1. **Pre-Work**
   1. Search for datasets
      * Understand the datasets
      * Read column description
      * Download datasets from source (Kaggle)
   2. Create database in Postgres
      * Create two tables with the following columns
        + ncov19\_data
          - country (set as primary key)
          - confirmed\_ncov
          - deaths\_ncov
          - recovered\_ncov
        + sars03\_data
          - Country (set as primary key)
          - confirmed\_sars
          - deaths\_sars
          - recovered\_sars
2. **Extract**
   1. Retrieve CSV data and pass into Pandas DataFrame
      * Check if datasets are correctly loaded
   2. Get standardized country list from pycountry module
3. **Transform**
   1. Create function to identify invalid country names within the datasets/sources
   2. Rename invalid country names to standardized country names per the list from pycountry module
   3. Check dataframe:
      * to make sure invalid country names are correctly updated
      * to make sure that province/states are assigned to the appropriate/same country
        + Hong Kong is assigned to both China and Hong Kong across different cases. For cases assigned to China, rename the country of those cases to ‘Hong Kong’
        + Taiwan and Macao are both assigned to China and their respective names as well. Rename the country of those cases to ‘Taiwan, Province of China’ and ‘Macao’
   4. Aggregate the dataframes by Country and Date (get the sum by Country and Date)
      * Filter for the latest date in the data-set
4. **Load**
   1. Create connection to SQL database
   2. Load dataframe into SQL tables
   3. Used a “**FULL OUTER JOIN**” to merge the two tables on Country:
      * Additional “**WHEN**” commands were needed to ensure the common columns weren’t showing up twice
      * Used “**COALESCE**” commands to replace null values with a zero