EDA (lab-10 / Part-1)

Bibek Sapkota

Exploratory data analysis

Importing csv files

```
MS_county_stops <- read.csv('MS_county_stops.csv')
MS_traffic_stops <- read.csv('MS_trafficstops_bw_age.csv')
```

Summarizing the data

```
str(MS_county_stops)
```

```
## 'data.frame':    82 obs. of  3 variables:
##  $ country_name: chr  "Adams County" "Alcorn County" "Amite County" "Attala County" ...
##  $ female      : num  36.7 33.3 38.3 36.7 32.1 ...
##  $ male        : num  38.4 34.1 40.3 38.1 34.4 ...
```

```
summary(MS_county_stops)
```

```
##  country_name           female           male      
##  Length:82          Min.   :29.55   Min.   :30.57  
##  Class :character   1st Qu.:33.16   1st Qu.:34.55  
##  Mode  :character   Median :34.34   Median :35.59  
##                     Mean   :34.29   Mean   :35.78  
##                     3rd Qu.:35.55   3rd Qu.:37.15  
##                     Max.   :38.30   Max.   :41.23
```

Plotting with ggplot2

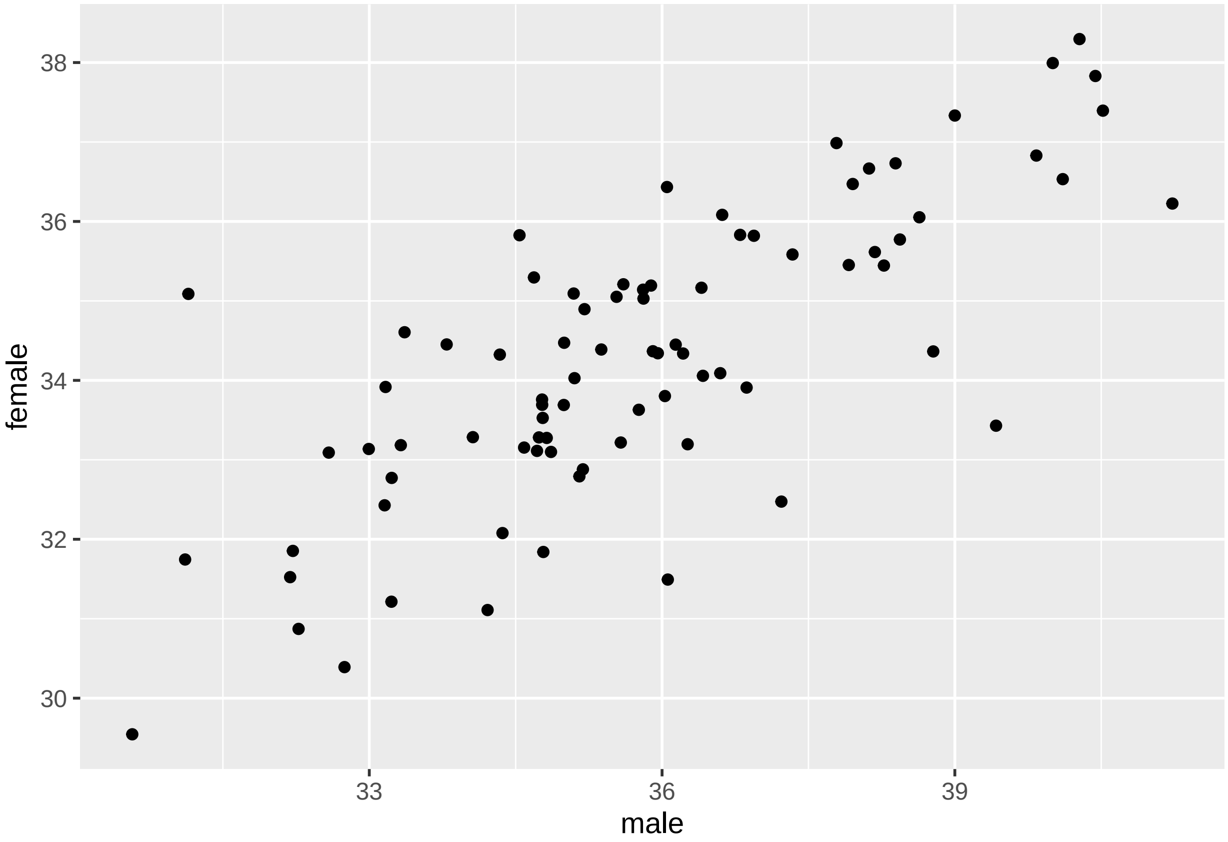
task 1: importing the library

```
library(ggplot2)
library(tidyverse)
```

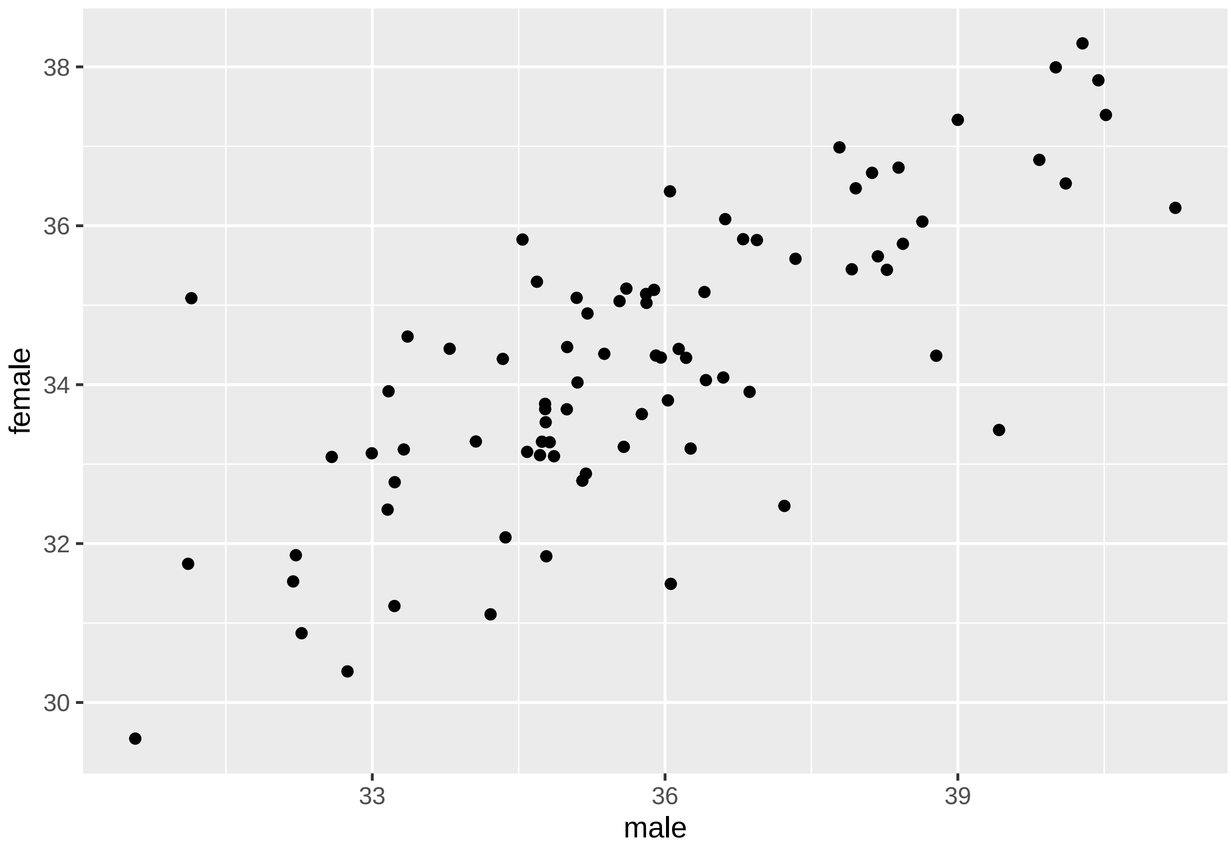
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

task 2: Ploting the data of ms_country_stops dataset.

```
ggplot(data = MS_county_stops, aes(x = male, y = female)) + geom_point()
```

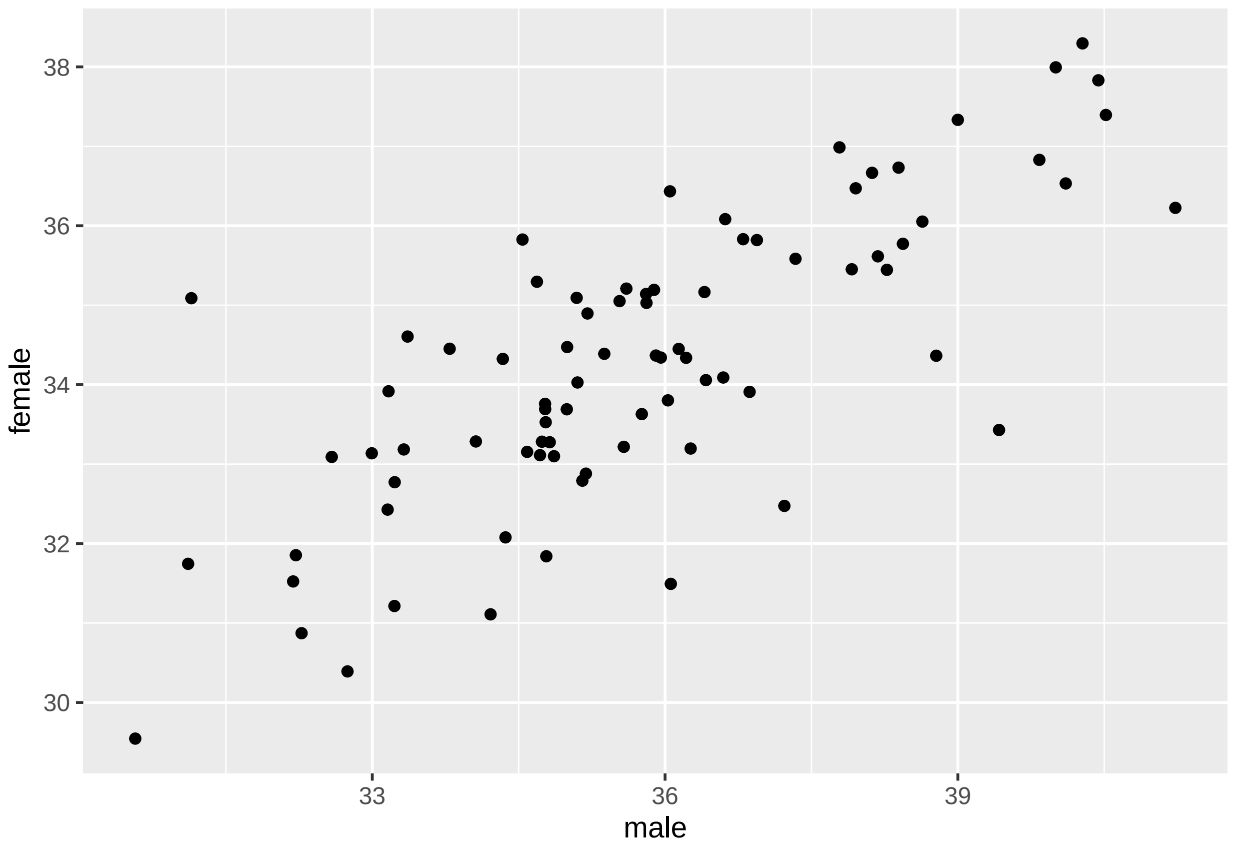


```
MS_county_stops %>% ggplot(aes(x = male, y = female)) + geom_point()
```

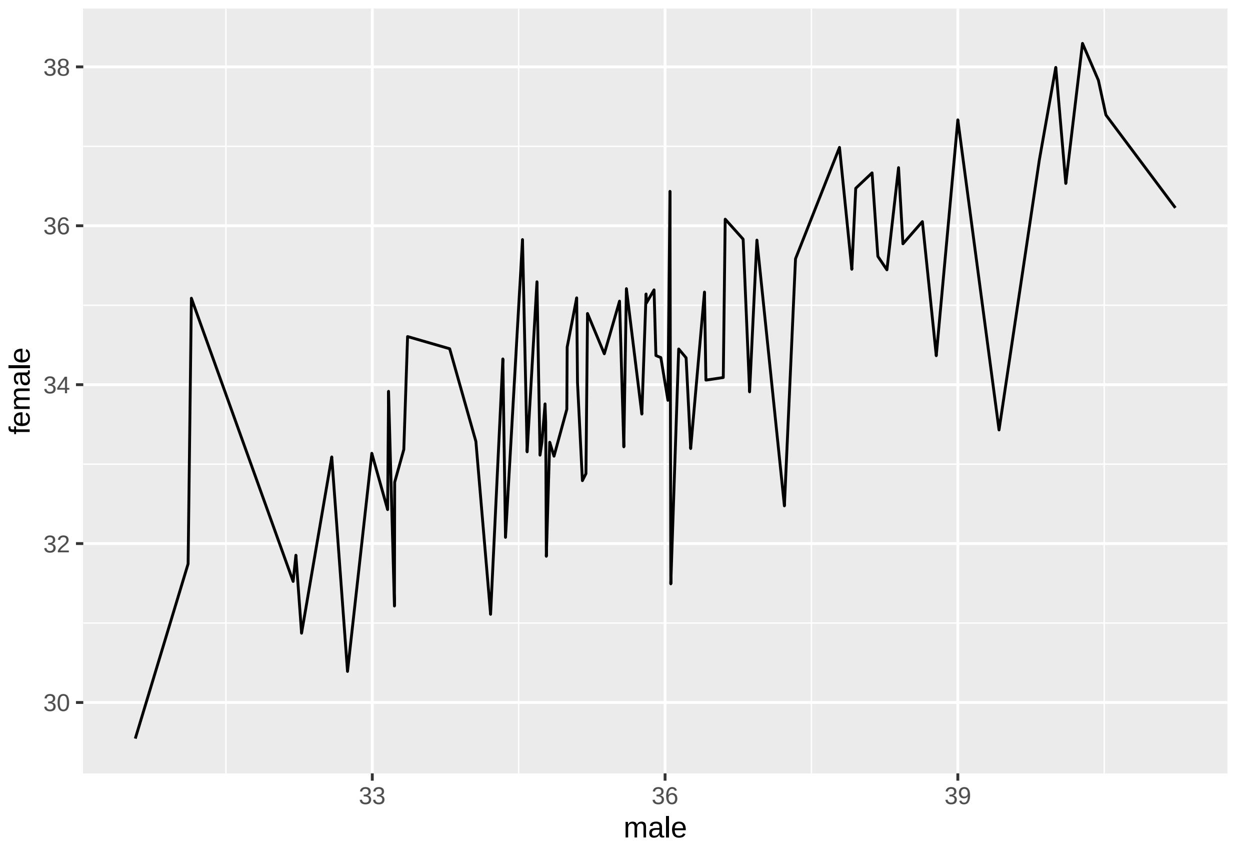


task 3: Assign plot to a variable and drawing data with datapoints and ploting it using lines

```
MS_plot <- ggplot(data = MS_county_stops, aes(x = male, y = female))
MS_plot + geom_point()
```



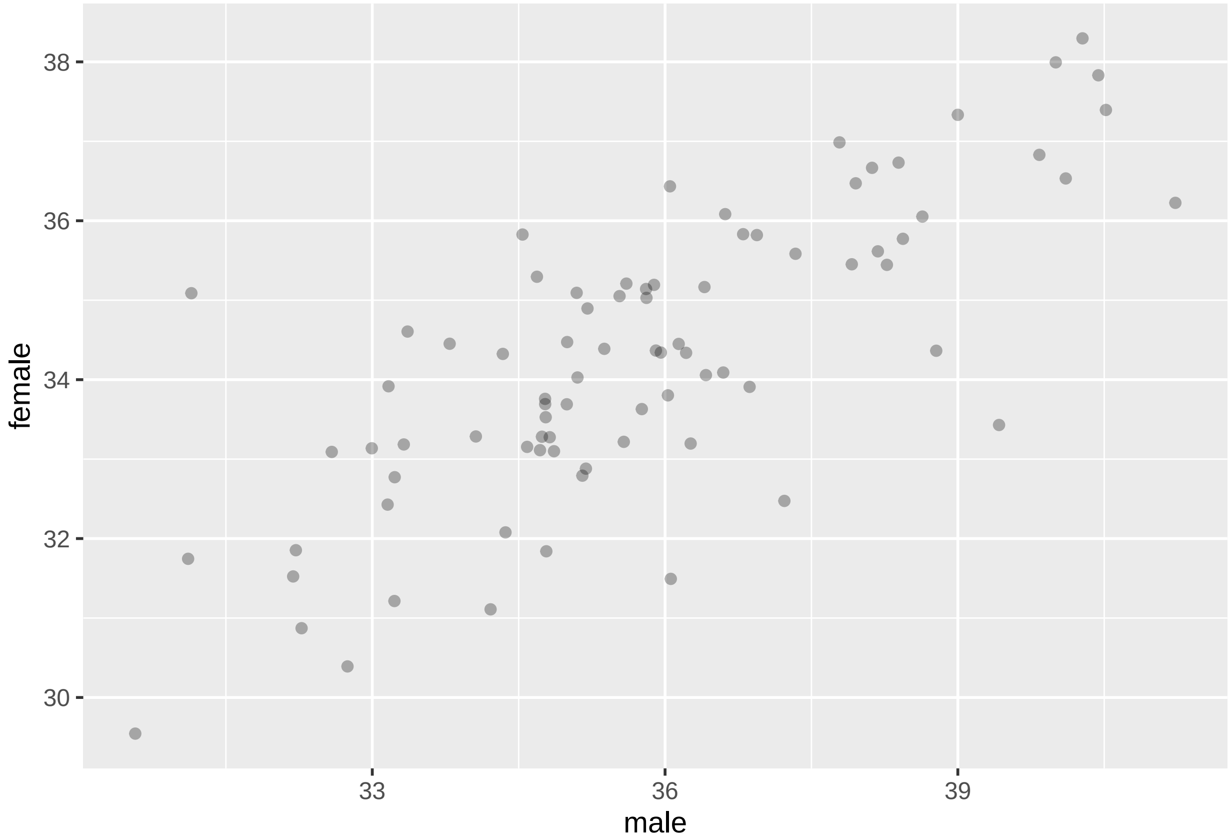
```
MS_plot + geom_line()
```



Scatter plot

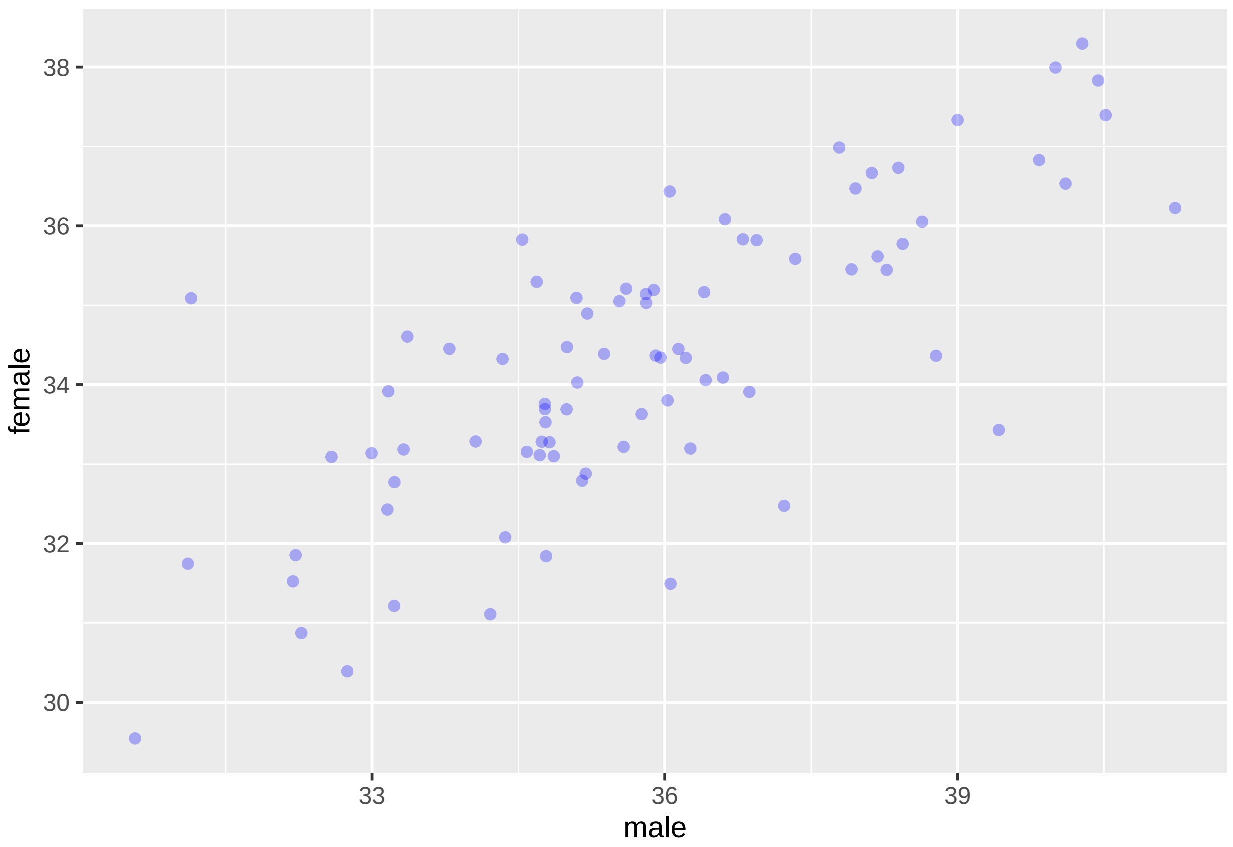
task 1:

```
ggplot(data = MS_county_stops, aes(x = male, y = female)) + geom_point(alpha= 0.3)
```



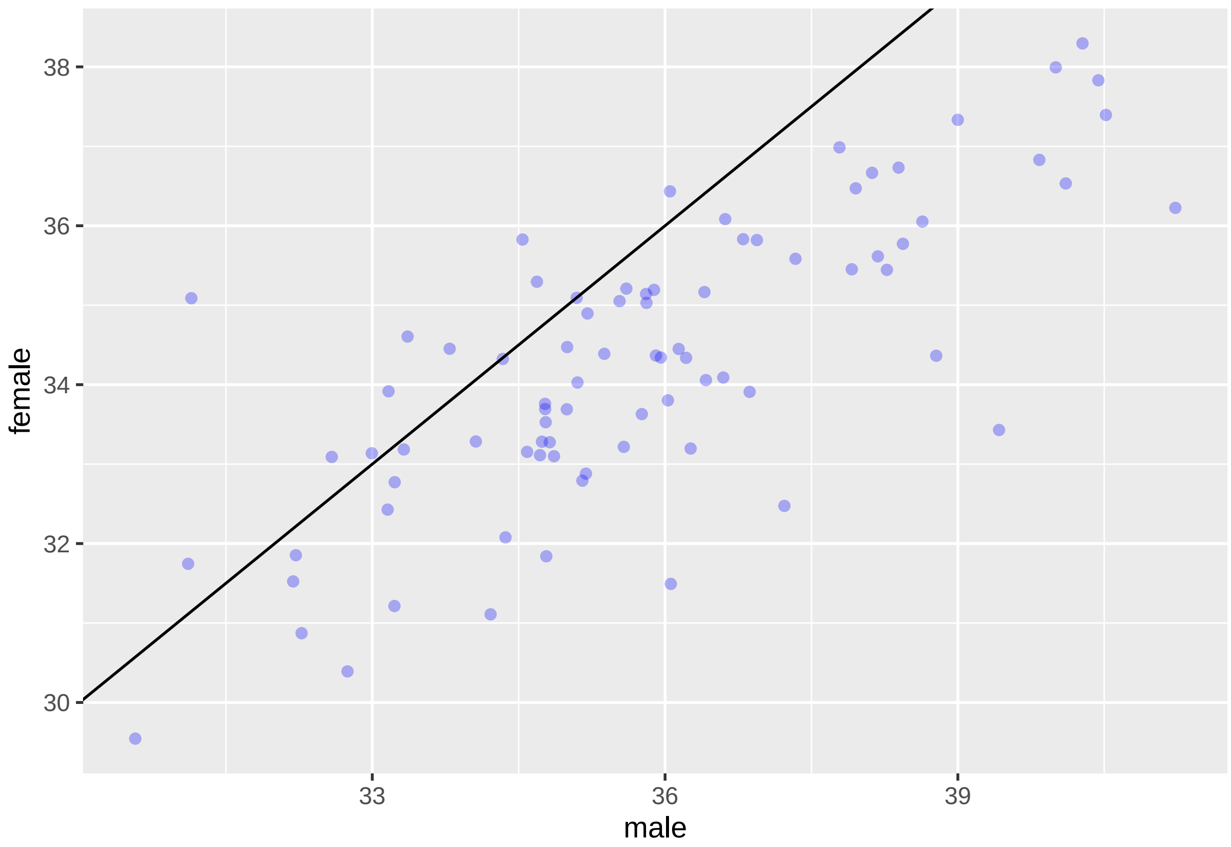
task 2: Adding blue color to the plot

```
ggplot(data = MS_county_stops, aes(x = male, y = female)) + geom_point(alpha= 0.3, color= "blue")
```



task 3:Adding another line graph in data plot.

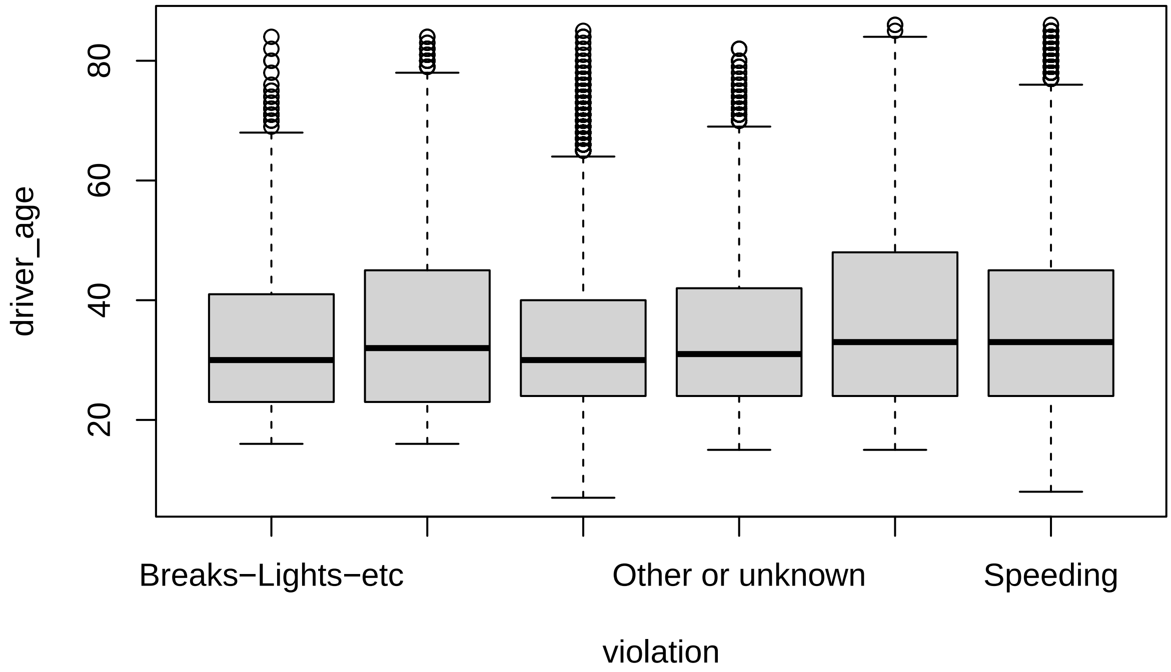
```
ggplot(data = MS_county_stops, aes(x = male, y = female)) + geom_point(alpha= 0.3, color= "blue")+ geom_
```



Boxplot

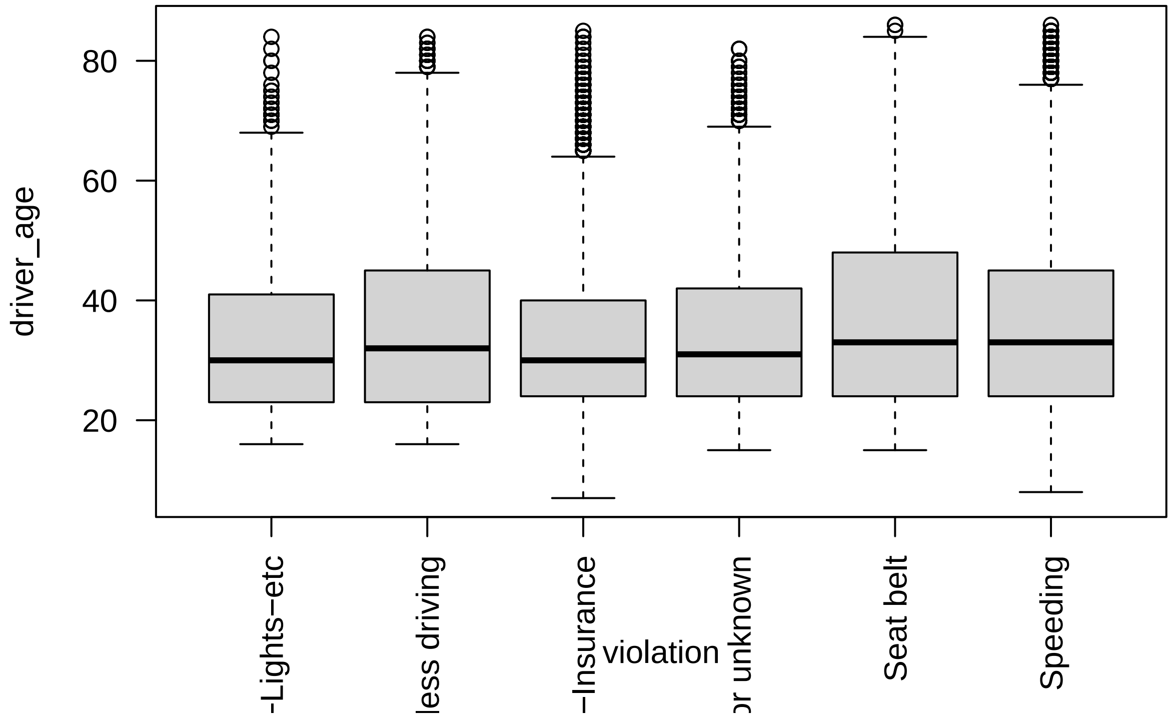
task 1:Ploting box plot of Ms_traffic_stops dataset

```
boxplot(driver_age~violation, data = MS_traffic_stops)
```



task 2:Rotating the x axis text

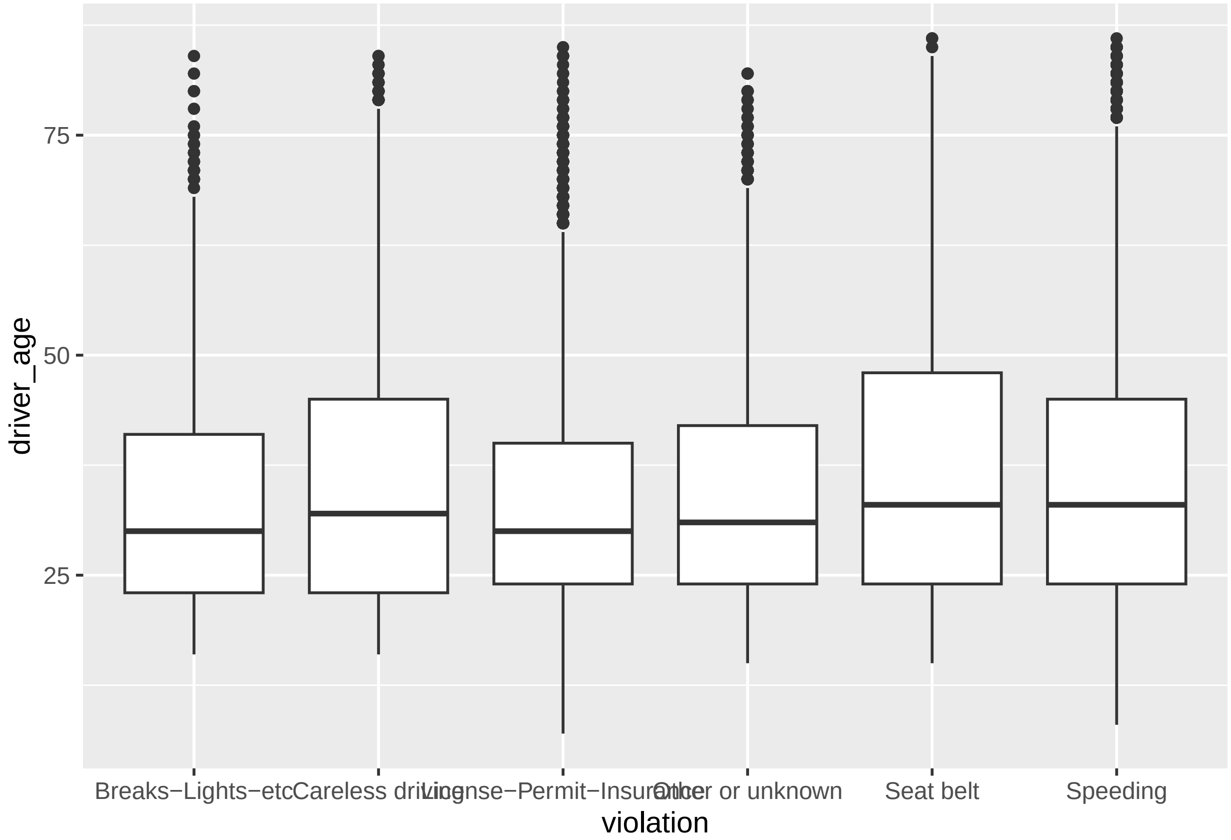
```r
boxplot(driver_age~violation, data = MS_traffic_stops, las = 2)
```



task 3:Drawing boxplots using the ggplot function.

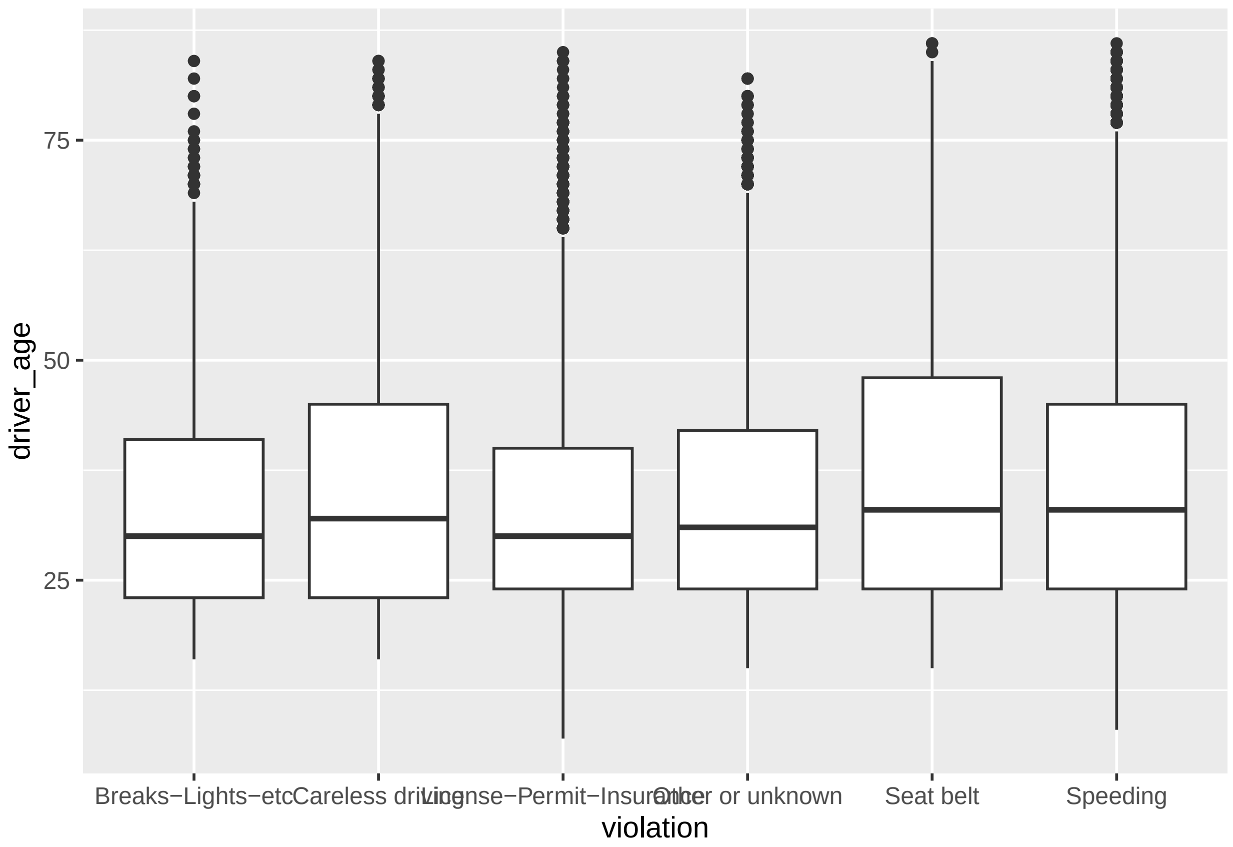
```
ggplot(MS_traffic_stops, aes(x = violation, y = driver_age)) + geom_boxplot()
```

```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



task 4:Filtering the missing values of driver age and Drawing boxplot

```
filtered_MS_traffic_stops <- MS_traffic_stops %>% filter(!is.na(driver_age))
ggplot(filtered_MS_traffic_stops, aes(x = violation, y = driver_age)) + geom_boxplot()
```

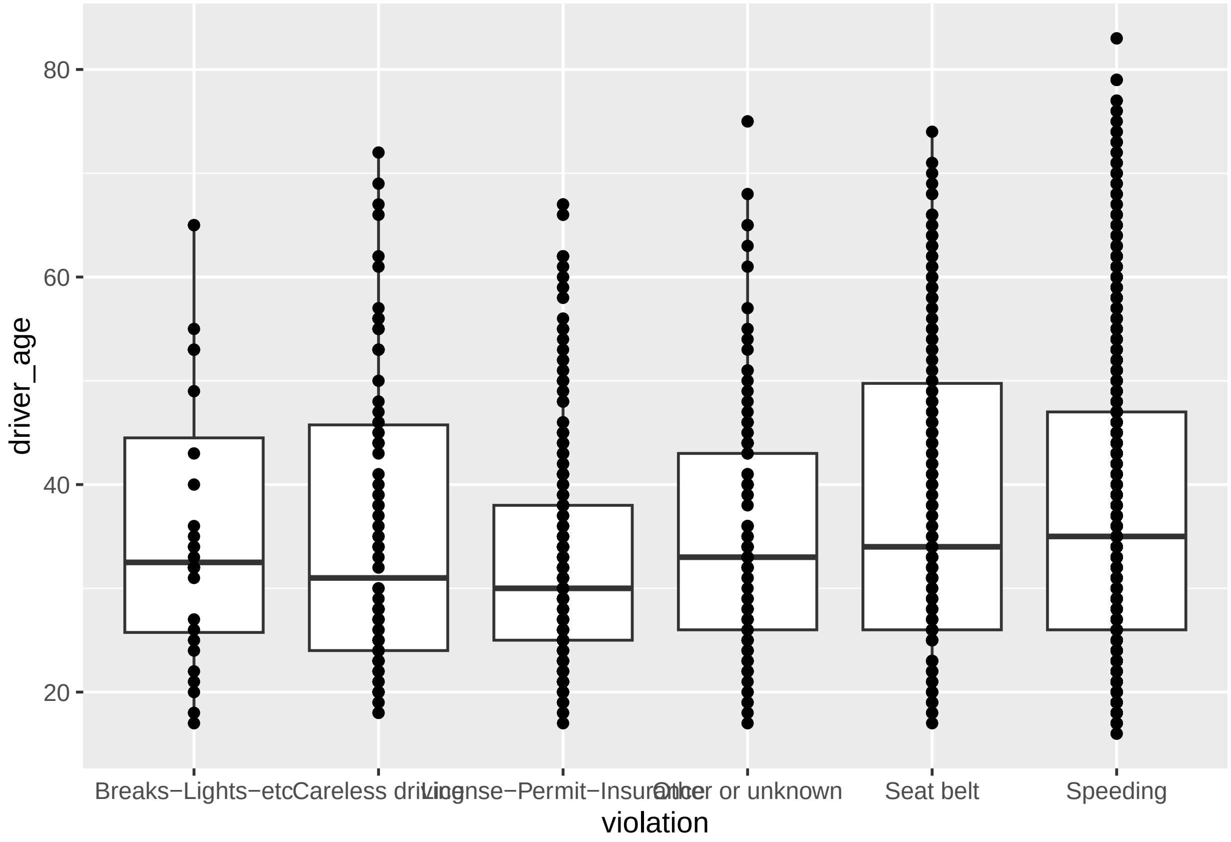


task 5:Filtering Yazoo country and removing its null values

```
Yazoo_stops <- MS_traffic_stops %>% filter(county_name == "Yazoo County", !is.na(driver_age))
```

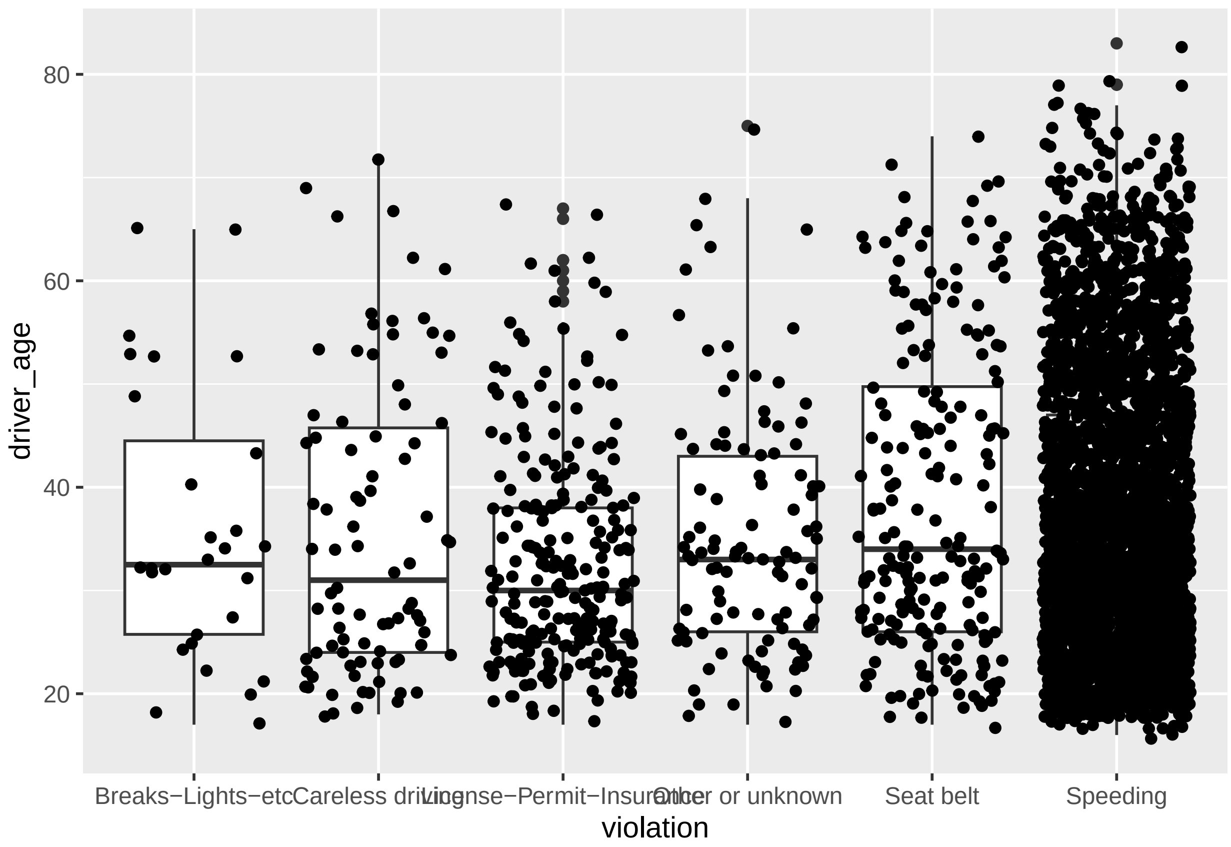
task 6:Using ggplot we can draw boxplots with data points on it.

```
ggplot(Yazoo_stops, aes(x = violation, y = driver_age)) + geom_boxplot() + geom_point()
```



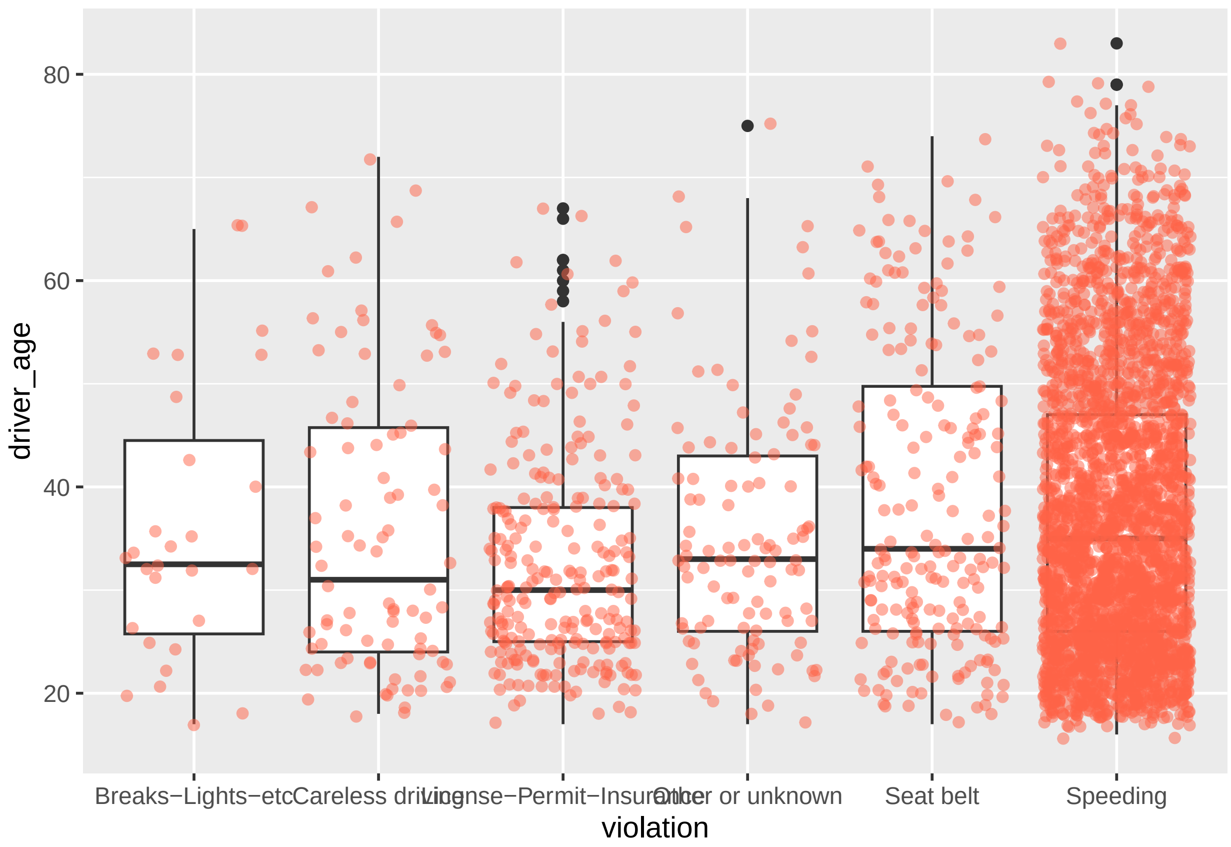
task 7:

```
ggplot(Yazoo_stops, aes(x = violation, y = driver_age)) + geom_boxplot() + geom_jitter()
```



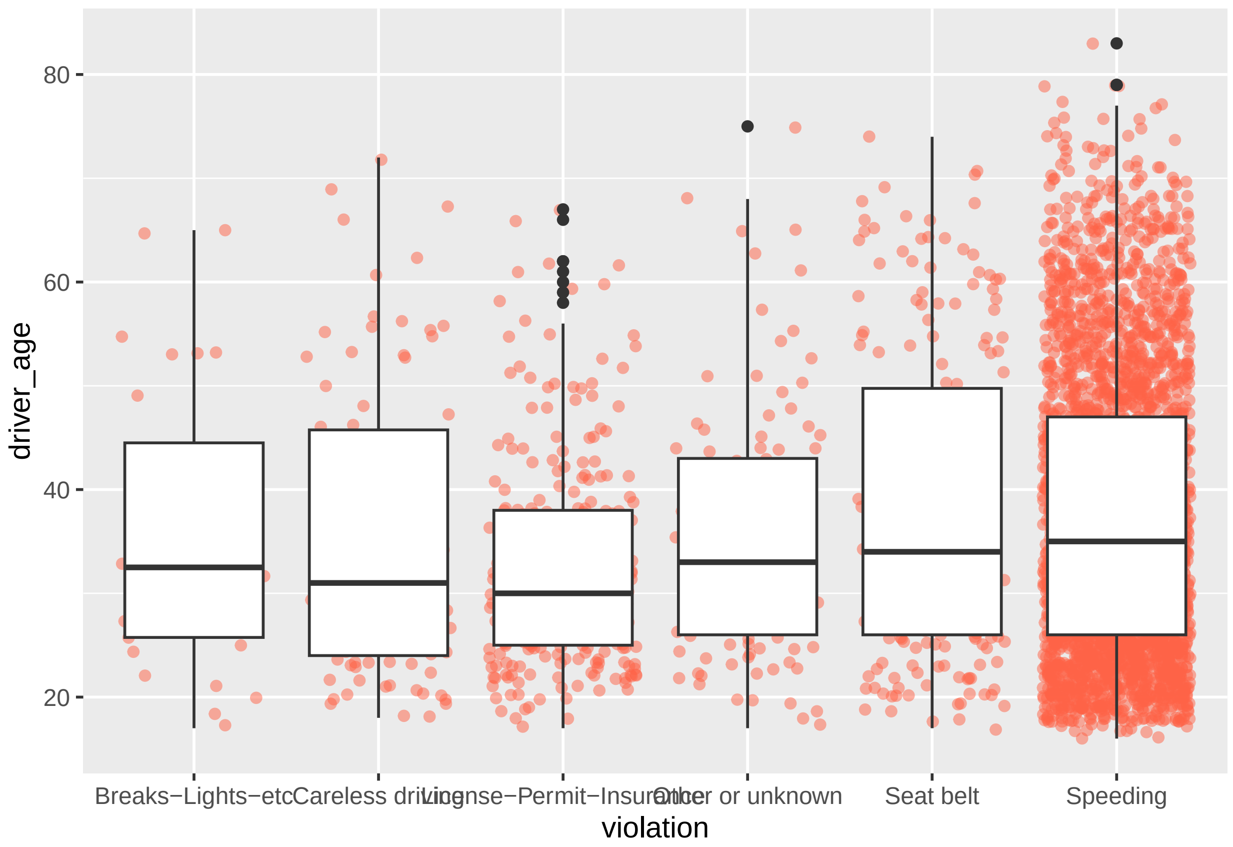
task 8:Coloring the noise

```
ggplot(Yazoo_stops, aes(x = violation, y = driver_age)) + geom_boxplot() + geom_jitter(alpha = 0.5, col
```



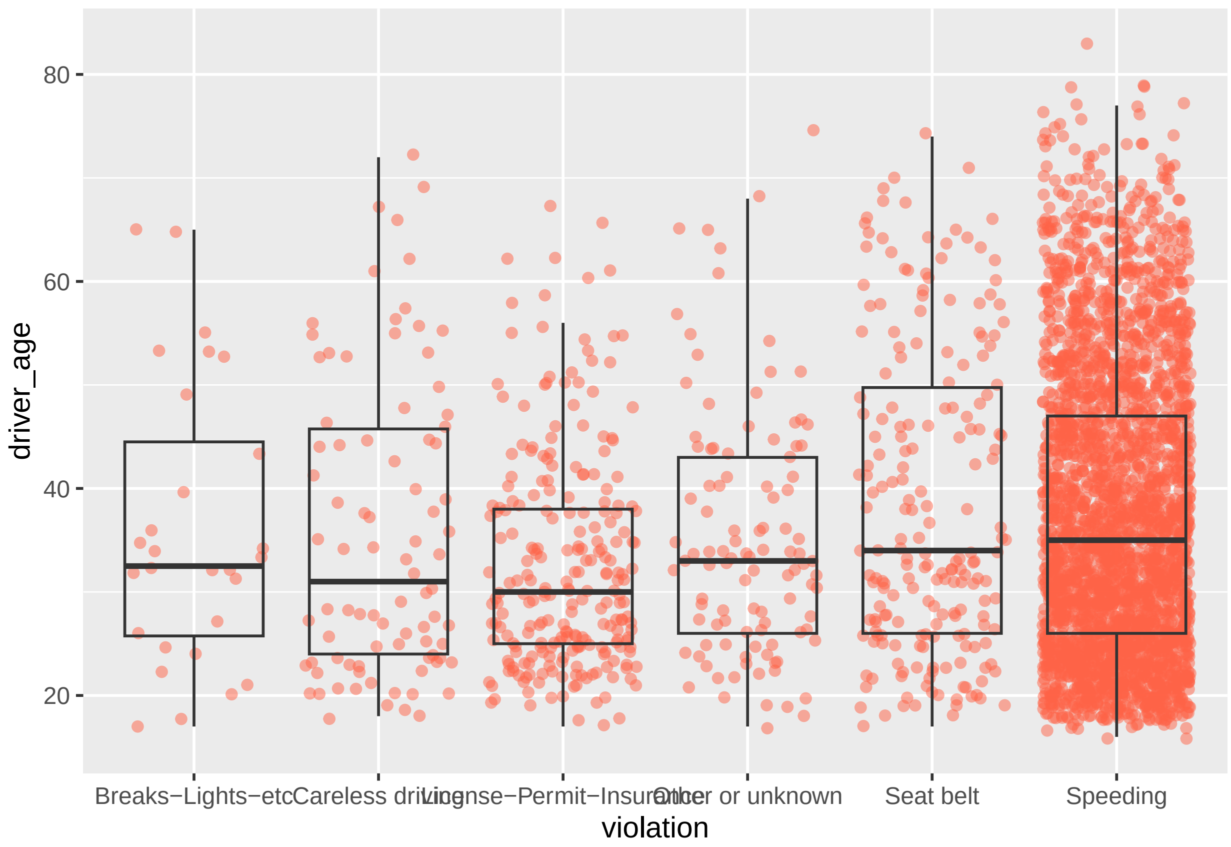
task 9:

```
ggplot(data = Yazoo_stops, aes(x = violation, y = driver_age)) + geom_jitter(alpha = 0.5, color = "toma
```



task 10:

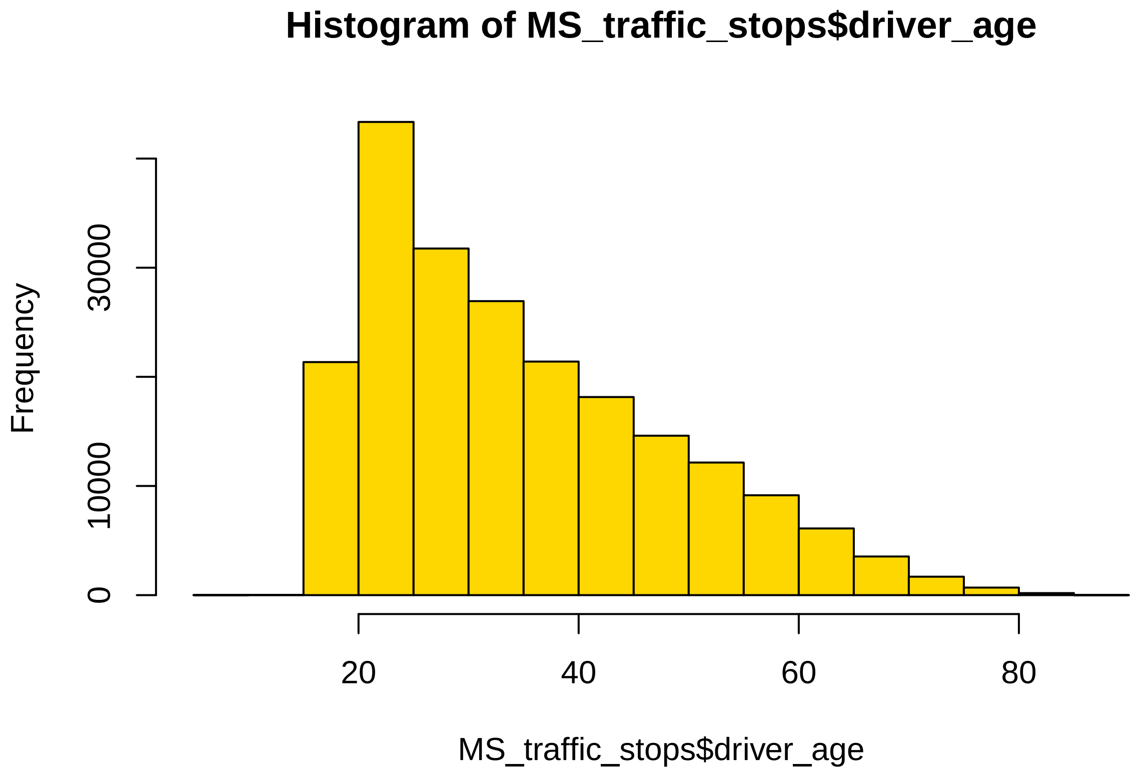
```
ggplot(data = Yazoo_stops, aes(x = violation, y = driver_age)) + geom_jitter(alpha = 0.5, color = "tomat
```



Histograms

task 1:Ploting a hist driagram of driver age

```
hist(MS_traffic_stops$driver_age,col="gold")
```

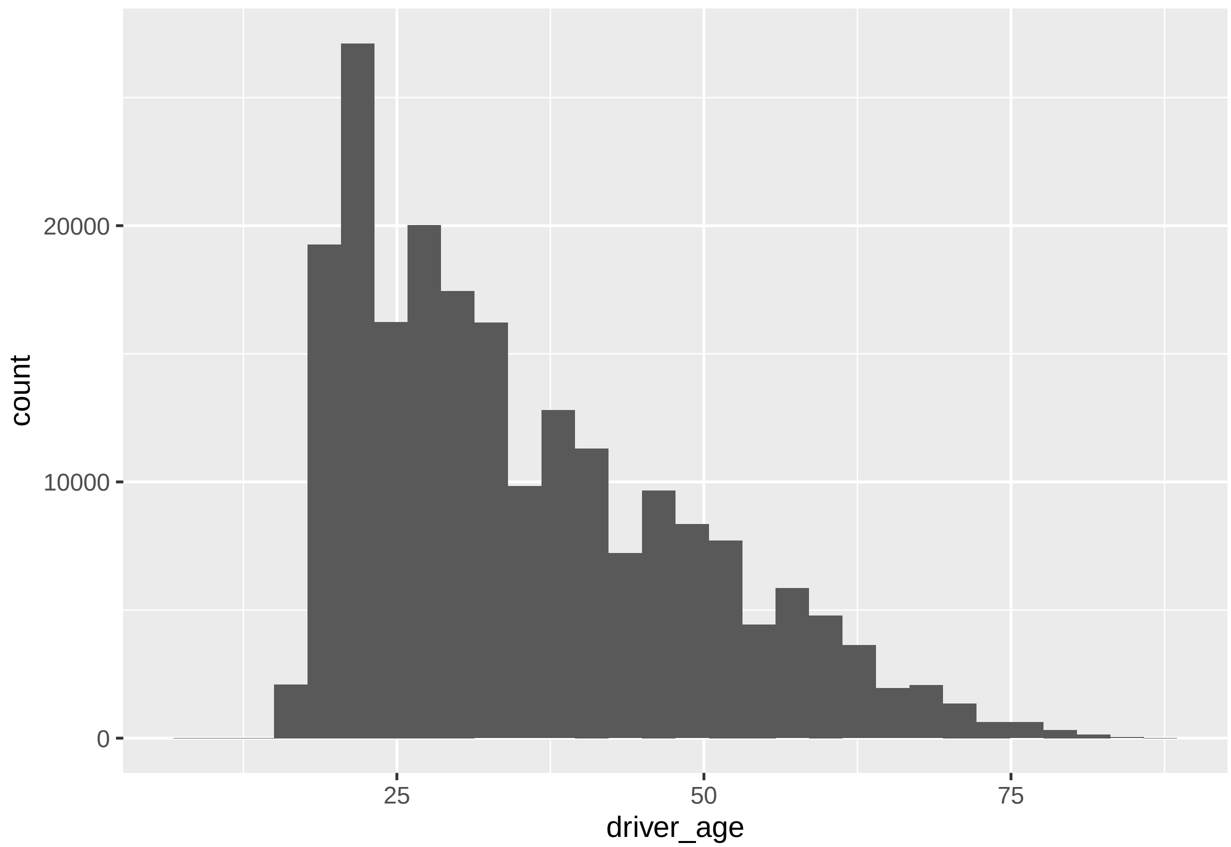


task 2:Drawing a histogram using the “ggplot” function.

```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram()
```

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

```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

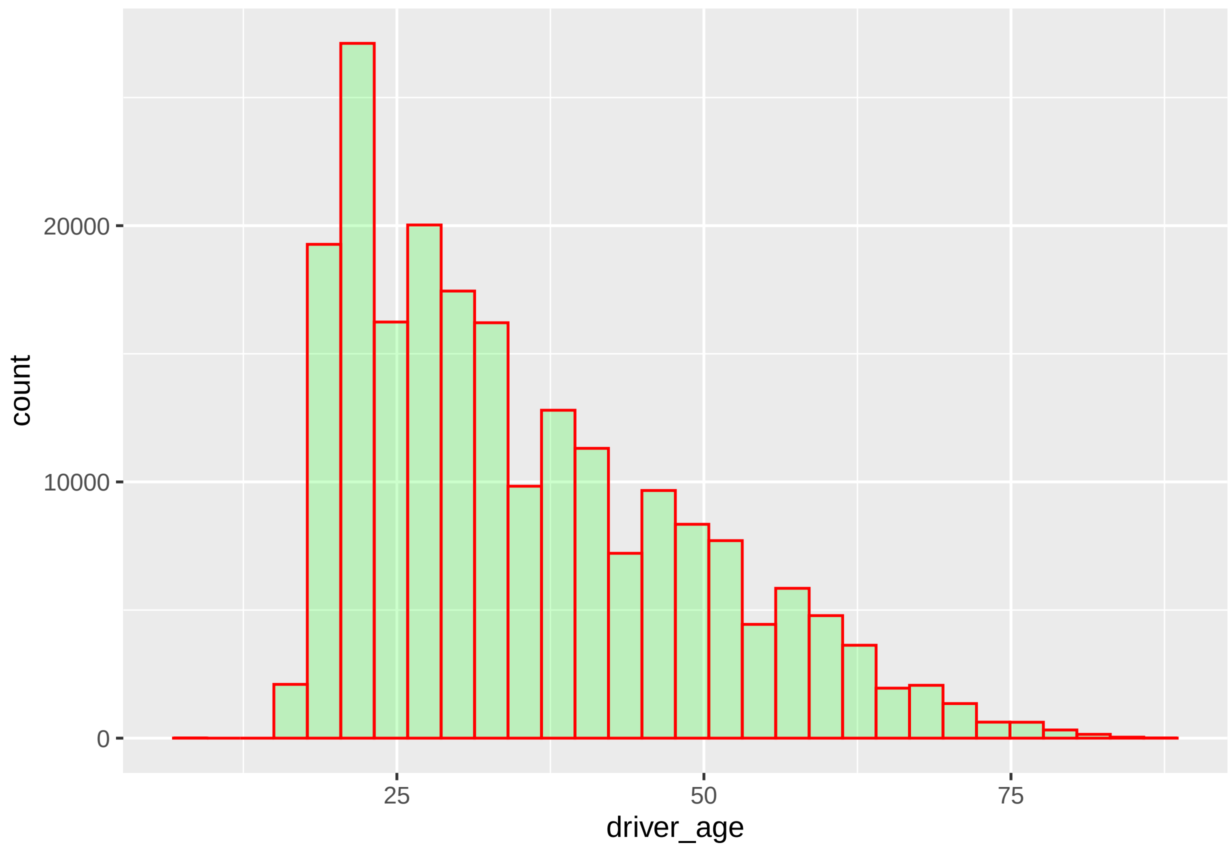


task 3:

```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram(col="red", fill="green", alpha=0.2)
```

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

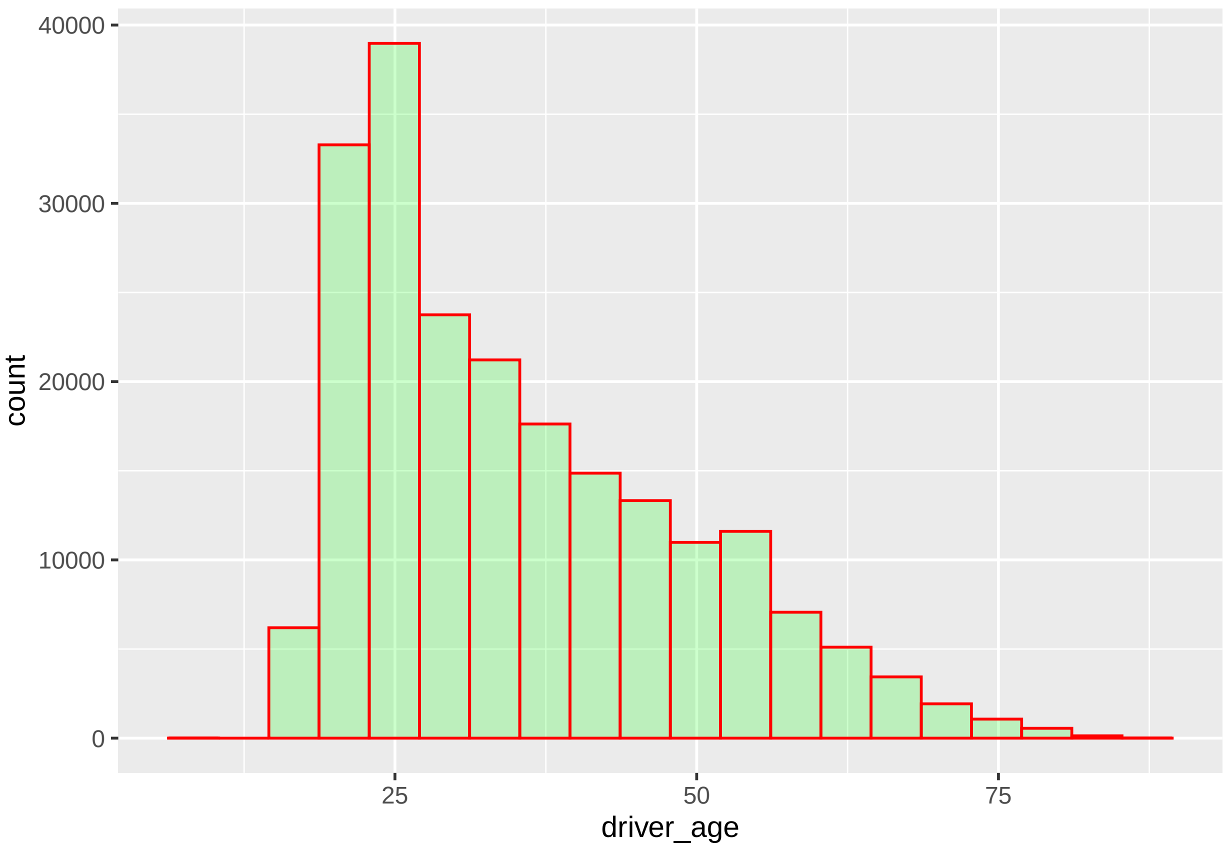
```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



task 4:Define bin count and Use break sequence with lower bound, upper bound and bin range

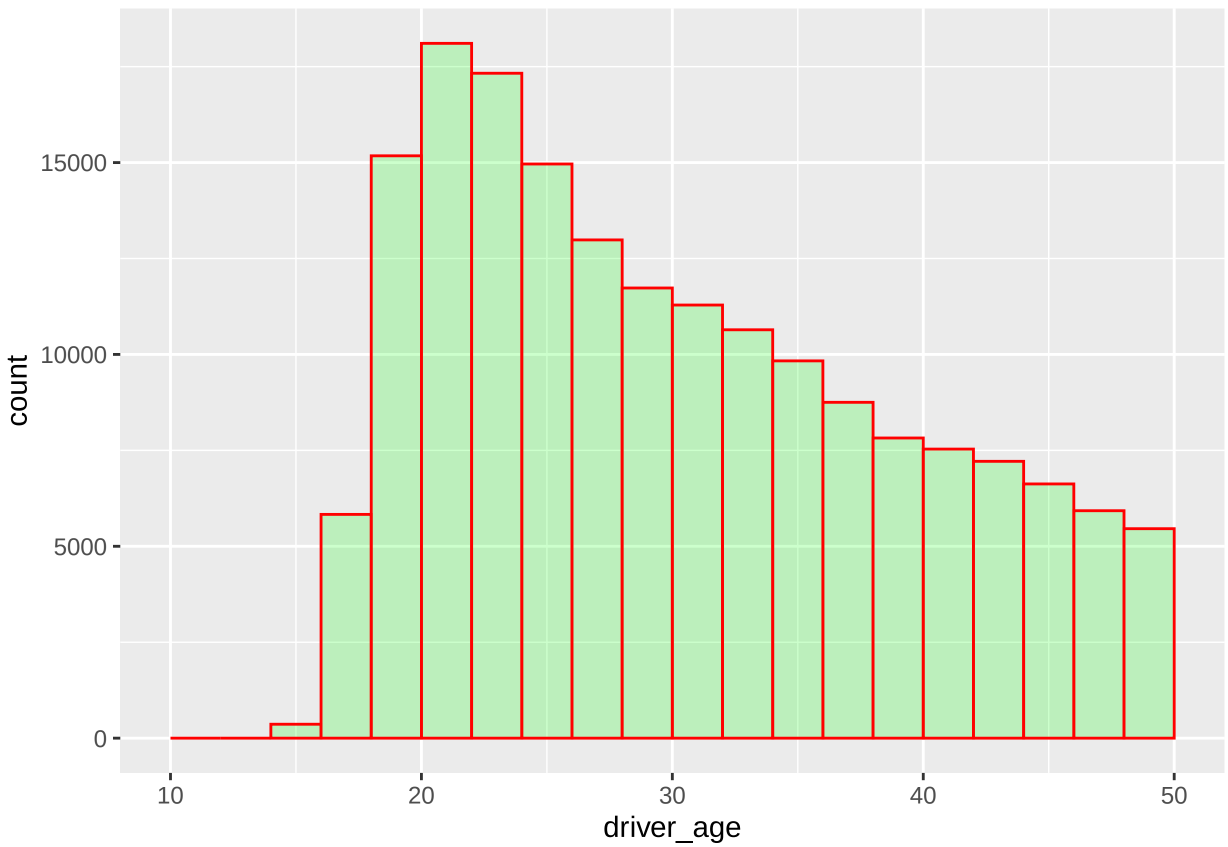
```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram(col="red", fill="green", alpha=0.2, bins=20)
```

```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram(col="red", fill="green", alpha=0.2, breaks=se
```

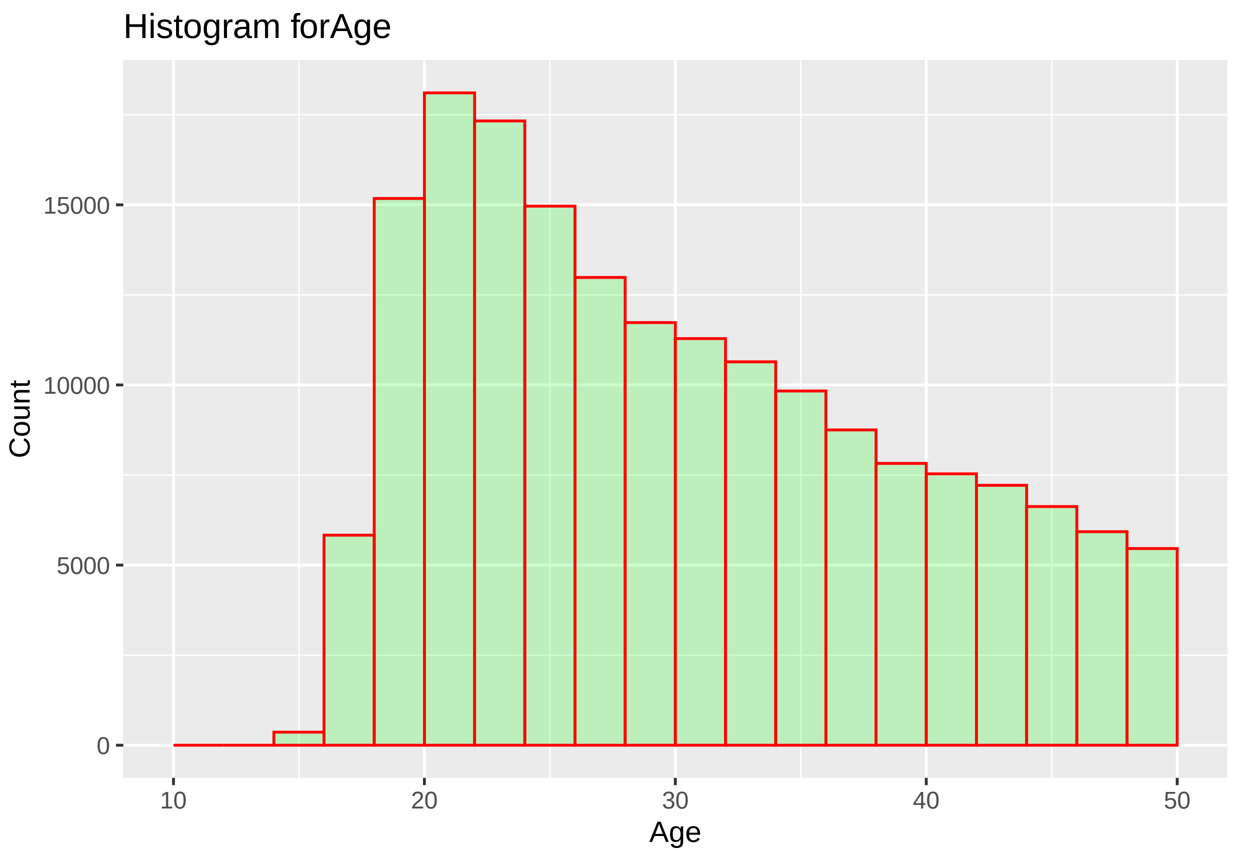
```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



task 5:Adding titles and x and y label

```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram(col="red", fill="green", alpha=0.2, breaks=s
```

```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```

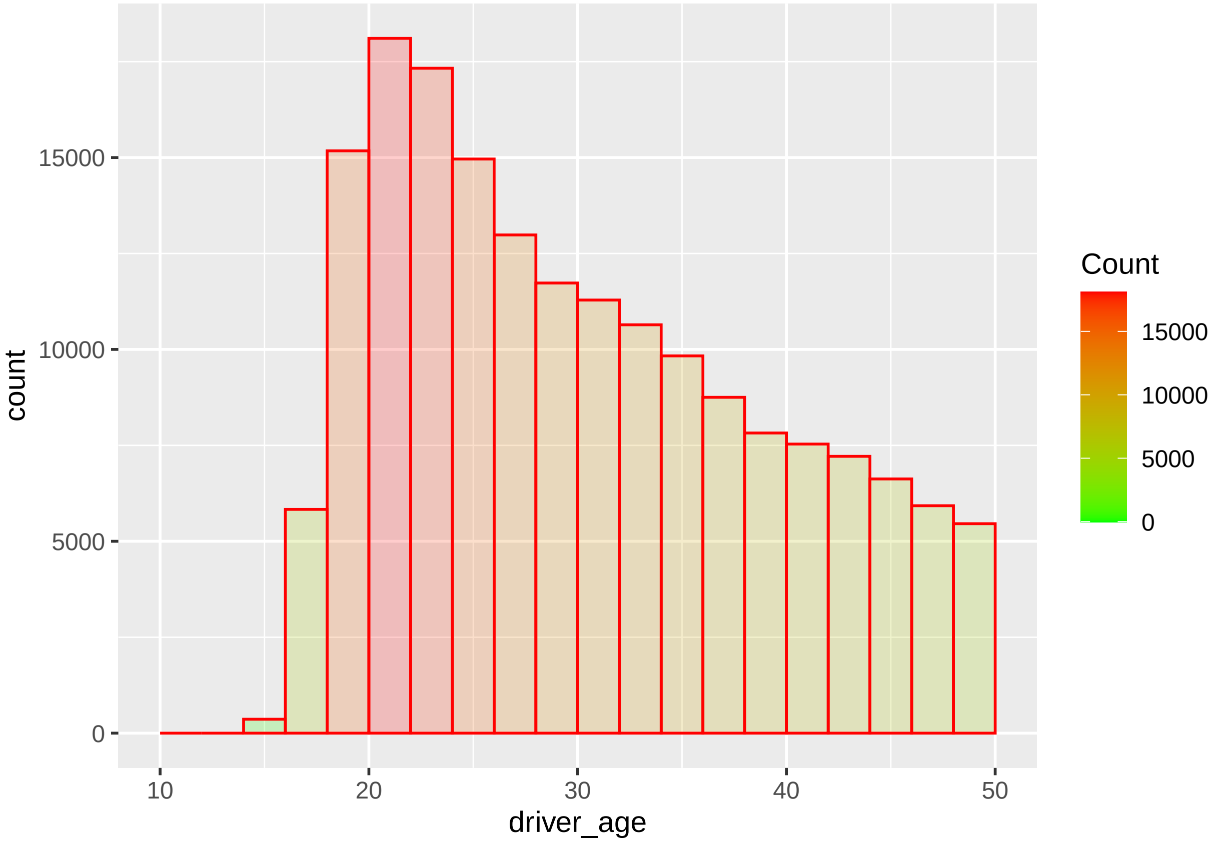


task 6: Following code segment count values from the y-axis low values should be in green and that the higher values should appear in red.

```
ggplot(MS_traffic_stops, aes(driver_age)) + geom_histogram(alpha=0.2, breaks=seq(10, 50, by=2), col="re
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## Warning: Removed 109 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



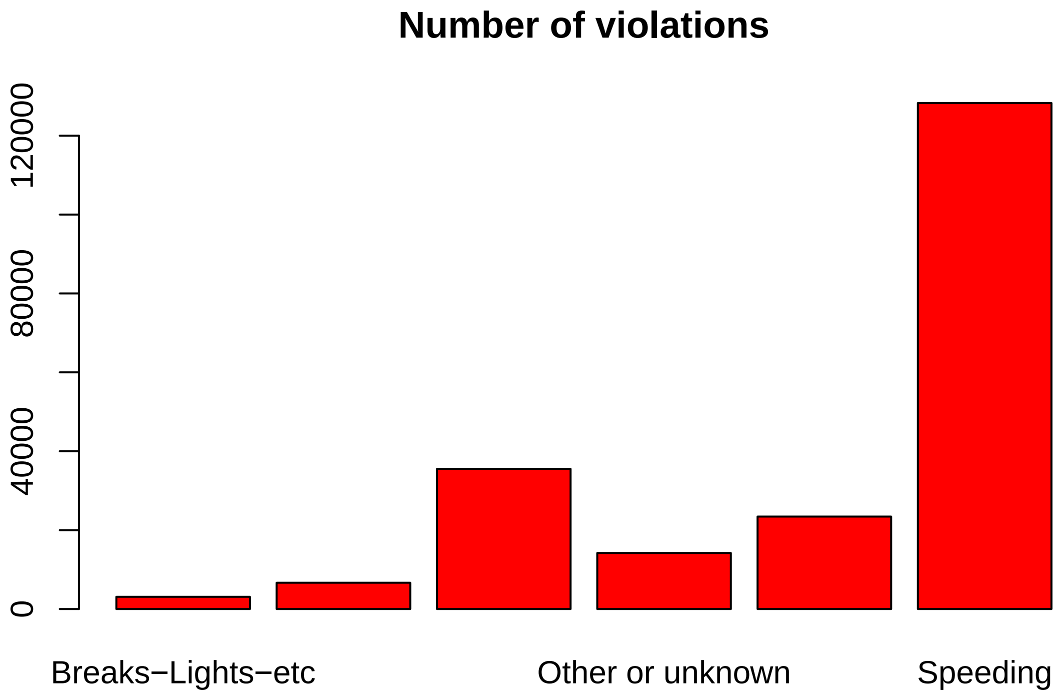
Bar charts / Bar plots

task 1:Display the frequency of each category and Draw bar chart

```
table(MS_traffic_stops$violation)
```

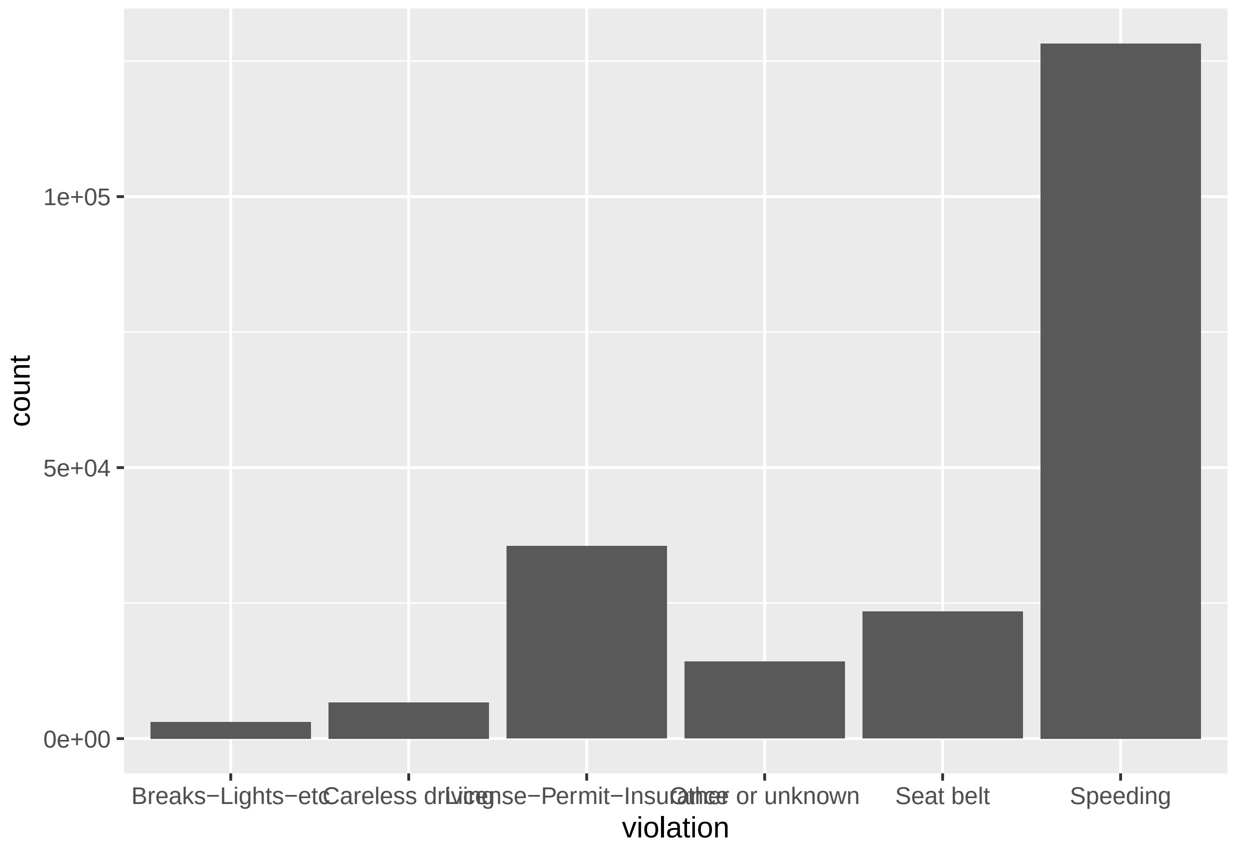
```
##
##        Breaks-Lights-etc         Careless driving License-Permit-Insurance
##                     3100                     6662                    35530
##         Other or unknown                Seat belt                 Speeding
##                    14207                    23435                   128277
```

```
barplot(table(MS_traffic_stops$violation),col = "red", main="Number of violations")
```



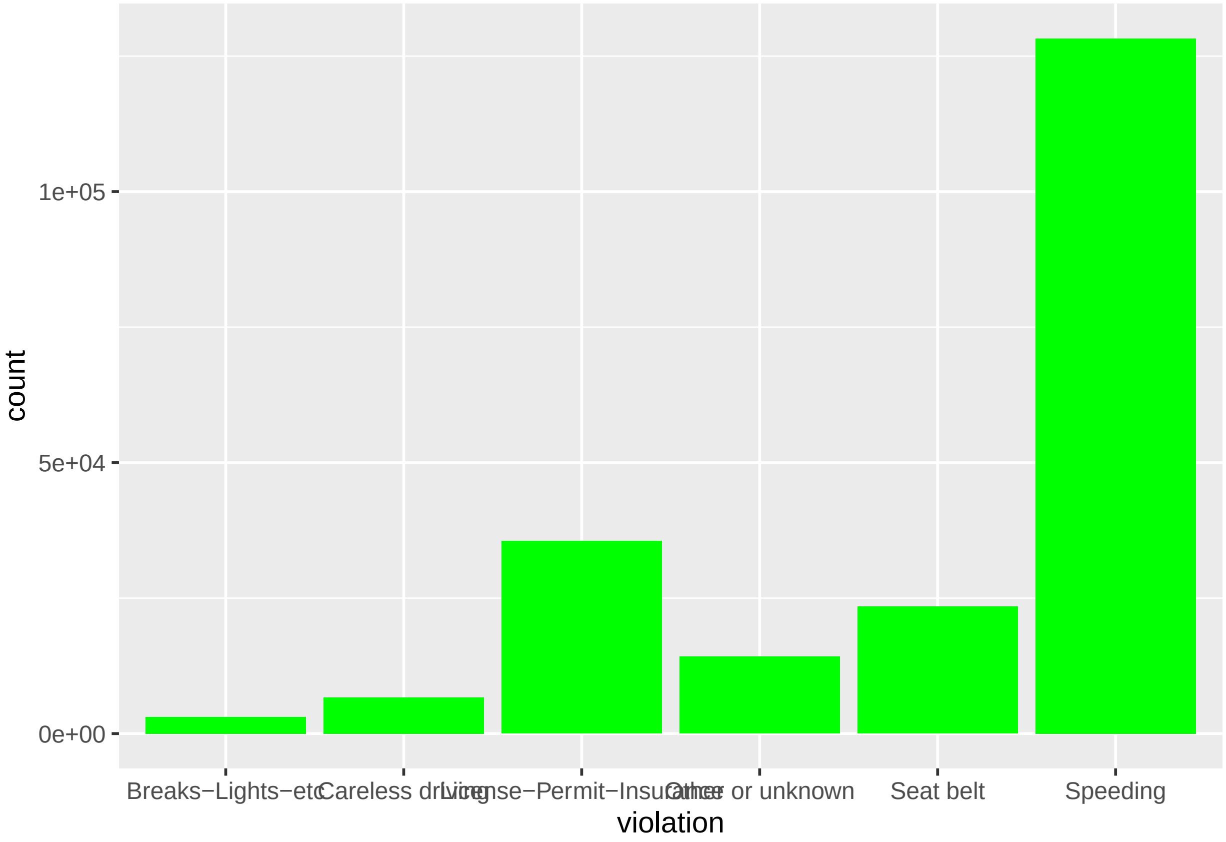
task 2:

```r
ggplot(MS_traffic_stops, aes(violation)) + geom_bar()
```



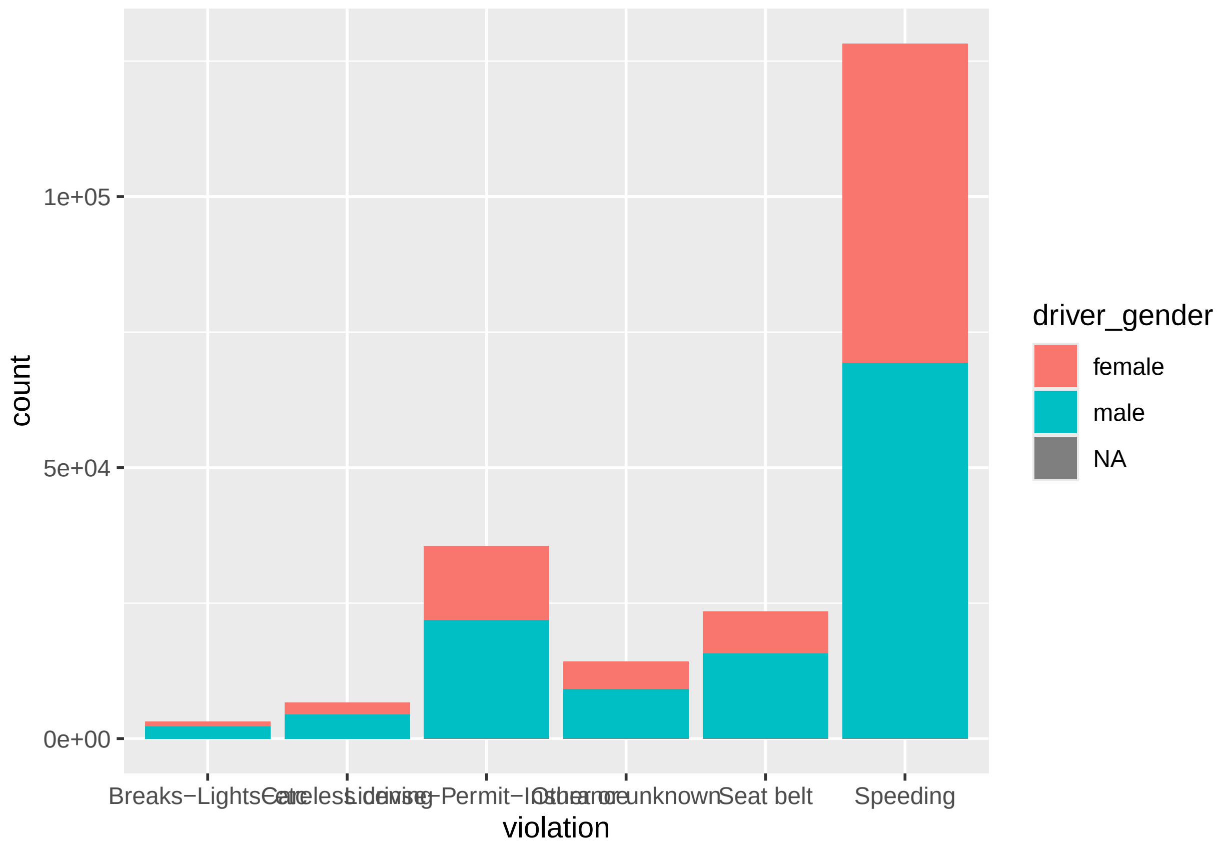
task 3:colouring the bars, we will use fill, instead of colour

```
ggplot(MS_traffic_stops, aes(violation)) + geom_bar(fill = "green")
```



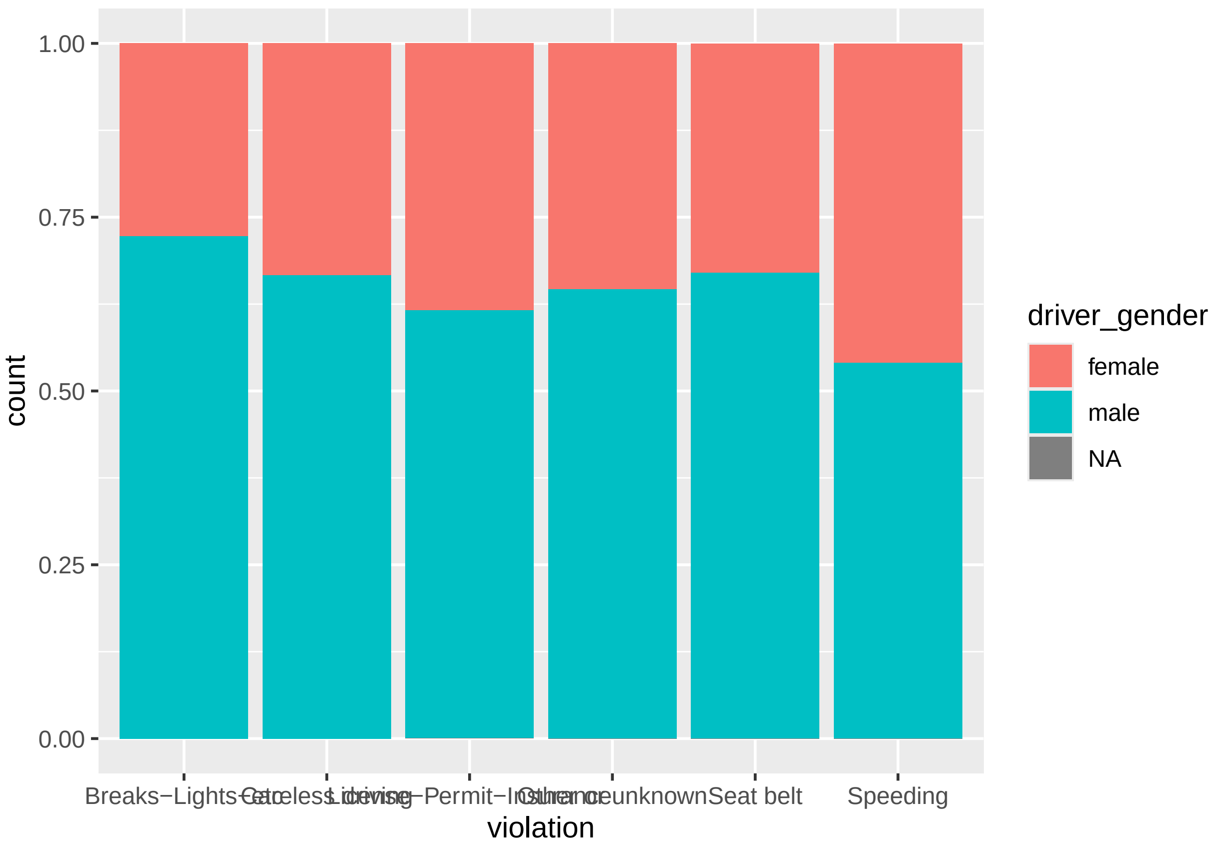
task 4:Mapping the values to different colours.

```
ggplot(MS_traffic_stops, aes(violation)) + geom_bar(aes(fill = driver_gender))
```



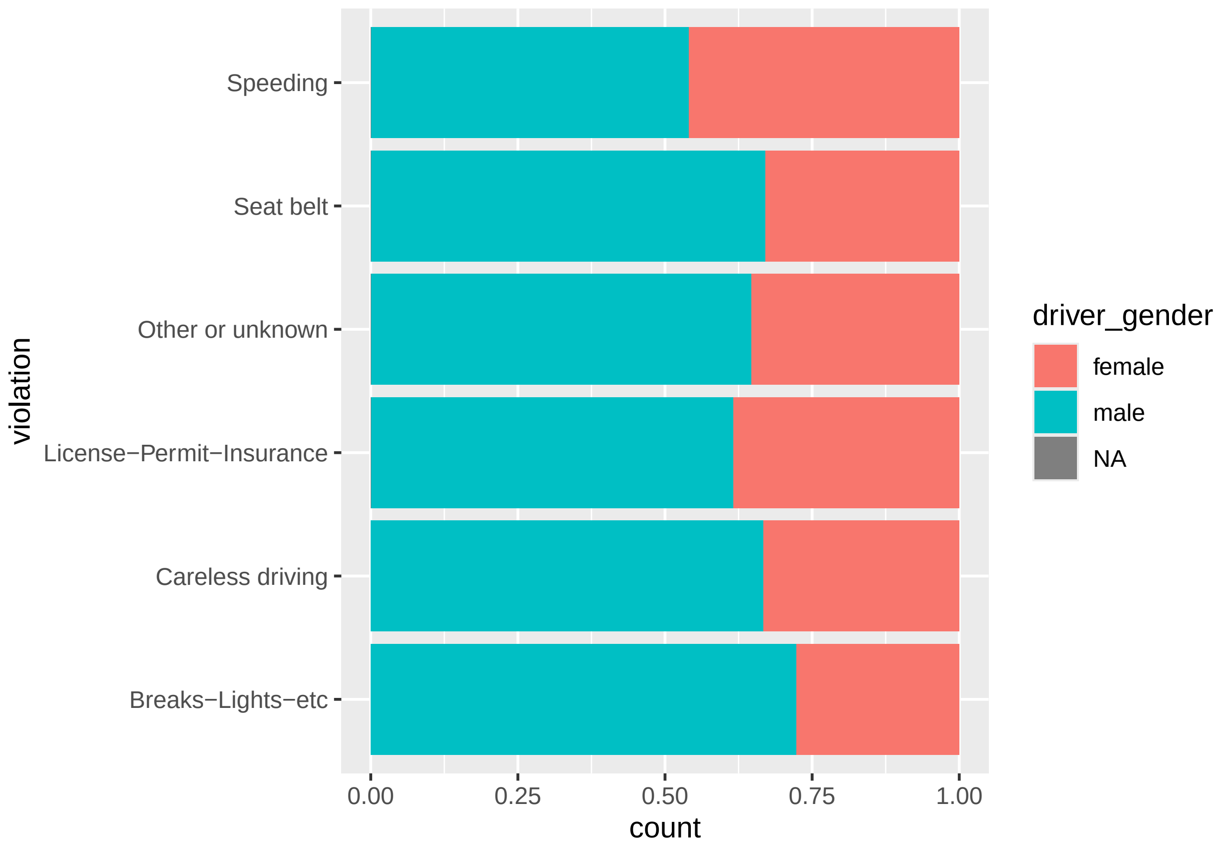
task 5:stretch the bars between 0 and 1, by setting the position parameter to ‘fill’.

```
ggplot(MS_traffic_stops, aes(violation)) + geom_bar(aes(fill = driver_gender), position = "fill")
```



task 6:Adding another function “coord_flip()”. In some scenarios, flipping will make the plot more readable.

```
ggplot(MS_traffic_stops, aes(violation)) + geom_bar(aes(fill = driver_gender), position = "fill") + coo
```

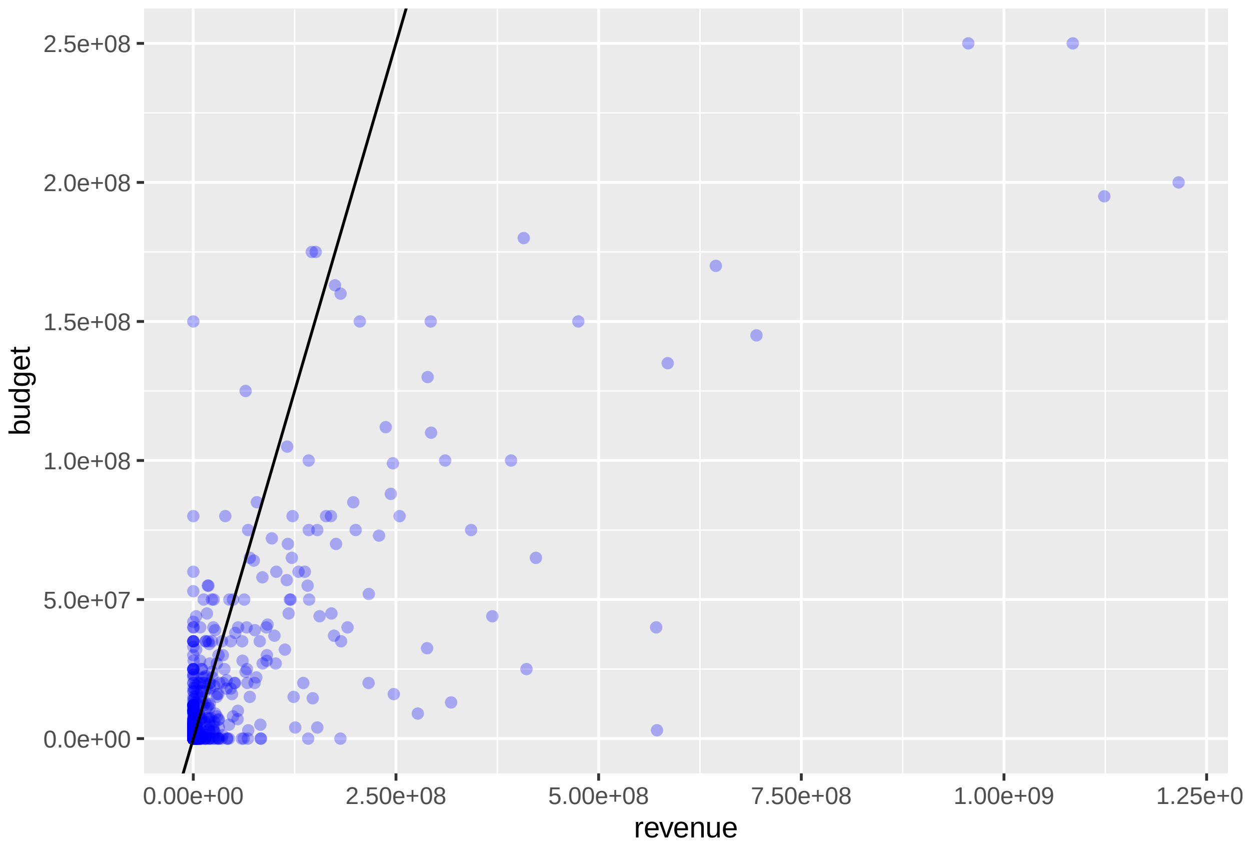


Correlation Analysis

```
# Load movies data set
movies <- read.csv('movies.csv')
```

Graphical Analysis

```
ggplot(data = movies, aes(x=revenue, y=budget)) + geom_point(alpha= 0.3, color=  "blue")+ geom_abline()
```



Quantitative Analysis

```
cor(movies$vote_average, movies$revenue)
```

```
## [1] 0.06986166
```

Correlation Matrix

task 1:Loading library

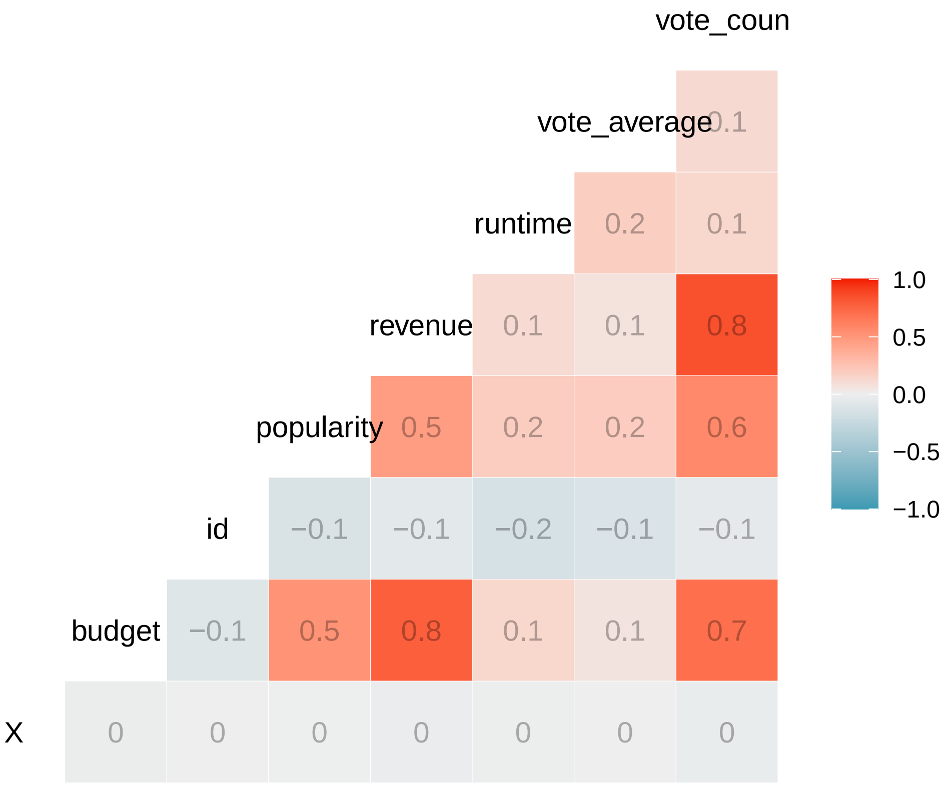
```
#install.packages("GGally")

library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

task 2L Displaying the heatmap

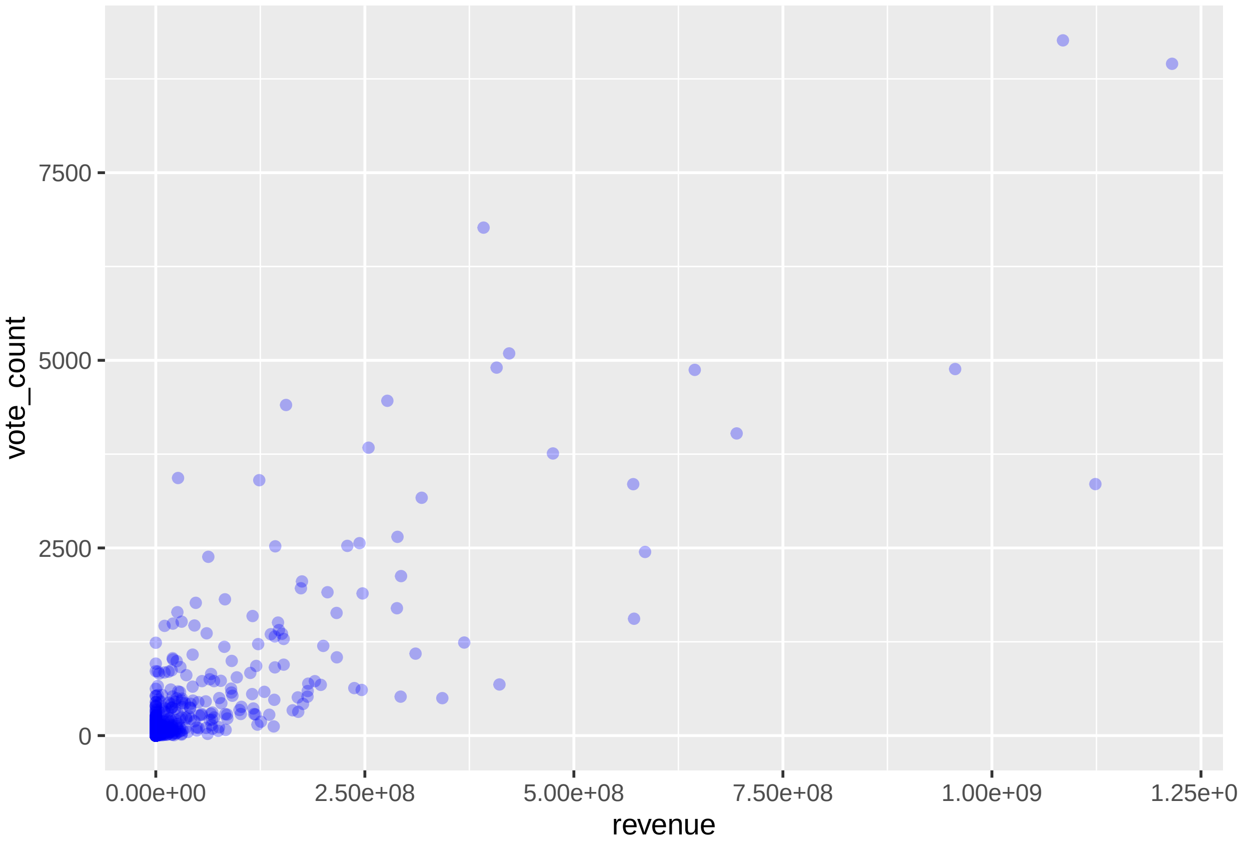
```
ggcorr(movies, label= TRUE , label_alpha= 0.3)
```

```
## Warning in ggcorr(movies, label = TRUE, label_alpha = 0.3): data in column(s)
## 'adult', 'original_language', 'release_date', 'status', 'title' are not numeric
## and were ignored
```



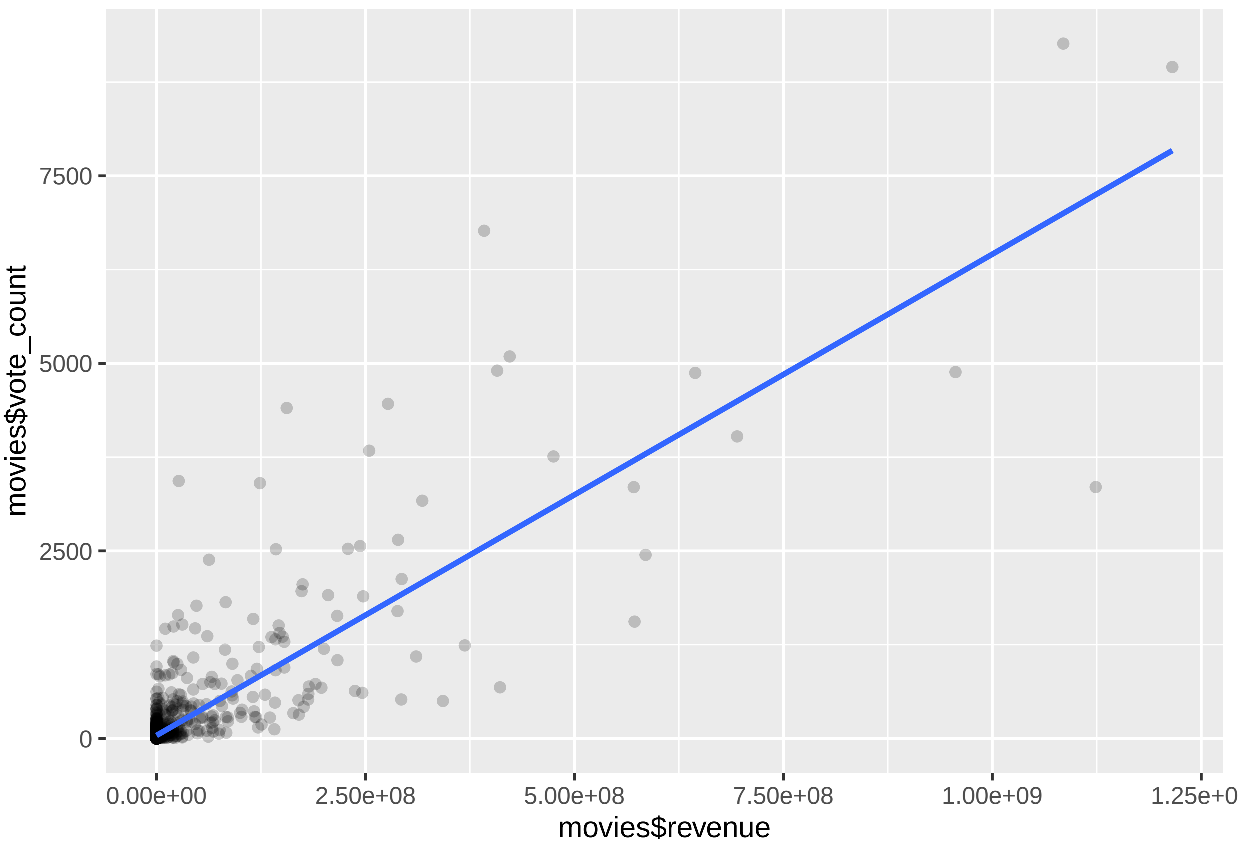
pairwise correlation

```
ggplot(data = movies, aes(x = revenue, y = vote_count)) + geom_point(alpha=
0.3, color= "blue")
```



```
qplot(movies$revenue, movies$vote_count, data=movies , geom= c("point",
"smooth" ), method= "lm", alpha= I (1/5), se= FALSE)
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning in geom_point(method = "lm", alpha = structure(0.2, class = "AsIs"), :
## Ignoring unknown parameters: ‘method‘ and ‘se‘

## Warning: Use of ‘movies$revenue‘ is discouraged.
## i Use ‘revenue‘ instead.

## Warning: Use of ‘movies$vote_count‘ is discouraged.
## i Use ‘vote_count‘ instead.

## Warning: Use of ‘movies$revenue‘ is discouraged.
## i Use ‘revenue‘ instead.

## Warning: Use of ‘movies$vote_count‘ is discouraged.
## i Use ‘vote_count‘ instead.

## ‘geom_smooth()‘ using formula = ’y ~ x’
```


7500
5000
2500
0
movies$vote_count
0.00e+00
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+0
movies$revenue