CASE STUDY NUMBER 2

FOR TRAINING
Realized by Nicolas BEHBAHANI

The 17th of september 2017

RÉSUMÉ. — It's a case study for a recruitment of a datascientist.

Remarque. — This case study was done in only two days with $\LaTeX$

Sommaire

Task

All these exercises were done with the language R using RStudio.

```
Environment   History   Spark
Import Dataset
Global Environment
Data
doc                        32561 obs. of 17 variables
  age : num 39 50 38 53 28 37 49 52 31 42 ...
  workclass : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
  fnlwgt : num 77516 83311 215646 234721 338409 ...
  education : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
  education_num : num 13 13 9 7 13 14 5 9 14 13 ...
  marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
  occupation : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
  relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
  race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
  sex : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
  capital_gain : num 2174 0 0 0 0 ...
  capital_loss : num 0 0 0 0 0 0 0 0 0 0 ...
  hours_per_week: num 40 13 40 40 40 40 16 45 50 40 ...
  native_country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
  seuils : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
  capital : num 2174 0 0 0 0 ...
  c.capital : num 0.0799 -0.0668 -0.0668 -0.0668 -0.0668 ...
 mat_1                     num [1:2, 1:2] 4 2 4 2
 mat_2                     num [1:2, 1:2] 2 1 2 1
test                       16281 obs. of 17 variables
Values
```

We load the two files : adult.data.txt and adult.test.txt.

The objectives

The task is to predict whether income exceeds $50K/yr based on census data. The data can be found at : https ://archive.ics.uci.edu/ml/datasets/Census+Income or more precisely at : https ://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

§ 1. Prepare data for analysis

Testing for missing values

```
> c.age <- sd(doc$age, na.rm=TRUE)
> c.age <- sd(doc$age, na.rm=TRUE)
> c.fnlwgt <- sd(doc$fnlwgt, na.rm=TRUE)
> c.education_num <- sd(doc$education_num, na.rm=TRUE)
> c.capital_gain<- sd(doc$capital_gain, na.rm=TRUE)
> c.capital_loss <- sd(doc$capital_loss, na.rm=TRUE)
> c.hours_per_week <- sd(doc$hours_per_week, na.rm=TRUE)
>
```

1.1 Cleaning Data

We put names for each column and change to numeric some columns.

```
> colNames
 [1] "age"            "workclass"      "fnlwgt"         "education"      "education_num"  "marital_status" "occupation"     "relationship"
 [9] "race"           "sex"            "capital_gain"   "capital_loss"   "hours_per_week" "native_country" "seuils"
>
```

Now, we can explore our data without missing values and with the correct numeric columns. For example, we can see the begininng and the end of our data :

```
> head(doc)
  age        workclass fnlwgt  education education_num     marital_status        occupation  relationship  race    sex capital_gain
1  39        State-gov  77516  Bachelors            13      Never-married      Adm-clerical Not-in-family White   Male         2174
2  50 Self-emp-not-inc  83311  Bachelors            13 Married-civ-spouse   Exec-managerial       Husband White   Male            0
3  38          Private 215646    HS-grad             9           Divorced Handlers-cleaners Not-in-family White   Male            0
4  53          Private 234721       11th             7 Married-civ-spouse Handlers-cleaners       Husband Black   Male            0
5  28          Private 338409  Bachelors            13 Married-civ-spouse    Prof-specialty          Wife Black Female            0
6  37          Private 284582    Masters            14 Married-civ-spouse   Exec-managerial          Wife White Female            0
  capital_loss hours_per_week native_country seuils capital   c.capital
1            0             40  United-States   <=50K    2174  0.07987968
2            0             13  United-States   <=50K       0 -0.06683404
3            0             40  United-States   <=50K       0 -0.06683404
4            0             40  United-States   <=50K       0 -0.06683404
5            0             40           Cuba   <=50K       0 -0.06683404
6            0             40  United-States   <=50K       0 -0.06683404
```

```
> tail(doc)
      age      workclass fnlwgt    education education_num     marital_status        occupation  relationship  race    sex capital_gain
32556  22        Private 310152 Some-college            10      Never-married   Protective-serv Not-in-family White   Male            0
32557  27        Private 257302   Assoc-acdm            12 Married-civ-spouse      Tech-support          Wife White Female            0
32558  40        Private 154374      HS-grad             9 Married-civ-spouse Machine-op-inspct       Husband White   Male            0
32559  58        Private 151910      HS-grad             9            Widowed      Adm-clerical     Unmarried White Female            0
32560  22        Private 201490      HS-grad             9      Never-married      Adm-clerical     Own-child White   Male            0
32561  52   Self-emp-inc 287927      HS-grad             9 Married-civ-spouse   Exec-managerial          Wife White Female        15024
      capital_loss hours_per_week native_country seuils capital   c.capital
32556            0             40  United-States   <=50K       0 -0.06683404
32557            0             38  United-States   <=50K       0 -0.06683404
32558            0             40  United-States    >50K       0 -0.06683404
32559            0             40  United-States   <=50K       0 -0.06683404
32560            0             20  United-States   <=50K       0 -0.06683404
32561            0             40  United-States    >50K   15024  0.94706976
>
```

So, the structure of our data is with the *str* command :
- 32561 observations
- 17 variables

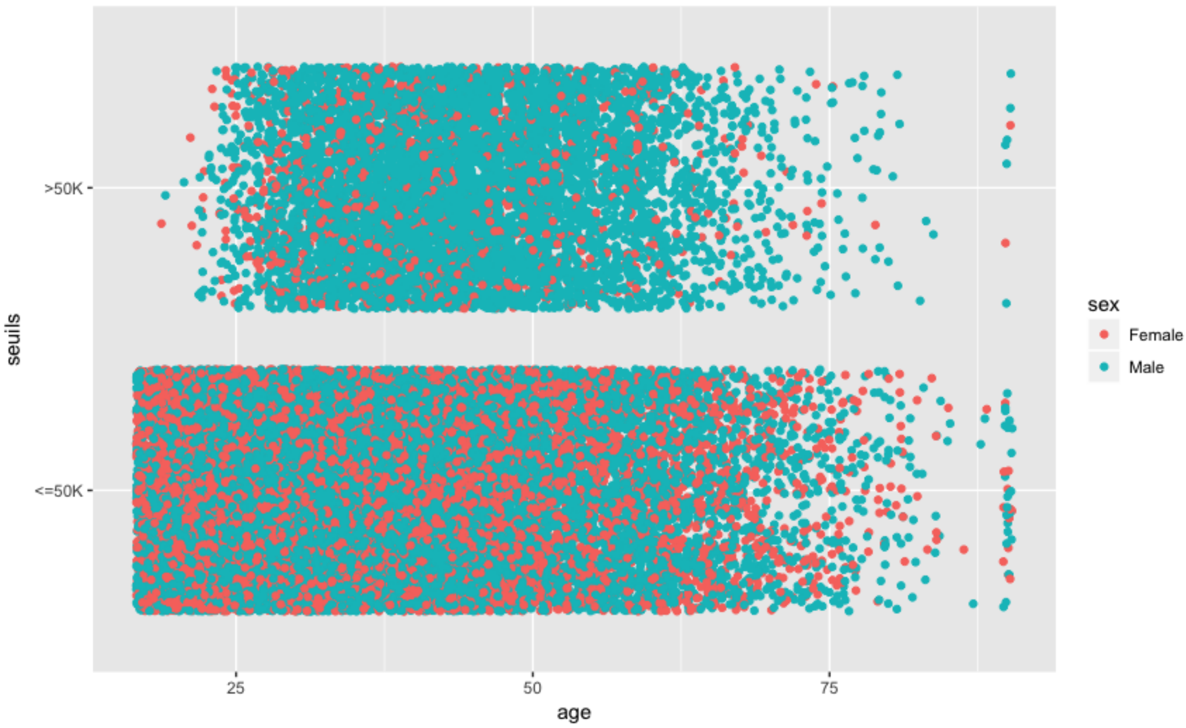
```
> str(doc)
'data.frame':   32561 obs. of  17 variables:
 $ age           : num  39 50 38 53 28 37 49 52 31 42 ...
 $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
 $ fnlwgt        : num  77516 83311 215646 234721 338409 ...
 $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
 $ education_num : num  13 13 9 7 13 14 5 9 14 13 ...
 $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
 $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
 $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
 $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
 $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
 $ capital_gain  : num  2174 0 0 0 0 ...
 $ capital_loss  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ hours_per_week: num  40 13 40 40 40 40 16 45 50 40 ...
 $ native_country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
 $ seuils        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
 $ capital       : num  2174 0 0 0 0 ...
 $ c.capital     : num  0.0799 -0.0668 -0.0668 -0.0668 -0.0668 ...
```

§ 2. Exploratory Analysis

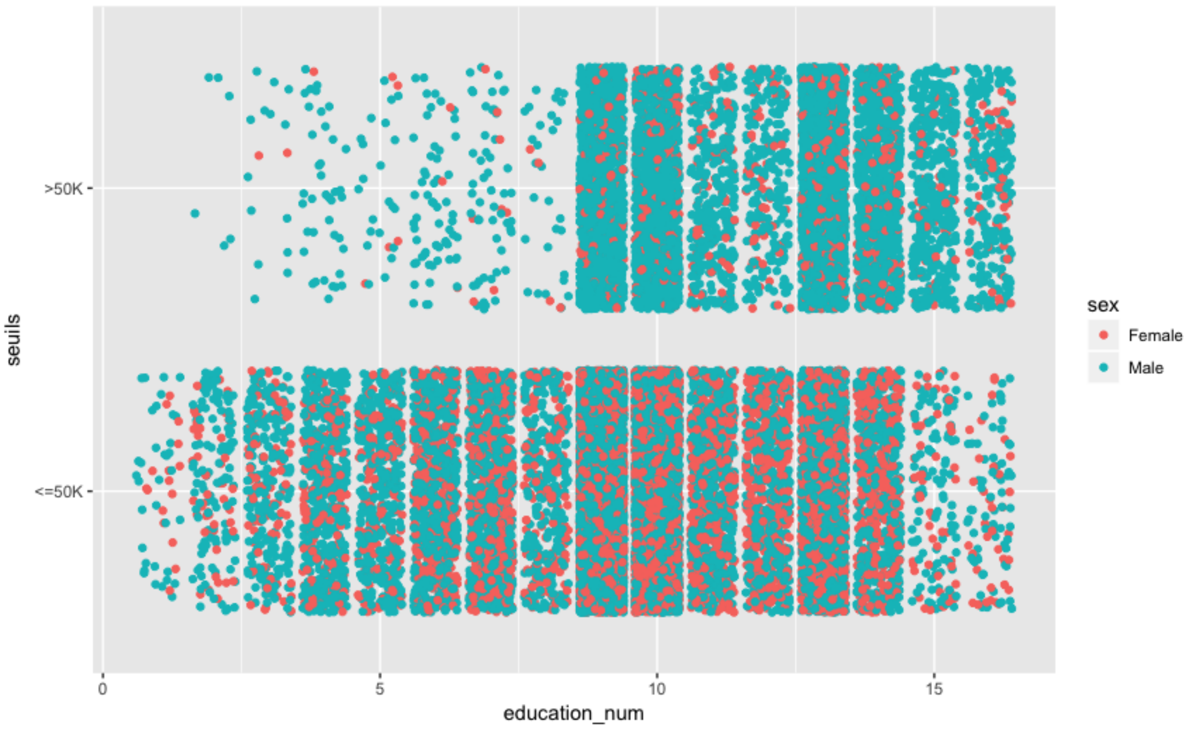
We can see a summary of our data in order to see possible abrasive values :

```
> summary(doc)
      age                   workclass          fnlwgt                 education      education_num                  marital_status
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
 Mean   :38.58   ?               : 1836   Mean   : 189778   Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
 3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051   Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
 Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705   11th        : 1175   Max.   :16.00   Separated            : 1025
                (Other)          :  981                    (Other)      : 5134                   Widowed              :  993
           occupation            relationship                   race            sex          capital_gain    capital_loss    hours_per_week
  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
  Exec-managerial:4066   Other-relative:  981   Black             : 3124                  Median :    0   Median :   0.0   Median :40.00
  Adm-clerical   :3770   Own-child     : 5068   Other             :  271                  Mean   : 1078   Mean   :  87.3   Mean   :40.44
  Sales          :3650   Unmarried     : 3446   White             :27816                  3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
  Other-service  :3295   Wife          : 1568                                             Max.   :99999   Max.   :4356.0   Max.   :99.00
 (Other)         :9541
       native_country      seuils         capital           c.capital
  United-States:29170   <=50K:24720   Min.   :-4356.0   Min.   :-0.36080
  Mexico       :  643   >50K : 7841   1st Qu.:    0.0   1st Qu.:-0.06683
  ?            :  583                 Median :    0.0   Median :-0.06683
  Philippines  :  198                 Mean   :  990.4   Mean   : 0.00000
  Germany      :  137                 3rd Qu.:    0.0   3rd Qu.:-0.06683
  Canada       :  121                 Max.   :99999.0   Max.   : 6.68166
 (Other)       : 1709
```

We use the *ggplot* to see the distribution of males and females :



Conlusion : It's clear that there are more males than females who earn more than 50K/yr. in a same graph, we can plot another parameter wich is education :



§ 3. Choose a model

3.1 Apply a model in our Data

We choose this model :

```
> model1 <- glm(seuils ~ sex + age + education_num + workclass + occupation + c.capital, data=doc, family=binomial(link="logit"))
```

and the result is :

```
> model1

Call:  glm(formula = seuils ~ sex + age + education_num + workclass +
    occupation + c.capital, family = binomial(link = "logit"),
    data = doc)

Coefficients:
                 (Intercept)                     sex Male                          age                education_num
                    -7.65858                      1.29168                      0.04071                      0.27348
        workclass Federal-gov          workclass Local-gov       workclass Never-worked            workclass Private
                     1.38839                      0.90014                     -9.63562                      1.00818
       workclass Self-emp-inc     workclass Self-emp-not-inc          workclass State-gov        workclass Without-pay
                     1.50904                      0.72993                      0.66106                    -10.90354
      occupation Adm-clerical       occupation Armed-Forces      occupation Craft-repair   occupation Exec-managerial
                    -0.14584                     -0.94994                      0.10621                      0.83545
   occupation Farming-fishing  occupation Handlers-cleaners  occupation Machine-op-inspct     occupation Other-service
                    -0.82166                     -0.98244                     -0.25816                     -1.20364
   occupation Priv-house-serv     occupation Prof-specialty   occupation Protective-serv             occupation Sales
                    -3.74351                      0.47984                      0.50512                      0.23683
      occupation Tech-support   occupation Transport-moving                    c.capital
                     0.45078                           NA                      3.48888

Degrees of Freedom: 32560 Total (i.e. Null);  32535 Residual
Null Deviance:      35950
Residual Deviance: 25740        AIC: 25800
```

finally, we have to calculate the prediction error :

```
> glm.pred = rep(" <=50K.", length(test$seuils))
> glm.pred[glm.probs >= 0.5] = " >50K."
> table(glm.pred, test$seuils)

glm.pred   <=50K.  >50K.
   <=50K.   11618   2029
   >50K.      817   1817
>
> # prediction error
> 1 - mean(glm.pred == test$seuils)
[1] 0.174805
```

Conclusion : We have a score of 17,48% of error for this linear model wich is a good result and we expect certainly a lower percentage error with an non linear-model.

§ 4. Task 2

The size of a matrix is defined by the number of rows and columns that it contains. A matrix with m rows and n columns is called an $m \; n$ matrix or m-by-n matrix, while m and n are called its dimensions. We use *solve*() for inverse the matrix.

```
> x1 = c(2, 4, 1)
> x2 = c(4, 1, 1)
> x3 = c(2, -1, 3)
> X = rbind(x1,x2,x3)
> X
   [,1] [,2] [,3]
x1    2    4    1
x2    4    1    1
x3    2   -1    3
> solve(X)
             x1         x2          x3
[1,] -0.1052632  0.3421053 -0.07894737
[2,]  0.2631579 -0.1052632 -0.05263158
[3,]  0.1578947 -0.2631579  0.36842105
>
```