UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2017/2018 : SEMESTER I

WQD7001 : Principles of Data Science

Jan 2018 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

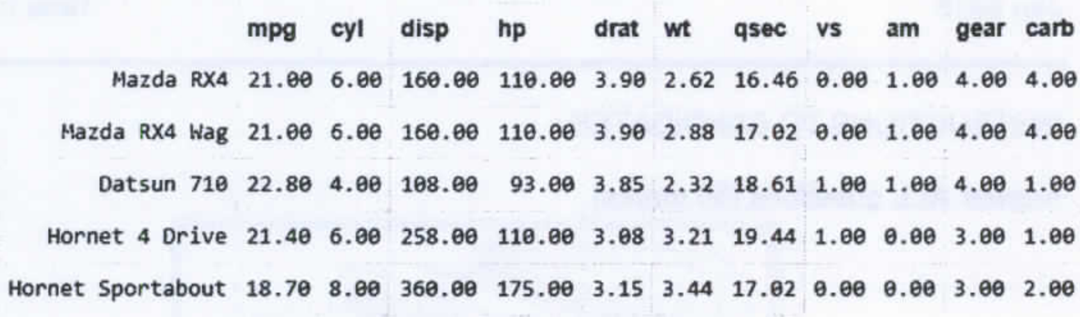
(This question paper consists of 4 questions on 4 printed pages)

1. Exploratory analysis is largely concerned with summarizing and visualizing data before performing formal modelling.

   a) Indicate **FOUR (4)** purpose of exploratory data analysis.

   (4 marks)

   b) The snapshot of the dataset "mtcars" is shown in Figure 1 below.



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.00	6.00	160.00	110.00	3.90	2.62	16.46	0.00	1.00	4.00	4.00
Mazda RX4 Wag	21.00	6.00	160.00	110.00	3.90	2.88	17.02	0.00	1.00	4.00	4.00
Datsun 710	22.80	4.00	108.00	93.00	3.85	2.32	18.61	1.00	1.00	4.00	1.00
Hornet 4 Drive	21.40	6.00	258.00	110.00	3.08	3.21	19.44	1.00	0.00	3.00	1.00
Hornet Sportabout	18.70	8.00	360.00	175.00	3.15	3.44	17.02	0.00	0.00	3.00	2.00

Figure 1 – mtcars dataset

   Show how you would perform the following exploration and manipulation on the given dataset.

   i) List names of variables in the dataset.

   ii) Show the number of rows/columns of the dataset.

   iii) See a summary statistics of the Gross horsepower (hp).

   iv) Display the structure of the dataset.

   v) Visually check possible correlated variables.

   vi) Return number of missing values in each variable of a dataset.

   vii) Replace all the 'am' transmission values with words representation. Transmission (0 = automatic, 1 = manual)

   (8 marks)

   c) Decide which descriptive statistics tool should you use according to the given situation.

Data type	Objective	Example	Statistical tools	Graphical tools
(i) Quantitative One variable	Estimate a frequency distribution	How many students per batch fail to graduate on time?	?	?

(ii) Quantitative One variable	Measure the central tendency of one sample	What is the average grade in a course?	?	?
(iii) Quantitative One variable	Measure the dispersion of one sample	How are the grades dispersed around the average grade in a course?	?	?
(iv) Quantitative Two variables	Describe the association between two variables	Does plant biomass increase or decrease with soil Pb content?	?	?
(v) Qualitative (univariate analysis)	Detect the most frequent category	Which is the most frequent eye color in Malaysia?	?	?

(8 marks)

2. a) Explain the problems reproducibility research **CAN** and **CANNOT** solve.

(6 marks)

b) Distinguish between weaving and tangling in literate programming.

(4 marks)

3. a) Machine learning (ML) is not a solution for every type of problem. There are certain cases where robust solutions can be developed without using ML techniques.
Discuss **TWO (2)** situations where machine learning is useful.

(4 marks)

b) Formal ML is defined as : A computer program is said to learn from experience E, with respect to some task T, and some performance measure P, if its performance on T as measured by P improves with experience E.

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam.
What is the task **T**, experience **E** and performance measure **P** in this setting?

(3 marks)

c) How would you write a program to distinguish a picture of yourself from a picture of someone else?

(3 marks)

4. The phrase "data storytelling" has been associated with many things - data visualizations, infographics, dashboards, data presentations, and so on. Too often data storytelling is interpreted as just visualizing data effectively. However, it is much more than just creating visually-appealing data charts.

Present a three key elements structured approach of data storytelling for communicating data insights. Summarize your presentation with a diagram.

(10 marks)

END