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(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, 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 <- na.omit(df)

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

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

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

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

```
# Drop some unnecessary columns
df = df %>% select(1:18)
glimpse(df)
```

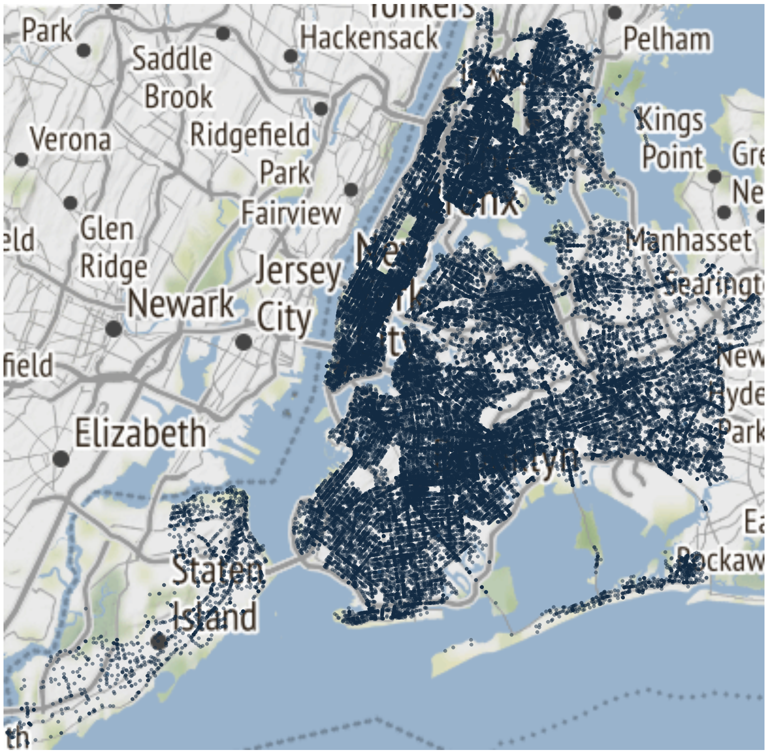
```
## Rows: 47,686
## Columns: 18
## $ CRASH.DATE                    <date> 2020-08-29, 2020-08-29, 2020-08-29, 202~
## $ CRASH.TIME                    <chr> "15:40:00", "21:00:00", "0:00:00", "17:1~
## $ BOROUGH                       <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN"~
## $ ZIP.CODE                      <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

```
qmplot(data = df, 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.

```
df %>% group_by(BOROUGH) %>% summarise(sum(NUMBER.OF.PERSONS.KILLED))
```

```
## # A tibble: 5 x 2
##   BOROUGH       ‘sum(NUMBER.OF.PERSONS.KILLED)‘
##   <chr>                                   <int>
## 1 BRONX                                      10
## 2 BROOKLYN                                   26
## 3 MANHATTAN                                   9
## 4 QUEENS                                     19
## 5 STATEN ISLAND                               6
```

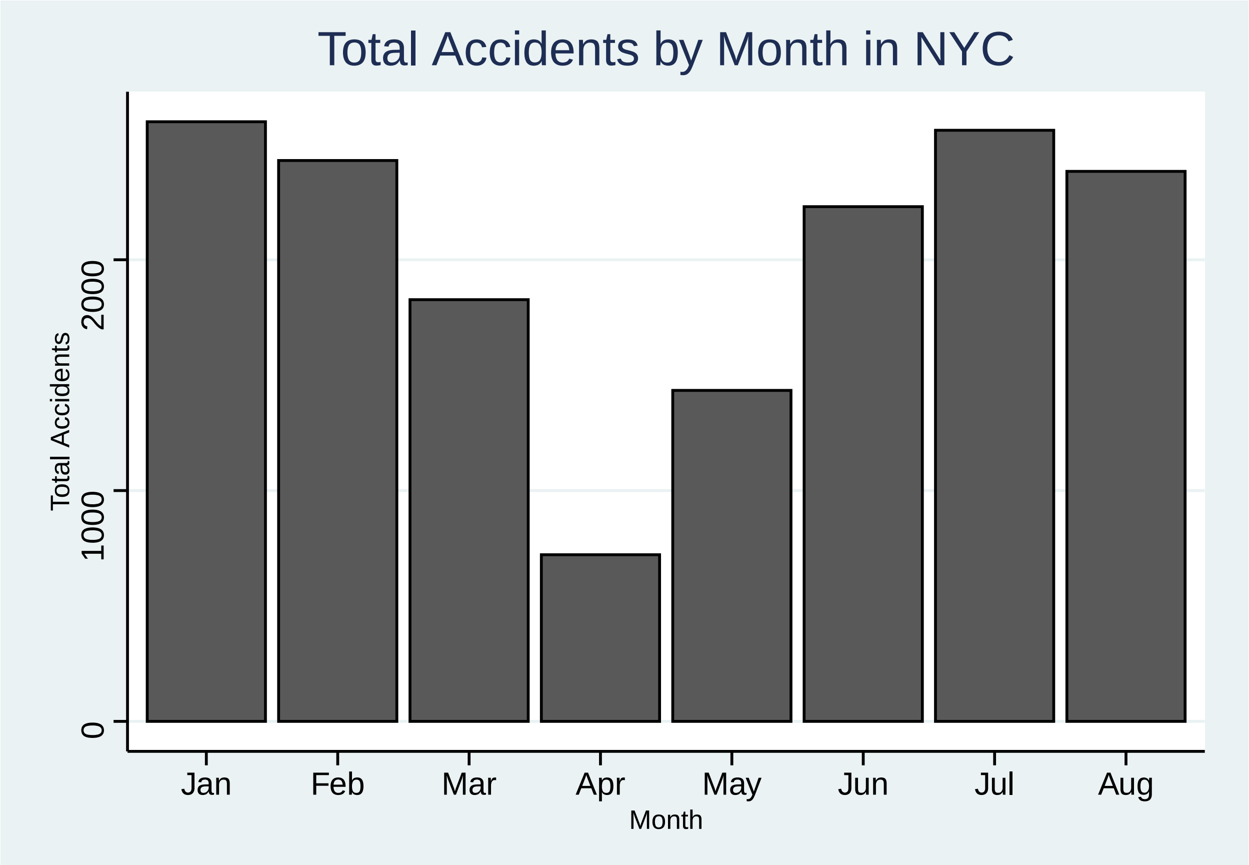
Regression analysis: Model the relationship between the number of accidents and borough.

```
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
## -0.4295 -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 ***
## BOROUGHBROOKLYN        0.014217   0.008889   1.599    0.110
## BOROUGHMANHATTAN      -0.047040   0.010777  -4.365 1.27e-05 ***
## BOROUGHQUEENS         -0.013156   0.009215  -1.428    0.153
## BOROUGHSTATEN 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
```

Accidents Per Month

```r
month_accidents = df %>% 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")) +
  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),
        legend.text = element_text(size = 12))
```


Total Accidents by Month in NYC
Total Accidents
0
1000
2000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Month