NYPD Shooting Incident Analysis

2022-06-05

Step 1: Load data

Data description:

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

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

head(nypd_si_src)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1     24050482 08/27/2006   05:35:00    BRONX       52                 0
## 2     77673979 03/11/2011   12:03:00   QUEENS      106                 0
## 3    203350417 10/06/2019   01:09:00 BROOKLYN       77                 0
## 4     80584527 09/04/2011   03:35:00    BRONX       40                 0
## 5     90843766 05/27/2013   21:16:00   QUEENS      100                 0
## 6     92393427 09/01/2013   04:17:00 BROOKLYN       67                 0
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
## 1                                  true
## 2                                 false
## 3                                 false
## 4                                 false
## 5                                 false
## 6                                 false
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1         25-44       F BLACK HISPANIC    1017542   255918.9 40.86906 -73.87963
## 2           65+       M          WHITE    1027543   186095.0 40.67737 -73.84392
## 3         18-24       F          BLACK     995325   185155.0 40.67489 -73.96008
## 4           <18       M          BLACK    1007453   233952.0 40.80880 -73.91618
## 5         18-24       M          BLACK    1041267   157133.5 40.59780 -73.79469
## 6           <18       M          BLACK    1001694   170112.9 40.63359 -73.93715
##                                          Lon_Lat
## 1  POINT (-73.87963173099996 40.86905819000003)
## 2 POINT (-73.84392019199998 40.677366895000034)
## 3 POINT (-73.96007501899999 40.674885741000026)
## 4  POINT (-73.91618413199996 40.80879780500004)
## 5 POINT (-73.79468553799995 40.597796249000055)
## 6  POINT (-73.93715330699996 40.63358818100005)
```

```
summary(nypd_si_src)
```

```
##   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
##
```

Step 2: Tidy and transform data

1. Select the required fields
2. Apply transformations on fields:

- PERP_AGE_GROUP, VIC_AGE_GROUP, PERP_RACE, VIC_RACE: Assign “UNKNOWN” to empty
- PERP_SEX, VIC_SEX: Assign “UNKNOWN” to empty and “U”

3. Filter out records having erroneous value in PERP_AGE_GROUP
4. Convert data type of fields

```
nypd_si_stg <- nypd_si_src %>%
               select(OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG,
                      PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                      VIC_AGE_GROUP, VIC_SEX, VIC_RACE)

nypd_si_stg$PERP_AGE_GROUP[nypd_si_stg$PERP_AGE_GROUP == ""] <- "UNKNOWN"
nypd_si_stg$VIC_AGE_GROUP[nypd_si_stg$VIC_AGE_GROUP == ""] <- "UNKNOWN"
nypd_si_stg$PERP_SEX[nypd_si_stg$PERP_SEX == "" | nypd_si_stg$PERP_SEX == "U"] <- "UNKNOWN"
nypd_si_stg$VIC_SEX[nypd_si_stg$VIC_SEX == "" | nypd_si_stg$VIC_SEX == "U"] <- "UNKNOWN"
nypd_si_stg$PERP_RACE[nypd_si_stg$PERP_RACE == ""] <- "UNKNOWN"
nypd_si_stg$VIC_RACE[nypd_si_stg$VIC_RACE == ""] <- "UNKNOWN"

nypd_si_stg <- nypd_si_stg %>%
               filter(PERP_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN"))


nypd_si <- nypd_si_stg %>%
           mutate(OCCUR_DATE=mdy(OCCUR_DATE),
                  OCCUR_TIME=hms(OCCUR_TIME),
                  BORO=factor(BORO),
                  PERP_AGE_GROUP=factor(PERP_AGE_GROUP),
                  PERP_SEX=factor(PERP_SEX),
                  PERP_RACE=factor(PERP_RACE),
                  VIC_AGE_GROUP=factor(VIC_AGE_GROUP),
                  VIC_SEX=factor(VIC_SEX),
                  VIC_RACE=factor(VIC_RACE))

summary(nypd_si)
```

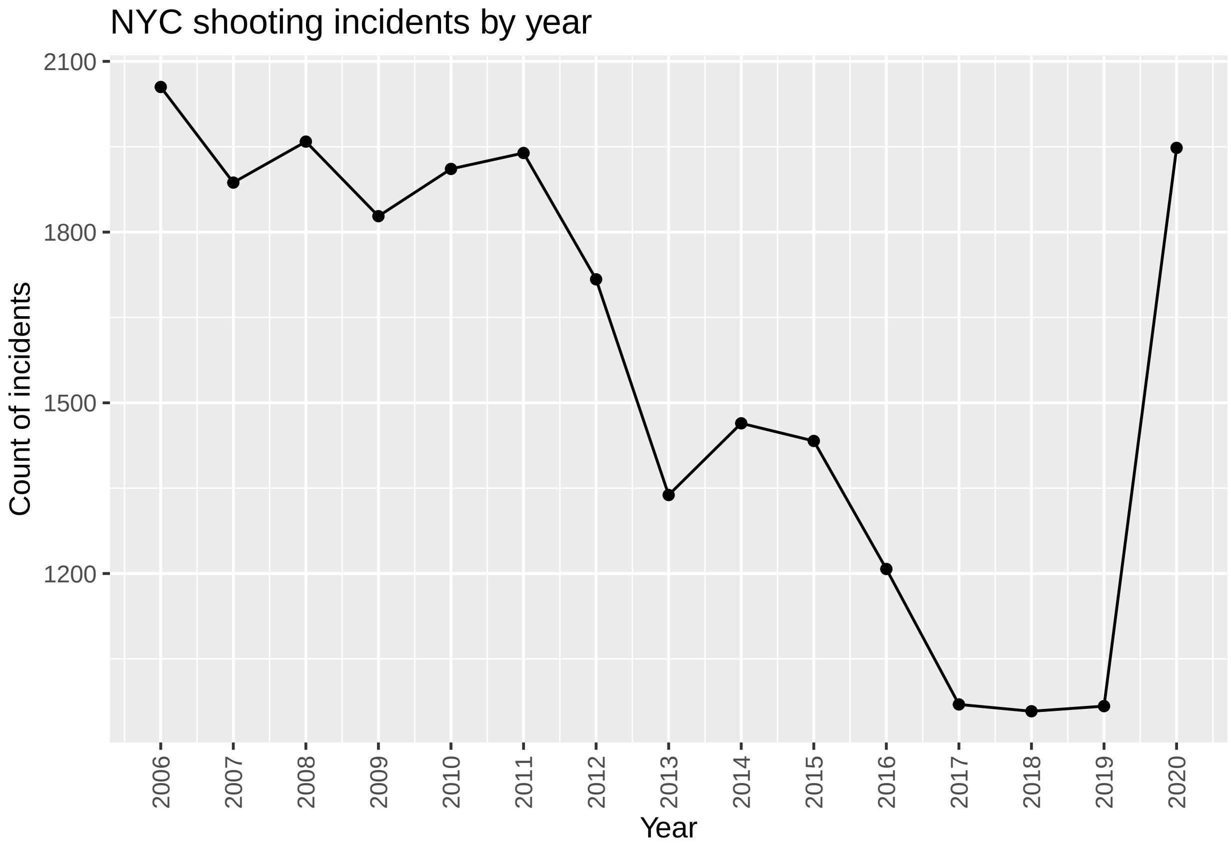
```
##    OCCUR_DATE           OCCUR_TIME                                   BORO
##  Min.   :2006-01-01   Min.   :0S                          BRONX        :6699
##  1st Qu.:2008-12-31   1st Qu.:3H 20M 0S                   BROOKLYN     :9733
##  Median :2012-02-27   Median :15H 0M 0S                   MANHATTAN    :2922
##  Mean   :2012-10-05   Mean   :12H 33M 9.14171825969242S   QUEENS       :3532
##  3rd Qu.:2016-03-03   3rd Qu.:20H 45M 0S                  STATEN ISLAND: 696
##  Max.   :2020-12-31   Max.   :23H 59M 0S
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Length:23582            <18    : 1368   F      :  335
##  Class :character        18-24  : 5508   M      :13487
##  Mode  :character        25-44  : 4714   UNKNOWN: 9760
##                          45-64  :  495
##                          65+    :   54
##                          UNKNOWN:11443
##
##                              PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2525   F      : 2204
##  ASIAN / PACIFIC ISLANDER      :  122   18-24  : 9002   M      :21367
##  BLACK                         :10024   25-44  :10301   UNKNOWN:   11
##  BLACK HISPANIC                : 1096   45-64  : 1541
##  UNKNOWN                       :10097   65+    :  154
##  WHITE                         :  255   UNKNOWN:   59  
##  WHITE HISPANIC                : 1986                  
##                            VIC_RACE    
##  AMERICAN INDIAN/ALASKAN NATIVE:    9  
##  ASIAN / PACIFIC ISLANDER      :  327  
##  BLACK                         :16868  
##  BLACK HISPANIC                : 2245  
##  UNKNOWN                       :   65  
##  WHITE                         :  620  
##  WHITE HISPANIC                : 3448  
```

Step 3: Analyse data

1. Overall trend of shooting incidents in NYC

Shootings in NYC declined steadily from 2006 to 2019. But the number spiked in 2020.

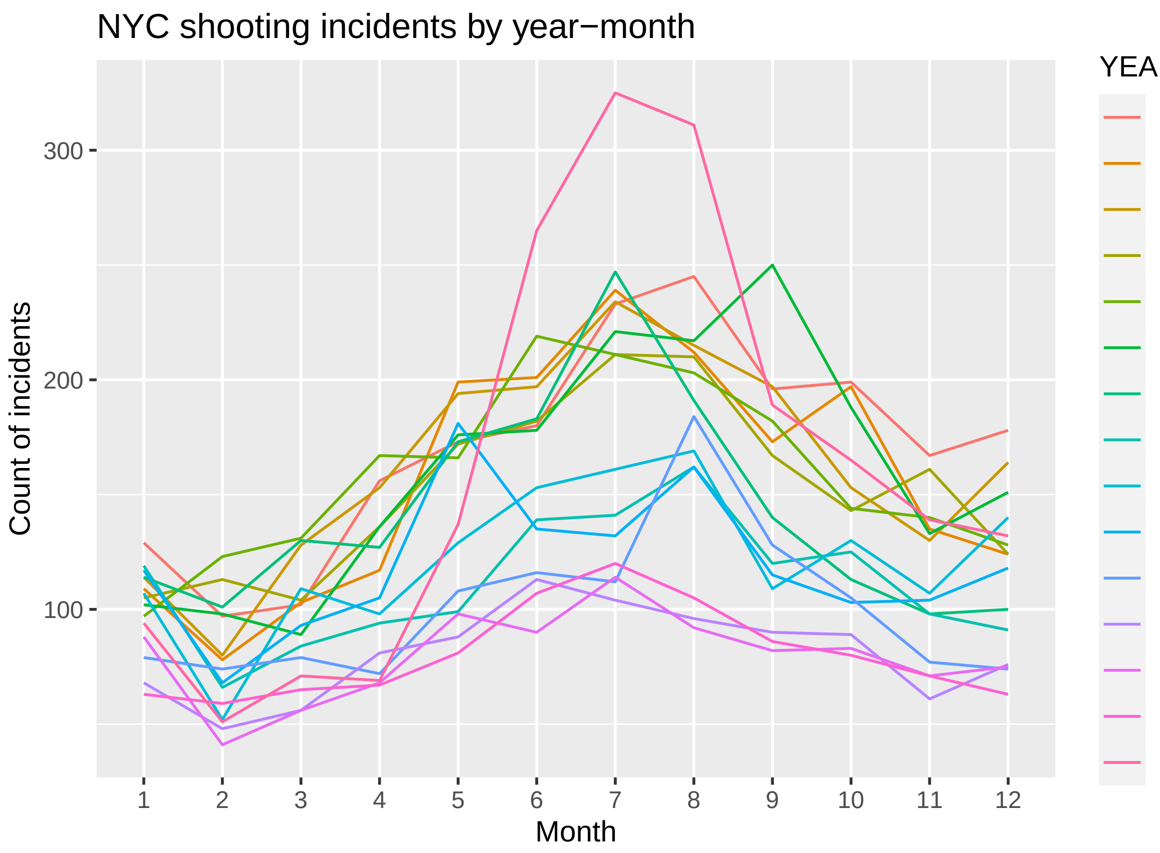
```
nypd_si_y <- nypd_si %>%
             mutate(OCCUR_YEAR=year(OCCUR_DATE)) %>%
             group_by(OCCUR_YEAR) %>%
             count()

nypd_si_y %>% ggplot(aes(x = OCCUR_YEAR, y = n)) +
              geom_line() +
              geom_point() +
              labs(title = "NYC shooting incidents by year", y = "Count of incidents") +
              theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
              scale_x_continuous("Year", breaks=nypd_si_y$OCCUR_YEAR)
```



2. Seasonality of shooting incidents

In general, there were more shooting cases in summer months.

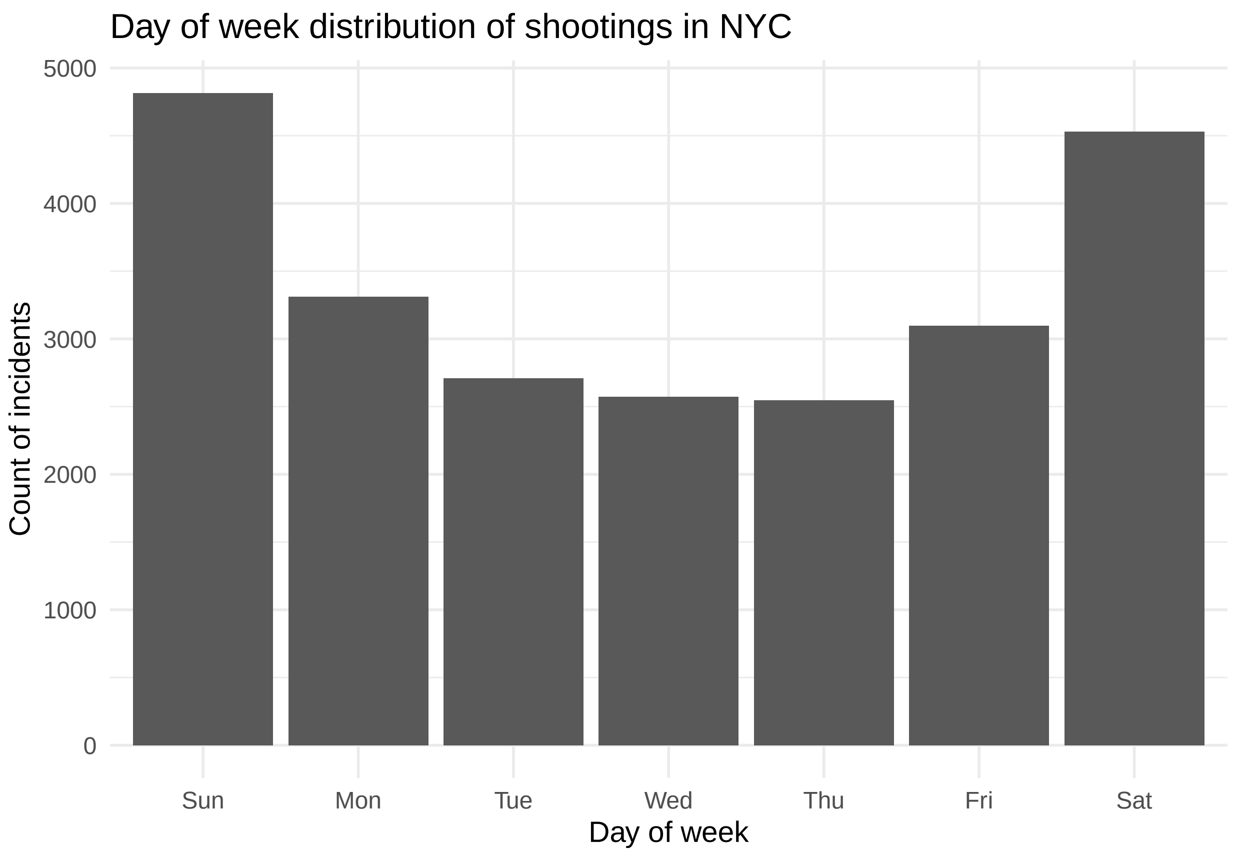
```
nypd_si_ym <- nypd_si %>%
              mutate(YEAR=factor(year(OCCUR_DATE)),
                     MONTH=factor(month(OCCUR_DATE))) %>%
              group_by(YEAR, MONTH) %>%
              count()

nypd_si_ym %>% ggplot(aes(x = MONTH, y = n, group = YEAR, colour = YEAR)) +
               geom_line() +
               labs(title = "NYC shooting incidents by year-month",
                    x = "Month", y = "Count of incidents")
```



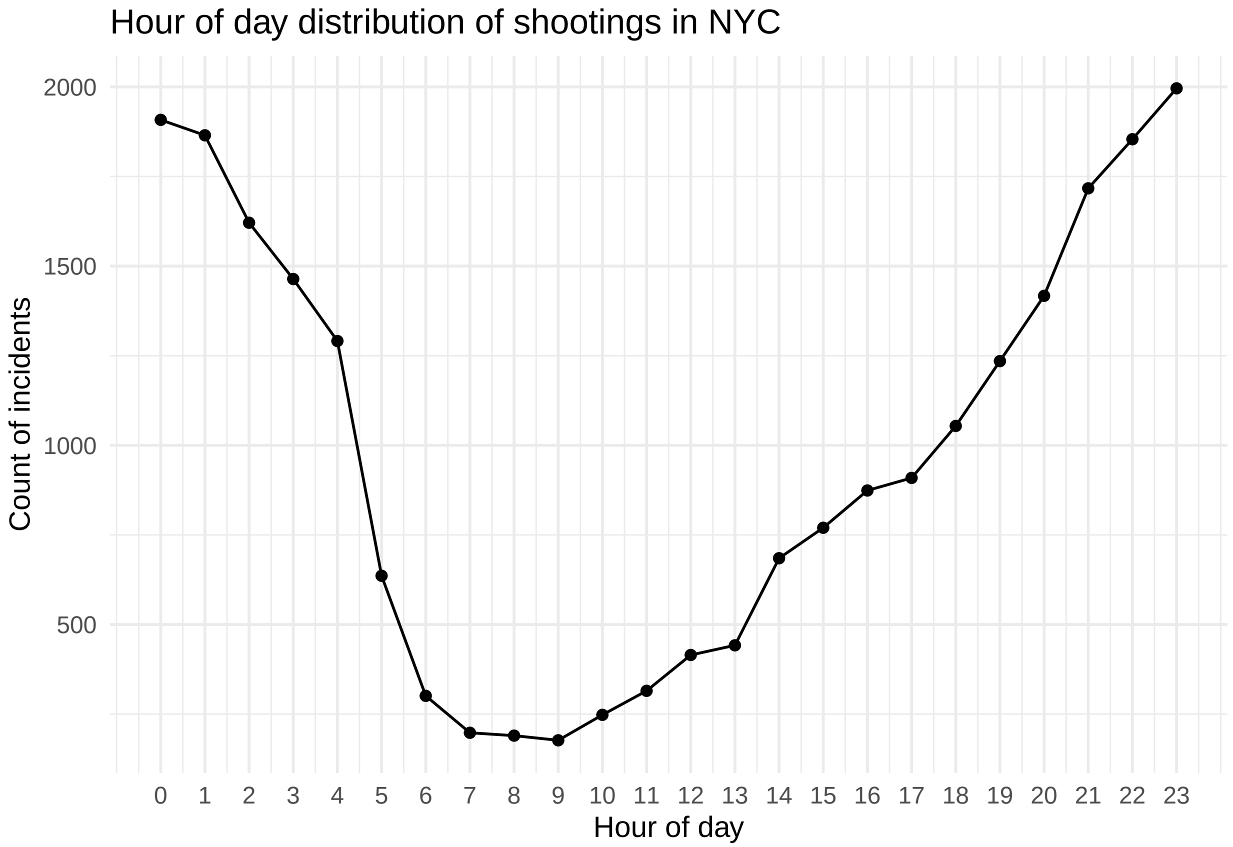
3. Day and time distribution of shootings

There were more shooting incidents in weekends and night time.

```
nypd_si_wd <- nypd_si %>%
              mutate(DAY_OF_WEEK=factor(wday(OCCUR_DATE, label = TRUE, locale="English_United States"))
              group_by(DAY_OF_WEEK) %>%
              count()

nypd_si_wd %>% ggplot(aes(x = DAY_OF_WEEK, y = n)) +
               geom_col() +
               labs(title = "Day of week distribution of shootings in NYC",
                    x = "Day of week", y = "Count of incidents") +
               theme_minimal()
```



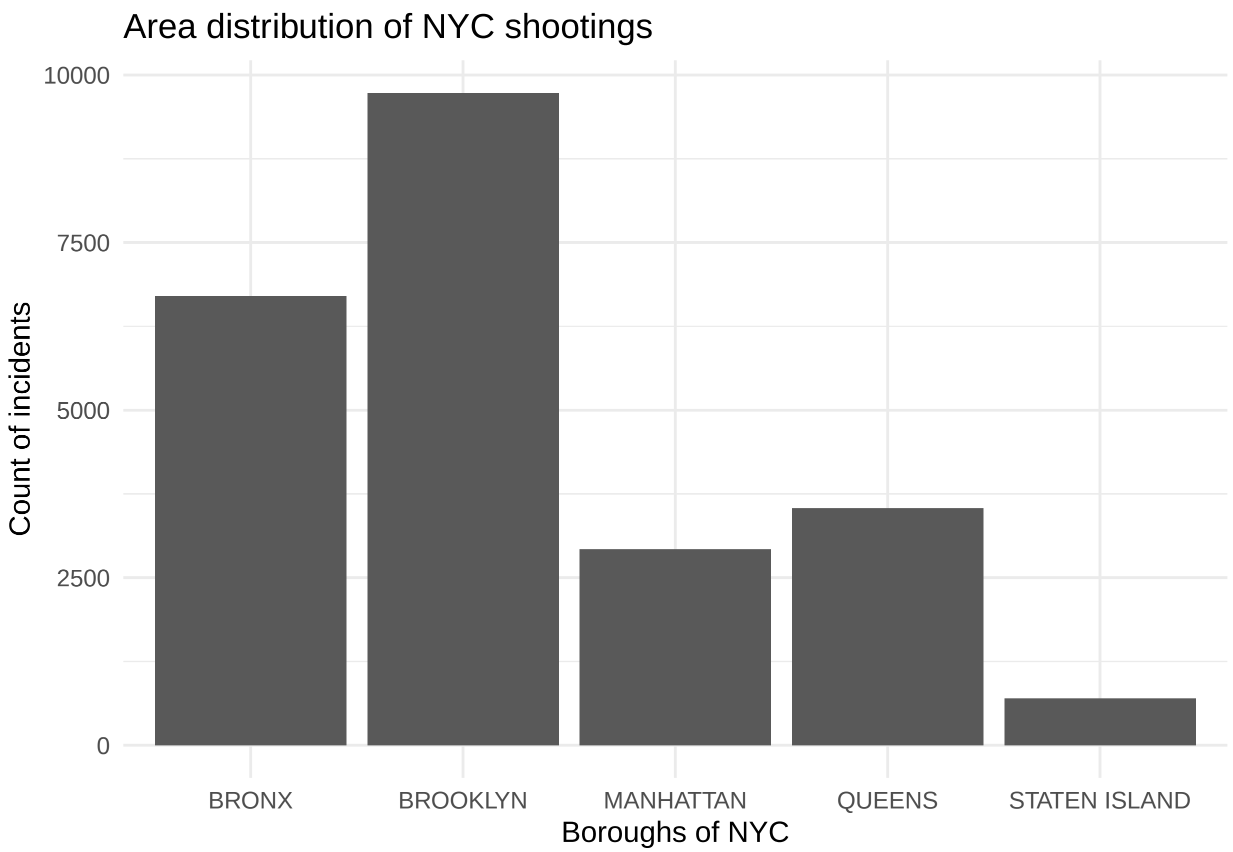
```
nypd_si_hour <- nypd_si %>%
                mutate(OCCUR_HOUR=hour(OCCUR_TIME)) %>%
                group_by(OCCUR_HOUR) %>%
                count()

nypd_si_hour %>% ggplot(aes(x = OCCUR_HOUR, y = n)) +
                 geom_line() +
                 geom_point() +
                 labs(title = "Hour of day distribution of shootings in NYC",
                      y = "Count of incidents") +
                 scale_x_continuous("Hour of day", breaks=nypd_si_hour$OCCUR_HOUR) +
                 theme_minimal()
```



4. Area distribution

Most of the shooting incidents happened in Brooklyn, followed by Bronx, Queens, and Manhattan.

```
nypd_si %>% ggplot(aes(x = BORO)) +
            geom_bar() +
            labs(title = "Area distribution of NYC shootings",
                 x = "Boroughs of NYC", y = "Count of incidents") +
            theme_minimal()
```



5. The profile of perpetrators and victims

- There were significantly more shooting incidents with male than those of female.
- Large portion of incidents involved age group 18-24 and 25-44.
- In terms of race, black accounted for the highest portion for both perpetrators and victims.

```
table(data.frame(nypd_si$PERP_SEX, nypd_si$VIC_SEX))
```

```
##                 nypd_si.VIC_SEX
## nypd_si.PERP_SEX     F     M UNKNOWN
##          F          50   284       1
##          M        1435 12046       6
##          UNKNOWN   719  9037       4
```

```
table(data.frame(nypd_si$PERP_AGE_GROUP, nypd_si$VIC_AGE_GROUP))
```

```
##                     nypd_si.VIC_AGE_GROUP
## nypd_si.PERP_AGE_GROUP  <18 18-24 25-44 45-64  65+ UNKNOWN
##                <18      417   554   324    63    8       2
##                18-24    713  2470  1988   291   34      12
##                25-44    234  1326  2691   392   38      33
##                45-64     18    60   265   136   11       5
##                65+        0     1    22    21   10       0
##                UNKNOWN 1143  4591  5011   638   53       7
```

```r
table(data.frame(nypd_si$PERP_RACE, nypd_si$VIC_RACE))
```

```
##                                  nypd_si.VIC_RACE
## nypd_si.PERP_RACE                AMERICAN INDIAN/ALASKAN NATIVE
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       4
##   BLACK HISPANIC                                              0
##   UNKNOWN                                                     5
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##                                  nypd_si.VIC_RACE
## nypd_si.PERP_RACE                ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             39    39             12
##   BLACK                                               126  7974            687
##   BLACK HISPANIC                                       17   448            279
##   UNKNOWN                                             102  7728            897
##   WHITE                                                11    29             18
##   WHITE HISPANIC                                       32   648            352
##                                  nypd_si.VIC_RACE
## nypd_si.PERP_RACE                UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0    11             21
##   BLACK                               24   165           1044
##   BLACK HISPANIC                       5    33            314
##   UNKNOWN                             24   176           1165
##   WHITE                                1   151             45
##   WHITE HISPANIC                      11    84            859
```

Step 4: Use linear model to analyse the relationship between shooting cases and shooting cases which counted as murder

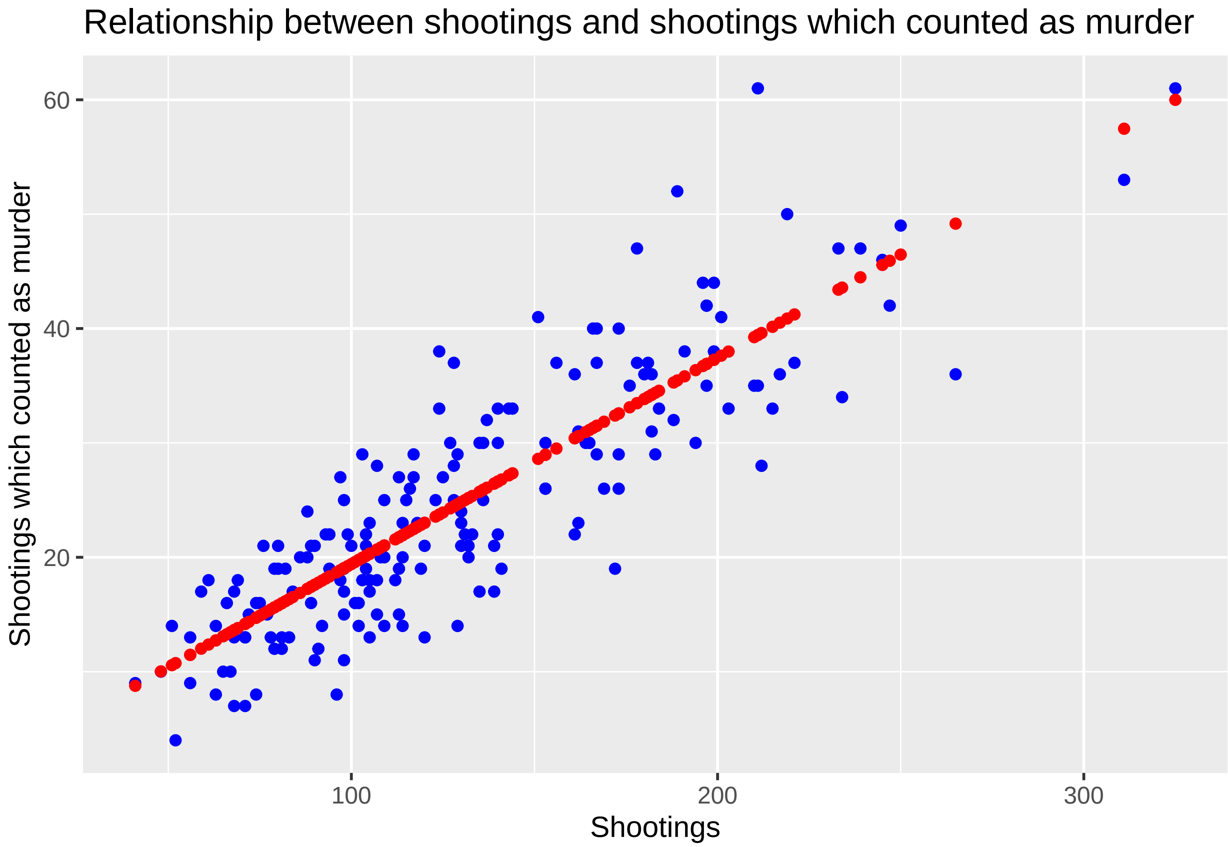
The model shows that the number of shooting cases and shooting cases which counted as murder is positively correlated.

```r
nypd_si_murder <- nypd_si %>%
                  mutate(YEAR=factor(year(OCCUR_DATE)),
                         MONTH=factor(month(OCCUR_DATE))) %>%
                  group_by(YEAR, MONTH) %>%
                  summarise(cases=n(),
                            murder_cases=sum(STATISTICAL_MURDER_FLAG == "true")) %>%
                  select(YEAR, MONTH, cases, murder_cases) %>%
                  ungroup() %>%
                  arrange(YEAR, MONTH)

mod = lm(murder_cases ~ cases, data = nypd_si_murder)

summary(mod)
```

```
##
```

```
## Call:
## lm(formula = murder_cases ~ cases, data = nypd_si_murder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3940  -3.8558  -0.0353   3.5520  21.5707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.366770   1.135379   1.204     0.23
## cases       0.180391   0.008038  22.444   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.696 on 178 degrees of freedom
## Multiple R-squared:  0.7389, Adjusted R-squared:  0.7374
## F-statistic: 503.7 on 1 and 178 DF,  p-value: < 2.2e-16
```

```
nypd_si_murder_with_predict <- nypd_si_murder %>%
                               mutate(murder_cases_pred = predict(mod))

nypd_si_murder_with_predict %>% ggplot() +
                                geom_point(aes(x = cases, y = murder_cases), color = "blue") +
                                geom_point(aes(x = cases, y = murder_cases_pred), color = "red") +
                                labs(title = "Relationship between shootings and shootings which counte
                                     x = "Shootings",
                                     y = "Shootings which counted as murder")
```



Step 5: Identify bias

- The data source failed to capture data for some perpetrators and victims. The UNKNOWN values may impact the result of the analyses.
- Above are just basic, simple analyses The characteristics of shooting cases can change over time. More thoughtful analyses are needed before drawing any conclusions.

Step 6: Conclusion

- Shootings in NYC declined steadily from 2006 to 2019. But the number spiked in 2020.
- Most of the shooting incidents happened in Brooklyn, followed by Bronx, Queens, and Manhattan.
- In general, there were more shooting cases in summer months.
- There were more shooting incidents in weekends and night time.
- There were significantly more shooting incidents with male than those of female.
- Large portion of incidents involved age group 18-24 and 25-44.
- In terms of race, black accounted for the highest portion for both perpetrators and victims.