Assignment 2

Inshal Naqvi

Question 1

This exercise relates to the College data set, which can be found in the file College.csv uploaded on the course'spublic webpage(https://scads.eecs.wsu.edu/wp-content/uploads/2021/09/College.csv). The dataset contains a numberof variables for 777 different universities and colleges in the US.

(a) Use the read.csv()function toread the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it as a list 2of lists or use the pandasdataframeto store your data. Call the loaded data college.Ensure that your column headers are not treated as a rowof data.

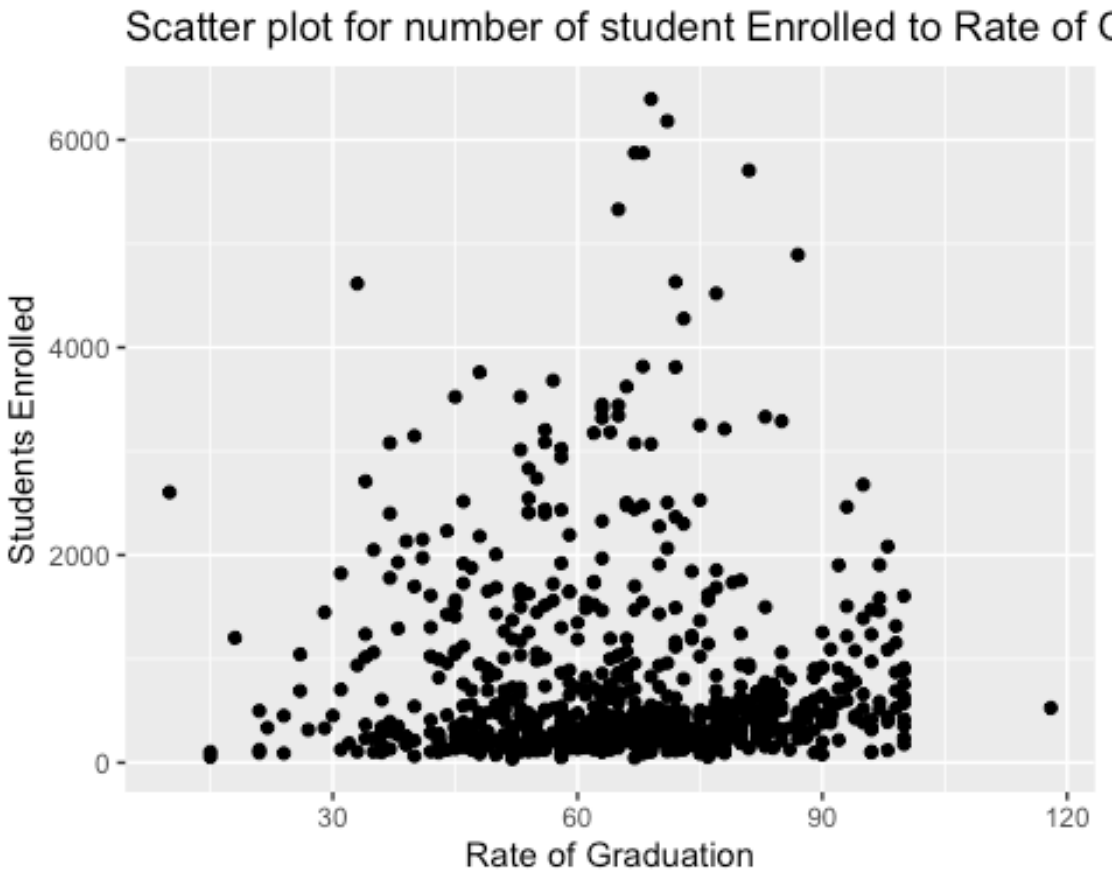
```
colleges = read.csv("college.csv", sep=",", header = TRUE)
```

(b) Find the median cost of room and board (Room.Board)for all schools in this dataset.Then find the median cost of room and board (Room.Board) for both public and private (Private) schools.

```
#Median cost of room and board for all colleges
median(colleges$Room.Board)

## [1] 4200

#Create a subset for private school for median
privmedian <- subset(colleges,Private=="Yes", na.rm=TRUE)
median(privmedian$Room.Board)

## [1] 4400

#Create a subset of public schools for median
pubmedian <- subset(colleges,Private=="No", na.rm=TRUE)
median(pubmedian$Room.Board)

## [1] 3708
```

(c) Produce a scatterplot that shows a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

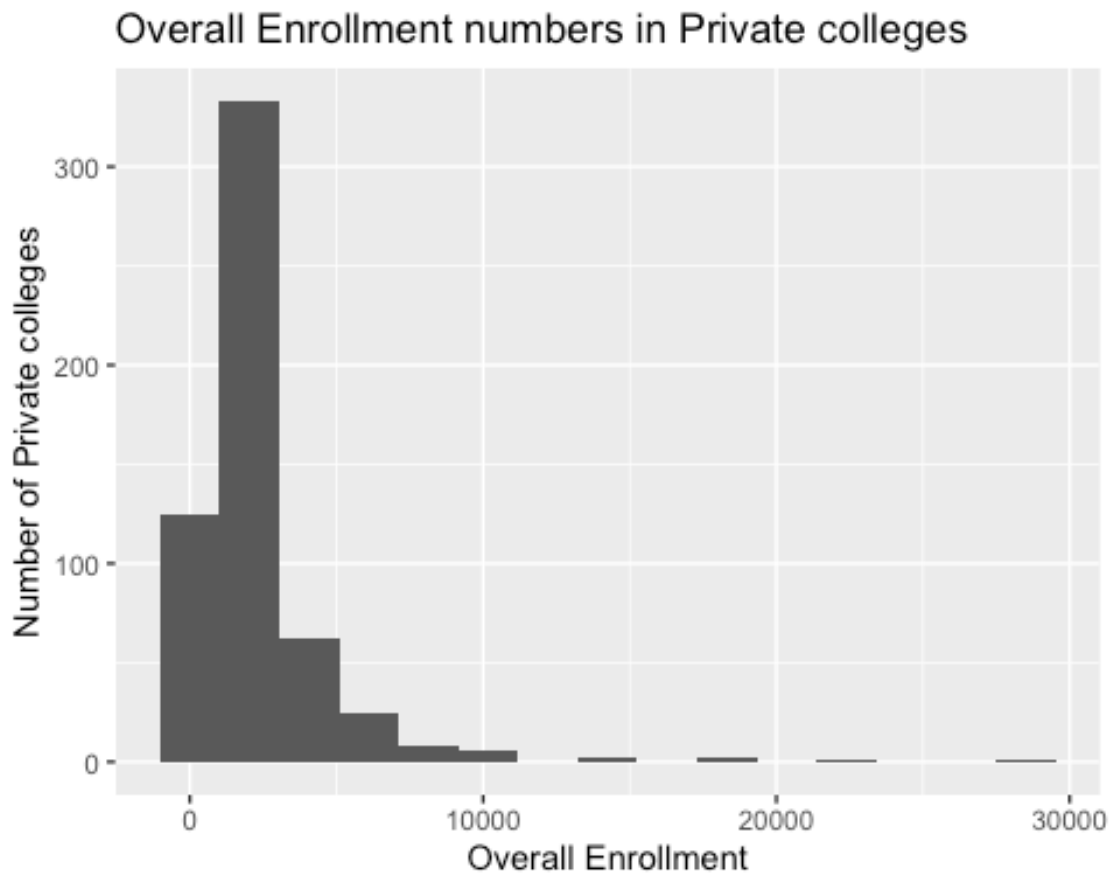
```
library(ggplot2)
#Scatter plot between Graduation rate on X axis and enrollment on y axis
sctrplt<-ggplot(colleges,aes(y=Enroll, x=Grad.Rate)) + geom_point()
sctrplt + labs(x="Rate of Graduation", y="Students Enrolled",
               title= "Scatter plot for number of student Enrolled to Rate of
Graduation")
```



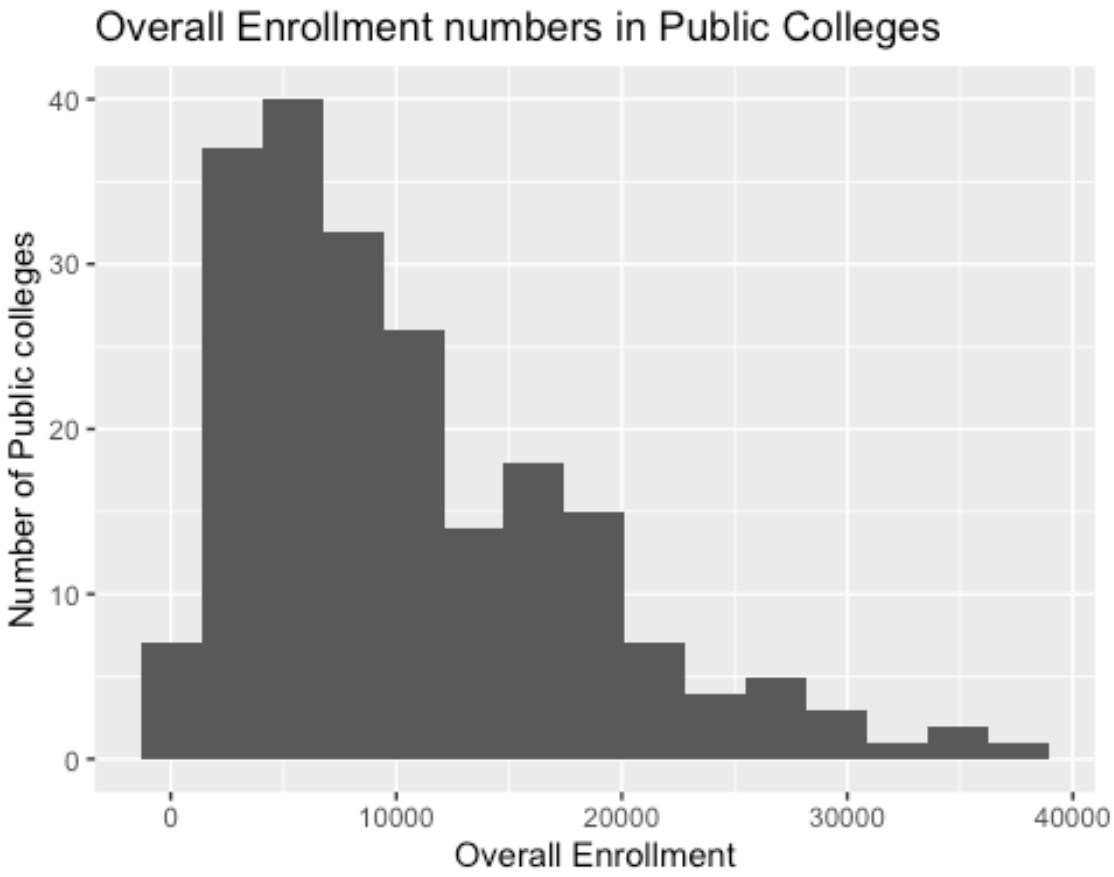
(d)Produce a histogram showing the overall enrollment numbers (P.UndergradplusF.Undergrad)for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools.Ensure whatever figures you produce haveappropriate axis labels and a title.

```
priv_colleges<-subset(colleges, Private=="Yes")

pub_colleges<- subset(colleges, Private=="No")

#Adding enrollment number for private colleges
priv_enrl <- (priv_colleges$P.Undergrad + priv_colleges$F.Undergrad)
#Add ^this col for total enrollment in priv_colleges
priv_colleges$fplusp <- priv_enrl#This line is not needed. For practice
purpose.

#Adding enrollment numbers for public colleges
pub_enrl <- (pub_colleges$P.Undergrad + pub_colleges$F.Undergrad)
#Add ^this col, to public colleges
pub_colleges$fplusp <- pub_enrl #This line is not needed. For practice
purpose.

#Histogram for private school against enrollment
```

```r
priv_hist <- ggplot(priv_colleges, aes(x=priv_enrl)) +
  geom_histogram(bins = 15)
#Adding labels
priv_hist + labs(x="Overall Enrollment", y="Number of Private colleges",
                 title = "Overall Enrollment numbers in Private colleges " )
```




```r
#Histogram for public school against enrollment
pub_hist <- ggplot(pub_colleges, aes(x=pub_enrl)) +
  geom_histogram(bins = 15)
#Adding labels
pub_hist +labs(x="Overall Enrollment", y="Number of Public colleges",
               title = "Overall Enrollment numbers in Public Colleges")
```

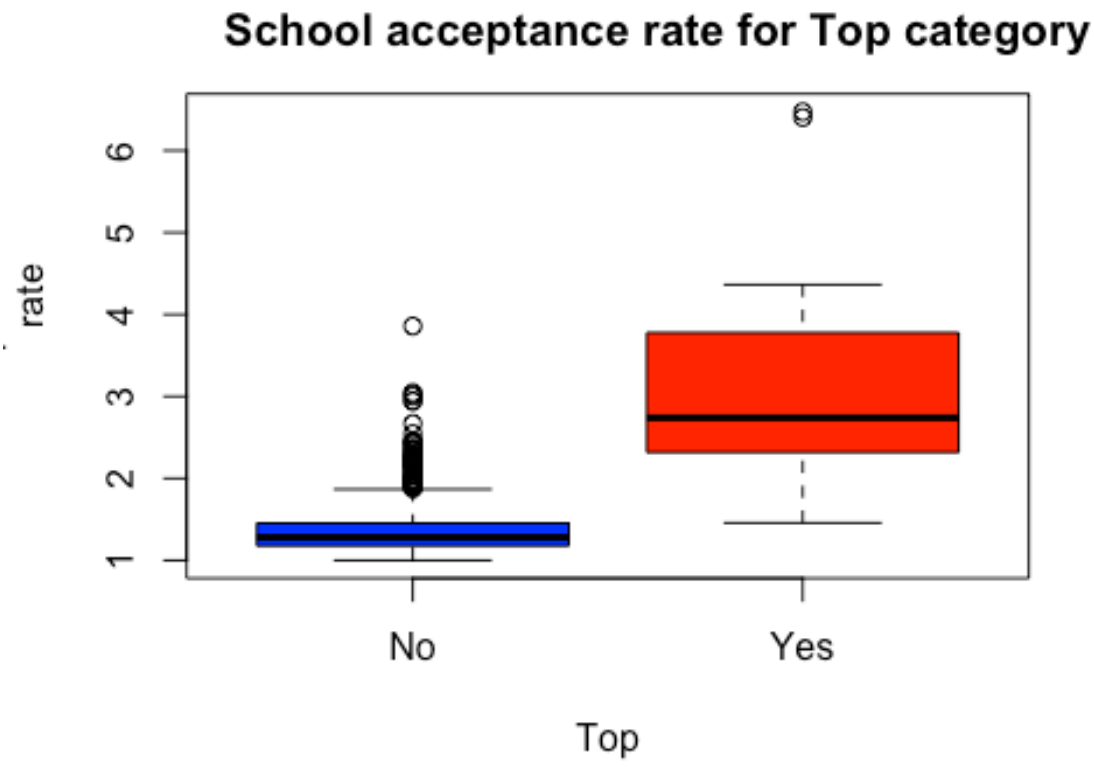


Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories(Yes and No). Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 75%.Now produceside-by-side boxplots ofthe schools' acceptance rates(based on Acceptand Apps)for each ofthe twoTopcategories. There should be two boxes on your figure, one for top schools and one for others. How many top universities are there?

Answer: 22

```
library(ggplot2)
#Binnig data into yes and no categories
Top = rep("No", nrow(colleges))
Top[colleges$Top10perc > 75] = "Yes"
Top = as.factor(Top)

colleges = data.frame(colleges, Top)
summary(colleges$Top)

##  No Yes
## 755  22

#Boxplot for acceptance rate in the above categories
#College acceptance rate
colleges$acceptance <- (colleges$Apps / colleges$Accept)
```

```
#boxplot
boxplot(colleges$acceptance ~ colleges$Top, col = c("blue", "red"), main ="
        School acceptance rate for Top category", xlab="Top",
ylab="Acceptance
        rate")
```

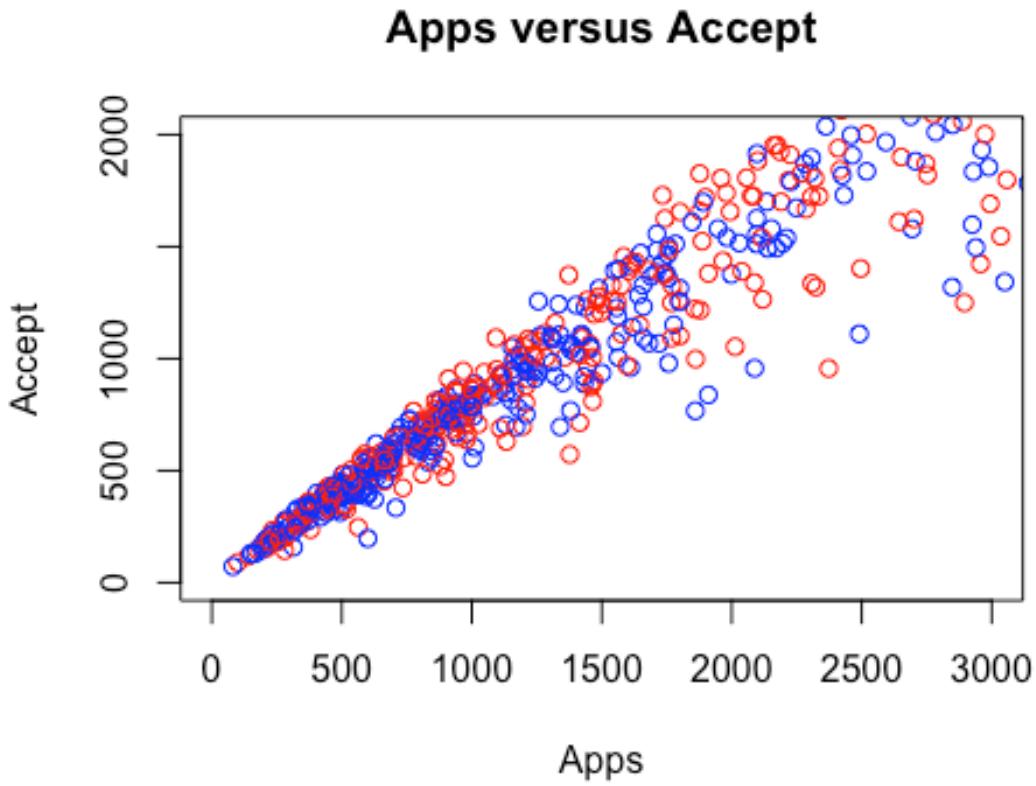


(f)

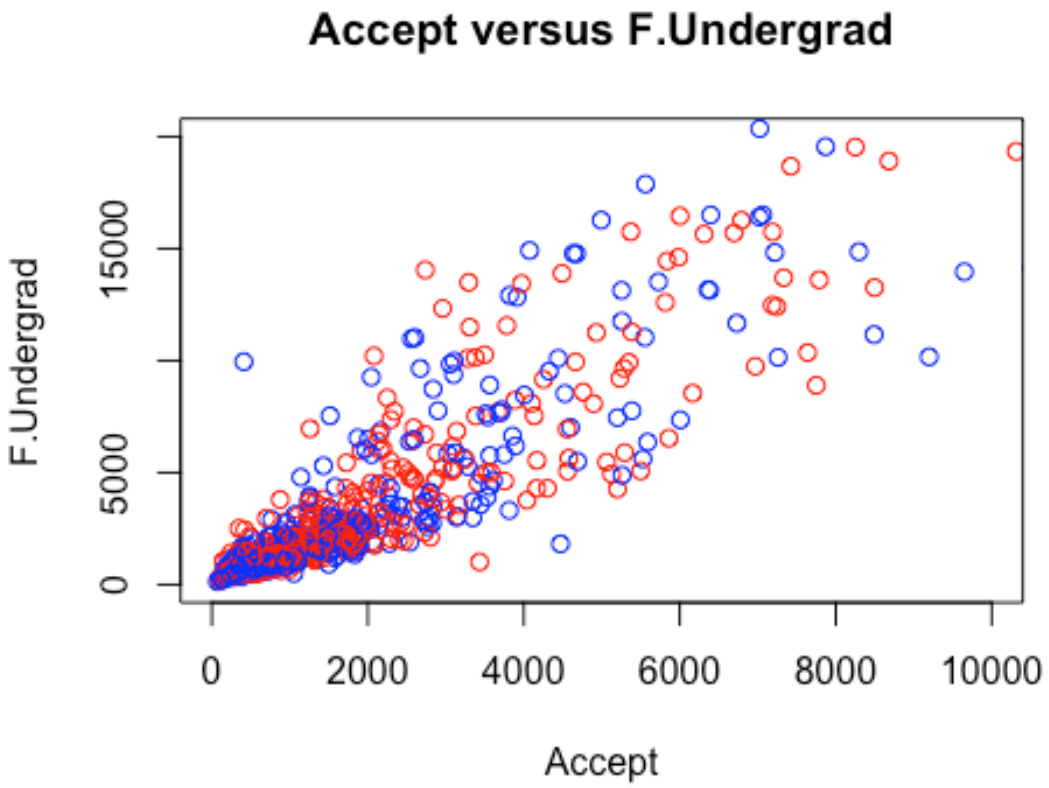
Continue exploring the data, producing two new plots of any type, and provide a brief(one to two sentence)summary of your hypotheses and what you discover.Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

There is almost a linear relationship between Apps and Accept. The relationship Accept and F.undergrad also seem linear but spreads out when it moves away from orign.

```
plot(x = colleges$Apps, y = colleges$Accept, xlim=c(0,3000), ylim=c(0,2000),
xlab ="Apps", ylab = "Accept", main = "Apps versus Accept", col =
c("blue","red"))
```




```
plot(x = colleges$Accept, y = colleges$F.Undergrad, xlim=c(0,10000),
ylim=c(0,20000), xlab ="Accept", ylab = "F.Undergrad", main = "Accept versus
F.Undergrad", col = c("blue","red"))
```




```
colnames(colleges)
```

```
##  [1] "X"           "Private"     "Apps"        "Accept"      "Enroll"
##  [6] "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"
## [11] "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"
## [16] "S.F.Ratio"   "perc.alumni" "Expend"      "Grad.Rate"   "Top"
## [21] "acceptance"
```

Question 2

Make sure that rows with missing values have been removed from the data. Show both the code you used and any relevant outputs.

```
forest_fire = read.csv("forestfires.csv")
forest_fire = na.omit(forest_fire)
summary(forest_fire)
```

```
##     month              day                FFMC            DMC
##  Length:517         Length:517         Min.   :18.70   Min.   :  1.1
##  Class :character   Class :character   1st Qu.:90.20   1st Qu.: 68.6
##  Mode  :character   Mode  :character   Median :91.60   Median :108.3
##                                        Mean   :90.64   Mean   :110.9
##                                        3rd Qu.:92.90   3rd Qu.:142.4
##                                        Max.   :96.20   Max.   :291.3
```

```
##        DC             ISI              temp             RH
##  Min.   :  7.9   Min.   : 0.000   Min.   : 2.20   Min.   : 15.00
##  1st Qu.:437.7   1st Qu.: 6.500   1st Qu.:15.50   1st Qu.: 33.00
##  Median :664.2   Median : 8.400   Median :19.30   Median : 42.00
##  Mean   :547.9   Mean   : 9.022   Mean   :18.89   Mean   : 44.29
##  3rd Qu.:713.9   3rd Qu.:10.800   3rd Qu.:22.80   3rd Qu.: 53.00
##  Max.   :860.6   Max.   :56.100   Max.   :33.30   Max.   :100.00
##       wind            rain              area
##  Min.   :0.400   Min.   :0.00000   Min.   :   0.00
##  1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:   0.00
##  Median :4.000   Median :0.00000   Median :   0.52
##  Mean   :4.018   Mean   :0.02166   Mean   :  12.85
##  3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:   6.57
##  Max.   :9.400   Max.   :6.40000   Max.   :1090.84
```

(a) Specify which of the predictors are quantitative (measuring numeric properties such as size, or quantity), and which are qualitative(measuring non-numeric properties such as color, appearance, type etc.), if any? Keep in mind that a qualitative variable may be representedas a quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings.

```
sapply(forest_fire, class)
```

```
##       month         day        FFMC         DMC          DC         ISI
## "character" "character"   "numeric"   "numeric"   "numeric"   "numeric"
##        temp          RH        wind        rain        area
##   "numeric"   "integer"   "numeric"   "numeric"   "numeric"
```

Qualitative or Quantitative variable:

1. month -> Qualitative (consist of name of the month and is treated as Character)
2. day -> Qualitative (consist of name of the day and is treated as Character)
3. FFMC -> Quantitative (consist data of numeric value)
4. DMC -> Quantitative (consist data of numeric value)
5. DC -> Quantitative (consist data of numeric value)
6. ISI -> Quantitative (consist data of numeric value)
7. temp -> Quantitative (consist data of numeric value)
8. RH -> Quantitative (consist data of integer value)
9. wind -> Quantitative (consist data of numeric value)
10. rain -> Quantitative (consist data of numeric value)
11. area -> Quantitative (consist data of numeric value)

(b) What is the range, mean and standard deviation of each quantitative predictor?

```
#Range for all Quantitative variables
sapply(forest_fire[,c(3:11)], range)
```

```
##      FFMC   DMC    DC  ISI temp  RH wind rain    area
## [1,] 18.7   1.1   7.9  0.0  2.2  15  0.4  0.0    0.00
## [2,] 96.2 291.3 860.6 56.1 33.3 100  9.4  6.4 1090.84
```

```
#Mean for all Quantitative variables
sapply(forest_fire[,c(3:11)], mean)

##         FFMC          DMC           DC          ISI         temp
RH
##  90.64468085 110.87234043 547.94003868   9.02166344  18.88916828
44.28820116
##         wind         rain         area
##   4.01760155   0.02166344  12.84729207

#Standard deviation for all Quantitative variables
sapply(forest_fire[,c(3:11)], sd)

##        FFMC         DMC          DC         ISI        temp          RH
##   5.5201108  64.0464822 248.0661917   4.5594772   5.8066253  16.3174692
##        wind        rain        area
##   1.7916526   0.2959591  63.6558185
```

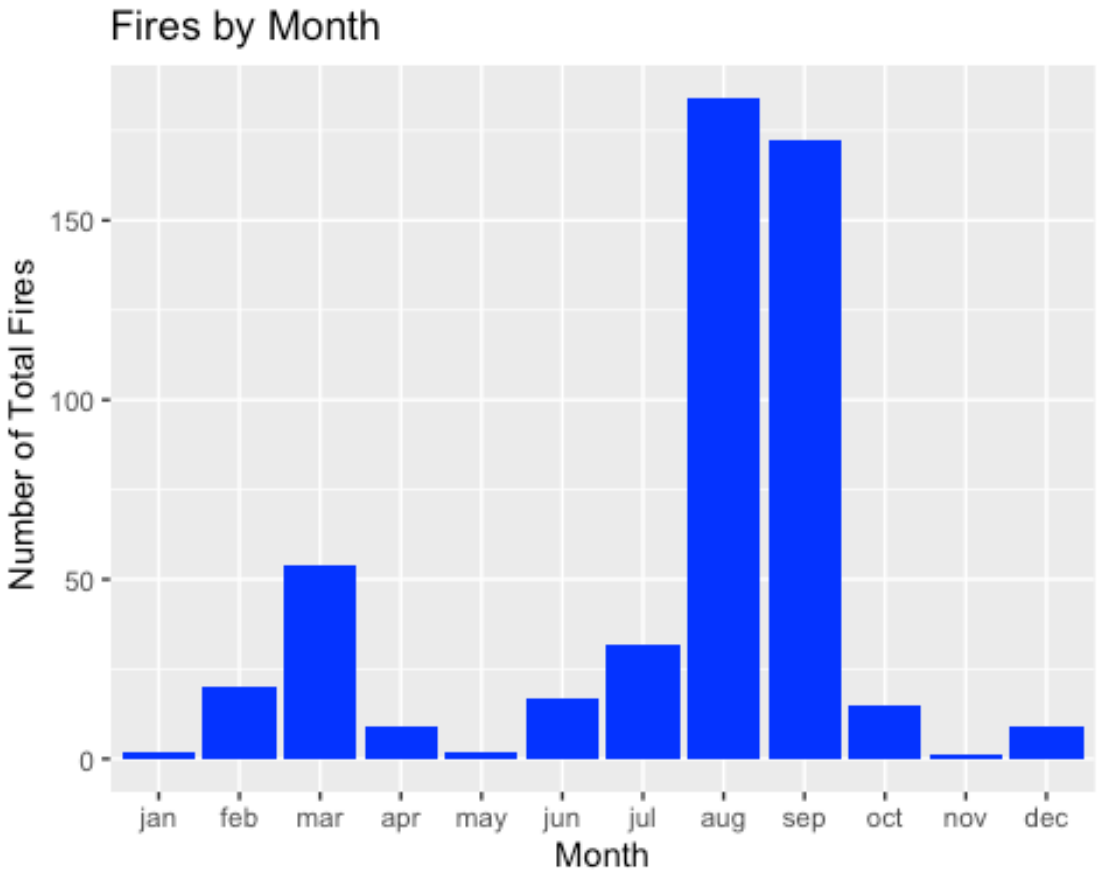
(c) Now remove the 20th through 70th (inclusive) observations from the dataset. What is the range, mean, and standard deviationof each predictor in the subset of the data that remains?

```
#subset of forest_fire from 20:70
sub.forest_fire = subset(forest_fire[-c(20:70),])

sapply(sub.forest_fire[,c(3:11)], range)

##      FFMC   DMC    DC  ISI temp  RH wind rain    area
## [1,] 18.7   1.1   7.9  0.0  2.2  15  0.4  0.0    0.00
## [2,] 96.2 291.3 860.6 22.7 33.3 100  9.4  6.4 1090.84

sapply(sub.forest_fire[,c(3:11)], mean)

##         FFMC          DMC           DC          ISI         temp
RH
##  90.62188841 113.52167382 548.04012876   8.98927039  18.94163090
44.59442060
##         wind         rain         area
##   4.01244635   0.02403433  14.25332618

sapply(sub.forest_fire[,c(3:11)], sd)

##        FFMC         DMC          DC         ISI        temp          RH
##   5.7429895  65.7845884 249.1977150   4.1109312   5.9027226  16.5912495
##        wind        rain        area
##   1.8179084   0.3116754  66.9058989
```

d) Produce a bar plot to show the count of forest fires in each month. During which months are forest fires most common? (Hint: group data by month and calculate count)

Answer: The most common months for forest fires are August and September

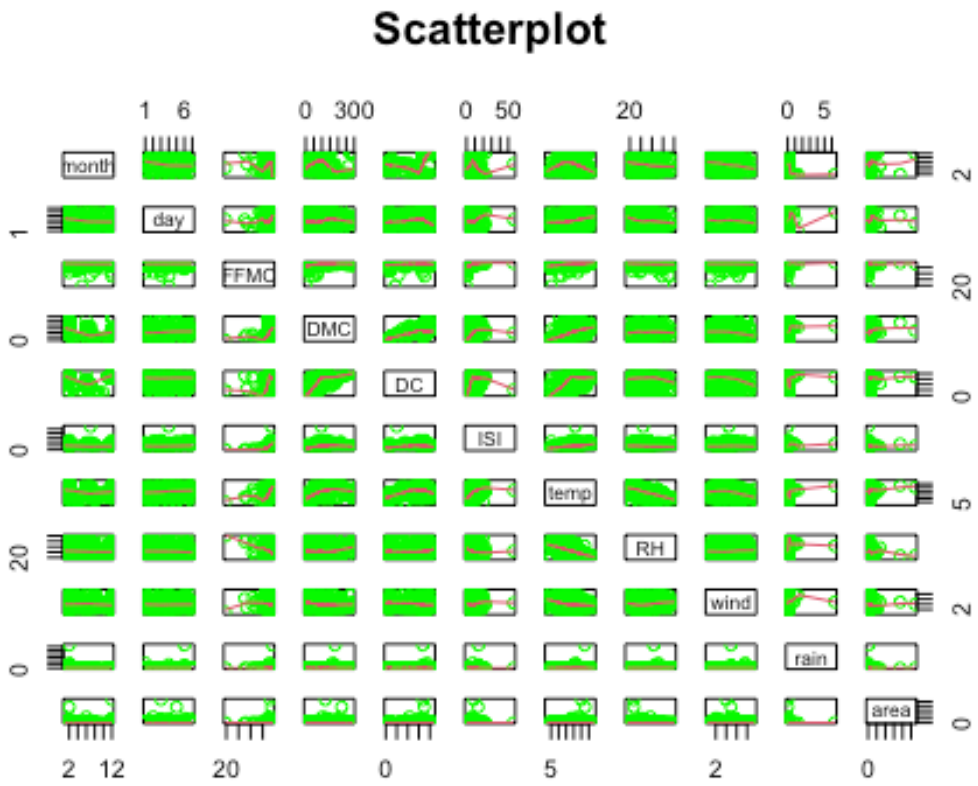
```
library(dplyr)

##
## 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

forestfiresmonths <- forest_fire %>%
  mutate(month = factor(month, levels = c("jan", "feb", "mar", "apr", "may",
"jun", "jul", "aug", "sep", "oct", "nov", "dec")))

fires_by_month<- forestfiresmonths %>% group_by(month) %>%
summarize(count_fires
                                                                    = n())
ggplot(fires_by_month) + aes(x = month, y = count_fires) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Fires by Month", x = "Month", y=" Number of Total Fires")
```



```
table(forest_fire$month)

##
## apr aug dec feb jan jul jun mar may nov oct sep
##   9 184   9  20   2  32  17  54   2   1  15 172
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or othertools of your choice. Create a correlation matrix for the relevant variables.

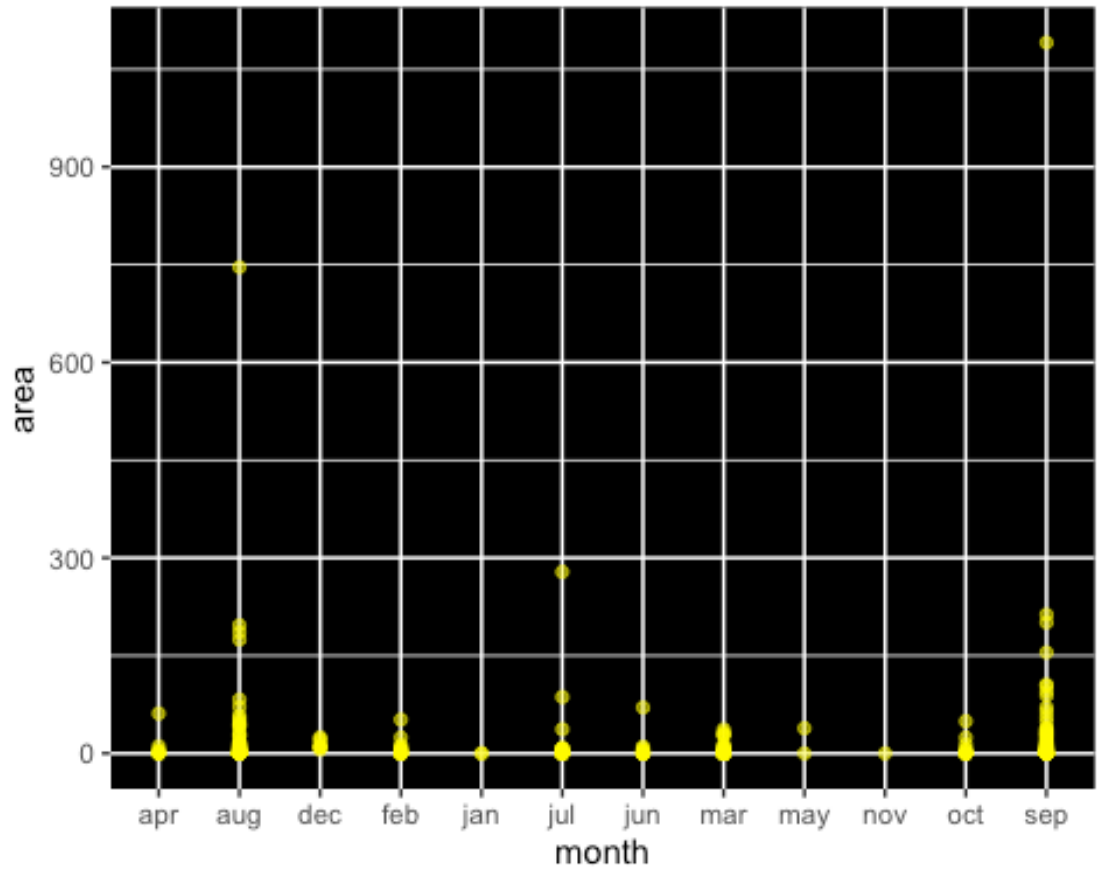
```
val = forest_fire[, !sapply(forest_fire, is.factor)]
 plot(val, panel = panel.smooth, main = "Scatterplot", col = "green")
```



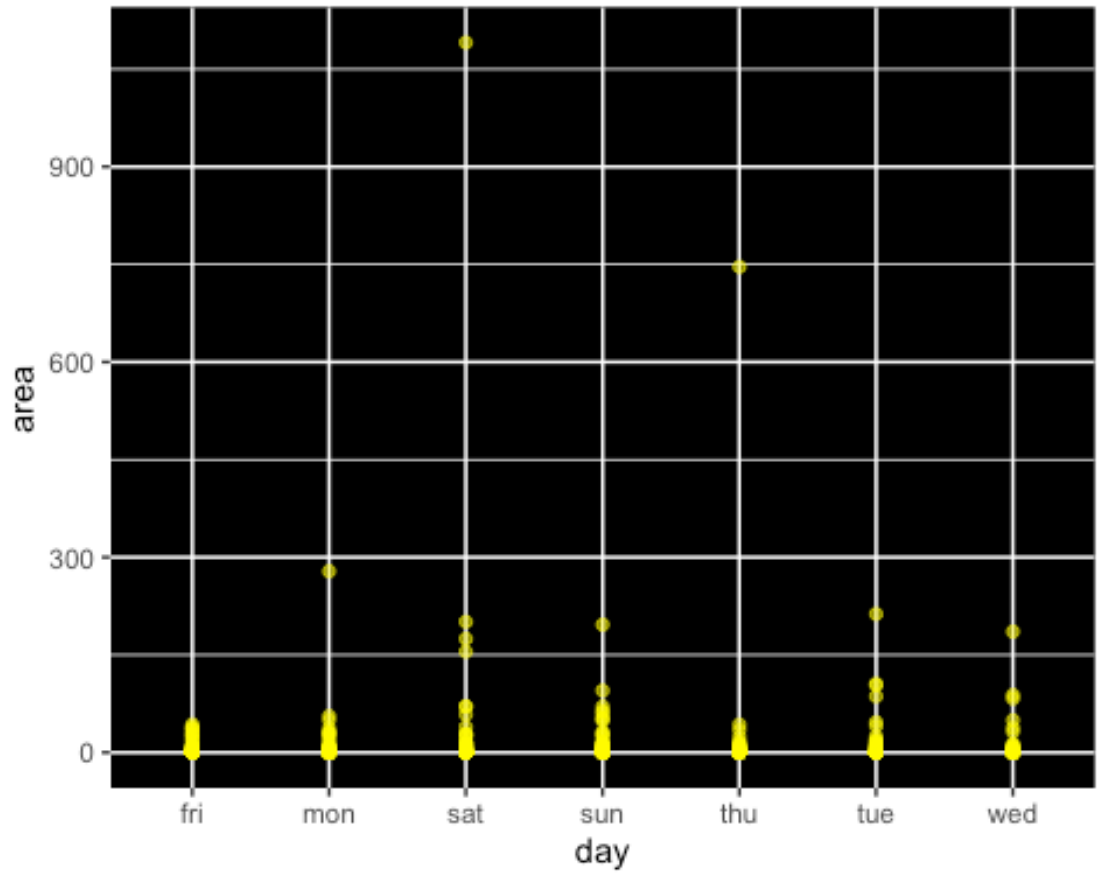
(f) Suppose that we wish to predict the area burned by the forest fire(area) on the basis of the other variables. Which, if any,of the other variables might be useful in predicting area? Justify your answer based on the prior correlations.

```
library(ggplot2)
library(purrr)
fire_area_scatter = function(x,y) {
  ggplot(data = forest_fire) +
    aes_string(x = x, y = y) +
    geom_point(alpha = 0.5, col="yellow") +
    theme(panel.background = element_rect(fill="black"))}
```

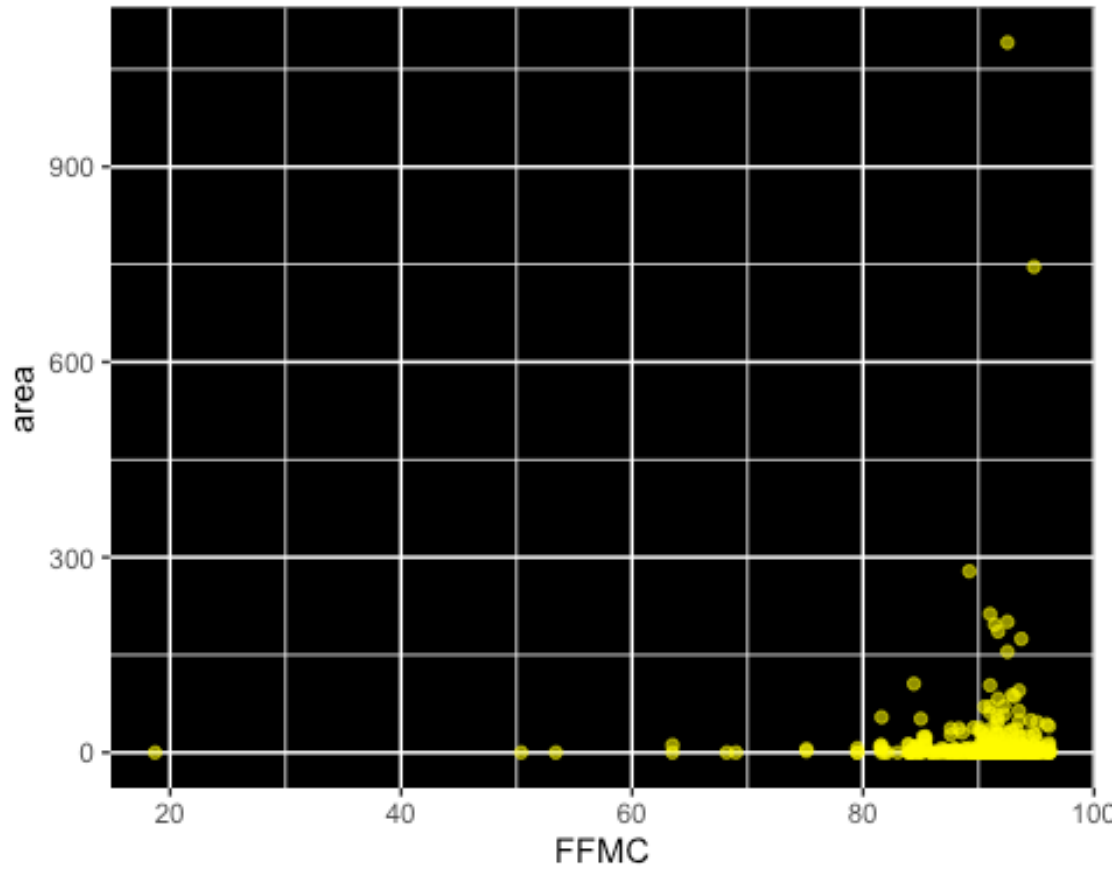
```
xvar <- names(forest_fire)[1:10]
yvar <- names(forest_fire)[11]
map2(xvar, yvar, fire_area_scatter)

## [[1]]
```



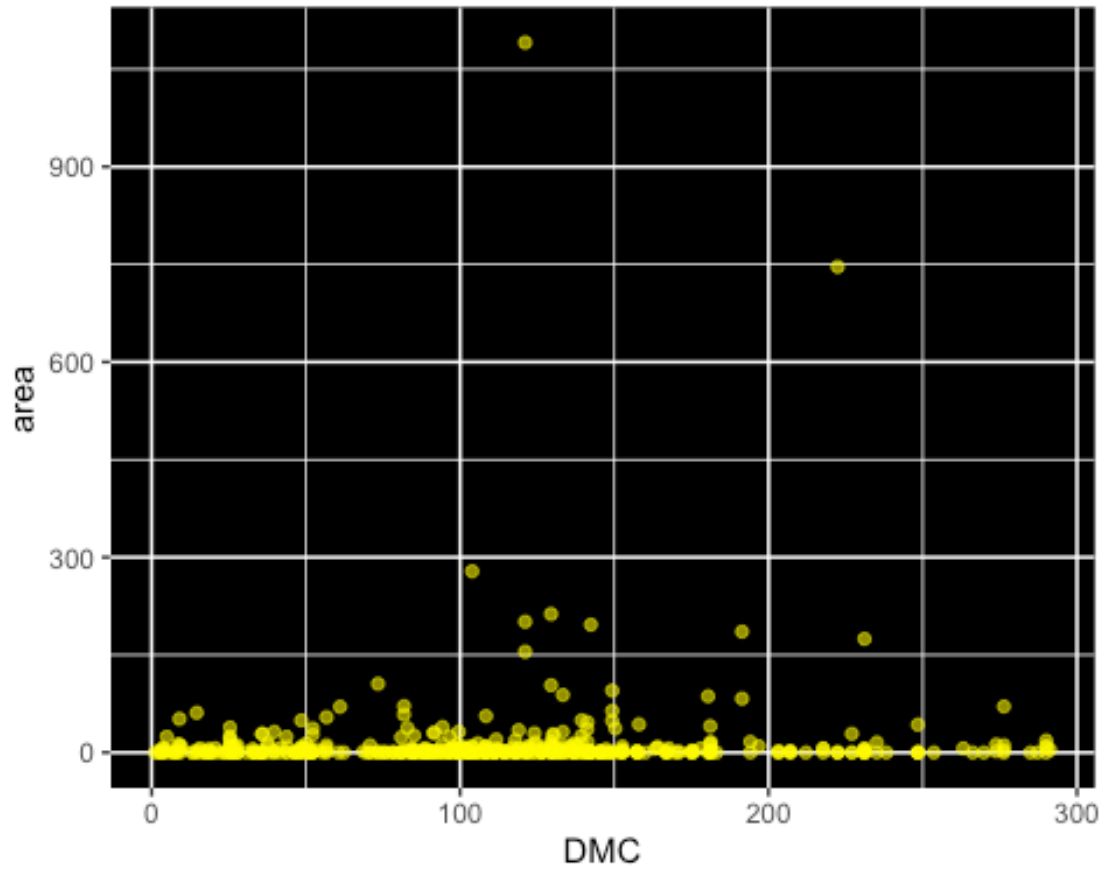

```
##
## [[2]]
```



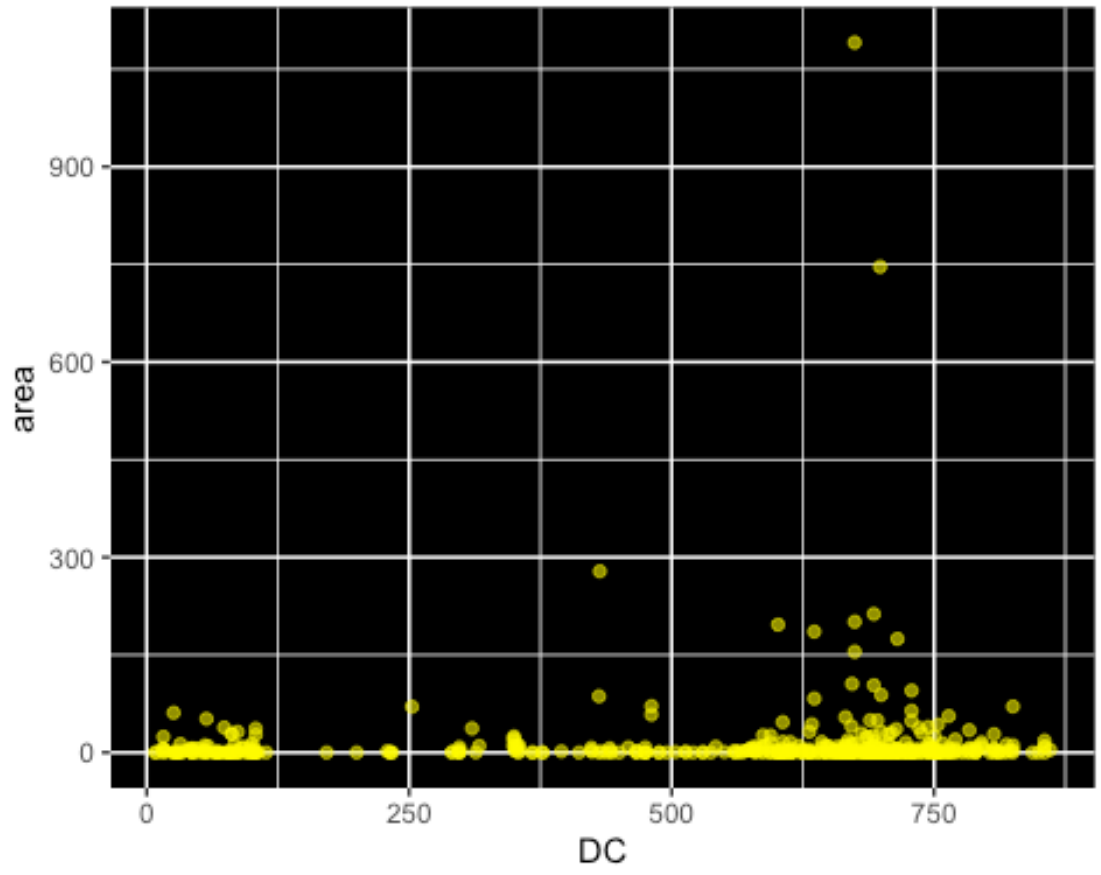

```
##
## [[3]]
```



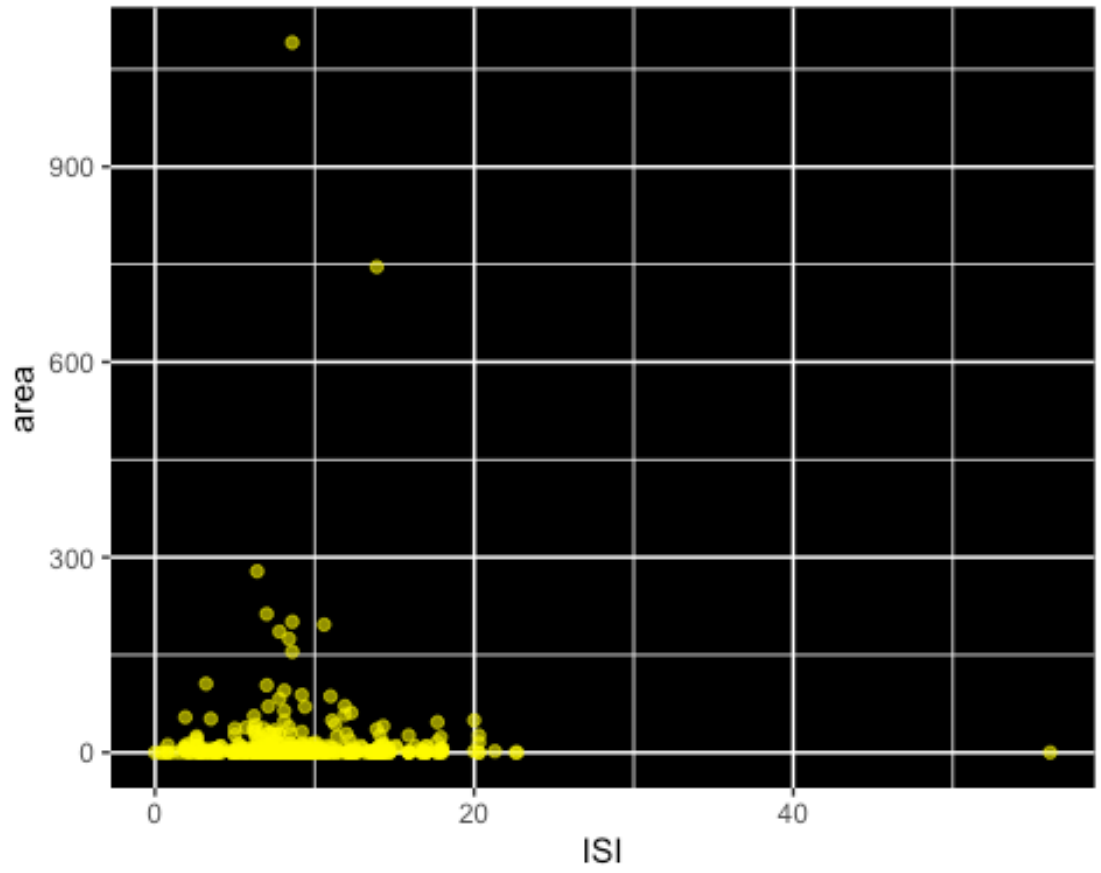

```
##
## [[4]]
```



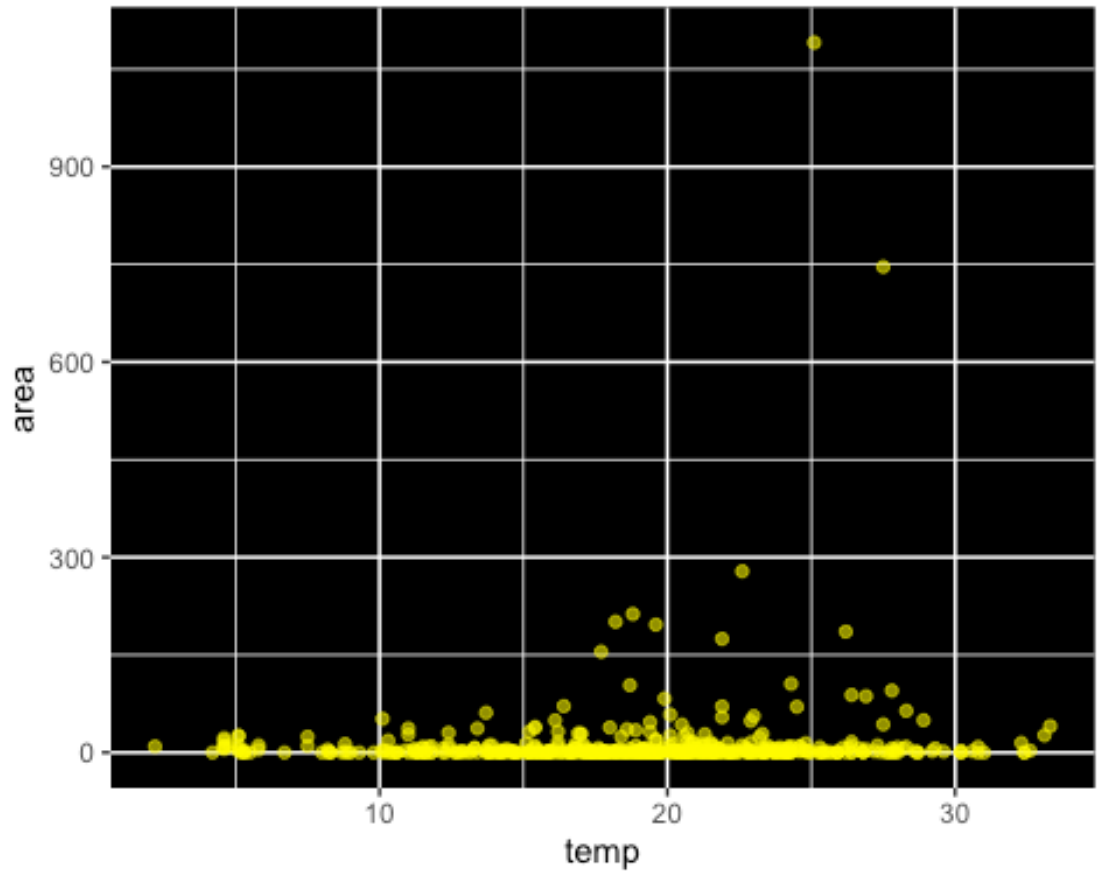

```
##
## [[5]]
```



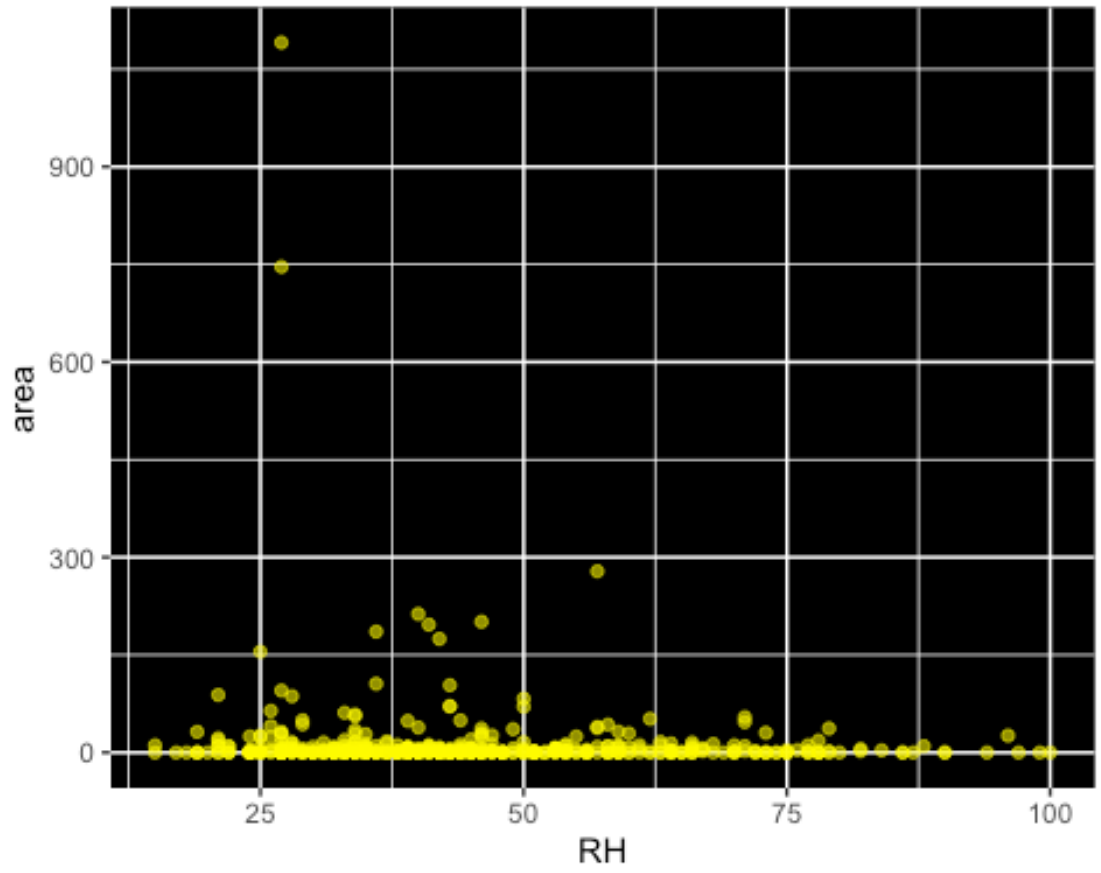

```
##
## [[6]]
```



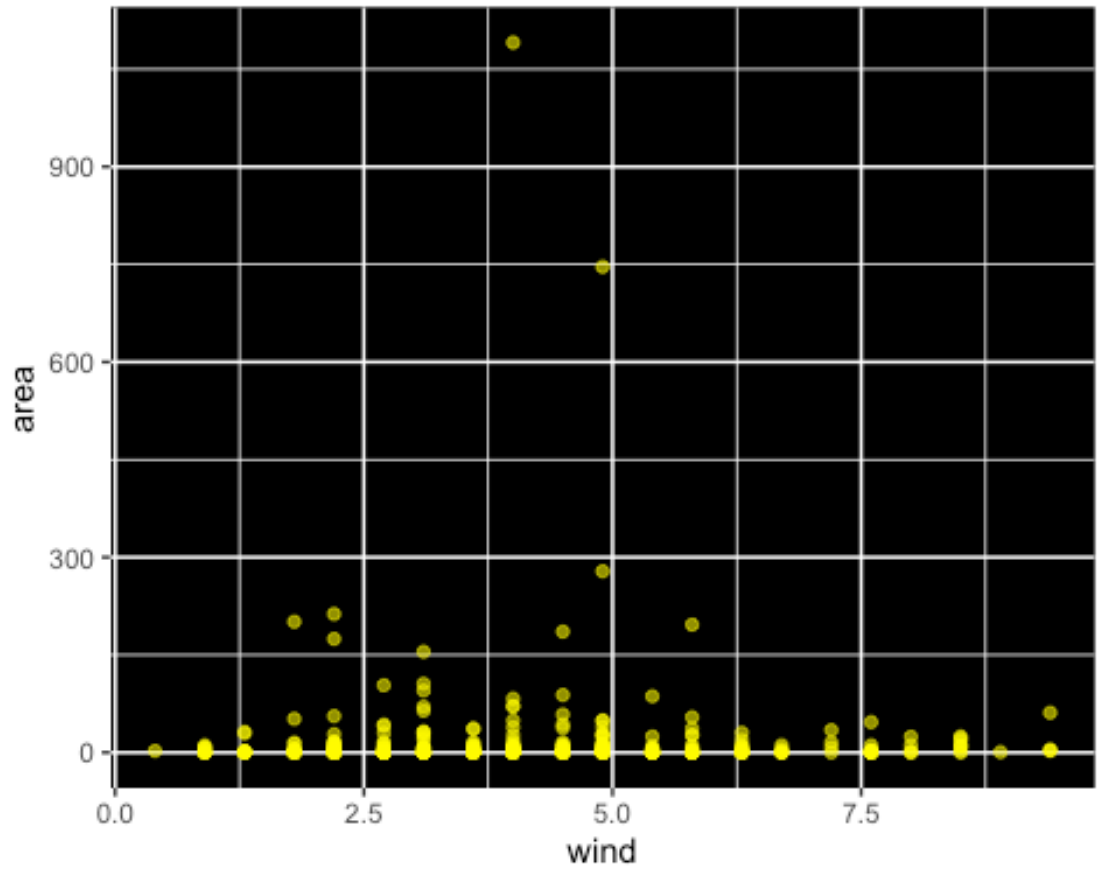

```
##
## [[7]]
```



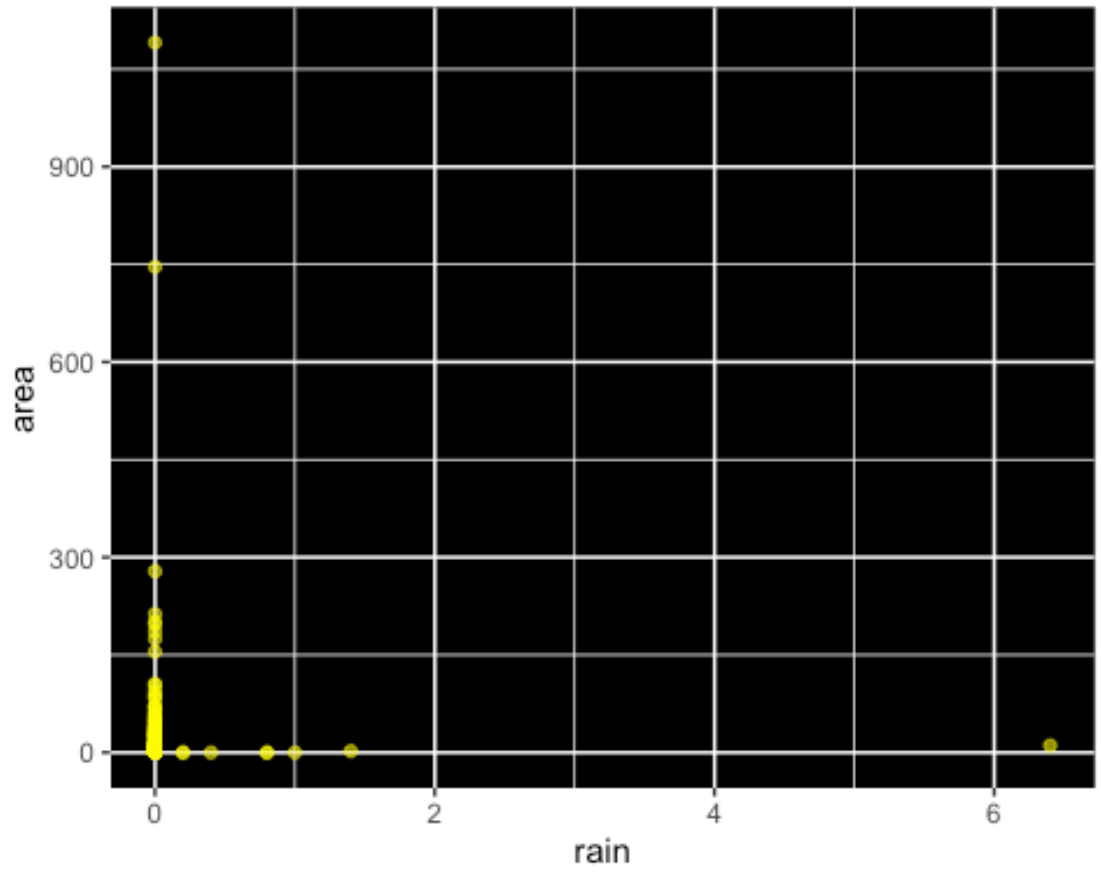

```
##
## [[8]]
```




```
##
## [[9]]
```




```
##
## [[10]]
```



I was hoping to find some correlation betweeen any variable and area but points representing area are either zero or close to zero. This tells me that there is no concrete relationship between area burned and any of the other variable.