Midterm 23/24 Sem 1

Introduction to Data Science (National University of Singapore)

NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 1 AY2023/2024)

Midterm Test

Time Allowed: 80 minutes

INSTRUCTIONS TO STUDENTS

1. Students are required to complete this test individually.
2. **Your submission should have only one file (.R) which includes the R code and your interpretation/analysis as comments**. Make sure that there is no error when the graders open and run your code.
3. Be sure to lay out systematically the various parts and steps in your code file.
4. At the end of the test, label your answer file by your student number (such as A0123456B.R), and upload it to Canvas/DSA1101/Assignments/Midterm Submission.

1. (*50 points*) The data file `data-midterm.csv` is extracted from an original dataset which was obtained from the WHO and United Nations website. The given file contains information on 89 countries in the year 2014.
We consider the the following variables:

Variable	Description
Status	status of a country (Developed, Developing)
Life_expectancy	life expectancy (year)
Adult_mortality	adult mortality rate
infant_deaths	number of infant deaths for that country
Alcohol	alcohol consumption (per capita (15+) consumption, in liters of pure alcohol)

For the questions below,

- **Status** is considered as the response variable.
- all four features are used to form models/classifiers.
- use `set.seed(1101)`.
- please report numerical answers to three significant figures if it's smaller than one and to three decimal places if it's larger than one.

Part I: Data Preparation

1. (2 points) Load the dataset into R and name it as **data**. Write code to remove the first three columns of the dataset which we will not use for any questions below.
2. (2 points) Report the name of five columns in **data**, in order as they appear in **data**.
3. (2 points) Write code to change the names of the four variables in **data**, Life_expectancy, Adult_mortality, infant_deaths and Alcohol, into **X1**, **X2**, **X3** and **X4**, respectively.

   Hint: After the step of name changing above, when you run `head(data)`, the output should show five columns with the names **Status**, **X1**, **X2**, **X3** and **X4**.

Part II: Data Exploration

4. (4 points) Report the number of developed countries and its proportion in the data given.
5. (6 points) Create a histogram for the sample of life expectancy (X1) with a red color normal density curve overlaying. Give your comment about this plot.
6. (2 points) Create a box plot for the sample of life expectancy (X1). Does it show any outliers? If yes, retrieve and report the full information of those outliers.
7. (4 points) Create a QQ plot for the sample of life expectancy (X1). Give your comment about this plot.

Part III: Linear Model

8. (2 points) Write code to transform the variable **Status** to numerical format where `Developing` = 0 and `Developed` = 1.

9. (6 points) Consider **Status** as a quantitative variable with values of 0 and 1. Fit a linear regression model for it, called M1, using all four features **X1**, **X2**, **X3** and **X4**. Report the regressors that are significant in the model.

10. (4 points) Report R^2 of model M1. Give your comments on the goodness-of-fit of M1.

11. (2 points) How many fitted values of model M1 are less than 0?

12. (4 points) A country has information listed below. Write code to predict the status of this country, using the rule of majority vote.

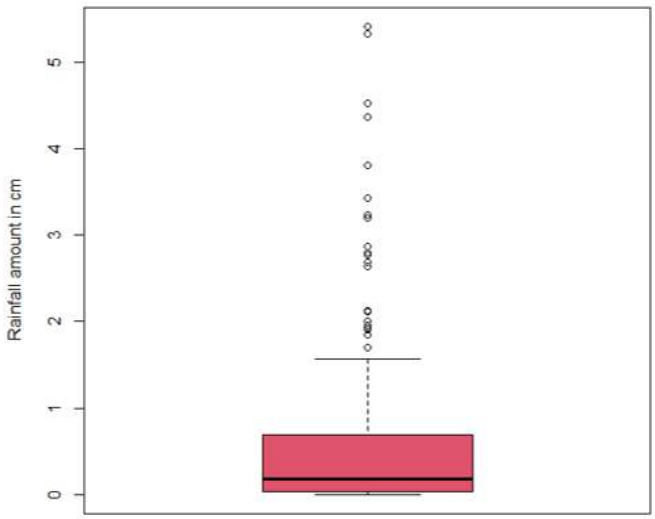
    `Life_expectancy` = 83, `Adult_mortality` = 57, `infant_deaths` = 2, and `Alcohol` = 3.

Part IV: KNN

13. (6 points) Use all the observations and standardized features to form the best k-NN classifier where $2 \leq k \leq 10$, based on accuracy. Report the best k found and the accuracy of the k-NN classifier with that k (called M2).

14. (4 points) With the country listed in Question 12, use the classifier M2 above to predict the status of this country.

2. (*10 points*) In a study of the natural variability of rainfall, the amount of rainfall (in cm) from each of 227 storms in Illinois (US) for the years 1960-1964 was collected. The boxplot of the rainfall sample is given in the figure below.



Choose the only one correct answer for the statements given below.
You should type your answer in the file of R code as comments. No explanation is required.

1. (3 points) The distribution of the sample of rainfall amount is

   a symmetric.

   b right skewed.

   c left skewed.

   d not determinable.

2. (3 points) For this sample, when comparing the mean to the median,

   a the mean is about the same as the median

   b the mean is much larger than the median

   c the mean is much smaller than the median

   d it's not determinable which one is larger.

3. (4 points) One would want to fit a linear model with the aim is to predict the rainfall amount.

   a It's not good to fit a linear model where rainfall amount is the response because the sample of rainfall does not satisfy the assumption for the response of a linear model.

   b It's good to fit a linear model where rainfall amount is the response because it is a quantitative variable.

   c It's good to fit a linear model where rainfall amount is the response as long as it has linear relationship with the regressors in the dataset.

   d Both (b) and (c) are correct.

END OF QUESTIONS