

# SIT112 | Data Science Concepts

Lecturer: Dr Sergiy Shelyag  
[Sergiy.shelyag@deakin.edu.au](mailto:Sergiy.shelyag@deakin.edu.au)

## ASSIGNMENT ONE

Due: 8pm, Friday 16 April 2021

**Note:** This assignment contributes 25% to your final SIT112 mark. It must be completed individually and submitted to *CloudDeakin* before the due date: **8pm, 16 April 2021**.

The theme for this assignment is to explore data related to COVID-19. In particular, we will use a public dataset (provided by OWID – *Our World In Data* website), which provides up-to-date information on COVID-19 epidemic parameters and evolution, as well as economic and sociological parameters for all countries of the world. Our data strategy and task specifications for this assignment will focus on the analysis and descriptive analytics of how COVID spreads in the world.

### 1. Data and Resources

In the Assignment 1 folder, you will find the following files:

Filename	Description
<b>datadictionary_template.xlsx</b>	This is the template for the data dictionary file in Excel.
<b>assignment1_notebook.ipynb</b>	This is the Jupyter Notebook, which has been prepared and pre-filled for you to complete the programming task.
<b>Assignment1_instructions.pdf</b>	These instructions.

Please visit the coronavirus source data website: <https://ourworldindata.org/coronavirus-source-data> and read the instructions. The dataset is updated daily, therefore you need to download the most recent **.csv** file from the following link:

<https://covid.ourworldindata.org/data/owid-covid-data.csv>

**Place the file in the assignment folder.**

The description of variables / fields in the data file is available at:

<https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv>

### 2. Task Description

There are two main tasks for this assignment:

- Construction of the data dictionary (**35 marks**) and
- Programming tasks to perform data analysis and descriptive analytics (**65 marks**).

## 2.1 Construction of the Data Dictionary (35 marks)

For a data scientist, after obtaining the dataset, the first most crucial task is to obtain a good understanding of the data they are dealing with. This includes: examining the data attributes (or, equivalently, data fields), seeing what they look like, what is the data type for each field, and, from this information, determining suitable analysis tools. *A systematic approach to this process, as we have learned from the lectures and practical sessions, is to construct a data dictionary for the dataset.*

You are required to prepare **two sheets** in your data dictionary Excel file:

- Dataset description (**5 marks**)
- Attribute dictionary (**30 marks**)

The total for this task is 35 marks. The data description sheet is worth 5 marks. The attribute dictionary is worth 30 marks, where each correct attribute specification is worth 2.5 marks. Name your solution as *[YourID]\_datadictionary.xls* and submit this file.

## 2.2 Programming task (65 marks)

A Python Jupyter Notebook file *assignment1\_notebook.ipynb* has been prepared for you to complete this task. Download this notebook, load it up to Jupyter and follow instructions inside the notebook to complete this task.

The total for this task is 65 marks.

*You are required to submit your solution in Jupyter Notebook format as well as its exported version in html.*

## 3. Summary for submission

This assignment is to be completed individually and submitted to CloudDeakin. By the due date, you are required to submit the following files to the corresponding *Assignment* (Dropbox) in *CloudDeakin*:

1. **[YourID]\_datadictionary.xls**: your solution for the data dictionary for the given dataset;
2. **[YourID]\_assignment1\_solution.ipynb**: your Jupyter Notebook solution source file;
3. **[YourID]\_assignment1\_output.html**: the output of your Jupyter Notebook solution in html.

For example, if your student ID is: 123456, you will then need to submit **three** files:

- **123456\_datadictionary.xls**
- **123456\_assignment1\_solution.ipynb**
- **123456\_assignment1\_output.html**

**END OF ASSIGNMENT DESCRIPTION**