Introduction to Data Science

Data Cleaning Practice

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

1. Transform the data into **tidy** one.

   (Korean Translation) dataset을 tidy 형태로 변환하시오.

2. Is there any unnecessary column in the dataset? If there is, what is it? Remove the unnecessary column from data frame.

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

3. 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**.

   (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.

4. 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.

   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.

5. Convert the data type of each variable into proper data type.

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

6. [Missing Values] Does the dataset contains any missing values? How may are they in the

dataset? How many missing values are in each variable?

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

7. [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.

   (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.

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

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

9. 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**".

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

10. 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

    (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.