**Project**

In this project, 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 to import and 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**:

Text, table

Description automatically generated with medium confidence

#### Task 1:

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.

Text, table

Description automatically generated

1. Ensure that all dates variables are of date data type and with the same format across the dataframes.
2. Considering the master dataframe is the one loaded from file “Covid19.csv”, 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, all.x=TRUE), where df1 and df2 are the dataframes to be merged]

1. Check for NAs in all dataframes and change them to Zero.
2. 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]

Chart

Description automatically generated with medium confidence

#### Task 2:

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]

1. 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).
2. 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)]

1. Find the day with the highest reported death toll 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]

1. Extract the data corresonding to the last day (05/05/2020) and save it in a separate dataframe and name it “lastDay\_data”.

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

1. Based on the data of the last day, extract the 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]

1. Based on the data of the last day, print the up to date confirmed, death, recovered cases as well as the tests for every continent.
2. 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]

1. 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 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]

#### Task 3:

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]

1. Compute the correlation matrix between the variables of the “cor\_data” and visualise this correlation matrix.
2. Divide the cor\_data into training and testing, where training data represent 65% of the number of rows.
3. 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.
4. 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:

Imagine you have been asked to plan for a dashboard that shall show the trends and the main figures of the different Covid19 waves that happened world-wide, so far. Given the current data in this assignment is only covering the first wave of the Covid19, 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?

Write the report as follows:

1. Objectives:

-----------------------------------------

-----------------------------------------

-----------------------------------------

2. List of data sources to augment the existing data:

-----------------------------------------

-----------------------------------------

-----------------------------------------

3. Set of figures/tables to show in the dashboard:

-----------------------------------------

-----------------------------------------

-----------------------------------------

4. Analysis strategy:

-----------------------------------------

-----------------------------------------

-----------------------------------------