Analysis of NYPD Shooting Data

Ethan DeJongh

2024-04-11

Introduction

The follow dataset from the NYPD contains information about shooting incidents in New York City from 2006 through 2022. Each entry contains a precise date, time, and location of a shooting, as well as demographic information (when known) about the perpetrator and the victim. In this report, I will examine long and short term trends in the number of shooting incidents over time and attempt to use the data to explain them.

The first step for this analysis is to import the data using the URL below. I then clean up the data by transforming "OCCUR_DATE" to a date variable type and renaming the date and time variables to make them more workable. I have also added an "incidents" column with a value of 1 for every entry; this will be useful for obtaining the sum number of incidents when grouping the data. Finally, I select the variables which I deem useful for this analysis and save to a dataframe.

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

```
## Rows: 28562 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.
```

```
shooting_data <- shooting_data %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE), incidents = 1) %>%
    rename(date = OCCUR_DATE, time = OCCUR_TIME) %>%
    select(date, time, BORO, PERP_AGE_GROUP, VIC_AGE_GROUP, incidents)
summary(shooting_data)
```

```
##       date                time             BORO           PERP_AGE_GROUP
##  Min.   :2006-01-01   Length:28562      Length:28562       Length:28562
##  1st Qu.:2009-09-04   Class1:hms        Class :character   Class :character
##  Median :2013-09-20   Class2:difftime   Mode  :character   Mode  :character
##  Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:2019-09-29
##  Max.   :2023-12-29
```

```
##  VIC_AGE_GROUP        incidents
##  Length:28562       Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
```

The summary of the data doesn't reveal any missing data; however, an inspection of the first few rows of data shows missing values for the age group of the perpetrator:

```
shooting_data
```

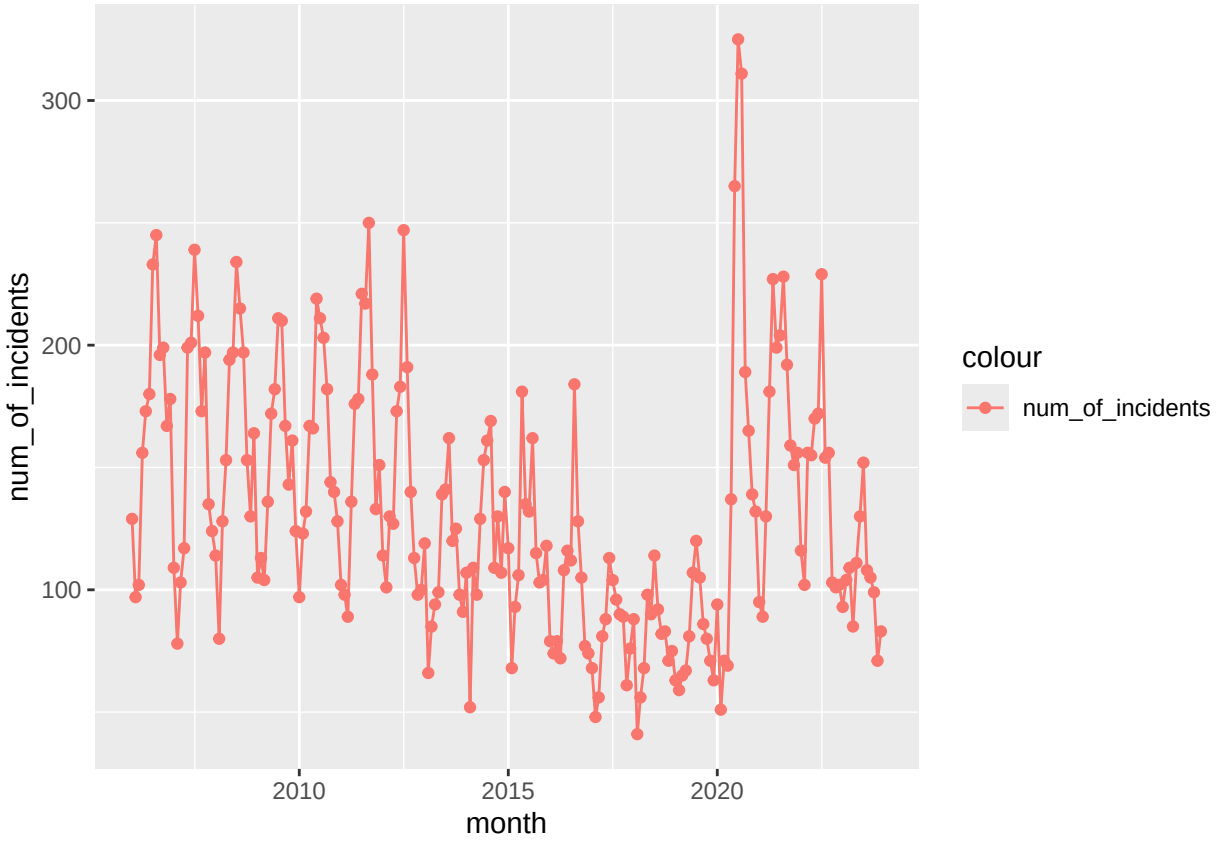
```
## # A tibble: 28,562 x 6
##    date       time   BORO      PERP_AGE_GROUP VIC_AGE_GROUP incidents
##    <date>     <time> <chr>     <chr>          <chr>              <dbl>
##  1 2022-05-05 00:10  MANHATTAN 25-44          25-44                  1
##  2 2022-07-04 22:20  BRONX     (null)         18-24                  1
##  3 2012-05-27 19:35  QUEENS    <NA>           18-24                  1
##  4 2019-09-24 21:00  BRONX     25-44          25-44                  1
##  5 2007-02-25 21:00  BROOKLYN  25-44          25-44                  1
##  6 2021-07-01 23:07  MANHATTAN <NA>           25-44                  1
##  7 2021-06-07 19:55  QUEENS    <NA>           45-64                  1
##  8 2021-07-22 01:47  BROOKLYN  <NA>           25-44                  1
##  9 2021-05-22 18:39  BRONX     <NA>           18-24                  1
## 10 2021-12-22 23:17  BRONX     25-44          25-44                  1
## # i 28,552 more rows
```

When considering age in my analysis, I will use only incidents in which the age group of the perpetrator is recorded.

Analysis

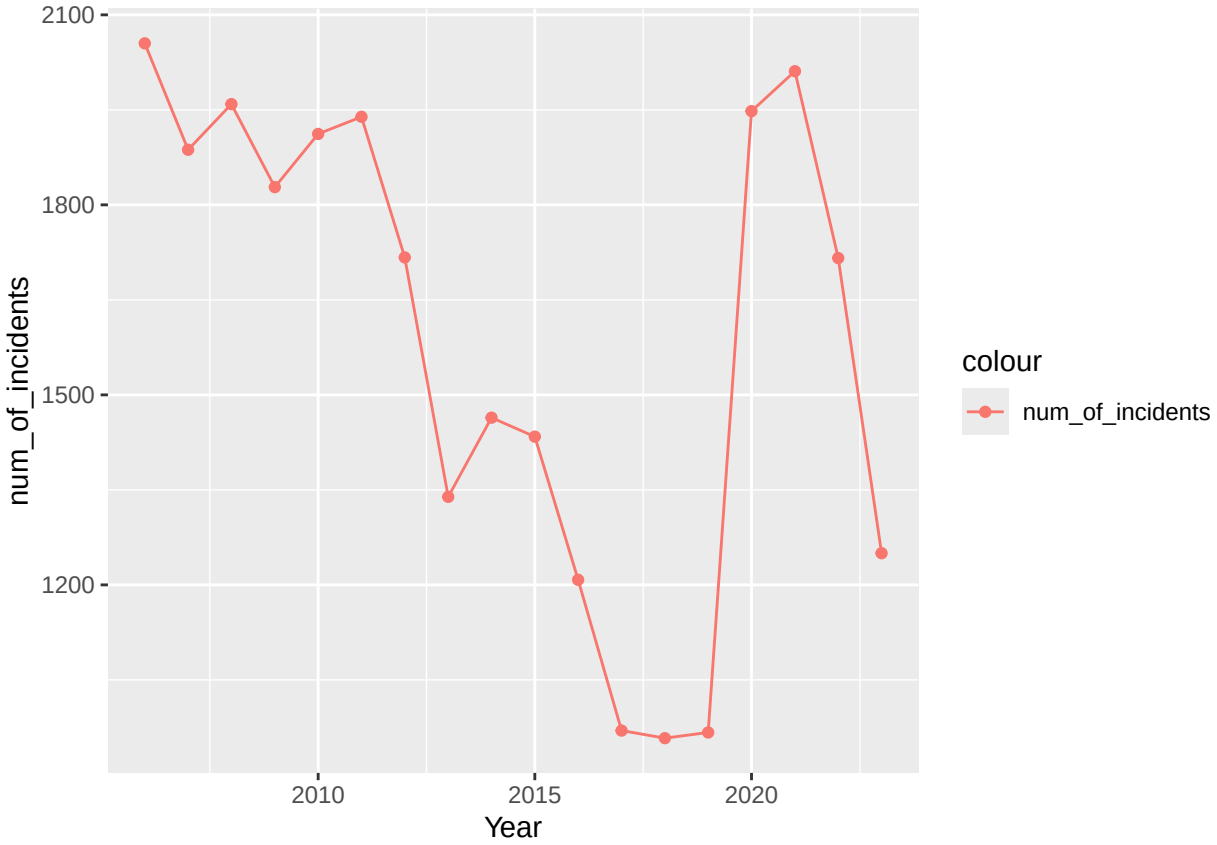
The first step in my analysis is to plot the number of incidents vs. time for the length of the dataset, grouping by month. The results are shown below:

```
shooting_data %>% group_by(month = lubridate::floor_date(date, 'month')) %>%
  summarize(num_of_incidents = sum(incidents)) %>%
  ggplot(aes(x = month, y = num_of_incidents)) +
  geom_line(aes(color = "num_of_incidents")) +
  geom_point(aes(color = "num_of_incidents"))
```



The first noticeable feature of this plot is the consistent up-and-down pattern in the number of shootings within each year. In order to more clearly visualize the long-term trend, I make a new plot grouping the number of incidents by year instead of month:

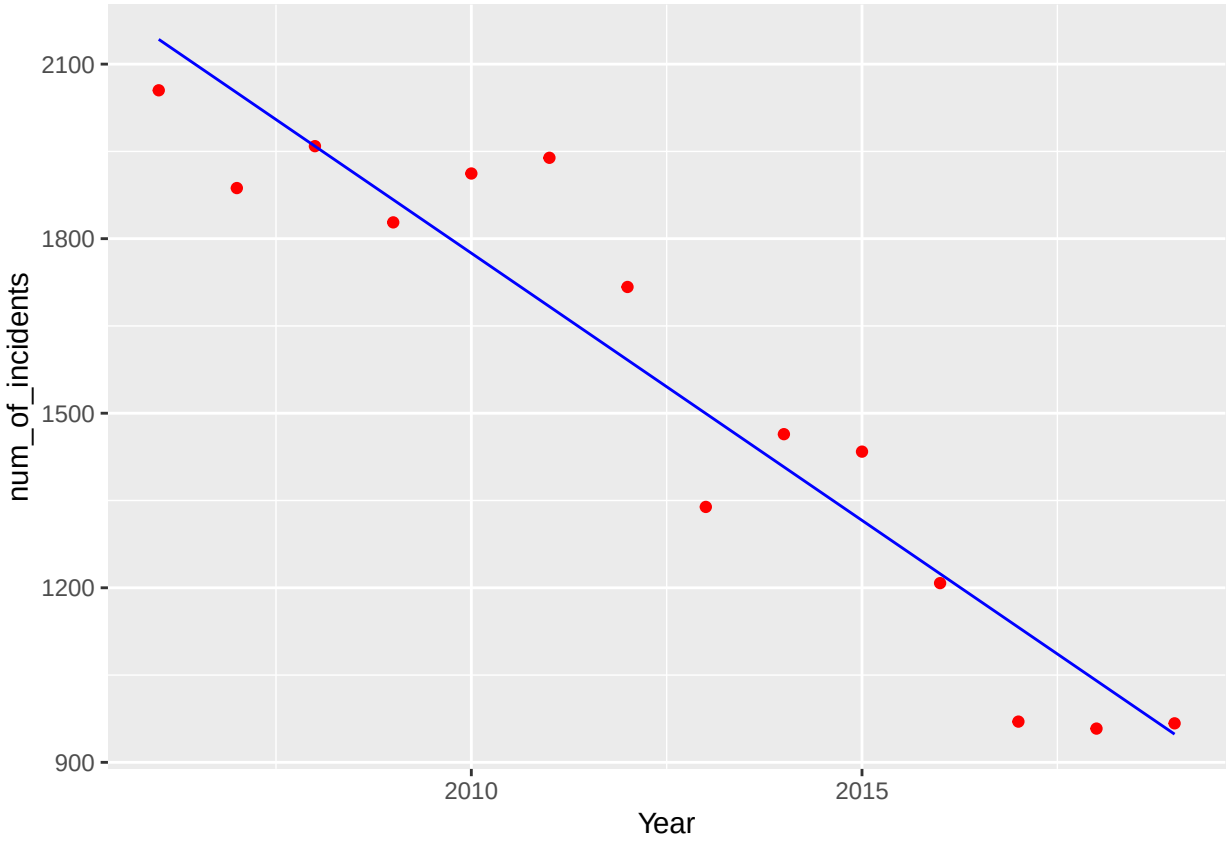
```
Incidents_by_year <- shooting_data %>%
  group_by(Year = lubridate::floor_date(date, 'year')) %>%
  summarize(num_of_incidents = sum(incidents))

Incidents_by_year %>%
  ggplot(aes(x = Year, y = num_of_incidents)) +
  geom_line(aes(color = "num_of_incidents")) +
  geom_point(aes(color = "num_of_incidents"))
```



This plot shows the number of shooting incidents per year declining steeply over most of the dataset before jumping back up in 2020, the first year of the COVID-19 pandemic. It appears that the city was on a successful path to reducing gun crime before the pandemic struck. We see incidents starting to come down again in 2022, but more years of data will be necessary to determine if this is a firm trend.

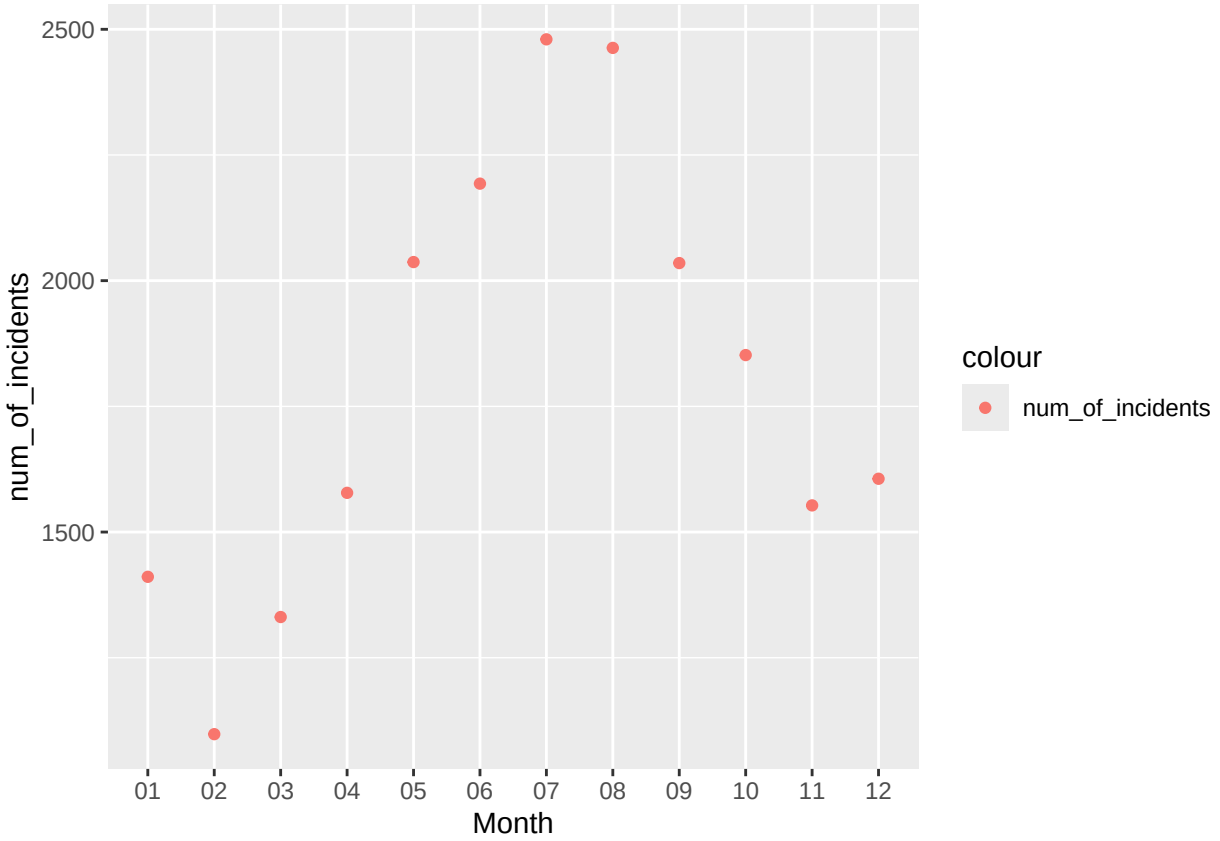
The decline in incidents from 2006 - 2019 can be represented surprisingly accurately with a linear model:

```
mod <- lm(num_of_incidents ~ Year,
          data = Incidents_by_year %>% filter(Year < '2020-01-01'))

Incidents_by_year %>% filter(Year < '2020-01-01') %>%
  mutate(pred = predict(mod)) %>%
  ggplot() +
  geom_point(aes(x = Year, y = num_of_incidents), color = "red") +
  geom_line(aes(x = Year, y = pred), color = "blue")
```



I am interested in learning more about the consistent intra-year pattern in the number of shootings. To examine this trend, I sum the number of incidents by month, independent of year. Because COVID-19 likely disrupted the normal pattern, I am looking at only data from before 2020 for this analysis.

```
Incidents_by_month <- shooting_data %>% filter(date < '2020-01-01') %>%
  group_by(Month = format(date, "%m")) %>%
  summarize(num_of_incidents = sum(incidents))

Incidents_by_month %>%
  ggplot(aes(x = Month, y = num_of_incidents)) +
  geom_point(aes(color = "num_of_incidents"))
```



This plot reveals a clear trend that more shootings occur during the summer months than the winter months. My first thought is that this could be related to teenagers being out of school during the summer. To investigate, I group the number of incidents by month and age group of the perpetrator. After pivoting the dataframe so that the monthly totals from different age groups are listed as columns, I sum the totals from all adult age groups and compare these to the totals from the <18 age group. In the plot below, youth incidents are shown in red, while adult incidents are shown in blue.

```
Age_Group_Incidents <- shooting_data %>%
  filter(date < '2020-01-01') %>%
  group_by(Month = format(date, "%m"), PERP_AGE_GROUP) %>%
  summarize(num_of_incidents = sum(incidents))
```

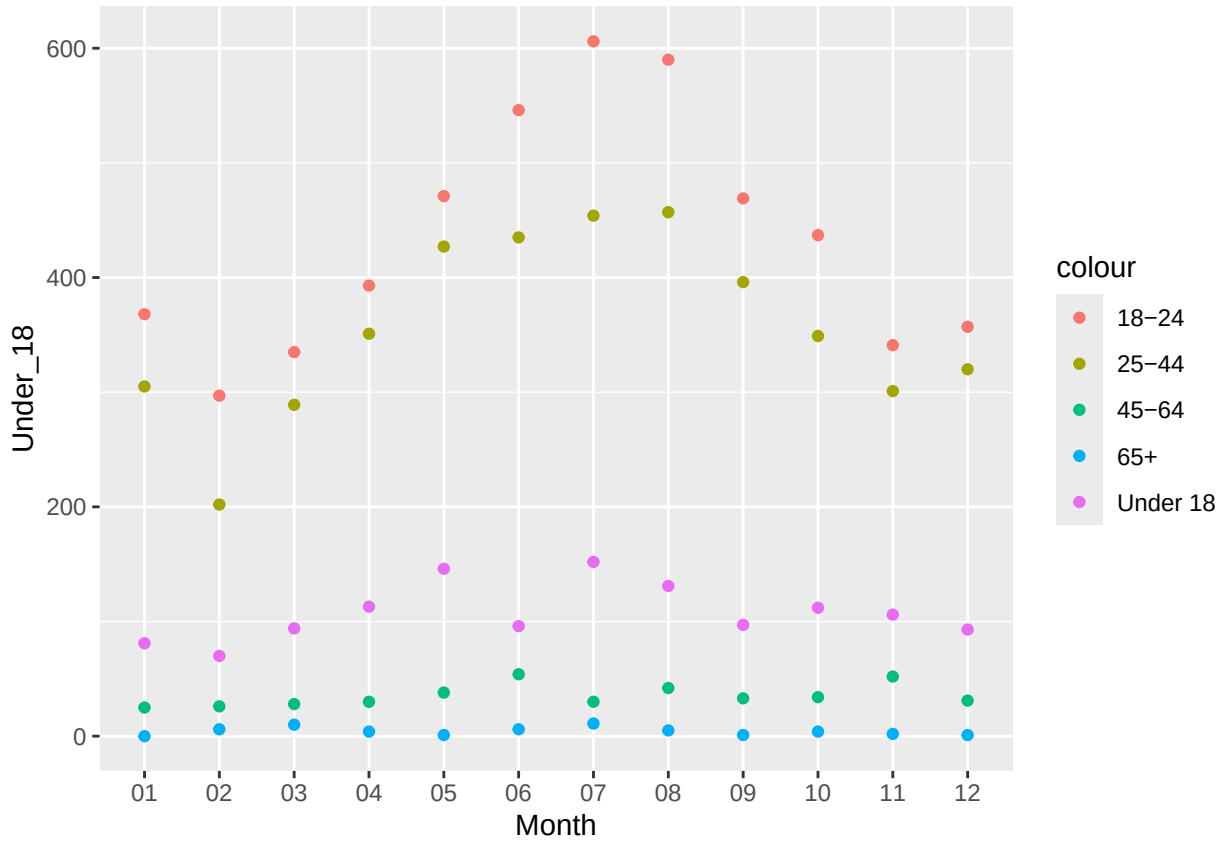
```
## `summarise()` has grouped output by 'Month'. You can override using the
## `.groups` argument.
```

```
Age_Group_Incidents <- Age_Group_Incidents %>%
        pivot_wider(names_from = PERP_AGE_GROUP,
                    values_from = num_of_incidents) %>%
  select("Month","<18","18-24","25-44","45-64","65+")
Age_Group_Incidents <- Age_Group_Incidents %>%
  replace(is.na(.),0) %>%
  rename(Under_18 = "<18", Age18_24 = "18-24", Age25_44 = "25-44", Age45_64 = "45-64", Over65 = "65+")
Age_Group_Incidents
```

```
## # A tibble: 12 x 6
## # Groups:   Month [12]
```

```
##    Month Under_18 Age18_24 Age25_44 Age45_64 Over65
##    <chr>    <dbl>    <dbl>    <dbl>    <dbl>  <dbl>
##  1 01          81      368      305       25      0
##  2 02          70      297      202       26      6
##  3 03          94      335      289       28     10
##  4 04         113      393      351       30      4
##  5 05         146      471      427       38      1
##  6 06          96      546      435       54      6
##  7 07         152      606      454       30     11
##  8 08         131      590      457       42      5
##  9 09          97      469      396       33      1
## 10 10         112      437      349       34      4
## 11 11         106      341      301       52      2
## 12 12          93      357      320       31      1
```

```
#Age_Group_Incidents <- Age_Group_Incidents %>%
#  replace(is.na(.), 0) %>%
#  mutate(Adult = sum(c_across(c(2:5)))) %>%
#  rename(Youth = "<18")

#Age_Group_Incidents %>%
#  ggplot() +
#  geom_point(aes(x = Month, y = Youth), color = "red") +
#  geom_point(aes(x = Month, y = Adult), color = "blue")

ggplot() +
  geom_point(data=Age_Group_Incidents, aes(Month, Under_18, color='Under 18'))  +
  geom_point(data=Age_Group_Incidents, aes(Month, Age18_24, color='18-24'))  +
  geom_point(data=Age_Group_Incidents, aes(Month, Age25_44, color='25-44'))  +
  geom_point(data=Age_Group_Incidents, aes(Month, Age45_64, color='45-64'))  +
  geom_point(data=Age_Group_Incidents, aes(Month, Over65, color='65+'))
```



This plot shows that the overall increase in shootings during the summer months can't be explained by a lack of school, as the bulk of the increase comes from adult perpetrators.

```
Month_time <- shooting_data %>% filter(date < '2020-01-01') %>%
  group_by(Hour = as.numeric(time) %/% 3600, Month = format(date, "%m")) %>%
  summarize(num_of_incidents = sum(incidents))
```

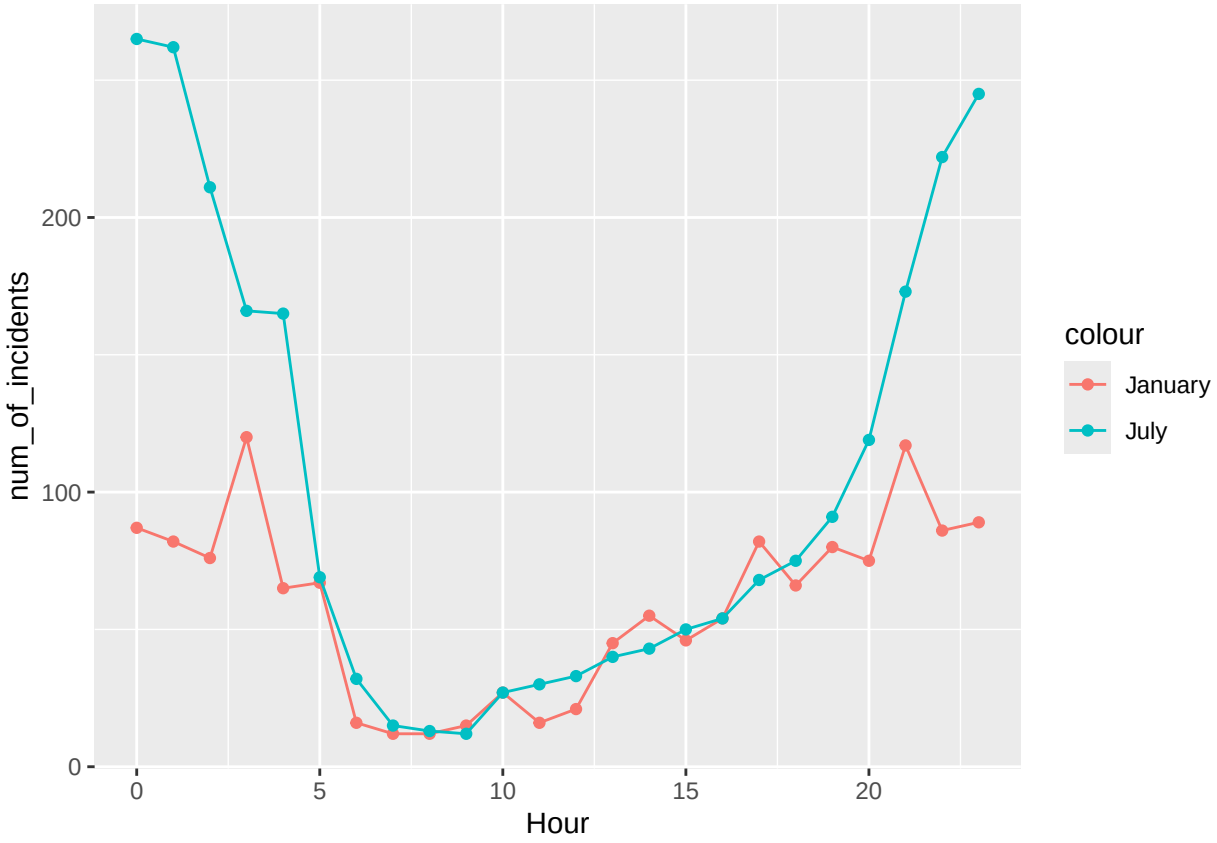
```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

```
Month_time
```

```
## # A tibble: 288 x 3
## # Groups:   Hour [24]
##     Hour Month num_of_incidents
##    <dbl> <chr>             <dbl>
##  1     0 01                   87
##  2     0 02                   69
##  3     0 03                  104
##  4     0 04                  107
##  5     0 05                  148
##  6     0 06                  225
##  7     0 07                  265
##  8     0 08                  250
##  9     0 09                  155
```

```
## 10     0 10                  126
## # i 278 more rows
```

```
ggplot() +
  geom_point(data=Month_time %>% filter(Month=="01"), aes(Hour, num_of_incidents, color="January"))  +
  geom_line(data=Month_time %>% filter(Month=="01"), aes(Hour, num_of_incidents, color="January")) +
  geom_point(data=Month_time %>% filter(Month=="07"), aes(Hour, num_of_incidents, color="July"))  +
  geom_line(data=Month_time %>% filter(Month=="07"), aes(Hour, num_of_incidents, color="July"))
```



Conclusion

This analysis of shooting incidents in New York City has revealed interesting trends. Prior to the COVID-19 pandemic, shootings were on a steep decline, but drastically increased at the start of the pandemic. Regardless of the yearly total, there is a consistent trend of increased shootings during the summer months, and this pattern holds across age groups. Since I am favorably inclined toward the public school system, my personal bias might have led me to jump to the conclusion that the lack of school during the summer was responsible for higher numbers of shootings. However, further analysis of the data did not support this hypothesis. At this point, the question of what causes the consistent summer increase remains unresolved. My next step would be to determine if the effect is weather-related. To do this, I would need to import additional data concerning the weather for the dates and locations in this dataset and look for correlations between weather patterns and the number of shootings.