UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2017/2018 : SEMESTER II

WQD7001 : Principles of Data Science

May/June 2018 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (60 marks).

(This question paper consists of 8 questions on 4 printed pages)

1. With the accelerated growth of tools allowing for easy implementation of powerful machine learning algorithms, it can become tempting for an amateur data scientist to skip the exploratory data analysis.

   Anticipate the effects of skipping exploratory data analysis in a data science project.

   (5 marks)

2. Suggest **FOUR (4)** ways you can use exploratory graphs to begin viewing what your own data can reveal. Your suggestion must include the type of exploratory graph, what it shows and its purpose in exploring your data.

   (12 marks)

3. Choose **ONE (1)** best answer for the following multiple choice questions.

   a) Which of the following statements best describes the relationship between a parameter and a statistic?
      i. A parameter has a sampling distribution with the statistic as its mean.
      ii. A parameter has a sampling distribution that can be used to determine what values the statistic is likely to have in repeated samples.
      iii. A parameter is used to estimate a statistic.
      iv. A statistic is used to estimate a parameter.

   b) A randomly selected sample of 1,000 university students was asked whether they had ever used the drug opium. Seventeen percent (17% or 0.17) of the 1,000 students surveyed said they had. Which one of the following statements about the number 0.17 is correct?
      i. It is a population proportion.
      ii. It is a sample proportion.
      iii. It is a margin of error.
      iv. It is a randomly chosen number.

   c) Null and alternative hypotheses are statements about:
      i. population parameters.
      ii. sample parameters.
      iii. sample statistics.
      iv. it depends - sometimes population parameters and sometimes sample statistics.

   d) Suppose a 95% confidence interval for the proportion of Malaysians who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?
      i. It is reasonable to say that more than 25% of Americans exercise regularly.

ii. It is reasonable to say that more than 40% of Americans exercise regularly.
iii. The hypothesis that 33% of Americans exercise regularly cannot be rejected.
iv. It is reasonable to say that fewer than 40% of Americans exercise regularly.

e) Which of the following is NOT true about the standard error of a statistic?
i. The standard error measures, roughly, the average difference between the statistic and the population parameter.
ii. The standard error is the estimated standard deviation of the sampling distribution for the statistic.
iii. The standard error can never be a negative number.
iv. The standard error increases as the sample size(s) increases.

f) A result is called "statistically significant" whenever
i. the null hypothesis is true.
ii. the alternative hypothesis is true.
iii. the p-value is less or equal to the significance level.
iv. the p-value is larger than the significance level.

g) Which of the following would be a legitimate reason for removing an outlier from a dataset?
i. The outlier is the result of natural variability in the measurement of interest.
ii. The outlier clearly belongs to a different population.
iii. The outlier is more than two standard deviations from the mean.
iv. The outlier is the only negative number in the dataset.

h) The Law of Large Numbers states that
i. individual occurrences are predictable and group occurrences are unpredictable.
ii. group data always follows a normal pattern.
iii. individual occurrences are unpredictable and group occurrences are predictable.
iv. the standard deviation of group data will always be greater than ten.

i) "Reno started eating a blueberry muffin for breakfast and his cholesterol level dropped. It must be because of the blueberry". Which is TRUE about this statement?
i. It is unclear which variable is the cause and which the effect.
ii. There are confounding variables.
iii. It is unreasonable to generalize from the sample actually studied to a larger study.
iv. The variables actually measured are not good stand-ins for the variables of interest.

j) The purpose of simple linear regression analysis is to:
   i. Predict one variable from another variable.
   ii. Replace points on a scatter diagram by a straight-line.
   iii. Measure the degree to which two variables are linearly associated.
   iv. Obtain the expected value of the independent random variable for a given value of the dependent variable.

(10 marks)

4. Clarify the difference between a training set, a test set and a validation set in the machine learning model.

(6 marks)

5. Evaluating the accuracy of predictive model is one of the most important tasks in the data science project. It indicates how good predictions are. In classification problems we look at metrics called precision and recall.

   Illustrate precision and recall using a confusion matrix.

(8 marks)

6. Using a diagram, explain the concept of reproducible research which you are going to adopt for your data science project.

(9 marks)

7. What are recommender systems?

(3 marks)

8. Data storytelling means different things to different people. There's no one way to tell a story. Data unlocks limitless possibilities, allowing you to shape facts and statistics into any form. For instance, an infographic can be used to illustrate a process, illuminate trends, support an argument, and drive emotions.

   Sketch **ONE (1)** narrative that showcase the power of data and explain how to do it.

(7 marks)

END