RMIT

Computer Science & IT, School of Science

COSC 2670 — Practical Data Science

Assignment 1: Data Cleaning and Summarising

Due: Midday, Thursday 30 March, 2017 (week 5)

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" blackboard 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`.

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 ``Tute/Lab 01'' under the `Learning Resources/Tutorials` section of the course BlackBoard.

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 `(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.
- You *must* include a plain text file called "readme.txt" with your submission. This file should include your name and student ID, and instructions for how to execute your submitted script files.
- 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 `TeachingRatings.csv`, which is available in the directory /KDrive/SEH/SCSIT/Students/Courses/COSC2670/2017/a1 on `(titan|saturn|jupiter).csit.rmit.edu.au`. A copy of the same file is also provided under the `Assessment tasks/Assignment 1` section of the course BlackBoard.

This data set is based on original data that was gathered by US researchers[1] who were interested in examining the question of whether the attractiveness of a lecturer has an influence on student ratings of course quality. The data set includes the following attributes:

- **prof**: a randomly assigned unique identifier for each instructor included in the study

[1]Daniel S. Hamermesh and Amy M. Parker. *Beauty in the Classroom: Professors Pulchritude and Putative Pedagogical Productivity*, NBER Working Paper No. 9853, July 2003, JEL No. J7, I2. Data also reported in Murtaza Haider. *Getting Started with Data Science*, IBM Press 2016.

- **eval**: course overall teaching evaluation score (1: very unsatisfactory, to 5: excellent)
- **beauty**: rating of instructor's physical appearance by a panel of six students, averaged across the panelists, and transformed to have a mean of zero
- **minority**: does the instructor belong to a minority (non-Caucasian)
- **age**: the instructor's age
- **gender**: the instructor's gender
- **native**: is the instructor a native English speaker
- **tenure**: is the instructor on a tenure track (this is similar to the difference between being a permanent versus contract staff member, in Australia)
- **credits**: is the course a single-credit elective (including special interest courses such as yoga or aerobics), or more than a single-credit elective course
- **division**: is the course lower division (typically a first or second year course), or upper division (a more adanced course)
- **students**: number of students who participated in the evaluation
- **allstudents**: number of students enrolled in the course

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, by carrying out the following steps:

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.
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.
3. Check whether there are *typos* in the data. If there are any typos, correct them by using masks.
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.
5. Demonstrate how to cast text data to lower-case, using an appropriate function.

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.

7. Check whether the loaded data has any *missing values*. If so, use an appropriate function to replace them with the *column-wise* mean value.

Task 2: Data Exploration (5%)

Explore the provided data based on the following steps:

1. Create a visualization for each column (except `prof`) by producing an appropriate type of graph.

    - You should explore each column with at least one type of graph, but you can explore with more than one type, including 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.

2. Explore the relationships between columns. You may choose which pairs of columns to focus on, but you need to generate at least 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 age of an instructor and the course quality as perceived by students? An appropriate visualisation for this could be to graph `age` against `eval` scores.

3. Build a *scatter matrix* for all numerical columns.

Task 3: Report (5%)

Write your report and save it in a file called `report.pdf` (it must be in PDF format) and answer the questions below. 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.
- For each numbered step 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). As part of this exercise, you must specifically list any data rows that you changed.

- 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.

Optional Extension: Analysis of Missing Values (Up to 1.5% bonus marks for practical component)

ONLY attempt this section if you have completed all previous sections of the assignment.

In Task 1, Step 7, you were asked to deal with missing values by including the column-wise mean.

Now, your task is to deal with the missing values in the data sets with other options:

- replacing them with a fixed value
- replacing with the median value (column-wise)
- ignoring all observations containing missing values

For each of these approaches, choose a data column, and produce a new graph (corresponding to the initial graph that youu produced in Task 2, Step 1).

In your `report.pdf` file, create a heading called "Extension". In this section, include your three graphs. Under each one, briefly discuss the impact that the different approachs to dealing with missing values have on what you observe from the visualisation.

What to Submit, When, and How

The assignment is due at

Midday, Thursday 30 March, 2017 (week 5).

Assignments submitted after this time will be subject to standard late submission penalties. There are three files you need to submit:

- Script file containing your python commands for Task 1.
- Script file containing your python commands for Task 2.
- Your `report.pdf` file.

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 Blackboard:

Assessment tasks/Assignment 1/Assignment 1 Submission.

Please do NOT submit other unnecessary files.