NYPD Shooting's Analysis

K.Smith

2/3/2022

"Effect's of the Pandemic on Murder/Shooting's in New York City"

Data

- The data for this analysis was sourced from the reports supplied by the NYPD.
- I Used the historical data and combined it with the year to date dataset.
  - The Historical data set can be found here.
  - The Year-to-date data set can be found here.

```
## Get Data needed
url_in <- "https://data.cityofnewyork.us/api/views/"
## This can be a vector of different csv files that start with the above url
file_names <- c("833y-fsy8/rows.csv?accessType=DOWNLOAD",
                "5ucz-vwe8/rows.csv?accessType=DOWNLOAD")

## Concantenate them together
urls <- str_c(url_in, file_names)
urls
```

```
## [1] "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
## [2] "https://data.cityofnewyork.us/api/views/5ucz-vwe8/rows.csv?accessType=DOWNLOAD"
```

Preview Data

- After scraping the data we can format it from csv to a DataFrame

```
################ FOR HISTORICAL ANALYSIS ################
## Create DataFrame from first file in urls for historical shootings
shootings_historical <- read_csv(urls[1])
```

```
## Rows: 23585 Columns: 19
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#shootings <- shootings_hist
shootings_historical <- shootings_historical[order(shootings_historical$OCCUR_DATE), ]
tail(shootings_historical)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>        <dbl>              <dbl>
## 1    206890929 12/31/2019 23:15      MANHATTAN       28                  0
## 2    206891917 12/31/2019 20:14      BROOKLYN        73                  0
## 3    222446417 12/31/2020 00:42      BRONX           44                  0
## 4    222468112 12/31/2020 14:59      QUEENS         103                  0
## 5    222466833 12/31/2020 19:27      QUEENS         113                  0
## 6    222473262 12/31/2020 23:45      MANHATTAN       33                  0
## # ... with 13 more variables: 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>
```

- Missing Data
  - Since one of the questions I am interested in is the impact of the pandemic on shootings, we can see from above that the historical data will not be sufficient to answer these questions.
  - I will be scraping a second data set from the same repository provided by the NYPD, and combining that with the historical data.

```
################ FOR CURRENT ANALYSIS ################
## Create DataFrame from second file in urls for current shootings
shootings_current <- read_csv(urls[2])
```

```
## Rows: 2011 Columns: 19
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#shootings <- shootings_current

################ FOR HISTORICAL THROUGH CURRENT ANALYSIS ################
## If we want to merge historical and current choose this path:
# Remove different columns that we don't need
shootings_hist <- shootings_historical %>%
  select(-c(Lon_Lat))

shootings_current <- shootings_current %>%
```

```
  select(-c("New Georeferenced Column"))

## Now they will play nice together when we join them
shootings <- rbind(shootings_hist, shootings_current)

## Preview Data
sample_n(shootings, 13)
```

```
## # A tibble: 13 x 18
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>        <dbl>              <dbl>
##  1     93053470 10/12/2013 23:08      BROOKLYN        67                  0
##  2     81042175 10/05/2011 15:00      QUEENS         114                  0
##  3     90215685 04/20/2013 04:09      MANHATTAN       34                  0
##  4     34720461 09/30/2007 00:03      BRONX           49                  0
##  5    236817419 11/22/2021 01:18      BRONX           52                  0
##  6     80399739 08/21/2011 23:10      BROOKLYN        67                  0
##  7     32914993 07/05/2007 01:27      BRONX           47                  2
##  8     92941489 10/04/2013 15:47      BRONX           47                  0
##  9     94337762 01/07/2014 09:40      MANHATTAN       25                  0
## 10    214925555 07/03/2020 00:25      BROOKLYN        79                  0
## 11     59660231 03/16/2009 23:55      QUEENS         113                  0
## 12    192259533 01/13/2019 06:00      BROOKLYN        70                  0
## 13     35890072 11/22/2007 18:01      BROOKLYN        73                  0
## # ... with 12 more variables: 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>
```

What variables are in the data?

- INCIDENT_KEY
  - Assuming this is some sort of record keeping identifier variable.
- OCCUR_DATE
  - Date of the shooting incident.
- OCCUR_TIME
  - Time of the shooting incident.
- BORO
  - New York City Borough where the shootings took place.
- PRECINCT
  - The responding New York City Police department precinct identifier.
- JURISDICTION_CODE
  - Code indentifying which jurisdiction the incidents occurred in.
- LOCATION_DESC
  - A brief description of the building/environment type where the incident occurred.
- STATISTICAL_MURDER_FLAG

  - Indicator variable for which shootings resulted in murder:TRUE for a fatal incident, FALSE for no fatality recorded.
- VIC_AGE_GROUP
  - Age range for victims of incident's.
- VIC_SEX
  - Sex of victims.
- VIC_RACE
  - Race of victims.
- X_COORD_
  - Geolocation coordinates identifier(function of Lat and Long).
- Y_COORD_
  - Geolocation coordinates identifier(function of Lat and Long).
- PERP_AGE_GROUP
  - Age range alleged of perpetrators.
- PERP_SEX
  - Sex of alleged of perpetrators.
- PERP_RACE
  - Race of alleged of perpetrators.
- Longitude
  - Longitudinal geographic coordinate where incident took place.
- Latitude
  - Latitudinal geographic coordinate where incident took place.

Cleaning Data

- After cleaning the DataFrame by removing some columns I was not interested in, and reformatting the OCCURR_DATE column to a date object and the values to datetime objects, and then sorting the dataframe in chronological order by this columns. It looks like this:

```
## Downsize to just the variables of interest (to me),
## and reformat OCCUR_DATE column values from a chr to date object
shootings <- shootings %>%
  mutate(shootings, OCCUR_DATE= as.Date(OCCUR_DATE, format= "%m/%d/%Y")) %>%
  select(c(OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG, OCCUR_TIME,
           VIC_AGE_GROUP, VIC_SEX, VIC_RACE, PERP_AGE_GROUP, PERP_SEX, PERP_RACE))

## Preview changes
shootings <- shootings[order(shootings$OCCUR_DATE), ]
head(shootings, 13)
```

```
## # A tibble: 13 x 10
##    OCCUR_DATE BORO    STATISTICAL_MUR~ OCCUR_TIME VIC_AGE_GROUP VIC_SEX VIC_RACE
##    <date>     <chr>   <lgl>            <time>     <chr>         <chr>   <chr>
##  1 2006-01-01 MANHAT~ TRUE             02:22      25-44         M       BLACK
##  2 2006-01-01 BROOKL~ FALSE            03:30      18-24         M       BLACK
##  3 2006-01-01 BRONX   FALSE            05:51      18-24         M       WHITE H~
##  4 2006-01-01 BROOKL~ TRUE             12:30      25-44         M       BLACK
##  5 2006-01-01 QUEENS  FALSE            19:00      18-24         M       BLACK
##  6 2006-01-01 QUEENS  TRUE             02:34      25-44         M       BLACK
##  7 2006-01-01 QUEENS  TRUE             02:34      25-44         M       BLACK
##  8 2006-01-01 BRONX   FALSE            02:00      <18           M       BLACK
##  9 2006-01-02 BROOKL~ TRUE             00:49      25-44         M       BLACK
## 10 2006-01-02 STATEN~ FALSE            10:53      18-24         M       BLACK
## 11 2006-01-02 BROOKL~ FALSE            03:59      18-24         M       BLACK
## 12 2006-01-02 BROOKL~ FALSE            03:59      25-44         M       BLACK H~
## 13 2006-01-03 QUEENS  FALSE            13:54      25-44         M       WHITE
## # ... with 3 more variables: PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>
```

Transforming Data

- I was interested in the murder rate per shooting. So I converted the values in the "STATISTICAL_MURDER_FLAG" column to indicator variables. FALSE becomes a 0 and TRUE becomes a 1. This is helpful for counting. So after summing up total murders and total shootings we can then compute the murder ratio (what proportion of shootings result in death). This is given simply by: $MurderRatio = \frac{TotalMurders}{TotalShootings}$

```
## I want to convert the STATISTICAL_MURDER_FLAG column to dummy variables
## so that TRUE's become 1 and FALSE's become 0 for counting methods.
shootings$STATISTICAL_MURDER_FLAG <-
  as.numeric(shootings$STATISTICAL_MURDER_FLAG == "TRUE")

## Tally the total shootings
total_shootings <- length(shootings$STATISTICAL_MURDER_FLAG)
## Tally the total murders
total_murders <- sum(shootings$STATISTICAL_MURDER_FLAG)



## What percentage of reported shootings result in murder?
murder_ratio <- total_murders/total_shootings
murder_percent <- round(murder_ratio * 100, digits = 2)

paste("Total Shootings:" ,total_shootings);
```

```
## [1] "Total Shootings: 25596"
```

```
paste("Total Murders:" ,total_murders);
```

```
## [1] "Total Murders: 4928"
```

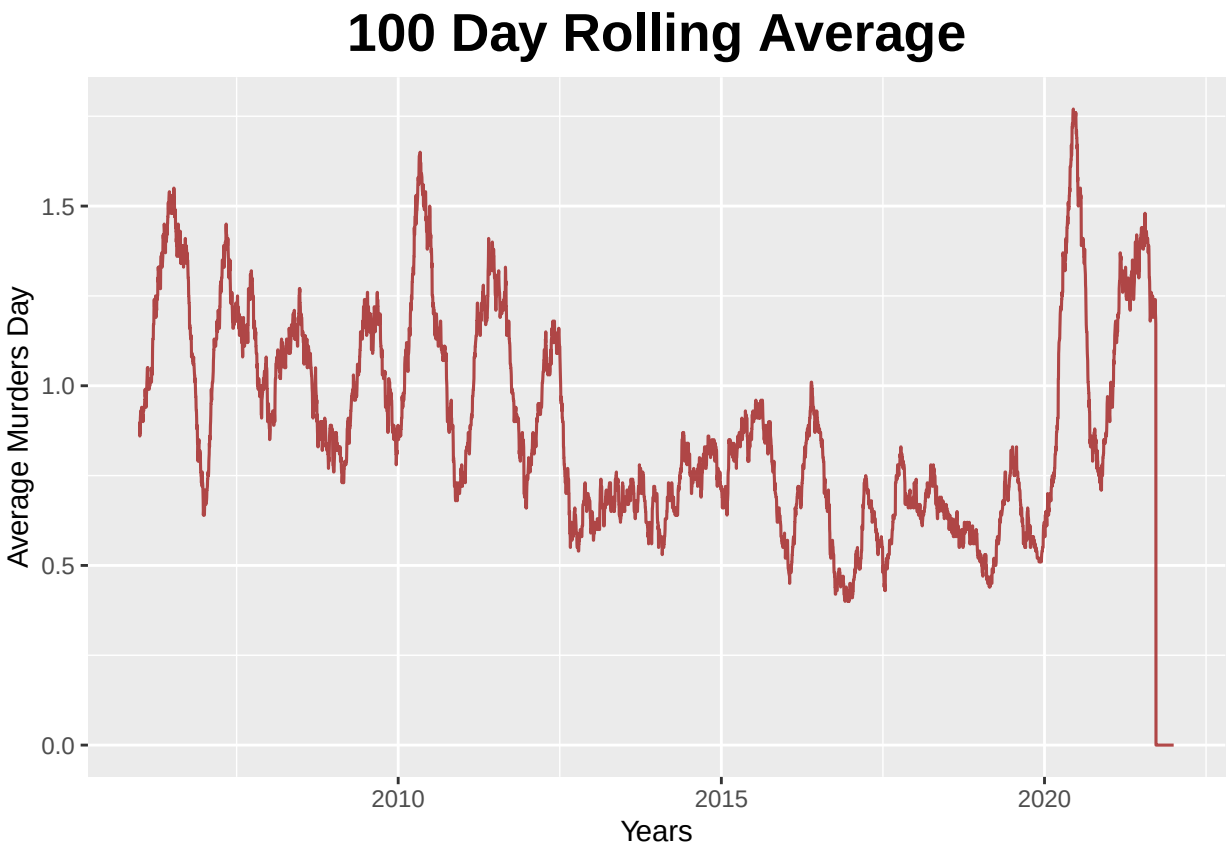
```r
paste("Murder Ratio:" , murder_percent, "%")
```

```
## [1] "Murder Ratio: 19.25 %"
```

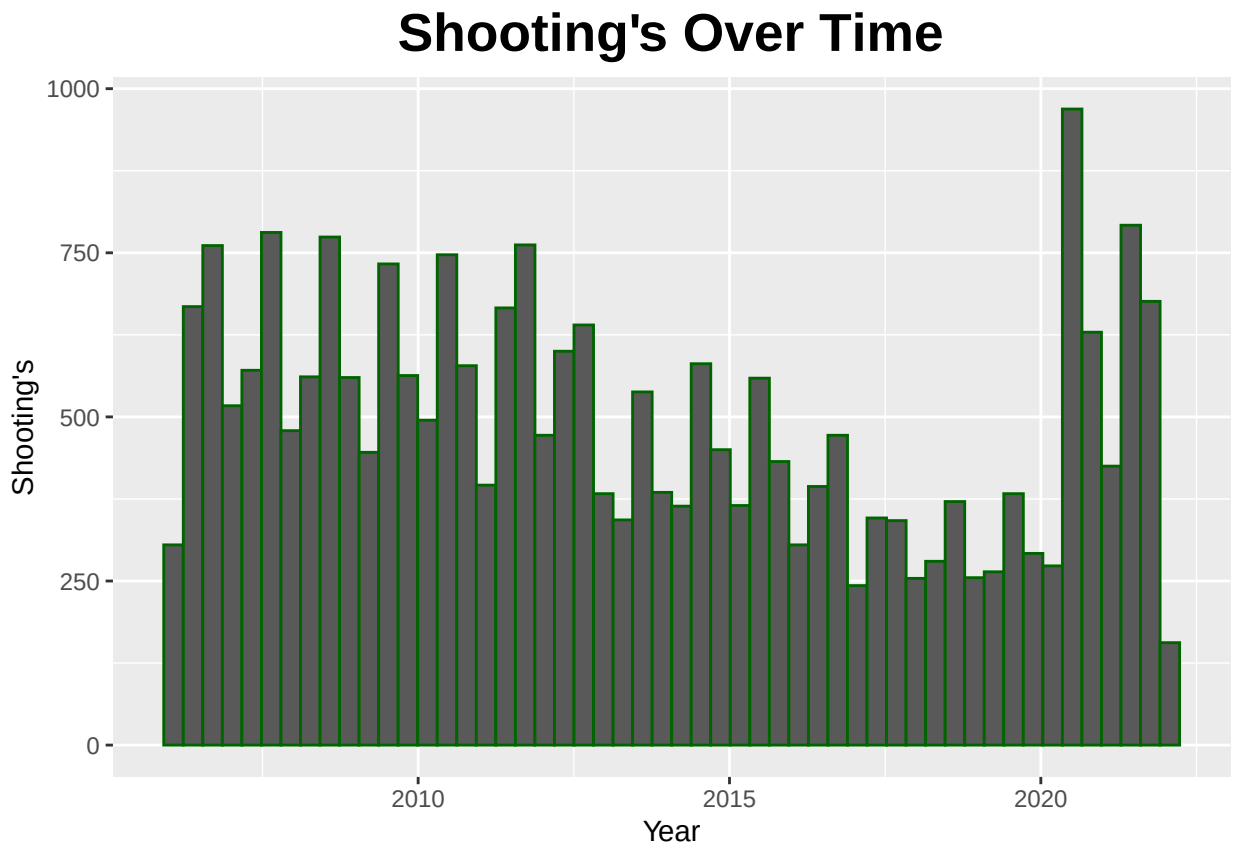
Plotting

- At the time of this analysis the murder ratio was 19.08%.
- It feels likely that roughly $\frac{1}{5}$ of shootings would result in murders
- I wanted to plot some things out so I decided to see what the average murder rates were over time.

```r
###################################
## Create a dataframe without variables to plot shootings/murders over time
shootings_time <- shootings %>%
  select(c(OCCUR_DATE, STATISTICAL_MURDER_FLAG))

## Create a dataframe without variables to plot shootings/murders over time of day
shootings_time_day <- shootings %>%
  select(c(OCCUR_DATE, STATISTICAL_MURDER_FLAG, OCCUR_TIME))

## Pivot table so that there is only one row per day and the murders are summed per day.
shootings_time <- shootings_time %>%
  group_by(OCCUR_DATE) %>%
  summarize(STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(OCCUR_DATE, STATISTICAL_MURDER_FLAG)

## Rolling average plot
avg_over_time_plot <- shootings_time %>%
  mutate(seven_avg = rollmean(STATISTICAL_MURDER_FLAG, 100,
                              align = 'left',
                              fill = 0)) %>%
  relocate(seven_avg) %>%
  ggplot(aes(x=OCCUR_DATE,
             y=STATISTICAL_MURDER_FLAG)) +
  #geom_col(fill = 'red') +
  geom_line(aes(y = seven_avg),
            color = 'brown',
            alpha = .85,
            size = .55) +
  labs(x = "Years", y = "Average Murders Day",
       title = "100 Day Rolling Average") +
  theme(plot.title = element_text(hjust=0.5, size=20, face="bold")) +
  geom_vline(xintercept = as.POSIXct(as.Date("2020-3-16")))

avg_over_time_plot
```



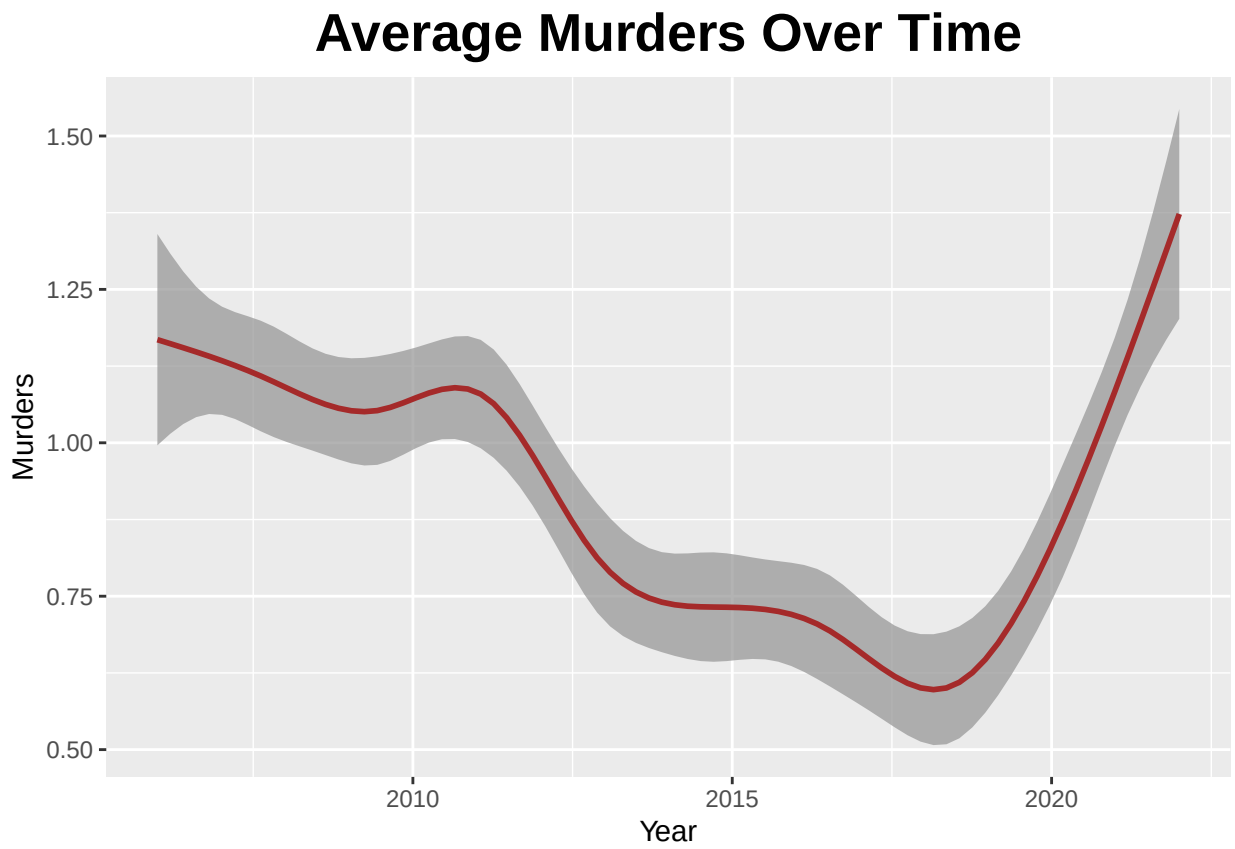

```
## Histogram of murders over time
hist_over_time <- shootings_time_day %>%
  ggplot(aes(x = OCCUR_DATE)) +
  geom_histogram(bins = 52, col='dark green') +
  labs(x="Year", y="Shooting's", title="Shooting's Over Time") +
       theme(plot.title = element_text(hjust=0.5, size=20, face="bold"))

hist_over_time
```



- We can see a steep drop off in murders on both graphs and this is simply due to the data missing beyond this date.
- It would appear that since 2006 the murder rate has trended downward. This was reversed in the pandemic era.
- This is a very alarming graph. The pandemic has impacted social statistics immensely and I am sure we will be discovering more as the data catches up.
- This last plot is admittedly a hard plot to look at, but I do believe it conveys some important data.
- So I made a different style plot with the same data, hopefully easier to look at:

```
plot_to_murder <- ggplot(shootings_time,
            aes(x=OCCUR_DATE,
                y = STATISTICAL_MURDER_FLAG)) +
  geom_smooth(color="brown", alpha = .75) +
  labs(x="Year", y="Murders",
       title="Average Murders Over Time") +
       theme(plot.title = element_text(hjust=0.5,
                                       size=20, face="bold"))
plot_to_murder
```

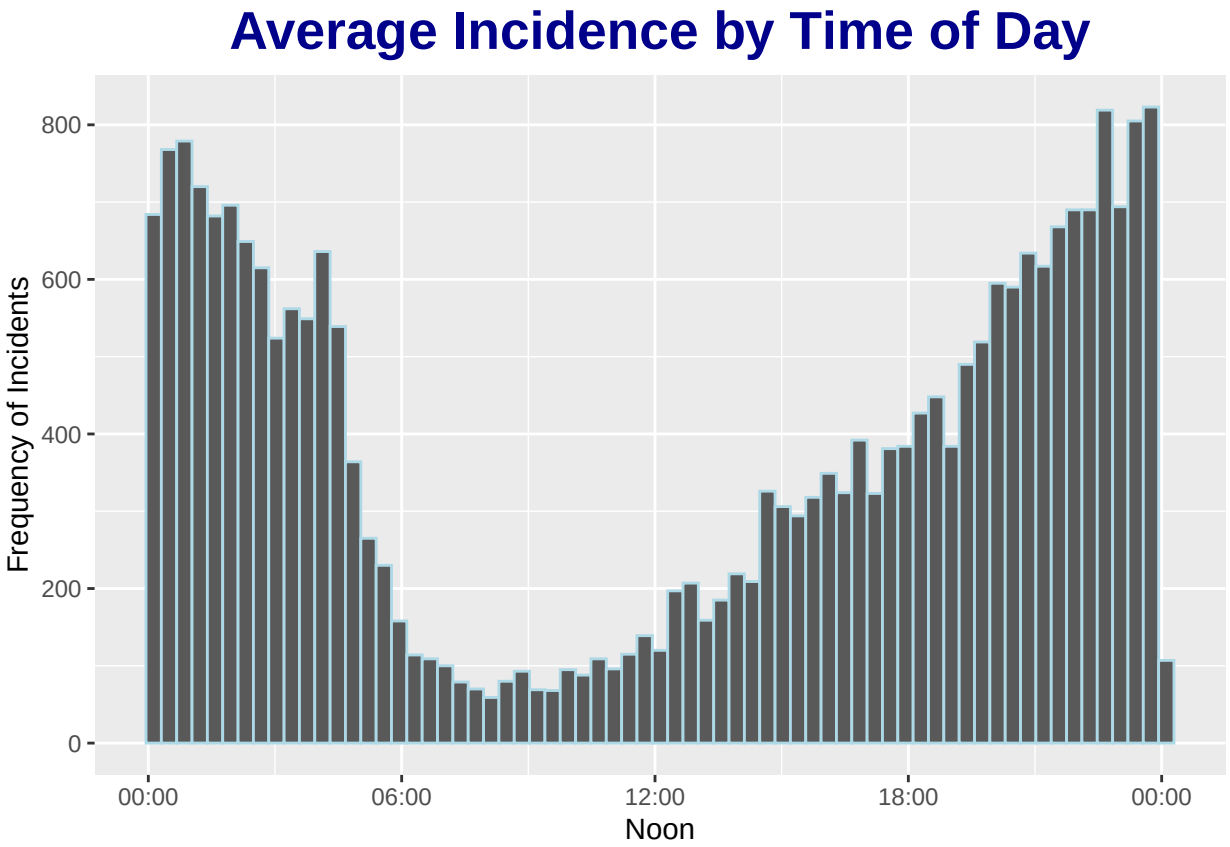
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



More Plotting

- Another Question I was interested in was what time of day most incident's are occuring.

```
murders_hourly_distribution <- function(x, split=24) {
  hours <- as.numeric(strftime(x, "%H"))
  years <- as.POSIXct(paste(ifelse(hours < split,
                                   "2020-01-02","2020-01-01"),
                            strftime(x, "%H:%M:%S")))
  }

hourly_distribution_plot2 <- shootings_time_day %>%
  mutate(time = murders_hourly_distribution(OCCUR_TIME)) %>%
  ggplot(aes(time)) +
  geom_histogram(bins = 67, col='light blue') +
  scale_x_datetime(labels = function(x)
    format(x, format = "%H:%M")) +
  xlab("Noon") +
  ylab("Frequency of Incidents") +
  theme(plot.title = element_text(hjust=0.5, size=20, face="bold",
                                  color ='dark blue')) +
  ggtitle("Average Incidence by Time of Day")

hourly_distribution_plot2
```



Bias

- When I was thinking about this question of how the pandemic would affect the rate of shootings I tried to approach this question without any expectations in order to limit the affect that bias would have to my approach or visualizations.
- For the sake of what bias may look like I asked friends and family what they thought the data would reveal and got very different predictions.
- Many people thought that the numbers would be lower after the shutdown, the logic being that people would be home and not out and about giving less opportunity for these incidences to take place.
- Of course the other side thought that lockdowns would cause a spike in these numbers, most reasoned along the lines of "idle hands are the devils playground" or just people being out of work or money and acting desperately.
- I mention this merely because if I was approaching this question from one of those angles I may be tempted to massage the data into telling the story I wanted to.

Modelling

Explainatory Modelling

- Using standard OLS I will model if time of day has an effect of fatality of a shooting.

```
mod1 <- lm(shootings_time_day$STATISTICAL_MURDER_FLAG ~ shootings_time_day$OCCUR_TIME)
summary(mod1)
```

```
##
## Call:
## lm(formula = shootings_time_day$STATISTICAL_MURDER_FLAG ~ shootings_time_day$OCCUR_TIME)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1929 -0.1928 -0.1923 -0.1921  0.8079
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.921e-01  4.410e-03  43.553   <2e-16 ***
## shootings_time_day$OCCUR_TIME 1.013e-08  8.027e-08   0.126      0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3943 on 25594 degrees of freedom
## Multiple R-squared:  6.227e-07,  Adjusted R-squared:  -3.845e-05
## F-statistic: 0.01594 on 1 and 25594 DF,  p-value: 0.8995
```

- It would appear that time of day does not have a significant effect of shooting's be more or less fatal.
- That seems like common sense, and the p-value confirms this intuition.
- Let's see of the year has any impact on how shootings being fatal.

```
mod2 <- lm(shootings_time_day$STATISTICAL_MURDER_FLAG ~ shootings_time_day$OCCUR_DATE)
summary(mod2)
```

```
##
## Call:
## lm(formula = shootings_time_day$STATISTICAL_MURDER_FLAG ~ shootings_time_day$OCCUR_DATE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1940 -0.1932 -0.1923 -0.1910  0.8092
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.011e-01  2.241e-02   8.973   <2e-16 ***
## shootings_time_day$OCCUR_DATE -5.395e-07  1.404e-06  -0.384    0.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3943 on 25594 degrees of freedom
## Multiple R-squared:  5.771e-06,  Adjusted R-squared:  -3.33e-05
## F-statistic: 0.1477 on 1 and 25594 DF,  p-value: 0.7007
```

- Again it would appear there is no real significant impact the year has on shootings being more or less fatal.
- It is worth noting that the p-value dropped quite a bit for this one, however I don't think this is enough evidence to suggest that shooting's are getting more deadly with time.

More Modelling

Predictive Modelling

```
## Let's look at the dataset since it has been a while.
glimpse(shootings)
```

```
## Rows: 25,596
## Columns: 10
## $ OCCUR_DATE              <date> 2006-01-01, 2006-01-01, 2006-01-01, 2006-01-0~
## $ BORO                    <chr> "MANHATTAN", "BROOKLYN", "BRONX", "BROOKLYN", ~
## $ STATISTICAL_MURDER_FLAG <dbl> 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0~
## $ OCCUR_TIME              <time> 02:22:00, 03:30:00, 05:51:00, 12:30:00, 19:00~
## $ VIC_AGE_GROUP           <chr> "25-44", "18-24", "18-24", "25-44", "18-24", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE HISPANIC", "BLACK", "~
## $ PERP_AGE_GROUP          <chr> "25-44", "UNKNOWN", "25-44", NA, "18-24", "18-~
## $ PERP_SEX                <chr> "M", "U", "M", NA, "M", "M", "M", "M", "U", "M~
## $ PERP_RACE               <chr> "BLACK", "UNKNOWN", "WHITE HISPANIC", NA, "BLA~
```

- We can see that most of our data is categorical so I will try to do some sort of classification regression.
- First we need to change some of the variables for this to work properly.

```
names <- c(1,2,5,6,7,8,9,10)
shootings[,names] <- lapply(shootings[,names] , factor)
glimpse(shootings)
```

```
## Rows: 25,596
## Columns: 10
## $ OCCUR_DATE              <fct> 2006-01-01, 2006-01-01, 2006-01-01, 2006-01-01~
## $ BORO                    <fct> MANHATTAN, BROOKLYN, BRONX, BROOKLYN, QUEENS, ~
## $ STATISTICAL_MURDER_FLAG <dbl> 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0~
## $ OCCUR_TIME              <time> 02:22:00, 03:30:00, 05:51:00, 12:30:00, 19:00~
## $ VIC_AGE_GROUP           <fct> 25-44, 18-24, 18-24, 25-44, 18-24, 25-44, 25-4~
## $ VIC_SEX                 <fct> M, M, M, M, M, M, M, M, M, M, M, M, M, M, M, M~
## $ VIC_RACE                <fct> BLACK, BLACK, WHITE HISPANIC, BLACK, BLACK, BL~
## $ PERP_AGE_GROUP          <fct> 25-44, UNKNOWN, 25-44, NA, 18-24, 18-24, 25-44~
## $ PERP_SEX                <fct> M, U, M, NA, M, M, M, M, U, M, M, M, M, U, M, ~
## $ PERP_RACE               <fct> BLACK, UNKNOWN, WHITE HISPANIC, NA, BLACK, BLA~
```

```
## New dataframe with selected variables
shootings_2 <- shootings %>%
  select(-c(OCCUR_DATE, OCCUR_TIME,
            STATISTICAL_MURDER_FLAG))

## Set the seed for reproducibility
set.seed(2022)

## Split the data up into training and testing sets
spl = sample.split(shootings_2$VIC_AGE_GROUP, SplitRatio = 0.7)
train = subset(shootings_2, spl==TRUE)
```

```r
test  = subset(shootings_2, spl==FALSE)

## See what dimensions our training and testing set's have
#print(dim(train)); print(dim(test))

## Let's look at the predictive value of each variable on Victim's Age Group
model_glm = glm(VIC_AGE_GROUP ~  . , family="binomial", data = shootings_2)
summary(model_glm)
```

```
##
## Call:
## glm(formula = VIC_AGE_GROUP ~ ., family = "binomial", data = shootings_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0405   0.3008   0.4576   0.5256   1.1874
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       12.46743 1020.11573   0.012  0.99025
## BOROBROOKLYN                       0.13818    0.06385   2.164  0.03045 *
## BOROMANHATTAN                      0.06117    0.08106   0.755  0.45044
## BOROQUEENS                         0.28252    0.08618   3.278  0.00104 **
## BOROSTATEN ISLAND                  0.12591    0.14759   0.853  0.39361
## VIC_SEXM                           0.65781    0.07293   9.020  < 2e-16 ***
## VIC_SEXU                          12.55077  426.10082   0.029  0.97650
## VIC_RACEASIAN / PACIFIC ISLANDER   2.31716    0.94679   2.447  0.01439 *
## VIC_RACEBLACK                      1.06291    0.88163   1.206  0.22797
## VIC_RACEBLACK HISPANIC             1.15089    0.88478   1.301  0.19334
## VIC_RACEUNKNOWN                   14.20375  200.61392   0.071  0.94356
## VIC_RACEWHITE                      1.93409    0.91240   2.120  0.03402 *
## VIC_RACEWHITE HISPANIC             1.28012    0.88376   1.448  0.14748
## PERP_AGE_GROUP1020                14.80130 1455.39753   0.010  0.99189
## PERP_AGE_GROUP18-24                1.08359    0.06978  15.528  < 2e-16 ***
## PERP_AGE_GROUP224                 14.70396 1455.39754   0.010  0.99194
## PERP_AGE_GROUP25-44                2.15010    0.08722  24.653  < 2e-16 ***
## PERP_AGE_GROUP45-64                2.42811    0.24227  10.022  < 2e-16 ***
## PERP_AGE_GROUP65+                 14.79590  189.27792   0.078  0.93769
## PERP_AGE_GROUP940                 14.56579 1455.39754   0.010  0.99201
## PERP_AGE_GROUPUNKNOWN              1.09652    0.09632  11.384  < 2e-16 ***
## PERP_SEXM                         -0.68833    0.22773  -3.023  0.00251 **
## PERP_SEXU                         -0.88419    0.28914  -3.058  0.00223 **
## PERP_RACEASIAN / PACIFIC ISLANDER -12.57426 1020.11541  -0.012  0.99017
## PERP_RACEBLACK                    -12.73505 1020.11535  -0.012  0.99004
## PERP_RACEBLACK HISPANIC           -12.86560 1020.11536  -0.013  0.98994
## PERP_RACEUNKNOWN                  -12.72477 1020.11537  -0.012  0.99005
## PERP_RACEWHITE                    -12.66859 1020.11540  -0.012  0.99009
## PERP_RACEWHITE HISPANIC           -12.85492 1020.11535  -0.013  0.98995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 11599  on 16251  degrees of freedom
## Residual deviance: 10699  on 16223  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 10757
##
## Number of Fisher Scoring iterations: 14
```

```
set.seed(2022)
# Predictions on the training set
predictTrain = predict(model_glm, data = train, type = "response")

training_accuracy <- sum(predictTrain >= 0.5) / length(predictTrain)
paste("Training Accuracy = ", round(training_accuracy, 3))
```

```
## [1] "Training Accuracy =  0.999"
```

```
#Predictions on the test set
predictTest = predict(model_glm, newdata = test, type = "response")

# Accuracy of our model
test <- table(test$VIC_AGE_GROUP, predictTest >= 0.5)

paste("Predictions on Test Set"); test
```

```
## [1] "Predictions on Test Set"
```

```
##
##           FALSE TRUE
##   <18         3  548
##   18-24       0 1807
##   25-44       1 2086
##   45-64       0  338
##   65+         0   35
##   UNKNOWN     0   16
```

Conclusion

- On murder trends:
  - It appears that with the data provided, going back almost two decades the shootings and murder rates were declining over time.
  - This trend the data revealed was reversed when the first shutdown of the pandemic began in early 2020.
- On popular times of day for shootings:
  - It would appear that most shootings and therefore murders as well happen in the hours leading up to and trailing midnight.
  - This doesn't come as a big surprise but it is always good to confirm intuitions with data.
- On modelling:
  - Using most all of the variables we were able to predict with very high accuracy the age group of the victim's.

– Although we did have quite a bit of training data for this small task, it is interesting to think about that with basically data on victim's sex and race, the perpetrator's age, sex, and race, and neighborhood data, we are able to predict the age group of the victim's.
– With more analysis we could decide how much if at all this model is overfitting.