Battle of the Neighborhoods

Coffee Shop in Vancouver, BC, Canada

Author: Rafael Vidigal de Moraes

Date: April 25th, 2021

1. Introduction

Crime is a distress to people in big cities and Vancouver is no exception to that. People who are willing to open a commercial establishments always take that information in consideration to open a new venue. In order to propose a study to open a new Coffee Shop in the city of Vancouver, I consider to address this issue by analyzing the Vancouver Crime Data available at Kaggle and pursue to find the a safe borough with a plenty of people.

1.1 Geographical Location

Vancouver is a coastal city located in the Lower Mainland region in the province of British Columbia, western Canada. The greater Vancouver, also known as Metro Vancouver, is made up of 21 municipalities and the City of Vancouver is just one of them. In this project I will focus on Vancouver City data. Vancouver have a population of approximately 606.746 habitants as in 2018. Vancouver is the third lasgest city in Canada.

1.2 Problem

The objective of this project is to find a safe borough with plenty of people to open a commercial establishment in Vancouver, British Columbia, Canada. This project is targeted to stakeholds willing to open a Coffee Shop in Vancouver. The first task is to find the cluster with Coffee Shops and their competitors, then find how many coffee shops are per neighborhood and borough, cross that data with a thousand habitants per neighborhood. After that it is necessary to find a safe borough in Vancouver with an opportunity to open a Coffee Shop. Questions to be answered:

1. Who are the competitors ?

2. How many Coffee Shops per neighborhood?

3. How many Coffee Shops per neighborhood per a thousand habitants?

4. Which is the safest borough?

5. Which is the best borough to open a Coffee Shop?

1.3 Stakeholders

People interest in open a Coffee Shop in Vancouver. It also can be used by habitants or tourists to find the best cluster for Coffee Shop and its competitors.

2. Data

In order to fetch the crime deatils of Vancouver it was used a real world dataset that can be found at Kaggle datasets, this dataset can be found here. Although this dataset was filled with important information regarding the criminal activity, all coordinates data are offset and in some cases it was not disclosed in order to provide privacy protection, as declared by the City of Vancouver's

website. In order to fetch the Boroughs it was necessary to do it from Wikipedia. Geocorder API and Foursquare APi were used in order to fetch venues listed in the neighborhoods.

The population dataset was also ferched from Wikipedia, saved in a csv file and uploaded to GitHub..

Properties of the Crime Report

- TYPE - Crime Type
- YEAR - Redorded year (2018)
- Month - Recorded Month
- DAY - Recorded Day
- HOUR - Recorded Hour
- MINUTE - Recorded Minute
- HUNDRED_BLOCK - Recorded Block
- NEIGHBORHOOD - Recorded Neighborhood
- X - GPS longitude
- Y - GPS latitide

Wikipidia's data set didn't require webscraping. The page has information regarding Neighborhood and boroughs.

Open Cage API

- **Neighborhood**: Name of the neighborhood in the borough.
- **Borough**: Name of the borough.
- **Latitude**: Neighborhood Latitude.
- **Longitude**: Neighborhood Longitude.

The dataset from Kaggle was a heavy file and GitHub couldn't handle it. The Vancouver Crime Dataset had approximately 600.000 rows. Therefore the only crimes there were the crimes that took place in 2018, in order to reduce the dataset. It was created a csv file out of the dataset and it was uploaded to GitHub.

The dataset was shapped to look like the image #1, as follows below.

	Type	Year	Month	Day	Hour	Neighbourhood
0	Break and Enter Commercial	2018	3	2	6	West End
1	Break and Enter Commercial	2018	6	16	18	West End
2	Break and Enter Commercial	2018	12	12	0	West End
3	Break and Enter Commercial	2018	4	9	6	Central Business District
4	Break and Enter Commercial	2018	10	2	18	Central Business District

Image #1

In order to provide privacy prootection the coordinates were improper encoded. This information was declared by the City of Vancouver and can be found in its website and in a disclose on Kaggle. Because of the misguided information in columns X and Y I decided to drop them along with columns Month, Hour, Day and Year.

After cleaning the crime dataset, it was necessary to create a new dataset from Wikipedia,as mentioned in the Data section. This dataset will be merged with the crime dataset. The image #2 is the data generated on the information from Wikipedia.

	Neighbourhood	Borough
0	West End	Central
1	Central Business District	Central
2	Hastings-Sunrise	East Side
3	Grandview-Woodland	East Side
4	Mount Pleasant	East Side

Image #2

The next step was to count how many neighborhoods have each borough. As shown in the image #3, displayed below.

```
Borough
West Side          10
East Side           8
Central             3
South Vancouver     3
dtype: int64
```

Image #3

Then, it was necessary to read the population data. That data was acquired from Wikipedia, saved as csv file and uploaded to GitHub in order to download and read it on the notebook. The image #4 shows how the dataset looks like.

	Neighborhood	Borough	Population
0	West End	Central	44543
1	Central Business District	Central	54690
2	Hastings-Sunrise	East Side	33992
3	Grandview-Woodland	East Side	27297
4	Mount Pleasant	East Side	26400

Image #4

With this dataset, it was able to know how many habitants per borough.

```
Borough
Central             99256
East Side          255905
South Vancouver     60821
West Side          190764
Name: Population, dtype: int64
```

Image #5

Next step was to fetch the coordinates, latitude and longitude, in order to plot the neighborhoods on the map, which would give better visualization. It is necessary to create a dataframe just like the image #6, as follows.

	Neighbourhood	Borough	Latitude	Longitude
0	West End	Central	49.284131	-123.131795
1	Central Business District	Central	49.271409	-123.101259
2	Hastings-Sunrise	East Side	49.278714	-123.039998
3	Grandview-Woodland	East Side	49.275849	-123.066934
4	Mount Pleasant	East Side	49.264048	-123.096249
5	Strathcona	East Side	49.277693	-123.088539
6	Shaughnessy	West Side	49.246305	-123.138405
7	Sunset	East Side	49.219093	-123.091665
8	Fairview	West Side	49.261956	-123.130408

Image #6

3. Methodology

In order to find the right borough and eventually a neighborhood, it was necessary to explore the demographics of Vancouver neighborhoods. The previous data frame, Image #6, was necessary to visualize the neighborhoods.

After merging the dataframe to a clean and new daraset. This dataset was used to explore the Neighborhoods and discover the top 5 most common venues in those neighborhoods through Foursquare location plataform. It was possible to indentify that Coffee Shop and possible competitors were among them in most neighborhoods.

It was necessary to explore again those neighborhoods, but with a diferente approach. The scope of this Project was Coffee Shop, so it was crucial to explore the neighborhood looking for Coffee Shops and direct competitors. It was chosen: Coffee Shop, Café, Cafeteria, Tea Bubble Shop, Bakery, Breakfast Sopt, Dessert Shop and Fast Food Restaurant.

After acquiring those information, it was mandatory to cluster using the location of those Coffe Shops. The clustering was using k-means and to find the best k-value was used Elbow and Silhouette Methods.

In order to find the best borough to open a Coffee Shop it was unavoidable to analize how many coffee shops there were in each neighborhood and borough,

and to analize how many coffee shops there were in each borough/neighborhood per a Thousand habitants.

With those information, it was fundamental to explore the Crime dataset to see which neighborhood/borough would be safe, and indicate were it would be a great place to open a Coffee Shop.

4. Foursquare API

The Foursquare API was used to show the neighborhoods, along with folium library.

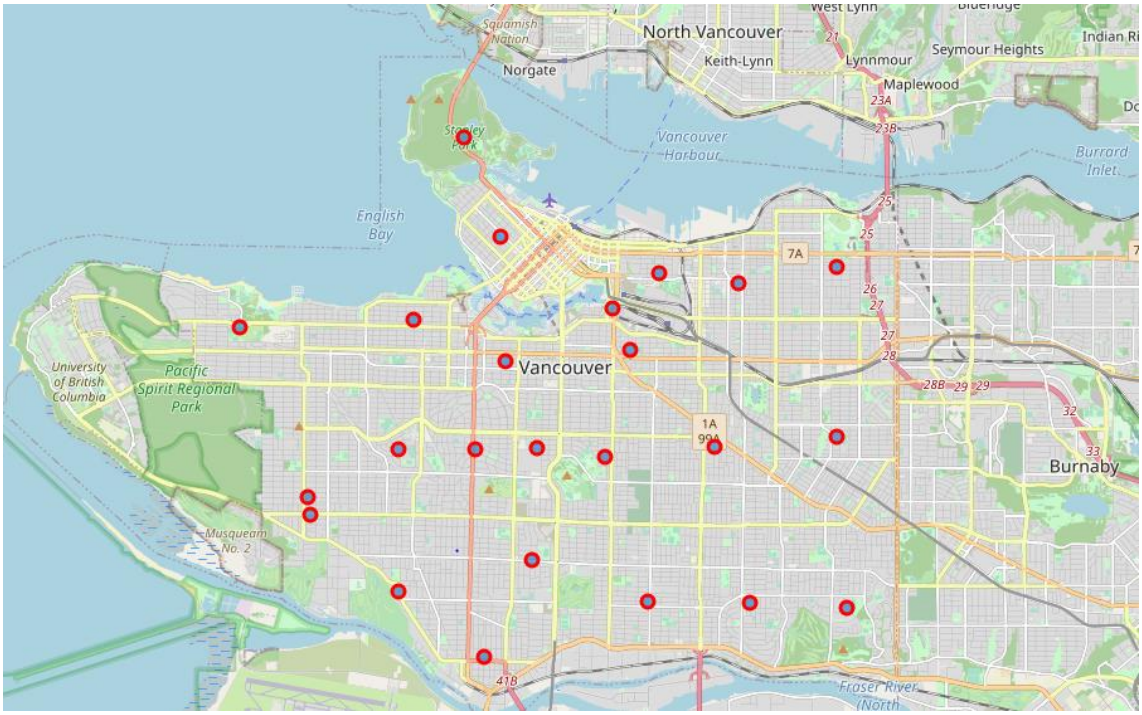

Image #7

After the neighborhoods were shown, the API was used to explore the area in a 2.500 radius in order to show the vênus in that area. The image below shows only the first 5 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	West End	49.284131	-123.131795	COBS Bread	49.281379	-123.132668	Bakery
1	West End	49.284131	-123.131795	Little Sister's Bookstore	49.282059	-123.134373	Bookstore
2	West End	49.284131	-123.131795	Ramen Danbo	49.287406	-123.129028	Ramen Restaurant
3	West End	49.284131	-123.131795	La Belle Patate	49.281977	-123.133623	Restaurant
4	West End	49.284131	-123.131795	Le Crocodile Restaurant	49.282658	-123.125287	French Restaurant

Image #8

One hot encoding was done on the venues data. One hot encoding is a process to convert categorical variables into a form that the Machine Learning algorithms will do better in prediction. After this the data was grouped by neighborhood, the mean was calcuated.

	Neighborhood	Accessories Store	American Restaurant	Amphitheater	Aquarium	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	BBQ Joint	Bagel Shop	Bakery	Bank	
0	Arbutus Ridge	0.0	0.020000	0.000000	0.0	0.0	0.000000	0.010000	0.0	0.000000	0.0	0.00	0.060000	0.010000	0.(
1	Central Business District	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.010000	0.0	0.010000	0.0	0.00	0.040000	0.000000	0.(
2	Dunbar-Southlands	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.00	0.034483	0.034483	0.(
3	Fairview	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.02	0.100000	0.000000	0.(
4	Grandview-Woodland	0.0	0.012346	0.012346	0.0	0.0	0.012346	0.024691	0.0	0.012346	0.0	0.00	0.000000	0.000000	0.(

Image #9

The Foursquare API was also used to explore the top 5 popular venues in each neighborhood in a radius of 2.500 meters. That returned a response in a json format containing those venues and in each neighborhoud, it was converted to a pandas dataframe, as follows.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Arbutus Ridge	Coffee Shop	Bakery	Café	Japanese Restaurant	Pharmacy
1	Central Business District	Coffee Shop	Restaurant	Taco Place	Brewery	Hotel
2	Dunbar-Southlands	Coffee Shop	Grocery Store	Italian Restaurant	Pharmacy	Bakery
3	Fairview	Bakery	Park	Coffee Shop	Restaurant	Café
4	Grandview-Woodland	Park	Brewery	Coffee Shop	Café	Mexican Restaurant

Image #10

5. Cluster Analysis

As shown in the image #10, Coffee Shop is among the top 5 venues. So it was decided to explore again the area, but looking for Coffee Shops and direct competitors in order to see were would be the best spot the open a new one.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	West End	49.284131	-123.131795	COBS Bread	49.281379	-123.132668	Bakery
8	West End	49.284131	-123.131795	Greenhorn Cafe	49.286628	-123.134038	Café
10	West End	49.284131	-123.131795	Breka Bakery & Café	49.285854	-123.127156	Bakery
15	West End	49.284131	-123.131795	Rocky Mountain Chocolate Factory	49.283626	-123.123306	Dessert Shop
22	West End	49.284131	-123.131795	PappaRoti	49.288716	-123.130916	Coffee Shop
33	West End	49.284131	-123.131795	Thierry Chocolaterie Patisserie	49.284892	-123.122951	Dessert Shop
55	West End	49.284131	-123.131795	Nero Belgian Waffle Bar	49.278451	-123.122024	Dessert Shop
62	West End	49.284131	-123.131795	Nero Belgian Waffle Bar	49.290543	-123.133905	Dessert Shop
66	West End	49.284131	-123.131795	Breka Bakery & Cafe	49.278496	-123.128062	Bakery
74	West End	49.284131	-123.131795	Delany's Coffee House	49.288195	-123.140433	Coffee Shop
77	West End	49.284131	-123.131795	Soirette Macarons & Tea	49.289654	-123.127993	Café

Image #11

In order to find similar clusters it was used the k-means method, which is a form of unsupervised machine learning algorithm that clusters data based on the predefined number of clusters. As shown below, it was utilized 8 clusters, this number was obtained after Elbow and Silhoutte Methods were applied.

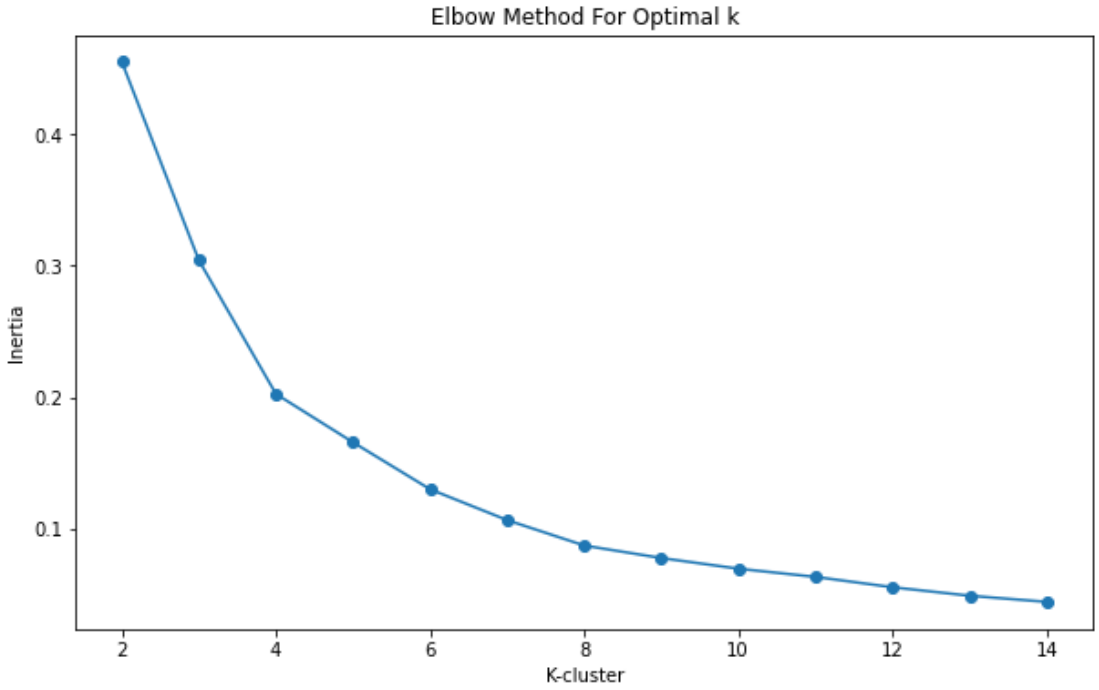

Image #12

```
N_cluster: 2, score: 0.4081282635647124
N_cluster: 3, score: 0.40381596664684355
N_cluster: 4, score: 0.4586725712185001
N_cluster: 5, score: 0.40909285301191994
N_cluster: 6, score: 0.4335069191844513
N_cluster: 7, score: 0.4575461278366868
N_cluster: 8, score: 0.46187968785821837
N_cluster: 9, score: 0.4576368537307136
```

Image #13

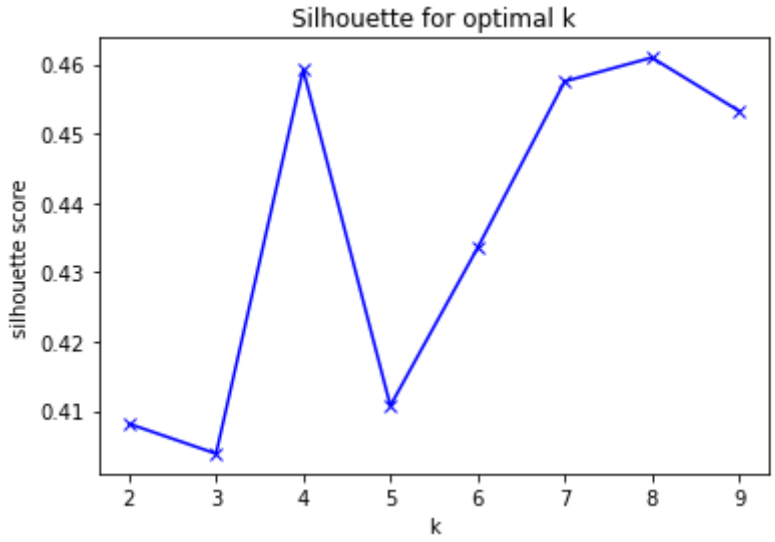

Image #14

The image below shows the 8 cluster in Vancouver.

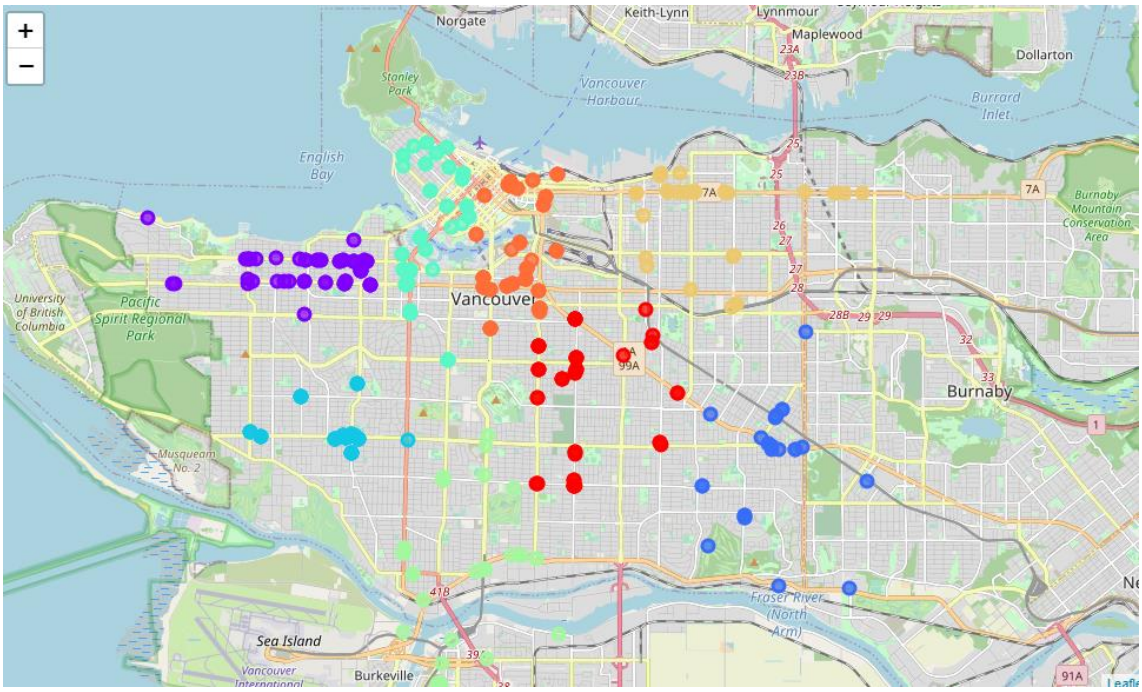

Image #15

A bar plot was created to show that Coffee Shop is the dominant venue among its competitors.

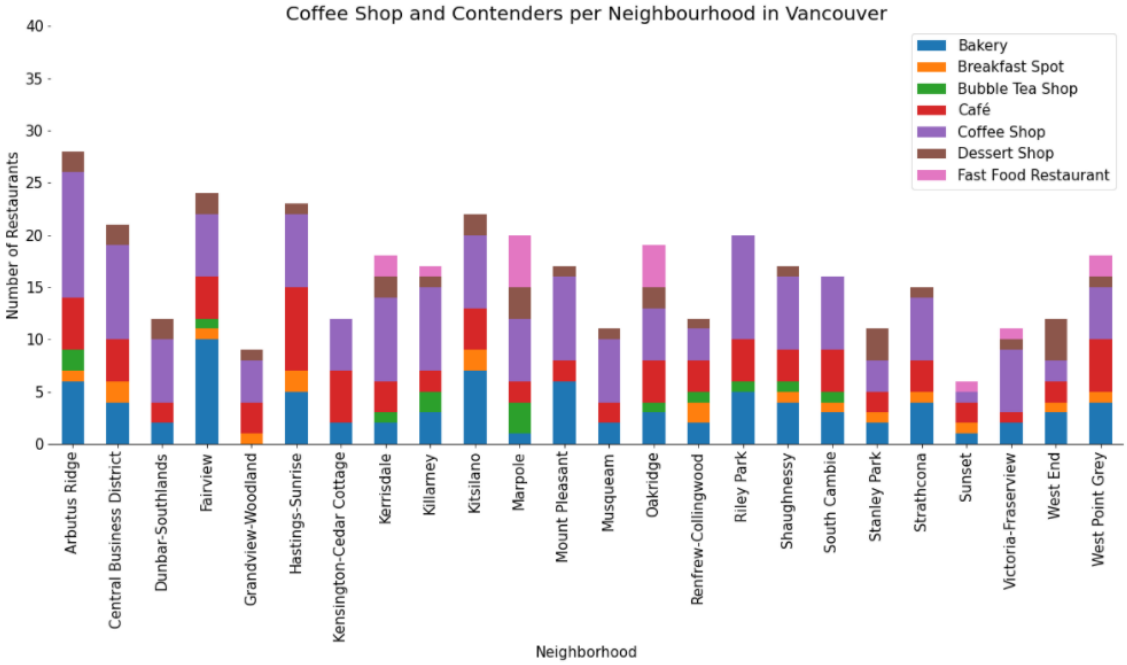

Image #16

Another bar plot was created to show the number of Coffee Shops per Neighborhood.

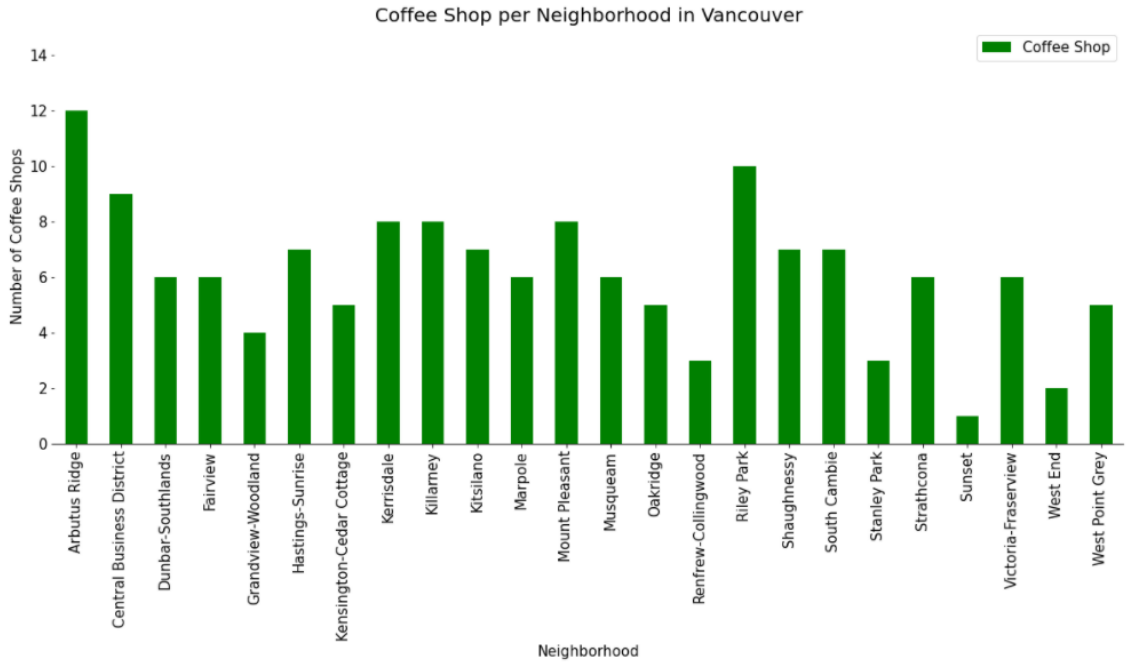

Image #17

This dataset was crossed with the population dataset in order to see how many coffee shops there are per a thousand habitants. This information would help to decide in which neighborhood is better.

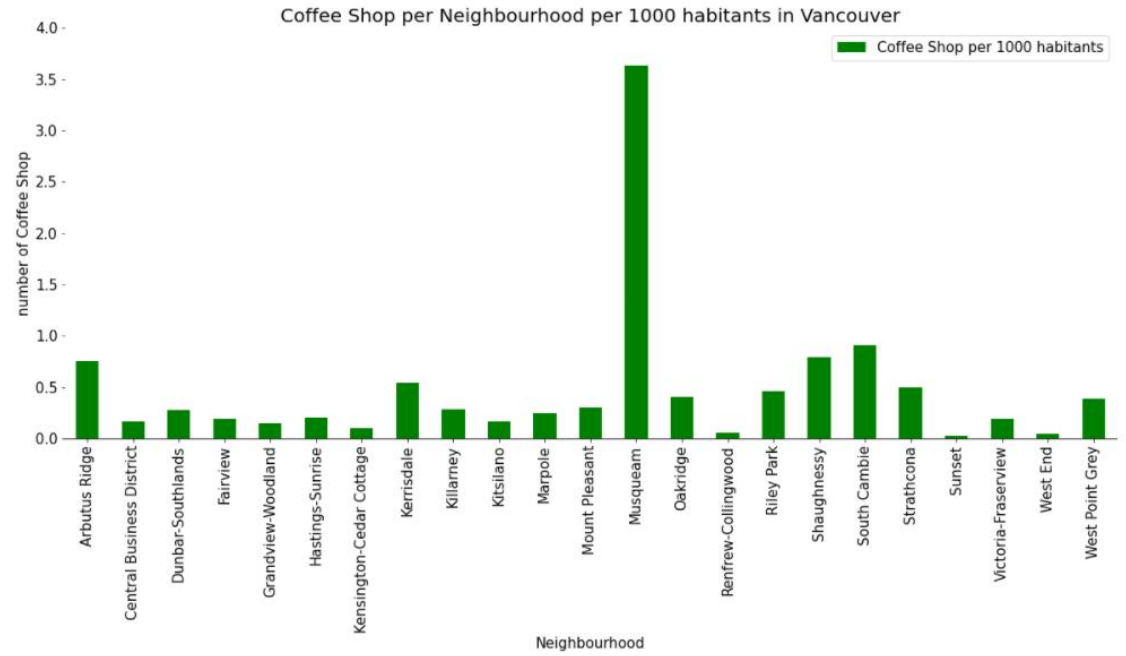

Image #18

In order to decide where to open a Coffee Shop it was mandatory to check the crime data set. The crime data set would help to choose a safe borough. Therefore the crime dataset was turned into a pivot table to show the sum of types of crime in each borough.

index	Borough	Break and Enter Commercial	Break and Enter Residential/Other	Mischief	Other Theft	Theft from Vehicle	YearTheft of Bicycle	Theft of Vehicle	YearVehicle Collision or Pedestrian Struck (with Fatality)	Vehicle Collision or Pedestrian Struck (with Injury)	Total
0	Central	787	198	2280	2489	6871	857	245	1	314	14042
1	East Side	786	1043	2192	1674	4754	678	605	8	660	12400
2	South Vancouver	49	156	187	88	483	36	71	1	111	1182
3	West Side	403	1000	1062	696	2838	588	225	3	389	7204

Image #19

A bar plot was created to help visualize the information above. It shows that the South Vancouver borough has the lowest crimes.

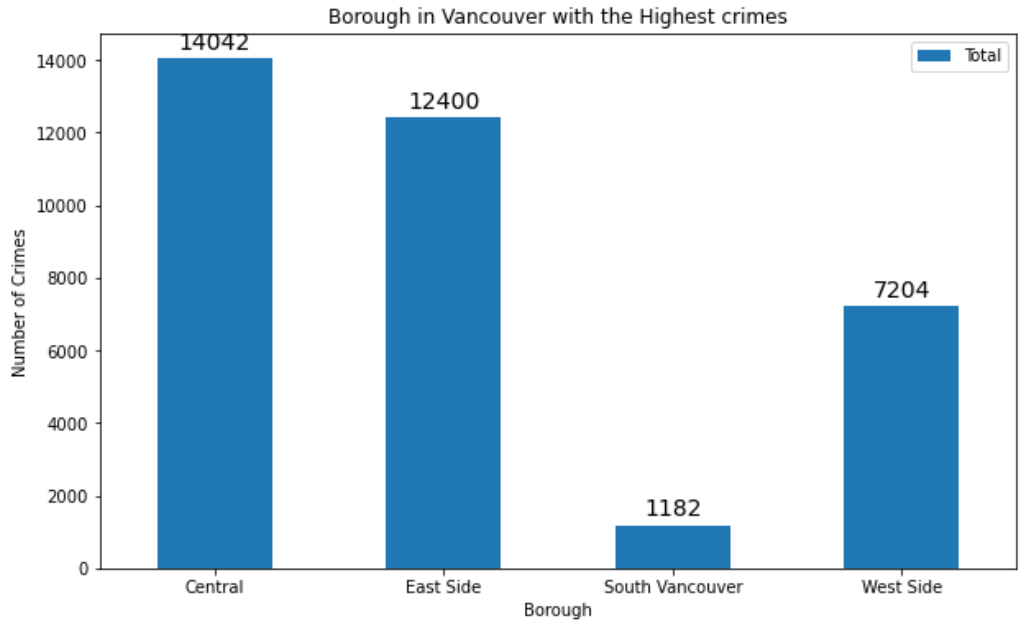

Image #20

6. Conclusion

More information were required to get a conclusion. Therefore a few bar plots were created to show how many Coffee Shops per Borough, how many habitants per Borough, and how many Coffee Shops per Borough per a thousand habitants.

Those information is requeired because the stakeholder need to know which borough has low crimes, what type of crimes, and they need the information regarding Coffee Shops and habitants.

Those information will be displayed below.

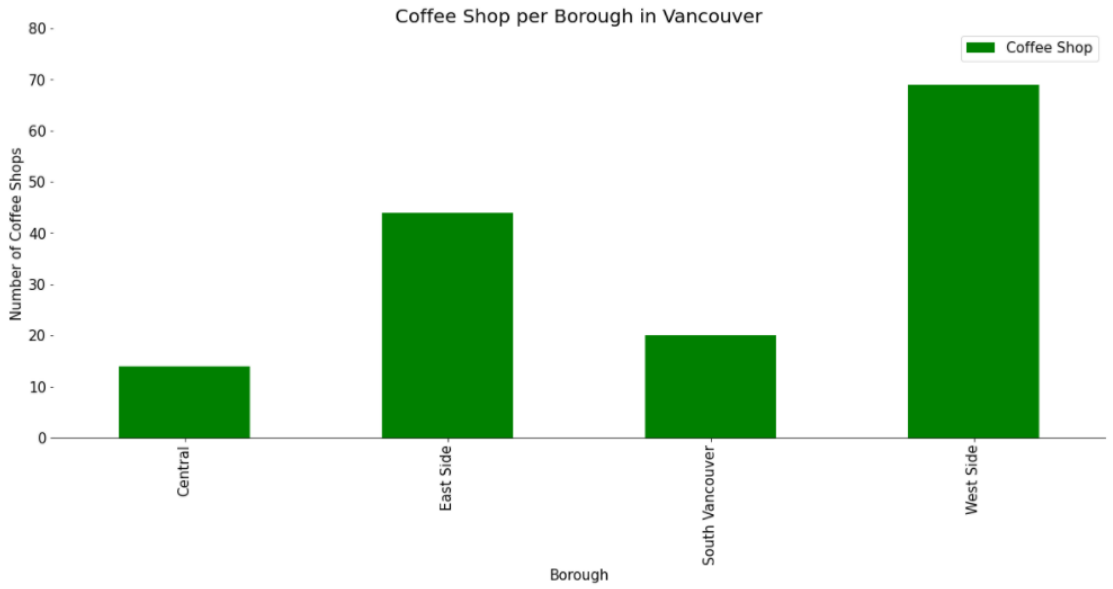

Image #21

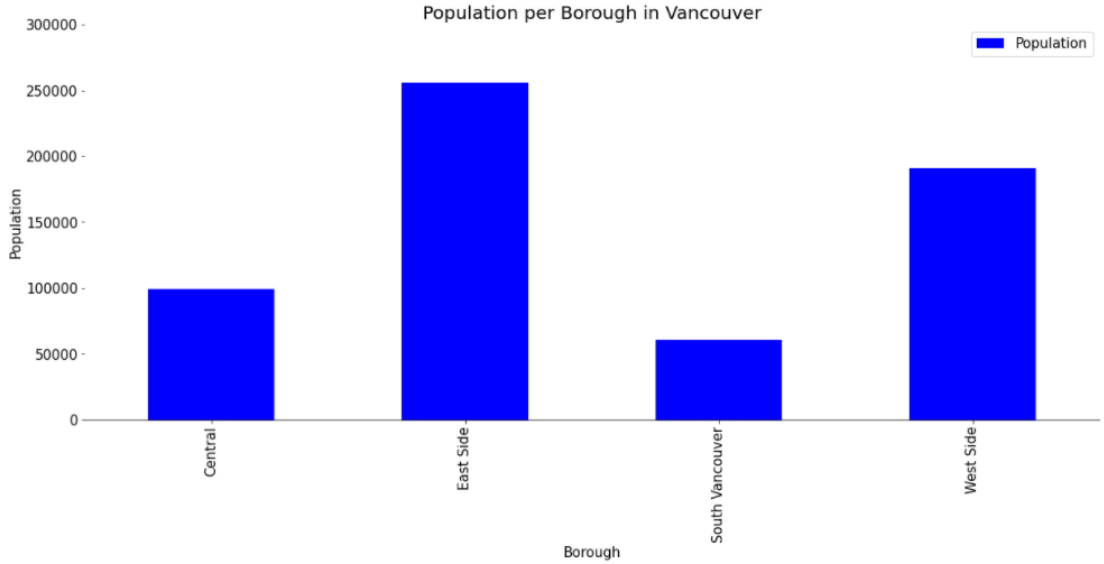

Image #22

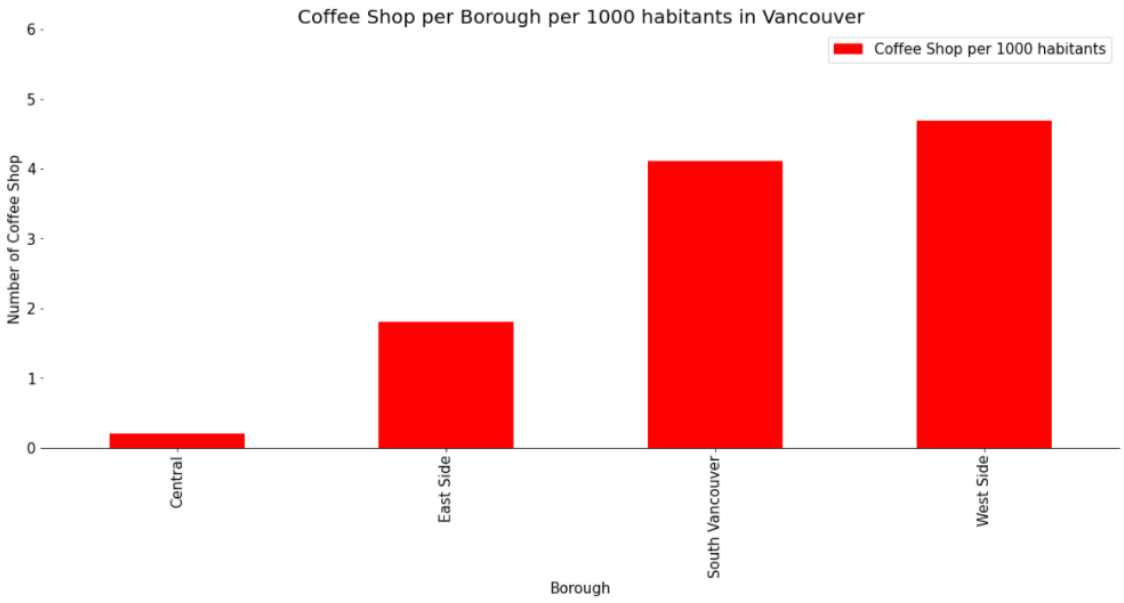

Image #23

Based on exploratory data analysis of the crimes per Borough, image #20, we can see that South Vancouver has the lowest crimes. Since South Vancouver has very little number of neighborhoods, only 3, and the number of habitants is 60.821. It is the smallest Cluster. When you compare Coffee Shop per Borough x 1000 Habitants you can see they are in second. Therefore the South Vancouver Cluster, Cluster N.3 is not a good option.

The next borough with lowest crime is West Side. West Side borough is the second borough with more habitants, 190.764 people live there. It is the first Borough in number of Coffee Shops per 1000 Habitants. Even though they have approximately 5 Coffee Shops per a thousand habitants the borough still have opportunities to have more Coffee Shops, as you can see in the Map with Clusters, between cluster N.2 and N.4, specially between Neighbourdhoods Dunbar-Southlands and Musqueam, there is a space with no venues related to Coffee Shops, Café, Cafeteria, Bubble Tea and other competitors, as shown in the image below. The largest Univeristy in the West Coast of Canada, UBC (University of British Columbia), is in the West Side Borough, also this borough has the richest neighbouhoods in Vancouver. Therefore, I see an opportunity to open a Coffee Shop there.

Last but not least, different types of crimes were recorded in the West Side Borough. Crimes like Breank and Enter Commercial is low among other types of crimes.

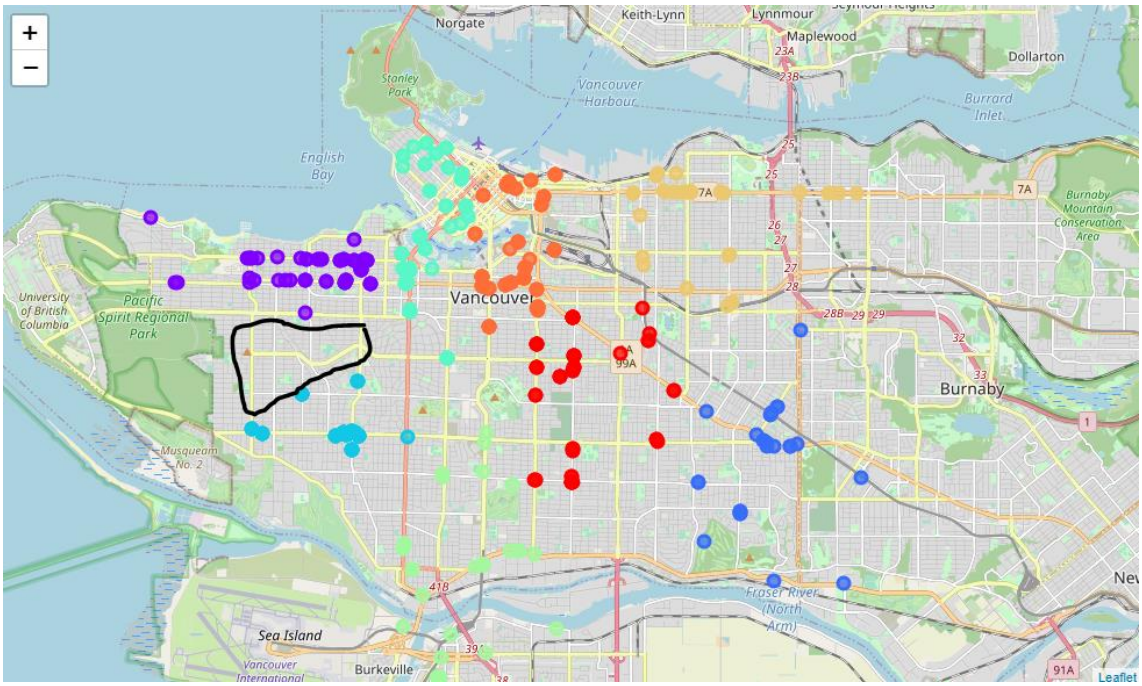

Image #24

7. Recommendation for future study

Even though it was possible to map the Coffee Shops and its competitors there are a lot of Coffee Shops that were not shown or located. Possibly because of the many limitations the API holds. Another API could help.

The crime data was another limitation, because of the privacy policies the coordinates were wrong and it made impossible to show on the map were the crimes occurred, that would help to choose a better place in that borough.

In order to overcome future incovenients such like those faced, another API could help find those missing Coffee Shops and reliable crime data would make a lot easier to identify to best borough.

8. References

- “Vancouver Crime Data”, Kaggle data;
- “List of Boroughs”, Wikipedia;
- “Vancouver Population per Neighborhood”, Wikipedia;
- “Foursquare API”, Foursquare;
- “IBM Data Science Notes”; Coursera