Dale Romero

Qinqin Zha

ETL Technical Report

**Extraction:**

The original sources were taken from Kaggle (<https://www.kaggle.com/parulpandey/covid19-clinical-trials-dataset>) and this dataset consists of clinical trials related to COVID 19 studies presented on the site.

The annual percent change GDP growth data was taken from International monetary fund IMF website (<https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD>).

We use pandas import both csv and excel file to dataframe.

**Transformation:**

Our plan for selecting the data we chose was to compare the number of clinical trials to the projected GDP by country. We thought it would be interesting to see how the number of Covid-19 clinical trials by country correlate to the projected GDP of that country.

Each dataframe have extra columns and not available data. We did drop some columns, dropna, renamed columns name and reset index.

To join the data, we had to perform cleanup on the country field (which was used to join the data). The Clinical Trial file had a “Location” field that listed the research institution, city, state, country, and in some cases, multiple institutions within that country. To normalize the data and get a clean “country” we had to perform a couple functions on the data.

* We had to pipe-delimit the location column (and grab the first location).
* We then had to parse the comma-delimited result, \and select the data after the last comma. This was then stored as the Country.

On the IMF data file, we had to trim/strip the country because it was being loaded with a leading space.

We then had to create a dictionary for a handful of records where the countries were not an exact match. For example, one file had “China” as the country, while the other had “People’s Republic of China.” We used replace function to get country name harmonized. After both data frames were cleaned, we joined together to form one dataframe that has combined data information.

**Load**

We then used Pandas SQL Alchemy to inject the final dataframe into a PostGres SQL database and confirmed data has been added by querying the covid\_CT\_GDP table.

The final database contains the Clinical trial number, study type, start data, completion data, country, 2016, 2017, 2018, 2019 and 2020 GDP growth rate. Any user of this database (contains a single table) can use a simple query to check the numbers of clinical trial in different country and country GDP growth rate, clinical trial study type as well as study duration days or get correlation of clinical trials with country GDP growth rate.