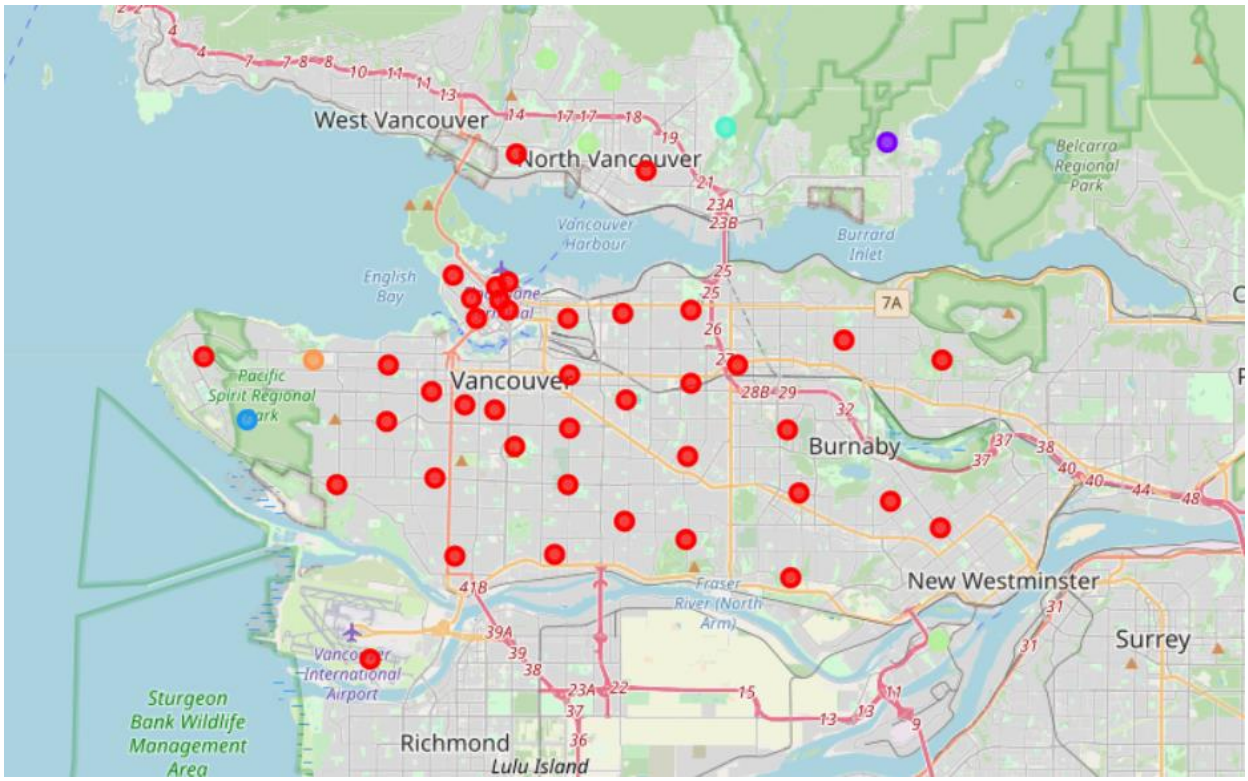
The Battle of Neighborhoods - Vancouver



Applied Data Science Capstone by IBM on Coursera

(Pooya Mirzabeygi)

1. INTRODUCTION: BUSINESS PROBLEM

The purpose of this project is to use machine learning techniques to explore neighborhoods of Vancouver Canada in order to propose a potential recommendation for the location of a new restaurant or a business office. This project would help new entrepreneurs moving to Vancouver Canada establishing their business in a premium spot.

The k-mean algorithm which is an un-supervised clustering technique is used to divide the neighborhoods into meaningful categories. So that, the best neighborhood is recommended for the new location of the new restaurant. The Foursquare API is used to acquire data about the venue in each neighborhood.

For visualization purposes, Folium visualization library is utilized to map the location of each neighborhood and the other neighborhood residing in the same category or other categories. This information can be used by various new businesses to locate a prime spot for their new restaurant, office, etc. The major stakeholders for this project are small business owners and planning to start their business at in Vancouver. This project would help them find the optimal location based on the category of their business such as:

- What's a best spot to open up a restaurant in Vancouver?
- What type of restaurant is most likely to be successful in different parts of the city?

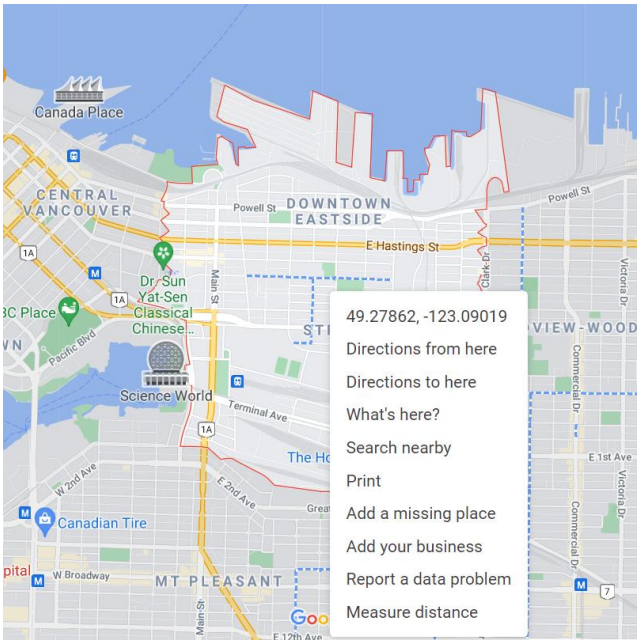
2. DATA

Vancouver has many neighborhoods. The Wikipedia website has the list of all postal codes starting with V. Only postal codes in and around Vancouver are selected amongst this data

"https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V"

	PostalCode	Borough	Neighborhood
0	V5A	Burnaby	Government Road, Lake City, SFU, Burnaby Mountain
1	V6A	Vancouver	Strathcona, Chinatown, Downtown Eastside
2	V5B	Burnaby	Parkcrest-Aubrey, Ardingley-Sprott
3	V6B	Vancouver	NE Downtown, Gastown, Harbour Centre, Internat...
4	V7B	Richmond	Sea Island, YVR

Unfortunately, the latitude and longitude values are missing for these postal codes in the website and geocode python module is very inaccurate. Therefore the latitude and longitude data has been procured from the google maps website manually and finally everything merged together.



	PostalCode	Latitude	Longitude
0	V5A	49.266519	-122.936557
1	V6A	49.277722	-123.090575
2	V5B	49.271882	-122.976632
3	V6B	49.279990	-123.115413
4	V7B	49.185816	-123.172296

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	V5A	Burnaby	Government Road, Lake City, SFU, Burnaby Mountain	49.266519	-122.936557
1	V6A	Vancouver	Strathcona, Chinatown, Downtown Eastside	49.277722	-123.090575
2	V5B	Burnaby	Parkcrest-Aubrey, Ardingley-Sprott	49.271882	-122.976632
3	V6B	Vancouver	NE Downtown, Gastown, Harbour Centre, Internat...	49.279990	-123.115413
4	V7B	Richmond	Sea Island, YVR	49.185816	-123.172296

Next, the venues information data for each neighborhood is extracted from the foursquare data as follows using the latitude and longitude:

https://foursquare.com/

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Government Road, Lake City, SFU, Burnaby Mountain	49.266519	-122.936557	Burnaby Mountain Golf Course	49.264878	-122.942871	Golf Course
1	Government Road, Lake City, SFU, Burnaby Mountain	49.266519	-122.936557	Clubhouse at Burnaby Mountain	49.264949	-122.943104	Burger Joint
2	Government Road, Lake City, SFU, Burnaby Mountain	49.266519	-122.936557	Burnaby Mountain Driving Range	49.263959	-122.942353	Golf Driving Range
3	Strathcona, Chinatown, Downtown Eastside	49.277722	-123.090575	Union Market	49.277371	-123.086989	Deli / Bodega
4	Strathcona, Chinatown, Downtown Eastside	49.277722	-123.090575	MacLean Park	49.278809	-123.088546	Park

3. METHODOLOGY

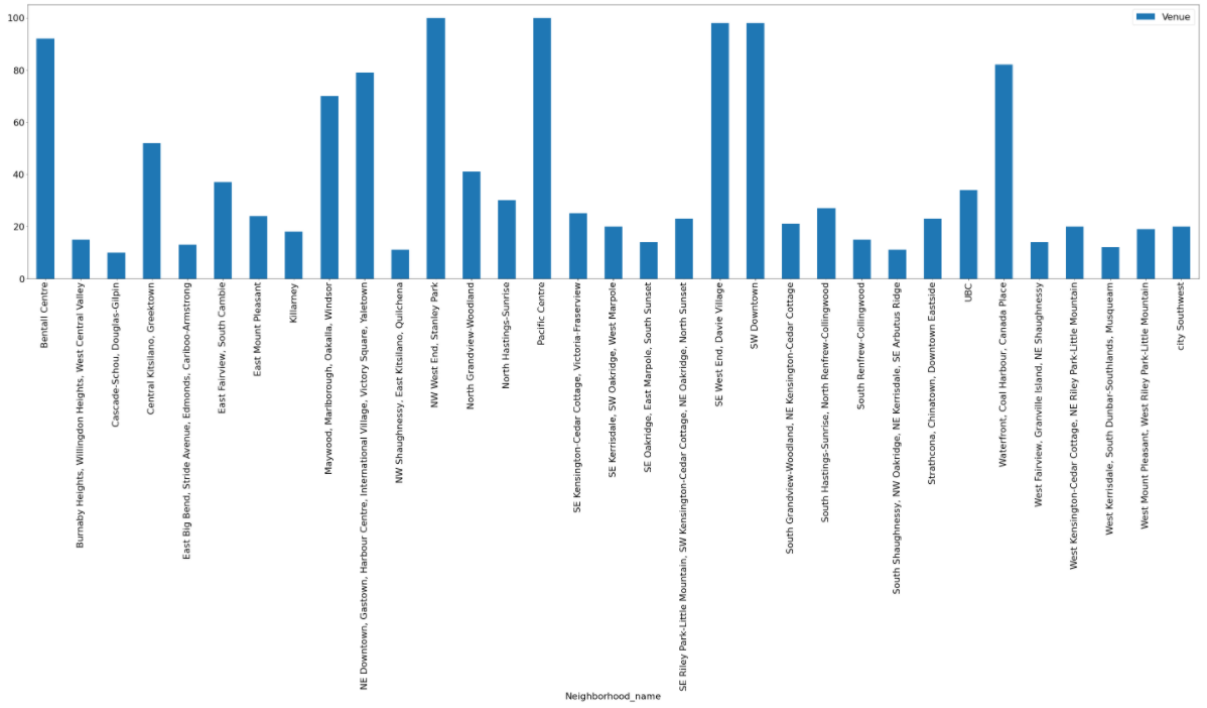
Now, we have the neighborhoods data of Vancouver (46 neighborhoods). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of 943 venues have been obtained in the whole city and 209 unique categories. But as seen we have multiple neighborhoods with less than 10 venues returned. In order to create a good analysis let's consider only the neighborhoods with more than 10 venues.

We can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighborhood. Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique have been used. The value of K is optimized to get the highest accuracy

The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

4. ANALYSIS

Looking into the dataset we found that there were many neighborhoods with less than 10 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 10 or more than 10 venues were obtained. The resultant dataset consists of 37 neighborhoods as shown below.



Filtered Neighborhood Dataset

Next, we will perform **one hot encoding** on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category. A sample output is shown below.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Auto Dealership	...	Thrift / Vintage Store	Toy / Game Store	Trade School	Udon Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Wine Shop
0	Bentall Centre	0.0	0.0	0.021739	0.0	0.0	0.01087	0.0	0.0	0.000000	...	0.0	0.010870	0.0	0.0	0.000000	0.0	0.000000	0.0
1	Burnaby Heights, Willingdon Heights, West Cent...	0.0	0.0	0.066667	0.0	0.0	0.00000	0.0	0.0	0.066667	...	0.0	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0
2	Cascade-Schou, Douglas-Gilpin	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.0	0.000000	...	0.0	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0
3	Central Kitsilano, Greektown	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.0	0.000000	...	0.0	0.019231	0.0	0.0	0.038462	0.0	0.019231	0.0
4	East Big Bend, Stride Avenue, Edmonds,	0.0	0.0	0.000000	0.0	0.0	0.00000	0.0	0.0	0.000000	...	0.0	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0

Mean of frequency of occurrence of each category

The above dataset is used to obtain the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence. A samplefor the first 5 neighborhoods is shown below:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bentall Centre	Hotel	Café	Dessert Shop	Food Truck	Coffee Shop	Plaza	Restaurant	Yoga Studio	Burger Joint	Cosmetics Shop
1	Burnaby Heights, Willingdon Heights, West Cent...	Bus Station	Design Studio	Auto Dealership	Sandwich Place	Deli / Bodega	Clothing Store	Motorcycle Shop	Restaurant	Hotel	Burger Joint
2	Cascade-Schou, Douglas-Gilpin	Chinese Restaurant	Food Court	Bus Stop	Electronics Store	Food & Drink Shop	Bookstore	Sandwich Place	Auto Garage	Snack Place	Accessories Store
3	Central Kitsilano, Greektown	Coffee Shop	Café	Pizza Place	Bank	Pub	Japanese Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant	Deli / Bodega
4	East Big Bend, Stride Avenue, Edmonds, Cariboo...	Restaurant	Chinese Restaurant	Convenience Store	Bar	Gas Station	Sandwich Place	Burger Joint	Pet Store	Eastern European Restaurant	Italian Restaurant

Ten Most Common Venues in each Neighborhood

This dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the bestvalue for K. Here, the silhouette score is used to find the best value for K. A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the value of 6 provides the best score. This K value is used for the K-Means Clustering Technique.

The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.

5. RESULTS

Let's examine the 6 clusters and find the discriminating venue categories that distinguish each cluster. For this purpose, lets also look into the five most common venue category in each cluster.

5.1. Cluster 1

The top venue categories in Cluster 1 are Bus stop and Chinese restaurants.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
17	0	SE Oakridge, East Marpole, South Sunset	Bus Stop	Chinese Restaurant	Indian Restaurant	Vietnamese Restaurant	Japanese Restaurant	Gas Station	Park	Sandwich Place	Motel	Fast Food Restaurant
24	0	South Shaughnessy, NW Oakridge, NE Kerrisdale,...	Chinese Restaurant	Bus Stop	Tea Room	Japanese Restaurant	Asian Restaurant	Electronics Store	Thrift / Vintage Store	Sushi Restaurant	Accessories Store	Motorcycle Shop

5.2. Cluster 2

The top venue categories in Cluster 2 are Restaurant and Asian restaurant

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	1	East Big Bend, Stride Avenue, Edmonds, Cariboo...	Restaurant	Chinese Restaurant	Convenience Store	Bar	Gas Station	Sandwich Place	Burger Joint	Pet Store	Eastern European Restaurant	Italian Restaurant
7	1	Killarney	Chinese Restaurant	Gas Station	Grocery Store	Pharmacy	Deli / Bodega	Sandwich Place	Farmers Market	Fast Food Restaurant	Coffee Shop	Shopping Mall
10	1	NW Shaughnessy, East Kitsilano, Quilchena	Restaurant	Café	Coffee Shop	Tennis Court	Art Gallery	Breakfast Spot	Cosmetics Shop	Chinese Restaurant	Grocery Store	Museum
16	1	SE Kerrisdale, SW Oakridge, West Marpole	Chinese Restaurant	Liquor Store	Noodle House	Sushi Restaurant	Coffee Shop	Massage Studio	Café	Thai Restaurant	Bar	Bank
18	1	SE Riley Park-Little Mountain, SW Kensington-C...	Chinese Restaurant	Asian Restaurant	Grocery Store	Dessert Shop	Baseball Field	Diner	Park	Sandwich Place	Field	Coffee Shop

5.3. Cluster 3

The top venue categories in Cluster 3 are Hotels and coffee shops.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	2	Bentall Centre	Hotel	Café	Dessert Shop	Food Truck	Coffee Shop	Plaza	Restaurant	Yoga Studio	Burger Joint	Cosmetics Shop
1	2	Burnaby Heights, Willingdon Heights, West Cent...	Bus Station	Design Studio	Auto Dealership	Sandwich Place	Deli / Bodega	Clothing Store	Motorcycle Shop	Restaurant	Hotel	Burger Joint
3	2	Central Kitsilano, Greektown	Coffee Shop	Café	Pizza Place	Bank	Pub	Japanese Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant	Deli / Bodega
5	2	East Fairview, South Cambie	Chinese Restaurant	Coffee Shop	Vietnamese Restaurant	Bank	Café	Bus Stop	Sushi Restaurant	Malay Restaurant	Movie Theater	Cantonese Restaurant
6	2	East Mount Pleasant	Vietnamese Restaurant	Ethiopian Restaurant	Sushi Restaurant	Yoga Studio	Liquor Store	Pizza Place	Outdoor Sculpture	Pub	Sandwich Place	Fast Food Restaurant
8	2	Maywood, Marlborough, Oakalla, Windsor	Coffee Shop	Clothing Store	Cosmetics Shop	Pharmacy	Bank	Japanese Restaurant	Ice Cream Shop	Sushi Restaurant	Department Store	Bookstore
9	2	NE Downtown, Gastown, Harbour Centre, Internat...	Hotel	Restaurant	Café	Coffee Shop	Seafood Restaurant	Taco Place	Bakery	Sandwich Place	Bar	Spa
11	2	NW West End, Stanley Park	Coffee Shop	Dessert Shop	Café	Japanese Restaurant	Grocery Store	Noodle House	Korean Restaurant	Ramen Restaurant	Pub	Vietnamese Restaurant
12	2	North Grandview-Woodland	Coffee Shop	Sushi Restaurant	Theater	Brewery	Café	Grocery Store	Asian Restaurant	Bakery	Chinese Restaurant	Breakfast Spot
13	2	North Hastings-Sunrise	Theme Park Ride / Attraction	Vietnamese Restaurant	Park	Theme Park	Beer Store	Liquor Store	Coffee Shop	Burger Joint	Sports Bar	Bridal Shop
14	2	Pacific Centre	Hotel	Food Truck	Dessert Shop	Restaurant	Seafood Restaurant	Japanese Restaurant	Cosmetics Shop	Steakhouse	Café	Coffee Shop
15	2	SE Kensington-Cedar Cottage, Victoria-Fraserview	Pizza Place	Fried Chicken Joint	Middle Eastern Restaurant	Gas Station	Motorcycle Shop	Bus Station	Bus Stop	Café	Fish Market	Sushi Restaurant

5.4. Cluster 4

The top venue categories in Cluster 4 are Grocery and liquor store.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
30	3	West Kerrisdale, South Dunbar-Southlands, Musq...	Grocery Store	Liquor Store	Golf Course	Café	Gym / Fitness Center	Coffee Shop	Gym	Pet Store	Japanese Restaurant	Music Store

5.5. Cluster 5

The top venue categories in Cluster 5 are Chinese restaurant and food court.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	4	Cascade-Schou, Douglas-Gilpin	Chinese Restaurant	Food Court	Bus Stop	Electronics Store	Food & Drink Shop	Bookstore	Sandwich Place	Auto Garage	Snack Place	Accessories Store

5.6. Cluster 6

The top venue categories in Cluster 6 are Bus Stop and Park.

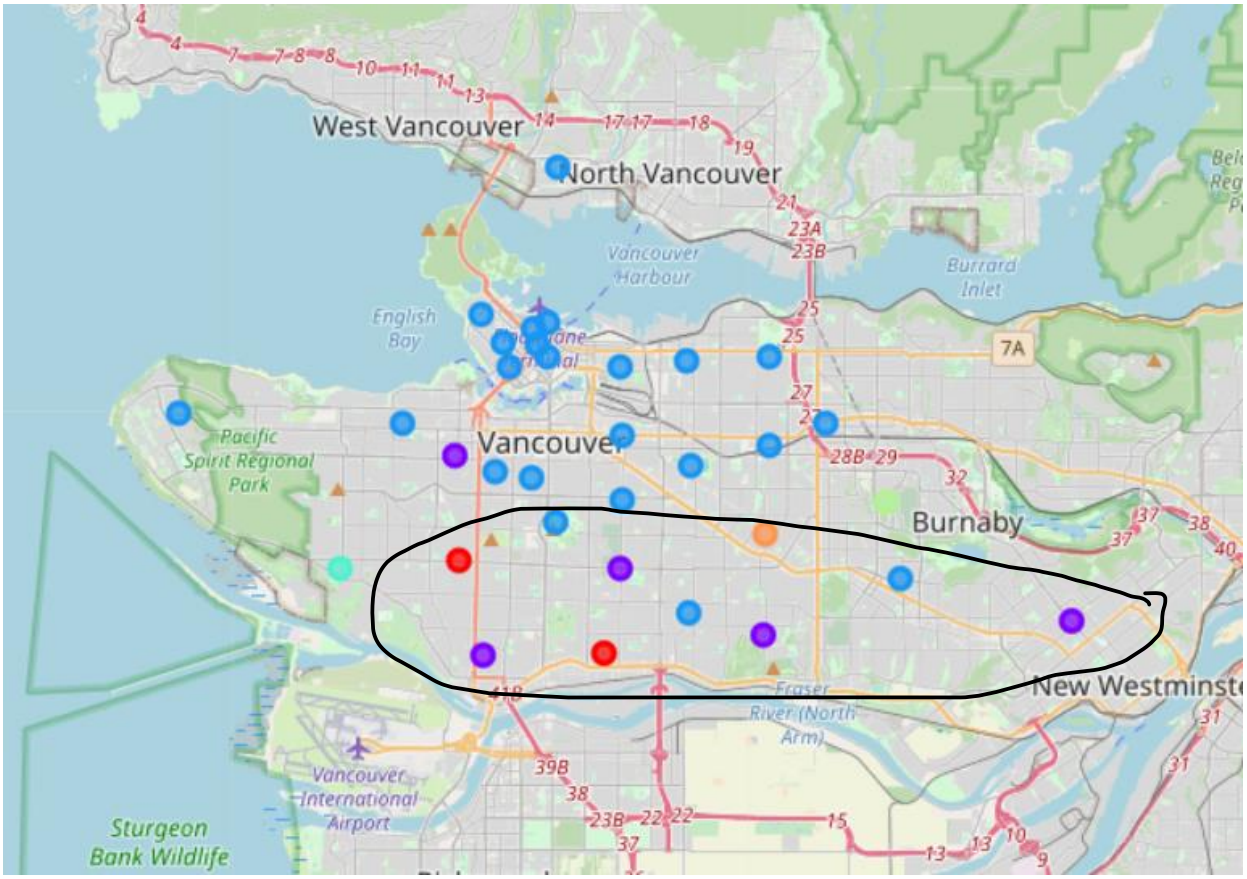
	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	5	South Renfrew-Collingwood	Bus Stop	Park	Ice Cream Shop	Bus Station	Hotel	Plaza	Asian Restaurant	Metro Station	Gift Shop	Fish & Chips Shop

6. DISCUSSION

Now let's use this data to provide the restaurant owner some recommendation to open restaurants or hotels.

1. Hotel

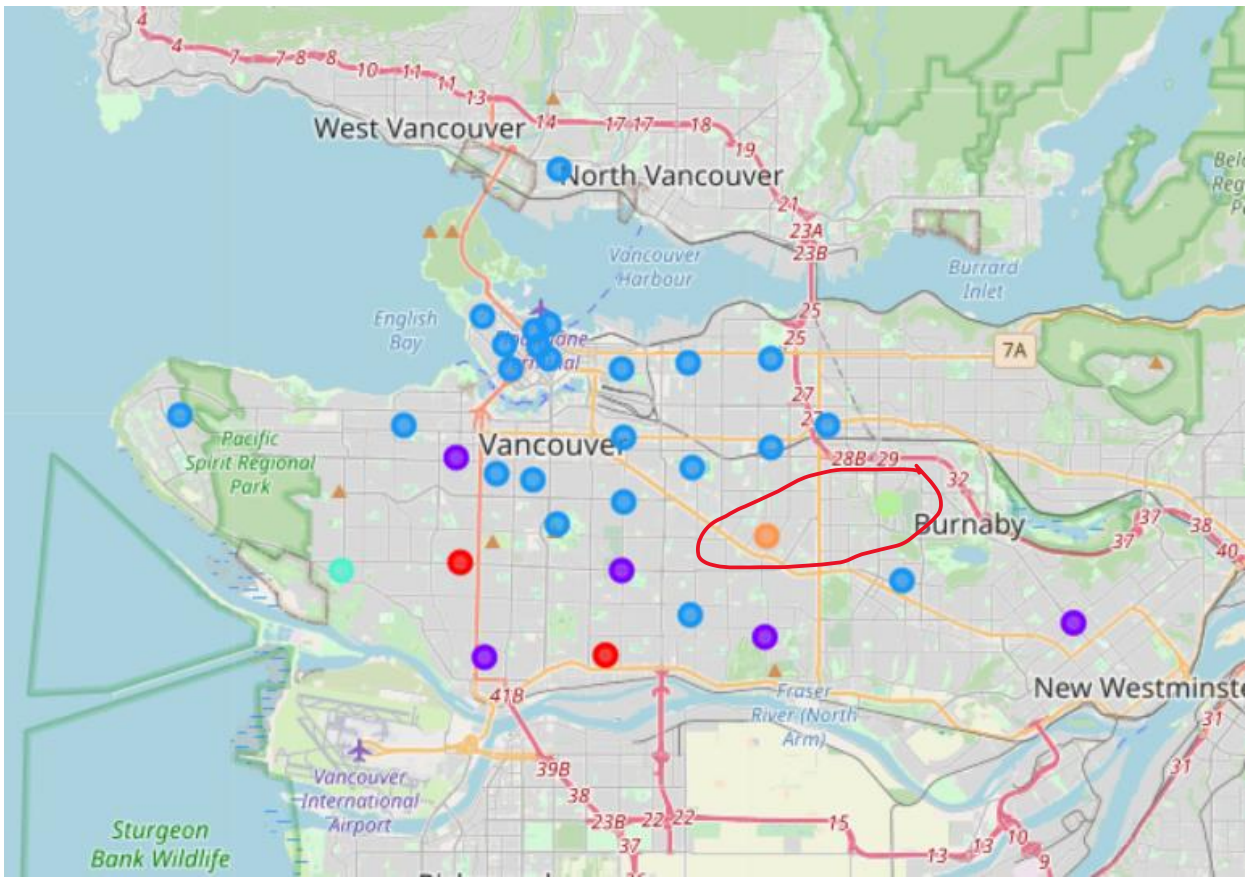
The neighborhoods in cluster 3 has the greatest number of hotels, hence opening one here is not the best choice. The optimal place would be one which has less hotels, but also have restaurants and other places to explore. Considering all these facts, Cluster 2 is the best option as shown below:



2. Restaurants

Clusters 1,2,3 and 5 has large number of Asian restaurants, so we recommend to the business owner if they're moving into these areas, they shouldn't think of any new Asian restaurants but to think about a new venue; it could be middle eastern or Mexican food as the prevalence of these type of restaurants are low.

Also they can be moving into cluster areas of 4 and 6 where the density of restaurants are very low and they can start thinking about starting a new Asian restaurant with less competition.



7. CONCLUSION

Purpose of this project was to analyze the neighborhoods of Vancouver and create a clustering model to suggest personals places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 10 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 86 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster. A few examples for the applications that the clusters can be used for have also beendiscussed. A map showing the clusters have been provided.

Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.