Extract:

We wanted to look at COVID-19 case and vaccination data by Arizona counties. We pulled data from four different sources: [Arizona Department of Health](https://www.azdhs.gov/preparedness/epidemiology-disease-control/infectious-disease-epidemiology/covid-19/dashboards/index.php), [United States Department of Agriculture](https://data.ers.usda.gov/reports.aspx?ID=17827), [Arizona Department of Health Services](https://www.azfamily.com/news/continuing_coverage/coronavirus_coverage/vaccine_headquarters/adhs-adds-map-showing-number-of-covid-19-vaccines-administered-in-each-zip-code/article_af4c16da-79ed-11eb-b3b7-73d01df565d5.html), and [United States Zip Code database](https://www.unitedstateszipcodes.org/zip-code-database/). The data all came from separate excel files that had to be converted into CSV. The most recent population data was from 2019 and the vaccination and case counts were updated on March 6. The data types were all text, so we later had to make them into integers.

Transform

We first renamed the columns so we could later merge the tables together. We later found out that we named the columns in a format that Postgres couldn’t read, so we had to go back and rename the columns again. We had tribal data, but that data was suppressed, so we dropped those columns. We also dropped rows that had no information in them. Some of the cases and vaccination data had ranges of 1-10, so we had to go in and replaced the ranges with a value of 5. We also changed the types of data into integers for later use.

Load

We loaded our four different tables into Postgres using table names we could easily identify. Used a config file to avoid calling out password/other machine specific information. We had to later go back to Jupyter Notebook and continue the cleaning process, so we changed “append” to “replace” our tables.

Other Considerations

After we loaded the tables, we decided to join them together on county, population, total cases, and total vaccinated. Building off of that table, we added the percentage of cases and people vaccinated based on population and grouped by county.

We found most of our issues after we tried to use test our different data sets and join them together. That is when we went back to Jupyter Notebook (after trying it in Postgres) to change column names, data types, and removed commas and range values.