NYPD_Shooting_Final_Version

Anonymous

8/13/2021

This is a data set that contains a list of every shooting incident that occurred in New York City from 2006 to December 2020. This data set includes details of every shooting, such as the date and time of the incident, and details about the victim and the perpetrator.

Step 1: Import the Data set

```
library(tidyverse)

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

```
## # A tibble: 23,568 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>            <dbl>              <dbl>
##  1    201575314 08/23/2019 22:10      QUEENS             103                  0
##  2    205748546 11/27/2019 15:54      BRONX               40                  0
##  3    193118596 02/02/2019 19:40      MANHATTAN           23                  0
##  4    204192600 10/24/2019 00:52      STATEN ISLAND      121                  0
##  5    201483468 08/22/2019 18:03      BRONX               46                  0
##  6    198255460 06/07/2019 17:50      BROOKLYN            73                  0
##  7    194570529 03/11/2019 16:30      BROOKLYN            81                  0
##  8    203211777 10/03/2019 01:45      BROOKLYN            67                  0
##  9    193694863 02/17/2019 03:00      QUEENS             114                  2
## 10    199582060 07/10/2019 02:56      BROOKLYN            69                  0
## # ... with 23,558 more rows, and 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

```
summary(url_info)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568      Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms        Class :character
##  Median : 83365370   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102218616                      Mode  :numeric
```

```
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  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: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           X_COORD_CD        Y_COORD_CD
##  Length:23568       Length:23568       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Step 2: Tidying the Data

Regarding NA, I was able to see that the columns that were filled with NA values were "Location Description", and all the columns pertaining to "Perpetrator". Originally I planned on removing the rows with these NA values, but instead, I decided to leave them in, unless the bulk of my analysis was going to focus on the perpetrator data.

Formatting the Columns:

For the most part, the data is in a usable format. I only went ahead and changed the format of the date and time columns:

```
library(lubridate)
```

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

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

```
Shooting_Data <- url_info %>%

rename(occur_date = `OCCUR_DATE`) %>%

rename(occur_time = `OCCUR_TIME`) %>%

mutate(occur_date = mdy(occur_date))
```

Once I had tidied the data, I had to decide what questions I wanted to answer from my analysis of the data set. After viewing the data set, I chose the following focus points:

- Seeing which Borough had the highest number of shootings
- Based on Question#1, seeing how the frequency of shootings changed in this borough over time?
- Seeing which Borough had the highest number of "Murder Shootings"

Step 3: Visualizing the Data

```
Shooting_Data %>% count(BORO, sort = TRUE)
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BROOKLYN       9722
## 2 BRONX          6700
## 3 QUEENS         3527
## 4 MANHATTAN      2921
## 5 STATEN ISLAND   698
```

```
#From this, its evident that Brooklyn is the borough with the highest number of shootings from 2006 to
#2020. Once I found this out, I needed to see how the frequency, or number of shootings
#changed over time in Brooklyn:

borough_data <- table(Shooting_Data$BORO)
borough_data <- as.data.frame(borough_data)

B <-subset(Shooting_Data, BORO=='BROOKLYN', select=c(BORO, occur_date))
n <- 5
B$YEAR <- substr(B$occur_date, nchar(B$occur_date) - n + 1, nchar(B$occur_date))

B <- subset(B, select = -c(occur_date))

BROOKLYN <- table(B$YEAR)
BROOKLYN <- as.data.frame(BROOKLYN)
BROOKLYN
```

```
##     Var1 Freq
## 1  2006-  850
## 2  2007-  833
## 3  2008-  785
## 4  2009-  770
## 5  2010-  805
## 6  2011-  839
## 7  2012-  651
## 8  2013-  593
## 9  2014-  614
## 10 2015-  583
## 11 2016-  499
## 12 2017-  357
## 13 2018-  362
## 14 2019-  372
## 15 2020-  809
```

4) Plotting/Analyzing the Data

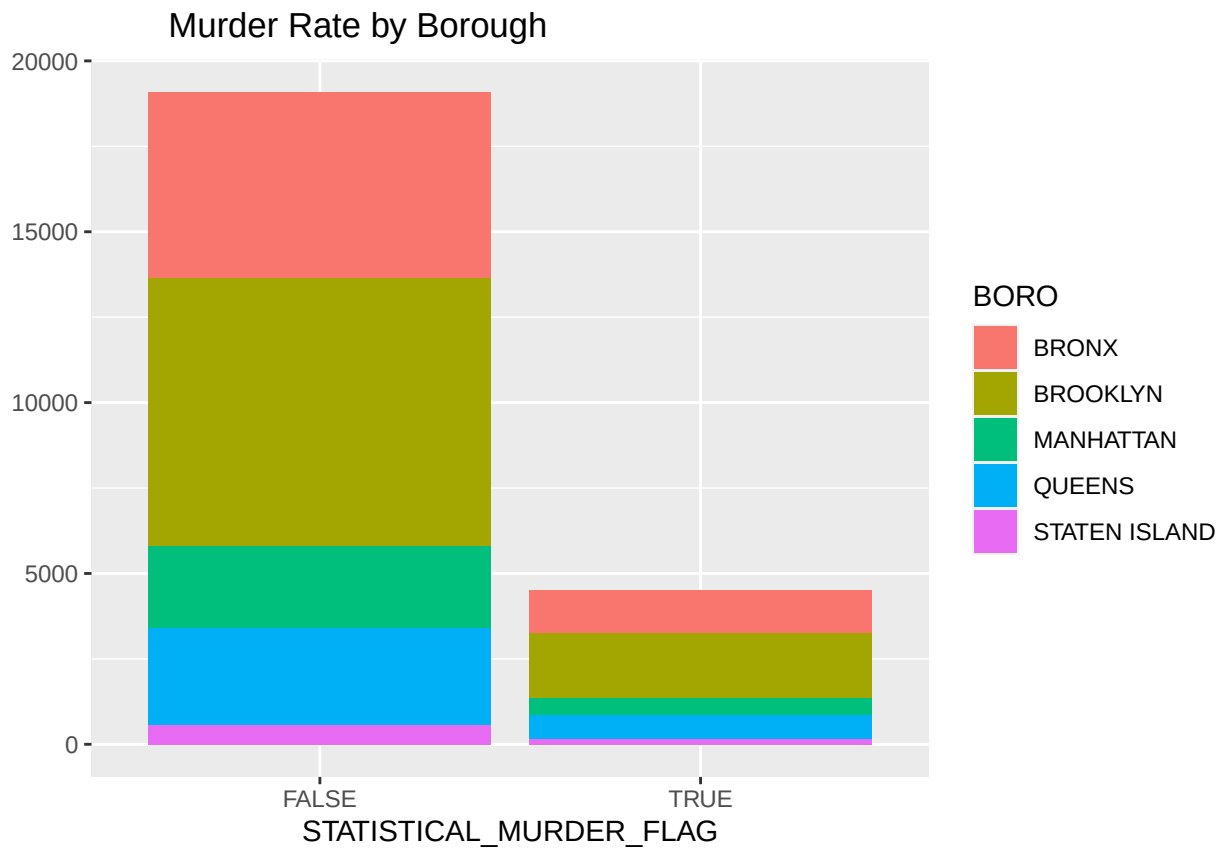
```
#Based off of the table above, I plotted the change in the number of shootings over the course of
#2006-2020 in Brooklyn:

ggplot(data=BROOKLYN, aes(x=Var1, y=Freq, group=1)) +
  geom_line()+
  geom_point()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(title = str_c("Frequency of Shootings In Brooklyn from 2006-2020 "), y = NULL)
```

From this graph, we can see that shootings started at a high number, then went down to almost nonexistent, and recently started to pick up again.

Secondly, I decided to plot the number of "Murder" vs. "Non-Murder" shootings per borough, to see which borough had the highest number of "Murder" shootings:

```
murder_counts <- Shooting_Data %>%
                 group_by(BORO, STATISTICAL_MURDER_FLAG, occur_date) %>%
                 tally()
ggplot(data = murder_counts, aes(x = STATISTICAL_MURDER_FLAG, y = n, fill = BORO)) +
     geom_bar(stat = "identity")+
  labs(title = str_c("        Murder Rate by Borough"), y = NULL)
```



5) Conclusion

1) Brooklyn had a significantly higher # of shootings compared to the other boroughs.

2) Shootings in Brooklyn started out at a high frequency in 2006, then went down to almost nonexistent in 2017. The numbers started to pick up again after 2019, and went up back to the original level in 2006.

3) Brooklyn not only had the highest # of shootings, but the highest # of murders as well.

Bias:

1) Personal Bias: My areas of interest and what I thought I already knew about the shootings in New York

2) Data bias: the data itself, or the website it was taken from