# Investigating Fire Incidents Severity Levels in Toronto: Enhancing Resource Allocation for Fire Response*

**Identifying Key Factors for Fire Severity and Evaluating the Effectiveness of Public Fire Response Systems**

Maggie Zhang

December 3, 2024

This study investigates fire severity levels in Toronto to improve resource allocation for Toronto Fire Services (TFS). The analysis identifies that higher-severity incidents are strongly correlated with incidents originating from mechanical, HVAC, and electrical areas, while cooking and lighting areas tend to lead to lower severity fires. The study also finds that fire incidents involving certain ignition sources, such as electrical units or lightning equipment, are more likely to be severe. The study also evaluates the effectiveness of public fire response systems, highlighting that fire alarms are successful in low and medium-severity fires, while sprinkler systems are often only partially operational. Results can guide TFS to enhance response times, prioritize resources, and optimize fire response operations, ultimately improving community safety in Toronto.

## Table of contents



*Code and data are available at: https://github.com/MaggieZ111119/TFS_Toronto_Fire_Incident.

# 1 Introduction

The study aims to estimate the severity of fires in the Toronto region. A fire incident, as defined by the Ontario Fire Marshal (OFM), refers to any event involving fire, explosion, or other situations suspected to lead to significant outcomes, such as injury or death (Government of Ontario n.d.).

The severity of a fire can be assessed through various factors, including the intensity of the flames, the size of the affected area, the duration of the fire, its impact on human activities, and its potential threat to animal habitats. There are numerous standards and considerations when discussing fire severity, but when it comes to prediction, it is unrealistic to consider all factors. Therefore, defining severity becomes the first task before any further analysis. In this paper, the severity of a fire is defined based on three aspects: casualty and financial loss, the

type of incident, and the number of apparatuses dispatched by Toronto Fire Services (TFS) in response to the fire. Further explanation of how severity is defined will be provided later in the paper.

This research also identifies specific high-priority response criteria that can enhance TFS operations. By examining the data, it has been found that the area of origin of a fire is a significant predictor of fire severity. Fires originating from mechanical, HVAC, and electrical areas are highly correlated with severe incidents, whereas fires originating from cooking and heating areas are more commonly associated with low-severity incidents.

Additionally, beyond investigating potential predictors of severe fire incidents, this study also explores the effectiveness of current public fire response mechanisms, such as fire alarms and sprinkler systems, in the Toronto region. It was discovered that sprinkler systems, across all levels of fire severity, were largely only partially operational. On the other hand, fire alarm systems were successfully activated in low and medium-severity fire incidents.

The Office of the Fire Marshal (OFM) plays a crucial role in overseeing fire safety regulations in Ontario, ensuring fire protection, prevention, and public safety across the province. They are responsible for setting safety standards and improving public safety for all municipalities. Under the leadership of the OFM, municipal fire departments, such as Toronto Fire Services (TFS), allocate resources and develop plans to protect shared communities (Government of Ontario n.d.). Therefore, by analyzing data within their standards, identifying factors that contribute to high-risk fire incidents, and evaluating current operational mechanisms, TFS can gain valuable insights to develop strategic response protocols that prioritize these incidents. This knowledge will help both the public and fire departments enhance response times, allocate resources more effectively, and reduce the occurrence of lower-severity incidents.

The remainder of this paper is structured as follows. Section 2 describes the data used in the analysis, including the sources and key variables and their visualizations.Section 3 outlines the models applied to estimate fire severity and identify key factors. Section 4 presents the results of the analysis. Section 5 provides an interpretation of the results in the context of current fire safety practices and suggests potential improvements. Finally, **?@sec-appendix** includes supplementary information and additional data details.

In this paper, I used several R packages for data analysis and visualization. The `tidyverse`(Wickham et al. 2023b), `dplyr`(Wickham et al. 2023a), and `tidyr`(Wickham, Bryan, et al. 2023) were used for data manipulation. `ggplot2`(Wickham 2016) facilitated data visualization, while `lubridate`(Grolemund and Wickham 2011) was used for date-time handling. `knitr`(Xie 2023) helped integrate code into reports. For modeling, I used `rstanarm`(Gelman et al. 2020) and `modelsummary`(Fox, Kropf, and Byrne 2021) for summaries, `caret`(Kuhn 2023) for model training, and `xgboost`(Chen and Guestrin 2016) for boosting. Additional tools like `rpart`(Therneau and Atkinson 2015), `fitdistrplus`(Delignette-Muller and Dutang 2023), and `pROC`(Robin, Turck, and H. 2023) supported partitioning, distribution fitting, and model evaluation.

# 2 Data

## 2.1 Overview

We use the statistical programming language R (R Core Team 2023)to investigate data about fire incidents(Fire Services 2024), downloaded from the Open Data Toronto Portal: https://open.toronto.ca/dataset/fire-incidents/ . The dataset was last updated on December 31, 2023. It includes fire incidents following the Ontario Fire Marshal's (OFM) definition, with descriptions according to OFM standards. The dataset also provides detailed information on Toronto Fire Services' (TFS) responses to each fire incident. The data is mostly complete, though some incidents are still under investigation and categorized as "no loss outdoor fire." The dataset includes information on the origin of fires, impacts on civilians and buildings, as well as geographical data (see Table 1). Additionally, it contains details on TFS response times and actions taken. Overall, the dataset is comprehensive, informative, and provides valuable insights for my later study.

| _id | area_of_origin | building_status | business_impact | civili |
|---|---|---|---|---|
| 3758854 | 81 - Engine Area | NA | NA | |
| 3758855 | 75 - Trash, rubbish area (outside) | NA | NA | |
| 3758856 | NA | NA | NA | |
| 3758857 | 75 - Trash, rubbish area (outside) | 01 - Normal (no change) | 1 - No business interruption | |
| 3758858 | NA | NA | NA | |
| 3758859 | 81 - Engine Area | NA | NA | |

## 2.2 Measurement

The dataset is clearly structured, with the first column as '_id,' which is a unique identifier for each record in the database. There are 43 columns in total, containing information ranging from Incident Type to Infrastructure. These columns can be mainly divided into four categories based on my understanding:

1. **Incident Type and Cause**: This category contains information directly related to each fire incident, including descriptive details about the incident, its nature, and origin. It covers aspects such as where the fire started, the materials involved, and potential causes. These variables are crucial for predicting fire behavior, and include columns like 'area of origin,' 'level of origin,' and 'smoke spread.'

2. **Impact on Humans and the Economy**: These variables describe the direct outcomes of fire incidents, including financial and environmental impacts, as well as effects on human life. This category is key for defining the severity of fires. Variables include: firefighter casualties, civilian casualties, building status after the fire, and the number of displaced people.

3. **Response Evaluation**: This category records TFS's reactions to each reported fire incident. It provides valuable insight into the effectiveness of the current system and helps when predicting fire severity. Variables include: 'fire control time,' 'incident location,' 'number of responding personnel,' and 'TFS arrival times.'

4. **Safety Infrastructure**: This category includes data related to the status of existing infrastructure during the fire incident. It helps us understand the current system and supports decision-making for future improvements. Variables include: presence of fire alarm systems, sprinkler systems, and their operational status.

This structured approach allows us to analyze and understand various aspects of fire incidents in Toronto. All variables in the dataset have the potential to provide valuable insights and perspectives for various analyses, contributing to fire prevention and safety improvements. However, for my analysis of fire incidents across Toronto, certain variables; such as latitude and longitude, incident numbers in the CAD system, Incident_Ward, and road intersections—are being excluded from the study. A more detailed explanation will be provided in the **?@sec-appendix**.

I have selected 17 variables to begin my exploration of the dataset:

1. **__id**: A unique identifier for each record, essential for distinguishing between incidents and for tracking and referencing.

2. **Area_of_Origin**: Represents the fire's origin location, useful for analyzing geographical patterns and fire causes.

3. **Building_Status**: Indicates whether the building was occupied, under construction, or vacant, providing insights into fire severity and response needs.

4. **Business_Impact**: Classifies the economic impact of the fire on businesses, helping to understand disruptions and guide recovery efforts.

5. **Civilian_Casualties**: The count of civilian casualties, reflecting the human impact and assessing fire safety and evacuation effectiveness.

6. **Estimated_Dollar_Loss**: The monetary loss due to fire damage, key for assessing the economic burden and informing prevention efforts.

7. **Exposures**: The number of secondary fires caused by the primary incident, useful for evaluating the fire's spread and containment success.

8. **Final_Incident_Type**: The final classification of the fire (e.g., accidental, arson), critical for identifying causes and informing preventive strategies.
9. **Fire_Alarm_System_Operation**: Indicates whether the fire alarm system functioned properly, evaluating the system's effectiveness in alerting and activating suppression.
10. **Ignition_Source**: The source of ignition (e.g., electrical faults, arson), crucial for identifying fire causes and guiding safety measures.
11. **Number_of_Responding_Appliances**: The number of fire trucks and equipment dispatched, reflecting the scale of the incident and resource allocation.
12. **Number_of_Responding_Personnel**: The number of fire personnel involved, providing insight into the severity of the fire and the required response intensity.
13. **Possible_Cause**: The potential cause of the fire, useful for identifying trends and informing fire prevention and safety campaigns.
14. **Sprinkler_System_Presence**: Indicates if a sprinkler system was present, highlighting its impact on fire severity and potential casualties.
15. **TFS_Alarm_Time**: The time when the fire service was notified, key for analyzing response time and minimizing damage.
16. **TFS_Arrival_Time**: The time when the first fire service unit arrived, crucial for evaluating response speed and fire control effectiveness.
17. **TFS_Firefighter_Casualties**: The number of firefighter casualties, useful for assessing the safety of firefighting operations and improving risk management.

By focusing on these 17 variables, I aim to uncover meaningful patterns and insights that can support fire safety improvements across the city.

## 2.3 Data Cleaning

The raw dataset was stored in its original csv format to ensure data integrity and reproducibility, and later analysis data after cleaning and processing is saved into parquet file using `write_parquet` function in `arrow` package (Wickham and Hester 2024).

After loading the original raw data, I selected the columns of interest as explained in the previous section. These columns were then renamed to lowercase, and spaces were replaced with underscores to ensure consistency. After reviewing the data, I kept in mind that this study focuses on the relationship between ignition sources and fire severity. As a result, I removed rows with "Undetermined" ignition sources.

Since there are time-related variables in the dataset, I converted them to the POSIX format. Additionally, I noticed the presence of several missing (NA) and null values. Given that the

dataset contains numerous categorical variables, I considered how to handle missing data differently for these types. I first examined whether any columns contained too little information to provide valuable insights into fire severity or its predictors. After calculating, I decided to remove columns with more than 20% missing data, carefully considering their relevance to the study.

To maintain data quality, rows with more than 30% missing data are removed. After ensuring that there are no missing values in the dataset, a new variable, Response_Time, is created to calculate the time difference (in minutes) between TFS_Alarm_Time and TFS_Arrival_Time (see Figure 1). This new variable will help analyze fire response efficiency and is a key indicator for the study.

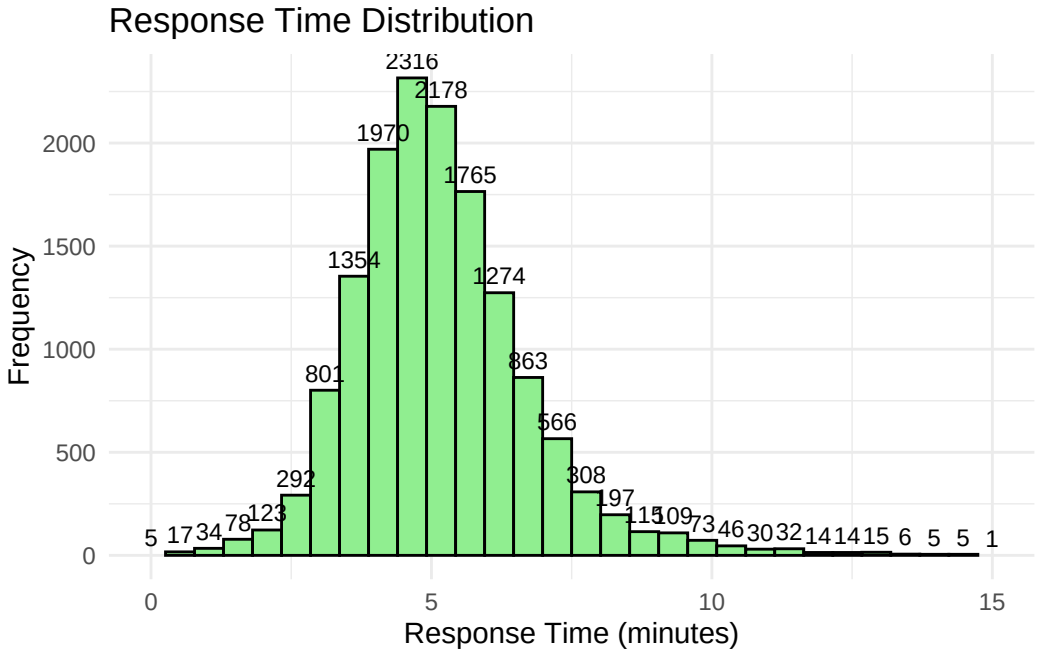


Figure 1: Distribution of response times for fire incidents in Toronto, showing the frequency of various response times (in minutes). This visualization highlights the efficiency of the Toronto Fire Service's (TFS) response. It's almost normally distributed with a mean of around 5, and the spread of not too far.

## 2.4 Data Explore

**Understanding Variables of Interest**:
To understand the distribution of various categorical and numerical variables in the fire incident dataset, I visualized each variable along with its distribution and reviewed its summary statistics. The goal is to gain insights into the key characteristics of the data, which will

support the later stages of analysis and modeling. Can check **?@sec-appendix** for visualization/summary of the distribution of all variables. Below are a few important variables considered:

1. **Area of Origin**:
   This is a categorical variable that indicates the location within a building or structure where a fire started. Initially, this variable had over 100 categories, so the goal was to categorize these locations into broader groups for easier interpretation and analysis. I grouped the areas into categories such as "Residential and Living Spaces," "Cooking and Dining Areas," and "Storage and Utility Areas," among others. These groups were then visualized to better understand the most common fire origins within the dataset (see Figure 2). Among all categories, Cooking and Dinning area stands out to be the area which most fire originate from. Living space and Transportation area also has a great amount to incidence occurred.

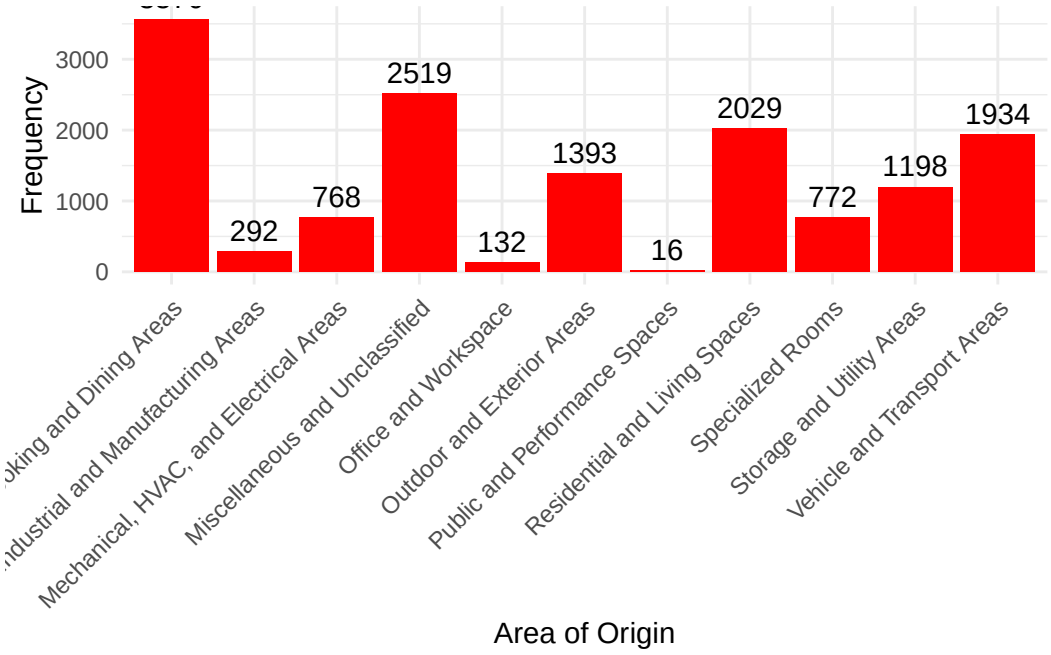


Figure 2: Distribution of the Area of Origin for fire incidents, categorized into various groups such as Residential, Office, Mechanical, and Industrial areas. The plot highlights the frequency of each area category, helping to understand the distribution of fire origins.

2. **Civilian Casualties**:
   This is a numerical variable that provides insight into the severity of the incidents. Basic descriptive statistics (e.g., mean, median, and range) were calculated using the `summary()` function. Additionally, a scatter plot was created to explore the relationship

between the number of civilian casualties and the incident ID (see Table 2). We can see that most of the time, there are no casualties. This data set ranges across over 10 years (see Figure 17 in **?@sec-appendix**), looking at all levels of casualties, and taking its average over 6 years, we can see that this indicates the city has done a good job in protecting civilians from fire places.

Table 2: Summary table showing the number of civilian casualties in fire incidents, including the frequency of each casualty count. This table summarizes the distribution of civilian casualties and provides an overview of the impact of fire incidents on civilians.

| Number_of_Casualties | Frequency |
|---|---|
| 0 | 13519 |
| 1 | 911 |
| 2 | 123 |
| 3 | 42 |
| 4 | 22 |
| 5 | 1 |
| 6 | 2 |
| 7 | 1 |
| 8 | 2 |

3. **Final Incident Type**:
   This is a categorical variable that classifies the type of incident (e.g., fire, explosion). A bar plot was created to visualize the distribution of different incident types.(see Figure 3). It's clear that there are more fire than explotion.

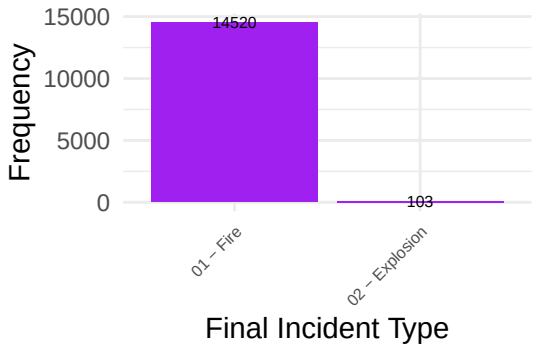


Figure 3: Bar plot illustrating the distribution of final incident types across fire incidents. The plot presents the frequency of each incident type, with the count displayed on top of the bars. We can see that there is much more fire than explotion.

4. **Ignition Source**:

This variable provides information about the cause of the fire. Categories of ignition sources, such as "Incandescent Lamp" or "Vehicle - Mechanical," were listed. The distribution of ignition sources was visualized with a bar plot (see Figure 4). We can tell that fire-related and open flame is one of the biggest source of fire incidence; cooking and heating equipment are second place as sources, which align with Figure 2 result, where Ketchen is the place most fire incidence starts off.

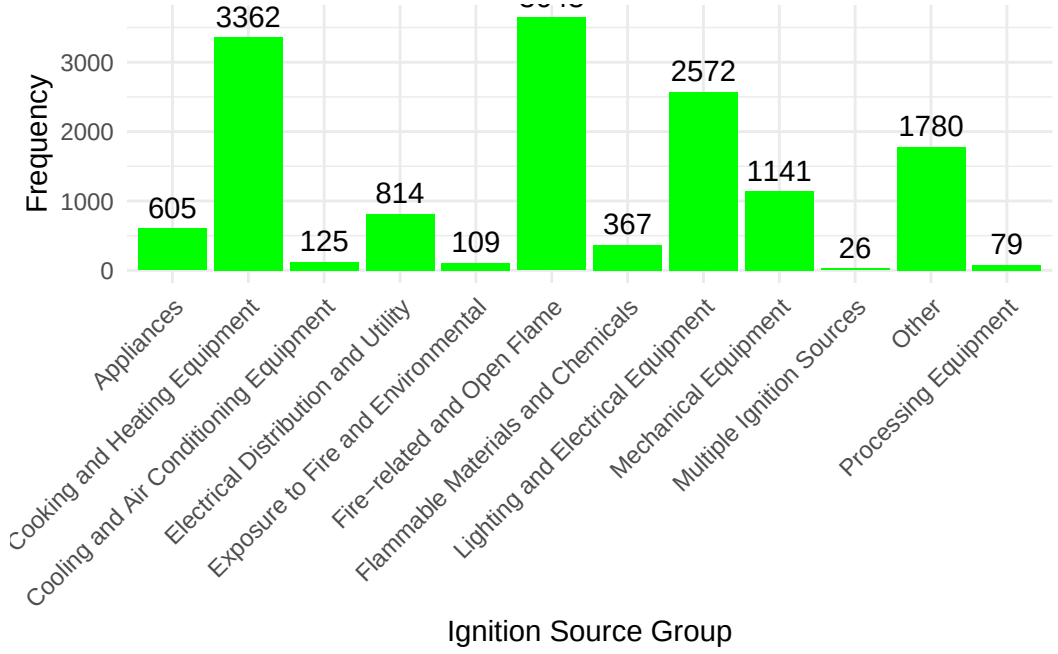


Figure 4: Bar plot illustrating the distribution of grouped ignition sources for fire incidents. The plot categorizes ignition sources into meaningful groups such as Lighting and Electrical Equipment, Cooking and Heating Equipment, and Mechanical Equipment, among others. The frequency of each group is displayed. Fire-related and Open Flame is the bigget ingnition source.

**Understanding Potential Indicators of Severity and Their Replashioships**:
To explore the relationship between financial losses and casualties, incidents were grouped into financial loss and casualty categories. A heatmap below (Figure 5) illustrates the interaction between these two indicators.

The scatter plot below showcases the relationship between the number of responding apparatuses and the logarithmic scale of financial losses (Figure 6). We can observe a slight trend between responding apparatus and financial loss A Pearson correlation coefficient was calculated to assess the strength of this relationship.

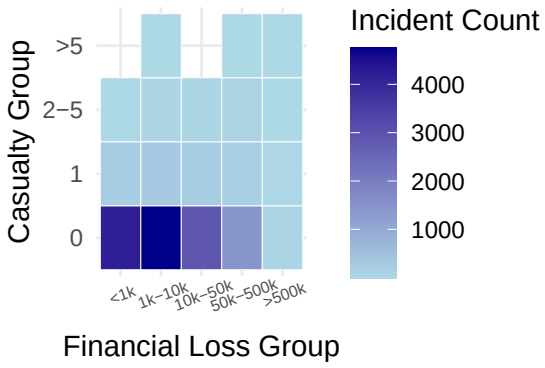


Figure 5: Heatmap illustrating the relationship between financial loss and casualty groups in fire incidents. The plot categorizes incidents based on financial loss and total casualties, with the color intensity representing the frequency of incidents in each combination of loss and casualty group. This provides insights into the severity of fire incidents, highlighting areas with higher financial losses and casualties. We can see that there are greater anount in 1k -10k region.

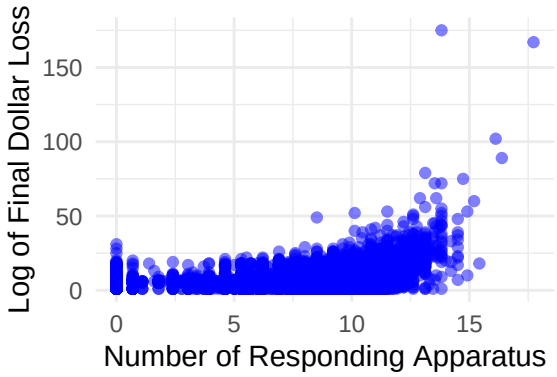


Figure 6: Scatter plot showing the relationship between the number of responding apparatus and the log-transformed financial loss in fire incidents. Each point represents a fire incident, with the x-axis displaying the log of the financial loss (with a 1 added for stability) and the y-axis showing the number of responding apparatus. There is a slight trend observed, indicating potential relationship.

# 3 Model

## 3.1 Define Severity

The severity of fire incidents was categorized based on three key factors: the number of responding fire apparatus, the estimated dollar loss, and the total number of casualties. First, the total number of responding apparatus from TFS was calculated and divided into bins after plotting and observing its distribution. Each incident was assigned to a bin based on how many apparatus responded, helping categorize incidents in terms of the scale of response.

Next, the estimated dollar loss for each fire incident was log-transformed. This was necessary because the data was extremely right-skewed, with highly influential outliers. The transformation compressed the scale of the data and managed extreme values better. The transformed data was stored in a new variable called `log_estimated_dollar_loss`. Quantiles (25th, 50th, 75th percentiles) were used to divide the dollar loss into categories to make severity thresholds more meaningful.

For casualty severity, the number of casualties, including civilians and firefighters, was categorized after observing its distribution. The categories were defined as:

- Low: No casualties
- Medium: 1 to 3 casualties
- High: 4 or more casualties

Another key factor was the incident type. If the type of incident was determined to be an explosion, it was automatically classified as High Severity regardless of other factors. The final severity of fire incidents was determined based on the following logic:

- If the incident type was an explosion, the severity was **High**.
- If casualty severity was "Low," apparatus used was in the lower bins (1-10 apparatus), and dollar loss was below the 25th percentile, the severity was **Low**.
- If casualty severity was "Medium," apparatus was in the mid-range (11-30 apparatus), or dollar loss was below the 50th percentile, the severity was **Medium**.
- If casualty severity was "High," apparatus response was in the higher bins (31-50 apparatus), or dollar loss was above the 50th percentile, the severity was **High**.

The analysis shows that most incidents fall under the Low Severity category (see Table 3), which aligns with the reality of everyday fire service responses.

Table 3: The table shows the distribution of fire incidents across Low, Medium, and High severity levels based on the number of responding apparatus, estimated dollar loss, and total casualties and incident type. Notice that there are musch more lower severe incidence.

| Severity | n |
|---|---|
| High | 105 |
| Low | 14003 |
| Medium | 515 |

## 3.2 Set Up Model For Predicting Severities

In this analysis, I aimed to predict the likelihood of high-severity fire incidents using a logistic regression model. The dataset includes several predictor variables that show potential connections to fire incident severities, as identified during prior data visualization. Among these variables, `area_of_origin_grouped` and `ignition_source_grouped` are categorical variables representing the area where the fire started and the source of ignition, respectively. The target variable for the model is the binary classification of fire severity, focusing on predicting high-severity incidents.

To prepare the target variable, the `severity` variable calculated earlier was transformed into a binary variable called `Severity_high`, where incidents classified as "High Severity" were coded as `1`. This binary transformation aligns with the analysis goal of predicting high-severity incidents. Since both predictor variables are categorical, they were converted into factors to ensure compatibility with logistic regression. Additionally, redundant spaces and commas in the variable names were cleaned to standardize the data and improve model interpretability.

To evaluate model performance, the dataset was split into a training set (80%) and a testing set (20%) using the `createDataPartition()` function from the `caret` package in R (R Core Team 2023).

## 3.3 Model Explnation

A logistic regression model was chosen because the outcome variable, Severity_high, is binary (0/1), and logistic regression is well-suited for binary classification tasks, especially when working with categorical predictor variables. The logistic regression model used the following predictors. The logistic regression model was initially fitted to the training dataset using the formula: ***Severity_high ~ area_of_origin_grouped + ignition_source_grouped.*** This initial model can be expressed as:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{j=1}^{n_{\text{area}}} \beta_{j,\text{area}} \cdot X_{j,\text{area}} + \sum_{k=1}^{n_{\text{ignition}}} \beta_{k,\text{ignition}} \cdot X_{k,\text{ignition}}$$

Where: $P$ is the probability of the event: $P(\text{Severity\_high} = 1)$.
$\text{logit}(P)$ is the log-odds of the probability.
$\beta_0$ is the intercept.
$\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the predictor variables $X_1, X_2, \ldots, X_k$.

## 3.4 Model Justification

The initial logistic regression model was evaluated for the significance of the predictors using the p-values from the model summary. Since a high p-value suggests that the variable does not significantly contribute to the model in prediction, it was removed to simplify the model and improve interpretability. This decision follows the principle outlined by Occam's Razor (Blumer et al. 1987), which proposes that people prefer simplicity.

To check for multicollinearity among the predictors, the Variance Inflation Factor (VIF) was calculated. According to standard practice, if any predictors had a high VIF (greater than 5), they were considered for removal to avoid multicollinearity, which could distort the model's estimates (Alin 2010).Based on the results, the variable `ignition_source_grouped` was removed from the model due to its non-significance (high p-value). The updated model retained only `area_of_origin_grouped`, which showed a significant relationship with the likelihood of high-severity fires. Model is:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{j=1}^{n_{\text{area}}} \beta_{j,\text{area}} \cdot X_{j,\text{area}}$$

The model's performance was also assessed using the Akaike Information Criterion (AIC), a measure of the model's goodness of fit. A lower AIC indicates a better-fitting model, and removing the insignificant variable `ignition_source_grouped` improved the AIC, providing more evidence that the updated, simpler model was a better fit for the data.

## 3.5 Model Evaluation

The updated logistic regression model was evaluated using the receiver operating characteristic (ROC) curve, with predictions made on the test dataset. The curve was plotted to assess the model's ability to discriminate between high and low-severity fire incidents. The area under the curve (AUC) is a key metric for model performance; higher AUC values indicate

better model performance (Metz 1978). The simpler model, after removing insignificant predictors, combined with its performance on the test set, provides a better understanding of the relationship between fire characteristics and severity.

# 4 Results

## 4.1 Logistic Regression Model Results

The logistic regression model was fitted to predict the probability of high severity fire incidents based on the categorical variables **area of origin**. The final model keep only significant predictors based on p-values and col-linearity diagnostics. below is a summary ( see Figure 7) of the model's estimates. The coefficients of each predictor. And Also included other information about model evaluation. It's noticeable that RMSE is not ideal, but this is already an improvement from the initial model.

The estimated coefficients for the significant predictor (`area_of_origin`) are visualized to interpret the relative influence of different categories below (Figure 8). We can tell most categories are relative to the severity of fire incidence. For example, Fire from mechanical, HVAC, and electrical Areas are highly contributing to the occurrence of ore sever fire incidence.

The ROC curve demonstrates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) for the logistic regression model. A higher AUC value signifies better predictive accuracy (Metz 1978). In this case, the curve is above the diagonal line and towards the left corners (Figure 9), indicating it's a meaningful model for prediction.

Therefore, `area_of_origin` is indeed a significant predictor of fire severity, better from it's initial model performance from comparison of AIC scores.

## 4.2 Fire infrastructure

There is no doubt that infrastructure such as fire alarm and sprinkler systems play a crucial role in controlling fire severity. The presence of these systems will reduces the risk and impact of fires. However, the key question remains: how well do these fire safety infrastructures perform in real fire incidents? We come across result of question by looking at the result we get from analyzing the data set. In this analysis, I processed the data and visualized the distribution of fire alarm system operations and sprinkler system presence across different levels of fire severity (Low, Medium, High). These visualizations provide valuable insights into how the presence of these systems correlates with the severity of fire incidents.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.45 | 0.29 | -18.84 | 0.00 |
| Area of Origin: Industrial_and_Manufacturing_Areas | 1.62 | 0.54 | 3.01 | 0.00 |
| Area of Origin: Mechanical_HVAC_and_Electrical_Areas | 2.23 | 0.36 | 6.21 | 0.00 |
| Area of Origin: Miscellaneous_and_Unclassified | 0.75 | 0.37 | 2.00 | 0.05 |
| Area of Origin: Office_and_Workspace | -13.12 | 636.54 | -0.02 | 0.98 |
| Area of Origin: Outdoor_and_Exterior_Areas | 0.23 | 0.50 | 0.45 | 0.65 |
| Area of Origin: Public_and_Performance_Spaces | -13.12 | 1809.05 | -0.01 | 0.99 |
| Area of Origin: Residential_and_Living_Spaces | 0.12 | 0.46 | 0.26 | 0.79 |
| Area of Origin: Specialized_Rooms | -0.28 | 0.77 | -0.36 | 0.72 |
| Area of Origin: Storage_and_Utility_Areas | 0.37 | 0.50 | 0.73 | 0.47 |
| Area of Origin: Vehicle_and_Transport_Areas | -0.51 | 0.58 | -0.89 | 0.37 |

Figure 7: Table showing the summary of logistic regression model estimates for predicting fire severity. The table includes coefficients for each significant predictor, their standard errors, z-values, and associated p-values. Significant predictors are highlighted, with p-values less than 0.05 indicating statistical significance. The results provide insights into the relationship between 'Area of Origin' and the likelihood of high-severity fire incidents.

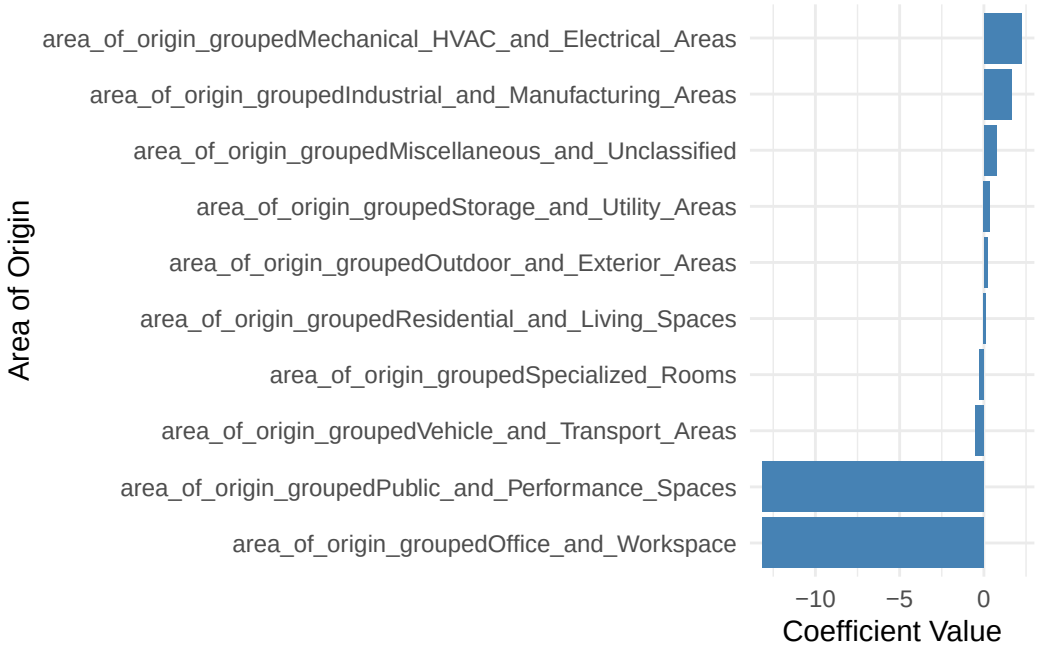


Figure 8: Bar plot of the logistic regression coefficients for the 'Area of Origin' variable, showing the effect of each location within a structure where the fire started on the severity of fire incidents. Positive coefficients suggest a higher likelihood of a high-severity fire in that area, while negative coefficients indicate a lower likelihood For example, the Electrical area is related to more severe fire incidence.

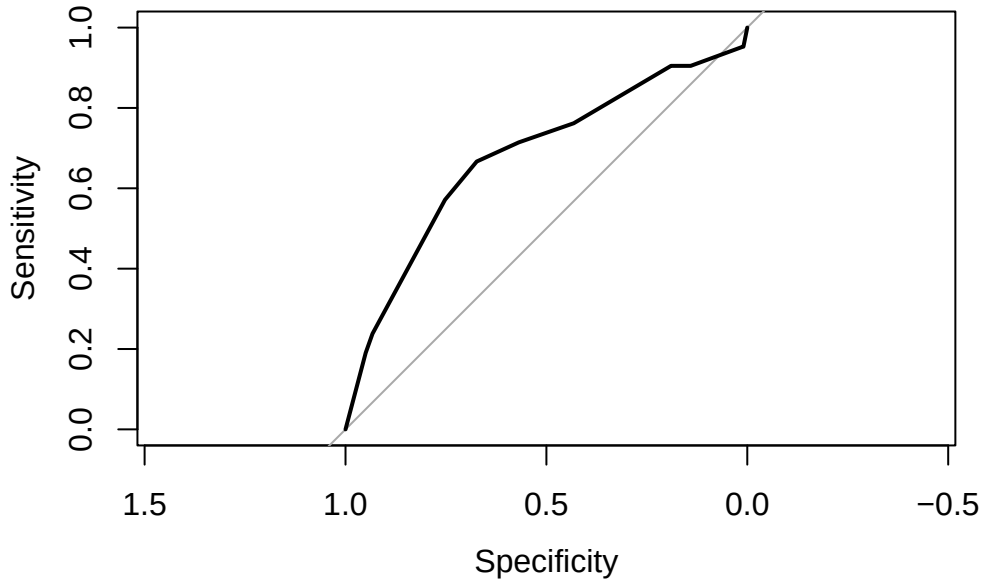


Figure 9: ROC curve for logistic regression model showing the trade-off between the true positive rate and false positive rate. This roc graph indicate model favor more False Negative.

### 4.2.1 Fire Alarm System Status

The results shown from the exploration of the proportion of fire alarm system statuses by severity (Figure 10) reveal that the fire alarm system was successfully operated during many low to medium severity fire incidents. However, it did not perform as reliably in high severity cases. It is important to note that the y-axis represents the proportion, and considering that high-severity incidents are much less frequent, their proportion might not be as large in comparison to the number of incidents. Nevertheless, this comparison provides useful insights, as we never want the system to fail when a real, highly severe incident occurs. This allows for a clearer understanding of how alarm system functionality varies. It is evident that our fire alarm system needs improvement in its operation during actual fire events.

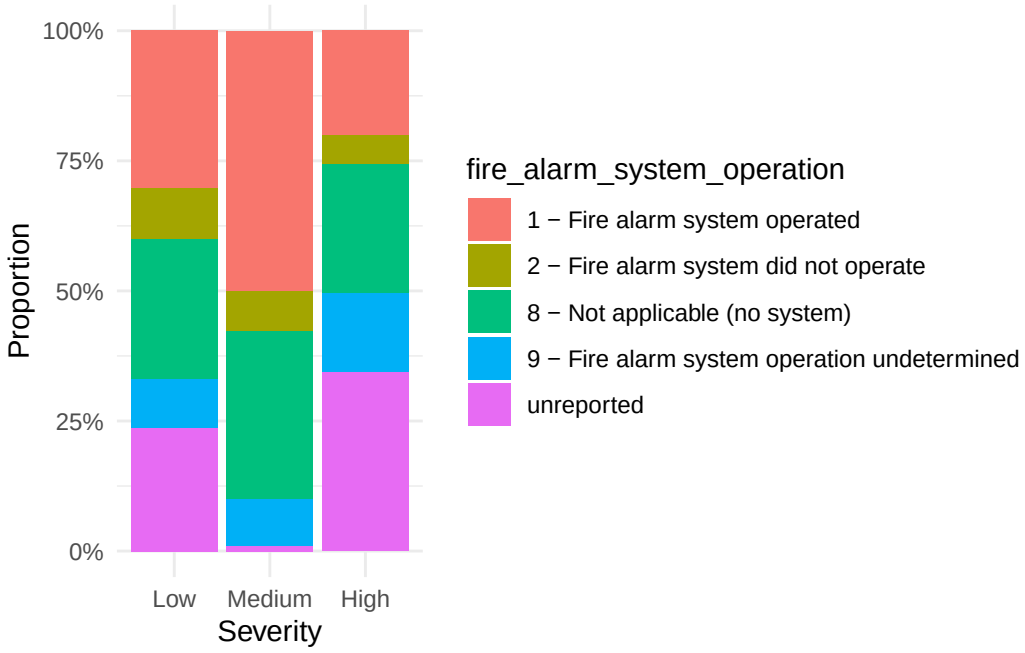


Figure 10: This shows a proportional distribution of fire alarm system operation status across different fire severity levels (Low, Medium, High). This stacked bar plot shows the relative frequencies of various fire alarm system operations within each severity group. We noctice that fire alarm was not always operated in high severity incidents, but not active in low severity fire.

### 4.2.2 Sprinkler System Status

As shown in (Figure 11), I have plotted the presence of the sprinkler system across different levels of fire severity. The operational status of the sprinkler system is represented by various

colors corresponding to statuses such as “Full sprinkler system present,” “Partial sprinkler system present,”. The results reveal that in High severity incidents, the “Undetermined” status is more common, indicating there may be a lack of clarity or understanding regarding the sprinkler system’s functionality. In Low severity incidents, the sprinkler system is most frequently reported as absent (“No sprinkler system”). In fact, this absence is the dominant status across all three levels of severity. This suggests that improvements in sprinkler system availability and implementation should be a focus for the future, particularly in higher severity fire scenarios, where the system’s functionality could play a critical role in preventing further negative impact.

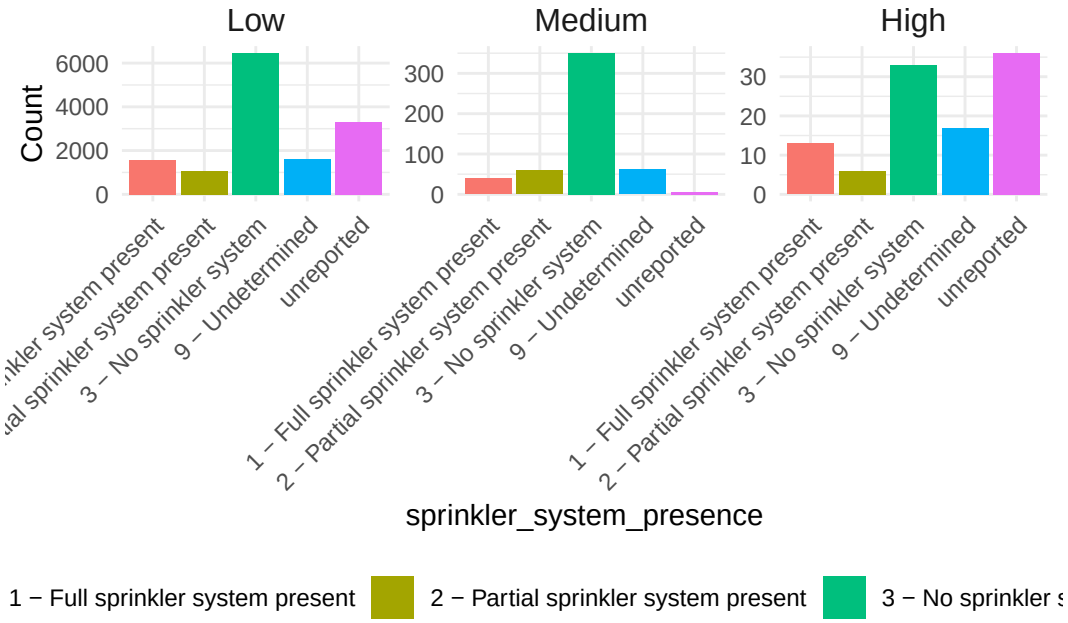


Figure 11: This joint set of three plots illustrates the distribution of sprinkler system presence across different fire severity levels (Low, Medium, High). Different colors represent the various operational statuses of the sprinkler system. The results emphasize the varying presence and functionality of sprinkler systems, showing that, on average, the sprinkler system did not operate across all levels of fire severity.

# 5 Discussion

## 5.1 Trade-off between False Positives and False Negatives

From the ROC curve, we can tell that it minimizes false negatives more in focus. Putting back into our context of fire severity production, It would be less harmful to tolerate some false

positives ( over-predicting the severity of fire); in order not to miss any positive cases (when real severe fire incidents came).

In fact, the cost of missing high-severity fires is far greater than the cost of false positives. While false positives may lead to overreaction, failing to detect a severe fire could escalate into a major disaster, resulting in significant financial losses and potentially causing more casualties. Therefore, a model that is tuned to minimize false negatives is preferable, as it prioritizes catching severe fires—even at the expense of some false positives. However, I will continue to fit additional models to strike a better balance between specificity and sensitivity.

## 5.2 Limitations and Biases in the Data

### 5.2.1 Potential Bias

The fire incidence dataset includes several categorical variables such as Ignition Source, Possible Cause, and Initial Types. During data collection, these variables, which have a limited range of categories, may introduce various types of bias.

1. **Classification Bias**: Classification bias occurs when events are mis-classified into categories that do not accurately represent the event itself (Brogden and Taylor 1950). This leads to incorrect or misleading conclusions. For example, in the dataset, there is a variable called Possible_Cause, which has over 100 categories, including categories like "Undetermined" and "Suspected Vandalism." When categorizing causes into these categories, ambiguity arises. There may not be enough evidence to definitively classify a cause, and this uncertainty can result in bias. If we were to generalize these 100 causes into fewer categories for prediction purposes, one of the generalized categories could be biased, misrepresenting the actual distribution of causes. This is one of the reasons why I decided to remove this variable from the dataset, besides the fact that it contains many missing values (NAs).
2. **Selective Reporting Bias**: Reporting bias occurs when data does not accurately reflect reality due to incomplete reporting (Saini et al. 2014). In our case, this is particularly relevant for civilian and firefighter casualties related to different fire incidents. It's possible that some fire incidents are not reported, especially smaller ones that don't require significant resources. On the other hand, large, tragic fire incidents are often over-reported or emphasized more than necessary. As a result, the reported number of casualties may be skewed. These discrepancies in the data can affect the definition of what is considered a "severe" fire. If we use these skewed measurements to make predictions or draw conclusions, the resulting insights may be biased.

Due to these biases, the quality of the data depends heavily on how the original data collectors set up their categorization systems. It's crucial to ensure that these systems are designed to capture sufficient and accurate data. By acknowledging the existence of classification and reporting biases, we can avoid overly relying on conclusions and predictions that may be skewed

by these factors. Fire incidents, by nature, have risks that should be carefully accounted for in the data collection process to ensure more reliable results.

### 5.2.2 Sampling Method in the Data

The fire incident dataset comes from the Ontario Fire Marshal (OFM) and includes incidents to which the Toronto Fire Service (TFS) responded. This data collection is not random, as it is not based on surveys or experimental studies. Instead, it is observational data collected by a specific organization. This can be considered **event-based sampling**.

Event-based sampling is a technique where data is collected when certain events occur, meeting predefined criteria (Sánchez, Guarnes, and Dormido 2009). This is different from systematically collecting data over a fixed time frame. In this case, data are recorded for every fire incident recognized by the OFM, with each fire incident treated as an event. Event-related data, such as the reason for the fire, time of occurrence, and the TFS's response actions, are recorded.

This is a **non-random sampling method**, meaning that events are not randomly selected from the entire population(Brislin and Baumgardner 1971). Here, the data is gathered based on the occurrence of specific events. As a result, the data may be skewed by the types of incidents that happen to be recorded. For example, areas with higher risks of fires, such as electrical factories according to our modelling, are more likely to be included in the dataset due to the higher frequency of fire incidents in these locations. Therefore, while the dataset provides valuable insights, it is important to acknowledge that the data are not randomly selected, which could introduce bias in the conclusions or predictions drawn from the dataset.

### 5.2.3 Privacy Protection in the Data

The fire incident dataset contains real, sensitive data collected from residents in Toronto. Therefore, it is essential to understand how the data is protected to ensure confidentiality.

1. **Excluded Data:** As mentioned on the data portal, some data have been excluded from the dataset to protect the privacy of individuals. This might explain why there are fewer high-severity cases in the dataset. For more severe incidents, it is more likely that the information could be recognized by the public, as media coverage and news publications often spread the details. As a result, these incidents may be easier to trace back to specific individuals or locations. To protect privacy, such data have been removed. However, this exclusion creates a natural incompleteness in the dataset, as these important and representative data cases, especially important to predict high severity, are no longer available for analysis.

2. **Aggregated Geographical Data**: Even though the study does not directly analyze fire incidents by district, geographical information remains important. To further protect privacy, the data collectors have aggregated exact locations into broader regions, such as large intersections or general areas. While this approach preserves privacy, it may introduce challenges for spatial analysis. Spatial analysis heavily relies on precise location data, so the aggregation of location information can limit the ability to perform detailed spatial studies and may affect the accuracy of geographical insights.

## 5.3 Government and Agency Actions for Improving Fire Safety Infrastructure

As the results from our analysis show, there is a significant need for improvement in both the facilitation and implementation of current fire safety systems. The fire alarm system, for example, demonstrated a good operational response in lower and medium severity incidents but failed to function effectively during high-severity situations. Similarly, sprinkler systems showed varying levels of performance, with a higher proportion of "Undetermined" and "Not applicable" statuses in high-severity incidents. These findings raise concerns, as high-severity fires, although less frequent, have a disproportionately severe impact. It is crucial for relevant agencies to ensure that these systems operate reliably during fire incidents, particularly in high-severity cases. Priority should be given to maintaining fire alarm systems and improving the implementation of sprinkler systems. These measures will be essential for mitigating the risks and impacts of severe fire incidents.

## 5.4 TFS Decision Making

The TFS (fire services) should prioritize decision-making before heading to the fire scene based on the predictor variables identified in the study, ensuring they reach the scene as quickly as possible without missing any high-severity incidents. It's also crucial to ensure the right amount of resources, such as firefighters, are allocated. By focusing on key variables like the area of origin, TFS can allocate resources more effectively.

For example, the analysis of fire data reveals a clear pattern in the area of origin across different severity levels. Looking at the graph (Figure 12), we can see that mechanical, HVAC, and electrical areas make up a significant portion of high-severity incidents, which aligns with the modeling results.

Thus, when TFS get reported of a fire, if there is a possibility that the fire originated from these high-risk areas, they should be prepared to deploy additional resources and firefighters to these locations. This proactive approach would not only help in saving civilian lives but also protect the firefighters themselves, ensuring a safer response. Additionally, this strategy would allow for more efficient use of social resources, minimizing the impact of the fire while ensuring a well-coordinated respose.

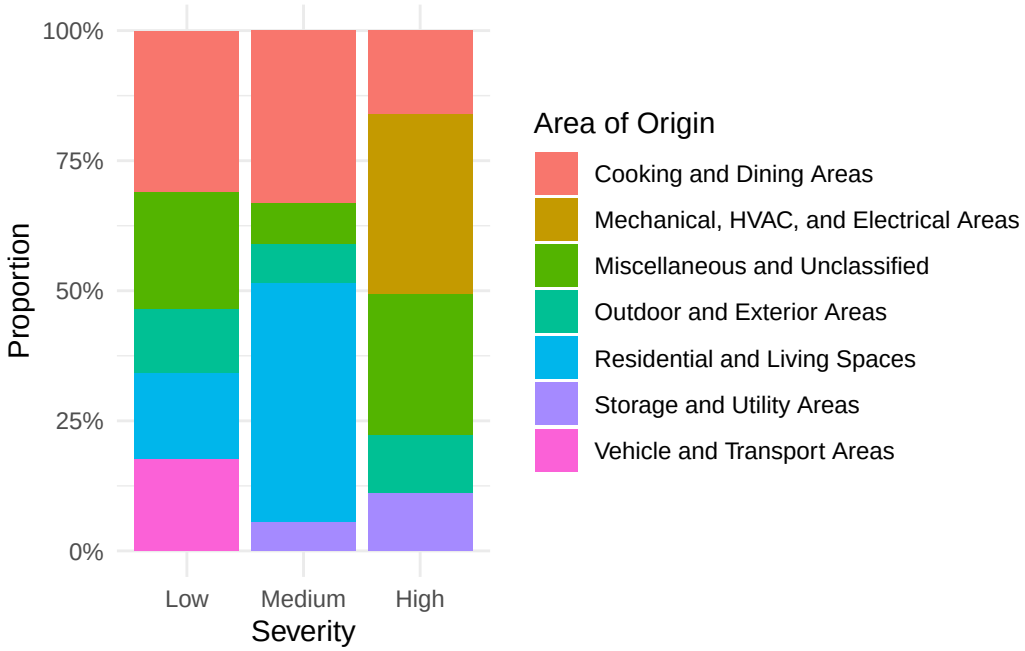


Figure 12

Using other factors, such as ignition sources (see Figure 13), which can potentially be determined before reaching the fire scene, can significantly aid in fire reporting. For example, if it's reported that the fire originated from electrical units or lightning equipment, there may be a higher likelihood of the incident being severe. Therefore, utilizing this information to make informed decisions will enhance the effectiveness of fire response efforts.

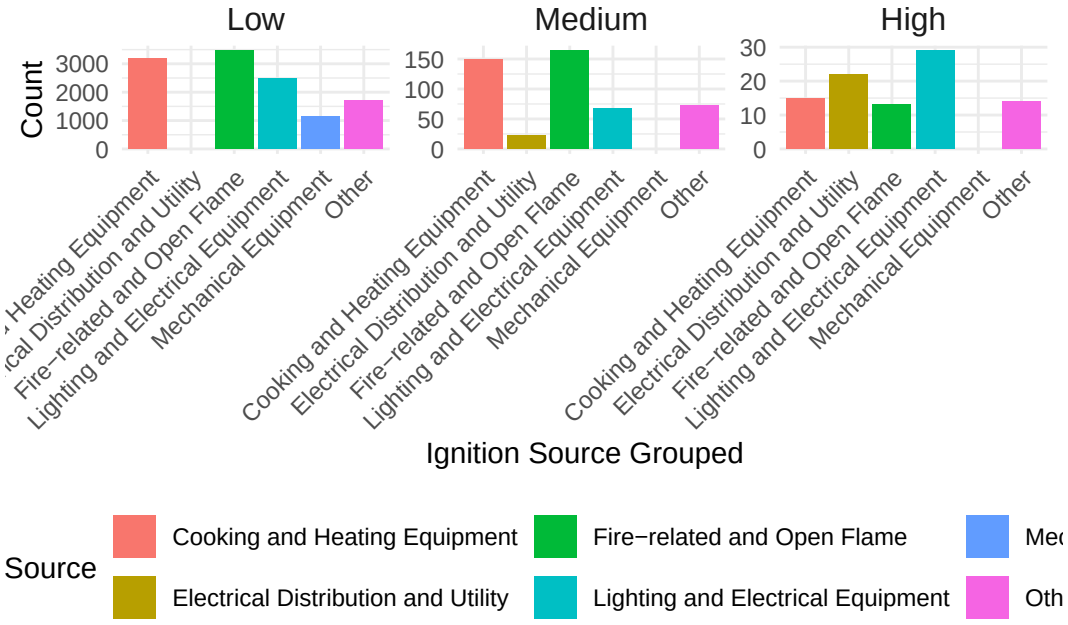


Figure 13: This plot shows the top 5 ignition sources for each fire severity level (Low, Medium, High). The bar chart illustrates how different ignition sources contribute to fire severity, with separate plots for each level of severity. For example, fires originating from electrical units or lightning equipment may have a higher chance of being severe incidents. By understanding these patterns, fire reporting can be more effective, allowing for better decision-making and resource allocation.

# Appendix

# A Variable Selection and Visualization

## A.1 Variables Selection Criterion

From this comprehensive dataset which included 43 vairables, I first selected the variables that are most relevant to my topic of study, leaving a total of 17 variables. Variables that are not essential to the study's focus include:

**Geographical Information of Incidence**: The study focuses on analyzing fire data for Toronto as a whole. Geographic details such as the specific station area, ward, or intersection are not critical to determining severity levels. Including these could even divert focus from more relevant variables.

- `Latitude`, `Longitude`, `Incident_Station_Area`, `Incident_Ward`, and `Intersection`.

**Some Safety System-Related Variables**: The study aims to study some infrasture system suhc as alarm and sprinkler systems. Some variables providing redundant information about these systems were excluded. The dataset also included variables on smoke alarm systems; however, due to the excessive number of null values, these variables were excluded as they made meaningful analysis challenging.

- `Fire_Alarm_System_Presence`, `Fire_Alarm_System_Impact_on_Evacuation`, `Smoke_Alarm_at_Fire_Origin`, `Smoke_Alarm_at_Fire_Origin_Alarm_Failure`, `Smoke_Alarm_at_Fire_Origin_Alarm_Type`, and `Sprinkler_System_Operation`.

**Repeated Information**: Certain variables contained information that overlapped with variables already chosen for inclusion. To simplify the analysis, these redundant features were removed.

- `Count_of_Persons_Rescued`, `Estimated_Number_Of_Persons_Displaced`, `Material_First_Ignited`, `Initial_CAD_Event_Type`, `Level_Of_Origin`, `Extent_Of_Fire`, `Status_of_Fire_On_Arrival`, and `Property_Use`.

**Unnecessary Details**: There are detailes in the dataset that is not directly related to the purpose of this study. Incident-specific details, such as how long firefighters took to respond or clear the fire, are more relevant for evaluating operational efficiency. These do not directly contribute to understanding the severity levels.

- `Ext_agent_app_or_defer_time`, `Fire_Under_Control_Time`, `Last_TFS_Unit_Clear_Time`, `Method_Of_Fire_Control`, and `Incident_Number`.

Upon reviewing the 17 initially selected variables of interest, I observed that some columns contain a significant proportion of missing values (NAs), see Table 5. As a result, certain columns (with more than 20% missing data) should be considered for removal if their relevance to the study topic is not as significant as other variables that should remain.

Table 5: This table shows the percentage of missing values for each variable of interest. Variables such as 'area_of_origin,' 'building_status,' and 'business_impact' exhibit high missing percentages, indicating potential areas for further investigation or data imputation, while others like '_id' and 'final_incident_type' have negligible missing values. Understanding these missing values is crucial for the completeness and quality of the analysis..

| | Missing_Percentage |
|---|---|
| _id | 0.00 |
| area_of_origin | 0.00 |
| building_status | 23.59 |
| business_impact | 23.60 |
| civilian_casualties | 13.95 |
| estimated_dollar_loss | 0.04 |
| exposures | 84.05 |
| final_incident_type | 0.00 |
| fire_alarm_system_operation | 23.60 |
| ignition_source | 0.00 |
| number_of_responding_apparatus | 0.00 |
| number_of_responding_personnel | 0.00 |
| possible_cause | 0.01 |
| sprinkler_system_presence | 23.60 |
| tfs_alarm_time | 0.00 |
| tfs_arrival_time | 0.00 |
| tfs_firefighter_casualties | 0.00 |

The variables removed are:

1. **Building_Status**: This column provides information on the condition of the building, but it is not directly related to the core objectives of the study, which focus on fire severity, casualties, and financial losses. With more than 20% missing data, it was deemed unnecessary.
2. **Business_Impact**: While this column tracks the impact of the fire on businesses, it overlaps with the **Estimated_Dollar_Loss** column, which provides more comprehensive financial data. Due to the high missing rate, this column was removed.

3. **Exposures**: This column captures the number of exposure fires but does not directly contribute to understanding fire severity or its impact on casualties and losses. With more than 20% missing data, it was excluded from the analysis.

Kept variables: These variables are essential for the completeness of the analysis. Since all of them are categorical data entries, missing values are replaced with "unreported."

- **Fire_Alarm_System_Operation**: This variable is crucial because the study aims to understand fire infrastructure.
- **Sprinkler_System_Presence**: This is also an important variable for understanding fire safety. Given that there are fewer variables under "Safety Infrastructure," it is essential to retain them.

After dealing with columns with a big proportion of missing values, removing rows with a certain amount of missing values would still remain in a relatively representative dataset. Therefore, I decided to remove any columns with more than 30% missing values to the usability of the dataset fro analysis.

## A.2 Distribution and Summary of Cleaned Variables

Below are the plots or summary statistics showing the distribution of all variables of interest in the analysis dataset, after cleaning and removing columns. This includes the created variable 'Response Time.'. The description of each variable is sourced from the data portal in Open Data Toronto [Fire Services (2024)].

1. **Area of Origin**: OFM Area of Origin code and description, specifying where the fire started. This information will contribute to the understanding what are some origins of fire and how it may influence fire severity or outcomes. 75 different recorded origins, grouped into 11 larger groups. (see Figure 2)
2. **Civilian Casualties**: Count of civilian casualties (injuries or fatalities). This is directly relates to the severity of fire incidents.(see Table 2)
3. **Estimated Dollar Loss**: Estimated financial loss caused by the fire. This is directly relates to the severity of fire incidents. (see Table 6)

```
Estimated of Loss   Frequency
0       :  1       Min.   :   1.00
1       :  1       1st Qu.:   1.00
2       :  1       Median :   2.00
3       :  1       Mean   :  67.39
4       :  1       3rd Qu.:  21.00
5       :  1       Max.   :1468.00
(Other):211
```

Table 6: The table summarizes the distribution of estimated dollar losses from fire incidents.

| Estimated of Loss | Frequency |
|---|---|
| 0 : 1 | Min. : 1.00 |
| 1 : 1 | 1st Qu.: 1.00 |
| 2 : 1 | Median : 2.00 |
| 3 : 1 | Mean : 67.39 |
| 4 : 1 | 3rd Qu.: 21.00 |
| 5 : 1 | Max. :1468.00 |
| (Other):211 | NA |

4. **Final Incidence Type**: Final classification of the fire incident This can give information about the nature of the incident, helpful to determine severity. (see Figure 3)
5. **Fire Alarm System Operation**: Describes the operation of fire alarm systems during the incident. This is important for assessing the role of alarms, this is a part of fire infrastructure.(see Figure 14)

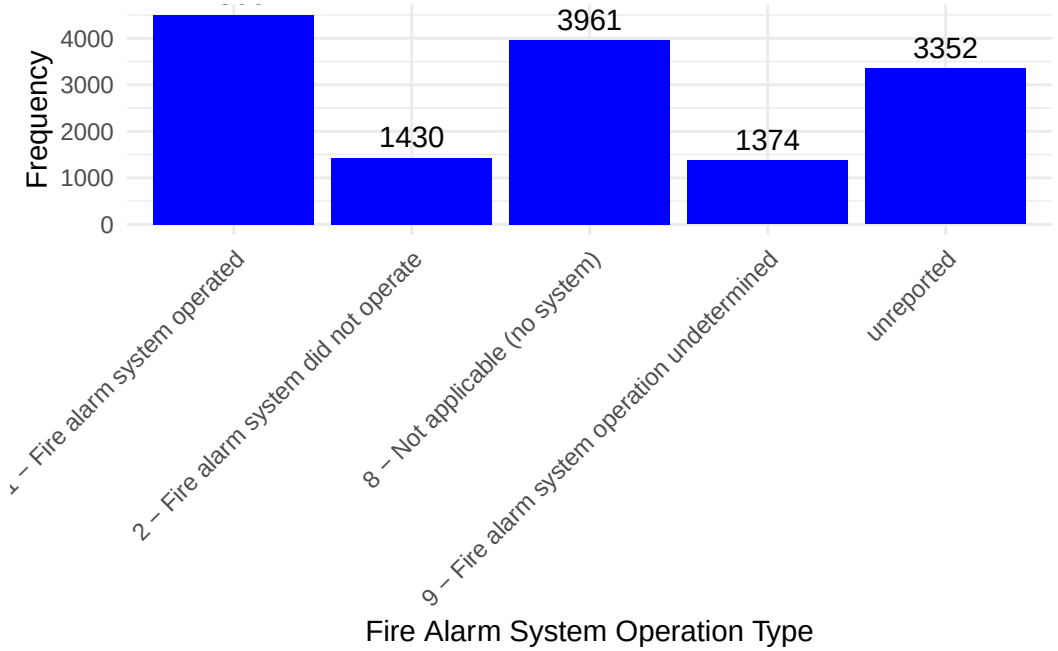


Figure 14: This bar chart shows the distribution of fire alarm system operation types. It highlights the frequency of each type of fire alarm system operation. Operated successfully most of the time.

6. **Ignition Source**: OFM Ignition Source code and description, detailing how the fire started. This is for understanding factors that contribute to fire ignition (see Figure 4)

7. **Number of Responding Apparatus**: Number of TFS apparatus that responded to the incident. This may correlate with the severity as it reflects the scale of response. (See Figure 15)

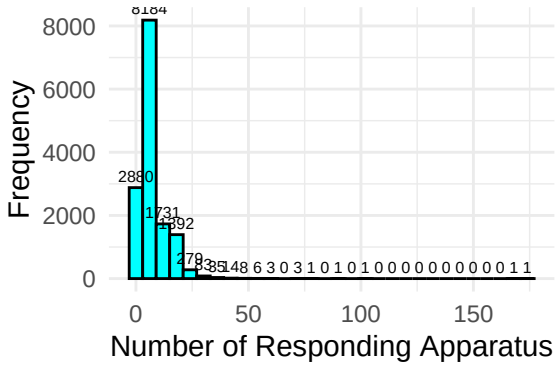


Figure 15: This histogram illustrates the distribution of the number of responding apparatus for fire incidents. The x-axis represents the number of apparatus dispatched, while the y-axis shows the frequency of incidents. The date is right skewed, closer to 1.

8. **Number of Responding Personnel**: Number of TFS personnel that responded to the incident. This reflect the magnitude of resources deployed, could relate with fire's severity similar as `Number_of_Responding_Apparatus`. (See Figure 16)

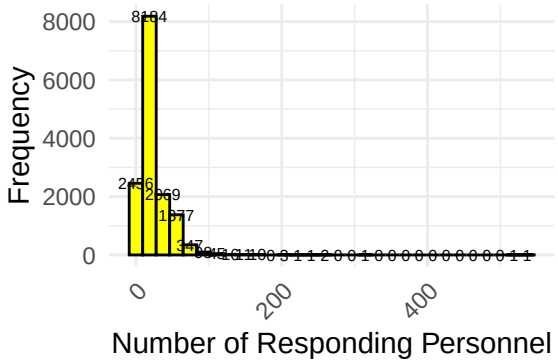


Figure 16: This histogram shows the distribution of the number of responding personnel for fire incidents. The x-axis indicates the number of personnel dispatched, while the y-axis represents the frequency of such incidents. Similar dsitribution as Figure 15.

9. **Possible Cause**: OFM Possible Cause code and description, explaining the potential cause of the fire. This tells the circumstances leading to the fire. (See Table 7)

Table 7: A comprehensive table showing top 6 possible fire casues.

| Possible Cause | Frequency |
|---|---|
| 52 - Electrical Failure | 3004 |
| 44 - Unattended | 1894 |
| 45 - Improperly Discarded | 1718 |
| 60 - Other unintentional cause, not classified | 1438 |
| 47 - Improper handling of ignition source or ignited material | 1035 |
| 99 - Undetermined | 810 |

11. **Sprinkler System Presence**: Describes whether a sprinkler system was present at the location. This provided information of the sprinkler system, which is also a part of fire infrastructure. (see Table 8)

Table 8: Table showing the frequency of sprinkler system presence across all fire incidents.

| sprinkler_system_presence | Frequncy |
|---|---|
| 1 - Full sprinkler system present | 1625 |
| 2 - Partial sprinkler system present | 1127 |
| 3 - No sprinkler system | 6835 |
| 9 - Undetermined | 1684 |
| unreported | 3352 |

12. **TFS Alarm Time**: Time stamp of when TFS was notified of the incident. This useful when analyzing response times and their impact on severity, as well as learning about the amount of incidents across times.(Amount of Fire Incidents by Year, see Figure 17) (Fire Incidents by Month and Year, see Figure 18)

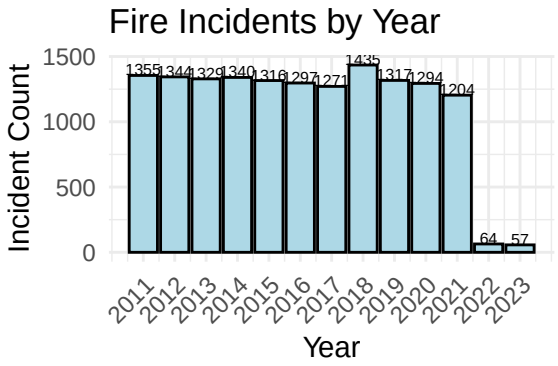


Figure 17: This bar chart shows the number of fire incidents by year, from 2011 to 2023.

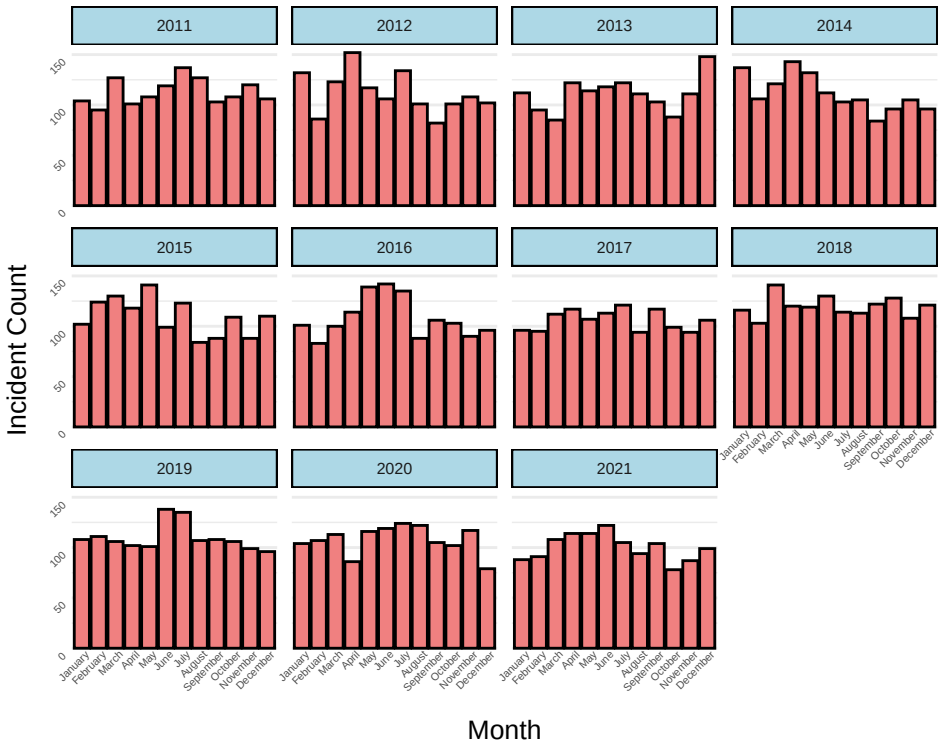


Figure 18: Monthly Incident Count by Year (Excluding 2022 and 2023 becasue from prvious plot, we can see they don't contribute much information)

14. **Firefighter Casualties**: Count of TFS firefighter casualties (injuries or fatalities). This is useful when determining the severity of fire. (See Figure 19)

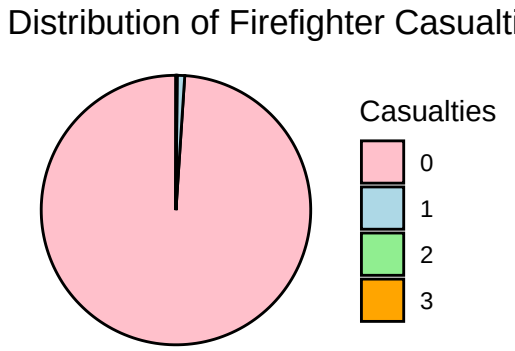


Figure 19: The Pie chart showing the distribution of firefighter casualties, rnaging from 0-3. With being the maximum amount of occurence.

15. **Response Time**: This is created variable. It is calculated as the difference between the `TFS_Alarm_Time` (when the fire service is notified of the incident) and the `TFS_Arrival_Time`. (See Figure 1)

## References


Alin, Aylin. 2010. "Multicollinearity." *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (3): 370–74.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1987. "Occam's Razor." *Information Processing Letters* 24 (6): 377–80.

Brislin, Richard W, and Steve R Baumgardner. 1971. "Non-Random Sampling of Individualsin Cross-Cultural Research." *Journal of Cross-Cultural Psychology* 2 (4): 397–400.

Brogden, Hubert E, and Erwin K Taylor. 1950. "The Theory and Classification of Criterion Bias." *Educational and Psychological Measurement* 10 (2): 159–83.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." https://CRAN.R-project.org/package=xgboost.

Delignette-Muller, D., and C. Dutang. 2023. "Fitdistrplus: An r Package for Fitting Distributions." https://CRAN.R-project.org/package=fitdistrplus.

Fire Services. 2024. *Fire Incidents*. https://open.toronto.ca/dataset/fire-incidents/.

Fox, John E., Matthew J. K. Kropf, and David J. Byrne. 2021. "Modelsummary: Model Summaries and Coefficient Tables." https://CRAN.R-project.org/package=modelsummary.

Gelman, A., P. Dunson, D. Vehtari, et al. 2020. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://CRAN.R-project.org/package=rstanarm.

Government of Ontario. n.d. "Office of the Fire Marshal." https://www.ontario.ca/page/office-fire-marshal#section-3.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://CRAN.R-project.org/package=lubridate.

Kuhn, Max. 2023. "Caret: Classification and Regression Training." https://CRAN.R-project.org/package=caret.

Metz, Charles E. 1978. "Basic Principles of ROC Analysis." In *Seminars in Nuclear Medicine*, 8:283–98. 4. Elsevier.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, L., C. Turck, and C. H. S. H. 2023. "pROC: A Package for AUC and ROC Curve Analysis." https://CRAN.R-project.org/package=pROC.

Saini, Pooja, Yoon K Loke, Carrol Gamble, Douglas G Altman, Paula R Williamson, and Jamie J Kirkham. 2014. "Selective Reporting Bias of Harm Outcomes Within Studies: Findings from a Cohort of Systematic Reviews." *Bmj* 349.

Sánchez, José, Miguel Ángel Guarnes, and Sebastián Dormido. 2009. "On the Application of Different Event-Based Sampling Strategies to the Control of a Simple Industrial Process." *Sensors* 9 (9): 6795–6818.

Therneau, Terry, and Beth Atkinson. 2015. "Rpart: Recursive Partitioning and Regression Trees." https://CRAN.R-project.org/package=rpart.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Jennifer Bryan, et al. 2023. “Tidyr: Tidy Messy Data.” https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023a. “Dplyr: A Grammar of Data Manipulation.” https://CRAN.R-project.org/package=dplyr.

———. 2023b. “The Tidyverse.” https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Jim Hester. 2024. “Arrow: R Interface to Apache Arrow.” https://cran.r-project.org/package=arrow.

Xie, Yihui. 2023. “Knitr: A General-Purpose Package for Dynamic Report Generation in r.” https://CRAN.R-project.org/package=knitr.