ETL Project

Greg O’Brien

Jin Dai

Meg Peterson

Michael Gonzalez

Volkan Kaya

Summary:

We used corona virus data that included the number of cases confirmed, people that recovered, and recorded deaths in different cities worldwide every day for the past few months. We extracted csv files, transformed them using pd.melt() and merge, and loaded them into a postgres database.

Extract:

Here is the link where we found our data: <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>. The data were 3 different csv files that we read into a .ipynb file where we used pandas to begin our transformation process.

Transform:

We analyzed the data and realized that we wanted to transform the tables to have cities and dates as row values where the dates originally were column headers. At first, we used pd.Transpose(), it accomplished half of what we wanted but our cities became the headers. To get our table the way we wanted to format, we used pd.melt(). This allowed for each city being the first column with each date associated with the city. Therefore, every date has the data for each city. Then we merged our 3 tables, so that our columns include:

* Province/State
* Country/Region
* Date
* Recovered Count
* Deaths Count
* Confirmed Count

Load:

We loaded our table into a postgresSQL database. We used sqlalchemy to connect our table to an existing database called corona\_db and created the corona\_virus table. Using create\_engine we connected to the database and pd.to\_sql() to insert our DataFrame into the database.

Extra:

We created an app.py file where you can see all the data or search by different province/state. We did this by using the pd.to\_html to create tables for the page. The province class uses the province to query search the postgres database and create pages for a certain province. We also made graphs to visualize our data set in our .ipynb file.