Homework3

Hongdou Li

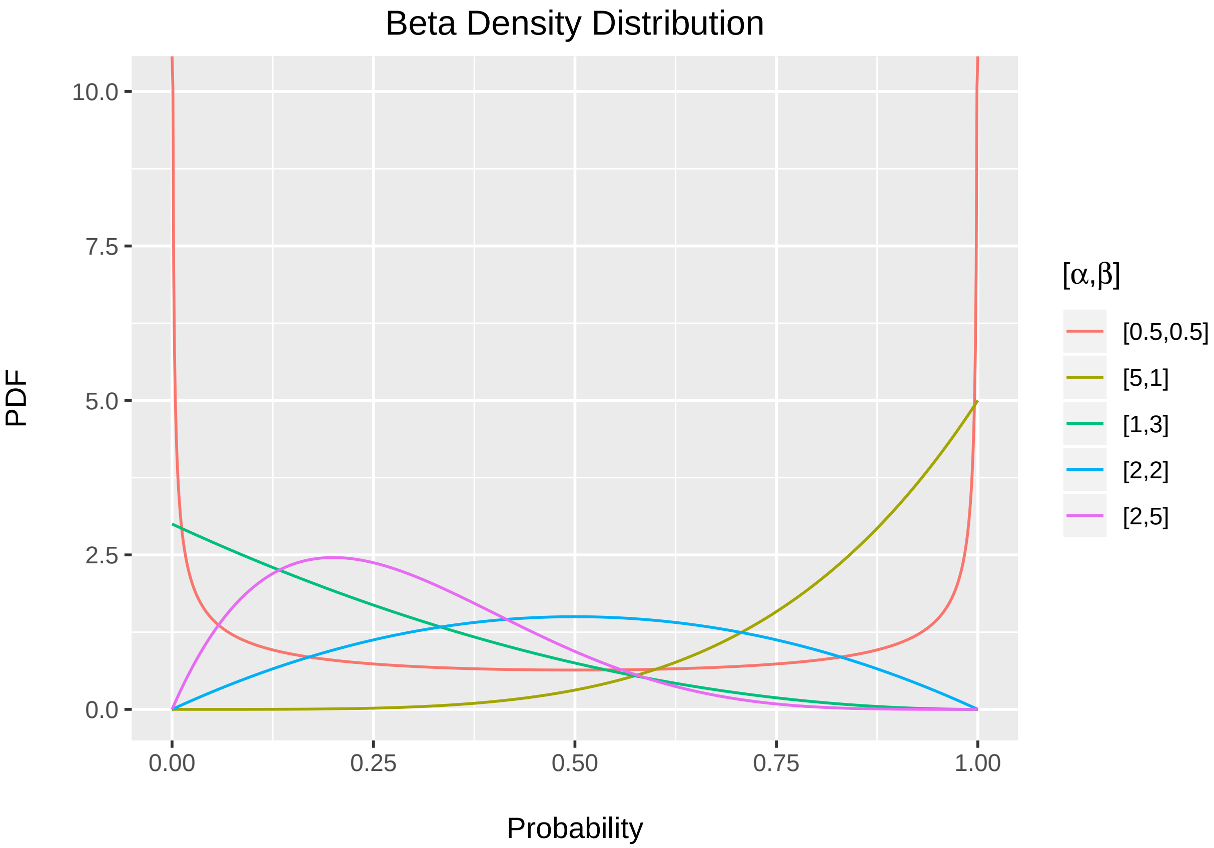
8/6/2018

Question 1

```
library(tidyverse) #load the packages we need
library(gridExtra)
library(grid)
library(magrittr)
library(scales)
library(reshape2)
```

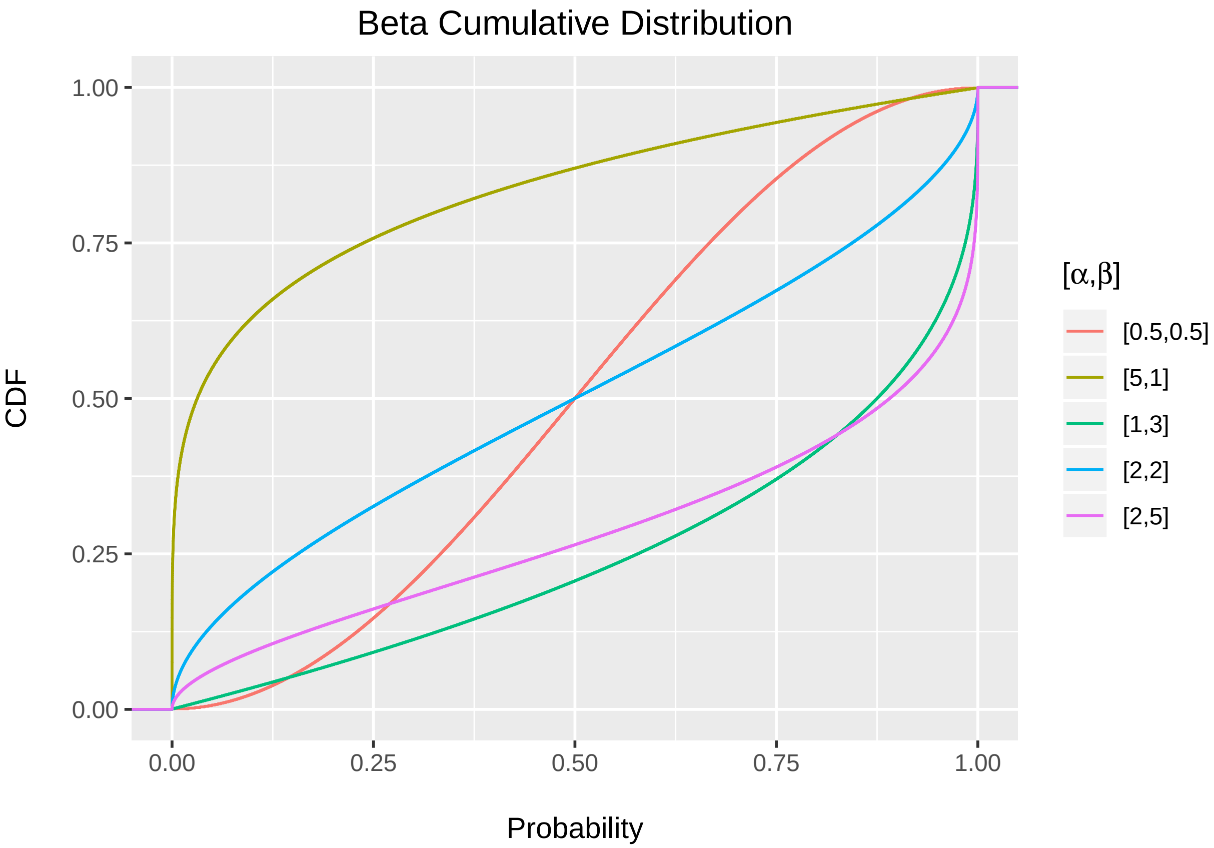
1.1

```
x <- seq(0,1,length=1000)
beta_pdf <- data.frame(cbind(x,dbeta(x,0.5,0.5),
                                   dbeta(x,5,1),
                                   dbeta(x,1,3),
                                   dbeta(x,2,2),
                                   dbeta(x,2,5)))
colnames(beta_pdf) <- c("x","[0.5,0.5]","[5,1]","[1,3]","[2,2]","[2,5]")

beta_pdf <- melt(beta_pdf,x)
beta_pdf %<>% mutate(label = "pdf")
p1 <- beta_pdf %>% ggplot( aes(x,value, color=variable))+
geom_line() +
ggtitle("Beta Density Distribution") +
xlab("\nProbability")+
ylab("PDF\n")+
theme(plot.title = element_text(hjust = 0.5))+
guides(color = guide_legend(title=expression(paste("[",alpha,",",beta,"]"))))
p1
```



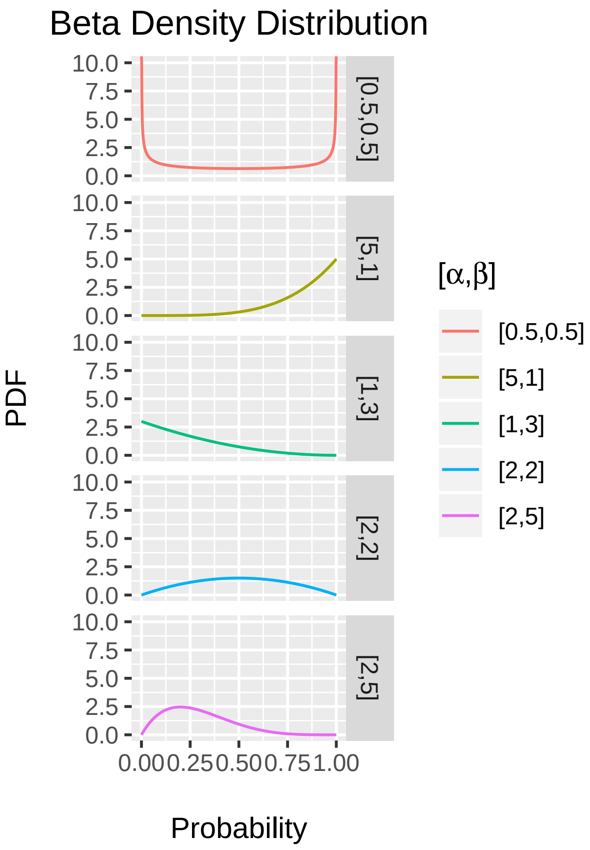
1.2

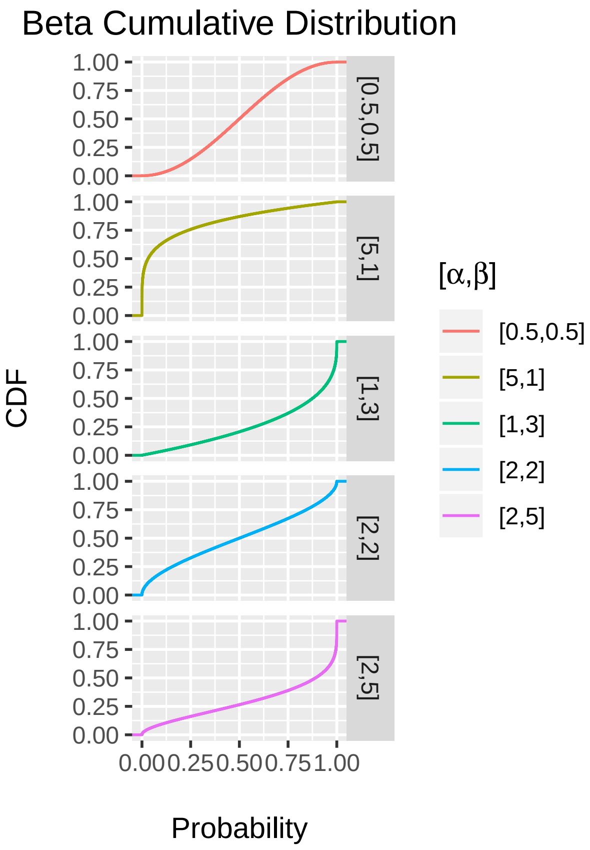
```
x <- seq(0,1,length=1000)
beta_cdf <- data.frame(cbind(x,pbeta(x,0.5,0.5),
                                   pbeta(x,5,1),
                                   pbeta(x,1,3),
                                   pbeta(x,2,2),
                                   pbeta(x,2,5)))
colnames(beta_cdf) <- c("x","[0.5,0.5]","[5,1]","[1,3]","[2,2]","[2,5]")

beta_cdf <- melt(beta_cdf,x)
beta_cdf %<>% mutate(label = "cdf")
p2 <- beta_cdf %>%
  ggplot(aes(value, colour=variable))+
  stat_ecdf(geom="step")+
  ggtitle("Beta Cumulative Distribution") +
xlab("\nProbability")+
ylab("CDF\n")+
theme(plot.title = element_text(hjust = 0.5))+
guides(color = guide_legend(title=expression(paste("[",alpha,",",beta,"]"))))
p2
```



1.3

```
p1 <- p1 +facet_grid(vars(variable))

p2 <- p2 + facet_grid(vars(variable))

grid.arrange(p1,p2,ncol=2)
```


Beta Density Distribution
PDF
10.0
7.5
5.0
2.5
0.0
[0.5,0.5]
[5,1]
[1,3]
[2,2]
[2,5]
0.00
0.25
0.50
0.75
1.00
Probability
[α,β]
[0.5,0.5]
[5,1]
[1,3]
[2,2]
[2,5]


Beta Cumulative Distribution
CDF
1.00
0.75
0.50
0.25
0.00
[0.5,0.5]
[5,1]
[1,3]
[2,2]
[2,5]
0.00
0.25
0.50
0.75
1.00
Probability
[α,β]
[0.5,0.5]
[5,1]
[1,3]
[2,2]
[2,5]

Question 2

```
#read the data
officerTrafficStops <- read.csv("~/classes/msan593/homework/hw3/Officer_Traffic_Stops.csv")
```

First, we take a quick look at the data

```
officerTrafficStops %>%
  glimpse()
```

```
## Observations: 79,884
## Variables: 17
## $ Month_of_Stop            <fct> 2016/01, 2016/01, 2016/01, 2016/01, 2...
## $ Reason_for_Stop          <fct> Speeding                     , Stop Light...
## $ Officer_Race             <fct> White, White, White, Black/African Am...
## $ Officer_Gender           <fct> Male, Male, Male, Male, Male, Male, M...
## $ Officer_Years_of_Service <int> 6, 6, 6, 2, 6, 6, 6, 2, 7, 9, 8, 5, 9...
## $ Driver_Race              <fct> White, Black, Black, Black, White, Wh...
## $ Driver_Ethnicity         <fct> Non-Hispanic, Non-Hispanic, Non-Hispa...
## $ Driver_Gender            <fct> Male, Male, Male, Female, Male, Femal...
## $ Driver_Age               <int> 63, 35, 30, 29, 45, 65, 40, 28, 57, 2...
## $ Was_a_Search_Conducted   <fct> No, No, No, No, No, No, No, No, No, Y...
## $ Result_of_Stop           <fct> Citation Issued, Verbal Warning, Cita...
## $ CMPD_Division            <fct> Eastway Division, Eastway Division, E...
## $ ObjectID                 <int> 1001, 1002, 1003, 1004, 1005, 1006, 1...
## $ CreationDate             <fct> 2016-12-20T23:49:30.533Z, 2016-12-20T...
## $ Creator                  <fct> charlottedata, charlottedata, charlot...
## $ EditDate                 <fct> 2016-12-20T23:49:30.533Z, 2016-12-20T...
## $ Editor                   <fct> charlottedata, charlottedata, charlot...
```

First, we look at a general distribution of traffic Stops among 12 months.

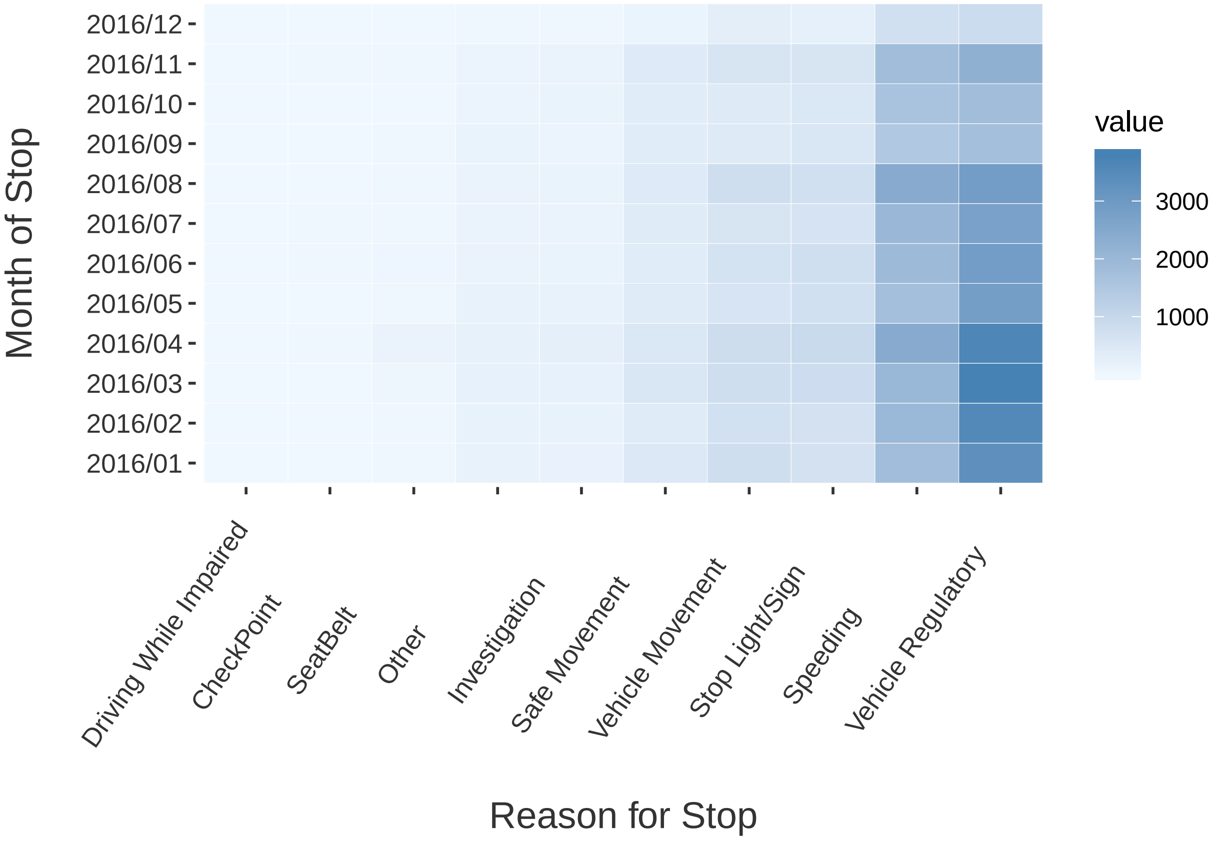
```
officerTrafficStops %>%
  group_by(Month_of_Stop) %>%
    count() %>%
      arrange(desc(n)) %>%
        knitr::kable()
```

Month_of_Stop	n
2016/04	9060
2016/03	8518
2016/02	7789
2016/08	7709
2016/01	7541
2016/06	7053
2016/05	6854
2016/07	6767
2016/11	6070
2016/10	5183
2016/09	4941
2016/12	2399

It is interesting to notice that April has the highest number of traffic stops, which is almost 4 times of the number in December.

Next, I want to know the distribution of those ten reasons among 12 months.Thus, I draw a heatmap showing the frequency of those reasons in each month.

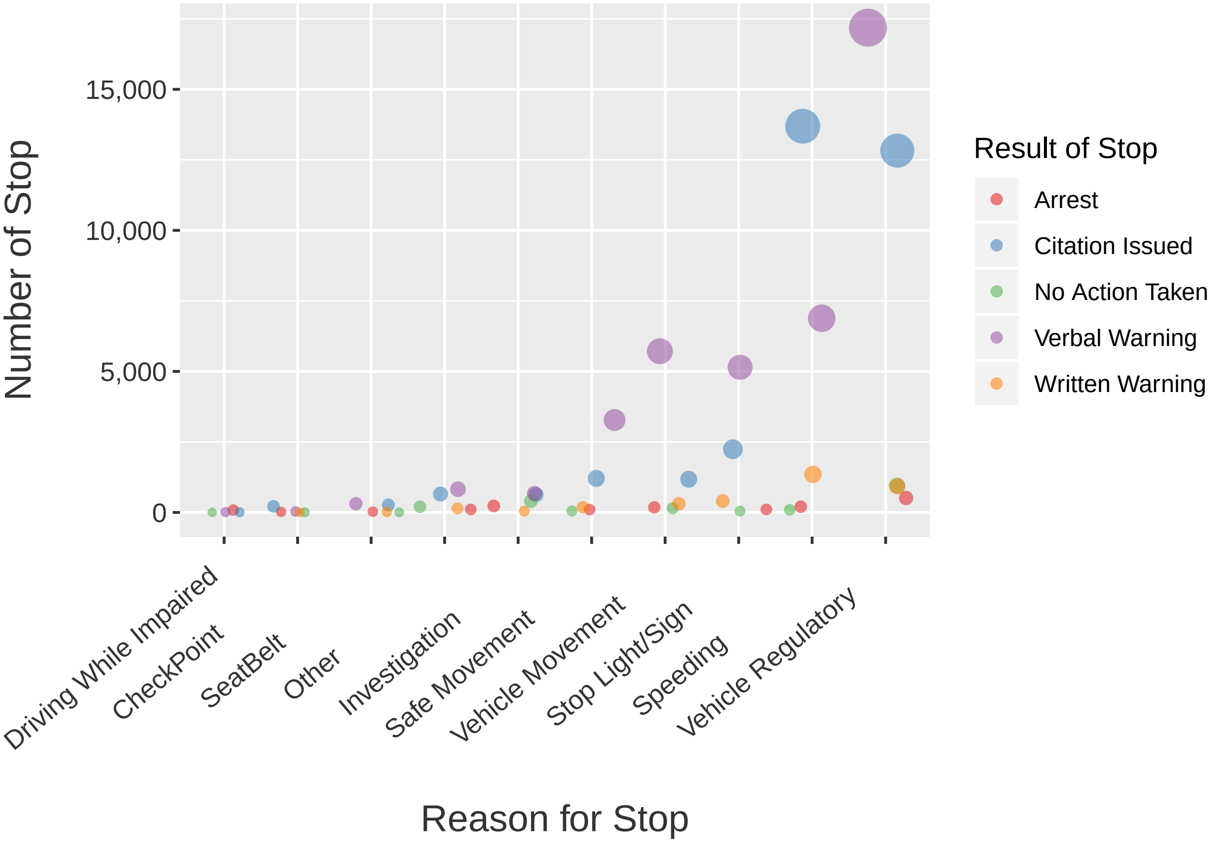
```
#reorder the reason variable based on its frequency
reasonOrder <- officerTrafficStops %>%
                    select(Reason_for_Stop) %>%
                      table()
officerTrafficStops$Reason_for_Stop <- factor(officerTrafficStops$Reason_for_Stop,
                                                     levels = names(reasonOrder[order(reasonOrder,
decreasing = F)]))
#draw the plot
df1 <- officerTrafficStops %>%
  group_by(Month_of_Stop,Reason_for_Stop) %>%
    count()
df.m <- melt(df1)
ggplot(df.m, aes(Reason_for_Stop, Month_of_Stop))+
  geom_tile(aes(fill = value), colour = "white")+
  ylab("Month of Stop\n")+
  xlab("\nReason for Stop")+
  scale_fill_gradient(low="aliceblue", high="steelblue")+
  theme(axis.text.x = element_text(angle=55, hjust=1,colour="grey20",size=10),
        axis.text.y = element_text(colour="grey20",size=10),
            axis.title.x = element_text(colour="grey20",size=14),
            axis.title.y = element_text(colour="grey20",size=14),
        panel.background = element_rect(fill = "white"))
```



It is clear that the number of stops in April is the highest, and there are far less stops happened in December. As for the reasons for stop, Vehicle Regulatory is the most common reason among all 12 months. However, The gap between numbers of stop caused by speeding and vehicle regulatory gets smaller at October, November and December, which lead to my guess that people tend to drive over speed during winter.

Then, we want to check what results will all those ten reasons lead to.

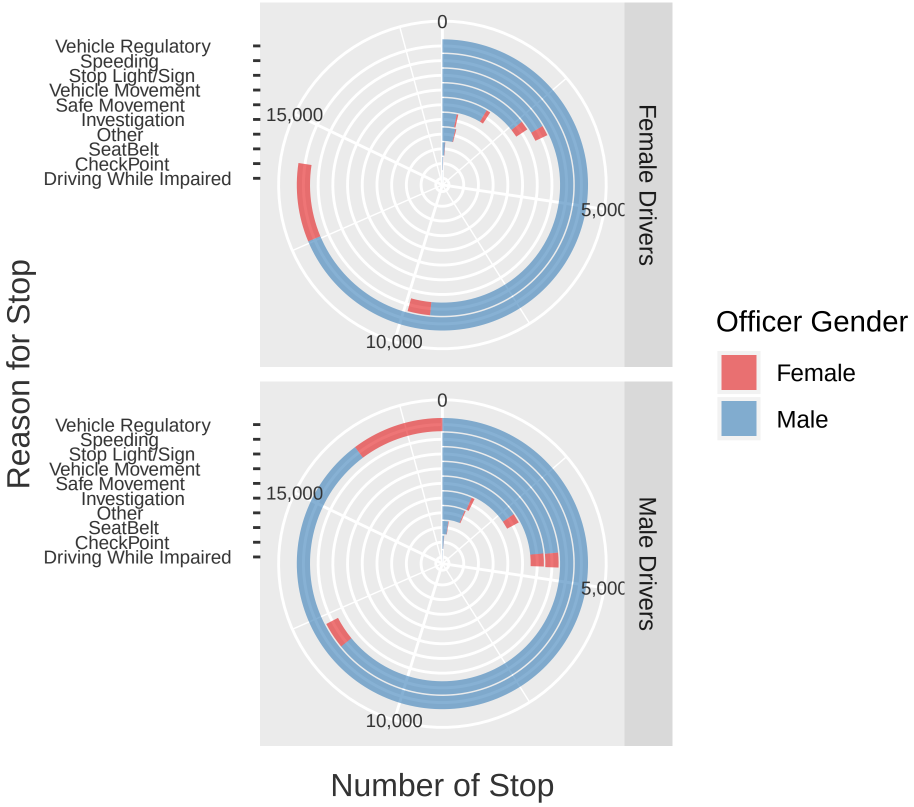
```
officerTrafficStops %>%
  group_by(Reason_for_Stop,Result_of_Stop) %>%
        count() %>%
          ggplot()+
          geom_point(aes(x=Reason_for_Stop, y = n, color = Result_of_Stop, size=n),
                     alpha=0.55,position='jitter')+
          xlab("\nReason for Stop")+
          ylab("Number of Stop\n")+
          theme(axis.text.x = element_text(angle=40, hjust=1,colour="grey20",size=10),
                axis.text.y = element_text(colour="grey20",size=10),
            axis.title.x = element_text(colour="grey20",size=14),
            axis.title.y = element_text(colour="grey20",size=14))+
          guides(colour = guide_legend("Result of Stop"), size=F)+
          scale_y_continuous(labels = comma)+
          scale_color_brewer(palette = "Set1")
```



As the scatter graph shown above, most speeding ended up with a citation Issued while most Vehicle Regulatory Issues may lead to a verbal warning. As for other reasons, Verbal warning is also the most common result. That is to say, although vehicle regulatory is the most common reason that a officer provides when stopping a vehicle, it is highly likely that the driver will only receive a verbal warning. Meanwhile, The probability of receiving a citation while driving over speed is higher than breaking the vehicle regulation. So, the punishment for speeding seems more severe. In addition, breaking vehicle regulation did have a slightly higher possibility in leading to an arrest.

Next, we would like to see whether the reasons of getting stopped are related with gender, both the officers' gender and drivers'.

```
labelNames <- c(`Female`="Female Drivers",
                `Male`="Male Drivers")
officerTrafficStops %>%
  group_by(Reason_for_Stop,Driver_Gender,Officer_Gender) %>%
  count() %>%
  ggplot()+
  geom_bar(aes(x=Reason_for_Stop, y=n,fill=Officer_Gender), stat='identity',
           alpha = 0.6, size = 3)+
  coord_polar("y")+
  facet_grid(rows = vars(Driver_Gender),labeller = as_labeller(labelNames))+
  xlab("\nReason for Stop")+
  ylab("Number of Stop\n")+
      theme(axis.text.x = element_text(colour="grey20",size=7),
            axis.text.y = element_text(colour="grey20",size=7),
            axis.title.x = element_text(colour="grey20",size=12),
            axis.title.y = element_text(colour="grey20",size=12))+
          scale_y_continuous(labels = comma)+
          guides(fill = guide_legend(title="Officer Gender"))+
          scale_fill_brewer(palette = "Set1")
```

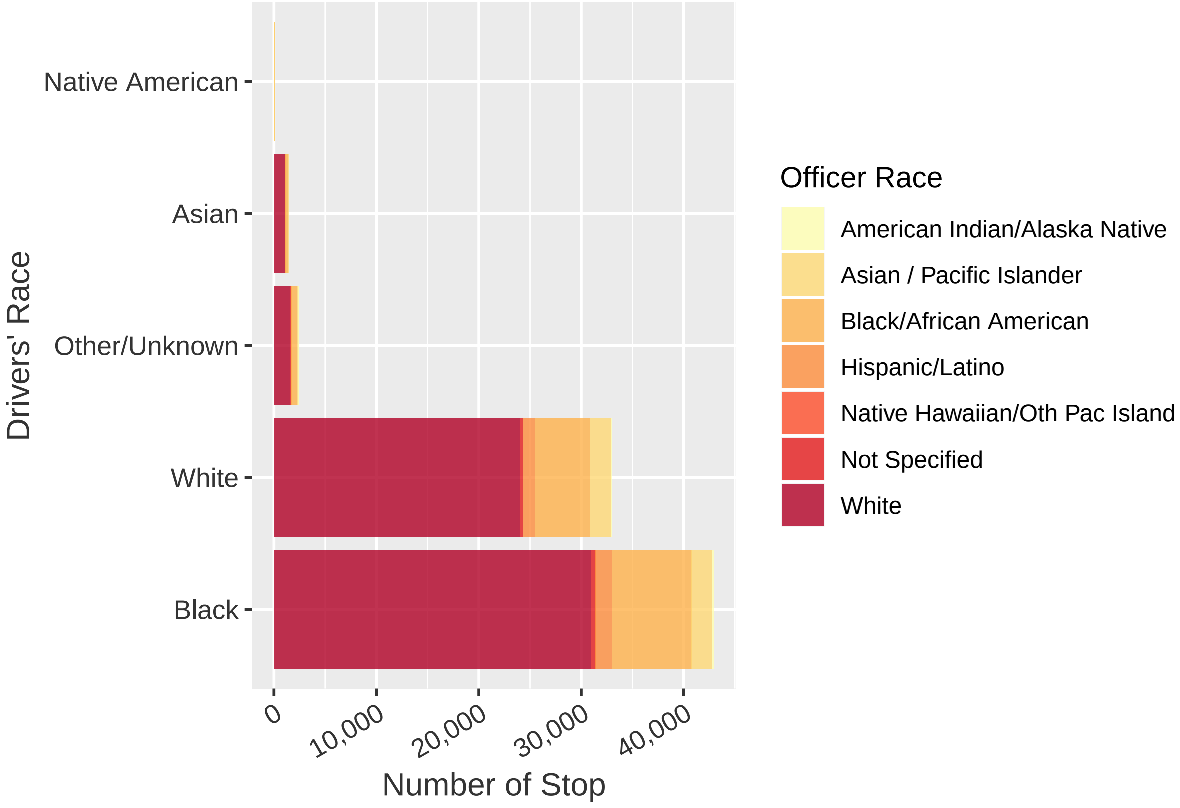


As is shown above, Vehicle Regulatory is the most common reason for stop for both male and female drivers. It is interesting to notice that for female drivers, not following the stop light/sign is more common than vehicle movement. For male drivers, however, the frequency for those two reasons are almost the same.

As for male and female officers, the differences are even bigger. It seems that male officers stopped more cars than female officers. But I assume the difference is due to the fact that there are far more male officers than female ones.

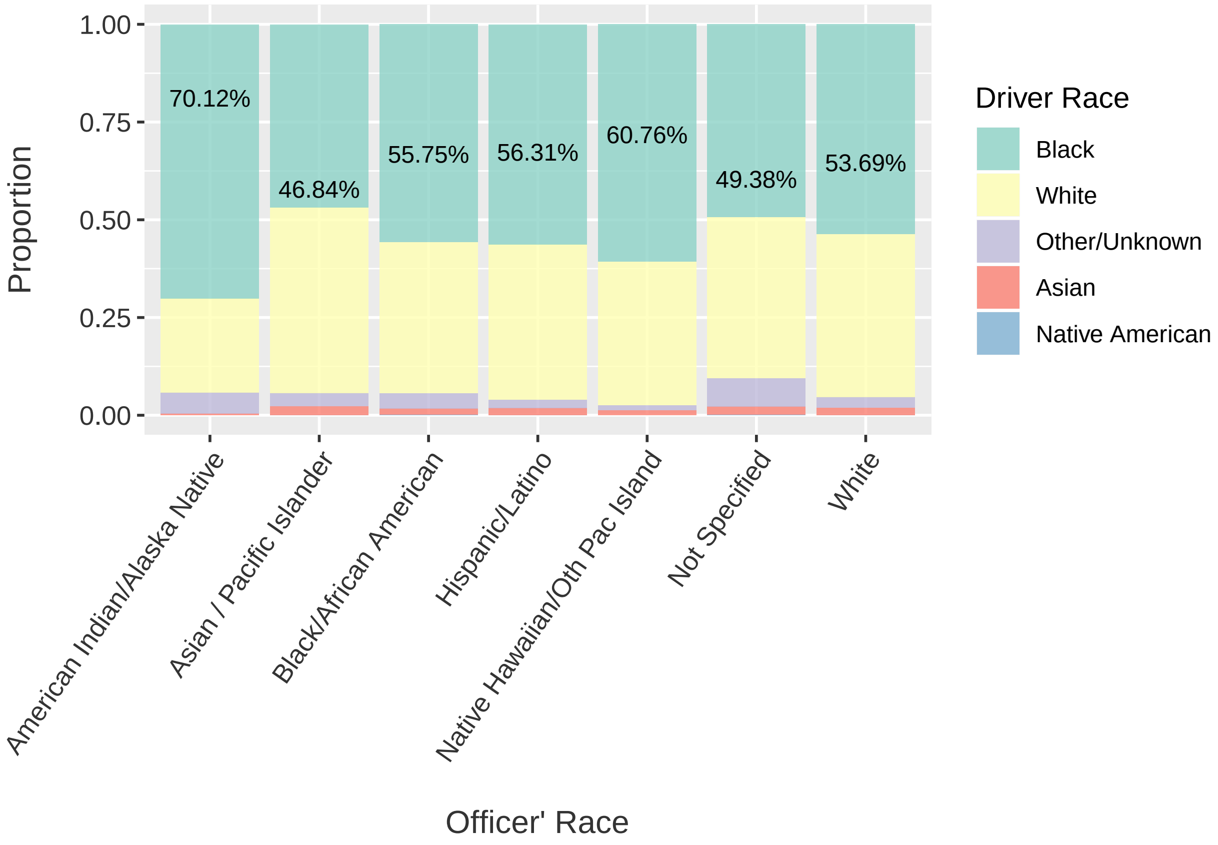
Then, we take a look at the relationship between officer race and driver race.

```
DriverRaceOrder <- officerTrafficStops %>%
                    select(Driver_Race) %>%
                      table()
officerTrafficStops$Driver_Race <- factor(officerTrafficStops$Driver_Race,
      levels = names(DriverRaceOrder[order(DriverRaceOrder,
                                           decreasing = T)]))

officerTrafficStops$Officer_Race[officerTrafficStops$Officer_Race==" "] <- "Not Specified"

officerTrafficStops %>%
  group_by(Officer_Race,Driver_Race) %>%
      ggplot()+
      geom_bar(aes(x=Driver_Race, fill=Officer_Race),position = position_stack(reverse=F),
               alpha=0.8)+
  xlab("\nDrivers' Race")+
  ylab("Number of Stop\n")+
      theme(axis.text.x = element_text(angle=30, hjust=1,colour="grey20",size=10),
            axis.text.y = element_text(colour="grey20",size=10),
            axis.title.x = element_text(colour="grey20",size=12),
            axis.title.y = element_text(colour="grey20",size=12))+
              scale_y_continuous(labels = comma)+
              guides(fill = guide_legend(title="Officer Race"))+
              coord_flip()+
      scale_fill_brewer(palette = "YlOrRd")
```



From the graph, we know that White Officer stoped the most cars. I believe that is because there are more white officers.Besides, among those cars, the amount of black and white drivers are much higher than drivers from other race. The gap is extremly high and we are not able to take a clear view for the behaviour of officers from minorities. Thus we may want to look at the proportion for car stopping frequency over different drivers' races for each race of officers. Particularly, we marked the proportion number on the bar chart for black drivers.

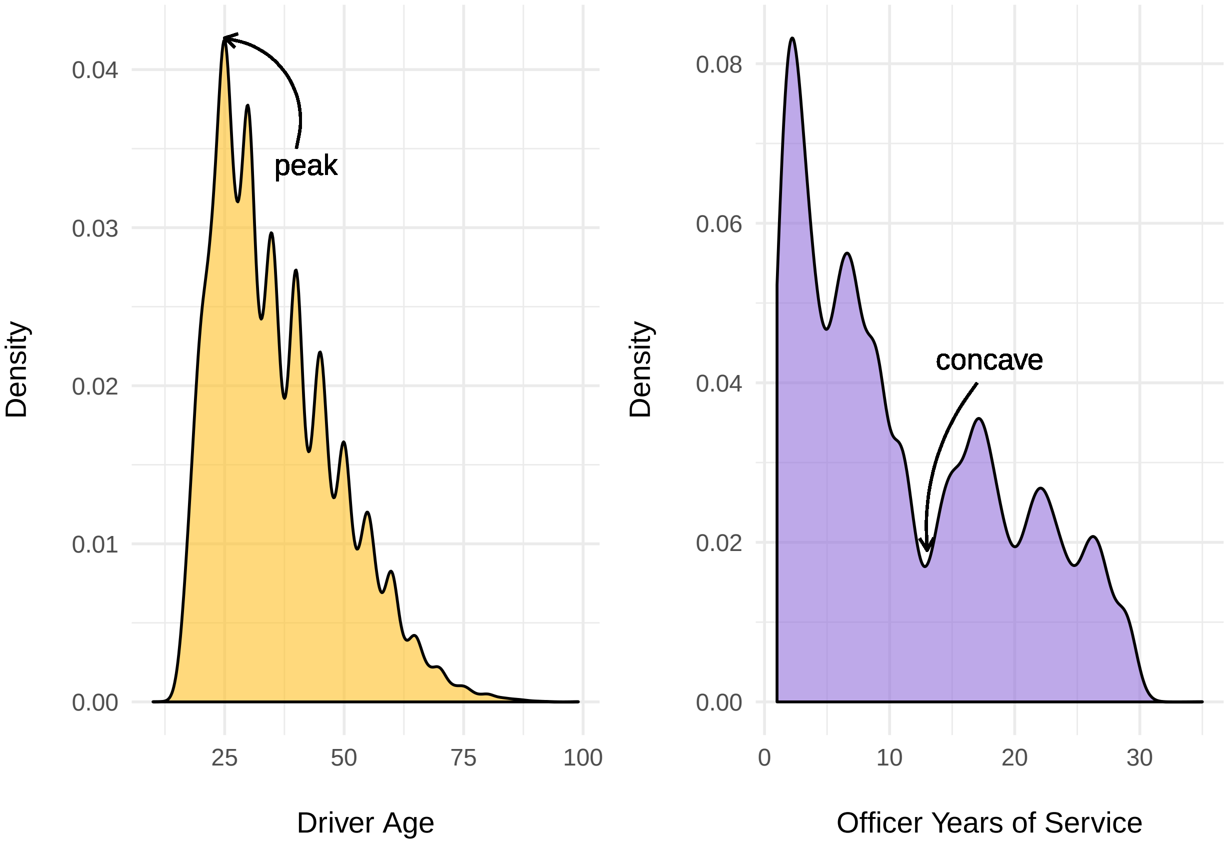
```r
officerTrafficStops %<>%
  group_by(Officer_Race) %>%
    mutate(total = n())

a <- officerTrafficStops %>%
  select(c("Driver_Race","Officer_Race","total")) %>%
  filter(Driver_Race=="Black") %>%
    group_by(Officer_Race, Driver_Race) %>%
      mutate(pct = n()/total) %>%
      group_by(Officer_Race, Driver_Race)
a <- unique(a)


officerTrafficStops %>%
  group_by(Officer_Race,Driver_Race) %>%
      ggplot()+
      geom_bar(aes(x=Officer_Race, fill=Driver_Race),position="fill",alpha=0.8)+
      geom_text(data = a, aes(x=Officer_Race,
                                y = pct,label=paste(round(100*pct,2), "%", sep="")),
                vjust=-2,check_overlap = T, size=3)+
      xlab("\nOfficer' Race")+
  ylab("Proportion\n")+
      theme(axis.text.x = element_text(angle=55, hjust=1,colour="grey20",size=10),
            axis.text.y = element_text(colour="grey20",size=10),
            axis.title.x = element_text(colour="grey20",size=12),
            axis.title.y = element_text(colour="grey20",size=12))+
      guides(fill = guide_legend(title="Driver Race"))+
      scale_fill_brewer(palette = "Set3")
```



As we can see, almost for every race of officers, nearly half or more of the drivers they stopped are black. This proportion is kind of abnormal. There might be a discrimination upon drivers when stopping their cars.

Next, we take a look at the relationship between drivers' age and the density of being stopped, as well as the relationship between officers' service length and their frequency of stopping a car.

```
p1 <- officerTrafficStops %>%
  ggplot()+
    geom_density(aes(Driver_Age),fill="goldenrod1", alpha=0.6)+
    geom_curve(aes(x=40, y=0.035, xend=25, yend=0.042),
               arrow = arrow(length = unit(0.03,"npc")))+
    geom_text(aes(x=42, y=0.034, label="peak"))+
    xlab("\nDriver Age")+
  ylab("Density\n")+
    theme_minimal()

p2 <- officerTrafficStops %>%
  ggplot()+
    geom_density(aes(Officer_Years_of_Service),fill="mediumpurple", alpha=0.6)+
 geom_curve(aes(x=17, y=0.04, xend=13, yend=0.019), curvature = 0.2,
            arrow = arrow(length = unit(0.03,"npc")))+
  geom_text(aes(x=18, y=0.043, label="concave"))+
  xlab("\nOfficer Years of Service")+
  ylab("Density\n")+
    theme_minimal()

grid.arrange(p1, p2, ncol=2)
```

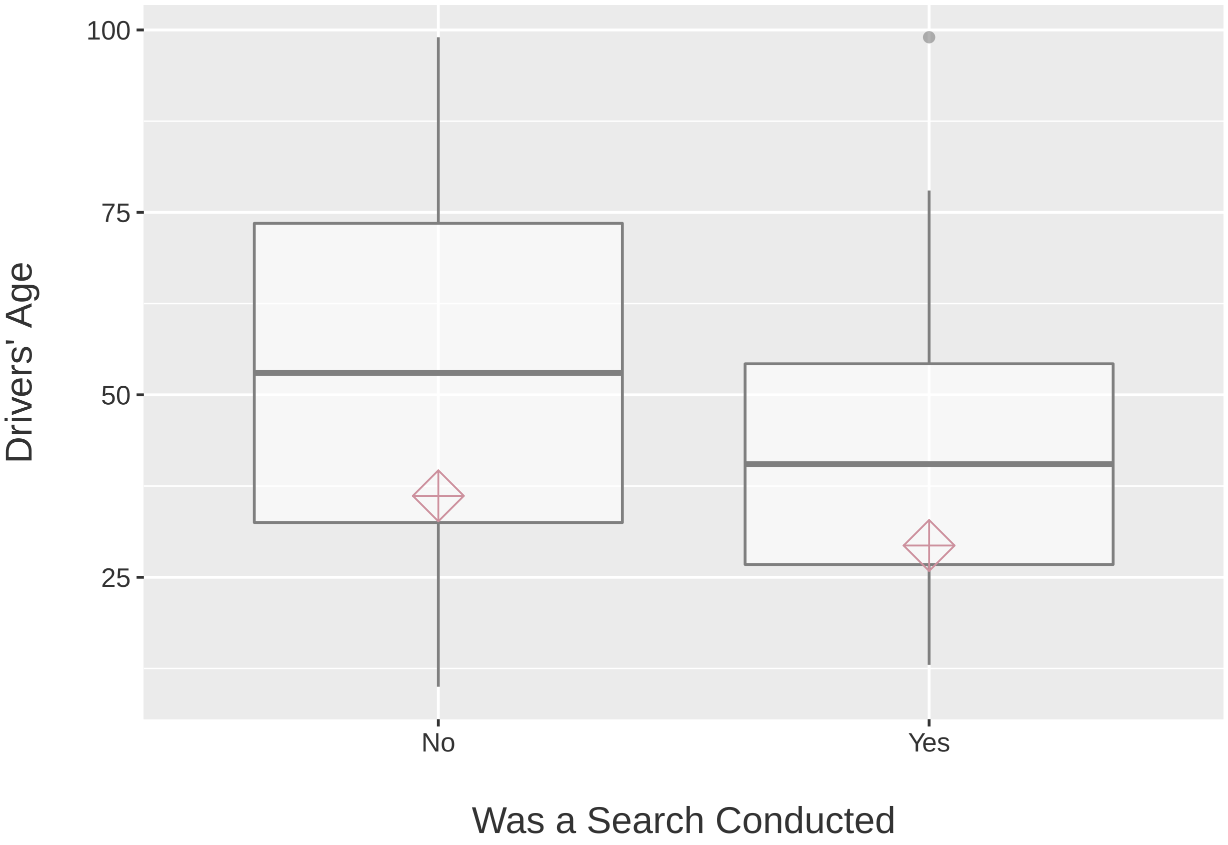


Notice that the distribution reaches its peak at 25 years old, after wards, it fluctuates downwards.

As for Officer years of service, It is interesting to find that the distribution peaks at year 2, it seems that a fresh Officer is more ambitious and tend to stop more cars. the number decreases rapidly and reached bottom for officers with 13 years experience. Thus, when their lengths of service get longer, officers tend to be nicer.

Next, I want to find out whether the drivers' age is associated with car search, the boxplot below shows the quartiles of drivers' age. Meanwhile, I also draw a point indicating the mean age for car search and no car search groups.

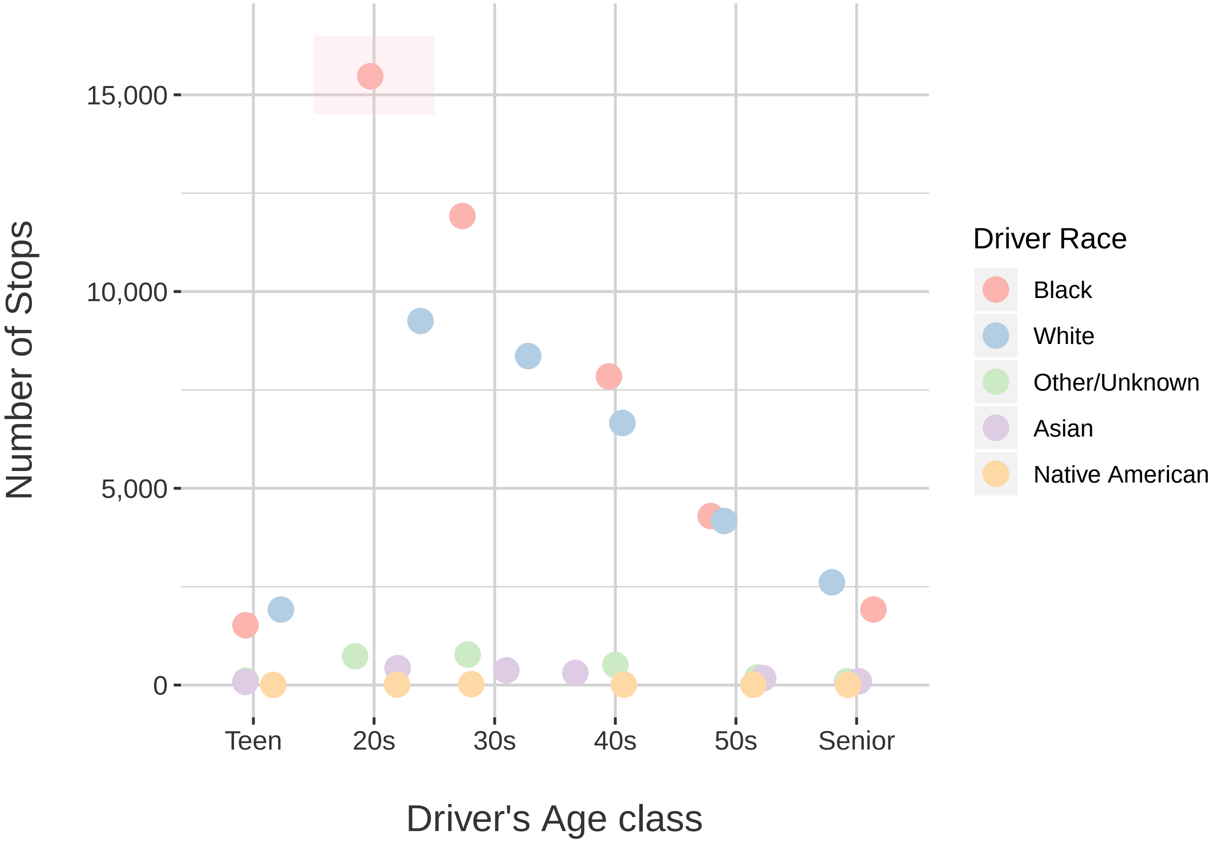
```
meanSearchAge <- officerTrafficStops %>%
                    group_by(Was_a_Search_Conducted) %>%
                      summarize(mean = mean(Driver_Age))

officerTrafficStops %>%
  group_by(Was_a_Search_Conducted,Driver_Age) %>%
    count() %>%
      ggplot()+
      geom_boxplot(aes(x = Was_a_Search_Conducted, y=Driver_Age), alpha = 0.6,
                   colour = "grey50")+
      geom_point(data = meanSearchAge, aes(x = Was_a_Search_Conducted, y = mean),
                 size = 6, shape=9, colour="pink3")+
  xlab("\nWas a Search Conducted")+
        ylab("Drivers' Age\n")+
        theme(axis.text.x = element_text(colour="grey20",size=10),
              axis.text.y = element_text(colour="grey20",size=10),
              axis.title.x = element_text(colour="grey20",size=14),
              axis.title.y = element_text(colour="grey20",size=14))
```



It is interesting to find that although the differences between mean age of drivers who got their car searched after being stopped and who didn't is not very obvious. The third quartile of age for those who got searched is much lower than those who did not. we can conclude that when people getting oldder(especially above 52 years old), they are less likely to got their car searched.

To have a closer examine about the relationship of drivers' age and their probability of being stopped, I divided the drivers into several groups based on their age. Then, I draw a scatter graph showing the number of stop related to drivers' age group and drivers' race.

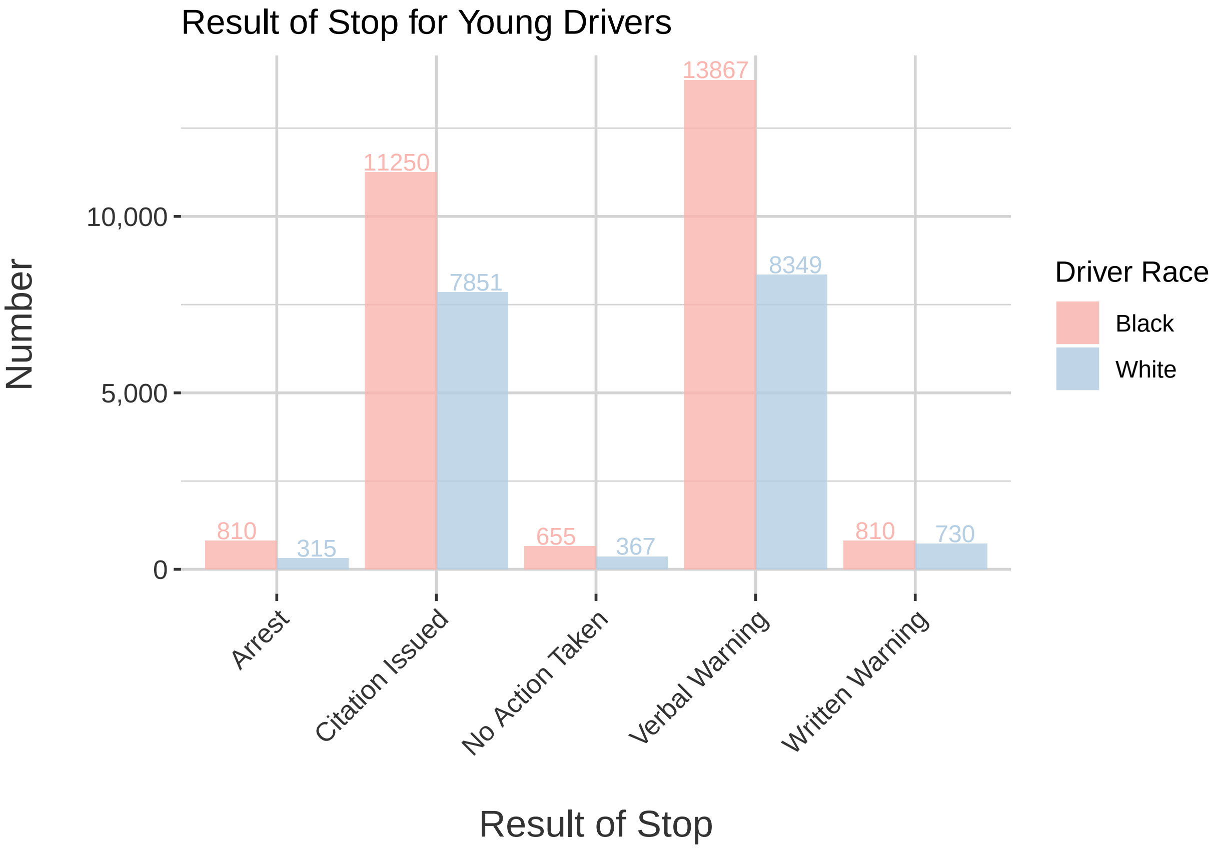
```
officerTrafficStops %<>% mutate(driverAgeClass= cut(Driver_Age,
                                                    breaks = c(0,19,29,39,49,59,Inf),
                 labels = c("Teen","20s","30s","40s","50s","Senior"),
                 levels = c("Teen","20s","30s","40s","50s","Senior"),
                 ordered = T))

officerTrafficStops %>%
  group_by(driverAgeClass,Driver_Race) %>%
    count() %>%
      ggplot()+
        geom_point(aes(x=driverAgeClass, y=n, colour= Driver_Race),
                   size = 4,position='jitter')+
        annotate("rect", xmin=1.5, xmax=2.5, ymin=14500, ymax=16500,
                 alpha=.2, fill= 'pink')+
        xlab("\nDriver's Age class")+
        ylab("Number of Stops\n")+
         theme(axis.text.x = element_text(colour="grey20",size=10),
              axis.text.y = element_text(colour="grey20",size=10),
              axis.title.x = element_text(colour="grey20",size=14),
              axis.title.y = element_text(colour="grey20",size=14),
              panel.background = element_rect(fill = "white"),
              panel.grid.major = element_line(size = 0.5,
                            linetype = 'solid',colour = "lightgrey"),
              panel.grid.minor = element_line(size = 0.25,
                            linetype = 'solid',colour = "lightgrey"))+
          scale_y_continuous(labels = comma)+
          guides(color = guide_legend(title="Driver Race"))+
          scale_color_brewer(palette = "Pastel1")
```



It is noticeable that Black drivers at their tewnties got stopped much more frequently compared with others. Also, the gaps between black and white drivers being stopped get smaller and smaller as drivers' age increases. For seniors, it seems more white senior drivers got stopped than black senior drivers.

Since we know that young black drivers are most likely to get stopped. Also, young drivers tend to receive a car search more often as well. I am interested in what are the most common results of stop for black drivers aged around 20~40 years old and whether there are any differences between black young drivers and white young drivers.

```
officerTrafficStops %>%
  filter(Driver_Race == "Black" | Driver_Race == "White") %>%
    filter(driverAgeClass == "20s" | driverAgeClass == "30s") %>%
    group_by(Result_of_Stop,Driver_Race) %>%
      count() %>%
        ggplot()+
        geom_bar(aes(x=Result_of_Stop, y = n, fill=Driver_Race),position="dodge",
                 stat='identity',alpha=0.8)+
        geom_text(aes(x=Result_of_Stop, y = n,label=n,color=Driver_Race),
                  position=position_dodge(width=1),vjust=-.1,size=3)+
        ggtitle("Result of Stop for Young Drivers")+
        xlab("\nResult of Stop")+
        ylab("Number\n")+
        theme(axis.text.x = element_text(angle=45, hjust=1,colour="grey20",size=10),
              axis.text.y = element_text(colour="grey20",size=10),
              axis.title.x = element_text(colour="grey20",size=14),
              axis.title.y = element_text(colour="grey20",size=14),
              panel.background = element_rect(fill = "white"),
              panel.grid.major = element_line(size = 0.5, linetype = 'solid',
                                colour = "lightgrey"),
  panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                colour = "lightgrey"))+
  scale_y_continuous(labels = comma)+
  guides(fill = guide_legend(title="Driver Race"), color =F)+
      scale_fill_brewer(palette = "Pastel1")+
      scale_color_brewer(palette = "Pastel1")
```



As we can see,for drivers aged 20~40 years old, black drivers received more verbal warning than citation issued when stopped by officers. The gap relatively small for young white drivers who got stopped. Young white drivers seem to receive citation and verbal warning at similar possibility.