NYPD Shooting Incident Data Project

EJDirenzo

2025-04-28

Import libraries (dependencies)

```
library( tidyverse )
library( lubridate )
```

Data

Description

The dataset used in this project was provided by the City of New York to the data.gov online catalog.

- Catalog url: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic
- Download url: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Title: NYPD Shooting Incident Data (Historic)
Updated: April 19, 2025
Description from data.gov:

> List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year . . . This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

Download

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

```
## # A tibble: 29,744 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
##  1    231974218 08/09/2021 01:06      BRONX    <NA>                    40
##  2    177934247 04/07/2018 19:48      BROOKLYN <NA>                    79
##  3    255028563 12/02/2022 22:57      BRONX    OUTSIDE                 47
##  4     25384540 11/19/2006 01:50      BROOKLYN <NA>                    66
##  5     72616285 05/09/2010 01:58      BRONX    <NA>                    46
##  6     85875439 07/22/2012 21:35      BRONX    <NA>                    42
##  7     79780323 07/12/2011 22:26      BROOKLYN <NA>                    71
##  8     85744504 07/14/2012 23:45      BROOKLYN <NA>                    69
##  9    142324890 04/21/2015 15:36      BROOKLYN <NA>                    75
## 10    152868707 05/07/2016 15:23      BROOKLYN <NA>                    69
## # i 29,734 more rows
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Tidy and transform

I have chosen to drop some columns:

- INCIDENT_KEY: administrative information not useful for this work
- LOC_OF_OCCUR_DESC: a lot of missing data, and what is present doesn't seem particularly useful
- JURISDICTION_CODE: administrative information not useful for this work, especially compared to other location data present
- LOC_CLASSFCTN_DESC: a lot of missing data, and what is present doesn't seem particularly useful
- LOCATION_DESC: a lot of missing data, and what is present doesn't seem particularly useful
- STATISTICAL_MURDER_FLAG: I could not find the footnotes on data.gov so I can't be sure exactly what this means
- X_COORD_CD: too precise to be useful for this work
- Y_COORD_CD: too precise to be useful for this work
- Latitude: too precise to be useful for this work
- Longitude: too precise to be useful for this work
- Lon_Lat: data type not useful for this work

The rest of the columns seemed valuable and were kept. The last operation was to typecast the date values to date objects. The other columns are okay in their given formats.

```
# drop INCIDENT_KEY
# keep OCCUR_DATE, OCCUR_TIME
# keep BORO maybe useful
# drop LOC_OF_OCCUR_DESC probably not useful
# keep PRECINCT maybe
# drop JURISDICTION_CODE
# drop LOC_CLASSFCTN_DESC, LOCATION_DESC not
# drop STATISTICAL_MURDER_FLAG
# keep PERP_AGE_GROUP, PERP_SEX, PERP_RACE
# keep VIC_AGE_GROUP, VIC_SEX, VIC_RACE
# drop X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat
tidy <- shoot %>% select(
```

```
  OCCUR_DATE,
  OCCUR_TIME,
  BORO,
  PRECINCT,
  # STATISTICAL_MURDER_FLAG,
  PERP_AGE_GROUP,
  PERP_SEX,
  PERP_RACE,
  VIC_AGE_GROUP,
  VIC_SEX,
  VIC_RACE
  ) %>% mutate( OCCUR_DATE = mdy( OCCUR_DATE ) )
# check
tidy
```

```
## # A tibble: 29,744 x 10
##    OCCUR_DATE OCCUR_TIME BORO     PRECINCT PERP_AGE_GROUP PERP_SEX PERP_RACE
##    <date>     <time>     <chr>       <dbl> <chr>          <chr>    <chr>
##  1 2021-08-09 01:06      BRONX          40 <NA>           <NA>     <NA>
##  2 2018-04-07 19:48      BROOKLYN       79 25-44          M        WHITE HISPAN~
##  3 2022-12-02 22:57      BRONX          47 (null)         (null)   (null)
##  4 2006-11-19 01:50      BROOKLYN       66 UNKNOWN        U        UNKNOWN
##  5 2010-05-09 01:58      BRONX          46 25-44          M        BLACK
##  6 2012-07-22 21:35      BRONX          42 18-24          M        BLACK
##  7 2011-07-12 22:26      BROOKLYN       71 <NA>           <NA>     <NA>
##  8 2012-07-14 23:45      BROOKLYN       69 <NA>           <NA>     <NA>
##  9 2015-04-21 15:36      BROOKLYN       75 25-44          M        BLACK
## 10 2016-05-07 15:23      BROOKLYN       69 18-24          M        BLACK
## # i 29,734 more rows
## # i 3 more variables: VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>
```

Missing data

There is missing data only under the PERP keys, which I think indicates shootings where the perpetrator is unknown (unsolved crimes). There is obviously no way to fill this data, but what data is known could be useful, and even the fact that some data is not known might be instructive if there seems to be a pattern in what crimes are solved vs unsolved. So, I have decided to keep these variables.

Analysis

Time

By year

```
byYear <- tidy %>%
  mutate( year=year(OCCUR_DATE) ) %>%
  group_by( year ) %>%
  summarize( count=n() )
#
byYear
```

```
## # A tibble: 19 x 2
##     year count
##    <dbl> <int>
##  1  2006  2055
##  2  2007  1887
##  3  2008  1959
##  4  2009  1828
##  5  2010  1912
##  6  2011  1939
##  7  2012  1717
##  8  2013  1339
##  9  2014  1464
## 10  2015  1434
## 11  2016  1208
## 12  2017   970
## 13  2018   958
## 14  2019   967
## 15  2020  1948
## 16  2021  2011
## 17  2022  1716
## 18  2023  1250
## 19  2024  1182
```

This is a good time to check the dates.

```
summary( tidy$OCCUR_DATE )
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2006-01-01" "2009-10-29" "2014-03-25" "2014-10-31" "2020-06-29" "2024-12-31"
```

The dates are as advertised.

By hour of the day

```
byHour <- tidy %>%
  mutate( hour=hour(OCCUR_TIME) ) %>%
  group_by( hour ) %>%
  summarize( count=n() )
#
byHour
```

```
## # A tibble: 24 x 2
##     hour count
##    <int> <int>
##  1     0  2337
##  2     1  2218
##  3     2  1921
##  4     3  1727
##  5     4  1538
##  6     5   770
##  7     6   410
```

```
##  8     7   254
##  9     8   268
## 10     9   254
## # i 14 more rows
```

Location

By borough

```
byBoro <- tidy %>%
  group_by( BORO ) %>%
  summarize( count=n() )
#
byBoro
```

```
## # A tibble: 5 x 2
##   BORO          count
##   <chr>         <int>
## 1 BRONX          8834
## 2 BROOKLYN      11685
## 3 MANHATTAN      3977
## 4 QUEENS         4426
## 5 STATEN ISLAND   822
```

Brooklyn and the Bronx have, by far, the greatest shooting incidents in the dataset. We don't have population data to make a pure comparison, but a quick Google search shows that the population of Brooklyn is very close to that of Queens, and the Bronx is similar to Manhattan. So this suggests the typical relationship between poverty and elevated crime.

By precinct

```
byPrecinct <- tidy %>%
  group_by( PRECINCT ) %>%
  summarize( count=n() )
#
byPrecinct
```

```
## # A tibble: 77 x 2
##    PRECINCT count
##       <dbl> <int>
##  1        1    29
##  2        5    74
##  3        6    29
##  4        7   127
##  5        9   128
##  6       10    76
##  7       13    64
##  8       14    69
```

```
##  9       17    10
## 10       18    48
## # i 67 more rows
```

Demographics

By perpetrator age

```
byPerpAge <- tidy %>%
  group_by( PERP_AGE_GROUP ) %>%
  summarize( count=n() )
#
byPerpAge
```

```
## # A tibble: 13 x 2
##    PERP_AGE_GROUP count
##    <chr>          <int>
##  1 (null)          1628
##  2 1020               1
##  3 1028               1
##  4 18-24           6630
##  5 2021               1
##  6 224                1
##  7 25-44           6342
##  8 45-64            775
##  9 65+               67
## 10 940                1
## 11 <18             1805
## 12 UNKNOWN         3148
## 13 <NA>            9344
```

Here we see there are some potential bad data rows, with age groups which seem to be incorrectly formatted. It isn't readily apparent how to correct the rows, because a few can be interpreted in slightly different ways which would put them into different categories. It does seem clear that the categories 18-24, 25-44, 45-64, and 65+ are correct. It also seems likely that the value (null) is correct because of the large amount of data classified as such, but I am unable to find the data footnotes promised online so I can only speculate that these are shootings commited by minors. But speculation isn't reliable, so we have to ignore the value (null).

By perpetrator sex

```
byPerpSex <- tidy %>%
  group_by( PERP_SEX ) %>%
  summarize( count=n() )
#
byPerpSex
```

```
## # A tibble: 5 x 2
##   PERP_SEX count
##   <chr>    <int>
## 1 (null)    1628
## 2 F          461
## 3 M        16845
## 4 U         1500
## 5 <NA>      9310
```

By perpetrator race

```
byPerpRace <- tidy %>%
  group_by( PERP_RACE ) %>%
  summarize( count=n() )
#
byPerpRace
```

```
## # A tibble: 9 x 2
##   PERP_RACE                       count
##   <chr>                           <int>
## 1 (null)                           1628
## 2 AMERICAN INDIAN/ALASKAN NATIVE      2
## 3 ASIAN / PACIFIC ISLANDER          184
## 4 BLACK                           12323
## 5 BLACK HISPANIC                   1487
## 6 UNKNOWN                          1838
## 7 WHITE                             305
## 8 WHITE HISPANIC                   2667
## 9 <NA>                             9310
```

By victim age

```
byVicAge <- tidy %>%
  group_by( VIC_AGE_GROUP ) %>%
  summarize( count=n() )
#
byVicAge
```

```
## # A tibble: 7 x 2
##   VIC_AGE_GROUP count
##   <chr>         <int>
## 1 1022              1
## 2 18-24         10677
## 3 25-44         13563
## 4 45-64          2118
## 5 65+             236
## 6 <18            3081
## 7 UNKNOWN          68
```

There is only one potential bad data row this time. And this time minors have an obvious classification: <18.

By victim sex

```
byVicSex <- tidy %>%
  group_by( VIC_SEX ) %>%
  summarize( count=n() )
#
byVicSex
```

```
## # A tibble: 3 x 2
##   VIC_SEX count
##   <chr>   <int>
## 1 F        2891
## 2 M       26841
## 3 U          12
```

By victim race

```
byVicRace <- tidy %>%
  group_by( VIC_RACE ) %>%
  summarize( count=n() )
#
byVicRace
```

```
## # A tibble: 7 x 2
##   VIC_RACE                         count
##   <chr>                            <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE      13
## 2 ASIAN / PACIFIC ISLANDER           478
## 3 BLACK                            20999
## 4 BLACK HISPANIC                    2930
## 5 UNKNOWN                             72
## 6 WHITE                              741
## 7 WHITE HISPANIC                    4511
```

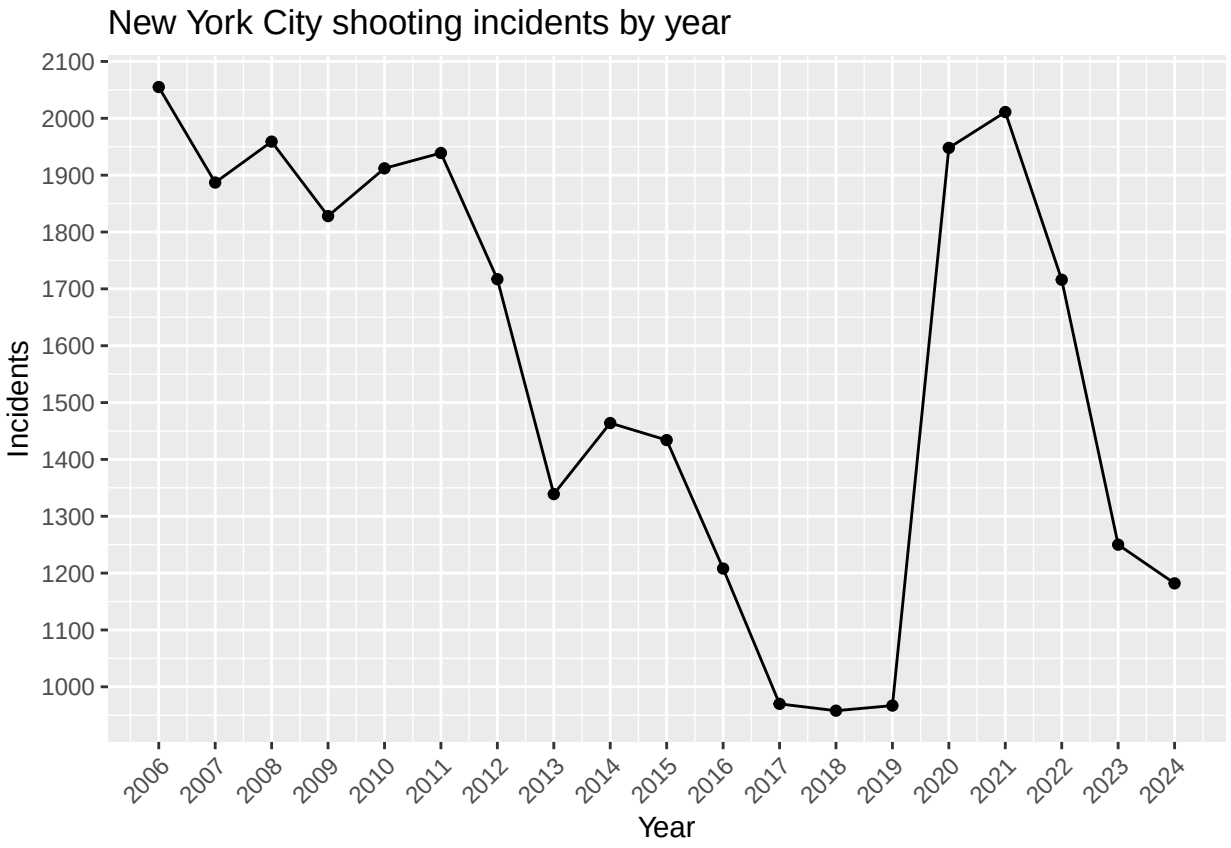
Visualization

By year, timeseries

A timeseries is the first and simplest idea for analysis and visualization, but it's always first for a reason.

```
byYear %>%
  ggplot( aes( x=year, y=count ) ) +
  geom_line( color="black" ) +
```

```
  geom_point( color="black" ) +
  scale_x_continuous( breaks=byYear$year ) +
  scale_y_continuous( breaks=seq(900,2100,100) ) +
  theme( axis.text.x=element_text( angle=45, hjust=1 ) ) +
  labs(
    title="New York City shooting incidents by year",
    x="Year",
    y="Incidents"
    )
```

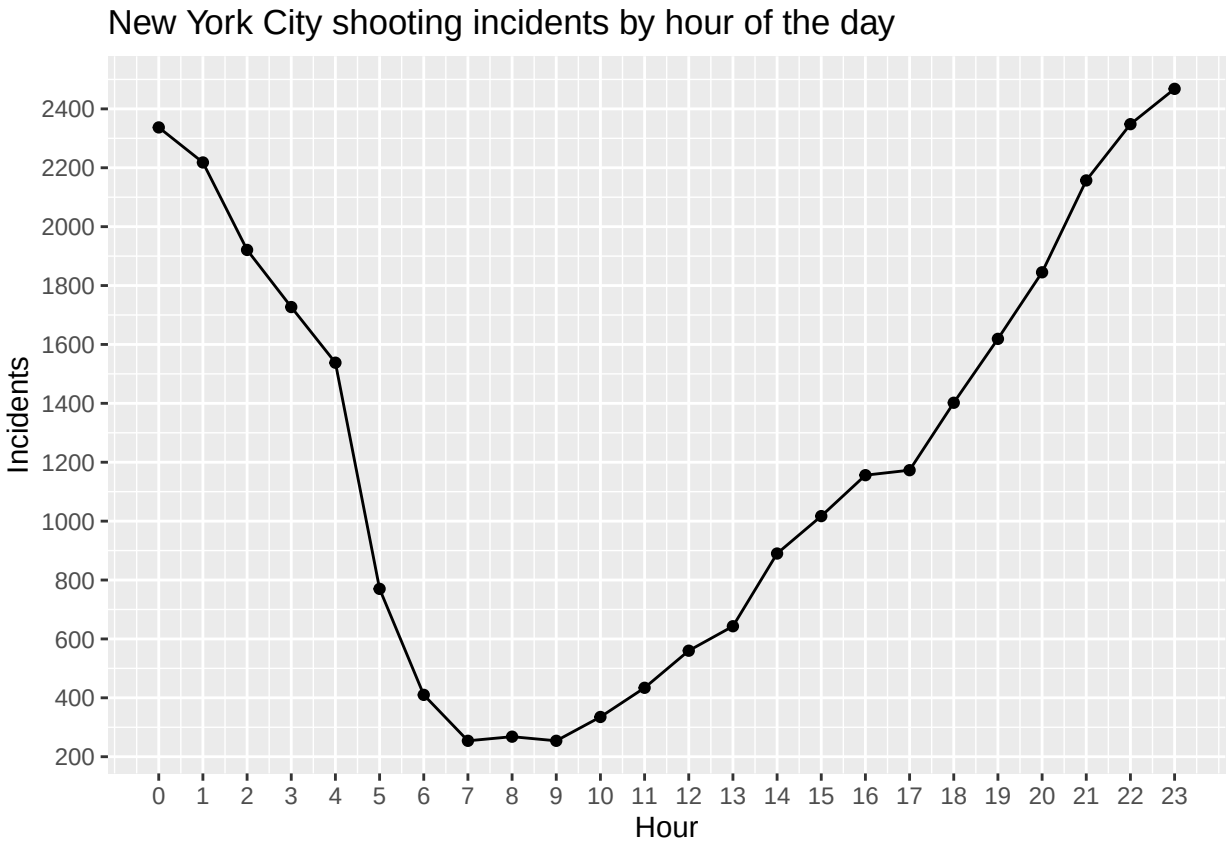


The plot actually does immediately invoke a question: what happened in 2019-2020? It is clear that shootings in New York are on a steady down trend, hitting a low in 2018, then suddenly they spike, doubling back to the levels of 2006. And in only a few years, the numbers have dropped back down almost a sharply. I don't think that addressing this question thoroughly is within the scope of this project, but if it were then I would definitely look for a link to the COVID-19 pandemic. Mass layoffs and unemployment led to a surge of crime in much of the nation, as many desperate people felt they were left with few other options to provide for themselves and their families.

By hour of the day

The next visualization that I want to do it shootings by hour of the day. Again, maybe a little obvious but worth doing.

```r
byHour %>%
  ggplot( aes( x=hour, y=count ) ) +
  geom_point( color="black" ) +
  geom_line( color="black" ) +
  scale_x_continuous( breaks=byHour$hour ) +
  scale_y_continuous( breaks=seq(200,2600,200) ) +
  theme( axis.text.x=element_text( angle=0, hjust=0.5 ) ) +
  labs(
    title="New York City shooting incidents by hour of the day",
    x="Hour",
    y="Incidents"
    )
```



The pattern is not very suprising. Shootings seem to peak around midnight and are at their lowest rate in the early morning. I'll try to do my model on this, it looks cubic.

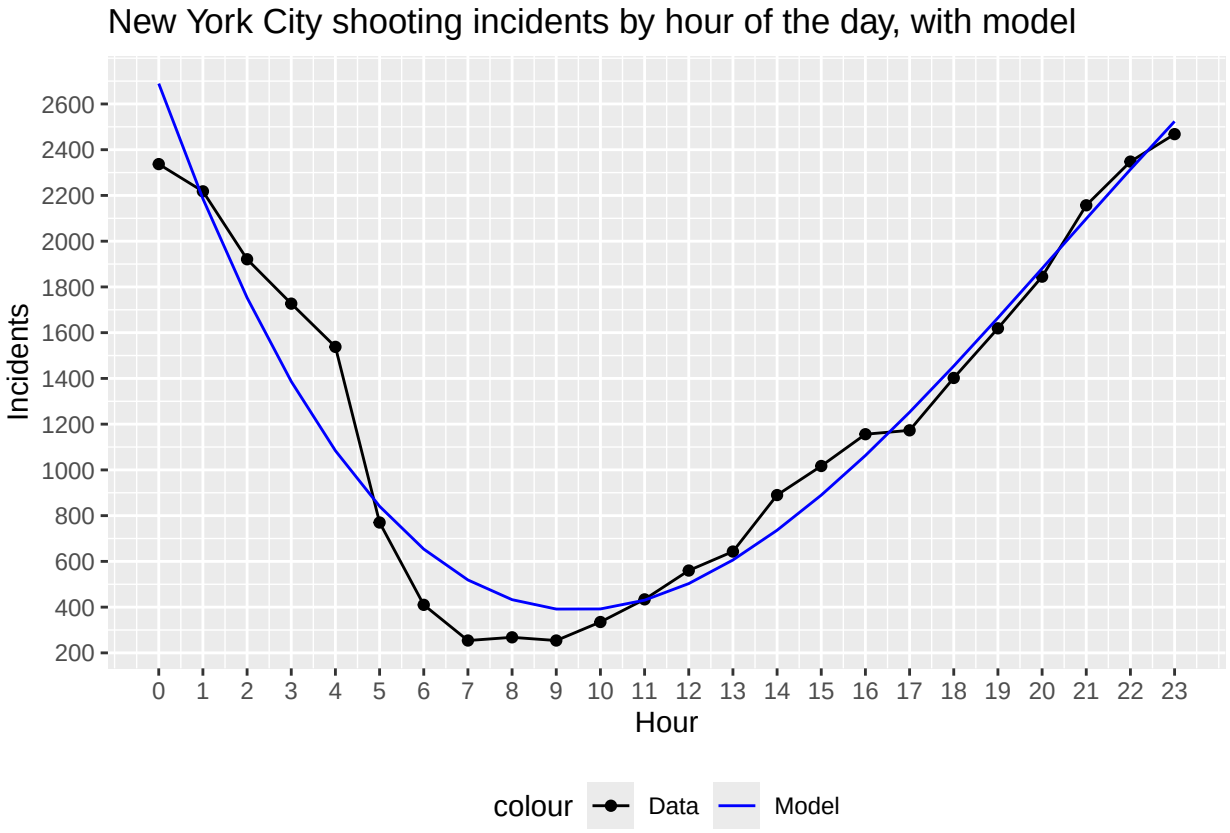
Modeling incidents by hour of the day

```r
# cubic model
mod <- lm( count ~ hour + I(hour^2) + I(hour^3), data=byHour )
# check summary
#summary( mod )
# add to new df
byHourWPred <- byHour %>% mutate( pred=predict(mod) )
```

```r
# check
byHourWPred
```

```
## # A tibble: 24 x 3
##     hour count  pred
##    <int> <int> <dbl>
##  1     0  2337 2689.
##  2     1  2218 2186.
##  3     2  1921 1753.
##  4     3  1727 1387.
##  5     4  1538 1084.
##  6     5   770  841.
##  7     6   410  654.
##  8     7   254  519.
##  9     8   268  433.
## 10     9   254  392.
## # i 14 more rows
```

```r
# visualize
byHourWPred %>% ggplot() +
  geom_point( aes(x=hour, y=count, color="Data" ) ) +
  geom_line( aes(x=hour, y=count, color="Data" ) ) +
  geom_line( aes(x=hour, y=pred, color="Model") ) +
  scale_color_manual( values=c("Data"="black","Model"="blue") ) +
  scale_x_continuous( breaks=byHour$hour ) +
  scale_y_continuous( breaks=seq(200,2600,200) ) +
  theme(
    legend.position="bottom",
    axis.text.x = element_text( angle=0, hjust=0.5 ) ) +
  labs(
    title="New York City shooting incidents by hour of the day, with model",
    x="Hour",
    y="Incidents"
    )
```



The cubic model appears to fit the data well. We can output details to support this observation:

```
# check summary
summary( mod )
```

```
##
## Call:
## lm(formula = count ~ hour + I(hour^2) + I(hour^3), data = byHour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -352.09  -72.95  -16.05   67.31  453.96
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2689.0866   134.6676  19.968 1.11e-14 ***
## hour        -539.7964    51.8075 -10.419 1.58e-09 ***
## I(hour^2)     37.0499     5.3001   6.990 8.77e-07 ***
## I(hour^3)     -0.6040     0.1513  -3.992 0.000717 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 191.3 on 20 degrees of freedom
## Multiple R-squared:  0.9451, Adjusted R-squared:  0.9369
## F-statistic: 114.7 on 3 and 20 DF,  p-value: 8.989e-13
```

Conclusion and bias discussion

The data shows a significant and sudden increase in shootings around 2019-2020 and a clear seasonality in shootings based on time of day.

Bias and bias mitigation

I already demonstrated some bias by speculating that the reason for the spike in shootings in 2019-2020 was likely a result of pandemic related poverty. I feel that this is not a bad hypothesis but we would need a lot more and different data to truly investigate that.

Another area in which I am biased concerns questions of perpetrator and victim demographics. Outside of superficial data exploration, I completely ignored them. These are areas of extreme controversy which many have strong opinions about. In my opinion, there are logical and well evidenced arguments concerning systemic bias affecting every aspect related to these incidents, from the underlying causes of the crimes to enforcement of the law and eventually reporting. This dataset is itself likely biased for those reasons, and we do not have enough data here for me to begin to make any kind of reasonable analysis using that part of the dataset. So, I chose to mitigate my bias on this topic by simply avoiding it and focusing elsewhere.

Session info

```
sessionInfo()
```

```
## R version 4.5.0 (2025-04-11)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3;  LAPACK version 3.9.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Chicago
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.4     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.2   tidyverse_2.0.0
##
```

```
## loaded via a namespace (and not attached):
##  [1] bit_4.6.0          gtable_0.3.6       crayon_1.5.3       compiler_4.5.0
##  [5] tidyselect_1.2.1   parallel_4.5.0     scales_1.4.0       yaml_2.3.10
##  [9] fastmap_1.2.0      R6_2.6.1           generics_0.1.3     curl_6.2.2
## [13] knitr_1.50         pillar_1.10.2      RColorBrewer_1.1-3 tzdb_0.5.0
## [17] rlang_1.1.6        utf8_1.2.4         stringi_1.8.7      xfun_0.52
## [21] bit64_4.6.0-1      timechange_0.3.0   cli_3.6.4          withr_3.0.2
## [25] magrittr_2.0.3     digest_0.6.37      grid_4.5.0         vroom_1.6.5
## [29] rstudioapi_0.17.1  hms_1.1.3          lifecycle_1.0.4    vctrs_0.6.5
## [33] evaluate_1.0.3     glue_1.8.0         farver_2.1.2       rmarkdown_2.29
## [37] tools_4.5.0        pkgconfig_2.0.3    htmltools_0.5.8.1
```