Midterm test

Introduction to Data Science

2018-04-18

7:00 ~ 10:00 PM

Answer the following questions. Put your answers in a document file (MS word, HWP, PDF, or any other text format). Submit the document with R script file and RData file. RData file should include the two result data frames of following questions. You can choose to answer in your proficient language from English and Korean. All the answers should be based on findings from the given datasets.

The total score of this exam is 50 + 15 + 70 = 135 pts. However you can get up to 100 pts which is full credit for this exam. If you get more than 100 pts, extra points will be summed up and distributed to your classmates equally. Literally this is chance to work hard to give.

You must submit your work to Edmodo in time (~10:00PM). The submission page will be locked after 10PM. Only the work that are submitted to Edmodo successfully in time will be graded. You can submit your work several times during the exam (7PM ~ 10PM) if necessary.

(Korean Translation) 아래의 질문에 대한 답을 문서파일로 작성(MS word, 한글파일, PDF 또는 원하는 형식 가능)하고 답을 찾기 위해 사용한 R script 파일과 함께 edmodo에 제출하세요. 결과로 얻게 되는 data frame(2개)는 RData 파일에 담아 함께 제출하시오. 답은 한글과 영어 중 편한 언어로 작성하면 됩니다. 모든 답은 데이터로부터 발견한 사실에 근거해야합니다.

총점은 135점이지만, 만점은 100점으로 100점까지 본인이 받을 수 있습니다. 100점을 넘게 받는다면 나머지 점수는 합산되어 반 동료들에게 똑같이 나누어지게 됩니다. "공부해서 남주자"를 실천해봅시다.

(주의) 10시까지가 시험이고 이 시간을 지나면 (10시 정각이 되면) 답안을 제출하지 못하게 됩니다. 답안이 제출되지 않은 경우 자동 0점 처리되고, 어떤 예외도 허용되지 않기 때문에 제시간에 제출하는 것을 주의하세요. 한번 제출한 이후에 시간 내라면 여러 번 제출하는 것은 가능합니다.

Basic R Skills (5 pts each)

"**Car04.csv**" file contains information of various car models. The dataset records characteristics on all of the new models of cars for sale in the US in a certain year.

(Korean Translation) "**Car04.csv**"파일은 미국에서 특정 연도에 판매된 여러 자동차 모델에 대한 정보를 담고 있습니다.

1. Load the data in the file into R as a data.frame named "**car_df**"

   (Korean Translation) 파일로부터 데이터를 R로 읽어오시오. 데이터 프레임 이름은 **car_df**

2. How many variables are in the dataset? And how many observations are there?

   (Korean Translation) 변수는 몇 개며 observation은 몇 개 입니까?

3. The "**name**" column represents the name of car models. What would be the proper data type of "**name**", character or factor? Convert the type into proper one if necessary.

   (Korean Translation) name columns은 자동차 모델의 이름을 담고 있습니다. character와 factor type 중 어떤 것이 적절할까요? 필요하다면 type 변환을 하시오.

4. "**msrp**" means Manufacturer Suggested Retail Price. Calculate average difference between **msrp** and **dealer_cost**.

   (Korean Translation) mrsp는 소비자 권장 가격입니다. dealer_cost와 비교해서 평균적으로 차이가 얼마나 나나요?

5. Find the car models with highest **city_mpg**. What is the difference of **city_mpg** and **hwy_mpg** for the car? Is it the same car with the car of highest **hwy_mpg**?

   (Korean Translation) city_mpg값이 가장 큰 자동차 모델은 무엇인가요? 그 차의 city_mpg와 hwy_mpg의 차이는 얼마인가요? 이 차는 hwy_mpg가 가장 높은 차와 같은 차종인가요?

6. How many cars models are in each car type of (**sport car, suv, wagon, minivan, pickup**)? How many cars are in none of these car types?

   (Korean Translation) 각 각의 자동차 종류(**sport car, suv, wagon, minivan, pickup**)마다 자동차 모델이 몇 개씩 있나요? 어떤 종류에도 속하지 않는 자동차는 몇 개나 있나요?

7. Compare the weight of **suv** and **minivan**. Which car type is heavier on average?

   (Korean Translation) SUV와 minivan 차종의 무게를 비교하시오. 평균적으로 어떤 차종이 더 무겁나요?

8. Add new column "**avg_mpg**" which represent average mpg value of **city_mpg** and **hwy_mpg**.

   (Korean Translation) 새로운 column "avg_mpg"를 추가하시오. avg_mpg는 city_mpg와 hwy_mpg의 평균입니다.

9. Add new column of "**eco_grade**" that has value of "**good**" for cars with high 20% avg_mpg, "**bad**" for ones with low 20% mpg, and "**normal**" for the rest.

   (Korean Translation) "eco_grade"라는 columns을 추가하시오. avg_mpg 상위 20%에는 "good" 하위 20%에는 "bad", 나머지에는 "normal" 값을 부여하시오.

10. Compare horse power of all wheel drive cars and rear wheel drive cars.

    (Korean Translation) 4륜 구동 자동차와 후륜구동 자동차의 마력을 비교하시오.

Finally, include the result data frame into an RData file.

(Korean Translation) 결과 데이터 프레임을 RData 파일에 포함시키시오.

Data Science Project Pipeline (5pts each)

1. Why do you think soft skills such as **communication** and **teamwork** is as important as other hard skills especially for "data scientist"?

   (Korean Translation) 데이터 과학자에게 **communication**과 **teamwork** 능력이 중요한 이유는 무엇인가요?

2. Explain about process of "**Data Exploration**". Why is it necessary?

   (Korean Translation) 데이터 탐색 단계에 대해 설명하시오. 그것이 왜 필요한가요?

3. Why do we need to **clean the data** given by Client?

   (Korean Translation) client에게서 받은 data를 cleaning 해야하는 이유는 무엇입니까?

Data Cleaning

Load **"midtermdata.RData"** into R. "**weather**" variable is a data frame that contains Historical weather information from Boston, USA collected for 12 months beginning Dec 2014. Answer the

following questions.

(Korean Translation) **"midtermdata.RData"**파일을 읽어오시오(load). weather 변수는 미국 보스턴에서 2014년 12월부터 12개월간 측정된 날씨 정보를 담고 있는 data frame이다. 다음 질문에 답하라.

1. [Data Exploration] Do you find the dataset is **tidy**? Explain why it is so. (5pts)

   (Korean Translation) dataset은 tidy인가 아닌가? 이유를 설명하시오.

2. [Data Transformation] If you believe the dataset is not **tidy**, transform it into **tidy** one. (15pts)

   (Korean Translation) dataset이 tidy가 아니라면 tidy한 형태로 변환하시오.

3. Is there any unnecessary column in the dataset? If there is, what is it? Remove the unnecessary column from data frame. (5pts)

   (Korean Translation) 데이터셋에 불필요한 column이 있는가? 있다면 무엇인가? 불필요한 column을 제거하시오.

4. There are **year**, **month**, and **day** column in the dataset. Combine these three column together to make a new column named "**date**" which is in **Date** data type. And remove the columns of **year**, **month**, and **day**. (10pts)

   (Korean Translation) 데이터에 year month day 세 column이 있는데 이를 하나로 합쳐서 date column을 추가하시오. date column은 Date data type이어야합니다. 그리고 year month day 세 column은 제거하시오.

5. Look at the variable "**PrecipitationIn**", there are several character value **"T"**, denoting a ***trace*** amount (i.e. too small to be accurately measured) of precipitation. To have this variable as numeric one, change all "T" to number zero. (5pts)

   PrecipitationIn(강수량) 변수를 보면 "T"라는 값이 있는데 이는 Trace 비가 아주 미량왔다는 의미이다. 해당 변수를 숫자형으로 변환할 수 있도록, "T"를 숫자 0으로 변환하시오.

6. Convert the data type of each variable into proper data type. (5pts)

   (Korean Translation) 각 변수의 data type을 적절한 것으로 변환하시오.

7. [Missing Values] Does the dataset contains any missing values? How may are they in the dataset? How many missing values are in each variable? (5pts)

   (Korean Translation) 데이터셋에 missing values가 있나요? 몇 개나 있나요? 각 변수 별로 몇 개씩 있나요?

8. [Outliers] Look at the variable **Max.Humidity**. Is there any outlier (extreme value) in the variable? Assuming that one more "0" was added accidently for the outlier, correct the outlier into proper value. (5pts)

   (Korean Translation) Max.Humidity(최대 습도) 변수를 보시오. outlier가 있나요? outlier 값이 실수로 0이 하나 더 붙어 나온 값이라고 합시다. 해당 outlier를 적절한 값으로 고치시오.

9. [Outliers] Look at the variable **Mean.VisibilityMiles**. Is there any outlier (extreme value) in the variable? Correct the outlier into proper value. (5pts)

   (Korean Translation) Mean.VisibilityMiles(평균시야거리) 변수를 보시오. outlier가 있나요? outlier를 적절한 값으로 고치시오.

10. The **Events** variable contains an **empty string ("")** for any day on which there was no significant weather event such as rain, fog, a thunderstorm, etc. However, if it's the first time you're seeing these data, it may not be obvious that this is the case, so it's best for us to be explicit and replace the empty strings with something more meaningful. Convert the **empty string** into "**None**". (5pts)

    (Korean Translation) Event변수를 보면 공백문자 ""가 포함되어있습니다. 비나 안개 같은 특별한 event가 없는 날이라는 표시인데, 더욱 명백하게 표현하는 것이 좋습니다. 공백문자를 "None"으로 바꾸시오.

11. For the column names of data frame, we prefer to have it in all lower-case letters. So we do not have to remember which letters are uppercase or lowercase. Convert all column names of data frame into lower case letters (5pts)

    (Korean Translation) data frame의 column name은 모두 소문자로 하는 것이 좋습니다. 나중에 대문자인지 소문자인지 기억하지 않아도 되기 때문입니다. data frame에서 column name을 모두 소문자로 바꾸시오.

    Include the result data frame into an RData file.

    결과 데이터 프레임을 RData 파일에 포함시키시오.