DESK No. □□□

FAMILY NAME: ______________________

GIVEN NAMES: ______________________

SIGNATURE: ______________________

STUDENT NUMBER: □□□□□□□□

SEMESTER 2, 2020 EXAMINATIONS

CITS4009

Physics, Mathematics & Computing

Department of Computer Science & Software Engineering

Computational Data Analysis

This paper contains: **4** Pages **(including title page)**

Time Allowed: **2:00** hours

INSTRUCTIONS:
* Answer all questions. The exam has seven (7) questions, worth a total of sixty (60) marks, which contributes to 60% of the total assessment in this unit.
* Answers do not need to be lengthy, be concise instead.
* If you believe that a question is ambiguous, state clearly any assumptions that you make in constructing your answer.
* For questions that require you to **"Write R code"**, minor syntactic errors will not be punished; but syntactic errors that obscure the meaning of an answer might cost you marks. Pseudo code may be given partial marks.

Students can bring one A4 page of notes to the examination which can be written on both sides.

THIS IS A CLOSED BOOK EXAMINATION (SEE ALLOWABLE ITEMS)

SUPPLIED STATIONERY	ALLOWABLE ITEMS
1 x Answer booklet 10 pages **1 x Student Notes Only. Please Specify Below.**	**UWA Approved Calculator with Sticker**

PLEASE NOTE

Examination candidates may only bring authorised materials into the examination room. If a supervisor finds, during the examination, that you have unauthorised material, in whatever form, in the vicinity of your desk or on your person, whether in the examination room or the toilets or en route to/from the toilets, the matter will be reported to the head of school and disciplinary action will normally be taken against you. This action may result in your being deprived of any credit for this examination or even, in some cases, for the whole unit. This will apply regardless of whether the material has been used at the time it is found. Therefore, any candidate who has brought any unauthorised material whatsoever into the examination room should declare it to the supervisor immediately. Candidates who are uncertain whether any material is authorised should ask the supervisor for clarification.

Candidates must comply with the Examination Rules of the University and with the directions of supervisors.
No electronic devices are permitted during the examination.
All question papers and answer booklets are the property of the University and remain so at all times.

This page has been left intentionally blank

1.

1. [2 marks] Use an example to illustrate that *Exploratory Data Analysis* is an iterative process.
2. [2 marks] What's main difference between a *histogram* and a *bar chart* for a single variable?
3. [3 marks] When comparing two categorical variables using bar charts, what are the three different types of *position adjustment*? Use the function *ggplot()* to **write R code** to illustrate.
4. [3 marks] Using two example categorical variables used in your project, explain what situation(s) each type of bar charts is most suitable for.

(Please write your essay question on a separate piece of paper)

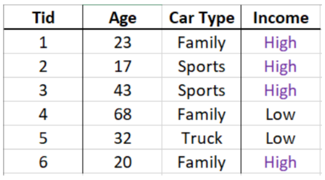
2.

1. [3 marks] What are the basic considerations and strategies for dealing with NAs in your data?
2. [3 marks] Use example scenarios to illustrate and justify *data-specific* and *domain-specific* treatments of NAs.
3. [4 marks] Explain how the *vtreat* package in R creates and applies a *treatment plan.*

(Please write your essay question on a separate piece of paper)

3.

Use the data frame below to answer the following questions. Variable *Age* is numerical, *Car Type* is categorical and *Income* is the target variable.



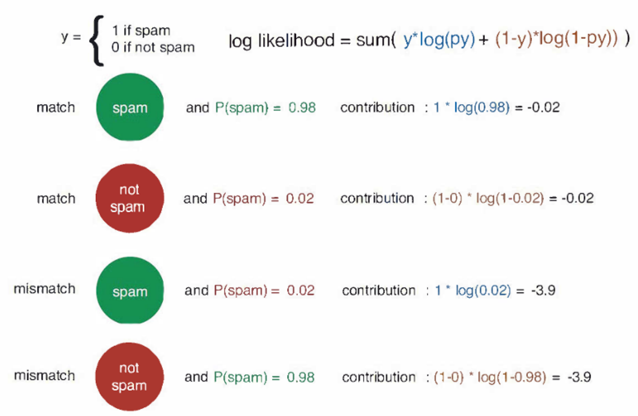
Tid	Age	Car Type	Income
1	23	Family	High
2	17	Sports	High
3	43	Sports	High
4	68	Family	Low
5	32	Truck	Low
6	20	Family	High

1. [3 marks] Explain the steps and intuitions of building a single variable model for categorical variables.
2. [3 marks] Write the contingency table for the *Car Type* variable with correct values for each cell.
3. [3 marks] What would be the predicated value for each *Car Type*?
4. [3 marks] **Write the R code** that can produce the contingency table.
5. [3 marks] **Write the R code** for a Null model in this scenario. What value will the null model use in this case?

(Please write your essay question on a separate piece of paper)

4.

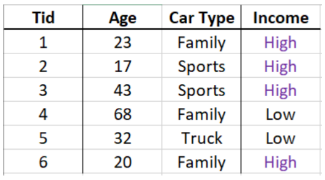
[5 marks] Given the following illustration of how to calculate the *Log Likelihood* of 4 observations of a binary classification task, **write an R function** to calculate the Log Likelihood of a model on *data* (the dataset) with two columns: *y* (binary ground truth values, 0 and 1) and (the predicated probability).


y = { 1 if spam
0 if not spam
log likelihood = sum(y*log(py) + (1-y)*log(1-py)))
match
spam
and P(spam) = 0.98
contribution : 1 * log(0.98) = -0.02
match
not spam
and P(spam) = 0.02
contribution : (1-0) * log(1-0.02) = -0.02
mismatch
spam
and P(spam) = 0.02
contribution : 1 * log(0.02) = -3.9
mismatch
not spam
and P(spam) = 0.98
contribution : (1-0) * log(1-0.98) = -3.9

```
loglikelihood <- function (data, y, py) {
   # This is where your code goes
   ...
}
```

(Please write your essay question on a separate piece of paper)

5.



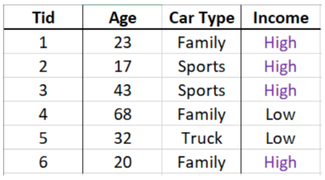
Tid	Age	Car Type	Income
1	23	Family	High
2	17	Sports	High
3	43	Sports	High
4	68	Family	Low
5	32	Truck	Low
6	20	Family	High

Use the above data frame to explain intuitively:

1. [4 marks] How a Decision Tree classifier *partitions* the dataset and *assigns* piecewise constant for each partition. You can use either Age or Car Type as the first variable.
2. [3 marks] How a k-Nearest Neighbour classifier works, and list two aspects that need to considered during data preparation.
3. [3 marks] How receiver operating characteristic curves and double density plots can be used to compare performance of binary classifiers.

(Please write your essay question on a separate piece of paper)

6.

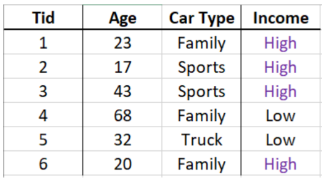


Tid	Age	Car Type	Income
1	23	Family	High
2	17	Sports	High
3	43	Sports	High
4	68	Family	Low
5	32	Truck	Low
6	20	Family	High

[3 marks] Using the above dataset, explain how the *dist()* function in R works out pairwise distances between data points.
[2 marks] Explain intuitively what the *Calinski-Harabasz Index* is, and how it is used in clustering tasks.

(Please write your essay question on a separate piece of paper)

7.



Tid	Age	Car Type	Income
1	23	Family	High
2	17	Sports	High
3	43	Sports	High
4	68	Family	Low
5	32	Truck	Low
6	20	Family	High

1. [1 mark] **Write R code** to build a *logistic regression* model on this dataset, assuming the target variable is *Income*.
2. [2 marks] Explain why in logistic regression we need to specify the *family* argument.
3. [2 marks] How do we obtain the *probability predications* of a logistic regression model on new observations? You can either explain or *write R code*.

(Please write your essay question on a separate piece of paper)