NYPD_Shooting_Incident_Data_historic

C. Chen

1/26/2022

Description:

2 visualizations are done with the NYPD Shooting Incident Data. The variables used are: the BORO, OCCUR_DATE and MURDER_FLAG. The data was transformed to categorize by borough and 2 graphs are created. One showing the Shootings per borough and the other one showing the murders per borough.

Later another data subset is created with the data from the Staten Island borough and a linear prediction is made comparing shootings and murders. As a result we see that the prediction is a linear trend which proves that there's a relationship between the shootings and murders but there's definitely other factors that need to included in order to have a better prediction due to the outliers.

Some Bias is that we are not taking into consideration the population proportions to compare the boroughs. Another bias can be racial because you can get into a conclusion but I would like to know more about the social status and environments related to each incident. I noticed that there's a lot of missing data for perpetrator. I wanted to do an analysis comparing the victims and the perpetrator but noticed there was a lot of missing data.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
## Rows: 23585 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.
```

```
shooting_incident_data<-subset (shooting_incident_data, select = -c(LOCATION_DESC,INCIDENT_KEY,OCCUR_TI
colnames(shooting_incident_data)[colnames(shooting_incident_data) == "BORO"] <- "BOROUGH"
colnames(shooting_incident_data)[colnames(shooting_incident_data) == "VIC_AGE_GROUP"] <- "VICTIM_AGE"
colnames(shooting_incident_data)[colnames(shooting_incident_data) == "VIC_SEX"] <- "VICTIM_SEX"
colnames(shooting_incident_data)[colnames(shooting_incident_data) == "VIC_RACE"] <- "VICTIM_RACE"
library(lubridate)
```

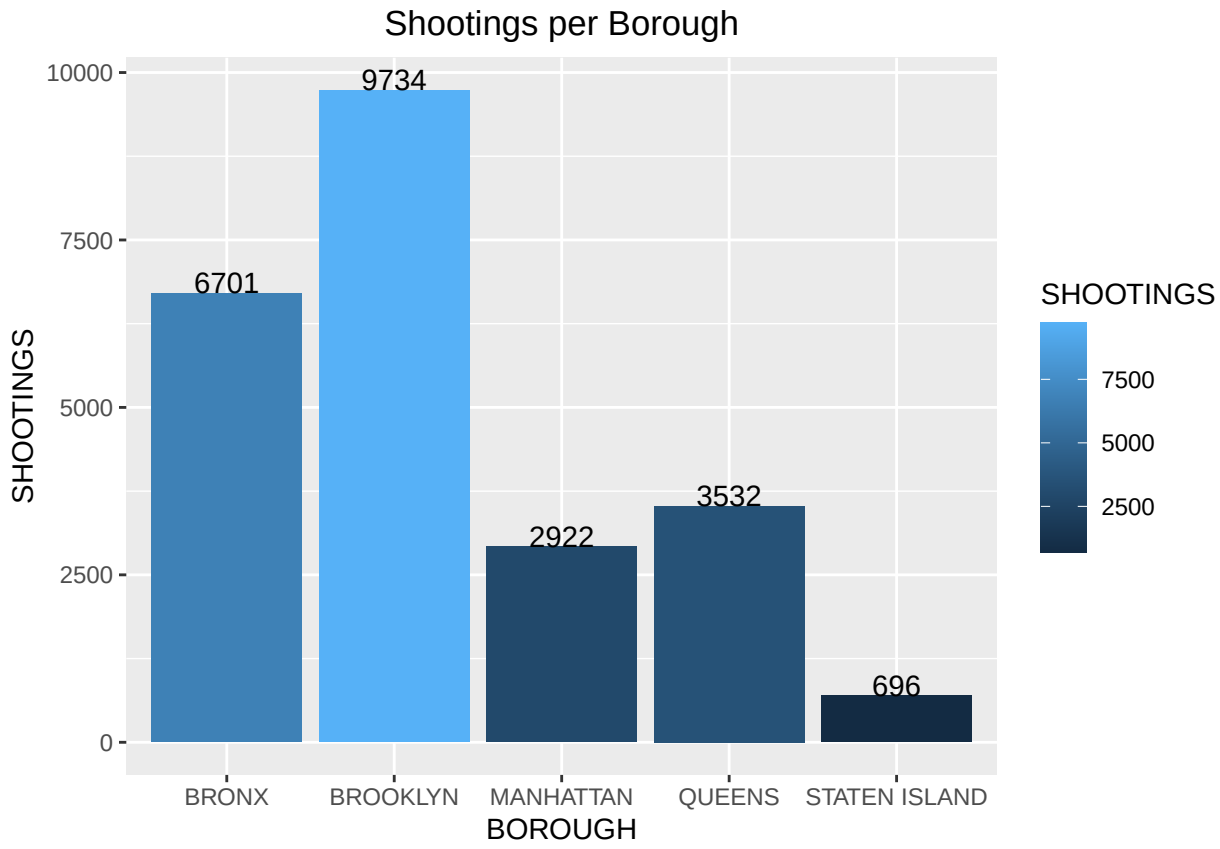
```
##
## Attaching package: 'lubridate'

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

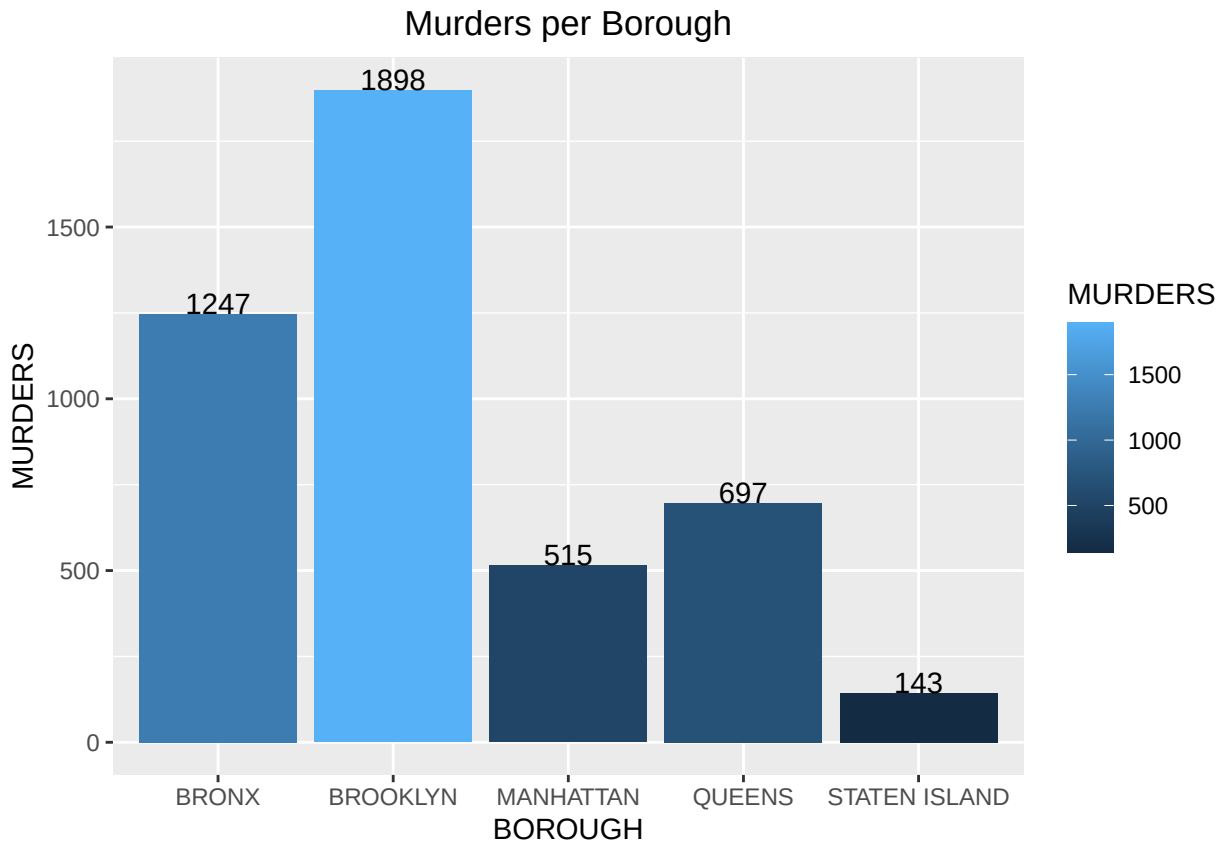
```
shooting_incident_data <- shooting_incident_data %>%
    mutate(DATE=mdy(OCCUR_DATE),MURDER=as.integer(shooting_incident_data$STATISTICAL_MURDER_FLAG),count
summary(shooting_incident_data)
```

```
##   OCCUR_DATE          BOROUGH          STATISTICAL_MURDER_FLAG
##  Length:23585       Length:23585       Mode :logical
##  Class :character   Class :character   FALSE:19085
##  Mode  :character   Mode  :character   TRUE :4500
##
##
##
##   VICTIM_AGE         VICTIM_SEX         VICTIM_RACE             DATE
##  Length:23585       Length:23585       Length:23585       Min.   :2006-01-01
##  Class :character   Class :character   Class :character   1st Qu.:2008-12-31
##  Mode  :character   Mode  :character   Mode  :character   Median :2012-02-27
##                                                           Mean   :2012-10-05
##                                                           3rd Qu.:2016-03-02
##                                                           Max.   :2020-12-31
##      MURDER       countOccurence
##  Min.   :0.0000   Min.   :1
##  1st Qu.:0.0000   1st Qu.:1
##  Median :0.0000   Median :1
##  Mean   :0.1908   Mean   :1
##  3rd Qu.:0.0000   3rd Qu.:1
##  Max.   :1.0000   Max.   :1
```

```
shootings_by_boro <-shooting_incident_data%>%
  group_by(BOROUGH)%>%
  summarize(MURDERS = sum(MURDER),SHOOTINGS = sum(countOccurence)) %>%
  select(BOROUGH,MURDERS,SHOOTINGS)%>%
  ungroup()
ggplot(shootings_by_boro, aes(x = BOROUGH, y = SHOOTINGS, fill = SHOOTINGS)) +
     geom_bar(stat = "identity") +
     geom_text(aes(label = SHOOTINGS), vjust = 0) +ggtitle("Shootings per Borough")+
     theme(plot.title=element_text(hjust=0.5))
```




```
ggplot(shootings_by_boro, aes(x = BOROUGH, y = MURDERS, fill = MURDERS)) +
     geom_bar(stat = "identity") +
     geom_text(aes(label = MURDERS), vjust = 0) +ggtitle("Murders per Borough")+
     theme(plot.title=element_text(hjust=0.5))
```

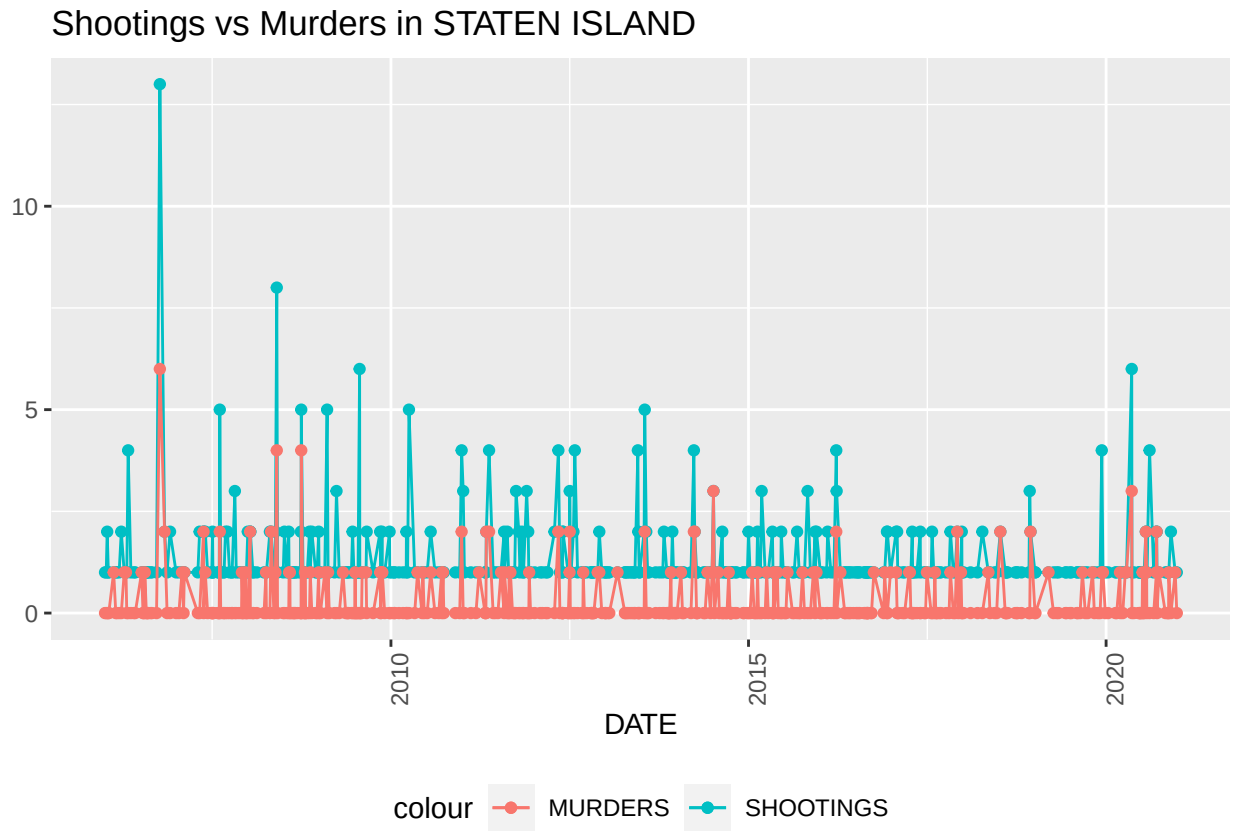



```
shootings_by_dates <-shooting_incident_data%>%
  group_by(DATE,BOROUGH)%>%
  summarize(MURDERS = sum(MURDER),SHOOTINGS = sum(countOccurence)) %>%
  select(BOROUGH,MURDERS,SHOOTINGS)%>%
  ungroup()
```

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

```
## Adding missing grouping variables: `DATE`
```

```
boro <-"STATEN ISLAND"
shootings_by_dates %>%
  filter(BOROUGH == boro) %>%
  ggplot(aes(x = DATE, y = SHOOTINGS))+
  geom_line(aes(color = "SHOOTINGS")) +
  geom_point(aes(color = "SHOOTINGS"))+
  geom_line(aes(y= MURDERS, color = "MURDERS")) +
  geom_point(aes(y=MURDERS, color = "MURDERS")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle =90,hjust=1)) +
  labs(title = str_c("Shootings vs Murders in ", boro), y=NULL)
```




```
shootings_by_Staten_Island<-shootings_by_dates %>%
     filter(BOROUGH == "STATEN ISLAND") %>%
     select(BOROUGH,MURDERS,SHOOTINGS,DATE)%>%
     ungroup()
mod<-lm( MURDERS ~ SHOOTINGS,data=shootings_by_Staten_Island)
summary(mod)
```

```
##
## Call:
## lm(formula = MURDERS ~ SHOOTINGS, data = shootings_by_Staten_Island)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0224 -0.1486 -0.1486 -0.1486  2.3524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.22616    0.03941  -5.739 1.62e-08 ***
## SHOOTINGS    0.37476    0.02402  15.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.522 on 519 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.3179
## F-statistic: 243.4 on 1 and 519 DF,  p-value: < 2.2e-16
```

```
shootings_by_Staten_Island_with_pred<-shootings_by_Staten_Island%>%
  mutate(pred=predict(mod))
shootings_by_Staten_Island_with_pred %>% ggplot() +
 geom_point(aes(x = SHOOTINGS, y = MURDERS ), color ="blue")+
 geom_point(aes(x = SHOOTINGS, y = pred), color = "red")
```