SECTION B: 5 QUESTIONS, 25 MARKS

Write your answers to this section in a separate answer book and write "Section B" on the first page of the book.

1. (5 points) Data Analytics

   (a) (2 points) A supermarket chain asks you to develop a system that would characterise types of shoppers. They have a record of past shopping transactions of shoppers, including the list of products bought in each shopping transaction, and the location. Answer the following questions.

      1. (1 mark) Is this an example of descriptive, diagnostic, predictive, or prescriptive analytics? Justify your answer.
      2. (1 mark) Given the data that the financial institution has, Would you apply supervised machine learning or unsupervised machine learning? Justify your answer.

   (b) (3 points) Explain the concepts of data, information, knowledge, wisdom. Pay attention to describe the distinguishing characteristics of each.

   (c) (3 points) Explain the three main steps in a supervised learning process: feature extraction, training, classification. Describe each step in two or three sentences.

   (d) (2 points) Explain the difference between numeric and categorical data, and how it affects whether we need to apply classification or regression techniques.

   (e) (2 points) In 3 to 5 sentences, explain the role of a Graphics Processing Unit (GPU) to help applying machine learning on data.

2. (5 points) Unstructured Data

   (a) (3 points) Describe three use cases of image analytics.

   (b) (2 points) Describe two common approaches to identify keywords from text.

   (c) (2 points) Explain what is the difference between stemming and lemmatisation.

   (d) (2 points) Explain what Part of Speech tagging is and how it could be useful for text analytics.

   (e) (0 points) ~~List the parts of speech in SAS Text Miner~~ (we would **not** ask this sort of question)

3. (5 points) Visual Analytics

   (a) (2 points) Explain how one can use a histogram to summarise the distribution of values in a numerical variable.

   (b) (2 points) Explain what is jittering and how one can use it to improve a plot.

   (c) (3 points) A car dealer wishes to determine the most popular car makes and people's opinions on them. Explain three visual analytic techniques that could be useful for this study. For each visual analytic technique, make sure that you specify the visual analytic technique, the reason for its use, and the data source on which it would be applied.

4. (5 points) Stream Processing

   (a) (2 points) Describe 2 applications of data stream analytics.

   (b) (2 points) Develop the issues of Velocity and Volume in stream processing and how they can be addressed.

(c) (3 points) Draw the Stream Model and explain all of its components.

(d) (2 points) Explain StreamSQL and its relation to event stream processing.

(e) (3 points) Sketch the online machine learning algorithm and explain how it can be used for stream processing.

5. (5 points) Big Data and Society

(a) (2 points) Explain what is the connection between bias in data and the occurrence of racist AI.

(b) (2 points) Explain Danah Boyd and Kate Crawford's comment "Big Data changes the definition of knowledge" in their 2012 paper "Critical Questions for Big Data".

END OF SECTION B. PLEASE REVISE YOUR ANSWERS.