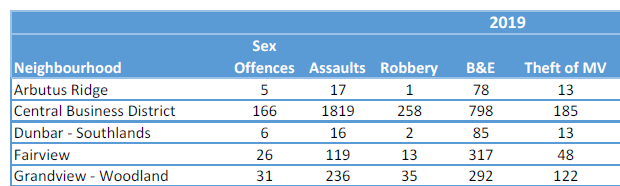
**Introduction/Business Problem**

In 1950, the percentage of the world population living in urban areas was just 30%. As urbanization runs its course in this fast-growing era, this number is projected to increase to 68% in 2050 as estimated by the UN. Amongst this, wealthy Canada is unsurprisingly at the higher end; it is estimated to have an urban population of 87% in 2050. That means that compared to 2021, 8 million *more* people will try to squish into its cities![1]

As someone living in Vancouver, one of the most populous cities in Canada, I believe urban municipalities need to be prepared to welcome these newcomers. One question I can imagine migrants will have is “there are so many neighborhoods, where should I live?”. Why I’m glad you asked! My goal for this project is to help migrants understand the different neighborhoods in the top Canadian cities. For this, I will be comparing and grouping neighborhoods with respect to its most popular venues and its crime rate.

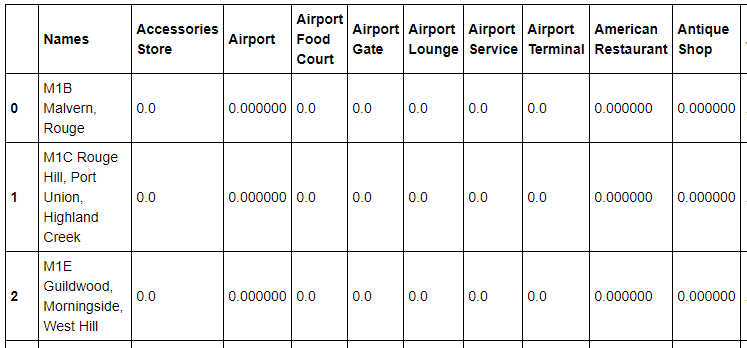
**Data**

I chose to analyse data from three big urban cities in Canada, namely Toronto, Montreal, and Vancouver [2]. I will be using the Foursquare API to gather data about popular venues. As for crime rate, the data for each city is openly accessible from their respective police departments’ websites. The neighborhoods that I am investigating are the ones that are characterized in the police data. Furthermore, I decided to retrieve data from the year 2019 because I expect 2020 data will differ from the norm due to COVID-19 and social distancing.

First, let’s dive into police data. My goal is to get the rate in which different types of crimes occur in each neighborhood. This specific information is conveniently shared in the annual report of all three police forces’ open dataset. Below is an example of the data format: 

The respective forces are Vancouver Police Department (VPD) [2] in Vancouver, Toronto Police Service (TPS) [3] in Toronto, and Service de police de la ville de Montreal (SPVM) [4] in Montreal. There are slight differences in how the data is presented between each police force. For example for VPD and not the others, homicide and attempted murders were excluded in the crime rate by neighborhood dataset, presumably to not cause distress in the public by knowing where homicide occurred. I will only include features that are shared by all datasets. The rate for a specific crime type in a specific neighborhood will be calculated as follows:

Next, let’s look at location data. I want to calculate the rate in which different venue types occur in each neighborhood, to judge how popular they are there. I will use the Foursquare API “venue/explore?” option to get a list of the popular venues by giving it the address of each neighborhood. I will then extract the “category type” variable for each venue and count the number of venues in each category type for each neighborhood. Below is an example of the data after re-organizing:



Furthermore, the rate for a specific venue type will be calculated as follows:

The location and crime data for all neighborhoods in all cities will then be merged. With rows consisting of neighborhoods and columns consisting of features. The data will be fed into unsupervised learning model to be clustered into groups.

[1] Hannah Ritchie (2018) - "Urbanization". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/urbanization' [Online Resource]

[2] https://vancouver.ca/police/organization/planning-research-audit/neighbourhood-statistics.html

[3] https://data.torontopolice.on.ca/pages/asr-archives

[4] https://spvm.qc.ca/en/Pages/Discover-SPVM/Organization/Annual-Reports