COSC 2670/2732 Practical Data Science with Python

Project Assignment 1, Semester 2, 2022

Marks : This assignment is worth 25% of the overall assessment for this course.

Due Date : Fri, 19 August 2022, 11:59PM (End Week 5), via `canvas`. Late penalties apply. A penalty of 10% of the total project score will be deducted per day. No submissions will be accepted 5 days beyond the due date.

Objective

The key objectives of this assignment are to learn how to process data using python and pandas. The assignment is broken into 5 separate problems, which you should solve in order as you may need functions you create from previous exercises. The problems are given to you as jupyter notebooks, and include many helper functions and clear instructions on what you need to do to solve each problem.

The data you will be working with is a dump from stackexchange, specifically the Security stack exchange. We will only need to use three files for input – Posts.feather, Comments.feather, and Users.feather. The feather format was designed as part of Apache Arrow, and is a clean way to serialize a dataframe in binary format. So, it should be easy for you to read in a file to a dataframe, and immediately start exploring it.

Provided files

The following template files are provided:

- **SXXXXX-A1-[Q0,Q5].ipynb** : The skeleton Jupyter notebook files you should use to solve the problems.
- **Posts.feather** : The postings dataframe.
- **Comments.feather** : The comments dataframe.
- **Users.feather** : The users dataframe.
- **A1.pdf** : This file.

See the provided notebook files for further instructions.

Creating Your Workspace

You will notice that the six notebooks we provide have the prefix "SXXXXX-". Your very first task is to create a folder called SXXXXX, but replace the Xs with your real student number. So for example if my student ID was S12345, I would name the folder S12345, and make a copy of each notebook into this folder with the name S12345-A1-Q1.ipynb, and so on. If I did this correctly, I should see something like:

```
S12345/S12345-A1-Q0.ipynb
S12345/S12345-A1-Q1.ipynb
S12345/S12345-A1-Q2.ipynb
S12345/S12345-A1-Q3.ipynb
S12345/S12345-A1-Q4.ipynb
S12345/S12345-A1-Q5.ipynb
S12345/Posts.feather
S12345/Comments.feather
S12345/Users.feather
```

Note that I have moved the three feather files into this working directory as well as I will need them in the same directory from where I run Jupyter. These 3 feather files will not be submitted as they are not tiny (they are real data!). More on what you need to submit later. Just remember, whenever I say SXXXXX henceforth, I mean your real student ID, and keep it case sensitive! That is, use a capital 'S' and not a lowercase 's' please. The case sensitive naming conventions may seem pedantic, but without them, we cannot possibly test the assignments from nearly 400 students in any reasonable amount of time!

Creating Anaconda Environment

The first task is to create the correct Anaconda working environment. The rules for this may differ a little depending on every OS, but you should be able to find the right invocation for your platform of choice.

Assuming you have properly setup jupyter to use Python 3.8.x, you should be able to just run the provided notebook SXXXXX-A1-Q0.ipynb once to install the necessary dependencies, and test your version of python. This will create a file called requirements.txt which should contain all of the library versioning information for your installation. You should submit a copy of the requirements.txt file created, and after you run the Q0 notebook, please use the option to save to notebook run as a PDF file from the pulldown menu in Jupyter. You will submit this file as well in case we need to check the output.

1 Marking Template

For clarity, I will briefly outline the marking policy for the assignment. There are really just six specific components you should consider. First, the five questions:

- Question 1 - 5 points
- Question 2 - 5 points
- Question 3 - 5 points
- Question 4 - 3 points
- Question 5 - 2 points

You will notice that the last two questions are worth less, as really these are designed as challenge questions. If you run out of time or can't get them working, they won't hurt much in the end. They are meant to push you a little bit and not to make your life miserable! The remaining 3 points are based on your ability to follow the guidelines. Submit the files as described below, name files case sensitive exactly as specified, and name everything correctly. If you make a mistake in naming files, it will fail the test harness, and it will cost you marks. These should be easy marks to get if you can follow the directions!

I will not repeat the detailed instructions contained in the Jupyter notebooks for each of the five challenge questions – you can look at the notebooks to find out more.

Submission

All submissions must be made through MyRMIT. A link will be provided within canvas, under the Assignments tab for submissions. Assignments submitted through any other method will not be marked. To submit your files, create a top level directory which is your student number, as described above. This is just a folder named SXXXXX, where SXXXXX is your student id. Inside that folder, there should be the six Jupyter notebook files, named as described above. You **should not submit the three feather data files**.

An example output is shown below of the files you should be submitting. If your student ID happened to be S101010, you would submit a file called `S101010.zip`, which contains the following, including the folder!

```
$ unzip S101010.zip
Archive:  S101010.zip
   creating: S101010/
  inflating: S101010/S10101-A1-Q0.ipynb
  inflating: S101010/S10101-A1-Q0.pdf
  inflating: S101010/S10101-A1-Q1.ipynb
  inflating: S101010/S10101-A1-Q1.py
  inflating: S101010/S10101-A1-Q2.ipynb
  inflating: S101010/S10101-A1-Q2.py
  inflating: S101010/S10101-A1-Q3.ipynb
  inflating: S101010/S10101-A1-Q3.py
  inflating: S101010/S10101-A1-Q4.ipynb
  inflating: S101010/S10101-A1-Q4.py
  inflating: S101010/S10101-A1-Q5.ipynb
  inflating: S101010/S10101-A1-Q5.py
  inflating: S101010/requirements.txt
```

NOTE 1 : You will notice that you must submit a python script which is generated automatically from your jupyter notebook. You **need** to run this file to ensure it runs correctly. We will walk you through how to do this for the first Question in the practicals to ensure that you know how to do it. The magic command is "jupyter nbconvert NOTEBOOK –to python" where NOTEBOOK is the name if the notebook you wish to convert. The PDF file is discussed above and will also be covered in your Practical sessions.

NOTE 2 : There is a discussion group available in the course canvas. Please do not post code snippets in this forum. Do ask questions if you have them! We will do our best to answer them as quickly as we can. This is the best medium to ask a question, as there is a very strong chance that if you are confused about something, your classmates are too.

Plagiarism Warning

University Policy on Academic Honesty and Plagiarism: It is University policy that cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student concerned. Please see the RMIT policy for more details: https://www.rmit.edu.au/students/my-course/assessment-results/academic-integrity. **THIS IS NOT A GROUP PROJECT.** Students are reminded that this assignment is to be attempted individually. Plagiarism of any form will result in zero marks being given for this assessment, and can result in disciplinary action. We routinely use plagiarism software on projects! Please, please don't do it. Be aware that paying someone on a coding site to do it for you is a form of plagiarism. If you are submitting work that is not your own, regardless of how you got it – you are in breach of this policy.

Extension Policy

Individual extensions will are unlikely to be considered or granted by the PDS team. If you have an Equitable Learning Plan in place, please discuss issues you may have with me at least two days prior to the deadline. An ELP does not mean you are eligible for an extension, it just means that if you have a good reason to need one, we will work it out. We are not monsters, but doing things on time matters, and not just for this course but for all of your courses. Getting behind creates more stress. We consider timeliness as a core component of the assessment. If we decide to extend the deadline, the fairest way we can do this is to extend it for *everyone*, and that may not be possible given the the size of the course and university policies. We have set deadlines as late as we can in order to meet the university timelines we cannot control. So try to be early and not late. If you procrastinate and suddenly have other priories, you will be in trouble, and that is not a valid reason for an extension.

If you have suffered a personal tragedy or illness, there is a University process in place to grant extensions. Our preference is that you go this route if you must, as they have very clear criteria to grant exemptions. For more information about applying for Special Consideration, see the rules and regulations at https://www.rmit.edu.au/students/my-course/assessment-results/special-consideration-extensions/special-consideration.

Getting Help

Come talk to us during office hours. Email us. Use the discussion board. Ask a question in a lectorial or a practical. There is help available if you need it. We do strongly urge that you start with the discussion board because, as mentioned above, if you are confused, there is a very good chance someone else has exactly the same question, and it is not a great use of the team's time to keep answering the same question over and over.