**Datasets Description**

There are three datasets used in my project. I downloaded the first dataset from Kaggle [1] and it stores the data about accumulative confirmed, death, and recovered cases of most cities in the world from January to August. I used the OpenCage Geocoder API to achieve the latitude and longitude of each city. I applied the BigQuery API to achieve another two datasets from the NOAA\_GSOD database. The second dataset is a table containing station information [2]. It records station number (saf), weather bureau army number (wban), country and its latitude and longitude. The third one is a weather table [3]. This dataset provides saf and wban the same as the second table. Besides, it contains information about temperature, dew point, sea level pressure, wind speed precipitation, and humidity.

Firstly, I merged two datasets retrieved from the NOAA\_GSOD dataset to a weather table based on the same columns, saf, and wban in both tables. I also added a new column called ‘day\_from Jan\_first’ to concatenate the weather table and the first table to a table called COVID-weather table. This table contains all the necessary information for my project. For each record in the dataset, it includes attributes like the date, city/state, country, and six weather features. For cleaning the dataset, I set zero to all empty cells in the tables and removed the records without information in the state column.

Datasets links:

[1] <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv>

[2] <https://www.kaggle.com/carinazhao/stations>

[3] <https://www.kaggle.com/carinazhao/gsod2020>