Assessment 2 CSE5DEV: Data Exploration and Analysis

Assignment Type	:	Coding
Weighting	:	25%
Due Date	:	Sunday, 11 October 2024 23:59 (Melbourne time)
Submitted files	:	.PDF, .Rmd, .html

Topic Overview

This assignment tests your knowledge and ability to understand the problem with the data, performs simple data preprocessing, and evaluates the impact of data preprocessing on data analysis. For this assignment, you will be given datasets for data exploration and analysis. You will utilise R as the primary tool, and your analytical, coding, creative and logical thinking skills will be used to overcome the challenges in this assignment.

Assessment Criteria

This assessment will measure your ability to:

- Accurately guide the AI-based tool (ChatGPT) to help you work with simple data preparation and analysis. **ChatGPT can only be used for Task 1**.
- Accurately apply data analysis concepts and implement steps as instructed to produce the correct output.
- Provide logical explanations with your findings.

Guidelines

- Use only the provided R Markdown template file for this assignment.
- ChatGPT can only be used for Task 1.
- Implement each question in a chunk (block) of code and provide a clear remark for the codes.
- A Significant deduction will be applied for any chunk without remark.
- You can only use libraries from the lecture notes or lab
- Pay attention to the number of words required for each explanation.
- You need to submit the R Markdown file and the knitted HTML file.

Submission format

Upload 3 files in the LMS submission space.

- A PDF file for your ChatGPT conversation with a valid URL link.
- An R Markdown file, and
- The generated HTML file

Inquiry about the assignment

Please send your inquiry regarding this assignment to:

- Instance coordinator Bendigo: Choiru Za'in (c.zain@latrobe.edu.au)
- Instance Coordinator Bundoora / Subject Coordinator: Kiki Adhinugraha (k.adhinugraha@latrobe.edu.au)

Detailed Instruction

Task 1: 50 Marks
Dataset: Studentmarks.csv

AI has been widely implemented in many applications and can help us solve problems efficiently. Despite the high benefits, AI-assisted tools can only produce the right response if the users can provide clear guidance and direction.
In this task, you must guide **ChatGPT in English** to create an R script that produces the right output. Remember the following notes:

- **Make sure you create a new topic to discuss Task 1 exclusively.**
- You must print the ChatGPT discussion in PDF, and share the conversation link in R Markdown. Check the following link on how to share and export your ChatGPT conversation topic. https://sway.cloud.microsoft/oYavOJKbeHY4ibSu?ref=Link
- Using a Screenshot to provide ChatGPT conversation **IS NOT ACCEPTABLE**.
- If you use multiple ChatGPT conversations, you must provide all shared links in the R file. You must merge all conversations into one PDF file.
- **You cannot alter the code from ChatGPT**. The code must be copied as is.
- You must put the R code in the provided R markdown file, and the R code must produce the identical output as shown in the task description below.
- Put each task in a separate chunk. Explain how the code works using your own words.
- Your work will be marked based on your guidance's authenticity and the correctness of the code and output.

Task description: You have been given a csv file named 'Studentmarks.csv' containing the annual student academic performance from 2020 until 2022. The file consists of six features, which are 'StudentID', 'Studentname', 'dob', '2021', '2022', and '2022'. Using ChatGPT, your task is specified as follow:

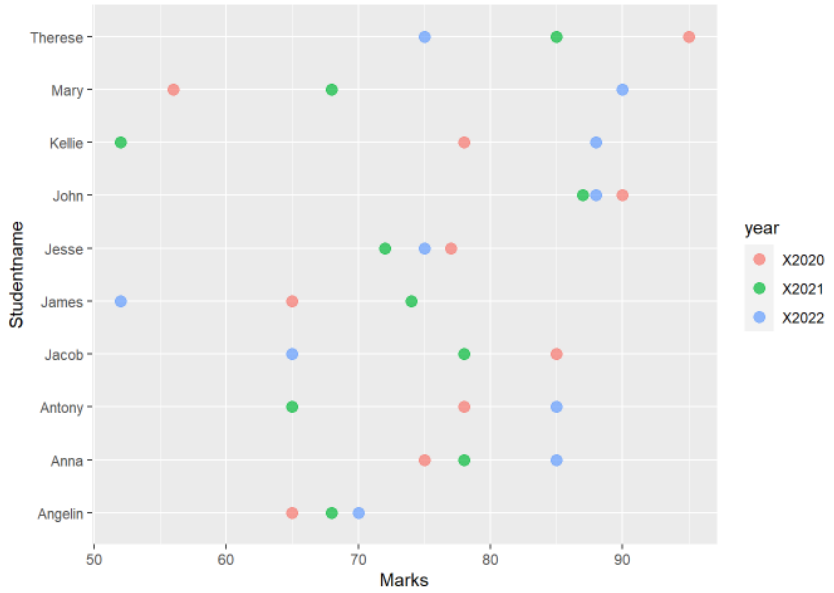
a. [20] Guide ChatGPT to create the code that will produce the following dataframe. Using your own words, please provide a 50-word explanation of how the code works.

```
##    StudentID Studentname date month year X2020 X2021 X2022     age1 age2
## 1          1        Anna   12    12 1998    75    78    85 24.79016   25
## 2          2       James   10    08 1999    65    74    52 24.13124   24
## 3          3        Mary   21    08 1998    56    68    90 25.09911   25
## 4          4      Antony   02    05 1999    78    65    85 24.40465   24
## 5          5       Jacob   22    07 1998    85    78    65 25.18113   25
## 6          6     Angelin   11    05 1998    65    68    70 25.37799   25
## 7          7      Kellie   19    04 1999    78    52    88 24.44019   24
## 8          8       Jesse   27    06 1998    77    72    75 25.24949   25
## 9          9        John   15    10 1998    90    87    88 24.94874   25
## 10        10     Therese   06    11 1998    95    85    75 24.88859   25
```

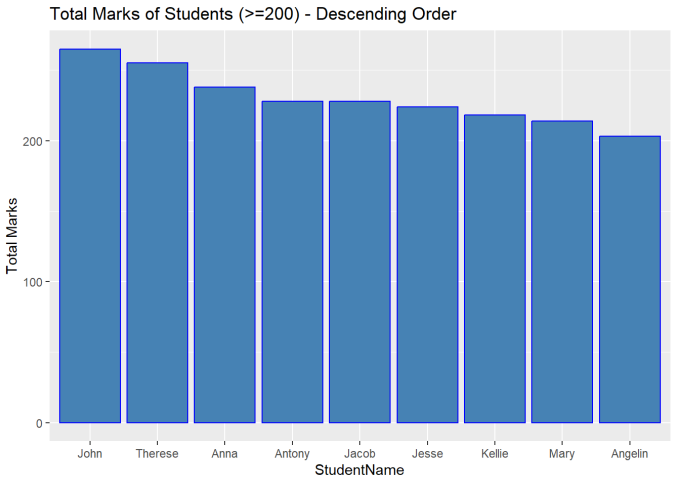
Notes:

- age1 : calculate the age using the current date. Please note that the value in age1 here might be different from yours.
- age2: Calculate the age using the year only.

b. [15] Guide ChatGPT to create the code to produce a scatter plot for Studentname versus Marks of all three years. Using your own words, please provide a 50-word explanation of how the code works.



c. [15] Student performance rank for students who get at least 200 marks in total. Using your own words, please provide a 50-word explanation of how the code works.



Task 2: 50 marks
Dataset: dirty_iris.csv

Warning: using AI-assisted tools to answer the following task is prohibited.

Preprocessing is one of the most critical tasks in data analysis due to the nature of raw data, which tends to be dirty. The dirtiness may vary from invalid value to missing value.

When substituting the empty data, it's hard to tell if the substitute data reflects the original data. In a Machine Learning environment, removing a value, applying a particular method to restore or predict the missing value, and comparing it with the original is called Supervised Learning. In this task, you will demonstrate the effectiveness of your NA

handling method on different data types. You must explain your method to overcome the missing data issue and its logical reason.

The evaluation will focus on statistical analysis, where deviations in the restored value will be compared with the original value. The mark will be given based on how far the restored value is from the original one and the impact of the substituted value on the overall trend of the data (if applicable). You can only use a statistical-based approach to restore the missing value. Some approaches you can use include but are not limited to Global/Local Aggregation (Mean, Median, Mode, Min, Max, Previous Value, Next Value, etc). Any Machine Learning methods are strictly prohibited.

a. [10] Load the data and identify the **NA**s features in the data. Provide an R code to show the statistical information about the dataset. Please follow the procedure to determine the actual number of NAs in the dataset.
   Identify the feature with NA, the data type and the strategy you want to apply to overcome the NA issue. Only statistical & non-ML methods are allowed. Write a 50-word explanation for your plan to overcome the NA in the dataset.

b. [15] Implement the approach for numerical data in R. Provide a statistical and graphical comparison with the clean data (specific and overall deviation) for this feature. Provide a 50-word explanation of your findings.

c. [15] Implement the approach for categorical data in R. Provide statistical and graphical comparison with the clean data (specific & overall deviation) to this feature. Provide a 50-word explanation for your findings.

d. [10] Perform a bivariate analysis for the numerical features of your choice using the original dataset and modified dataset in separate charts. Provide R code to compare the trends between these two charts. Provide a 50-word explanation about why you evaluate those features and your method's impact on overcoming NA issues with the bivariate analysis result.

This is an individual Assignment. You are not permitted to work as a group when writing this assignment.

Copying, Plagiarism: Plagiarism is the submission of **AI-generated content other than in Task 1 or somebody else's work** in a manner that gives the impression that the work is your own. The Department of Computer Science and Information Technology treats plagiarism very seriously. When it is detected, penalties are strictly imposed. Students are referred to the Department of Computer Science and Information Technology's Handbook and policy documents with regard to plagiarism and assignment return, and also to the section of 'Academic Integrity' on the subject learning guide.

No extensions will be given: Penalties are applied to late assignments (5% of total assignment mark given is deducted per day, accepted up to 5 days after the due date only). If there are circumstances that prevent the assignment being submitted on time, an application for special consideration may be made. See Student Handbook for details. Note that delays caused by computer downtime cannot be accepted as a valid reason for a late submission without penalty. Students must plan their work to allow for both scheduled and unscheduled downtime.

Academic integrity and plagiarism
'Academic integrity means being honest in academic work and taking responsibility for learning the conventions of scholarship . . . Academic integrity education is integral to the learning experience at La Trobe University . . . The University requires its academic staff and students to observe the highest ethical standards in all aspects of academic work, and it demonstrates its commitment to these values by awarding due credit for honestly conducted scholarly work, and by penalising academic misconduct and all forms of cheating' (La Trobe University, n.d.).

The penalty for submitting an assignment as your own that is the work of a third-party may be as severe as 'exclusion from the University without readmission' (La Trobe University, 2020, p. 4).

Refer to the *Academic integrity – schedule of responses and penalties for academic misconduct*.

You should familiarise yourself with the Academic integrity website and complete the academic integrity module (AIM) in your LMS.

If you have any questions regarding academic integrity, your Subject or Course Coordinator will be able to assist.

Students are also referred to the Department of Computer Science and Computer Engineering's Handbook and policy documents about plagiarism and assignment return, and to the document on 'Academic Misconduct' in the subject learning guide.

https://latrobe.libguides.com/academic-integrity