Homework 3

Question 1

Louise Lai

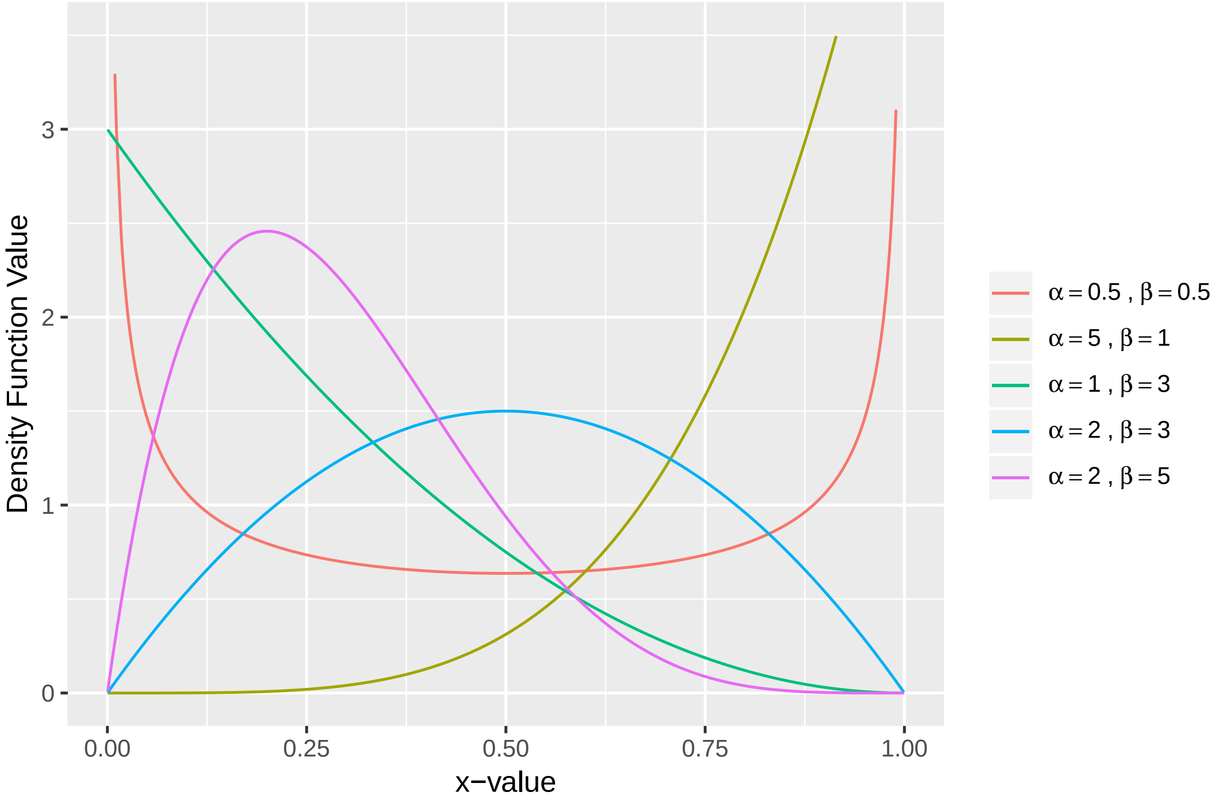
August 2, 2018

1.1

PDFs

```
# generate initial dataframe, arranged by smallest xval first
PDFs <- as.data.frame(runif(1000, 0.0, 1.0))
colnames(PDFs) <- "xValue"

PDFs %<>%
  arrange(xValue)

# store given alphas and betas
alphas <- c(0.5, 5, 1, 2, 2)
betas <- c(0.5, 1, 3, 2, 5)

# start filling DF!
generatePDFs <- function(df, alphaArray, betaArray){

  # loop through all 5 given alphas/betas
  for(i in 1:5){
    a <- alphaArray[i]
    b <- betaArray[i]

    # extract x values
    xVals <- df$xValue
    betaDistribution <- c()

    # start filling the distribution vector
    for(k in 1:length(xVals)){
       betaDistribution[k] <- dbeta(xVals[k], a, b)
    }

    # convert distribution vector into df column
    df[[i+1]] <- betaDistribution
  }
  names(df) <-  c("xValue", "B1", "B2", "B3", "B4", "B5")
  return(df)
}

PDFs <- generatePDFs(PDFs, alphas, betas)

PDFs %>%
  ggplot( aes(x=xValue)) +
    geom_line(aes(y=B1, color='1')) +
```

```r
  geom_line(aes(y=B2, color='2')) +
  geom_line(aes(y=B3, color='3')) +
  geom_line(aes(y=B4, color='4'))+
  geom_line(aes(y=B5, color='5')) +
  scale_color_discrete(name="", labels=c(bquote(alpha==0.5~","~beta==0.5),
                                         bquote(alpha==5~","~beta==1),
                                         bquote(alpha==1~","~beta==3),
                                         bquote(alpha==2~","~beta==3),
                                         bquote(alpha==2~","~beta==5))) +
xlab("x-value") +
ylab("Density Function Value") +
ylim(0, 3.5) +
labs(caption="Beta Distributions Governed by Alpha and Beta")
```

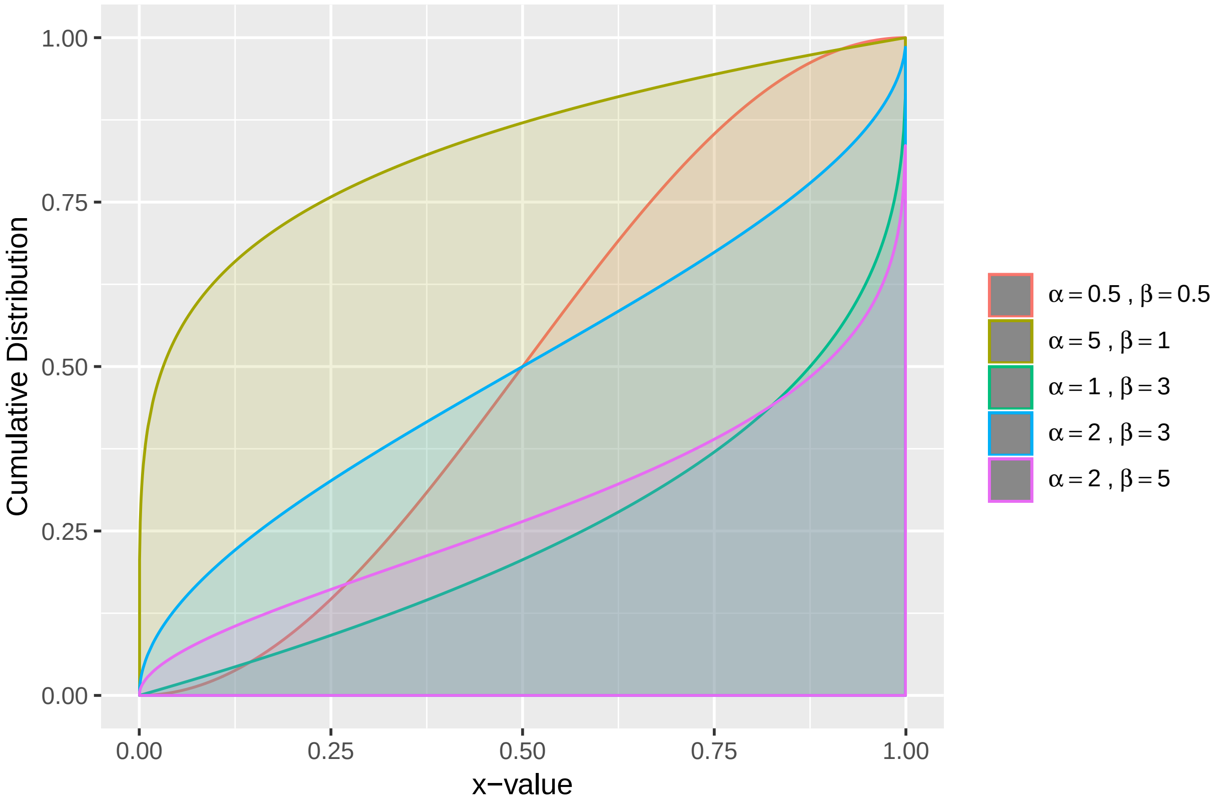


Beta Distributions Governed by Alpha and Beta

1.2

CDFs

```
generateCDFs <- function(df, alphaArray, betaArray){

  # loop through all 5 given alphas/betas
  for(i in 1:5){
    a <- alphaArray[i]
    b <- betaArray[i]

    # extract x values
    xVals <- df$xValue
    betaCumulative <- c()

    # start filling the cumulative vector
    for(k in 1:length(xVals)){
       betaCumulative[k] <- qbeta(xVals[k], a, b)
    }

    # convert distribution vector into df column
    df[[i+5+1]] <- betaCumulative
  }

  names(df) <-  c("xValue", "B1", "B2", "B3", "B4", "B5",
                  "BC1", "BC2", "BC3", "BC4", "BC5")
  return(df)
}

PDFCDF <- generateCDFs(PDFs, alphas, betas)

ggplot(data=PDFCDF, aes(x=xValue)) +
  geom_area(aes(y=BC1, color="1", fill="1"), alpha=0.15) +
  geom_area(aes(y=BC2, color="2", fill="2"), alpha=0.15) +
  geom_area(aes(y=BC3, color= "3", fill="3"), alpha=0.15) +
  geom_area(aes(y=BC4, color ="4", fill="4"), alpha=0.15) +
  geom_area(aes(y=BC5, color = "5", fill="5"), alpha=0.15) +
  scale_color_discrete(name="", labels=c(bquote(alpha==0.5~","~beta==0.5),
                                         bquote(alpha==5~","~beta==1),
                                         bquote(alpha==1~","~beta==3),
                                         bquote(alpha==2~","~beta==3),
                                         bquote(alpha==2~","~beta==5))) +
  scale_fill_discrete(guide=FALSE) +
  xlab("x-value") +
  ylab("Cumulative Distribution") +
  labs(caption="Cumulative Density of Beta Function Governed by Alpha and Beta")
```

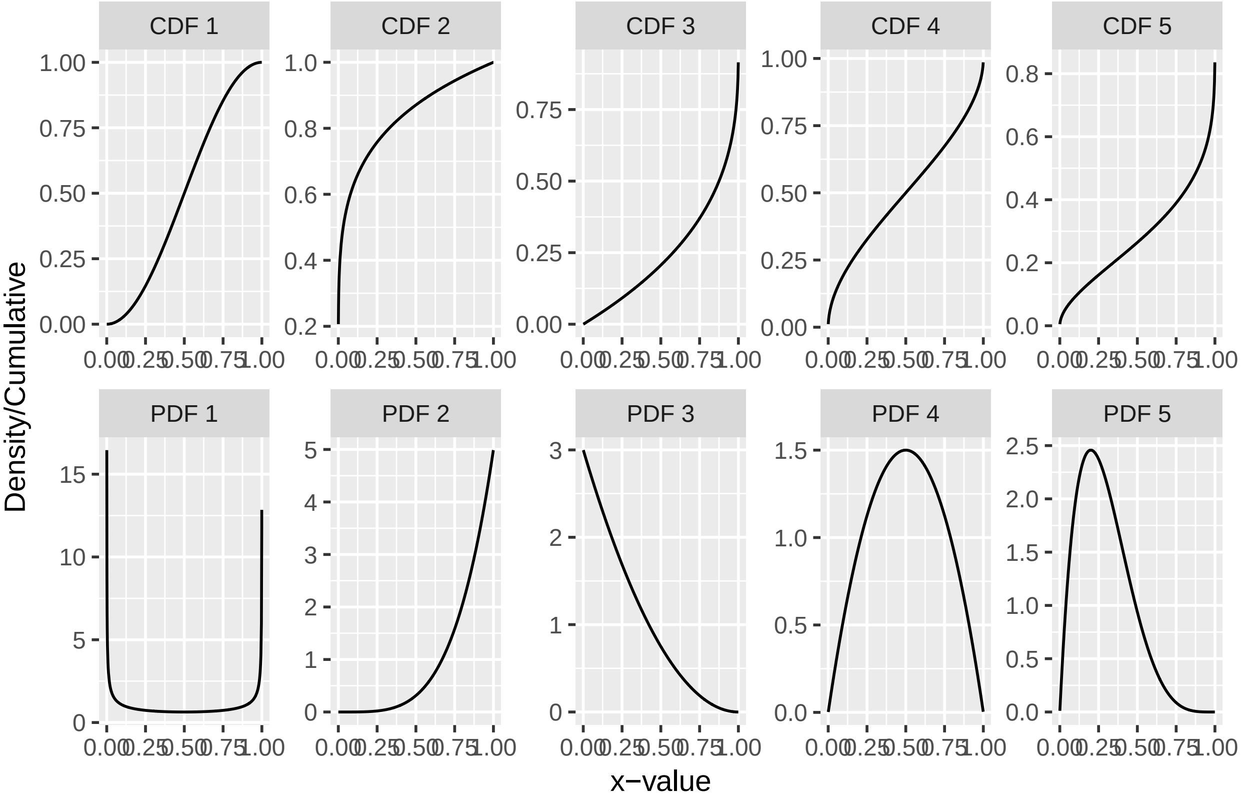


Cumulative Density of Beta Function Governed by Alpha and Beta

1.3

PDFs & CDFs

```
names(PDFCDF) <-  c("xValue", "PDF 1", "PDF 2", "PDF 3", "PDF 4", "PDF 5",
                    "CDF 1", "CDF 2", "CDF 3", "CDF 4", "CDF 5")

PDFCDF %>%
  gather(-xValue, key='var', value="value") %>%
  ggplot(aes(x=xValue, y = value)) +
    geom_line() +
    scale_color_discrete(name="", labels=c(bquote(alpha==0.5~","~beta==0.5),
                                           bquote(alpha==5~","~beta==1),
                                           bquote(alpha==1~","~beta==3),
                                           bquote(alpha==2~","~beta==3),
                                           bquote(alpha==2~","~beta==5))) +
  facet_wrap(~ var, scales="free", ncol=5) +
  xlab("x-value") +
  ylab("Density/Cumulative") +
  labs(caption="PDFs and CDFs of Beta Governed by Alpha and Beta")
```



PDFs and CDFs of Beta Governed by Alpha and Beta

Homework 3

Question 2: Traffic Exploration

Louise Lai

August 4, 2018

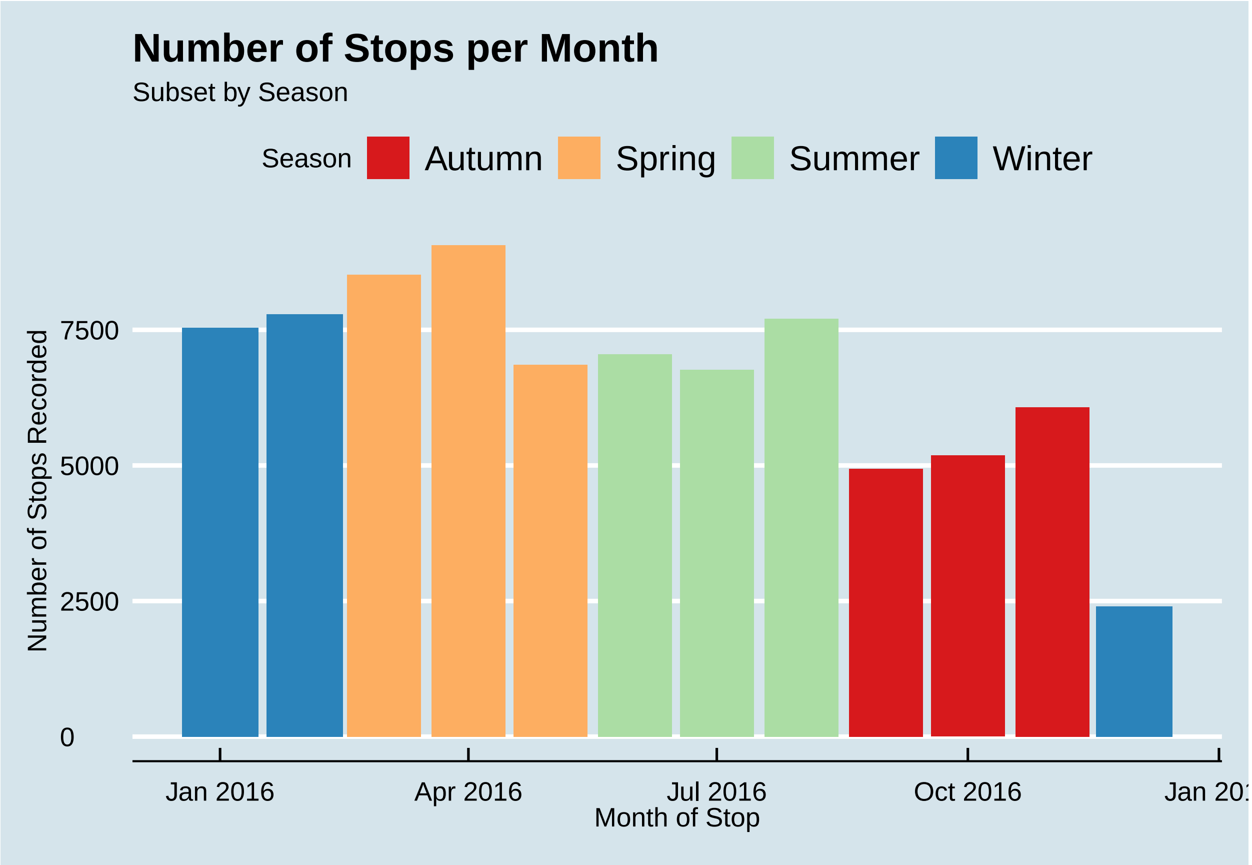
```r
# load files
setwd("~/Desktop/programming/eda/homework/hw3")
original <- as.data.frame(read.csv("./Officer_Traffic_Stops.csv", na.strings = c("", " ", "  ", "NA")))
```

Data Cleaning

```r
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-
                            Month_of_Stop == as.Date("2016-03-01") | Month_of_Stop == as.Date("2016-04-
                            Month_of_Stop == as.Date("2016-12-01") | Month_of_Stop == as.Date("2016-01-
                            Month_of_Stop == as.Date("2016-09-01") | Month_of_Stop == as.Date("2016-10-
  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="Subset by Season") +
    theme_economist()
```

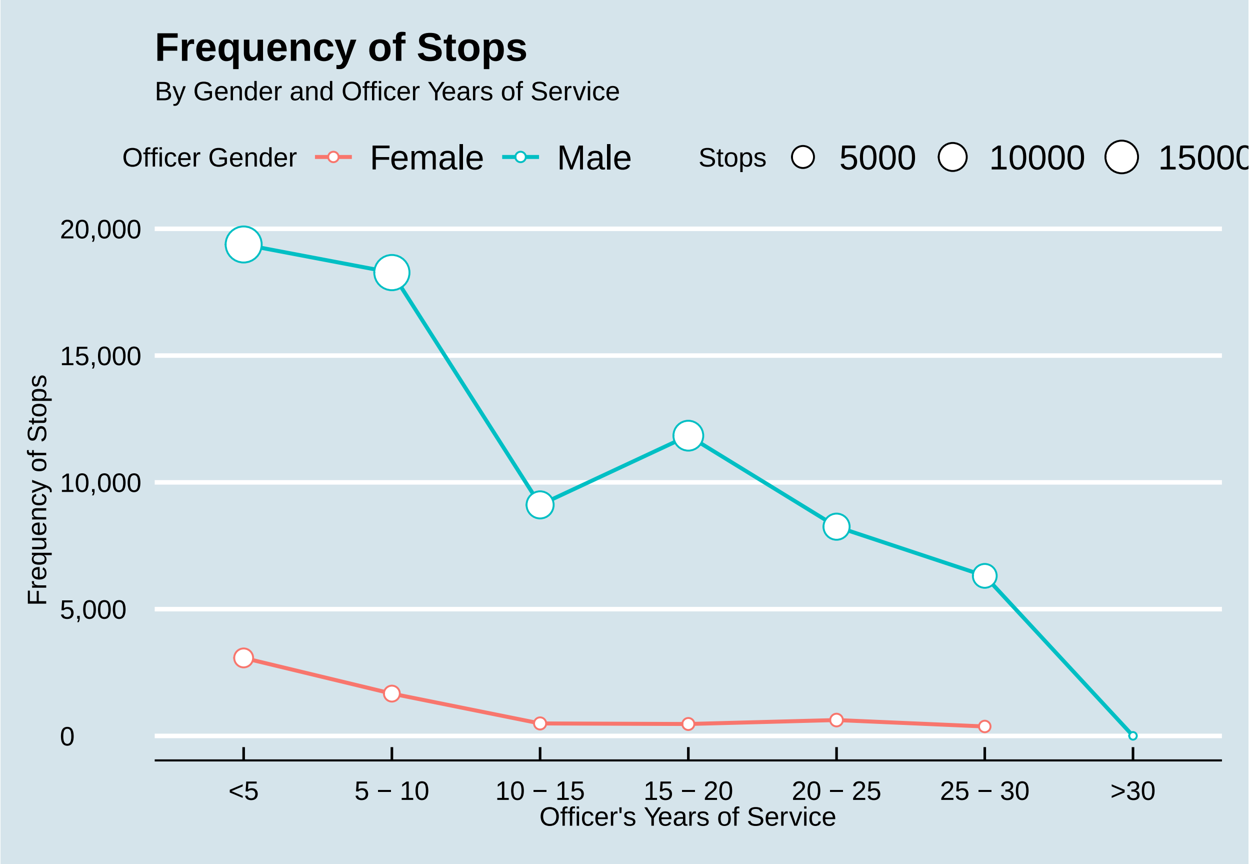


The Spring and Summer months have the highest levels of stops recorded, and it appears that Autumn and Winter are, on average, the lowest months.

This could be becuase people are more likely to drive around when the weather is nice, such as in the Spring or the end of Winter. Conversely, at the start of Winter, there are far fewer stops made.

Exploring Years of Service

```r
# 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) ~ "10
                                (Officer_Years_of_Service >= 15) & (Officer_Years_of_Service < 20) ~ "15
                                (Officer_Years_of_Service >= 20) & (Officer_Years_of_Service < 25) ~ "20
                                (Officer_Years_of_Service >= 25) & (Officer_Years_of_Service <= 30) ~ "2
                                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 - 2

# 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"))
```



We can clearly see that the bulk of stops are made by younger officers, and the least by senior officers. This could be due to the enthusiasm of younger officers, or that senior officers are doing more important work.

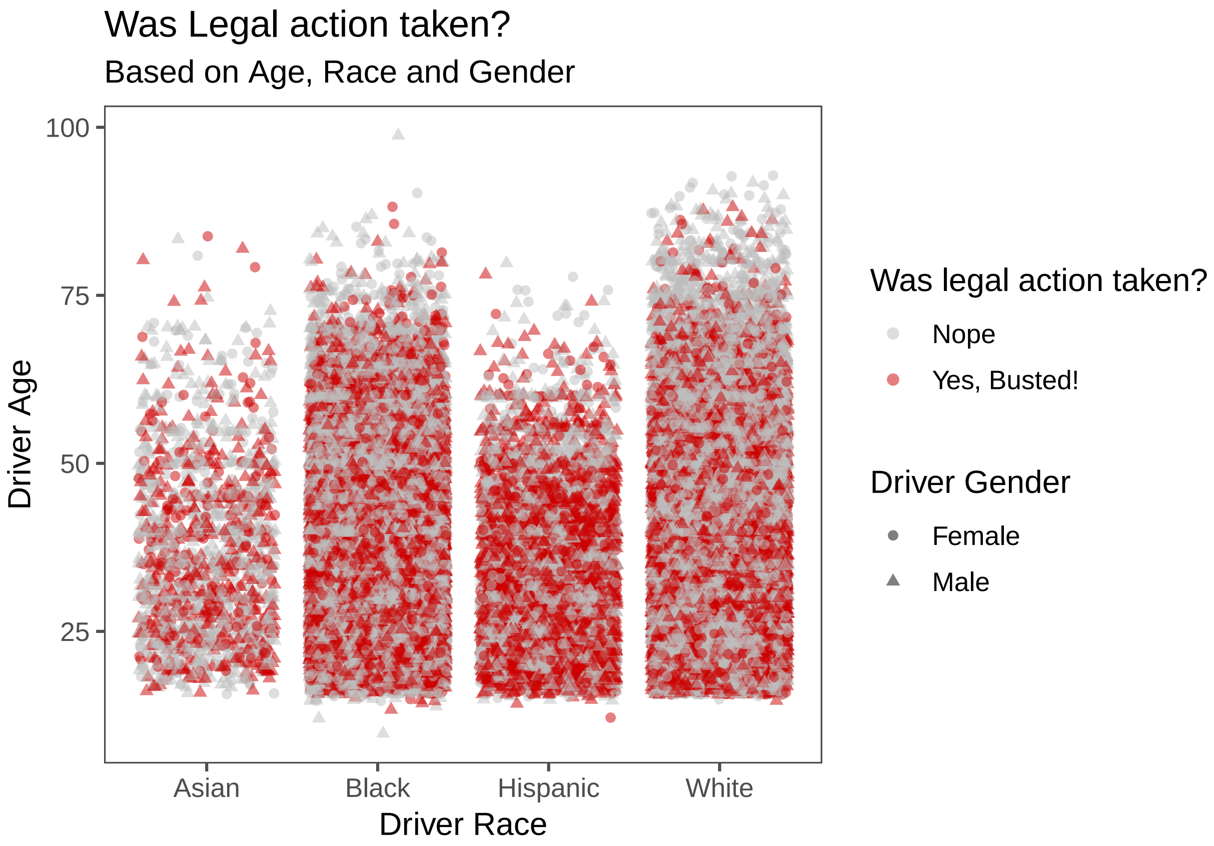
The *average* number of stops made would be an ideal statistic; however, though seemingly straightforward, there is no actual way to derive this information from the given dataset because we were only given the number of stops that were made. In order to get an average, we would need to count the total number of officers - no only officers that made a stop - in a particular year bucket and take that as our denominator to get the average.

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.

```r
# 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()
```



On a whole, we see that white drivers that are stopped are generally older than other drivers. It appears that older Asian and Hispanic people are rarely getting stopped!

There could be a host of reasons for this. Perhaps there are just fewer Asian and Hispanic people over the age of 50, and this sample reflects the general population. Other theories include: older Asian and Hispanic people don't tend to get stopped becuase they are law-abiding, they prefer public transport, they don't leave the house often. These could all be valid hypotheses to this visualization.

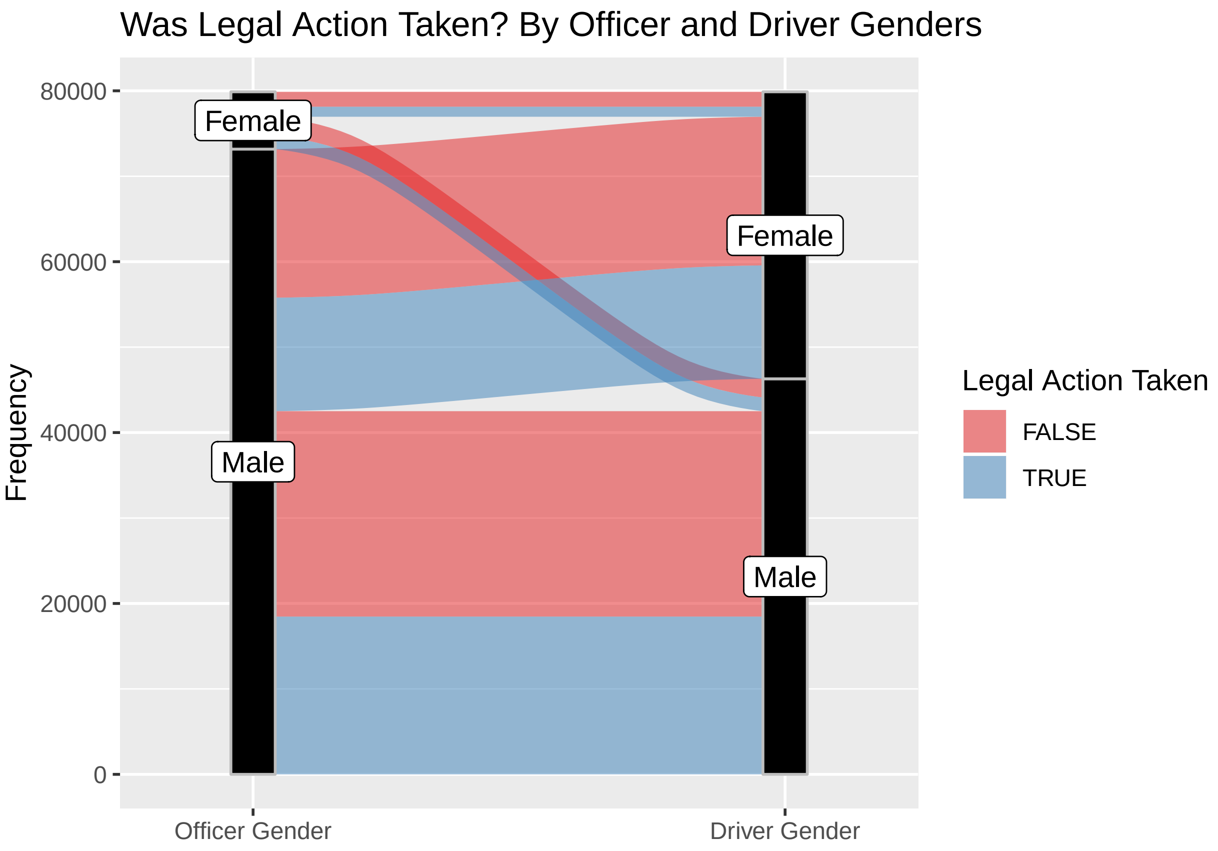
Funnily enough, if you look closely, the older drivers that get arrested tend to be female instead of male.

Exploring Officer Genders and Driver Genders

```
# 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")
```

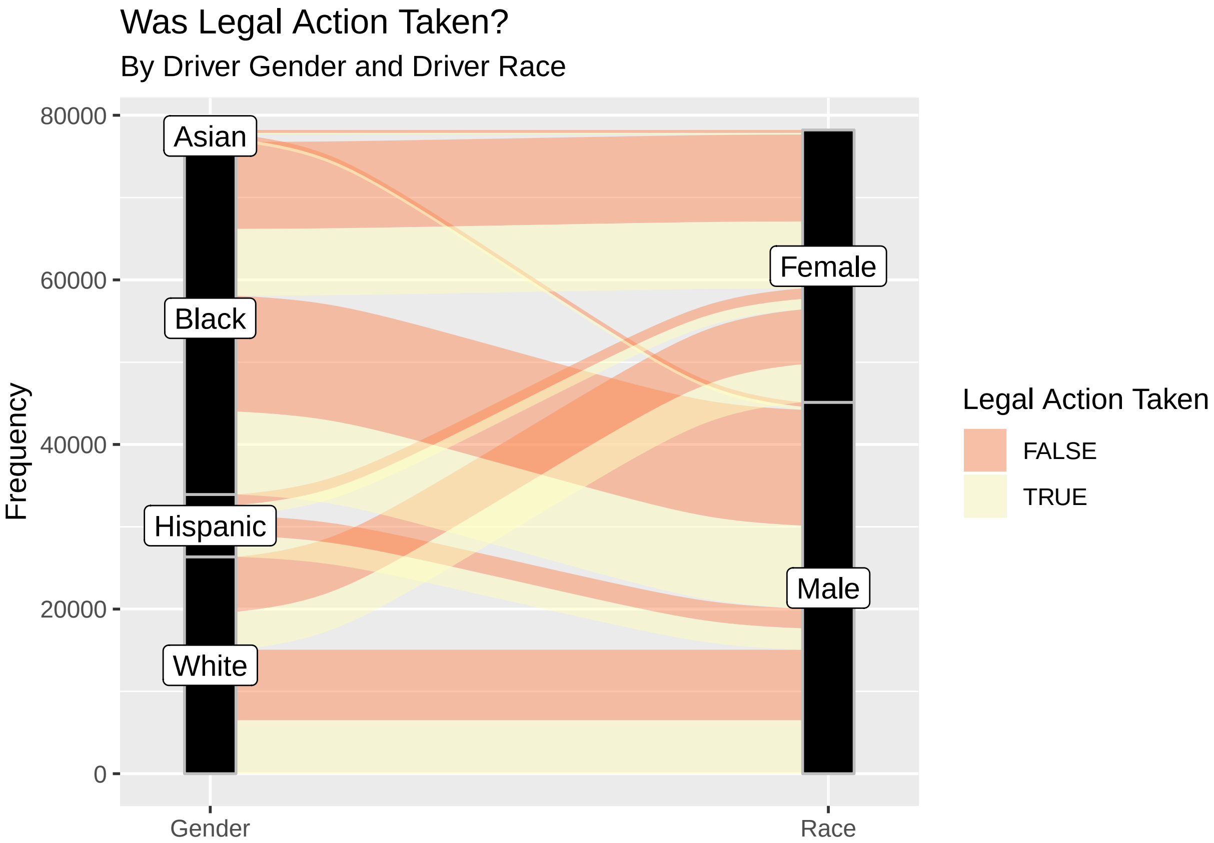


The first observation is that there are an overwhelming amount of male officers. Second, both male and female officers take legal action to about fifty percent of the drivers they stop. Third, male and female officers tend *ever so slightly* to stop male drivers more than female drivers, explaining the overall *slight* majority of female vs. male drivers stopped.

```r
# 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
```

```r
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") +
  labs(title="Was Legal Action Taken?",
       subtitle="By Driver Gender and Driver Race")
```



The most fascinating two points here are the Black and Hispanic outcomes. It appears that black and hispanic people are stopped very frequently (they make up a dispoportionately large pool of total drivers stopped) and yet, the outcome is less legal action.

This might mean that officers are being over-cautious of black and hispanic people, and one they have been stopped and searched, they average a lower rate of legal action taken because they were over-sampled in the first place. This a large statement, and should be backed up with more evidence.