Homework 2

"**homework2_2020.RData**" contains the dataset you need for this homework. After loading this file into R, perform the following tasks. Submit your homework, with **reporting document** (containing all the answers and explanations, can be word, PDF, PPT or any of your preferred format), **R script file** containing your r code, and **RData file** that you are requested to submit.

"**homework2_2020.RData**" 파일은 이번 중간고사에서 사용하게 될 dataset을 저장하고 있는 RData 파일입니다. 이 파일을 R에 loading한 후 아래 task를 수행하시오. **R script파일**과 문제의 답 및 실행 결과가 포함된 **보고서** (word, 한글, ppt, pdf 형식 무관), 아래 문제에서 요구하는 **RData** 파일을 함께 edmodo에 제출하시오.

1. Data Exploration and Cleaning

Suppose that you are working for a health insurance company who are trying to promote a new product and increase the number of insurant for your company by analyzing customer information. "**cust.df**" is a data frame that contains customer information of your company. Attributes are described as follows

당신이 일하고 있는 (미국)보험회사에서는 최근 출시된 보험상품을 고객에게 홍보하고 가입을 유도하기 위하여 여러분이 갖고 있는 고객 정보를 분석하기로 했다. **"cust.df"** data frame은 회사가 가지고 있는 고객들의 다양한 정보를 포함한 데이터이다. 데이터가 가지고 있는 고객 정보는 아래와 같다.

state.of.res: state of residence

custid: id of customer

sex

is.employed: is this customer employed

Income: yearly income in USD

marital.stat: marital state

health.ins: is this customer insured by other insurance product

housing.type

recent.move: has this customer recently moved ?

num.vehicles: the number of vehicles this customer has

age

1. Describe the structure of the dataset. How many observations and variables does it have? What are the types of the variables?

   데이터의 구조를 파악하시오. 데이터의 observation의 수와 variable의 수는 어떻게 되나요? 각 variable의 type은 어떠한가요?

2. Examine if all the types of variables are prospered set. If there are any variable not properly typed. Explain why you think so, and convert the type into proper one.

   각 variable의 type이 적절한지 확인하시오. 만일 적절하지 않은 type이 있다면 **이유를 설명**하고, **적절한 type으로 변경**하시오.

3. for the variable **custid**, it is numeric variable that is typically treated as numbers, but we want it to be in character type. When converting them, we want to make the id in combination of alphabet letter and numbers in length 8 in letters. For example, if the id was **"7339"**, then we want it to be "**c0007339**" so all the IDs start with alphabet (small) letter "**c**" and the empty positions are padded with "0" (zeros). Convert the variable "**custid**" as described.

   custid의 경우는 고객 id로서 숫자로 표현되어 있으나 character type으로 변경하고 싶다고 한다. 그리고 character type으로 변경할 때, id는 총 8자리 글자와 숫자의 조합으로 표시하고 싶다. 예를 들어 고객 id가 원래 **"7339"**였다면 **"c0007339"**와 같이 첫 글자는 **"c"**를 넣고 나머지 비어있는 앞자리는 **"0"**으로 채워 넣고자 한다. 이와 같이 고객 id 변수를 변경하시오.

4. Examine if there are any missing values in each variable, how may missing values are in each variable, and what are the ratio of missing value for each variable.

   각 변수 별로 missing value가 존재하는지, 존재한다면 몇 개나 존재하며 비율은 어떻게 되는지 설명하시오.

5. Assume that for a set of variables**(A)** having missing values less than 10%, we are considering to remove observations with missing values. Examine whether the variables**(A)**' values are missing on the same customers or different set of customers. How many

customers' valuesare missing for those variables**(A)**?

Missing Value의 비율이 10%가 넘지 않는 변수들(A)에 대해서, Missing Value를 포함하는 행(고객)을 삭제하는 것을 고려할 수 있다고 한다. Missing Value가 존재하지만 비율이 10%가 넘지 않는 변수들**(A)**에 대해서 동일한 고객에게 동시에 Missing Value가 발생하였는지 아니면 서로 다른 고객에게 Missing Value가 발생하였는지 확인하시오. 이 변수들**(A)**에 대해서 Missing Value를 갖고 있는 고객의 수는 총 몇 명인가?

6. For the variable **is.employed**, we may treat missing values as valid values by having character value "**missing**" instead of **NA**s. (You might need to change the type of the variable which was logical). Convert value **TRUE** into "**employed**", **FALSE** into "**not employed**", and **NA**s into "**missing**" for **is.employed** variable.

   is.employed 변수의 경우는 존재하는 missing value들을 "missing" 이라는 새로운 category를 하나 추가해서 missing value를 의미 있는 데이터로 취급할 수 있을 것이다. (이를 위해서는 **is.employed**는 logical vector이기 때문에, 다른 type으로 변환해야 할 것이다.) **is.employed**의 TRUE 값은 "employed", FALSE 값은 "not employed", NA 값은 "missing"으로 변경하여 저장하여라.

7. Find the average and median income for each group of residential state. Make a new data frame named "**avg_income**" that has 3 columns of "**state.of.res", "median.income", "mean.income"**.

   거주 state 별로 income의 중간값과 평균값을 계산하시오. 계산 결과를 새로운 data.frame "**avg_income**"으로 만드시오. "**avg_income**"는 3개의 column을 가지며 column의 이름은 순서대로 "**state.of.res", "median.income", "mean.income"** 이다.

8. Change the missing value of **Income** variable into average income of the state where the corresponding customer lives.

   Income 변수에 존재하는 Missing Value들을 해당 고객이 거주하는 state의 평균 연봉 mean.income으로 치환하시오.

9. It is known that there is large income disparity among states and cities in US. Hence we might want to know **relative income** among their state of residence instead of absolute value of income in USD. Find the value of relative income compared to their median income of state of residence and make it as a new variable "**income.relative**". For example, if there is a person who live in Alabama and whose income exactly in the middle of all people in Alabama, there the value of relative income would be 1.0. If a certain person earns 50% more yearly than the median income of his state of residence, his relative income will be 1.5.

미국은 각 주와 도시 별로 소득 수준의 격차가 크다고 알려져있다. 그렇다면 고객의 income의 절대값을 아는 것도 중요하지만 그보다 그 고객이 사는 지역의 소득 수준과 비교하여 상대적으로 income이 많은지 적은지를 아는 것이 더 유용할 수 있을 것이다. 고객이 거주하고 있는 주(state)의 소득의 중간 값(median)과 비교한 비율을 계산하여 "**income.relative**" 라는 변수로 추가하여라. 예를 들어 Alabama 주에 사는 고객이 Alabama 주의 소득의 중간 값 만큼 정확히 번다면 이 변수의 값은 1.0이 될 것이며, 중간값의 1.5배 만큼 번다면 이 변수의 값은 1.5가 될 것이다.

10. Examine if there are any outliers in the dataset (at least 2). Explain why you think they are outliers.

    Data에 Outlier가 존재하는지 확인하시오 (최소 2개 이상). 왜 그 값들이 outlier인지 설명하시오.

11. Explain whether the outliers you found in Question 10 are obvious errors, valid values, or something you have to clarify with. Describe your answer with explanation.

    11번에서 발견한 outlier가 명백한 오류인지, 정상적인 값인지, 확인이 필요한지 대답하고 그 이유를 설명하시오.

2. Tidy Data

"**bankruptcy_df**" is a data frame that shows various information of 250 companies to study business bankruptcy. Some went bankrupt and some did not, among 250 companies. Followings are attributes described in the dataset.

"**bankruptcy_df**" data.frame은 기업의 도산(bankcruptcy) 여부를 예측하기 위해서, 250개 기업의 여러 지표들을 표현한 데이터이다. 표현된 기업의 지표들은 아래와 같다.

1 Industry risk (IR) :

Government policies and International agreements,

Cyclicality,

Degree of competition,

The price and stability of market supply,

The size and growth of market demand,

The sensitivity to changes in macroeconomic factors,

Domestic and international competitive power,

Product Life Cycle.

2 Management risk(MR):

Ability and competence of management,

Stability of management,

The relationship between management/ owner,

Human resources management,

Growth process/business performance,

Short and long term business planning,

achievement and feasibility.

3 Financial Flexibility(FF):

Direct financing,

Indirect financing,

Other financing

4 Credibility (CR):

Credit history,

reliability of information,

The relationship with financial institutes.

5 Competitiveness (CO):

Market position,

The level of core capacities,

Differentiated strategy,

6 Operating Risk (OP):

The stability and diversity of procurement,

The stability of transaction,

The efficiency of production,

The prospects for demand for product and service,

Sales diversification,

Sales price and settlement condition,

Collection of A/R,

Effectiveness of sale network.

7. Class: NB = Non-Bankruptcy, B = Bankruptcy

1. Explain whether the "**bankruptcy_df**" data frame is tidy or not and describe the reason of your answer.

   "**bankruptcy_df**" 데이터가 tidy data인지 아닌지 답하고, 그 이유를 설명하시오.

2. If the answer of Question 1 is "not tidy", then transform the dataset into tidy one. Submit the transformed (tidy) data frame in a "RData" file. (use save() command in R)

   1번의 답이 tidy가 아니라고 한다면, tidy 형태로 변환하시오. 결과는 RData에 저장(save 명령 사용)하여 제출하시오.

3. The data value is represented in abbreviation as described follows. To make the data more understandable and clear to human, convert the abbreviations into full name (full text).

   각 지표의 값들이 P, A, N, B, NB와 같이 약어로 표시되어있다. 데이터의 가독성을 높이고 의미를 명확히 하기 위해서 약어를 모두 아래와 같이 변환하시오.

   A. P: Positive

   B. A: Average

   C. N: negative

   D. B: Bankruptcy

   E. NB: Non-Bankruptcy