NYPD Shooting Incident Project

anonymous student

2025-04-17

NYPD Shooting Incident Project

The data that will be loaded in is data directly provided by the city of New York that provides a list of every shooting incident that occurred in New York City starting from 2006 to 2024. There is data provided including, but not limited to, the borough in which it occurred, victim and perpetrator age group, victim and perpetrator sex, as well as information about race. For this project, I am interested in seeing if there is a correlation between victim sex or age group and the number of shooting occurrences within each of these groups.

Get the current data in

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

Load libraries.

Note: The TinyTex package will need to be installed prior to running this.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tinytex)
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 3.5.2     v tibble  3.2.1
## v purrr   1.0.4     v tidyr   1.3.1
## -- 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(ggplot2)
```

Data cleanup.

Clean the date up by aligning the format to yyyy-mm-dd and add code for changing the format. The columns that I am not interested in visualizing will also be excluded. I am mostly interested in seeing information about victim sex and age group.

```
nyc_data <- nyc_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
test3 <- nyc_data %>%
  select(-c(INCIDENT_KEY, BORO, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE,
            LOC_CLASSFCTN_DESC, LOCATION_DESC, STATISTICAL_MURDER_FLAG, X_COORD_CD,
            Y_COORD_CD, Latitude, Longitude, Lon_Lat, PERP_RACE, VIC_RACE))
```

Note: There are missing data for some of the age ranges and perpetrator sex. This can be ignored as it will not apply to the final data to be visualized.

Convert the time given in the data set to time that can be graphed.

There are blanks and N/As in some time values that will need to be filtered out.

Note: There is a warning for timezone but it can be ignored as this is all data for New York, so it is in the same timezone.

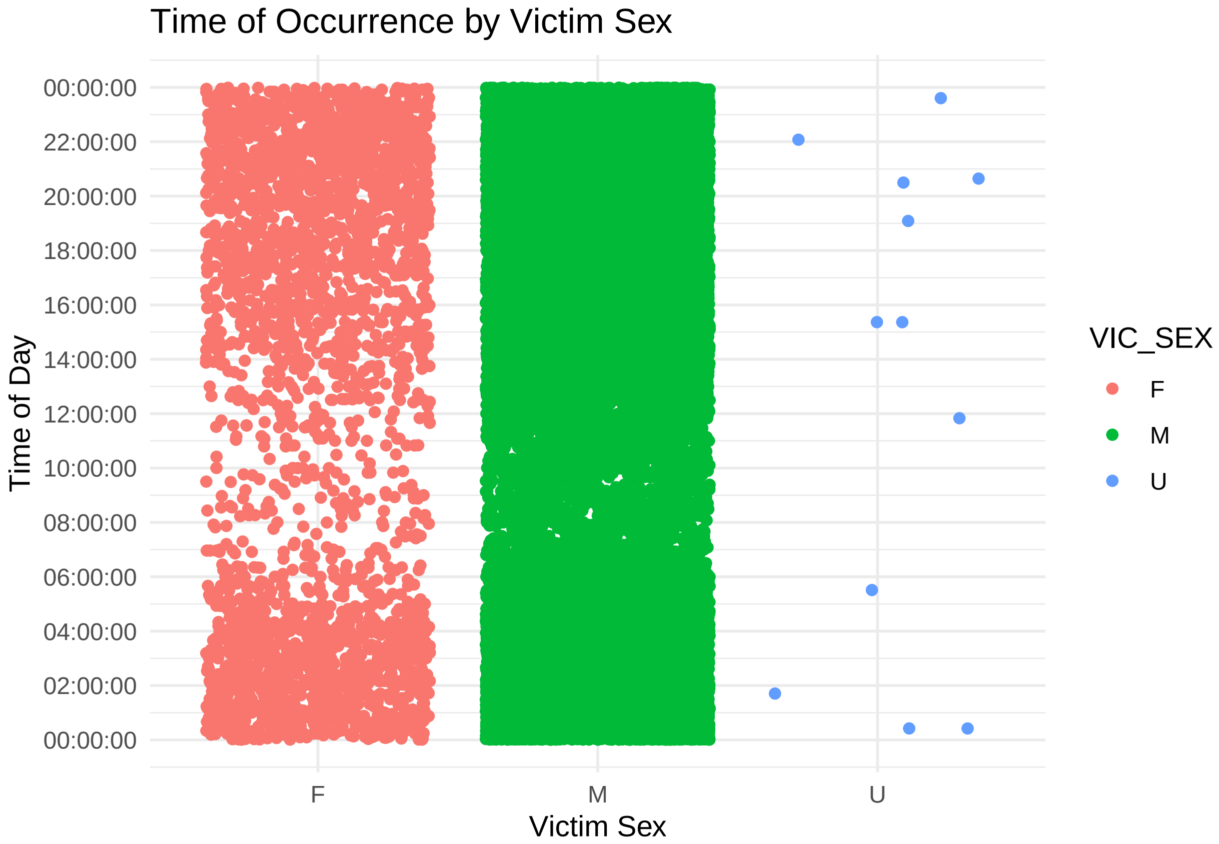
```
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME))
test3 <- test3 %>%
  filter(!is.na(OCCUR_TIME))
test3 <- test3 %>%
  mutate(OCCUR_TIME = as.numeric(OCCUR_TIME)) %>%
  filter(OCCUR_TIME>0)
test3 <- test3 %>%
  mutate(OCCUR_TIME = paste(sprintf("%02d", hour(OCCUR_TIME)), sprintf("%02d",
  minute(OCCUR_TIME)), sprintf("%02d", second(OCCUR_TIME)), sep = ":"))
```

```
## Warning: There were 3 warnings in ‘mutate()‘.
## The first warning was:
## i In argument: ‘OCCUR_TIME = paste(...)‘.
## Caused by warning:
## ! tz(): Don’t know how to compute timezone for object of class numeric; returning "UTC".
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 2 remaining warnings.
```

Create a graph to visualize the sex of victims and the times the shootings occur.

A jitter is used for visual clarity.

```
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME), OCCUR_SEC = as.numeric(OCCUR_TIME)) %>%
  filter(!is.na(OCCUR_SEC) & OCCUR_SEC > 0)
test3 %>%
  ggplot(aes(x = VIC_SEX, y = OCCUR_SEC, color = VIC_SEX)) +
  geom_jitter() +
  scale_y_continuous(name = "Time of Day", breaks = seq(0, 86400, by = 7200),
    labels = function(x) format(as.POSIXct(x, origin = "1970-01-01", tz = "UTC"), "%H:%M:%S")) +
  labs(title = "Time of Occurrence by Victim Sex", x = "Victim Sex") +
  theme_minimal()
```



Analysis for the hour at which the most shootings occurred by victim sex.

```
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME), OCCUR_HOUR = hour(OCCUR_TIME))
```

```
## Warning: There was 1 warning in ‘mutate()‘.
## i In argument: ‘OCCUR_TIME = hms(OCCUR_TIME)‘.
## Caused by warning in ‘.parse_hms()‘:
## ! Some strings failed to parse
```

```
test3 <- test3 %>%
  filter(!is.na(VIC_SEX))
highest_hour <- test3 %>%
  group_by(VIC_SEX, OCCUR_HOUR) %>%
  summarize(count = n()) %>%
  group_by(VIC_SEX) %>%
  slice_max(count, n=1) %>%
  ungroup()
```

```
## ‘summarise()‘ has grouped output by ’VIC_SEX’. You can override using the
## ‘.groups‘ argument.
```

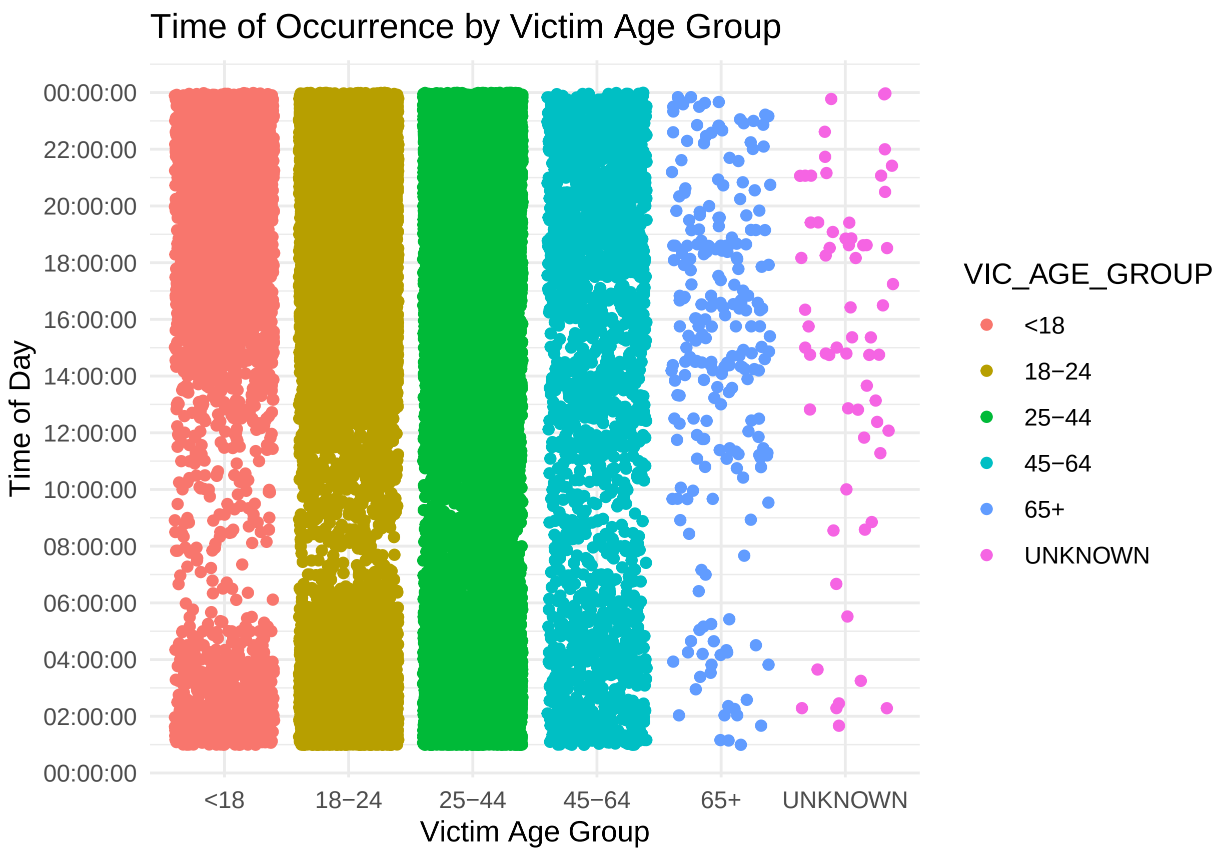
```
print(highest_hour)
```

```
## # A tibble: 5 x 3
##   VIC_SEX OCCUR_HOUR count
##   <chr>        <dbl> <int>
## 1 F               22   249
## 2 M               23  2250
## 3 U               15     2
## 4 U               20     2
## 5 U               NA     2
```

Create a graph for the victim age groups versus time of occurrences

An additional filter will be added to account for possible blanks and NAs in the victim age group category. There was also a random point at “1022” that needed to be removed as it was likely a typo.

```
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME), OCCUR_SEC = as.numeric(OCCUR_TIME)) %>%
  filter(!is.na(OCCUR_SEC) & OCCUR_SEC > 0)
test3 <- test3 %>%
  filter(!is.na(VIC_AGE_GROUP) & VIC_AGE_GROUP != "1022")
test3 %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = OCCUR_SEC, color = VIC_AGE_GROUP)) +
  geom_jitter() +
  scale_y_continuous(name = "Time of Day", breaks = seq(0, 86400, by = 7200),
    labels = function(x) format(as.POSIXct(x, origin = "1970-01-01", tz = "UTC"), "%H:%M:%S")) +
  labs(title = "Time of Occurrence by Victim Age Group", x = "Victim Age Group") +
  theme_minimal()
```



Analysis for the hour at which the most shootings occurred for each age group

```
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME), OCCUR_HOUR = hour(OCCUR_TIME))
highest_hour <- test3 %>%
  group_by(VIC_AGE_GROUP, OCCUR_HOUR) %>%
  summarize(count = n()) %>%
  group_by(VIC_AGE_GROUP) %>%
  slice_max(count, n = 1) %>%
  ungroup()
```

```
## ‘summarise()‘ has grouped output by ’VIC_AGE_GROUP’. You can override using the
## ‘.groups‘ argument.
```

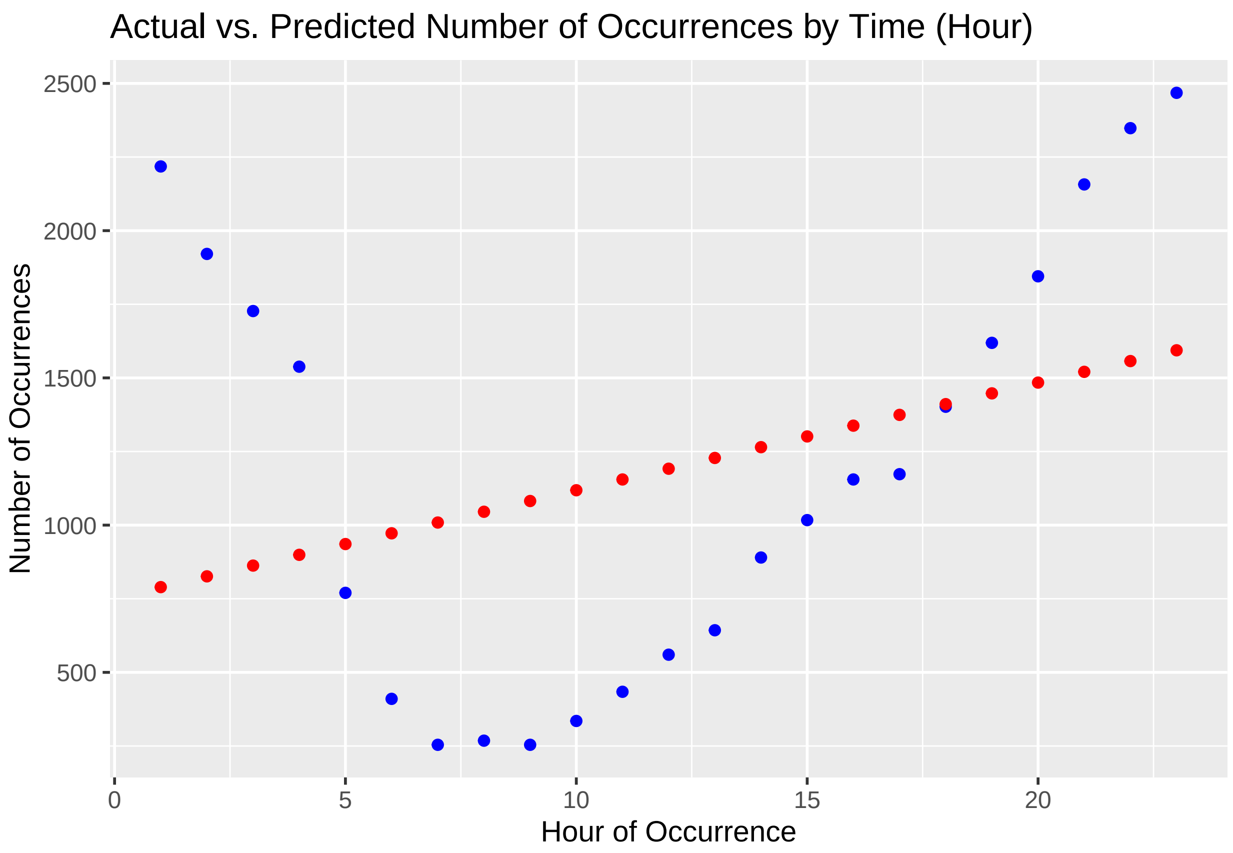
```
print(highest_hour)
```

```
## # A tibble: 6 x 3
##   VIC_AGE_GROUP OCCUR_HOUR count
##   <chr>              <dbl> <int>
## 1 18-24                 23   906
## 2 25-44                 23  1099
## 3 45-64                 22   153
## 4 65+                   14    26
## 5 <18                   23   296
## 6 UNKNOWN               18    10
```

Creating a model to see the correlation between time (hour) and amount of occurrences.

After running the model for the first time and seeing that the data is parabolic, I found that a parabolic rather than linear fit may be more appropriate.

```r
test3 <- test3 %>%
  filter(!is.na(OCCUR_TIME) & OCCUR_TIME > 0)
test3 <- test3 %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME), OCCUR_HOUR = hour(OCCUR_TIME))
counts_per_hour <- test3 %>%
  group_by(OCCUR_HOUR) %>%
  summarize(count = n()) %>%
  ungroup()
mod <- lm(count ~ OCCUR_HOUR, data = counts_per_hour)
counts_per_hour <- counts_per_hour %>%
  mutate(pred = predict(mod, newdata = counts_per_hour))
counts_per_hour %>%
  ggplot() +
  geom_point(aes(x = OCCUR_HOUR, y = count), color = "blue") +
  geom_point(aes(x = OCCUR_HOUR, y = pred), color = "red") +
  labs(title = "Actual vs. Predicted Number of Occurrences by Time (Hour)",
       x = "Hour of Occurrence", y = "Number of Occurrences")
```



Conclusion

Overall, the potential connections between both the victim sex and victim age groups affiliated with their times of occurrences showed similar correlations where these shootings are more likely to occur later in the night, not during the daytime while the perpetrators can be easily caught. It was interesting to see the density of times of occurrences for females and males differing, where much less shootings were reported during the day for females than males, however, both showed a similar trend. Another interesting case shown by the data is that for older age groups, there are less occurrences of shootings overall. Seeing how the data was more evenly spread out for individuals of older age groups than for younger age groups, there is a potential correlation between age group and likely time for them to be affected by a shooting. In analyzing and modeling the data, it is also seen that there is a parabolic trend leaning towards more shootings occurring at night. For females, the highest hour of occurrence was 10 PM while for males, the highest occurrence was at 11 PM. This clearly shows that there is a similarity in the data for males and females, and the hour of occurrence as reflected in the model. The linear regression does not fit the data perfectly as the data itself is parabolic, however, there is an upward trend in cases occurring more the later it is in the day. There were possible sources of bias affiliated with race that I did not want to include for this, which is why I focused on overall ages and sexes. Based on the media, a potential source of personal bias I could have seen is that the victim race may affect data where individuals of certain races are more or less likely to be a victim of a shooting than others. I mitigated this by not including it in the data.