

SSC 442 - Final Project

Peter Fu Chen , Mike Liu

2023-02-03

PowerPoint reference: https://www.linkedin.com/in/hoyuli/overlay/1593491717863/single-media-viewer/?profileId=ACoAABGobQoBad3aoz8Aka2DErr_wx3A9UJhl50

Analysis reference: <https://www.kaggle.com/datasets/mysarahmadbhat/nyc-traffic-accidents/code>

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, obtained from Kaggle, 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)
```

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, 1, 0~
## $ NUMBER.OF.PERSONS.KILLED <int> 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, 1, 0~
## $ NUMBER.OF.PEDESTRIANS.KILLED <int> 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~
## $ NUMBER.OF.CYCLIST.KILLED <int> 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~
## $ NUMBER.OF.MOTORIST.KILLED <int> 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> "", "", "", "", "", "", "", "", "", "", "", ~
```

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

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

```
# 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 dropped
sum(is.na(df_withoutna))
```

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

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

```
# 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, 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, 0~
## $ NUMBER.OF.PEDESTRIANS.INJURED <int> 0, 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, 0~
## $ NUMBER.OF.CYCLIST.INJURED <int> 0, 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, 0~
## $ NUMBER.OF.MOTORIST.INJURED <int> 0, 2, 0, 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, 0~
```

Map plotting the people killed by motor vehicle in NYC

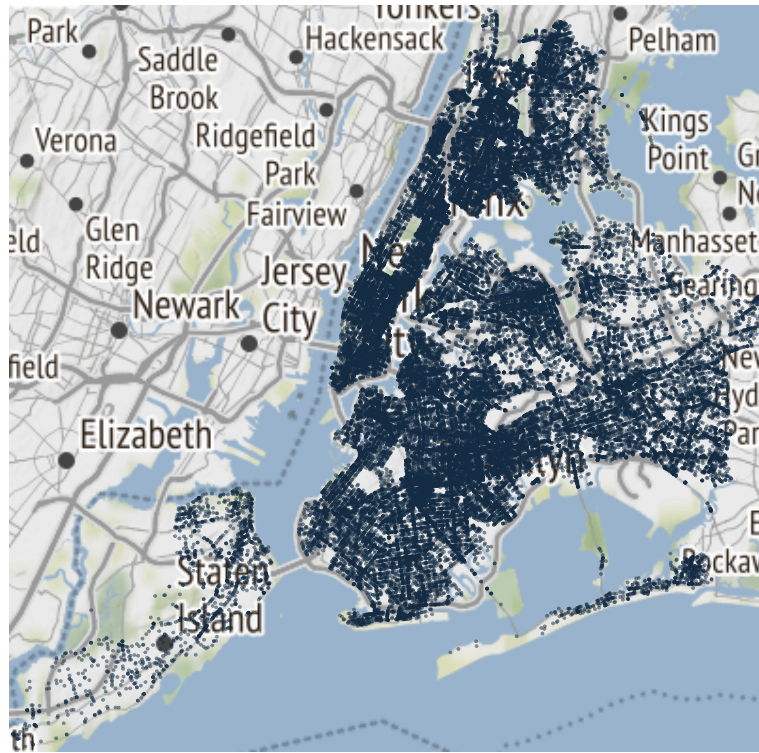
First and foremost, let's examine the number of fatalities resulting from motor vehicle accidents across New York City in 2020. To represent each death, we have used dots on a map and it's evident that the concentration of these dots varies among different boroughs. Upon initial inspection, it appears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

```
qplot(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.
```

NYC Traffic Accidents Killed



Descriptive statistics: Summarize the main characteristics of the data set, such as the mean, median, and standard deviation of the number of accidents and injuries.

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

```
model <- lm(NUMBER.OF.PERSONS.INJURED ~ BOROUGH, data = df)

# Summarize the model
summary(model)

##
## Call:
## lm(formula = NUMBER.OF.PERSONS.INJURED ~ BOROUGH, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4219 -0.4208 -0.3432  0.5792 14.6437
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.420768   0.004518  93.135  <2e-16 ***
## BOROUGHBRONX   -0.077559   0.008729  -8.885  <2e-16 ***
## BOROUGHBROOKLYN -0.064466   0.007175  -8.984  <2e-16 ***
## BOROUGHMANHATTAN -0.128235   0.009585 -13.379  <2e-16 ***
```

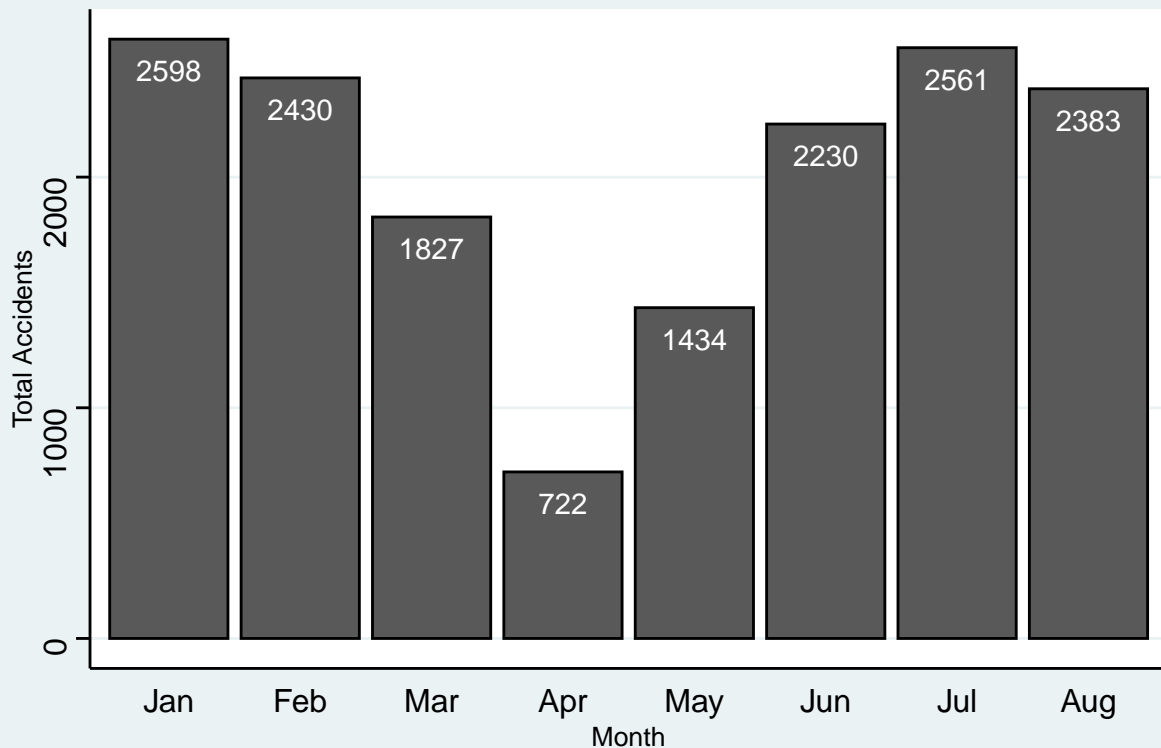
```
## BOROUGHQUEENS          -0.092595   0.007609 -12.170   <2e-16 ***
## BOROUGHSTATEN ISLAND  0.001085   0.019590   0.055     0.956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7248 on 74875 degrees of freedom
## Multiple R-squared:  0.003746,   Adjusted R-squared:  0.00368
## F-statistic: 56.31 on 5 and 74875 DF,  p-value: < 2.2e-16
```

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 = 'black') +
  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())
```

Total Accidents by Month in NYC

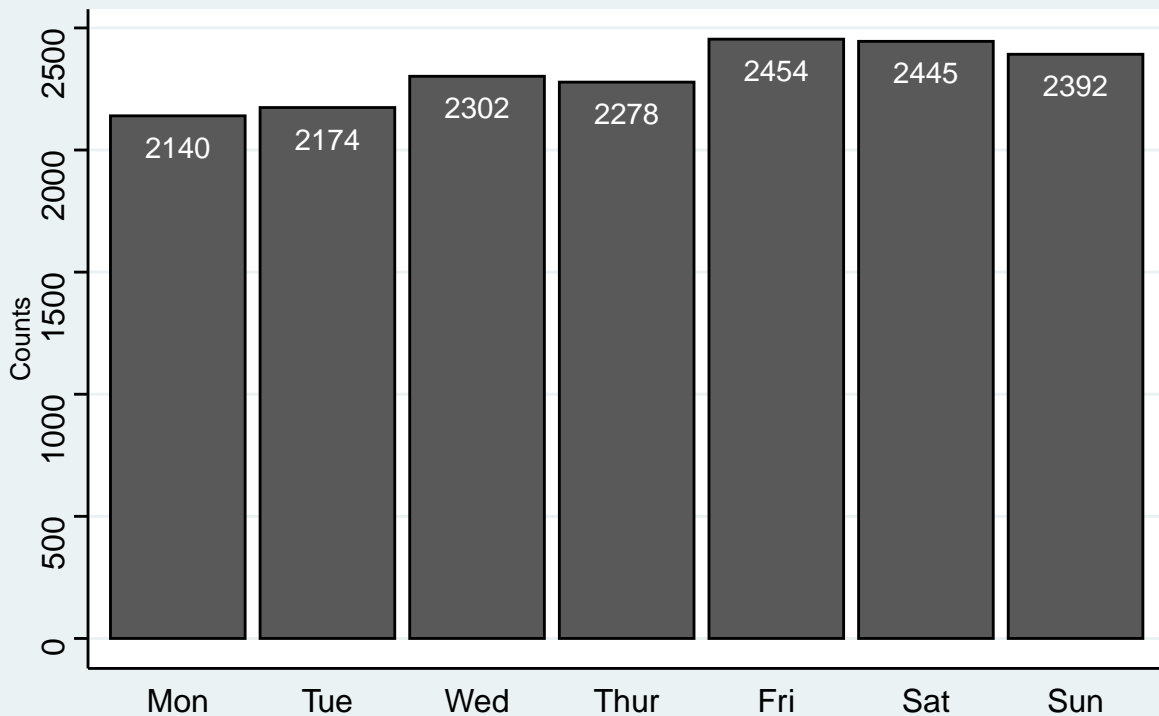


Accidents Per day

```
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 accidents 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())
```

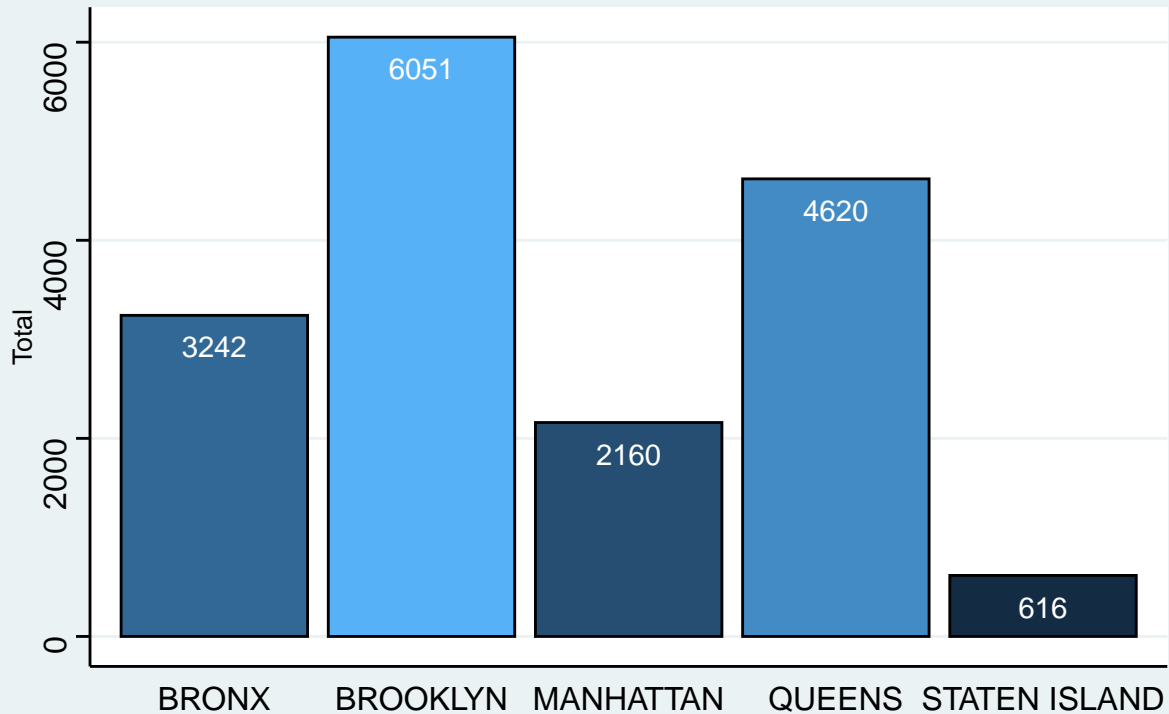
Total accidents per day of the week in NYC



Accidents, injured and killed per borough

```
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 accidents 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())
```

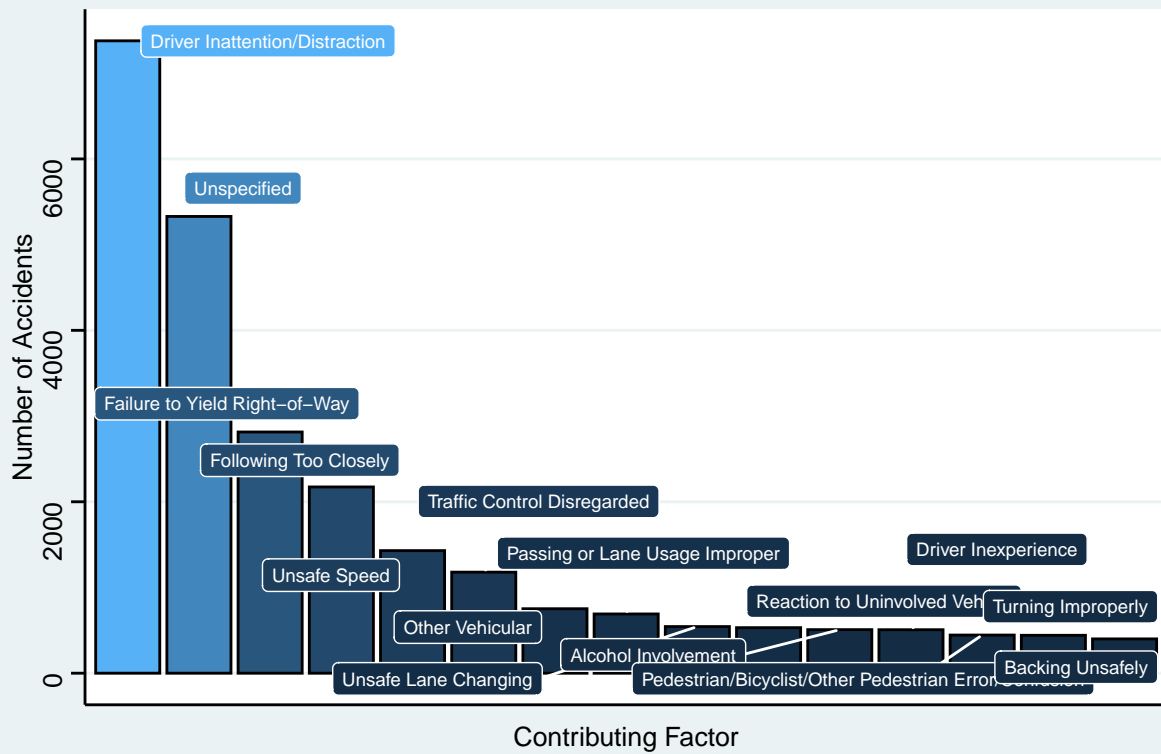
Total accidents by borough



Work in progress.....

```
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.overlap = 5) +
  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())
```


Total Number of Accidents by Contributing Factor



df