2024-Spring

Data Science
Mid-term Exam

[Data→ weather.rds]

1. (4 points) "weather.rds" 데이터를 R 에 loading 하고, 위 데이터 프레임이 대략적으로 어떠한 특징을 가지고 있는지 확인해 보시오. (본인이 생각하는 명령어를 사용하여 확인해 보고, 각 명령에 대해 왜 그런 과정을 수행했는지 주석으로 설명할 것)
   Load the "weather.rds" data into R and check roughly what characteristics the above data frame has. (Check using the commands you think of, and explain in comments why you performed that process for each command.)

2. (5 points) 위 데이터가 tidy 가 아닌 이유를 설명 하시오. 또한, 불필요한 column 을 제거하고 그 이유를 주석으로 설명하시오.
   Please explain why the above data is not tidy. Also, remove unnecessary columns and explain the reason in comments.

3. (5 points) 위 데이터를 tidy 한 형태로 변환하고, 그 과정을 주석으로 설명하시오.
   Convert the above data into tidy form and explain the process in comments.

[Data → bmi_clean.csv]

4. (4 points) bim.clean.csv 파일의 데이터는 1980 년부터 2008 년까지 측정된 각 국가 표분인구의 평균 bmi 를 나타낸다. 이 데이터는 tidy 한가? 만약 tidy 하지 않다면, tidy 한 형태로 변환하고 그 과정을 주석으로 설명하시오.
   The "bmi_clean.csv" file represent average bmi of sample population of each country measured in year for 1980 to 2008.
   Is this data tidy? If it is not tidy, convert it to tidy form and explain the process in comments.

[Data → students2.csv]

5. (2 points) students2.csv 파일을 R 에 로드하고, 구조(structure)를 확인하여라.
   Load data of students2.csv into R. Preview students2's structure.

6. (4 points) dob 열을 연-월-일 date 형식으로 변환하고, nurse_visit 열을 연-월-일-시-분-초 형식으로 변환하여라.
   Coerce dob to a date (with no time) and Coerce nurse_visit to a date and time.

7. (2 points) 변환된 데이터의 구조(structure)를 확인하고, 어떤 변화가 있는지 주석으로 설명하시오.
   See the changes to students2, and explain changes in comments.

[Data → winequality-red.csv]

8. (3 points) 위 데이터를 R 에 데이터 프레임으로 loading 하시오.
   Load the above data into R as a data frame.

9. (2 points) 위 데이터프레임 에서 총 몇 개의 관측치와 변수가 있는지 확인하시오.
   Check how many observations and variables there are in total in the data frame above.

10. (5points) volatile.acidity 변수와 fixed.acidity 변수의 합으로 연산되는 total.acidity 열을 생성하여 데이터프레임에 추가하시오.
    Create a total.acidity column that is calculated as the sum of the volatile.acidity variable and the fixed.acidity variable and add it to the data frame.

11. (4 points) 본 데이터프레임에는 결측치가 있는가? 확인해 보아라.
    Are there any missing values in this data frame? Check it out.

12. (4 points) 본 데이터프레임에서 quality 변수의 분포를 확인하기 위한 히스토그램을 그려보시오.
    Draw a histogram to check the distribution of the quality variable in this data frame.

13. (4 points) quality 변수에 outlier 가 있는가? outlier 를 확인하기 위해 필요한 그래프를 그려 보고, 주석으로 설명하시오.
    Are there outliers in the quality variable? Draw the graph needed to check the outlier and explain it with comments.

14. (3 points) quality 변수를 기준으로 데이터셋을 내림차순으로 정렬하는 코드를 작성하시오.
    Write code to sort the dataset in descending order based on the quality variable.

15. (3 points) "sample" 함수를 사용하여 임의로 30 개의 데이터만 추출하고, 추출한 데이터들의 quality, density, pH, alcohol 의 평균을 각각 계산하시오.
    Use the "sample" function to randomly extract only 30 pieces of data, and calculate the average of quality, density, pH, and alcohol of the extracted data.

16. (6 points) pH 와 alcohol 변수 사이에 어떤 관계가 있는가? 데이터로 추론할 수 있는 내용을 주석으로 설명하시오. (둘 사이의 관계를 확인할 수 있는 그래프도 그려 보시오.)
    Is there any relationship between pH and alcohol variables? Explain what can be inferred from the data. (Please also draw a graph to confirm the relationship between the two.)

[Data → 'https://github.com/hbchoi/SampleData/raw/master/adult.RData']

17. (4 points) 위 데이터를 R 에 load 하고, train 데이터셋(70%)과 test 데이터셋(30%)으로 분류하시오.
    Load the above data into R and classify it into train dataset (70%) and test dataset (30%).

18. (6 points) 위 데이터를 사용하여 "income_mt_50k"여부를 판단하는 single variable classification model 을 만들고자 한다. "education"를 통해 "income_mt_50k"를 판별하는 모델을 만들고, 정확도를 구해 보아라. (threshold setting 은 본인 마음대로 하면 됩니다.)
    Using the above data, we want to create a single variable classification model that determines whether it is "income_mt_50k". Create a model that determines "income_mt_50k" through "education" and find the accuracy. (You can set the threshold as you wish.)

---------- (여기서부터는 오픈북 금지! 주관식 문제입니다. 답안은 주석으로 작성하시오.)----------

No open books from here on out! Please write your answers in comments.

19. (6points) 나이브 베이즈 모델과 정확한 베이즈 모델의 차이에 대해 아는 대로 설명하시오.
Explain to the best of your knowledge the difference between a naive Bayes model and an exact Bayes model.

20. (5points) single variable vs multiple variable 모델과 classification vs regression 모델에 대해 아는 대로 설명하시오.
Explain to the best of your knowledge the single variable vs multiple variable model and classification vs regression model.

21. (5points) 모델 분석시에 train 데이터와 test 데이터를 나누는 이유는 무엇인가?
Why are train data and test data divided when analyzing a model?

22. (7points) 분석 모델의 정확도(accuracy)를 높이기 위해 시도할 수 있는 방법은 어떤 것 들이 있는지 설명하시오.
Please explain what methods can be tried to increase the accuracy of the analysis model.

23. (7points) Overfitting 현상에 대해 아는 대로 설명하시오.
Explain the overfitting to the best of your knowledge.

고생 많으셨습니다. 😊
Thank you for your hard work. 😊

(실습실 컴퓨터를 사용한 1 분반 학생들은 문제지, 답안지, 실습파일을 사용 컴퓨터에서 삭제하고 퇴장하여 주시기 바랍니다.)