NYPD Shooting Incident - Week 3

6/18/2021

NYPD Shooting Incidents_Historic Data

Purpose:

1. Understand, clean and analyze NYPD shooting incidents historic data from the year 2006 to 2020.
2. Focus on understanding the shooting incidents occurring in different BOROs, which can help us understand the safety of different neighborhoods in New York around NYC.
3. Investigate any anomaly in the location wise incidents.
4. Find a pattern in the day and hour wise shooting incidents for the neighborhood with highest number of shootings.

Data source & description:

1. Using the NYPD Shooting Incidents Historical data containing reported incidents across 5 boroughs from the year 2006 to 2020
2. The link to the .csv file used for this analysis: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD
3. This data contains reports on the shooting incidents across 5 boroughs, their time and date of occurrence, the perpetrator race, sex and age group, the victim race, sex and age group, location description of the incident, precinct, jurisdiction code, statistical murder flag, latitude and longitude information of incident.

Library the necessary packages

- install.packages("tidyverse")
- install.packages("lubridate")
- install.packages("ggplot2")
- install.packages("kableExtra")
- install.packages("dplyr")

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(kableExtra)
library(dplyr)
```

Importing NYPD Shooting Incident Dataset

```
##Get NYPD Shooting Incident csv data

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

Reading and Storing the data in nypd_csvData

```
#Reading the nypd csv from the url and storing it in nypd_csvData data frame

nypd_csvData <- read.csv(url_in)
summary (nypd_csvData)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME              BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.3323
##  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: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       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     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
##
```

Tidying data

```
## Removed 5 columns (Lat, Long, Lat_Lon, X-coord, Y-coord) which were not needed for the analysis, usi
## Convert date column from chr data type to date data type in YYYY:MM:DD format, using lubridate packag

nypd_csvData <-
select(nypd_csvData, -c(Lon_Lat, Latitude, Longitude, X_COORD_CD, Y_COORD_CD)) %>%
mutate(OCCUR_DATE = mdy(`OCCUR_DATE`))
summary (nypd_csvData)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:23568       Length:23568
##  1st Qu.: 55317014   1st Qu.:2008-12-30   Class :character   Class :character
##  Median : 83365370   Median :2012-02-26   Mode  :character   Mode  :character
##  Mean   :102218616   Mean   :2012-10-03
##  3rd Qu.:150772442   3rd Qu.:2016-02-28
##  Max.   :222473262   Max.   :2020-12-31
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.3323
##  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: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
##  Length:23568       Length:23568
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

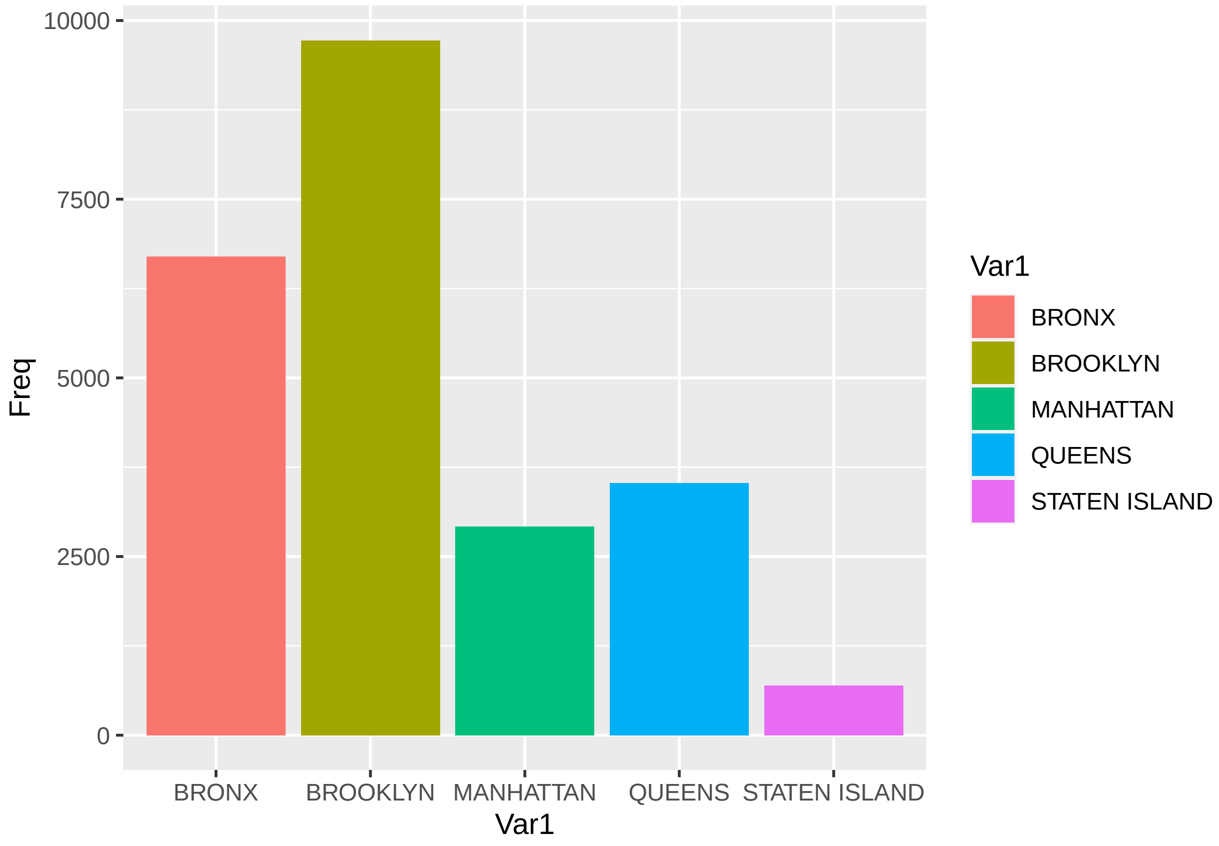
Data Analysis and Visualization

Analysis of incidents in each BORO

```r
#Preparing data for plotting and generating a frequency table for BORO using kableExtra package
#Created a table of BORO from the nypd_csvData data set and assigned it to 'borough'
#Converted the table 'borough' into data frame 'borough' using as.data.frame
#Calculating the percentage of the frequency of shootings in each BORO
#Using kable library got the frequency and frequency % table for incidents in each BORO

borough <- table(nypd_csvData$BORO)
borough <- as.data.frame(borough)
borough$Percent <- round((borough$Freq / sum(borough$Freq)*100),2)
kable(borough)
```

Var1	Freq	Percent
BRONX	6700	28.43
BROOKLYN	9722	41.25
MANHATTAN	2921	12.39
QUEENS	3527	14.97
STATEN ISLAND	698	2.96

```r
#From the above frequency table we understand that Brooklyn BORO had the highest shootings
#Plotting bar graph for number of incidents in BORO vs BORO using ggplot

ggplot(borough, aes(x=Var1, y=Freq, fill=Var1)) + geom_bar(stat="identity")
```

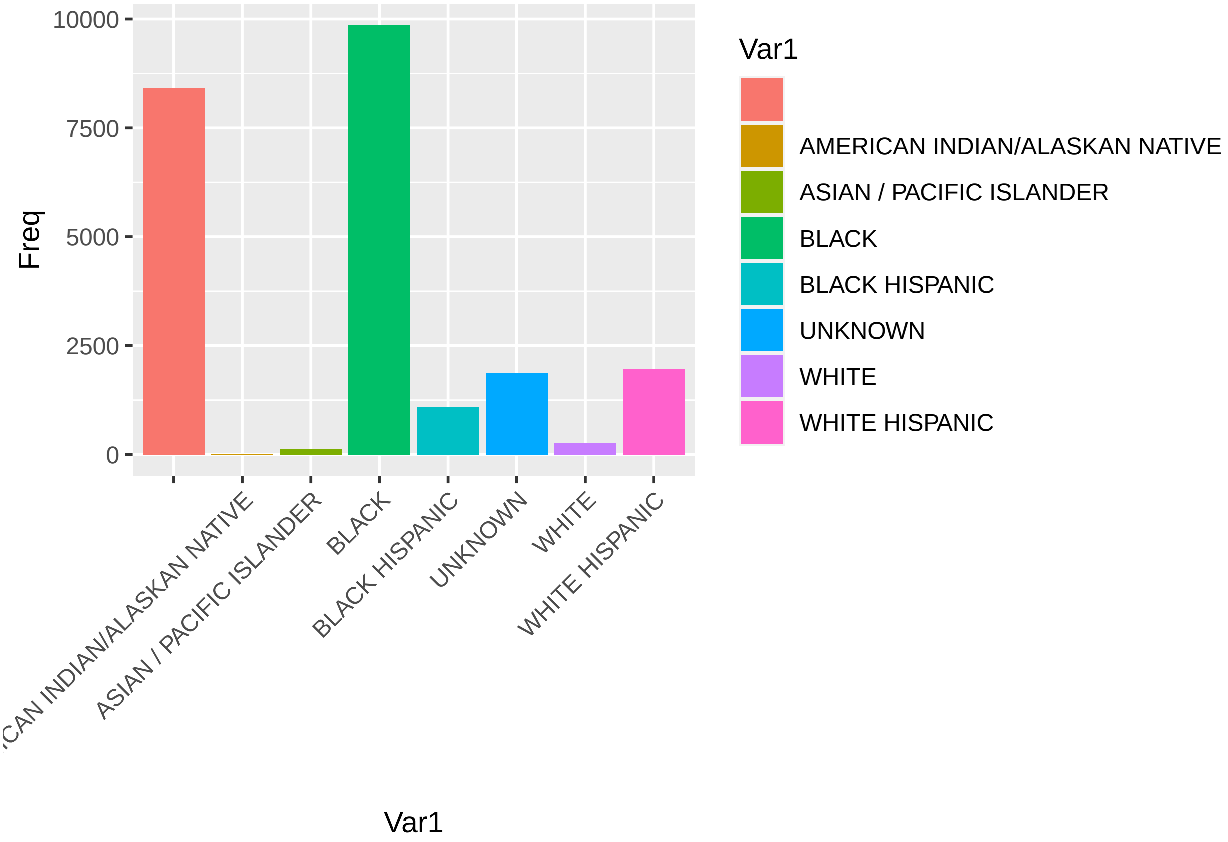


Analysis of incidents for each PERP_RACE

```r
#Preparing data for plotting and generating a frequency table for PERP_RACE using kableExtra package
#Created a table of PERP_RACE from the nypd_csvData data set and assigned it to 'perp_race'
#Converted the table 'perp_race' into data frame 'perp_race' using as.data.frame
#Calculating the percentage of the frequency of shootings for each PERP_RACE
#Using kable library got the frequency and frequency % table for incidents for each PERP_RACE

perp_race <- table(nypd_csvData$PERP_RACE)
perp_race <- as.data.frame(perp_race)
perp_race$Percent <- round((perp_race$Freq / sum(perp_race$Freq)*100),2)
kable(perp_race)
```

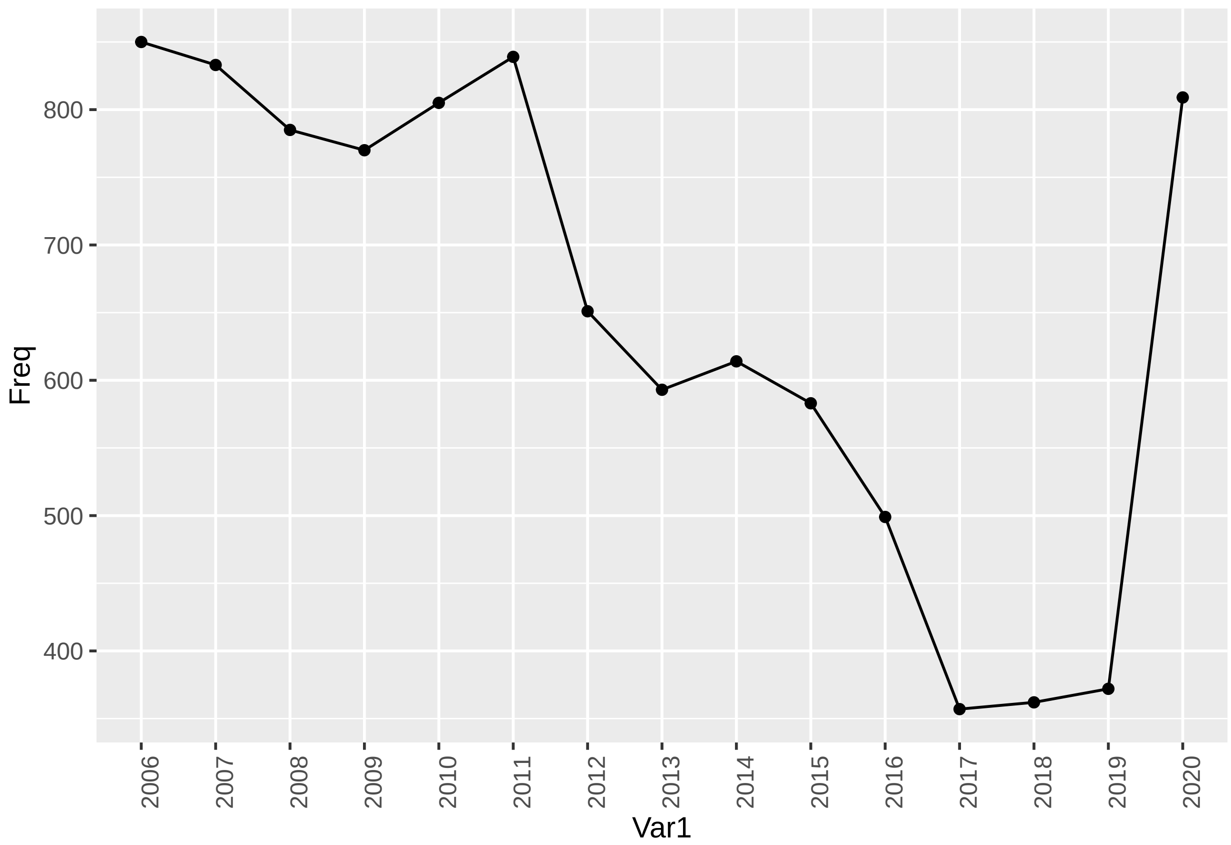
Var1	Freq	Percent
	8425	35.75
AMERICAN INDIAN/ALASKAN NATIVE	2	0.01
ASIAN / PACIFIC ISLANDER	120	0.51
BLACK	9855	41.82
BLACK HISPANIC	1081	4.59
UNKNOWN	1869	7.93
WHITE	255	1.08
WHITE HISPANIC	1961	8.32

```r
#From the above data we understand that for 35.75% of shooting incidents the perp_race was not reported
#Plotting bar graph of # of incidents by each PERP_RACE

ggplot(perp_race, aes(x=Var1, y=Freq, fill=Var1)) + geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Analysis of total shootings in BORO- Brooklyn in each year

```
#From the BORO analysis we understand that Brooklyn BORO had the highest number of shootings hence eval

#Creating a subset for all the incidents in Brooklyn by Occurrence date (OCCUR_Date) from nypd_csvData

brooklyn <-subset(nypd_csvData, BORO=='BROOKLYN', select=c(BORO, OCCUR_DATE))

#Extracting the year component from the date from OCCUR_DATE column using substr
brooklyn$YEAR <- substr(brooklyn$OCCUR_DATE, (nchar(brooklyn$OCCUR_DATE) - 4), nchar(brooklyn$OCCUR_DAT

#De-selecting OCCUR_DATE from the subset
brooklyn <- subset(brooklyn, select = -c(OCCUR_DATE))

#creating table of shooting incidents in brooklyn and the year of occurrence and assigning it to BRLYN
#Converting BRLYN table into  dataframe using as.data.frame
BRLYN <- table(brooklyn$YEAR)
BRLYN <- as.data.frame(BRLYN)

#plotting the data on a line & point graph using ggplot

ggplot(data=BRLYN, aes(x=Var1, y=Freq, group=1)) +
  geom_line()+
  geom_point()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Analysis of incidents in Brooklyn by day and hour

```
#Creating a subset of all the incidents in Brooklyn from nypd_csvData
#Reformatting the date to mm-dd-yyyy
#Creating and extracting the weekday from the year$OCCUR_DATE and assigning it to year$DAY

year <- subset(nypd_csvData, BORO =="BROOKLYN")
year$OCCUR_DATE <- as.Date(year$OCCUR_DATE,format = "%m/%d/%Y")
year$DAY<- wday(year$OCCUR_DATE, label=TRUE)
```

```
#Created a function 't' to split OCCUR_TIME string at the hour value hence delimiter ':'and converting
t <- function(x) {
    if(!is.null(x)) {
        return (as.numeric(strsplit(x,":")[[1]][1]))
    }
}

#Creating HOUR column to the year data set and grouping by HOUR and DAY
hour = year %>% mutate(HOUR = sapply(OCCUR_TIME, t)) %>% group_by(DAY, HOUR) %>% summarize(count = n())
```

```
## ‘summarise()‘ has grouped output by ’DAY’. You can override using the ‘.groups‘ argument.
```

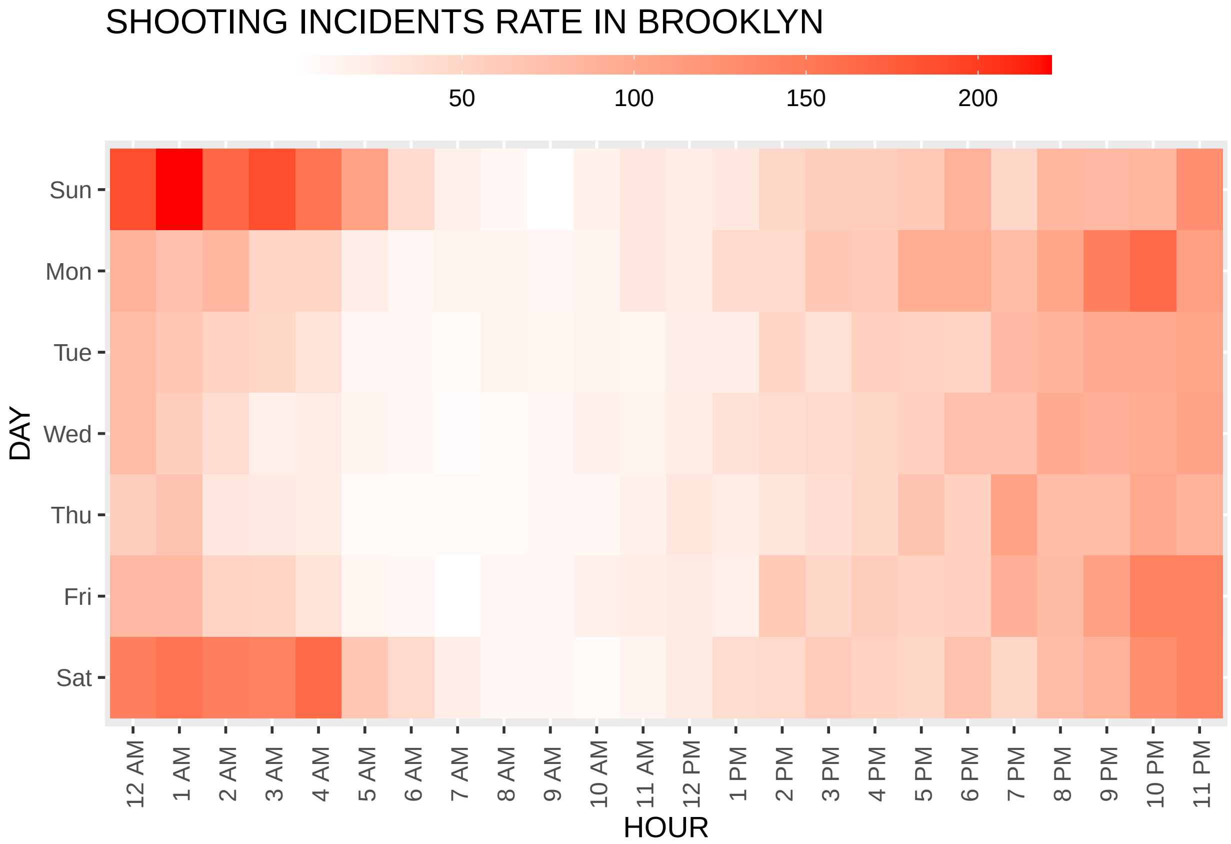
```
#Created specific vectors for day and hr
day <- c("Sun","Mon","Tue","Wed","Thu","Fri","Sat")
hr <- c(paste(c(12,1:11),"AM"), paste(c(12,1:11),"PM"))
```

```
#Converting HOUR and DAY variable to specific variables above
hour$DAY <- factor(hour$DAY, level = rev(day))
hour$HOUR <- factor(hour$HOUR, level = 0:23, label = hr)
```

Plotting a heat map of all the shooting incidents in Brooklyn in the year 2020, to understand the highest incidents occurrence times and days

```
#Plotting a heat map of all the shooting incidents in Brooklyn by day and time

ggplot(hour, aes(x = HOUR, y = DAY, fill = count)) +geom_tile() + theme(axis.text.x = element_text(angl
```

```
## Warning: ‘legend.margin‘ must be specified using ‘margin()‘. For the old
## behavior use legend.spacing
```



Conclusion:

Insights:

1. The BORO with highest shooting incident was Brooklyn.
2. There was a drastic increase in the shooting incidents in Brooklyn in the year 2020 from the year 2019. That could be attributed to: 2.a. The COVID-19 pandemic 2.b. Lockdowns 2.c. Daily wage workers compelled to leave their jobs which led to a rise in unemployment crisis 2.d. Rise in domestic violence in the country.

3. Also, the number of shootings in 2020 were in the same range of the number of shootings in the year 2008 when USA was struck by financial recession.
4. One can say that, there is a correlation between unemployment and frequency of shooting incidents.
5. In Brooklyn, the most frequent days and times of shootings were on weekends between 11pm to 4am, as attributed by the heat map.

Data Bias

1. During the analysis of perpetrator race, due to a lot of unknown values and unreported race data there is a huge bias to consider the Black race as the highest number of perpetrators
2. The other source of bias could be introduced by the number of incidents reported. If shootings in other neighborhoods are not reported or under reported unlike they are in the Brooklyn neighborhood then our analysis might be misguided.