NYShootingData

A. Coles

2024-10-14

Loading NYPD Shooting Incident Data

The data for this project comes from data.gov and is described as a "list of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset." https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

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

ny_shooting <- read_csv(url_in)
```

```
## Rows: 28562 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.
```

Tidying NYPD Shooting Incident Data

Data was tidied to remove unnecessary rows and convert the date to a date format. All "null" or "unknown" values were changed to NA for consistency. For NA or missing data in the details on the locations, perpetrators, and victims, the conclusions will acknowledge the missing data.

```
# Remove columns not needed for analysis
ny_shooting <- ny_shooting %>%
  select(-INCIDENT_KEY, -LOC_OF_OCCUR_DESC, -LOC_CLASSFCTN_DESC,
         -LOCATION_DESC, -PRECINCT, -JURISDICTION_CODE,
         -STATISTICAL_MURDER_FLAG, -X_COORD_CD, -Y_COORD_CD,
```

```
         -Latitude, -Longitude, -Lon_Lat)

# Convert the date from a character to a date

ny_shooting <- ny_shooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

# Convert "null" or "unknown" to NA

ny_shooting <- ny_shooting %>%
  mutate(across(c(PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE),
                ~ na_if(., "null")),
         across(c(PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE),
                ~ na_if(., "unknown")))

# Show data
ny_shooting
```

```
## # A tibble: 28,562 x 9
##    OCCUR_DATE OCCUR_TIME BORO    PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
##    <date>     <time>     <chr>   <chr>          <chr>    <chr>     <chr>
##  1 2022-05-05 00:10      MANHAT~ 25-44          M        BLACK     25-44
##  2 2022-07-04 22:20      BRONX   (null)         (null)   (null)    18-24
##  3 2012-05-27 19:35      QUEENS  <NA>           <NA>     <NA>      18-24
##  4 2019-09-24 21:00      BRONX   25-44          M        UNKNOWN   25-44
##  5 2007-02-25 21:00      BROOKL~ 25-44          M        BLACK     25-44
##  6 2021-07-01 23:07      MANHAT~ <NA>           <NA>     <NA>      25-44
##  7 2021-06-07 19:55      QUEENS  <NA>           <NA>     <NA>      45-64
##  8 2021-07-22 01:47      BROOKL~ <NA>           <NA>     <NA>      25-44
##  9 2021-05-22 18:39      BRONX   <NA>           <NA>     <NA>      18-24
## 10 2021-12-22 23:17      BRONX   25-44          M        WHITE HI~ 25-44
## # i 28,552 more rows
## # i 2 more variables: VIC_SEX <chr>, VIC_RACE <chr>
```

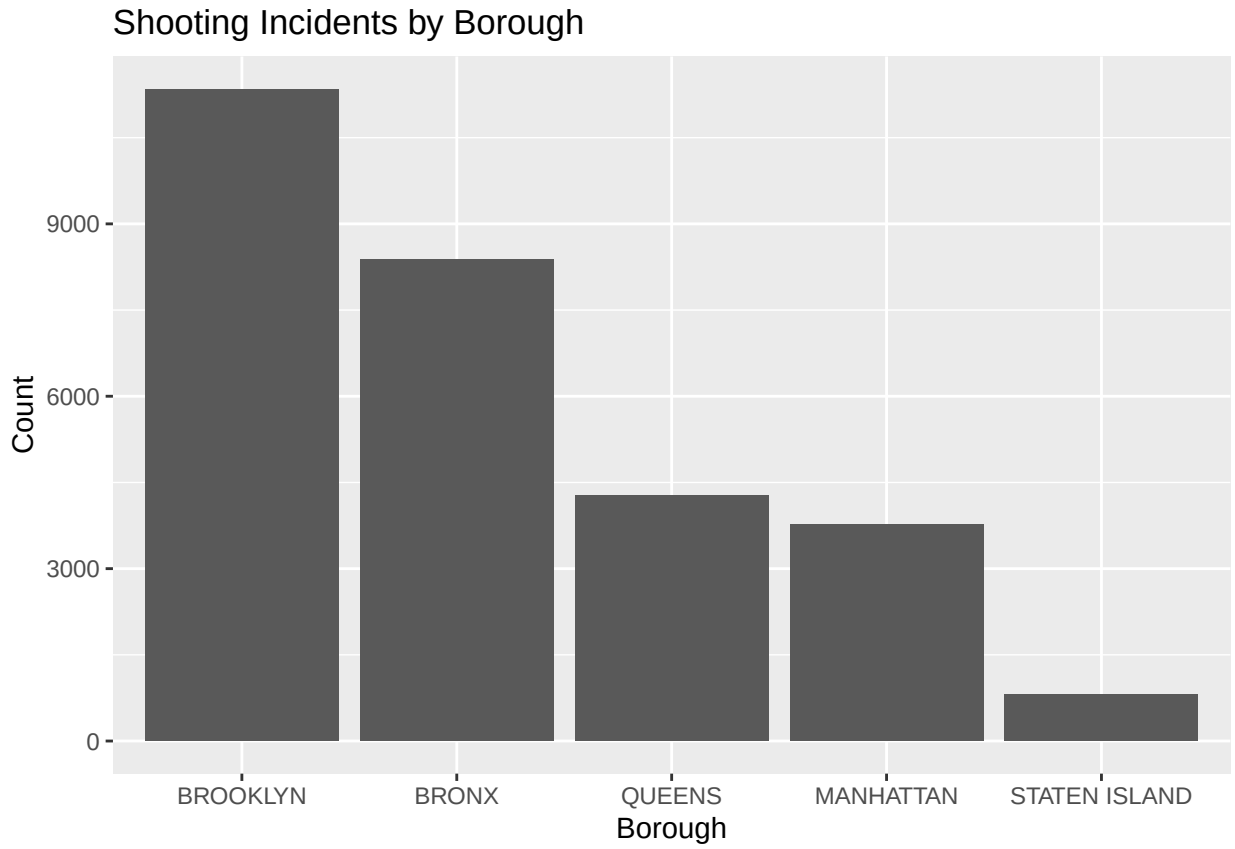
Analysis and Visualization

The following sections count the overall number of incidents and then plot them by borough.

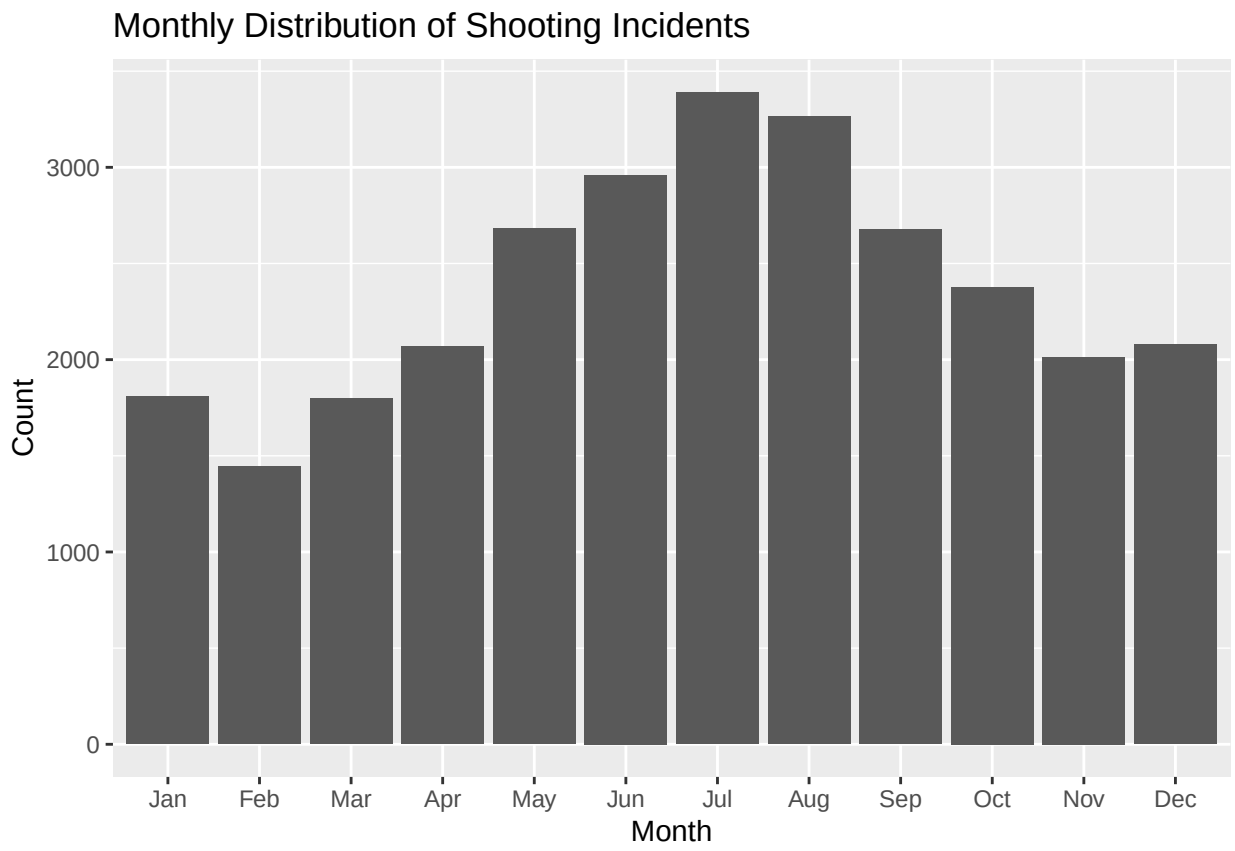
```
# Summarizing the data by borough.
summary_stats <- ny_shooting %>%
  summarize(
    total_incidents = n(),
    unique_boros = n_distinct(BORO)
    )
summary_stats
```

```
## # A tibble: 1 x 2
##   total_incidents unique_boros
##             <int>        <int>
## 1           28562            5
```

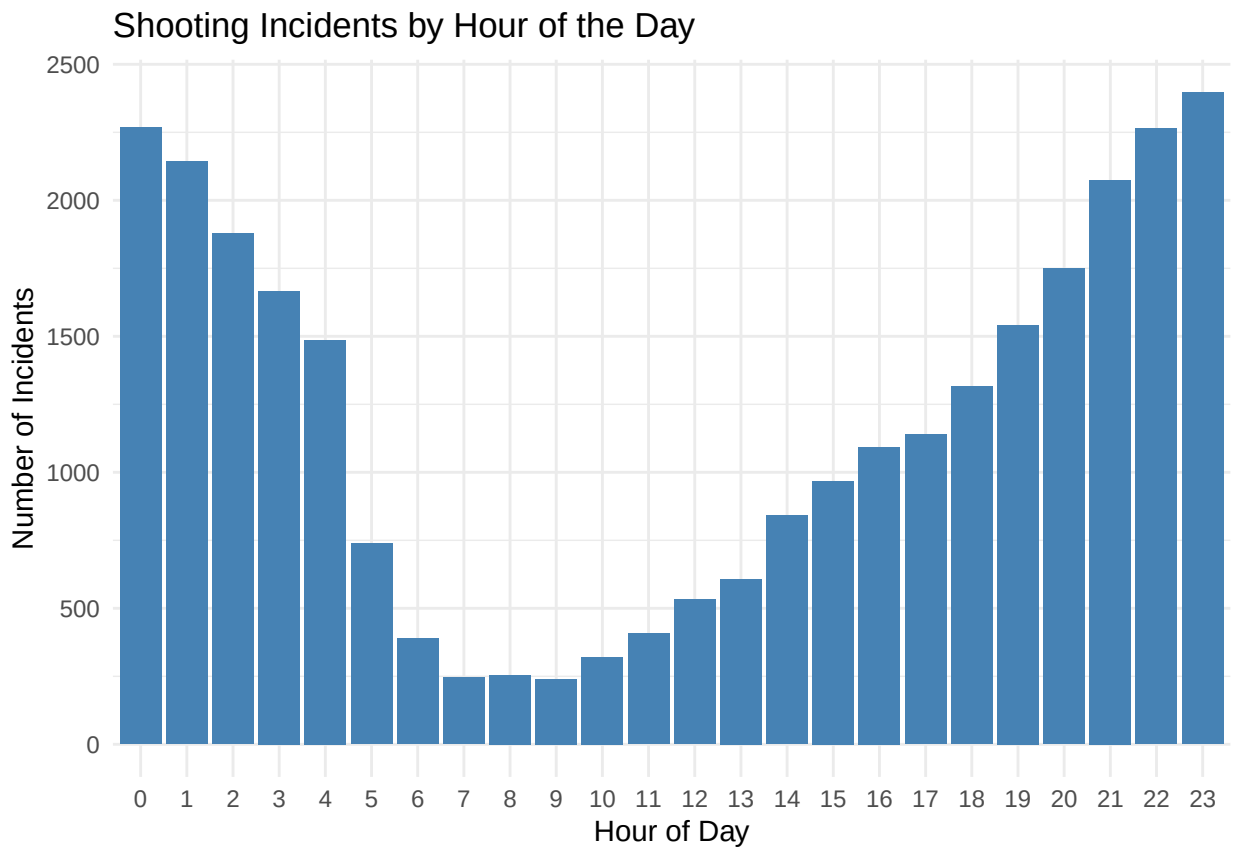
```r
# Counting the incidents per borough.
incidents_by_boro <- ny_shooting %>%
  group_by(BORO) %>%
  summarize(count = n())

# Plotting the incidents per borough.
ggplot(incidents_by_boro, aes(x = reorder(BORO, -count),
                              y = count)) +
  geom_bar(stat = "identity") +
  labs(title = "Shooting Incidents by Borough",
       x = "Borough", y = "Count")
```



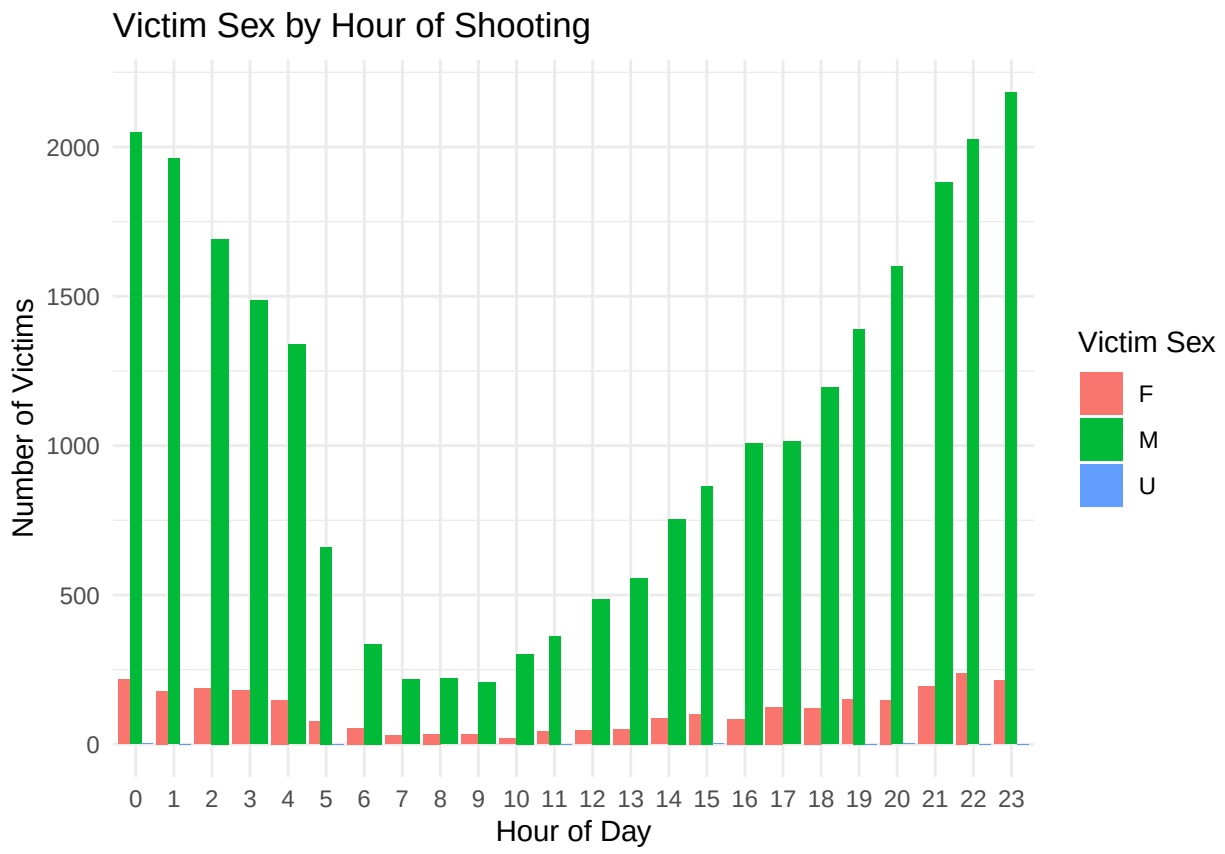

```r
# Analyzing how many incidents happened per month
ny_shooting <- ny_shooting %>%
    mutate(month = month(OCCUR_DATE, label = TRUE))

incidents_by_month <- ny_shooting %>%
  group_by(month) %>%
  summarize(count = n())

# Plotting incidents per month
ggplot(incidents_by_month, aes(x = month, y = count)) +
   geom_bar(stat = "identity") +
   labs(title = "Monthly Distribution of Shooting Incidents",
        x = "Month", y = "Count")
```



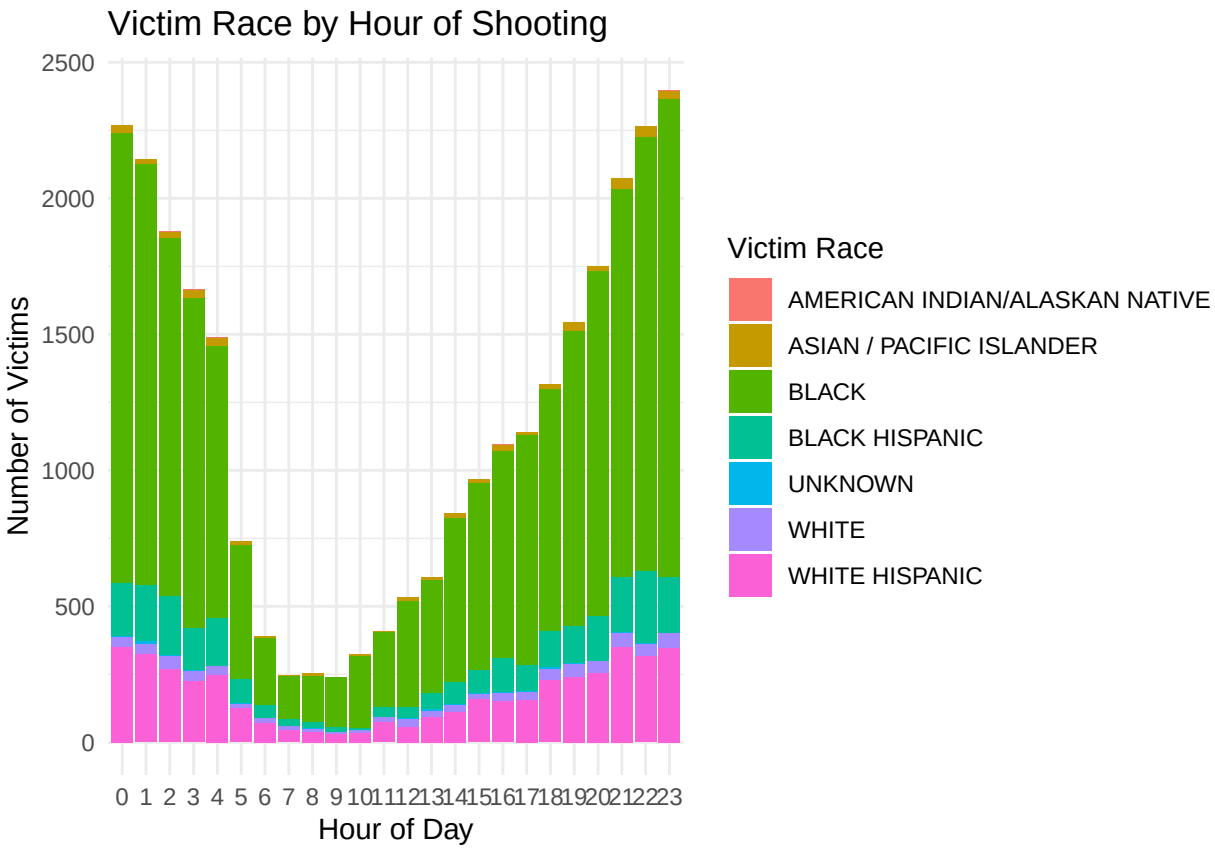

```
# Analyzing how many incidents happened at different times of day
ny_shooting <- ny_shooting %>%
  mutate(hour = hour(OCCUR_TIME))

hourly_victims <- ny_shooting %>%
  group_by(hour) %>%
  summarize(total_victims = n(), .groups = 'drop')

# Visualizing hourly incidents
ggplot(hourly_victims, aes(x = factor(hour), y = total_victims)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Shooting Incidents by Hour of the Day",
       x = "Hour of Day",
       y = "Number of Incidents") +
  theme_minimal()
```




```
# Visualizing Victom Sex by Hour of Shooting
ggplot(ny_shooting, aes(x = factor(hour), fill = VIC_SEX)) +
  geom_bar(position = "dodge") +
  labs(title = "Victim Sex by Hour of Shooting",
       x = "Hour of Day",
       y = "Number of Victims",
       fill = "Victim Sex") +
  theme_minimal()
```




```r
#Visualizing Victim Race by Hour of Shooting
ggplot(ny_shooting, aes(x = factor(hour), fill = VIC_RACE)) +
  geom_bar() +  # Default is stacked
  labs(title = "Victim Race by Hour of Shooting",
       x = "Hour of Day",
       y = "Number of Victims",
       fill = "Victim Race") +
  theme_minimal()
```

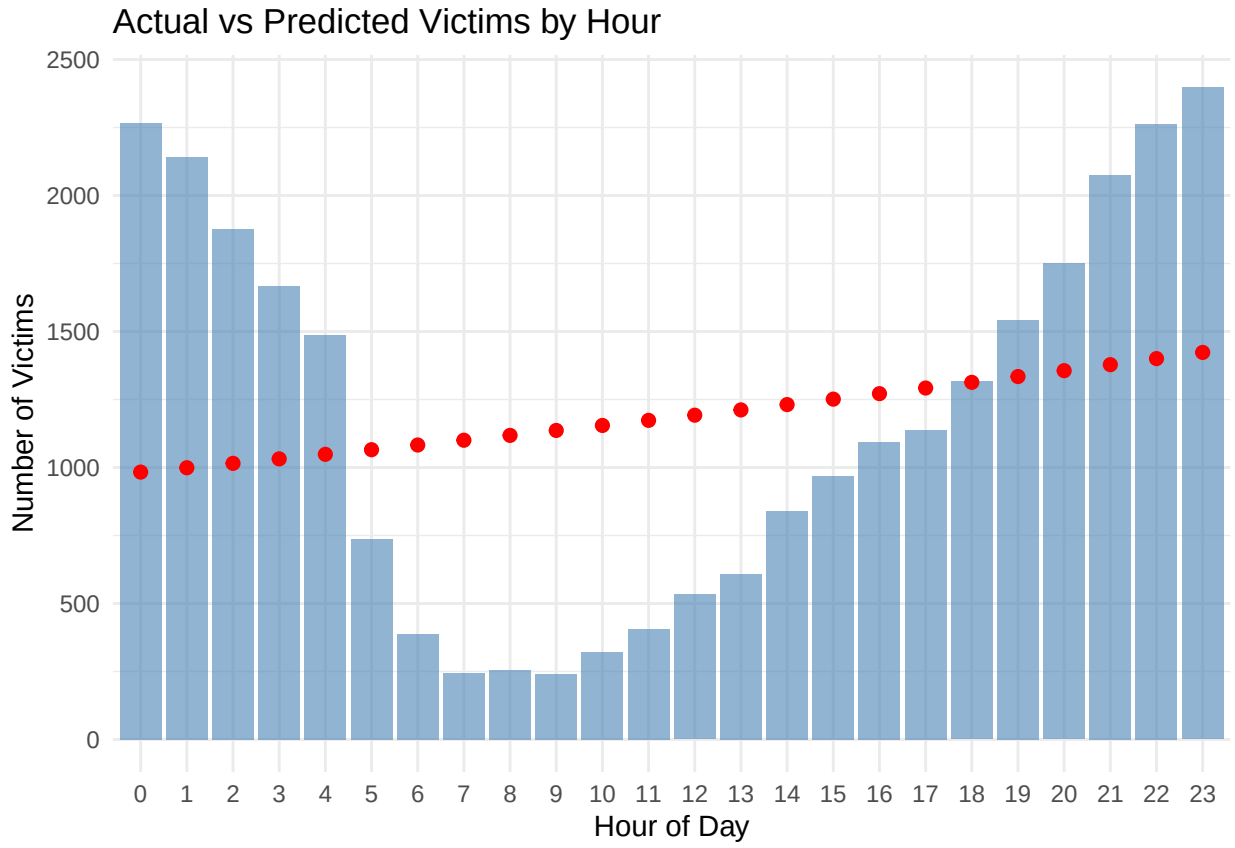


Modeling Hourly Victims

```r
# Fit the Poisson regression model
poisson_model <- glm(total_victims ~ hour, family = "poisson", data = hourly_victims)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = total_victims ~ hour, family = "poisson", data = hourly_victims)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.890640   0.012075  570.63   <2e-16 ***
## hour        0.016083   0.000858   18.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 11362  on 23  degrees of freedom
## Residual deviance: 11010  on 22  degrees of freedom
## AIC: 11222
##
## Number of Fisher Scoring iterations: 5
```

```r
# Create a dataframe for predictions
predictions <- hourly_victims %>%
  mutate(predicted_victims = predict(poisson_model, newdata = hourly_victims, type = "response"))

# Plot actual vs predicted values
ggplot() +
  geom_bar(data = hourly_victims, aes(x = factor(hour), y = total_victims),
           stat = "identity", fill = "steelblue", alpha = 0.6) +
  geom_point(data = predictions, aes(x = factor(hour), y = predicted_victims),
             color = "red", size = 2) +
  labs(title = "Actual vs Predicted Victims by Hour",
       x = "Hour of Day",
       y = "Number of Victims") +
  theme_minimal()
```



Conclusions and Bias

The analysis of NY police department reports of shooting incidents shown above demonstrates first that a simple poisson model of the number of shootings per hour misses the variation in the number of shootings across the course of a day. The descriptive statistics, however, demonstrate that a greater number of victims are shot during the evening hours of 5pm through 5am than during the daylight hours. The variation between numbers of female victims shot at different times of day is far less than the variation between male victims. This suggests that male victims are engaged in different activities or are shot for different reasons than female victims. Finally, the racial breakdown of victims by hour shows a greater number of shooting incidents across victims of all recorded races during the night-time hours, but most of all among black victims.

Bias: It is possible that bias creeps into the reporting statistics for shooting incidents either at the level of the witness, who might falsely identify a perpetrator according to their own biases, or the officer who recorded the report not recording race or gender as accurately as the perpetrator or victim would self-identify. To mitigate this sort of bias, I analyzed the victims' demographic profiles rather than the perpetrators'. The victim (or medical examiner) is less likely to skew their own (or their patients') reported demographic information. My personal bias that people are up to no good in the wee hours of the morning led to my curiosity about when shooting incidents occur. I would further my analysis about where incidents occurred to help me understand whether I was biased about shooters being out on the town, up to no good, or at home while shooting their victims.