NYPDShootingData

6/9/2021

```
knitr::opts_chunk$set(echo = TRUE)
library(lubridate)
```

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

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

```
library(ggplot2)

library(tidyverse)
```

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

```
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

Questions

I wanted to see what the shootings per 100,000 looked like in each borough to compare and see how close they are, and I wanted to see if seasonality or time of the year affected the number of shootings.

##Summary This data set is a list of every shooting incident that occurred in New York City going back to 2006 through the end of the previous calender year. Each record is labeled by an Incident key, and originally contains the following information on the shooting:

-The date of the shooting

-The time the shooting occurred

-Where is NYC the shooting took place(i.e Manhattan, Queens, e.t.c.)

-The Police Precinct Number

-The Jurisdiction Code

-A description of the location

-A Statistical murder flag

-The Perpetrator's age group

-The Perpetrator's Sex

-The Perpetrator's Race

-The Victim's age group

-The Victim's Sex

-The Victim's Race

-The X coordinate

-The Y coordinate

-The Latitude

-The Longitude

-The longitude and latitude point

Reading in the CSV file*

```
#This reads in the data set
shooting_data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLO

summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  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       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     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
##
```

Tidying up the data

With the data read in, we can now tidy it up and remove, name, and combine columns so that the data is much easier to use and understand.

First, We are going to get rid of the following columns:

-Incident Key

-Precinct

-Jurisdiction code

-location description

-And all of the longitude and latitude data, including the X and Y coordinates

We are also going to rename some of the columns so that they can be easier read and understood. I am going to keep the new names all capitals as well, just so that they are similar to what is already in the data.

-OCCUR_DATE = DATE

-OCCUR_TIME = TIME

-STATISTICAL_MURDER_FLAG = MURDER_FLAG

Looking at the summary of the data we can also see that DATE and TIME are character variable types, we want to change that to date and time variable types. We are also going to change the MURDER_FLAG column from a character type to a logical type.

```
shooting_data <- shooting_data %>%
    select(-c(INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, LOCATION_DESC,
              X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat)) %>%
    rename(DATE = 'OCCUR_DATE',
           TIME = 'OCCUR_TIME',
           MURDER_FLAG = 'STATISTICAL_MURDER_FLAG')%>%
    mutate(DATE = mdy(DATE),
           TIME = hms(TIME),
           MURDER_FLAG = as.logical(MURDER_FLAG))

summary(shooting_data)
```

```
##       DATE                TIME                            BORO
##  Min.   :2006-01-01   Min.   :0S                        Length:23568
##  1st Qu.:2008-12-30   1st Qu.:3H 20M 0S                 Class :character
##  Median :2012-02-26   Median :15H 0M 0S                 Mode  :character
##  Mean   :2012-10-03   Mean   :12H 32M 59.1318737270849S
##  3rd Qu.:2016-02-28   3rd Qu.:20H 44M 15S
##  Max.   :2020-12-31   Max.   :23H 59M 0S
##  MURDER_FLAG     PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  Mode :logical   Length:23568       Length:23568       Length:23568
##  FALSE:19080     Class :character   Class :character   Class :character
##  TRUE :4488      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Missing data from the set

```
# This prints the first 10 lines of shooting_data
head(shooting_data, 10)
```

```
##          DATE       TIME          BORO MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1  2019-08-23 22H 10M 0S        QUEENS       FALSE
## 2  2019-11-27 15H 54M 0S         BRONX       FALSE            <18        M
## 3  2019-02-02 19H 40M 0S     MANHATTAN       FALSE          18-24        M
## 4  2019-10-24     52M 0S STATEN ISLAND        TRUE          25-44        M
## 5  2019-08-22  18H 3M 0S         BRONX       FALSE          25-44        M
## 6  2019-06-07 17H 50M 0S      BROOKLYN       FALSE          45-64        M
## 7  2019-03-11 16H 30M 0S      BROOKLYN       FALSE          18-24        M
## 8  2019-10-03  1H 45M 0S      BROOKLYN        TRUE
## 9  2019-02-17   3H 0M 0S        QUEENS       FALSE          18-24        M
## 10 2019-07-10  2H 56M 0S      BROOKLYN       FALSE          25-44        M
##         PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                        25-44       M          BLACK
## 2           BLACK        25-44       F          BLACK
## 3  WHITE HISPANIC        18-24       M BLACK HISPANIC
## 4           BLACK        25-44       F          BLACK
## 5  BLACK HISPANIC        18-24       M          BLACK
## 6  WHITE HISPANIC        25-44       M          BLACK
## 7           BLACK        25-44       M          BLACK
## 8                        25-44       M          BLACK
## 9           BLACK        25-44       M          BLACK
## 10          BLACK        25-44       M          BLACK
```

Looking at the first 10 rows of the data set above, we can see that we are actually missing data in a few places. We are missing some entries in the PERP_AGE_GROUP, PERP_SEX, and PERP_RACE Columns. We

don't exactly know why this data is missing, the most likely though is that the police either don't know who the perpetrator is. So, for the missing data we are just going assume that missing values means that the police don't have any information on the perpetrator.

Visualizing the Data

To start visualizing the data, I am going to create a new data frame called shooting_loc, which has the sum of instances of the particular borough.

```
shooting_loc <- count(shooting_data, BORO = shooting_data$BORO)

shooting_loc
```

```
##            BORO    n
## 1         BRONX 6700
## 2      BROOKLYN 9722
## 3     MANHATTAN 2921
## 4        QUEENS 3527
## 5 STATEN ISLAND  698
```

Now that we have an understanding of the total number of shootings in each borough, I want to see if there is a correlation with the population of each borough. Meaning, I want to visualize the shootings per 100,000 in each borough.

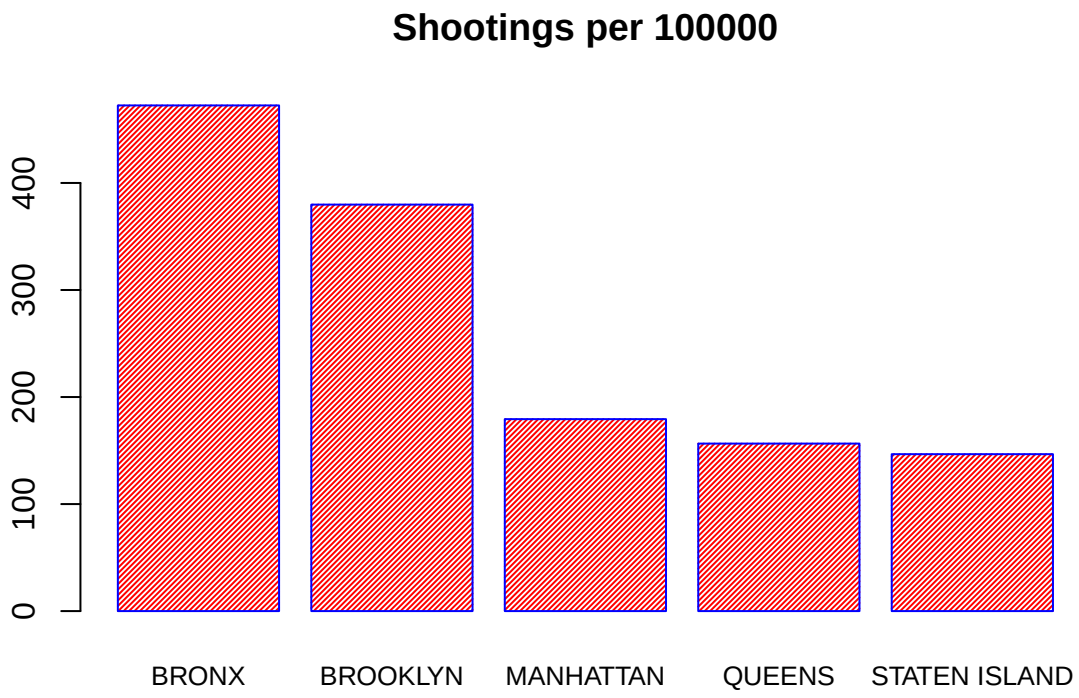
So, first we have to add population data to the shooting_loc data frame, and then we can do the calculations for the visualization. The population data was obtained from census.gov, and is the official estimates based on the most recent census data

```
# This creates a population vector so that it can easily be added to the data frame
pop_vec <- c(1418187, 2559903, 1628706, 2253858, 476143)
shooting_loc$population <- pop_vec

#This creates a new column that is the shootings per 100,000
shooting_loc <- transform(shooting_loc, per_100000 = ((n/population) * 100000))

shooting_loc
```

```
##            BORO    n population per_100000
## 1         BRONX 6700    1418187   472.4342
## 2      BROOKLYN 9722    2559903   379.7800
## 3     MANHATTAN 2921    1628706   179.3448
## 4        QUEENS 3527    2253858   156.4872
## 5 STATEN ISLAND  698     476143   146.5946
```

We are going to show a bar graph of the shootings per 100,000

```
barplot(shooting_loc$per_100000,
        main = "Shootings per 100000",
        names.arg = shooting_loc$BORO,
        border="blue",
        cex.names=0.8,
        col="red",
        density=50)
```



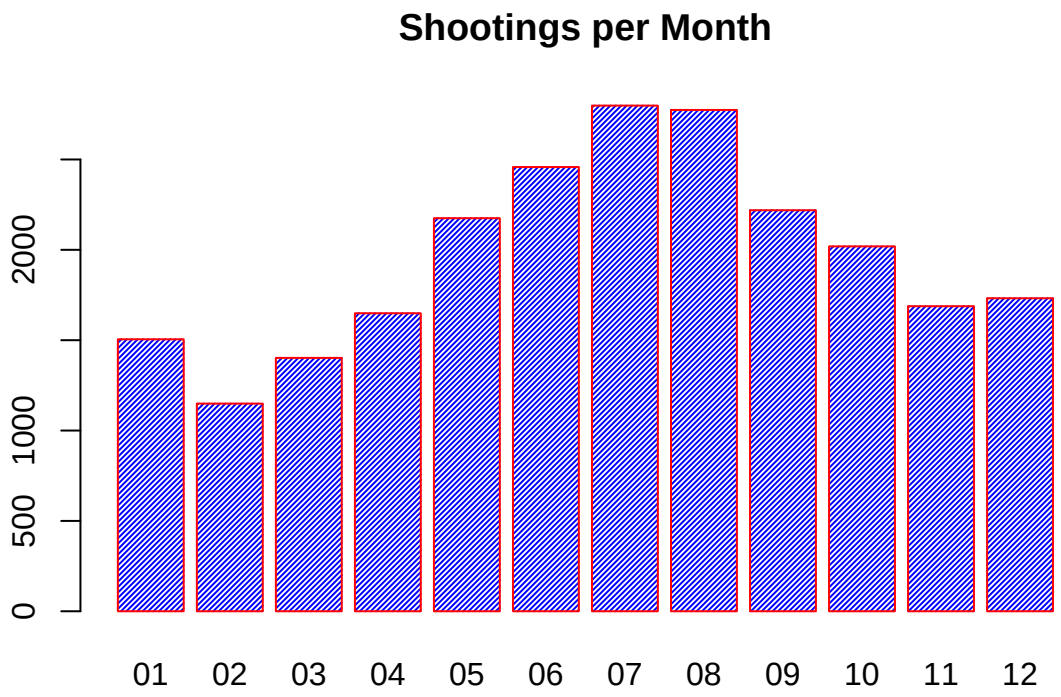
From the above graph, we can see that the shootings per 100,000 are not equal for every borough, meaning that the population is not the only factor in the frequency. There must be other factors. Things like population density, socio-economic status, gang membership, or a combination of multiple factors.

Now, I want to see if time of year has any impact on shootings, we are going to create a scatter plot based on the dates of the shootings. First, we have to create a new column of the dates, specifically, it is going to show what month the entry took place in to make creating the plot easier.

```
monthly <- shooting_data %>%
           mutate(month = format(DATE, "%m"), year = format(DATE, "%Y"))

monthly <- count(monthly, Month = monthly$month)
monthly
```

```
##    Month    n
## 1     01 1505
## 2     02 1149
## 3     03 1402
## 4     04 1649
## 5     05 2175
## 6     06 2458
## 7     07 2798
## 8     08 2774
## 9     09 2219
## 10    10 2019
## 11    11 1688
## 12    12 1732
```

```r
barplot(monthly$n,
        main = "Shootings per Month",
        names.arg = monthly$Month,
        border="red",
        col="blue",
        density=50)
```



As we can see in the above graph, it appears that more shootings happen in the summer time, which makes sense, as more people would be out and about the city. We see a dip in frequency in the winter months, as more people would stay in doors.

Conclusion and Bias

For Bias, my assumption that socio-economic status might play a roll in the shootings per 100,000 of a borough is defenitly a bias, it is not something that is backed up in the data that we have or used. Another source of bias would be my selection and analysis based on the months of the year, I assumed that there would be a difference over the year, because of the changing weather.

So, based of all the data we had gathered, and the types of we have concluded that the frequency of per captia shootings in a borough is not reliant on just the population and that the months and seasons of the year do play a part in the frequency of shootings. We can see that population is not the only factor, because if it was than the shootings per 100,000 would be about the same for each borough. We also saw that the summer had a much higher frequency of shootings compared to the winter.

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] forcats_0.5.1    stringr_1.4.0    dplyr_1.0.6      purrr_0.3.4
##  [5] readr_1.4.0      tidyr_1.1.3      tibble_3.1.2     tidyverse_1.3.1
##  [9] ggplot2_3.3.3    lubridate_1.7.10
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1  xfun_0.23         haven_2.4.1       colorspace_2.0-1
##  [5] vctrs_0.3.8       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.2.1        rlang_0.4.11      pillar_1.6.1      glue_1.4.2
## [13] withr_2.4.2       DBI_1.1.1         dbplyr_2.1.1      modelr_0.1.8
## [17] readxl_1.3.1      lifecycle_1.0.0   munsell_0.5.0     gtable_0.3.0
## [21] cellranger_1.1.0  rvest_1.0.0       evaluate_0.14     knitr_1.33
## [25] fansi_0.5.0       highr_0.9         broom_0.7.6       Rcpp_1.0.6
## [29] scales_1.1.1      backports_1.2.1   jsonlite_1.7.2    fs_1.5.0
## [33] hms_1.1.0         digest_0.6.27     stringi_1.6.1     grid_4.1.0
## [37] cli_2.5.0         tools_4.1.0       magrittr_2.0.1    crayon_1.4.1
## [41] pkgconfig_2.0.3   ellipsis_0.3.2    xml2_1.3.2        reprex_2.0.0
## [45] rstudioapi_0.13   assertthat_0.2.1  rmarkdown_2.8     httr_1.4.2
## [49] R6_2.5.0          compiler_4.1.0
```