Opening an Indian Restaurant in Vancouver

Abhijith Ravishankar

April 2021

1. Introduction

1.1 Background

Vancouver, a bustling west coast seaport in British Columbia, is among Canada's densest, most ethnically diverse cities known for its art and scenery. With a population of over 2.5 million, there are more than 200 ethnic groups in the country, making up more than 16 per cent of the total population. South Asians account for a quarter of all visible minorities in Canada, or four per cent of the total population. South Asians have lived in the Vancouver region since the late 19th century; at first, mainly working in the forestry industry. After an initial first wave of immigration during the early 20th century, government policies aimed at curtailing immigration from the Indian subcontinent resulted in a populated stagnation through the 1950s. At that time, the relaxing of racial and national immigration restrictions by the federal government initiated a new wave of immigration into Vancouver and has continued into the present day. The vast majority of South Asians in Greater Vancouver and in adjacent cities are Punjabi Sikhs (Indians), differing greatly from the diverse ethnic and religious composition of South Asians in Canada. A lot of Punjabi Sikhs now own and run their very own restaurants in and around parts of Vancouver area focusing on various cuisines of the Indian subcontinent. The majority of these Indo-Canadian restaurants focus on the cuisine of northern India. In 2013 Alexandra Gill of The Globe and Mail wrote that in regards to area food critics the Indian restaurant scene was "a largely unknown dining landscape".

1.2 Problem

Seeing the popularity of Indian food in Vancouver, the objective of the project is to use Foursquare API and clustering to best determine a neighborhood in Vancouver suitable to open an Indian restaurant. This will help business owners in deciding a convenient and apt location to open up a restaurant.

1.3 Interest

This is aimed towards business owners and entrepreneurs, basically anyone who wants to open an Indian restaurant or grow their current business into a chain. At the end, several locations would be given out as ideal spots to open up an Indian restaurant

2. Data acquisition and cleaning

2.1 Data sources

This project will use Vancouver neighborhood data, Geolocation of Vancouver and Foursquare API to figure out which neighborhood will be an ideal location to open an Indian restaurant.

- Data scraping and cleaning up the British Columbia Province Neighborhood Data from Wikipedia here.
- Geolocation of Vancouver - Will be determined via the Geocoder Package here
- Foursquare API – to show the location of various Indian restaurants and the ones lacking of such restaurants using this website

2.2 Data cleaning

This is a list of postal codes in Canada where the first letter is V. Postal codes beginning with V are located within the Canadian province of British Columbia. These will be narrowed down to only the ones that include 'Vancouver' in their Borough. Via webscraping, using BeautifulSoup library and parsing, we narrow down a dataframe which includes PostalCode, Borough and Neighborhood as the columns. Total number of Postal Codes comes out to be 43 where the Boroughs include 'Vancouver'.

	PostalCode	Borough	Neighborhood
0	V6A	Vancouver	Strathcona,Chinatown,Downtown Eastside
1	V6B	Vancouver	Downtown,Gastown,Harbour Centre,International ...
2	V6C	Vancouver	Waterfront,Coal Harbour,Canada Place
3	V6E	Vancouver	West End,Davie Village
4	V6G	Vancouver	West End,Stanley Park

Table1: Web scraping Wikipedia

With the help of Postal Code Database, the file in CSV will be then transformed into a data frame with 3 columns: Postal Codes, Latitudes and Longitudes for further analysis. After cleanup, and arranging the Postal codes with their latitudes and longitudes, they are then merged onto the dataframe above.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	V6A	Vancouver	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062
1	V6B	Vancouver	Downtown,Gastown,Harbour Centre,International ...	49.278755	-123.112479
2	V6C	Vancouver	Waterfront,Coal Harbour,Canada Place	49.285471	-123.116483
3	V6E	Vancouver	West End,Davie Village	49.283854	-123.127602
4	V6G	Vancouver	West End,Stanley Park	49.288790	-123.135592

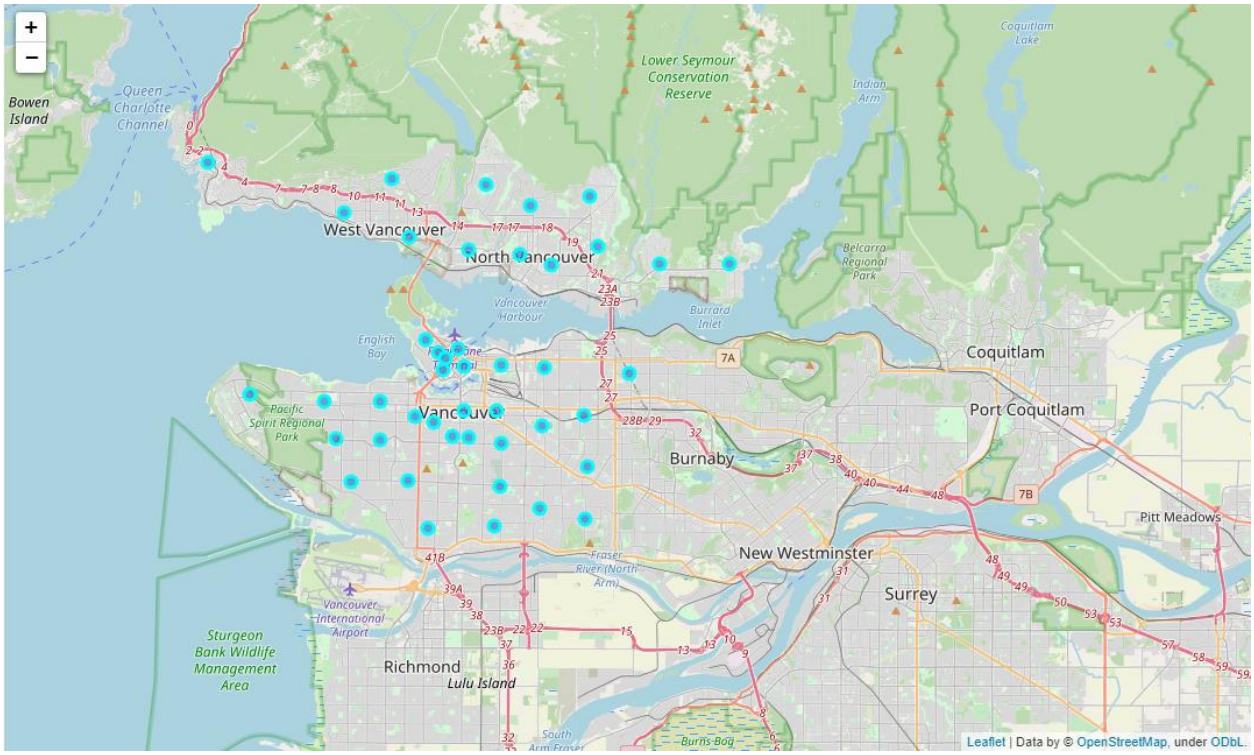
Table2: Adding CSV information to Wikipedia scraping

Using the above dataframes and my API credentials in Foursqaure, we will further explore the various venues in and around Vancouver before narrowing it down to only Indian restaurants.

3. Methodology

3.1 Visual of the neighborhood

First, we create a map of Vancouver that includes the 43 postal codes. This will help to understand our boundaries of the boroughs for future analysis. The map is generated using Folium and Table2 while using the average latitude and longitude of Vancouver



Map1: Vancouver neighborhoods via Folium

3.2 Foursquare API

Using my credentials from the developer account of Foursquare, nearby venues for various neighborhoods are extracted using a limit of 500 and a radius around the neighborhoods as 2km. The version of the extracted data was last updated end of 2020. Once this has been extracted, a dataframe is created where all venues are grouped by the various neighborhoods. 248 unique venue categories are found in the new dataframe. Some of the venue categories are Ice Cream Shop, Trail, Bakery, Gym, etc.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062	Union Market	49.277371	-123.086989	Deli / Bodega
1	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062	Phnom Penh	49.278517	-123.098214	Asian Restaurant
2	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062	The Mackenzie Room	49.283168	-123.094911	Restaurant
3	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062	Matchstick Coffee Roasters	49.278626	-123.099303	Café
4	Strathcona,Chinatown,Downtown Eastside	49.278849	-123.091062	Boxcar	49.276613	-123.100076	Bar

Table3: showing some of the venues and their category after using Foursquare API

3.3 One-Hot encoding

Often, machine learning tutorials will recommend or require that you prepare your data in specific ways before fitting a machine learning model. Categorical data are variables that contain label values rather than numeric values. And categorical data must be converted to a numerical form to work with machine learning. One-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

Here we use one-hot encoding to create a grouped dataframe that consists of the neighborhoods and provides a mean of frequency of occurrences of various venue categories.

	Neighborhood	Yoga Studio	Accessories Store	American Restaurant	Amphitheater	Aquarium	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	...
1	Arbutus Ridge,Dunbar-Southlands	0.010000	0.0	0.00	0.0	0.0	0.0	0.00	0.000000	0.010000	...
2	Bentall Centre	0.010000	0.0	0.01	0.0	0.0	0.0	0.00	0.010000	0.000000	...
3	Downtown	0.010000	0.0	0.00	0.0	0.0	0.0	0.01	0.000000	0.000000	...
4	Downtown,Gastown,Harbour Centre,International ...	0.000000	0.0	0.01	0.0	0.0	0.0	0.01	0.000000	0.020000	...
5	Dunbar-Southlands,Chaldecutt,University Endowm...	0.010526	0.0	0.00	0.0	0.0	0.0	0.00	0.010526	0.010526	...

Table4: showing some of the venue categories and their mean occurrences after using one-hot encoding

3.4 Narrowing down the Venue Categories

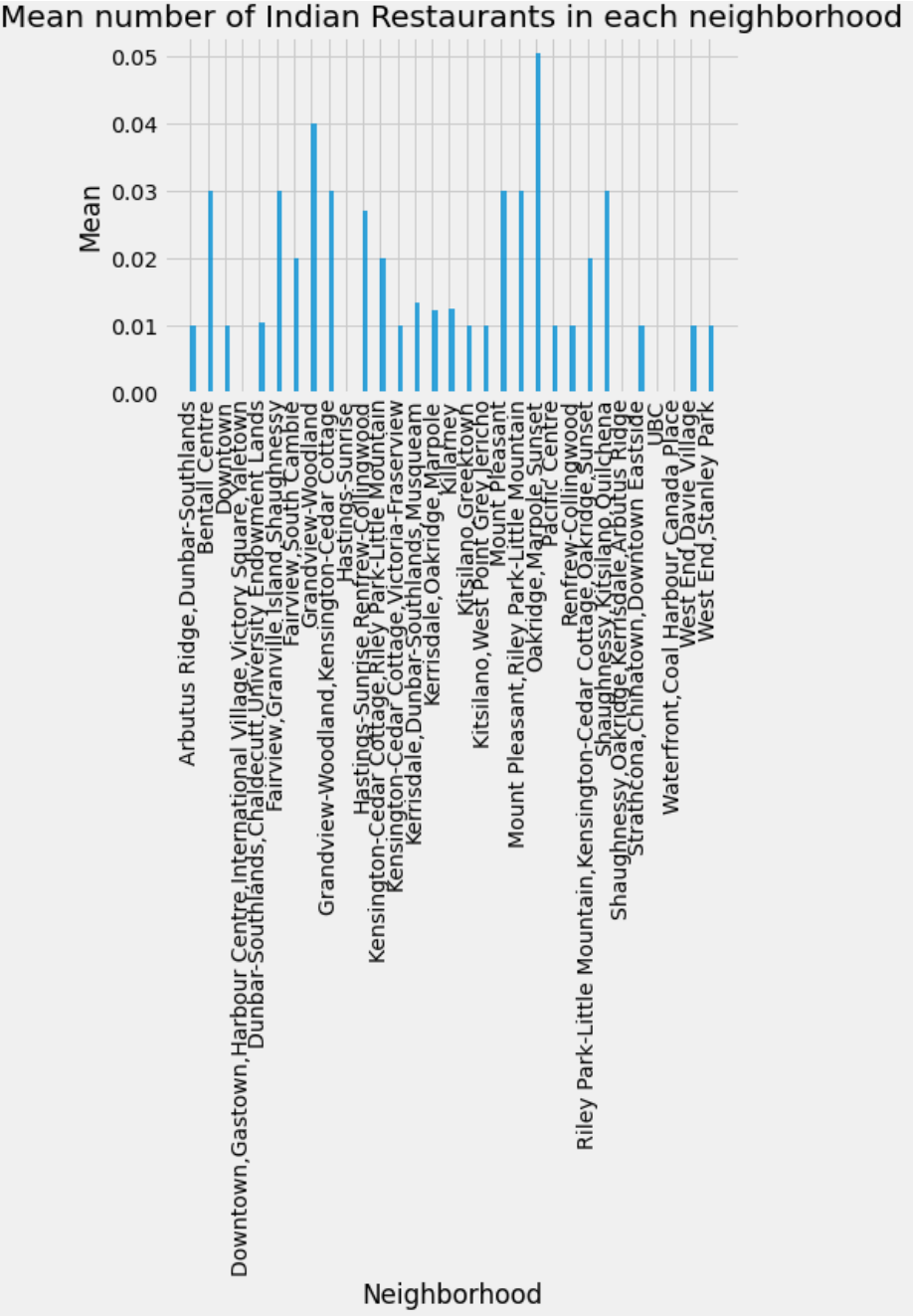
Using the above, another dataframe was created while limiting the venue categories to only 'Indian Restaurant' and including it along with the 'Neighborhood' column.

	Neighborhood	Indian Restaurant
1	Arbutus Ridge,Dunbar-Southlands	0.010000
2	Bentall Centre	0.030000
3	Downtown	0.010000
4	Downtown,Gastown,Harbour Centre,International ...	0.000000
5	Dunbar-Southlands,Chaldecutt,University Endowm...	0.010526

Table5: mean of frequency of Indian restaurant in each neighborhood

The mean of frequency of occurrences of Indian restaurants was outputted in the 2nd column. This was then plotted via a bar graph to show how Indian restaurants placed in each neighborhood.

As shown in the next page (Graph1), Oakridge, Marpole Sunset had the highest mean number with Grandview-Woodland trailing. UBC, Waterfront, Coal Harbour, Canada Place etc. were some neighborhoods which had no Indian restaurants



Graph1: Mean number of Indian restaurants in each neighborhood

Now that we have the necessary information, we use k-means clustering algorithm to find groups of neighborhoods which will then be used for analysis via the mean frequency to see where would be suitable locations to open shop. This is also used to confirm the above suspicions of locations not having Indian restaurants.

3.5 K-means Clustering Algorithm

Before we move ahead with K-means clustering, we have to first find the optimal number of 'k' to be used in our analysis. To determine this, we will use elbow method and the kneed library.

Elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

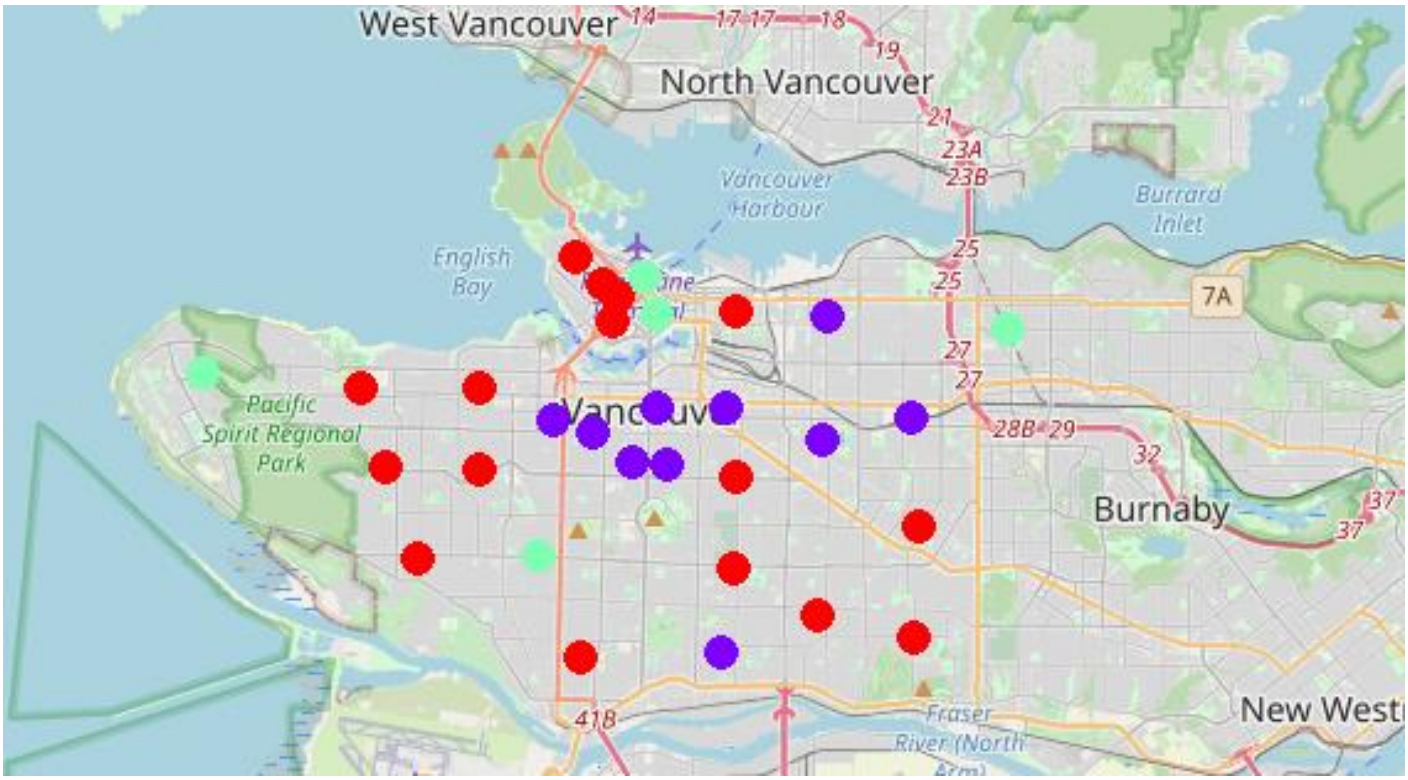
Running the k-means algorithm 10 times with different centroid seeds for a maximum of 300 iterations while determining random number generation for centroid initialization, we create a plot to visualize sum of squares errors against the total number of clusters. Below is the graph:

Graph2: Number of clusters vs Error sum of squares

One can see that the elbow effect happens and grows steeper around 3. To confirm this, we import Kneelocator from Kneed library. For the similar range of number of clusters shown above, i.e., (1,12), we are outputted 3 as the optimal number of clusters.

Now with the 3 clusters, we create a map to show where the clusters fall. Displayed below is the map where,

- Cluster 1 is red
- Cluster 2 is purple
- Cluster 3 is cyan



Map2: Showing the 3 clusters via K-means

Once these clusters are formed, we then breakdown and create dataframes for each of the clusters accordingly to their neighborhoods. With the help of this we are able to discern that there are

- 180 unique venue categories in the first cluster which is marked as red dots
- 152 unique venue categories in the first cluster which is marked as purple dots
- 127 unique venue categories in the first cluster which is marked as cyan dots

3.6 Cluster breakdowns for Indian Restaurants

The 3 above clusters are then broken down into 3 new dataframes that only contain ‘Venue Category’ as ‘Indian Restaurant’. This in turn gives us the below dataframes for each of the clusters

There are duplicate values for some Indian restaurants, we use ‘.unique’ to exactly display the total number of Indian restaurants then in each of the clusters

	Borough	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
12	Vancouver	Strathcona,Chinatown,Downtown Eastside	0.01	0	49.278849	-123.091062	Tandoori Palace	49.271653	-123.069721	Indian Restaurant
165	Vancouver	West End,Davie Village	0.01	0	49.283854	-123.127602	Crave India	49.279056	-123.123368	Indian Restaurant
248	Vancouver	West End,Stanley Park	0.01	0	49.288790	-123.135592	Mumbai Local	49.280910	-123.132244	Indian Restaurant
328	Vancouver	Kitsilano,Greektown	0.01	0	49.265073	-123.162119	The Indian Oven	49.268021	-123.150553	Indian Restaurant

Table 6: Cluster 1 where k = 0

	Borough	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
63	Vancouver	Fairview,Granville Island,Shaughnessy	0.03	1	49.256903	-123.130947	Indian Roti Kitchen	49.258950	-123.115087	Indian Restaurant
69	Vancouver	Fairview,Granville Island,Shaughnessy	0.03	1	49.256903	-123.130947	Vij's	49.257525	-123.115083	Indian Restaurant
98	Vancouver	Fairview,Granville Island,Shaughnessy	0.03	1	49.256903	-123.130947	The Indian Oven	49.268021	-123.150553	Indian Restaurant
117	Vancouver	Shaughnessy,Kitsilano,Quilchena	0.03	1	49.259470	-123.141797	The Indian Oven	49.268021	-123.150553	Indian Restaurant
194	Vancouver	Shaughnessy,Kitsilano,Quilchena	0.03	1	49.259470	-123.141797	Vij's	49.257525	-123.115083	Indian Restaurant

Table7: Cluster 2 where k = 1

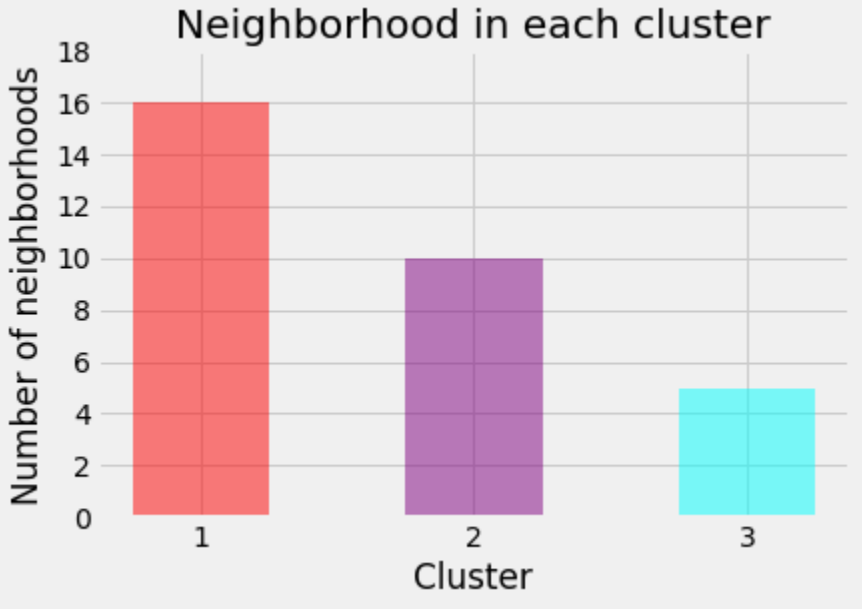
Borough	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category

Table8: Cluster 3 where k = 2

In the dataframes, you notice that all the venue categories are ‘Indian Restaurant’ and they all have similar ‘Cluster Labels’ which correspond to those neighborhoods in the clusters.

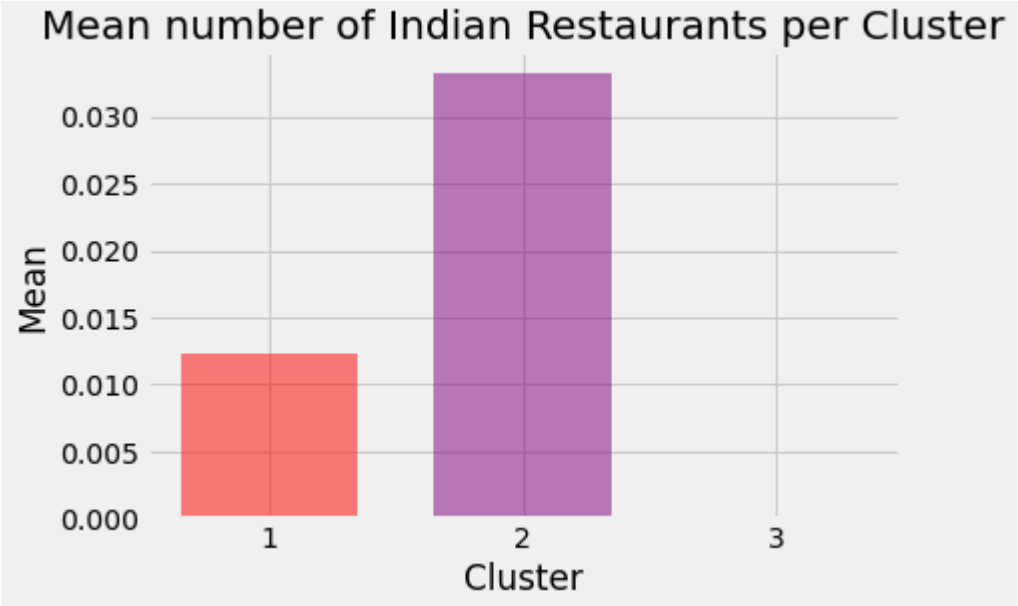
4. Results

- First a data frame was created to analyze all the Postal Codes in British Columbia that included 'Vancouver' in the Borough
    - Unique number of Postal Codes came out to be 43
    - All of them begun with the letter 'V'
- 248 unique venue categories were to be found via the Foursquare API with a limit of 500 and a radius of 2 km
- Mean number of Indian Restaurants in each neighborhood was displayed via a graph as shown in the methodology section
    - Oakridge, Marpole Sunset had the highest mean number with Grandview-Woodland trailing
    - UBC, Waterfront, Coal Harbour, Canada Place etc. were some neighborhoods which had no Indian restaurants
- Via K-means, 3 clusters were used and a map was created. Using the 'Cluster Labels', the value was counted for each cluster
    - Cluster1 has 16 neighborhoods
    - Cluster2 has 10 neighborhoods
    - Cluster3 has 5 neighborhoods



Graph3: Number of neighborhoods in each cluster

- Number of unique venue categories in each of the clusters came out to be:
    - Cluster1 had 180 unique categories
    - Cluster2 had 152 unique categories
    - Cluster3 had 127 unique categories
- Mean of the frequency of occurrence of Indian restaurant in each Cluster
    - Cluster1 had a mean of 0.0118
    - Cluster2 had a mean of 0.0329
    - Cluster3 has no Indian restaurants



Graph4: Mean number of Indian restaurants per Cluster

5. Discussions

Cluster 1 which occupies 16 different neighborhoods currently has 12 Indian restaurants. Cluster 2 which occupies 10 different neighborhoods currently has 16 Indian restaurants. Cluster3 which occupies 5 different neighborhoods has no Indian restaurants. According to the mean of frequency of occurrences of Indian restaurants, cluster 3 lacks the most with cluster 1 trailing behind. Cluster 2 consists of neighborhoods that have a lot of Indian restaurants. Ideal location would be the neighborhoods in cluster 3.

Cluster 3 neighborhoods include Downtown, Gastown, Harbour Centre, International Village, Victory Square, Yaletown, UBC, Shaughnessy, etc. which according to Graph1 confirms our theory.

6. Conclusion

After analyzing the mean of the frequency of occurrence of Indian restaurant in each Cluster and the number of neighborhoods in each cluster

- Cluster3 which occupies 5 different neighborhoods and has no Indian restaurants would be an ideal spot to start up an Indian restaurant!
- In Cluster3, Coal Harbour, UBC, Canada Place neighborhoods and some others would be an optimal location to startup an Indian Restaurant!