Stats_Lab1

Danielle Salah, Peter Trenkwalder

May 16, 2018

```
library(car)
```

```
## Loading required package: carData
```

Introduction

Our analysis serves to explore this design question: "what factors lead to particularly damaging forest fires?

By performing exploratory analysis on this data we aim to discover commonalities in areas with sizable fire damange. The insights discovered will help inform a ditection system that provides early warnings to these regions. In order to identify these characteristics, we followed the following steps:

- examine the data and make any necessary adjustments to improve its quality and usability
- perform initial analysis on the included variables to identify key indicators
- use multivariate analysis on the material variables to converge upon some commonalities

Add: data source, restrictions

The Data

Upon importing the data, we can see that we have 517 observations and 13 variables. These variables are mostly of the numerical type, with a few integers and two factors included as well. All of these variable types appear to be appropriate given the data.

```
nrow(df)
```

```
## [1] 517
```

```
str(df)
```

```
## '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 ...
```

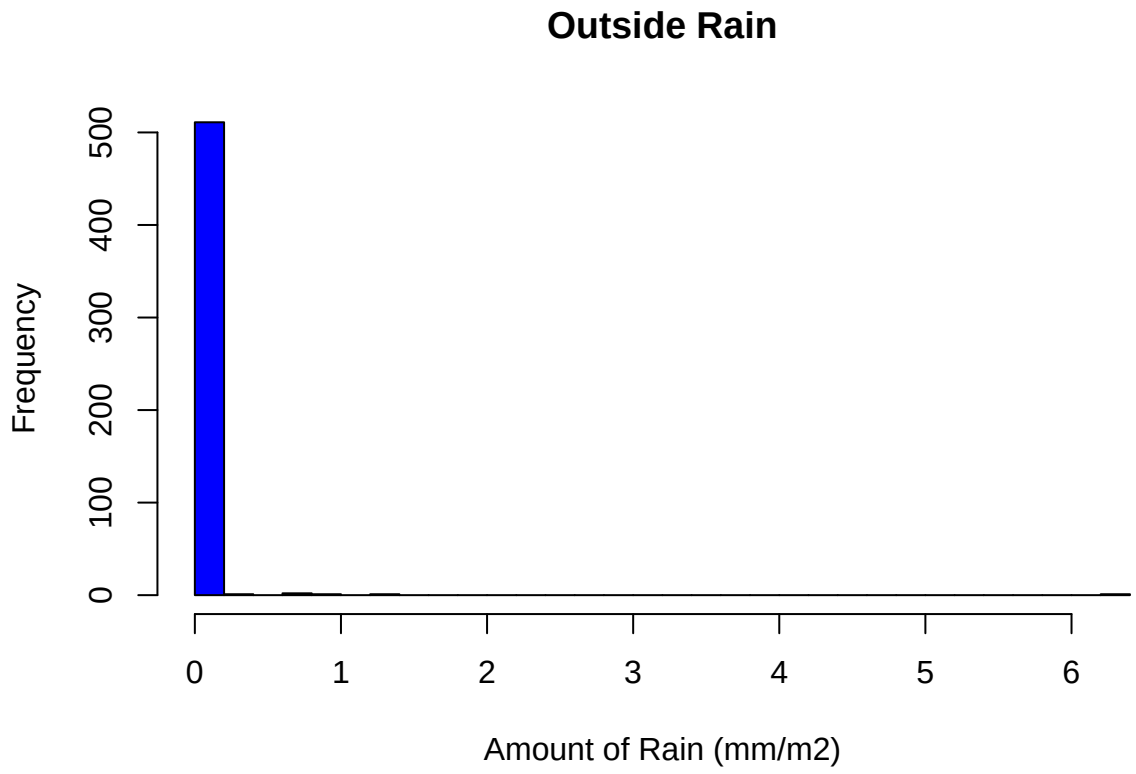
Add: table of definitions

The Cleanup

Initially in the process, we can see high level details on the 13 variables. None of them appear to have unreadable or otherwise unusable values, there are no negative minimums, and all of the index variables have minimums and maximums within the appropriate ranges for each specific index.

One cariable of interest based on these statistical summaries is rain. When we look at the histogram of this variable we can start to uncover what is happening.

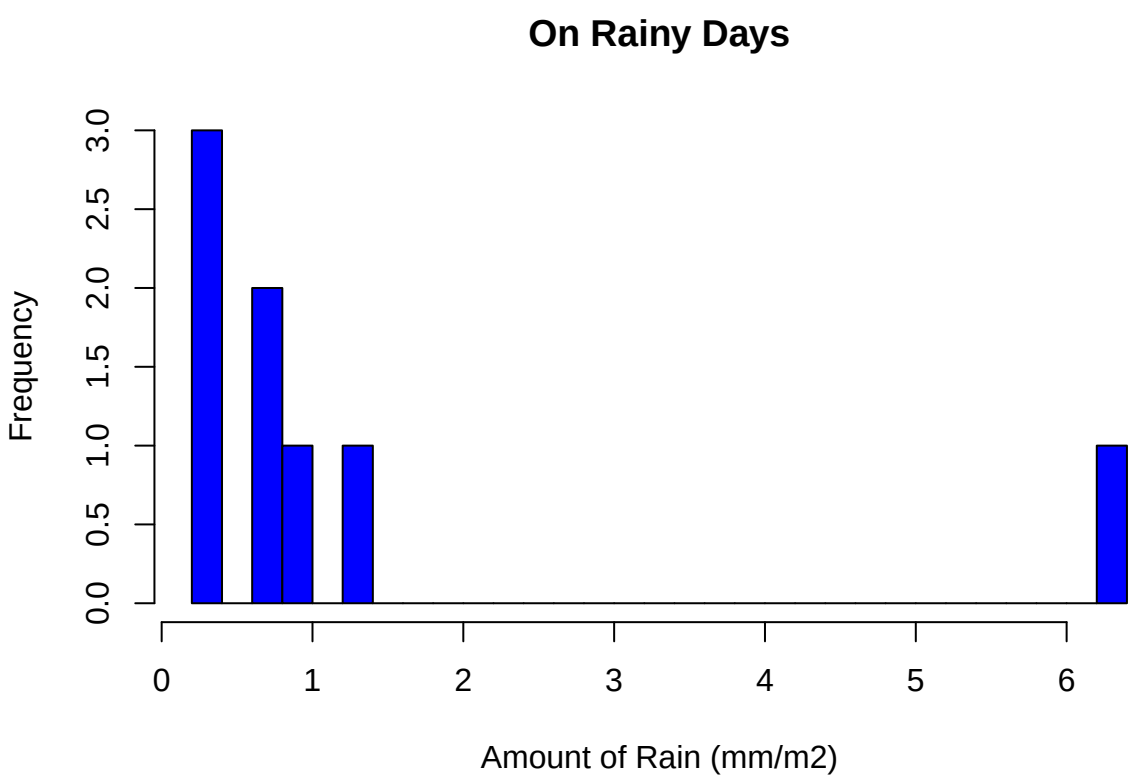
```
hist(df$rain, breaks = 36, col = "blue",
     xlab = "Amount of Rain (mm/m2)",
     main = "Outside Rain")
```



We immediately notice that there is a signficiant number of observations with no rain at all. While this detail may be important to the analysis in the future, it initially entirely obscures our observations of what the data looks like. If we graph a histogram of the data removing any 0 values, we can

```
rainNonZero <- subset(df$rain, df$rain > 0.00)

hist(rainNonZero, breaks = 36, col = "blue",
     xlab = "Amount of Rain (mm/m2)",
     main = "On Rainy Days")
```



From here, a few things become clear about the variable:

- There are only 8 observations on days with any rain.
- All but one of these observations falls below 1.5 mm/m2.
- This single observation is actually quite an outlier at 6.4 mm/m2.

```
rainyDay <- subset(df, df$rain > 6)
rainyDay
```

```
##     X Y month day FFMC   DMC    DC  ISI temp RH wind rain  area
## 500 7 5   aug tue 96.1 181.1 671.2 14.3 27.3 63  4.9  6.4 10.82
```

When we isolate the data for this observation, we can see that it occurred in aug which is a reasonable time of the year for a very rainy day. Interestingly this observation also includes 10.82 hectares of burned forest.

The Time

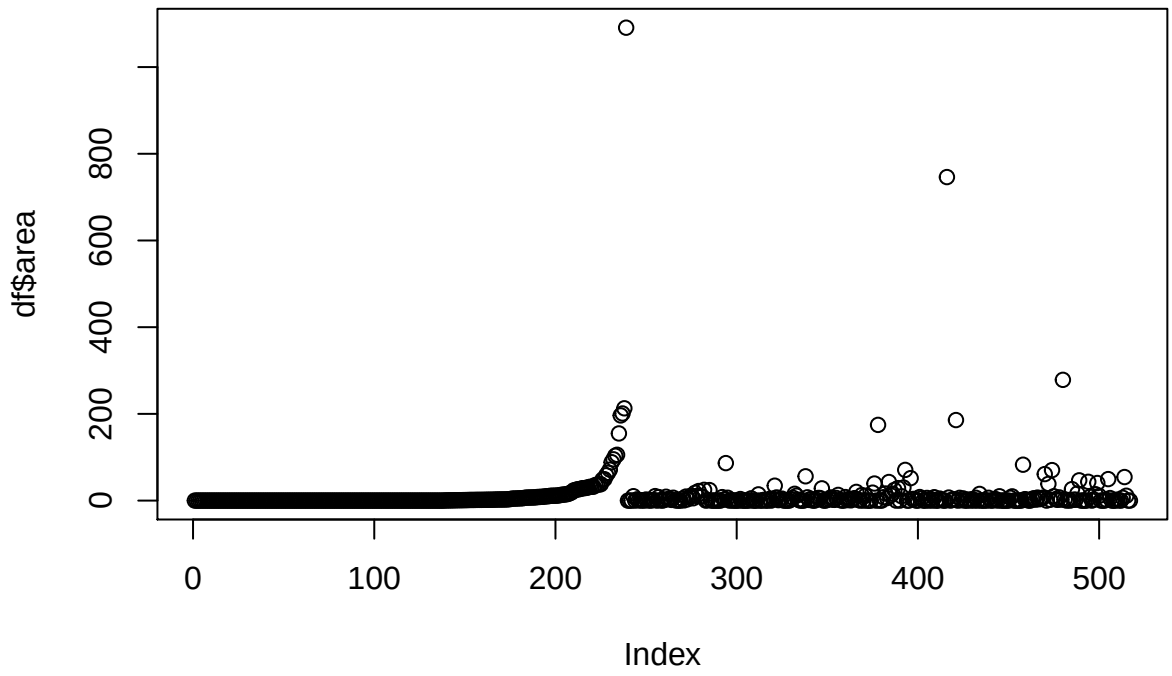
To aid in later analysis, we decided to subset the data according to season under the hypothesis that the factors contributing to fire risk differed at different times of the year.

```
spring <- subset(df, df$month == "mar" | df$month == "apr" | df$month == "may")
summer <- subset(df, df$month == "jun" | df$month == "jul" | df$month == "aug")
fall <- subset(df, df$month == "sep" | df$month == "oct" | df$month == "nov")
winter <- subset(df, df$month == "dec" | df$month == "jan" | df$month == "feb")
```

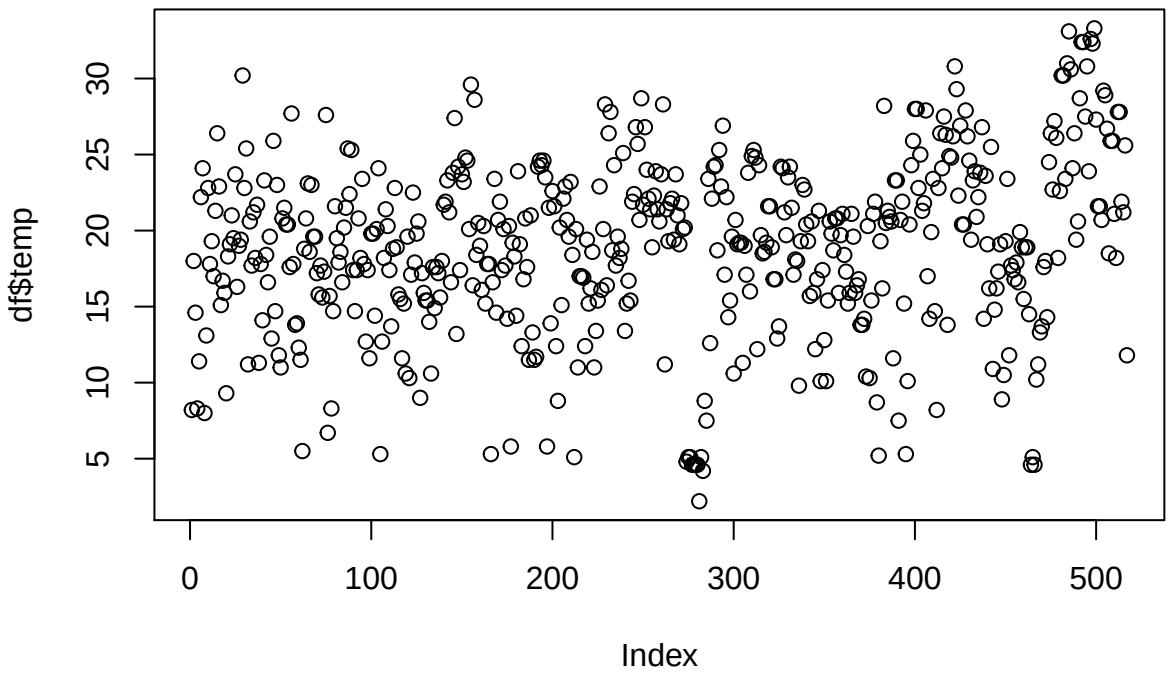
This gives up four subsets of the data frame: 65 in spring, 233 in summer, 188 in fall, and 31 in winter.

Target Variables

```
plot(df$area)
```



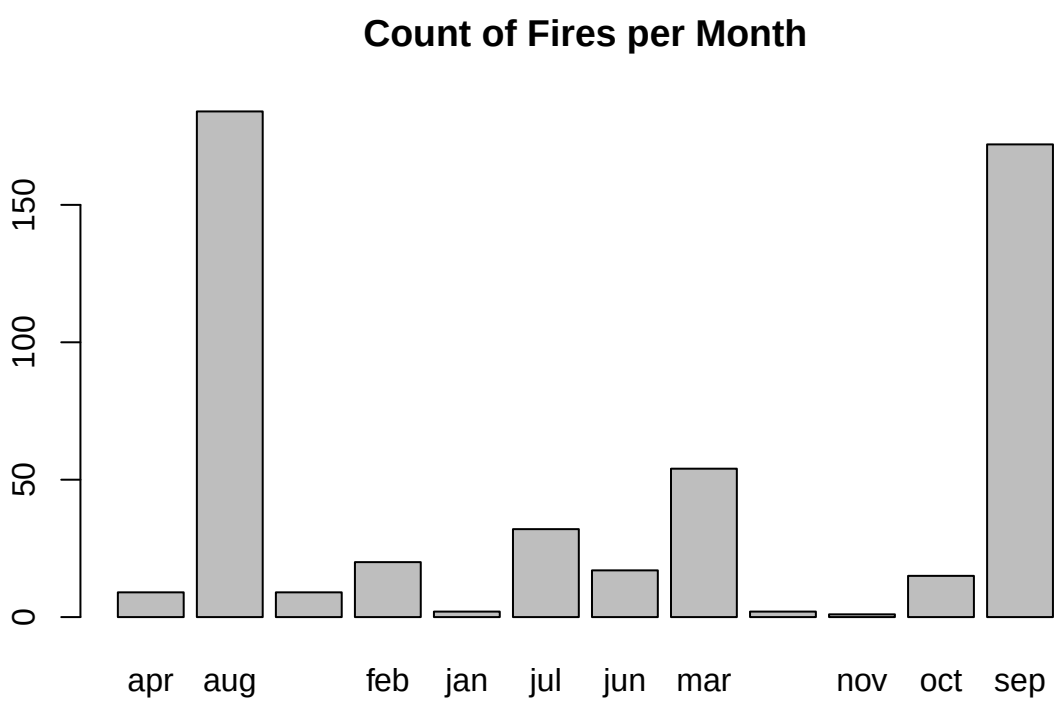

```
plot(df$temp)
```



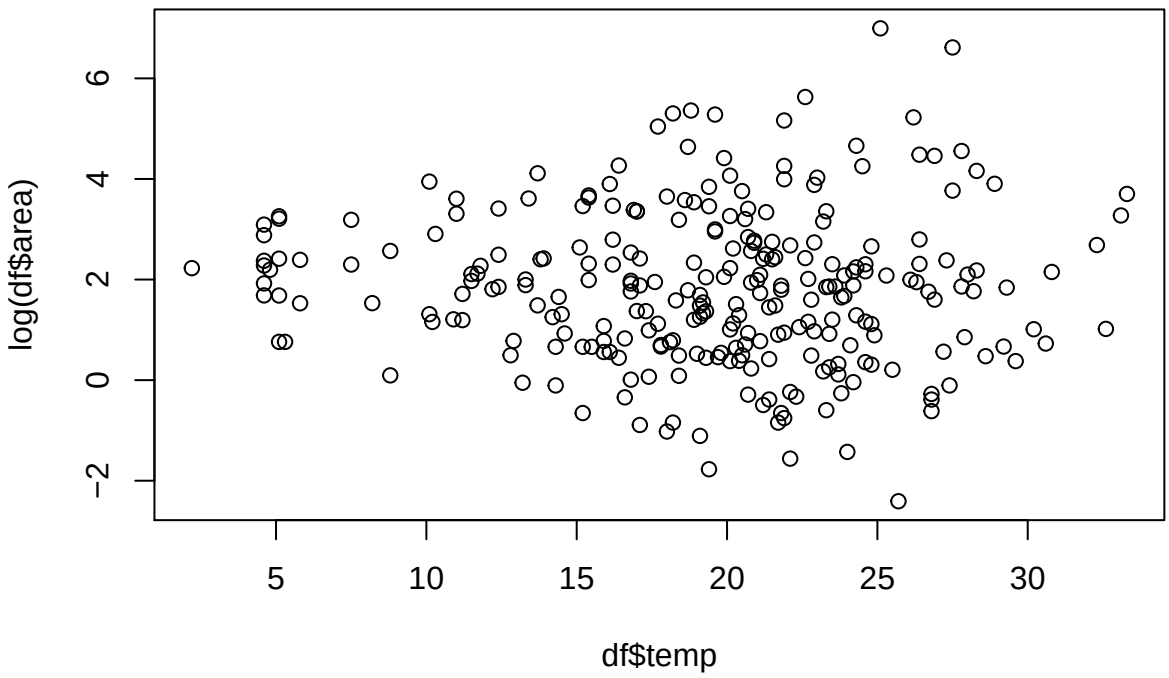
Months

** shift order ** break into seasons

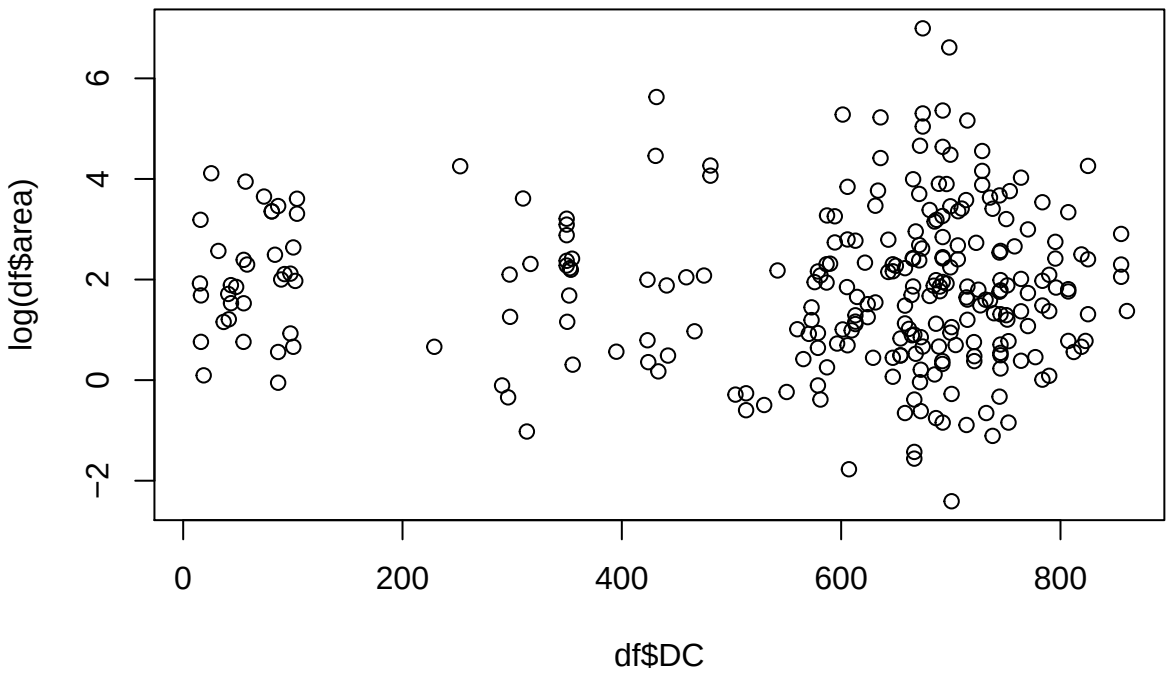
```
plot(df$month,
     main = "Count of Fires per Month")
```



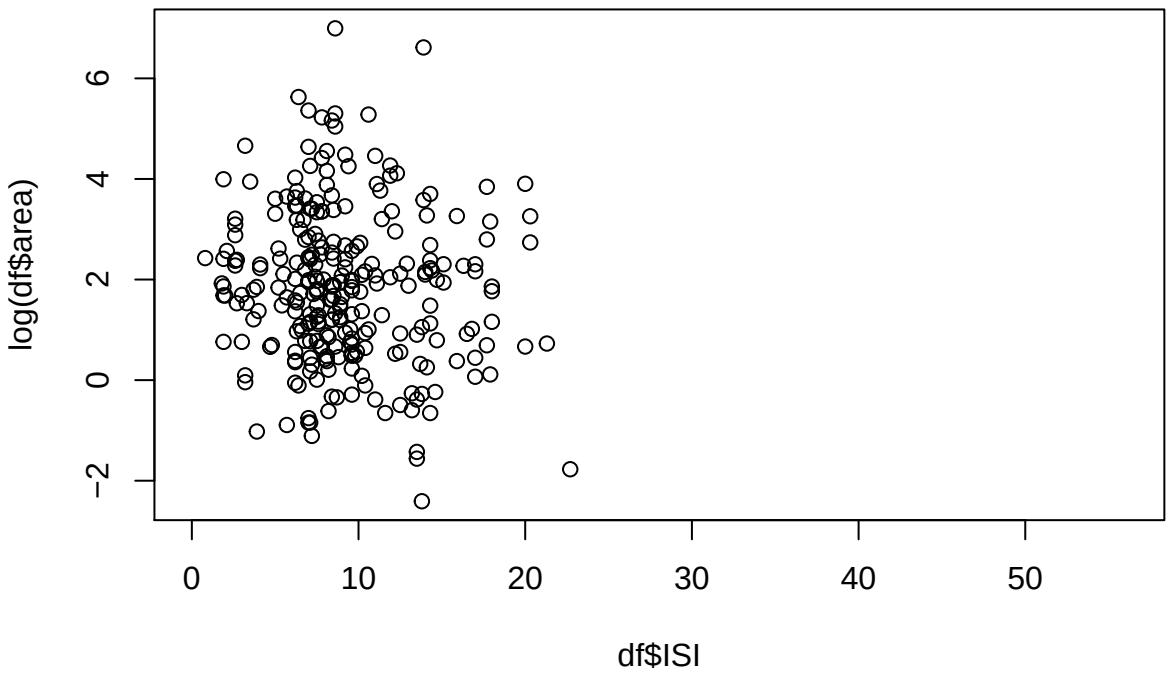

```
plot(df$temp, log(df$area))
```




```
plot(df$DC, log(df$area))
```




```
plot(df$ISI, log(df$area))
```




```
cor(df$DC, df$area)
```

```
## [1] 0.04938323
```

```
cor(df$ISI, df$area)
```

```
## [1] 0.008257688
```