Capstone Project - The Battle of Neighborhoods

Part I (Week 1)

Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.

This submission will eventually become your **Introduction / Business Problem** section in your final report. So I recommend that you push the report (having your Introduction/Business Problem section only for now) to your Github repository and submit a link to it.

Part 2 [Week 1]

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.

Section 1: Introduction

Utilizing the Foursquare location data, crime statistics and restaurants availability, schools and other convenient shopping stores in Vancouver for homebuyer's decisions and house price predictions.

Background

FourSquare is one of the best sites that will tell you all about places to go, things to see, restaurants to eat at, bars to drink in, nightclubs to part the night away in and then where to go in the morning to get breakfast and a strong coffee. However, Foursquare alone can't give you hints of which areas are perfect to settle or live. It also cant predict the real price of the home in an area. If we combine data from several sources we might be able to build a model that will help homebuyers to take such a life decision with accurate if not close to accurate about the real price. That is no need even to consult a real state agent or spend a lot of time searching if that is the real value of the house you decide to buy or not. Imagine the following scenario:

1. You like to plan ahead and always review your options and make your choices about where you will settle and live.
2. You are in a city and you plan to move to a different one (example Vancouver) to settle after you got a job offer.
3. You start exploring the chances to buy a home in an area in Vancouver.
4. But you know no one in Vancouver to show you around to all the best areas and even if so they don't have much of information about the house prices or they don't know what you really looking for.
5. Also you want to avoid directly asking a real state agent and even if so you need a full and accurate estimate of the prices based on your own criteria.

What do you do ... ?

Project Idea

My idea for the Capstone Project is to show that when driven by venue and location data from FourSquare, backed up with open source crime data in Vancouver, that it is possible to predict the price of a home based on your own criteria.

A high level approach is as follows:

1. The homebuyers decides on a city location [in this case Vancouver]
2. The ForeSquare website is scrapped for the top venues in the city
3. From this list of top venues the list is augmented with additional geographical data
4. Using this additional geographical data the top nearby restaurants, schools, shopping mall are selected and scored.
5. The historical crime within a predetermined distance of all venues are obtained
6. A map is presented to the to the homebuyers showing the selected venues and crime statistics of the area.
7. The base estimate of a home in a neighborhood is predicted based on some criteria.

Note: The prediction model can be interactive and the homebuyers can add futures to see how the price can be affected (example number of rooms, bathrooms, size,…,etc). But because of time limitation of this project I will only achieve till point six where a basic estimation of a single family home can be visualized from the map.

Who is this solution targeted at

This solution is targeted regular homebuyers and any real state investors. They want to visualize the home prices on the map based on the prediction model giving all previously mentioned criteria

There are many data science aspect of this project including:

1. Data Acquisition
2. Data Cleansing
3. Data Analysis
4. Data visualization

Section 2: Data

Data Description

In this section, I will describe the data used to solve the problem as described previously.

As noted below in the Further Development Section, it is possible to attempt quite complex and sophisticated scenarios when approaching this problem. However, given the size of the project and for simplicity only the following scenario will be addressed:

1. Query the FourSqaure website for the top sites in Vancouver
2. Use the FourSquare API to get supplemental geographical data about the top sites
3. Use the FourSquare API to get top restaurant recommendations closest to each of the top site
4. Use open source Vancouver Crime data to provide the user with additional crime data.
5. Use the average price in the city of Vancouver as a basic price.
6. Then use the criteria to estimate the predicted price based on all previously mentioned factors.
7. Visualize the home prices on the map.

Section 3: Methodology

1. Get and clean FourSquare Data

We will use the FourSquare website to request the top 50 sites in Vancouver and then use BeautifulSoup to scrape the data. We will focus our search only on the following major Venues.

Feature	ID
Shopping mall	4bf58dd8d48988d1fd941735
University	4bf58dd8d48988d1ae941735
school	4bf58dd8d48988d13b941735
Park	5bae9231bedf3950379f89d0
Trail	4bf58dd8d48988d159941735
Transportation	4d4b7105d754a06379d81259

Define Foursquare Credentials and Version

[5]:
```
CLIENT_ID = '5JH4RNH4BWX0DHRIZ1B3HCPGJ5FFZRBB4ANWHWUMIH5GOGTG' #  Foursquare ID
CLIENT_SECRET = 'UL4PP3BO54X3VBFR02W5X42F0E52UTM4FQ4J1LIF0QN2XFI5' #  Foursquare Secret
VERSION = '20180605' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: 5JH4RNH4BWX0DHRIZ1B3HCPGJ5FFZRBB4ANWHWUMIH5GOGTG
CLIENT_SECRET:UL4PP3BO54X3VBFR02W5X42F0E52UTM4FQ4J1LIF0QN2XFI5
```

[8]:
```
# create an agent
address = 'Vancouver, canada'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Vancouver are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Vancouver are 49.2608724, -123.1139529.
```

[ ]:
```
# We will focus our search only on the following major Venues.
# Shopping mall: 4bf58dd8d48988d1fd941735 University : 4bf58dd8d48988d1ae941735
# school:4bf58dd8d48988d13b941735 Park: 5bae9231bedf3950379f89d0
# Trail: 4bf58dd8d48988d159941735 Transportation: 4d4b7105d754a06379d81259
```

[9]:
```
Mall_id = '4bf58dd8d48988d1fd941735'
University_id ='4bf58dd8d48988d1ae941735'
school_id='4bf58dd8d48988d13b941735'
Park_id= '5bae9231bedf3950379f89d0'
Trail_id='4bf58dd8d48988d159941735'
Transportation_id='4d4b7105d754a06379d81259'
url_1 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,Mall_id)
url_2 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,University_id)
url_3 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,school_id)
url_4 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,Park_id)
url_5 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,Trail_id)
url_6 = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,latitude,longitude,Transportation_id)
```

1. The Foursquare Credential and version are defined as can be seen above
2. The IDs of all identified features are extracted from the Foursquare website.
3. A URL get requested procedure is initiated for each feature.
4. The generated data from each URL request is converted into a Jason file.

5. Then only the related data fields are extracted and stored into a separate data frame. See below an example:

[60]:

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	CF Pacific Centre	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.282980	-123.118361
1	City Square Shopping Centre	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.260947	-123.116535
2	Centrepoint Shopping Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.259030	-123.101381
3	Harbour Centre Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.284562	-123.112296
4	Granville Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.281772	-123.120302
5	King Edward Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.248629	-123.125933
6	Hudson's Bay	[{'id': '4bf58dd8d48988d1f6941735', 'name': 'D...	49.282483	-123.117514
7	Kingsgate Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.262375	-123.097736
8	Apple Pacific Centre	[{'id': '4bf58dd8d48988d122951735', 'name': 'E...	49.283404	-123.117292
9	International Village Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.280425	-123.106967
10	International Village	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.280238	-123.106614
11	Vancouver Centre Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.282383	-123.118634
12	Prada	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.284055	-123.116317
13	Chinatown Plaza	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.279458	-123.100673
14	Broadway & Oak	[{'id': '4bf58dd8d48988d1f6931735', 'name': 'G...	49.263423	-123.126310
15	Arbutus Mall	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.248724	-123.154292
16	Arbutus Village	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.248806	-123.154121
17	Il Mercato	[{'id': '4bf58dd8d48988d1fd941735', 'name': 'S...	49.269969	-123.069841
18	Pacific Centre Food Court	[{'id': '4bf58dd8d48988d120951735', 'name': 'F...	49.283124	-123.118434

6. Each dataframe is filtered based on the category field, example shown below

[61]:

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	CF Pacific Centre	Shopping Mall	49.282980	-123.118361
1	City Square Shopping Centre	Shopping Mall	49.260947	-123.116535
2	Centrepoint Shopping Mall	Shopping Mall	49.259030	-123.101381
3	Harbour Centre Mall	Shopping Mall	49.284562	-123.112296
4	Granville Mall	Shopping Mall	49.281772	-123.120302
5	King Edward Mall	Shopping Mall	49.248629	-123.125933
6	Hudson's Bay	Department Store	49.282483	-123.117514
7	Kingsgate Mall	Shopping Mall	49.262375	-123.097736
8	Apple Pacific Centre	Electronics Store	49.283404	-123.117292
9	International Village Mall	Shopping Mall	49.280425	-123.106967
10	International Village	Shopping Mall	49.280238	-123.106614
11	Vancouver Centre Mall	Shopping Mall	49.282383	-123.118634
12	Prada	Shopping Mall	49.284055	-123.116317
13	Chinatown Plaza	Shopping Mall	49.279458	-123.100673
14	Broadway & Oak	General Travel	49.263423	-123.126310
15	Arbutus Mall	Shopping Mall	49.248724	-123.154292
16	Arbutus Village	Shopping Mall	49.248806	-123.154121
17	Il Mercato	Shopping Mall	49.269969	-123.069841
18	Pacific Centre Food Court	Food Court	49.283124	-123.118434

2. Get and clean Crime Data

1. Check the content of the crime data

```
[17]: DF_Crime.head()
```

[17]:

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y	Latitude	Longitude
0	Other Theft	2003	5	12	16.0	15.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
1	Other Theft	2003	5	7	15.0	20.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
2	Other Theft	2003	4	23	16.0	40.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
3	Other Theft	2003	4	20	11.0	15.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763
4	Other Theft	2003	4	12	17.0	45.0	9XX TERMINAL AVE	Strathcona	493906.5	5457452.47	49.269802	-123.083763

2. Keep only the related columns of the data as no need for any other column in feature analysis.

[19]:

	TYPE	YEAR	NEIGHBOURHOOD	Latitude	Longitude
99207	Break and Enter Commercial	2005	Strathcona	49.270524	-123.087369
99208	Theft from Vehicle	2005	Oakridge	49.233524	-123.118504
99209	Break and Enter Commercial	2005	West End	49.285132	-123.123461
99210	Theft from Vehicle	2005	Oakridge	49.233524	-123.118504
99211	Theft of Vehicle	2005	Renfrew-Collingwood	49.258095	-123.028596

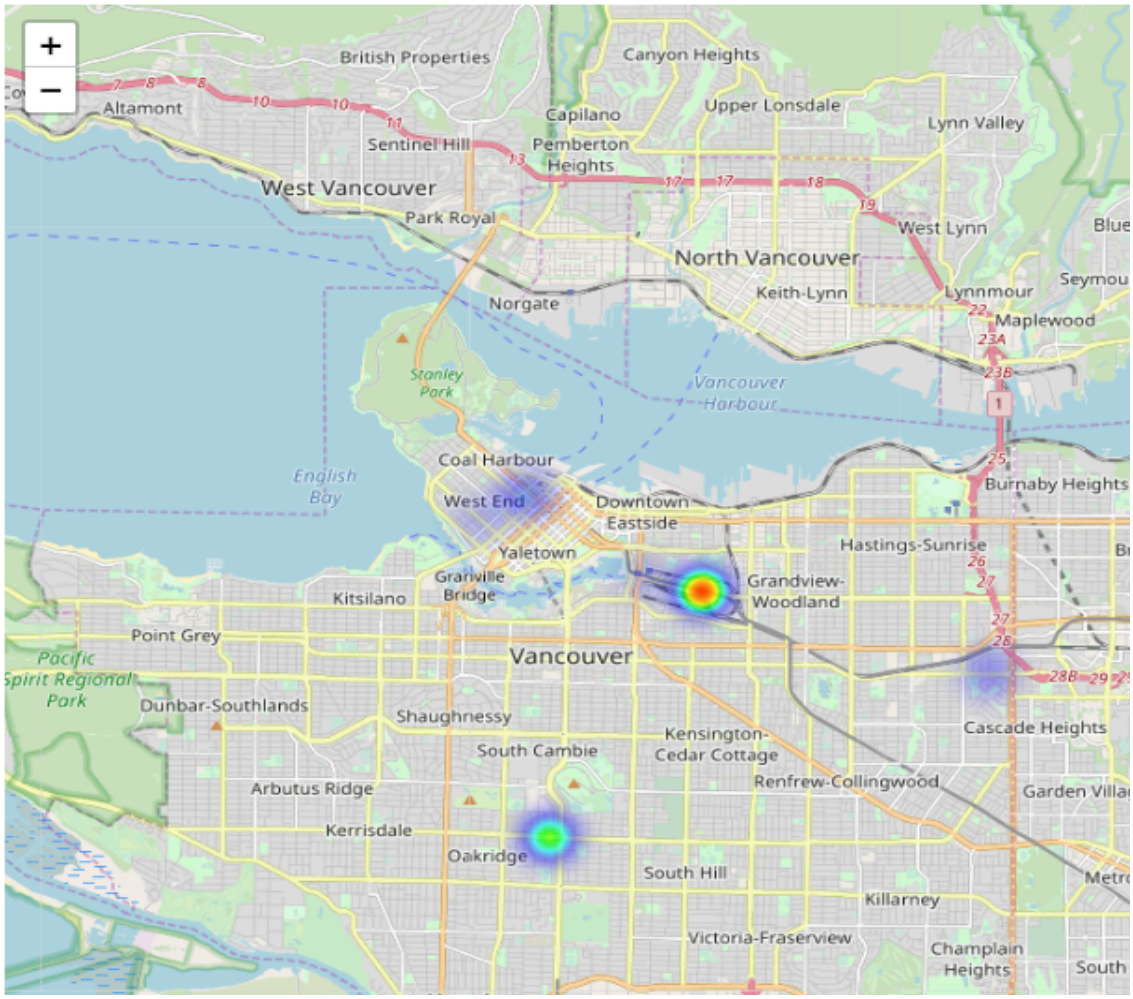
3. The number of unique Crimes categories based on the Primary type is extracted and then the most common crime is written to a new dataframe.

[25]:

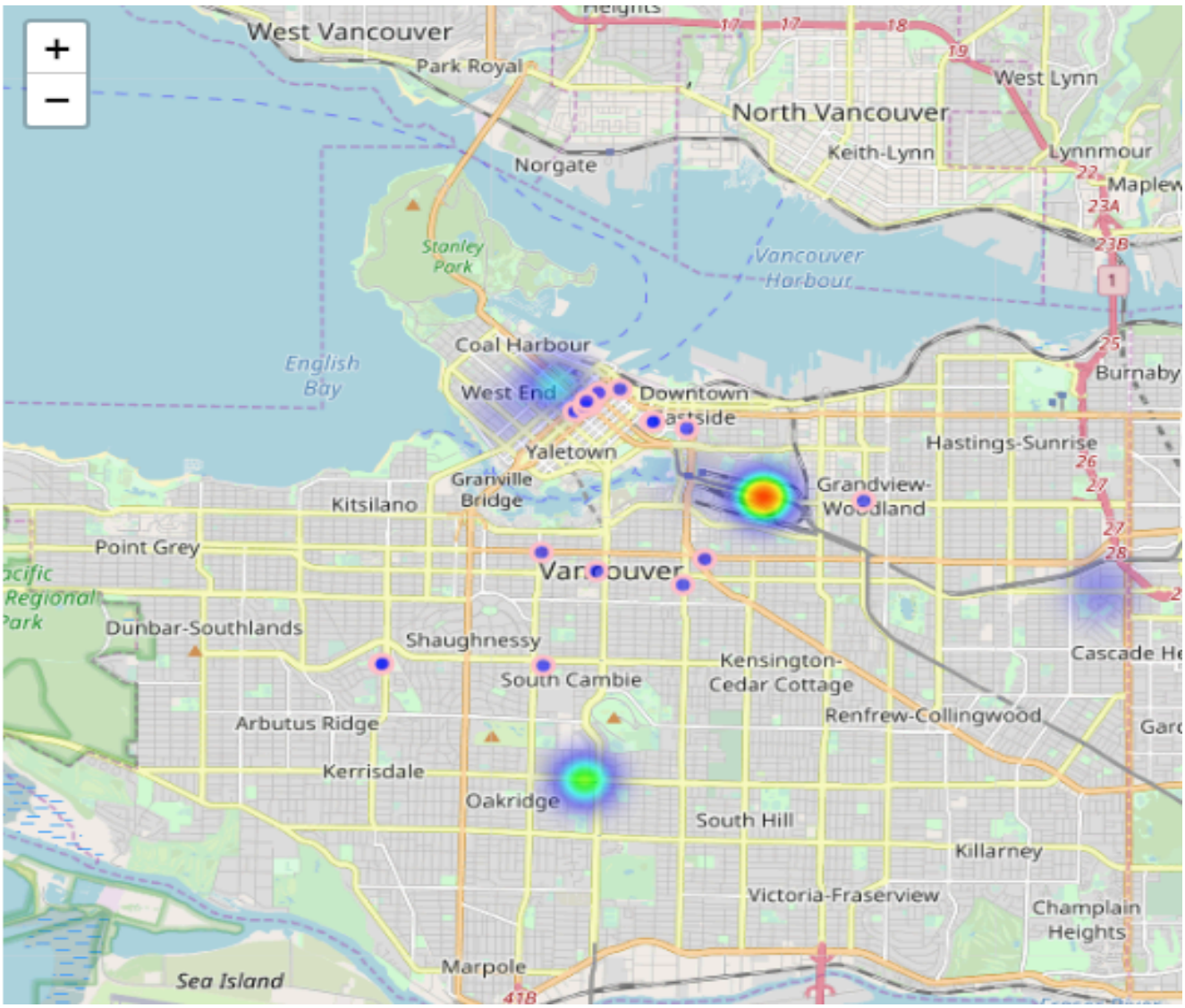
	TYPE	Count
6	Theft from Vehicle	136752
3	Mischief	58421
1	Break and Enter Residential/Other	47441
5	Other Theft	46980
4	Offence Against a Person	46831
0	Break and Enter Commercial	27365
8	Theft of Vehicle	25955
7	Theft of Bicycle	23082
10	Vehicle Collision or Pedestrian Struck (with I...	18230
9	Vehicle Collision or Pedestrian Struck (with F...	208
2	Homicide	180

Data Visualization

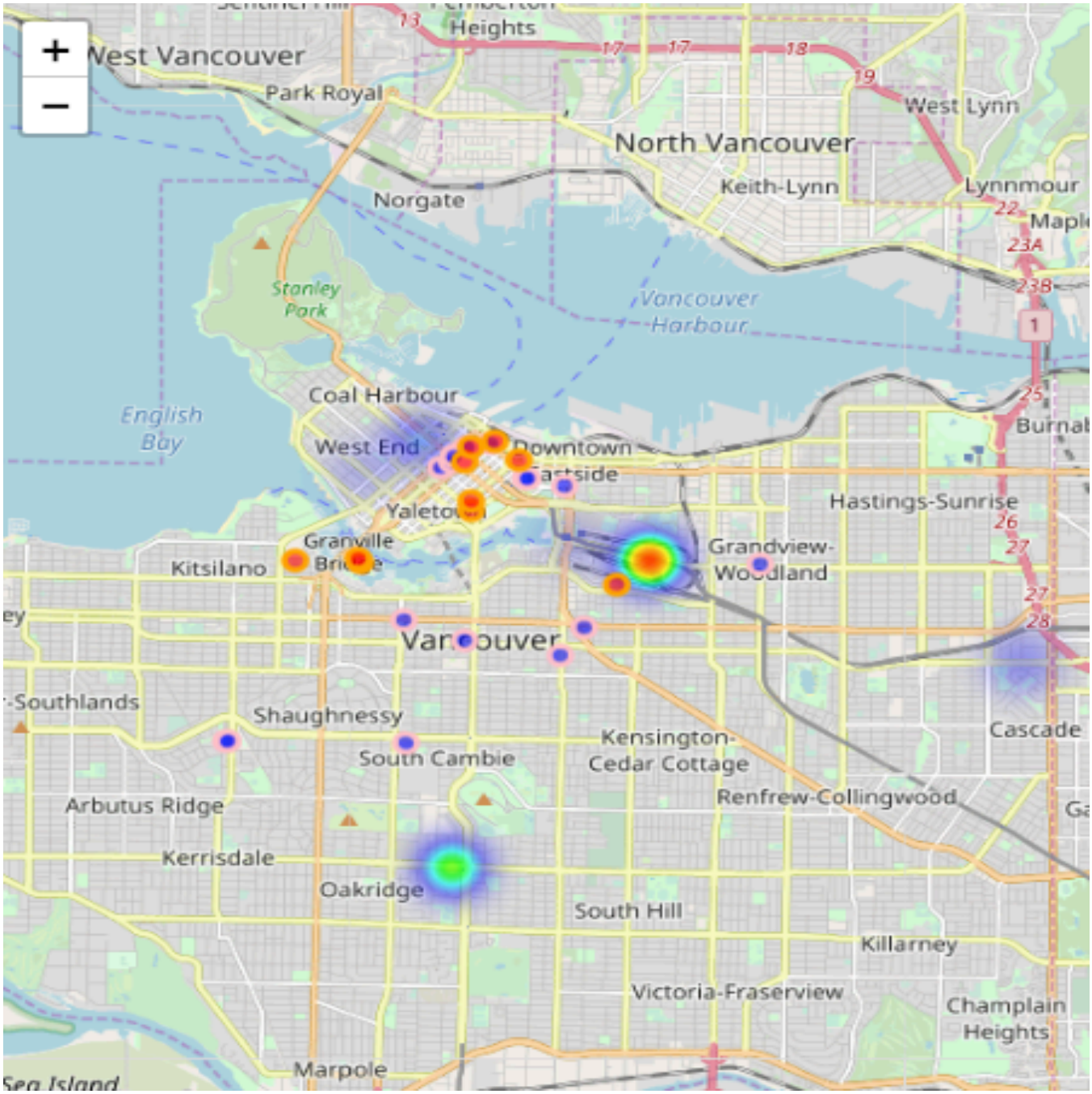
1. A heat map is generated based on the number of crimes and its type.



2. The extracted features will be presented on the same map based on their latitude and longitude. I started with the first date frame (Feature_1) and then I added all other features one by one.

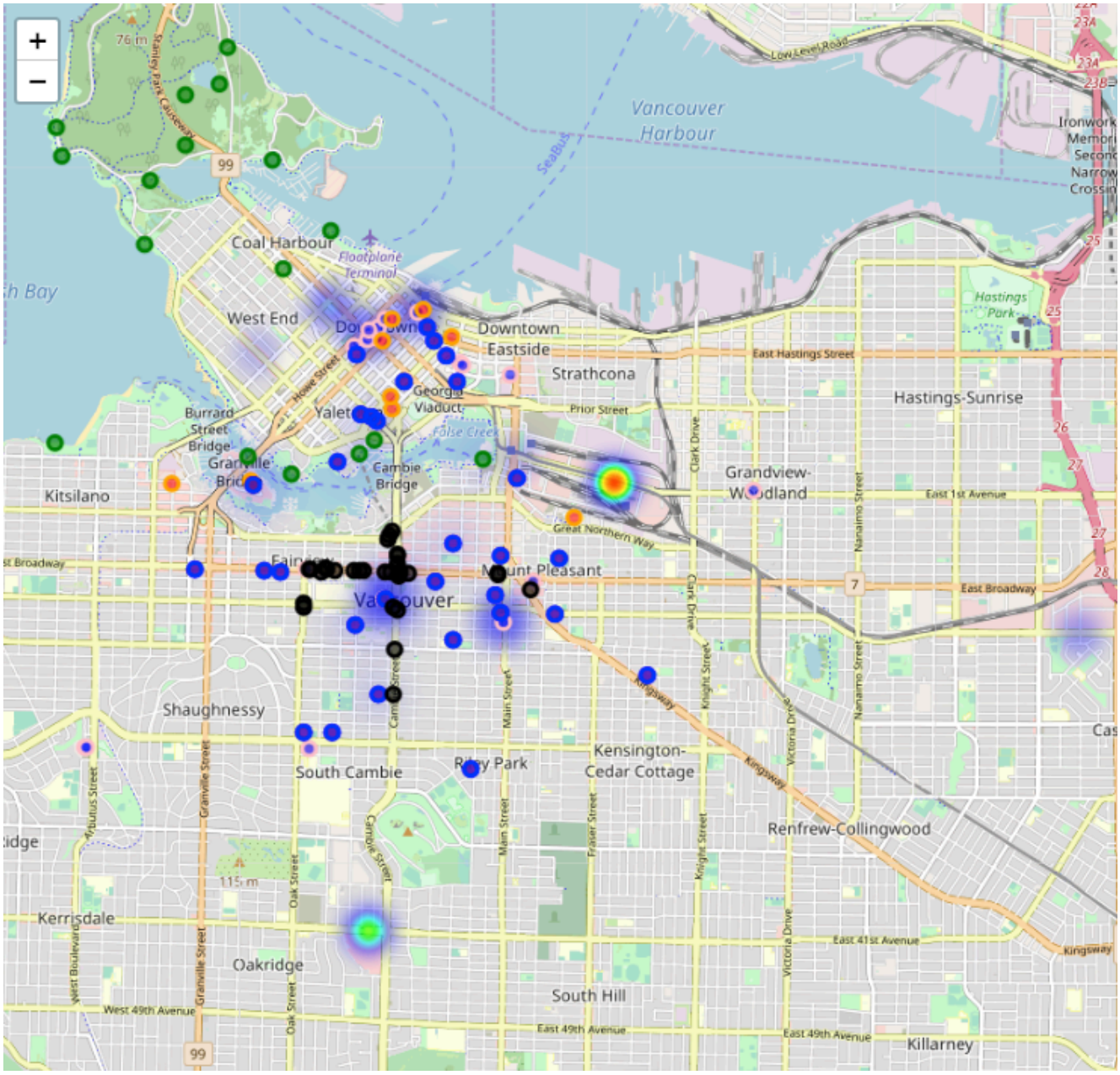

West Vancouver
Park Royal
North Vancouver
West Lynn
Keith-Lynn
Lynnmour
Norgate
Stanley Park
Vancouver Harbour
English Bay
Coal Harbour
West End
Downtown
Yaletown
Granville Bridge
Kitsilano
Point Grey
Grandview-Woodland
Hastings-Sunrise
Burnaby
Dunbar-Southlands
Shaughnessy
South Cambie
Kensington-Cedar Cottage
Renfrew-Collingwood
Arbutus Ridge
Kerrisdale
Oakridge
South Hill
Killarney
Victoria-Fraserview
Champlain Heights
Marpole
Sea Island

2. The second feature is added



Results

After all features have been added; a heat map is generated for each data frame a long with the crime data. Please refer to the code for more clarification. The final result is presented below.



Discussion:

As we can see from the visualized map above the homebuyer can obviously see which area of the city has the most attractive features or which area is more dangerous. The homebuyer or even real state agent can definitely utilize this map with all features to make an initial guess of the price of a home.

Conclusion

The use of data science can help homebuyers to find, select and have an initial estimate of a home based on the extracted features, crime data and offered prices.

They would like to select locations that are nearby attraction areas, safe to stay and walk in as well as near to famous restaurants and transportations.

In this project, we analyzed and clustered Vancouver city based on tourism attractions, crime rates, public transportations and famous restaurants and parks. We also provided recommendations on the top best hotel locations that optimizes homebuyers needs and requirements.

Future development:

This project can be considered a basic for a future development for the same idea. It can be expanded to be more dynamic and as long as the user hover over the map it can predict the price of that land or house.

It might be expanded to include major cities in Canada or US. It also can be used as basic for an app in a cell phone.