COVID Vaccination Data ETL

# description

In my first Bootcamp project – “Anxiety and Depression in the time of COVID-19” we looked at how COVID case counts, and indicators of anxiety and depression progressed over the course of the pandemic, and the relationship between the two. As a follow on, we wanted to look at how vaccinations changed the picture once the vaccine rollout started. In this ETL project I will collect COVID vaccination data at the state and national level, clean it up, re-organize it and load it into a PostgrSQL database. If time permits I will do the same for COVID case counts and possibly death rates, as this was never done in a repeatable manner for our original project (we just basically hacked the Excel file).

# data sources

The key datasets that will be used are the following:

1. **COVID Vaccination Data**: Unfortunately, I am not able to find consolidated vaccine data on the CDC websites that would give me everything that I want – I can get cumulative totals by state, and I can get daily counts at the national level, but I want daily vaccination counts by State. So I will us the “Our World in Data'” site <https://ourworldindata.org/us-states-vaccinations> and download several different files to merge them:
   * us-covid-number-fully-vaccinated-in-US.csv (by state)
   * us-covid-share-fully-vaccinated.csv (by state)
   * us-daily-covid-vaccine-doses-administered-by-state.csv
   * us-daily-covid-vaccine-doses-per-million.csv (by state)
2. **COVID Case Counts**: CDC US COVID-19 Cases and Deaths by State over Time
   * <https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36>
3. **COVID Deaths**:
   * <https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36>

# rough task breakdown

Key tasks that will need to be performed are outlined below.

## Data Exploration & extract

1. Find all of the files that I need
2. Download and extract to Pandas dataframes in a Jupyter notebook
3. Remove empty columns
4. Merge into a single file

## Data Cleanup & transformation

1. Clean up the merged Vaccination data
   * Remove duplicate date (Federal Agencies)
   * Replace remaining NaN values with zeros
   * Change the datatypes of some of the fields (remove unnecessary decimal places)
   * Change column names
   * Split out the data in a State level df and a National df
2. Clean up the COVID Case data
   * Transform state codes into state names so that I can join with the vaccination data
   * Merge the NYC and NYS data (it is currently split out in the original dataset)
3. Clean up the COVID deaths (BONUS)
   * Transform state codes into state names so that I can join with the vaccination data
   * Merge the NYC and NYS data (it is currently split out in the original dataset)

## data engineering

1. Design PostgreSQL DB schema
2. Create COVID database and write SQL to create the tables

## data loading

1. Write the Python, Pandas & SQL-Alchemy code to populate the database with the data that was cleaned up in the first step.