task1

Manvydas Sokolovas

10/29/2018

```
library(data.table) # data.table package will be used to complete most of the tasks
data_full <- fread("bank/bank-full.csv")
data <- fread("bank/bank.csv")

## 1. -----
n <- round(nrow(data_full)*0.1, 0) # 10% of data
smpl <- data_full[sample(.N, n)]

## 2. -----
duom <- data_full[pdays != (-1)] # choosing customers who were contacted before
duom[duom == "unknown"] <- NA
duom <- na.omit(duom)

x1 <- duom[!job %in% c("unemployed", "retired", "student") & balance > 0 & housing == "no" & loan == "n
head(x1)
```

```
##    age         job  marital education default balance housing loan
## 1:  33      admin.  married  tertiary      no     882      no   no
## 2:  51      admin.   single secondary      no    3132      no   no
## 3:  51  management divorced  tertiary      no     119      no   no
## 4:  49  management  married  tertiary      no    1533      no   no
## 5:  47 blue-collar  married secondary      no    1484      no   no
## 6:  38  management  married  tertiary      no     494      no   no
##      contact day month duration campaign pdays previous poutcome  y
## 1: telephone  21   oct       39        1   151        3  failure no
## 2: telephone   5   nov      449        1   176        1  failure no
## 3:  cellular  17   nov      200        1   165        2  failure no
## 4:  cellular  17   nov      324        1   172        1  failure no
## 5:  cellular  17   nov      297        1   119        3  failure no
## 6:  cellular  17   nov      146        1   104        2    other no
```

```
## 3. -----
x2 <- duom[, !c("housing", "default")]
head(x2)
```

```
##    age         job marital education balance loan   contact day month
## 1:  33      admin. married  tertiary     882   no telephone  21   oct
## 2:  42      admin.  single secondary    -247  yes telephone  21   oct
## 3:  33    services married secondary    3444   no telephone  21   oct
## 4:  36  management married  tertiary    2415   no telephone  22   oct
## 5:  36  management married  tertiary       0   no telephone  23   oct
## 6:  44 blue-collar married secondary    1324   no telephone  25   oct
##    duration campaign pdays previous poutcome   y
## 1:       39        1   151        3  failure  no
## 2:      519        1   166        1    other yes
## 3:      144        1    91        4  failure yes
## 4:       73        1    86        4    other  no
## 5:      140        1   143        3  failure yes
## 6:      119        1    89        2    other   no
```

```
a <- NULL; a$ncol1 <- ncol(duom); a$ncol2 <- ncol(x2); a # two columns were deleted
```

```
## $ncol1
## [1] 17
##
## $ncol2
## [1] 15
```

```
setnames(x1, c("housing", "y"), c("housingloan", "termdep")) # renaming two variables

## 4. -----

table(duom$y) # shows how many people are subsribed a term deposit and how many are not
```

```
##
##   no  yes
## 6056 1786
```

```
round(prop.table(table(duom$job, duom$y), margin = 1), 3) # percentages by type of job (each row sums t
```

```
##
##                    no   yes
##   admin.        0.773 0.227
##   blue-collar   0.887 0.113
##   entrepreneur  0.882 0.118
##   housemaid     0.781 0.219
##   management    0.719 0.281
##   retired       0.581 0.419
##   self-employed 0.758 0.242
##   services      0.837 0.163
##   student       0.561 0.439
##   technician    0.790 0.210
##   unemployed    0.611 0.389
```

```
round(prop.table(table(duom$education, duom$y), margin = 1), 3) # percentages by education (each row su
```

```
##
##                no   yes
##   primary   0.830 0.170
##   secondary 0.800 0.200
##   tertiary  0.705 0.295
```

```
dat <- as.data.frame(duom)
q <- sapply(dat, class)
x3 <- dat[, noquote(q == "numeric") | (q == "integer")] # choosing numeric or integer class variables f
summary(x3, digits = 5)
```

```
##       age            balance            day           duration
##  Min.   :18.000   Min.   :-1884.0   Min.   : 1.00   Min.   :   5.00
##  1st Qu.:32.000   1st Qu.:  162.0   1st Qu.: 7.00   1st Qu.: 113.00
##  Median :38.000   Median :  595.0   Median :14.00   Median : 194.00
##  Mean   :40.784   Mean   : 1552.3   Mean   :14.26   Mean   : 261.29
##  3rd Qu.:47.000   3rd Qu.: 1733.8   3rd Qu.:20.00   3rd Qu.: 324.00
##  Max.   :89.000   Max.   :81204.0   Max.   :31.00   Max.   :2219.00
##     campaign           pdays           previous
##  Min.   : 1.0000   Min.   :   1.00   Min.   : 1.0000
##  1st Qu.: 1.0000   1st Qu.:133.00   1st Qu.:  1.0000
##  Median : 2.0000   Median :195.00   Median :  2.0000
##  Mean   : 2.0643   Mean   :223.25   Mean   :  3.1843
##  3rd Qu.: 2.0000   3rd Qu.:326.00   3rd Qu.:  4.0000
##  Max.   :16.0000   Max.   :871.00   Max.   :275.0000
```

```r
pp <- seq(0.1, 0.90, 0.1)
sapply(x3, quantile, probs = pp) # quantiles
```

```
##     age balance day duration campaign pdays previous
## 10%  29       0   4       67        1    91        1
## 20%  31      96   6       99        1   109        1
## 30%  33     234   9      129        1   160        1
## 40%  36     392  12      159        1   181        2
## 50%  38     595  14      194        2   195        2
## 60%  41     917  16      236        2   258        3
## 70%  45    1390  18      290        2   300        3
## 80%  50    2212  20      370        3   342        4
## 90%  57    3990  27      532        4   361        6
```

```r
duom[, .(median(duration), mean(balance)), by = .(housing, job)] #  median of last contact durations an
```

```
##     housing           job    V1        V2
##  1:      no        admin. 204.0 1463.5182
##  2:     yes        admin. 178.0 1005.2786
##  3:     yes      services 185.0 1099.4763
##  4:     yes    management 173.0 1866.5844
##  5:     yes   blue-collar 180.0 1053.9047
##  6:      no    technician 209.0 1835.6516
##  7:     yes    unemployed 184.5 1579.0889
##  8:     yes  entrepreneur 184.0  951.8742
##  9:      no    management 209.0 2183.0380
## 10:     yes    technician 177.0 1256.0484
## 11:     yes     housemaid 169.0 2044.0333
## 12:      no   blue-collar 209.5 1719.9756
## 13:     yes       retired 180.0 1324.6901
## 14:      no  entrepreneur 230.5 1746.0000
## 15:      no      services 244.0 1350.4581
## 16:      no       retired 263.0 3034.7726
## 17:     yes self-employed 174.0 1493.6101
## 18:      no self-employed 222.0 2790.4857
## 19:      no     housemaid 178.0 1484.3372
## 20:      no    unemployed 302.5 1507.9407
## 21:      no       student 221.5 1483.6111
## 22:     yes       student 153.0 1769.1754
##     housing           job    V1        V2
```

```r
## 5. -----
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
## The following object is masked from 'package:base':
##
##     date
```

```
duom[, "date" := ymd(paste(2011, duom$month, duom$day))] # creating date variable
duom[, "day_of_week" := weekdays(duom$date)] # day of week variable

duom[, "birthdate" := 2011-age] # calculating birth date from age variable

duom[, "ageclass" := factor(cut(age, breaks = c(min(age), 34, 50, 70, max(age))))] # new variable with
table(duom$ageclass)
```

```
##
## (18,34] (34,50] (50,70] (70,89]
##    2757    3547    1372     164
```

```
## 6. -----
x4 <- duom[order(rank(job), -balance, age)] # ordering data by job, balance and age. Jobs by alphabet,
head(x4)
```

```
##    age    job  marital education default balance housing loan   contact
## 1:  29 admin.  married secondary      no   22171     yes   no  cellular
## 2:  57 admin.  married secondary      no   16873      no   no  cellular
## 3:  42 admin.  married secondary      no   16517      no   no  cellular
## 4:  42 admin.  married secondary      no   16517      no   no  cellular
## 5:  60 admin.  married secondary      no   12980      no   no  cellular
## 6:  60 admin. divorced secondary      no   12039      no   no telephone
##    day month duration campaign pdays previous poutcome   y       date
## 1:  18   may       44        1   355        3  failure  no 2011-05-18
## 2:  14   oct      219        3   372        1  failure  no 2011-10-14
## 3:  24   aug      497        2   279        2  failure  no 2011-08-24
## 4:  15   mar      549        5   203        4  failure  no 2011-03-15
## 5:   3   sep      177        2   182        1  success  no 2011-09-03
## 6:  12   oct      261        1   187        1  success yes 2011-10-12
##    day_of_week birthdate ageclass
## 1:   Wednesday      1982  (18,34]
## 2:      Friday      1954  (50,70]
## 3:   Wednesday      1969  (34,50]
## 4:     Tuesday      1969  (34,50]
## 5:    Saturday      1951  (50,70]
## 6:   Wednesday      1951  (50,70]
```