NYPD Shooting Incident Data

Michiel Schinkel

24-11-2021

Loading in the data

Using the code below, we will read in the NYPD Shooting Incident Data from the city of New York for further analysis. We will use a direct link to the dataset to aid reproducibility.

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

Tidying the data

With the code below, we will tidy the data. Using `length(unique(as.factor(data$INCIDENT_KEY)))`, we noticed that there were just 18562 unique incidents, while there are 23568 records. We therefore started cleaning the data by removing duplicates. In the next step, we removed all columns that would not be used in the final analysis, and added **UNKNOWN** or **U** to missing values. Finally, we set the correct column types.

```
tidy_data <- data %>%
  distinct(INCIDENT_KEY, .keep_all = TRUE) %>%
  select(-c("JURISDICTION_CODE",  "X_COORD_CD", "Y_COORD_CD", "Latitude", "Longitude", "Lon_Lat")) %>%
  mutate(LOCATION_DESC = fct_recode(LOCATION_DESC, "UNKNOWN" = ""),
         PERP_SEX = fct_recode(PERP_SEX, "U" = ""),
         PERP_RACE = fct_recode(PERP_RACE, "UNKNOWN" = ""),
         PERP_AGE_GROUP = replace(PERP_AGE_GROUP, PERP_AGE_GROUP %in% c("", "1020", "224", "940"), "UNKN
         OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         PRECINCT = as.factor(PRECINCT))
print(summary(tidy_data))
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 57282440   1st Qu.:2009-02-12   1st Qu.:3H 21M 0S
##  Median : 84353591   Median :2012-04-22   Median :14H 59M 0S
##  Mean   :103140450   Mean   :2012-10-30   Mean   :12H 34M 43.1060116354201S
##  3rd Qu.:152014878   3rd Qu.:2016-04-10   3rd Qu.:20H 45M 0S
##  Max.   :230611229   Max.   :2020-12-31   Max.   :23H 59M 0S
##
##             BORO         PRECINCT                         LOCATION_DESC
##  BRONX        :5103   75     : 1080   UNKNOWN                  :10867
##  BROOKLYN     :7838   73     : 1029   MULTI DWELL - PUBLIC HOUS: 3401
##  MANHATTAN    :2274   67     :  906   MULTI DWELL - APT BUILD  : 1926
##  QUEENS       :2795   79     :  729   PVT HOUSE                :  617
```

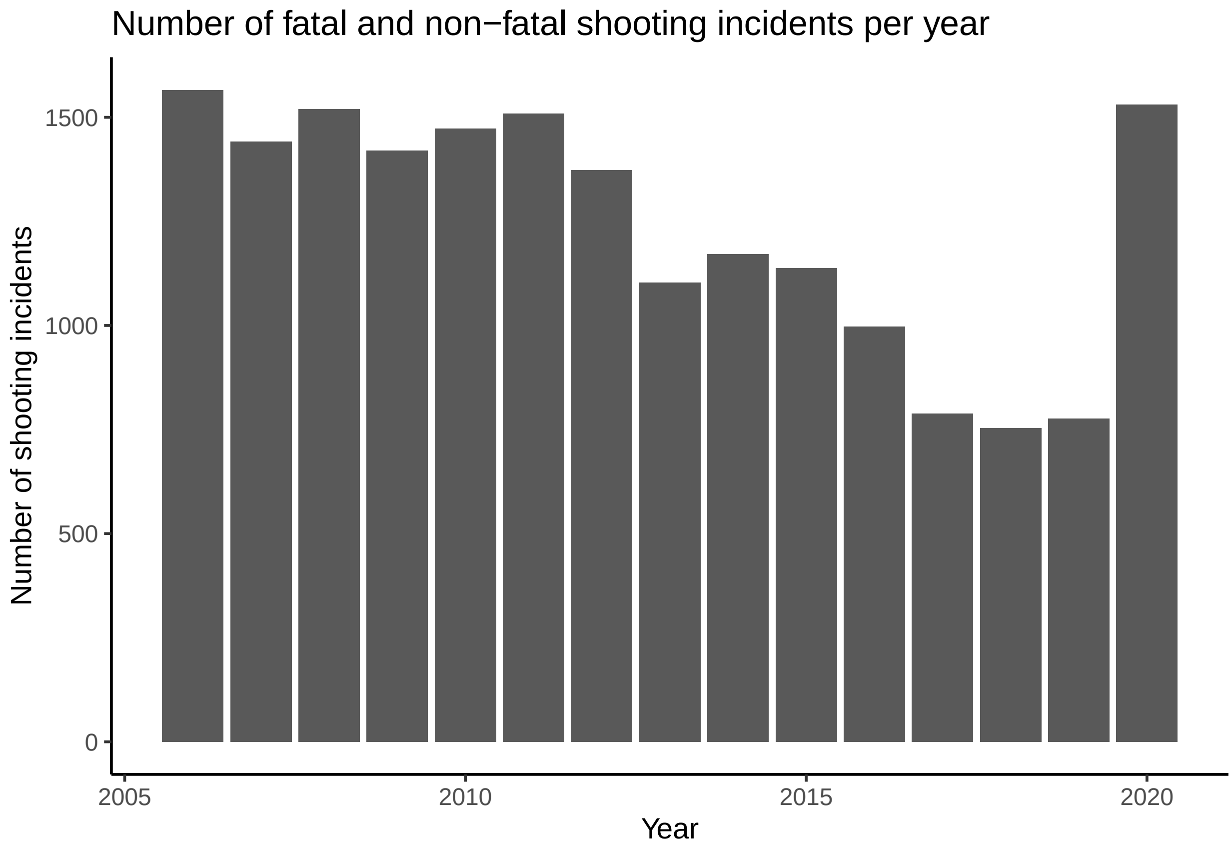
```
##  STATEN ISLAND: 554   47      :  647   GROCERY/BODEGA           :  431
##                       44      :  624   BAR/NIGHT CLUB           :  389
##                       (Other):13549   (Other)                  :  933
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##  false:15325             UNKNOWN:9951   U:8537
##  true : 3239             18-24  :3871   F: 190
##                          25-44  :3429   M:9837
##                          <18    : 927
##                          45-64  : 345
##                          65+    :  41
##                          (Other):   0
##                           PERP_RACE     VIC_AGE_GROUP  VIC_SEX
##  UNKNOWN                       :8809   <18    :1865   F: 1399
##  AMERICAN INDIAN/ALASKAN NATIVE:   2   18-24  :7129   M:17159
##  ASIAN / PACIFIC ISLANDER      :  76   25-44  :8264   U:    6
##  BLACK                         :7417   45-64  :1154
##  BLACK HISPANIC                : 739   65+    : 116
##  WHITE                         : 200   UNKNOWN:  36
##  WHITE HISPANIC                :1321
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    8
##  ASIAN / PACIFIC ISLANDER      :  241
##  BLACK                         :13601
##  BLACK HISPANIC                : 1686
##  UNKNOWN                       :   50
##  WHITE                         :  488
##  WHITE HISPANIC                : 2490
```

Visualizing and analyzing the data

With this project, we aim to study the factors that may influence the fatality rates of shooting incidents (STATISTICAL_MURDER_FLAG == TRUE). To understand the data, we first visualize several aspects.

We start of the exploratory analysis by looking at the shooting incidents over time since the beginning of the data collection.

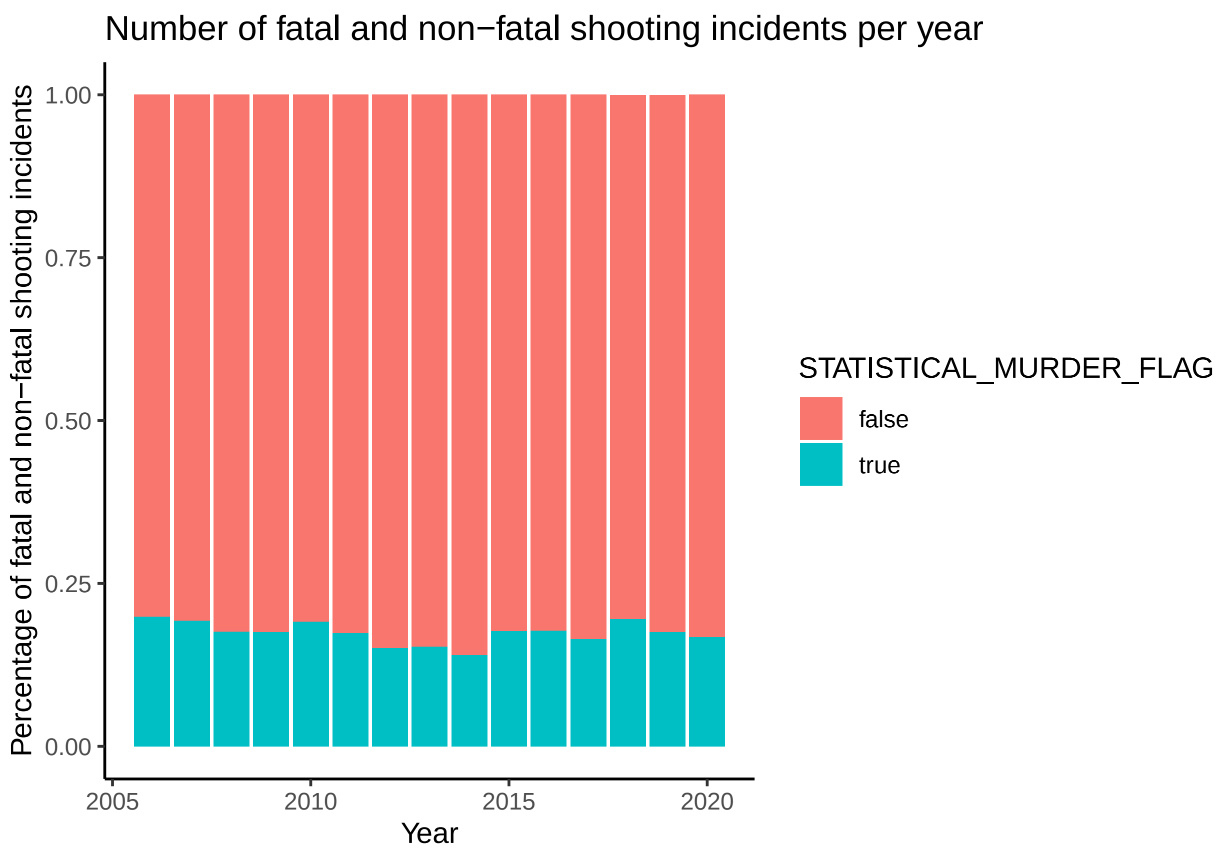
```r
ggplot(data = tidy_data, aes(x = year(OCCUR_DATE))) +
geom_bar(stat = "count") +
xlab("Year") +
ylab("Number of shooting incidents") +
ggtitle("Number of fatal and non-fatal shooting incidents per year") +
theme_classic()
```



From this visualization, we learn that the number of shooting incidents had been decreasing until 2019. The COVID-19 pandemic has suddenly brought us back to where we were in 2005. Also notable, the numbers of shooting incidents per year range from about 700 to 1500, indicating that there are between 2-4 shooting incidents in New York each day.

In the next visualization, we look at the percentage of fatal incidents per year.

```
ggplot(data = tidy_data, aes(x = year(OCCUR_DATE), fill=STATISTICAL_MURDER_FLAG)) +
geom_bar(stat = "count", position="fill") +
xlab("Year") +
ylab("Percentage of fatal and non-fatal shooting incidents") +
ggtitle("Number of fatal and non-fatal shooting incidents per year") +
theme_classic()
```

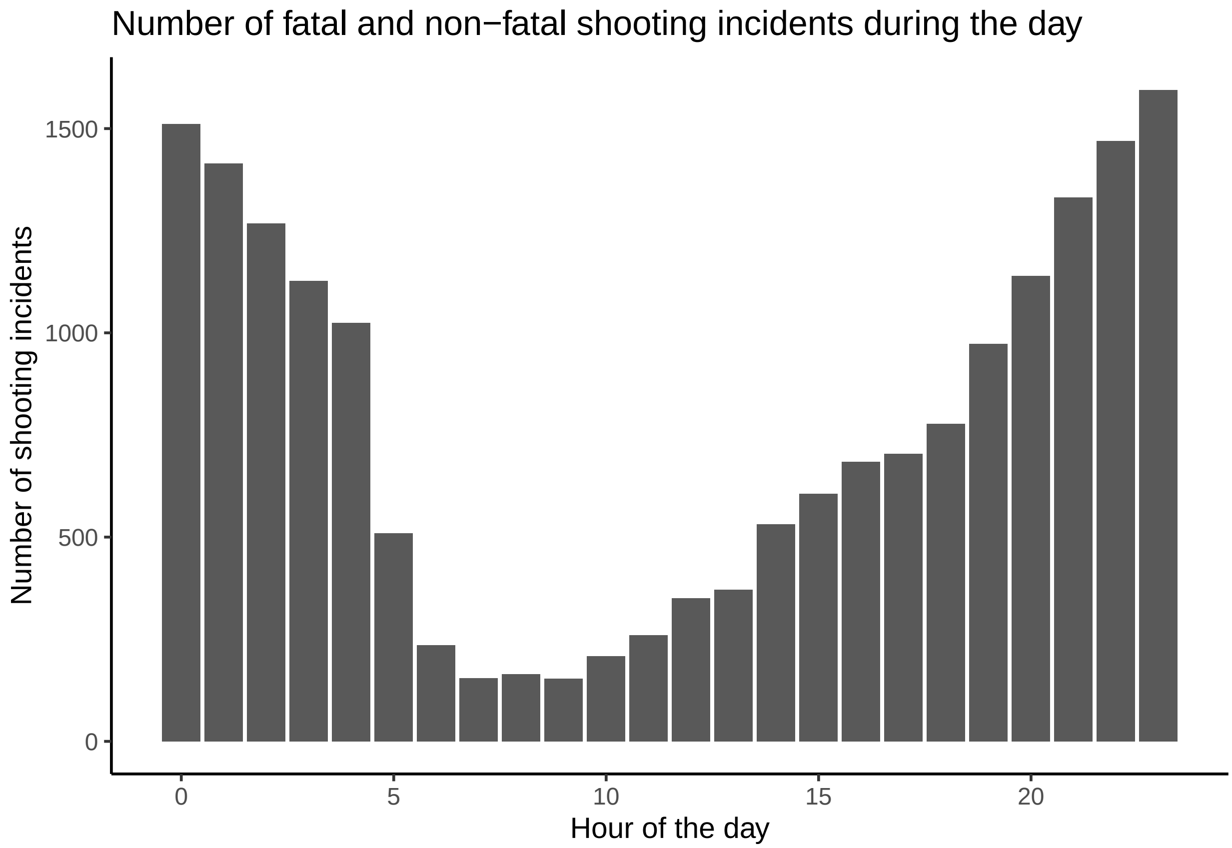


From this visualization, we learn that the percentage of fatal incidents has remained rather similar, even with the overall decreases in shooting incidents over time.

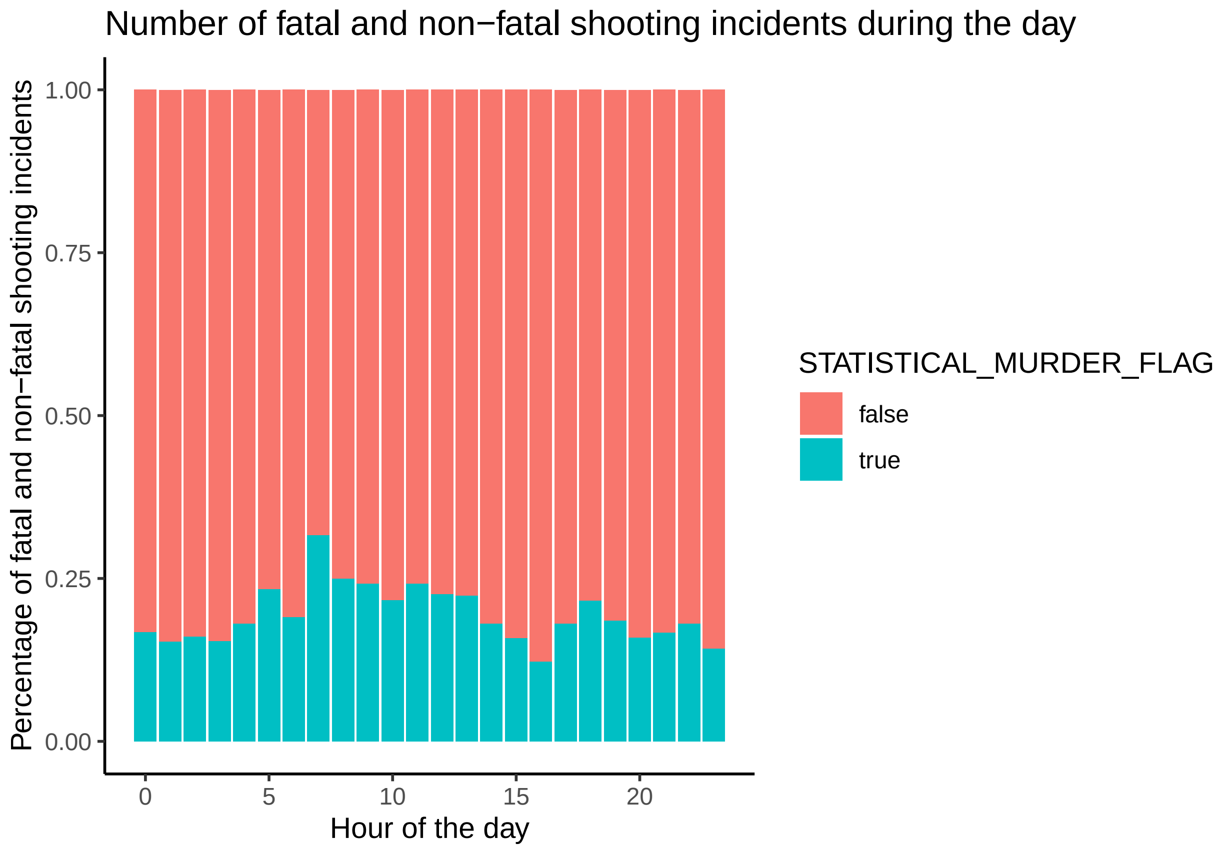
Daily patterns in shooting incidents

Next, we zoom on on daily patterns of the shooting incidens. We then also plot the percentage of fatal and non-fatal shooting incidents during all hours of the day.

```
ggplot(data = tidy_data, aes(x = hour(OCCUR_TIME))) +
geom_bar(stat = "count") +
xlab("Hour of the day") +
ylab("Number of shooting incidents") +
ggtitle("Number of fatal and non-fatal shooting incidents during the day") +
theme_classic()
```




```r
ggplot(data = tidy_data, aes(x = hour(OCCUR_TIME), fill=STATISTICAL_MURDER_FLAG)) +
geom_bar(stat = "count", position="fill") +
xlab("Hour of the day") +
ylab("Percentage of fatal and non-fatal shooting incidents") +
ggtitle("Number of fatal and non-fatal shooting incidents during the day") +
theme_classic()
```



From the visualizations, it seems that there is a clear distribution of the shooting incidents during every 24-hour period. Most shooting incidents happen at night, while few happen between 6am and 11am. In the second visualization, we can see that the percentage of fatal shooting incidents does not seem to be influenced by the timing of the incident.

We further set out to study factors other than timing during the day for their influence on the fatality rates of shooting incidents. We created a logistic regression model to predict the status of the STATISTICAL_MURDER_FLAG based on the sex and age of the perpetrator, as well as the boro in which they happened:

```
mod <- glm(STATISTICAL_MURDER_FLAG ~ PERP_SEX + PERP_AGE_GROUP + BORO, data = tidy_data, family="binomi
exp(coef(mod))
```

```
##          (Intercept)            PERP_SEXF            PERP_SEXM
##            1.6097701            0.1334726            0.1081277
##    PERP_AGE_GROUP18-24    PERP_AGE_GROUP25-44    PERP_AGE_GROUP45-64
##            1.2521375            1.7875320            2.5874861
##     PERP_AGE_GROUP65+ PERP_AGE_GROUPUNKNOWN          BOROBROOKLYN
##            3.4230110            0.1080382            1.1468474
##        BOROMANHATTAN           BOROQUEENS     BOROSTATEN ISLAND
##            0.9404828            1.0959838            1.0195619
```

```
exp(confint(mod))
```

```
## Waiting for profiling to be done...
```

```
##                           2.5 %    97.5 %
## (Intercept)           1.09792206 2.4197693
## PERP_SEXF             0.08135805 0.2133613
## PERP_SEXM             0.07499333 0.1506985
## PERP_AGE_GROUP18-24   1.03371297 1.5253007
## PERP_AGE_GROUP25-44   1.47763296 2.1749796
## PERP_AGE_GROUP45-64   1.94112598 3.4456828
## PERP_AGE_GROUP65+     1.75020777 6.5196985
## PERP_AGE_GROUPUNKNOWN 0.07244241 0.1569578
## BOROBROOKLYN          1.04320321 1.2614042
## BOROMANHATTAN         0.82011589 1.0767924
## BOROQUEENS            0.96853107 1.2392375
## BOROSTATEN ISLAND     0.80798416 1.2760665
```

```r
summary(mod)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_SEX + PERP_AGE_GROUP +
##     BORO, family = "binomial", data = tidy_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7080  -0.6335  -0.6031  -0.2066   2.8469
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.47609    0.20106   2.368  0.01789 *
## PERP_SEXF             -2.01386    0.24560  -8.200 2.41e-16 ***
## PERP_SEXM             -2.22444    0.17753 -12.530  < 2e-16 ***
## PERP_AGE_GROUP18-24    0.22485    0.09918   2.267  0.02339 *
## PERP_AGE_GROUP25-44    0.58084    0.09856   5.894 3.78e-09 ***
## PERP_AGE_GROUP45-64    0.95069    0.14628   6.499 8.08e-11 ***
## PERP_AGE_GROUP65+      1.23052    0.33306   3.695  0.00022 ***
## PERP_AGE_GROUPUNKNOWN -2.22527    0.19669 -11.314  < 2e-16 ***
## BOROBROOKLYN           0.13702    0.04845   2.828  0.00468 **
## BOROMANHATTAN         -0.06136    0.06945  -0.884  0.37693
## BOROQUEENS             0.09165    0.06287   1.458  0.14486
## BOROSTATEN ISLAND      0.01937    0.11647   0.166  0.86790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17187  on 18563  degrees of freedom
## Residual deviance: 16580  on 18552  degrees of freedom
## AIC: 16604
##
## Number of Fisher Scoring iterations: 6
```

From this data we can learn various things. First of all, we see that sex and age of the perpetrator have a significant association with the chance of a fatal incident. To start with sex, we can see that both male and female perpetrators have lower risks of a fatal accident as opposed to the "UNKNOWN" reference groups.

We will further discuss the possible reasons for this is the next paragraph on bias identification. We also see that various age groups of perpetrators are significantly associated with higher fatality rates. Compared with the reference group of perpetrators under the age of 18, all age groups above 24 are significantly associated with higher mortality rates. Finally, we see that the location of the shooting has a far smaller influence on the fatality rate, although shooting incidents in Brooklyn are significantly associated with a slightly higher odds ratio for fatality.

Bias identification

From the logistic regression analysis, we learned that being either a female or male perpetrator were both associated with lower fatality rates compared with the "UNKNOWN" group. The bias here may be that it is more difficult to identify the perpetrator of a fatal shooting incident, since the victim cannot identify the person in question. Therefore, the percentage of UNKNOWN sex labels in the fatal shooting incident may be much higher. Interestingly, in case of the perpetrators age, the UNKNOWN label is associated with a lower mortality rate. This contradicts the first hypothesis, which should thus be carefully evaluated in future analyses.

Conclusion

In conclusion, we looked at factors that are associated with fatal shooting incidents in New York. We learned that most shootings happen in the evening and night, while few happen in the morning. However, the fatality rates of the shooting do not seem to be influenced by the time of the day. We did find that perpetrators in older age categories have a significantly higher odds ratio for a fatal shooting incident. The same is true for shooting incidents that happen in Brooklyn. Lastly, we found that shooting incidents by perpetrators of whom the sex is known, are associated with a much lower fatality rate than when the sex is unknown. This final observation may be caused by a selection bias and needs to be studied further in future analyses.

Session info

To ensure this work is reproducible, we here add the session info.

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Dutch_Netherlands.1252  LC_CTYPE=Dutch_Netherlands.1252
## [3] LC_MONETARY=Dutch_Netherlands.1252 LC_NUMERIC=C
## [5] LC_TIME=Dutch_Netherlands.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.7.10 knitr_1.31       magrittr_2.0.1   forcats_0.5.1
##  [5] stringr_1.4.0    dplyr_1.0.5      purrr_0.3.4      readr_1.4.0
##  [9] tidyr_1.1.3      tibble_3.1.1     ggplot2_3.3.5    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
```

```
##  [1] tidyselect_1.1.0  xfun_0.22         haven_2.4.1       colorspace_2.0-0
##  [5] vctrs_0.3.8       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.1.4        rlang_0.4.11      pillar_1.6.4      glue_1.4.2
## [13] withr_2.4.1       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     labeling_0.4.2
## [25] fansi_0.4.2       highr_0.8         broom_0.7.6       Rcpp_1.0.6
## [29] scales_1.1.1      backports_1.1.10  jsonlite_1.7.2    farver_2.1.0
## [33] fs_1.5.0          hms_1.0.0         digest_0.6.27     stringi_1.5.3
## [37] grid_3.6.3        cli_2.5.0         tools_3.6.3       crayon_1.4.1
## [41] pkgconfig_2.0.3   MASS_7.3-51.5     ellipsis_0.3.2    xml2_1.3.2
## [45] reprex_2.0.1      assertthat_0.2.1  rmarkdown_2.7     httr_1.4.2
## [49] rstudioapi_0.13   R6_2.5.0          compiler_3.6.3
```