M3Project

Kane

2025-11-21

Statement of Question/Interest

This analysis will focus on the number of incidents over time and when they occur though the day

Questions: How do the number of incidents change over the the years? What times of day are most likely for shootings to occur?

Description of Data

Source: City of New York

Description: This data is provided by the City of New York and contains the all reported shooting incidents from 2006 to 2024. The reported data is when the incident took place, what boro, which precinct, the jurisdiction code, whether or not it was statistically linked to a murderer, a basic description of the perpetrator (age, sex, race), a basic description of the victim (age, sex, race), global coordinators, and local map coordinators. The data also included descriptions of the incident scene, which as been ommited from this import due to lack of entries.

```r
library(tidyverse)
library(zoo)
library(lubridate)
library(plyr)
library(ggplot2)
library(scales)
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

incident_data <- read_csv(url_in)
```

```r
# Convert all unknown data to 'UNKNWON' for consitency
incident_data$PERP_AGE_GROUP[incident_data$PERP_AGE_GROUP %in%
    c("(null)", NA)] = "UNKNOWN"
incident_data$PERP_SEX[incident_data$PERP_SEX %in% c("(null)",
    "U", NA)] = "UNKNOWN"
incident_data$PERP_RACE[incident_data$PERP_RACE %in% c("(null)",
    NA)] = "UNKNOWN"

incident_data$VIC_AGE_GROUP[incident_data$VIC_AGE_GROUP %in%
    c("(null)", NA)] = "UNKNOWN"
incident_data$VIC_SEX[incident_data$VIC_SEX %in% c("(null)",
    "U", NA)] = "UNKNOWN"
incident_data$VIC_RACE[incident_data$VIC_RACE %in% c("(null)",
```

```r
    NA)] = "UNKNOWN"


# Transform the data to the appropriate data type
incident_data = transform(incident_data, PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP))

incident_data = transform(incident_data, OCCUR_DATE = as.Date(OCCUR_DATE,
    format = "%m/%d/%Y"))
incident_data = transform(incident_data, BORO = as.factor(BORO))
incident_data = transform(incident_data, PRECINCT = as.factor(PRECINCT))
incident_data = transform(incident_data, JURISDICTION_CODE = as.factor(JURISDICTION_CODE))

incident_data = transform(incident_data, PERP_SEX = as.factor(PERP_SEX))
incident_data = transform(incident_data, PERP_RACE = as.factor(PERP_RACE))

incident_data = transform(incident_data, VIC_AGE_GROUP = as.factor(PERP_AGE_GROUP))
incident_data = transform(incident_data, VIC_SEX = as.factor(VIC_SEX))
incident_data = transform(incident_data, VIC_RACE = as.factor(VIC_RACE))


# Duplicate Data, lat and lon both already in table
incident_data = subset(incident_data, select = -Lon_Lat)

# Removed because most of the data did not include entries
# for these descriptions --- Probably optional when the
# report was filed or change to how reports were filed at
# some point to add this information
incident_data = subset(incident_data, select = -c(LOC_OF_OCCUR_DESC,
    LOC_CLASSFCTN_DESC, LOCATION_DESC))

summary(incident_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :00:00:00.000000
##  1st Qu.: 67321140   1st Qu.:2009-10-29   1st Qu.:03:30:45.000000
##  Median :109291972   Median :2014-03-25   Median :15:15:00.000000
##  Mean   :133850951   Mean   :2014-10-31   Mean   :12:46:10.874798
##  3rd Qu.:214741917   3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000
##  Max.   :299462478   Max.   :2024-12-31   Max.   :23:59:00.000000
##
##            BORO          PRECINCT     JURISDICTION_CODE
##  BRONX        : 8834   75      : 1680   0   :24957
##  BROOKLYN     :11685   73      : 1561   1   :  109
##  MANHATTAN    : 3977   67      : 1288   2   : 4676
##  QUEENS       : 4426   44      : 1159   NA's:    2
##  STATEN ISLAND:  822   79      : 1073
##                        47      : 1048
##                        (Other):21935
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           UNKNOWN:14120   F      :  461
##  FALSE:23979             18-24  : 6630   M      :16845
##  TRUE :5765              25-44  : 6342   UNKNOWN:12438
##                          <18    : 1805
```

```
##                          45-64   :  775
##                          65+     :   67
##                          (Other):    5
##                            PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   UNKNOWN:14120   F       : 2891
##  ASIAN / PACIFIC ISLANDER      :  184   18-24   : 6630   M       :26841
##  BLACK                         :12323   25-44   : 6342   UNKNOWN:    12
##  BLACK HISPANIC                : 1487   <18     : 1805
##  UNKNOWN                       :12776   45-64   :  775
##  WHITE                         :  305   65+     :   67
##  WHITE HISPANIC                : 2667   (Other):    5
##                             VIC_RACE       X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:   13   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  478   1st Qu.:1000094   1st Qu.:183042
##  BLACK                         :20999   Median :1007826   Median :195506
##  BLACK HISPANIC                : 2930   Mean   :1009442   Mean   :208722
##  UNKNOWN                       :   72   3rd Qu.:1016739   3rd Qu.:239980
##  WHITE                         :  741   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 4511
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.91
##  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
```

Number of Incidents Over Time

```r
frequency_over_time <- incident_data %>%
    subset(select = OCCUR_DATE)  #%>% rename(Date = OCCUR_DATE)

year <- frequency_over_time$OCCUR_DATE %>%
    year()
quarter <- frequency_over_time$OCCUR_DATE %>%
    quarters()

frequency_over_time <- data.frame(year, quarter) %>%
    ddply(.(.$year, .$quarter), nrow)
names(frequency_over_time) <- c("Year", "Quarter", "Count")

FoT_graph <- frequency_over_time %>%
    ggplot(aes(fill = Quarter, x = Year, y = Count)) + geom_bar(position = "dodge",
    stat = "identity") + labs(title = "NYPD Reported Shooting Incidents Overtime by Quarter")

model_incidents_over_time = glm(Count ~ Year + Quarter + Year:Quarter,
    data = frequency_over_time, family = "poisson")
model_incidents_over_time_noint = glm(Count ~ Year + Quarter,
    data = frequency_over_time, family = "poisson")
# Poisson selected due to modeling counts of something,
# looking for an overall trend, therefore simple glm with
```

```r
# interaction and not any loess or ksmooth


tend_line_width = 1

FoT_graph = FoT_graph + geom_line(aes(y = fitted(model_incidents_over_time),
    group = Quarter), size = tend_line_width * 1.5, linetype = "solid",
    lineend = "round", color = "#000") + geom_line(aes(y = fitted(model_incidents_over_time),
    color = Quarter), size = tend_line_width, linetype = "solid",
    lineend = "round") + theme_bw()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
summary(model_incidents_over_time)
```

```
##
## Call:
## glm(formula = Count ~ Year + Quarter + Year:Quarter, family = "poisson",
##     data = frequency_over_time)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    32.212750   5.062212   6.363 1.97e-10 ***
## Year           -0.013194   0.002513  -5.251 1.52e-07 ***
## QuarterQ2      15.753293   6.526072   2.414 0.015783 *
## QuarterQ3      23.539532   6.305611   3.733 0.000189 ***
## QuarterQ4      37.120675   6.786191   5.470 4.50e-08 ***
## Year:QuarterQ2 -0.007613   0.003240  -2.350 0.018771 *
## Year:QuarterQ3 -0.011386   0.003130  -3.638 0.000275 ***
## Year:QuarterQ4 -0.018309   0.003369  -5.435 5.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3587.6  on 75  degrees of freedom
## Residual deviance: 1668.8  on 68  degrees of freedom
## AIC: 2273.6
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(model_incidents_over_time)
```

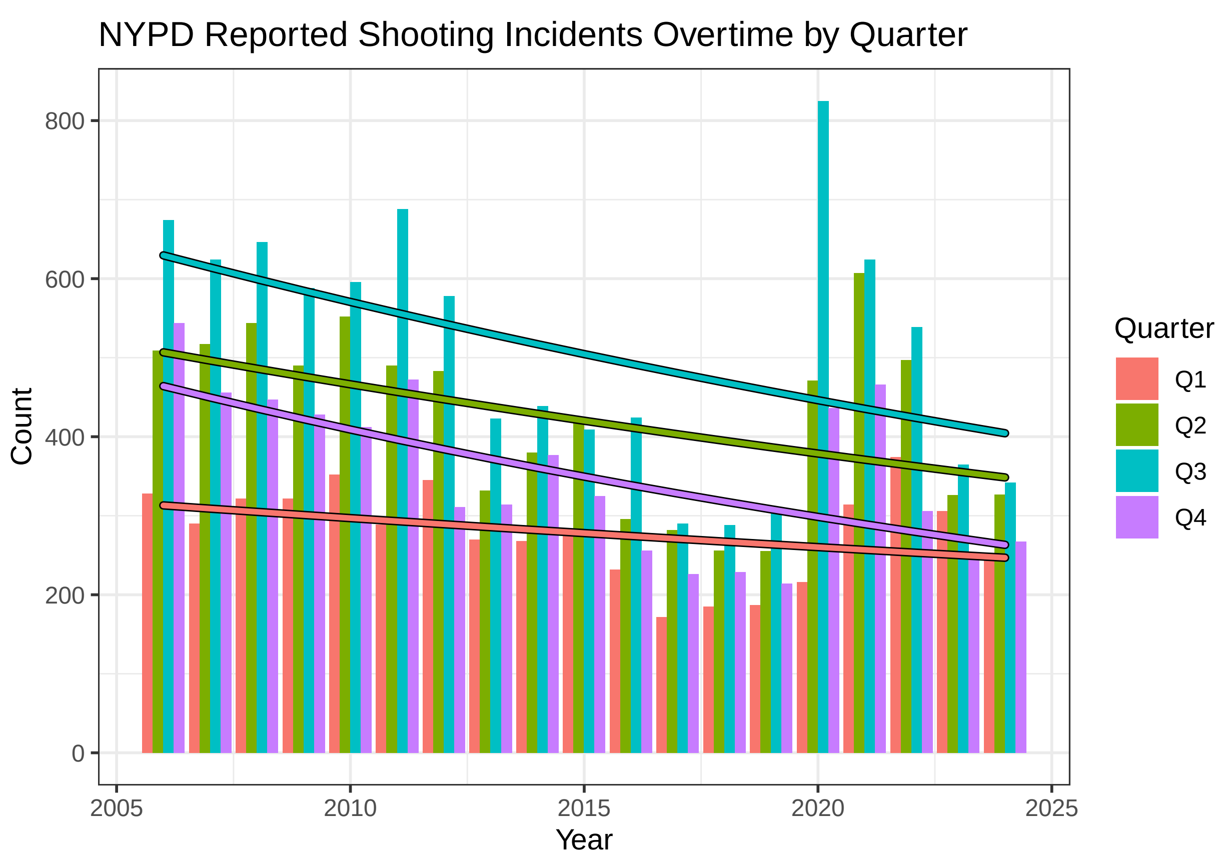
```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Count
```

```
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             75     3587.6
## Year           1   473.33        74     3114.3 < 2.2e-16 ***
## Quarter        3  1414.01        71     1700.3 < 2.2e-16 ***
## Year:Quarter   3    31.46        68     1668.8 6.795e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_incidents_over_time_noint, model_incidents_over_time)
```

```
## Analysis of Deviance Table
##
## Model 1: Count ~ Year + Quarter
## Model 2: Count ~ Year + Quarter + Year:Quarter
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        71     1700.3
## 2        68     1668.8  3   31.462 6.795e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
FoT_graph
```



```r
mu <- function(year, q2, q3, q4) return(32.21275 - 0.013194 *
    year + 15.753293 * q2 + 23.539532 * q3 + 37.120675 * q4 -
    0.007613 * year * q2 - 0.011386 * year * q3 - 0.018309 *
    year * q4)


total <- function(year) exp(mu(year, 0, 0, 0)) + exp(mu(year,
    1, 0, 0)) + exp(mu(year, 0, 1, 0)) + exp(mu(year, 0, 0, 1))

decrease_percentage = mean(1 - total(2007:2024)/total(2006:2023))

print(decrease_percentage)  # [1] 0.02281219 -> 2.28% year/year on average
```

```
## [1] 0.02281219
```

According to the model there is an average of around 2.28% fewer shooting incidents every year. However this data seems to indicate that there was a noticeable increase in incidents in 2020, which coincides with the HCoV-19 (COVID-19) global pandemic. This may skew the data, but not enough time has past for a valid model to be produced with the data post 2020. Further observation needs to be carried out in the following years. The data also shows that quarter 3 seems to have a significantly higher rate of shooting incidents while quarter 4 has the fewest. I am not sure what causes the drastic differences in the quarters. Quarter 1: Jan, Feb, Mar; quarter 2: Apr, May, Jun; quarter 3: Jul, Aug, Sep; quarter 4: Oct, Nov, Dec. With the months of each quarter, the Q3 increase in incidents line up with the beginning of school/end of summer. Both q2 and q3 are spring/summer months, where q4 and q1 are winter months. There may be more which align with these months, but I do not know and more research would be required to determine what the correlated events are which may cause the disparity in cases. A large source of bias is that I am using a poisson linear model for this analysis, I am assuming that there is a trend with the number of incidents based on the year and quarter. According to the ANOVA there is statistically significant interaction between the year and the quarter of the year predictors.

Rate of Incidents Throughout the Day

```r
time_of_event <- incident_data %>%
    subset(select = c(OCCUR_TIME, OCCUR_DATE)) %>%
    transform(OCCUR_TIME = as.POSIXct(OCCUR_TIME), OCCUR_DATE = weekdays(OCCUR_DATE))

time_of_event2 <- transform(time_of_event, `Hour of Day` = cut(as.numeric(hour(OCCUR_TIME)) +
    1, breaks = seq(0, 24, 1), label = c(0:23)))
time_of_event2 <- ddply(time_of_event2, .(time_of_event2$`Hour of Day`,
    time_of_event2$OCCUR_DATE), nrow)
names(time_of_event2) <- c("Hour of Day", "Day of Week", "Count")

time_of_event2 <- transform(time_of_event2, `Hour of Day` = as.integer(`Hour of Day`) -
    1, `Day of Week` = as.factor(`Day of Week`))
time_of_event2
```

```
##     Hour of Day Day of Week Count
## 1             0      Friday   255
## 2             0      Monday   280
## 3             0    Saturday   534
```

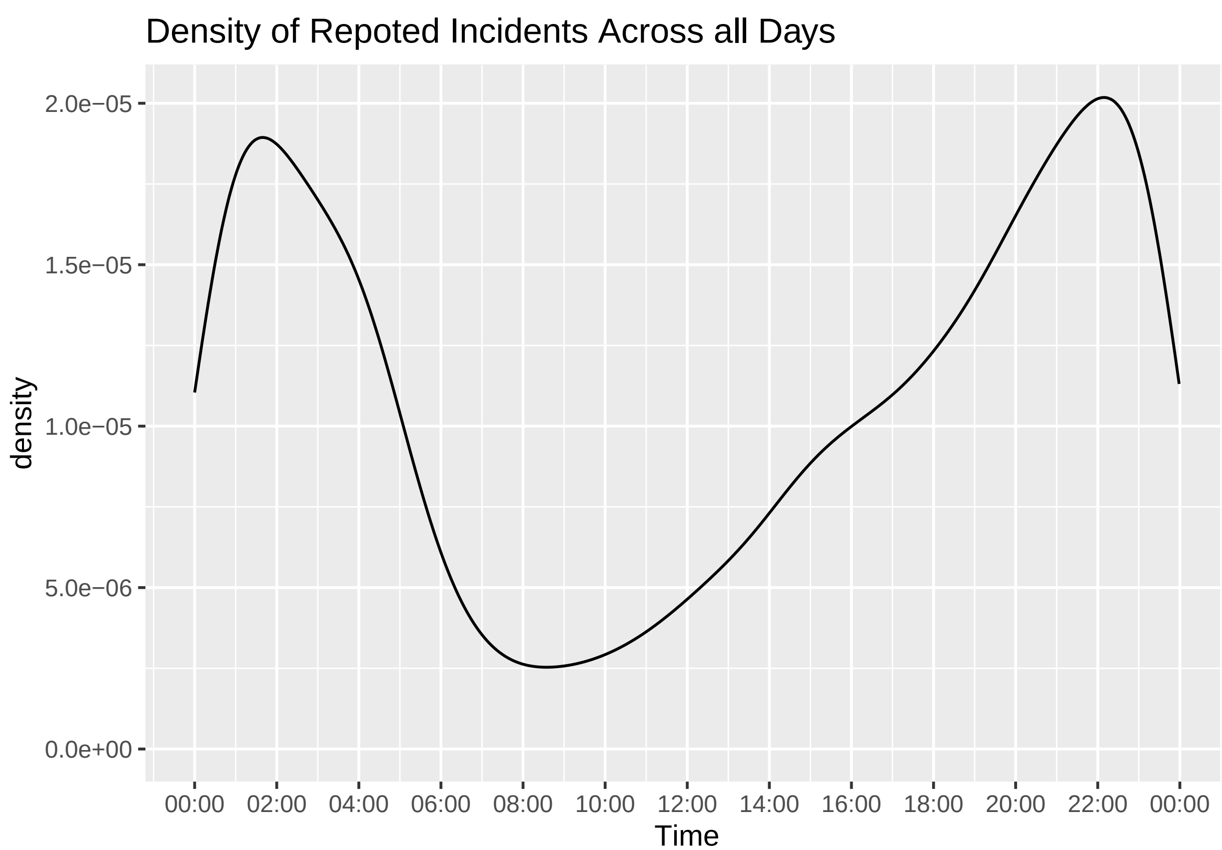
```
## 4              0    Sunday 555
## 5              0  Thursday 222
## 6              0   Tuesday 262
## 7              0 Wednesday 229
## 8              1    Friday 236
## 9              1    Monday 228
## 10             1  Saturday 512
## 11             1    Sunday 633
## 12             1  Thursday 204
## 13             1   Tuesday 217
## 14             1 Wednesday 188
## 15             2    Friday 184
## 16             2    Monday 206
## 17             2  Saturday 527
## 18             2    Sunday 552
## 19             2  Thursday 123
## 20             2   Tuesday 201
## 21             2 Wednesday 128
## 22             3    Friday 151
## 23             3    Monday 160
## 24             3  Saturday 509
## 25             3    Sunday 579
## 26             3  Thursday 104
## 27             3   Tuesday 129
## 28             3 Wednesday  95
## 29             4    Friday 139
## 30             4    Monday 174
## 31             4  Saturday 461
## 32             4    Sunday 511
## 33             4  Thursday  91
## 34             4   Tuesday  79
## 35             4 Wednesday  83
## 36             5    Friday  53
## 37             5    Monday  87
## 38             5  Saturday 214
## 39             5    Sunday 291
## 40             5  Thursday  28
## 41             5   Tuesday  46
## 42             5 Wednesday  51
## 43             6    Friday  32
## 44             6    Monday  38
## 45             6  Saturday 112
## 46             6    Sunday 131
## 47             6  Thursday  30
## 48             6   Tuesday  33
## 49             6 Wednesday  34
## 50             7    Friday  24
## 51             7    Monday  42
## 52             7  Saturday  58
## 53             7    Sunday  60
## 54             7  Thursday  23
## 55             7   Tuesday  23
## 56             7 Wednesday  24
## 57             8    Friday  31
```

```
## 58              8    Monday  49
## 59              8  Saturday  36
## 60              8    Sunday  48
## 61              8  Thursday  39
## 62              8   Tuesday  35
## 63              8 Wednesday  30
## 64              9    Friday  44
## 65              9    Monday  43
## 66              9  Saturday  25
## 67              9    Sunday  22
## 68              9  Thursday  41
## 69              9   Tuesday  45
## 70              9 Wednesday  34
## 71             10    Friday  40
## 72             10    Monday  48
## 73             10  Saturday  43
## 74             10    Sunday  53
## 75             10  Thursday  49
## 76             10   Tuesday  46
## 77             10 Wednesday  56
## 78             11    Friday  56
## 79             11    Monday  82
## 80             11  Saturday  67
## 81             11    Sunday  54
## 82             11  Thursday  58
## 83             11   Tuesday  64
## 84             11 Wednesday  53
## 85             12    Friday  64
## 86             12    Monday  84
## 87             12  Saturday  94
## 88             12    Sunday  76
## 89             12  Thursday  88
## 90             12   Tuesday  76
## 91             12 Wednesday  78
## 92             13    Friday  81
## 93             13    Monday 110
## 94             13  Saturday  94
## 95             13    Sunday  82
## 96             13  Thursday  97
## 97             13   Tuesday  89
## 98             13 Wednesday  90
## 99             14    Friday 162
## 100            14    Monday 139
## 101            14  Saturday 108
## 102            14    Sunday 143
## 103            14  Thursday 116
## 104            14   Tuesday 108
## 105            14 Wednesday 114
## 106            15    Friday 153
## 107            15    Monday 179
## 108            15  Saturday 135
## 109            15    Sunday 152
## 110            15  Thursday 123
## 111            15   Tuesday 131
```

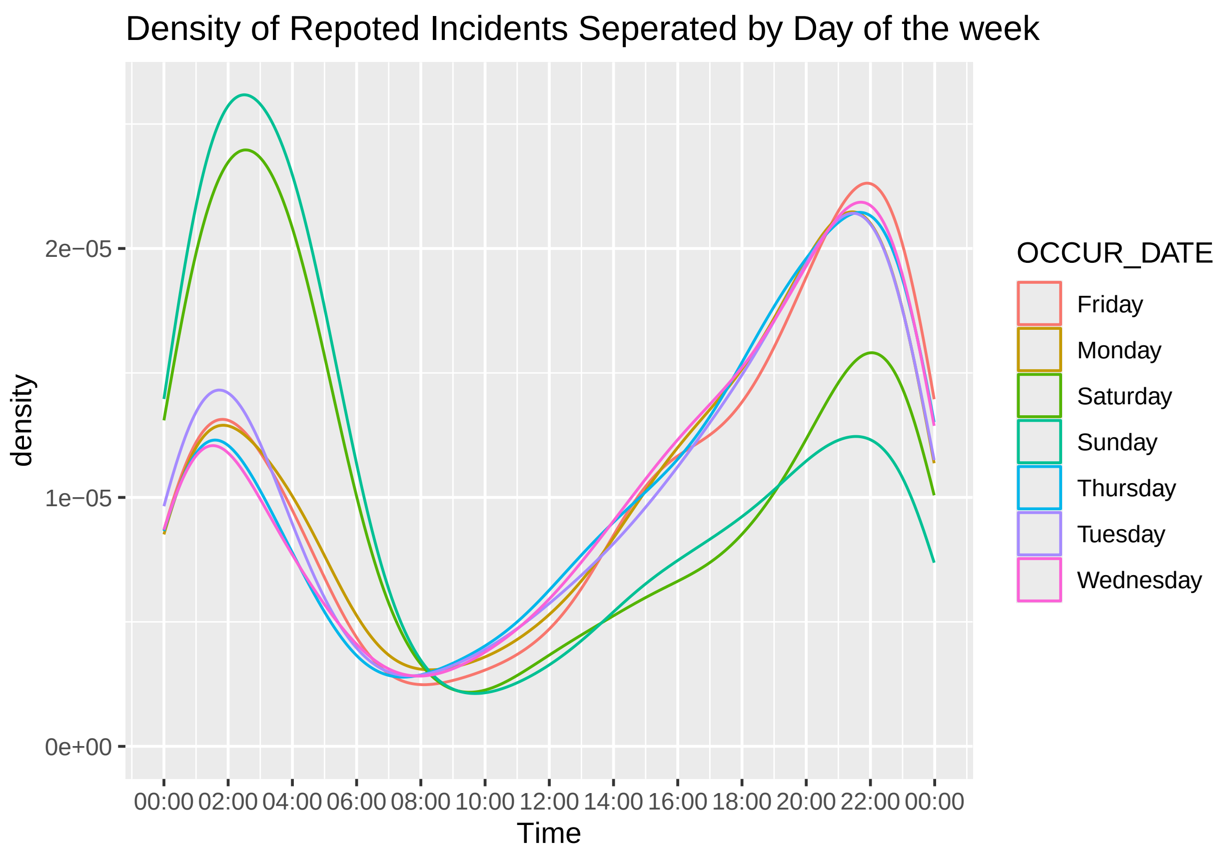
```
## 112          15  Wednesday  144
## 113          16     Friday  191
## 114          16     Monday  194
## 115          16   Saturday  146
## 116          16     Sunday  168
## 117          16   Thursday  139
## 118          16    Tuesday  156
## 119          16  Wednesday  162
## 120          17     Friday  155
## 121          17     Monday  223
## 122          17   Saturday  132
## 123          17     Sunday  183
## 124          17   Thursday  154
## 125          17    Tuesday  173
## 126          17  Wednesday  153
## 127          18     Friday  192
## 128          18     Monday  237
## 129          18   Saturday  191
## 130          18     Sunday  200
## 131          18   Thursday  218
## 132          18    Tuesday  176
## 133          18  Wednesday  188
## 134          19     Friday  244
## 135          19     Monday  268
## 136          19   Saturday  218
## 137          19     Sunday  210
## 138          19   Thursday  232
## 139          19    Tuesday  233
## 140          19  Wednesday  214
## 141          20     Friday  290
## 142          20     Monday  318
## 143          20   Saturday  234
## 144          20     Sunday  272
## 145          20   Thursday  236
## 146          20    Tuesday  256
## 147          20  Wednesday  239
## 148          21     Friday  316
## 149          21     Monday  353
## 150          21   Saturday  370
## 151          21     Sunday  284
## 152          21   Thursday  247
## 153          21    Tuesday  306
## 154          21  Wednesday  281
## 155          22     Friday  388
## 156          22     Monday  397
## 157          22   Saturday  391
## 158          22     Sunday  260
## 159          22   Thursday  310
## 160          22    Tuesday  305
## 161          22  Wednesday  297
## 162          23     Friday  420
## 163          23     Monday  335
## 164          23   Saturday  439
## 165          23     Sunday  353
```

```
## 166           23    Thursday   323
## 167           23     Tuesday   280
## 168           23   Wednesday   318
```

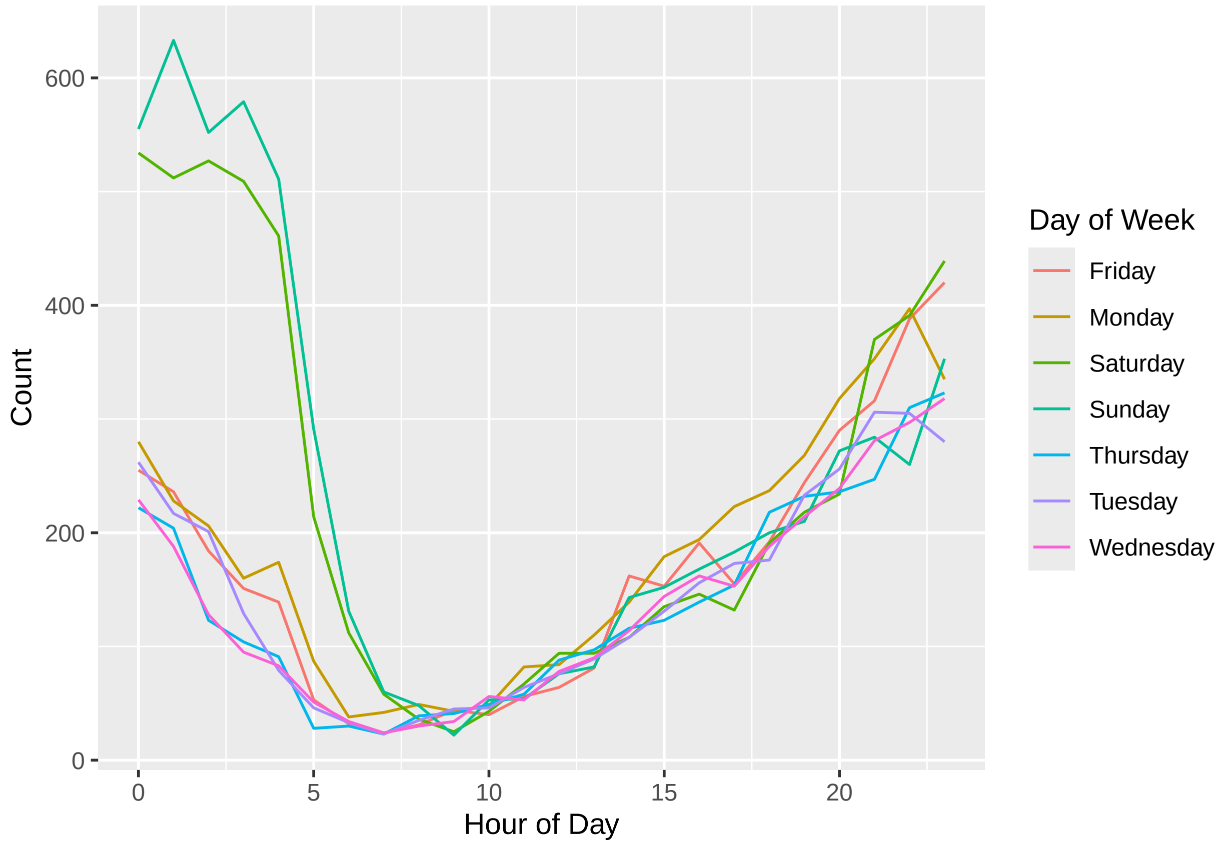
```
time_of_event %>%
    ggplot(aes(x = OCCUR_TIME)) + geom_density() + scale_x_datetime(breaks = date_breaks("2 hours"),
    labels = date_format("%H:%M")) + labs(title = "Density of Repoted Incidents Across all Days",
    x = "Time")
```




```
time_of_event %>%
    ggplot(aes(x = OCCUR_TIME, color = OCCUR_DATE)) + geom_density() +
    scale_x_datetime(breaks = date_breaks("2 hours"), labels = date_format("%H:%M")) +
    labs(title = "Density of Repoted Incidents Seperated by Day of the week",
        x = "Time")
```




```
ggplot(aes(y = Count, x = `Hour of Day`, color = `Day of Week`),
    data = time_of_event2) + geom_line()
```



It appears that there are 2 peaks at which shootings occur and that is around 10 at night and 1:30-2:00 in the morning. If I were to guess at the correlated events, that would be domestic events in the evening and intruder events in the early morning. Shootings occur the least during the morning and mid day, but there is an interesting increase during the mid afternoon of 2:00-4:00. The data needed to be able to correlate the events and the peaks would require there to be either a conviction or reason column for the data which gives what the type/motivation is for the shootings.

Conclusion

Based on the model of the data the number of incidents has decreased over time, there appears to be a roughly 2.28% decrease in the number of incidents every year. There is a statistical anomaly stating in 2020 which aligns with the 2020 global pandemic and may have skewed the average decease upwards towards a smaller decrease. There also appears to be some form of linking of the number of incidents to the quarter of the year, more research and analysis is required to determine what though.

There are a couple of times of day when the density of shooting incidents are high, in the early night/late afternoon, around 10:00pm, and again in the early morning, around 1:30am-2:00am. This could be linked to domestic incidents and intruder incidents respectively, but there is not data in the data set to be able to determine this connection. There also happens to be an increase in density in the early afternoon, 2:00pm-4:00pm, and I don't have any idea as to what could cause this, more research is need to be able to make connections on the correlation and causation of these points.

Bias

Sources of bias could be: - Survivor bias, as only know incidents would be reported - NAs, nulls, unknowns reduce the possible size of the data

Personal Bias - I expected shootings to occur more in the day then they did | I graphed the data about as raw as I could - I expected there to be a negative trend over time with the shootings | I graphed all the data before making any trends overall I went into this project with very little knowledge on the data before hand and the data I found interesting. - I have some ideas of when shooting should occur, e.g. any gang stuff during the day, domestic stuff when people get home, intruders during the night; but this may not be right and would need data to back it up. - I know that COVID lock down occurred in 2020 and that that forced people into close proxcimity to each other for far longer than they were accustom too and that cuase many problems.