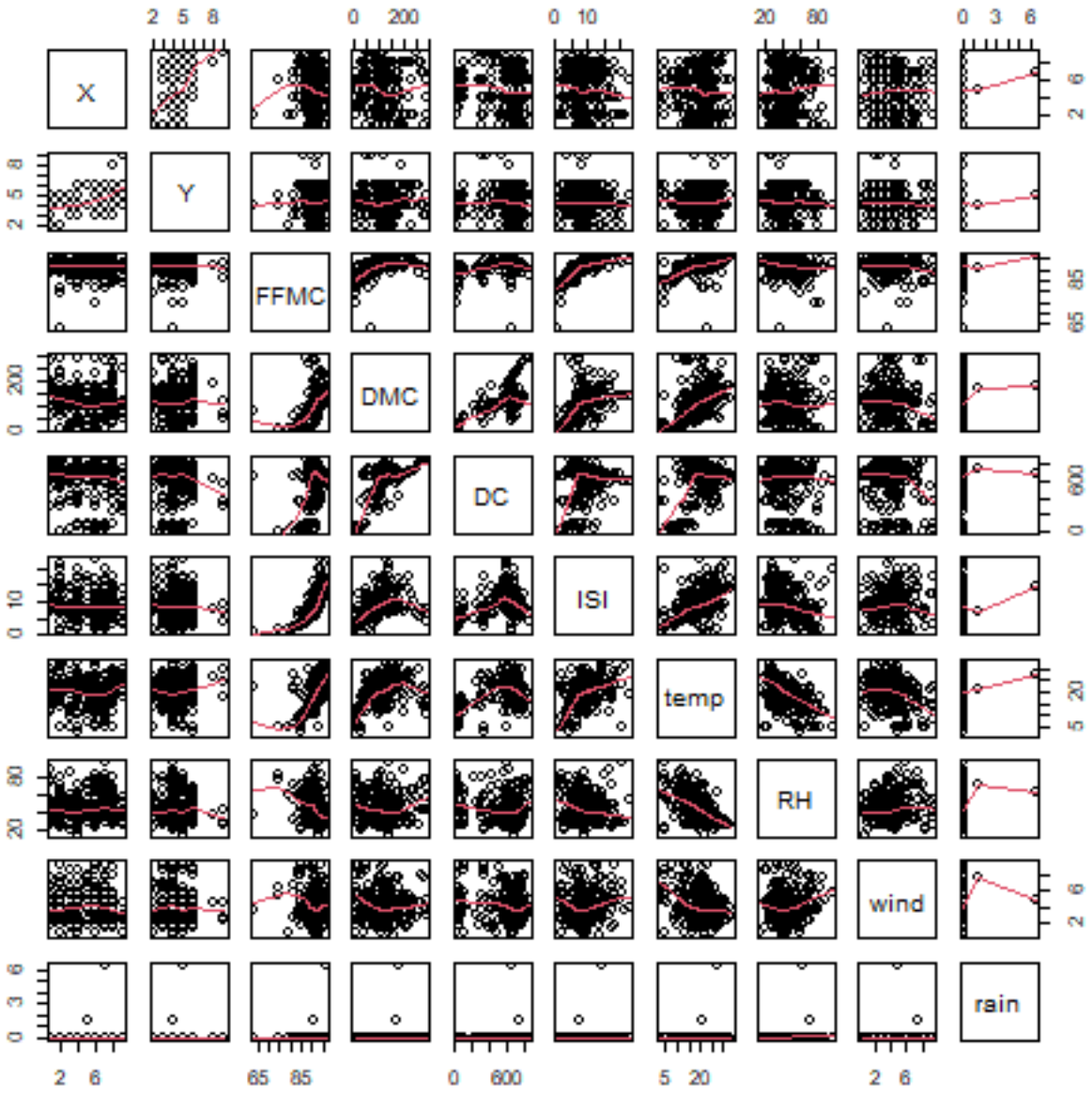
Median : 6.37

Mean : 24.60

3rd Qu.: 15.42

Max. :1090.84

>pairs(pos_steam[,c(1,2,seq(5,12))],panel = panel.smooth)



Picture 3: The plots of all variable and their correlation in forest-fire's dataset.

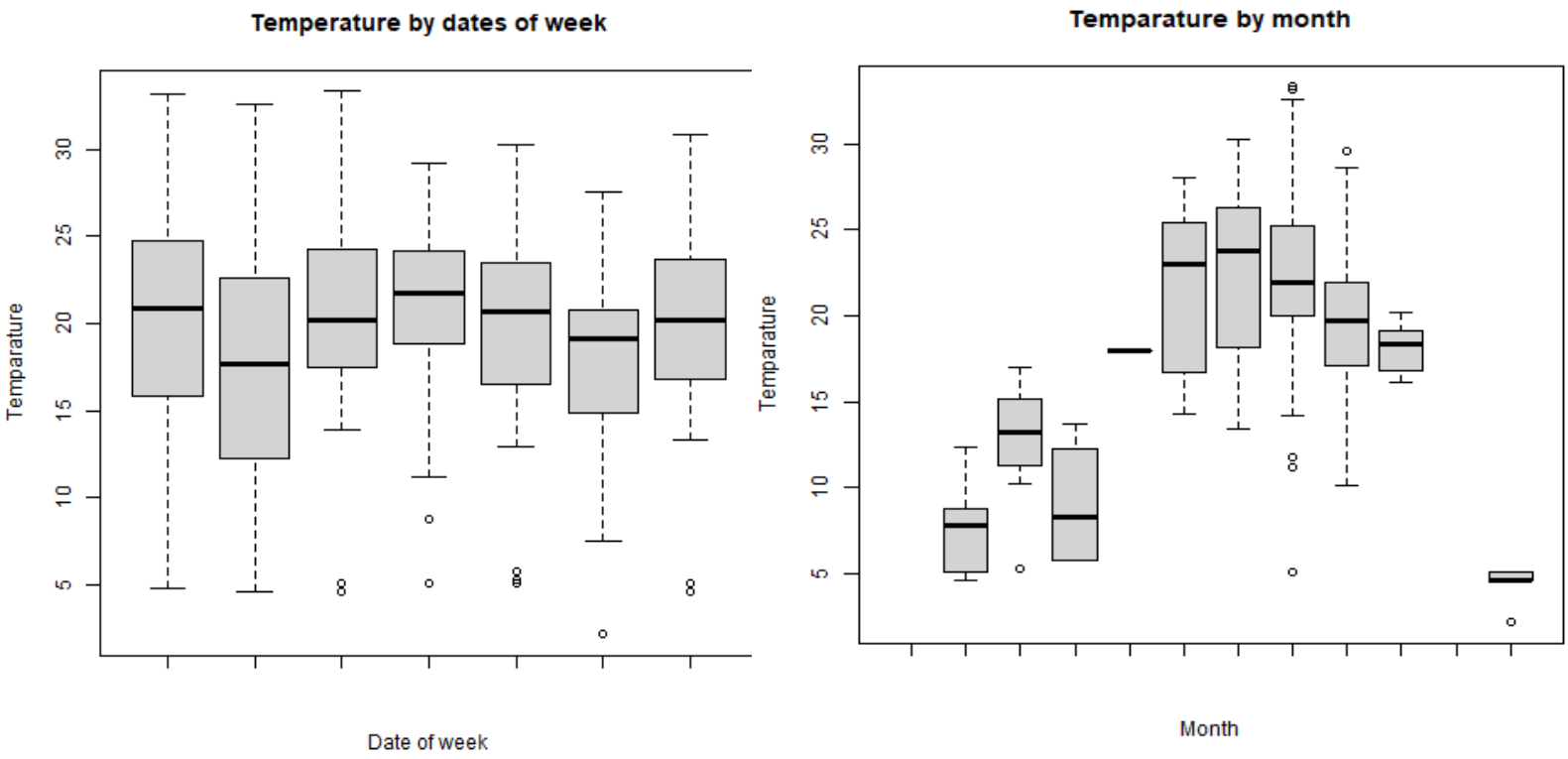
As we can see from the plot that the correlation between DMC and DC seems high. On the other hand, temp and RH variables seem correlated. Now I want to plot the temperature by months and dates of a week.

```
> Mon<- subset(pos_steam,day=="mon")

> Tue<- subset(pos_steam,day=="tue")

> Wed<- subset(pos_steam,day=="wed")

> Thu<- subset(pos_steam,day=="thu")

> Fri<- subset(pos_steam,day=="fri")

> Sat<- subset(pos_steam,day=="sat")

> Sun<- subset(pos_steam,day=="sun")
```

```
> Jan<- subset(pos_steam,month=="jan")

> Feb<- subset(pos_steam,month=="feb")

> Mar<- subset(pos_steam,month=="mar")

> Apr<- subset(pos_steam,month=="apr")

> May<- subset(pos_steam,month=="may")

> Jun<- subset(pos_steam,month=="jun")

> Jul<- subset(pos_steam,month=="jul")

> Aug<- subset(pos_steam,month=="aug")

> Sep<- subset(pos_steam,month=="sep")

> Oct<- subset(pos_steam,month=="oct")

> Nov<- subset(pos_steam,month=="nov")

> Dec<- subset(pos_steam,month=="dec")
```

```
>boxplot(Sun$temp,Mon$temp,Tue$temp,Wed$temp,Thu$temp,Fri$temp,Sat$temp,main="Te
mperature by dates of week",xlab="Date of week",ylab="Temparature")

> boxplot
(Jan$temp,Feb$temp,Mar$temp,Apr$temp,May$temp,Jun$temp,Jul$temp,Aug$temp,Sep$tem
```

p,Oct$temp,Nov$temp,Dec$temp,main="Temparature by month",xlab="Month",ylab="Temparature")



Picture 4 The boxplot of temperature by dates of week and months of year

It seems that there is no obvious relationship between the burned area and the days of week but month of a year does.

References:

[1] Introduction to R : https://humblelu.github.io/IntroR/index.html

[2] UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php