UNIVERSITY OF CANBERRA
AUSTRALIA'S CAPITAL UNIVERSITY

University of Canberra

Faculty of Science and

Technology FINAL ASSESMENT

SEMESTER 2, 2021

UNIT NAME: Introduction to Data Science

UNIT NUMBER: 11372

Due date: 7th of November, 23:59

EXAMINER'S NAME: Dr. Ibrahim Radwan

CONTACT details: ibrahim.radwan@canberra.edu.au

INSTRUCTIONS FOR STUDENTS

1. **This is a take-home assignment.**
2. **The assignment is organised into two parts, where the first part is composed of four general questions to assess your understanding of the Data Science principles. This part is expected to be delivered as a PDF file. The other part is code-based tasks, which you will need to submit R-code script and a final report. The final report should contain the output of running your code for those questions that are asking you to generate summaries or graphs from the data.**
3. **The PDF files and R script are expected to be submitted as a one compressed file (e.g. *.ZIP) on the Canvas website of the unit by the due date of the assignment.**
4. **The submitted file should be renamed as [studentID_lastname_final_assessment.zip], where "studentID" is your university ID and "lastname" is your lastname.**
5. **The assignment will be open from Tuesday, the 19th of October (17:30) until Sunday, the 7th of November (23:59).**
6. **This assignment has 100 marks in total and weighs 40% of the final grade of the unit, where 50% of the assignment marks is compulsory to pass the unit.**
7. **The assignment will cover all the learning outcomes and the taught contents of the unit.**

There are no errors deliberately placed in any of the questions in this assignment, unless explicitly stated. If you think you have identified any errors, please contact the supervisor.

Part A – Data Science Questions (15 *marks*)

There are four questions in this part, with differing marks. All answers must be recorded in a MS Word and then exported to PDF file.

Q1) (3 marks)

From your understanding of ethical data science, mention three principles of a code of ethics that any data scientist should consider.

Write your answer as:

P1: Collect relevant data, and avoid an abundance of data collection which might leads to data privacy concerns. A large collection of data does not mean they are applicable for the target which the data scientists aim at gaining. Therefore, the data has to be usable and concise enough for further uses.

P2: Identify sensitive data and the best way to deal with the delicate data. If the data is personal, this will be the moment when the data scientist decides on the best solution to withdraw an insight out of those data without violating ethical issues.

P3: Considering solutions toward the possible risks when conducting any insights of the data. Moreover, the process of decision-making should be regarded deliberately, not depending solely on machine learing decision making process.

Q2) (4 marks)

To build a visualisation using the ggplot2 library, we use the following template:

```
ggplot(data= [dataset], mapping = aes(x = [x-variable], y = [y-variable]))+
  geom_xxx() +
  other options
```

Based on the above template, mention the main components of building a graph using ggplot2 and describe the meaning of each of these components.

Write your answer as:

Main components:

- data=[dataset]: this is to bind the plot to a specific data with data function
- mapping=aes(x=[x-variable], y=[y-variable])): identify the aesthetic mapping with the aes() function.

This function is used for defining the relevant variables to be plotted and which ways

to plot them regarding

size, shape, colours, etc. x will be for the variable on the x axis, y will be for the variable on the y axis.

- Geom_xxx(): this is the plotting styles, which are the graphical representation of the data. These

Styles include points, line, bar category. The functions for some common styles are geom_point(),

geom_line(), geom_bar(), geom_boxplot(), etc.

- Other function: To make the illustration more interactive and coherent, some extra functions

Are added as a result. These functions could be geom_vline, geom_density, facet_wrap, ggtitle, theme styles, etc.

For example, theme_bw can add the theme for the entire illustration as black and white. Meanwhile,

facet_graph can be used for creating multiple sub-graphs for different elements.

- These continuous functions are connected by “+” to display the elements in a row.

Q3) (3 marks)

Describe three properties of the correlation coefficient of two variables

Write your answer as:

1. The magnitude (absolute value-R) of the correlation coefficient shows the strength of the linearity between the explanatory and the target variable. R will be represented between (+1) and (-1).
2. The sign of coefficient will indicate the direction of the relationship. (-) is a negative relationship, and (+) will show a positive direction of the relationship.
3. The correlation coefficient is a unit-free relationship, between x variable and y variable.

Q4) (5 marks)

Imagine we have a dataset that lists the heights of the fathers and their sons. You have built a linear model that encodes the relationship between the fathers' heights and the sons' heights as follows:

```
lm(son ~ father, data = heights_data)


Call:
lm(formula = son ~ father, data = heights_data)


Coefficients:
(Intercept)    father
    70.45       0.50
```

The estimated coefficient (i.e. intercept and slope), which describes the relationship between the fathers' and sons' heights can be interpreted as:

In this case, the sons' height will be assigned as y (dependant variable), and the fathers' height will be

assigned as x (independent variable). The relationship of x and y can be interpreted as below:

y = intercept + slope*x => y = 70.45 + 0.50*x

Therefore, with one cm in the height of the father, the height of the son will increase by 0.50 cm.