PYTHON FOR DATA – SUMMARY MATERIALS

INTRODUCTION TO LIBRARIES

Libraries, modules, and packages are all umbrella terms (with minor distinctions) for any code which you or others have written which you may be able to re-use. The ecosystem is:

- The "standard libraries" (stdlib) that come packaged with Python
- The wider ecosystem – all user-created packages in Python – if you can think of it, someone will have made a library for it. The full list of easily accessible packages is at the Python Package Index (https://pypi.org/).
- Your own libraries – for example the 903 validator.

To use these external packages you

- Install them (if they are not stdlib packages).
    - Repl.It does this automatically, but local Python may not.
- Import them using the **import** statement

There are a few ways to import information from a module – either

- **import pandas** – imports the module and calls it pandas (e.g. **pandas.DataFrame** in code)
- **import pandas as pd** – renames the module (e.g. **pd.DataFrame** in code)
- **from pandas import DataFrame** – specifically imports the **DataFrame** only, with no namespace (e.g. **DataFrame** in code)

For context on how modules and packages are created and used, and how to create your own, a continued excellent resource is the tutorial here https://docs.python.org/3/tutorial/modules.html.

INTRODUCTION TO PANDAS

The Pandas User Guide cover all concepts introduced in this video in much more detail – and will be your friend in doing any future development in this course.

For the topics covered in this video, read

- Intro to DataFrame and Series https://pandas.pydata.org/docs/user_guide/dsintro.html
- Indexing and selecting data https://pandas.pydata.org/docs/user_guide/indexing.html

But the whole user guide is excellent, and you will need some other parts not covered in the video for the exercises, and for doing any further Python analysis work: https://pandas.pydata.org/docs/user_guide/

PANDAS DATA ANALYSIS

For this section we talk about reading in files, working with dates and working with strings. Again, the user guide continues to be a great resource for details here, so I will signpost appropriate sections.

- Reading csv files (and anything else) https://pandas.pydata.org/docs/user_guide/io.html
- Working with dates https://pandas.pydata.org/docs/user_guide/basics.html#dt-accessor
- Working with strings https://pandas.pydata.org/docs/user_guide/text.html

Otherwise, continue reading the User Guide for more information. Some of the advanced concepts such as Merging and Grouping we will cover again later in the course – but there is no harm in delving into those now.

EXERCISES

For the exercises in this section you will be using synthetic 903 data and answering some analysis questions. This should get you used to filtering the data and some basic functionality of Pandas.

On Repl.It you should see the project has been assigned to you. As usual – work on these exercises, and let me know if there are any questions or you want me to take a look at the code.

Your code should initially read in the CSV files I have provided in the Repl (**header.csv** and **episodes.csv**) and then write code which prints out answers to each of the following questions:

- How many children are mothers?
- How many episodes are present in the episodes data for child ID 465367?
- How many children in the dataset were born in July and are Male?
- What are the index labels (the entries in the DataFrame index) for all episodes which had a DECOM (date commenced) of 4th March, 2017?
    - Hint: your answer should be a list of integers. The length of the list is less than 10.
- Stretch: what is the average episode length in days?
    - Hint: You will need to use DECOM and DEC from the Episodes data – the start and end date of the episodes.
    - Hint: You will need to work with **timedelta** objects (these are differences between two **datetime** objects) – Pandas has details on this in the user guide.
    - Hint: Pandas has a built in "mean" function for finding the average of the columns.
    - Hint: Use google and stackoverflow to find out more.