APPLIED DATA SCIENCE CAPSTONE PROJECT REPORT

MADHUMITA DAS

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Vancouver is a major city in western Canada, located in the Lower Mainland region of British Columbia. As the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada. Vancouver has the highest population density in Canada, with over 5,400 people per square kilometer. The **objective** of this Capstone project is to propose to the stakeholders a safe place to start a mini-grocery store business venture in the city of Vancouver. Opening a mini-grocery store can be a lucrative business venture if it can be opened in a secure place with less crime and competition. When people are looking for specialty foods or ingredients that can't be found at the corner store or neighborhood supermarket, they typically head to small grocers. Such retail establishments sell food and items that are uncommon or not carried by bigger stores. The problem will be approached in two phases: First, the safe borough has been selected for opening the store by analyzing the crime data of Vancouver city neighborhoods. Second, the data science tools learned in the course, has been used to explore the neighborhood of the safest borough and the ten most common venues in each neighborhood. Based on the results of most common venues, proposal has been placed to the stakeholders as to in which neighborhood the store can be opened.

1.2 TARGET AUDIENCE

This project will be of interest to the stakeholders who would like to invest in a mini-grocery store business in the Vancouver City and would like to find out a safe and secure place with less competition to start with.

2. DATA

2.1 DATA SOURCE

The data sources that have been used are:

I. Vancouver City Crime data from Kaggle: Since the data set is huge, for the purpose of this project, I have considered only the 2019 crime data from Here is the source:

II. Further data has been scraped from Wikipedia to gather information on Boroughs in Vancouver: Vancouver Boroughs Data

III. A consolidated data set will be created of neighborhoods, boroughs, coordinates and the crime data gathered before.

IV. Foursquare API has been used to fetch that data and to find the most common venues in each neighborhood.

V. Machine learning algorithm to be used to cluster the neighborhoods and finally select the best neighborhood to open the store.

2.2 DATA CLEANING

Data is read from the Vancouver Crime Data Set from Kaggle. Since it is a huge data set, I have filtered the 2019 Crime data and worked on it. A snapshot of the data in the Pandas dataframe is shown below.

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Break and Enter Commercial	2019	3	7	2	6	10XX SITKA SQ	Fairview	490612.9648	5457109.822
1	Break and Enter Commercial	2019	8	27	4	12	10XX ALBERNI ST	West End	491007.7798	5459174.338
2	Break and Enter Commercial	2019	6	9	10	58	10XX BEACH AVE	West End	490232.6157	5458203.356
3	Break and Enter Commercial	2019	1	6	1	36	10XX BEACH AVE	West End	490234.4136	5458201.015
4	Break and Enter Commercial	2019	7	21	11	21	10XX BEACH AVE	Central Business District	490249.2307	5458166.833

For the purpose of this project, the MINUTE, HUNDRED_BLOCK, X, Y columns have been dropped. Snapshot of the data frame below:

	TYPE	YEAR	MONTH	DAY	HOUR	NEIGHBOURHOOD
0	Break and Enter Commercial	2019	3	7	2	Fairview
1	Break and Enter Commercial	2019	8	27	4	West End
2	Break and Enter Commercial	2019	6	9	10	West End
3	Break and Enter Commercial	2019	1	6	1	West End
4	Break and Enter Commercial	2019	7	21	11	Central Business District

The uppercase column names have been changed into lowercase in the next step.

	Type	Year	Month	Day	Hour	Neighbourhood
0	Break and Enter Commercial	2019	3	7	2	Fairview
1	Break and Enter Commercial	2019	8	27	4	West End
2	Break and Enter Commercial	2019	6	9	10	West End
3	Break and Enter Commercial	2019	1	6	1	West End
4	Break and Enter Commercial	2019	7	21	11	Central Business District

The second data set is fetched from Wikipedia as mentioned above. The objective is to find out which neighborhood belongs to which borough. Snapshot below:

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

3. METHODOLOGY

Exploratory data analysis has been carried out to find the borough with the lowest crime rate. The total number of crimes in each neighborhood has been calculated. Snapshot below:

```
Central Business District    10618
West End                      2864
Mount Pleasant                2219
Strathcona                    1928
Fairview                      1827
Grandview-Woodland            1613
Renfrew-Collingwood           1503
Kitsilano                     1426
Kensington-Cedar Cottage      1363
Hastings-Sunrise              1229
Sunset                         893
Riley Park                     828
Marpole                        771
Victoria-Fraserview            530
Killarney                      507
Oakridge                       491
Kerrisdale                     382
South Cambie                   367
Dunbar-Southlands              329
West Point Grey                323
Shaughnessy                    266
Arbutus Ridge                  263
Stanley Park                   167
Musqueam                        10
Name: Neighbourhood, dtype: int64
```

The Crime data table and the boroughs table have been merged and used to calculate the total number of crimes in each borough as shown below:

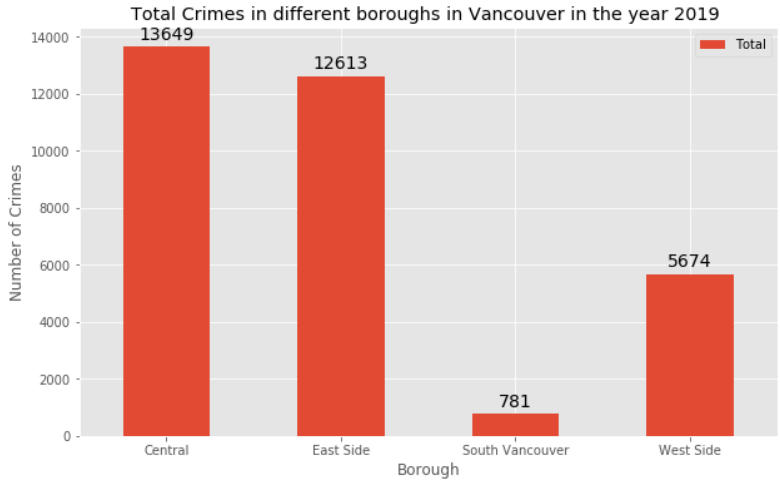
	Type	Year	Month	Day	Hour	Neighbourhood	Borough
0	Break and Enter Commercial	2019	3	7	2	Fairview	West Side
1	Break and Enter Commercial	2019	4	21	16	Fairview	West Side
2	Break and Enter Commercial	2019	10	26	0	Fairview	West Side
3	Break and Enter Commercial	2019	3	27	8	Fairview	West Side
4	Break and Enter Commercial	2019	7	13	1	Fairview	West Side

```
Central            13649
East Side          12613
West Side           5674
South Vancouver      781
Name: Borough, dtype: int64
```

Next, the different crime types in each borough has been obtained as shown below:

Type	Year Break and Enter Commercial	Break and Enter Residential/Other	Mischief	Other Theft	Theft from Vehicle	Theft of Bicycle	Theft of Vehicle	Vehicle Collision or Pedestrian Struck (with Fatality)	Vehicle Collision or Pedestrian Struck (with Injury)	All
Borough										
Central	774	192	2003	2473	7041	696	251	1	218	13649
East Side	787	961	2141	1574	5215	745	714	8	468	12613
South Vancouver	56	104	96	118	290	35	33	1	48	781
West Side	347	697	749	701	2273	506	202	2	197	5674
All	1964	1954	4989	4866	14819	1982	1200	12	931	32717

A Bar graph has been created for the visual representation of the total crimes in each borough.



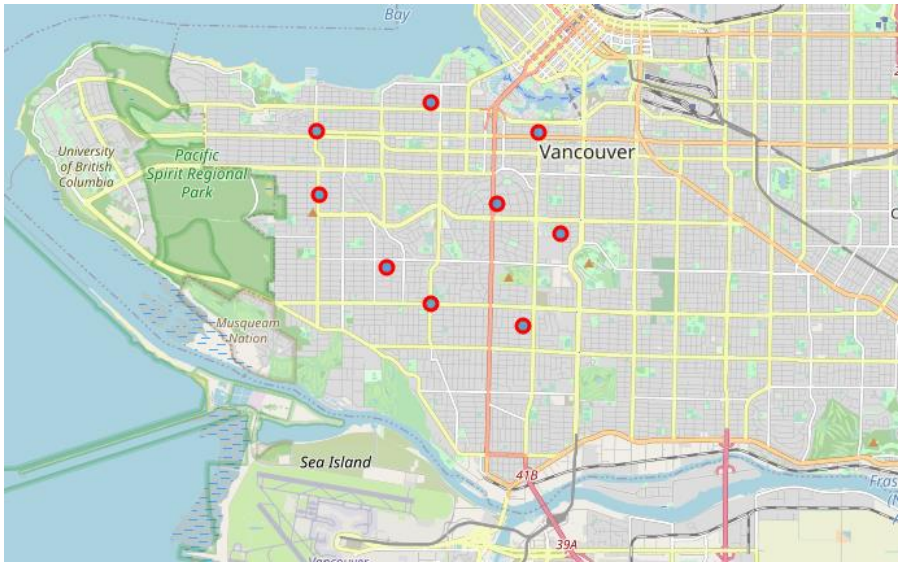
FINDINGS

As the results are analyzed, South Vancouver has the lowest crime rate out of the 4 boroughs in Vancouver. This may be because the number of neighborhoods is minimum there. Hence, we will consider the borough West Side as it has the next minimum crime rate after South Vancouver. Also, it has many neighborhoods and the 'Break and Enter Commercial' type crime is less comparatively.

Next, the geographical coordinates of each neighborhood in West Side Borough has been found using Opencage Geocoder API. Snapshot below:

	Neighbourhood	Borough	Latitude	Longitude
0	Fairview	West Side	49.264113	-123.126835
1	Shaughnessy	West Side	49.251863	-123.138023
2	Kerrisdale	West Side	49.234673	-123.155389
3	Kitsilano	West Side	49.269410	-123.155267
4	Oakridge	West Side	49.230829	-123.131134
5	West Point Grey	West Side	49.264484	-123.185433
6	Arbutus Ridge	West Side	49.240968	-123.167001
7	South Cambie	West Side	49.246685	-123.120915
8	Dunbar-Southlands	West Side	49.253460	-123.185044

Next, the coordinates of Vancouver were retrieved, and Folium library has been used to plot the neighborhoods of West Side in Vancouver. Snapshot below:



4. ANALYSIS AND DISCUSSIONS

The venues of each neighborhood have been retrieved using the Foursquare API. The snapshot is below.

	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
0	Fairview	49.264113	-123.126835	Gyu-Kaku Japanese BBQ	BBQ Joint
1	Fairview	49.264113	-123.126835	CRESCENT nail and spa	Nail Salon
2	Fairview	49.264113	-123.126835	Salmon 'n' Bannock	Restaurant
3	Fairview	49.264113	-123.126835	Finlandia Pharmacy	Pharmacy
4	Fairview	49.264113	-123.126835	Charleson Park	Park

The number of venues in each neighborhood have been calculated as below:

Neighbourhood	Venue
Arbutus Ridge	5
Dunbar-Southlands	6
Fairview	26
Kerrisdale	40
Kitsilano	48
Oakridge	8
Shaughnessy	3
South Cambie	15
West Point Grey	40

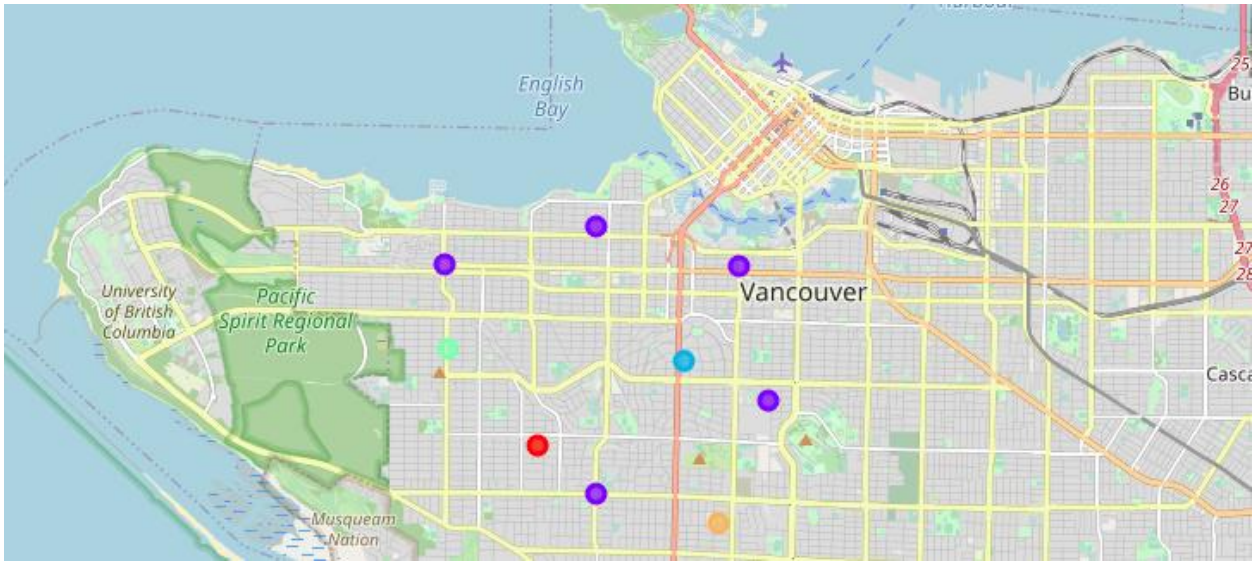
The 10 most common venues of each neighborhood have been calculated then:

	Neighbourhood	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	Arbutus Ridge	Nightlife Spot	Bakery	Pet Store	Grocery Store	Spa	Gastropub	Dessert Shop	Diner	Falafel Restaurant	Fast Food Restaurant
1	Dunbar-Southlands	Sushi Restaurant	Italian Restaurant	Indian Restaurant	Ice Cream Shop	Coffee Shop	Yoga Studio	French Restaurant	Deli / Bodega	Dessert Shop	Diner
2	Fairview	Coffee Shop	Park	Asian Restaurant	Malay Restaurant	Pharmacy	Chinese Restaurant	Diner	Nail Salon	Falafel Restaurant	Restaurant
3	Kerrisdale	Chinese Restaurant	Coffee Shop	Tea Room	Sandwich Place	Pharmacy	Sushi Restaurant	Convenience Store	Hobby Shop	Pizza Place	Liquor Store
4	Kitsilano	Bakery	American Restaurant	Coffee Shop	Japanese Restaurant	Restaurant	Ice Cream Shop	French Restaurant	Sushi Restaurant	Food Truck	Thai Restaurant

K Means Clustering has been used to group the neighborhoods into five clusters and cluster labels have been associated.

	Neighbourhood	Borough	Latitude	Longitude	Cluster Labels	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	Fairview	West Side	49.264113	-123.126835	1	Coffee Shop	Park	Asian Restaurant	Malay Restaurant	Pharmacy	Chinese Restaurant	Diner	Nail Salon	Falafel Restaurant	Restaurant
1	Shaughnessy	West Side	49.251863	-123.138023	2	Park	Bus Stop	French Restaurant	Yoga Studio	Dessert Shop	Diner	Falafel Restaurant	Fast Food Restaurant	Food Truck	Gastropub
2	Kerrisdale	West Side	49.234673	-123.155389	1	Chinese Restaurant	Coffee Shop	Tea Room	Sandwich Place	Pharmacy	Sushi Restaurant	Convenience Store	Hobby Shop	Pizza Place	Liquor Store
3	Kitsilano	West Side	49.269410	-123.155267	1	Bakery	American Restaurant	Coffee Shop	Japanese Restaurant	Restaurant	Ice Cream Shop	French Restaurant	Sushi Restaurant	Food Truck	Thai Restaurant
4	Oakridge	West Side	49.230829	-123.131134	4	Israeli Restaurant	Fast Food Restaurant	Café	Pharmacy	Sandwich Place	Sushi Restaurant	Convenience Store	Vietnamese Restaurant	Falafel Restaurant	Deli / Bodega

The Folium map below shows the clusters.



5. RESULTS

Each cluster has then been analyzed to find out the results:

```
In [36]: vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 2, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

Out[36]:

	Borough	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
1	West Side	Park	Bus Stop	French Restaurant	Yoga Studio	Dessert Shop	Diner	Falafel Restaurant	Fast Food Restaurant	Food Truck	Gastropub

```
In [37]: vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 3, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

Out[37]:

	Borough	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
8	West Side	Sushi Restaurant	Italian Restaurant	Indian Restaurant	Ice Cream Shop	Coffee Shop	Yoga Studio	French Restaurant	Deli / Bodega	Dessert Shop	Diner

```
In [38]: vancouver_merged.loc[vancouver_merged['Cluster Labels'] == 4, vancouver_merged.columns[[1] + list(range(5, vancouver_merged.shape[1]))]]
```

Out[38]:

	Borough	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	West Side	Israeli Restaurant	Fast Food Restaurant	Café	Pharmacy	Sandwich Place	Sushi Restaurant	Convenience Store	Vietnamese Restaurant	Falafel Restaurant	Deli / Bodega

6. CONCLUSION

Based on the above analysis, the grocery stores are much less common in clusters(with labels 2, 3, 4) and therefore ideal for setting up one.

Thank you to the wonderful Coursera Team for this great exposure to Data Science !