Understanding the Drivers of Fire Frequency and Severity in Toronto: A Data-Driven Approach

Shijun Yu

April 30, 2025

1 Introduction

1.1 Background

Fire incidents pose significant risks to public safety and can result in substantial economic losses. Understanding when and where fires happen, and what might be contributing to their frequency and severity, is essential for prevention, preparedness, and resource planning. Motivated by this broader goal, I decided to explore fire incidents in the city of Toronto using publicly available data from the City of Toronto's open data portal. This dataset contains key details about each fire, including the alarm time, geographic location, and estimated dollar loss.

My initial curiosity centered around the role of weather conditions. It seemed intuitive that certain environmental factors, such as hot, dry, and windy, might be linked to a greater likelihood or severity of fires. To test this idea, I retrieved historical weather data from the Open-Meteo API. This source provides hourly data on temperature, precipitation, and wind speed, which I matched to the time and location of each fire incident.

However, after conducting some preliminary analysis, I found that the relationship between weather conditions and fire severity wasn't as strong as I had expected. While weather may play a role, it became evident that other factors, including the time of day, season, and neighborhood, might be more informative. This observation led me to expand my original focus to consider a wider range of variables.

1.2 Research Question

To better understand the dynamics of fire incidents in Toronto, I refined my research question to:

"What factors affect the frequency and severity of fires in Toronto?"

To answer this question, I focus on fire incidents with valid location and timestamp data, and I use estimated dollar loss as a proxy for fire severity. Recognizing that variables like dollar loss can be highly skewed, appropriate transformations were applied during the analysis. This project aims to investigate how spatial, temporal, and environmental variables relate to fire occurrence and impact across the city.

2 Methods

2.1 Data Acquisition

Two primary datasets were used in this analysis:

1. Fire incident dataset was downloaded from the City of Toronto's Open Data Portal (https://open.toronto.ca/dataset/fire-incidents/). This dataset contains records for fire incidents within Toronto, providing comprehensive details such as the exact date and time of incident, geographical coordinates, and the estimated dollar loss incurred from each incident.

2. Weather dataset was obtained using the Open-Meteo Historical Weather API (https://open-meteo.com/en/docs/historical-weather-api). The Open-Meteo API provides detailed meteorological information including temperature (°C), wind speed (m/s), and precipitation (mm) at an hourly granularity. For each incident, the exact hourly weather conditions matching the incident's alarm hour were fetched using latitude, longitude, and incident date as query parameters.

These two datasets were initially stored as separate tibbles and later merged into a single comprehensive dataset. This final merged dataset contains data from January 2011 to June 2016.

2.2 Data Cleaning and Wrangling

Following data acquisition, the merged dataset was first inspected to verify variable structure and data types. Several data type conversions and renaming procedures were conducted for consistency and clarity. Missing values were checked during the inspection, with only one observation containing NA. This incomplete record was removed from the dataset.

After that, several new variables were created to support the analysis of fire frequency and economic impact. A binary DayNight variable was created by splitting incidents into daytime (6AM to 6PM) and nighttime (6PM to 6AM). A Season variable was also created by mapping each incident's month to its corresponding season, along with a labeled Month column for monthly trend analysis.

Exploratory checks of variable distributions were conducted next, using histograms to assess potential skewness. Temperature was approximately normally distributed and required no transformation. However, wind speed exhibited significant right-skewness, prompting a log transformation to achieve a more symmetric distribution suitable for subsequent analysis. Due to the highly skewed nature of precipitation data (predominantly composed of zero values), this numerical variable was converted into a binary categorical variable, indicating the presence or absence of rainfall at the time of the incident. In addition, estimated dollar loss exhibited significant right-skewness, driven by a small number of incidents with extremely high losses. To address this, a log transformation was applied to reduce skewness, improve interpretability, and maintain valuable information regarding incident severity.

Finally, outlier detection and removal were carried out for temperature and log of wind speed using the 1.5*IQR method. These extreme values were safely removed. Conversely, outliers in the log of estimated dollar loss were intentionally retained, recognizing that extreme values in this variable reflect genuinely severe fire incidents critical to the study's objectives.

After completing these data cleaning and wrangling procedures, the cleaned dataset comprised 9931 rows and 12 columns.

2.3 Data Exploration Tools

To investigate the frequency and economic impact of fire incidents in Toronto, a variety of visualizations were created using the cleaned dataset. These visual tools supported the exploration of temporal, spatial, and environmental patterns and provided insight into how various factors may be related to the occurrence and cost of fires.

Environmental Patterns:

To assess whether environmental factors influence how often fires occur, histograms were used to compare fire counts under different temperature and log of wind speed. Additionally, a bar chart was used to compare

fire frequency under rainy or non-rainy conditions. These plots allowed for a clear visual comparison of how fire frequency varies under different weather conditions.

Temporal Analysis:

Fire frequency was first examined across temporal dimensions. Bar charts were used to visualize the total number of incidents by season, month, and time of the day. These visualizations helped to reveal how fire frequency changes across time.

Spatial Analysis:

To understand spatial differences in fire occurrence, fire incidents were mapped onto Toronto neighborhoods using a choropleth map. The number of fire incidents was aggregated by neighborhood, and neighborhoods were shaded by total incident count. This visualization highlighted geographic areas with higher or lower fire activity.

Economic Loss Patterns:

To explore the economic impact of fire incidents under different conditions, boxplots were used to compare fire losses across several categorical variables, including time of day, season, month, and precipitation status. These boxplots offered insights into when and under what conditions fires tend to result in greater financial damage.

Numerical Variable Relationships:

To assess relationships among numerical variables, a correlation heatmap was generated to summarize the strength and direction of linear associations between variables such as temperature, log-transformed wind speed, and log-transformed estimated dollar loss. To further explore these relationships visually, scatterplots with fitted linear trend lines were used to visualize the actual form and variability of the associations between continuous predictors and economic loss.

This combination of graphical tools provided a comprehensive exploratory understanding of when, where, and under what conditions fire incidents tend to occur and result in higher economic damage. Insights from this step informed the selection of variables and modeling strategies used in later stages of analysis.

2.4 Modeling Approach and Evaluation

The full dataset was randomly split into a training set (70%) and a test set (30%). All transformations, outlier removals, and feature engineering were completed prior to the split to prevent data leakage.

To explore what factors influence how often fires occur, a count-based modeling approach was used. The number of fire incidents was aggregated monthly and modeled using a Generalized Additive Model (GAM) with a Poisson link. Model performance was evaluated using R^2 on the test set, which measures how well the model explains variation in monthly fire counts.

To model the log-transformed estimated dollar loss, five regression models were trained using the same set of predictors: temperature, log-transformed wind speed, precipitation condition, time of day, month and season. The following models were fit:

1. Pruned Tree (rpart)

- Optimal complexity parameter selected by minimizing cross-validation error

2. Bagging (random forest with mtry = number of predictors)

- Built using the randomForest package with all predictors considered at each split

3. Random Forest (randomForest library)

- Ensemble of decision trees with default mtry value

4. Boosting (gbm via caret)

- 1000 trees
- 10-fold cross-validation
- Best shrinkage value selected from: (0.001, 0.005, 0.01, 0.05, 0.1)
- Fixed interaction depth = 1
- Gaussian error distribution

5. XGBoost (caret)

- 10-fold cross-validation with parallel processing
- Grid search
- max depths: (1, 3, 5)
- number of iterations: (100, 150, . . . , 500)
- learning rates: (0.01, 0.05, 0.1)

Each model was trained on the training set and evaluated on the test set using Root Mean Squared Error (RMSE), which provides an interpretable measure of prediction accuracy in the same units as the response variable.

3 Results

3.1 Exploratory Data Analysis

First of all, summary statistics of numeric variables are presented in Table 1 below. These statistics reflect the cleaned and transformed dataset after outlier removal and transformation steps described in the Methods section.

Table 1: Summary Statistics of Numeric Variables

Statistic	Temperature	Wind_Speed_Log	Estimated_Dollar_Loss_Log
Min	-24.600	0.405	0.000
Q1	0.700	1.233	5.525
Median	8.900	1.522	7.824
Mean	8.748	1.507	7.029
Q3	17.900	1.795	9.210
Max	34.800	2.612	16.380

Table 1 reveals that fire incidents occurred across a wide range of temperatures, from -24.6°C to 34.8°C, with a median of 8.9°C. Wind speed, after log transformation, ranged from approximately 0.4m/s to 2.6m/s, with a median of 1.5m/s. The log-transformed estimated dollar loss varied considerably, with a minimum value of 0 and a maximum value of 16.38, indicating substantial variability in fire-related financial damages.

Next, the following Figure 1 shows the frequency of fire incidents by temperature, log-transformed wind speed, and precipitation condition.

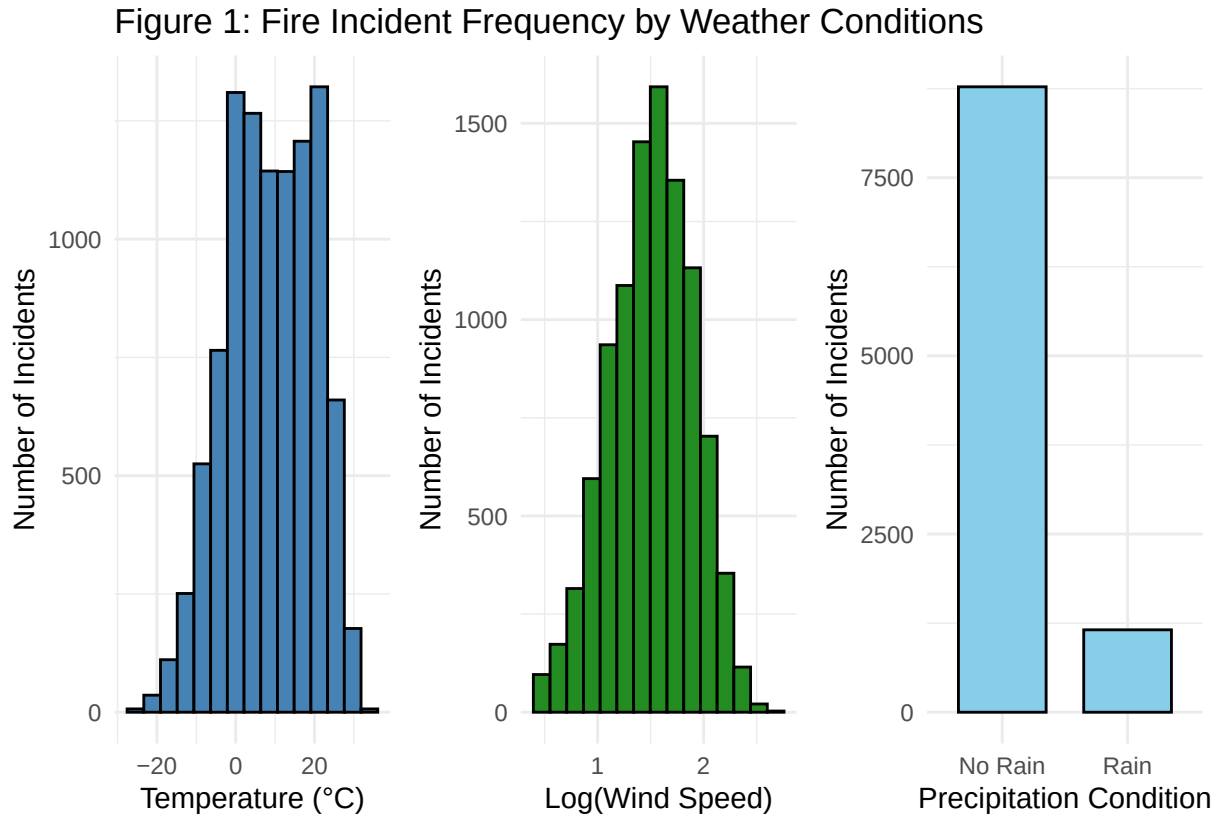
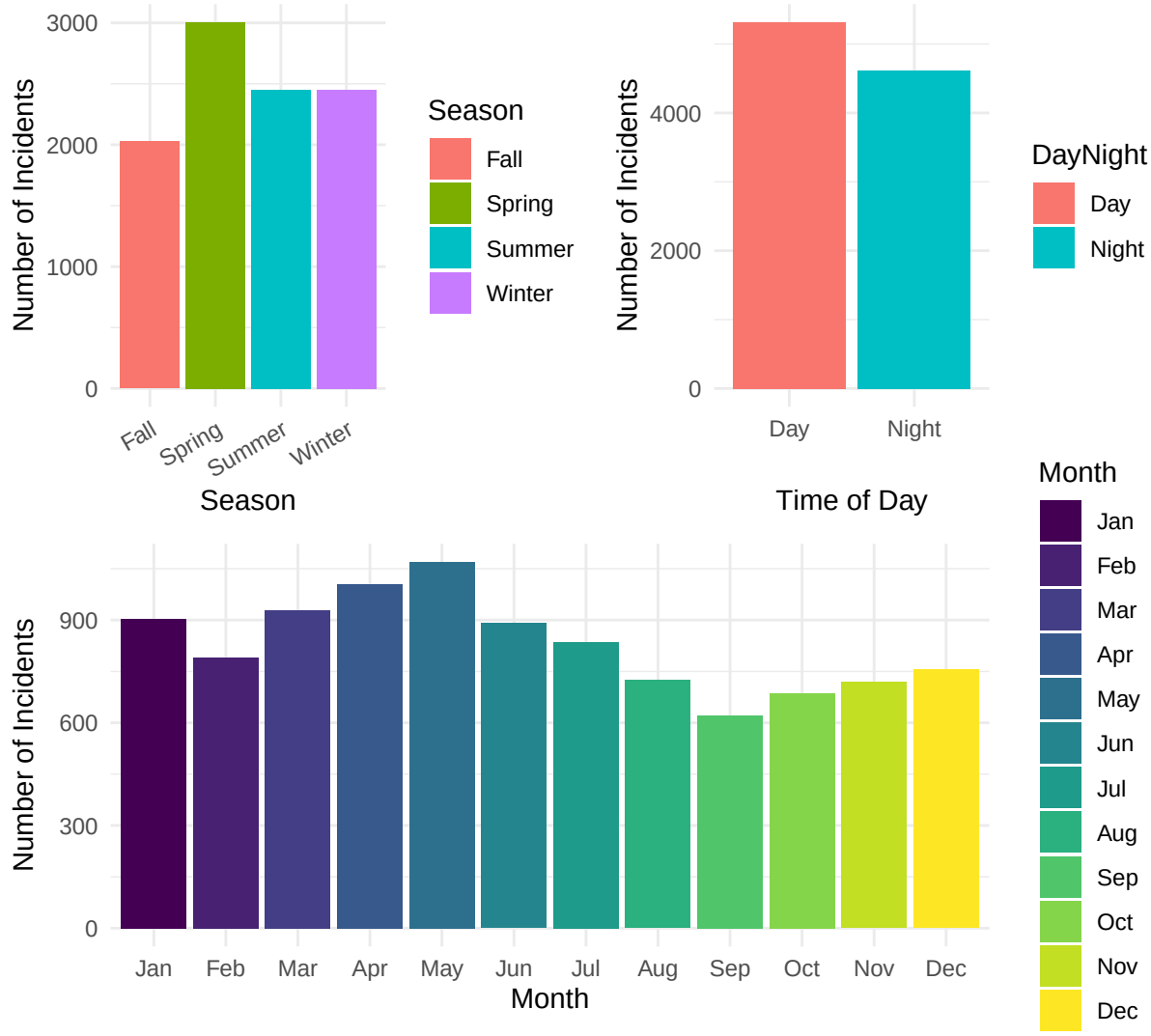


Figure 1 indicates that fire incidents occurred across a wide temperature range, with a higher concentration during moderate temperatures. Incidents also clustered around moderate wind speeds when examined on a log-transformed scale, suggesting a relatively normal distribution. Additionally, the majority of fire incidents took place under dry conditions, illustrating that the absence of precipitation may play a role in the increased likelihood of fire occurrences.

Then, in order to explore the temporal characteristics of fire incidents in Toronto, the following Figure 2 was generated to summarize incident frequency across different time dimensions.

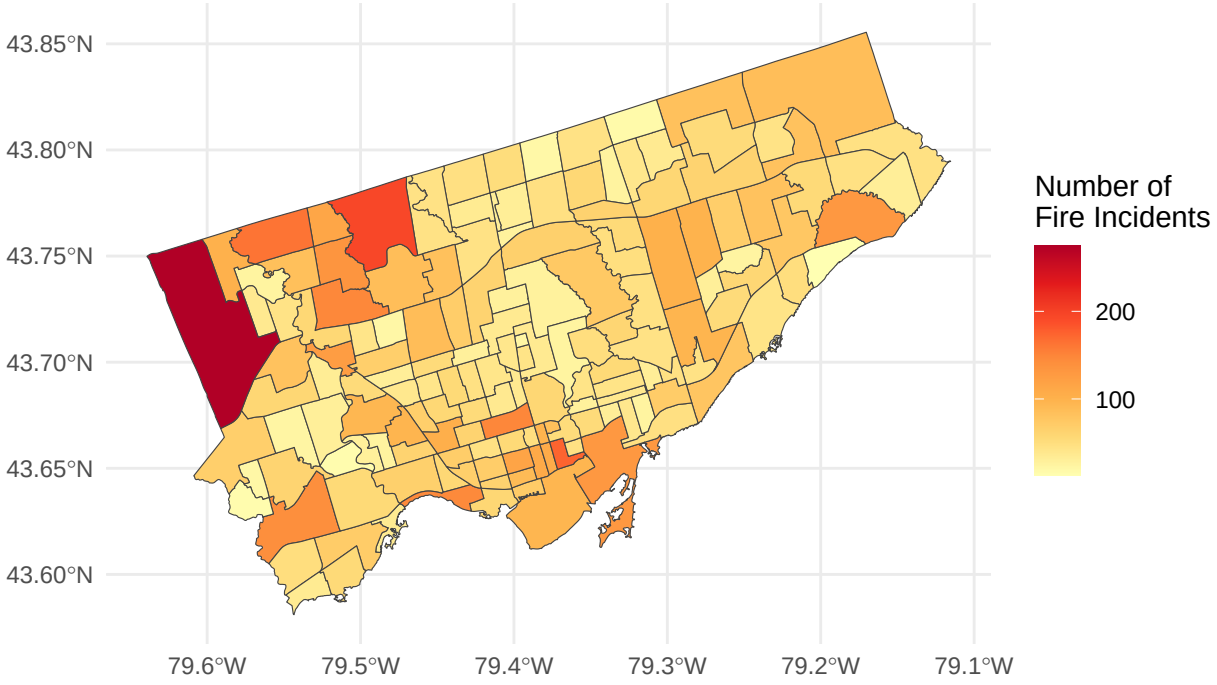
Figure 2: Fire Incident Frequency by Season, Month, and Time of Day



As shown in Figure 2, spring experienced the highest number of fire incidents among all seasons, followed by summer and winter, while fall had the fewest. On a monthly scale, May had the greatest number of incidents, while September recorded the lowest. In terms of time of day, daytime incidents were more frequent than nighttime ones. These patterns suggest that temporal variation may play a role in influencing fire frequency, with warmer and more active months such as May potentially leading to increased fire risk.

To understand where fire incidents occur most frequently, each incident was geocoded to a Toronto neighborhood, and the total count of incidents was mapped onto a choropleth. Figure 3 below highlights geographic hot spots, offering insight into how certain neighborhoods may experience more fires (an interactive version of this plot can be found on the website).

Figure 3: Choropleth Map of Fire Incident Frequency by Neighborhoods



As shown in Figure 3, several neighborhoods in the northwestern region of Toronto appear to experience notably higher fire frequencies. These spatial disparities suggest that neighborhood-level differences, such as population, building age, or building density, could play an important role in shaping where fires are more likely to occur.

After that, in order to assess patterns in fire severity, log-transformed estimated dollar loss was compared across time of day, season, month, and precipitation condition using boxplots. Figure 4 below highlights when fires tend to result in greater economic damage.

Figure 4: Log−Transformed Estimated Dollar Loss by Categorical Variables

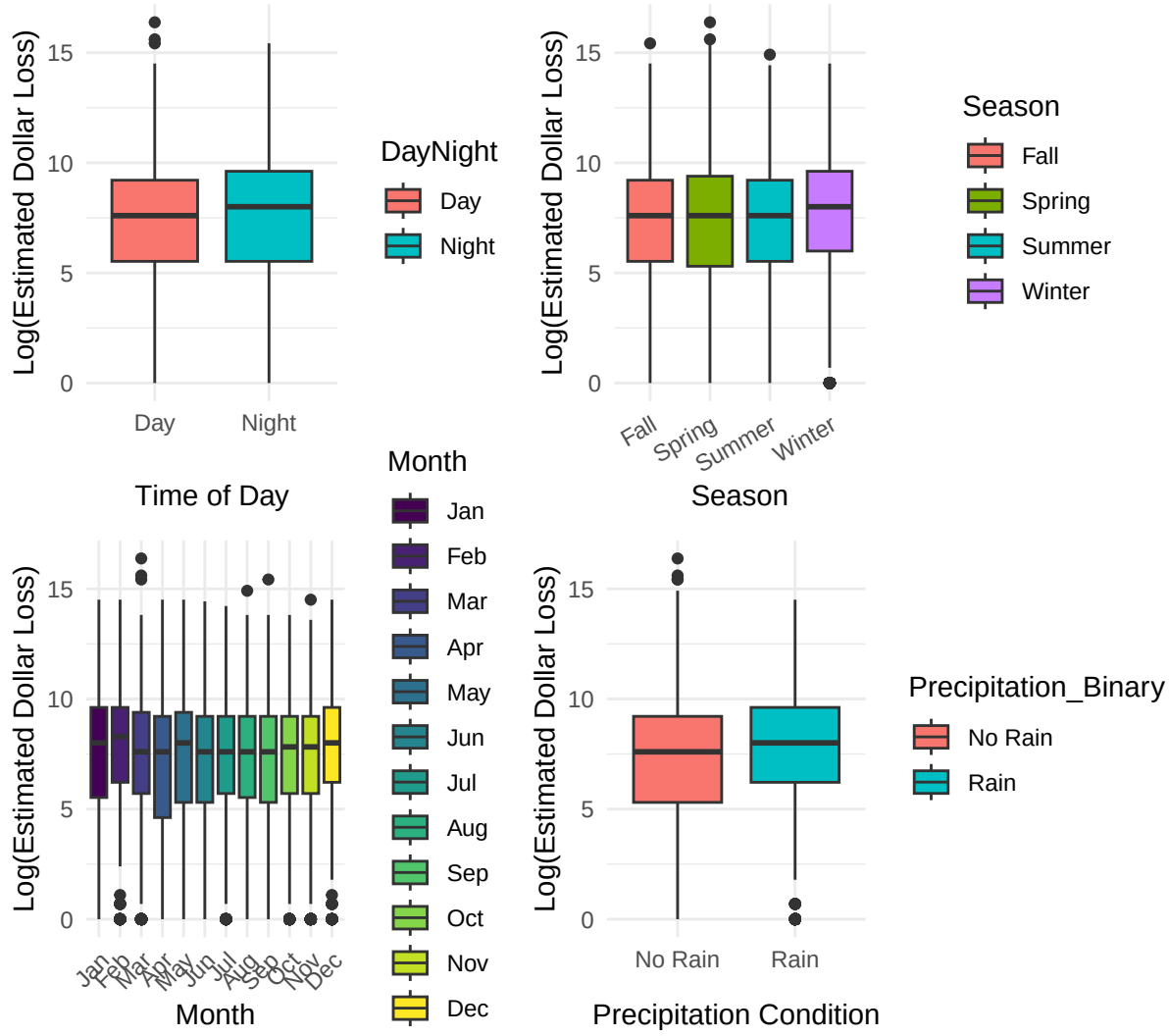
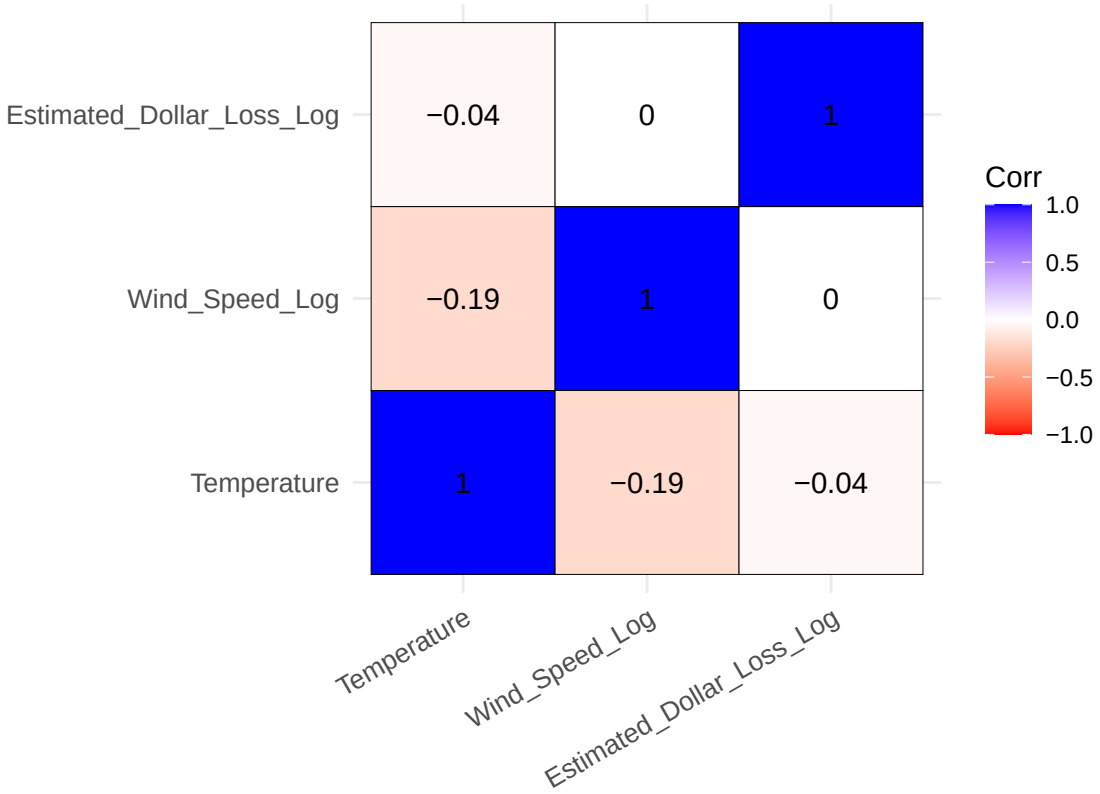


Figure 4 suggests that fire incidents occurring at night generally have slightly higher financial losses compared to those during the day. Seasonal and monthly variations appear modest, though winter shows slightly higher medians. Additionally, incidents that occurred during rainfall tend to be associated with higher losses, indicating a potential link between precipitation condition and fire severity.

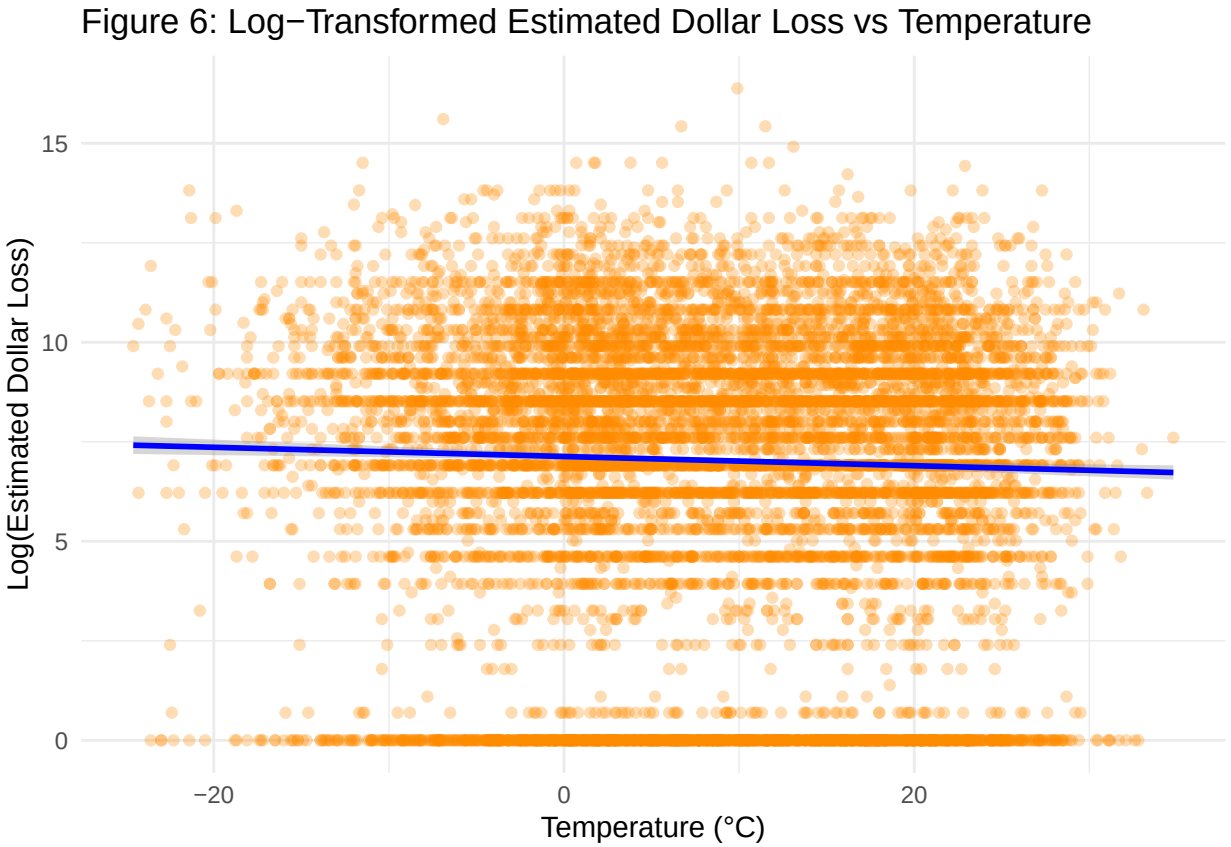
Finally, Figure 5 below provides a concise summary of correlations among numerical variables.

Figure 5: Correlation Heatmap of Numerical Variables



As shown in Figure 5, the weak correlations between these numerical variables indicate potential non-linear relationships or interactions that may be explored further.

Thus, the following scatterplots are shown to provide visual insights. Figure 6 illustrates a scatterplot of temperature versus log-transformed estimated dollar loss, showing a weak negative linear relationship.

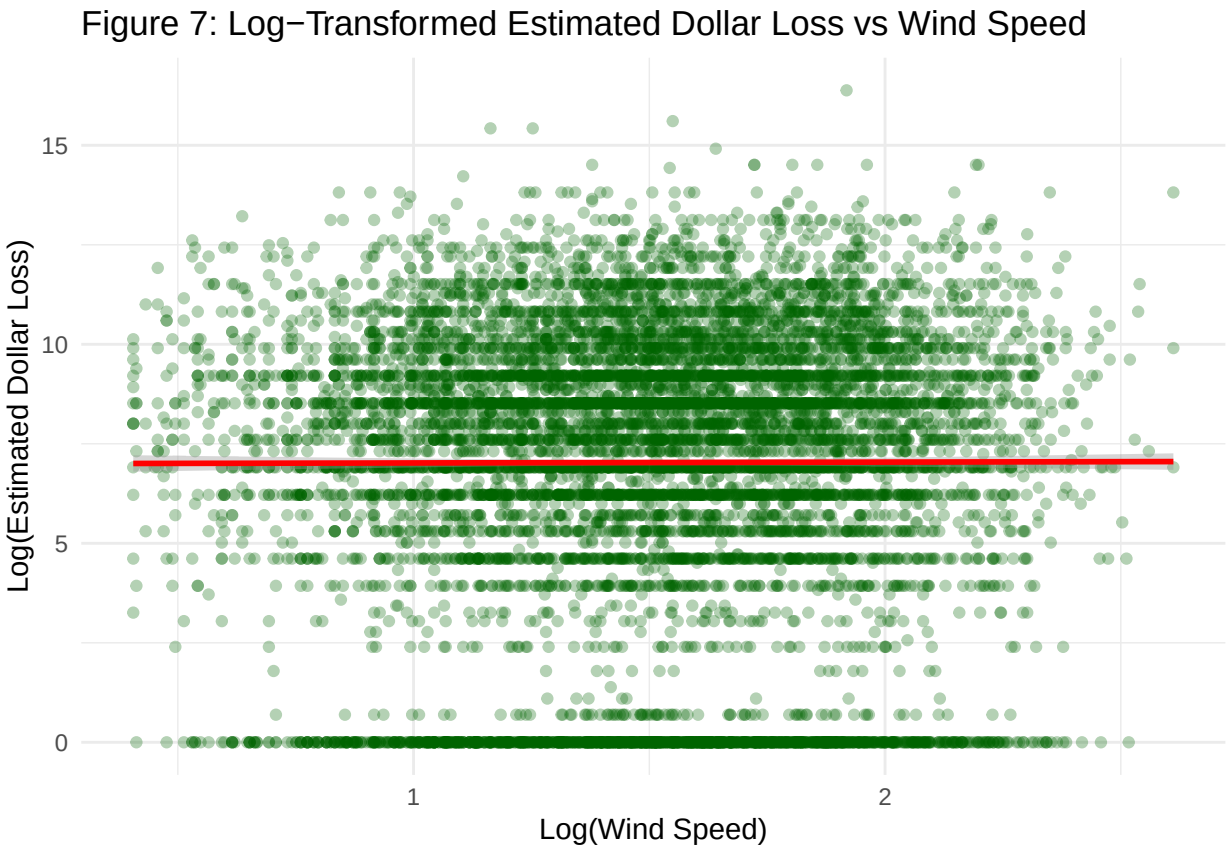


The summary of this linear model is shown in Table 2. It confirms the previous indication of a weak but statistically significant negative relationship between temperature and log-transformed estimated dollar loss. Specifically, the estimated coefficient for Temperature is approximately -0.012 (p-value < 0.05), suggesting that for each 1°C increase in temperature, the expected log-transformed dollar loss decreases by about 0.012 units. Interpreted on the natural scale, this translates to roughly a 1% decrease in estimated dollar loss per degree Celsius increase.

Table 2: Summary Table of First Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.130	0.045	157.592	0
Temperature	-0.012	0.003	-3.582	0

Similarly, Figure 7 below depicts the relationship between log-transformed wind speed and log-transformed estimated dollar loss, suggesting no strong linear relationship.



The summary of this linear model is shown in Table 3. In contrast to the previous model, Wind_Speed_Log exhibits a small positive coefficient (approximately 0.021) with a high p-value (0.813), indicating no statistically significant linear relationship between the log-transformed wind speed and the log-transformed estimated dollar loss.

Table 3: Summary Table of Second Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.997	0.14	49.809	0.000
Wind_Speed_Log	0.021	0.09	0.236	0.813

3.2 Modelling Analysis

To begin with, the GAM model was fit to the monthly-aggregated fire counts in order to quantify the effects suggested by the exploratory analyses. A Poisson-link GAM was used to allow for smooth, nonlinear effects of average temperature, average log-transformed wind speed, and calendar month (to capture seasonality), while also including categorical terms for season and linear terms for the fractions of rainy and daytime hours. Tables 4–6 below first report the estimated parametric coefficients, then the significance of the smooth terms, and finally the model's R^2 on the test set.

Table 4: Parametric Coefficients for the GAM

Term	Estimate	Std_Error	z_value	p_value
(Intercept)	5.119	0.285	17.955	0.000

Term	Estimate	Std_Error	z_value	p_value
SeasonSpring	0.445	0.201	2.218	0.027
SeasonSummer	0.251	0.169	1.488	0.137
SeasonWinter	0.117	0.164	0.714	0.475
RainyFrac	-0.247	0.351	-0.704	0.481
DayFrac	-0.568	0.475	-1.196	0.232

Table 5: Smooth Terms for the GAM

Smooth	EDF	Ref_df	Chi_sq	p_value
s(AvgTemp)	8.713	8.943	29.226	0.001
s(AvgWind)	1.000	1.000	2.445	0.118
s(as.numeric(Month))	4.662	7.000	20.994	0.000

Table 6: Test Set R^2 for the GAM

Model	R_squared
GAM	0.32

Table 4 shows that, among the parametric terms, only the spring indicator reaches statistical significance (p-value = 0.027), while summer and winter do not differ significantly from the fall baseline; neither the fraction of rainy hours nor the fraction of daytime hours are significant predictors of monthly fire counts. In Table 5, both the temperature spline and the cyclic month spline are highly significant, indicating strong nonlinear effects of average temperature and seasonality on monthly fire frequency, whereas average log-transformed wind speed does not contribute significantly (p-value = 0.118). Finally, $R^2 = 0.32$ (Table 6) indicates that the model explains about one-third of the variability in monthly fire counts. These results suggest that temperature and seasonal timing are the primary drivers of fire frequency, while wind speed and precipitation pattern play a much smaller role.

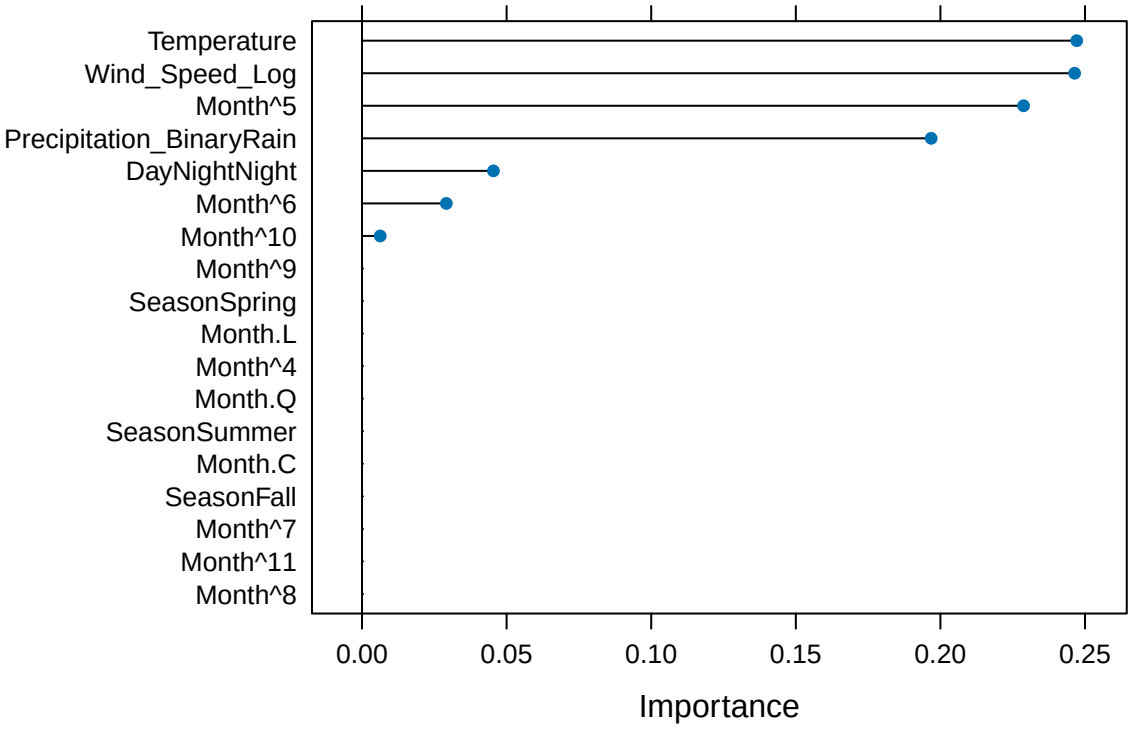
After examining the factors influencing the frequency of fire incidents using a GAM, the next stage of analysis focuses on understanding the severity of fire incidents, as measured by the log-transformed estimated dollar loss. To evaluate how well different modeling approaches can predict fire-related economic loss, five machine learning regression models are trained and compared using the same set of predictors (details can be found in section 2.4). Model performance is assessed using RMSE on test set to ensure fair comparison.

Table 7: Test Set RMSE for All Regression Models

Model	RMSE
Pruned Tree	3.587
Bagging	3.760
Random Forest	3.611
Boosting	3.604
XGBoost	3.583

Among the five regression models evaluated, XGBoost achieved the lowest test RMSE, indicating the best predictive performance for estimating fire-related economic losses. It slightly outperformed boosting model and pruned tree model, while bagging model showed the highest error. Given XGBoost's superior accuracy, its variable importance plot is presented in Figure 8 below to highlight the most influential predictors contributing to the model's performance.

Figure 8: Variable Importance Plot (XGBoost)



The variable importance plot reveals that temperature and log-transformed wind speed are the two most influential predictors of fire-related economic losses. These are followed by precipitation status and certain month indicator (Month^5). In contrast, some other month indicators don't contribute, as well as the season. Overall, weather conditions appear to play key roles in determining the economic impact of fire incidents in Toronto.

4 Conclusions and Summary

4.1 Findings

This study investigated the factors that influence the frequency and economic severity of fire incidents in the City of Toronto, using a combination of fire incident data from the City's open data portal and historical weather data from the Open-Meteo API. The analysis explored temporal, spatial, and environmental variables, and employed both statistical and machine learning models to assess their predictive power.

For fire frequency, the results from GAM suggest that temperature and calendar month have strong nonlinear effects on the number of fires per month, while other factors like precipitation, wind speed, and time of day appear less influential. Seasonal effects were somewhat limited, with only spring showing a statistically significant difference relative to fall. These results highlight the importance of seasonality and weather conditions in predicting when fires are more likely to occur, with the model explaining approximately 32% of the variance in monthly fire counts.

For fire severity, measured as log-transformed estimated dollar loss, a series of machine learning models were trained and compared. Among the five models tested, XGBoost model achieved the best predictive accuracy with the lowest test RMSE. Variable importance from XGBoost model revealed that temperature, log-transformed wind speed, and precipitation condition were the most influential predictors of fire-related

financial loss. This reinforces the notion that environmental conditions play a key role not only in the occurrence but also the costliness of fire incidents.

4.2 Limitations

While the models provided valuable insights, several limitations should be noted. First, estimated dollar loss is a proxy for fire severity and may be subject to reporting inconsistencies. Second, the weather data was joined based on hourly alignment, which may not perfectly reflect the local conditions at the time of each fire. Third, some potentially relevant variables such as building type, fire cause, or population density were not included due to data availability. Lastly, although the models demonstrated reasonable performance, there remains substantial unexplained variability, suggesting the need for richer datasets and additional features.

4.3 Summary

In summary, this analysis shows that weather and time-based variables offer meaningful predictive power in understanding both the occurrence and severity of fire incidents. The findings could support more targeted public safety interventions, such as allocating fire prevention resources more effectively across seasons, or raising awareness during high-risk periods. Future research could extend this work by integrating more detailed spatial and structural data or by incorporating real-time weather feeds for predictive deployment.