NYPD Shooting Incident Data Report

Linpeng Sun

2021/10/7

Import Data

```
library(tidyverse)
```

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

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
```

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
file_name <- c("NYPD_Shooting_Incident_Data.csv")
```

```
NYPD_shooting <- read_csv(url_in)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
```

```
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Tidy and Transform Data

```
NYPD_shooting <- NYPD_shooting %>%
  select(-c(INCIDENT_KEY, JURISDICTION_CODE, LOCATION_DESC, STATISTICAL_MURDER_FLAG))
```

```
NYPD <- NYPD_shooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

summary(NYPD)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO              PRECINCT
##  Min.   :2006-01-01   Length:23568      Length:23568       Min.   :  1.00
##  1st Qu.:2008-12-30   Class1:hms        Class :character   1st Qu.: 44.00
##  Median :2012-02-26   Class2:difftime   Mode  :character   Median : 69.00
##  Mean   :2012-10-03   Mode  :numeric                       Mean   : 66.21
##  3rd Qu.:2016-02-28                                        3rd Qu.: 81.00
##  Max.   :2020-12-31                                        Max.   :123.00
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23568       Length:23568       Length:23568       Length:23568
##  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:23568       Length:23568       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
```

```
##  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
```

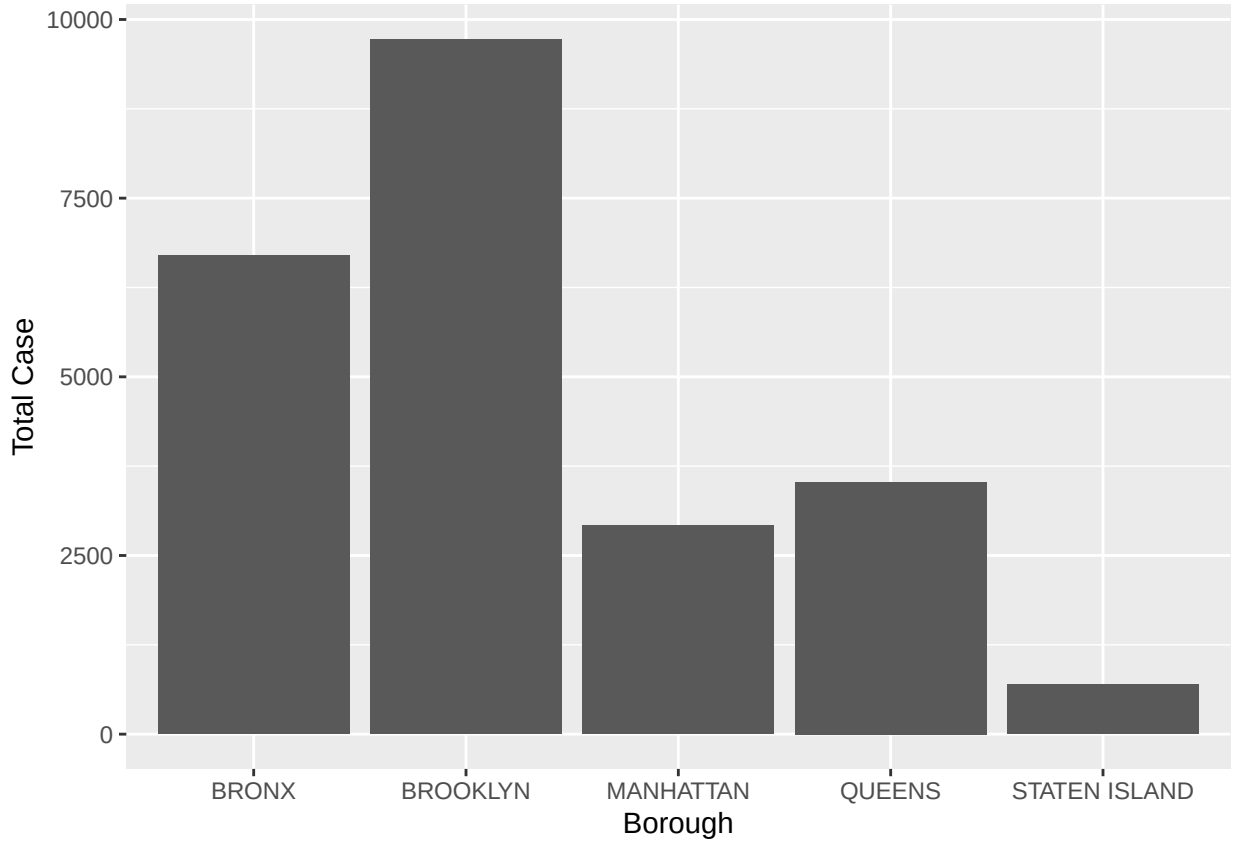
```
NYPD_Borough <- NYPD %>%
  group_by(BORO, OCCUR_DATE) %>%
  select(BORO, OCCUR_DATE, OCCUR_TIME, PERP_AGE_GROUP, PERP_SEX, VIC_AGE_GROUP, VIC_SEX) %>%
  ungroup()

summary(NYPD_Borough)
```

```
##      BORO             OCCUR_DATE          OCCUR_TIME       PERP_AGE_GROUP
##  Length:23568       Min.   :2006-01-01   Length:23568      Length:23568
##  Class :character   1st Qu.:2008-12-30   Class1:hms        Class :character
##  Mode  :character   Median :2012-02-26   Class2:difftime   Mode  :character
##                     Mean   :2012-10-03   Mode  :numeric
##                     3rd Qu.:2016-02-28
##                     Max.   :2020-12-31
##    PERP_SEX         VIC_AGE_GROUP        VIC_SEX
##  Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Visualizing Data

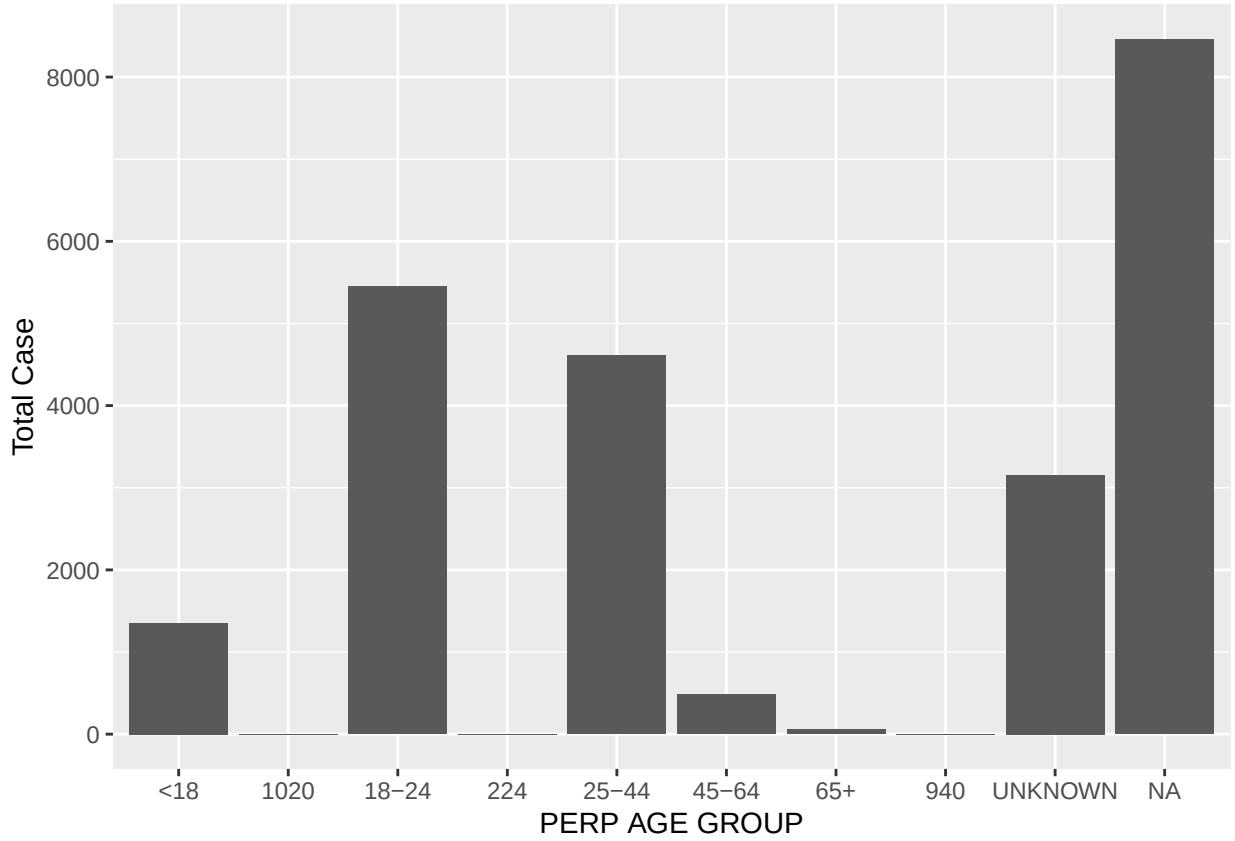
```
Borough <- NYPD_Borough %>% count(BORO)
PERP_AG <- NYPD %>% count(PERP_AGE_GROUP)
VIC_AGE <- NYPD %>% count(VIC_AGE_GROUP)
```

```
# Map of total case in each borough
Borough <- data.frame(Borough)
ggplot(Borough, aes(x=BORO, y=n)) +
  geom_bar(stat = "identity") +
  xlab("Borough") + ylab("Total Case")
```



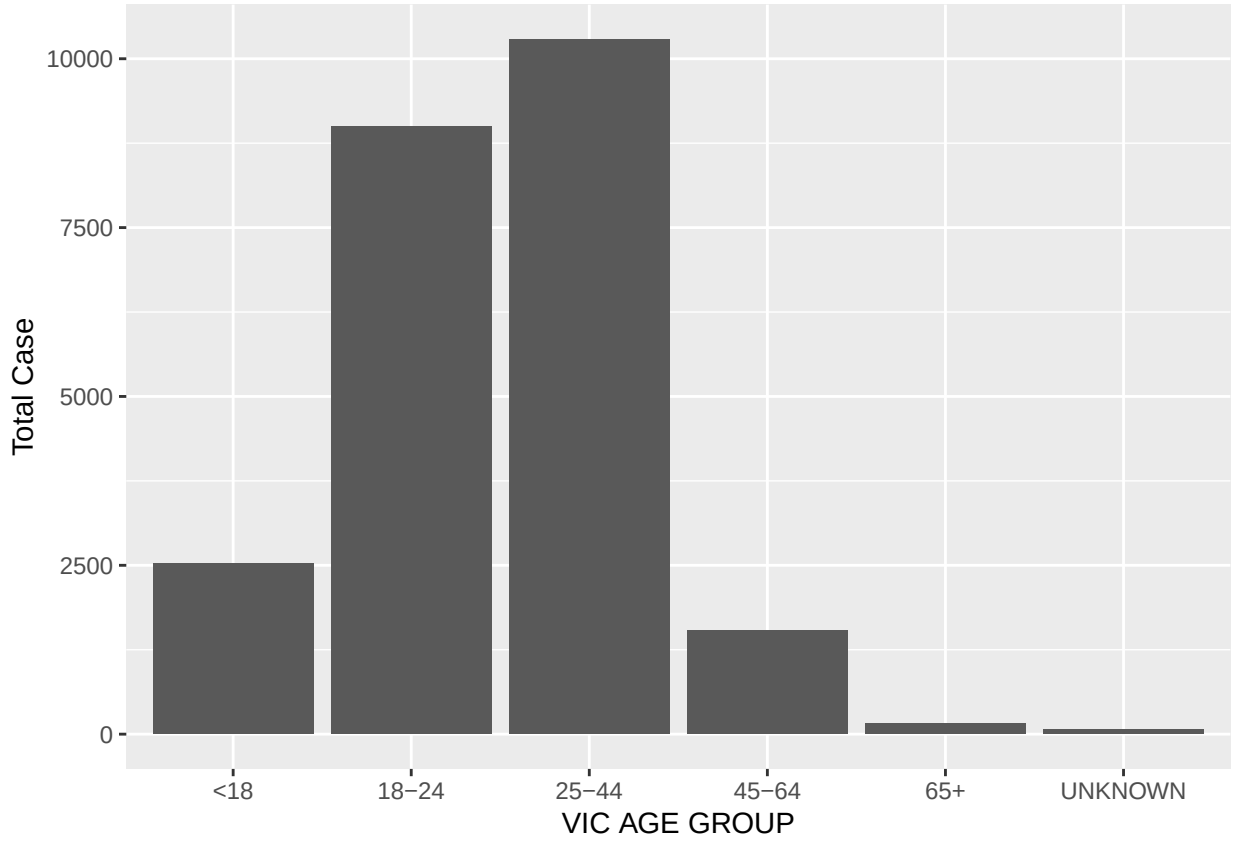
We could see that Brooklyn has the highest number of total cases, Bronx has the second-highest number of total cases.

```
# Map of the total case for perpetrator's age group
PERP_AG <- data.frame(PERP_AG)
ggplot(PERP_AG, aes(x=PERP_AGE_GROUP, y=n)) +
  geom_bar(stat = "identity") +
  xlab("PERP AGE GROUP") + ylab("Total Case")
```



From the bar plot we could conclude that other than Unknown and NA, 18-24 is the largest perpetrator's age group.

```
# Map of the total case for victim's age group
VIC_AGE <- data.frame(VIC_AGE)
ggplot(VIC_AGE, aes(x=VIC_AGE_GROUP, y=n)) +
  geom_bar(stat = "identity") +
  xlab("VIC AGE GROUP") + ylab("Total Case")
```



From the second bar plot we could conclude that other than Unknown, 25-44 is the largest victim's age group.

Analyzing Data

```
# Linear model of precinct and victim's age group
NYPD_shooting <- na.omit(NYPD_shooting)
summary(lm( PRECINCT ~ VIC_AGE_GROUP, data = NYPD_shooting))
```

```
##
## Call:
## lm(formula = PRECINCT ~ VIC_AGE_GROUP, data = NYPD_shooting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.496 -22.195   2.664  14.805  59.876
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            63.1236     0.6652  94.898  < 2e-16 ***
## VIC_AGE_GROUP18-24      3.0715     0.7622   4.030 5.61e-05 ***
## VIC_AGE_GROUP25-44      3.2120     0.7524   4.269 1.97e-05 ***
## VIC_AGE_GROUP45-64      3.3720     1.0992   3.068  0.00216 **
## VIC_AGE_GROUP65+        4.5858     2.6841   1.709  0.08756 .
## VIC_AGE_GROUPUNKNOWN    4.2975     3.7844   1.136  0.25615
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.13 on 15103 degrees of freedom
## Multiple R-squared:  0.001368,   Adjusted R-squared:  0.001037
## F-statistic: 4.137 on 5 and 15103 DF,  p-value: 0.0009337
```

From the p-value of each different age groups we could conclude victim's age group of 18-24, 25-44 are statistically significant to precinct.

Bias Sources

My last two plots did not move the UNKNOWN and NA value, which could cause possible visual confusion. And due to the development, political reasons, population composition of each borough, the total cases would be different.