Group10 - Analysis on Motor Theft in Toronto neighbourhoods

Authors

Group 10 - Thibault Clement, Kuo, Sheng Peng, Mohan, Shriraksha, Parthasarathy, Sathyanarayanan, Sharma, Rijul

Contents

[**1.Background and Objectives 2**](#_5hwlnrmchg97)

[**2. Exploratory Analysis 2**](#_g2jh8anmc731)

[2.1 Dataset 2](#_papxa4z82o1a)

[2.2 Time-Series Analysis 3](#_3b4a80yrq48l)

[2.3 Geo-Spatial Analysis 5](#_vrnx6knel3h5)

[**3. Predictive Modeling 6**](#_mzdyrhj42toa)

[3.1 Approach](#_f52etubdn6ms) [7](#_s68h8z1uas30)

[3.2 Model Evaluation 7](#_c95o16u16jfv)

[**4. Results & Discussion 8**](#_moufm0mqwjlf)

[**5. Conclusion 8**](#_u2fyo7pixnyj)

[**Appendices 8**](#_nynw5u91n81v)

### 

### 1.Background and Objectives

This study aims to investigate trends and patterns in vehicle theft across Toronto, with the goal of developing a predictive model to identify high-risk neighborhoods for future theft incidents. Initially, we will perform an exploratory analysis, incorporating both time-series and geo-spatial methods to gain a deeper understanding of how vehicle thefts have evolved over time and how they vary by neighborhood. The time-series analysis will allow us to track theft trends across different time intervals, while the geo-spatial analysis will identify geographical hotspots and spatial patterns of theft.

Building upon these findings, we will then develop a predictive model that can assess the likelihood of vehicle theft in specific neighborhoods at any given time. This model will integrate historical theft data, time-of-day, seasonal patterns, and spatial variables to generate risk scores for each area, offering a quantitative approach to forecasting future theft incidents.

To achieve this, we will leverage data from Toronto’s Open Data portal, which includes detailed information on vehicle theft incidents, including locations, timestamps, and crime types. By analyzing this data, the study aims to provide valuable insights into underlying crime patterns, which can inform more effective law enforcement strategies. Specifically, the predictive model will enable the optimization of patrol schedules and resource allocation, helping law enforcement agencies focus on high-risk areas and times, ultimately enhancing crime prevention efforts.

In addition to supporting tactical crime prevention, the findings from this study can inform broader policy decisions related to urban planning, neighborhood safety, and public awareness campaigns. By anticipating where and when thefts are most likely to occur, Toronto’s police force can better allocate resources, prevent crime, and enhance public safety.

### 2. Exploratory Analysis

#### 2.1 Dataset

The dataset utilized for this analysis encompasses all reported incidents of *Theft from Motor Vehicles* in Toronto, offering detailed data on each individual occurrence. It includes records for two types of offenses: *Theft from Motor Vehicle Under* and *Theft from Motor Vehicle Over*, both categorized by the date the theft was reported. The dataset is organized at the offense and/or victim level, meaning a single theft incident may be represented across multiple rows if it involves different offense types linked to the same event. This structure allows for a more granular analysis of each theft, accounting for all aspects of the incident.

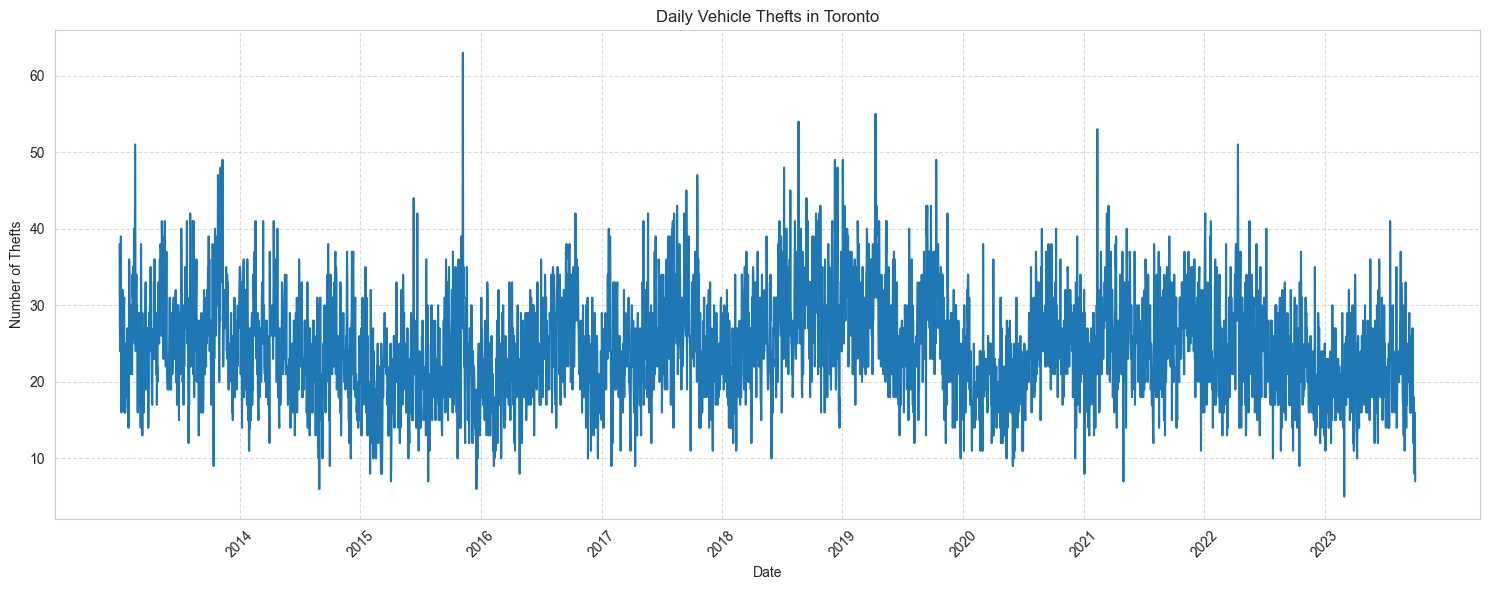
This comprehensive dataset serves as the foundation for several key aspects of the study, including the analysis of long-term trends in vehicle theft, the identification of spatial and temporal variations in theft rates across different neighborhoods, and the development of predictive models for future theft incidents. Importantly, the dataset excludes cases that have been classified as unfounded—those where law enforcement determined that the offense either did not occur or was not attempted.

In addition to data on theft incidents, the dataset also includes all *Motor Vehicle Crime Investigation* (MCI) occurrences reported to the Toronto Police Service. This ensures that even cases where location data may be incomplete or unverifiable are included in the analysis. As a result, some records may contain missing geographic coordinates or reference incidents that occurred outside the City of Toronto’s boundaries. While this can introduce some noise into the dataset, it remains valuable for understanding broader patterns and trends.

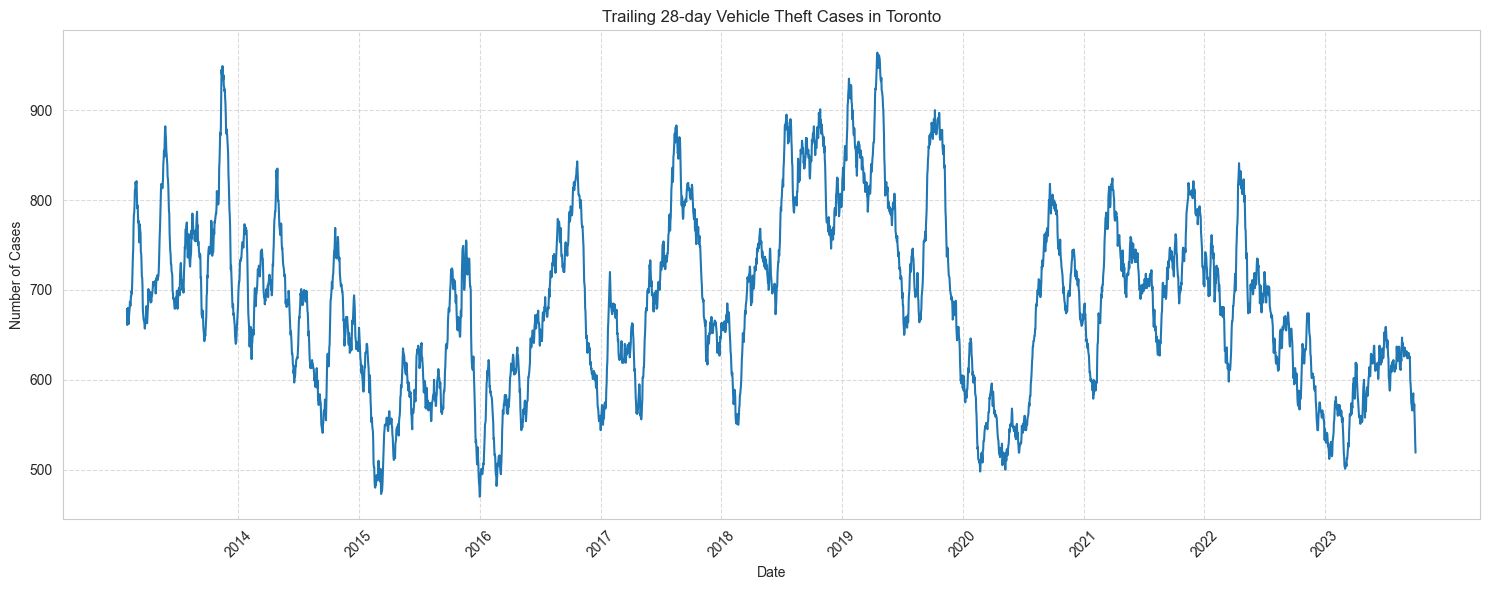
#### 2.2 Time-Series Analysis

**Figure 1** below displays the daily count of vehicle thefts, which highlights significant day-to-day fluctuations, making the data appear quite noisy. To better understand long-term trends, **Figure 2** shows the trailing 28-day data, which smooths out short-term variations like those caused by the day of the week. This reveals clear trends, including a sharp peak around 2020, likely driven by the COVID-19 pandemic. After this spike, thefts declined substantially, possibly due to changes in societal behavior and law enforcement efforts. The data doesn't follow a consistent seasonal pattern, but instead shows distinct phases of increase and decrease. The most recent data from 2023-2024 indicates a downward trend, which could be linked to improved security measures, law enforcement strategies, or broader socioeconomic factors.

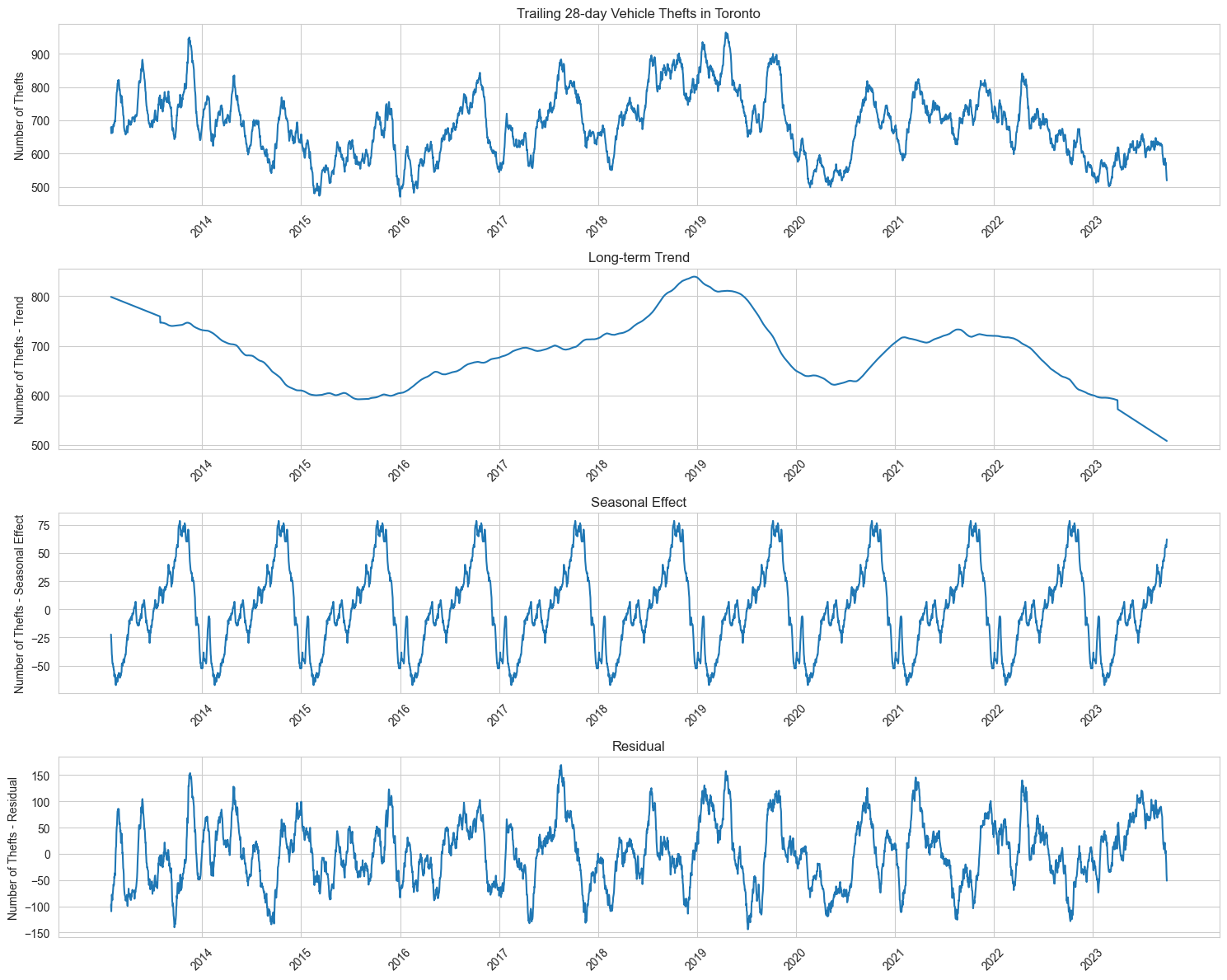
**Figure 1 - Daily Vehicle Thefts in Toronto**



**Figure 2 - Daily Vehicle Thefts in Toronto - Cumulative Trailing 28-day**



As a second step, we perform a seasonal decomposition analysis to gain a deeper understanding of long-term trends and seasonal effects. The chart in **Figure 3** highlights several important patterns. First, the long-term trend reveals significant variations, with a peak around 2019 at approximately 850 cases every 28 days, followed by a sharp decline. There was a brief rebound in 2021, but since 2022, there has been another decline. The COVID-19 pandemic and its socio-economic effects likely played a major role in these fluctuations, influencing both criminal behavior and broader societal conditions. The seasonal component of the data, also shown in **Figure 3**, demonstrates a strong yearly cycle, with thefts peaking in the summer and dropping sharply in winter. This makes sense in a city like Toronto, where extreme winter conditions can create physical barriers to theft, while milder weather in the summer may facilitate more criminal activity. Lastly, the residuals (representing variations not explained by long-term trends or seasonality) remain significant. This suggests that there is still a level of randomness and unpredictability in the data, with some thefts occurring due to factors not captured by the trend or seasonal patterns.

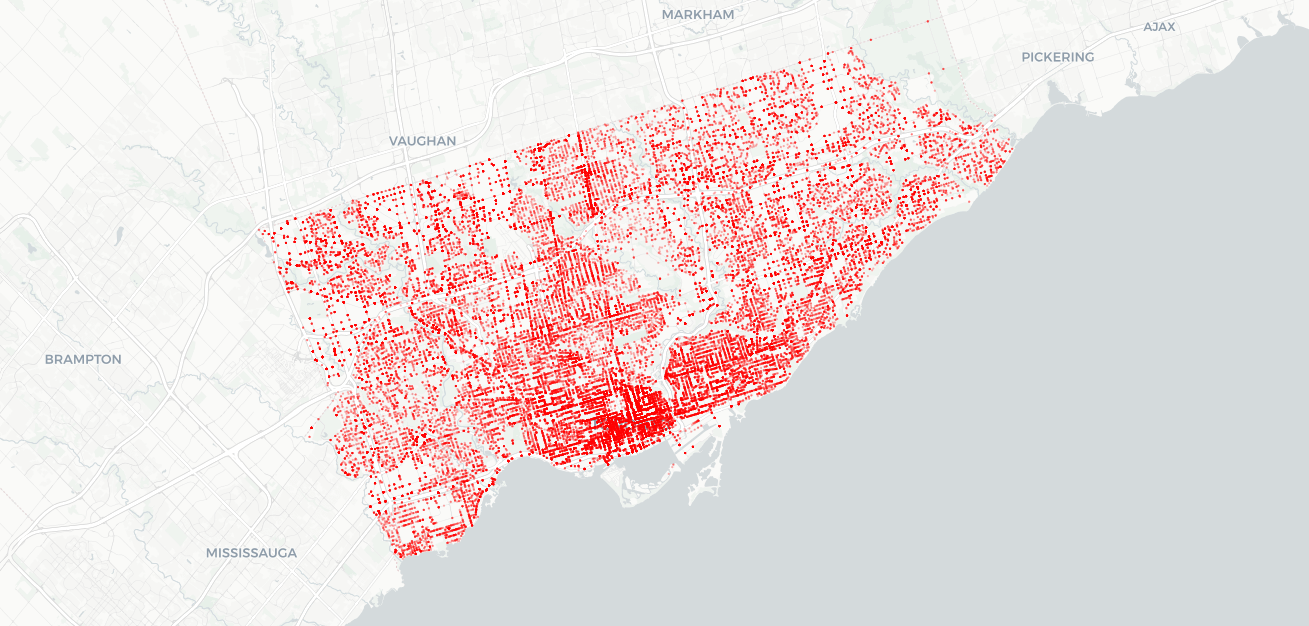
**Figure 3 - Seasonal Decomposition**

In addition to the seasonal decomposition analysis, we also analyzed Autocorrelation and Partial Autocorrelation Function plots (not included here). This analysis showed a strong correlation up to lag 100, indicating that some underlying factors might have lingering effects on the number of thefts for several months. This is in line with the presence of strong trends identified through the seasonal decomposition analysis.

#### 2.3 Geo-Spatial Analysis

In this section, we analyze the geo-spatial distribution of vehicle thefts in Toronto. Using the available longitude and latitude data from the dataset, we plot each theft incident on a map. **Figure 4** illustrates that thefts are concentrated in specific areas, with a notable clustering in the downtown core. This suggests that vehicle thefts are more prevalent in densely populated urban regions, likely influenced by factors such as higher vehicle density, increased foot traffic, and potentially lower levels of surveillance or security.

**Figure 4 - Geographical distribution**



Next, we look at the distribution of the number of thefts by neighborhood using the ‘HOOD\_158’ column of the dataset, which includes a unique identifier for each of the 158 Toronto neighborhoods.

**Figure 5 - Distribution by Neighborhood**

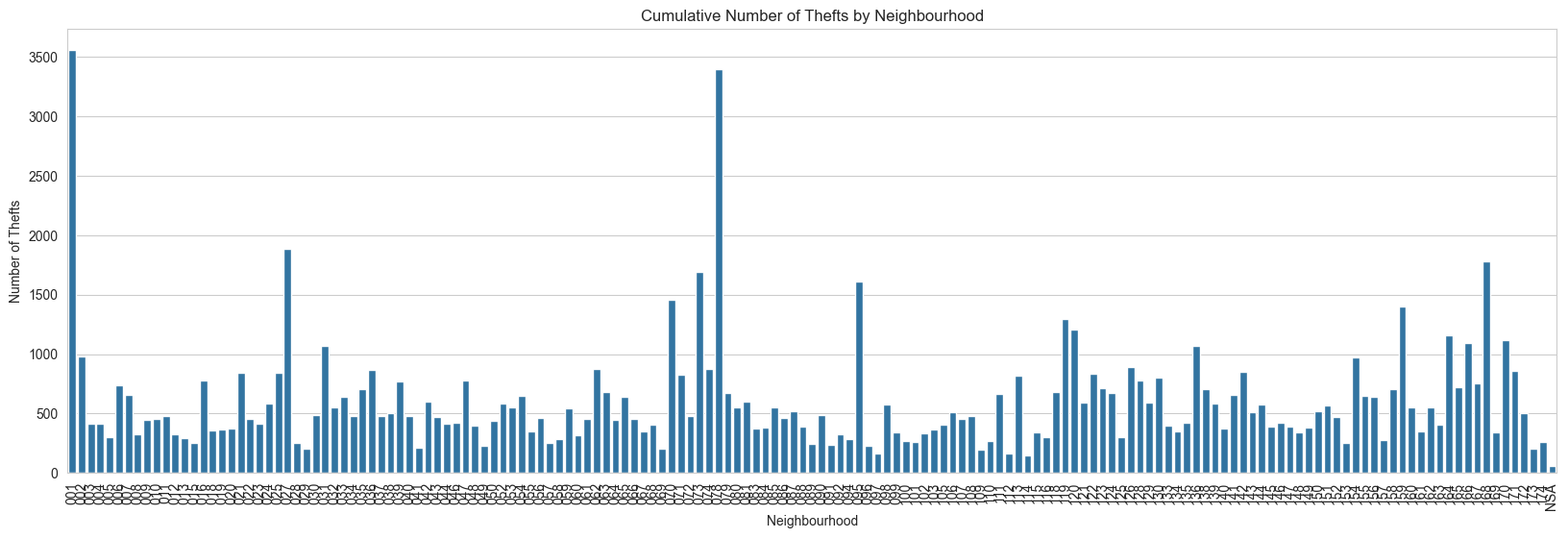


Figure 5 shows that certain neighbourhoods concentrate a much higher number of thefts over the years, with neighborhood 001 (West Humber-Clairville) totalling 3,500 cases, and neighborhood 114 (Lambton Baby Point) only 144 cases.

### 3. Predictive Modeling

#### 3.1 Approach

In order to develop a model to predict the count of motor vehicle thefts, the data was processed to achieve the following outputs:

* Grouped by location and date: Aggregated the number of incidents (incident\_count) at each

unique location (longitude, latitude) and date.

* Calculated a 28-day trailing count: For each location, a rolling window of 28 days was used

to compute the cumulative number of incidents.

Next, data was prepared for regression analysis by using the trailing 28-day count as the target

variable and identifying potential predictors.

Based on the data, the potential predictors were found to be:

* Location of thefts represented by Latitude and Longitude parameters
* Report Date represented by Year,Month and, Day

Once identified, the data was split for testing and training purposes. Using linear regression, a model was trained on the predictors to predict the trailing 28-day theft count.

#### 3.2 Model Evaluation

**Analyzing the Regression Model**

The linear regression model tries to predict the trailing 28-day count of motor vehicle theft incidents based on several features (longitude, latitude, year, month, and day of the week).

1. Model Fit:

Coefficients: By examining the coefficients of the model, we can see how each feature influences the prediction. For example, if the latitude coefficient is positive, it means that higher latitude values are associated with a higher trailing 28-day theft count.

Intercept: The intercept gives the baseline level of theft when all predictors are zero.

By examining the coefficients of the regression model, we can identify which features have the most significant impact on the trailing 28-day theft count. For instance, if the day of the week has a large coefficient, it means that certain days may have higher theft incidents.

2. Predictions:

On comparing the predicted values (`y\_pred`) with the actual values (`y\_test`), we get an idea of how well the model is performing.

Mean Squared Error (MSE)

The MSE is a measure of how well the model's predictions match the actual data. It calculates the average squared difference between the predicted and actual values.

Interpretation: Lower MSE values indicate better performance as they mean that the predictions are closer to the actual values. A higher MSE indicates a larger average error between the predicted and actual values. This could be due to several reasons, such as the model not capturing all the important factors, or the relationship between the predictors and the target variable not being linear.

R Squared (Coefficient of determination): The R squared value explains the amount of variance that the model is able to explain or account for. Lower R squared values indicate that the linear regression model is not capturing much of the variance. This could be due to the limited set of predictors or the nature of the problem (possibly non-linear relationships).

### 4. Results & Discussion

We have successfully trained a **Linear Regression** model to predict the 28-day trailing theft count at specific locations. Here are the evaluation metrics for the test set:

* **Mean Squared Error (MSE)**: 74.76 — This is the average squared difference between predicted and actual values.
* **R² (coefficient of determination)**: 0.123 — This indicates that the model explains about **12.3%** of the variance in the 28-day trailing theft counts.The R² score (0.123) indicates that the linear regression model is not capturing much of the variance. This could be due to the limited set of predictors or the nature of the problem (possibly non-linear relationships).
* **Possible Improvements**:
  + Include additional predictors (like neighborhood, premises type, and prior counts).
  + Use non-linear models such as Random Forest or Gradient Boosting Regressors.
  + Investigate potential seasonal or cyclical patterns.

### 5. Conclusion

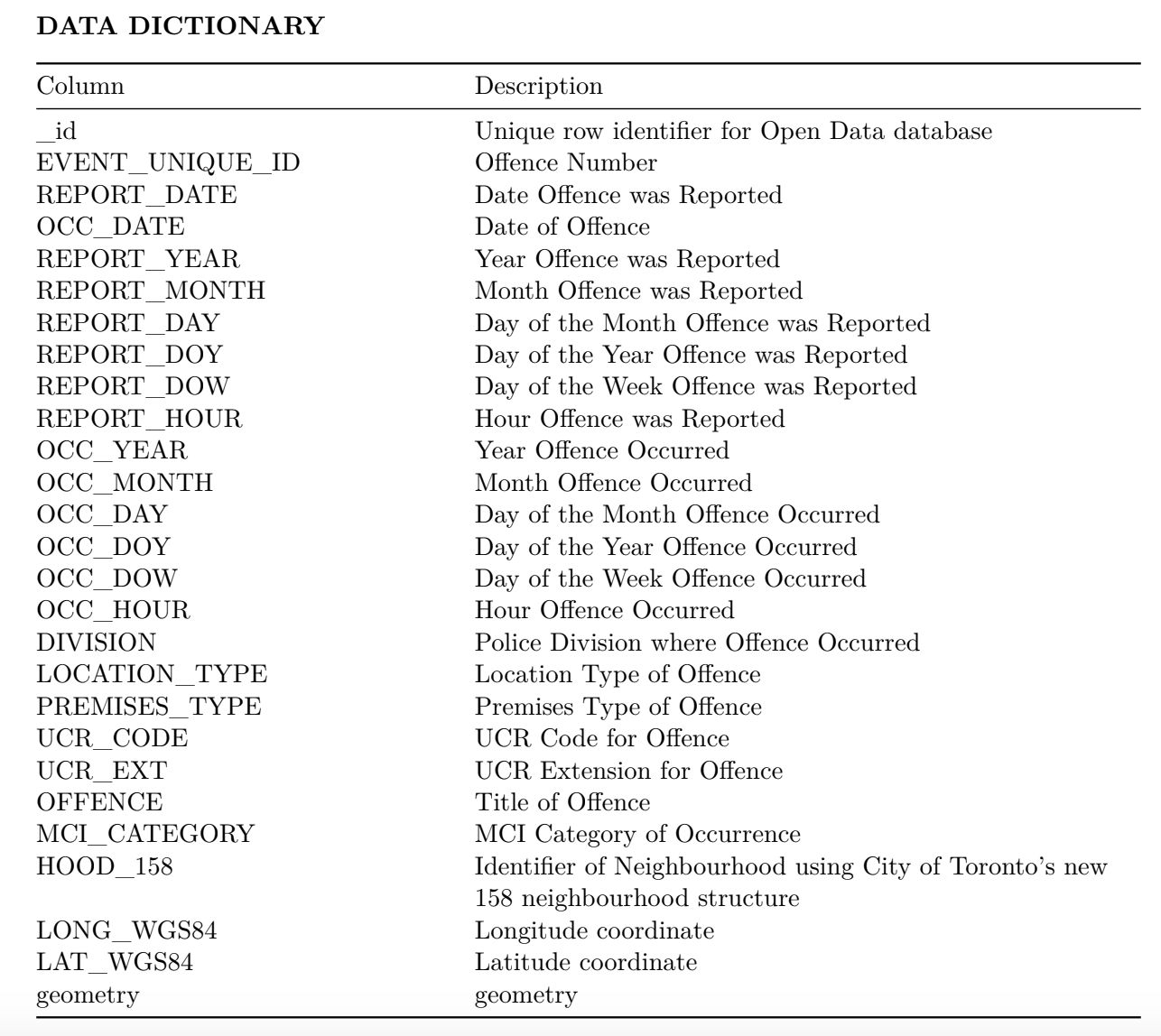
The temporal analysis reveals significant fluctuations over the years, with a notable peak occurring around 2020, followed by a marked decline that coincided with societal changes during the COVID- 19 pandemic. Through seasonal decomposition, we observed that while seasonal patterns exist, they are relatively modest, with seasonal effects ranging from -67 to +78 incidents. The most recent data from 2023-2024 shows an encouraging downward trend in theft cases, which may reflect the success of improved security measures or enhanced law enforcement strategies.

The linear regression model developed to predict trailing 28-day theft counts achieved an R-squared value of 0.123, indicating that approximately 12.3% of the variance in theft counts can be explained by our current set of predictors. While the model’s Mean Squared Error of 74.76 suggests moderate prediction accuracy, the relatively low R-squared value indicates substantial room for improvement in predictive power. This suggests that vehicle theft patterns are influenced by additional factors not captured in our current model.

Geographic location emerged as a significant factor in theft patterns, as evidenced by the substantial coefficients for longitude (-2.99) and latitude (-18.03). The yearly trend coefficient of 1.03 indicates a slight upward trend in thefts over the years when controlling for other variables. Interestingly, monthly and day-of-week variables showed minimal impact on theft patterns, with coefficients of 0.047 and 0.069 respectively, suggesting that temporal cycles play a less significant role than location in predicting theft incidents.

### 

### Appendices



#### 