The Battle of Neighborhoods - Vancouver

Rafael Bastos

April 23, 2020

1. Introduction: Business Problem

Starting a business can be an extremely ardous task. When an individual decides to open a company, one faces a multitude of challenges and doubts, namely the type of business, the location, which costs are involved and whether your product will attract enough customers in order to your company to succeed, just to mention a few. Now imagine that a client is moving abroad and looks forward to open a branch of his company in a new country.

This project will discuss a case in which a stakeholder already has an expertise in the field of bike rental in their country of origin. The fact is that a member of the group is moving to Vancouver, Canada, and is observing a opportunity to expand the company abroad, therefore he is looking forward to open a branch of the bike rental company in the City of Vancouver. However, although this individual has the know-how of running such a business, he has insufficient knowledge about Vancouver's neighborhoods and its peculiarities. The stakeholder wants the firm to be opened in a location where it will have a higher chance of prospering. Moreover, the individual who will be running the branch wants to live in the vicinity of the company, if possible, so he can commute by foot or riding a bike.

The challenge of this project is to is to find the most suitable location, utilizing data science tools, for this individual to open his bike rental shop, considering that the chosen neighborhood should also be a good place to live.

2. Data

To solve the problem and find the perfect location for the business, the City of Vancouver will be segmented in neighborhoods acording to their postal codes. To do so, we will utilize the website geonames.org, which has a free and extensive geographical database, to extract information about the neighborhoods, such as their names, postal codes and geographic coordinates.

We can observe below a slice of the raw table extracted from the website, before being treated.

	N	Place	Code	Country	Admin1	Admin2	Admin3
0	1	Vancouver (Killarney)	V5S	Canada	British Columbia	Vancouver	
1		49.218/-123.038	None	None	None	None	None
2	2	Vancouver (North Hastings-Sunrise)	V5K	Canada	British Columbia	Vancouver	
3		49.281/-123.04	None	None	None	None	None
4	3	Vancouver (North Grandview-Woodlands)	V5L	Canada	British Columbia	Vancouver	

Table 1. Data extracted from geonames.org

The informations in the dataframe above are tangled, which means we need to clean e prepare the data to be suitable for our work.

After handling the data, this is a sample of the table we will start working with:

	PostalCode	Neighborhood	Latitude	Longitude
0	V5S	Vancouver (Killarney)	49.218	-123.038
1	V5K	Vancouver (North Hastings-Sunrise)	49.281	-123.04
2	V5L	Vancouver (North Grandview-Woodlands)	49.279	-123.067
3	V5P	Vancouver (SE Kensington / Victoria-Fraserview)	49.222	-123.068
4	V5R	Vancouver (South Renfrew-Collingwood)	49.24	-123.041

Table 2. Dataframe created after cleaning the data

After extracting and cleaning the data, we will be using a Python tool called **folium** to make an interactive map of the region, identifying all the neighborhoods and, in the following step, the Foursquare API will be utilized so we can determine the most common types of venues on each neighborhood.

After our data is ready, we use **folium** to generate the map of Vancouver with the neighborhoods.

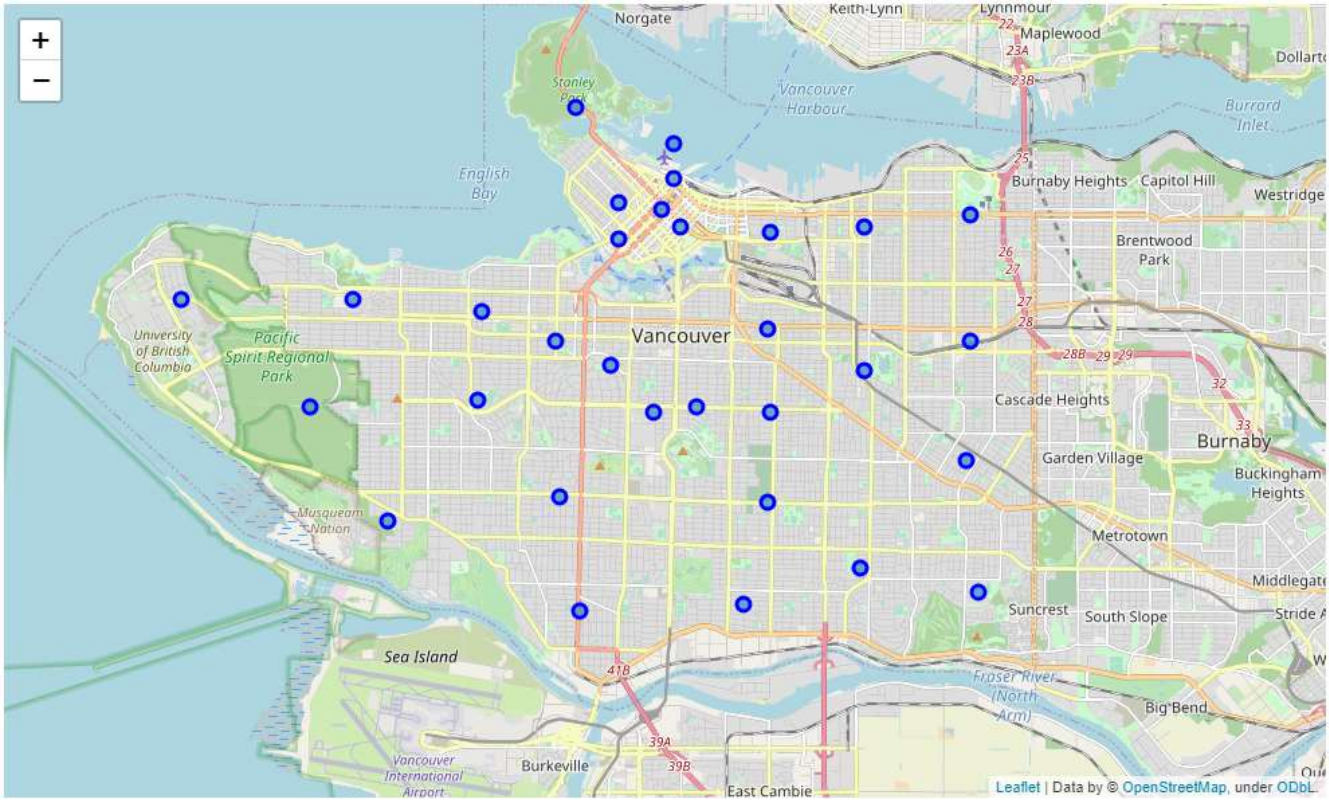

Figure 1. Map of Vancouver signalizing the neighborhoods

With the support of the data obtained from the **Foursquare API** we will be able to segment the neigborhoods into different clusters, according to their similarities, and identify a location that matches the stakeholder's requirements. We will be preferably looking for a neighborhood with enough parks, trails and easy to access, where the demand for bikes are probably higher than in other regions of the city. Furthermore, it would be desirable that such neighborhood is also a good place to live, with a considerable amount of restaurants, stores and other services nearby.

It is important to notice that our dataframe has 31 rows of postal codes and 4 columns (PostalCode, Neighborhood, Latitude and Logitude). Each postal code contains one up to four neighborhoods. Since

the neighborhoods combined in a postal code are adjacent, it is convenient to recognize them as a unity. Thus, we are considering 31 neighborhoods for this project.

3. Methodology

The focus of this project is to detect regions of Vancouver with abundance of parks and trails as well as plenty of services such as restaurants, cafés, bars and good transport system. To achieve our objective we will utilize the Foursquare API to search for the most common venues in each neighborhood. The venues will be identified to a neighborhood if they are located within 500 meters from the neighborhood coordinates.

First, we are testing our model exploring the first neighborhood in our dataframe. We will search for the top 100 venues located in the region, with the assistance of the Foursquare API, and analyze the results.

This is a sample of the JSON file generated from the Foursquare API data.


```
{'meta': {'code': 200, 'requestId': '5ea2063798205d7478cb4387'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Killarney',
  'headerFullLocation': 'Killarney, Vancouver',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 17,
  'suggestedBounds': {'ne': {'lat': 49.222500004500006,
    'lng': -123.03112351337492},
   'sw': {'lat': 49.2134999955, 'lng': -123.04487648662507}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4b636f14f964a5207f792ae3',
       'name': 'Champlain Square',
       'location': {'address': '7180 Kerr St.',
        'crossStreet': '@ 54th Ave.',
        'lat': 49.21877353130896,
        'lng': -123.04038966866126,
        'labeledLatLngs': [{'label': 'display',
          'lat': 49.21877353130896,
          'lng': -123.04038966866126}],
```


Figure 2. Slice of the data extracted from the Foursquare API

The JSON file will be cleaned and the data will be converted into a dataframe.

	name	categories	lat	lng
0	Champlain Square	Shopping Mall	49.218774	-123.040390
1	A&W	Fast Food Restaurant	49.219269	-123.040876
2	Kin's Farm Market	Farmers Market	49.219534	-123.040562
3	Sushi Go	Sushi Restaurant	49.219544	-123.041000
4	Subway	Sandwich Place	49.218948	-123.039908

Table 3. Sample of the top 100 venues at a neighborhood

Then, we are going to replicate this concept to all the neighborhoods in Vancouver. At this point we are able to inspect the quantity of venues found for each neighborhoods, limited to 100, and the amount of unique types of venues.

```
Neighborhood
 (Bentall Centre)                                                          9
 (Central Kitsilano)                                                      29
 (Chaldecutt / South University Endowment Lands)                           2
 (Dunbar-Southlands / Musqueam)                                            3
 (East Fairview / South Cambie)                                           20
 (East Mount Pleasant)                                                    25
 (Killarney)                                                              17
 (NE Downtown / Harbour Centre / Gastown / Yaletown)                      37
 (NW Arbutus Ridge)                                                        3
 (NW Shaughnessy / East Kitsilano / Quilchena)                            15
 (North Grandview-Woodlands)                                              37
 (North Hastings-Sunrise)                                                 25
 (North West End / Stanley Park)                                           4
 (Pacific Centre)                                                        100
 (SE Kensington / Victoria-Fraserview)                                    22
 (SE Kerrisdale / SW Oakridge / West Marpole)                             18
 (SE Oakridge / East Marpole / South Sunset)                               4
 (SE Riley Park-Little Mountain / SW Kensington / NE Oakridge / North Sunset)  20
 (SW Downtown)                                                            50
 (South Grandview-Woodlands / NE Kensington)                              15
 (South Hastings-Sunrise / North Renfrew-Collingwood)                     24
 (South Renfrew-Collingwood)                                               6
 (South Shaughnessy / NW Oakridge / NE Kerrisdale / SE Arbutus Ridge)      7
 (South West End)                                                        100
 (Strathcona / Chinatown / Downtown Eastside)                             15
 (UBC)                                                                    37
 (Waterfront / Coal Harbour / Canada Place)                               92
 (West Fairview / Granville Island / NE Shaughnessy)                       7
 (West Kensington / NE Riley Park-Little Mountain)                        29
 (West Kitsilano / Jericho)                                                1
 (West Mount Pleasant / West Riley Park-Little Mountain)                   5
Name: Venue, dtype: int64
```

Figure 3. Number of venues found per neighborhood, limited to 100

In the following step, we are using one-hot encoding to convert categorical data to numerical data, so our machine learning model can properly deal with the data. All neighborhoods will be analyzed and grouped taking the mean of the frequency of occurrence of each unique venue category. After this process, we are able to observe the most common categories of venues per neighborhood.

```
---- (Bentall Centre)----
                 venue  freq
0                Plaza  0.22
1    Airport Terminal  0.11
2  American Restaurant  0.11
3            Irish Pub  0.11
4            Gastropub  0.11


---- (Central Kitsilano)----
                           venue  freq
0                    Coffee Shop  0.10
1                          Café  0.07
2  Vegetarian / Vegan Restaurant  0.07
3                    Yoga Studio  0.03
4                  Farmers Market  0.03
```

Figure 4. Sample of the most common venues in each neigborhood per frequency

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	(Bentall Centre)	Plaza	Airport Terminal	American Restaurant	Outdoor Sculpture	Irish Pub
1	(Central Kitsilano)	Coffee Shop	Vegetarian / Vegan Restaurant	Café	Yoga Studio	Restaurant
2	(Chaldecutt / South University Endowment Lands)	Park	Yoga Studio	Falafel Restaurant	Food Court	Fish & Chips Shop
3	(Dunbar-Southlands / Musqueam)	Vietnamese Restaurant	Home Service	Fast Food Restaurant	Yoga Studio	Falafel Restaurant
4	(East Fairview / South Cambie)	Bus Stop	Coffee Shop	Chinese Restaurant	Grocery Store	Malay Restaurant

Table 4. Sample of the most common venue categories per neighborhood

Finally, the 31 neighborhoods will be segmented in 8 clusters utilizing the method **k-means**, which is an unsupervised machine learning algorithm that will group similar neighborhoods in a specific cluster. After all 8 clusters are established, a new dataframe containing the top 10 venue category per neighborhood will be created and we are able to plot a map and view the results. At this point we should have enough information in our hand to detect the optimal location for the stakeholder.

	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	V5S	(Killarney)	49.218	-123.038	1	Deli / Bodega	Coffee Shop	Shopping Mall	Fast Food Restaurant	Liquor Store
1	V5K	(North Hastings-Sunrise)	49.281	-123.04	1	Park	Event Space	Theme Park	Gas Station	Fair
2	V5L	(North Grandview-Woodlands)	49.279	-123.067	1	Asian Restaurant	Pizza Place	Sushi Restaurant	Café	Brewery

Table 5. Sample of the final dataframe with all the main informations

With all the information gathered we can generate a map showing the clustered neighborhoods

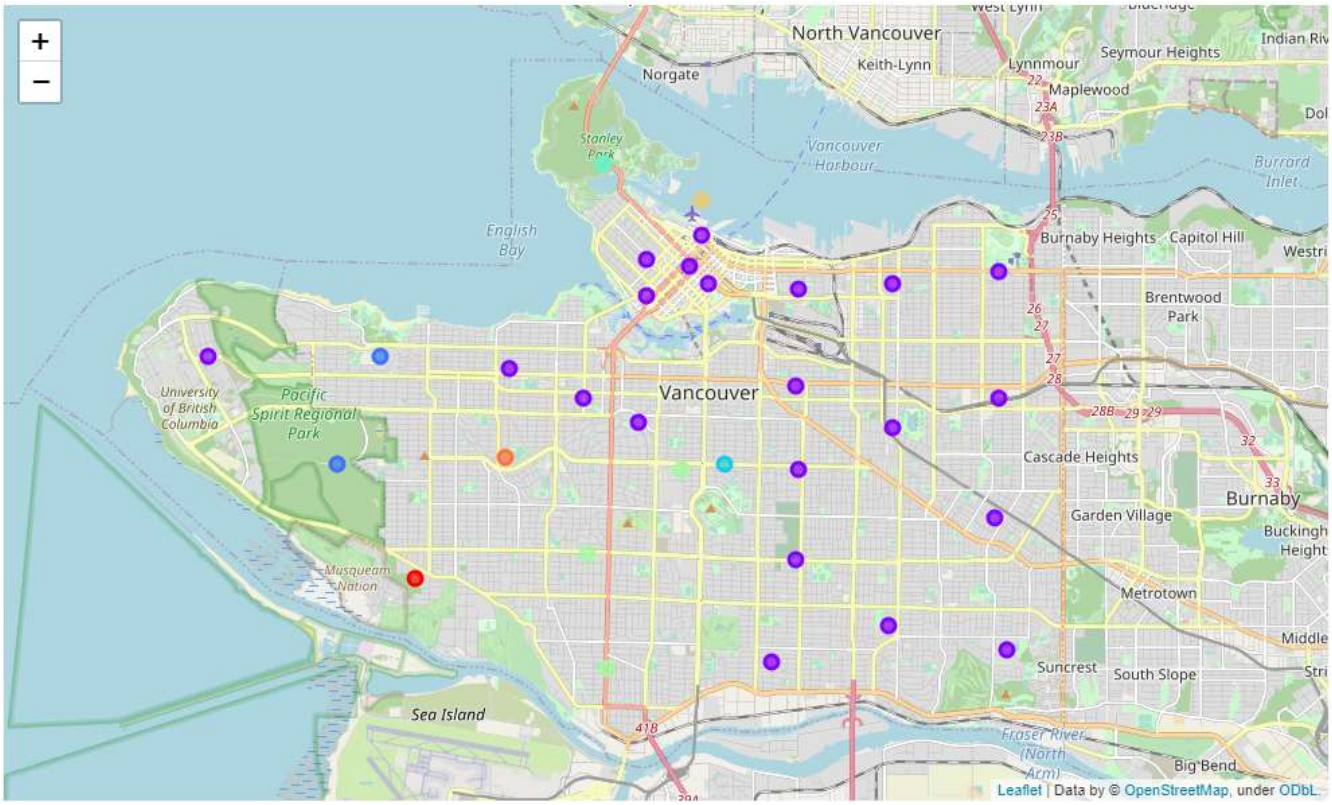

Figure 5. Map of Vancouver with its neighborhoods segmented in 8 clusters

The tables below represent the two cluster that better suited our demands

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
(West Kitsilano / Jericho)	Park	Yoga Studio	Falafel Restaurant	Food Court	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Field
(Chaldecutt / South University Endowment Lands)	Park	Yoga Studio	Falafel Restaurant	Food Court	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Field

Table 6. Sample of cluster 2

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
(North West End / Stanley Park)	Park	Outdoor Sculpture	Trail	Garden	Yoga Studio	Falafel Restaurant	Fish & Chips Shop	Financial or Legal Service

Table 7. Sample of cluster 4

4. Results and Discussion

After analyzing the results of our model, we detected that Vancouver has a multitude of outdoor venues that could be suitable for bikers. In many clusters we found parks, gardens and fields among the top 10 most common categories. Of course some specific clusters caught our attention, considering the characteristics we were searching for.

Of all eight clusters generated by the model, I can say that two of them stood out, namely clusters 2 and 4. Both clusters include neighborhoods surrounded by parks and trails and they would probably be a smart choice when deciding a location to open a bike rental company.

Trying to solve a possible dilemma between clusters the stakeholder might face, I will narrow the analysis between clusters 2 and 4. In addition to the positive aspects both regions presented, we can observe that cluster 4, which includes the neighborhoods North End West and Stanly Park, has parks, outdoor sculptures, trails, gardens and fields among the top 10 most common venues. These are great indicators that this region is one of the most suitable for biking, and as a consequence, for starting a

bike rental company. Furthermore, North West End and Stanley Park are located close to cluster 1, which covers some neighborhoods in the downtown region of the city, also containing some parks, as well as hotels, restaurants, cafés, stores, bus and metro stations. The region has a large variety of venues and is easily accessible via public transport which may attract both locals and tousrists. Finally, these are also clear signals that the region and its surroundings might be a good place to live, an important aspect for the shareholder.

5. Conclusion

The main purpose of this project was to identify the optimal neighborhood or region for a stakeholder to open a branch of a bike rental company in Vancouver. The idea was to search for locations with parks and trails nearby, with easy access and good infrastructure. Through data analysis and machine learning we could segment the City of Vancouver in clusters narrowing our focus on specific regions that matches the stakeholder's demands. The final decision should be made by the stakeholder, considering our recommendation in section 4 and based on other aspects of each region that may attract our client.