Week3_NYPD_ShootingAnalysis_Wiggall_VFinal.rmd

A. Wiggall

2023-06-15

Week 3 - NYPD Shooting Data

By Alex Wiggall

```
install.packages("readr")  # Install required package
```

```
##
## The downloaded binary packages are in
##  /var/folders/f6/6ql5jf9j4n75l_pfw0dcqm4r0000gn/T//RtmpqYlHvG/downloaded_packages
```

```
library(readr)             # Load the package

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

shooting.data <- read_csv(url)
```

```
## 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.
```

Download the Necessary Packages

```
install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/f6/6ql5jf9j4n75l_pfw0dcqm4r0000gn/T//RtmpqYlHvG/downloaded_packages
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v purrr     1.0.1
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## -- 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 errors
```

Tidy the Data & Create an Analysis

By reviewing the data

```
incident_summary <- shooting.data %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y")) %>%
  filter(OCCUR_DATE > as.Date("2020-12-31")) %>%
  group_by(BORO) %>%
  summarize(total_incidents = n())

print(incident_summary)
```

```
## # A tibble: 5 x 2
##   BORO          total_incidents
##   <chr>                   <int>
## 1 BRONX                    1236
## 2 BROOKLYN                 1199
## 3 MANHATTAN                 650
## 4 QUEENS                    562
## 5 STATEN ISLAND              80
```

####I started my analysis first by looking at the shootings that occurred by NYC Borough. As someone living in NYC, these data interested me as a source of overall safety information based on location. Without population included, or more specific occurrence data (i.e., by Zip Code) it is difficult to determine the true underlying rate of shootings by region.

Summary of Kill Rate By Shooting

```
selected_data <- shooting.data %>%
  select(BORO, PERP_RACE, STATISTICAL_MURDER_FLAG)
kill_rate <- selected_data %>%
  group_by(BORO, PERP_RACE) %>%
  summarize(kill_rate = mean(STATISTICAL_MURDER_FLAG, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
print(kill_rate)
```

```
## # A tibble: 42 x 3
## # Groups:   BORO [5]
##    BORO     PERP_RACE                     kill_rate
##    <chr>    <chr>                             <dbl>
##  1 BRONX    (null)                           0.0874
##  2 BRONX    AMERICAN INDIAN/ALASKAN NATIVE   0
##  3 BRONX    ASIAN / PACIFIC ISLANDER         0.182
##  4 BRONX    BLACK                            0.238
##  5 BRONX    BLACK HISPANIC                   0.215
##  6 BRONX    UNKNOWN                          0.0590
##  7 BRONX    WHITE                            0.4
##  8 BRONX    WHITE HISPANIC                   0.247
##  9 BRONX    <NA>                             0.148
## 10 BROOKLYN (null)                           0.186
## # i 32 more rows
```

####Printing the kill rate allows us to analyse numerically how often a shooting resulted in a death. This is an interesting point to follow, and after reviewing I decided that looking at the kill rate by location and race would be best viewed graphically.

Create A Plot Visualizing The Kill Rate By Boro

```
install.packages("ggplot2")  # for data visualization
```

```
##
## The downloaded binary packages are in
##  /var/folders/f6/6ql5jf9j4n75l_pfw0dcqm4r0000gn/T//RtmpqYlHvG/downloaded_packages
```

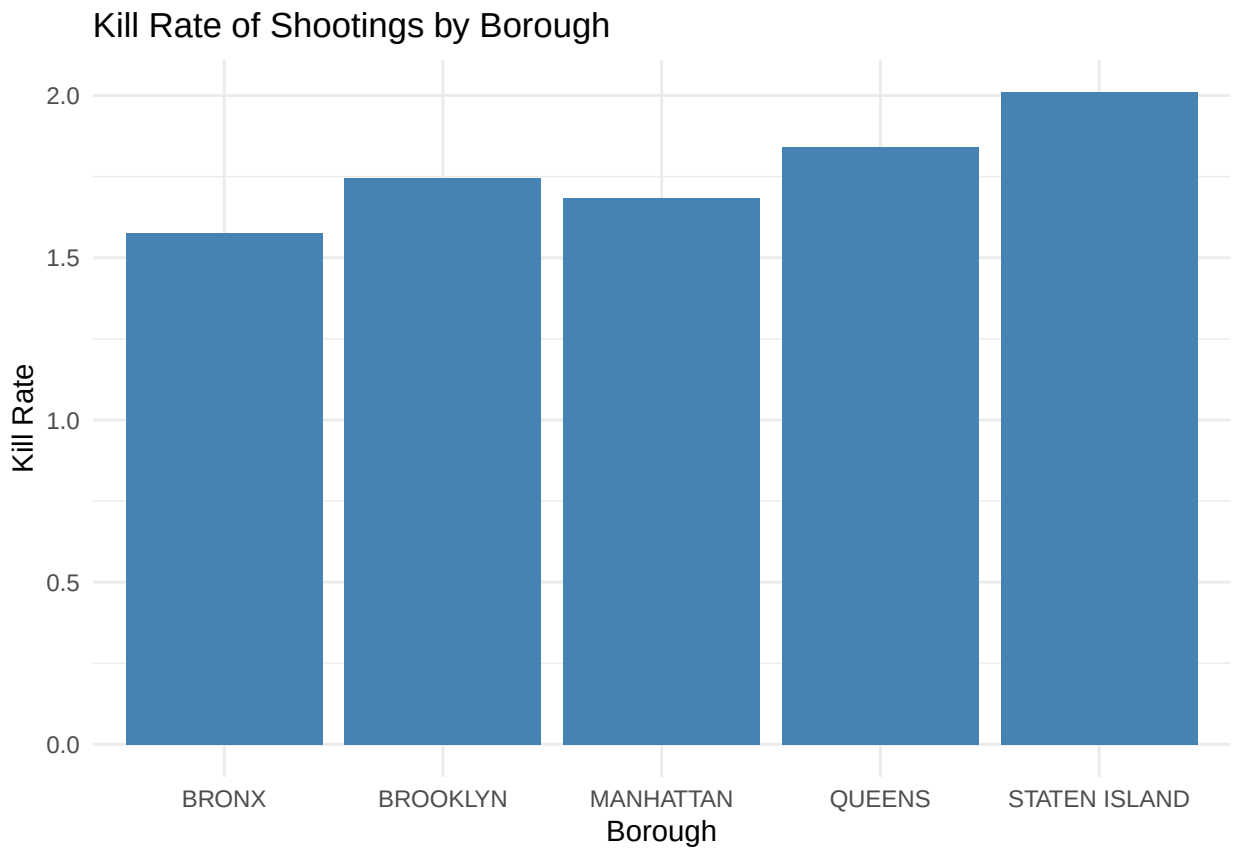
```
install.packages("ggpubr")   # for grouped bar plot
```

```
##
## The downloaded binary packages are in
##  /var/folders/f6/6ql5jf9j4n75l_pfw0dcqm4r0000gn/T//RtmpqYlHvG/downloaded_packages
```

```
library(ggplot2)
library(ggpubr)

bar_plot <- ggplot(kill_rate, aes(x = BORO, y = kill_rate)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Borough", y = "Kill Rate", title = "Kill Rate of Shootings by Borough") +
  theme_minimal()

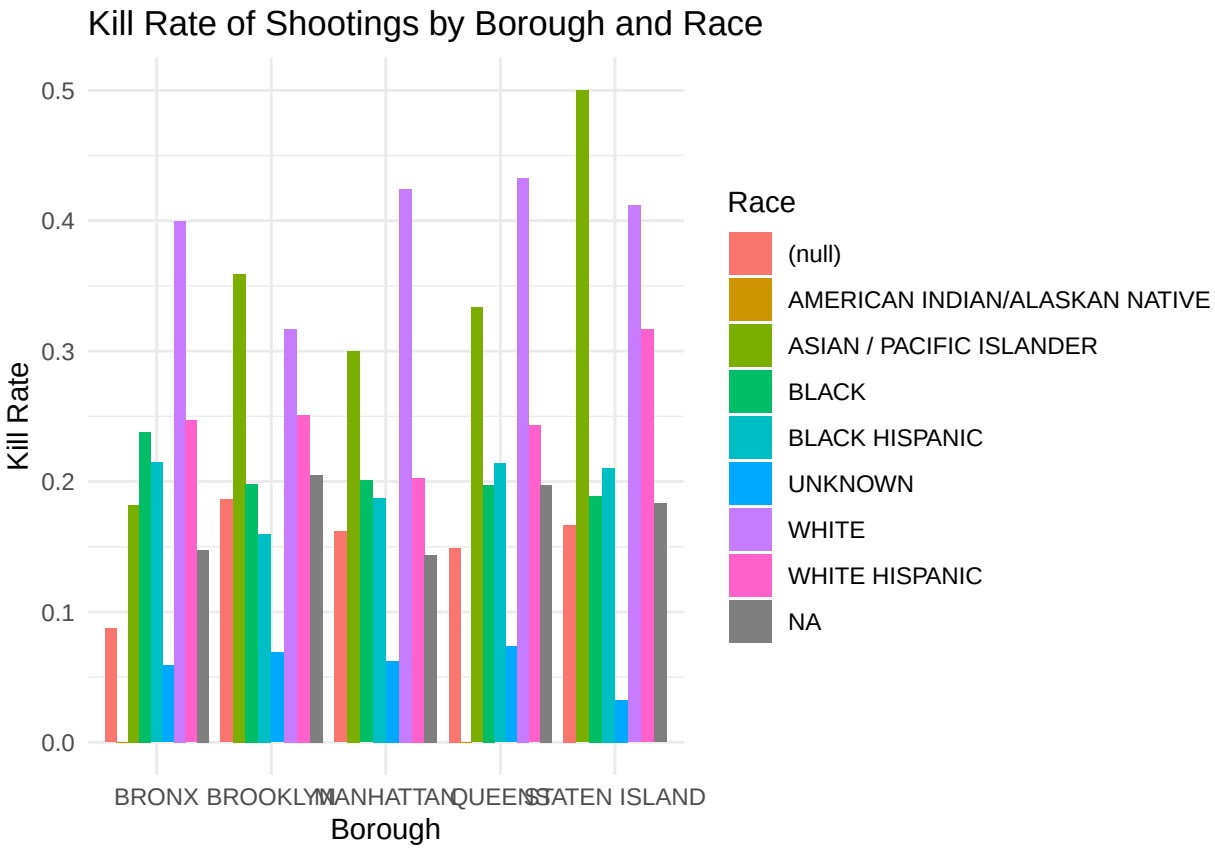
print(bar_plot)
```



Create a Visual Showing the Kill Rate By Boro and By Race

```
grouped_bar_plot <- ggplot(kill_rate, aes(x = BORO, y = kill_rate, fill = PERP_RACE)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Borough", y = "Kill Rate", title = "Kill Rate of Shootings by Borough and Race") +
  scale_fill_discrete(name = "Race") +
  theme_minimal()

print(grouped_bar_plot)
```



####The data indicates that white perpetrators, in almost all boroughs, have a much higher rate of murder when involved in a shooting versus the others. These data indicate that the only exception to this is in Staten Island. I would be interested to know if a potential source of bias in the collection of these data is an underlying lack of reporting on shootings that occur which do not involve murder when Black or Hispanic individuals are involved.

```
## Create a Linear Model

```r
# Fit a linear model
linear_model <- lm(kill_rate ~ BORO + PERP_RACE, data = kill_rate)

# Print the summary of the linear model
summary(linear_model)

##
## Call:
## lm(formula = kill_rate ~ BORO + PERP_RACE, data = kill_rate)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.130995 -0.018499  0.004083  0.021884  0.132095
##
## Coefficients:
```
```

```
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            0.12825    0.02835   4.523 0.000128
## BOROBROOKLYN                           0.01415    0.02699   0.524 0.604642
## BOROMANHATTAN                          0.01395    0.02699   0.517 0.609817
## BOROQUEENS                             0.02686    0.02565   1.047 0.305113
## BOROSTATEN ISLAND                      0.05509    0.02699   2.041 0.051952
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -0.14168    0.04422  -3.204 0.003682
## PERP_RACEASIAN / PACIFIC ISLANDER       0.18457    0.03245   5.688 6.37e-06
## PERP_RACEBLACK                         0.05437    0.03245   1.675 0.106336
## PERP_RACEBLACK HISPANIC                0.04694    0.03245   1.447 0.160427
## PERP_RACEUNKNOWN                      -0.09108    0.03245  -2.807 0.009555
## PERP_RACEWHITE                         0.24692    0.03245   7.609 5.78e-08
## PERP_RACEWHITE HISPANIC                0.10180    0.03245   3.137 0.004333
##
## (Intercept)                           ***
## BOROBROOKLYN
## BOROMANHATTAN
## BOROQUEENS
## BOROSTATEN ISLAND                     .
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE **
## PERP_RACEASIAN / PACIFIC ISLANDER      ***
## PERP_RACEBLACK
## PERP_RACEBLACK HISPANIC
## PERP_RACEUNKNOWN                      **
## PERP_RACEWHITE                        ***
## PERP_RACEWHITE HISPANIC               **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05131 on 25 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.8824, Adjusted R-squared:  0.8307
## F-statistic: 17.06 on 11 and 25 DF,  p-value: 5.658e-09
```

Session Info

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
##  [1] ggpubr_0.6.0    lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0
##  [5] dplyr_1.1.2     purrr_1.0.1     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.2   tidyverse_2.0.0 readr_2.1.4
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0 xfun_0.39        carData_3.0-5    colorspace_2.1-0
##  [5] vctrs_0.6.2      generics_0.1.3   htmltools_0.5.5  yaml_2.3.7
##  [9] utf8_1.2.3       rlang_1.1.1      pillar_1.9.0     glue_1.6.2
## [13] withr_2.5.0      bit64_4.0.5      lifecycle_1.0.3  munsell_0.5.0
## [17] ggsignif_0.6.4   gtable_0.3.3     evaluate_0.21    labeling_0.4.2
## [21] knitr_1.43       tzdb_0.4.0       fastmap_1.1.1    parallel_4.2.1
## [25] curl_5.0.0       fansi_1.0.4      highr_0.10       broom_1.0.4
## [29] scales_1.2.1     backports_1.4.1  vroom_1.6.3      abind_1.4-5
## [33] farver_2.1.1     bit_4.0.5        hms_1.1.3        digest_0.6.31
## [37] stringi_1.7.12   rstatix_0.7.2    grid_4.2.1       cli_3.6.1
## [41] tools_4.2.1      magrittr_2.0.3   crayon_1.5.2     car_3.1-2
## [45] pkgconfig_2.0.3  timechange_0.2.0 rmarkdown_2.21   rstudioapi_0.14
## [49] R6_2.5.1         compiler_4.2.1
```