NYPD Shooting Data Analysis

Bojan Jovanović

24 january 2024

Prerequisites

To be able to successfully complete our analysis, we have to include some R libraries we will need in the process. Those include **tinytex, tidyverse, dplyr and ggplot2**.

```
library("tinytex")
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library("dplyr")
library("ggplot2")
```

Importing NYPD Shooting Incident Data

Before we continue with our analysis, we have to import the data first. In this report, we will be analyzing "*NYPD Shooting Incident Data (Historic)*", which is provided to the public by the city of New York. The dataset includes **data about shootings in New York** from 2006, all the way up until the end of the last calendar year. The dataset was last updated on 27th of April 2023, meaning the last included year is 2022.

More info about this dataset can be found at this link.

The dataset includes information about shooting event, location and time, as well as demographic information about the suspect and the victim. Each shooting event is stored in a separate row of a table.

```
## NOTE: IF ERROR HAPPENS IN THIS CHUNK ON FIRST RUN, RUN IT ONCE AGAIN!
## FOR SOME WEIRD REASON, DATA IMPORT CAN FAIL ON FIRST RUN
## HOWEVER, IT SUCCEDD ON SUBSEQUENT ATTEMPTS
data_loc <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read_csv(data_loc)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

Tidying Our Data

Unnecessary Variables

The dataset we have loaded consists of 21 variables related to shooting incidents. However, we won't need all of those for our analysis. So, we will **select only those columns we will actually use** in further analysis.

From the original data, we will select the following columns: *INCIDENT_KEY, STATISTICAL_MURDER_FLAG, VIC_AGE_GROUP, VIC_SEX, VIC_RACE*. Remaining columns won't be needed in our following analysis, so we will not include them in our cleaned dataset.

Formatting and Missing Values

In our data, victim sex is indicated as either M of F (Male and Female). To make our data a bit more user-friendly, we will convert those values into Male and Female.

Also, for our further analysis, we will convert categorical variables *VIC_AGE_GROUP, VIC_SEX* and *VIC_RACE* into factors. We will also convert *STATISTICAL_MURDER_FLAG* into integer values 0 and 1.

We will also **eliminate records with missing data**. We have checked if there are any NA values in our columns. There are no NAs. However, by taking a closer look at our data, we can see that missing data is coded by word "UNKNOWN", instead of NA. Thus, we will remove any row containing "UNKNOWN" in any of its columns.

Removing Nonsensical/Insufficient Data

Also, by analyzing values in our column, we have found some nonsensical data. For example, we have an **age group labeled "1022"**, which makes no sense. We will remove this single row.

"AMERICAN INDIAN/ALASKAN NATIVE" race members were identified as victims in 10 cases only. Compared to a total of 27312 records, this is a very small numbers. As we worry this might skew our analysis and make it inaccurate, we will remove those records.

```
## FILTER OUT COLUMNS
nypd_data_tidy <- nypd_data %>% select(INCIDENT_KEY, STATISTICAL_MURDER_FLAG,
                    VIC_AGE_GROUP, VIC_SEX, VIC_RACE)

## ELIMINATE ROWS NOT CORRESPONDING TO DEATHS
nypd_data_tidy <- nypd_data_tidy %>% mutate(VIC_SEX=ifelse(VIC_SEX=="M","Male","Female"))
```

```r
## CHECK FOR NAs
sum(is.na(nypd_data_tidy$STATISTICAL_MURDER_FLAG))
```

```
## [1] 0
```

```r
sum(is.na(nypd_data_tidy$VIC_AGE_GROUP))
```

```
## [1] 0
```

```r
sum(is.na(nypd_data_tidy$VIC_SEX))
```

```
## [1] 0
```

```r
sum(is.na(nypd_data_tidy$VIC_RACE))
```

```
## [1] 0
```

```r
## ANALYZE VALUES IN COLUMNS
nypd_data_tidy %>% group_by(STATISTICAL_MURDER_FLAG) %>% count(STATISTICAL_MURDER_FLAG)
```

```
## # A tibble: 2 x 2
## # Groups:   STATISTICAL_MURDER_FLAG [2]
##   STATISTICAL_MURDER_FLAG     n
##   <lgl>                   <int>
## 1 FALSE                   22046
## 2 TRUE                     5266
```

```r
nypd_data_tidy %>% group_by(VIC_AGE_GROUP) %>% count(VIC_AGE_GROUP)
```

```
## # A tibble: 7 x 2
## # Groups:   VIC_AGE_GROUP [7]
##   VIC_AGE_GROUP     n
##   <chr>         <int>
## 1 1022              1
## 2 18-24         10086
## 3 25-44         12281
## 4 45-64          1863
## 5 65+             181
## 6 <18            2839
## 7 UNKNOWN          61
```

```r
nypd_data_tidy %>% group_by(VIC_SEX) %>% count(VIC_SEX)
```

```
## # A tibble: 2 x 2
## # Groups:   VIC_SEX [2]
##   VIC_SEX     n
##   <chr>   <int>
## 1 Female   2626
## 2 Male    24686
```

```
nypd_data_tidy %>% group_by(VIC_RACE) %>% count(VIC_RACE)
```

```
## # A tibble: 7 x 2
## # Groups:   VIC_RACE [7]
##   VIC_RACE                           n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE    10
## 2 ASIAN / PACIFIC ISLANDER         404
## 3 BLACK                          19439
## 4 BLACK HISPANIC                  2646
## 5 UNKNOWN                           66
## 6 WHITE                            698
## 7 WHITE HISPANIC                  4049
```

```
## REMOVE ROWS WITH UNKNOWN DATA AND AMERICAN INDIAN/ALASKAN NATIVE
nypd_data_tidy <- nypd_data_tidy %>% filter(VIC_AGE_GROUP!="UNKNOWN" & VIC_RACE!="UNKNOWN"
                    & VIC_RACE!="AMERICAN INDIAN/ALASKAN NATIVE")

## REMOVE ROW WITH WEIRD AGE DATA
nypd_data_tidy <- nypd_data_tidy %>% filter(VIC_AGE_GROUP!="1022")

## CONVERT COLUMNS TO FACTORS AND INTEGER
nypd_data_tidy$STATISTICAL_MURDER_FLAG <- as.integer(nypd_data_tidy$STATISTICAL_MURDER_FLAG)
nypd_data_tidy$VIC_AGE_GROUP <- as.factor(nypd_data_tidy$VIC_AGE_GROUP)
nypd_data_tidy$VIC_SEX <- as.factor(nypd_data_tidy$VIC_SEX)
nypd_data_tidy$VIC_RACE <- as.factor(nypd_data_tidy$VIC_RACE)

summary(nypd_data_tidy)
```

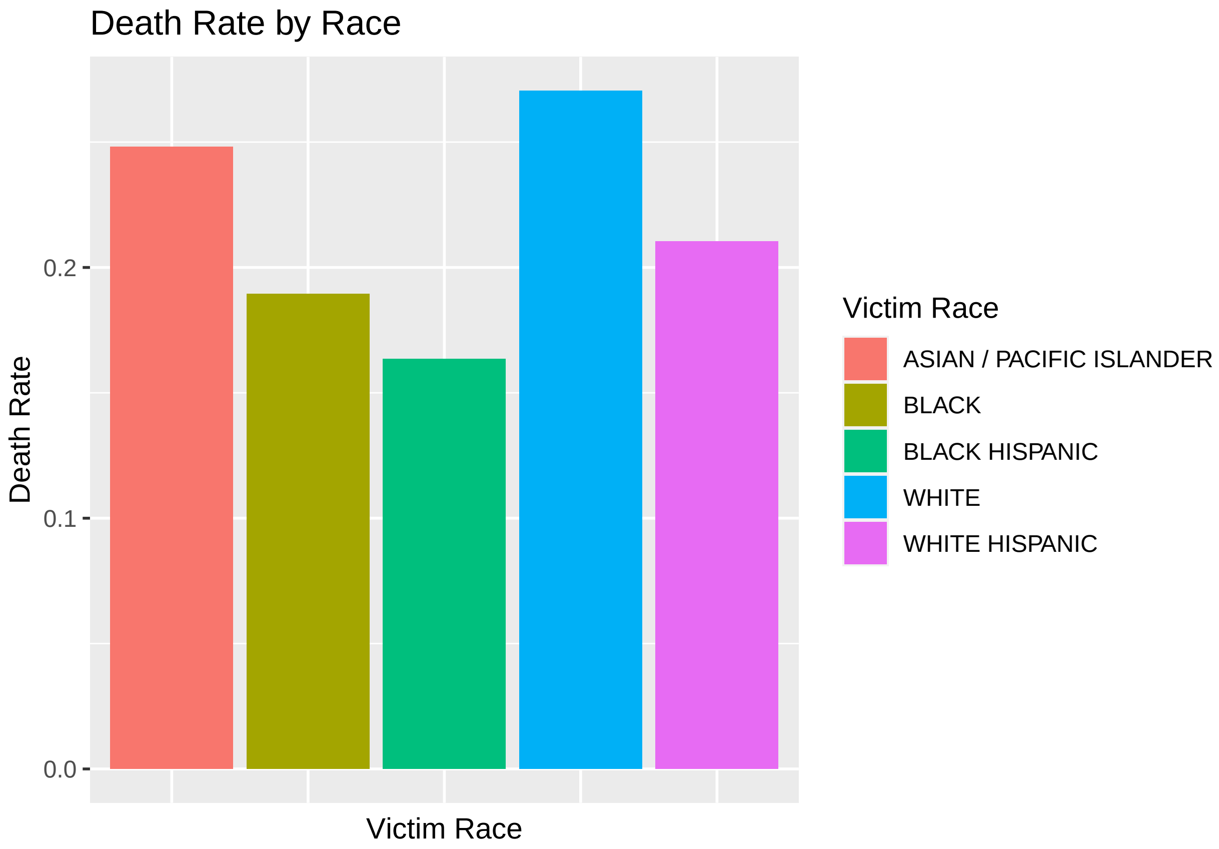
```
##   INCIDENT_KEY       STATISTICAL_MURDER_FLAG VIC_AGE_GROUP   VIC_SEX
##  Min.   :  9953245   Min.   :0.0000          <18   : 2833   Female: 2612
##  1st Qu.: 63859932   1st Qu.:0.0000          18-24:10061   Male  :24578
##  Median : 90451951   Median :0.0000          25-44:12257
##  Mean   :120945186   Mean   :0.1929          45-64: 1858
##  3rd Qu.:189136471   3rd Qu.:0.0000          65+  :  181
##  Max.   :261190187   Max.   :1.0000
##                         VIC_RACE
##  ASIAN / PACIFIC ISLANDER:  403
##  BLACK                   :19420
##  BLACK HISPANIC          : 2642
##  WHITE                   :  684
##  WHITE HISPANIC          : 4041
##
```

Data Analysis

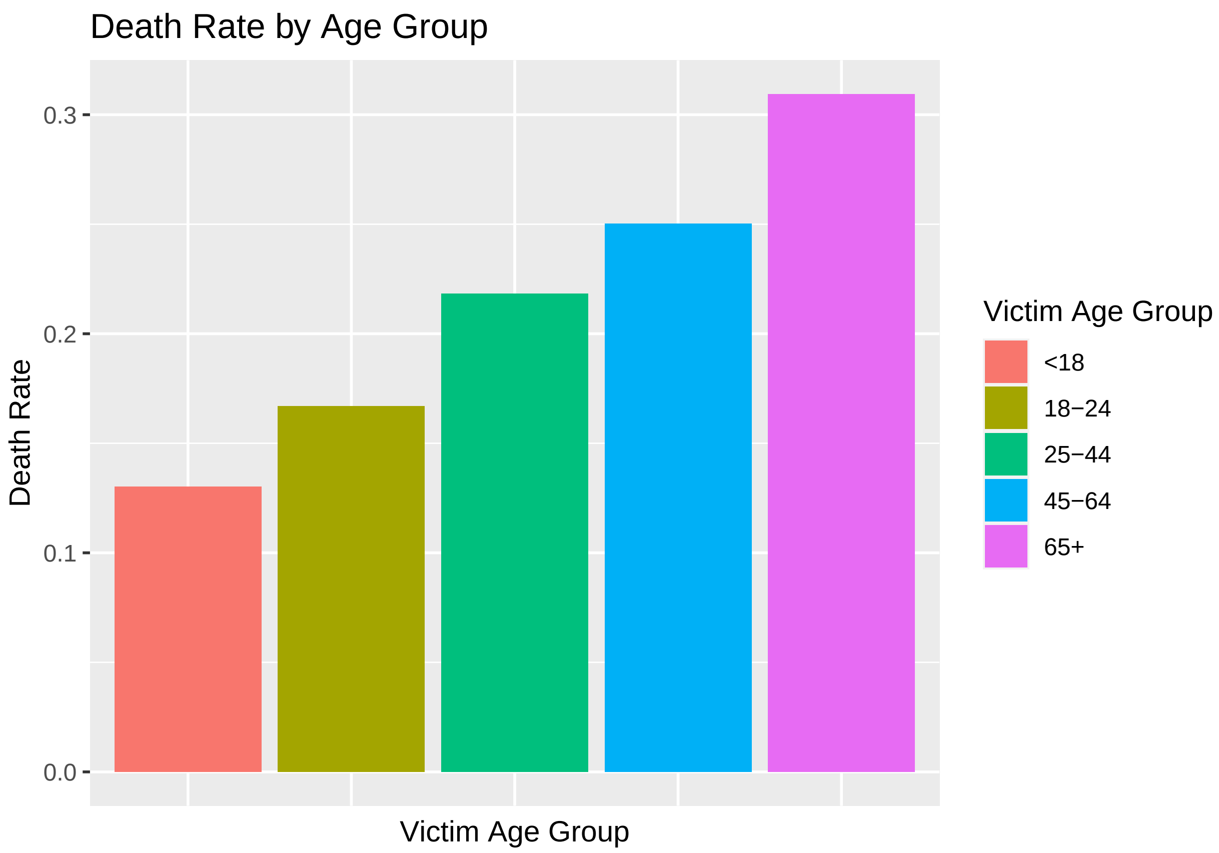
In this part of our report, we will analyze our prepared data. Our main research question will be **whether demographics affect one's likelihood to die in a case of being shot**.

First, we will create some boxplots to roughly visualize if there are differences in death rate between the races, age groups and sexes. Then, we will create a logistic regression model in the effort to better understand the factors affecting the likelihood of death during a shooting event.

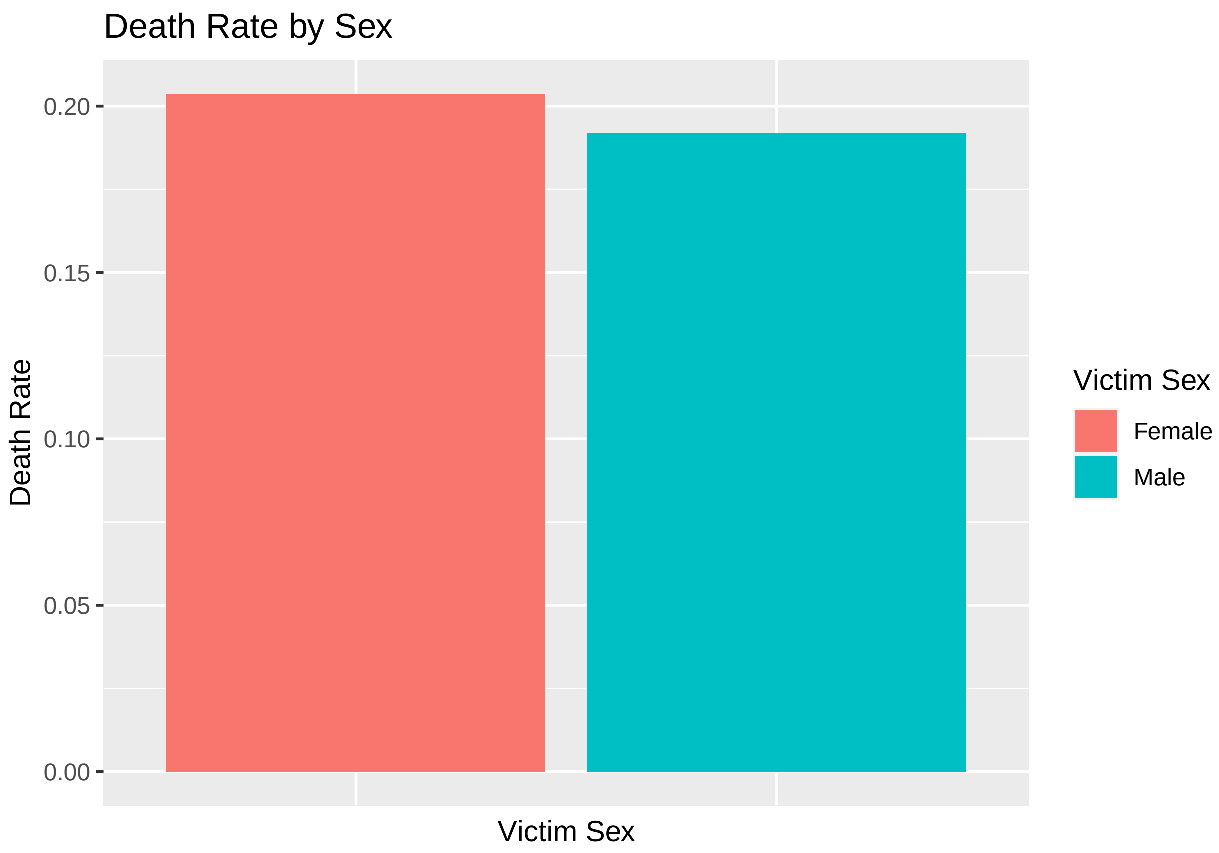
```r
## PLOT DEATH RATE AGAINT THE RACE
ggplot(data=nypd_data_tidy) +
  stat_summary(aes(x=VIC_RACE, y=STATISTICAL_MURDER_FLAG, fill=VIC_RACE),
               geom="bar", fun="mean") +
  ggtitle("Death Rate by Race") +
  xlab("Victim Race") +
  ylab("Death Rate") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  labs(fill="Victim Race")
```



```r
##PLOT DEATH RATE AGAINST THE AGE GROUP
ggplot(data=nypd_data_tidy) +
  stat_summary(aes(x=VIC_AGE_GROUP, y=STATISTICAL_MURDER_FLAG, fill=VIC_AGE_GROUP),
               geom="bar", fun="mean") +
  ggtitle("Death Rate by Age Group") +
  xlab("Victim Age Group") +
  ylab("Death Rate") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  labs(fill="Victim Age Group")
```



```
##PLOT DEATH RATE AGAINST THE SEX
ggplot(data=nypd_data_tidy) +
  stat_summary(aes(x=VIC_SEX, y=STATISTICAL_MURDER_FLAG, fill=VIC_SEX),
               geom="bar", fun="mean") +
  ggtitle("Death Rate by Sex") +
  xlab("Victim Sex") +
  ylab("Death Rate") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
  labs(fill="Victim Sex")
```



Analyzing out Plots

From the plots we have created, we can clearly see that **there is a difference in death rates between different races of victims**. It sseems like black hispanic victims have the greatest chances of survival, while white victims fared the worst.

We have also comfirmed something we already believed to be true. Namely, **the older a victim is, the greater the chances of dying are**. As we know our bodies are getting weaker with the age, this finding is absolutely to be expected.

One interesting finding is that **male and female victims have roughly the same chances of surviving a shooting**. I honestly expected men to handle shootings better, due to larger bodies and higher amounts of blood in them, which could be critical for a survival. However, I wasn't right.

Our data told us there are differences in death rate between different races. However, it didn't tell us why this might be the case and if race itself caused this difference, or some other factors. It is possible that white victims tend to be older than black victims, so their age actually increases their death rate instead of their race.

To analyze this issue further, we will **fit a logistic regression model on our data**. It will have *STATISTICAL_MURDER_FLAG* as the outcome, and *VIC_RACE* and *VIC_AGE_GROUP* as two predictors.

```
## FITTING LOGISTIC REGRESSION
model <- glm(formula=STATISTICAL_MURDER_FLAG~VIC_RACE+VIC_AGE_GROUP,
             family="binomial", data=nypd_data_tidy)
summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_RACE + VIC_AGE_GROUP,
##     family = "binomial", data = nypd_data_tidy)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.63576    0.12889 -12.691  < 2e-16 ***
## VIC_RACEBLACK           -0.26823    0.11737  -2.285 0.022293 *
## VIC_RACEBLACK HISPANIC -0.44694    0.12736  -3.509 0.000449 ***
## VIC_RACEWHITE            0.06298    0.14470   0.435 0.663392
## VIC_RACEWHITE HISPANIC -0.14711    0.12220  -1.204 0.228646
## VIC_AGE_GROUP18-24       0.28512    0.06193   4.604 4.15e-06 ***
## VIC_AGE_GROUP25-44       0.61130    0.06002  10.185  < 2e-16 ***
## VIC_AGE_GROUP45-64       0.76651    0.07785   9.846  < 2e-16 ***
## VIC_AGE_GROUP65+         1.03162    0.17125   6.024 1.70e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26668  on 27189  degrees of freedom
## Residual deviance: 26405  on 27181  degrees of freedom
## AIC: 26423
##
## Number of Fisher Scoring iterations: 4
```

```r
## GET PVALUE FOR RESIDUAL DEVIANCES
1 - pchisq(summary(model)$deviance, summary(model)$df[2])
```

```
## [1] 0.9996052
```

We fitted our model as described previously. However, after we fit any model, we should diagnose how well it fitted our data. To do so, we will calculate p-value for our Resudual Deviance. From the theory, we know that it follows Chi Squared distribution with degrees of freedom equal to number of data points minus the number of parameter.

As we can see when we run our code, we have a very high p-value. This is indicative of **a model which successfully fits the data**. Thus, we will proceed with interpretation of our results.

Analysis Conclusions

Just as we already saw on our plots, our model confirmed that **increased age also increases your chances of dying** from a shooting wound. There are statistically significant differences between the base case (under 18) and every other age group.

Things are a bit more interesting when it comes to racial differences. From what we see in our model summary, it seems like **black and black hispanic persons have better chances of survival** than Asian / Pacific, White or White Hispanic person.

We see this by statistically significant negative coefficients for *VIC_RACEBLACK* and *VIC_RACEBLACK_HISPANIC*, indicating log odds of dying if you get shot decrease for Black and Black Hispanic people compared to base case (Asian / Pacific).

Our model didn't identify statistically significant difference between the base case and White or White Hispanic people, indicating a possibility of no differences between those groups.

Bias Identification

Data Reporting Bias

The most obvious bias for me is the **bias related to data reporting**. We don't know much about how shooting incidents were reported, so we have to recognize the possibility of bias there.

For example, white people in New York have historically been richer and better connected than black people. Thus, it is usually easier for a white offender to bribe police officers or call his connections for protection than a black person. Thus, policemen might decide not to report a shooting if it involves a person with high status, skewing our data.

I hope such dishonest work of police is very rare. However, we have absolutely heard of such cases in media, so we have to recognize this bias possibility.

Undesigned Experiment Bias

When performing an research, the best way to go about it is to properly design a study before going out to collect the exact data we need. However, this wasn't the case in this report.

We were given this dataset and we were forced to do our best with what we have. Thus, **it is possible we missed many variables which could have affected death rate**. For example, we don't have information about physical fitness of victims, which is a variable that could affect someone's likelihood of surviving if being shot.

It is possible that black people have better fitness than white people, so their fitness actually make them more likely to survive instead of their race.

Personal Biases

Previously, I have mentioned better fitness of black people as possible factor for their lower death rate. This idea, however, might be a product of my personal biases.

In my country, it is **a common stereotype that black people are physically dominant** over the white race. For such reason, this was the first idea that came to my mind when thinking about what might make black people more likely to survive if they get shot.

To mitigate this bias, **I refrain from making claims without significant statistical proof**. I did mention this factor of superior fitness as a possibility. However, I never claimed it is actually what makes black people more likely to survive, as I don't have the data to back up my claim.

Session Info

To make this report reproducible, the configuration used will be shown below. This list includes R version, Operating system, locale and packages in use.

```r
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
```

```
##
## locale:
## [1] LC_COLLATE=Serbian (Latin)_Serbia.utf8
## [2] LC_CTYPE=Serbian (Latin)_Serbia.utf8
## [3] LC_MONETARY=Serbian (Latin)_Serbia.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=Serbian (Latin)_Serbia.utf8
##
## time zone: Europe/Belgrade
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.4.4   tidyverse_2.0.0 tinytex_0.49
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.4      highr_0.10        crayon_1.5.2
##  [5] compiler_4.3.2    tidyselect_1.2.0  parallel_4.3.2    scales_1.3.0
##  [9] yaml_2.3.8        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
## [13] generics_0.1.3    curl_5.2.0        knitr_1.45        munsell_0.5.0
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4
## [21] stringi_1.8.3     xfun_0.41         bit64_4.0.5       timechange_0.3.0
## [25] cli_3.6.2         withr_3.0.0       magrittr_2.0.3    digest_0.6.34
## [29] grid_4.3.2        vroom_1.6.5       rstudioapi_0.15.0 hms_1.1.3
## [33] lifecycle_1.0.4   vctrs_0.6.5       evaluate_0.23     glue_1.7.0
## [37] farver_2.1.1      fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.25
## [41] tools_4.3.2       pkgconfig_2.0.3   htmltools_0.5.7
```