## ETL Covid

## By Peter Pages

## Extraction:

Data was loaded into a Jupyter Notebook for analysis and transformation.

**Data Sources:**

**COVID 19 -** Number of Confirmed, Death and Recovered cases every day across the globe

<https://www.kaggle.com/imdevskp/corona-virus-report>

**Format:** CSV

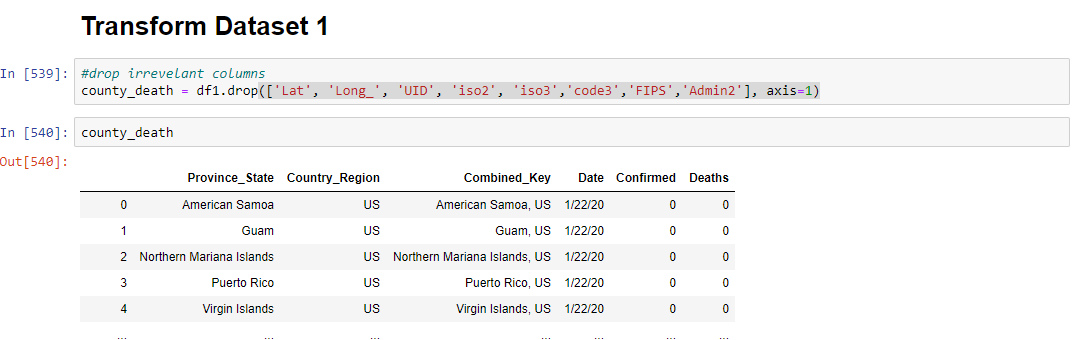
**Data Source: Diversity Index of US counties**

[**https://www.kaggle.com/mikejohnsonjr/us-counties-diversity-index**](https://www.kaggle.com/mikejohnsonjr/us-counties-diversity-index)

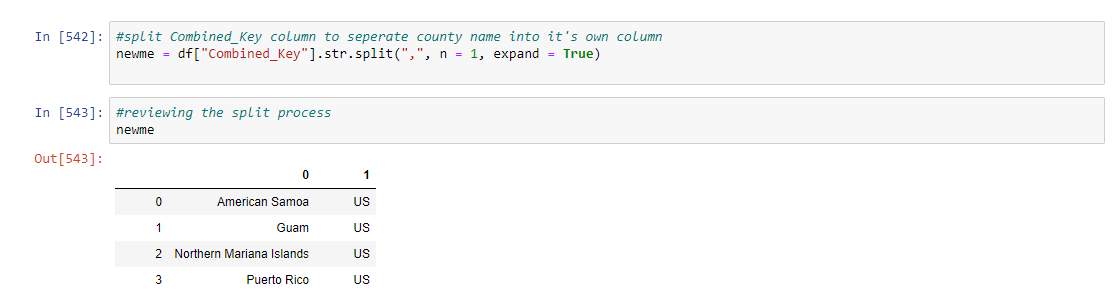
**Format:** XLS

## Transform:

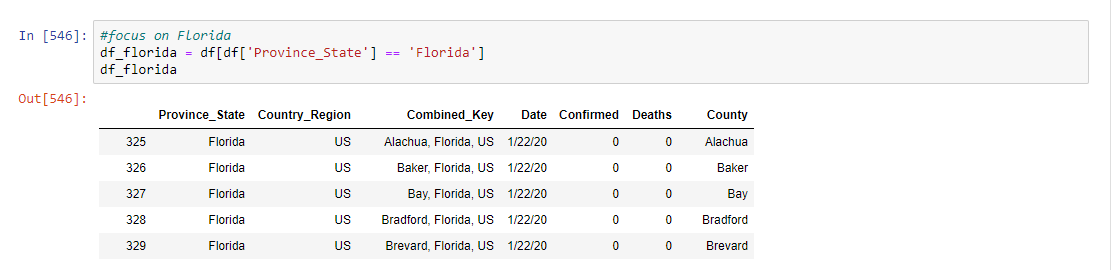
Cleaning the datasets by first dropping variables deemed irrelevant.



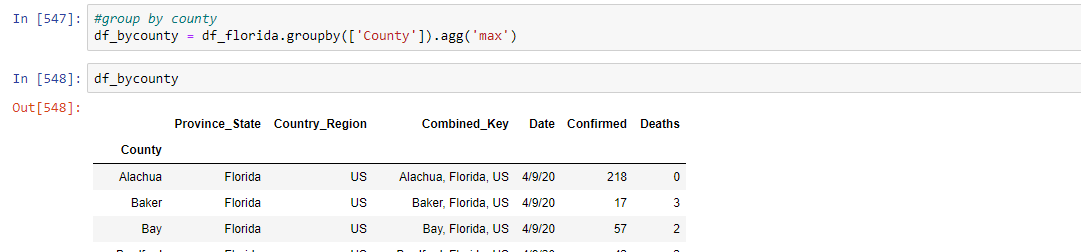
Split **Combined\_Key** column to separate the **county** from city/state. Items are comma separated so a a pandas split method was used.



Focus on Florida. This was accomplished by filtering and reassigning to a variable.

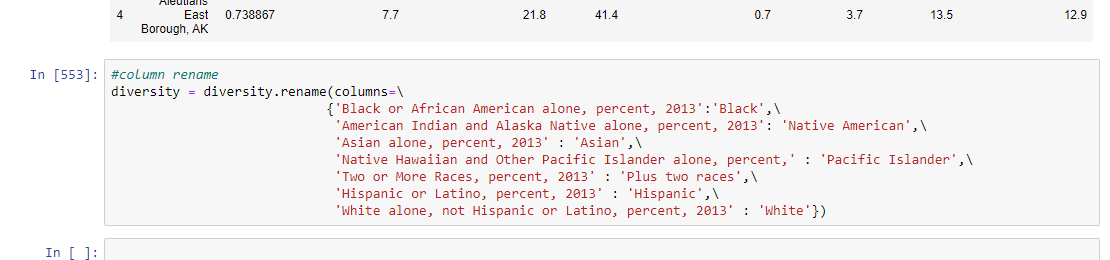


Group By County and using the aggregate function and (‘max’) calculation specified. The totals are running totals so max is appropriate.

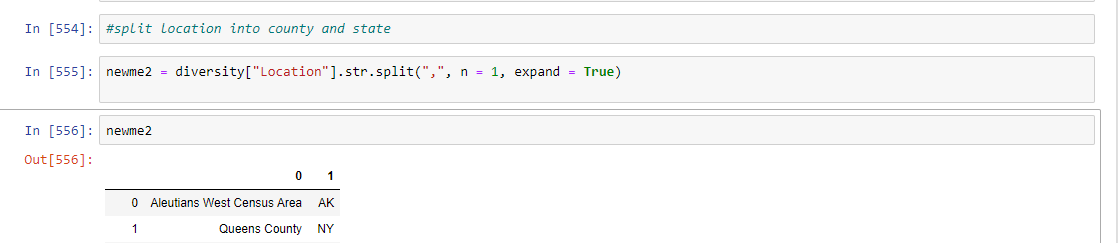


DataSet 2

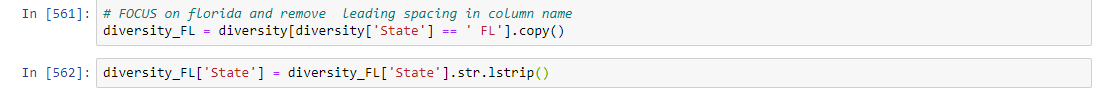
Rename Columns as they’re too long



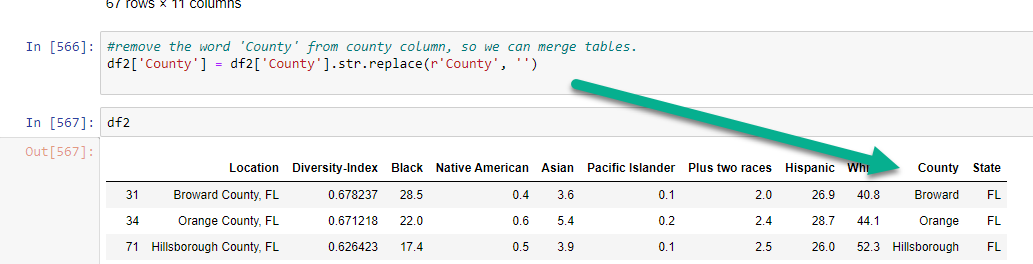
Split Location into County & State with split method.



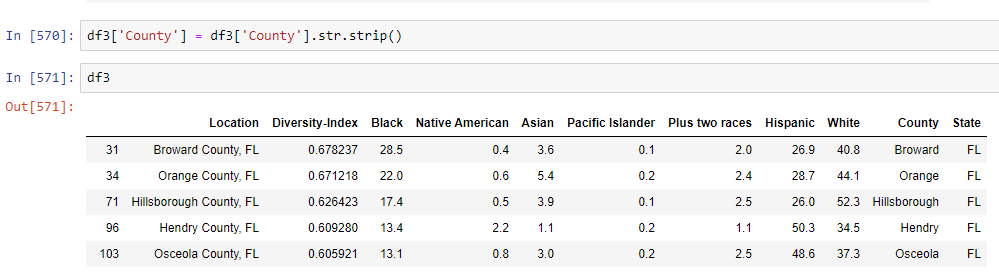
Removing Leading Space in State column



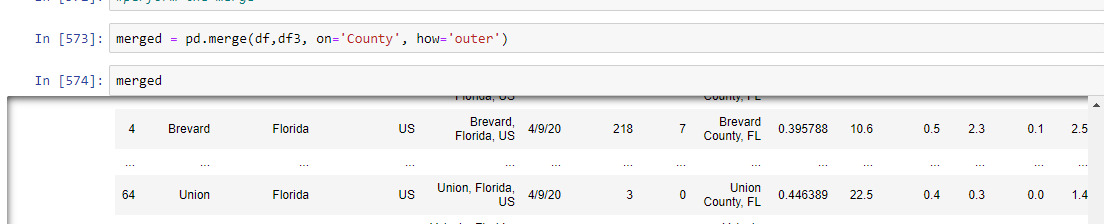
Remove word “**county** “from County column. It’s redundant and prevents a merge.



Clean any white space in the ‘County’ column



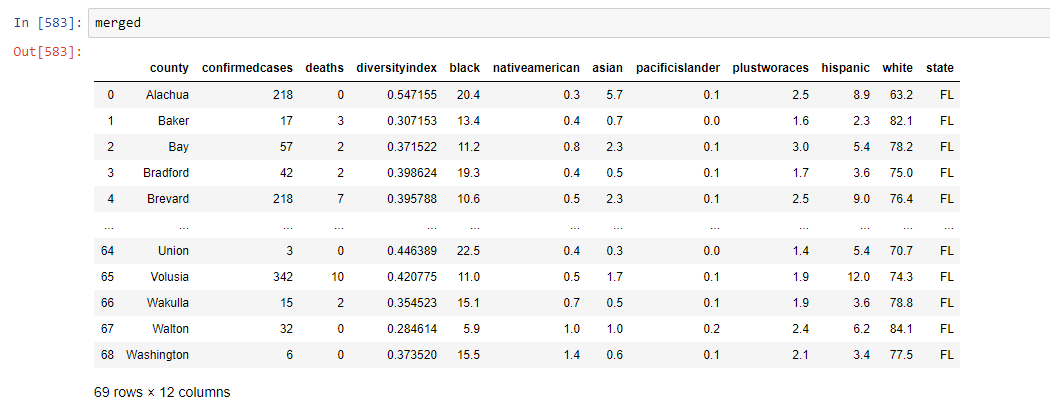
Merge



Rename some more columns and drop unnecessary columns in order to prep for loading.



Final form ready to load



## Load

The final database was created in Postgres called countycovid.

## Limitations:

The diversity metric dataset is from 2013 and applies to the general population. It is based on a formula called the Simpson Diversity Index: D = 1 - ∑(n/N)^2 (where n = number of people of a given race and N is the total number of people of all races, to get the probability of randomly selecting two people and getting two people of different races (ecological entropy).