COMP1832	**Programming Fundamentals forData Science**	**Faculty HeaderID:**	**Contribution 100% of course**
Course Leader **Dr. Jia Wang**	**COMP1832 Portfolio (resit)**		**Deadline Date:** **April 12, 2024** **23:30 UK time**
This coursework should take an average student who is up-to-date with tutorial work approximately 50 hours Feedback and grades are normally made available within 15 working days of the coursework deadline			
Learning Outcomes: 1. Demonstrate knowledge and understanding of commonly used data structures and processing techniques for Data Science. 2. Understand and implement processing pipelines for use in Data Science applications. 3. Build efficient solutions for data manipulation.			

Plagiarism is presenting somebody else's work as your own. It includes: copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as your own work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by theUniversity. Please see your student handbook for further details of what is / isn'tplagiarism.

All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.

Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.

Coursework Submission Requirements

- An electronic copy of your work for this coursework must be fully uploaded on **Friday 12/04/2024 by 11:30pm UK time** using the links on the module Moodle page.
- For this coursework you must submit **a single PDF** document which include solutions to both Python and R portfolio tasks. All resulting plots are required to provide plot descriptions and explanations. In addition, please also upload **a single .zip file** including both Python and R source code. **Your source code files must be error-free (does not require debugging to run the code).**
- Make sure that any files you upload are virus-free and not protected by a password or corrupted otherwise they will be treated as null submissions.
- You must NOT submit a paper copy of this coursework.
- All coursework must be submitted as above. Under no circumstances can they be accepted by academic staff.

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See http://www2.gre.ac.uk/current- students/regs

Detailed Coursework Specification

This Coursework is to be completed individually.

Portfolio Tasks

<<<<<<<<<<<<<<<<<<<<<Python part>>>>>>>>>>>>>>>>>>>>>

The topic of the following tasks is processing and visualising one-dimensional and multidimensional data. The coursework will generate data visualisation products, which allow the customers or the users of these products to obtain insights from the data. All tasks are expected to be developed by using the Python programming language. Any integrated development environment for Python can be utilised for the development of the coursework.

Task 1 (15 marks)

Dataset: Download the Iris dataset from the Moodle page of the module. The dataset is organised as a single Comma Separated Values (CSV) file. The file contains data about three species of iris.

Read the dataset into a suitable data structure and print on the screen its first 10 rows **(5 marks)**.

By considering only the Petal width, obtain the following statistical information for each one of the three species of iris from the dataset:

- Measures of Central Tendency (mean and median) **(5 marks)**
- Measures of Dispersion (standards deviation, range) **(5 marks)**

Provide a complete Python code and the resulting statistical values.

Task 2 (15 marks)

Identify which one of the following types of charts would provide optimal visualisation for a single attribute of a single species from the Iris dataset **(5 marks)**:

- Line Chart
- Bar Chart
- Box Plot Chart

Justify your choice with maximum 5 sentences **(5 marks)**. Choose one species and one attribute from the Iris dataset and visualise its values by using the selected type of chart **(5 marks)**. In this case, the data will be one-dimensional. Provide the Python code and the resulting visualisation.

Task 3 (20 marks)

Visualise the entire Iris dataset by using Parallel Coordinates diagram. Provide the Python code and the resulting visualisation. The result will be evaluated as follows:

Use different colours for the multi-lines **(5 mark)**. Add appropriate labels explaining the elements of the chart **(5 mark)**. By observing the Parallel Coordinates diagram, identify which property can be used to classify the different species of iris and justify your choice with maximum 3 sentences **(5 mark)**. Add appropriate comments to the source code explaining the functionality of the program **(5 mark)**.

<<<<<<<<<<<<<<<<<<<<<R part>>>>>>>>>>>>>>>>>>>>>

Task 1 Basic Statistics and plotting using R (25 mark)

For this task, we'll use the built-in R dataset named "Nile" - no need to download any data.

Q1. Compute the mean, median, mode, variance, and standard deviation of the dataset (**5 mark**)

Q2. Compute how "spread out" the data are. Here you need to calculate the minimum, maximum and range (**3 mark**).

Q3. Calculate the interquartile (IQR) range (**1 mark**). Use the function quantile() to measure quantiles for the same dataset, and comment on the difference and relation of these two functions (**1 mark**).

Q4. Use the in-built R basic functions (no need to import any library) to create a histogram. Make sure you add the following arguments (**4 mark**) and interpret this histogram regarding the frequency and variability **(4 mark).**

main: Add a title for this plot to reflect what Nile dataset is about (avoid label it as Nile) as well as the type of the plot created.

xlab: Add a meaningful label for the x axis

ylab: Add a meaningful label for the y axis

col: set a colour of the bars

Q5. Use qqnorm() and qqplot()to produce quantile-quantile plot.

You will need to set the reference line colour as blue and its width as 3 (**2 marks**).

Interpret the plot (what are the points, and what is the purpose of the line?) and comment on normality of the dataset (**2 marks**).

Q6. Use plot() to further explore the dataset including arguments such as xlab, ylab, main and type (**3 marks**).

Task 2 Visualisation with ggplot2 (25 marks)

Dataset

The dataset *mpg* used in this task is a data frame and can be found in the package ggplot2 (aka ggplot2::mpg).

For all three questions, you will need to provide:

- **code used to generate the plot (screenshot of code not allowed)**
- **high resolution plot with title, and labels for x, y axes (screenshot of plot not allowed)**
- **plot interpretations**
- **answers to the questions raised by the task**

Q1. Plot and explain: Which vehicle brand (or *manufacturer*), offers the best *mpg* in both city and in the highway? (**5 marks**)

Q2. Plot and explain: Which type of car, regarding their *displ* range (size of engine) has the lowest *mpg* in the city categorised by the vehicle type (e.g., compact, suv or 2seaters defined in the variable *class*)? Display the resulting plot categorised by the vehicle type. (**8 marks**) **Hint: facet_wrap() for the categorisation.**

Q3. Plot and explain: Which type of car, regarding their *displ* range (size of engine) has the best *mpg* performance in both city and highway?

Display the resulting plot categorised by the number of cylinders and the drive type (where f = front-wheel drive, r = rear wheel drive, 4 = 4wd). If you are a buyer who wants a high litre engine vehicle and drives mostly in the highway, which type of car would you choose? (**12 marks**) **Hint: facet_grid() for the categorisation.**

Grading Criteria

80%-100% Exceptional	You will need to have: an excellent implementation and reflection on understanding your tasks. All requirements are implemented to a higher standard.
70-79% Excellent	A very good implementation showing your solutions with all requirements implemented. Code and plots are clear and readable with justified descriptions and explanations.
60-69% Very good	A good implementation showing your solutions to all the tasks: all required plots are implemented, and both python and R code are working. Explanations of your solutions are provided which reflect good understanding of given data analytics tasks.
50-59% Good	An implementation showing your basic understanding and programming skills of data transformation, basic statistical analyses, and visualisation. Providing solutions with minimum requirements implemented with some justifications.
0-49% Fail	A portfolio with very limited tasks solved. No solution or very few solutions provided for the assignment. A portfolio that fails in reflecting the understanding of the basics of processing and visualising data in Python and R.