Final Project

Section 2

A’di Dust

2021-10-06

```
knitr::opts_chunk$set(echo = TRUE, error = TRUE, message = FALSE)
# load packages
library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(mosaic)
library(ggmosaic)
```

```
police_stops <- read.csv('mn_saint_paul_2020_04_01.csv')
```

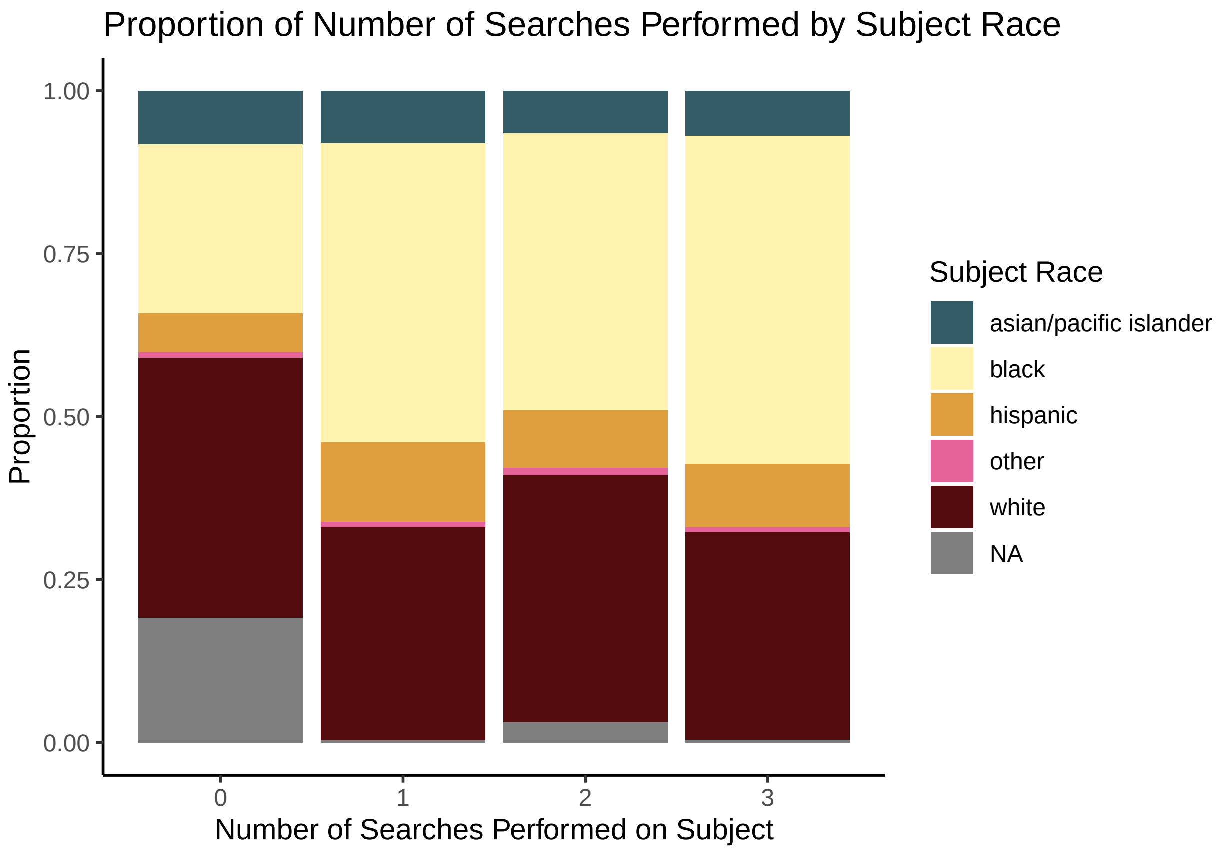
Cleaning And Variable Manipulation

```
#transform data

#sum searches
police_stops <- police_stops %>%
rowwise() %>%
mutate(sum_searches = sum(frisk_performed, search_conducted, search_vehicle))

#transform citation
police_stops <- police_stops %>%
rowwise() %>%
mutate(cited = as.numeric(citation_issued))
```

Question 1

Exploratory Data Analysis

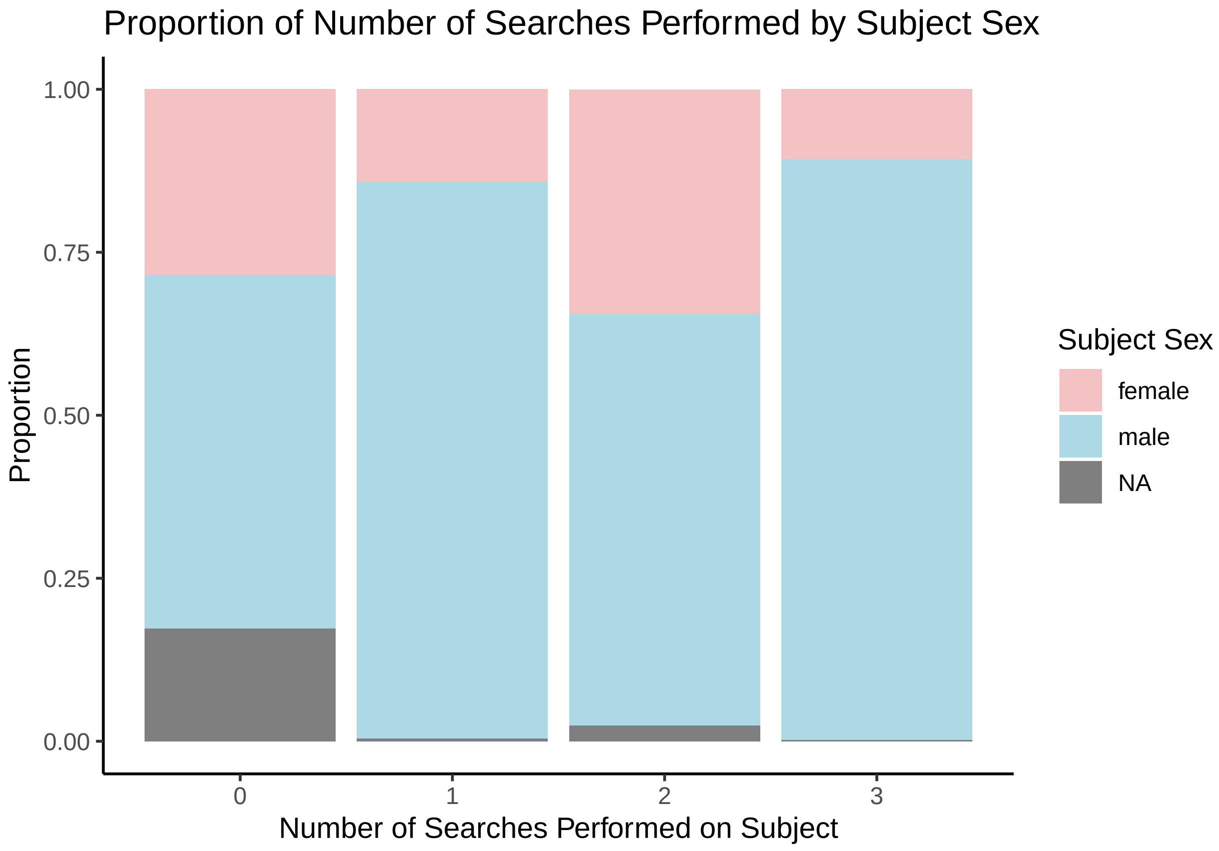
```
# race versus searches

# Visualization

police_stops %>%
    ggplot(aes(x = sum_searches, fill = subject_race)) +
    geom_bar(position = "fill") +
    scale_fill_manual(values = c("#335C67", "#FFF3B0", "#E09F3E", "#E56399", "#540B0E")) +
    ggtitle('Proportion of Number of Searches Performed by Subject Race') +
        labs(x = 'Number of Searches Performed on Subject', y = 'Proportion', fill = 'Subject Race') +
    theme_classic()
```



```
# summary

police_stops %>%
  group_by(sum_searches) %>%
  count(subject_race) %>%
  mutate(relfreq = n / sum(n))
```

```
## # A tibble: 24 x 4
## # Groups:   sum_searches [4]
##    sum_searches subject_race                n relfreq
##           <int> <chr>                   <int>   <dbl>
##  1            0 asian/pacific islander  50643 0.0822
##  2            0 black                  159845 0.259
##  3            0 hispanic                36951 0.0600
##  4            0 other                    4836 0.00785
##  5            0 white                  246138 0.399
##  6            0 <NA>                   117743 0.191
##  7            1 asian/pacific islander   1272 0.0809
##  8            1 black                    7215 0.459
##  9            1 hispanic                 1914 0.122
## 10            1 other                     128 0.00814
## # ... with 14 more rows
```

```
# sex versus searches

# Visualization

police_stops %>%
    ggplot(aes(x = sum_searches, fill = subject_sex)) +
    geom_bar(position = "fill") +
    scale_fill_manual(values = c("#F4C2C2", "#ADD8E6")) +
    ggtitle('Proportion of Number of Searches Performed by Subject Sex') +
    labs(x = 'Number of Searches Performed on Subject', y = 'Proportion', fill = 'Subject Sex') +
    theme_classic()
```



```
# summary

police_stops %>%
  group_by(sum_searches) %>%
  count(subject_sex) %>%
  mutate(relfreq = n / sum(n))
```

```
## # A tibble: 12 x 4
## # Groups:   sum_searches [4]
##    sum_searches subject_sex      n relfreq
##           <int> <chr>        <int>   <dbl>
##  1            0 female      175432 0.285
##  2            0 male        334074 0.542
##  3            0 <NA>        106650 0.173
##  4            1 female        2223 0.141
##  5            1 male         13437 0.854
##  6            1 <NA>            68 0.00432
##  7            2 female         356 0.344
##  8            2 male           653 0.632
##  9            2 <NA>            25 0.0242
## 10            3 female        4526 0.107
## 11            3 male         37630 0.891
## 12            3 <NA>            82 0.00194
```

Model Creation

```
mod1 <- lm(data=police_stops, sum_searches~subject_race * subject_sex)

mod1
```

```
##
## Call:
## lm(formula = sum_searches ~ subject_race * subject_sex, data = police_stops)
##
## Coefficients:
##                          (Intercept)                      subject_raceblack
##                              0.04639                                0.07312
##                 subject_racehispanic                      subject_raceother
##                              0.05552                                0.09784
##                    subject_racewhite                        subject_sexmale
##                              0.03570                                0.18573
##    subject_raceblack:subject_sexmale   subject_racehispanic:subject_sexmale
##                              0.18295                                0.10956
##    subject_raceother:subject_sexmale      subject_racewhite:subject_sexmale
##                             -0.08756                               -0.03920
```

Model Evaluation

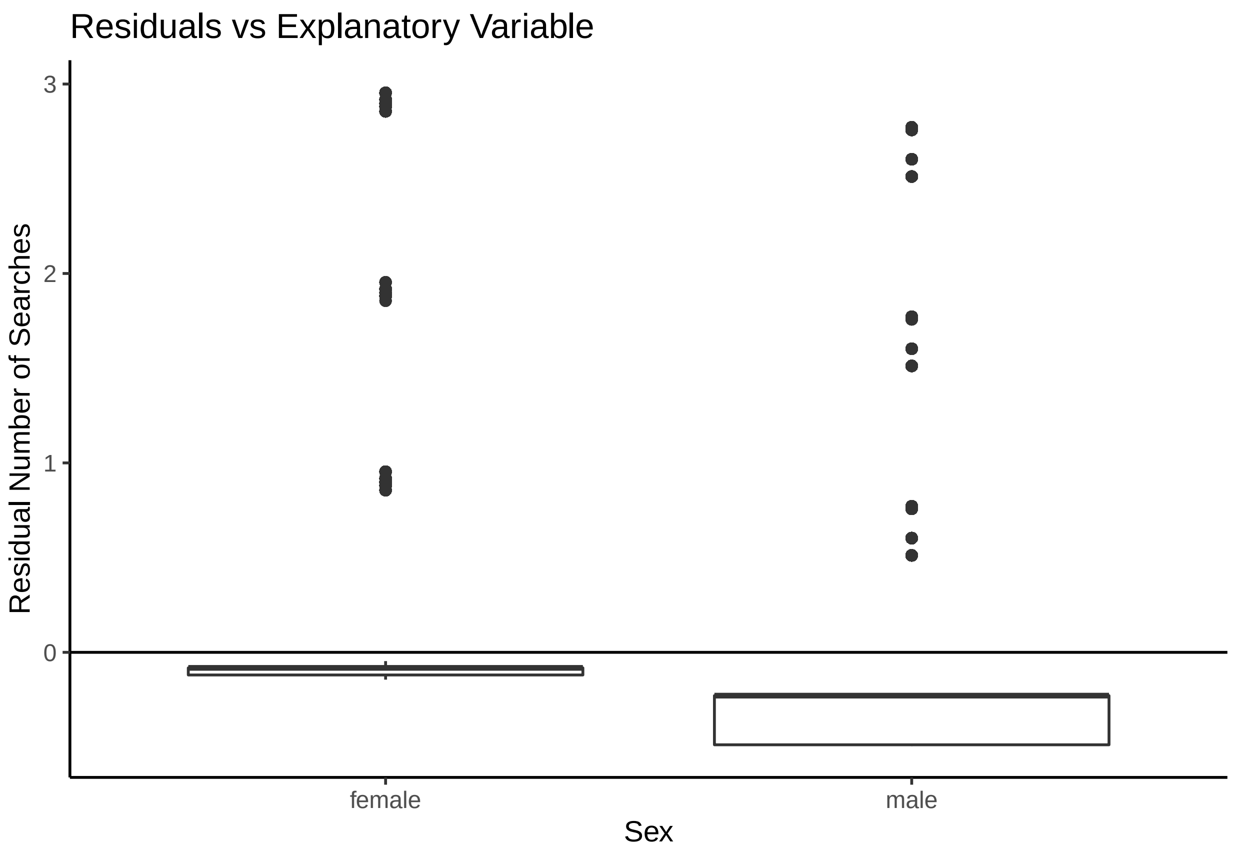
```
confint(mod1)
```

```
##                                             2.5 %      97.5 %
## (Intercept)                            0.03324188  0.05954014
## subject_raceblack                      0.05841091  0.08783802
## subject_racehispanic                   0.03462281  0.07642103
## subject_raceother                      0.05969460  0.13598492
## subject_racewhite                      0.02164884  0.04975636
## subject_sexmale                        0.17051353  0.20094309
## subject_raceblack:subject_sexmale     0.16582767  0.20006685
## subject_racehispanic:subject_sexmale  0.08577149  0.13335756
## subject_raceother:subject_sexmale    -0.13470287 -0.04042049
## subject_racewhite:subject_sexmale    -0.05564009 -0.02275030
```

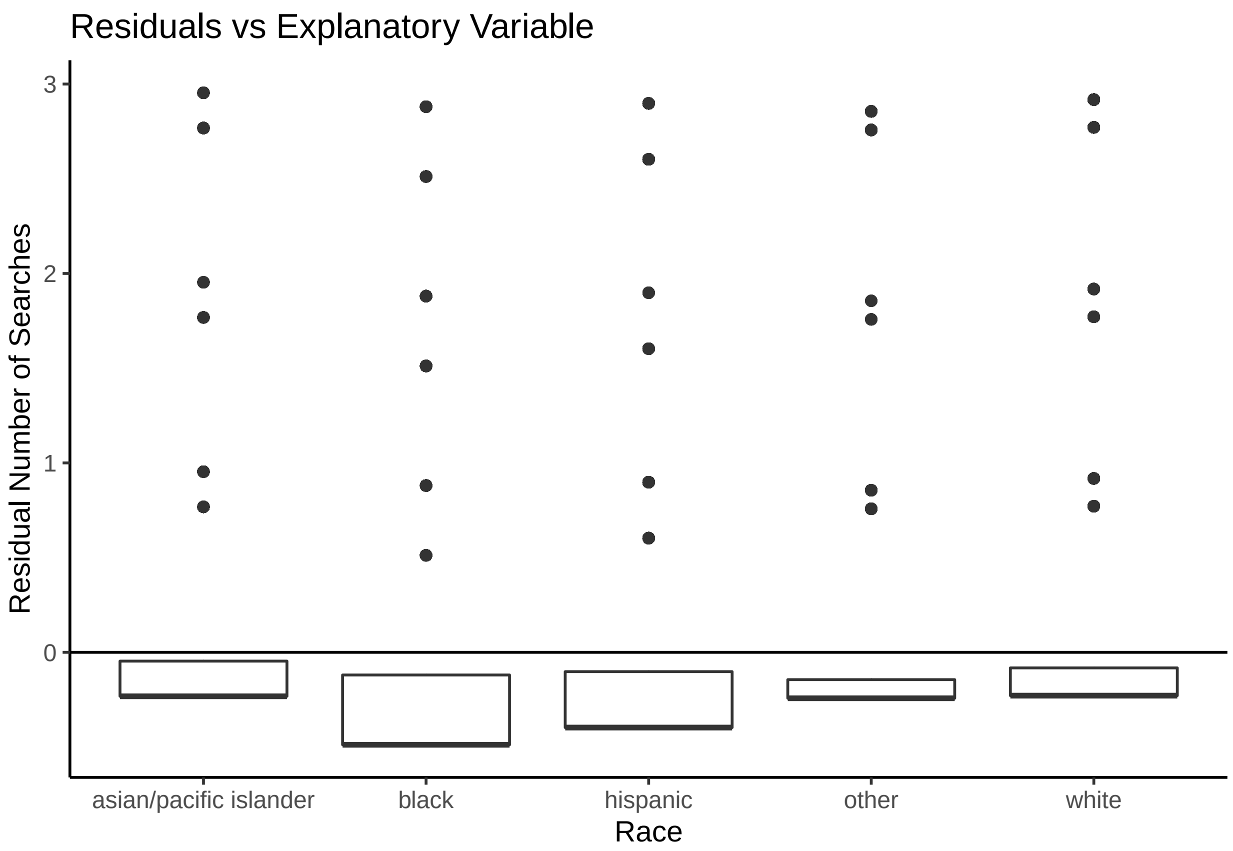
```
tidy(mod1)
```

```
## # A tibble: 10 x 5
##    term                                 estimate std.error statistic   p.value
##    <chr>                                   <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                            0.0464   0.00671      6.91 4.69e- 12
##  2 subject_raceblack                      0.0731   0.00751      9.74 2.03e- 22
##  3 subject_racehispanic                   0.0555   0.0107       5.21 1.92e-  7
##  4 subject_raceother                      0.0978   0.0195       5.03 4.98e-  7
##  5 subject_racewhite                      0.0357   0.00717      4.98 6.39e-  7
##  6 subject_sexmale                        0.186    0.00776     23.9  1.93e-126
##  7 subject_raceblack:subject_sexmale      0.183    0.00873     20.9  2.27e- 97
##  8 subject_racehispanic:subject_sexmale   0.110    0.0121       9.03 1.79e- 19
##  9 subject_raceother:subject_sexmale     -0.0876   0.0241      -3.64 2.72e-  4
## 10 subject_racewhite:subject_sexmale     -0.0392   0.00839     -4.67 2.99e-  6
```

```
# residuals vs sex
augment(mod1) %>%
  ggplot(aes(y = .resid, x = subject_sex)) +
  geom_boxplot() +
  geom_smooth(se = FALSE) +  # add smooth line (without gray uncertainty interval)
  geom_hline(yintercept = 0) + # add horizontal line at y = 0
  labs(x = 'Sex', y = 'Residual Number of Searches', title = 'Residuals vs Explanatory Variable') + # u
  theme_classic()
```



```
# residuals vs race
augment(mod1) %>%
  ggplot(aes(y = .resid, x = subject_race)) +
  geom_boxplot() +
  geom_smooth(se = FALSE) +  # add smooth line (without gray uncertainty interval)
  geom_hline(yintercept = 0) + # add horizontal line at y = 0
  labs(x = 'Race', y = 'Residual Number of Searches', title = 'Residuals vs Explanatory Variable') + #
  theme_classic()
```



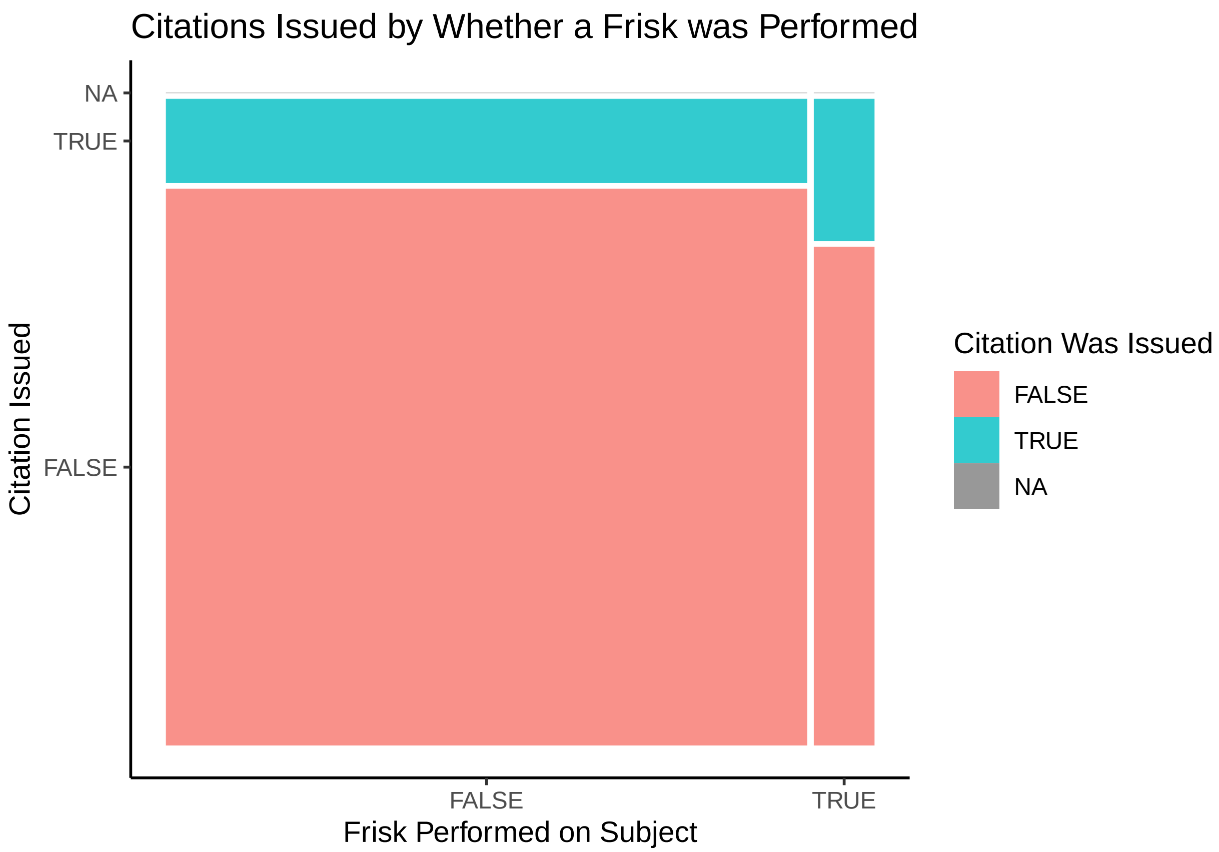
```
#measures of goodness
glance(mod1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df   logLik      AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>    <dbl>    <dbl>  <dbl>
## 1    0.0356        0.0356 0.790     2283.       0     9 -658744. 1317510. 1.32e6
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Question 2

Exploratory Data Analysis

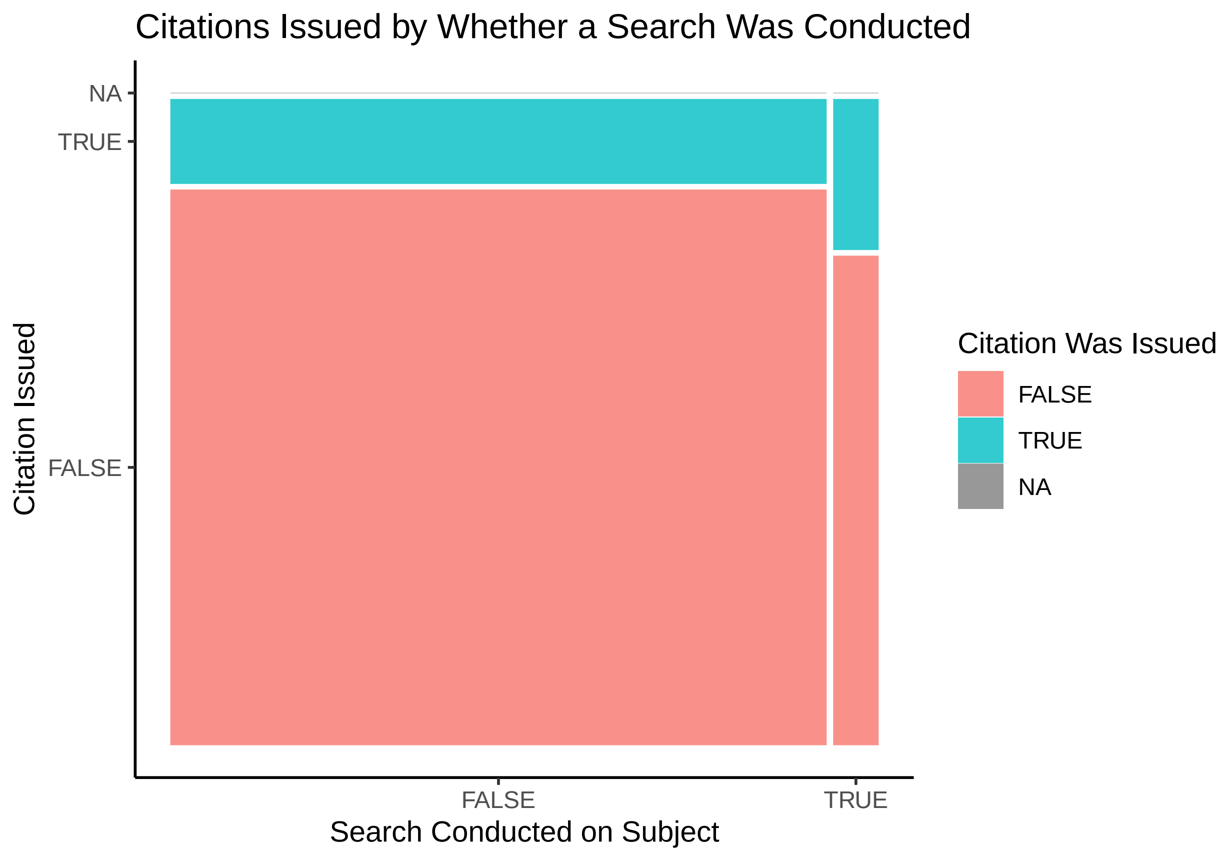
```
# frisks performed

# Visualization

police_stops %>%
    ggplot() +
    geom_mosaic(aes(x = product(citation_issued, frisk_performed), fill = citation_issued)) +
    ggtitle('Citations Issued by Whether a Frisk was Performed') +
    labs(y = 'Citation Issued', x = 'Frisk Performed on Subject', fill = 'Citation Was Issued') +
    theme_classic()
```



```
# summary

police_stops %>%
  group_by(frisk_performed) %>%
  count(citation_issued) %>%
  mutate(relfreq = n / sum(n))
```

```
## # A tibble: 5 x 4
## # Groups:   frisk_performed [2]
##   frisk_performed citation_issued      n    relfreq
##   <lgl>           <lgl>            <int>      <dbl>
## 1 FALSE           FALSE           536474 0.869
## 2 FALSE           TRUE             80713 0.131
## 3 FALSE           NA                   3 0.00000486
## 4 TRUE            FALSE            45124 0.778
## 5 TRUE            TRUE             12842 0.222
```

```
# searches conducted

# Visualization

police_stops %>%
    ggplot() +
    geom_mosaic(aes(x = product(citation_issued, search_conducted), fill = citation_issued)) +
    ggtitle('Citations Issued by Whether a Search Was Conducted') +
    labs(y = 'Citation Issued', x = 'Search Conducted on Subject', fill = 'Citation Was Issued') +
    theme_classic()
```



```
# summary

police_stops %>%
  group_by(search_conducted) %>%
  count(citation_issued) %>%
  mutate(relfreq = n / sum(n))
```

```
## # A tibble: 5 x 4
## # Groups:   search_conducted [2]
##   search_conducted citation_issued      n    relfreq
##   <lgl>            <lgl>            <int>      <dbl>
## 1 FALSE            FALSE           548510 0.868
## 2 FALSE            TRUE             83371 0.132
## 3 FALSE            NA                   3 0.00000475
## 4 TRUE             FALSE            33088 0.765
## 5 TRUE             TRUE             10184 0.235
```

Model Creation

```r
# fit model 2
mod2 <- glm(data=police_stops, family='binomial', citation_issued~frisk_performed+search_conducted)

coef(mod2) %>%
  exp()
```

```
##          (Intercept)  frisk_performedTRUE search_conductedTRUE
##            0.1503531            1.4648352            1.4090136
```

Fitted Model

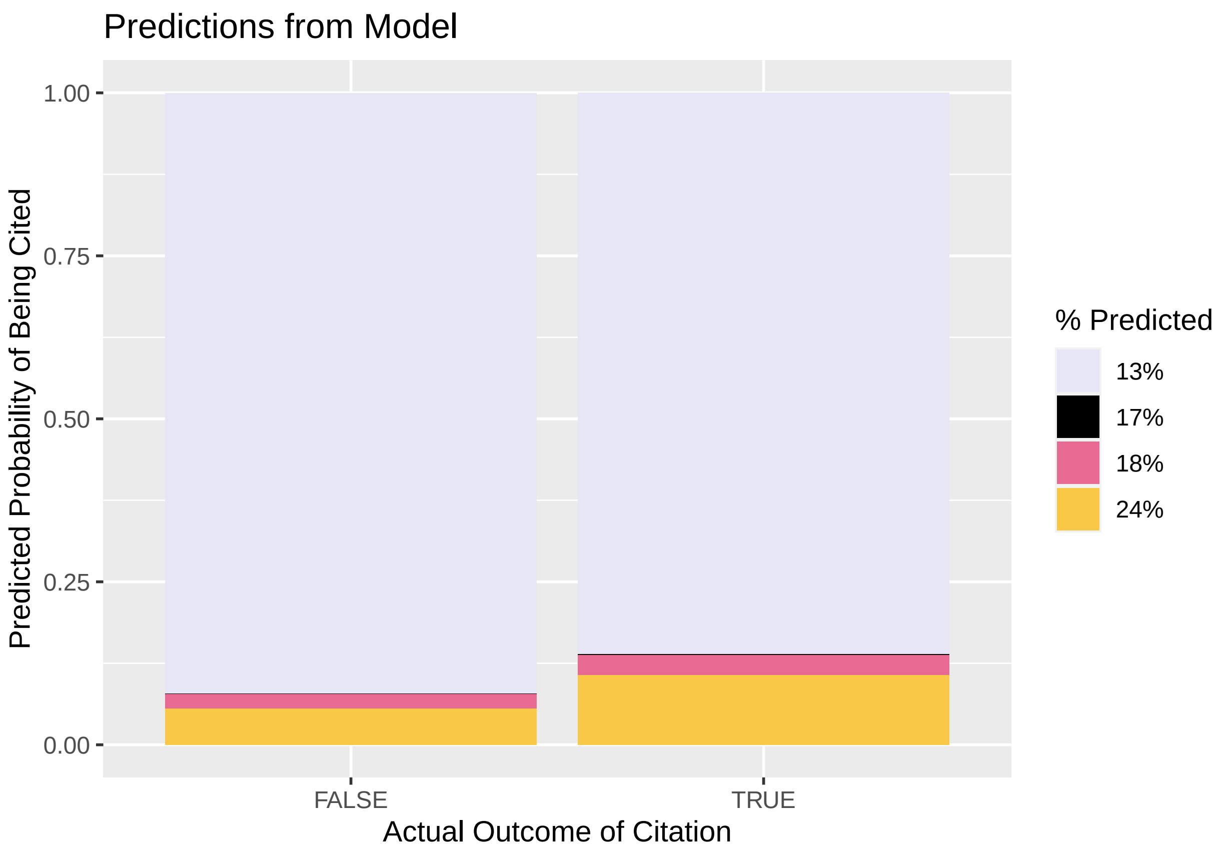
```r
confint(mod2)%>%
  exp()
```

```
##                          2.5 %    97.5 %
## (Intercept)          0.1492435 0.1514689
## frisk_performedTRUE  1.4071641 1.5244432
## search_conductedTRUE 1.3476825 1.4733999
```

```r
tidy(mod2)
```

```
## # A tibble: 3 x 5
##   term                 estimate std.error statistic  p.value
##   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            -1.89    0.00378    -502.  0
## 2 frisk_performedTRUE     0.382   0.0204       18.7 5.62e-78
## 3 search_conductedTRUE    0.343   0.0228       15.1 2.51e-51
```

```r
# Visualization

threshold <- 0.131

# predicted probability stacked relative frequency barchart
mod2 %>%
  augment(type.predict = 'response') %>%
  ggplot(aes(fill=factor(.fitted), x = factor(citation_issued))) +
  geom_bar(position = "fill") +
  ylab('Predicted Probability of Being Cited') +
  scale_fill_manual(labels = c("13%", "17%", "18%", "24%"), values = c("#e7e6f7", "black", "#e86a92", "#
  xlab('Actual Outcome of Citation') +
  ggtitle('Predictions from Model')
```



```
# predictions for model

mod2 %>%
  augment(type.predict = 'response') %>%
  mutate(predictCitation = .fitted >= threshold) %>%
  count(citation_issued, predictCitation)
```

```
## # A tibble: 4 x 3
##   citation_issued predictCitation      n
##   <lgl>           <lgl>            <int>
## 1 FALSE           FALSE           535636
## 2 FALSE           TRUE             45962
## 3 TRUE            FALSE            80517
## 4 TRUE            TRUE             13038
```