

EDA ANALYSIS OF ROAD ACCIDENTS IN INDIA – 2018

##1. ACCIDENTS IN 2018 WITHOUT WEARING HELMET

```
##loading dataset
```

```
non_wearing_helmet<-read.csv("C:/Users/Tenisha/Documents/acci_data/Road-Accidents-2018-non wearing of helmet.csv")
```

```
#checking null values
```

```
is.null(non_wearing_helmet) %>% sum()
```

```
## [1] 0
```

```
##removing unnecessary rows and columns
```

```
##removing total rows
```

```
data1<-non_wearing_helmet[-37,]
```

```
##removing unwanted columns which contains
```

```
data2<-data1[-c(1,4,7,9,12)]
```

```
##calculating mean
```

```
apply(data2[,2:7],2,mean)
```

```
##           Drivers...Persons.Killed...Number
##                                           784.7222
##   Drivers...Persons.Injured...Greviously.Injured
##                                           631.1111
##           Drivers...Persons.Injured...Minor.Injury
##                                           1326.2500
##           Passengers...Persons.Killed...Number
##                                           426.7778
##   Passengers...Persons.Injured...Greviously.Injured
##                                           526.6944
##           Passengers...Persons.Injured...Minor.Injury
##                                           851.2778
```

```
##calculating median
```

```
apply(data2[,2:7],2,median)
```

```
##           Drivers...Persons.Killed...Number
##                                           138.5
##   Drivers...Persons.Injured...Greviously.Injured
##                                           201.0
##           Drivers...Persons.Injured...Minor.Injury
```

```
##                                181.5
##          Passengers...Persons.Killed...Number
##                                71.0
## Passengers...Persons.Injured...Greviously.Injured
##                                106.5
##          Passengers...Persons.Injured.  Minor.Injury
##                                114.0
```

```
##boxplot
```

```
library(ggplot2)
```

```
data3<-data2[-1]
```

```
library(reshape2)
```

```
colnames(data3)<-c("Drivers.killed","Drivers.grev.injured","Drivers.minorinjured",
"Passenger.killed","Passengers.grev.injured","Passengers.minorinjured")
```

```
data4<-melt(data3)
```

```
## No id variables; using all as measure variables
```

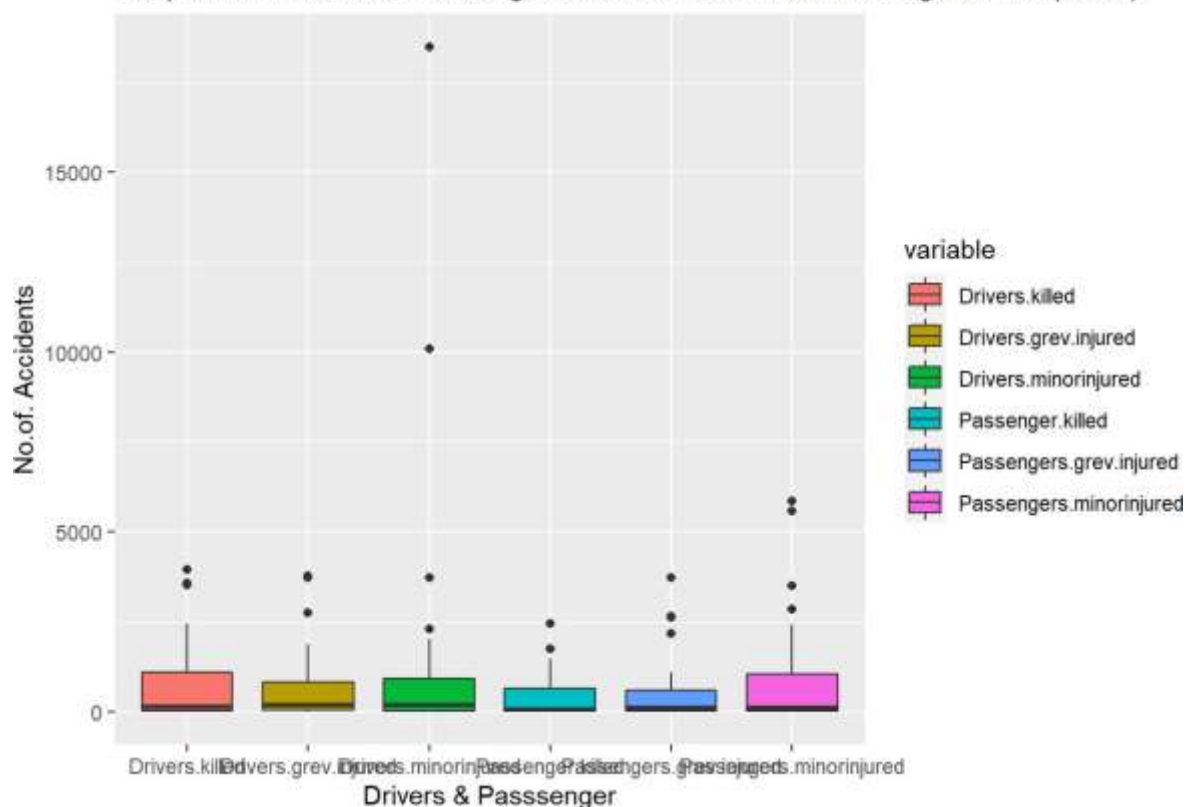
```
##boxplot
```

```
a<-ggplot(data4,aes(x=variable,y=value,fill=variable))+geom_boxplot()+ggtitle("Boxplot for Drivers & Passengers Accidents without wearing Helmet (2018)")+
```

```
  xlab("Drivers & Passsenger")+ylab("No.of. Accidents")
```

```
a
```

Boxplot for Drivers & Passengers Accidents without wearing Helmet (2018)



Inference:

The above box plot shows the minimum and maximum, first quartile(lower), third quartile(upper) and median values for all the variables. In the above plot by comparing drivers and passenger's drivers minor injured are most killed without wearing helmet.

```
##plotting histogram
##histogram for drivers not wearing helmet
par(mfrow=c(3,1))
par(mar=rep(2,4))

hist(data3$Drivers.killed,col = "blue",breaks=30,main = "Drivers.Killed.wit
hout wearing helmet",xlab = "no.of.drivers.killed",ylab = "Frequency")

abline(v=mean(data3$Drivers.killed),col = "black",lwd=3)
abline(v=median(data3$Drivers.killed),col = "yellow",lwd=3)

legend(x="topright",c("Density","mean","median"),col=c("blue","black","yell
ow"),cex=0.75,lwd=c(3,3,3))

hist(data3$Drivers.grev.injured,breaks=30,col = "orange",main = "Drivers.gr
ievously.injured.without wearing helmet",xlab = "no.of.drivers.killed",ylab
= "Frequency")

abline(v=mean(data3$Drivers.grev.injured),col = "green",lwd=3)
abline(v=median(data3$Drivers.grev.injured),col = "red",lwd=3)
```

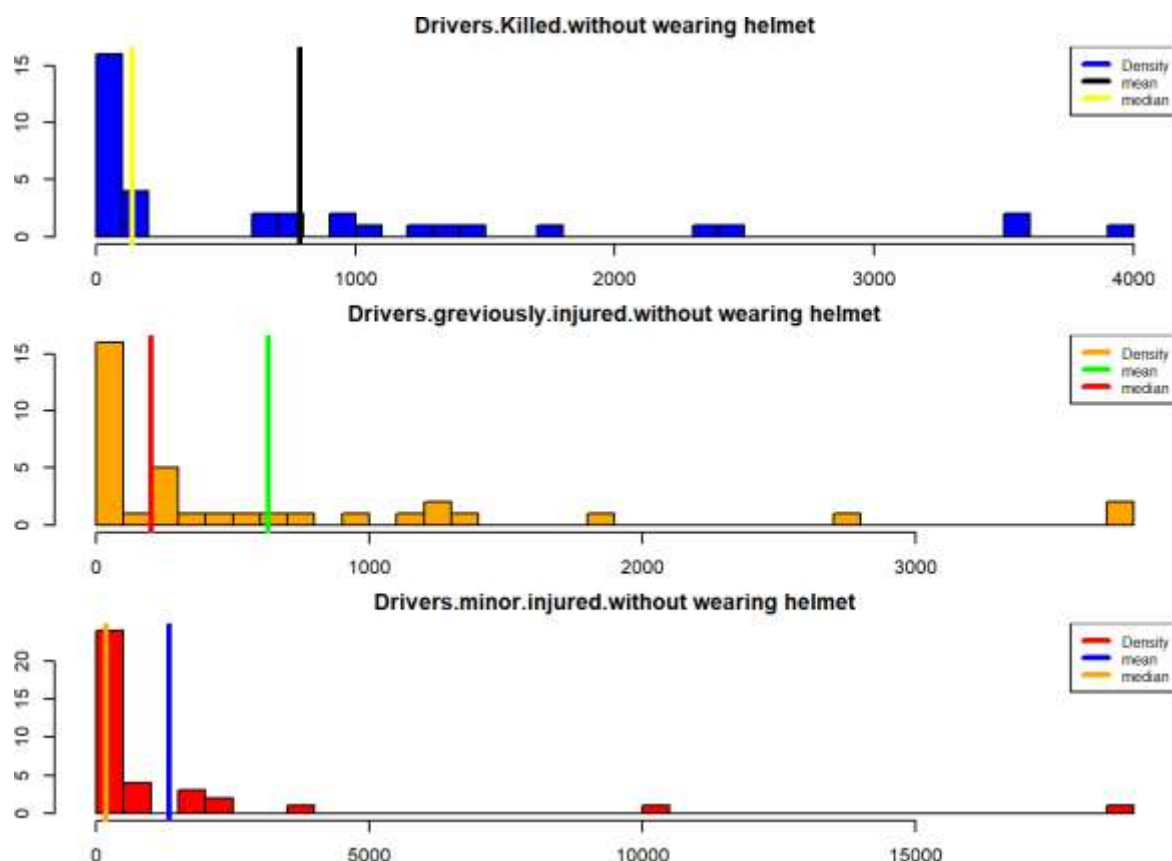
```
legend(x="topright",c("Density","mean","median"),col=c("orange","green","red"),cex=0.75,lwd=c(3,3,3))
```

```
hist(data3$Drivers.minorinjured,breaks=30,col = "red",main = "Drivers.minor.injured.without wearing helmet",xlab = "no.of.drivers.killed",ylab = "Frequency")
```

```
abline(v=mean(data3$Drivers.minorinjured),col = "blue",lwd=3)
```

```
abline(v=median(data3$Drivers.minorinjured),col = "orange",lwd=3)
```

```
legend(x="topright",c("Density","mean","median"),col=c("red","blue","orange"),cex=0.75,lwd=c(3,3,3))
```



Inference:

The above plot is right skewed distribution. It shows the histogram for drivers killed ,drivers grievously injured and drivers minor injured. In this drivers killed median value is greater than drivers grievously injured. By comparing drivers grievously injured and drivers minor injured the median value is less than drivers minor injured.

```
##histogram for passenger not wearing helmet
```

```
par(mfrow=c(3,1))
```

```
par(mar=rep(2,4))
```

```
hist(data3$Passenger.killed,breaks=30,col = "green",main = "Passengers.Killed.without wearing helmet",xlab = "no.of.Passengers.killed",ylab = "Frequency")
```

```

abline(v=mean(data3$Passenger.killed),col = "blue",lwd=3)
abline(v=median(data3$Passenger.killed),col = "black",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("green","blue","black"),cex=0.75,lwd=c(3,3,3))

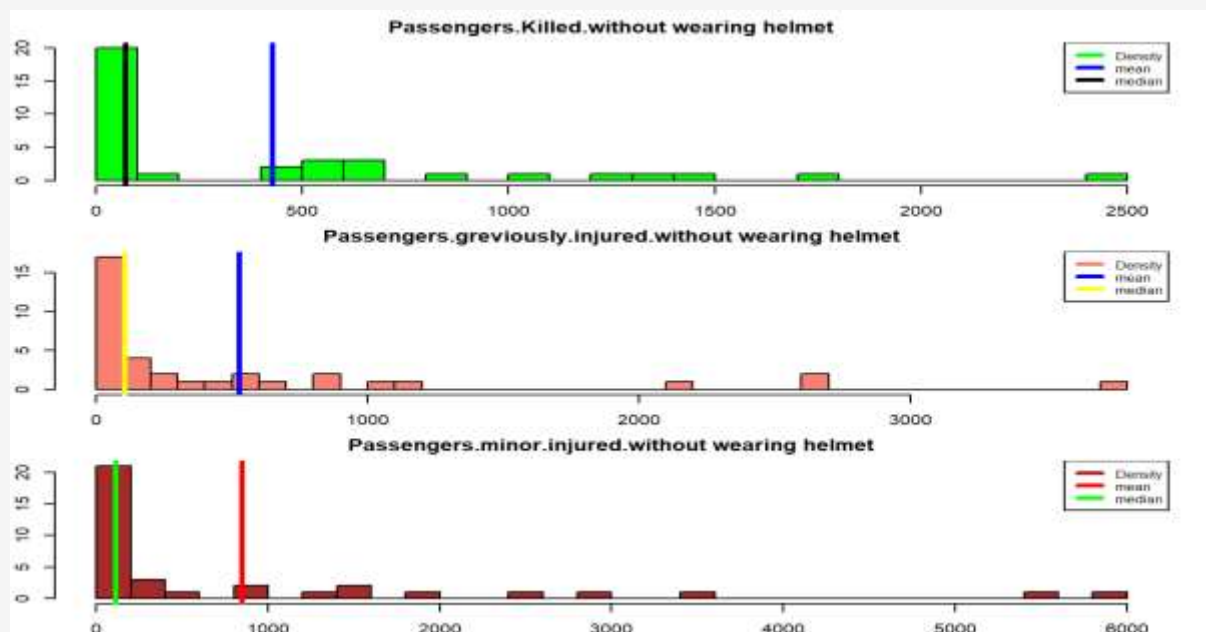
hist(data3$Passengers.grev.injured,breaks=30,col = "salmon",main = "Passengers.grievously.injured.without wearing helmet",xlab = "no.of.Passengers.killed",ylab = "Frequency")

abline(v=mean(data3$Passengers.grev.injured),col = "blue",lwd=3)
abline(v=median(data3$Passengers.grev.injured),col = "yellow",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("salmon","blue","yellow"),cex=0.75,lwd=c(3,3,3))

hist(data3$Passengers.minorinjured,breaks=30,col = "brown",main = "Passengers.minor.injured.without wearing helmet",xlab = "no.of.Passengers.killed",ylab = "Frequency")

abline(v=mean(data3$Passengers.minorinjured),col = "red",lwd=3)
abline(v=median(data3$Passengers.minorinjured),col = "green",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("brown","red","green"),cex=0.75,lwd=c(3,3,3))

```



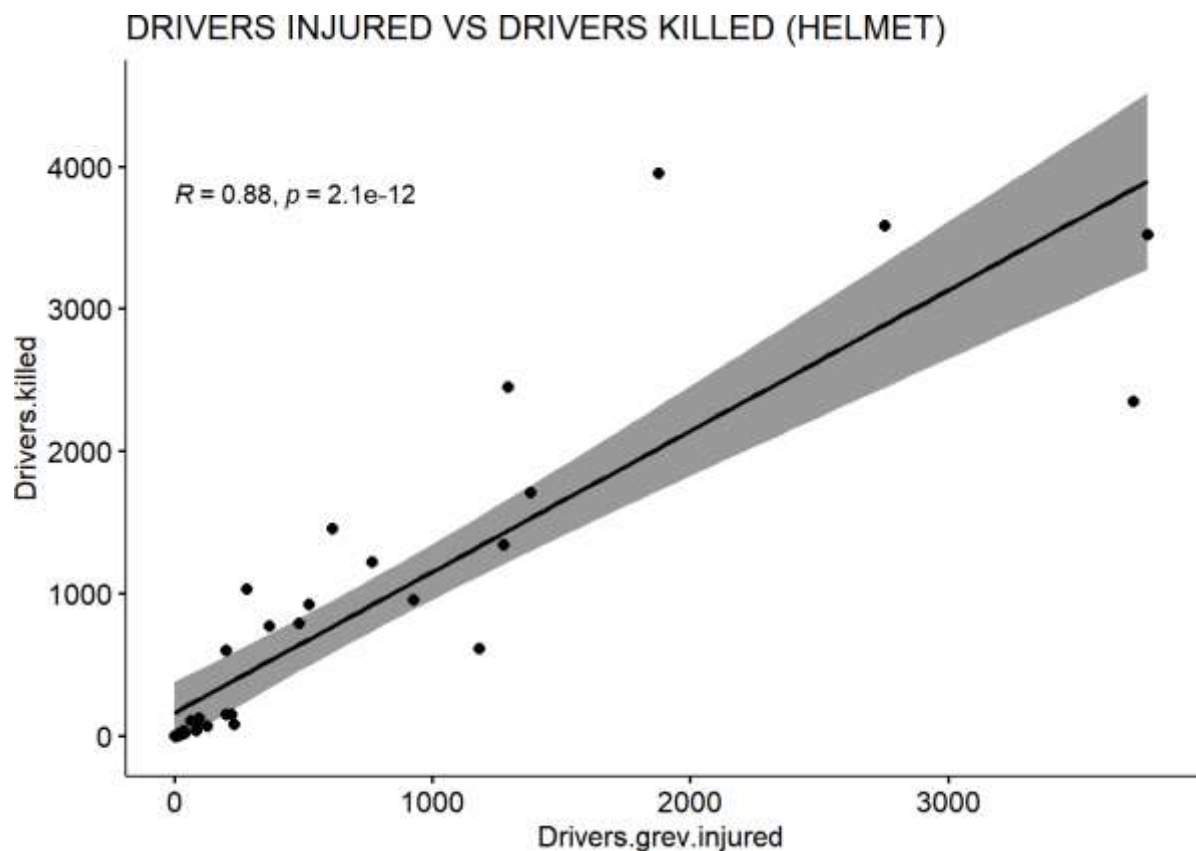
Inference:

The above plot is right skewed distribution. It shows the histogram for passengers killed ,passengers grievously injured, passengers minor injured. In this passengers killed mean value 426.778 and passengers grievously injured mean value 526.6944 .By comparing the passengers grievously killed median value is greater than passengers minor injured.

```
#SCATTER PLOT

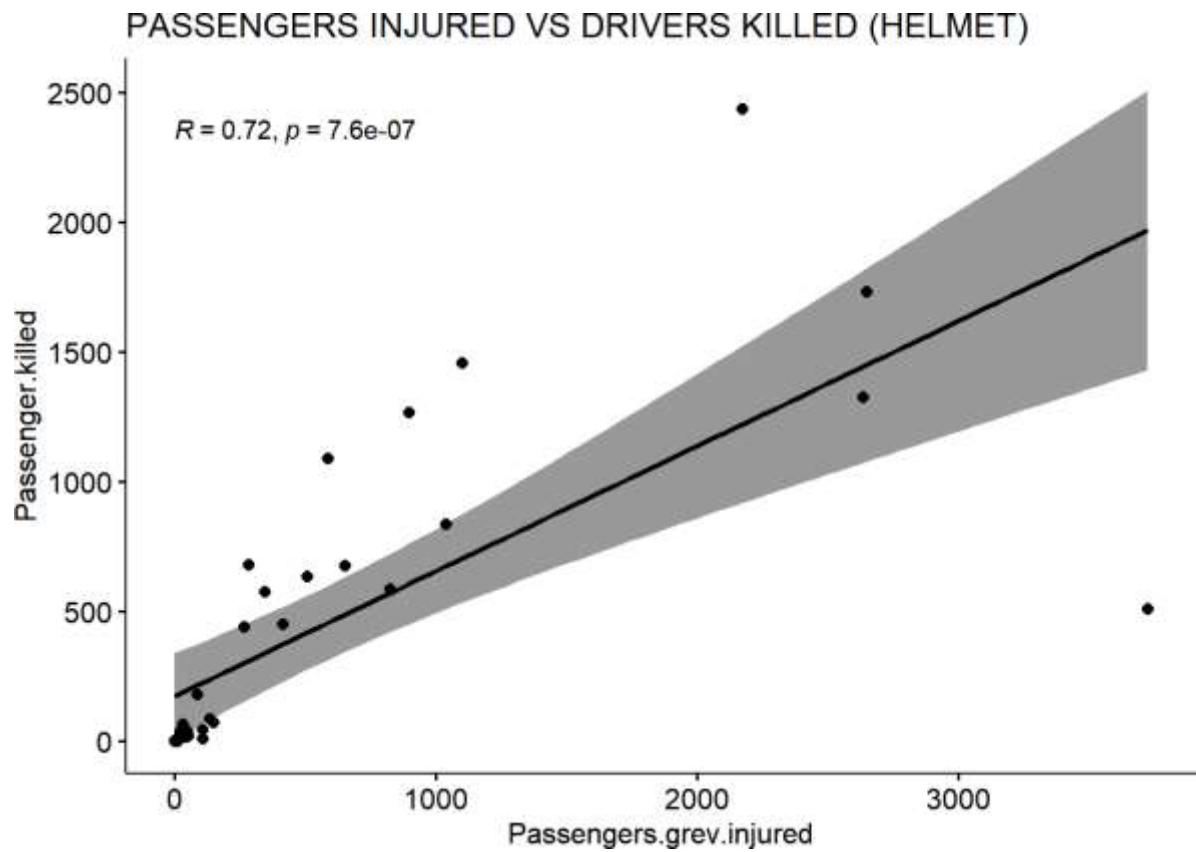
#Drivers

ggscatter(data33,x="Drivers.grev.injured",y="Drivers.killed",add = "reg.line",conf.int = TRUE,cor.coef = TRUE,method = "pearson")+ggtitle("DRIVERS INJURED VS DRIVERS KILLED (HELMET)")
```



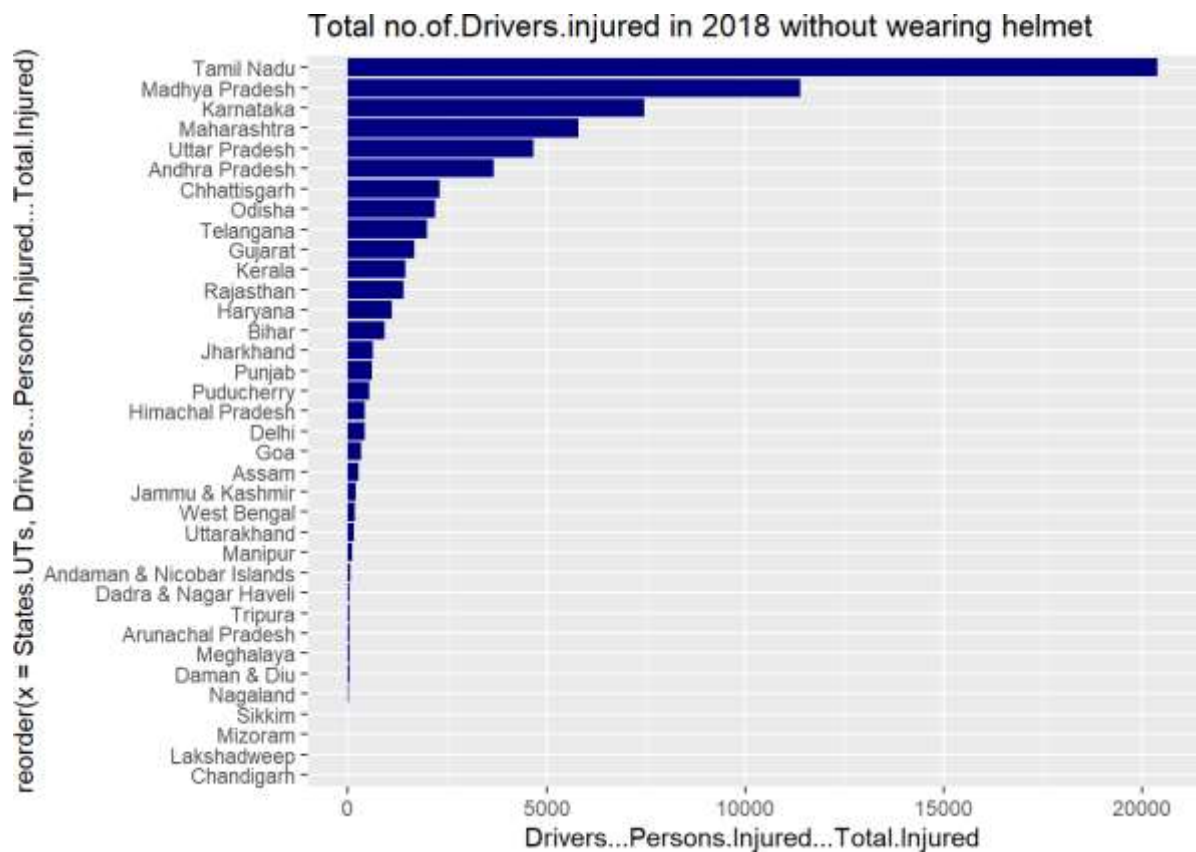
```
#Passengers

ggscatter(data33,x="Passengers.grev.injured",y="Passenger.killed",add = "reg.line",conf.int = TRUE,cor.coef = TRUE,method = "pearson")+ggtitle("PASSENGERS INJURED VS DRIVERS KILLED (HELMET)")
```



```
b<-ggplot(data=data1,aes(reorder(x=States.UTs,Drivers...Persons.Injured...Total.Injured),y=Drivers...Persons.Injured...Total.Injured))+coord_flip()+geom_bar(stat="identity",fill = "navy")+ggtitle("Total no.of.Drivers.injured in 2018 without wearing helmet")
```

b



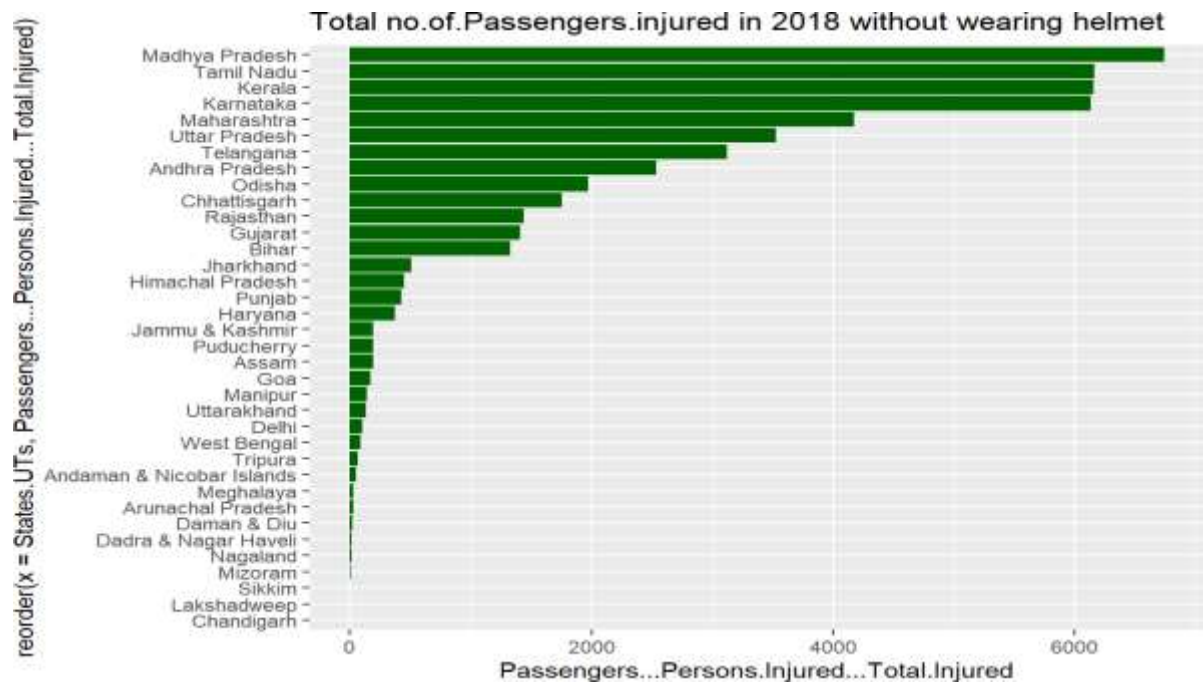
Inference:

The above plot shows the Total number of drivers injured in 2018 without wearing helmet. The total number of drivers are mostly injured in Tamilnadu. The top five injured states are Tamilnadu, Madhya Pradesh, Karnataka, Maharashtra, Uttar Pradesh.

```
##barplot for not wearing helmet(Passengers)

c<-ggplot(datal,aes(reorder(x=States.UTs,Passengers...Persons.Injured...Total.Injured),y=Passengers...Persons.Injured...Total.Injured))+coord_flip()+geom_bar(stat="identity",fill = "dark green")+ggtitle("Total no.Of.Passengers.injured in 2018 without wearing helmet")
```

c



Inference:

The above plot shows the Total number of passengers injured in 2018 without wearing helmet. The total number of passengers are mostly injured in Madhya Pradesh. The top five injured states are Madhya Pradesh, Tamilnadu, Kerala, Karnataka, Maharashtra.

```
##barplot for Accidents happened for Drivers & Passengers in 2018-without helmet

library(RColorBrewer)

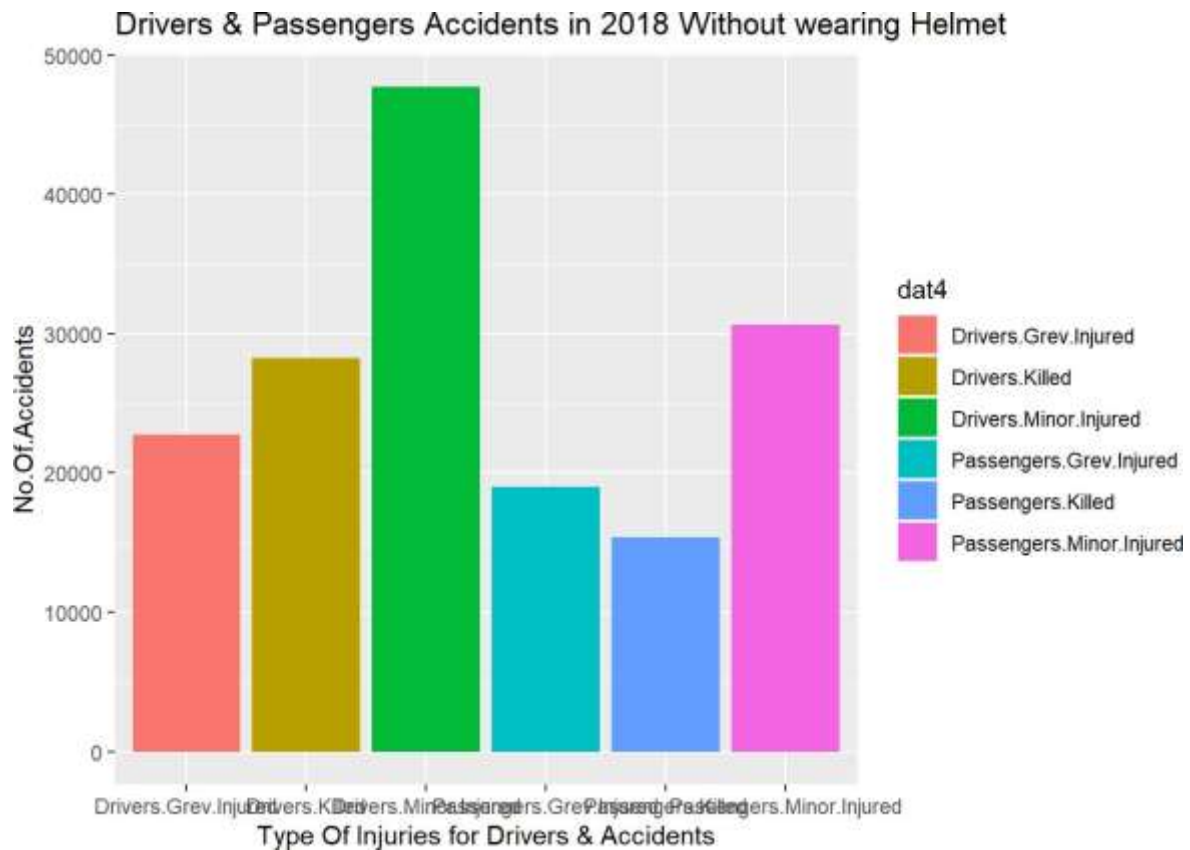
dat1<-non_wearing_helmet[37,]
dat2<-dat1[c(3,5,6,8,10,11)]
dat3<-data.frame(t(dat2))

dat4<-c("Drivers.Killed","Drivers.Grev.Injured","Drivers.Minor.Injured","Passengers.Killed","Passengers.Grev.Injured","Passengers.Minor.Injured")
dat5<-data.frame(dat4,dat3$X37)

coll<-brewer.pal(6,"Set1")

ggplot(dat5,aes(x=dat5$dat4,y=dat5$dat3.X37,fill=dat4))+geom_bar(stat="identity",)+theme()+ggtitle("Drivers & Passengers Accidents in 2018 Without wearing Helmet")+xlab("Type Of Injuries for Drivers & Accidents")+ylab("No.Of. Accidents")

## Warning: Use of `dat5$dat4` is discouraged. Use `dat4` instead.
## Warning: Use of `dat5$dat3.X37` is discouraged. Use `dat3.X37` instead.
```



Inference:

The above plot shows the drivers and passengers Accidents in 2018 without wearing helmet. In this we can see the highest value was drivers Minor injured.

##2. ACCIDENTS IN 2018 WITHOUT WEARING SEATBELT

```
##loading dataset
```

```
non_wearing_seatbelt<-read.csv("C:/Users/Tenisha/Documents/acci_data/Non-Use of Safety Device (Non-Wearing of Seat Belt) by Victims during 2018.csv")
```

```
library(dplyr)
```

```
##checking missing values
```

```
##removing unnecessary rows and columns
```

```
##removing total rows
```

```
data1<-non_wearing_seatbelt[-37,]
```

```
##removing unwanted columns which contains
```

```
data2<-data1[-c(1,4,7,9,12)]
```

```
##boxplot
```

```
library(ggplot2)
```

```
dat3<-data2[-1]
```

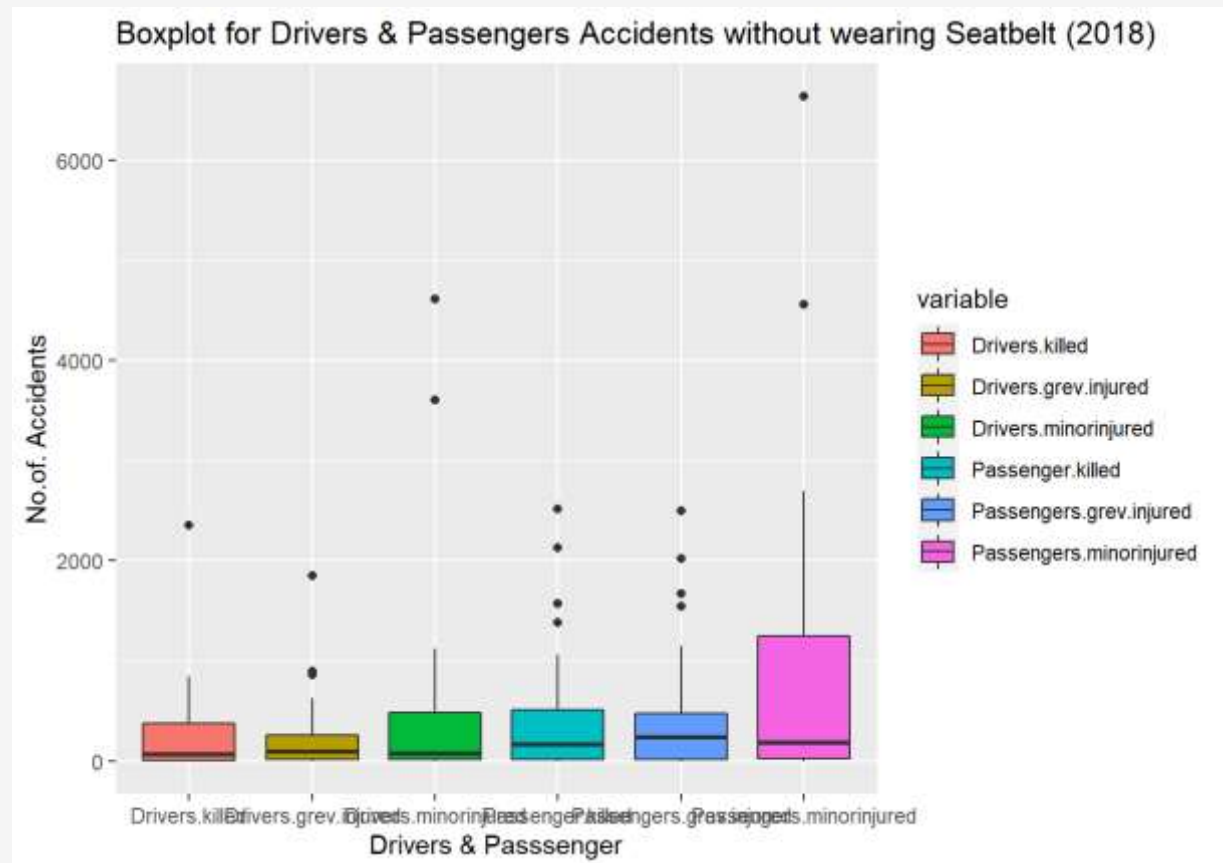
```
library(reshape2)

colnames(dat3)<-c("Drivers.killed","Drivers.grev.injured","Drivers.minorinj
ured","Passenger.killed","Passengers.grev.injured","Passengers.minorinj
ured")

data4<-melt(dat3)

## No id variables; using all as measure variables

a<-ggplot(data4,aes(x=variable,y=value,fill=variable))+geom_boxplot()+ggtit
le("Boxplot for Drivers & Passengers Accidents without wearing Seatbelt (20
18)")+xlab("Drivers & Passsenger")+ylab("No.of. Accidents")
```



Inference:

In the above plot by comparing drivers and passengers accidents without wearing seatbelt passengers minor injured are killed the most without wearing seatbelt.

```
##plotting histogram
##histogram for drivers not wearing helmet
par(mfrow=c(3,1))
par(mar=rep(2,4))
```

```
hist(data3$Drivers.killed,breaks=30,col = "gold",main = "Drivers.Killed.wit
hout wearing Seatbelt",xlab = "no.of.drivers.killed",ylab = "Frequency")

abline(v=mean(data3$Drivers.killed),col = "black",lwd=3)

abline(v=median(data3$Drivers.killed),col = "green",lwd=3)

legend(x="topright",c("Density","mean","median"),col=c("gold","black","gree
n"),cex=0.75,lwd=c(3,3,3))
```

```
hist(data3$Drivers.grev.injured,breaks=30,col = "deep sky blue",main = "Dri
vers.grievously.injured.without wearing Seatbelt",xlab = "no.of.drivers.kil
led",ylab = "Frequency")

abline(v=mean(data3$Drivers.grev.injured),col = "brown",lwd=3)

abline(v=median(data3$Drivers.grev.injured),col = "orange",lwd=3)

legend(x="topright",c("Density","mean","median"),col=c("deep sky blue","bro
wn","orange"),cex=0.75,lwd=c(3,3,3))
```

```
hist(data3$Drivers.minorinjured,breaks=30,col = "wheat",main = "Drivers.min
or.injured.without wearing Seatbelt",xlab = "no.of.drivers.killed",ylab = "
Frequency")

abline(v=mean(data3$Drivers.minorinjured),col = "red",lwd=3)

abline(v=median(data3$Drivers.minorinjured),col = "maroon",lwd=3)

legend(x="topright",c("Density","mean","median"),col=c("wheat","red","maroo
n"),cex=0.75,lwd=c(3,3,3))
```



Inference:

The above plot is right skewed distribution. The above histogram shows the drivers killed ,drivers grievously killed and drivers minor injured. In this drivers grievously injured median value is greater which is 91.50.

```
##histogram for passenger not wearing seatbelt
par(mfrow=c(3,1))
par(mar=rep(2,4))

hist(data3$Passenger.killed,breaks=30,col = "tan",main = "Passengers.Killed
.without wearing Seatbelt",xlab = "no.of.Passengers.killed",ylab = "Frequency")

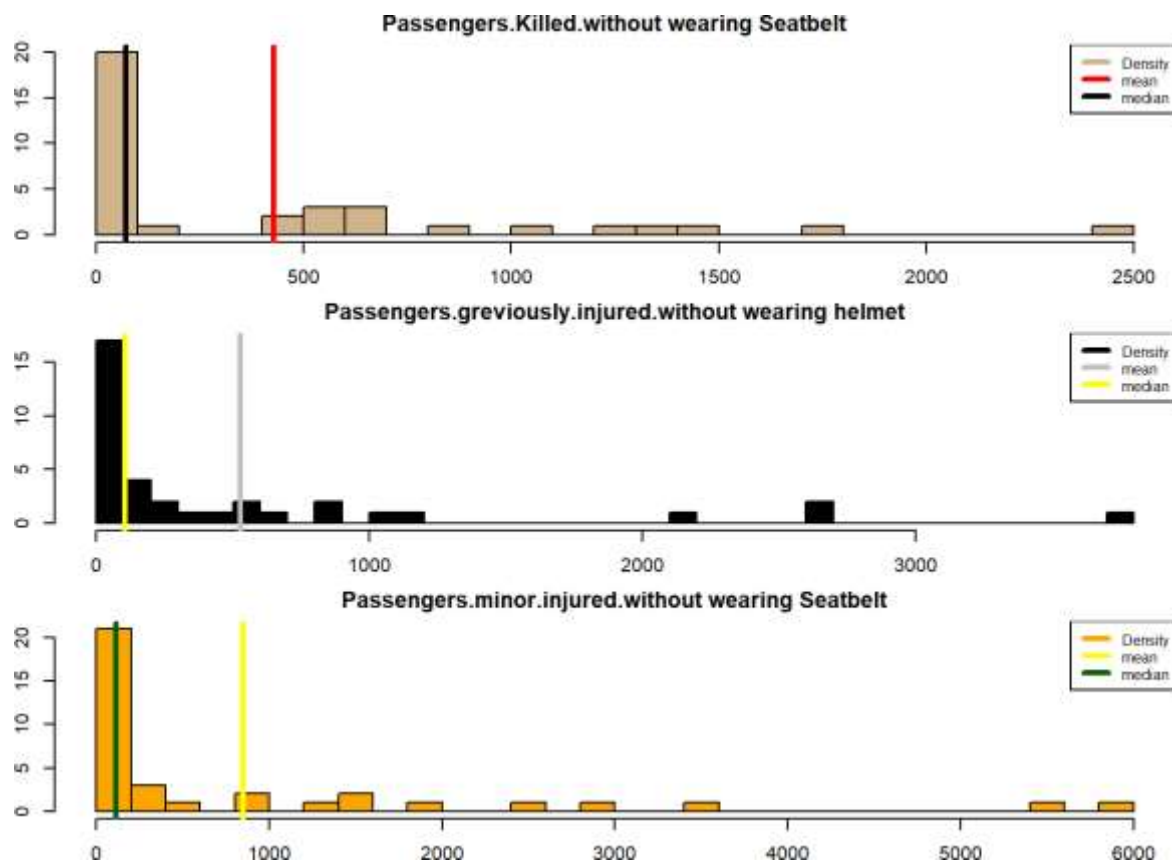
abline(v=mean(data3$Passenger.killed),col = "red",lwd=3)
abline(v=median(data3$Passenger.killed),col = "black",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("tan","red","black"),
,cex=0.75,lwd=c(3,3,3))

hist(data3$Passengers.grev.injured,breaks=30,col = "black",main = "Passengers.grievously.injured.without wearing helmet",xlab = "no.of.Passengers.killed",ylab = "Frequency")

abline(v=mean(data3$Passengers.grev.injured),col = "gray",lwd=3)
abline(v=median(data3$Passengers.grev.injured),col = "yellow",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("black","gray","yellow"),cex=0.75,lwd=c(3,3,3))

hist(data3$Passengers.minorinjured,breaks=30,col = "orange",main = "Passengers.minor.injured.without wearing Seatbelt",xlab = "no.of.Passengers.killed",ylab = "Frequency")

abline(v=mean(data3$Passengers.minorinjured),col = "yellow",lwd=3)
abline(v=median(data3$Passengers.minorinjured),col = "dark green",lwd=3)
legend(x="topright",c("Density","mean","median"),col=c("orange","yellow","dark green"),cex=0.75,lwd=c(3,3,3))
```



Inference:

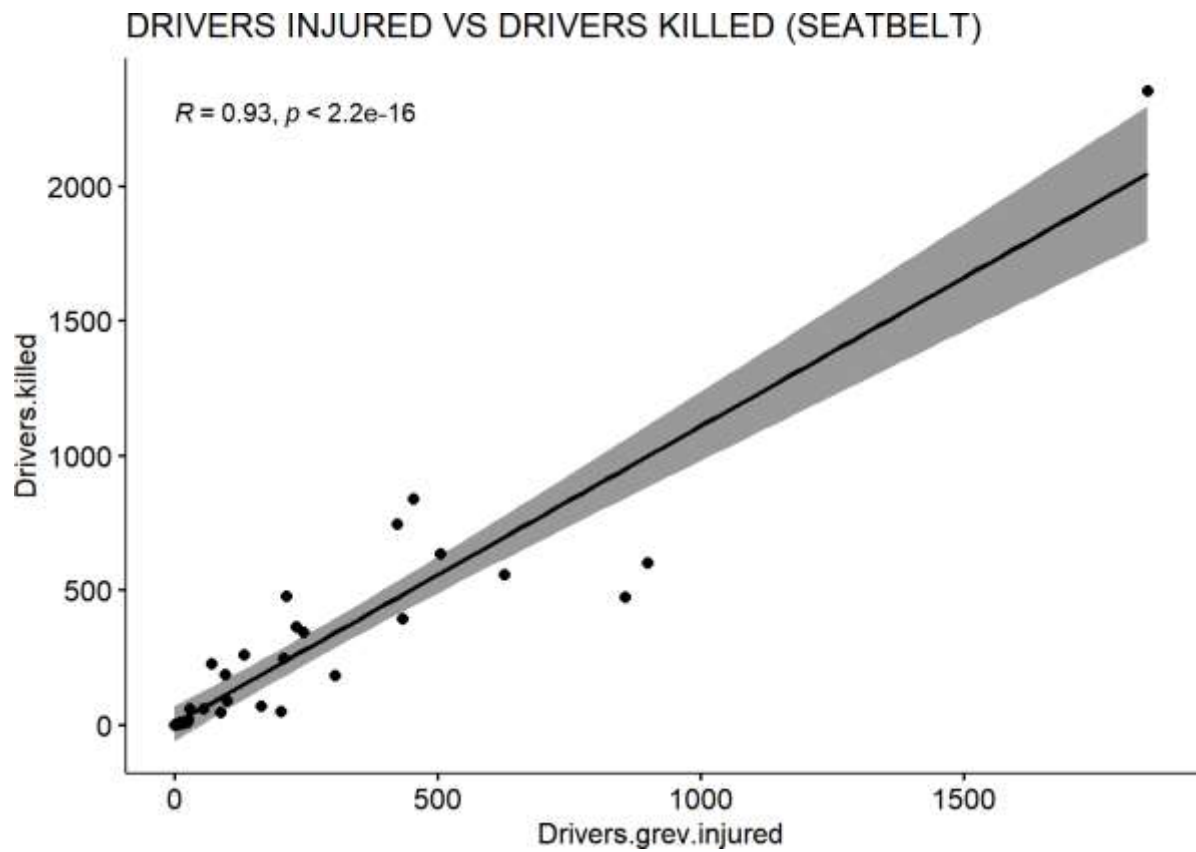
The above plot is right skewed distribution. The above plot shows histogram for passengers killed, passengers grievously injured and passengers minor injured. In this passengers killed mean value is 419.1 and passengers grievously injured mean value is 441.1. By comparing the median value of passengers grievously injured and passengers minor injured, passengers grievously injured median value is greater than passengers minor injured.

```
#SCATTER PLOT DRIVERS
```

```
ggscatter(data3,x="Drivers.grev.injured",y="Drivers.killed",add = "reg.line",
  conf.int = TRUE,cor.coef = TRUE,method = "pearson")+ggtitle("DRIVERS INJURED VS DRIVERS KILLED (SEATBELT)")
```

```
## Warning: Ignoring unknown parameters: method
```

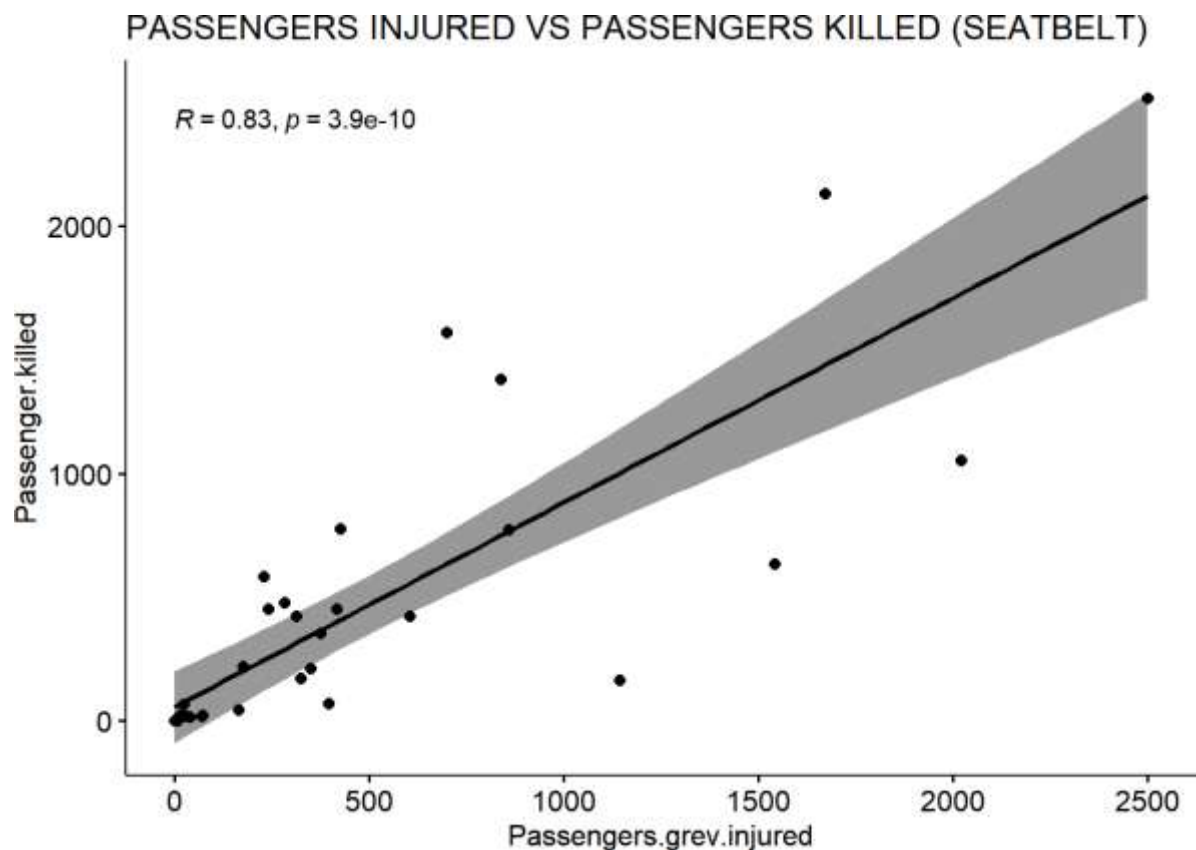
```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggscatter(data3,x="Passengers.grev.injured",y="Passenger.killed",add = "reg  
.line",conf.int = TRUE,cor.coef = TRUE,method = "pearson")+ggtitle("PASSENG  
ERS INJURED VS PASSENGERS KILLED (SEATBELT)")
```

```
## Warning: Ignoring unknown parameters: method
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
##barplot for not wearing Seatbelt(Drivers)
```

```
b<-ggplot(data=data1,aes(reorder(x=States.UTs,Drivers...Persons.Injured...Total.Injured),y=Drivers...Persons.Injured...Total.Injured,fill=States.UTs))
+coord_flip()+geom_bar(stat="identity")+ggtitle("Total no.of.Drivers.injured in 2018 without wearing Seatbelt")+xlab("States&Union Territories")+ylab("No.of.Accidents")
```

b

Inference:

This shows the total number of drivers injured in 2018 without wearing seatbelt Here Tamil nadu tops the list which is followed by Madhya Pradesh, Uttar Pradesh, Karnataka, Rajasthan.

```
##barplot for not wearing Seatbelt(Passengers)
```

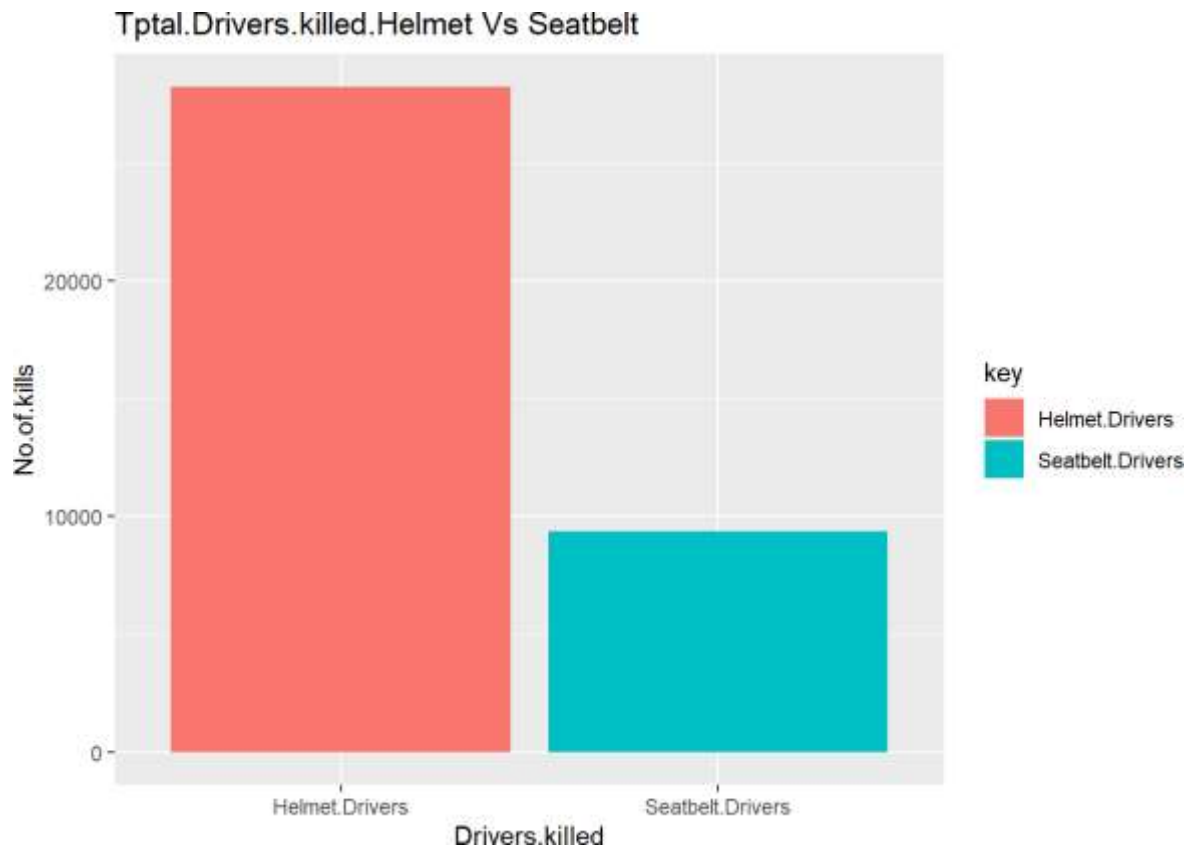
```
c<-ggplot(data1,aes(reorder(x=States.UTs,Passengers...Persons.Injured...Total.Injured),y=Passengers...Persons.Injured...Total.Injured,fill=States.UTs))
+coord_flip()+geom_bar(stat="identity")+ggtitle("Total no.of.Passengers.injured in 2018 without wearing Seatbelt")+xlab("States&Union Territories")+ylab("No.of.Accidents")
```

c


```
##interactive plot  
ggplotly(c)
```

0200040006000ChandigarhLakshadweepDadra & Nagar HaveliSikkimMizoramDaman &

```
# TYPE OF ACCIDENTS  
  
Type_of_Accidents_2018<-read.csv("C:/Users/Tenisha/Desktop/Accidents_dataset/Road-Accidents-2018--type of accidents in 2018.csv")  
  
##barplot  
  
library(ggplot2)  
  
library(plotly)  
  
a<-ggplot(data = tota_accidents1,aes(reorder(x=States.UTs>Total.Accidents),  
y=Total.Accidents))+geom_bar(stat="identity",fill="salmon")+coord_flip()+gg  
title("Total Road Accidents in 2018")+xlab("States & Union Territories")+yl  
ab("Total No.of Road Accidents")  
  
a  
  
  
helmet.killed<-data3[1]  
seatbelt.killed<-d3[1]  
  
Drivers.killed.helmetvseatbelt<-cbind.data.frame(helmet.killed,seatbelt.kil  
led)  
  
colnames(Drivers.killed.helmetvseatbelt)<-c("Helmet.Drivers", "Seatbelt.Driv  
ers")  
  
df<-gather(Drivers.killed.helmetvseatbelt)  
  
ggplot(df,aes(x=key,y=value,fill=key))+geom_bar(stat="identity")+ggtitle("T  
ptal.Drivers.killed.Helmet Vs Seatbelt")+  
xlab("Drivers.killed")+ylab("No.of.kills")
```

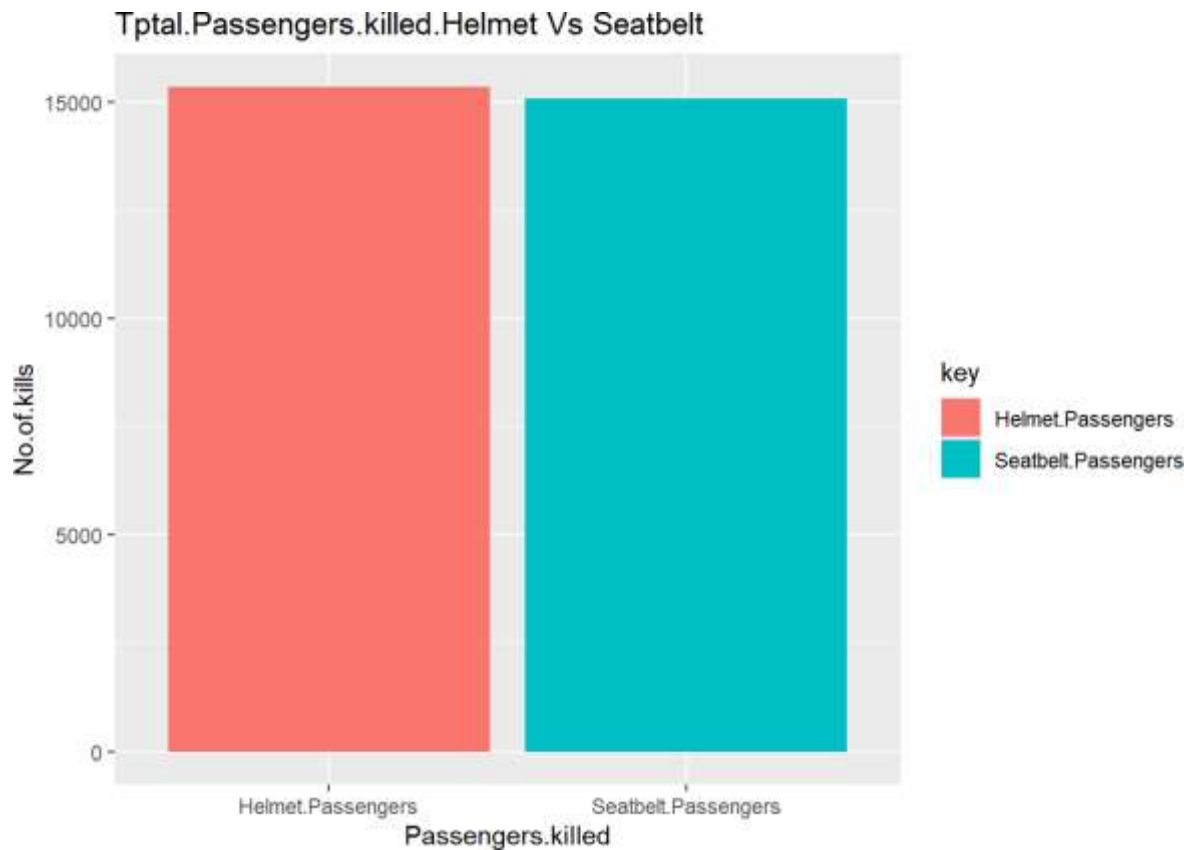


Inference:

This is a comparison between drivers killed without helmet and seatbelt. It is less that the drivers without helmet are killed the most compared to drivers without seatbelt.

```
##total killed Passengers helmet vs Seatbelt
helmet.passengers.killed<-data3[4]
seatbelt.passengers.killed<-d3[4]
Passengers.killed.helmetvsseatbelt<-cbind.data.frame(helmet.passengers.killed,seatbelt.passengers.killed)
colnames(Passengers.killed.helmetvsseatbelt)<-c("Helmet.Passengers","Seatbelt.Passengers")
df1<-gather(Passengers.killed.helmetvsseatbelt)

ggplot(df1,aes(x=key,y=value,fill=key))+geom_bar(stat="identity")+ggtitle("Tptal.Passengers.killed.Helmet Vs Seatbelt")+
  xlab("Passengers.killed")+ylab("No.of.kills")
```

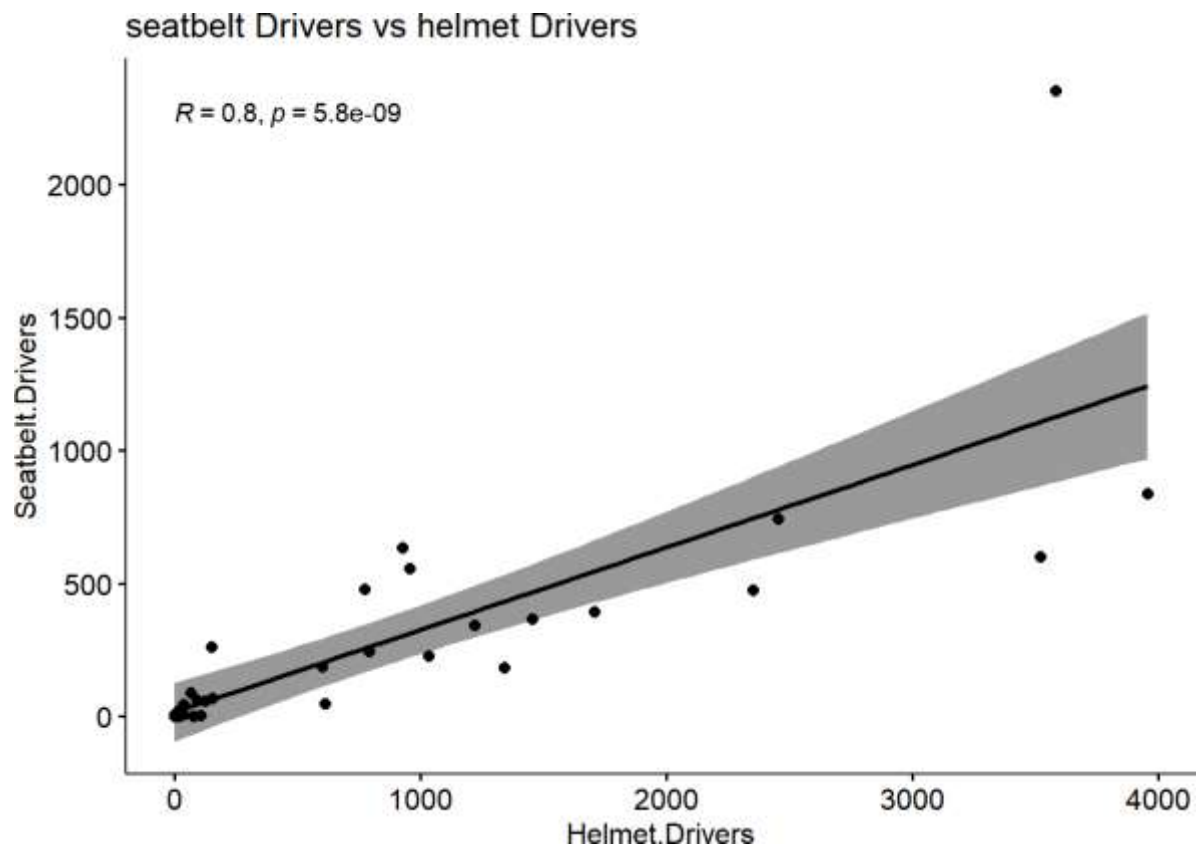


Inference:

Here is the comparison between the total passengers killed without helmet and seatbelt. Passengers without helmet are most likely killed.

```
ggscatter(Drivers.killed.helmetvseatbelt,x="Helmet.Drivers",y="Seatbelt.Drivers",add = "reg.line",conf.int = TRUE,cor.coef=TRUE,method = "pearson")+gg
title("seatbelt Drivers vs helmet Drivers")

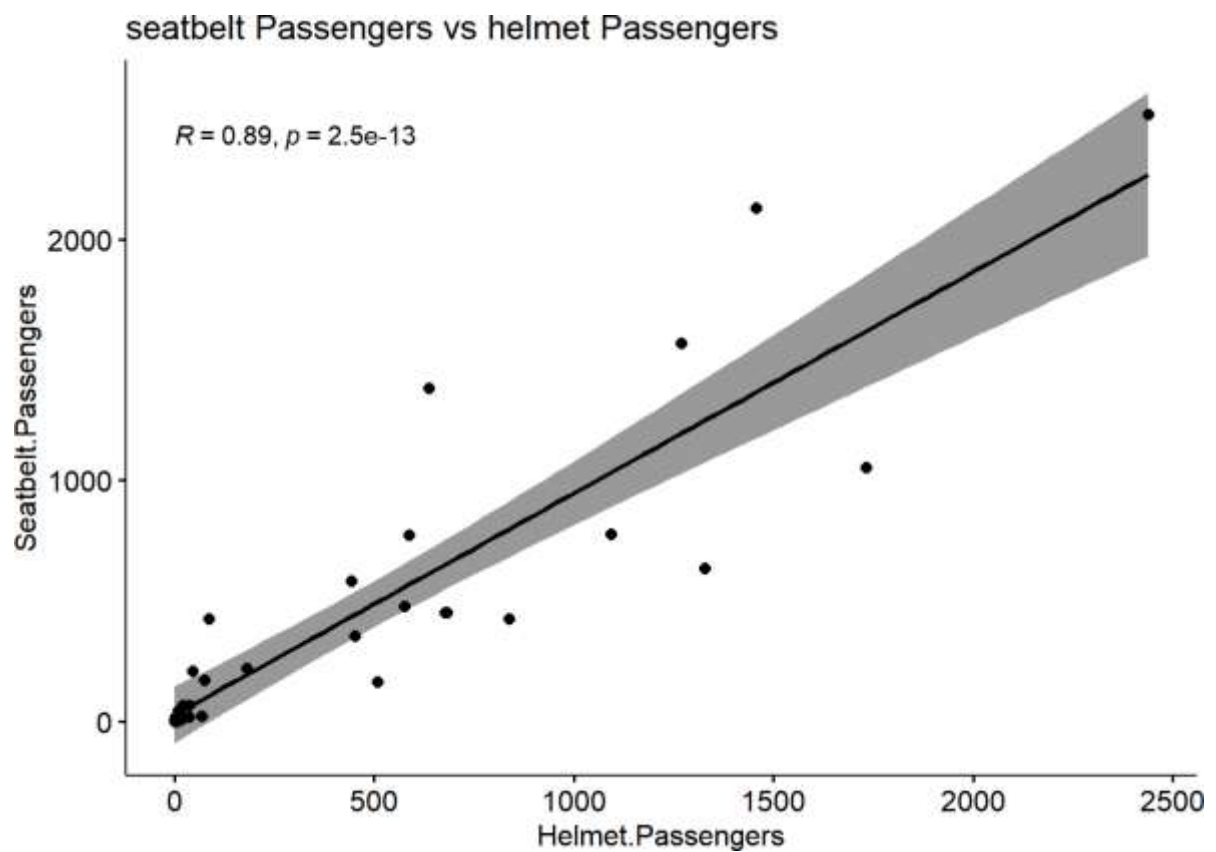
## Warning: Ignoring unknown parameters: method
## `geom_smooth()` using formula 'y ~ x'
```



```
ggscatter(Passengers.killed.helmetvsseatbelt,x="Helmet.Passengers",y="Seatbelt.Passengers",add = "reg.line",conf.int = TRUE,cor.coef=TRUE,method = "pearson")+ggtitle("seatbelt Passengers vs helmet Passengers")
```

```
## Warning: Ignoring unknown parameters: method
```

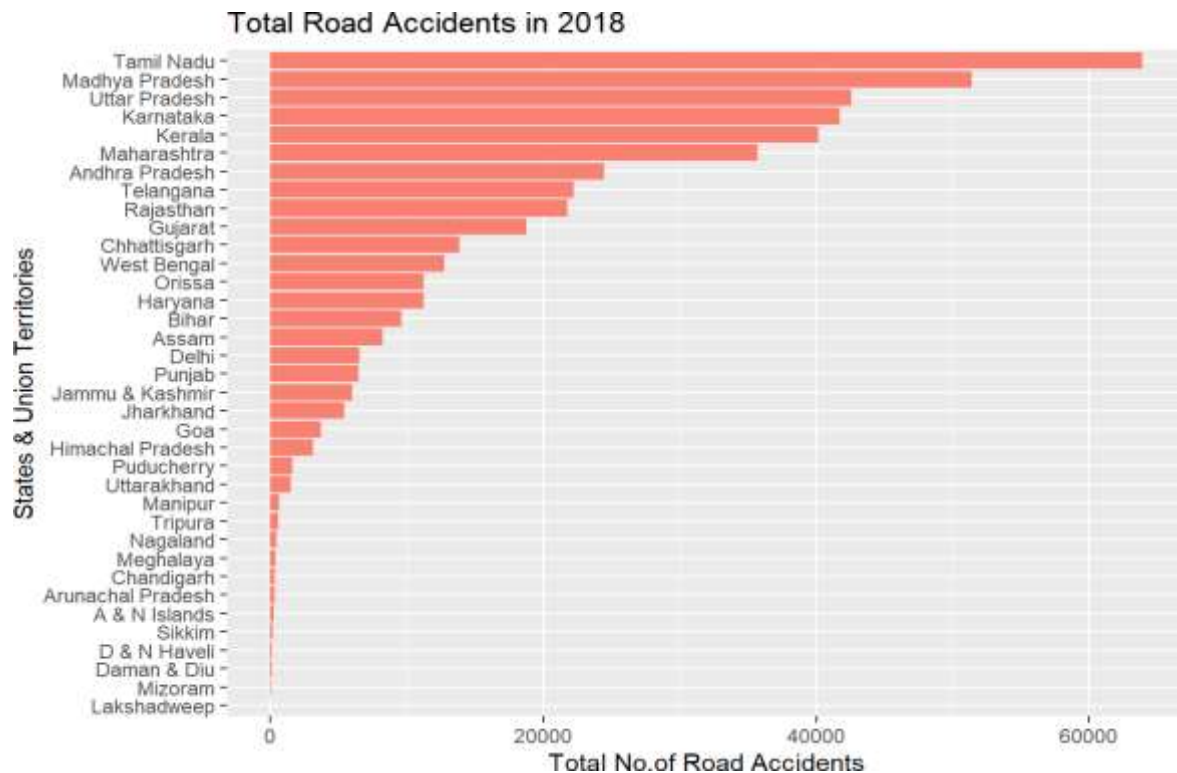
```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(Passengers.killed.helmetvsseatbelt, aes(x=Helmet.Passengers, y=Seatbelt.Passengers)) + ggtitle("seatbelt Passengers vs helmet Passengers") + geom_point() + geom_line()
```

Inference:

The above plot shows the relationship between the Passengers killed in road accident without wearing seatbelt and helmet.



```
##Pie_chart for type of accidents
```

```
piechart<-Type_of_Accidents_2018[37,]
```

```
head(piechart)
```

```
##      S.No States.UTs Fatal.Accidents Grievous.Injury.Accidents
## 37 Total      Total      137726      125311
##      Minor.Injury.Accidents Non.Injury.Accidents Total.Accidents
## 37      169920      34087      467044
```

```
piechart1<-select(piechart,(3:6))
```

```
View(piechart1)
```

```
z<-t(piechart1)
```

```
z
```

```
##      37
## Fatal.Accidents      137726
## Grievous.Injury.Accidents 125311
## Minor.Injury.Accidents      169920
## Non.Injury.Accidents      34087
```

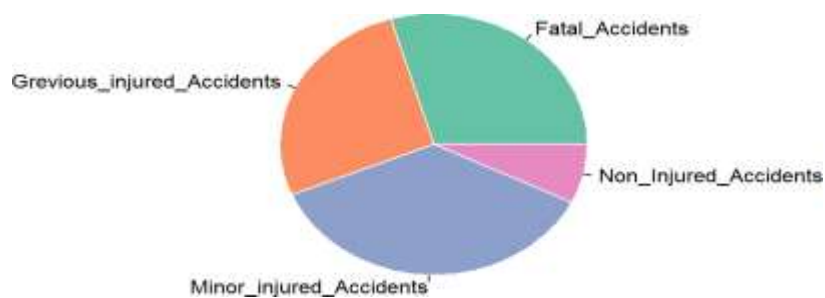
```
df1<-c("Fatal_Accidents","Grievously_injured_Accidents","Minor_injured_Accidents","Non_injury_Accidents")
```

```
df1<-c("Fatal_Accidents","Grievously_injured_Accidents","Minor_injured_Accidents","Non_injury_Accidents")
```

```
df2<-data.frame(z)
```

```
df3<-data.frame(df1,df2$X37)
View(df3)

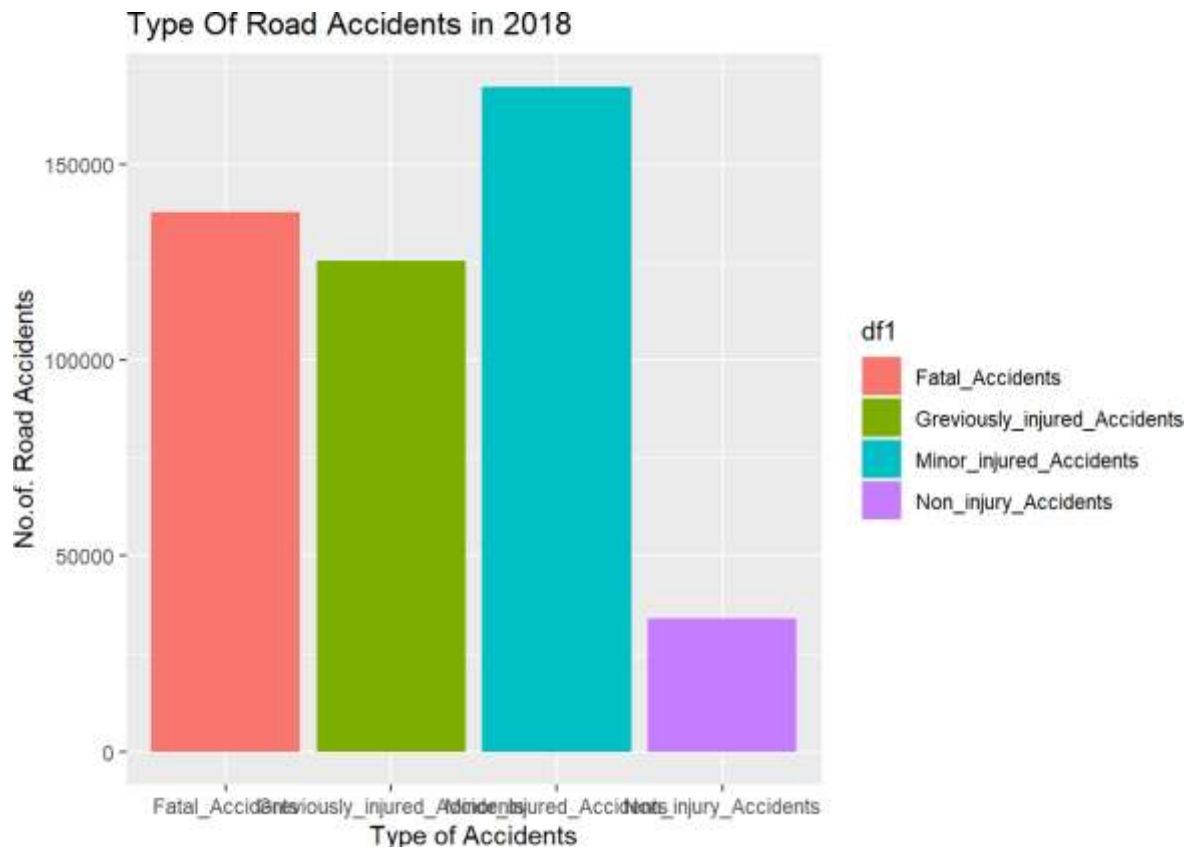
##plotting Pie chart
mypal<-brewer.pal(4,"Set2")
pie(df3$df2.X37,border = "white",col=mypal,labels = c("Fatal_Accidents","Gr
evious_injured_Accidents","Minor_injured_Accidents","Non_Injured_Accidents"
))
```



Inference:

The pie chart shows the various relationship between the parts divided about Total Accidents.

```
#barplot
mycol<-brewer.pal(4,"Set1")
s<-ggplot(df3,aes(x=df1,y=df2.X37,fill=df1))+geom_bar(stat="identity")+ggtitle("Type Of Road Accidents in 2018")+xlab("Type of Accidents")+ylab("No.of . Road Accidents")
s
```



Inference:

The above figure is the Type of Accidents (fatal, Greivous, Minor injure and non-injury accidents).

Accidents with Minor Injury has occurred more in India in 2018.

```
##AGE AND GENDER WISE ROAD ACCIDENT

age_gender_data<-read.csv("C:/Users/Tenisha/Desktop/Accidents_dataset/Road-
Accidents-2018--age_and_gender.csv")

age.gender1<-age_gender_data[37,]
age.gender2<-age.gender1[3:16]
age.gender3<-data.frame(t(age.gender2))

age.gender4<-c("lessthan.18-M","lessthan.18-F","18-25.M","18-25.F","25-35.M",
,"25-35.F","35-45.M","35-45.F","45-60.M","45-60.F","60&above.M","60&above.
F","Age notknown.M","Age notknown.F")

age.gender5<-data.frame(age.gender4,age.gender3$X37)

#barplot

p<-ggplot(age.gender5,aes(reorder(x=age.gender4,age.gender3.X37,fill=age.ge
nder4),y=age.gender3.X37))+geom_bar(stat="identity",fill="orange")+coord_fl
ip()+ggtitle("Age & Gender wise Accidents in 2018")+xlab("Age & Gender")+yl
ab("No Of Accidents")

## Male and Female accidents

##Male
```



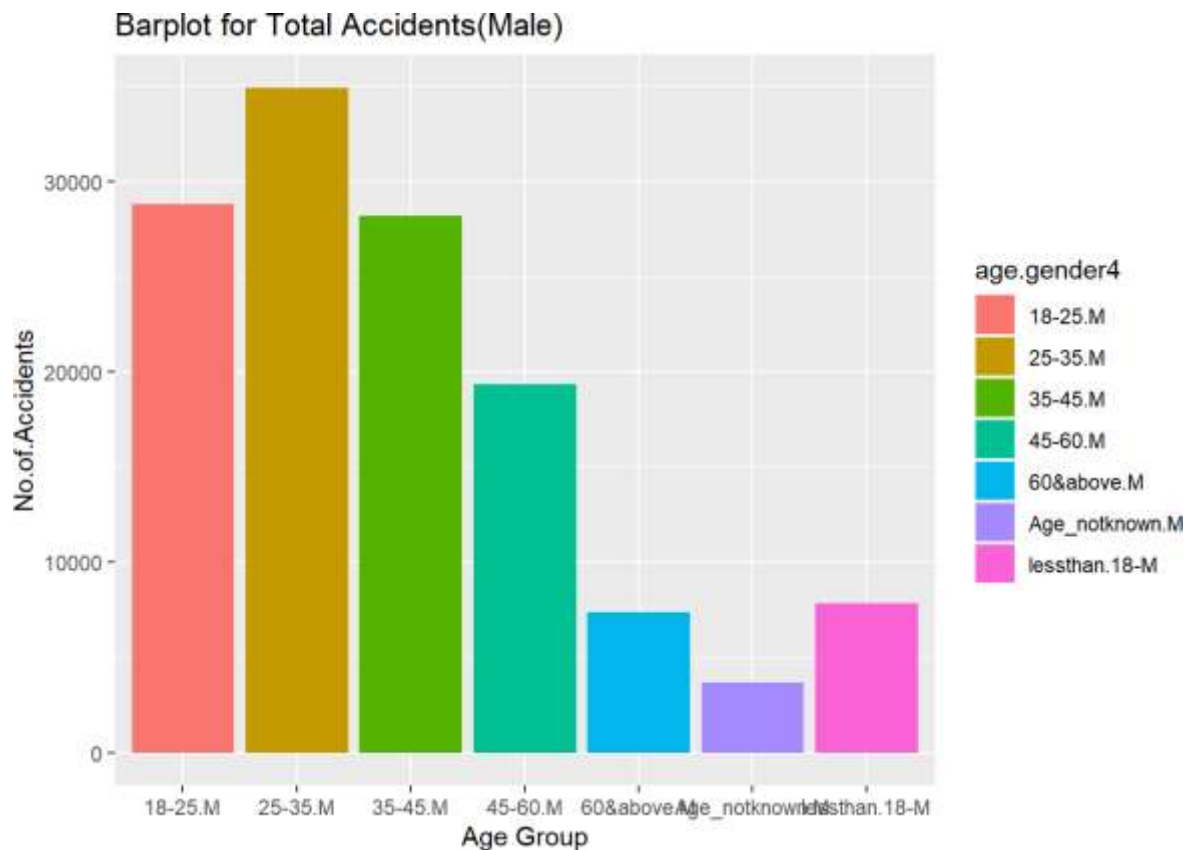
```
male_data<-age.gender5[c(1,3,5,7,9,11,13),]

ggplot(male_data,aes(x=male_data$age.gender4,y=male_data$age.gender3.X37,fill=age.gender4))+geom_bar(stat = "identity")+ggtitle("Barplot for Total Accidents(Male)")+xlab("Age Group")+ylab("No.of.Accidents")

## Warning: Use of `male_data$age.gender4` is discouraged. Use `age.gender4` instead.

## Warning: Use of `male_data$age.gender3.X37` is discouraged. Use `age.gender3.X37` instead.

## instead.
```



Inference:

This Bar plot shows the the no of Accidents (Male)

25 – 35 Male is the age category where more Accidents occurred.

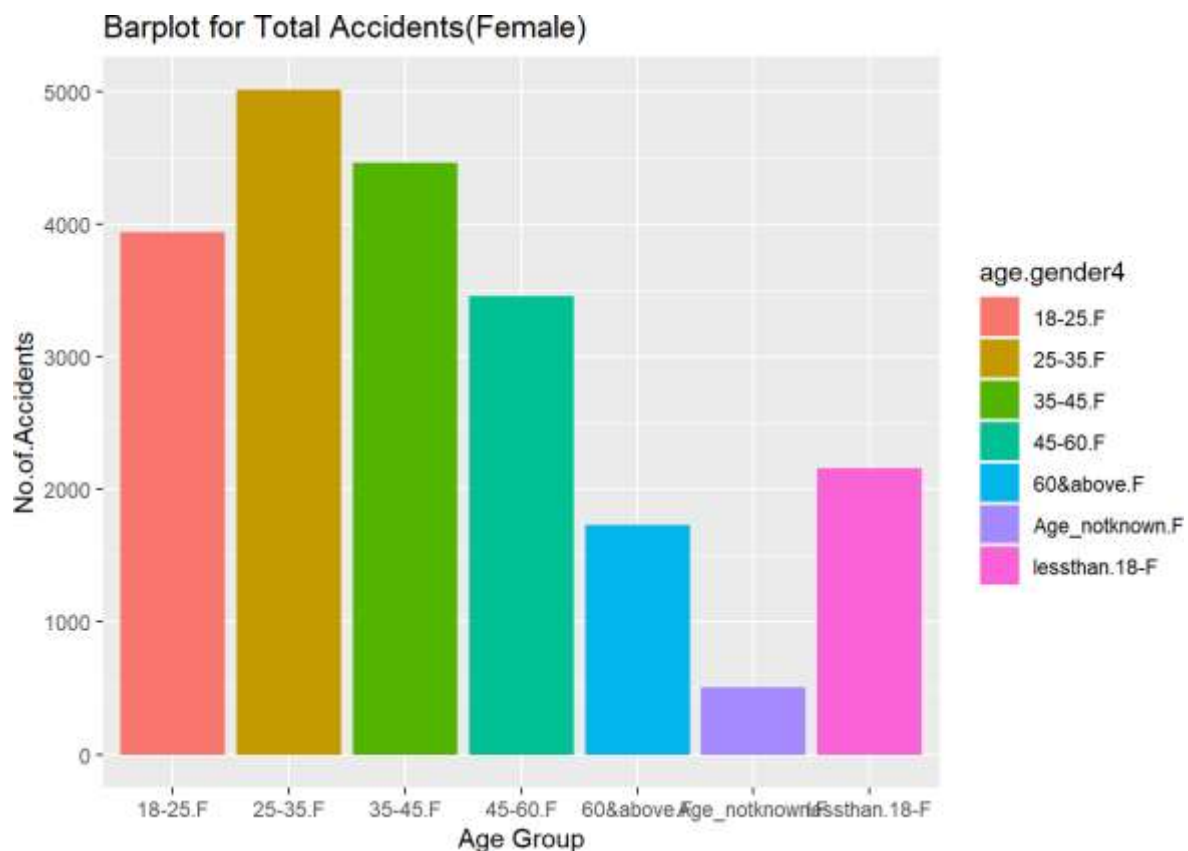
```
##Female

female_data<-age.gender5[c(2,4,6,8,10,12,14),]

ggplot(female_data,aes(x=female_data$age.gender4,y=female_data$age.gender3.X37,fill=age.gender4))+geom_bar(stat = "identity")+ggtitle("Barplot for Total Accidents(Female)")+xlab("Age Group")+ylab("No.of.Accidents")

## Warning: Use of `female_data$age.gender4` is discouraged. Use `age.gende
r4` instead.
```

```
## Warning: Use of `female_data$age.gender3.X37` is discouraged. Use `age.g
ender3.X37`
## instead.
```



Inference:

The above figure is the bar plot for Total Accidents (Female).

25 -35 Female is the age category where more accidents occurred.

#TRAFFIC VIOLATIONS

```
##loading dataset
```

```
traffic_violations<-read.csv("C:/Users/Tenisha/Documents/acci_data/Road-Accidents-2018-Traffic violations.csv")
```

```
View(traffic_violations)
```

```
library(dplyr)
```

```
is.na(traffic_violations)%>%sum()
```

```
## [1] 2
```

```
##removing unwanted rows and columns
```

```
#removing rows which contains total
```

```
data1<-traffic_violations[-37,]
```

```
##checking missing values
```

```
is.na(data1)%>%sum()
```

```
## [1] 0

##subsetting important columns for analysis
View(traffic_violations)
data2_total_accidents<-select(data1,c(2,3,10,15,20,25))
View(data2_total_accidents)%>%head()

## NULL

data3_total_killed<-select(data1,c(2,5,11,16,21,26))
View(data3_total_killed)

##boxplot

library(reshape2)
data5<-data2_total_accidents[-1]
colnames(data5)<-c("Over.speeding","Drunken&Drug.Driving","Driving.wrongside",
"jumping.redlight","Use.of.mobile")
colnames(data5)

## [1] "Over.speeding"          "Drunken&Drug.Driving" "Driving.wrongside"
## [4] "jumping.redlight"        "Use.of.mobile"

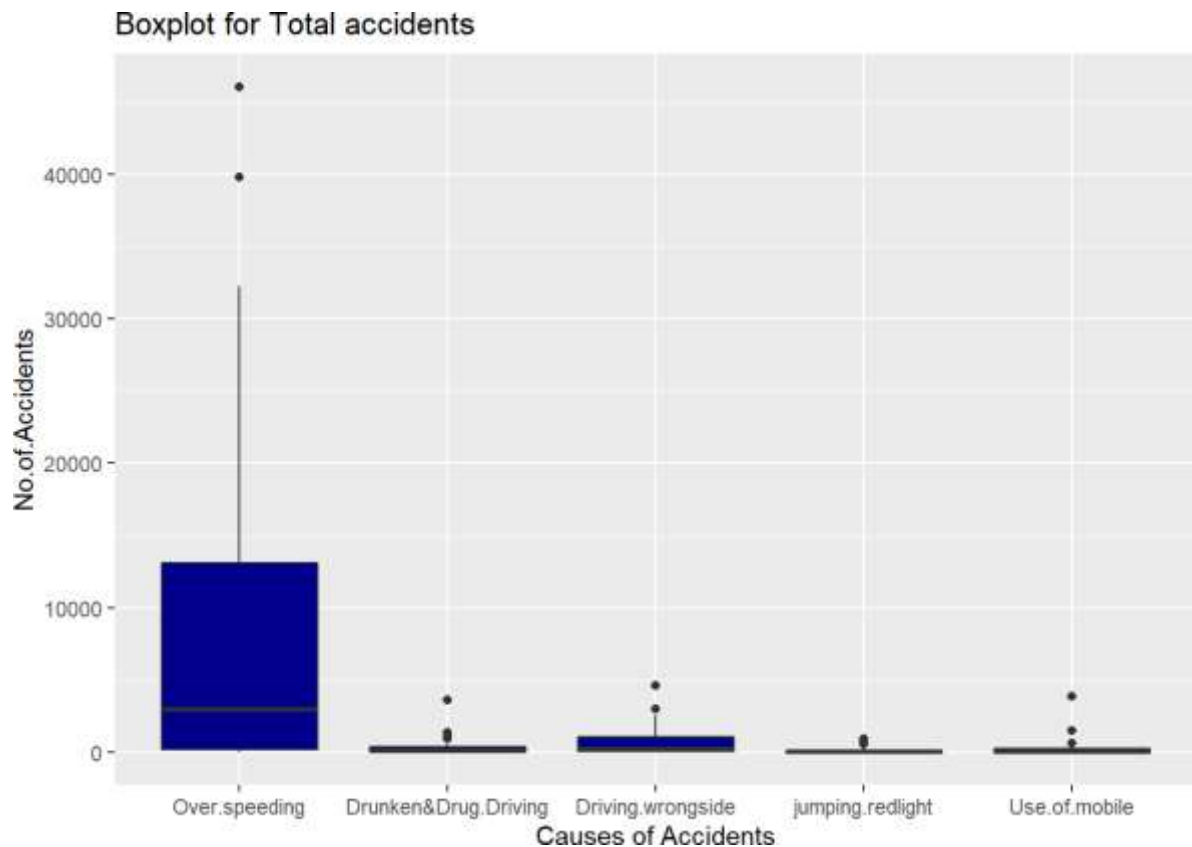
data6<-melt(data5)

## No id variables; using all as measure variables

a<-ggplot(data6,aes(x=variable,y=value))+geom_boxplot(fill = "dark blue")+g
gtitle("Boxplot for Total accidents")

a

a+xlab("Causes of Accidents")+ylab("No.of.Accidents")
```



Inference:

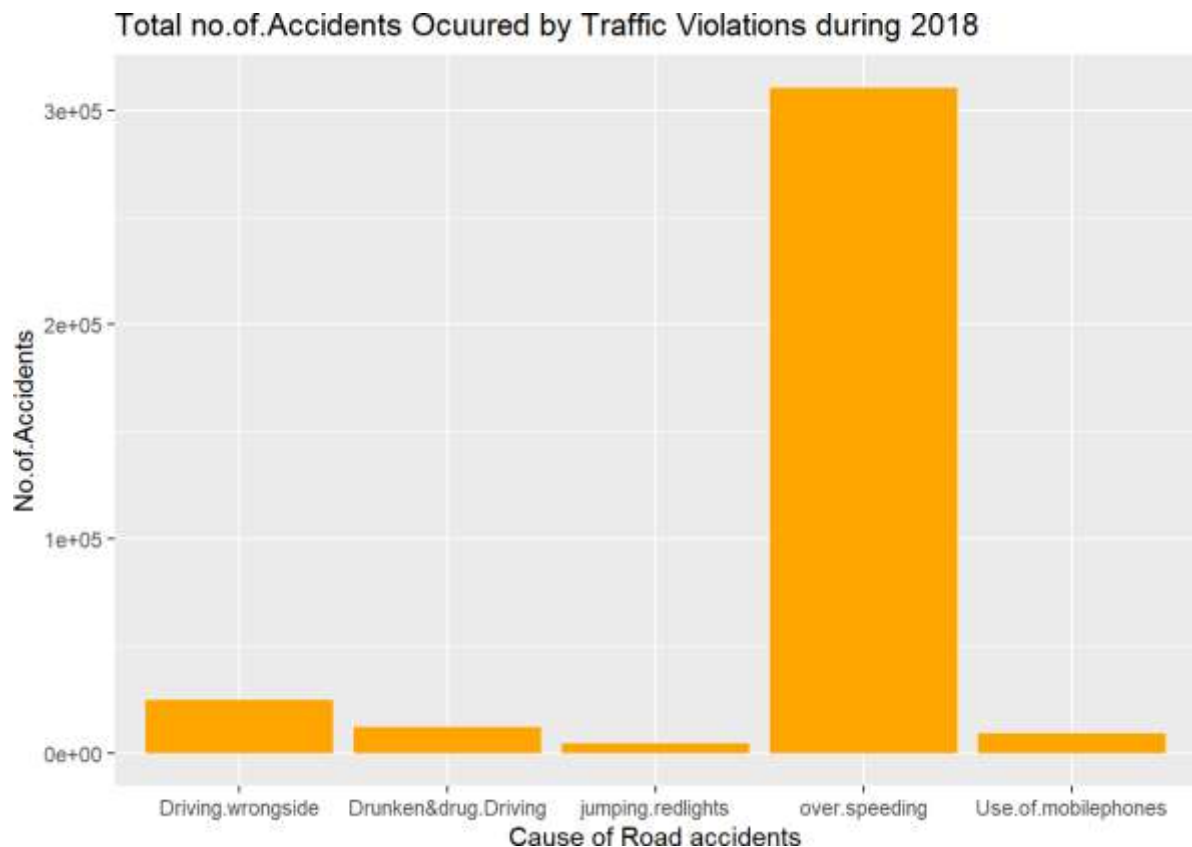
The above figure is the Boxplot for Total Accidents (Fatal, Grievously and minor injured) for the year 2018.

```
##point plot
b<-ggplot(data6,aes(x=variable,y=value))+geom_point()
b
```

```
##barplot for total accidents occurred by violating traffic signals
data7<-traffic_violations[37,]
##subsetting data
total_accidents<-select(data7,c(3,10,15,20,25))
View(total_accidents)
#transposing the dataset
total_accident1<-t(total_accidents)
df1<-c("over.speeding","Drunken&drug.Driving","Driving.wrongside","jumping.
redlights","Use.of.mobilephones")
df2<-data.frame(total_accident1)
df3<-data.frame(df1,df2$X37)
```

```
a<-ggplot(df3,aes(x=df1,y=df2.X37))+geom_bar(stat = "identity",fill="orange")
+labs(x="Type of Cause",y = "No. Of Accidents Occured")

a+ggtitle("Total no.of.Accidents Occured by Traffic Violations during 2018")
+xlabs("Cause of Road accidents")+ylab("No.of.Accidents")
```



Inference:

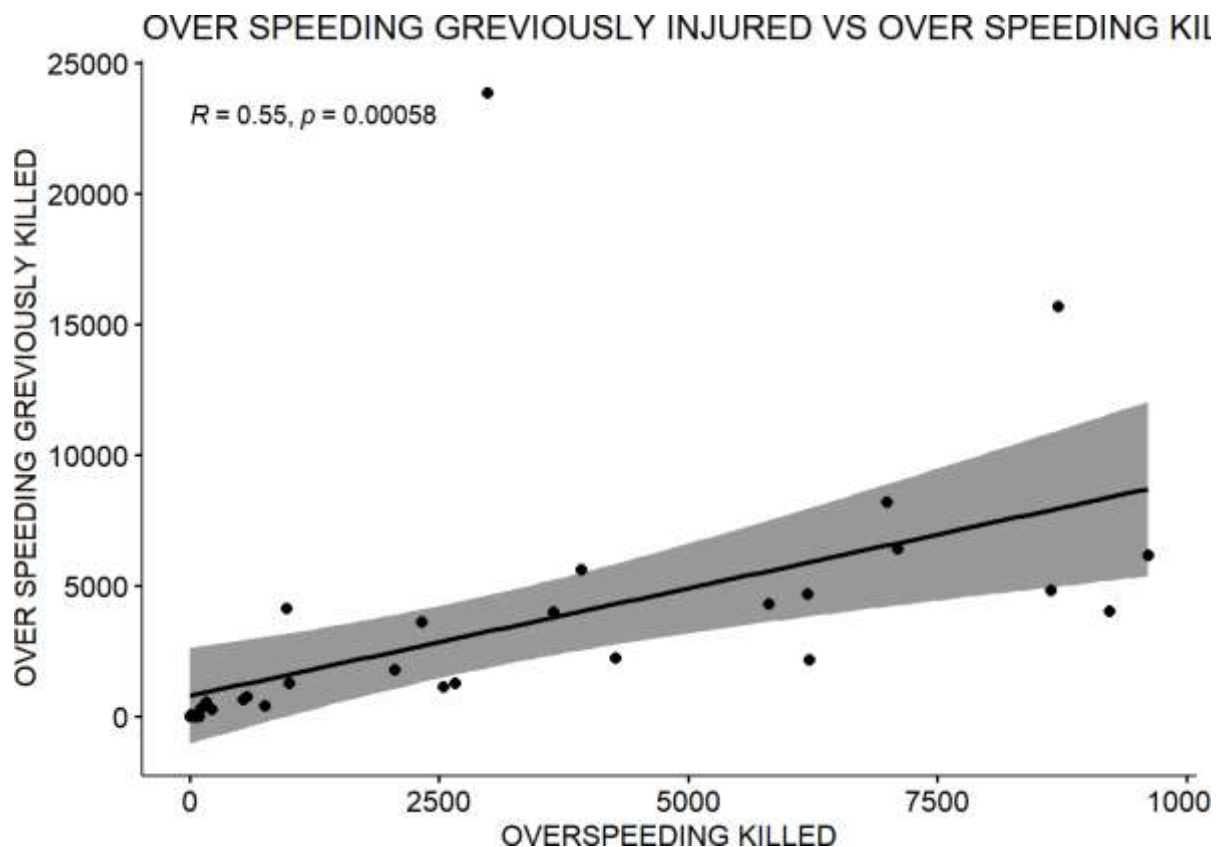
The above bar plot shows the Total no of Accidents Occurred because of Traffic Violations.

```
##scatterplot

#Overspeedng killed Vs Injured

ggscatter(datal,x="Over.Speeding...Persons.Killed...Number",y="Over.Speeding...Persons.Injured...Greviously.Injured",add = "reg.line",conf.int = TRUE,cor.coef = TRUE,cor.method = "pearson")+ggtitle("OVER SPEEDING GREVIOUSLY INJURED VS OVER SPEEDING KILLED")+xlab("OVERSPEEDING KILLED")+ylab("OVER SPEEDING GREVIOUSLY KILLED")

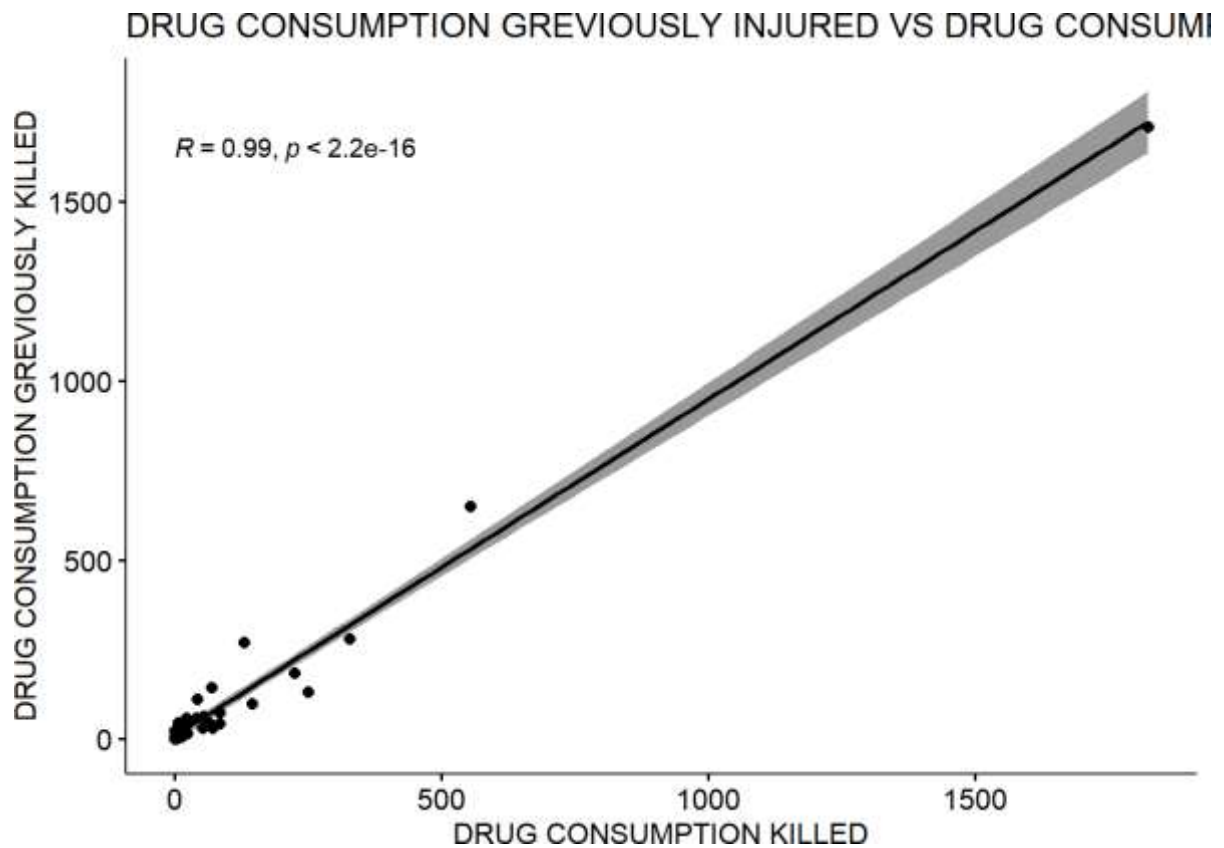
## `geom_smooth()` using formula 'y ~ x'
```



```
#Drug consumption killed Vs Greviously killed
```

```
ggscatter(data1,x="Drunken.Driving.Consumption.of.Alcohol...Drug...Persons.
Killed",y="Drunken.Driving.Consumption.of.Alcohol...Drug...Persons.Injured.
..Greviously.Injured",add = "reg.line",conf.int = TRUE,cor.coef = TRUE,cor.
method = "pearson")+ggtitle("DRUG CONSUMPTION GREVIOUSLY INJURED VS DRUG CO
NSUMPTION KILLED")+xlab("DRUG CONSUMPTION KILLED")+ylab("DRUG CONSUMPTION G
REVIOUSLY KILLED")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
### barplot for no.of.persons killed
```

```
##subsetting data
```

```
##barplot for total accidents occured by violating traffic signals
```

```
data7<-traffic_violations[37,]
```

```
data7
```

```
]
```

7878

```
##subsetting data
```

```
total_accidents<-select(data7,c(3,10,15,20,25))
```

```
View(total_accidents)
```

```
#transposing the dataset
```

```
total_accident1<-t(total_accidents)
```

```
df1<-c("over_speeding","Drunken.driving.drug_consumption","Driving.wrongside",  
"jumping.red.lights","Use.of.mobile.phones")
```

```
df2<-data.frame(total_accident1)
```

```
df3<-data.frame(df1,df2$X37)
```

```
a<-ggplot(df3,aes(x=df1,y=df2.X37))+geom_bar(stat = "identity")+labs(x="Type  
of Cause",y = "No. Of Accidents Occured")
```

a

```
a+ggtitle("Total no.of.Accidents Ocuured by Traffic Violations during 2018")
)+xlab("Cause of Road accidents")+ylab("No.of.Accidents")
```

###NATIONAL & STATE HIGHWAYS

```
NH_killed<-read.csv("C:/Users/Tenisha/Documents/acci_data/national-highways-
killed.csv")
```

```
SH_killed<-read.csv("C:/Users/Tenisha/Documents/acci_data/statehighways-kil
led.csv")
```

##SLICING DATA

```
NH_killed.persons<-NH_killed[c(2,6)]
```

```
SH_killed.persons<-SH_killed[6]
```

##COMBINING DATA

##NH&SH

```
combined.SH.NH.killed<-cbind(NH_killed.persons,SH_killed.persons)
```

```
colnames(combined.SH.NH.killed)<-c("STATES/UTs","NH.KILLED","SH.KILLED")
```

```
combined.SH.NH.killed<-combined.SH.NH.killed[-37,]
```

##histogram for no of people killed in national highways

```
par(mfrow=c(2,1))
```

```
par(mar=rep(2,4))
```

```
hist(combined.SH.NH.killed$NH.KILLED,col = "green",breaks = 30,main = "Hist
ogram for No.of.People.killed In National Highways")
```

```
abline(v=mean(combined.SH.NH.killed$NH.KILLED),col="red",lwd=3)
```

```
abline(v=median(combined.SH.NH.killed$NH.KILLED),col="yellow",lwd=3)
```

```
legend(x="topright",c("Density","mean","median"),col=c("green","red","yello
w"),cex=0.75,lwd=c(3,3,3))
```

##histogram for no of people killed in State Highways

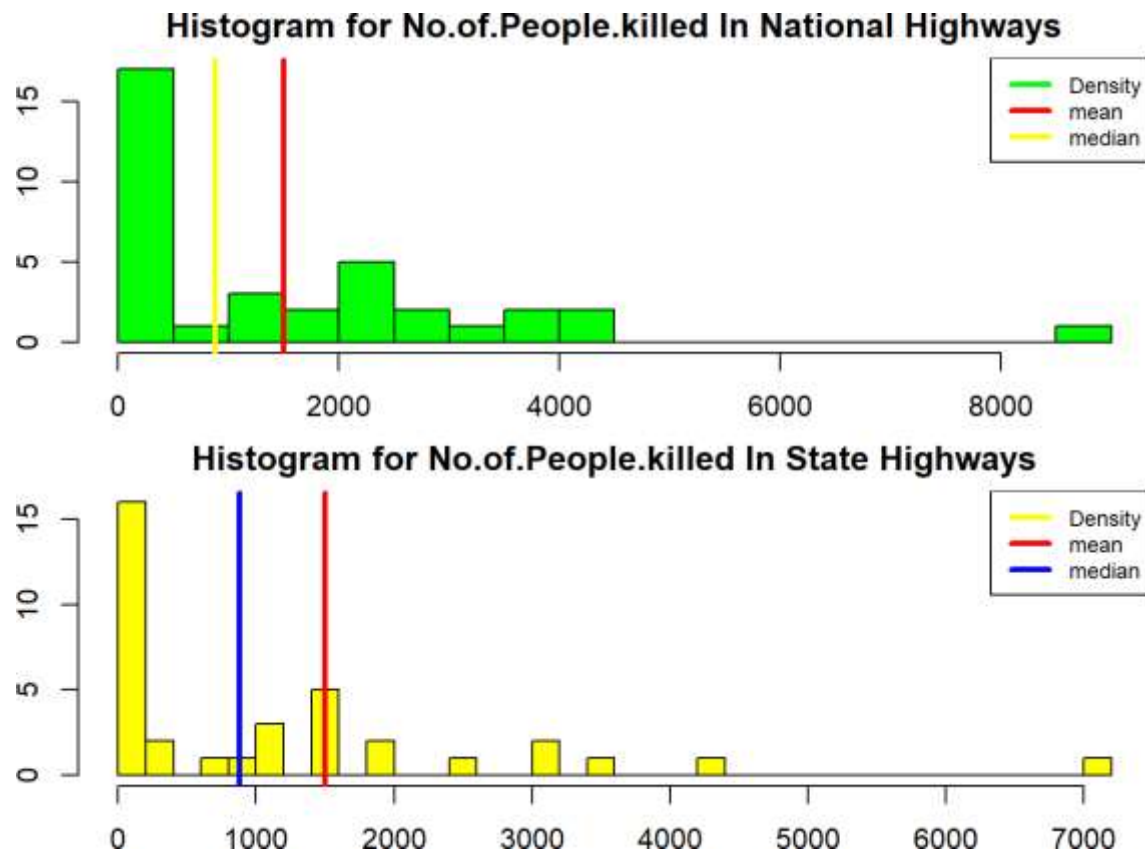
```
hist(combined.SH.NH.killed$SH.KILLED,col = "yellow",breaks = 30,main = "His
togram for No.of.People.killed In State Highways")
```

```
abline(v=mean(combined.SH.NH.killed$NH.KILLED),col="red",lwd=3)
```

```
abline(v=median(combined.SH.NH.killed$NH.KILLED),col="blue",lwd=3)
```



```
legend(x="topright",c("Density","mean","median"),col=c("yellow","red","blue"),cex=0.75,lwd=c(3,3,3))
```



Inference:

The above figure show the histogram of total people killed in State highways and Nations Highways in 2018.

Both the Histogram are Positively skewed or right skewed which means most values are clustered around the left tail.

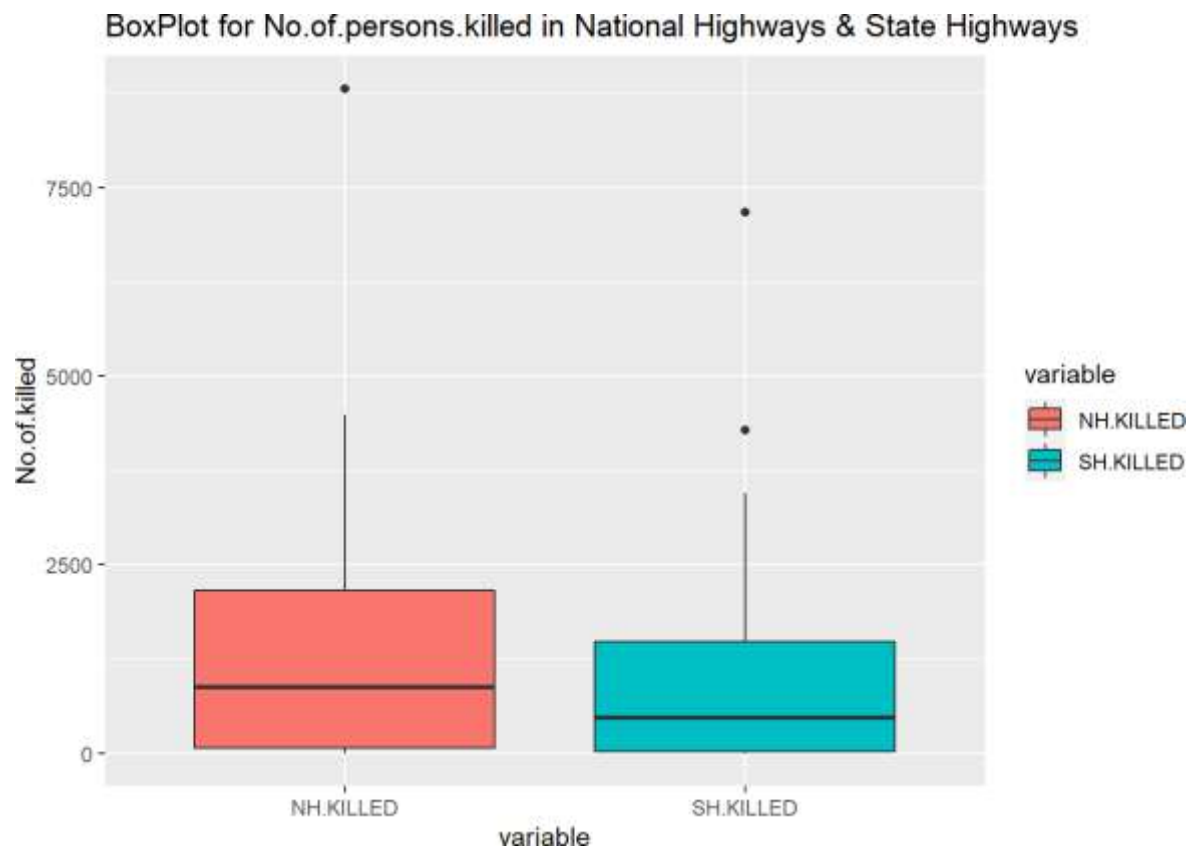
In the Histogram of people killed in National Highways, the value of mean is higher than median

In the Histogram of people killed in State Highways, the value of mean is higher than median

```
##Boxplot
library(reshape2)
View(combined.SH.NH.killed)
df<-combined.SH.NH.killed[-1]
df1<-melt(df)
## No id variables; using all as measure variables
```

```
a<-ggplot(dfl,aes(x=variable,y=value,fill=variable))+geom_boxplot()+ggtitle
("BoxPlot for No.of.persons.killed in National Highways & State Highways")+
ylab("No.of.killed")
```

a



Inference:

This is the Boxplot for Total no of people killed in National Highways and State Highways.

From the above observation it is clear that, the boxplot of State Highways is more consistent than National Highways.

```
##barplot

df2<-gather(df)

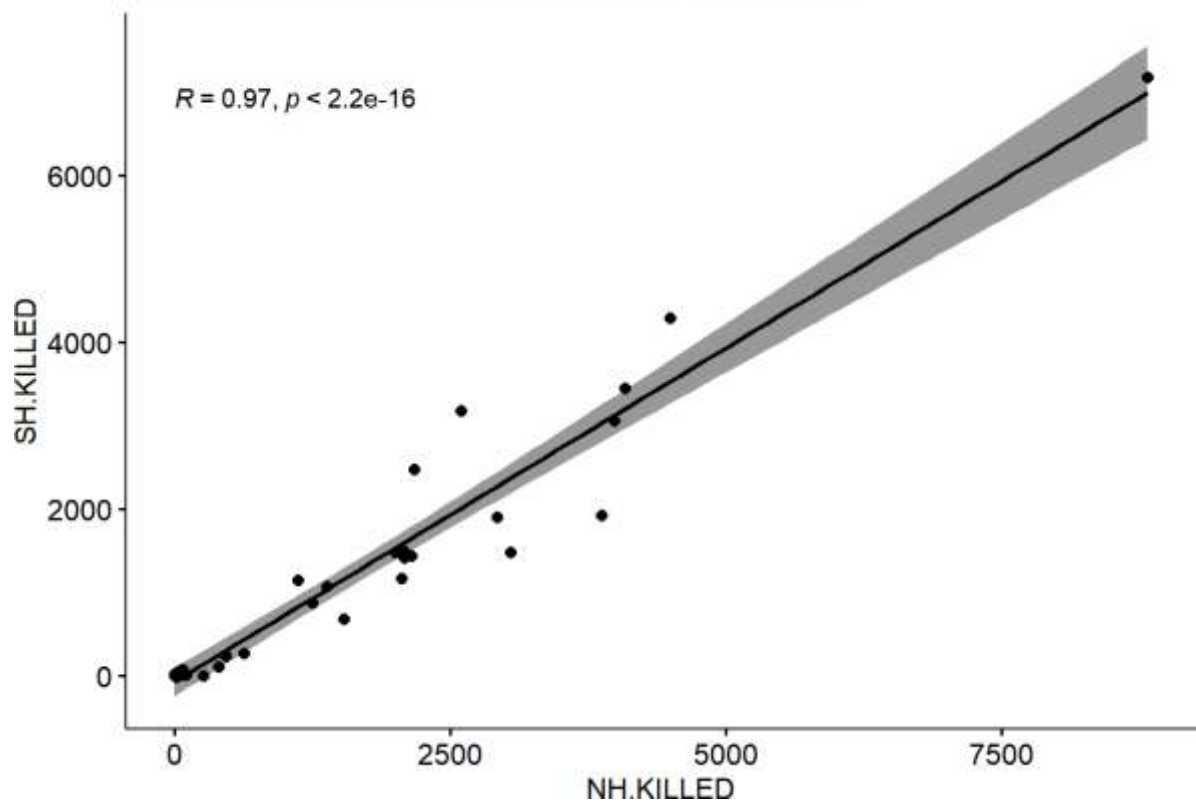
ggplot(df2,aes(x=df2$key,y=df2$value,fill=key))+geom_bar(stat="identity")+g
gtitle("Barplot for comparing NH VS SH - Killed")

## Warning: Use of `df2$key` is discouraged. Use `key` instead.
## Warning: Use of `df2$value` is discouraged. Use `value` instead.

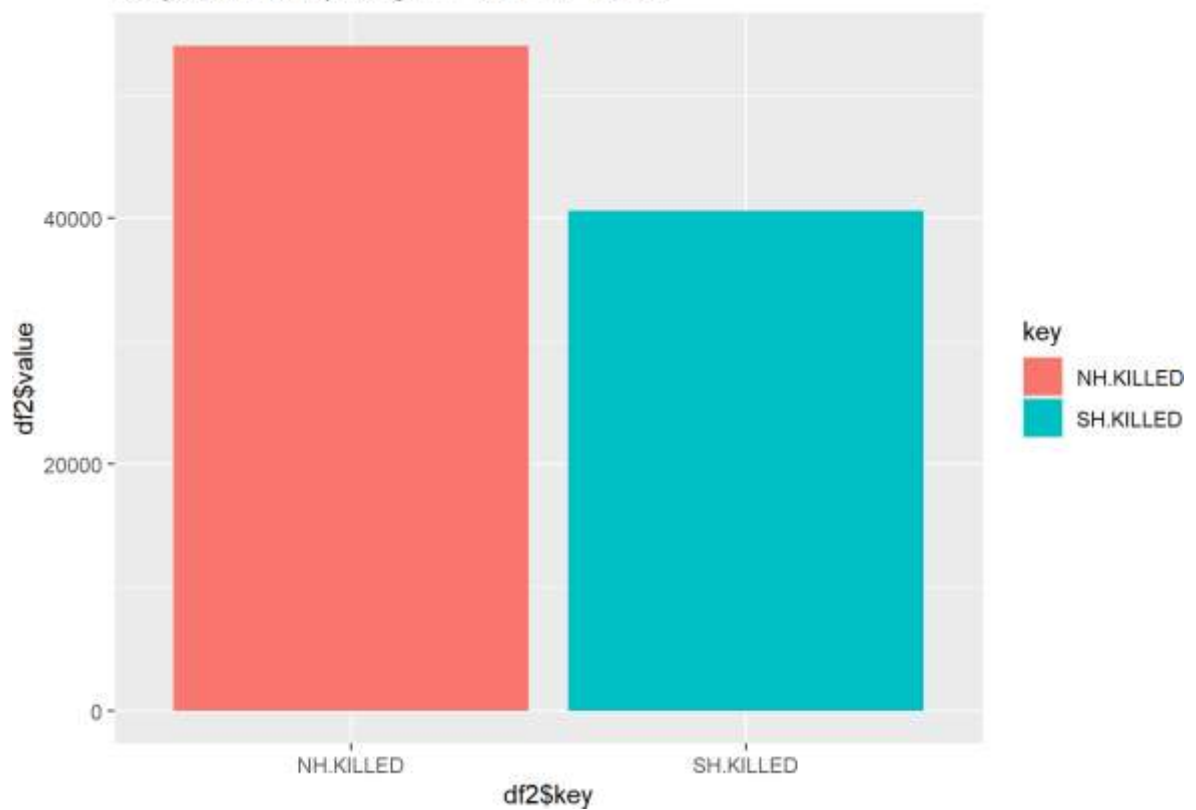
ggscatter(df,x="NH.KILLED",y="SH.KILLED",add = "reg.line",conf.int = TRUE,c
or.coef = TRUE,cor.method = "pearson")+ggtitle("NATIONAL HIGHWAYS VS STATE
HIGHWAYS")

## `geom_smooth()` using formula 'y ~ x'
```

NATIONAL HIGHWAYS VS STATE HIGHWAYS



Barplot for comparing NH VS SH - Killed



...

Inference:

The above bar plot clearly shows the difference between the number of people killed in National highways and State Highways.

Total no of people killed in National Highways is more than State Highways.

INSIGHTS:

1. **Tamil Nadu, Madhya Pradesh, Karnataka, Uttar Pradesh and Andhra Pradesh** are top five states with large number of **Drivers injured without wearing helmet** in 2018
2. **Madhya Pradesh, Tamil Nadu, Kerala, Karnataka, Maharashtra** are the top five states with large number of **passengers injured without wearing helmet** in 2018.
3. By analysing the data, Without wearing helmet, Drivers Minor Injury Accidents occurred more in 2018
4. Less no of passengers without wearing helmet killed in 2018 comparing Drivers and Passengers data.
5. The histogram analysis of Accidents occurred without helmet (both drivers and passengers) is all positively skewed
6. **Tamil Nadu, Madhya Pradesh, Uttar Pradesh, Karnataka and Rajasthan** are top five states with large number of **Drivers injured without wearing Seatbelt** in 2018.
7. **Madhya Pradesh, Tamil Nadu, Uttar Pradesh, Karnataka and Maharashtra** are top five states with large number of **Passengers injured without wearing Seatbelt** in 2018.
8. By comparing Drivers killed (both helmet & seatbelt), Drivers without wearing helmet are killed more than Drivers killed without wearing seatbelt.
9. By comparing Passengers killed (both helmet and Seatbelt), there is not much difference between the no of passengers killed.
10. By analysing Total Road Accidents (injured & Killed), **Tamil Nadu, Kerala, Madhya Pradesh, Uttar Pradesh and Karnataka** are the top five states with large no of people affected by road accidents.

11. In 2018, there are more no of Minor Injured Accidents that occurred frequently. Grievously Injured accidents and fatal accidents are also more in 2018.
12. Non – Injured Accidents occurred very less in 2018.
13. **25 – 35 years Male age groups** are more affected by accidents in 2018
14. **60 & above years, less than 18 years** and **age unknown** are the Victim's age categories that are less affected by road accidents in 2018.
15. **25 – 35 years Female age groups** are more affected by road accidents in 2018.
16. **60 & above years** and **Age unknown** are Female age groups that are less affected by road Accidents in 2018.
17. **Over speeding** had been the serious traffic violation in 2018. As many people are injured and killed due to Over speeding.
18. More People are killed in National Highways compared to State highways.
19. According to our analysis, many accidents are occurred in the month of May which is 9.49 %. And less accidents are occurred in the month of August which is 7.3 %.