Data Mining - project (Adults)

Wiktoria Fimińska 262283, Julia Grzegorzewska 262314

2024-04-01

Spis treści

1 Introduction

2 Data

```
##   age workclass fnlwgt  education education.num      marital.status
## 1  52   Private 178983  Doctorate            16  Married-civ-spouse
## 2  27   Private 137063  Bachelors            13  Married-civ-spouse
## 3  41   Private 207779    HS-grad             9           Separated
## 4  50   Private 130780    HS-grad             9  Married-civ-spouse
## 5  36   Private 160120    HS-grad             9  Married-civ-spouse
##           occupation   relationship                race     sex capital.gain
## 1     Prof-specialty        Husband               White    Male            0
## 2              Sales        Husband               White    Male            0
## 3      Other-service  Not-in-family               White  Female            0
## 4  Handlers-cleaners        Husband               Black    Male            0
## 5       Adm-clerical        Husband  Asian-Pac-Islander    Male            0
##   capital.loss hours.per.week native.country income
## 1            0             40              ?   >50K
## 2            0             50  United-States  <=50K
## 3            0             40  United-States  <=50K
## 4            0             40  United-States  <=50K
## 5            0             40        Vietnam  <=50K
```

```
# which columns have '?' values
columns_with_question <- names(data)[sapply(data, function(col) any(col == " ?"))]

print(columns_with_question)
```

```
## [1] "workclass"      "occupation"     "native.country"
```

2.1 Data characteristics

Dataset contains 14 variables, which are:

(a) **age**
age of a person - continuous variable.

(b) **workclass**
workclass of a person: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

(c) **fnlwgt**
NO IDEA WHAT IT IS

(d) **education**
level of education of a person: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

(e) **education-num**
continuous version of education (ONE OF THEM WON'T BE NECESSARY TO USE - WE JUST DROP IT)

(f) **marital-status**
marital status of a person: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

(g) **occupation**
occupation of a person: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

(h) **relationship**
relationship status of a person: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

(i) **race**
race of a person: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

(j) **sex**
sex of a person: Female, Male.

(k) **capital-gain**
amount of money a person gained - continuous variable.

(l) **capital-loss**
amount of money a person loosed - continuous variable.

(m) **hours-per-week**
XYZ idk

(n) **native-country**
native-country of a person: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

2.2 Data preparation

```
##    age     workclass fnlwgt     education education.num      marital.status
## 1   30       Private 231620     Bachelors            13       Never-married
## 2   55       Private 160362     Bachelors            13  Married-civ-spouse
## 3   55       Private 189528       5th-6th             3  Married-civ-spouse
## 4   46     Local-gov 111558  Some-college            10           Divorced
## 5   46       Private 250821   Prof-school            15           Divorced
## 6   43  Self-emp-inc  83348       HS-grad             9  Married-civ-spouse
## 7   41       Private 237452     Bachelors            13  Married-civ-spouse
## 8   17       Private 130125          10th             6       Never-married
## 9   34       Private  49469     Bachelors            13       Never-married
## 10  36       Private 144154       HS-grad             9           Divorced
##             occupation   relationship                race     sex capital.gain
## 1                Sales  Not-in-family               White  Female            0
## 2         Adm-clerical        Husband               White    Male            0
## 3         Craft-repair        Husband               White    Male            0
## 4   Machine-op-inspct      Own-child               White  Female            0
## 5      Farming-fishing      Unmarried               White    Male            0
## 6         Craft-repair        Husband               White    Male            0
## 7       Prof-specialty           Wife               White  Female            0
## 8        Other-service      Own-child  Amer-Indian-Eskimo  Female         1055
## 9                Sales  Not-in-family               White    Male        99999
## 10    Exec-managerial  Not-in-family               White  Female            0
##    capital.loss hours.per.week native.country income
## 1             0             40         Mexico  <=50K
## 2             0             40  United-States   >50K
## 3             0             50  United-States  <=50K
## 4             0             40  United-States  <=50K
## 5             0             48  United-States  <=50K
## 6             0             60  United-States  <=50K
## 7             0             15           Cuba   >50K
## 8             0             20  United-States  <=50K
## 9             0             50  United-States   >50K
## 10            0             40  United-States  <=50K
```

Missing observations?

```
sum(is.na(data))
```

```
## [1] 4262
```

WE NEED TO FIND A METHOD TO FILL THE NANS

```
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.58   48.00   90.00
```

```
summary(capital.gain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1078       0   99999
```

```
summary(capital.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   87.31    0.00 4356.00
```

```
summary(hours.per.week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.44   45.00   99.00
```

```
table(workclass)
```

```
## workclass
##                 ?       Federal-gov         Local-gov      Never-worked
##              1836               960              2093                 7
##           Private      Self-emp-inc  Self-emp-not-inc         State-gov
##             22696              1116              2541              1297
##       Without-pay
##                14
```

```
table(education)
```

```
## education
##          10th          11th          12th       1st-4th       5th-6th
##           933          1175           433           168           333
##       7th-8th           9th    Assoc-acdm     Assoc-voc     Bachelors
##           646           514          1067          1382          5354
##     Doctorate       HS-grad       Masters     Preschool   Prof-school
##           413         10501          1723            51           576
##  Some-college
##          7291
```

```
table(marital.status)
```

```
## marital.status
##               Divorced      Married-AF-spouse     Married-civ-spouse
##                   4443                     23                  14976
##  Married-spouse-absent          Never-married              Separated
##                    418                  10682                   1025
##                Widowed
##                    993
```

```
table(occupation)
```

```
## occupation
##                   ?       Adm-clerical       Armed-Forces       Craft-repair
##                1843               3769                  9               4099
##    Exec-managerial    Farming-fishing  Handlers-cleaners  Machine-op-inspct
##                4066                994               1370               2002
##      Other-service    Priv-house-serv     Prof-specialty    Protective-serv
##                3295                149               4140                649
##               Sales       Tech-support   Transport-moving
##                3650                928               1597
```

```
table(relationship)
```

```
## relationship
##         Husband   Not-in-family  Other-relative       Own-child       Unmarried
##           13193            8304             981            5068            3446
##            Wife
##            1568
```

```
table(race)
```

```
## race
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black               Other
##                 311                1039                3124                 271
##               White
##               27815
```

```
table(sex)
```

```
## sex
##  Female    Male
##   10771   21789
```

```
table(income)
```

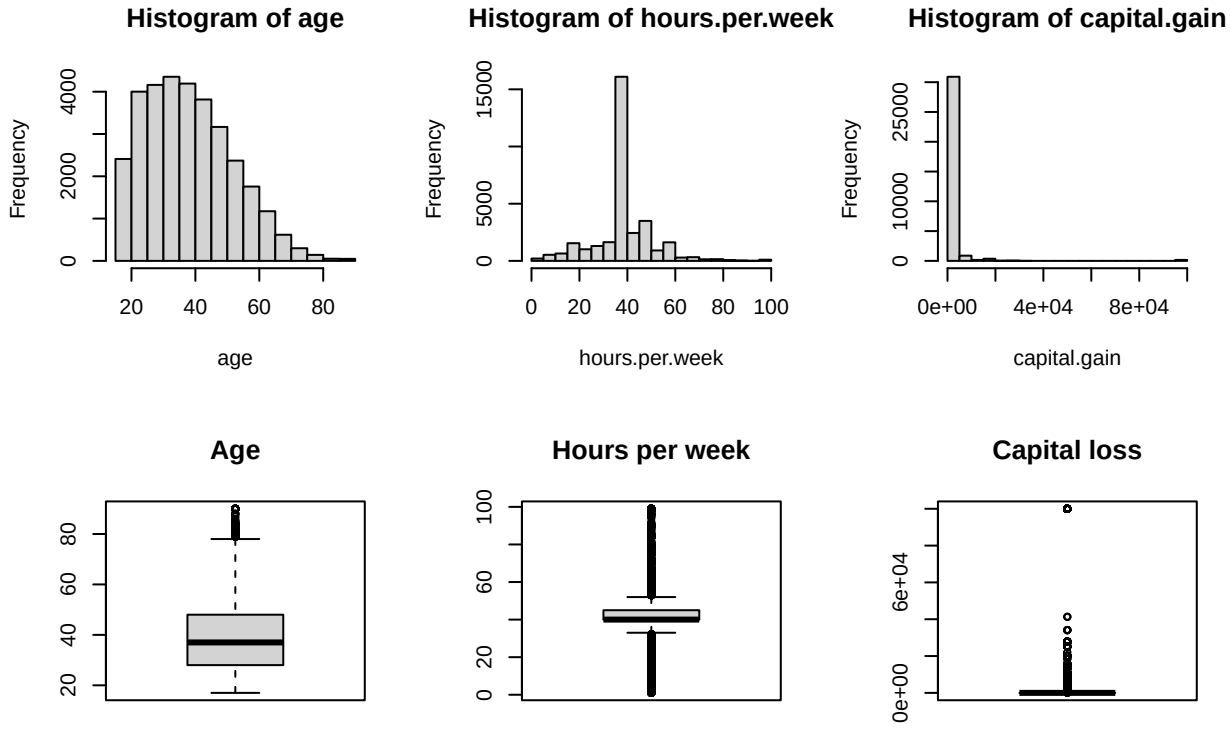
```
## income
##  <=50K   >50K
##  24719   7841
```

```
table(native.country)
```

```
## native.country
##                             ?                      Cambodia
##                           583                            19
##                        Canada                         China
##                           121                            75
##                      Columbia                          Cuba
##                            59                            95
##           Dominican-Republic                       Ecuador
##                            70                            28
##                   El-Salvador                       England
##                           106                            90
##                        France                       Germany
##                            29                           137
##                        Greece                     Guatemala
##                            29                            64
##                         Haiti           Holand-Netherlands
##                            44                             1
##                      Honduras                          Hong
##                            13                            20
##                       Hungary                         India
##                            13                           100
##                          Iran                       Ireland
##                            43                            24
##                         Italy                       Jamaica
##                            73                            81
##                         Japan                          Laos
##                            62                            18
##                        Mexico                     Nicaragua
##                           643                            34
##  Outlying-US(Guam-USVI-etc)                          Peru
##                            14                            31
##                   Philippines                        Poland
##                           198                            60
##                      Portugal                   Puerto-Rico
##                            37                           114
```

```
##                   Scotland                      South
##                         12                         80
##                     Taiwan                   Thailand
##                         51                         18
##           Trinadad&Tobago             United-States
##                         19                      29169
##                    Vietnam                 Yugoslavia
##                         67                         16
```

```
par(mfrow=c(2,3))
hist(age)
hist(hours.per.week)
hist(capital.gain)
boxplot(age, main="Age")
boxplot(hours.per.week, main="Hours per week")
boxplot(capital.gain, main="Capital loss")
```



Rysunek 1: Charts: continuous quantitative data

```r
par(mfrow=c(1,1))
```

3 Analysis

4 Discussion