

Description of the data - How it will be used to solve the problem.

COVID-19 Cases in Toronto (<https://open.toronto.ca/dataset/covid-19-cases-in-toronto/>)

Dataset provided by open data portal of the city of Toronto, Canada, contains weekly-refreshed information about the confirmed and probable cases of patients infected with SARS-CoV-2 virus, which causes COVID-19, since January 2020.

The data features include Unique row identifier, unique assigned ID of cases by Toronto Public Health, Outbreak Associated Type, Age Group, Neighborhood Name, Forward sortation area code, Source of Infection, Confirmed Classification, Episode Date, Reported Date, Client Gender, Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU and Ever Intubated.

Since here we mainly concern with occurrence of case based on location factor, we will extract confirmed cases with neighborhood and other possible elements which might have influence on the cause of disease for further analysis.

Neighborhoods (<https://open.toronto.ca/dataset/neighbourhoods/>)

Dataset provided by open data portal of the city of Toronto, Canada, contains geospatial information about City of Toronto Neighborhoods, especially the boundary information of neighborhoods. The last refreshed date of dataset is on Mar 15, 2021.

The data features include Unique row identifier, Area ID, Area Alter ID, Parent Area ID, Area short code, Area long code, Area name, Area desc, X, Y, Longitude, Latitude, Object ID, Area Shape, Shape Area, Shape Length, Geometry, Classification,

Classification Code.

We need the geometry information which contains the boundaries of neighborhoods and area names contained in this dataset. These data will give us basic geospatial reference for both graph plots and association analysis.

Foursquare Place API (<https://developer.foursquare.com/docs/places-api/>)

Provides access to Foursquare's global database which contains location-based data of venues, including venue trending, venue categories, venue latitude and longitudes and so on. One of the advantages of this database is that we could retrieve location data given specific date which have a significant influence on our analysis. We would like to apply categories defined and given by Foursquare database and assume venues of the same category have the same influence on case occurrence. One of the disadvantages of this API is that we could only get data on a limit of 50 venues with each call of any endpoint. Therefore, here we mainly call for category and geospatial information of 50 trending venues on an assumption that population mobility has a prior influence on virus transmission based on the obvious phenomenon that closure significantly reduce the increase of confirmed case number.