UNIVERSITY OF MALAYA

MIDDLE SEMESTER EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2018/2019 : SEMESTER 1

WQD7001 : Principles of Data Science

OCTOBER 2018 Time : 1½ hours

INSTRUCTIONS TO CANDIDATES :

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

**NAME**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**MATRIC NO.**: **WQD 1\_\_ \_\_ \_\_ \_\_ \_\_**

(This question paper consists of 5 questions on 6 printed pages)

1. Multiple-choice questions. **Underline** the **ONE** best answer

a) Which of the following is performed by a data scientist?

i. Create reproducible code.

ii. Analyze data for actionable insights.

iii. Challenge results.

iv. All of the functions mentioned.

b) The Command Line Interface (CLI) can help you do the following, EXCEPT?

i. Navigate folders.

ii. Create files, folders, and programs.

iii. Choose an action using menus.

iv. Run computer programs.

c) Suppose you have forked a repository called datascientist on Github but it isn’t on your local computer yet.

Which of the following is the command to bring the directory to your local computer? (For this question assume that your user name is wqd7001)

1. git pull https://github.com/ wqd7001/datascientist.git
2. git clone https://github.com/ wqd7001/datascientist.git
3. git init
4. git pull datascientist master

d) Which of the following big data characteristics are more concerned to data science?

i. Variety ii. Velocity iii. Volume iv. Volatility

e) Point out the **wrong** statement:

i. The big volume indeed represents Big Data.

ii. Big Data is just about lots of data.

iii. The data growth and social media explosion have changed how we look at the data.

iv. All of the Mentioned.

f) Which of the following is characteristic of Raw Data?

i. Data is ready for analysis.

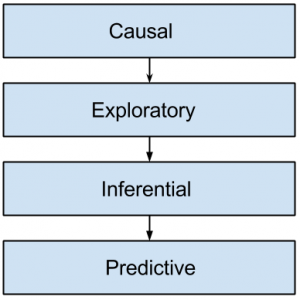
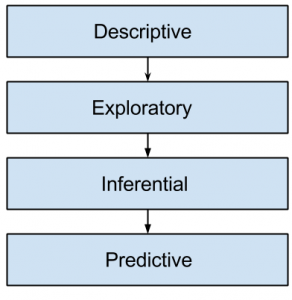
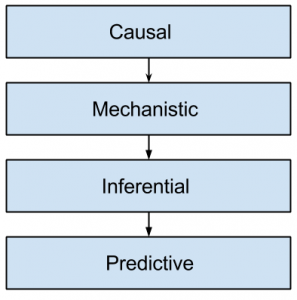
ii. Original version of data.

iii. Easy to use for data analysis.

iv. None of the Mentioned.

g) Which of the following figure correctly shows approximate order of difficulty?

i. ii iii

h) Which of the following is common goal of statistical modelling?

i. Inference ii. Summarizing iii. Subsetting iv. Descriptive

i) All data science investigations start with an existing dataset.

i. TRUE ii. FALSE

j) Most data scientists spend the majority of their time developing new models.

i. TRUE ii. FALSE

[10 marks]

1. (a) Why was it not possible for data science to exist 20 years ago?.

[4 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(b) Distinguish between data analysis, data analytics and data mining. Include a diagram in your explanation.

[6 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

1. (a) Digital disruption happens when advances in technology change our

markets and our societies. Digital disruption has already happened.

Digital disruption is threatening the survival of many businesses and industries.

Identify **TWO** industries most vulnerable to digital disruption and show **THREE** successful digital disruption examples.

[6 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(b) Why is datafication important? Provide **TWO** examples of datafication.

[4 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

1. Data science is an ocean of information. This is one of the major challenges faced by many aspiring data scientists. As a self-made data scientist, you are invited to give enlightenment on this subject.

(a) What are **TWO** useful learning resources a data scientist can refer to?

[4 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(b) Who is an inspiring data scientist that you would like to recommend people to follow? State his/her organization. (e.g. DJ Patil from Devoted Health, San Francisco)

[2 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(c) Where would you recommend the newbies to join free training programs?

[2 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

(d) Where can they find datasets to do a data science project?

[2 marks]

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------

1. Argue why dropping observations that have missing values and imputing the missing values based on other observations are not recommended. Suggest the solutions on how to handle missing values for categorical and numerical data.

*Use the blank page to write your answer.*

[10 marks]

**END**