

Data

To analyze data and obtain any meaningful insight from it, raw data must first be collected.

Calgary's pedestrian and bike traffic data were collected from the city of Calgary's Open Data Portal. Location information (Latitude and Longitude) include postal codes, borough and neighborhood in Calgary where collected (Web-Scrapped) from Wikipedia. Data regarding Calgary's top venues and location were collected using FourSquare API. Hence, the following are data sources used for this project:

1. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T) – To retrieve postal codes, borough, Latitude and Longitude data.
2. (<https://data.calgary.ca/Transportation-Transit/Pedestrian-Traffic-Count/nz3v-2tau>) – To retrieve pedestrian traffic count.
3. (<https://data.calgary.ca/Transportation-Transit/Bike-Traffic-Count/pubj-un4t>) – To retrieve bike traffic count.
4. (<https://developer.foursquare.com>) – To retrieve the top venues or hotspots in Calgary.

Data Cleaning and Feature Extraction

The first data was obtained from Wikipedia. Wikipedia webpage was web-scraped using pandas read_html syntax. The table already had all the information we needed for our analysis; therefore we won't be using Geolocator to obtain both longitude and latitude.

The second data was a csv file downloaded from city Calgary's open data portal. The file had 31,862 rows and 102 columns. The columns contained information on: **Date** the data was collected, the **location** (address), time the information was collected from (12am to 11:45pm).

The third data was also csv file downloaded from city Calgary's open data portal. The file had 60,000+ rows and 102 columns. The columns contained information on: **Date** the data was collected, the **location** (address), time the information was collected from (12am to 11:45pm).

The last data will be retrieved using FourSquare API and a personal login credential. Data from the API will be used to determine the top venues and hotspots in the city. Neighbors will then be clustered based on their most popular services.