Analyze and Compare Tech Hubs in Canada

Shivani Sheth

4 June 2020

1. Introduction

1.1 Background

Canada is one of the fastest-growing technical hubs in the world, with an increasing number of resources provided by the government to attract bright minds all over the world. This is especially in the case of STEM students looking for study opportunities in the country. One of the most difficult questions that one faces while deciding upon a university/ college is its location. I myself being a prospective student in Canada, feel that a suitable neighborhood is an important factor as it not only enhances a student's college experience but also helps decide a preferable location for internships or exchange programs. Hence, here we explore and analyze the province of British Columbia, which is one of the top technical hubs of the country, in terms of science and engineering, and compare it with the previously analyzed data for Toronto, which is also famous for its prime opportunities in the field.

1.2 Proposal

In this project, we take leverage of the Foursquare API to find the most popular places in each of the neighborhoods consisting of British Columbia. A place is marked as "happening" by the Foursquare API according to the number of people present at a given place and hence the place is updated in real-time; it might change every few minutes. We then cluster the neighborhoods based upon their preferred places in the surrounding area. This will give a clear picture of the aura or the vibe of the place, which can help an individual know what to expect in the neighborhood, and hence decide upon a suitable location according to their preference. Finally, we also compare it with Toronto, another widely preferred city among prospective students, and brief upon the similarities and dissimilarities based upon its neighborhood.

1.3 Interest

This analysis can be useful for many people in different domains. First, it could be

used by business owners who might want to target students in the STEM domain, such as manufactures in the IT industry. Second, the analysis could be useful to the entrepreneurs in the engineering field, who are likely to open up their startups in the technical hubs of a country, to attract bright young minds freshly out of college. Third, it can also be used by prospective STEM-based students like me, to decide upon a preferable location for their universities, colleges, or exchange programs. Last but not the least, it can be used to compare the environment between the top two technical hubs of Canada and the type of crowd in the different neighborhoods, which could interest companies planning to shift their headquarters or open up a new branch.

2. Data

2.1 Acquisition

The data acquired for the purpose of this project can be obtained in many ways. The primary datasets that we need are the postal codes of the province of British Columbia, along with their latitude and longitude coordinates. After many tries with various datasets, I came upon the complete dataset used in this project on 'Postal Codes British Columbia, Canada - GeoNames'. We scrape the table from the given HTML page using the read_html function of the Pandas library. This will give us the primary table needed for the project, as shown in the figure below.

	Unnamed: 0	Place	Code	Country	Admin1	Admin2	Admin3
0	1.0	Port Moody	V3H	Canada	British Columbia	NaN	NaN
1	NaN	49.323/-122.863	49.323/-122.863	49.323/-122.863	49.323/-122.863	49.323/-122.863	49.323/-122.863
2	2.0	Pitt Meadows	V3Y	Canada	British Columbia	NaN	NaN
3	NaN	49.221/-122.69	49.221/-122.69	49.221/-122.69	49.221/-122.69	49.221/-122.69	49.221/-122.69
4	3.0	White Rock	V4B	Canada	British Columbia	NaN	NaN

2.2 Cleaning and Preprocessing

As we see, the scrapped table contains the address information in every alternate record, and following that is a row that contains the coordinates of the location. The coordinates are given in the form of 'lat/long' values, and the entire row consists of the same values. Our aim here is to fetch the alternate rows for the address values and just one cell from its next row for the coordinates.

We begin our data cleaning and preprocessing by extracting alternate address rows of this table in a separate data frame, say 'temp_df'. For this, we loop over

the number of records in a multiple of two and fetch each row. This will give us a total of 192 rows containing the neighborhood address of British Columbia, which looks like this:

	index	Place	Code	Country	Admin1
374	188.0	Vancouver (Strathcona / Chinatown / Downtown E...	V6A	Canada	British Columbia
376	189.0	Vancouver (NE Downtown / Harbour Centre / Gast...	V6B	Canada	British Columbia
378	190.0	Richmond South	V7A	Canada	British Columbia
380	191.0	Duncan	V9L	Canada	British Columbia
382	192.0	Parksville	V9P	Canada	British Columbia

Next, we delete the rows that have a float value in the index column of the original data frame, say 'original_df' and hence we are left with only the coordinate values in the table. Now, we keep any one of the columns and remove all the other ones as all the cells of a row contain the same data, and hence it is redundant. Hence, we obtain the coordinates for each address in one column. As the columns in both data frames, the 'original_df' and the 'temp_df', have data ordered according to their sequences in the table, we can merge both of them back into the original_df. We thus have all our data, a total of 192 records, in a single data frame as shown below:

	Coordinates	Place	Code	Country	Admin1
0	49.323/-122.863	Port Moody	V3H	Canada	British Columbia
1	49.221/-122.69	Pitt Meadows	V3Y	Canada	British Columbia
2	49.026/-122.806	White Rock	V4B	Canada	British Columbia
3	49.481/-119.586	Penticton	V2A	Canada	British Columbia
4	49.866/-119.739	Westbank	V4T	Canada	British Columbia

Now, for processing convenience, we convert the coordinate values into different columns consisting of the latitude and longitude values separately. We do this by storing the coordinate values into a list and by iterating through each value. We split the string by the "/" character, and hence obtain two parts of the same string. The former would be the latitude values and the latter would be the longitude values. We initially append both lat and long values in a list and after the iteration process completes, we assign it to the data frame column. Finally, we rename our columns appropriately, which completes our data cleaning and preprocessing. Our final data frame looks like this:

	Coordinates	Place	Code	Country	Province	Latitude	Longitude
0	49.323/-122.863	Port Moody	V3H	Canada	British Columbia	49.323	-122.863
1	49.221/-122.69	Pitt Meadows	V3Y	Canada	British Columbia	49.221	-122.69
2	49.026/-122.806	White Rock	V4B	Canada	British Columbia	49.026	-122.806
3	49.481/-119.586	Penticton	V2A	Canada	British Columbia	49.481	-119.586
4	49.866/-119.739	Westbank	V4T	Canada	British Columbia	49.866	-119.739

2.3 Alternatives

The same data can also be obtained from various other sources such as the Wikipedia pages that contain the postal codes for all the provinces in Canada, ordered by their regions. The page that can be used to obtain the postal codes for British Columbia, along with their neighborhoods is given here. Although this has the complete information needed for the scope of the project, the table is not ordered in a scraping-friendly manner. Hence, upon scraping, we obtain the information in a manner shown below.

	V1AKimberley	V2APenticton	V3ALangley Township(Langley City)	V4ASurreySouthwest
0	V1BVernonEast	V2BKamloopsNorthwest	V3BPort CoquitlamCentral	V4BWhite Rock
1	V1CCranbrook	V2CKamloopsCentral and Southeast	V3CPort CoquitlamSouth	V4CDeltaNortheast
2	V1ESalmon Arm	V2EKamloopsSouth and West	V3ECoquitlamNorth	V4EDeltaEast

This happens because the data stored for a complete row is in the same box, and hence for each cell in the table, we have the complete address of the region. To iterate through this table and make this data readable, we first store the table values in a list format using the df.values.tolist() function. This gives us a nested list structure that contains the rows and columns of the table which looks like this:

```
[['V1BVernonEast',
  'V2BKamloopsNorthwest',
  'V3BPort CoquitlamCentral',
  'V4BWhite Rock',
  'V5BBurnaby(Parkcrest-Aubrey / Ardingley-Sprott)',
  'V6BVancouver(NE Downtown / Gastown / Harbour Centre / International Village
  'V7BRichmond(Sea Island / YVR)',
  'V8BSquamish',
  'V9BVictoria(West Highlands / North Langford / View Royal)'],
 ['V1CCranbrook',
  'V2CKamloopsCentral and Southeast',
  'V3CPort CoquitlamSouth',
  'V4CDeltaNortheast',
  'V5CBurnaby(Burnaby Heights / Willingdon Heights / West Central Valley)',
  'V6CVancouver(Waterfront / Coal Harbour / Canada Place)',
  'V7CRichmondNorthwest',
  'V8CKitimat',
  'V9CVictoria(Colwood / South Langford / Metchosin)'],
```

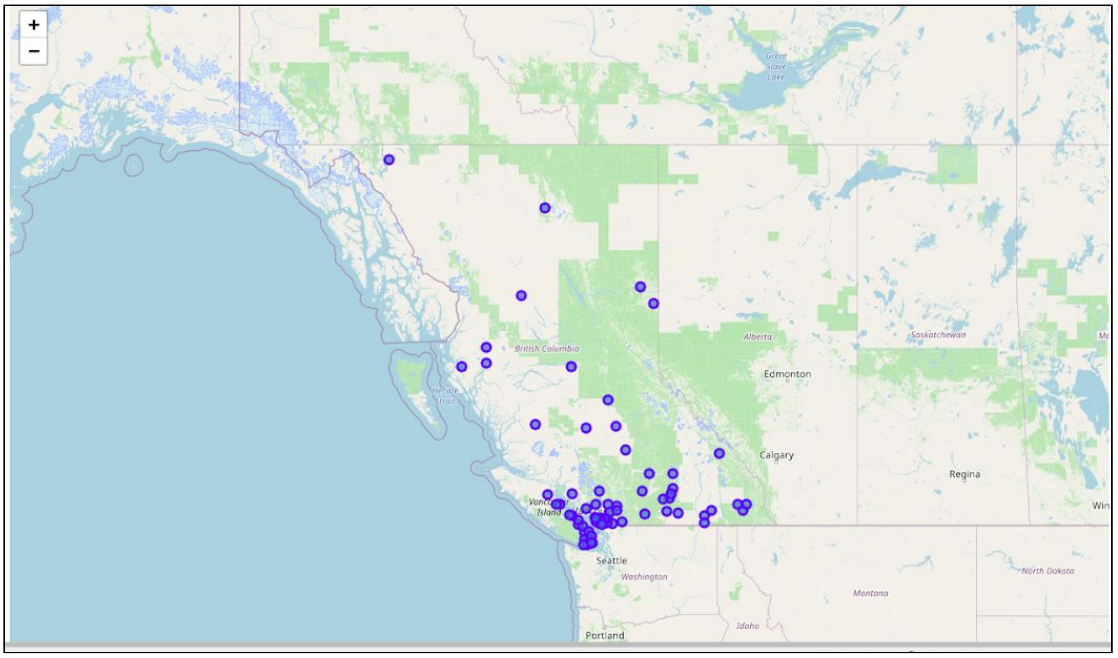
Now, we iterate through this list, fetch the first three alphabets of each string as the postal code, and the remaining characters as the neighborhood value. We store this information in two separate lists known as the 'postal_code' and 'neighborhood' and add them to a data frame. Now, our data frame consists of postal codes and neighborhoods in an orderly and readable manner, which looks something like this:

	Postal Codes	Neighborhood
0	V1B	VernonEast
1	V2B	KamloopsNorthwest
2	V3B	Port CoquitlamCentral
3	V4B	White Rock
4	V5B	Burnaby

Once we have obtained the addresses and the postal codes, we can use geocoder from the geopy library to find the latitude and longitude values for each pair of postal code and address and merge that in this data frame. Hence, we will have the same data as the previous method through different datasets.

3. Methodology

We begin our methodology section by outlining a map of British Columbia and superimposing all of the neighborhoods, from our data frame, onto the map. For this purpose, we utilize the Folium library. We also need the coordinates of the British Columbia province to map its outline, and hence, similar to the data acquisition process, we use the geocoder package to input its address and extract its coordinates. The map thus visualized using the information mentioned above, looks like this:



These are the neighborhoods in British Columbia that we aim to explore, analyze, and cluster.

3.1 Explore Neighborhoods

Now, using the addresses and coordinates of each neighborhood, we will leverage the Foursquare API in this section to explore the nearby venues of each neighborhood. Foursquare is a social location service that allows users to explore the world around them, and the Foursquare API allows us to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which we can send requests and it returns the data in a JSON file.

We begin our exploration process by defining our Foursquare credentials and the version number. Next, for each set of neighborhood names and coordinates, we prepare a URL request string and send a GET request. The output is the JSON file,

from which we extract the nearby locations and store them in a data frame. We define a function to repeat this process for each neighborhood from our data and store it in a new data frame named ‘neighborhood_venues’, which now contains the nearby venues for each neighborhood from our data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbotsford	49.0625	-122.3125	Discovery Trail	49.060245	-122.315565	Trail
1	Abbotsford	49.0625	-122.3125	Grandmas Market Gladwin Rd	49.066149	-122.313659	Grocery Store
2	Burnaby	49.2500	-123.0000	BCITSA's Stand Central SE2	49.251424	-123.001384	Snack Place
3	Burnaby	49.2500	-123.0000	BCIT Bookstore	49.251548	-123.001364	Bookstore
4	Burnaby	49.2500	-123.0000	The Rix @ BCIT	49.251153	-123.000636	Coffee Shop

We also check the number of venues returned for each neighborhood and observe that we have 76 unique venues.

3.2 Analyze Each Neighborhood

Now to analyze each neighborhood, we use the one-hot encoding technique and map the nearby venue categories against each neighborhood into a series of 0’s and 1’s. The categories that are present in the neighborhood are marked by 1, and the ones that aren't present in that neighborhood are marked by 0. Next, we add the neighborhood names corresponding to their values in the data frame and calculate the mean of venue categories present. The mean values typically range from 0 to 1.

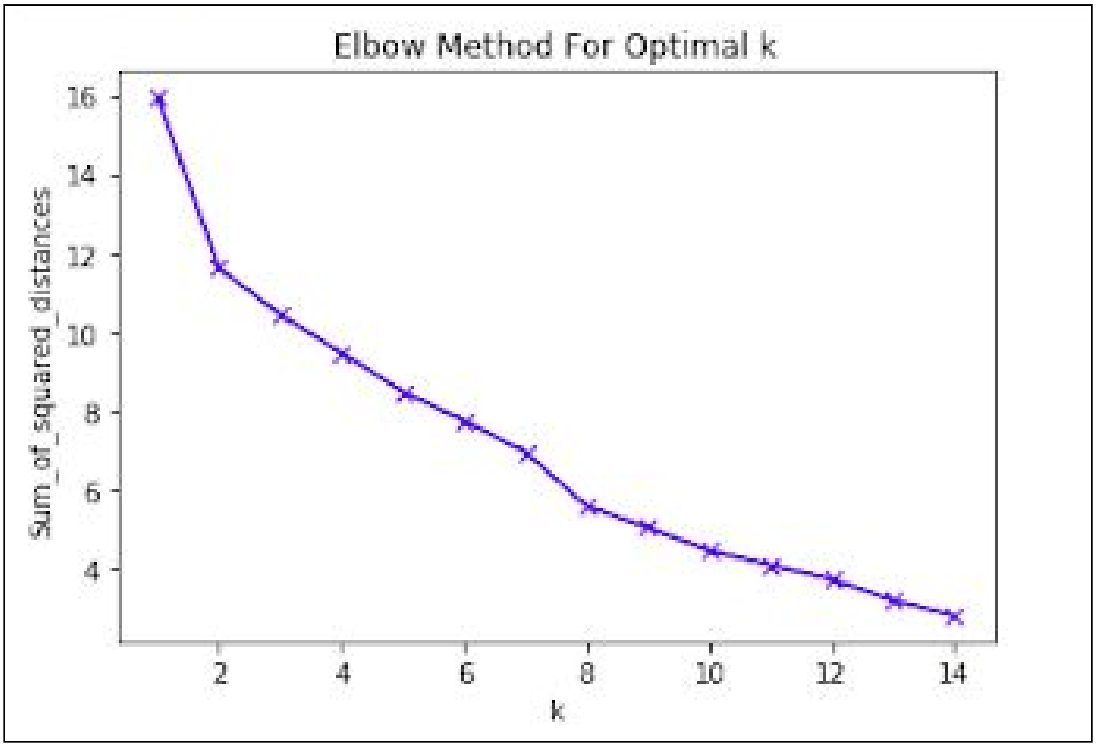
Neighborhood	American Restaurant	Asian Restaurant	Athletics & Sports	Auto Workshop	Bakery	Bank	Baseball Field	Beach	Boat or Ferry	Bookstore	Breakfast Spot
Abbotsford	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
Burnaby	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.111111	0.0
Comox	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
Coquitlam	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
Cranbrook	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0

This gives us an overview of the venue categories. To make it more clear and appealing, we can arrange the top ten venues of each neighborhood and display it in a separate data frame. To achieve this, we rearrange the mean values of the venue categories for each neighborhood and extract the top ten venue categories. We rename the columns and the new data frame that we now have is:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Abbotsford	Grocery Store	Trail	Falafel Restaurant	Coffee Shop	Construction & Landscaping
Burnaby	Bus Stop	Bookstore	Snack Place	Park	Bus Station
Comox	Fast Food Restaurant	Coffee Shop	Pharmacy	Sandwich Place	Juice Bar
Coquitlam	Asian Restaurant	Convenience Store	Golf Course	Gas Station	Coffee Shop
Cranbrook	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop
Duncan	Convenience Store	Gas Station	Dog Run	Zoo	Fast Food Restaurant
Esquimalt	Boat or Ferry	Fish & Chips Shop	Convenience Store	Dessert Shop	Dim Sum Restaurant

3.3 Cluster Neighborhoods

After analyzing each neighborhood, we cluster them according to their top most common venues. We use the k-means clustering algorithm to achieve this. One of the most important factors in K-Means Clustering is to choose the value of 'k', or the number of clusters for the dataset. To decide this, we first calculate the sum of squared distances for each k value and then plot a graph with the 'sum of squared distances' on the y-axis and the 'k' value on the x-axis. Ideally, from this graph, the best 'k' value would be the one with the least sum of squared distance but that usually results in a very high k value. Hence, we apply the elbow method to choose the optimum k value.



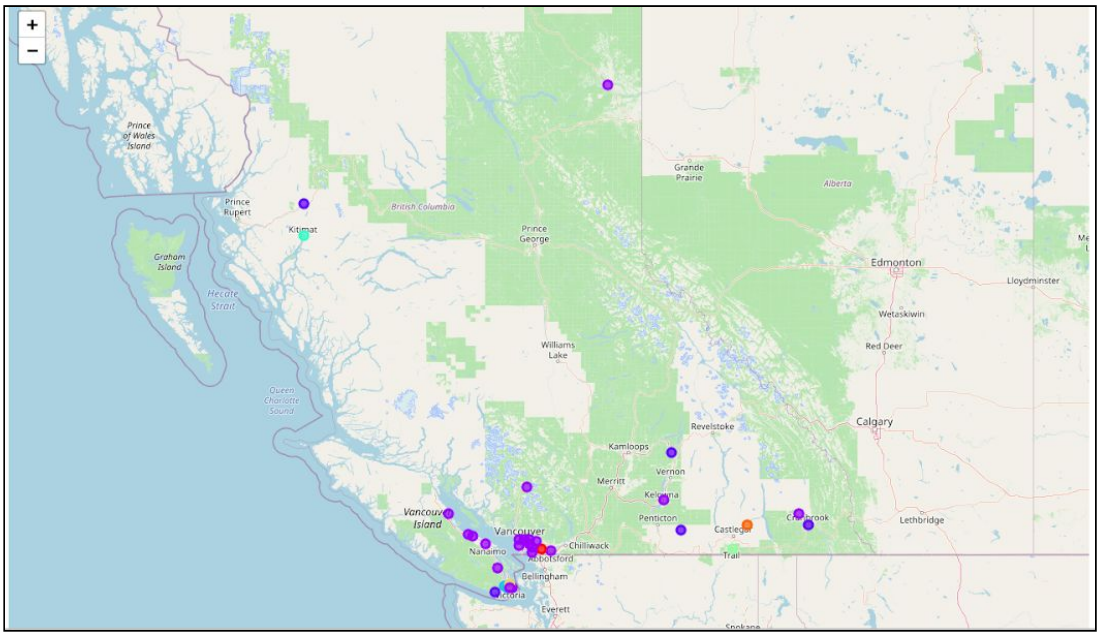
In the graph, as we increase the value of k, the distance decreases. The point which shows a steep decline in the distance, after which the decrease in the distance is comparatively less is known as the optimal value of k (also known as the elbow point). In this case, as we see from the graph, the elbow point is k=8, and we run the run k-means to cluster the neighborhood into eight different clusters. We extract the cluster labels for each neighborhood and append it to the 'top venues' data frame.

Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Abbotsford	Grocery Store	Trail	Falafel Restaurant	Coffee Shop	Construction & Landscaping
1	Burnaby	Bus Stop	Bookstore	Snack Place	Park	Bus Station
1	Comox	Fast Food Restaurant	Coffee Shop	Pharmacy	Sandwich Place	Juice Bar
1	Coquitlam	Asian Restaurant	Convenience Store	Golf Course	Gas Station	Coffee Shop
2	Cranbrook	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop

We can also calculate the number of neighborhoods in each cluster by the .value_counts() function.

```
1    20
2     5
3     2
7     1
6     1
5     1
4     1
0     1
Name: Cluster Labels, dtype: int64
```

Finally, we append the postal codes and coordinates from our original data frame into this new 'top venues' data frame, and using this, visualize the clusters superimposed onto the British Columbia map. As we did earlier, we use the Folium library to achieve this except we mark the different clusters in different colors to differentiate.



4. Results

Thus, the neighborhoods are grouped into eight different clusters based on the similarity of their nearby venues. We now examine each cluster and determine the discriminating venue categories that distinguish each cluster.

British Columbia Cluster 1:

(1, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Langley City	Baseball Field	Zoo	Fish & Chips Shop	Convenience Store	Dessert Shop

British Columbia Cluster 2:

(20, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Abbotsford	Grocery Store	Trail	Falafel Restaurant	Coffee Shop	Construction & Landscaping
1	Burnaby	Bus Stop	Bookstore	Snack Place	Park	Bus Station
2	Comox	Fast Food Restaurant	Coffee Shop	Pharmacy	Sandwich Place	Juice Bar
3	Coquitlam	Asian Restaurant	Convenience Store	Golf Course	Gas Station	Coffee Shop
4	Duncan	Convenience Store	Gas Station	Dog Run	Zoo	Fast Food Restaurant

British Columbia Cluster 3:

(5, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Cranbrook	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop
1	Salmon Arm	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop
2	Sooke	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop
3	South Okanagan	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop
4	Terrace	Construction & Landscaping	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop

British Columbia Cluster 4:

(2, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Esquimalt	Boat or Ferry	Fish & Chips Shop	Convenience Store	Dessert Shop	Dim Sum Restaurant
1	Highlands	Zoo	Theme Park	Boat or Ferry	Wine Shop	Auto Workshop

British Columbia Cluster 5:

(1, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Kitimat	Business Service	Zoo	Fast Food Restaurant	Convenience Store	Dessert Shop

British Columbia Cluster 6:

(1, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Trail	Pub	Falafel Restaurant	Coffee Shop	Construction & Landscaping	Convenience Store

British Columbia Cluster 7:

(1, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Saanich Central	Bank	Zoo	Fish & Chips Shop	Convenience Store	Dessert Shop

British Columbia Cluster 8:

(1, 11)

	Place	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Nelson	Trail	Zoo	Fast Food Restaurant	Construction & Landscaping	Convenience Store

5. Discussions - Comparison with Toronto

The neighborhoods in Toronto, another one of the top tech hubs in Canada, were explored and analyzed in our previous assignment. Those were clustered into four different groups, as illustrated below:

Toronto Cluster 1:

(32, 12)

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Downtown Toronto	0	Coffee Shop	Bakery	Pub	Park	Breakfast Spot
1	Downtown Toronto	0	Coffee Shop	Sushi Restaurant	Diner	Yoga Studio	Creperie
2	Downtown Toronto	0	Clothing Store	Coffee Shop	Café	Restaurant	Japanese Restaurant
3	Downtown Toronto	0	Coffee Shop	Café	Gastropub	Cocktail Bar	American Restaurant
4	Downtown Toronto	0	Coffee Shop	Cocktail Bar	Italian Restaurant	Beer Bar	Restaurant

Toronto Cluster 2:

(1, 12)

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Central Toronto	1	Garden	Yoga Studio	Department Store	Ethiopian Restaurant	Electronics Store

Toronto Cluster 3:

(1, 12)

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Central Toronto	2	Playground	Restaurant	Yoga Studio	Deli / Bodega	Eastern European Restaurant

Toronto Cluster 4:

(1, 12)

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	East Toronto	4	Trail	Health Food Store	Pub	Yoga Studio	Deli / Bodega

As we see, there is a similarity between the Toronto Cluster 1 and British Columbia Cluster 2. These two clusters have restaurants and coffee shops as their top common venues, which are generally visited by the working middle-aged population, on a very frequent basis. Hence the neighborhoods in these clusters would be a preferable choice for companies to set up offices, and for fresh graduates to look for work opportunities.

Another similarity is between Toronto Cluster 3 and British Columbia Cluster 4, which have common areas such as zoos, playgrounds, parks, and restaurants, that are more suitable for families with younger children. Hence, these neighborhoods would be highly populated by families, and sparsely populated by bachelors.

Other clusters include neighborhoods with Health Food Stores, Yoga Studios, Pubs etc. which might be suitable for different people based on their current situations.

6. Conclusion

The neighborhood clusters define the type of locality and can be used by different people to choose their preferred area of interest accordingly. For example, clusters involving a locality that has baseball fields, zoos in their most popular nearby venues and hence would be more suitable for a younger generation, preferably kids in their middle schools. On the contrary, the clusters involving a locality that has grocery shops, fast food shops, and convenience stores as their popular nearby locations would be more suitable for college students, who would need all of these on an almost daily basis.

Hence, these clusters provide a basic guide or a map on the different types of neighborhoods in British Columbia and give an idea of what to expect in that neighborhood.