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

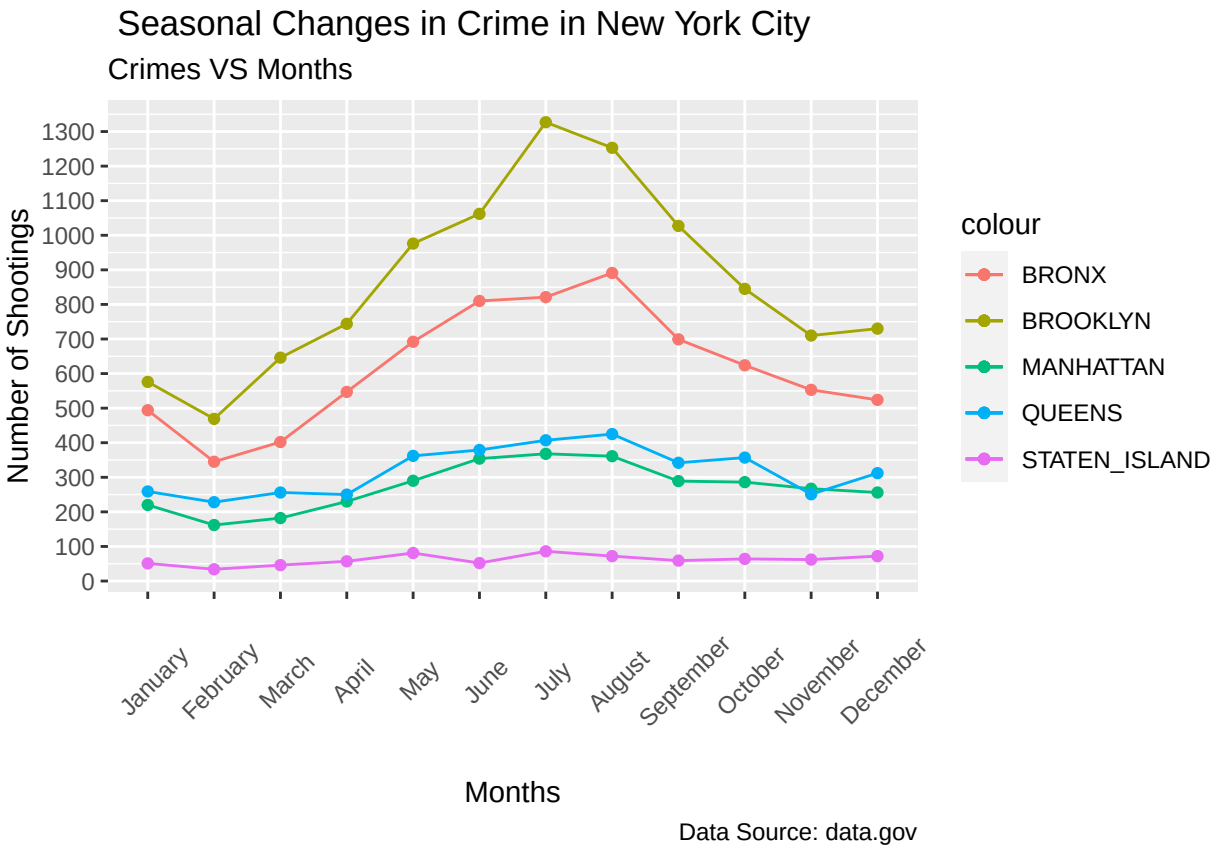
New York City is a massive area consisting of several Boroughs. It is also the county seat of New York County. With a 2023 population of 7,888,121, it is the largest city in New York and the largest city in the United States.(*) This makes it a great candidate for obtaining large amounts of data. This 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*

$$H_0 = \text{There is an increase in the number of shootings seasonally.}$$

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

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.



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 being trended over many years. When columns of data are parsed or cleaned like this, if there is an outlier it may and will often skew results and go unnoticed. Additionally any year over year changes can often be missed.

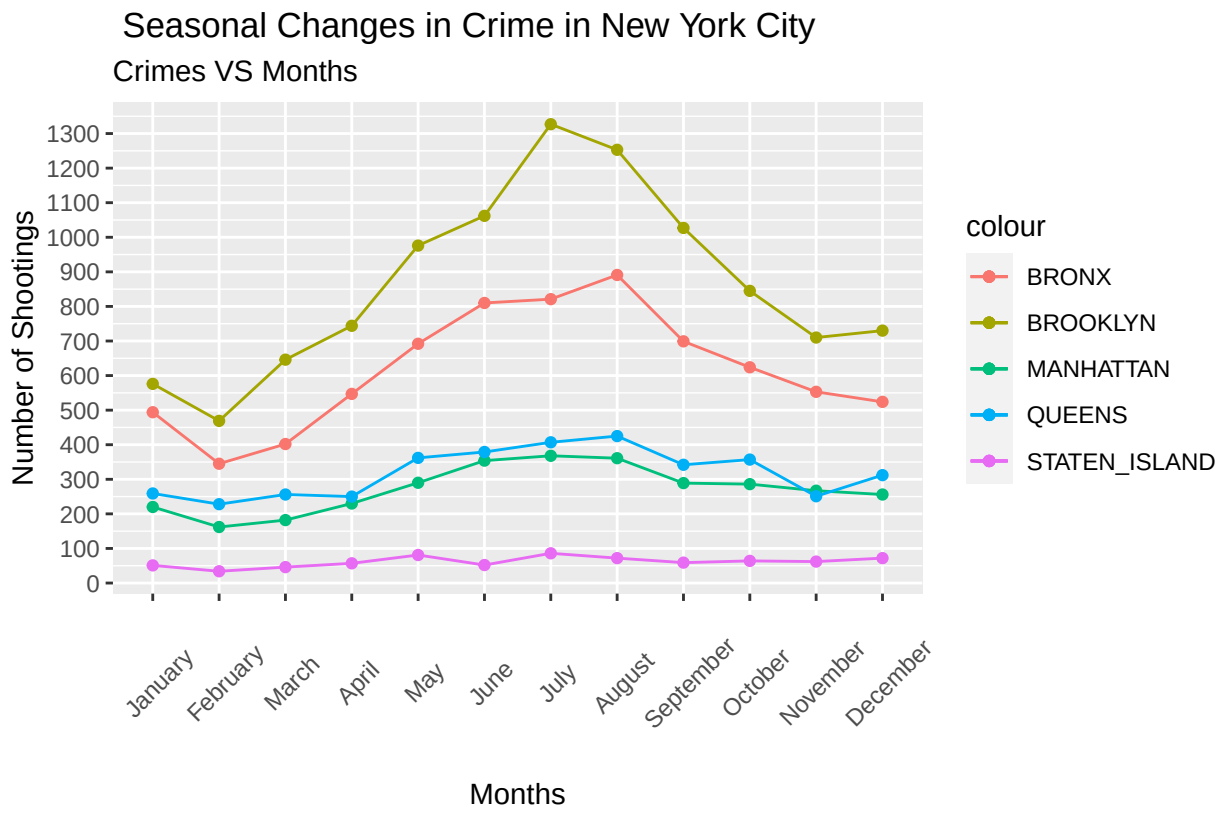
When the data was divided further and inspected more closely. Some visual differences became quickly apparent. For example when the data for the Borough of Brooklyn was further explored.(fig2)Some variables that would skew results were quickly seen.The years 2019,2020,2021 all had very large spikes in numbers of shootings. So those years at least for the purposes of this study 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 re plotted excluding more recent years. An difference between summer months and winter months is not readily apparent if there is one at all.(fig3,fig4,fig5,fig6) The previous visualization(fig1) was also re-plotted (fig7) this time, using data that is a slightly better fit. And excluding the data from very recent years. The overall visual effect seemed to be the same. With a large jump in shootings during the summer months for most boroughs. But this data was a simple summation, and not reliable for trending.

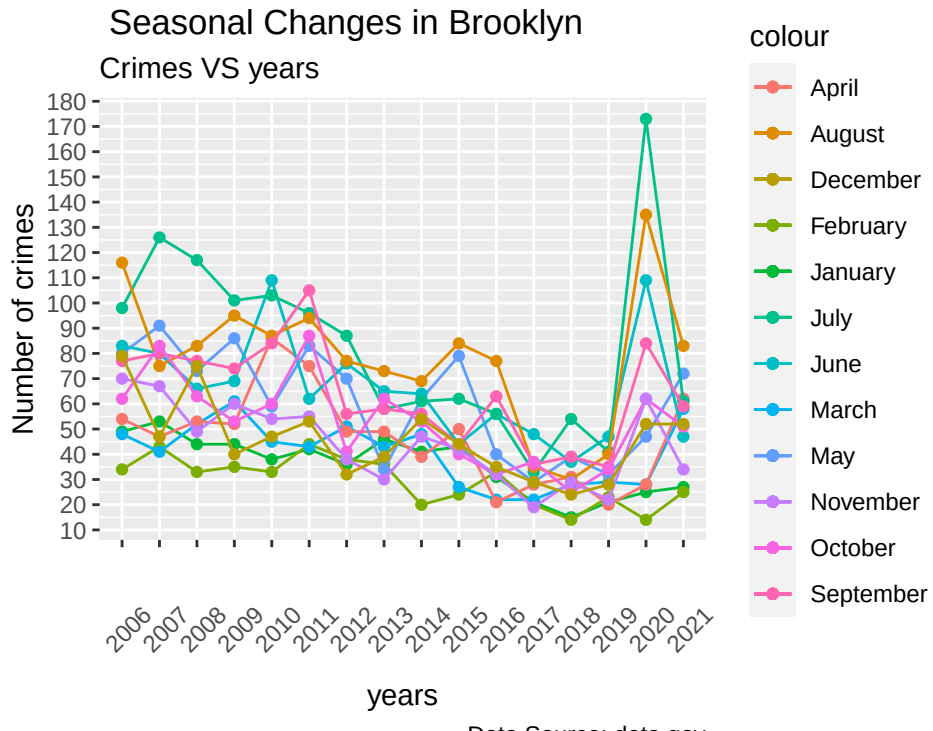
The months of January and July were chosen to be plotted to see if there were any obvious differences between the way a plot of all shootings vs a plot of just a small slice of time would look.(fig8,fig9,fig10,fig11) When the data was even more subdivided so there is less to confuse the eye. The data seems more flatenned and there appeared to be fewer differences. While data for Brooklyn shows a generally higher rate of shootings during summer, the other Boroughs show mixed results. Some locations 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. (fig8,fig9,fig10,fig11)

This confusion can be quickly refuted with the use of simple linear models.(fig12,fig13,fig14,fig15) When monthly crime rates are trended and linear modeling is applied. Specifically trending the frequency of appearances of specific dates, so two shootings in one day, would mean two counts of said date, etc.The resulting visualizations were 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?

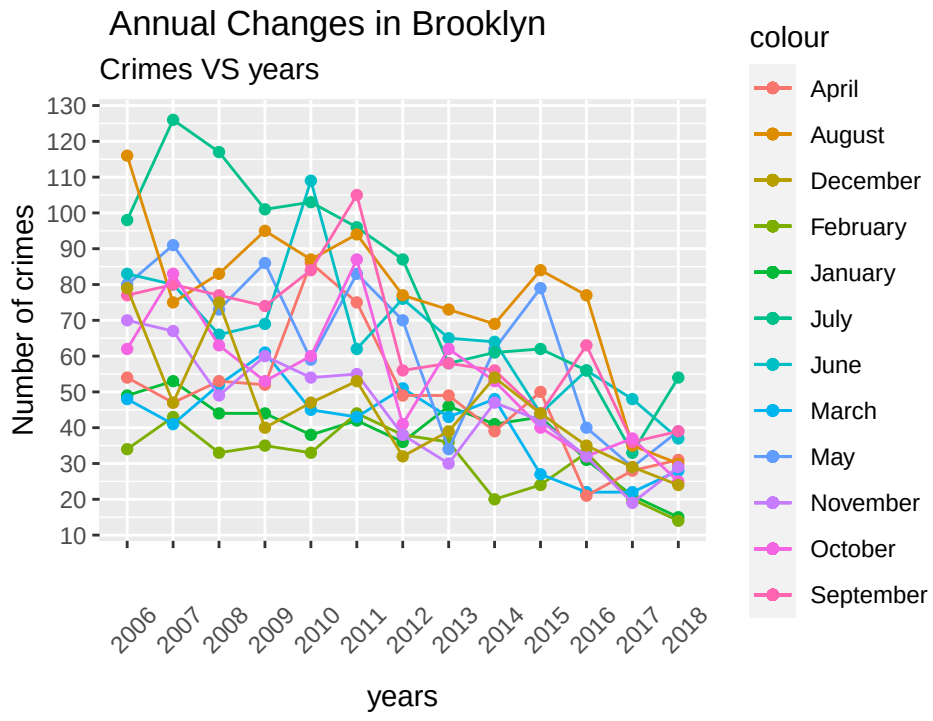
```
## [1] "Figure 1"
```


Seasonal Changes in Crime in New York City
Crimes VS Months
Number of Shootings
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
January
February
March
April
May
June
July
August
September
October
November
December
Months
colour
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN_ISLAND
Data Source: data.gov

```
## [1] "figure2"
```

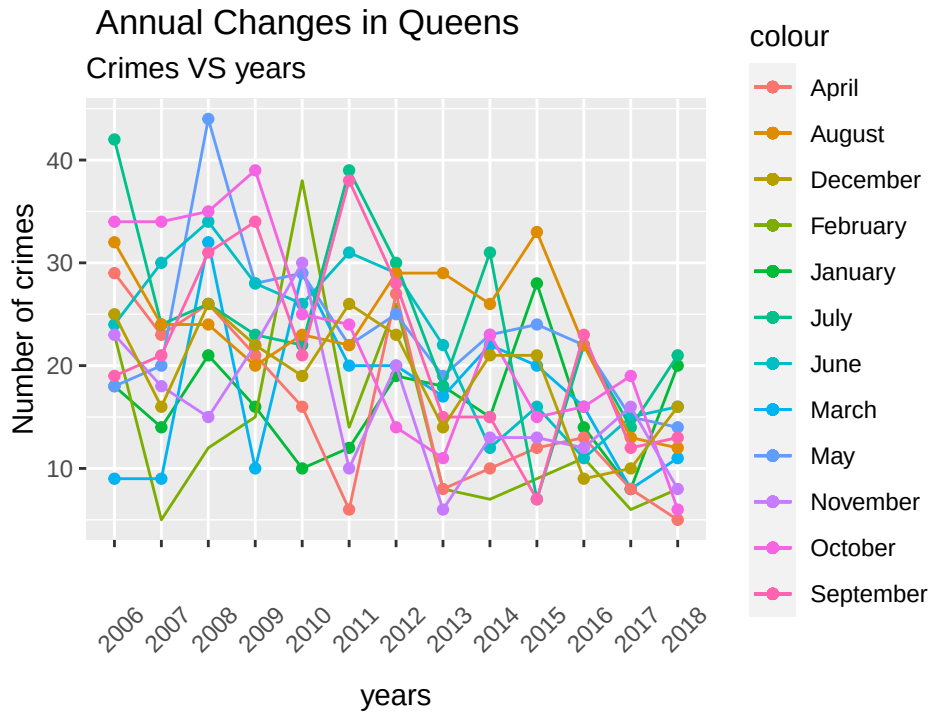

Seasonal Changes in Brooklyn
Crimes VS years
Number of crimes
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
years
colour
April
August
December
February
January
July
June
March
May
November
October
September
Data Source: data.gov

```
## [1] "Figure3"
```



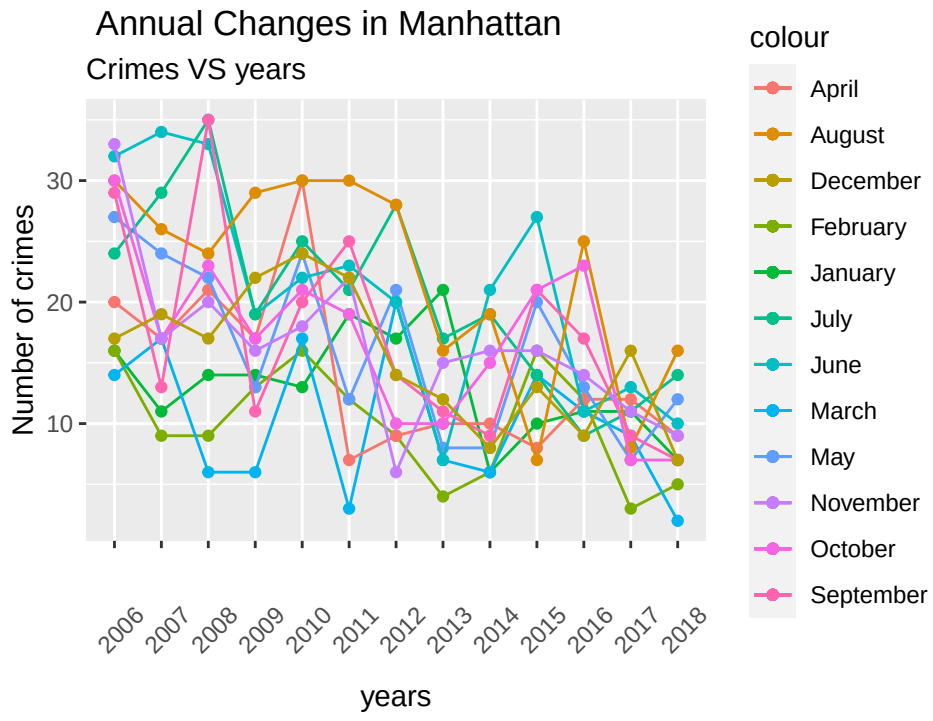
Data Source: data.gov

```
## [1] "Figure 4"
```

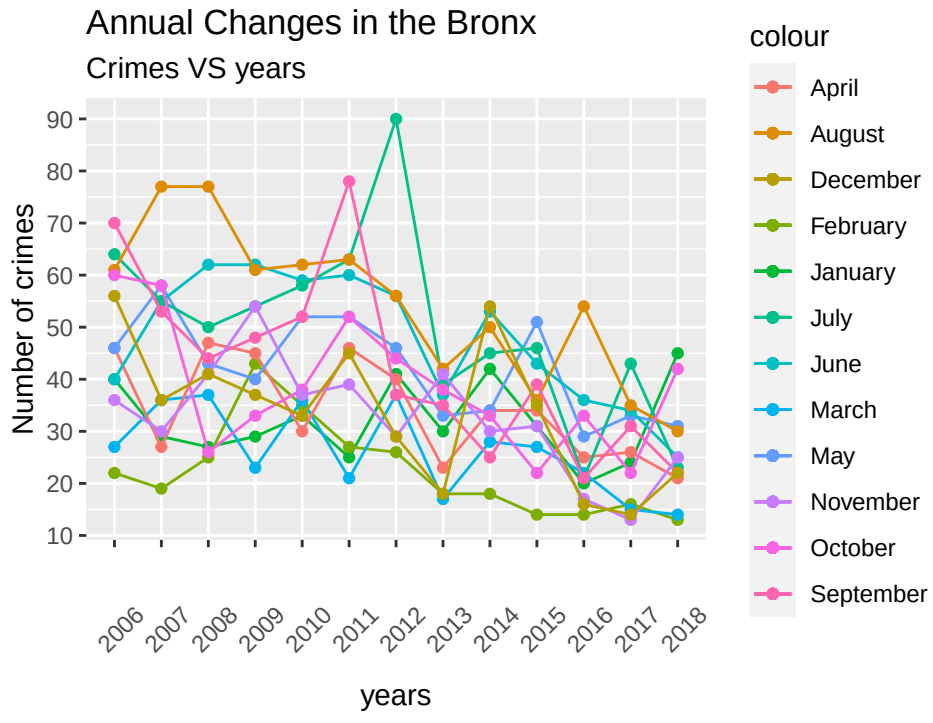


Data Source: data.gov

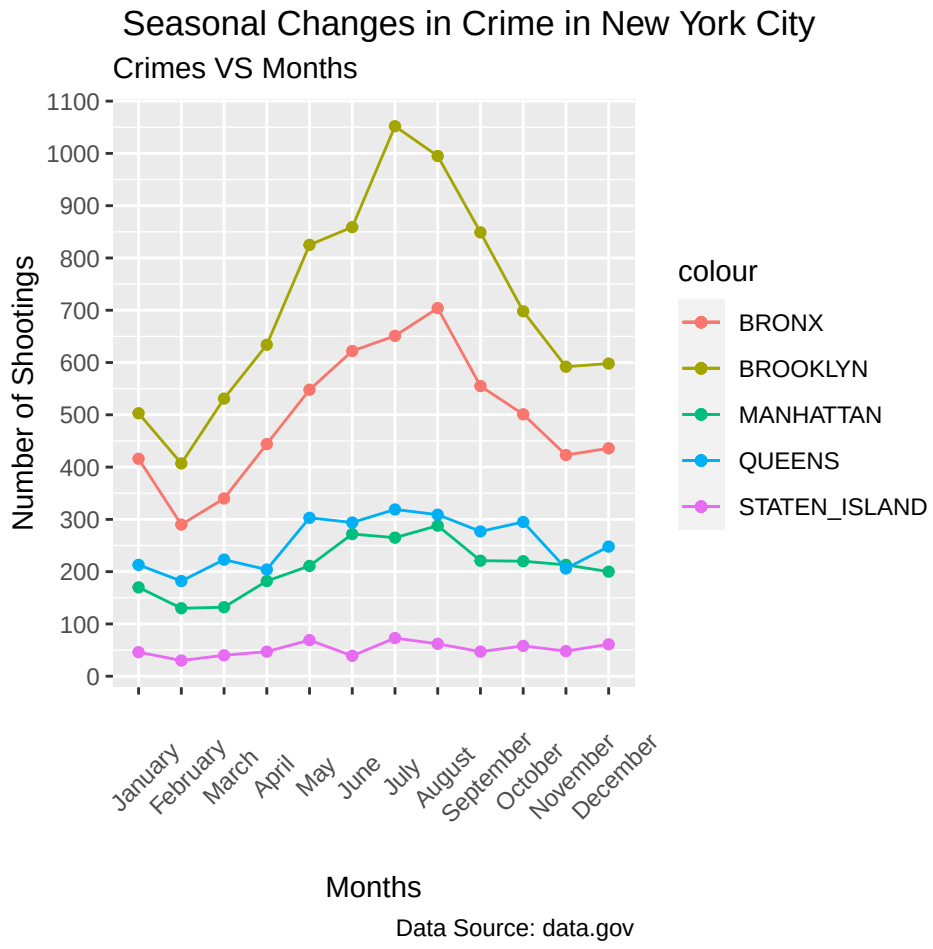
```
## [1] "Figure 5"
```



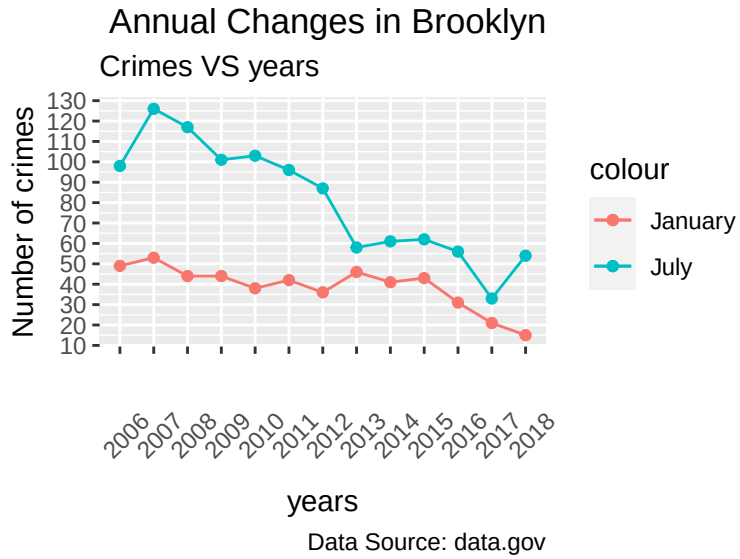

```
## [1] "Figure 6"
```



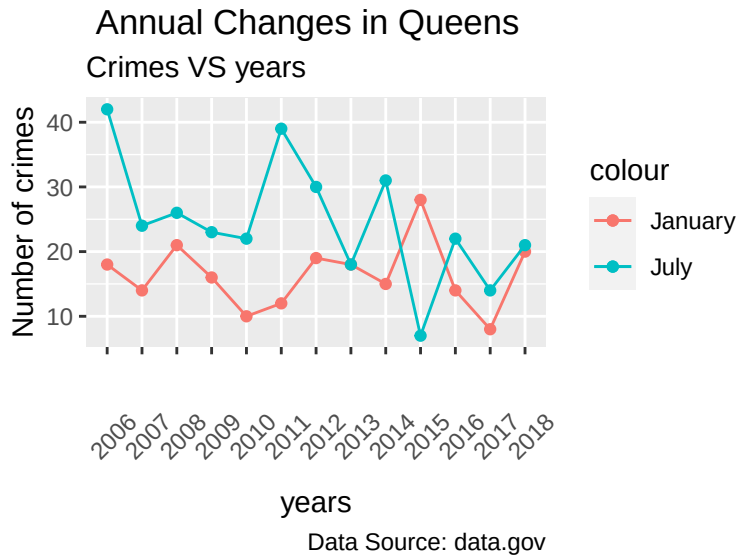
```
## [1] "Figure 7"
```


Seasonal Changes in Crime in New York City
Crimes VS Months
Number of Shootings
1100
1000
900
800
700
600
500
400
300
200
100
0
January
February
March
April
May
June
July
August
September
October
November
December
Months
Data Source: data.gov
colour
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN_ISLAND

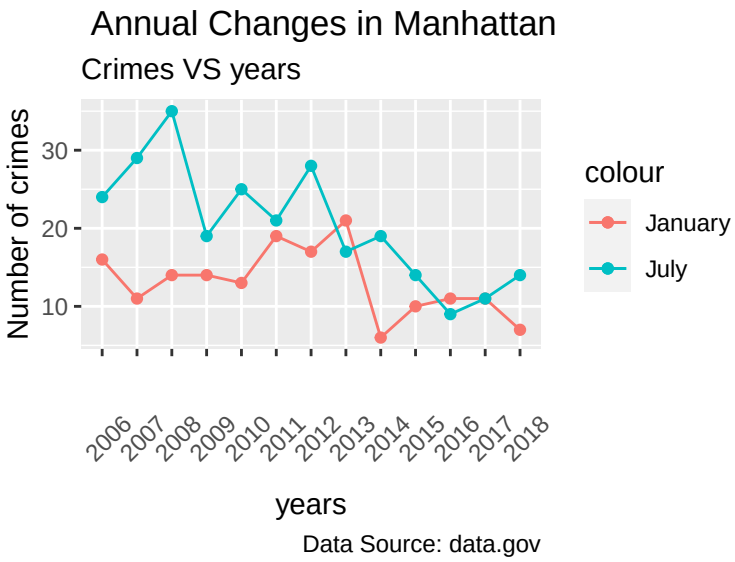
```
## [1] "Figure 8"
```



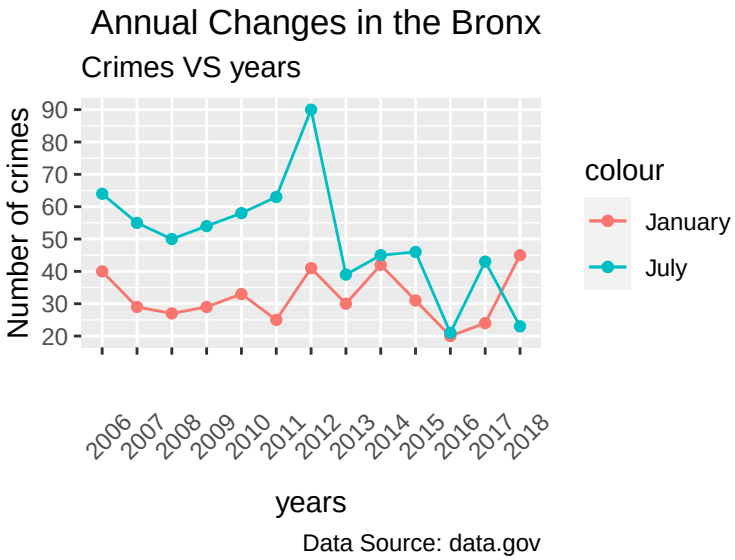

```
## [1] "Figure 9"
```



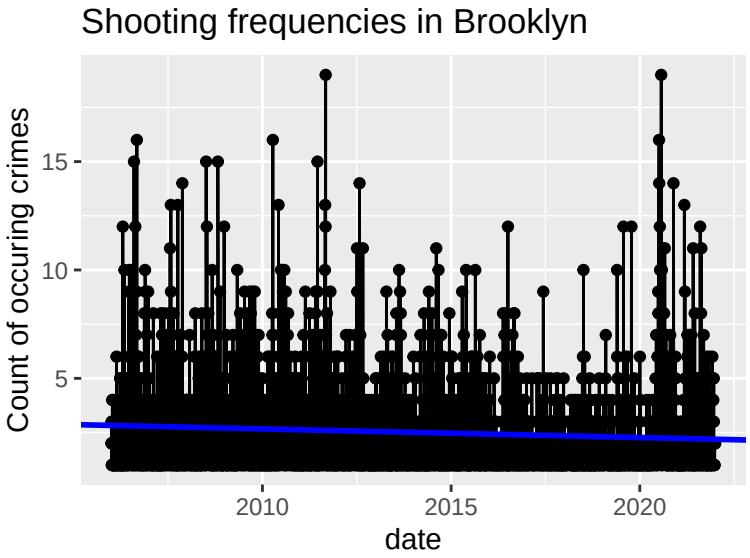

```
## [1] "Figure 10"
```


Annual Changes in Manhattan
Crimes VS years
Number of crimes
30
20
10
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
years
Data Source: data.gov
colour
January
July

```
## [1] "Figure 11"
```


Annual Changes in the Bronx
Crimes VS years
Number of crimes
90
80
70
60
50
40
30
20
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
years
Data Source: data.gov
colour
January
July

```
## [1] "Figure 12"
```

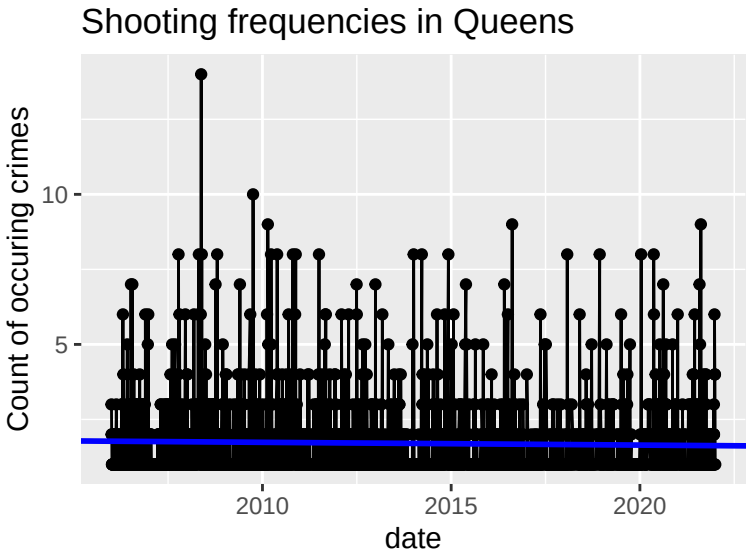



```
## [1] "Linear model output for Brooklyn"

## (Intercept)
##    4.260597

##   OCCUR_DATES
## -0.0001090344

## [1] "Figure 13"
```

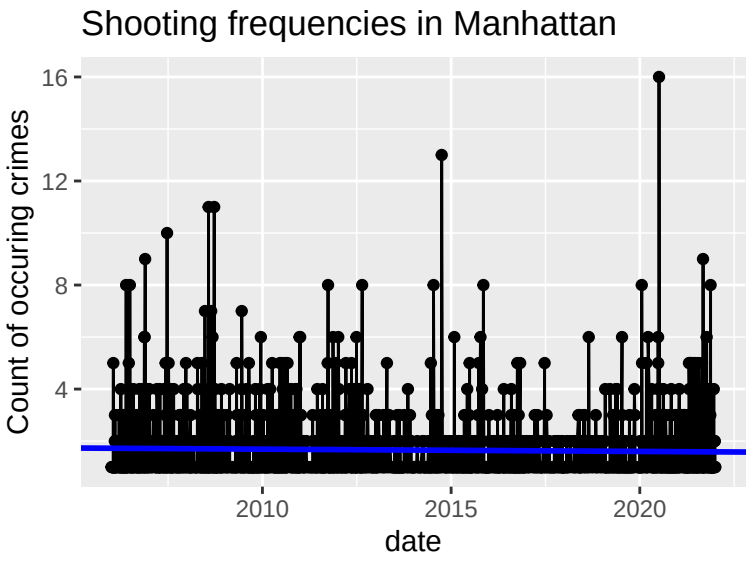



```
## [1] "Linear model output for Queens"
```

```
## (Intercept)
##    2.106139

##   OCCUR_DATES
## -2.524722e-05

## [1] "Figure 14"
```

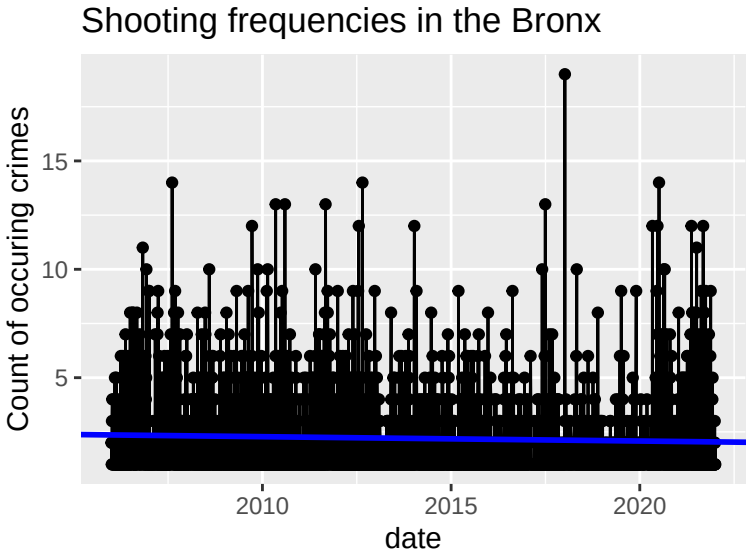



```
## [1] "Linear model output for Manhattan"

## (Intercept)
##    2.038181

##   OCCUR_DATES
## -2.382857e-05

## [1] "Figure 15"
```



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

## (Intercept)
##    3.081127

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