DESK No. □□□

FAMILY NAME: ______________________

GIVEN NAMES: ______________________

SIGNATURE: ______________________

STUDENT NUMBER: □□□□□□□□

THE UNIVERSITY OF
WESTERN
AUSTRALIA

SEMESTER 2, 2018 EXAMINATIONS

CITS4009

Physics, Mathematics & Computing

Department of Computer Science & Software Engineering

Introduction to Data Science

This paper contains: **5** Pages **(including title page)**

Time Allowed: **2:00** hours

INSTRUCTIONS:

This paper consists of 6 questions, each is worth 10 marks.

Answer all questions.

All questions need to be answered in the provided Answer Booklet.

Nothing written in the Question Booklet will be considered during marking.

THIS IS A CLOSED BOOK EXAMINATION

SUPPLIED STATIONERY	**ALLOWABLE ITEMS**
1 x Answer Booklet 18 Pages	**UWA Approved Calculator with Sticker**

This page has been left intentionally blank

Q1. Data Science Fundamentals

a) [6 marks] Use an example domain (e.g. the dataset you worked with in your projects) to describe the full life cycle of a data science project.

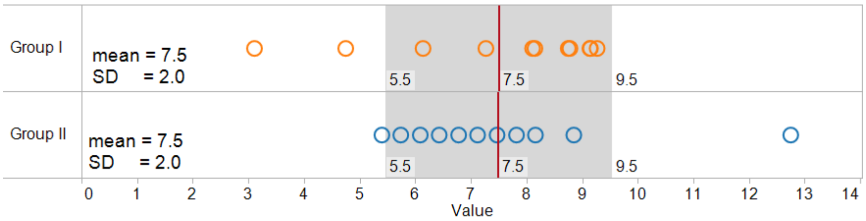
b) [4 marks] List four different ways of sub-setting a data frame.

Q2. Visualisation and Exploratory Data Analysis

a) [3 marks] What are the three essential components of the layered grammar of graphics that ggplot implements?

b) [4 marks] List two sensible questions to ask when doing EDA, and illustrate using sketches to show which types of visualisations are best suited to answer each question.

c) [3 marks] Given the following data distribution of two groups, explain why boxplot stats are better at capturing the central tendency in this case.

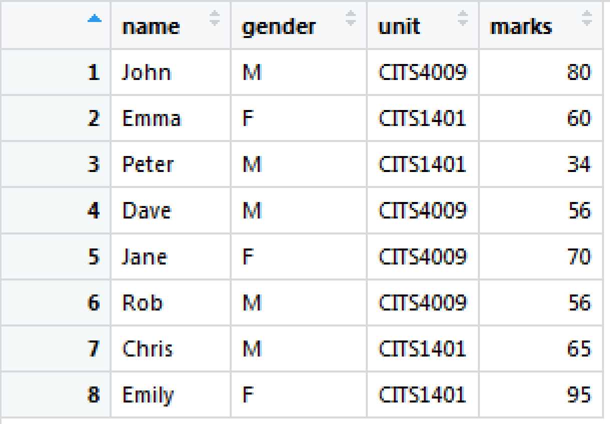


Q3. Data Cleansing and Transformation

a) [4 marks] What are the basic considerations and strategies for dealing with NAs in your data? Use example scenarios to illustrate data-specific and domain-specific treatments of NAs.

b) [3 marks] Given the following dataframe `df`, write R code to insert a new column that converts the marks in to grade.

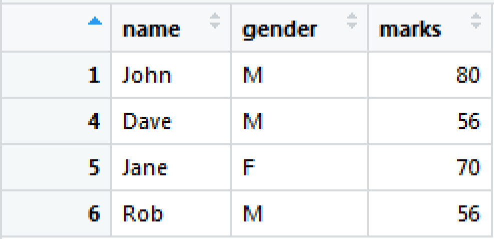
- [0, 50) to "NN"
- [50, 60) to "PA"
- [60, 70) to "CR"
- [70, 80) to "D"
- [80, 100] to "HD"



	name	gender	unit	marks
1	John	M	CITS4009	80
2	Emma	F	CITS1401	60
3	Peter	M	CITS1401	34
4	Dave	M	CITS4009	56
5	Jane	F	CITS4009	70
6	Rob	M	CITS4009	56
7	Chris	M	CITS1401	65
8	Emily	F	CITS1401	95

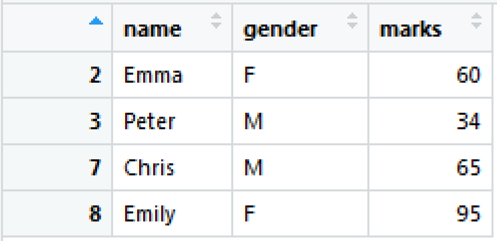
c) [3 marks] Explain why it is important to have tidy data. Given the two data frames (`df_cits4009` and `df_cits1401`) below, write R code to create the data frame above (`df`).

df_cits4009



	name	gender	marks
1	John	M	80
4	Dave	M	56
5	Jane	F	70
6	Rob	M	56

df_cits1401



	name	gender	marks
2	Emma	F	60
3	Peter	M	34
7	Chris	M	65
8	Emily	F	95

Q4. Classification

You have been employed as a data scientist to analyse the historical data of home credit default risk, in an attempt to predict how capable each applicant is of repaying a loan. For each historical application, the applicant's *age at the application*, *gender*, the *loan amount*, the *annual repayment amount*, the *education level*, *marital status*, the *annual family combined income*, the *number of children* are recorded.

a) [5 marks] Given the response variable as *binary*, yes for "risky", no for "no risk", choose one of the basic memorisation methods that you are familiar with, and explain clearly how the model works.

b) [5 marks] Explain the general strategies of model evaluation. Select metrics that you are familiar with (e.g. precision and recall), describe what they are measuring and how they can be calculated.

Q5. Regression

Use the same dataset in Q4, but now the response variable is a *score* indicating how likely the applicant is at risk of not repaying a loan.

a) [5 marks] Describe how one can build a linear regression model using the credit risk dataset.

b) [5 marks] Describe how one can build a logistic regression model of the same dataset and explain how it is different from the linear regression model.

Q6. Unsupervised Methods

Use the same dataset in Q4.

c) [5 marks] Select a clustering techniques that you are familiar with, and describe how it works in this scenario. Select two relevant features and discuss the appropriate distance measures to use for each feature.

d) [5 marks] When applying association rule mining to this dataset, we can treat each applicant as a basket, describe how you would apply the `apriori` function to find association rules for gender, the education level, and marital status. First illustrate what the transaction set looks like, and then explain how to measure the interestingness of a rule by writing down the formula to calculate `confidence` and `support`.

Assume gender has two levels ‘`M`’, ‘`F`’. Education level takes three values ‘`Incomplete`’, ‘`Secondary`’, ‘`Tertiary`’, and marital status has three levels ‘`Married`’, ‘`Separated`’, ’`Single`’.