Cereals – correlation and regression analysis

Made by majafoi

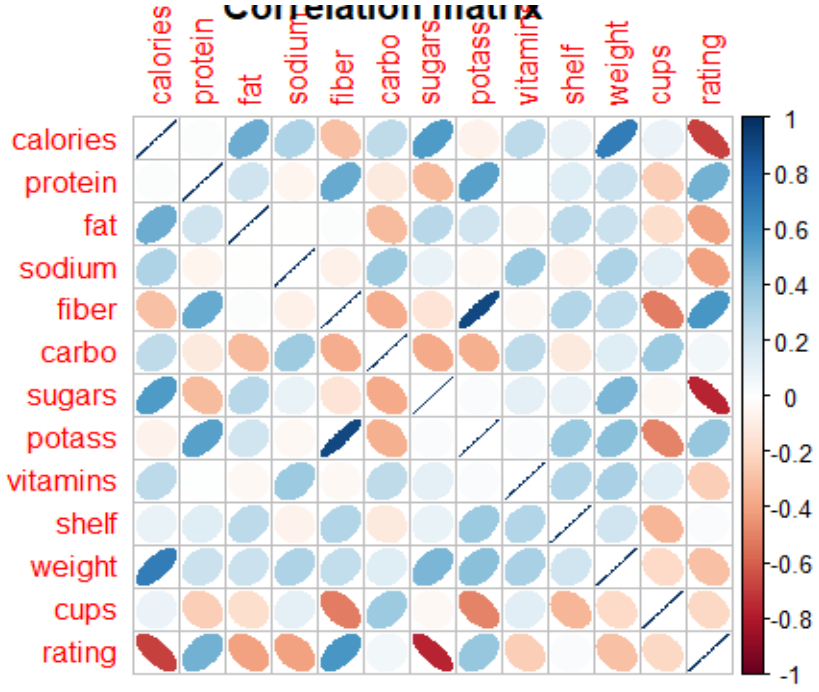
4/25/2021

In this part, I am going to show you correlation and simple regression analysis for the numeric variables in the dataset. Since there were some variable that were character, I created/saved a new CSV file containing only numeric values.

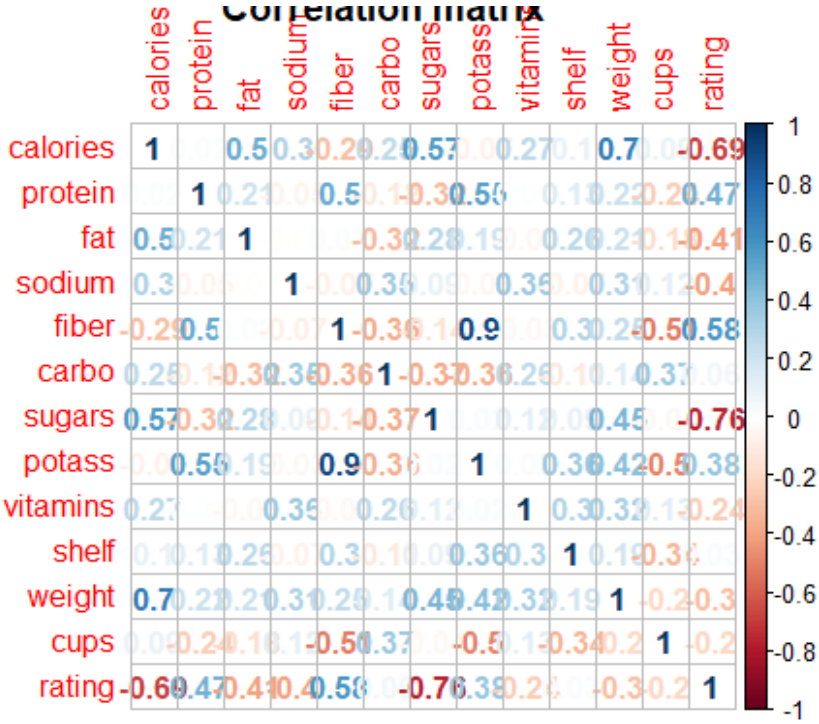
```
cereals <- read.csv ("cereal-cor.csv") # to Load dataset called Cereals
attach (cereals) # to omit the need of dataset-name$variable-name

library (corrplot)

## corrplot 0.84 loaded

library (ggpubr)

## Loading required package: ggplot2

cereals.correlation = cor (cereals)
corrplot (cereals.correlation, method = c("ellipse"), title = "Correlation matrix", sig.level = 0.05) #for the correl
ation matrix 1
```

Here we can see a correlation matrix, showing relationship between all numeric variables. A very important prerequisite of performing correlation analysis is that the variables included must be outlier-free, if we take that correlation in consideration.

From EDA, we could see that only variables that had 0 or 1 outlier were fat, sodium, carbo, sugars, shelf, cups and rating.



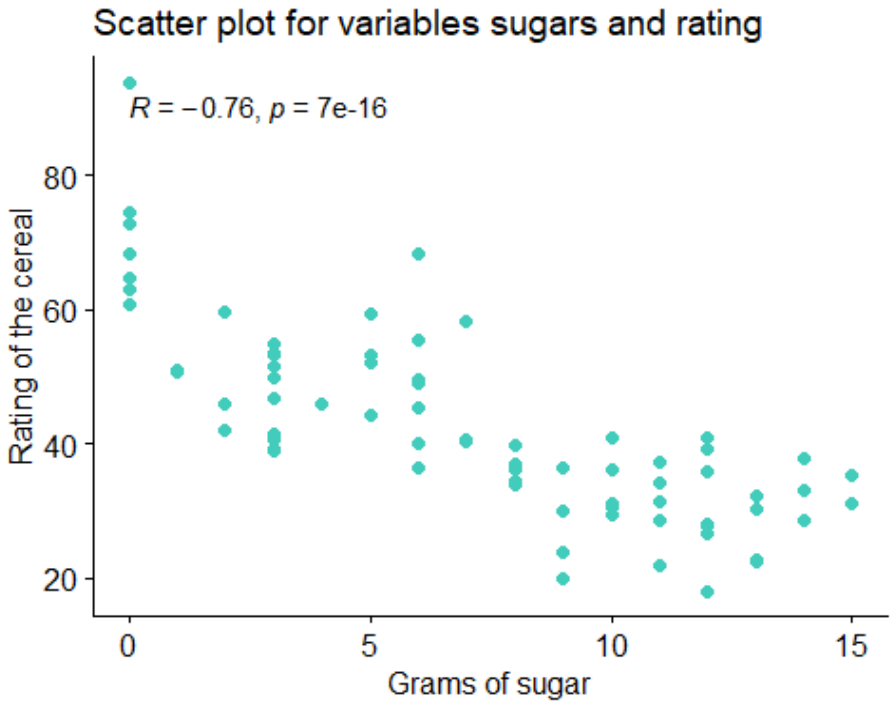

```
corrplot (cereals.correlation, method = c("number"), title = "Correlation matrix", sig.level = 0.05) #for the correlation matrix 2
```



These correlation matrices above show some significant correlation coefficients, but those variables included aren't outlier free. The only significant correlation is between variables named sugars and rating (-0.76).

To see what kind of relationship they have, I can create a scatter plot, which is below.

```
ggscatter (cereals, x = "sugars", y = "rating", color = "#41CCBC", title = "Scatter plot for variables sugars and r
ating", xlab = "Grams of sugar", ylab = "Rating of the cereal", cor.coef = TRUE)
#scatter plot to see the relationship between the two variables
```



As you can see only from the scatter plot above, the relationship between sugars and rating of the cereals is pretty negative and strong, with some outliers showing up. The coefficient is -0.76, with p-value of lower than 0.05, which makes this model significant.

Let's proceed with the regression analysis below, between those two variables.

```
reg1 = lm (rating~sugars)
print (summary (reg1))

##
## Call:
## lm(formula = rating ~ sugars)
##
## Residuals:
##    Min     1Q  Median     3Q    Max
## -17.849  -5.619  -1.359   4.926  34.117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.584     1.954   30.50  < 2e-16 ***
## sugars        -2.435     0.238  -10.23 6.99e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.135 on 75 degrees of freedom
## Multiple R-squared:  0.5827, Adjusted R-squared:  0.5771
## F-statistic: 104.7 on 1 and 75 DF,  p-value: 6.992e-16



ggplot (cereals, aes (x=sugars, y=rating)) + geom_point (alpha=.7, color = "#41CCBC") + stat_smooth (metho
d=lm, level = 0.95, color = "#DA4A62")

## `geom_smooth()` using formula 'y ~ x'
```

From the regression model we can see that the distribution of the residuals don't show a normal distribution (there is a higher variability present, must've been because of outliers), but the model shows that the sugars variable is significant in the model itself (associated p-value is lower than 0.05). Adjusted R-squared is 0.57, which means that 57% of the variability of the variable rating is being explained by the variable sugars. That is understandable, as there are way more variables taking into consideration by the customer when he/she is giving the rating, than the amount of sugar in the cereals.

If we want to predict the rating given by the customer using the amount of sugar in the cereals, here is how you proceed (take into account this is only suitable for this sample, not every cereal!, and that the R-squared is still low enough to give precise predictions):

Y= ax+b → linear equation model

Rating = -2.435*sugars + 59.584

So, for example, let's see which rating the cereals with 15 grams of sugar in it have:

Rating = -2.435*15 + 59.584 = 23

How about zero sugar?

Rating = -2.435*0 + 59.584 = 59.58