Week 3 Assignment

Joshua Trahan

11/12/2021

```
library(tidyverse)
library(chron)
library(lubridate)
url <- 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
nypd_df <- read.csv(url)
```

New York City: Analysis of Shooting Incidents

The dataset includes all of the shooting incidents that occured in NY City from 2006 until the end of 2020.

```
nypd_df <- nypd_df %>%
  select(-c(INCIDENT_KEY, JURISDICTION_CODE, LOCATION_DESC)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE), PRECINCT = as.factor(PRECINCT),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         OCCUR_TIME = chron(times = OCCUR_TIME))
```

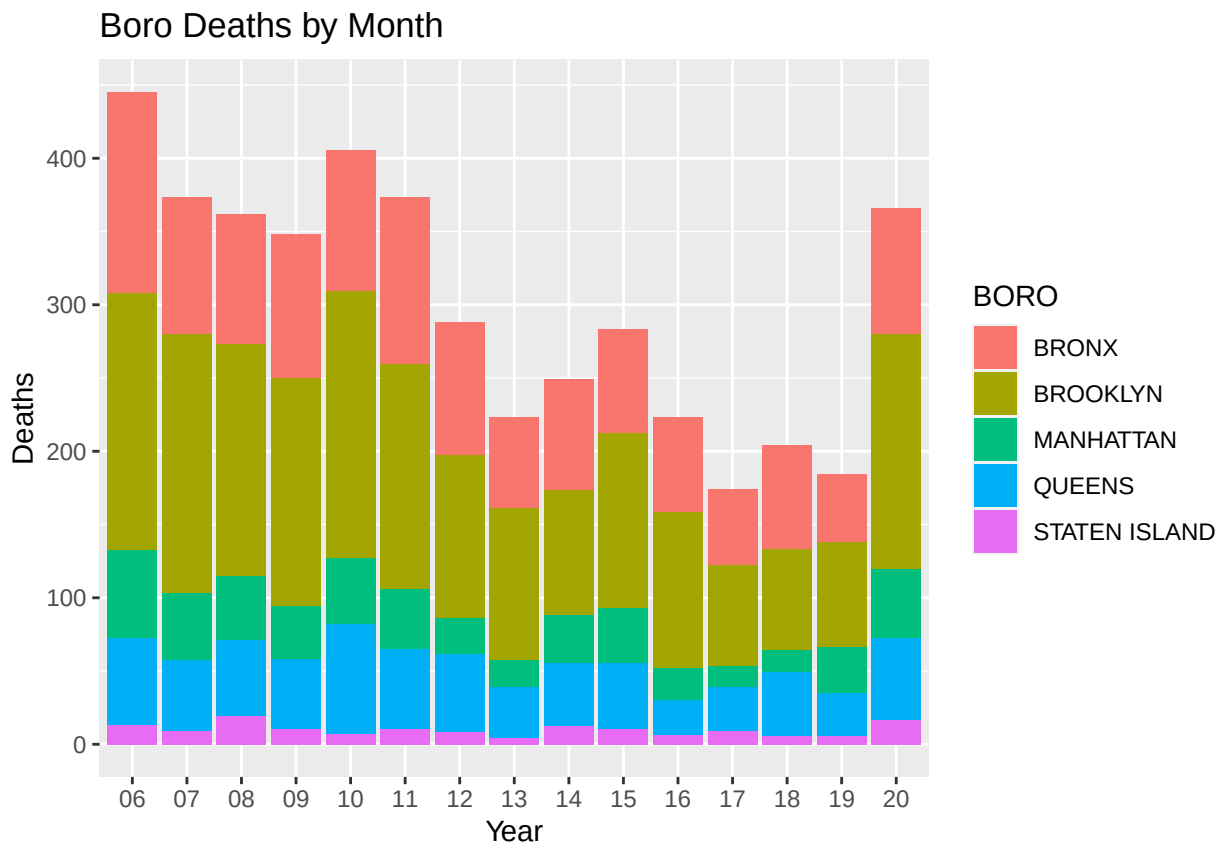
The chart below depicts the number of murders by New York City boro

Statin Island and Queen have reported the fewest while Brooklyn and the Bronx the most.

```
boro_shootings_deaths <- nypd_df %>%
  select(OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG)%>%
  mutate(STATISTICAL_MURDER_FLAG = as.integer(nypd_df$STATISTICAL_MURDER_FLAG))

boro_shootings_deaths <- boro_shootings_deaths %>%
  mutate(MM_YY = format(as.Date(OCCUR_DATE), '%m-%y'))
boro_shootings_deaths1 <- boro_shootings_deaths %>%
  mutate(Year = format(as.Date(OCCUR_DATE),'%y'))
```

```
ggplot(boro_shootings_deaths1, aes(fill=BORO, y=STATISTICAL_MURDER_FLAG, x = Year))+
  geom_bar(position="stack", stat="identity")+
  labs(title = 'Boro Deaths by Month', x = 'Year', y = 'Deaths')
```




```
view(boro_shootings_deaths)
```

Data Summary

Observations are missing in perpetrator data

> This data will be omitted for logistic regression. Rows are included for accurate shooting and death comparisons.

```
NY_shooting_v_Death <- boro_shootings_deaths %>%
  mutate(Shooting = is.integer(boro_shootings_deaths$STATISTICAL_MURDER_FLAG))
NY_shooting_v_Death <- NY_shooting_v_Death %>%
  select(MM_YY, STATISTICAL_MURDER_FLAG, Shooting) %>%
  mutate(MM_YY = factor(NY_shooting_v_Death$MM_YY))%>%
  group_by(MM_YY) %>%
  summarize(STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG), Shooting = sum(as.numeric(Shooting))
str(NY_shooting_v_Death)
```

```
## tibble [180 x 3] (S3: tbl_df/tbl/data.frame)
##  $ MM_YY                  : Factor w/ 180 levels "01-06","01-07",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ STATISTICAL_MURDER_FLAG: int [1:180] 29 14 23 18 18 16 14 19 28 27 ...
##  $ Shooting               : num [1:180] 129 109 114 105 97 102 114 119 107 117 ...
```

```
summary(NY_shooting_v_Death)
```

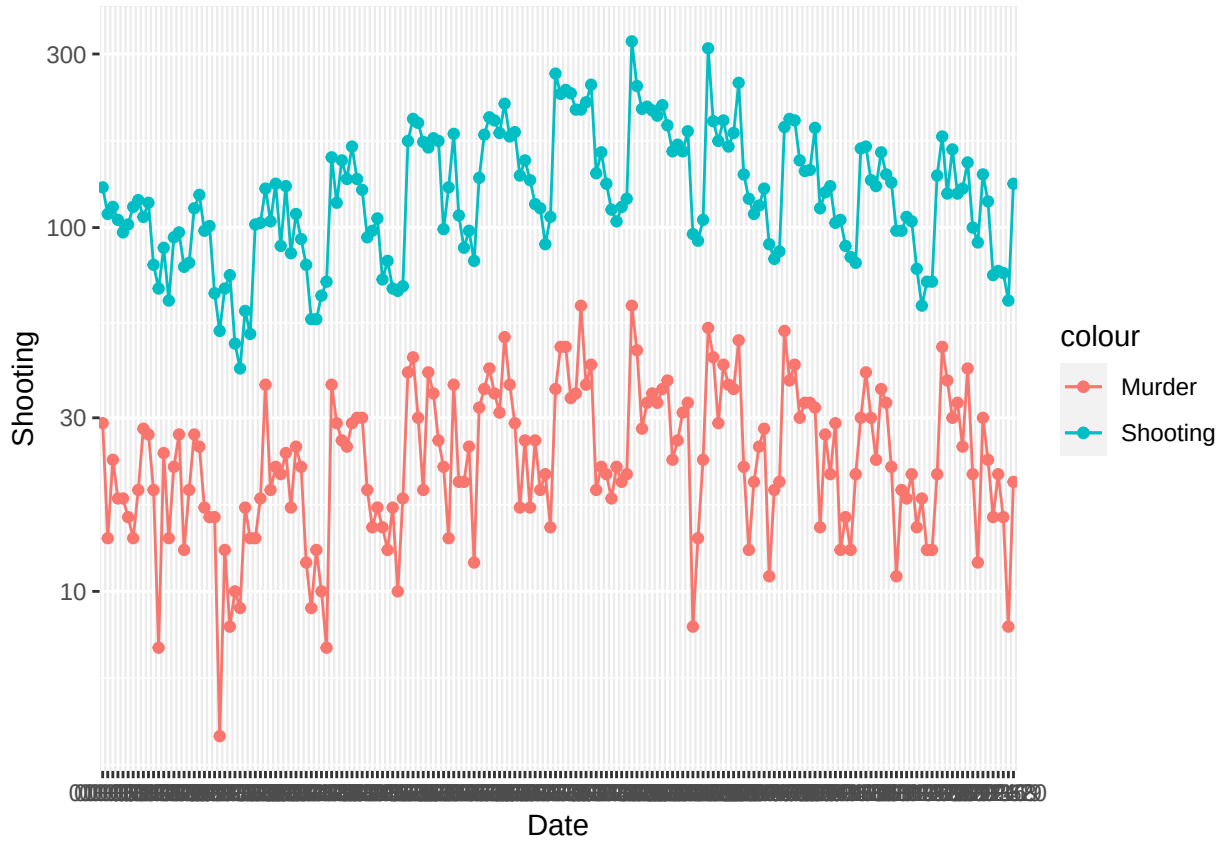
```
##      MM_YY     STATISTICAL_MURDER_FLAG    Shooting
##  01-06  :  1   Min.   : 4              Min.   : 41.00
##  01-07  :  1   1st Qu.:17              1st Qu.: 92.75
##  01-08  :  1   Median :22              Median :119.50
##  01-09  :  1   Mean   :25              Mean   :131.03
##  01-10  :  1   3rd Qu.:33              3rd Qu.:167.00
##  01-11  :  1   Max.   :61              Max.   :325.00
##  (Other):174
```

```
colnames(NY_shooting_v_Death)[2] <- 'Murder'
colnames(NY_shooting_v_Death)[1] <- 'Date'
```

```
NY_shooting_v_Death_plot <- NY_shooting_v_Death %>%
  ggplot(aes(x=Date, y=Shooting, group=1))+
  geom_line(aes(color = 'Shooting'))+
  geom_line(aes(y=Murder, color = 'Murder'))+geom_point(aes(color='Shooting'))+
  geom_point(aes(y=Murder, color = 'Murder'))+
  scale_y_log10()
```

Trends between shooting incidents and corresponding death

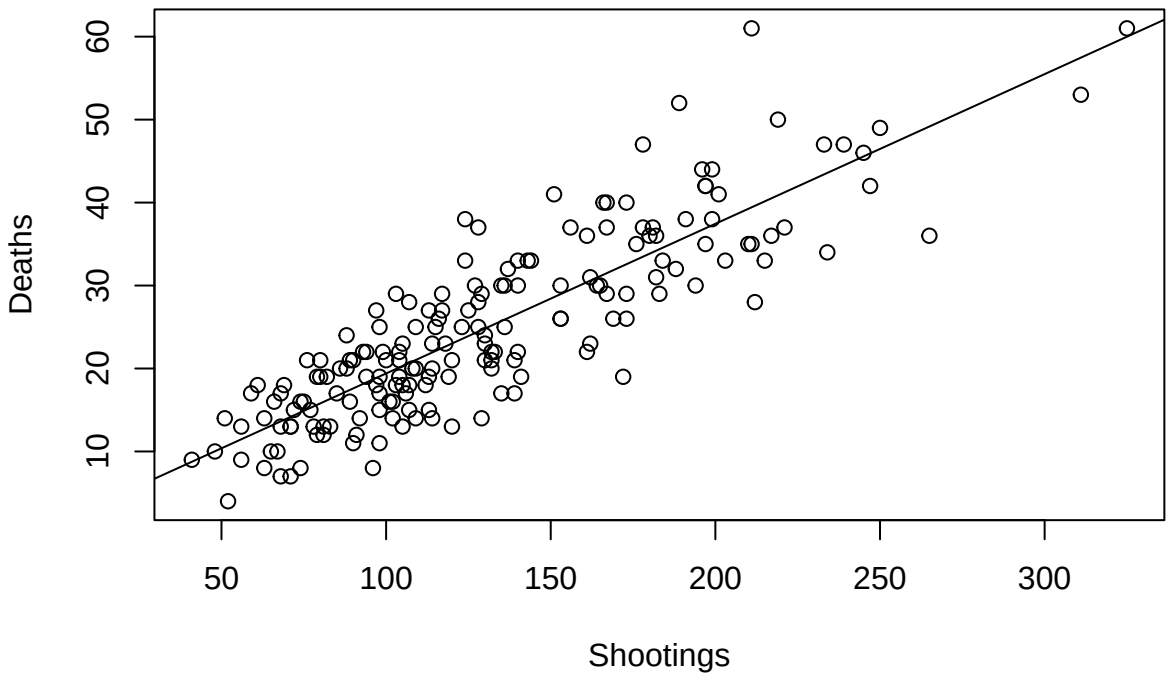
```
NY_shooting_v_Death_plot
```



Simple Linear Regression

This regression analysis depicts the relationship between shootings and shooting deaths. Shooting deaths increase along with shootings. As the number of shootings increase it becomes more difficult to predict resulting deaths. Analysis of outliers may help with understanding the nature of shootings.

```
lm_death <- lm(Murder ~ Shooting, data = NY_shooting_v_Death)
plot(x = NY_shooting_v_Death$Shooting, y= NY_shooting_v_Death$Murder, xlab = 'Shootings', ylab = 'Deaths
  abline(lm_death)+
  geom_text(aes(y = Deaths, label = Murder), size = 4)
```




```
## NULL
```

Identifying Bias

In order to avoid a bias dataset, additional data is needed about the population of the provided areas. The data cannot answer the *why*. No additional information is provided about the precincts, state economic investment, or detailed demographics.