NYPD Shooting Report

I have included all my commands that lead to everything as code blocks. I showed only the output of the most integral ones. I hope you enjoy :D

Beginnings

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.7     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(RcppRoll)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

The URL our data is located in

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

Loading our data into R

```
nyc_data <- read_csv(url_in)
```

```
## Rows: 25596 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Changing the date column to be an actual datetime object

```
nyc_data <- nyc_data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Remove unnecessary columns that don't help our analysis

```
nyc_data <- nyc_data %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude, Longitude,Lon_Lat))

nyc_data
```

INCIDENT_KEY <dbl>	OCCUR_DATE <dbl>	OCCUR_TIME <time>	BORO <chr>	PRECINCT <dbl>	JURISDICTION_CODE <dbl>
236168668	2021-11-11	15:04:00	BROOKLYN	79	0
231008085	2021-07-16	22:05:00	BROOKLYN	72	0
230717903	2021-07-11	01:09:00	BROOKLYN	79	0
237712309	2021-12-11	13:42:00	BROOKLYN	81	0
224465521	2021-02-16	20:00:00	QUEENS	113	0
228252164	2021-05-15	04:13:00	QUEENS	113	0
226950018	2021-04-14	21:08:00	BRONX	42	0
237710987	2021-12-10	19:30:00	BRONX	52	0
224701998	2021-02-22	00:18:00	MANHATTAN	34	0
225295736	2021-03-07	06:15:00	BROOKLYN	75	0

1-10 of 10,000 rows | 1-6 of 14 columns

Previous 1 2 3 4 5 6 ... 1000 Next

Quick Summary of the data

```
summary(nyc_data)
```

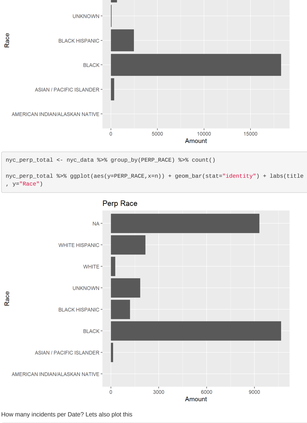
```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME                 BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596       Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms         Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime    Mode  :character
##  Mean   :123382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238499103   Max.   :2021-12-31
##
##    PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000   Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000   Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000   Mode  :character   TRUE :4928
##  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
##  Length:25596       Length:25596
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##
```

Visualisation

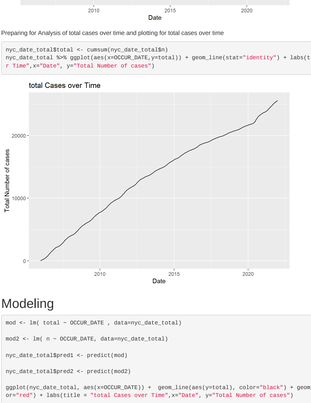
Let's do some visualisation, for that we will first create a couple of interesting tables

- The number of records by Boro->Precint
- The total victim race
- The total perpetrator race

```
nyc_boro <- nyc_data %>% group_by(BORO,PRECINCT) %>% count()


nyc_vic_total <- nyc_data %>% group_by(VIC_RACE) %>% count()

nyc_vic_total %>% ggplot(aes(y=VIC_RACE, x=n)) + geom_bar(stat="identity") + labs(title = "Victim Race",x="Amount" , y="Race")
```

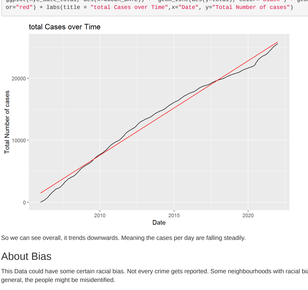



```
nyc_perp_total <- nyc_data %>% group_by(PERP_RACE) %>% count()

nyc_perp_total %>% ggplot(aes(y=PERP_RACE, x=n)) + geom_bar(stat="identity") + labs(title = "Perp Race",x="Amount" , y="Race")
```



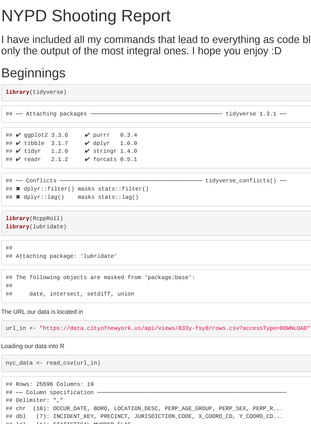
How many incidents per Date? Lets also plot this

```
nyc_date_total <- nyc_data %>% group_by(OCCUR_DATE) %>% count()


nyc_date_total %>% ggplot(aes(x=OCCUR_DATE,y=n)) + geom_bar(stat="identity") + labs(title = "Cases over Time",x="Date", y="Number of cases")
```



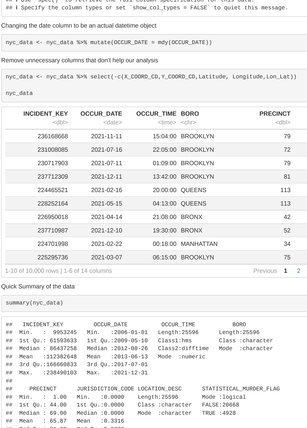
Preparing for Analysis of total cases over time and plotting for total cases over time

```
nyc_date_total$total <- cumsum(nyc_date_total$n)
nyc_date_total %>% ggplot(aes(x=OCCUR_DATE,y=total)) + geom_line(stat="identity") + labs(title = "total Cases over Time",x="Date", y="total Number of cases")
```



Modeling

```
mod <- lm( total ~ OCCUR_DATE , data=nyc_date_total)

mod2 <- lm( n ~ OCCUR_DATE, data=nyc_date_total)

nyc_date_total$pred1 <- predict(mod)

nyc_date_total$pred2 <- predict(mod2)

ggplot(nyc_date_total, aes(x=OCCUR_DATE)) +  geom_line(aes(y=total), color="black") + geom_line(aes(y=pred1), color="red") + labs(title = "total Cases over Time",x="Date", y="total Number of cases")
```



So we can see overall, it trends downwards. Meaning the cases per day are falling steadily.

About Bias

This Data could have some certain racial bias. Not every crime gets reported. Some neighbourhoods with racial bias might be profiled more. Or in general, the people might be misidentified.