NYPD Shooting Incident Data Report

2024-04-28

Get the NYPD shooting incident data.

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

nypd_sid <- 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.
```

This dataset contains many columns that I'm not interested in exploring right now, so I'm going to remove them.

```
nypd_sid <- nypd_sid %>%
  select(-c(JURISDICTION_CODE, PRECINCT, LOC_CLASSFCTN_DESC, LOC_OF_OCCUR_DESC, LOCATION_DESC))
```

```
nypd_sid <- nypd_sid %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, INCIDENT_KEY, STATISTICAL_MURDER_FLAG
```

The dates are formatted as characters, so I'll convert them into proper dates.

```
nypd_sid <- nypd_sid %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))
```

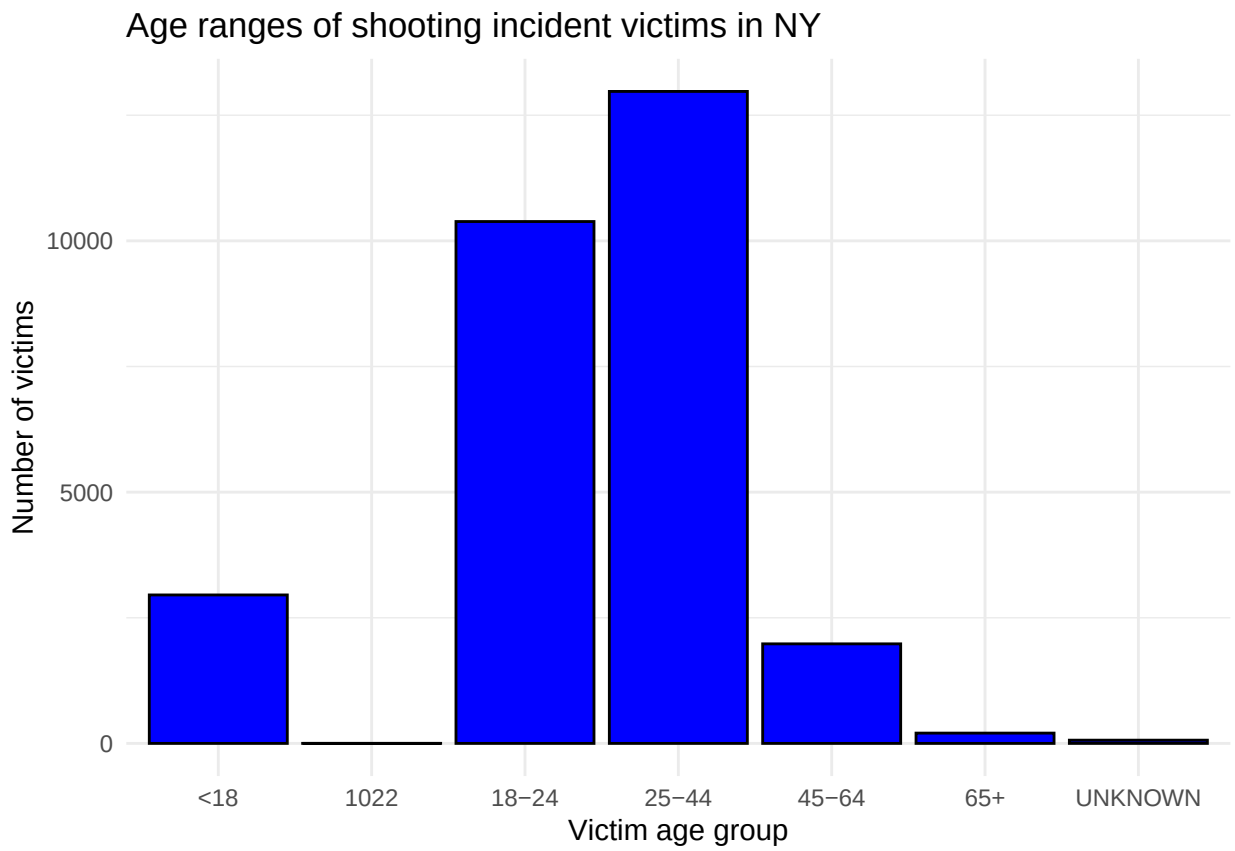
I want to figure out which age group has the most victims of shooting incidents in NYPD. So, I need to count the total number of victims and use the age group values as categories.

```
total_vic <- nrow(nypd_sid$VIC_AGE_GROUP)

nypd_sid$VIC_AGE_GROUP <- as.factor(nypd_sid$VIC_AGE_GROUP)
```

Let's visualize!

```
nypd_sid %>%
  ggplot(aes(x = VIC_AGE_GROUP)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Age ranges of shooting incident victims in NY",
       x = "Victim age group",
       y = "Number of victims")+
  theme_minimal()
```



The age group that is most frequently impacted by NY shooting incidents are those in the 25-44 year-old range.

I'm going to convert the time data into categorical time data for a heat map comparison.

```
nypd_sid$OCCUR_TIME <- ifelse(hour(nypd_sid$OCCUR_TIME) %in% 0:5, "Night",
                      ifelse(hour(nypd_sid$OCCUR_TIME) %in% 6:11, "Morning",
                        ifelse(hour(nypd_sid$OCCUR_TIME) %in% 12:17, "Afternoon",
                               "Evening")))
```

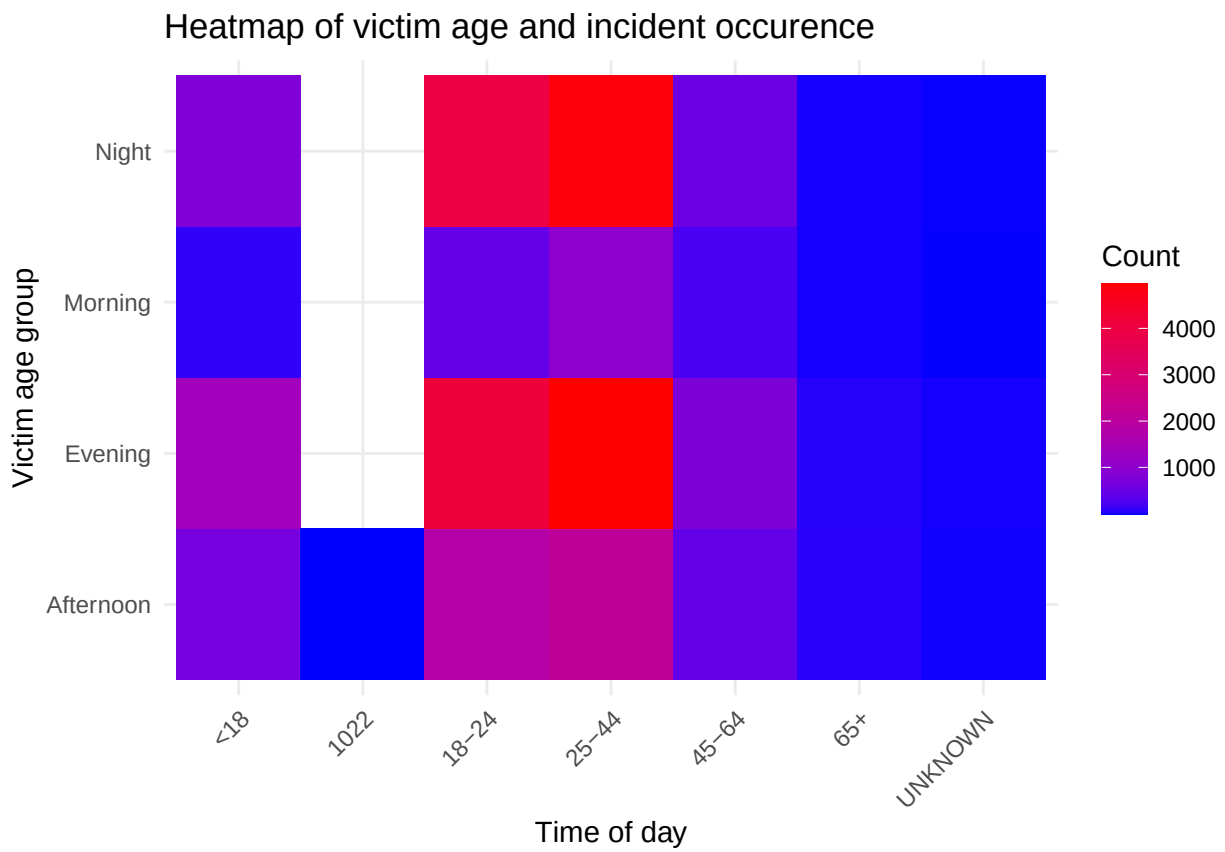
I need to count the occurences in each category for comparison.

```
counts <- nypd_sid %>%
  group_by(VIC_AGE_GROUP, OCCUR_TIME) %>%
  summarise(Count = n(), .groups = 'drop')

head(counts)
```

```
## # A tibble: 6 x 3
##   VIC_AGE_GROUP OCCUR_TIME Count
##   <fct>         <chr>      <int>
## 1 <18           Afternoon    679
## 2 <18           Evening     1380
## 3 <18           Morning      108
## 4 <18           Night        787
## 5 1022          Afternoon      1
## 6 18-24         Afternoon   1820
```

```
counts %>%
ggplot(aes(x = VIC_AGE_GROUP, y = OCCUR_TIME, fill = Count))+
  geom_tile()+
  scale_fill_gradient(low = "blue", high = "red")+
  labs(title = "Heatmap of victim age and incident occurence",
       x= "Time of day",
       y = "Victim age group",
       fill = "Count")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The highest concentration of victims are in the age groups 18-24 and 25-44.

```
mod <- lm(Count ~ VIC_AGE_GROUP + OCCUR_TIME, data = counts)
summary(mod)
```

```
##
```

```
## Call:
## lm(formula = Count ~ VIC_AGE_GROUP + OCCUR_TIME, data = counts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1344.9  -511.3     0.0   620.1  1112.5
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              412.1      555.1   0.742   0.4693
## VIC_AGE_GROUP1022       -411.1     1063.0  -0.387   0.7044
## VIC_AGE_GROUP18-24      1857.5      641.0   2.898   0.0110 *
## VIC_AGE_GROUP25-44      2504.7      641.0   3.908   0.0014 **
## VIC_AGE_GROUP45-64      -243.3      641.0  -0.379   0.7096
## VIC_AGE_GROUP65+        -687.3      641.0  -1.072   0.3006
## VIC_AGE_GROUPUNKNOWN    -722.5      641.0  -1.127   0.2774
## OCCUR_TIMEEvening       1026.8      523.4   1.962   0.0686 .
## OCCUR_TIMEMorning       -554.0      523.4  -1.059   0.3066
## OCCUR_TIMENight          832.7      523.4   1.591   0.1325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 906.5 on 15 degrees of freedom
## Multiple R-squared:  0.8001, Adjusted R-squared:  0.6802
## F-statistic: 6.672 on 9 and 15 DF,  p-value: 0.0007084
```

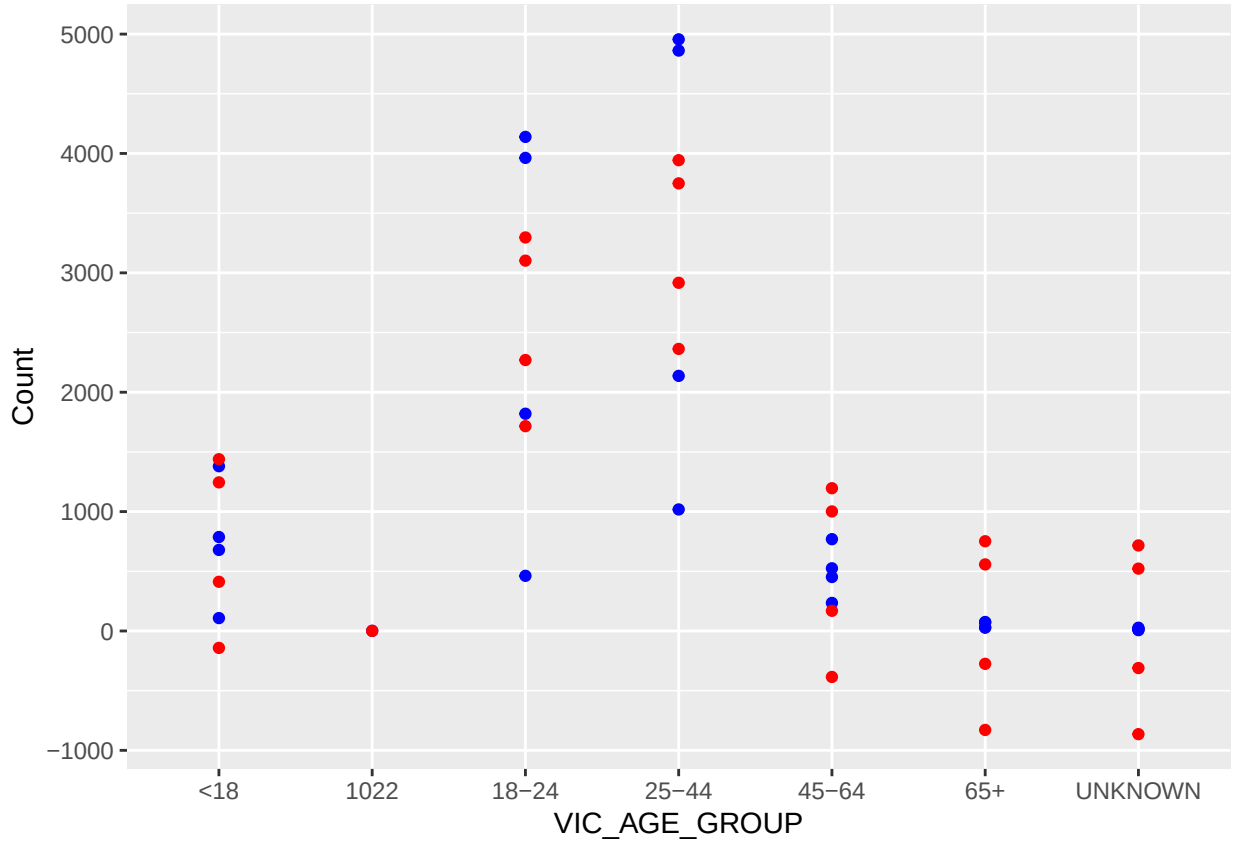
Now I'll attempt to predict the number of victims based on each given age range.

```
counts %>% mutate(pred = predict(mod))
```

```
## # A tibble: 25 x 4
##    VIC_AGE_GROUP OCCUR_TIME Count    pred
##    <fct>         <chr>      <int>   <dbl>
##  1 <18           Afternoon    679  412.
##  2 <18           Evening     1380 1439.
##  3 <18           Morning      108 -142.
##  4 <18           Night        787 1245.
##  5 1022          Afternoon      1    1.00
##  6 18-24         Afternoon   1820 2270.
##  7 18-24         Evening     4139 3296.
##  8 18-24         Morning      462 1716.
##  9 18-24         Night       3963 3102.
## 10 25-44         Afternoon   2137 2917.
## # i 15 more rows
```

```
counts_w_pred <- counts %>% mutate(pred = predict(mod))
```

```
counts_w_pred %>%
  ggplot() +
  geom_point(aes(x = VIC_AGE_GROUP, y = Count), color = "blue") +
  geom_point(aes(x = VIC_AGE_GROUP, y = pred), color = "red")
```



Biases:

Going into this study I held the assumption that most of the victims of shooting incidents would be people in their 20s late at night.

Although the highest concentration of shooting incidents took place in the evening and at night, the age range of 25-44 is too broad to give an accurate picture of if the majority of victims were in their 20s.I tried to mitigate this bias by allowing the data to lead the way. This data also might be skewed towards 18-44 year-olds because there is a larger population of them in NYC.