# EDA ANALYSIS FOR ROAD ACCIDENTS DATA IN 2018

## Road Accidents Data in 2018:

*"In life, more than in anything else, it isn't easy to end up alive"*

**India is one of the developing business hubs in the world where it nerve is the connectivity between the nodes. The connecting nerves are the national highways, the State highways, the International highways and the local roads. The roads are not only used for the development of the country but also it's a place that takes the life of many people. In 2019 alone, the country reported over 151 thousand fatalities due to road accidents. So, we did the Exploratory Data Analysis on the Road Accidents Data in the year 2018 and drawn some insights and inference which can reduce the number of road Accidents.**

*Main Objective:*

**To reduce the number of the Road Accidents in India**

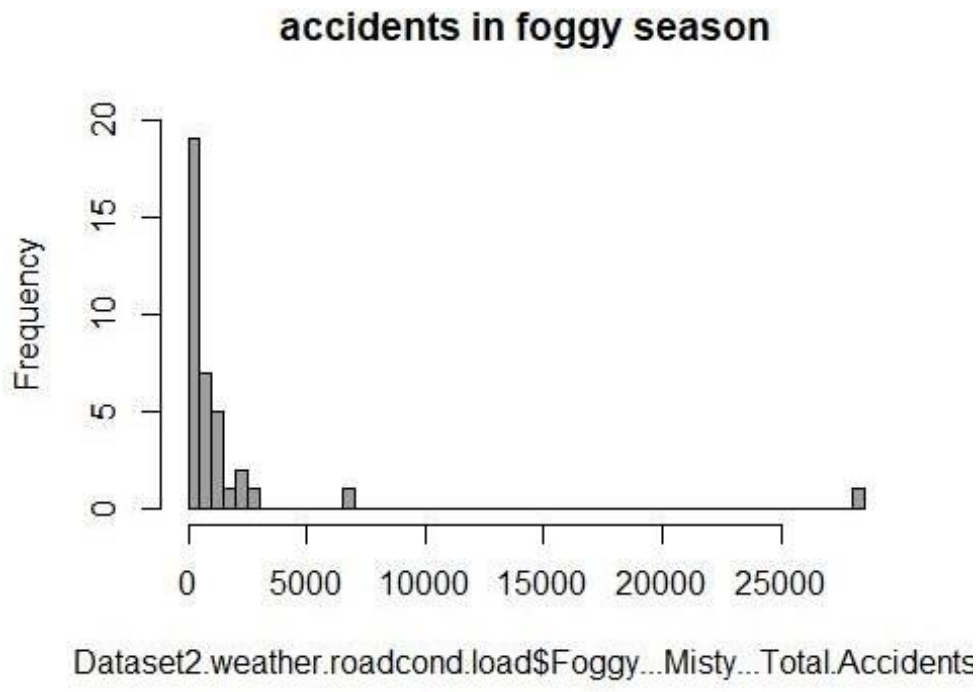
**The data is been collected from the government portal data.gov.in. in the year 2018. The list of 15 datasets been collected which comprises of the following things,**

- **Number of Deaths, Injured (Grievous Injuries & Minor Injuries)**
- **Number of deaths based on Age Categories (From less than 5 years to above 60 years).**
- **Number of Deaths based on the Sex.**
- **Non-wearing the Safety Measures.**
- **Road Conditions**
- **Weather Conditions**
- **Type of Vehicles**
- **Loaded Vehicles**
- **Licence details**
- **Time of Occurrence of Accidents**
- **Type of Accidents**
- **Type of Traffic Violations**

**The Slicing and merging of data are done in the analysis wherever necessary. The final dashboard has been created and attached to the report.**

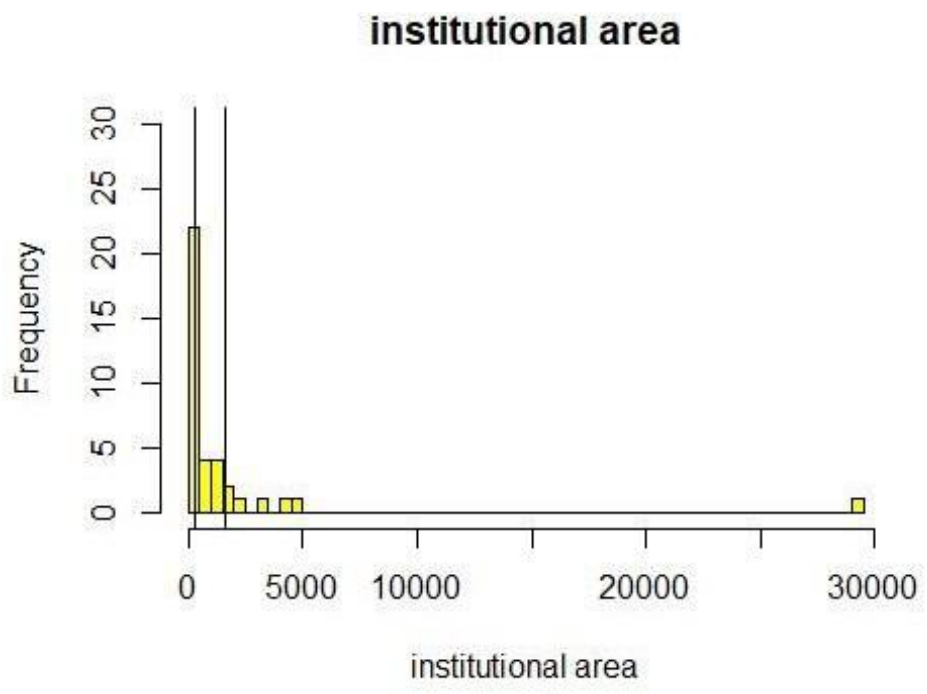
## Histogram Analysis:

```
Dataset2.weather.roadcond.load <- read.csv("C:/Users/kaviya
subramanian/Deskt op/Dataset2-weather-roadcond-load.csv")
View(Dataset2.weather.roadcond.load)

hist(Dataset2.weather.roadcond.load$Foggy...Misty...Total.Accidents,col =
8,y lim = c(0,20),main ='accidents in foggy season',breaks=50)
```




**Inference:** Total accidents occurred more than 25000, but highest frequency of accidents occurred between 0-2500 in foggy season.

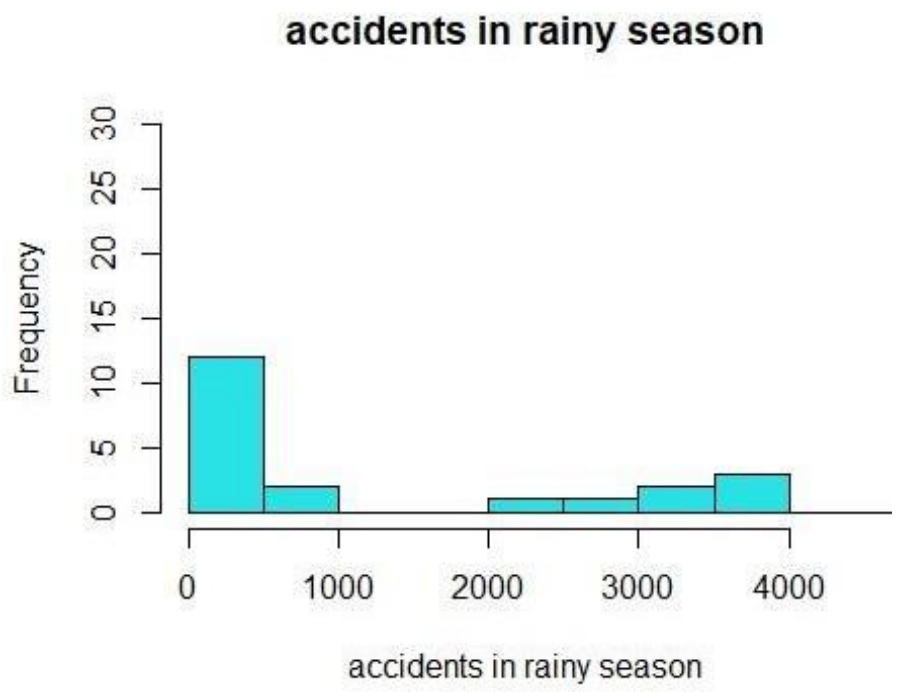
```
hist(Dataset2.weather.roadcond.load$Institutional.Area...Total.Accidents,b
rea ks = 75,ylim=c(0,30),col="yellow",main="institutional
area",xlab="institution al area")
abline(v=mean(Dataset2.weather.roadcond.load$Institutional.Area...Total.Ac
cid ents))
abline(v=median(Dataset2.weather.roadcond.load$Institutional.Area...Total.
Acc idents))
```




**Inference:** The mean and median of the institutional area are shown in the plot. The mean is 1586 and median is 342. Most number of frequencies are less than the mean value. The highest number of accidents are in range (1000-2000).

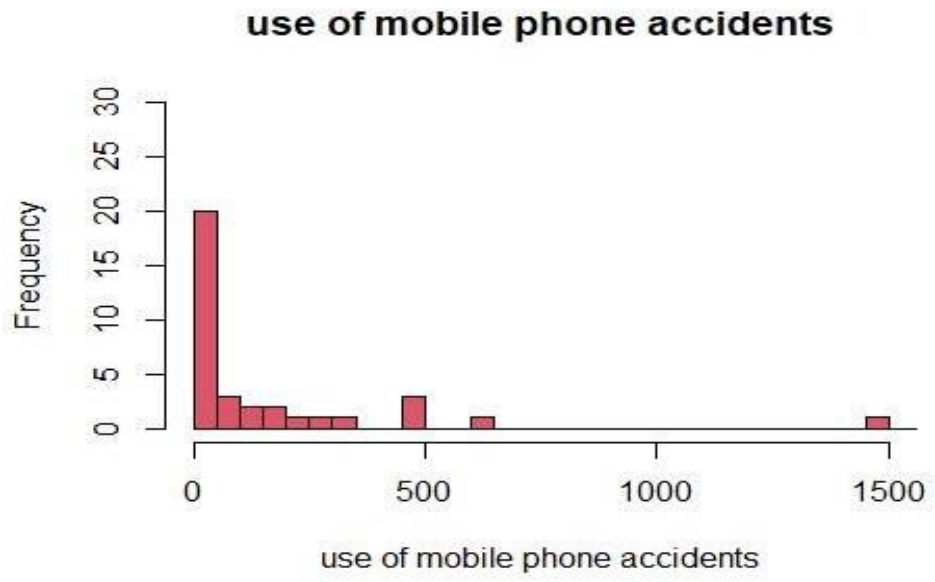
```
 dataset4.typeofaccident.license.traffic.violation <-
read.csv("C:/Users/kavi ya subramanian/Desktop/dataset4-typeofaccident-
license-traffic-violation.csv"
)
 View(dataset4.typeofaccident.license.traffic.violation)
```

```
hist(Dataset2.weather.roadcond.load$Sunny.Clear...Total.Accidents...Number
,br eaks = 500,col=5,xlab = "accidents in rainy season",ylim =
c(0,30),xlim=c(0,4 500),main="accidents in rainy season")
```



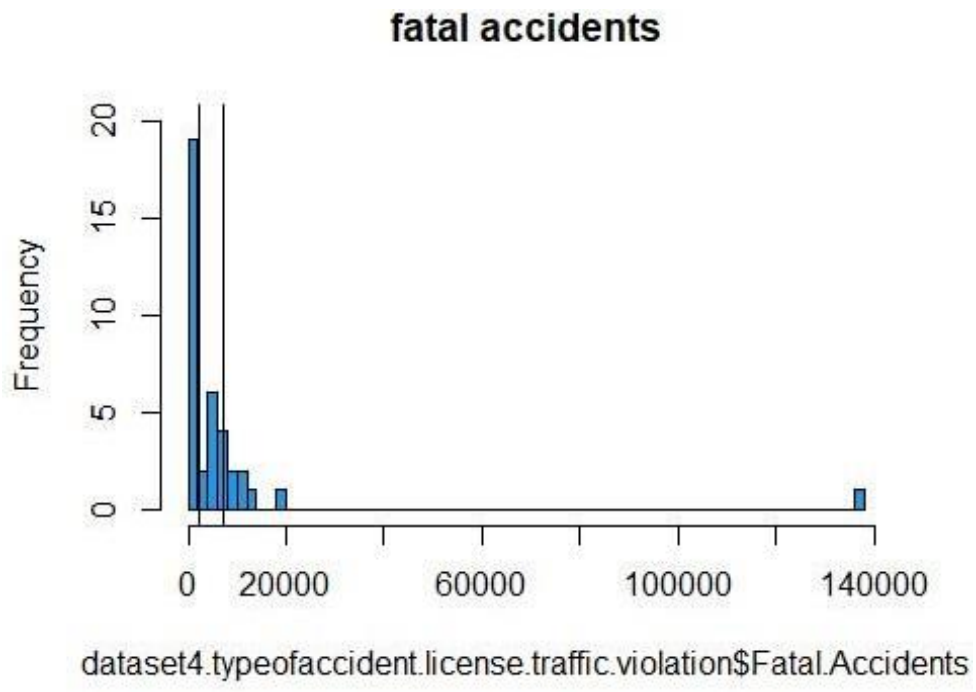

**Inference:** The accident occurred in rainy season in all states its takes from 0-4000 persons killed. Here is the graph,major persons died are range of 2000-4000.

```
hist(dataset4.typeofaccident.license.traffic.violation$Use.of.Mobile.Phone
... Number.of.Accidents, ylim =
c(0,30),xlim=c(0,1500),breaks=150,col=10,border = "black",xlab="use of
mobile phone accidents",main="use of mobile phone accidents")
```



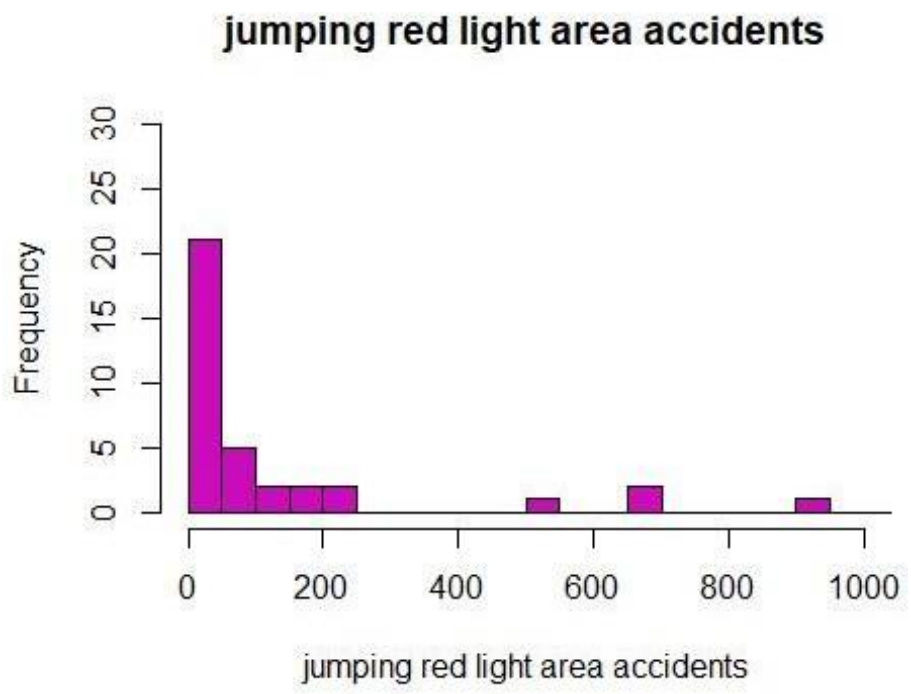

**Inference:** The major accidents occurred due to use of mobile phones while driving is below 500 in major states.

```
hist(dataset4.typeofaccident.license.traffic.violation$Fatal.Accidents,breaks
= 50,col=4,ylim = c(0,20),main="fatal accidents")
abline(v=mean(dataset4.typeofaccident.license.traffic.violation$Fatal.Acciden ts))
abline(v=median(dataset4.typeofaccident.license.traffic.violation$Fatal.Accid ents))
```

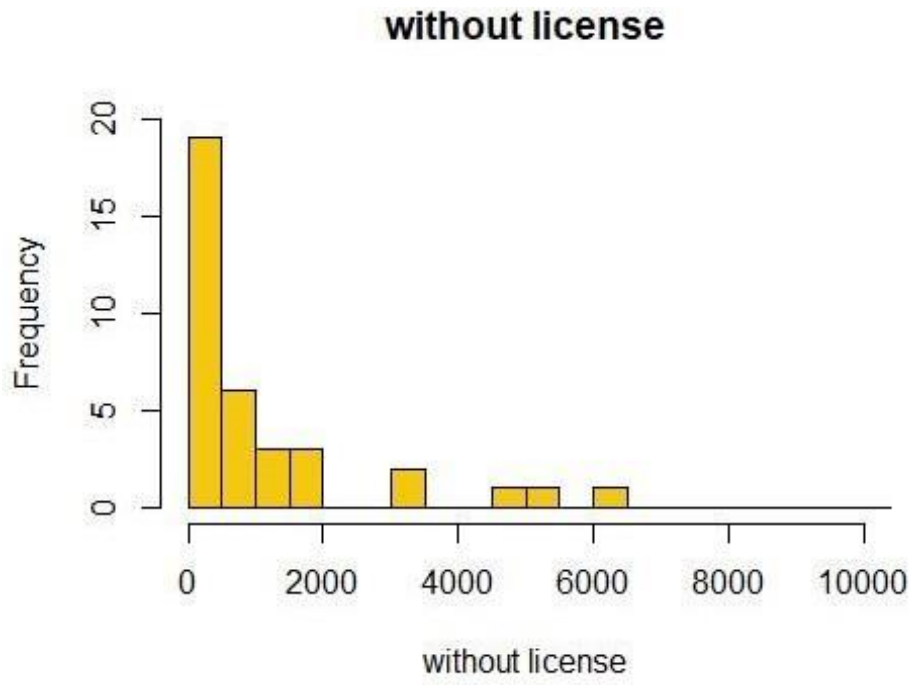



**Inference:** The mean and median of the accidents shows in the plot. the mean is 7250 and median is 2243.three states there are 6000-7000 accidents occurred.

```
hist(dataset4.typeofaccident.license.traffic.violation$Jumping.Red.Light...Number.of.Accidents, breaks = 90,col=6,ylim=c(0,30),xlim=c(0,1000),main="jumping red light area accidents",xlab="jumping red light area accidents")
```

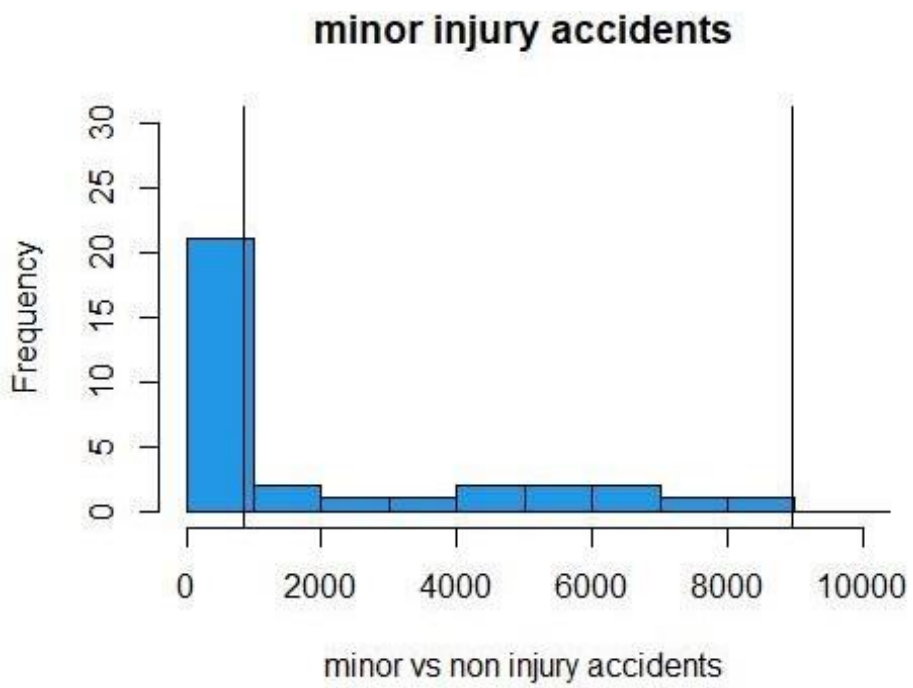



**Inference:** The major accidents ocuured due to use of jumping red light area accidents while driving is below 300 in major states.




**Inference:** The persons not holding license below the age 18 years are the major cause of accidents occurred in India. Nearly 2000 person killed or died in accident 2018.

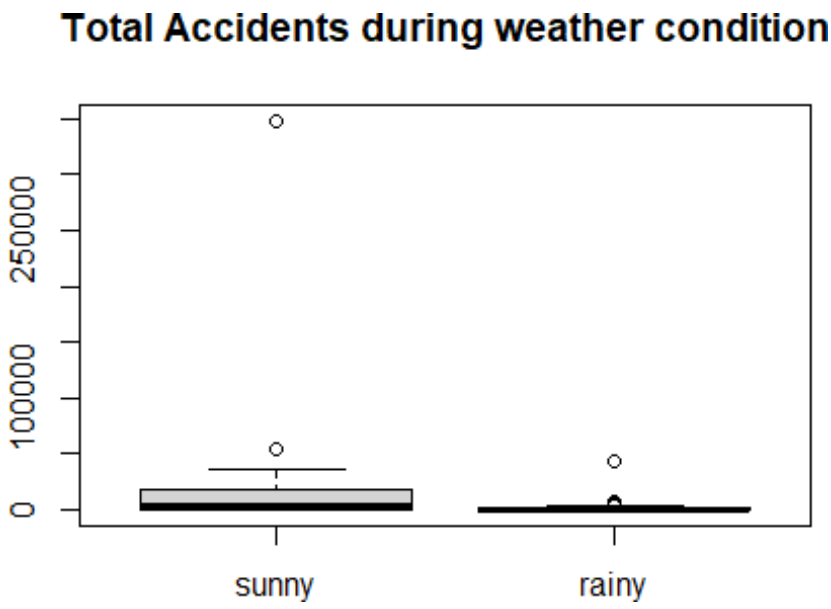
```
hist(dataset4.typeofaccident.license.traffic.violation$Minor.Injury.Accidents
,breaks = 195,xlim=c(0,10000),ylim=c(0,30),col=20,xlab="minor vs non injury a
ccidents",main="minor injury accidents")
abline(v=median(dataset4.typeofaccident.license.traffic.violation$Minor.Injur
y.Accidents))
abline(v=mean(dataset4.typeofaccident.license.traffic.violation$Minor.Injury.
Accidents))
```




**Inference:** The mean and median is between the range of 1000-9000 in minor injury accidents, maximum states of accidents occurred in minimum rates in 2018.

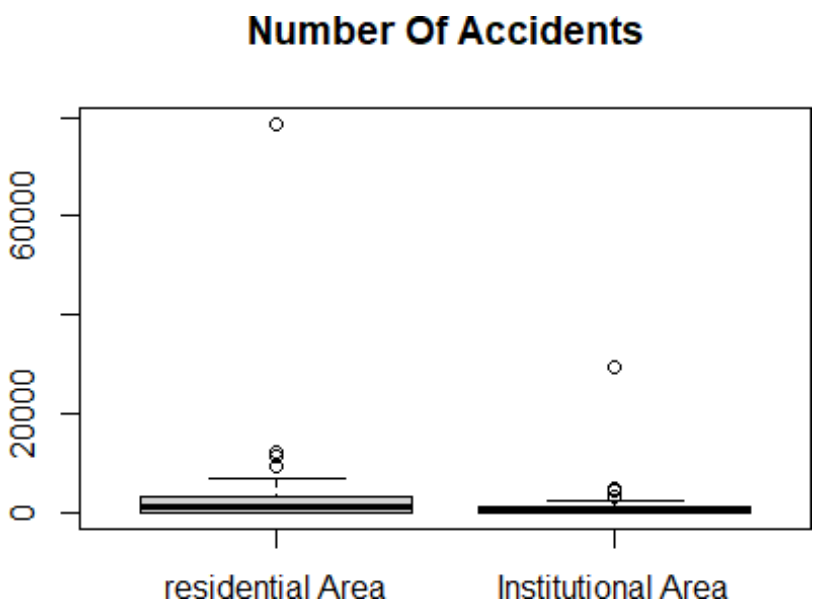
**Boxplot Analysis:**

```
library(tidyverse)

library(lattice)
library(readr)


Dataset2_weather_roadcond_load <- read.csv("Dataset2-weather-roadcond-load
.csv")
Dataset3_typevechicle_safety_time<-read.csv("Dataset3-typevechicle-safety-
time.csv")
dataset4_typeofaccident_license_traffic_violation <- read.csv("dataset4-ty
peofaccident-license-traffic-violation.csv")
Dataset1_dies_injured_Age_sex <- read.csv("Dataset1-dies-injured-Age_sex.c
sv")

boxplot(Dataset2_weather_roadcond_load$Sunny.Clear...Total.Accidents...Num
ber,Dataset2_weather_roadcond_load$Rainy...Total.Accidents,las=0,names=c("
sunny","rainy"),main="Total Accidents during weather condition")
```



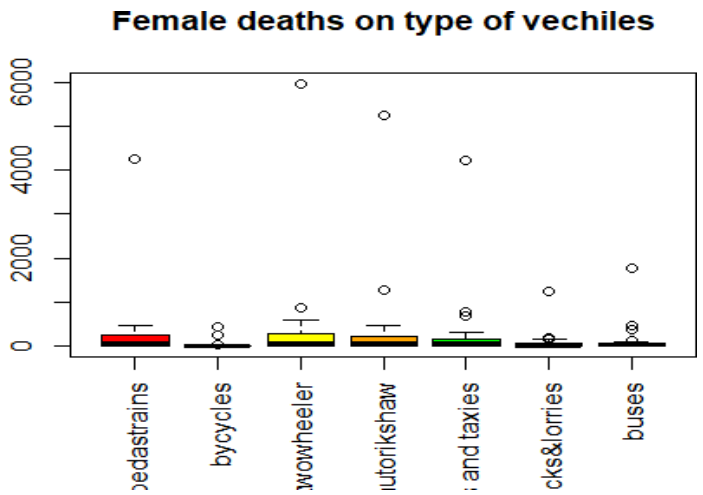

**Inference:** The above plot shows the Accidents in different weather conditions like sunny and rainy seasons. The IQR range for the sunny data is more than the rainy season.

```
boxplot(Dataset2_weather_roadcond_load$Residential.Area...Total.Accidents,
Dataset2_weather_roadcond_load$Institutional.Area...Total.Accidents,las=0,
names=c("residential Area","Institutional Area"),main="Number Of Accidents
")
```



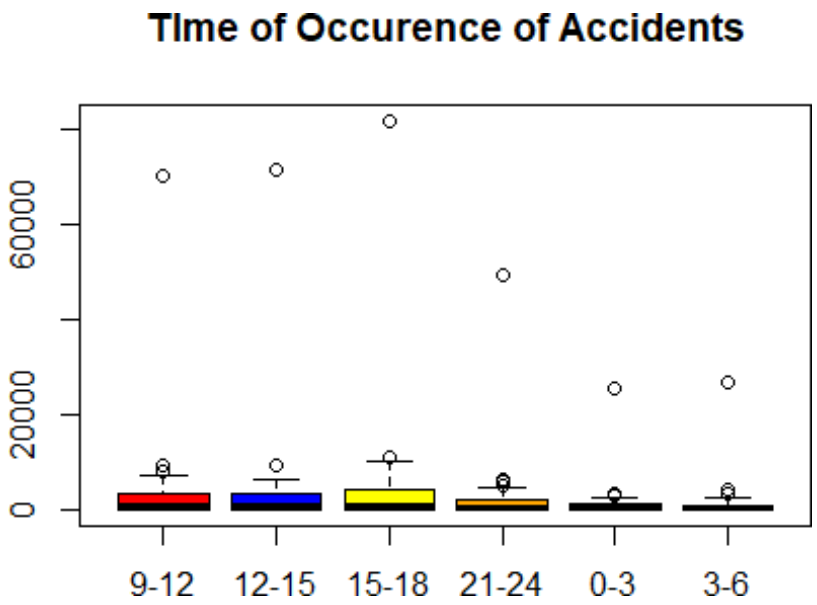

**Inference:** This plot shows that the number of accidents of are maximum in the residential area compared to the Institutional. The more of Accidents are may be due to jam.

```
boxplot(Dataset3_typevechicle_safety_time$Pedestrian...Female,Dataset3_typ
evechicle_safety_time$Bycycles...Female,Dataset3_typevechicle_safety_time$
Two.Wheelers...Female,Dataset3_typevechicle_safety_time$Auto.Rickshaws...M
ale,Dataset3_typevechicle_safety_time$Cars..taxies.Vans...LMV...Female,Dat
aset3_typevechicle_safety_time$Trucks.Lorries...Female,las=3,Dataset3_type
vechicle_safety_time$Buses...Female,main="Female deaths on type of vechile
s",at = c(1,2,3,4,5,6,7),names = c("pedastrains","bycycles","twowheeler","
autorikshaw","cars and taxies","trucks&lorries","buses") ,col=c("red","blu
e","yellow","orange","green","violet"))
```



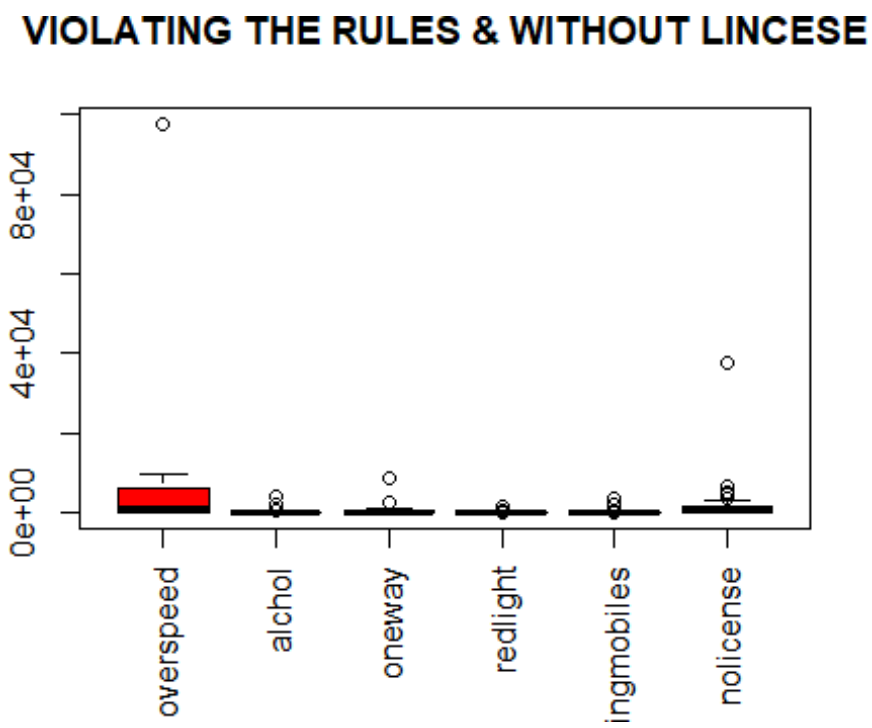
**Inference:** This plot shows that female deaths on the different type of Vehicles. The outliers are in the two wheelers vehicle which is nearer to 6000.

```
boxplot(Dataset3_typevechicle_safety_time$X09.1200hrs ...Day.,Dataset3_typ
evechicle_safety_time$X12.1500hrs ...Day.,Dataset3_typevechicle_safety_tim
e$X15.1800hrs....Day.,Dataset3_typevechicle_safety_time$X21.2400hrs... Nig
ht.,Dataset3_typevechicle_safety_time$X00.300hrs ...Night.,Dataset3_typeve
chicle_safety_time$X03.600hrs....Night.,main="TIme of Occurence of Acciden
ts",at = c(1,2,3,4,5,6),names = c("9-12","12-15","15-18","21-24","0-3","3-
6"),col=c("red","blue","yellow","orange","green","brown","violet"))
```



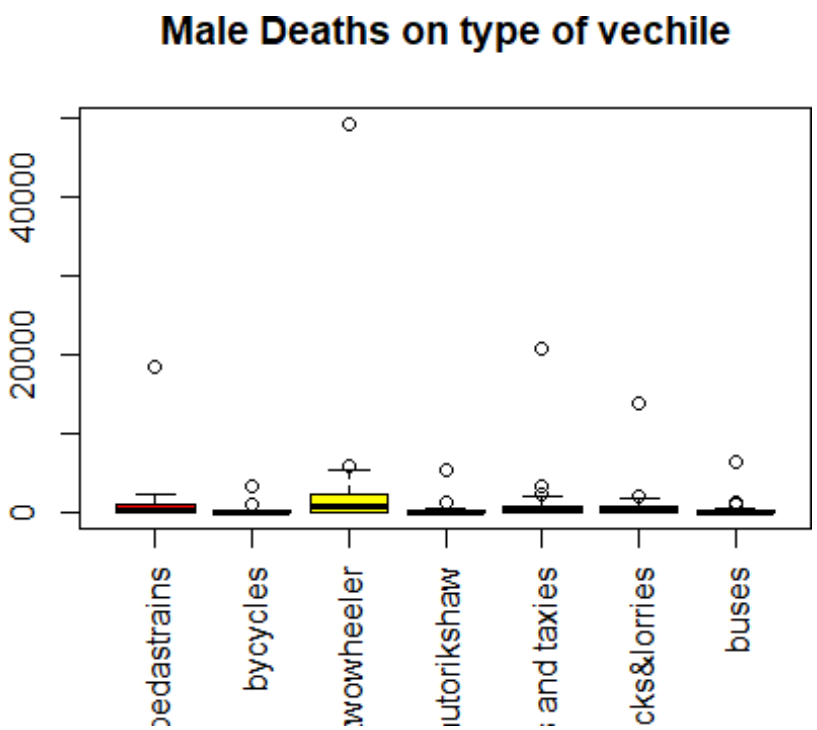

**Inference:** This plot shows the timeline of occurrence of Accidents . The time interval between 15.00 hrs to 18.00 hrs is more and also contains the maximum outliers.

```
 boxplot(dataset4_typeofaccident_license_traffic_violation$Over.Speeding..
 .Persons.Killed...Number,dataset4_typeofaccident_license_traffic_violation
$Drunken.Driving.Consumption.of.Alcohol...Drug...Persons.Killed,dataset4_t
ypeofaccident_license_traffic_violation$Driving.on.Wrong.Side...Persons.Ki
lled,dataset4_typeofaccident_license_traffic_violation$Jumping.Red.Light..
.Persons.Killed,dataset4_typeofaccident_license_traffic_violation$Use.of.M
obile.Phone...Persons.Killed,dataset4_typeofaccident_license_traffic_viola
tion$Without.Licence,main="VIOLATING THE RULES & WITHOUT LINCESE",at = c(1
,2,3,4,5,6),las=3,names = c("overspeed","alchol","oneway","redlight","usin
gmobiles","nolicense"), col=c("red","blue","yellow","orange","green","viol
et"))
```



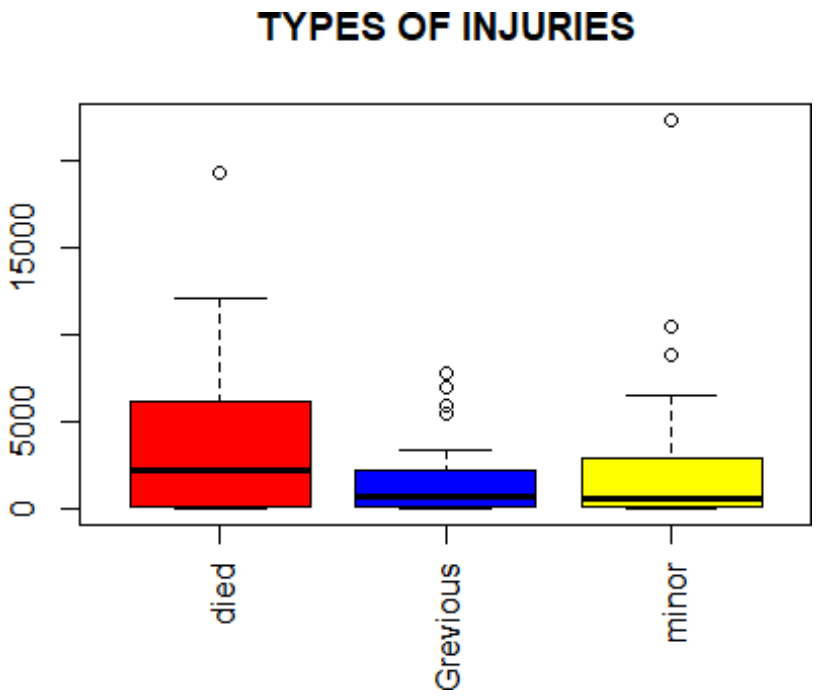

**Inference:** This plot shows the various reasons for the occurrence of Accidents. The Overspeed is the main reason which leads to more number of deaths.

```
pedastrain=Dataset3_typevechicle_safety_time$Pedestrian...Male
Bycycle=Dataset3_typevechicle_safety_time$Bycycles...Male
twowheeler=Dataset3_typevechicle_safety_time$Two.Wheelers...Male
cars_taxies_vans=Dataset3_typevechicle_safety_time$Cars..taxies.Vans...LMV
...Male
trucks_lorries=Dataset3_typevechicle_safety_time$Trucks.Lorries...Male
buses=Dataset3_typevechicle_safety_time$Buses...Male

boxplot(pedastrain,Bycycle,twowheeler,Dataset3_typevechicle_safety_time$Au
to.Rickshaws...Male,cars_taxies_vans,trucks_lorries,buses,col=c("red","blu
e","yellow","orange","green","violet"),main="Male Deaths on type of vechil
e",at = c(1,2,3,4,5,6,7),las=3,names = c("pedastrains","bycycles","twowhee
ler","autorikshaw","cars and taxies","trucks&lorries","buses"))
```



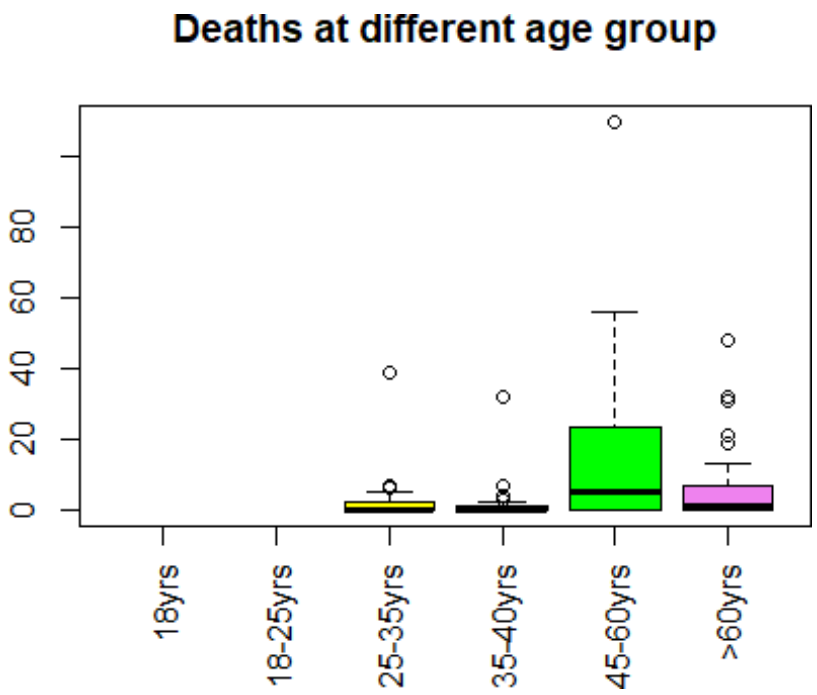

**Inference:** This plot shows the Male deaths on different type of vehicles. The maximum number of deaths occurred in two wheeler.

```
boxplot(Dataset1_dies_injured_Age_sex$person.died.in.2018,Dataset1_dies_in
jured_Age_sex$Greviously.Injured.persons,Dataset1_dies_injured_Age_sex$Min
or.Injured.persons,main="TYPES OF INJURIES",at = c(1,2,3),las=3,names = c(
"died","Grevious","minor"),col = c("red","blue","yellow"))
```



**Inference:** This plot shows the level of injuries (died, Grevious_injured, Minor_injured). The grevious injured people are less than the people who are injured minorly.

```
boxplot(Dataset1_dies_injured_Age_sex$X18.Yrs...Killed...Male,Dataset1_dies_injured_Age_sex$X18.25.Yrs...Killed...Male,Dataset1_dies_injured_Age_sex$X25.35.Yrs...Killed...Female,Dataset1_dies_injured_Age_sex$X35.40.Yrs...Killed...Female,Dataset1_dies_injured_Age_sex$X45.60.Yrs...Killed...Male,Dataset1_dies_injured_Age_sex$X60.Yrs.above...Killed...Male,main="Deaths at different age group",at = c(1,2,3,4,5,6),las=3,names = c("18yrs","18-25yrs","25-35yrs","35-40yrs","45-60yrs",">60yrs"),col=c("red","blue","yellow","orange","green","violet"))
```




**Inference:** This shows death rate of different age groups. The maximum of deaths occurred in the age group of 45 – 60 years. Above 60 years, there are more outliers.

**Scatterplot Analysis:**

## Accident by persons having license, without license

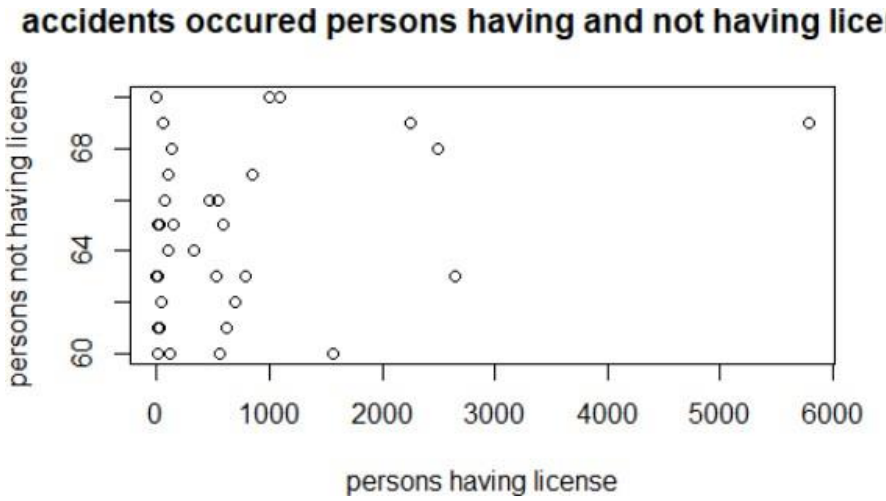
>library(readxl)

> dataset4 <- read_excel("D:/Desktop/Desktop/sharmila/dataset4.xlsx")

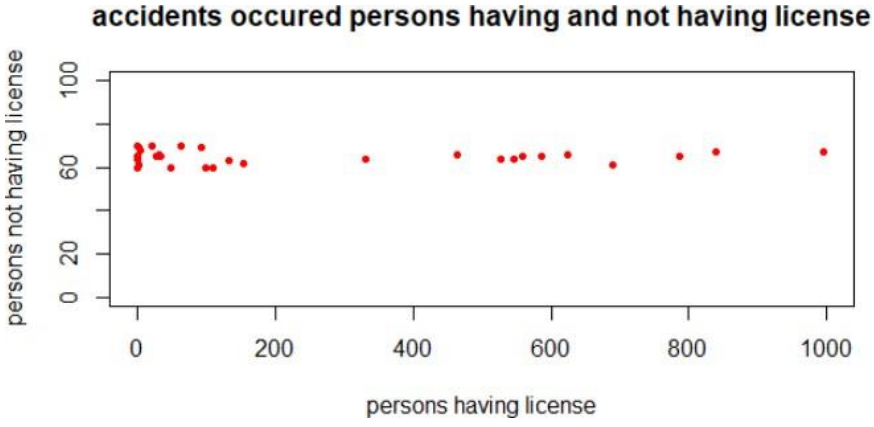
>View(dataset4)

>plot(dataset4$`Licence`,dataset4$`Without Licence`)

>plot(dataset4$`Licence`,dataset4$`Without Licence`,xlab="persons having license",ylab="persons not having license",main="accidents occured persons having and not having license")



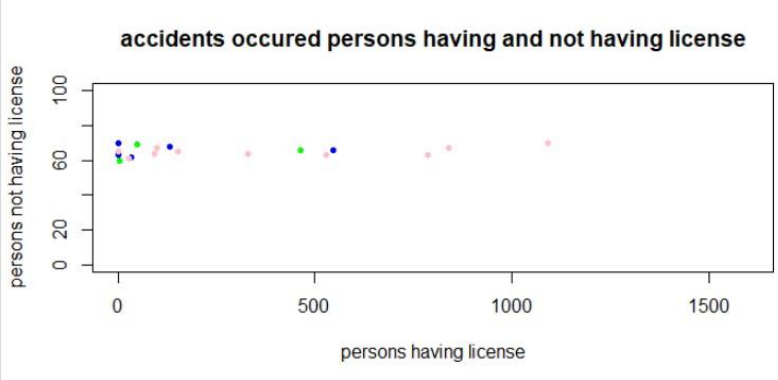

> plot(dataset4$`Licence`,dataset4$`Without Licence`,xlab="persons having license",ylab="persons not having license",main="accidents occured persons having and not having license",xlim =c(0,1000),ylim = c(0,100),pch=20,col="red")



> points(dataset4$`Licence`[dataset4$`States/UTs`=="Uttar Pradesh"],dataset4$`Without Licence`[dataset4$`States/UTs`=="Uttar Pradesh"],xlab="persons having license",ylab="persons not having license",main="accidents occured persons having and not having license",xlim =c(0,1600),ylim = c(0,100),pch=20,col="blue")

> points(dataset4$`Licence`[dataset4$`States/UTs`=="Tamil Nadu"],dataset4$`Without Licence`[dataset4$`States/UTs`=="Tamil Nadu"],xlab="persons having license",ylab="persons not having license",main="accidents occured persons having and not having license",xlim =c(0,1600),ylim = c(0,100),pch=20,col="green")

> points(dataset4$`Licence`[dataset4$`States/UTs`=="Chhattisgarh"],dataset4$`Without Licence`[dataset4$`States/UTs`=="Chhattisgarh"],xlab="persons having license",ylab="persons not having license",main="accidents occured persons having and not having license",xlim =c(0,1600),ylim = c(0,100),pch=20,col="pink")




**Inference:** Accidents occurred by persons having license and not having license is taken and compared with three states Uttarpradesh, Chattisgarh and Tamilnadu. The limit is taken (0,1600) and (0,100) respectively. Persons having license below 500 and not having license above in 60-80 range accident is occurred more

## Minor Injury, Non-Injury in accidents

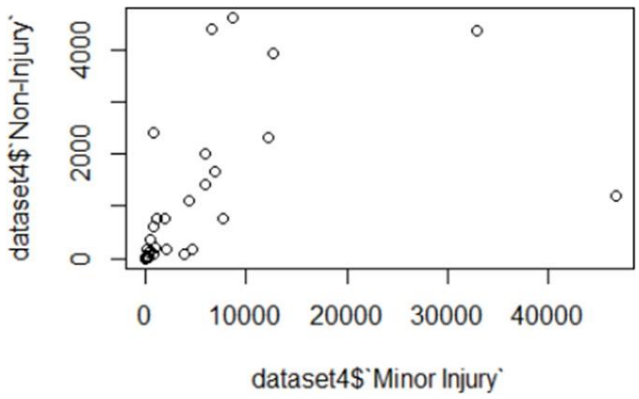
library(readxl)

> dataset4 <- read_excel("D:/Desktop/Desktop/sharmila/dataset4.xlsx")

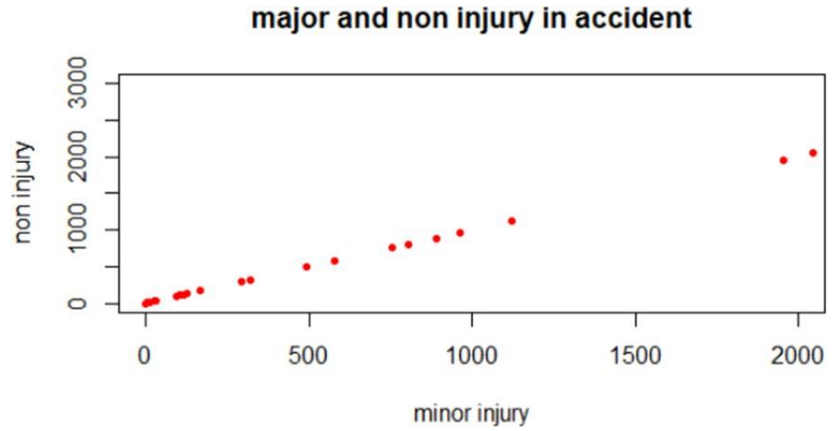
>View(dataset4)

>plot(dataset4$`Minor Injury`,dataset4$`Non-Injury`)

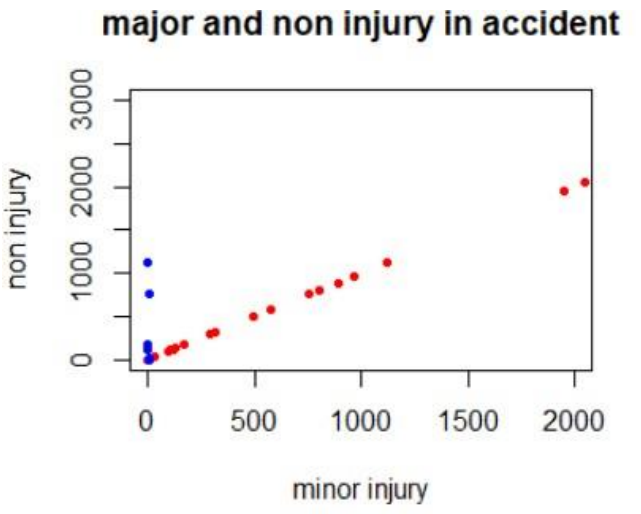
>plot(dataset4$`Minor Injury`,dataset4$`Non-Injury`,xlab="minor injury",ylab="non injury",main="major and minor injury in accident")



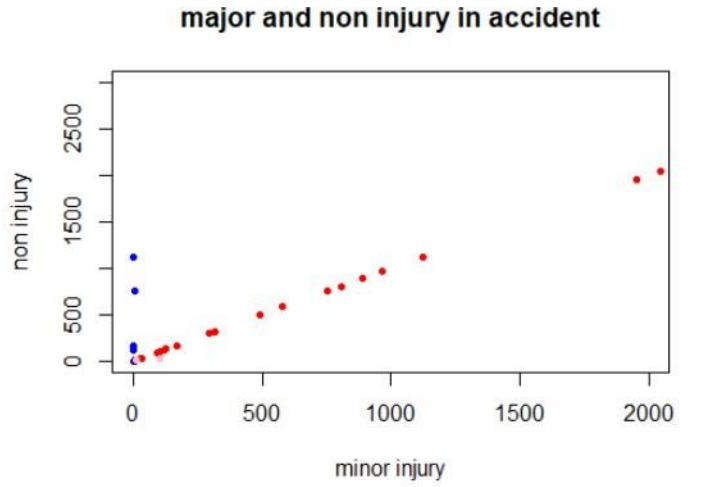

> plot(dataset4$`Minor Injury`,dataset4$`Minor Injury`,xlab="minor injury",ylab="non injury",main="major and non injury in accident",xlim =c(0,2000),ylim = c(0,3000),pch=20,col="red")



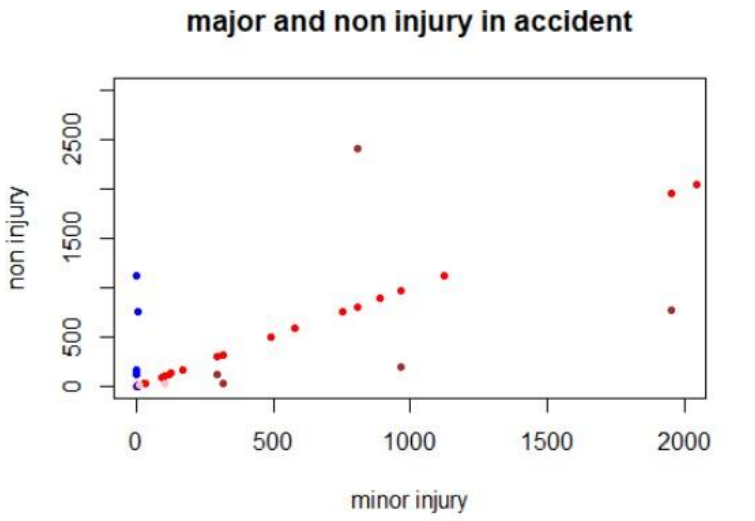

>points(dataset4$`Minor Injury``[dataset4$`States/UTs`=="Uttar Pradesh"],dataset4$`Non-Injury `[dataset4$`States/UTs`=="Uttar Pradesh"],xlab="minor injury",ylab="non injury",main="major and non injury in accident",xlim =c(0,2000),ylim = c(0,3000),pch=20,col="blue")




>points(dataset4$`Minor Injury`[dataset4$`States/UTs`=="Tamil Nadu"],dataset4$`Non-Injury`[dataset4$`States/UTs`=="Tamil Nadu"],xlab="minor injury",ylab="non injury",main="major and non-injury in accident",xlim =c(0,2000),ylim = c(0,3000),pch=20,col="pink")




>points(dataset4$`Minor Injury`[dataset4$`States/UTs`=="Chhattisgarh"],dataset4$`Non-Injury`[dataset4$`States/UTs`=="Chhattisgarh"],xlab="minor injury",ylab="non injury",main="major and non injury in accident",xlim =c(0,2000),ylim = c(0,3000),pch=20,col="brown")




**Inference:** Major and no injury in accident is taken in all states and with specific states like Uttarpradesh and Tamilnadu and Chattisgarh are compared .in all state's injury is in increasing trend .in uttarpradesh non injury is in increasing trend and in Tamilnadu (0-250) range minor injury is occurred

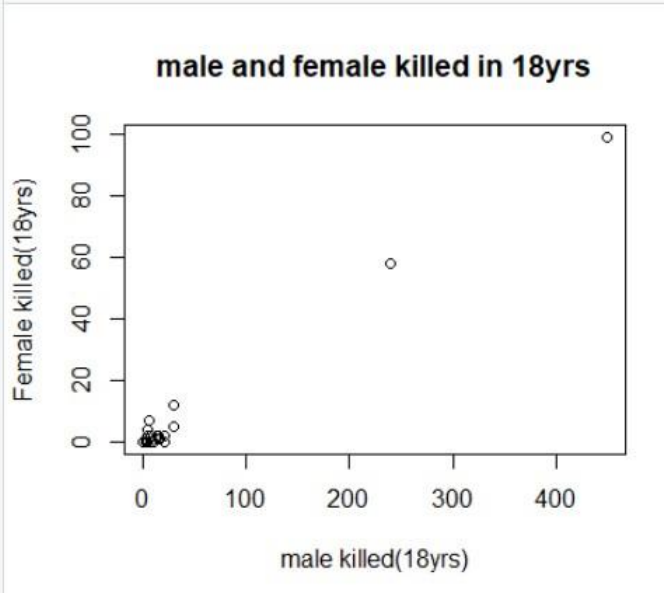
## Male and female(18yrs) killed in accident

>library(readxl)

> Dataset1 <- read_excel("D:/Desktop/Desktop/Dataset1.xlsx")

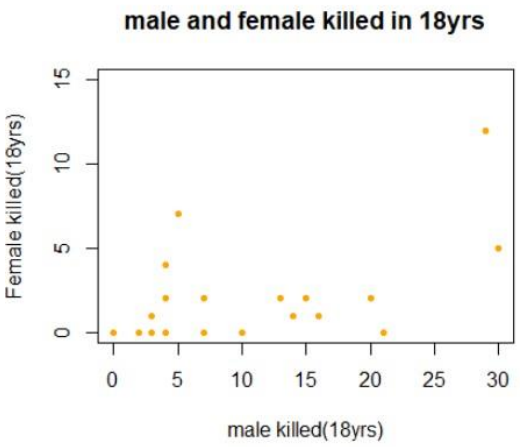
>View(Dataset1)

>plot(Dataset1$`18 Yrs - Killed - Male`,Dataset1$`18 Yrs - Killed - Female`)



>plot(Dataset1$`18 Yrs - Killed - Male`,Dataset1$`18 Yrs - Killed - Female`,xlab="male killed(18yrs)",ylab = "Female killed(18yrs)",main = "male and female killed in 18yrs")

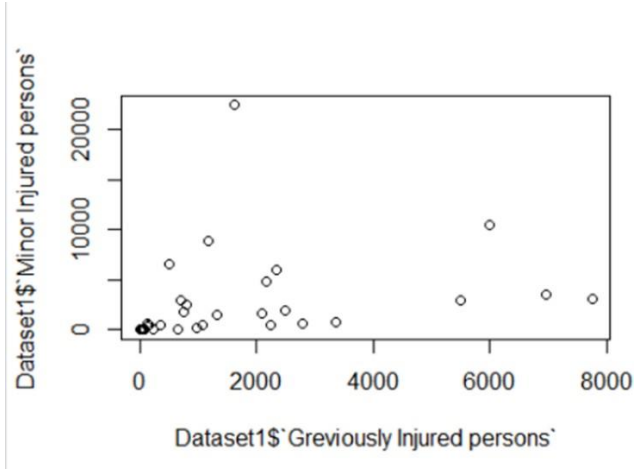
> plot(Dataset1$`18 Yrs - Killed - Male`,Dataset1$`18 Yrs - Killed - Female`,xlab="male killed(18yrs)",ylab = "Female killed(18yrs)",main = "male and female killed in 18yrs",xlim =c(0,30),ylim=c(0,15),pch=20,col="orange")




## MINOR AND GREVIOUSLY INJURED PERSON in accident

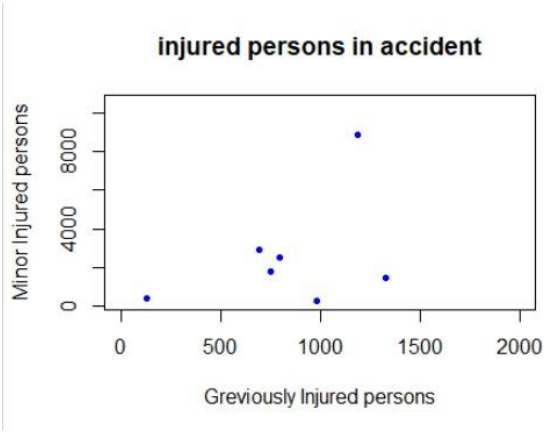
>plot(Dataset1$`Greviously Injured persons`,Dataset1$`Minor Injured persons`)

>plot(Dataset1$`Greviously Injured persons`,Dataset1$`Minor Injured persons`,xlab="Greviously Injured persons",ylab="Minor Injured persons",main="injured persons in accident",pch=60,col="black")

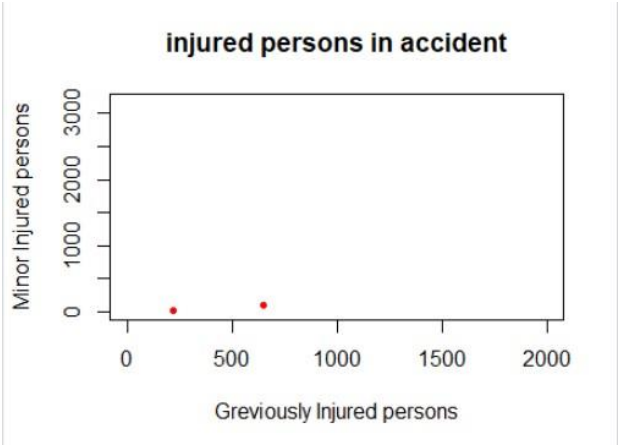



> plot(Dataset1$`Greviously Injured persons`[Dataset1$`States/UTs`=="Assam"],Dataset1$`Minor Injured

persons`[Dataset1$`States/UTs`=="Assam"],xlab="Greviously Injured persons",ylab="Minor Injured persons",main="injured persons in accident",pch=20,col="blue",xlim=(0,2000)




> plot(Dataset1$`Greviously Injured persons`[Dataset1$`States/UTs`=="Tripura"],Dataset1$`Minor Injured persons`[Dataset1$`States/UTs`=="Tripura"],xlab="Greviously Injured persons",ylab="Minor Injured persons",main="injured persons in accident",pch=60,col="red",xlim=c(0,2000))



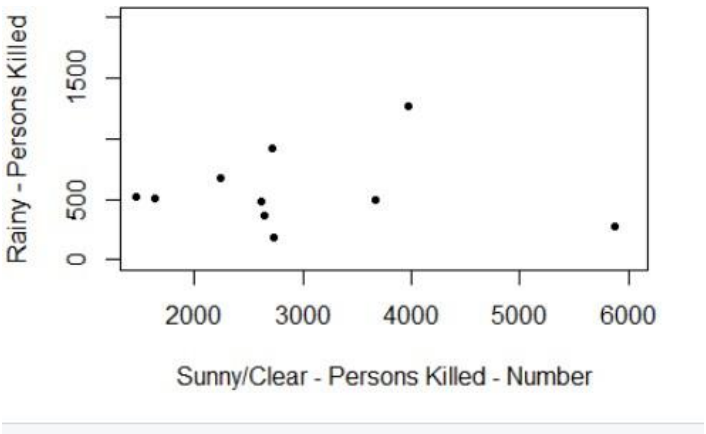

**Inference:** In minor injured persons and grievously injured persons in assam is taken from (0,2000) and (0,above 8000) respectively.in the range of 500-1000 major persons is injured .In Tripura same range is taken 0-1000 and minimum persons is injured

## persons killed in sunny and rainy season

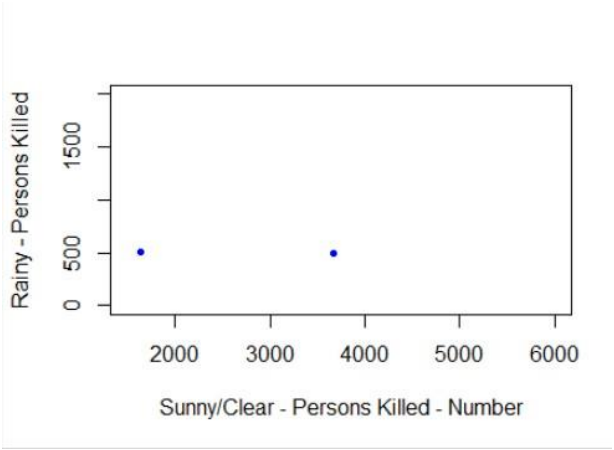
View(Dataset2)

>plot(Dataset2$`Sunny/Clear - Persons Killed - Number`,Dataset2$`Rainy - Persons Killed`)

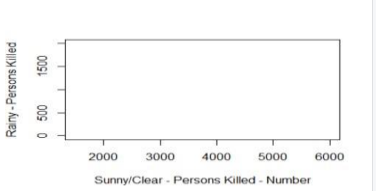
> plot(Dataset2$`Sunny/Clear - Persons Killed - Number`,Dataset2$`Rainy - Persons Killed`,xlab="Sunny/Clear - Persons Killed - Number",ylab="Rainy - Persons Killed",xlim=c(1500,6000),ylim = c(0,2000),pch=20,col="black")




```
> plot(Dataset2$`Sunny/Clear - Persons Killed - Number`[Dataset2$`States/UTs`=="Jharkhand"],Dataset2$`Rainy - Persons Killed`[Dataset2$`States/UTs`=="Jharkhand"],xlab="Sunny/Clear - Persons Killed - Number",ylab="Rainy - Persons Killed",xlim=c(1500,6000),ylim = c(0,2000),pch=20,col="blue")
```




```
> plot(Dataset2$`Sunny/Clear - Persons Killed - Number`[Dataset2$`States/UTs`=="Delhi"],Dataset2$`Rainy - Persons Killed`[Dataset2$`States/UTs`=="Delhi"],xlab="Sunny/Clear - Persons Killed - Number",ylab="Rainy - Persons Killed",xlim=c(1500,6000),ylim = c(0,2000),pch=20,col="green")
```



**Inference:** In sunny/clear weather and in rainy weather persons killed in Jharkhand the limit is taken from (1500,6000) and (0,2000) respectively. This state has the minimum persons killed in both the weather and in Delhi same range so no persons is killed

## Accidents caused by male and female in two wheelers

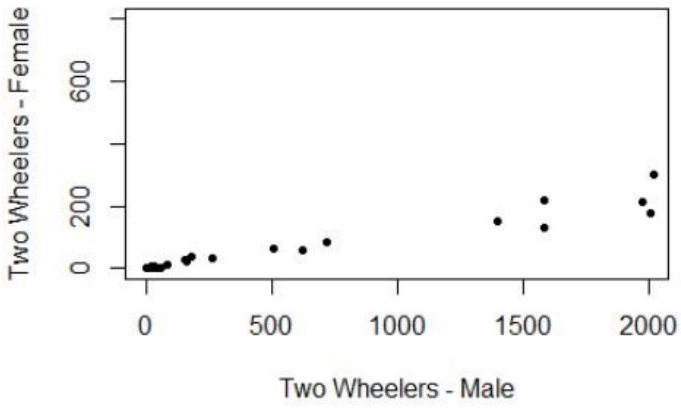
>library(readxl)

> Dataset3 <- read_excel("D:/Desktop/Desktop/Dataset3.xlsx")

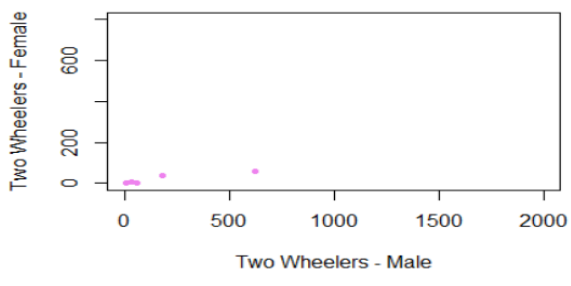
>View(Dataset3)

>plot(Dataset3$`Two Wheelers - Male`,Dataset3$`Two Wheelers - Female`)

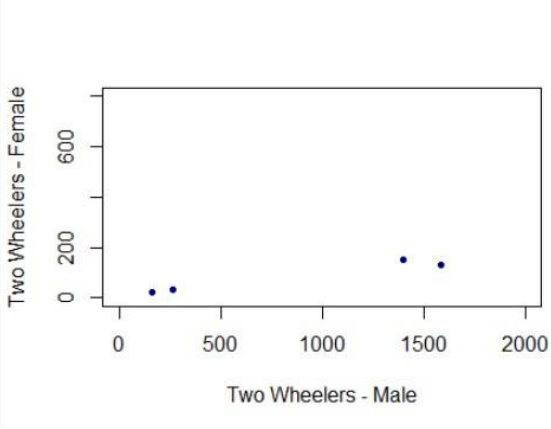
> plot(Dataset3$`Two Wheelers - Male`,Dataset3$`Two Wheelers - Female`,xlab = "Two Wheelers - Male",ylab = "Two Wheelers - Female",xlim=c(0,2000),ylim =c(0,800),pch=20,col="black")




> plot(Dataset3$`Two Wheelers - Male`[Dataset3$`States/UTs`=="Sikkim"],Dataset3$`Two Wheelers - Female`[Dataset3$`States/UTs`=="Sikkim"],xlab = "Two Wheelers - Male",ylab = "Two Wheelers - Female",xlim=c(0,2000),ylim =c(0,800),pch=20,col="violet")



```
> points(Dataset3$`Two Wheelers -
Male`[Dataset3$`States/UTs`=="Haryana"],Dataset3$`Two Wheelers -
Female`[Dataset3$`States/UTs`=="Haryana"],xlab = "Two Wheelers - Male",ylab = "Two
Wheelers - Female",xlim=c(0,2000),ylim =c(0,800),pch=20,col="navy")
```




**Inference:** Accidents occurred due to two-wheeler are taken in two states are taken Tamilnadu and Sikkim. In Sikkim state accidents occurred by both female and male are below 500 persons and in Tamilnadu only two accidents occurred.

***Insights:***

**Causes:**

- **From the Violating the rules dataset, we could observe that the overspeed is one of the violations which causes more Accidents.**
- **In Sex wise Categories of deaths, we could find that a greater number of deaths in Male and female is caused by Two-Wheeler type of Accidents.**
- **In Time of Occurrence dataset, we can find that a most road Accidents occurs between the time interval (18.00hour – 21.00hour) busy hours.**

**Remedies:**

- **The possibility of Accidents in Public Transport is lower so we could use the public transports more to avoid Accidents.**
- **Implementing Speed Limiter in the Two-Wheeler so that we could avoid the overspeed violation and following strict traffic rules could also reduce it.**
- **The more accidents occur in the time period (18.00hours - 21.00hours) so by taking precautions we could avoid this.**

**----------XXX----------**