Applied Data Science Capstone Project

The Battle of Neighbourhoods



Analysis of Neighbourhoods and Demographics of Calgary, Alberta, Canada.

Final Report

Submitted By:

Dhananjay Saxena

Table of Contents

|  |  |  |
| --- | --- | --- |
| S No. | Topic | Pg no |
| 1. | Discussion of the background | 3 |
| 2. | Description of the problem | 4 |
| 3. | Target Audience | 4 |
| 4. | Data and use | 5 |
| 5. | Methodology | 8 |
| 6. | Results | 12 |
| 7. | Discussion | 15 |
| 8. | Conclusion | 16 |

Discussion of the background

**Calgary** is a city in the western Canadian province of [Alberta](https://en.wikipedia.org/wiki/Alberta). It is situated at the confluence of the Bow River and the Elbow River in the south of the province, in an area of foothills and [prairie](https://en.wikipedia.org/wiki/Canadian_Prairies), about 80 km (50 mi) east of the front ranges of the Canadian Rockies, roughly 299 km (186 mi) south of the provincial capital of [Edmonton](https://en.wikipedia.org/wiki/Edmonton) and approximately 240 km (150 mi) north of the [Canada–United States border](https://en.wikipedia.org/wiki/Canada%E2%80%93United_States_border). The city anchors the south end of the [Statistics Canada](https://en.wikipedia.org/wiki/Statistics_Canada)-defined urban area, the [Calgary–Edmonton Corridor](https://en.wikipedia.org/wiki/Calgary%E2%80%93Edmonton_Corridor).

The city had a population of 1,285,711 in 2019, making it Alberta's [most-populous city](https://en.wikipedia.org/wiki/List_of_cities_in_Alberta#List) and the most-populous in [western Canada](https://en.wikipedia.org/wiki/Western_Canada). In 2016, Calgary had a metropolitan population of 1,392,609, making it the [fourth-largest](https://en.wikipedia.org/wiki/List_of_census_metropolitan_areas_and_agglomerations_in_Canada) census metropolitan area (CMA) in Canada and second-largest in western Canada (after [Vancouver](https://en.wikipedia.org/wiki/Vancouver)).

The downtown region of the city consists of five neighbourhoods: Eau Claire (including the Festival District), the Downtown West End, the Downtown Commercial Core, Chinatown, and the Downtown East Village (also part of the Rivers District). The commercial core is itself divided into a number of districts including the Stephen Avenue Retail Core, the Entertainment District, the Arts District and the Government District. Distinct from downtown and south of 9th Avenue is Calgary's densest neighbourhood, the Beltline. The area includes a number of communities such as Connaught, Victoria Crossing and a portion of the Rivers District. The Beltline is the focus of major planning and rejuvenation initiatives on the part of the municipal government to increase the density and liveliness of Calgary's centre.

The city of Calgary is diverse in terms of culture, ethnicity, religion which implies there is diversity in terms of venues people like to visit and spend their leisure time.

Description of the problem

The objective of this capstone is analysing the population distribution in the city of Calgary as well as use the Foursquare API to analyse neighbourhoods in Calgary. Using Data Science Methodology and Machine Learning technique Like K Means Clustering this project aims to provide solutions to the question “How similar are neighbourhoods in terms of venues visited?”. We will also be able to suggest most frequently visited venues for every neighbourhood.

Target Audience

The project will come in handy to anybody who wishes to analyse the population distribution in the city of Calgary. Anyone who wishes to move to the city can look through the project to figure out the neighbourhood they would like to settle in as well the venues they will have access to.

About the data

1.**Wikipedia:**

The dataset to be used in this project shall be scraped from the website [Calgary Data](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary). The Wikipedia web page contains a list of all neighbourhoods within the City of Calgary including residential communities, industrial areas, major parks and residual areas by electoral ward. Calgary, Alberta, Canada, as of 2016, has 197 neighbourhoods, which are referred to as "communities" by the City of Calgary, and 42 industrial areas.



2**.Foursquare API:**

Foursquare, Foursquare is a technology company that built a massive dataset of location data. What is interesting about Foursquare is that they were very smart about building their dataset. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

Please click the here: [**Foursquare**](https://foursquare.com/) to visit their website.



3.**Geopy:**

The geographic co-ordinates for the neighbourhoods/communities has been incorporated into the dataset with the help of Geopy. Geopy is a Python client for several popular geocoding web services.

Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

Geopy includes geocoder classes for the OpenStreetMap, Nominatim, Google Geocoding API (V3), and many other geocoding services. The full list is available on the Geocoders doc section. Geocoder classes are located in geopy.geocoders.

Geopy is tested against CPython (versions 3.5, 3.6, 3.7, 3.8) and PyPy3. Geopy 1.x line also supported CPython 2.7, 3.4 and PyPy2.

Please click here: [**GeoPy**](https://geopy.readthedocs.io/en/stable/) to visit the GeoPy documentation page.



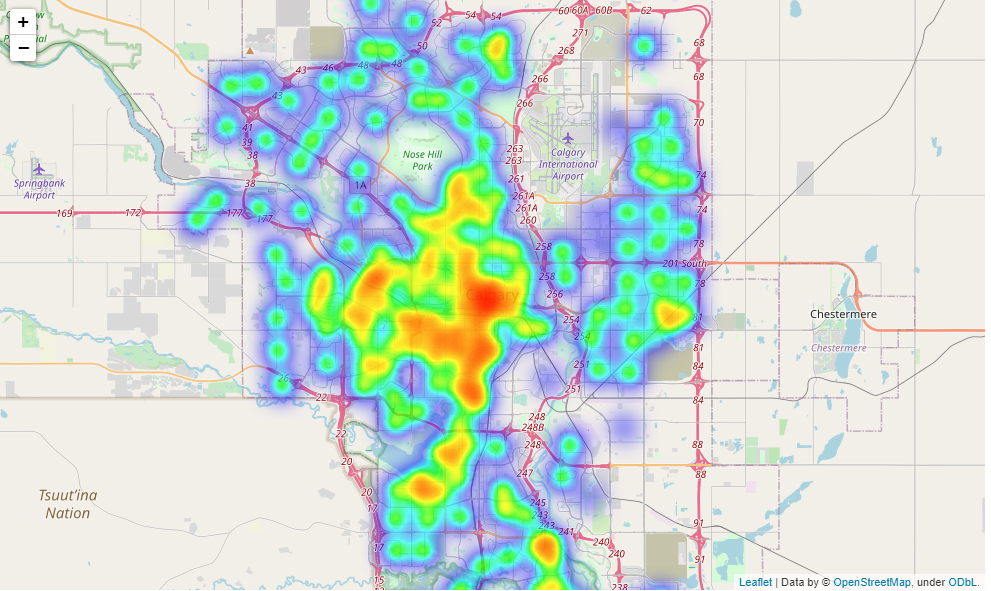
The use of data

The table containing Calgary’s neighbourhood data will be scraped from the Wikipedia page and will be combined with GeoPy module’s coordinate to prepare a dataset which will contain data regarding the city’s neighbourhood, its population density and its latitude and longitude.

FoursquareTM API will be used to import the data regarding the venues belonging to all the neighbourhoods. This data will be cleaned and prepared to obtain data to analyse

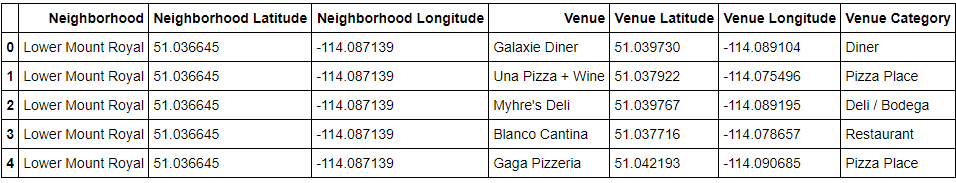
Methodology

In order to understand the population distribution throughout the city, we will import the dataset and then clean the data of unecessary data as well as values with absurd/no context. This will be followed by integration of geometric co-ordinates into our dataframe which will eventually help in plotting of the dataframe to the map of Calgary.

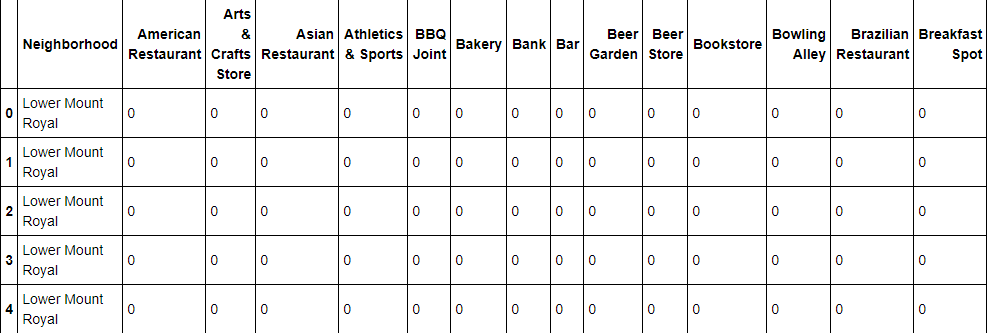


We will then follow the steps below :

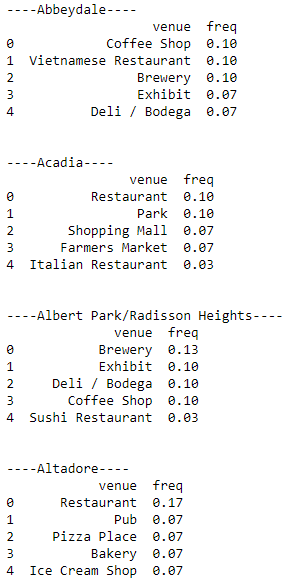
1. Explore the city venues with the help of FourSquareTM API.



1. Using One-hot encoding we will turn categorical variables into numeric data



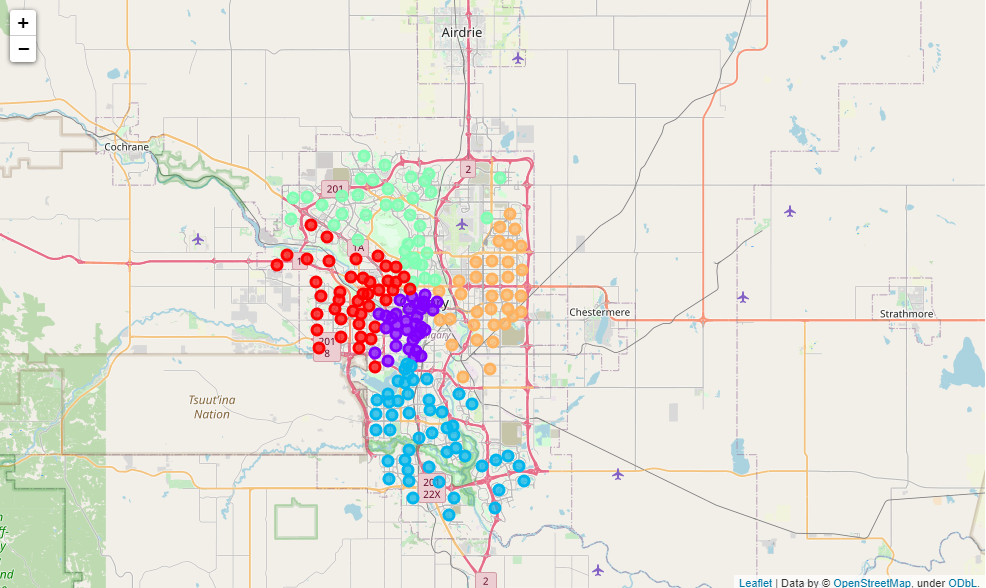
1. Get top 5 venues for each spot



1. Cluster te neighborhoods : Using K Means Clustering algorithm and merge it into the table :



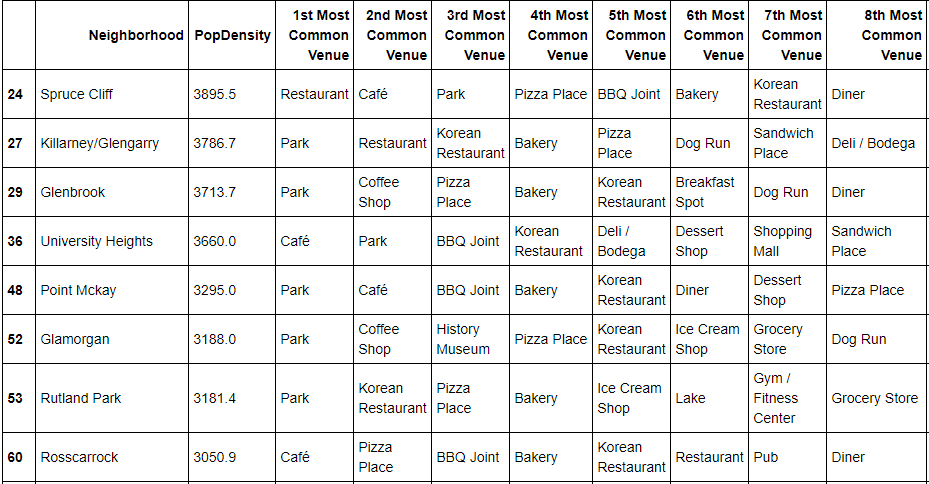
1. Map the clusters into neighborhood:



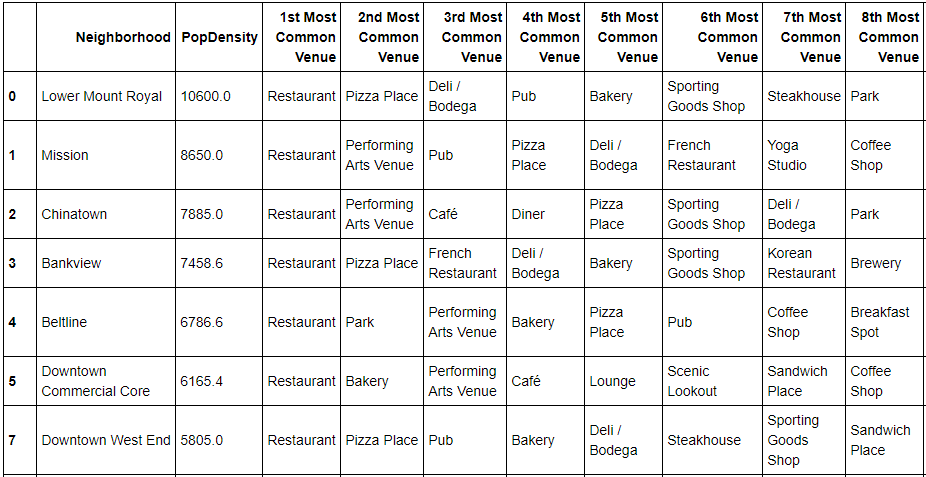
Results

The results of 5 clusters are specified in this section. The neighborhoods are sorted based on the most common venues visited.

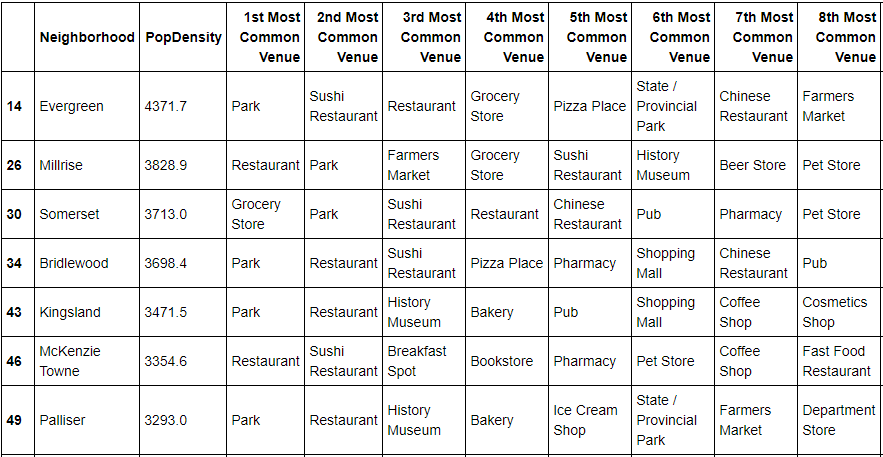
1. Cluster 1:



1. Cluster 2:



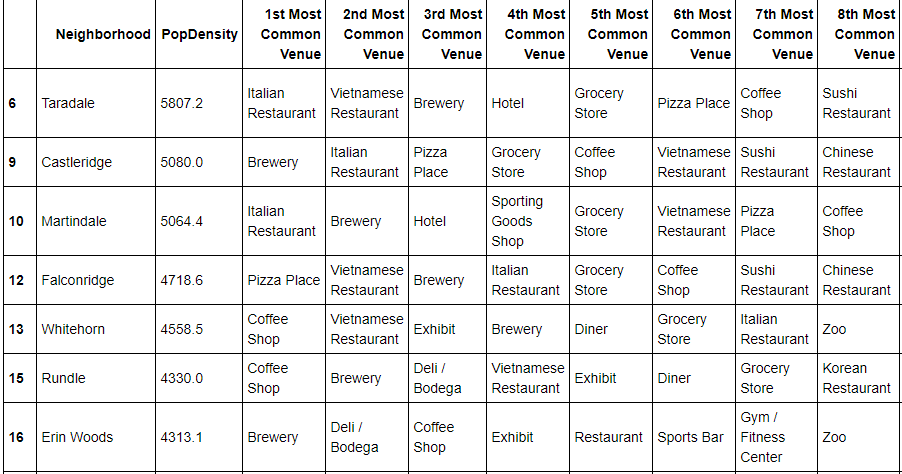
1. Cluster 3:



1. Cluster 4:



1. Cluster 5:



Discussion

According to the analysis, Cluster 1 is good for anyone who wishes to reside in an area where population is neither too large nor too small. With alot of parks and a Cafe this place is remcommended for someone who loves going out alot. Cluster 2 has plenty of restaurants and pubs with higher population density than most throughout the city. A person who like nightlife and social interactions would be advised to choose this cluster.Cluster 3 is on the lower end of population spectrum with more amenity based venues like supermarket. This place would be advised to families due to plethora of convenience stores. Cluster 4 has alot of speciality restaurants and would be advised for tourists and hotel owners due to the fact it has alot of tourist attracting avenue. Cluster 5 would also be advised to families since it is slightly higher on the scale of population than normal as well has alot of fitness-related as well as family oriented venues.

Conclusion

We used the result of analysis of demographics and neighborhoods to advise people moving into the city of Calgary, the neighborhoods that would be suited for them based on their likings and dislikings. This project has some limitations :

1. The dataset has been taken from WikiPedia which last updated the data in 2016 regarding the neighborhoods of the city.
2. The venues have been selected entirely from the FourSquare API. Any venues specified outside of this API which weren’t included/which have started their buisness recently may not be included in this project.