Week3 NY SHOOTING REPORT

Data imported from “https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv”

Introduction

We have here a table of 23,585 shooting incidents in New York occurring from 2006 to 2020. Data are from government publications.
Even if both information are available, for this report I will focus on the victim rather than the perpetrator. This is a slightly biased report in that I choose to consider the victim side, but this will have no impact on the outcome of the analysis.

```
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()
```

```
data_collected = read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

```
## Rows: 23585 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.
```

```
head(data_collected)
```

```
## # A tibble: 6 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    203350417 10/06/2019 01:09      BROOKLYN       77                 0
## 4     80584527 09/04/2011 03:35      BRONX          40                 0
## 5     90843766 05/27/2013 21:16      QUEENS        100                 0
## 6     92393427 09/01/2013 04:17      BROOKLYN       67                 0
## # ... with 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(data_collected)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23585       Length:23585      Length:23585
##  1st Qu.: 55322804   Class :character   Class1:hms        Class :character
##  Median : 83435362   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102280741                      Mode  :numeric
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   FALSE:19085
##  Median : 69.00   Median :0.000     Mode  :character   TRUE :4500
##  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
##
```

Comments

The data are pretty clean, there are no inconvenient “NAs” or missing useful values for our analyses.

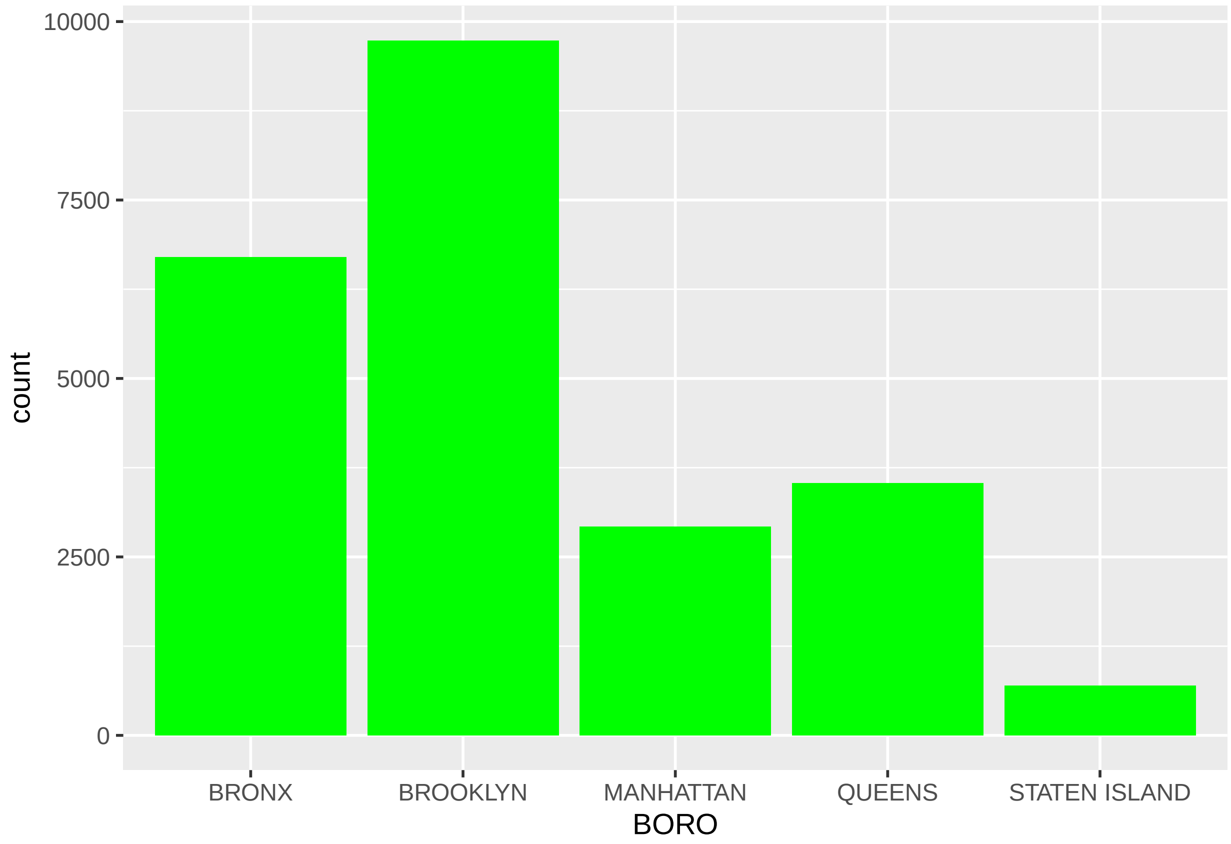
Let’s select variables for our analyses. We will focus on the victims rather than the perpetrators, the “bias” here is to highlight the likelihood to be a potential victim of shootings in New York.

```
data_collected=data_collected %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME,BORO,VIC_AGE_GROUP, VIC_S
data_collected
```

```
## # A tibble: 23,585 x 7
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     VIC_AGE_GROUP VIC_SEX VIC_RACE
##           <dbl> <chr>      <time>     <chr>    <chr>         <chr>   <chr>
##  1     24050482 08/27/2006 05:35      BRONX    25-44         F       BLACK HISP~
##  2     77673979 03/11/2011 12:03      QUEENS   65+           M       WHITE
##  3    203350417 10/06/2019 01:09      BROOKLYN 18-24         F       BLACK
##  4     80584527 09/04/2011 03:35      BRONX    <18           M       BLACK
##  5     90843766 05/27/2013 21:16      QUEENS   18-24         M       BLACK
##  6     92393427 09/01/2013 04:17      BROOKLYN <18           M       BLACK
##  7     73057167 06/05/2010 21:16      BROOKLYN <18           M       BLACK
##  8    211362213 03/20/2020 21:27      BROOKLYN 25-44         M       BLACK
##  9    137564752 07/04/2014 00:25      QUEENS   18-24         M       BLACK
## 10    147024011 10/18/2015 01:33      QUEENS   18-24         M       BLACK
## # ... with 23,575 more rows
```

A) Analysis of general trends
1)Number of victims per borough

```
ggplot(data_collected,aes(x=BORO))+geom_bar(position="stack", fill="green")
```



Let's import some population data from another government website to further compare these numbers with the population levels.

Source = https://www1.nyc.gov/assets/planning/download/pdf/planning-level/nyc-population/projections_briefing_booklet.pdf

There are no significant increase in the populations from 2005 to 2020, so we can use 2020 population for each borough as a reference.

$BRONX = 1,420,277$

$BROOKLYN = 2,628,211$

$MANHATTAN = 1,729,530$

$QUEENS = 2,396,949$

$STATEN\ ISLAND = 517,597$

$TOTAL\ NEW\ YORK = 8,692,564$

```
sum(data_collected$BORO=="BRONX")/23585
```

```
## [1] 0.2841213
```

```
sum(data_collected$BORO=="BROOKLYN")/23585
```

```
## [1] 0.4127199
```

```
sum(data_collected$BORO=="MANHATTAN")/23585
```

```
## [1] 0.1238923
```

```
sum(data_collected$BORO=="QUEENS")/23585
```

```
## [1] 0.1497562
```

```
sum(data_collected$BORO=="STATEN ISLAND")/23585
```

```
## [1] 0.02951028
```

Comparing with population ratios

```
1420277/8692564
```

```
## [1] 0.1633899
```

```
2628211/8692564
```

```
## [1] 0.3023516
```

```
1729530/8692564
```

```
## [1] 0.1989666
```

```
2396949/8692564
```

```
## [1] 0.2757471
```

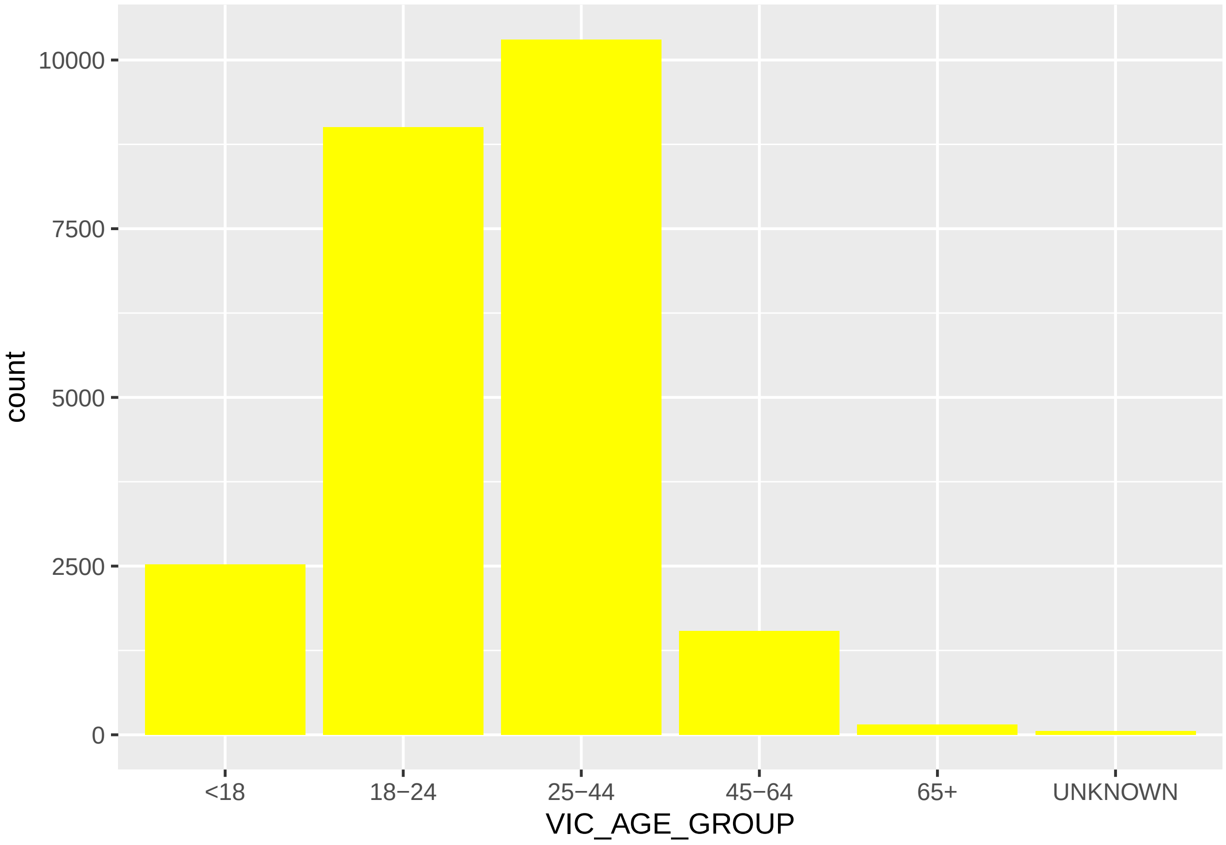
```
517597/8692564
```

```
## [1] 0.0595448
```

Comments

So the relative crime rates for Brooklyn and also Bronx are clearly higher with respect to the population levels, 41% vs 30% and 28% vs 16%. 2)Number of victims per age category

```
ggplot(data_collected,aes(x=VIC_AGE_GROUP))+geom_bar(position="stack", fill="yellow")
```



Comments

No surprise there, violent crime victims mostly belong to the young and relative young population in every city.

Let's isolate time, month and year variables.

```
glimpse(data_collected)
```

```
## Rows: 23,585
## Columns: 7
## $ INCIDENT_KEY  <dbl> 24050482, 77673979, 203350417, 80584527, 90843766, 92393~
## $ OCCUR_DATE    <chr> "08/27/2006", "03/11/2011", "10/06/2019", "09/04/2011", ~
## $ OCCUR_TIME    <time> 05:35:00, 12:03:00, 01:09:00, 03:35:00, 21:16:00, 04:17~
## $ BORO          <chr> "BRONX", "QUEENS", "BROOKLYN", "BRONX", "QUEENS", "BROOK~
## $ VIC_AGE_GROUP <chr> "25-44", "65+", "18-24", "<18", "18-24", "<18", "<18", "~
## $ VIC_SEX       <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "F", "~
## $ VIC_RACE      <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "BLACK", "B~
```

```
data_collected=data_collected %>% separate(OCCUR_TIME,c("Crime_hour","Crime_min")) %>% separate(OCCUR_DA
```

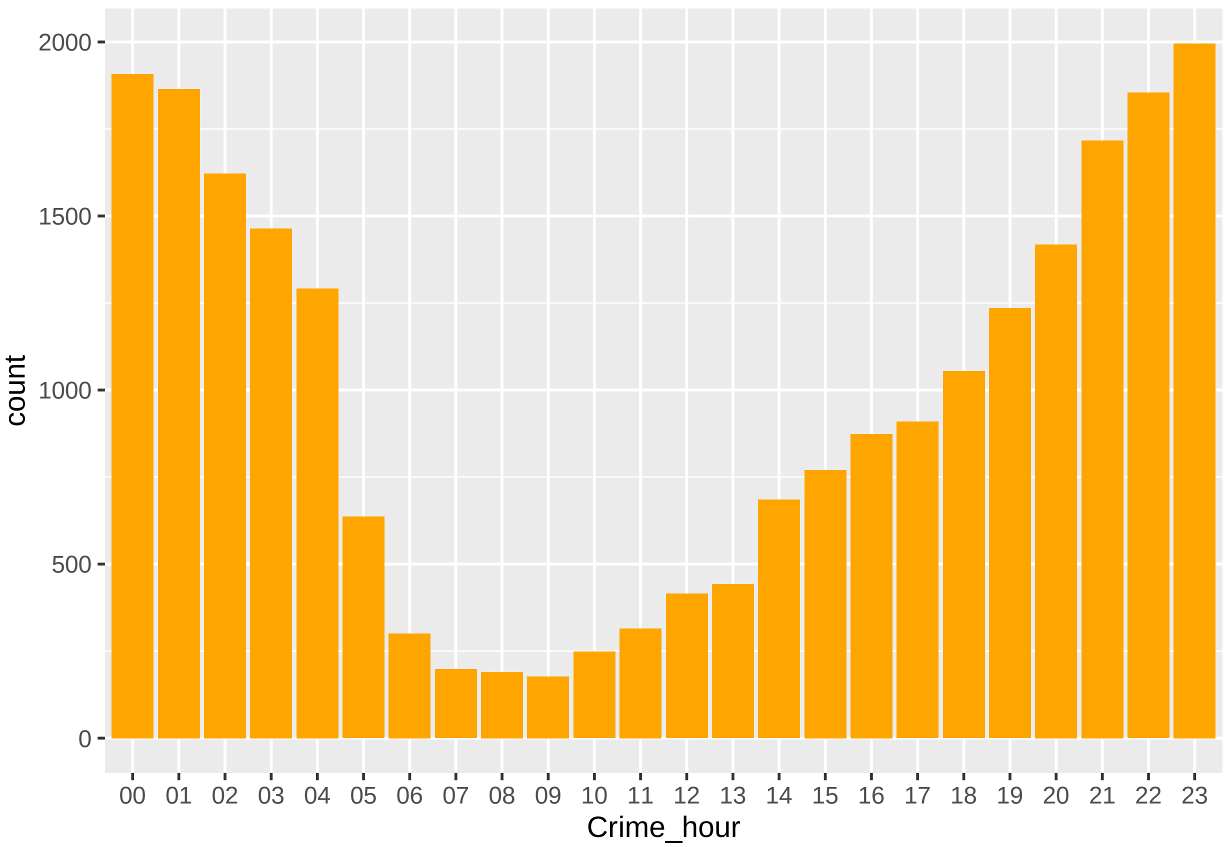
```
## Warning: Expected 2 pieces. Additional pieces discarded in 23585 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
data_collected
```

```
## # A tibble: 23,585 x 10
##    INCIDENT_KEY Crime_month Crime_day Crime_year Crime_hour Crime_min BORO
##           <dbl> <chr>       <chr>     <chr>      <chr>      <chr>     <chr>
##  1     24050482 08          27        2006       05         35        BRONX
##  2     77673979 03          11        2011       12         03        QUEENS
##  3    203350417 10          06        2019       01         09        BROOKLYN
##  4     80584527 09          04        2011       03         35        BRONX
##  5     90843766 05          27        2013       21         16        QUEENS
##  6     92393427 09          01        2013       04         17        BROOKLYN
##  7     73057167 06          05        2010       21         16        BROOKLYN
##  8    211362213 03          20        2020       21         27        BROOKLYN
##  9    137564752 07          04        2014       00         25        QUEENS
## 10    147024011 10          18        2015       01         33        QUEENS
## # ... with 23,575 more rows, and 3 more variables: VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>
```

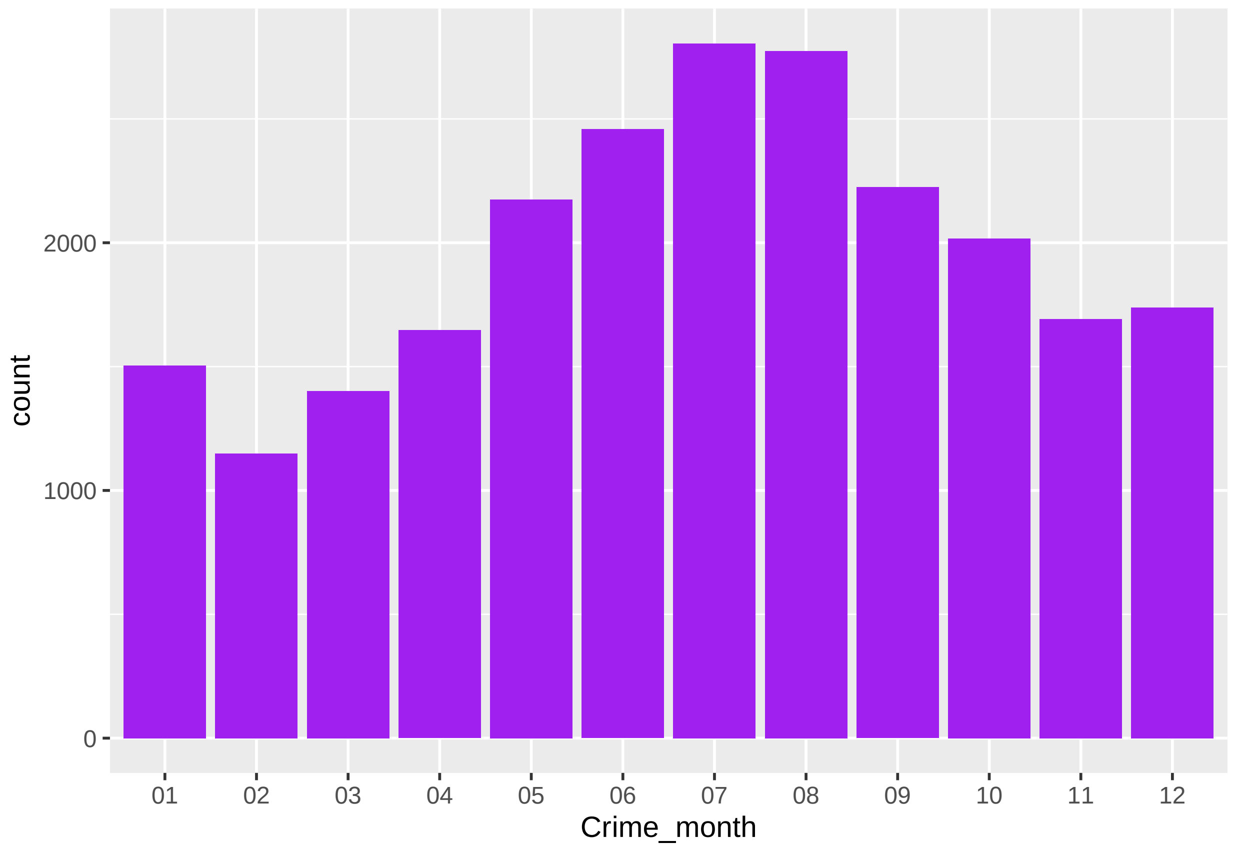
3)Number of victims per hour

```
ggplot(data_collected,aes(x=Crime_hour))+geom_bar(position="stack", fill="orange")
```



Comments
The slopes are definitely clear, the shootings increase the later it gets up to 23:00, the "crime ideal time", then start to decrease as sunrise gets closer. 4)Number of victims per month

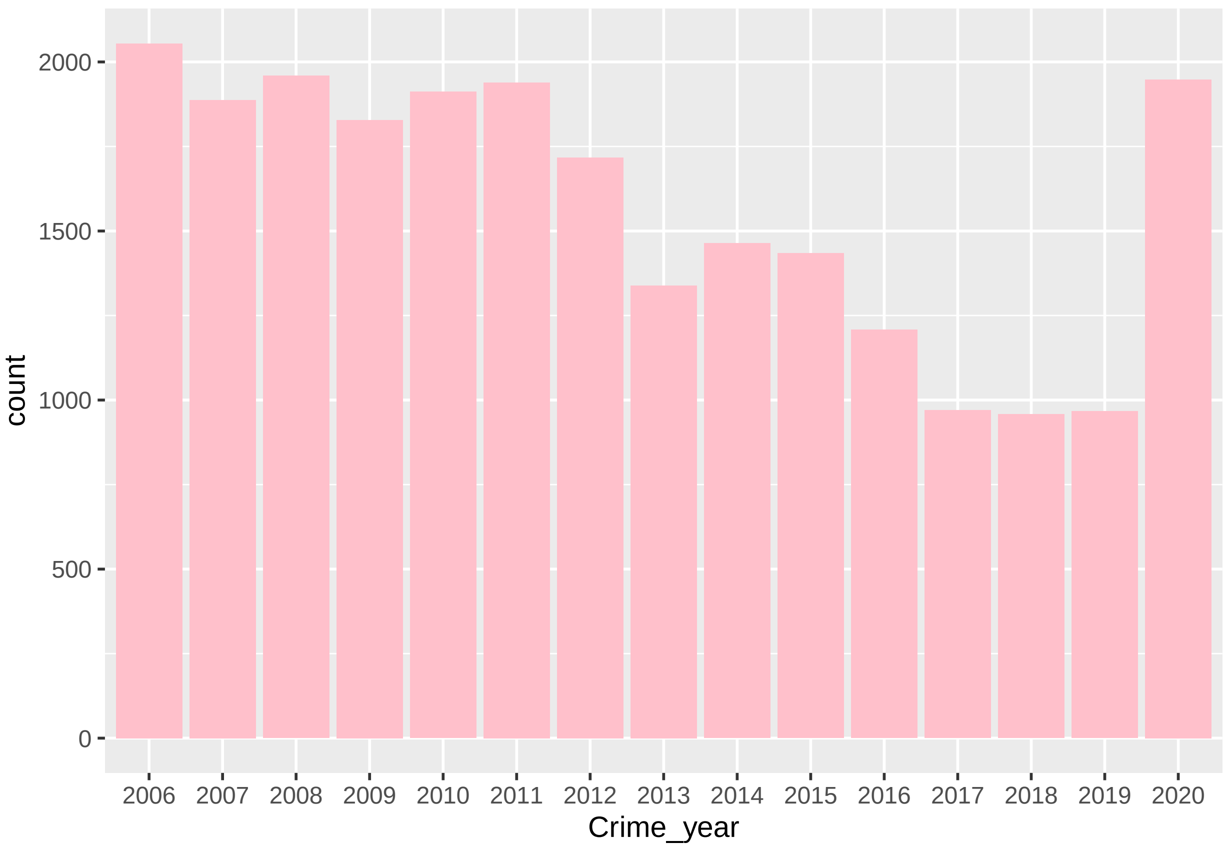
```
ggplot(data_collected,aes(x=Crime_month))+geom_bar(position="stack", fill="purple")
```



Comments

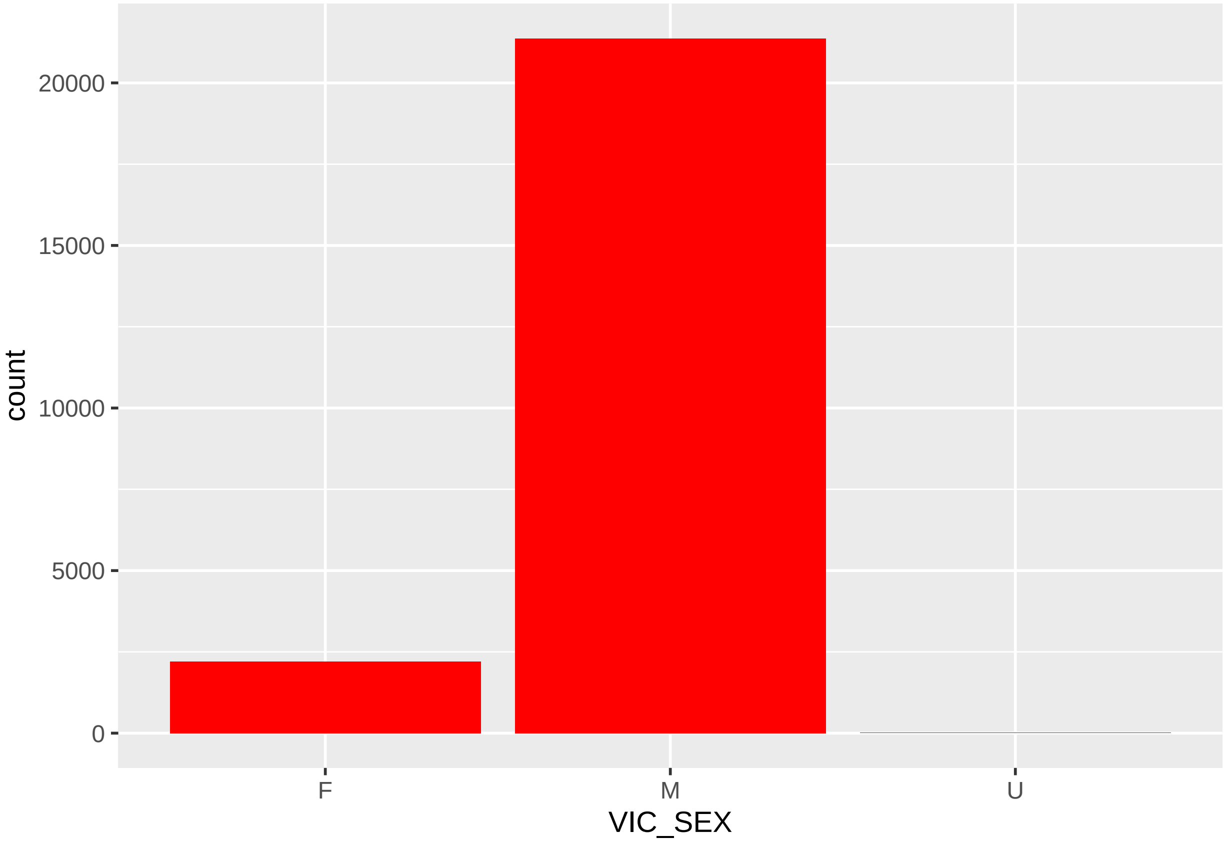
Same as for the day analysis, there is also a clear trend throughout the year.
It's surprising to see an increase as the weather gets nicer as if crime rate was boosted by the sunny days!!
This is a "funny" paradox given that the shooting peak is at night time when there is no sun anymore.
5)Number of victims per year

```
ggplot(data_collected,aes(x=Crime_year))+geom_bar(position="stack", fill="pink")
```



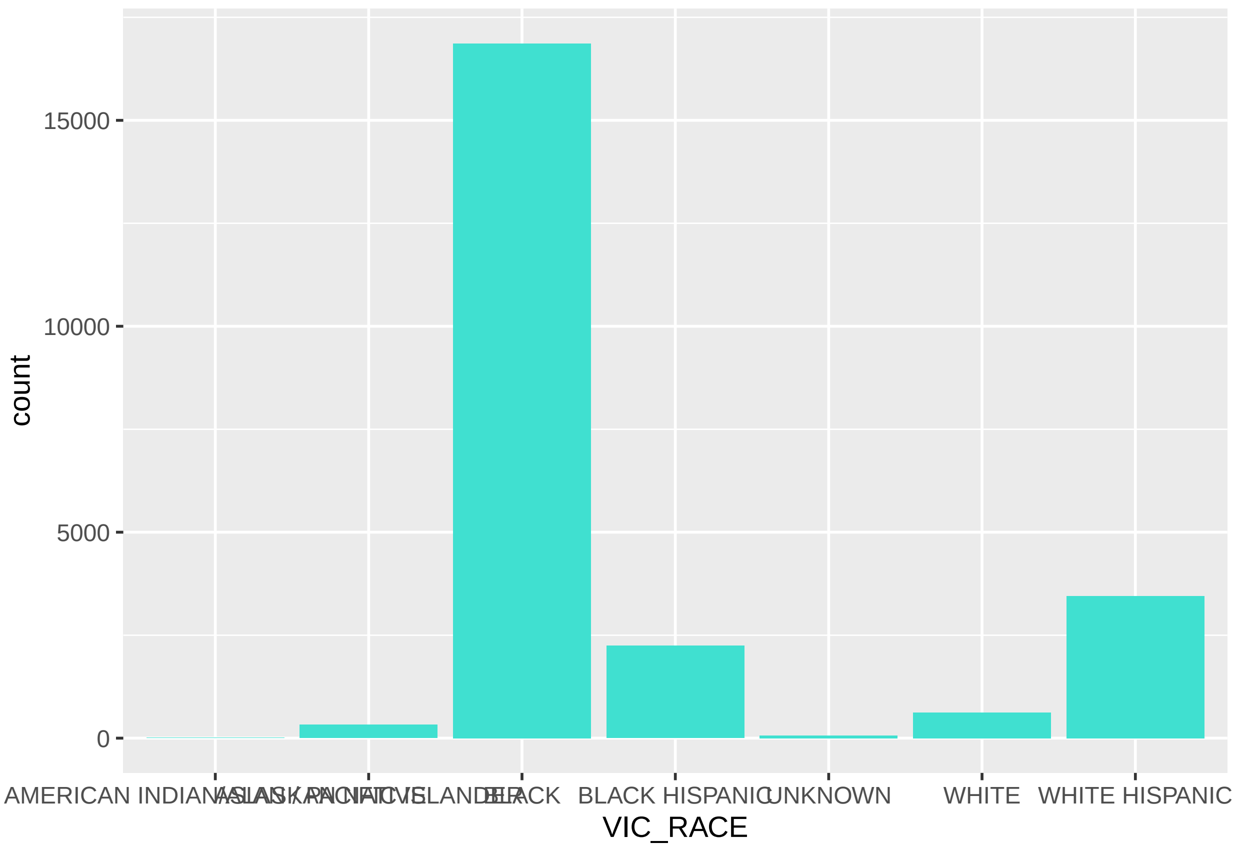
Comments
It might be interesting to understand the rationale behind the shocking rise in shootings in 2020, the first covid year. We would expect the steady trend from 2017 to 2019 to keep going or even decrease, but we are facing twice as much shootings as during each of all three previous years. Has covid crisis definitely ruined so many years of improvement in violent crime numbers in New York?
6)Number of victims per gender

```
ggplot(data_collected,aes(x=VIC_SEX))+geom_bar(position="stack", fill="red")
```



Comments
No surprise there, violent crime rate has always been higher within male population. 7)Number of victims per race

```
ggplot(data_collected,aes(x=VIC_RACE))+geom_bar(position="stack", fill="turquoise")
```



Comments
The highest potential victim is a black man, aged between 25 and 44, living in or visiting Brooklyn, at night time specifically around 23:00 in July. Can we then safely conclude that any man meeting those race and age criterion should avoid at all costs Brooklyn in summer at night time? Additional variables such as the circumstances of the shootings might be necessary to further conclusions. **B) Further analysis for Brooklyn**

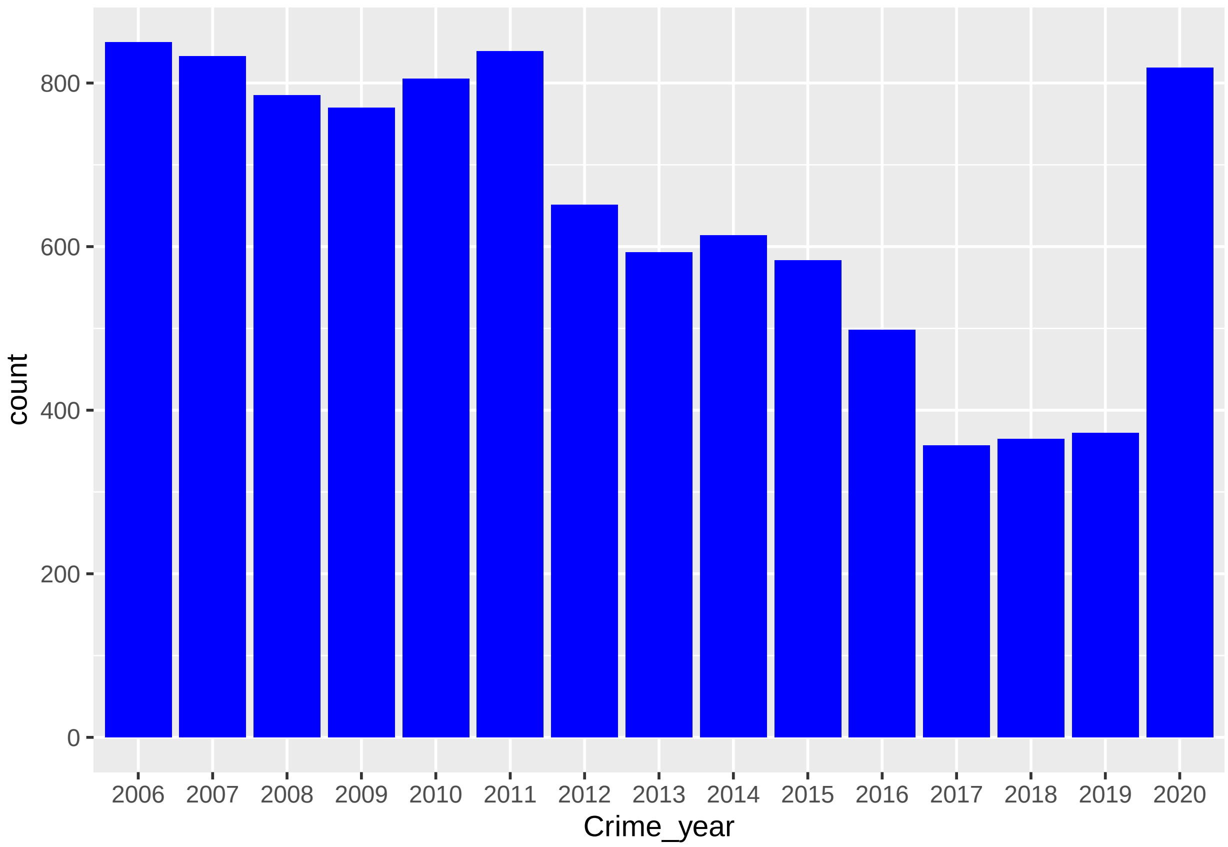
Since Brooklyn won the award of the absolute shooting cases numbers and also was outstanding in the relative shooting cases numbers with regards to population levels, let's do a deeper analysis for this borough.

```
data_Brooklyn = data_collected %>% filter(data_collected$BORO=="BROOKLYN")
head(data_Brooklyn)
```

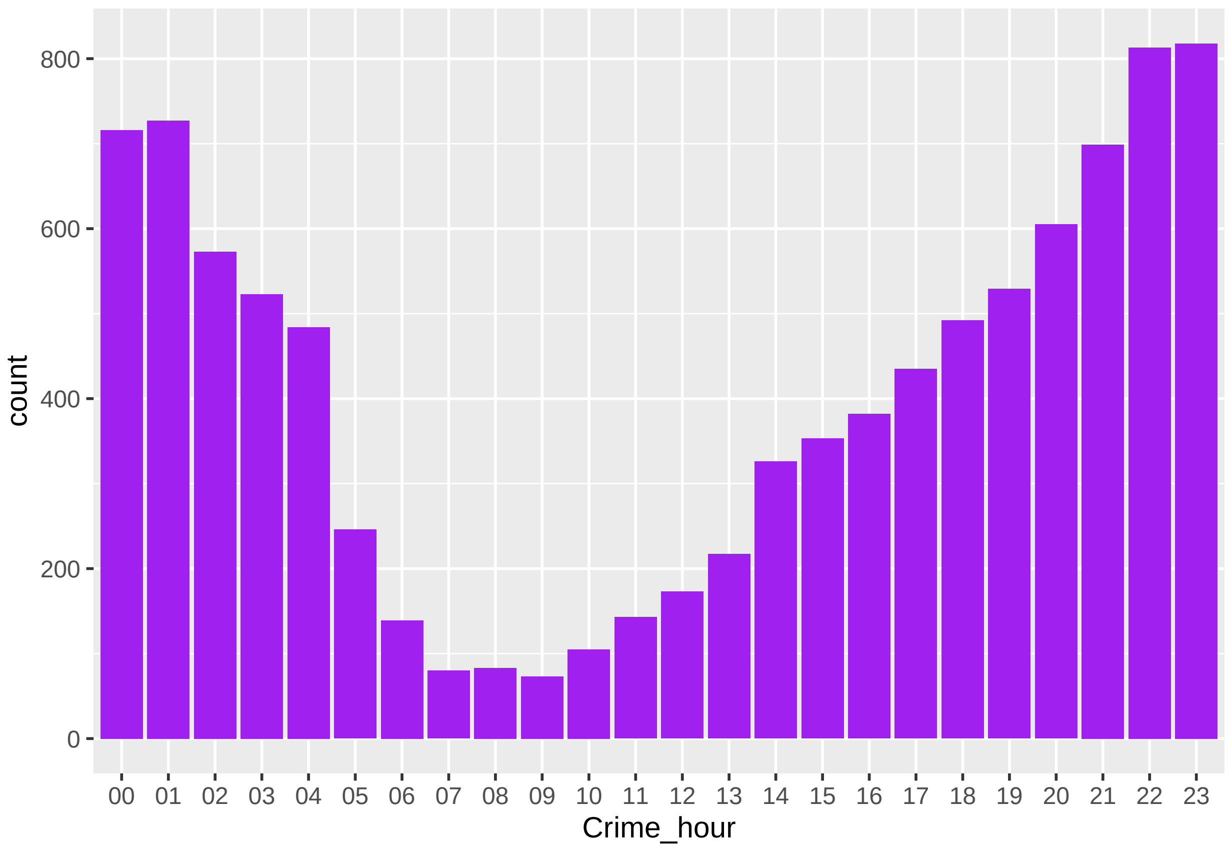
```
## # A tibble: 6 x 10
##   INCIDENT_KEY Crime_month Crime_day Crime_year Crime_hour Crime_min BORO
##          <dbl> <chr>       <chr>     <chr>      <chr>      <chr>     <chr>
## 1    203350417 10          06        2019       01         09        BROOKLYN
## 2     92393427 09          01        2013       04         17        BROOKLYN
## 3     73057167 06          05        2010       21         16        BROOKLYN
## 4    211362213 03          20        2020       21         27        BROOKLYN
## 5     82333894 12          26        2011       03         00        BROOKLYN
## 6    214693508 06          27        2020       00         35        BROOKLYN
## # ... with 3 more variables: VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>
```

1)Brooklyn shooting cases trend from 2006 to 2020
2)Brooklyn shooting cases day trend
3)Brooklyn shooting cases month trend

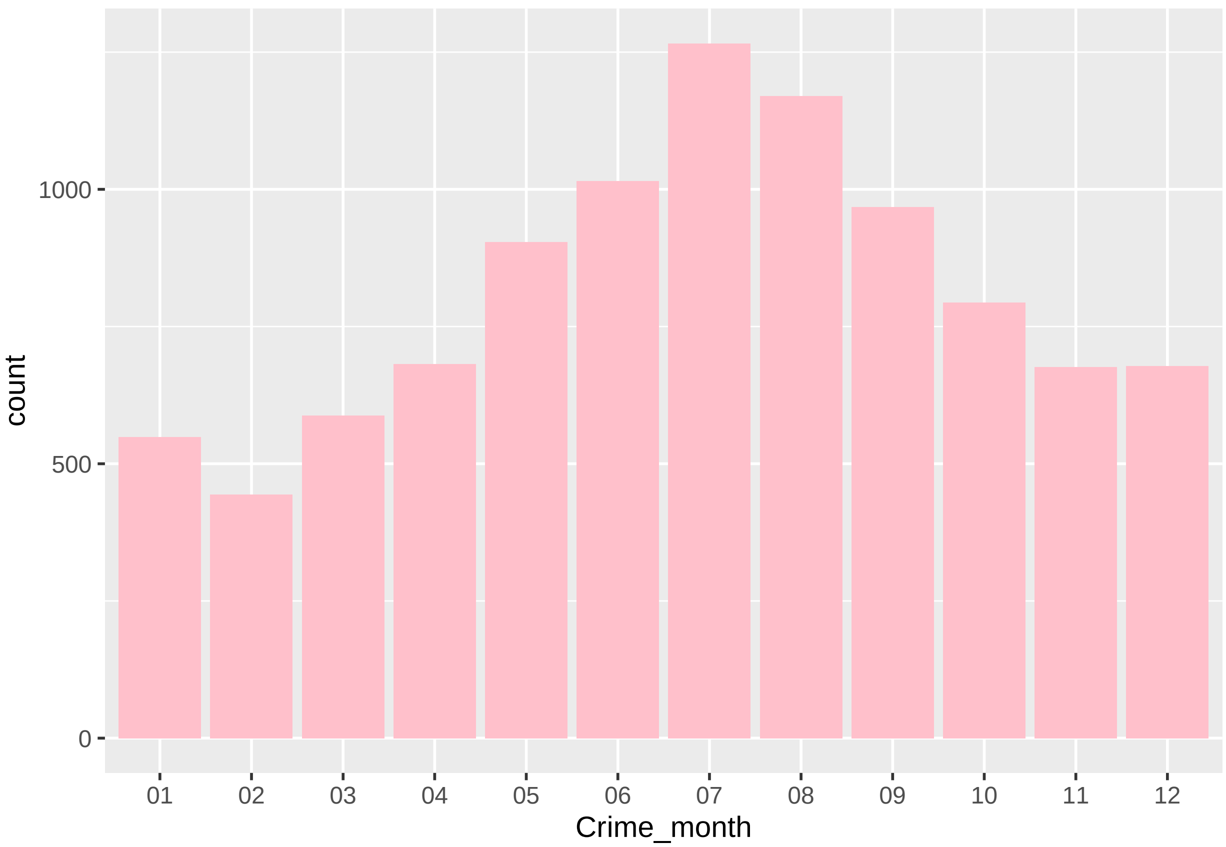
```
ggplot(data_Brooklyn,aes(x=Crime_year))+geom_bar(position="stack", fill="blue")
```



```
ggplot(data_Brooklyn,aes(x=Crime_hour))+geom_bar(position="stack", fill="purple")
```



```
ggplot(data_Brooklyn,aes(x=Crime_month))+geom_bar(position="stack", fill="pink")
```



Conclusion

I chose to focus on the likelihood to be victim of a shooting in New York because I was interested in knowing whether it was a safe place to visit. But the variables here are not sufficient for a thorough analysis. It would be interesting to know the circumstances of the shootings, meaning whether they occurred within drug trafficking or in the middle of the city etc.
We see that the trends for Brooklyn with respect to day, month and year are the same as those for total population. We do not need to run the plots for gender and race, as they will certainly have the same trends as for total New York population. It might be interesting to see if these trends are the same on the country level.