# NYPD Shooting Incident

SK

08/07/22

## Importing Data

Data was downloaded from : https://catalog.data.gov/dataset Imports the shooting project data set in a reproducible manner.

## Tidy and Transform your data

```
summary(df_nypd)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  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:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
```

```
##                                       Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  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
##
```

```
colnames(df_nypd)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "JURISDICTION_CODE"
##  [7] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
##  [9] "PERP_AGE_GROUP"          "PERP_SEX"
## [11] "PERP_RACE"               "VIC_AGE_GROUP"
## [13] "VIC_SEX"                 "VIC_RACE"
## [15] "X_COORD_CD"              "Y_COORD_CD"
## [17] "Latitude"                "Longitude"
## [19] "Lon_Lat"
```

```
dim(df_nypd)
```

```
## [1] 25596    19
```

```
str(df_nypd)
```

```
## 'data.frame':    25596 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  24050482 77673979 226950018 237710987 224701998 225295736 231190175
##  $ OCCUR_DATE             : chr  "08/27/2006" "03/11/2011" "04/14/2021" "12/10/2021" ...
##  $ OCCUR_TIME             : chr  "05:35:00" "12:03:00" "21:08:00" "19:30:00" ...
##  $ BORO                   : chr  "BRONX" "QUEENS" "BRONX" "BRONX" ...
##  $ PRECINCT               : int  52 106 42 52 34 75 32 26 41 67 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 2 2 0 ...
##  $ LOCATION_DESC          : chr  "" "" "COMMERCIAL BLDG" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "65+" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "F" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1017542 1027543 1009489 1017440 1005426 ...
##  $ Y_COORD_CD             : num  255919 186095 243050 256046 254690 ...
##  $ Latitude               : num  40.9 40.7 40.8 40.9 40.9 ...
##  $ Longitude              : num  -73.9 -73.8 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73.84392019
```

**Cleaning up data sets**

## Removing zero variance using nearZerVar function.

```
non_zer_var <- nearZeroVar(df_nypd)
nypd_clean <- df_nypd[, -non_zer_var]
dim(nypd_clean)
```

```
## [1] 25596     0
```

## Getting rid of any columns not needed for future analysis and graph report.

```
nypd_data <- select(df_nypd, -c(PRECINCT,JURISDICTION_CODE))
sapply(nypd_data, function(x) sum(is.na(x)))
```

```
##            INCIDENT_KEY             OCCUR_DATE             OCCUR_TIME
##                       0                      0                      0
##                    BORO          LOCATION_DESC STATISTICAL_MURDER_FLAG
##                       0                      0                      0
##          PERP_AGE_GROUP               PERP_SEX              PERP_RACE
##                       0                      0                      0
##           VIC_AGE_GROUP               VIC_SEX               VIC_RACE
##                       0                      0                      0
##              X_COORD_CD             Y_COORD_CD               Latitude
##                       0                      0                      0
##               Longitude               Lon_Lat
##                       0                      0
```

```
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:25596       Length:25596            Length:25596
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:25596       Length:25596       Length:25596       Length:25596
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:25596       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000011   1st Qu.:182782   1st Qu.:40.67
##  Mode  :character   Median :1007715   Median :194038   Median :40.70
##                     Mean   :1009455   Mean   :207894   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239429   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:25596
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
```

## Removing all NA values in nypd data set.

```
na_val <- sapply(nypd_data,function(x) mean(is.na(x))) > 0.95
nypd_data <- nypd_data[,na_val == FALSE]
dim(nypd_data)
```

```
## [1] 25596    17
```

## Date format is changed for future use as graph presentation.

```
nypd_data$y_month <- strftime(as.Date(df_nypd$OCCUR_DATE,"%m/%d/%Y"), "%Y-%m")
```

## Visualizations and Analysis.

## Analysing the nypd data set using correlation of date and time

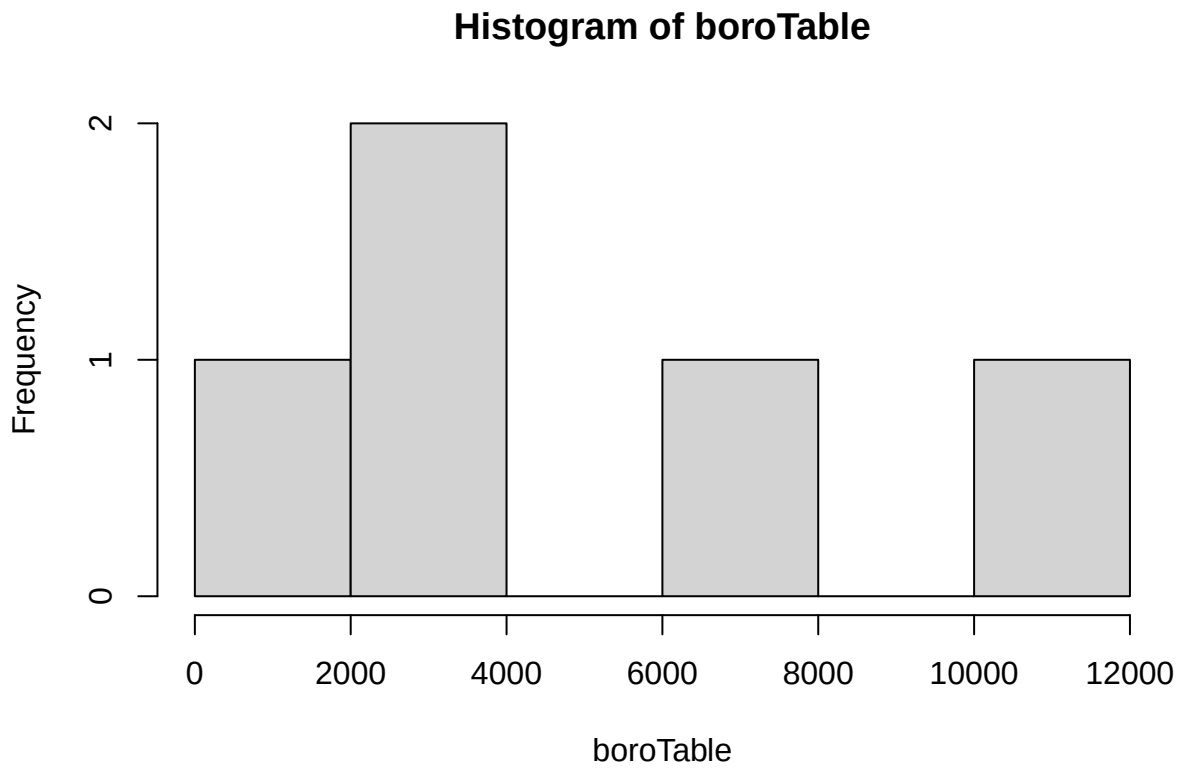
```
d_date <- as.numeric(as.factor(nypd_data$OCCUR_DATE))
d_time <- as.numeric(as.factor(nypd_data$OCCUR_TIME))
corr.date.time <- cor.test(x = d_date,
                           y = d_time)
corr.date.time
```

```
##
##  Pearson's product-moment correlation
```

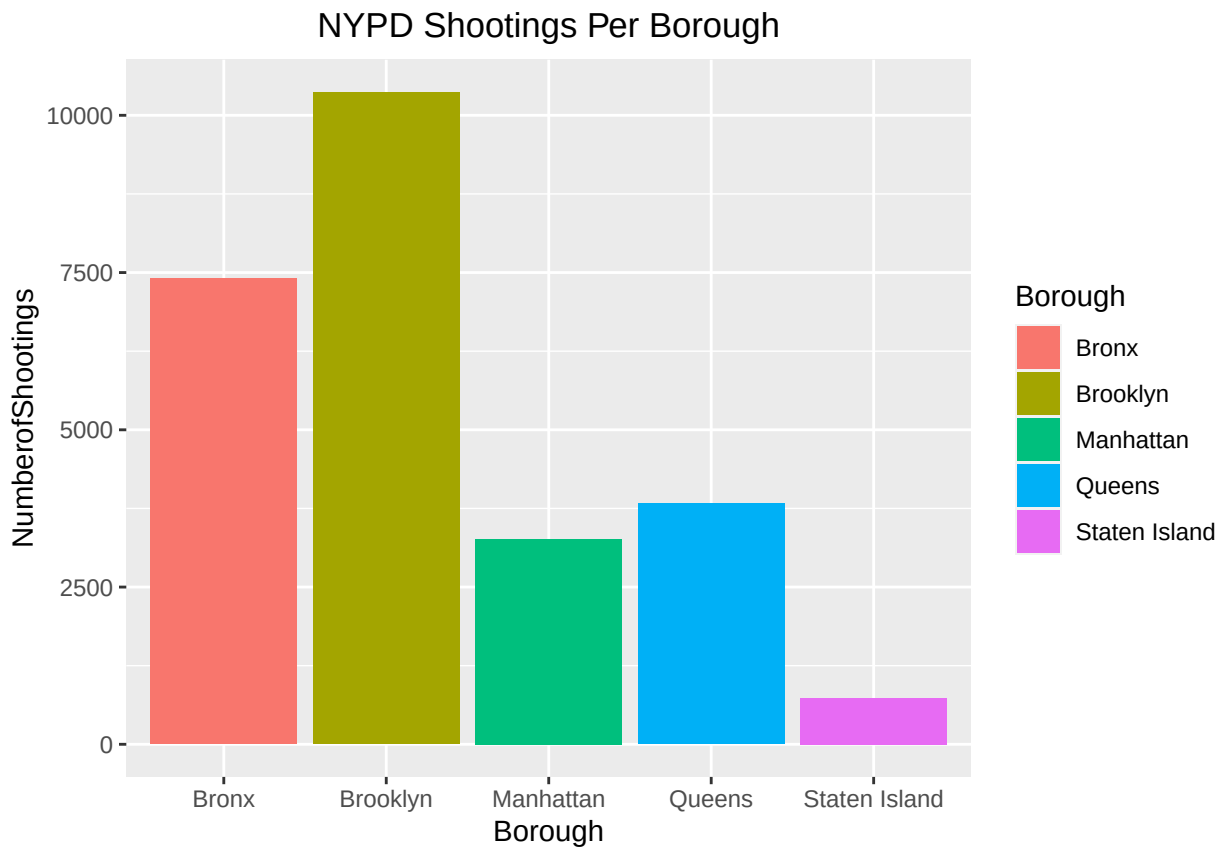
```
##
## data:  d_date and d_time
## t = -0.37329, df = 25594, p-value = 0.7089
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.014583730  0.009917811
## sample estimates:
##          cor
## -0.002333309
```

## Visualization 1: NYPD shootings by borough plot.

```
boroTable <- table(nypd_data$BORO)
hist(boroTable)
```




```
boros <- data.frame(
                    Borough = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"),
                    NumberofShootings = c(7402, 10365, 3265, 3828, 736)
)
theme_update(plot.title = element_text(hjust = 0.5))
ggplot(boros, aes(x = Borough, y = NumberofShootings,fill = Borough)) +
    geom_bar(stat = "identity") +
    ggtitle("NYPD Shootings Per Borough")
```




More than 10000 shootings occurred in Brooklyn city of NewYork. Least occurred in Staten Island as per dataset given.

## Visualization 2: Perp age vs victims age plot with the year of 2021.

```
perp_age<- nypd_data$PERP_AGE_GROUP
vic_age <- nypd_data$VIC_AGE_GROUP
filter(nypd_data, y_month == 2021) %>%
  ggplot( aes(perp_age, vic_age, color = nypd_data$LOCATION_DESC), xlab = perp_age, ylab = vic_age) +
  geom_point()
```

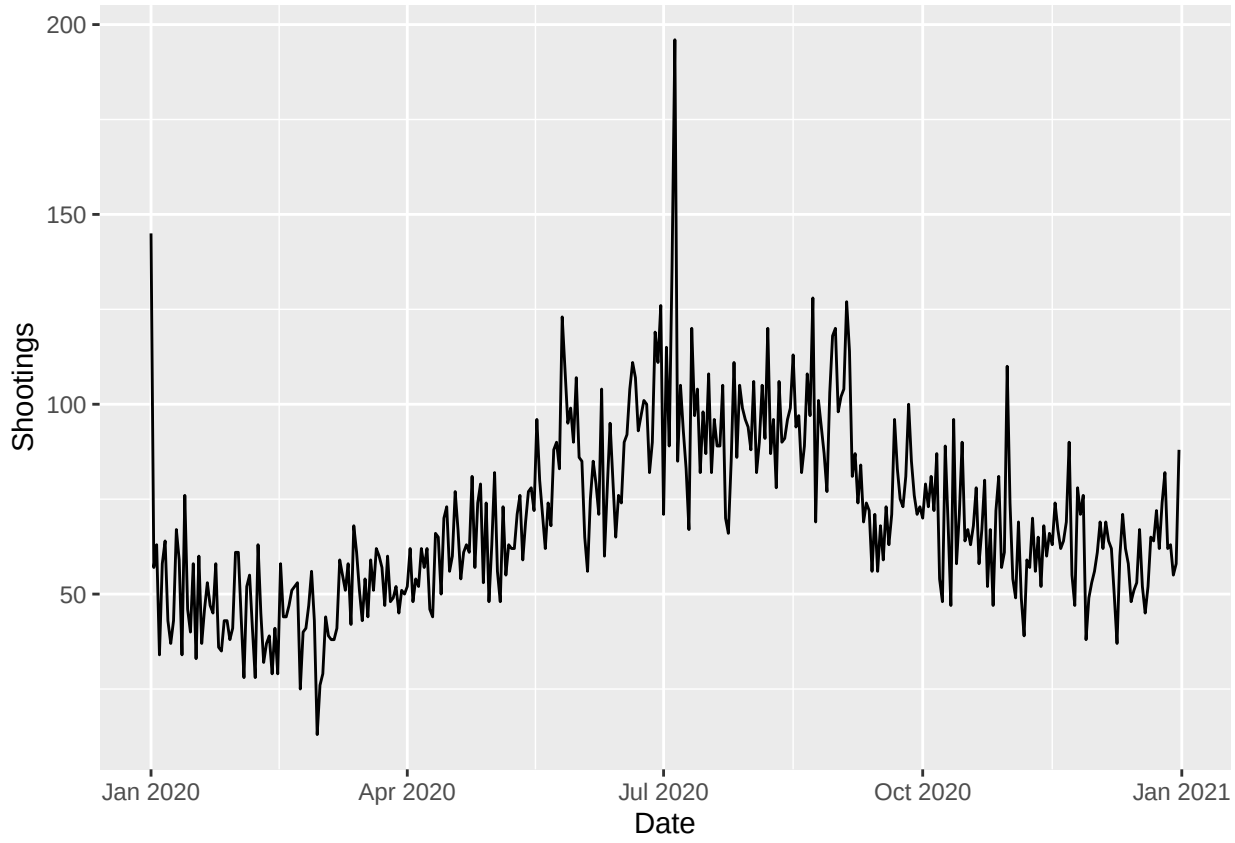

Most of the perp targets victims by their younger/senior citizen in Multi dwell apartments at the age of 25-44. From 2019 Covid season, most perp race was black hispanic. Spike happened in mid july 2020 from brooklyn city.

## Linear Regression model to built

```
nypd_data <- nypd_data %>% select(OCCUR_DATE,BORO,LOCATION_DESC,PERP_AGE_GROUP,PERP_SEX,PERP_RACE,VIC_A
head(nypd_data)
```

```
##   OCCUR_DATE      BORO   LOCATION_DESC PERP_AGE_GROUP PERP_SEX      PERP_RACE
## 1 2020-08-27     BRONX
## 2 2020-03-11    QUEENS
## 3 2020-04-14     BRONX COMMERCIAL BLDG
## 4 2020-12-10     BRONX
## 5 2020-02-22 MANHATTAN
## 6 2020-03-07  BROOKLYN                          25-44        M BLACK HISPANIC
##   VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1         25-44       F BLACK HISPANIC
## 2           65+       M          WHITE
## 3         18-24       M          BLACK
## 4         25-44       M          BLACK
## 5         25-44       M BLACK HISPANIC
## 6         25-44       M WHITE HISPANIC
```

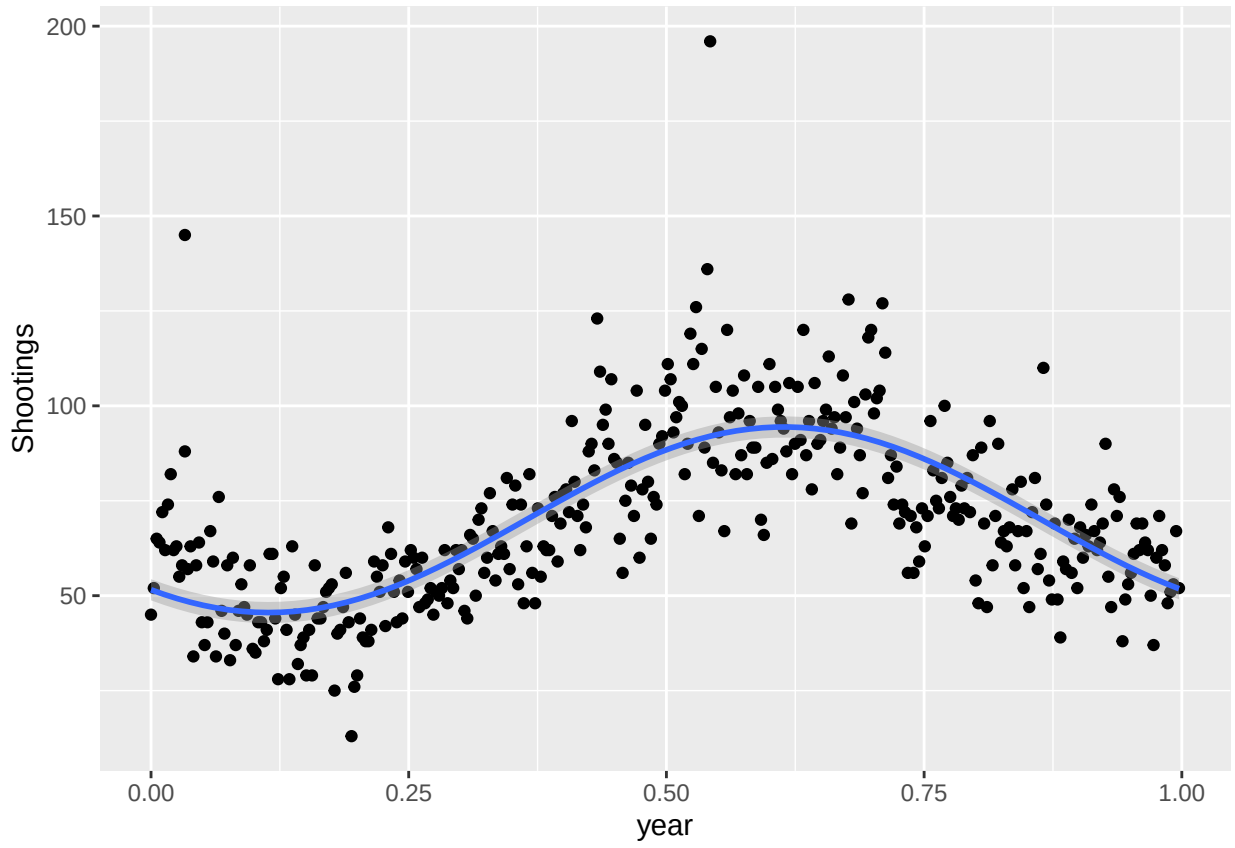
```
shootings_by_date <- nypd_data %>% group_by(OCCUR_DATE) %>% summarise(COUNT = n())
ggplot( data = shootings_by_date, aes( OCCUR_DATE, COUNT )) +
geom_line() +
xlab("Date") + ylab("Shootings")
```




```
filter(shootings_by_date, COUNT > 150)
```

```
## # A tibble: 1 x 2
##   OCCUR_DATE COUNT
##   <date>     <int>
## 1 2020-07-05   196
```

```
ggplot( data = shootings_by_date, aes( x=(julian(OCCUR_DATE)%%365)/365, y=COUNT )) +
  geom_point() +
  geom_smooth(method="lm", formula= y ~ sin(2*pi*x)+cos(2*pi*x) ) +
  xlab("year") + ylab("Shootings")
```




## Analysis with the report for NYPD shootings:

We can now summarize the conclusions of all the NYPD analysis and visualizations performed. The number of shooting incidents in NYPD had been decreasing steadily since 2006 until recently mid of 2021, when a significant spike was observed. Most of the shootings occurred locations were multi dwell apartments and ATM,BANK during Covid season. Suspect age targets younger and senior citizen victims age group.

The spike observed in mid 2020, coincides with the COVID-19 lockdown situation, which probably suggests that the increase in number of shootings was the result of the higher number of unemployment caused by the economic impact of the lockdown. The boroughs of Bronx and Brooklyn are the areas with the highest number of incidents per million inhabitants. These two areas are also the boroughs that experience higher rates of poverty, which suggests a possible correlation between the two data points.

Most of the perp targets victims by their younger/senior citizen in Multi dwell apartments at the age of 25-44. From 2019 Covid season, most perp race was black hispanic. Spike happened in mid july 2020 from brooklyn city.

## Bias Identification:

The data given was written by NYPD department. They might have accidentally lost some documents or many shootings long back at time goes unreported.