University of Canberra

Faculty of Science and Technology

FINAL ASSESMENT

SEMESTER 1, 2021

UNIT NAME: Introduction to Data Science

UNIT NUMBER: 11372

STUDENT NAME: Greene (Thao) Vu

STUDENT NUMBER: 3209044

Part A: Data Science Questions

1. From understanding of ethical data science, mention 3 principles of a code of ethics that any data scientist should consider:

    - Data scientist has responsibility on keeping the information confidential. This might include all kind of information that are not generally known by the public about any stakeholder. Any acts of using data to the public needs a written consent from the person that information belongs or was provided.
    - A data scientist may reveal the information about the client and proper authorities to protect the client from any harm or death, as well as prevent anyone on committing crime or fraud.
    - In communication with clients, data scientist shall provide competent data science services to client. During the services, clients have the rights to be informed about the status of the data science service, what happen to the data and potential risks. Data scientist is encouraged to discuss with the customer about reasonable request for information, and also the limitation of a data scientist on using information which is overestimated by client.

2. 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.

Answer:

- gglplot: is a R package dedicated to data visualisation. We use ggplot to create graphics efficiently.

- [dataset]: contains the variables that we want to present
- aes(): is appreciation of aesthetic, is a function where presents the variables on x and/or y=asis, and the aesthetic element such as colour, size, fill, shape, transparency.
- Symbol "+": is used at the end of the line of the code to indicate the different layers that will be added to the plot.
- Geom_xxx: type of graphical representation. For example: geom_point for scatter plot, geom_line for line plot, geom_bar for barplot,ect
- "other options": further personalise the plot such as axis's labels and graph's title.

3. Describe three properties of the correlation coefficient of two variables

- The coefficient of correlation (R) is a unit-free measure: This means that if x denotes the temperature express in C and y denotes the electricity bill expresses in Rs, then the correlation coefficient between temperature and electricity bill will be free from any unit.
- The coefficient of correlation is very sensitive to outliner: one small change would lead to significant change of R.
- The coefficient of correlation always lies between -1 and 1 including both limiting values: If R>0, y increases, R<0, y decreases and R=0, y never changes.

4. *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:

Son's height = 70.45 + 0.50*father's height

The slope of 0.5. This means if Dad A is taller than Dad B 1(height measurement), then Dad's A son is taller than Dad B's son 7.045 (height measurement) in height. This also means the kid's height is 7.045 (height measurement) higher for every 1(height measurement) Dad's height is.

The interception is 70.45. This means if two dads have the same height, then the kids are 0.5 (height measurement) different in height.