MATH 9102 - Probability and Statistical Inference
Assignment – 1

1. **Understanding data** (3 marks)

   (a) Assume that you have a table with variables that describe a person. Name, age, height, weight and profession. Identify variables that are discrete, continuous, and categorical. (1 mark)

   (b) Assume that you have a table with variables that describe a lecturer. Name, gender, subject, semester, and staff number. Identify variables that are ordinal, nominal, interval, and ratio. (1 mark)

   (c) You and a friend wonder if it is ”normal” that some bottles of your favourite beer contain more beer than others although the volume is stated as 0.33L. You find out from the manufacturer that the volume of beer in a bottle has a mean of 0.33L and a standard deviation of 0.03. If you now measure the beer volume in the next 100 bottles that you drink with your friend, how many of those 100 bottles are expected to contain more than 0.39L given that the information of the manufacturer is correct? (1 mark)

2. **Descriptive statistics** (6 marks)
   Use **salary.rds** dataset from lecture 1.

   (a) Install the following packages **Hmisc, pastecs, psych** (1 mark)

   (b) Describe the data using installed packages and identify the differences in description by different package. (1 mark)

   (c) Generate summary statistics by using grouping by Gender. (1 mark) *Hint: use package psych*

   Load ***iris*** dataset into workspace.

   (d) Identify mean, median, range, 98th percentile of Petal.Length (1 mark)

   (e) Draw the histogram for Septal.Width, mention which measure of dispersion method suits the best? (1 mark)

   Load ***HairEyeColor*** dataset into workspace.
   Hint: dataHairEye $<-$ *as.data.frame(HairEyeColor)*

   (f) As a customer, I would like to know the total number of people with various color combinations of hair and eyes. Which chart suits best for this task? Plot the same. (1 mark)

3. **Visualisation** (6 marks)

   (a) A meteorologist wants to compare the annual average rain fall between two cities for the past 20 years. Which plot is most suitable? Plot the graph by generating 20 random data points between 0 to 28 for Dublin and Cork. (2 marks)

   (b) Load the provided world-small.csv file. (2 marks)

      i. Draw histogram for 'gdppcap08'

      ii. Draw boxplot for 'polityIV'

      iii. Identify the region that has highest gdpcap.

      iv. Which country has lowest polityIV ?

   (c) Table 1 represents people in Dublin who like to own certain types of pets. (2 marks)

Table 1: Pet Lovers

Pet	**Number of people**
Dogs	2034
Cats	492
Fish	785
Macaw	298

      i. Plot the most suitable graph for the given dataset.

      ii. Is it a good idea to choose a pie chart (in case you have not chosen it in (i))? Why is it a good idea or why is it not a good idea?