Research Proposal

Haoluan Chen

Research questions of interest

As our population grow, the larger number of people live and work in bigger buildings all round the world. However, the risk and the cost of indoor fire is also increasing. For my project, I am interested in understanding the relationship between the estimated dollar loss for indoor fire based on fires characteristics, presence of fire prevention systems, time for the Toronto Fire Services to arrive and control the fire.

Main independent variables of interest

My main dependent variable of interest is the estimated cost of indoor fire.

Dependent variable of interest

My dependent variable can be group into three main category: fire characteristics, presence of fire prevention systems and the time for the Toronto Fire Services to arrive and control the fire.

Fire characteristics includes following variables: Extent of fire, material first ignited.

Presence of fire prevention system includes: presence of fire alarm, smoke alarm, and sprinkler system.

Lastly, included the alarm time for Toronto Fire Service(TFS), time for TFS to arrive and the time when fire is under control. With these variables, we are able to derive the time took for TFS to arrive after the alarm, and the time took for TFS to control the fire.

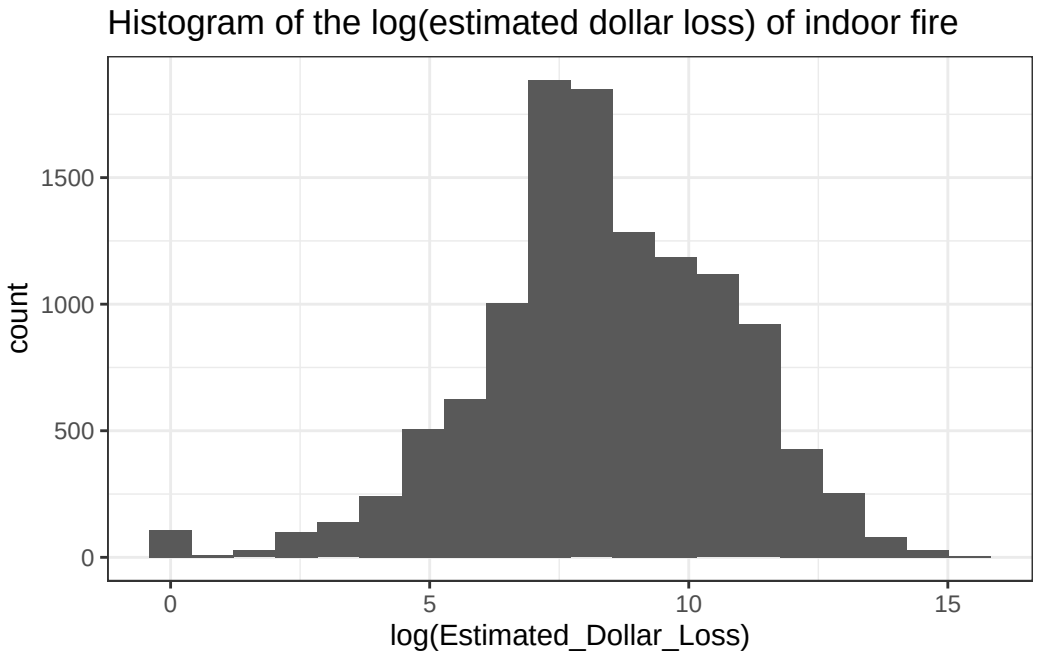
Dataset

The dataset is available in Toronto Open (Data.https://open.toronto.ca/dataset/fire-incidents/). The dataset include only fire incidents as defined by the Ontario fire Marshal(OFM) up to Decemeber 31, 2021.

The original dataset contains 25,860 observations and 43 variables. However, it includes any observations related to outdoor fire which are not our focus. All the outdoor fire are remove from the data. There are 216 observations with missing Estimated_Dollar_Loss are also removed because it is our variable of interest.

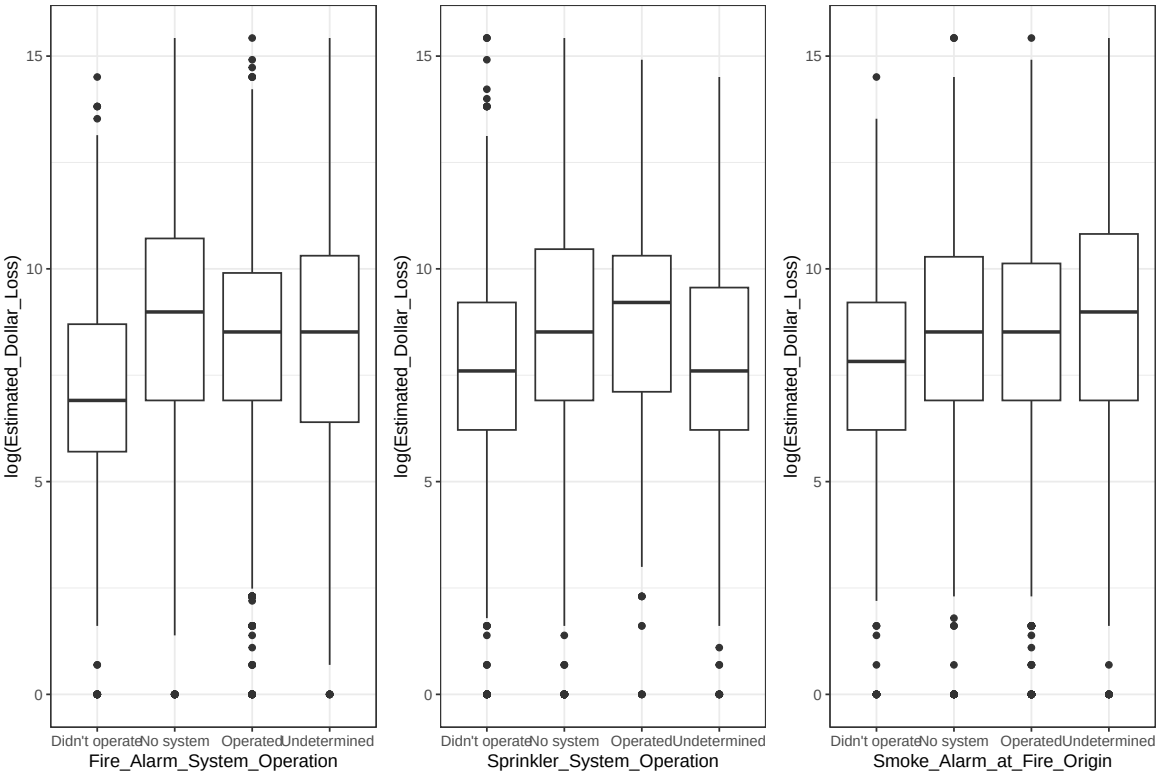
After data cleaning process described above, there are still 4855 observations with missing values for the following variables: Extent_Of_Fire, Fire_Alarm_System_Operation, Smoke_Alarm_at_Fire_Origin, Sprinkler_System_Operation. As in the Toronto Open Data: Incidents with incomplete data may be under investigation or is classified as a no loss outdoor fire. Therefore, I decided to remove these observations. Lastly, I remove 12 outlier with estimated cost above 5 Million for better plots.

EDA

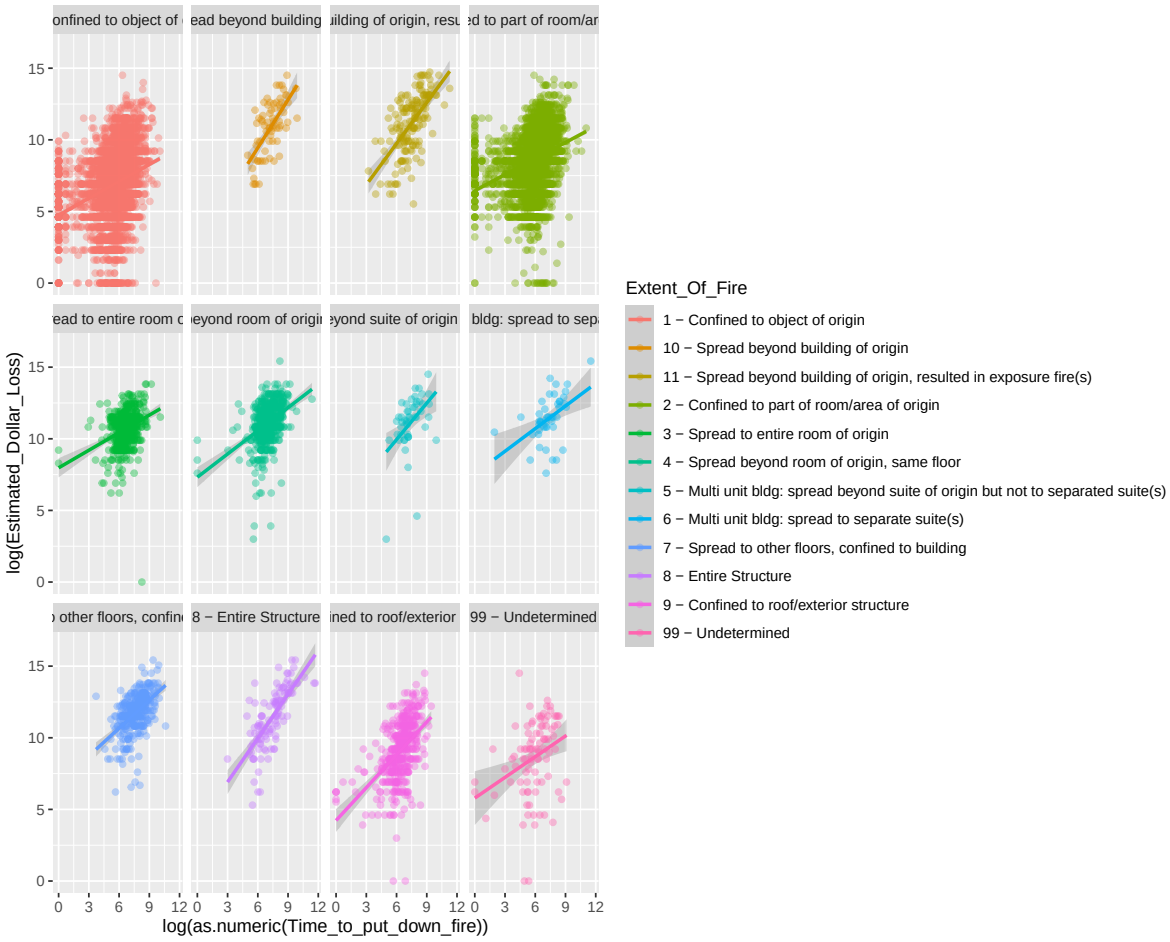


The log estimated dollar loss is close to nromal.

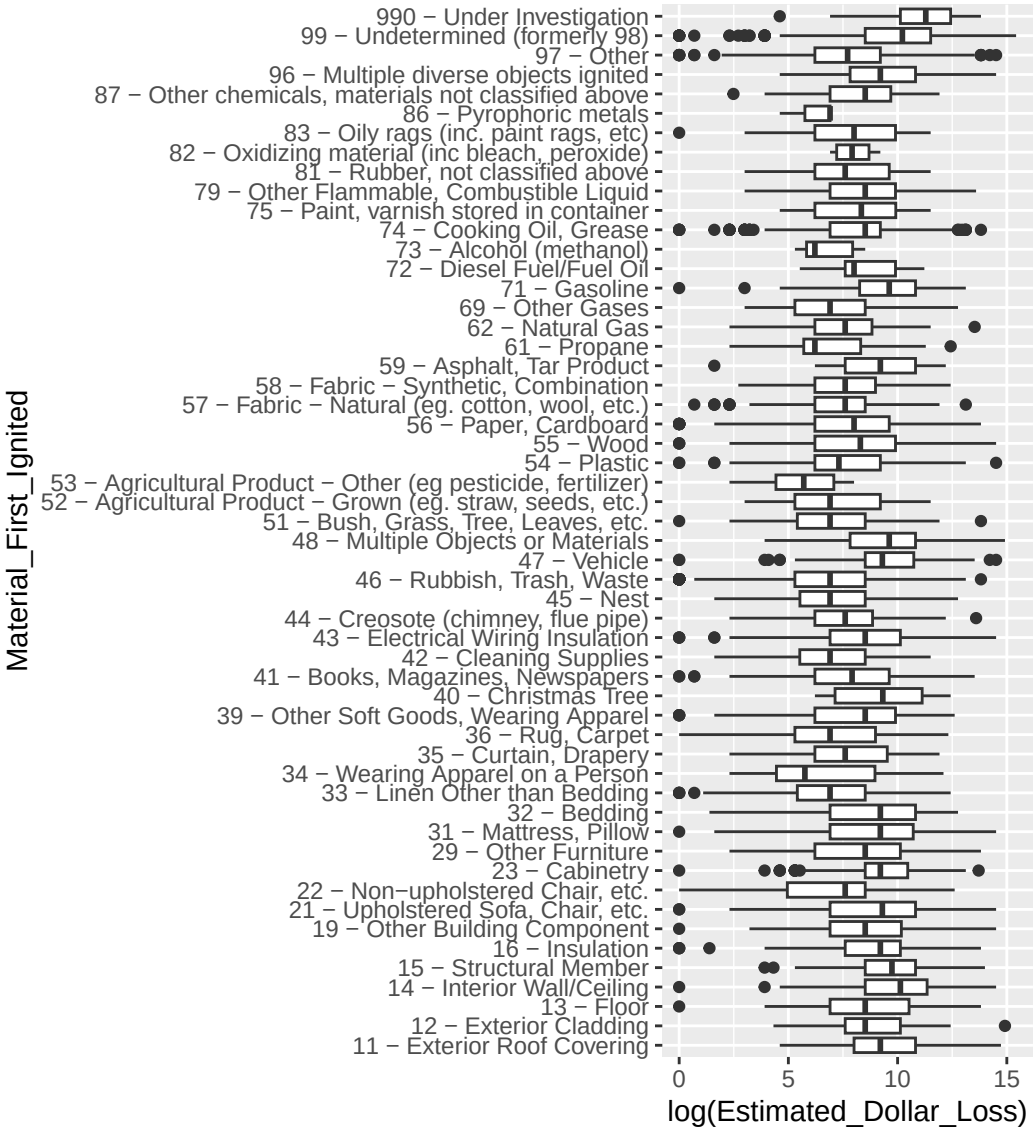
Boxplot of log(Estimated_Dollar_Loss) vs the presence of the fire prevention system



Looks like if the fire prevention system did not operated, the cost is lower. This is likely that because fire is too small to trigger the system. For fire alarm, the median of log cost when there is no system is slightly higher than when the system operated. Where as for sprinkler system when system operated has higher cost compared to no system. No significant difference for smoke alarm between no system and operated. Also notice that when the median when system operated has higher cost compared to when the system did not operate for all three system.



Different extent of fire lead to different slope when we have the log cost on the y axis and the time to put down the fire on x axis. Maybe we can put an interaction on the extent of fire and the time took to put down fire.



Different material first ignited has different cost distribution, suggesting an random effect.

Model

Based on my EDA, I'm going to fit the following model:

$$y_{[i]}|\alpha_{j[i]}, \beta_{[i]} \sim N(\alpha_{j[i]} + \beta_{[i]}x_i)$$

With normal noninformative on the $\alpha_{j[i]}$ and β

Where $\alpha_{j[i]}$ is the Material First Ignited specific intercept (Random effect)

β are the coefficients of the fixed effect including variables listed in Dependent variable of interest and an interaction between extent of fire and the time took to put down fire.

x_i is the dependent variable