Crime Hotspots Throughout the Year

Joshua Luo, Parth Patel, and Thimira Wijepala

December 2, 2023

1 Introduction

1.1 Abstract

All sorts of crimes occur throughout the year ranging from homicide to theft and police forces are sent out to minimise these crimes. It is important to know where a majority of these crimes happen to avoid crime hotspots and to send police officers to help prevent these crimes. In this report, we investigate how the time of year affects crime rates in Vancouver. More specifically, we look into which neighbourhoods crimes occur in and what sort of crimes are being committed throughout the year.

1.2 Data Acquisition:

The data obtained for this report is from the Vancouver Police Department Open Data website (https://geodash.vpd.ca/opendata/#), where they contain CSV files about crimes in Vancouver. The dataset contains the list of crimes committed with the type of crime committed, when it occurred, and the location. Government data is usually considered fairly reliable and accurate.

2 Analysis

2.1 Data cleaning

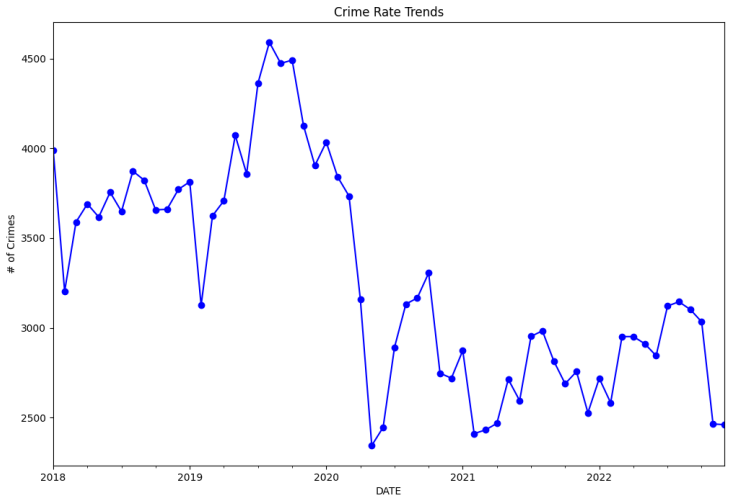
From the VPD Open Data dataset, we dropped the DAY, HOUR, MINUTE and HUNDRED_BLOCK columns since they were not of much use for our investigation. We also found there were many instances of NULL values in the NEIGHBOURHOOD, X, and Y columns. Since the NEIGHBOURHOOD values were crucial to our investigation and results, we dropped all instances where the NEIGHBOURHOOD value was NULL. Considering that we were not using the X and Y values to conclude results but only to visualise the map, we approximated missing X and Y values based on the NEIGHBOURHOOD of the crime.

We also added a new column to our dataset called SEASON to analyse seasonal crime trends (in addition to just the monthly ones) to gain a broader perspective, where each season is defined by its meteorological start date. For visualising the map, we needed the latitude and longitude values, but since our X

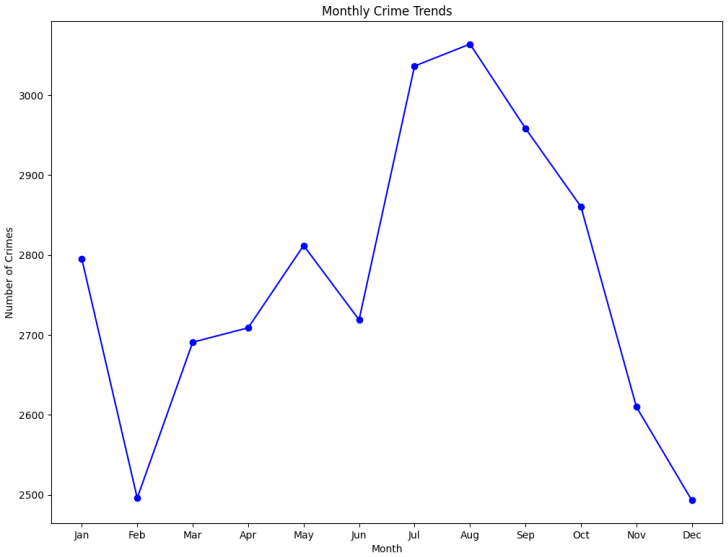
and Y values were coordinate values projected in UTM Zone 10, we converted our X and Y values to latitude (LAT) and longitude (LON) values. Seeing that we now have the new columns LAT and LON for precisely locating the crime spots, we dropped X and Y from the dataset.

Also, a lot of the categorical data in the plots like MONTHS, NEIGHBOURHOOD, and TYPES has been abbreviated so labels are much easier to read. Finally, our dataset is left with the columns YEAR, MONTH, SEASON, TYPE, NEIGHBOURHOOD, LAT, and LON and is also sorted in that same order.

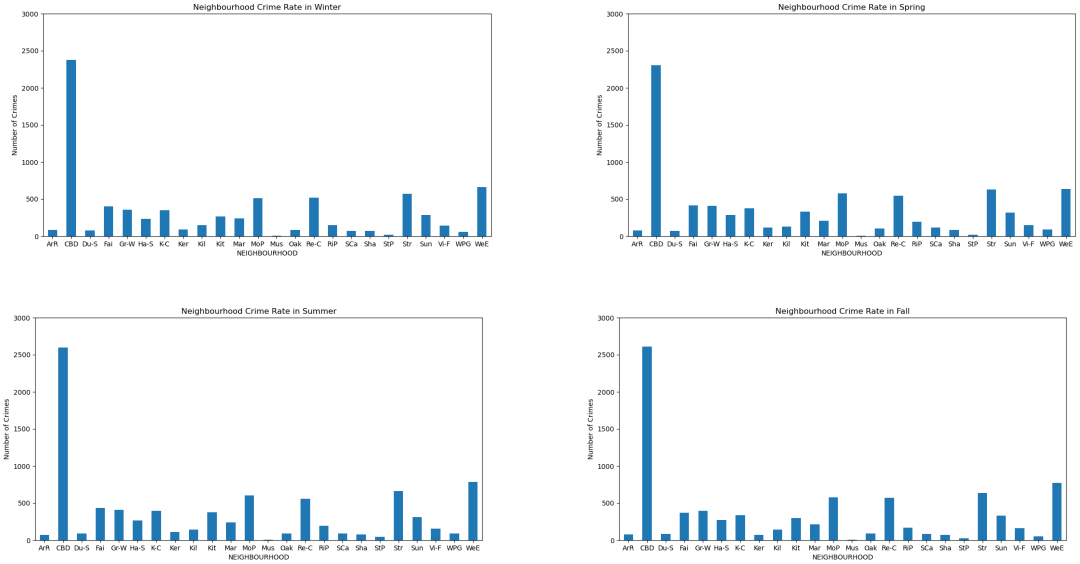
2.2 Data Visualization



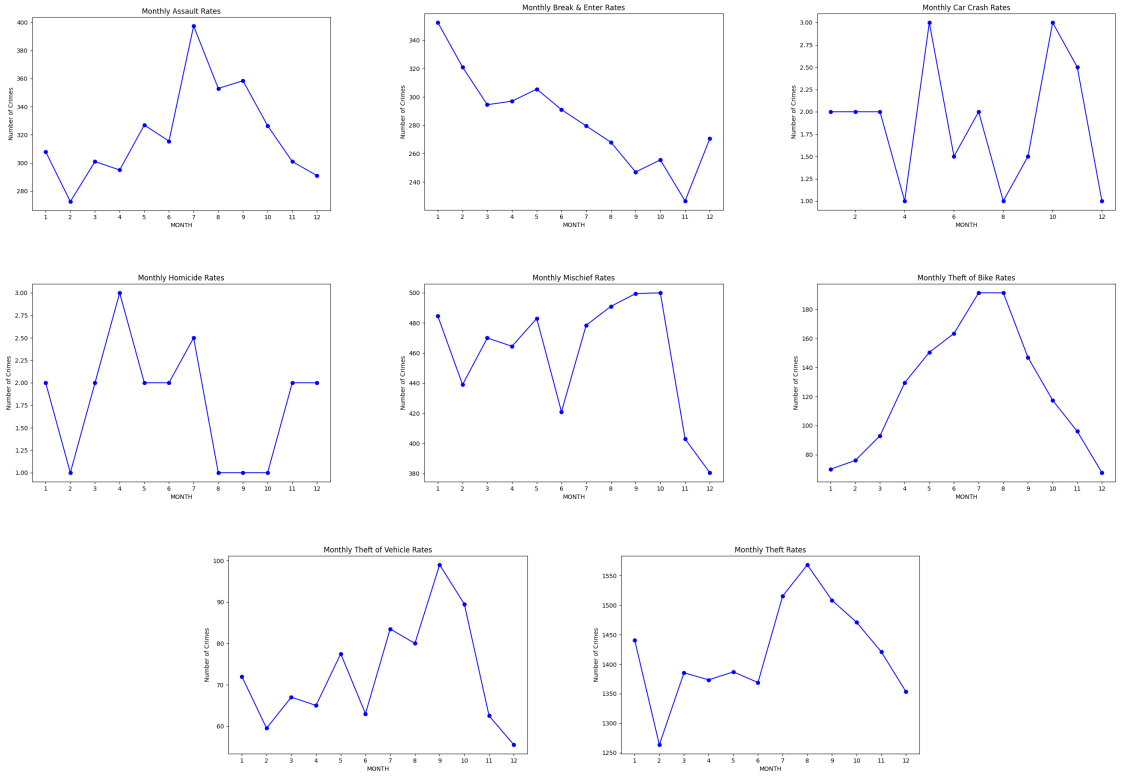
This figure shows the general crime rates from the start of 2018 to the end of 2022. As we can see from the graph, there is a sudden decline in crime after the first quarter of 2020, this is most likely due to the start of the COVID-19 Pandemic and the state of emergency declared on March 18th, 2020 till June 30th, 2020. People are in lockdown and staying at home. Since we want our results to show the trends of more recent data, we will base our results on 2021 and 2022 VPD crime data.



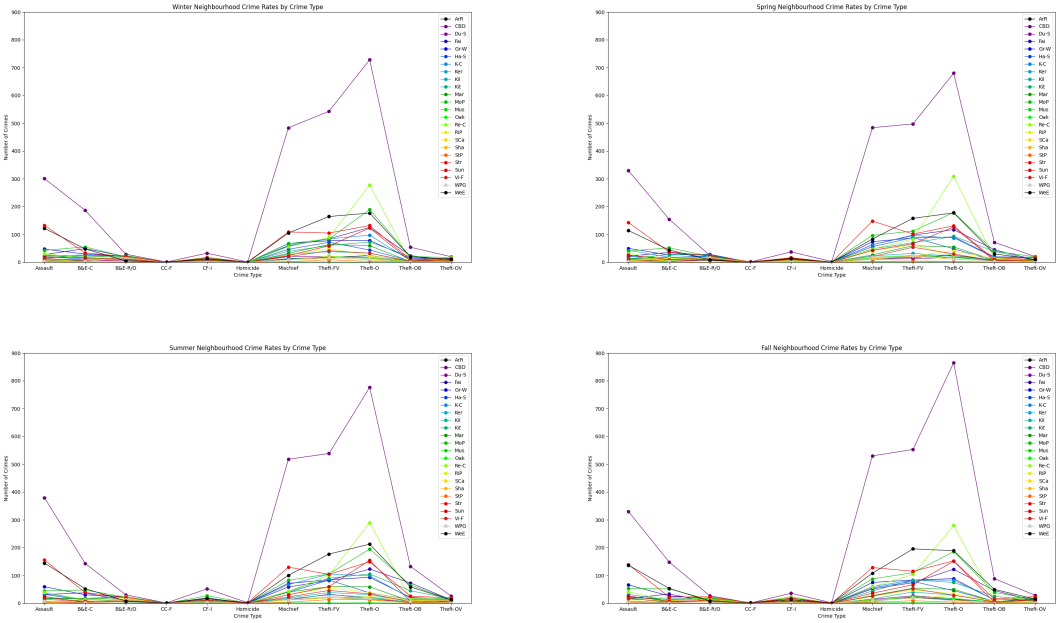
This figure shows a graph of the average number of crimes per month for the years 2021 and 2022. Looking at the graph we can see that there is a significant increase in crimes during July and August, with crimes surpassing the 3000 mark. Conversely, months like February and December showed subdued crime rates, with criminal activity staying below the 2500 threshold.



These figures show the crime rate in each neighbourhood for each season. We see that in each season there is a high crime rate in the Central Business District compared to others, but especially during the summer and fall, over 2500 crimes were committed in the Central Business District. While it is hard to see in these figures, there was also a significant increase in other neighbourhoods during Summer and Fall.



These figures show the monthly crimes committed for each crime type. As we can see, for crimes like Assault, Mischief, and Theft, there is a significant increase in crimes committed towards the end of summer and the start of fall when people go outside more often. While for crimes like Break and Enter, there is an increase in crimes committed during mid-winter. But for crimes like Car Crashes and Homicide, the crimes committed seem a bit more random and do not necessarily matter on the time of year.



These figures show the seasonal neighbourhood crime rates for each crime type. We can see that neighbourhood crime trends for each crime type are consistent across seasons. When it comes to neighbourhoods, the Central Business District has, on average, the most number of crimes committed across all crime types, whereas places like Musqueam and Stanley Park have significantly lower crime rates. With regards to crime types , we can also see a consistent pattern across all neighbourhoods with crimes like theft and mischief being the most committed throughout the year. On the other hand, all the neighbourhoods witnessed significantly less instances of homicides and fatal car crashes.

2.3 Inferential Tests

For inferential tests, we want to test if there is a change in the crimes committed throughout the year, which is also just testing if its dependent on the month. So we can use the Chi-Squared test to see if crimes committed are dependent on the month.

To test, if the number of crimes committed is dependent on the month, we can use the chi-square test to see if the distribution throughout the year is uniform.

MONTH	1	2	3	4	5	6	7	8	9	10	11	12
Number of Crimes	2795.5	2496.0	2691.0	2709.0	2812.0	2719.0	3036.5	3064.0	2958.5	2860.5	2610.0	2493.0

Running this data on the scipy.stats chisquare(), it tests the null hypothesis that the data is uniform by comparing it to an array where the categories are all equally likely. From the result of the chi-square test, we got the p-value, 6.21e-25, which implies that the number of crimes is not independent of the month.

To test if the number of crimes committed in each neighbourhood is dependent on the month we can use the scipy.stats chi2_contingency(), on the following contingency table.

NEIGHBOURHOOD MONTH	ArR	CBD	Du-S	Fai	Gr-W	Ha-S	K-C	Ker	Kil	Kit	Mar	MoP	Mus	Oak	Re-C	RiP	SCa	Sha	StP	Str	Sun	Vi-F	WPG	WeE
1	33.5	825.0	35.5	148.0	143.5	83.0	123.5	31.5	46.0	94.0	96.0	195.0	1.0	32.5	197.0	62.5	21.5	23.0	2.0	212.5	94.0	45.5	26.0	224.0
2	27.0	716.5	26.0	140.0	105.5	91.5	118.0	31.0	53.5	99.0	68.5	156.0	2.5	25.0	171.5	48.5	30.0	26.5	3.0	187.0	92.0	47.0	22.5	208.0
3	25.5	768.5	20.0	134.0	137.0	83.0	134.5	39.0	45.0	92.0	81.5	202.0	1.0	32.0	176.5	70.5	36.0	22.0	3.0	203.0	109.5	47.5	28.0	200.0
4	25.0	764.5	22.0	130.5	138.0	105.0	116.0	41.0	38.0	128.5	55.0	186.5	2.5	38.5	176.5	72.0	41.0	38.5	7.0	215.0	86.5	48.0	26.5	207.0
5	29.5	770.5	28.5	152.0	131.5	99.5	123.5	35.5	49.5	112.5	73.5	188.0	4.0	35.0	195.5	54.0	41.0	26.0	6.5	209.5	124.0	54.5	35.5	232.5
6	24.5	763.0	21.5	132.0	132.0	93.5	147.0	39.0	49.5	124.0	76.5	191.0	2.0	29.0	175.5	56.5	30.0	24.5	12.5	192.5	97.5	48.0	41.0	217.5
7	25.0	879.5	38.0	139.5	148.5	93.0	138.5	34.0	41.0	136.0	89.0	203.5	4.0	31.5	194.0	75.5	24.5	29.5	19.5	230.5	103.5	55.0	22.0	283.5
8	25.0	957.0	31.0	165.5	127.0	82.5	109.5	36.0	51.0	116.0	76.5	207.5	3.5	28.0	191.5	63.5	38.5	25.0	11.0	239.0	112.5	54.5	28.5	284.0
9	33.5	900.5	27.5	128.5	137.5	96.5	113.5	32.0	48.5	107.0	82.0	224.0	2.0	32.0	190.5	65.0	38.0	25.5	7.5	217.0	111.0	53.5	19.5	266.0
10	24.5	881.5	31.5	134.0	134.0	87.5	122.0	26.5	43.5	108.5	70.0	186.0	3.0	34.5	181.0	51.0	24.5	25.5	10.0	228.0	118.5	56.5	17.5	261.0
11	19.0	827.5	25.0	108.0	127.0	87.5	99.5	16.0	49.5	86.0	63.0	166.0	2.5	21.5	197.5	56.0	25.0	20.0	7.0	194.0	103.0	50.5	16.0	243.0
12	24.5	838.0	19.0	115.0	107.5	57.5	108.0	27.5	48.0	74.0	78.0	162.0	1.5	24.5	152.0	40.5	19.5	22.5	12.5	169.5	99.0	48.5	11.5	232.5

This function tests if the crimes committed in each neighbourhood are independent of the month. From the test, we got the p-value of 7.55e-4, which implies the number of crimes committed in each neighbour is not independent of the month.

To test if the number of each crime type is dependent on the month, we once again use the scipy.stats chi2_contigency(), but now on a different contingency table.

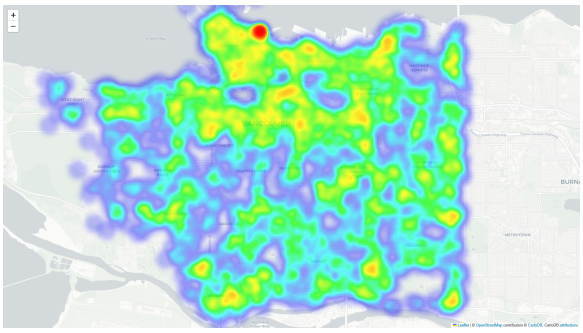
TYPE MONTH	Assault	B&E-C	B&E-R/O	CC-F	CF-I	Homicide	Mischief	Theft-FV	Theft-O	Theft-OB	Theft-OV
1	308.0	231.0	121.5	2.0	65.5	2.0	484.5	652.0	789.0	70.0	72.0
2	272.5	194.0	127.0	2.0	62.0	1.0	439.0	553.0	710.5	76.0	59.5
3	301.0	177.0	117.5	2.0	77.0	2.0	470.0	621.5	764.0	93.0	67.0
4	295.0	170.0	127.0	1.0	82.0	3.0	464.5	596.5	777.0	129.5	65.0
5	327.0	167.0	138.5	3.0	76.5	2.0	483.0	602.0	785.0	150.5	77.5
6	315.5	163.0	128.0	1.5	92.5	2.0	421.0	601.0	768.0	163.5	63.0
7	397.5	153.5	126.0	2.0	86.0	2.5	478.5	632.5	883.0	191.5	83.5
8	353.0	166.5	101.5	1.0	111.0	1.0	491.0	678.5	890.0	191.5	80.0
9	358.5	158.0	89.0	1.5	97.0	1.0	499.5	698.0	810.5	147.0	99.0
10	326.5	149.0	106.5	3.0	97.0	1.0	500.0	642.5	828.5	117.5	89.5
11	301.0	138.0	88.5	2.5	95.5	2.0	403.0	565.0	856.0	96.0	62.5
12	291.0	164.5	106.0	1.0	73.0	2.0	380.5	546.5	807.0	67.5	55.5

On this new contingency table, the function tests if the crime types are independent of the month. From the test, we get the p-value 1.60e-21, which implies that the number of each crime type committed is not independent of the month.

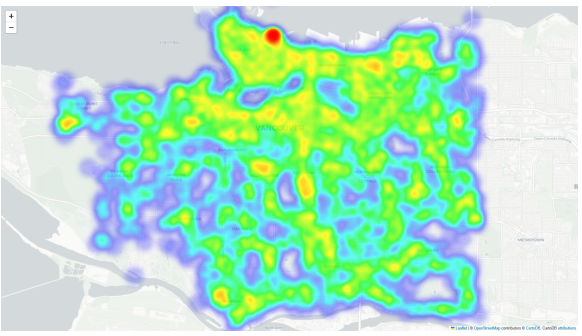
3 Conclusions

From the inferential tests, we know that there is a significance between the time of year and the crimes being committed, from the number of crimes to where and what crime is being committed. From the inferential test results, we can start to make conclusions about the data.

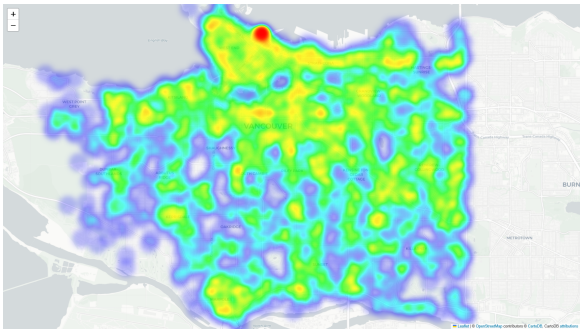
From the graphs, we see that there is an increased crime rate in late summer and early fall. The majority of crimes happen in the Central Business District throughout the year with also an increase in crime rate in late summer and early fall. We can also see where some of these crime hotspots are from heat maps that were made.



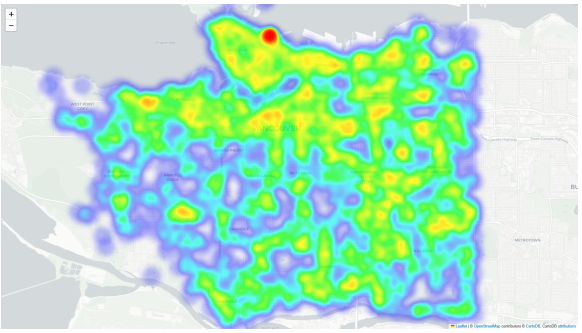
Winter Heatmap



Spring Heatmap



Summer Heatmap



Fall Heatmap

From the crime-type trends, we also notice that during the end of summer and early spring there is an increase in crimes like Assault, Mischief, and Theft being committed. While during mid-winter, there is an increase in Break & Enters being committed. We also notice that the majority of crimes being committed are mischief and theft which usually occur in the Central Business District.

4. Limitations

One of the biggest limitations is the amount of data and the span of years. Since we decided to use more recent data (post-covid data) to look at where current crimes are being committed, it was harder to build classifiers that could build decent predictions.

Also since our data was primarily categorical, there were not many inferential tests that we could use so we had to stick with the chi-squared tests to test if our data was independent or not.

Accomplishment Statements:

Joshua Luo

- Trained machine learning models to predict the type of crime based on location and time with ~30% accuracy.
- Created heatmaps showing density of crime based on time of the year.

Parth Patel

- Created multiple plots number of crimes types committed and crime types committed in each neighbourhood over each season.
- Made approximations for X & Y values by using the most frequent values to replace missing values

Thimira Wijepala

- Made chi-squared tests for the number of overall crimes committed over months, crimes committed in each neighbourhood over the months, and each crime type committed over the months.
- Created multiple plots to show monthly trends in crimes, from trends in overall crime rate to trends for each crime type.