Dublin Business School
excellence through learning

QQI

Master of Science (MSc) in Data Analytics

SUMMER 2019 EXAMINATIONS

Module Code: **B9DA100**

Module Description: **Programming for Data Analysis, Processing and Visualisation**

Examiner: **John O'Sullivan**

Internal Moderator: **Darren Redmond**

External Examiner: **Dr. Andrew Parnell**

Date: Thursday, 2nd May 2019
Time: 09:30-11:30

INSTRUCTIONS TO CANDIDATES
Time allowed is 2 hours.
Answer all questions in an R script.
The marks available for each part are shown clearly: [X marks]
Comment your answer script appropriately. Put your student number as a comment at the top of each script. At the end of the exam, submit your script to the *Exam_One_40%* object on Moodle.

1. This question is worth **50 marks.**

   The dataset **Chile** is available from the **carData** package.

   a. Install the **carData** package, load the library, and access the **Chile** dataset contained in it. Load the help file and read about the dataset. In what year were the data collected?

   [6 marks]

   b. Look at the structure of the dataset and describe it briefly. What unit is the *income* variable measured in?

   [6 marks]

   c. Find the mean income of all respondents for which data is available.

   [6 marks]

   d. Create a two-way table to examine the relationship between *region* and *sex*. Describe the results. How many female respondents from the city of Santiago are there?

   [5 marks]

   e. Remove all rows with missing values to create a new dataset **Chile2**. What is the size of the dataset now? Work with this reduced dataset for the remainder of Question 1.

   [5 marks]

   f. Find the mean and standard deviation of the *age* of the respondents, grouped by the *sex* variable. Describe your findings.

   [6 marks]

   g. The *education* variable is a factor, but it is not ordered. Convert it into a sensibly ordered factor (primary education is 'less than' secondary education which is 'less than' post-secondary education).

   [8 marks]

   h. Use the aggregate function to aggregate the *income* and *age* variables by the factors of *sex* and *vote*, returning a single object.

   Which sex/vote combination has the highest mean income? What is the mean age of this group? You must use code to find these answers.

   [8 marks]

2. This question is worth **25 marks.**

   The dataset **Cars93** is available from the **MASS** library.

   a. Load the **MASS** library and access the **Cars93** dataset contained in it. Load the help file and read about the dataset. How many variables does the dataset contain?

   [4 marks]

   b. Produce a histogram of the Price variable, and make the graph look neat and presentable (paying particular attention to labels, colours, titles etc.). Comment on the resulting graph.

   [9 marks]

   c. Use the type='n' argument (or otherwise) to help you to create a scatterplot of the Length variable against the Price variable where there are three distinct groups in the plot, depending on the *DriveTrain* type of the cars.

   You should:

   - Select a different plotting character than the default pch
   - Colour the three groups differently
   - Include a legend to explain these groups
   - Include sensible x- and y-axis labels and a main title
   - Rotate the numbers on the y-axis so they appear horizontal

   Comment on the resulting graph.

   [12 marks]

3. This question is worth **25 marks.**

   The code below is used to take a matrix of numeric data, find the rows in this matrix which have at least one positive number, and then return the index of these rows.

   (To save you time, you can find the code in the file **Q3.R** contained in the **Exam** folder.)

```
fun1 <- function(mat) {

  out <- NULL

  for(i in 1:nrow(mat)) {
    if(any(mat[i,] > 0)) {
      out <- c(out, i)
    }
  }

  return(out)
}

set.seed(13)
my.mat <- matrix(rnorm(20), 10, 2)
my.mat
res1 <- fun1(my.mat)
res1
```

   a. Check how the function **fun1()** works, and add comments to explain what is happening in each line.

      [10 marks]

   b. Comment on the relative advantages and disadvantages of using the benchmark() function (from the rbenchmark library) *vs.* the Rprof() function.

      [5 marks]

   c. The function **fun1()** is quite slow. Write a faster version **fun2()** and confirm that it produces the same output. Benchmark **fun1()** and **fun2()** in order to compare their performance, and comment on the results. Copy and paste the benchmark output into your script, in order to support your answer.

      [10 marks]

END OF EXAMINATION