NYPD_Shooting

Henry Clay Ver Valen

5/21/2021

First, read in the NYPD Shooting incident data. The CSV file is located in ./data (use setwd to set your working directory to the same folder as this Rmd file is located) and the original file can be downloaded at https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic.

You will need tidyverse and lubridate install.packages("tidyverse") install.packages("lubridate") library(tidyverse) library(lubridate)

```
shooting_data <- read_csv("./data/NYPD_Shooting_Incident_Data__Historic_.csv")
```

Now eliminate INCIDENT_KEY and all columns after VIC_RACE and convert OCCUR_DATE to a date data type

```
shooting_data <- shooting_data %>%
  select(OCCUR_DATE:VIC_RACE) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Show summary of the data

```
summary(shooting_data)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO              PRECINCT
##  Min.   :2006-01-01   Length:23568      Length:23568       Min.   :  1.00
##  1st Qu.:2008-12-30   Class1:hms        Class :character   1st Qu.: 44.00
##  Median :2012-02-26   Class2:difftime   Mode  :character   Median : 69.00
##  Mean   :2012-10-03   Mode  :numeric                       Mean   : 66.21
##  3rd Qu.:2016-02-28                                        3rd Qu.: 81.00
##  Max.   :2020-12-31                                        Max.   :123.00
##
##  JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.:0.0000    Class :character   FALSE:19080
##  Median :0.0000    Mode  :character   TRUE :4488
##  Mean   :0.3323
##  3rd Qu.:0.0000
##  Max.   :2.0000
##  NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##    VIC_SEX            VIC_RACE
##  Length:23568       Length:23568
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

The visualizations I will be using do not require any filtering of missing values,but if it did I could do it with:

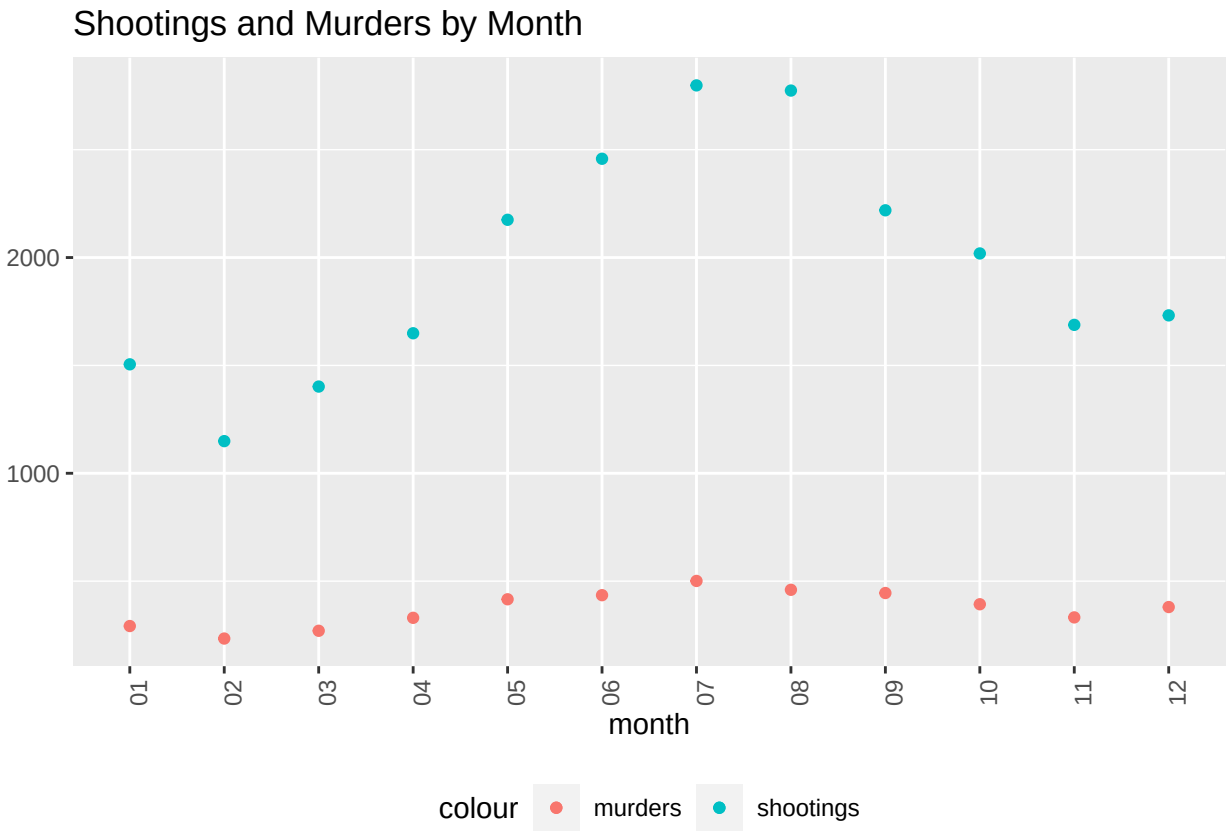
```
shooting_data_no_missing <- shooting_data %>%
  filter(PERP_AGE_GROUP != "NA" & PERP_AGE_GROUP != "UNKNOWN" & PERP_SEX != "NA" &
           PERP_RACE != "NA" & VIC_AGE_GROUP != "NA" & VIC_AGE_GROUP != "UNKNOWN" &
           VIC_SEX != "NA" & VIC_RACE != "NA")
```

Group the data by month for both murders and shootings for the first visualization

```
month_group <- shooting_data %>% group_by(month=format(floor_date(OCCUR_DATE, "month"), format="%m")) %
  summarize(murders = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE), shootings = length(STATISTICAL_MURDER_
  ungroup()
```

First visualization - Shootings and Murders by Month

```
month_group %>%
  ggplot(aes(x = month, y = murders)) +
  geom_point(aes(color = "murders")) +
  geom_point(aes(y = shootings, color = "shootings")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Shootings and Murders by Month", y = NULL)
```



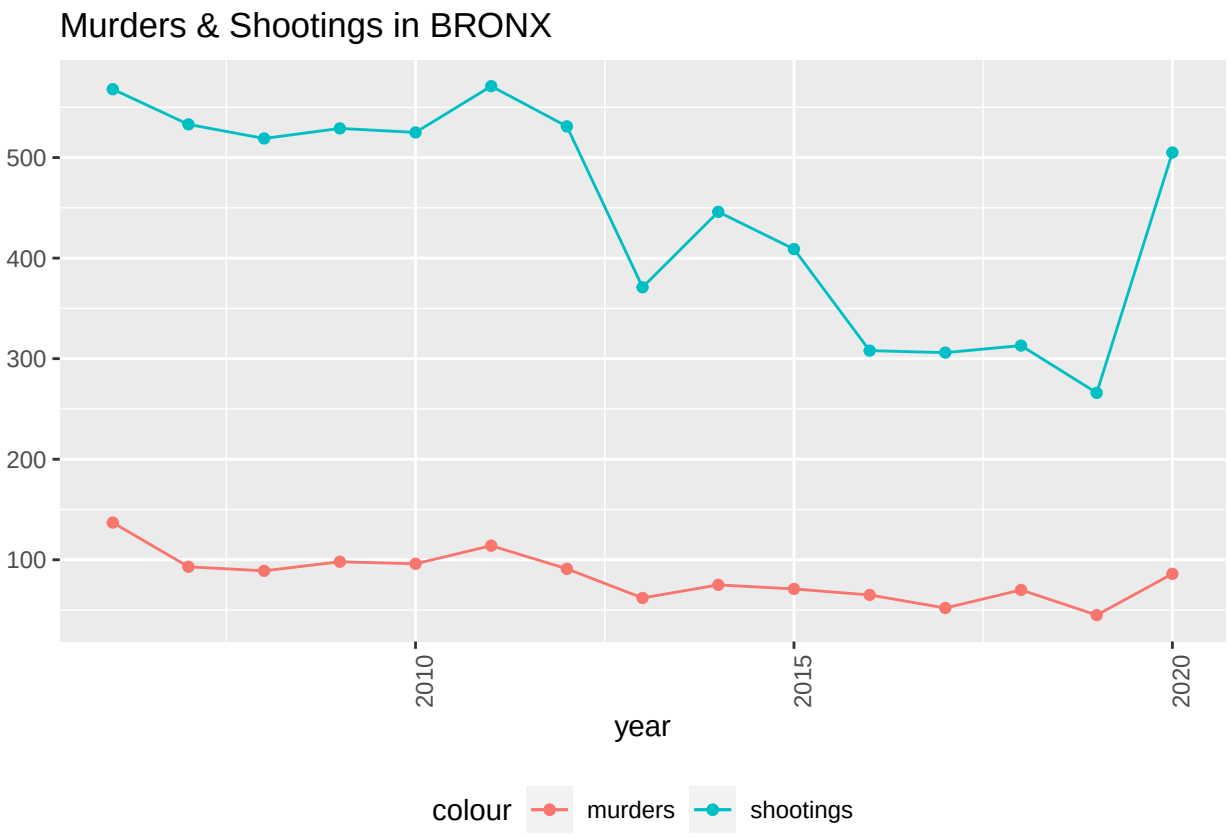
Group the data by borough and year for murder and shootings for the second visualization

```
boro_group <- shooting_data %>% group_by(BORO, year=year(OCCUR_DATE)) %>%
  summarize(murders = sum(STATISTICAL_MURDER_FLAG, na.rm = TRUE), shootings = length(STATISTICAL_MURDER_
  select(BORO, year, murders, shootings) %>%
  ungroup()
```

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

Second visualization: Murders and shootings by year for Bronx

```
boro <- "BRONX"
boro_group %>%
  filter(BORO == boro) %>%
  ggplot(aes(x = year, y = murders)) +
  geom_line(aes(color = "murders")) +
  geom_point(aes(color = "murders")) +
  geom_line(aes(y = shootings, color = "shootings")) +
  geom_point(aes(y = shootings, color = "shootings")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("Murders & Shootings in ", boro), y = NULL)
```



Find the three boroughs with the highest murder counts since 2010

```
boro_group %>% filter(year >= 2010) %>% group_by(BORO) %>% summarize(murders = sum(murders)) %>% slice_
```

```
## # A tibble: 3 x 2
##   BORO     murders
##   <chr>      <int>
## 1 BROOKLYN    1225
## 2 BRONX        827
## 3 QUEENS       490
```

Bias identification: At first I was very interested in seeing how race and age might play out in these shooting incidents, but then realized how fraught with biases both of these were, both my own and in the race identifications available in the data as well as the very broad age groupings that were used.

So to avoid these biases both my own and in the data, I looked only at murders and shootings as they relate to time, either month of the year or year over year. The exception to this is the analysis of the boroughs with the highest number of murders. One might think that Manhattan is a safer place from this, but instead it could be that most murders happen in the evenings and Manhattan has more businesses than residences. To find out if this could be biasing the results would require further research and data.