NYPD_Shooting_hist

L. Black

2023-12-02

Import Library and Dataset

Libraries to be used in this report, tidyverse The data set of the Historic NYPD shooting is imported with the URL as a CSV file. The URL variable links the raw data that is read into as "data". The question I am starting with is there any correlation in lethal and non-lethal shootings in communities of high and low wealth and is the poorest borough and percentage of overall .

```
#install.packages("tidyverse")

library("tidyverse")
library("lubridate")
```

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
raw_data <- read_csv(url,show_col_types = FALSE)

summary(raw_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO          
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312      
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character  
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character  
##  Mean   :120860536                      Mode  :numeric                      
##  3rd Qu.:188810230                                                          
##  Max.   :261190187                                                          
##                                                                             
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312      
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character  
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character  
##                     Mean   : 65.64   Mean   :0.3269                      
##                     3rd Qu.: 81.00   3rd Qu.:0.0000                      
##                     Max.   :123.00   Max.   :2.0000                      
##                                      NA's   :2                           
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP    
##  Length:27312       Mode :logical           Length:27312      
##  Class :character   FALSE:22046             Class :character  
##  Mode  :character   TRUE :5266              Mode  :character  
##                                                               
##                                                               
##                                                               
##                                                               
```

```
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                         NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Tidy Dataset

Summary and clean up the data. First, the dates within the data set needs to be standardized for easy viewing. Data in relation to reports with shooting incidents that aren't labelled as murders are of interest for the first test set. The other data set is where shooting incidents end in lethally. Any missing data is removed after the interested data is selected, the missing data can be seen in age, race, sex, & most things in relation to identification of a perpetrator.

```
nonlethal_data <- raw_data %>% filter(STATISTICAL_MURDER_FLAG == FALSE) %>% select(OCCUR_DATE,BORO) %>%
nonlethal_data$OCCUR_DATE <- mdy(nonlethal_data$OCCUR_DATE)

summary(nonlethal_data)
```

```
##    OCCUR_DATE             BORO
##  Min.   :2006-01-01   Length:22046
##  1st Qu.:2009-07-21   Class :character
##  Median :2013-05-13   Mode  :character
##  Mean   :2014-01-07
##  3rd Qu.:2018-10-01
##  Max.   :2022-12-31
```

```
lethal_data_t <- raw_data %>% filter(STATISTICAL_MURDER_FLAG == TRUE) %>% select(OCCUR_DATE,BORO) %>% d

lethal_data_t$OCCUR_DATE <- mdy(lethal_data_t$OCCUR_DATE)

summary(lethal_data_t)
```

```
##    OCCUR_DATE             BORO          
##  Min.   :2006-01-01   Length:5266       
##  1st Qu.:2009-06-30   Class :character  
##  Median :2013-03-06   Mode  :character  
##  Mean   :2014-01-03                     
##  3rd Qu.:2018-12-30                     
##  Max.   :2022-12-30                     
```

Transform and Visualize Dataset

Grouping the data by date and counted to start to take a look at the trends in shooting victims over the years that didn't result in murder. I am looking to test the data against the wealthiest and poorest boroughs in NYC to see any differences between them in lethal and non-lethal shootings. In these sets we are not looking in specific regions but by dates and count of incidents.

Non-Lethal vs Lethal Shooting Reports by NYPD

Transforming the data sets to reflect lethal and non-lethal reports out of the overall data set.The sets are grouped by date and counted. The total of the victims of both reports are of interest not the number of incidents reported.

```r
non_lethal_data <- nonlethal_data %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()
lethal_data <- lethal_data_t %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()

head(non_lethal_data)
```

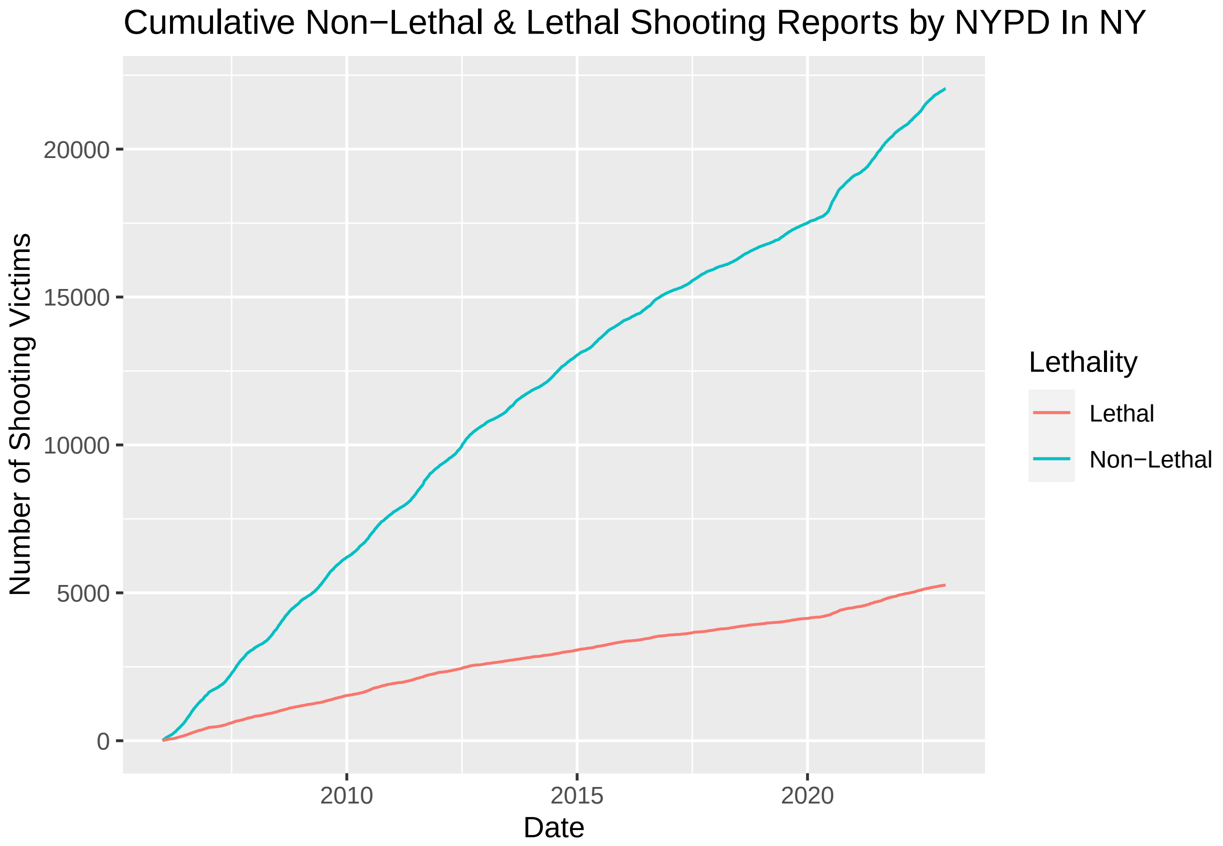
```
## # A tibble: 6 x 2
##   OCCUR_DATE COUNT
##   <date>     <int>
## 1 2006-01-01     4
## 2 2006-01-02     3
## 3 2006-01-03     3
## 4 2006-01-04     4
## 5 2006-01-05     4
## 6 2006-01-06     4
```

```r
head(lethal_data)
```

```
## # A tibble: 6 x 2
##   OCCUR_DATE COUNT
##   <date>     <int>
## 1 2006-01-01     4
## 2 2006-01-02     1
## 3 2006-01-03     1
## 4 2006-01-07     1
## 5 2006-01-08     1
## 6 2006-01-09     5
```

The visual below is to show how linear the data is over the years for lethal and non-lethal incidents is. Shown in the visual, non-lethal shootings have been on the rise much higher than lethal shooting incidents which can be seen as a positive. Lethal shootings have been on the rise, in a less than 2,500 every 5 years while non-lethal shootings are on the rise on average more than 5,500 every 5 years.

```r
ggplot() +
  geom_line(data = non_lethal_data, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Non-Lethal')) +
  geom_line(data = lethal_data, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Lethal')) +
  labs(title = "Cumulative Non-Lethal & Lethal Shooting Reports by NYPD In NY") +
  labs(y ="Number of Shooting Victims", x = "Date", color = "Lethality")
```



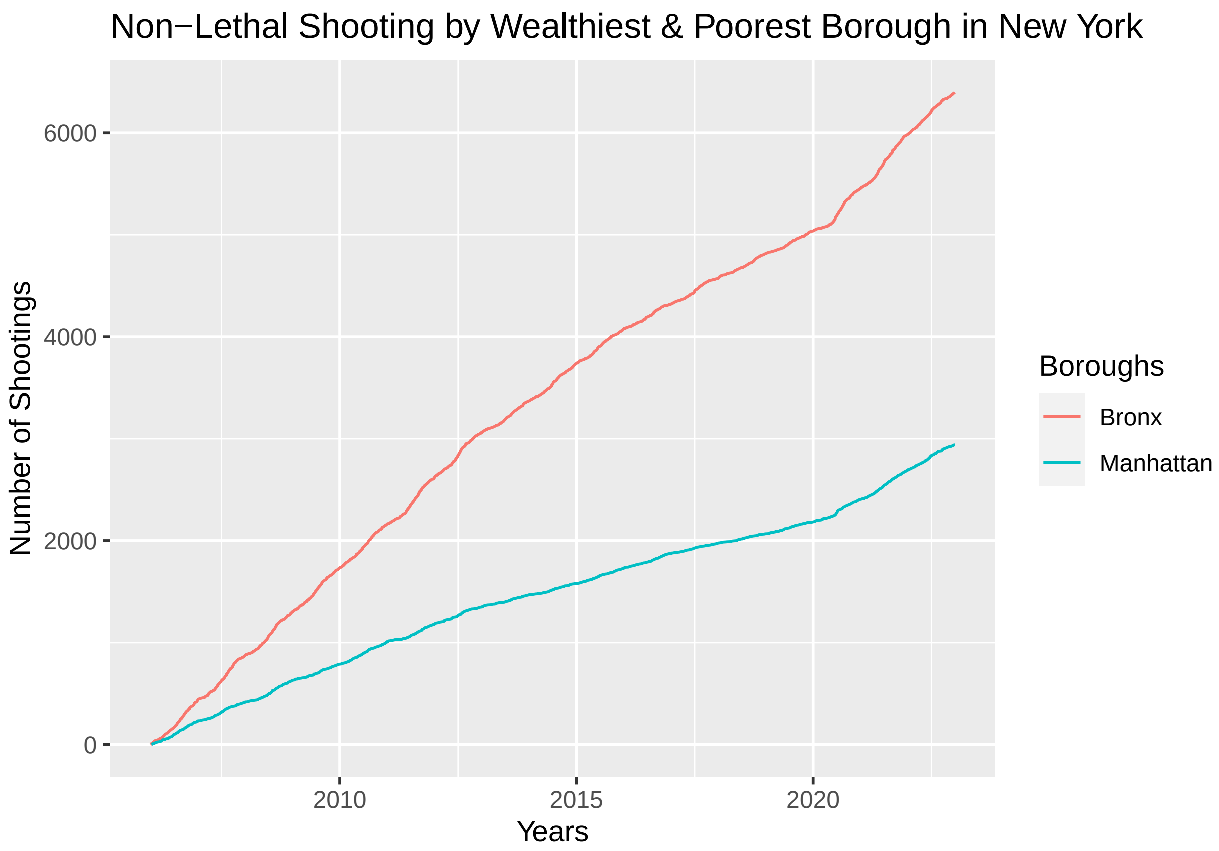
Non-Lethal Shooting Reports by Wealthiest & Poorest Borough in New York

This visual for this set in only for Non-lethal shooting reports based on statistically overall most wealthy and most poor of the boroughs in NYC. It is widely believed that the more poor a community is, the higher rate of crime there is but that can also be said for wealth communities being full of crime and shooting because of robberies & burglaries. Overall the Bronx, poor community, has a significantly more non-lethal shootings than Manhattan perhaps by a few thousand.

```r
bronx_n <- nonlethal_data %>% filter(BORO == "BRONX") %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n(
manhattan_n <- nonlethal_data %>% filter(BORO == "MANHATTAN") %>% group_by(OCCUR_DATE) %>% summarise(COU
```

```r
ggplot() +
  geom_line(data = bronx_n, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Bronx')) +
  geom_line(data = manhattan_n, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Manhattan')) +

  labs(title = "Non-Lethal Shooting by Wealthiest & Poorest Borough in New York") +
  labs(y = "Number of Shootings", x = "Years", color = "Boroughs")
```



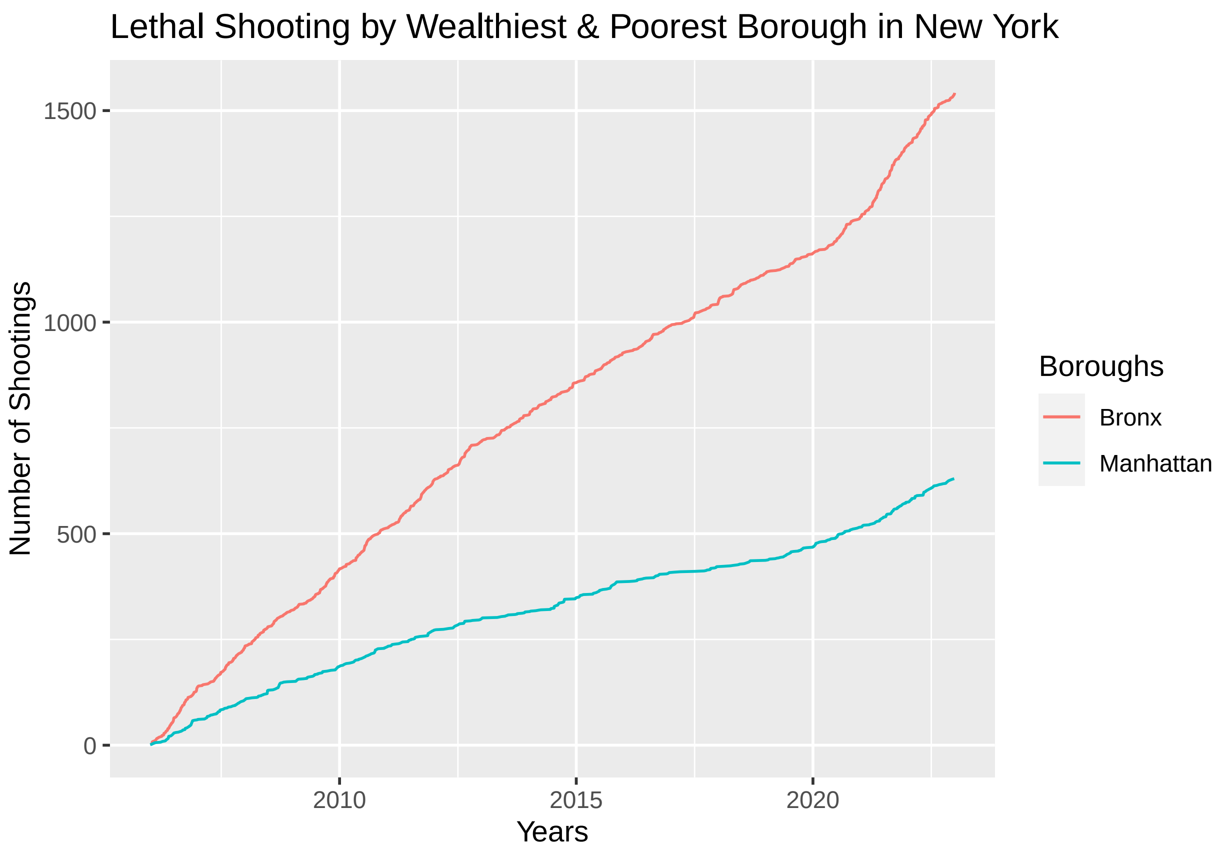
Lethal Shooting Reports by Wealthiest & Poorest Borough in New York

This visual for this set in only for Lethal shooting reports based on statistically overall most wealthy and most poor of the boroughs in NYC. It is widely believed that the more poor a community is, the higher rate of crime there is but that can also be said for wealth communities being full of crime and shooting because of robberies & burglaries. Overall the Bronx, poor community, has a significantly more lethal shootings than Manhattan perhaps by a several hundred.

```r
bronx_l <- lethal_data_t %>% filter(BORO == "BRONX") %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()
manhattan_l <- lethal_data_t %>% filter(BORO == "MANHATTAN") %>% group_by(OCCUR_DATE) %>% summarise(COU
```

```r
ggplot() +
  geom_line(data = bronx_l, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Bronx')) +
  geom_line(data = manhattan_l, aes(x = OCCUR_DATE, y = cumsum(COUNT), color = 'Manhattan')) +

  labs(title = "Lethal Shooting by Wealthiest & Poorest Borough in New York") +
  labs(y = "Number of Shootings", x = "Years", color = "Boroughs")
```



Yearly percent change in Lethal & Non-Lethal Shooting by Wealthiest & Poorest Borough

Here we wanted to see the yearly changes between the boroughs of most to least wealth in comparison to the overall percentage from all boroughs. These comparisons will be done for each of lethal and non-lethal shootings for visual purposes.

```
overall_yearly_nl <- non_lethal_data
overall_yearly_nl$OCCUR_DATE <- overall_yearly_nl$OCCUR_DATE %>% year()
overall_yearly_nl <- overall_yearly_nl %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>% ungroup()
overall_pct_nl <- overall_yearly_nl %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

summary(overall_pct_nl)
```

```
##    OCCUR_DATE       COUNT          pct_change
##  Min.   :2006   Min.   :285.0   Min.   :-6.990881
##  1st Qu.:2010   1st Qu.:328.0   1st Qu.:-2.571379
##  Median :2014   Median :333.0   Median :-0.440398
##  Mean   :2014   Mean   :326.9   Mean   : 0.006725
##  3rd Qu.:2018   3rd Qu.:342.0   3rd Qu.: 1.265511
##  Max.   :2022   Max.   :350.0   Max.   :14.186851
##                                 NA's   :1
```

```
bronx_yearly <- bronx_n
bronx_yearly$OCCUR_DATE <- bronx_yearly$OCCUR_DATE %>% year()
bronx_yearly <- bronx_yearly %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()
bronx_pct <- bronx_yearly %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

manhat_yearly <- manhattan_n
manhat_yearly$OCCUR_DATE <- manhat_yearly$OCCUR_DATE %>% year()
manhat_yearly <- manhat_yearly %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()
manhat_pct <- manhat_yearly %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

summary(bronx_pct)
```

```
##    OCCUR_DATE       COUNT         pct_change
##  Min.   :2006   Min.   :143.0   Min.   :-24.7368
##  1st Qu.:2010   1st Qu.:170.0   1st Qu.: -3.1293
##  Median :2014   Median :202.0   Median :  0.2646
##  Mean   :2014   Mean   :191.7   Mean   :  1.5281
##  3rd Qu.:2018   3rd Qu.:208.0   3rd Qu.:  5.4040
##  Max.   :2022   Max.   :249.0   Max.   : 42.7586
##                                 NA’s   :1
```

```
summary(manhat_pct)
```

```
##    OCCUR_DATE       COUNT         pct_change
##  Min.   :2006   Min.   : 62.0   Min.   :-30.909
##  1st Qu.:2010   1st Qu.: 88.0   1st Qu.:-17.147
##  Median :2014   Median :110.0   Median : -7.327
##  Mean   :2014   Mean   :110.8   Mean   :  2.747
##  3rd Qu.:2018   3rd Qu.:131.0   3rd Qu.: 23.659
##  Max.   :2022   Max.   :166.0   Max.   : 52.273
##                                 NA’s   :1
```

```
overall_yearly_l <- lethal_data
overall_yearly_l$OCCUR_DATE <- overall_yearly_l$OCCUR_DATE %>% year()
overall_yearly_l <- overall_yearly_l %>% group_by(OCCUR_DATE) %>% summarise(COUNT=n()) %>% ungroup()
overall_pct_l <- overall_yearly_l %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

summary(overall_pct_l)
```

```
##    OCCUR_DATE       COUNT         pct_change
##  Min.   :2006   Min.   :110.0   Min.   :-23.6111
##  1st Qu.:2010   1st Qu.:144.0   1st Qu.:-12.7678
##  Median :2014   Median :182.0   Median :  0.7895
##  Mean   :2014   Mean   :169.8   Mean   :  0.6326
##  3rd Qu.:2018   3rd Qu.:193.0   3rd Qu.:  6.4499
##  Max.   :2022   Max.   :217.0   Max.   : 64.6552
##                                 NA’s   :1
```

```
bronx_yearly_l <- bronx_l
bronx_yearly_l$OCCUR_DATE <- bronx_yearly_l$OCCUR_DATE %>% year()
bronx_yearly_l <- bronx_yearly_l %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()
```

```r
bronx_pct_l <- bronx_yearly %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

manhat_yearly_l <- manhattan_l
manhat_yearly_l$OCCUR_DATE <- manhat_yearly_l$OCCUR_DATE %>% year()
manhat_yearly_l <- manhat_yearly_l %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n()) %>% ungroup()
manhat_pct_l <- manhat_yearly_l %>% mutate(pct_change = ((COUNT/lag(COUNT) - 1) * 100))

summary(bronx_pct_l)
```

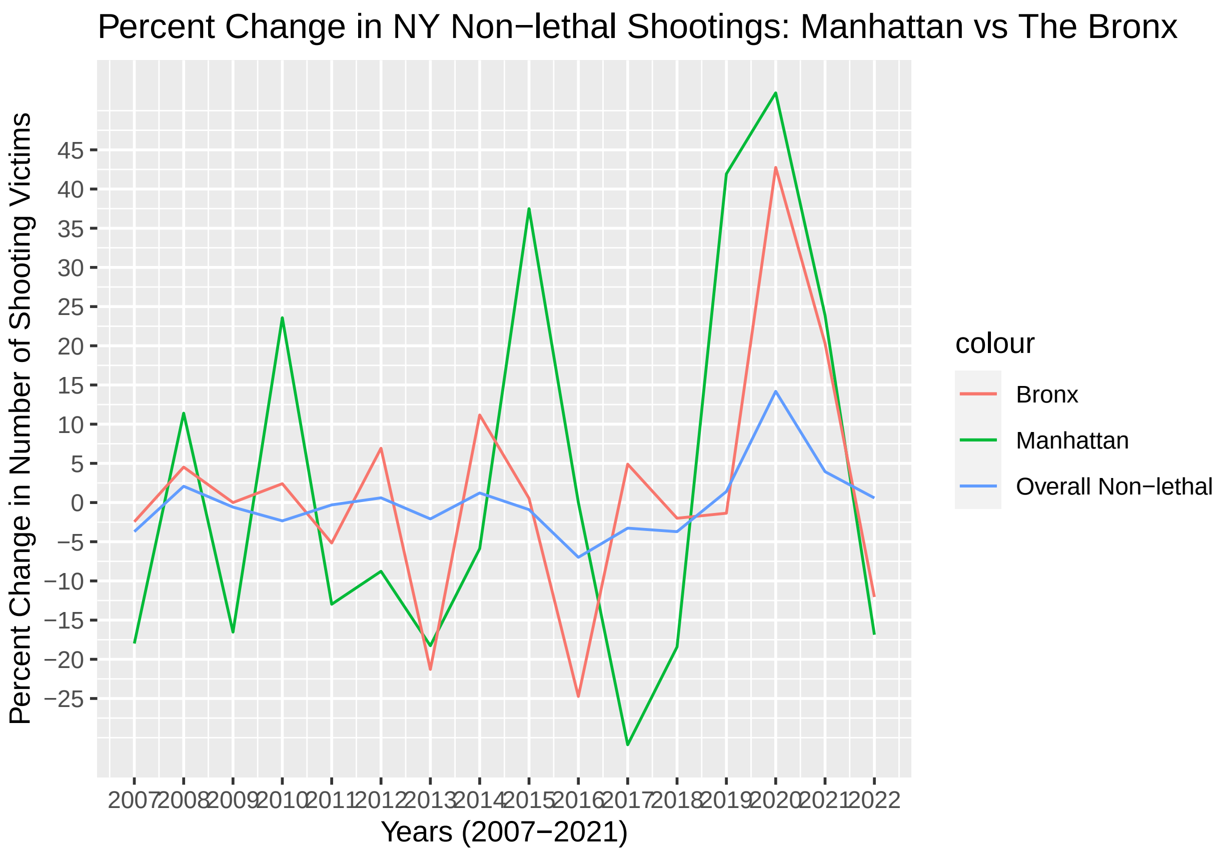
```
##    OCCUR_DATE       COUNT         pct_change
##  Min.   :2006   Min.   :143.0   Min.   :-24.7368
##  1st Qu.:2010   1st Qu.:170.0   1st Qu.: -3.1293
##  Median :2014   Median :202.0   Median :  0.2646
##  Mean   :2014   Mean   :191.7   Mean   :  1.5281
##  3rd Qu.:2018   3rd Qu.:208.0   3rd Qu.:  5.4040
##  Max.   :2022   Max.   :249.0   Max.   : 42.7586
##                                 NA's   :1
```

```r
summary(manhat_pct_l)
```

```
##    OCCUR_DATE       COUNT         pct_change
##  Min.   :2006   Min.   :13.00   Min.   :-29.630
##  1st Qu.:2010   1st Qu.:19.00   1st Qu.:-21.155
##  Median :2014   Median :27.00   Median : -7.895
##  Mean   :2014   Mean   :26.82   Mean   :  3.172
##  3rd Qu.:2018   3rd Qu.:34.00   3rd Qu.: 26.797
##  Max.   :2022   Max.   :44.00   Max.   : 92.308
##                                 NA's   :1
```

```r
ggplot() +
  geom_line(data = manhat_pct[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Manhattan")) +
  geom_line(data = bronx_pct[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Bronx")) +
  geom_line(data = overall_pct_nl[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Overall Non-letl

  labs(title = "Percent Change in NY Non-lethal Shootings: Manhattan vs The Bronx") +
  labs(y = "Percent Change in Number of Shooting Victims", x = "Years (2007-2021)") +
  scale_x_continuous(breaks = pretty(bronx_pct$OCCUR_DATE, n = 20)) +
  scale_y_continuous(breaks = pretty(bronx_pct$pct_change, n = 15))
```

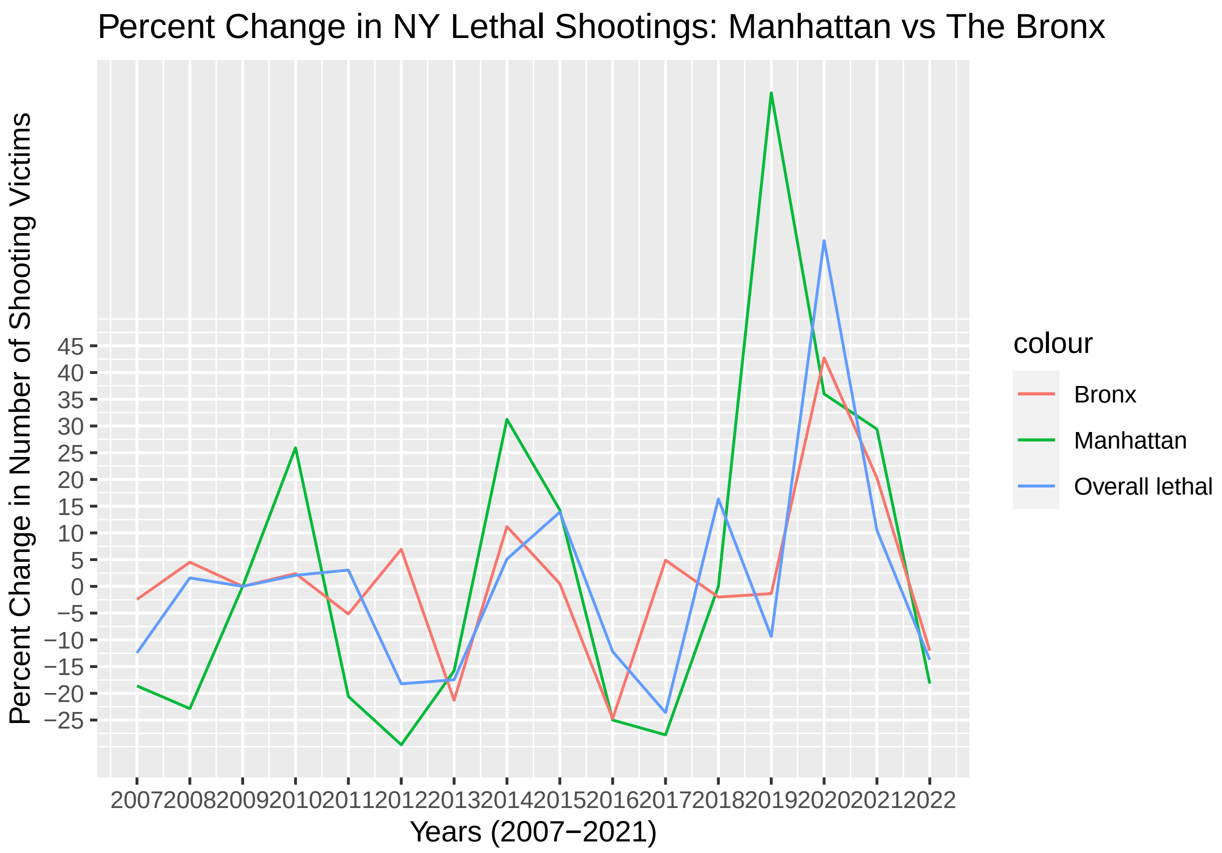


Analysis

The overall data percent is relatively consistent except for the anomaly of 2020 with a big spike in shootings. The year 2020 was a globally hard time and the spike was higher in Manhattan than The Bronx by over 10% from another. Also of note, Manhattan has larger spikes (dips and peaks) in percentages than the Bronx.

```
ggplot() +
  geom_line(data = manhat_pct_l[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Manhattan")) +
  geom_line(data = bronx_pct_l[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Bronx")) +
  geom_line(data = overall_pct_l[-1,],aes(x = `OCCUR_DATE`, y = `pct_change`, color = "Overall lethal")

  labs(title = "Percent Change in NY Lethal Shootings: Manhattan vs The Bronx") +
  labs(y = "Percent Change in Number of Shooting Victims", x = "Years (2007-2021)") +
  scale_x_continuous(breaks = pretty(bronx_pct_l$OCCUR_DATE, n = 20)) +
  scale_y_continuous(breaks = pretty(bronx_pct_l$pct_change, n = 15))
```



Analysis

Interestingly enough The Bronx follows the Overall lethal trend of percentage while Manhattan has highs peaks and dips not following the percetange of overall. I think it is fair to make the assumption that Manhattan has a greater percentage of all shooting incidents than that of the poorer boroughs.

Modeling

Here, a linear regression model is used to compare the correlation between the percent change in shooting incidents in the Overall and the Bronx. There is little to no correlation between the two, which is evidence that the coincidence in similar percent change in 2020 is an anomaly, likely linked to external factors or global events, like the pandemic.

```
both_data_pct_nl <- merge(overall_pct_nl[-1,],bronx_pct[-1,], by="OCCUR_DATE")

mod <- lm(pct_change.x ~ pct_change.y, data = both_data_pct_nl)

summary(mod)
```

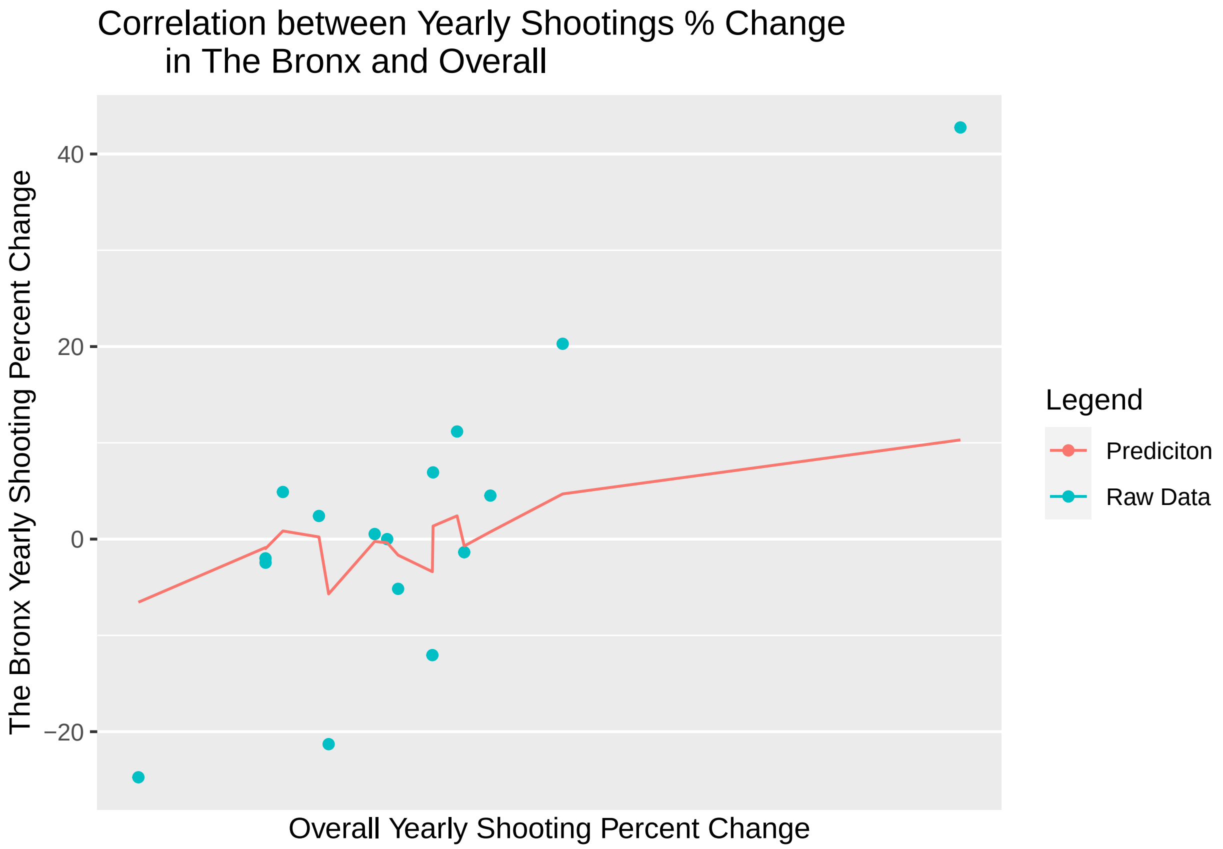
```
##
## Call:
## lm(formula = pct_change.x ~ pct_change.y, data = both_data_pct_nl)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.1158 -1.5392 -0.5488  1.5538  3.9677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.37498    0.65168  -0.575    0.574
## pct_change.y  0.24980    0.04278   5.839  4.3e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.594 on 14 degrees of freedom
## Multiple R-squared:  0.7089, Adjusted R-squared:  0.6881
## F-statistic: 34.09 on 1 and 14 DF,  p-value: 4.302e-05
```

```
preds <- both_data_pct_nl %>% mutate(pred = predict(mod))

preds %>% ggplot() + geom_point(aes(x = pct_change.x, y = pct_change.y, color = "Raw Data")) +
  geom_line(aes(x = pct_change.x, y = pred, color = "Prediciton")) +
  scale_x_continuous(breaks = pretty(both_data_pct_nl$OCCUR_DATE, n = 10)) +

  labs(title = "Correlation between Yearly Shootings % Change
       in The Bronx and Overall") +
  labs(y="The Bronx Yearly Shooting Percent Change",
       x="Overall Yearly Shooting Percent Change", color="Legend")
```



Bias

Bias here can be on the stereotypes on the type of communities, let alone how fiances play a part in it. Also, the data collection itself may have bias because some areas are over or under reported which can skew the data every which way. Shifts in socio-behavior during the pandemic can be noted throughout the report.