Programming for Data Analysis, Processing and Visualisation

Assignment 1

John O'Sullivan

Instructions:

- This assignment is due on Friday the 14th of June at 11:59pm
- You should submit your assignment to the 'Assignment 1' object on Moodle
- You should submit two files:
  - (i) a single .Rmd script file containing all of the commented code you used to obtain your answers
  - (ii) the HTML (or pdf) file which you produced from the .Rmd script
- The marks available for each question are shown in brackets
- You may need to find some new functions in order to do some of these tasks. Remember to use R's search engine, as well as checking online.
- Make sure that your file is readable and has a neat presentation and clear flow. The HTML (or pdf) output file should be a stand-alone document containing the answers to all questions and showing all necessary code and all necessary output.
- I advise you to first create an R script with all of your answers. When you are happy with this, convert it piece-by-piece into an .Rmd file.
- You must use code for all answers - e.g., reading the values from a table is not sufficient to get full marks.

Question 1: [35 marks]

The dataset **GSSvocab** is available in the R package called **carData**.

(i) Load the package carData, and load the dataset **GSSvocab**. Type the command needed to see the top 6 rows of the dataset. [2 marks]

(ii) What commands are used to access the description and the structure of the dataset? Briefly describe the data contained in **GSSvocab**. [5 marks]

(iii) Print the 2,000th row of **GSSvocab**. What is the age of this person? (Use code to find this age - looking at the row is not sufficient.) [3 marks]

(iv) Create a new column called vocab.pct which is the percentage of the number of words out of 10 correct on the vocabulary test. Print the head of GSSvocab to confirm that this column has attached correctly. [5 marks]

(v) Two of the factors should be *ordered* factors. Use the help file to help you decide which two. Convert them to ordered factors, and check the structure again to confirm that this has worked. [5 marks]

(vi) Create a table of educational group against age group. What is the combination of Education Group/Age Group with the smallest number of people? [5 marks]

(vii) Produce a clustered barchart of the table in the previous question. Explain in a few lines what the clustered barchart tells us about the relationship between the two variables. (The plot should be neat and presentable.) [10 marks]

Question 2: [65 marks]

The dataset **diamonds** is available in the R package called **ggplot2**.

(i) Install and load the package ggplot2, and load the dataset diamonds. Look at the structure and the top of the dataset. Briefly describe the contents of the diamonds dataset. What kind of variables does it contain? How many observations are there? [5 marks]

(ii) Are there any missing values in the dataset? [2 marks]

(iii) Which row contains the diamond with a *depth* of 70? What colour is this diamond? [4 marks]

(iv) Use the summary() function on **diamonds**. Describe the results for any two of the variables. [5 marks]

(v) Create a table of *color* against *cut*. Write some code to find the colour/cut combination with the smallest number of diamonds. (i.e., you must find this using code, and not just by looking at the table.) [5 marks]

(vi) Using the table from the previous question, produce a table showing proportions (instead of counts) and marginal sums. Print this new table, and comment on it. [5 marks]

(vii) According to this dataset, what is the size (length, width and depth in mm) of the most expensive diamond? [4 marks]

(viii) Which are the 7 most expensive prices for diamonds of *clarity* IF? [8 marks]

(ix) How many diamonds are of Ideal cut, best colour and best clarity? [5 marks]

(x) Create a new column called *price.per.carat* which contains the price of a diamond divided by its carat. [3 marks]

(xi) Create a subset of diamonds called sub.diamonds which contains only those diamonds of cut Ideal. How many observations are there in sub.diamonds? [3 marks]

(xii) Use the aggregate() function to aggregate the *price.per.carat* column of sub.diamonds to find the mean of this variable for every *clarity* and *color* combination. Save this output as a dataframe called *df1*. [5 marks]

(xiii) Order *df1* by descending price per carat. Comment on the results. [3 marks]

(xiv) Be creative - produce an interesting table, a plot, or create a new variable which helps to tell us something new about the diamonds dataset. Describe your findings. [8 marks]