



Practical Data Science with Python
COSC 2670/2738
Assignment 1

| | | |
|---|-----------------|--------------------------------|
|  | Assessment Type | Individual |
|  | Due Date | 23:59, the 15th of April, 2020 |
|  | Marks | 15 |

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 → Anaconda3 (64-bit) → Jupyter Notebook

Then,

- Select New → Python 3
- The new created ‘*.ipynd’ is created at the following location:
 - C:\Users\sXXXXXXXX
 - where sXXXXXXXX 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.

Academic integrity and plagiarism (standard warning)

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 following: <https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity>.

All submission will be checked by TurnedIn.

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.

Assessment details

Task 1: Data Preparation (5%)

Have a look at the file **StarWars.csv**, which is available in Canvas under the **Assignments** -> **Assignment 1** section of the course Canvas.

This file contains data behind the story America's Favorite 'Star Wars' Movies (And Least Favorite Characters)¹. The author collected the data by running a poll through SurveyMonkey Audience, surveying 1,186 respondents. The description of the questions asked in the survey is given below.

- Have you seen any of the 6 films in the Star Wars franchise?
- Do you consider yourself to be a fan of the Star Wars film franchise?
- Which of the following Star Wars films have you seen? Please select all that apply. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)
- Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)
- Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her. (Han Solo, Luke Skywalker, Princess Leia Organa, Anakin Skywalker, Obi Wan Kenobi, Emperor Palpatine, Darth Vader, Lando Calrissian, Boba Fett, C-3P0, R2-D2, Jar Jar Binks, Padme Amidala, Yoda)
- Which character shot first?
- Are you familiar with the Expanded Universe?
- Do you consider yourself to be a fan of the Expanded Universe?
- Do you consider yourself to be a fan of the Star Trek franchise?
- Gender
- Age

¹<https://github.com/fivethirtyeight/data/tree/master/star-wars-survey>

- Household Income
- Education
- Location (Census Region)

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.

Task 2: Data Exploration (5%)

Explore the provided data based on the following steps:

1. Explore the survey question: *Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi)*, then analysis how people rate Star Wars Movies.
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. Explore whether there are relationship between people's demographics (Gender, Age, Household Income, Education, Location) and their attitude to Start War characters.

Note, each visualization (graph) should 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.

What to Submit, When, and How

The assignment is due at

23:59, the 15th of April, 2020.

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 2, ‘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.

A Marking Guidelines

| Data Preparation (Maximum = 5 marks) | Data Exploration (Maximum = 5 marks) | Report (Maximum = 5 marks) |
|--|--|--|
| <p>5 marks</p> <p>Data preparation is well designed, systematic and well explained. All potential errors/issues have been completely examined and properly treated</p> | <p>5 marks</p> <p>Analysis is thorough and demonstrates understanding and critical analysis. Well-reasoned exploration are provided for all sub-tasks. All analysis, comparisons and conclusions are evidenced by data (e.g. in well-formatted figures and/or tables).</p> | <p>5 marks</p> <p>Very clear, well structured and accessible report, an undergraduate student can pick up the report and understand it with no difficulty.</p> |
| <p>4 marks</p> <p>Data preparation is reasonably designed, systematic and explained. There are at least one obvious missing issue/error. Each examined error/issue have been completely checked and properly treated.</p> | <p>4 marks</p> <p>Analysis is thorough and demonstrates good understanding and critical analysis. Adequate exploration are made for all sub-tasks. Most analysis are supported by evidence by data (e.g. in well-formatted figures and/or tables).</p> | <p>4 marks</p> <p>Clear and structured for the most part, with a few unclear minor sections.</p> |
| <p>2-3 mark</p> <p>Data preparation is somewhat adequately designed, systematic and explained. There are several obvious missing issues/errors. Each identified issue/error might not be checked completely and/or treated properly.</p> | <p>2-3 marks</p> <p>Analysis is adequate and demonstrates some understanding and critical analysis. Some exploration are given for some sub-tasks. A portion of analysis and comparisons are supported by some evidence.</p> | <p>2-3 mark</p> <p>Generally clear and well structured, but there are notable gaps and/or unclear sections.</p> |
| <p>0-1 marks</p> <p>Data preparation is poorly designed, systematic and explained. There are many obvious missing issues/errors. Each type of identified issues/errors have not be checked completely and treated properly.</p> | <p>0-1 mark</p> <p>Analysis is poor and demonstrates minimal understanding and critical analysis. Few exploration are made and illustrated partially for tasks. Little analysis are supported by some evidence.</p> | <p>0-1 marks</p> <p>The report is unclear on the whole and the reader has to work hard to understand.</p> |