CRIME DATA ANALYSIS

- DATA CLEANING
- DATA ANALYSIS
- DATA VISUALIZATION

TEAM MEMBERS :

- SWATI MEHTA (3765212)
- JASKARAN SINGH(3761860)

Introduction

Crime affects the safety and well-being of people and communities across Canada. However, the types and rates of crime can be very different depending on the city. In this project, we focus on two major cities—Toronto and Vancouver—to study and compare crime trends from 2020 to 2024.The goal of this project is to understand how crime varies between these cities and how factors like income, education, and employment may influence crime rates. By analyzing open data from government sources, we explore which crimes are most common, when they happen most often, and where they occur.

To carry out this analysis, we used several Python libraries such as Pandas, NumPy, Matplotlib, and ThinkStats. These tools helped us clean and organize the data, perform statistical analysis, and create clear visualizations. We used bar charts, line graphs, and heatmaps to highlight crime patterns across different neighborhoods, times of the year, and property types. By combining data analysis with visual storytelling, we aim to present insights that are both informative and easy to understand.

In conclusion, this project provides a detailed look into the crime patterns of Toronto and Vancouver, highlighting both differences and similarities in their crime rates and related social factors. By using data and analytical tools, we were able to uncover valuable insights about when and where crimes occur, and how they may relate to factors like income, education, and employment.

Related Work

Crime analysis has been explored in many previous studies using open data from Canadian cities. Researchers often focus on finding patterns in crime types, locations, and timings to support better decision-making .Many of these studies also look at how social factors like income, education, and employment affect crime rates.

In our project, we worked with large datasets covering crime data from 2014 to 2024 for both Toronto and Vancouver. Since the raw data was very detailed and included thousands of records, we applied data cleaning techniques to remove duplicates, handle missing values, and filter out irrelevant information. This helped us prepare the data for accurate analysis.

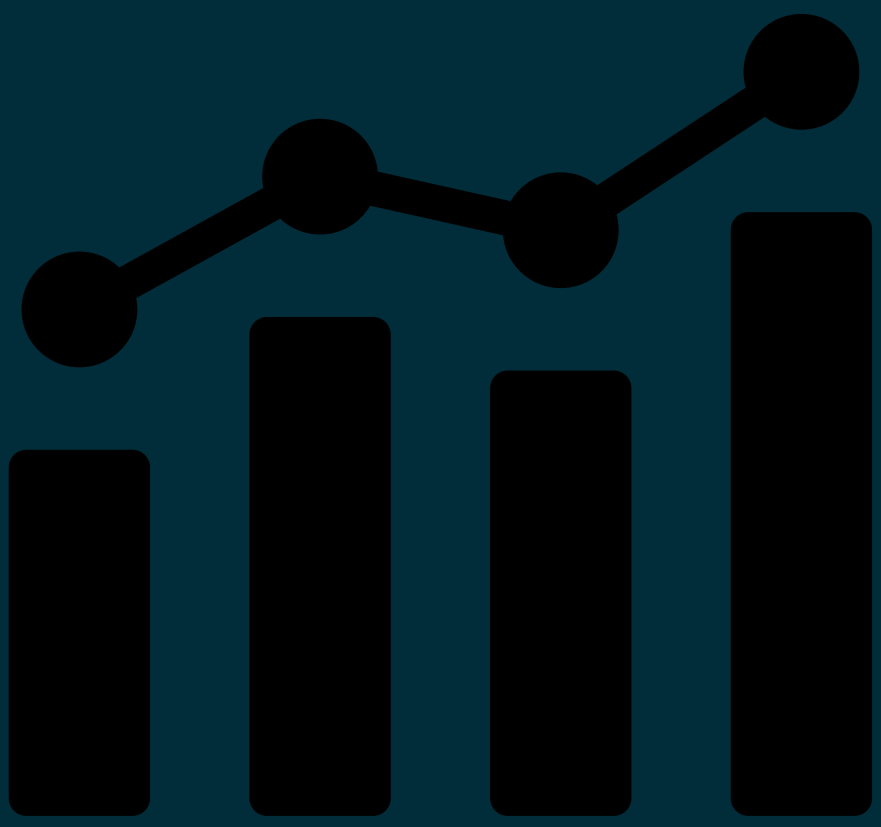
Using tools like Pandas, NumPy, Matplotlib, and ThinkStats, we explored key patterns such as the most and least common crimes, seasonal trends, and peak times of criminal activity. Unlike some past studies that focus only on a single year or city, our work provides a broader 10-year comparison between two major Canadian cities, making our analysis both comprehensive and insightful

Problem Statement

CRIME SCENE - DO NOT CROSS CRIME SCENE - DO NOT CROSS

This project focuses on conducting a comparative crime analysis between Toronto and Vancouver, using a data (2014–2024) to cover deep insights into crime patterns. Our goal is to answer key questions such as:

- What are the most common and least common types of crimes in both cities?
- What is the total number of crimes reported up to 2025, and how have these numbers changed over time?
- During which months or seasons is crime most frequent?
- Is there a relationship between crime occurrences and the time of day?
- What types of properties—such as houses, apartments, or commercial spaces—are most affected by specific crimes like break-and-enter, assault, and robbery?
- What are the income and education levels in neighborhoods where high crime rates are observed?



Project Requirements

To successfully carry out the project, the following requirements were identified:

Crime Data: Crime records from 2014 to 2024 for Toronto and Vancouver, were taken from open government portals and trusted datasets (e.g., Kaggle, Toronto Police Open Data, Vancouver Open Data).
Socioeconomic Data: While neighborhood-level socioeconomic datasets (including income, education, and housing) were initially planned for analysis, the data available was incomplete or outdated. To overcome this, we identified the top 10 high-crime neighborhoods in both Toronto and Vancouver. From these, we selected the top 3 locations in each city and manually researched their living conditions using reliable open data sources and government reports.

Data Requirements

- Import and clean large datasets, handling missing, inconsistent, or duplicate data entries.
- Analyze crime frequency by:
- Type of crime (e.g., theft, assault, robbery)
- Time of year and time of day
- Property type (e.g., residential, commercial)
- Identify trends and seasonal patterns over the 10-year period.
- Perform comparative analysis between Toronto and Vancouver.
- Generate meaningful visualizations

Functional Requirements

Libraries and Tools:

- Pandas – for data manipulation and analysis
- NumPy – for numerical operations
- Matplotlib – for data visualization
- ThinkStats – for statistical exploration

Data Storage: CSV format datasets

- Development Environment: Jupyter Notebook

Technical Requirements

Approach & Analysis

The key techniques and methods used throughout the project:

1. Data Loading

[1.1] We started the project by importing two large datasets—one for Vancouver and one for Toronto—using the Pandas library. The files contained detailed crime records over multiple years. Due to the large size of the datasets (several megabytes), there were performance issues during loading

```
import pandas as pd

# Load Vancouver and Toronto crime data from CSV files into DataFrames
df_vancouver = pd.read_csv('VancouverFinal.csv')
df_toronto   = pd.read_csv('Toronto.csv')

# As the dataset is huge so we are displaying the first 5 rows of each DataFrame
display(df_vancouver.head())
display(df_toronto.head())
```

[1.2] Once the data was loaded, we used the head() function to view the first few rows and confirm that the datasets were loaded correctly. This allowed us to verify the structure and contents of the files. It also helped us identify key columns such as crime type, location, and date.

Result

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Break and Enter Commercial	2012	12	14	8	52	NaN	Oakridge	491285.0000	5.453433e+06
1	Break and Enter Commercial	2019	3	7	2	6	10XX SITKA SQ	Fairview	490612.9648	5.457110e+06
2	Break and Enter Commercial	2019	8	27	4	12	10XX ALBERNI ST	West End	491004.8164	5.459177e+06
3	Break and Enter Commercial	2021	4	26	4	44	10XX ALBERNI ST	West End	491007.7798	5.459174e+06
4	Break and Enter Commercial	2014	8	8	5	13	10XX ALBERNI ST	West End	491015.9434	5.459166e+06

	OBJECTID	EVENT_UNIQUE_ID	REPORT_DATE	OCC_DATE	REPORT_YEAR	REPORT_MONTH	REPORT_DAY	REPORT_DOY	REPORT_DOW	REPORT_HOUR	...	OFFENCE	MCI_CATEGORY	HOOD_158
0	1	GO-20141263217	1/1/2014 5:00:00 AM	12/31/2013 5:00:00 AM	2014	January	1	1	Wednesday	16	...	Theft Of Motor Vehicle	Auto Theft	043
1	2	GO-20141260715	1/1/2014 5:00:00 AM	1/1/2014 5:00:00 AM	2014	January	1	1	Wednesday	3	...	Assault	Assault	092
2	3	GO-20141260730	1/1/2014 5:00:00 AM	1/1/2014 5:00:00 AM	2014	January	1	1	Wednesday	3	...	Assault	Assault	105
3	4	GO-20141260597	1/1/2014 5:00:00 AM	1/1/2014 5:00:00 AM	2014	January	1	1	Wednesday	2	...	Assault	Assault	080
4	5	GO-20141259762	1/1/2014 5:00:00 AM	12/31/2013 5:00:00 AM	2014	January	1	1	Wednesday	2	...	Assault	Assault	164

5 rows × 31 columns

2. Data Cleaning

[2.1] After loading the datasets for Toronto and Vancouver, we performed a series of cleaning and other steps to ensure the data was accurate, consistent, and ready for analysis.

[2.2] Toronto Dataset

- First, we converted the 'OCC_DATE' column to datetime format to allow for time-based filtering .
- We then filtered the data to include only crimes that occurred between 2020 and 2024.
- Some new columns were created to include specific components such as Date, Time, and Year from the datetime values.
- Then we selected only relevant columns such as the type of offence, neighbourhood, division, and coordinates.
- For better understanding, we renamed several columns (e.g., MCI_CATEGORY to Category, LAT_WGS84 to Latitude and many more).
- We removed any duplicate or missing entries to ensure better data .
- Finally, we separated the dataset into two parts: one containing crime-related details (toronto_main) and another containing location data (toronto_location)

```
# ---------------- Vancouver ----------------
df_vancouver = pd.read_csv('VancouverFinal.csv')
print("Cleaning and filtering Vancouver dataset:")

# Create date columns
df_vancouver['DATE'] = pd.to_datetime(df_vancouver[['YEAR', 'MONTH', 'DAY']])
df_vancouver = df_vancouver[(df_vancouver['DATE'].dt.year >= 2020) & (df_vancouver['DATE'].dt.year <= 2024)]

# Droping all nulls and duplicates values
df_vancouver = df_vancouver.dropna(subset=['DATE', 'TYPE', 'NEIGHBOURHOOD'])
df_vancouver = df_vancouver.drop_duplicates()
df_vancouver.index.name = 'Index'
df_vancouver['YEAR'] = df_vancouver['YEAR'].astype(int)
df_vancouver['HOUR'] = df_vancouver['HOUR'].astype(int)

# Split into two different tables
vancouver_main = df_vancouver[['TYPE', 'YEAR', 'MONTH', 'DAY', 'HOUR', 'MINUTE', 'HUNDRED_BLOCK', 'NEIGHBOURHOOD',
                               'DATE']]
vancouver_location = df_vancouver[['X', 'Y']]

print("\nFirst 10 rows of Vancouver's Crime:\n")
print(vancouver_main.head(10).to_string())
print("\nDisplaying location - (X and Y):\n")
print(vancouver_location.head(10).to_string())
```

```
First 10 rows of Toronto's Crime :

            Date      Time  Year Division                    Neighbourhood                      Offence
Category
Index
211487  2020-01-01  05:00:00  2020      D41                 Eglinton East (138)                      Assault
Assault
211488  2020-01-01  05:00:00  2020      D43                   Morningside (135)  Unlawfully In Dwelling-House  Br
eak and Enter
211489  2020-01-01  05:00:00  2020      D23                 Rexdale-Kipling (4)                      Assault
Assault
211490  2020-01-01  05:00:00  2020      D51  Cabbagetown-South St.James Town (71)          Assault With Weapon
Assault
211491  2020-01-01  05:00:00  2020      D11     Runnymede-Bloor West Village (89)                      B&E  Br
eak and Enter
211493  2020-01-01  05:00:00  2020      D55              Greenwood-Coxwell (65)          Robbery - Mugging
Robbery
211496  2020-01-01  05:00:00  2020      D41           Birchcliffe-Cliffside (122)                      Assault
Assault
211498  2020-01-01  05:00:00  2020      D52             Yonge-Bay Corridor (170)          Robbery - Mugging
Robbery
211499  2020-01-01  05:00:00  2020      D55              Danforth East York (59)       Theft Of Motor Vehicle
Auto Theft
211503  2020-01-01  05:00:00  2020      D42          Tam O'Shanter-Sullivan (118)                      Assault
Assault

Displaying Location - (Latitude and Longitude):

         Latitude  Longitude
Index
211487  43.734555 -79.258136
211488  43.775996 -79.212020
211489  43.716804 -79.566790
211490  43.662332 -79.367179
211491  43.656097 -79.487243
211493  43.678624 -79.327902
211496  43.701525 -79.252928
211498  43.658675 -79.381912
211499  43.692962 -79.333390
211503  43.780512 -79.300396
```

[2.3] Vancouver Dataset

- The original dataset had columns for YEAR, MONTH, and DAY, which were combined into a single DATE column and converted to datetime format and then was filtered
- Rows with missing or null values in key fields like DATE, NEIGHBOURHOOD, and TYPE were removed, duplicate values were removed .
- The cleaned dataset was split into two parts:
  - vancouver_main – crime-related details.
  - vancouver_location – containing location

```
# ---------------- Toronto ----------------
df_toronto = pd.read_csv('Toronto.csv')

print("Cleaning and filtering Toronto dataset: ")

# Converting date column to actual date format
df_toronto['OCC_DATE'] = pd.to_datetime(df_toronto['OCC_DATE'])
df_toronto = df_toronto[(df_toronto['OCC_DATE'].dt.year >= 2020) & (df_toronto['OCC_DATE'].dt.year <= 2024)]
df_toronto['Date'] = df_toronto['OCC_DATE'].dt.date
df_toronto['Time'] = df_toronto['OCC_DATE'].dt.time
df_toronto['Year'] = df_toronto['OCC_DATE'].dt.year

df_toronto = df_toronto[['Date', 'Time', 'Year', 'DIVISION', 'NEIGHBOURHOOD_158', 'OFFENCE', 'MCI_CATEGORY',
                         'LAT_WGS84', 'LONG_WGS84']]
cols = df_toronto.columns.tolist()

# Modify specific columns names to improve readability
cols[cols.index('DIVISION')] = 'Division'
cols[cols.index('NEIGHBOURHOOD_158')] = 'Neighbourhood'
cols[cols.index('OFFENCE')] = 'Offence'
cols[cols.index('MCI_CATEGORY')] = 'Category'
cols[cols.index('LAT_WGS84')] = 'Latitude'
cols[cols.index('LONG_WGS84')] = 'Longitude'
df_toronto.columns = cols

# Remove all rows with missing/(NaN) values
df_toronto = df_toronto.dropna()
df_toronto = df_toronto.drop_duplicates()
df_toronto.index.name = 'Index'

toronto_main = df_toronto[['Date', 'Time', 'Year', 'Division', 'Neighbourhood', 'Offence', 'Category']]
toronto_location = df_toronto[['Latitude', 'Longitude']]

print("\nFirst 10 rows of Toronto's Crime :\n")
print(toronto_main.head(10).to_string())
```

```
Cleaning and filtering Vancouver dataset:

First 10 rows of Vancouver's Crime:

                              TYPE  YEAR  MONTH  DAY  HOUR  MINUTE    HUNDRED_BLOCK NEIGHBOURHOOD       DATE
Index
3       Break and Enter Commercial  2021      4   26     4      44  10XX ALBERNI ST      West End 2021-04-26
5       Break and Enter Commercial  2020      7   28    19      12  10XX ALBERNI ST      West End 2020-07-28
6       Break and Enter Commercial  2021     11   21     6      33  10XX ALBERNI ST      West End 2021-11-21
7       Break and Enter Commercial  2022      1    5     7      34  10XX ALBERNI ST      West End 2022-01-05
14      Break and Enter Commercial  2022      1    3    16      19  10XX ALBERNI ST      West End 2022-01-03
24      Break and Enter Commercial  2020      6   19     3      40  10XX ALBERNI ST      West End 2020-06-19
25      Break and Enter Commercial  2023      9   14     3      30  10XX ALBERNI ST      West End 2023-09-14
27      Break and Enter Commercial  2022      6   17     5      16  10XX ALBERNI ST      West End 2022-06-17
35      Break and Enter Commercial  2020      1    3     6      43  10XX ALBERNI ST      West End 2020-01-03
42      Break and Enter Commercial  2020      9   27    20       0  10XX ALBERNI ST      West End 2020-09-27

Displaying location - (X and Y):

                X             Y
Index
3       491007.7798  5.459174e+06
5       491015.9434  5.459166e+06
6       491015.9434  5.459166e+06
7       491015.9434  5.459166e+06
14      491036.0799  5.459146e+06
24      491059.4824  5.459122e+06
25      491065.2962  5.459130e+06
27      491067.3433  5.459115e+06
35      491068.6829  5.459126e+06
42      491073.0884  5.459109e+06
```

3. Data Wrangling

[3.1] Key Steps:

- A new column called city was added to both datasets to clearly distinguish between Toronto and Vancouver records.
- We changed column names like TYPE, MCI_CATEGORY, YEAR, and NEIGHBOURHOOD in both datasets so that they use the same names, making it easier to combine and compare the data
  For example, TYPE in Vancouver and MCI_CATEGORY in Toronto were both renamed to Category
- NEIGHBOURHOOD and NEIGHBOURHOOD_158 were both renamed to Neighbourhood.
- We then selected only the common and relevant columns: Category, Year, Neighbourhood, and city.
- Finally, the two datasets were combined into a single DataFrame using pd.concat() for easier analysis across both cities.

```python
# Data Wrangling
df_toronto['city'] = 'Toronto'
df_vancouver['city'] = 'Vancouver'

# Modify specific columns names to improve readability
vancouver_columns = df_vancouver.columns.tolist()
if 'TYPE' in vancouver_columns:
    vancouver_columns[vancouver_columns.index('TYPE')] = 'Category'

if 'NEIGHBOURHOOD' in vancouver_columns:
    vancouver_columns[vancouver_columns.index('NEIGHBOURHOOD')] = 'Neighbourhood'

if 'YEAR' in vancouver_columns:
    vancouver_columns[vancouver_columns.index('YEAR')] = 'Year'

df_vancouver.columns = vancouver_columns

# Modify specific columns names to improve readability
toronto_columns = df_toronto.columns.tolist()
if 'MCI_CATEGORY' in toronto_columns:
    toronto_columns[toronto_columns.index('MCI_CATEGORY')] = 'Category'

if 'NEIGHBOURHOOD_158' in toronto_columns:
    toronto_columns[toronto_columns.index('NEIGHBOURHOOD_158')] = 'Neighbourhood'

if 'YEAR' in toronto_columns:
    toronto_columns[toronto_columns.index('YEAR')] = 'Year'

df_toronto.columns = toronto_columns
common_columns = ['Category', 'Year', 'Neighbourhood', 'city']

df_toronto_selected = df_toronto[common_columns]
df_vancouver_selected = df_vancouver[common_columns]

# Combining the Two Datasets
df_combined = pd.concat([df_toronto_selected, df_vancouver_selected])
```

```
Toronto Data:
              Category  Year                          Neighbourhood     city
Index
211487         Assault  2020                    Eglinton East (138)  Toronto
211488  Break and Enter  2020                      Morningside (135)  Toronto
211489         Assault  2020                    Rexdale-Kipling (4)  Toronto
211490         Assault  2020  Cabbagetown-South St.James Town (71)  Toronto
211491  Break and Enter  2020     Runnymede-Bloor West Village (89)  Toronto

Vancouver Data:
                        Category  Year Neighbourhood       city
Index
3      Break and Enter Commercial  2021      West End  Vancouver
5      Break and Enter Commercial  2020      West End  Vancouver
6      Break and Enter Commercial  2021      West End  Vancouver
7      Break and Enter Commercial  2022      West End  Vancouver
14     Break and Enter Commercial  2022      West End  Vancouver

 Record Counts by City:
Toronto      186946
Vancouver    165306
Name: city, dtype: int64
```

4. Standardize and Group Crime Types

[4.1] Key Steps:

- We created a dictionary (crime_mapping) that matches crime types (like Assault, Auto Theft, or Break and Enter) to a crime group (e.g., Violent Crime or Property Crime).
  For example:
  - Assault, Homicide, and Robbery were all grouped under "Violent Crime".
  - Auto Theft, Theft from Vehicle, and Break and Enter grouped under "Property Crime".
- Traffic-related cases were grouped under "Traffic Incident", and crimes like Mischief were labeled as "Other Crime".
- We then applied this mapping to the combined dataset using the .map() function, creating a new column called standard_crime.

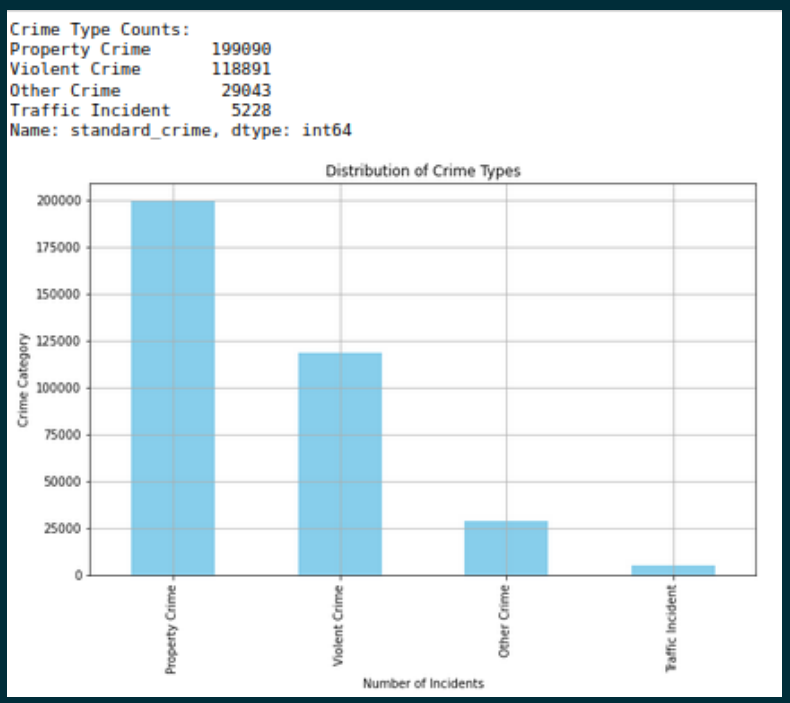
```
Sample standardized crime categories:
                                        Category    standard_crime
0                                        Assault     Violent Crime
1                                Break and Enter    Property Crime
5                                        Robbery     Violent Crime
8                                     Auto Theft    Property Crime
133                                   Theft Over    Property Crime
186946             Break and Enter Commercial       Property Crime
196952          Break and Enter Residential/Other    Property Crime
203960                                   Homicide     Violent Crime
```

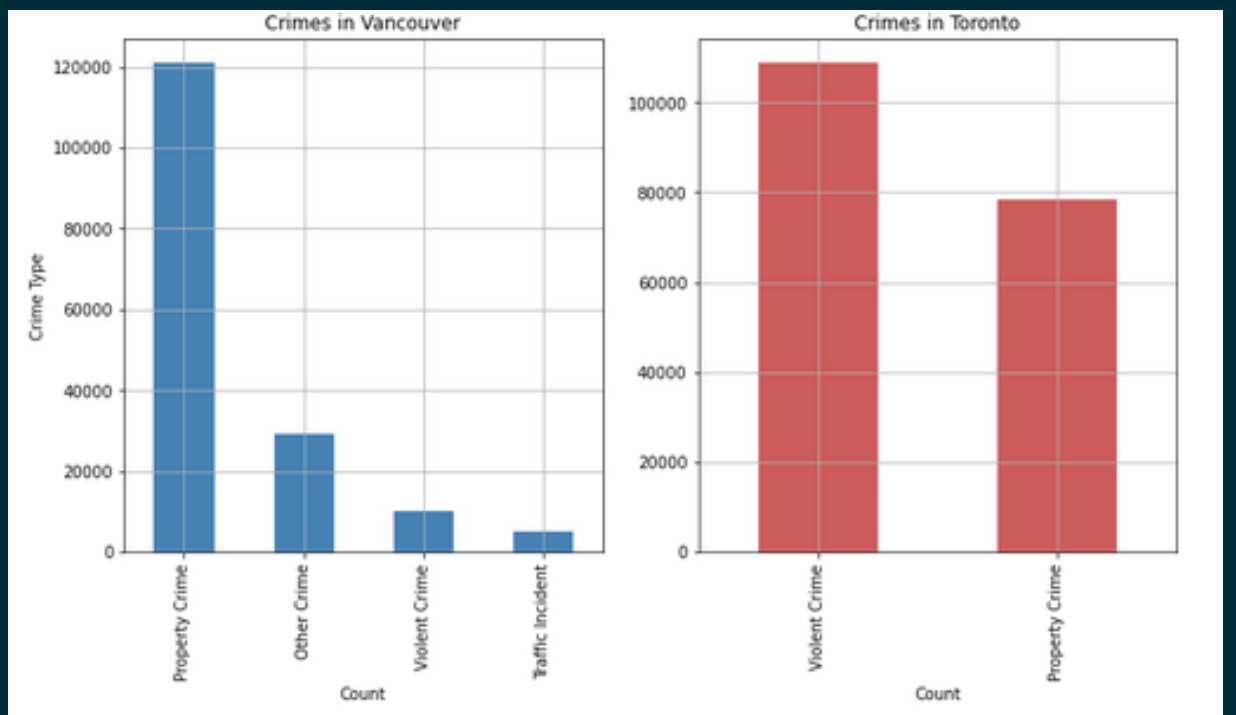
5. Data Visualization

After standardizing the crime types into broader categories (like Violent Crime, Property Crime, etc.), we plotted how these categories were distributed across Toronto and Vancouver.

- We used value_counts() to count how many times each standard crime type appeared in the combined dataset.
- We then created a horizontal bar chart using Matplotlib .
- To compare the crime patterns between Vancouver and Toronto, we filtered the dataset by city and counted the number of crimes for each standard crime type.
- We then created side-by-side bar charts to visualize how crime categories differ between the two cities.

```
Crime Type Counts:
Property Crime      199090
Violent Crime       118891
Other Crime          29043
Traffic Incident      5228
Name: standard_crime, dtype: int64
```





Conclusion

- **Property Crime** is the most common crime category in **Vancouver** showing significantly higher property crime rates compared to Toronto. In Vancouver, property crimes accounts for more than 120,000 incidents.
- **Violent Crime** is the leading crime category in **Toronto,** surpassing property crimes. This indicates a different crime pattern where Toronto has a higher occurrence of personal offenses like assault, robbery, and homicide.
- Other Crimes (such as mischief) and Traffic Incidents appear less frequently in both cities, though **Vancouver shows slightly more incidents in these categories than Toronto**.

When combining both cities:

- Property Crime accounts for nearly 200,000 incidents, making it the most widespread crime type.
- Violent Crime follows with around 119,000 incidents.
- Other Crime and Traffic Incident categories are much less common.

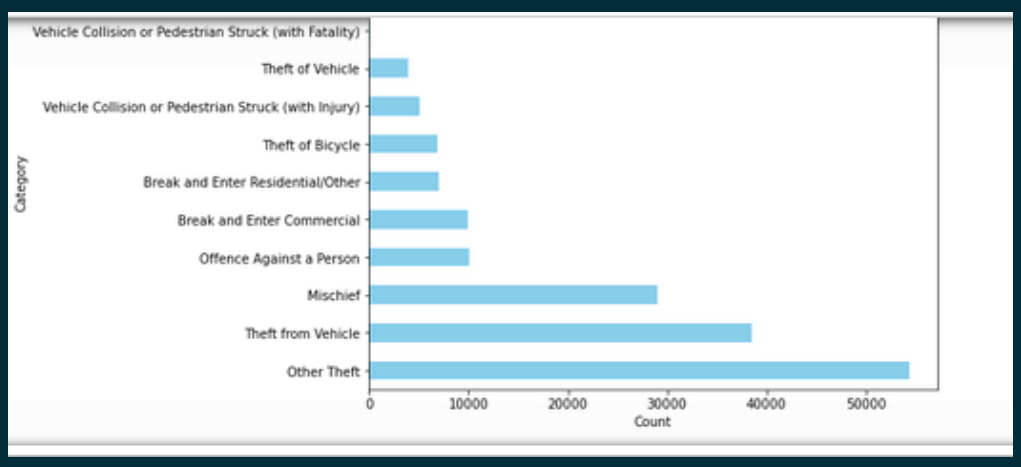
Total number of incidents

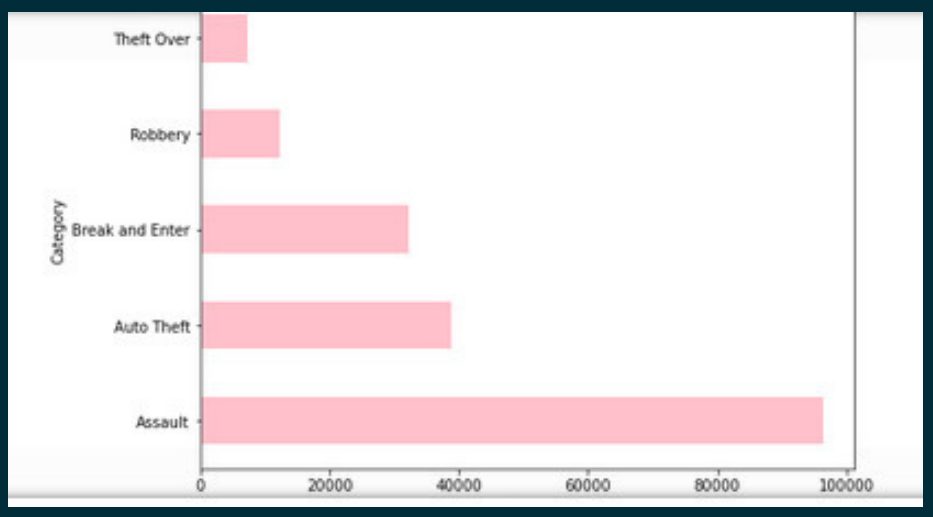
Vancouver (Blue Graph):

- The most common crime category is Other Theft, followed by Theft from Vehicle and Mischief.
- Break and Enter (both Residential/Other and Commercial) is also relatively frequent.
- Violent crimes such as Offence Against a Person are less frequent than property-related crimes

Toronto (Pink Graph):

- Assault is by far the most reported crime type in Toronto, with nearly 100,000 incidents.
- This is followed by Auto Theft, Break and Enter, and Robbery.
- **Overall, Toronto shows a higher concentration of violent crimes compared to Vancouver**.





Key Insights

- Vancouver faces a higher volume of property-related crimes, particularly theft and mischief.
- Toronto, on the other hand, reports significantly more violent crimes, especially assault.

Crime Count Summary by City and Year

- We grouped the data by city, year, and crime category to see how many times each type of crime occurred each year.
- We selected the top most common crime categories .
- We filtered the data to include only those top crimes for a cleaner, focused summary.
- Rows show each city and year.
- Columns show the top 10 crime types.
- Values show the number of crimes reported.

Observations

Toronto

- Assaults steadily increased each year from 2020 to 2023, then dropped in 2024.
- Auto thefts had a sharp increase between 2021–2023 and dropped in 2024.
- Break and enter showed a fluctuating but overall slightly increasing trend

Vancouver

- Other Theft in Vancouver has consistently increased year by year.
- Theft from Vehicle shows a gradual decline from 2020 to 2024.

Key Insights

- Toronto rising trend in violent crimes, especially assault and auto theft.
- Vancouver is affected by property-related crimes, particularly theft, with "Other Theft" rise and "Theft from Vehicle" declining .

Crime Counts for Top Crime Types by City and Year:

Category	city	Year	Assault	Auto Theft	Break and Enter	Other Theft	\
0	Toronto	2020	16600.0	5113.0	6749.0	0.0	
1	Toronto	2021	17317.0	5880.0	5530.0	0.0	
2	Toronto	2022	19209.0	8594.0	5910.0	0.0	
3	Toronto	2023	21703.0	10841.0	7454.0	0.0	
4	Toronto	2024	21597.0	8317.0	6592.0	0.0	
5	Vancouver	2020	0.0	0.0	0.0	8649.0	
6	Vancouver	2021	0.0	0.0	0.0	8583.0	
7	Vancouver	2022	0.0	0.0	0.0	10759.0	
8	Vancouver	2023	0.0	0.0	0.0	12831.0	
9	Vancouver	2024	0.0	0.0	0.0	13607.0	

Category	Theft from Vehicle
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
5	10431.0
6	7505.0
7	7281.0
8	7378.0
9	5934.0

Descriptive Statistics of Selected Crime Types (2020–2024)

- We grouped crime data by city, year, and crime type to count how often each crime occurred.
- Then, we pivoted the data so that each crime type became a column, and each row shows crime counts for one city and one year.
- We used .describe() to calculate summary statistics for each crime type, such as:
- Mean (average number of crimes)
- Min and Max (lowest and highest counts)
- Standard deviation (how much the crime count varies)
- Percentiles
- This gives us a quick overview of which crimes are most common, how crime levels vary

Descriptive Statistics of Crime Counts:

Category	Assault	Auto Theft	Break and Enter
count	10.000000	10.000000	10.000000
mean	9642.600000	3874.500000	3223.500000
std	10285.250715	4360.890231	3434.547657
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	8300.000000	2556.500000	2765.000000
75%	18736.000000	7707.750000	6421.500000
max	21703.000000	10841.000000	7454.000000

Category	Break and Enter Commercial	Break and Enter Residential/Other
count	10.000000	10.000000
mean	1000.600000	700.800000
std	1112.014009	790.748703
min	0.000000	0.000000
25%	0.000000	0.000000
50%	660.000000	488.000000
75%	1948.000000	1247.750000
max	2788.000000	2084.000000

Category	Homicide	Mischief	Offence Against a Person	Other Theft
count	10.000000	10.00000	10.000000	10.000000
mean	7.400000	2904.30000	1012.300000	5442.900000
std	8.248906	3076.52387	1067.839673	5941.965415
min	0.000000	0.00000	0.000000	0.000000
25%	0.000000	0.00000	0.000000	0.000000
50%	5.500000	2704.00000	963.500000	4291.500000
75%	13.250000	5577.50000	2024.750000	10231.500000
max	19.000000	6458.00000	2087.000000	13607.000000

Category	Robbery	Theft Over	Theft from Vehicle	Theft of Bicycle
count	10.000000	10.00000	10.00000	10.000000
mean	1226.800000	727.20000	3852.90000	690.700000
std	1309.947395	794.60903	4207.69933	783.645619
min	0.000000	0.00000	0.00000	0.000000
25%	0.000000	0.00000	0.00000	0.000000
50%	985.000000	543.50000	2967.00000	440.000000
75%	2416.500000	1387.50000	7353.75000	1357.500000
max	2748.000000	1811.00000	10431.00000	1987.000000

Highest Averages

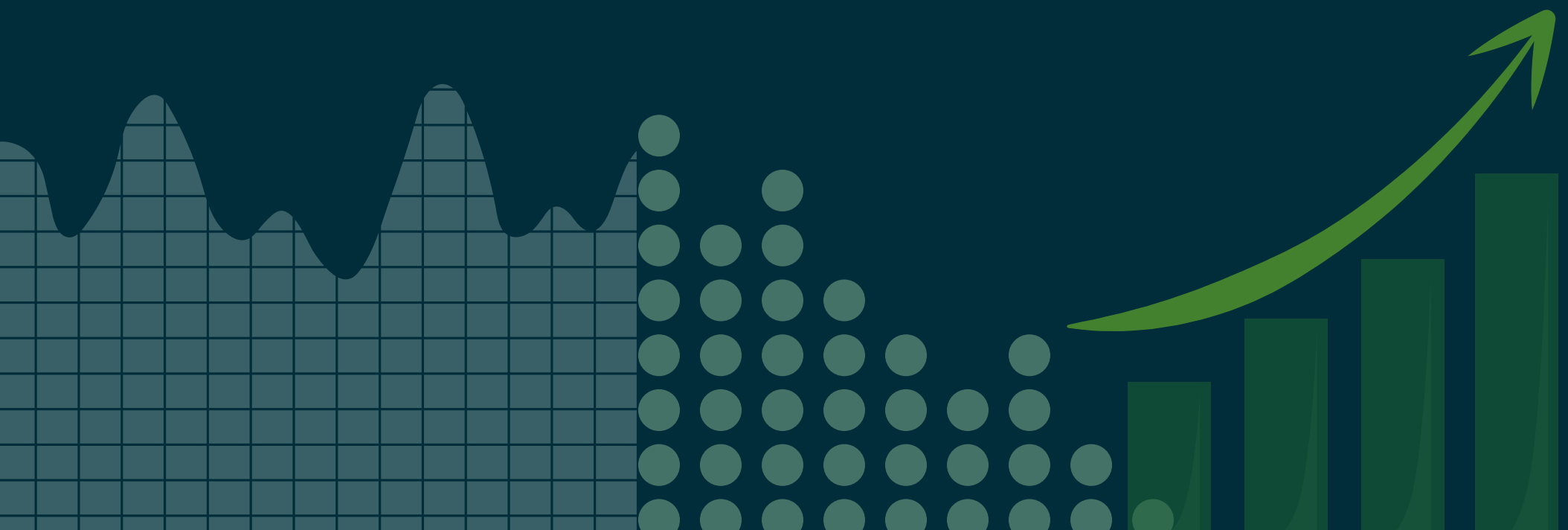
Assault has the highest average number of incidents (9642 per year) with a maximum of 21,703 cases, showing it’s the most frequent violent crime.

Less Frequent Crimes:

Homicide has the lowest counts overall (mean = 7, max = 22)

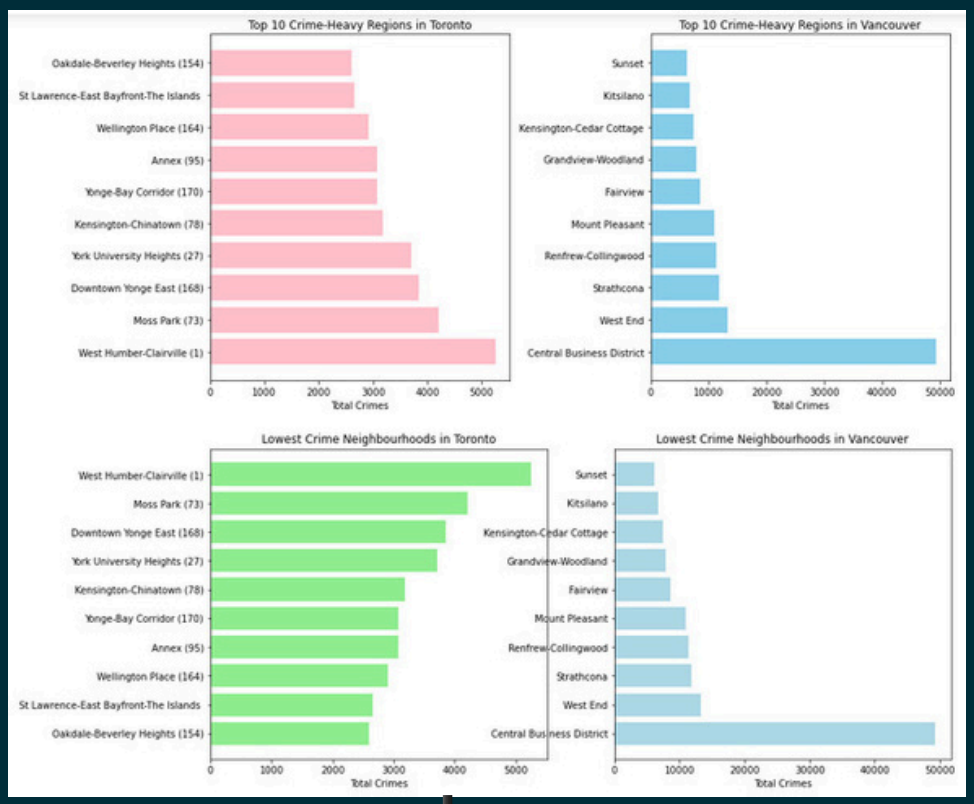
Important Take Away

Some crime types, like Auto Theft and Theft from Vehicles, go up and down a lot each year. This could be because of changes in how crimes are reported, how police handle them, or even changes in the economy



Regional Crime Analysis in Toronto and Vancouver

- We grouped the data by city and Neighbourhood to count the total number of crimes reported in each region.
- The counts were then sorted to find the Top 10 neighbourhoods with the highest and lowest crime rates in each city.
- We created horizontal bar charts to visualize these neighbourhoods, making it easy to compare areas with:
  - The most crime ("hotspots")
  - The least crime ("low-risk zones")



Top 10 Crime-Heavy Neighbourhoods (2020–2024)

In both Toronto and Vancouver, crime is not evenly distributed. Neighbourhoods account for a large number of total reported incidents

In Toronto, these high-crime areas seems to represent densely populated regions, possibly with limited access to safety .

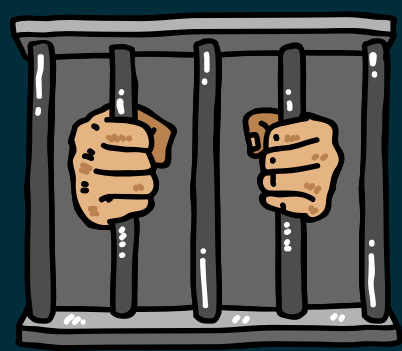
We can say that these areas require :

- Increased police presence
- Improved street lighting and surveillance

Lowest Crime Neighbourhoods

Several neighbourhoods in both cities consistently reported very low crime counts over five years.
These regions may have :

- Higher socioeconomic stability
- Better access to education, health, and support systems

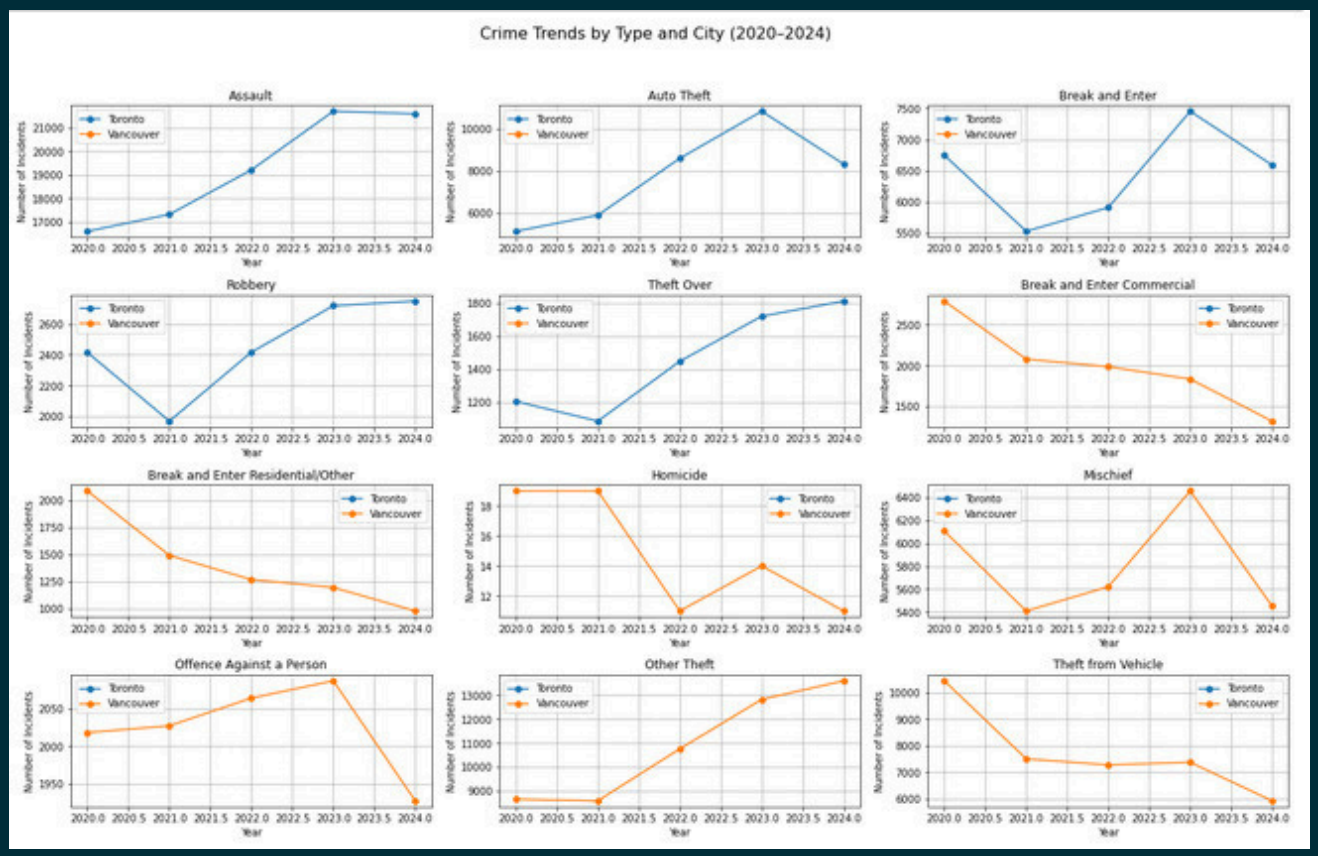


Crime Trend Analysis by Type and City

- Groups data by Year, City, and Crime Category (like Violent Crime, Property Crime).
- Counts how many times each crime category occurred per year and city.
- Converts grouped data into a DataFrame (crime_summary) for plotting.
- Prepares a 4x3 grid layout for plotting up to 12 charts.

For each crime category:

- Plots a line graph showing trends over the years.
- Uses different lines for Toronto and Vancouver.
- Adds titles, labels, gridlines, and legends to each chart.
- Displays all charts with a main title at the top.

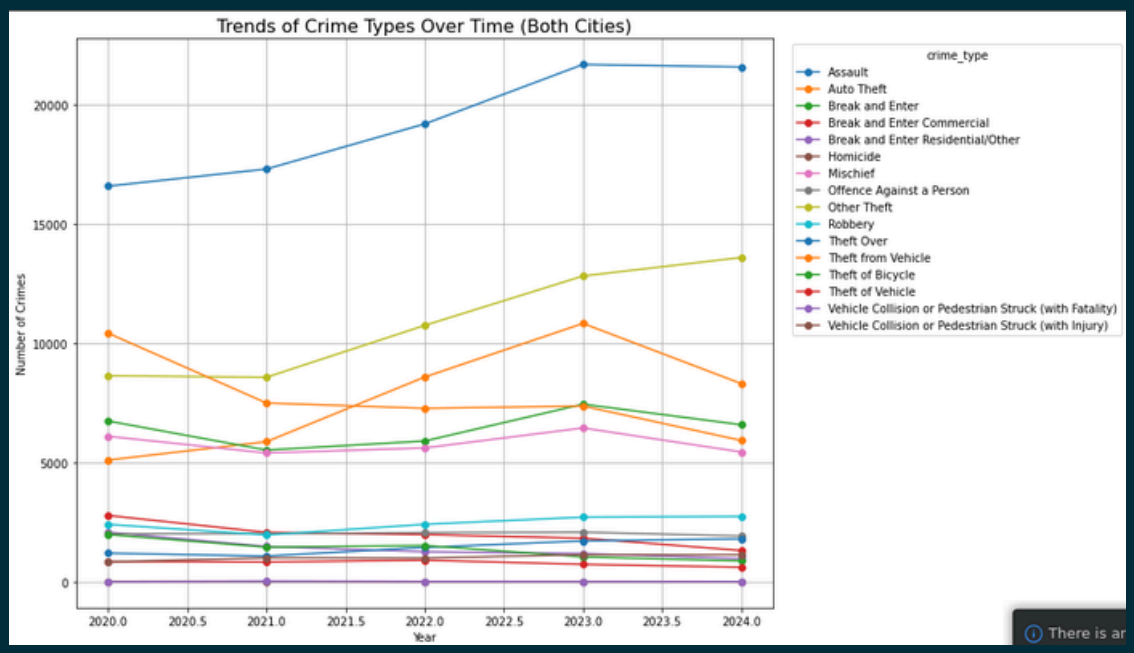


Result

A	B	C
Crime Type	**Toronto Trend**	**Vancouver Trend**
Assault	Steady rise until 2023, slight dip in 2024	Low and stable throughout
Auto Theft	Sharp rise (2021–2023), drop in 2024	Very low, almost flat
Break and Enter	Increased till 2023, declined in 2024	Low and steady
Robbery	Gradual and consistent increase	Very few or no cases
Theft Over	Rising trend across all years	Rare or not reported
Break and Enter (Commercial)	Slightly fluctuating, mild increase	Declining steadily
Break and Enter (Residential)	Mild decrease over years	Declining every year
Homicide	Low and mostly stable	Low and flat
Mischief	Gradual increase	Sharp in 2023, drop in 2024
Offence Against a Person	Increased till 2023, sudden drop in 2024	Very low, flat
Other Theft	Mild and steady trend	Consistent rise, peaking in 2024
Theft from Vehicle	Moderate with small fluctuations	Gradual decrease year by year

Crime Types Over Time (Both Cities Combined)

- Groups the data by Year and crime_type to count how many times each crime happened per year.
- Pivots the data so each crime type becomes a separate column, with years as rows and crime counts as values
- Plots line graphs for each crime type to show how they change over time.
- Adds labels, title, legend, and grid to make the chart clear and readable

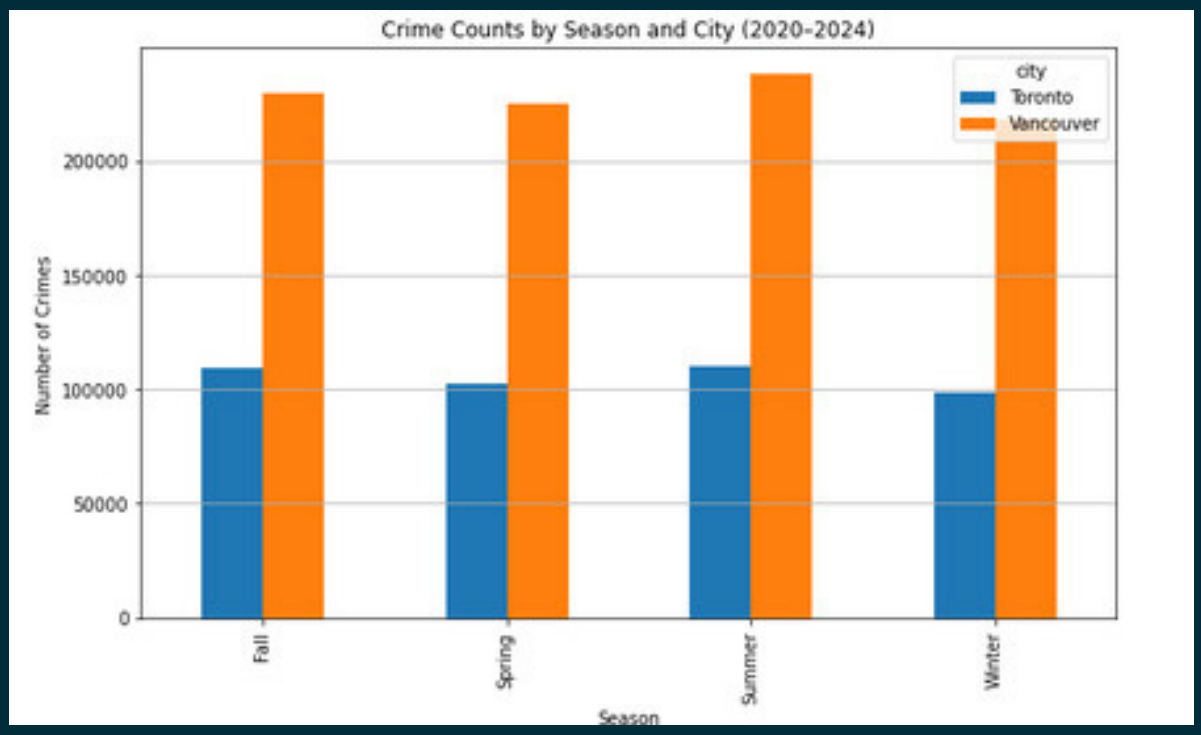


Key Outcomes

- Assault shows a steady and significant increase, in 2023 and slightly decrease in 2024 .
- Auto Theft rose sharply from 2021 to 2023, then dropped in 2024 — showing high year-to-year fluctuation.
- Other Theft has shown a consistent increase, becoming the second most reported crime by 2024.
- Break and Enter and Theft from Vehicle slightly decreased in 2024 after rising in earlier years.
- Traffic Incidents (Injury/Fatality) and Homicides remain consistently low throughout all years.
- Most other crimes like Mischief, Robbery, and Theft of Bicycle have remained relatively stable with slight ups and downs.

Seasonal Crime Analysis

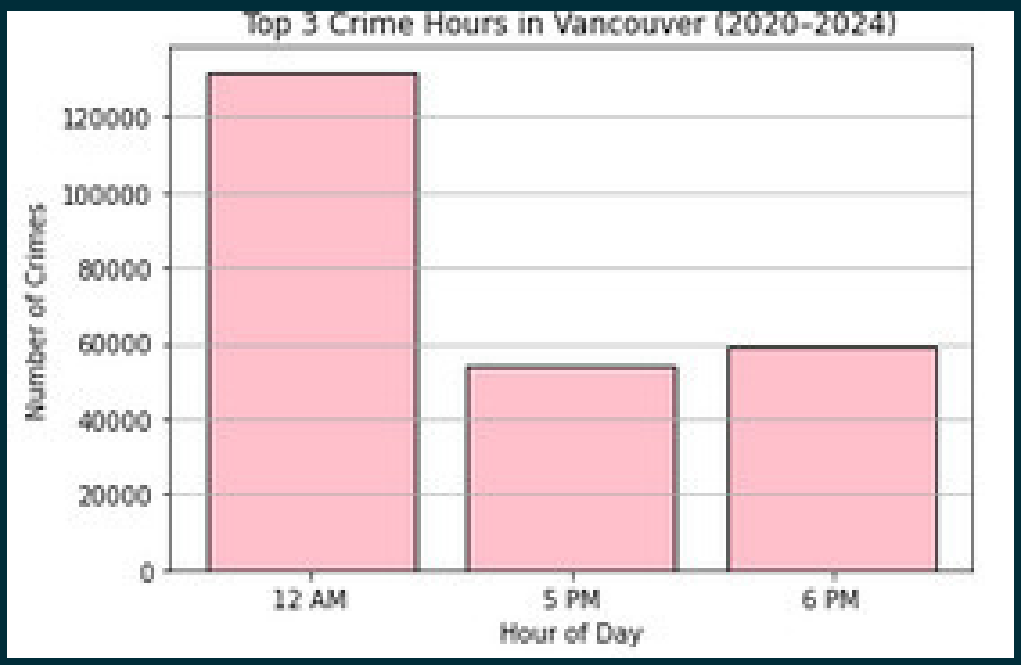
- Combined the two datasets into one using pd.concat() for unified analysis.
- Defined a custom function get_season() to convert each month to a season:
- Winter: Dec, Jan, Feb
- Spring: Mar, Apr, May
- Summer: Jun, Jul, Aug
- Fall: Sep, Oct, Nov
- Applied the function to the Month column to create a new Season column.
- Grouped the combined data by city and season, and counted the number of crimes in each group.
- Used pivot() to format the grouped data for comparison between cities.
- Plotted a bar chart to visualize the number of crimes per season for both cities.

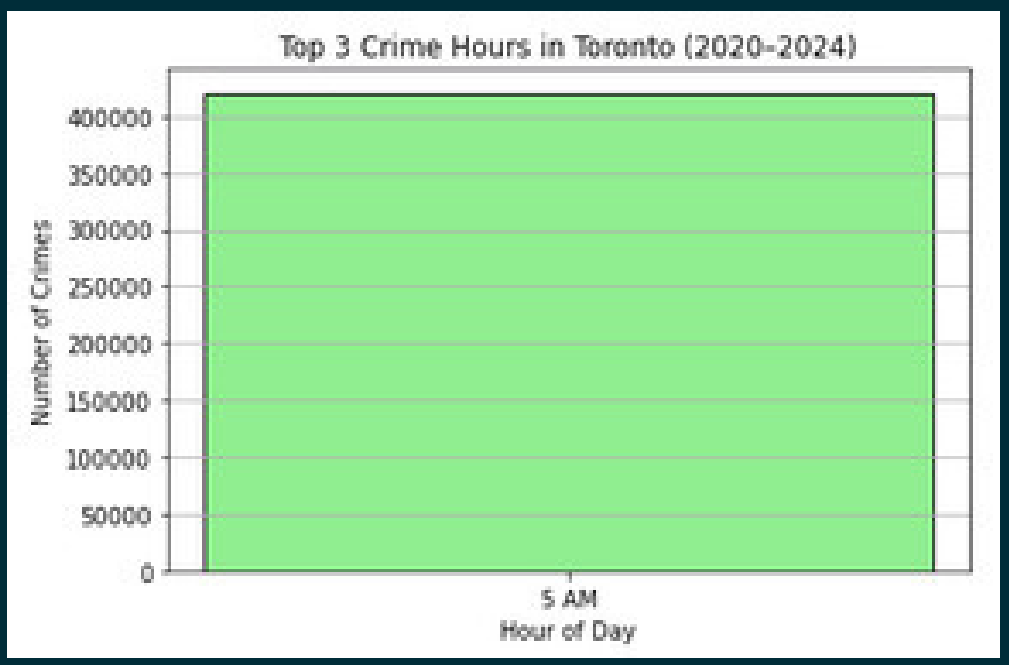


- Vancouver has consistently higher crime counts than Toronto across all seasons from 2020 to 2024.
- Summer is the peak season for crimes in both cities, while Fall appears lowest in Toronto, likely due to **missing data**.

Crime by Time of Day

- The code starts by converting the date and time columns into proper datetime format.
- It extracts the hour from each crime record to find out what time of day the crime happened.
- A new column is added to identify whether the data is from Toronto or Vancouver.
- Both city datasets are combined into one for analysis.
- The code then counts how many crimes happened at each hour of the day for Toronto.
- From this, it selects the top 3 hours with the most crimes.
- The hours are converted to AM/PM time labels (like 1 PM, 12 AM) .
- A bar chart is created to show the top 3 crime hours for Toronto.
- The same steps are repeated for Vancouver: count crimes by hour, get top 3, convert time labels, and plot the chart





Toronto:

5 AM is shown as the only and peak crime hour with over 400,000 recorded incidents.

Vancouver:

The top 3 hours for crimes are:

- 12 AM (midnight) – the highest, over 120,000 incidents.
- 5 PM and 6 PM – with about 60,000 incidents each.
-

IMPORTANT NOTICE: Vancouver’s data appears valid,Toronto’s data needs review may have data quality issue .

Socioeconomic Factors

The purpose of this section was to examine how income levels, education, and other socioeconomic indicators relate to crime rates in different neighbourhoods in Toronto and Vancouver.

Data Limitation & Approach Taken

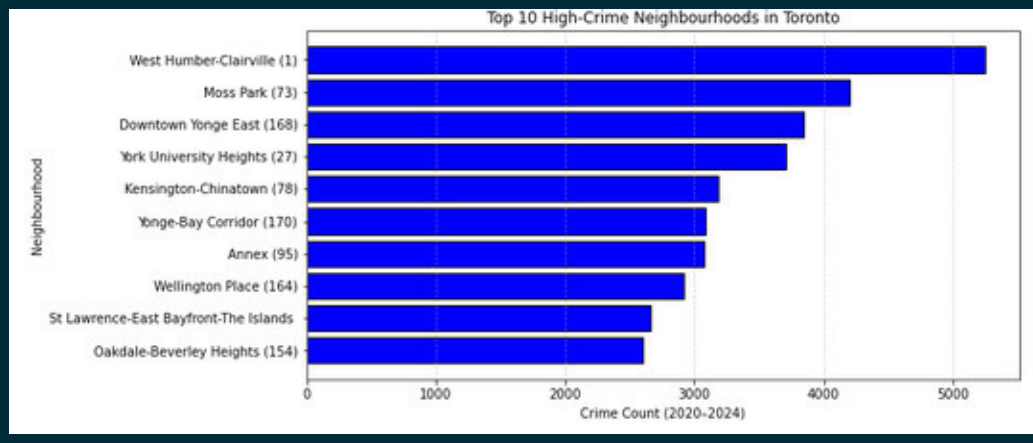
While our main datasets did not contain direct socioeconomic details (like income or education levels).We seached for the data set but we didn’t got . We researched external sources to manually collect relevant information about each neighbourhood.
We used : City of Toronto Neighbourhood Profiles (2021)

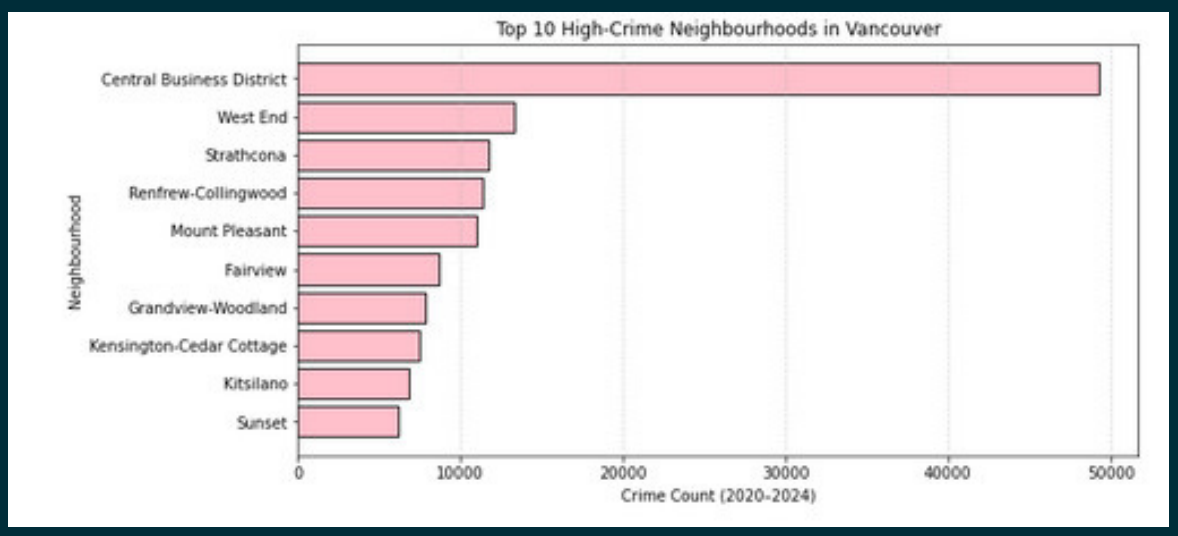
- City of Vancouver Community Profiles
- Statistics Canada Census 2021
- City of Toronto Neighbourhood Profiles (2021)

We used dataset showing selected high-crime and low-crime neighbourhoods in both cities, and added:

- Estimated income level (Low to Very High)
- Estimated education level
- General notes about the area (e.g., crowded, wealthy, popular) taken from the above sources .
- Although not from direct dataset columns, these were estimates were based on research and city sources.

Neighbourhood	City	Crime Level	Income Level	Education Level	Notes
Downtown Yonge East	Toronto	High	Low	Medium	Crowded, busy area, more crimes
West End	Vancouver	Very High	Medium	Medium	Popular area, many people
Kitsilano	Vancouver	High	High	High	Rich area but some crimes still
Moss Park	Toronto	Very High	Low	Low	Poor area, many crimes
Cabbagetown	Toronto	High	Medium	Medium	Changing area, mix of people
Woodbine-Lumsden	Toronto	Low	High	High	Quiet, safe, family area
Kerrisdale	Vancouver	Low	Very High	Very High	Very rich, very safe
West Point Grey	Vancouver	Low	High	High	Safe and rich neighborhood
Bridle Path	Toronto	Very Low	Very High	Very High	Super rich, very safe
Leaside	Toronto	Low	High	High	Rich and peaceful area





Crime Level Analysis and Future Projection

Histogram: Crime Level Distribution by City

- A histogram compares how many neighbourhoods in each city fall into each crime level category.
- It shows how Toronto and Vancouver differ in terms of overall safety distribution

Boxplot: Crime vs Income and Education

- A boxplot compares distributions of crime level, income, and education.
- Helps visualize variation and median values across all three dimensions.
- Shows that crime levels are more spread out, while income and education are skewed toward higher values.

CDF: Crime Level

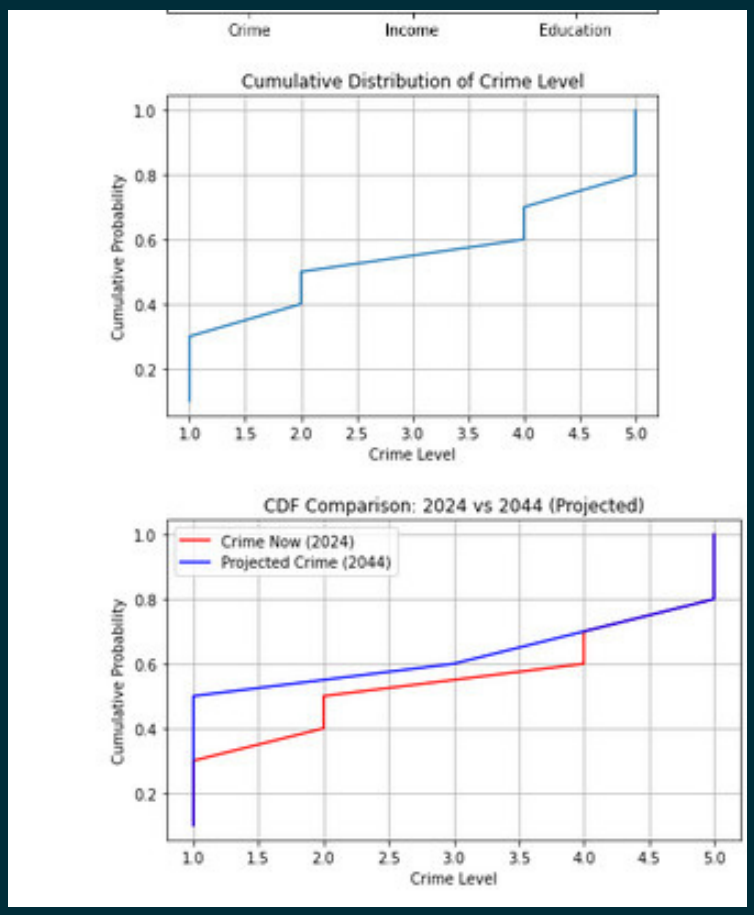
- A cumulative distribution plot of crime levels is shown.
- Helps understand what proportion of neighbourhoods fall below or at a certain crime level

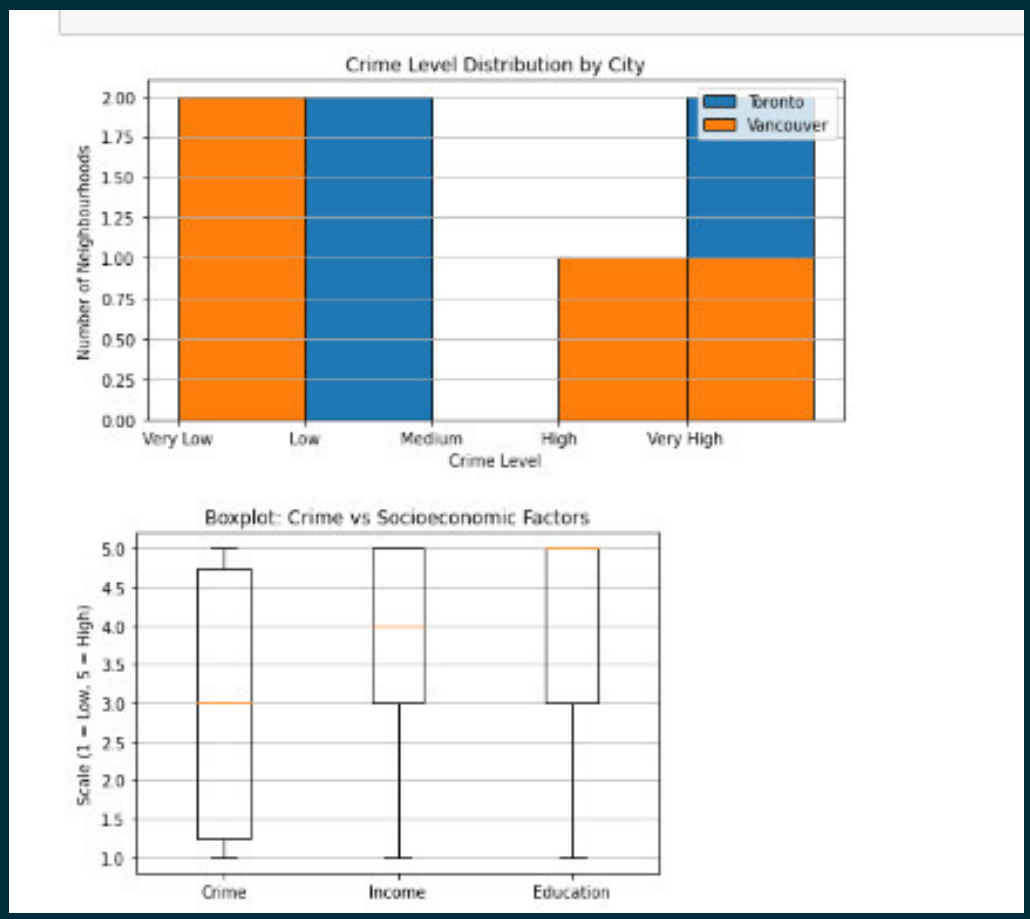
CDF Comparison: 2024 vs Projected 2044

The code plots two CDFs:

- Current Crime Level (2024)
- Projected Crime Level (2044)

The shift in the curve to the left in 2044 indicates an overall reduction in crime levels, assuming improvements in socioeconomic conditions.





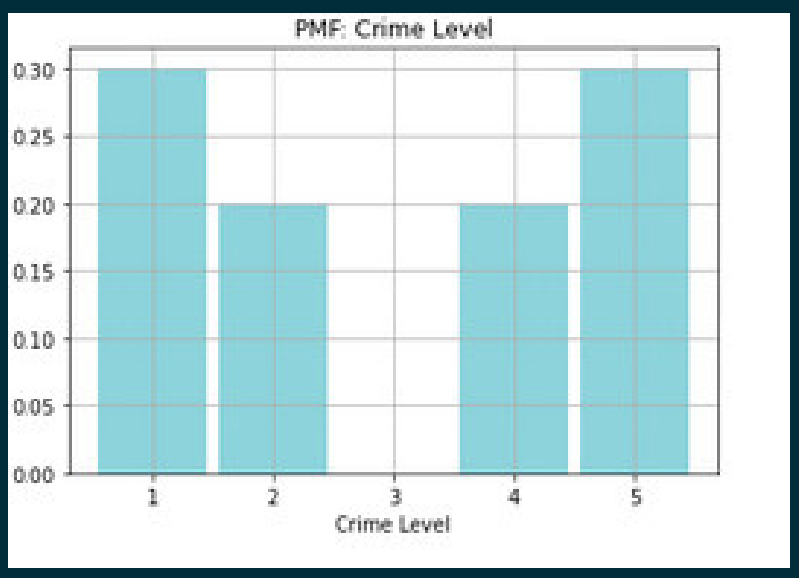
Socioeconomic Overview : Crime, Income, and Education

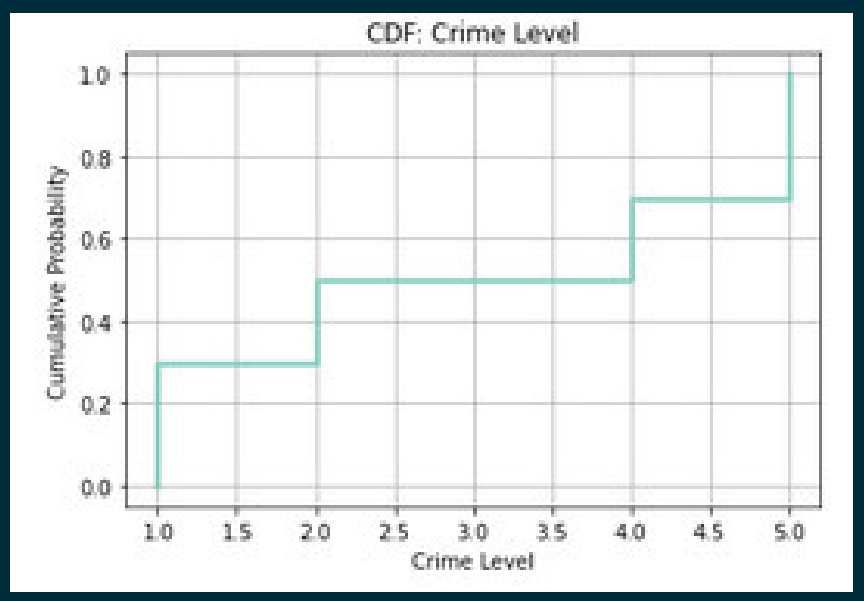
- A custom dataset is created with 10 neighbourhoods from Toronto and Vancouver.
- Each neighbourhood is scored manually on:
- Crime Level (1 = Very Low, 5 = Very High)
- Income Level (1 = Low, 5 = Very High)
- Education Level (1 = Low, 5 = Very High)
- The data is grouped by city using .groupby("City"), and the average scores for Crime_Level, Income, and Education are calculated for Toronto and Vancouver.
- The result (grouped_means) shows the mean values .

City	Crime_Level	Income	Education
Toronto	3.166667	3.333333	3.666667
Vancouver	2.750000	4.250000	4.500000

Probability and Cumulative Distribution Analysis

- We created a list of crime_levels based on manual scores from 10 neighborhoods.
- **PMF** (Probability Mass Function) was computed using thinkstats2.Pmf():
- It shows the relative frequency (probability) of each crime level (from 1 to 5).
- **CDF** (Cumulative Distribution Function) was computed using thinkstats2.Cdf():
- It shows the cumulative probability that a crime level is less than or equal to a given value.



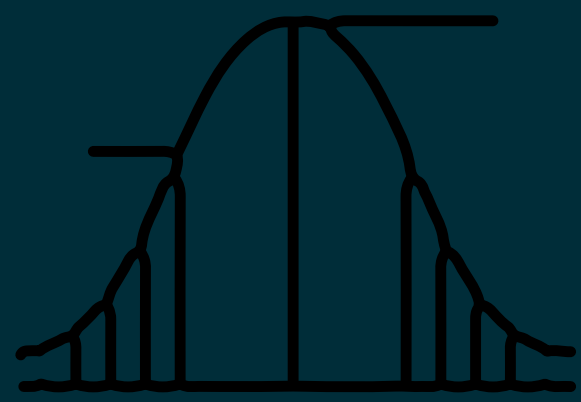


PMF

- Crime Level 1 and 5 both have the highest probability (30%)
- Crime Levels 2 and 4 each have a 20% probability.
- Crime Level 3 is not present, hence it shows 0 probability.

CDF

- About 30% of neighbourhoods have a crime level ≤ 1.
- CDF jumps again at level 4 and 5, indicating that:
- 70% of areas have a crime level ≤ 4.
- 100% have a crime level ≤ 5

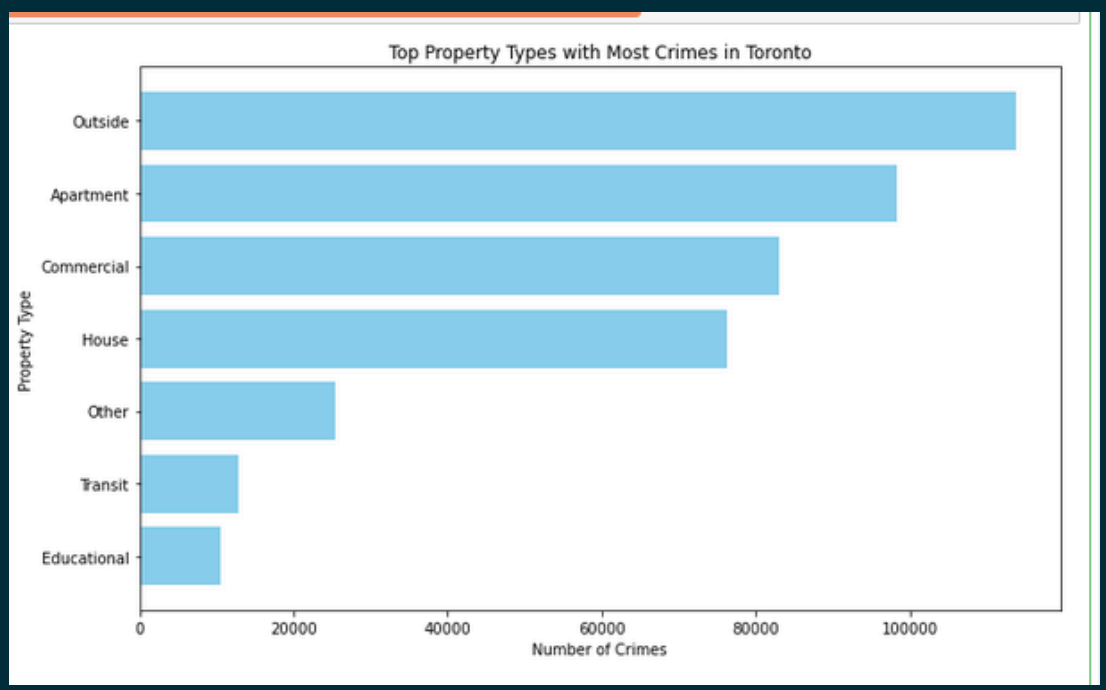


Property Analysis

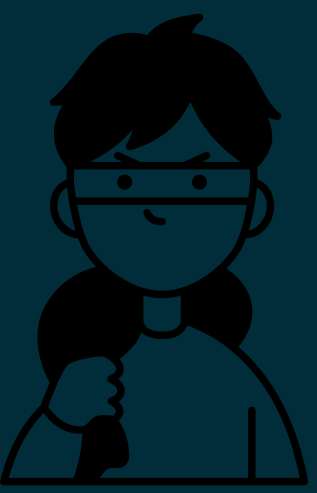
- Loaded the Toronto crime dataset and grouped incidents by PREMISES_TYPE (e.g., Apartment, Commercial, Outside).
- Calculated the number of crimes for each property type.
- Displayed the top 10 property types with the most reported crimes using a horizontal bar chart.

Vancouver Data Limitation!!!!

The Vancouver dataset does not contain a PREMISES_TYPE or LOCATION_TYPE column, so it was not possible to directly analyze property types.



- Outside" locations had the highest number of reported crimes, indicating that public and open areas (such as streets, parks, and open lots) are the most vulnerable
- "Apartments" and "Commercial" properties also show high crime counts, suggesting that both residential and business areas are significantly affected.
- "Houses" rank just below commercial properties, indicating private residences also face considerable crime risk



Challenges & References

Initial Data Limitations

At first, our plan was to compare Halifax and Toronto, but we were unable to find sufficient and complete crime data for Halifax. As a result, we shifted our focus to Toronto and Vancouver, which had more publicly available data.

Date and Time Formatting

One of the biggest technical challenges was converting and extracting useful information from date and time columns. We were not previously taught how to handle this in class, so we had to research and teach ourselves how to use datetime formatting in Python.

Missing Socioeconomic Data

Our crime datasets did not include any information about income, education, or living conditions. To complete the socioeconomic analysis, we manually researched each neighbourhood using external sources like city reports and census data.

Incomplete Fields in Vancouver Dataset

Unlike Toronto's dataset, Vancouver did not have a "Premises Type" or property-type column, which made it impossible to analyze crime by property type for Vancouver. So, this analysis was only done for Toronto

Large and Complex Datasets

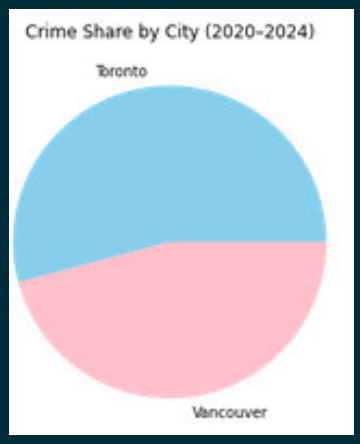
The datasets we found—especially for Toronto—were very large and detailed, with many columns and data points. Cleaning and filtering this data to extract relevant fields (like crime type, date, location, and neighbourhood) took a significant amount of time.

References

The rename() method in Pandas is used to rename the labels (column names or index values) of a DataFrame.
Converts the 'OCC_DATE' column (which may be a string) into datetime objects.
This enables to extract components like date, year, month, hour, etc., easily using .Using ignore_index with pandas.concat(): This article explains how to concatenate.inplace=True: Makes the changes directly to the original DataFrame (df_toronto) without needing to assign it to a new variable.

Conclusion

	City	Crime_Count
0	Toronto	200097
1	Vancouver	174480



This project explored and compared crime trends in Toronto and Vancouver from 2020 to 2024 using publicly available data. Toronto reported a total of **200,097 crimes**, slightly more than Vancouver's **174,480,** indicating that both cities experience high and consistent levels of urban crime. Seasonal analysis showed that crime rates peaked during the summer months in both cities, while Toronto had unusually low recorded incidents in the fall—likely due to missing or incomplete data. Time-of-day analysis revealed that Vancouver had logical crime peaks at 12 AM, 5 PM, and 6 PM, while Toronto showed a spike at 5 AM, suggesting possible data entry or default issues.In terms of location, Toronto saw the highest number of crimes in outside areas, followed by apartments, commercial properties, and houses. Vancouver lacked a property-type column.The cumulative distribution curve for projected 2044 crime levels showed a positive shift toward lower crime, reinforcing the importance of addressing root causes such as education and economic stability.

Overall, based on total reported incidents, **Toronto had the highest crime count during the 2020–2024** period. However, crime types and patterns varied significantly across both cities, influenced by geography, urban infrastructure, and social factors.
Additionally, visualizations such as histograms, boxplots, bar charts, and CDFs provided iinsights into crime concentration and its relationship with socioeconomic factors.The results show that improving education and income can help reduce crime, and that using data can help cities make better decisions to improve public safety.