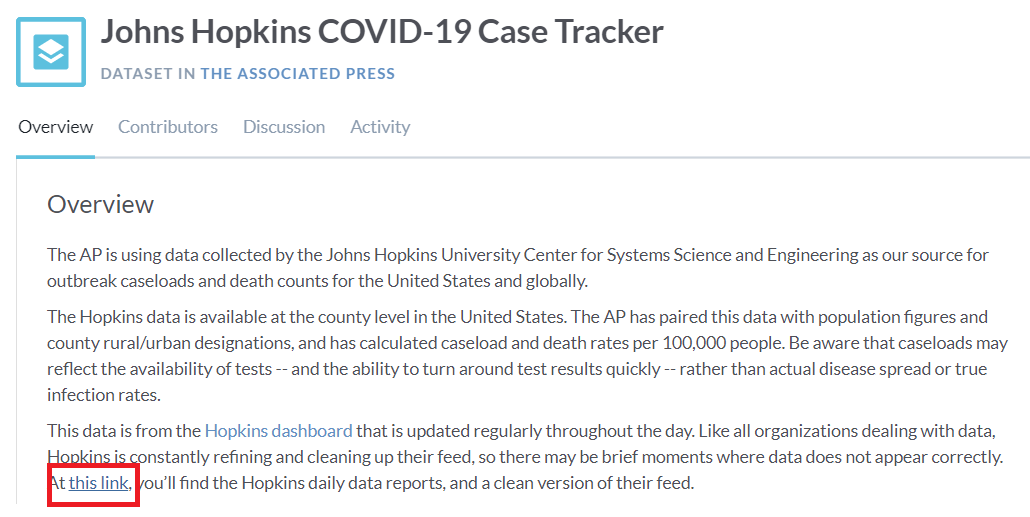
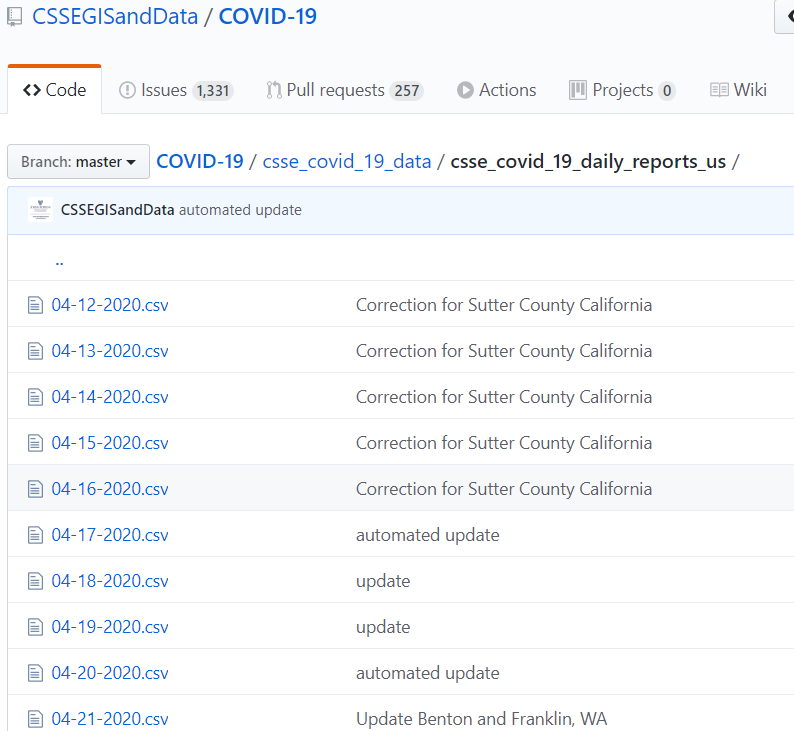
# **\* \*\*E\*\*xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc)**

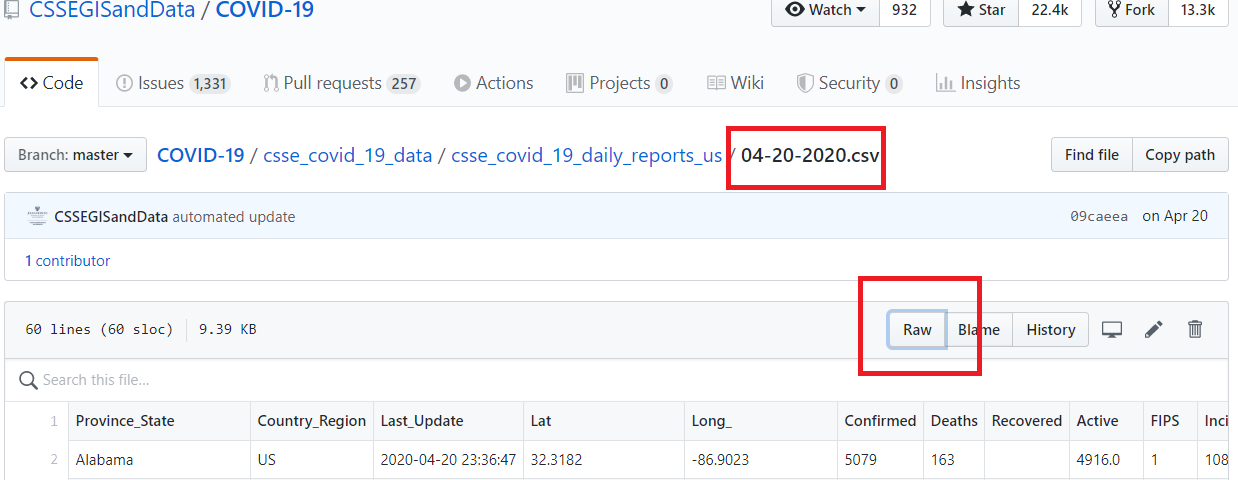
We have extracted the data from the following sources:

1. <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.





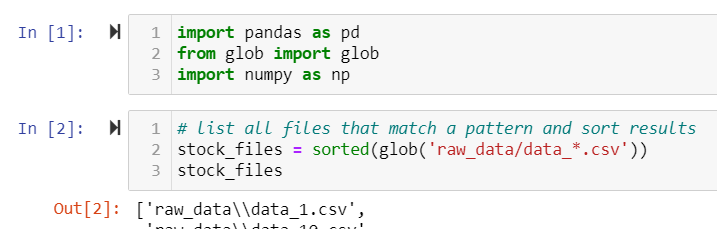
2. Download the CSV files one by one: Go to each daily dataset for April 20 to May 20. You will see "Raw" button on the top right side of the dataset. Press "Alt" and then left click the "Raw" button. The whole CSV will download in your system and save to the designated folder where the jupyter notebook is also saved.



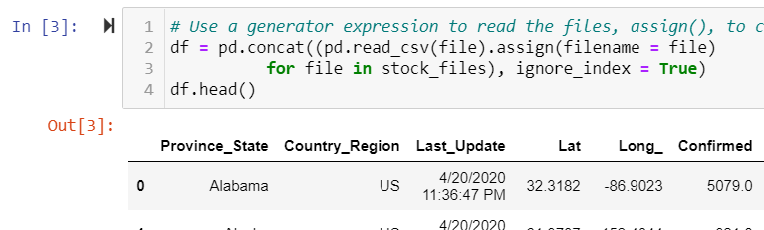
# **\* \*\*T\*\*ransform: what data cleaning or transformation was required.**

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

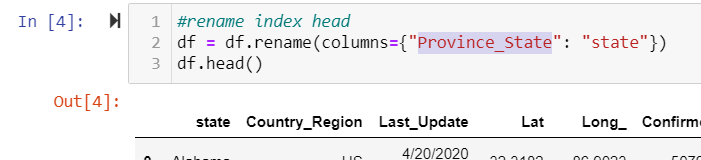
1. Use pandas, glob, and numpy modules on jupyter notebook. Pull the files using glob from the CSV file local folder location.



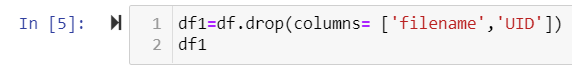
1. Use a generator expression to read the files, assign(), to create a new column, and concat Dataframes



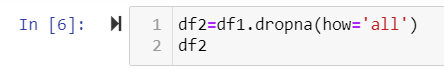
1. Rename the index column from ‘Province\_State’ to ‘state’ to align index data with my teammates file



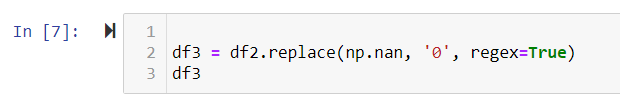
1. Drop the columns that are not needed for analysis (e.g. 'filename','UID' )



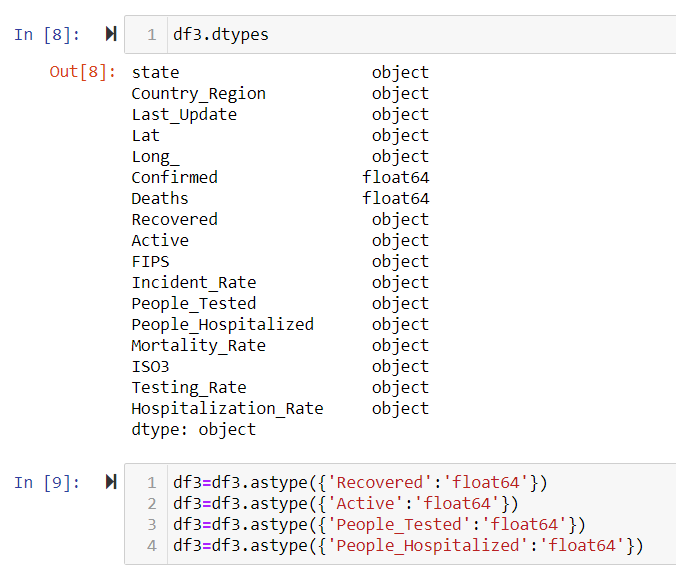
1. Drop rows that does not contain data



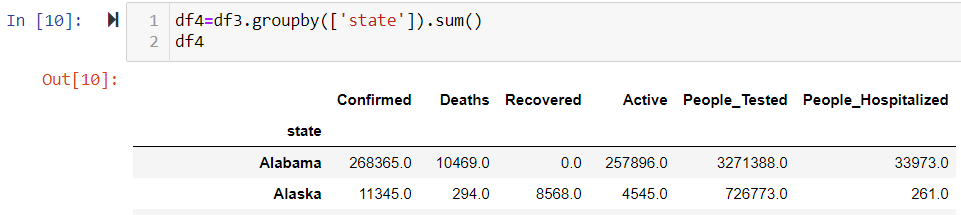
1. Replace all “NaN” cells with zeros



1. Change the data type of columns that contains numbers from object to float64 to be able to perform numeric calculations.



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



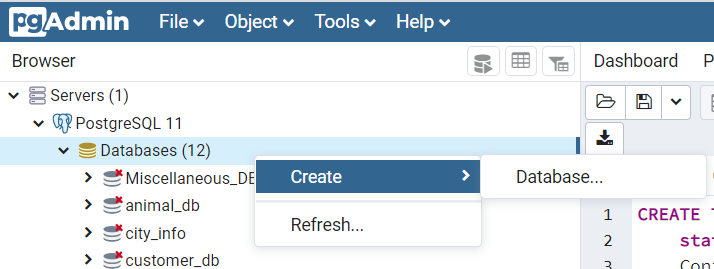
1. Save the dataframe into CSV file as john\_hopkins.csv

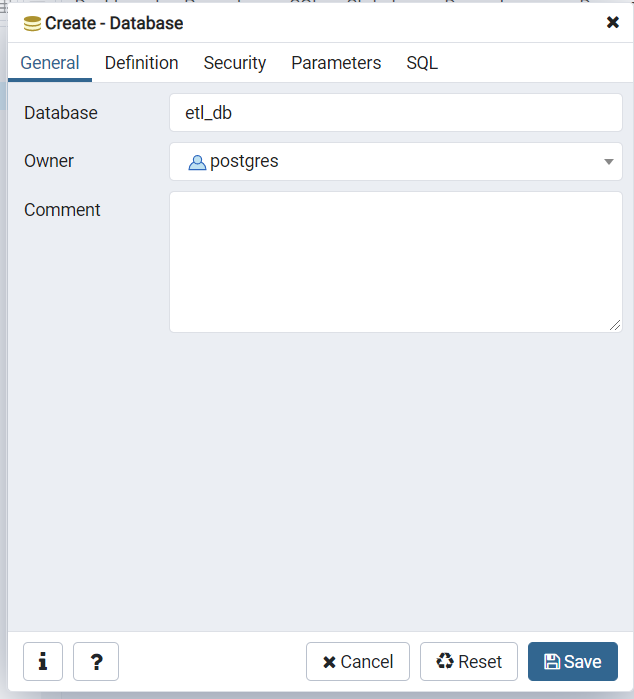


1. Merge file to teammates file

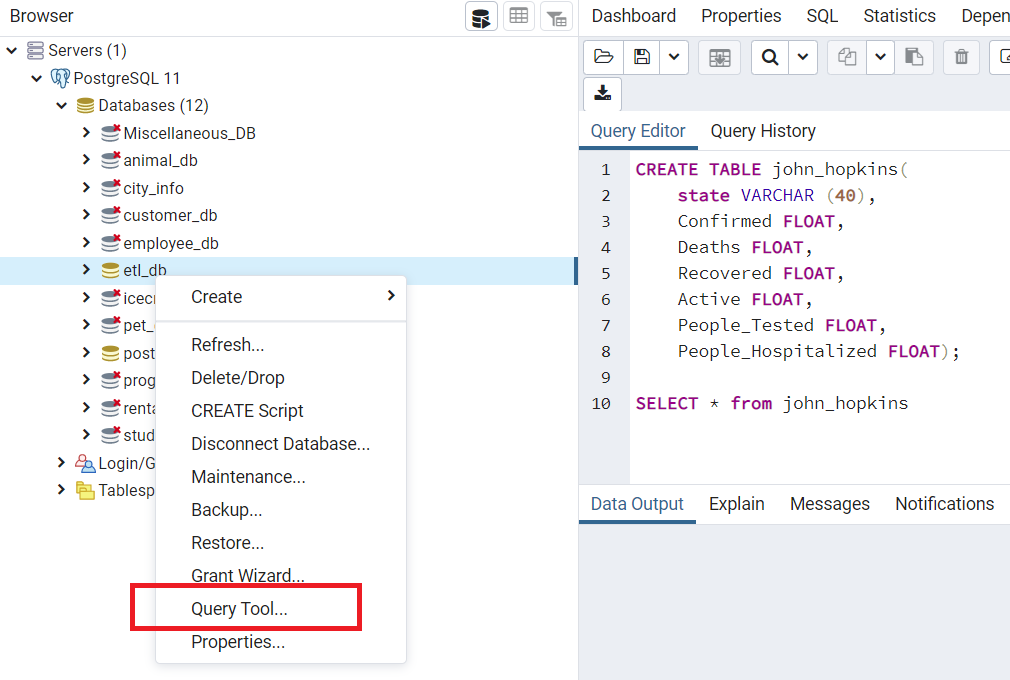
# **\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.**

1. Create etl\_db in postgres

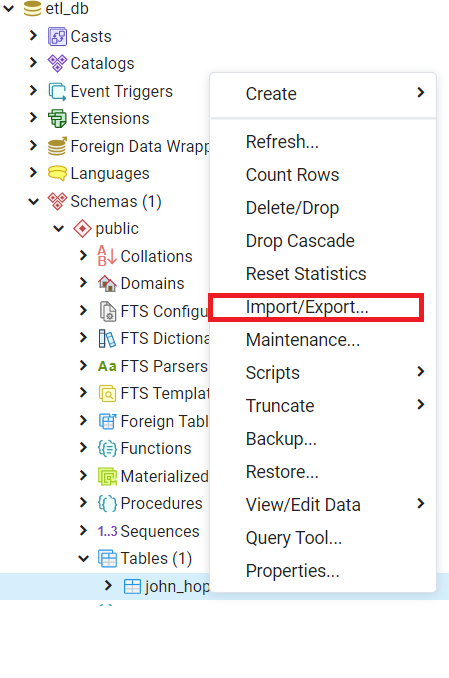
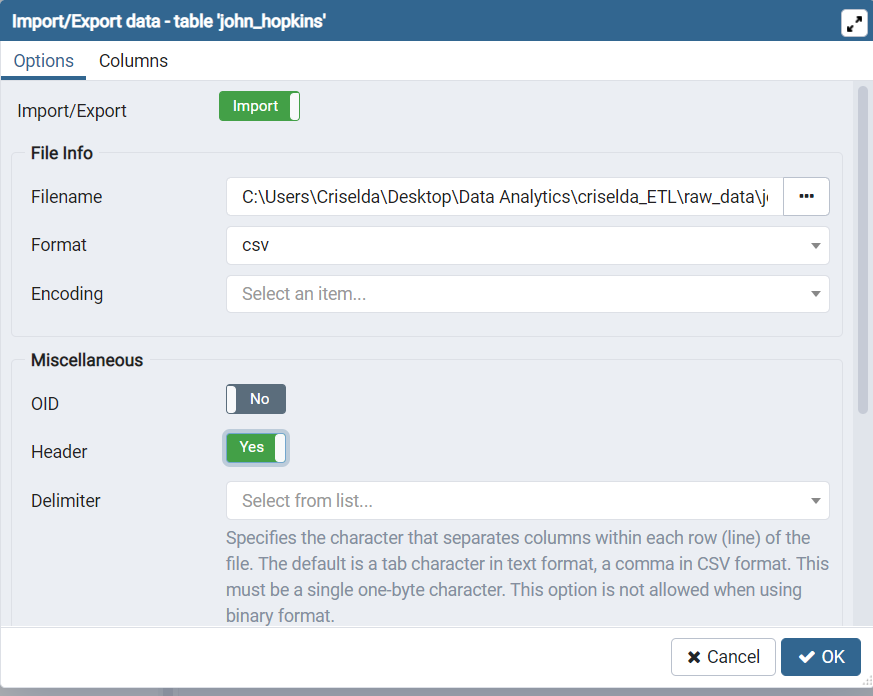




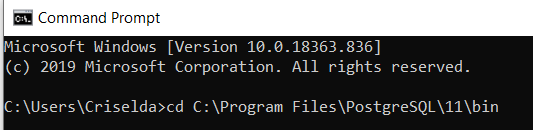
1. Create table in etl\_db named as john\_hopkins with columns and datatypes aligned with the john\_hopkins.csv file.



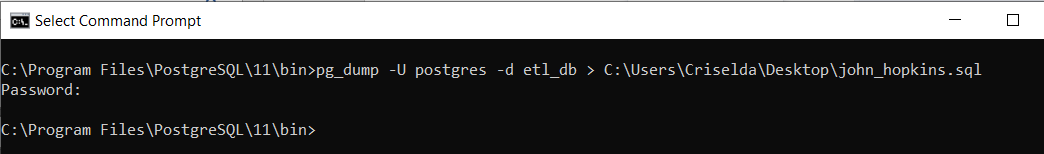
1. Import john\_hopkins.csv file into the table

1. Given that this is a team project, we have to Backup our table using Postgres by performing the following:
   1. Open Command Prompt
   2. Enter the folder location of my Postgres bin:



* 1. Perform the pg\_dump by providing username (-U) as postgres, database name (-d) as etl\_db and send to (>) location where I want it saved and my designated filename and format. Afterwards, I enter my Postgres password to execute the Backup file.



The final tables or collections that will be used in the production database.