Temporal and Spatial Dynamics of Firearm Discharges in Toronto

JSC370 Final Project

Yiteng Zhang

Contents

Introduction

Background and Motivation

The motivation from conducting an analysis on the shooting dataset comes from the fact that people are suffering from more and more unsafe communities around them. Therefore the need for a data-driven analysis is crucial, and can provide valuable insights for relative authorities. This information can be used for these authorities to identify factors to gun violence, and thus enhance the management of Toronto Police, or maybe affect future law enforcement related to gun control. I believe these measures would reduce the risk of gun violence faced by people in Toronto, and improve the overall safety in the city.

Overview

I will be working with the Shooting and Firearm Discharges Open Data dataset provided by Toronto Police Service. The shooting dataset records all reported shooting-related occurrences in Toronto since 2004. By looking at the official data portal provided by Toronto Police service, I come out with the primary questions of the project: When and where do most shooting and firearm discharges occur in Toronto? What is the trend of change for shooting incidents over the years?

Methods

Data

Data Description

The original dataset consists of 6051 rows with 27 columns. Each row in the dataset records an occurrence of shooting event and each column is an attribute of the incident. Some important attributes are:

- OCC_DATE: The specific timestamp of the occurrence of the shooting incident.
- OCC_TIME_RANGE: The time range in a day where the incident occurs. e.x.: Afternoon
- DEATH: The number of deaths counted from the incident.
- INJURIES: The number of injuries counted from the incident.
- NEIGHBOURHOOD_158: Name of Neighbourhood using City of Toronto's new 158 neighbourhood structure.
- LONG_WGS84: Longitude Coordinates.
- LAT_WGS84: Latitude Coordinates.

Data Cleaning and Preprocessing

First, load in the dataset by the fread function in the data.table package. Filter the columns to only keep the columns that are relevant to our analysis, which are id of the incident, the datetime value of the occurrence of the incident, the count of deaths and injuries of the incident, the police division of the location and the coordinates of the incident. Then factor columns with string type, which in this case are OCC_TIME_RANGE, OCC_MONTH, OCC_DOW, NEIGHBOURHOOD_158. The missing values of longitudes and latitudes are recorded as 0, and other missing values are recorded as " " or "NSA", so I filter out rows with longitude and latitude being 0 and mark all "NSA" values to NA. Then Rename all variables in a better way for later analysis. I would like to both consider deaths and injuries of an incident, so I created another column summing up the counts of deaths and injuries. Mutate another column called "season" that represents the month of the occurrence. I would like to look into more positional attributes of related to the location, so I calculate the geographical midpoint of the datapoints and separate the points into NorthEast, NorthWest, SouthWest and SouthEast categories. Factor the two mutated columns as well.

I would like to look into data in each police division as well.

Poisson Regression Model

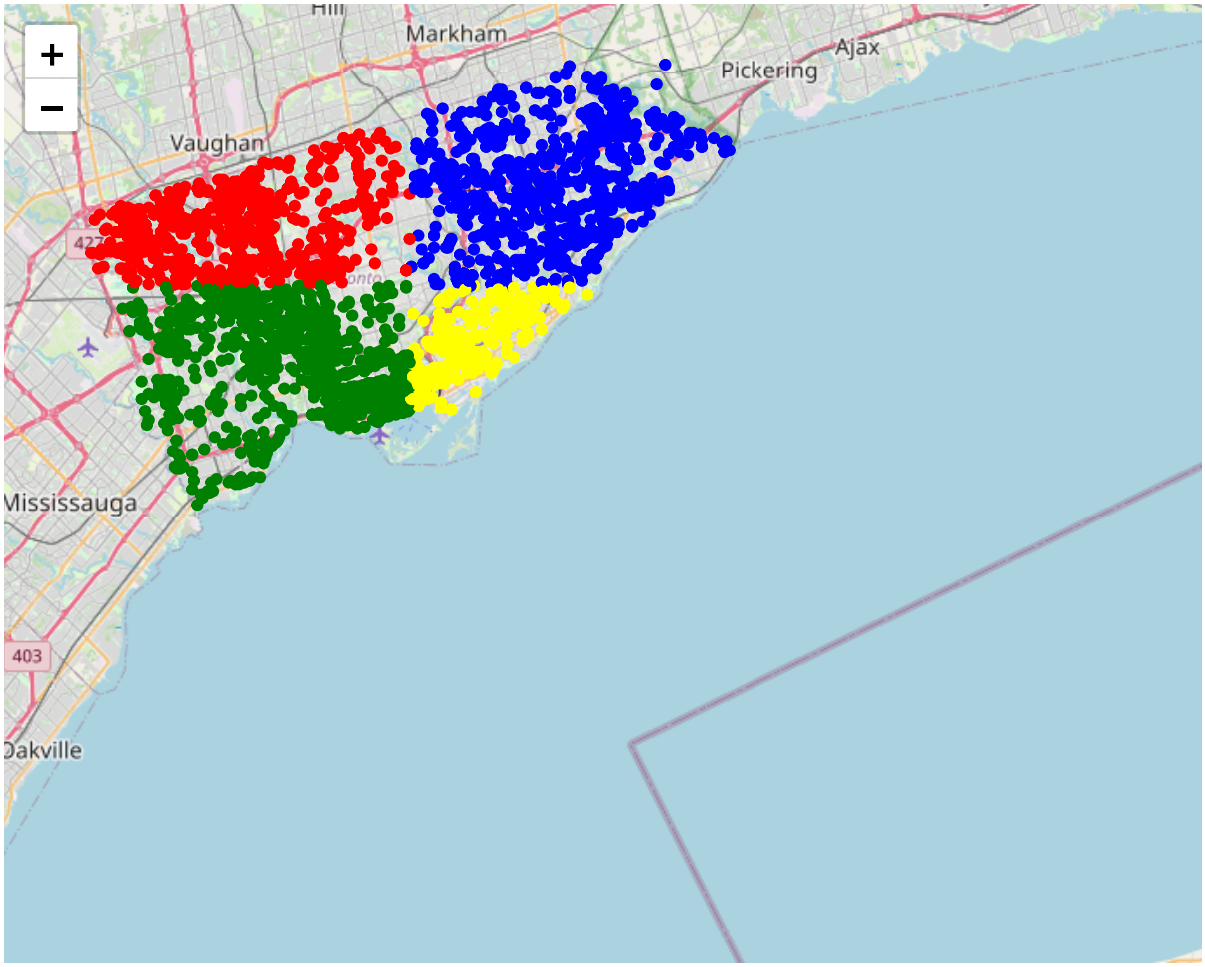
Apart from the EDA above, to investigate the factors influencing the number of total deaths and injuries reported in shooting incidents, I am employing a poisson regression model. The response variable total.deaths.injuries represents the deaths and injuries of shooting events. The predictors are year, position, and day of the week. Then we can write the full model:

$$\log(\lambda_i) = \beta_0 + \beta_1 \times Year_i + \beta_2 \times Position_i + \beta_3 \times DOW_i$$

where λ_i is the expected number of total deaths and injuries for the ith observation. β_0 represents the intercept, β_1, β_2, β_3 are coefficients of year, position, and day of the week, respectively.

Results

Exploratory Data Analysis



First, take a look at how the records are distributed in terms of longitude and latitude on the map. We can see that there are more sparse data points in the middle area of Toronto, probably because the area is less inhabited.

Table 1: Top 10 Divisions with Highest Deaths and Injuries

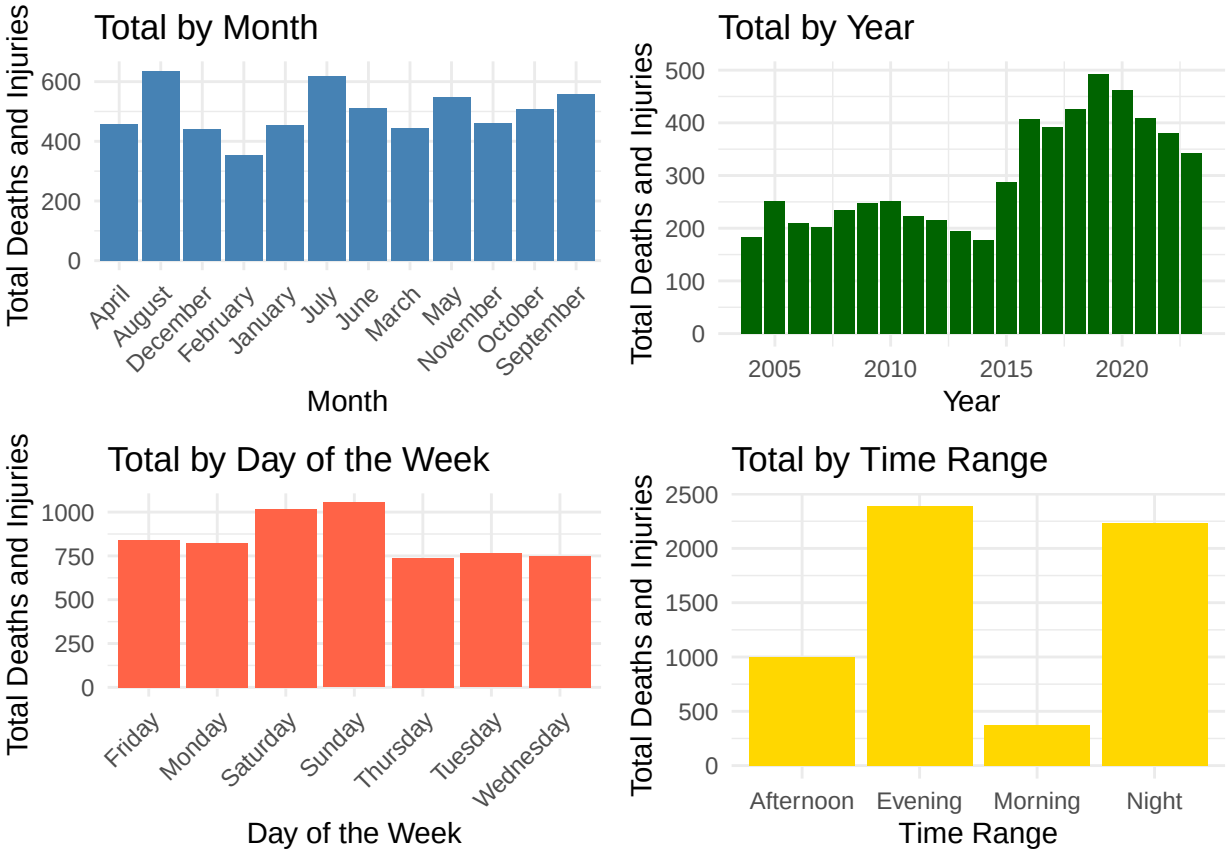
division	total_deaths	total_injuries	total_casualties
D31	109	474	583
D23	82	300	382
D43	59	245	304
D12	59	221	280
D42	66	210	276
D14	46	202	248
D51	56	180	236
D41	32	179	211
D32	33	153	186
D55	36	123	159

From the table we can take a look at the top 10 divisions that have the highest deaths and injuries over the ten years. We can see that division D31 suffers from obviously high casualties, compared even to the second high division D23.

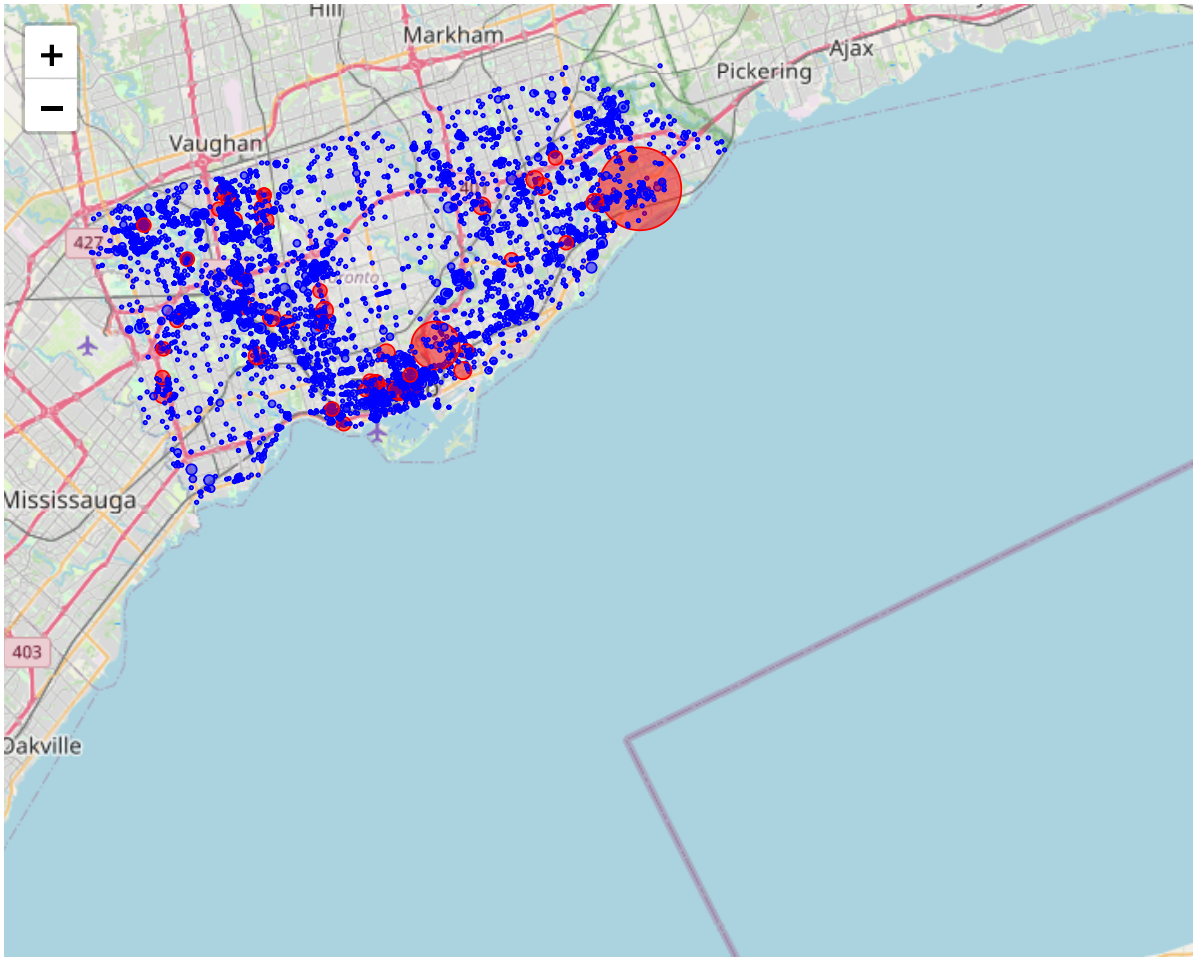
Table 2: Top 10 Neighbourhoods with Highest Deaths and Injuries

neighbourhood	total_deaths	total_injuries	total_casualties
Glenfield-Jane Heights	22	148	170
Mount Olive-Silverstone-Jamestown	32	95	127
Black Creek	25	97	122
West Humber-Clairville	12	79	91
Wellington Place	12	76	88
York University Heights	14	65	79
West Hill	12	66	78
Yorkdale-Glen Park	13	62	75
Kensington-Chinatown	13	60	73
Rockcliffe-Smythe	8	63	71

From this table we can see the top 10 neighbourhoods with highest deaths and injuries. Notice that the top 3 neighbourhoods have relatively high numbers compared to the rest in the table.



Here are four histograms that show the distribution of total deaths and injuries by month, year, day of week and time range. 1. We can see that over the ten years, the highest month of deaths and injuries is August, and the lowest month is February. 2. Accounting for year, this number shows a significant increase from the year 2015, reaching the peak value of almost 500 at the year 2019, then experience a decrease to the year of 2024. Before 2015, it shows a steady trend with a slight decrease after 2010. 3. The barplot of total deaths and injuries by day of week shows that the numbers are approximately evenly distributed during the weekdays, however the numbers are higher during the weekends. 4. For the barplot of total deaths and injuries by time range, we can see that total deaths and injuries are higher during evening and at night, and are lower during afternoon and morning.



This is a leaflet map of spatial information of total deaths and injuries over the ten years. Here, I define high

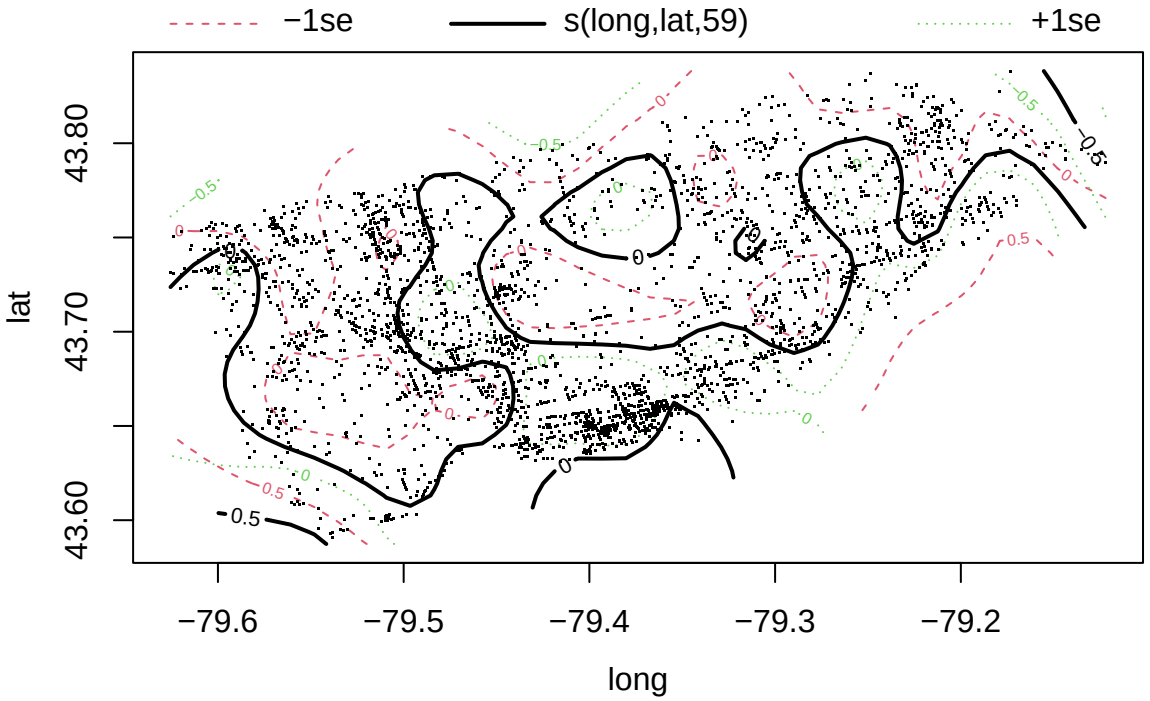
casualties to be the number of casualties to be greater than 3, and low casualties to be <= 3. The larger radius means larger number of casualties in that shooting event. We can see that an overall trend is that shooting events are more distributed along the highways. Also there are more shooting events with high and low causualties in the downtown area and scarbough area.

Preliminary Results

Table 3: Summary of GLM Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	66.8127103	5.7791860	11.5609205	0.0000000
year	-0.0334422	0.0028694	-11.6548488	0.0000000
positionNW	-0.0559572	0.0454016	-1.2324937	0.2177647
positionSE	0.1025665	0.0581759	1.7630422	0.0778934
positionSW	0.0964601	0.0438411	2.2002211	0.0277912
dowMonday	0.0683455	0.0628165	1.0880190	0.2765867
dowSaturday	0.0667278	0.0598143	1.1155833	0.2646006
dowSunday	0.0959950	0.0589483	1.6284599	0.1034274
dowThursday	-0.0440068	0.0668169	-0.6586183	0.5101409
dowTuesday	-0.0260539	0.0656168	-0.3970609	0.6913225
dowWednesday	-0.0669653	0.0667969	-1.0025202	0.3160924

From the coefficients summary, we can see that the predictor position and dow is not significant in the model. Therefore the results are not so satisfiying.



From the plot we can see that the black line is the estimated value of response, the dotted lines represent confidence interval within one standard error.

Table 4: Tidy Summary of GAM Model

term	edf	ref.df	statistic	p.value
s(long,lat)	59	59	1.810133	0.0001553

From the summary of the coefficients we can see that the smooth term s(long,lat) has an estimated degrees of freedom of 59 with high significance, indicating a complex smooth with a lot of flexibility to fit the data. Therefore it suggests that there is a spatial structure in the data that the smooth term is capturing. Also the coefficient of intercept is also significant. However, there is still much space for improvement of the model because R^2 is 0.00792, explaining only a small portion of the deviance.

Conclusion

In conclusion, the project aims to analyze the spatial and timely distribution of deaths and injuries caused by firearm shooting events in Toronto over the past ten years. We used a public dataset provided by Toronto Police that offers the exact same content. For analyzing methods, we utilized several visualization techniques including leaflet plots, barplots and histograms, as well as poisson regression model and spline regression model to model the space and time structure that might affect deaths and injuries in shooting events.

Considering time, the value seems to be higher during the evening and night in a day, higher during the weekend. This value doesn't seem to be so related with month, but showed a significant increase each year after 2015. Considering space, we find out that for police divisions, the most casualties happen in D31. For neighbourhoods, the most casualties happen in Glenfield-Jane Heights. Also there are more shooting events with relatively higher casualty level beside some of the main highways of Toronto. Also downtown and scarbough have higher number of shooting events than other areas.

For the poisson regression model, only the year predictor is significant in the model, the result tend not to be significant. So I tried the spline regression model to capture the spatial information in the map of Toronto. The model is significant and captures some of the spatial information, but only a small portion of it.

Therefore my two research questions are answered by the above analysis.

Showcasing plots

Figure 1

Figure 2

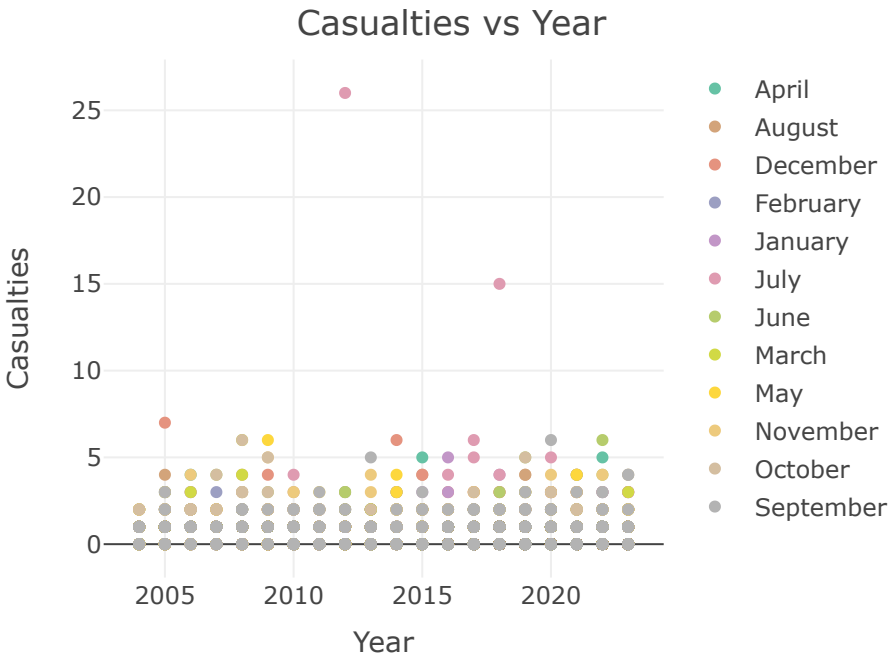
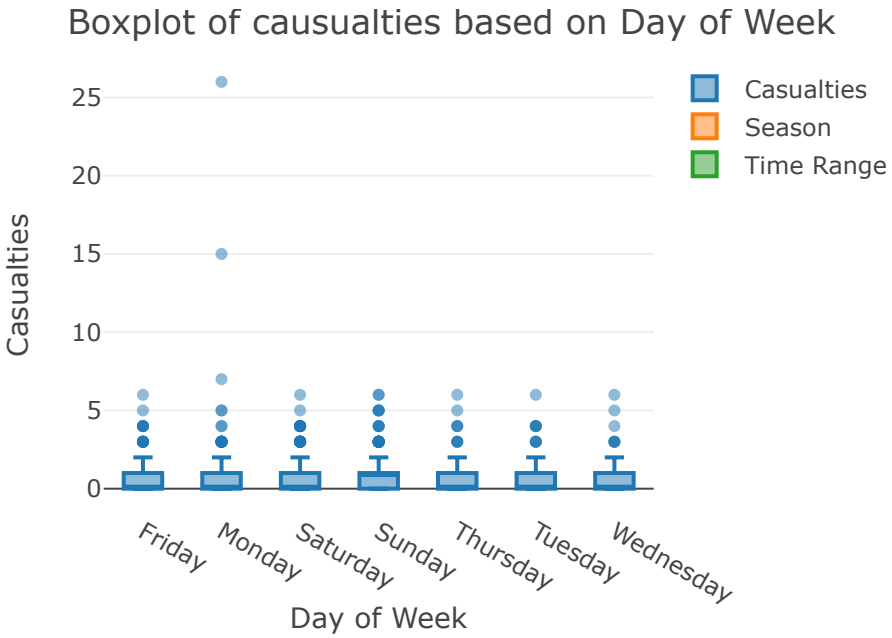


Figure 3



Resources

https://data.torontopolice.on.ca/datasets/TorontoPS::shooting-and-firearm-discharges-open-data/about