Project 1 - DS-600

## Introduction

Every year, there are many forest fires occur worldwide. To protect the natural wildlife, numerous attempts has been made to prevent forests from getting fire. In this project, we will be outlining a summary in about a forest fire that occurred in a region at Portugal. The data is collected by using metheorological and other data. For more information about the dataset, this [link](http://www3.dsi.uminho.pt/pcortez/forestfires/forestfires-names.txt) can be followed.

The following is the dataset information.

# reading   
forestfires = read.csv('forestfires.csv')

## Data Cleaning

With the dataset, we will first be investigating the variables’ data type and general shape of occurance. We will be examining data in terms of the following objectives:

1. Variable data types
2. Empty or NA values

Below, we are presenting the general outline of the dataset to understand the structure of our dataset. To use the some functions, we have to install the library “dplyr” first.

head(forestfires)

## 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 NA 35.4 669.1 6.7 18.0 33 0.9 0.0 0  
## 3 7 4 oct sat NA 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

class(forestfires)

## [1] "data.frame"

dim(forestfires)

## [1] 517 13

names(forestfires)

## [1] "X" "Y" "month" "day" "FFMC" "DMC" "DC" "ISI"   
## [9] "temp" "RH" "wind" "rain" "area"

glimpse(forestfires)

## Observations: 517  
## Variables: 13  
## $ X <int> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6,...  
## $ Y <int> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4,...  
## $ month <fct> mar, oct, oct, mar, mar, aug, aug, aug, sep, sep, sep, s...  
## $ day <fct> fri, tue, sat, fri, sun, sun, mon, mon, tue, sat, sat, s...  
## $ FFMC <dbl> 86.2, NA, NA, 91.7, 89.3, 92.3, 92.3, 91.5, 91.0, 92.5, ...  
## $ DMC <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, ...  
## $ DC <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 69...  
## $ ISI <dbl> 5.1, 6.7, 6.7, 9.0, 9.6, 14.7, 8.5, 10.7, 7.0, 7.1, 7.1,...  
## $ temp <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8,...  
## $ RH <int> 51, 33, 33, 97, 99, 29, 27, 86, 63, 40, 51, 38, 72, 42, ...  
## $ wind <dbl> 6.7, 0.9, 1.3, 4.0, 1.8, 5.4, 3.1, 2.2, 5.4, 4.0, 7.2, 4...  
## $ rain <dbl> 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0...  
## $ area <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

summary(forestfires)

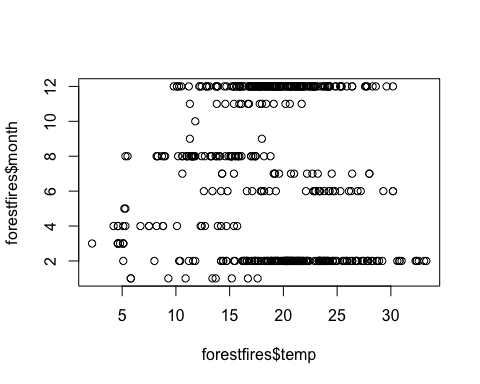
## X Y month day FFMC   
## Min. :1.000 Min. :2.0 aug :184 fri:85 Min. :18.70   
## 1st Qu.:3.000 1st Qu.:4.0 sep :172 mon:74 1st Qu.:90.20   
## Median :4.000 Median :4.0 mar : 54 sat:84 Median :91.65   
## Mean :4.669 Mean :4.3 jul : 32 sun:95 Mean :90.65   
## 3rd Qu.:7.000 3rd Qu.:5.0 feb : 20 thu:61 3rd Qu.:92.90   
## Max. :9.000 Max. :9.0 jun : 17 tue:64 Max. :96.20   
## (Other): 38 wed:54 NA's :7   
## DMC DC ISI temp   
## Min. : 1.1 Min. : 7.9 Min. : 0.000 Min. : 2.20   
## 1st Qu.: 68.6 1st Qu.:437.7 1st Qu.: 6.500 1st Qu.:15.50   
## Median :108.3 Median :664.2 Median : 8.400 Median :19.30   
## Mean :110.9 Mean :547.9 Mean : 9.022 Mean :18.89   
## 3rd Qu.:142.4 3rd Qu.:713.9 3rd Qu.:10.800 3rd Qu.:22.80   
## Max. :291.3 Max. :860.6 Max. :56.100 Max. :33.30   
##   
## RH wind rain area   
## Min. : 15.00 Min. :0.400 Min. :0.00000 Min. : 0.00   
## 1st Qu.: 33.00 1st Qu.:2.700 1st Qu.:0.00000 1st Qu.: 0.00   
## Median : 42.00 Median :4.000 Median :0.00000 Median : 0.52   
## Mean : 44.29 Mean :4.018 Mean :0.02166 Mean : 12.85   
## 3rd Qu.: 53.00 3rd Qu.:4.900 3rd Qu.:0.00000 3rd Qu.: 6.57   
## Max. :100.00 Max. :9.400 Max. :6.40000 Max. :1090.84   
##

We observed that FFMC variable has some missing values. Therefore, we decided to update values with replacing NAs with mean of it.

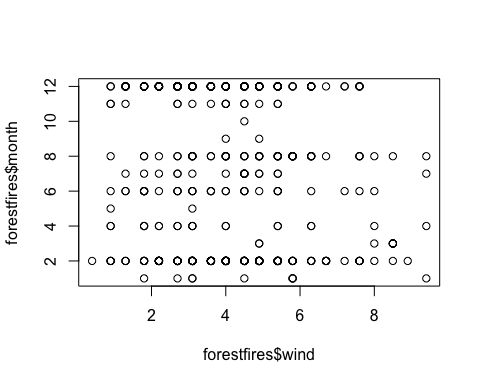
# finding indices of na  
ind = which(is.na(forestfires$FFMC))  
  
#replacing empty data with 0  
forestfires$FFMC[ind]=0  
  
#replacing empty data with 0  
forestfires$FFMC[ind]=mean(forestfires$FFMC)  
  
# removing outliers  
forestfires <- filter(forestfires, area < 150)

Now, we want to see how is the behavior of the weather through the time. Let’s take a look what are the values of some variables, as temperature, humidity, wind and rain in the months of this year. Doing this, we can know more about the weather of this zone of Portugal.

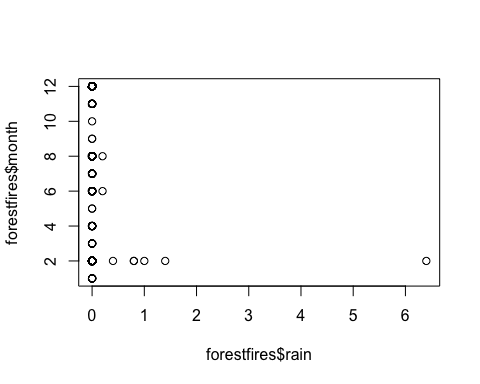
# making plots of temperature in Celsius, wind in km/h, rain in mm/m2, and humidty (RH) in %   
plot(forestfires$temp, forestfires$month)



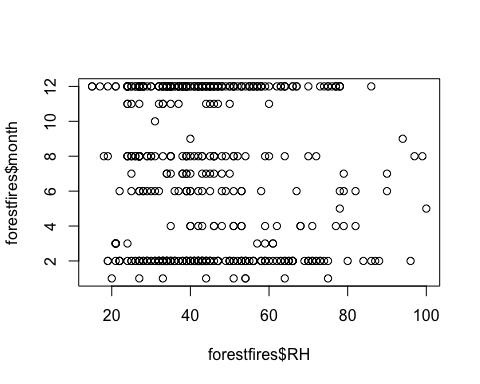
plot(forestfires$wind, forestfires$month)



plot(forestfires$rain, forestfires$month)

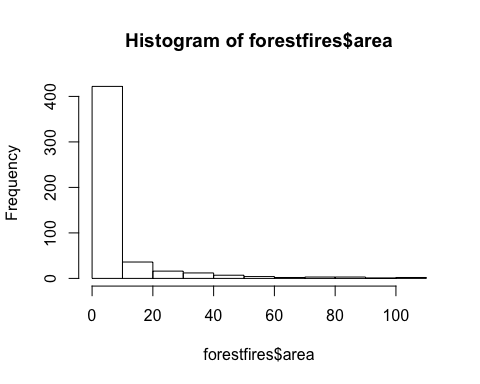


plot(forestfires$RH, forestfires$month)



Another way to see the variables is through the histograms. Let’s see how is the behavior of some variables, as burned area of the forest.

# making histogram of burned area of the forest  
hist(forestfires$area)



### Gathering the Information

The data presents the events of fires in x, y spatial coordinates within the Montesinho map. Other columns represents the features of each observed variable. For the sake of analysis, we may need to do some operations within the data. The following will present some manipulation examples that may occur during data analysis.

# unite coordinates  
forestfires\_t1 <- unite(forestfires, "coordinates", c(X, Y))  
  
# changing data  
forestfires\_t2 <- mutate(forestfires\_t1, coordinates = paste("coor", coordinates, sep = "\_"))  
  
# filtering the data with some parameters  
forestfires\_t3 <- filter(forestfires\_t2, month == "mar", day == "fri")  
  
# summarize all the columns with mean  
summarise\_if(forestfires\_t3, is.numeric, funs(round(mean(., na.rm = T), 2)))

## FFMC DMC DC ISI temp RH wind rain area  
## 1 90.01 34.46 81.19 8.36 14 38.36 5.86 0.02 0.99

# doing multiple things at once  
forestfires %>%   
 mutate(ISIandTemp = ISI \* temp, rain = exp(rain), area = log(2)) %>%   
 select(X:day,ISIandTemp, ISI, temp, area) %>%  
 arrange(desc(ISIandTemp)) %>%  
 filter(temp > 32)

## X Y month day ISIandTemp ISI temp area  
## 1 4 5 aug mon 547.68 16.8 32.6 0.6931472  
## 2 6 5 aug tue 476.19 14.3 33.3 0.6931472  
## 3 2 5 aug sun 466.71 14.1 33.1 0.6931472  
## 4 3 4 aug tue 461.89 14.3 32.3 0.6931472  
## 5 4 4 aug thu 447.12 13.8 32.4 0.6931472  
## 6 1 3 aug fri 366.12 11.3 32.4 0.6931472

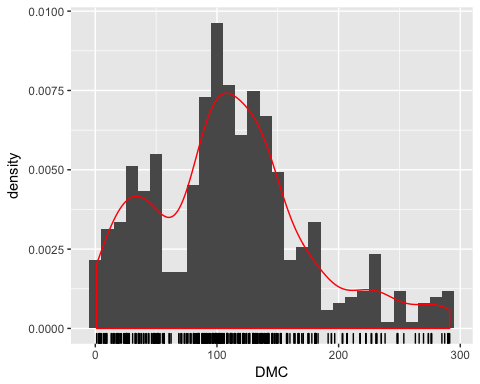
# grouping the data  
forestfires %>%  
 group\_by(month) %>%  
 summarise(  
 n = n(),  
 RainTotal = sum(rain),  
 WindAverage = mean(wind),  
 areaTotal = sum(area)  
 )

## # A tibble: 12 x 5  
## month n RainTotal WindAverage areaTotal  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 apr 9 0 4.67 80.0  
## 2 aug 180 10.8 4.08 995   
## 3 dec 9 0 7.64 120   
## 4 feb 20 0 3.76 126   
## 5 jan 2 0 2.00 0   
## 6 jul 31 0.200 3.70 181   
## 7 jun 17 0 4.14 99.3  
## 8 mar 54 0.200 4.97 235   
## 9 may 2 0 4.45 38.5  
## 10 nov 1 0 4.50 0   
## 11 oct 15 0 3.46 99.6  
## 12 sep 168 0 3.58 1427

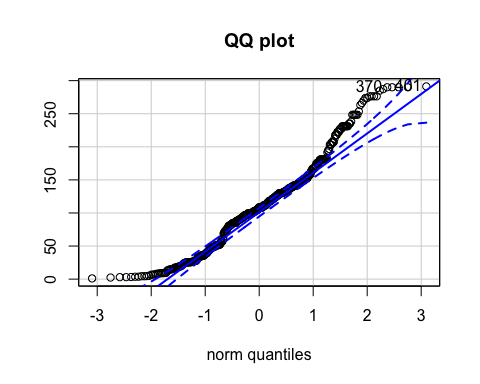
### Using GGPlot

ggplot(forestfires, aes(x=DMC)) +   
 geom\_histogram(aes(y = ..density..)) +   
 geom\_density(color = "red") +   
 geom\_rug()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



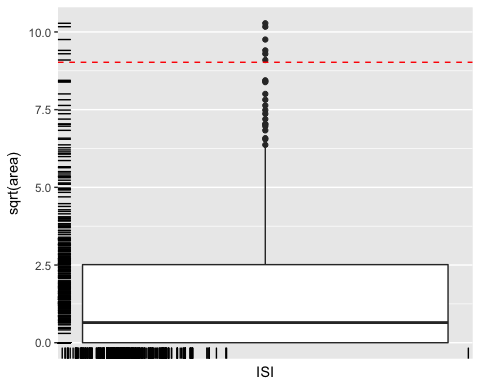
qqPlot(forestfires$DMC, main = "QQ plot", ylab = "")



## [1] 401 370

ggplot(forestfires, aes(x=ISI, y=sqrt(area))) +  
 geom\_boxplot() +   
 geom\_rug() +   
 geom\_hline(aes(yintercept=mean(forestfires$ISI, na.rm = T)), linetype = 2, color = "red") +  
 scale\_x\_discrete(breaks = NULL)

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



### Correlation between features

symnum(cor(forestfires[5:12], use = "complete.obs"))

## F DM DC I t R w r  
## FFMC 1   
## DMC . 1   
## DC . , 1   
## ISI . . 1   
## temp . . . . 1   
## RH . 1   
## wind 1   
## rain 1  
## attr(,"legend")  
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '\*' 0.95 'B' 1

### Correlation Plot

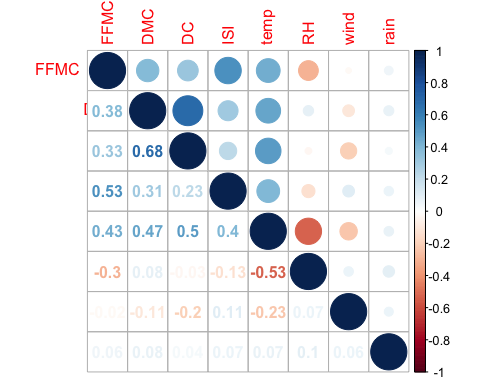
cm <- cor(forestfires[,5:12], use="complete.obs")  
corrplot(cm, type="upper", tl.pose="d")

## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt  
## = tl.srt, : "tl.pose" is not a graphical parameter

## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col  
## = tl.col, : "tl.pose" is not a graphical parameter

## Warning in title(title, ...): "tl.pose" is not a graphical parameter

corrplot(cm,add=TRUE, type="lower", method="number",diag=FALSE, tl.pos="n", cl.pos="n")



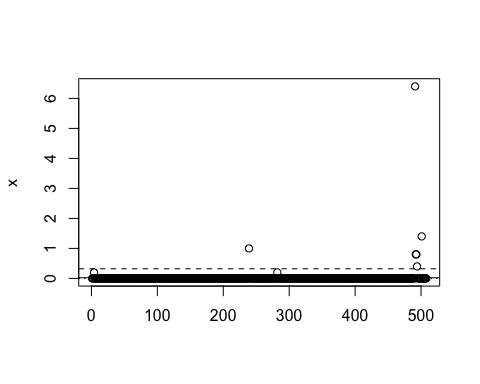
### Linear Model

fit <- lm(area ~ ., forestfires)   
summary(fit)

##   
## Call:  
## lm(formula = area ~ ., data = forestfires)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.835 -6.977 -3.681 0.716 89.216   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.66891 17.99215 -0.371 0.7111   
## X 0.56999 0.35332 1.613 0.1074   
## Y -0.50643 0.67260 -0.753 0.4519   
## monthaug 6.58057 8.98867 0.732 0.4645   
## monthdec 16.72265 8.69402 1.923 0.0550 .  
## monthfeb -0.60077 6.09565 -0.099 0.9215   
## monthjan -0.78949 13.24999 -0.060 0.9525   
## monthjul 1.90263 7.80402 0.244 0.8075   
## monthjun -1.04387 7.13962 -0.146 0.8838   
## monthmar -4.13685 5.49967 -0.752 0.4523   
## monthmay 10.53541 11.96215 0.881 0.3789   
## monthnov -10.73012 16.06261 -0.668 0.5044   
## monthoct 15.19918 10.70342 1.420 0.1562   
## monthsep 15.41145 10.08857 1.528 0.1273   
## daymon 1.00248 2.47156 0.406 0.6852   
## daysat 1.65235 2.39276 0.691 0.4902   
## daysun 3.28983 2.30547 1.427 0.1542   
## daythu -0.63423 2.62049 -0.242 0.8089   
## daytue 5.13023 2.56920 1.997 0.0464 \*  
## daywed 2.96581 2.69129 1.102 0.2710   
## FFMC 0.05962 0.18073 0.330 0.7416   
## DMC 0.04647 0.02056 2.261 0.0242 \*  
## DC -0.03291 0.01388 -2.370 0.0182 \*  
## ISI -0.23348 0.19585 -1.192 0.2338   
## temp 0.49965 0.24467 2.042 0.0417 \*  
## RH 0.02173 0.06842 0.318 0.7509   
## wind 0.59722 0.42068 1.420 0.1564   
## rain -1.08735 2.33327 -0.466 0.6414   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15 on 480 degrees of freedom  
## Multiple R-squared: 0.06624, Adjusted R-squared: 0.01371   
## F-statistic: 1.261 on 27 and 480 DF, p-value: 0.1736

### Distribution of the Values of Columns

plot1 = function(x) {plot(x,xlab="")  
 abline(h=mean(x,na.rm=T),lty=1)  
 abline(h=mean(x,na.rm=T)+sd(x,na.rm=T),lty=2)  
 abline(h=median(x,na.rm=T),lty=3)}  
  
plot1(forestfires$rain)



## Conclusion

The data after cleaning operations is ready for statistical analysis. After operations of tidying, restructing and removing outliers, we will have the data that is ready for further analysis.

After analysis on the report, we evaluate the model in terms of correlation between features. After, we run a linear model to see the results. Above, we represent our findings and plot features according to predictors.