Noor Habib

QinQin Zha

Lei Sabarre

ETL Technical Report

**Extraction:**

The original sources were taken from Kaggle (<https://www.kaggle.com/imdevskp/corona-virus-report>) and there were a total of five CSV files that had data of COVID-19 recoveries, confirmations, deaths from 1/22/2020 through 2/20/2020. The original data also noted the geographical locations and contained columns for latitude, longitude, country and/or province. Only three out of five CSVs were used to create an aggregate that only displays the number of deaths, confirmations and recoveries that happened daily for the entire span of the month.

The database was created using SQL - only one simple table was created to present the tabulation.

**Transformation**:

Each CSV chosen noted the geographical locations and contained columns for latitude, longitude, country and/or province. Each table had 76 rows and 35 columns and were cleaned to a simple one column that had the dates and another column that reported the incidence (death/confirmation/recovery). After all three data frames were created (one per incidence), all were joined together to form one data frame that has a simple aggregate of these incidences daily.

**Load**:

The final database contains the dates, deaths, recoveries and confirmations from 01/22/2020 to 02/20/2020. Any user of this database (contains a single table) can use a simple query to check the number of deaths/recoveries/confirmations per day, as well as manipulate the data to either create more aggregates, or calculate other statistical analysis (mean/mode/median) across the span of one month. Dropping the other columns that indicated geographical data simplified the data frame and allows for simple statistical analysis without interference or confusion due to the various locations. Some locations on the table also were missing data points, or there were too many locations to create a cohesive table.