**ETL-Project Report**

Harmony Mosby, Robert McLauchlan, Daniel Love

Harmony, Robert and Daniel's baller transformation of COVID-19/Weather data which could potentially save lives.

What did you do with your Saturday morning?

We created a Database of confirmed cases for the Corona Virus (Covid – 19) and combined that with Weather Data from those areas to see if there is a correlation between weather and the spread of confirmed cases from China to other countries.

**Extract - Data Sets used for our Database**

**Data Set #1 used is from https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases**

The CSV gathered includes the Novel Corona Virus (COVID – 19) epidemiological data from January 22, 2020 to March 6, 2020.

Fields available within the data include Province/State, Country/Region, Latitudeand Longitude, Date of Last  Update, Confirmed, Suspected, Recovered, Deaths.  For our purposes for this project we are only using the Confirmed cases.

**Data Set #2 used is from http://ncdc.noaa.gov/cdo-web/search**

The CSV gathered includes Weather Data during the same sampled time frame.

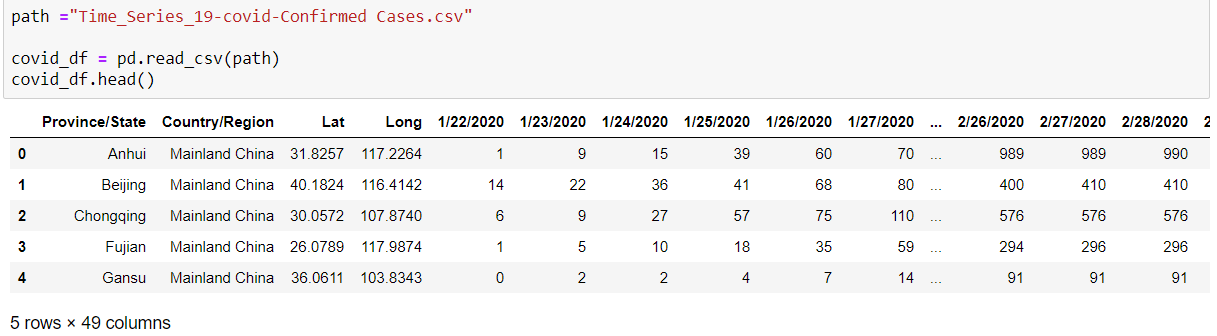
Fields available include State/Country, Latitude and Longitude, Precipitation, Minimum, Maximum and Average Temp for our chosen date ranges.

**Transform - Data Cleanup and Analysis**

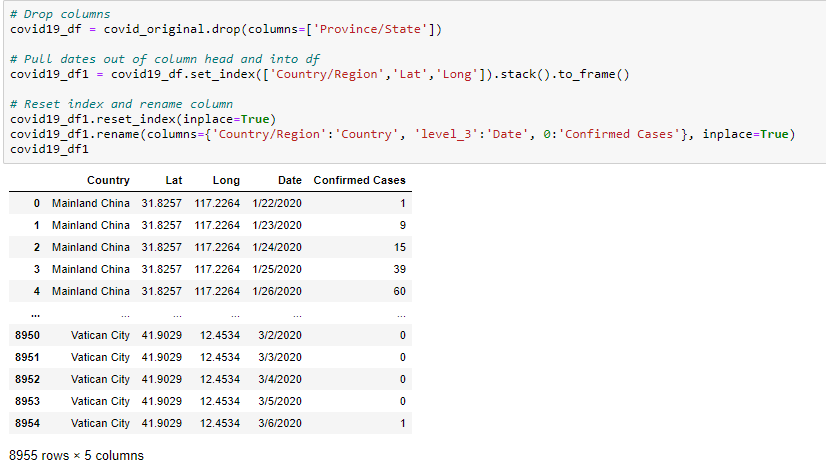
We used Pandas to import our CSVs into a dataframe to clean up the data.

First, we cleaned the Covid - 19 data - confirmed cases.

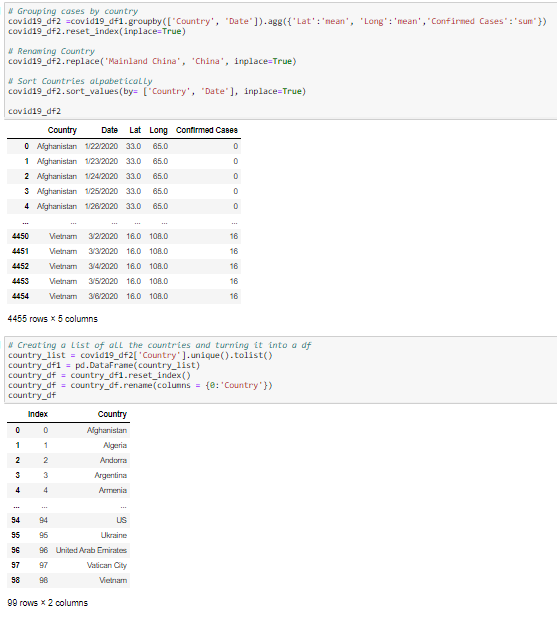
The initial CSV was broken out over time with confirmed cases spanning individual dates from 1-22 through 3-6 which produced 49 columns.



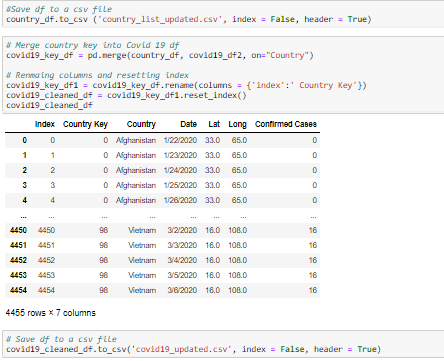
Our next step was to take the individual dates listed as columns in the above table and create rows instead. This would allow for adding in additional data later.



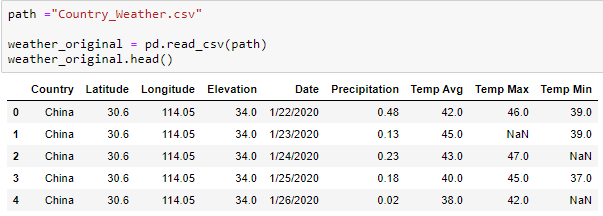
We knew that the Country would be our Foreign key within the final SQL tables created along with a 3rd table for ID and Country data using the ID as our Primary key within the ERD used for the database. We worked to clean up the country listing to correspond to our weather data country listing. For example, changing Mainland China to just be China. Then we grouped cases by country and created a country dataframe resulting in 99 unique country IDs.



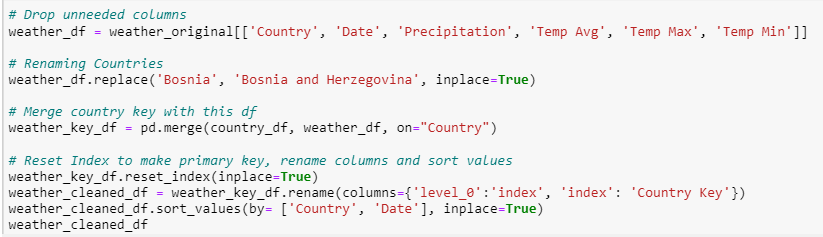
We saved the dataframe to a csv to be imported into pgadmin and used in our database. We then merged the country key into our covid-19 data, renamed columns and reset the index.

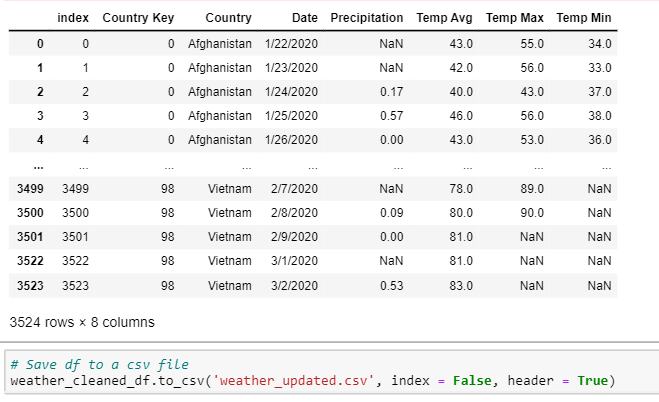


Second, we cleaned the Weather Data – our initial data set was set up to look like the following, as you can see with the date column having individual dates in rows much like our Covid-19 data.

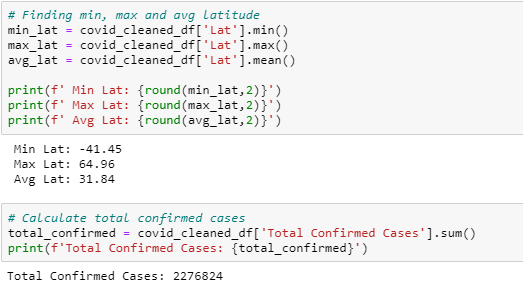


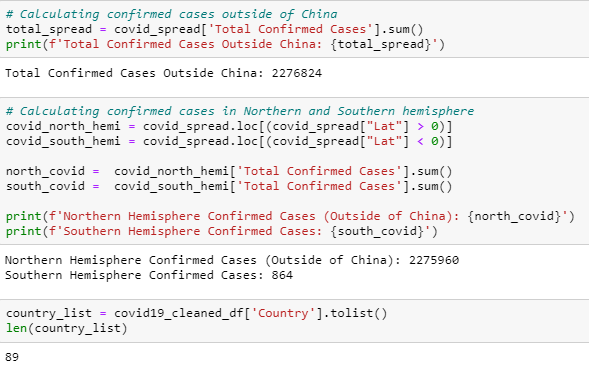
We began by dropping some unneeded columns ­­­­for example elevation and Longitude/Latitude and then merged our country key with our weather data.





Further analysis found in our previous workbooks: we also did some further analysis of total cases and calculated those numbers for Mainland China vs the spread of Covid 19 to  other countries as well as the calculations of cases of Northern Hemisphere vs Southern Hemisphere to be used in our final analysis.

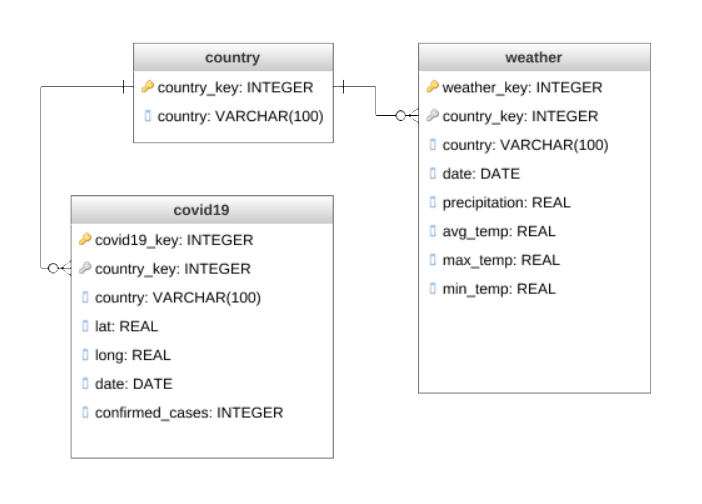




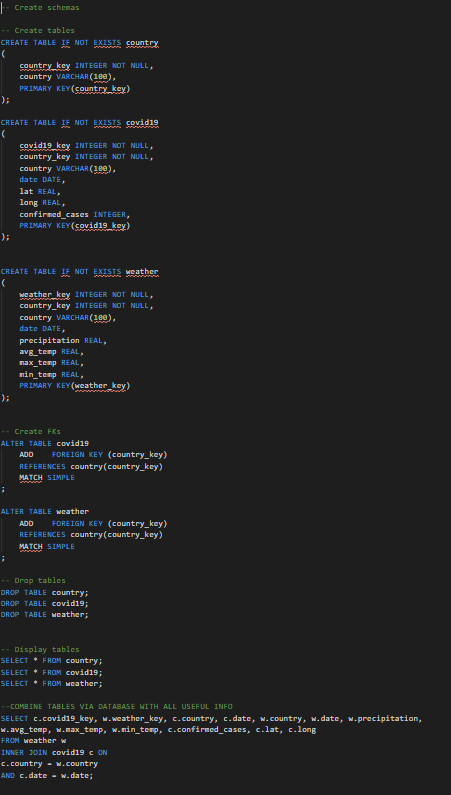
Once both data sets (Covid – 19 and Weather) were cleaned via Pandas in a dataframe, they were saved back to a csv to be used with Postgres and PgAdmin.

**Load – Uploaded data to SQL (GenMyModel and PgAdmin)**

We began by creating an ERD within GenMyModel to connect our Covid - 19 data with our Weather data using the Country as the foreign key and the ID as the primary key.



Next using PgAdmin we created a Database called ETL - Project and imported our cleaned CSVs. We created a schema shown below and 3 tables, a country table, Covid – 19 table and weather data table.



We chose the relational database method so that we could represent and store data in tables and rows as opposed to storing data in collections in the JSON format.

**Final Analysis**

Yes, there is a correlation between the weather and the spread of Covid – 19 into other countries. The total confirmed cases in the Northern Hemisphere are 2,275,960 with only 864 confirmed cases within the Southern Hemisphere. This suggests that since the Northern Hemisphere is experiencing its winter and the Southern Hemisphere its Summer that in warmer weather the spread of this virus is less rampant.

