NYPD Shootings EDA

Swapnil Bhatta

2022-11-15

How to get the data

Navigate to the dataset link https://catalog.data.gov/dataset and search for a dataset titled NYPD Shooting Incident Data (Historic). The repo has the csv file, but the script automatically downloadeds it from the website when ran.

```
nypd_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
police_shootings <- read.csv(nypd_url,
                             header=TRUE,
                             sep=",")
```

```
summary(police_shootings)
```

```
##   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
##
```

What cleaning process was used?

- Changed the type for OCCUR_DATE to date
- Removed unwanted columns

```
# Remove unwanted columns
police_shootings <- select(police_shootings, -c(LOCATION_DESC, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP,
```

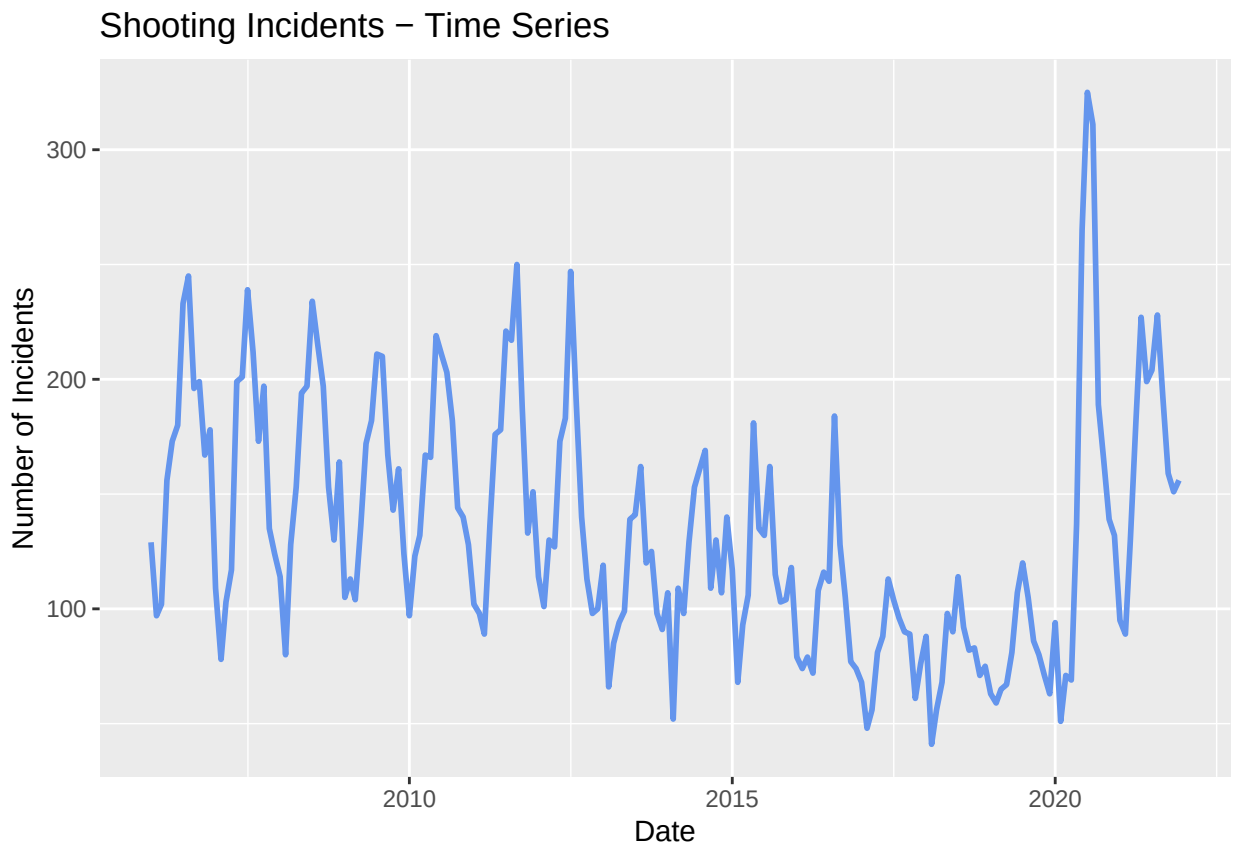
```
# Convert string to datetime
police_shootings$ OCCUR_DATE <- mdy(police_shootings$ OCCUR_DATE)
```

Daily frequency of shootings

```
# Add columns for monthly, yearly averages
police_shootings <- police_shootings %>%
    mutate(YEAR_MONTH = floor_date(police_shootings$ OCCUR_DATE, 'month')) %>%
    mutate(MONTH = strftime(police_shootings$ OCCUR_DATE, format='%m'))
```

```
date_value_counts <- police_shootings %>% count(YEAR_MONTH)
ggplot(date_value_counts, aes(x=YEAR_MONTH, y=n)) +
                        geom_line(color = "cornflowerblue", size=1) +
                        labs(x = "Date", y = "Number of Incidents", title='Shooting Incidents - Time Se
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

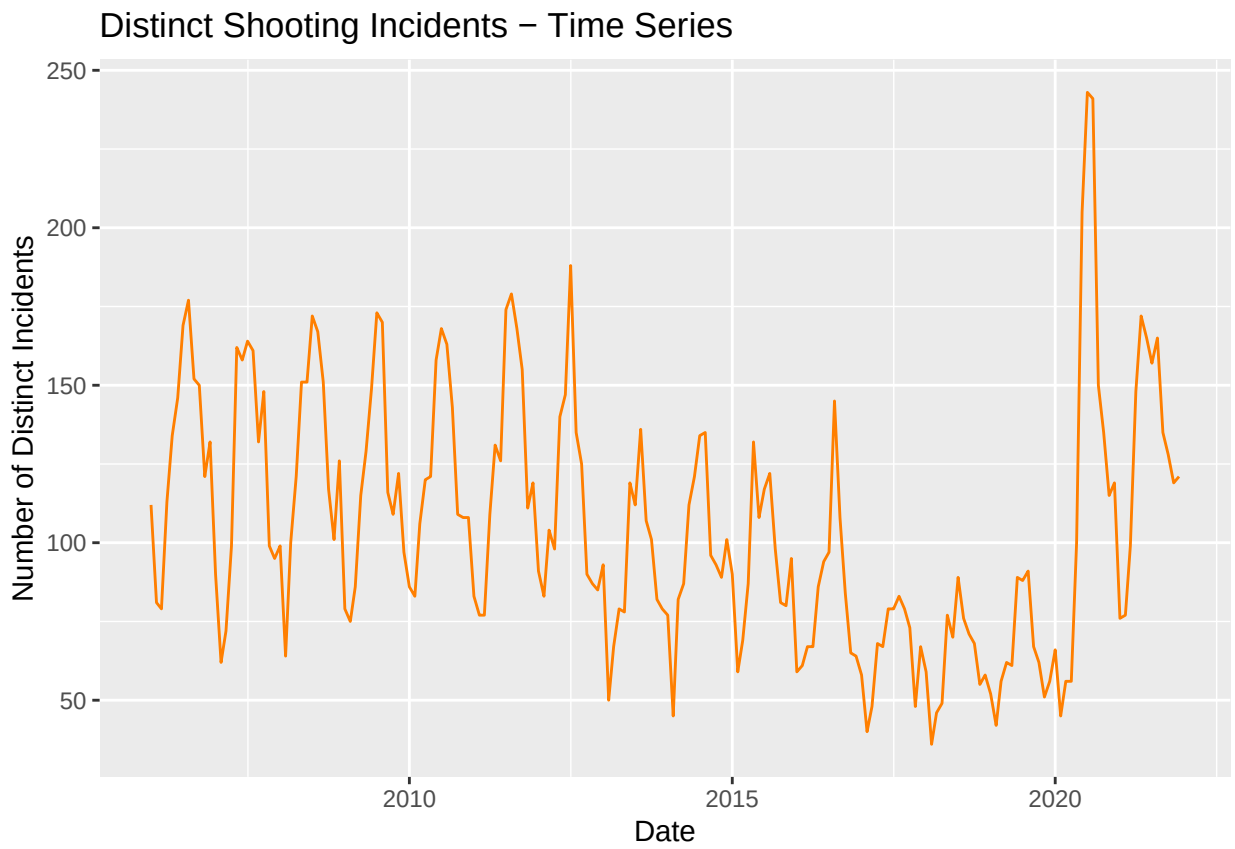



```
# Checking if some incidents are reported additional times for multiple shootings
incident_counts <- police_shootings %>% count(INCIDENT_KEY)
incident_counts[order(incident_counts$ n, decreasing = TRUE),] %>% head(10)
```

```
##       INCIDENT_KEY  n
## 15510    173354054 18
## 867       23749375 12
## 1209      24717013 12
## 2358      33478089 12
## 2428      33706902 12
## 2876      35803777 12
## 5578      66027258 12
## 6279      72195829 12
## 6362      72616285 12
## 7966      79378503 12
```

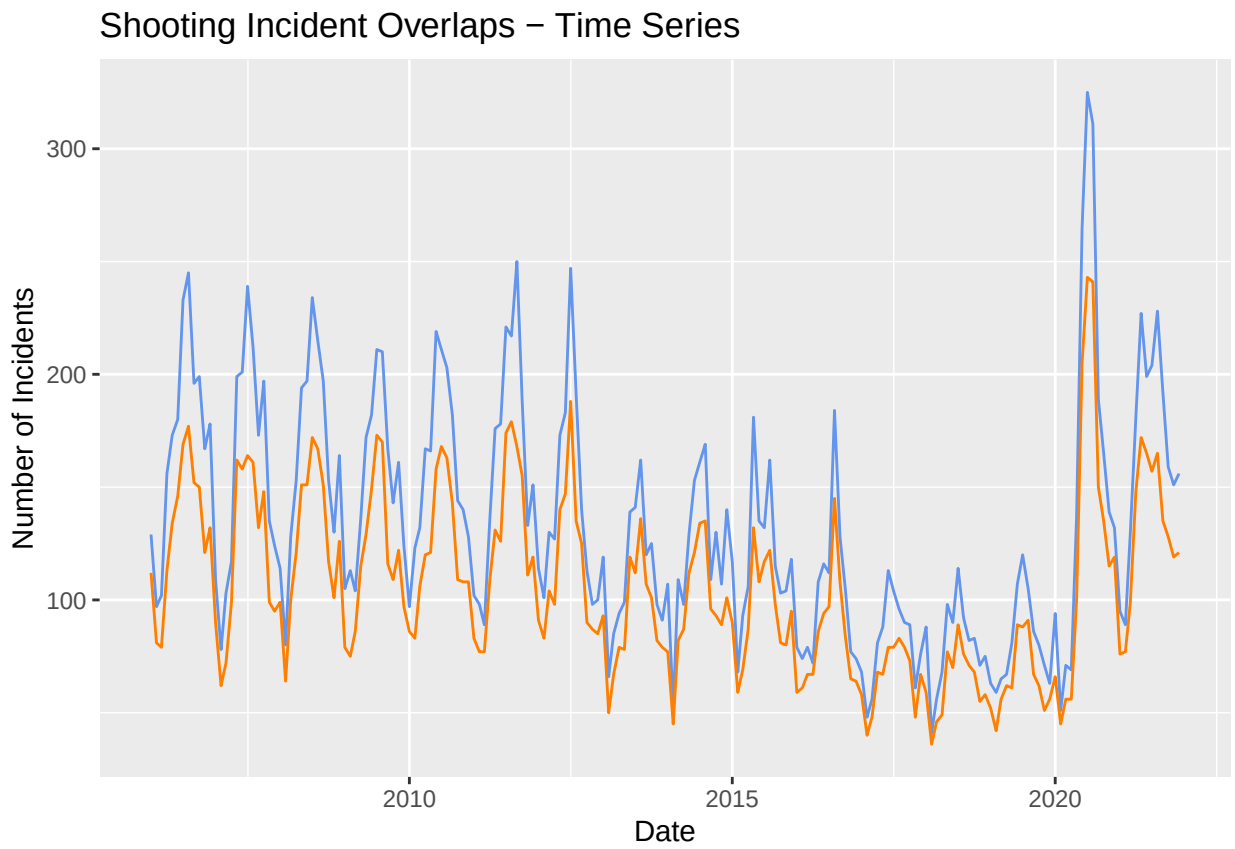
We see that an incident can have multiple data points, plotting only unique counts confirms that there might've been some outliers

```
distinct_year_month <- police_shootings %>% group_by(YEAR_MONTH) %>%
                    summarize(distinct_incident = n_distinct(INCIDENT_KEY))
ggplot(distinct_year_month, aes(x=YEAR_MONTH, y=distinct_incident)) +
                geom_line(color = "darkorange1") +
                labs(x = "Date", y = "Number of Distinct Incidents", title='Distinct Shooting Incidents
```



Visualizing overlap of incidents

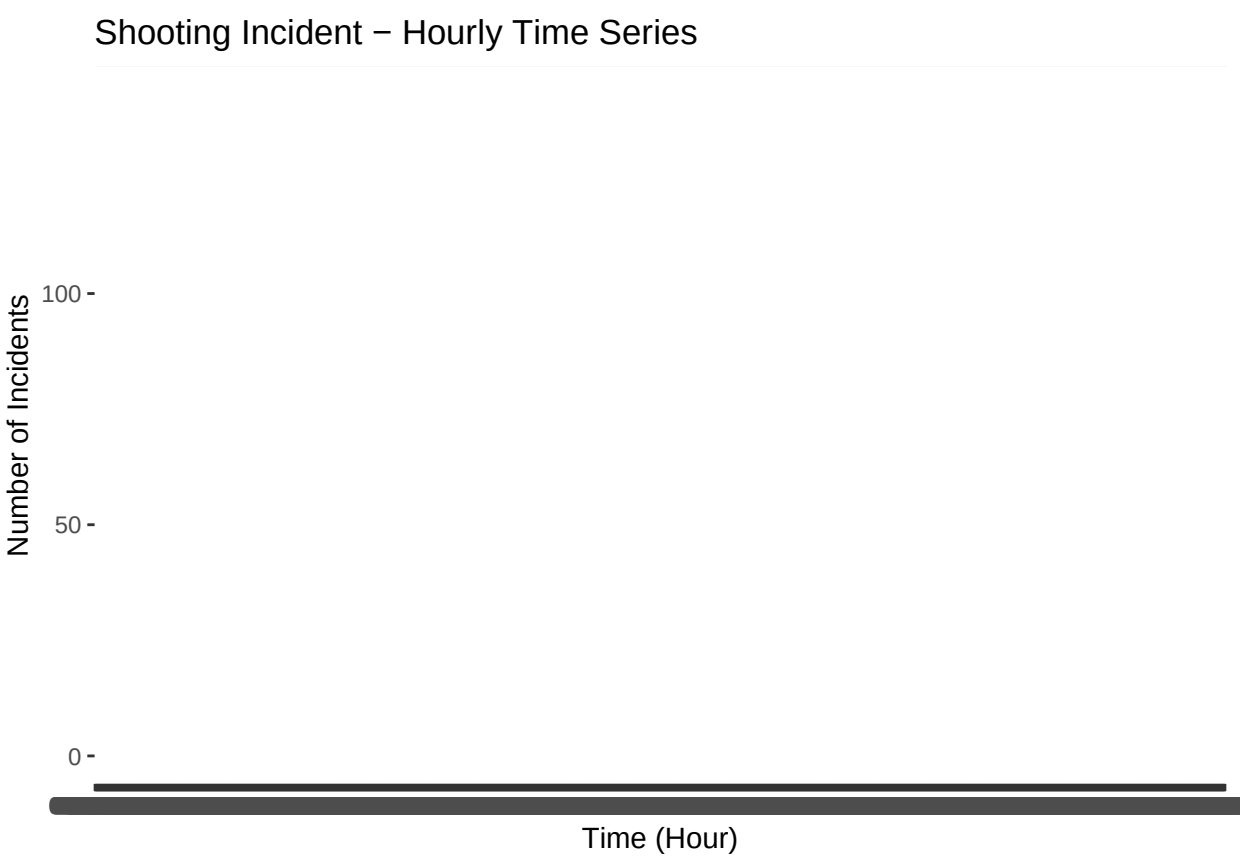
```
ggplot(date_value_counts, aes(x=YEAR_MONTH, y=n)) +
                        geom_line(color = "cornflowerblue") +
                        geom_line(data = distinct_year_month, aes(x=YEAR_MONTH, y=distinct_incident), co
                    labs(x = "Date", y = "Number of Incidents", title='Shooting Incident Overlaps - Time
```



Hourly breakdown for shooting incidents

```
distinct_hours <- police_shootings %>% group_by(OCCUR_TIME) %>%
                    summarize(distinct_incident = n_distinct(INCIDENT_KEY))
ggplot(distinct_hours, aes(x=OCCUR_TIME,
                           y=distinct_incident)) + geom_line(color = "blue")+
                    labs(x = "Date", y = "Number of Incidents", title='Shooting Incident Overlaps - Tim
                    labs(x = "Time (Hour)", y = "Number of Incidents", title='Shooting Incident - Hourl
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

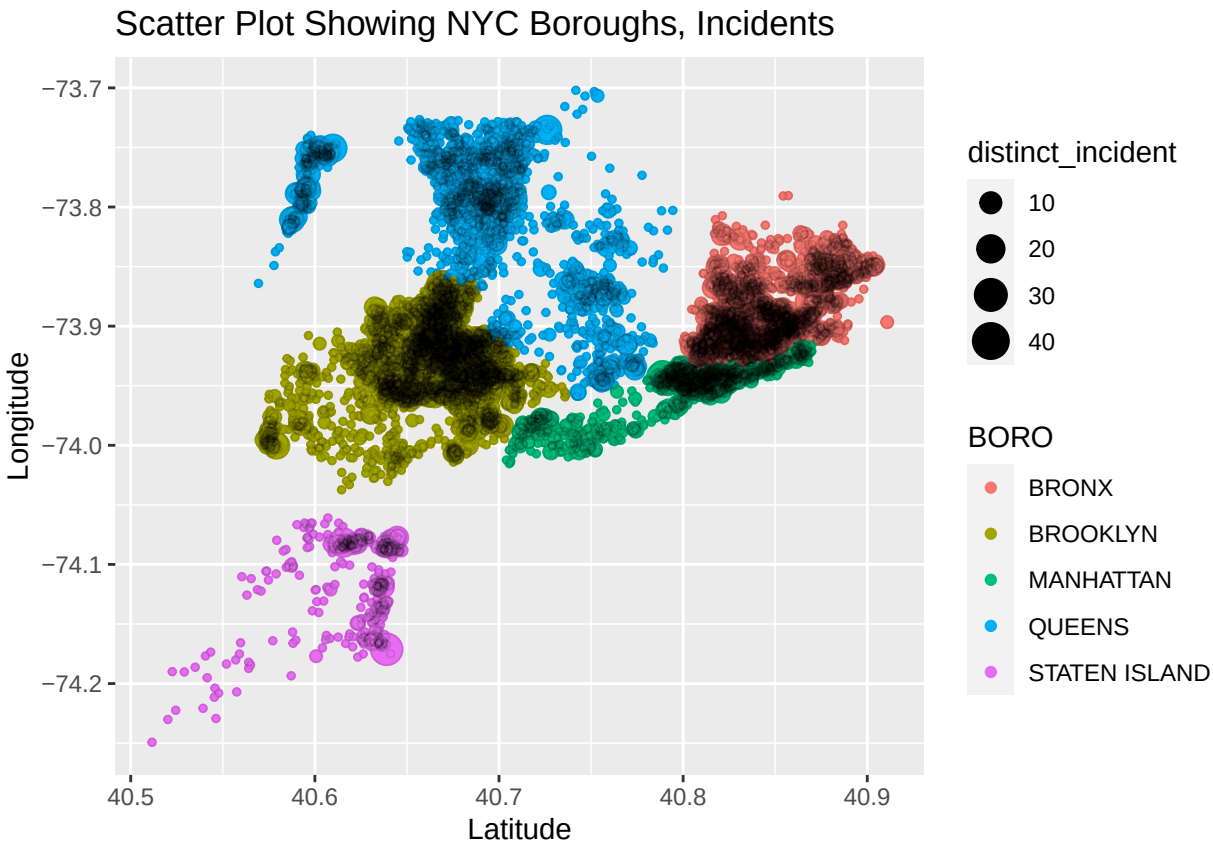


Most shooting incidents take place after hours, between 8 PM and 4 AM

```
distinct_lat_long <- police_shootings %>% group_by(Latitude, Longitude, BORO) %>%
                     summarize(distinct_incident = n_distinct(INCIDENT_KEY))
```

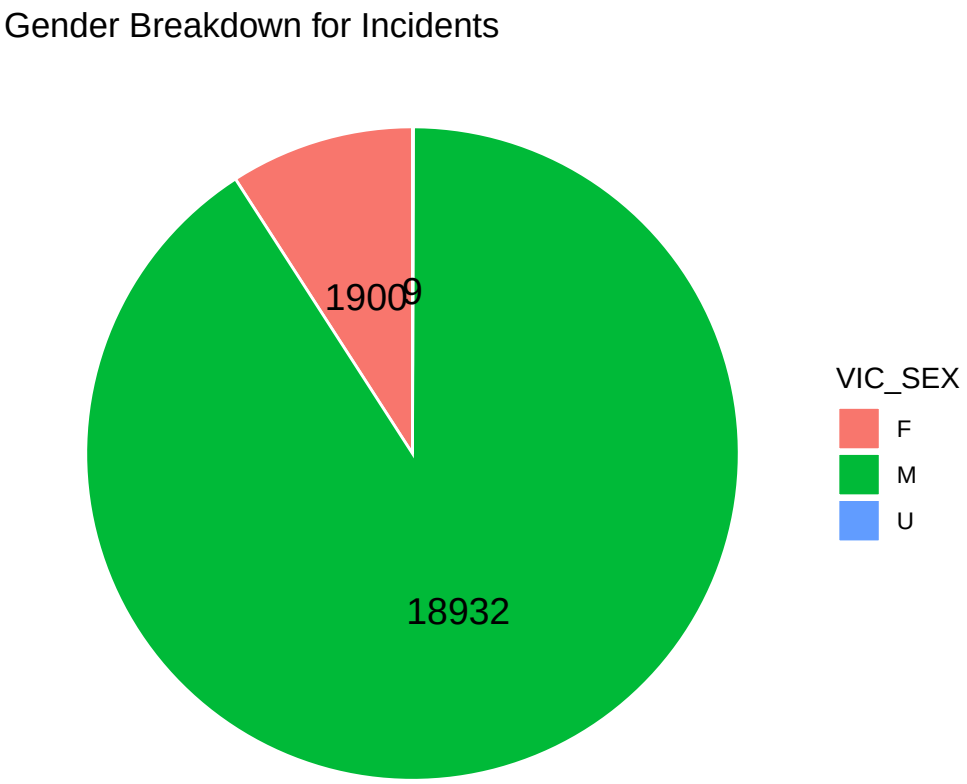
```
## `summarise()` has grouped output by 'Latitude', 'Longitude'. You can override
## using the `.groups` argument.
```

```
ggplot(distinct_lat_long, aes(x=Latitude, y=Longitude, color=BORO)) +
        geom_point(aes(size=distinct_incident)) +
        geom_point(shape = 1, aes(size=distinct_incident), alpha = 0.1, colour = "black") +
                     labs(x = "Latitude", y = "Longitude", title='Scatter Plot Showing NYC Boroughs, Inci
```



We can visualize the approximate location of boroughs and the size of incidents for various locations

```
distinct_gender <- police_shootings %>% group_by(VIC_SEX) %>%
                    summarize(distinct_incident = n_distinct(INCIDENT_KEY))
ggplot(distinct_gender, aes(x="", y=distinct_incident, fill=VIC_SEX, label = distinct_incident)) +
                geom_bar(stat="identity", width=1, color="white") +
                          coord_polar("y", start=0) + theme_void() +
                geom_text(aes(label = distinct_incident), position = position_stack(vjust = 0.5), size=5
                    labs(title='Gender Breakdown for Incidents')
```

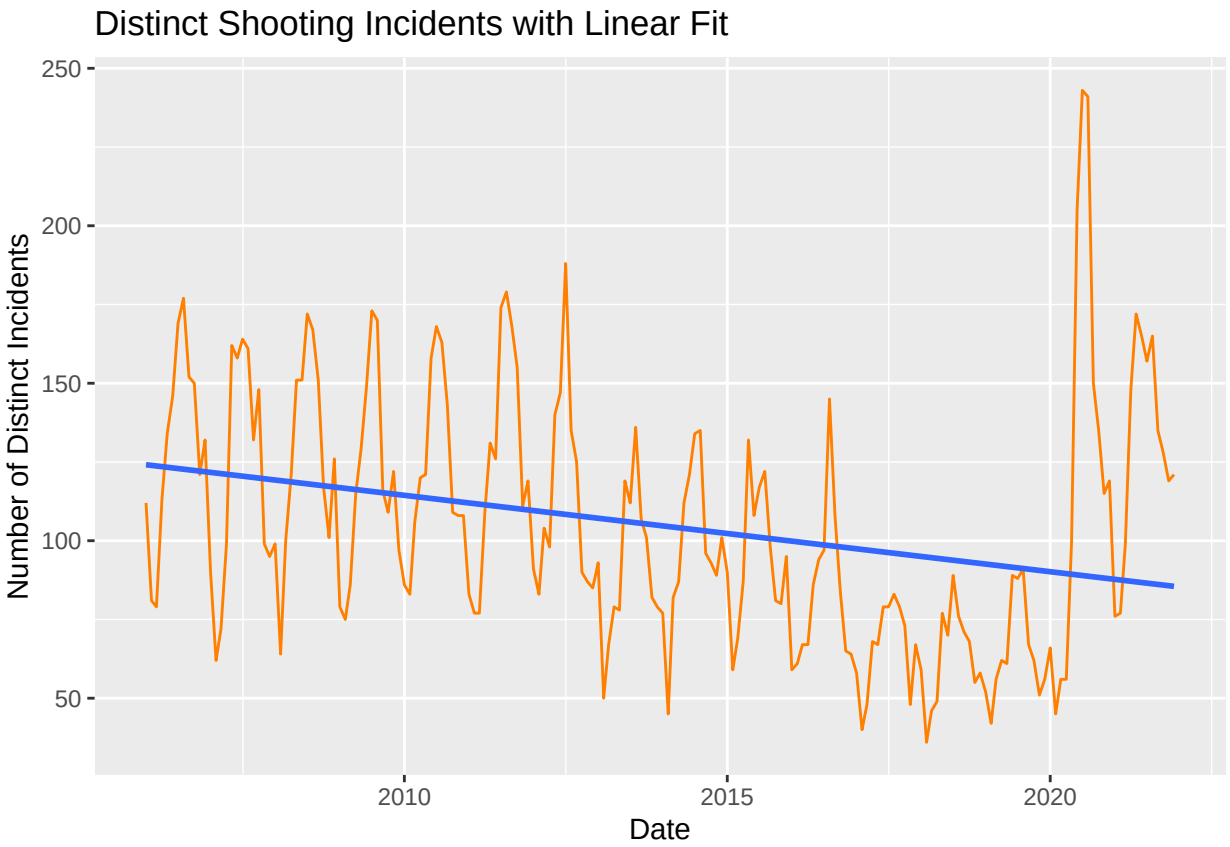


Add a linear model to see if number of shootings have been increasing

```
linear_incidents = lm(distinct_incident~YEAR_MONTH, data=distinct_year_month)
```

```
ggplot(distinct_year_month, aes(x=YEAR_MONTH,
                                y=distinct_incident, group = 1)) +
                                geom_line(color = "darkorange1") +
                                geom_smooth(method='lm', se = FALSE) +
                    labs(x = "Date", y = "Number of Distinct Incidents", title='Distinct Shooting Incid
```

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



The linear model suggests the number of shooting has decreased

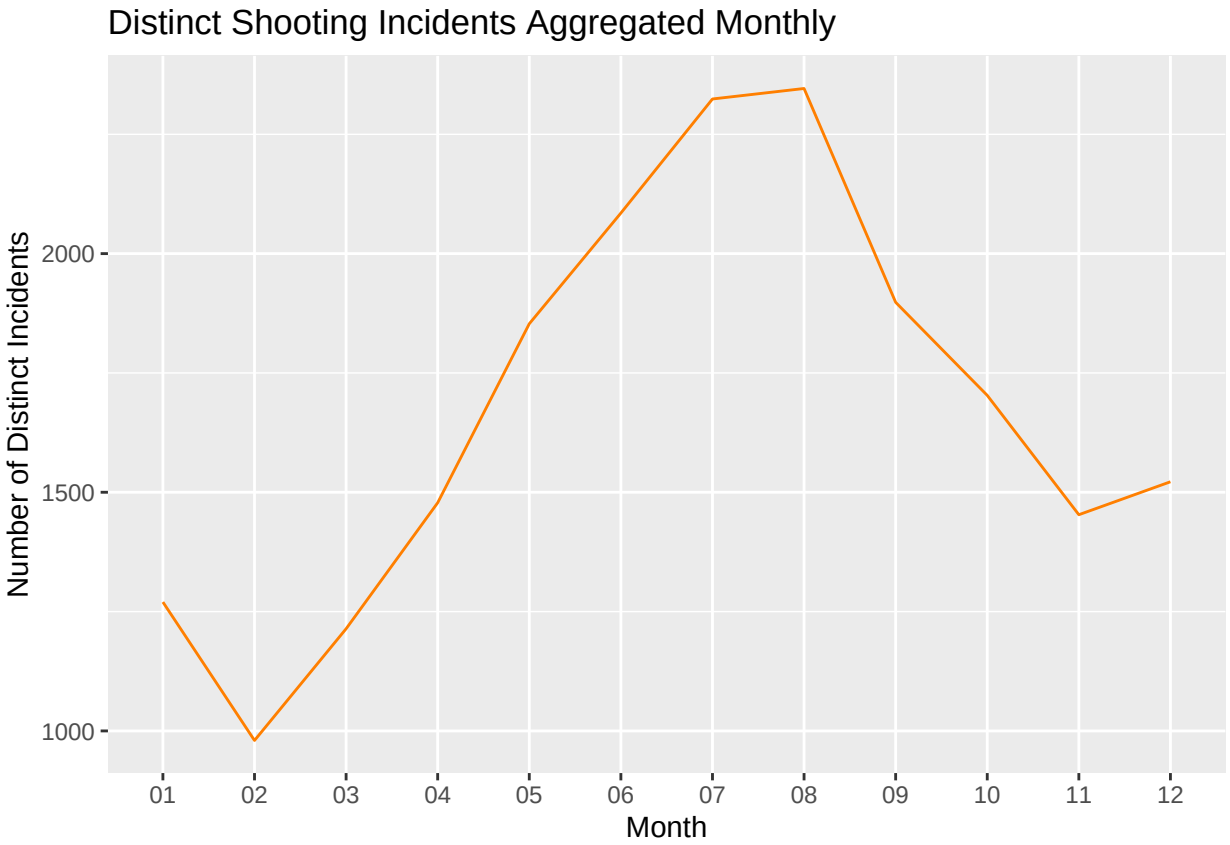
```
summary(linear_incidents)
```

```
##
## Call:
## lm(formula = distinct_incident ~ YEAR_MONTH, data = distinct_year_month)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.511 -28.362  -6.566  27.068 154.043
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 211.448032  26.170392   8.080 7.36e-14 ***
## YEAR_MONTH   -0.006641   0.001621  -4.097 6.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.9 on 190 degrees of freedom
## Multiple R-squared:  0.08116,    Adjusted R-squared:  0.07633
## F-statistic: 16.78 on 1 and 190 DF,  p-value: 6.2e-05
```

The R value suggests a poor fit, which is expected given the seasonality in the data and the outlier for the pandemic years.

```
distinct_monthly <- police_shootings %>% group_by(MONTH) %>%
                                summarize(distinct_incident = n_distinct(INCIDENT_KEY))
```

```
ggplot(distinct_monthly, aes(x=MONTH,
                             y=distinct_incident, group = 1)) +
                             geom_line(color = "darkorange1") +
                  labs(x = "Month", y = "Number of Distinct Incidents", title='Distinct Shooting Inci
```



Summer months have the highest number of shootings.

Possible Bias

The collected data itself can be biased due to higher police presence and incidents in communities that have predominantly minority population, further exploration of the PERP_RACE fields and understanding of New York City's racial distribution areas can help shed more light into it.