Weather and Crime in Toronto: Analyzing the Impact of Meteorological Conditions on Criminal Activity

Richard Guo

2025-03-14

Introduction

Understanding the factors that influence crime rates is a critical area of study in criminology, urban planning, and public policy. Among the various potential influences, weather conditions have been hypothesized to play a role in influencing crime rates. Certain environmental factors—such as temperature, precipitation, and visibility—may influence human behavior, affecting both criminal activity and law enforcement response times. For example, warmer temperatures might lead to an increase in violent crimes due to heightened social interactions, while heavy rainfall or snowfall could deter outdoor criminal activities. This project aims to explore the relationship between outdoor weather conditions and crime rates in Toronto (specifically Downtown Yonge East and the Bay-Yonge Corridor) by using a combination of Random Forests and XGBoost models to explore a merged dataset of both crime and weather statistics.

Data

To conduct this analysis, I will use two primary datasets:

- Toronto Police Service's Major Crime Indicators Dataset – The Toronto Police Service (TPS) provides publicly accessible crime data through its open data portal. The Major Crime Indicators dataset includes records of reported criminal incidents across the city, categorized into key crime types such as assault, robbery, and theft. Each entry contains information on the type of crime, the date and time it was reported, and the geographic location (latitude and longitude). This dataset is sourced from official police reports, making it a highly reliable indicator of crime trends. However, some limitations exist: crime reports may be influenced by reporting biases, police resource allocation, and differences in how crimes are classified over time. Each API call is also restricted to 2000 entries, meaning multiple API calls may be required. Additionally, minor crimes or unreported incidents are not captured, meaning the dataset represents a subset of all criminal activity in the city.
- Open-Meteo Historical Weather Data – The second dataset is obtained from Open-Meteo's historical weather API. This dataset provides high-resolution hourly weather measurements for Toronto, including temperature, humidity, precipitation, snowfall, wind speed, and visibility. It is arranged by the hour (across all of 2024), and restricted to the coordinates (43.7, -79.4), which correspond to the downtown Toronto area. The data is derived from a combination of meteorological stations, satellite observations, and numerical weather prediction models, ensuring accuracy and consistency across different time periods. The Open-Meteo API aggregates data from multiple sources, including Environment Canada and global meteorological networks, making it a reliable source for historical climate conditions. However, it is important to note that some weather variables—such as localized precipitation or sudden wind gusts—may not always be captured with perfect precision, especially in highly variable urban microclimates.

By merging these two datasets using the date and hour as the common key, I will explore how different weather conditions correspond to changes in crime patterns. This study will employ statistical analysis and data visualization techniques to identify potential correlations and trends. The findings may provide valuable insights into the extent to which environmental conditions influence criminal activity, contributing to a broader understanding of crime dynamics in urban settings. These insights could be of interest to law enforcement agencies, policymakers, and urban planners aiming to improve public safety through data-driven decision-making.

Methods

Data Collection

The datasets used in this analysis were collected from two publicly available sources:

1. Toronto Police Service's Major Crime Indicators Dataset: This dataset was accessed through the Toronto Police Service Open Data API. It provides detailed information on reported crimes, including the type of crime, date and time of occurrence, and geographical location. The data was retrieved in JSON format and converted into an R data frame.
2. Open-Meteo Historical Weather Data: The weather data was obtained from the Open-Meteo API, which provides high-resolution hourly historical weather measurements. The dataset includes temperature, humidity, wind speed, precipitation, and other meteorological variables for Toronto in 2024. The data was also retrieved in JSON format and converted into a structured data frame for analysis.

Data Preprocessing

Before merging the datasets, several preprocessing steps were performed to clean and format the data.

For the crime dataset:

- The Toronto Police Service Open Data API only allows for queries that return a maximum of 2000 rows. In order to capture a full year's worth of data, the API was called twice; once with neighborhood Downtown Yonge East, and once with Bay-Yonge Corridor. The two resulting dataframes were then binded together.
- The OCC_MONTH field was originally stored as a month name and was converted into its corresponding numeric value.
- A new date variable was created by combining year, month, day, and hour into a single datetime format.

For the weather dataset:

- The weather data was structured into a cleaned dataframe containing only relevant meteorological variables.
- The datetime column, representing the timestamp of each weather record, was formatted as a POSIXct datetime object for consistency.

Data Cleaning and Wrangling

After collecting the crime and weather datasets, several preprocessing steps were necessary to ensure data integrity and prepare the data for analysis. The cleaning and wrangling process involved handling missing values, standardizing variable formats, and filtering unnecessary data.

Handling Missing Values Missing data can impact analysis, so it was important to identify and address any missing values in both datasets. The summary() function was used to check for NA values, and missing values were handled by either removing incomplete records or imputing reasonable estimates when appropriate.

If missing values were found in key variables (e.g., date, crime type, temperature), they were removed to avoid data inconsistencies:

For weather-related variables, missing values were imputed using median values, which assumes that weather patterns remain relatively consistent during short time frames.

Standardizing and Formatting Variables The crime dataset contained categorical fields such as crime type and location, while the weather dataset included numerical meteorological variables. To ensure consistency, categorical variables were converted to factors, and date-time formats were standardized using lubridate.

Filtering and Selecting Relevant Variables To streamline analysis, unnecessary variables (e.g., unique IDs, administrative fields) were removed. Only essential variables related to crime occurrence and weather conditions were retained.

Data Merging

The two datasets were merged using the datetime variable as the common key. An inner join was performed to retain only records that had matching timestamps in both datasets. Before the join, the crime dataset had 1651 observations, and the meteorological data had 8784. After the merge, the combined dataset retains 1651 entries, meaning no data was lost during the merge.

Preliminary Results

We start by taking a look at the number of crimes and average temperature by month, as well as the distributions of temperatures and crime.

Table 1: Table 1: Average Temperature and Number of Crimes per Month (Toronto, ON, 2024)

Month	Average Temperature (°C)	Number of Crimes
January	-2.51	131
February	0.18	134
March	3.22	113
April	7.78	119
May	15.79	145
June	19.35	167
July	22.31	152
August	20.75	163
September	18.11	119
October	11.49	148
November	6.13	139
December	-0.89	121

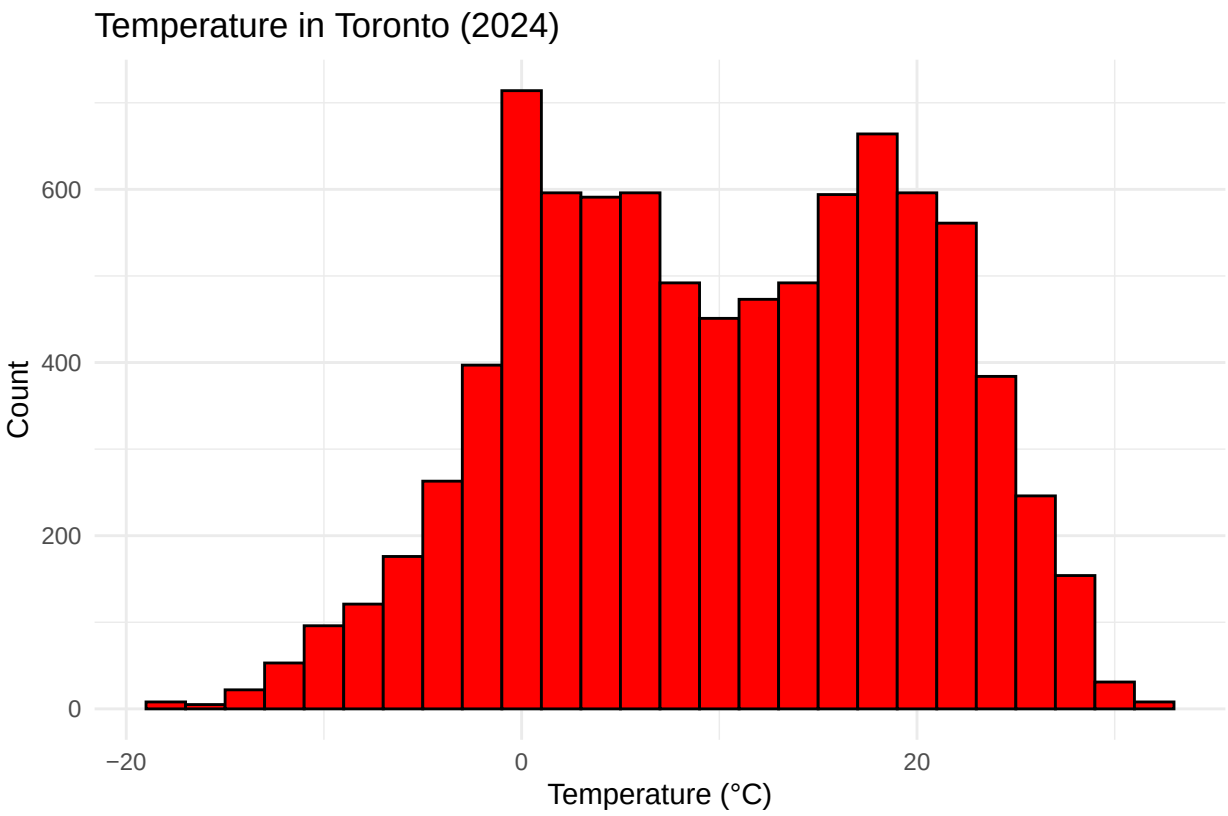


Figure 1: Histogram of Temperature in Toronto (2024)

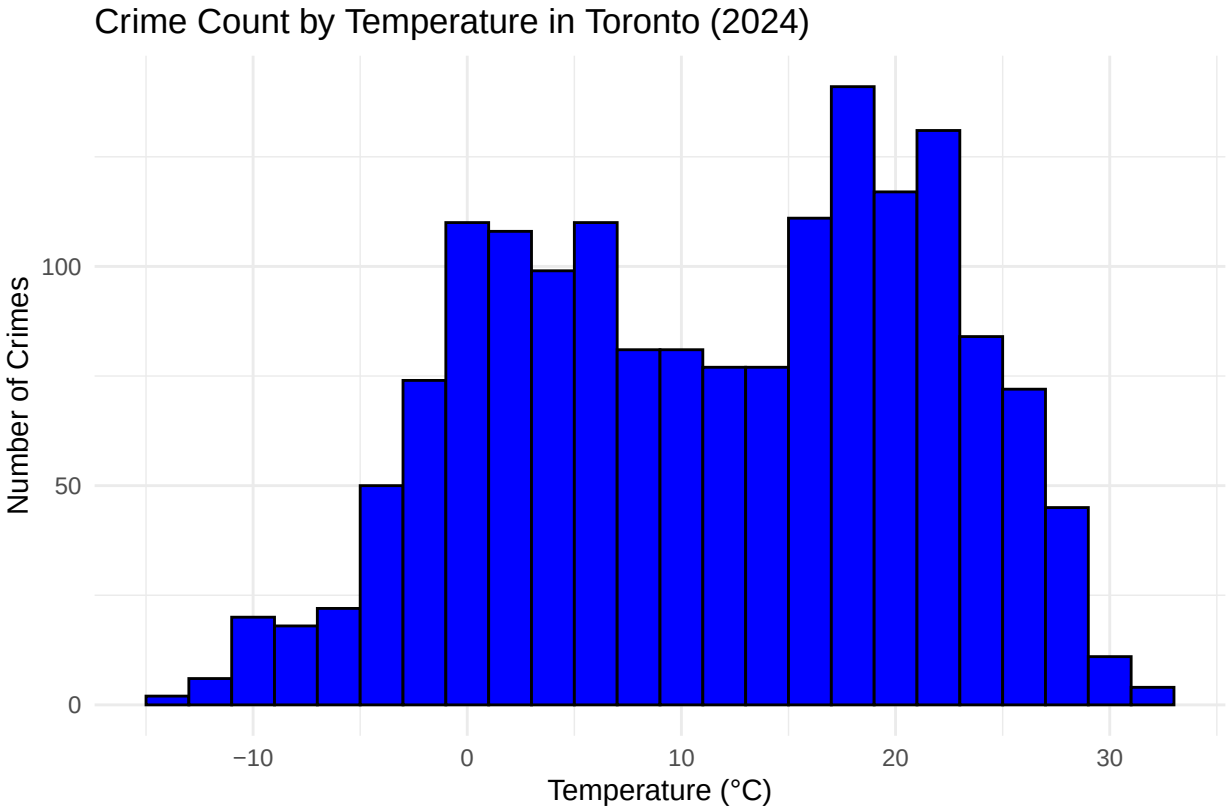


Figure 2: Crime Count by Temperature in Toronto (2024)

It appears that more crimes are committed during higher temperatures. However, it is impossible to deduce any functional relationship between temperature and crime count at this stage, as we have not considered any other factors. For example, it may be easier to commit certain types of offences, such as assault, during warmer seasons, as there are likely more people outside and a lack of snow makes maneuverability significantly easier. It can be difficult to visualize the relationship between temperature and crime from these plots, so we can plot both on separate axis' to more closely examine the relationship.

(See website for plots)

While it does seem as though the number of crimes climbs between the months of April and August, there is an unexpected drop in crimes during September, even though October is when the temperature takes a steep dive. We can also begin to examine the types of crime committed and how they change with the temperature.

Crime Category	Number of Crimes
Assault	1139
Break and Enter	178
Robbery	170
Theft Over	84
Auto Theft	80

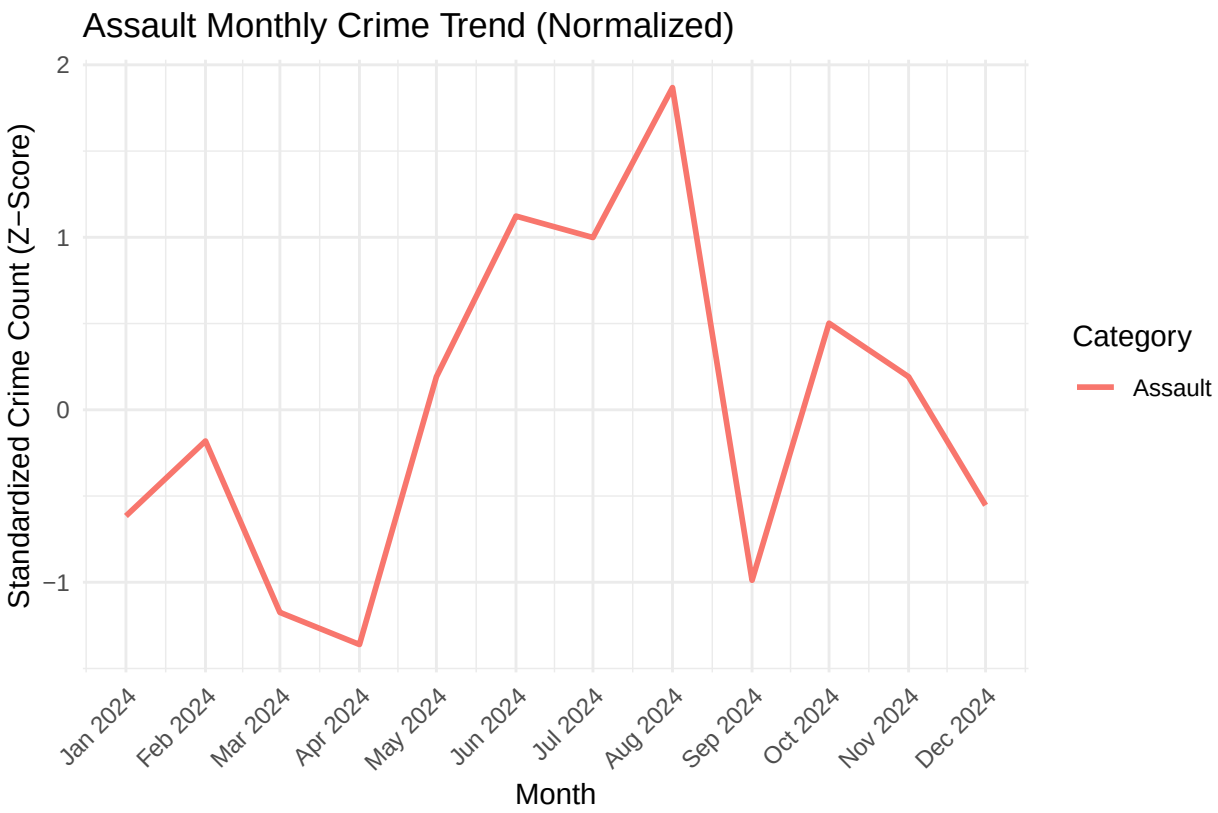


Fig 3: Trend in Assaults during 2024

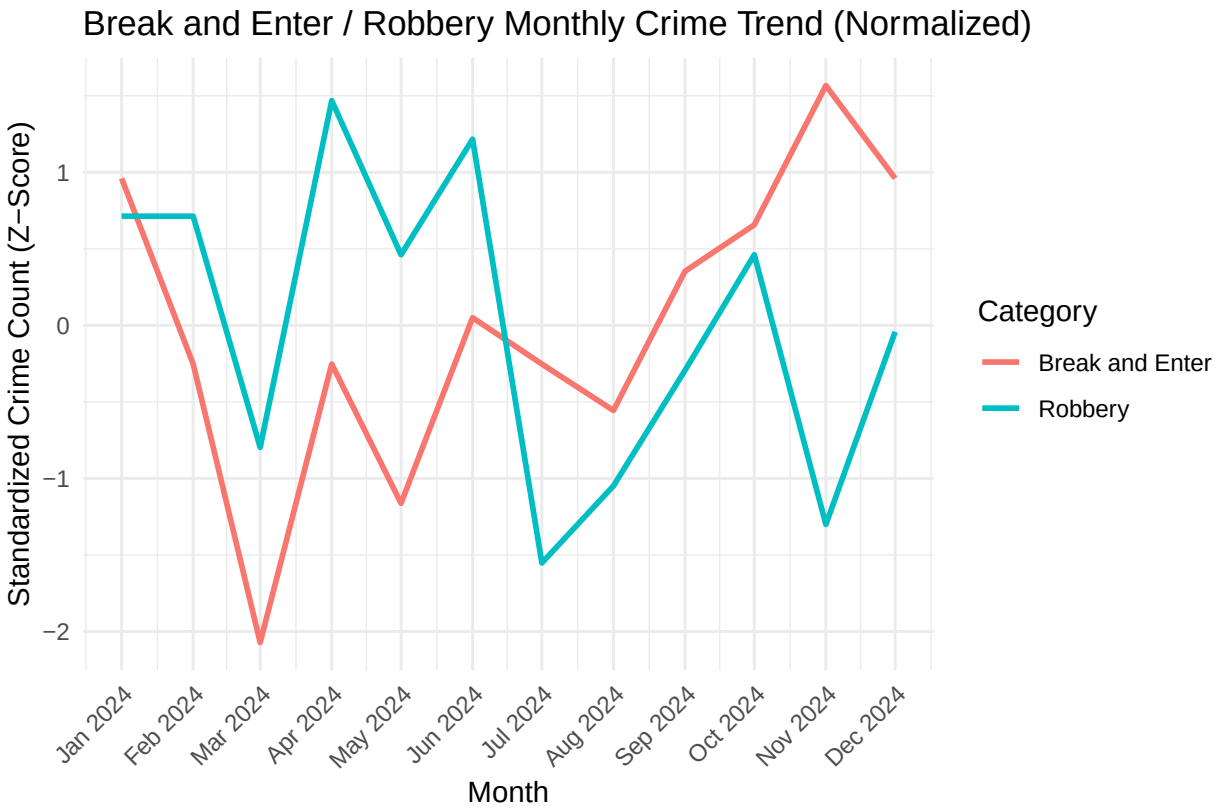


Fig 4: B&E and Robbery Trend

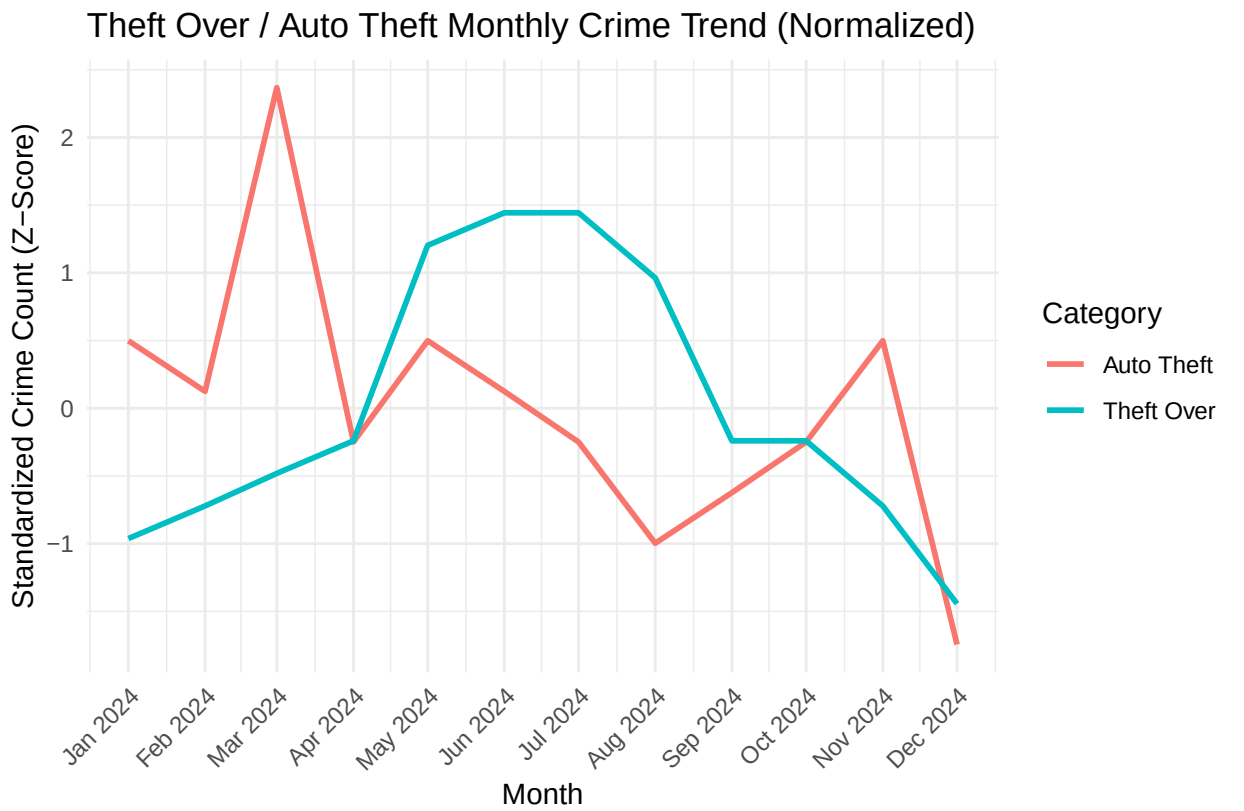


Fig 5: Auto Thefts and Theft

After splitting the categories into 3 distinct sections, based on the number of crimes in each category, their standardized counts, which represent the general trends over the months, were plotted. The "Assault" and "Theft Over" categories show a distinct rise in the crimes during the warmer months, while "Auto Theft" crimes actually dip in the warmer months and peak during the Spring. The other categories show no clear trends and have very volatile counts, likely due to a lack of data (working around the 2000 row limit set by the API).

Methods (Cont')

Choosing Models

It is important to note that predicting levels of crime is an extremely difficult task which is affected by a ton of factors, which range from social, economic, demographic, geographic, and more. When choosing and evaluating these statistical models, I do not expect them to predict crime rates with high accuracy. However, the models can tell us which variables share a statistically significant relationship with crime rates; for example, these models may show us that temperature is a better indicator of crime than precipitation, and vice versa. While something like a Generative Additive Model is good for exploring seasonality and non-linear trends, I prefer to use models with more interpretable results; as the goal of these models is not accuracy or predicting crime, but rather the relationship between weather factors and crime.

This analysis will use a Random Forest model and an XGBoost model. A Random Forest is an ensemble of decision trees built on bootstrapped samples of the data; it reduces variance and increases stability while naturally capturing non-linear effects and interactions. XGBoost is a gradient boosting method that builds trees sequentially, where each new tree aims to correct the errors of the previous ones. It is highly flexible and often achieves strong performance on structured data. Both RF and XGBoost provide variable importance

metrics that rank which predictors contribute most to reducing error, and when combined with partial dependence plots, they can help visualize the marginal effect of each weather variable on crime. These methods thus serve not only as predictive models but also as interpretable frameworks for identifying which environmental conditions may be linked to variations in criminal activity.

Fitting Models

To explore the relationship between weather variables and crime using Random Forest (RF) and XGBoost (XGB), I treat these models as tools for identifying important predictors rather than for precise forecasting. After splitting the data into a training set and a testing set, I fit each model using features such as temperature, humidity, precipitation, and visibility, along with seasonal variables like sine and cosine transformations of time. I examine the variable importance metrics they produce. For Random Forest, I use measures like %IncMSE (increase in mean squared error when a variable is permuted), and for XGBoost, I look at metrics such as gain, cover, and frequency to evaluate which variables most contribute to reducing prediction error across trees.

Although these models are typically designed for high-performance prediction tasks, they are valuable here because they can automatically capture complex, nonlinear relationships and interactions between features. To gain a deeper understanding of how each variable affects predicted crime levels, I generate partial dependence plots (PDPs), which visualize the marginal effect of each predictor on the outcome while averaging over other variables. This helps reveal whether, for instance, crime increases steadily with temperature or shows more nuanced patterns. Even in the presence of low overall prediction accuracy—expected due to the many social drivers of crime not present in the dataset—RF and XGB offer robust ways to explore and visualize the potential associations between climate conditions and criminal activity.

Results

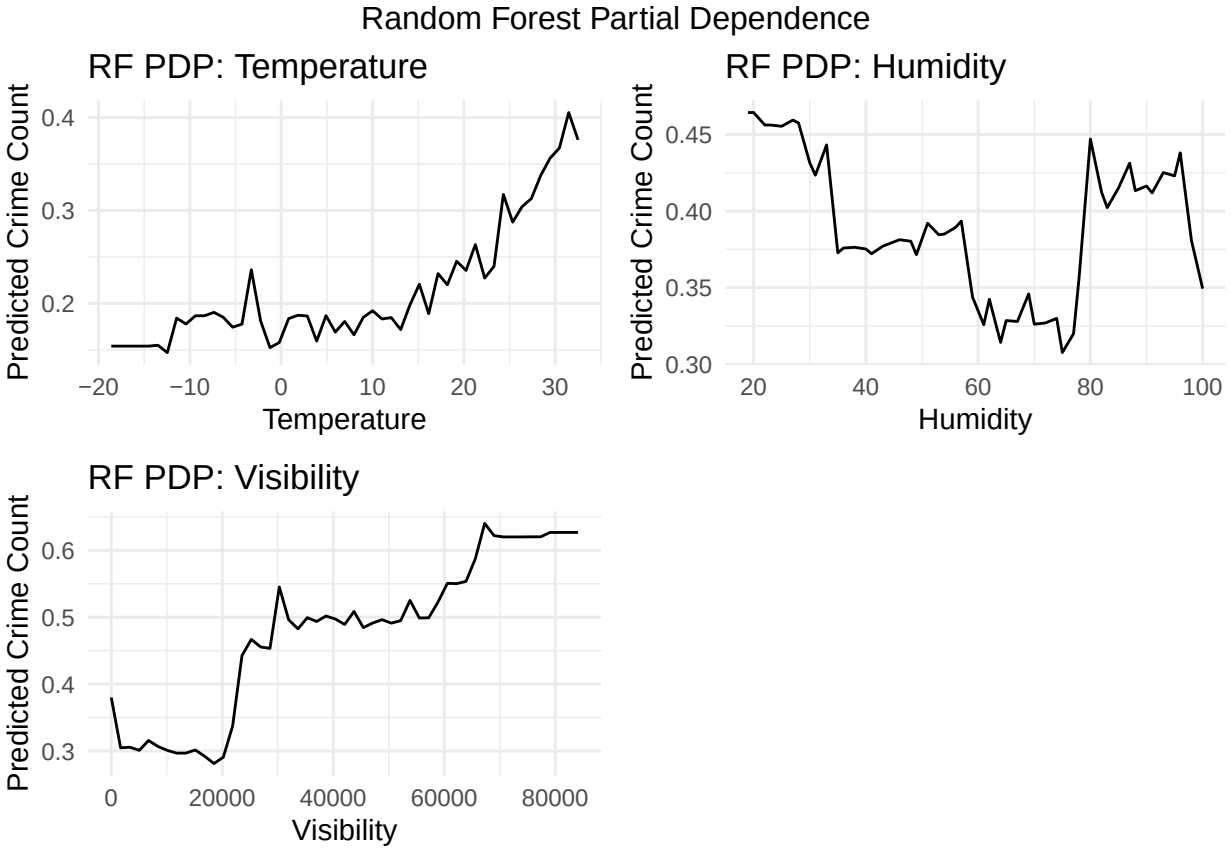
(See website for plots)

From the variable importance charts of both Random Forest and XGBoost models, several weather features seem to be essential in their ability to predict hourly crime counts. In the Random Forest model, the most important predictors are humidity, visibility, and temperature, as evidenced by the highest increases in mean squared error when the variables are permuted. This suggests that the model relies heavily on these features to provide predictions. Wind speed and precipitation are moderately predictive, while snowfall has very little or no predictive ability. The most influential variables in gain in the XGBoost model are wind speed, temperature, and visibility. Although lower-ranked in XGBoost than in Random Forest, humidity remains a contributing factor in the model. Interestingly, dew point is an influential variable in XGBoost but was absent in the Random Forest model and could be indicative of model-specific sensitivities. Both models agree mainly on temperature, visibility, and wind speed as being important predictors, while snowfall and precipitation are less so in terms of ranking in both models.

Partial dependence plots (PDPs) provide more insight into how each predictor influences predicted crime count when all other variables remain constant. Temperature for Random Forest and XGBoost both display a straight positive correlation with predicted crime: the higher the temperature, the higher the rate of predicted crime. This is as predicted by criminological theory that suggests warmer conditions facilitate higher outdoor activity and more opportunities for certain types of crime. Visibility also has a high positive influence, which would suggest clearer weather conditions are associated with higher levels of crime—perhaps because more people have opportunity to be out when it is clear.

Other variables exhibit less distinctive or less consistent patterns. In the Random Forest PDP for humidity, crime predictions are lower between around 55–70% humidity, with increases outside of that range, suggesting that very high or very low humidity might be linked to more crime. XGBoost's PDP for wind speed suggests that crime increases with wind speed but levels off after some point, and might decrease a bit, perhaps showing that light winds do not deter people from being outside, but increased winds might. Rain and snow

have little or no clear correlation to crime in either model, as one would expect from their low ratings of relevance.



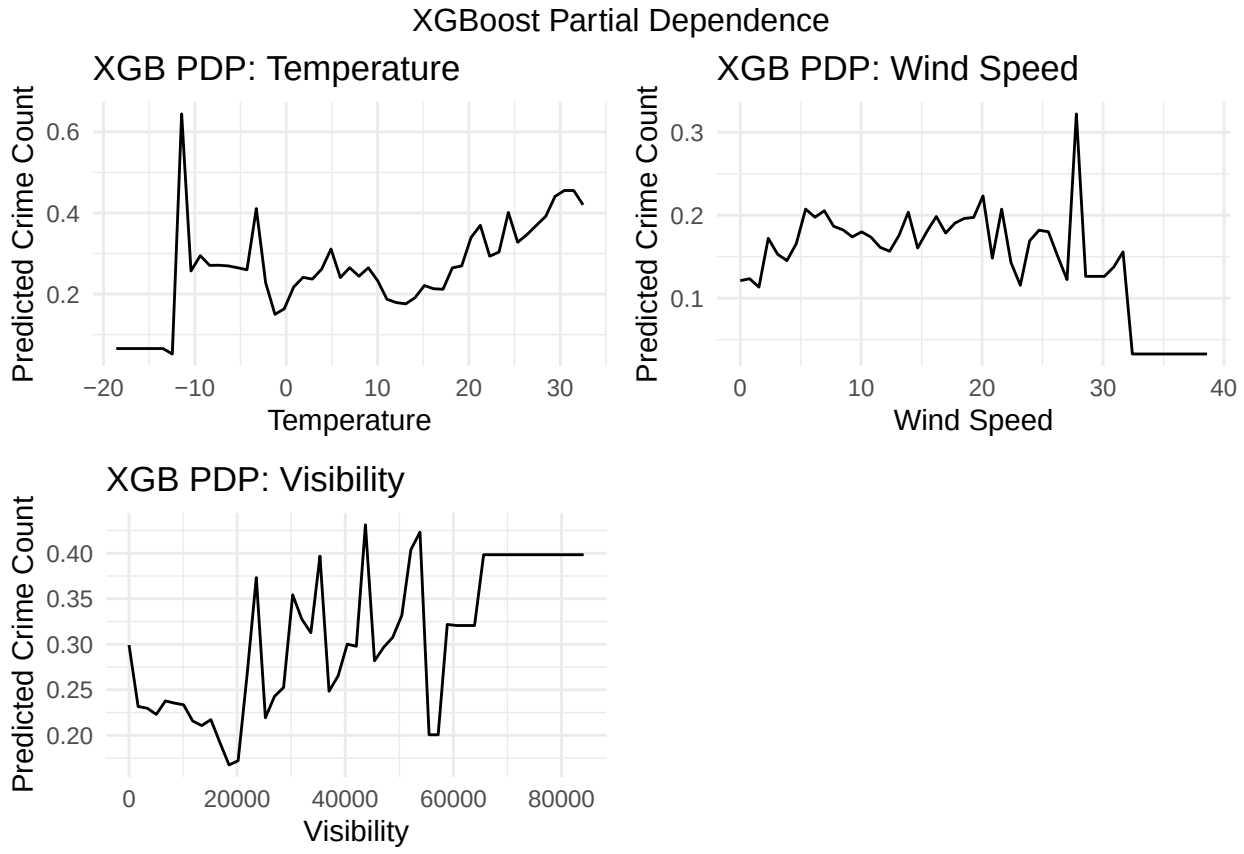


Fig 8: PDP's for Random Forest and XGB. Somewhat volatile trends; trends are more visible and concrete in the random forest than the XGB.

Combined, these results suggest that temperature and visibility are the most enduring weather variables in this sample with a uniform impact on crime. The findings support the notion that certain environmental conditions can facilitate crime by increasing activity and interaction in public spaces. While RMSE for both models were both ~0.53, meaning our predictions were off by ~0.53 crimes per hour on average, prediction accuracy was not an extremely important area of concern in this analysis, these machine learning models offer easy-to-understand resources for uncovering non-linear relations and identifying which aspects of weather are most correlated with crime fluctuations.

Summary

This analysis used Random Forest and XGBoost models to explore potential associations between weather conditions and hourly crime counts in Toronto. While all three models highlighted certain weather features—particularly temperature, visibility, and wind speed—as having measurable influence on predicted crime rates, the overall predictive accuracy remained low. This aligns with expectations, given that crime is a complex, highly multivariate phenomenon influenced by a wide array of social, economic, spatial, and behavioral factors that are not captured in weather data alone.

Although some patterns emerged—such as higher crime rates in warmer and clearer conditions—the modest model performance and inconsistent importance rankings across methods reinforce the idea that weather plays a secondary or enabling role in shaping crime patterns. It may influence when and where people go outside, potentially affecting opportunities for certain types of crime, but it is unlikely to be a primary causal factor. These results highlight the limitations of modeling crime in isolation from broader social dynamics, and underscore the need to integrate weather with other explanatory variables—such as time of day, location

characteristics, population density, socioeconomic indicators, or policing patterns—for a more comprehensive understanding of crime behavior.