NYPD shootings project

Bruno Lecona

4/23/2022

NYPD Shooting Incident Data (Historic) list every shooting incident that occurred in NYC, going back to 2006 through March 2022.

This data is reviewed by the Office of Management Analysis and Planning, before being posted on the NYPD website and each record includes information of the event, location and time of occurrence as well as information related to the suspect and victim.

Being said that, my main question is. Where have occurred the most cases in NYC? To have some insight of the most dangerous neighborhoods along the years.

Let's start by loading the data from the database mentioned above

```
NYPD_shooting_data <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=D
```

```
## 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.
```

And reading it to know more about the variables of the data set and look what are we going to use and what we don't need for this easy analysis

```
NYPD_shooting_data
```

```
## # A tibble: 23,585 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>       <dbl>              <dbl>
##  1     24050482 08/27/2006 05:35      BRONX          52                  0
##  2     77673979 03/11/2011 12:03      QUEENS        106                  0
##  3    203350417 10/06/2019 01:09      BROOKLYN       77                  0
##  4     80584527 09/04/2011 03:35      BRONX          40                  0
##  5     90843766 05/27/2013 21:16      QUEENS        100                  0
##  6     92393427 09/01/2013 04:17      BROOKLYN       67                  0
##  7     73057167 06/05/2010 21:16      BROOKLYN       77                  0
##  8    211362213 03/20/2020 21:27      BROOKLYN       81                  0
##  9    137564752 07/04/2014 00:25      QUEENS        101                  0
```

```
## 10    147024011 10/18/2015 01:33      QUEENS        106                 0
## # ... with 23,575 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>
```

```
str(NYPD_shooting_data)
```

```
## spec_tbl_df [23,585 x 19] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ INCIDENT_KEY           : num [1:23585] 2.41e+07 7.77e+07 2.03e+08 8.06e+07 9.08e+07 ...
##  $ OCCUR_DATE             : chr [1:23585] "08/27/2006" "03/11/2011" "10/06/2019" "09/04/2011" ...
##  $ OCCUR_TIME             : 'hms' num [1:23585] 05:35:00 12:03:00 01:09:00 03:35:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                   : chr [1:23585] "BRONX" "QUEENS" "BROOKLYN" "BRONX" ...
##  $ PRECINCT               : num [1:23585] 52 106 77 40 100 67 77 81 101 106 ...
##  $ JURISDICTION_CODE      : num [1:23585] 0 0 0 0 0 0 0 0 0 0 ...
##  $ LOCATION_DESC          : chr [1:23585] NA NA NA NA ...
##  $ STATISTICAL_MURDER_FLAG: logi [1:23585] TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ PERP_AGE_GROUP         : chr [1:23585] NA NA NA NA ...
##  $ PERP_SEX               : chr [1:23585] NA NA NA NA ...
##  $ PERP_RACE              : chr [1:23585] NA NA NA NA ...
##  $ VIC_AGE_GROUP          : chr [1:23585] "25-44" "65+" "18-24" "<18" ...
##  $ VIC_SEX                : chr [1:23585] "F" "M" "F" "M" ...
##  $ VIC_RACE               : chr [1:23585] "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num [1:23585] 1017542 1027543 995325 1007453 1041267 ...
##  $ Y_COORD_CD             : num [1:23585] 255919 186095 185155 233952 157134 ...
##  $ Latitude               : num [1:23585] 40.9 40.7 40.7 40.8 40.6 ...
##  $ Longitude              : num [1:23585] -73.9 -73.8 -74 -73.9 -73.8 ...
##  $ Lon_Lat                : chr [1:23585] "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73
##  - attr(*, "spec")=
##   .. cols(
##   ..   INCIDENT_KEY = col_double(),
##   ..   OCCUR_DATE = col_character(),
##   ..   OCCUR_TIME = col_time(format = ""),
##   ..   BORO = col_character(),
##   ..   PRECINCT = col_double(),
##   ..   JURISDICTION_CODE = col_double(),
##   ..   LOCATION_DESC = col_character(),
##   ..   STATISTICAL_MURDER_FLAG = col_logical(),
##   ..   PERP_AGE_GROUP = col_character(),
##   ..   PERP_SEX = col_character(),
##   ..   PERP_RACE = col_character(),
##   ..   VIC_AGE_GROUP = col_character(),
##   ..   VIC_SEX = col_character(),
##   ..   VIC_RACE = col_character(),
##   ..   X_COORD_CD = col_double(),
##   ..   Y_COORD_CD = col_double(),
##   ..   Latitude = col_double(),
##   ..   Longitude = col_double(),
##   ..   Lon_Lat = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Transform Occur_Date to a date format

```
NYPD_shooting_data$OCCUR_DATE<- anydate(NYPD_shooting_data$OCCUR_DATE)
```

Clean the data set to have only the necessary keys for the analysis, needing:INCIDENT_KEY PRECINT, Longitude, Latitude BORO, and OCCUR_DATE

```
NYPD_shooting_data <- NYPD_shooting_data %>% select(-c(JURISDICTION_CODE,LOCATION_DESC,STATISTICAL_MURDE

NYPD_shooting_data
```

```
## # A tibble: 23,585 x 7
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT Latitude Longitude
##           <dbl> <date>     <time>     <chr>       <dbl>    <dbl>     <dbl>
##  1     24050482 2006-08-27 05:35      BRONX          52     40.9     -73.9
##  2     77673979 2011-03-11 12:03      QUEENS        106     40.7     -73.8
##  3    203350417 2019-10-06 01:09      BROOKLYN       77     40.7     -74.0
##  4     80584527 2011-09-04 03:35      BRONX          40     40.8     -73.9
##  5     90843766 2013-05-27 21:16      QUEENS        100     40.6     -73.8
##  6     92393427 2013-09-01 04:17      BROOKLYN       67     40.6     -73.9
##  7     73057167 2010-06-05 21:16      BROOKLYN       77     40.7     -73.9
##  8    211362213 2020-03-20 21:27      BROOKLYN       81     40.7     -73.9
##  9    137564752 2014-07-04 00:25      QUEENS        101     40.6     -73.8
## 10    147024011 2015-10-18 01:33      QUEENS        106     40.7     -73.8
## # ... with 23,575 more rows
```

Taking a look to the Precincts with the most and less cases

4 of the 5 with higher cases are in Brooklyn(75,73,67,79) and one is in the Bronx (44)

For less cases we divide them between Manhattan (19,17,22) and Queens(112,111)

Now we know why everyone wants to live in Manhattan. (Just kidding)

These are the 5 Precincts with the most historical cases in NYC

```
precinct_head<-NYPD_shooting_data %>%
  group_by(PRECINCT) %>%
  tally()
 precinct_head<-precinct_head %>%
   arrange(desc(n)) %>%
   head(n=5)
 precinct_head
```

```
## # A tibble: 5 x 2
##   PRECINCT     n
##      <dbl> <int>
## 1       75  1375
## 2       73  1284
## 3       67  1101
## 4       79   921
## 5       44   841
```

These are the 5 Precincts with less historical cases in NYC

```
precinct_tail<-NYPD_shooting_data %>%
 group_by(PRECINCT) %>%
 tally()
precinct_tail<-precinct_tail %>%
  arrange(desc(n)) %>%
  tail(n=5)
precinct_tail
```

```
## # A tibble: 5 x 2
##   PRECINCT     n
##      <dbl> <int>
## 1      112    19
## 2       19    11
## 3       17     6
## 4      111     6
## 5       22     1
```

But I feel that I need to go deeper because having Precincts with the most cases does not necessarily mean they are indeed where there have been the most cases in general.

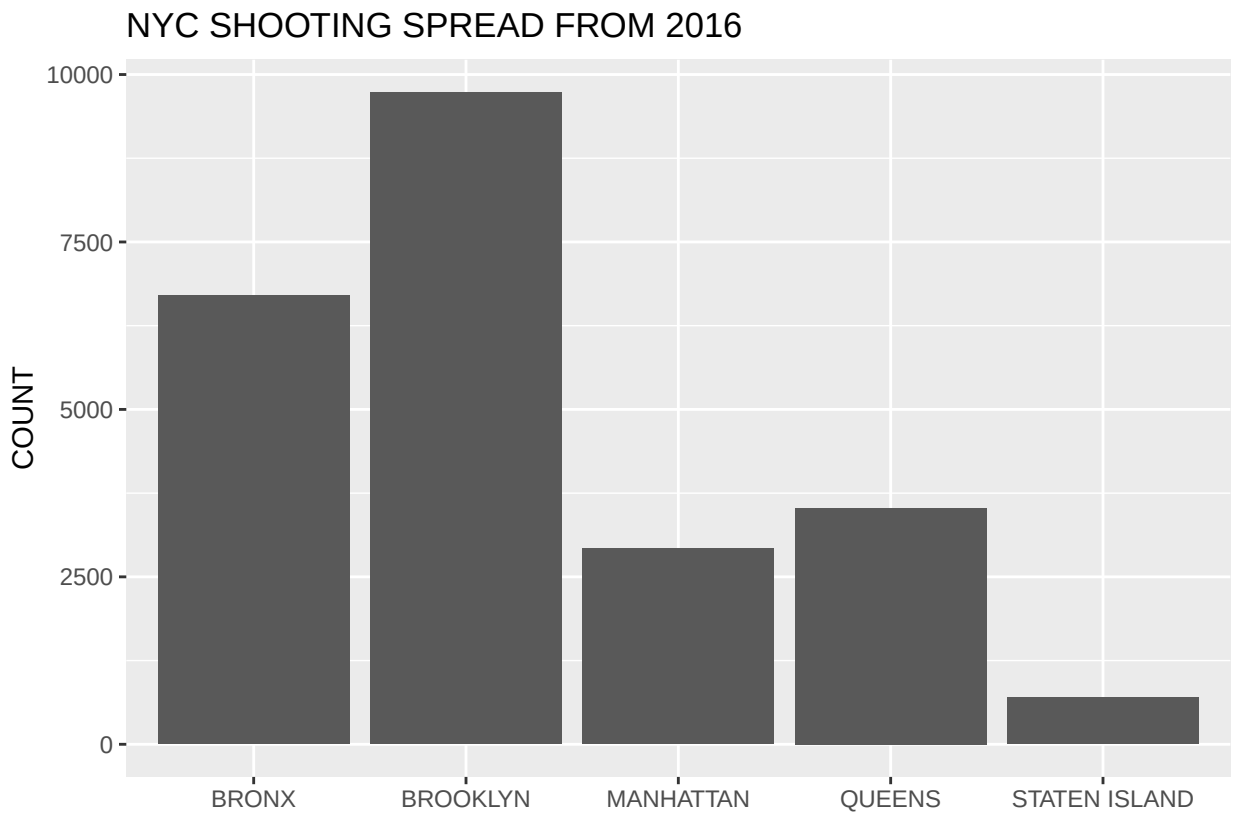
By creating this Bar plot of NYC shootings it is clearer how the cases are distributed.

Looking at you Brooklyn, our first place.

But we should be careful, more cases does not mean it is more dangerous. The size of the Borough and population, between other data that we could add to make a deeper analysis could be part to get rid of any possible bias.

But what we do know is that Brooklyn is the Borough with the highest number of shootings since 2006.

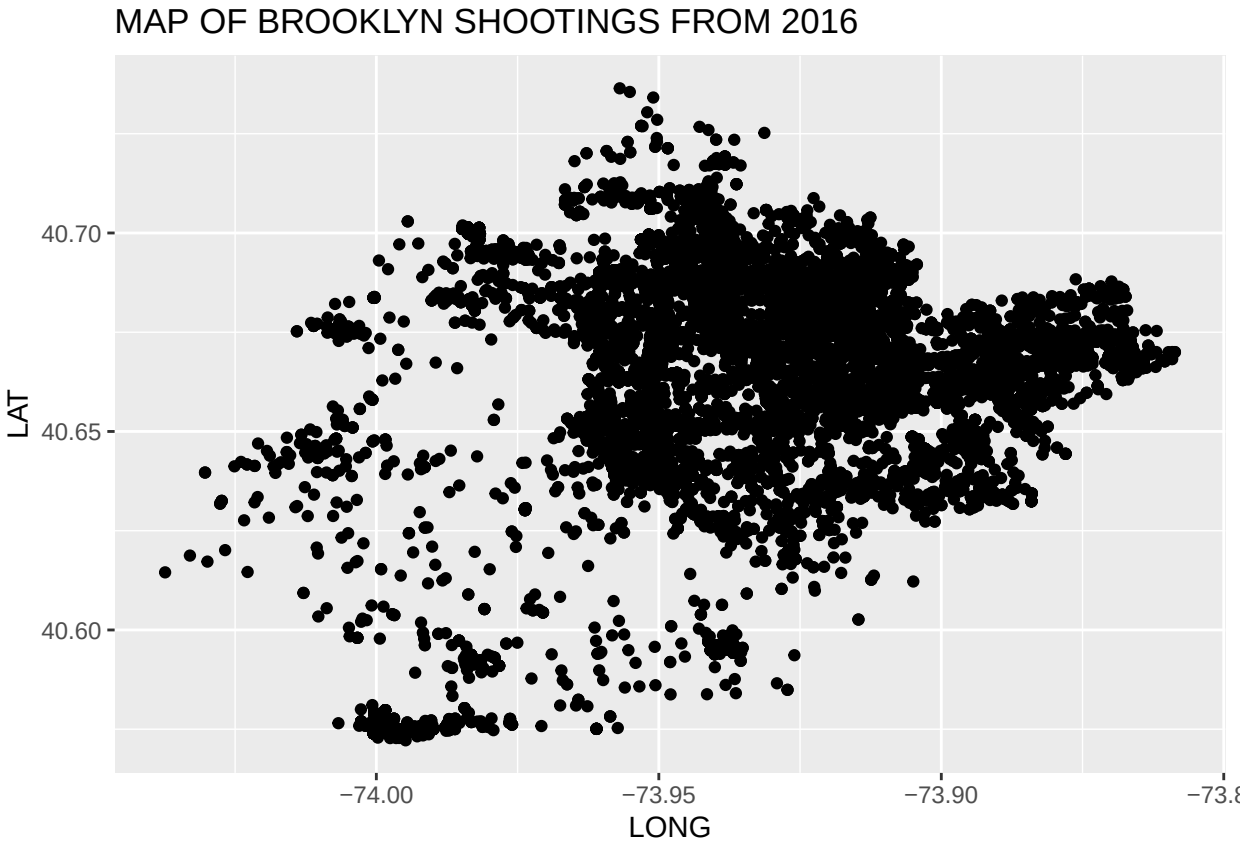
```
barplot_NYC<-NYPD_shooting_data %>%
  ggplot(aes(x=BORO))+geom_bar()+xlab("")+ ylab("COUNT")+ggtitle("NYC SHOOTING SPREAD FROM 2016")
barplot_NYC
```



For the last step we will visualize a map of Brooklyn to have an idea of which part of it has the most density

Basically, in the upper right quadrant of our point graph is where it is concentrated the majority of the cases registered so far.

```
brooklyn_map<-NYPD_shooting_data %>%
  filter(BORO=="BROOKLYN") %>%
  ggplot(aes(x=Longitude, y=Latitude))+geom_point()+xlab("LONG")+ ylab("LAT")+ggtitle("MAP OF BROOKLYN S
brooklyn_map
```



I'm left with some questions.

I would like to know if this last area in Brooklyn where we saw the majority of cases is an area with many businesses, malls or if it is an area where there has been or has criminal groups. I would also like to know, historically, in which approximate months of the year NYC has had the most cases, that could be a tendency to look at in a future analysis. Being influenced by tourism, weather, population, any elections, etc.

At the start of the project I would've thought that the most incidents could have been in Manhattan based on being one of the zones with most tourists in the world all of the year. And I am glad I was wrong, because trying to look at it from the start could have been a bias to worry about.