NYPD Shooting Incident Data (Historic)

Claire Robbins

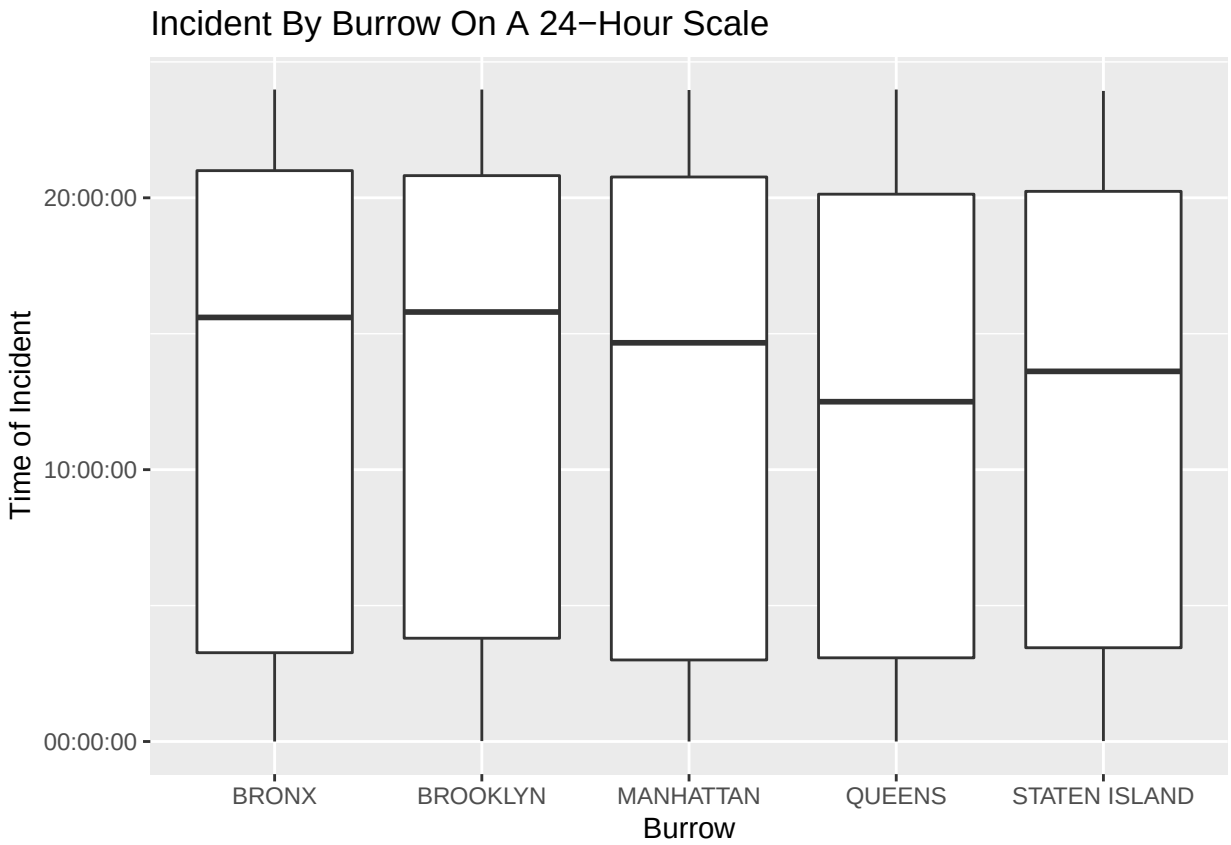
2022-06-07

R Markdown

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

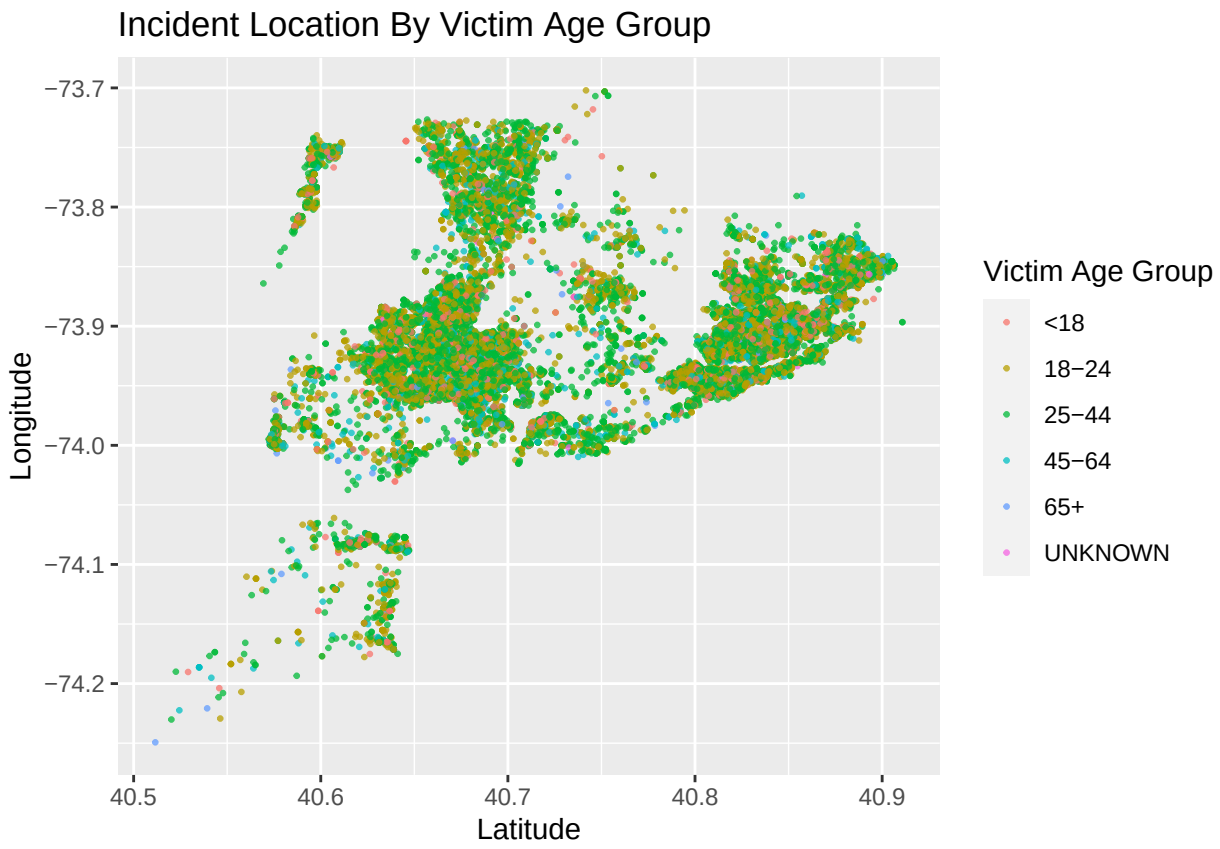
```
The first visualization I chose to explore is a visual representation of the time of
incident across each burrow. Although there did not seem to be a significant
difference between each of the burrows, it was a good first step in data
exploration.
```

```
data %>% ggplot(aes(x=BORO, y=OCCUR_TIME)) +
  geom_boxplot() +
  labs(title = 'Incident By Burrow On A 24-Hour Scale',
    x = 'Burrow', y = 'Time of Incident')
```



The second visualization I created was a scatterplot of latitude vs longitude, broken down by victim age group. The goal of this visualization was to see if there is any obvious correlation between location and victim age, but the plot was generally overwhelmed by the 25-44 age group. Although initially unintended, you can see a fairly clear representation of each of the burrows as a function of longitude and latitude.

```
data %>% ggplot(aes(Latitude, Longitude, color=VIC_AGE_GROUP)) +
  geom_point(alpha=0.75, size=0.5) +
  labs(title = 'Incident Location By Victim Age Group') +
  scale_colour_discrete(name="Victim Age Group")
```



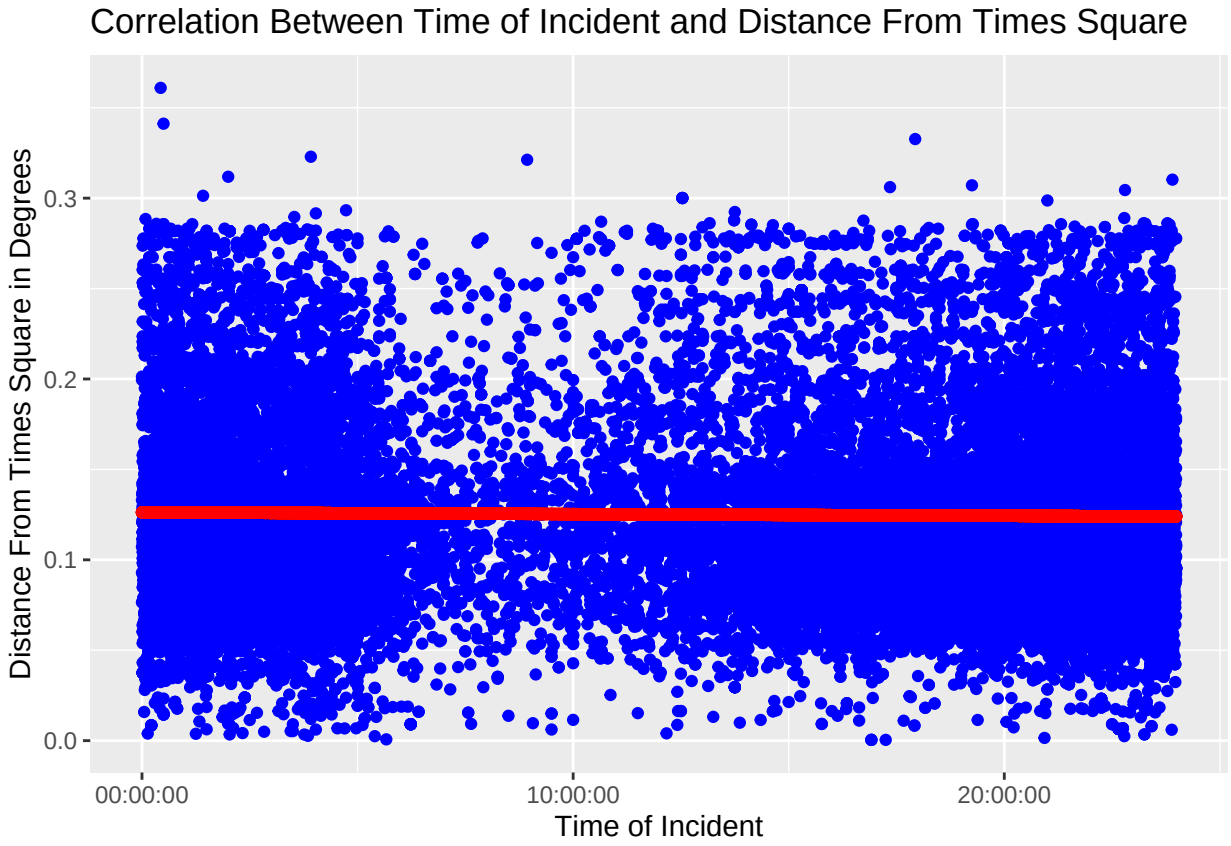
The analysis and linear model I chose to explore was distance from Times Square, as a function of time of incident. I predicted that incidents would increase closer to the heart of the city, such as Times Square, in the late night and early morning. However, with such an overwhelming amount of incidents across the different burrows, there was no real correlation between distance from times square and time of incident. You can clearly see a minimization of incidents between 7 and 9am.

```
data$LAT_FROM_TISQ <- 40.7580 - data$Latitude
data$LON_FROM_TISQ <- -73.9855 - data$Longitude
data$DIST_FROM_TISQ <- sqrt(data$LAT_FROM_TISQ**2 + data$LON_FROM_TISQ**2)

mod <- lm(DIST_FROM_TISQ ~ OCCUR_TIME, data = data)
summary(mod)
```

```
##
## Call:
```

```
## lm(formula = DIST_FROM_TISQ ~ OCCUR_TIME, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.124983 -0.034371 -0.008031  0.021411  0.234790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.262e-01  5.813e-04 217.151   <2e-16 ***
## OCCUR_TIME  -2.645e-08  1.058e-08  -2.499   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05198 on 25594 degrees of freedom
## Multiple R-squared:  0.000244,   Adjusted R-squared:  0.0002049
## F-statistic: 6.246 on 1 and 25594 DF,  p-value: 0.01245
```

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

predict %>% ggplot() +
  geom_point(aes(x = OCCUR_TIME, y = DIST_FROM_TISQ), color = "blue") +
  geom_point(aes(x = OCCUR_TIME, y = pred), color = "red") +
  labs(title = 'Correlation Between Time of Incident and Distance From Times Square',
    x = 'Time of Incident', y = 'Distance From Times Square in Degrees')
```




```
I imagine there to be bias in the Statistical Murder Flag column, as that is based
in subjectivity and policy on policing and safety. I focused my analysis primarily
on time and location, which are two minimally biased metrics. Depending on how the
```

```
data was collected, there could be bias in variables such as race, location
description, and binary sex.
```