Assigment 5

Jeff Moise

9/30/2019

This assignment works with the c2015 dataset.

1. Clean the data for easy graphing. Do the follows to clean and reduce the size of the data Remove all observations that have a cell being either (1) NA, (2)'Unknown', (3)'Not Rep', or (4)'Not Reported' Remove all observations that have a cell containing either (1)'Unknown', (2)'Not Rep', or (3)'Not Reported'. For instance, observations with DRINKING variable being Unknown (Police Reported) will be removed. Fix TRAV_SP and AGE (following previous assignments) so that they are both numerics. Filter so that there are only drivers in the data

```
path <- "C:/Users/student/Documents/RStudio/c2015.xlsx"
library(readxl)
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(dplyr)
library(stringr)
d=read_excel(path)
head(d)
```

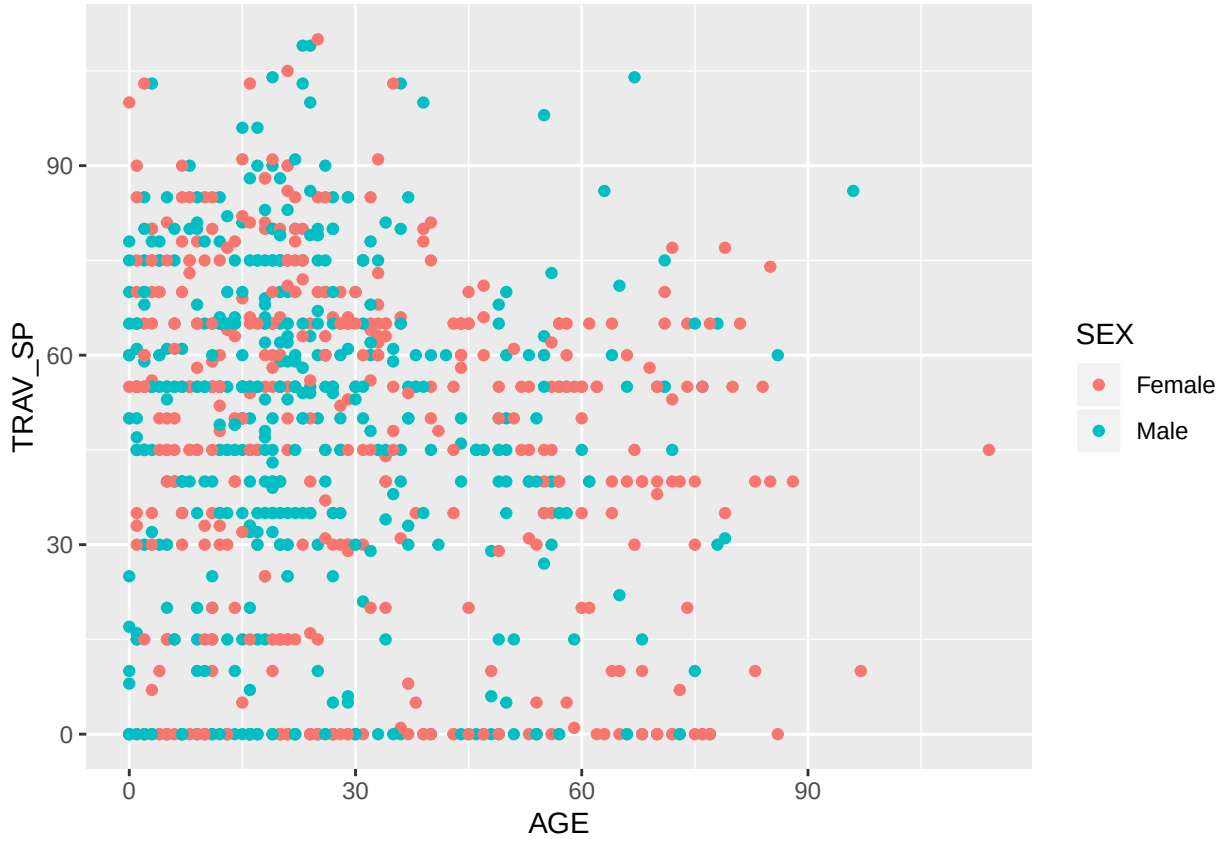
```
## # A tibble: 6 x 28
##   STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE AGE   SEX
##   <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <chr> <chr>
## 1 Alab~   10001      1      1    127     1 Janu~     2     40 68    Male
## 2 Alab~   10002      1      1     83     1 Janu~    22     13 49    Male
## 3 Alab~   10003      1      1     11     1 Janu~     1     25 31    Male
## 4 Alab~   10003      1      2     11     1 Janu~     1     25 20    Fema~
## 5 Alab~   10004      1      1     45     4 Janu~     0     57 40    Male
## 6 Alab~   10005      1      1     45     7 Janu~     7      9 24    Male
## # ... with 17 more variables: PER_TYP <chr>, INJ_SEV <chr>,
## #   SEAT_POS <chr>, DRINKING <chr>, YEAR <dbl>, MAN_COLL <chr>,
## #   OWNER <chr>, MOD_YEAR <chr>, TRAV_SP <chr>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

```r
#Remove NA, Unknown, Not Rep, Not Reported
d = d %>% filter_all(~!is.na(.))
d = d %>% filter_all(~!(.=="Unknown"))
d = d %>% filter_all(~!(.=="Not Rep"))
d = d %>% filter_all(~!(.==str_detect(.,"Not Rep")))
d = d %>% filter_all(~!(.==str_detect(.,"Unknown")))
d = d %>% filter_all(~!(.=="Not Reported"))
d = d %>% filter_all(~!(d$SEAT_POS == "Front Seat, Left Side"))

d$TRAV_SP[d$TRAV_SP=='Stopped'] <- '0'
d$TRAV_SP<- stringr::str_replace(d$TRAV_SP," MPH", "")
d$TRAV_SP <- as.numeric(d$TRAV_SP)


d<-d %>%
  mutate(AGE=case_when(
    AGE=='Less than 1' ~ '0',
    TRUE ~ (AGE)))
d$AGE <- as.numeric(d$AGE)

head(d)
```

```
## # A tibble: 6 x 28
##   STATE ST_CASE VEH_NO PER_NO COUNTY   DAY MONTH  HOUR MINUTE   AGE SEX
##   <chr>   <dbl>  <dbl>  <dbl>  <dbl> <dbl> <chr> <dbl>  <dbl> <dbl> <chr>
## 1 Alab~   10687      1      2     93    17 Dece~    21     15    18 Fema~
## 2 Alas~   20028      2      2     20     3 July      1      4    29 Fema~
## 3 Ariz~   40118      1      2     13    15 Febr~    20     46    34 Fema~
## 4 Ariz~   40189      1      2     25    26 March    22      7    44 Male
## 5 Ariz~   40245      1      2     19    28 April    23      9    22 Fema~
## 6 Ariz~   40245      1      3     19    28 April    23      9    32 Male
## # ... with 17 more variables: PER_TYP <chr>, INJ_SEV <chr>,
## #   SEAT_POS <chr>, DRINKING <chr>, YEAR <dbl>, MAN_COLL <chr>,
## #   OWNER <chr>, MOD_YEAR <chr>, TRAV_SP <dbl>, DEFORMED <chr>,
## #   DAY_WEEK <chr>, ROUTE <chr>, LATITUDE <dbl>, LONGITUD <dbl>,
## #   HARM_EV <chr>, LGT_COND <chr>, WEATHER <chr>
```

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

```r
library(ggplot2)
ggplot(d, aes(AGE, TRAV_SP, color=SEX)) +
  geom_point()
```

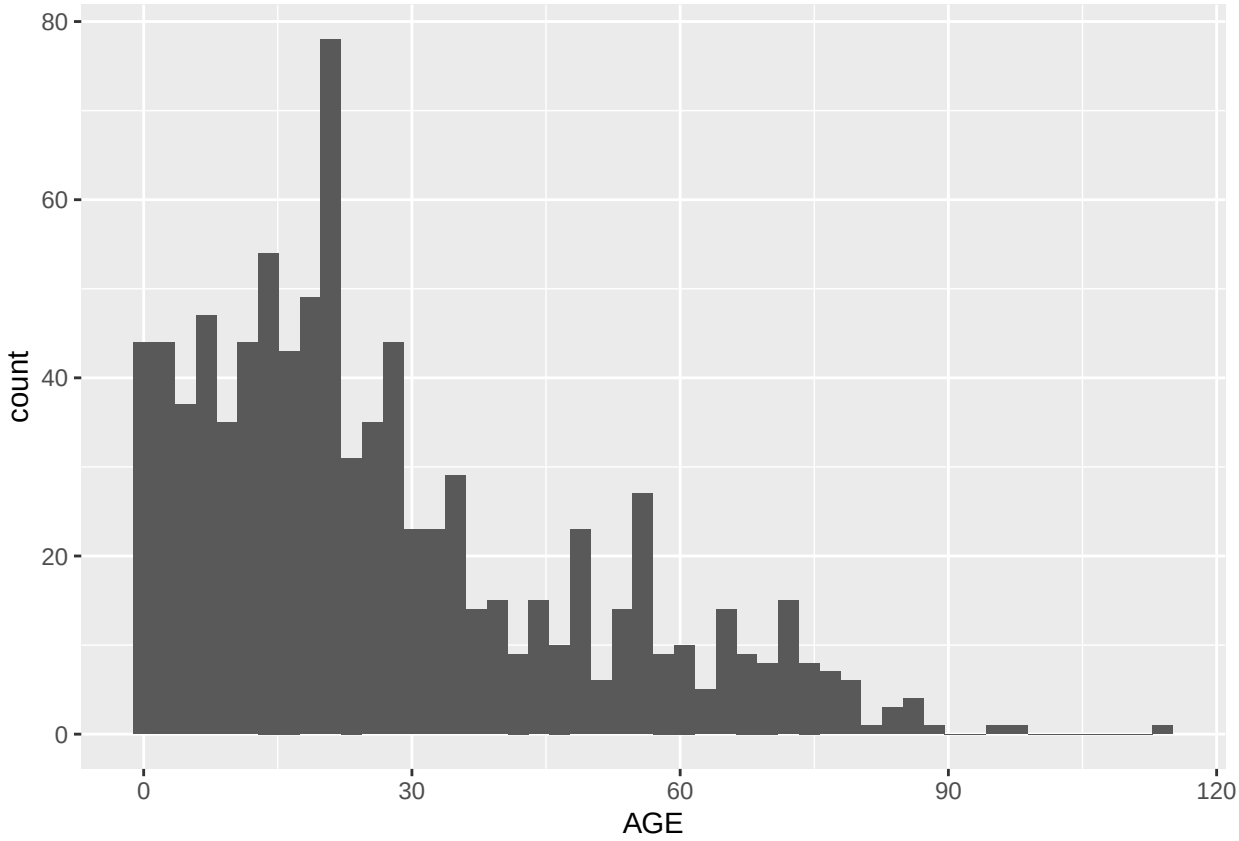


3. There is overplotting in 2. Overplotting is when many points are duplicated on the graph. Use geom_jitter instead of geom_point for 2. to avoid overplotting.

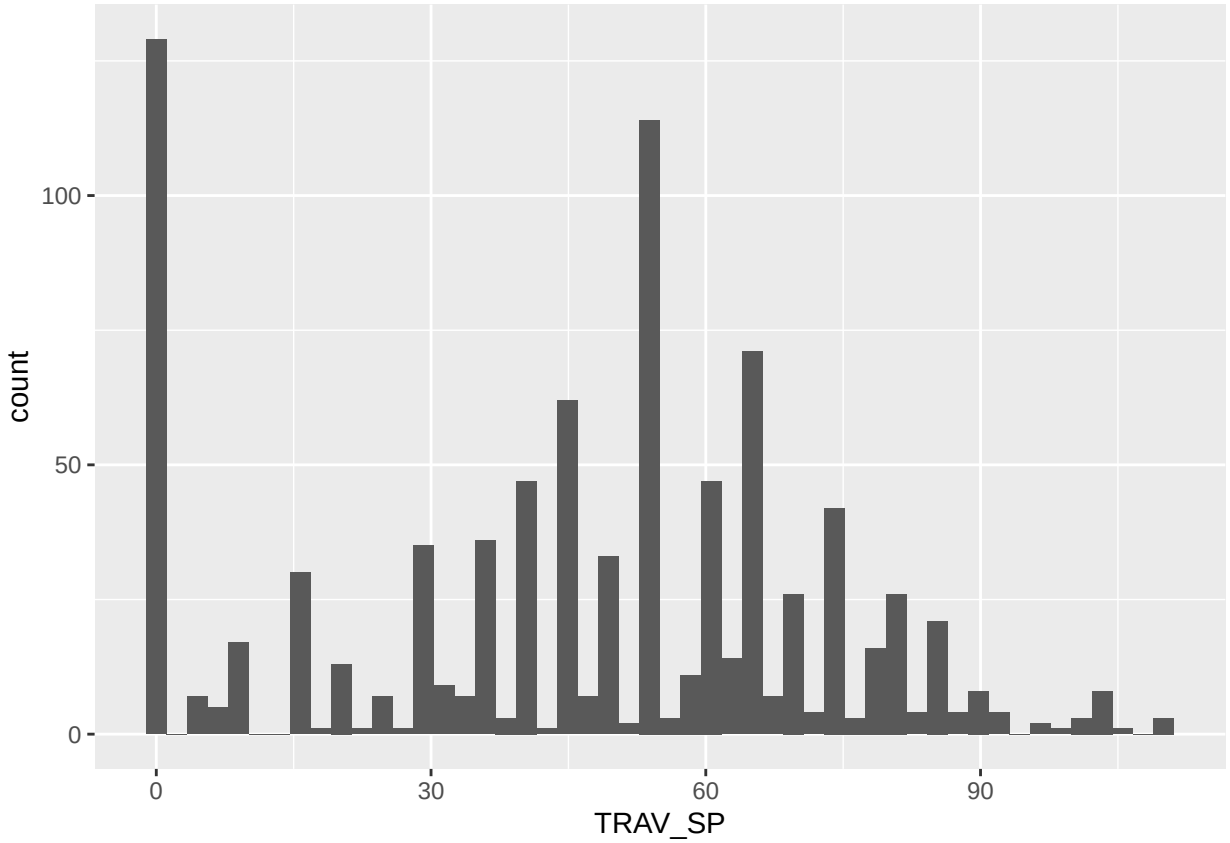
```
library(ggplot2)
ggplot(d, aes(AGE, TRAV_SP, color=SEX)) +
  geom_jitter()
```

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

```
library(ggplot2)
ggplot(d, aes(AGE)) +
  geom_histogram(bins= 50)
```




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



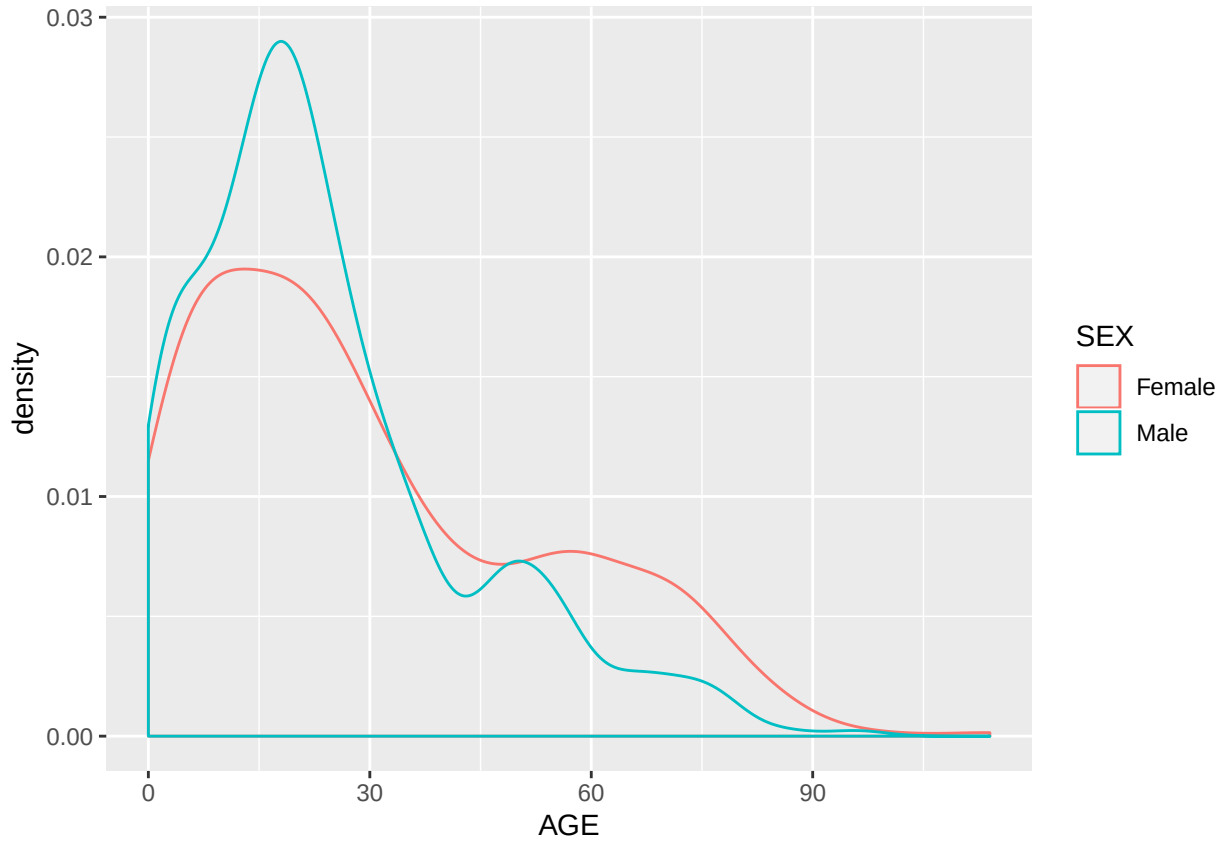
5. Plot a histogram of AGE coloring (fill) by SEX.

```
ggplot(d, aes(AGE, fill=SEX)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

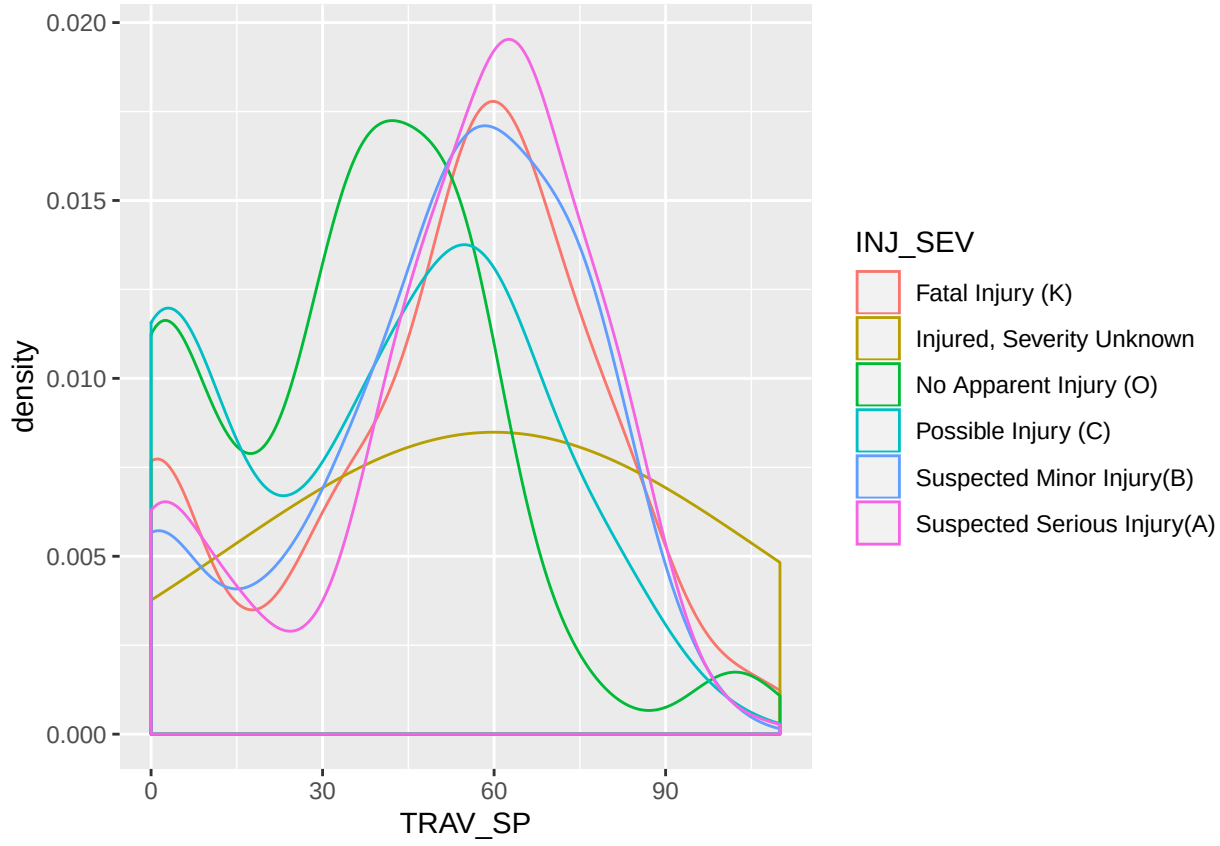
6. Using geom_density to plot estimated densities of AGE colored by SEX.

```
ggplot(d, aes(AGE, color=SEX)) +
  geom_density()
```



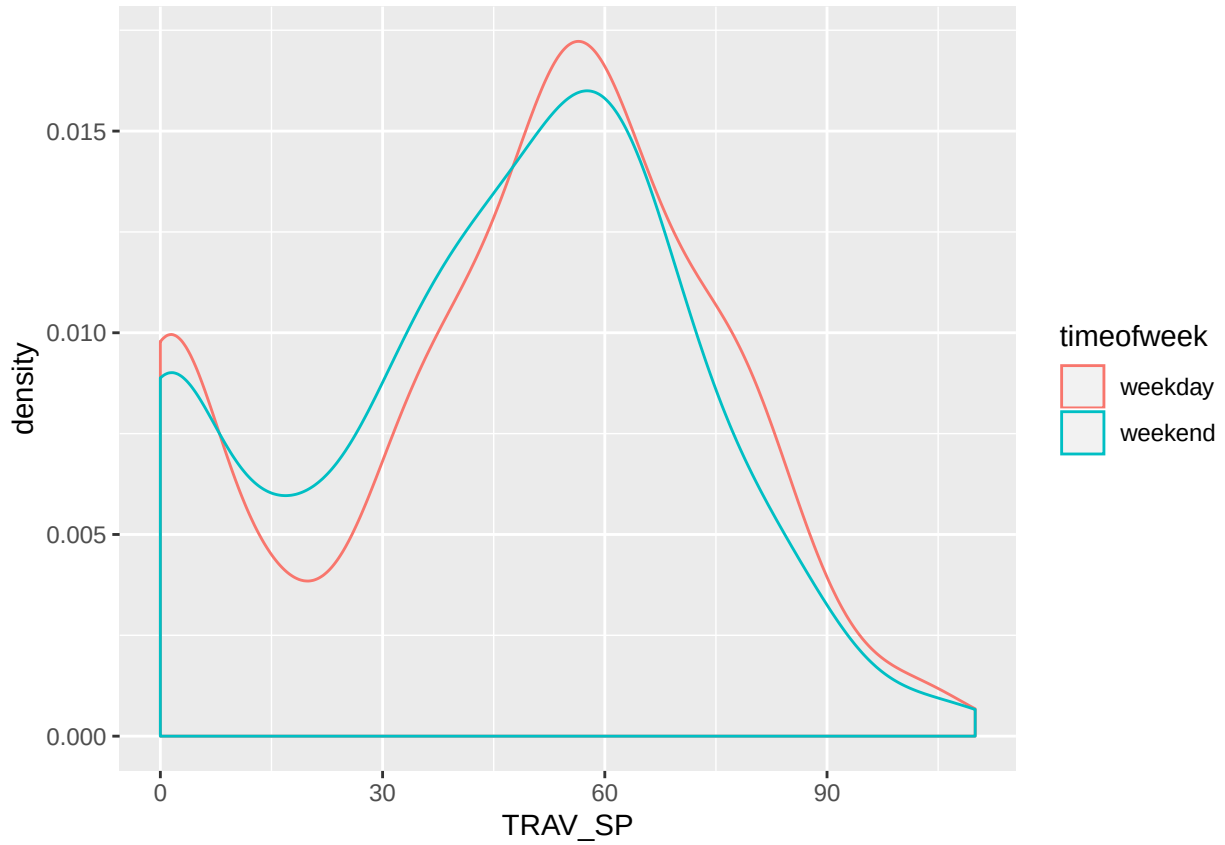
7. Plot estimated densities of TRAV_SP colored by INJ_SEV.

```
ggplot(d, aes(TRAV_SP, color=INJ_SEV)) +
  geom_density()
```



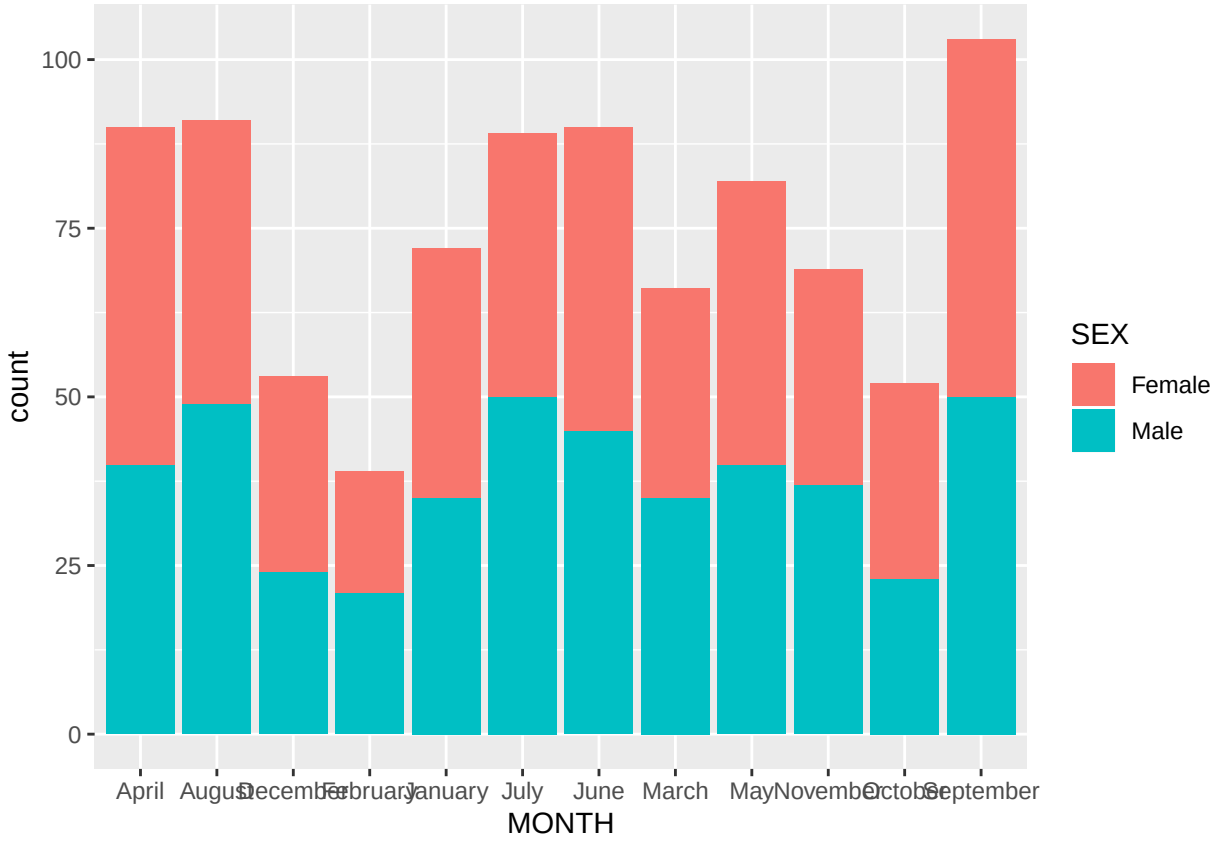
8. Plot estimated densities of TRAV_SP seperated (colored) by weekdays and weekends.

```
d<- d %>%
  mutate(
   timeofweek = case_when(
    DAY_WEEK=="Monday" ~ "weekday",
        DAY_WEEK=="Tuesday" ~ "weekday",
        DAY_WEEK=="Wednesday" ~ "weekday",
        DAY_WEEK=="Thursday" ~ "weekday",
        DAY_WEEK=="Friday" ~ "weekday",
    DAY_WEEK=="Saturday" ~ "weekend",
      DAY_WEEK=="Sunday" ~ "weekend"))

ggplot(d, aes(TRAV_SP, color=timeofweek)) +
  geom_density()
```



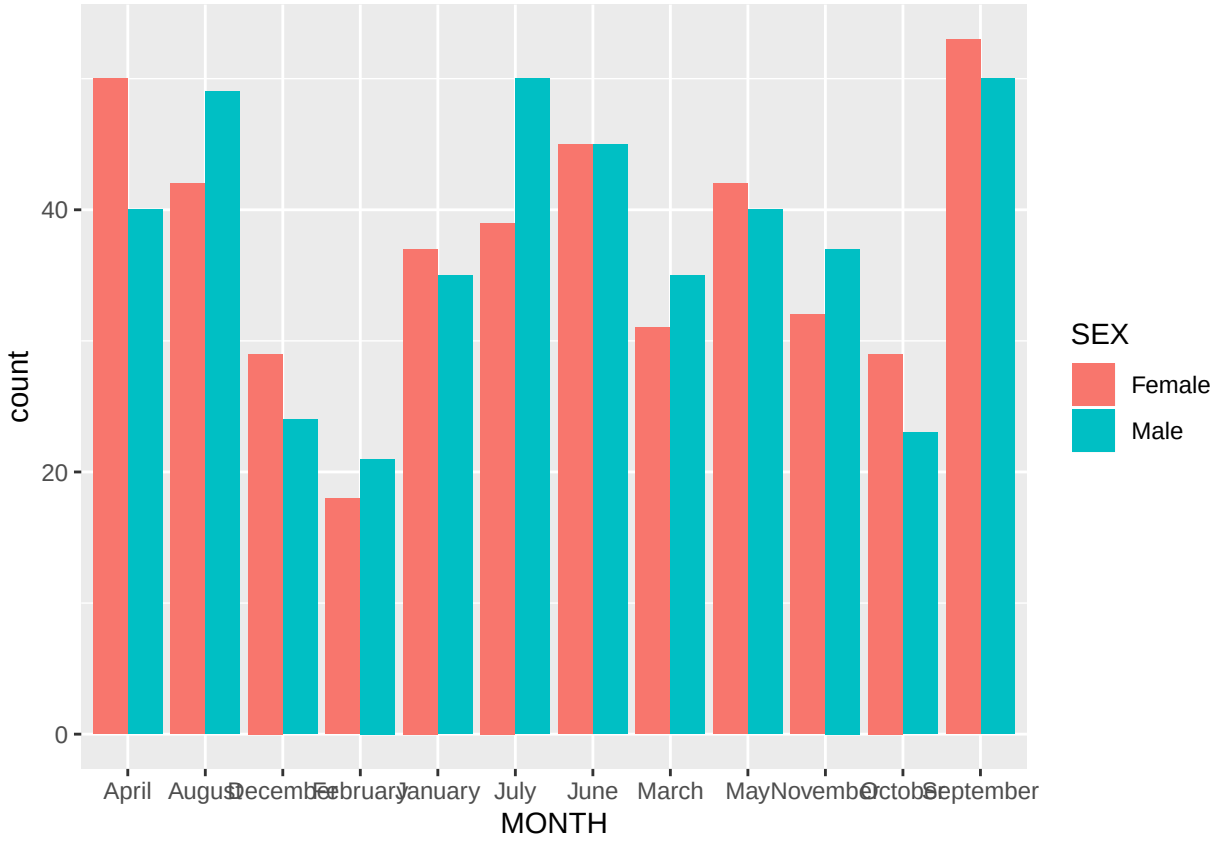
9. Implement geom_bar on MONTH. Implement geom_bar on MONTH filled by SEX

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



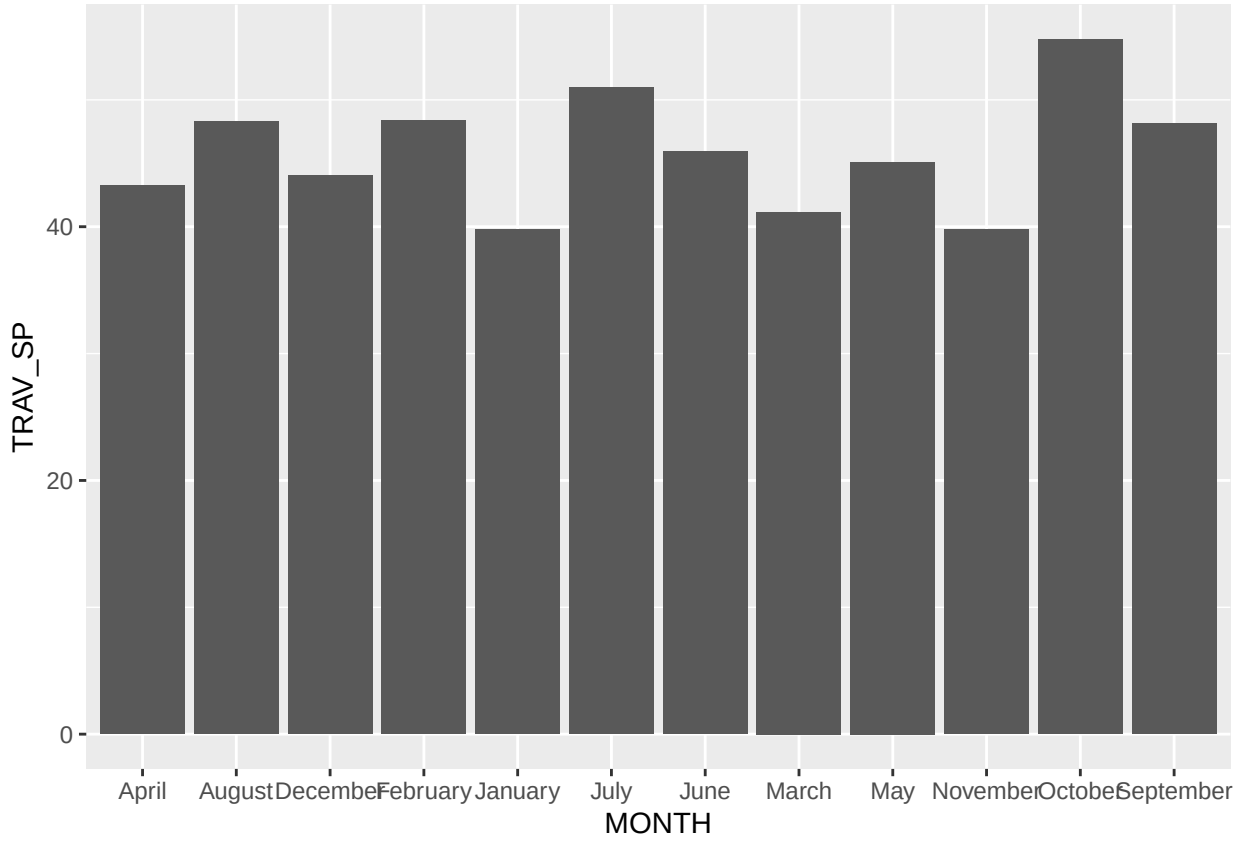
10. Implement geom_bar on MONTH and SEX with position='dodge'

```
ggplot(d, aes(MONTH, fill=SEX)) +
  geom_bar(position="dodge")
```



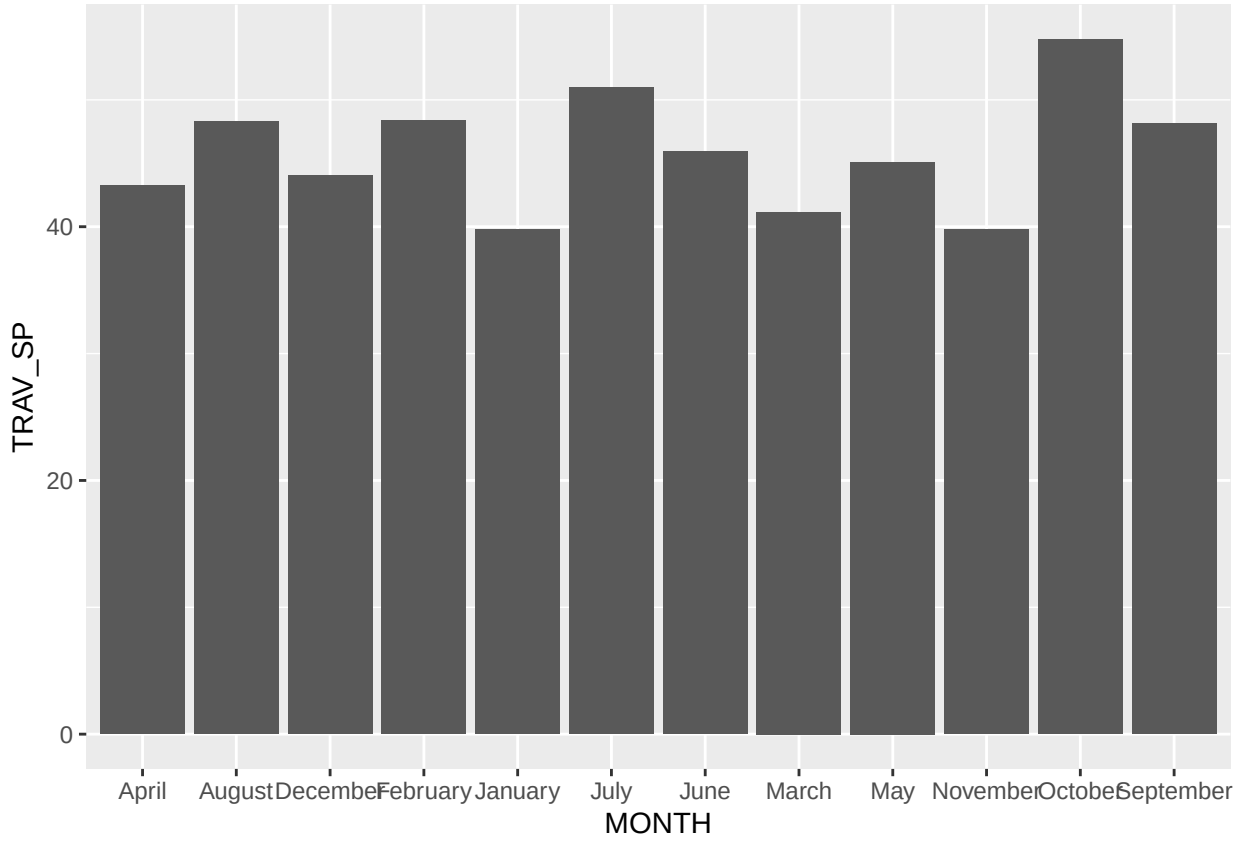
11. Plot a bar chart of average speeds in months using geom_col

```
ggplot(d, aes(MONTH, TRAV_SP)) +
  stat_summary(fun.y = "mean", geom = "col")
```



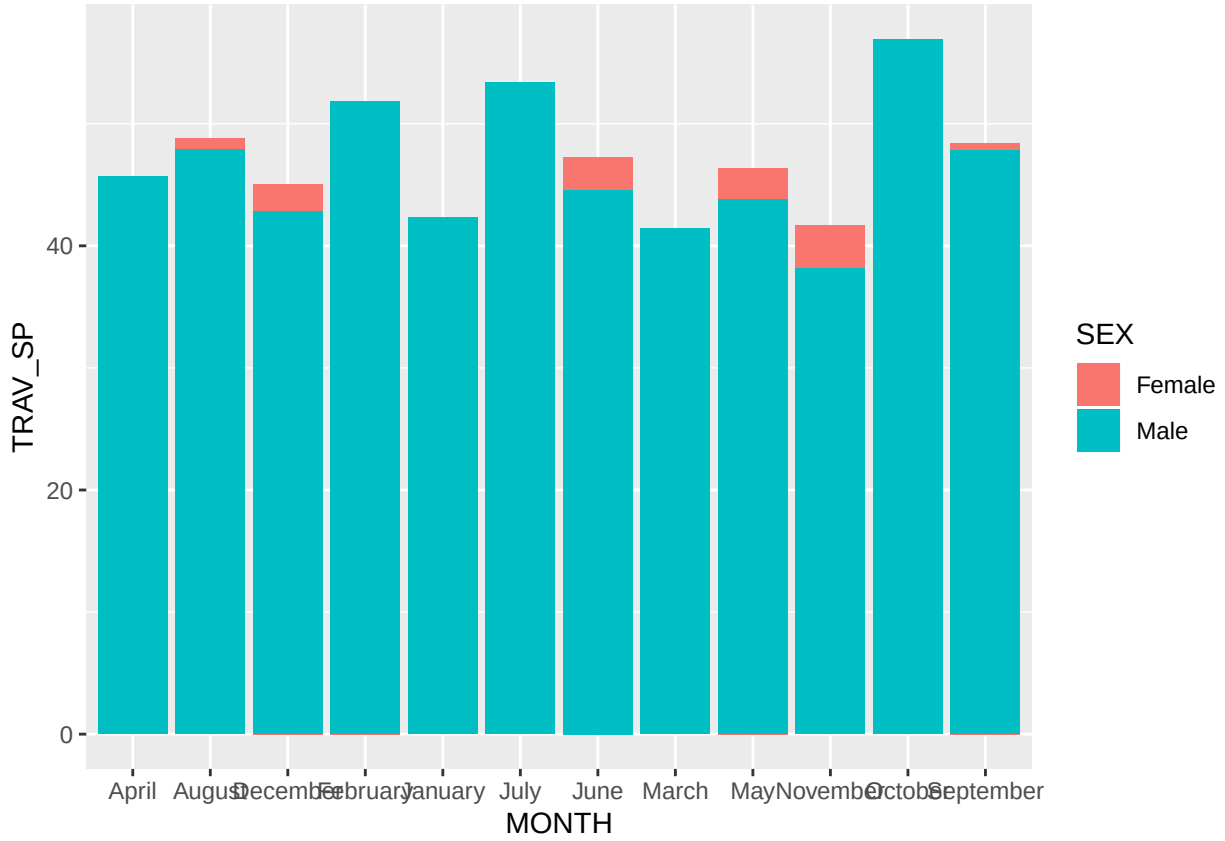
12. Plot a bar chart of average speeds in months using geom_bar

```
ggplot(d, aes(MONTH, TRAV_SP)) +
  stat_summary(fun.y = "mean", geom = "bar")
```



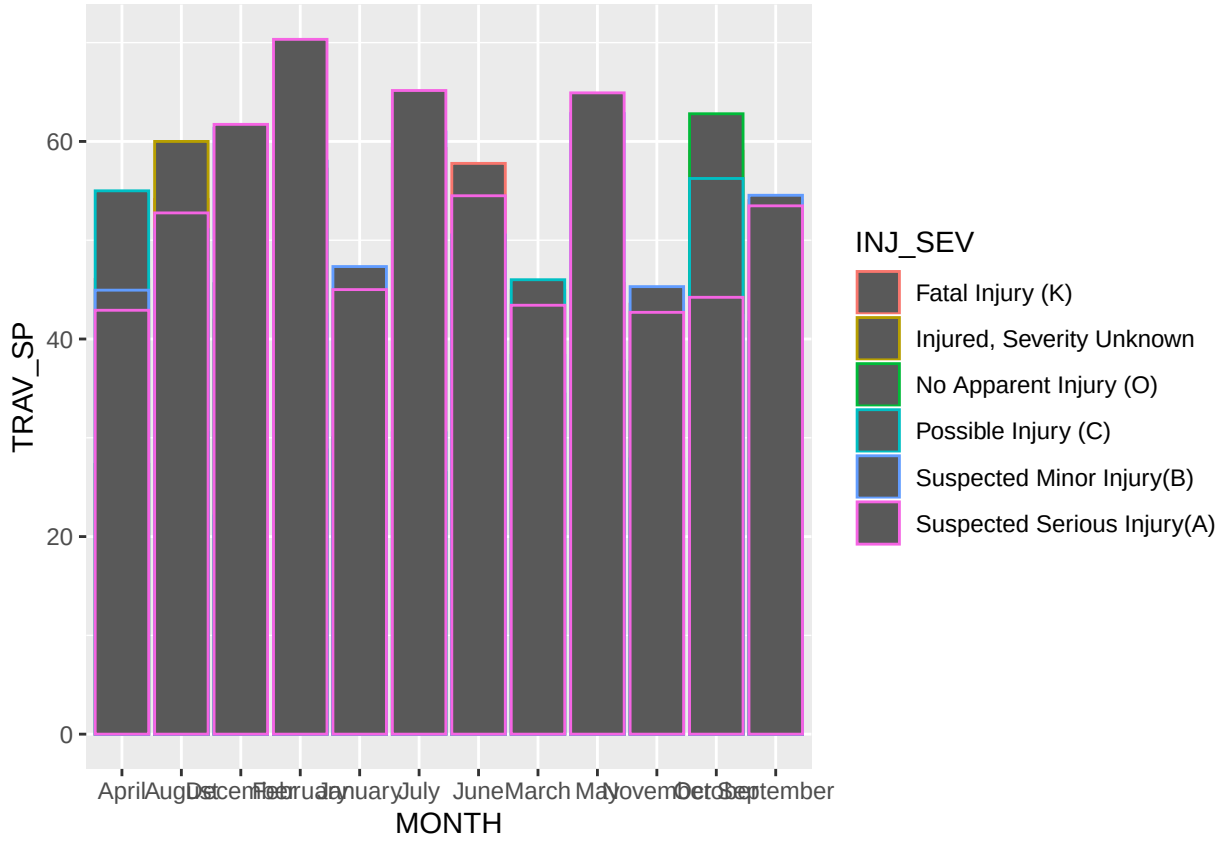
13. Plot a bar chart of average speeds in months filled by SEX

```
ggplot(d, aes(MONTH, TRAV_SP, fill=SEX)) +
  stat_summary(fun.y = "mean", geom = "col")
```



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

```
ggplot(d, aes(MONTH, TRAV_SP, color=INJ_SEV)) +
  stat_summary(fun.y = "mean", geom = "col")
```

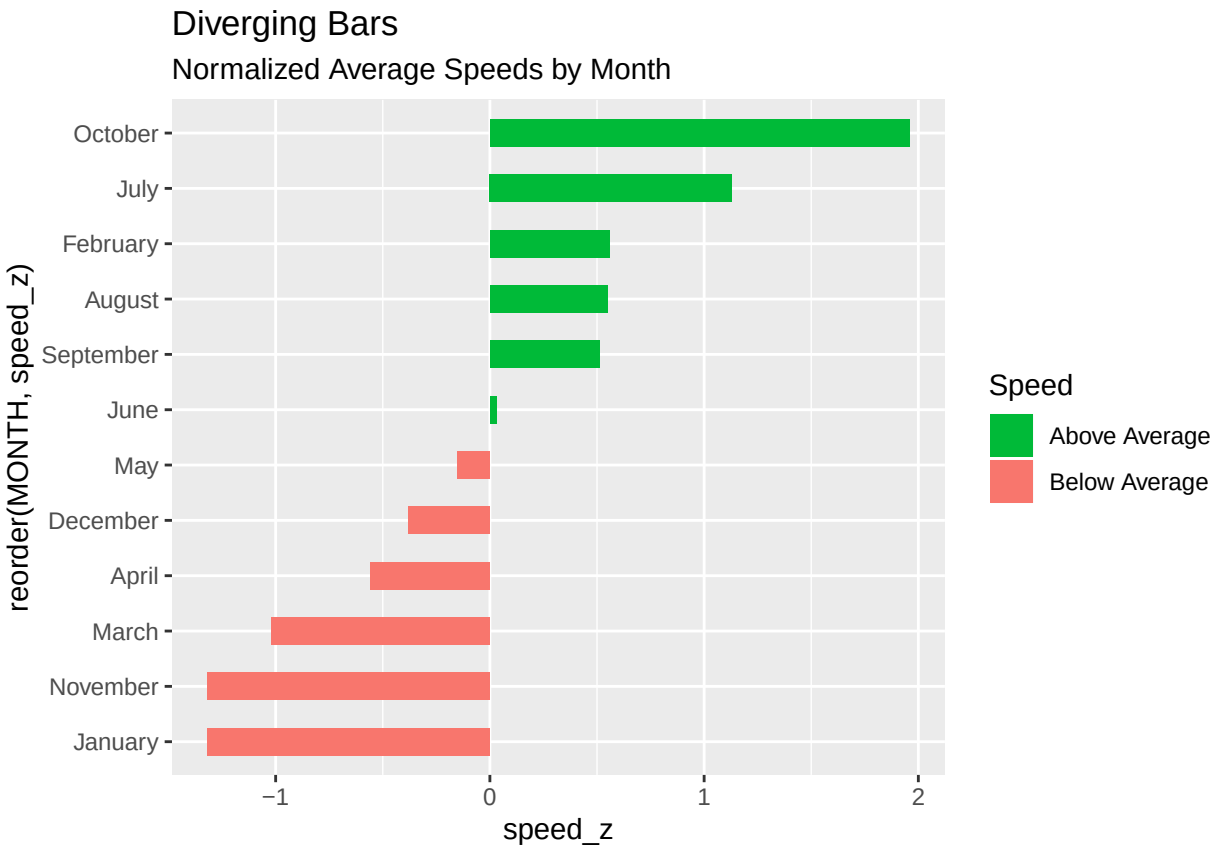


Refer to this link to have a similar following plot:

15.

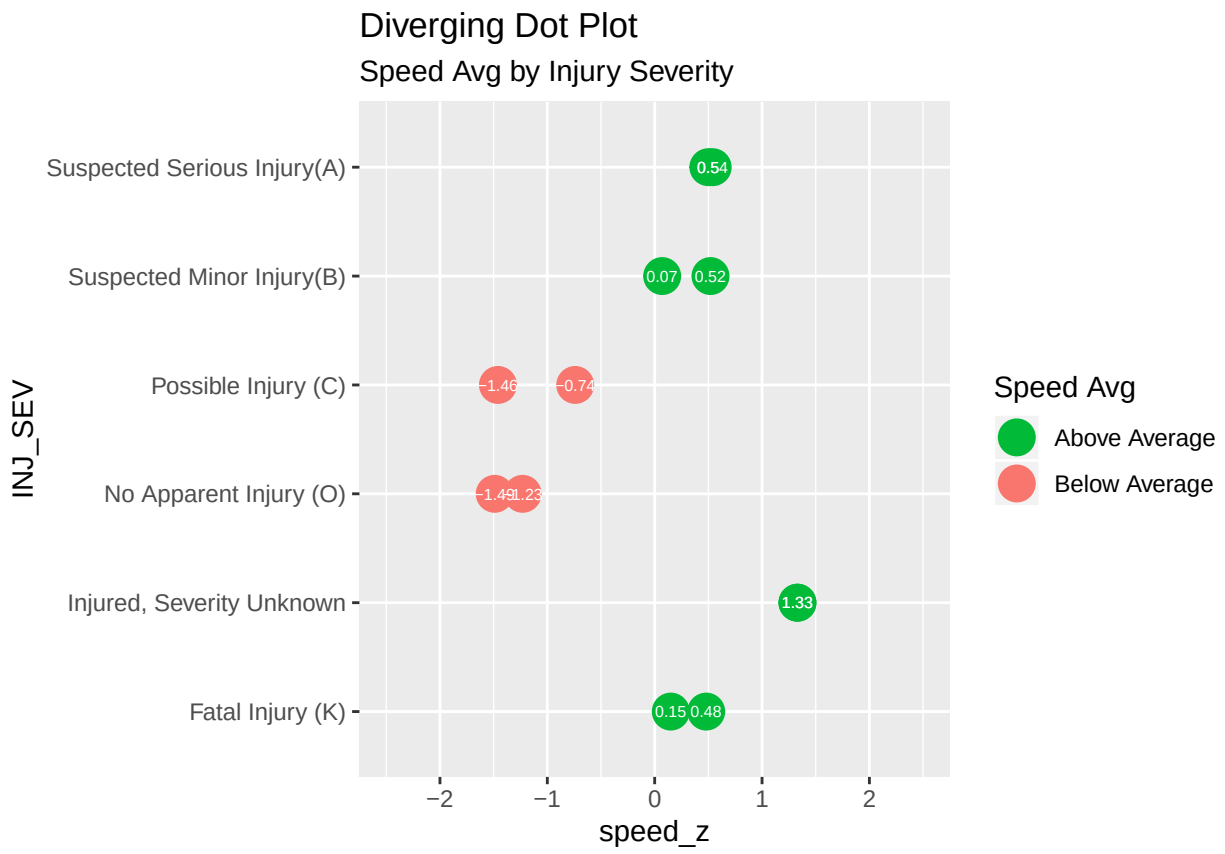
Horizontal axis is for (monthly) average speed The vertical axis is for months Color by two colors: one for above overall average speed and the other for below the avarage speed The speed on the horizontal axis is standardized

```
library(ggplot2)

d1<- d %>%
  group_by(MONTH) %>%
  summarize(speed_avg = mean(TRAV_SP, na.rm=TRUE))
d1$speed_z= round((d1$speed_avg - mean(d1$speed_avg))/sd(d1$speed_avg),2)
d1$speed_type = ifelse(d1$speed_z<0, "below","above")
ggplot(d1,aes(x=reorder(MONTH, speed_z), y=speed_z, label=speed_z)) +
  geom_bar(stat='identity', aes(fill=speed_type), width=.5) + scale_fill_manual(name="Speed",
           labels = c("Above Average", "Below Average"),
           values= c("above"="#00ba38", "below"="#f8766d")) + labs(subtitle= "Normalized Average Speeds
       title= "Diverging Bars") +
  coord_flip()
```



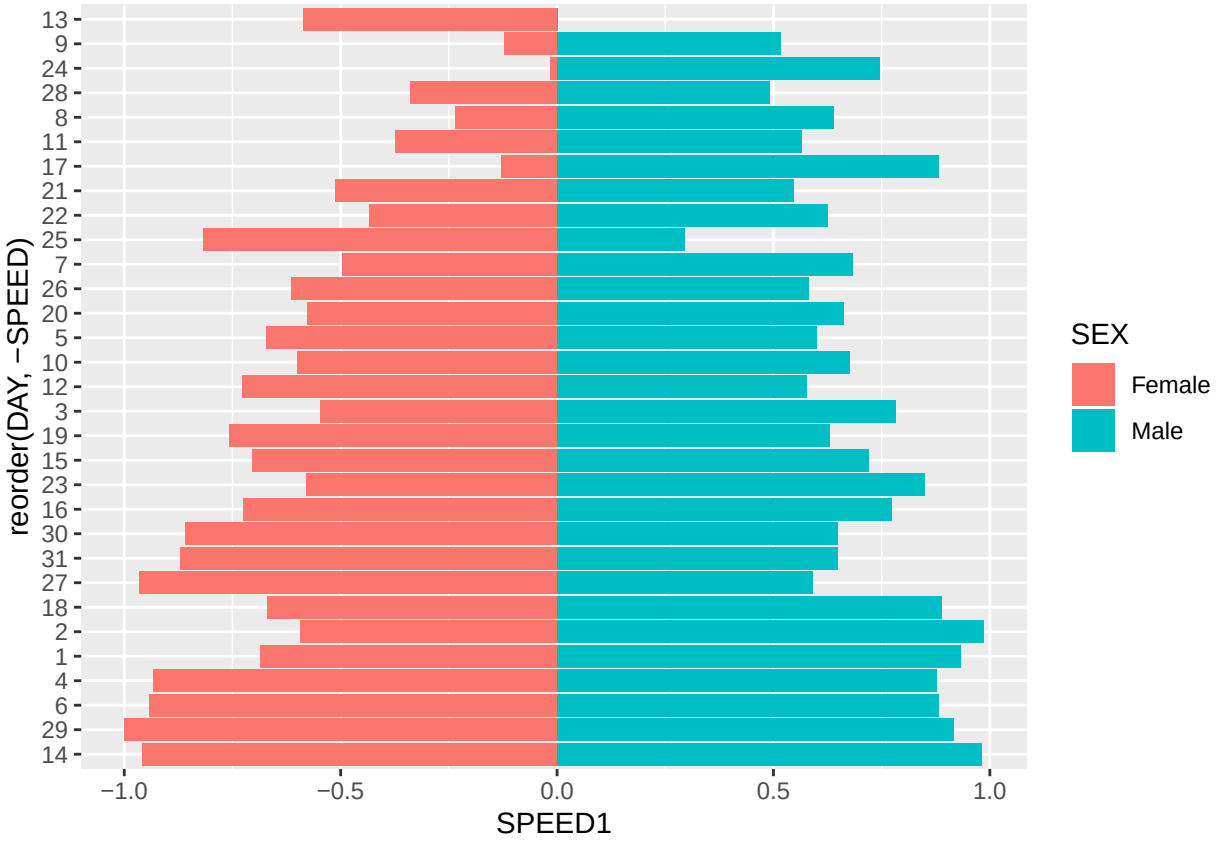
16. Refer to this link to have a similar following plot: Horizontal Axis is for mean speed Vertical Axis is for INJ_SEV Color by SEX The numbers of speed are shown in points.

```
d1<- d %>%
  group_by(INJ_SEV, SEX) %>%
  summarize(speed_avg = mean(TRAV_SP, na.rm=TRUE))
d1$speed_z= round((d1$speed_avg - mean(d1$speed_avg))/sd(d1$speed_avg),2)
d1$speed_type = ifelse(d1$speed_z<0, "below","above")
ggplot(d1, aes(x=INJ_SEV, y=speed_z, label=speed_z, color=SEX))+
  geom_point(stat="identity", aes(col=speed_type), size=6) +
  scale_color_manual(name="Speed Avg",
          labels = c("Above Average", "Below Average"),
          values= c("above"="#00ba38", "below"="#f8766d"))+ geom_text(color="white", size=2) +
  labs(title="Diverging Dot Plot", subtitle= "Speed Avg by Injury Severity") + ylim(-2.5,2.5) + coord_fl
```



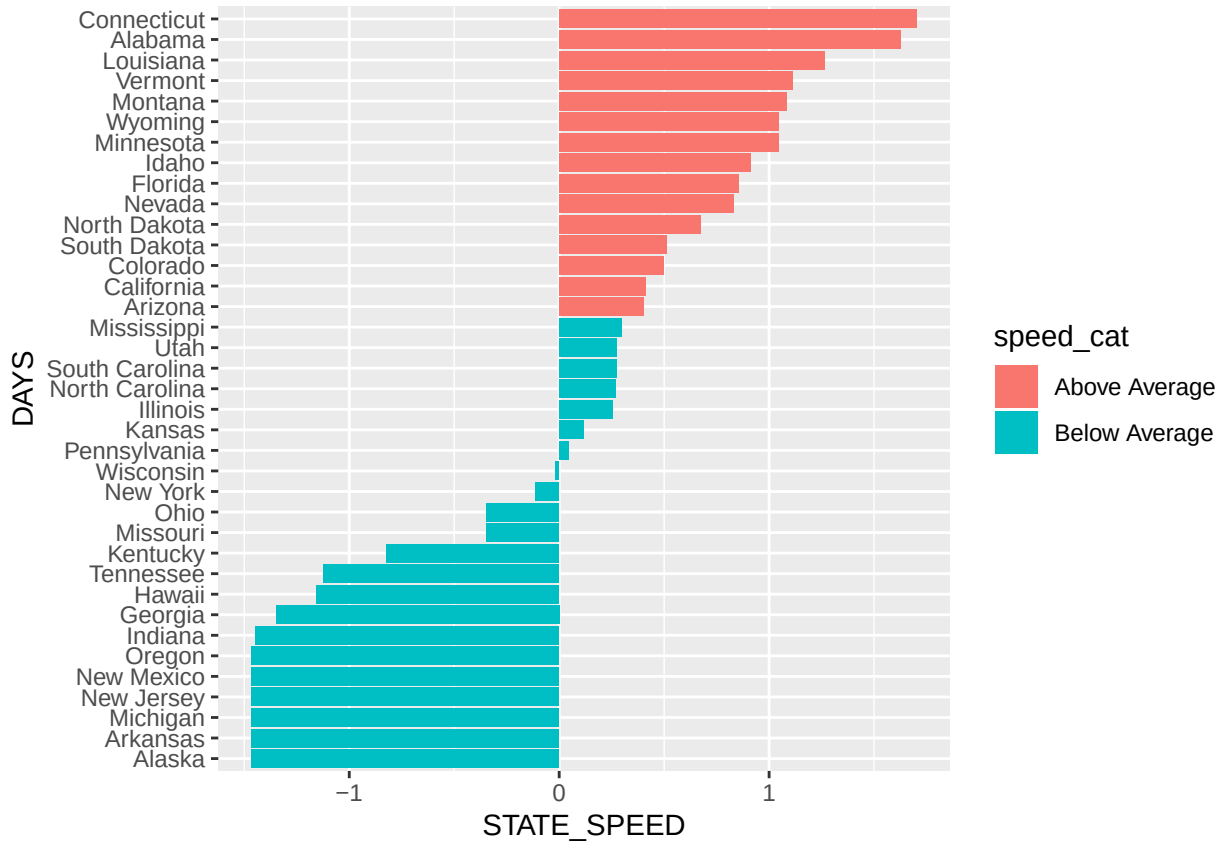
17. Refer to this link to have a similar following plot: Horizontal Axis is for speed Vertical Axis is for DAY Color by SEX The should be a invisible vertical line seperating the two sexes.

```
df2= d %>% group_by(DAY,SEX) %>% summarize(SPEED=mean(TRAV_SP))
a=min(df2$SPEED)
b=max(df2$SPEED)

df2%>% mutate(SPEED1 = ifelse(SEX=='Female', (SPEED-a)/(a-b), (SPEED-a)/(b-a))) %>%
  ggplot(aes(x=reorder(DAY,-SPEED),y=SPEED1, fill=SEX)) + geom_col() + coord_flip()
```

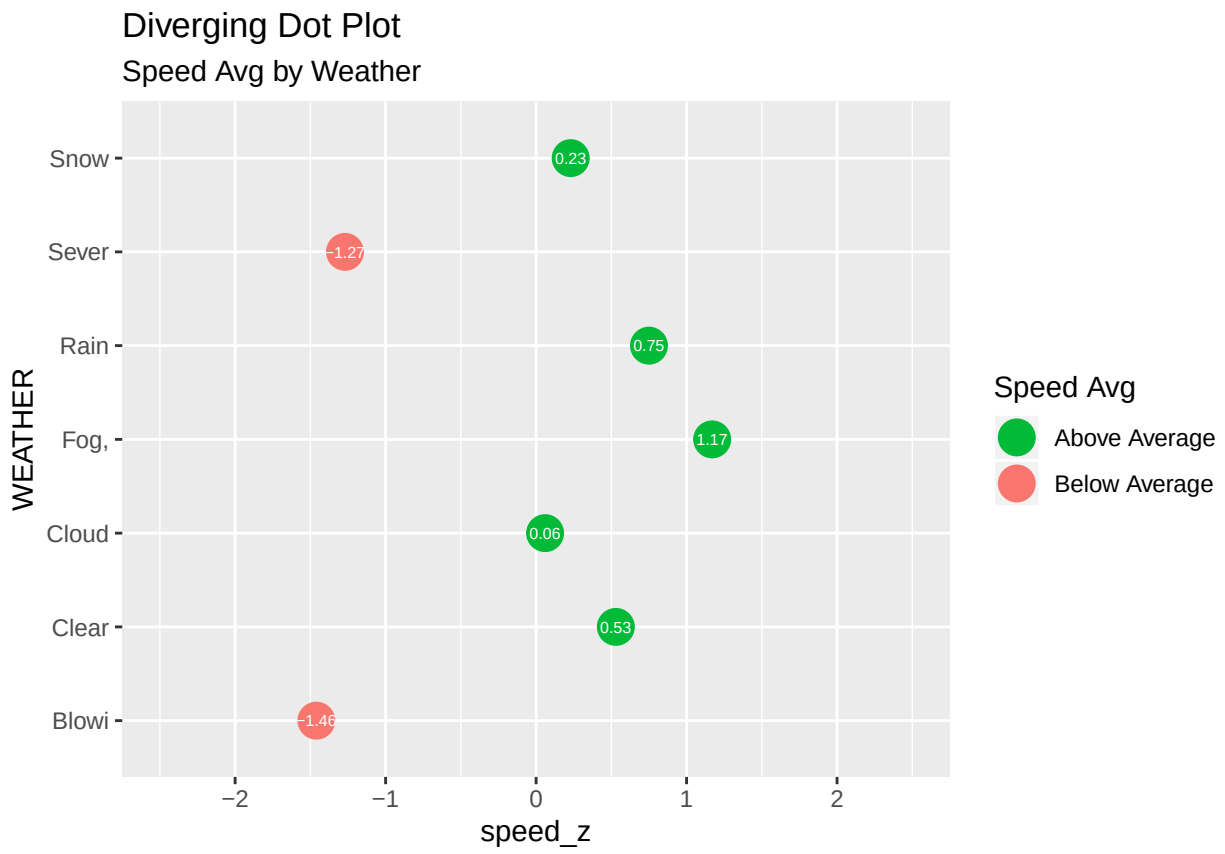


18-20. Generate three other interesting graphs from the dataset.

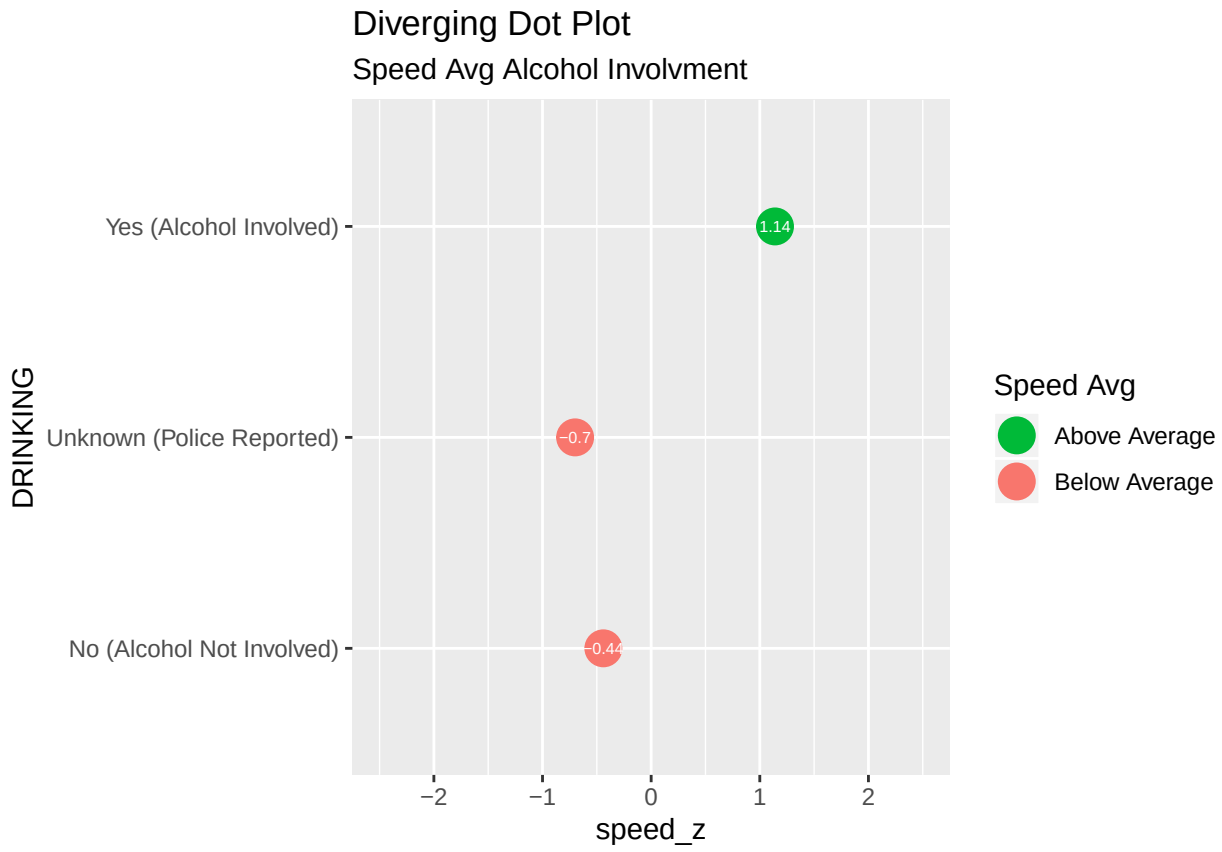
```
df2= d %>% group_by(STATE) %>% summarize(STATE_SPEED=mean(TRAV_SP)) %>% arrange(STATE_SPEED) %>%
  mutate(speed_cat=ifelse(STATE_SPEED>mean(d$TRAV_SP), 'Above Average', 'Below Average'))

df2 %>% mutate(STATE_SPEED = (STATE_SPEED - mean(STATE_SPEED))/sd(STATE_SPEED)) %>%
  ggplot(aes(x=reorder(STATE,STATE_SPEED), y=STATE_SPEED, fill=speed_cat))+
  geom_col() +
  labs(x='DAYS') + coord_flip()
```




```
d1<- d %>%
  group_by(WEATHER) %>%
  summarize(speed_avg = mean(TRAV_SP, na.rm=TRUE))
d1$speed_z= round((d1$speed_avg - mean(d1$speed_avg))/sd(d1$speed_avg),2)
d1$speed_type = ifelse(d1$speed_z<0, "below","above")
ggplot(d1, aes(x=WEATHER, y=speed_z, label=speed_z))+
  geom_point(stat="identity", aes(col=speed_type), size=6) +
  scale_color_manual(name="Speed Avg",
          labels = c("Above Average", "Below Average"),
          values= c("above"="#00ba38", "below"="#f8766d"))+ geom_text(color="white", size=2) +
  labs(title="Diverging Dot Plot", subtitle= "Speed Avg by Weather") + ylim(-2.5,2.5) + coord_flip()
```




```
d1<- d %>%
  group_by(DRINKING) %>%
  summarize(speed_avg = mean(TRAV_SP, na.rm=TRUE))
d1$speed_z= round((d1$speed_avg - mean(d1$speed_avg))/sd(d1$speed_avg),2)
d1$speed_type = ifelse(d1$speed_z<0, "below","above")
ggplot(d1, aes(x=DRINKING, y=speed_z, label=speed_z))+
  geom_point(stat="identity", aes(col=speed_type), size=6) +
  scale_color_manual(name="Speed Avg",
          labels = c("Above Average", "Below Average"),
          values= c("above"="#00ba38", "below"="#f8766d"))+ geom_text(color="white", size=2) +
  labs(title="Diverging Dot Plot", subtitle= "Speed Avg Alcohol Involvment") + ylim(-2.5,2.5) + coord_fl
```


Diverging Dot Plot
Speed Avg Alcohol Involvment
DRINKING
Yes (Alcohol Involved)
Unknown (Police Reported)
No (Alcohol Not Involved)
1.14
−0.7
−0.44
−2
−1
0
1
2
speed_z
Speed Avg
Above Average
Below Average