COMP 5070
Statistical Programming for Data Science

Assessment 1 SP2 2023

- You must submit your assessment through LearnOnline. Every submission should include two files – Jupyter Notebook file with all cells executed (that is, with all outputs, graphs, discussions) and PDF printout for the same file.
- Please think about your reader. Take care about readability of your code and your discussions. Make it easy to see that your job has been done well.
- No compressed files (e.g. .zip, .rar, .tar, .gz, .7z) are allowed for submission. These files will be ignored during marking.
- You do NOT need to include the data files provided to you, as it can be safely assumed I have them too.
- The assessment is out of 100 marks. To obtain the maximum available marks you should aim to:
    1. Code all requested components (60%)
    2. Aim for optimised code in terms of computational overhead (5%). It is not always possible to avoid loops, however you should aim to avoid loops where possible.
    3. Use a clear coding style (5%). Code clarity is an important part of your submission. Thus, you should choose meaningful variable names and adopt the use of comments - you don't need to comment every single line, as this will affect readability - however you should aim to comment at least each section of code.
    4. Have the code run successfully (5%).
    5. Output the information in a presentable manner as decided by yourself and present the requested statistical analyses/discussions (25%).

• *Plagiarism is a specific form of academic misconduct.* Although the University encourages discussing work with others and the Social Forum will support this, ultimately this assessment is to represent your individual work. If plagiarism is found, all parties will be penalised.

You should retain copies of all assignment computer files used during development of the solution to the assessment. These files must remain unchanged after submission, for the purpose of checking if required.

Late submission will be penalized by 10-point deduction for each day or part of it after the due date.

COVID-19 in Australia

In this assessment you will analyse data about a development of COVID-19 in Australia. The data was collected by **COVID Live** (https://covidlive.com.au/).

Data

The data cover 5 main states in Australia – NSW, Vic, QLD, SA, WA – and represent information about the numbers of new cases (files daily_cases) and new deaths (files daily_death).
Important note: despite files names there is a mix of data – daily numbers before September 9, 2022, and weekly data after that. This is due to the Australian government decision regarding reporting rules. You will be doing all analysis on weekly bases, so you will need to aggregate daily data into weekly.

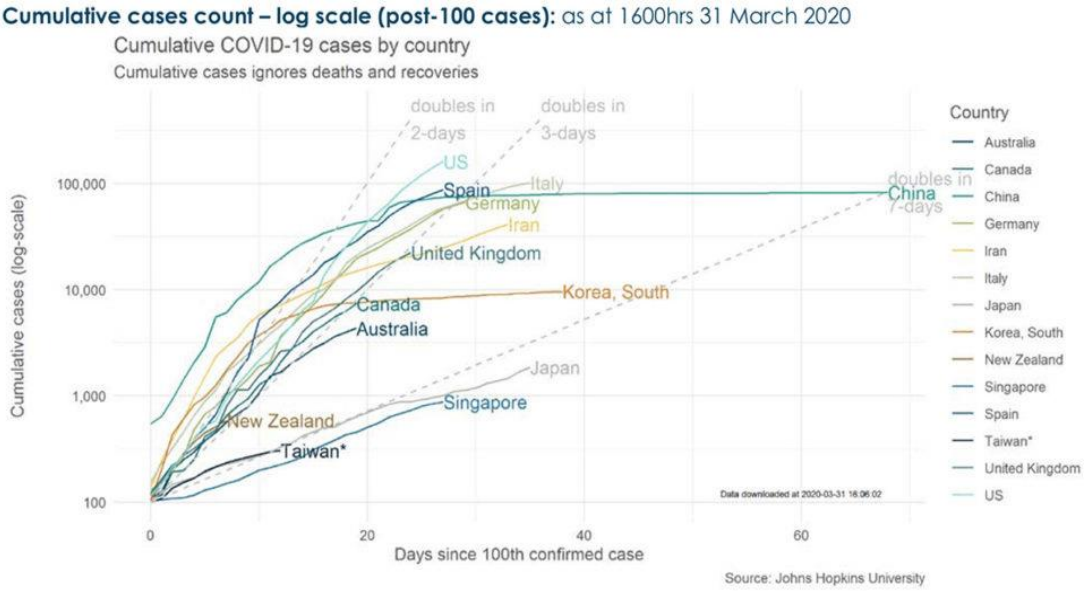
Data set for new cases has two columns which look similar NEW and NET. You are going to use column/variable NEW – the number officially reported by authorities. As a result, you will need to calculate your own number for total cases. Column NET contains adjusted numbers.

When you need to load data files, you can work with zip file or with individual unpacked TSV-files. Both versions of the data will be in the working directory during marking process.

Analysis

You must execute the following steps and present answers for the research questions below.

1. Write an introduction about what type of analysis you plan to execute. Provide a brief data description: what are your data about; how many variables and observations there?
2. Report and discuss distributions of new cases and deaths weekly numbers in five states.
3. Create a graph similar to the example below to plot the history of COVID-19 in different states. Pay attention that the graph for each state starts on different calendar day as it starts on a day after 100 reported cases. Your graph should start on the week after 1000 cases were reported and then show cumulative weekly numbers as we do a weekly-based analysis. Provide brief comments on the progress of COVID-19.



4. Normalise numbers of new cases by population (see table in the appendix) and plot a calendar-based historical graph of weekly cases. Provide brief comments on similarities and differences.
5. Study a relationship between number of new cases and deaths in five states.
6. Write a conclusion outlining your analysis and results.

You do not need to write too much as everyone hates reading long and/or boring reports. At the same time, if you need more space for your discussion then you can take it. There are no limits for word count or number of pages.

As a general rule, every question requires to present result in three ways – as a graph, as numbers and as text describing both. There might be exceptions when you cannot have all three, but these situations are not so common.

Submission

You must do your work using Jupyter Notebook. There will be programming cells to import packages, load and clean data, and run analysis; then there will be outputs for the results of the analysis and your discussions about the results as Markdown cells.

You must submit Jupyter Notebook file and a copy of this file as PDF printout. Any other files will be ignored. These two files should be submitted individually – no zip of other archives are allowed.

Appendix

Australian population from Australian Bureau of Statistics (ABS) – https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/jun-2022

State	Population at 30 Jun 2022 ('000)	Change over previous year ('000)	Change over previous year (%)
New South Wales	8153.60	59.8	0.7
Victoria	6613.70	65.7	1.0
Queensland	5322.10	104.4	2.0
South Australia	1820.50	17.3	1.0
Western Australia	2785.30	35.4	1.3
Tasmania	571.50	3.6	0.6
Northern Territory	250.60	1.4	0.6
Australian Capital Territory	456.70	3.1	0.7