NYPD Shooting Occurences Report

D. Fritts

2025-08-30

Data Load and Setup

```
# Loading in necessary libraries
library(tidyverse)
```

```
# This imports the nypd shooting incident data
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd_raw <- read_csv(url)
```

Introduction

This report will analyze the following data:

1. Yearly rates of shooting occurrences in the Boroughs of New York.
2. The Victim and Perpetrator Gender Dynamics of shooting occurrences in New York.

In the hopes of getting some insight into the following questions:

1. Have shootings in New York been increasing or decreasing?
2. What do the gender dynamics of shooting occurrences look like?

Summary of the raw NYPD Shooting Occurrence Data

The following is a summary of the raw NYPD Shooting Occurrence dataset used in this analysis. This dataset consists of historical reports of shooting incidents that occured in New York. These include various details about the incident such as when and where it occurred, the race, gender and age group of the perpetrator and victim, wether it was a murder or not, and what police force was responsible for the case. As you can see, most of the data is in character format. We will not be using most of the columns in this analysis.

```
summary(nypd_raw)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Length:29744       Min.   :00:00:00.000000
##  1st Qu.: 67321140   Class :character   1st Qu.:03:30:45.000000
##  Median :109291972   Mode  :character   Median :15:15:00.000000
##  Mean   :133850951                      Mean   :12:46:10.874798
##  3rd Qu.:214741917                      3rd Qu.:20:44:00.000000  
##  Max.   :299462478                      Max.   :23:59:00.000000  
##                                                                   
##      BORO           LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  Length:29744       Length:29744       Min.   :  1.00   Min.   :0.0000   
##  Class :character   Class :character   1st Qu.: 44.00   1st Qu.:0.0000   
##  Mode  :character   Mode  :character   Median : 67.00   Median :0.0000   
##                                        Mean   : 65.23   Mean   :0.3181   
##                                        3rd Qu.: 81.00   3rd Qu.:0.0000   
##                                        Max.   :123.00   Max.   :2.0000   
##                                                         NA's   :2        
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:29744       Length:29744       Mode :logical          
##  Class :character   Class :character   FALSE:23979            
##  Mode  :character   Mode  :character   TRUE :5765             
##                                                               
##                                                               
##                                                               
##                                                               
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP     
##  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_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD    
##  Length:29744       Length:29744       Min.   : 914928   Min.   :125757  
##  Class :character   Class :character   1st Qu.:1000094   1st Qu.:183042  
##  Mode  :character   Mode  :character   Median :1007826   Median :195506  
##                                        Mean   :1009442   Mean   :208722  
##                                        3rd Qu.:1016739   3rd Qu.:239980  
##                                        Max.   :1066815   Max.   :271128  
##                                                                          
##     Latitude       Longitude        Lon_Lat         
##  Min.   :40.51   Min.   :-74.25   Length:29744      
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character  
##  Median :40.70   Median :-73.91   Mode  :character  
##  Mean   :40.74   Mean   :-73.91                     
##  3rd Qu.:40.83   3rd Qu.:-73.88                     
##  Max.   :40.91   Max.   :-73.70                     
##  NA's   :97      NA's   :97
```

Initial Data Preparation

Here we filter out any unused columns, keeping only the following:

1. OCCUR_DATE: What day the shooting occurred on.
2. BORO: The Borough the shooting occurred in.
3. PERP_SEX: The shooter's gender.
4. VIC_SEX: The victim's gender.

For transformations, OCCUR_DATE's type is changed to the date type and BORO, PERP_SEX and VIC_SEX's types are changed to the factor type.

```
# Here I have filtered down the data, changed OCCUR_DATE to date type, and the other variables BORO, PE
nypd_filtered <- nypd_raw %>%
  select(-c(OCCUR_TIME, INCIDENT_KEY,
            LOCATION_DESC, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC,
            X_COORD_CD, Y_COORD_CD,
            Latitude, Longitude, Lon_Lat,
            PRECINCT, JURISDICTION_CODE, STATISTICAL_MURDER_FLAG,
            PERP_AGE_GROUP,PERP_RACE,
            VIC_AGE_GROUP, VIC_RACE)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         BORO = as.factor(BORO),
         PERP_SEX = as.factor(PERP_SEX),
         VIC_SEX = as.factor(VIC_SEX),
         )


summary(nypd_filtered)
```

```
##    OCCUR_DATE                     BORO         PERP_SEX     VIC_SEX
##  Min.   :2006-01-01   BRONX        : 8834   (null): 1628   F: 2891
##  1st Qu.:2009-10-29   BROOKLYN     :11685   F     :  461   M:26841
##  Median :2014-03-25   MANHATTAN    : 3977   M     :16845   U:   12
##  Mean   :2014-10-31   QUEENS       : 4426   U     : 1500
##  3rd Qu.:2020-06-29   STATEN ISLAND:  822   NA's  : 9310
##  Max.   :2024-12-31
```

As you can see, the only column which we have NA's or (null) values in is PERP_SEX. For our analysis, we will exclude any rows which contain NA or (null) values.

Analysis 1: Shooting Occurrences by Year and Borough

Here, we analyze the number of shooting occurrences over time for each Borough.

Borough Data Preparations and Visualization

Here we create two derived columns:

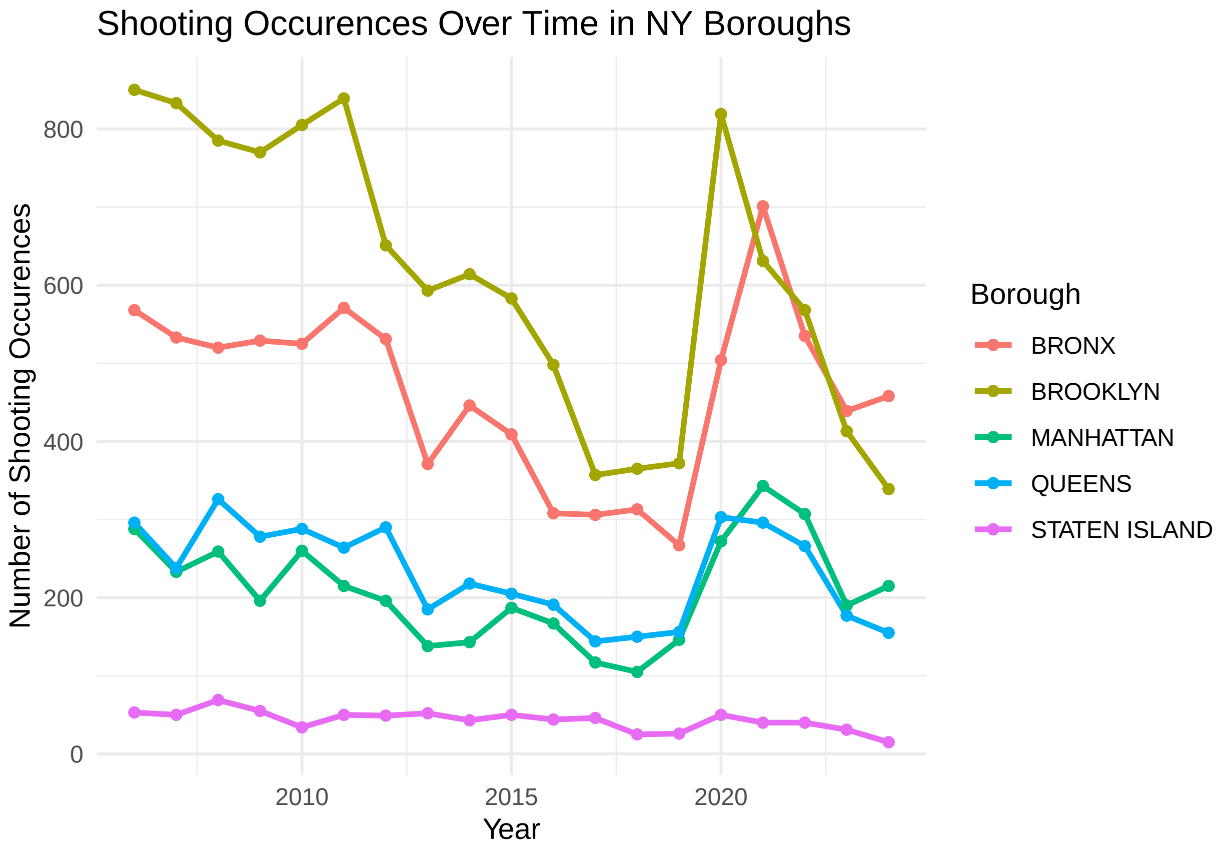
1. OCCUR_YEAR, which extracts the year element from the OCCUR_DATE column
2. NUM_OCCURENCES, which counts the number of shooting per Borough per Year

The reason for a yearly analysis and not a monthly or daily analysis is to make the overall trends clearer and reduce noise in the graph.

```
# Here we will analyze how many shooting occurrences there were in each Borough for each year
boro_analysis <- nypd_filtered %>%
  mutate(OCCUR_YEAR = year(OCCUR_DATE)) %>%
  group_by(BORO, OCCUR_YEAR) %>%
  summarise(NUM_OCCURENCES = n(), .groups="drop") %>%
  ungroup()

summary(boro_analysis)
```

```
##             BORO      OCCUR_YEAR   NUM_OCCURENCES
##  BRONX        :19   Min.   :2006   Min.   : 15.0
##  BROOKLYN     :19   1st Qu.:2010   1st Qu.:145.0
##  MANHATTAN    :19   Median :2015   Median :272.0
##  QUEENS       :19   Mean   :2015   Mean   :313.1
##  STATEN ISLAND:19   3rd Qu.:2020   3rd Qu.:478.0
##                     Max.   :2024   Max.   :850.0
```

```
#This graph shows the shooting occurrences in each borough for each year
boro_analysis %>%
  ggplot(aes(x=OCCUR_YEAR, y=NUM_OCCURENCES, color=BORO)) +
  geom_line(linewidth=1) +
  geom_point() +
  labs(x="Year", y="Number of Shooting Occurences", title="Shooting Occurences Over Time in NY Boroughs
  theme_minimal()
```



This is a visualization which shows the number of shooting occurrences per year for each Borough.

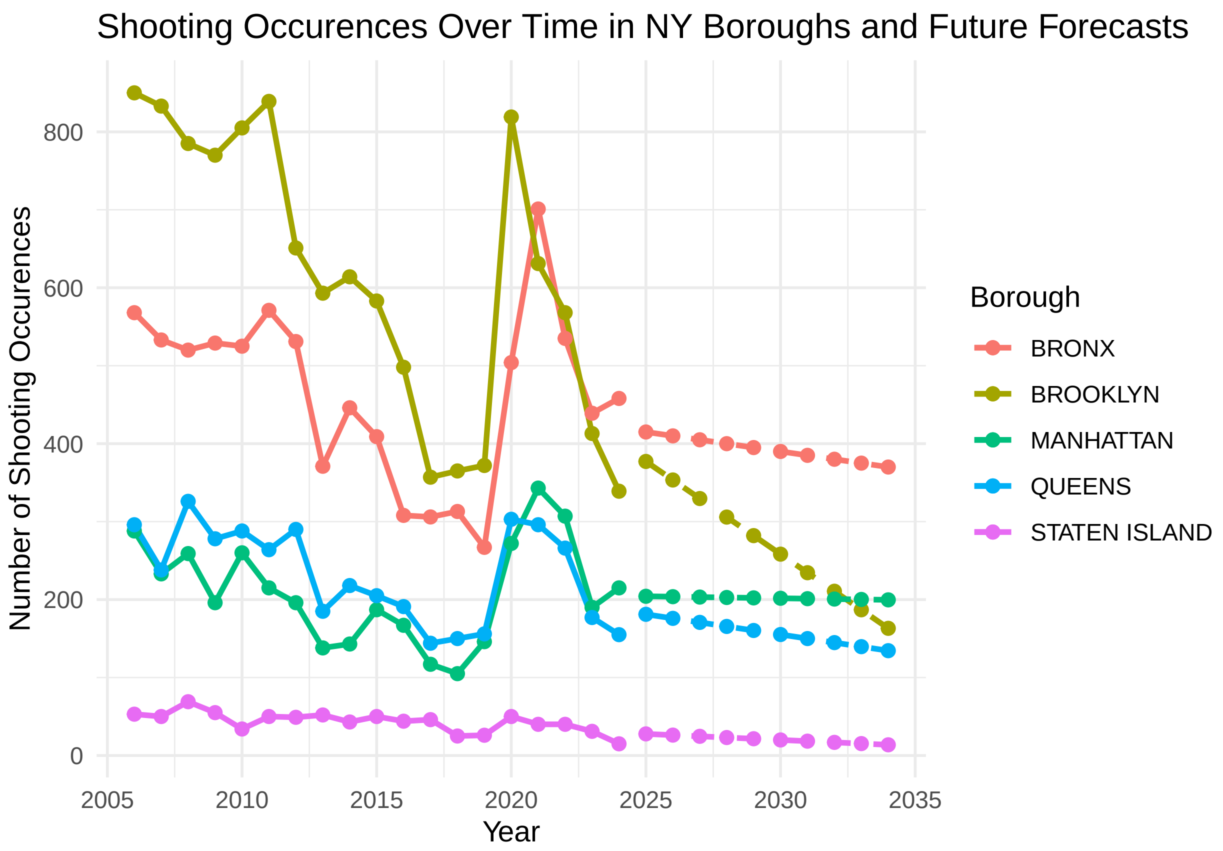
Future Shooting Occurence Modeling

In this section we will be using a basic linear model to predict the trend for the next 10 years for the number of shooting occurrences in each Borough.

The solid-lines represent historical data taken from the NYPD dataset. The dotted lines represent predictions made by a linear model which takes in the historical trend for each Borough and extrapolates it out 10 years.

```r
unique_boros <- unique(boro_analysis$BORO)
boro_forecast <- tibble()

for (boro in unique_boros) {

  boro_data <- boro_analysis %>% filter(BORO == boro) #Obtains data for just a single Borough

  model <- lm(NUM_OCCURENCES ~ OCCUR_YEAR, data = boro_data) #Fits a linear model for each borough

  future_years <- tibble(OCCUR_YEAR = seq(max(boro_data$OCCUR_YEAR) + 1, max(boro_data$OCCUR_YEAR) + 10

  forecast_results <- predict(model, future_years)

  # Creates a tibble with the boro name, the future year, the forecast for that year, and a signifier t
  individual_boro_forecast <- tibble(
    BORO = boro,
    OCCUR_YEAR = future_years$OCCUR_YEAR,
    NUM_OCCURENCES = forecast_results,
    TYPE = "Forecast"
  )

  # Appends the predictions for that Boro
  boro_forecast <- bind_rows(boro_forecast, individual_boro_forecast)
}


boro_history <- boro_analysis %>%
  mutate(TYPE = "Historical")

combined_borough_data <- bind_rows(boro_history, boro_forecast)
```

```r
# Plotting out both the hisorical and future predicted yearly shooting occurences per Borough
combined_borough_data %>%
  ggplot(aes(x=OCCUR_YEAR, y=NUM_OCCURENCES, color=BORO, group=BORO)) +
  geom_line(data=filter(combined_borough_data, TYPE == "Historical"), linewidth=1) +
  geom_point(data=filter(combined_borough_data, TYPE == "Historical"), size=2) +
  geom_line(data=filter(combined_borough_data, TYPE == "Forecast"), linetype = "dashed", linewidth=1) +
  geom_point(data=filter(combined_borough_data, TYPE == "Forecast"), size=2) +
  labs(x="Year", y="Number of Shooting Occurences", title="Shooting Occurences Over Time in NY Boroughs
  theme_minimal()
```



Shooting Occurrences by Year and Borough Conclusion

Since the year 2006 shooting incidents have generally declined (as in the Bronx and Brooklyn) or stayed roughly the same (as in Manhattan, Queens and Staten Island). The exception to this is 2020 and the following years, which saw a marked increase in the number of shootings. It seems a reasonable assumption to make that this spike in activity is due to COVID and it's knock-on effects. Perhaps there was reduced police presence due to the danger of the virus, or increased rates of mental problems due to lock downs, but I don't have enough data or analysis to say this definitively. The only exception to that 2020 spike is Staten Island, which has remained relatively stable throughout the whole period.

Our forecast, while not extremely sophisticated, does indicate that given the historical trend one would expect that shooting occurences will either stay stable or decrease as time goes on, as that has been the overall trend for all Boroughs. An expanded feature set and more advanced modeling techniques would likely give a more insightful answer as to what the future holds.

Without population data, it is hard to say whether the proportion of shooting occurrences relative to the population has increased or decreased, but it seems like that the policies enacted by New York to stop shooting occurrences have been effective enough to not increase the occurrences (except for the time during COVID). More analysis is needed to determine why places like Staten Island have such low and consistent rates of shooting occurrences, whereas places like the Bronx and Brooklyn are so much higher.

Potential Data Biases Some potential data biases I have identified are:

1. Population differences are not shown in the data set, so even though it seems like Brooklyn has the highest number of shooting occurrences, it is unknown what it's shooting rate is in relation to the total population?

2. The tracking and reporting of occurrences could have gotten better or worse over time.
3. There may have been ways in which “What counts as an occurrence?” has changed over time, skewing reporting.
4. There could be discrepancies in reporting between Boroughs.

Mitigating Data Biases In the hope of mitigating these biases, one could do the following:

1. Population data could be found for the Boroughs from another data set in order to get a more precise measure of shooting occurrences per-capita. This measure can then be used to more accurately compare Boroughs.
2. Research can be done on how the NYPD reports a shooting occurrence to understand the details of an occurrence report. If those details have changed over time, data slicing or manipulation can be used to obtain a more accurate view. Additionally, verification that the reporting is done in the same manner for all Boroughs could be obtained from such research.

Potential Personal Biases For my own personal bias in this analysis, as I don’t live in New York nor a big city, it is hard for me to personally relate to these numbers. 500 shooting incidents a year seems like a lot to me, but in a city as big as New York that might be a very small amount. It’s important, then, to not make any statements about the safety of the city due to this data (i.e., New York is dangerous because it has X amounts of shootings per year) without further population data and analysis comparing it to other cities.

Another reason I would hesitate to make conclusions about the safety of the city is that I am from the suburbs and generally find cities too overwhelming. I don’t particularly enjoy being in cities, and I would not like to live in one. This means I likely have a subconscious negative bias towards big cities like New York, which could affect my conclusions about the safety and quality of the city without proper mitigation strategies.

Mitigating Personal Biases By recognizing that my own experience of life is likely very different than someone living in New York, acknowledging my innate lack of affinity for big cities, and by explaining how not enough data is available to make conclusions about the safety and quality of New York, I hope to mitigate any personal biases and prevent any hasty conclusions from being made.

Potential Future Rates of Shooting Occurences

As an exercise, we will use a simple linear model to attempt to model the future rates of shooting occurences in each Borough, given the historical data available to us.

Analysis 2: Shooting Occurrences by Gender Dynamics

Here, we analyze the different gender dynamics that are involved in shooting occurrences. By gender dynamics, what I mean is: “What are the genders of the shooting perpetrator and the shooting victim, where there is a perpetrator and victim?”

Gender Data Preparation and Visualization

In this step, we keep two columns from the set:

1. “PERP_SEX”: the gender of the perpetrator.

2. “VIC_SEX”: the gender of the victim.

All nulls and NA values are removed from these two columns, as we are only concerned with incidents where there is a perpetrator and a victim. These two columns are then combined into one column, called “PERP_AND_VIC_TYPE”. The structure of this new column is:

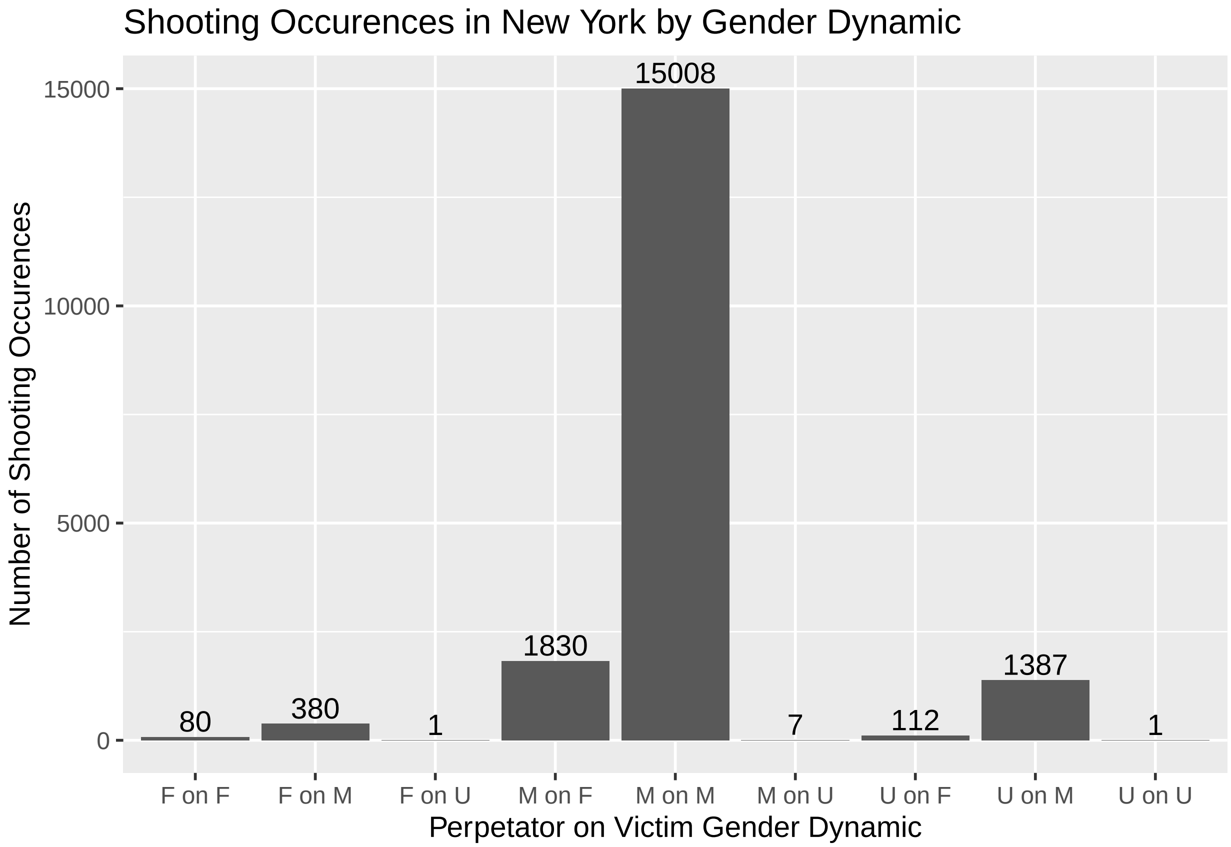
- “{PERP_SEX} on {VIC_SEX}”

This data is then counted based on the number of occurrences of each PERP_AND_VIC_TYPE category, giving us the next derived column: “NUM_OCCURRENCES”. This results in our final dataset used for this analysis.

```
# Here we will analyze the amount of shooting occurrences attributed to a given gender against another
## The categories will be like "male on female" or "female on unknown"
# To handle nulls, we will throw out any null occurrences in both the perpetrator gender and victim gen

gender_analysis <- nypd_filtered %>%
  select(c(PERP_SEX, VIC_SEX)) %>%
  filter(!is.na(PERP_SEX) & !is.na(VIC_SEX), PERP_SEX != "(null)", VIC_SEX != "(null)") %>%
  mutate(PERP_AND_VIC_TYPE = str_c(PERP_SEX, VIC_SEX, sep=" on ")) %>%
  group_by(PERP_AND_VIC_TYPE) %>%
  summarize(NUM_OCCURRENCES = n()) %>%
  ungroup()

summary(gender_analysis)
```

```
##  PERP_AND_VIC_TYPE  NUM_OCCURRENCES
##  Length:9           Min.   :    1
##  Class :character   1st Qu.:    7
##  Mode  :character   Median :  112
##                     Mean   : 2090
##                     3rd Qu.: 1387
##                     Max.   :15008
```

```
gender_analysis %>%
  ggplot(aes(x=PERP_AND_VIC_TYPE, y=NUM_OCCURRENCES)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = NUM_OCCURRENCES, vjust = -0.25)) +
  labs(x = "Perpetator on Victim Gender Dynamic", y = "Number of Shooting Occurences",  title = "Shootin
```



This is a visualization which shows the number of New York shooting occurrences for each Gender Dynamic.

Gender Dynamic Analysis Conclusion

From the data we can clearly see that Men commit significantly more shooting occurrences than Women or the "U" category, but it also is clear that Men are disproportionately the victims of shooting occurrences as well. The reasons for these are likely multifaceted and interesting, but there is not enough data here to definitively claim one reason or another. We can say, however, that as long as the data can be trusted, Men commit more shooting occurrences than any other gender and are more victimized by shooting occurrences than any other gender.

Potential Data Biases Some potential data biases I have identified are:

1. Population demographic differences (male vs female vs "u" ratios) are not defined, so it's impossible to get an accurate per-capita measure. If there were significantly more of one gender than the others, it would need to be accounted for in the analysis.
2. Is access to guns in any way influenced by gender? For example, if 50% of men had a gun and only 10% of women had a gun, it would follow that men would be likely to commit more shooting occurrences than women.

Mitigating Data Biases In the hopes of mitigating these biases, one could do the following:

1. Find the population numbers and demographic gender ratios for NY to obtain accurate per-capita data to include in the analysis.

2. Look up the firearm ownership rate for men and women in New York and consider that data in the analysis as well.

Potential Personal Biases I can identify two major personal biases which could impact any kind of conclusions drawn from this data. Firstly, there is a common cultural assumption that Men are more violent than Women - but is this true? The data above would seem to suggest such a thing, as the shooting occurrences are much higher with a male initiator than with any other group, but the reasons for this can be multifaceted.

Secondly, as I am a Man, does that mean I have a positive in-group bias? It's possible that I do. But, it's also possible that I have a negative in-group bias, in favor of Women. These unconscious biases can lead to inaccurate conclusions if not recognized and accounted for.

Mitigating Personal Biases In order to mitigate these personal biases, further investigations into the assumptions of "why" Men seem to commit more shooting occurrences is warranted. It could be gender ratios, access to firearms, economic opportunity, or intrinsic nature. Unfortunately, I just do not have the data here to make any descriptive statements about the character or natures of Men, Women or otherwise. Next, by recognizing that I am a Man and may have biases for and against any given gender, I can hope to mitigate that by strictly following the data and not letting my personal feelings come into the equation when drawing a conclusion.

```r
sessionInfo()
```

```
## R version 4.5.1 (2025-06-13 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##   LAPACK version 3.12.1
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.1.0     readr_2.1.5     tidyr_1.3.1     tibble_3.3.0
##  [9] ggplot2_3.5.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.6.0          gtable_0.3.6       crayon_1.5.3       compiler_4.5.1
##  [5] tidyselect_1.2.1   parallel_4.5.1     scales_1.4.0       yaml_2.3.10
##  [9] fastmap_1.2.0      R6_2.6.1           labeling_0.4.3     generics_0.1.4
## [13] curl_7.0.0         knitr_1.50         pillar_1.11.0      RColorBrewer_1.1-3
## [17] tzdb_0.5.0         rlang_1.1.6        stringi_1.8.7      xfun_0.52
## [21] bit64_4.6.0-1      timechange_0.3.0   cli_3.6.5          withr_3.0.2
## [25] magrittr_2.0.3     digest_0.6.37      grid_4.5.1         vroom_1.6.5
## [29] rstudioapi_0.17.1  hms_1.1.3          lifecycle_1.0.4    vctrs_0.6.5
## [33] evaluate_1.0.4     glue_1.8.0         farver_2.1.2       rmarkdown_2.29
## [37] tools_4.5.1        pkgconfig_2.0.3    htmltools_0.5.8.1
```