NYPD Shooting Incident

Deccription

Shooting incident data in New York Since 2006.

Source

https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Add Library

and library which needed

```
library(tidyverse)
```

Import Data

Import data from website

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

Read Data

Read data from csv

```
data <- read.csv(url)
```

Clean Up Data

Select only focus columns and formatting date

```
data <- data %>%
  select(OCCUR_DATE, BORO, STATISTICAL_MURDER_FLAG) %>%
  mutate(date = as.Date(OCCUR_DATE, "%m/%d/%Y")) %>%
  select(-c(OCCUR_DATE))
```

Show summary to check if there is missing data.

```
summary(data)
```

```
##       BORO           STATISTICAL_MURDER_FLAG      date
##  Length:23568       Length:23568            Min.   :2006-01-01
##  Class :character   Class :character        1st Qu.:2008-12-30
##  Mode  :character   Mode  :character        Median :2012-02-26
##                                             Mean   :2012-10-03
##                                             3rd Qu.:2016-02-28
##                                             Max.   :2020-12-31
```

If there is missing BORO data label as unknown

```
data <- data %>% mutate(BORO = ifelse(BORO != "", BORO, "unknown"))
```

Group Data By Area

Group data by date and area. Then, count number of case and murder case

```
area <- data %>% mutate(cases = 1,
                        murder_cases = ifelse(STATISTICAL_MURDER_FLAG == "true", 1, 0)) %>%
  group_by(date, BORO) %>%
  summarize(cases = sum(cases),
            murder_cases = sum(murder_cases))
```

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

Transform Data

prepare data for visualization

Total NY Case

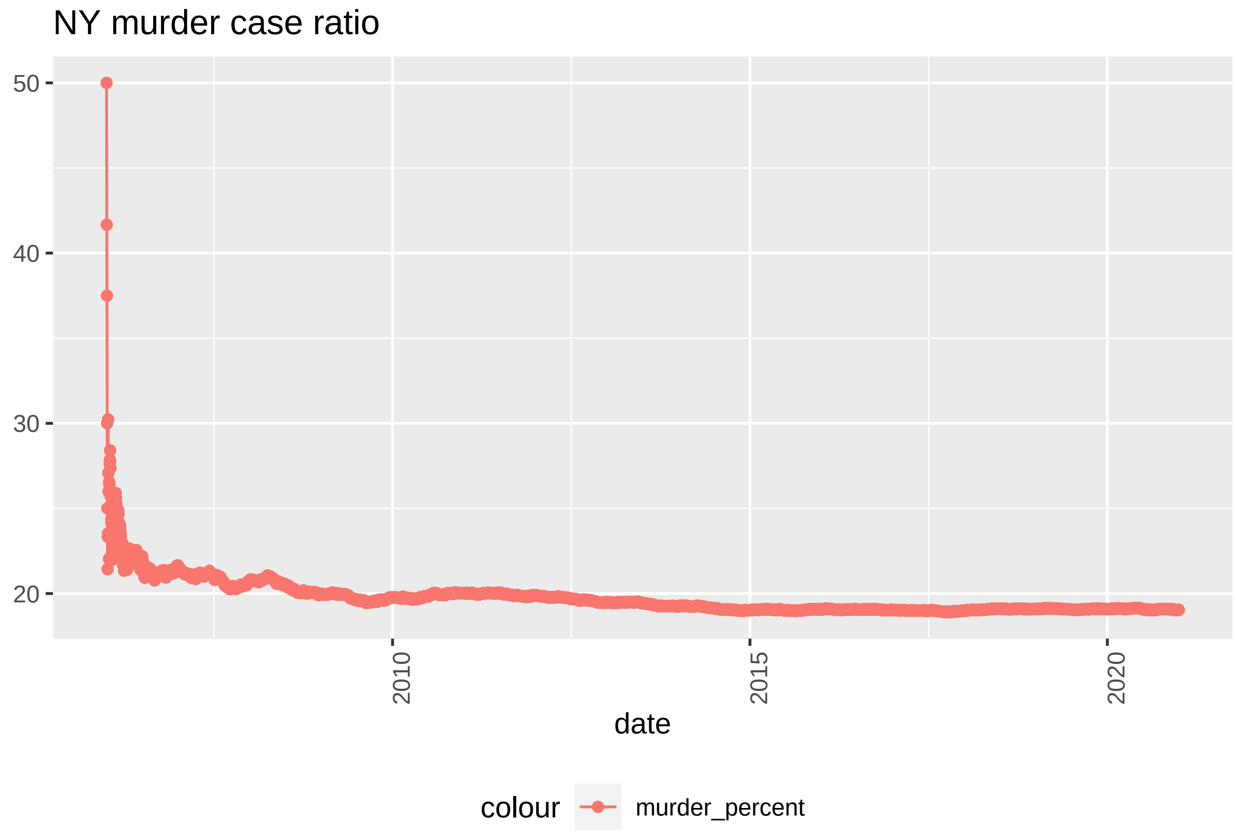
```
NY_total <- area %>% group_by(date) %>%
  summarize(cases = sum(cases),
            murder_cases = sum(murder_cases)) %>%
  mutate(cases = cumsum(cases),
         murder_cases = cumsum(murder_cases)) %>%
  mutate(murder_percent = murder_cases/cases*100)
```

Queens Case

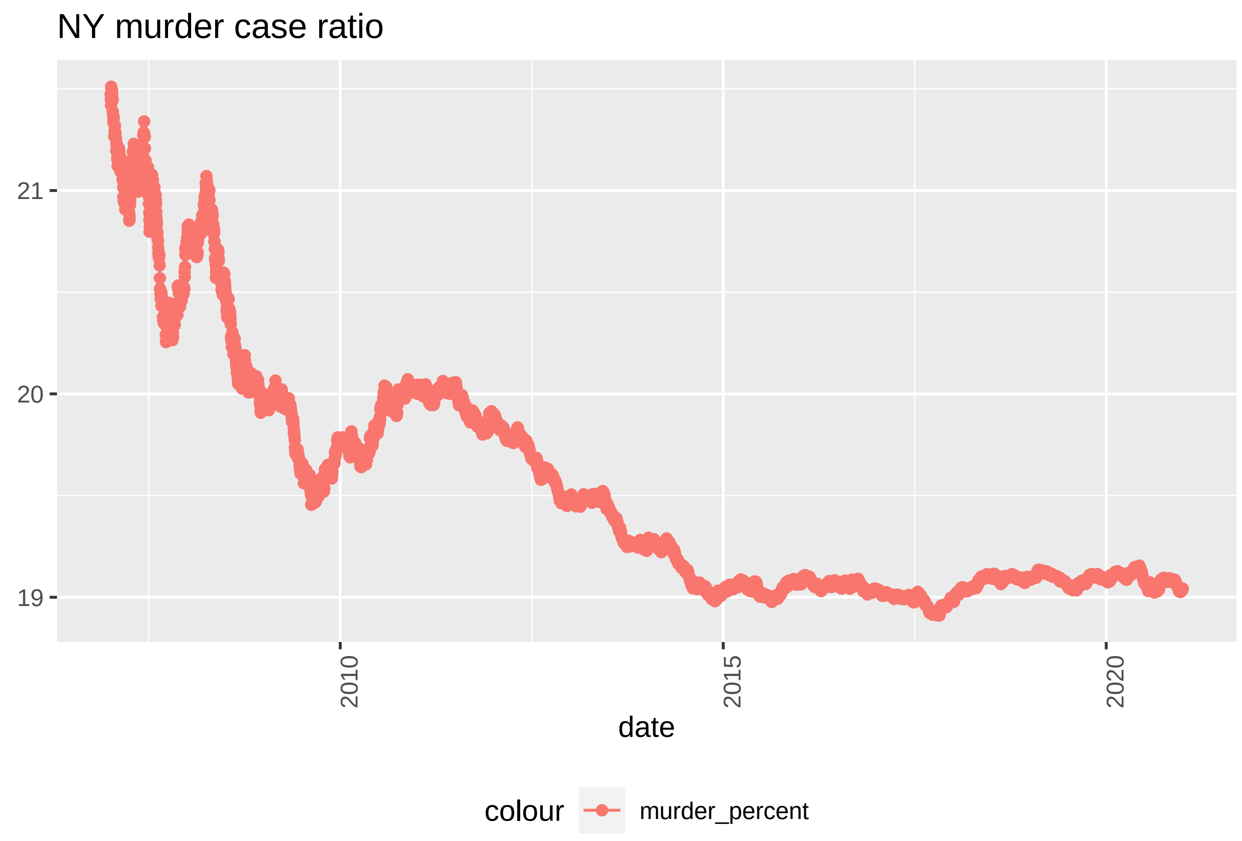
```
queens_case <- area %>% filter(BORO == 'QUEENS') %>%
  select(-c(BORO)) %>%
  ungroup %>%
  mutate(cases = cumsum(cases),
         murder_cases = cumsum(murder_cases)) %>%
  mutate(murder_percent = murder_cases/cases*100)
```

visualizations

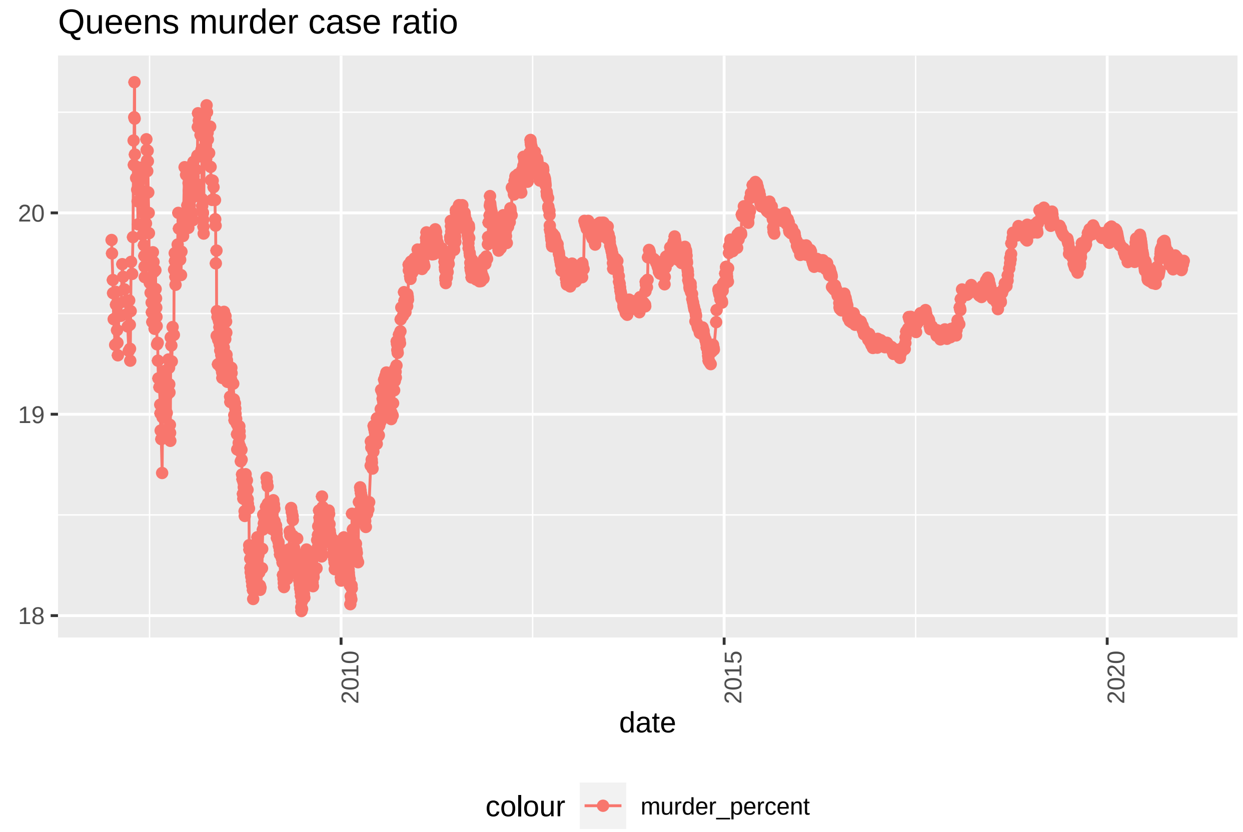
New York cases



New York cases since 2007



cases in Queens since 2007



Modeling

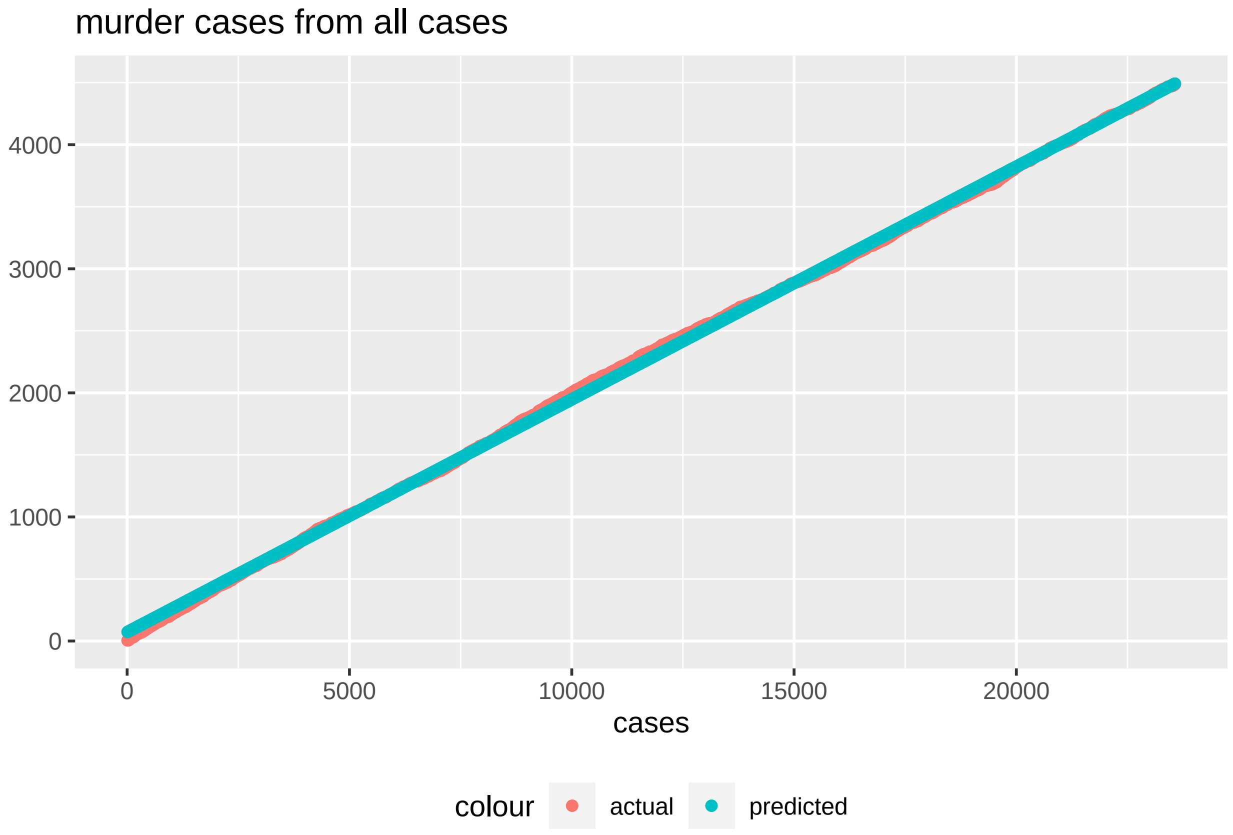
Create model for predict murder cases by cases

```r
mod <- lm(murder_cases ~ cases, data = NY_total)
```

Predict murder cases with model

```r
NY_total <- NY_total %>% mutate(predict = predict(mod))
```

Plot predict result



Analysis

As you can see from graph 2 (NY cases since 2007) murder rate decrease until 2015. Since 2015 to 2020 murder rate is stable around 19%. But, if you look at graph 3 (Queens cases since 2007) rate is decrease from 2007 then stable at about 18.3% for a while. Until 2010 murder rate is increase to 19-20%.

From this information, I thinks there was something happened in Queens around 2010 which cause murder rate in Queens not keep decreasing until 2015 as same as overall New York City.

conclusion

New York murder rate overall decrease compare to 2007. But in Queens is difference.

Bias

Bias of this analysis might be how I calculate murder percentage. Because in very early date there are only few case to calculate compare to latest date. As you can see there is 50% murder rate in graph 1 (NY cases since 2006). So, I prevent this by using data since 2007 which I assumes 2007 have enough cases.

Also, there's personal bias That race dose not effect murder rate at all. I just looked back what is impression with the data and I found that I remove columns about race out because my bias. actually it might be a variable which effect murder rate.