An analysis of Toronto's Fire Incident situation in 2018

A study of potential factors that impact financial cost caused by a fire incident, using data set from Open Data Toronto, 2018

Xiao Bai

27/04/2022

Abstract

In the past, fire incidents has incurred enormous losses and threntened many lives. The nature, charactertistics and response performances of fire department is important when emergency occured. Using the 2018 fire incident data set from Open Data Toronto, I performed a confidence interval and a multiple linear regression to study the potential factors that may cause the variation of financial loss of different fire incidents. The result shows that the we are 95% confident that the true average time that firefighters take to arrive the location of occurrence is between 297.44 and 304.15 second and different sprinkler system operation, building status, fire alarm system operation and arrival time have potential impact on the variation of estimated dollar loss of different fire incidents

Contents

Code and data are available at[1]

[1]https://github.com/XiaoBai-blip/304-Final-paper

1. Introduction

The Great Fire of Toronto of 1904 April, also known as the Cathedral Fire, was the first major fire in the history of Toronto, Ontario, Canada. The fire remains the largest fire ever to have occurred in Toronto. It swept through 20 acres of Toronto's industrial core. By the time firefighters arrived the occurring place, the blaze had destroyed at least 98 buildings. Canadian Bank of Commerce general manager Edmund Walker states that the fire "was merely a halting moment in the prosperity of Toronto." (Bradburn 2020). The exact cause of the fire was never determined, however, it was thought that the fire started in the Currie Neckwear Company building after a stove pipe became red hot. Hundreds of firefighters from other cities were announced to help douse the flames. Since the fire was so massive that it could be seen from all the way in Buffalo, New York. American firefighters were also dispatched to provide help in dealing with fire. The fire was finally put out 9 hours after it had first been noticed with the help from multiple different cities (Nakatani 2022).

This hazard seriously injury people and caused damage and destroy buildings. Even though Downtown Toronto experienced a construction boom during the 1890s and 1900s, and buildings up to six storeys high arose during this period, the fire safety standards had not caught up until the occurring of Great Fire of Toronto (Bradburn 2020). It exposed the city's need for safer building codes and a high-pressure water system. Right after the disaster happened, four years later, the city promoted high-pressure water system. Landlords were encouraged to install sprinkler systems. As a consequency, the Great Fire of Toronto caused severe financial loss. Total estimated losses were $10,000,000 in 1904 dollars. Most businesses had insurance, but even so, some lost tens of thousands of dollars. Five thousand workers lost their jobs, temporarily or permanently (Toronto, n.d.).

Fire hazard presents in all areas of life. According to Reno et al. (2000), civil authorities have recognized that the threat of fire incident not only affects the well-being of individuals, but also negatively influence the welfare and security of the community. Ghassempour (2021) argues that fire disaster is considered as the fourth most common cause of unintentional trauma all over the world, and residential fire incident accounts for a large percentage of fire-related injuries. In addition, even though there are many reasons that can cause a fire-related incident, one of the most major reasons is due to people's lack of consciousness. For example, a stove left burning at the end of the work day can be a cause. If people being more conscious in the use of fire, the number of fire incident cases would reduce. Therefore, it is necessary to provide a basic understanding of the fire incident situation in Toronto, and the OFM's possible methods for fire control.

In this paper, I used 2018 Fire Incident cases collection from Open Data Toronto to study the factors behind that may potentially increase the fire risk. Specifically, in the essence of many people's concerns about the financial cost of a fire event, if people are given an idea of possible total loss due to a fire incident, they will be more careful in fire usage in their daily life, which will then potentially reduce the number of fire-related incident cases happen to some extent. In this case, the main topic of this paper will be more focusing on the potential factors that cause the dollar loss due to the fire. According to Harvey (2020), researchers define total cost of a fire as the directly and indirectly cost of fire plus the cost of equipment used to prevent the spread of that fire. Specifically, according to Zhuang (2017), the total cost of a fire is divided into two parts: expenditure and loss. Expenditure includes indirect (passive fire protection) expenditure and direct (active fire protection) expenditure, and loss refers to the human loss. Using the given dataset, I will investigate if the dollar loss depends on different number of responding firefighters, different number of equipment used as well as a few other factors in interest. To have a general understanding of the data set, I made explanatory data analysis with the data. I classify all potential predictors into numerical variables and categorical variables, and summarized numerical variable for our study into tables. Further, I made some plots to visualize the distribution of categorical variables

Regarding statistical inference and modelling, I used confidence interval and multiple linear regression. The bootstrap will performed to estimated the sampling distribution about a given population. The result shows that we are 95% confident that true true average time that firefighters take to arrive an occurrence location is between 297.44 and 304.15 second. As for linear regression, I am interested in what factors affect the financial cost of a fire. To build an accurate model, I apply statistical model selection methods for selecting predictors that are more likely to explain the estimated dollar loss. I assume that factors of arrival time,

different sprinkler system types, building status and fire alarm system type will drive the estimated dollar loss to be higher.

The analysis will be conducted in R (R Core Team 2020), and the package we will use is tidyverse (Wickham et al. 2019). All graphs will be created using function ggplot2 (Wickham 2016). The packages knitr (Friendly et al. 2020) are also used to generate the R markdown report.

2. Data

2.1 Data Sources

The dataset is collected from 'Open Data Toronto' (Gelfand 2020), and it contains information about Toronto's 2018 fire incident case record. Open Data Toronto is a digital data that is initiative by the City of Toronto government and it is made available with the technical characteristics necessary for it to be freely used. The portal provides a variety of tabular datasets relating to the city's services, infraustructions and development. There are in total more than ten different data categories in this portal, and the data that will be used in this paper relates to the information of Toronto's fire incident.

The dataset relates to Toronto's 2018 Fire incident cases and it includes only Fire incidents as defined by the Ontario Fire Marshal. The Fire Marshal is the principal adviser to government on public fire protection policy and fire safety issues. Both the Fire Marshal and Deputy Fire Marshal are statutory positions, appointed by the Lieutenant Governor in Council(Ontario 2022). The Office of the Fire Marshal (OFM) promotes development to reduce the fire cases and minimize the negative impact of fire happen in Ontario. It also provides improvement on fire safety and other public safety hazards on people, property and the environment in Ontario. The OFM is responsible for encouraging fire protection, fire prevention and public safety in Ontario through administering provincial legislation. They works to ensure that all fire departments in Ontario provide the right levels of fire prevention and protection based on the needs and circumstances of the areas they serve and the provisions of the FPPA. Moreover, OFM provides support such as training for firefighters and other fire department personnel, and professional development seminars. This 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. For privacy purposes personal information is not provided and exact address have been aggregated to the nearest major or minor intersection. Besides, the dataset receive a silver quality score and it is refreshed annually. The dataset is characterize under the topic of 'Public safety, Locations and mapping, Community services' and all data are displayed in table form.

2.2 Data Cleaning and Data Overview

The dataset contains 117,536 observation and 47 variables. Within all variables, only three of them are numerical which represent the estimated dollar loss due to a fire, the number of responding apparatus and the number of responding personnel respectively. All categorical variables can be classified into five segments: location of fire incident, occurring time, fire incident type, information about buildings, and how fire department responds to the fire. For the convenience of our study, I filtered out missing values and omitted these missing values so that only meaningful numerical numbers are left. Also, after the data was imported, it is not perfectly clean. I first clean the names using janitor's clean name function, and made all the numbers numeric. To make the data more readable, I created a new data frame which consists only variables in interest. These variables are estimated dollar loss, number of responding apparatus, number of responding personnel, TFS alarm time, TFS arrival time, method of fire control, sprinkler system operation, building status, and fire alarm system operation. I mutate a new variable by subtracting TFS arrival time to TFS alarm time for illustrating how long it takes for fire department to arrive fire occurring location. This new variable was then converted from minute to second for simplicity. Similarly, the variable estimated dollar loss is divided by 10000. Moreover, all digit numbers will display as two significant figures. The detailed description of these variables are shown in Table 1.

Table 1: Description of variables

Variable	Description	Type
Dollar Loss	Estimated dollar loss for each fire incident (in $10,000)	Number
Responding Apparatus	Number of equipment used	Number
Responding Personnel	Number of responding firefighters	Number
Fire Control Method	Different methods of fire control - Extinguished by fire department/automatic system/occupant/self extinguished/unclassified	Character
Building Status	OFM Building status code and description - Normal/under renovation/construction/demolition/abandoned/not applicable/undetermined	Character
Fire Alarm System Operation	OFM Fire Alarm System Operation code and description - Fire alarm system operated/did not operated/not applicable /undetermined	Character
TFS alarm time	Timestamp of when TFS was notified of the incident	Date Time
TFS arrival time	Timestamp of first arriving unit to incident	Date Time
Sprinkler system operation	OFM Sprinkler System Operation code and description - activated/did not activate/non activation/not applicable/undetermined	Character

In addition, to perform the model validation process, I split the given dataset randomly and into two independent sets of data: training and testing datasets. The proportion to split the dataset can be arbitrary, and 80/20 is the most common split proportion. The training dataset will be used to perform all model building and diagnostics until a final model is built. The testing dataset will only then be used to evaluate the performance of the model. To have an overview of our data, Table 2 and 3 shows the basic summary for numerical variables in both training and testing dataset.

Table 2: Numerical summaries of the variables (Training dataset)

Variable	Min	Max	Mean	Standard Deviation
Arrival Time (in second)	24	17871	300.7	203.45
Responding Apparatus	1	175	7.42	7.13
Responding Personnel	1	537	24.48	22.04
Estimated Dollar Loss (in $10,000)	1e-07	5000	3.52	48.63

Table 3: Numerical summaries of the variables (Testing dataset)

Variable	Min	Max	Mean	Standard Deviation
Arrival Time (in second)	26	735	297.98	80.74
Responding Apparatus	1	436	7.51	10.79
Responding Personnel	3	1275	24.8	32.69
Estimated Dollar Loss (in $10,000)	0	1300	3.43	45.53

Table 2 and 3 shows the numerical summary of variables that will be used for building regression model. The dataset is split into training and testing dataset, and the first table shows summaries of variables in training dataset while the second table shows summaries using testing dataset. As can be observed from

this table, the mean of four variable in both training and testing datasets looks similar, even though their maximum values in two dataset have a large difference. The information in these two tables will be applied to check model validation in the following section.

Figure 1: Number of fire incidents according to different categories

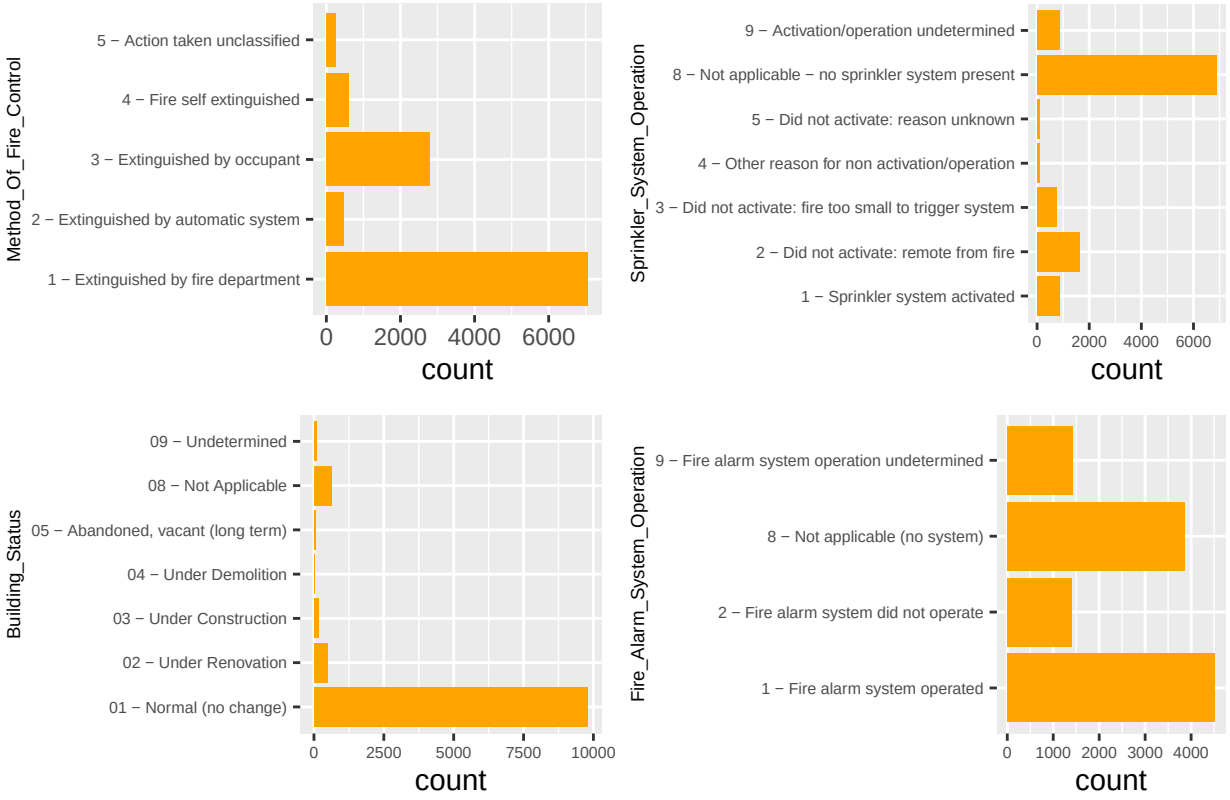


Figure 1 shows the barplots of our categorical variables. The four bar charts represent the number of cases according to different categories. The count is the number of cases and is represented by the horizontal axis. Overall, it can be seen that there are five different methods of fire control, seven sprinkler system operations, seven different building status and four fire alarm system operations in this figure, and each categorical is represented by a specific index for simplification. The method of fire control that the fire was extinguished by fire department accounts for the largest proportion among all five categorical of control methods. Since it may be more costly for fire controlled with the intervention of fire department than fire extinguished without using any fire control method, I assume that this variable would be directly influence the variation of the fire cost. By contrast, except the percentage for action taken unclassified, the number of cases that fire controlled by self extinguished and extinguished by automatic system account for two smallest percentage of data. Moreover, buildings that have not installed sprinkler system and under normal building status have largest proportion of data.

Figure 2: Relationship between Estimated Dollar Loss and the Number of Apparatus/Personnel

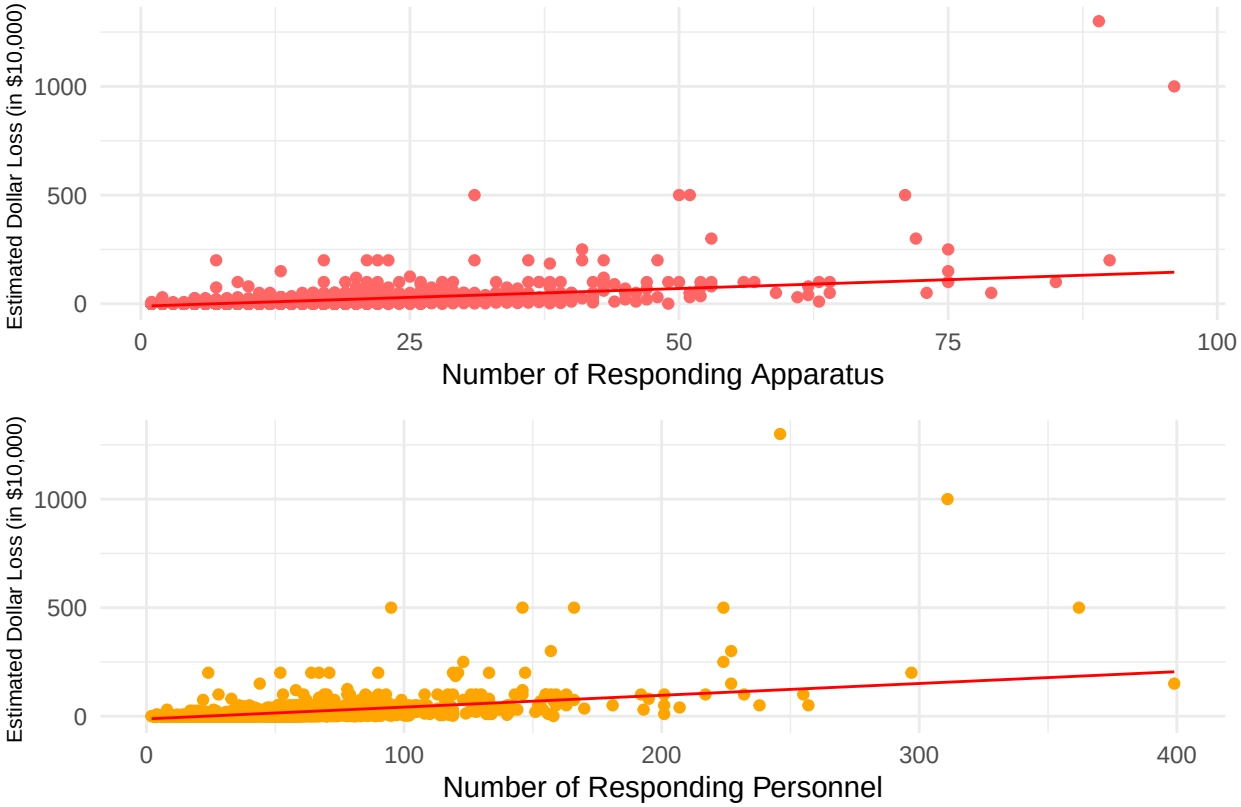


Figure 2 shows relationship between two numerical variables. Each graph represents the relationship with estimated dollar loss by the number of responding apparatus and number of responding personnel respectively. The overall pattern of two plots is in linear form, which means there might exist a positive relationship between two predictors and response variable (an increase in one of the variables is associated with an increase in the other) because the data points make a straight line going from near the origin out to high y-values. Moreover, the strength in two graphs is moderate as most of the points are slightly spread out. In practical aspect, we assume that as the number of responding apparatus or personnel increase by one unit, the estimated dollar loss may also increase, holding other variables unchanged.

3. Method and Model

Statistical methods are applied in this section to help observing and interpreting the data in an alternative way. In this paper, two statistical methods will be applied to help exploring deeply about our data. These methods are simple linear regression model and confidence interval.

3.1 Confidence interval

To analyze our data more deeply, we narrow down our topic to be more focusing on the average percentage of Tanzania communities within all background characteristics that had no education experience. We calculated the estimated mean percentage above using this dataset, which is 25.644. However, since the dataset is just one sample, there might be a problem about how we obtain a measure of precision and confidence about our estimate. Therefore, in order to describes the uncertainty surrounding an estimate, we will perform statistical inference and apply bootstrap method to get the confidence interval in this section.

Bootstrap is a statistical method that is used to estimate the sampling distribution about a given population. It creates multiple resamples (with replacement) from a single set of observations, and then computes the effect size of interest on each of these resamples. This bootstrap resamples of the effect size can then be used to determine the confidence interval. One type of bootstrap is empirical bootstrap, which samples from an estimator's sampling distribution without specifying the data distribution. In this paper, we will use empirical bootstrap. Besides, each confidence interval has a percentage associated with it, called a confidence level. More specificity, if we perform 95% confidence interval, 95% indicates that any such confidence interval will capture the population mean difference 95% of the time. Alternatively, it means that when repeating an experiment or survey over and over again, 95 percent of the time the results will match the results we get from a population. Moreover, with a 95 percent confidence interval, we have a 5 percent chance of being wrong. In addition, for a given dataset, increasing the confidence level of a confidence interval will only result in larger intervals (or at least not smaller). With the small sample, we expect to see that the 95% confidence interval is similar to the range of the data. But only a tiny fraction of the values in the large sample lie within the confidence interval. This is because the 95% confidence interval defines a range of values that you can be 95% certain contains the population mean. With large samples, we know that mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample. In our dataset, since the sample size is too small (n=43) and I think a wider confidence level might give an accurate result than a narrower one, I will use a relatively wider confidence level (95%) in this report. Before applying the bootstrap, there are some assumptions that need to be concerned. We assume that all samples are independent, and the parameter will be the true mean of percentage of people had no education experience.

3.2 linear Regression Model

The purpose is to investigate the "best" linear regression model that predict the estimated dollar loss caused by a fire, including the number of responding apparatus, responding personnel, arrival time, uesed fire control method, sprinkler system operation, building status and fire alarm system. We will build full model using these several predictors at the beginning. Then we will select only few predictors that can explain response variable the most by applying model selection methods for instance by checking assumption, detecting multicollinearity, and applying automatic model selection. The detailed description will be illustrated in the following few subsections.

3.2.1 Model Setup

The multiple linear regression (MLR) is used to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. The general form for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Where β_i represents the coefficients need to estimated, $x_i, i = 1...p$ is the predictor variables, y is the response variable, ϵ is the error term. Mathematically, β_0 measures where the line intercept y-axis, and $\beta_1 - \beta_p$ measures the slope of the line. More practically, β_o is the value of y when x equals the zero, and $\beta_1 - \beta_p$ is the average change of y when x increase by 1. I will first build the full model using seven predictors that I chose at the beginning, then I will apply the model selection methods to select only few predictors that can explain our response variable the most.

3.2.2 Model Selection

Check linear Regression assumptions

Regression violation testing is based on the statistical theory about assumptions of linear regression model. The linear regression model has four assumptions: linearity, uncorrelated errors, constant variance and

normality. When deriving the unbiasedness and the covariance of our estimator, the assumptions may need to be used many times to obtain results. When all the model assumptions are satisfied, we can then be sure that the estimators will behave in a nice way and have all these lovely properties. However, if even one assumption is violated, this can have a large impact on how we can use our estimates.

Residual plots can be used to determine whether there are violations of model assumptions. Residual plots allow us to visually inspect the model assumptions. Moreover, we work with residual plots because the data can sometimes be too noisy to see model violations clearly. There are three main types of residual scatter plots that we use: residuals versus predictor plots, residuals versus fitted values plots, and normal QQ plots. Both residual versus predictors and residual versus fitted value plots can be used to assess whether the first three assumptions hold. We can check by observing from the residuals plot and if there is no discernible pattern seen in the residual's plots, then the assumptions hold. In other words, to satisfy the assumptions, residual verses fitted value and predictors plots should not have any pattern or large clusters of residuals. If the model does not violate any of these assumptions, we will consider current model as full model and proceed to the next step of model selection. By contrast, if our model does not satisfy these assumptions, we will apply Box-cox transformation to correct it. The transformed model will be the full model if violations are fixed, otherwise, we will record the violation. Once we have decided the full model, we proceed to the next step for model selection.

Check multicollinearity

A model may provide contradictory information without noticing that this model required a transformation on the response or predictors to correct model violations. In fact, one such situation would be if the predictors are too strongly correlated with one another. This issue of multicollinearity and can result in a number of problems with the model. For example, coefficients may have the wrong sign, compared to existing knowledge of the relationship and many predictors may be non-significant individually, but the overall F-test is highly significant.Thus, there is needed to discover all the variables are highly correlated and solve the multicollinearity issue of the model. One tool that can be used to detect possible multicollinearity in the predictors that both takes into account the conditional nature of regression and the possibility multiple predictors are correlated to each other is VIF. We conduct by calculating the VIF and we remove the predictors that have a high VIF. In general, a VIF larger than 5 should be removed, and the process should be conducted multiple times until none of predictors have a VIF that is larger than 5.

Apply two model selection methods

After detecting multicollinearity, I will use remaining predictors and perform both automated selection and manual selection methods to build two models. For automated selection, I will use stepwise selection based on AIC. This method is a combination of forward selection and backward elimination, testing at each step for variables to be included or excluded. It will automatically add or remove predictors until AIC will not increase in the next step. However, the result might not be trustable as it can be used even when model has violations. Therefore, it is needed to make sure that the model satisfies the assumptions first before applying AIC. For manual selection, I will choose appropriate predictors based on how significance they are to the response variable. I will manually remove the predictors with high p-values from full model, and build a new model using remaining few predictors. The model that was built using automatic selection will be the candidate model 1 while the model with manual selection will be candidate model 2.

After we built two candidate models, we will compare these models according to their AIC, BIC and adjusted R^2 to decide the most likely appropriate one as our preferred model. In general, the one with smaller AIC and BIC value, and higher adjusted R^2 is better. We also compare the assumption violations of each model using residual and QQ plots.

Problematic observations

Once we have decided preferred model, we investigate problematic observations by checking through leverage point, outlier, and influential point. Leverage point can be verified using leverage formula while influential point can be checked using Cook's Distance, DFFITS, and DFBETA. If there are no contextual reasons to remove those problematic observations, we will test data validation using testing dataset. We will build the same model using 20% testing dataset and then compare its adjusted R^2, predictor significance, and value of coefficient with the result shown in training dataset. If the difference of adjusted R^2, predictor significance, coefficient is not significant, then this model is likely validated. If this model is not validated, we will go back and test the validation of candidate 2 model.

4 Results

4.1 Confidence interval

The graph is the result of bootstrap confidence interval for mean time that firefighters take to arrive an fire occurrence location. Values between the 2 red lines are in the 95% interval. We rounded our result to three significant digits (refer to the Table 4). We are 95% confident that true mean is between 297.44 and 304.15. The confidence interval is meaningful because both number (297.44 and 304.15) is close to and bounded around the sample mean we calculated above. Specifically, we are 95% sure that the true average time that firefighters take to arrive an occurrence location is between 297.44 and 304.15 second.

Figure 3: Histogram of boostrapped mean

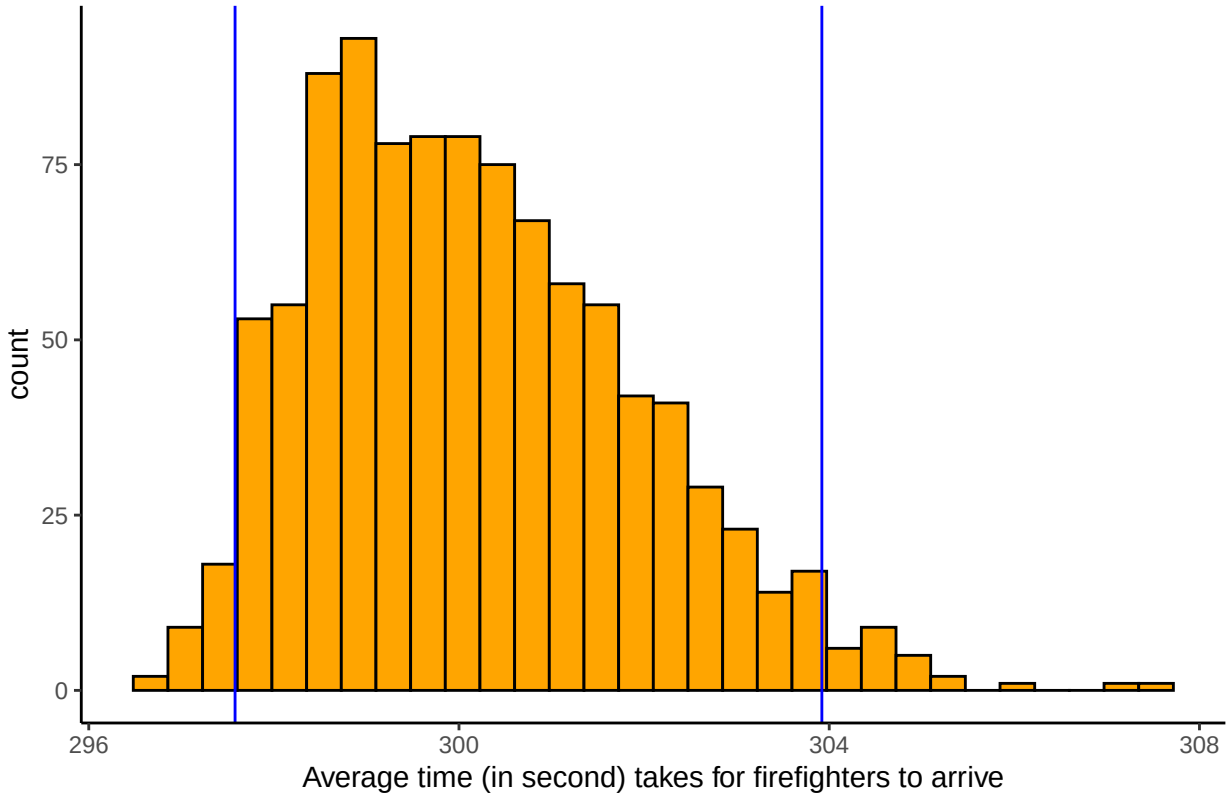


Table 4: Confidence interval Result

2.5%	97.5%	CI
297.44	304.15	(297.44, 304.15)

4.2 Linear Regression Model

4.2.1 Starting model

I build starting model using seven predictors that I am interested in. These predictors are: the number of responding apparatus, the number of responding personnel, arrival time, method of fire control, sprinkler system operation, building status, and fire alarm system operation. Our response variables is: estimated dollar loss. The estimated model will present in the form of multiple linear regression:

$$\hat{Estimated\ dollar\ loss} = \beta_0 + \beta_1 \hat{Arrival\ time} + \beta_2 \hat{Sprinkler\ system\ operation} + \beta_3 \hat{Building\ status}$$

$$+\beta_4 \hat{Fire\ alarm\ system\ operation} + \beta_5 \hat{Number\ of\ apparatus}$$

$$+\beta_6 \hat{Number\ of\ personnel} + \beta_7 \hat{Method\ of\ fire\ control}$$

4.2.2 Model Selection Result

The model satisfies both condition 1 and 2. However, since there exist a large cluster in residual-fitted plot, the model may violate independence assumption. The normal QQ-plot shows that the normality assumption is quite satisfied as most of points are close to the diagonal. We apply Box-cox transformation to let it automatically generate appropriate transformed value for our variables, and we take log of response variable (estimated dollar loss), and log of other three numerical predictors in transformed model. By constructing residual plots and QQ-plot again on transformed model, there is a significant improvement on model assumption of independence (Figure 3-5). Therefore, in the following sections, we will consider transformed model as full model and perform model selection process to choose the most appropriate combination of predictors to our model.

Figure 3: Residuals VS. Fitted

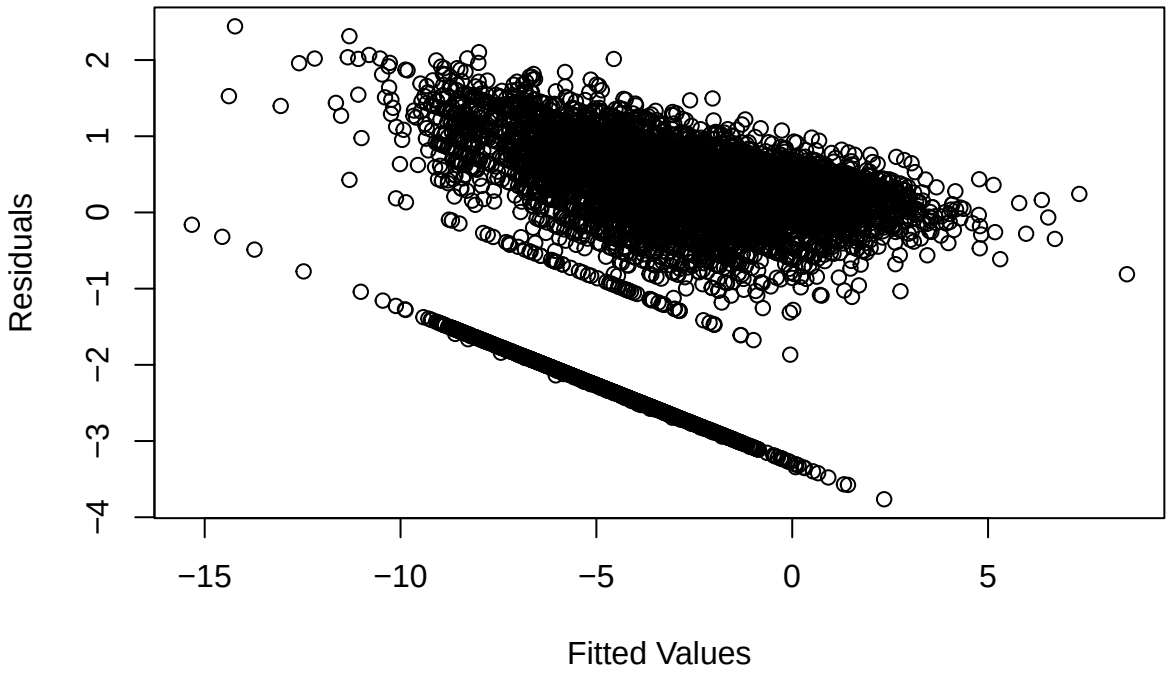


Figure 4: Residual VS. Predictors

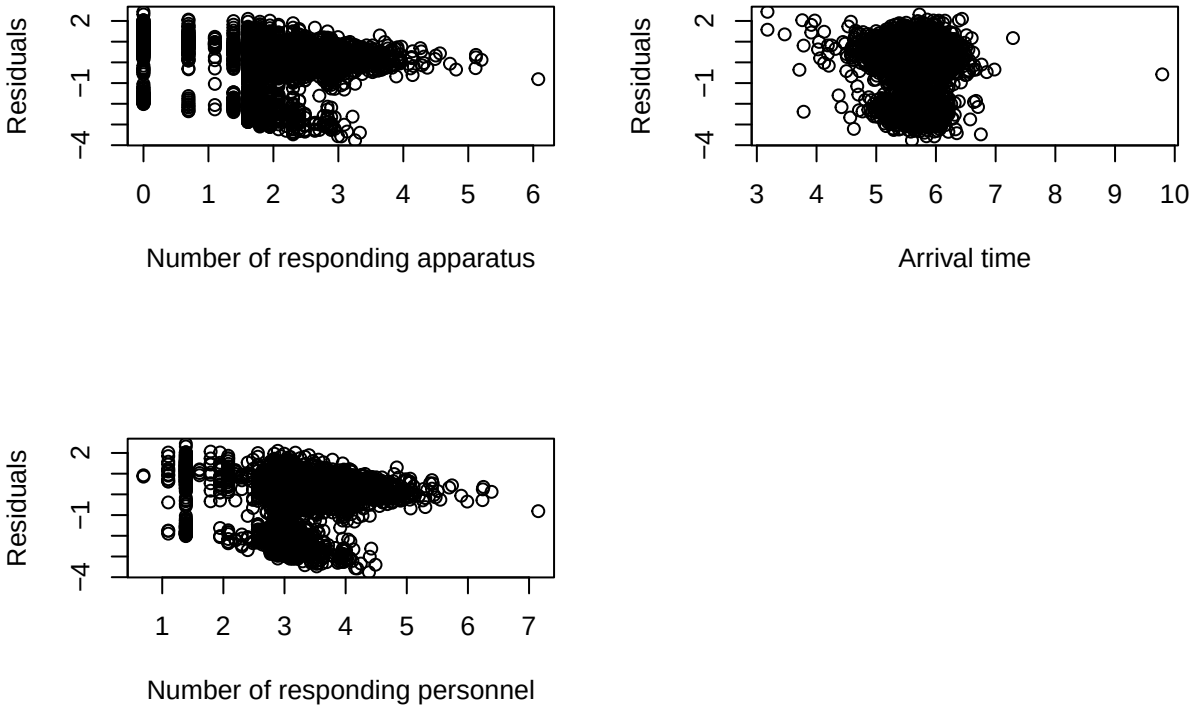
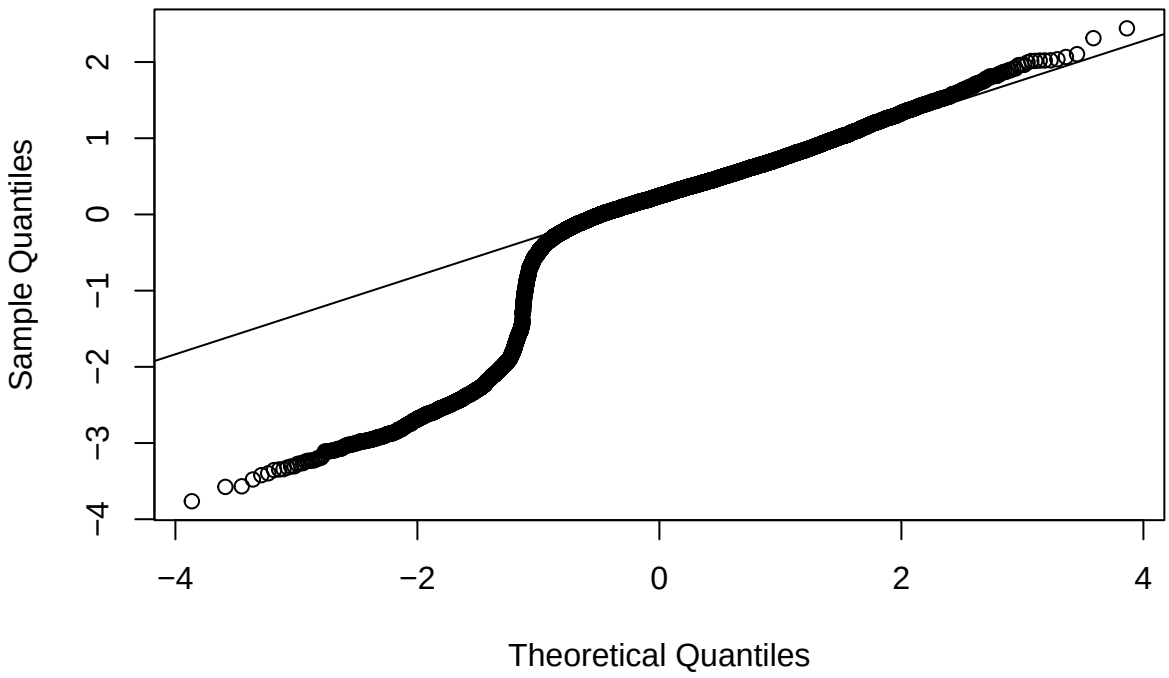


Figure 5: Normal QQ Plot



Since the predictors "number of apparatus," "number of personnel" and "method of fire control" have VIF greater than 5 (Table 5), we remove those predictors. We will use the remaining predictors to build a model as full model. In addition, the stepwide model selection producded the same full model, which means adding or removing any variables will increase AIC in the full model. Then we compare the full model and manual selected model that consist less predictors. Predictors in manual selected model are chosen based on their p-values. The predictors "arrival" and "Sprinkler System Operation" have p-value less than 0.05, so we build the model that only contains these two predictors. The model that was built using automatic selection will be the candidate model 1 while the model with manual selection will be candidate model 2. Now, since two candidate models were constructed, we move to next step of choosing the best one as final model.

Table 5: Multicollinearity (VIF) Result

Variable	GVIF	DF
Number of responding apparatus	80.95	1
Number of responding personnel	80.25	1
Arrival time	1.02	1
Method Of Fire Control	5.22	4
Sprinkler System Operation	2.81	6
Building Status	1.19	6
Fire Alarm System Operation	1.62	3

Determining a better model through examining model assumptions may not work here as residual and QQ plot in both two models look similar, that is, both models seem have worse violation in linearity and normality. It is because there is a significant linear pattern in their residual plots and points are not close to diagonal in QQ plot (see Appendix 1-4). However, candidate 1 model can still be the preferred model as it has a

smaller adjusted R^2 and larger AIC/BIC (Table 6). There are 567 leverage point and no contextual reasons to remove them. In addition, there are no outlier or influential point in the training dataset. Therefore, the final preferred model will be the candidate 1 model.

Table 6: Model Selection Criteria

Model	Adjusted R^2	AIC	BIC
Auto Selected Model	0.09	55511.72	55639.55
Manual Selected Model	0.08	55629.42	55693.33

To test the validation of Candidate 1 model, Table 1 and 2 in Data section shows that the mean of predictors and response variable in both training and testing dataset are similar. However, in Table 6, although the adjusted R^2 are similar and the significant predictors are the same, the coefficients have big difference between two models. Moreover, we have checked the assumption violation for model using training dataset and it shows that model may violate normality. However, the model fitted by using testing dataset, it seems not violate normality assumption. Therefore, this model is not likely validated. Similarly, Candidate 2 model is also not validated by examining residual and QQ plots. Thus, we will record this limitation in discussion section and still consider candidate model 1 as our final model.

The final model with four predictors is:

$$log(Estimated\ dollar\ loss) = \beta_0 + \beta_1 log(Arrival\ time) + \beta_2 log(Sprinkler\ system\ operation)$$

$$+\beta_3 log(Building\ status) + \beta_4 log(Fire\ alarm\ system\ operation)$$

P-values and coefficients in regression analysis work together to be used for showing which relationships in the model are statistically significant and the nature of those relationships. The p-values for the coefficients indicates whether these relationships are statistically significant. The sign of a regression coefficient can tell us whether there is a positive or negative correlation between each independent variable and the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase.

Our regression output can be seen in Table 7, where all the coefficients of our predictors in both training and testing datasets are listed. The coefficients describe the mathematical relationship between each independent variable and the dependent variable. From the previous analysis, we found that dollar loss is related to firefighter's arrival time, sprinkler system operation, building system and fire alarm system. If one unit increase in log of arrival time, we expect 1.08 increase in log of dollar loss, holding other variables constant. As for variable building status, the estimated dollar loss for building status under renovation (02) is 1.21 higher than building status that is normal (01). Similar interpretation can be applied to other variables. These findings explain the research question of how these four predictors influence dollar loss caused by a fire.

Table 7: Parameter Estimates for Final Model (Response: log(Estimated dollar loss))

Variable	Coefficient Estimate (Training)	Standard Error (Training)	Coefficient Estimate (Testing)	Standard Error(Testing)
(Intercept)	-9.80	1.12	-4.42	1.00
arrival time	1.08	0.20	0.50	0.17
Sprinkler System Operation 2	-0.50	0.25	-1.11	0.22
Sprinkler System Operation 3	-3.26	0.30	-1.27	0.27
Sprinkler System Operation 4	1.25	0.58	-1.03	0.61
Sprinkler System Operation 5	0.36	0.59	-0.75	0.68
Sprinkler System Operation 8	1.65	0.23	-0.49	0.20
Sprinkler System Operation 9	-0.20	0.29	-1.33	0.26
Building Status 02	1.21	0.28	0.75	0.23
Building Status 03	0.17	0.48	-0.51	0.49
Building Status 04	-6.04	1.48	-4.14	1.25
Building Status 05	-0.60	0.83	-1.25	0.73
Building Status 08	-1.70	0.25	-0.64	0.23
Building Status 09	0.93	0.63	1.08	0.73
Fire Alarm System Operation 2	-1.03	0.19	-0.98	0.16
Fire Alarm System Operation 8	0.49	0.15	0.12	0.13
Fire Alarm System Operation 9	0.16	0.19	-0.11	0.17

5. Discussion

5.1 Findings

Toronto had a significant amount of fire incidents in 2018, reaching about 11,214 cases in total throughout the whole year. The potential reason behind is that people's lack of consciousness causes unexpected kitchen fires outbreak. Accordingly, if people are given basic understanding of how severe financial loss is caused by a fire, it can enhance people from being more careful about fire usage in their daily life.

In this paper, I analyze the potential factors behind that can influence the financial cost of a fire using the dataset relates to Toronto Fire Incident in 2018. I analyze by using statistical inference and modelling such as confidence interval and multiple linear regression. I discovered average time it takes for firefighters to arrive the fire occurring location through simulation and calculate confidence interval. The result shows that we are 95% confident that true true average time for firefighters to arrive an occurring location is between 297.44 and 304.15 seconds. As for factors affect the financial cost of a fire, I use several predictors to construct linear regression model, and apply several statistical model selection methods to select the "best" linear regression model that can predict the estimated dollar loss caused by a fire. I discover that four predictors: arrival time, different sprinkler system types, building status and fire alarm system type will drive the estimated

dollar loss to be higher. The finding shows that it is crucial important for a building to have an advanced fire safety equipment as it directly influence the financial loss caused by fire

5.2 Limitation and Next steps

The final model is barely validated as the results are different when building model using training dataset and testing dataset. For instance, normal QQ plot for training dataset and testing dataset are different, as one plot shows violation in normality and the other one does not. Moreover, in general, we can expect some differences between the coefficient of predictors in two model, but the differences should not be much bigger than the standard error of each coefficient. However, in this case, the coefficient of predictors have big difference between two models, and some coefficients even have a totally different signs (e.g.Sprinkler System Operation 4).

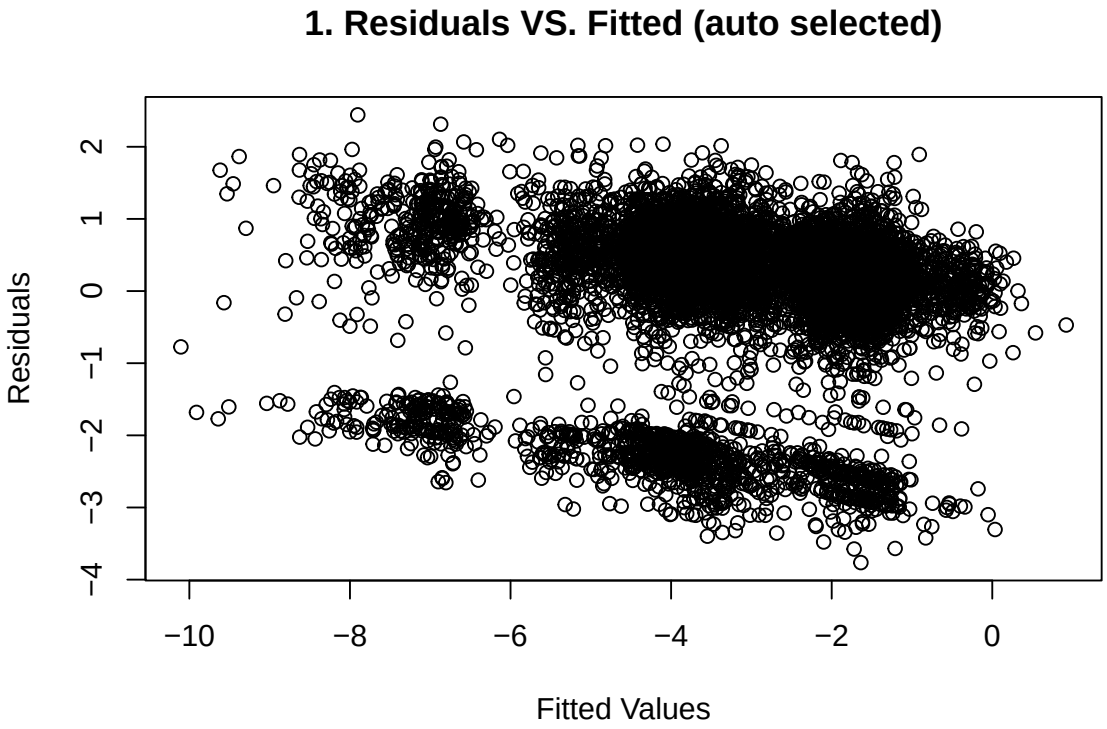
The first potential reason is that the test dataset is very different from the training dataset as we can see the minimum and maximum value of testing and training dataset is quite different even though they have a similar mean value (Table 1-2). The second reason is that there are a lot of influential observations in one or both of the datasets. This may not be true as in this data, we have checked that there is no outlier or influential point. However, there are 567 leverage points in our data, which means that we have lots of data points that have x-values with an unusually large effect on the estimated regression model. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values. Since outliers and high leverage data points have the potential to be influential, we can assume that large cluster of leverage points can be considered as large influential points even though we generally have to investigate further to determine whether or not they are actually influential. In addition, as we know that influential observations can impact how the regression surface is estimated. This can result in estimated coefficients being different depending on whether the observations are used to estimate the model or not. According to the result of predictor coefficients that are listed above, we indeed see that some coefficients have a totally different signs between train dataset and test dataset. This evidence shows that coefficients might be impacted by influential points. Therefore, this can be one possible reason for model not being validated. The third reason might be our testing dataset is too small so there is too much variation present. In general, the more data used to train model, the better the predictive power, and the more data that had been hold-out to later test the model, the better the performance estimation. Since 80% of data are chosen as training dataset, we may see too much variation in model built using testing dataset. The fourth reason might be the Box-cox transformation applied to correct model assumptions were too specific to the training dataset and don't help in the test dataset. Since we used a slightly complicated transformation, it may tailor the model too closely to the training data. This complicated transformation may work really badly in the test models even though transformation works really well in the training data models.

On the other hand, we use residual plots for full model at the beginning of model selection process, and we apply Box-cox transformation to correct model assumption as we see the starting full model violate independence assumption. We take log of response variable (estimated dollar loss), and log of other three numerical predictors in transformed model and we see that there is a significant improvement on model assumption of independence after the transformation. However, we can only say that model violations were improved but it is not good enough to say that these violations have not been fully corrected. This is because we can still observe that our model does not satisfy normality assumption in residual QQ-plot. In addition, even if we observed that influential points affect the accuracy of our result, we can not remove these observations as we do not have contextual reason to remove them.

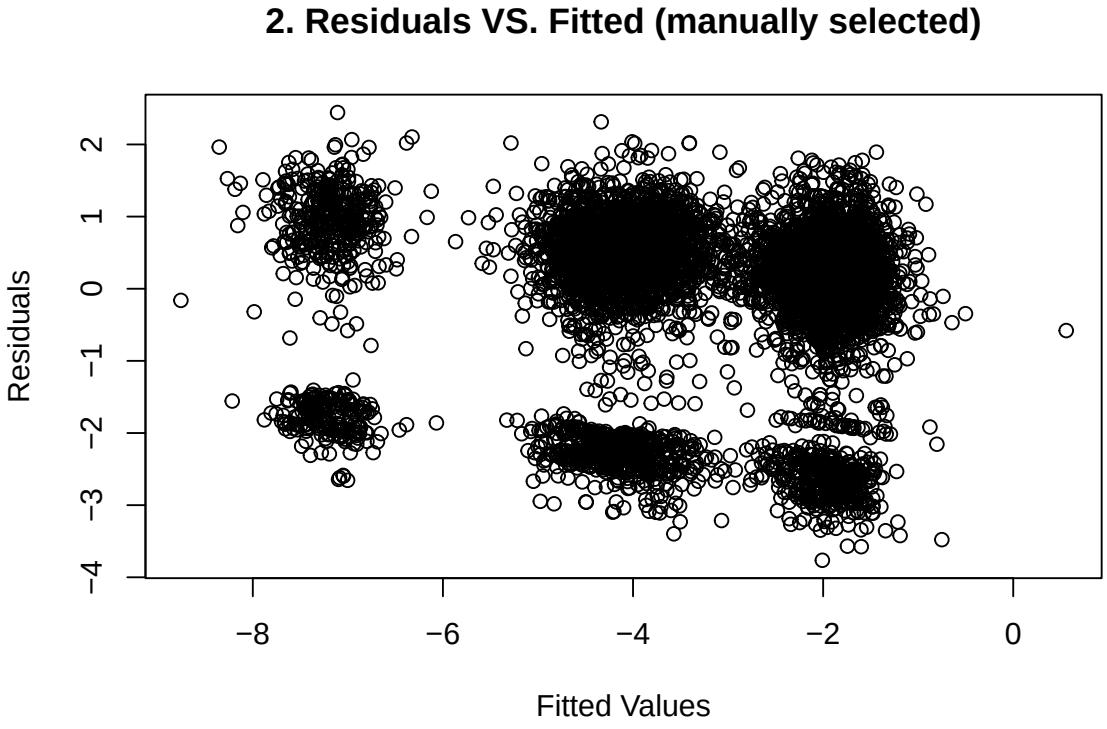
In the model selection section, we often attempt to compare models with various combinations of predictors to each other. Nevertheless, too many predictors is considered over-fitting (i.e. the model is only good on the data used to build it), but too few is considered under-fitting. Our final model may have under fitting problem as we only include four predictors in the final model. That means the model is unable to capture the relationship between the all predictors and the target value, which is estimated dollar loss. Therefore, it is hard to create a perfect model to demonstrate our analysis and the linear model we fit should be expanded with more variables.

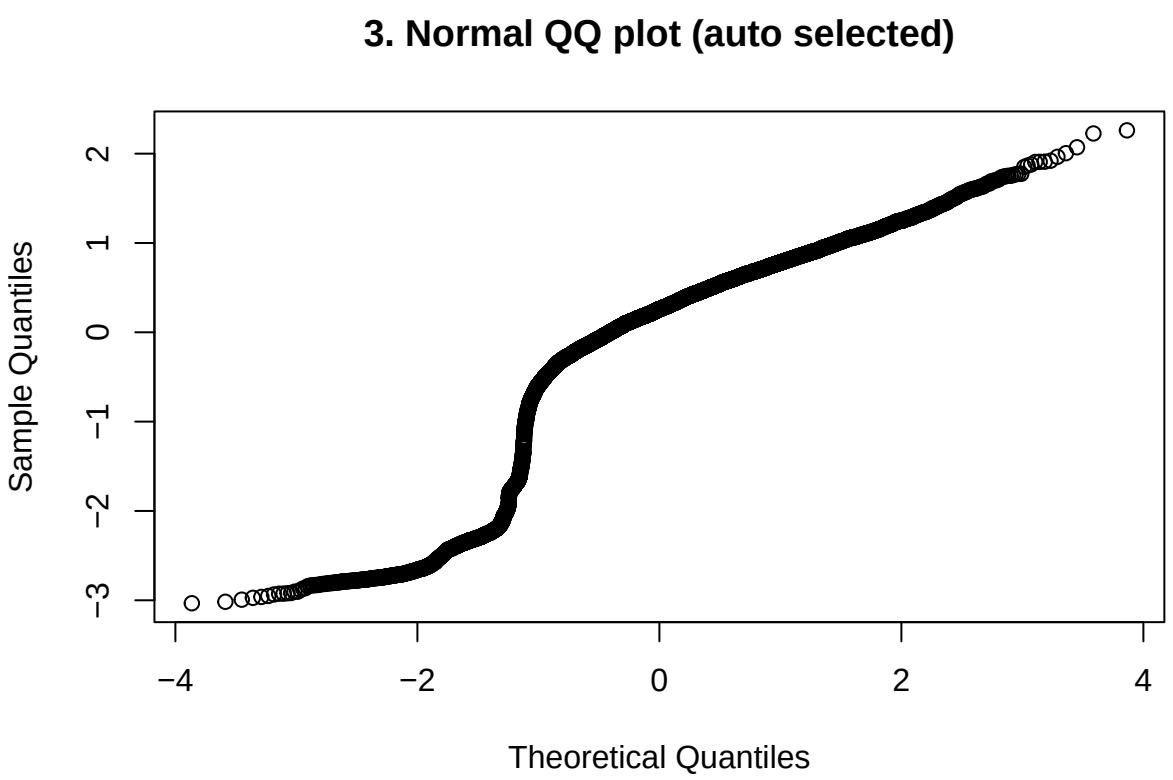
Regarding the dataset itself, since the dataset is not collected by myself, we are limited information to only those provided by fire department. The data was downloaded from the website, but the description of datset is not provided with enough information. For instance, how the data was collected and the methodology used to collect was not mentioned. Also, even though there are about 17,536 observations in this dataset, some parts of information are missing. There are only 11,520 left after we filter out all missing values. Furthermore, some observations contain meaningless information. For instance, it is less likely that only one personnel was used in a fire incident, but many fire incidents have only one recorded in the dataset. Our knowledge of statistical modelling is limited, and there should be more possible studies we can do to fully use this dataset. There should be lots more information to be provided together with this dataset.

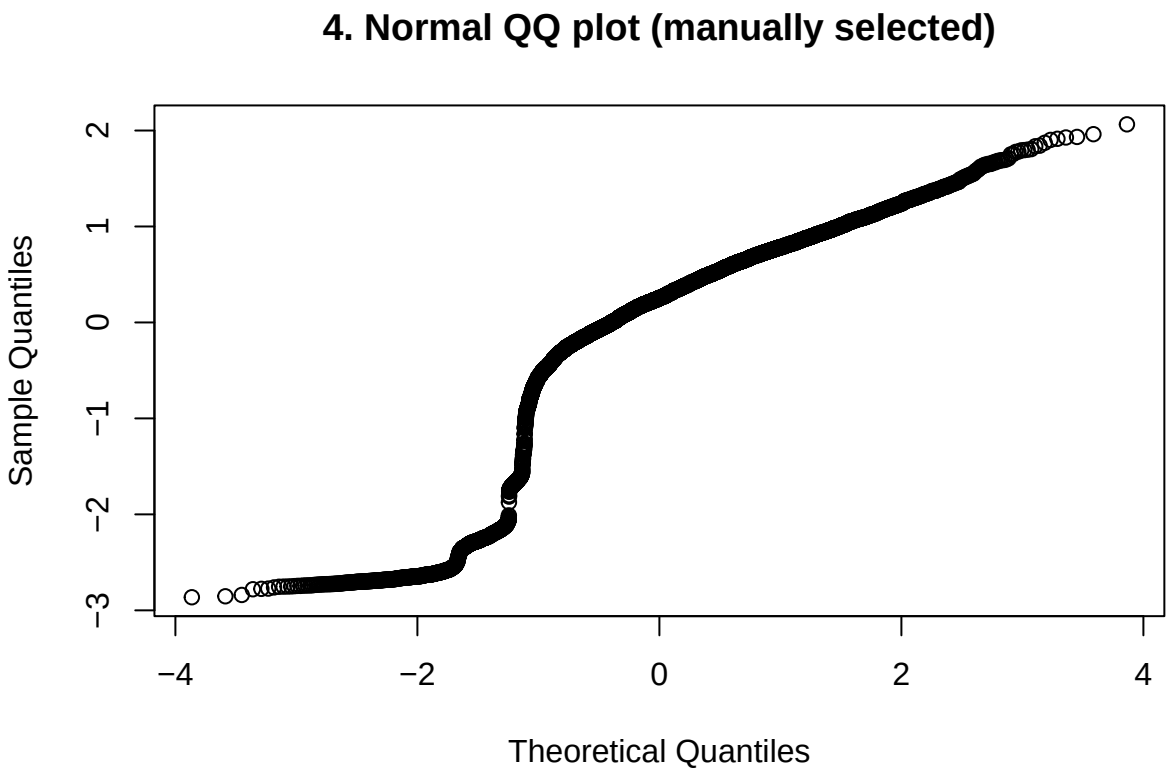
6. Appendix


1. Residuals VS. Fitted (auto selected)
Residuals
Fitted Values
2
1
0
−1
−2
−3
−4
−10
−8
−6
−4
−2
0

1.


2. Residuals VS. Fitted (manually selected)
Residuals
Fitted Values
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0


3. Normal QQ plot (auto selected)
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
−4
−2
0
2
4



2. Shiny application is available at "output" folder in https://github.com/XiaoBai-blip/304-Final-paper/tree/main/outputs

3. Extract of the questions from Gebru (2021)

Motivation

1. *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.*
   - The dataset was created to record the fire incident condition in Toronto as it generally record the number of incidents happened each year. The dataset contains enough variable for analyzing, however, the description for each variable is not properly provided or lack of explanation. There are lots of missing value in this dataset.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by City of Toronto Community and is collected under City of Toronto's Open Data Portal.
3. *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.*
   - The dataset was funded by open data community, which includes developers, policymakers, academics and civic advocates.
4. *Any other comments?*

- These oen data communities mobilize around a common goal: that some data should be freely available for everyone to use and re-publish as they wish.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.* -Dollar Loss: Estimated dollar loss for each fire incident (in $10,000) -Responding Apparatus: Number of equipment used -Responding Personnel: Number of responding firefighters -Fire Control Method: Different methods of fire control -Building Status: OFM Building status code and description -Fire Alarm System Operation: OFM Fire Alarm System Operation code and description -TFS alarm time:Timestamp of when TFS was notified of the incident -TFS arrival time: Timestamp of first arriving unit to incident -Sprinkler system operation: OFM Sprinkler System Operation code and description

2. *How many instances are there in total (of each type, if appropriate)?* -The dataset contains 117,536 observation and 47 variables. Within all variables, only three of them are numerical which represent the estimated dollar loss due to a fire, the number of responding apparatus and the number of responding personnel respectively. All categorical variables can be classified into five segments: location of fire incident, occurring time, fire incident type, information about buildings, and how fire department responds to the fire.

3. *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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset contain all possible instances from a larger set. It was collected directly from City of Toronto Community which record the information relates to fire incidents in Toronto throughout the whole year.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - This data mostly contain data frames with characters summarized. The data are mostly characteristics. There are 47 variables in total, only three of them are numerical which represent the estimated dollar loss due to a fire, the number of responding apparatus and the number of responding personnel respectively. The rest variables are categorical representing fire type and building status.

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

    - Yes. The whole dataset describes the information about each fire incident such as occurring time and occurring location. Columns are divided according to different categories (eg: building status, fire control method, number of responding personnel, etc. . .).

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

    - Yes. In the raw data, they indeed had some missing values in some variables. However, we removed those missing values in the data cleaning process, and the cleaned data now should contain values for most of the observations. The information is missing primary because fire department did not record those information or those information are undetermined.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- Each fire incident occurring location is identifies by a specific id, and they are group according to different regions. In addition, the exact location (city, ctreet) are provided.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - Yes. The dataset was split randomly into two independent sets of data: training and testing datasets. The proportion to split the dataset can be arbitrary, and 80/20 is the most common split proportion. The training dataset will be used to perform all model building and diagnostics until a final model is built. The testing dataset will only then be used to evaluate the performance of the model.

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

    - Yes. Some observations contain meaningless information in the dataset. For instance, it is less likely that only one personnel was used in a fire incident, but many fire incidents have only one recorded in the dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, 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.*

    - No. The dataset was collected directly by Open toronto community and send to OFM fire department annually.

11. *Does the dataset contain data that might be considered confidential (for example, 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.*

    - NO. confidential information has already been hiden when the information is provided. The dataset retrived contain no confidential information.

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

    - Yes. After performing EDA to our dataset, I found that the proportion of dataset that have no sprinkler system present is extremely larger than other five categories. It will be difficult to perform statistically inference and make comparison with other variables that have similar characteristics.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No. The dataset contains only information about Toronto fire incident in 2018. It records number of cases in the exact location and exact year.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - NO, we can not. Confidential information are hiden in the report.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, 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.*

    - NO. By reviewing the website that provides this dataset, we can see that some sensitive information, and ethics and income are asked, but the raw data is hiden from the report, and our study is based on the dataframe provided in the report.

16. *Any other comments?*

    - NO.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, 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.*

    - This data is collected by Open Toronto data community. It records the each fire incident happened throughout the whole year in Toronto and send directly to OFM, and summarized by characteristics. The information relates to validation is not provided.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The absolute probability of selecting an EA (product of the probability of selecting a ward/branch and the conditional probability of selecting an EA within a ward/branch

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset is not a sample from a larger set.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data was collected by open data community, which involves developers, policymakers, academics and civic advocates to ensure that some data should be freely available for everyone to use and re-publish as they wish.

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

    - The data is collected in 2018. It matches the timeframe of the data associated with the instances.

6. *Were any ethical review processes conducted (for example, 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.*

    - Ethical review processes conducted is undetermined.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was provided under the website 'Open Data Toronto,' and it was collected directly from recording the number of fire incidents in Toronto in 2018.

8. *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.*
    - The website emphasize that when started designing the new portal, they spoke to the individuals best informed to direct how it should function.
9. *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.*
    - The City complies with the freedom of information and privacy laws and will only be releasing public information layers. Every data set goes through a privacy lens before publishing to ensure it has met MFIPPA laws.
10. *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).*
    - NO, there is not.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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.* -The programme has the potential of doubling the current national contraceptive prevalence rate.
12. *Any other comments?*
    - NO

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, 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.*
    - Yes, the dataset provided under website was raw data, so I have done lots of cleaning process for making the data readable. For example, we remove all the meaningless characteristics and change the name of some variables. We also used some packages such as tidyverse for analyzing our data. We also remove all the missing values and mutate new variables.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes, raw data can be found under input folder of my Github. It can be found at this link: https://github.com/XiaoBai-blip/304-Final-paper/tree/main/inputs/data.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - We only use packages from R to help preprocessing the data such as read.csv.
4. *Any other comments?*
    - No

Uses

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

- Yes, the dataset must had been used by other people or organization. Since datasets are made freely available for anyone under the City's Open Government License, which is based on version 1.0 of the Open Government Licence – Toronto. This licence allows worldwide, royalty-free, perpetual, non-exclusive use of the City's open datasets, for both commercial and non-commercial use.

2. *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.*
   - Code and data are available at: https://github.com/XiaoBai-blip/304-Final-paper
3. *What (other) tasks could the dataset be used for?*
   - The dataset could be used to analyze fire incident condition in Toronto. It can also be used to idnetify potential factors behind that can increase the risk of fire and financial loss caused by a fire.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact 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 (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - Yes. Even though there are about 17,536 observations in this dataset, some parts of information are missing. There are only 11,520 left after we filter out all missing values. Some consumer should pay attention on these missing values as it may distract the futher analysis.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No.
6. *Any other comments?*
   - No. **Distribution**
7. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - Yes. Everyone can use this dataset for any purpose. Datasets are made freely available for anyone under the City's Open Government License, which is based on version 1.0 of the Open Government Licence – Toronto. This licence allows worldwide, royalty-free, perpetual, non-exclusive use of the City's open datasets, for both commercial and non-commercial use.
8. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The data can be find by searching on the open data portal. Consumer can filter their search further by topic, division, and other attributes in the catalogue's left sidebar. Consume can download the data directly under the website as csv, xlsm form. In some cases, consumer may need to submit an FOI request for more specific data or historical data that may not be available through our open data catalogue.
9. *When will the dataset be distributed?*
   - The dataset was distributed after 2018, and it was renew annually.
10. *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.*

- Yes. This website and the materials and images appearing within it are protected by Canadian copyright law. Except as otherwise provided for under Canadian copyright law, such website, materials and images may not be copied, published, distributed, downloaded or otherwise stored in a retrieval system, transmitted or converted, in any form or by any means, electronic or otherwise, without the prior written permission of the copyright owner.

11. *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.*
    - Yes, by Government, Statistics Act.
12. *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.*
    - NO
13. *Any other comments?*
    - NO

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset is supported by The Open Data team. City partners provides support through the portal development process. In particular, key team members from Common Components, Digital Communications, Technical Infrastructure Services, Digital Technology Services, and Strategic Communications provides partnership and contributions.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Website page: https://open.toronto.ca/
    - Email Adress: opendata@toronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - This given dataset is conducted in 2018, and the report is set not to be final. The dataset will be renew annually up to date. Questions are accepted: opendata@toronto.ca.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, 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.*
    - No
6. *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.*
    - The dataset will be renewed annually by Open Data team. The data and errors can be reported using email.
7. *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.*

- Publishing or contribute to the dataset request should be done by sending an email, and someone from the Open Data team will reach out with assistance within 2 business days.

8. *Any other comments?*
    - NO

7. Reference

Bradburn, Jamie. 2020. "Great Fire of Toronto (1904)." https://www.thecanadianencyclopedia.ca/en/article/great-fire-of-toronto-1904.

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean 'Lahman' Baseball Database*. https://CRAN.R-project.org/package=Lahman.

Gebru, Jamie Morgenstern, Timnit. 2021. "Datasheets for Datasets." Communications of the ACM."

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Ghassempour, N. 2021. "Estimating the Total. Number of Residential Fire-Related Incidents and Underreported Residential Fire Incidents in New South Wales, Australia by Using Linked Administrative Data."

Harvey, Tannous, L. 2020. "Health Impacts and. Economic Costs of Residential Fires (RESFIRES Study): Protocol for a Population-Based Cohort Study Using Linked Administrative Data." *BMJ*.

Nakatani, Mina. 2022. "THE TRAGIC STORY OF THE GREAT FIRE OF TORONTO." https://www.grunge.com/823163/the-tragic-story-of-the-great-fire-of-toronto/.

Ontario. 2022. "The Office of the Fire Marshal." https://www.ontario.ca/page/office-fire-marshal.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reno, J., Marcus D., Leary M., and Samuels J. 2000. "Fire and Arson Scene Evidence. A Guide. For Public. Safety Personnel."

Toronto, City of. n.d. "The Great Fire of 1904." https://www.toronto.ca/explore-enjoy/history-art-culture/online-exhibits/web-exhibits/web-exhibits-significant-events/the-great-fire-of-1904/.

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

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.

Zhuang, Payyappalli, J. 2017. "Total Cost of Fire in the. United States." *2017 Fire Protection Research Foundation*. https://www.nfpa.org/-%20/media/Files/News-and-Research/Fire-statistics-and-%20reports/Executive-%20summaries/RFTotalCostExSummary.ashx.