ETL Project

Members: Saibal, Benjamin

ETL (extract, transform, load)

**Proposal**

**Our proposal was to download, clean, and process COVID data from two sources to compare US statewide vaccination data and CDC COVID confirmed, probable cases, and deaths.**

* Source 1: https://ourworldindata.org/explorers/coronavirus-data-explorer?zoomToSelection=true&time=2020-03-01..latest&pickerSort=asc&pickerMetric=location&Metric=Confirmed+cases&Interval=7-day+rolling+average&Relative+to+Population=true&Align+outbreaks=false&country=USA~GBR~CAN~DEU~ITA~IND
* Source 2: <https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data>

**Purpose**

The purpose of this was to analyze:

1. Correlation between the total vaccination and total deaths (confirmed and speculated)
2. Correlation between the total vaccination and total cases (confirmed and speculated)
3. Query between the states and total cases (confirmed and speculated)
4. Query between the states and total deaths (confirmed and speculated)
5. Plot bar chart to show vaccination vs. total deaths by state
6. Plot bar chart to show cases vs. total deaths by state
7. Plot bar chart to show vaccination vs. cases by state
8. Null Hypothesis: The more vaccinated the fewer deaths per state
   1. Run p-value test to confirm/reject null hypothesis

**Process**

**Extract**:

* Download data from sources and save in folder for jupyter notebook to access.
* Create jupyter notebook and import necessary libraries for data manipulation.
* Ingest data into panda dataframes.

**Transform**:

* Filtered by most current date listed in both dataframes (4/26/2021).
* Dropped unnecessary columns.
* Due to one dataframe having a state name and another state code, we downloaded a lookup file for state province and codes and loaded it into a third dataframe.
* Replaced all null values with 0 using fillna() function.

**Load**:

* Used SQLalchemy to create an engine object.
* Used PGadmin to create a single database with three tables with primary key.
* Used engine to connect to database
* Used the to\_sql function to load data from dataframe into database tables.