NYPD Shooting Incident Data

Vaughn Pourchot

2023-10-16

NYPD Shooting Incident Data

Analyzing police department shoot data provides us with a vast amount of information that can be used for good. Any time data analysis can be used for good then it affects us all. If shooting data is analysed optimally then communities can operate more smoothly and efficiently. They can do this because they can feel safe if the police departments know and understand the trends in their area. The most important of which is identifying specific locations where the incident rate is abnormally high and having the police department and the city officials develop ways to make those areas safer that doesn't involve stationing half of the police department there.

What better place to analyse then New York City. The diversity of both the population and the environment makes for an ideal data trove to be examined. Hopefully, we can discover elements that transcend jurisdictions and can help city officials and police departments in other states, or even countries.

Import Data

Tidy Data - Remove Unwanted Variables

```
clean_data <- data %>%
  select(-Longitude,-Latitude,-INCIDENT_KEY,-STATISTICAL_MURDER_FLAG,
         -X_COORD_CD,-Y_COORD_CD,-Lon_Lat,-JURISDICTION_CODE,-PRECINCT)
```

Tidy Data - Combining like catagories with Variable

```
clean_data <- clean_data %>%
  mutate(
    PERP_SEX = case_when(
    PERP_SEX %in% c("", "(null)", "U") ~ "Unknown",
    TRUE ~ as.character(PERP_SEX)
  ),
  LOCATION_DESC = case_when(
    LOCATION_DESC %in% c("","(null)","Other") ~ "Other",
    TRUE ~ as.character(LOCATION_DESC)
  ),
  LOC_CLASSFCTN_DESC = case_when(
    LOC_CLASSFCTN_DESC %in% c("","OTHER","Other") ~ "Other",
    TRUE ~ as.character(LOC_CLASSFCTN_DESC)
  ),
```

```
  PERP_AGE_GROUP = case_when(
    PERP_AGE_GROUP %in% c("","(null)","UNKNOWN") ~ "Unknown",
    TRUE ~ as.character(PERP_AGE_GROUP)
  ),
  PERP_RACE = case_when(
    PERP_RACE %in% c("","(null)","UNKNOWN") ~ "Unkonwn",
    TRUE ~ as.character(PERP_RACE)
  ),
  VIC_AGE_GROUP = case_when(
    VIC_AGE_GROUP %in% c("1022","UNKNOWN") ~ "Unknown",
    TRUE ~ as.character(VIC_AGE_GROUP)
  ),
  LOC_OF_OCCUR_DESC = case_when(
    LOC_OF_OCCUR_DESC %in% c("") ~ "Unknown",
    TRUE ~ as.character(LOC_OF_OCCUR_DESC)
  )
  )
```

Tidy Data - Correct Data Types

```
library(lubridate)
clean_data$OCCUR_DATE <- mdy(clean_data$OCCUR_DATE)
clean_data$OCCUR_TIME <- hms(clean_data$OCCUR_TIME)
clean_data$BORO <- as.factor(clean_data$BORO)
clean_data$LOC_OF_OCCUR_DESC <- as.factor(clean_data$LOC_OF_OCCUR_DESC)
clean_data$LOC_CLASSFCTN_DESC <- as.factor(clean_data$LOC_CLASSFCTN_DESC)
clean_data$LOCATION_DESC <- as.factor(clean_data$LOCATION_DESC)
clean_data$PERP_AGE_GROUP <- as.factor(clean_data$PERP_AGE_GROUP)
clean_data$PERP_SEX <- as.factor(clean_data$PERP_SEX)
clean_data$PERP_RACE <- as.factor(clean_data$PERP_RACE)
clean_data$VIC_AGE_GROUP <- as.factor(clean_data$VIC_AGE_GROUP)
clean_data$VIC_SEX <- as.factor(clean_data$VIC_SEX)
clean_data$VIC_RACE <- as.factor(clean_data$VIC_RACE)
summary(clean_data)
```

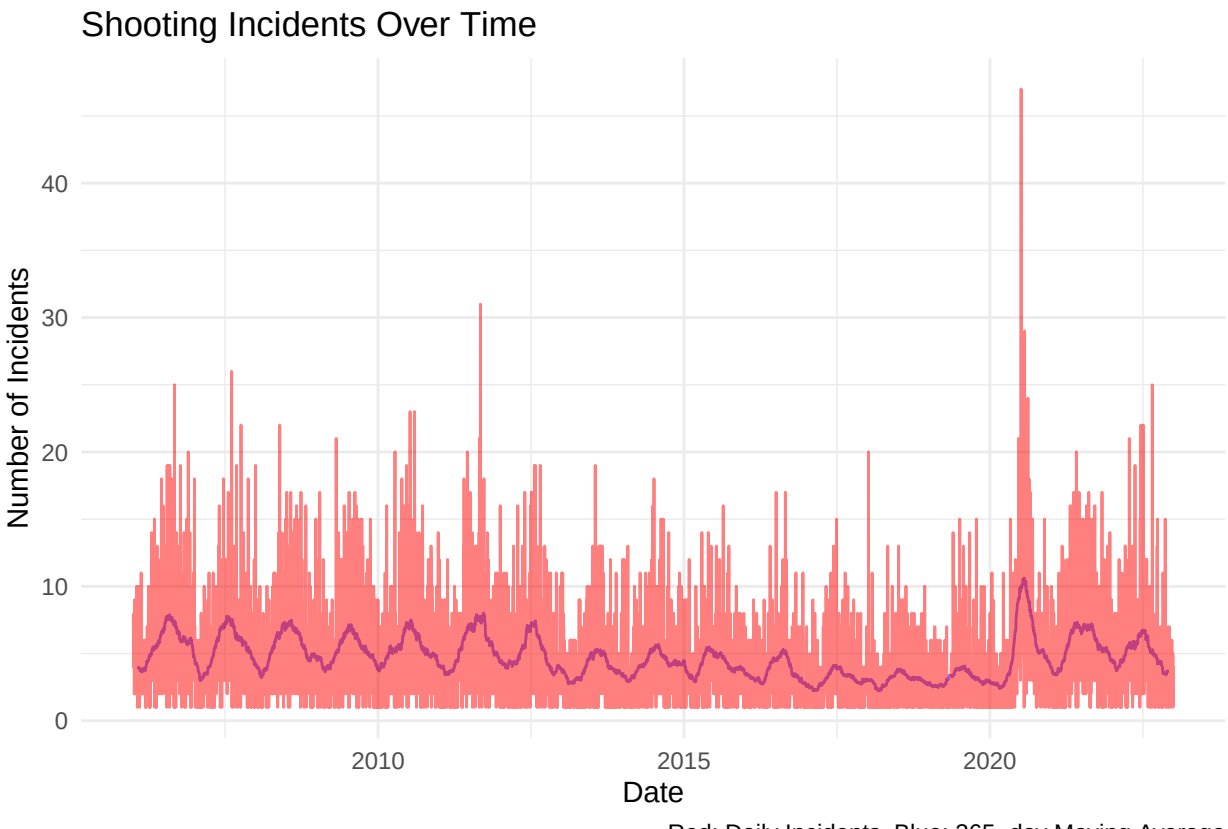
```
##    OCCUR_DATE           OCCUR_TIME                                    BORO
##  Min.   :2006-01-01   Min.   :0S                          BRONX        : 7937
##  1st Qu.:2009-07-18   1st Qu.:3H 27M 0S                   BROOKLYN     :10933
##  Median :2013-04-29   Median :15H 11M 0S                  MANHATTAN    : 3572
##  Mean   :2014-01-06   Mean   :12H 41M 31.7091388400731S   QUEENS       : 4094
##  3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S                  STATEN ISLAND:  776
##  Max.   :2022-12-31   Max.   :23H 59M 0S
##
##  LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC                  LOCATION_DESC
##  INSIDE :  242      Other     :25627    Other                    :15954
##  OUTSIDE: 1474      STREET    : 1103    MULTI DWELL - PUBLIC HOUS: 4832
##  Unknown:25596      HOUSING   :  280    MULTI DWELL - APT BUILD  : 2835
##                     DWELLING  :  127    PVT HOUSE                :  951
##                     COMMERCIAL:  100    GROCERY/BODEGA           :  694
##                     PLAYGROUND:   30    BAR/NIGHT CLUB           :  628
##                     (Other)   :   45    (Other)                  : 1418
```

```
##  PERP_AGE_GROUP     PERP_SEX                                PERP_RACE
##  Unknown:13132   F       :  424   AMERICAN INDIAN/ALASKAN NATIVE:    2
##  18-24  : 6222   M       :15439   ASIAN / PACIFIC ISLANDER      :  154
##  25-44  : 5687   Unknown:11449    BLACK                         :11432
##  <18    : 1591                    BLACK HISPANIC                : 1314
##  45-64  :  617                    Unkonwn                       :11786
##  65+    :   60                    WHITE                         :  283
##  (Other):    3                    WHITE HISPANIC                : 2341
##  VIC_AGE_GROUP   VIC_SEX                              VIC_RACE
##  <18    : 2839   F: 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10
##  18-24  :10086   M:24686   ASIAN / PACIFIC ISLANDER      :  404
##  25-44  :12281   U:   11   BLACK                         :19439
##  45-64  : 1863             BLACK HISPANIC                : 2646
##  65+    :  181             UNKNOWN                       :   66
##  Unknown:   62             WHITE                         :  698
##                            WHITE HISPANIC                : 4049
```

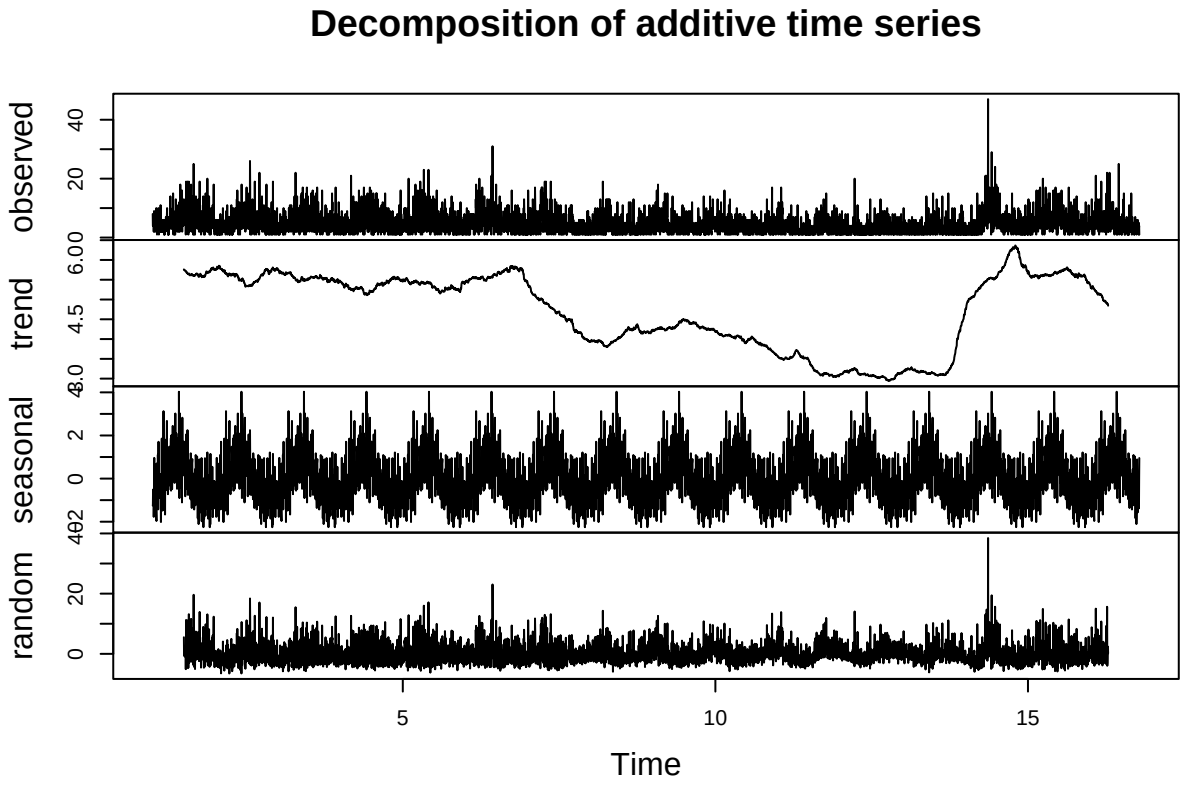
Visualizing the Data

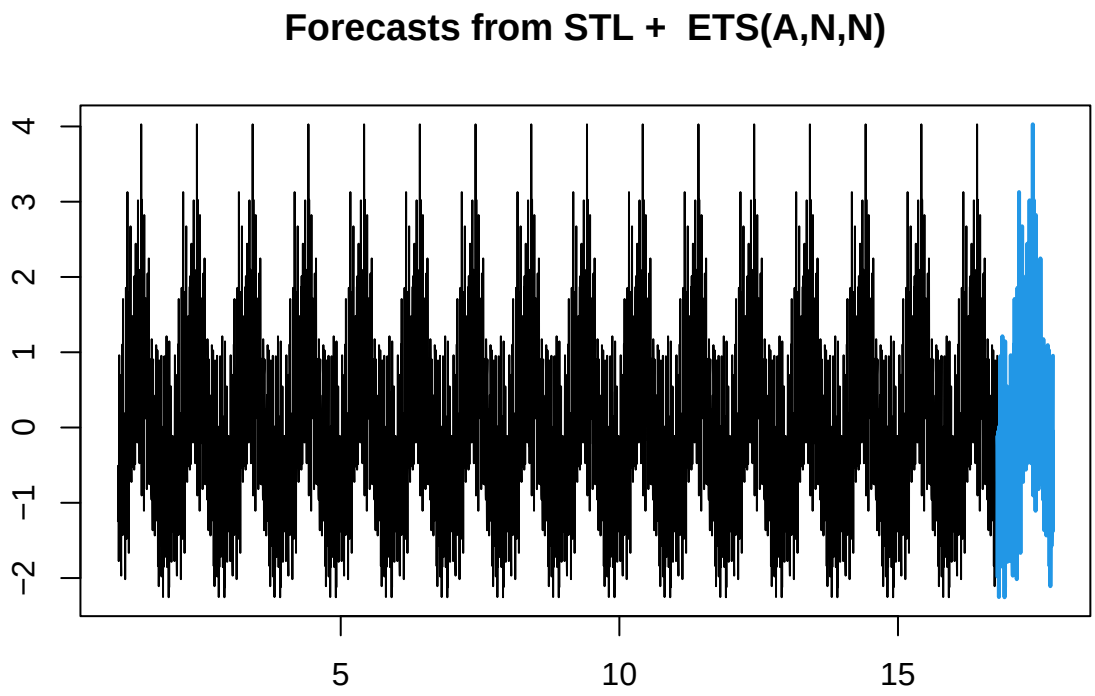
Annual Trend related to number of Daylight Hours

```
## Warning: Removed 59 rows containing missing values (`geom_line()`).
```



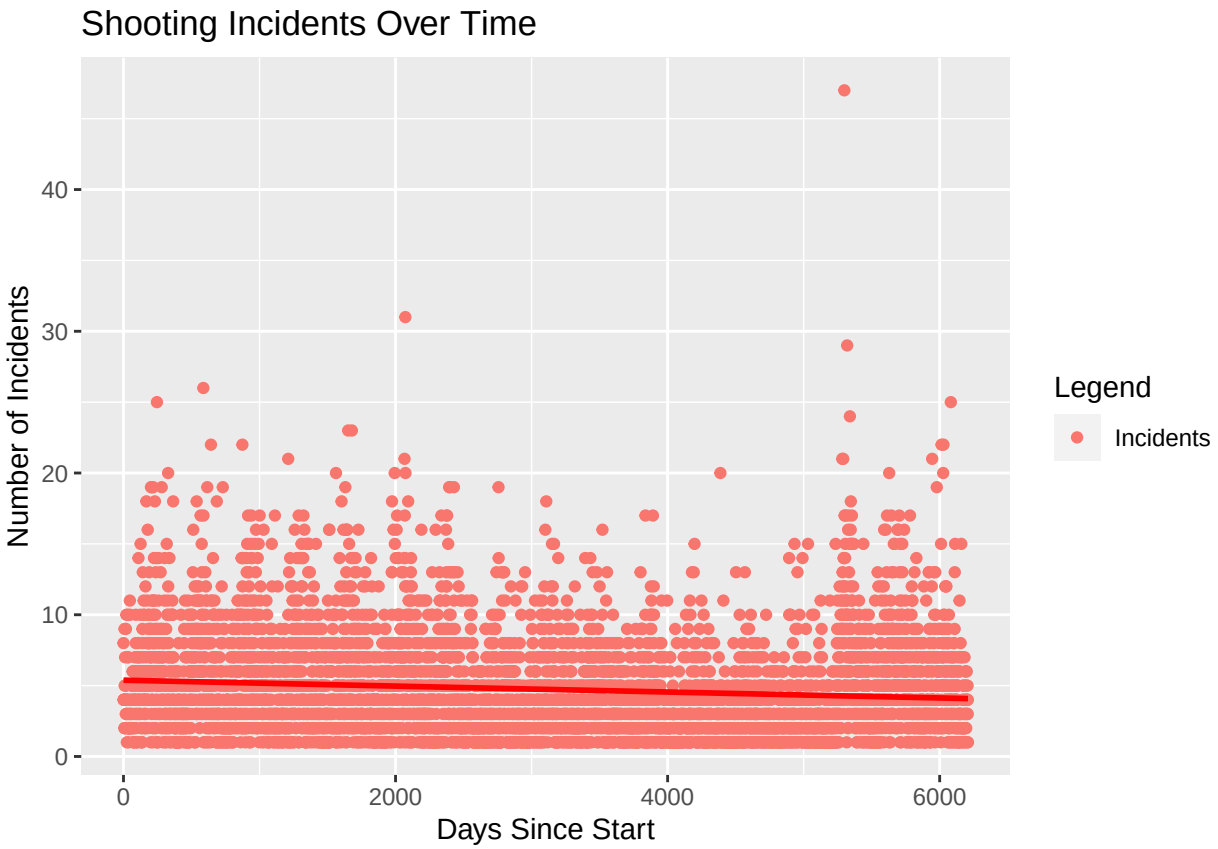
Red: Daily Incidents, Blue: 365−day Moving Average


Decomposition of additive time series
observed
trend
seasonal
random
Time
5
10
15




```
##
## Call:
## lm(formula = incidents ~ day_num, data = incidents_by_day)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.378 -2.500 -0.965  1.684 42.725
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.383e+00  9.242e-02  58.244  < 2e-16 ***
## day_num     -2.091e-04  2.592e-05  -8.064 8.88e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.559 on 5759 degrees of freedom
## Multiple R-squared:  0.01117,    Adjusted R-squared:  0.01099
## F-statistic: 65.03 on 1 and 5759 DF,  p-value: 8.884e-16

## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion

There is evidence of seasonality in the rate of shooting incidents throughout the year. However, the reason for the trend is yet to be determined. One could argue that there are more hours of daylight in the summer months which means more people are out later than other times of year. This thesis could be tested by analyzing the time of day for each incident but I would believe that most incidents occur in the cover of night. Some bias that I'm experiencing with this data is that I would prefer that the incidents are more nightly than in the daytime because I am someone who is rarely out at night and would thus prefer the daytime to be a safer portion of the day. You could also argue that the hotter temperatures cause more irritability in people which leads to more aggressive encounters. This theory would be difficult to test without employing someone with a psychology background. Ultimately, there is plenty of data here to comb through and find meaningful correlations. The seasonality of the incidents is only the tip of the iceberg.