NYPD Shooting Incident (Historic)

2022-07-14

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

```
data <- read_csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv')
```

The source data is provided by the New York Police Department listing shooting incidents from 2006 to 2021. The data set includes information on the the perpetrator, the victim, the location of the incident and when the incident took place.

The Question

After the a brief examination of the data, a question that comes to mind is:

Which boroughs in New York are safe to live in?

It is important to note that the data used is historic data of shooting incidents in New York from 2006 to 2021. Therefore, "safety" from the question is going to be defined as the chance of not getting involved in a shooting incident and dying.

With that out the way, it's time to start tidying up the data set.

Data Cleaning

```
# took out some unnecessary columns
# and changed the date column to a date data type
data <- data %>%
  select(c(OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG, VIC_AGE_GROUP,
           VIC_SEX, VIC_RACE)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

# made some categorical columns into factors
cols_to_factor <- c('BORO', 'VIC_AGE_GROUP', 'VIC_SEX', 'VIC_RACE')
data <- data %>%
  mutate_at(cols_to_factor, factor)

# No NA values seen for these columns
summary(data)
```

```
##    OCCUR_DATE                   BORO       STATISTICAL_MURDER_FLAG
##  Min.   :2006-01-01   BRONX        : 7402   Mode :logical
```

```
##  1st Qu.:2009-05-10   BROOKLYN     :10365   FALSE:20668
##  Median :2012-08-26   MANHATTAN    : 3265   TRUE :4928
##  Mean   :2013-06-13   QUEENS       : 3828
##  3rd Qu.:2017-07-01   STATEN ISLAND:  736
##  Max.   :2021-12-31
##
##  VIC_AGE_GROUP   VIC_SEX                              VIC_RACE
##  <18    : 2681   F: 2403   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  18-24  : 9604   M:23182   ASIAN / PACIFIC ISLANDER      :  354
##  25-44  :11386   U:   11   BLACK                         :18281
##  45-64  : 1698             BLACK HISPANIC                : 2485
##  65+    :  167             UNKNOWN                       :   65
##  UNKNOWN:   60             WHITE                         :  660
##                            WHITE HISPANIC                : 3742
```

```
head(data)
```

```
## # A tibble: 6 x 6
##   OCCUR_DATE BORO      STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX VIC_RACE
##   <date>     <fct>     <lgl>                   <fct>         <fct>   <fct>
## 1 2006-08-27 BRONX     TRUE                    25-44         F       BLACK HISP~
## 2 2011-03-11 QUEENS    FALSE                   65+           M       WHITE
## 3 2021-04-14 BRONX     TRUE                    18-24         M       BLACK
## 4 2021-12-10 BRONX     FALSE                   25-44         M       BLACK
## 5 2021-02-22 MANHATTAN FALSE                   25-44         M       BLACK HISP~
## 6 2021-03-07 BROOKLYN  TRUE                    25-44         M       WHITE HISP~
```

Visualization and Analysis

To start looking at safety, I want to look at the chance that a person may die from a shooting incident from any borough.

Chance of Death

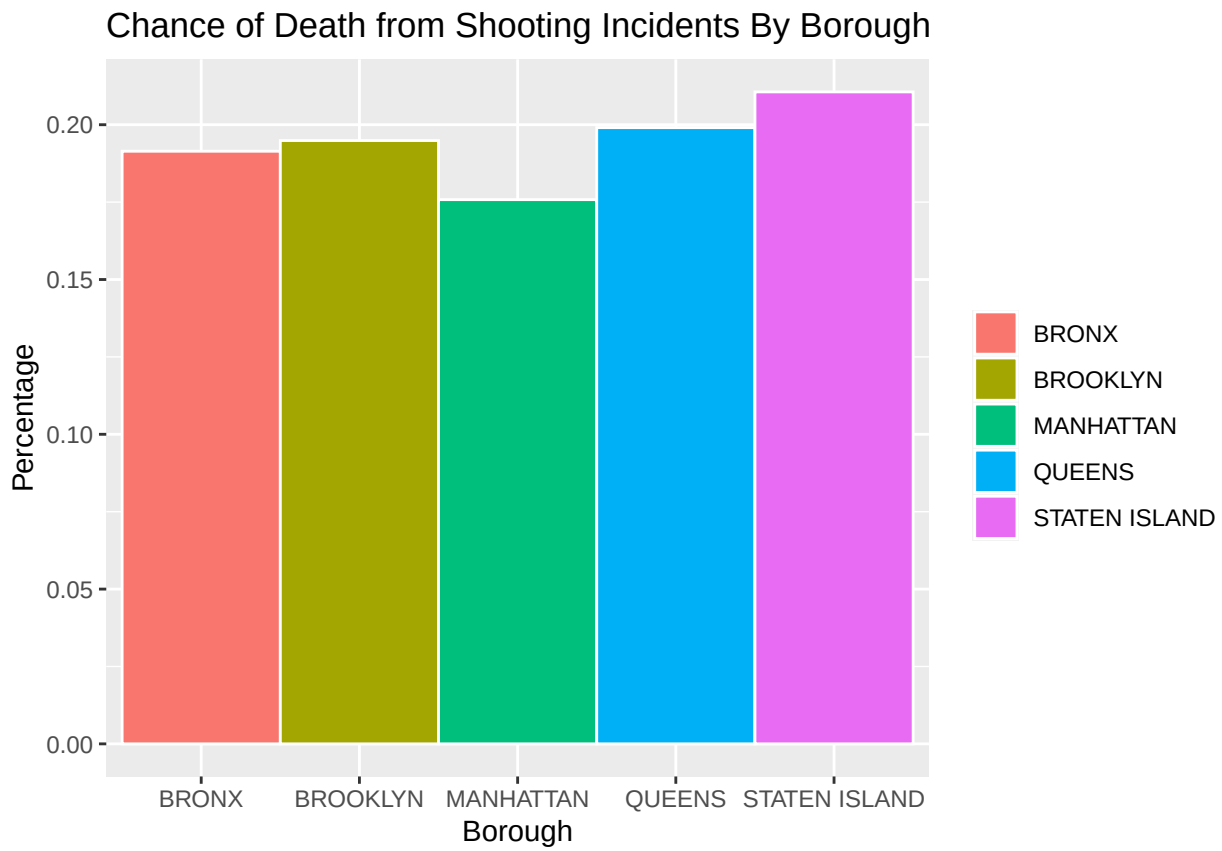
To find the likelihood that a person may die from a shooting, I decided to look at the column named 'STATISTICAL_MURDER_FLAG'. According to NYPD who provided this data set, this column is a conditional column, where TRUE would indicate a victim's death resulting from a shooting incident, i.e. a murder.

The task is then simple, to find the chance of death from getting involved in any shooting incident in New York would then be the sum of the TRUE values in divided by total incidents.

```
# grouping by borough and calculating the chance of death from
# dividing the TRUE values from total incidents of each group.
boro <- data %>%
  group_by(BORO) %>%
  summarize(murders = sum(STATISTICAL_MURDER_FLAG) / n())

# creating bar graph
boro %>%
  ggplot(aes(x = BORO, y = murders, fill = BORO)) +
```

```
  geom_bar(stat = 'identity', width = 1, color = 'white') +
  labs(title = "Chance of Death from Shooting Incidents By Borough",
       y = 'Percentage', x = 'Borough', color = 'Boroughs') +
  scale_fill_discrete(name=NULL)
```



It seems that all five boroughs in New York have roughly a 20% chance of murder from any shooting incident. That's every 1 in 5 shooting incident the victim is likely to get killed. That does not sound very good.

Now let's talk about the victim.

The Victim

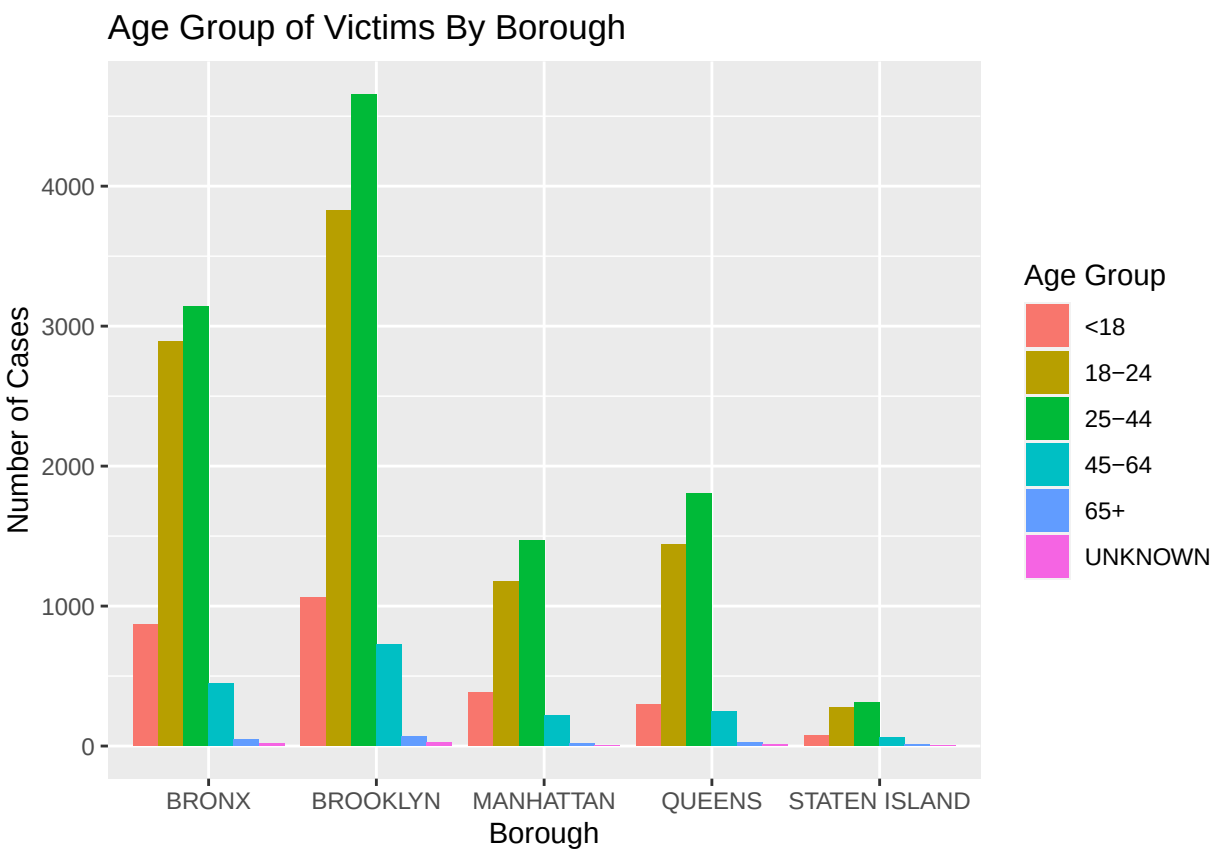
To find the safety of the boroughs, we are going to look at the demographics of the victims. More specifically, we are going to examine their age, race, and sex.

Age

```
# group by victim age showing total for each age group
victim_ages <- data %>%
  group_by(BORO, VIC_AGE_GROUP) %>%
  summarize(num_cases = n())

# creating bar chart
```

```
victim_ages %>%
  ggplot(aes(x = BORO, y = num_cases, fill = VIC_AGE_GROUP)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(title = 'Age Group of Victims By Borough',
       x = 'Borough', y = 'Number of Cases') +
  scale_fill_discrete(name='Age Group')
```



From the chart displayed above, there's a few things that can be glimpsed. Initially, we spoke about the probability that the victim is killed from any shooting in each borough. Here we can see that Brooklyn has the most cases and Bronx coming in close. Staten Island has the least amount of incidents in comparison.

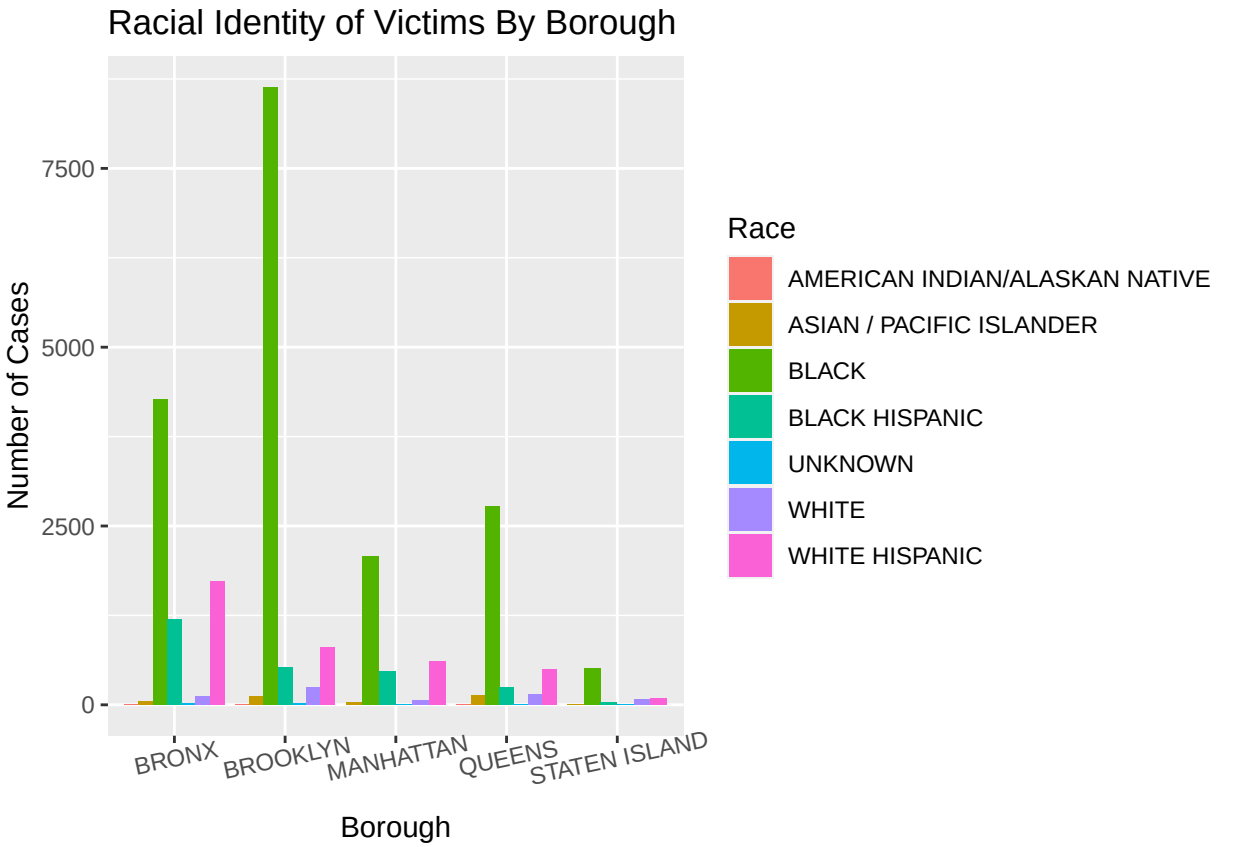
Next, we notice that there is a large discrepancy in the age groups of the victim. For every borough, the vast majority of them were between ages 18-44, the age range for young adults. People falling into that age group would be at a higher risk of being involved in shooting incidents, supposedly. The reason could be that that age group is where the most social interactions would occur.

Race

```
# group by borough and race of victims and get the total count for grouping
victim_race <- data %>%
  group_by(BORO, VIC_RACE) %>%
  summarize(num_cases = n())

# bar chart
victim_race %>%
```

```
ggplot(aes(x = BORO, y = num_cases, fill = VIC_RACE)) +
geom_bar(stat = 'identity', position = 'dodge') +
labs(title = 'Racial Identity of Victims By Borough',
     x = 'Borough', y = 'Number of Cases') +
scale_fill_discrete(name='Race') +
theme(axis.text.x = element_text(angle = 12))
```



The first thing that pops out from the chart is the green bars. The green bars indicate victims that are considered Black. There is an overwhelming number of cases where Blacks end up as victims to shootings. White Hispanics are the second-most typical race to be victims but not by a long shot. Furthermore, at a quick glance, the distribution between races for each borough looks very similar to one another ranking:

1. Black
2. White Hispanic
3. Black Hispanic
4. White
5. Other

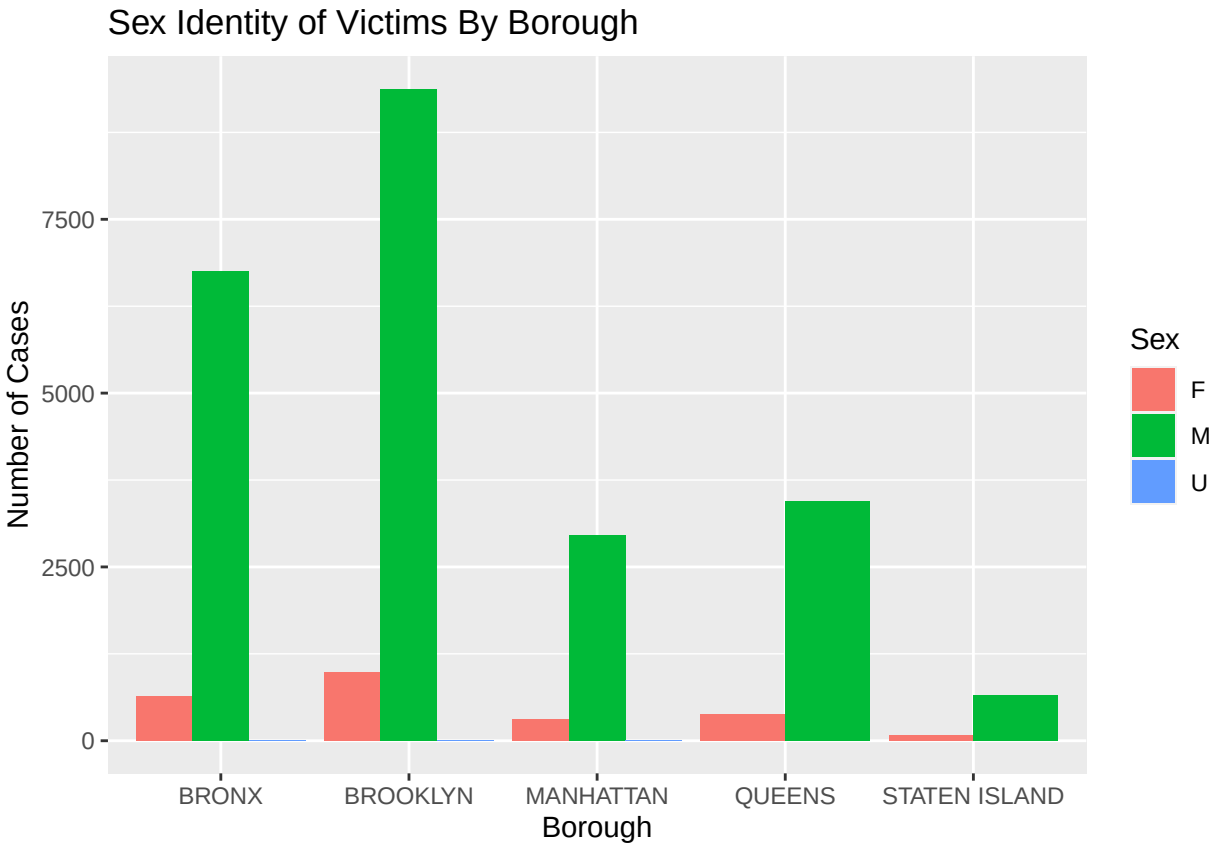
People from the lower end of the socioeconomic spectrum is more likely to be victims of shootings whether it be by chance or involvement of some sort.

Sex

So far we have seen that victims tend to be Black young adults that are getting involved in the shooting incidents in New York. From this we can sort of infer that the sex of the victims are most likely going to be predominantly male. But to confirm this speculation, let's look at the data.

```r
# grouping by borough and sex identity of victims
victim_sex <- data %>%
  group_by(BORO, VIC_SEX) %>%
  summarize(num_cases = n())

# bar chart
victim_sex %>%
  ggplot(aes(x = BORO, y = num_cases, fill = VIC_SEX)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(title = 'Sex Identity of Victims By Borough',
       x = 'Borough', y = 'Number of Cases') +
  scale_fill_discrete(name='Sex')
```



We can see that the speculation that males would be the predominant sex for the victims was correct. We see that every borough is the same in that males are extremely more likely to be victims to shootings.

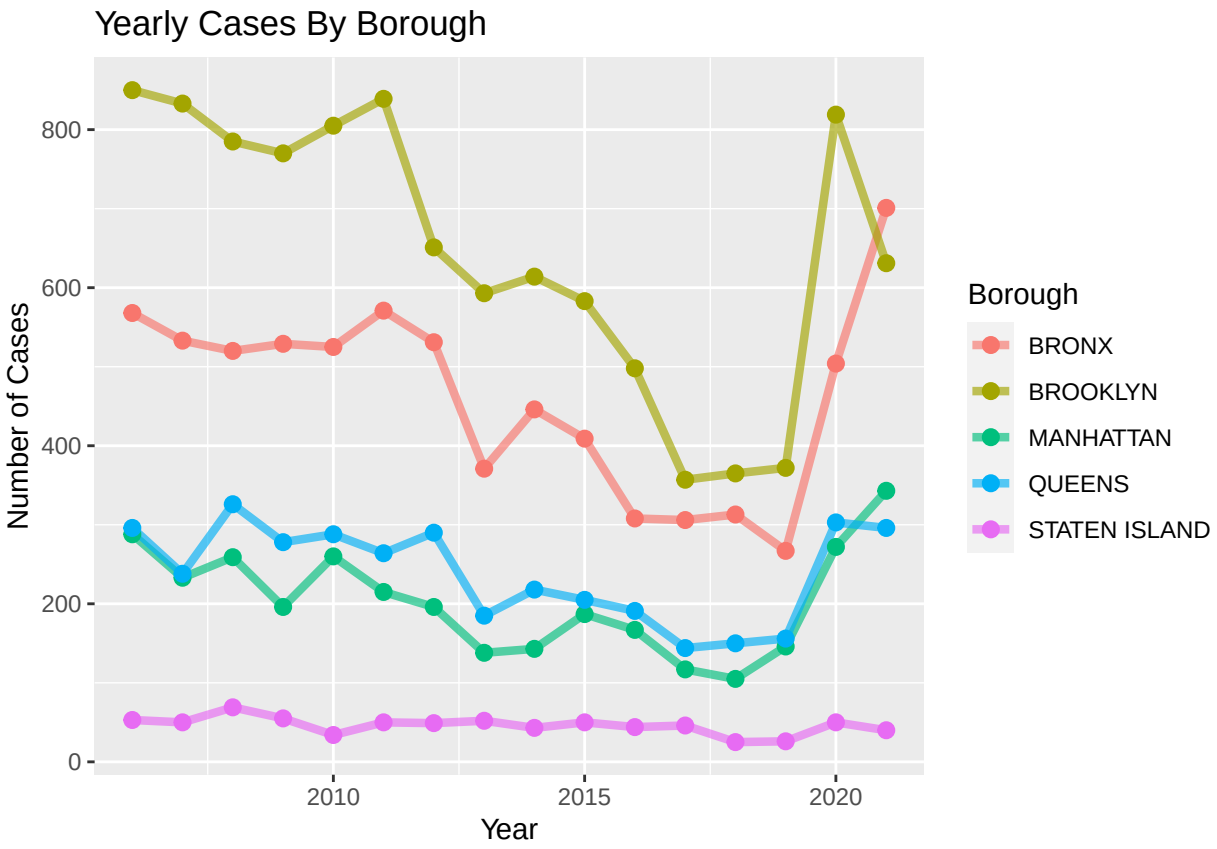
Cases Over Time

So we have a general identification for the victims and we find that there have been a significant amount of shooting cases in each borough. But have they changed over the years? Are shootings still rampant or have they decreased over time?

```r
# round to closest year
data$OCCUR_DATE <- floor_date(data$OCCUR_DATE, unit = 'year')
```

```r
# group by year and borough
time_data <- data %>%
  group_by(BORO, OCCUR_DATE) %>%
  summarize(num_cases = n())

# line graph
time_data %>%
  ggplot(aes(x = OCCUR_DATE, y = num_cases, color = BORO)) +
  geom_line(size = 1.5, alpha = 0.65) +
  geom_point(size = 2.5) +
  labs(title = 'Yearly Cases By Borough',
       x = 'Year', y = 'Number of Cases') +
  scale_color_discrete(name = 'Borough')
```



Starting from 2006 when the data set started, there has been a general decrease in shootings in every borough except Staten Island which remains relatively constant albeit their low number of cases. Interestingly, we see that shootings suddenly spiked in 2020 and boroughs like the Bronx and Manhattan setting new highs in the number of cases in 2021.

Generally the cases were decreasing in areas where shootings were a serious problem. This may be due to some gun regulations over the years from the state of New York or just in terms of education. Yet there was a spike in recent years. Coincidentally, 2020 to 2021 was the year COVID-19 emerged and was a major problem. Within that context, the spike in cases may be a result of several things such as mental health issues or disputes from a loss of income source.

Modeling Predicted Murders

We are going to model predicted murders by the number of total shootings that occurred in New York.

```
# creating a data set for total shootings and incidents considered murders
murders <- data%>%
  group_by(BORO) %>%
  summarize(total_murders = sum(STATISTICAL_MURDER_FLAG == TRUE),
            total_alive = sum(STATISTICAL_MURDER_FLAG == FALSE),
            total_cases = n()) %>%
  ungroup() %>%
  mutate(cum_shootings = cumsum(total_cases)) %>%
  mutate(cum_murders = cumsum(total_murders))

# creating linear model
model <- lm(cum_murders ~ cum_shootings, data = murders)

# creating prediction
pred <- murders %>%
  mutate(pred = predict(model))

# graphing line graph of model
pred %>%
  ggplot(aes(x = cum_shootings, y = cum_murders, color = 'blue')) +
  geom_point(alpha = 0.65, size = 1.5) +
  geom_line(aes(x = cum_shootings, y = pred, color = 'red'), alpha = 0.65, size = 1.5) +
  labs(title = 'Model Predicting Amount Cases Considered Murders',
       x = 'Number of Shootings', y = 'Number of Murders') +
  scale_color_manual(labels = c('Actual', 'Predicted'), values = c('blue', 'red'))
```

The original data showed a relatively linear relationship once the cumulative sum was taken to find the total shootings and total murder incidents. A linear model was then used to try to predict the number of murders given a number of total shootings. The model was shown to be very close to what was perceived in the original data, landing almost on top of each other. Looking at the graph, again it looks like the number of murders compared to the total cases in around 1 in 5.

Conclusion

Going back to the question: which boroughs in New York are safe to live in, the answer probably Staten Island. According to cases over time, they have a relatively low number of shooting incidents in comparison to the other boroughs and did not have a heavy resurgence in recent years.

However, regardless where one might live, there is a 1 in 5 chance a shooting case results in death of the victim. Victims tend to be young adults (ages 18-44) black males. So if you don't belong to these demographics then the risk of danger should be reduced.

It is true that these characteristics indicate a high chance of becoming a victim to shootings according to the data. However, we do not know whether there was a relationship between the perpetrator and victims nor the motive behind the incidents according to the data provided. Therefore, we can't say exactly which borough is safer because we do not know whether these incidents are personal or random occurrences. One would be better than the other simply by just not getting involved.

Furthermore, we defined safety in this case to be not getting involved in a shooting incident and dying however safety is more than just shootings. It would be interesting to see what percent of criminal charges were shootings and their motives. Also it would be interesting to find why Brooklyn has such a high number of shooting incidents.

Personal Bias

Coming into this data exploration, I may have had confirmation bias where I thought the victims to be mostly young black adults and I was tempted to look solely on that data. In terms of bias management, I tried to look objectively at the data and not try to overly emphasize any demographic to create a story that pointed heavily in one direction. I tried to make the story more like a layer of discovery and included as much source of comparison as the data can provide to give a more generalized picture.

Session Info

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.9
##  [5] purrr_0.3.4     readr_2.1.2     tidyr_1.2.0     tibble_3.1.6
##  [9] ggplot2_3.3.6   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2 xfun_0.30        haven_2.5.0      colorspace_2.0-3
##  [5] vctrs_0.4.1      generics_0.1.3   htmltools_0.5.2  yaml_2.3.5
##  [9] utf8_1.2.2       rlang_1.0.2      pillar_1.7.0     glue_1.6.2
## [13] withr_2.5.0      DBI_1.1.3        bit64_4.0.5      dbplyr_2.2.1
## [17] modelr_0.1.8     readxl_1.4.0     lifecycle_1.0.1  munsell_0.5.0
## [21] gtable_0.3.0     cellranger_1.1.0 rvest_1.0.2      evaluate_0.15
## [25] labeling_0.4.2   knitr_1.39       tzdb_0.3.0       fastmap_1.1.0
## [29] curl_4.3.2       parallel_4.1.3   fansi_1.0.2      highr_0.9
## [33] broom_1.0.0      backports_1.4.1  scales_1.2.0     vroom_1.5.7
## [37] jsonlite_1.8.0   farver_2.1.1     bit_4.0.4        fs_1.5.2
## [41] hms_1.1.1        digest_0.6.29    stringi_1.7.6    grid_4.1.3
## [45] cli_3.3.0        tools_4.1.3      magrittr_2.0.2   crayon_1.5.1
## [49] pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.3       reprex_2.0.1
## [53] assertthat_0.2.1 rmarkdown_2.14   httr_1.4.3       rstudioapi_0.13
## [57] R6_2.5.1         compiler_4.1.3
```