RMIT UNIVERSITY

Computer Science & IT, School of Science

COSC2791 Practical Data Science with Python

Assessment 1: Data Cleaning and Summarising

Assessment Type: Practical assignment & report

Due date: Sunday of Week 3, 23:59 (Melbourne time)

Weighting: 30%

Word limit: Maximum six pages (in single-column format including figures and references) with a font size between 10 and 12 points.

Overview

In this assignment, you will examine a data file and carry out the first steps of the data science process, including the cleaning and exploring of data. You will need to develop and implement appropriate steps, in IPython, to load a data file into memory, clean, process, and analyse it.

This assignment provides practical experience with the typical first steps of the data science process.

Assessment Learning Outcomes

This assessment will measure your ability to:

- prepare the provided data for analysis
- explore the provided data and build a scatter matrix for all numerical columns
- write a report to explain and justify how you dealt with different kinds of errors

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 Extract an interpretation and visualisation of data using exploratory data analysis in Python
- CLO4 Construct and document an experimental methodology for analysis of data
- CLO5 Select appropriate models, and apply simple machine learning tools and feature selection strategy for a defined data science problem
- CLO6 Apply professional standards to allow reproducibility of analysis

Assessment details

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 IPython.
- Parts of this assignment will include a written report; this must be in PDF format.
- 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.

Task 1: Data Preparation (10%)

Have a look at the file Automobile.csv, which is available in Canvas under the Assignments/Assignment 1 section of the course Canvas.

This Automobile Dataset consists of the specification of an auto in terms of various characteristics, its assigned insurance risk rating along with its normalised losses in use as compared to other cars. The original dataset was created/donated to UCI repository by Jeffrey C. Schlimmer

Below is a description of the attributes.

- Symboling: Insurance risk rating (+3 indicates high-risk auto; -3 indicates safe).
- Normalised-losses: Normalised losses in use as compared to other cars.
- Make: Make of the car.
- Fuel-type: Fuel type of the car.
- Aspiration: Aspiration of the car.
- Num-of-doors: Number of doors of the car.
- Body-style: Body of the car.
- Drive-wheels: Drive-type of the car.
- Engine-location: Location of the engine.
- Wheel-base: Measurement of wheel-base.
- Length: Length of the car.
- Width: Width of the car.
- Height: Height of the car.
- Curb-weight: The curb-weight of the car.
- Engine-type: The type of engine used in the car.
- Num-of-cylinders: Number of cylinders the engine has.
- Engine-size: The size of the engine.
- Fuel-system: The fuel system of the car.
- Bore: The bore of the cylinder.
- Stroke: Number of strokes.
- Compression-ratio: Compression ratio of the car.
- Horsepower: Engine power.
- Peak-rpm: Peak Revolutions Per Minute.
- City-mpg: Miles Per Gallon for city-drive.
- Highway-mpg: Miles Per Gallon for highway-drive.
- Price: price of the car.

Being a careful data scientist, you know that it is vital to carefully check, any available data before starting to analyse it. Your task is to prepare the provided data for analysis. You will begin by loading the CSV data from the file (using appropriate pandas functions) and checking whether the loaded data is equivalent to the data in the source CSV file. Then, you need to clean the

data by using the knowledge we taught in the lectures. You need to deal with all the potential issues/errors in the data appropriately (such as typos, extra whitespaces, sanity checks for impossible values, and missing values, etc.).

Task 2: Data Exploration (10%)

Explore the provided data based on the following steps:

1. Choose 1 column with nominal values, 1 column with ordinal Values, and 1 column with numerical values. (Please try to explore the columns/attributes of potential importance to the analysis, not just a random choice). Then, create a visualisation for each of them.
2. Explore the relationships between columns. You need to choose three pairs of columns to focus on, and you need to generate one visualisation for each pair. Each pair of columns that you choose should address a plausible hypothesis for the data concerned.
3. Build a scatter matrix for all numerical columns.

Note, each visualisation (graph) should be complete and informative in itself and should be clear for readers to read and obtain information.

Task 3: Report (10%)

Write your report and save it in a file called report.pdf, and it must be in PDF format and must be **at most 6 (in single-column format) pages (including figures and references) with a font size between 10 and 12 points.**

Penalties will apply if the report does not satisfy the requirement. The quality of the report is also being considered, e.g. clarity, grammar mistakes, the flow of the presentation.

Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs.

- Create a heading called “Data Preparation” in your report.
  - Provide a brief explanation of how you addressed the task.
  - For the steps dealing with the potential issues/errors, create a sub-section for each type of errors dealt with. (E.g. typos, extra whitespaces, sanity checks for impossible values, and missing values etc.)
  - Explain and justify how you dealt with each kind of errors.

- Create a heading called “Data Exploration” in your report.
  For each numbered step in Task 2 above, create a sub-section with the corresponding numbering.
  - **In subsection 1,** include all of your graphs from Task 2, Step 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 subsection 2,** include your plots from Task 2, Step 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 visualisation.
  - **In subsection 3,** present your scatter matrix and analyse what you observe from the graph.

Referencing guidelines

Use Harvard referencing style for this assessment. If you are using secondary sources, include these as a final slide in your PowerPoint deck, or as the final section of your report.

You must acknowledge all the courses 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.

Submission format

You need to submit the following files:

- Notebook file containing your python commands for Task 1 and Task, 'assignment1.ipynb'. Please use the provided solution template to organise your solutions: assignment1 TEMPLATE.ipynb
- # For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:
- Main menu → Kernel → Restart & Run All
- Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.
- Your report.pdf file: at most 6 (in single column format) pages (including figures and references) with a font size between 10 and 12 points. Penalties will apply if the report does not satisfy the requirement.
- They must be submitted as 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.

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.

RMIT UNIVERSITY

Assessment Criteria

Criteria		HD	DI	CR	PA	NN	NF
Task 1: Data Preparation (10%)	Points						
Data Retrieving Point 1: Load the CSV data from the file. You need to use an appropriate pandas function to load the csv data, and make use of the correct arguments including sep, decimal, header, names, if needed.	1.0						
Check data types Point 2: Check whether the loaded data is equivalent to the data in the source (CSV) file. That is, 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.	1.0						
Typos Point 3: Check whether there are typos in the data. If there are any typos, correct them by using masks.	1.0						
Extra-whitespaces Point 4: 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.	1.0	10.00 > 7.99	7.99 > 6.99	6.99 > 5.99	5.99 > 4.99	4.99 > 0.00	0.00 Pts
Upper/Lower-case Point 5: Cast all text data to upper-case by using an appropriate function.	1.0						
Sanity checks Point 6: 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.	2.0						
Missing values Point 7: 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.	3.0						
Task 2: Data Exploration (10%)		**HD**	**DI**	**CR**	**PA**	**NN**	**NF**
Choosing 3 columns	3.0	10.00 > 7.99	7.99 > 6.99	6.99 > 5.99	5.99 > 4.99	4.99 > 0.00	0.00 Pts

Choose 1 column with nominal values, 1 column with ordinal Values, and 1 column with numerical 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, or boxplots. - 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.							
Relationships between columns Explore the relationships between columns. You may choose which pairs of columns to focus on, but you need to generate 3 visualisations for this subtask. These should address a plausible hypothesis for the data concerned. For example, you might wonder: is there a relationship between the city-mpg and the price of a car? An appropriate visualisation for this could be to graph price against city-mpg scores. Please also format the graph as required in Task 2.1.	6.0						
Scatter matrix Build a scatter matrix for all numerical columns.	1.0						
Task 3: Report (10%)		**HD**	**DI**	**CR**	**PA**	**NN**	**NF**
Data Preparation Create a heading called "Data Preparation" in your report. For step 2 to 7 in Task 1 above, create a sub-section with corresponding numbering, and provide a brief explanation of how you addressed the task, and explain any choices that you made (if appropriate).	3.0	10.00 > 7.99	7.99 > 6.99	6.99 > 5.99	5.99 > 4.99	4.99 > 0.00	0.00 Pts
Data Exploration Create a heading called "Data Exploration" in your report. For each numbered step in Task 2 above, create a sub-section with corresponding numbering. - In subsection 1, include all of your graphs from Task 2, Step 1. Under each graph, include a brief explanation of why you chose this graph type(s) to represent the data in a particular column.	3.0						
Plots In subsection 2, include your plots from Task 2, Step 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 visualisation.	3.0						
Scatter matrix and analysis	1.0						

In subsection 3, present your scatter matrix and analyse what you observe from the graph.							