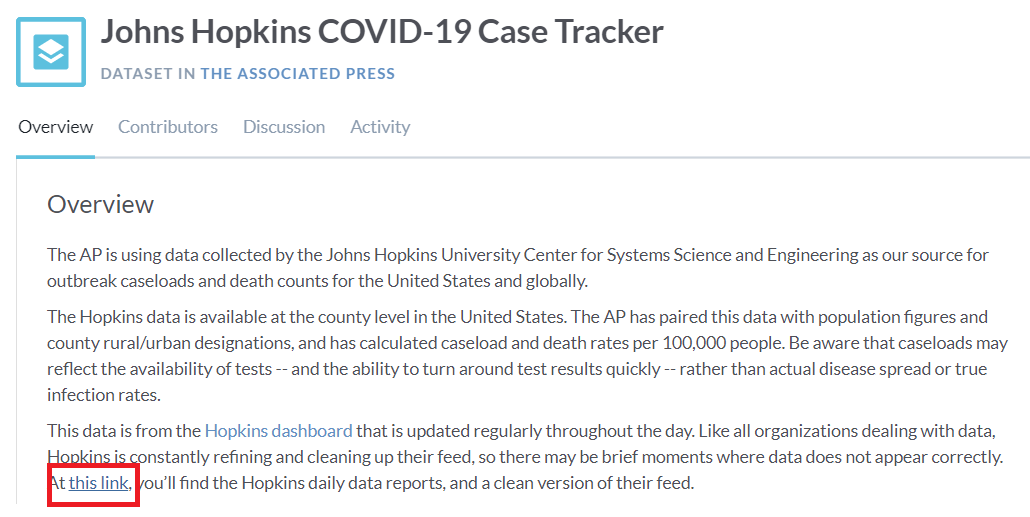
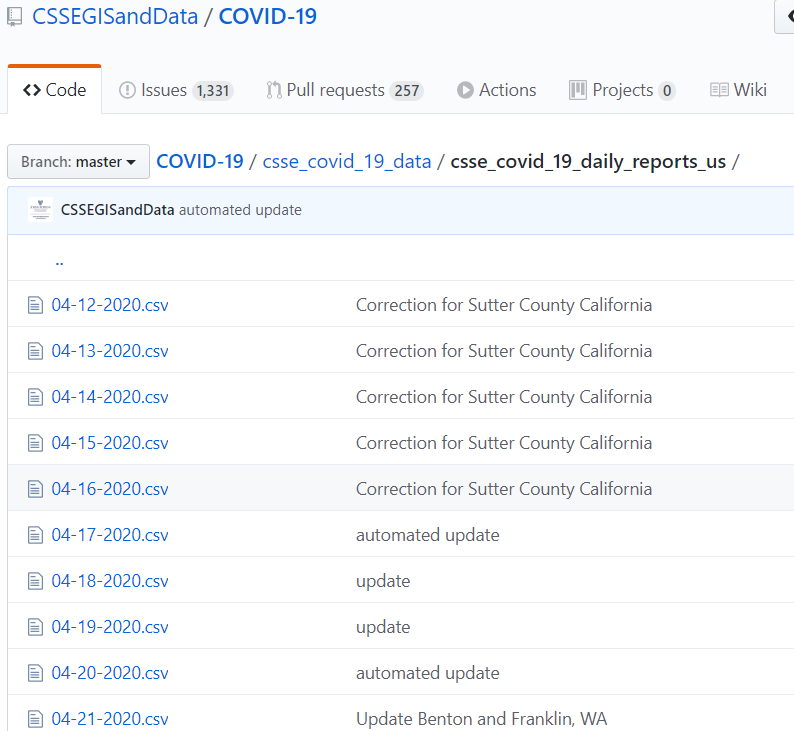
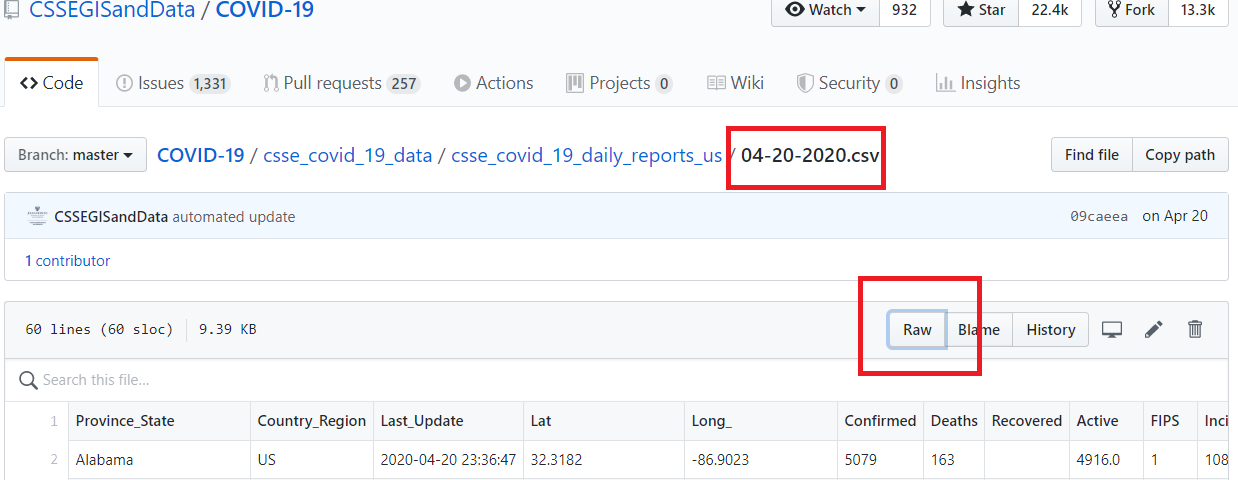
# **Extract:**

1. We have extracted the John Hopkins data from <https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker>, where an embedded [link](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data) on its overview page have led us to find the daily CSV files on git hub.





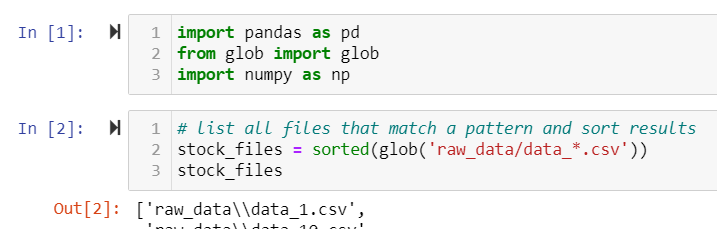
1. We downloaded **30 CSV** files one by one by doing the following steps:
   1. Go to each daily dataset for April 20 to May 20.
   2. Place the cursor to the "Raw" button on the top right side of the dataset.
   3. Press "Alt" and then left click the "Raw" button.
   4. The whole CSV downloaded in the system automatically
   5. Saved to the designated folder where the jupyter notebook is also saved.



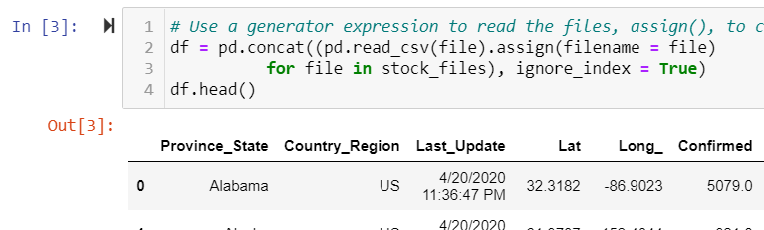
# **Transform:**

The type of transformation needed for this data are executed in the following steps:

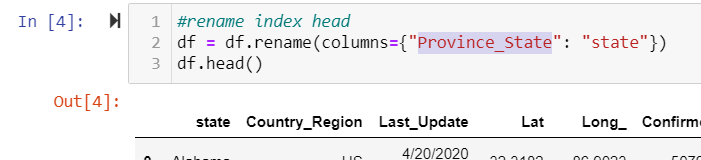
1. Dependencies for John Hopkins data are pandas, glob, and numpy modules on jupyter notebook. The we listed all files using glob from the CSV file local folder.



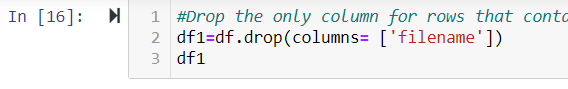
1. Used generator expression to read the files, assign() to create a new column, and concatenate Dataframes

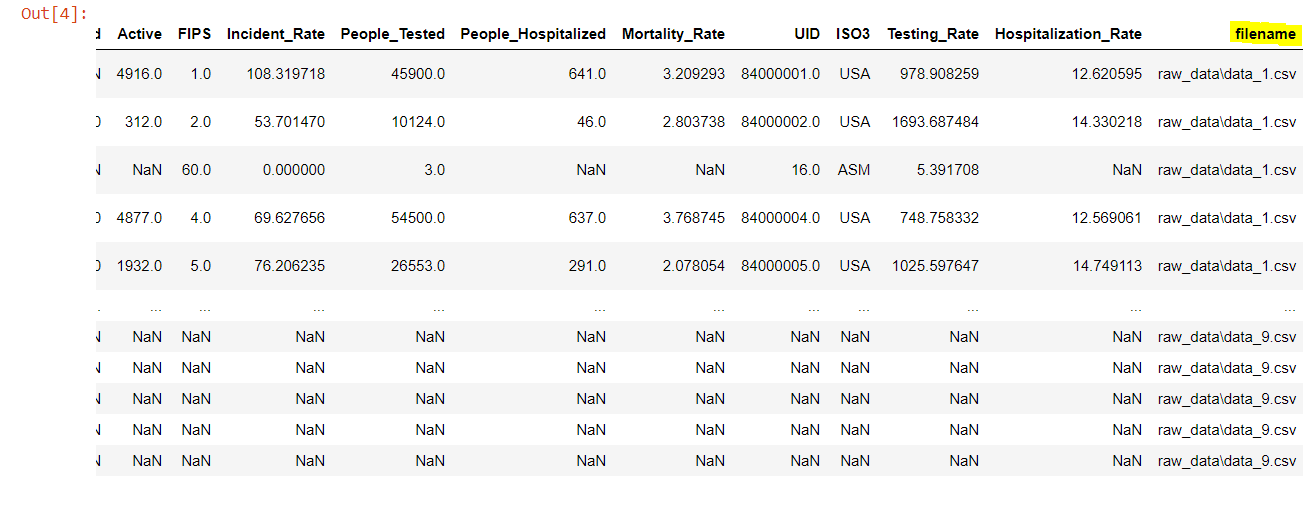


1. Renamed the index column from ‘Province\_State’ to ‘state’ to align index data with my teammates file

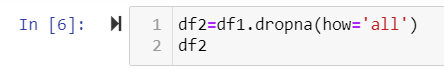


1. Dropped the column for rows that contains another data type besides Nan

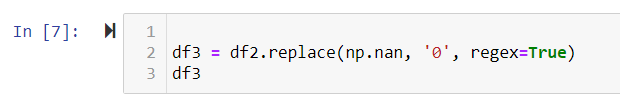




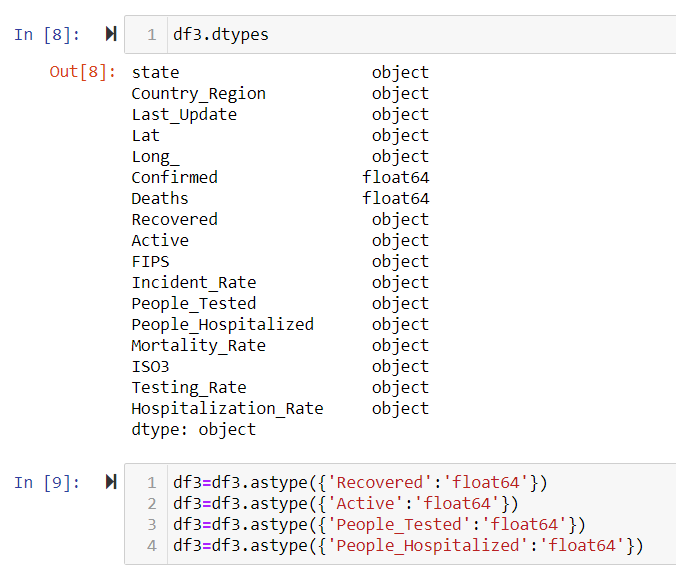
1. Dropped rows that does not contain data



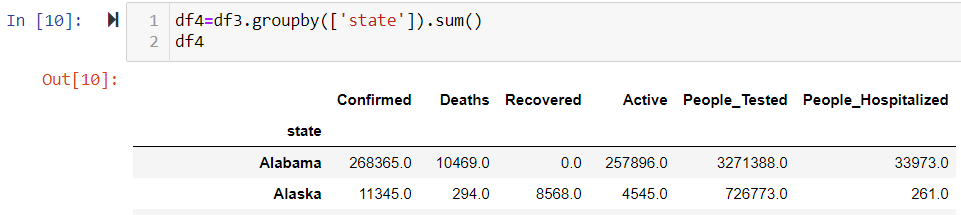
1. Replaced all “NaN” cells with zeros



1. Changed the data type of columns ,that contain numbers, from ‘object ‘to ‘float64‘ to be able to perform numeric calculations.



1. Groupped the data by ‘state’ by summing up the values.

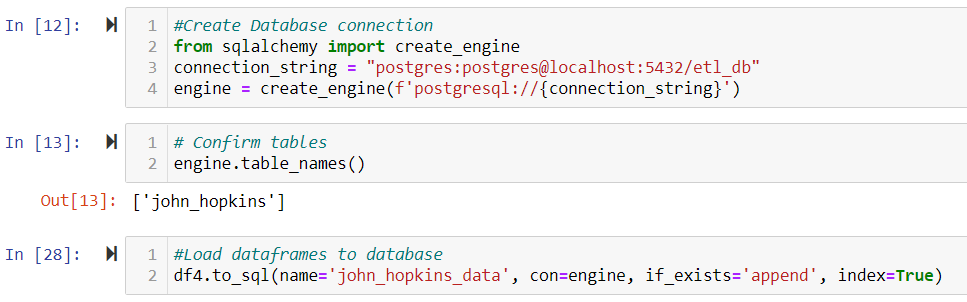


1. Saved the dataframe into CSV file as john\_hopkins.csv

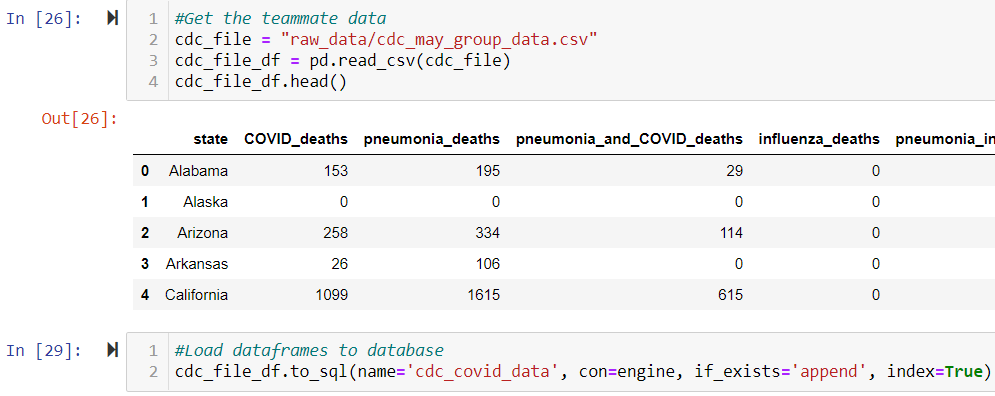


# **Merge:**

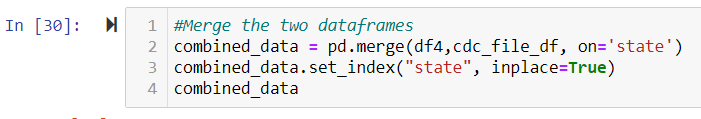
1. Create database connection and load the dataframes



1. Get the teammate’s CSV file, convert to dataframe and load to my database



1. Merge the two dataframes



# **Load final tables to database:**

1. Save the merged dataframes as CSV and load to database to be the final table for production.



## This type of data was chosen due to the uprising numbers of COVID-19 deaths and infections. My partner and I are mothers who are concerned and looking to learn more on finding the right sources that can provide correct and organized data.