Data Science

Homework 1

"bank_hw.csv" file contains a data related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

(Korean translation) "bank_hw.csv" 파일은 포르투갈의 어떤 은행의 마케팅 캠페인 결과를 담고 있습니다. 마케팅 캠페인은 전화로 이루어졌으며 정기 예금 상품에 가입하도록 권유 하기 위한 것입니다. 필요에 따라 한 고객에 대해서 여러 번 연락이 되기도 하였습니다.

- Attribute information: (각 변수 설명)

Input variables:

bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

17 - y - has the client subscribed a term deposit? (binary: "yes","no")

Write a R script code to answer following questions. You should submit your R script file with a document that has your answers (and explanation of how you derived the answers) to the Edmodo course page.

(Korean translation) 다음 질문들에 대답하기 위한 R 스크립트 코드를 작성하시오. 여러분은 R 스크립트 코드와 답을 작성한 보고서를 같이 제출해야 합니다. 제출은 에드모도에 해주세요.

Question 1.

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

(Korean translation) 데이터에는 몇 명의 고객에 대한 정보가 담겨 있나요? 나이가 30보다 어린 고객들의 수와 50보다 많은 고객들의 수는 각각 어떻게 되나요?

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

(Korean translation) balance 변수는 고객의 잔고 정보를 유로 단위로 표시하고 있습니다. balance_kw라는 새로운 변수를 데이터에 추가해주세요. 한국(원) 단위로 변환한 값을 가지고 있어야 합니다. 환율은 1유로에 1200원이라고 가정합니다.

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?

(Korean translation) 얼마나 많은 고객들이 정기 예금 상품에 가입하였나요? 전체 고객 대비 상품 가입에 yes로 응답한 비율은 어떻게 되나요?

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.

(Korean translation) pdays 변수의 -1의 값은 해당 고객에게 이전에 연락한 적이 없다는 것을 의미합니다. 이 -1값을 NA값으로 바꾸어 보세요. 얼마나 많은 NA가 pdays 변수에 있나요?

Question 5.

Count the numbers of clients for each job type.

(Korean translation) 각 직업 군에 속한 고객들의 수는 어떻게 되나요?

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?

(Korean translation) 새로운 변수 "age_group"을 추가하시오. 가장 많은 고객이 속한 연령대는 어떤 연령대인가요?

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?

(Korean translation) "age_group" 변수를 참고하여, 연령대 별 정기 예금 가입 비율을 계산하시오. 정기 예금 가입 비율(마케팅 성공율)이 가장 높은 연령대는 어디인지 말해보시오.

Question 8.

Calculate average contact duration ("duration" field) for each contact type ("contact" field).

(Korean translation) contact type ("contact" field) 별로 평균 contact duration ("duration" field)를 계산하시오.

Question 9.

Sort the data in ascending order of client age.

(Korean translation) 데이터를 고객의 나이의 오름차순으로 정렬하시오.

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.

(Korean translation) 이제까지 작업한 data.frame을 RData 파일로 저장하여 R 스크립트와 보고서 파일과 함께 제출하시오.

(가급적 압축하지 말고 제출하세요)