My title*

My subtitle if needed

Luxuan Zhu

02 February 2022

Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction (talk about dataset, findings, and conclusion)

2 1st paragraph: motivation and broad

3 2nd paragraph: what I did and findings

4 3 paragraph: implications

5 precisely discuss the source of data and the bias it brings (ethical, statistical, etc.)

Fire incidents occur everyday around the world. In most societies, these incidents are a source of threat to safety, capital, buildings, and societal development. They occur in different locations, with different causes, and result in damages to different extends. To further understand the causes of unintentional house fires, Turner and her peers conducted literature research and found that demographic factors such as age, gender, and socio-economic status can contribute to fire incidents. Moreover, social factors such as household composition, property characteristics, and fire safety practices can potentially cause house fires. An important observation from the study is that older and disabled residents are more likely to suffer an injury or fatality in the event of a fire, while they are less likely to ignite the fire in the first place. (Turner et al. 2016) Therefore, It is important for a society to look at the characteristics across fire incidents and further prevent future events, especially to protect those who are more in need for help and support.

An inevitable consideration while working with a dataset of real incidents is ethics. One needs to consider who are represented by the dataset and who are excluded. Under ethical considerations, what does it mean to make conclusions and thus influences based on the dataset? Looking at the source of the dataset, one can assume that the dataset does not record all fire incidents in Toronto. For instance, some incidents were excluded due to privacy concerns, while some may simply not be reported. Then, it is important to consider the groups of people influenced by the fire incidents. The dataset does not reflect the demographic of the persons involved, thus each incident may associate with any one person or group, with any demographic information, under any circumstance. Thus, for instance, when a statistician concludes that one type of fire incident is more significant than another type, the same affect may apply to the individuals associated behind the two types of incidents. Does that mean the group of individuals associated with the more severe incident type “deserve” more support? These questions are crucial to people such as statisticians and policy

*Code and data are available at: LINK.

Table 1: First ten rows of a dataset of fire incidents in Toronto

Area of Origin	Fire Alarm System Operation	Extent of Fire	Building Status	Latitude	Longitude	Estimated Dollar Loss
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	1 - Confined to object of origin	01 - Normal (no change)	43.7	-79.4	500
24 - Cooking Area or Kitchen	8 - Not applicable (no system)	1 - Confined to object of origin	01 - Normal (no change)	43.7	-79.4	0
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	1 - Confined to object of origin	01 - Normal (no change)	43.8	-79.4	0
24 - Cooking Area or Kitchen	8 - Not applicable (no system)	2 - Confined to part of room/area of origin	01 - Normal (no change)	43.7	-79.4	60000
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	1 - Confined to object of origin	01 - Normal (no change)	43.7	-79.6	3000
24 - Cooking Area or Kitchen	8 - Not applicable (no system)	1 - Confined to object of origin	01 - Normal (no change)	43.7	-79.4	200
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	2 - Confined to part of room/area of origin	03 - Under Construction	43.8	-79.3	15000
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	2 - Confined to part of room/area of origin	01 - Normal (no change)	43.8	-79.4	10000
24 - Cooking Area or Kitchen	1 - Fire alarm system operated	1 - Confined to object of origin	01 - Normal (no change)	43.7	-79.4	5000
24 - Cooking Area or Kitchen	8 - Not applicable (no system)	2 - Confined to part of room/area of origin	01 - Normal (no change)	43.7	-79.6	5000

setters who might rely on results derived from the dataset to make certain decisions that would, directly or indirectly, influence the society.

This paper provides an in-depth analysis on the fire incidents occurred in cooking areas or kitchens located in Toronto, ON. Specifically, the author is interested in exploring the differences in estimated money loss as a result of the fire incidents, categorized by whether the fire alarm system was operated. The author's hypothesis is that fire incidents where the fire alarm system did not operate or did not exist would result in higher estimated dollar loss compared to those where the fire alarm system operated. Such a hypothesis

The author found that…

The remainder of the paper will follow the structure of: Section **??** explains the data; Section **??** showcases the models to support the analysis; Section **??** covers the results of the models; Section…

6 Data (generate a table and talk about the variables)

The dataset is obtained from Open Data Toronto (Gelfand 2020) using R (R Core Team 2020). The dataset provides information on 17536 fire incidents to which Toronto Fire responds in more detail, as displayed by the 43 variables including Area of Origin, Extend of Fire, and Estimated Dollar Loss. The dataset includes only fire incidents defined by the Ontario Fire Marshal, which means it does not capture all fire incidents in Toronto. Since the fire incidents recorded are actual events related to citizens of Toronto and the society as a whole, personal information is not provided and the exact address have been approximated to the nearest intersections for privacy purposes. Moreover, the dataset follows exemptions under Section 8 of Municipal Freedom of Information and Protection of Privacy Act (MFIPPA) and excludes certain incidents.

For the purpose of the paper, the analysis will focus on the fire incidents occurred in cooking areas or kitchens. Moreover, the author is interested in exploring the relationship between the estimated dollar loss and the fire alarm system operation among the selected kitchen fire incidents. To do so, the author uses Tidyverse (Wickham et al. 2019) to select the variables of "Area of Origin," "Fire Alarm System Operation," "Latitude," "Longitude," and "Estimated Dollar Loss." The author also removes all entrants with missing values for any of the variables. The following shows an extract of the cleaned dataset (Table 1).

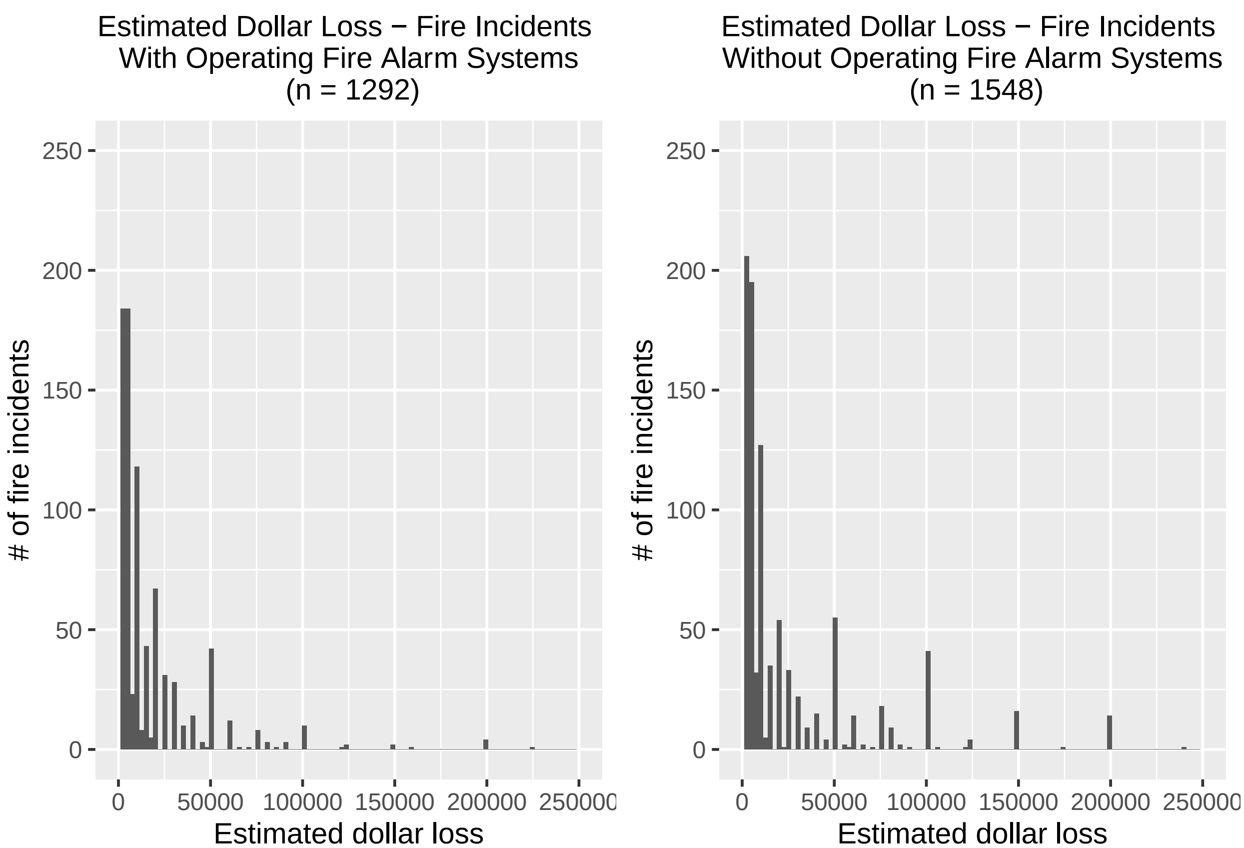
The first variable, Area of Origin, shows the initial cause of the fire incidents. In the context of this paper, all observations should contain the Area of Origin of "24 - Cooking Area or Kitchen." The second variable of Fire Alarm System Operation shows the status of the fire alarm system during the fire incident. The four statuses include: Fire alarm system operated; Fire alarm system did not operate; not applicable (no system); and Fire alarm system operation undetermined. The next two variables, longitude and latitude, are used in mapping the locations of the fire incidents. Lastly, the Estimated Dollar Loss showcases the financial consequences of each fire incident.

fire_alarm_system_operation	N	%
1 - Fire alarm system operated	1292	45.5
2 - Fire alarm system did not operate	364	12.8
8 - Not applicable (no system)	936	33.0
9 - Fire alarm system operation undetermined	248	8.7

7 Need a paragraph to describe the dataset

8 Histogram

From observing the dataset, the author is interested in seeing the difference between the distributions of estimated dollar loss relative to the fire alarm system. The author's hypothesis is that fire incidents where the fire alarm did not operate, was absent, or was undetermined, would contribute to a higher estimated dollar loss. In contrast, fire incidents where the fire alarm system operated would result in faster reactions and thus lower estimated dollar loss. Therefore, to test such a hypothesis, the author creates two histograms with estimated dollar loss on the x-axis and the number of fire incidents on the y-axis. The data input is the same cleaned dataset where the area of origin focuses on cooking area/kitchen. The histogram on the left represents the distribution of estimated dollar loss for fire incidents with operating fire alarm systems, whereas the graph on the right represents the remaining fire incidents where the fire alarm system was either undetermined, absent, or did not operate. To get a closer observation of the distributions, the author excluded the fire incidents with estimated dollar loss higher than $250,000. The detailed breakdown of the outliers are represented by the table below (Table 8), followed by a summary of the number of fire incidents under each fire alarm system status (Table @ref(tab:summary_table)), and lastly the resulting histograms (Figure **??**).



\begin{table}

\caption{Fire Incidents with Estimated Dollar Loss Exceeding $250,000}

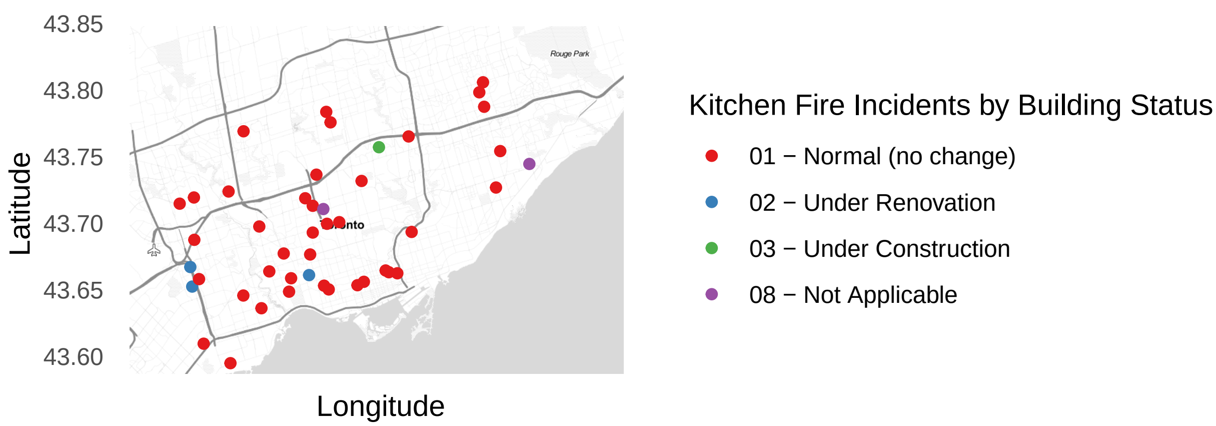
Fire Alarm System Operation	# of Fire Incidents
1 - Fire alarm system operated	6
2 - Fire alarm system did not operate	4
8 - Not applicable (no system)	22
9 - Fire alarm system operation undetermined	13

\end{table}

The outlier summary table shows the number of fire incidents with estimated dollar losses exceeding $250,000, which are excluded from the histograms. As one may observe, the number of outliers with operating fire alarm systems is much lower than the rest of the three categories.

9 Map

To further analyze the outliers, one can create a visual representation of the geographical distribution of these fire incidents with high estimated dollar loss. The author uses ggmap (Kahle and Wickham 2013), mapproj (R by Ray Brownrigg, Minka, and Plan 9 codebase by Roger Bivand. 2022), and maps (Richard A. Becker, Ray Brownrigg. Enhancements by Thomas P Minka, and Deckmyn. 2021) to construct a map pinpointing the locations of selected data points. Besides the location, the author is also interested in seeing the building status of each incident, taking possible values of normal, under renovation, under construction, and unknown. The building status variable is shown via color.



10 Results

Previously, the author used two histograms to illustrate the distributions of estimated dollar loss between fire incidents with operating fire alarm systems and those without operating fire alarm systems. Moreover, the author explores the geographical distribution of fire incidents with more than $250,000 of estimated dollar loss within Toronto. From the histograms, one may conclude that both fire incidents with or without operating fire alarm systems result in mostly estimated dollar loss below $25,000. However, when the operating fire alarm system is absent, there are more fire incidents with estimated dollar loss above $500,000. For instance, around 40 fire incidents without operating fire alarm systems had estimated dollar losses around $100,000, a few with $150,000 and $200,000, compared to a much lower number of fire incidents with operating fire alarm system that had the same amount of estimated dollar loss. One can also observe the same pattern from the outlier summary table, where most outliers do not have operating fire alarm systems.

In addition, the geographical distribution of the outliers suggests that most fire incidents with large dollar losses occurred around central to southern part of Toronto, with the building status of “normal.” It is coherent with the assumption that fire incidents from the kitchen or cooking area are more likely to occur in operating or occupied buildings.

10.1 Weaknesses and next steps

Appendix

A Additional details

References

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Kahle, David, and Hadley Wickham. 2013. "Ggmap: Spatial Visualization with Ggplot2." *The R Journal* 5 (1): 144–61. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

R by Ray Brownrigg, Doug McIlroy. Packaged for, Thomas P Minka, and transition to Plan 9 codebase by Roger Bivand. 2022. *Mapproj: Map Projections*. https://CRAN.R-project.org/package=mapproj.

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

Richard A. Becker, Original S code by, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2021. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Turner, Samantha L, Rhodri Johnson, Alison L Weightman, Sarah E Rodgers, Geri Arthur, Rowena Bailey, and Ronan A Lyons. 2016. "RISK FACTORS FOR UNINTENTIONAL HOUSE FIRE INCIDENTS, INJURIES AND DEATHS: A SYSTEMATIC REVIEW." *Injury Prevention* 22: A301.

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.