# 

# COMP 5070 Statistical Programming for Data Science

## Take-home Exam DUE by 10 am (CST), Tuesday, June 28

* This take-home exam is worth 25% of your overall grade.
* The exam is to be submitted as uncompressed files using LearnOnline (no zip, tar, rar, etc). Your submission should include (see more details at the end of the instructions):

1. well-presented report in Word or PDF format
2. R-scripts used to produce presented results/graphs
3. data set you scrape from the internet.

* The exam is out of 100 marks. There are three main parts in the exam:

1. downloading data using *rvest* package (25%)
2. run data analysis and prepare a written report (50%)
3. create an interactive dashboard using package *shiny* (25%).

* To obtain the maximum available marks you should aim to:

1. Code all requested components
2. Aim for optimised code in terms of computational overhead
3. Use a clear coding style. 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. I should be able to run your R-code without any modifications, that is, I open your code in RStudio and press Ctrl+Shift+S. If there are any errors – you lose points.
5. Output the information in a presentable manner and present the requested statistical analyses/discussions.

* This is an exam, so there is no provision for a late submission. Please don’t send me your submission by email – I am not allowed to accept them. All submission should go through LearnOnline.
* 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 submission 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. These files must remain unchanged after submission, for the purpose of checking if required.

# COVID-19 in Australia

## Downloading data

You have to get your own data for this exam. You must use *rvest* package to scrape data from <https://covidlive.com.au/>. There are multiple pages for different variables and states. For example,

* historical data about **new cases** for Victoria is <https://covidlive.com.au/report/daily-cases/vic> and for NSW is <https://covidlive.com.au/report/daily-cases/nsw>
* historical data about **vaccinations** in SA is <https://covidlive.com.au/report/daily-vaccinations/sa>

You don’t want to put too much pressure on the “COVID Live” website. Please, review all questions, make a plan of what data you need, then create one R-script to download and clean all required data and save it on the hard-drive using functions save() or saveRDS(). Later you will run analysis and create a dashboard that will load your data file by functions load() or readRDS().

Download data just once, store it and then use for analysis stored data. Don’t download the same data multiple times.

## Data analysis

**Question 0**. Introduction in the business case and analysis to be presented later. Relevant descriptive statistics and discussions.

It might be beneficial if you write an introduction *after* you completed all other tasks. This way you would have a good understanding of what to put in this section as you already know what you’ve done and what data you’ve used.

**Question 00**. Conclusion at the end of the report briefly summarising all your findings.

## “New” pandemic of COVID-19

**Question 1.** You are going to analyse the latest development in the total number of COVID-19 cases in Victoria, NSW and WA (**New and Cumulative Confirmed Cases**). There are two similar columns – NEW and NET. You must take column NEW and then calculate your own variable for total cases. It will be different to the reported total cases due to some errors and correction – that is OK. You work with officially reported numbers that are presented in column NEW.

For a very long time WA remained completely locked from other states. Make a historical graph of total cases in NSW, Vic and WA. Provide your comments about similarities and differences between overall trajectories for COVID-19 cases in these three states. For a better comparison you might consider states population provided in Appendix.

## Vaccination rates

**Question 2.** In this question you should analyse a relationship between vaccinations and new cases for the same three states as in question 1 (**Daily Doses Administered** and **New and Cumulative Confirmed Cases**). One hypothesis (which might be incorrect) is that higher number of reported cases might encourage more vaccinations. We consider here “Total doses administered” without any separation for the first or second or booster shots.

There is a very strong pattern in the number of vaccinations related to the weekend, so to eliminate its effect you should do a weekly-based analysis.

First, analyse distribution of weekly cases and weekly vaccinations. Then, study the relationship between new cases and vaccinations. Provide a brief discussion on the results.

## Hospitalisations

**Question 3.** Investigate an effect of COVID-19 on the health system for the same three states as above. Study the relationship between active cases and cases admitted to hospital (**Cases Admitted to Hospital** and **Daily Active Cases**). Then, get a historical graph of a proportion of daily active cases admitted to hospital with added 7-day moving average line. Discuss your findings. Can you identify an effect of the vaccination campaign (you can check yourself in **Daily Second Doses Administered** but no need to download or analyse these data)?

## Interactive dashboard

You will create a dashboard using package *shiny* as an alternative delivery to the written report. The dashboard should cover the same topics as a report. Take care about a good design for the dashboard. One graph per screen is probably a bad idea (unless your graph is very detailed, and you need a lot of space for it). All graphs on one page could be a bad idea too (unless your graphs are very simple).

All graphs on the dashboard should be interactive. Your dashboard is not a simple copy of the written report but an enhanced version of it.

## Submission structure

Here is a suggested structure for the exam submission. While it is not compulsory to follow this structure, it should help you to better organise your exam, avoid missing any relevant files and speed-up marking process. These are five files suggested for your submission:

1. R-script file *download.R* with all code used to scrape data from the internet, clean it and store on the hard drive ready for the analysis.
2. R data file – *examdata.RData* or *examdata.rds* – were use store all information prepared by *download.R* script. If you have multiple variables, then you can use function *save()* or you can put all your variables in one list and use function *saveRDS()*.
3. R-script *analysis.R* with all code used to load and analyse the data file and to output all statistics and data visualisations for the written report.
4. MSWord or PDF file with a written report. It should be well-organised and well-prepared report with introduction, conclusion, and all required discussions. Good report should have a title, headings and sub-heading, page numbers, titles for figures and tables. Report should NOT include R-code or R-outputs but results of your analysis. I can see your code in the submitted R-script.
5. R-script *app.R* with the code of your interactive dashboard. The dashboard will include mostly the same visualisations as your report, as the dashboard is an enhanced copy of the report but without discussion. However, data visualisations should be interactive, e.g., add/remove a line of a one-week moving average; make an average line adjustable to cover 5, 7, 11, … days; add/remove states; adjust time intervals presented on the graph; etc. Table is a type of data visualisation too, so it can be included in the dashboard. Like the report, the dashboard should have appropriate titles and headings – it should be easy for the reader to understand what is what on the dashboard.

## Getting help

You are allowed to ask questions during exam. I will be monitoring the forum and try to respond as quick as possible. At the same time, I will not answer direct questions like “how to do question 1?” or “I’ve done this – is it correct?”. When you ask questions – remember about academic integrity and don’t reveal a solution or a part of it.

## Appendix - Population

Analysing raw counts might be not appropriate as there are small and large states. Use a percentage of population and/or numbers per 100,000 of population. Population numbers for Australia and all states are provided by ABS (<https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release#states-and-territories>)

|  |  |
| --- | --- |
|  | **Population at 30 Jun 2022 ('000)** |
| New South Wales | 8,153.6 |
| Victoria | 6,613.7 |
| Queensland | 5,322.1 |
| South Australia | 1,820.5 |
| Western Australia | 2,785.3 |
| Tasmania | 571.5 |
| Northern Territory | 250.6 |
| Australian Capital Territory | 456.7 |
| Australia (a) | 25,978.9 |

No need to use *rvest* to download population numbers, you can use the table above.

## Appendix – If there are problems with internet access

Here is something you don’t need but it is good to have a backup plan.

If you get error messages while trying to scrape data from <https://covidlive.com.au/> ; if you don’t have access to this website (e.g., overseas students or poor internet connection) – you can use an archive *covide\_html.zip*. It contains html pages you need for the exam.

You use *rvest* and everything is the same – you just use a different address to load data

# normal way – with internet access

read\_html(″https://covidlive.com.au/report/daily-cases/vic″)

# offline - no internet access, html files are in working directory

read\_html(″https\_\_\_covidlive.com.au\_report\_daily-cases\_vic.html″)

As you can see, it is the same address but special symbols / and : are replaced by underscore, as they cannot be a part of the file name), and an extension “.html” is added.

Html pages look different in the internet browser comparing to the original web design as they don’t have access to server-based CSS and images, but they are identical to pages on the web server. In fact, the same web scraping code will work for both – internet-based and offline files.

There are more files than you need for the exam. For example, I included all states. It is not expected that you do analysis for states beyond NSW, Victoria and WA. However, you can if you want. When the code is ready, it can run for more than three times.

Most probably, you don’t need this archive and this section. But if you have problems with internet – they should help.