Data Mining & Analysis

Cumulative Assignment Report

Toronto Police Major Crime Indicators

Prepared By:

Ahmed Al-Mahdi

Abdallah Seyam

Muhammad Osama

Objective of Project:

The objective of the project is to analyze the reported crimes across the city of Toronto. We will be retrieving the live data from REST API, the goal of the assignment will be to visualize and predict crime patterns using methods like time series and geospatial in Power BI.

We have taken the dataset from https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about

We will be using REST API using HTTP GET with JSON output format, it will be via Python code and JSON parsing. So, we are using public API to get the data. The dataset consists of crime data from 2014 to 2025 till March. It has 400000+ rows and 31 columns. However, for the scope of this assignment we will be choosing the data of the year 2025 only which is from January till March. So, the total rows are approximately 10000 that we are using for the visualization and the columns are 6.

Why our group choose this dataset:

Our group chose the Toronto Police Major Crime Indicators data because it offers detailed time-stamped geographically referenced data on major offence categories like Assault, Break and Enter, Auto Theft, Robbery, and Theft Over. The data is data-intensive and granular, available by offence and/or by victim for multi-dimensional exploration by time and place of spatial variation of crimes. While one event may occur repeatedly of more than one classification, this formatting offers more insightful understanding of the nature of crimes reported. This dataset is publicly available and is perfect for use in software like Power BI for data visualization as well as forecasting. With public safety relevance, real world applicability, as well as possible building predictive models, it is an ideal choice for assignment by our group. We will be using Clustering for this task and identifying the crime hotspots across the city of Toronto based upon the data we have.

Column Name and Explanation:

OCC Date: Date and time of crime happened.

OFFENCE: What type of offense was committed.

MCI_CATEGORY: The major crime group (like Assault, Theft, Robbery or Homicide)

Neighborhood_140: The graphic area of Toronto where crime took place.

OCC_DOW: The day of the week the crime happened.

PREMISES_TYPE: The type of location where crime took place like apartment, commercial, house.

CRIME_SEVERITY: A scale from 1- 5 that determines the severity of the crime

Data Transformation:

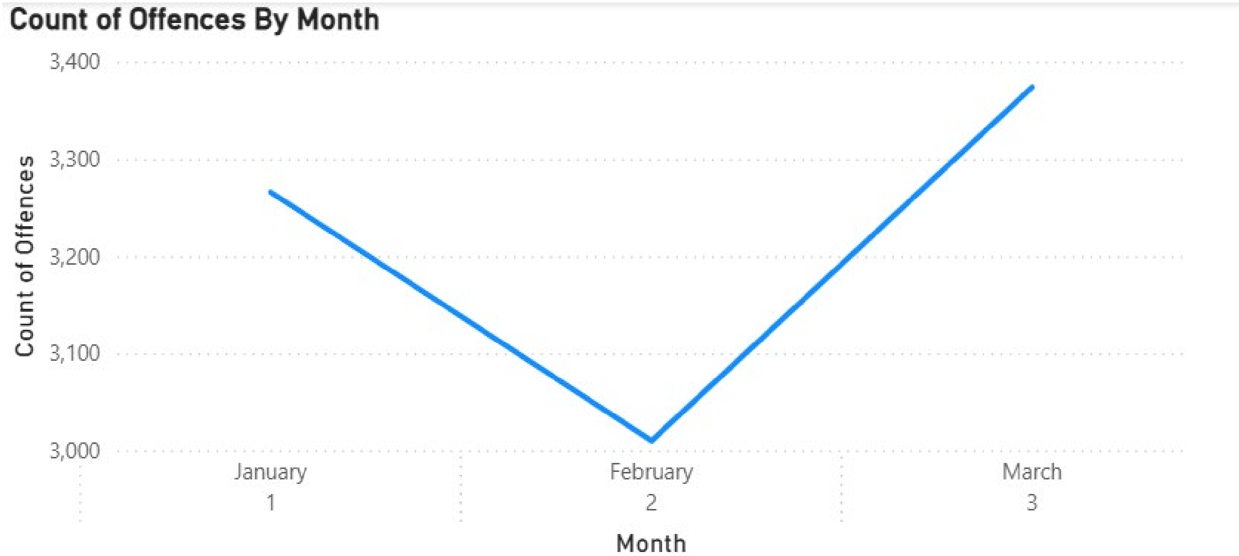
First, we have used Python script to connect to Toronto Police Rest API from there we have collected data for the months of Jan 2025 to March 2025. After that we have dropped the columns which are not meaningful in our analysis. We created some new features like crime severity which scores between 1 - 5 based upon MCI_CATEGORY. We have also cleaned some missing values in the columns of NEIGHBOURHOOD_140 and PREMISES_TYPE. After that we have dropped original timestamp and temporary time columns, and the date column is renamed for clarity. The data is then passed into Power BI where we have applied multiple transformations like column types are standardized, null values are filtered out, and months are converted from text to numeric format using a series of replacements so that we can visualize the data more accurately.

Machine Learning:

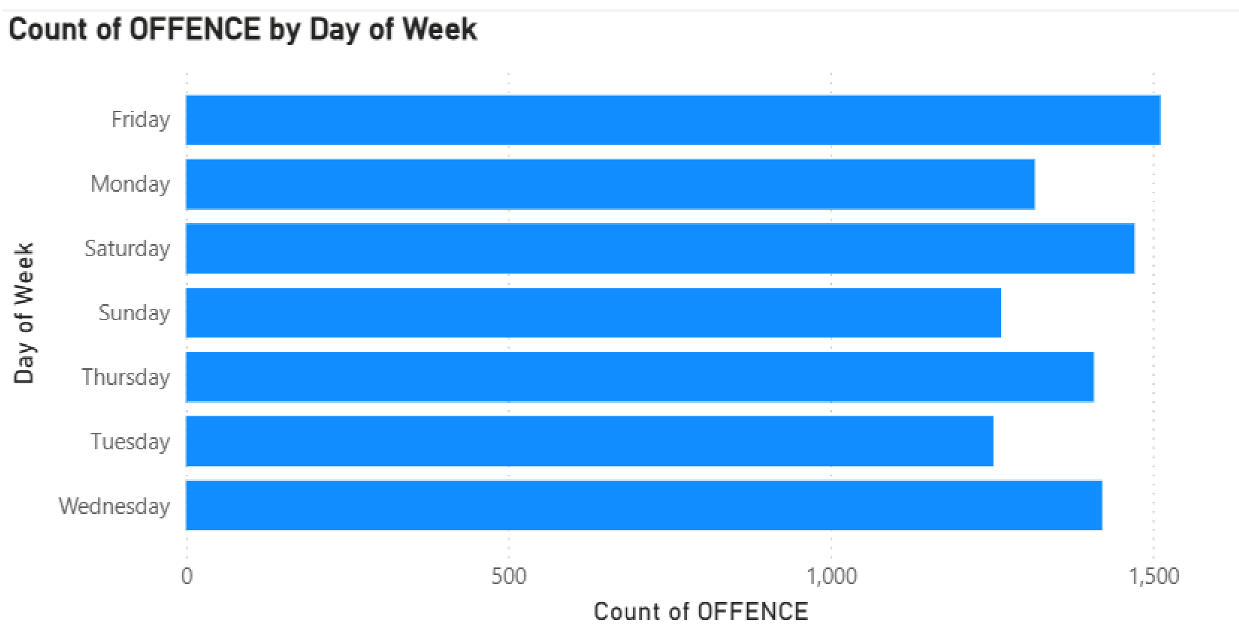
In this assignment we have used machine learning clustering model which we have implemented through PyCaret's unsupervised learning module. It identifies the crimes hotspots across the city of Toronto. We have applied this K-means algorithm to our cleaned and transformed data, which includes features such as crime date, type, severity, and location. The model is trained on similar crime incidents based in our data and once the model is trained each assign each record to cluster. This helps us to identify the segments which have similar crime characteristics. This enables us to visualize and identify high risk areas across the city of Toronto.

Analysis of Visuals:

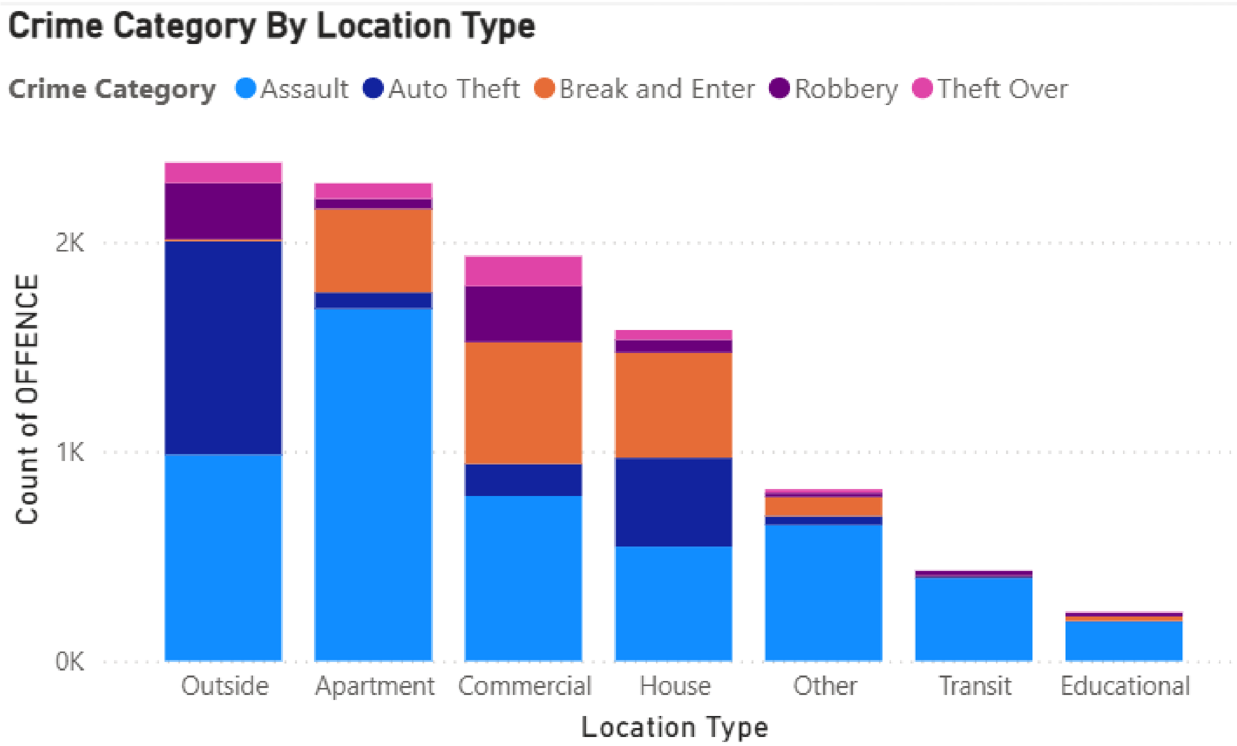
Data Storytelling:



The “Count of Offences by Month” visual shows us that the total number of crimes that took place in each month. As we can see the crime dropped from January to a February which is the lowest among the 3 months, then rose sharply in March to the highest level in the three-month period. The February dip may relate to winter conditions or seasonal patterns, while March’s increase could reflect more public activity. This trend highlights the need for targeted police resource allocation during expected seasonal spikes.

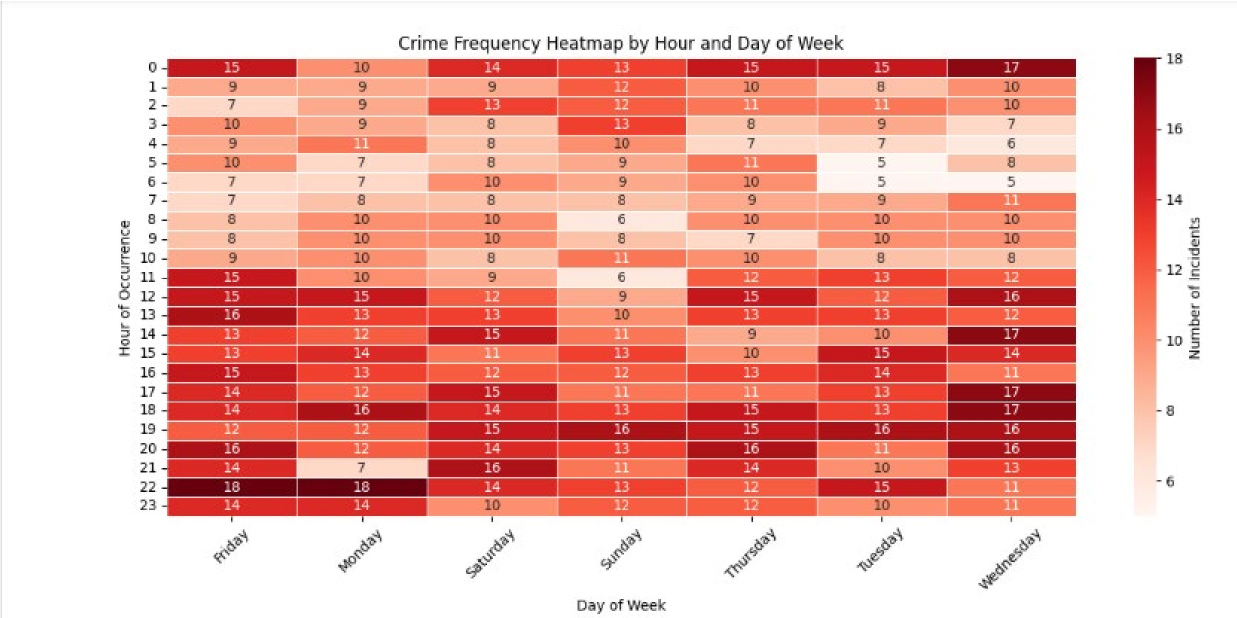


The “Count of Offence by Day of Week” tells us how each day of the day have the most crimes, it is distributed among all three months and further separated in days of the week. We can conclude from the visual Friday is the day on which most of the crime activities take place, closely followed by Saturday and Wednesday. The lowest counts occur on Tuesday and Sunday. This suggests crime peaks toward the end of the workweek like Friday has the highest crime count and weekend like Saturday, potentially due to increased social activity.

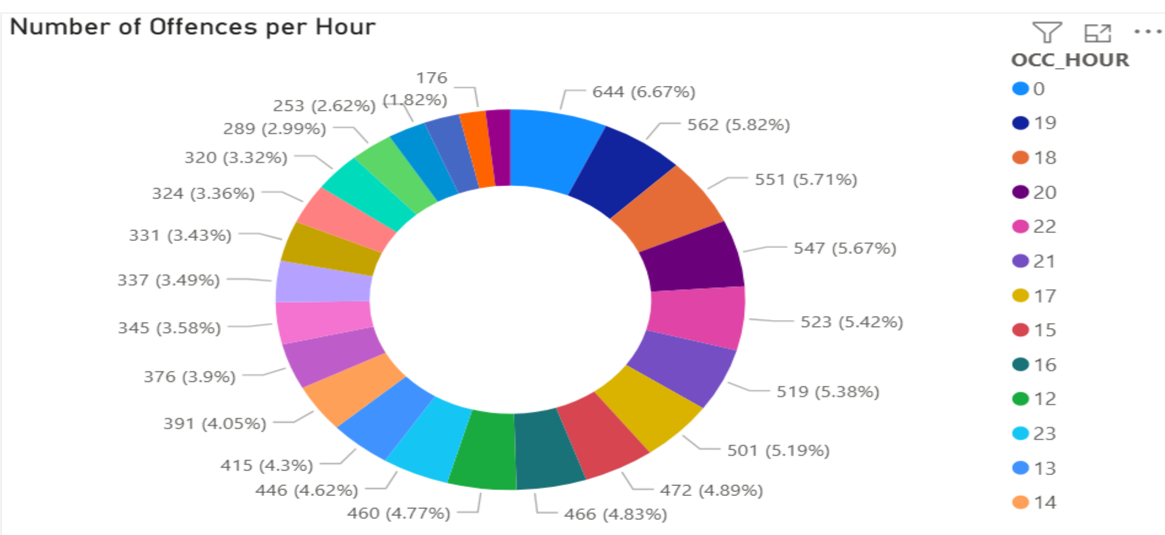


The “Crime Category by Location Type” visual shows that what types of offences are being done in what sort of premises and what is the total count of those offences. From visuals we can conclude outdoor or outside are most prone to crime activities which have a larger chunk of Assault and Auto Theft. The second most is apartment where we see assault is happening most of the times. Moving towards the lowest affected vicinities educational institutions have the least number of offences, which is followed by Transit. From this we can draw conclusion that overall seeing at the data Assault tends to be the most reported offence amongst all categories.

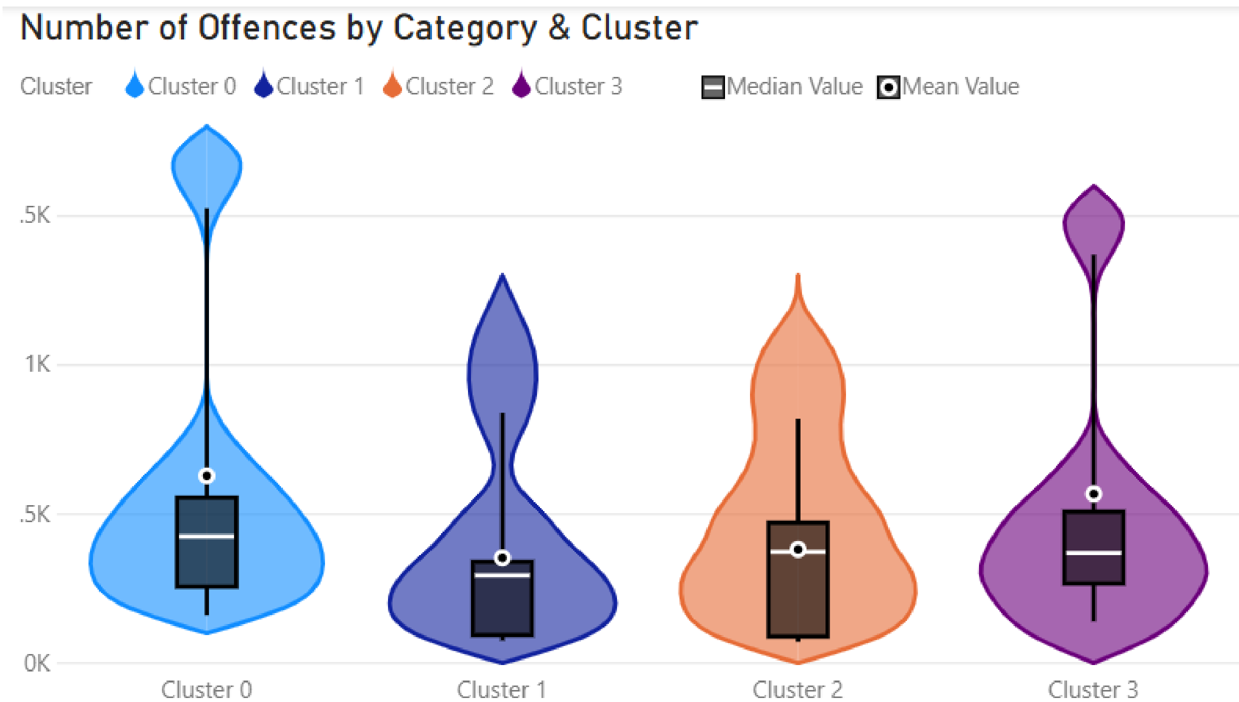
Data Showcasing:



In this visual of "Crime Frequency Heatmap by Hour and Day of Week" it shows us during the days which time is the hotspots for offences. We can see how on Friday at 22:00 hour the greatest number of crimes are taking place. This can enable us to detect the time which is more frequent in offences. Early morning hours (02:00–06:00) generally show the lowest activity across all days. And the hours from 1800 onwards till midnight are severe in terms of offences and require heightened law enforcement presence and community safety measures.

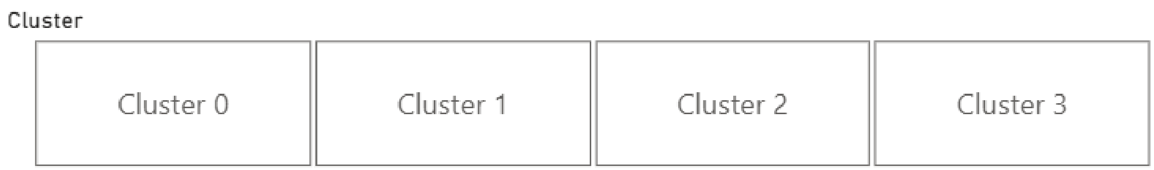


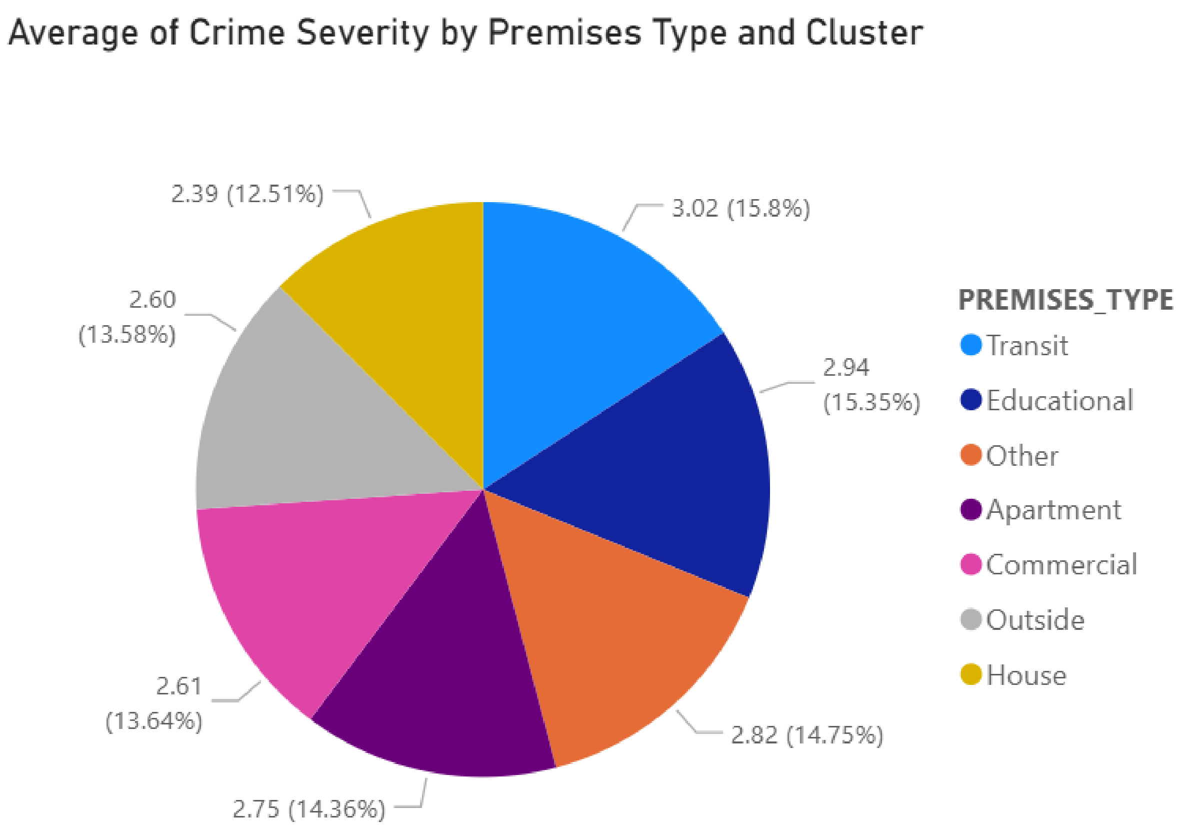
In this donut chart we can see the “Number of Offences per Hour” it clearly highlights how the midnight (00:00 hours) is the time when most of the offences are happening. It is followed by 19:00 hours which have the second most count for reported offences. Moving on to 1800 hours and so on. We can also see that hours like 5, 4, and 6 when least number of offences are committed. From this we can conclude that hours 18 to 00 are the most vulnerable times when a crime is happening.



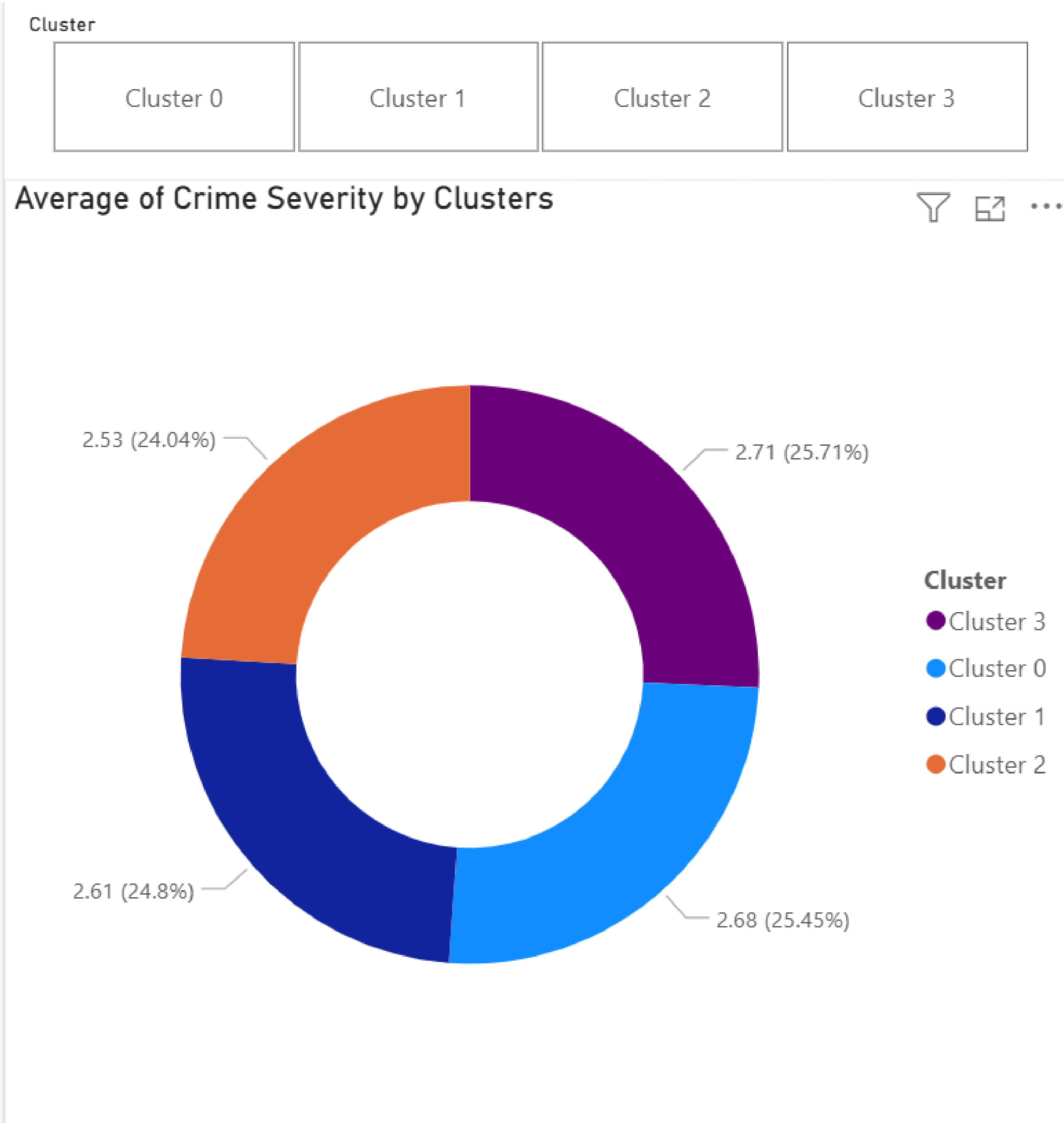
For this visualization we have used machine learning algorithm. This violin plot shows the distribution of counts of offences among four different clusters, each of them colored differently: light blue (Cluster 0), dark blue (Cluster 1), orange (Cluster 2), and purple (Cluster 3). The silhouette of each violin displays the concentration of data points, and where offences are denser or sparser in each cluster. A box plot inside each violin displays the interquartile range, and a median and a mean are shown with a black line and white dot, respectively. Most of the offences are bunched together around the lower end of the spectrum (below 1,000), which means that high counts of offences are comparatively an uncommon feature throughout the whole range of clusters. This visualization provides an opportunity to compare the way offences are distributed and centered in each group and get information about future patterns or abnormalities of the data set

Data Art:





This pie chart provides the average severity of crimes in different types of premises, divided by clusters. We have used slicers for clusters if we need to see the statistics per cluster. Each sector of the chart displays a separate location type like Transit, Educational, Apartment, Commercial, and House with its related average severity score and percentage of the whole. Transit locations have the highest average crime severity (3.02), followed by educational locations (2.94), while Houses have the minimum (2.39). The legend and cluster labels (Cluster 0 to Cluster 3) in the chart further hint at the fact that these averages could be studied further in particular groupings. As such, the visualization portrays effectively how crime severity varies according to the location context and provides useful information about location-specific focused safety interventions and policy development.



This donut chart gives an indication of the average severity of crimes by cluster, and each cluster has a distinct color: light blue for Cluster 0, dark blue for Cluster 1, orange for Cluster 2, and purple for Cluster 3. Chart segments show average severity score and percentage data contribution of each cluster to the entirety. Cluster 3 has the highest average severity with 2.71 (25.71%), Cluster 0 has the second highest with 2.68 (25.45%), and Cluster 2 has the lowest with 2.53 (24.04%). Relatively small gaps between clusters show a relatively even division of severity of crimes, but the visualization still shows subtle differences potentially useful for directional intervention or allocation of resources plans.

Conclusion:

In summary, examination of the January–March 2025 Toronto data on major crimes presents obvious temporal, geographic, and categorical crime patterns to guide focused policing responses. Crime reaches highest frequencies in March, on late evenings and weekends, with highest concentrations in outdoors and apartment locations. Cluster results indicate some neighborhoods (Clusters 0 and 3) have highest offence frequencies and somewhat elevated average severity, and transit and education facilities are the locations of highest-severity crimes. These findings indicate prioritization of law enforcement efforts toward evening patrols and protecting vulnerable types of premises to maximize public safety.

Resources In APA Format:

1) Toronto Police Service. (n.d.). Major crime indicators open data. City of Toronto Open Data Portal. Retrieved August 12, 2025, from https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about
2) Ali, M. (2020). PyCaret: An open-source, low-code machine learning library in Python. PyCaret. Retrieved August 12, 2025, from https://pycaret.org
3) Microsoft Corporation. (n.d.). Power BI. Retrieved August 12, 2025, from https://powerbi.microsoft.com
4) Fielding, R. T. (2000). Architectural styles and the design of network-based software architectures (Doctoral dissertation, University of California, Irvine). Retrieved from https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
5) Bray, T. (2017). The JavaScript Object Notation (JSON) Data Interchange Format (RFC 8259). Internet Engineering Task Force. https://doi.org/10.17487/RFC8259