MODULE 5 R PRACTICE

Week 5

Abstract

In this assignment I will use regression modeling for some data

Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

The databse I choose to use is from Kaggle (MsSmartyPants, 2021) it is about the quality of water and particles in water samples

First Of all, I import and install required libraries

```
 print('Mohammad Hossein Movahedi')

#install and load packages

install.packages('FSA')

install.packages('FSAdata')

install.packages('magrittr')

install.packages('dplyr')

install.packages('tidyr')

install.packages('plyr')

install.packages('tidyverse')

install.packages('outliers')

install.packages('ggplot2')

install.packages('lubridate')

install.packages('corrplot')



library(ggplot2)

library(outliers)

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(tidyr)

library(dplyr)

library(tidyverse)

library(scales)

library(lubridate)

library(corrplot)
```

then I import the dataset and delete non-numeric columns.

```
 #loading data

data <- read.csv("waterQuality1.csv")



datan <- (data[, unlist(lapply(data, is.numeric))])
```

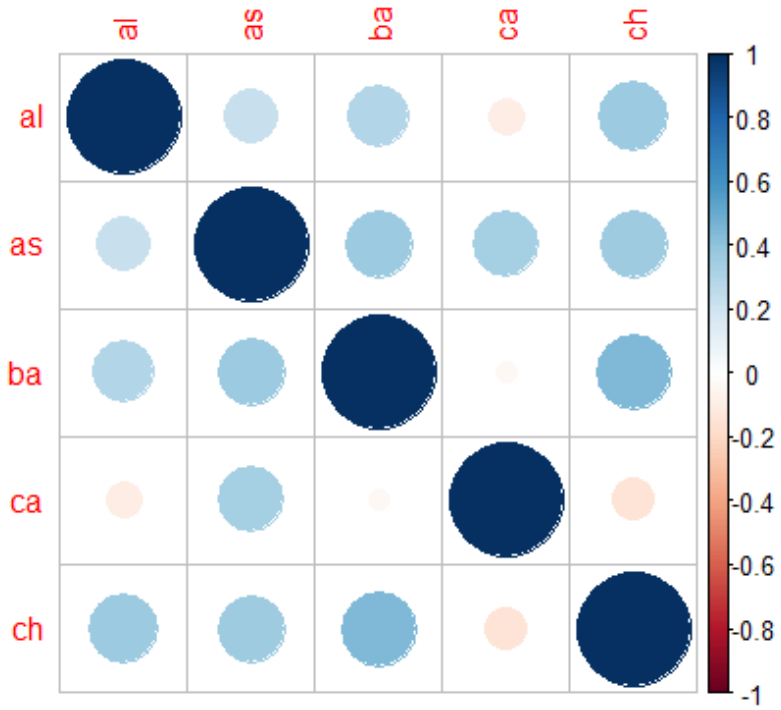
Then Because the dataset is vast I choose only first 5 columns for the regression test.

```
 #subseting the first 5 coloumns for analyze

datan <- data.frame(al = datan$aluminium,as = datan$arsenic,ba =datan$barium,ca =
datan$cadmium,ch = datan$chloramine)
```

now I create a correlation matrix and create a corrplot to understand the correlation between the variables . the reason why I only use 5 columns for this part is adding each columns multiples the number correlations

```
 m = cor(datan)

corrplot(m)
```

the resulting plot looks like this

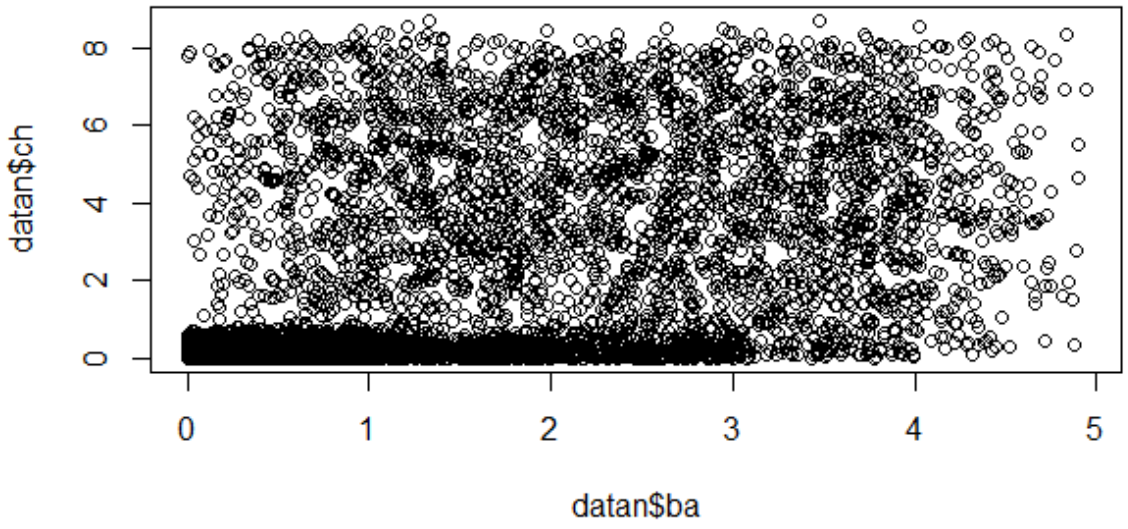


As It can be seen, nearly all variables have a positive correlation; the reason is that when a water sample is from a polluted place, It will get all kinds of pollution, and when it is a clean sample, it doesn't have any pollution however there is a weak negative correlation between chloramine and cadmium maybe the reason is that these particles react to each other.

Based on the Corrplot, there is a strong relation between chloramine and barium; at first, we create the plot to see if there is a visible relation or not.

```
 SELECT *

FROM interview

WHERE person_id = "14887" OR person_id = "16371"
```

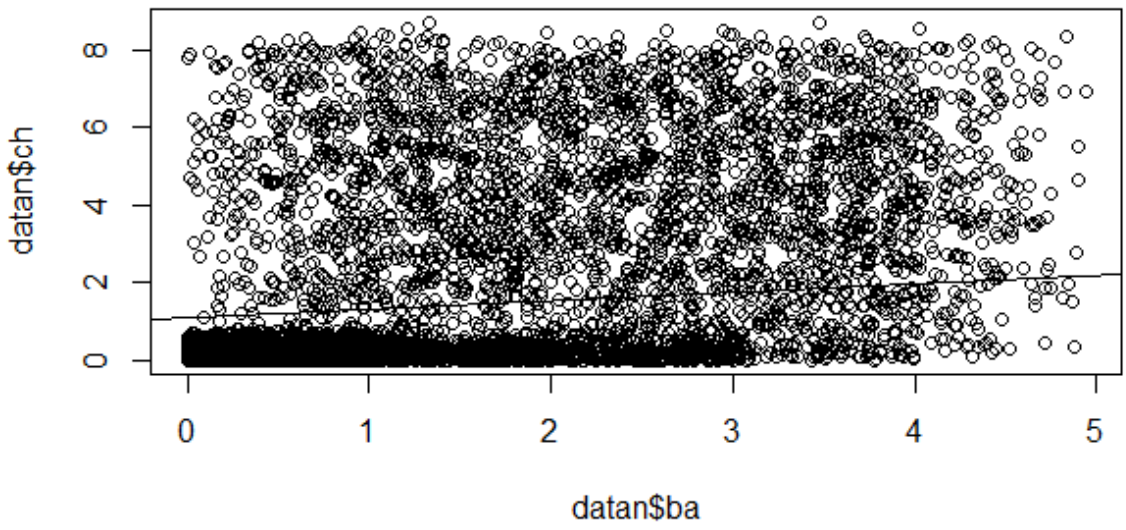
the resulting plot looks like this



There is no visible relation in this chart, so our regression will probably fail

```
 reg <- lm(ba~ch,data = datan)

summary(reg)

abline(reg)
```

the resulting line looks like this



And the result of summery is listed below

```
 > summary(reg)


Call:

lm(formula = ba ~ ch, data = datan)


Residuals:

    Min      1Q  Median      3Q     Max

-2.7552 -0.7904 -0.2732  0.7364  3.6970


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.106824   0.015950   69.39   <2e-16 ***

ch          0.211726   0.004739   44.68   <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.088 on 7997 degrees of freedom

Multiple R-squared:  0.1997, Adjusted R-squared:  0.1996

F-statistic:  1996 on 1 and 7997 DF,  p-value: < 2.2e-16
```

The R -squared of this regression is just 19% showing it is not a good estimate. this example shows some times there is a correlation between the variables, but you cant create regression for those variables.

Bibliography

MsSmartyPants (2021). Water quality. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/mssmartypants/water-quality [Accessed 30 Mar. 2022].