

# A review of previous data on trends in shootings in NYC

A.M

2023-04-09

## Abstract

Looking at trends in data can help make decisions about budgeting and allocating resources for government entities, businesses etc. It is why it is important to properly parse, summarize and analyze such data. This study revisits an earlier study(fictional) and reviews the methods by which the data was analyzed, and challenges the conclusions drawn by the previous study. The 4 Major boroughs were included in this study. Brooklyn, Bronx, Queens, and Manhattan. Staten Island was excluded due to having many missing values for the needed study, more data points were needed than were present. Data was compiled and trended in order to infer whether or not there were changes in the levels of crime seasonally in the area of New York City.

## Introduction

New York City is a massive area consisting of several Boroughs. It is also the county seat of New York County. With a 2023 population of 7,888,121, it is the largest city in New York and the largest city in the United States.(\*). This makes it a great candidate for obtaining large amounts of data. This paper details the analysis of one such data set. Namely the historical data of NYC shootings starting in 2006 to 2018. And whether or not the amounts of homicides is changing seasonally. The hypothesis is stated below:

<https://worldpopulationreview.com/us-cities/new-york-city-ny-population>\*

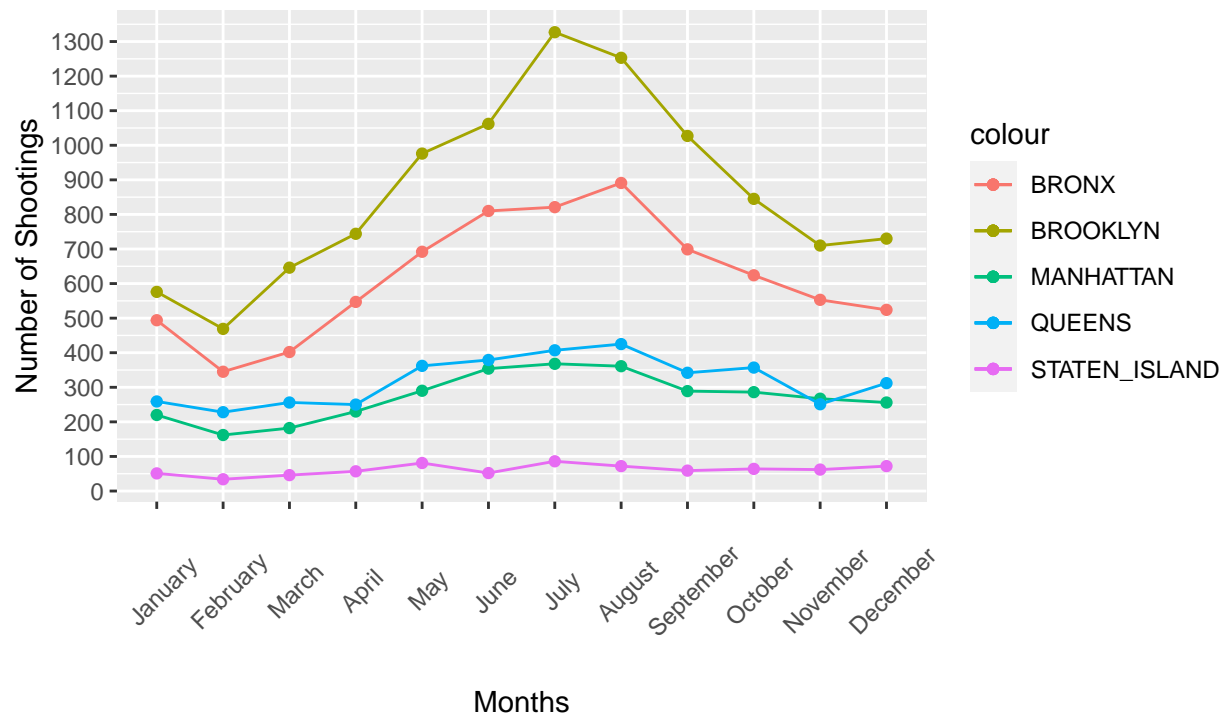
$H_0$  = There is an increase in the number of shootings seasonally.

$H_a$  = There is no change in the number of shootings during different seasons.

This paper is written in order to further explore a previous finding where an apparent trend was observed in the numbers of shootings versus the months of the year. With an upswing during the summer months and a trend downward during the winter and fall. A figure from the previous study is shown below.

## Seasonal Changes in Crime in New York City

### Crimes VS Months



Data Source: data.gov

## Materials and methods

The data set came from the .gov data catalog, which is a huge repository of information that is free for all to use. <https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic> The data was in the form of a CSV file.

### The libraries used:

```
dplyr
lubridate
tidyverse
stringr
data.table
ggplot2
forecast
```

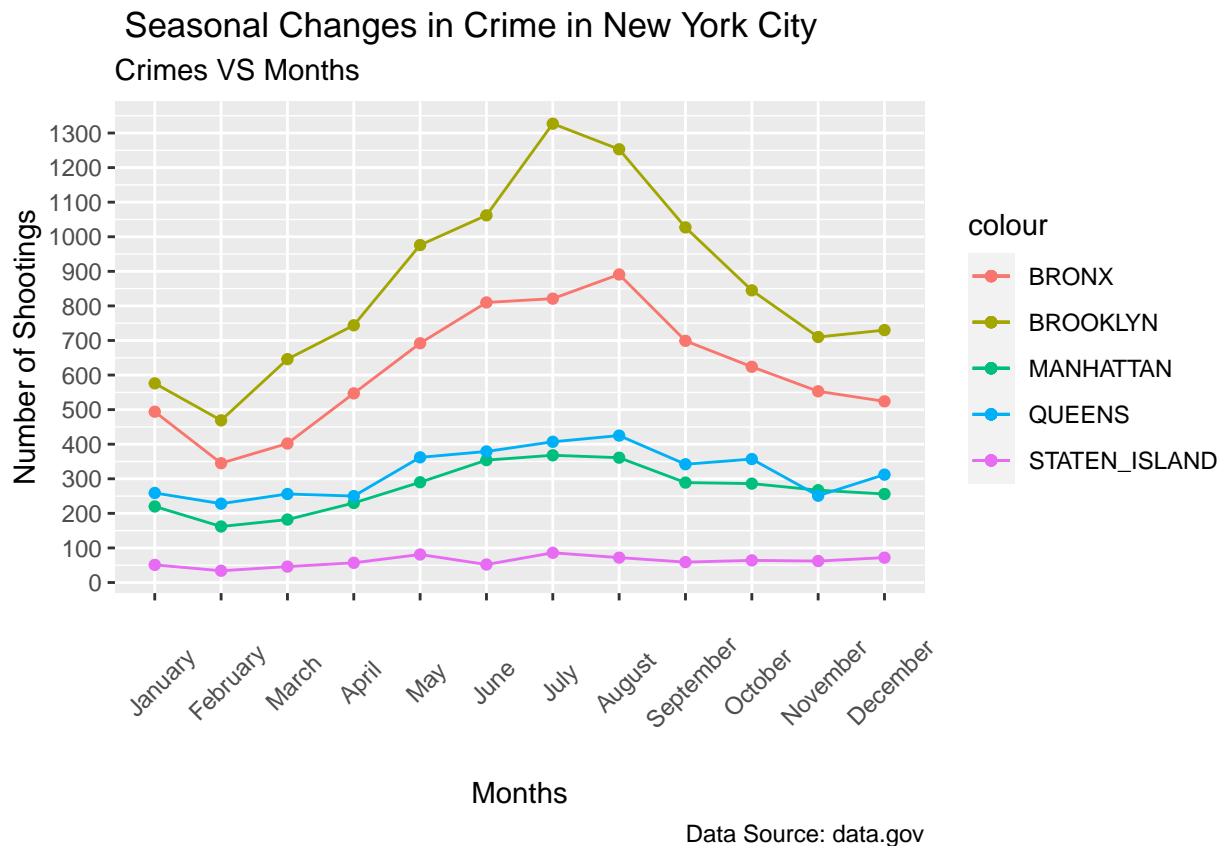
The data was loaded into R using a built in function and filtered for the occurrence date and the Borough. The data was then further subdivided into a separate section for each Borough. The stringr library and built in functions were used. This was done in order to simplify data visualization and make it easier to spot obvious differences and outliers in each data set. Ggplot2 was utilized for the data visualization portion. Each individual data set then had linear modeling applied to the resulting data.

```
## Warning in cbind(cause, annual_deaths, odds_one_out_of, odds_2021_Brooklyn, :
## number of rows of result is not a multiple of vector length (arg 4)
```

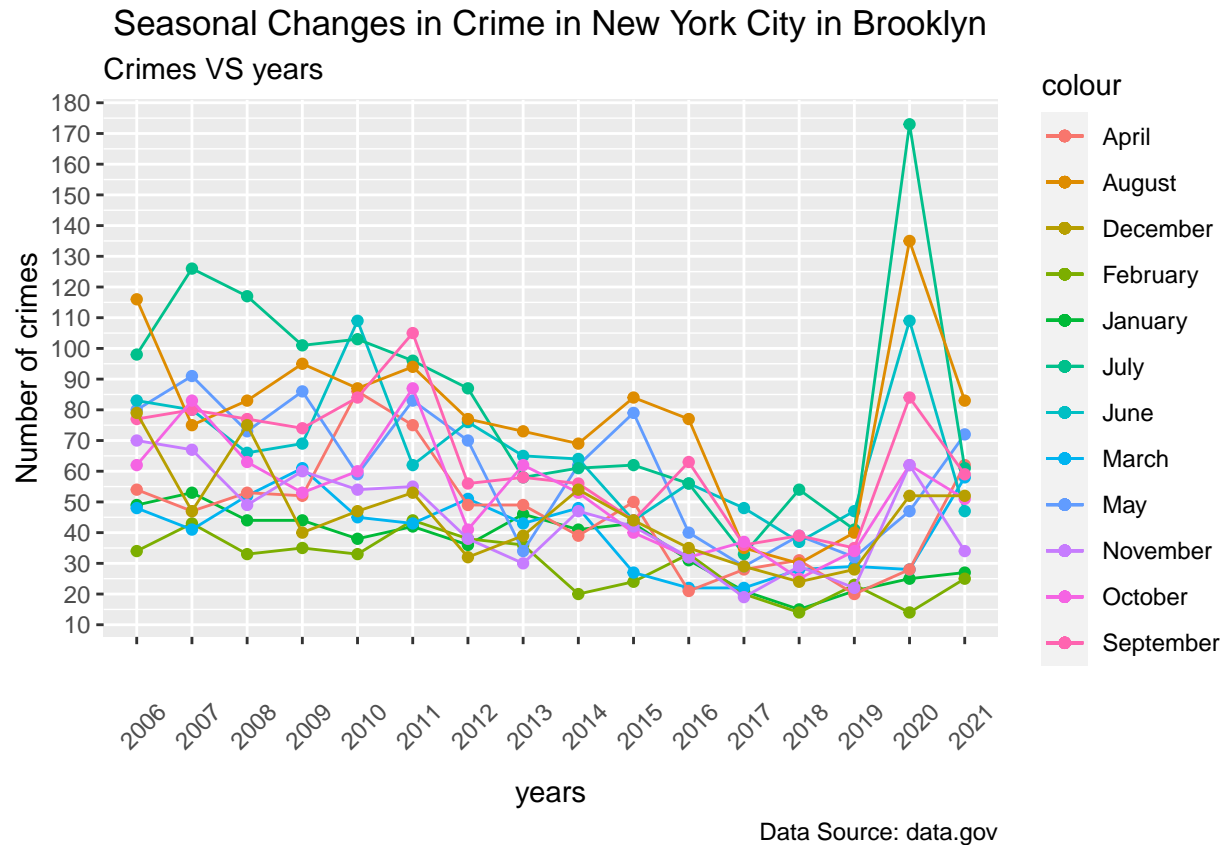
## Results and discussion:

This data set is a perfect example of how numerical data can be interpreted in many ways. And why it is important to look over results and be cautious and mindful when working with data to come to some conclusion. The analyst saw a first supportive result and drew a conclusion from said result. instead of exploring the data further. And performing further analysis. At first glance the numbers seem alarming. But upon closer inspection. It became quite clear that the data was summed over many years, rather than trending. When columns are all summed simply by month instead of using any type of trending, if there is an outlier it may and will often skew results. Additionally what this visualization does not show would be the changes from year to year, not just season to season. Or simply, more data points in general.

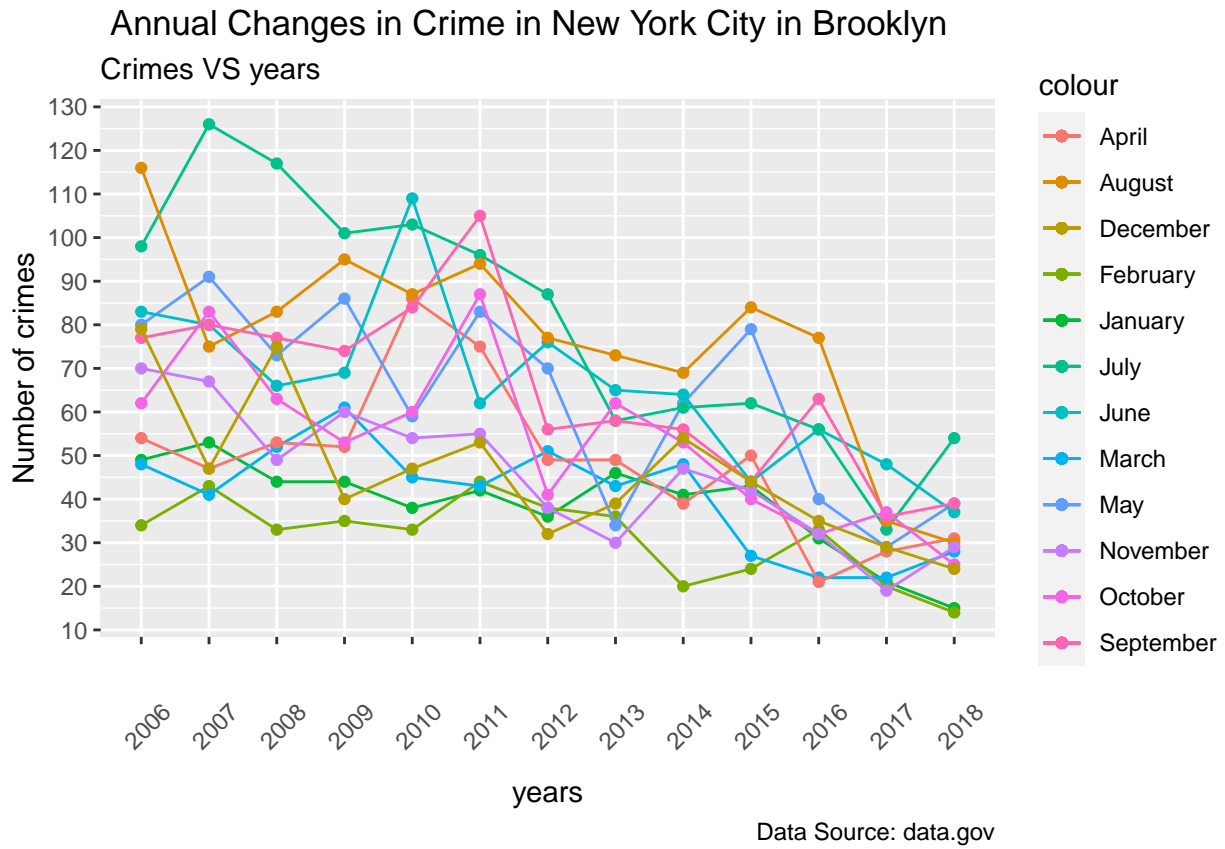
```
## [1] "Figure 1"
```



When the data was divided further and inspected more closely. Some visual differences became quickly apparent. For example when the data for the Borough of Brooklyn was further explored. Some variables that would skew results were quickly seen. So those years at least for now were excluded. The data should of course be revisited and retrended once more data becomes available.

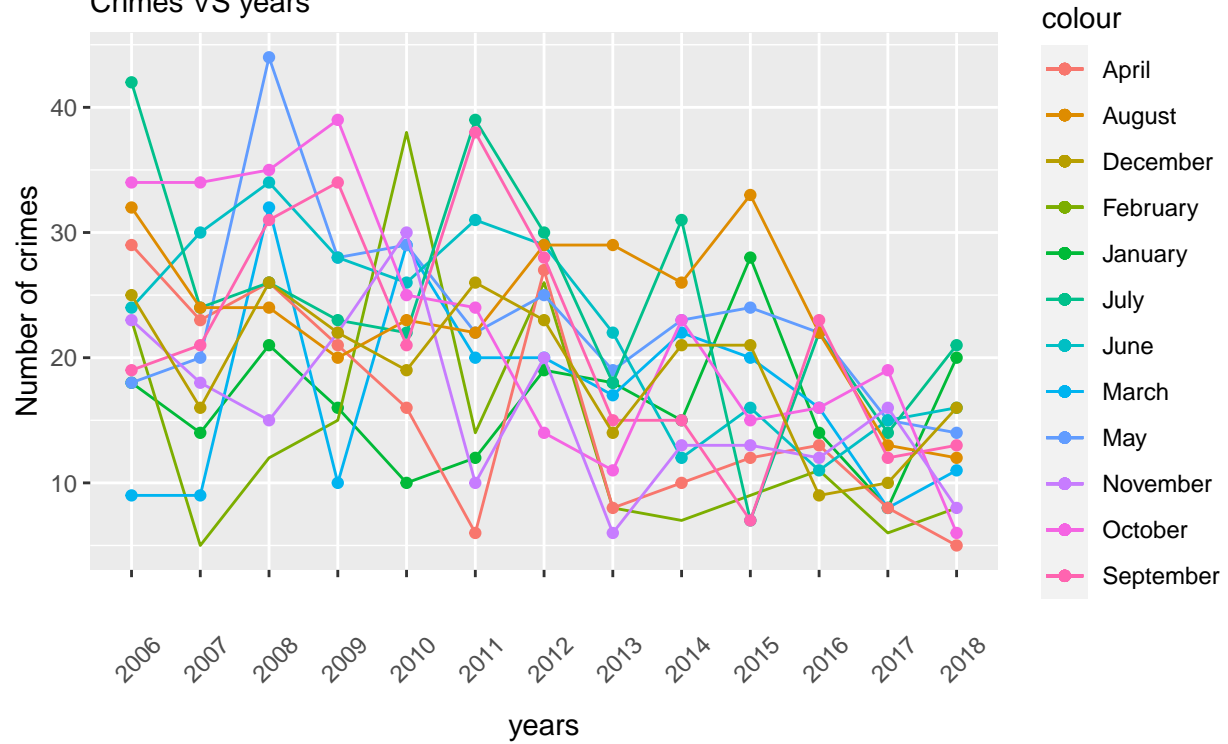


But what can be seen even from this basic visualization, is that most of the data is closely packed and trends in a generally downward direction. When the data is replotted excluding more recent years. An overall difference can be seen between summer months and winter months.



# Annual Changes in Crime in New York City in Queens

## Crimes VS years



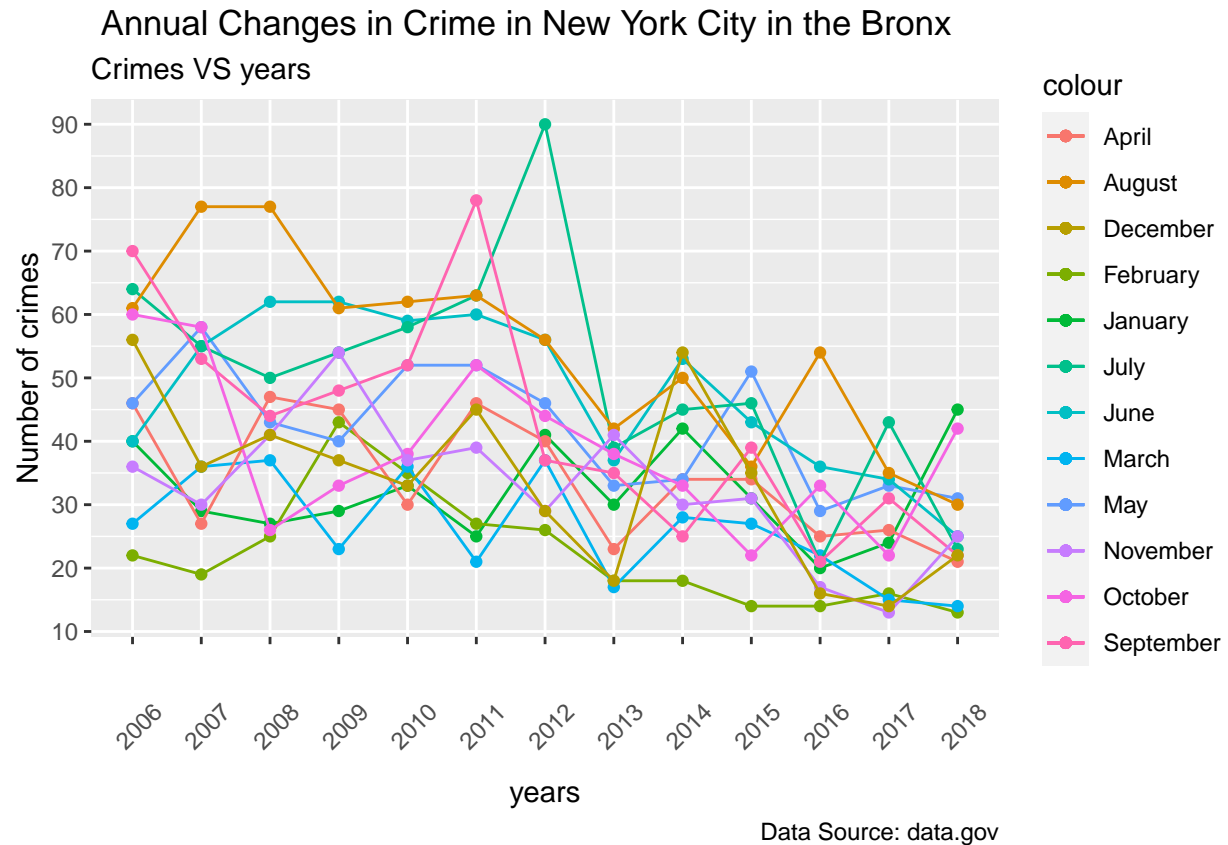
Data Source: data.gov

# Annual Changes in Crime in New York City in Manhattan

## Crimes VS years

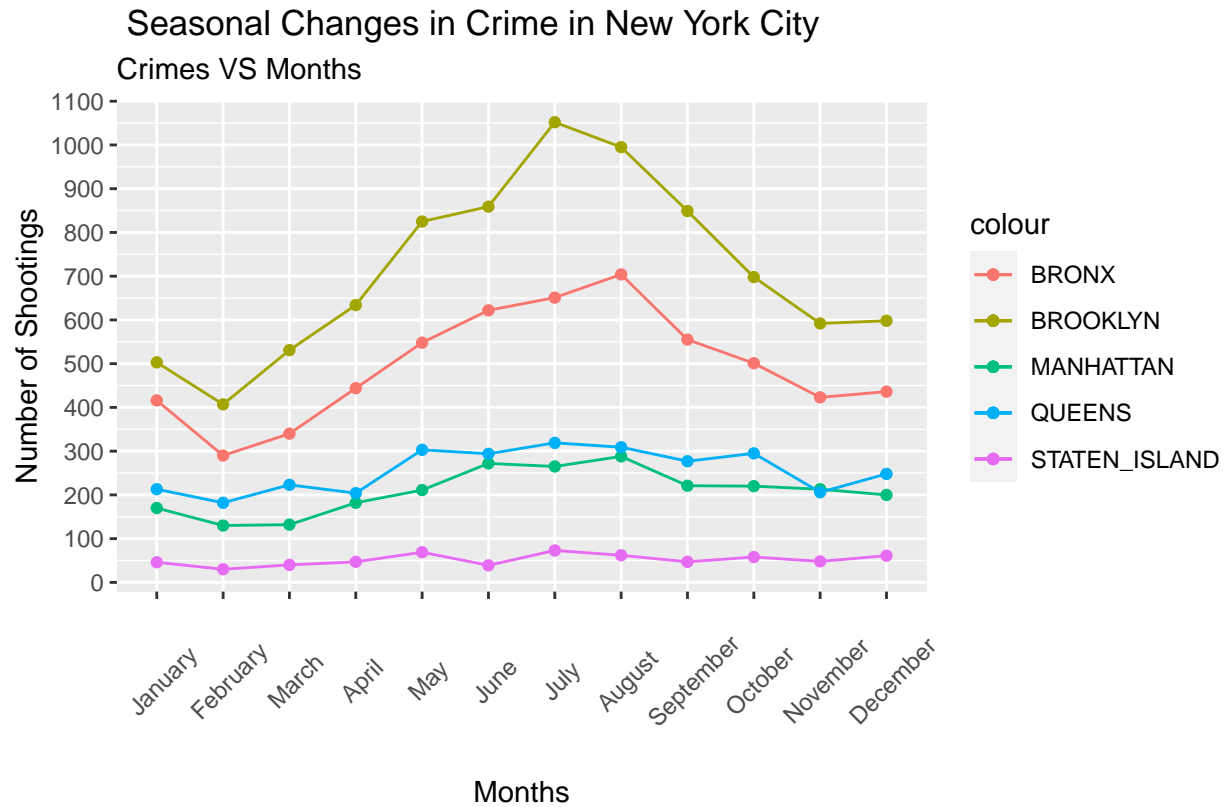


Data Source: data.gov



When the data is further adjusted for even more extraneous data and replotted to the same visualization from before it does not actually change very much. As you can see the plot simply moves down slightly. So could this data still be right and valid?





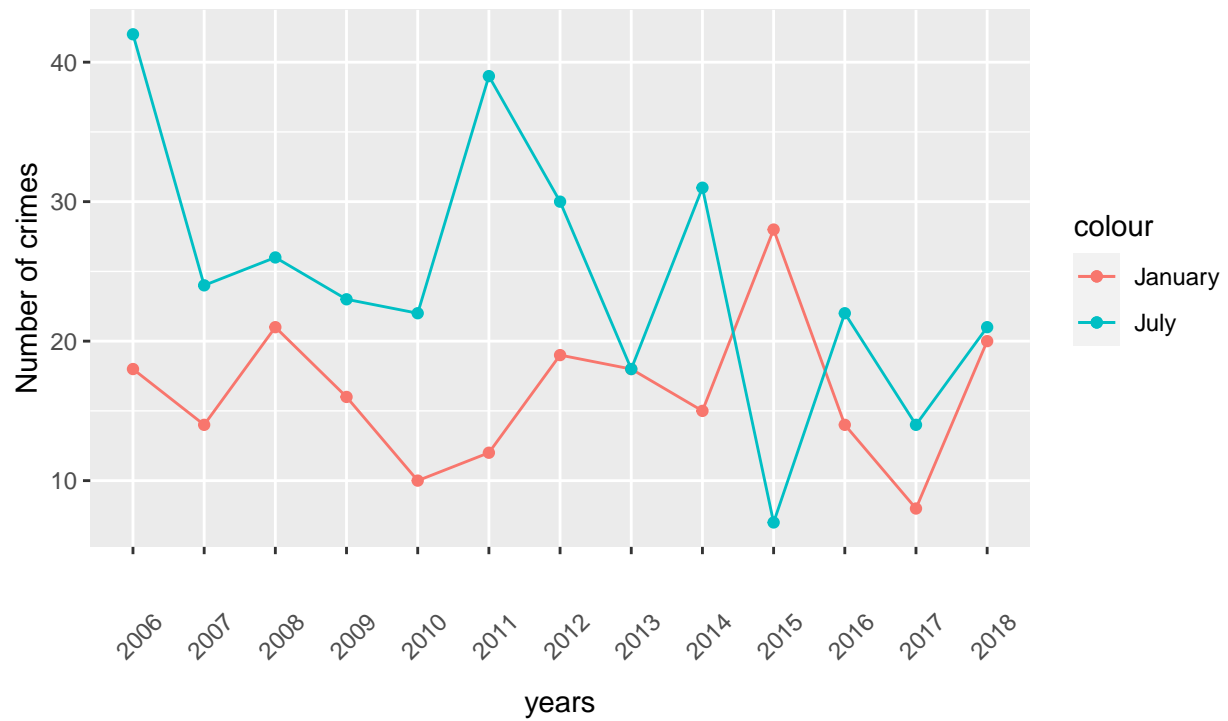
Data Source: data.gov

A good thing to do would be to look over some smaller sections. The months of January and July were chosen to be plotted to see if there were any obvious differences between the way a plot of all shootings vs a plot of just a small slice of time would look. When the data is even more subdivided so there is less to confuse the eye. The data seems more flatenned and there appears to be fewer differences.

While data for Brooklyn shows a generally higher rate of shootings during summer, the other Boroughs show mixed results. Some places have higher shooting counts in July and others have lower numbers. This kind of confounds and muddles the conclusions to be drawn from this analysis.

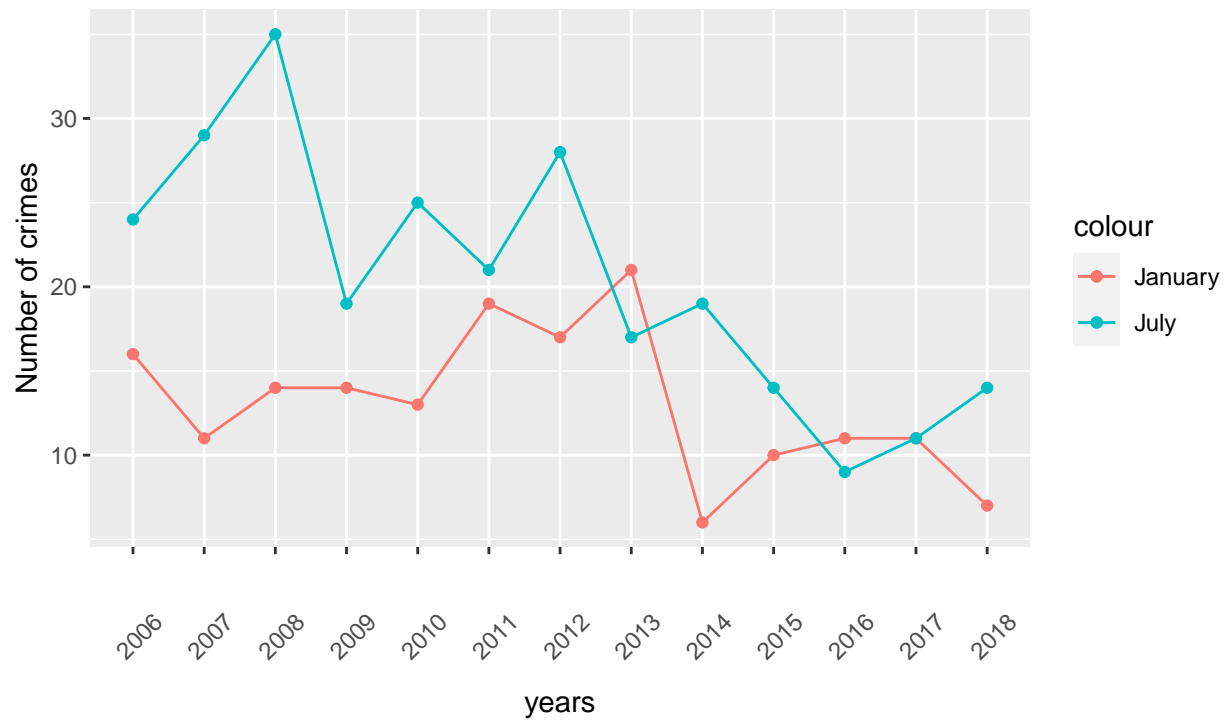


Annual Changes in Crime in New York City in Queens  
Crimes VS years



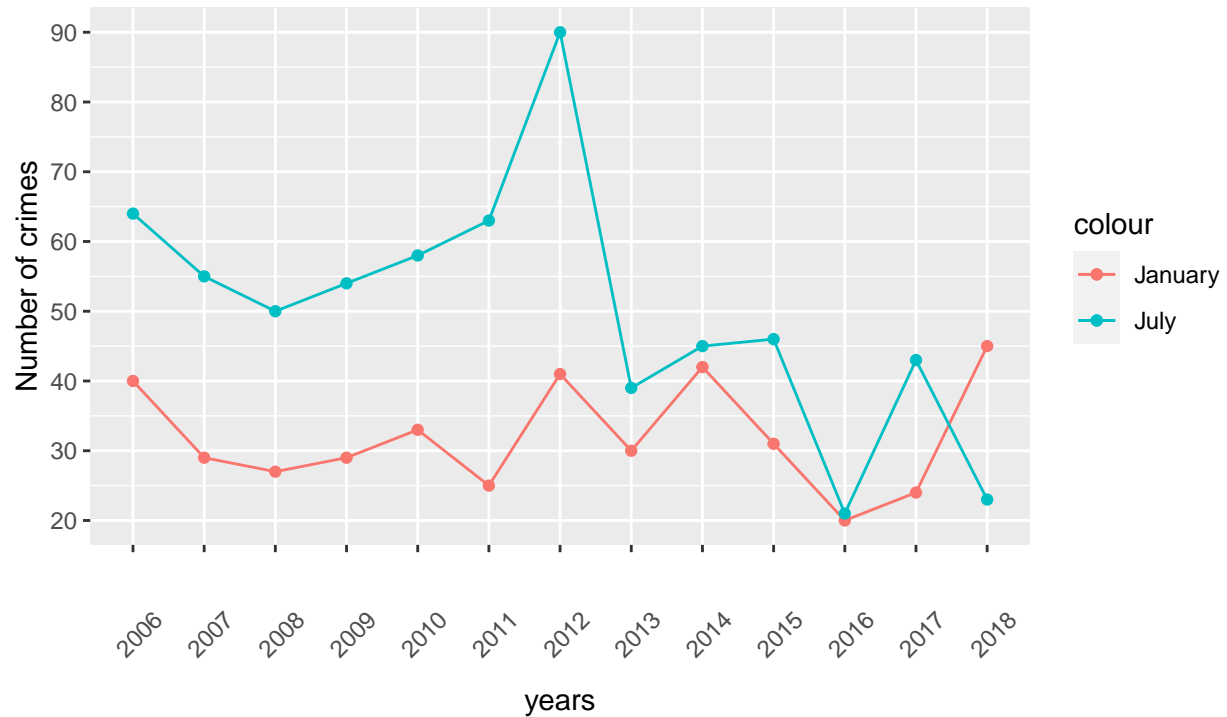
Data Source: data.gov

Annual Changes in Crime in New York City in Manhattan  
Crimes VS years



Data Source: data.gov

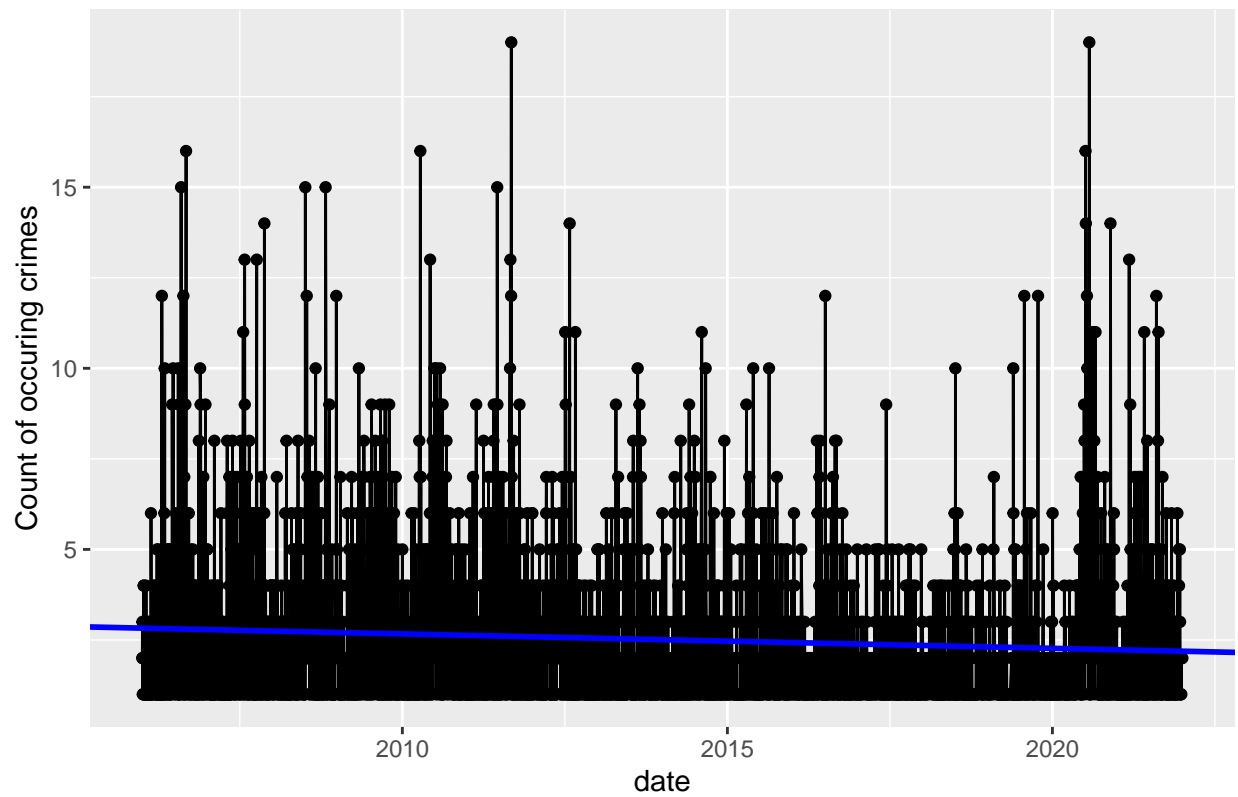
## Annual Changes in Crime in New York City in the Bronx Crimes VS years



Data Source: data.gov

This confusion can be further supported with the use of simple linear models. When monthly crime rates are trended and linear modeling is applied. Specifically trending the frequency of appearances of specific dates. The resulting visualization is very flat. With almost no change between different seasons, or years. And this is common to all Boroughs that were included in the study. All had close Y intercepts, and slopes showed to be very flat. Given all of this data and the results obtained from said data. The null hypothesis is rejected. Overall the summer and winter months show the same rates of shootings over time. This also highlights an important point when looking over data sets with a large variety of data. Should data like this be further subdivided into smaller groups and looked at in individual pieces? Or since the data shows an overall rejection of the null hypothesis should the results be taken at face value and not explored further?

Shooting frequencies in Brooklyn



```
## [1] "Linear model output for Brooklyn"
```

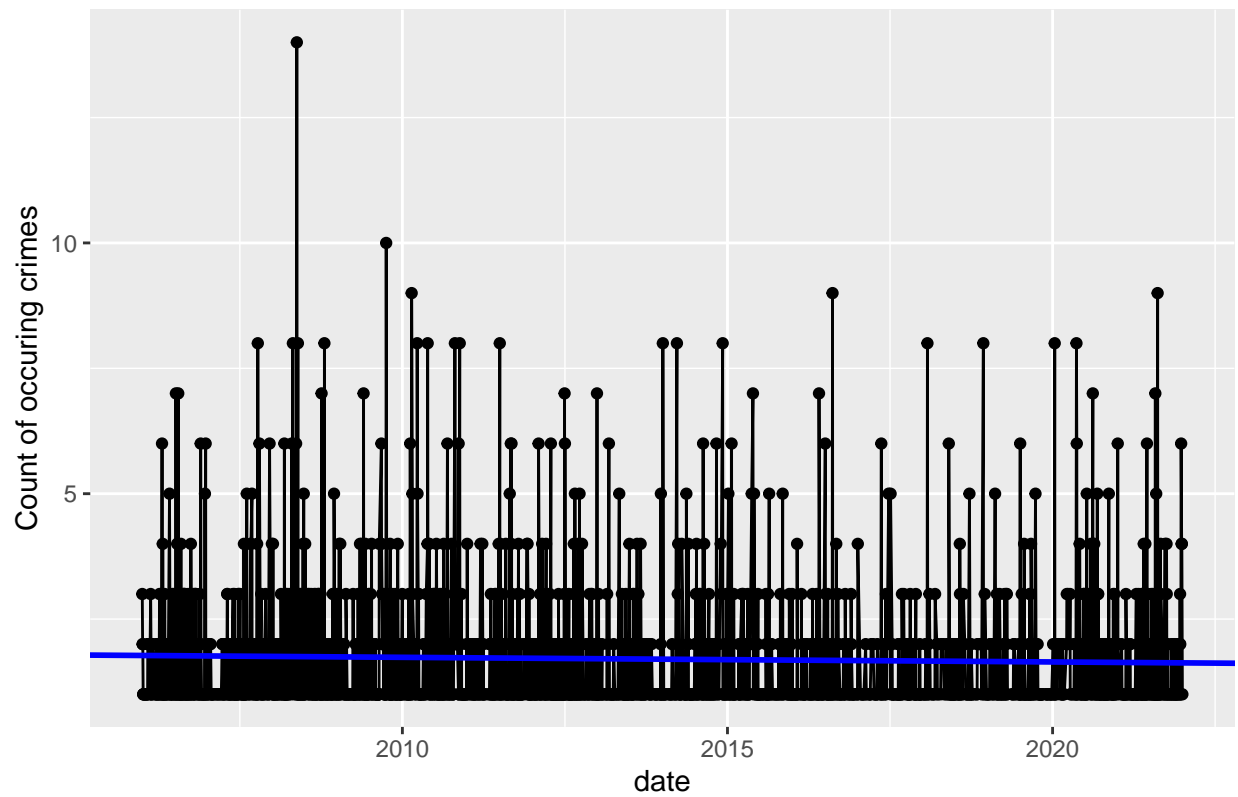
```
## (Intercept)
```

```
## 4.260597
```

```
## OCCUR_DATES
```

```
## -0.0001090344
```

Shooting frequencies in Queens



```
## [1] "Linear model output for Queens"
```

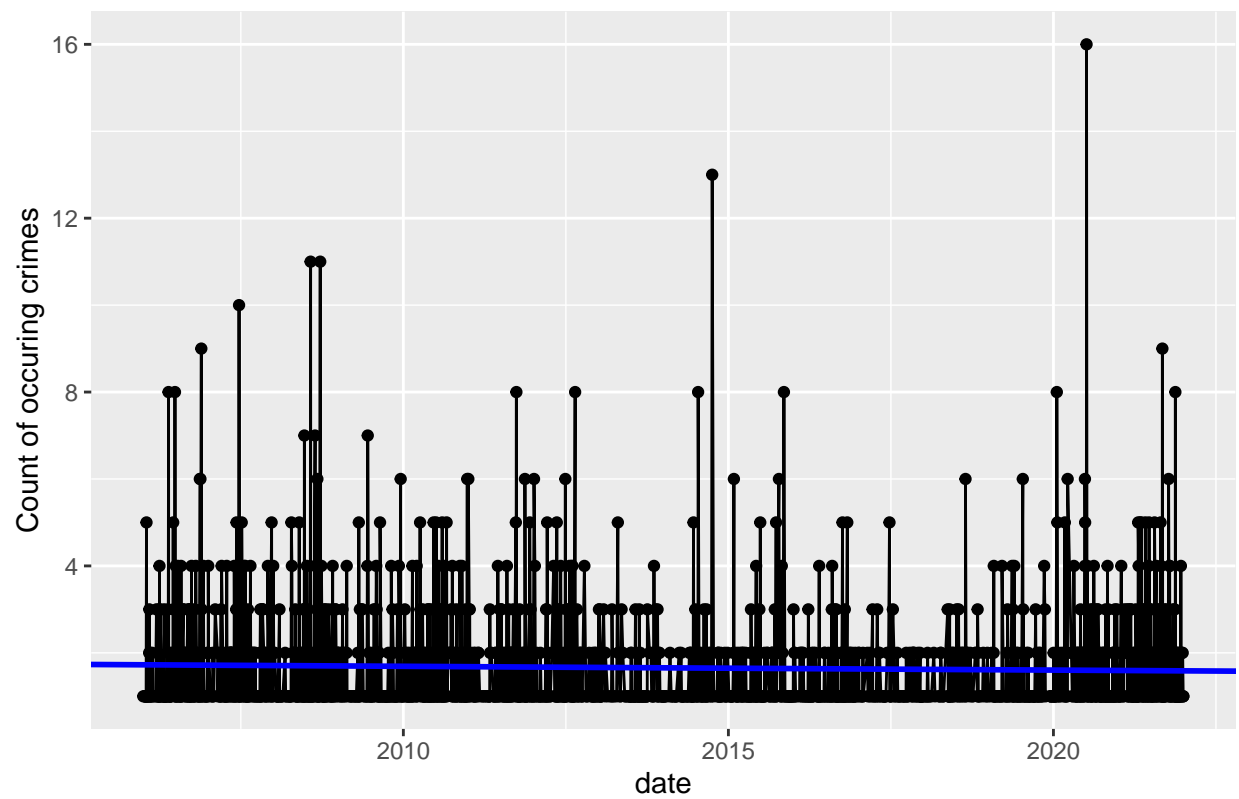
```
## (Intercept)
```

```
## 2.106139
```

```
## OCCUR_DATES
```

```
## -2.524722e-05
```

## Shooting frequencies in Manhattan



```
## [1] "Linear model output for Manhattan"
```

```
## (Intercept)
```

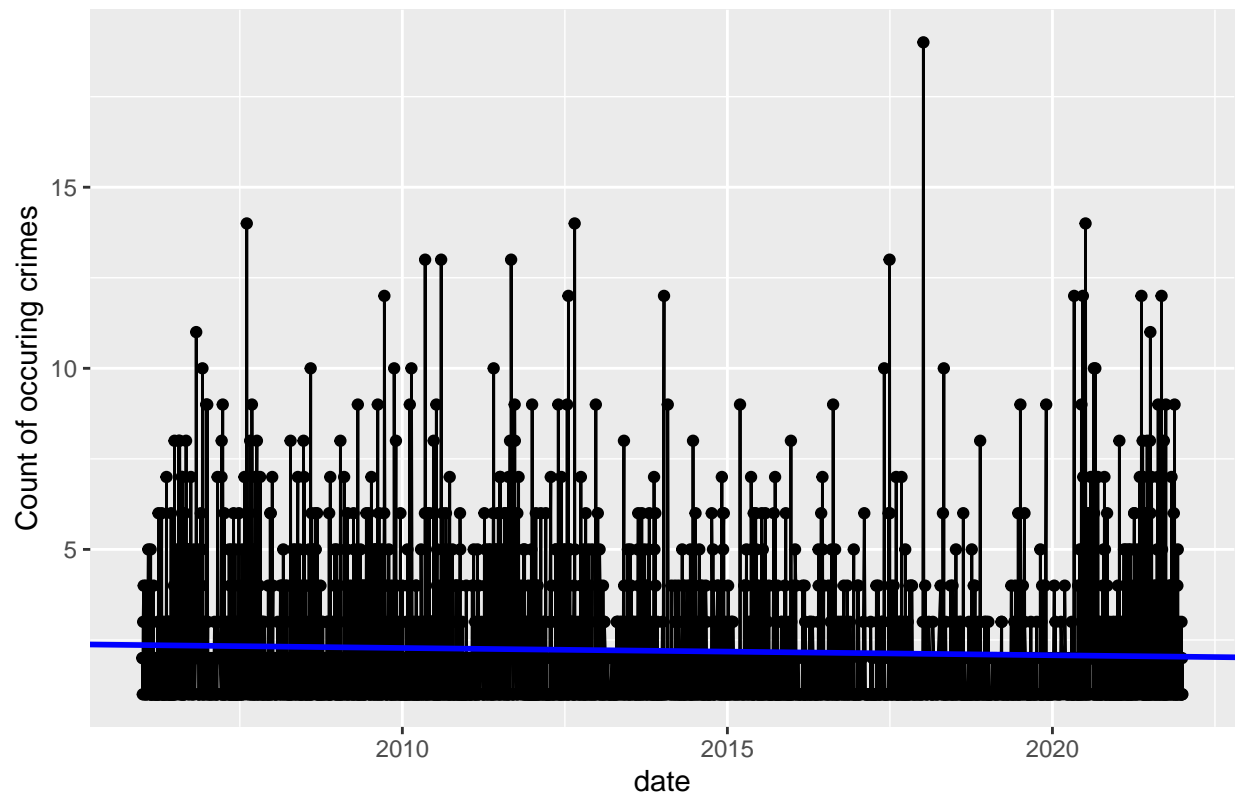
```
## 2.038181
```

```
## OCCUR_DATES
```

```
## -2.382857e-05
```



## Shooting frequencies in the Bronx



```
## [1] "Linear model output for the Bronx"
```

```
## (Intercept)
```

```
## 3.081127
```

```
## OCCUR_DATES
```

```
## -5.492149e-05
```