RMIT Vietnam University
School of Science, Engineering and Technology

COSC2999/2789/2809
Practical Data Science (with Python)
Assignment 1: Data Preparation and Exploration

Due on 22nd November 2025 (Week 4)
This assignment is worth 25% of your overall mark.

Introduction

In this assignment, you will examine a data file and carry out the first steps of the data science process, including cleaning and exploring the data. You are required to develop and implement appropriate steps, in IPython (Jupyter Notebook), to load a data file into memory, clean, process, and analyze it. This assignment is intended to give you practical experience with the typical first steps of the data science workflow.

The "Practical Data Science with Python" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis - it is your responsibility to stay informed regarding any announcements or changes.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: Jupyter Notebook (or Jupyter Lab) on Lab PCs or your laptop.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarized will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalized; there are no exceptions and no excuses. More information on Academic Integrity is available at

https://www.rmit.edu.vn/students/my-studies/assessment-and-results/academic-integrity

Note that "**Any ideas or outputs generated by AI must be referenced accurately in your academic work, otherwise it is considered plagiarism**. This means you must use citations crediting the tool/s where appropriate, as well as including it in your reference list – just like you would any other resource. See the Library's AI referencing guide for specific AI referencing information."

General Requirements

This section contains information about the general requirements that your assignment must meet. Please read all requirements carefully before you start.

- You must do the analysis in Python Jupyter Notebook/Jupyter Lab.
- Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e., if it is specified that the file name is gryphon, then that is exactly the file name you should submit; Gryphon, GRYPHON, griffin, and anything else but gryphon will be rejected.

For **tasks 1 to 3** you will use the data file **student_data_25s3.csv** given with this specification. You can find it under the Assignments/ Assignment 1 section of the course Canvas.

Task 1: Data Preparation (5%)

For this assignment, you need to work with data file **student_data_25s3.csv**, which is available in Canvas under the **Assignments / Assignment 1** section of the course Canvas. The data file student_data_25s3 used in this Assignment is a modified version of the Student Performance dataset from UCI Machine Learning Repository: https://doi.org/10.24432/C5TG7T. A description of the dataset and attributes is given on the website (link). Note that not all attributes in the original dataset are used in the dataset for this assignment.

Dataset Information and attribute Information: please read the information given in the link.

In a data science project, it is vital to set the goal of the project, then thoroughly pre-process any available data (each attribute) before the next steps. You will start by loading the data from the file. Then, you need to clean the data by dealing with potential issues (or errors, such as typos, impossible values, missing values, etc.) in the data appropriately.

Task 2: Data Exploration (5%)

Carry out the following tasks to explore the data provided.

2.1. Choose 3 variables (attributes), 1 attribute with nominal values, 1 attribute with ordinal values, and 1 attribute with numerical values. For each selected variable, construct an appropriate visualization using a suitable type of graph. Provide a concise explanation for each variable, addressing the following points: The rationale for selecting the variable (attribute); and the key observations revealed by the visualization.

2.2. Explore the relationships between attributes. You need to choose 3 pairs of columns to focus on, and you need to generate 1 visualization for each pair. Each pair of columns that you choose should address a plausible hypothesis for the data concerned. Give a short description for each hypothesis.

2.3. Build a scatter matrix for 3 numerical columns. Note, each visualization (graph) should be complete and informative in itself and should be clear for readers to read and obtain information.

Task 3: Analysis of Missing Values and Outliers (5%)

3.1. In Task 1 you checked whether the loaded data has any missing values. You might ignore all observations containing missing values or replace them with a fixed value. Now, your task is to deal with the missing values in the data set using other techniques, such as:

- replacing them with the column-wise mean value using an appropriate function,

- replacing them with the median value (column-wise).

For each of these approaches, you will create a new data file in which the missing values are replaced with new values. Save these newly generated data files with names student_fix1.csv, student_fix2.csv, etc. You must submit these newly generated data files with a clear explanation for each file in your final report.

For each of these approaches, choose a data column, and produce a new graph (corresponding to the initial graph that you produced in Task 2.1. Under each one, briefly discuss the impact that the different approaches to dealing with missing values have on what you observe from the visualization.

3.2. Give a brief explanation for each of the following questions:

- Do outliers affect standard deviation?

- When should an outlier not be removed?

- Consider the numeric columns in the given file data, are there outliers in these columns of the given data set? How do you detect them?

3.3. [Only students registered for course **COSC2999** are required to answer this question]

What can be hidden data quality issues? What approaches can be applied to detect those issues before formal preprocessing begins?

Task 4: External data (5%)

In this Task you will work with the MovieLens 100K dataset given in this link:

https://grouplens.org/datasets/movielens/100k/

Your task is to retrieve two data files named **u.data** and **u.item** from this data source. You will combine (merge) the 2 files to form one data file, then perform the first steps of exploring the data to find potential issues and prepare it (make it ready) for the next steps of a data science process.

Task 5: Report (5%)

Write your report within the Jupyter notebook using proper format either with HTML or with Jupyter Notebook formatting guide (or a word processor), then save it in a PDF format with a file named **report.pdf**. All the text within the Jupyter Notebook and the report must **not be in red** color as the color will be used for marking. Penalties will apply if the report does not satisfy the requirements. Moreover, the quality of the report will be considered, e.g. clarity, grammar mistakes, etc. Remember to clearly cite any sources (including books, research papers, course notes, data, AI tools, etc.) that you referred to while designing aspects of your solution.

- Create a heading called "**Data Preparation**" in your report. Create a sub-section with corresponding numbering when needed. Provide a brief explanation of how you addressed the task and explain any choices that you made (if appropriate). As part of this assignment, you must specifically list any data rows that you changed. For the potential issues/errors of the data you must deal with typos, extra whitespaces, sanity checks for impossible values, missing values, etc.
- Create a heading called "**Data Exploration**" in your report. For each numbered step in Task 2, create a sub-section with corresponding numbering.
    - In sub-section 1, include all your graphs from Task 2.1. Under each graph, include a brief explanation of why you chose this graph type(s) to represent the data in a particular column.

  - In sub-section 2, include your plots from Task 2.2. With each plot, state the hypothesis that you are investigating. Then, briefly discuss any interesting relationships (or lack of relationships) that you can observe from your visualization.
  - In sub-section 3, present your scatter matrix and analyze what you observe from the graph.
- Create a heading called "**Analysis of Missing Values and Outliers**" in your report. For each numbered step in Task 3, create a sub-section with corresponding numbering.
  - In sub-section 1, include a brief explanation of each approach for dealing with missing values, include all of your newly produced graphs, data files, and discussions.
  - In sub-section 2, include an answer and a brief explanation of each given question, and your approach for dealing with outliers in the given data set.
- Create a heading called "**External data**" for Task 4. Number steps that you use to solve this task.
- Length of the report: the length of your report should be less than 15 pages (A4 page size with standard font size and margins).

Important information

Corrections: From time to time, students or staff find errors (e.g., typos, unclear instructions, etc.) in the assignment specification. In that case, a corrected version of this file will be produced, announced, and distributed to all students. Because of that, you are NOT to modify the given files in any way to avoid conflicts.

Late submissions & extensions: A penalty of 10% of the maximum mark per day will apply to late assignments up to a maximum of five days, and 100% penalty thereafter (Extensions will only be permitted in *exceptional* circumstances under the University's rules).

Academic Dishonesty: This is an advanced course, so we expect full professionalism and ethical conduct. Plagiarism is a serious offense. We will pursue the strongest consequences available according to the **University Academic Integrity policy**. In a nutshell, **never look at a solution done by others**, either in (e.g., classmate) or outside (e.g., web, **AI tools**) the course.

Silent Policy: A silent policy will take effect **24 hours** before this assignment is due. This means that no question about this assignment will be answered, whether it is asked on the newsgroup, by email, or in person.

Code of Honour

We expect every RMIT student taking this course to adhere to the **Code of Honour** under which every learner-student should:

- Submit their own original work.
- Do not share answers with others.
- Report suspected violations.
- Not engage in any other activities that will dishonestly improve their results or dishonestly improve or damage the results of others.

What to Submit, When, and How

The assignment is due at 21:00, the 22nd, November 2025 in Week 4. Solutions submitted after this time will be subject to standard late submission penalties. Requests for special consideration must be provided in written form and must follow the University regulations.

You need to submit the following files:

- One notebook file containing your Python commands for Task 1, Task 2, Task 3, and Task 4, “assignment1.ipynb”. Please use the provided solution template (assignment1_TEMPLATE.ipynb) to organize your solutions. For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells).
- One report file in PDF format.
- The generated datasets and python file(s) must be included in the submission. A README.txt file should be included if needed.

The files must be submitted in ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas: Assignments/Assignment 1. Please do NOT submit other unnecessary files.