PARTA:

Q1) (3 marks) From your understanding of ethical data science, mention three principles of a code of ethics that any data scientist should consider.

Answe:

1. Doing no harm.
2. Respecting privacy of data collected.
3. Remember that we remain a member of the society.
   Examples of these values are as follows.
   a. Respecting privacy of the data:
   Not doing data mining on someone's personal data, if we don't want to do it on our data, Anonymizing the data whenever possible. And not using data beyond project goals or for any personal use.
   b. Producing truthful, interpreted results.
   Whether data analysis Valid, Interpretation of the results should be fair and making sense. Any social consequences of the outcome.

##

Q2) (4 marks) To build a visualization using the ggplot2 library, we use the following template:

ggplot(data= [dataset], mapping = aes(x = [x-variable], y = [yvariable]))+

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:

1. Data: Set of variables that we need to represent in terms of graph.
   "ggplot(data= [dataset], mapping = aes(x = [x-variable], y = [yvariable])) "
   represent the Data part
2. Geometry: This is for type of plot like line, scatter, boxplot etc.. which can be generated
   By using respective functions like geom_line(), geom_boxplot() and so on (**geom_xxx()**)
3. Aesthetic mapping: The coordinate map and other visual cues such as color, scale, size , group etc..(Other options)

##

Q3) (3 marks) Describe three properties of the correlation coefficient of two variables.

Answer:

1. The correlation coefficient is referred as R .
2. Magnitude/absolute value of the correlation coefficient, which represents the strength of the linear association between explanatory and target variable.
3. If the magnitude is higher, the relationship between the variables is stronger.

   R will be always between +1 and -1. + means increasing y with increasing x. – means decreasing, with increasing x. 0 means Y values never change with increase in x.

4. The sign of the correlation coefficient indicates the direction of relationship between explanatory and target variable.
5. The correlation coefficient is unitless, even with normalizing /scaling it stays same.
6. The correlation coefficient is very sensitive to outliers.

##

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:

Answer:

Son Height = 0.50*father Height+70.45

##