Introduction to Data Scientist 11372 (UG)

Final Assessment Part A – Data Science Questions

Tuan Anh (Vincent) Nguyen – u3196825

Q1: From your understanding of ethical data science, mention three principles of code of ethics that any data scientist should consider.

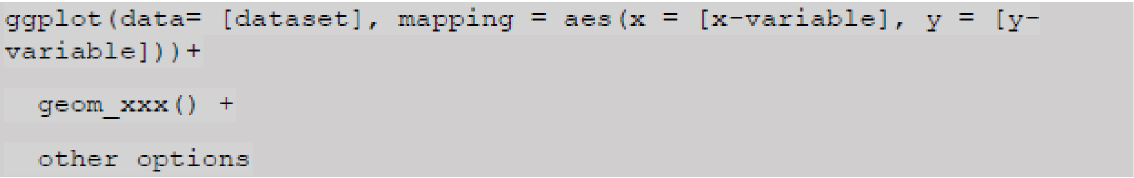
Answer:

P1: Data scientists should extract good and meaningful things from the data, it is all about what is good for people and what is better for lives. For example, use to predict the early cancers symptoms or the natural disaster forecast (tsunami, hurricane)

P2: Respect the privacy of the data, such as not using personal data for commercial purposes or personal gain without the consent of the owner of the data.

P3: Produce a trustful result. For example, US Government using data and misinterpret the statistic and number to misleading the US citizen during Vietnam War that they are winning the war to gain political benefit and support from voter.

Q2: To build a visualization 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:

There are 3 main components to build a graph:

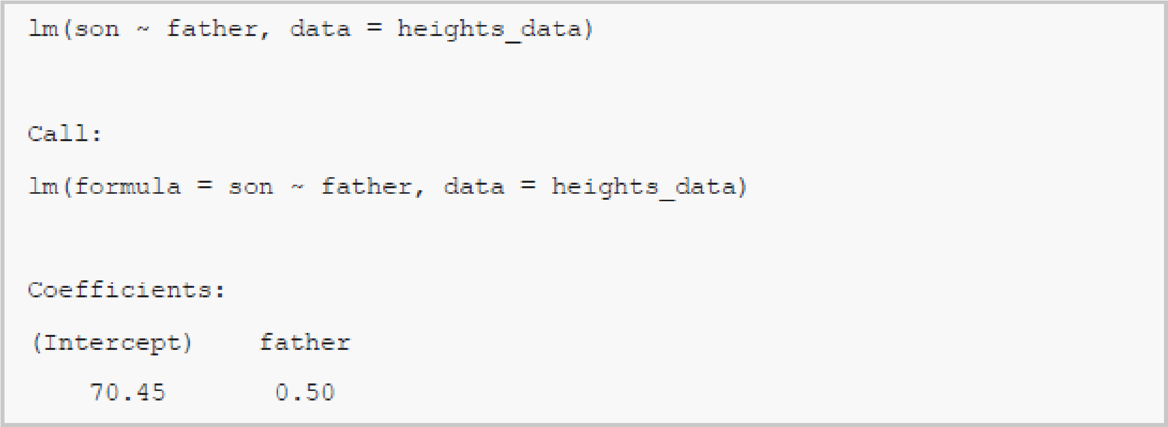
- **Data**: which is input data [dataset], usually the set of records / variables that will be presented with a graph.
- **Geometry**: type of plot that will be generated, usually function (scatter plot, boxplot, barplot, histogram…)
- **Aesthetic mapping**: the coordinate map (x and y in the template) and other visual cues (size, scale, colour, etc..).
- **Other options**: is to add other functions or options like add an additional variables, colour, layers,etc..

Q3: Describe three properties of the correlation coefficient of two variables.

Answer:

- **The magnitude (absolute value) of correlation coefficient (referred as R)** represent the of strength of the linear association between the explanatory and the response variable. The higher it is, the stronger the relation between them. R is always between +1 and -1:
    - **R = +1** means y increase the same amount as x increase
    - **R = -1** means y decrease the same amount as x increase
    - **R = 0** means y never changes when x does, i.e., the relationship is represented by horizontal line.
- **The sign of the correlation coefficient** indicates the direction of the relationship between variables. For example, a positive correlation is the relationship between the amount of coal burnt and the air quality.
- **The correlation coefficient is unit-less.** The value is stay very similar even with the scaling or normalising any of the variables.

Q4: 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's heights and the sons's heights as follow:



```
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:

Answer: (I not sure the measurement of height used in this context)

The estimated regression can be written as follow:

- Son's height = 70.45 (b0) + 0.5 (b1) x Father's height

The intercept (b0) is **70.45**. It can be interpreted as the predicted son's height for a father of height 1 is 70.45.

The slope (b1) is **0.5**. This meant that, for every 1 (height) increase in father height, the height of son is getting increased by 0.5.

References

Radwan, I., 2021. *University of Canberra - Introduction to Data Science.* [Online]
[Accessed 09 05 2021].