Forest Fires

## R Markdown

#1. Data Loading  
getwd()

## [1] "/Users/fawwazrizvi/Downloads"

#setwd("C:/Users/M93p-7/Downloads")  
setwd("/Users/fawwazrizvi/Downloads")  
ff<-read.csv("forestfires.csv", header = T)

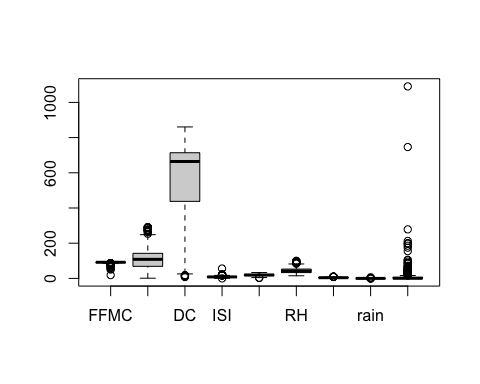
#2. Exploratory Data Analysis - EDA  
str(ff)

## '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 ...

#2. EDA continued...  
#Our main attributes in this dataset are the Forest Fire measurement indices (FFMC, DMC, DC, ISI) and the climate related measures (temp, RH (relative humidity), wind, rain) and our classifier attribute (area (burned)).  
#Therefore, the attributes X, Y, month and day will be removed from the dataset  
ff<-ff[-c(1:4)]  
str(ff)

## 'data.frame': 517 obs. of 9 variables:  
## $ 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 ...

#2. EDA Continued...  
boxplot(ff)

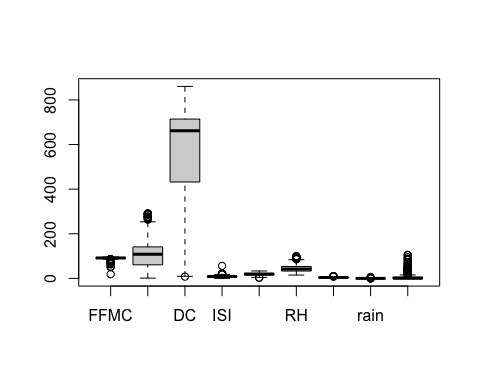


#as we can see the classifier attribute area has a number of outliers that we need to remove  
summary(ff$area)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.52 12.85 6.57 1090.84

#Additionally, we can see a large portion of the data in the area attribute consists of 0 values, which attribute to zero burned areas (in hectares).

#3. Data Pre-Processing.  
#As per above, clearly we can see a bunch of outliers exist for the area attribute, so we will remove these.  
#We will use an upper bound of 0.98 for the upper quartlile as very few values exist above 100 in the area attribute.  
#As we saw using summary() on the area attribute that zero values exist in the dataset up to the 1st quartile, and since the burned area can not be lower than 0, we do not have any lower bound outliers.  
#We will use the quantile function to identify our outliers and then remove them:  
Q <- quantile(ff$area, probs=c(.25, .98), na.rm = FALSE)  
iqr<-IQR(ff$area)  
up <- Q[2]+1.5\*iqr # Upper Range  
low<- Q[1]-1.5\*iqr # Lower Range  
ff<- subset(ff, ff$area > (Q[1] - 1.5\*iqr) & ff$area < (Q[2]+1.5\*iqr))  
  
#Now we can replot our boxplot of the dataset and see that the upper bound outliers are now removed from the area attribute  
boxplot(ff)



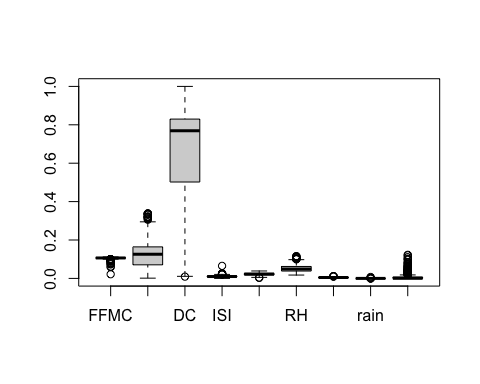
#now we have a better measure for burned area with outliers removed

#3. Data Pre-Processing cont...  
#Now that the outliers have been removed, we will create a normalization function and apply it to our dataset  
  
normalize <- function(x) {  
 return ((x-min(x)) / (max(x)-min(x)))  
}  
#the data set will now be normalized  
norm\_ff<-normalize(ff)

#3. Data Pre-Processing cont...  
#Now our dataset is normalized and area attribute free of large outliers.  
#All values are between 0 and 1.  
summary(norm\_ff)

## FFMC DMC DC ISI   
## Min. :0.02173 Min. :0.001278 Min. :0.00918 Min. :0.000000   
## 1st Qu.:0.10481 1st Qu.:0.070852 1st Qu.:0.50276 1st Qu.:0.007437   
## Median :0.10644 Median :0.125842 Median :0.76900 Median :0.009761   
## Mean :0.10530 Mean :0.127948 Mean :0.63471 Mean :0.010486   
## 3rd Qu.:0.10795 3rd Qu.:0.164188 3rd Qu.:0.82965 3rd Qu.:0.012782   
## Max. :0.11178 Max. :0.338485 Max. :1.00000 Max. :0.065187   
## temp RH wind rain   
## Min. :0.002556 Min. :0.01743 Min. :0.0004648 Min. :0.000e+00   
## 1st Qu.:0.017894 1st Qu.:0.03835 1st Qu.:0.0031373 1st Qu.:0.000e+00   
## Median :0.022426 Median :0.04880 Median :0.0046479 Median :0.000e+00   
## Mean :0.021886 Mean :0.05159 Mean :0.0046747 Mean :2.562e-05   
## 3rd Qu.:0.026493 3rd Qu.:0.06188 3rd Qu.:0.0058389 3rd Qu.:0.000e+00   
## Max. :0.038694 Max. :0.11620 Max. :0.0109226 Max. :7.437e-03   
## area   
## Min. :0.000000   
## 1st Qu.:0.000000   
## Median :0.000488   
## Mean :0.007779   
## 3rd Qu.:0.007338   
## Max. :0.122775

boxplot(norm\_ff)



#3. Data Pre-Processing cont...  
#We can now apply our labels to our classifier attribute 'Area'.  
#Since a large portion of our dataset contains 0 burned area quantity, we will label it as 'Zero Burn'. Burned area between are 1st quartile of 0 and our 3rd quartile, can be considered 'Low/Med Burn', and anything from our 3rd quartile to our Max value is considered 'High Burn'  
norm\_ff$area <- cut(norm\_ff$area,  
 breaks = c(-1,0,0.00734,0.13),  
 labels = c("Zero Burn","Low/Med Burn","High Burn"), right = TRUE)  
summary(norm\_ff$area)

## Zero Burn Low/Med Burn High Burn   
## 247 134 127

#As we can see below, we have 247 combination of attributes that result in Zero Burn area, 134 result in Low/Med Burn areas and 127 that result in High Burn area.

#4. Predictive Modeling  
#we can now begin our predictive modeling on this dataset...

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.