Module Code:	**B9DA100**
Module Description:	**Programming for Data Analysis, Processing and Visualisation**
Examiner:	**John O'Sullivan**
Internal Moderator:	**Dr. Shahram Azizi Sazi**
External Examiner:	**Prof. Andrew Parnell**

Date: *Monday, 29th July 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 your 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 `tips` is located in the `reshape2` package in R. It contains information on the tips received by a waiter while working in a restaurant for several months. Each row/record in the dataset corresponds to a different table of customers.

a. Install the `reshape2` package, load the library, and access the dataset `tips`. Include the commands needed to look at both the structure of the dataset and its help file. Summarise the dataset.

[8 marks]

b. The `smoke` variable records if there were any smokers at the table. How many tables had smokers present?

[4 marks]

c. What is the size of the largest group of diners? How many groups of this size dined at the restaurant? What was the largest bill amongst these groups?

[8 marks]

d. It's difficult to analyse the amount left as a tip, without taking the size of the corresponding bill into consideration. In order to do this, form a new column called `percentage.tip` which contains the percentage of the bill which the tip constitutes (e.g., if a bill is €50 and the tip is €5, then this new column would record that the `percentage.tip` is 10 - i.e., 10%). What is the average percentage tip?

[6 marks]

e. Which sex/day combination left the smallest mean `percentage.tip`?

[8 marks]

f. Create a new column called `rating` which converts `percentage.tip` to an ordered factor using the **cut()** function. Use bins of 0 - 10% ("Normal"), 10 - 20% ("Generous"), and 20 - 50% ("Very generous").

[8 marks]

g. The `time` variable records whether a table of diners sat at Dinner or Lunch. Make a two-way table of `time` vs. `rating`. How many tables at lunch are considered generous?

[8 marks]

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

Load the `dublin.Rdata` object, which you will find in the **Exam** folder on Moodle. It includes a dataframe with daily measurements of the following variables: mean temperature (`meantp`); maximum temperature (`maxtp`); minimum temperature (`mintp`); precipitation (`rain`); mean windspeed (`wdsp`); and sunshine duration (`sun`). Temperature is measured in degrees Celsius; precipitation in millimetres (mm); mean windspeed in knots; and sunshine duration is measured in hours.

a. A data analyst wants to study the daily mean windspeed (`wdsp`) variable. Produce a histogram of this variable, and make the graph look neat and presentable (paying particular attention to labels, colour, titles etc.).

   Comment on the resulting histogram.

   [10 marks]

b. The data analyst is also interested in learning more about the relationship between the mean temperature (`meantp`) variable and the `month` variable. Produce a plot showing boxplots of the mean temperature variable grouped by month.

   In addition to producing this plot, you should:

   - colour the boxplots
   - include x- and y-axis labels and a main title
   - make the plot look neat (e.g., change the numbers on the y-axis to appear horizontal rather than vertical)

   Comment on the resulting plot.

   [15 marks]

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

   This questions uses the `dublin.RData` object from Question 2. Reload this object. Using the class() function, assign the class `weather` to the list `dublin`.

   a. Write a `summary` method for an object of class `weather` which includes the following information in its output:

      ■ The beginning and end years in the dataset
      ■ The minimum of the `mintp` variable
      ■ The maximum of the `maxtp` variable
      ■ One other piece of information from the dataset

      Your summary method should be neat and clear and easy to read.

      Test your `summary` method on the object `dublin`.

      [17 marks]

   b. Explain **in your own words** what the following object-oriented programming (OOP) terms mean in the R language:

      ■ Polymorphism
      ■ Inheritance

      [8 marks]

   End of Exam