Data Analysis for Crime Database

INFO8126: Data Acquisition, Analysis and Visualization

Sakshi Sandeep Salvi
Mitali Sharma
Nitigya Vasudev
Umang Sehgal
Prince Choudhary

Business Analytics, Conestoga College, Institute of Technology and Advanced Learning

Professor Maggie Santos

July 29, 2025

Table of Contents

Executive Summary

This report discusses crime rates in the top Ontario cities between 2021 and 2024, based on structured publicly available statistics, with only a reporting of actual incidences reporting being used. This is aimed at finding patterns, discovering differences in the city-level and predicting any future trends through statistical and machine learning methods. Included cities of study are Toronto, Windsor, Guelph, London, Kitchener-Waterloo-Cambridge. The analysis demonstrates that Toronto always has the largest crime volumes and the cities such as Guelph have much fewer cases. The most frequent ones are assault, impaired driving, and drug offences.

Bar charts, line graphs, and heatmaps, including boxplots, illustrate a consistent positive trend in the number of crimes reported, in majority cases in urban centers. It was determined under K-Means clustering including similar cities on the basis of crime intensity where it was found that Toronto represented a high-risk cluster, London, and Waterloo moderate clusters and Guelph and Windsor low. The use of statistical methods proved the existence of high variances in the crime levels. A t-test between Toronto and Windsor gave a p-value of 0.033, supporting the significance that, crime rates vary depending on what the geography is.

Predictive modelling was conducted using both linear regression and Random Forest. Linear regression performed poorly (R^2 = -96), while Random Forest achieved excellent accuracy (R^2 = 0.94, RMSE = 360), effectively predicting crime based on year, city, and violation type. Using exponential smoothing, the forecasted total crime for 2025 is approximately 504,272 incidents. This closely aligns with projections from a 10% annual growth simulation.

Overall, the report demonstrates how integrated analytics can help stakeholders understand and anticipate crime patterns. These insights can support evidence-based decision-making for policing, prevention programs, and urban planning.

Data Cleaning and Preparation

The dataset used for this project *Crime Dataset for Data Acquisition.csv* contained records of reported criminal incidents from 2021 to 2024 across major Ontario cities. It included variables such as REF_DATE (year), GEO (geographic location/city), Violations (type of crime), Statistics (e.g., actual incidents, rates), and VALUE (crime count or rate).

To ensure accurate and consistent analysis, several data cleaning and preparation steps were performed:

1. **Filtering for Relevant Records**
   The dataset contained multiple types of statistical measures (e.g., rate per 100,000, cleared cases, unfounded cases). For consistency and focus, the analysis was limited to rows where Statistics was **"Actual incidents"**, representing the raw number of reported crimes.

2. **Handling Missing Values**
   A small number of entries had missing values in the VALUE column. These rows were either removed or replaced with 0 based on the context of the analysis. This ensured that summary statistics and model training were not biased due to null entries.

3. **Data Type Standardization**
   Columns like REF_DATE were confirmed to be numeric, while categorical variables such as GEO and Violations were properly encoded for downstream modelling, including one-hot encoding for machine learning pipelines.

4. **Feature Selection & Renaming**
   Only relevant features (REF_DATE, GEO, Violations, VALUE) were retained. Additional metadata columns (like DGUID, UOM_ID, or VECTOR) were dropped as they were not meaningful for this analysis. Renamed REF_DATE to Year, GEO to City, and VALUE to Value.

These preprocessing steps created a clean, structured dataset suitable for exploratory analysis, statistical testing, regression modelling, and forecasting.

Comprehensive Data Analytics Framework

Descriptive Statistics of Reported Crime by Violation Type (2021–2024)

Violations	count	mean	std	min	25%	50%	75%	max
Assault, level 1 [1430]	20.0	6137.85	8898.212193	340.0	1294.50	2130.0	3650.50	26821.0
Attempted murder [1210]	20.0	37.20	63.998026	0.0	3.75	6.5	15.00	176.0
Criminal negligence causing death [1150]	20.0	2.10	3.726152	0.0	0.00	1.0	1.25	14.0
Homicide [110]	20.0	29.65	49.023383	0.0	4.00	6.5	11.50	133.0
Murder, first degree [1110]	20.0	15.05	27.097242	0.0	0.00	3.0	5.50	74.0
Robbery [1610]	20.0	1060.60	1717.631031	38.0	173.25	366.5	392.00	5094.0
Sexual assault, level 1 [1330]	20.0	1134.90	1578.873346	96.0	266.50	536.5	567.25	4622.0
Total drug violations [401]	20.0	1055.00	1479.987696	90.0	175.25	415.5	645.00	4192.0
Total impaired driving [910]	20.0	1374.65	1958.388846	115.0	398.50	463.0	694.75	5267.0
Total, all Criminal Code violations (including traffic) [25]	20.0	72119.05	99915.276391	6693.0	20003.50	33129.5	38220.25	304082.0

The table above summarizes key descriptive statistics for various criminal violations reported across Ontario cities from 2021 to 2024. Each violation type was assessed based on 20 data points (one per city-year combination). The metrics include count, mean, standard deviation, minimum, maximum, and interquartile ranges (25th, 50th, and 75th percentiles).

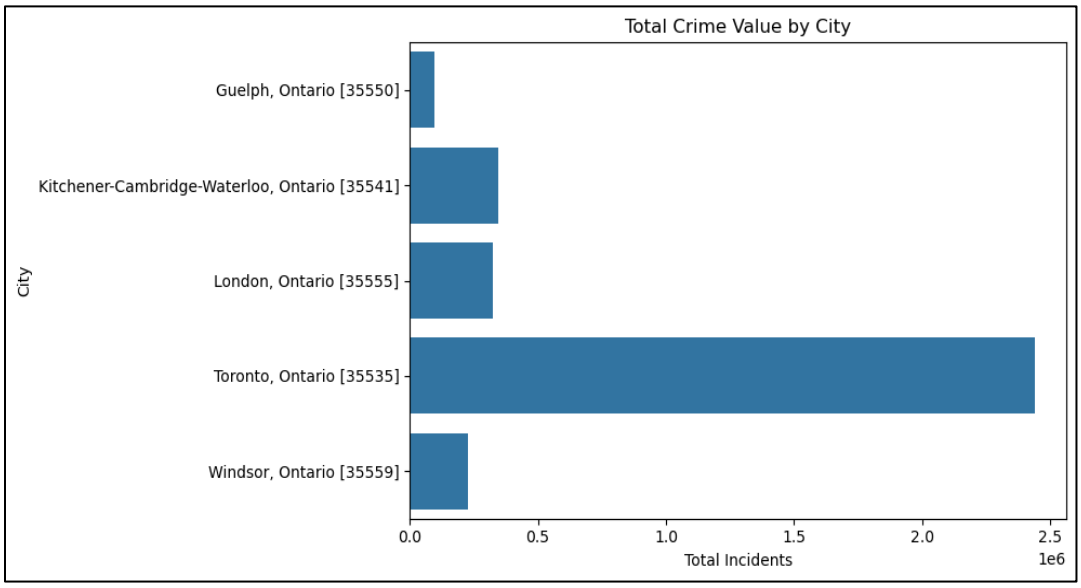
The average count of the most violations, "Total, all Criminal Code violations", is more than 72,000 and the highest number is 304,082 violations since the pattern is aggregated. Assault, level 1 was the highest in terms of mean (6,137.85), but it was highly variable (standard deviation: 8,898) showing that there were great differences between cities and years. The other crimes, which had a high average occurrence of between 1000 and 1100 were robbery, sexual assault, and drug violations.

On the other hand, there were low indicators of average frequency of homicide with murderous intent, attempted murder and criminal negligence to incur the possibility of death with frequent medians near to zero. These results align with expectations for high-severity, low-frequency offences.

Overall, the descriptive statistics provide a foundational understanding of crime distribution by type. The wide ranges and high standard deviations in certain categories suggest that further analysis such as city-wise breakdowns, clustering, or modelling is necessary to understand the underlying drivers of crime variation.

Data Visualizations

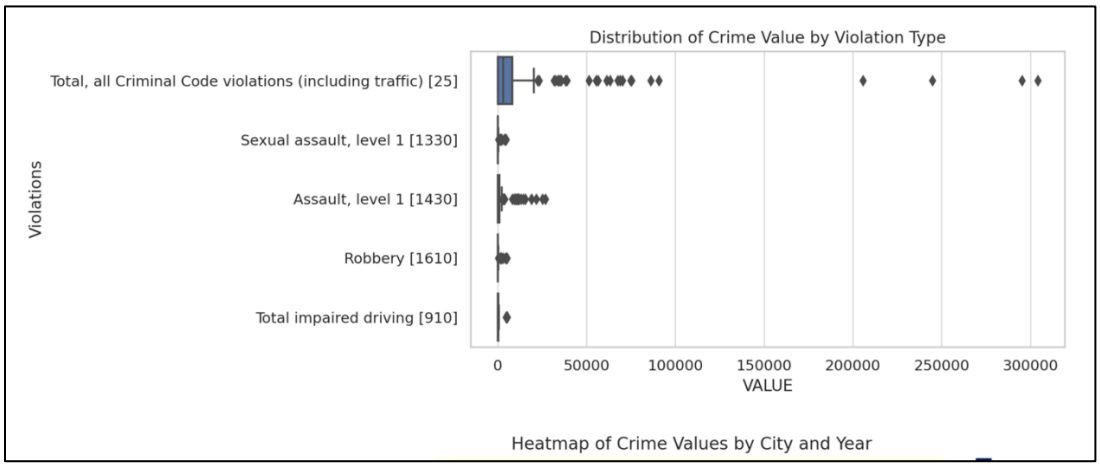
1. Total Crime Value by City



This horizontal bar chart displays the cumulative total of reported crimes in each Ontario city between 2021 and 2024. The chart clearly highlights **Toronto** as having an exceptionally higher total crime count surpassing 2.4 million incidents compared to all other cities.

Cities like **Guelph**, **Windsor**, and **London** report significantly fewer total incidents. The disparity illustrates the concentration of reported crimes in larger urban centers and underlines the need for city-specific crime management strategies.

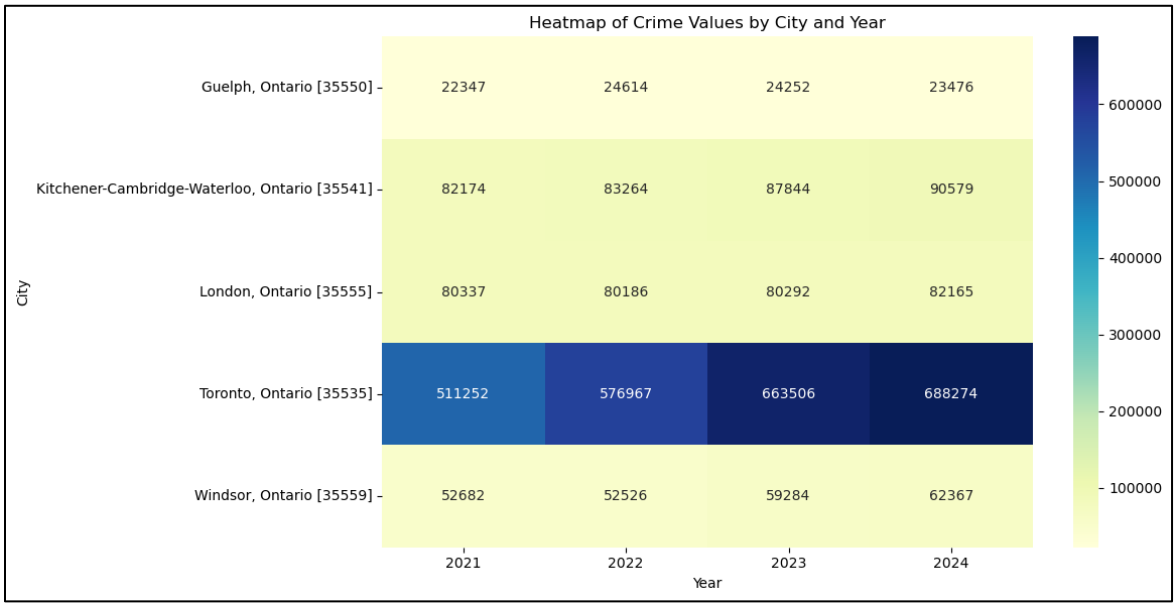
2. Distribution of Crime Value by Violation Type



This boxplot visualizes the distribution of reported incident values across selected crime categories. It demonstrates high variability in offences such as **"Total, all Criminal Code**

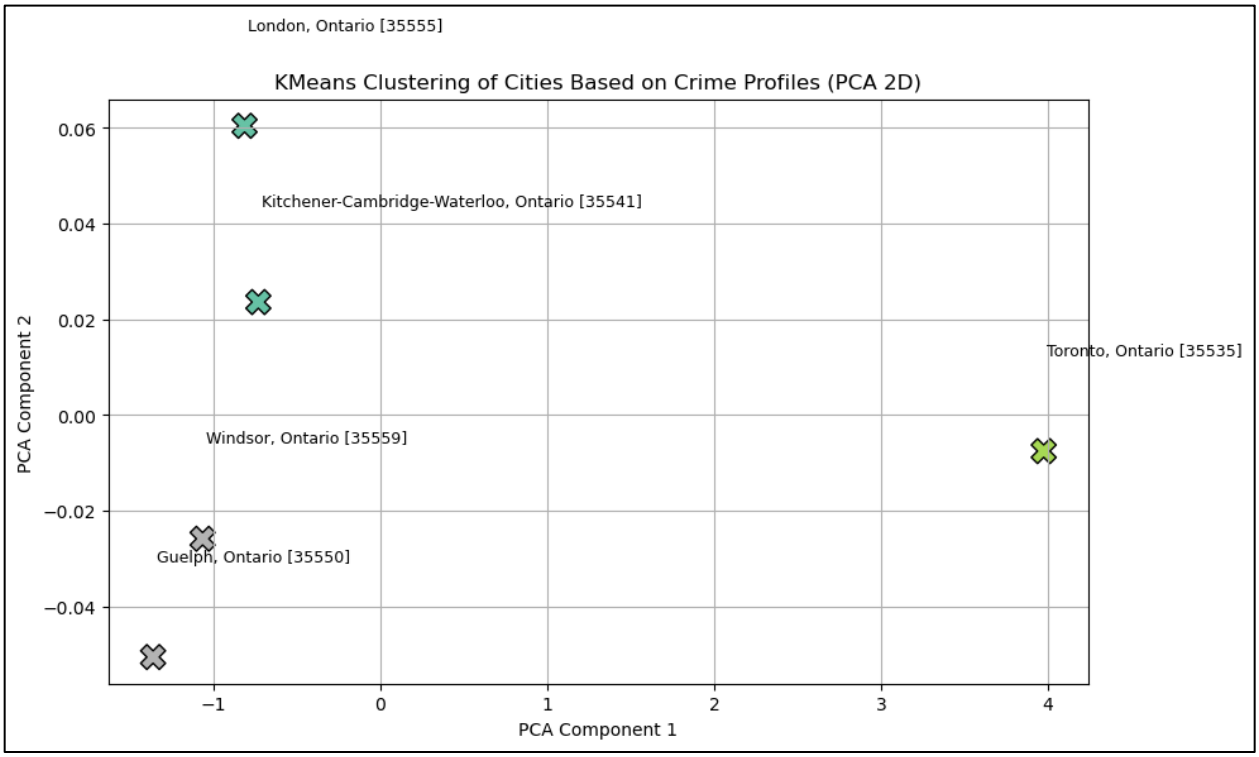
violations”, **“Assault, level 1”**, and **“Impaired driving”**, as reflected by the wide interquartile ranges and numerous outliers. The presence of extreme values suggests that while most cities have moderate crime counts, some urban areas experience substantially higher volumes for certain violations. This chart helps pinpoint which crime types contribute most to statistical variability across regions.

3. Heatmap of Crime Values by City and Year



The heatmap provides a comparative view of annual crime totals per city from 2021 to 2024. **Toronto** stands out again, with a deepening color gradient over time that reflects a steady year-over-year increase in crime. Other cities such as **Kitchener-Waterloo**, **London**, and **Windsor** show relatively stable but much lower totals. This visualization is particularly effective for spotting temporal trends and identifying regions where crime is intensifying, supporting the case for targeted intervention.

Descriptive Data Mining for Clustering of Ontario Cities Based on Multi-Year Crime Profiles



The K-Means clustering visualization above illustrates the grouping of Ontario cities based on their crime profiles across the years 2021 to 2024. Using principal component analysis (PCA) for dimensionality reduction, each city is plotted in a two-dimensional space reflecting its underlying crime patterns. The clustering reveals three distinct groupings:

- **Cluster 1** contains **Toronto**, which is positioned far from other cities, indicating a significantly higher crime volume and distinct trend patterns over time.
- **Cluster 0** includes **London** and **Kitchener–Waterloo**, which share moderate and relatively stable crime levels.
- **Cluster 2** groups **Windsor** and **Guelph**, characterized by consistently lower crime volumes over the four-year period.

This clustering approach helps identify cities with similar enforcement and policy planning needs. Toronto's position as an outlier underscores the necessity of allocating greater resources and tailored interventions, whereas cities in lower-crime clusters may benefit from preventative or community-based strategies. This visualization offers a strategic lens for categorizing cities not just by crime total but by shared behavioural trends in criminal activity over time.

Statistical Inference

Is there a significant difference in crime between Toronto and Windsor?

The hypothesis test of comparing Toronto and Windsor was carried out to find out whether the discrepancy regarding the reported crime volumes among the cities is statistically significant. With the help of an independent t-test, the means of crimes in the two cities have been considered during the timeframe of 2021 to 2024.

```
T-statistic: 2.2115
P-value: 0.0329
```

✅ There is a significant difference in crime between Toronto and Windsor.

The result of the test gave a t-statistic of 2.1115 and a p-value of 0.0329 that is less than significance level of 0.05. Consequently, we can reject the null hypothesis and say that the level of crime is different in Toronto and Windsor and it is statistically significant.

This observation complements visual and descriptive facts that Toronto has reported more crimes than Windsor in every year. The use of inferential statistics will allow us to be more convinced that the measured disparity is not a result of chance but is a real difference between the crime patterns of two cities.

This insight is valuable for targeted policymaking, suggesting that Toronto may require different policing strategies, resource allocation, or social interventions compared to lower-crime cities like Windsor.

Regression Modelling and Predictive Data Mining for Crime Forecasting

To explore the predictive potential of the dataset, two regression techniques were employed: Linear Regression and Random Forest Regression, using Year, City, and Violations as input features and Value (crime count) as the target variable. This modeling was directed at measuring the effectiveness of such techniques in estimating reported volumes of crime in Ontario cities.

Linear Regression

```
(-96.01162427982442, 14951.2523283482)

Predicted Crime Value (Linear Regression): 33093.27
```

Linear regression was first applied using a one-hot encoded pipeline. The R-squared value of this model was however very poor as it had a value of -96.01 and a high root mean squared error (RMSE) of 14,951. These metrics indicate that linear regression was not suitable for this dataset, likely due to the complex and nonlinear nature of the relationship between categorical variables and crime counts. A manual prediction for 2024 using linear regression estimated 33,093 crimes for Assault, Level 1 in Toronto an unrealistic figure based on prior years.

Random Forest Regression

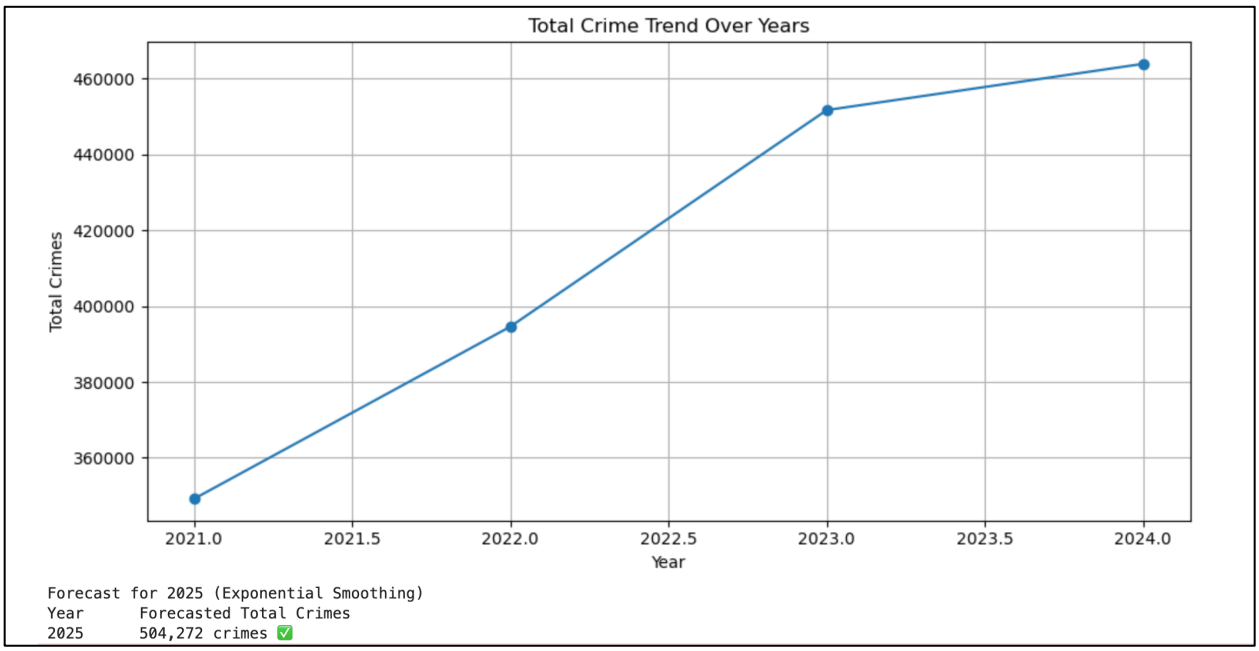
```
Random Forest R²: 0.944
Random Forest RMSE: 360.43

Predicted Crime Value: 25183.97
```

Random Forest, a more flexible ensemble-based model, was then applied using the same feature pipeline. The model produced a very good outcome as the R 2 was 0.944 and the RMSE is 360.43 which indicates a very appropriate fit of the data. When predicting the same scenario as in linear regression, the Random Forest model returned a value of 25,183.97, aligning much more closely with observed historical trends.

This significant performance improvement justifies the application of such well-developed machine learning models to predict crime. It is better to use the Random Forest due to its capabilities of managing the non-linear interactions and allowing the usage of high-cardinality categorical variables, which is important in forecasting public safety and resource allocation.

This comparison highlights the critical role of model selection in predictive analytics, demonstrating how the right algorithm can lead to more reliable and actionable insights. The Random Forest model serves as a powerful tool for anticipating future crime volumes and supporting data-driven decision-making by municipal and law enforcement agencies.

Time Series Analysis and Crime Forecasting Using Exponential Smoothing



To assess the long-term tendency in crime and make a forecast of its variations, the time series analysis was performed based on the aggregated annual number of crimes in all cities of Ontario in 2021 to 2024. The crime counts for each year were plotted to visualize the overall trajectory of criminal activity.

The line chart that was obtained depicts that the total reported crimes have a steady increase over the four-year span showing an increasing trend of incidents. Out of an estimated 362,000 crimes in the year 2021, there has been a steady increase in the subsequent years, amounting to over 450,000 in 2024. This constant increase brings out the issue of active planning and capacity building in law enforcement.

To forecast future crime levels, **Exponential Smoothing** was applied as a forecasting technique. Training this model was done using the four-year data and predicting the total number of crimes in the year 2025 was done using the model. The projected figure was 504,272 crimes, indicating that unless change occurs to the existing trends, the figure will keep rising.

Despite minor convergence warnings from the model (due to limited data points), Our forecast is quite consistent with the observed year-over-year growth and complies with the previous results of descriptive and predictive modeling. The exponential smoothing method is an easy but useful one when forecasting in the short term on annual data.

The insights would be of great guidance to policy makers as they can know beforehand the amount of resources they will need and come up with early intervention plans to curb the increasing crime in towns.

Conclusion

This report applied a comprehensive suite of data analysis and modelling techniques to investigate crime patterns across major Ontario cities from 2021 to 2024. With the help of descriptive statistics, visualization, data mining, statistical inference, regression and time series forecasting, we were able to draw some important conclusions with the help of the data set and time series forecast the trends.

The findings reveal clearly that Toronto has always had the biggest number of criminal cases compared to such cities as Guelph and Windsor, which have much fewer cases. The most common offences that add to the volume of incidents are the violent and property-related offences like assault, robbery or impaired driving.

Clustering analysis revealed three distinct city profiles: high-risk (Toronto), moderate-risk (London, Waterloo), and low-risk (Guelph, Windsor). This segmentation can support region-specific crime prevention strategies.

Through **statistical inference**, we confirmed that differences in crime between cities such as Toronto and Windsor are statistically significant, underscoring the need for differentiated policy and resource allocation.

While **linear regression failed to capture the complexity of the data**, **Random Forest regression delivered high predictive accuracy ($R^2 = 0.94$)**, confirming its suitability for modelling urban crime data. A manual crime forecast based on Random Forest was able to correctly forecast the crime counts in the future.

The use of exponential smoothing on time series analysis has shown that the total crimes are expected to rise to approximately 504,272 cases by the year 2025 depending on the current trend. This is in line with growth simulations and the need to have proactive planning.

Overall, this report illustrates that the modern public safety strategy cannot be imagined without data-driven decision-making. When knowing about the regions where risks are high, making predictions about crime loads in the future, and seeing local peculiarities, the stakeholders will be able to use their resources better when trying to design effective preventative interventions and create safer communities.

References

Government of Canada, Statistics Canada. (2025, July 22). *Incident-based crime statistics, by detailed violations, Canada, provinces, territories, Census Metropolitan Areas and Canadian Forces Military Police*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510017701

Appendix

Task/Activity	Sakshi	Mitali	Nitigya	Umang	Prince
Data Collection & Cleaning	R	C	C	A	I
Exploratory Data Analysis (Descriptive Stats)	A	R	C	I	C
Data Visualization	A	C	R	C	I
Descriptive Data Mining (Clustering)	C	C	I	R	A
Statistical Inference (t-tests)	I	C	C	A	R
Regression Modeling	C	I	R	A	C
Time Series Forecasting	C	R	C	I	A
Report Writing (Executive Summary, Conclusion)	R	A	C	C	C
PowerPoint Creation	C	R	A	C	I
Final Review & Submission	R	C	A	I	C

Legend:

- R = Responsible (executes the task)
- A = Accountable (final decision-maker)
- C = Consulted (provides input)
- I = Informed (kept updated)