

Domain/Dataset

Our Domain is the City of Toronto's COVID 19 case and Homeless shelter occupancy.

Case Data: <https://open.toronto.ca/dataset/covid-19-cases-in-toronto/>

Homeless shelter occupancy data: <https://open.toronto.ca/dataset/daily-shelter-occupancy/>

For the Case Data, we would want:

- Assigned_ID, which would be our key
- Outbreak Associated, which would help us find the setting the case occurred in
- Age group
- Neighborhood Name
- Postal Code
- Source of infection
- Episode Date (Important to compare with shelter data)
- Outcome
- Ever Hospitalized
- Ever in ICU
- Ever intubated

For the shelter data, we would want:

- _id, Our key
- OCCUPANCY_DATE (to link with Case Data)
- SHELTER_NAME (possibly)
- FACILITY_NAME
- SHELTER_POSTAL_CODE
- OCCUPANCY
- CAPACITY

Understanding how postal codes work geographically may be a struggle in finding larger regions and their data

There are many extra columns we may not need. Also the shelter data has some rows where the occupancy and capacity are 0 so they may need pruning.

Questions

Our 3 questions are:

- What populations (age, region, etc) are being disproportionately affected by COVID 19
- How has (or hasn't) homeless shelter occupancy changed during the pandemic
- Does level of occupancy effect COVID 19 rates either city wide or per shelter

Schema

Shelters(FACILITY_NAME, SHELTER_POSTAL_CODE, CAPACITY)

- A tuple of shelters. The Key is the Facility name, and it is linked with their unchanging information like postal code and max capacity

Neighborhoods(Postal Code, Neighborhood name)

- Neighborhoods are how we are managing location, so collecting only the neighborhoods being used is important
- A tuple of postal codes and their neighborhood name

Cases(Assigned_ID, outbreakAssociated, ageGroup, neighbourhoodName, FSA, sourceOfInfection, classification, episodeDate, clientGender, outcome, everHospitalized, everInICU, everIntubated)

- Assigned_ID is our key per case. This relation is a list of every covid case in Toronto and the relevant details of it. It outlines the general characteristics of the case including source, patient details, and location

Occupancy(_id, OCCUPANCY_DATE, ORGANIZATION_NAME, SHELTER_NAME, SHELTER_POSTAL_CODE, FACILITY_NAME, PROGRAM_NAME, OCCUPANCY, CAPACITY)

- _id is our key for one shelter program per day. Each row includes the shelter information including location, name, and shelter details along with its occupancy and capacity

Occupancy[OCCUPANCY_DATE] \subseteq Cases[episodeDate]

Shelters[SHELTER_POSTAL_CODE] \subseteq Neighborhoods[Postal_Code]

Attribute	Description	Data Type
Assigned_ID	A unique ID assigned to cases by Toronto Public Health for the purposes of posting to Open Data, to allow for tracking of specific cases.	Int
Outbreak Associated	Outbreak associated cases are associated with outbreaks of COVID-19 in Toronto healthcare institutions and healthcare settings (e.g. long-term care homes, retirement homes, hospitals, etc.) and other Toronto congregate settings (such as homeless shelters).	Str
Age Group	Age at time of illness. Age groups (in years): ≤ 19 , 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90+, unknown (blank)	int
Neighbourhood Name	Toronto is divided into 140 geographically distinct neighborhoods. These are the names of neighborhoods cases occur in.	Str

FSA	Forward sortation area (i.e. first three characters of postal code) based on the case's primary home address. FSA values are generated from client postal codes. One FSA can span multiple neighbourhoods.	Str
Source of Infection	The suspected / reported source of COVID infection	Str
Classification	Whether or not the case is confirmed or probable	Str
Episode Date	The episode date is a derived variable that best estimates when the disease was acquired, and refers to the earliest available date from: symptom onset (the first day that COVID-19 symptoms occurred), laboratory specimen collection date, or reported date.	Date (Str)
Client Gender	Self-reported gender. Gender is a system that operates in a social context and generally classifies people based on their assigned biological sex.	Str
Outcome	Fatal: Cases with a fatal outcome reported. Resolved: Cases not reported as deceased, and who are either: Reported as 'recovered' OR Where the report date is more than 14 days from symptom onset AND the case is not currently hospitalized. This number is underreported due to a lag in data entry.Active: All other cases	Str
Ever Hospitalized	Cases that were hospitalized related to their COVID-19 infection(includes cases that are currently hospitalized and those that have been discharged or are deceased).	Bool
Ever in ICU	Cases that were admitted to the intensive care unit (ICU) related to their COVID-19 infection (includes cases that are currently in ICU and those that have been discharged or are deceased).	Bool
Ever Intubated	Cases that were intubated related to their COVID-19 infection (includes cases that are currently intubated and those that have been discharged or deceased)	Bool
_id	Unique row identifier for Open Data database	int
OCCUPANCY_DATE	Date of the data	Date (str)

ORGANIZATION_NAME	Organization Name. The registered name of the non-profit entity responsible for the shelter operation.	Str
SHELTER_NAME	Name of the shelter site. The shelter name generally delineates the subgroup within the organization that is responsible for the shelter. For example the City of Toronto is the organization responsible for the shelters named: Women's Residence, Family Residence, Seaton House etc.	Str
SHELTER_POSTAL_CODE	Postal code of the shelter	Str
FACILITY_NAME	Name of the facility. In some cases shelters share the same name but the actual service is provided in multiple facilities. For example Women's Residence has 3 facilities: the Alexandra Hotel, 63 Bellwoods and 674 Dundas Street West	Str
PROGRAM_NAME	Name of the program within the facility. Some shelters offer different programs or services within the same facility. For example Seaton House offers emergency shelter on certain floors of 339 George Street but other floors within the same facility are dedicated to longer staying clients. In this example, the program name is the "Long Term" program.	Str
SECTOR	Clientele of the shelter. Sector can be women, men, youth, co-ed, or family.	Str
OCCUPANCY	Number of homeless clients in the program at 4:00 AM. Occupancy is a point in time number at 4 AM of the next day. For example, the occupancy count of January 1st would be taken on January 2nd at 4 AM.	Int
CAPACITY	Maximum number of spaces available for homeless clients within the program.	Int

Justification

Splitting the data into these 4 tables has a few reasons. The two larger tables keep track of the larger case data and shelter data by date which is important to keep track of. The first of the smaller tables let us understand which postal codes we have access to. This is important because the two larger tables have different postal codes so we need to keep track of these. The second smaller table keeps all of the shelter data, ignoring the multiple instances over many dates to get an exhaustive list of each shelter we're going to look at.