NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

DSA1101 Introduction to Data Science

(Semester 1 AY2025/2026)

Midterm Test

Time Allowed: 80 minutes

INSTRUCTIONS TO STUDENTS

1. Students are required to complete this test individually.
2. This is an OPEN BOOK BLOCKED INTERNET exam. You may refer to your lecture notes, tutorials, textbooks or any notes that you have made, but you are not allowed to perform any online search.
3. The use of offline Large Language Models (LLMs) is prohibited for this test.
4. Both programmable calculators and non-programmable calculators are allowed.
5. Students are required to answer ALL questions. Total mark is 60.
6. At the end of the test,
    - Copy and paste your R code into the Examplify text box. Indentation and alignment may not be retained when pasting, but that is acceptable.
    - Save an exact copy of your R file on your laptop, and upload it to Canvas/DSA1101/Assignments/Midterm Submission immediately after the test.
7. Do not modify the code in your submission file. Any difference found (except for indentation or alignment) between Examplify and Canvas submissions will be penalized.
8. **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.
9. Be sure to lay out systematically the various parts and steps in your code file by the comments in R code file.

1. (*45 points*) The data file `ford.csv` contains data from a random sample of 17965 used Ford car listings in the UK. The used Ford cars being sold belong to cars registered from 1996 to 2020.

The given file contains columns with their names listed below.

Column's Name	Description
model	model of the car
year	year the car was registered
price	price of the car listing (in £)
transmission	transmission type (Automatic/Semi-Auto/Manual)
mileage	miles travelled
fuelType	type of fuel used (Petrol/Diesel/Electric/Hybrid/Other)
tax	road tax of the car (in £)
mpg	miles per gallon
engineSize	size of engine (in litres)

- please run function `setwd()` in a separate line (if you need it) when importing the data set into R.
- **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 absolute is smaller than one (e.g. 0.0123) and to three decimal places if it's absolute is larger than one (e.g. -2.345).

Part I: Data Exploration (14 points)

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

1. (2 points) Write code to change all listings with **fuelType** equals to "Other" into "Hybrid" instead.
2. (2 points) Write code to remove all listings with invalid **year**. After the removal, report the number of rows in **data**. *Hint*: it should be an even number.
3. (2 points) Create a contingency table (named **tab1**) for **transmission** and **fuelType**. Report the number of cars that use petrol and also have an automatic transmission.
4. (2 points) Write code to create box plots of car's price by groups of **year** for years 2016 to 2020. Give your comments.
5. (3 points) Write code to create a new column for **data**, named **priceHL**, which equals to "high" if the **price** of the car is above £12500, and equals to "low" otherwise. Create a contingency table (named **tab2**) for **transmission** and **priceHL**, but only for "Automatic" and "Manual" cars.

6. (3 points) Using **tab2**, write one command in R to find the two probabilities below.

   (i) the probability of having high price car in the group of cars with automatic transmission

   (ii) the probability of having high price car in the group of cars with manual transmission

   Report the difference of the two probabilities above and interpret the meaning of that difference.

Part II: Linear Regression (6 points)

For the questions in this Part II, we would want to form a linear regression using variables **fuelType**, **transmission**, **mileage** and **mpg** to predict a car's **price**.

7. (2 points) Write code to find the correlation between **price** and **mileage**. Compare with the correlation between **1/price** and **mileage**. Give your comments.

8. (2 points) Write code to create a linear model, called **model1**, which uses **fuelType**, **transmission**, **mileage** and **mpg** to predict a car's **price**. Show the coefficients of **model1**.

9. (2 points) Using **model1**, predict the **price** of a car which has a manual transmission, runs on petrol at 60 miles per gallon, and has a mileage of 30000.

Part III: KNN (25 points)

For the questions in this Part III, we would want to form KNN classifiers using numeric input features **mileage**, **tax**, **mpg** and **engineSize** which help to predict if a car will be sold as high-price or not. Hence, we would consider **priceHL** as the response variable and high-price is considered as positive.

10. (2 points) Write code to create a new data frame, called **data.KNN** which its first column is the column of the response and other columns are **mileage**, **tax**, **mpg**, and **engineSize** after standardization for all observations;

11. (2 points) Run the command **set.seed(210)**.

    Then, write code to randomly split **data.KNN** into two groups of equal size: one group is named as **train.set** and other group has the rest of rows which is named as **test.set**.

12. (8 points) Write R code to create a vector of odd values, from 3 to up to 25, named **K**.

    Use **train.set** to form the KNN classifiers where each classifier has the value of nearest neighbours k from **K**, to predict for **test.set**. The values of FNR (Type 2 error rate) and accuracy for all classifiers are kept in two vectors, named **fnr** and **accuracy**, respectively.

    Hint: Use "for" loop; and run the command **set.seed(210)** before running the "for" loop to get stable results. Computer might take a few minutes to finish the loops.

13. (5 points) Write R code to find all the values of k that gives FNR value (Type 2 error Rate) smaller than 0.1.

    Report the matrix that consists those values of k with the corresponding FNR and accuracy values (as comments in your answer file). Denote that matrix as **good.fnr**.

Among the values of k that gives FNR smaller than 0.1, write code to find the best value of k that has highest accuracy, and name it as **best.K**.

Report (as a comment in your answer file) the values of FNR and accuracy for that **best.K**.

14. (2 points) Write R code to form a KNN classifier using the best k found in Question 13, where the classifier is formed based on **test.set**; and is used to predict the outcomes for **train.set**, where the output is named as **pred.best.K**.

15. (2 points) Write R code to find the values of FNR and accuracy for the prediction in Question 14, and report these values.

16. (4 points) Use the KNN classifier formed in Question 14 with the best k to predict the label for a used Ford car with information given below. Report the output.

    mileage = 30000, **tax** = 110, **mpg** = 60, **engineSize** = 1.

    Hint: You may standardize the values for the new observation using the mean and the standard deviation of all the observations in the given data set.

2. (*15 points*) Adam and Eve wants to take up a housing loan from the Housing & Development Board (HDB). The intended loan amount is \$700,000. At the start of each month, the remaining loan amount increases by an interest of 2.6% divided by 12. At the end of each month, they want to pay \$2,500 together to cover the loan.

1. (5 points) Using a while loop, calculate the number of months that Adam and Eve will take to pay back their loan. Since housing loans from HDB need to be paid back in 25 years, will their plan be possible?

2. (5 points) Define a function in R, named **F**, which helps them to calculate the maximum 25-year loan that they can take for a certain monthly payment amount. Using the function, what is the maximum 25-year loan that they can take with a monthly payment of \$2,500?

   Note: Function **F** must have at least one argument, named **payment**, to specify Adam and Eve's starting monthly payment amount.

3. (5 points) Using any loop together with the function **F**, find the minimum monthly payment amount needed to get the \$700,000 loan. The minimum monthly payment should be **rounded UP** to the nearest integer.

END OF QUESTIONS