**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.

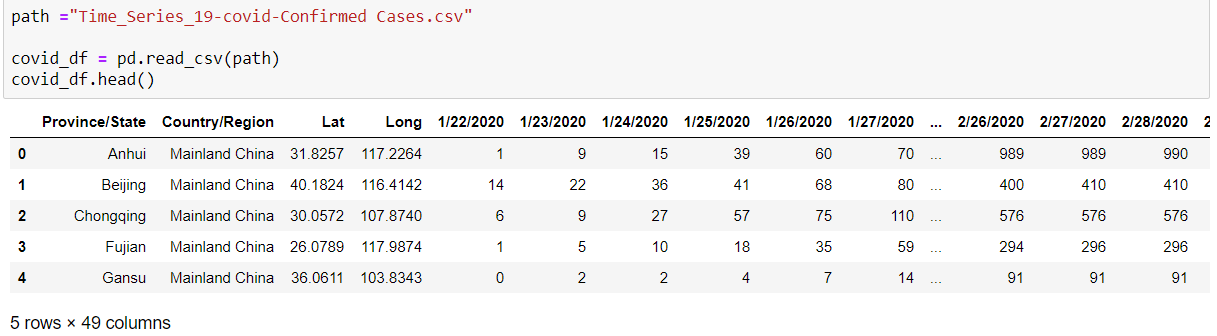
Fields available include State/Country, Latitude and Longitude, Precipitation, Minimum, Maximum and Average Temp for our chosen date ranges.

**Transform - Data Cleanup and Analysis**

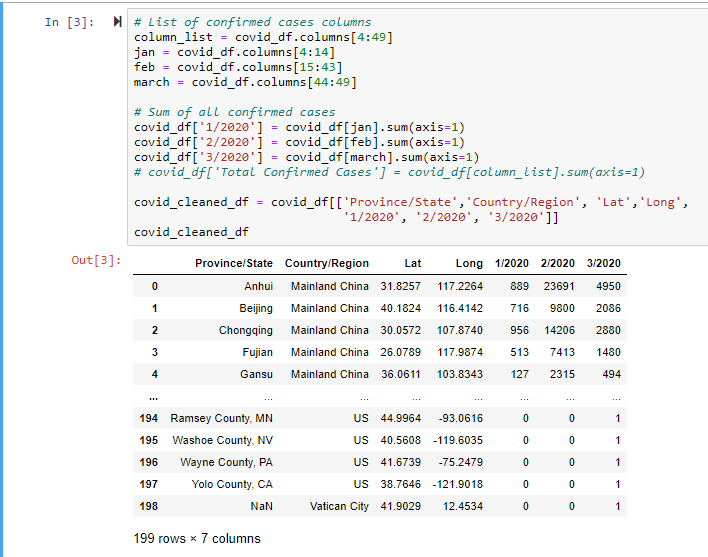
We used Pandas to import our CSVs into a dataframe to clean up the data.

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.



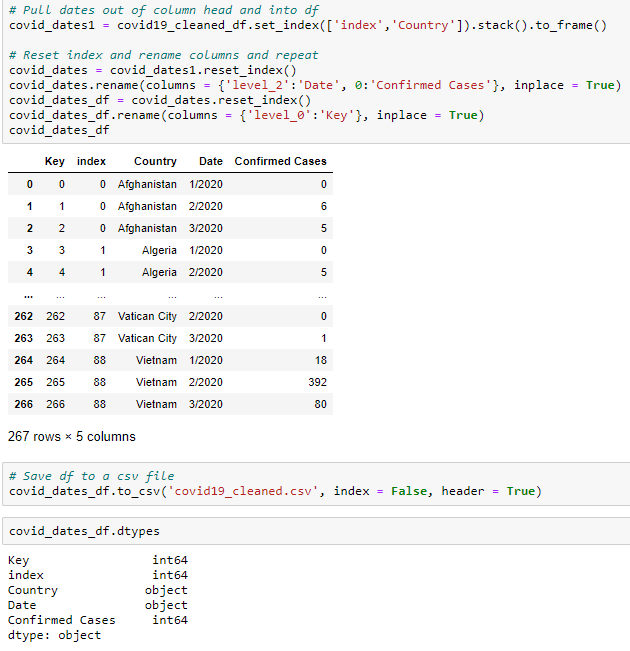
Our next step was to combine the individual dates into monthly data for each Province/State and Country/Region.



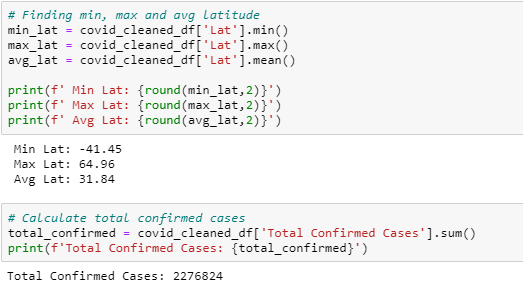
We knew that the Country would be our Primary Key within the final SQL tables created along with a 3rd table for ID and Country data using the ID as our Foreign key within the ERD used for the database. We worked to clean up the country listing to correspond to our weather data country listing. Some examples that we changed  are both San Marino and Vatican City which are micro-states within Italy, we changed US to be United States and changed UK to be United Kingdom. We ended up with a total Country listing of 89 countries.

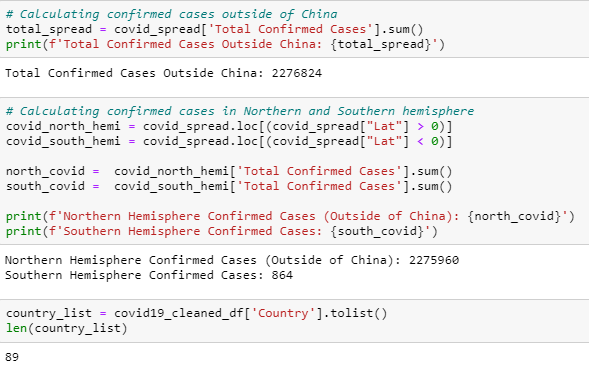




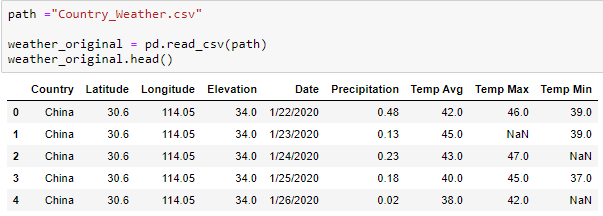


We also did some further analysis of total cases and calculated those numbers for Mainland China vs the spread of Covid 19 to other countries as well as the calculations of cases of Northern Hemisphere vs Southern Hemisphere

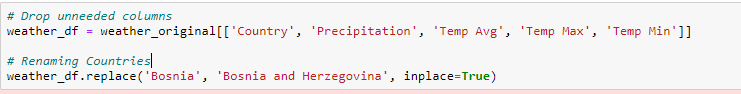


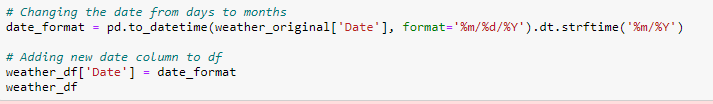


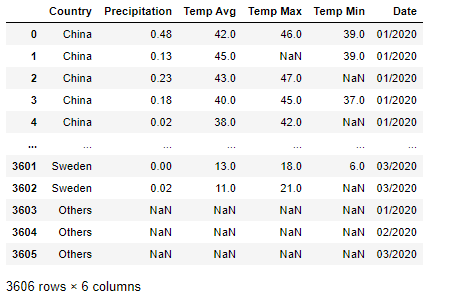
Second, we cleaned the Weather Data – our initial data set was set up to look like the following, as you can see with the date column having individual dates this produced many line items per country and therefore thousands of line items total.



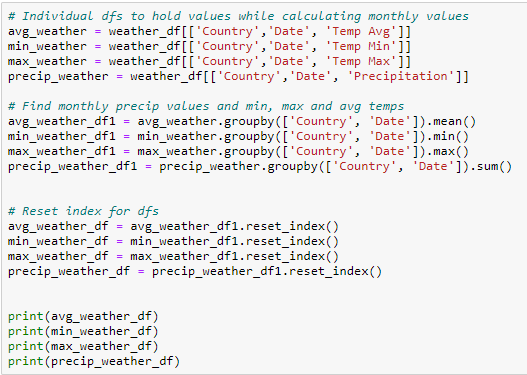
We began by dropping some unneeded columns ­­­­for example elevation and Longitude/Latitude and determined we need to change the individual days within the data to months.





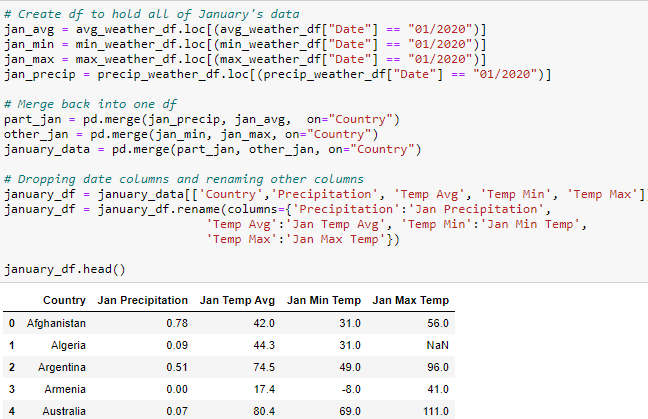


As you can see from the output, we still had 3,606 rows. We then went through a process of calculating the average, minimum and maximum temps for each month for  each country and then printed this data to check our information.

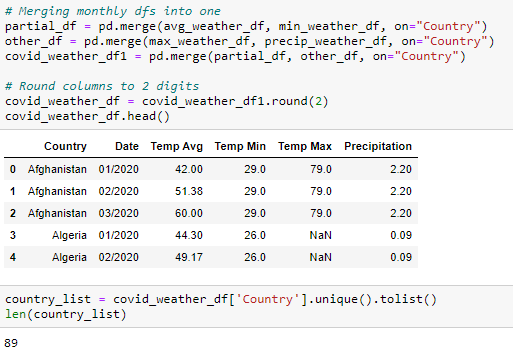


Once that was complete different dataframes were created for each month to combine all data for each month  and then merged back together that way we went from listing out individual monthly dates to a listing of the  countries with an Average temp, min temp and max temp for each month.

The following example is for January, each month can be seen in our Jupyter Notebook in our ETL – Project Repository.



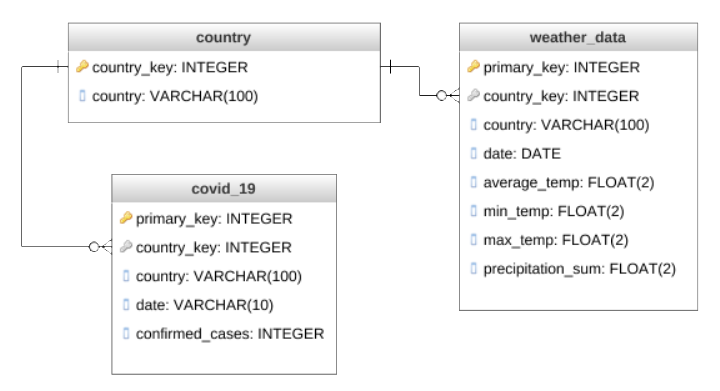
Once each monthly dataframe was created, they were merged back into one so now we had 3 entries for each Country.



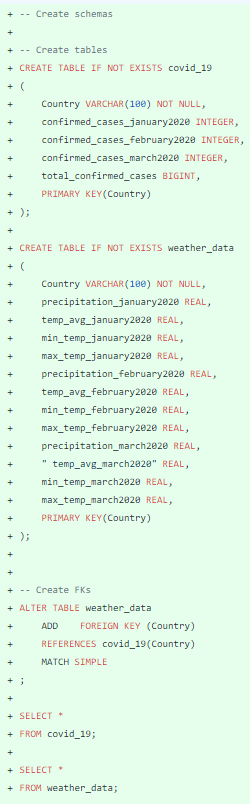
Once both data sets (Covid – 19 and Weather) were cleaned via Pandas in a dataframe, they were saved back to a csv to be used with Postgres and PgAdmin.

**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 primary key and an ID as the foreign 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.

