NYPDShootingIncidentData

Theerawan Cox

2022-07-02

Load Libraries

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     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
```

Load Data

We will get the data from "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD" And we will store it in the NYPDShooting variable

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

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
NYPDShooting
```

```
## # A tibble: 25,596 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>        <dbl>             <dbl>
##  1     24050482 08/27/2006 05:35      BRONX           52                 0
##  2     77673979 03/11/2011 12:03      QUEENS         106                 0
##  3    226950018 04/14/2021 21:08      BRONX           42                 0
##  4    237710987 12/10/2021 19:30      BRONX           52                 0
##  5    224701998 02/22/2021 00:18      MANHATTAN       34                 0
##  6    225295736 03/07/2021 06:15      BROOKLYN        75                 0
##  7    231190175 07/21/2021 00:40      MANHATTAN       32                 0
##  8    233429421 09/11/2021 20:20      MANHATTAN       26                 2
##  9    227950661 05/09/2021 02:50      BRONX           41                 2
## 10    227344198 04/23/2021 13:25      BROOKLYN        67                 0
## # ... with 25,586 more rows, and 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

```
summary(NYPDShooting)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  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
##
```

Data Summary

As you can see, the data has columns INCIDENT_KEY, OCCUR_DATE(in the char type), OCCUR_TIME, BORO, PRECINCT, and 14 more. I will keep only the columns we need and also change the char type to date type on OCCUR_DATE. I also added a “case” column with the number “1” to represent one case for each row.

```
NYPDShooting <- NYPDShooting %>%
  select(c(1,2,3,4)) %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%Y"),
         case = 1)

NYPDShooting = NYPDShooting%>%
  mutate(OCCUR_MONTH = as.numeric(format(NYPDShooting$OCCUR_DATE, '%m')))
summary(NYPDShooting)
```

```
##   INCIDENT_KEY          OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596      Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms        Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##       case     OCCUR_MONTH
##  Min.   :1   Min.   : 1.000
##  1st Qu.:1   1st Qu.: 5.000
##  Median :1   Median : 7.000
##  Mean   :1   Mean   : 6.857
##  3rd Qu.:1   3rd Qu.: 9.000
##  Max.   :1   Max.   :12.000
```

```
sum(is.na(NYPDShooting))
```

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

Checking Data

From the summary above, we see there is no Null in our data. Other columns also look good. Now the data is ready to use.

Plotting Graph Between Number of Case and Month on Each City

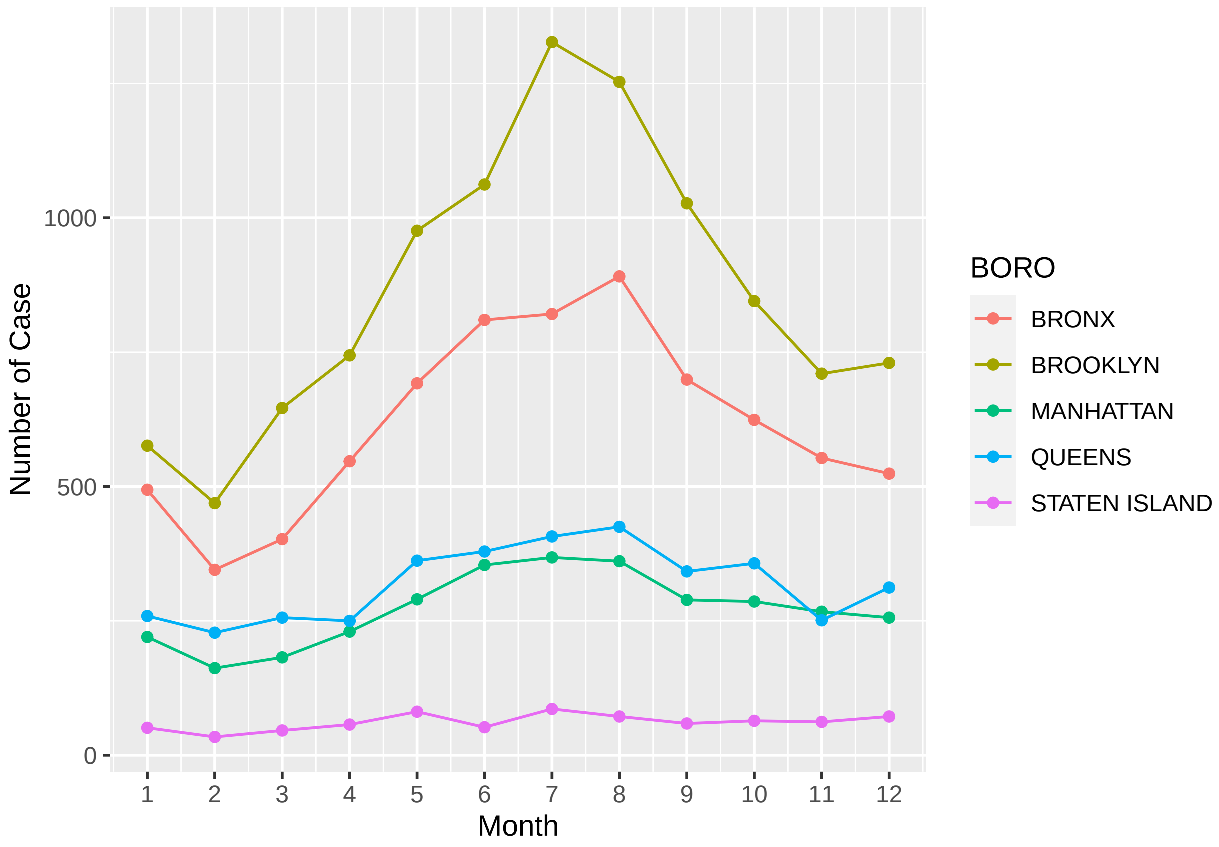
```
NYPDShootingInMonth = NYPDShooting%>%
  group_by(OCCUR_MONTH, BORO)%>%
  summarise(case = sum(case))
```

```
## `summarise()` has grouped output by 'OCCUR_MONTH'. You can override using the
## `.groups` argument.
```

```
sum(is.na(NYPDShooting))
```

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

```
NYPDShootingInMonth %>%
  ggplot(aes(x = OCCUR_MONTH, y = case)) +
  geom_point(aes(color = BORO)) +
  geom_line(aes(color = BORO)) +
  scale_x_continuous(breaks=c(1:12)) +
  labs(x = "Month", y = "Number of Case")
```



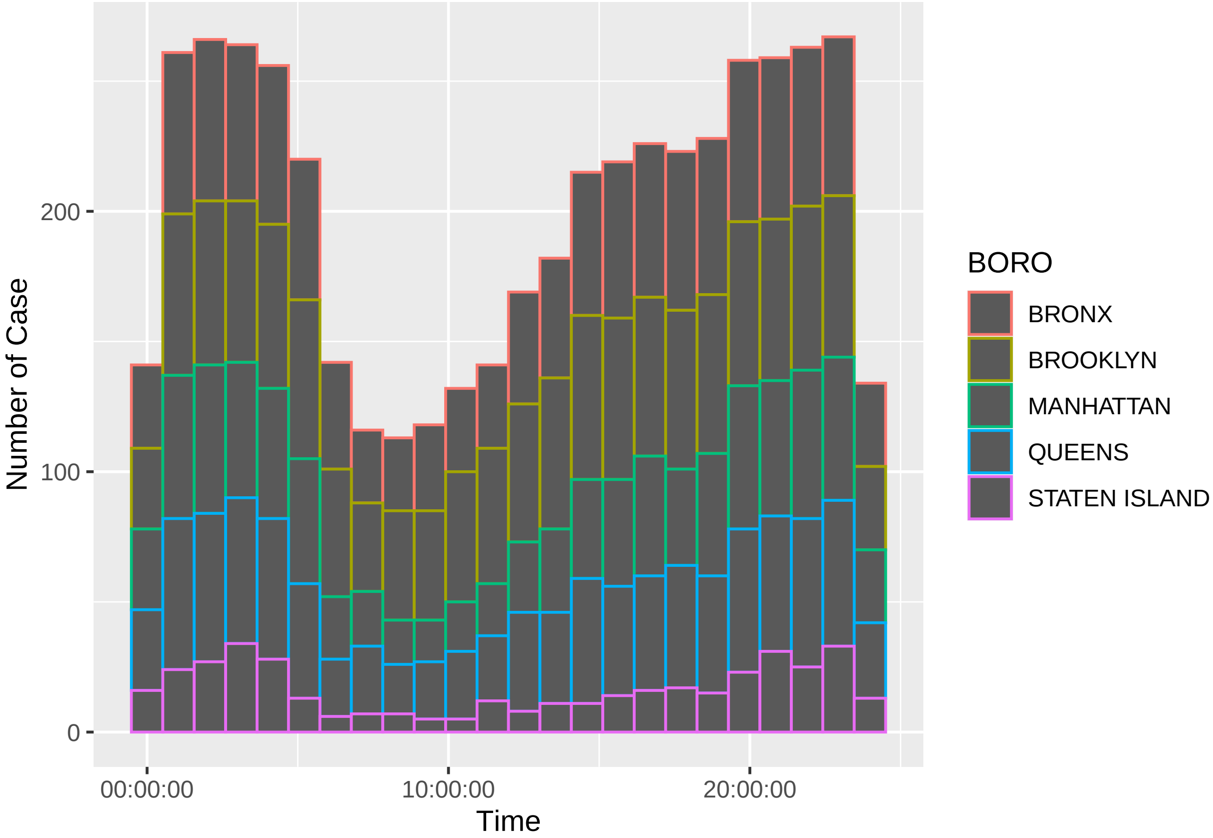
# Conclusion Per the data visualization above, we see that July and August have the most cases for all cities in our data. And the cases are lowest during December, January, February. Why? Is this something to do with summer and winter? Would the temperature relate to people's emotions?

Plotting Histogram Between Number of Case and Time by Cities

```
NYPDShootingTime = NYPDShooting%>%
  group_by(OCCUR_TIME, BORO)%>%
  summarise(case = sum(case))
```

```
## `summarise()` has grouped output by 'OCCUR_TIME'. You can override using the
## `.groups` argument.
```

```
NYPDShootingTime %>%
  ggplot(aes(x = OCCUR_TIME)) +
  geom_histogram(aes(color = BORO), bins = 24) +
  labs(x = "Time", y = "Number of Case")
```



# Conclusion Per the data visualization above, it seems like the cases increase during the evening to early morning and drop down during the daylight hours, which makes sense that the perpetrators mostly commit crime after sunset and before sunrise. However, the data shows that the cases drop a lot during 23:00 - 1:00 o'clock. Does the data has an error?

Bias sources

Regarding this data, I didn't choose to do race or age analysis due to Asian hated crime that I heard all over the news. This would introduce more bias to the conclusions of the report. For my analysis I chose month and time because it would help us know when the crime rate is high and we can try to avoid travelling to or going out during the risky months and time.