SSC 442 - Final Project (With R Code)

Peter Fu Chen

2023-04-21

NYC Traffic Accidents Analysis (policy conclusion project)

Background and Motivation

The focus of this project is the examination of motor vehicle collisions that were reported by the New York City Police Department from January to August of 2020. Each record in the dataset, represents a unique collision and includes various details such as the date, time, and location of the accident, as well as additional data.

The analysis will compare the percentage of total accidents by month, providing a snapshot of the trend over time. These statistics will be visually represented through the use of maps, which will demonstrate the frequency of accidents in different boroughs of New York City. Furthermore, the analysis will also determine the most common streets, days, and times when accidents are likely to occur.

The ultimate goal of this project is to provide recommendations to the city of New York on how this analysis can be used to prevent future accidents. By highlighting the most common accident hotspots, times, and contributing factors, city planners and law enforcement officials will be equipped with the information necessary to implement effective safety measures and reduce the number of accidents in the city.

Methodology

Loading libraries

```
library(tidyverse)
library(ggthemes)
library(ggmap)
library(ggrepel)
library(lubridate)
library(lmtest)
library(sandwich)
library(AER)
```

Data preparation

```
# First of all, that's read the data and take a glimpse what it contains
df = read.csv("data.csv")
```

```r
# Check for missing values
sum(is.na(df))
```

```
## [1] 37643
```

```r
# Drop all the NAs
df_withoutna = na.omit(df)

# We also need to drop LATITUDE equal 0, otherwise it will cause problems when we plot
df_withoutna =  df_withoutna[df_withoutna$LATITUDE != 0,]

#Check if all the NAs are droped
sum(is.na(df_withoutna))
```

```
## [1] 0
```

```r
# Use lubridate() to change the date format
df_withoutna$CRASH.DATE = mdy(df_withoutna$CRASH.DATE)
```

```r
# Drop some unnecessary columns
df_withoutna = df_withoutna %>% select(1:18)
```

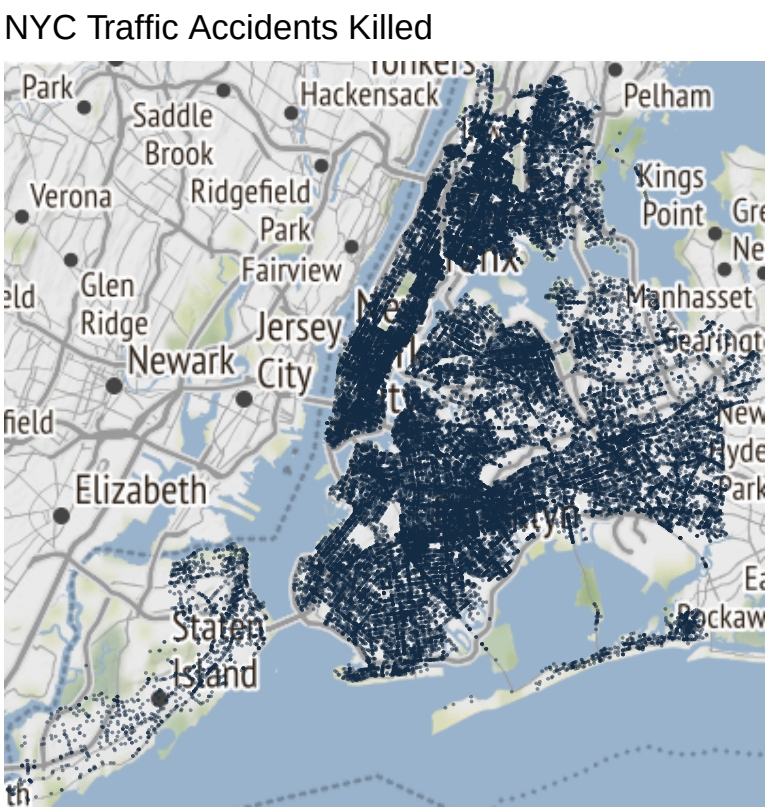
Map plotting the people killed by motor vehicle in NYC

The number of fatalities resulting from motor vehicle accidents across New York City in 2020 was visualized using a map. Each dot on the map represents a death, and it's evident that the concentration of these dots varies among different boroughs. Manhattan, the Bronx, and Brooklyn appear to have the highest density of dots, indicating a higher rate of fatalities in these areas. This information can help policymakers target their interventions to specific locations to reduce the number of motor vehicle-related fatalities in New York City.

```
# Plot a map showing the locations of accidents where people were killed

qmplot(data = df_withoutna, x = LONGITUDE, y = LATITUDE, maptype = "terrain",
      darken = 0, geom = "auto", color = NUMBER.OF.PERSONS.KILLED ,
      alpha=I(.5), size = I(0.0000001),
      zoom = 10,extent = "panel",f = 0.005,
      xlab = "", ylab = "", main = "NYC Traffic Accidents Killed") +
      theme(axis.ticks = element_blank(), axis.text = element_blank()) +
      theme(legend.position="none")
```

```
## i Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.
```

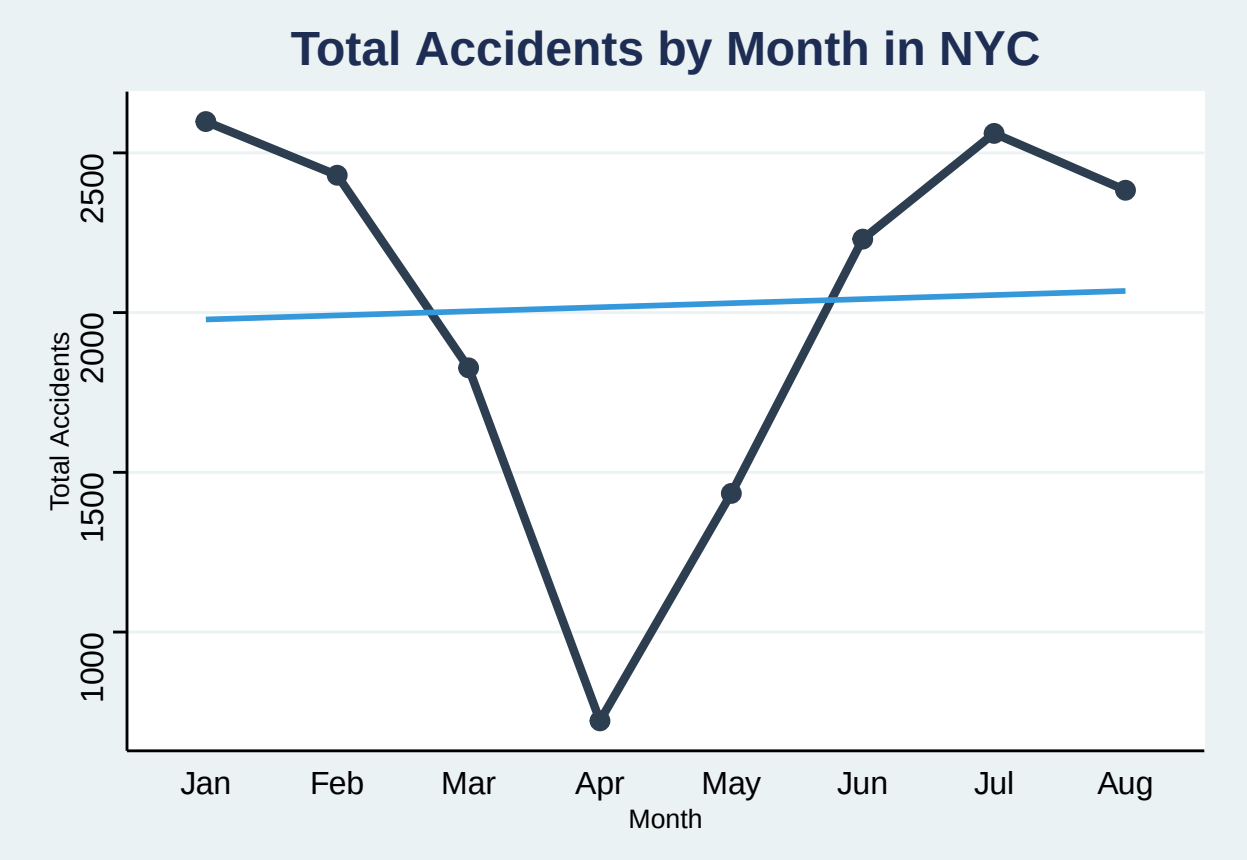


Accidents Per Month

```
month_accidents = df_withoutna %>% group_by(month(CRASH.DATE)) %>%
  summarise(sum(NUMBER.OF.PERSONS.INJURED , NUMBER.OF.PERSONS.KILLED)) %>%
  rename("Month" = "month(CRASH.DATE)") %>%
  rename("Total_accidents" = "sum(NUMBER.OF.PERSONS.INJURED, NUMBER.OF.PERSONS.KILLED)")

ggplot(month_accidents, aes(x = Month, y = Total_accidents)) +
  geom_line(size = 1.5, color = "#2c3e50") +
  geom_point(size = 3, color = "#2c3e50") +
  geom_smooth(method = "lm", se = FALSE, color = "#3498db", fullrange = TRUE) +
  scale_x_discrete(limits = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug")) +
  xlab("Month") + ylab("Total Accidents") +
  ggtitle("Total Accidents by Month in NYC") +
  theme_stata() +
  theme(legend.position = "none",
        axis.text = element_text(size = 12),
        plot.title = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 12),
        axis.ticks.x = element_blank())
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



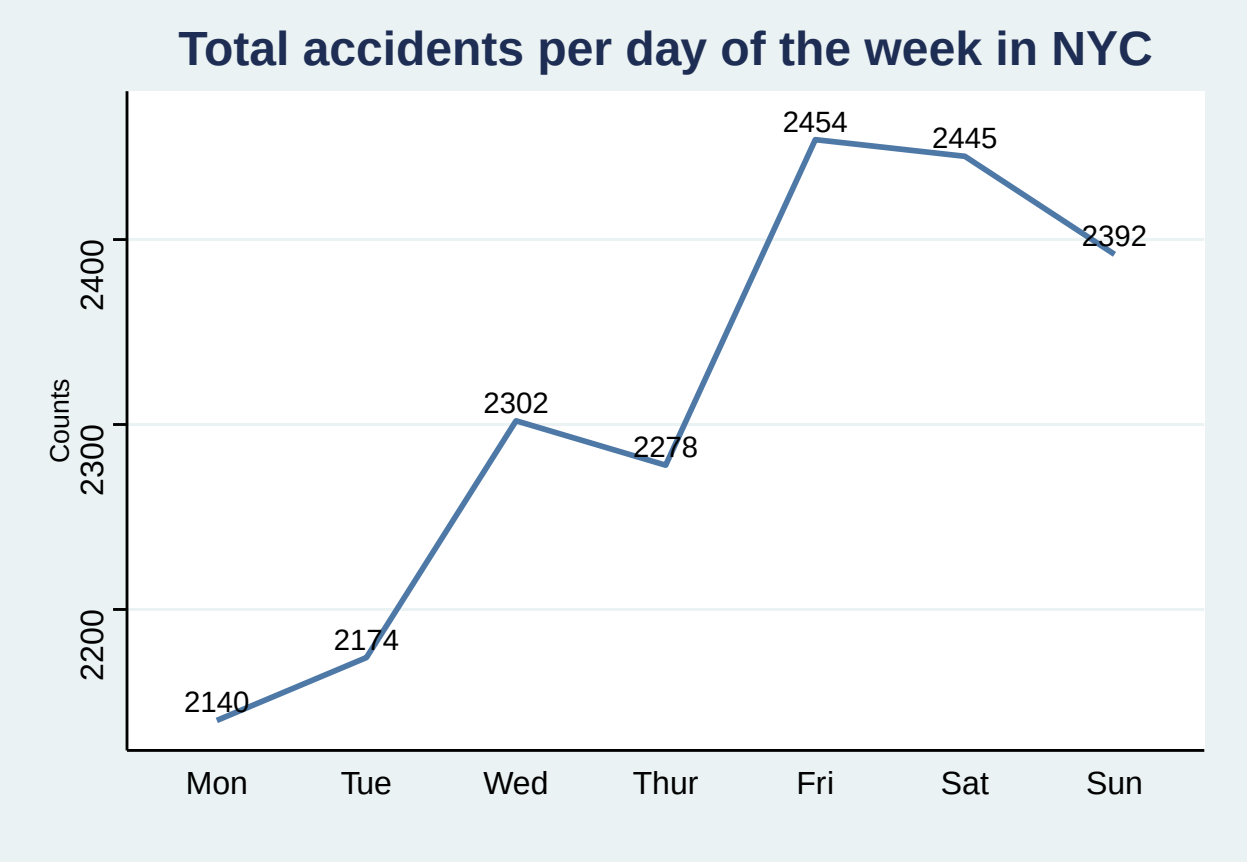
The line chart illustrates the total number of accidents per month in NYC from January to August 2020. April had the lowest number of accidents, with a total of 11,234, which was significantly lower than the

previous month, March, which had 15,315 accidents. Several factors may have contributed to this decrease, such as the COVID-19 lockdown and weather conditions. It's important to note that this decrease may not reflect a long-term trend, and analyzing the types and factors of accidents may provide further insight.

Accidents Per day

```
day_accidents = df_withoutna %>%
  mutate(dotw = wday(CRASH.DATE)) %>%
   group_by(dotw) %>%
   summarise(total_accidents = sum(sum(NUMBER.OF.PERSONS.INJURED , NUMBER.OF.PERSONS.KILLED)))


ggplot(day_accidents, aes(x = dotw, y = total_accidents, color = total_accidents)) +
  geom_line(size = 1, color = "#4e79a7") +
  scale_x_discrete(limits = c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun")) +
  ggtitle("Total accidents per day of the week in NYC") +
  xlab("") +
  ylab("Counts") +
  geom_text(aes(label = total_accidents), nudge_y = 10, size = 4, color = "black") +
  theme_stata() +
  theme(legend.position = "none",
        axis.text = element_text(size = 12),
        plot.title = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 12),
        axis.ticks.x = element_blank())
```



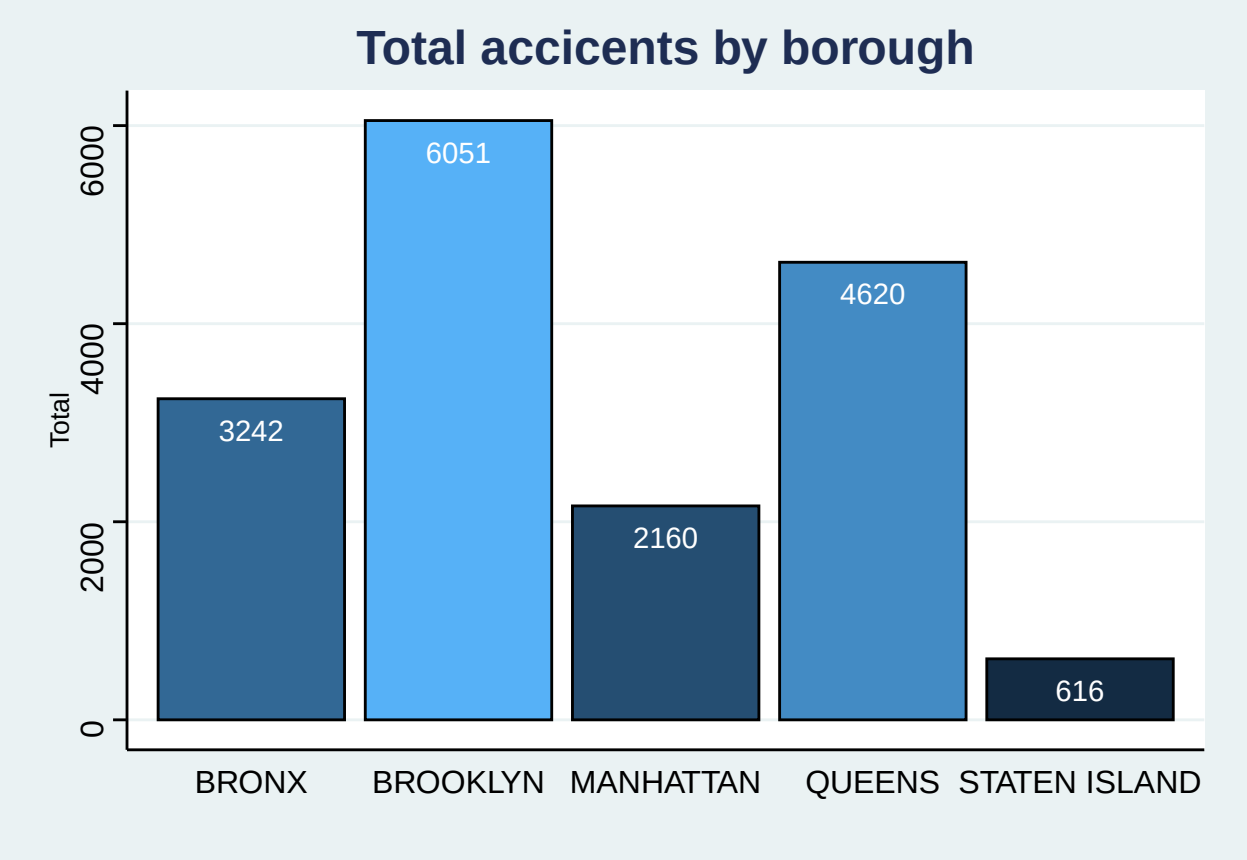
This plot above shows the total number of accidents per day of the week in NYC. We observed that Friday and Saturday had a slightly higher number of accidents, but this difference was not significant enough. There could be various reasons for this, including increased weekend traffic, alcohol consumption, or time of day. Further analysis is needed to understand the underlying factors that contribute to this pattern.

Total injured and killed per borough

```
total_borough = df %>%
  group_by(BOROUGH) %>%
  summarise(accidents = n(),
            injuries = sum(NUMBER.OF.PERSONS.INJURED),
            deaths = sum(NUMBER.OF.PERSONS.KILLED))
```

BOROUGH	accidents	injuries	deaths
	25741	10831	72
BRONX	9417	3232	10
BROOKLYN	16907	6024	27
MANHATTAN	7353	2151	9
QUEENS	14017	4600	20
STATEN ISLAND	1446	610	6

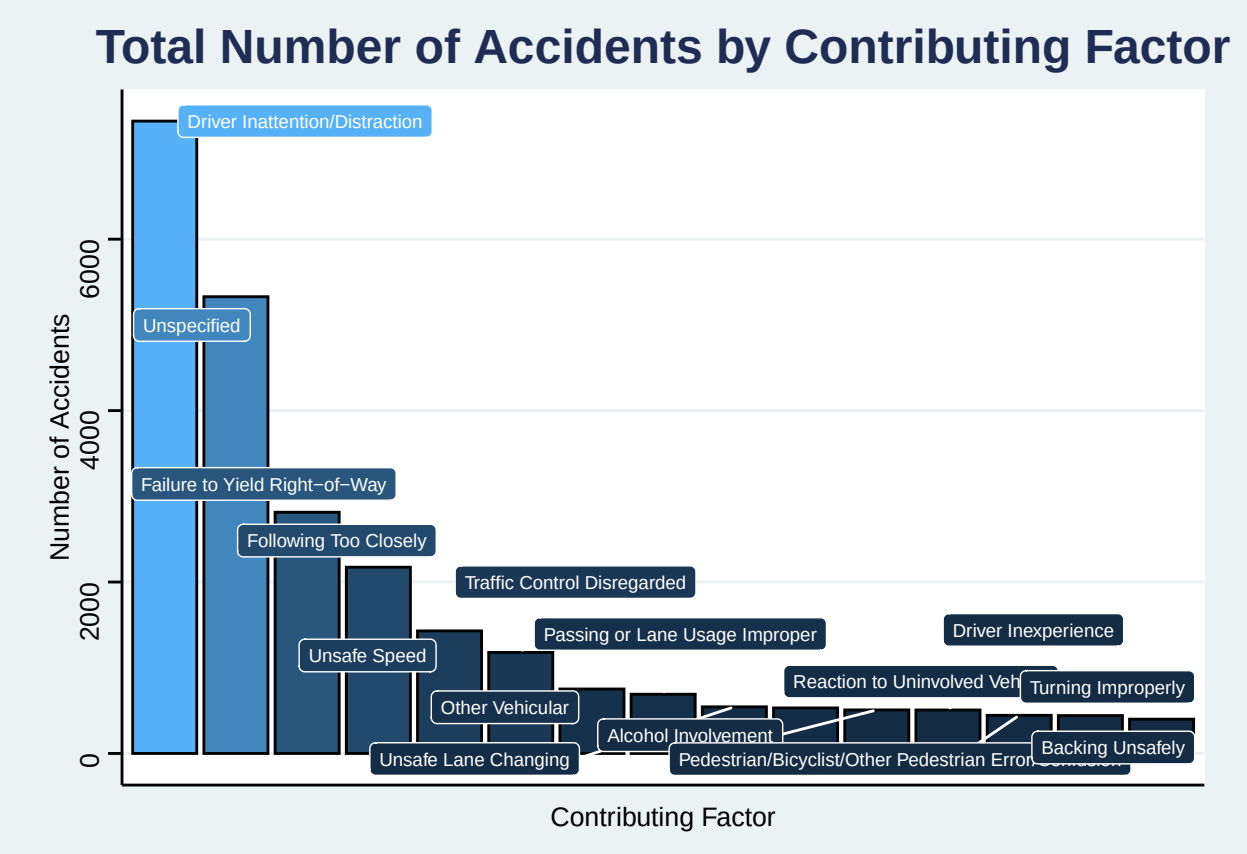
```r
df %>% group_by(BOROUGH) %>%
  summarise(injuries = sum(NUMBER.OF.PERSONS.INJURED),
            deaths = sum(NUMBER.OF.PERSONS.KILLED),
            total = sum(injuries, deaths)) %>%
  filter(row_number() != 1) %>%
  ggplot(aes(x = BOROUGH, y = total, fill = total)) +
  geom_bar(stat = "identity", colour = "black") +
  ggtitle("Total accicents by borough") +
  xlab("") +
  ylab("Total") +
  geom_text(aes(label = total), position = position_dodge(0.9), vjust = 2, size = 4, color = "#ffffff") +
  theme_stata() +
  theme(legend.position = "none",
        axis.text = element_text(size = 12),
        plot.title = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 12),
        axis.ticks.x = element_blank())
```



This analysis examines the total number of accidents, injuries, and deaths by borough in NYC. Brooklyn had the highest number of accidents, followed by Queens. Manhattan had a lower number of accidents, which could be due to differences in traffic patterns, population density, or road infrastructure. External factors such as weather, road conditions, or driver behavior may have also influenced the number of accidents in each borough.

Total Number of Accidents by Contributing Factor

```
df %>% group_by(CONTRIBUTING.FACTOR.VEHICLE.1) %>%
  summarise(injuries = sum(NUMBER.OF.PERSONS.INJURED),
            deaths = sum(NUMBER.OF.PERSONS.KILLED),
            total = sum(injuries, deaths)) %>%
  filter(row_number() != 1) %>%
  arrange(desc(total)) %>%
  slice(1:15) %>%
  ggplot(aes(x = reorder(CONTRIBUTING.FACTOR.VEHICLE.1, total, FUN = desc), y = total, fill = total)) +
  geom_bar(stat = "identity",color = "black") +
  ggtitle("Total Number of Accidents by Contributing Factor") +
  xlab("Contributing Factor") +
  ylab("Number of Accidents") +
  # geom_text(aes(label = total), position = position_dodge(0.9), vjust = 2, size = 4, color = "#ffffff
  geom_label_repel(aes(label = CONTRIBUTING.FACTOR.VEHICLE.1), color = "white", size = 2.5, max.overlaps
  theme_stata() +
  theme(axis.text.x = element_blank(),
        plot.title = element_text(size = 18, face = "bold"),
        legend.text = element_text(size = 12),
        legend.position = "none",
        axis.ticks.x = element_blank())
```



The bar chart shows the total number of accidents by contributing factor, with distraction being the highest followed by failed to yield and others. To reduce the number of accidents, policies and interventions can be

implemented such as public awareness campaigns, enforcement of traffic laws, infrastructure improvements, and driver education.

Conclusion

This project analyzed the 2020 New York City motor vehicle accidents dataset to gain insights into the frequency, distribution, and causes of accidents. Although our analysis of the New York City motor vehicle accidents dataset from 2020 identified patterns and trends in accident frequency, distribution, and causes, there are several caveats to consider. While certain factors were associated with a higher probability of accidents, it is challenging to establish a causal relationship between these factors and accidents. Additionally, as a policy conclusion project, our model's accuracy is limited. However, our findings provide policymakers, law enforcement, and public safety officials with valuable information to develop targeted interventions and strategies for how and when to reduce the number of accidents in the city.