**Project Summary:**

In our quest to build our database, we ambitiously aggregated data underpinning covid-19 at both the county and state level. We collected data regarding covid-19 specifically (cases/deaths by state/county and cases/deaths by ethnicity), amongst several seemingly non-related county/state level attributes (voter breakdown, unemployment rates, median household incomes, land area, internet access etc.)

**Extraction Process**

In order to gather the data we were searching for, we had to use six different data sources and the data was acquired in several different ways. In all, we used beautifulsoup, splinter, API calls, and simple CSV files to gather our data.

* Data Sources
  + USA Facts (<https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>
    - This data was extracted via BeautifulSoup web scraping
  + United States Census Bureau (<https://www.census.gov/geographies/reference-files/2019/demo/popest/2019-fips.html>)
    - This data was extracted via csv download
  + The Covid Tracking Project (<https://covidtracking.com/race>)
    - This data was extracted via csv download
  + FIPS Codes
  + World Population Review (<https://worldpopulationreview.com/states/state-abbreviations>)
    - This data was extracted via JSON through API calls
  + USDA.gov (<https://data.ers.usda.gov/reports.aspx?ID=17828>)
    - This data was extracted via Splinter web scraping
  + MIT Election Data (<https://electionlab.mit.edu/data>)
    - This data was extracted via csv download

**Transform Process**

This part of the project is probably where the majority of our time was spent. In order to get our data in the correct formats we had to: separate larger datasets into multiple tables (i.e state table and county table), rename data, reorganize/pivot data, merge data, drop nulls, drop duplicates, strip unnecessary strings etc. We based the majority of our transformation work on our anticipated ERD sketch of what our database should/would look like. Once we completed this process and could map it to succinct tables generated in our ERD diagram, we were ready for the Load process.

**Load Process**

Once the data was wrangled and organized – the load process was likely the simplest portion of the ETL project. We simply extracted the PostgreSQL script from our ERD generated in QuickDBD and ran the query in PostgreSQL to generate our database tables. Once these were in place – we simply loaded the tables we created in the transform process. This process ran mostly without hiccups but there were instances where slightly more transformations were required as the datasets were not perfect (i.e we found rows with mislabeled FIPS codes in our original dataset that we were unaware of until our foreign keys alerted us to these issues). Once the hiccups were cleared – the data loaded in fairly easily.