Historic NYPD Shooting Incident

S. Ravi

2022-04-08

Import the data

This document analyzes shooting incidents in New York City from 2006-2020. The data was obtained from https://catalog.data.gov/.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
incidents <- read.csv(url)
```

All necessary packages listed below:

```
#install.packages("egg")
install.packages("dplyr")
install.packages("formatR")
install.packages("ggplot2")
install.packages("ggrepel")
install.packages("reshape2")
install.packages("tinytex")
install.packages("tidyverse")
library(dplyr, formatR, tinytex)
```

Clean the data

The summary of the data doesn't tell us much, but it shows there are columns such as characteristics of the people involved, coordinates, and jurisdiction related things that are unnecessary to the analysis.

```
summary(incidents)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME             BORO
##  Min.   :  9953245   Length:23585       Length:23585       Length:23585
##  1st Qu.: 55322804   Class :character   Class :character   Class :character
##  Median : 83435362   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102280741
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Length:23585
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   Class :character
##  Median : 69.00   Median :0.000     Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.333
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX           VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                         Mean   :1009379   Mean   :207300
##                                         3rd Qu.:1016782   3rd Qu.:239163
##                                         Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude       Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

After manually skimming through the data, I noticed there were also several data points where the sex, age group, or race of the perpetrator were missing. These columns and these rows with missing values were removed to make handling the data simpler and analyzing it easier.

```
incidents <- dplyr::select(incidents, -c(INCIDENT_KEY, STATISTICAL_MURDER_FLAG,
↪  JURISDICTION_CODE, PRECINCT, LOCATION_DESC, Latitude, Longitude, X_COORD_CD,
↪  Y_COORD_CD, Lon_Lat))

cleaned <- incidents[!(is.na(incidents$PERP_RACE) | incidents$PERP_RACE==""),]
cleaned <- cleaned[!(is.na(cleaned$PERP_AGE_GROUP) | cleaned$PERP_AGE_GROUP==""),]
```

Looking at the grouping of the sex of the perpetrators across the original and cleaned dataset, it's clear that at least some 8000 data points have to be omitted as a result of having incomplete data.

```
table(incidents$PERP_SEX); table(cleaned$PERP_SEX)
```

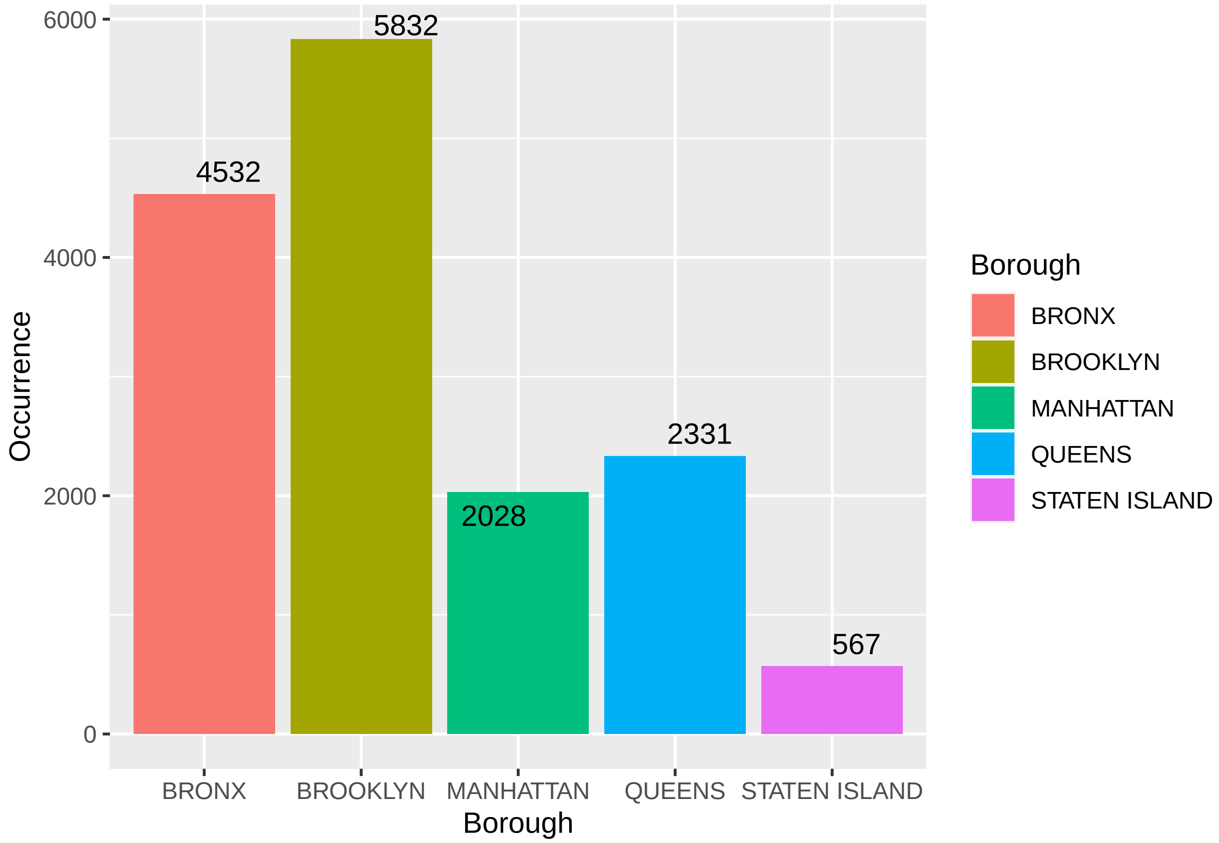
```
##
##           F     M     U
##  8261   335 13490  1499
```

```
##
##     F     M     U
##   335 13490  1465
```

Analyze the data

Grouping the data by borough can give us a rough overview of where the shootings are localized in New York. Looking at the plot below, there already seems to be some sort of pattern amongst four of the boroughs, with Staten Island being the obvious outlier (which could be because it's quite separated from the others).

```
library(ggplot2); library(ggrepel)
grp <- as.data.frame(table(cleaned$BORO, dnn=list('Borough')), responseName='Occurrence')

ggplot(grp, aes(x=Borough, y=Occurrence, fill=Borough)) + geom_bar(stat="identity") +
↪  geom_text_repel(data=grp, aes(label=Occurrence))
```



Over time

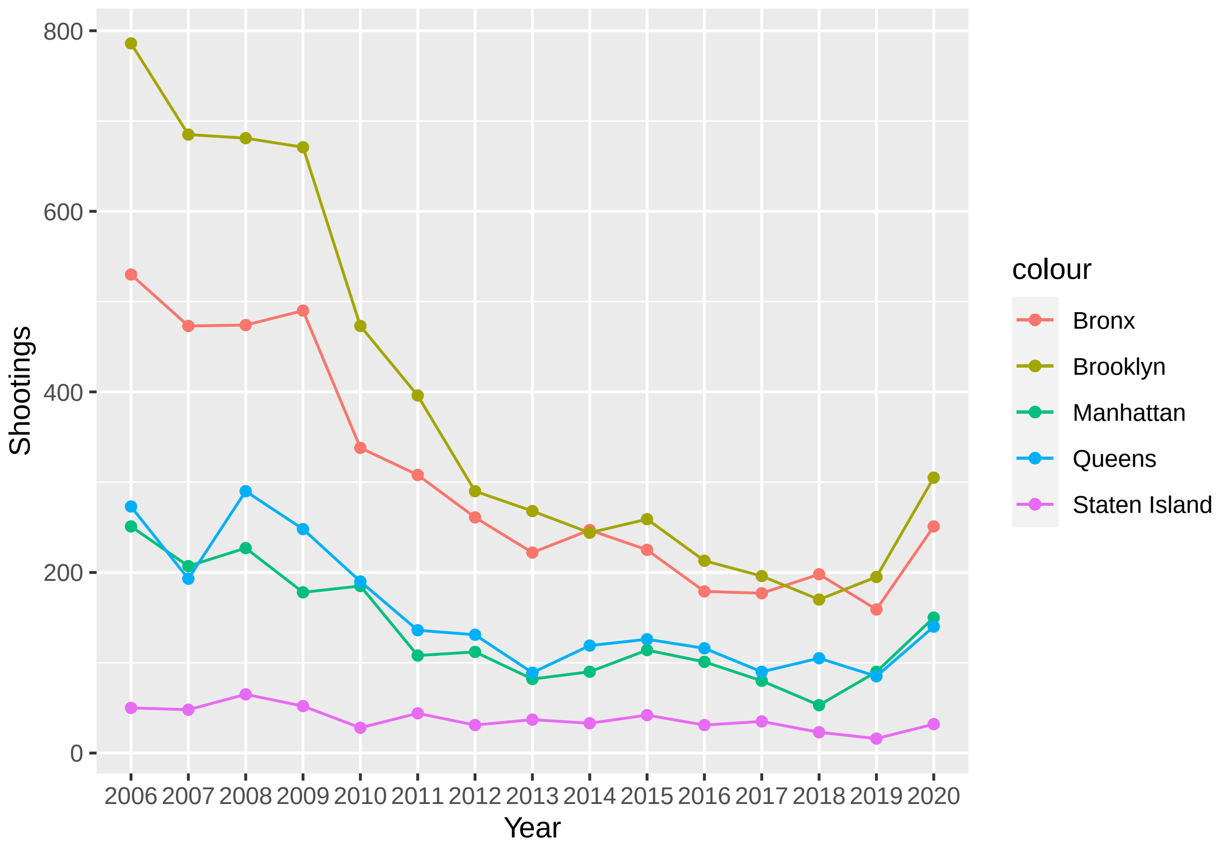
Plotting the shootings over time can give a clearer perspective. However, while attempting to do so, I ran into some issues with converting the date format to something R-friendly despite the built-in function I used seemingly converting some dates while throwing "NA" for others. See below for an example.

```
dt <- as.Date(cleaned[which(cleaned$BORO=='QUEENS'),]$OCCUR_DATE, format='%d/%m/%Y')
head(dt, n=50)
```

```
##  [1] NA           "2010-11-03" "2018-12-03" "2006-04-07" NA
##  [6] "2016-03-07" NA           "2009-05-11" NA           NA
## [11] "2008-06-01" "2018-11-01" NA           "2011-02-09" NA
## [16] "2010-04-07" NA           "2020-04-08" NA           NA
## [21] "2009-08-11" NA           NA           "2017-03-04" "2013-07-01"
## [26] NA           NA           NA           "2020-08-09" NA
## [31] NA           "2020-06-06" "2009-04-04" NA           NA
## [36] "2006-01-10" NA           NA           "2009-01-12" "2015-04-02"
## [41] NA           "2019-03-07" "2006-09-09" NA           NA
## [46] "2020-08-09" NA           NA           "2011-08-06" "2007-02-06"
```

This went on for the rest of the Queens dataset, as well as the other four boroughs. Unfortunately, I couldn't figure out why so I ended up taking a standard approach.

```r
bx <- cleaned[which(cleaned$BORO=='BRONX'),]
by <- cleaned[which(cleaned$BORO=='BROOKLYN'),]
mh <- cleaned[which(cleaned$BORO=='MANHATTAN'),]
qn <- cleaned[which(cleaned$BORO=='QUEENS'),]
si <- cleaned[which(cleaned$BORO=='STATEN ISLAND'),]

bx_y <- as.data.frame(table(substring(bx$OCCUR_DATE, 7, 11)))
by_y <- as.data.frame(table(substring(by$OCCUR_DATE, 7, 11)))
mh_y <- as.data.frame(table(substring(mh$OCCUR_DATE, 7, 11)))
qn_y <- as.data.frame(table(substring(qn$OCCUR_DATE, 7, 11)))
si_y <- as.data.frame(table(substring(si$OCCUR_DATE, 7, 11)))

over_time <- data.frame('Year'=bx_y$Var1, 'Bronx'=bx_y$Freq, 'Brooklyn'=by_y$Freq,
↪  'Manhattan'=mh_y$Freq, 'Queens'=qn_y$Freq, 'Staten Island'=si_y$Freq)

ggplot(over_time, aes(x=Year, y=Shootings, group=1)) +
  geom_line(aes(y=Bronx, color='Bronx')) +
  geom_point(aes(y=Bronx, color='Bronx')) +
  geom_line(aes(y=Brooklyn, color='Brooklyn')) +
  geom_point(aes(y=Brooklyn, color='Brooklyn')) +
  geom_line(aes(y=Manhattan, color='Manhattan')) +
  geom_point(aes(y=Manhattan, color='Manhattan')) +
  geom_line(aes(y=Queens, color='Queens')) +
  geom_point(aes(y=Queens, color='Queens')) +
  geom_line(aes(y=Staten.Island, color='Staten Island')) +
  geom_point(aes(y=Staten.Island, color='Staten Island'))
```



The fluctuations in the shooting rate and the discrete grouping of the boroughs is now quite evident but it's also clear that crime has been steadily decreasing, although I think we all know the reason behind the collective uptick in 2020. The pattern that seemed to be there in the bar chart is definitely strong in this plot: Bronx and Brooklyn seem to follow each other just like Manhattan and Queens.

Fitting a standard regression model solidifies the argument that shootings are decreasing as shown below.

```
library(reshape2); library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

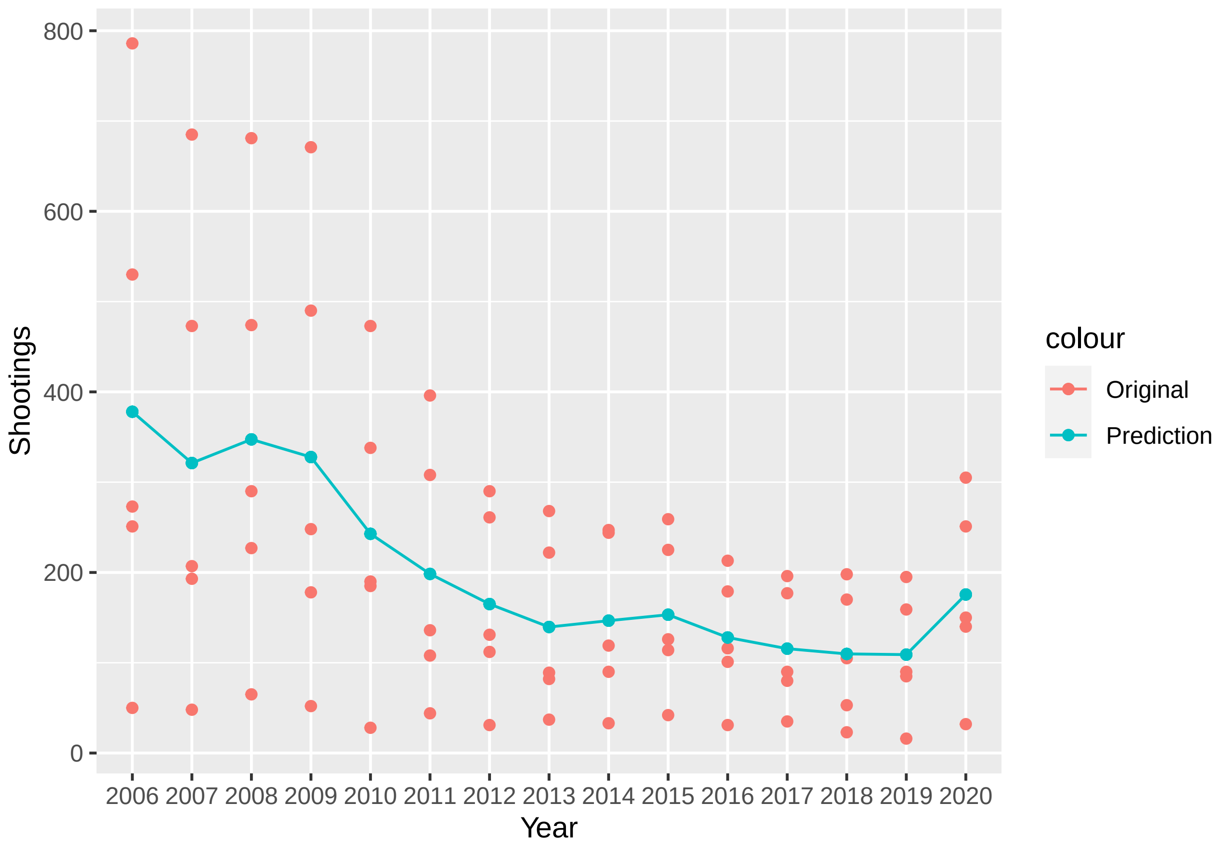
```
mt <- melt(over_time); md <- lm(value ~ Year, data=mt) ## shootings ~ year
```

```
## Using Year as id variables
```

```
mt <- mutate(mt, pred=predict(md))

summary(md)
```

```
##
## Call:
## lm(formula = value ~ Year, data = mt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -328.0  -91.7  -27.2   91.7  408.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   378.00      71.61   5.279 1.89e-06 ***
## Year2007      -56.80     101.27  -0.561   0.5770
## Year2008      -30.60     101.27  -0.302   0.7636
## Year2009      -50.20     101.27  -0.496   0.6219
## Year2010     -135.20     101.27  -1.335   0.1869
## Year2011     -179.60     101.27  -1.774   0.0812 .
## Year2012     -213.00     101.27  -2.103   0.0396 *
## Year2013     -238.40     101.27  -2.354   0.0219 *
## Year2014     -231.40     101.27  -2.285   0.0259 *
## Year2015     -224.80     101.27  -2.220   0.0302 *
## Year2016     -250.00     101.27  -2.469   0.0164 *
## Year2017     -262.40     101.27  -2.591   0.0120 *
## Year2018     -268.20     101.27  -2.648   0.0103 *
## Year2019     -269.00     101.27  -2.656   0.0101 *
## Year2020     -202.40     101.27  -1.999   0.0502 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 160.1 on 60 degrees of freedom
## Multiple R-squared:  0.2894, Adjusted R-squared:  0.1236
## F-statistic: 1.745 on 14 and 60 DF,  p-value: 0.06988
```

```
ggplot(mt, aes(x=Year, y=Shootings, group=1)) +
  geom_point(aes(y=value, color='Original')) +
  geom_line(aes(y=pred, color='Prediction')) +
  geom_point(aes(y=pred, color='Prediction'))
```



Age group

Finally, I grouped the data by age group of the perps and victims to see if there is anything noteworthy. The two bar charts shows the total number of people involved from 2006 to 2020. The perpetrators of one age group obviously didn't target victims who are of the same age group, so a direct comparison is inconsequential but it can give historical insight nevertheless.

```
bx_pa <- as.data.frame(table(bx$PERP_AGE_GROUP))[-c(2, 4),] ## omit erroneous datapoints
by_pa <- as.data.frame(table(by$PERP_AGE_GROUP))[-c(6),]    ## omit erroneous datapoint
mh_pa <- as.data.frame(table(mh$PERP_AGE_GROUP))
qn_pa <- as.data.frame(table(qn$PERP_AGE_GROUP))
si_pa <- as.data.frame(table(si$PERP_AGE_GROUP))

bx_va <- as.data.frame(table(bx$VIC_AGE_GROUP))
by_va <- as.data.frame(table(by$VIC_AGE_GROUP))
mh_va <- as.data.frame(table(mh$VIC_AGE_GROUP))
qn_va <- as.data.frame(table(qn$VIC_AGE_GROUP))
si_va <- as.data.frame(table(si$VIC_AGE_GROUP))

perp_age <- data.frame('Age'=bx_pa$Var1, 'Bronx'=bx_pa$Freq, 'Brooklyn'=by_pa$Freq,
↪  'Manhattan'=mh_pa$Freq, 'Queens'=qn_pa$Freq, 'Staten Island'=si_pa$Freq)

vict_age <- data.frame('Age'=bx_va$Var1, 'Bronx'=bx_va$Freq, 'Brooklyn'=by_va$Freq,
↪  'Manhattan'=mh_va$Freq, 'Queens'=qn_va$Freq, 'Staten Island'=si_va$Freq)
```

```
perp_age <- melt(perp_age, value.name='Amount', variable.name='Perpetrator')
```

```
## Using Age as id variables
```

```
vict_age <- melt(vict_age, value.name='Amount', variable.name='Victim')
```

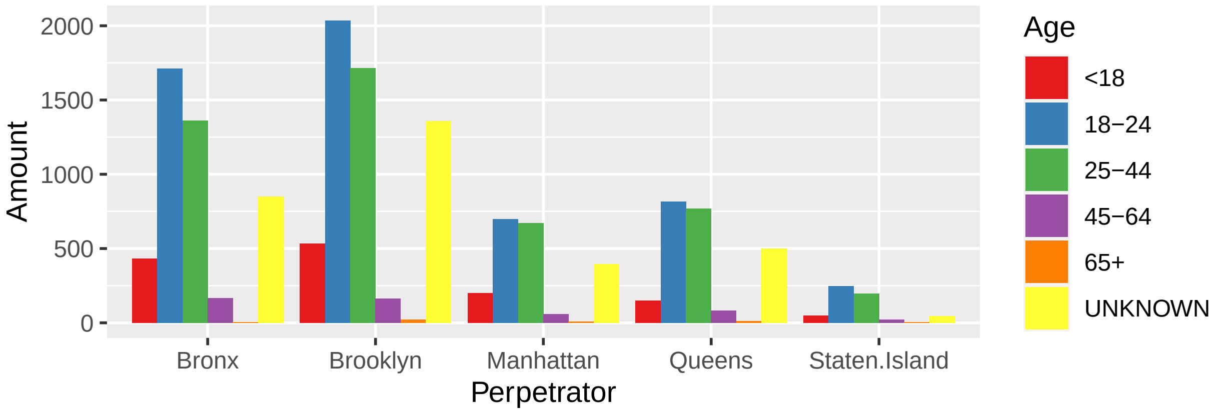
```
## Using Age as id variables
```

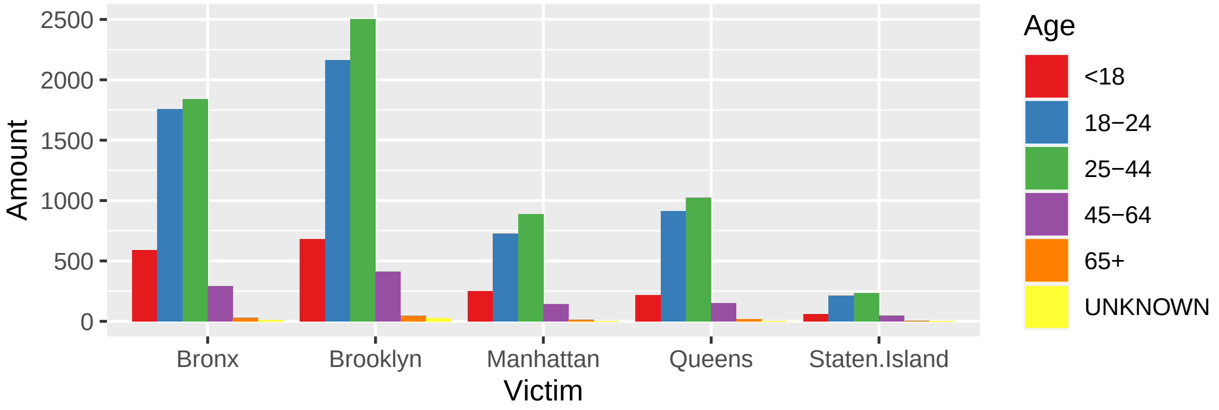
```
pl1 <- ggplot(perp_age, aes(Perpetrator, Amount, fill = Age)) +
 geom_bar(stat="identity", position = "dodge") +
 scale_fill_brewer(palette = "Set1")

pl2 <- ggplot(vict_age, aes(Victim, Amount, fill = Age)) +
 geom_bar(stat="identity", position = "dodge") +
 scale_fill_brewer(palette = "Set1")

require(gridExtra); grid.arrange(pl1, pl2, nrow=2)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```





Results

A simple plot like the one on page 5 could go a long way in helping bring shooting even lower. Comparing a more detailed version of the shootings over time with current events related of that area and general public discourse during that timeframe, politicians and people in power can better understand what changes in society either directly or indirectly caused the shooting rate to rise or fall in the days that followed.

One thing I found interesting in the bar chart above is that what's consistent across all boroughs is the amount of young adult perpetrators (18-24) is higher or substantially higher than of adults (25-44) whereas it's the exact opposite for the victims. If lots of these shootings can be attributed to gang violence, it's possible that the young perpetrators are pressured due to gang responsibilities in order to fit in. Something worth noting is that across all boroughs and age groups, the number of victims are higher than perpetrators, which implies the perps are often *too* successful.

As this data was obtained directly from the city government, I don't expect reasons for bias in this data as all information are purely factual descriptors. However, as I am from New York, I am easily biased in the way I would approach the analysis. For example, I am aware that shootings in Brooklyn or the Bronx are more likely to be related to gang violence or ongoing crimes as opposed to other boroughs where it's more likely to be standalone and one-offs. So, I might come to different conclusions because of subconscious bias I injected during the data analysis or the writeup.