Bachelor of Informatics Honours

Information & Knowledge in Organisations: Data Analytics

Assignment 1: Analysing the spread of Covid-19 in South Africa

Covid-19 had a profound impact on the world over the past couple of years. In this assignment, you will use R to clean and visualize data related to the spread of the virus. You will obtain data from a publicly available source and use R to transform it into a format that can be used for analysis. After cleaning the data, you will create visualizations that provide insights into the spread of the virus over time, in different countries and regions.

This assignment focuses on data wrangling and visualisation. For this assignment you need to produce a Quarto notebook that includes the code to wrangle and process the data necessary for visualising the spread of Covid-19 in South Africa.

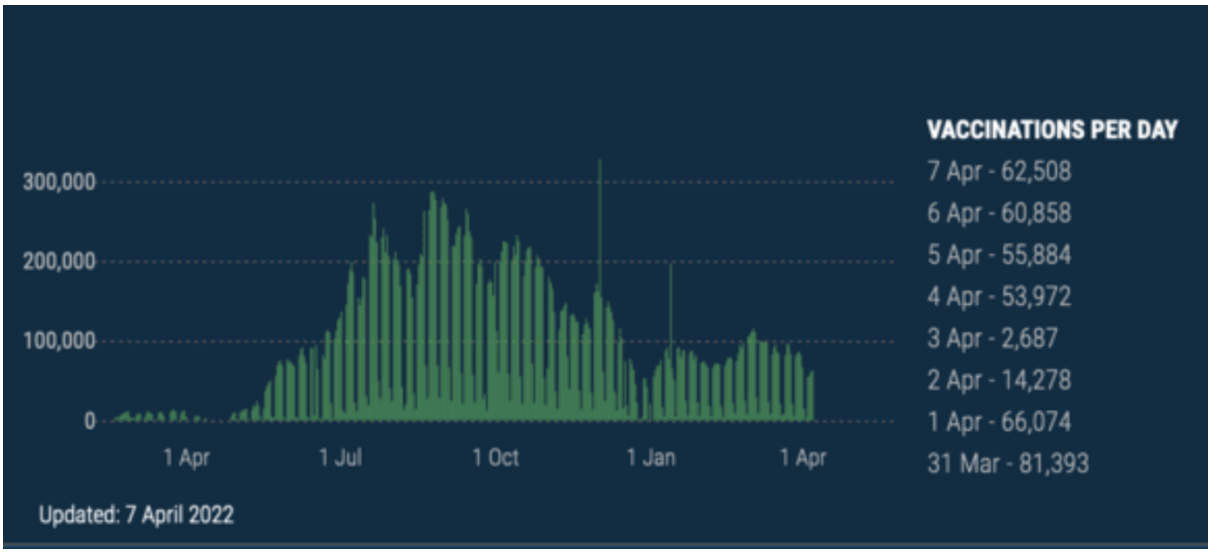
The assignment has two parts:

- **Part A: Figure reproduction**
  - You need to produce the **seven** visualisations presented on the next pages.[1] This will involve collecting and processing the data prior to producing the visualisations.
- **Part B: Figure design**
  - Using this same data (or any other data) you need to produce **seven** further visuals that enable us to understand aspects of the Covid-19 pandemic beyond South Africa. These visuals cannot be the same as those produced in Part A.

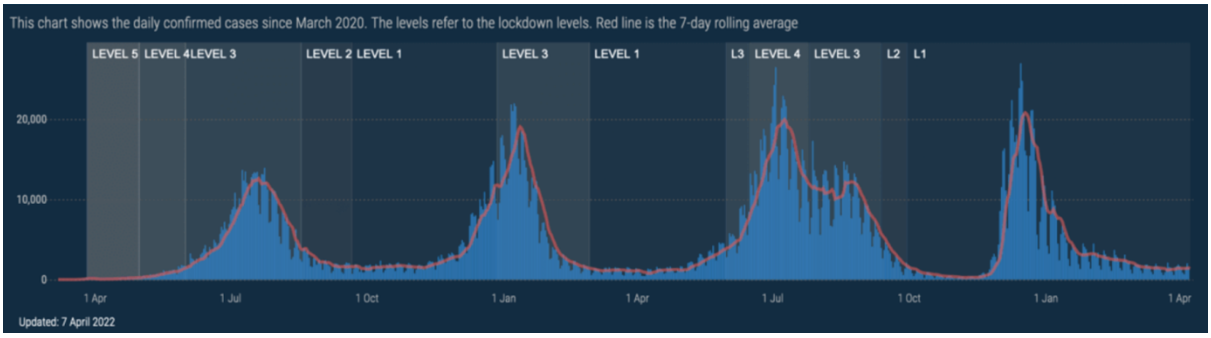
[1] These visualisations are based on the now paused dashboard available at: https://mediahack.co.za/datastories/coronavirus/dashboard/

Figures for part A

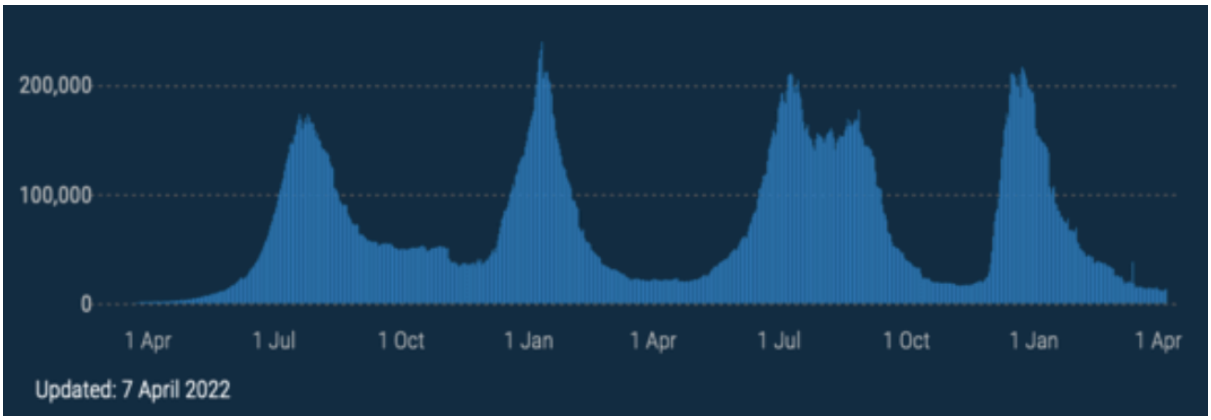
Daily vaccinations



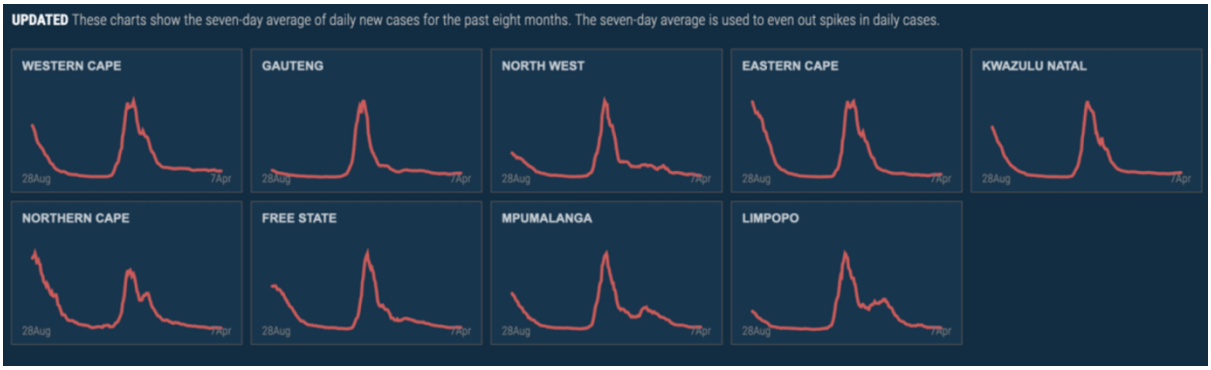
Daily confirmed cases



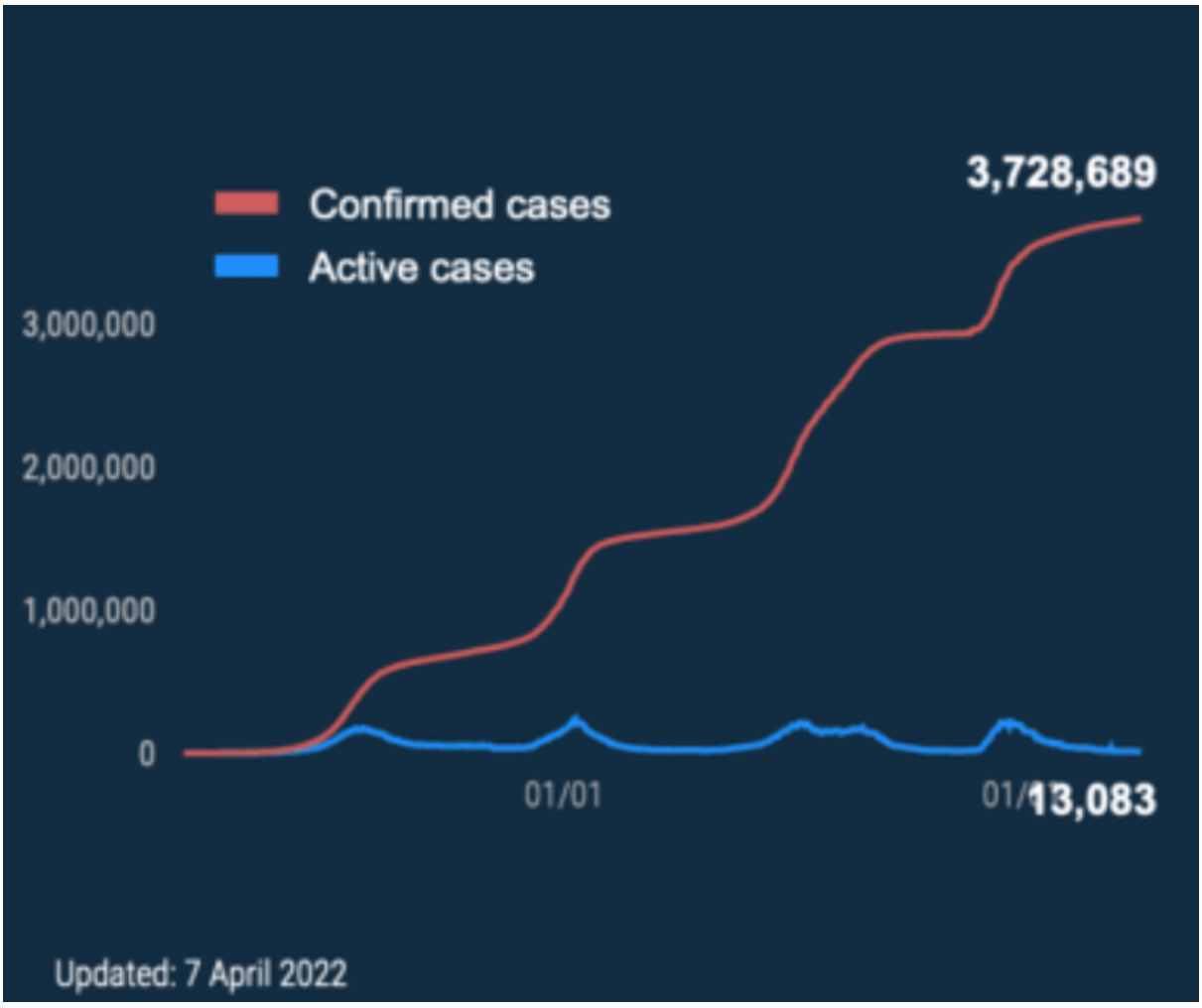
Active case total by day



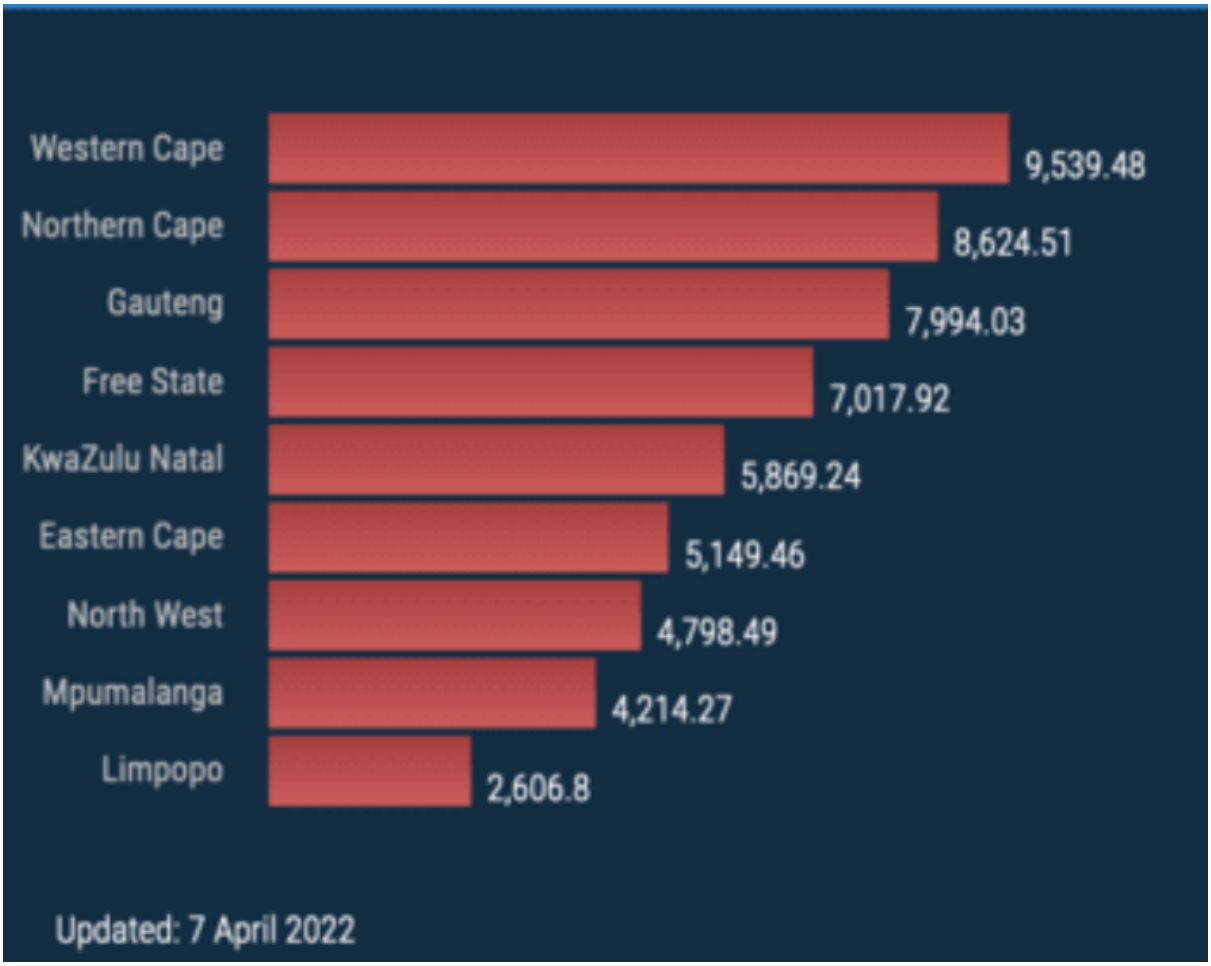
Daily case trends



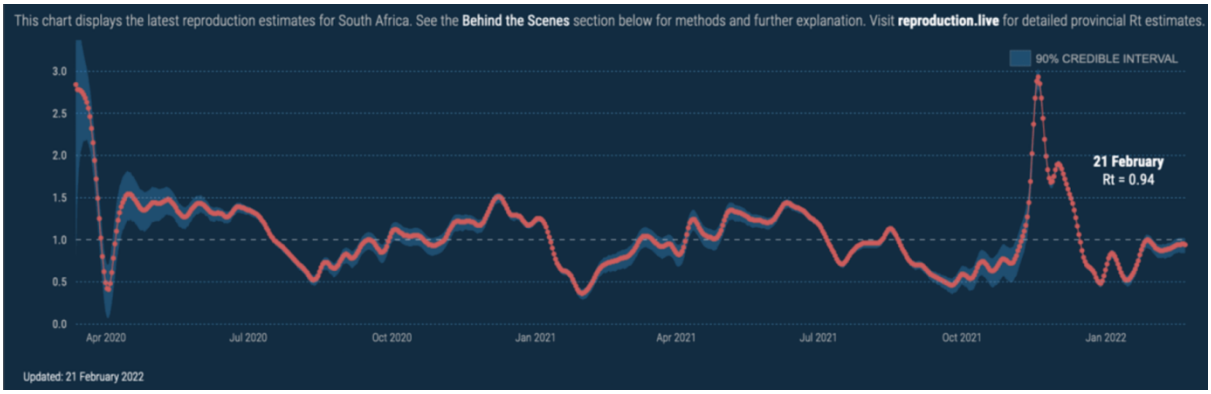
Cases vs. active cases



Provincial infection per 100,000 population



Rt estimates for South Africa



Your notebook needs to include the code used to wrangle and process the data, as well as the outputs and visualisations produced for parts A and B. Additionally, there needs to be a logical and coherent flow to your document, with headings and descriptions. For this assignment, you do not need to programmatically collect the data (although you *can*). You should be able to find various openly accessible datasets to use for your analyses.

For data sources I recommend starting with the following, but you are free to look for and use any other sources if these are insufficient:

- https://mediahack.co.za/datastories/coronavirus/data/

- https://reproduction.live/about

This task primarily focuses on data cleaning, wrangling, transformation, and visualisation with R. You need to reproduce all seven figures and produce a further five of your own. While you need to create the same figures, you are free to choose a different colour scheme. In fact, it is recommended that you select a different theme and colour scheme that more clearly communicates the data (see: https://ggplot2-book.org/scale-colour.html or https://www.shanelynn.ie/themes-and-colours-for-r-ggplots-with-ggthemr/).

Assessment criteria

Part A: Figure Reproduction (50 marks)
- Quality and accuracy of data cleaning and processing code (15 marks)
- Visualization re-production (21 marks)
- Quality of data visualization code (14 marks)

Part B: Figure Design (45 marks)
- Quality and accuracy of data cleaning and processing code (10 marks)
- Originality and creativity of visual design (14 marks)
- Clarity and readability of visual representation (14 marks)
- Quality of data visualization code (7 marks)

Overall:
- Notebook organisation and structure (5 marks)

Due Date: Friday 3 May, 13h00 (via SUNLearn link)