

NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 1 AY2024/2025)

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 file.
3. Be sure to lay out systematically the various parts and steps in your code file.
4. At the end of the test, please upload the answer file to CANVAS/DSA1101/Assignments/Midterm Submission.

1. (50 points) The data file `patient_satisfaction.csv` is a random sample collected from a hospital about the satisfactory of patients about the hospital's service when they were discharged. The data set includes information about patient's age, the score of illness severity, the anxiety score, and the status if the patient went through surgery (1) or only had medical treatment (0). The given file contains columns with names listed below.

Column's Name	Description
Satisfaction	satisfaction score about hospital's service
Age	age of abalone (year)
Severity	severity score
Surgical.Medical	1 = had surgery; 0 = no surgery
Anxiety	anxiety score

- please run function `setwd()` in a separate line (if you need it) when importing the data set into R.
- use `set.seed(310)`. You get penalty of (-2) points if you don't have it.
- **the names given in bold below MUST be used in your R code.**
- please report numerical answers to at least three significant figures if it's smaller than one (e.g. 0.0123) and to three decimal places if it's larger than one (e.g. 2.345).

Part I: Data Exploration (18 points)

Load the file `patient_satisfaction.csv` into R and name it as **data**.

1. (2 points) Write code to change the name of the 4th column of **data** to **Surgical**.
2. (2 points) Write code to create a table of proportions for column **Surgical**. Report the percentage of patients that had surgery.
3. (2 points) Write code to create a histogram with normal density curve overlay-ed for the sample of satisfaction.
4. (2 points) Write code to create a QQ plot for the sample of satisfaction. Give your comments.
5. (4 points) Write code to create box plots of the satisfaction scores by groups of surgical status. Give your comments.
6. (6 points) Write code to create a scatter plot of satisfaction score against the age of patients for which the points are classified by the surgical status: points of patients had surgical is in red color and points of patients with no surgical is in blue color. Give your comments.

Part II: Linear Model (8 points)

7. (4 points) Fit a linear regression model for the response variable **Satisfaction**, named as **M**, using all input features, **Age**, **Severity**, **Surgical** and **Anxiety**. Report p-values of the regressors that are NOT significant in the model, at significance level 0.1.
8. (4 points) Two patients, A and B, that have information listed below. Write code to predict the satisfaction score for both of them. Report the prediction values.
A: **Age** = 35, **Severity** = 45, **Anxiety** = 2.5 , and had no surgical.
B: **Age** = 60, **Severity** = 40, **Anxiety** = 3 , and had surgical.

Part III: KNN (24 points)

9. (2 points) Write R code to create a new column in **data**, names as **S** where **S** has two categories: *Good* and *Not Good*.
S is *Good* if the satisfaction score in **Satisfaction** is larger than 70; and **S** is *Not Good* if the satisfaction score in **Satisfaction** is ≤ 70 .
10. (2 points) Write code to create a new data frame, called **data.X** which has the three columns **Age**, **Severity** and **Anxiety** after standardization for all patients.
11. (4 points) Write code to randomly split the total patients into two groups: one group has 15 patients (will be the train set) and other group has the rest of patients (will be the test set).
For the questions below, we consider **S** as the response variable. We would want to form KNN classifiers using input features **Age**, **Severity** and **Anxiety** which helps to predict if a patient ranks “Good” or “Not Good” when the patient discharges from the hospital.
12. (6 points) Use the train set with three standardized features to form the KNN classifiers where **k** = 3, 5, 7, 9, 11, and accuracy value for each classifier is kept in a vector named **accuracy**.
13. (2 points) Write code to plot **accuracy** against **k** in the question above.
14. (2 points) Using accuracy as the criterion, report the best **k** found and the accuracy of the KNN classifier with that value of **k**.
15. (6 points) Use the KNN with the best **k** found in Question 14 to predict the rank (*Good*, *Not Good*) that patient A listed in Question 8 will evaluates the hospital service.

Hint: you can use the same train set as in Question 12 above. You may standardize the values for the new patient using the mean and the standard deviation of all the patients in the data set given.

2. (*10 points*) Alena Lee is working for a company where her current salary is \$50,000 annually and the salary is increased 5% every year. So far, until end of 2024, she has a saving amount of \$30,000. She plans to form a start-up company which will requires an initial amount of \$100,000.

1. Define a function, **function.year**, which helps to calculate and return the number of years that Alena Lee could save up enough money if the proportion of annual salary she puts for savings is fixed at 0.2 (20%).
2. Using function **function.year** defined above, write R codes that help to calculate the smallest proportion of annual salary that she should save so that she could have enough money for her start-up at the end of 2029, (5 years counting from start of 2025 to end of 2029).

Note: The proportions of saving salary should be up to 2 decimal places only.

END OF QUESTIONS