Traffic Exploration

Louise Lai

August 4, 2018

```
# load files
setwd("~/Desktop/programming/eda/homework/hw3")
original <- as.data.frame(read.csv("./Officer_Traffic_Stops.csv", na.strings = c("", " ", "  ", "NA")))
```

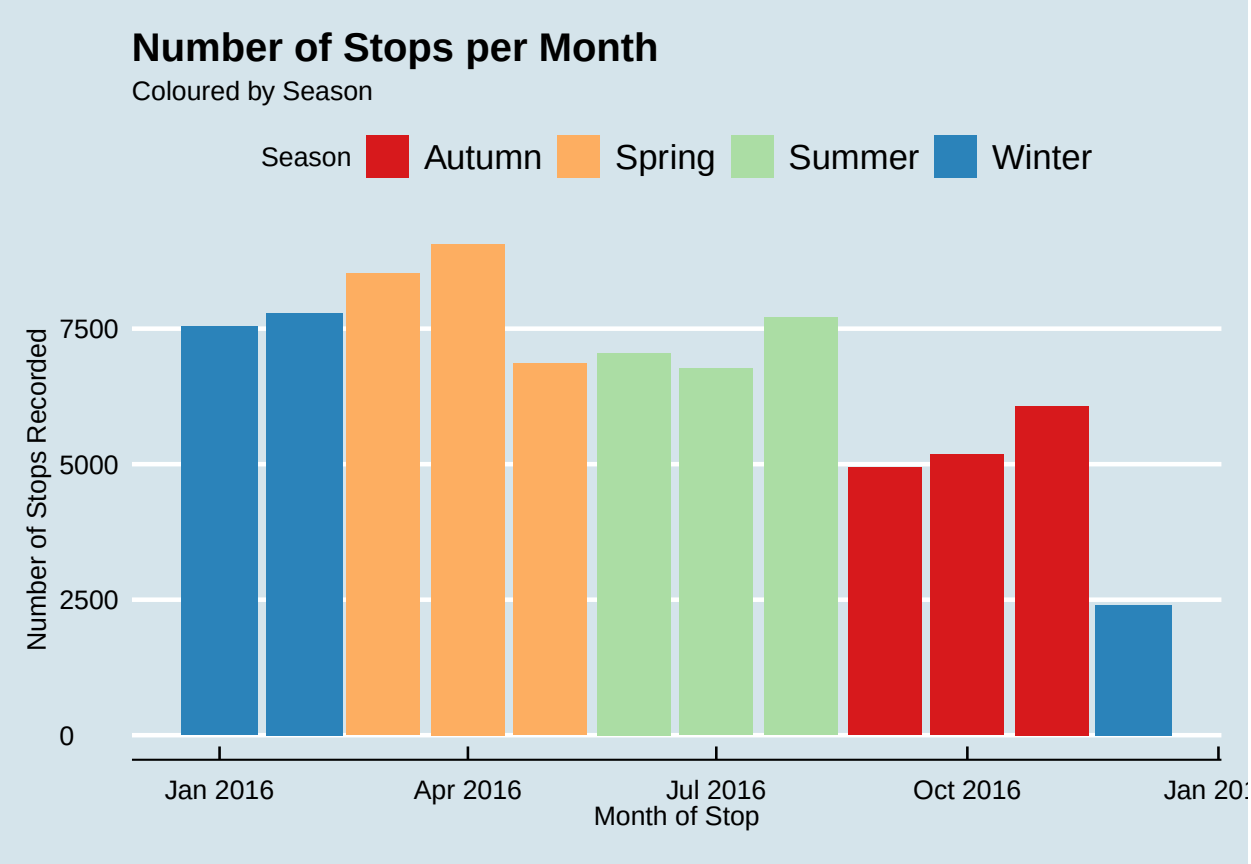
Data Cleaning

```
df <- original

##### Data Cleaning

returnCleanData <- function(){

  # count NAs
  anyNA(df) # TRUE
  colnames(df)[apply(df, 2, anyNA)] # returns colnames that have NAs

  ## NAs
  df$Officer_Race <- as.character(df$Officer_Race)
  df %<>% mutate(Officer_Race = replace(Officer_Race, is.na(Officer_Race), "Missing")) # replace
  df$Officer_Race <- as.factor(df$Officer_Race)

  # unsure what to do with CMPD_Division NAs yet!

  ## Dates
  df$Month_of_Stop <- parse_date(df$Month_of_Stop, "%Y/%m")

  ## Logicals
  levels(df$Was_a_Search_Conducted) <- c(FALSE, TRUE) # set 'No' to FALSE, 'Yes' to TRUE
  df$Was_a_Search_Conducted <- as.logical(df$Was_a_Search_Conducted)

  ## Order Factors
  levels(df$Result_of_Stop) # come back~ <<

  # Change White/Hispanic
  df %>%
    filter(Driver_Ethnicity == "Hispanic") %>%
    nrow() # There are 6,623 drivers that are White and Hispanic. 7,578 total == 955 unidentified other

  df$Driver_Race <- as.character(df$Driver_Race)

  df %<>%
    mutate(Driver_Race = replace(Driver_Race, Driver_Ethnicity == "Hispanic", "Hispanic")) # replace Hi

  # Filter truly unnecessary columns
  df %<>%
```

```
    select(-13, -14, -15, -16, -17)

  return(df)
}

df <- returnCleanData()
```

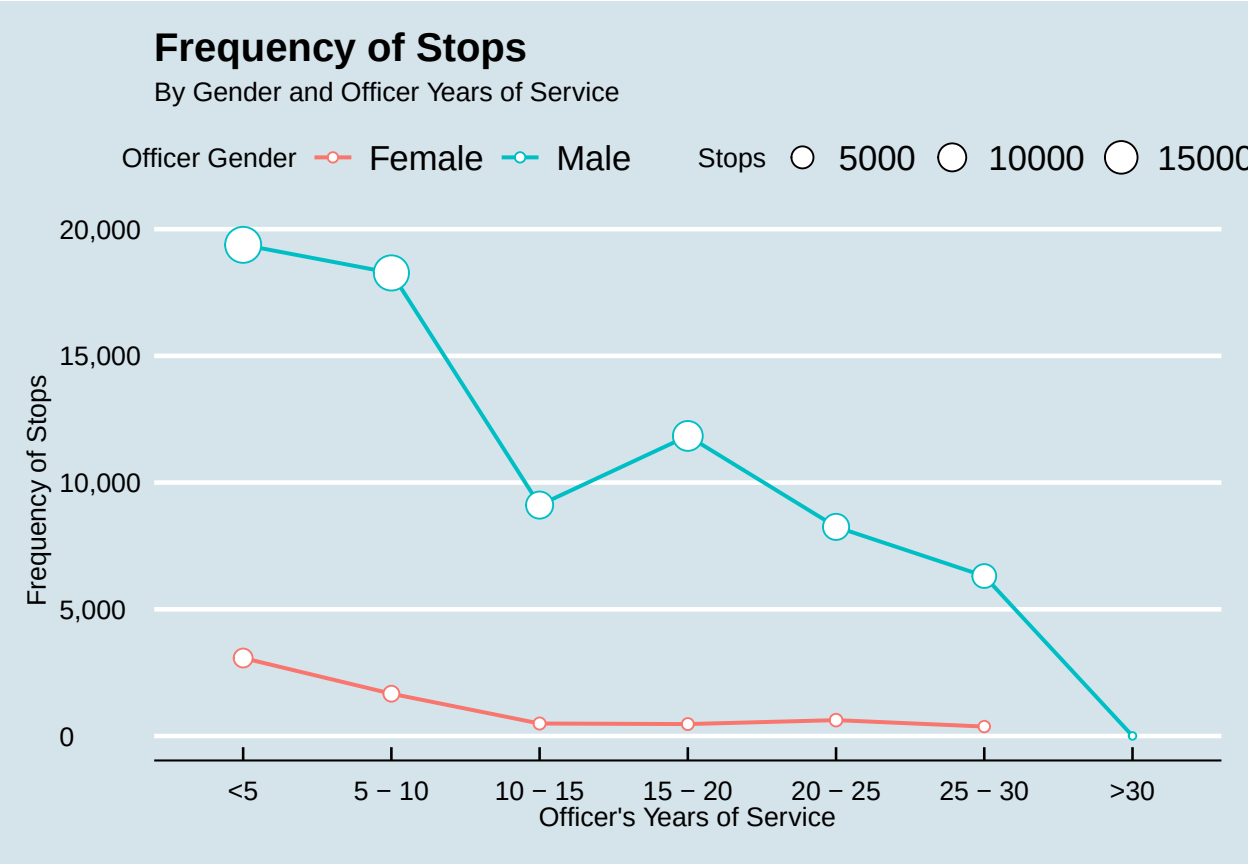
Exploring Months

```
# Frequency of Stops by Month
monthStops <- df %>%
  group_by(Month_of_Stop) %>%
  summarize(Freq=n())

df %>%
  mutate(season = case_when(Month_of_Stop == as.Date("2016-06-01") | Month_of_Stop == as.Date("2016-07-0
                            Month_of_Stop == as.Date("2016-03-01") | Month_of_Stop == as.Date("2016-04-0
                            Month_of_Stop == as.Date("2016-12-01") | Month_of_Stop == as.Date("2016-01-0
                            Month_of_Stop == as.Date("2016-09-01") | Month_of_Stop == as.Date("2016-10-0
  group_by(Month_of_Stop) %>%
  ggplot() +
    geom_bar(aes(Month_of_Stop, fill=factor(season))) +
    scale_fill_brewer(palette = "Spectral", name="Season") +
    xlab("Month of Stop") +
    ylab("Number of Stops Recorded") +
    labs(title= "Number of Stops per Month",
         subtitle="Coloured by Season") +
    theme_economist()
```



Exploring Years of Service

```
# years of service ~ average stops
serviceStops <- df %>% # creating bucket in 5 yr increments
  mutate(yearBucket = case_when(Officer_Years_of_Service < 5 ~ "<5",
                                (Officer_Years_of_Service >= 5) & (Officer_Years_of_Service < 10) ~ "5 -
                                (Officer_Years_of_Service >= 10) & (Officer_Years_of_Service < 15) ~ "1
                                (Officer_Years_of_Service >= 15) & (Officer_Years_of_Service < 20) ~ "1
                                (Officer_Years_of_Service >= 20) & (Officer_Years_of_Service < 25) ~ "2
                                (Officer_Years_of_Service >= 25) & (Officer_Years_of_Service <= 30) ~ "
                                Officer_Years_of_Service > 30 ~ ">30")) %>%
  select(Officer_Gender, yearBucket)

# order year buckets
serviceStops$yearBucket <- ordered(serviceStops$yearBucket, levels= c("<5", "5 - 10", "10 - 15", "15 - 

# avg stops
serviceStops %>%
  group_by(Officer_Gender, yearBucket) %>%
  summarize(stops = n()) %>%
  ggplot() +
    geom_line(aes(x=yearBucket, y=stops, group=Officer_Gender, color=Officer_Gender), size=.7) +
    geom_point(aes(x=yearBucket, y=stops, color=Officer_Gender, size=stops), shape=21, fill="white") +
    xlab("Officer's Years of Service") +
    ylab("Frequency of Stops") +
    scale_color_discrete(name = "Officer Gender") +
    #scale_color_manual(labels = c("Nope", "Yes, Busted!"), values = c("grey", "red3"), name="Was legal
    scale_size_continuous(name="Stops") +
    scale_y_continuous(labels=comma) +
    labs(title="Frequency of Stops", subtitle="By Gender and Officer Years of Service") +
    theme_economist() +
    theme(text=element_text(family="Helvetica"))
```

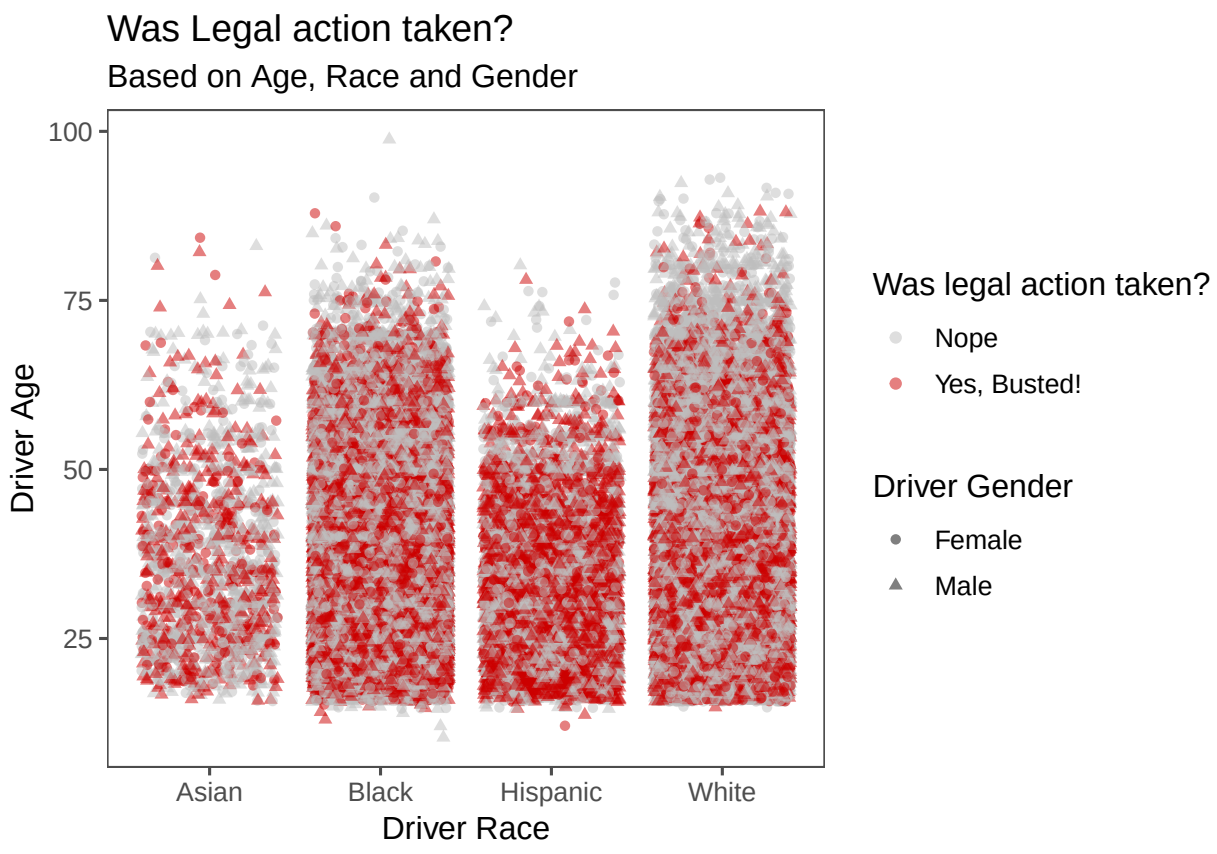

Frequency of Stops
By Gender and Officer Years of Service
Officer Gender
Female
Male
Stops
5000
10000
15000
20,000
15,000
10,000
5,000
0
Frequency of Stops
<5
5 − 10
10 − 15
15 − 20
20 − 25
25 − 30
>30
Officer's Years of Service

Exploring Legal Action

Not every stop leads to an arrest. Some stops can result in a mere verbal warning, with the most severe being an arrest on the spot. In this case study, we definte a legal action as receiving a citation or getting arrested.

These fields are found in the `Result_of_Stop` column.

```
# Exploring legal action by race, age and gender
ageGender <- df %>%
  filter(!Driver_Race == "Other/Unknown" & !Driver_Race == "Native American") %>%
  mutate(legal_action_taken = Result_of_Stop == "Citation Issued" | Result_of_Stop == "Arrest") %>%
  select(Driver_Age, Driver_Gender, Driver_Race, legal_action_taken)

ageGender %>%
  ggplot() +
  geom_point(aes(x=Driver_Race, y=Driver_Age, color=legal_action_taken, shape=Driver_Gender), alpha=0.5
  scale_color_manual(labels = c("Nope", "Yes, Busted!"), values = c("grey", "red3"), name="Was legal ac
  scale_shape_manual(values=c(1, 2)) +
  scale_shape(name = "Driver Gender") +
  xlab("Driver Race") +
  ylab("Driver Age") +
  labs(title="Was Legal action taken?", subtitle= "Based on Age, Race and Gender") +
  theme_few()
```

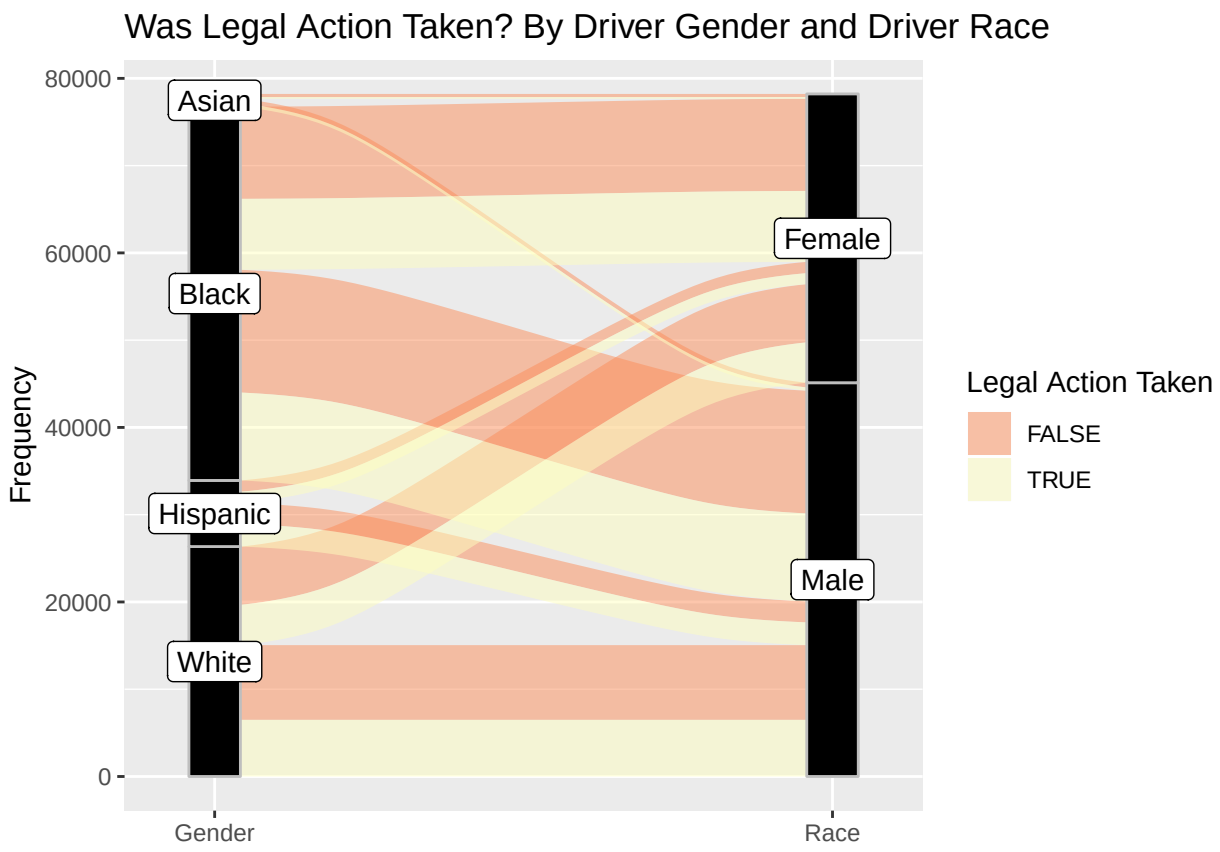


Exploring Relationships Between Columns

```
# Using alluvia, which is great for categoricla data, which we have a lot of
library(ggalluvial)

likelyAlluvia <- df %>%
  filter(!Driver_Race=="Other/Unknown" & !Driver_Race=="Native American") %>%
  mutate(legal_action_taken = Result_of_Stop == "Citation Issued" | Result_of_Stop == "Arrest") %>%
  group_by(legal_action_taken, Driver_Gender, Driver_Race) %>%
  summarise(Frequency=n())

is_alluvia_form(likelyAlluvia, axes=1:3, silent=TRUE)
```

```
## [1] TRUE
```

```
likelyAlluvia %>%
  ggplot(aes(y = Frequency, axis2 = Driver_Gender, axis1=Driver_Race)) +
  geom_alluvium(aes(fill = legal_action_taken), width = 1/12) +
  geom_stratum(width = 1/12, fill = "black", color = "grey") +
  geom_label(stat = "stratum", label.strata = TRUE) +
  scale_x_discrete(limits = c("Gender", "Race"), expand = c(.05, .05)) +
  scale_fill_brewer(type = "qual", palette = "Spectral",name="Legal Action Taken") +
  ggtitle("Was Legal Action Taken? By Driver Gender and Driver Race")
```




```
# What about just genders?
likelyAlluvia3 <- df %>%
  mutate(legal_action_taken = Result_of_Stop == "Citation Issued" | Result_of_Stop == "Arrest") %>%
```

```
  group_by(legal_action_taken, Officer_Gender, Driver_Gender) %>%
  summarise(Frequency=n())

is_alluvia_form(likelyAlluvia3, axes=1:3, silent=TRUE)
```

```
## [1] TRUE
```

```
likelyAlluvia3 %>%
  ggplot(aes(y = Frequency, axis1 = Officer_Gender, axis2=Driver_Gender )) +
  geom_alluvium(aes(fill = legal_action_taken), width = 1/12) +
  geom_stratum(width = 1/12, fill = "black", color = "grey") +
  geom_label(stat = "stratum", label.strata = TRUE) +
  scale_x_discrete(limits = c("Officer Gender", "Driver Gender"), expand = c(.1, .1)) +
  scale_fill_brewer(type = "qual", palette = "Set1", name="Legal Action Taken") +
  ggtitle("Was Legal Action Taken? By Officer and Driver Genders")
```