NYPD-shooting

2024-09-20

Introduction

This is an analysis for every shooting incident that occurred in NYC going back to 2006. This analysis focuses on if crime occur more often on certain hours and areas.

1.Load and tidy dataset

let import the dataset from https://catalog.data.gov/ but since this link is only for downloading the dataset directly to local machine. Instead I have uploaded the dataset to my Github and used the link below.

```
rawData <- read.csv("https://raw.githubusercontent.com/xuannguyen1206/NYPD-shooting-incident/main/NYPD_
```

I notice that column OCCUR_DATE and OCCUR_TIME is chr typed which is not correct so I will change them to approriate type

```
library(chron)
rawData$OCCUR_DATE <- as.Date(rawData$OCCUR_DATE, format = "%m/%d/%Y")
rawData$OCCUR_TIME <- times(rawData$OCCUR_TIME)
```

2.Transform data

I add 3 more columns for easier data manipulation.

-`count` is used instead of `INCIDENT_KEY`(too arbitrary).

-`decimal_hour` is decimal number representation of hours in`OCCUR_TIME`

-`hour_incident`categorizes each incidient into a 1-hour interval

```
library(dplyr)
rawData <- rawData %>%  mutate(count=row_number() ,decimal_hour =as.numeric(rawData$OCCUR_TIME) * 24)
rawData$hour_incident <-cut(rawData$decimal_hour,
               breaks = seq(0, 24, by = 1),
               labels = seq(0, 23),
               include.lowest = TRUE)
```

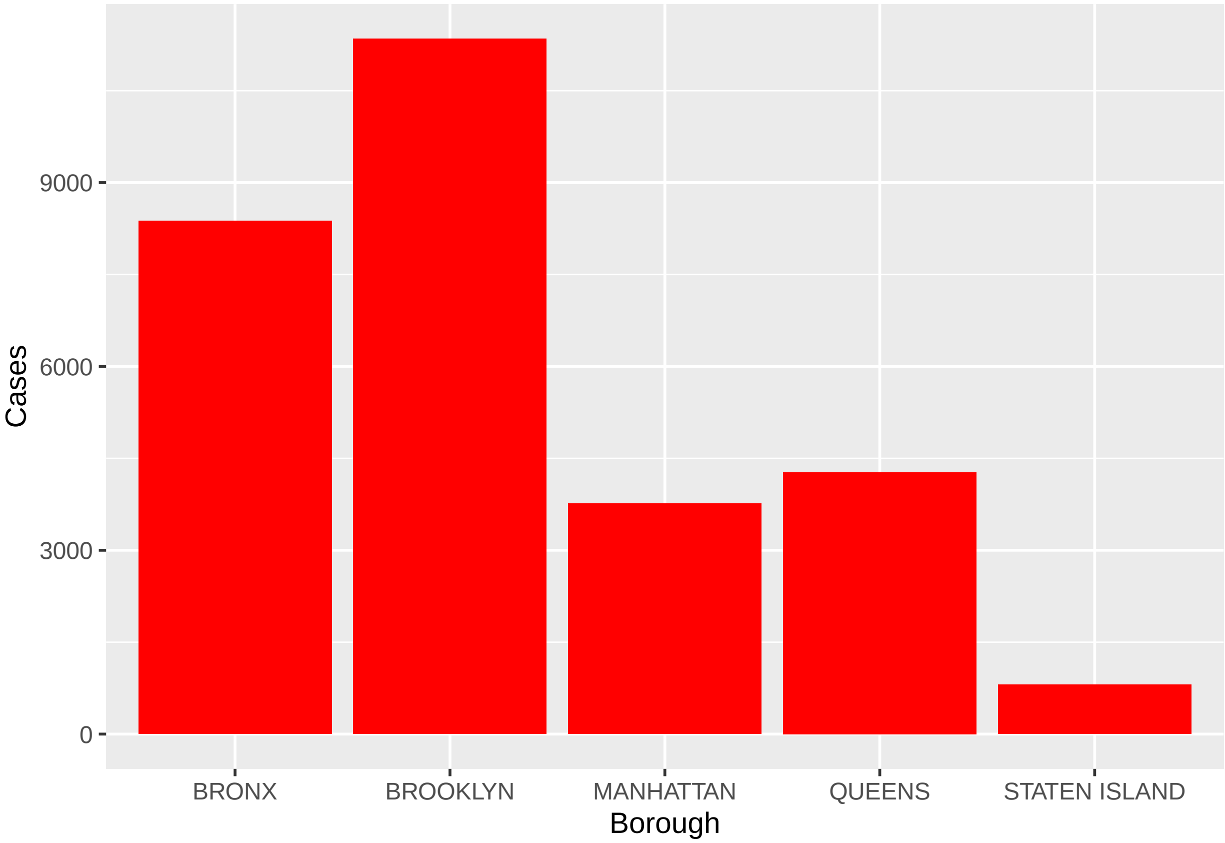
the final data format will be

```
## 'data.frame':    28562 obs. of  24 variables:
##  $ INCIDENT_KEY           : int  244608249 247542571 84967535 202853370 27078636 230311078 229224142
##  $ OCCUR_DATE             : Date, format: "2022-05-05" "2022-07-04" ...
##  $ OCCUR_TIME             : 'times' num  00:10:00 22:20:00 19:35:00 21:00:00 21:00:00 ...
##   ..- attr(*, "format")= chr "h:m:s"
##  $ BORO                   : chr  "MANHATTAN" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "INSIDE" "OUTSIDE" "" "" ...
##  $ PRECINCT               : int  14 48 103 42 83 23 113 77 48 49 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 2 0 0 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  "COMMERCIAL" "STREET" "" "" ...
##  $ LOCATION_DESC          : chr  "VIDEO STORE" "(null)" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "true" "false" "false" ...
##  $ PERP_AGE_GROUP         : chr  "25-44" "(null)" "" "25-44" ...
##  $ PERP_SEX               : chr  "M" "(null)" "" "M" ...
##  $ PERP_RACE              : chr  "BLACK" "(null)" "" "UNKNOWN" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "18-24" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  986050 1016802 1048632 1014493 1009149 ...
##  $ Y_COORD_CD             : num  214231 250581 198262 242565 190105 ...
##  $ Latitude               : num  40.8 40.9 40.7 40.8 40.7 ...
##  $ Longitude              : num  -74 -73.9 -73.8 -73.9 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.9935 40.754692)" "POINT (-73.88233 40.854402)" "POINT (
##  $ count                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ decimal_hour           : num  0.167 22.333 19.583 21 21 ...
##  $ hour_incident          : Factor w/ 24 levels "0","1","2","3",..: 1 23 20 21 21 24 20 2 19 24 ...
## NULL
```

3.DATA OVERVIEW

First we look at total cases reported by NYC Boroughs:

```r
library(ggplot2)
ggplot(rawData, aes(x = BORO)) + geom_bar(fill = "red") + labs(x="Borough",y = "Cases")
```

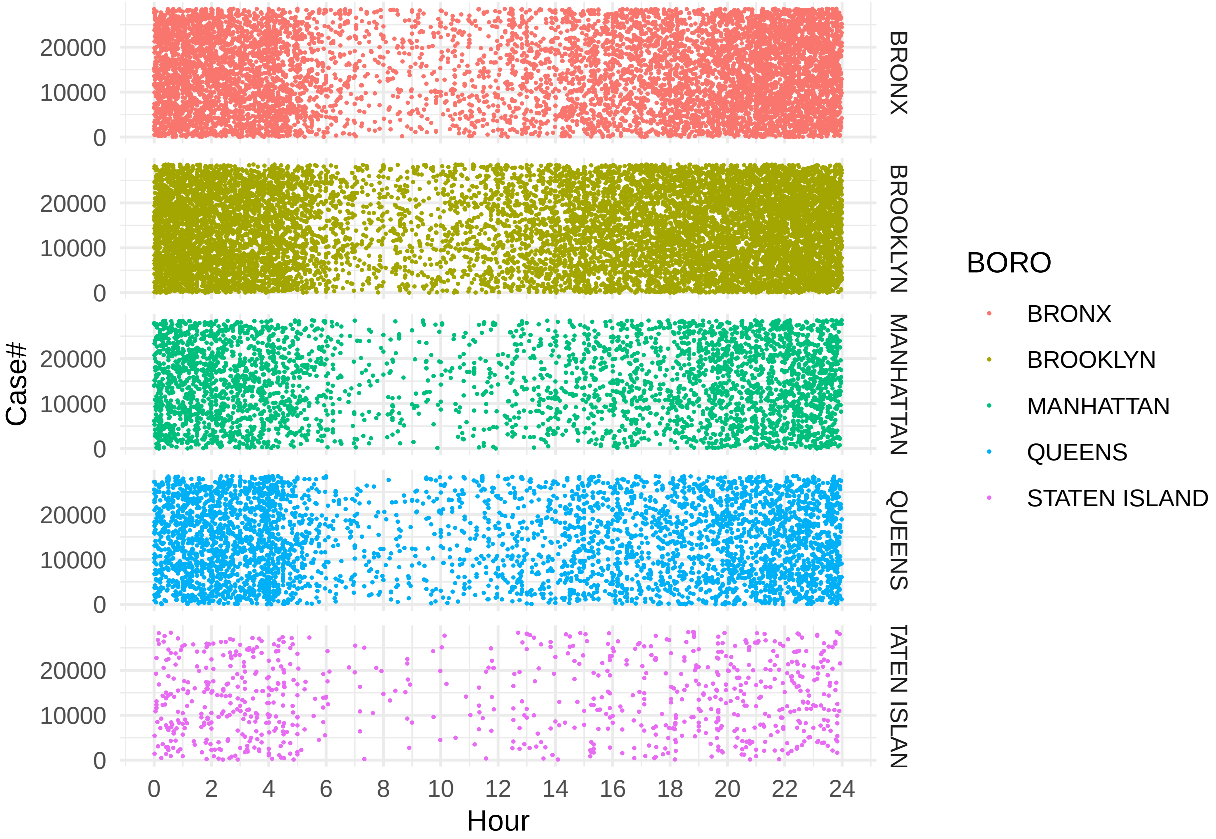


We can clealy see the majority of reports concentrated in 2 boroughs which are Bronx and Brooklyn with

Bronx nearly approach 9000 cases and Brooklyn even exceeds that. .Whilst Staten Island has the lowest number of reports with no more than 1500 cases reported.

Next, I plot the crime cases number (using `count` column) against the time of the day at which it occurs (using `decimal_hours`) to see the if there are any relationsionship between time and crime.

```
crimeVtime <- ggplot(rawData, aes(x=decimal_hour,y = rawData$count,color=BORO )) +
     scale_x_continuous(limits = c(0, 24), breaks = seq(0, 24, by = 2)) +
    geom_point    (size = 0.1) +
    labs(y = "Case#",x= "Hour") + facet_grid(BORO~.)+
    theme_minimal()
print(crimeVtime)
```



We can vaguely guess the relations between time and crime. The trend seem to be clearer in less reported boroughs. Shooting-incidents tend to occur around night hours and decreases as the morning come.

We will look deeper into this trend in the following section

4.Relationship between shootings and time

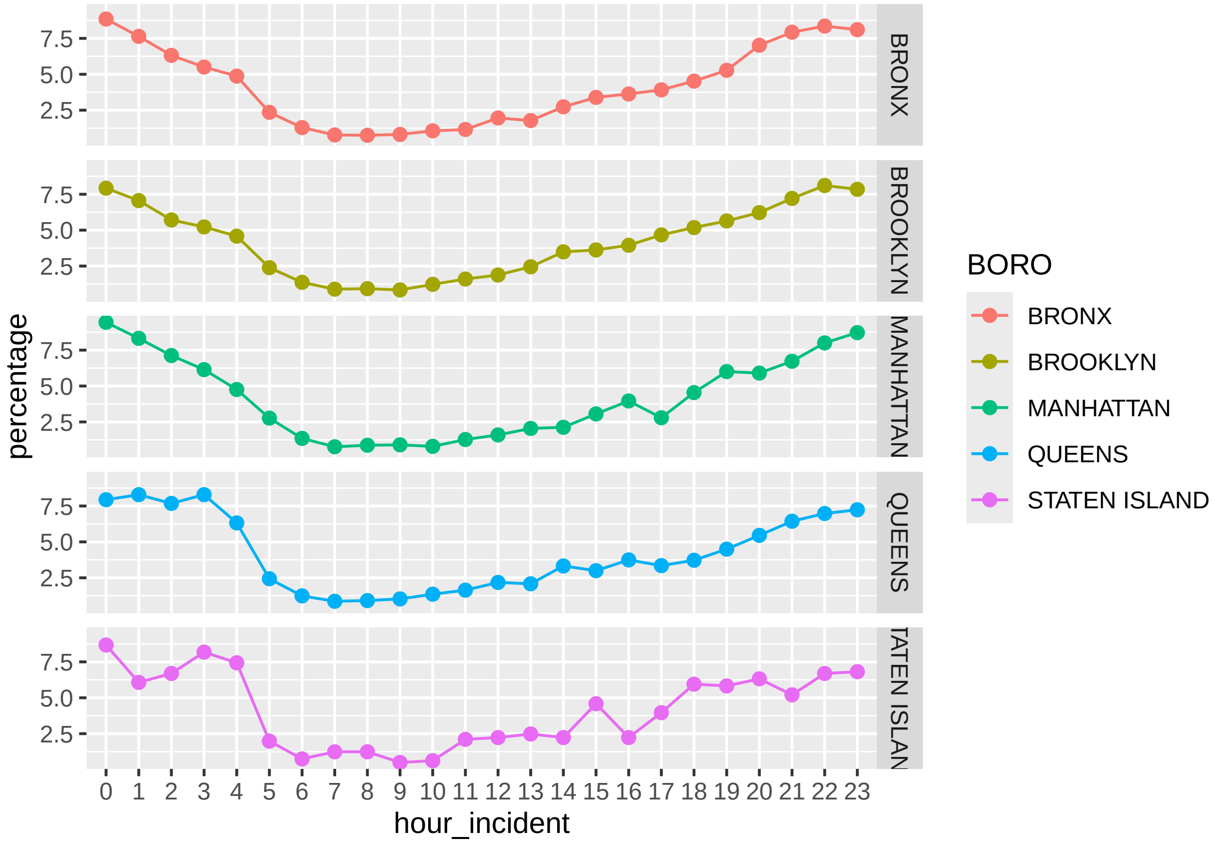
I want to look into more detail of which time does the shooting occurs more. To do that, I will plot the percentage of crime occur at each hour of the day. This is done by taking the sum of shooting incidients at each hour divided by the sum of shooting. Each borough will have their own hourly shooting incidents rate.

```
library(purrr)
boroughs <- unique(rawData$BORO)
calculate_borough_incident_percentages <- function(data, borough) {
  borough_data <- data %>%
    filter(BORO == borough)
  borough_sum_count <- nrow(borough_data)

  borough_hourly_percentage <- borough_data %>%
    group_by(hour_incident) %>%
    summarise(count = n(), .groups = 'drop') %>%
    mutate(percentage = count / borough_sum_count * 100,
           BORO = borough)

  return(borough_hourly_percentage)
}
all_borough_data <- map_df(boroughs, ~calculate_borough_incident_percentages(rawData, .))
```

```
hourly_incident_rate_plot <- ggplot(all_borough_data, aes(x = hour_incident, y = percentage,color=BORO,
print(hourly_incident_rate_plot)
```



We can see similar trend irrespective of areas. They all peaked around the time between 22-23 PM and gradually decrease and hit lowest at 7AM. The shooting rise again rather soon at around 13PM.

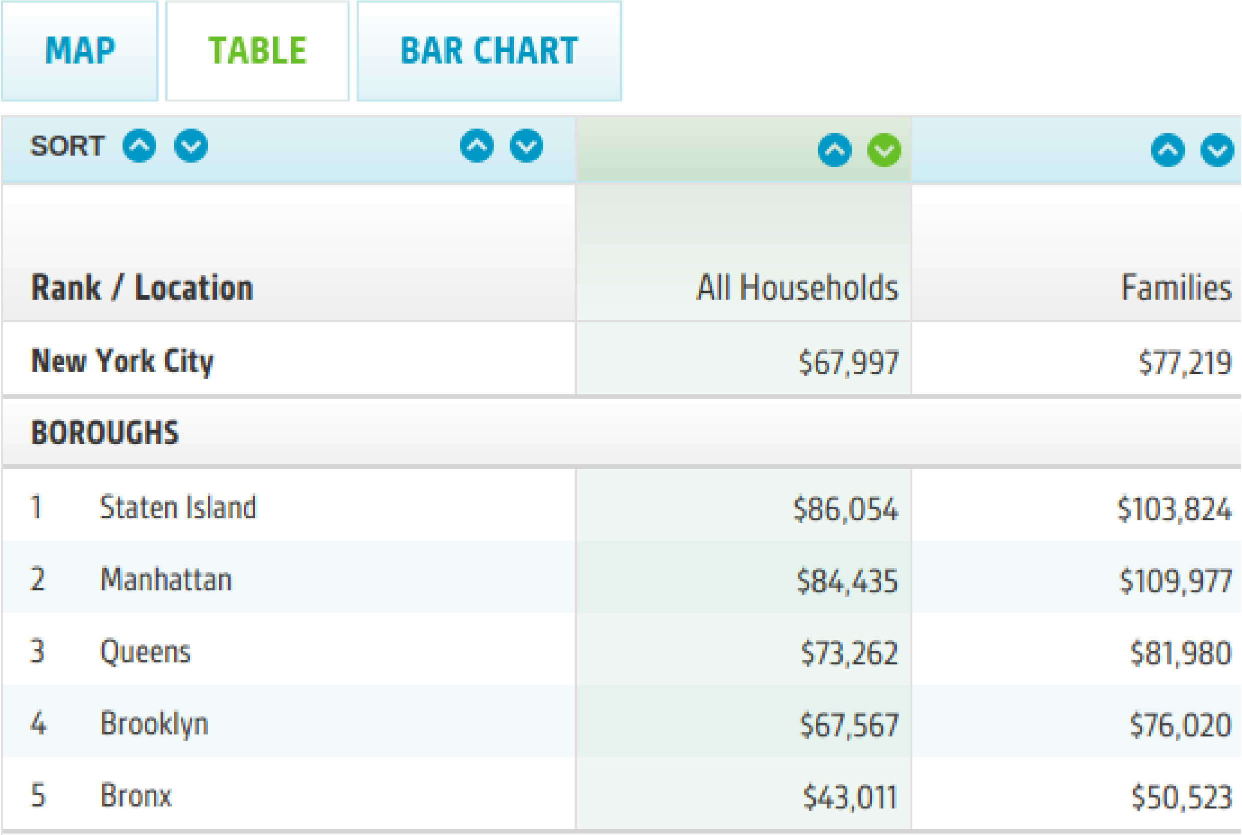
5.Relationship between shootings and areas

We will look into how borough affects the shootings occurence rates, espcially the relationship between median income vs shootings incident.

From the image, we can come up with a similar data frame in R

```
medianBoroIncome <- data.frame( Rank = c(1, 2, 3, 4, 5), BORO = c("STATEN ISLAND", "MANHATTAN", "QUEENS
```

We will plot the total of shooting incidents along side with the median income of each borough to see the



Rank / Location		All Households	Families
New York City		$67,997	$77,219
BOROUGHS			
1	Staten Island	$86,054	$103,824
2	Manhattan	$84,435	$109,977
3	Queens	$73,262	$81,980
4	Brooklyn	$67,567	$76,020
5	Bronx	$43,011	$50,523

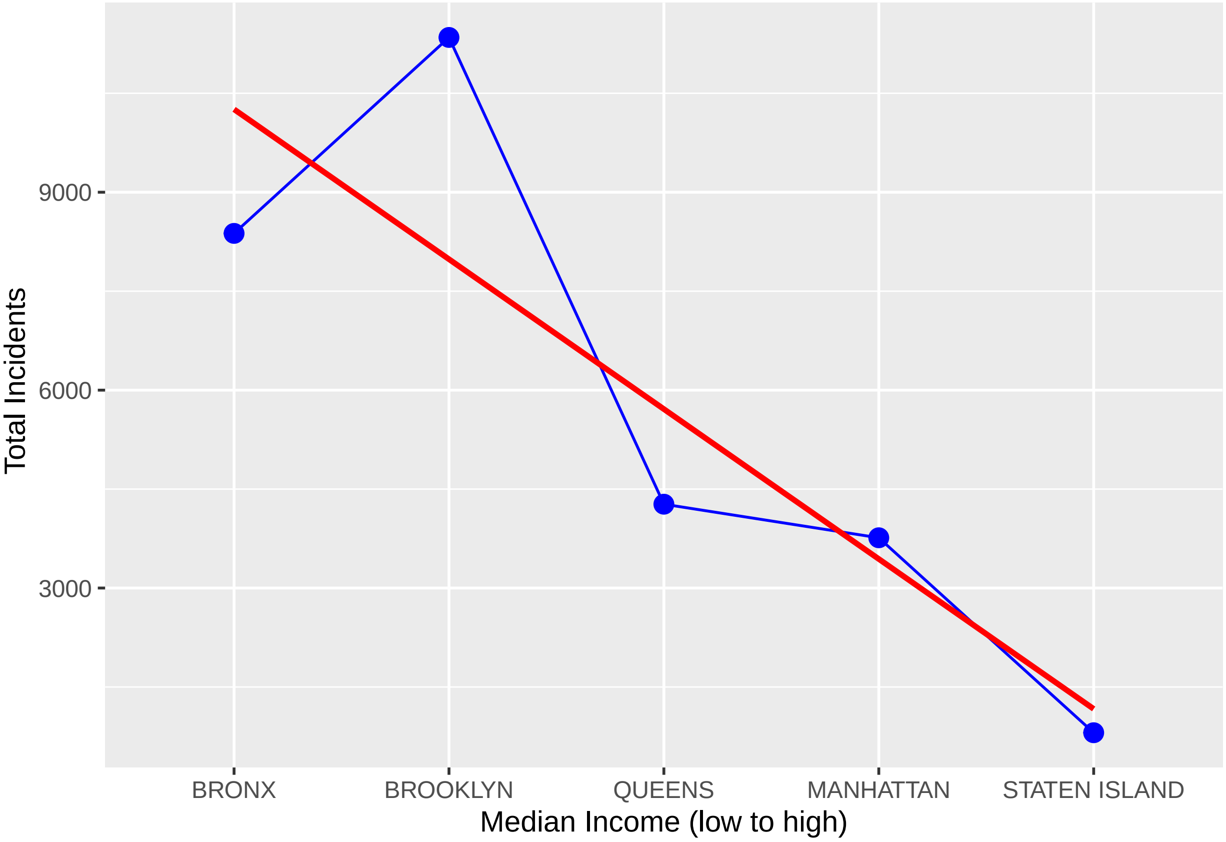
Figure 1: household income by borough at https://data.cccnewyork.org/data/table/66/median-incomes#66/107/127/a/a

relationship

```
medianBoroIncome <- medianBoroIncome %>%
  left_join(rawData %>% count(BORO, name = "incidents"),by = "BORO" )

incomeVcrime <- ggplot(medianBoroIncome, aes(x= reorder(BORO,All_Households),y=incidents,group=1)) +
  geom_point(color = "blue", size = 3) +
  geom_line(color="blue") +
  geom_smooth(method = "lm",se=FALSE,colour = "Red") +
  labs(x="Median Income (low to high)",y = " Total Incidents")

print(incomeVcrime)
```



There's a clear inverse relationship between median income and the number of shootings in each borough as shown by the red line (using regression model) in the graph.

6.Conlcusion

According to the data, the likelihood of shootings increases in neighborhoods with lower income levels, particularly during late-night hours.