


STA238 - Winter 2021

Assignment 1

[Sunny (Wei-Han) Wang - 

January 30, 2021

Part 1

Question:

Suppose you are working for Canada Life Downtown Toronto office and want to know how did last decade insurance plans sale go. Canada life is an assurance company with services in assurance (insurance for insurance companies), insurance, investment, and more.

Let X_1, X_2, \dots, X_n be independent identically distributed and represent the number of insurance plan sold every day in Downtown Toronto office. We are not sure of the distribution but we know the mean is 23 and variance is 156.

Approximate: $P(X_n < 25)$ if n is 90

Solution:

We use Central Limit Theorem (CLT) to calculate sample mean and standard deviation for a population that is too big. We could also use CLT to calculate the probability of mean greater than, less than, equal to a value given a sample size. From MIPS textbook, we know the formula of CLT.

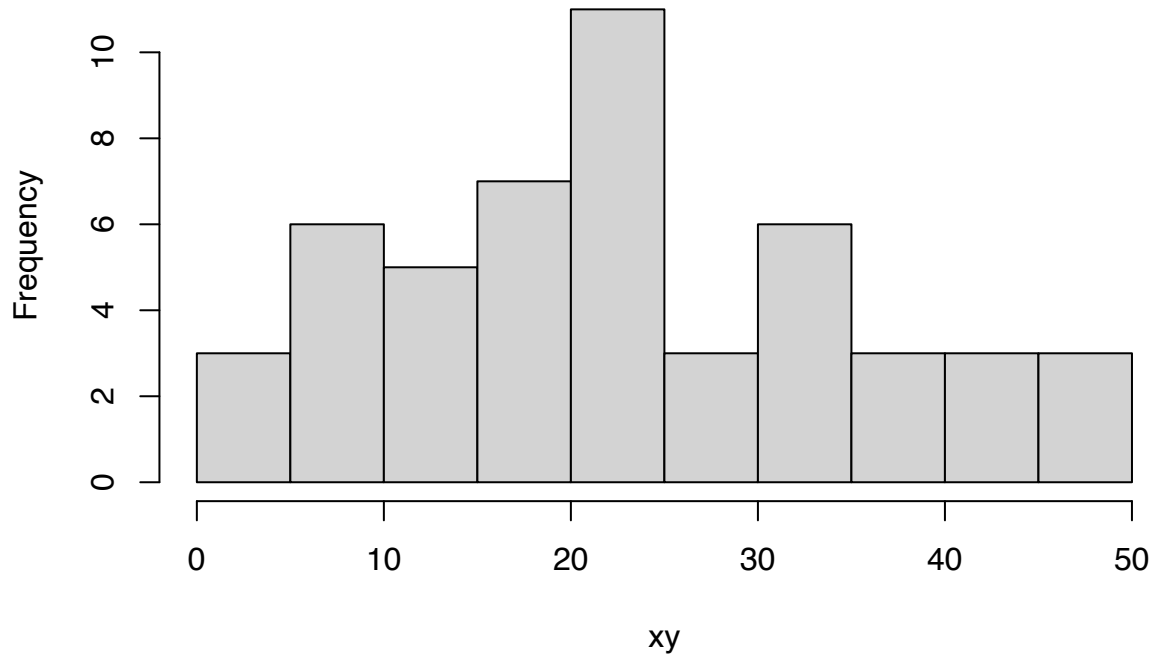
```
xn = 25
m = q
v = var(xy)
n = 90
abs(xn-m)/(sqrt(n)*sqrt(v))
```

```
## [1] 0.01626246
```

There are 1.6% that when n is 90, X_n is less than 25 plans sold.

```
hist(xy, main = "Distribution of Insurance Plans")
```

Distribution of Insurance Plans



The histogram shows distribution of insurance plans that Canada Life has sold if we were to collect random 50 days' sales report and find its mean.

Reference

1. *Canada Life*, [<http://www.canadalife.com/>]. (Last accessed: January 26, 2021)
2. Glen, Stephanie. “*Central Limit Theorem: Definition and Examples in Easy Steps.*” *Statistic-showto.com*, 11 Jan. 2021, [<http://www.statisticshowto.com/probability-and-statistics/normal-distributions/central-limit-theorem-definition-examples/>]. (Last accessed: January 26, 2021)
3. Yang, Yi. “*An Example R Markdown.*” *Math.mcgill.ca*, 5 Oct. 2017, [<http://www.math.mcgill.ca/yyang/regression/RMarkdown/example.html>] (Last accessed: January 26, 2021)

Part 2

Data Introduction

This dataset, Fire Incident, is collected through Open Data in Toronto's city website (<http://open.toronto.ca>). Once the link is opened, you can either search for dataset or click the button that says “*explore the catalogue*” right below the scenery picture. When you enter the catalogue, you will be able to filter types of datasets you are interested in. To find Fire Incident data, see the group called “Publisher”, expand it and click *Fire Services*. Fire Incidents will appear on first page of result. Click on the data and you will be redirected to a page containing preview, features, exploration of Fire Incident dataset. You can click on the *download data* bar and choose the format you'd like your dataframe in and click *download*.

The format of the dataset is similar to what is sent to Ontario Fire Marshal. Whenever a new fire incident is reported, a new data is imported onto this dataset. The table is refreshed annually for Torontonians to view. There are many variables in here so it is clear that the table goes in deep details when documenting fire incidents.

#Data Cleaning Process (*combining columns*)

The following code inputs our data in file type csv from the directory we are working in. Next, I used the function *glimpse(data)* to see the variables I am dealing with. This code also displays the first few observations of variables. This is very important as it could provide some ideas of how to begin the process of tidying the data.

```
Fire_Incidents_Data <- read_csv("/home/jovyan/STA238-Assignment1.git/Fire Incidents Data.csv")
glimpse(Fire_Incidents_Data)
```

```
## Rows: 17,536
## Columns: 43
## $ '_id' <dbl> 12279...
## $ Area_of_Origin <chr> "81 -...
## $ Building_Status <chr> NA, N...
## $ Business_Impact <chr> NA, N...
## $ Civilian_Casualties <dbl> 0, 0,...
## $ Count_of_Persons_Rescued <dbl> 0, 0,...
## $ Estimated_Dollar_Loss <dbl> 15000...
## $ Estimated_Number_Of_Persons_Displaced <dbl> NA, N...
## $ Exposures <dbl> NA, N...
## $ Ext_agent_app_or_defer_time <dtm> 2018...
## $ Extent_Of_Fire <chr> NA, N...
## $ Final_Incident_Type <chr> "01 -...
## $ Fire_Alarm_System_Impact_on_Evacuation <chr> NA, N...
## $ Fire_Alarm_System_Operation <chr> NA, N...
## $ Fire_Alarm_System_Presence <chr> NA, N...
## $ Fire_Under_Control_Time <dtm> 2018...
## $ Ignition_Source <chr> "999 ...
## $ Incident_Number <chr> "F180...
## $ Incident_Station_Area <dbl> 441, ...
## $ Incident_Ward <dbl> 1, 18...
## $ Initial_CAD_Event_Type <chr> "Vehi...
## $ Intersection <chr> "Dixo...
## $ Last_TFS_Unit_Clear_Time <dtm> 2018...
## $ Latitude <dbl> 43.68...
## $ Level_Of_Origin <chr> NA, N...
## $ Longitude <dbl> -79.5...
```

```
## $ Material_First_Ignited <chr> "47 -...
## $ Method_Of_Fire_Control <chr> "1 - ...
## $ Number_of_responding_apparatus <dbl> 1, 1,...
## $ Number_of_responding_personnel <dbl> 4, 4,...
## $ Possible_Cause <chr> "99 -...
## $ Property_Use <chr> "896 ...
## $ Smoke_Alarm_at_Fire-Origin <chr> NA, N...
## $ Smoke_Alarm_at_Fire-Origin_Alarm_Failure <chr> NA, N...
## $ Smoke_Alarm_at_Fire-Origin_Alarm_Type <chr> NA, N...
## $ Smoke_Alarm_Impact_on_Persons_Evacuating_Impact_on_Evacuation <chr> NA, N...
## $ Smoke_Spread <chr> NA, N...
## $ Sprinkler_System_Operation <chr> NA, N...
## $ Sprinkler_System_Presence <chr> NA, N...
## $ Status_of_Fire_On_Arrival <chr> "7 - ...
## $ TFS_Alarm_Time <dtm> 2018...
## $ TFS_Arrival_Time <dtm> 2018...
## $ TFS_Firefighter_Casualties <dbl> 0, 0,...
```

Now I know what variables I have and how many of them are in this big data frame. When I look at these variables, I noticed *smoke alarm at fire origin*, *smoke alarm at fire origin failure*, and *smoke alarm at fire origin type* are related. Smoke alarm is an important variable in this dataset and in real life as it should help civilians from escaping the fire in the early stage.

Now we see all three columns are under the column name *smoke alarm at fire origin alarm failure* using the code `head()`.

Examine data variables and I've found two other columns that could be merged into one given the direct relationship to their values. Look at the values under *fire alarm system presence* and *fire alarm system operation* and I spot the relationship between those two variables. Again, I used the code `unite()` under tidyR package to combine these two columns.

Delete Empty Rows of Similar Columns

I found the observations under column *Area of Origin*, *Material first ignited*, *possible case*, and *ignition source* are very similar and if there is no input in one of those four columns, the other three will not have results either. All the empty observations do not serve good purpose for data viewing so I want to get rid of rows under these four columns that are all empty.

As one of the purpose of this data is to inform the viewers of areas that are more likely to have fire incidents. Therefore, the column *Area of Origin* is an important variable with educational observations. In order to have a visual presentation of which areas are more prone to accidents, I want to create multiple barplots for them.

I'll start by grouping the observations under column *Area of Origin* in order for R to calculate the frequency of each unique observation within the large dataset. Then I add a column to the data called *frequency* using the `mutate` function in the package `dplyr`. After doing so, I pick out two columns from the Fire Incident Data: *Area of Origin* and *frequency*, the newly added column. Finally, I write *distinct* to have R only show the unique observations of the column *Area of Origin*.

I save this two-column data under the name `p`.

```
p <- Fire_Incidents_Data %>%
  group_by(Area_of-Origin)%>%
  mutate(frequency = n())%>%
  select(Area_of-Origin, frequency)%>%
```

```
distinct()%>%
ungroup()
```

```
#Now we know how many times a fire has started in the same area (not same address)
#Order the area from a ascending order so barplots would be clear
p <- arrange(p)
head(p)
```

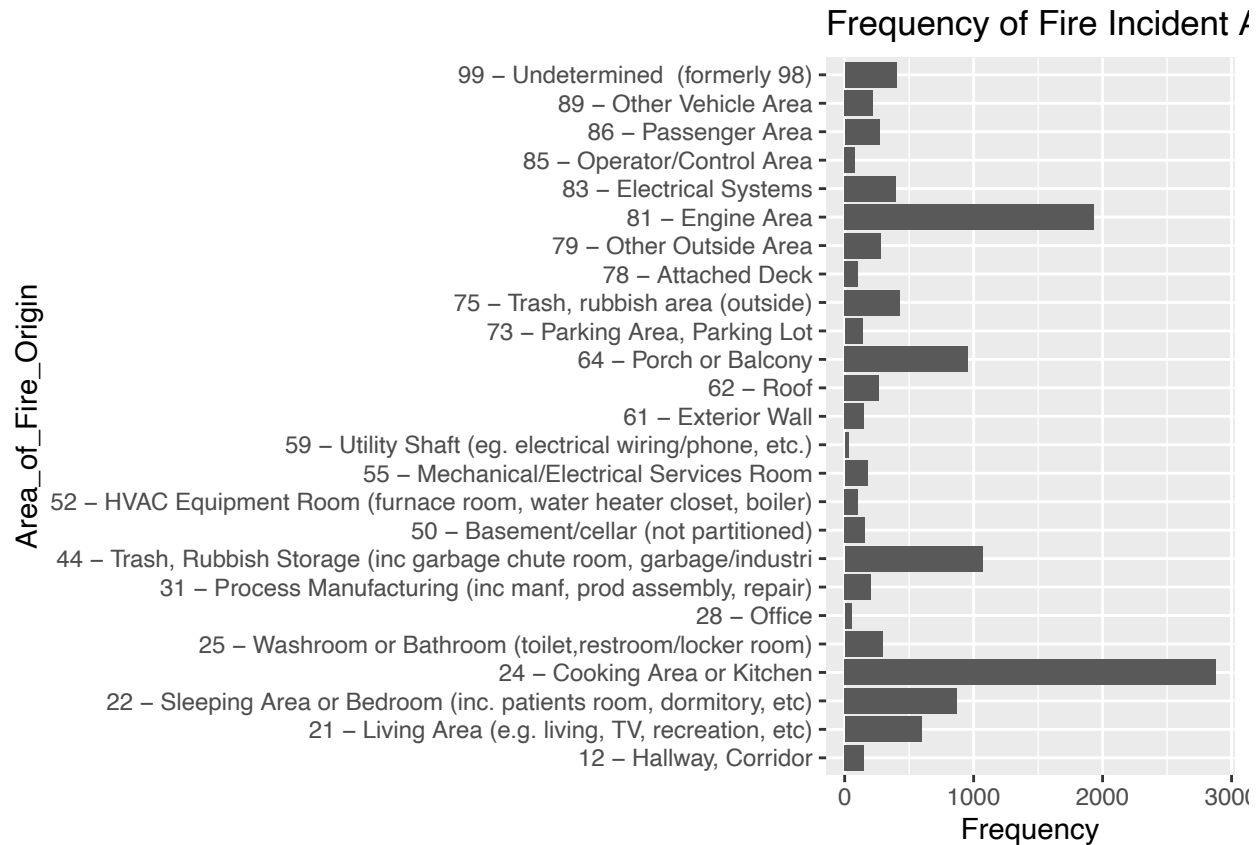
```
## # A tibble: 6 x 2
##   Area_of_Origin frequency
##   <chr>          <int>
## 1 81 - Engine Area      1931
## 2 75 - Trash, rubbish area (outside)    429
## 3 22 - Sleeping Area or Bedroom (inc. patients room, dormitory, etc)    871
## 4 55 - Mechanical/Electrical Services Room    177
## 5 28 - Office          54
## 6 24 - Cooking Area or Kitchen      2878
```

There are 73 observations within data *p* and it would be very chaotic if we were to graph them all in one barplot. So we will be dissecting this data into 3.

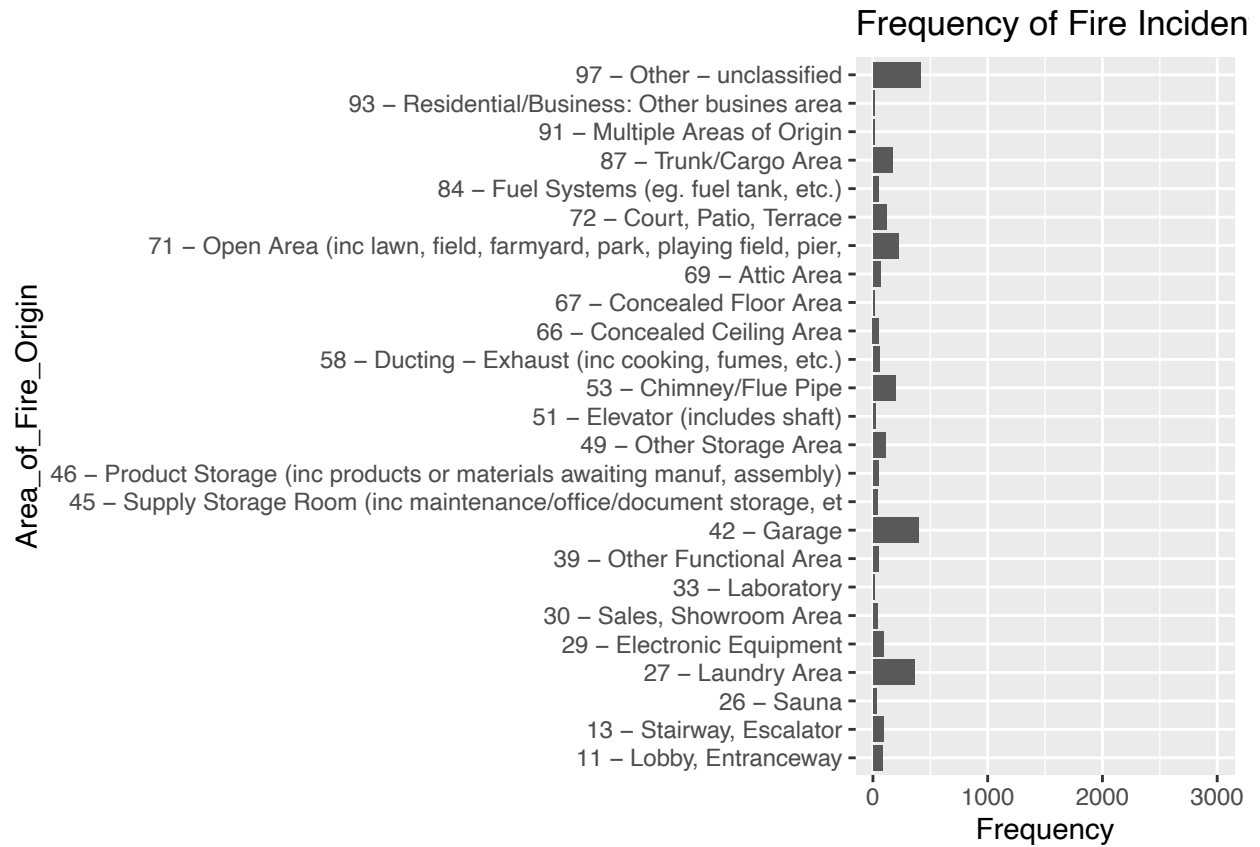
```
#The code we are doing here allows R to select certain rows
#from our data using its corresponding row number.
one <- p[1:25,]
two <- p[26:50,]
three <- p[51:73,]
```

Using ggplot we plot each data onto barplot. First barplot has higher frequency of fire origin so I have to change the limit of y-axis for the second and third plot to make comparison of three plots more convenient.

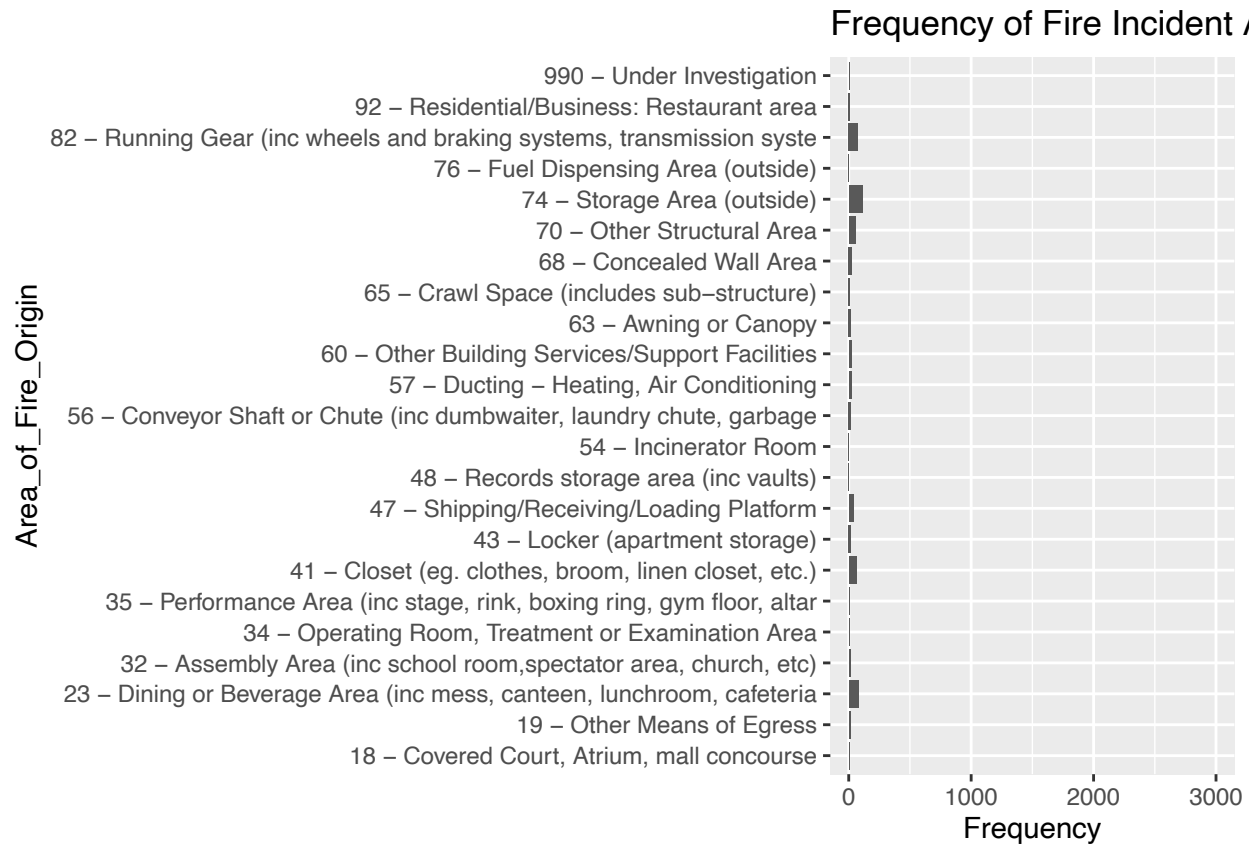
```
ggplot(one, aes(Area_of_Origin, frequency))+
  geom_bar(stat = "identity")+
  coord_flip()+
  labs(title = "Frequency of Fire Incident Area",
       x = "Area_of_Fire-Origin", y = "Frequency")
```



```
ggplot(two, aes(Area_of_Origin, frequency))+
  geom_bar(stat = "identity")+
  coord_flip(ylim = c(0,3000))+
  labs(title = "Frequency of Fire Incident Area",
        x = "Area_of_Fire_Origin", y = "Frequency")
```



```
ggplot(three, aes(Area_of_Origin, frequency))+
  geom_bar(stat = "identity")+
  coord_flip(ylim = c(0,3000))+
  labs(title = "Frequency of Fire Incident Area",
       x = "Area_of_Fire_Origin", y = "Frequency")
```



We have plotted the frequency of fire incidents area. We see peaks in cooking areas and kitchen and engine areas on the three barplots. The barplots are informative to us because they tell us areas where we need to be more careful of our surroundings. This is one of ways we could interpret this data.

#Numerical Summary

Another way to evaluate the data is find the numerical distribution. As mentioned above, locations and areas of fire incidents are important to this data. We've acquired the higher frequencies fire origins and now we can acquire another accurate area location of these fire incidents. That is: longitude and Latitude of the fire location. This would be more accurate than street names as provided in the data.

We will need to remove rows under Latitude and Longitude columns that are empty because R cannot factor them into calculations.

The following function calculates the quantile of longitude and latitude excluding the missing values as well as NaN values in our original dataframe by using the function na.rm.

```
Longitude_Q75 <- quantile(Fire_Incidents_Data$Longitude,
                          probs = 0.75, na.rm = TRUE)
Longitude_Q25 <- quantile(Fire_Incidents_Data$Longitude,
                          probs = 0.25, na.rm = TRUE)
Latitude_Q75 <- quantile(Fire_Incidents_Data$Latitude,
                         probs = 0.75, na.rm = TRUE)
Latitude_Q25 <- quantile(Fire_Incidents_Data$Latitude,
                         probs = 0.25, na.rm = TRUE)
Longitude_mean <- mean(Fire_Incidents_Data$Longitude, na.rm = TRUE)
Latitude_mean <- mean(Fire_Incidents_Data$Latitude, na.rm = TRUE)
Longitude_Med <- median(Fire_Incidents_Data$Longitude, na.rm = TRUE)
Latitude_Med <- median(Fire_Incidents_Data$Latitude, na.rm = TRUE)

Summary <- data.frame(Longitude_Q25, Longitude_Q75,
                      Latitude_Q25, Latitude_Q75,
                      Longitude_mean, Latitude_mean,
                      Longitude_Med, Latitude_Med)

glimpse(Summary)
```

```
## Rows: 1
## Columns: 8
## $ Longitude_Q25 <dbl> -79.48914
## $ Longitude_Q75 <dbl> -79.33561
## $ Latitude_Q25 <dbl> 43.66381
## $ Latitude_Q75 <dbl> 43.75247
## $ Longitude_mean <dbl> -79.40597
## $ Latitude_mean <dbl> 43.70867
## $ Longitude_Med <dbl> -79.40772
## $ Latitude_Med <dbl> 43.70322
```

The average latitude and longitude of fire incidents are (43.70867°, -79.406597°). As a fire incident data available to public, one of its purpose is to remind civilians areas that are easier to cause a fire if not careful. Other purposes of the data are also about the importance of having smoke alarm and fire alarm systems installed and operating within households. We could create a side-to-side bar chart to see if there is a link between having alarm systems and casualties from fire incidents. Knowing the geographical coordinate of fire incidents show people a direct and exact location. Perhaps this could be beneficial for Toronto Fire Department to increase officers around the more frequent fire incident location. There are a lot of variables to work with and there are many hypothesis that could result from this dataset.

All analysis for this report was programmed using R `version 4.0.2`.

Bibliography

1. Grolemond, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. <https://rmarkdown.rstudio.com/docs/>. (Last Accessed: January 15, 2021)
4. Palavuzlar, Mehper C. (2011, Oct 13) “*List Distinct Values in a Vector in R.*” Stack Overflow. [<http://stackoverflow.com/questions/7755240/list-distinct-values-in-a-vector-in-r>] (<https://stackoverflow.com/questions/7755240/list-distinct-values-in-a-vector-in-r>) (Last Accessed: January 28, 2021)
5. Kumar, Ajitesh. (2014, Sept 8) “*Learn R: How to Extract Rows and Columns From Data Frame - DZone Big Data.*” DZone. [<http://dzone.com/articles/learn-r-how-extract-rows>] (<http://dzone.com/articles/learn-r-how-extract-rows>). (Last Accessed: January 28, 2021)
6. Bharani. (2018, Mar 26) “*Finding Frequency of Observations in R.*” Edureka Community. [<http://www.edureka.co/community/201/finding-frequency-of-observations-in-r>] (<http://www.edureka.co/community/201/finding-frequency-of-observations-in-r>). (Last Accessed: January 28, 2021)
7. Bates, Casey. (2020, July 28) “*Tutorial: Loading and Cleaning Data with R and the Tidyverse.*” Dataquest [<http://www.dataquest.io/blog/load-clean-data-r-tidyverse/>] (<http://www.dataquest.io/blog/load-clean-data-r-tidyverse/>) (Last Accessed: January 27, 2021)
8. (2021, Jan 17) “*Data Import Cheatsheet.*” RStudio. [<http://https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf>] (<https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf>) (Last Accessed: January 28, 2021)
9. (2021, Jan 17) “*Data Transformation Cheatsheet.*” RStudio. [<http://https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf>] (<https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf>) (Last Accessed: January 28, 2021)
10. (2020, Nov 16) “*Data Visualization Cheatsheet.*” RStudio. [<http://https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf>] (<https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf>) (Last Accessed: January 28, 2021)