Week 5 Shooting

SL

2024-03-17

Introduction & Description of Data

In this report, we will be analyzing the NYPD Shooting Incident data sourced from the NYC OpenData website: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic. The data shows the breakdown of every shooting incident that occurred in NYC from 2006 - 2022. Every record represents a shooting incident and includes information about the event, such as details regarding the perpetrator, details regarding the victim, and the location of the incident. The data was last updated on September 2023.

For this analysis, we will be looking to see if the perpetrators' demographics can be used to predict if the shooting is fatal.

Step 1

Import Libraries

```
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse
2.0.0 --
## v dplyr     1.0.8     v readr     2.1.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.1.6
## v lubridate 1.8.0     v tidyr     1.2.0
## v purrr     0.3.4
## -- 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

library(aod)
library(ggplot2)
```

Upload data and show summary

```
df = read_csv('https://data.cityofnewyork.us/api/views/833y-
fsy8/rows.csv?accessType=DOWNLOAD')
```

```
## 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.
```

```
summary(df, show_col_types = FALSE)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            X_COORD_CD        Y_COORD_CD        Latitude
```

```
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude          Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Step 2

Tidy & Transform

To tidy and transform the data, I will do a few things:

- Turn appropriate variables to factors
- Subset data
- Create a new binary variable for the statistical murder flag.
- Format applicable variables to dates.
- Replace null and U categories to unknown in variables
- Deal with NAs in in variables. I replaced the missing data with 'unknown.' Given the type of column, this data may be missing because certain details were not collected. In other words, relabeling the NA to 'unknown' may be more appropriate. A few of the other columns already use this 'unknown' label for uncollected data.

```
df <- df %>%
    replace_na(list(PERP_SEX = 'UNKNOWN', PERP_AGE_GROUP = 'UNKNOWN',
PERP_RACE = 'UNKNOWN'))

cols = c('STATISTICAL_MURDER_FLAG', 'PERP_AGE_GROUP', 'PERP_SEX',
'PERP_RACE', 'OCCUR_DATE')
shooting_df = df[cols]

shooting_df= shooting_df %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate_at(cols, factor) %>%
  mutate(murder_binary=case_when(
    STATISTICAL_MURDER_FLAG==TRUE ~ 1,
    STATISTICAL_MURDER_FLAG==FALSE ~ 0
  ))
```

```r
#Transform PERP_SEX
levels(shooting_df$PERP_SEX)[levels(shooting_df$PERP_SEX)=="(null)"] <-
"UNKNOWN"
levels(shooting_df$PERP_SEX)[levels(shooting_df$PERP_SEX)=="U"] <- "UNKNOWN"

#Transform PERP_AGE_GROUP
levels(shooting_df$PERP_AGE_GROUP)[levels(shooting_df$PERP_AGE_GROUP)=="(null
)"] <- "UNKNOWN"
levels(shooting_df$PERP_AGE_GROUP)[levels(shooting_df$PERP_AGE_GROUP)=="1020"
] <- "UNKNOWN"
levels(shooting_df$PERP_AGE_GROUP)[levels(shooting_df$PERP_AGE_GROUP)=="224"]
<- "UNKNOWN"
levels(shooting_df$PERP_AGE_GROUP)[levels(shooting_df$PERP_AGE_GROUP)=="940"]
<- "UNKNOWN"

#Transform PERP_RACE
levels(shooting_df$PERP_RACE)[levels(shooting_df$PERP_RACE)=="(null)"] <-
"UNKNOWN"


table(shooting_df$PERP_AGE_GROUP)

##
## UNKNOWN     <18   18-24   25-44   45-64     65+
##   13135    1591    6222    5687     617      60

table(shooting_df$PERP_SEX)

##
## UNKNOWN       F       M
##   11449     424   15439

table(shooting_df$PERP_RACE)

##
##                        UNKNOWN AMERICAN INDIAN/ALASKAN NATIVE
##                          11786                              2
##       ASIAN / PACIFIC ISLANDER                          BLACK
##                            154                          11432
##                 BLACK HISPANIC                          WHITE
##                           1314                            283
##                 WHITE HISPANIC
##                           2341

table(shooting_df$STATISTICAL_MURDER_FLAG)

##
## FALSE  TRUE
## 22046  5266
```

```
summary(shooting_df)
```

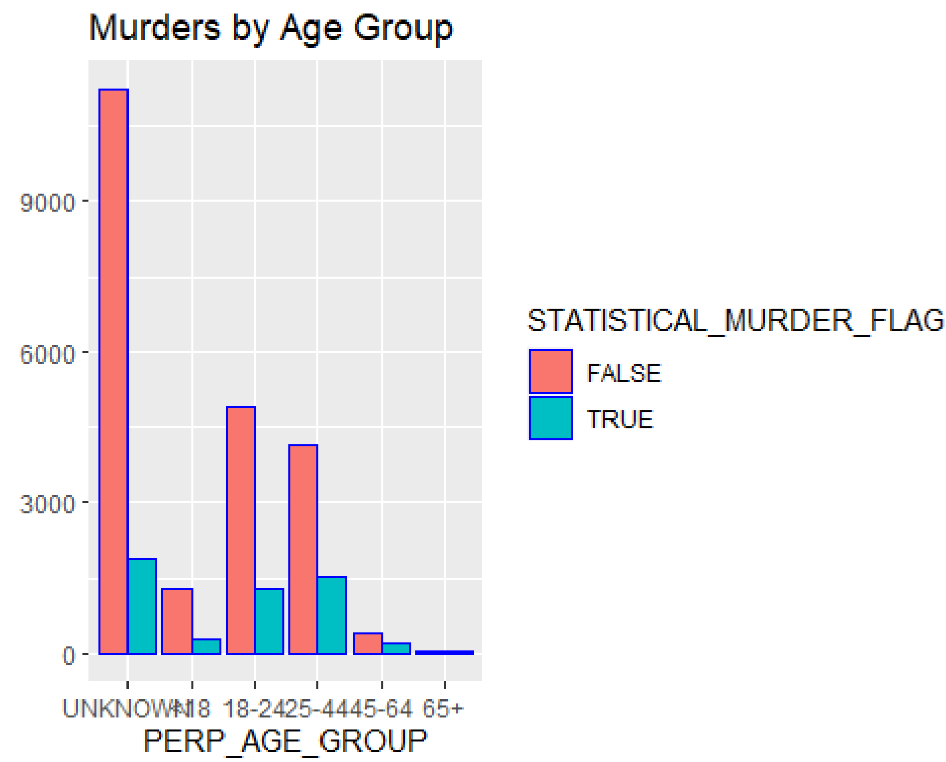
```
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  FALSE:22046             UNKNOWN:13135   UNKNOWN:11449
##  TRUE : 5266             <18     : 1591   F      :  424
##                          18-24   : 6222   M      :15439
##                          25-44   : 5687
##                          45-64   :  617
##                          65+     :   60
##
##                            PERP_RACE          OCCUR_DATE    murder_binary
##  UNKNOWN                        :11786   2020-07-05:   47   Min.   :0.0000
##  AMERICAN INDIAN/ALASKAN NATIVE:     2   2011-09-04:   31   1st Qu.:0.0000
##  ASIAN / PACIFIC ISLANDER      :   154   2020-07-26:   29   Median :0.0000
##  BLACK                         :11432   2007-08-11:   26   Mean   :0.1928
##  BLACK HISPANIC                : 1314   2006-09-04:   25   3rd Qu.:0.0000
##  WHITE                         :   283   2022-08-27:   25   Max.   :1.0000
##  WHITE HISPANIC                : 2341   (Other)   :27129
```

Step 3 Add Visuals and Analysis

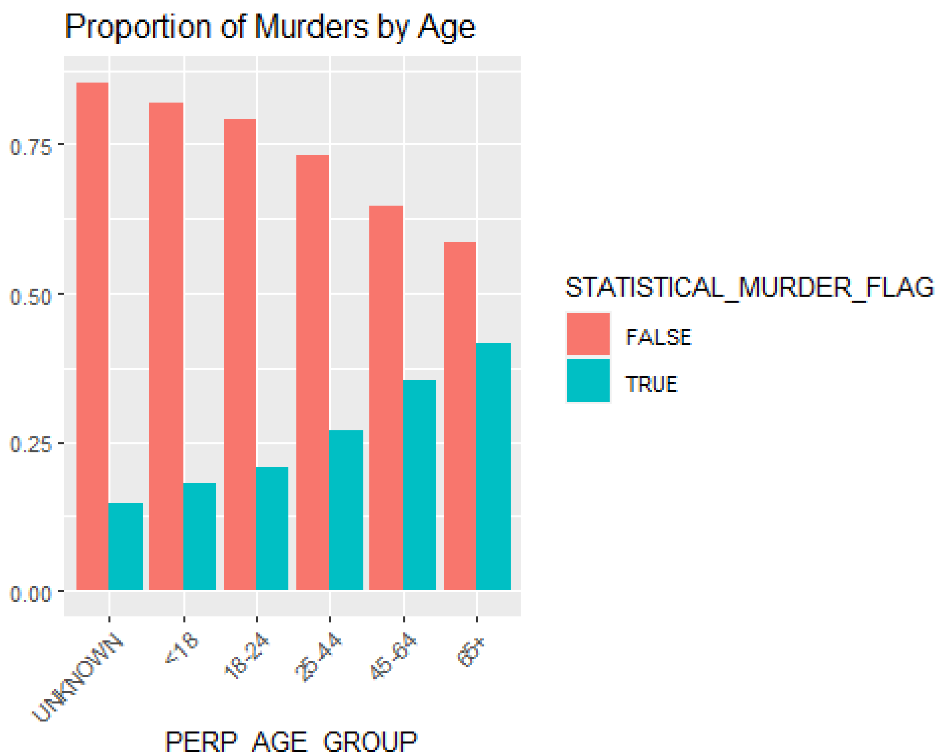
Visuals

The plots below show different views of looking at the demographic variables (age, sex, race of perpetrator) and the statistical murder flag variable. I believe it's important to look at both the overall counts of each group and the proportions. For example, the Murders by Gender plot shows that a majority of murders are done by males, from a count perspective. However, the Proportion of Murders by Gender plot shows that of crimes committed by each gender, females have a slightly higher proportion of murders.

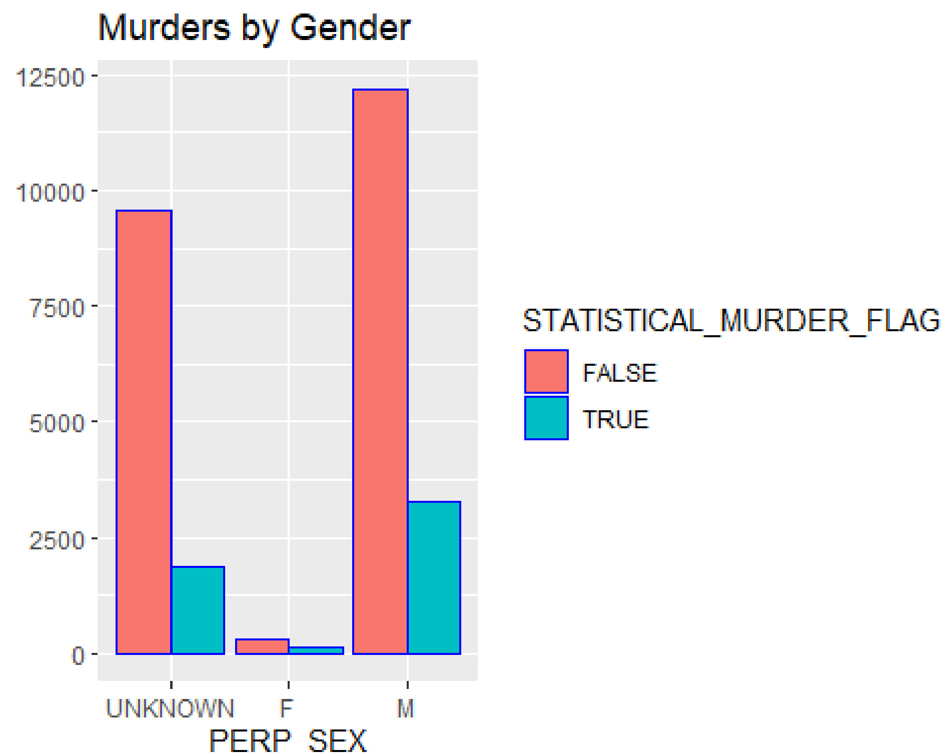
```
shooting_df %>% ggplot() +
  geom_bar(aes(PERP_AGE_GROUP, fill = STATISTICAL_MURDER_FLAG), color =
'blue',position=position_dodge())+
  labs(title = str_c('Murders by Age Group'), y = NULL)
```



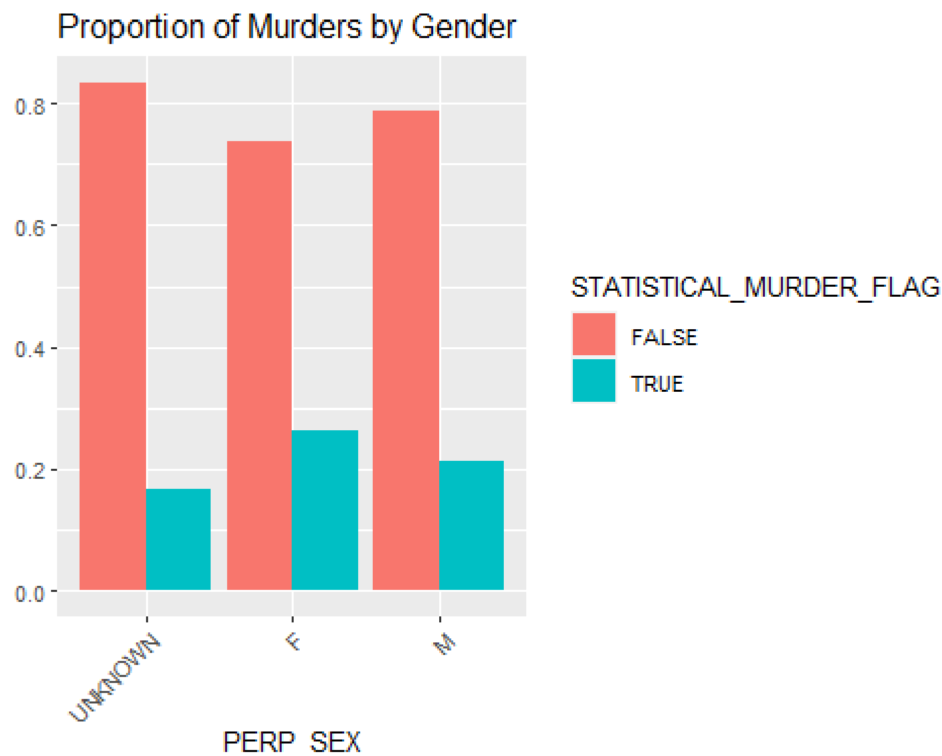

```
shooting_df %>%
  count(STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP) %>%
  group_by(PERP_AGE_GROUP) %>%
   mutate(Sum=sum(n)) %>%
   mutate(proportion = n/Sum) %>%
  ggplot(aes(y=proportion, x=PERP_AGE_GROUP, fill=STATISTICAL_MURDER_FLAG)) +
   geom_col(position = "dodge")+
  theme(text = element_text(size = 10), axis.text.x = element_text(angle =
45, hjust = 1)) +
  labs(title = str_c('Proportion of Murders by Age'), y = NULL)
```



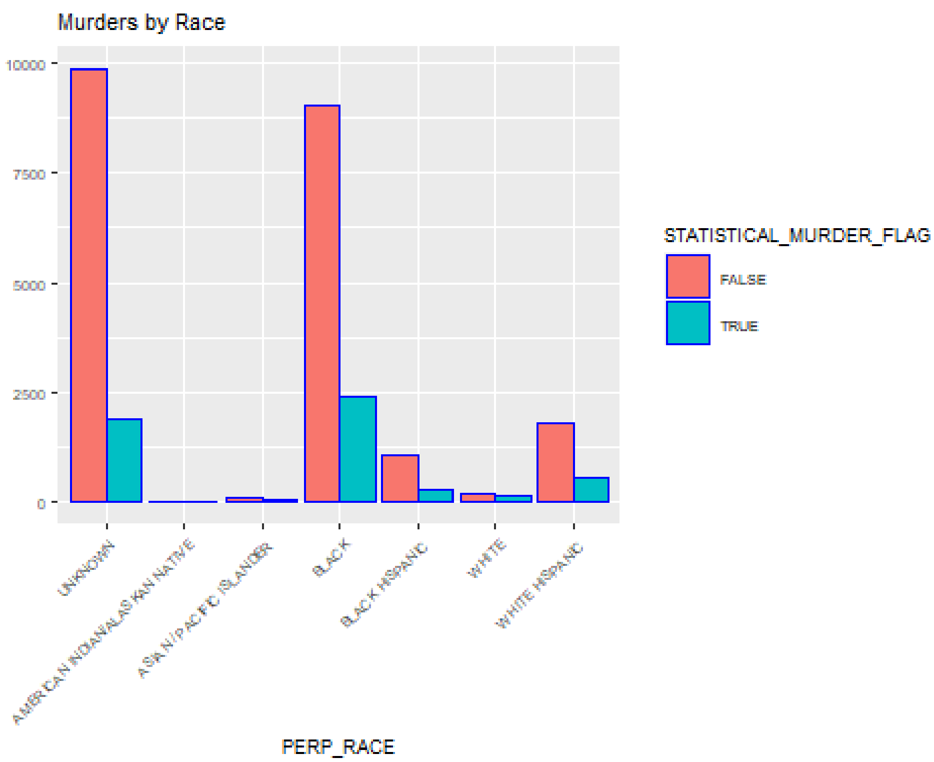

```
shooting_df %>% ggplot() +
  geom_bar(aes(PERP_SEX, fill = STATISTICAL_MURDER_FLAG), color =
'blue',position=position_dodge())+
  labs(title = str_c('Murders by Gender'), y = NULL)
```



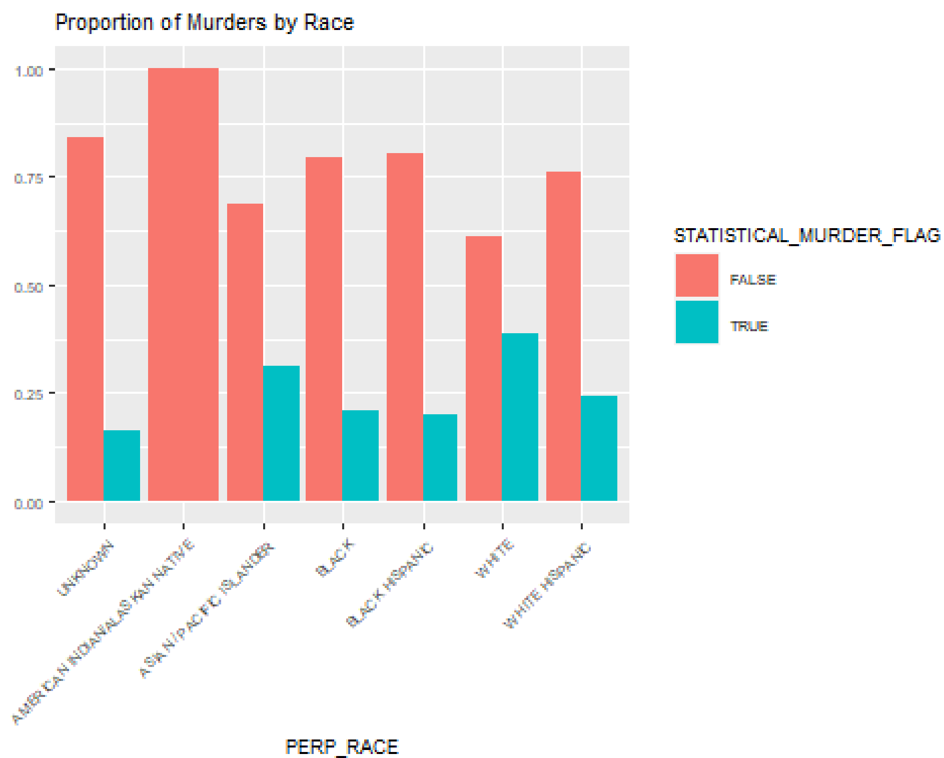

```
shooting_df %>%
  count(STATISTICAL_MURDER_FLAG, PERP_SEX) %>%
  group_by(PERP_SEX) %>%
   mutate(Sum=sum(n)) %>%
   mutate(proportion = n/Sum) %>%
  ggplot(aes(y=proportion, x=PERP_SEX,fill=STATISTICAL_MURDER_FLAG)) +
   geom_col(position = "dodge")+
  theme(text = element_text(size = 10), axis.text.x = element_text(angle =
45, hjust = 1)) +
  labs(title = str_c('Proportion of Murders by Gender'), y = NULL)
```




```
shooting_df %>% ggplot() +
  geom_bar(aes(PERP_RACE, fill = STATISTICAL_MURDER_FLAG), color =
'blue',position=position_dodge())+
  labs(title = str_c('Murders by Race'), y = NULL)+
  theme(text = element_text(size = 7), axis.text.x = element_text(angle = 45,
hjust = 1))
```




```
shooting_df %>%
  count(STATISTICAL_MURDER_FLAG, PERP_RACE) %>%
  group_by(PERP_RACE) %>%
   mutate(Sum=sum(n)) %>%
   mutate(proportion = n/Sum) %>%
  ggplot(aes(y=proportion, x=PERP_RACE,fill=STATISTICAL_MURDER_FLAG)) +
   geom_col(position = "dodge")+
  theme(text = element_text(size = 7), axis.text.x = element_text(angle = 45,
hjust = 1)) +
  labs(title = str_c('Proportion of Murders by Race'), y = NULL)
```



Regression

```
options(scipen=999)
## Give summary of log odds explanation of those who commit murder, tend ot
be older

mod = glm(murder_binary ~ PERP_SEX + PERP_AGE_GROUP + PERP_RACE , data =
shooting_df, family = 'binomial')
summary(mod)

##
## Call:
## glm(formula = murder_binary ~ PERP_SEX + PERP_AGE_GROUP + PERP_RACE,
##     family = "binomial", data = shooting_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8785  -0.6762  -0.5983  -0.2276   2.9206
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                               -1.62953    0.02524 -64.559
## PERP_SEXF                                 -2.46257    0.26502  -9.292
## PERP_SEXM                                 -2.62138    0.23942 -10.949
## PERP_AGE_GROUP<18                          2.22749    0.17028  13.081
## PERP_AGE_GROUP18-24                        2.40937    0.16032  15.028
## PERP_AGE_GROUP25-44                        2.72387    0.16032  16.990
```

```
## PERP_AGE_GROUP45-64                         3.08530    0.17926  17.212
## PERP_AGE_GROUP65+                           3.25082    0.30987  10.491
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -8.96266   84.41341  -0.106
## PERP_RACEASIAN / PACIFIC ISLANDER           0.94600    0.27273   3.469
## PERP_RACEBLACK                              0.48219    0.20808   2.317
## PERP_RACEBLACK HISPANIC                     0.38012    0.21850   1.740
## PERP_RACEWHITE                              1.08441    0.24268   4.468
## PERP_RACEWHITE HISPANIC                     0.61010    0.21299   2.865
##                                                       Pr(>|z|)
## (Intercept)                             < 0.0000000000000002 ***
## PERP_SEXF                               < 0.0000000000000002 ***
## PERP_SEXM                               < 0.0000000000000002 ***
## PERP_AGE_GROUP<18                       < 0.0000000000000002 ***
## PERP_AGE_GROUP18-24                     < 0.0000000000000002 ***
## PERP_AGE_GROUP25-44                     < 0.0000000000000002 ***
## PERP_AGE_GROUP45-64                     < 0.0000000000000002 ***
## PERP_AGE_GROUP65+                       < 0.0000000000000002 ***
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE             0.915443
## PERP_RACEASIAN / PACIFIC ISLANDER                   0.000523 ***
## PERP_RACEBLACK                                      0.020488 *
## PERP_RACEBLACK HISPANIC                             0.081917 .
## PERP_RACEWHITE                                    0.00000788 ***
## PERP_RACEWHITE HISPANIC                             0.004176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26781  on 27311  degrees of freedom
## Residual deviance: 25855  on 27298  degrees of freedom
## AIC: 25883
##
## Number of Fisher Scoring iterations: 9
```

```
exp(coef(mod))
```

```
##                              (Intercept)
PERP_SEXF
##                             0.1960208102
0.0852158740
##                                PERP_SEXM
PERP_AGE_GROUP<18
##                             0.0727026736
9.2765718188
##                      PERP_AGE_GROUP18-24
PERP_AGE_GROUP25-44
##                            11.1269622803
15.2391257665
##                      PERP_AGE_GROUP45-64
PERP_AGE_GROUP65+
```

```
##                              21.8740537078
25.8115156328
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE       PERP_RACEASIAN / PACIFIC
ISLANDER
##                               0.0001281049
2.5753910447
##                             PERP_RACEBLACK                   PERP_RACEBLACK
HISPANIC
##                               1.6196140006
1.4624541493
##                             PERP_RACEWHITE                   PERP_RACEWHITE
HISPANIC
##                               2.9576904503
1.8406221971
```

```
exp(cbind(coef(mod), confint(mod)))
```

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

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
##                                                                2.5 %     97.5
%
## (Intercept)                                0.1960208102  0.18650769
0.2059065
## PERP_SEXF                                  0.0852158740  0.04988018
0.1411763
## PERP_SEXM                                  0.0727026736  0.04462646
0.1142653
```

```
## PERP_AGE_GROUP<18                          9.2765718188  6.71649218
13.1120146
## PERP_AGE_GROUP18-24                       11.1269622803  8.22959321
15.4495981
## PERP_AGE_GROUP25-44                       15.2391257665 11.27166303
21.1600085
## PERP_AGE_GROUP45-64                       21.8740537078 15.54359577
31.4292110
## PERP_AGE_GROUP65+                         25.8115156328 14.00098072
47.3739560
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE  0.0001281049          NA
72.4852448
## PERP_RACEASIAN / PACIFIC ISLANDER        2.5753910447  1.51715533
4.4301885
## PERP_RACEBLACK                           1.6196140006  1.09354797
2.4780724
## PERP_RACEBLACK HISPANIC                  1.4624541493  0.96593104
2.2799994
## PERP_RACEWHITE                           2.9576904503  1.85732079
4.8191191
## PERP_RACEWHITE HISPANIC                  1.8406221971  1.23005050
2.8413826
```

Thoughts based on regression results:

- Based on the regression results, many of the demographic variable are statistically significant predictors of fatality. Also a few of the categories have such small sample sizes that the regression had trouble modelling them.
- To begin digging into practical significance and for interpretability, I calculated the odds ratio of each variable and the 95% CI for each odds ratio.
- Based on the odds ratio and the CIs, although gender is statistically significant, it does not seem to be a strong predictor of fatality. Age seems to be the strongest predictor, and shows that the older the perpetrator is, the more likely the shooting is fatal–Where the odds of a shooting being fatal is almost 26 times higher if the perpetrator is 65+ vs not, if all other variables are constant, given an odds ratio of 25.8. When looking at race, shootings with White perpetrator are more likely to be fatal—Where the odds of a shooting being fatal is almost 3 times higher if the perpetrator is white vs not, if all other variables are constant, given an odds ratio of 2.96.

Questions raised by this analysis: - I wonder if older perpetrators were more likely to target older victims, thus lowering the likelihood of the victim surviving. - I wonder if the motivations behind the shootings vary by demographics. For example, maybe younger, black, male perpetrators use shootings as an intimidation technique but do not purposely try to kill their victims, thus lowering the likelihood of those shootings being fatal.

Conclusion

In this report, I endeavored to understand if the perpetrators' demographics can be used to predict if the shooting is fatal. When looking at the overall counts in the data, shootings, including fatal shootings, seem to be associated with younger, male, and black perpetrators. This interpretation of the count data is disregarding the shootings where demographics are unknown. However, after looking at the proportion of fatal shooting per demographic, it seems like fatal shootings seem to be associated with older and white perpetrators. The logistic regression I conducted also supports this interpretation. Based on the regression results, a shooting is more likely be fatal if the perpetrator is older and White. That being said, there is a large amount of shootings where demographics were not collected.

In terms of personal biases, as a young, Hispanic, female, I could be more sympathetic towards perpetrators in my age range, which is the 25-44 age range. Also, I can be more forgiving toward perpetrators who are classified as white Hispanic

Session Info

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## 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
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] aod_1.3.3       lubridate_1.8.0 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.0.8     purrr_0.3.4     readr_2.1.2     tidyr_1.2.0
##  [9] tibble_3.1.6    ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] highr_0.10        pillar_1.9.0      compiler_4.0.5    tools_4.0.5
##  [5] bit_4.0.4         digest_0.6.27     evaluate_0.23     lifecycle_1.0.4
##  [9] gtable_0.3.4      pkgconfig_2.0.3   rlang_1.1.3       cli_3.6.2
## [13] DBI_1.2.2         rstudioapi_0.15.0 curl_4.3.2        parallel_4.0.5
## [17] yaml_2.3.5        xfun_0.42.4       fastmap_1.1.0     withr_3.0.0
## [21] knitr_1.45        generics_0.1.3    vctrs_0.6.5       hms_1.1.3
## [25] bit64_4.0.5       grid_4.0.5        tidyselect_1.2.0  glue_1.6.2
```

```
## [29] R6_2.5.1          fansi_1.0.3       vroom_1.5.7       rmarkdown_2.26
## [33] farver_2.1.0      tzdb_0.3.0        magrittr_2.0.3    MASS_7.3-53.1
## [37] scales_1.2.0      ellipsis_0.3.2    htmltools_0.5.7
assertthat_0.2.1
## [41] colorspace_2.0-3  labeling_0.4.3    utf8_1.2.2        stringi_1.7.6
## [45] munsell_0.5.0     crayon_1.5.2
```