Forest Fire Exploration

Shan He

## Background

In some areas, forest fires are a major environmental concern, endangering human lives and causing substantial economic damage. Your team has been hired by an agency that is developing an early warning system to identify particularly damaging forest fires.

## Data

You are provided with the file **forestfires.csv**, containing measurements taken of recent fires in a Portuguese park.

As a proxy for how damaging a fire is, you should use the area variable, representing the region burned in hectares.

The dataset includes a number of meteorological variables. Some of these come from the forest Fire Weather Index (FWI), a Canadian system for rating fire danger. These include Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI).

The following codebook summarizes each variable:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "jan" to "dec"
4. day - day of the week: "mon" to "sun"
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in hectares): 0.00 to 1090.84

## Objective

Conduct an exploratory analysis with the aim of understanding what factors lead to particularly damaging forest fires.

## Setup

First, we load the car library, which gives us a convenient scatterplotMatrix function.

library(car)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

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

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

library(ggplot2)

### Introduction (20 pts)

State the research question that motivates your analysis.

Load your data set into R.

# Load the data   
getwd()

## [1] "/Users/shanhe/Desktop/W203/Lab/Lab\_1/Forest\_Fire\_EDA"

f <- read.csv("forestfires.csv")

Describe your data set. What types of variables does it contain? How many observations are there?

nrow(f)

## [1] 517

str(f)

## '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: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12 ...  
## $ day : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...  
## $ 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 ...

Evaluate the data quality. Are there any issues with the data? Explain how you handled these potential issues.

# factor month has undesired order  
# when we display graph involving month, we want Jan -> Dec  
levels(f$month)

## [1] "apr" "aug" "dec" "feb" "jan" "jul" "jun" "mar" "may" "nov" "oct"  
## [12] "sep"

#reorder month factor  
f$month = factor(f$month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec"))  
# double check reordered levels  
levels(f$month)

## [1] "jan" "feb" "mar" "apr" "may" "jun" "jul" "aug" "sep" "oct" "nov"  
## [12] "dec"

# factor day has undesired order  
# when we display graph involving day, we want Mon -> Sun  
levels(f$day)

## [1] "fri" "mon" "sat" "sun" "thu" "tue" "wed"

# reorder day factor  
f$day = factor(f$day, levels = c("mon", "tue", "wed", "thu", "fri", "sat", "sun"))  
# double check reordered levels  
levels(f$day)

## [1] "mon" "tue" "wed" "thu" "fri" "sat" "sun"

# data frame looks good  
summary(f)

## X Y month day FFMC   
## Min. :1.000 Min. :2.0 aug :184 mon:74 Min. :18.70   
## 1st Qu.:3.000 1st Qu.:4.0 sep :172 tue:64 1st Qu.:90.20   
## Median :4.000 Median :4.0 mar : 54 wed:54 Median :91.60   
## Mean :4.669 Mean :4.3 jul : 32 thu:61 Mean :90.64   
## 3rd Qu.:7.000 3rd Qu.:5.0 feb : 20 fri:85 3rd Qu.:92.90   
## Max. :9.000 Max. :9.0 jun : 17 sat:84 Max. :96.20   
## (Other): 38 sun:95   
## 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   
##

Explain whether any data processing or preparation is required for your data set.

# create categorical variables  
f\_new <- mutate(f,  
 sun = ifelse(f$day=="sun", 1, 0),  
 mon = ifelse(f$day=="mon", 1, 0),  
 tue = ifelse(f$day=="tue", 1, 0),  
 wed = ifelse(f$day=="wed", 1, 0),  
 thu = ifelse(f$day=="thu", 1, 0),  
 fri = ifelse(f$day=="fri", 1, 0),  
 sat = ifelse(f$day=="sat", 1, 0),  
 time.of.week = ifelse(f$day == "sat" | f$day == "sun", "Weekend", "Weekday"),  
 weekend\_n = ifelse(f$day == "sat" | f$day == "sun", 1, 0),  
 season = ifelse(f$month == "mar" | f$month == "apr" | f$month == "may", "spring",  
 ifelse(f$month == "jun" | f$month == "jul" | f$month == "aug", "summer",  
 ifelse(f$month == "sep" | f$month == "oct" | f$month == "nov", "fall",  
 ifelse(f$month == "dec" | f$month == "jan" | f$month == "feb", "winter", "NA")))),  
 spring\_n = ifelse(f$month == "mar" | f$month == "apr" | f$month == "may", 1, 0),  
 summer\_n = ifelse(f$month == "jun" | f$month == "jul" | f$month == "aug", 1, 0 ),  
 fall\_n = ifelse(f$month == "sep" | f$month == "oct" | f$month == "nov", 1, 0 ),  
 winter\_n = ifelse(f$month == "dec" | f$month == "jan" | f$month == "feb", 1, 0),  
 fire = ifelse(f$area == 0, 0, 1)  
)

### Univariate Analysis of Key Variables (20 pts)

Use visualizations and descriptive statistics to perform a univariate analysis of each key variable. Be sure to describe any anomalies, coding issues, or potentially erroneous values. Explain how you respond to each issue you identify. Note any features that appear relevant to statistical analysis. Discuss what transformations may be appropriate for each variable.

## Key Variable - Area

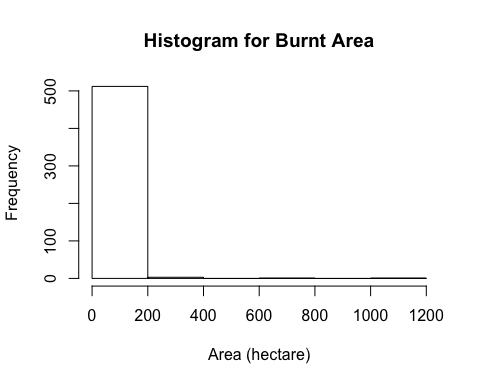
summary(f$area)

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

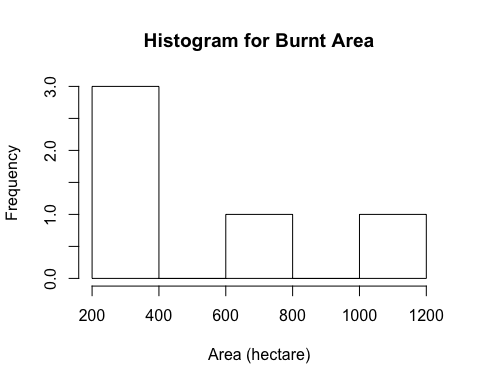
summary(sqrt(f$area)) #positive skewness correction

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.7211 1.8009 2.5632 33.0279

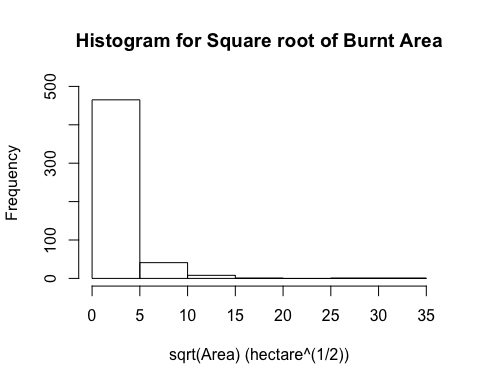
# the majority of the fires caused burnt area < 200 hectares, Strong positive skewness  
hist(f$area,   
 main="Histogram for Burnt Area",   
 xlab="Area (hectare)",   
 xlim=c(0,1200),  
 breaks=6)



# if we look at only the fires that caused burnt area >= 200 hectares  
hist(subset(f$area, f$area >= 200),  
 main="Histogram for Burnt Area",   
 xlab="Area (hectare)",   
 xlim=c(200,1200),  
 breaks=6)



#if we transform to square root of area  
hist(sqrt(f$area),  
 main="Histogram for Square root of Burnt Area",   
 xlab="sqrt(Area) (hectare^(1/2))",  
 ylim = c(0,500)  
 )

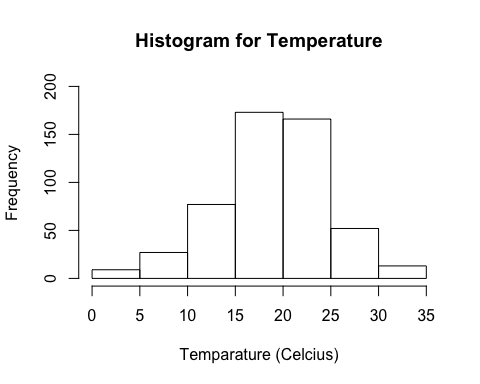


## Key Variable - temp

summary(f$temp)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.20 15.50 19.30 18.89 22.80 33.30

# seems to follow a normal distribution  
hist(f$temp,   
 main="Histogram for Temperature",   
 ylim = c(0, 200),  
 xlab="Temparature (Celcius)")

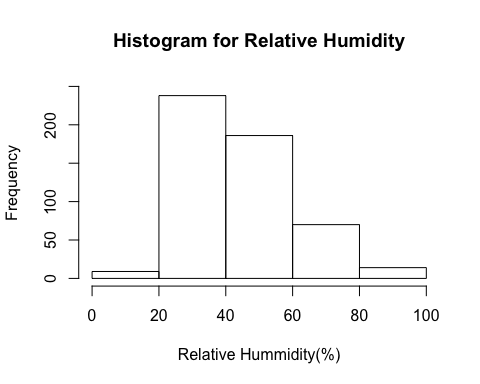


## Key Variable - RH (relative humidity)

summary(f$RH)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.00 33.00 42.00 44.29 53.00 100.00

hist(f$RH,   
 main="Histogram for Relative Humidity",   
 xlab="Relative Hummidity(%)",  
 xlim = c(0, 100),  
 ylim = c(0,250),  
 breaks = 0:5 \* 20  
 )

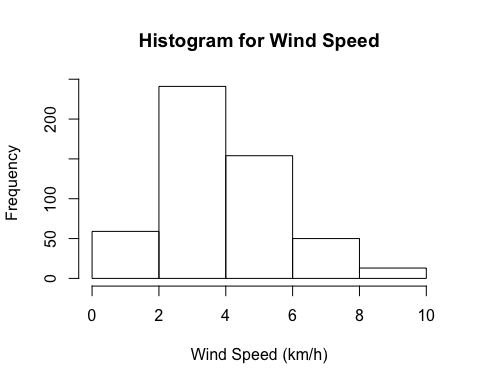


## Key Variable - Wind

summary(f$wind)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.400 2.700 4.000 4.018 4.900 9.400

hist(f$wind,   
 main="Histogram for Wind Speed",   
 xlab="Wind Speed (km/h)",  
 #xlim = c(0, 100),  
 #ylim = c(0,250),  
 breaks = 0:5 \* 2  
 )

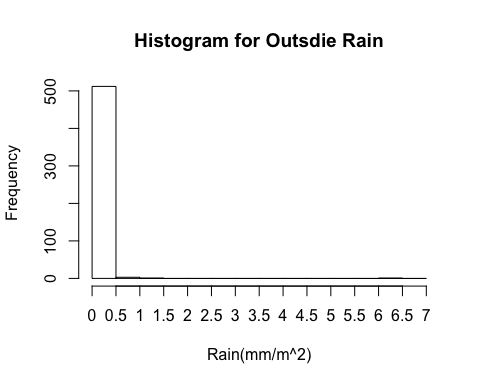


## Key Variable - Rain

summary(f$rain)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.02166 0.00000 6.40000

# majority are 0 values  
hist(f$rain,   
 main="Histogram for Outsdie Rain",   
 xlab="Rain(mm/m^2)",  
 breaks = 0:14 \* 0.5  
 )  
axis(1, at = 0:6 + 0.5)

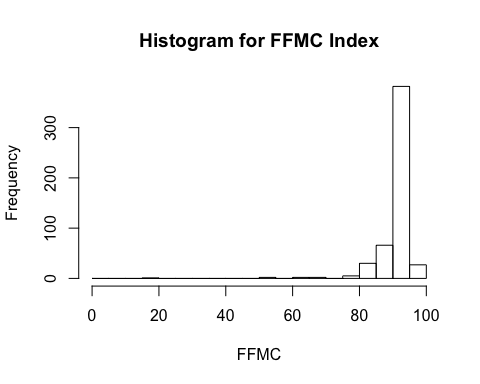


## Key Variable - FFMC

summary(f$FFMC)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.70 90.20 91.60 90.64 92.90 96.20

hist(f$FFMC,   
 main="Histogram for FFMC Index",   
 xlab="FFMC",  
 xlim = c(0,100),  
 breaks = 0:20\*5  
 )

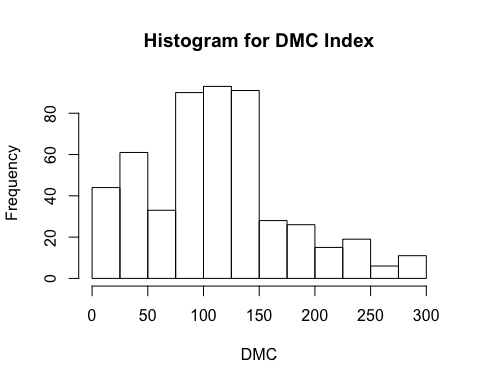


## Key Variable - DMC

summary(f$DMC)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.1 68.6 108.3 110.9 142.4 291.3

hist(f$DMC,   
 main="Histogram for DMC Index",   
 xlab="DMC",  
 breaks = 0:12 \* 25  
 )

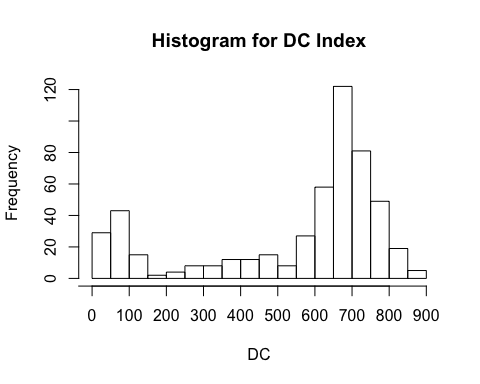


## Key Variable - DC

summary(f$DC)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.9 437.7 664.2 547.9 713.9 860.6

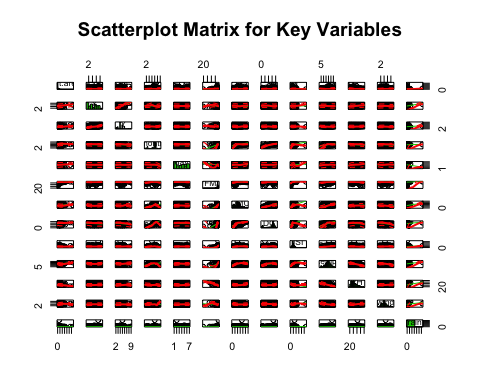
hist(f$DC,   
 main="Histogram for DC Index",   
 xlab="DC",  
 xlim = c(0, 900),  
 breaks = 0:18 \* 50  
 )  
axis(1,at = 0:5\*200 - 100)



### Analysis of Key Relationships (30 pts)

Explore how your outcome variable is related to the other variables in your dataset. Make sure to use visualizations to understand the nature of each bivariate relationship.

scatterplotMatrix(~ sqrt(area) + X + Y + month + day + FFMC + DMC + DC + ISI + temp + RH + wind + rain, data = f,  
 main = "Scatterplot Matrix for Key Variables",   
 diagonal = "histogram")



f\_num <- f  
f\_num$month <- as.numeric(f\_num$month)  
f\_num$day <- as.numeric(f\_num$day)  
  
f\_num\_sqrt <- f  
f\_num\_sqrt$month <- as.numeric(f\_num$month)  
f\_num\_sqrt$day <- as.numeric(f\_num$day)  
f\_num\_sqrt$area <- sqrt(f\_num$area)  
  
cor(f\_num, use= "complete.obs")

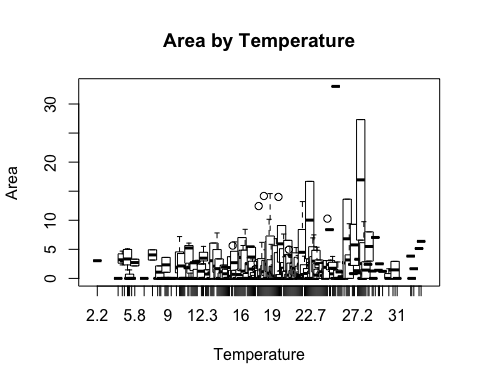
## X Y month day FFMC  
## X 1.000000000 0.539548171 -0.06500303 -0.0249218945 -0.02103927  
## Y 0.539548171 1.000000000 -0.06629179 -0.0054533368 -0.04630755  
## month -0.065003032 -0.066291786 1.00000000 -0.0508365920 0.29147677  
## day -0.024921895 -0.005453337 -0.05083659 1.0000000000 -0.04106833  
## FFMC -0.021039272 -0.046307546 0.29147677 -0.0410683308 1.00000000  
## DMC -0.048384178 0.007781561 0.46664525 0.0628703973 0.38261880  
## DC -0.085916123 -0.101177767 0.86869776 0.0001049027 0.33051180  
## ISI 0.006209941 -0.024487992 0.18659697 0.0329092595 0.53180493  
## temp -0.051258262 -0.024103084 0.36884151 0.0521903410 0.43153226  
## RH 0.085223194 0.062220731 -0.09528038 0.0921514374 -0.30099542  
## wind 0.018797818 -0.020340852 -0.08636797 0.0324781638 -0.02848481  
## rain 0.065387168 0.033234103 0.01343813 -0.0483401530 0.05670153  
## area 0.063385299 0.044873225 0.05649577 0.0232263716 0.04012200  
## DMC DC ISI temp RH  
## X -0.048384178 -0.0859161229 0.006209941 -0.05125826 0.08522319  
## Y 0.007781561 -0.1011777674 -0.024487992 -0.02410308 0.06222073  
## month 0.466645252 0.8686977586 0.186596974 0.36884151 -0.09528038  
## day 0.062870397 0.0001049027 0.032909260 0.05219034 0.09215144  
## FFMC 0.382618800 0.3305117952 0.531804931 0.43153226 -0.30099542  
## DMC 1.000000000 0.6821916120 0.305127835 0.46959384 0.07379494  
## DC 0.682191612 1.0000000000 0.229154169 0.49620805 -0.03919165  
## ISI 0.305127835 0.2291541691 1.000000000 0.39428710 -0.13251718  
## temp 0.469593844 0.4962080531 0.394287104 1.00000000 -0.52739034  
## RH 0.073794941 -0.0391916472 -0.132517177 -0.52739034 1.00000000  
## wind -0.105342253 -0.2034656909 0.106825888 -0.22711622 0.06941007  
## rain 0.074789982 0.0358608620 0.067668190 0.06949055 0.09975122  
## area 0.072994296 0.0493832253 0.008257688 0.09784411 -0.07551856  
## wind rain area  
## X 0.01879782 0.065387168 0.063385299  
## Y -0.02034085 0.033234103 0.044873225  
## month -0.08636797 0.013438129 0.056495774  
## day 0.03247816 -0.048340153 0.023226372  
## FFMC -0.02848481 0.056701533 0.040122004  
## DMC -0.10534225 0.074789982 0.072994296  
## DC -0.20346569 0.035860862 0.049383225  
## ISI 0.10682589 0.067668190 0.008257688  
## temp -0.22711622 0.069490547 0.097844107  
## RH 0.06941007 0.099751223 -0.075518563  
## wind 1.00000000 0.061118880 0.012317277  
## rain 0.06111888 1.000000000 -0.007365729  
## area 0.01231728 -0.007365729 1.000000000

cor(f\_num\_sqrt, use= "complete.obs")

## X Y month day FFMC  
## X 1.000000000 0.539548171 -0.06500303 -0.0249218945 -0.02103927  
## Y 0.539548171 1.000000000 -0.06629179 -0.0054533368 -0.04630755  
## month -0.065003032 -0.066291786 1.00000000 -0.0508365920 0.29147677  
## day -0.024921895 -0.005453337 -0.05083659 1.0000000000 -0.04106833  
## FFMC -0.021039272 -0.046307546 0.29147677 -0.0410683308 1.00000000  
## DMC -0.048384178 0.007781561 0.46664525 0.0628703973 0.38261880  
## DC -0.085916123 -0.101177767 0.86869776 0.0001049027 0.33051180  
## ISI 0.006209941 -0.024487992 0.18659697 0.0329092595 0.53180493  
## temp -0.051258262 -0.024103084 0.36884151 0.0521903410 0.43153226  
## RH 0.085223194 0.062220731 -0.09528038 0.0921514374 -0.30099542  
## wind 0.018797818 -0.020340852 -0.08636797 0.0324781638 -0.02848481  
## rain 0.065387168 0.033234103 0.01343813 -0.0483401530 0.05670153  
## area 0.070530444 0.041901967 0.09964574 0.0111913097 0.05159795  
## DMC DC ISI temp RH  
## X -0.048384178 -0.0859161229 0.006209941 -0.05125826 0.08522319  
## Y 0.007781561 -0.1011777674 -0.024487992 -0.02410308 0.06222073  
## month 0.466645252 0.8686977586 0.186596974 0.36884151 -0.09528038  
## day 0.062870397 0.0001049027 0.032909260 0.05219034 0.09215144  
## FFMC 0.382618800 0.3305117952 0.531804931 0.43153226 -0.30099542  
## DMC 1.000000000 0.6821916120 0.305127835 0.46959384 0.07379494  
## DC 0.682191612 1.0000000000 0.229154169 0.49620805 -0.03919165  
## ISI 0.305127835 0.2291541691 1.000000000 0.39428710 -0.13251718  
## temp 0.469593844 0.4962080531 0.394287104 1.00000000 -0.52739034  
## RH 0.073794941 -0.0391916472 -0.132517177 -0.52739034 1.00000000  
## wind -0.105342253 -0.2034656909 0.106825888 -0.22711622 0.06941007  
## rain 0.074789982 0.0358608620 0.067668190 0.06949055 0.09975122  
## area 0.083270309 0.0674723473 -0.002292634 0.09060837 -0.07401997  
## wind rain area  
## X 0.01879782 0.065387168 0.070530444  
## Y -0.02034085 0.033234103 0.041901967  
## month -0.08636797 0.013438129 0.099645741  
## day 0.03247816 -0.048340153 0.011191310  
## FFMC -0.02848481 0.056701533 0.051597949  
## DMC -0.10534225 0.074789982 0.083270309  
## DC -0.20346569 0.035860862 0.067472347  
## ISI 0.10682589 0.067668190 -0.002292634  
## temp -0.22711622 0.069490547 0.090608374  
## RH 0.06941007 0.099751223 -0.074019966  
## wind 1.00000000 0.061118880 0.043214922  
## rain 0.06111888 1.000000000 0.006216205  
## area 0.04321492 0.006216205 1.000000000

# area is most strongly correlated with temp and DMC. But they are not strong correlation (<0.1).   
# we can see that FFMC, DMC, DC, and ISI have strong correlation with each other and Month  
# X and Y are strongly correlated  
# rain is not very important in the correlation since it's mostly 0  
# temp has a strong correlation with Month, FFMC, DMC, DC, ISI, RH, and wind

boxplot(area ~ temp, data = f\_num\_sqrt,  
 at = sort(unique(f$temp)),  
 main = "Area by Temperature",  
 xlab = "Temperature",  
 ylab = "Area"  
 )



What tranformations can you apply to clarify the relationships you see in the data? Be sure to justify each transformation you use.

### Analysis of Secondary Effects (10 pts)

What secondary variables might have confounding effects on the relationships you have identified? Explain how these variables affect your understanding of the data.

### Conclusion (20 pts)

Summarize your exploratory analysis. What can you conclude based on your analysis?

# Evaluation

We will evaluate your report for technical correctness, but also clarity and overall effectiveness. A point distribution is provided with the above outline. In addition to these point totals, we will impose penalties for output dumps, unclear language, and other errors.

# Submission

Only one student in the team needs to submit via the ISVC. Make sure that you include the names of all group members in your report.

You must turn in

Use the following naming convention for your files:

* lastname1\_lastname2\_lab1.pdf
* lastname1\_lastname2\_lab1.Rmd