

|  |  |
| --- | --- |
| **CMM020: Data Visualisation and Analysis Coursework** | |
|  |  |
| **School** | Computing Science and Digital Media |
| **Course** | MSc IT (all routes)  MSc IT for the Oil and Gas Industry |
| **Year of Study** | Postgraduate |
| **Academic Year** | 2019-2020 |
| **Semester** | TWO |
| **Issue Date** | 9 April 2020 |
| **Submission deadline** | **16:00 on 11 May 2020**  **By 16:00** |
| **Feedback** | **By 8 June 2020** |

|  |
| --- |
| **Instructions to Candidates**  **Complete ALL sections of this coursework.**  Important notes:   * This is an **individual** piece of work. * You should **show all your workings/code as well as plots and critical comments/discussions as required.** * Submission details appear at end and should be read **before** you start working on the tasks. * **The grading scheme appears on page 5.** * If, for a specific task, you are unable to undertake data preparation using R but you are able prepare your data by other means, explain how you have achieved this and complete the rest of the task in R. * The Turnitin plagiarism service will be used on all submissions. |

**Software requirements**

An R programming environment such as RStudio. Details on how to access RStudio appear on page 6.

## Downloads

## Download the file *child.csv* from CampusMoodle and load it into RStudio. It is a modified dataset (adapted from the autism dataset at <https://www.kaggle.com>) which contains the following data columns (attributes/variables) for a number of children who were tested for autism.

|  |  |
| --- | --- |
| **ATTRIBUTE** | **DESCRIPTION** |
| **score** | A numeric value obtained from the standard tests for autism. |
|
| **score2** | A numeric value obtained using alternative (non-standard) tests for autism. |
|
| **age** | The child’s age. |
|
| **gender** | m or f (male or female) |
|
| **ethnicity** | Child’s ethnicity |
|
| **jaundice** | Whether the child was born with jaundice. |
|
| **cost** | Total cost of testing the child in pounds. |
| **autismFM** | Whether there is family history of autism. |
| **residence** | Country of residence. |
| **relation** | Who completed the test for the child. |
|
| **autism** | Whether the child has autism (YES or NO). |
|

## TASKS

You are required to conduct the following tasks using R where coding is required. If, for a specific task, you are unable to undertake data preparation using R but you are able prepare your data by other means, explain how you have achieved this and complete the rest of the task in R.

1. Produce a plot with the relative proportion of children residing in Australia, Germany, Italy and India. Comment on your visualisation and suggests one alternative to your plot, highlighting its advantages. There is no need to plot the alternative.

[10 marks]

1. Use univariate statistics on at least the first 4 attributes to describe the data. Discuss the results obtained, highlighting any result which you consider particularly useful. Use visualisations if needed.

[15 marks]

1. Apply data analysis techniques in order to answer each of the questions below, justifying the steps you have followed and the limitations (if any) of your analysis. If a question cannot be answered explain why.
   * Is the score mean different for children with autism and children without autism using a significance value of 0.05?
   * Is there a difference of at least 1 in mean scores between children with a family history of autism and those without a family history of autism?
   * What is the predicted value of the alternative score (score2) for a child with a standard score of 7?
   * What is the predicted value of the alternative score (score2) for a child with a standard score of 12?

[25 marks]

1. Create a dataset which contains all the data in child plus a new column “ageGroup” with values “Five and under” and “6 and over”. Use one or more visualisations to compare the standard score against the cost for each age group. The visualisation(s) should also show whether there was a family history of autism. Comment on your visualisations.

[15 marks]

1. Assume that, in addition to the child dataset supplied with this coursework (dataset 1), you also have another 19 independent datasets with the same number of observations about children tested for autism with the following distribution for attribute autism:

|  |  |
| --- | --- |
| **Dataset**  **number** | **Percentage of autism= YES** |
| 2 | 51 |
| 3 | 47 |
| 4 | 48 |
| 5 | 45 |
| 6 | 44 |
| 7 | 49 |
| 8 | 54 |
| 9 | 50 |
| 10 | 51 |
| 11 | 53 |
| 12 | 50 |
| 13 | 47 |
| 14 | 48 |
| 15 | 50 |
| 16 | 51 |
| 17 | 50 |
| 18 | 49 |
| 19 | 50 |
| 20 | 52 |

Demonstrate that the size of the confidence intervals for the average percentage of positive cases of autism (autism=YES) increases as the confidence level increases. Use 90%, 95% and 98% confidence. Discuss any improvements which may enhance your demonstration.

[15 marks]

1. Critically discuss the following statement, using at most 3 plot examples to illustrate your explanations [ Word limit 300].

“*There are different methods of disp*lay*ing data, with no method being suitable for the visualisation of all types of data. Some visualisations easily convey the information they are designed to communicate whereas others fail to adequately show the data. Data-ink ratio and lie factor also play a part in the quality of a visualisation.”*

Note: your plot examples must relate to the child dataset.

[20 marks]

**Submission**

You are required to use the drop box on the CampusMoodle to submit the following TWO files:

1. **A word, html or pdf file containing all the required information for all tasks, i.e.:**
   * **Code**
   * **Plots**
   * **Descriptions/justification of choices**
   * **Analysis of results**
   * **Explanation of any data preparation task achieved by means other than R.**

Your answers to all the tasks should be included in this file.

1. **An Rmd or R file containing all your code, suitably labelled (with the question number) and commented.**

Your two files should be submitted to the appropriate coursework box on CampusMoodle.

Note: the Turnitin plagiarism service will be used on all submissions.

**Grading**

A: Mark >=70

B: 75 > Mark >= 60

C: 63 > Mark >= 50

D: 50 > Mark >= 40

E: 40 > Mark >= 35

F: 35 > Mark

**Obtaining RStudio**

There are 2 main ways of accessing RStudio:

* Installing it in your own computer – available at [https://rstudio.com](https://rstudio.com/)

Note that you will first need to install R – available at <https://www.r-project.org/>

* Accessing Citrix RStudio from MyApps (see <https://www.rgu.ac.uk/it-helpdesk/myapps>). Note that this choice requires you to save your Rmd files into your H: drive in order to run your code. You can then copy your files elsewhere. It is recommended that within MYAPPS you use the DESKTOPS view (see below). You will not be able to install new libraries, but the libraries available are sufficient for this coursework. Note that library “here” is not available so you will need to set your working directory manually. For example, if you are working on a folder called CMM020Coursework on your H: drive you can set your working directory with setwd(“H:/CMM020Coursework”)

