NYPD Shooting Incident Analysis

ADP

2023-07-24

Introduction

This is an analysis of NYPD shooting incidents that occurred from the start of 2006 to the end of 2022. This project will be an effort to understand and analyze the provided data and to draw some meaningful conclusions as part of the "Data Science As A Field Course" at UC Boulder in the Summer 2 session in 2023. It is also motivated by a desire to understand recent trends in NYC shootings based on a number of factors. Please be advised that the topic discussed is a sensitive one and may be upsetting to some readers.

Importing Libraries

The only libraries I needed to run the following analysis were tidyverse and lubridate, but that was accessing the data via the link specified in Week 3 on Coursera (listed below as url1). If there are any issues with that website, data can also be accessed via my personal Github (listed below as url2).

Note that I have had some issues accessing the file via Github and installed then loaded the httr and readr libraries to rectify this. You may also need to do this to access the Github file.

If that fails, you can download the file manually from my Github and save it locally to access and analyze.

Please ensure that you have tidyverse and lubridate installed before running the following code (which loads the library but does not install it). If you need httr and readr to access the Github code, please install those as well before loading. Otherwise you may not be able to run the code.

```
## import your libraries in this code chunk
library(tidyverse)
library(lubridate)
```

Part 1 Importing Data

As specified above, the primary data source for the project can be found at url1 below. A backup data source can be found at url2 below (please see instructions above for how to access). More detailed information on the data gathered can be found via this data_info link.

data_info <- "https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8")

```
url1 <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
url2 <- "https://raw.githubusercontent.com/apretko11/Data_Science_As_A_field/main/NYPC_Shooting_Data.cs
shots <- read_csv(url1, show_col_types=TRUE)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

```
## intentionally showing column types to help with analysis
## can set to FALSE if you want to save space in the output or don't need the info
```

Part 2 Transform, Tidy, and Summarize

Since much of this analysis will focus on the timing of the incidents, I will transform the variables into more readily usable forms such as hour, year, and adjusted date (since the original date is not formatted as a date).

```
## get the hour when it occurred
shots$hour <- hour(shots$OCCUR_TIME)

## get the year it occurred
shots$adj_date <- as.Date(shots$OCCUR_DATE, format="%m/%d/%Y")
shots$year <- year(shots$adj_date)
```

My next task is to remove the columns that are not necessary in my analysis. Note the columns included are the data points I wish to use for my analysis. This does not mean the other points are inherently without value. They are simply not the focus of this analysis but could add value in a subsequent analysis.

```
shots<- shots %>%
  select(-c(OCCUR_DATE:VIC_SEX, X_COORD_CD:Lon_Lat))
```

Note on Missing Data

There are several missing values in the original data set. In particular, the data about the perpetrator has a lot of missing information. The features that I am working with, however, have no missing values and require no further imputation of values. Since the missing values are all in columns that are not featured in my analysis, they should not have a significant impact on the outcome.

Note that these columns were chosen to facilitate a temporal analysis - not simply to avoid missing data points in the other columns.

If we wanted to include those variables with missing values, we could choose between dropping the rows with missing data, imputing missing values, only including the rows when they have a value for that variable, or using some other method to adjust the data.

```
if(sum(colSums(is.na(shots)) ==0)){
  print("No missing values")
} else {
  print("Warning! Missing Values - Double Check Data")
}
```

```
## [1] "No missing values"
```

The end result of importing, transforming, and tidying is a data set with five columns (features/columns) for 27,312 data points. Note that figure is current as of 2023.07.19 but that does not mean no corrections or amendments can occur in the data.

While I do not anticipate a major increase or decrease in data points, please note that some small data shifts may occur. If any major shifts in data occur, please contact me to address the issue.

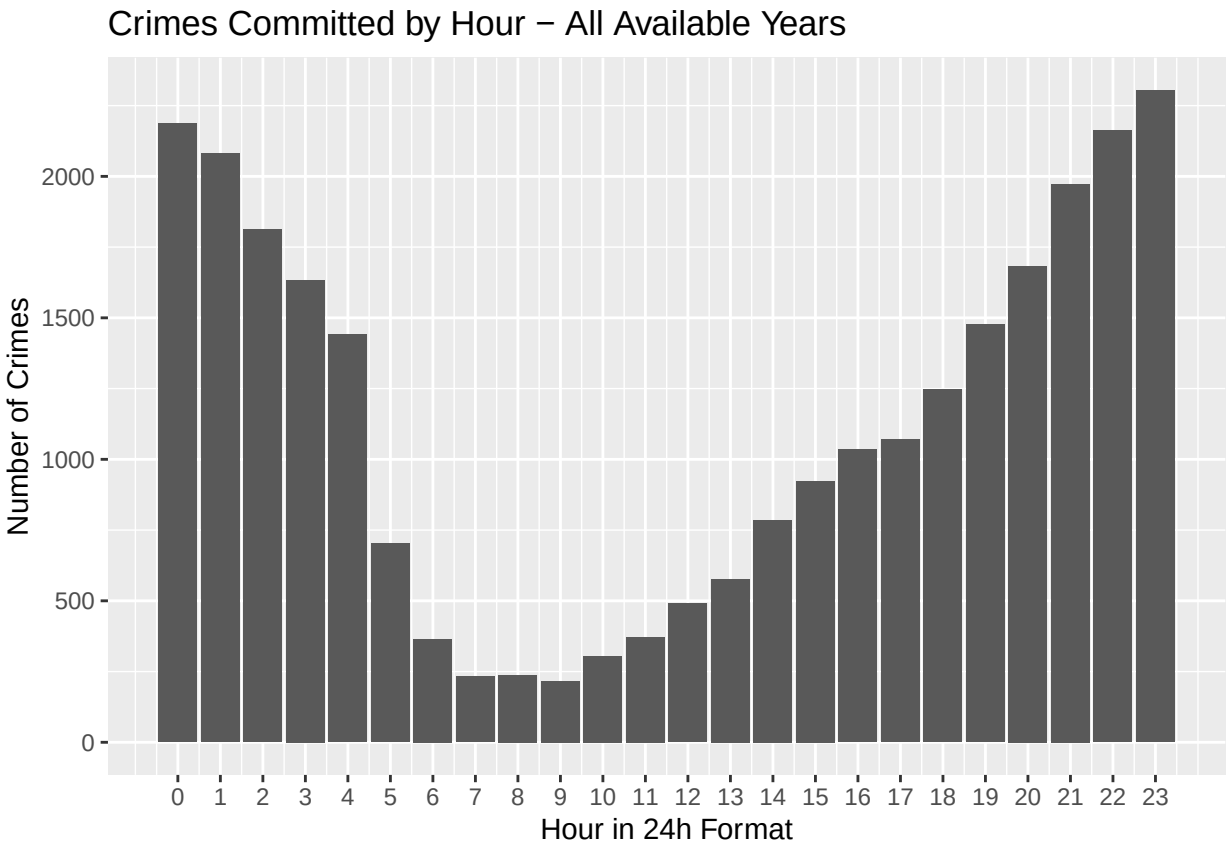
```
summary(shots)
```

```
##   INCIDENT_KEY         VIC_RACE              hour          adj_date
##  Min.   :  9953245   Length:27312       Min.   : 0.00   Min.   :2006-01-01
##  1st Qu.: 63860880   Class :character   1st Qu.: 3.00   1st Qu.:2009-07-18
##  Median : 90372218   Mode  :character   Median :15.00   Median :2013-04-29
##  Mean   :120860536                      Mean   :12.22   Mean   :2014-01-06
##  3rd Qu.:188810230                      3rd Qu.:20.00   3rd Qu.:2018-10-15
##  Max.   :261190187                      Max.   :23.00   Max.   :2022-12-31
##       year
##  Min.   :2006
##  1st Qu.:2009
##  Median :2013
##  Mean   :2013
##  3rd Qu.:2018
##  Max.   :2022
```

Part 3: Visualizations and Relevant Analysis

Shootings do not occur evenly throughout the day. Using the full data set, we see that shootings increase throughout the evening and reach a peak around midnight before gradually decreasing until mid morning. It then begins to rise again steadily throughout the day until its midnight peak

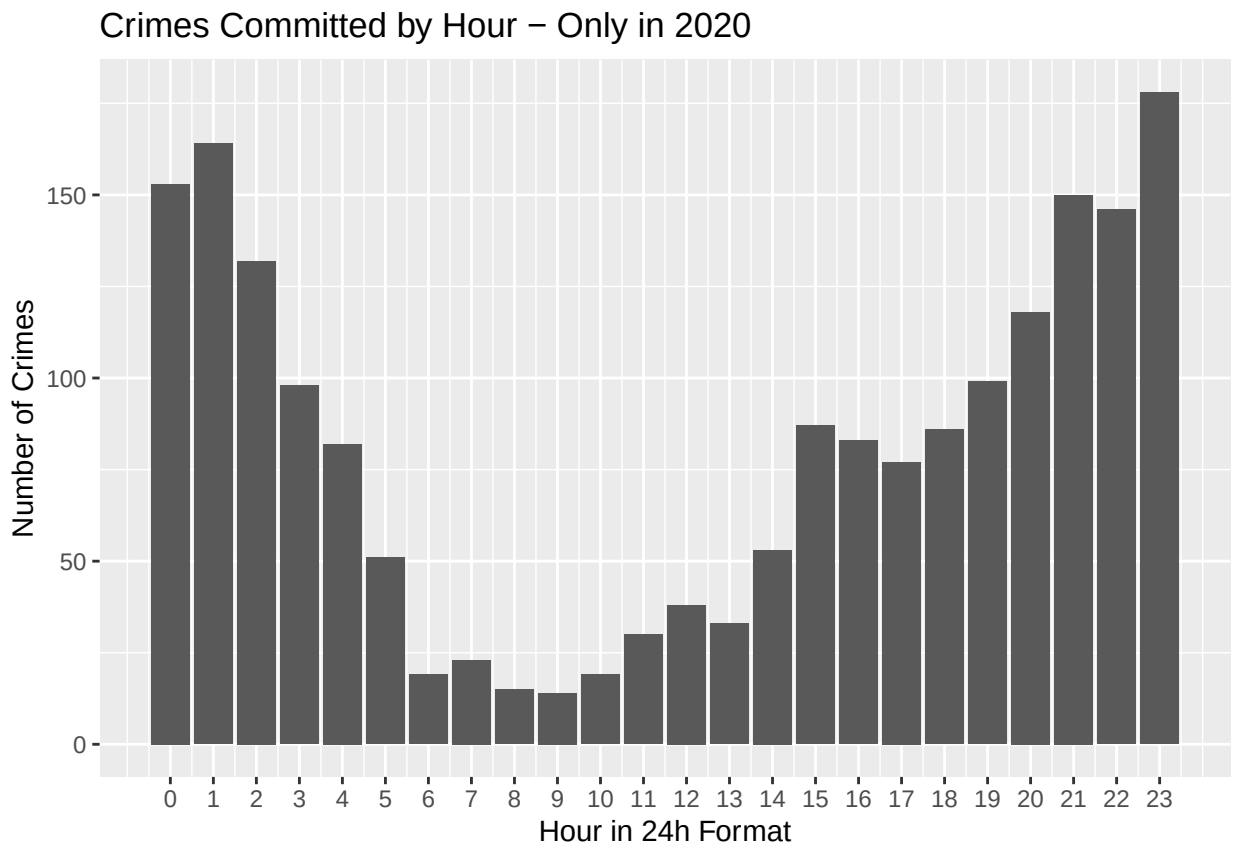
```
ggplot(shots, aes(x=hour)) +
  geom_bar() +
  labs(title="Crimes Committed by Hour - All Available Years",
       x="Hour in 24h Format",
       y="Number of Crimes") +
  scale_x_continuous(breaks=seq(0,23,1))
```



This trend is seen in every year in the data from 2006 through 2022, which is the total data set. The below graph shows 2020, which I selected as the year most heavily impacted by Covid. Note that the shape of the results is still substantially the same. Curious readers are invited to try filtering for other years (such as 2006 through 2019 or 2021-2022) in the code block below, which will reveal much the same results.

While the pattern of more shootings late at night and early morning has remained the same over the years, the exact number of shootings per year and per period has changed and will be shown in the next graph.

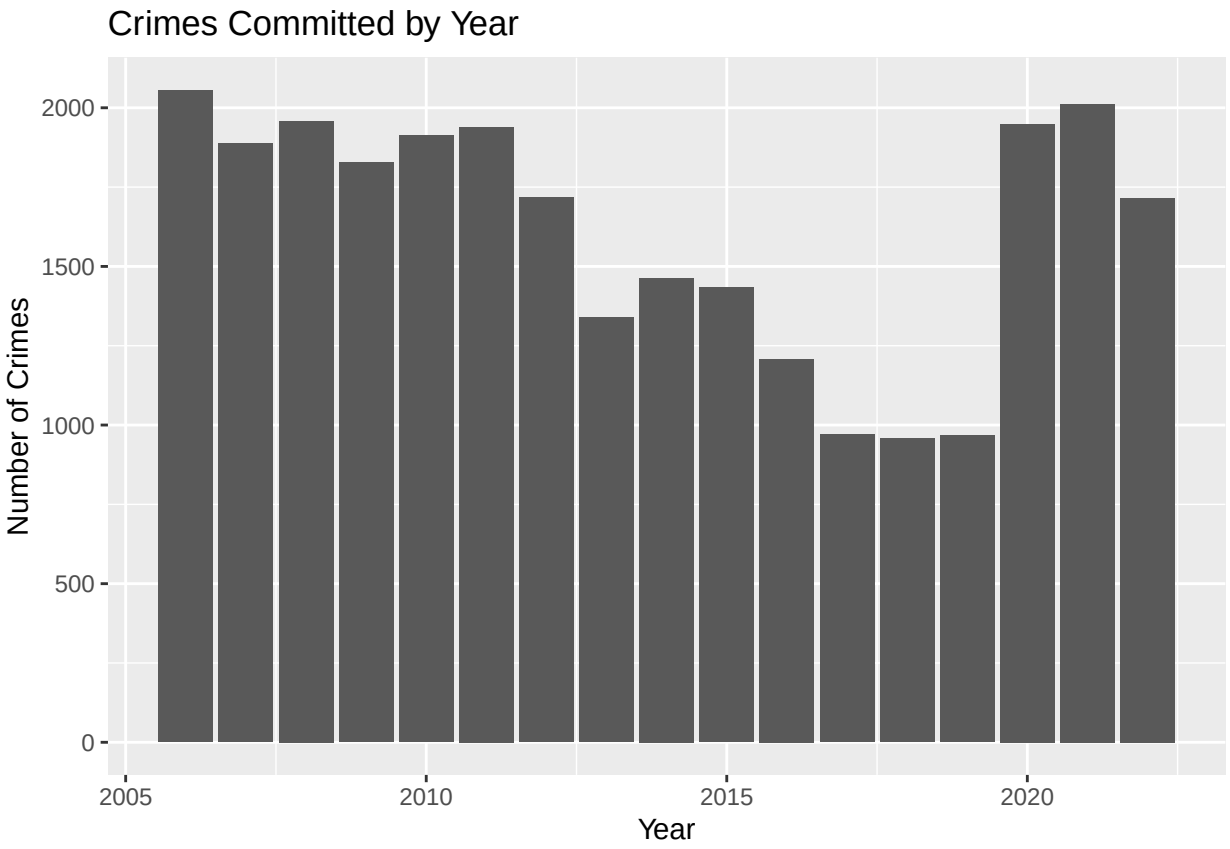
```
## only 2020 data
shots_2020 <- shots %>% filter(year==2020)

ggplot(shots_2020, aes(x=hour)) +
  geom_bar() +
  labs(title="Crimes Committed by Hour - Only in 2020",
       x="Hour in 24h Format",
       y="Number of Crimes") +
  scale_x_continuous(breaks=seq(0,23,1))
```



The dataset shows a significant decrease in shootings, with values decreasing from an average of 1930 per year from 2006 through 2011 to an average of 1257 in 2012 through 2019 and dipping as low as 965 in the years 2017 through 2019 before a huge increase during Covid.

The data shows that 2020 and 2021 on average experience more than double the incident rate in the 2017 through 2019 period. 2020 through 2021 averaged 1980 incidents per year before decreasing slightly to 1716 incidents in 2022. While these figures wouldn't be out of line with 2006 through 2011 figures, they certainly are an abrupt reversal in the trend.

```
ggplot(shots, aes(x=year)) +
  geom_bar() +
  labs(title="Crimes Committed by Year", x="Year", y="Number of Crimes")
```




```
## 2006 through 2011 Average
data_2006_to_2011 <- shots %>% filter(year(adj_date) >= 2006 & year(adj_date) <= 2011)
data_2006_to_2011 <- data_2006_to_2011 %>% group_by(year) %>% summarise(Crime_Count = n())
mean(data_2006_to_2011$Crime_Count)
```

```
## [1] 1930
```

```
## 2012 through 2019 Average
data_2012_to_2019 <- shots %>% filter(year(adj_date) >= 2012 & year(adj_date) <= 2019)
data_2012_to_2019 <- data_2012_to_2019 %>% group_by(year) %>% summarise(Crime_Count = n())
mean(data_2012_to_2019$Crime_Count)
```

```
## [1] 1257.125
```

```
## 2017 through 2019 Average
data_2017_to_2019 <- shots %>% filter(year(adj_date) >= 2017 & year(adj_date) <= 2019)
data_2017_to_2019 <- data_2017_to_2019 %>% group_by(year) %>% summarise(Crime_Count = n())
mean(data_2017_to_2019$Crime_Count)
```

```
## [1] 965
```

```
## 2020 through 2021 Average
data_2020_to_2021 <- shots %>% filter(year(adj_date) >= 2020 & year(adj_date) <= 2021)
data_2020_to_2021 <- data_2020_to_2021 %>% group_by(year) %>% summarise(Crime_Count = n())
mean(data_2020_to_2021$Crime_Count)
```

```
## [1] 1979.5
```

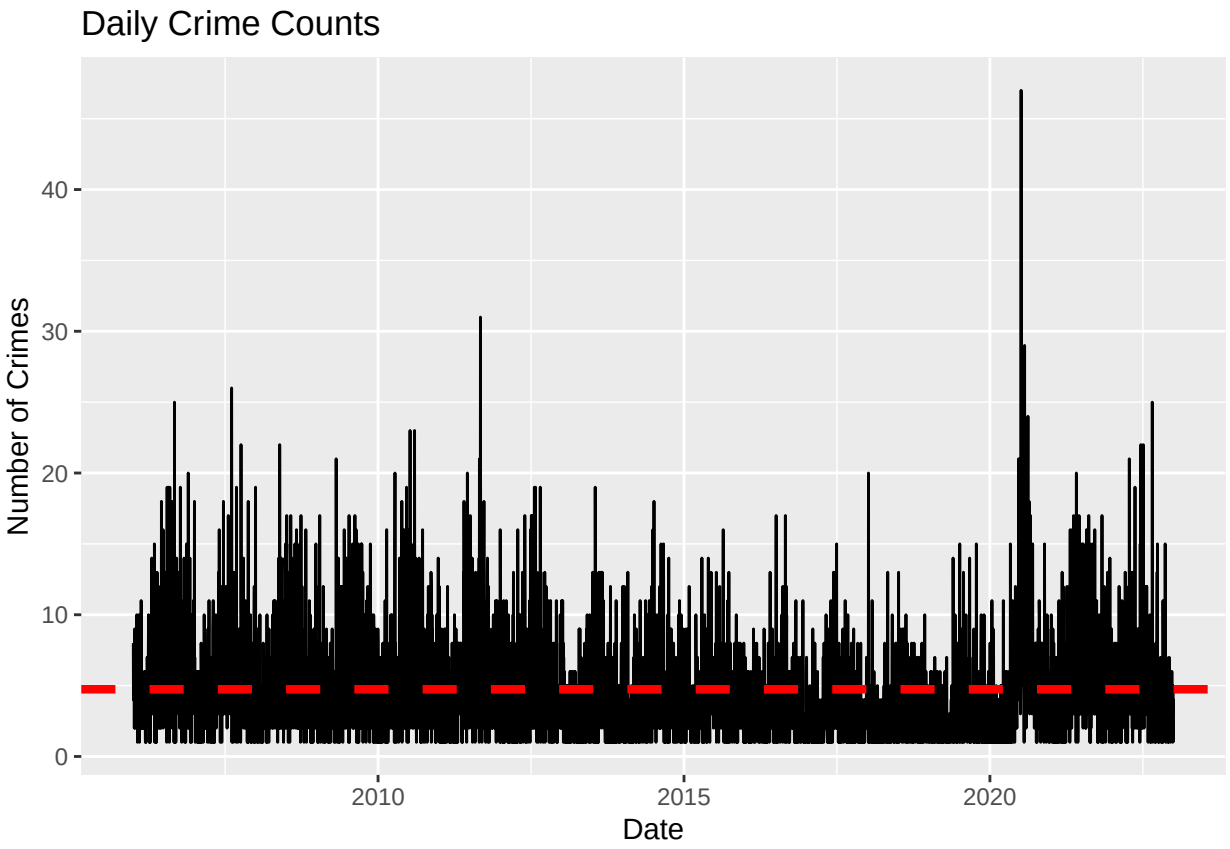
```
## 2022 Count
data_2022 <- shots %>% filter(year(adj_date)==2022) %>% summarise(Crime_Count=n())
data_2022
```

```
## # A tibble: 1 x 1
##   Crime_Count
##         <int>
## 1        1716
```

I now consider how many crimes are occurring per day. Average daily count is fairly low but frequent and significant upward deviations are clearly visible. The maximum value is 47 and occurs on 2020.07.05 Additional information at the abc link below, which explains a significant and aberrant spike in shootings.

abc <- "https://abc7ny.com/nyc-shootings-2020-last-night-this-week-in/6299513/"

```
daily_crime_counts <- shots %>% group_by(adj_date) %>% summarise(Crime_Count = n())

## graph of crime count by day with the average in a dotted line in red
ggplot(daily_crime_counts, aes(x=adj_date,y=Crime_Count)) +
  geom_line() +
  geom_hline(yintercept=mean(daily_crime_counts$Crime_Count),
             linetype="dashed", color="red", linewidth=1.5)+
  labs(title="Daily Crime Counts",x="Date",y="Number of Crimes")
```



```r
daily_max <- daily_crime_counts %>%
  filter(daily_crime_counts$Crime_Count == max(daily_crime_counts$Crime_Count))

daily_top_25 <- daily_crime_counts %>%
  arrange(desc(Crime_Count)) %>% head(25)

daily_top_25
```

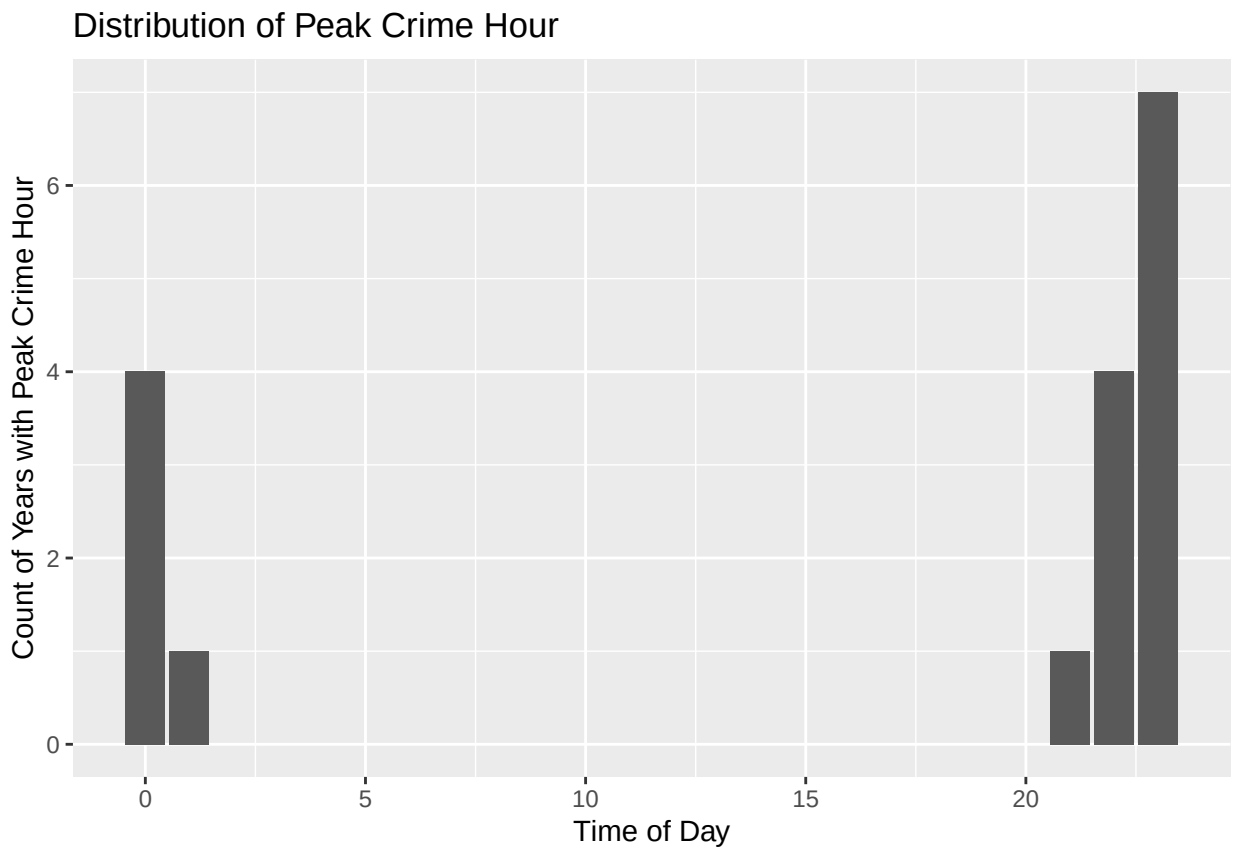
```
## # A tibble: 25 x 2
##    adj_date   Crime_Count
##    <date>           <int>
##  1 2020-07-05          47
##  2 2011-09-04          31
##  3 2020-07-26          29
##  4 2007-08-11          26
##  5 2006-09-04          25
##  6 2022-08-27          25
##  7 2020-08-15          24
##  8 2010-07-11          23
##  9 2010-08-07          23
## 10 2007-10-06          22
## # i 15 more rows
```

We first calculate for each year the the one hour period during the day when the most incidents occur. Then we create a bar chart showing the counts of how many years have their maximum during this time period. 7 out of 17 years experience the most shootings between 11pm and midnight, which is reflected by the largest bar appearing at the value 23 (indicating 23:00 in 24h format). Midnight to 1am and 10pm to 11pm are both equally likely with 4 counts. Only one year has its most intense hour for shooting between 1am and 2am, with another outlier being one year with its most shootings between 9pm and 10pm.

```r
hourly_counts <- shots %>% group_by(year, hour) %>% summarise(Crime_Count = n())
```

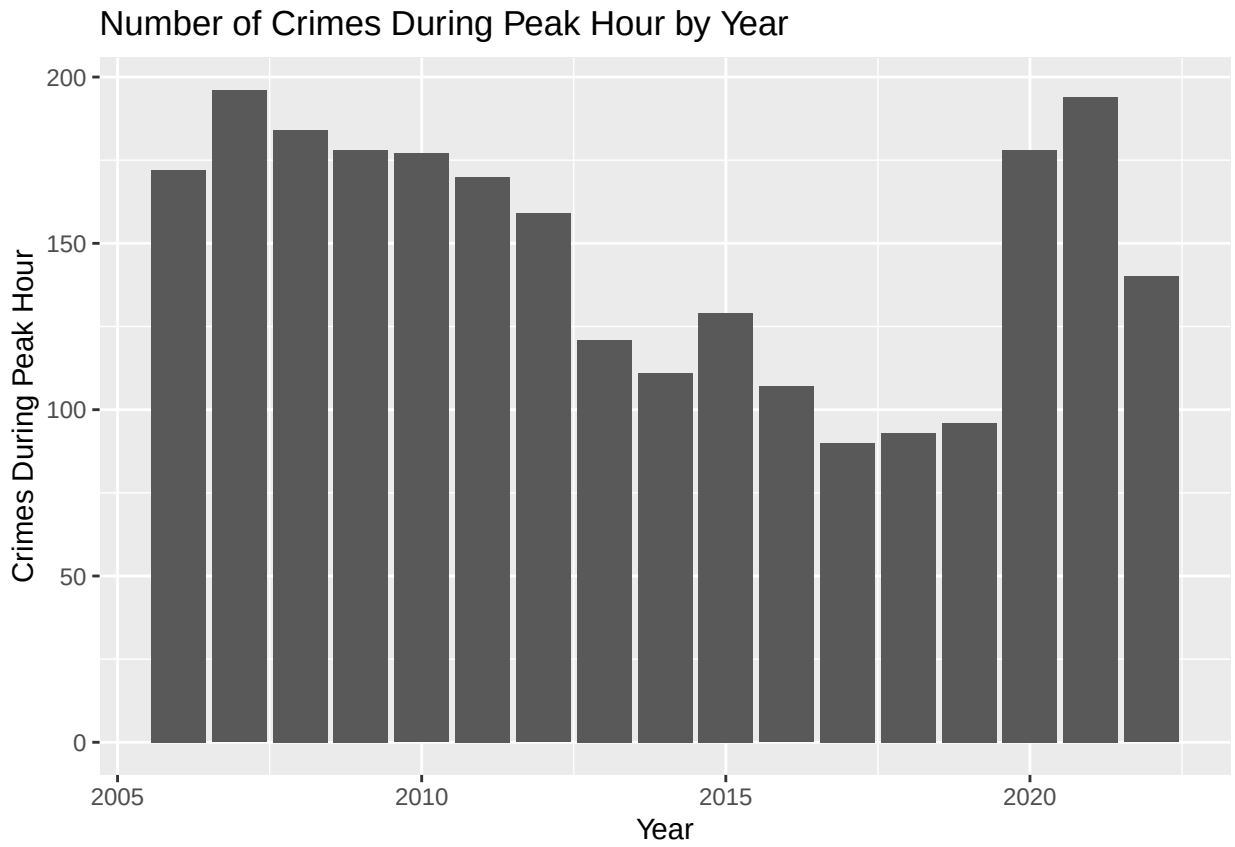
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```r
max_hour_per_year <- hourly_counts %>% group_by(year) %>% slice(which.max(Crime_Count))
hour_counts <- max_hour_per_year %>% count(hour)
ggplot(hour_counts, aes(x=hour, y=n)) +
  geom_bar(stat="identity") +
  labs(title="Distribution of Peak Crime Hour",x="Time of Day",y="Count of Years with Peak Crime Hour")
```



While the main hours that the shootings occur are consistently late at night or very early morning, the number of shootings that occur during that peak hour tracks the trends in overall shootings. The number of shootings that occurred during peak times was higher in 2006 through 2011 followed by a steady decline to the lowest period of 2017 through 2019. This trend then rapidly reversed course in 2020 and 2021 with a small decline in 2022. The most obvious high-level explanation would be Covid, but I would caution against drawing a causal inference from this preliminary analysis.

```
ggplot(max_hour_per_year, aes(x=year, y=Crime_Count)) +
  geom_bar(stat="identity") +
  labs(title="Number of Crimes During Peak Hour by Year ",x="Year",y="Crimes During Peak Hour")
```

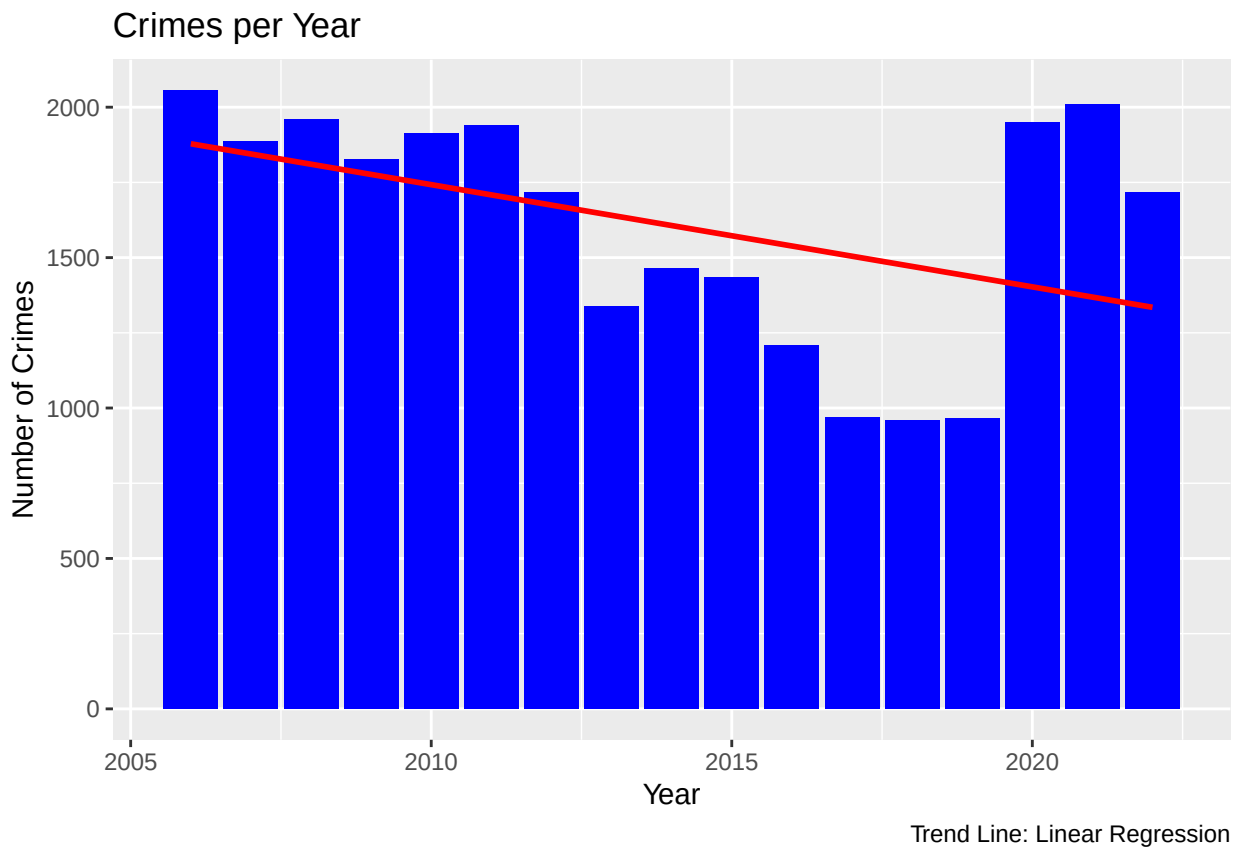


Part 4: Model Using Linear Regression to Smooth

The model below uses linear regression to smooth out the trend and shows how markedly the 2020-2022 years exceed the trend. Note that this model includes 2020 through 2022 to calculate the trend line. Excluding those three years (i.e. showing a trend line based on 2006 through 2019 data results in an even steeper decline in trend line)

```
crimes_per_year <- shots %>% group_by(year) %>% summarise(number_of_crimes = n())

ggplot(crimes_per_year, aes(x=year, y=number_of_crimes)) +
  geom_col(fill="blue") +
  geom_smooth(method="lm", se=FALSE, color="red") +
  labs(title="Crimes per Year",
       x="Year",
       y= "Number of Crimes",
       caption = "Trend Line: Linear Regression")
```

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

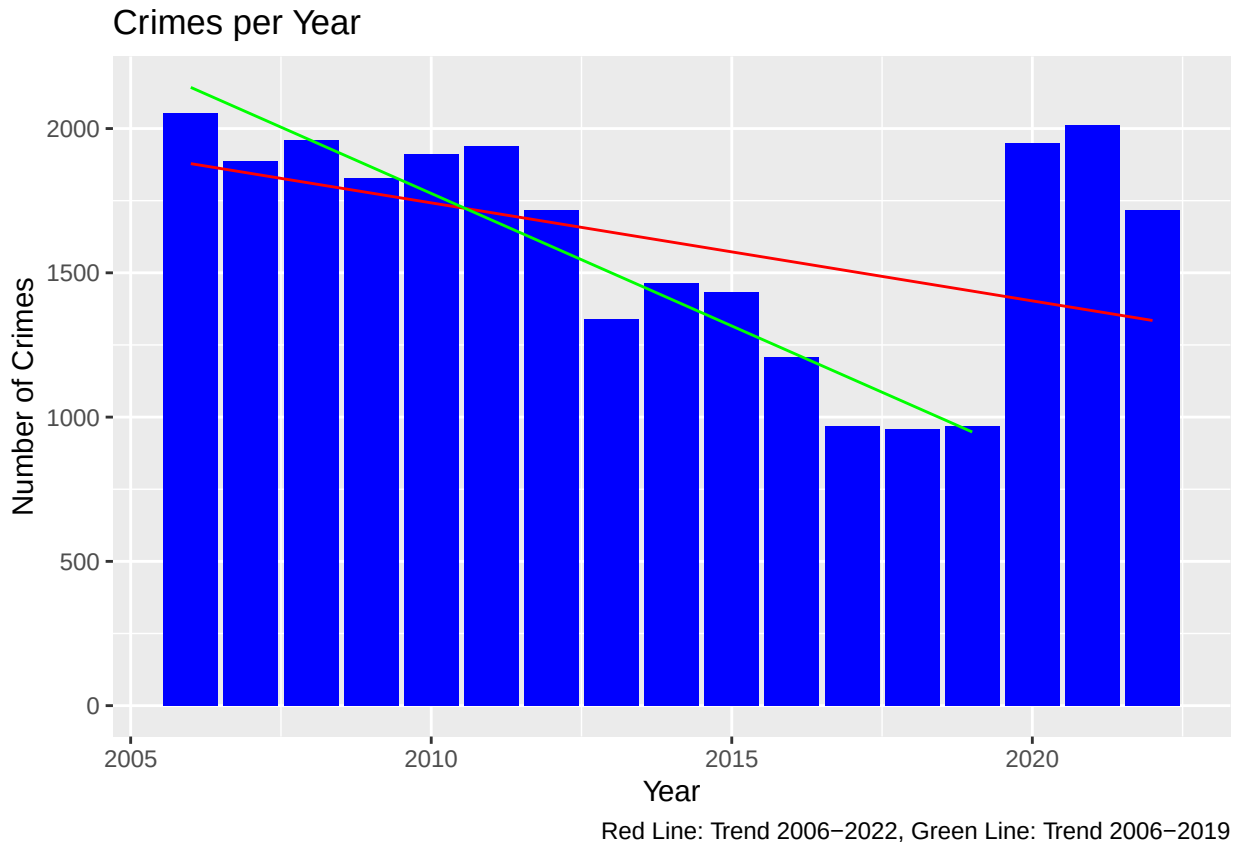


The below model uses linear regression to smooth out the trend but trains the data exclusively on 2006 through 2019, resulting in a much steeper decline (shown in green). The line trained on all data from 2006 through 2022 is included in red above.

This shows that a trend line through the pre-Covid data indicates a much sharper trajectory than a best fine line that includes the 2020 through 2022 data. Consequently, the actual 2020 through 2022 values are an even bigger deviation from the trend when considering the trends based solely on pre-Covid data

```
# calculate number of crimes per year for 2006 through 2022
crimes_per_year_2006_to_2022 <- shots %>%
  group_by(year) %>% summarise(number_of_crimes=n())

# trend line with full period LM
trend_line_2006_to_2022 <- lm(number_of_crimes ~ year, data=crimes_per_year_2006_to_2022)

# data for 2006 through 2019
data_2006_to_2019 <- shots %>%
  filter(year >= 2006 & year <= 2019)

# calculate number of crimes per year for 2006 through 2022
crimes_per_year_2006_to_2019 <- data_2006_to_2019 %>%
  group_by(year) %>% summarise(number_of_crimes=n())

# trend line with smaller period
trend_line_2006_to_2019 <- lm(number_of_crimes ~ year, data=crimes_per_year_2006_to_2019)

# now for the plot
```

```
ggplot() +
  geom_col(data=crimes_per_year_2006_to_2022,
           aes(x=year, y=number_of_crimes), fill="blue") +
  geom_line(data=crimes_per_year_2006_to_2022,
            aes(x=year, y=predict(trend_line_2006_to_2022)), color="red") +
  geom_line(data=crimes_per_year_2006_to_2019,
            aes(x=year, y=predict(trend_line_2006_to_2019)), color="green") +
  labs(title="Crimes per Year",
       x = "Year",
       y = "Number of Crimes",
       caption = "Red Line: Trend 2006-2022, Green Line: Trend 2006-2019",
       fill = "Actual Data")
```



Part 5 Bias Indentification and Conclusion

While I chose to focus on the temporal aspect of the data, I'm aware that there are some potential gaps or biases in the data. For example, while the information on the victims was fairly thorough, information on the perpetrators had a large number of gaps. In many cases, information such as gender, race, and age characteristics were missing in the data, which could introduce certain biases.

For example, if we only have data on half of the perpetrators and of that half, X% are the same race as their victim, can we generalize that to the sample as a whole? This assumption might be valid if the missing data is random, but it is a poor assumption if there is another variable that impacts this. What if people who are the victims committed by someone of the same race are less likely to report it? What if they're more likely to report it?

Another source of bias is how the data is sourced and defined. For example, this is data that comes directly from the NYPD. What if some groups of people are less likely to contact the police when they're the victim? Even if the data is reported and recorded accurately, this would mean the data is biased and does not necessarily reflect what shootings occurred - only those that were reported.

What if the police response and intensity of the investigation varies depending on when / where the crime occurred? For example, different officers and precincts might have different responses to similar situations, or the same precinct and officers might respond differently based on where and when the crime is. In other words, an officer in a Staten Island precinct might respond differently from one based in Tribeca or one in Williamsburg. More importantly, the same officer might respond to a 3pm Monday call in Tribeca (a very wealthy neighborhood and a time of day with limited crime) very differently than they would respond to a midnight call in Brownsville or Bushwick (both more empoverished areas at a time where crime is more likely to occur.) While this is not ideal, police are humans too, which can impact the data.

The author is also not immune from bias and must do his best to acknowledge and mitigate this. For example, the data seems to show a significant shift from 2019 to the 2020-2022 period. While my first reaction is to suspect this is from Covid, I should not rule other factors that were not "top of mind." For example, my bias towards suspecting governmental and public policy decisions might prevent me from noticing trends that might be more evident to someone more familiar with racial, gender, geographic, or other potential discrepancies in NYC crime and policing.

Obviously self-reflection and honesty are two useful techniques to help limit personal bias in the work, another helpful strategy is to seek the input of other individuals (particularly those or varying backgrounds). As the other is a heterosexual, cisgendered white male, getting the perspective of people from different communities would be highly beneficial. While eliminating 100% of personal bias is difficult to achieve, the goal should be to eliminate or at least strongly reduce the impact of bias while acknowledging any sources of bias that could potentially remain in the data or the analysis.

In conclusion, the data shows two particularly interesting trends. First, these incidents tend to occur most often late at night or very early morning. While the hour shifts slightly by year, the general trend still holds. Even in years with lower or higher total incident rates, the distribution of when the shootings occur is roughly the same. Second, crime rates were highest in the earlier part of the data set before declining in 2011/2012. They reach a nadir in 2019 but rebound strongly in 2020 and 2021 before dipping somewhat in 2022.

While Covid seems at least a probable explanation, I would recommend more analysis before drawing any sweeping causal conclusions. For example, one potential avenue of expansion would be to add in overall NYC demographic data. While we currently (in most cases) have information on the age, gender, and race of the victim, we don't know their proportion of the population. If two groups both account for 20% of the victims in the data set, but one group only accounts for 2% of the population, this suggests they are proportionally more likely to be the victim of a shooting than the other group. Another option is to expand the analysis with more detailed time-series analysis (perhaps SARIMA or ETS models) that is supplemented by Covid related variables (infection rate, death rate, binary variables encoding lockdown policies, etc.).

In any case, the data so far has revealed interesting trends, but more analysis is needed to reach a more comprehensive interpretation of the facts.