If You need to Speed, Avoid These Streets! An Analysis of the Red Light Camera Distribution in Toronto

Faria Khandaker

02/09/2020

Abstract

In this paper, I attempted to conduct Exploratory Data Analysis on the Red-Light Camera dataset to answer the question: Does Scarborough contain the highest number of red-light cameras in Toronto? After running frequency analysis on the selected variables from the dataset as well on as their relation to each other, I found that yes, Scarborough does in fact have the highest number of red-light cameras compared to other districts in Toronto. Although the reasoning behind this finding is not crystal clear, it does raise some interesting questions. Given that residents of Scarborough have lower average incomes than the other three districts used in the dataset, findings like this can lead to discussions and further research on the effect of policy on marginalized communities and better ways to tackle simple issues like traffic violations without unintentionally punishing the already disadvantaged.

Introduction

Red light cameras are automated cameras that takes pictures of vehicles crossing an intersection when the traffic light is red. The picture captures the vehicle's license plate number which is used to find the driver's address, where the hefty $325 ticket is sent. The dataset contains the names of the intersections where the cameras are found in Toronto as well as their latitude and longitude for ease of mapping. It is provided by Open Data Toronto and is refreshed annually, last update being February 3rd, 2020. The original dataset contains 32 variables with 148 instances of data. The values in the dataset are often repeated between columns but are written in different formats. This redundancy helped in confirming consistency of values across the table. The missing values in a number of columns along with the qualitative nature of the dataset resulted in the usage of only three columns for the exploratory data analysis; Main, which is thought to contain the names of major streets in Toronto, District which thought to divide Toronto into 4 regions (North York, Toronto and East York, Etobicoke York and Scarborough), and Year of Activation which most likely represents the day the camera first started working on a specific street. Table 1 in the appendix summarizes the findings for this dataset and provides many of the numbers on which the graphs in this paper are based on.

Research Question:

As a driver living in Scarborough, I was very interested in exploring this dataset. I see many red-light cameras when driving around and the appearance of new ones at intersections that never had them, always catch me by surprise. This analysis was very much question driven. My initial research questions were: Are red light cameras found in higher numbers in different wards around Toronto? Is the volume of traffic a reason for this? Is there a disparity in income between the areas where there are more red-light cameras versus where there are few? After seeing mostly empty columns for Wards 1-4 (there were only 4 ward columns, with only ward 1 being mostly full) and seeing a completed column for district (with one of the district classifications being Scarborough), the research questions changed to: Does Scarborough contain the highest number of red-light cameras in Toronto? Is the volume of traffic a reason for this? Is there a disparity in income between the areas where there are more red-light cameras versus where there are few? Streets with high traffic volume can have more people running red lights just because of the number of people who use those streets (higher probability of people crossing on red). It can also be that because of the sheer number of people using the streets, drivers speed through intersections so as not to get caught up in traffic. Disparity in income can often be traced back to many social issues like ill physical and mental health, chances of success in a post-secondary environment, etc. Those who work two jobs to make ends meet are probably the ones who are rushing to their second workplace causing them to speed on the streets and potentially run a few red lights.

Discussion

Initially the qualitative in nature of the data concerned me in terms of the kind of analysis that could possibly be run However, I noticed that each row in the dataset had unique ID and that the names of streets repeated throughout the "MAIN" column. I created a smaller dataset and ran analysis of those values using a variety of graphs and plots. Counting the frequency of each street in the dataset resulted in the number of red-light camera present per street (see Figure A2 in the Appendix). Figure 1 below highlights the streets with four or more red light cameras. The worst offenders are Lakeshore Blvd W and Steeles Ave W. Since the district each street resided was also present in the dataset, grouping by district and counting the frequency of the streets gave me the number of red-light cameras per district. As per my assumption, at 43, Scarborough had the highest number of red-light cameras between the four districts, but not by much, which was surprising (see figure 2 below). Etobicoke York came in second with 42 cameras, followed by Toronto and East York at 35 cameras and North York had the lowest amount with 28 cameras. Now was this distribution because of external social factors or did one district have a greater number of "main streets" in it than others? After running another frequency analysis, it was found that although Scarborough many major streets running through it (19 major streets), the Toronto and East York district had the most (23) with Etobicoke York and North York tied for third place (17) (See the appendix).

Figure 1: streets with Four or more Red Light Cameras

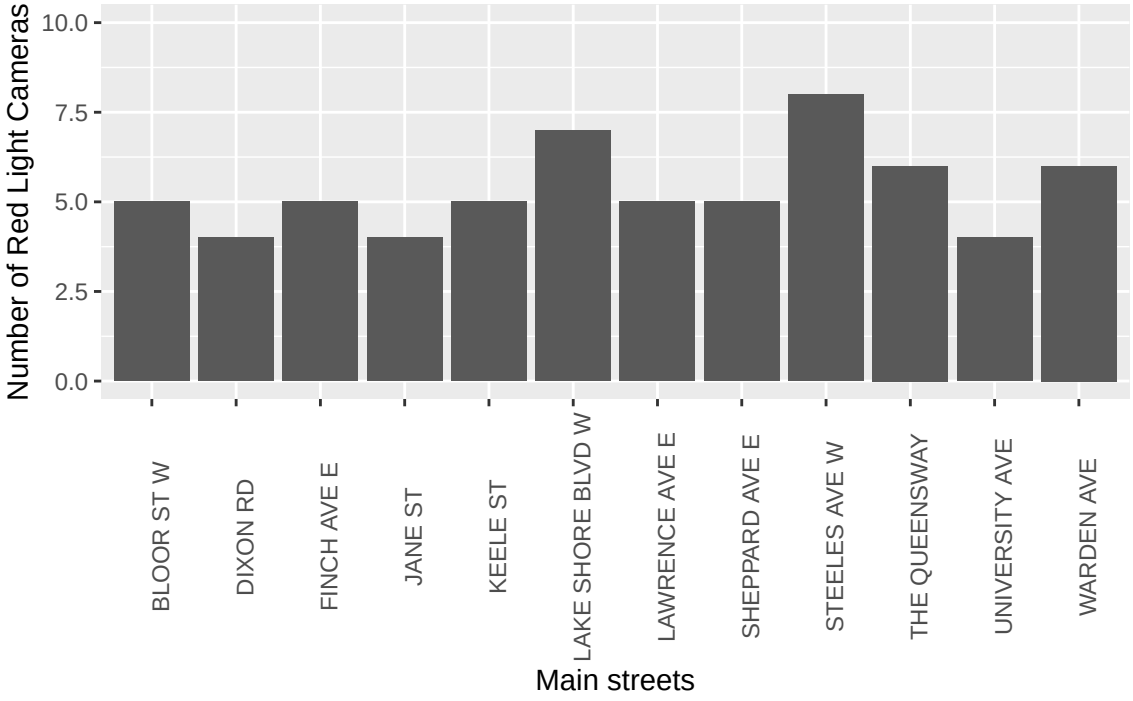


Figure 2: Number of Red light Cameras at each District

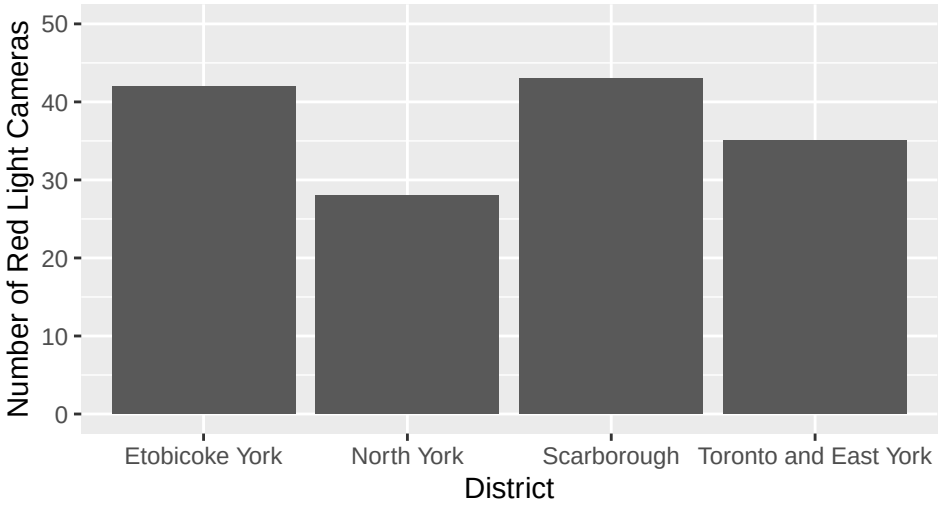


Figure 3 below portrays the installation/activation of red-light cameras between 2006 and 2020. There is a very clear gap in the graph approximately between the 2009 and 2016. Then, the installation of cameras increases in 2017 for all four districts but especially for Scarborough and Etobicoke York. A quick search on the Toronto Police's data portal reveals high numbers of vehicle involved fatalities in 2015 and 2016. How much of a correlation there is between those numbers and this increase in camera activation is unclear requires further investigation.

The biggest limitation of this dataset is the lack of descriptions of the column variables.

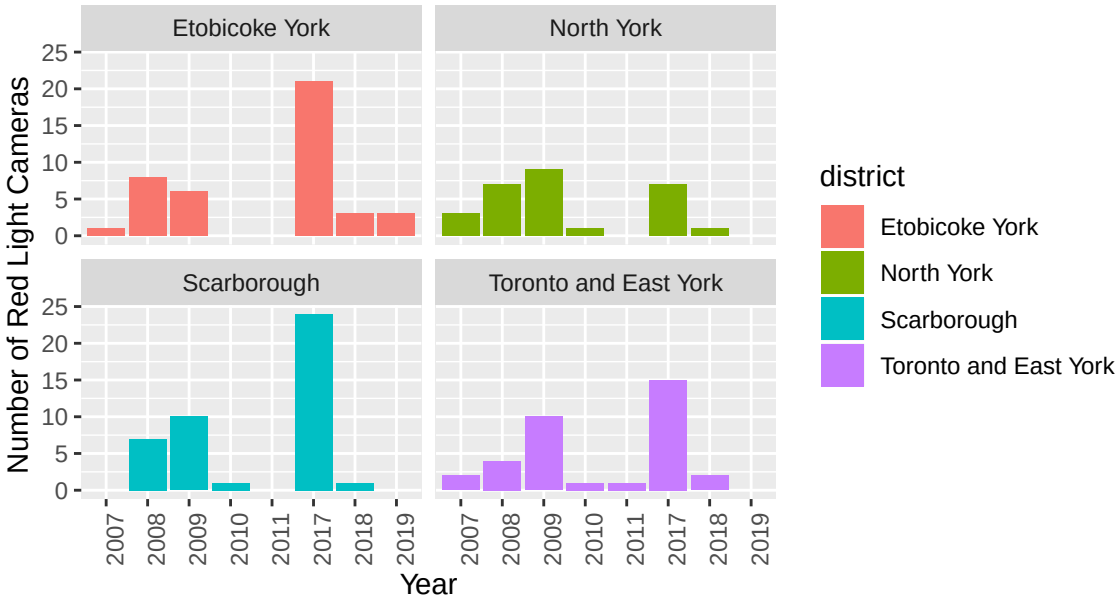
Much of the analysis relies on how well researchers can infer details from different columns. What is the difference between a ward and a district? Why are the street names repeated multiple times under multiple columns? The dataset is highly qualitative and there are only a handful of columns with complete information; this limits the kind of analytical tools that can be used for analysis. The biggest limitation to this research is the use of a singular dataset to infer many conclusions; thus creating biased results. It would allow for a more in-depth analysis of policy and use of tax-payer money if the efficacy of the camera was included as a column: how offenders captured, fatalities before and after camera installation, volume of the traffic passing each street, etc.

Although the answer is YES to the first part of my research question (Does Scarborough have the highest number of red-light cameras in Toronto?), it is unclear why that is. Is it because of traffic? The answer is unclear because the dataset revealed that Scarborough has the second highest number of major streets running through it (first being Toronto and East York district who had the third highest number of red-light cameras). It is difficult to pinpoint a relation. Can the higher camera numbers be traced to economic disparity? This is also unclear because Etobicoke and York district only lags behind Scarborough by one camera and average income for residents in the Etobicoke and York district is higher than residents in Scarborough (according to Wikipedia). There are pockets of low and high average income in all districts. An extension of this project requires the use of other datasets containing income data or traffic data to bring more depth to these findings.

There are no ethical issues in the compilation or use of this dataset or the results of the analysis as no personal information identifier exists nor can any private information be extrapolated.

In Conclusion, according to the findings presented in this paper, Scarborough is not a great place to accelerate on a yellow traffic light and if you are running late, best to avoid Steeles Ave W and Lakeshore Blvd W.

Figure 3: Red Light Camera Installation by District per Year



References

City of Toronto (2020). Red Light Cameras dataset. In: Toronto Open Data Portal.https://open.toronto.ca/dataset/red-light-cameras/. Accessed 01 Feb 2020.

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1. https://CRAN.R-project.org/package=skimr

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham and Lionel Henry (2020). tidyr: Tidy Messy Data. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.4. https://CRAN.R-project.org/package=dplyr

Ontario Traffic Tickets (2020). Red Light Traffic Tickets. In: Ontario Traffic Tickets. https://www.ontariotraffictickets.com/traffic-tickets/red-light-ticket/. Accessed 07 Feb 2020.

Sharla Gelfand (2019). opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.1. https://CRAN.R-project.org/package=opendatatoronto

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

R version 3.6.2 (2019-12-12) – "Dark and Stormy Night" Copyright (C) 2019 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit)

Toronto Police Service (2019). Public Safety Data Portal. In: Toronto Police Service. http://data.torontopolice.on.ca/pages/fatalities. Accessed 09 Feb 2020.

Wikipedia (2019). Demographics of Toronto neighorhoods. In: Wikipedia. https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods. Accessed 09 Feb 2020.

Appendix

categories of analysis	results
number of main streets	63
number of districts	4
number of redlight cameras in the dataset	148

Year with the highest number of redlight cameras installed	2017
District with the high number of redlight camera	Scarborough
Street(s) with the highest number of redlight cameras	Steeles Ave W (8) & Lakeshore Blvd (7)
District containing highest number of main streets	Toronto and East York (23)

Table 1: Summary of findings for the dataset, numbers in parenthesese() indicate amount

```
library(opendatatoronto)
library(tidyr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(lubridate)
library(skimr)
knitr::opts_chunk$set(fig.width=6, fig.height=4)

citation("skimr")
```

```
##
## To cite package 'skimr' in publications use:
##
##   Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la
##   Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible
##   Summaries of Data. R package version 2.1.
##   https://CRAN.R-project.org/package=skimr
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {skimr: Compact and Flexible Summaries of Data},
##     author = {Elin Waring and Michael Quinn and Amelia McNamara and Eduardo {Arino de
##     year = {2020},
##     note = {R package version 2.1},
##     url = {https://CRAN.R-project.org/package=skimr},
##   }
```

```r
red_light_camera <- search_packages("Red Light Cameras")
red_light_resources <- red_light_camera %>%
  list_package_resources()

red_light_camera_table <-
  red_light_resources %>%
  get_resource()
```

```r
summary(red_light_camera_table)
skim(red_light_camera_table)
```

```r
main <- red_light_camera_table$MAIN
```

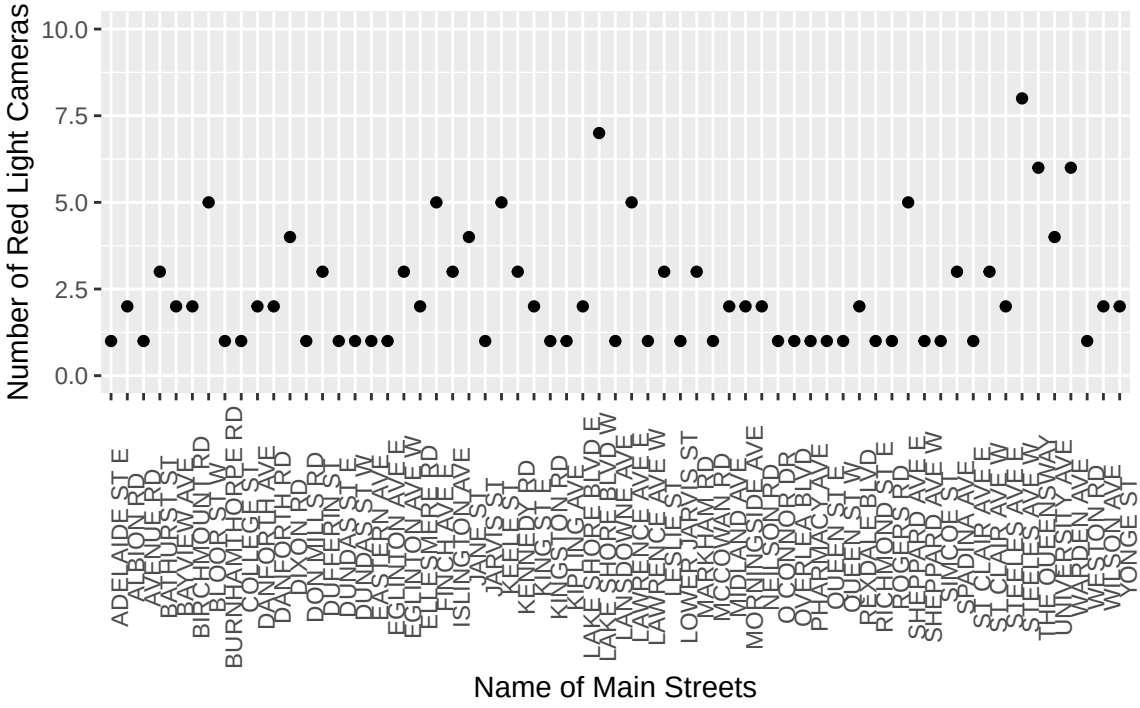
```r
#main is assumed to be names of main streets where the cameras appear
#counting the occurance of main is assumed to give us the number of cameras
count_table <-red_light_camera_table %>%
  group_by(MAIN) %>%
  count()

count_table <- tibble(main_streets = count_table$MAIN,
       num_of_rlightcam = count_table$n)
head(count_table)
```

```
## # A tibble: 6 x 2
##   main_streets  num_of_rlightcam
##   <chr>                    <int>
## 1 ADELAIDE ST E                1
## 2 ALBION RD                    2
## 3 AVENUE RD                    1
## 4 BATHURST ST                  3
## 5 BAYVIEW AVE                  2
## 6 BIRCHMOUNT RD                2
```

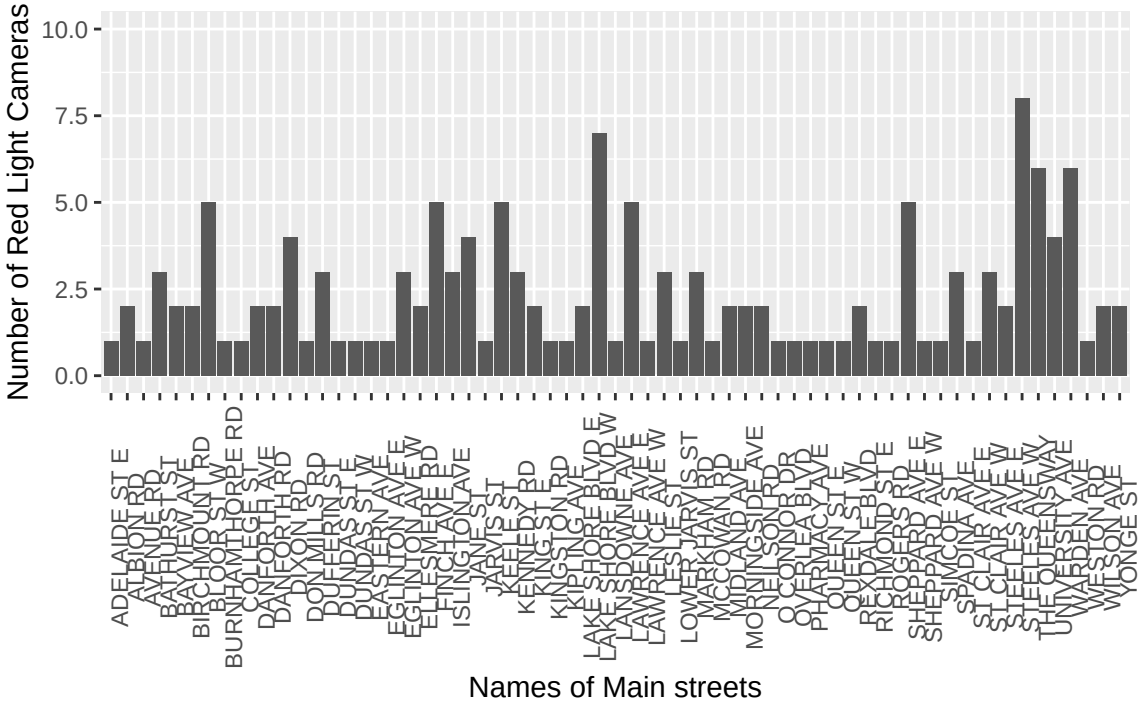
```r
ggplot(data = count_table)+
  geom_point(mapping = aes(x = main_streets, y = num_of_rlightcam))+
  theme(axis.text.x = element_text(angle = 90)) +ylim(0,10)+labs(
    title = "Figure A1:Red light cameras on major streets in Toronto",
       x = "Name of Main Streets",
       y = "Number of Red Light Cameras")
```

Figure A1:Red light cameras on major streets in Toronto




```
ggplot(data = count_table)+
  geom_col(mapping = aes(x = main_streets, y = num_of_rlightcam))+
  theme(axis.text.x = element_text(angle = 90)) +ylim(0,10)+labs(
    title = "Figure A2: Red light cameras on major streets in Toronto",
       x = "Names of Main streets",
       y = "Number of Red Light Cameras")
```

Figure A2: Red light cameras on major streets in Toronto




```
#main streets cross districts
#num of rlight cameras per street per district
count_unique_table <-red_light_camera_table %>%
  group_by(MAIN) %>%
  count(DISTRICT)

count_unique_table <- tibble(district = count_unique_table$DISTRICT,
       main_streets = count_unique_table$MAIN,
       number_of_streets = count_unique_table$n)
head(count_unique_table)
```

```
## # A tibble: 6 x 3
##   district              main_streets  number_of_streets
##   <chr>                 <chr>                     <int>
## 1 Toronto and East York ADELAIDE ST E                1
## 2 Etobicoke York        ALBION RD                    2
## 3 North York            AVENUE RD                    1
## 4 North York            BATHURST ST                  2
## 5 Toronto and East York BATHURST ST                  1
## 6 North York            BAYVIEW AVE                  2
```

```r
#number of main streets in district
MinD <- count_unique_table %>%
  group_by(district) %>%
  count(district)

MinD <- tibble(district = MinD$district,
       number_of_streets = MinD$n)
MinD
```

```
## # A tibble: 4 x 2
##   district              number_of_streets
##   <chr>                             <int>
## 1 Etobicoke York                       17
## 2 North York                           17
## 3 Scarborough                          19
## 4 Toronto and East York                23
```

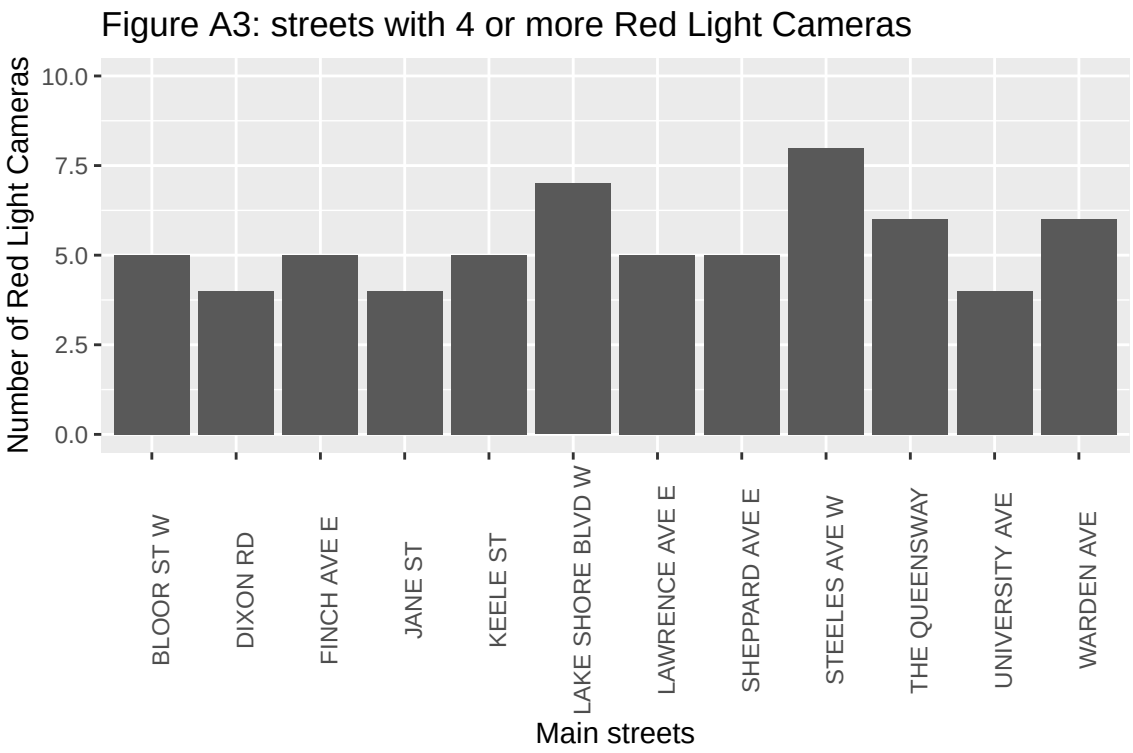
```r
#a few redlight cameras are not worrysome.
#Anything greater than three on single street can be annoying

avoidstreets <- subset(count_table, num_of_rlightcam > 3)
avoidstreets <- tibble(main_streets = avoidstreets$main_streets,
       num_of_rlightcam = avoidstreets$num_of_rlightcam)
avoidstreets
```

```
## # A tibble: 12 x 2
##    main_streets      num_of_rlightcam
##    <chr>                        <int>
##  1 BLOOR ST W                       5
##  2 DIXON RD                         4
##  3 FINCH AVE E                      5
##  4 JANE ST                          4
##  5 KEELE ST                         5
##  6 LAKE SHORE BLVD W                7
##  7 LAWRENCE AVE E                   5
##  8 SHEPPARD AVE E                   5
##  9 STEELES AVE W                    8
## 10 THE QUEENSWAY                    6
## 11 UNIVERSITY AVE                   4
## 12 WARDEN AVE                       6
```

```r
#streets with more than 3 redlight cameras
ggplot(data = avoidstreets)+
```

```
  geom_col(mapping = aes(x = main_streets, y = num_of_rlightcam))+theme(axis.text.x = el
    title = "Figure A3: streets with 4 or more Red Light Cameras",
       x = "Main streets",
       y = "Number of Red Light Cameras")
```




```
#table with only 3 attributes
cleanrlight <- red_light_camera_table %>% select(DISTRICT, MAIN,ACTIVATION_DATE)
head(cleanrlight)
```

```
## Simple feature collection with 6 features and 3 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: -79.60009 ymin: 43.64146 xmax: -79.36402 ymax: 43.79601
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
## # A tibble: 6 x 4
##   DISTRICT             MAIN             ACTIVATION_DATE                geometry
##   <chr>                <chr>            <chr>                       <POINT [°]>
## 1 Toronto and East Yo~ RICHMOND ST E    2007-11-09T05:00:00 (-79.36402 43.65456)
## 2 Toronto and East Yo~ LAKE SHORE BLVD~ 2007-11-09T05:00:00 (-79.38087 43.64146)
## 3 North York           STEELES AVE W    2007-11-09T05:00:00 (-79.44759 43.79201)
## 4 North York           STEELES AVE W    2007-11-09T05:00:00 (-79.42927 43.79601)
## 5 Etobicoke York       ALBION RD        2007-11-09T05:00:00 (-79.60009 43.74295)
```

```
## 6 Etobicoke York       DIXON RD         2008-03-18T04:00:00 (-79.55865 43.69425)
```

```
#get rid of geometry data column by converting tables to tibble

#turn date from 'character' to 'date' format
fix_date <- cleanrlight%>%
  separate(ACTIVATION_DATE, into = c("year", "rest"), sep="T")

fix_date <- fix_date %>% select(-rest)

formatted_date <-ymd(fix_date$year)


class(ymd(fix_date$year))
```

```
## [1] "Date"
```

```
#create new table with formatted dates in
summary_stats_table <- fix_date %>% cbind(formatted_date)
summary_stats_table <- summary_stats_table %>%  select(-year, -geometry)

#head(summary_stats_table)

#class(summary_stats_table)

attable <- tibble(district = summary_stats_table$DISTRICT,
       main = summary_stats_table$MAIN,
       formatted_date = summary_stats_table$formatted_date)

#class(attable)

#head(attable)
```

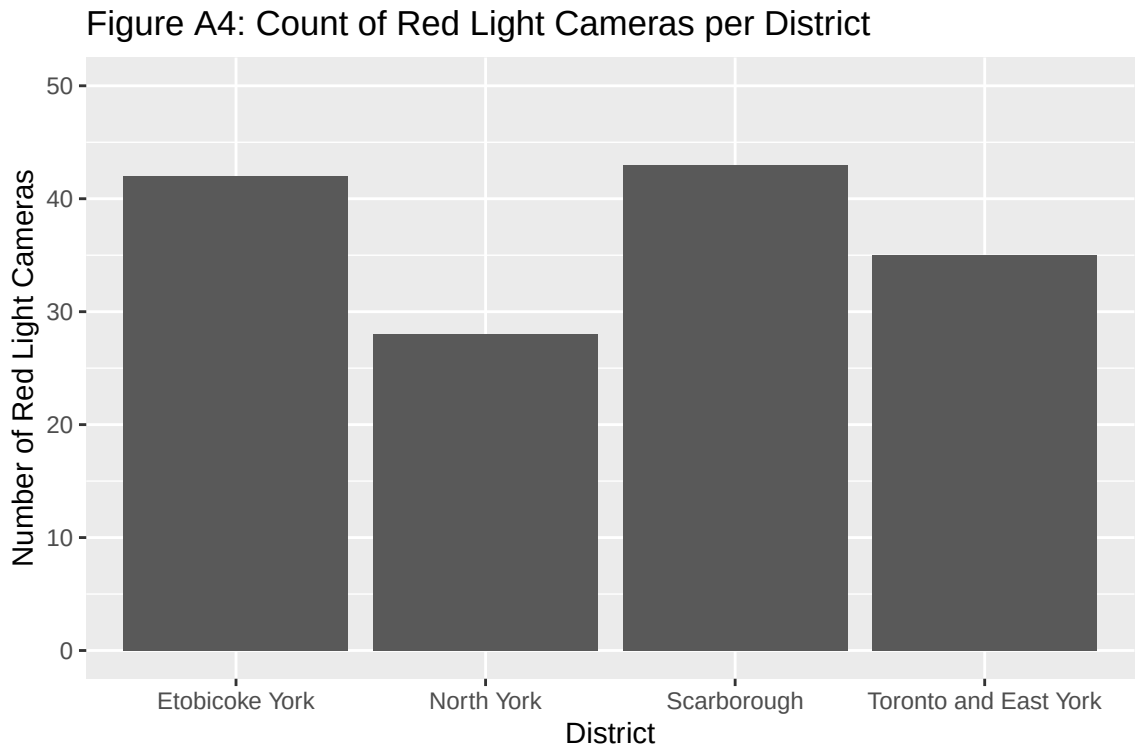
```
#number of redlight cameras per district
 district_count_table <-cleanrlight %>%
  group_by(DISTRICT) %>%
  count()
district_count_table <- tibble(district = district_count_table$DISTRICT,
       num_of_rlightcam = district_count_table$n)

district_count_table
```

```
## # A tibble: 4 x 2
```

```
##   district              num_of_rlightcam
##   <chr>                            <int>
## 1 Etobicoke York                      42
## 2 North York                          28
## 3 Scarborough                         43
## 4 Toronto and East York               35
```

```
ggplot(data = district_count_table, mapping =
         aes(x = district, y=num_of_rlightcam))+
  geom_col()+ylim(0,50)+labs(
    title = "Figure A4: Count of Red Light Cameras per District",
       x = "District",
       y = "Number of Red Light Cameras")
```




```
#getting only year from date format
#didn't work, date turned back in to char datatype
attable2 <- attable %>%
  separate(formatted_date, into = c("year","m","d"), sep="-")
attable2
```

```
## # A tibble: 148 x 5
##    district              main              year  m     d
##    <chr>                 <chr>             <chr> <chr> <chr>
```

```
##  1 Toronto and East York RICHMOND ST E     2007  11    09
##  2 Toronto and East York LAKE SHORE BLVD W 2007  11    09
##  3 North York            STEELES AVE W     2007  11    09
##  4 North York            STEELES AVE W     2007  11    09
##  5 Etobicoke York        ALBION RD         2007  11    09
##  6 Etobicoke York        DIXON RD          2008  03    18
##  7 Etobicoke York        STEELES AVE W     2008  10    28
##  8 North York            SHEPPARD AVE W    2009  03    17
##  9 North York            BATHURST ST       2007  11    09
## 10 North York            LAWRENCE AVE W    2008  03    19
## # ... with 138 more rows
```

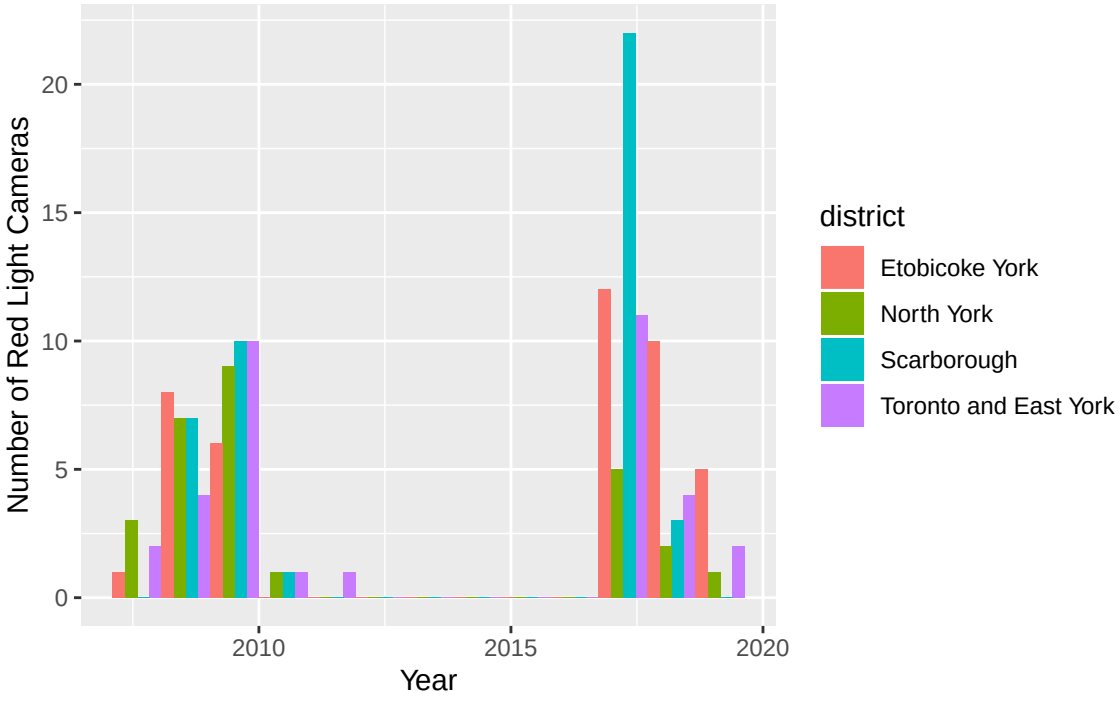
```
attable2 <- attable2 %>% select(-m,-d)
```

```
#count number of occurances of district per year
#gives us number of red light camera installations per district per year
NperYperD <-attable2 %>%
  arrange(district) %>%
  count(district,year)

#count number of installations of redlight cam per year
NperY <- attable2 %>%
  group_by(year)%>%
  count()
```

```
ggplot(data=attable)+
  geom_histogram(aes(x=formatted_date, fill= district),position="dodge",bins = 13)+
  labs(
    title = "Figure A5: Red light camera installation at each district by Year",
       x = "Year",
       y = "Number of Red Light Cameras")
```

Figure A5: Red light camera installation at each district by Year




```
#plot bar graph show installation of redlight cameras over the years per district
#faceted by district for more detail
ggplot(data=NperYperD)+
  geom_col(aes(x=year,y=n, fill= district)) + facet_wrap(~district)+
  theme(axis.text.x = element_text(angle = 90))+labs(
    title = "Figure A6: Red light camera at each district by Year",
       x = "Year",
       y = "Number of Red Light Cameras")
```

Figure A6: Red light camera at each district by Year

Workflow of project, inspired by the Data-Cleaning Brownbags

1. Selected columns of interest
2. tried to make sense of what information the columns hold
3. realized that each red light camera had a unique ID so that those on the same streets could be differentiated.
4. counting the number of main streets and the number of times those names are repeated in the dataset can tell us the number of redlight camera=148. 148 is also the total number of observances in the dataset, as revealed by skimr

5.plotted all 63 streets with their count on a scatter plot. some streets have more redlight cameras than others but too much information in one graph to say much

6.created seperate dataset for streets with more than 3 cameras and plotted those

7. I got the answer of which streets to avoid but how to connect them back to the districts
8. was playing around with arrange, count and group by when I got a weird number of occurances for a dataset:76. 63 was on skimr summary for main, so i knew there were 63 unique values for main. 148 was the total number of rows in the dataset. Where did 76 come from? Started scrolling through the data and saw streets being repeated for different districts: duhhh, major streets cross into other districts. Now I knew I

had to analyse whether there are more red light cameras in a district because there are more of them installed or is it because that district has more major intersections and streets running through them.

9. grouped streets by district and counted the occurance of each main in each district: got back number of redlight cameras per district

10. grouped streets by district and counted the number of occurances of district in the dataset: got back number of main streets present in each district.

11. Now came the question of what else I can do with the data? What kind of summary stats can I include? Asked rohan, He said to do something with the dates. okayyy

12. dates were in char data type so needed to convert to date datatype.

13. used "separate" to detach y-m-d from time, with separator of "T"

14. used lubridate to convert y-m-d from string to date data type

15. created new table wit district, main and the date the cameras were installed

16. created histogram, but histogram x-axis was not detailed enough and google's suggestions weren't helping. There were many districts that had cameras installed in the same year but the scale of the x-axis wasn't providing much detail.Clearly scarborough had a year where many redlight cameras were installed but the scale was too broad to pinpoint the exact year. Realized the dates were "too precise" because the scale was in years but the datapoints were being graphed down to the exact date as well. Thats why the x-axis scale was so broad, there were still too many datapoints

17. decided to graph only by year. but when i tried to separate y-m-d by "-", the year turned into a char. tried using as.Date() to turn it into a date datatype. It didn't work.back to square one

18. Then i thought, why not just count the frequency of the years and graph those. I graphed a bar graph and faceted it by disctrict. Lo and Behold, redlight camera installation per year per district.