Data Science

Homework 11

<Homework 11 Introduction>

Lastly, we're going to review each of the previous chapters one by one... to summarize!
As you study with your team, think logically and comprehensively about which outcome is most appropriate!
The process (logical development) will be largely reflected in the score. The evaluation will be based on the logicality of the better result + the fit of the model through question 11.
자. 사실상 데이터 과학의 꽃.. 여기를 위해 한학기를 달려왔다고 봐도 됩니다. 마지막 과제는 이전 챕터에 있는 내용을 하나씩 복습하면서.. 총 정리하는 느낌으로 진행해 볼까 합니다! 팀원들과 함께 공부하면서, 어떤 결과값이 가장 적절할지에 대해 논리적이고 종합적으로 생각해 보세요!
그 과정(논리전개)이 점수에 크게 반영될 예정입니다. 평가는 문제 11 을 통해 나온 결과값에 대한 논리성+모형의 적합도를 기준으로 진행됩니다.

실습 데이터: SeoulBikeData.csv

- Do not modify csv itself, but modify file name, type conversion, etc. after being imported from R studio
- 원칙: csv 자체를 수정하지 말고, R 스튜디오에서 불러온 이후에 파일명, 타입변환 등을 수행할 것!

데이터 설명 Data Description

SeoulBike Data contains information related to the use of sharing bike system in Seoul.
SeoulBikeData 는 서울시 따릉이 이용과 관련된 정보를 담고 있습니다.

```
> str(SeoulBike)
 'data.frame':    8760 obs of 14 variables:
 $ Date                      : chr  "01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017"...
 $ Rented.Bike.Count         : int  254 204 173 107 78 100 181 460 930 490 ...
 $ Hour                      : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Temperature               : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5
 $ Humidity.percent.         : int  37 38 39 40 36 37 35 38 37 27 ...
 $ wind.speed                : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
 $ visibility                : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1927 ...
 $ Dew.point.temperature     : num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
 $ Solar.Ratiation..MJ.m2.   : num  0 0 0 0 0 0 0 0 0.01 0.23 ...
 $ Rainfall.mm.              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Snowfall..cm.             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seasons                   : chr  "winter" "winter" "winter" "winter" ...
 $ Holiday                   : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
 $ Functioning.Day           : chr  "Yes" "Yes" "Yes" "Yes" ...
```

• 문제 1 (데이터 불러오기. 특성 파악하기): 위 데이터프레임을 R 스튜디오에 불러오고. 해당 데이터프레임의 특성이 어떠한지 파악해 보세요.
Question 1 (Importing data. Identify characteristics/structures): Import the above data frame to the R studio. Figure out the characteristics of the data frame.
• 문제 2 (변수 이름 수정하기): 변수 중 이름이 너무 길고 복잡한 변수들은 적절한 형태로 수정하여 주세요.
Question 2 (modifying variable names): Please modify the variables whose names are too long and complicated to the appropriate form.
• 문제 3 (데이터 타입 바꾸기): 여러 변수 중 데이터 타입의 변환이 필요한 케이스가 몇 가지 있습니다. Date 변수를 chr 타입에서 날짜 타입으로 변환시키고, Holiday 변수는 더미로 변환하여 주세요.
Question 3 (Change data type): There are several cases that require data type conversion among variables.
Please convert “Date” variable from chr type to date type and convert “Holiday” variable to dummy.
• 문제 4 (실전연습): 총 13 개의 변수 중에 필요 없어 보이는 변수는 무엇인가요? 그리고 그 이유는 무엇인가요?
Question 4 (practice): out of a total of 13 variables, which one does not seem necessary and why?
• 문제 5 (인덱스 활용): 문제 4 에서 필요 없다고 판단된 변수를 제외하여 다시 데이터 프레임을 구성하여 주세요.
Question 5 (index): Please reconfigure the data frame except for the variables that are considered unnecessary in Question 4.

Now, you are going to perform a multilinear regression analysis related to the use of Sharing bike system in Seoul.
If the research problem is to figure out which factors influence the use of Sharing bike system in Seoul, which variable is appropriate as the dependent variable?

here, Rented.Bike.Count can be used as a dependent variable, right?
Then let's make a research hypothesis now.

자, 이제 여러분은 서울시 따릉이 이용과 관련하여 다중선형회귀분석을 수행해 보려고 합니다. 서울시 따릉이 이용에 어떤 요인들이 영향을 미치는지 파악하는 것이 연구 문제라면, 어떤 변수가 종속변수로 적절할까요?
그렇죠.. 여기서는 Rented.Bike.Count (만약에 변수명 바꿨다면 여러분이 바꾼 변수명으로 쓰셔도 됩니다)가 종속변수로 사용이 되겠지요?
그렇다면 이제는 연구 가설을 세워봅시다.

- 문제 6 (가설 세우기) 따릉이 이용수에 영향을 미치는 요인들은 어떤것들이 있을까요? 그리고 각각 어떤 (정+ 부-)영향을 끼칠까요? 가설을 세워서 주석으로 달아주세요. (ex. *Rainfall 변수와는 부- 의관계를 보일것으로 예상된다. 왜냐하면 비가 오면 자전거 운행이 쉽지 않기 때문이다. 등등 변수별로 작성하세요.)

Question 6 (set Hypothesis) What are the factors that can affect the number of users of Sharing bike system? Will it affect positive or negative direction? Make a hypothesis and annotate it. (ex. *It is expected to show a negative relationship with the Rainfall variable and the number of uses of Sharing bike system, because it is not easy to drive a bicycle when it rains.)

- 문제 7 (lm 모델) 문제 6 에서 세운 가설을 확인해볼 차례입니다. 다중선형회귀분석을 수행해 보세요.

Question 7 (lm model) It's time to check out the hypothesis we established in problem 6. Try doing a multilinear regression analysis.

- 문제 8 (다중공선성확인) 위 모델에서 다중공선성을 확인해 보세요. HINT: package "car", vif)

Question 8 (Check multicollinearity) Check multicollinearity in the above model. HINT: package "car", vif)

- 문제 9 (성능확인, 시각화) 학습 모델의 RMSE 와 R^2값을 구해봅시다. 모델성능을 그래프로도 확인해봅시다 (hint: ggplot). Question 9 (Performance Check, Visualization) Let's find the RMSE and R^2 values of the learning model. Let's check the model performance with a graph (hint: ggplot).
- 문제 10 (중간평가) 문제 6~9 에서 도출해낸 결과에 대해 해석해봅시다. (변수별, 모델 전체 성능 모두) Question 10 (Interim Evaluation) Let's interpret the results derived from questions 6-9. (Both variable and model overall performance)
- 문제 11 (모델성능향상): 모델 성능 향상을 위해 어떤 아이디어를 적용할 수 있을까요? 우리조에서 시도한 방법을 상세하게 설명하고 "문제 6~9 과정을 반복하며 "최적의 결과를 도출해야 합니다.

Question 11 (Model performance improvement): What ideas can we apply to improve model performance? Explain in detail the method we tried in our group and repeat the steps 6 to 9 to "get the best results."