**ETL PROJECT**

**Team Members:** Maryam Torabi Kia & Ahmad Mahroo

**Project Proposal:**

The purpose of this project is to collect data related to COVID-19 cases in the province of Ontario and further match this data to medical centers in the infected patient’s neighbourhoods (first three characters of the postal codes).

**Data Sources:**

1. Provincial COVID-19
   1. <https://covid-19.ontario.ca/>
2. Provincial Medical Centers
   1. <https://data.ontario.ca/dataset/hospital-locations>

**Data Source Type:**

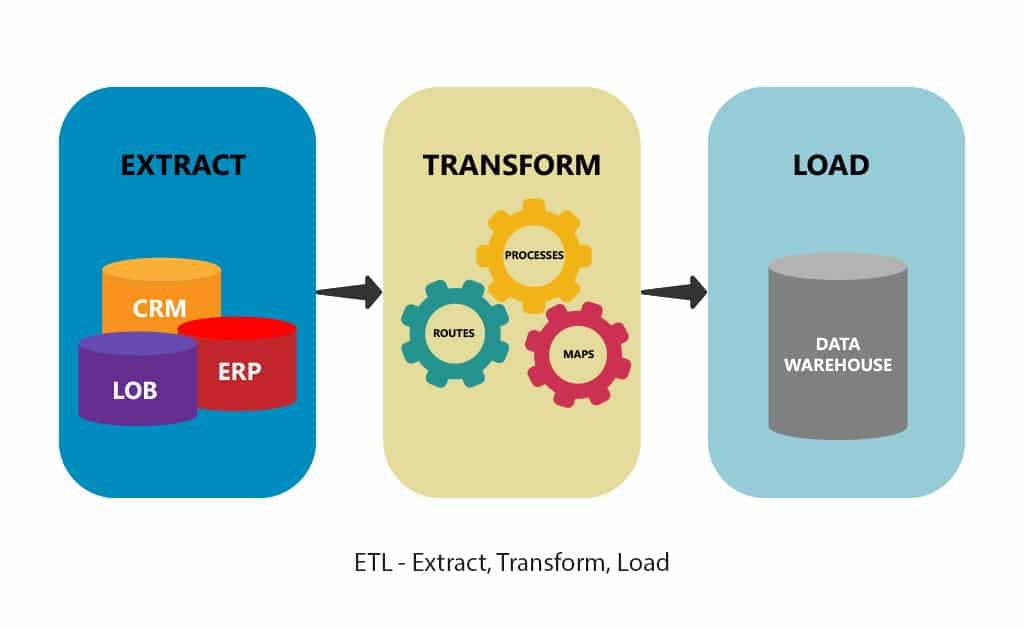
* CSV

**Database:**

* SQL

**Data Transformation strategy:**

1. Break down data by medical center types
2. Multiple tables will be created based on medical center types
3. Narrow the neighbourhoods by using the first three characters of the postal code

****

**Project Report:**

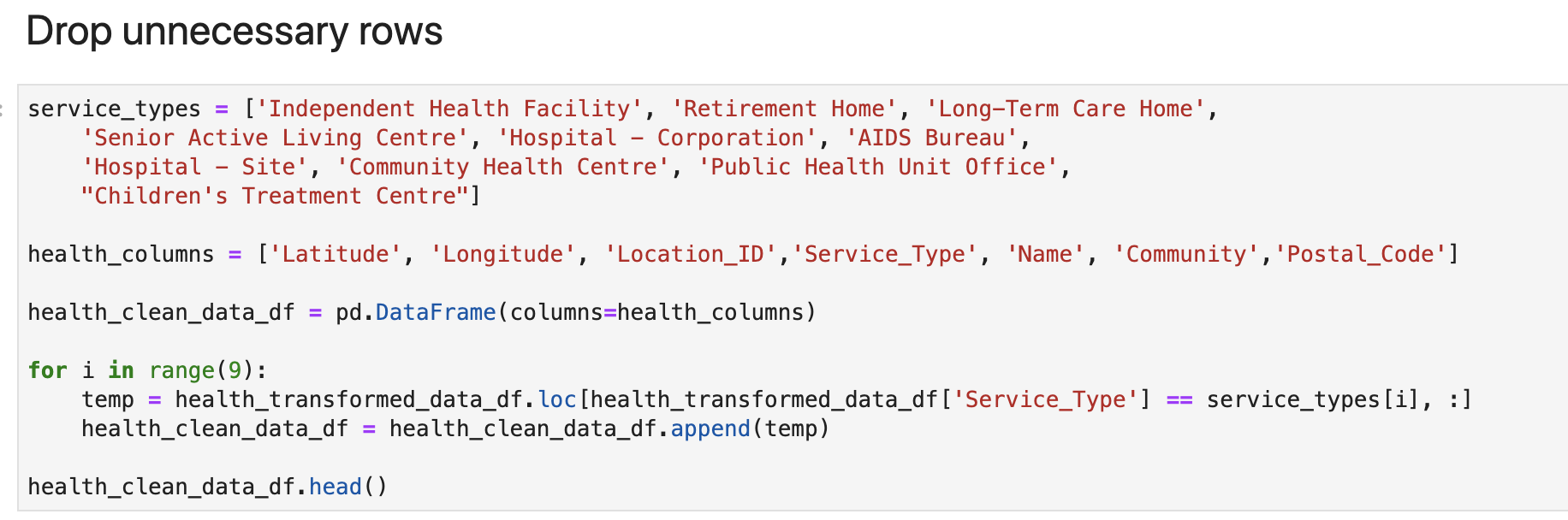
**E: Extraction**

1. Read Province of Ontario Covid-19 Case data into a Pandas data frame
2. Read Ministry of Health Service Providers’ location data into a Pandas data frame

**T: Transformation**

*Data Frame 1: Covid-19*

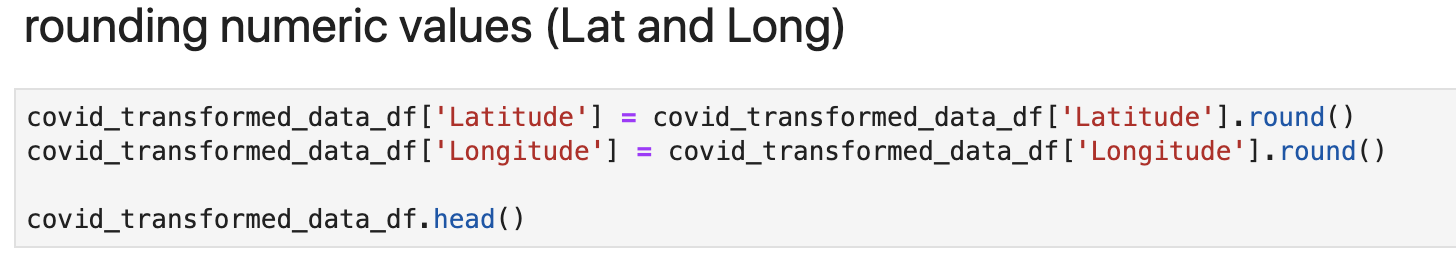
1. Dropped unnecessary columns from Covid-19 dataset. Removed columns include: Public Health Unit street address, Public Health Unit website url, and etc.



1. Renamed columns to more meaningful titles. Sample code shown below:

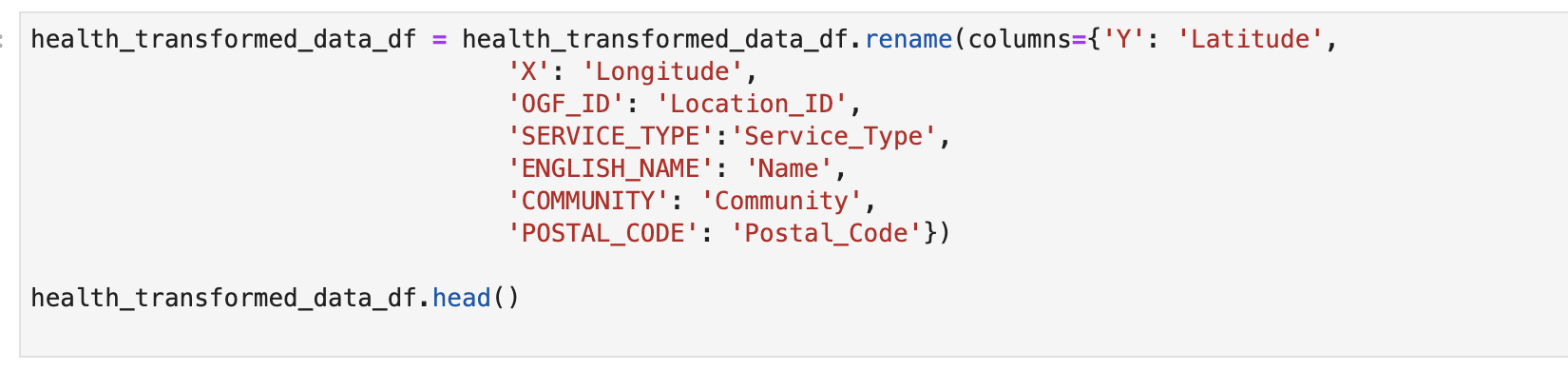


1. Rounded the latitude and longitude data to 2 digits with no decimals in order to achieve a larger coverage area. Code shown below:



*Data Frame 2:* Ministry of Health Service Providers’ location dataset

1. Dropped unnecessary columns. Removed columns include: Effective Date, French Alternate Name, Street Address, and etc.
2. Renamed columns for specific data to make a more clear data frame. An example is renaming Y and X to Latitude and Longitude respectively.

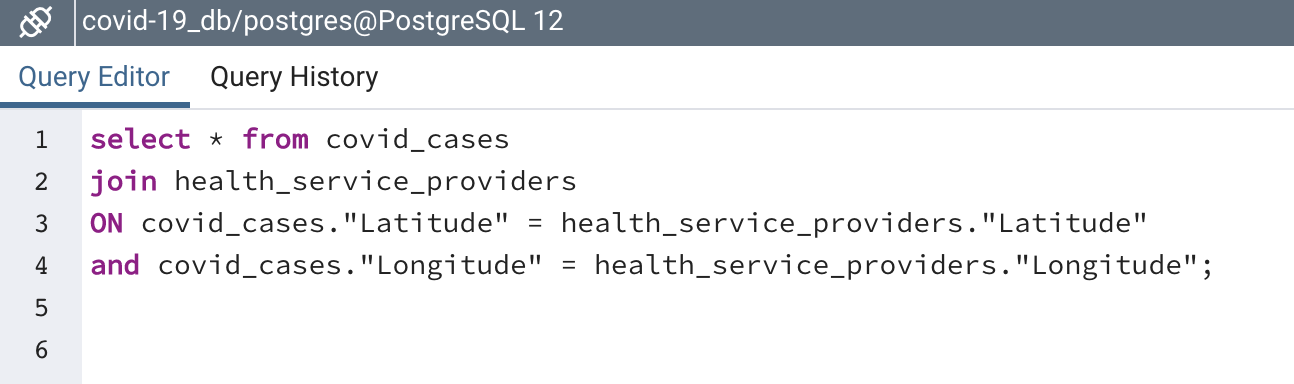


1. Dropped Unnecessary Rows such as service type. Some service types such as pharmacy were removed from the dataset as Pharmacies do not report test results for Covid-19.
2. Rounded the Latitude and Longitude to 2 digits with no decimals.
3. Set the location id as index

**L: Load**

Prior to loading data, a database was created in SQL Postgres. subsequently the data from both data frames were loaded into two separate tables. The 2 tables were then joined using latitude and longitude data.

SQL code:



Output data:

