School of Computing Technologies

COSC2791 | Practical Data Science with Python

Assessment 1: Dataset preparation report

Key assessment information

Weighting: 25%

Due Date: Sunday Week 3, 23:59PM (Melbourne time)

Assessment type: Written report (submitted with code file and dataset file)

Word limit: Up to 6 pages (report only)

Overview

In this assessment you'll replicate the first few steps of a standard data science project. Using the case study and raw dataset provided, you'll need to define the research goal, retrieve the data, and prepare it for preliminary analysis. You'll document your thinking and process in a report that you'll use to uncover the business problem. Your report and supporting materials will also form the groundwork that enables this fictional organisation to reproduce your same process on similar types of data they have collected.

Purpose

This assessment will give you transferrable knowledge and skills in data science that you can apply to future data-driven projects. Having clean data enables you to execute the next steps in the data science process. Without clean data you may form wrong and meaningless conclusions.

What do you need to deliver?

For this assessment you'll need to produce three separate deliverables:

Deliverable 1: Code
Write and save your code in notebook format (.ipynb file) that you'll use to clean the data.

Deliverable 2: Dataset
Provide a cleaned dataset as a .csv file.

Deliverable 3: Report
Write a PDF report using the supplied template to document and justify your process.

Supporting materials

To assist you with completing your assessment, you'll need to:

- Read the supplied case study as it includes questions you must address in your report
- Review the linked Jupyter Notebook (which includes your dataset)
- Use the supplied word template to complete your report

Refer to the 'Assessment details' section below for more information about your supporting materials.

Tools required for this assessment

You're encouraged to use the following tools to develop and complete your assessment deliverables:

- **Jupyter Notebook** to write code and modify and clean your dataset
- **Microsoft Office suite** to complete your .pdf report

Marking criteria

Your assessment will be marked according to the following criteria:

- Prepare raw data to ensure it is clean and free of errors (5 marks)
- Conduct exploratory data analysis in Python to form an interpretation of data (7 marks)
- Document and justify the process by which you cleaned, analysed, and interpreted the data (10marks)
- Apply professional standards for reproducibility of analysis (3 marks)

Course learning outcomes

This assessment is relevant to the following course learning outcomes:

CLO1	Use industry and evidence-based tools and approaches to transform raw data into a format suitable for a data science pipeline
CLO3	Generate an interpretation and visualisation of data using exploratory data analysis in Python
CLO4	Construct and document an experimental methodology for analysis of data
CLO6	Apply professional standards to allow reproducibility of analysis

Assessment details

To successfully complete this assessment, you'll need to check the case study page in Canvas and undertake the following activities:

- Read the related case study and pay particular attention to the question(s) provided
- Access the dataset through the linked Jupyter notebook in Canvas
- Write your code in the notebook file that you'll use to develop your solutions for your presentation
- Download and use the report template and follow its structure
- Generate an error-free dataset (.csv file) and include it with your submission (see instructions in the Getting Started Guides, and your notebook file)
- Answer the questions provided with the case study
- Prepare a report (.pdf file) of up to 6 pages (including figures and references) with a font size between 10-12 points
- Submit your work as per the submission guidelines

Refer to the rubric table below for a detailed breakdown of what tasks are expected as well as a clear guide as to how your work will be assessed by your facilitator.

Submission format

Follow these steps when submitting your work:

- Go to the Assessment 1 page in the Assignments section in Canvas.
- Click on 'Start Assignment' at the top right corner and upload one zip file named as your student number, e.g., s1234567.zip.
- Your zip file must include the following files with case sensitive names and formats:
  - Deliverable 1 – Code: notebook file saved as '**assignment1.ipynb**'
  - Deliverable 2 – Dataset: cleaned data saved as '**assignment1.csv**'
  - Deliverable 3 – Report: document saved as '**report.pdf**'

Course activities you'll need to complete for this assessment

The following sections in Canvas build the skills you'll need to complete your assessment if you would like to go back and review before beginning the assessment:

- 1.3.0 Activity: Data curation: reading data
- 1.4.0 Activity: Data preparation
- 2.1.0 Activity: Data summarisation
- 2.3.0 Activity: Descriptive statistics and data visualisation

Referencing guidelines

Use RMIT Harvard referencing style for this assessment.

You must acknowledge all sources of information you have used in your assessments.

Refer to the RMIT Easy Cite referencing tool to see examples and tips on how to reference in the appropriated style. You can also refer to the library referencing page for more tools such as EndNote, referencing tutorials and referencing guides for printing.

Academic integrity and plagiarism

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas.

You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct.

Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the University website.

Assessment declaration

When you submit work electronically, you agree to the assessment declaration.

Criteria	Ratings						Pts
	HD	D	C	P	N	DNS	
Preparation of raw data Prepare raw data to ensure it is clean and free of errors. Marks will be awarded as follows: 1. **Data Retrieving (0.5 points)** Load the CSV data from the file. Use an appropriate pandas function to load the csv data, and make use of the correct arguments including sep, decimal, header, names, if needed. 2. **Check data types (0.5 points)** Check whether the loaded data is equivalent to the data in the source (CSV) file. You will need to ensure that the loaded data has appropriate data types assigned or take steps to ensure that the appropriate types are used. 3. **Typos (0.5 points)** Check whether there are typos in the data. If there are any typos, correct them by using masks. 4. **Extra white spaces (0.5 points)** Check whether there are instances of extra whitespaces in the data, and if so, demonstrate how to remove them by calling on an appropriate function. 5. **Upper/Lower-case (0.5 points)** Cast all text data to upper-case by using an appropriate function. 6. **Sanity checks (1 point)** Design and run a small test-suite, consisting of a series of sanity checks to test for the presence of impossible values for each attribute. 7. **Missing values (1.5 points)** Check whether the loaded data has any missing values. If so, use an appropriate function to replace them with one of the following values: - a fixed value - the column-wise median value - the column-wise mean value - or ignoring all observations containing missing values.							
	5 to >3.99 pts	3.99 to >3.49 pts	3.49 to >2.99 pts	2.99 to >2.49 pts	2.49 to >0 pts	0 pts	5 pts

Exploratory data analysis

The appropriate columns were used and selected from the data to conduct exploratory data analysis. Marks will be awarded as follows:

1. **Selecting three columns for analysis (3 points)**
   Choose one column each for nominal, ordinal, and interval/ratio values. Then, create a visualisation for each of them: You should explore each column with the appropriate type of graphs, for example histograms, barcharts, pie graphs, scatter plots, boxplots or scatter matrix. Format each graph carefully.
   You need to include appropriate labels on the x-axis and y-axis, a title, and a legend. The fonts should be sized for good readability. Components of the graphs should be coloured appropriately, if applicable.

2. **Relationships between columns (4 points)**
   Explore the relationships between columns. You may choose which pairs of columns to focus on, but you need to generate three visualisations for this subtask. These should address a plausible hypothesis for the data concerned. Please format each graph carefully.

	7 to >5.59 pts	**5.59 to >4.89 pts**	**4.89 to >4.19 pts**	**4.19 to >3.49 pts**	**3.49 to >0 pts**	**0 pts**	**7 pts**

Document and justify the process

Document and justify the process by which you cleaned, analysed, and interpreted the data. Marks will be awarded as follows:

1. **Data Preparation (2 points)**
   Use the student template to describe the data and the process you used to prepare it. Explain any choices that you made (if appropriate).

2. **Data Exploration (3 points)**
   Use the student template to describe the process you used to explore your data and explain any choices that you made (if appropriate, e.g. choice of graph type(s) to represent the data in a particular column).

3. **Plots (4 points)**
   Include each plot from your exploration and state the questions that you are investigating. Then, briefly discuss any interesting relationships (or lack of relationships) that you can observe from your visualisation.

4. **Structure and referencing (1 point)**
   Presents the main points in a logical and clear way and follows the report structure.
   Consistently uses accurate references, appropriately positioned.
   Communicates meaning through use of clear and unambiguous language.

	10 to >7.99 pts	**7.99 to >6.99 pts**	**6.99 to >5.99 pts**	**5.99 to >4.99 pts**	**4.99 to >0 pts**	**0 pts**	**10 pts**

Reproducibility of standards

Apply professional standards for reproducibility of analysis. The submitted assignment code should work for the same case study even with a new set of data. If the data is for the given case study, the code would work and does not need any manual interaction/correction/revision. (3 points)

3 to >2.39 pts	2.39 to >2.09 pts	2.09 to >1.79 pts	1.79 to >1.49 pts	1.49 to >0 pts	0 pts	3pts
				Total:		**25 pts**