Lecture7

Sabrina

9 April 2019

## Quiz - Time

D\_Q <- Depredations

### a) What would be the very first step if someones gives you a dataset?

Answer: Learn more about what each column means. For that use the command:

??Depredations

and to show the first couple of entries use:

head(D\_Q)

## longitude latitude number early late  
## 1 -94.5 46.1 1 0 1  
## 2 -93.0 46.6 2 0 2  
## 3 -94.6 48.5 1 1 0  
## 4 -92.9 46.6 2 0 2  
## 5 -95.9 48.8 1 0 1  
## 6 -92.7 47.1 1 0 1

### b) How do you determine the number of observations?

nrow(D\_Q)

## [1] 434

### c) Which of the variables are continuous which ones are factors?

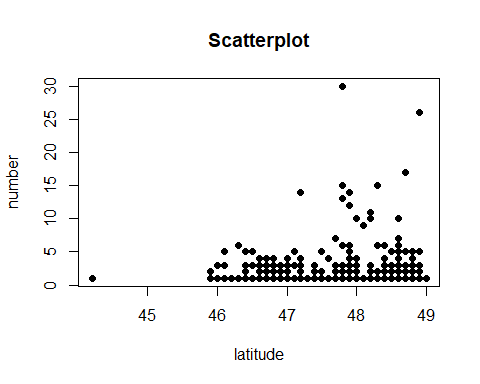
longitude, latitude are continuous variables

number, early, late are factors (ordinal factors)

### d) If you want to investigate the distribution of the latitude with respect to number of depredations, what type of plot (and what R-Code) would you use?

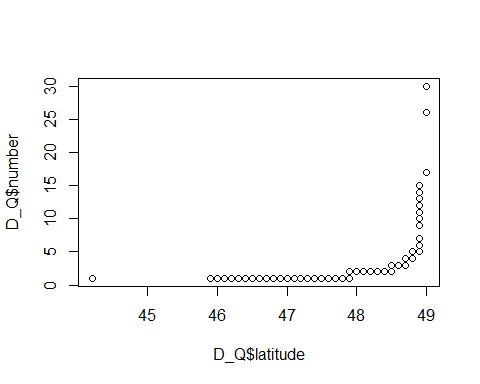
One could for example use a scatter plot:

plot(D\_Q$latitude, D\_Q$number, xlab = 'latitude', ylab = 'number', main = 'Scatterplot', pch=16)



One could also ask, whether the random variable: latitude and the random variable: number are from the same distribution. For that, we use a qqplot:

qqplot(D\_Q$latitude, D\_Q$number)



they don’t seem to be from the same distribution!

### e) What variables do you suspect to be related and how would you test this?

Knowing the data, we could suspect that Number, Early, and Late might be correlated, since they all represent the number of depredations but for different year-classes. To check that intuition, we can create a correlation matrix:

D\_Q1 <- D\_Q %>% select(number, early, late)  
round(cor(D\_Q1), digits =3)

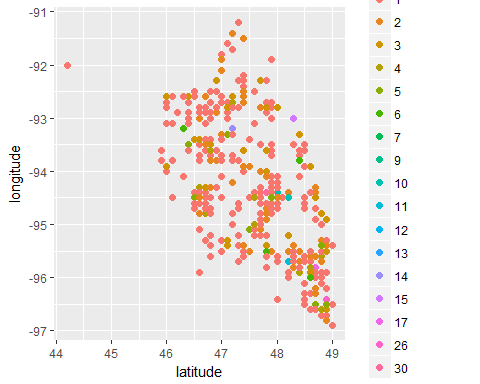
## number early late  
## number 1.000 0.855 0.768  
## early 0.855 1.000 0.324  
## late 0.768 0.324 1.000

As to be expected, there is a high correlation between the number and early and late. However, early and late are not highly correlated.

### f) Can you think of some other questions you would like to answer with that data set?

We could for example create a 2-D plot to see if there are specific regions (longitude and latitude) where most of the depredations occured. That is:

ggplot(D\_Q, aes(x=latitude, y=longitude, col=factor(number)))+geom\_point(size=2)



We see that specific regions are targeted in this data set. However, we can not see that specific high values of depredations (variable = number) appear in one specific region.