NYPD Shooting Incident

Akikat

5/21/2021

NYPD Shooting Incident Analysis

This is a template for analysis of NYPD Shooting Incident data, that can be downloaded from site https://catalog.data.gov/dataset.

Note: Prior to analysis the following packages should be installed: tidyverse, lubridate, janitor, details and ggplot2.

```
library(lubridate)
```

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

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

```
library(tidyverse)
```

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

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- 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()
```

```
library(dplyr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(details)
```

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

file <- read.csv(url_in)
```

Objectives of the analysis

There can be a lot of question asked from this data. I chose three:

- In which parts of the city shootings happen more often?
- What is the distribution of shooting incidents by days of a week?
- What are the age groups of most of the victims?

Initial acquitance with data

First let's have a look at the data:

```
str(file)
```

```
## 'data.frame':    23568 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  201575314 205748546 193118596 204192600 201483468 198255460 1945705
##  $ OCCUR_DATE             : chr  "08/23/2019" "11/27/2019" "02/02/2019" "10/24/2019" ...
##  $ OCCUR_TIME             : chr  "22:10:00" "15:54:00" "19:40:00" "00:52:00" ...
##  $ BORO                   : chr  "QUEENS" "BRONX" "MANHATTAN" "STATEN ISLAND" ...
##  $ PRECINCT               : int  103 40 23 121 46 73 81 67 114 69 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 2 0 ...
##  $ LOCATION_DESC          : chr  "" "" "" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "false" "true" ...
##  $ PERP_AGE_GROUP         : chr  "" "<18" "18-24" "25-44" ...
##  $ PERP_SEX               : chr  "" "M" "M" "M" ...
##  $ PERP_RACE              : chr  "" "BLACK" "WHITE HISPANIC" "BLACK" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "25-44" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "M" "F" "M" "F" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK HISPANIC" "BLACK" ...
##  $ X_COORD_CD             : chr  "1037451" "1006789" "999347" "938149" ...
##  $ Y_COORD_CD             : chr  "193561" "237559" "227795" "171781" ...
##  $ Latitude               : num  40.7 40.8 40.8 40.6 40.9 ...
##  $ Longitude              : num  -73.8 -73.9 -73.9 -74.2 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.80814071699996 40.697805308000056)" "POINT (-73.9185706
```

```
summary(file)
```

```
##   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
##
```

There are 19 variables in this file. Not all of them are useful for our analysis. Therefore, let us choose few important columns that represent city parts, age of victims and dates of incidents:

```
  newfile <- file  %>%
      select(1:4, 12)
```

Lets us make sure that we have got columns we wanted:

```
colnames(newfile)
```

```
## [1] "INCIDENT_KEY"  "OCCUR_DATE"    "OCCUR_TIME"    "BORO"
## [5] "VIC_AGE_GROUP"
```

```
summary(newfile)
```

```
##   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
##  VIC_AGE_GROUP
##  Length:23568
##  Class :character
##  Mode  :character
##
##
##
```

```
colSums(is.na(newfile)) # making sure there is no missing data
```

```
##  INCIDENT_KEY    OCCUR_DATE    OCCUR_TIME          BORO VIC_AGE_GROUP
##             0             0             0             0             0
```

We can derive day_of_week variable from date of an incident using lubridate library.

```
newfile$OCCUR_DATE <- mdy(newfile$OCCUR_DATE)

newfile$OCCUR_DATE <- as.Date(newfile$OCCUR_DATE)

newfile$day_of_week <- format(as.Date(newfile$OCCUR_DATE), "%A")
```

Data analysis

Now we can have a look at the distribution of incidents by the place, day of the week and age group.

```
table(newfile$VIC_AGE_GROUP)
```

```
##
##     <18   18-24   25-44   45-64     65+ UNKNOWN
##    2525    9000   10287    1536     155      65
```

```
table(newfile$BORO)
```

```
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          6700          9722          2921          3527           698
```

```
table(newfile$day_of_week)
```

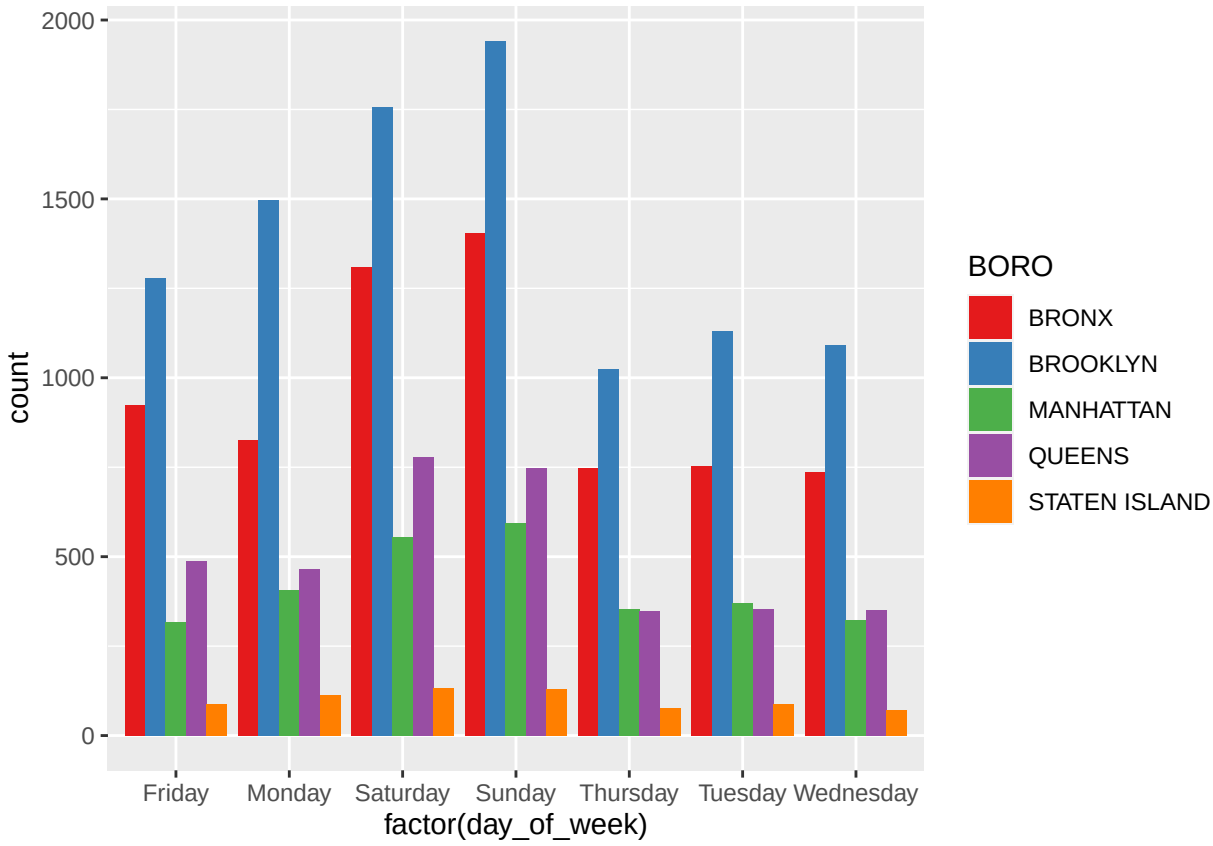
```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##      3095      3309      4532      4817      2549      2694      2572
```

```
tabyl(newfile, BORO, day_of_week)
```

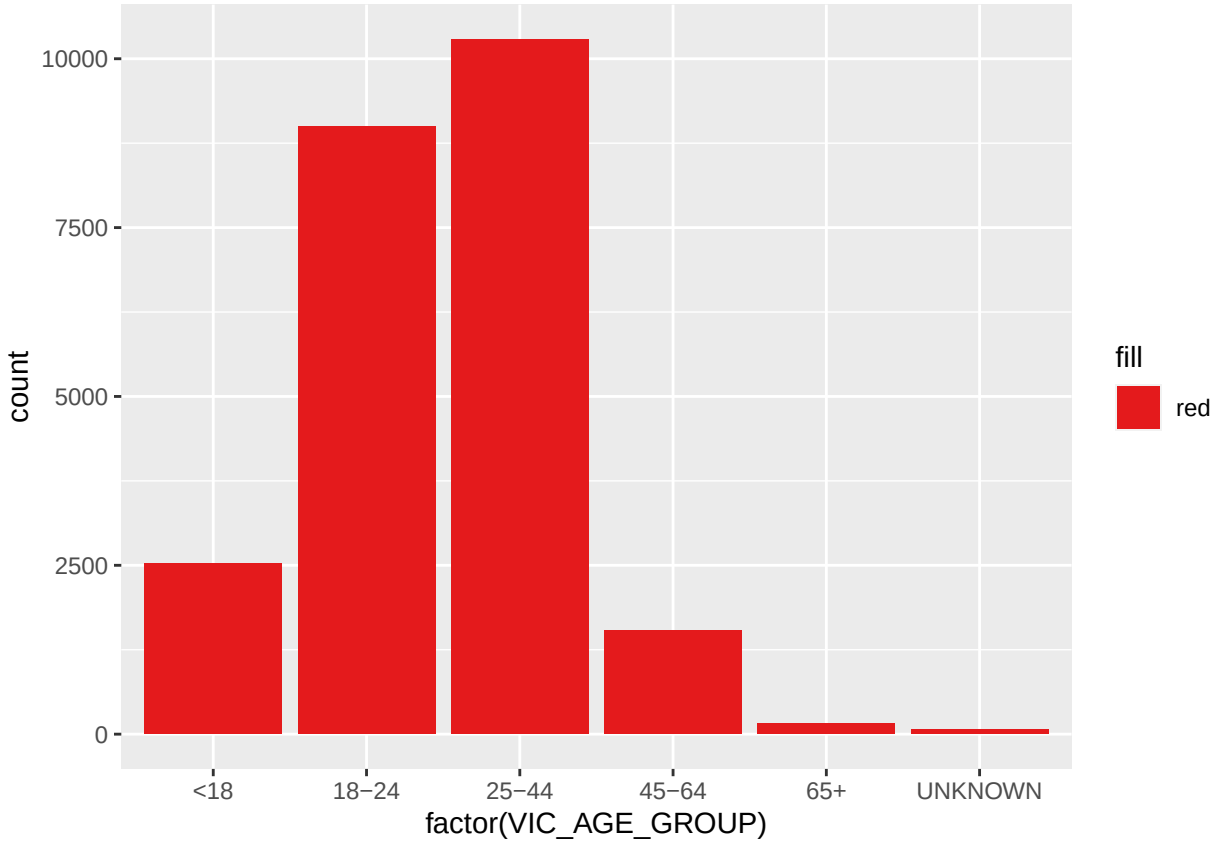
```
##           BORO Friday Monday Saturday Sunday Thursday Tuesday Wednesday
##          BRONX    923    827     1309   1405      747     753       736
##       BROOKLYN   1280   1498     1756   1942     1024    1130      1092
##      MANHATTAN    318    406      556    594      354     371       322
##         QUEENS    487    464      778    747      348     353       350
##  STATEN ISLAND     87    114      133    129       76      87        72
```

We can also make visuals to make the data more presentable.

```
ggplot(newfile, aes(factor(day_of_week), fill=BORO )) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
```




```
ggplot(newfile, aes(factor(VIC_AGE_GROUP), fill="red")) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
```



Results

It is obvious that shootings tend to happen on weekends and people between 18-44 are more likely to be victims. Also Brooklyn and Bronx seem to be more dangerous areas of New York city.

Source of bias

Since the data that I have chosen is not interpetable - date, age, place - it is not much prone to bias. There are some unknown data points in the database, but they are few comapred to main bulk of the data, so most likely does not affect it.

It was expected that in NY kids and elders are much less likely to be involved with shooting, and data confirmed it. I am not very familiar with NY boroughs. But judging from general stereotypes of Brooklyn being rough place and Staten Island an area for more wealthy citizens, looks like the data correlates with expectations. It is more likely to encounter shootings in poorer areas of any city ont he world.

And of course, weekends being traditionally time for more partying and social encounters, it is not surprising to see the rise of incidents on weekends and drop in mid-week.

Disclaimer: this is very basic analysis done as an exercise. Inferences could vary with more time spent on working on data.

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
```

```
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## 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
## system code page: 1251
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] details_0.2.1    janitor_2.1.0    forcats_0.5.1    stringr_1.4.0
##  [5] dplyr_1.0.5      purrr_0.3.4      readr_1.4.0      tidyr_1.1.3
##  [9] tibble_3.1.1     ggplot2_3.3.3    tidyverse_1.3.1  lubridate_1.7.10
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6         png_0.1-7          assertthat_0.2.1   rprojroot_2.0.2
##  [5] digest_0.6.27      utf8_1.2.1         R6_2.5.0           cellranger_1.1.0
##  [9] backports_1.2.1    reprex_2.0.0       evaluate_0.14      httr_1.4.2
## [13] highr_0.9          pillar_1.6.0       rlang_0.4.10       readxl_1.3.1
## [17] rstudioapi_0.13    rmarkdown_2.7      desc_1.3.0         labeling_0.4.2
## [21] munsell_0.5.0      broom_0.7.6        compiler_4.0.5     modelr_0.1.8
## [25] xfun_0.22          pkgconfig_2.0.3    clipr_0.7.1        htmltools_0.5.1.1
## [29] tidyselect_1.1.0   fansi_0.4.2        crayon_1.4.1       dbplyr_2.1.1
## [33] withr_2.4.2        grid_4.0.5         jsonlite_1.7.2     gtable_0.3.0
## [37] lifecycle_1.0.0    DBI_1.1.1          magrittr_2.0.1     scales_1.1.1
## [41] cli_2.4.0          stringi_1.5.3      farver_2.1.0       fs_1.5.0
## [45] snakecase_0.11.0   xml2_1.3.2         ellipsis_0.3.1     generics_0.1.0
## [49] vctrs_0.3.7        RColorBrewer_1.1-2 tools_4.0.5        glue_1.4.2
## [53] hms_1.0.0          yaml_2.2.1         colorspace_2.0-0   rvest_1.0.0
## [57] knitr_1.33         haven_2.4.1
```