**COVID-19 in the United States: Infection Rates and Election Results**

Caitlin Berryman

Scope

The purpose of this ETL project was to put together a database for an analyst to look at some of the influencing factors of the 2020 United States Presidential Election. Held November 3rd 2020 as the country was ravaged by the COVID-19 pandemic, this looks at the potential relationship between a state’s voting tendencies and the infection rates of their population.

The database is comprised of tables containing three sets of information: population, election statistics, and infection rates.

Population

Population data was taken from Wikipedia, “List of states and territories of the United States” (<https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States>).

The data was scraped using Python library Beautiful Soup, where the information was extracted in HTML format and subsequently transformed into a Pandas DataFrame.

As the webpage is very heavily formatted, the information had to be filtered to search initially for the correct table, and then for the specific elements within. Once these had been located, the script looped to append each state and its population number into lists; state abbreviations had “\n” stripped from their entries, and populations were converted from string to integer, removing commas in the process. These cleaned lists were then combined into a DataFrame.

Covid

COVID-19 data was taken from the Kaggle dataset “COVID-19 in USA” by SRK, the specific file “us\_states\_covid19\_daily.csv” (<https://www.kaggle.com/sudalairajkumar/covid19-in-usa>).

The csv was loaded into a DataFrame for transformation. The file contains many metrics about each state’s cumulative test results, positive cases, hospitalisations, and deaths, amongst others. The final table intended to take a snapshot of the 50 states on Election Day, and was mainly concerned with the numbers for positive cases and deaths.

In addition to the 50 states, this dataset contains information on the federal District of Colombia and five inhabited territories, which was filtered out to remain consistent with the contents of the other tables.

Election

Election data was taken from the Kaggle set “US Elections Dataset” by Bojan Tunguz, file “1976-2020-president.csv” (<https://www.kaggle.com/tunguz/us-elections-dataset>).

Similar to the COVID data, the csv was loaded into a DataFrame for filtering. This dataset contains voting results by state for each election from 1976 onwards; the first step was to filter for only 2020. From there, only the candidate that won the state was to be included – the shorthand method to achieve this involved sorting the “candidatevotes” column by descending values, and then dropping duplicates by their state abbreviation. The table was then re-sorted into alphabetical order by state.

Voting information for DC was dropped to maintain consistency. The columns were trimmed to keep state abbreviation (renamed from “state\_po” to “state” to match the other tables), winning candidate name and party, their number of votes, and the total votes of each state.

Loading

The tables were loaded into a SQL database, interfaced by pgAdmin4. The schema has been included in this repository for reference. The index for each table was set to the two-letter state abbreviations, as they are unique and common to all three sets of information.