Mengye Wei
05-25-2020

ETL Project Report

- **Extract**
  1. *US_comfirmed_cases.csv* is the dataset that was obtained from dataworld, which has all the confirmed and death records about COVID-19 until 05-23-2020. This dataset was ordered by counties, has a total population for each county and contains Federal Information Processing Standards(FIPS) code.
     (Reference: https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker/workspace/file?filename=1_county_level_confirmed_cases.csv)

  2. *Hospital_bed.csv* also obtained from dataworld, updated in March 2020. This dataset has lots of different columns, including licensed beds, icu beds, staffed beds records, and licensed beds mean all the beds that are in the hospital. Also, it has a potential increase of beds in the hospital for each county.
     (Reference: https://data.world/qventus/covid-19-localized-scenario-planner/workspace/file?filename=Definitive_Healthcare%253A_USA_Hospital_Beds.csv)

  3. The last file *annual_aqi_by_county_2020.csv* was obtained from the air quality index official website. It describes the average Air Quality Index(AQI) for each county in 2020, the last update is 05-19-2020. This has most of the records for the counties. There is another file that has a daily record for the counties, but not all counties have the same record date, therefore taking the average AQI is enough.
     (Reference: https://aqs.epa.gov/aqsweb/airdata/download_files.html)

- **Transform:**

  hospital_bed.csv has all the information for all the hospitals in the US, hence I need to add all the hospital beds together and potential increase beds for each county. Also, I included the fips code for this dataset to connect with the confirmed and deaths dataset. For the *US_comfirmed_cases.csv* and csv for AQI, I copy the columns from the datasets to make it into a new data frame. Then, I convert the data that I need into a new data frame.

- **Load**

  I use postgre sql to store the three datasets, confirmed and hospital bed datasets can connect with fips. And confirmed and aqi datasets can connect with county name and state name. For confirmed_case, I choose state and county to connect with AQI, fips connect with hospital_beds. Total_population, confirmed and deaths data need to be used in the further project. Included total population can calculate the density and percentage of people who are confirmed. AQI data frame has data for aqi_days records, good_days, pm2.5 and pm10, with all these data we can calculate bad days data and others.