NYC Shooting

R P

2025-04-15

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

Importing and Describing

This is my attempt at importing and describing the shooting project dataset in a reproducible manner.

I am:

- Assigning a name to the URL
- Reading the data in
- Showing the data that was read in
- Providing a summary of that data

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

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): 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.
```

```
shooting_data
```

```
## # A tibble: 29,744 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
```

```
## 1     231974218 08/09/2021 01:06      BRONX    <NA>                        40
## 2     177934247 04/07/2018 19:48      BROOKLYN <NA>                        79
## 3     255028563 12/02/2022 22:57      BRONX    OUTSIDE                     47
## 4      25384540 11/19/2006 01:50      BROOKLYN <NA>                        66
## 5      72616285 05/09/2010 01:58      BRONX    <NA>                        46
## 6      85875439 07/22/2012 21:35      BRONX    <NA>                        42
## 7      79780323 07/12/2011 22:26      BROOKLYN <NA>                        71
## 8      85744504 07/14/2012 23:45      BROOKLYN <NA>                        69
## 9     142324890 04/21/2015 15:36      BROOKLYN <NA>                        75
## 10    152868707 05/07/2016 15:23      BROOKLYN <NA>                        69
## # i 29,734 more rows
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   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>
```

```r
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME            BORO
##  Min.   :  9953245   Length:29744       Length:29744      Length:29744
##  1st Qu.: 67321140   Class :character   Class1:hms        Class :character
##  Median :109291972   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :133850951                      Mode  :numeric
##  3rd Qu.:214741917
##  Max.   :299462478
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:29744       Min.   :  1.00   Min.   :0.0000    Length:29744
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.00   Median :0.0000    Mode  :character
##                     Mean   : 65.23   Mean   :0.3181
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:29744       Mode :logical           Length:29744
##  Class :character   FALSE:23979             Class :character
##  Mode  :character   TRUE :5765              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE          VIC_AGE_GROUP         VIC_SEX
##  Length:29744       Length:29744       Length:29744       Length:29744
##  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:29744       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67
##  Mode  :character   Median :1007826   Median :195506   Median :40.70
```

```
##                    Mean   :1009442   Mean   :208722   Mean   :40.74
##                    3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83
##                    Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                       NA's   :97
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:29744
##  1st Qu.:-73.94   Class :character
##  Median :-73.91   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :97
```

Factors

Next, we need to assess which of my variable are factors, and for ones that are not, determine if they should be.

```r
data.frame(
variable = names(shooting_data),
is_factor = sapply(shooting_data, is.factor),
class = sapply(shooting_data, function(x) class(x)[1]))
```

```
##                                        variable is_factor     class
## INCIDENT_KEY                       INCIDENT_KEY     FALSE   numeric
## OCCUR_DATE                           OCCUR_DATE     FALSE character
## OCCUR_TIME                           OCCUR_TIME     FALSE       hms
## BORO                                       BORO     FALSE character
## LOC_OF_OCCUR_DESC             LOC_OF_OCCUR_DESC     FALSE character
## PRECINCT                               PRECINCT     FALSE   numeric
## JURISDICTION_CODE             JURISDICTION_CODE     FALSE   numeric
## LOC_CLASSFCTN_DESC           LOC_CLASSFCTN_DESC     FALSE character
## LOCATION_DESC                     LOCATION_DESC     FALSE character
## STATISTICAL_MURDER_FLAG STATISTICAL_MURDER_FLAG     FALSE   logical
## PERP_AGE_GROUP                   PERP_AGE_GROUP     FALSE character
## PERP_SEX                               PERP_SEX     FALSE character
## PERP_RACE                             PERP_RACE     FALSE character
## VIC_AGE_GROUP                     VIC_AGE_GROUP     FALSE character
## VIC_SEX                                 VIC_SEX     FALSE character
## VIC_RACE                               VIC_RACE     FALSE character
## X_COORD_CD                           X_COORD_CD     FALSE   numeric
## Y_COORD_CD                           Y_COORD_CD     FALSE   numeric
## Latitude                               Latitude     FALSE   numeric
## Longitude                             Longitude     FALSE   numeric
## Lon_Lat                                 Lon_Lat     FALSE character
```

As you can see, none of the variables are factors. However, some of them should be as long as they are categorical.

Relevant variables

Before we do that, lets look at the entire dataset to see which variables are irrelevant to any sort of analysis.

The below chunk will allow us to view the first several rows in RMD for all variables.

```r
shooting_data %>%
  head() %>%
```

```r
  print(width = Inf)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
## 2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
## 3    255028563 12/02/2022 22:57      BRONX    OUTSIDE                 47
## 4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
## 5     72616285 05/09/2010 01:58      BRONX    <NA>                    46
## 6     85875439 07/22/2012 21:35      BRONX    <NA>                    42
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC
##               <dbl> <chr>              <chr>
## 1                 0 <NA>               <NA>
## 2                 0 <NA>               <NA>
## 3                 0 STREET             GROCERY/BODEGA
## 4                 0 <NA>               PVT HOUSE
## 5                 0 <NA>               MULTI DWELL - APT BUILD
## 6                 2 <NA>               MULTI DWELL - PUBLIC HOUS
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE      VIC_AGE_GROUP
##   <lgl>                   <chr>          <chr>    <chr>          <chr>
## 1 FALSE                   <NA>           <NA>     <NA>           18-24
## 2 TRUE                    25-44          M        WHITE HISPANIC 25-44
## 3 FALSE                   (null)         (null)   (null)         25-44
## 4 TRUE                    UNKNOWN        U        UNKNOWN        18-24
## 5 TRUE                    25-44          M        BLACK          <18
## 6 FALSE                   18-24          M        BLACK          18-24
##   VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
##   <chr>   <chr>         <dbl>      <dbl>    <dbl>     <dbl>
## 1 M       BLACK      1006343     234270      40.8     -73.9
## 2 M       BLACK      1000083.    189065.     40.7     -73.9
## 3 M       BLACK      1020691     257125      40.9     -73.9
## 4 M       BLACK       985107.    173350.     40.6     -74.0
## 5 F       BLACK      1009854.    247503.     40.8     -73.9
## 6 M       BLACK      1011047.    239814.     40.8     -73.9
##   Lon_Lat
##   <chr>
## 1 POINT (-73.92019278899994 40.80967347200004)
## 2 POINT (-73.94291302299996 40.685609672000055)
## 3 POINT (-73.868233 40.872349)
## 4 POINT (-73.99691224999998 40.642489932000046)
## 5 POINT (-73.90746098599993 40.84598358900007)
## 6 POINT (-73.90317908399999 40.82487781900005)
```

It looks like there are some variables we wont need. Let's remove the ones that offer precise geographical location data. We do not need those.

```r
shooting_data_reduced<- shooting_data %>%
select(-c(X_COORD_CD:Lon_Lat))
```

There is a date variable, but the class is classified as a character. We need to change that to a date class. This is why we libraried in Lubridate earlier.

```r
shooting_data_reduced$OCCUR_DATE<- mdy(shooting_data_reduced$OCCUR_DATE)
shooting_data_reduced
```

```
## # A tibble: 29,744 x 16
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <date>     <time>     <chr>    <chr>                <dbl>
##  1    231974218 2021-08-09 01:06      BRONX    <NA>                    40
##  2    177934247 2018-04-07 19:48      BROOKLYN <NA>                    79
##  3    255028563 2022-12-02 22:57      BRONX    OUTSIDE                 47
##  4     25384540 2006-11-19 01:50      BROOKLYN <NA>                    66
##  5     72616285 2010-05-09 01:58      BRONX    <NA>                    46
##  6     85875439 2012-07-22 21:35      BRONX    <NA>                    42
##  7     79780323 2011-07-12 22:26      BROOKLYN <NA>                    71
##  8     85744504 2012-07-14 23:45      BROOKLYN <NA>                    69
##  9    142324890 2015-04-21 15:36      BROOKLYN <NA>                    75
## 10    152868707 2016-05-07 15:23      BROOKLYN <NA>                    69
## # i 29,734 more rows
## # i 10 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   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>
```

We've removed unnecessary columns and ensured the OCCUR_DATE was accurately represented as a date class. Next we need to determine which variable should be treated as factors. None of the variables look like they would be needed for any computational analysis and all look like they are categorical, therefore each variable can be turned into a factor with the exception of Incident Key, Occur Date, and Occur Time.

Adding Variables

One thing I noticed first before making these factors: There is currently no way of using Occur Time as a category. So if we create a new variable using three time periods of the day, the time can be a useful tool in understanding do more shootings occur during certain time periods. Let us create a new variable, separating the times into these four groups:

1. 00:00 - 05:59 = Early Morning
2. 06:00 - 11:59 = Late Morning
3. 12:00 - 17:59 = Afternoon
4. 18:00 - 23:59 = Night

We saw in the earlier assess_factors chunk that Occur_time is in the hms class, and time format. We do not have to do anything else to that column to prepare it. Let's create the new variable next to Occur_Time labeled Time_Block.

```r
shooting_data_reduced$TIME_BLOCK <- case_when(
  hour(shooting_data_reduced$OCCUR_TIME) >= 0  & hour(shooting_data_reduced$OCCUR_TIME) < 6  ~ "Early Mo
  hour(shooting_data_reduced$OCCUR_TIME) >= 6  & hour(shooting_data_reduced$OCCUR_TIME) < 12 ~ "Late Mor
  hour(shooting_data_reduced$OCCUR_TIME) >= 12 & hour(shooting_data_reduced$OCCUR_TIME) < 18 ~ "Afterno
  hour(shooting_data_reduced$OCCUR_TIME) >= 18 & hour(shooting_data_reduced$OCCUR_TIME) <= 23 ~ "Night")

shooting_data_reduced <- shooting_data_reduced %>%
  relocate(TIME_BLOCK, .after = OCCUR_TIME)

shooting_data_reduced
```

```
## # A tibble: 29,744 x 17
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME TIME_BLOCK    BORO     LOC_OF_OCCUR_DESC
##           <dbl> <date>     <time>     <chr>         <chr>    <chr>
##  1    231974218 2021-08-09 01:06      Early Morning BRONX    <NA>
##  2    177934247 2018-04-07 19:48      Night         BROOKLYN <NA>
```

```
## 3     255028563 2022-12-02 22:57      Night         BRONX    OUTSIDE
## 4      25384540 2006-11-19 01:50      Early Morning BROOKLYN <NA>
## 5      72616285 2010-05-09 01:58      Early Morning BRONX    <NA>
## 6      85875439 2012-07-22 21:35      Night         BRONX    <NA>
## 7      79780323 2011-07-12 22:26      Night         BROOKLYN <NA>
## 8      85744504 2012-07-14 23:45      Night         BROOKLYN <NA>
## 9     142324890 2015-04-21 15:36      Afternoon     BROOKLYN <NA>
## 10    152868707 2016-05-07 15:23      Afternoon     BROOKLYN <NA>
## # i 29,734 more rows
## # i 11 more variables: PRECINCT <dbl>, JURISDICTION_CODE <dbl>,
## #   LOC_CLASSFCTN_DESC <chr>, 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>
```

Now that we have an a way to use time of day to categorize that data, lets move on to making the variables factors (except the three mentioned before).

```
shooting_data_reduced$TIME_BLOCK <- as.factor(shooting_data_reduced$TIME_BLOCK)
shooting_data_reduced$BORO <- as.factor(shooting_data_reduced$BORO)
shooting_data_reduced$LOC_OF_OCCUR_DESC <- as.factor(shooting_data_reduced$LOC_OF_OCCUR_DESC)
shooting_data_reduced$PRECINCT <- as.factor(shooting_data_reduced$PRECINCT)
shooting_data_reduced$JURISDICTION_CODE <- as.factor(shooting_data_reduced$JURISDICTION_CODE)
shooting_data_reduced$LOC_CLASSFCTN_DESC <- as.factor(shooting_data_reduced$LOC_CLASSFCTN_DESC)
shooting_data_reduced$LOCATION_DESC <- as.factor(shooting_data_reduced$LOCATION_DESC)
shooting_data_reduced$STATISTICAL_MURDER_FLAG <- as.factor(shooting_data_reduced$STATISTICAL_MURDER_FLAG
shooting_data_reduced$PERP_AGE_GROUP <- as.factor(shooting_data_reduced$PERP_AGE_GROUP)
shooting_data_reduced$PERP_SEX <- as.factor(shooting_data_reduced$PERP_SEX)
shooting_data_reduced$PERP_RACE <- as.factor(shooting_data_reduced$PERP_RACE)
shooting_data_reduced$VIC_AGE_GROUP <- as.factor(shooting_data_reduced$VIC_AGE_GROUP)
shooting_data_reduced$VIC_SEX <- as.factor(shooting_data_reduced$VIC_SEX)
shooting_data_reduced$VIC_RACE <- as.factor(shooting_data_reduced$VIC_RACE)

shooting_data_reduced
```

```
## # A tibble: 29,744 x 17
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME TIME_BLOCK    BORO     LOC_OF_OCCUR_DESC
##           <dbl> <date>     <time>     <fct>         <fct>    <fct>
## 1     231974218 2021-08-09 01:06      Early Morning BRONX    <NA>
## 2     177934247 2018-04-07 19:48      Night         BROOKLYN <NA>
## 3     255028563 2022-12-02 22:57      Night         BRONX    OUTSIDE
## 4      25384540 2006-11-19 01:50      Early Morning BROOKLYN <NA>
## 5      72616285 2010-05-09 01:58      Early Morning BRONX    <NA>
## 6      85875439 2012-07-22 21:35      Night         BRONX    <NA>
## 7      79780323 2011-07-12 22:26      Night         BROOKLYN <NA>
## 8      85744504 2012-07-14 23:45      Night         BROOKLYN <NA>
## 9     142324890 2015-04-21 15:36      Afternoon     BROOKLYN <NA>
## 10    152868707 2016-05-07 15:23      Afternoon     BROOKLYN <NA>
## # i 29,734 more rows
## # i 11 more variables: PRECINCT <fct>, JURISDICTION_CODE <fct>,
## #   LOC_CLASSFCTN_DESC <fct>, LOCATION_DESC <fct>,
## #   STATISTICAL_MURDER_FLAG <fct>, PERP_AGE_GROUP <fct>, PERP_SEX <fct>,
## #   PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>, VIC_RACE <fct>
```

Success. We can see the variables have turned to fct.

NA's

Next we need to account for any missing data. Lets find out which variables have NAs in their set, and how many.

```r
colSums(is.na(shooting_data_reduced))
```

```
##             INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                        0                       0                       0
##               TIME_BLOCK                    BORO       LOC_OF_OCCUR_DESC
##                        0                       0                   25596
##                 PRECINCT       JURISDICTION_CODE      LOC_CLASSFCTN_DESC
##                        0                       2                   25596
##            LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                    14977                       0                    9344
##                 PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                     9310                    9310                       0
##                  VIC_SEX                VIC_RACE
##                        0                       0
```

Some of these variables will be very useful as they give complete/near complete data for all 28K+ rows. However, there are some variables with significant amounts of missing data that will make those variables unreliable in any meaningful analysis. I'm inclined to keep the variables with the NAs, but it will be unlikely I will use the ones with high amounts (i.e. LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, and LOCATION_DESC). Similiarly, PERP_SEX, PERP_RACE, PERP_AGE_GROUP have approx 30% missing data, which also renders them unreliable, but we might find some use for them.

Additional Variable

There is one more variable I would like to add. I want to include Month_Occur and Year. I hypothesize that warmer months will show an increase in shootings. Adding this variable will allow us to determine that.

```r
shooting_data_reduced$MONTH_OCCUR <- format(shooting_data_reduced$OCCUR_DATE, "%b")
shooting_data_reduced$YEAR <- format(shooting_data_reduced$OCCUR_DATE, "%Y")
shooting_data_reduced$MONTH_OCCUR <- factor(
shooting_data_reduced$MONTH_OCCUR,
levels = month.abb, ordered = TRUE)
shooting_data_reduced <- shooting_data_reduced %>%
relocate(MONTH_OCCUR, .after = OCCUR_DATE) %>%
relocate(YEAR, .after = MONTH_OCCUR)
shooting_data_reduced
```

```
## # A tibble: 29,744 x 19
##    INCIDENT_KEY OCCUR_DATE MONTH_OCCUR YEAR  OCCUR_TIME TIME_BLOCK    BORO
##           <dbl> <date>     <ord>       <chr> <time>     <fct>         <fct>
##  1    231974218 2021-08-09 Aug         2021  01:06      Early Morning BRONX
##  2    177934247 2018-04-07 Apr         2018  19:48      Night         BROOKLYN
##  3    255028563 2022-12-02 Dec         2022  22:57      Night         BRONX
##  4     25384540 2006-11-19 Nov         2006  01:50      Early Morning BROOKLYN
##  5     72616285 2010-05-09 May         2010  01:58      Early Morning BRONX
##  6     85875439 2012-07-22 Jul         2012  21:35      Night         BRONX
##  7     79780323 2011-07-12 Jul         2011  22:26      Night         BROOKLYN
##  8     85744504 2012-07-14 Jul         2012  23:45      Night         BROOKLYN
##  9    142324890 2015-04-21 Apr         2015  15:36      Afternoon     BROOKLYN
## 10    152868707 2016-05-07 May         2016  15:23      Afternoon     BROOKLYN
## # i 29,734 more rows
## # i 12 more variables: LOC_OF_OCCUR_DESC <fct>, PRECINCT <fct>,
```

```
## #   JURISDICTION_CODE <fct>, LOC_CLASSFCTN_DESC <fct>, LOCATION_DESC <fct>,
## #   STATISTICAL_MURDER_FLAG <fct>, PERP_AGE_GROUP <fct>, PERP_SEX <fct>,
## #   PERP_RACE <fct>, VIC_AGE_GROUP <fct>, VIC_SEX <fct>, VIC_RACE <fct>
```

Success. We now have a Month and Year variable.

Questions

Let us consider some questions we might want answers to:

1. Do shootings tend to increase or decrease in certain months?
2. Are there more shootings in certain time blocks/Boro combinations than others?

Analysis

1. Do shootings tend to increase or decrease in certain months?

My hypothesis for this question would be that, since this is a city in the Northeast part of the US that experiences all four seasons, there would be more shootings during warmer months than during colder ones. This would be due to the very nature of more people (both perps and victims) would be out and about during the summer, and not have the cold factor keeping them indoors.

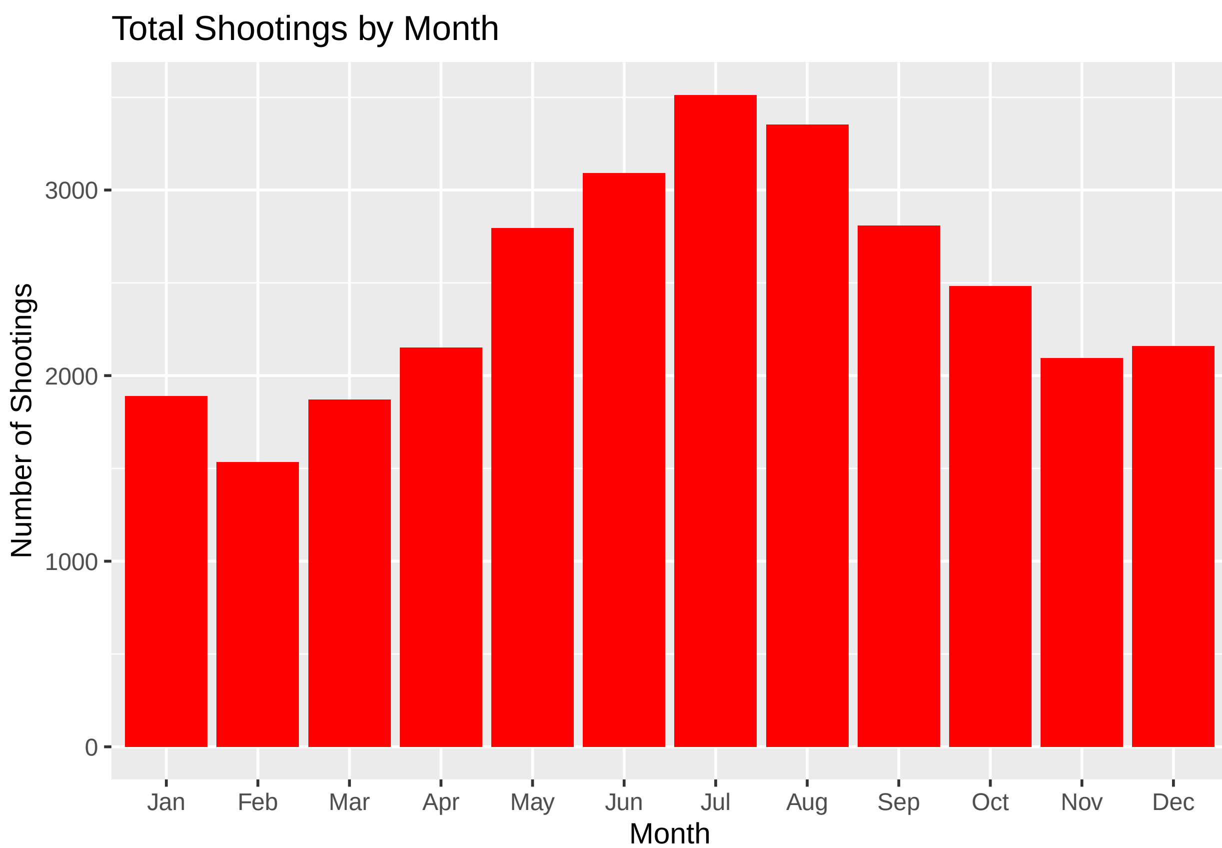
We can assess this hypothesis with a simple table and look at the shootings per month:

```r
shootings_by_month <- shooting_data_reduced %>%
count(MONTH_OCCUR)
shootings_by_month
```

```
## # A tibble: 12 x 2
##    MONTH_OCCUR     n
##    <ord>       <int>
##  1 Jan          1891
##  2 Feb          1533
##  3 Mar          1872
##  4 Apr          2150
##  5 May          2795
##  6 Jun          3091
##  7 Jul          3513
##  8 Aug          3352
##  9 Sep          2808
## 10 Oct          2483
## 11 Nov          2096
## 12 Dec          2160
```

Looking through each month in the table, you can certainly tell that there is a difference between some seasons, but it might be better with a histogram:

```r
ggplot(shooting_data_reduced, aes(x= MONTH_OCCUR)) +
  geom_bar(fill = "red") +
  labs(title = "Total Shootings by Month", x = "Month", y = "Number of Shootings")
```



This histogram confirms the hypothesis that there are more shootings in warmer months.

We can go further with this. Let's model this out to get a deeper understanding.

```
# Create a monthly summary dataset
monthly_shootings <- shooting_data_reduced %>%
  count(YEAR, MONTH_OCCUR)

# Ensure MONTH_OCCUR is a factor in Jan-Dec order
monthly_shootings$MONTH_OCCUR <- factor(
  monthly_shootings$MONTH_OCCUR,
  levels = month.abb,
  labels = month.abb,
  ordered = FALSE
)

# Set dummy coding (default base is Jan)
contrasts(monthly_shootings$MONTH_OCCUR) <- contr.treatment(12, base = 1)

# Fit linear model
month_model <- lm(n ~ MONTH_OCCUR, data = monthly_shootings)

# Show model summary
summary(month_model)
```

```
##
## Call:
```

```
## lm(formula = n ~ MONTH_OCCUR, data = monthly_shootings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.421 -31.250  -0.684  27.013 140.105
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     99.526      9.383  10.607  < 2e-16 ***
## MONTH_OCCUR2   -18.842     13.270  -1.420 0.157072
## MONTH_OCCUR3    -1.000     13.270  -0.075 0.939999
## MONTH_OCCUR4    13.632     13.270   1.027 0.305446
## MONTH_OCCUR5    47.579     13.270   3.586 0.000416 ***
## MONTH_OCCUR6    63.158     13.270   4.760 3.55e-06 ***
## MONTH_OCCUR7    85.368     13.270   6.433 7.92e-10 ***
## MONTH_OCCUR8    76.895     13.270   5.795 2.40e-08 ***
## MONTH_OCCUR9    48.263     13.270   3.637 0.000345 ***
## MONTH_OCCUR10   31.158     13.270   2.348 0.019777 *
## MONTH_OCCUR11   10.790     13.270   0.813 0.417065
## MONTH_OCCUR12   14.158     13.270   1.067 0.287196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.9 on 216 degrees of freedom
## Multiple R-squared:  0.3895, Adjusted R-squared:  0.3584
## F-statistic: 12.53 on 11 and 216 DF,  p-value: < 2.2e-16
```

What does this model analysis mean?

Intercept represents January (this can be altered) average shootings over the dataset. Every subsequent Month_Occur is the next month (i.e. 2=Feb, 3=Mar, etc).

The estimate is how many shootings, on average, can you expect in that month in relation to January. February is 20.3 less. March is almost flat. July is 87.8 more.

The P values are important. The Months with the asterisks to the right have P values less than .05. These months have significantly more shootings than January.

2. Are there more shootings in certain time blocks/Boro combinations than others?

I want to determine whether or not there are certain time block (based on the aforementioned timeframes) / Boro combinations that stand out as outliers compared to others. I hypothesize that there would likely be more shootings in Boros that have a higher rate of poverty, and during either the Night or Early Morning time blocks.

We can approach this question in the same way as question 1. Let's create a table.

```
time_boro_counts <- shooting_data_reduced %>%
     count(TIME_BLOCK, BORO)

time_boro_counts
```

```
## # A tibble: 20 x 3
##    TIME_BLOCK    BORO              n
##    <fct>         <fct>         <int>
##  1 Afternoon     BRONX          1556
##  2 Afternoon     BROOKLYN       2338
```

```
##  3 Afternoon     MANHATTAN       620
##  4 Afternoon     QUEENS          779
##  5 Afternoon     STATEN ISLAND   146
##  6 Early Morning BRONX          3075
##  7 Early Morning BROOKLYN       3804
##  8 Early Morning MANHATTAN      1511
##  9 Early Morning QUEENS         1804
## 10 Early Morning STATEN ISLAND   317
## 11 Late Morning  BRONX           531
## 12 Late Morning  BROOKLYN        806
## 13 Late Morning  MANHATTAN       250
## 14 Late Morning  QUEENS          314
## 15 Late Morning  STATEN ISLAND    54
## 16 Night         BRONX          3672
## 17 Night         BROOKLYN       4737
## 18 Night         MANHATTAN      1596
## 19 Night         QUEENS         1529
## 20 Night         STATEN ISLAND   305
```

Based on this table, it's clear to see that the Night time block and Brooklyn carries the most shootings, but lets find out the subtotals.

```r
#This shows number of shootings by time block
shooting_data_reduced %>%
     count(TIME_BLOCK) %>%
     arrange(desc(n))
```

```
## # A tibble: 4 x 2
##   TIME_BLOCK        n
##   <fct>         <int>
## 1 Night         11839
## 2 Early Morning 10511
## 3 Afternoon      5439
## 4 Late Morning   1955
```

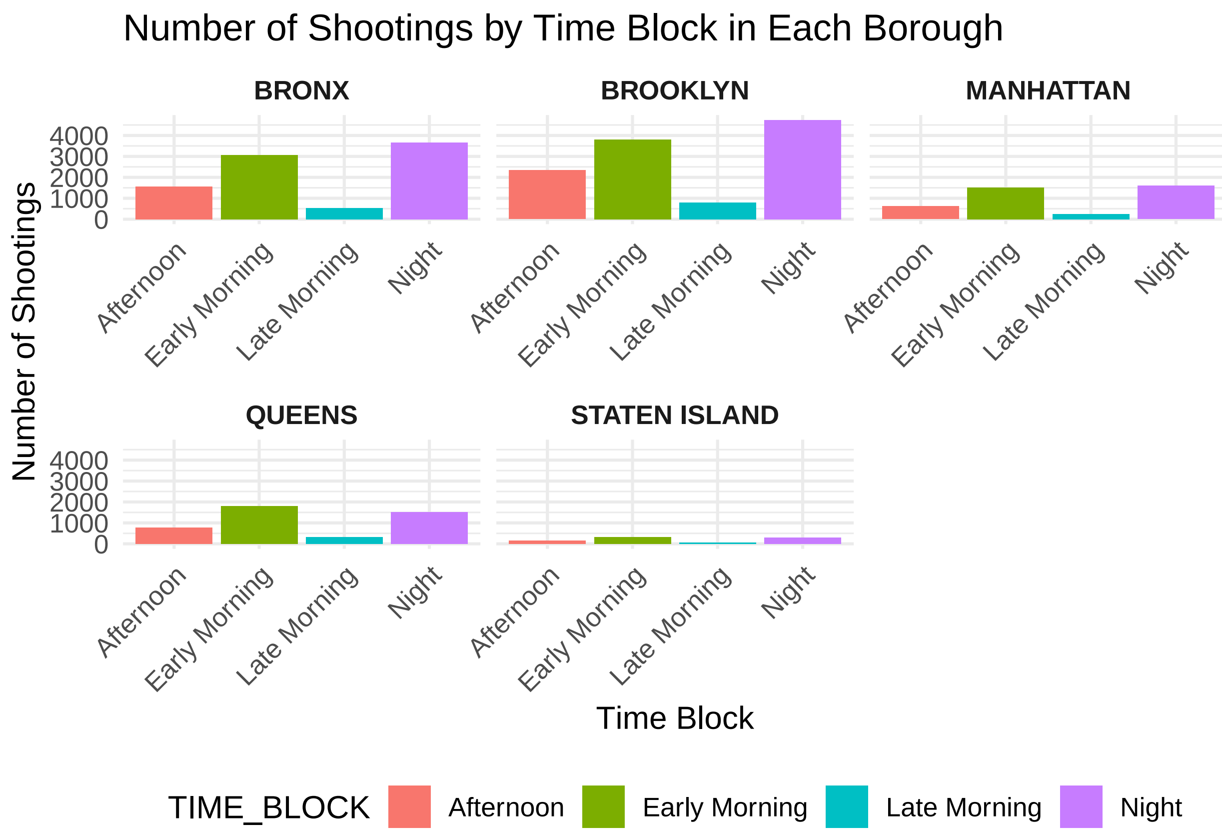
```r
#This shows number of shootings by boro
shooting_data_reduced %>%
     count(BORO) %>%
     arrange(desc(n))
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <fct>         <int>
## 1 BROOKLYN      11685
## 2 BRONX          8834
## 3 QUEENS         4426
## 4 MANHATTAN      3977
## 5 STATEN ISLAND   822
```

These tables are helpful, but a histogram might be better to show the difference.

```r
ggplot(shooting_data_reduced, aes(x = TIME_BLOCK, fill = TIME_BLOCK)) +
  geom_bar() +
  facet_wrap(~ BORO, scales = "free_x") +
  labs(title = "Number of Shootings by Time Block in Each Borough",
    x = "Time Block", y = "Number of Shootings") +
  theme_minimal(base_size = 12) +
```

```r
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    strip.text = element_text(face = "bold"),
    legend.position = "bottom")
```



These histograms show by boro, which time blocks have the most shootings occur.

The hypothesis was that the most shootings would occur during dark hours, and likely in the boros with highest poverty. A cursory review of the website: https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,richmondcountynewyork,bronxcountynewyork,newyorkcountynewyork,kingscountynewyork,queenscountynewyork/PST045223 shows Bronx and Brooklyn with the highest poverty levels, followed by Manahattan and Queens, and lastly Staten Island. This shows there is a correlation between poverty and shooting counts.

Conclusion and Biases

The data shown in this dataset is not much different than similar datasets I have seen in the past regarding crime and urban environments. Having lived in an urban environment my entire life, I suspected that my environment was not much different than NYC. In mine, I knew that the warm weather brought about much more seasonal violent crime, particularly in financially disaffected areas. Hence, the hypotheses that I made. Turns out, NYC followed the same trend as my own very large home city.

With regard to avoiding bias, I stuck to questions that had complete data to back up an answer. I did my absolute best to leave out any analysis or hypothesis that was culturally or socially sensitive (i.e. racial analysis), particularly because I know my R expertise is minimal, and I would not be able to conduct further analysis that would need to stand up to increased scrutiny, due to the sensitivity of the topic, and I did not want to risk making unsupported or overly simplistic conclusions.