Assignment 5: Data Visualization with ggplot2

Louis Dion

9/30/2019

#Using c2015 dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
c2015<-read_excel('C:/Users/student/Documents/MATH421/data/c2015.xlsx')
```

1. Clean the data for easy graphing

##Remove observations that are unknown,fix age and TRAV_SP,filter for drivers

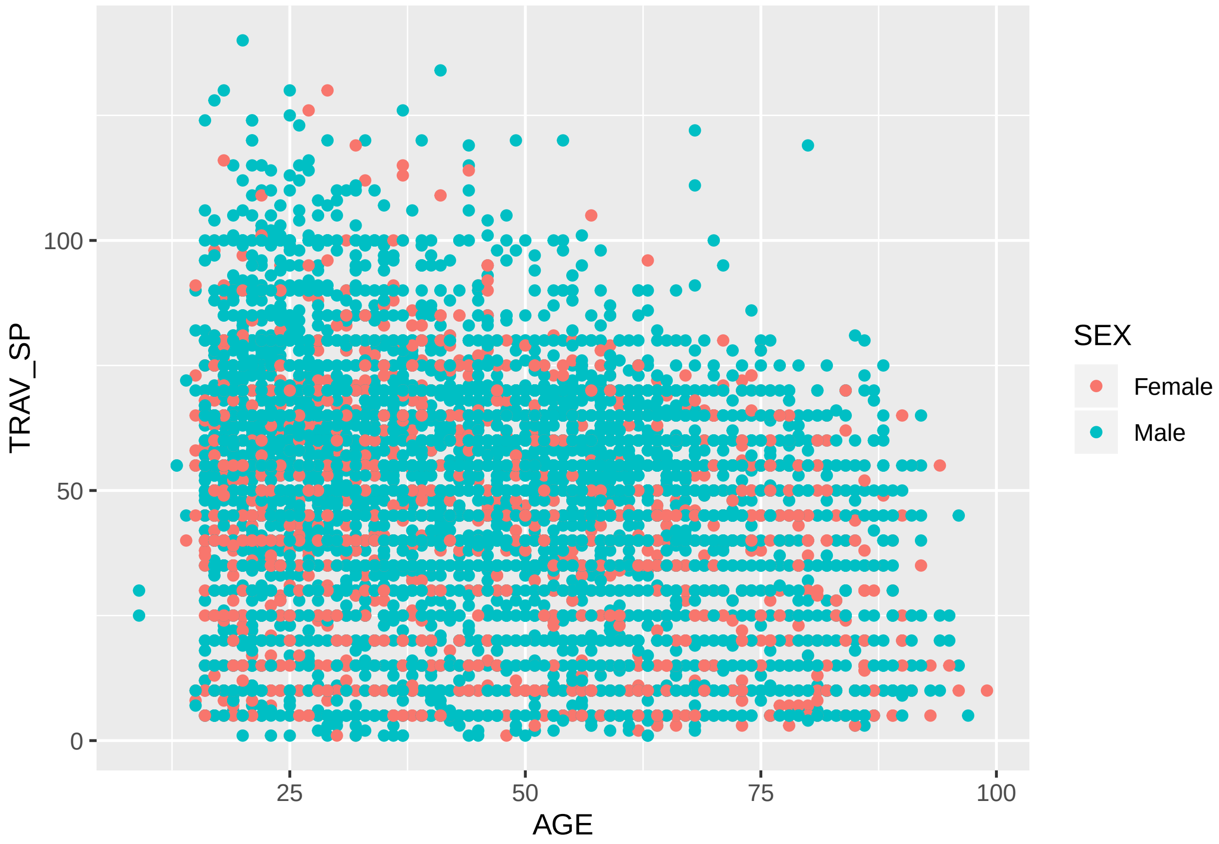
```
c2015<-c2015%>%mutate(TRAV_SP=sapply(strsplit(TRAV_SP,split=" ",fixed = TRUE),function(x) (x[1])), TRAV_
```

```
## Warning: NAs introduced by coercion
```

```
c2015<-c2015%>%filter_all(~!is.na(.))
c2015<-c2015%>%filter_all(~!(.=='Unknown'))
c2015<-c2015%>%filter_all(~!(.=='Other'))
c2015<-c2015%>%filter_all(~!(.=='Unknown (Police Reported)'))
c2015<-c2015%>%filter_all(~!(.=='Injured, Severity Unknown'))
c2015<-c2015%>%filter_all(~!(.=='Not Rep'))
c2015<-c2015%>%filter_all(~!(.=='Not Reported'))
c2015<-c2015%>%filter_all(~!(.==str_detect(.,'Not Rep')))
c2015<-c2015%>%filter_all(~!(.==str_detect(.,'Unknown')))
c2015<-c2015%>%mutate(AGE=replace(AGE,AGE=='Less than 1','0'),AGE=as.numeric(AGE))
c2015<-c2015%>%filter(SEAT_POS=='Front Seat, Left Side')
```

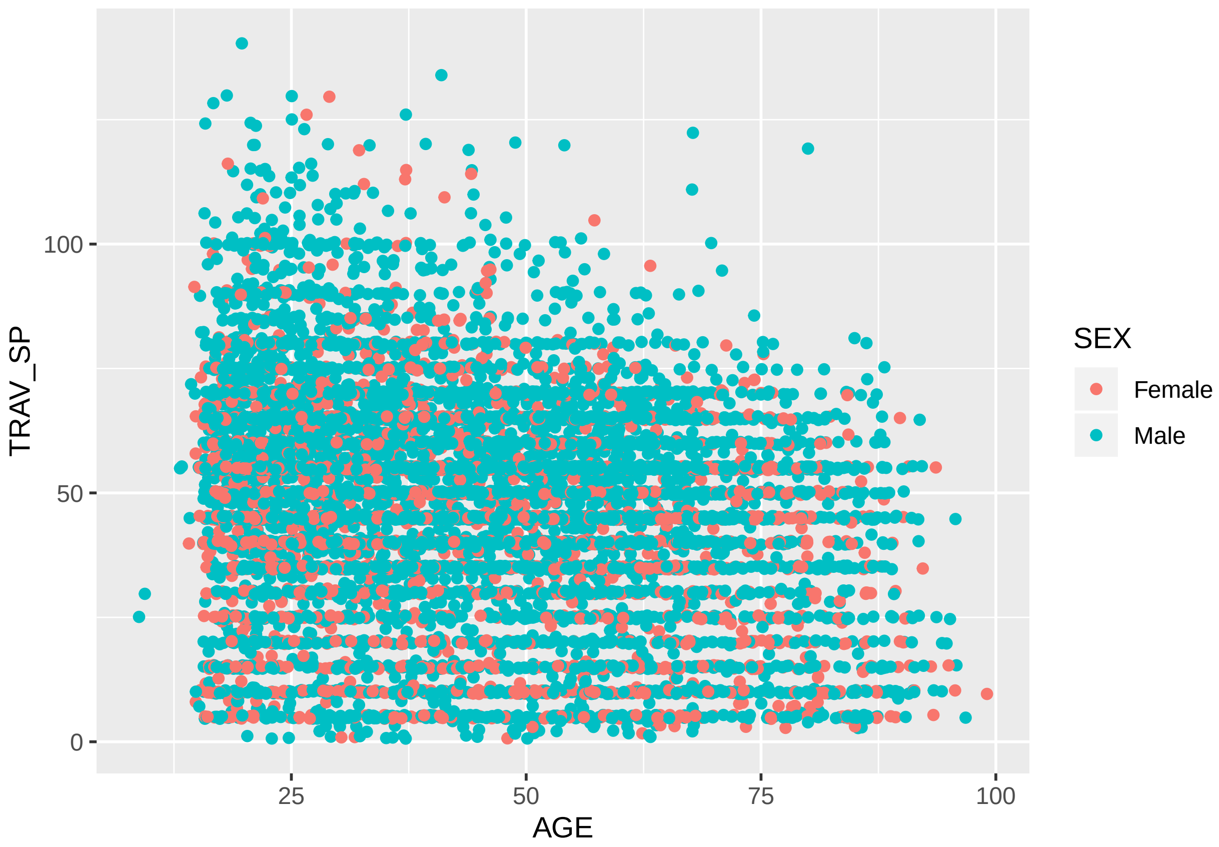
#2.Use geom_point to plot AGE and TRAV_SP coloring by SEX.

```
ggplot(c2015,aes(AGE,TRAV_SP,col=SEX))+
  geom_point()
```



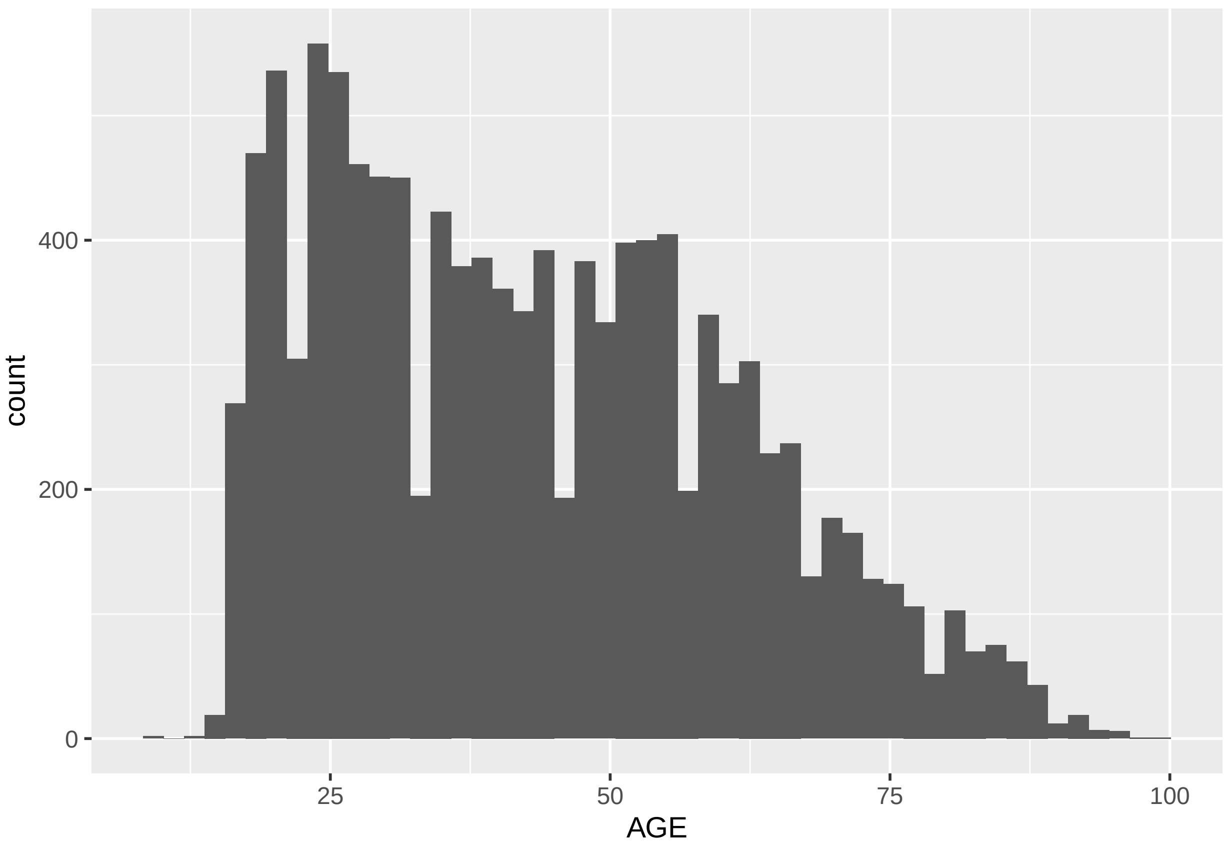
#3.There is overplotting in 2. Overplotting is when many points are duplicated on the graph. Use geom_jitter instead of geom_point for 2. to avoid overplotting.

```
ggplot(c2015,aes(AGE,TRAV_SP,col=SEX))+
  geom_jitter()
```

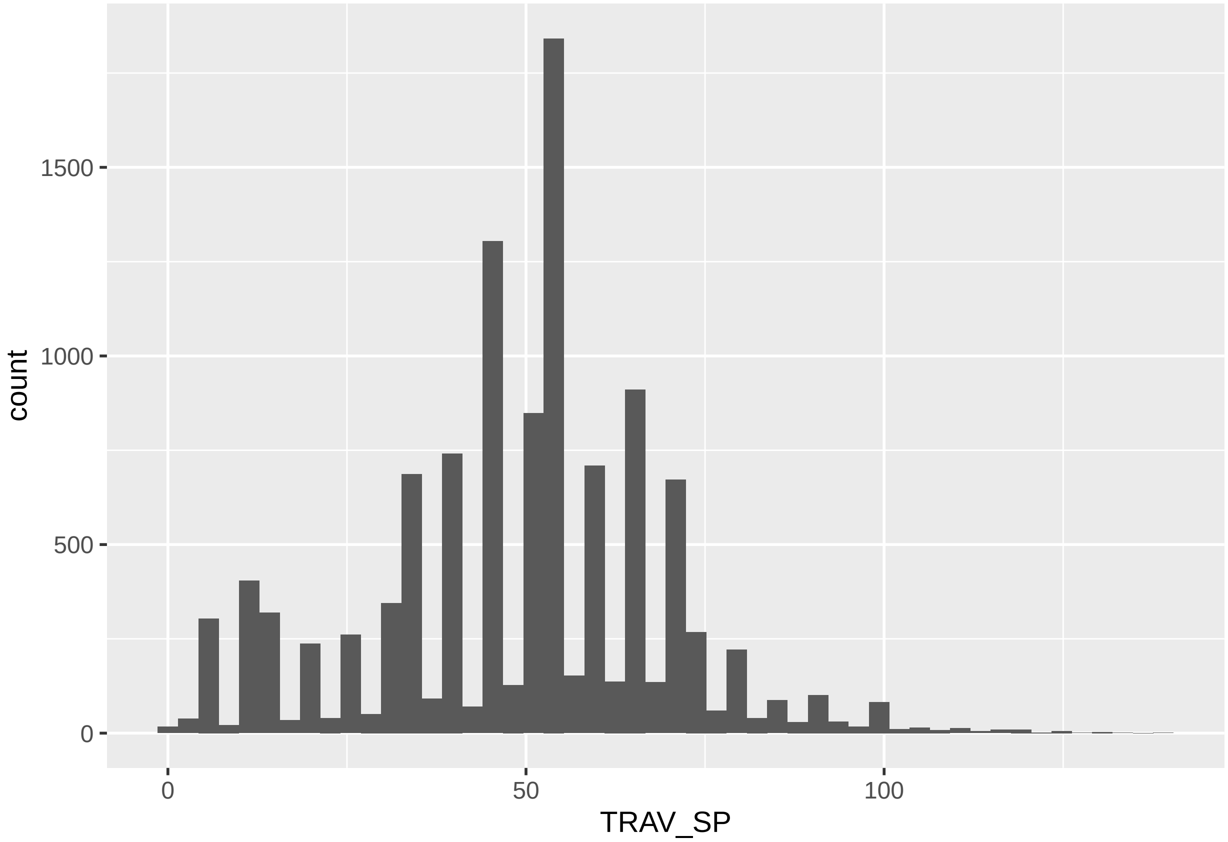


#4.Plot histograms of AGE, TRAV_SP with bins = 50.

```
ggplot(c2015,aes(x=AGE))+
  geom_histogram(bins=50)
```

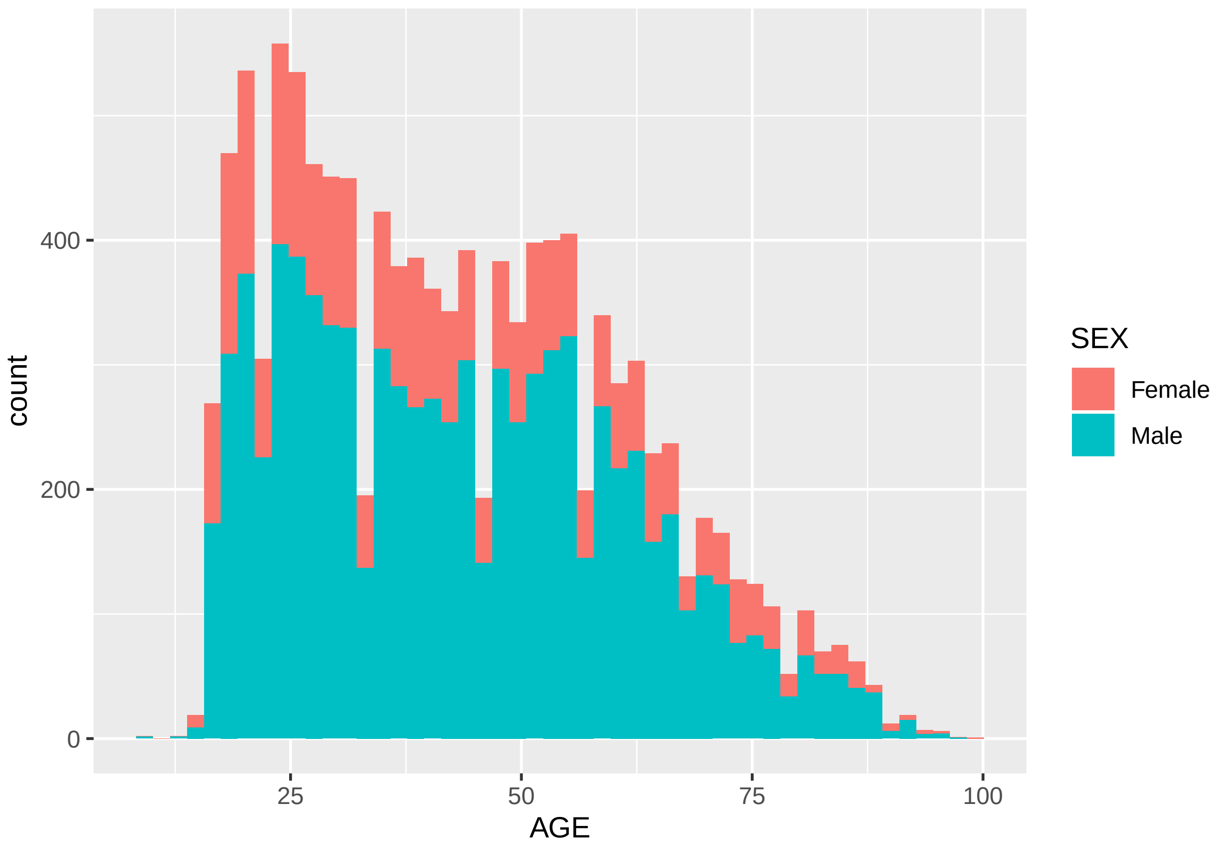


```
ggplot(c2015,aes(x=TRAV_SP))+
  geom_histogram(bins=50)
```



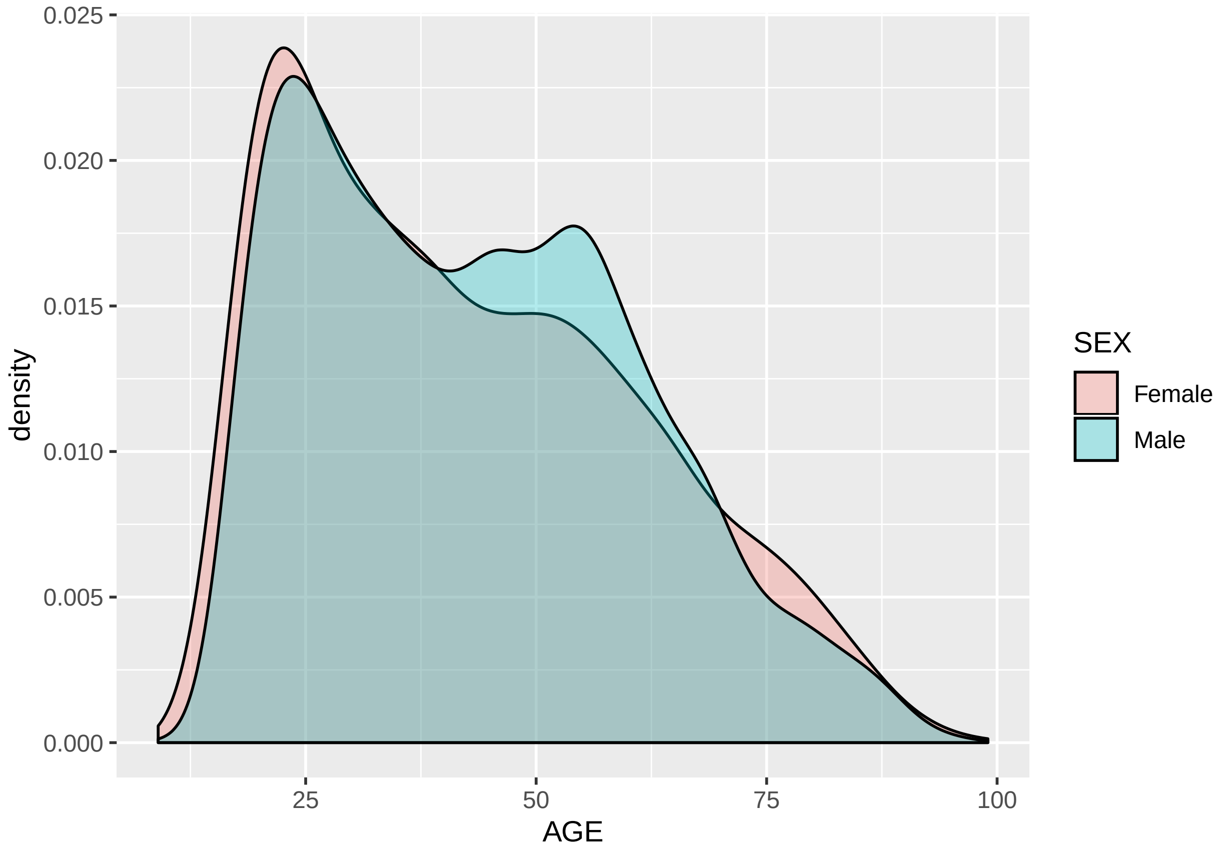
#5.Plot a histogram of AGE coloring (fill) by SEX.

```
ggplot(c2015,aes(x=AGE,fill=SEX))+
  geom_histogram(bins=50)
```



#6.Using geom_density to plot estimated densities of AGE colored by SEX.

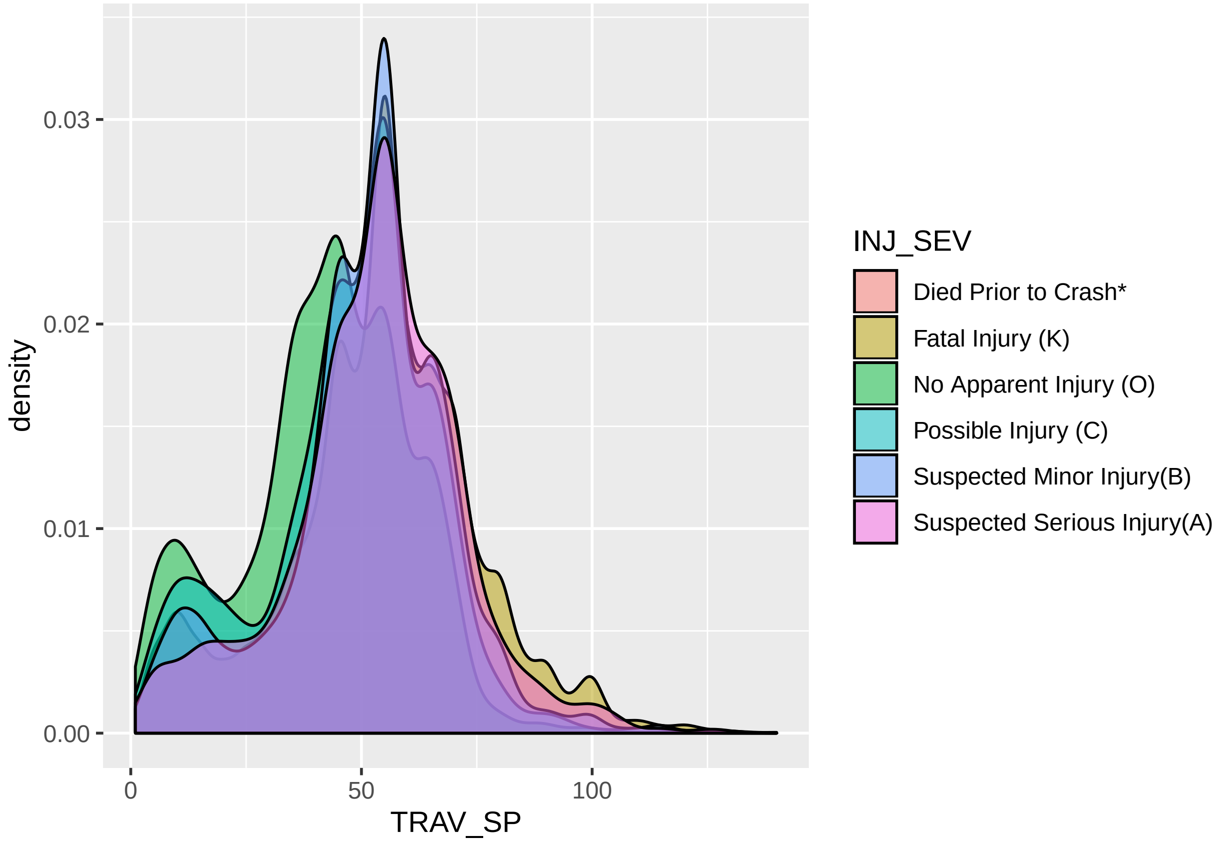
```
ggplot(c2015,aes(x=AGE,fill=SEX))+
  geom_density(alpha=0.3)
```



#7.Plot estimated densities of TRAV_SP colored by INJ_SEV.

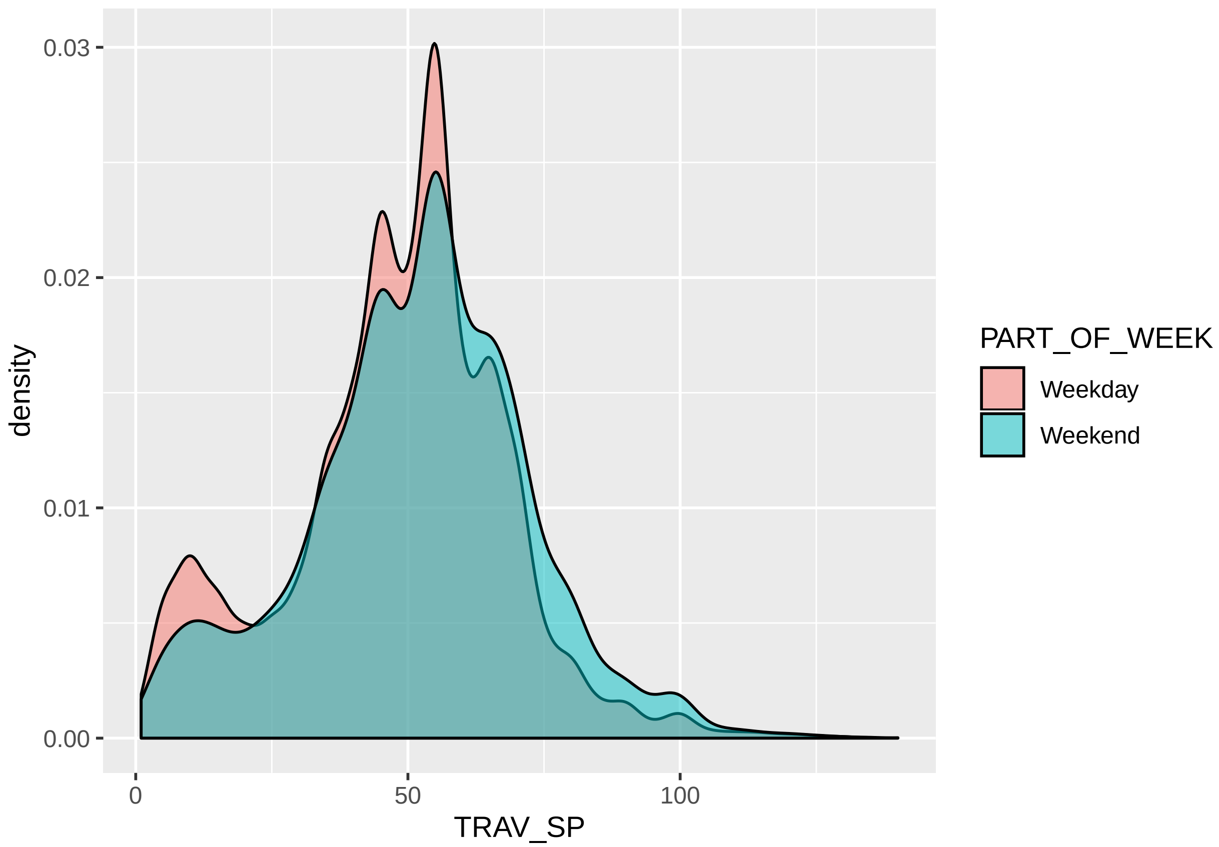
```
ggplot(c2015,aes(x=TRAV_SP,fill=INJ_SEV))+
  geom_density(alpha=0.5)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```



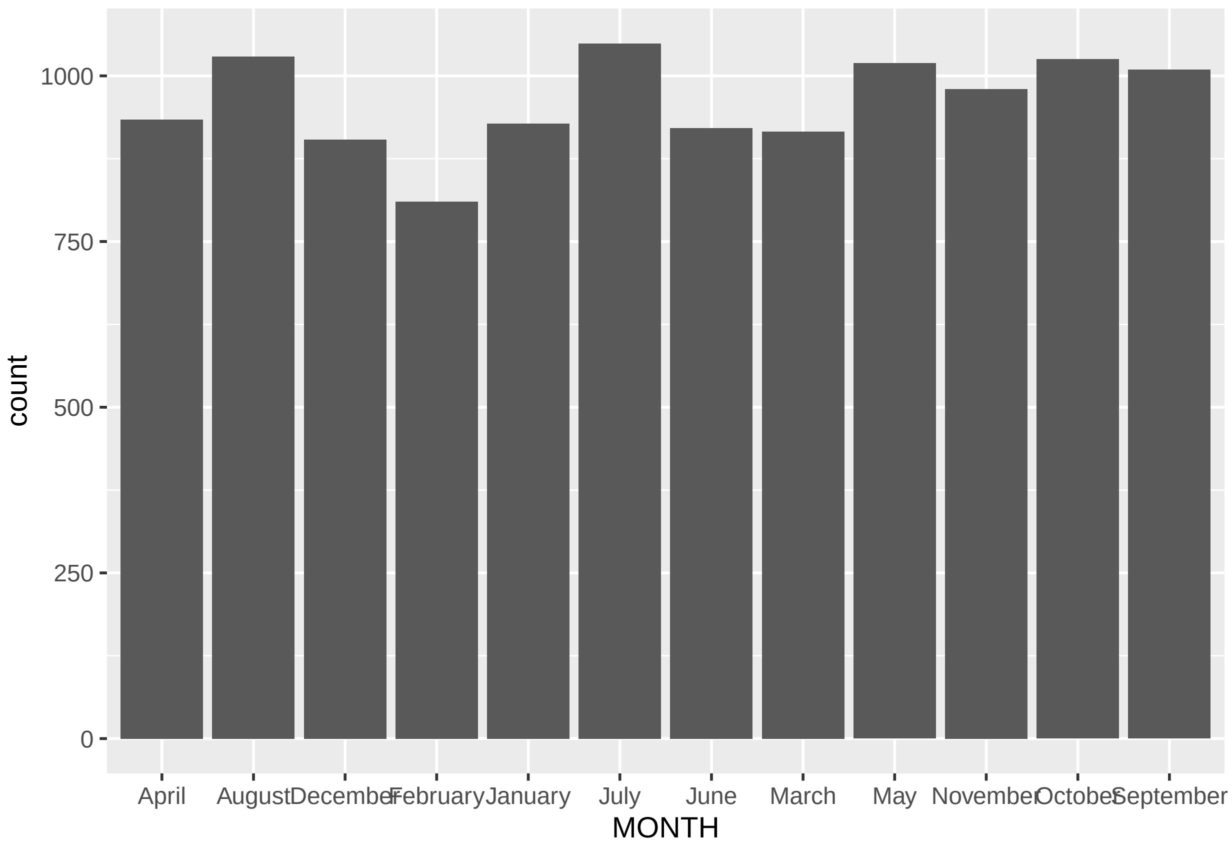
#8.Plot estimated densities of TRAV_SP seperated (colored) by weekdays and weekends.

```
#recode to make weekday and weekend category
c2015<-c2015%>%mutate(PART_OF_WEEK=recode(DAY_WEEK,'Monday'='Weekday','Tuesday'='Weekday','Wednesday'='
ggplot(c2015,aes(x=TRAV_SP,fill=PART_OF_WEEK))+
  geom_density(alpha=0.5)
```

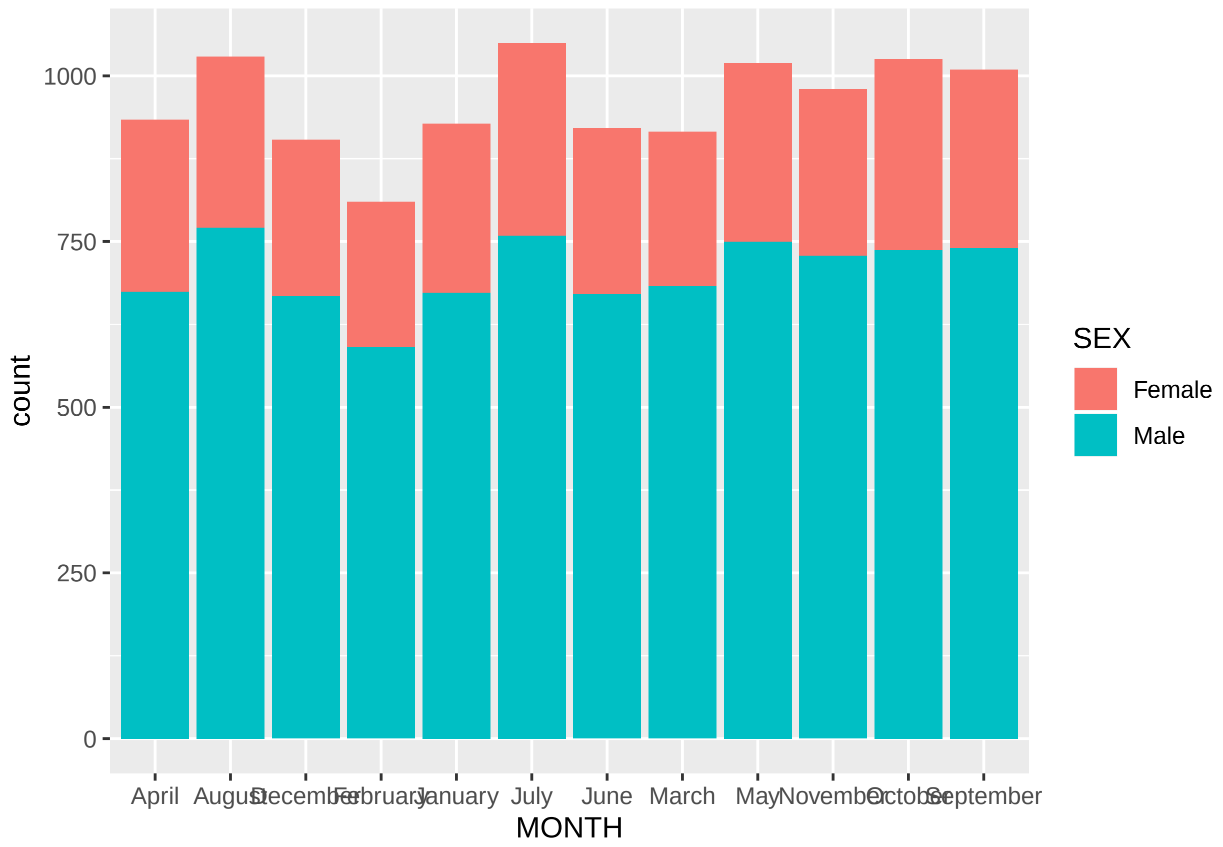


#9.Implement geom_bar on MONTH. Implement geom_bar on MONTH filled by SEX

```
ggplot(c2015,aes(MONTH))+
  geom_bar()
```

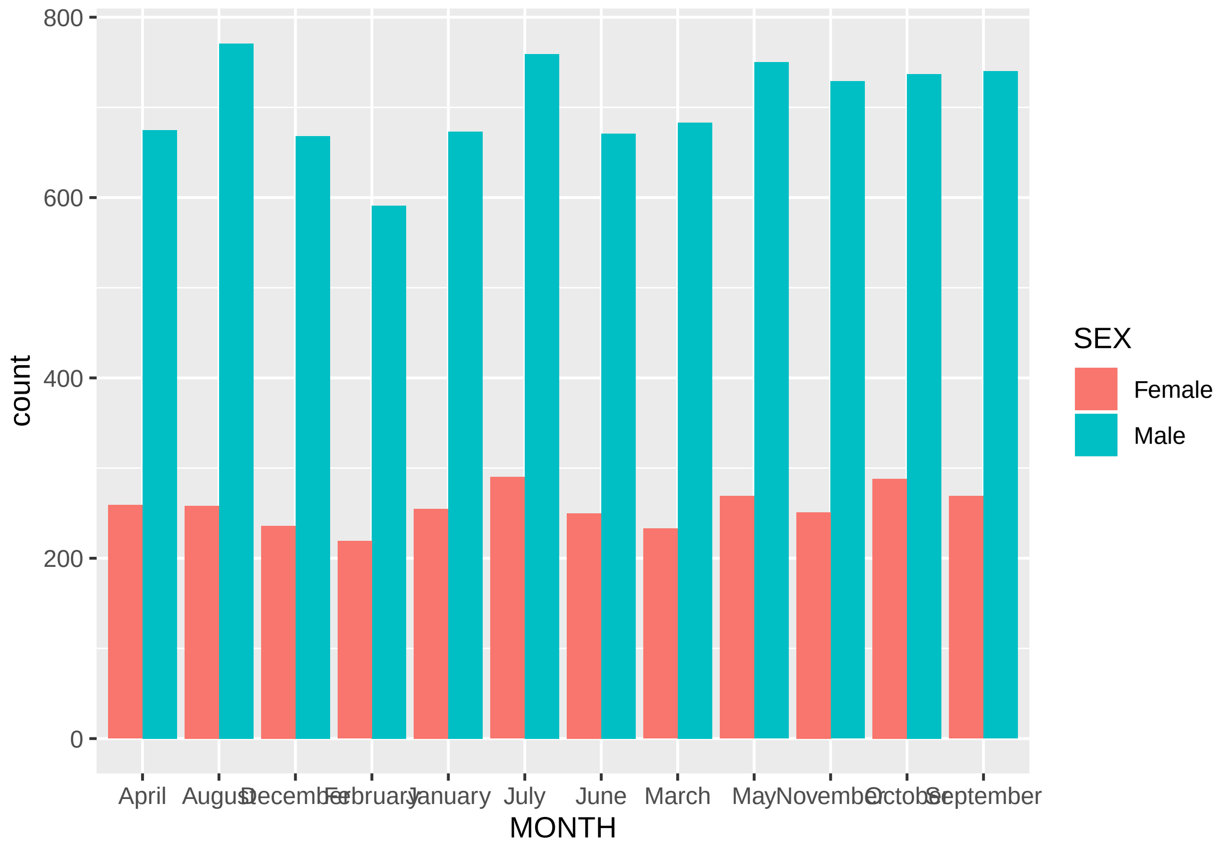


```
ggplot(c2015,aes(MONTH,fill=SEX))+
  geom_bar()
```



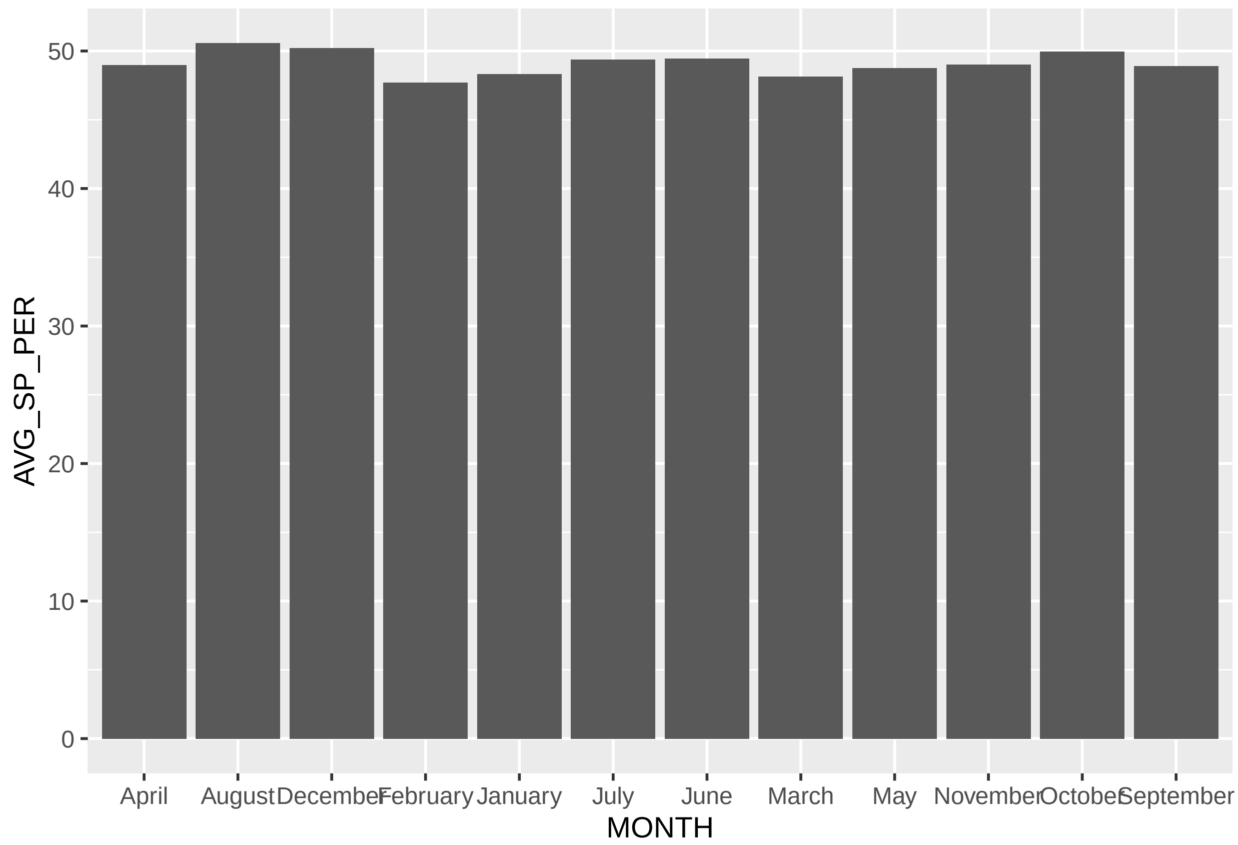
#10.Implement geom_bar on MONTH and SEX with position='dodge'

```
ggplot(c2015,aes(MONTH,fill=SEX))+
  geom_bar(position='dodge')
```



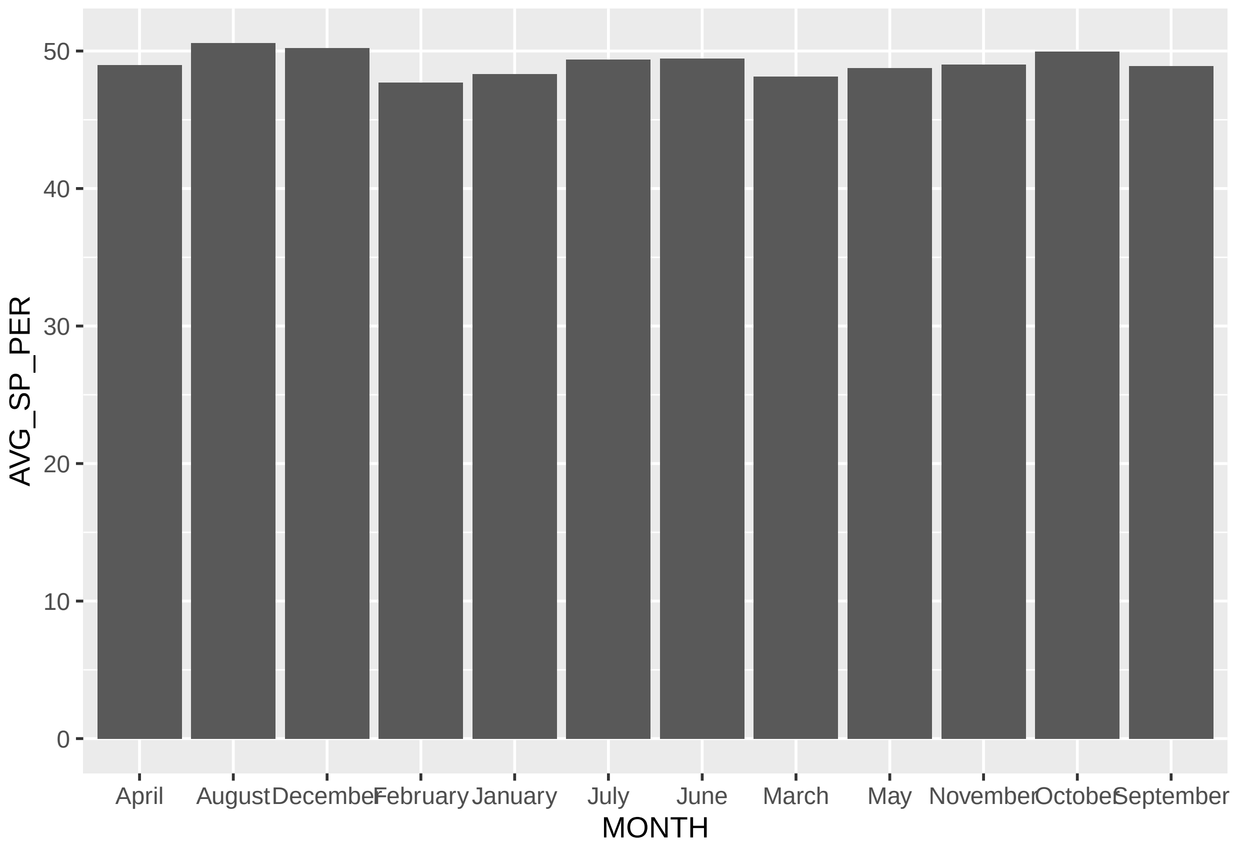
#11.Plot a bar chart of average speeds in months using geom_col

```
monthdata=c2015%>%group_by(MONTH)%>%mutate(AVG_SP_PER=mean(TRAV_SP)/n())
ggplot(monthdata,aes(MONTH))+
  geom_col(aes(y=AVG_SP_PER))
```



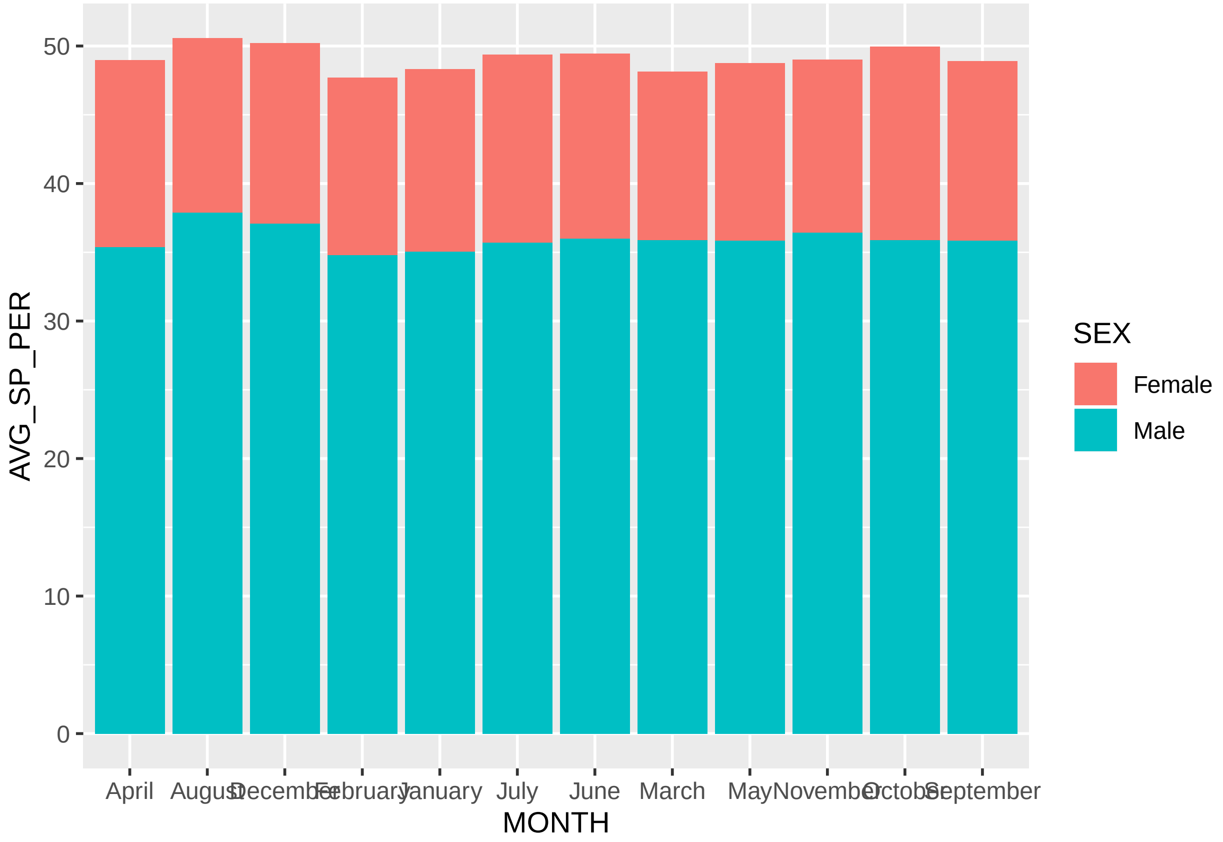
#12.Plot a bar chart of average speeds in months using geom_bar

```
ggplot(monthdata,aes(MONTH))+
  geom_bar(aes(y=AVG_SP_PER),stat='identity')
```



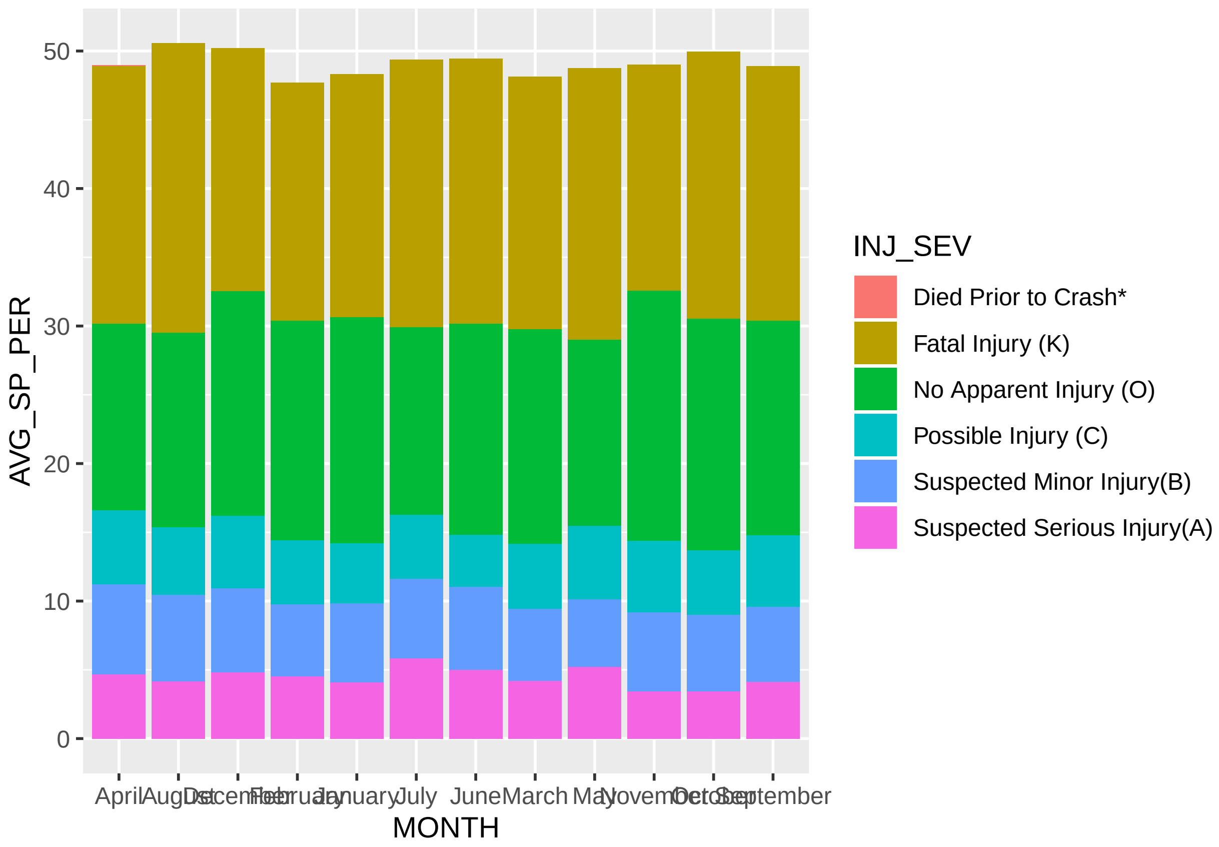
#13.Plot a bar chart of average speeds in months filled by SEX

```
ggplot(monthdata,aes(MONTH,fill=SEX))+
 geom_col(aes(y=AVG_SP_PER))
```



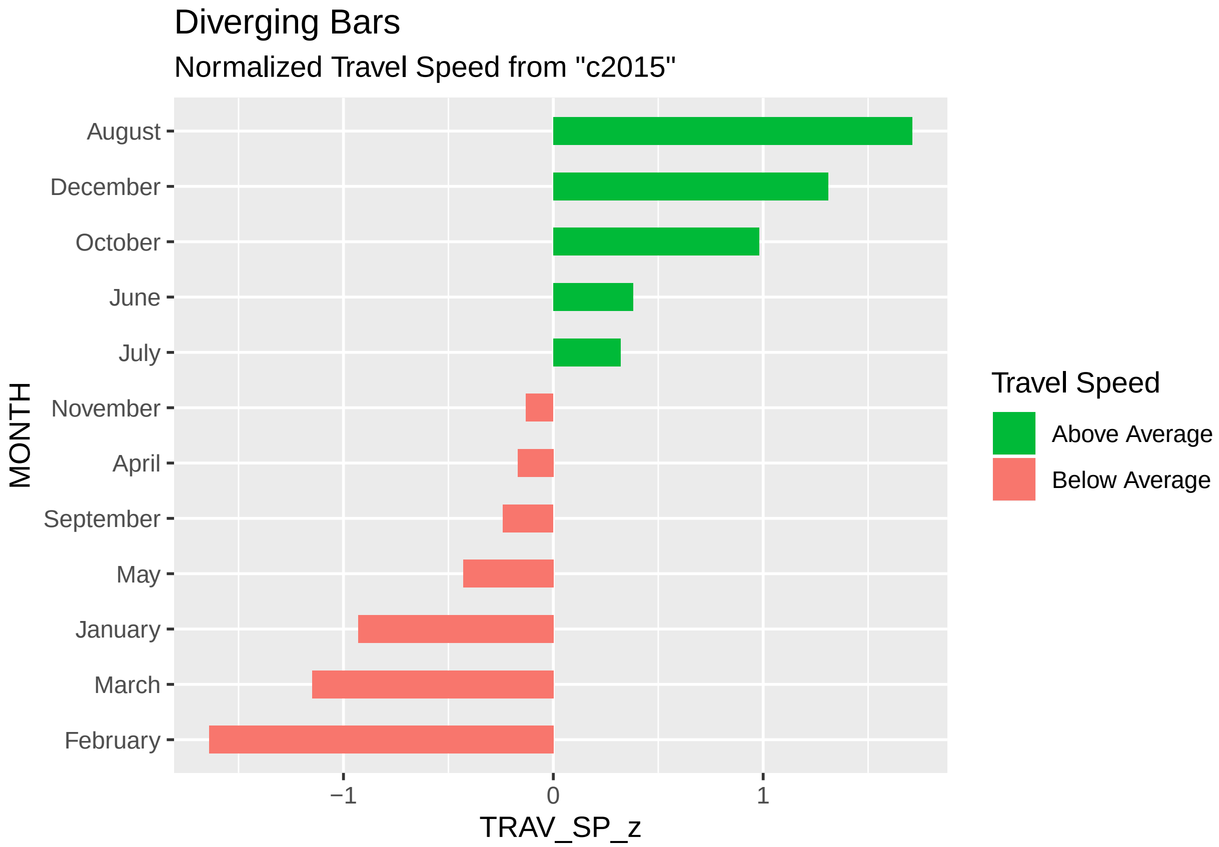
#14.Plot a bar chart of average speeds in months colored by INJ_SEV

```
ggplot(monthdata,aes(MONTH,fill=INJ_SEV))+
  geom_col(aes(y=AVG_SP_PER))
```



#15.Refer to this link to have a similar following plot: ##Horizontal axis is for (monthly) average speed ##The vertical axis is for months ##Color by two colors: one for above overall average speed and the other for below the avarage speed ##The speed on the horizontal axis is standardized

```
#data prep
months<-c2015%>%group_by(MONTH)%>%summarize(mean=mean(TRAV_SP))%>%arrange(mean)
months$TRAV_SP_z<-round((months$mean-mean(months$mean))/sd(months$mean),2)
months$TRAV_SP_type<-ifelse(months$TRAV_SP_z < 0, 'below','above')
months$MONTH<-factor(months$MONTH,levels=months$MONTH)

#plot
ggplot(months,aes(x=MONTH ,y=TRAV_SP_z ,label=TRAV_SP_z))+
  geom_bar(stat='identity',aes(fill=TRAV_SP_type),width=0.5)+
  scale_fill_manual(name='Travel Speed',
                    labels = c("Above Average", "Below Average"),
                    values = c("above"="#00ba38", "below"="#f8766d"))+
  labs(subtitle = 'Normalized Travel Speed from "c2015"',
       title = 'Diverging Bars')+
  coord_flip()
```

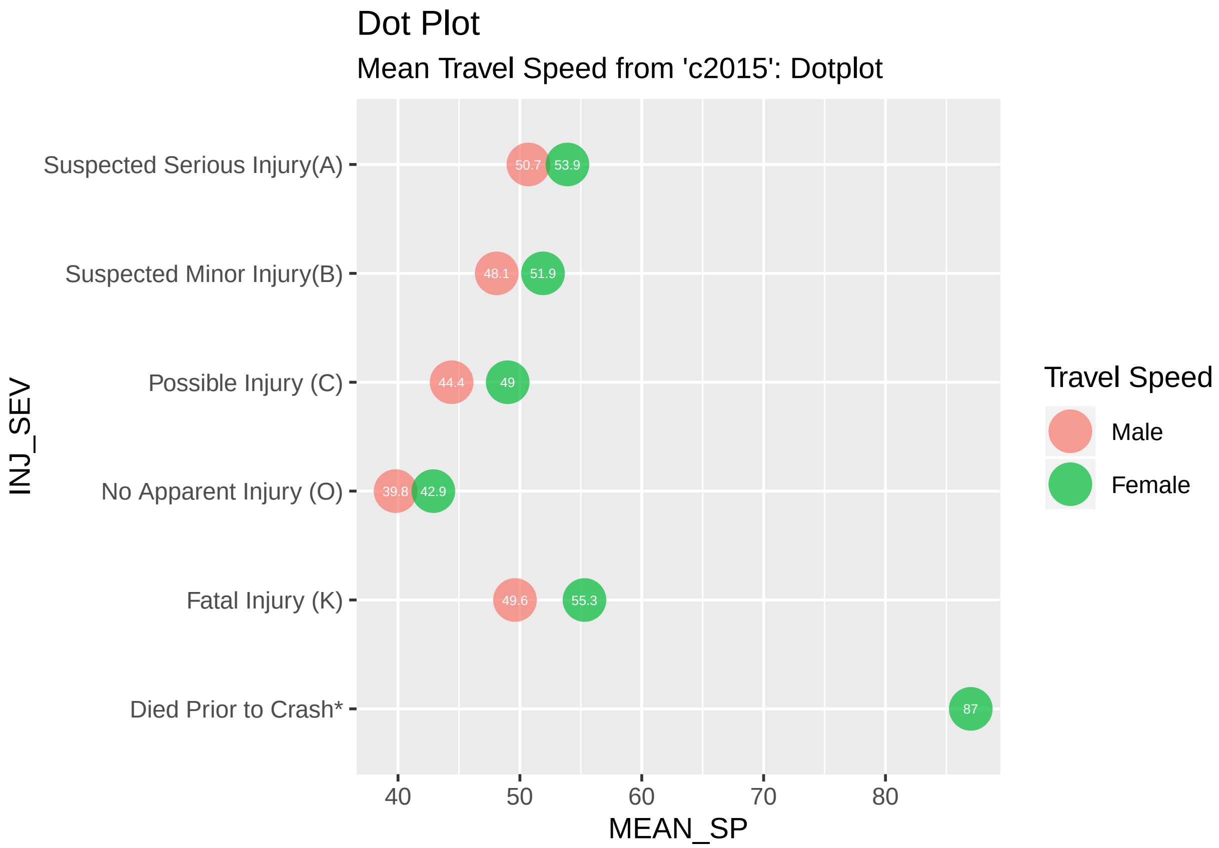


#16.Refer to this link to have a similar following plot: ##Horizontal Axis is for mean speed ##Vertical Axis is for INJ_SEV ##Color by SEX ##he numbers of speed are shown in points.

```
#data prep
injury<-c2015%>%group_by(INJ_SEV,SEX)%>%summarize(mean=mean(TRAV_SP))%>%ungroup()
injury$MEAN_SP<-round(injury$mean,1)
injury
```

```
## # A tibble: 11 x 4
##    INJ_SEV                      SEX     mean MEAN_SP
##    <chr>                        <chr>  <dbl>   <dbl>
##  1 Died Prior to Crash*         Male    87      87
##  2 Fatal Injury (K)             Female  49.6    49.6
##  3 Fatal Injury (K)             Male    55.3    55.3
##  4 No Apparent Injury (O)       Female  39.8    39.8
##  5 No Apparent Injury (O)       Male    42.9    42.9
##  6 Possible Injury (C)          Female  44.4    44.4
##  7 Possible Injury (C)          Male    49.0    49
##  8 Suspected Minor Injury(B)    Female  48.1    48.1
##  9 Suspected Minor Injury(B)    Male    51.9    51.9
## 10 Suspected Serious Injury(A) Female  50.7    50.7
## 11 Suspected Serious Injury(A) Male    53.9    53.9
```

```
#plot
ggplot(injury, aes(x=INJ_SEV, y=MEAN_SP, label=MEAN_SP)) +
  geom_point(stat='identity', aes(col=SEX), size=7,alpha=0.7)  +
  scale_color_manual(name="Travel Speed",
                     labels = c("Male", "Female"),
                     values = c("Male"="#00ba38", "Female"="#f8766d")) +
  geom_text(color="white", size=1.8) +
  labs(title="Dot Plot",
       subtitle="Mean Travel Speed from 'c2015': Dotplot") +
  ylim(39,87) +
  coord_flip()
```



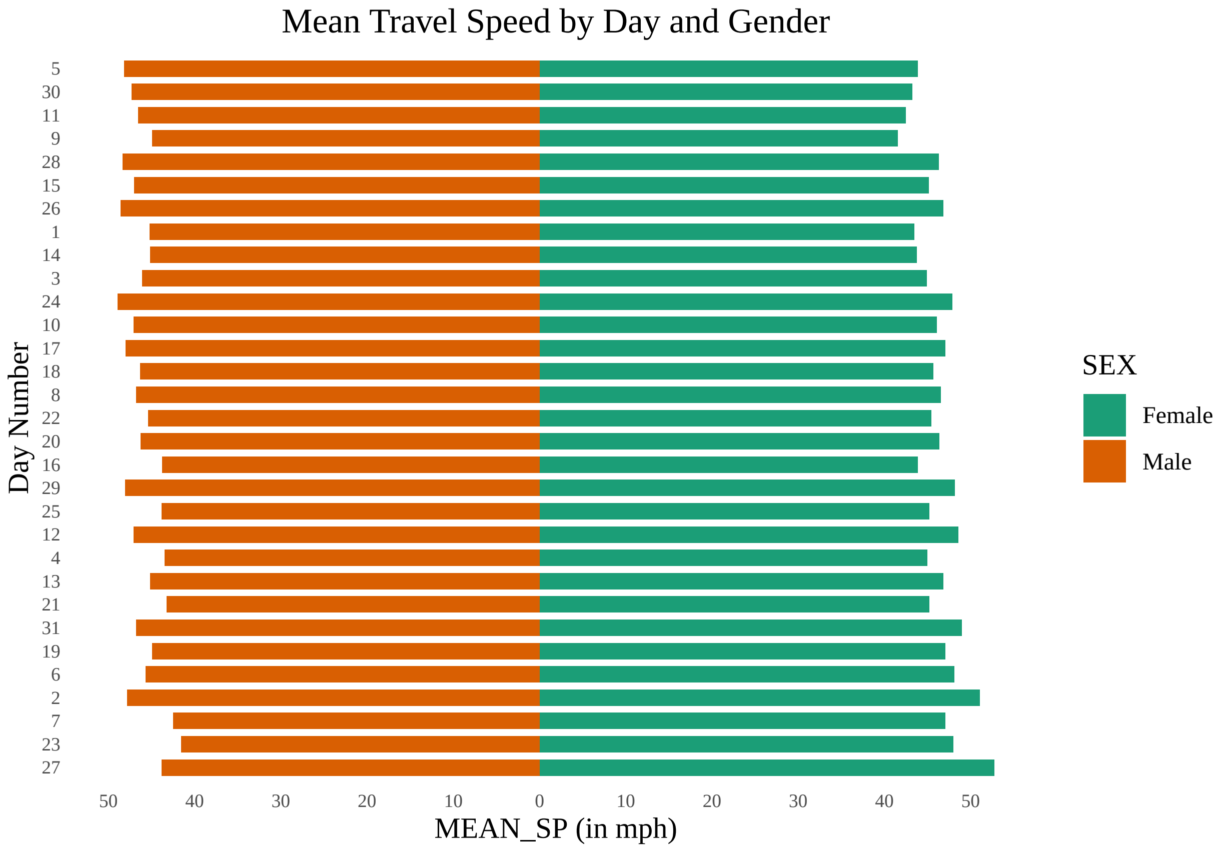
```
######TRY TO FIX THIS BETTER TO MAKE IT DIVERGE!!!!!!!!!!!
```

#17.Refer to this link to have a similar following plot: ##Horizontal Axis is for speed ##Vertical Axis is for DAY ##Color by SEX ##The should be a invisible vertical line seperating the two sexes.

```
#data/plot prep
library(ggthemes)
days<-c2015%>%group_by(DAY,SEX)%>%summarize(mean=mean(TRAV_SP))%>%arrange(mean)%>%ungroup()
days$MEAN_SP<-round(days$mean,2)
days$MEAN_SP[days$SEX=='Male']=-days$MEAN_SP
```

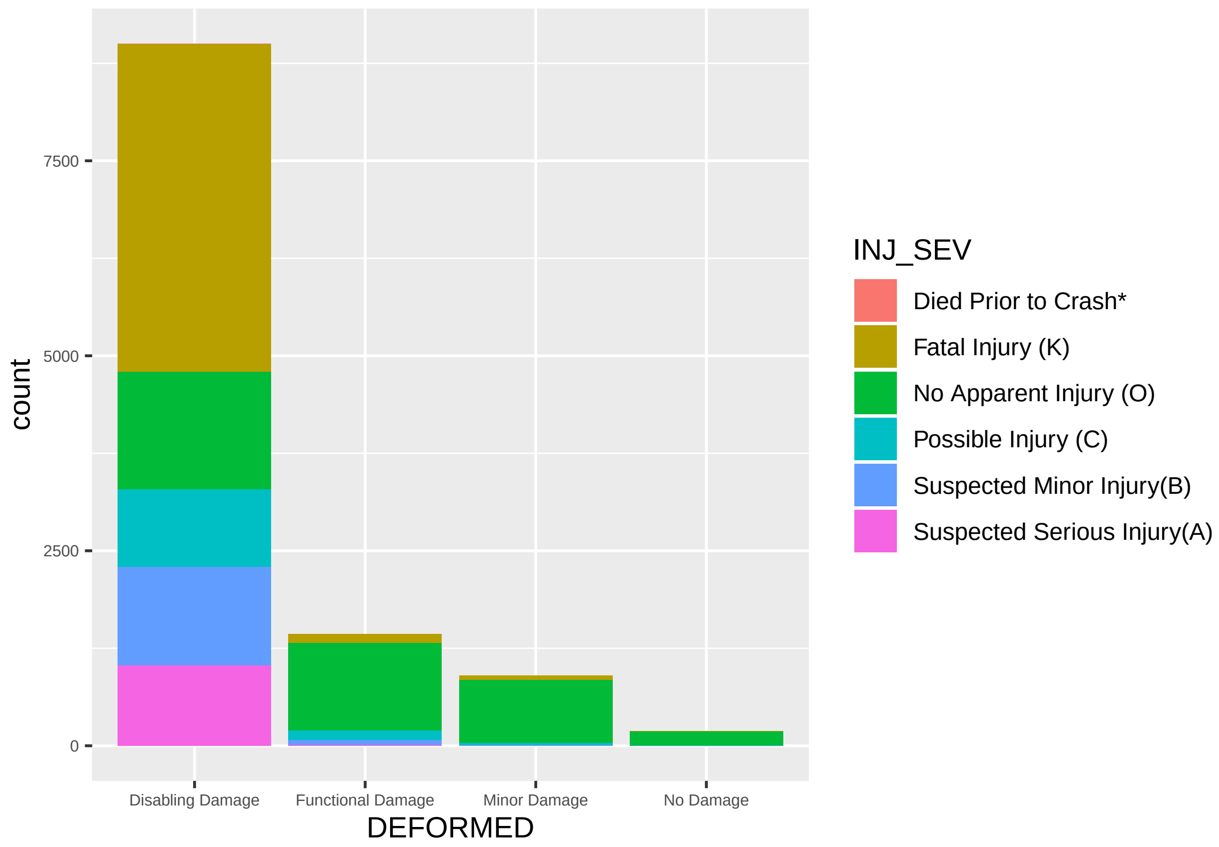
```
## Warning in days$MEAN_SP[days$SEX == "Male"] = -days$MEAN_SP: number of
## items to replace is not a multiple of replacement length
```

```
brks <- seq(-50,50, 10)
lbls <- paste0(as.character(c(seq(50, 0,-10), seq(10, 50, 10)), "mph"))

#plot
ggplot(days,aes(x=reorder(DAY,-MEAN_SP),y=MEAN_SP,fill=SEX))+
  geom_bar(stat='identity',width=0.7)+
  scale_y_continuous(breaks=brks,
                     labels=lbls)+
  coord_flip()+
  labs(title = 'Mean Travel Speed by Day and Gender')+
  theme_tufte()+
  theme(plot.title=element_text(hjust=0.5),
        axis.ticks=element_blank(),
        axis.text=element_text(size=7))+
  scale_fill_brewer(palette='Dark2')+
  xlab('Day Number')+
  ylab('MEAN_SP (in mph)')
```



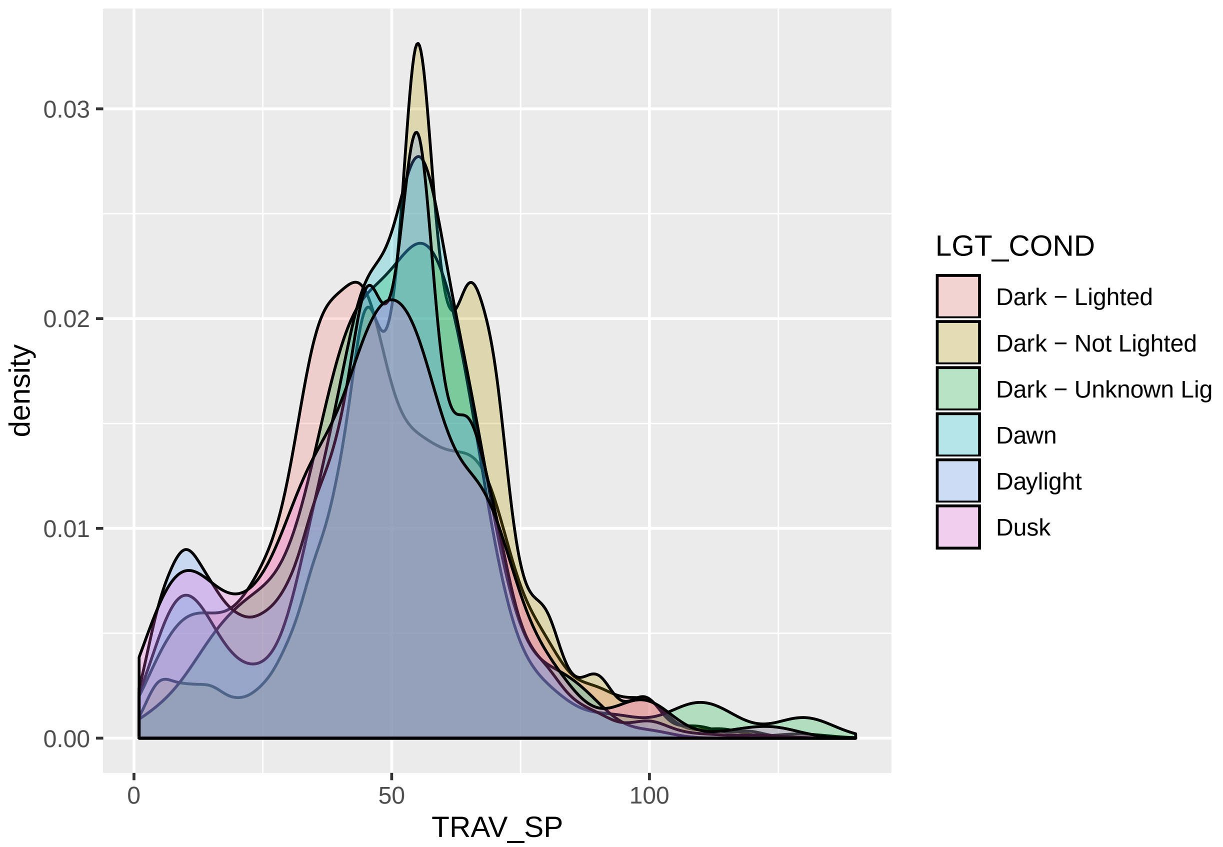
#18-20. Generate three other interesting graphs from the dataset.

```
#Question 18

ggplot(c2015,aes(x=DEFORMED,fill=INJ_SEV))+
  geom_bar()+
  theme(axis.text=element_text(size=6))
```



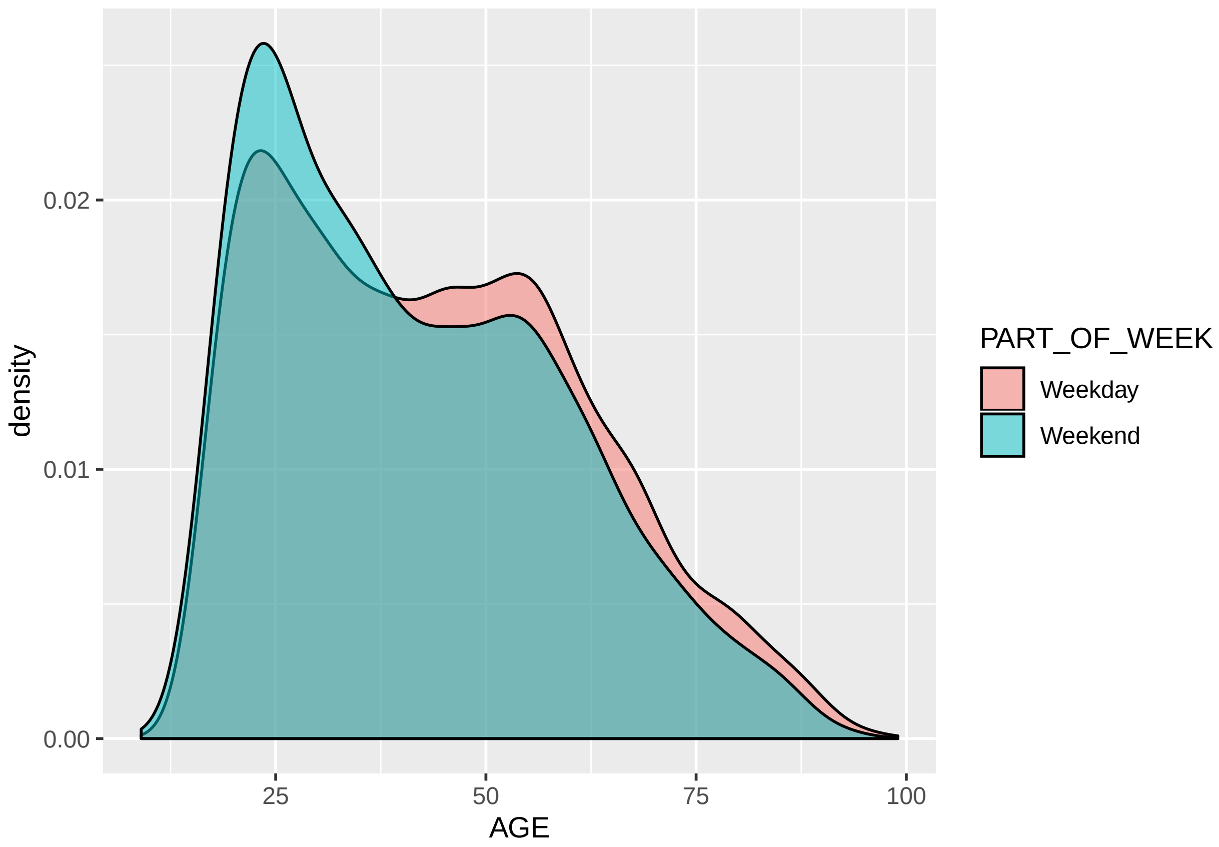
```
#This graph shows the count for each type of damage of the car colored by the severity of the injury of
```

```
#Question 19

ggplot(c2015,aes(x=TRAV_SP,fill=LGT_COND))+
  geom_density(alpha=0.25)
```



```
#This density graph shows the travel speed of drivers in accidents colored by the lighting condition at
```

```
#Question 20
ggplot(c2015,aes(AGE,fill=PART_OF_WEEK))+
  geom_density(alpha=0.5)
```



```
#This graph shows the distribution of ages of drivers in accidents based on whether the accident occurr
```