Exploratory Data Analysis of soil data

Mor ndiaye

05/09/2020

This dataset is provided by SOTER and I have make some cleaning in order to obtain data in which i can carry out somme exploratory data analysis without any problems.

Import dataset

```
data_sol<-read.csv2("C:\\Users\\pc\\Downloads\\Modelisation_Soter_data_Senegal-master\\Modelisation_Sot
```

Structure of datast

let's take a look in the data set

```
str(data_sol)
```

```
## 'data.frame':    418 obs. of  28 variables:
##  $ LNGI             : num  -16.3 -16.3 -16.3 -16.3 -16 ...
##  $ LATI             : num  14.3 14.3 14.3 14.3 15.6 ...
##  $ PRID             : Factor w/ 124 levels "SN002/KAL","SN008C/A001",..: 1 1 1 1 2 2 2 2 2 2 ...
##  $ Horizon          : int  1 2 3 4 1 2 3 4 5 6 ...
##  $ Nom_horizon      : Factor w/ 51 levels "(B)","A","A0",..: NA NA NA NA 15 4 7 16 42 43 ...
##  $ epais_hor        : int  13 24 36 110 6 17 59 90 103 135 ...
##  $ transition_distin: Factor w/ 4 levels "A","C","D","G": 2 4 2 NA NA NA NA NA NA NA ...
##  $ Munshell_col_hud : Factor w/ 80 levels "10YR 6/1,5","10YR2,5/3",..: 11 35 33 17 13 9 9 8 13 19 ..
##  $ sable_gros       : int  NA NA NA NA NA 37 34 33 NA NA ...
##  $ sable_moy        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ sable_fin        : int  NA NA NA NA NA 52 55 55 NA NA ...
##  $ sable_tr_fin     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ sable_total      : int  86 90 85 43 NA 90 88 88 NA NA ...
##  $ Limon            : int  8 7 7 37 NA 4 4 4 NA NA ...
##  $ Argile           : int  3 1 4 20 NA 6 7 8 NA NA ...
##  $ classe_TT        : Factor w/ 12 levels "C","CL","L","LS",..: 5 5 4 3 NA 5 5 5 NA NA ...
##  $ PH               : num  3.61 3.77 3.67 3.06 8.7 ...
##  $ PHKC             : num  3.41 3.63 3.37 2.85 NA ...
##  $ SO4              : num  0.78 0.25 0.78 7.51 NA ...
##  $ EXCA             : num  NA NA NA 0.68 NA ...
##  $ EXMG             : num  0.4 0.25 0.79 3.72 NA ...
##  $ EXNA             : num  0.51 0.38 0.53 4.53 NA ...
##  $ EXCK             : num  0.01 0.01 0.01 0.07 NA ...
##  $ EXAL             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CECS             : num  2.8 2.6 3.2 13.6 NA ...
##  $ total_carbone    : num  NA NA NA NA 2.3 ...
##  $ total_azote      : num  NA NA NA NA 0.22 ...
##  $ Phosphore        : int  NA NA NA NA 6 1 1 1 NA NA ...
```

Cut some variable

To make sure that the other analysis run without problem we decided to cut some variable that doesn't use in the future .

```
varq<-data_sol[,c("sable_total","Limon","Argile","PH","PHKC","EXCA","EXMG","EXNA","epais_hor","total_ca
```

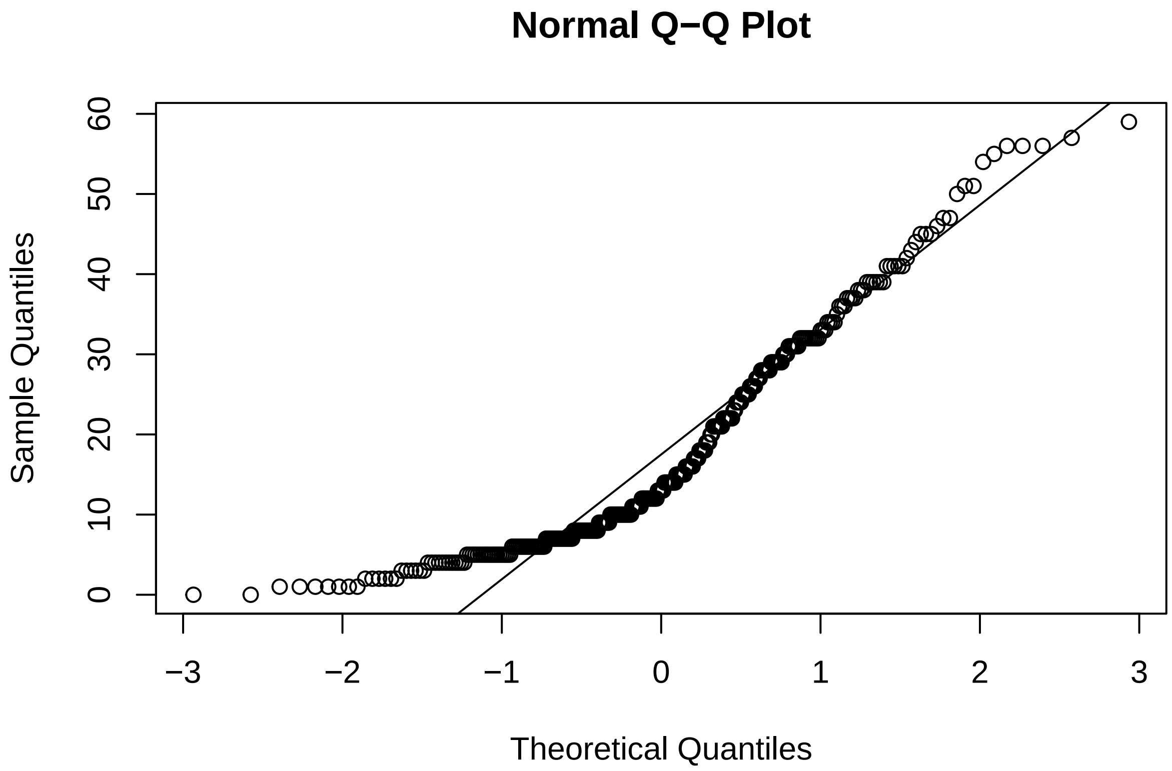
we use summary to take a look in certain parameter of differents variables

```
print(summary(varq))
```

```
##   sable_total        Limon           Argile            PH             PHKC      
##  Min.   : 1.00   Min.   : 0.00   Min.   : 0.00   Min.   :2.40   Min.   :2.850  
##  1st Qu.:53.00   1st Qu.: 4.00   1st Qu.: 7.00   1st Qu.:5.10   1st Qu.:4.213  
##  Median :74.50   Median : 9.00   Median :13.00   Median :5.70   Median :4.700  
##  Mean   :68.38   Mean   :13.37   Mean   :17.97   Mean   :5.92   Mean   :4.764  
##  3rd Qu.:89.00   3rd Qu.:19.00   3rd Qu.:28.00   3rd Qu.:6.60   3rd Qu.:5.175  
##  Max.   :99.00   Max.   :67.00   Max.   :59.00   Max.   :9.20   Max.   :7.600  
##  NA's   :116     NA's   :119     NA's   :118     NA's   :92     NA's   :304    
##       EXCA             EXMG              EXNA          epais_hor     
##  Min.   : 0.050   Min.   :  0.000   Min.   : 0.0000   Min.   :  1.00  
##  1st Qu.: 0.765   1st Qu.:  0.330   1st Qu.: 0.0550   1st Qu.: 25.00  
##  Median : 1.500   Median :  0.610   Median : 0.1200   Median : 64.00  
##  Mean   : 3.397   Mean   :  1.752   Mean   : 0.5641   Mean   : 70.03  
##  3rd Qu.: 2.775   3rd Qu.:  1.210   3rd Qu.: 0.2800   3rd Qu.:101.00  
##  Max.   :46.200   Max.   :109.000   Max.   :40.0000   Max.   :235.00  
##  NA's   :155      NA's   :141       NA's   :151       NA's   :2       
##  total_carbone        CECS         
##  Min.   : 0.069   Min.   :   0.700  
##  1st Qu.: 2.078   1st Qu.:   2.600  
##  Median : 3.350   Median :   5.000  
##  Mean   : 5.692   Mean   :  13.664  
##  3rd Qu.: 6.325   3rd Qu.:   9.855  
##  Max.   :62.400   Max.   :1340.000  
##  NA's   :170      NA's   :156       
```

check the normality of some variables with function qqnorm and qqline

```
qqnorm(varq$Argile)
qqline(varq$Argile)
```



I make this for one argile we can make this for th othe r variable in the dataset

```
mod_1<-lm(EXMG~Limon+CECS,data=na.omit(varq))
summary(mod_1)
```

```
##
## Call:
## lm(formula = EXMG ~ Limon + CECS, data = na.omit(varq))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4705 -0.3946  0.0646  0.2401  4.1161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.38615    0.17679  -2.184   0.0323 *
## Limon        0.10344    0.01039   9.952 5.68e-15 ***
## CECS         0.11491    0.02052   5.600 4.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9714 on 69 degrees of freedom
## Multiple R-squared:  0.7515, Adjusted R-squared:  0.7443
## F-statistic: 104.3 on 2 and 69 DF,  p-value: < 2.2e-16
```