Homework1

YoungJin Lee

3/28/2018

### Question 1. How many clients are included in the data? How many clients are younger than 30 and how many are older than 50?

## [1] 4521

## [1] 482

## [1] 927

#### ANSWER: 4521 clients are included in the data. 482 clients are younger than 30 and 927 clients are older than 50

### Question 2. “balance” field represents bank account balance in euros. Add new field named “balance\_kw” that shows the balance in Korean won. Let us assume the exchange rate of currency is 1200 kw = 1 euro

## age job marital education default balance housing loan contact  
## 1 30 unemployed married primary no 1787 no no cellular  
## 2 33 services married secondary no 4789 yes yes cellular  
## 3 35 management single tertiary no 1350 yes no cellular  
## 4 30 management married tertiary no 1476 yes yes unknown  
## 5 59 blue-collar married secondary no 0 yes no unknown  
## 6 35 management single tertiary no 747 no no cellular  
## day month duration campaign pdays previous poutcome y balance\_kw  
## 1 19 oct 79 1 -1 0 unknown no 2144400  
## 2 11 may 220 1 339 4 failure no 5746800  
## 3 16 apr 185 1 330 1 failure no 1620000  
## 4 3 jun 199 4 -1 0 unknown no 1771200  
## 5 5 may 226 1 -1 0 unknown no 0  
## 6 23 feb 141 2 176 3 failure no 896400

## [1] "age" "job" "marital" "education" "default"   
## [6] "balance" "housing" "loan" "contact" "day"   
## [11] "month" "duration" "campaign" "pdays" "previous"   
## [16] "poutcome" "y" "balance\_kw"

#### ANSWER: new column; balance\_kw added with values(balance x 1200)

### Question 3. How many clients have subscribed a term deposit? In “y” field, what is the proportion of “yes” to all clients in the data?

## [1] 521

## [1] 0.11524

#### ANSWER: 521 clients have subscribed a term deposit. the proportion of ‘yes’ to all clients in the data is 0.11524(approximately 11.52%)

### Question 4. In “pdays” field, “-1” value means “the client was not previously contacted”, change the value “-1” to NA value in the field. Find the how many NAs the field has.

## [1] -1 339 330 -1 -1 176

## [1] NA 339 330 NA NA 176

## [1] 3705

#### ANSWER: chaneged the value ‘-1’ to NA in the pday field. There are 3705 NAs in the field.

### Question 5. Count the numbers of clients for each job type.

## admin. blue-collar entrepreneur housemaid management   
## 478 946 168 112 969   
## retired self-employed services student technician   
## 230 183 417 84 768   
## unemployed unknown   
## 128 38

#### ANSWER: admin. : 478, blue-collar : 946, entrepreneur: 168, housemaid: 112, management: 969, retired: 230, self-employed: 183, services: 417, student: 84, technician: 768, unemployed: 128, unknown : 38

### Question 6. Add new field “age\_group” that represents categorical age groups “under 20”, “20~29”, “30~39”, “40~49”, “50~59”, “over 60”. Which age group has the largest number of clients?

## Var1 Freq  
## 1 <20 4  
## 2 20-29 478  
## 3 30-39 1808  
## 4 40-49 1203  
## 5 50-59 854  
## 6 60+ 174

## [1] 30-39  
## Levels: <20 20-29 30-39 40-49 50-59 60+

#### ANSWER: new field; ‘age\_group’ added. the age group with the largest number of clients is ‘30-39’; the number is 1808.

### Question 7. From the “age\_group” field, calculate campaign success rate for each age group (the portion of “yes” in “y” field). Which age group has the highest success rate?

## <20 20-29 30-39 40-49 50-59 60+   
## 0.50000000 0.15062762 0.10232301 0.10224439 0.09836066 0.31609195

## [1] "<20"

#### ANSWER: ‘<20’: 0.50000000, ‘20-29’: 0.15062762, ‘30-39’: 0.10232301, ‘40-49’: 0.10224439,‘50-59’: 0.09836066, ‘60+’: 0.31609195. ‘<20’ group has the highest rate at 0.5.

### Question 8. Calculate average contact duration (“duration” field) for each contact type (“contact” field).

## [1] 267.1126

## [1] 243.3555

## [1] 261.753

#### ANSWER: ‘cellular’: 267.1126,‘telephone’: 243.3555,‘unknown’: 261.753

### Question 9. Sort the data in ascending order of client age.

## age job marital education default balance housing loan contact  
## 504 19 student single primary no 103 no no cellular  
## 1901 19 student single unknown no 0 no no cellular  
## 2781 19 student single secondary no 302 no no cellular  
## 3234 19 student single unknown no 1169 no no cellular  
## 14 20 student single secondary no 502 no no cellular  
## 1000 20 student single secondary no 291 no no telephone  
## day month duration campaign pdays previous poutcome y balance\_kw  
## 504 10 jul 104 2 NA 0 unknown yes 123600  
## 1901 11 feb 123 3 NA 0 unknown no 0  
## 2781 16 jul 205 1 NA 0 unknown yes 362400  
## 3234 6 feb 463 18 NA 0 unknown no 1402800  
## 14 30 apr 261 1 NA 0 unknown yes 602400  
## 1000 11 may 172 5 371 5 failure no 349200  
## age\_group2  
## 504 <20  
## 1901 <20  
## 2781 <20  
## 3234 <20  
## 14 20-29  
## 1000 20-29

#### ANSWER: sorted in acsending order of client age.

### Question 10. Save the data.frame that you have worked through this homework as “.RData” file and submit the file to Edmodo as well as R script file and reporting document.