Data Preparation Methods

The purpose of this document is to outline the steps that were taken to create the cic_challenge_data_sets from the raw downloaded reports.

KFF

1. Created a new row within the economic_measures raw download excel sheets and calculated the mean unemployment rate for every state and the given year.
2. All variables were copied and special pasted (transpose) onto the cic_challenge_data_set.
3. Empty 'year' rows were filled with their respective years.
4. Dummy variable was created for the 50 contiguous states.
5. Renamed columns to an abbreviated Python friendly format.
6. Set all percentage columns to number and special paste multiplied by 100 to get numbers with integer parts >1.
7. Used Excel's Find and Replace feature to replace N/A entries with blanks.

CDC

1. Filtered and removed US territories from the data set. (Federated States of Micronesia (FM), Guam (GU), Marshall Islands (MH), Northern Mariana Islands (MP), Puerto Rico (PR), and Palau (PW). I also removed the UNK rows.
2. Recoded MMWR weeks to their respective year.
3. Summed the annual total for each state for a given year.
4. Recoded State abbreviations to their full name.
5. Added Hawaii (they weren't able to report because of system differences) and left Cov_Vacc_one cell as null (blank).

Google Dataset Publishing Language –

1. Removed Puerto Rico (PR) from the data set.
2. Made a copy of the cic_challenge_data_set and pasted data for respective states and years. (This was more expedient and saved processing time as there is only ~150 rows and I wouldn't have to filter them out of every single query).