SSC 442 - Final Project

Peter Fu Chen , Mike Liu

2023-03-19

NYC Traffic Accidents Analysis

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("NYC Accidents 2020.csv")
glimpse(df)
```

```
## Rows: 74,881
## Columns: 29
## $ CRASH.DATE                    <chr> "8/29/20", "8/29/20", "8/29/20", "8/29/2~
## $ CRASH.TIME                    <chr> "15:40:00", "21:00:00", "18:20:00", "0:0~
## $ BOROUGH                       <chr> "BRONX", "BROOKLYN", "", "BRONX", "BROOK~
## $ ZIP.CODE                      <int> 10466, 11221, NA, 10459, 11203, NA, 1045~
## $ LATITUDE                      <dbl> 40.89210, 40.69050, 40.81650, 40.82472, ~
## $ LONGITUDE                     <dbl> -73.83376, -73.91991, -73.94656, -73.892~
## $ LOCATION                      <chr> "POINT (-73.83376 40.8921)", "POINT (-73~
## $ ON.STREET.NAME                <chr> "PRATT AVENUE", "BUSHWICK AVENUE", "8 AV~
## $ CROSS.STREET.NAME             <chr> "STRANG AVENUE", "PALMETTO STREET", "", ~
## $ OFF.STREET.NAME               <chr> "", "", "", "1047 SIMPSON STREET", "4609~
## $ NUMBER.OF.PERSONS.INJURED     <int> 0, 2, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0~
## $ NUMBER.OF.PERSONS.KILLED      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.PEDESTRIANS.INJURED <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
## $ NUMBER.OF.PEDESTRIANS.KILLED  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.CYCLIST.INJURED     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.CYCLIST.KILLED      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.MOTORIST.INJURED    <int> 0, 2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.MOTORIST.KILLED     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CONTRIBUTING.FACTOR.VEHICLE.1 <chr> "Passing Too Closely", "Reaction to Unin~
## $ CONTRIBUTING.FACTOR.VEHICLE.2 <chr> "Unspecified", "Unspecified", "", "Unspe~
## $ CONTRIBUTING.FACTOR.VEHICLE.3 <chr> "", "", "", "Unspecified", "", "", "", "~
## $ CONTRIBUTING.FACTOR.VEHICLE.4 <chr> "", "", "", "Unspecified", "", "", "", "~
## $ CONTRIBUTING.FACTOR.VEHICLE.5 <chr> "", "", "", "", "", "", "", "", "", "", ~
## $ COLLISION_ID                  <int> 4342908, 4343555, 4343142, 4343588, 4342~
## $ VEHICLE.TYPE.CODE.1           <chr> "Sedan", "Sedan", "Station Wagon/Sport U~
## $ VEHICLE.TYPE.CODE.2           <chr> "Station Wagon/Sport Utility Vehicle", "~
## $ VEHICLE.TYPE.CODE.3           <chr> "", "", "", "Sedan", "", "", "", "", "Se~
## $ VEHICLE.TYPE.CODE.4           <chr> "", "", "", "Motorcycle", "", "", "", ""~
## $ VEHICLE.TYPE.CODE.5           <chr> "", "", "", "", "", "", "", "", "", "", ~
```

```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)
glimpse(df_withoutna)
```

```
## Rows: 47,686
## Columns: 18
## $ CRASH.DATE                    <date> 2020-08-29, 2020-08-29, 2020-08-29, 202~
## $ CRASH.TIME                    <chr> "15:40:00", "21:00:00", "0:00:00", "17:1~
## $ BOROUGH                       <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN"~
## $ ZIP.CODE                      <int> 10466, 11221, 10459, 11203, 10459, 10466~
## $ LATITUDE                      <dbl> 40.89210, 40.69050, 40.82472, 40.64989, ~
## $ LONGITUDE                     <dbl> -73.83376, -73.91991, -73.89296, -73.933~
## $ LOCATION                      <chr> "POINT (-73.83376 40.8921)", "POINT (-73~
## $ ON.STREET.NAME                <chr> "PRATT AVENUE", "BUSHWICK AVENUE", "", "~
## $ CROSS.STREET.NAME             <chr> "STRANG AVENUE", "PALMETTO STREET", "", ~
## $ OFF.STREET.NAME               <chr> "", "", "1047 SIMPSON STREET", "4609 SNY~
## $ NUMBER.OF.PERSONS.INJURED     <int> 0, 2, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 2~
## $ NUMBER.OF.PERSONS.KILLED      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.PEDESTRIANS.INJURED <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ NUMBER.OF.PEDESTRIANS.KILLED  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.CYCLIST.INJURED     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.CYCLIST.KILLED      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ NUMBER.OF.MOTORIST.INJURED    <int> 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2~
## $ NUMBER.OF.MOTORIST.KILLED     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

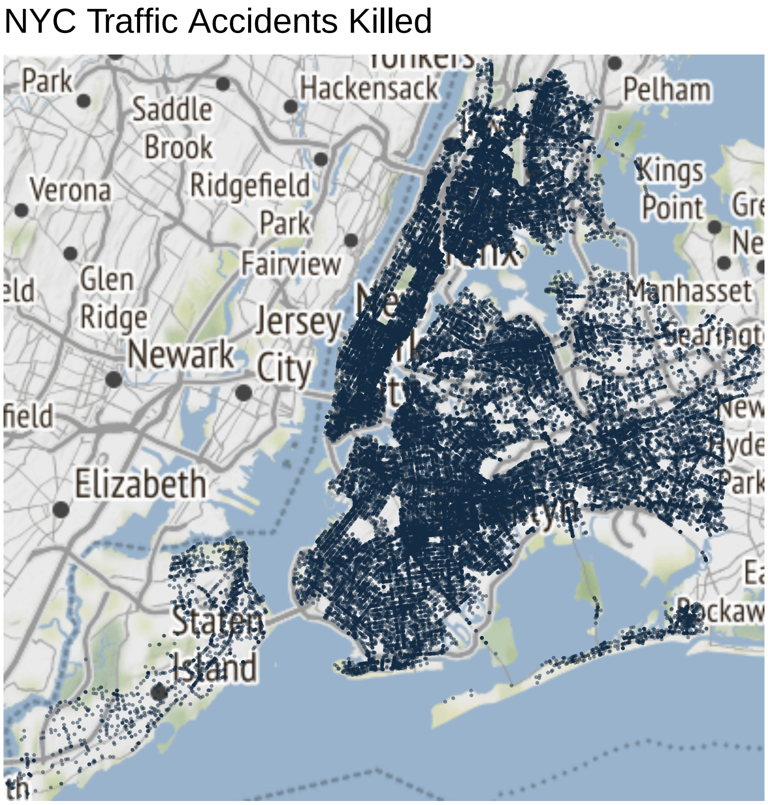
Map plotting the people killed by motor vehicle in NYC

First and foremost, let's examine the number of fatalities resulting from motor vehicle accidents across New York City in 2020. To represent each death, we have used dots on a map and it's evident that the concentration of these dots varies among different boroughs. Upon initial inspection, it appears that Manhattan, the Bronx, and Brooklyn have the highest density of dots, indicating a higher rate of fatalities in these areas.

Next, we will employ various analytical techniques to determine which borough is experiencing the most severe problem and to understand how public policy can be utilized to address this issue. By analyzing the data and identifying trends, we can work towards developing effective strategies that can help reduce the number of motor vehicle-related fatalities in New York City.

```
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.
```



Regression analysis: Model the relationship between the number of accidents and borough.

```
# Summarize the model
myReg = lm(NUMBER.OF.PERSONS.INJURED ~ as.factor(BOROUGH), data = df_withoutna)
coeftest(myReg, vcov = vcovHC(myReg, "HC1"))
```

```
##
## t test of coefficients:
##
##                                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                      0.3412076  0.0072879 46.8184 < 2.2e-16 ***
## as.factor(BOROUGH)BROOKLYN       0.0142173  0.0091336  1.5566 0.1195722
## as.factor(BOROUGH)MANHATTAN     -0.0470403  0.0100631 -4.6745 2.955e-06 ***
## as.factor(BOROUGH)QUEENS        -0.0131560  0.0092557 -1.4214 0.1552080
## as.factor(BOROUGH)STATEN ISLAND  0.0882888  0.0233044  3.7885 0.0001517 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We use the linear regression model that is used to predict the number of persons injured in a traffic accident based on the borough in which the accident occurred. The borough is treated as a categorical variable with five levels (Bronx, Brooklyn, Manhattan, Queens, and Staten Island).

In this model, Bronx is the base level, also known as omitted level.

The coefficients of the model are shown in the table produced by the coeftest() function with heteroscedasticity-consistent standard errors computed using the vcovHC() function.

The first coefficient is the intercept, which represents the expected number of persons injured when the accident occurs in the Bronx. The estimate of the intercept is 0.3412076, which means that on average, the number of persons injured in a traffic accident in the Bronx is 0.3412076.

The next four coefficients are the differences between the expected number of persons injured in each of the other boroughs and the expected number in the Bronx. For example, the coefficient for Brooklyn (0.0142173) means that, on average, the number of persons injured in a traffic accident in Brooklyn is 0.0142173 higher than the number in the Bronx, holding all other variables constant (ceteris paribus).

The t-values and p-values associated with each coefficient indicate whether each coefficient is statistically significant. The null hypothesis for each t-test is that the coefficient is equal to zero, and the alternative hypothesis is that the coefficient is different from zero. A p-value less than 0.05 suggests that the coefficient is statistically significant at the 5% level. In this model, the intercept and the coefficients for Manhattan and Staten Island are statistically significant, with p-values less than 0.05. The coefficients for Brooklyn and Queens are not statistically significant, with p-values greater than 0.05. This suggests that there is not enough evidence to conclude that the number of persons injured in a traffic accident differs between the Bronx and Queens or Brooklyn, after adjusting for other variables. However, there is evidence to suggest that the number of persons injured in a traffic accident is lower in Manhattan and higher in Staten Island compared to the Bronx, after adjusting for other variables.

Do we need heteroskedasticity-consistent errors?

We might want to use Breusch-Pagan test to see if we need heteroskedasticity for this model.

```
df_withoutna$uhat = residuals(myReg)
df_withoutna$uhat2 = (df_withoutna$uhat)^2
```

```
# Estimate a linear regression model with uhat2 as the dependent variable
bpReg = lm(NUMBER.OF.PERSONS.INJURED ~ as.factor(BOROUGH), data = df_withoutna)
summary(bpReg)
```

```
## 
## Call:
## lm(formula = NUMBER.OF.PERSONS.INJURED ~ as.factor(BOROUGH), 
##     data = df_withoutna)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.4295 -0.3554 -0.3281  0.6446 14.6446 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                       0.341208   0.007143  47.770  < 2e-16 ***
## as.factor(BOROUGH)BROOKLYN        0.014217   0.008889   1.599    0.110    
## as.factor(BOROUGH)MANHATTAN      -0.047040   0.010777  -4.365 1.27e-05 ***
## as.factor(BOROUGH)QUEENS         -0.013156   0.009215  -1.428    0.153    
## as.factor(BOROUGH)STATEN ISLAND   0.088289   0.019584   4.508 6.55e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.6798 on 47681 degrees of freedom
## Multiple R-squared:  0.001438,   Adjusted R-squared:  0.001354 
## F-statistic: 17.16 on 4 and 47681 DF,  p-value: 4.481e-14
```

We get F-statistic: 17.16 on 4 and 47681 DF, p-value: 4.481e-14

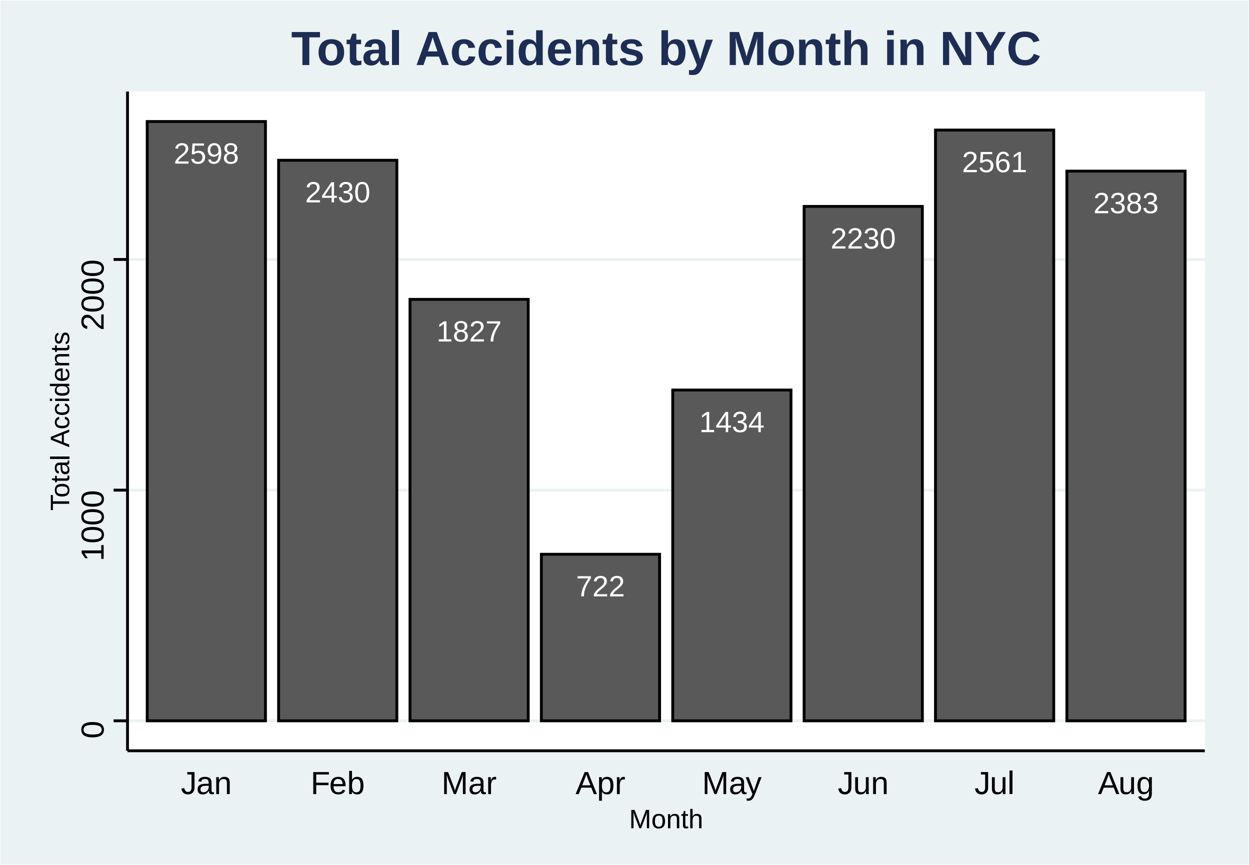
In this case, the Breusch-Pagan test statistic is F=17.16, with 4 and 47681 degrees of freedom, and a very small p-value of 4.481e-14. This indicates strong evidence against the null hypothesis of homoscedasticity (i.e., constant variance of the errors) and suggests that there is heteroscedasticity in the model.

In other words, the variance of the errors in the regression model is not constant across all levels of the independent variable (borough), which means that the residuals are not equally distributed. Therefore, the standard errors of the coefficients may not be accurate, and the results of the t-tests may not be reliable. To correct for heteroscedasticity, we could consider using a robust standard error estimator, such as the HC1 estimator, which we used in our regression model. This estimator can provide more accurate standard errors and t-statistics even in the presence of heteroscedasticity.

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, color = Total_accidents)) +
  geom_bar(stat = "identity", position = "dodge", colour = "black") +
  scale_x_discrete(limits = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug")) +
  geom_text(aes(label = Total_accidents), position = position_dodge(0.9), vjust = 2, size = 4, color =
  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())
```



After we obtained the visualization of total accidents by month in NYC in 2020. We found that April had the lowest number of accidents per month, there could be several reasons for this:

- COVID-19 lockdown: One of the most significant factors that could have contributed to a decrease in the number of accidents is the COVID-19 lockdown in New York City, which began in March 2020.

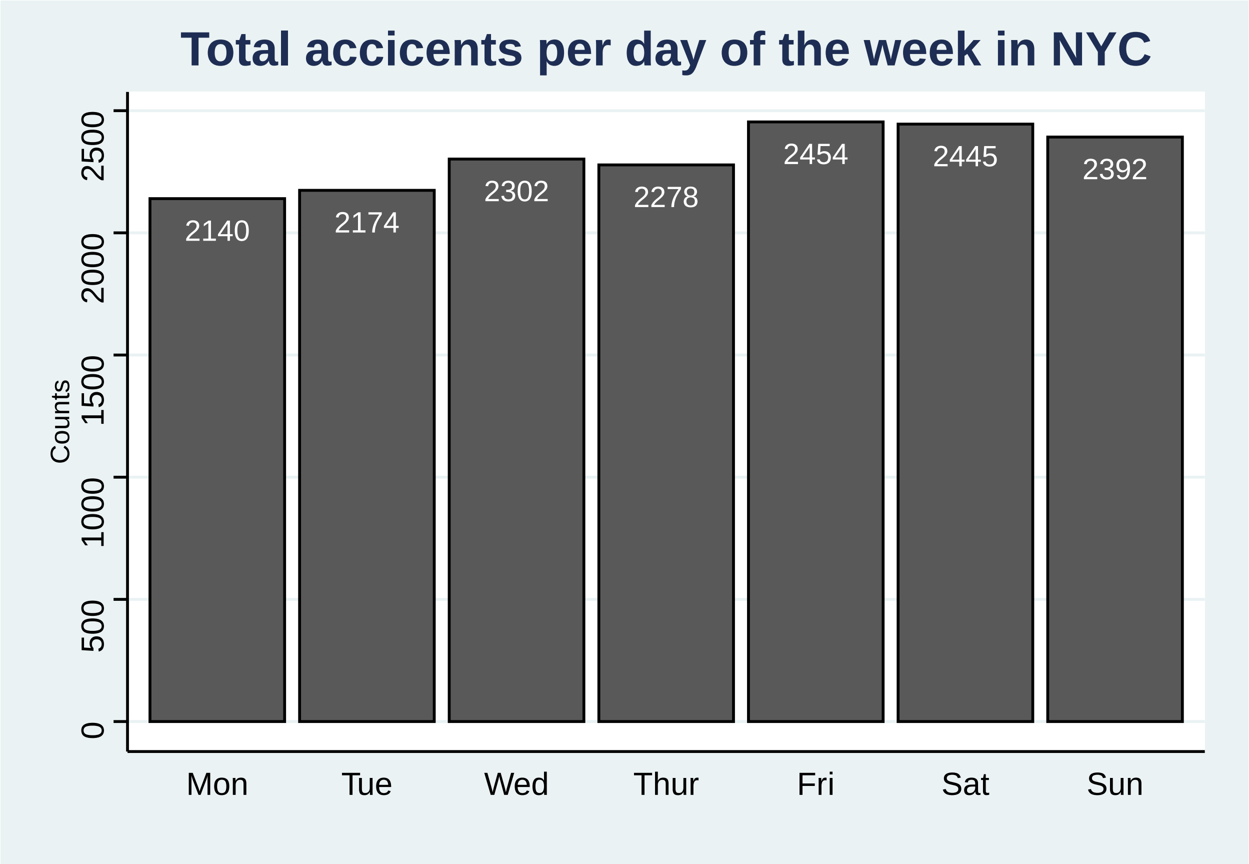
The lockdown resulted in many people staying at home, which would have led to a decrease in the number of cars on the road and a lower probability of accidents.

- Weather: Another possible factor that could have contributed to a decrease in the number of accidents in April is the weather. April is typically a month with moderate temperatures in New York City, and the weather conditions may have been less conducive to accidents.
- Other factors: There could be several other factors that contributed to the decrease in the number of accidents in April. For example, there could have been changes in traffic patterns, road closures, or road maintenance activities that reduced the number of accidents.

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_bar(stat = "identity", colour = "black") +
  scale_x_discrete(limits = c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun")) +
  ggtitle("Total accicents per day of the week in NYC") +
  xlab("") +
  ylab("Counts") +
  geom_text(aes(label = total_accidents), position = position_dodge(0.9), vjust = 2, size = 4, color =
  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())
```



We found that Friday and Saturday had higher numbers of accidents per day, but overall not significant enough, there could be several reasons for this:

- Weekend traffic: One of the most significant factors that could have contributed to a higher number of accidents on Fridays and Saturdays is weekend traffic. On weekends, more people may be out and about, and more vehicles may be on the road, leading to a higher probability of accidents.

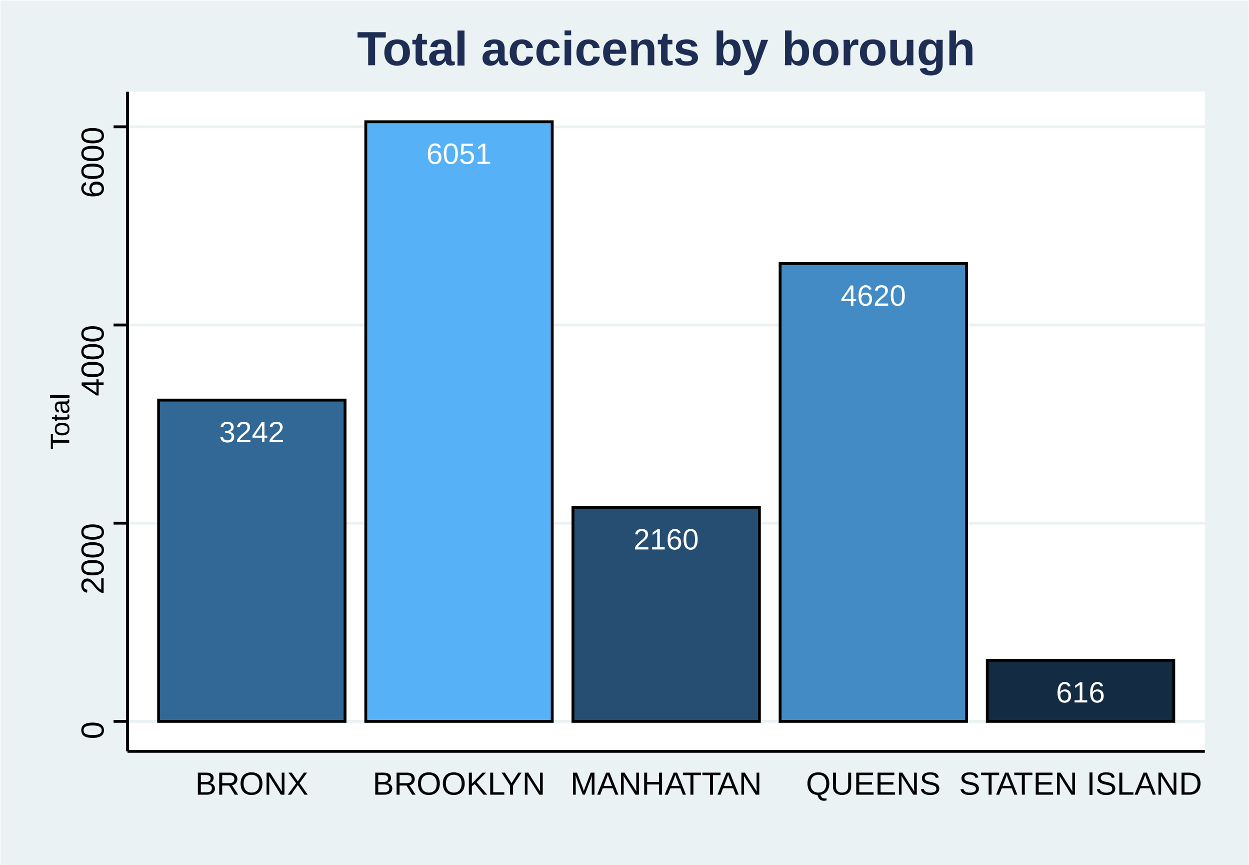
- Alcohol consumption: Another factor that could have contributed to a higher number of accidents on Fridays and Saturdays is alcohol consumption. On weekends, people may be more likely to consume alcohol, which can impair their driving ability and increase the risk of accidents.
- Time of day: It is also possible that the time of day could have contributed to the higher number of accidents on Fridays and Saturdays. People may be more likely to travel at night or later in the day on weekends, which can increase the risk of accidents due to reduced visibility or fatigue.

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

```
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())
```

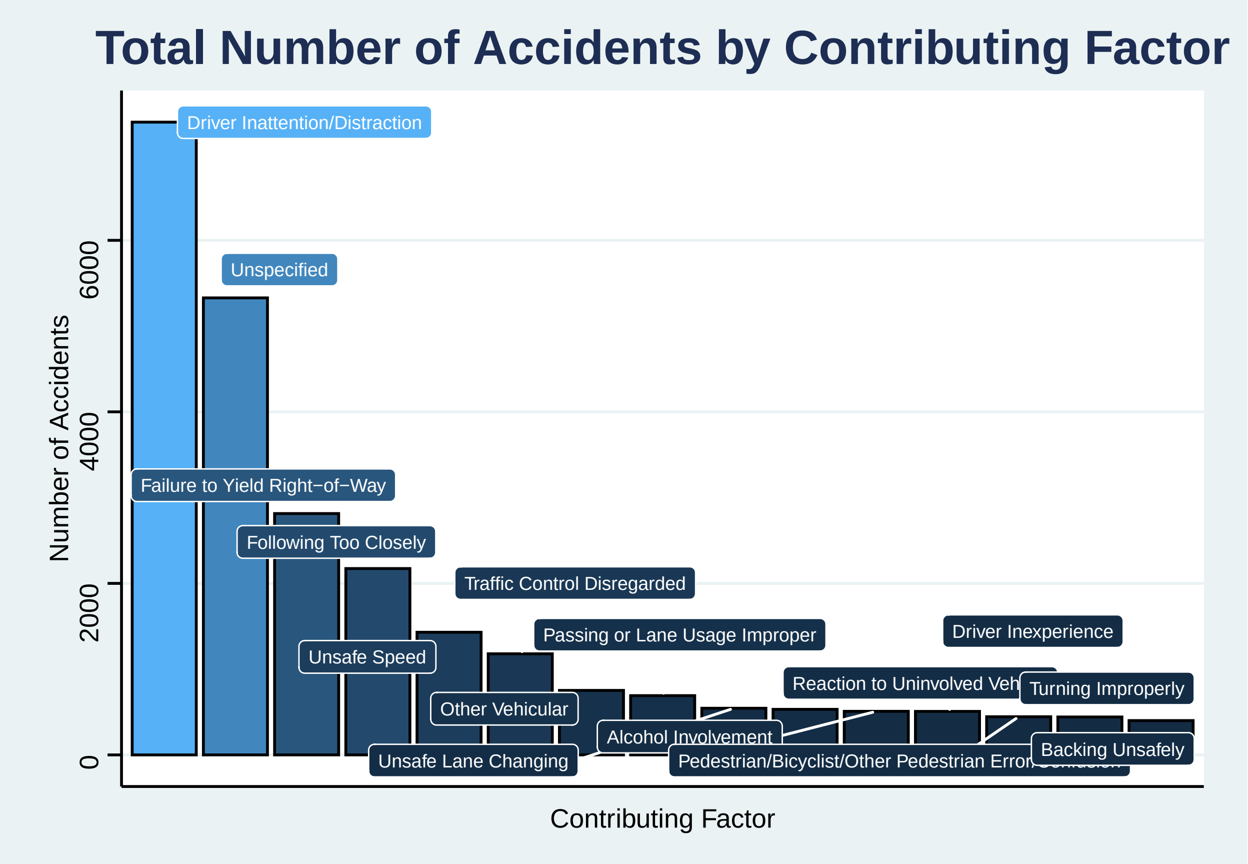


We found that Brooklyn had the highest number of accidents, followed by Queens, and we are surprised that Manhattan did not have the highest number, there could be several reasons for this:

- Differences in traffic patterns: Brooklyn and Queens are both large boroughs with a lot of traffic and major highways running through them, which could explain the higher number of accidents. Manhattan, on the other hand, is a smaller borough with fewer highways, which could result in fewer accidents.
- Differences in population density: Another factor that could contribute to differences in the number of accidents is population density. Brooklyn and Queens have a higher population density than Manhattan, which means there are more people and vehicles on the road, increasing the likelihood of accidents.
- Differences in road infrastructure: Road infrastructure can also play a role in the number of accidents. Some areas may have poorly designed roads, confusing intersections, or inadequate signage, which can increase the risk of accidents.
- External factors: It is also possible that external factors such as weather, road conditions, or driver behavior could have 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())
```



We found out that distraction is the higest followed by failed to yield and others, this suggests that policies and interventions aimed at reducing distracted driving and improving yielding behavior could help reduce the number of accidents.

- Public awareness campaigns: The city could launch a public awareness campaign to educate drivers about the risks of distracted driving and the importance of focusing on the road. The campaign could use different channels such as social media, radio, TV, and billboards to reach as many drivers as possible.
- Enforcement of traffic laws: The police department could increase the enforcement of traffic laws related to distracted driving and failing to yield. This could include targeted patrols in areas with high accident rates or increased use of traffic cameras to capture and penalize drivers who violate these laws.
- Infrastructure improvements: The city could invest in infrastructure improvements such as better road markings, traffic signals, and signage to help drivers navigate safely and avoid accidents.
- Driver education: The city could consider incorporating education about distracted driving and yielding behaviors into driver education programs, to ensure that new drivers are aware of the risks and are better prepared to drive safely.

By implementing policies and interventions that address the leading contributing factors to motor vehicle collisions, the city can work towards reducing the number of accidents, injuries, and fatalities on the roads.