NYPD Shooting Incidents Project Delivery

Learner

2023-01-24

Introdution:

We are asked to apply the Data Science Process we have learned from lectures of Dr. J Wall, we have New York City Shooting Incidents Data Set from year 2006 up to year 2021. I am going to add my comments describing each code under the 4 Steps as follows:

Step 1 : Start an Rmd Document.

Start an Rmd document that describes and imports the shooting project dataset in a reproducible manner

library(tidyverse)

```
\section{table}
── Attaching packages
─────────────────────────────────────────────────────────────────────── tidyverse
1.3.2 ──
✓ ggplot2 3.4.0      ✓ purrr   1.0.1
✓ tibble  3.1.8      ✓ dplyr   1.0.10
✓ tidyr   1.3.0      ✓ stringr 1.5.0
✓ readr   2.1.3      ✓ forcats 0.5.2
── Conflicts
──────────────────────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

library(tinytex)

library(lubridate)

```
\section{table}
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

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

nypd <- read_csv(url)

```
\section{table}

Rows: 25596 Columns: 19
── Column specification
────────────────────────────────────────────────────────────────────────────────
_______
Delimiter: ","
chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX,
PERP_RACE, VIC_AGE_GROUP, VIC_SEX,...
dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD,
Latitude, Longitude
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.
```

nypd

```
\section{table}

# A tibble: 25,596 × 19
   INCIDENT_…¹ OCCUR…² OCCUR…³ BORO  PRECI…⁴ JURIS…⁵ LOCAT…⁶ STATI…⁷ PERP_…⁸
PERP_…⁹ PERP_…˟ VIC_A…˟ VIC_SEX
         <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>
<chr>   <chr>   <chr>   <chr>
 1   236168668 11/11/… 15:04   BROO…      79       0 NA      FALSE   NA
NA      NA      18-24   M
 2   231008085 07/16/… 22:05   BROO…      72       0 NA      FALSE   45-64
M       ASIAN … 25-44   M
 3   230717903 07/11/… 01:09   BROO…      79       0 NA      FALSE   <18
M       BLACK   25-44   M
 4   237712309 12/11/… 13:42   BROO…      81       0 NA      FALSE   NA
NA      NA      25-44   M
 5   224465521 02/16/… 20:00   QUEE…     113       0 NA      FALSE   NA
NA      NA      25-44   M
 6   228252164 05/15/… 04:13   QUEE…     113       0 NA      TRUE    NA
NA      NA      25-44   M
 7   226950018 04/14/… 21:08   BRONX      42       0 COMMER… TRUE    NA
NA      NA      18-24   M
 8   237710987 12/10/… 19:30   BRONX      52       0 NA      FALSE   NA
NA      NA      25-44   M
 9   224701998 02/22/… 00:18   MANH…      34       0 NA      FALSE   NA
NA      NA      25-44   M
10   225295736 03/07/… 06:15   BROO…      75       0 NA      TRUE    25-44
M       BLACK … 25-44   M
# … with 25,586 more rows, 6 more variables: VIC_RACE <chr>, X_COORD_CD
```

```
<dbl>, Y_COORD_CD <dbl>,
#   Latitude <dbl>, Longitude <dbl>, Lon_Lat <chr>, and abbreviated variable
names ¹INCIDENT_KEY,
#   ²OCCUR_DATE, ³OCCUR_TIME, ⁴PRECINCT, ⁵JURISDICTION_CODE, ⁶LOCATION_DESC,
⁷STATISTICAL_MURDER_FLAG,
#   ⁸PERP_AGE_GROUP, ⁹PERP_SEX, ˟PERP_RACE, ˟VIC_AGE_GROUP
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all
variable names
```

Step 2 : Tidy and Transform Your Data

Add to your Rmd document a summary of the data and clean up your dataset by changing appropriate variables to factor and date types and getting rid of any columns not needed. Show the summary of your data to be sure there is no missing data. If there is missing data, describe how you plan to handle it.

summary(nypd)

```
\section{}

  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
PRECINCT
 Min.   :  9953245   Length:25596       Length:25596       Length:25596
Min.   :  1.00
 1st Qu.: 61593633   Class :character   Class1:hms         Class :character
1st Qu.: 44.00
 Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
Median : 69.00
 Mean   :112382648                      Mode  :numeric
Mean   : 65.87
 3rd Qu.:166660833
3rd Qu.: 81.00
 Max.   :238490103
Max.   :123.00

 JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
PERP_SEX
 Min.   :0.0000    Length:25596       Mode :logical           Length:25596
Length:25596
 1st Qu.:0.0000    Class :character   FALSE:20668             Class
:character   Class :character
 Median :0.0000    Mode  :character   TRUE :4928              Mode
:character   Mode  :character
 Mean   :0.3316
 3rd Qu.:0.0000
 Max.   :2.0000
 NA's   :2
  PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
X_COORD_CD
 Length:25596       Length:25596       Length:25596       Length:25596
```

```
Min.   : 914928
 Class :character   Class :character   Class :character   Class :character
1st Qu.:1000011
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
Median :1007715

Mean   :1009455

3rd Qu.:1016838

Max.   :1066815

   Y_COORD_CD        Latitude       Longitude        Lon_Lat
 Min.   :125757   Min.   :40.51   Min.   :-74.25   Length:25596
 1st Qu.:182782   1st Qu.:40.67   1st Qu.:-73.94   Class :character
 Median :194038   Median :40.70   Median :-73.92   Mode  :character
 Mean   :207894   Mean   :40.74   Mean   :-73.91
 3rd Qu.:239429   3rd Qu.:40.82   3rd Qu.:-73.88
 Max.   :271128   Max.   :40.91   Max.   :-73.70
```

*Removing unwanted columns

nypd = nypd %>% select(-c(INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, STATISTICAL_MURDER_FLAG, PERP_RACE, VIC_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

summary(nypd)

```
\section{table}

  OCCUR_DATE            BORO           LOCATION_DESC      PERP_AGE_GROUP
PERP_SEX
 Length:25596       Length:25596       Length:25596       Length:25596
Length:25596
 Class :character   Class :character   Class :character   Class :character
Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character
 VIC_AGE_GROUP        VIC_SEX
 Length:25596       Length:25596
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Changing the "OCCUR_DATE" column data type into datefor correct analysis

class(nypd$OCCUR_DATE) `\section{table} [1] "character"`

nypd <- nypd %>% mutate (OCCUR_DATE = mdy(OCCUR_DATE))

class(nypd$OCCUR_DATE)

```
 \section{table}
 [1] "Date"
```

changing names for some columns to better analysis

colnames(nypd)[1] <- “Date” colnames(nypd)[2] <- “Borough” colnames(nypd)[3] <- “place” colnames(nypd)[4] <- “Perp_Age” colnames(nypd)[5] <- “Perp_Gender” colnames(nypd)[6] <- “Vic_Age” colnames(nypd)[7] <- “Vic_Gender”

Step 3 Add Visualizations and Analysis

Add at least two different visualizations & some analysis to your Rmd. Does this raise additional questions that you should investigate?

nypd_Borough <- nypd %>% count(Borough)

nypd_Borough

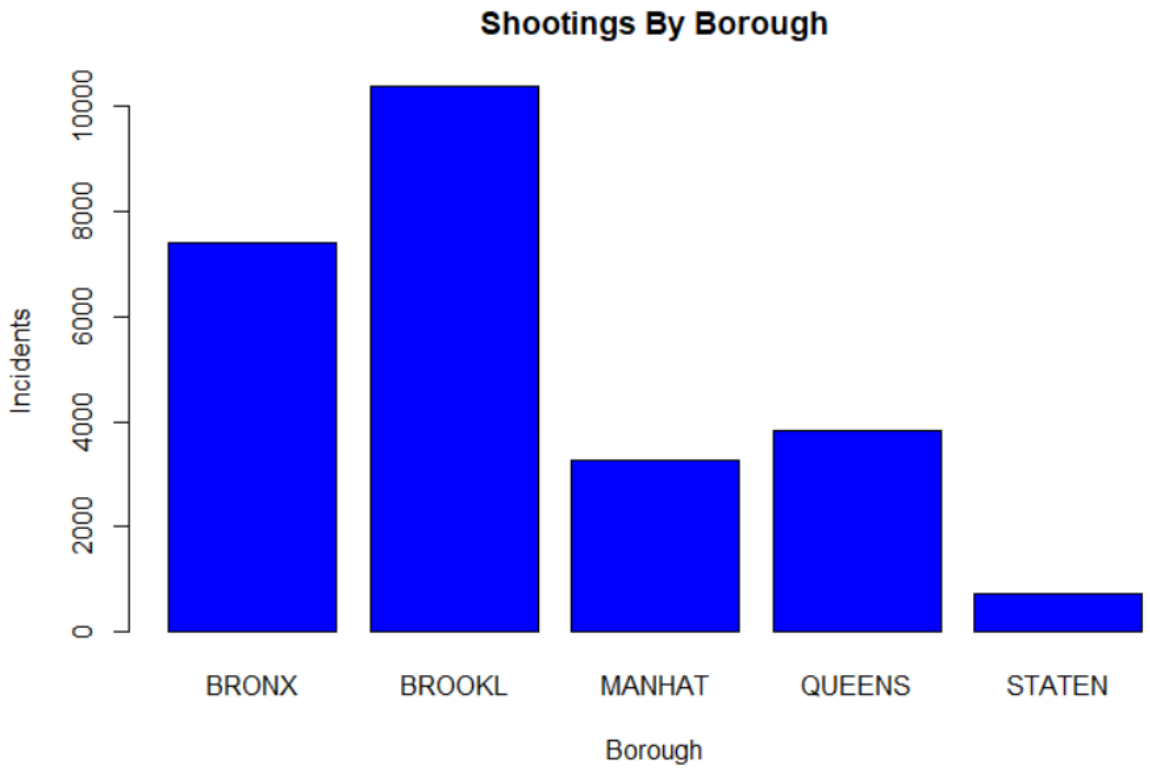
```
\section{table}

# A tibble: 5 × 2
  Borough           n
  <chr>         <int>
1 BRONX          7402
2 BROOKLYN      10365
3 MANHATTAN      3265
4 QUEENS         3828
5 STATEN ISLAND   736
```

a <- nypd_Borough %>% mutate(across(everything(), str_sub, 1,6))

a <- a %>% mutate(n=as.numeric(n))

barplot(a$n, names = a$Borough, xlab=“Borough”, ylab=“Incidents”, main=“Shootings By Borough”, col=“blue”)



Analysis Conclusion: Brookl has the highest shooting incidents

nypd_Vic_Age <- nypd %>% count(Vic_Age)

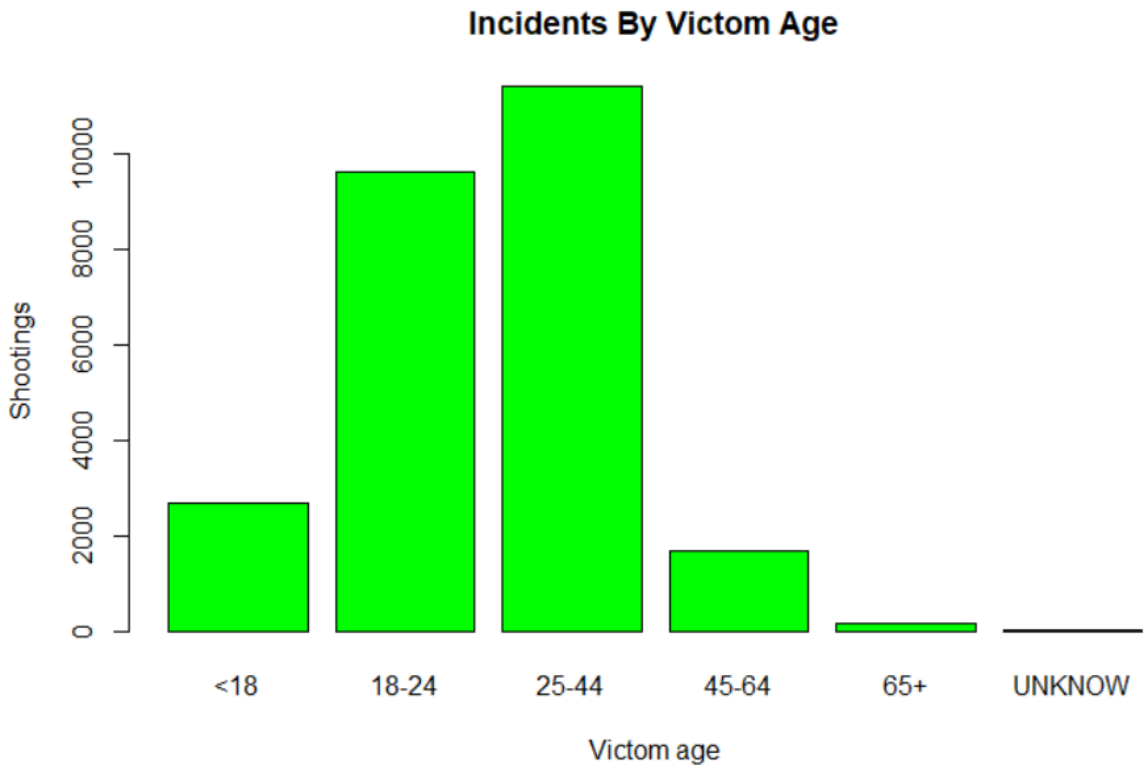
nypd_Vic_Age

```
\section{table}
A tibble: 6 × 2
  Vic_Age     n
  <chr>   <int>
1 <18      2681
2 18-24    9604
3 25-44   11386
4 45-64    1698
5 65+       167
6 UNKNOWN    60
```

b <- nypd_Vic_Age %>% mutate(across(everything(), str_sub, 1,6))

b <- b %>% mutate(n=as.numeric(n))

barplot(b$n, $names = b$Vic_Age, xlab="Victom age", ylab="Shootings", main="Incidents By Victom Age", col="green")



Analysis Conclusion: the Age Group 25-44 has the highest shooting incidents

Step 4: Add Bias Identification

Write the conclusion to your project report and include any possible sources of bias. Be sure to identify what your personal bias might be and how you have mitigated that.

This data set has much amount of work to manipulate with, and I did only very basic analysis, also I would like to mention that this project was the first project I make analysis in R , and I am sure that It will be the first step forward to me for being a Data Scientist. To avoid bias I did not like to analyze the data in terms of race.