NYPD Data Project

2023-11-22

Introduction

This is for the MSDS at CU assignment in the Data Science as a Field course. I left all code displayed to help illustrate what I did and to demonstrate my thought process, even if the document would be aesthetically better off otherwise.

Primary Question

What trends exist between victims who are male and boroughs/precincts?

Project Setup

After downloading the data from https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic, it was imported with:

```
raw_shooting_data = read_csv("./NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

The raw data shown as above. Note that this command assumes the CSV is in the same folder as the Rmd script.

Summary of Data

As we can see in the summary, the data contains information about perpetrators and victims, such as age, sex, and race. It also contains data as to when and where the shooting occurred. The data contains some incomplete columns, such as the location description.

```r
summary(raw_shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     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:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA’s   :10
##    Longitude       Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA’s   :10
```

Data Cleanup

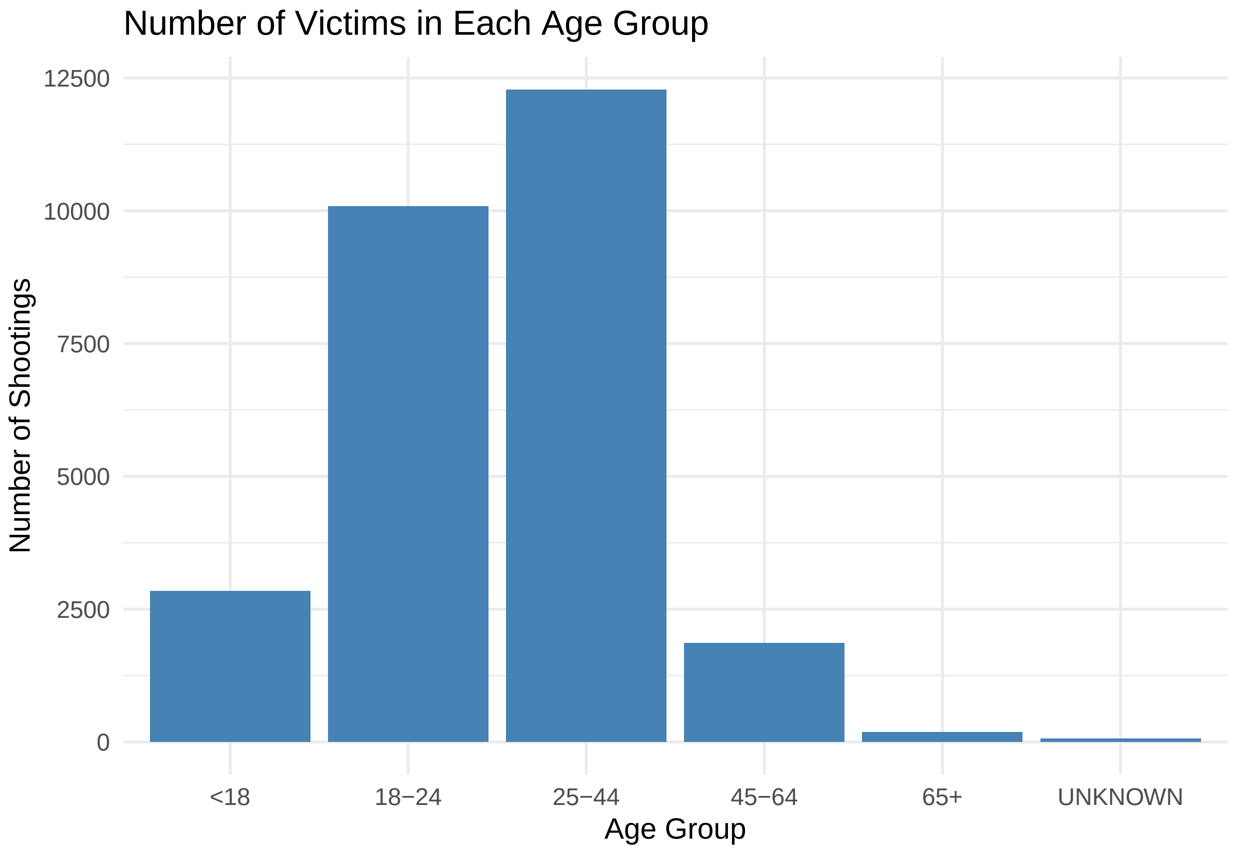
Since the victim age group “1022” is ambiguous and few in number, it will be excluded from this report.

```
age_group_data_filtered = raw_shooting_data %>%
                    filter(VIC_AGE_GROUP != "1022")
```

Notable Visualizations

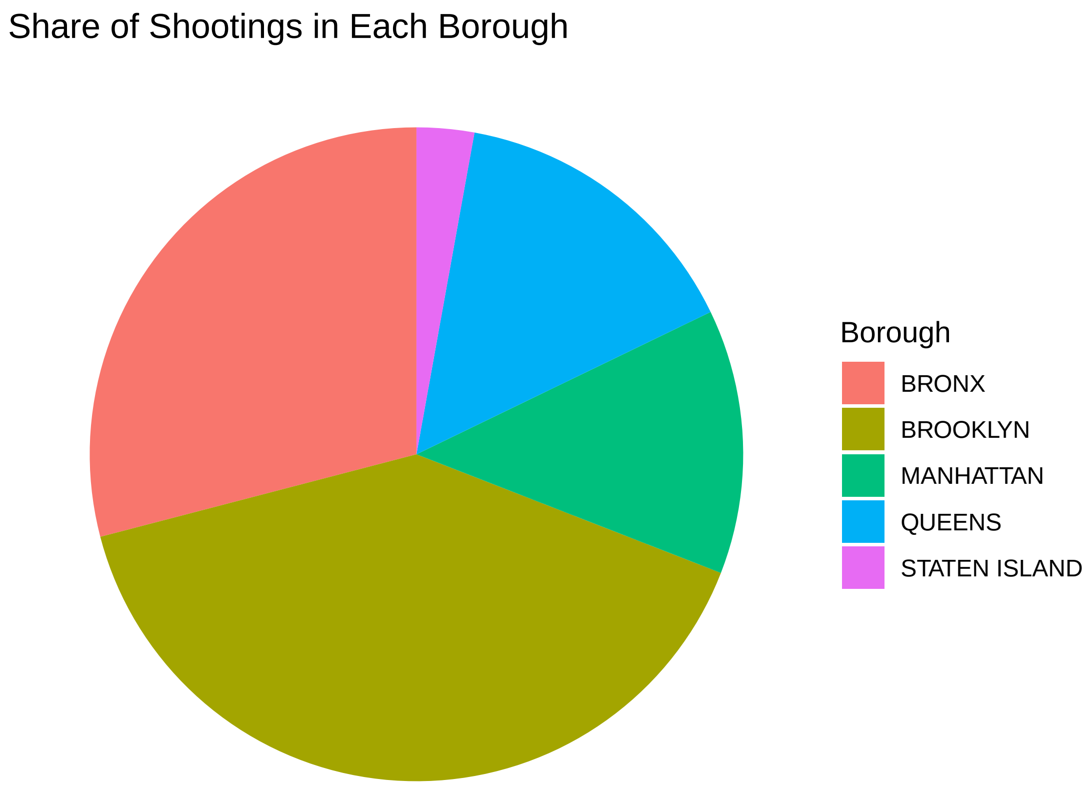
```
age_group_counts <- age_group_data_filtered %>%
                    group_by(VIC_AGE_GROUP) %>%
                    summarise(Count = n())

ggplot(age_group_counts, aes(x = VIC_AGE_GROUP, y = Count)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    theme_minimal() +
    labs(title = "Number of Victims in Each Age Group",
         x = "Age Group",
         y = "Number of Shootings")
```



```
burrow_counts <- raw_shooting_data %>%
                    group_by(BORO) %>%
                    summarise(Count = n())

ggplot(burrow_counts, aes(x = "", y = Count, fill = BORO)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar(theta = "y") +
    theme_void() +
    labs(title = "Share of Shootings in Each Borough",
         fill = "Borough")
```



These charts demonstrate some different queries, as well as some different colors and themes that could be applied.

Data Analysis

Next, we will analyze shootings by precinct and sex (how many shootings of each sex (victims), in each precinct):

```
precinct_sex_counts <- raw_shooting_data %>%
                       group_by(PRECINCT, VIC_SEX) %>%
                       summarise(Count = n(), .groups = "drop")

# Now format it cleaner
better_format <- precinct_sex_counts %>%
                pivot_wider(names_from = VIC_SEX, values_from = Count, values_fill = list(Count = 0))
```

```r
options(tibble.print_max = Inf) # Since we have 77 rows, we need to
print(better_format)
```

```
## # A tibble: 77 x 4
##    PRECINCT     F     M     U
##       <dbl> <int> <int> <int>
##  1        1     4    21     0
##  2        5     6    52     0
##  3        6     2    26     0
##  4        7    16    93     0
##  5        9     8   101     0
##  6       10    12    61     0
##  7       13    16    44     0
##  8       14    11    45     0
##  9       17     1     9     0
## 10       18     3    31     0
## 11       19     2    18     0
## 12       20     3    37     0
## 13       22     0     1     0
## 14       23    43   442     2
## 15       24     4   101     0
## 16       25    47   414     0
## 17       26     8   141     0
## 18       28    35   308     0
## 19       30    12   217     0
## 20       32    68   566     0
## 21       33    19   206     0
## 22       34    32   284     0
## 23       40    64   844     0
## 24       41    49   445     0
## 25       42    65   785     0
## 26       43    74   684     0
## 27       44    92   927     1
## 28       45    10   172     0
## 29       46    86   809     0
## 30       47    92   861     0
## 31       48    73   712     2
## 32       49    32   321     0
## 33       50    13   141     0
## 34       52    50   533     0
## 35       60    49   323     0
## 36       61    18   135     0
## 37       62     7    63     0
## 38       63    27   255     0
## 39       66     2    44     0
## 40       67   118  1098     0
## 41       68     2    30     0
## 42       69    38   426     2
## 43       70    58   400     1
## 44       71    60   519     0
## 45       72     8   101     0
## 46       73   161  1289     2
## 47       75   140  1417     0
## 48       76     8    159     0
## 49       77    75    720     0
## 50       78     3     58     1
## 51       79   103    909     0
## 52       81    66    733     0
## 53       83    38    462     0
## 54       84     9    115     0
## 55       88    22    258     0
## 56       90    37    278     0
## 57       94     8     78     0
## 58      100    18    152     0
## 59      101    43    446     0
## 60      102    25    185     0
## 61      103    55    538     0
## 62      104    19     83     0
## 63      105    40    439     0
## 64      106    22    202     0
## 65      107    11     90     0
## 66      108     7     60     0
## 67      109    14    101     0
## 68      110    17    143     0
## 69      111     1     10     0
## 70      112     5     18     0
## 71      113    83    719     0
## 72      114    41    328     0
## 73      115    14    165     0
## 74      120    68    504     0
## 75      121    13     99     0
## 76      122     5     56     0
## 77      123     5     26     0
```

```
options(tibble.print_max = 10) # Setting the limit back to default... just in case.
```

. . . and yes, that would've been better as a visualization. It is clear that some precincts have less shootings than others, and males are the victim more often in general. Some precincts, like precinct 22, have barely any shootings. Others have thousands.

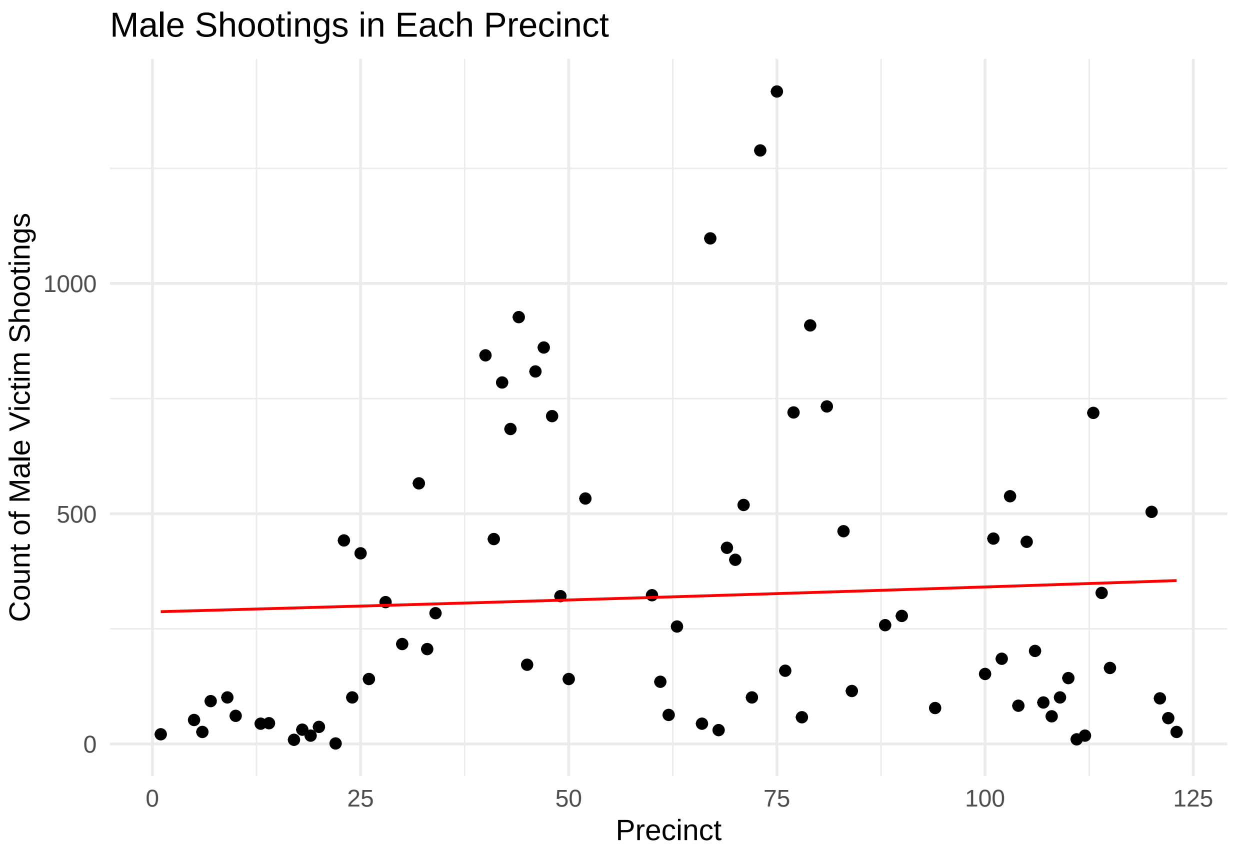
Notable Model

Inspired by the last analysis, we will build a model. For simplicity, we are going to look just at male victim shootings by precinct, and drop females out of the model. We will be looking to prove (fictional) claims that higher number precincts are more dangerous.

```
male_shootings <- raw_shooting_data %>%
                filter(VIC_SEX == "M") %>%
                group_by(PRECINCT) %>%
                summarise(MaleShootings = n(), .groups = "drop")

model <- glm(MaleShootings ~ PRECINCT, data = male_shootings, family = poisson)
male_shootings$PredictedCounts <- predict(model, type = "response")

ggplot(male_shootings, aes(x = PRECINCT, y = MaleShootings)) +
    geom_point() +
    geom_line(aes(y = PredictedCounts), color = "red") +
    theme_minimal() +
    labs(title = "Male Shootings in Each Precinct",
         x = "Precinct",
         y = "Count of Male Victim Shootings")
```



In this case, our model actually suggests that as your precinct number goes up, so do your shootings (slightly). This is, likely not true in reality, and a great case of how poor data analysis can lead to false claims. Truthfully, since the slope is pretty linear, I'd say this actually suggests that there is no trend between precinct number and number of shootings.

Potential Bias in the Data

I will explain one bias I suspect is in the data, and then a bias I think I may have. A bias I think might be in the data is that police officers may mis-report shooters or victims race. They may do this intentionally if they are racially biased, and they may do this by accident if they do not get a good look at the shooter or victim (e.g. the shooter escapes before the police can get a look at who they are). A bias I have is that I am expecting police to file a report without due diligence, and therefore expecting the data to be inaccurate in some way.

Conclusion

This project went over some data cleaning, visualizations, analysis and a model construction. It also explored bias, and how data models aren't always best taken at face value.