Overview

Now that you have been equipped with the knowledge and skills to process and analyze data with R, you will have the opportunity in this final project to practice and apply it by analyzing real-world COVID-19 testing data online.

Scenario

In this scenario, you will play the role of a data analyst who has recently joined the data science team of a news channel. Your channel's documentary team channel's recently accepted an assignment
to create a feature story on global COVID-19 testing by country and they need some data-driven insights from your team.Your team has been using IBM Watson Studio, which is a powerful and easy-to-use platform for data analysis and data science tasks. As a new team member, your team leader has assigned this task to you. You need to leverage your R skills to acquire the relevant datasets, as well as process and analyze them using Watson Studio.

Watson Studio

For this project, you will use Watson Studio. Watson Studio is a component of IBM Cloud Pak for Data, is a suite of tools and a collaborative environment for data scientists, data analysts, AI and machine learning engineers, and domain experts to develop and deploy your projects.

Step-By-Step Labs

In this final project, you will webscrape a global COVID-19 dataset from a public wiki page, and conduct various data analysis tasks on the data set.

To help you finish the final project, we have prepared several readings and labs:

- Create IBM Cloud account and Watson Studio instance
- Download a skeleton Jupyter notebook with tasks, sample code, and comments via provided URL
- Add the skeleton Jupyter notebook to Watson Studio
- Analyze global COVID-19 testing data
- Share your Jupyter notebook via Watson Studio for peer-review

Submission

A screenshot in JPEG format is required to be submitted for solution to each of the problems. The screenshot for each task should clearly show the code and output.

The screenshots will be uploaded in the following sections.

Example Submissions

Here is an example of a submission clearly showing both the code and output, when executed from a Jupyter notebook.

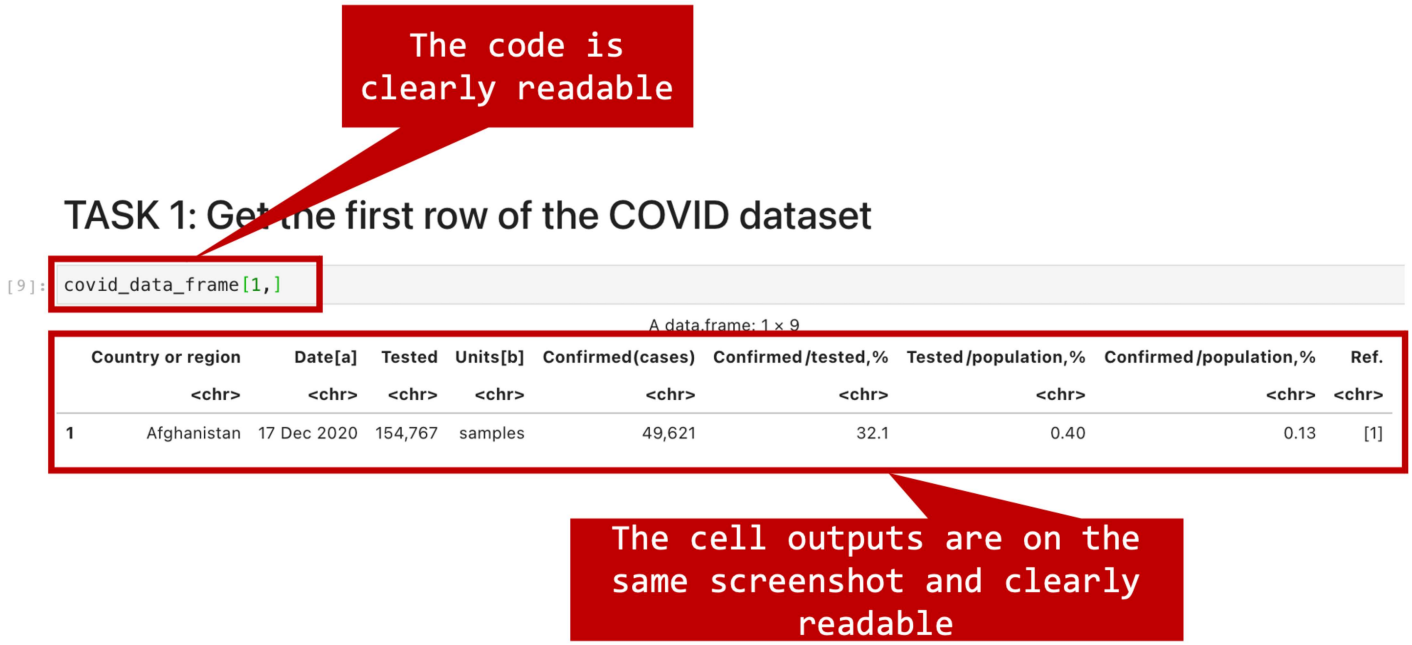
TASK 1: Get the first row of the COVID dataset

```
[9]: covid_data_frame[1,]
```

A data.frame: 1 × 9

	Country or region	Date[a]	Tested	Units[b]	Confirmed(cases)	Confirmed/tested,%	Tested/population,%	Confirmed/population,%	Ref.
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	Afghanistan	17 Dec 2020	154,767	samples	49,621	32.1	0.40	0.13	[1]

Important aspects of the submission are highlighted below for illustrative purposes:



Grading scheme

There are **20** total points possible for this assignment. Here is the breakdown:

- Task 1: Get a COVID-19 pandemic Wiki page using HTTP request (2 pts)
- Task 2: Extract COVID-19 testing data table from the wiki HTML page (2 pts)
- Task 3: Pre-process and export the extracted data frame (2 pt)
- Task 4: Get a subset of the extracted data frame (2 pt)
- Task 5: Calculate worldwide COVID testing positive ratio (2 pts)
- Task 6: Get a sorted name list of countries that reported their testing data (2 pts)
- Task 7: Identify country names with a specific pattern (2 pts)
- Task 8: Pick two countries you are interested in, and then review their testing data (2 pts)
- Task 9: Compare which one of the selected countries has a larger ratio of confirmed cases to population (2 pts)
- Task 10: Find countries with confirmedcases to population ratio rate less than a threshold (2 pts)

Author(s)

Yan Luo

Change log

Date	Version	Changed by	Change Description
2021-03-31	1.0	Yan Luo	Created the initial version