The Presence of Fire Alarm Systems Mitigates the Losses of Residential Fire - Research Based on Toronto Fire Incidents Data From 2011 to 2019

Yuzhi Pi

2022-04-27

Abstract

Data on fire incidents responded by Toronto Fire from 2011 to 2019 was pulled from Open Data Toronto. The dataset is cleaned and streamlined so that a Multiple Linear Regression model and four Simple Linear Regression Model can be established to analyze how the presence and the working condition of the fire alarm system (including the smoke detector) and the sprinkler system have an affect on estimated damage (in CAD) for residential fire cases. Through the result of this research, there is a weak relationship between the presence/operating status of the two systems and the estimated economic loss of fire - households with a working system on average incur less economic loss compared to households without. Analyzing the dataset also siginifies the importance of installing fire prevention systems in the Kitchen, as it is the most common area for fire to occur among all other living areas. Code and data supportiing this analysis is available at https://github.com/YuzhiPi/Final-Paper.

Introduction

1.1 Motivation

In 2019, Toronto Fire Services (TFS) responded to 35,334 reported fires/activated fire alarms, of which 1,682 cases (only 5%) were actual working fires. Out of the working fires, 53% were residential property fires, with case numbers totalling 891 (Toronto Fire Services 2020). Due to the lack of professional fire hazard training, lack of fire extinguishing equipment, the randomness of the fire incidents, and other factors, residential fires are the most common fire incident type and could lead to significant economic and property damages, in addition to severe injuries, casualties of both civilian and firefighters.

Two types of mechanisms have been widely implemented to mitigate the adverse consequences of residential fire: detective and corrective mechanisms. In a typical residential house or apartment, the fire alarm system acts as a detective control, while the sprinkler system acts as a corrective control. A fire alarm system consists of three components: 1) the initiating devices, including smoke detectors, heat detectors, or manual initiating devices that can detect fire in the first place; 2) fire notification devices such as bells, chimes, and horns that notify the residence when a fire is detected; 3) a fire alarm control panel that manages the initiating devices and monitor the alarm system, sending notifications to authorities (building front desk, or the fire services) to ask for additional assistance (Fire Protection Author 2020). The alarm system is triggered in the very early stages of fire, while false alarms are very common when smoke detectors are placed in the kitchen. On the other hand, a sprinkler system is considered a corrective mechanism as it could actually suppress or extinguish a fire. The sprinkler system is only triggered when there is an actual fire, as it is heat activated.

The installation of the two fire systems could be costly and requires regular maintenance. Although according to The Ontario Fire Code, every home should at least have one or more working smoke alarms (City of Toronto 2021), the presence of the alarm does not guarantee the working condition of the alarm. The alarm could be disabled by the residents or was not functioning due to lack of maintenance. As a resident of

a high-rise apartment building located in downtown Toronto, I have personally heard many complaints regarding the fire alarm system from other high-rise apartment residences. False alarms are often triggered by building residences, which could be irritating and frustrating, especially at night. Consequently, many have decided to disable their smoke detector to avoid false alarms. Moreover, the sprinkler system is more costly to install and is not mandatory in all residential homes except in high-rise apartment buildings.

With this being said, this report aims to investigate the actual benefit of having a working fire alarm system and a sprinkler system in residential buildings by constructing a Multiple Linear Regression (MLR) model. The "benefit" of a working system is quantified and observed through the estimated loss of the fire (in CAD). Comparisons will be made between residential fire cases with/without the presence of a fire alarm system and a sprinkler system. The result of this research should raise awareness of the actual benefit of implementing such fire control mechanisms. In addition to the model, this report will provide a primary insight into 1) the most common origins of residential fire, 2) possible causes of residential fire, and 3) how the fire was eventually controlled. These three insights should better assist the residents in deciding where to install the fire system at home and how effective the sprinkler system is at controlling the fire.

1.2 Paper Outline

This paper will include the following sections: a data section where the dataset used for the analysis will be presented and explained in detail; a model section that would provide an insight into the building process of the MLR model and justifies the model choice; a result section where the result of the model will be discussed, accompanying by appropriate visualizations and explanations of the result; finally a discussion section where implications, limitations, and future research outlook will be discussed.

Data

2.1 R and Data Processing

All data included in this report is cleaned, analyzed, and visualized with R (R Core Team 2021). Several packages need to be installed for data processing purposes. For data cleaning and manipulation, tidyverse(Wickham et al. 2019), dplyr(Wickham et al. 2021), and stringr(Wickham 2019) were used. kableExtra (Zhu 2021) is used for constructing tables presented in this report, while ggplot2(Wickham 2016) is used for data visualization. For model building and other data manipulation involving statistical analysis, car(Fox and Weisberg 2019) package is used.

2.2 Dataset

The dataset used in this report is available on the Open Data Portal hosted by the City of Toronto. The data was first downloaded from the portal in csv format and then imported into R Studio for further analysis using default R package. The data is listed on the portal titled "Fire Incidents"(City of Toronto 2022a), published and maintained by Toronto Fire Services, which is the only organization in Toronto City that responses to emergencies relates to all hazards (City of Toronto 2022b). The dataset was last refreshed on April 18, 2022.

The dataset aims to include all information related to fire incidents that were sent to the Ontario Fire Marshal. Only fire incidents that were responded by Toronto Fire were included in this dataset. The data is refreshed and maintained annually, and it currently includes all fire incidents from 2011 to 2019. The Fire Marshal records all demographic, time, location, types of fire, presence and working conditions of fire prevention mechanisms, damage of the fire, and all other information relates to the response from Toronto Fire.

Table 1: Preview of Cleaned Dataset

	X_id	Area_of_Origin	Estimated_Dollar_Loss	Fire_Alarm_System_Operation	Fire_Alarm_System_Presence	Method_Of_Fire_Control	Possible_Cause	Sprinkler_System_Operation	Sprinkler_System_Presence
7	52615	22 - Sleeping Area or Bedroom (inc. patients room, dormitory, etc)	2000	0	0	1 - Extinguished by fire department	20 - Design/Construction/Installation/Maintenance Deficiency	0	0
8	52616	55 - Mechanical/Electrical Services Room	100000	1	1	1 - Extinguished by fire department	52 - Electrical Failure	0	0
12	52620	24 - Cooking Area or Kitchen	500	1	1	3 - Extinguished by occupant	44 - Unattended	0	0
17	52625	25 - Washroom or Bathroom (toilet,restroom/locker room)	15000	0	0	1 - Extinguished by fire department	52 - Electrical Failure	0	0
18	52626	24 - Cooking Area or Kitchen	0	1	1	4 - Fire self extinguished	98 - Unintentional, cause undetermined	0	0
19	52627	24 - Cooking Area or Kitchen	60000	0	0	1 - Extinguished by fire department	80 - Exposure fire	0	0
23	52631	64 - Porch or Balcony	100	0	1	3 - Extinguished by occupant	99 - Undetermined	0	1
24	52632	24 - Cooking Area or Kitchen	3000	1	1	2 - Extinguished by automatic system	51 - Mechanical Failure	1	1
36	52644	24 - Cooking Area or Kitchen	200	0	0	3 - Extinguished by occupant	60 - Other unintentional cause, not classified	0	0
45	52653	50 - Basement/cellar (not partitioned)	500000	0	0	1 - Extinguished by fire department	46 - Used or Placed too close to combustibles	0	0
46	52654	64 - Porch or Balcony	1000	0	1	1 - Extinguished by fire department	98 - Unintentional, cause undetermined	0	1
49	52657	64 - Porch or Balcony	1000	1	1	1 - Extinguished by fire department	45 - Improperly Discarded	0	0
52	52660	78 - Attached Deck	1000	0	0	1 - Extinguished by fire department	45 - Improperly Discarded	0	0
55	52663	24 - Cooking Area or Kitchen	10000	1	1	1 - Extinguished by fire department	44 - Unattended	0	0
63	52671	24 - Cooking Area or Kitchen	5000	0	0	1 - Extinguished by fire department	47 - Improper handling of ignition source or ignited material	0	0

2.3 Dataset Preparation

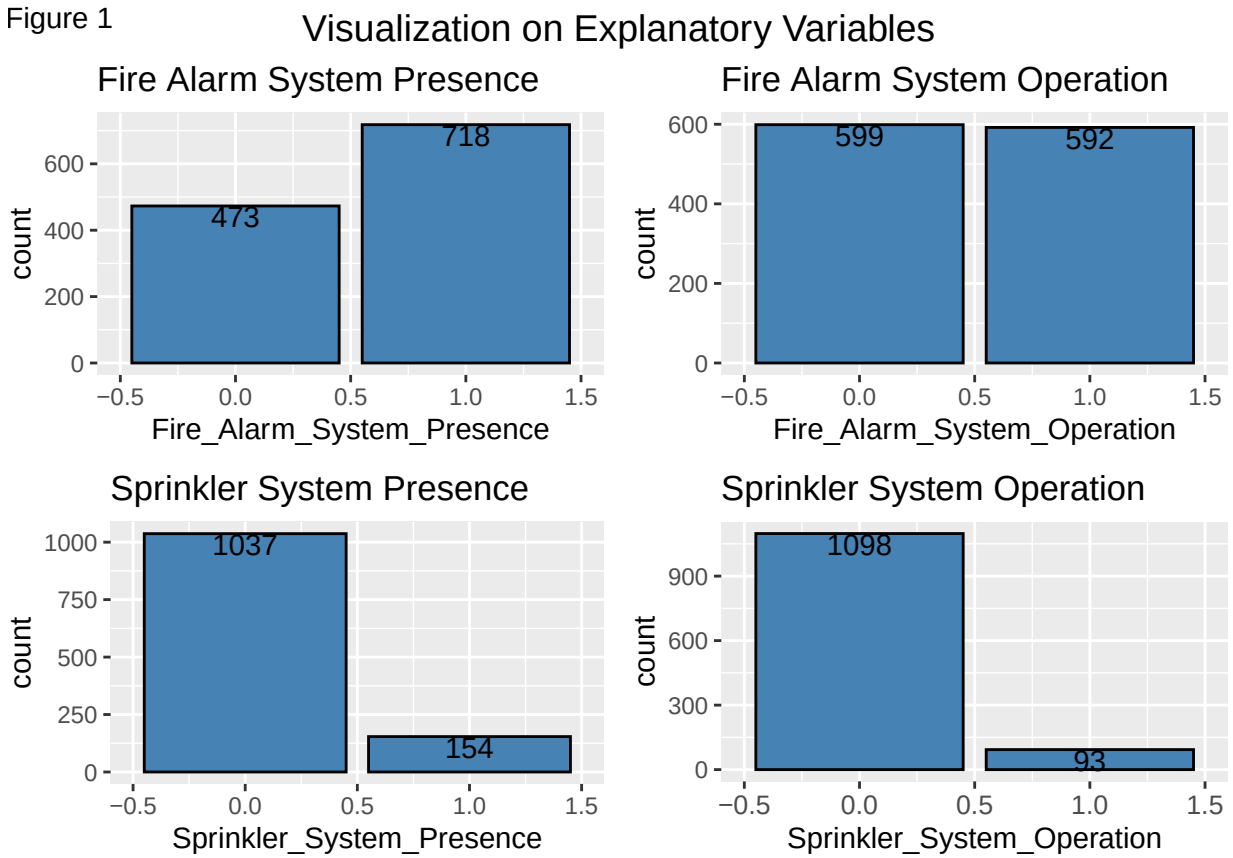
Given the main purpose of this research, which is to investigate how the presence of fire alarm system affects property loss due to residential fire by constructuring a MLR model, I am particularly interested in these factors: estimated loss for each fire incidence (dependent variable in this research), the working condition of the fire prevention mechanisms (fire alarm system and the sprinkler system), which are the explanatory factors in this research. On the top of that, for the three additional research criteria, columns titled 1) Area_of_Origin, 2) Method_Of_Fire_Control and 3) Possible_Cause need to be included in the dataset.

With this, the raw dataset needs to be thoroughly cleaned and prepared for the purpose of model building and further analysis. Table 1 is a preview of the result dataset - the result dataset contains only residential fire cases by fuzzy matching fire event type with keyword "residential"; rows containing blanks and N/As were removed, and columns relating to the presences and operating status of the fire alarm/sprinklers system were reconstructed to dummy variables - 1 represents the presence/operation of the system while 0 means otherwise. All rows where the status of the system are classified as "undefined" were removed to ensure the quality of the model. After data cleaning, the dataset contains 1191 rows of observations.

2.4 Dataset Visualization

To better understand the dataset, I conducted a preliminary visualization on the explanatory variables (the presence/working condition of the fire alarm and sprinkler system) of the model. Referring to Figure 1, we observe that among the sample of Toronto residential fire cases, while 60% of the household has fire alarm system in place, barely 50% of households' system were in working condition. On the other hand, only 12.93% of the household has installed a sprinkler system, while only 7.8% of households' sprinkler system were in working condition. We see that although some households have a fire prevention mechanism in place, they are not in working condition. This is the primary reason why all four variables will be included in the model as they should have different effect on the estimated loss of fire.

Figure 1

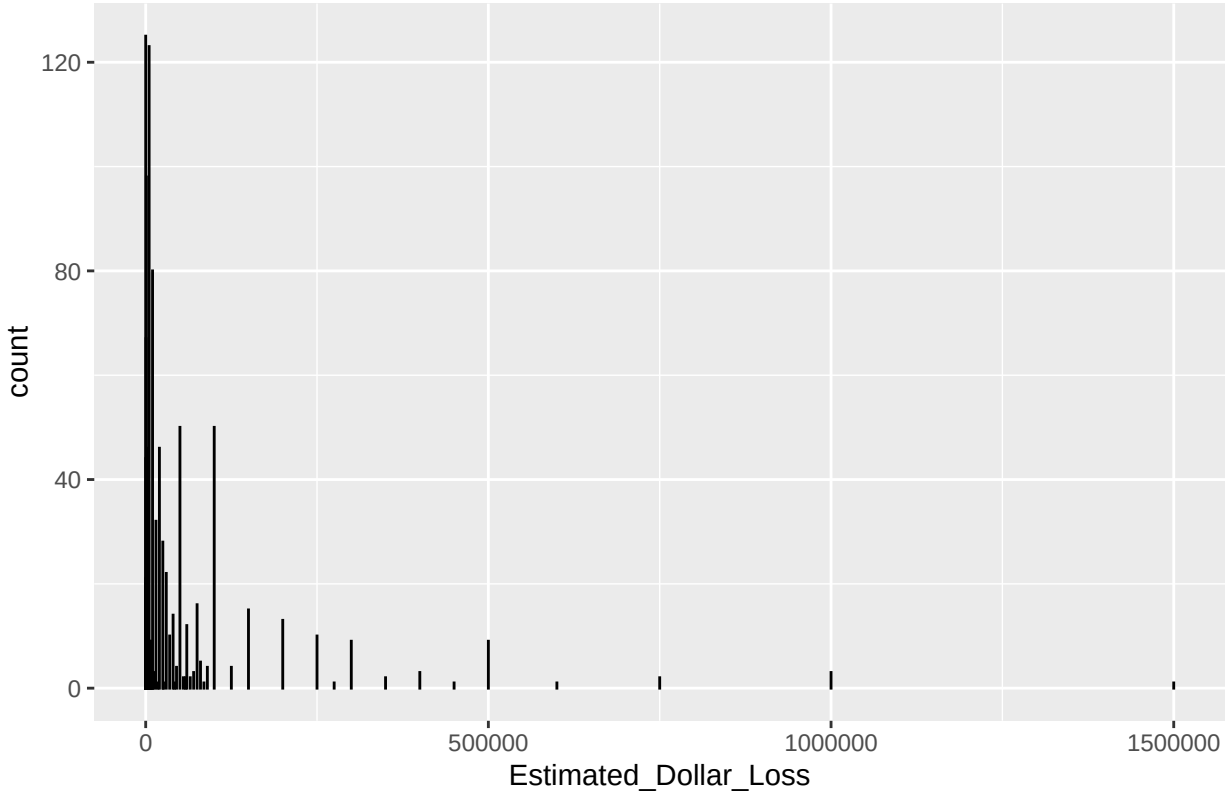


On the other hand, by plotting the estimated loss for each residential fire case, we observe that the estimated loss amount is extremely right-skewed, which signifies that most of the fire losses are of relatively smaller amounts. Table 2 provides a better insight into the frequency of the estimated fire losses - out of 1191 observations, 493 residential fire cases have losses that are between 0 to 10,000 CAD. This trend is reflected in Figure 2, as most of the observations are clustered around the left side of the bar chart.

Table 2: Most Common Fire Damage (in CAD)

	Estimated_Loss	Frequency
1	0	125
31	5000	123
24	1000	98
39	10000	80
18	500	67

Figure 2: Estimated Loss Due to Fire (in CAD)



2.5 Data Limitaion

Although the dataset is available through Toronto Open Data Portal, there is no further clarification on the actual collection method of the data. We can only assume that this is a primary data provided by Toronto Fire to the Ontario Fire Marshal. According to the data description on Toronto Open Data Portal, "Some incidents have been excluded pursuant to exemptions under Section 8 of Municipal Freedom of Information and Protection of Privacy Act (MFIPPA)" (City of Toronto 2022a). Consequently, the dataset obtained from the source is a subset of all fire incidence took place in the City of Toronto. Additionally, as the raw dataset contains missing data, rows with missing data were removed and therefore excluded from the analysis. This further limits the dataset's ability to represent all Toronto fire incidences.

On the other hand, during the preliminary data visualization process, Figure 2 has demonstrated a strong right-skew trend in the dependent variable. This is an indication that for model building purposes, transformation needs to be performed on the dependent variable to improve the accuracy of the MLR model. The reasons behind the non-normality and its implication will be further explained in the discussion section later in this report.

With this being said, there is no other similar datasets available for use given the main goal of this research. The fact that this is a official dataset posted on Toronto Open Data Portal provides some assurance for the data's accuracy, existence, and occurance. However, the completeness of the dataset remains unassured.

Model

3.1 Model Overview

A Multiple Linear Regression (MLR) Model is established to investigate how the presence/working condition of the fire prevention mechanism impacts the economic loss of residential fire in Toronto. The MLR analysis is performed using R (R Core Team 2021). The dependent variable (Y_1) is extracted from the column titled "Estimated_Dollar_Loss", while the explanatory variables are "Fire_Alarm_System_Operation" (x_1), "Fire_Alarm_System_Presence" (x_2), "Sprinkler_System_Operation" (x_3), and "Sprinkler_System_Presence" (x_4). The model takes the simplest MLR model form as follows:

$$Y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

A MLR Model is suitable for this research as all of the variables are of numerical values, although they are self-constructed dummy variables. The MLR will provide useful results that allow us to quantify the average difference between the economic loss for each dummy status (0 or 1), thus drawing further discussions on which fire prevention mechanism is more effective at reducing losses on average, and the importance of maintaining the system, ensuring the system's operative status. For example, β_0 will represent the average estimated loss (in CAD) for Toronto residential fire if there were no fire alarm system nor sprinklers (as x_2 and x_4 takes on the value of 0, x_1 and x_2 will naturally become zero). All other coefficients (β_1 to β_4) will represent the average difference between economic loss for each dummy status - in other words, β_1 will represent the average difference in economic loss for a house with or without an operating fire alarm system, while holding other variables constant.

3.2 Model Building Process and Validation

During the model building process, given that this is a MLR model, assumptions such as multicollinearity and normalities were checked. See appendix for the details on the tests. For validating the models, a test and a training were set up by equally splitting the cleaned dataset in half - however, the performance of the model is poor for both test and training sets. Table 3 presents the R-squared and the residual standard error (RSE) for the model in both sets, which confirms its poor performance. The extremely small R-squared and large RSE shows that the model barely explains the variations in the dependent variable. Nevertheless, I will continue to examine the results and the implications I have obtained from the model and other analysis, including additional simple linear regression models, with the hope that they will motivate similar research and better data collection processes in the future.

Table 3: R-squared and RSE for Training and Testing Set

	Training Set	Testing Set
Adjusted R-squared	0.04325	0.01635
Residual Standard Error	72270	119600

3.3 Supplemental Model Analysis - Simple Linear Regression

As discussed previously, with a small r-squared and large RSE value, the MLR does a fairly poor job at exploring the true relationship between the dependent and the independent variables. With this, I decided

to introduce simpler models, which are simple linear regression models that only examine one independent variable at a time so that the relationship could be examined without the influence of other independent variables. Below is an example of the SLR that will be generated in this report:

$$Y_1 = \beta_0 + \beta_1 x_1 + \varepsilon$$

In the SLR model, the dependent variable (Y_1) is still the estimated loss in CAD, while the independent variable is one of the four explanatory variables for each model. Hence, four SLR models were constructed during the analysis processes. The SLR models were also built based on the training set established previously and were tested using the testing set. The result of the four SLR models will be discussed later in the result section alongside with implications in the discussion section.

3.4 Threats to Model Validity

The poor performance of the statistical models could be explained by the quality of the dataset and this applies to both the MLR and the SLR models.

Although this dataset is provided by Toronto Fire service and posted on Open Data Toronto hosted by the City of Toronto, the dependent variable itself is only an estimation recorded by the firefighters after arriving at the location. As a result, the value of the dependent variable falls onto a few specific values, as shown per Table 2 and Figure 2. This could lead to significant bias on the data itself, as there is no further explanation on how the estimations were generated and their accuracy. As biases and inaccuracies exist in the raw data, the predictive power of the model will be largely impacted. The estimation approach also led to a large amount of outliers in the dependent variable, as per demonstrated in Figure 3. The large amount of outliers, which again, is caused by the poor estimation process as smaller loss tends to be estimated, potentially had a huge impact on the accuracy of the model, as the model was heavily impacted by the outliers.

Result

4.1 Results from the MLR

From the result of the MLR from the training set (Table 4), we observe that there is some relationship between the economic loss and the presence of the fire alarm system for residential fire cases in the City of Toronto. The coefficient (β_2) is significant with 99.9% confidence with a value of -43585. This means that on average, while holding other conditions constant, in an event of fire, residential houses with a fire alarm system will save 43,585 CAD in losses compared to those without a fire alarm system.

Other variables included in the model seem to produce no significant coefficients. Therefore, we cannot conclude that there is a relationship between the rest of the variables and the dependent variable. To put into context, the model even suggests a positive relationship between the operation of the two systems with the estimated loss, that houses with an operating fire prevention system will lead to greater loss - both β_1 and β_3 are positive as shown per Table 4.

Other than the coefficients, the intercept of the model is also significant with a value of 59415. The interpretation is that on average, in the event of a fire, residential houses without the presence of any fire prevention system will incur an estimated loss of 59,415 CAD. A loss of almost 60k CAD is huge as the average dollar loss in the dataset is 34,582 CAD while the median is only 5,000 CAD. The result of the intercept suggests that households without both systems will on average bear more economic losses than other households that at least have one presence.

Table 4: Summary Statistics for MLR (Training Set)

	Estimate	Std. Error	t-value	Pr(>t)
Intercept	50468	4736	10.657	<2e-16 ***
Fire_Alarm_System_Operation	12122	9704	1.249	0.212
Fire_Alarm_System_Presence	-40362	9950	-4.056	5.65e-05 ***
Sprinkler_System_Operation	-11890	14332	-0.830	0.407
Sprinkler_System_Presence	-2206	11167	-0.198	0.843

Signif. codes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

4.2 Supplemental Analysis Result

As the model provides little information on whether a true relationship exists between the presence/operation status of the two systems and the estimated loss (in CAD), I have constructed a few other analyses that would enhance the findings of this research.

Result of the SLR

Referring to Figure 4, which shows the scatter plot of all four SLR models constructed, we observe a moderate upward sloping trendline in all SLR models. That is, if we observe each explanatory variable individually, there exists a relationship between the status of the alarm/sprinkler system and the estimated loss. In contrast to the result of the MLR for β_1 and β_3, all of the SLR suggests that on average, in the event of fire, the presence and the working of fire alarm/sprinkler systems will lead to smaller economic loss for the household. For instant, according to Table 5, which is a summary of all the coefficients and intercepts of the SLR, the simple linear regression model suggests that on average, households with a non-operating fire alarm system will incur an estimated loss of 41,208 CAD while the loss 20,762 CAD less for a house with an operating alarm system. All of the coefficients are negative for the SLR models, and all of the intercepts and the coefficients are significant with 99% confidence.

Figure 4

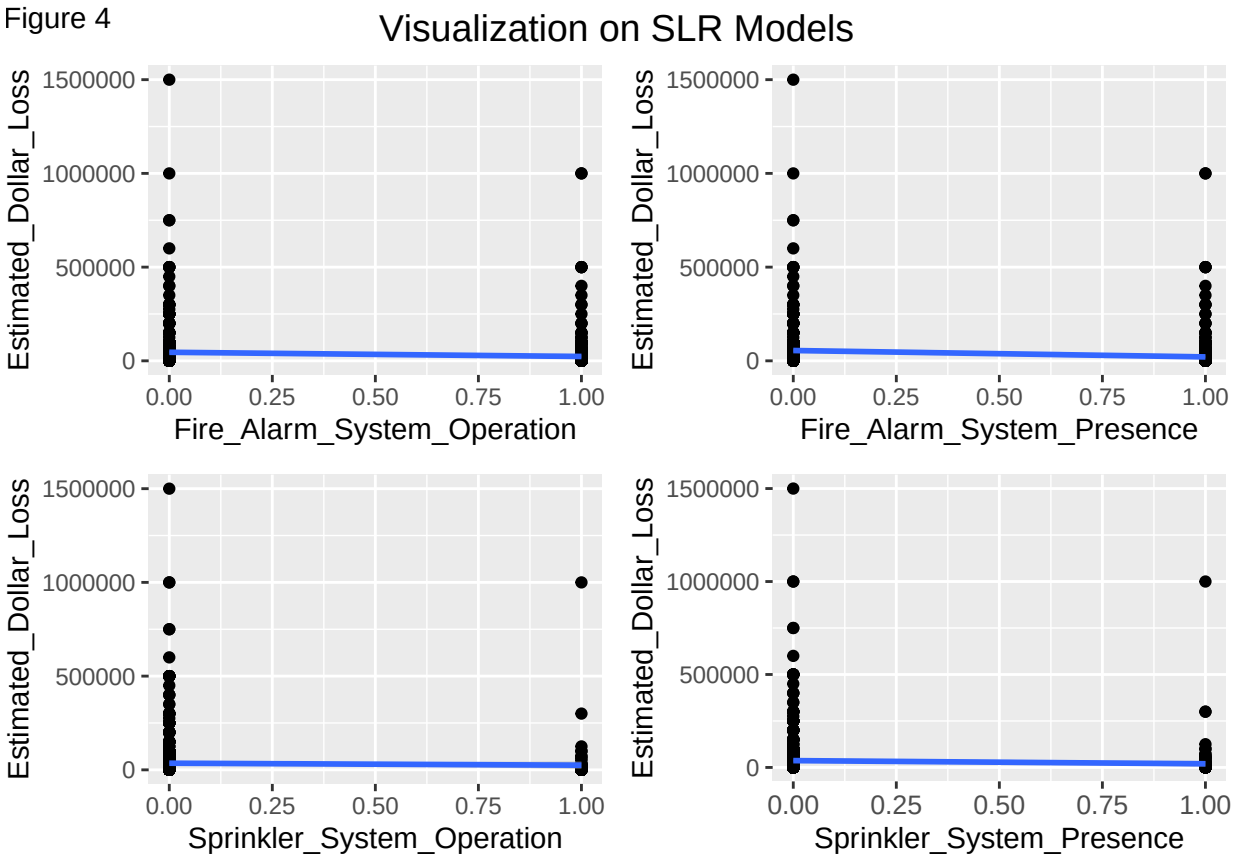


Table 5: Summary Statistics for SLR (Training Set)

Independent Variable	Estimate	Std. Error	t-value	Pr(>t)
Fire_Alarm_System_Operation				
Coefficient	41208	4324	9.732	<2e-16 ***
β_1	-20762	6003	-3.458	0.000582 ***
Fire_Alarm_System_Presence				
Coefficient	50572	4732	10.686	<2e-16 ***
β_1	-32368	6067	-5.335	1.36e-07 ***
Sprinkler_System_Operation				
Coefficient	32483	3129	10.382	<2e-16 ***
β_1	-23855	12067	-1.977	0.0485 *
Sprinkler_System_Presence				
Coefficient	33308	3218	10.352	<2e-16 ***
β_1	-20356	9315	-2.185	0.0293

Signif. codes: *** p < 0.001; ** p < 0.01; * p < 0.05

Most Common Fire Origins and Causes

Table 6 and Table 7 is the result of a preliminary sorting by occurrence of the two variables: 1) Area_of_Origin and 2) Possible_Cause.

Among all 1,191 residential fire cases included in the final dataset, 336 (28%) cases of fire stemmed from the cooking area or kitchen, while 117 (10%) and 110 (9%) cases took place in the sleeping area and the porch, respectively. The result shows that compared to other areas in a residential household, the kitchen is more susceptible to fire hazard. It is surprising that trash and rubbish storage ranked the fifth in the table, as

Table 6: Most Common Fire Origin

	Area of Origin	Frequency
8	24 - Cooking Area or Kitchen	336
6	22 - Sleeping Area or Bedroom (inc. patients room, dormitory, etc)	117
37	64 - Porch or Balcony	110
5	21 - Living Area (e.g. living, TV, recreation, etc)	87
20	44 - Trash, Rubbish Storage (inc garbage chute room, garbage/industri	84

Table 7: Most Common Fire Control Method

	Area of Origin	Frequency
1	1 - Extinguished by fire department	761
3	3 - Extinguished by occupant	273
4	4 - Fire self extinguished	86
2	2 - Extinguished by automatic system	54
5	5 - Action taken unclassified	17

compared to other areas where residences will spend longer time in the area, residences' frequency of visiting the trash and rubbish storage is significantly lower. However, among 1,191 cases, still 84 (7%) cases of fire originated from this area.

On the other hand, looking at the most common causes of fire, 201 cases of fire were caused by attendance which amounts to 17% of the residential fire cases in the dataset. Other than that, 163 cases were caused by improper disposal and 155 cases were caused by electrical failure. As a result, negligence is the primary cause of residential fire.

Most Common Fire Control Method

Table 8 outlines the most common fire control method. Among 1,191 residential fire cases, 761 cases (64%) were controlled by the fire department, while 272 (23%) cases were extinguished by the occupant. In contrast, only 54 (4.5%) cases were controlled by the automatic system, in this case, most likely referring to the sprinkler system. This means that despite the working status of the system, its ability to control fire remains weak as most of the time other forces are required to control the fire.

Table 8: Most Common Causes of Fire

	Area of Origin	Frequency
9	44 - Unattended	201
10	45 - Improperly Discarded	163
17	52 - Electrical Failure	155
22	99 - Undetermined	139
18	60 - Other unintentional cause, not classified	97

Discussion

5.1 General Findings

1) Despite poor model fitting given the value of the r-squared and MSE, there is a weak relationship between the estimated loss of the residential fire and the working status of the fire control system. The MLR model claims that the presence of the fire alarm system has the most effect on the loss amount, as the model suggests on average a house with a fire alarm system will incur a loss that is 43,585 CAD less than a household without. However, this effect for a sprinkler system is only 9,219 CAD.
2) Although the MLR model suggests the operation of the alarm and the sprinkler system could even lead to more economic loss as the coefficients (β_2 and β_4) were positive in both model, the SLR model constructed on each variable produced all negative coefficients, which contradicts the results provided by the MLR model. The SLR model indicates that the presence and the operation of the two systems on average do prevent the residence from incurring larger economic loss in the event of fire.
3) Through the additional research conducted for the three additional categorical variables: 1) Area_of_Origin, 2) Method_Of_Fire_Control and 3) Possible_Cause included in the cleaned dataset, it is not surprising that kitchen is the most common area for residential fire to occur. In fact, when looking at the most common fire causes, all of the possible causes are more likely to take place in the kitchen compared to other living areas. For instance, "unattended" ranks the first place for possible causes of fire - it is extremely common for residences' to leave an operating stove or oven unattended and causing fire. Electrical failure, which ranks the third common causes of fire, can also take place in the kitchen as kitchens are often plugged with small cooking appliances and larger appliances such as stove and oven. These factors all contribute and support the previous observation that the kitchen is the most common area for residential fire to take place.
4) Finally, the result of the analysis conducted on the most common method of fire control poses the question of whether the sprinkler system is truly effective at controlling the fire. Only 4.5% of the residential fire cases included in the dataset were controlled by the sprinkler system, while 64% were controlled by Toronto Fire. However, according to the statistical research conducted by the National Fire Protection Association (Marty Ahrens 2021), sprinkler systems are effective 88% of the time at controlling the fire, as long as they were present and the fire was large enough to trigger them. Hence, the result in this research could be explained by the fact that the data was not sorted based on the operation status of the sprinkler system, which is a future research outlook to consider.

5.2 Implications

The installation and the regular maintenance of the fire prevention systems should be encouraged

Combining the results of both the MLR and SLR models, a downward sloping trend associated with a negative coefficient can be observed across models. That is, the systems on average do help the residence to mitigate economic losses due to fire. Therefore, the government should provide more educational programs or subsidies to encourage the installation of the two systems. On the other hand, the result of the SLR states that compared to the presence of the sprinkler system, its operation creates more benefit in terms of mitigating losses. In other words, it is not enough to just have a sprinkler system in place, regular maintenance is crucial for fully exploiting the benefit associated with it. However, due to the limitation of the model accuracy, a cost-benefit analysis cannot be conducted to fully understand the benefit of the systems. This will be further explained in the sections below.

The Fire Prevention Systems should be primarily installed in the Kitchen

Based on the analysis result, the kitchen is the most common place of residential fire, alongside higher exposure to the most common causes of fire. Therefore, if a household decides to install the fire prevention

system (the alarm system and the sprinkler system), the primary area of focus should be the kitchen. On the other hand, given the overall environment of the kitchen area, including smoke and grease, alarm systems installed in the kitchen are susceptible to higher rates of false alarms.

5.3 Ethics and biases

There exists an ethical dilemma for this research, as fire is dangerous - it will be unethical for researchers to gather experimental data to better analyze the relationship. Consequently, the quality of the dataset cannot be assured and the model built could contain preliminary bias from the data collection method, which is an explanation of the poor r-squared and other model fitting metrics.

Moreover, the limitation of this dataset is the huge amount of missing data and the method itself for obtaining the dependent variable. From the data cleaning process, rows with empty cells were removed as they failed to contain useful information on the independent variables. Therefore, the cleaned datasets' ability to represent the residential fire cases in Toronto was weakened. The loss of data points also decreased the accuracy of the models, which ultimately had an impact on the predictive power of the models. Additionally, the dependent variable in this report is based on an estimation provided by the onsite fire fighters, who does not have the full knowledge of evaluating the true loss of the fire. Hence, from the exploratory data analysis process, I observed that the dependent variables were extremely right skewed and does not take on a lot of values. The lack of continuity and randomness lead to bias in the dependent variable itself. The estimations tend to take on smaller values, while the accuracy and the reasonableness of the estimations remain unknown.

Another bias exists in this research is the omitted variable bias. Given the limitation of the data type, the four independent variables selected in this research could be influenced by other factors, which could also impact the loss of the fire. For example, the income of the household could influence the decision on whether a system will be installed - while it could also impact the loss of the fire as households with higher income will own properties with higher values. Another example will be the duration and the size of the fire. For smaller fires, the sprinkler systems may not be triggered and therefore the loss could be smaller.

5.4 Research Outlook

The key implication on future research outlook is the quality of the dataset - for the analysis purpose of this report, the dataset should contain more quantitative variables that could provide further insight into the demographics of the fire households, numbers of the alarm system and the sprinkler systems installed, the size of the fire, the extent of the fire, etc. These variables provide greater versatility of the independent variables, which potentially allows the model to better explain the variations in the dependent variable. The additional variables could also be incorporated into the model as control variables, which will mitigate the omitted variable bias and improve model accuracy.

Another improvement is that the estimation on the loss could be performed by a professional appraiser, as it will significantly decrease the estimation bias embedded in the dependent variable. However, a downside of this is that the cost of the research will be increased, and additional funding of the research will be required.

Appendix

Datasheet

A.1 Motivation

Q: For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A: The dataset was created to investigate the relationship between fire prevention methods and its impact on the estimated loss

Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A: The dataset was originally created by Toronto Fire Service and was cleaned and manipulated by Yuzhi Pi

Q: Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A: N/A

Q: Any other comment?

A: For privacy purposes personal information is not provided and exact address have been aggregated to the nearest major or minor intersection. Some incidents have been excluded pursuant to exemptions under Section 8 of Municipal Freedom of Information and Protection of Privacy Act for the raw dataset (City of Toronto 2022a)

A.2 Composition

Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A: The dataset is composed of all fire cases controlled by Toronto Fire from 2011 to 2019. The dataset provides information similar to what is sent to the Ontario Fire Marshal relating to only fire Incidents to which Toronto Fire responds in more detail than the dataset including all incident types. The Dataset includes only Fire incidents as defined by the Ontario Fire Marshal (City of Toronto 2022a)

Q: How many instances are there in total (of each type, if appropriate)?

A: The raw dataset contains 17536 unique fire cases in total

Q: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A: It could be thought so as the dataset is a population of all fire cases controlled by Toronto Fire. However there is a possibility of active fire and Toronto Fire was not notified

Q: What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

A: The Fire Marshal records all demographic, time, location, types of fire, presence and working conditions of fire prevention mechanisms, damage of the fire, and all other information relates to the response from Toronto Fire.

Q: Is there a label or target associated with each instance? If so, please provide a description.

A: Yes, the first column is an unique ID number and it is associated with each case

Q: Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A: Yes, for a number of cases information on the working status and presence of fire prevention system were missing therefore the cell is left blank

Q: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A: There exists no explicit relationship between individual instances.

Q: Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A: For model building purposes the data is split into half after sorting into training set and testing set

Q: Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A: There are no errors, sources of noise, or redundancies in the dataset presented.

Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A: The dataset is self-contained

Q: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

A: Yes, the data in the dataset relating to geometric and demographic information is confidential

Q: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A: No.

Q: Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

A: No

Q: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

A: Yes, the dataset contains geometric location of the fire incidences

Q: Any other comments?

A: N/A.

A.3 Collection Process

Q: How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A: The data was collected as a part of the record provided by Toronto Fire to the Ontario Fire Marshal

Q: What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

A: Not enough information is provided on the collection of the data, but it could be assumed that information is collected through both manual human curation and internal software used by Toronto Fire

Q: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A: N/A

Q: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A: Data were collected by Toronto Fire

Q: Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

A: The dataset was collected over the period of 2011 to 2019

Q: Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

A: N/A

Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

A: The data was collected by Toronto Fire, therefore it is obtained via a third party

Q: Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

A: It was unclear whether the individual is aware of the data collection

Q: Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

A: No information was provided by the Open Data Toronto Portal

Q: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

A: N/A

Q: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

A: N/A

Q: Any other comments?

A: N/A

A.4 Preprocessing/cleaning/labelling

Q: Was any preprocessing/cleaning/labelling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

A: Yes. The data was first screen for missing data and all rows containing missing data were removed. Afterwards, the remaining data was matched by location to sort out all the residential fire cases, and then the data was further manipulated to create dummy variables used in the model building process. All other columns that are irrelevant to the purpose of this research was removed

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labelled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

A: https://github.com/YuzhiPi/Final-Paper

Q: Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

A: https://www.rstudio.com

Q: Any other comments?

A: N/A

A.5 Uses

Q: Has the dataset been used for any tasks already? If so, please provide a description.

A: No information was provided on the previous use of the dataset

Q: Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A: N/A

Q: What (other) tasks could the dataset be used for?

A: The dataset could be used to constructed visualization on the characteristics of the Fire cases in Toronto over the period

Q: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labelled that might im- pact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

A: The users need to be aware of the causes of fire and the type of fire (residential, commercial, etc) as they will have impact on the consequences of the fire (economic loss, casualty, etc)

Q: Are there tasks for which the dataset should not be used? If so, please provide a description.

A: N/A

Q: Any other comments?

A: N/A

A.6 Distribution

Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A: N/A

Q: How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A: it does not have DOI, but it is distributed through Open Data Toronto Portal and will be available on Github https://github.com/YuzhiPi/Final-Paper in the folder input

Q: When will the dataset be distributed?

A: N/A, it is available at all times

Q: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A: N/A

Q: Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A: N/A

Q: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A: N/A

Q: Any other comments?

A.7 Maintenance

Q: Who will be supporting/hosting/maintaining the dataset?

A: Fire Services, the contact information is kevin.ku@toronto.ca

Q: How can the owner/curator/manager of the dataset be contacted(e.g., email address)?

Q: Is there an erratum? If so, please provide a link or other access point.

A: kevin.ku@toronto.ca

Q: Will the dataset be updated (e.g., to correct labelling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

A: According to Open Data Toronto, the dataset is refreshed annually.

Q: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

A: N/A

Q: Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

A: N/A

Q: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

A: N/A

Q: Any other comments?

A: N/A

Appendix 1.1 Testing Set and Training Set Summary

Variable	Mean (S.D) in Training	Mean (S.D) in Test
Estimated_Dollar_Loss	30879.1244(73889.746)	38278.9983(120556.961)
Fire_Alarm_System_Operation	0.4975(0.500)	0.4966(0.500)
Fire_Alarm_System_Presence	0.6084(0.489)	0.5973(0.491)
Sprinkler_System_Operation	0.0672(0.251)	0.0889(0.285)
Sprinkler_System_Presence	0.1193(0.324)	0.1393(0.347)

Appendix 1.2 Summary Statistic for MLR, Testing Set

Coefficients:

	Estimate Std.	Error	t value	Pr (> t)
(Intercept)	59415	7723	7.693	6.05e-14***
Fire_Alarm_System_Operation	9356	16467	0.568	0.57014
Fire_Alarm_System_Presence	-43585	16505	-2.641	0.00849**
Sprinkler_System_Operation	17264	21438	0.805	0.42097
Sprinkler_System_Presence	-9219	18006	-0.512	0.60882

Signif. codes: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Reference

City of Toronto. 2021. *Smoke Alarms*. https://www.toronto.ca/community-people/public-safety-alerts/safety-tips-prevention/safety-equipment-devices/smoke-alarms/.

———. 2022a. *About Fire Incidents*. https://open.toronto.ca/dataset/fire-incidents/.

———. 2022b. *Fire Services*. https://www.toronto.ca/city-government/accountability-operations-customer-service/city-administration/staff-directory-divisions-and-customer-service/fire-services/.

Fire Protection Author. 2020. "Basic Components of Fire Alarm & Detection Systems." https://www.wsfp.com/blog/what-are-the-basic-components-of-fire-alarm-detection-systems/.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Marty Ahrens. 2021. *US Experience with Sprinklers*. https://www.nfpa.org/-/media/files/news-and-research/fire-statistics-and-reports/suppression/ossprinklers.pdf.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto Fire Services. 2020. "Toronto Fire Services 2019 Annual Report."

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

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.