RMIT University
Computer Science & IT, School of Science

COSC 2670/2738 — Practical Data Science

Assignment 1: Data Cleaning and Summarising

Due: 23:59, Thursday 11 April, 2019 (week6)

This assignment is worth 15% 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 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 is intended to give you practical experience with the typical first steps of the data science process.

The "Practical Data Science" 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 with regards to any announcements or changes. Login through `https://learninghub.rmit.edu.au`.

Where to Develop Your Code

You are encouraged to develop and test your code in two environments: **Jupyter Notebook on Lab PCs** and **Teaching Servers**.

Jupyter Notebook on Lab PCs

On Lab Computer, you can find Jupyter Notebook via:

Start → All Programs → Anaconda2 (64-bit) → Jupyter Notebook

Then,

- Select New → Python 2
- The new created '*.ipynd' is created at the following location:
  - C:\Users\sXXXXXXX
  - where sXXXXXXX should be replaced with a string consisting of the letter "s" followed by your student number.

Teaching Servers

Three CSIT teaching servers are available for your use:

`(titan|saturn|jupiter).csit.rmit.edu.au`.

Details for how to access these servers are available in `‘‘Extra: Run Anaconda on RMIT Coreteaching Servers’’` under the `Modules/Week2: Data Curation` section of the course Canvas. You are encouraged to develop your code on these machines.

If you choose to develop your code elsewhere, it is your responsibility to ensure that your assignment submission can be successfully run using the version of IPython installed on Lab PCs or `(titan|saturn|jupiter).csit.rmit.edu.au`, as this is where your code will be run for marking purposes.

Important: You are required to make regular backups of all of your work. This is good practice, no matter where you are developing your assignment solutions.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised 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 penalised; there are no exceptions and no excuses. For further information, please see the *Academic Integrity* information at `http://www1.rmit.edu.au/academicintegrity`.

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 (5%)

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 normalized losses in use as compared to other cars. The original dataset was created/donated to UCI repository by Jeffrey C. Schlimmer [1]. The description of the attributes is given below.

- **symboling:** Insurance risk rating (+3 indicates high risk auto; -3 indicates safe).
- **normalized-losses:** Normalized 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.

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

- **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 start 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 (5%)

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 visualization for each of them.
2. Explore the relationships between columns. You need to choose **3** pairs of columns to focus on, and you need to generate **1** 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 visualization (graph) shoul be complete and informative in itself, and should be clear for readers to read and obtain information.

Task 3: Report (5%)

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. Moreover, the quality of the report will be 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 of dealing with the potential issues/errors, please create a sub-section for each type of errors you dealt with (e.g. typos, extra whitespaces, sanity checks for impossible values, and missing values etc), and also 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 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 analyze what you observe from the graph.

What to Submit, When, and How

The assignment is due at

23:59, Thursday 11 April, 2019 (week6).

Assignments submitted after this time will be subject to standard late submission penalties.
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:

1. Main menu → Kernel → Restart & Run All
2. 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.