University of Canberra

Faculty of Science and Technology

FINAL ASSESMENT

SEMESTER 1, 2021

UNIT NAME:	Introduction to Data Science
UNIT NUMBER:	11372
TIME ALLOWED:	8 days
EXAMINER'S NAME:	Dr. Ibrahim Radwan
CONTACT details:	ibrahim.radwan@canberra.edu.au

INSTRUCTIONS FOR STUDENTS

1. **This is a take-home assignment.**
2. **The assignment is organised into two parts, where the first part is composed of four general questions to assess your understanding of the Data Science principles. This part is expected to be delivered as a PDF file. The other part is code-based tasks, which you will need to submit R-code script and a final report. The final report should contain the output of running your code for those questions that are asking you to generate summaries or graphs from the data.**
3. **The PDF files and R script are expected to be submitted as a one compressed file (e.g. *.ZIP) on the Canvas website of the unit by the due date of the assignment.**
4. **The submitted file should be renamed as [studentID_lastname_final_assessment.zip], where "studentID" is your university ID and "lastname" is your lastname.**
5. **The assignment will be open from Friday, the 7th of May (17:30) until Friday, the 14th of May (17:30).**
6. **This assignment has 100 marks in total and weighs 40% of the final grade of the unit, where 50% of the assignment marks is compulsory to pass the unit.**
7. **The assignment will cover all the learning outcomes and the taught contents of the unit.**

There are no errors deliberately placed in any of the questions in this assignment, unless explicitly stated. If you think you have identified any errors, please contact the supervisor.

Part A – Data Science Questions (15 *marks*)

There are four questions in this part, with differing marks. All answers must be recorded in a MS Word and then exported to PDF file.

Q1) (3 marks)

From your understanding of ethical data science, mention three principles of a code of ethics that any data scientist should consider.

Write your answer as:

P1:____

P2:____

P3:____

Q2) (4 marks)

To build a visualisation using the ggplot2 library, we use the following template:

```
ggplot(data= [dataset], mapping = aes(x = [x-variable], y = [y-variable]))+
  geom_xxx() +
  other options
```

Based on the above template, mention the main components of building a graph using ggplot2 and describe the meaning of each of these components.

Write your answer as:

____

____

____

____

____

Q3) (3 marks)

Describe three properties of the correlation coefficient of two variables

Write your answer as:

1. ____
2. ____
3. ____

   ____

Q4) (5 marks)

Imagine we have a dataset that lists the heights of the fathers and their sons. You have built a linear model that encodes the relationship between the fathers' heights and the sons' heights as follows:

```
lm(son ~ father, data = heights_data)


Call:

lm(formula = son ~ father, data = heights_data)


Coefficients:

(Intercept)    father

    70.45       0.50
```

The estimated coefficient (i.e. intercept and slope), which describes the relationship between the fathers' and sons' heights can be interpreted as:

____

____

____

____

____

Part B –Data Preparation, exploring and modelling (85 *marks*)

In this part, you are given four data files in CSV format, that are collected from the internet for the COVID-19 pandemic for different regions and countries around the world. The data are described below. The data are quite messy. You are asked to write R scripts to import and to wrangle these data files and put them in reasonable format to conduct analysis and do data-driven modelling on them.

This part is consisted of four tasks, that are listed below in details. You are asked to write R-code for all of questions in each task.

Data Description:

The four CSV files are described in the following table:

File Name	Ordered New Column Names
Covid19.csv	**This is the master file that include 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:

1. **https://github.com/owid/covid-19-data/tree/master/public/data/**
2. **https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases**
3. **https://www.worldometers.info/coronavirus/**
4. **https://en.wikipedia.org/wiki/Gross_domestic_product**

Task 1: Data Preparation and Wrangling: (20 marks)

1. **Load and read the data from the CSV files and store them into dataframes named appropriately.**

2. **Tidy up the dataframe driven from the file "Recovered.csv" to be compatible with the dataframe driven from the file "Covid19.csv", 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 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**

4. **Ensure that all dates variables are of date data type and with the same format across the dataframes.**

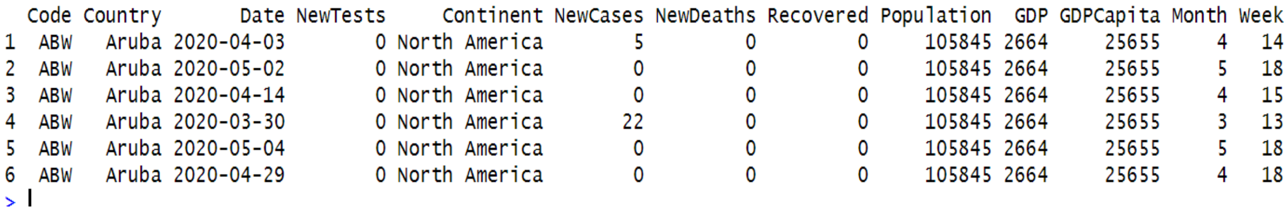
5. **Considering the master dataframe is the one loaded from file “Covid19.csv”, add new 5 variables to it from other files (Recovered.csv, Tests.csv, Countries.csv). The 5 new added variables should be named (“Recovered”, “NewTests”, “Population”, “GDP”, “GDPCapita”) accordingly.**

   [Hint: you can use the merge function to facilitate the alignment of the data in the different dataframes.]

6. **Check for Nas in all dataframes 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]**

[Hint: To ensure that this task has been finished correctly, when you run head(covid19_data), you should get results such as in the below image]



```
  Code Country       Date NewTests     Continent NewCases NewDeaths Recovered Population  GDP GDPCapita Month Week
1  ABW   Aruba 2020-04-03        0 North America        5         0         0     105845 2664     25655     4   14
2  ABW   Aruba 2020-05-02        0 North America        0         0         0     105845 2664     25655     5   18
3  ABW   Aruba 2020-04-14        0 North America        0         0         0     105845 2664     25655     4   15
4  ABW   Aruba 2020-03-30        0 North America       22         0         0     105845 2664     25655     3   13
5  ABW   Aruba 2020-05-04        0 North America        0         0         0     105845 2664     25655     5   18
6  ABW   Aruba 2020-04-29        0 North America        0         0         0     105845 2664     25655     4   18
> 
```

Task 2: Exploratory Data Analysis: (40 marks)

1. **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]

2. **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).**

3. **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)**
   [Hint: Cases_1M_Pop = CumCases(10^6) / Population)]*

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

5. **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, use log for Y axis for better presentation, Use different colour to distinguish between new cases, deaths, and recovered]

6. **Extract the last day (05/05/2020) data and save it in a separate dataframe called "lastDay_data".**
[Hint: use filter function with Date = "2020-05-05"]

7. **Based on the last day data, extract the whole records of the top 10 countries worldwide that have current active cases, total confirmed cases, and 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]

8. **Based on the last day data, print the up to date confirmed, death, recovered cases as well as the tests for every continent.**

9. **Build a graph to show the total number of cases over the time for the top 10 countries that have been obtained in question 7 (Use log for Y axis for better presentation).**
[Hint: first you need to get the data of the top-10 countries and then plot their lines]

10. **Build a graph for the top 10 countries with current highest active cases which was obtained previously in question 7. The graph should have one subgraph (i.e. using facet function) for each of these countries, every subgraph should show how the new cases, new deaths, and new recovered cases were changing over 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]

Task 3: Data-Driven Modelling: (15 marks)

1. **Based on the data of the last day, that you have extracted in the previous task, create a separate dataframe named "cor_data" with the data of these variables (CumCases, CumTests, Population, GDP, GDPCapita).**
[Hint: you can use select function on the lastday_data dataframe]
2. **Compute the correlation matrix between the variables of the "cor_data" and visualise this correlation matrix.**

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

4. **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.**

5. **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.**

Task 4: Documentation and Reporting: (10 marks)

You are required to build a report (e.g. using MS Word) for the results of Task 2 and Task 3. The report is basically composed of the answers to those questions that asked you to generate or print summaries or asked you to build graphs for some variables. Then you need to export this MS word file into a PDF file to be submitted.

Deliverables

You are required to submit a compressed (e.g. ZIP) file to Canvas with the following files:

1- A PDF document for the answers to questions of part A

2- A PDF document for the summaries and graphs as described in task 4 of part B

3- A single R script file with the code for the tasks in part B.