UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER II

WQD7001 : Principles of Data Science

June 2019 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (40 marks).

(This question paper consists of 5 questions on 3 printed pages)

1. (a) Explain why exploratory data analysis (EDA) is valuable to data science projects.

(3 marks)

(b) A survey was conducted for 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, purchaser were asked the age they were when they purchased their car. The following box plots shown in Figure1 display the results.

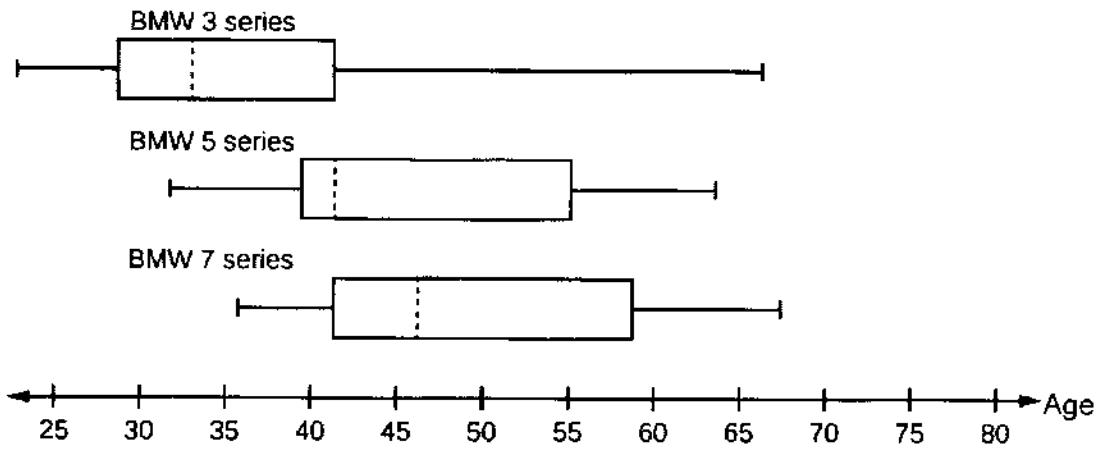

Figure 1: Box plots for car purchaser survey

i. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
ii. Which group is most likely to have an outlier? Explain how you determined that.
iii. Compare the box plots in Figure 1. What do they imply about the age of purchasing a BMW from the series when compared to each other?
iv. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
v. Look at the BMW 7 series. Estimate the interquartile range (IQR).

(5 marks)

2. (a) Illustrate the concept of bootstrap sampling with a diagram.

(3 marks)

(b) Determine the concept and goal of A/B Testing.

(3 marks)

(c) Is more data always better? Explain your rational perspective on this matter.

(2 marks)

3. (a) Provide an example of a dataset to which you could apply linear regression. State what is the goal of applying linear regression to that dataset, and make up a few numerical values of dummy data points to make it clear what the data could look like.

(4 marks)

(b) Propose suitable R tools that can aid reproducible research. Name the tool and write its function for each of these categories:

i. Authoring
ii. Typsetting / Formatting Languages
iii. R Packages
iv. Version control

(4 marks)

4. (a) Express how machine learning work using a diagram.

(5 marks)

(b) A typical question asked by a beginner, when facing a wide variety of machine learning algorithms, is "which algorithm should I use?" How would you answer this question? Justify your answer.

(3 marks)

5. (a) Determine **THREE (3)** types of data products, and give a specific example for each type and explain the contribution of data in the products.

(4 marks)

(b) Suppose you have text-based data collected from social media users on their words of appreciation to their favorite teachers. Create **ONE (1)** qualitative visualization for that text-based data and explain the technique that you present.

(4 marks)

END