ROBERT GORDON UNIVERSITY ABERDEEN

COURSEWORK OVERVIEW

Module Number & Title	CMM535 Data Science Development
Coursework Part	PARTS 1 of a 2-part coursework. Coursework PART 1 for the module will be detailed in this brief. This is an INDIVIDUAL piece of work.
Submission Method	Coursework must be submitted electronically via the designated Assessment Dropbox on CampusMoodle. Submission instructions appear at the end.
Deadline	**PART 1: 11 March 2022 @ 16:00** PART 2: 29 April 2022 @ 16:00 **The submission deadline is 4pm** – whilst a 30-minute grace period has been added for technical issues, you should aim to submit by 4pm.
Module Co-ordinator	Module Coordinator details – Ines Arana and David Lonie Email: i.arana@rgu.ac.uk , d.p.lonie@rgu.ac.uk For Staff Office Hours, visit: http://campusmoodle.rgu.ac.uk/course/view.php?id=96625
LOs Assesse	This coursework part assesses: LO1 Discuss the main concepts and tools for a data science project. LO2 Load, explore, model and visualise data using off-the-shelf tools and packages.

AUXILIARY INFORMATION

Please refer to the coursework brief below for detailed information on your submission, including any software packages/versions that should be used, word counts and limits.

Coursework received after the submission deadline indicated above will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost. Coursework extended due to extenuating circumstances shall be assessed in the normal way.

If the **word count** of an assessment is considered critical, then this will be reflected in the assessment criteria for that assessment together with any consequent penalties.

In line with the RGU Assessment Policy, **this coursework has been moderated** by the School of Computing Moderation Panel, which comprised a detailed technical review and panel overview to schedule staggered submissions.

ACADEMIC INTEGRITY

All work is expected to be completed by yourself, unless you have been clearly instructed to work as part of a team or group. Before submitting assignments, you should check your submission to ensure that it complies with Academic Regulation A3-2 Student Conduct Procedure and Academic Integrity, including but not limited to:

- *all material identified as originally from a previously published source has been properly attributed by the inclusion of an appropriate citation at the point of use in the text;*
- *direct quotations are marked as such (using "quotation marks" at the beginning and end of the selected text), and*
- *full details of the reference citations have been included in the list of references.*

EXTENSIONS AND DEFERRALS

The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so. For information about extensions and deferrals, please familiarise yourself with the extenuating circumstances defined therein.

Introduction

Aquaculture (fish farms) are an important part of the Scottish economy. Fish are grown in large cages, either in the sea or in lakes. The industry involves the use of animals and has an impact on the environment, so there is an increased focus in monitoring the fish farms, and keeping records of any incidents there may be.

In this coursework, you will use fish farming data. While the data is based on real data, some of it has been modified for this coursework and some of it is made up.

Downloads

Download the following files:

- **escapes.csv:** contains records of fish escapes, i.e. incidents where some of the fish in the cages have escaped into the wild. This is undesirable and is reported.
- **analysis.csv**: the results of water analysis using a number of components. This file has been selected to contain records which somehow align with some of the escapes data.

Software

You should use R (Rstudio) to complete this coursework. All data preparation should be undertaken in R. your code, as well as output results and discussions, should be included in an R markdown file.

If some of the data preparation task is undertaken outside R (e.g. in excel), this should be clearly stated in the report. You won't get credit for that specific data preparation task, but it will allow you to proceed with the rest of the coursework.

PART 1

In this part, you will prepare the datasets for learning before you carry out two different learning tasks. As with lots of real data, the datasets require a substantial amount of improvement before the learning can be undertaken.

Carry out the following tasks in R:

1. Prepare each individual dataset for learning. You will need to perform some exploratory data analysis in order to justify your data preparation. This may include:
    a. Cleaning the data.
    b. Transforming data / attributes.
    c. Removing data.
    d. Imputing data.
    e. Discarding features/attributes.
    f. Deleting instances.

    You will need to justify any data preparation action and any assumption made such as the learning task(s) you are preparing the data for. In the escapes file, the Escaped Species, the Age and the Average Weight are of special interest.

2. Integrate the 2 datasets together into a merged dataset which you will call *escapesPlus* and you will save to a file called **escapesPlus.csv**.
3. Undertake additional exploratory data analysis of the dataset, highlighting any interesting information. Note that data exploration may involve the application of statistical functions and/or the use of visualisations. Prepare this new dataset for learning (if needed).
4. Undertake ONE other learning task covered during the CMM535 lab sessions. The task should be different from tasks 1-3 and, ideally, it complements the work undertaken in previous tasks.

Submission

Professional presentation is part of what is being assessed, so submissions missing any of the required files will receive a limited grade (as defined in the marking criteria).

Text comments in the markdown file should be concise and relevant and included as markdown text. Irrelevant or excessive comments may be penalised.

You should submit the following files to the appropriate dropbox on CampusMoodle:

1. An **Rmd file** with ALL the code, suitably labelled and commented.
2. An **escapesPlus.csv** file containing the combined dataset ***escapesPlus***.
3. A **Word or pdf** file containing
   - Code.
   - Results and plots.
   - Descriptions/justification of choices and discussions.
   - Critical discussion 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.

   This file should normally be generated by knitting the Rmd file.

   PAGE LIMIT: 25 pages.

Grading

Grade	Criteria (award highest grade for which the work meets the requirements
A	An Rmd, a csv and a word/pdf file are submitted containing ALL the required information and data. The page limit for the word/pdf file is respected. The exploratory data analysis carried out is thorough and is explained clearly and concisely with the use of visualisations where needed. The data is correctly prepared for learning and the actions taken to obtain it are clearly explained. An excellent critical discussion of options available and justification for decisions taken is included. A merged dataset is created and prepared appropriately. The actions taken are clearly explained and justified. Data processing is reproducible. The additional task adds value to the work carried out in previous tasks. The task is well justified and the results critically evaluated. Excellent report presentation with critical justification for choices.
B	An Rmd a csv and a word/pdf file are submitted containing ALL the required information and data. The page limit for the word/pdf file is respected. The exploratory data analysis carried out is thorough and is explained clearly and concisely with the use of visualisations where needed. The data is correctly prepared for learning with some deficiencies and the actions taken to obtain it are clearly explained. A strong critical discussion of options available and justification for decisions taken is included. A merged dataset is created and prepared appropriately with a few deficiencies. The actions taken are explained and justified. Data processing is reproducible. The additional task is well justified, and the results critically evaluated with a few deficiencies. Very good report presentation with critical justification for choices.
C	An Rmd, a csv and a word/pdf file are submitted containing ALL the required information and data. The page limit for the word/pdf file is respected. The exploratory data analysis carried out at least moderately thorough and is explained clearly and concisely with the use of visualisations where needed. The data is correctly prepared for learning with some deficiencies. There is some justification for the actions taken. A merged dataset is created and prepared, possibly with some deficiencies. The actions taken are explained and justified. Data processing ais reproducible. The additional task has been carried out and there is some evaluation of results with some deficiencies. Good report presentation with some justification for choices.
D	An Rmd, a csv and a word/pdf file are submitted containing ALL the required information. The page limit for the word/pdf file is respected. The exploratory data analysis carried out is mainly correct and includes a few key highlights. Some other key highlights might have been omitted. At least one visualisation is used.

	Some of the data has been prepared for learning, possibly with a few errors. Some of the data might have been prepared outside R. with some deficiencies and the actions taken are clearly explained. A shallow justification for decisions taken is included, possibly with some errors. Significant progress is made towards obtaining a merged dataset with some deficiencies. The actions taken are explained and justified. The additional task may have errors and the discussion of results may lack critical appraisal. Adequate report presentation.
E	An Rmd and a word/pdf file are submitted. The page limit for the word/pdf file is respected. The exploratory data analysis carried out is limited with serious errors/omissions. Some of the data is prepared for learning, possibly with some errors. Some of the data might have been prepared outside R with some deficiencies. The discussion of options may be missing and justification for decisions taken is included. A merged dataset may not have been created due to code errors or may have serious deficiencies. The additional task imay have serious deficiencies.
F	Very limited effort.

GRADE COMBINATION OF PARTS 1 AND 2

The grades from Part 1 and Part 2 of the CMM535 assessment are each awarded a grade A -F. These are combined as follows to give the overall CMM535 module grade. The final grade for each pair of grades combination is as follows:

The minimum requirements for a grade are as follows:

- Grade A: 2 As.
- Grade B: 1B + 1C.
- Grade C: 1C + 1D.
- Grade D: 1D + 1E.
- Grade E: 1E + 1F
- Grade F: where none of the above requirements are met.