New York Shooting (Historic)

Solomon

2024-07-25

This project sumerizes the NYPD shooting data. Additional information will be found in the below link.

R Markdown

Import NYPD Shooting Incident Data (Historic)

The URL have a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset. https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Step 1 Start an Rmd document and loading libraries

```
url_in = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd <- read_csv(url_in)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

##Step 2: Tydying and transform our data. visualization and analysis: under this step it is crucial to uncover insights and make informed decisions. This stage

We can also embed plots. In the following chunk of code I will identify the first six line of data set in order to learn about the table.

```r
head(nypd)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

##Step 2.1 **Cleaning the data aka Tidying** ## Data cleaning includes handling missing values, removes duplicates, correct errors and standardize formats. Cleaning the data aka Tidying ## Data cleaning includes handling missing values, removes duplicates, correct errors and standardize formats.

We can also embed plots, for example:

```r
nypd <- subset(nypd, select = -c(JURISDICTION_CODE, Latitude, Longitude, Lon_Lat))
```

```r
nypd_2 = nypd %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, LOC_OF_OCCUR_DESC, PRECINCT,
                         LOC_CLASSFCTN_DESC, STATISTICAL_MURDER_FLAG, PERP_SEX, PERP_RACE,
                         PERP_AGE_GROUP, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, X_COORD_CD, Y_COORD_CD)
```

Return the new dataset

```r
head(nypd)
```

```
## # A tibble: 6 x 17
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 11 more variables: LOC_CLASSFCTN_DESC <chr>, LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>
```

```
nypd_2 = nypd %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, LOC_OF_OCCUR_DESC, PRECINCT,
                         LOC_CLASSFCTN_DESC, STATISTICAL_MURDER_FLAG, PERP_SEX, PERP_RACE,
                         PERP_AGE_GROUP, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, X_COORD_CD, Y_COORD_CD)
```

```
library(sf)
library(spData)
```

```
## Warning: package ’spData’ was built under R version 4.3.3
```

```
## To access larger datasets in this package, install the spDataLarge
## package with: ‘install.packages(’spDataLarge’,
## repos=’https://nowosad.github.io/drat/’, type=’source’)‘
```

```
library(tmap)
```

```
## Warning: package ’tmap’ was built under R version 4.3.3
```

```
## Breaking News: tmap 3.x is retiring. Please test v4, e.g. with
## remotes::install_github(’r-tmap/tmap’)
```

```
library(mapview)
```

```
## Warning: package ’mapview’ was built under R version 4.3.3
```

```
library(viridis)
```

```
## Warning: package ’viridis’ was built under R version 4.3.3
```

```
## Loading required package: viridisLite
```

```
library(ggplot2)
library(RColorBrewer)
library(knitr)
```

```
## Warning: package ’knitr’ was built under R version 4.3.3
```

```
lapply(nypd_2, function(x) sum(is.na(x)))
```

```
## $INCIDENT_KEY
## [1] 0
##
## $OCCUR_DATE
## [1] 0
##
## $OCCUR_TIME
## [1] 0
##
## $BORO
## [1] 0
##
## $LOC_OF_OCCUR_DESC
## [1] 25596
##
## $PRECINCT
## [1] 0
##
## $LOC_CLASSFCTN_DESC
## [1] 25596
##
## $STATISTICAL_MURDER_FLAG
## [1] 0
##
## $PERP_SEX
## [1] 9310
##
## $PERP_RACE
## [1] 9310
##
## $PERP_AGE_GROUP
## [1] 9344
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
##
## $X_COORD_CD
## [1] 0
##
## $Y_COORD_CD
## [1] 0
```

Identifying data types are essentials for accurate analysis, effective data cleaning, appropriate data transformation and insightful visualization and optimization. There are afair amount of unidentifiable amount

# of data in the data set. I will replace NA with "UNKNOWN"

##The data type need to be converted are the following: **INCIDENT_KEY** *SHOULD BE TREATED AS A STRING* **OCCUR_DATE** *SHOULD BE TRATED AS A FACTOR* **OCCUR_TIME** *SHOULD BE TRATED AS A FACTOR* **BORO** *SHOULD BE TRATED AS A FACTOR* **PREP_AGE_GROUP** *SHOULD BE TRATED AS A FACTOR* **PREP_SEX** *SHOULD BE TRATED AS A FACTOR* **PREP_RACE** *SHOULD BE TRATED AS A FACTOR* **VIC_AGE_GROUP** *SHOULD BE TRATED AS A FACTOR* **VIC_SEX** *SHOULD BE TRATED AS A FACTOR* **VIC_RACE** *SHOULD BE TRATED AS A FACTOR* **X_COORD_CD** *SHOULD BE TRATED AS A FACTOR* **Y_COORD_CD** *SHOULD BE TRATED AS A FACTOR*

```r
unique_values <- sapply(lapply(nypd_2, unique), length)
print(unique_values)
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                   22394                    6095                    1423
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       5                       3                      77
##      LOC_CLASSFCTN_DESC STATISTICAL_MURDER_FLAG                PERP_SEX
##                      11                       2                       5
##               PERP_RACE          PERP_AGE_GROUP           VIC_AGE_GROUP
##                       9                      12                       7
##                 VIC_SEX                VIC_RACE              X_COORD_CD
##                       3                       7                   12706
##              Y_COORD_CD
##                   12918
```

```r
nypd_2 = nypd_2 %>%
  replace_na(list(OCCUR_DATE = "UNKNOWN",
             OCCUR_TIME = "UNKNOWN",
             BORO = "UNKNOWN",
             PERP_AGE_GROUP = "UNKNOWN",
             PERP_SEX = "UNKNOWN",
             PERP_RACE = "UNKNOWN",
             VIC_AGE_GROUP = "UNKNOWN",
             VIC_SEX = "UNKNOWN",
             VIC_RACE = "UNKNOWN"))
```

```r
head(nypd_2)
```

```
## # A tibble: 6 x 16
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## 6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
## # i 10 more variables: LOC_CLASSFCTN_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, PERP_AGE_GROUP <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>
```

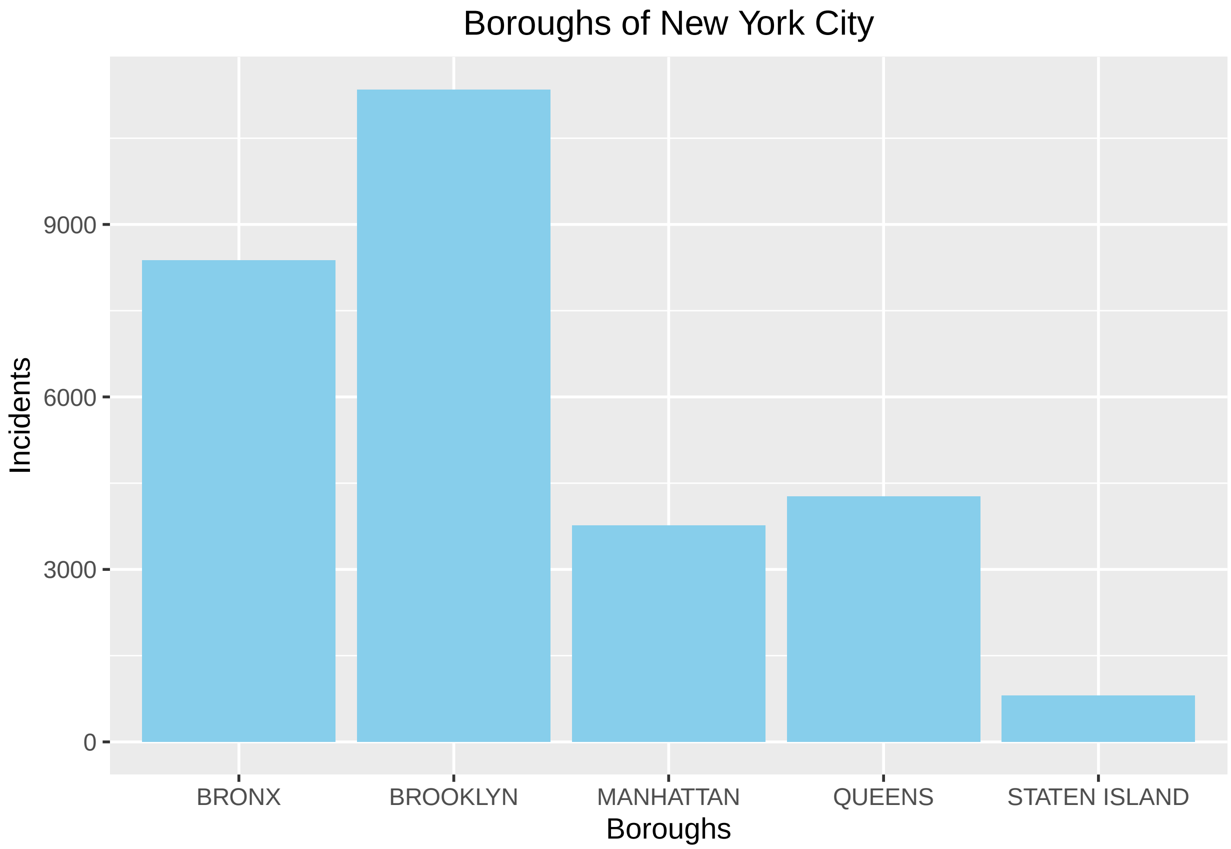
```r
nypd_2 = nypd_2 %>%
  mutate(INCIDENT_KEY = as.character(INCIDENT_KEY),
         OCCUR_DATE = as.factor(OCCUR_DATE),
         OCCUR_TIME = as.character(OCCUR_TIME),
         BORO = as.factor(BORO),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
         PERP_SEX = as.factor(PERP_SEX),
         PERP_RACE = as.factor(PERP_RACE),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE),
         X_COORD_CD = as.factor(X_COORD_CD),
         Y_COORD_CD = as.factor(Y_COORD_CD))
```

```
summary(nypd_2)
```

```
##  INCIDENT_KEY            OCCUR_DATE     OCCUR_TIME                    BORO
##  Length:28562       07/05/2020:   47   Length:28562       BRONX        : 8376
##  Class :character   09/04/2011:   31   Class :character   BROOKLYN     :11346
##  Mode  :character   07/26/2020:   29   Mode  :character   MANHATTAN    : 3762
##                     08/11/2007:   26                      QUEENS       : 4271
##                     08/27/2022:   25                      STATEN ISLAND:  807
##                     09/04/2006:   25
##                     (Other)   :28379
##  LOC_OF_OCCUR_DESC     PRECINCT     LOC_CLASSFCTN_DESC STATISTICAL_MURDER_FLAG
##  Length:28562       Min.   :  1.0   Length:28562       Mode :logical
##  Class :character   1st Qu.: 44.0   Class :character   FALSE:23036
##  Mode  :character   Median : 67.0   Mode  :character   TRUE :5526
##                     Mean   : 65.5
##                     3rd Qu.: 81.0
##                     Max.   :123.0
##
##     PERP_SEX              PERP_RACE     PERP_AGE_GROUP  VIC_AGE_GROUP
##  (null) : 1141   BLACK          :11903   UNKNOWN:12492   <18    : 2954
##  F      :  444   UNKNOWN        :11147   18-24  : 6438   1022   :    1
##  M      :16168   WHITE HISPANIC: 2510   25-44  : 6041   18-24  :10384
##  U      : 1499   BLACK HISPANIC: 1392   <18    : 1682   25-44  :12973
##  UNKNOWN: 9310   (null)        : 1141   (null) : 1141   45-64  : 1981
##                  WHITE         :  298   45-64  :  699   65+    :  205
##                  (Other)       :  171   (Other):   69   UNKNOWN:   64
##  VIC_SEX                            VIC_RACE            X_COORD_CD
##  F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11   1017119.4375:   66
##  M:25790   ASIAN / PACIFIC ISLANDER      :  440   1008276     :   47
##  U:   12   BLACK                         :20235   1026387     :   47
##            BLACK HISPANIC                : 2795   936721.6875 :   44
##            UNKNOWN                       :   70   1017141     :   44
##            WHITE                         :  728   1006434     :   42
##            WHITE HISPANIC                : 4283   (Other)     :28272
##         Y_COORD_CD
##  183909.34375:   66
##  183623      :   47
##  262634      :   47
##  172119.4375 :   44
##  183798      :   44
##  244344      :   43
##  (Other)     :28271
```

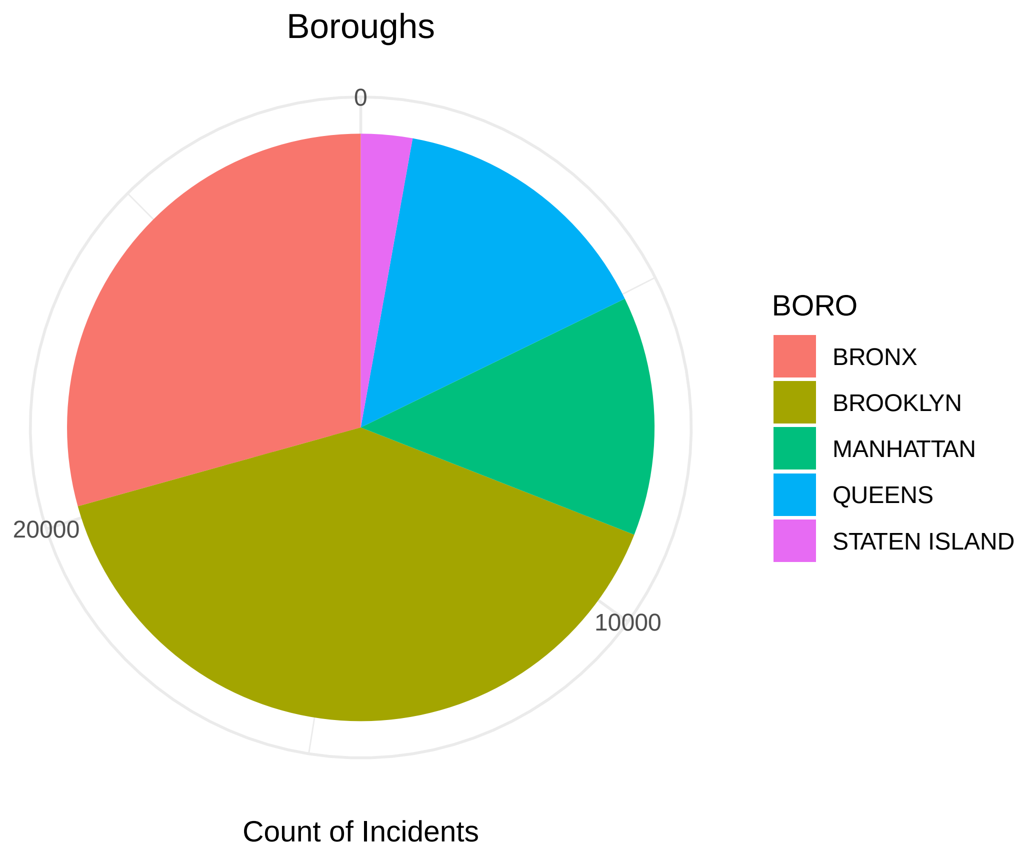
#Step 3 Visualization. Visuals can transform complex datasets into understandable and actionable insights. Charts, graphs, and maps make it easier to see patterns, trends, and outliers that might not be apparent in raw data.

```r
g <- ggplot(nypd_2, aes(x = BORO)) + geom_bar(fill = "skyblue") +
  labs(title = "Boroughs of New York City",
       x = "Boroughs",
       y = "Incidents") +
  theme(plot.title = element_text(hjust = 0.5))
g
```



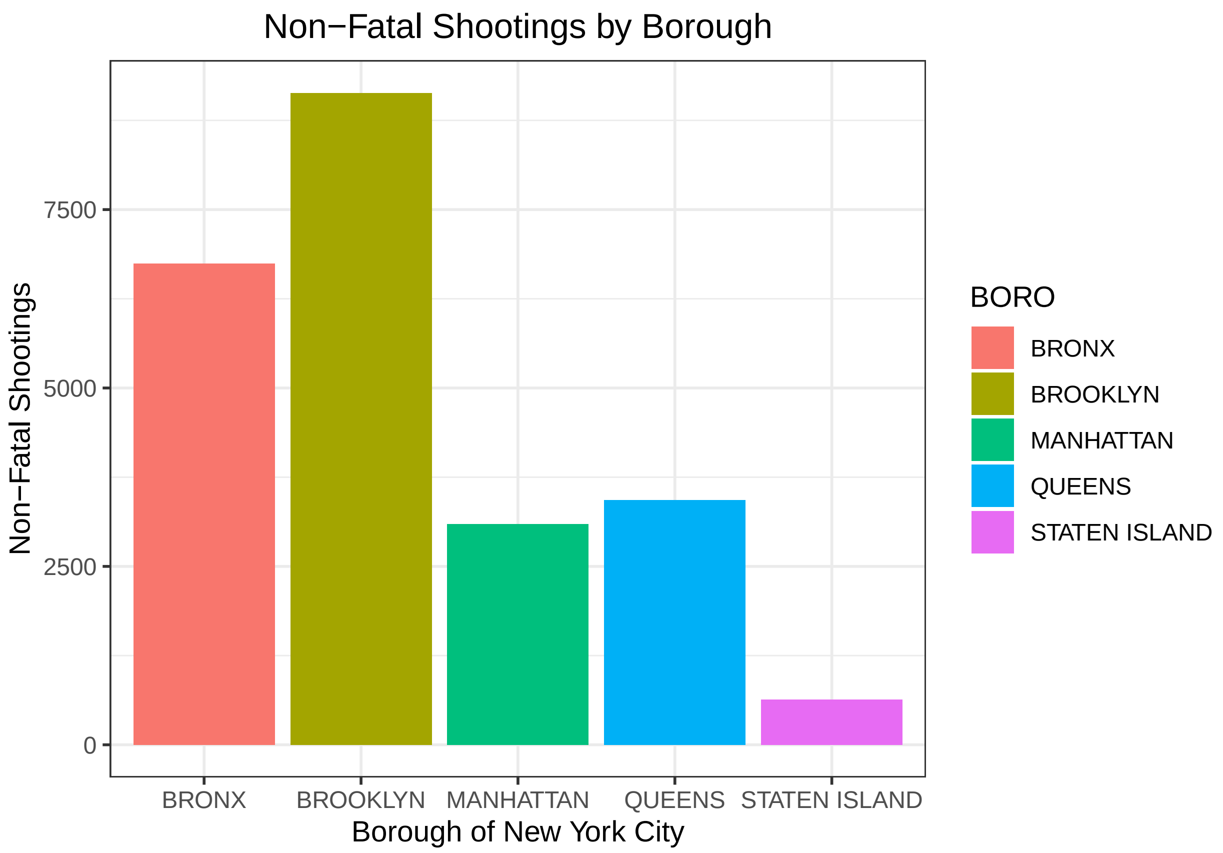
##More shooting incidents occur in Brooklyn and Bronx than the other boroughs. Staten Island has the fewest shooting incidents as you may see it in the bar chart.

```r
library(ggplot2)

g <- ggplot(nypd_2, aes(x = "", fill = BORO)) +
  geom_bar(width = 1, stat = "count") +
  coord_polar(theta = "y") +
  labs(title = "Boroughs",
       x = "",
       y = "Count of Incidents") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))

g
```



##Let's look at some bar charts over time per borough. I want to see if maybe the excess shootings are due to an outlier time period where the number of shootings was way up, or if there's just a steady amount of shootings in Brooklyn that's higher than the other boroughs. So it looks like Brooklyn has the highest number of shootings with Bronx second in line.

```r
nypd_2 %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = BORO, fill = BORO)) +
  geom_bar() +
  theme_bw() +
  labs(x = "Borough of New York City",
       y = "Non-Fatal Shootings",
       title = "Non-Fatal Shootings by Borough") +
  theme(plot.title = element_text(hjust = 0.5))
```

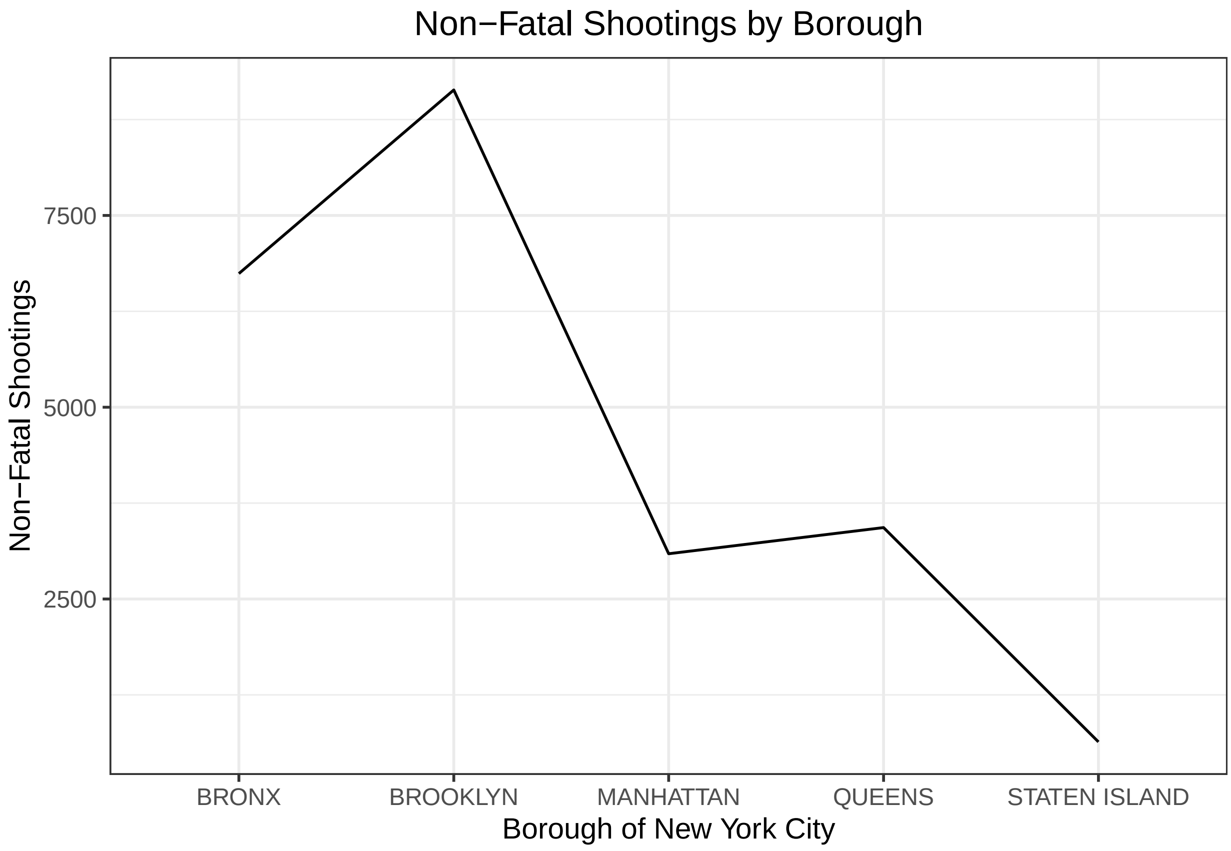


This code should produce a line chart where each line represents the trend of shootings over time for a specific borough.

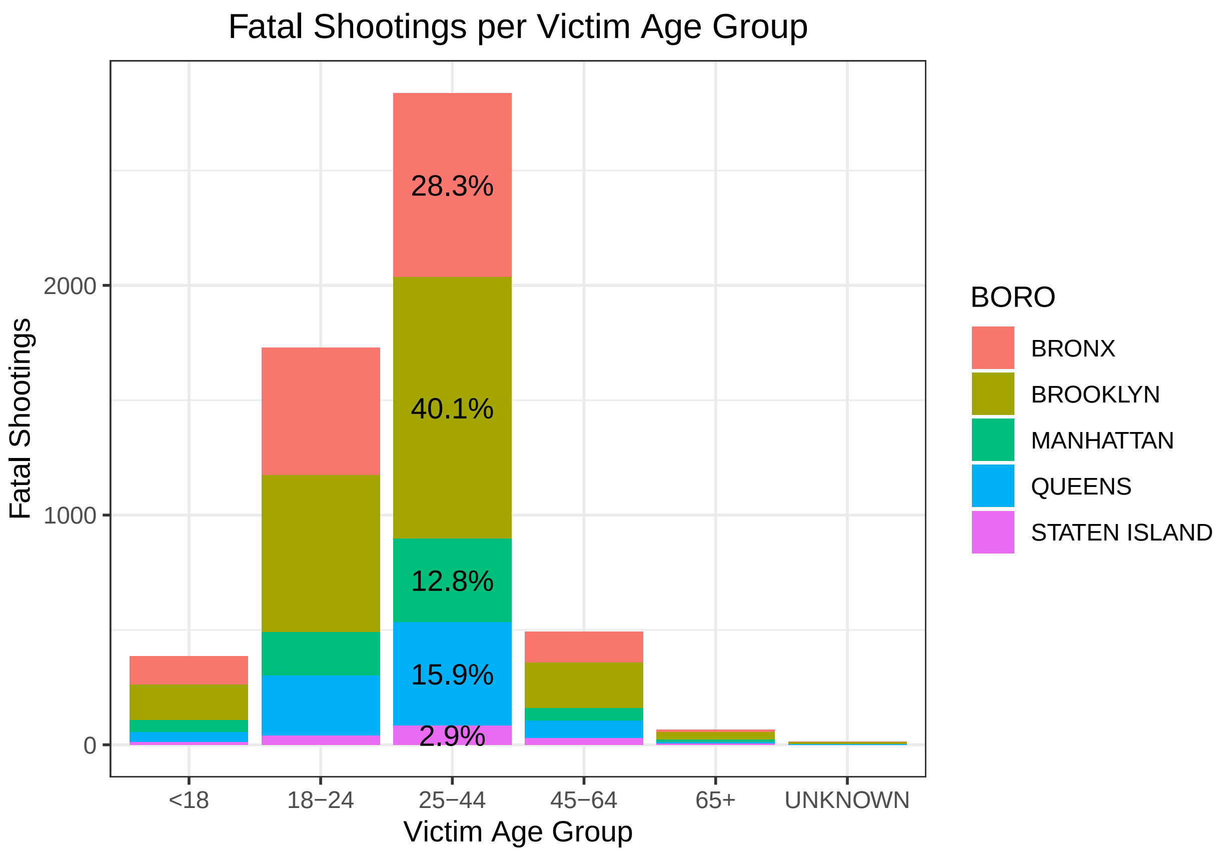
```r
nypd_2 %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = BORO, group = 1)) +  # group = 1 ensures a single line
  geom_line(aes(y = ..count.., color = BORO), stat = "count") +
  theme_bw() +
  labs(x = "Borough of New York City",
       y = "Non-Fatal Shootings",
       title = "Non-Fatal Shootings by Borough") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

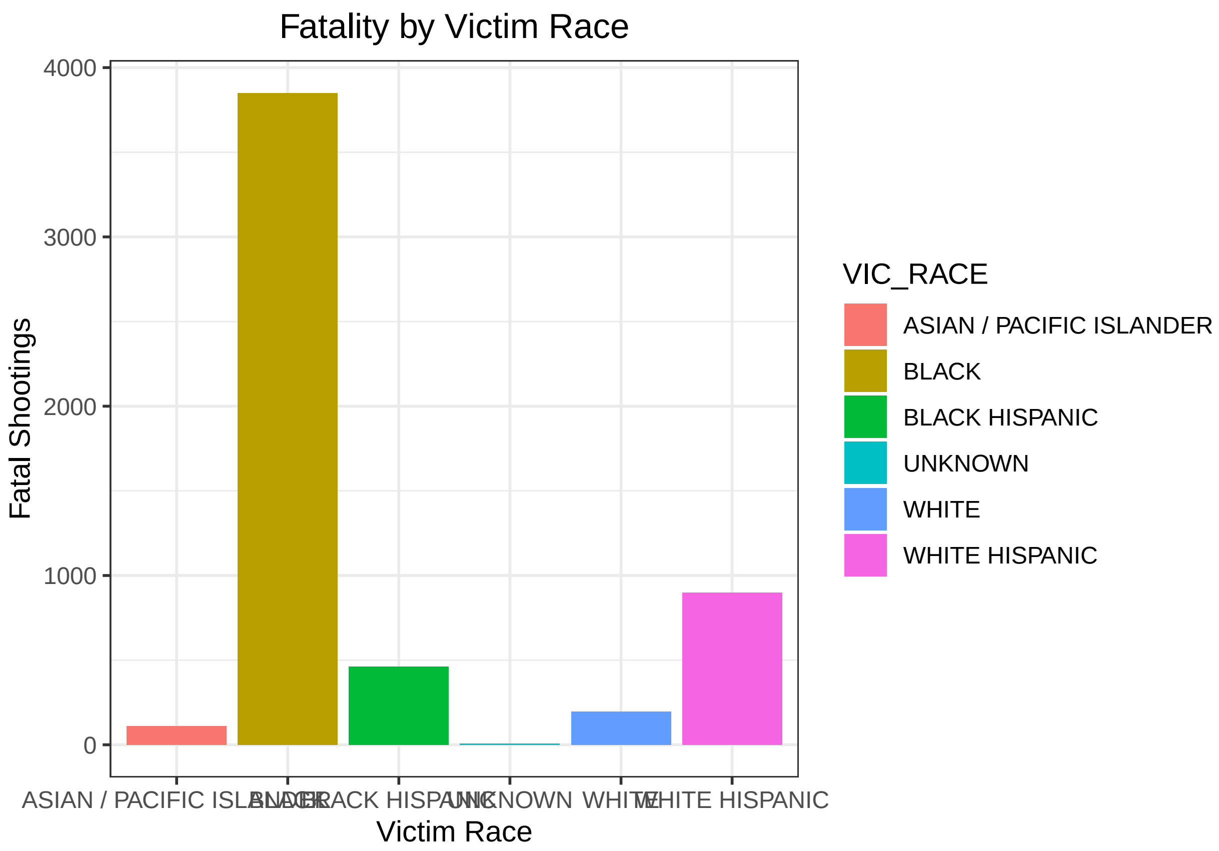
```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



```
nypd_2 %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  count(VIC_AGE_GROUP, BORO) %>%
  group_by(VIC_AGE_GROUP) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = n, fill = BORO)) +
  geom_bar(stat = "identity") +
  geom_text(data = . %>% filter(VIC_AGE_GROUP == "25-44"),
            aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5)) +
  theme_bw() +
  labs(x = "Victim Age Group",
       y = "Fatal Shootings",
       title = "Fatal Shootings per Victim Age Group") +
  theme(plot.title = element_text(hjust = 0.5))
```



```
nypd_2 %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  theme_bw() +
  labs(x = "Victim Race",
       y = "Fatal Shootings",
       title = "Fatality by Victim Race") +
  theme(plot.title = element_text(hjust = 0.5))
```



We can conclude that the majority of shooting victims are black males aged 25-44.

Step 4. Bias discussion

##I believe the data seems very inclusive and representative. However, the boroughs were mostly represented ##by a group of particular race and gender. overall the statistical analysis represents transparency around and ##important data.

Step 5. Model Discussion

##As you may see it in the pie chart Brooklyn represent more counts of incidents ##where more black are residing. In any given data set, we expect bias in the ##sampling of data, the demographic and reporting it. we will see unfortunate ##disparity and biases in the representation of the data.

Summary

##More shooting incidents occur in summer months. The number of these incidents was lower between 2013 and 2019 compared to the period between 2006 and 2012. However, there was a ##significant increase in shooting incidents in 2020. While unemployment is slightly associated with these incidents, it does not fully account for the variation. Other potential ##social and environmental factors related to the COVID-19 pandemic, such as school closures, reduced availability of social services, and the impacts of social isolation, should ##be explored.