CISC 878

PROJECT REPORT

An exploratory data analysis and visualization of crime in Vancouver

Zili Luo

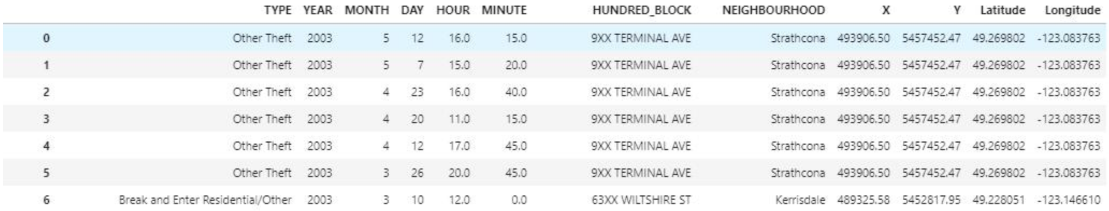
20001744

Introduction

This report is cover my data analytic project, which did an exploratory data analysis and visualization of crime in Vancouver from 2003 to 2017. It will cover My dataset, Initial preprocessing of the dataset, Visualization results, potential correlations between reality and data points, predictive model, preprocessing for the model, and the predictive results.

Dataset

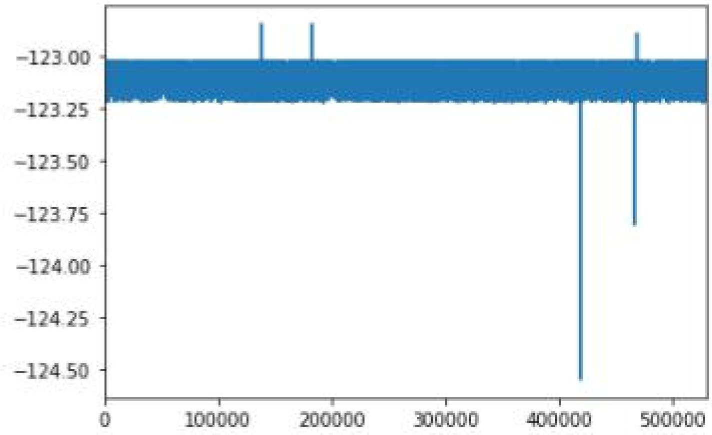
The dataset is available on Kaggle[1] named "crimes in Vancouver". It contains 530,652 records of crimes that happened in Vancouver between January 2003 and June 2017. Each row in the dataset includes the type of the crime, which has nine categories, when it has happened, the neighborhood where it has happened, and the spatial information represented in latitude and longitude. The dataset itself does not contain a rich information.In addition, about 10% of the data is anonymous, and 1% of the data is missing.



	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	493906.50	5457452.47	49.269802	-123.083763
1	Other Theft	2003	5	7	15.0	20.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
2	Other Theft	2003	4	23	16.0	40.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
3	Other Theft	2003	4	20	11.0	15.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
4	Other Theft	2003	4	12	17.0	45.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
5	Other Theft	2003	3	26	20.0	45.0	9XX TERMINAL AVE	Strathcona	493906.50	5457452.47	49.269802	-123.083763
6	Break and Enter Residential/Other	2003	3	10	12.0	0.0	63XX WILTSHIRE ST	Kerrisdale	489325.58	5452817.95	49.228051	-123.146610

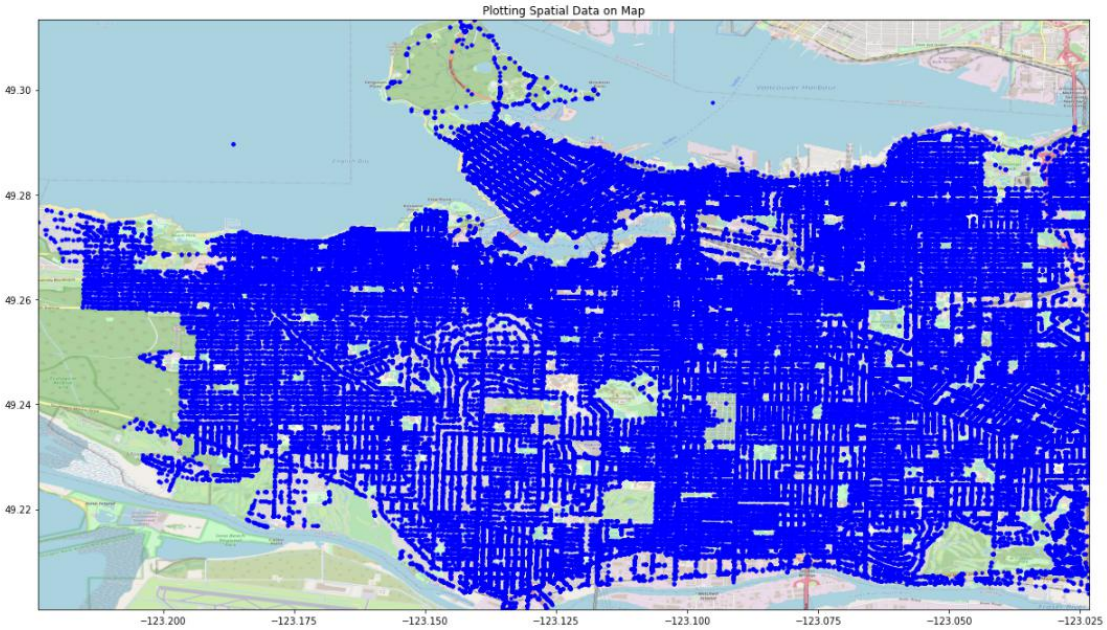
Preprocessing

Since the data is simple, my preprocessing does not take many steps.
I dropped the missing data and anonymous data since both of them do not have a spatial information to visualize them. And I also dropped some outliers ,which are outside of Vancouver, I have found that might be caused by input mistakes.

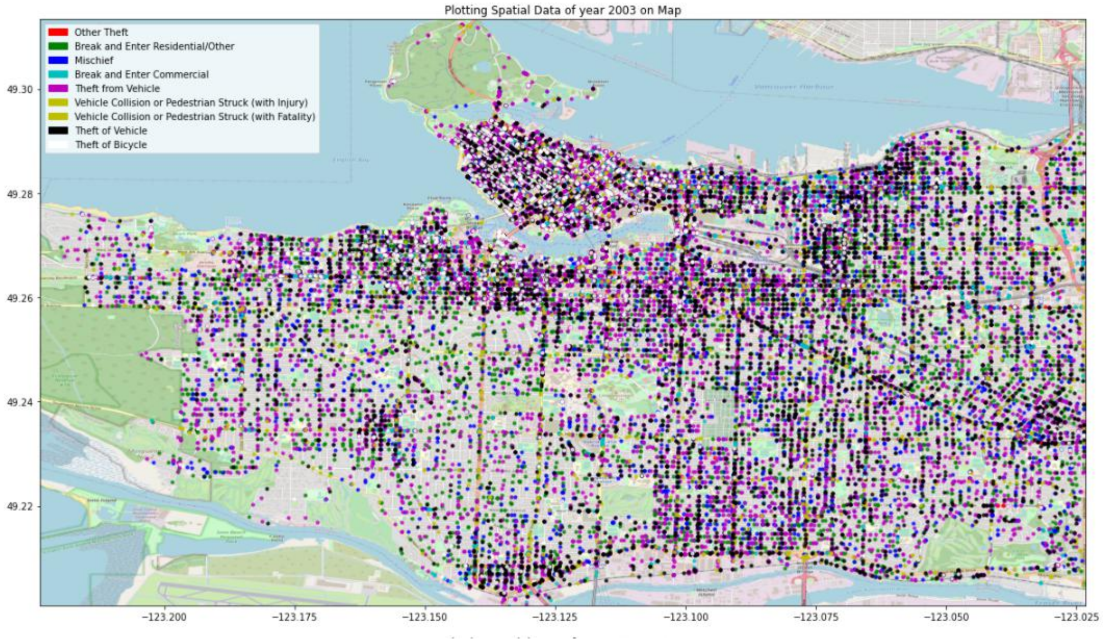


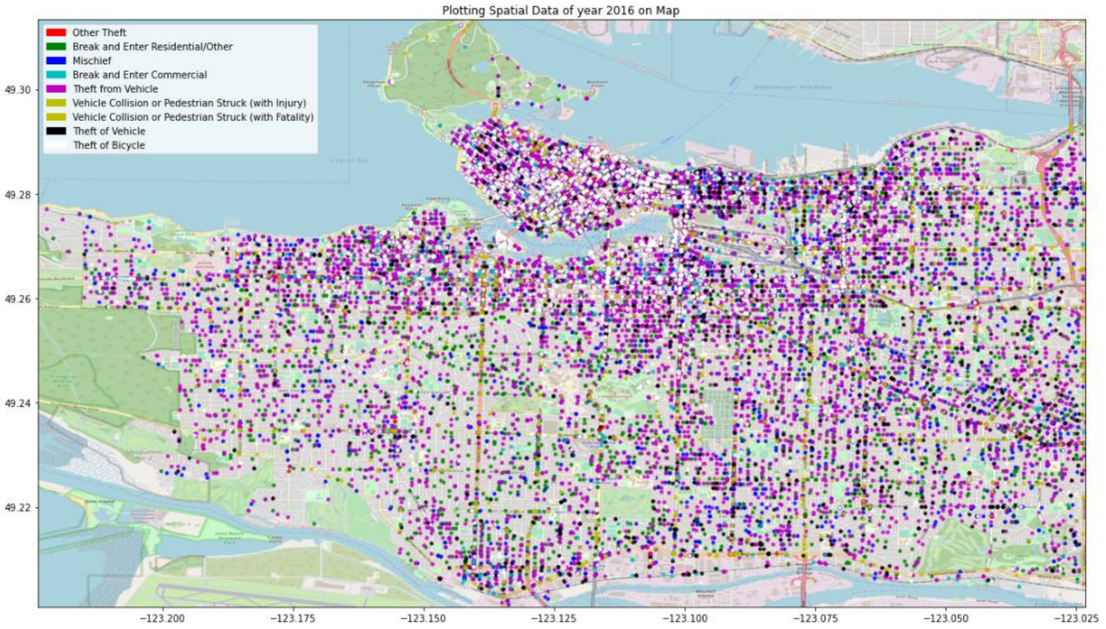
Visualization

The first visualization aims to find any hotspots that crimes are more likely to happen, which visualize all the data points on the map at once. But the number of data points is too large to find any trace; the result shows crimes happen everywhere in Vancouver during those years.



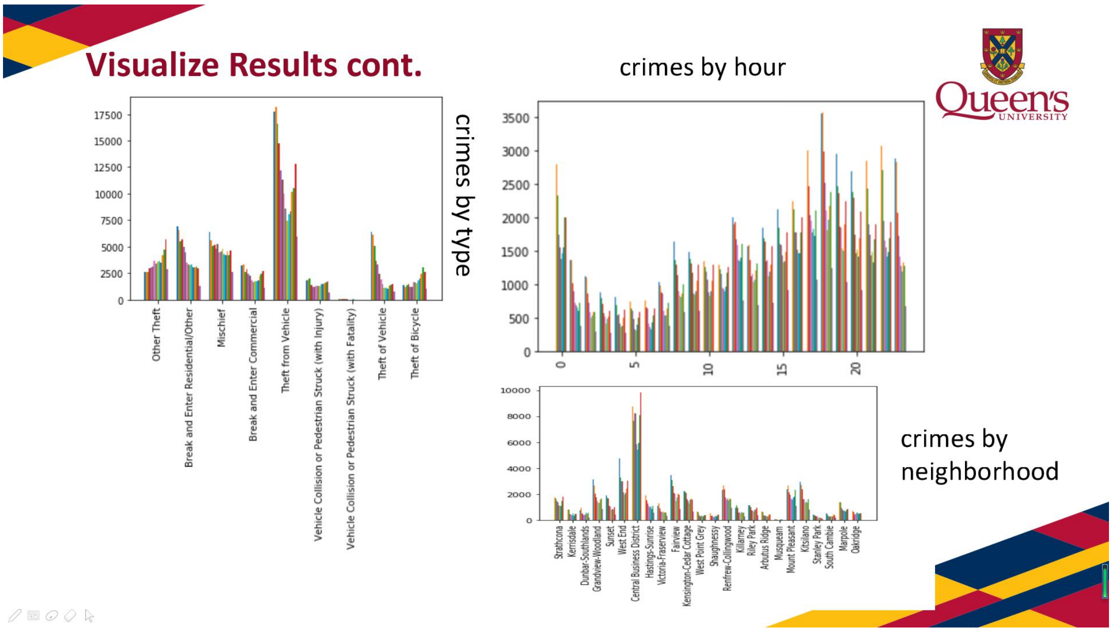
In the second visualization, the data is split by each year and also by the crime types. In this try, some traces of crime cluster could be observed in the map. All the map visualization result could be found at appendix.


Plotting Spatial Data of year 2003 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2016 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

The third visualization is on the data itself. Some histograms is created to find if anything in data itself are interesting in those grams.

Three histograms to plot crimes every year by a different type, hour, and neighborhood index.



Visualize Results cont.

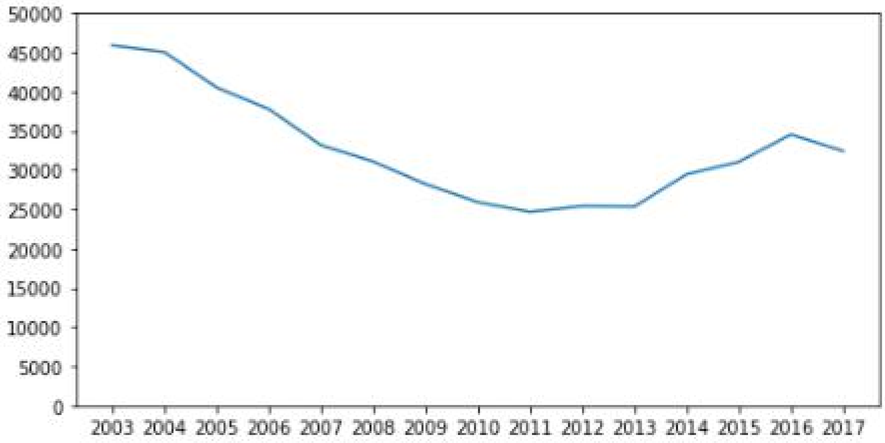
crimes by type

crimes by hour

crimes by neighborhood

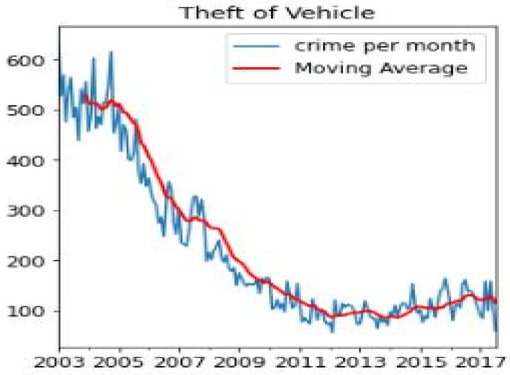
There isn't anything specific in the first two histograms, but in the crime by neighborhood gram, It seems in some neighborhoods, the crime rate has a significant change.

Then the total amount of crimes every year are plotted, since the data in 2017 is only a half-year data, the data in 2017 is doubled; we could observe that the number of crimes decreased from 2003 to 2010 and have gradually increased with a bump in 2016.



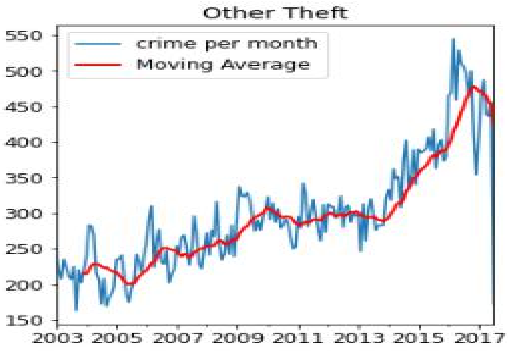
The next visualization is the theft crimes categories, which includes "Theft of Vehicle", "Other Theft", "Theft from Vehicle", "Theft of Bicycle", since the theft crimes occupies the majority of the dataset.

- First, lets start with "Theft of Vehicle":



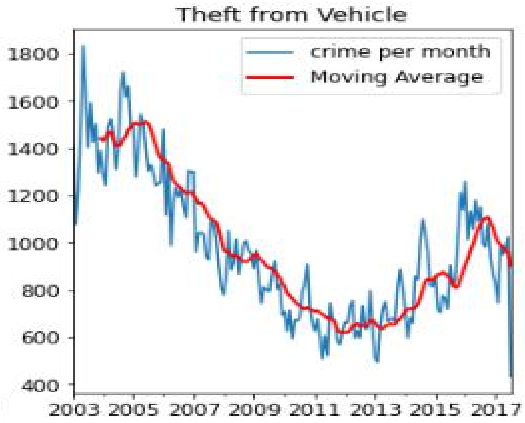
- It had a major decrease, from an average of around 520 crimes per month in 2003 to around 100 in 2012.
- Although the average has been increasing in the past years, it's way below 2003.
- In 2002, the "Bait Car" program was launched and in 2003 the IMPACT group was formed in response to this peak in theft. It looks like they've been doing a great job!

- Second, about "Other Theft":



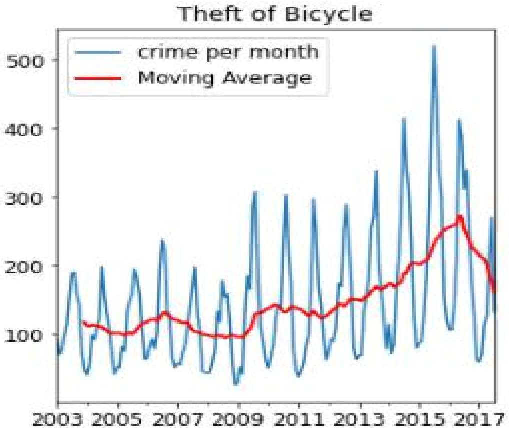
- On the opposite trend, other theft has been increasing, from around 200 to almost 500 crimes per month.
- With the decrease of "Theft of Vehicle", This means the theft still happens. While the vehicle itself might be well protected from being theft, the number of thieves does not decrease, and they still do their work.

- About "Theft from Vehicle":



- It is the most frequent type.
- It decreased along with "Theft of Vehicle" until 2012, but then it increased significantly.

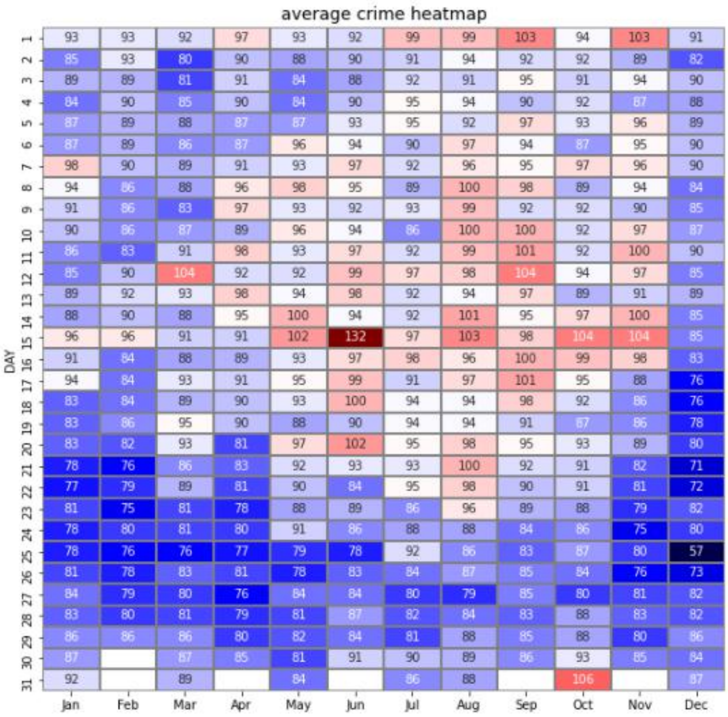
- Finally, about "Theft of Bicycle":



  - We can see a clear trend within the year. It has peaks during the summer months, which is expected.
  - The average has also been increasing.

The last work I did in the visualization part is the Heat map. I have made two heat maps representing the average crimes that happen each day and average crimes in each hour.

The first heatmap to be introduced is average crimes in each day heatmap.



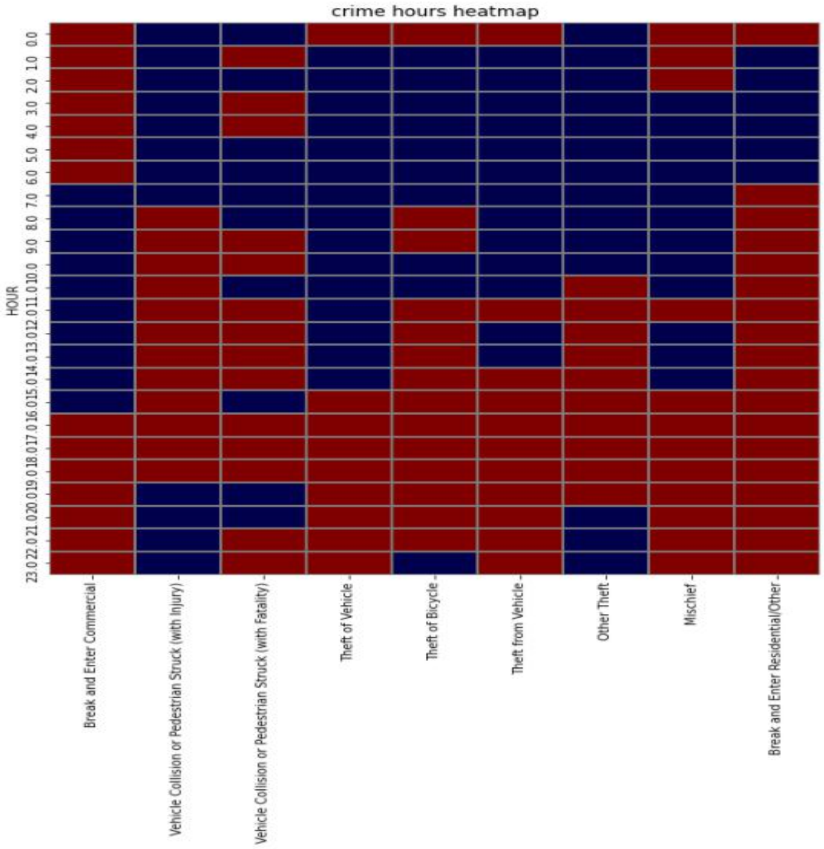
Blue means good days. Red bad days. White average days.

The calmest day of crime is Christmas Day, December 25. That might because people all stay with their families, so no one walk in the street to be theft or criminals might also celebrate with families.

The worst day is June 15, which is caused by The Stanley Cup Riot in 2016. Over 600 crimes

happened on that day.
The first day of the month is usually a busy day for all months.
Halloween is also dangerous.
The second week of the summer months is usually the most dangerous.
BC Day (August 7) long weekend also has high averages.

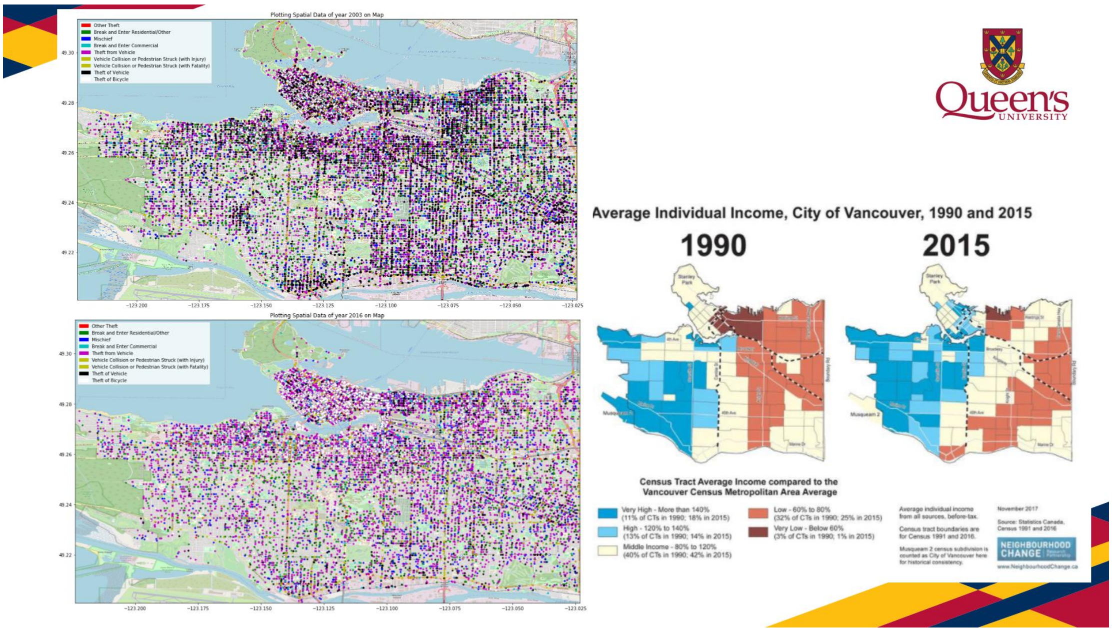
The second heatmap to be introduced is average crimes in each hour heatmap.



Most crimes happen between 5:00 pm and 8:00 pm
Break-and Enter-crimes happen almost all day.
All types of theft doesn't occur much in those between 3 am and 5 am as expected, that might because of most people are sleeping

After all the visualization, I tried to find if there are anything in reality could explain the changes in the dataset visualization.
At first, I checked the Vancouver Police Department website[4], seems they have the same strategy goal every year, which aims to decrease the amount of crime in Vancouver by 2.5%. But this can explain neither the increase from 2010 to 2016 and the bump in 2016. Then I searched for the big event that happened in Vancouver[3] and found Vancouver hosted the 2010 Winter Olympics and the election for the host city in 2003. This could explain the decrease from 2003 to 2010. And In August 2015, the 33-year-old businessman was killed by six Vancouver police officers. This news still has high clout in recent months. This event might cause a decrease in the trust of police at a significant amplitude.

Recall that the findings of some crimes are clustered in some neighborhoods. I went to the government of Vancouver website and found some evidence for this phenomenon[2]. The graph at right shows the average income of each area; we could observe the crime become sparser as the average income goes higher.


Plotting Spatial Data of year 2003 on Map

Plotting Spatial Data of year 2016 on Map

Average Individual Income, City of Vancouver, 1990 and 2015

Census Tract Average Income compared to the Vancouver Census Metropolitan Area Average

- Very High - More than 140% (11% of CTs in 1990; 18% in 2015)
- High - 120% to 140% (13% of CTs in 1990; 14% in 2015)
- Middle Income - 80% to 120% (40% of CTs in 1990; 42% in 2015)
- Low - 60% to 80% (32% of CTs in 1990; 25% in 2015)
- Very Low - Below 60% (3% of CTs in 1990; 1% in 2015)

Average individual income from all sources, before-tax.

Census tract boundaries are for Census 1991 and 2016.

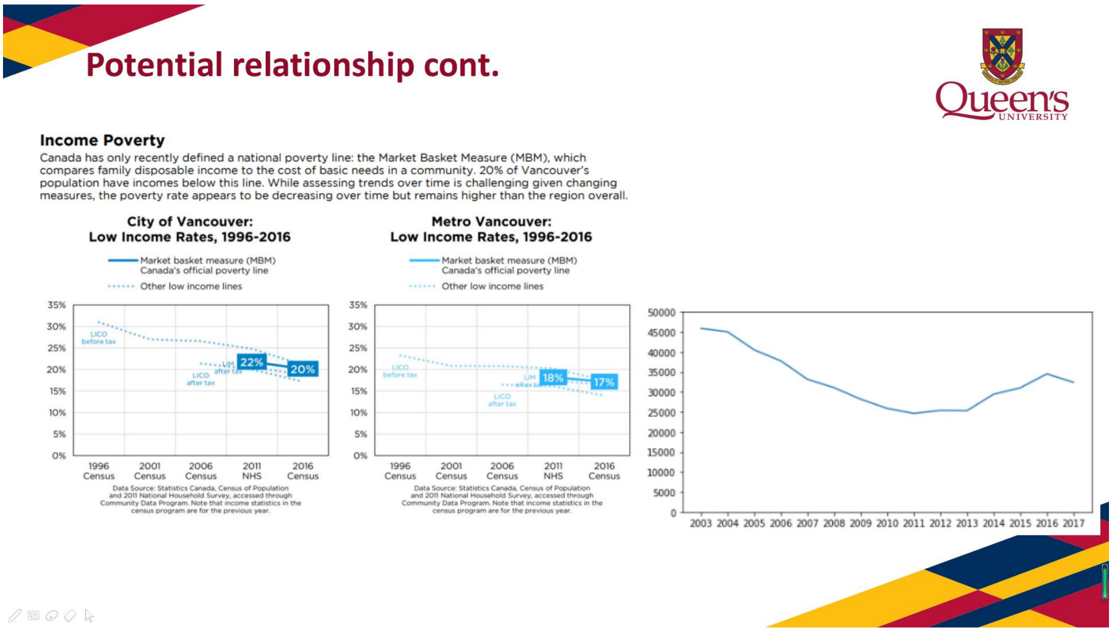
Musqueam 2 census subdivision is counted as City of Vancouver here for historical consistency.

November 2017

Source: Statistics Canada, Census 1991 and 2016

www.NeighbourhoodChange.ca

Also, except for the bump in 2016, the total crime curve each year has similar trends to the low-income rates curve in Vancouver[2].


Potential relationship cont.

Income Poverty

Canada has only recently defined a national poverty line: the Market Basket Measure (MBM), which compares family disposable income to the cost of basic needs in a community. 20% of Vancouver's population have incomes below this line. While assessing trends over time is challenging given changing measures, the poverty rate appears to be decreasing over time but remains higher than the region overall.

City of Vancouver: Low Income Rates, 1996-2016

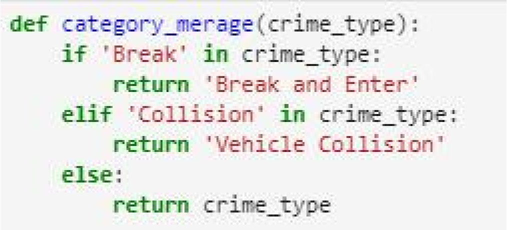
Metro Vancouver: Low Income Rates, 1996-2016

Data Source: Statistics Canada, Census of Population and 2011 National Household Survey, accessed through Community Data Program. Note that income statistics in the census program are for the previous year.

predictive model

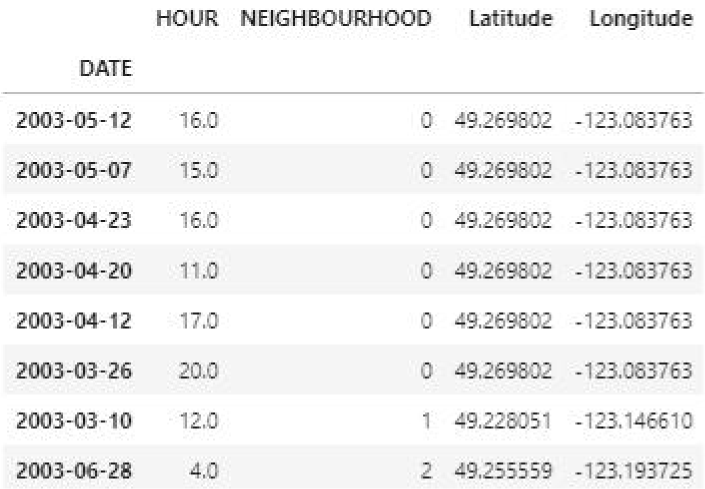
After all the visualization tasks, I'm curious about if a well-performed predictive model could be established to classify those crimes since there are some potential features of those criminal records. But as most data is related to theft, and the visualization results are not clear enough to support the idea. A bad result is expected.

The preprocessing part is still simple, as most features are already nominal. I merged some crime types, such as vehicle collision, break-and-enter. Select some essential elements such as month, hour, neighborhood, etc. And convert all ordinal data into categorical data.



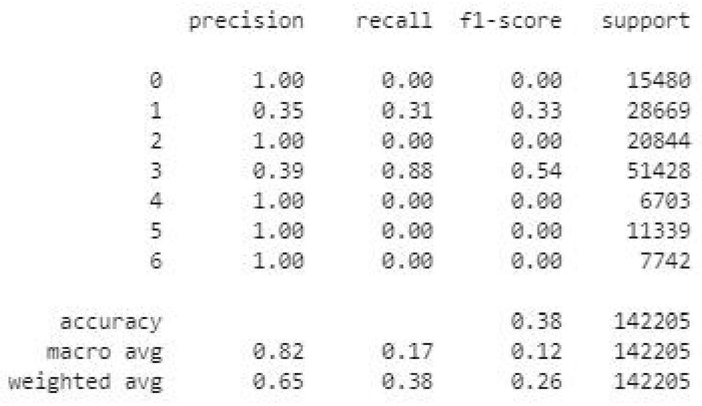
```
def category_merage(crime_type):
    if 'Break' in crime_type:
        return 'Break and Enter'
    elif 'Collision' in crime_type:
        return 'Vehicle Collision'
    else:
        return crime_type
```

The data feeding into the model is consist of hour when it has happened, the neighbourhood where it has happened, latitude and longtidude.



	HOUR	NEIGHBOURHOOD	Latitude	Longitude
DATE				
2003-05-12	16.0	0	49.269802	-123.083763
2003-05-07	15.0	0	49.269802	-123.083763
2003-04-23	16.0	0	49.269802	-123.083763
2003-04-20	11.0	0	49.269802	-123.083763
2003-04-12	17.0	0	49.269802	-123.083763
2003-03-26	20.0	0	49.269802	-123.083763
2003-03-10	12.0	1	49.228051	-123.146610
2003-06-28	4.0	2	49.255559	-123.193725

The result of the predictive model is not good at all. With 38% training accuracy and similar testing accuracy. The model classified all the cases into class 1 and class 3, which are "Break and enter" and "Theft from vehicle". This result is under my expectation as those two cases happen everywhere and occupy the majority of the records.



```
              precision    recall  f1-score   support

           0       1.00      0.00      0.00     15480
           1       0.35      0.31      0.33     28669
           2       1.00      0.00      0.00     20844
           3       0.39      0.88      0.54     51428
           4       1.00      0.00      0.00      6703
           5       1.00      0.00      0.00     11339
           6       1.00      0.00      0.00      7742

    accuracy                           0.38    142205
   macro avg       0.82      0.17      0.12    142205
weighted avg       0.65      0.38      0.26    142205
```

I also modified the model several times but got a similar result. Sometimes the model can recognize the "Mischief" class, which also has a good proportion in the dataset and happens almost everywhere.

Conclusion

The dataset itself does not contain that much information, but we could still observe some potential rules for crimes to happen via explore and visualization the data. And found something significant in reality have a correlation with crimes happening. Although it is nearly impossible to make a well-performed predictive model to predict when and where a crime will happen, which is also a problem in other crime data analytic with a dataset that contains hundreds of features, the police could still benefit from the findings to adjust their policy. That's what Intelligence-led policing encourages.

Reference

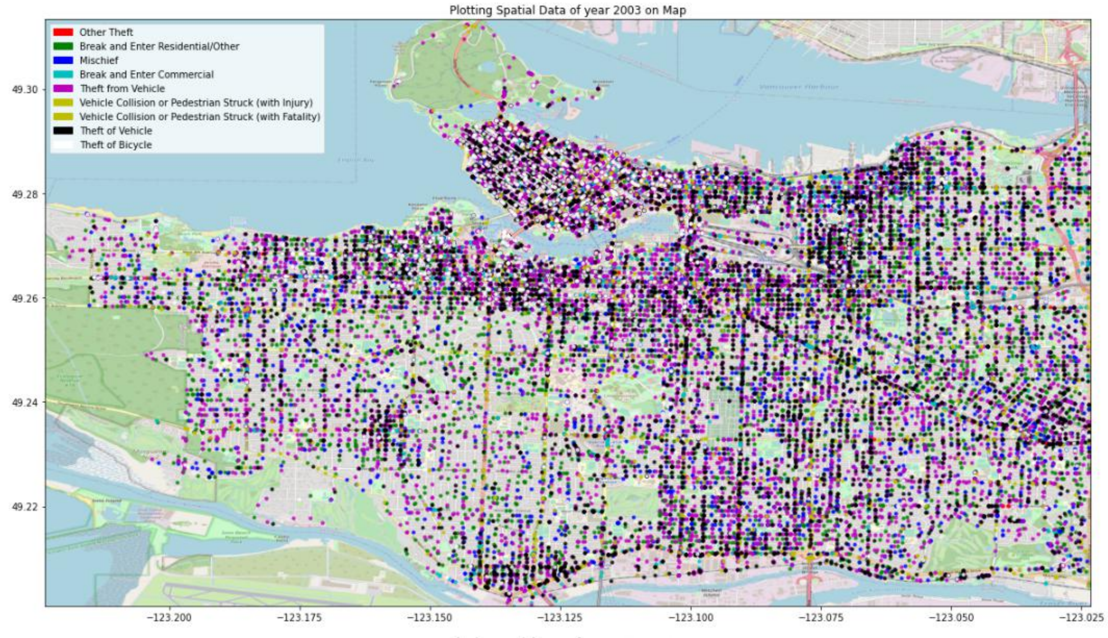
[1] https://www.kaggle.com/wosaku/crime-in-vancouver

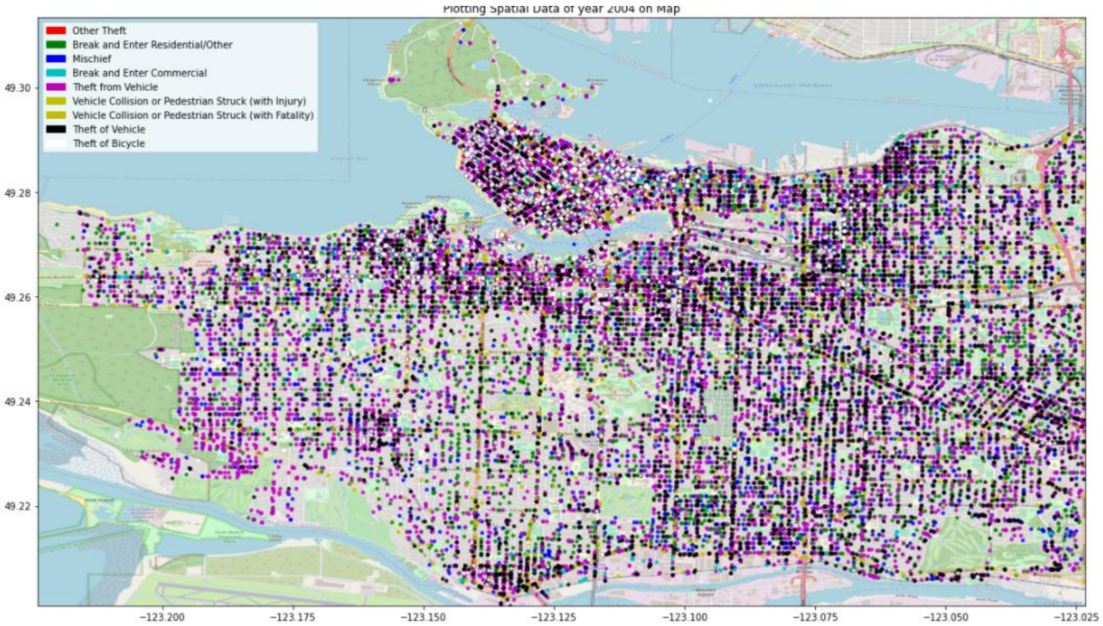
[2] https://vancouver.ca/files/cov/social-indicators-profile-city-of-vancouver.pdf

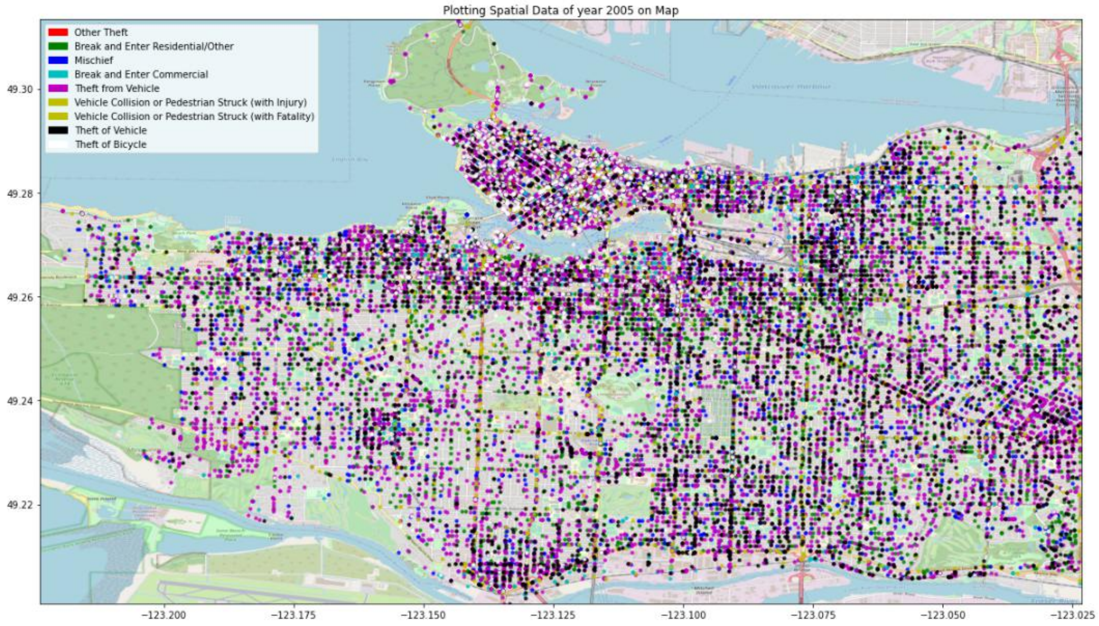
[3] https://dailyhive.com/vancouver/34-biggest-moments-vancouver-history

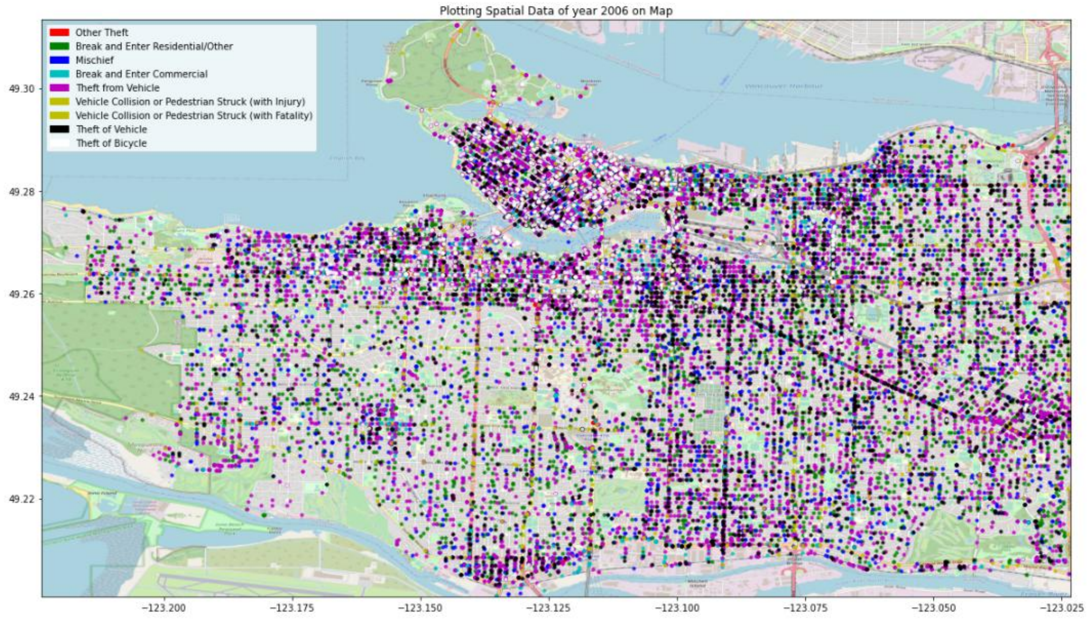
[4] https://vpd.ca/police/about/strategic-planning/index.html

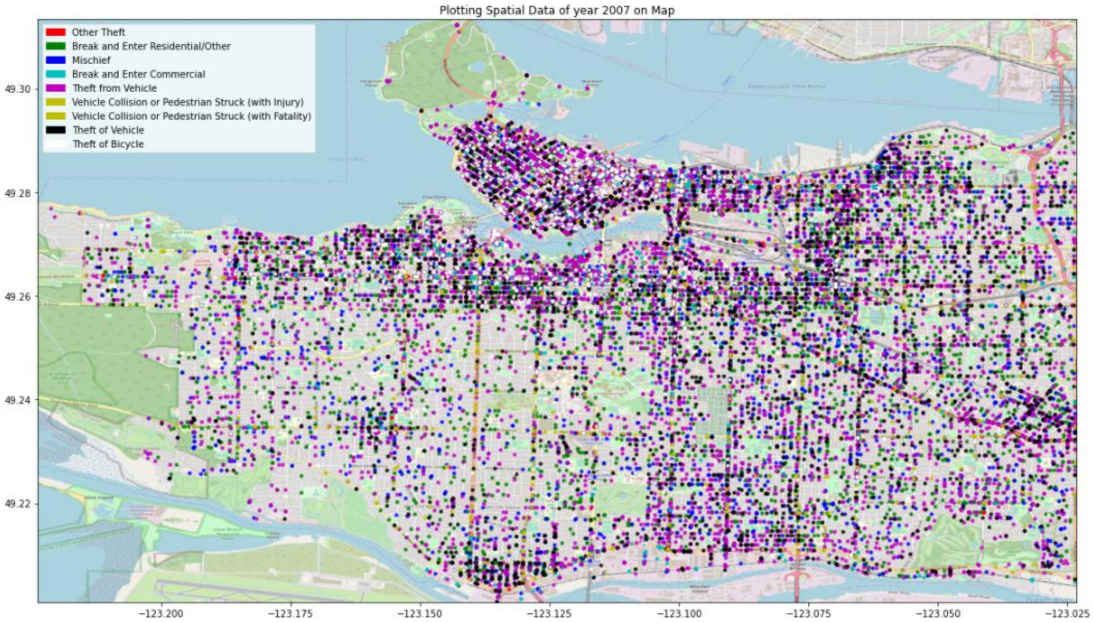
Appendix

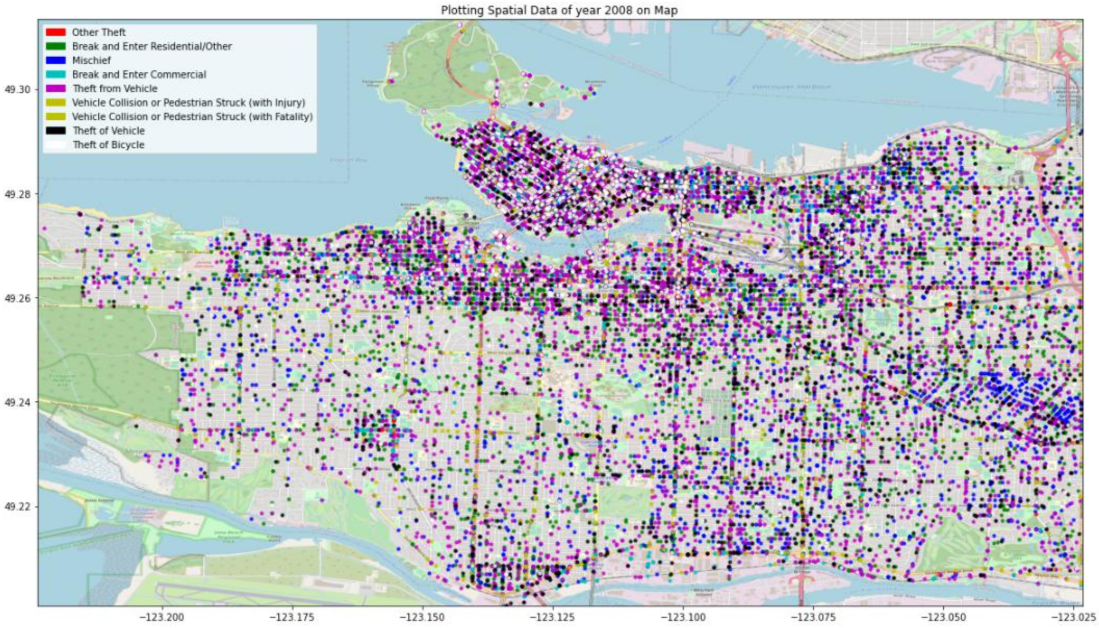

Plotting Spatial Data of year 2003 on Map
🟥 Other Theft
🟩 Break and Enter Residential/Other
🟦 Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
⬛ Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

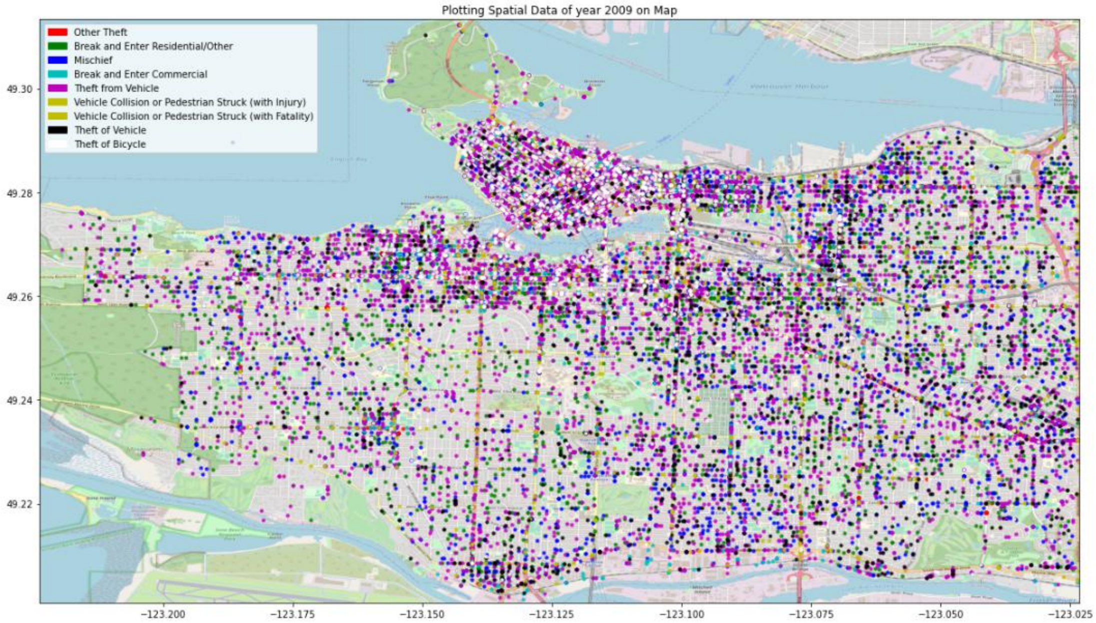

Plotting Spatial Data of year 2004 on Map
🟥 Other Theft
🟩 Break and Enter Residential/Other
🟦 Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
⬛ Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

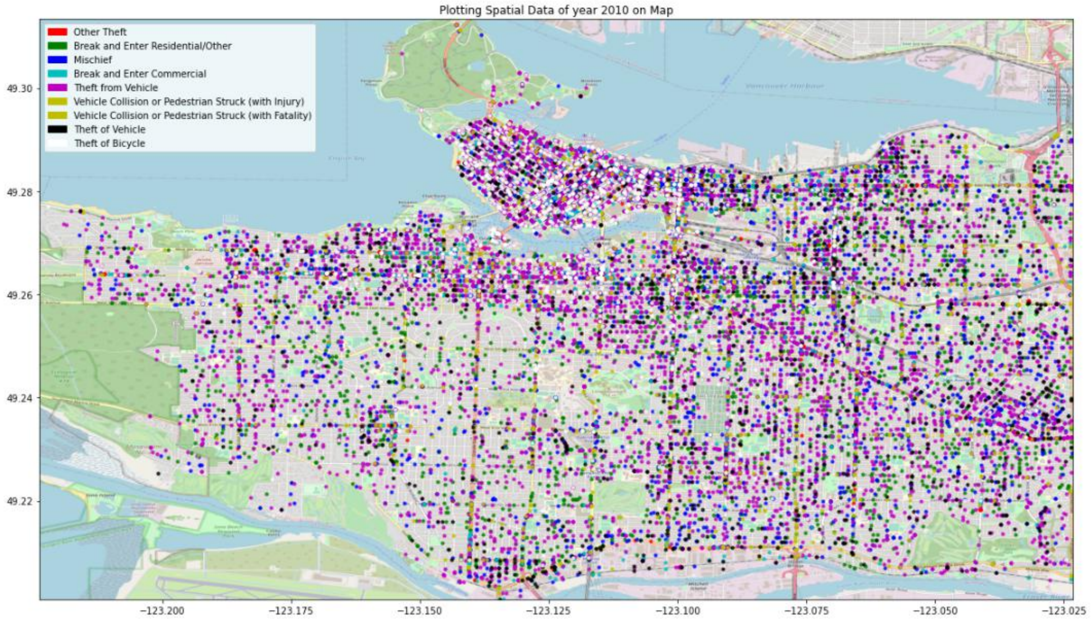

Plotting Spatial Data of year 2005 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

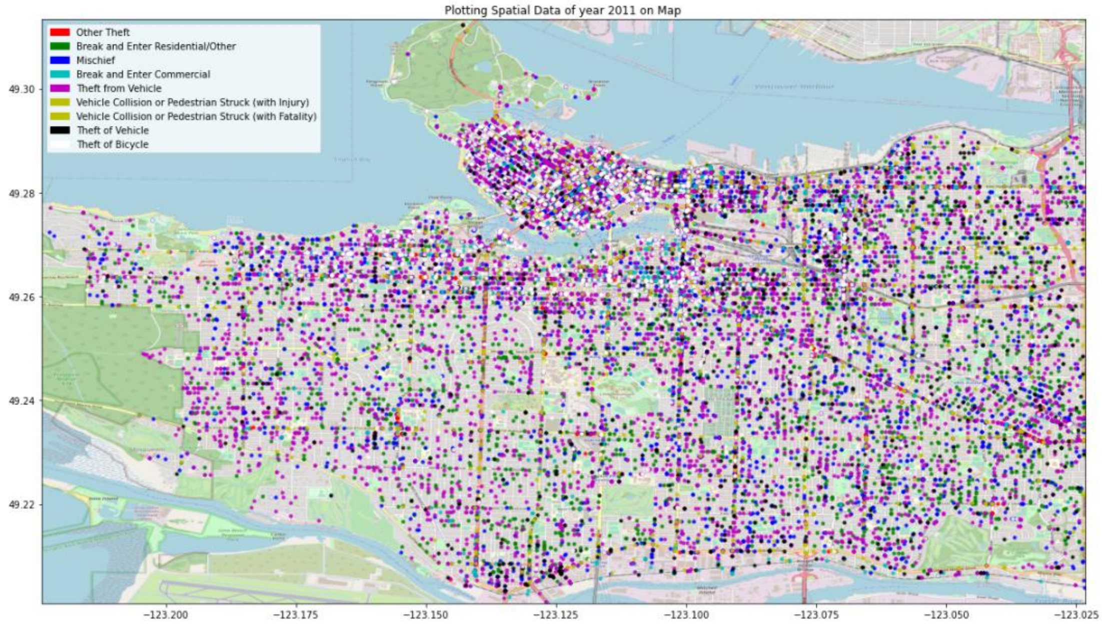

Plotting Spatial Data of year 2006 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

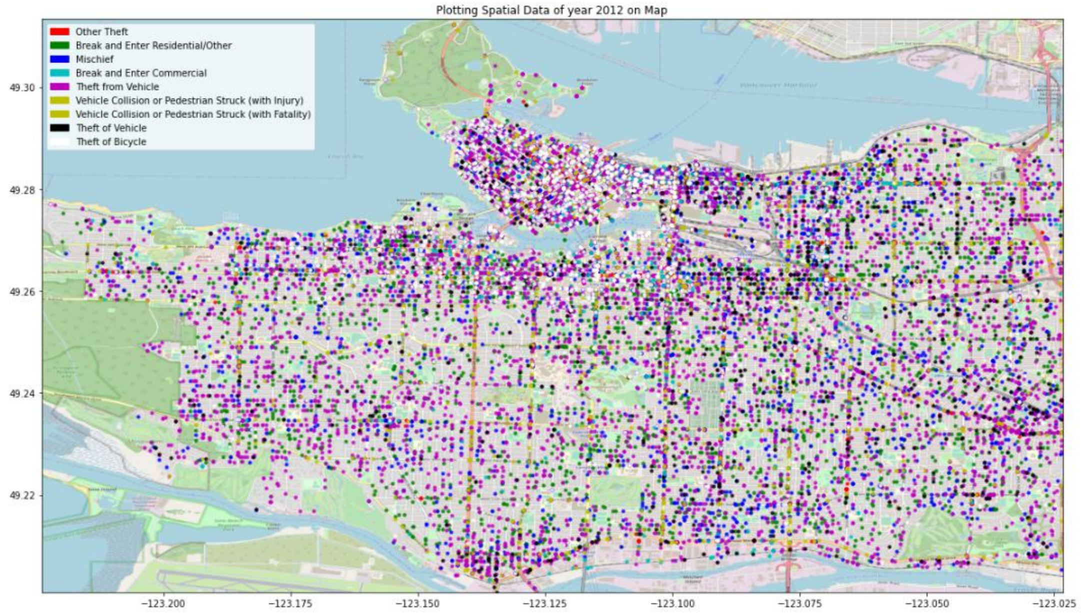

Plotting Spatial Data of year 2007 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

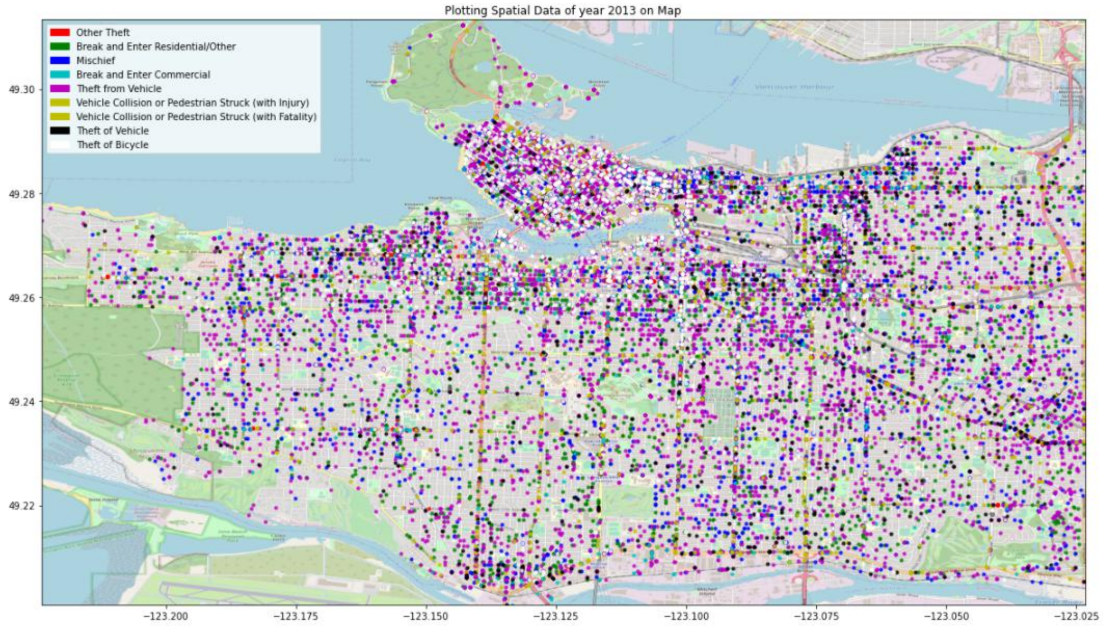

Plotting Spatial Data of year 2008 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

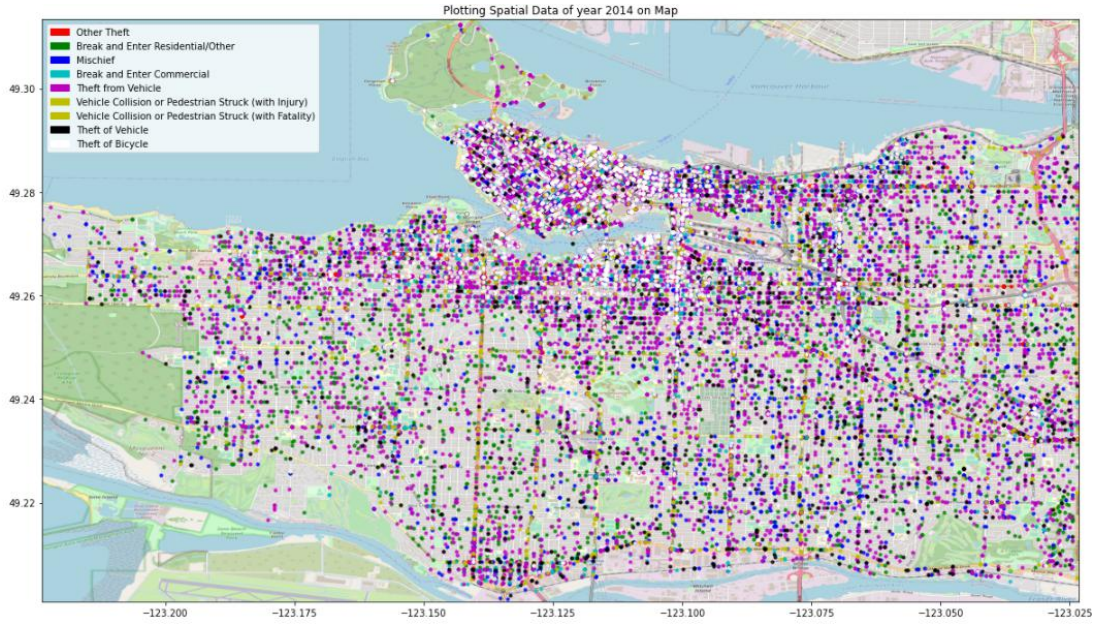

Plotting Spatial Data of year 2009 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

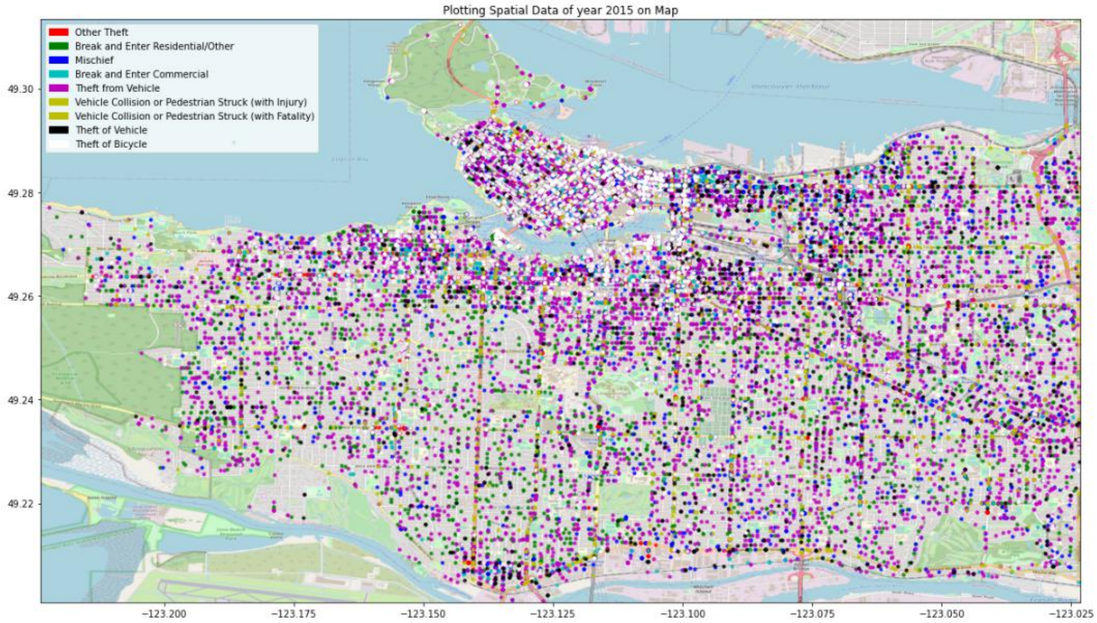

Plotting Spatial Data of year 2010 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

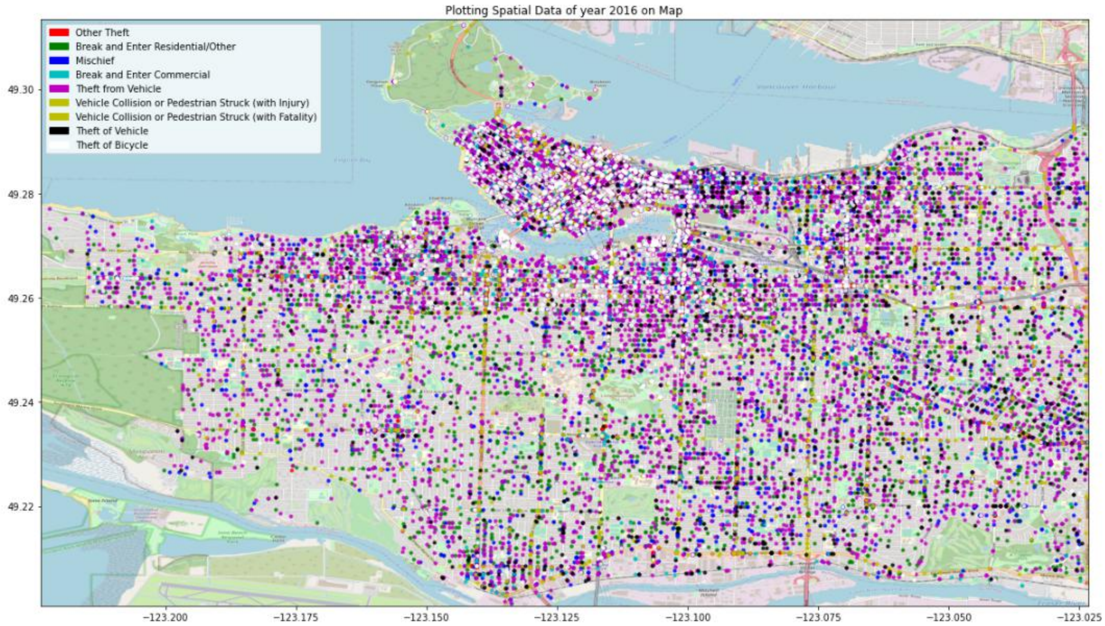

Plotting Spatial Data of year 2011 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025

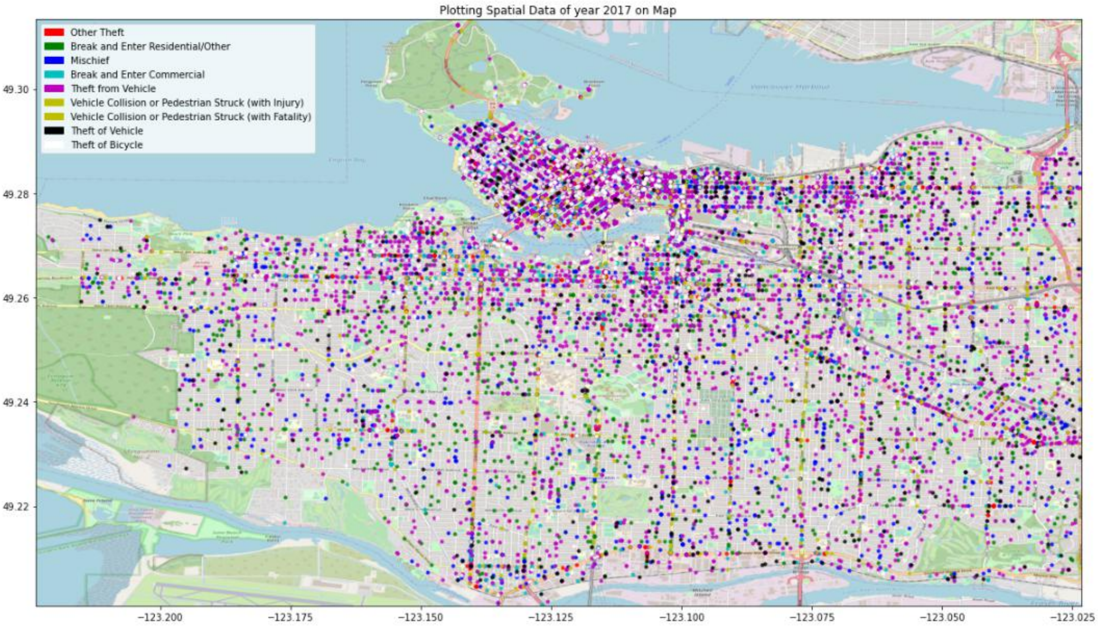

Plotting Spatial Data of year 2012 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2013 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2014 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2015 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2016 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025


Plotting Spatial Data of year 2017 on Map
Other Theft
Break and Enter Residential/Other
Mischief
Break and Enter Commercial
Theft from Vehicle
Vehicle Collision or Pedestrian Struck (with Injury)
Vehicle Collision or Pedestrian Struck (with Fatality)
Theft of Vehicle
Theft of Bicycle
49.30
49.28
49.26
49.24
49.22
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025