Analyzing Traffic Stops Data in North Carolina

Nicha Ruchirawat

August 1, 2017

```r
# load all libraries
library(ggplot2)
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

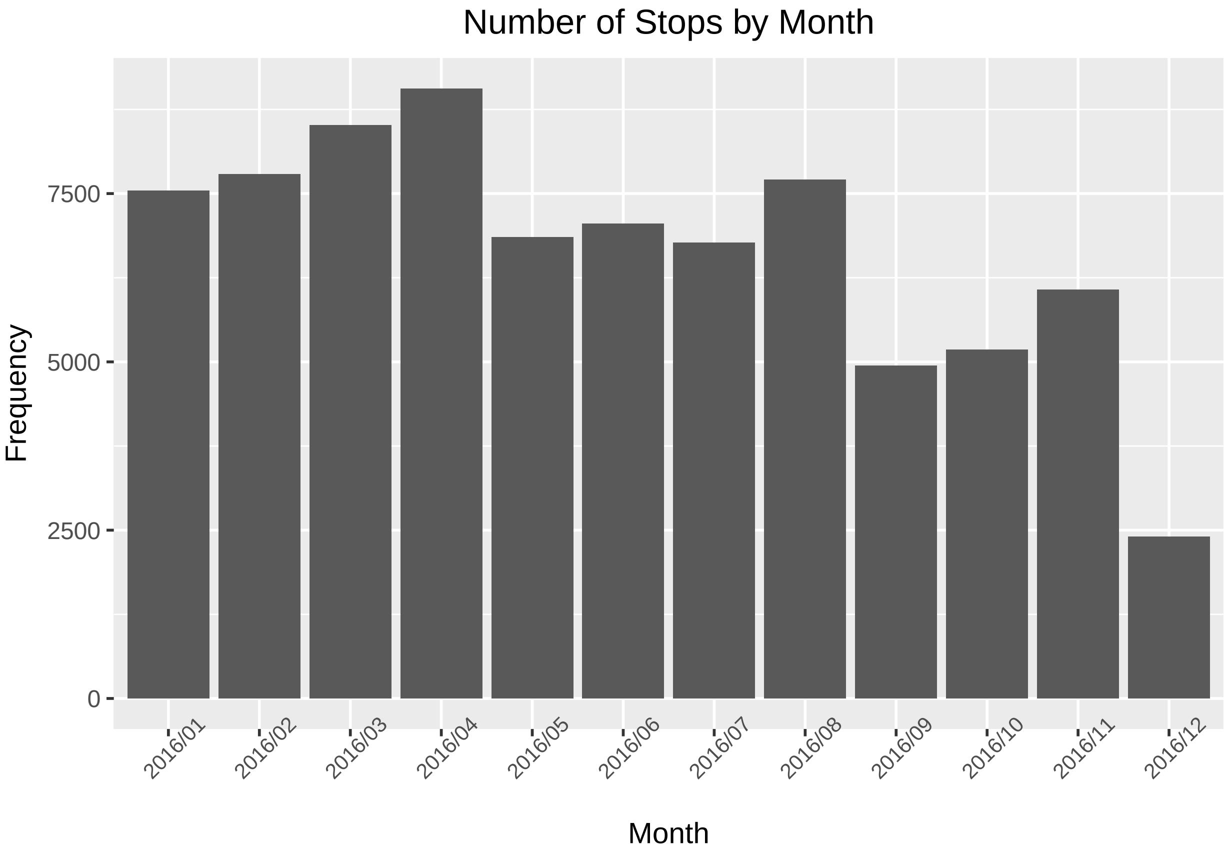
```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

This report is on exploratory data analysis of traffic stops data in North Carolina.

```r
traffic_stops <- read.csv("~/Downloads/Officer_Traffic_Stops.csv")
str(traffic_stops)
```

```
## 'data.frame':    79884 obs. of  17 variables:
##  $ Month_of_Stop           : Factor w/ 12 levels "2016/01","2016/02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Reason_for_Stop         : Factor w/ 10 levels "CheckPoint               ",..: 7 8 7 10 10 8 8 10
##  $ Officer_Race            : Factor w/ 8 levels " ","American Indian/Alaska Native",..: 8 8 8 4 8 8
##  $ Officer_Gender          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 1 2 ...
##  $ Officer_Years_of_Service: int  6 6 6 2 6 6 6 2 7 9 ...
##  $ Driver_Race             : Factor w/ 5 levels "Asian","Black",..: 5 2 2 2 5 5 5 2 2 5 ...
##  $ Driver_Ethnicity        : Factor w/ 2 levels "Hispanic","Non-Hispanic": 2 2 2 2 2 2 1 2 2 2 ...
##  $ Driver_Gender           : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 1 2 2 1 2 ...
##  $ Driver_Age              : int  63 35 30 29 45 65 40 28 57 29 ...
##  $ Was_a_Search_Conducted  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Result_of_Stop          : Factor w/ 5 levels "Arrest","Citation Issued",..: 2 4 2 4 2 2 2 4 2 1 .
##  $ CMPD_Division           : Factor w/ 14 levels "","Central Division",..: 3 3 3 7 3 3 5 7 7 14 ...
##  $ ObjectID                : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ CreationDate            : Factor w/ 560 levels "2016-12-20T23:49:27.408Z",..: 2 2 2 2 2 2 2 2 2 2
##  $ Creator                 : Factor w/ 1 level "charlottedata": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EditDate                : Factor w/ 560 levels "2016-12-20T23:49:27.408Z",..: 2 2 2 2 2 2 2 2 2 2
##  $ Editor                  : Factor w/ 1 level "charlottedata": 1 1 1 1 1 1 1 1 1 1 ...
```

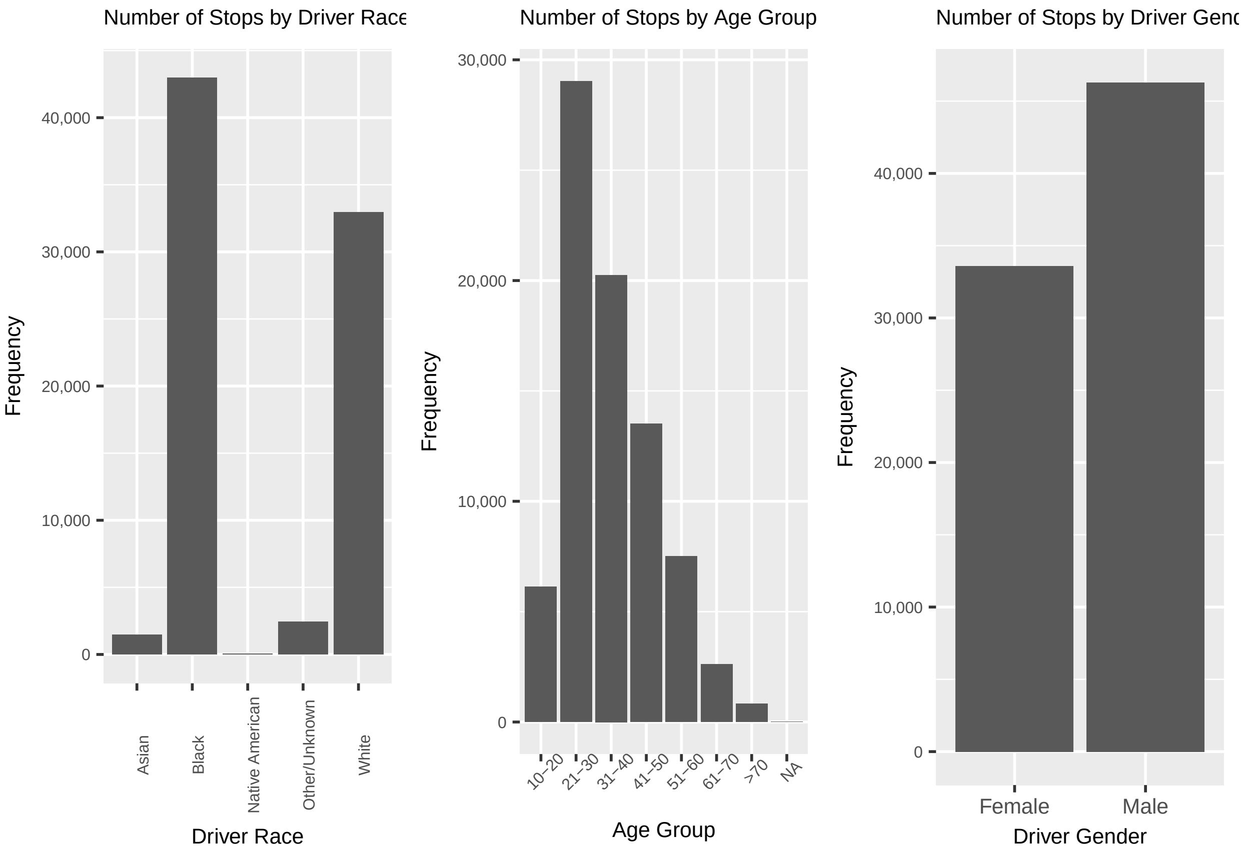
```r
# ggplot for number of stops grouped by month
traffic_stops %>% group_by(Month_of_Stop) %>% count(.) %>% ggplot(aes(Month_of_Stop,
    n)) + geom_bar(stat = "identity") + theme(axis.text.x = element_text(size = 8,
    angle = 45)) + labs(title = "Number of Stops by Month", x = "Month",
    y = "Frequency") + theme(plot.title = element_text(hjust = 0.5))
```



As shown by the graph above, the number of traffic stops seem cyclical every period of 4 months, peaking at every 4th month and falling at the start of the next period after, with the exception of December. There are more traffic stops at the start of the year, with a gradual decline in each period towards the end of the year.

```r
# ggplot for number of stops by Driver Race
a <- traffic_stops %>% group_by(Driver_Race) %>% count(.) %>% ggplot(aes(x = Driver_Race,
    y = n)) + geom_bar(stat = "identity") + labs(title = "Number of Stops by Driver Race",
    x = "Driver Race", y = "Frequency") + theme(axis.text.x = element_text(size = 6,
    angle = 90), axis.text.y = element_text(size = 6), plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)

# ggplot for number of stops by Age Group
b <- traffic_stops %>% mutate(age_group = cut(Driver_Age, breaks = c(10,
    20, 30, 40, 50, 60, 70, 99), labels = c("10-20", "21-30", "31-40",
    "41-50", "51-60", "61-70", ">70"), levels = c("10-20", "21-30",
    "31-40", "41-50", "51-60", "61-70", ">70"), ordered_result = T)) %>%
    group_by(age_group) %>% count(.) %>% ggplot(aes(x = age_group,
    y = n)) + geom_bar(stat = "identity") + labs(title = "Number of Stops by Age Group",
    x = "Age Group", y = "Frequency") + theme(axis.text.x = element_text(size = 6,
    angle = 45), axis.text.y = element_text(size = 6), plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)
```

```r
# ggplot for number of stops by Driver Gender
c <- traffic_stops %>% group_by(Driver_Gender) %>% count(.) %>% ggplot(aes(x = Driver_Gender,
    y = n)) + geom_bar(stat = "identity") + labs(title = "Number of Stops by Driver Gender",
    x = "Driver Gender", y = "Frequency") + theme(axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 6), plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)

# combine 3 plots into one side-to-side plot
gridExtra::grid.arrange(a, b, c, ncol = 3)
```



Driver demographics contributing to the most number of traffic stops:

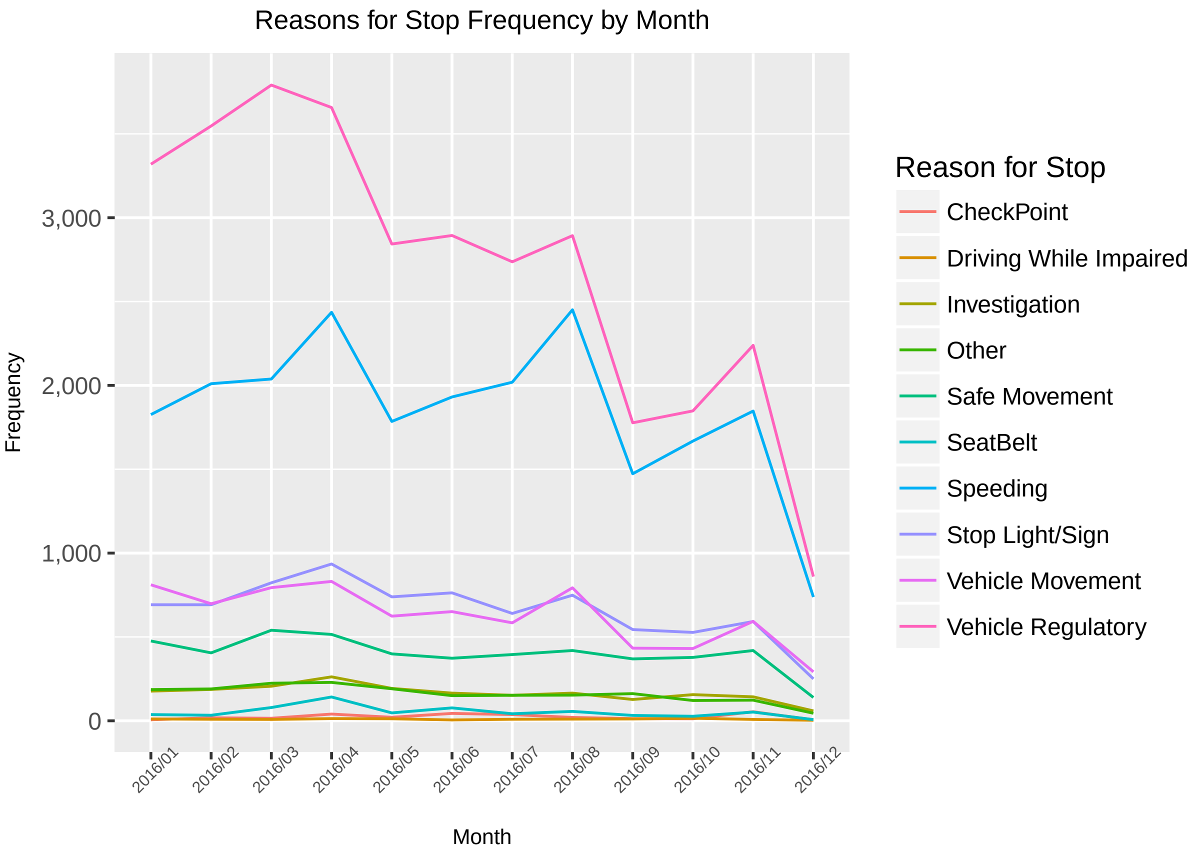
1) **Race**: Black dominates, followed by Whites
2) **Age**: Mostly 21-30 year olds, followed by 31-40 and 41-50 year olds
3) **Gender**: More males than females

```r
# create table of top 10 demographic profiles for number of stops
traffic_stops %>% mutate(age_group = cut(Driver_Age, breaks = c(10,
    20, 30, 40, 50, 60, 70, 99), labels = c("10-20", "21-30", "31-40",
    "41-50", "51-60", "61-70", ">70"), levels = c("10-20", "21-30",
    "31-40", "41-50", "51-60", "61-70", ">70"), ordered_result = T)) %>%
    group_by(Driver_Race, Driver_Gender, age_group) %>% count(.) %>%
    arrange(desc(n)) %>% ungroup() %>% slice(1:10) %>% knitr::kable(format.args = list(big.mark = ","),
    col.names = c("Driver Race", "Driver Gender", "Age Group", "Frequency"),
    align = "c")
```

Driver Race	Driver Gender	Age Group	Frequency
Black	Male	21-30	9,962
Black	Female	21-30	7,380
White	Male	21-30	6,133
Black	Male	31-40	5,810
Black	Female	31-40	5,209
White	Male	31-40	4,840
White	Female	21-30	4,259
Black	Male	41-50	3,819
White	Male	41-50	3,425
White	Female	31-40	3,273

In fact, the table above shows top 10 demographic profiles contributing to traffic stops. As shown, Blacks, both Females and Males, aged 21-30 are the biggest contributors. Blacks, both Females and Males, aged 31-40 are also substantial contributors. White Males and Females, aged 21-30, are also substantial contributors, with Males contributing slightly more than Females. Targeted efforts towards these top demographics can be useful when designing initiatives to lower traffic stops. In general, we should focus efforts towards Black and Whites aged 21-40 and Black and White Males aged 41-50.

```
# create scatter plot showing trend of number of stops by reason
# over the months
traffic_stops %>% group_by(Reason_for_Stop, Month_of_Stop) %>% count(.) %>%
    ggplot() + geom_line(aes(x = Month_of_Stop, y = n, color = Reason_for_Stop,
    group = Reason_for_Stop)) + labs(title = "Reasons for Stop Frequency by Month",
    x = "Month", y = "Frequency", color = "Reason for Stop") + theme(axis.text.x = element_text(size =
    angle = 45), plot.title = element_text(size = 10), axis.title = element_text(size = 8)) +
    theme(plot.title = element_text(hjust = 0.5)) + scale_y_continuous(labels = scales::comma)
```



The trend for Reasons for Stops by Month for Vehicle Regulatory and Speeding also follow the same cyclical trends as the total number of traffic stops over the months. These two reasons also remained dominant reasons for stops throughout the year. This suggests that number of traffic stops is mainly driven by these two reasons, as the other reasons have remained relatively low and stable throughout the year.

Therefore, we will investigate the demographics that contribute to number of the traffic stops from these two biggest reasons.

```
# create top 10 demographic profiles for stops due to Vehicle
# Regulatory
traffic_stops %>% filter(as.character(Reason_for_Stop) == "Vehicle Regulatory       ") %>%
    mutate(age_group = cut(Driver_Age, breaks = c(10, 20, 30, 40,
        50, 60, 70, 99), labels = c("10-20", "21-30", "31-40", "41-50",
        "51-60", "61-70", ">70"), levels = c("10-20", "21-30", "31-40",
        "41-50", "51-60", "61-70", ">70"), ordered_result = T)) %>%
    group_by(Driver_Race, Driver_Gender, age_group) %>% count(.) %>%
    arrange(desc(n)) %>% ungroup() %>% slice(1:10) %>% knitr::kable(format.args = list(big.mark = ","),
    col.names = c("Driver Race", "Driver Gender", "Age Group", "Frequency"),
    align = "c")
```

Driver Race	Driver Gender	Age Group	Frequency
Black	Male	21-30	4,632
Black	Female	21-30	3,591
Black	Male	31-40	2,869
Black	Female	31-40	2,664
White	Male	21-30	2,202
Black	Male	41-50	1,831
White	Male	31-40	1,746

Driver Race	Driver Gender	Age Group	Frequency
White	Female	21-30	1,584
Black	Female	41-50	1,511
White	Female	31-40	1,255

```
# create top 10 demographic profiles for stops due to Speeding
traffic_stops %>% filter(as.character(Reason_for_Stop) == "Speeding                  ") %>%
    mutate(age_group = cut(Driver_Age, breaks = c(10, 20, 30, 40,
        50, 60, 70, 99), labels = c("10-20", "21-30", "31-40", "41-50",
        "51-60", "61-70", ">70"), levels = c("10-20", "21-30", "31-40",
        "41-50", "51-60", "61-70", ">70"), ordered_result = T)) %>%
    group_by(Driver_Race, Driver_Gender, age_group) %>% count(.) %>%
    arrange(desc(n)) %>% ungroup() %>% slice(1:10) %>% knitr::kable(format.args = list(big.mark = ","),
    col.names = c("Driver Race", "Driver Gender", "Age Group", "Frequency"),
    align = "c")
```

Driver Race	Driver Gender	Age Group	Frequency
Black	Male	21-30	2,020
White	Male	21-30	1,776
Black	Female	21-30	1,720
White	Male	31-40	1,488
White	Female	21-30	1,342
Black	Female	31-40	1,228
Black	Male	31-40	1,138
White	Female	31-40	1,102
White	Male	41-50	1,079
White	Female	41-50	979

The above demographics align with the overall top demographics contributing to traffic stop, with some demographics more pronounced in each reason. For example, Blacks aged 21-40 dominate Vehicle Regulatory, while Black and White Males aged 21-30 dominate Speeding. Perhaps, initiatives can be slightly tailored to these target groups to lower traffic stops for each reason appropriately.

```
# create Result of Stops for stops due to Vehicle Regulatory
traffic_stops %>% filter(as.character(Reason_for_Stop) == "Vehicle Regulatory       ") %>%
    group_by(Result_of_Stop) %>% count(.) %>% arrange(desc(n)) %>%
    knitr::kable(format.args = list(big.mark = ","), col.names = c("Result of Stop",
        "Frequency"), align = c("l", "c"))
```

Result of Stop	Frequency
Verbal Warning	17,186
Citation Issued	12,827
No Action Taken	951
Written Warning	929
Arrest	512

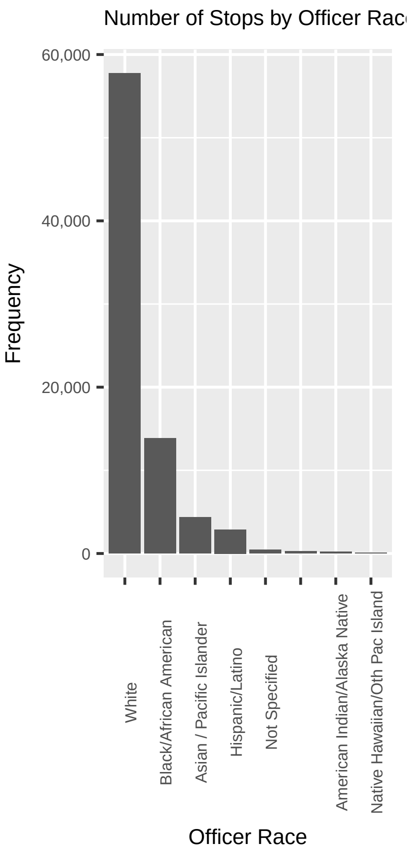
```
# create Result of Stops for stops due to Speeding
traffic_stops %>% filter(as.character(Reason_for_Stop) == "Speeding                  ") %>%
    group_by(Result_of_Stop) %>% count(.) %>% arrange(desc(n)) %>%
    knitr::kable(format.args = list(big.mark = ","), col.names = c("Result of Stop",
```

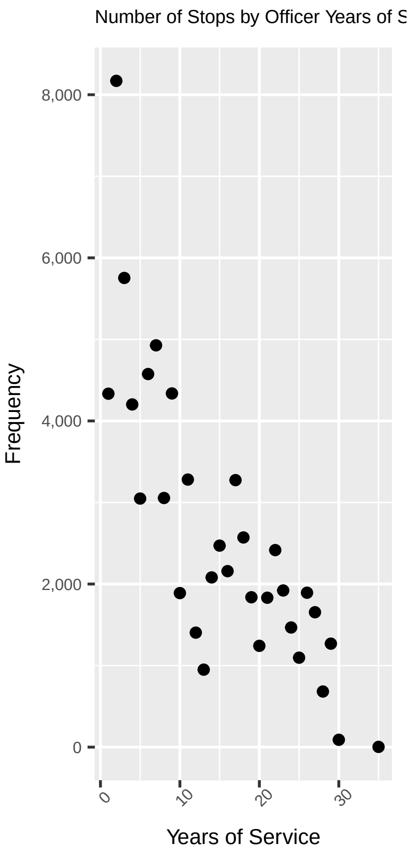
```r
        "Frequency"), align = c("l", "c"))
```

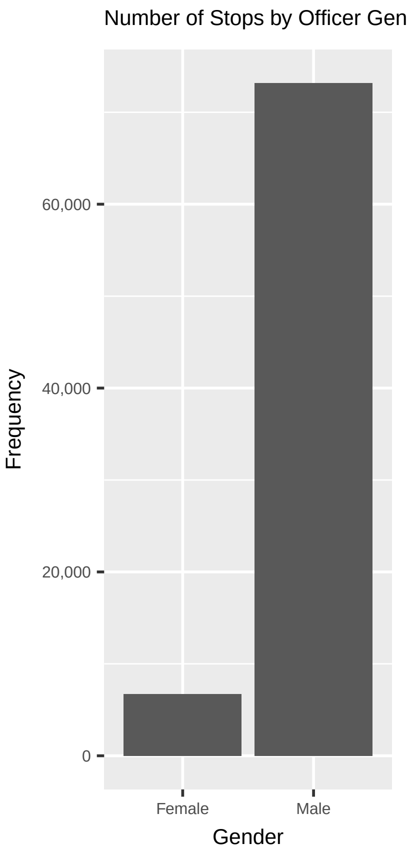
Result of Stop	Frequency
Citation Issued	13,693
Verbal Warning	6,884
Written Warning	1,350
Arrest	203
No Action Taken	92

Top result of stop for Vehicle Regulatory and Speeding are Citation Issued and Verbal Warning. This can be useful to evaluate effectiveness of action taken on traffic stops on lowering occurences of these types of stops. For example, being stricter with action than merely verbal warning in Vehicle Regulatory will decrease stops for this reason.

```r
# create frequency histogram for number of stops by officer race
d <- traffic_stops %>% group_by(Officer_Race) %>% count(.) %>% ggplot(aes(x = reorder(Officer_Race,
    -n), y = n)) + geom_bar(stat = "identity") + labs(title = "Number of Stops by Officer Race",
    x = "Officer Race", y = "Frequency") + theme(axis.text.x = element_text(size = 6,
    angle = 90), axis.text.y = element_text(size = 6), plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)

# create scatter plot for number of stops by officer years of age
e <- traffic_stops %>% group_by(Officer_Years_of_Service) %>% count(.) %>%
    ggplot() + geom_point(aes(x = Officer_Years_of_Service, y = n)) +
    labs(title = "Number of Stops by Officer Years of Service", x = "Years of Service",
        y = "Frequency") + theme(axis.text.x = element_text(size = 6,
    angle = 45), axis.text.y = element_text(size = 6), plot.title = element_text(size = 6.5),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)

# create histogram for number of stops by officer gender
f <- traffic_stops %>% group_by(Officer_Gender) %>% count(.) %>% ggplot(aes(x = Officer_Gender,
    y = n)) + geom_bar(stat = "identity") + labs(title = "Number of Stops by Officer Gender",
    x = "Gender", y = "Frequency") + theme(axis.text.x = element_text(size = 6),
    axis.text.y = element_text(size = 6), plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)

# combine 3 plots into side-by-side plot
gridExtra::grid.arrange(d, e, f, ncol = 3)
```





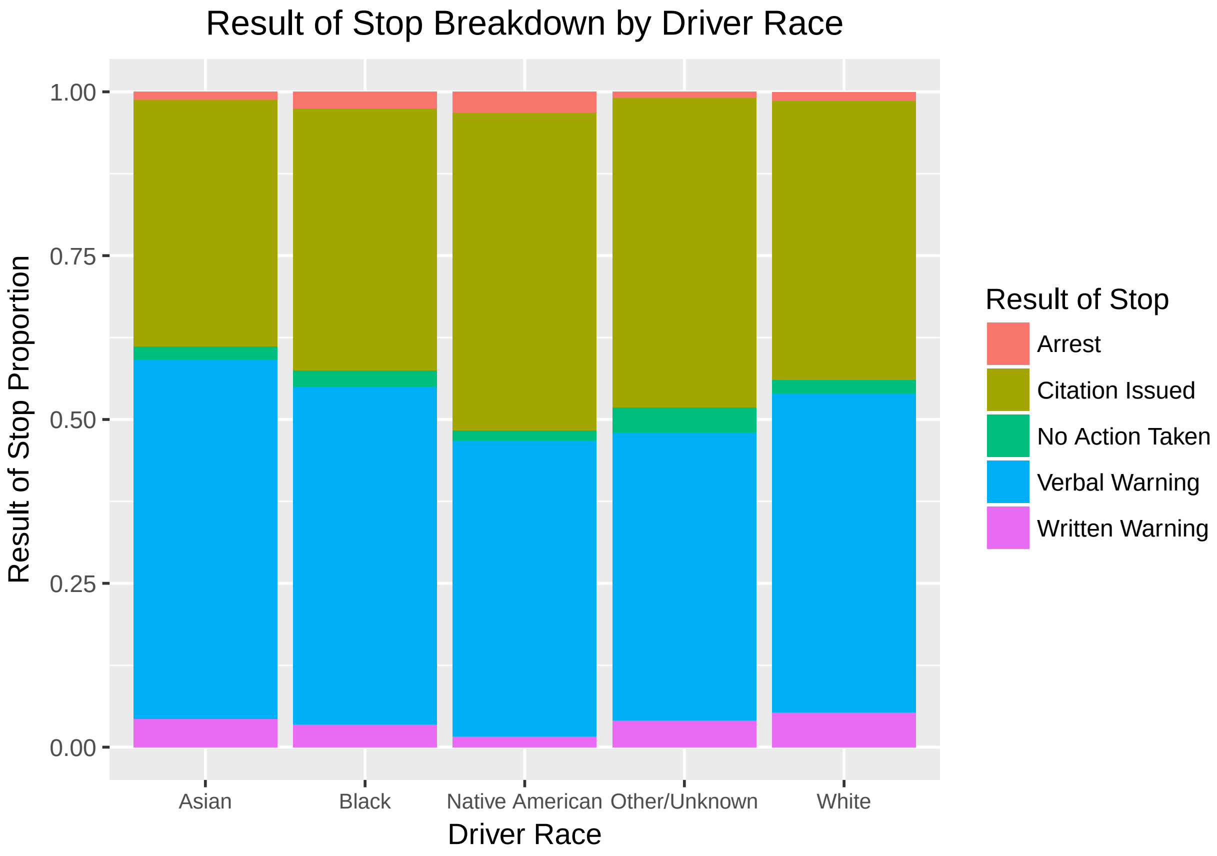


Demographics of officers contributing to most traffic stops are:

1) **Race**: Whites mostly, followed by Blacks
2) **Years of Service**: Most stops are from officers with lower years of service. We might want to investigate whether this is due to there being more younger officers, or if younger officers are stricter.
3) **Gender**: Most stops are from Male officers

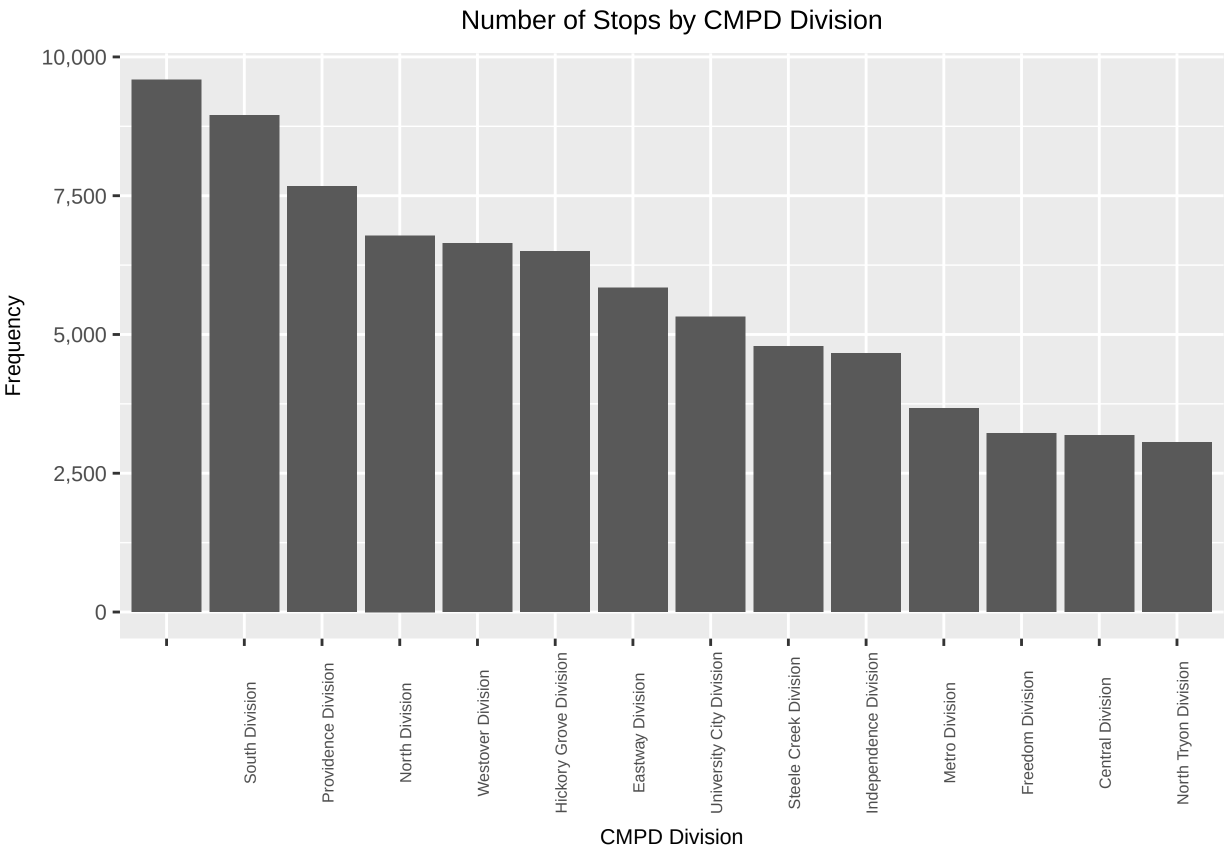
To evaluate whether there is racial bias in an officer taking action on a stop towards various driver races:

```r
# create plot breaking down result of stop by Driver Race
traffic_stops %>% group_by(Result_of_Stop, Driver_Race) %>% count(.) %>%
    ggplot(aes(fill = Result_of_Stop, y = n, x = Driver_Race)) + geom_bar(stat = "identity",
    position = "fill") + theme(axis.text.x = element_text(size = 8)) +
    labs(title = "Result of Stop Breakdown by Driver Race", x = "Driver Race",
        y = "Result of Stop Proportion", fill = "Result of Stop") +
    theme(plot.title = element_text(hjust = 0.5))
```



It seems that officers have been slightly stricter with Native Americans with stop actions such as Citation Issue and Arrest than other groups. We might want to investigate why.

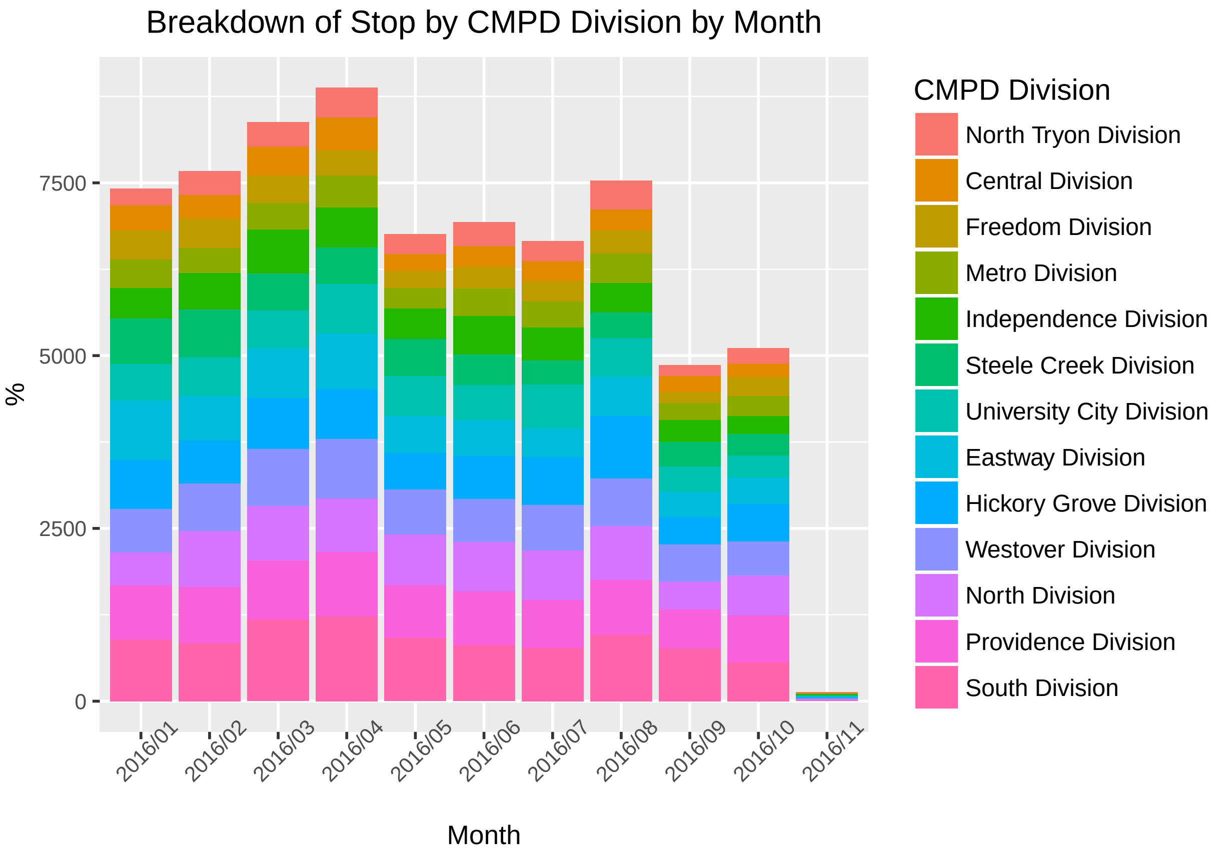
```
# create histogram of number of stops by CMPD Division
traffic_stops %>% group_by(CMPD_Division) %>% count(.) %>% ggplot(aes(x = reorder(CMPD_Division,
    -n), y = n)) + geom_bar(stat = "identity") + theme(axis.text.x = element_text(size = 8,
    angle = 90)) + labs(title = "Number of Stops by CMPD Division",
    x = "CMPD Division", y = "Frequency") + theme(axis.text.x = element_text(size = 6,
    angle = 90), axis.text.y = element_text(size = 8), plot.title = element_text(size = 10,
    hjust = 0.5), axis.title = element_text(size = 8)) + scale_y_continuous(labels = scales::comma)
```



We can also target CMPD Divisions that contribute the most to traffic stops, top 5 being:

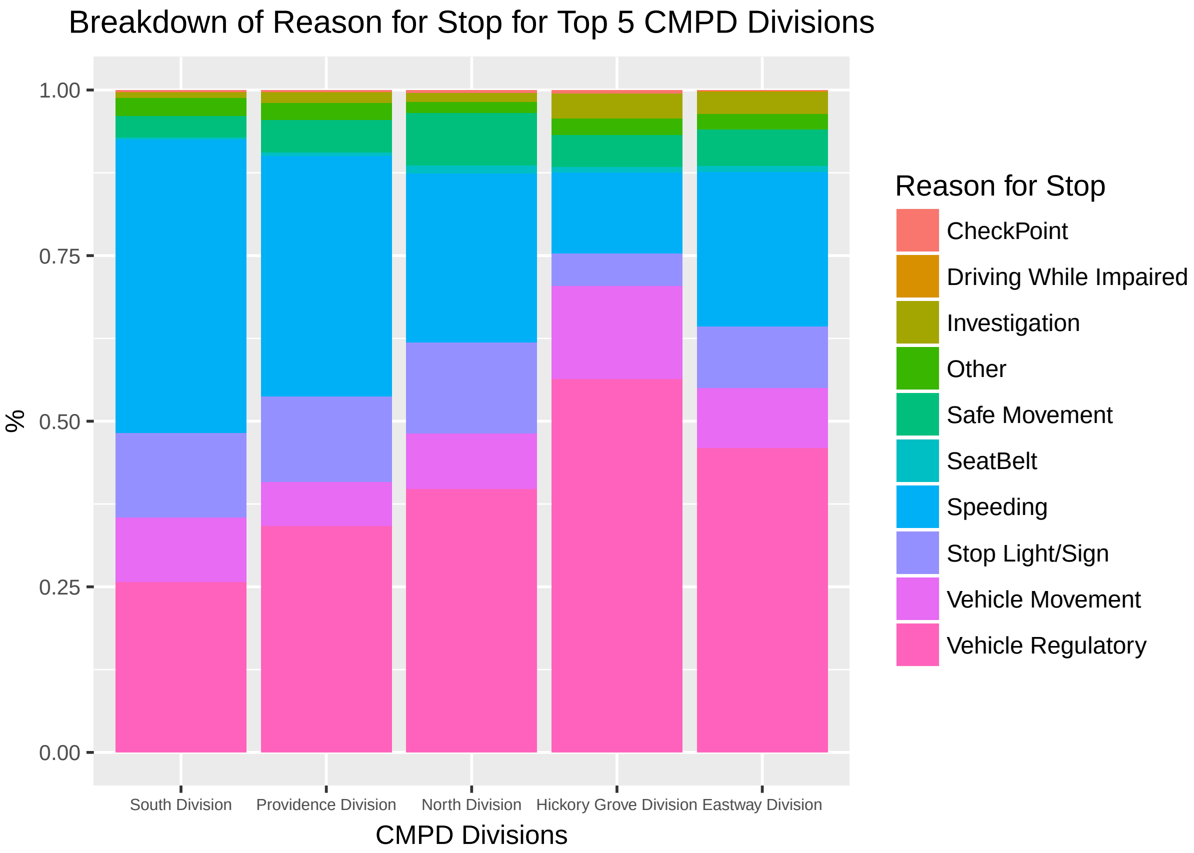
1) South Division
2) Providence Division
3) North Division
4) Westover Division
5) Hickory Grove Division

```
# create 100% plots of number of stops by CMPD Division over the
# months
traffic_stops %>% filter(as.character(CMPD_Division) != "") %>% group_by(CMPD_Division,
    Month_of_Stop) %>% count(.) %>% ggplot(aes(fill = reorder(CMPD_Division,
    n), y = n, x = Month_of_Stop)) + geom_bar(stat = "identity") +
    labs(title = "Breakdown of Stop by CMPD Division by Month", x = "Month",
        y = "%", fill = "CMPD Division") + theme(axis.text.x = element_text(size = 8,
    angle = 45), axis.text.y = element_text(size = 8), plot.title = element_text(size = 12,
    hjust = 0.5), axis.title = element_text(size = 10))
```



The breakdown by month, however, show that certain CMPD Divisions such as Eastway Division, Independence Division, and University City Division grew in stop traffics particularly in December. These may need attention as well as to why it grew substantially towards the year's end.

```
# create 100% plots of number of stops by reason of stop for top 5
# CMPD Divisions
traffic_stops %>% filter(as.character(CMPD_Division) == "South Division" |
    as.character(CMPD_Division) == "Providence Division" | as.character(CMPD_Division) ==
    "North Division" | as.character(CMPD_Division) == "Hickory Grove Division" |
    as.character(CMPD_Division) == "Eastway Division") %>% group_by(Reason_for_Stop,
    CMPD_Division) %>% count(.) %>% ggplot(aes(fill = Reason_for_Stop,
    y = n, x = reorder(CMPD_Division, -n))) + geom_bar(stat = "identity",
    position = "fill") + labs(title = "Breakdown of Reason for Stop for Top 5 CMPD Divisions",
    x = "CMPD Divisions", y = "%", fill = "Reason for Stop") + theme(axis.text.x = element_text(size =
    axis.text.y = element_text(size = 8), plot.title = element_text(size = 12,
        hjust = 0.5), axis.title = element_text(size = 10))
```



Each top CMPD Divisions has different distributions of reasons for stop traffic. For example, Hickory Grove Division is dominated by Vehicle Regulatory, while South Division is dominated by Speeding. Actions can be taken differently towards these divisions to lower stops for their dominant reasons.