Assessment instructions

This assessment is organised into four parts which are code-based questions. You are required to write RMarkdown style of code and description on real-time data. You must attempt all the parts and submit four `.rmd` files that include aspects of your formal reporting (e.g., an introduction, assumptions) using the markdown syntax, mixed with your R Code and results.

Assessment overview

In this assignment, you are given four data files in the CSV format. These files are collected from the internet for the COVID-19 pandemic for different regions and countries around the world for the period between 1/1/2020 and 5/5/2020. The data are described below. However, please note that the data is quite messy. You are asked to write R code (in the RMarkdown format) to import and wrangle these data files and put them in reasonable format to conduct analysis and do data-driven modelling on them.

- Data files in .ZIP formatDownload Data files in .ZIP format
- Template Submission for Part A (RMD Download RMD) (datadesc.PNG)
- Template Submission for Part B (RMD Download RMD)
- Template Submission for Part C (RMD Download RMD)
- Template Submission for Part D (RMD Download RMD)

Data description

The four CSV files are described in the following table.

Assessment 2 data description (UC, 2023)

File name	Description
Covid19.csv	This is the master file that includes information about the countries, continents and the daily new cases and daily new deaths in each country.
Tests.csv	This file lists information about the daily COVID-19 tests for each country.
Countries.csv	This file provides information about the countries.
Recovered.csv	This file presents information about the daily recovered cases in each country.

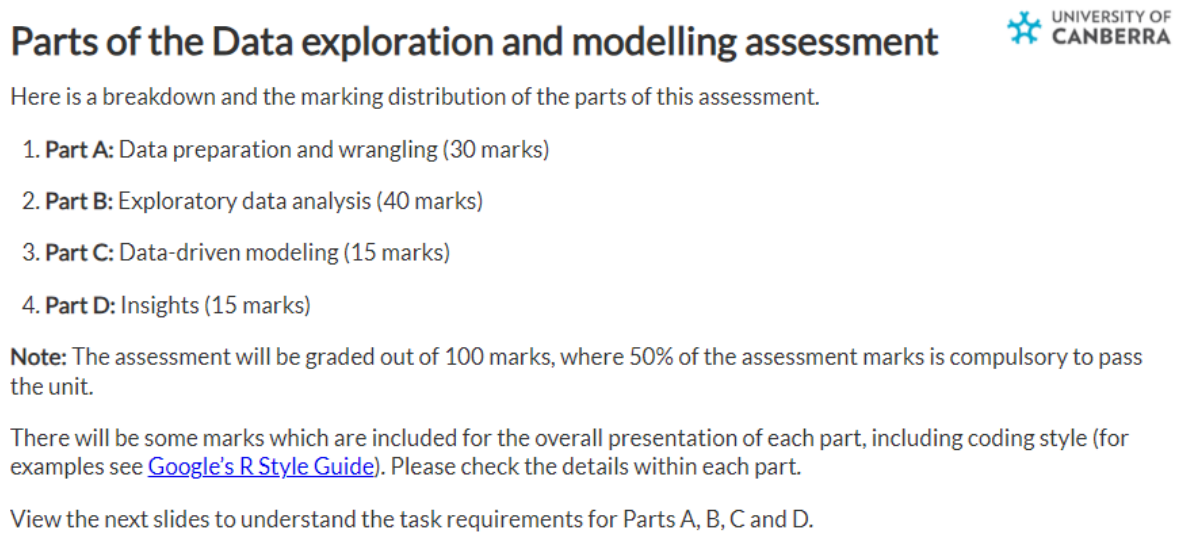
Data copyright

The data has been scrapped from different places on the internet with a focus on the following four sites:

1. COVID-19 public dataLinks to an external site. (GitHub, 2023)
2. Novel Coronavirus (COVID-19) Cases DataLinks to an external site. (John Hopkins, 2023)
3. Coronavirus casesLinks to an external site. (worldometer, 2023)
4. Gross domestic productLinks to an external site. (Wikipedia, 2023)

The assessment is divided into four distinct parts, each comprising multiple sub-parts that necessitate the creation of R scripts within the RMarkdown format. These tasks are deliberately designed to be interconnected, with the output of one task serving as input for subsequent steps in the assessment. You are asked to answer all questions in each part.

Go through the slider below to review the tasks in detail.



Part A: Data preparation and wrangling (30 marks)

UNIVERSITY OF CANBERRA

Your tasks for Part A are:

1. Import the data from the CSV files and store them into dataframes named appropriately.

2. Tidy up the dataframe driven from the "Recovered.csv" files to be compatible with the dataframe driven from the "Covid19.csv" file, *i.e.*, every observation should have a record of recovered patients in one country in a single day.

3. Change the column names in the dataframes that were loaded from the following files accordingly:

File Name	Ordered New Column Names
Covid19.csv	**Code, Country, Continent, Date, NewCases, NewDeaths**
Tests.csv	**Code, Date, NewTests**
Countries.csv	**Code, Country, Population, GDP, GDPCapita**
Recovered.csv	**Country, Date, Recovered**

Note: Tasks 4 to 7 are continued in the next slide.

Part A: Data preparation and wrangling (30 marks) continued

UNIVERSITY OF CANBERRA

4. Ensure that all dates variables are of the same date format across all dataframes.

5. Considering the master dataframe is the one loaded from the "Covid19.csv" file, add new 5 variables to it from the other files (Recovered.csv, Tests.csv, Countries.csv). The 5 new added variables should be named ("Recovered", "NewTests", "Population", "GDP", "GDPCapita") accordingly.

[Hint: you may use the `merge` function to facilitate the alignment of the data of the different dataframes. You may use this format: `merge(x=df1,y=df2, [specify the merging dimension if needed])` , where df1 and df2 are the dataframes to be merged]

6. Check NAs in the merged dataframe and change them to `Zero` .

7. Using existing "Date" variable; add month and week variables to the master dataframe.

[Hint: you may use functions from `lubridate` package]

Note: Tasks 8 to 10 are continued in the next slide.

Part A: Data preparation and wrangling (30 marks)

UNIVERSITY OF CANBERRA

8. Add four new variables to the master dataframe ("CumCases", "CumDeaths", "CumRecovered", "CumTests"). These variables should reflect the cumulative relevant data up to the date of the observation; *i.e.*, CumCases for country "X" at Date "Y" should reflect the total number of cases in country "X" since the beginning of recording data till the date "Y".

[Hint: first arrange by date and country, then for each new variable to be added you need to group by country and mutate the new column using the cumsum function]

9. Add two new variables to the master dataframe ("Active", "FatalityRate"). Active variable should reflect the infected cases that has not been closed yet (by either recovery or death), and it could be calculated from (CumCases – (CumDeaths + CumRecovered)). On the other hand, FatalityRate variable should reflect the percentages of death to the infected cases up to date and it could be calculated from (CumDeaths / CumCases).

10. Add four new variables to the master dataframe ("Cases_1M_Pop", "Deaths_1M_Pop", "Recovered_1M_Pop", "Tests_1M_Pop") These variables should reflect the cumulative relevant rate per one million of the corresponding country population, (i.e Cases_1M_Pop for country "X" at Date "Y" should reflect the total number of new cases up to date "Y" per million people of country "X" population)

Part B: Exploratory data analysis (40 marks)

UNIVERSITY OF CANBERRA

Your tasks for Part B are as follows:

1. Find the day with the highest death toll reported across the world. Print the date and the death toll of that day.

2. Build a graph to show how the cumulative data of (Infected Cases, Deaths, Recovered, Tests) change over the time for the whole world collectively.

[Hint: Use geom_line as a geometry function, use log for the Y axis for better presentation, Use different colour to distinguish between new cases, deaths, and recovered]

3. Extract the data corresonding to the last day *(05/05/2020)* for all countries and save it in a separate dataframe and name it *"lastDay_data"*.

[Hint: use filter function with Date = "2020-05-05"]

Note: Tasks 4 to 6 are continued in the next slide.

Part B: Exploratory data analysis (40 marks) continued

UNIVERSITY OF CANBERRA

4. Based on the data in *"lastDay_data"* dataframe, extract the corresponding records of the top 10 countries worldwide with current active cases, total confirmed cases, or fatality rate in separate dataframes (*i.e.*, top10activeW, top10casesW, top10fatalityW, top10testsMW).

[Hint: you can use head(arranged_data, n=10) to get the top 10 records and pass the records of these 10 countries in newly created data frames]

5. Based on the data of the in *"lastDay_data"* dataframe, print total confirmed cases, death, recovered cases as well as the total tests per every `continent`.

6. Build a graph to show the total number of cases over the time for the top 10 countries that have been obtained in question 4 (Use log transformation for the values in Y axis for better presentation).

[Hint: first you need to get the data of the top-10 countries and then plot their lines, ie, one line per country]

Note: Tasks 7 to 10 are continued in the next slide.

Part B: Exploratory data analysis (40 marks) continued

UNIVERSITY OF CANBERRA

7. Build a graph for the top 10 countries with current highest active cases which was obtained previously in question 4. The graph should have one sub-graph (*i.e.*, using facet function) for each of these countries, every sub-graph should show how the new cases, new deaths, and new recovered cases were changing over the time (Use log for Y axis for better presentation, Use different colour to distinguish between new cases, deaths, and recovered).

[hint: geom_line function with date on x_axis and each of the values of the variables in y_axis]

8. Build a graph for the top 10 countries with current highest total tests per one million of the population which was obtained previously in question 4. This graph should present total number of infected cases, total tests so far, and the total tests per million of the population for each country.

[hint: you can use bar chart to achieve this task]

9. Build a graph to present the statistics total, average, median of confirmed cases of the continents. (you may use log for Y axis for better presentation, Use Continent in the legend, make sure x-axis labels does not overlap).

10. Based on the data of the *"lastDay_data"* dataframe, list the top 2-countries of each continent that report the highest death toll.

Part C: Data-driven modelling (15 marks)

UNIVERSITY OF CANBERRA

Your tasks for Part C are as follows:

1. Based on the covid19_data dataframe, that you have wrangled and used in the previous tasks, create a separate dataframe named "cor_data" with the data of these variables (CumCases, CumTests, Population, GDP, GDPCapita) variables.

[Hint: you can use select function on the covid19_data dataframe]

2. Compute the correlation matrix between the variables of the "cor_data" and visualise this correlation matrix.

3. visualise the distribution of the cumulative cases in the cor_data with and without changing the scale of the x axis to log transformation.

4. Divide the cor_data into training and testing, where training data represent 65% of the number of rows.

Note: Tasks 5 to 7 are continued in the next slide.

Part C: Data-driven modelling (15 marks) continued

UNIVERSITY OF CANBERRA

5. Train a linear regression model to predict cumulative cases from the GDP of the countries. Then, evaluate this model on the test data and print the root mean square error value.

6. Train another linear regression model to predict cumulative cases from all the other variables. Then, evaluate this model on the test data and print the root mean square error value.

7. Interpret the two models and write a small report of highlighting the differences between using the two models. For example, in which cases we should use the first model and in which cases the second one is better to use.

Part D: Insights (15 marks)

UNIVERSITY OF CANBERRA

Imagine you have been asked to plan for a dashboard that shall show the trends and the main figures of the different COVID-19 waves that happened worldwide in the last pandemic. Given that the current data in this assignment is only covering the first wave of COVID-19,

- how would you augment this data?
- What are the other sources of data that you will rely on?
- What types of figures will you be focusing on to show in your dashboard? and why?

You are expected to pay close attention to several key aspects:

1. **Code Structure, Quality, and Functionality:** You should ensure that your code is well-structured, maintains high quality, and functions as intended. This includes adhering to best coding practices and standards.
2. **Objective Clarity:** It's crucial to express the assessment's objectives from your perspective. This clarity helps demonstrate a deep understanding of the tasks at hand.

3. **Documentation:** When necessary, you should provide appropriate documentation within your code to elucidate complex sections or methods, making it easier for others to comprehend your work.
4. **Output presentation:** The final output of the code should be seamlessly **rendered into** an HTML or PDF document. This document should include not only code snippets but also comprehensive descriptions and reflections on the overall findings derived from the analysis.
5. **Innovative approaches**: You are encouraged to go beyond the required tasks and explore unconventional methods of accomplishing them. Think creatively and consider alternative approaches that may lead to unique and insightful data analysis.

In summary, you are encouraged to approach this assessment with a focus on code quality, clear communication of objectives, effective documentation, and a well-structured presentation of your results in the final document.

Deliverables and Assessment formatting guidelines

To submit your work, you must compile all the components into a single .ZIP file and upload it to the designated section on the UC LearnOnline platform for this specific unit.

1. The four `.rmd` files with the markdown reports and code snippets.
2. The four HTML or PDF documents generated by knitting your `.rmd` files.

Please follow the following structure to name the submitted zip file:

`[studentID_assignment1.zip]`

- Students' names are not to be included on any assessment tasks/submissions. Only student ID numbers should be included (as per the Assessment Policy and Assessment Procedures which can be found in the University Policy Library: AcademicLinks to an external site.).
- Please note this assessment will be reviewed by the University's plagiarism checking software (Turnitin) and, with reasonable grounds, be subject to further inquiry through the Office of the Associate Dean of Education.

Oral Evaluation Session:

- After submission, students are required to attend an oral evaluation session (via Virtual Room) with the lecturer/tutor to present their assignment. Each session is expected to last 15 minutes. The schedule for these sessions will be provided to students shortly and will be booked on a first-come-first-served basis.