**DSC-540\_Data-Preparation**

**COVID-19-Program**

COVID-19 has infected more than 9000 people in South Korea. KCDC (Korea Centers for Disease Control & Prevention) announces the information of COVID-19. The following are the structured datasets based on the report materials of KCDC and local governments. I am going to analyze and visualize the data using various data mining or visualization techniques.

**The three different datasets:**

* **Dataset1:** PatientInfo.csv

1. **Source :** [**https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv**](https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv)
2. **Introduction:** PatientInfo is the epidemiological data of COVID-19 patients in South Korea. Showing 8 to 33 of 2,771 entries(observations), and 18 total columns(variables)

* **Dataset2:** Region.csv

1. **Source:**[**https://www.kaggle.com/kimjihoo/coronavirusdataset#Region.csv**](https://www.kaggle.com/kimjihoo/coronavirusdataset#Region.csv)
2. **Introduction:** location and statistical data of the regions in South Korea. This dataset has 244 observations(rows), and 12 variables(columns).

* **Dataset3:** Weather.csv

1. **Source:** [**https://www.kaggle.com/kimjihoo/coronavirusdataset#Weather.csv**](https://www.kaggle.com/kimjihoo/coronavirusdataset#Weather.csv)
2. **Introduction:** data on the weather in the regions of South Korea. This dataset has 24,799 observations(rows), and 10 variables(columns).

* **Website** : [https://datacatalog.worldbank.org/](https://slack-redir.net/link?url=https%3A%2F%2Fdatacatalog.worldbank.org%2F&v=3)
* **API**: <https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic_by_country_and_territory>

**The relationships between the datasets**

The datasets are related, so I do not need to manipulate the data to create a relationship between them. They meet the requirements of having a minimum of 1000 rows across all datasets. And each of the dataset should have a minimum of 10 columns(variables). The column named "province" is common to all three chosen datasets.

Importing and cleaning data by performing exploratory data analysis (EDA). And construct questions that lead to deeper analysis.

**Program use: Jupiter Notebook**