NYPD Shooting Data Incident Analysis

iffig

2022-06-14

Objective

Through this data set I will be looking to answer if and how time plays a factor in these shooting incidents in New York City. Primarily I will try to answer the following questions:

- When historically have these shootings occurred during the day?
- On what day of the week are shootings most likely to occur?
- Does time of year affect the number of shootings?
- How have these trends shifted through the years?
- What is the overall trend in shooting incidents throughout the life of the data set?

Reviewing the Dataset

For this analysis the following the data was retrieved from: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

Raw Data Summary

The following `summary` and `glimpse` commands give a good overview of the columns of data and some general statistics on the data (max/min/mean/median/data_type/length/etc.).

```
summary(incident_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
```

```
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

```
glimpse(incident_data)
```

```
## Rows: 25,596
## Columns: 19
## $ INCIDENT_KEY            <int> 24050482, 77673979, 226950018, 237710987, 2247~
## $ OCCUR_DATE              <chr> "08/27/2006", "03/11/2011", "04/14/2021", "12/~
## $ OCCUR_TIME              <chr> "05:35:00", "12:03:00", "21:08:00", "19:30:00"~
## $ BORO                    <chr> "BRONX", "QUEENS", "BRONX", "BRONX", "MANHATTA~
## $ PRECINCT                <int> 52, 106, 42, 52, 34, 75, 32, 26, 41, 67, 43, 6~
## $ JURISDICTION_CODE       <int> 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0~
## $ LOCATION_DESC           <chr> "", "", "COMMERCIAL BLDG", "", "", "", "", "MU~
## $ STATISTICAL_MURDER_FLAG <chr> "true", "false", "true", "false", "false", "tr~
## $ PERP_AGE_GROUP          <chr> "", "", "", "", "", "25-44", "25-44", "", "25-~
## $ PERP_SEX                <chr> "", "", "", "", "", "M", "M", "", "M", "", "",~
## $ PERP_RACE               <chr> "", "", "", "", "", "BLACK HISPANIC", "BLACK",~
## $ VIC_AGE_GROUP           <chr> "25-44", "65+", "18-24", "25-44", "25-44", "25~
## $ VIC_SEX                 <chr> "F", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1017542, 1027543, 1009489, 1017440, 1005426, 1~
## $ Y_COORD_CD              <dbl> 255918.9, 186095.0, 243050.0, 256046.0, 254690~
## $ Latitude                <dbl> 40.86906, 40.67737, 40.83376, 40.86941, 40.865~
## $ Longitude               <dbl> -73.87963, -73.84392, -73.90880, -73.88000, -7~
## $ Lon_Lat                 <chr> "POINT (-73.87963173099996 40.86905819000003)"~
```

On initial glance it seems the data set provides us with an id, date/time, and location information for each incident. Each incident record also has information about the perpetrator and victims. It is important to note (per the footnotes), a single incident key can represent multiple victims, so the key can be duplicate.

The landing page and data footnotes PDF are helpful in providing additional information on the various columns in the data set:

- https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8
- https://data.cityofnewyork.us/api/views/833y-fsy8/files/e4e3d86c-348f-4a16-a17f-19480c089429?download=true&filename=NYPD_Shootings_Incident_Level_Data_Footnotes.pdf

Data Cleaning

Updating Variable Types

- Convert OCCUR_DATE from String to Date object

```
incident_data <- incident_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Removing Columns

Looking at the summary and glimpse of the data, it would likely make sense to drop most of the location data, except for `Boro` as the lat/lon and x/y coordinates would take more interpretation to become relevant.

- Lon_Lat
- Latitude
- Longitude
- X_COORD_CD
- Y_COORD_CD

```
incident_data <- incident_data %>% select(-c(Lon_Lat,X_COORD_CD,Y_COORD_CD, Latitude, Longitude))
```

Identifying Missing Data

In the glimpse above, notice a handful of variables have " " as the input to a field. To make this dataset more meaningful, updating those to something more consistent with the rest of the dataset would be helpful. The following code can be used to identify columns that had missing data:

```
colNames <- names(incident_data)
for (i in colNames){
  values <- unique(incident_data[[i]])
  missing <- "" %in% values
  if( missing == TRUE){
    print(i)
  }
}
```

```
## [1] "LOCATION_DESC"
## [1] "PERP_AGE_GROUP"
## [1] "PERP_SEX"
## [1] "PERP_RACE"
```

Handling Missing Data

The following code describes the possible entries for each of the columns with missing data:

```
unique(incident_data[["LOCATION_DESC"]])
```

```
##  [1] ""                          "COMMERCIAL BLDG"
##  [3] "MULTI DWELL - PUBLIC HOUS" "GROCERY/BODEGA"
##  [5] "MULTI DWELL - APT BUILD"   "BAR/NIGHT CLUB"
##  [7] "PVT HOUSE"                 "HOSPITAL"
##  [9] "HOTEL/MOTEL"               "GAS STATION"
## [11] "DEPT STORE"                "BEAUTY/NAIL SALON"
## [13] "RESTAURANT/DINER"          "BANK"
## [15] "FAST FOOD"                 "DRY CLEANER/LAUNDRY"
## [17] "NONE"                      "CLOTHING BOUTIQUE"
## [19] "SOCIAL CLUB/POLICY LOCATI" "SMALL MERCHANT"
## [21] "LIQUOR STORE"              "SUPERMARKET"
## [23] "SHOE STORE"                "SCHOOL"
## [25] "STORE UNCLASSIFIED"        "CHAIN STORE"
## [27] "DRUG STORE"                "TELECOMM. STORE"
## [29] "JEWELRY STORE"             "FACTORY/WAREHOUSE"
## [31] "CANDY STORE"               "VARIETY STORE"
## [33] "ATM"                       "GYM/FITNESS FACILITY"
## [35] "VIDEO STORE"               "DOCTOR/DENTIST"
## [37] "LOAN COMPANY"              "PHOTO/COPY STORE"
## [39] "CHECK CASH"                "STORAGE FACILITY"
```

```
unique(incident_data[["PERP_AGE_GROUP"]])
```

```
##  [1] ""        "25-44"   "18-24"   "<18"     "45-64"   "65+"     "UNKNOWN"
##  [8] "1020"    "940"     "224"
```

```
unique(incident_data[["PERP_SEX"]])
```

```
## [1] ""  "M" "F" "U"
```

```
unique(incident_data[["PERP_RACE"]])
```

```
## [1] ""                              "BLACK HISPANIC"
## [3] "BLACK"                         "WHITE HISPANIC"
## [5] "WHITE"                         "ASIAN / PACIFIC ISLANDER"
## [7] "UNKNOWN"                       "AMERICAN INDIAN/ALASKAN NATIVE"
```

To handle the missing points, here are the updates that will be made:

- LOCATION_DESC: This already has a NONE category, we can update the empty entries to be NONE
- PERP_AGE_GROUP: This already has an UNKNOWN category, we can update the empty entries to be UNKNOWN
- PERP_SEX:This already has an UNKNOWN (U) category, we can update the empty entries to be U
- PERP_RACE: This already has an UNKNOWN category, we can update the empty entries to be UNKNOWN

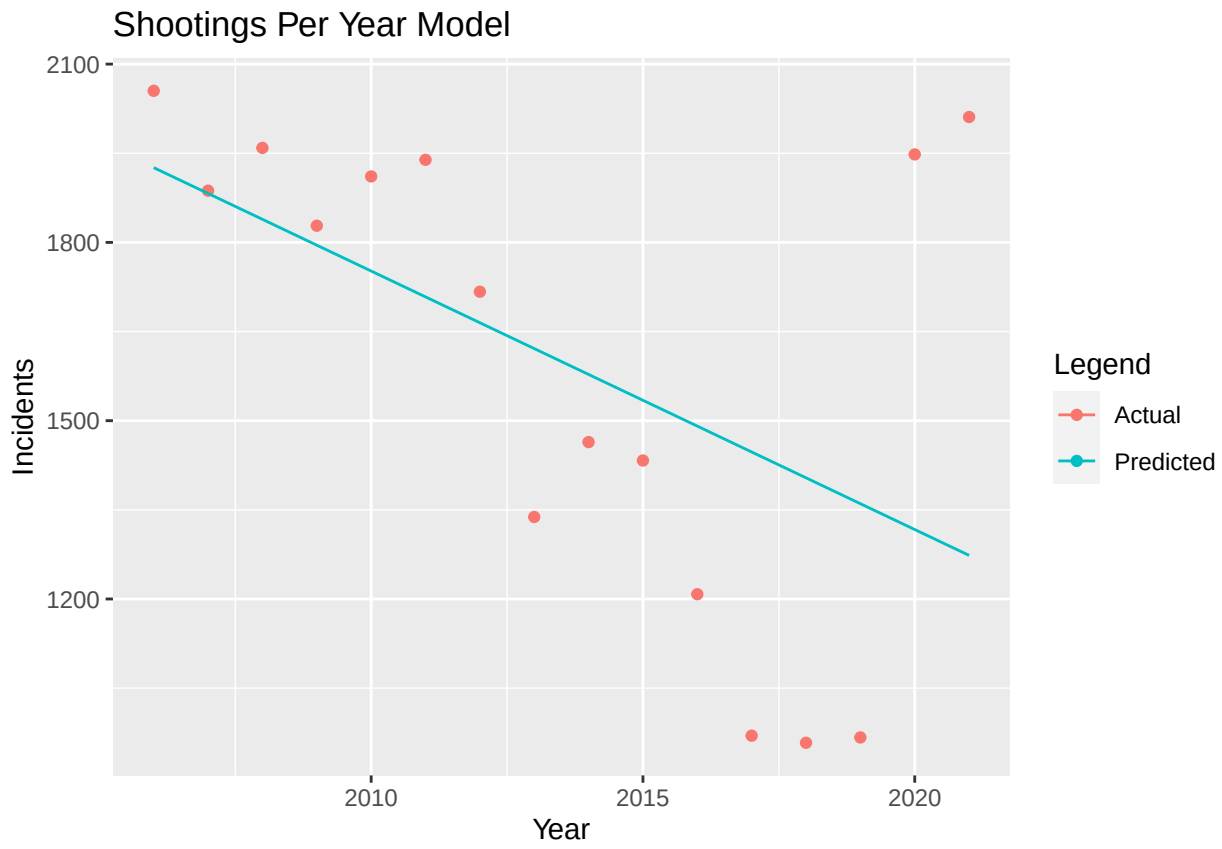
```r
incident_data <- incident_data %>%
  mutate(LOCATION_DESC = ifelse(LOCATION_DESC == "", "NONE", LOCATION_DESC)) %>%
  mutate(PERP_AGE_GROUP = ifelse(PERP_AGE_GROUP == "", "UNKNOWN", PERP_AGE_GROUP)) %>%
  mutate(PERP_SEX = ifelse(PERP_SEX == "", "U", PERP_SEX)) %>%
  mutate(PERP_RACE = ifelse(PERP_RACE == "", "UNKNOWN", PERP_RACE))

incident_data <- incident_data %>%
  filter(PERP_AGE_GROUP != "1020", PERP_AGE_GROUP != "940", PERP_AGE_GROUP != "224")
```

Now the data set has been updated to reflect more consistently when certain details about an incident are unknown. Because there is still useful information in these rows, they will remain in the data set.

Time Analysis

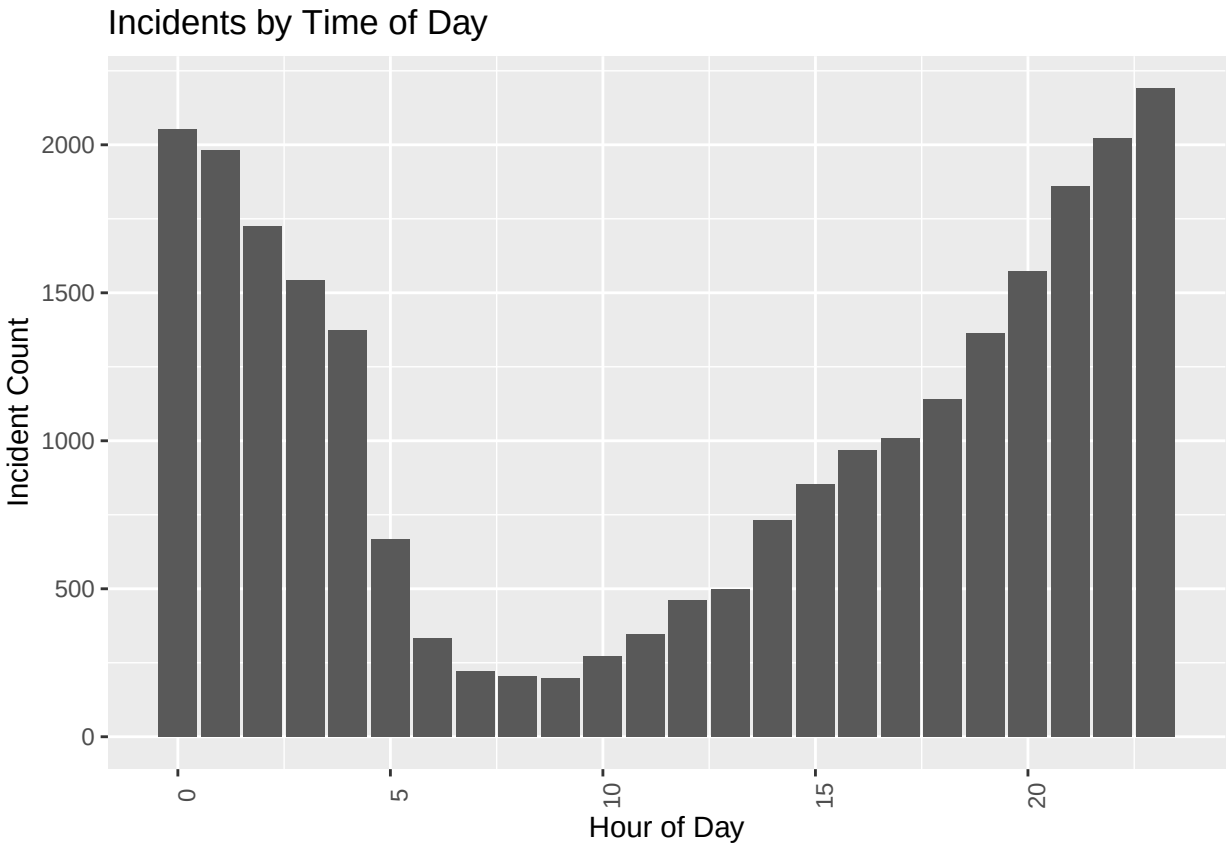
Yearly Trends



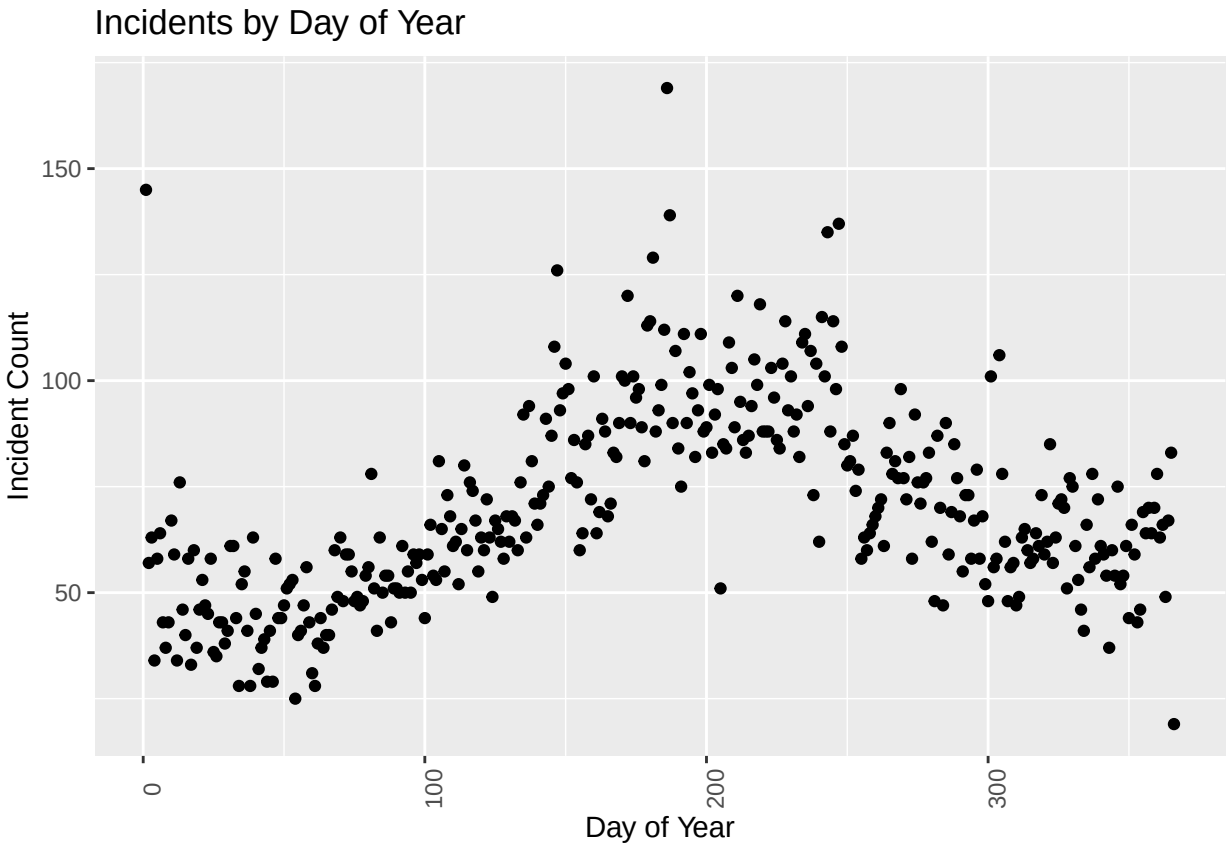

```
##
## Call:
## lm(formula = incidents ~ year, data = incidents_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -477.33 -282.95   18.71  136.72  737.65
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89177.93   39380.68   2.265   0.0399 *
## year          -43.50      19.56  -2.224   0.0431 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 360.6 on 14 degrees of freedom
## Multiple R-squared:  0.261,  Adjusted R-squared:  0.2083
## F-statistic: 4.946 on 1 and 14 DF,  p-value: 0.04312
```

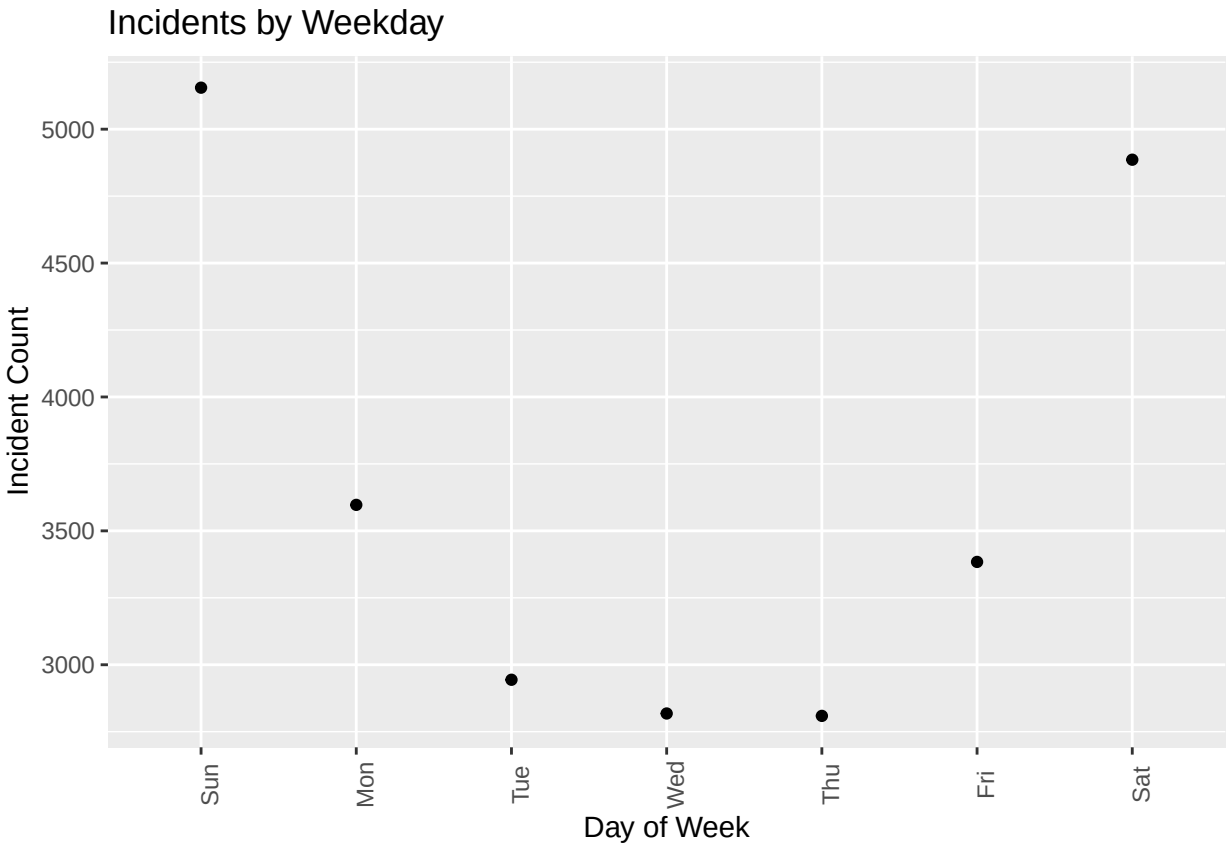
Time of Year



Time of Day


Incidents by Day of Year
Incident Count
150
100
50
0
100
200
300
Day of Year

Day of Week



Conclusions

In the above analysis there were several interesting findings. It does appear that time does have some influence over the number of incidents that occur. When looking at time of day, incidents tend to occur with more frequency in the early morning hours or in the later hours of the day. There is a steady decline from 12am to about 8am then steadily increases through the end of the day. Time of year seems to have a pattern as well. There is a steady arc of increasing incidents from January through about July, that begins to decline from July through the end of the year. Similarly, for days of the week, you see peaks on the weekend days and a steady decline as the week starts, and a steady rise heading into the weekend. I found it interesting that most of these data points have the similar arc patterns.

The other finding of interest was in looking at the trends of shooting incidents from 2006 to present. Rates of shootings has been on a pretty steady decline since the beginning of the data set. But in the past year, shootings nearly doubled, seeming to coincide with the onset of the COVID-19 pandemic.