20122065_ Lab 13

Code

Load the dataset

Hide

```
data <- read.csv("accident.csv", stringsAsFactors = T)
head(data)
```

	id <int>	weight <dbl>	dead <fctr>	airbag <fctr>	seatbelt <fctr>	frontal <int>	sex <fctr>	ageOFocc <int>	yearacc <int>
1	1	25.069	alive	none	belted	1	f	26	1997
2	2	25.069	alive	airbag	belted	1	f	72	1997
3	3	32.379	alive	none	none	1	f	69	1997
4	4	495.444	alive	airbag	belted	1	f	53	1997
5	5	25.069	alive	none	belted	1	f	32	1997
6	6	25.069	alive	none	belted	1	f	22	1997

6 rows | 1-10 of 14 columns

Hide

```
attach(data)
```

```
The following object is masked from df (pos = 4):

    sex

The following object is masked from df (pos = 13):

    sex
```

Summary of the data

Hide

```
summary(data)
```

```
       id            weight           dead          airbag      Min.
:    1   Min.   :    0.00   alive:14282   airbag:7165
 1st Qu.: 3750   1st Qu.:   31.80   dead :  717   none  :7834
 Median : 7500   Median :   82.98
 Mean   : 7500   Mean   :  440.38
 3rd Qu.:11250   3rd Qu.:  342.74
 Max.   :14999   Max.   :57871.60
seatbelt        frontal       sex         ageOFocc
belted:10512   Min.   :0.0000   f:7021   Min.   :16.00    none  :
```

```
 4487   1st Qu.:0.0000   m:7978   1st Qu.:22.00
 Median :1.0000            Median :33.00
                Mean   :0.6362            Mean   :37.48
                3rd Qu.:1.0000            3rd Qu.:48.00
                Max.   :1.0000            Max.   :97.00
 yearacc          abcat        occRole          deploy
 Min.   :1997   deploy  :4410   driver:11789   Min.   :0.000
  1st Qu.:1997   nodeploy:2755   pass  : 3210   1st Qu.:0.000
  Median :1998   unavail :7834                  Median :0.000
  Mean   :1998                                  Mean   :0.294
  3rd Qu.:1999                                  3rd Qu.:1.000
  Max.   :2000                                  Max.   :1.000
 injSeverity            caseid       Min.   :0.000   0.556261574:
 12
  1st Qu.:1.000   0.584039352:   12
  Median :2.000   0.590289352:   12
  Mean   :1.746   0.54931713 :   11
  3rd Qu.:3.000   0.587511574:   11
  Max.   :6.000   0.588900463:   11
  NA's   :76      (Other)    :14930
```

Structure of the data

Hide

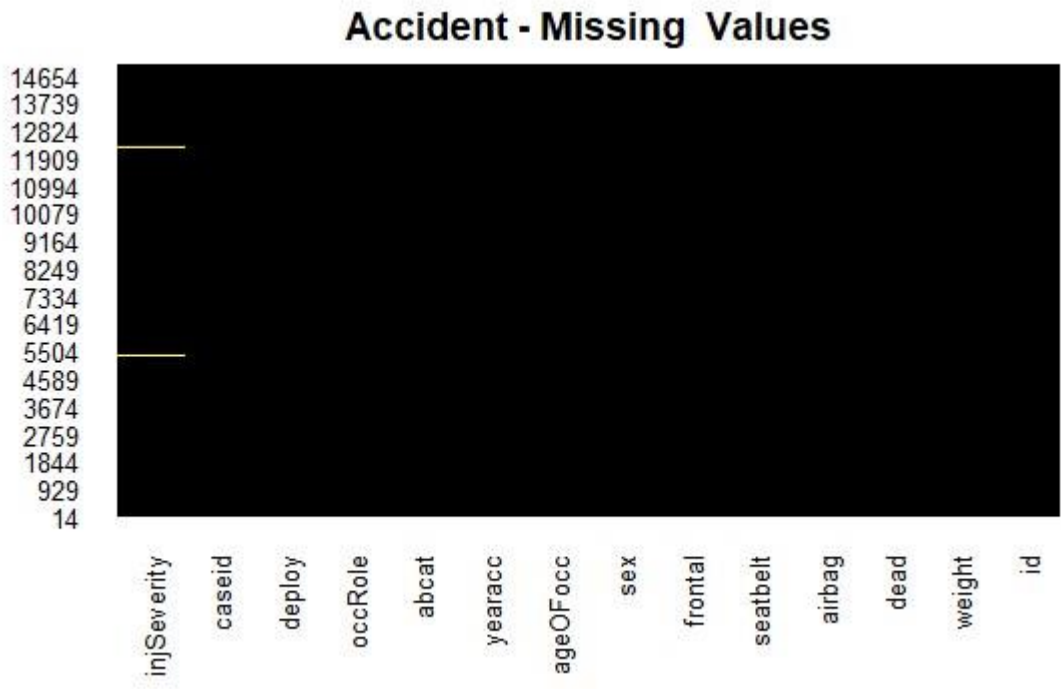
```
str(data)
```

```
'data.frame':   14999 obs. of  14 variables:
 $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ weight     : num  25.1 25.1 32.4 495.4 25.1 ...
 $ dead       : Factor w/ 2 levels "alive","dead": 1 1 1 1 1 1 1 2 1 1 ...
 $ airbag     : Factor w/ 2 levels "airbag","none": 2 1 2 1 2 2 2 2 2 2 ...
 $ seatbelt   : Factor w/ 2 levels "belted","none": 1 1 2 1 1 1 1 2 1 1 ...
 $ frontal    : int  1 1 1 1 1 1 1 1 0 1 ...  $ sex        : Factor
w/ 2 levels "f","m": 1 1 1 1 1 1 2 2 2 1 ...
 $ ageOFocc   : int  26 72 69 53 32 22 22 32 40 18 ...
 $ yearacc    : int  1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 ...
 $ abcat      : Factor w/ 3 levels "deploy","nodeploy",..: 3 1 3 1 3 3 3 3 3 3 ...
 $ occRole    : Factor w/ 2 levels "driver","pass": 1 1 1 1 1 1 1 1 1 1 ...
 $ deploy     : int  0 1 0 1 0 0 0 0 0 0 ...
 $ injSeverity: int  3 1 4 1 3 3 3 4 1 0 ...
 $ caseid     : Factor w/ 6673 levels "0.125011574",..: 1889 1890 1893 1900 1902 1903 1905 1906
1907 1908 ...
```

Using the library AMELIA for the visulaization of missing values for analysis in our dataset

Hide

```
library(Amelia)
missmap(data, main =  "Accident - Missing  Values" ,
        col = c( "yellow", "black"), legend =  FALSE)
```



Hide

```
any(is.na(data))
```

```
[1] TRUE
```

Hide

```
summary(is.na(data))
```

```
     id            weight           dead           airbag         Mode
:logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:14999     FALSE:14999     FALSE:14999     FALSE:14999
seatbelt        frontal           sex           ageOFocc
Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:14999     FALSE:14999     FALSE:14999     FALSE:14999
yearacc          abcat          occRole          deploy
Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:14999     FALSE:14999     FALSE:14999     FALSE:14999
injSeverity       caseid         Mode :logical   Mode :logical
 FALSE:14923     FALSE:14999      TRUE
:76
```

Since there is a present in NULL VALUES we are removing it.

Now using lib DPLYR we are removing the NULL VALUES

Hide

```
library(dplyr)
```

```
package 'dplyr' was built under R version 4.0.4
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

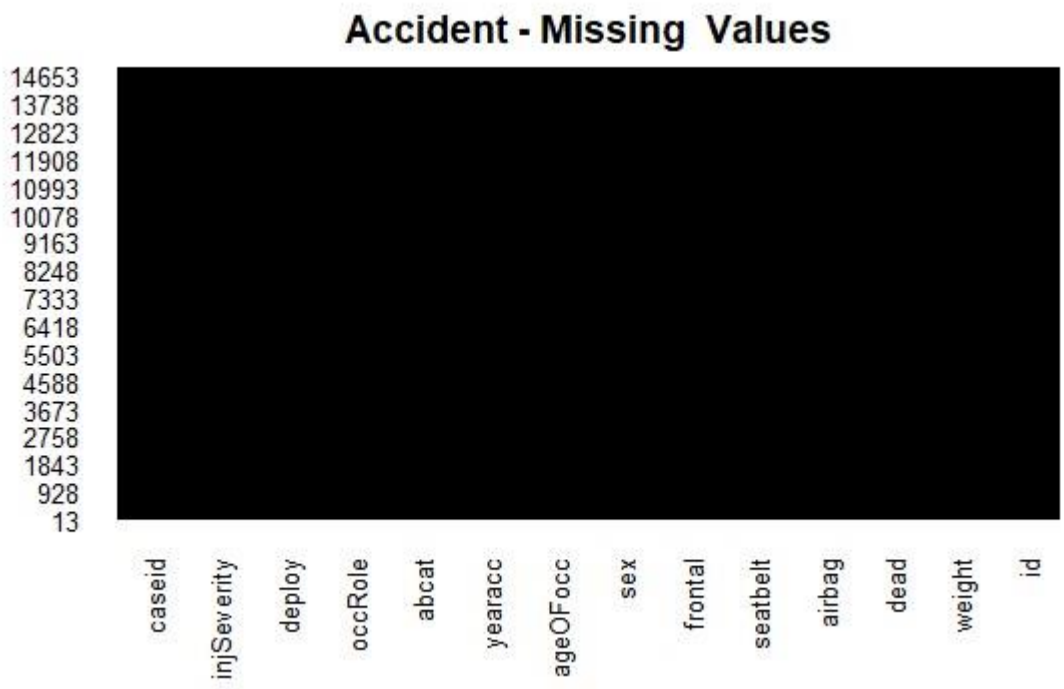
Hide

```
data1 <- data %>% na.omit()
dim(data1)
```

```
[1] 14923    14
```

Hide

```
missmap(data1, main = "Accident - Missing Values",
        col = c("yellow", "black"), legend = FALSE)
```

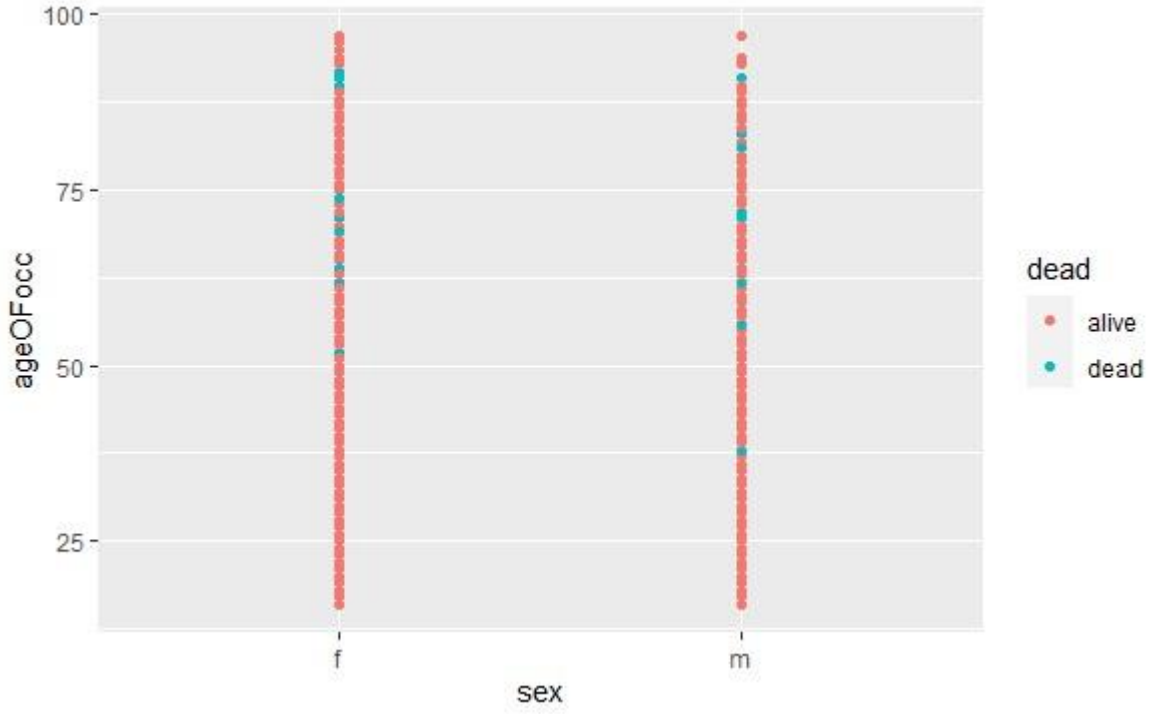


The na values are removed.

Using the lib GGPLOT2 fr the visualizatin of the relationship b/w the variable for analysis in our dataset

Hide

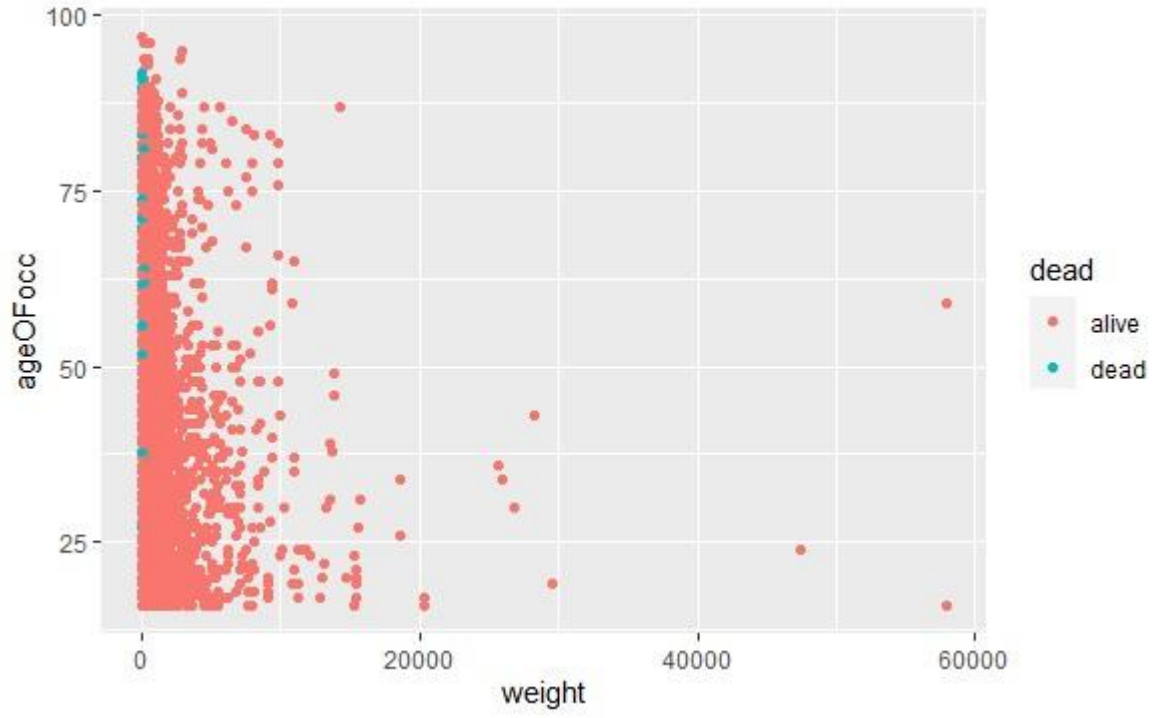
```
library(ggplot2)
ggplot(data1, aes(sex, ageOFocc)) + geom_point(aes(color = dead))
```



Most of the females who are dead are of the above 50. Men, below the age of 30 survived, but the ones who are dead, are of age >30.

Hide

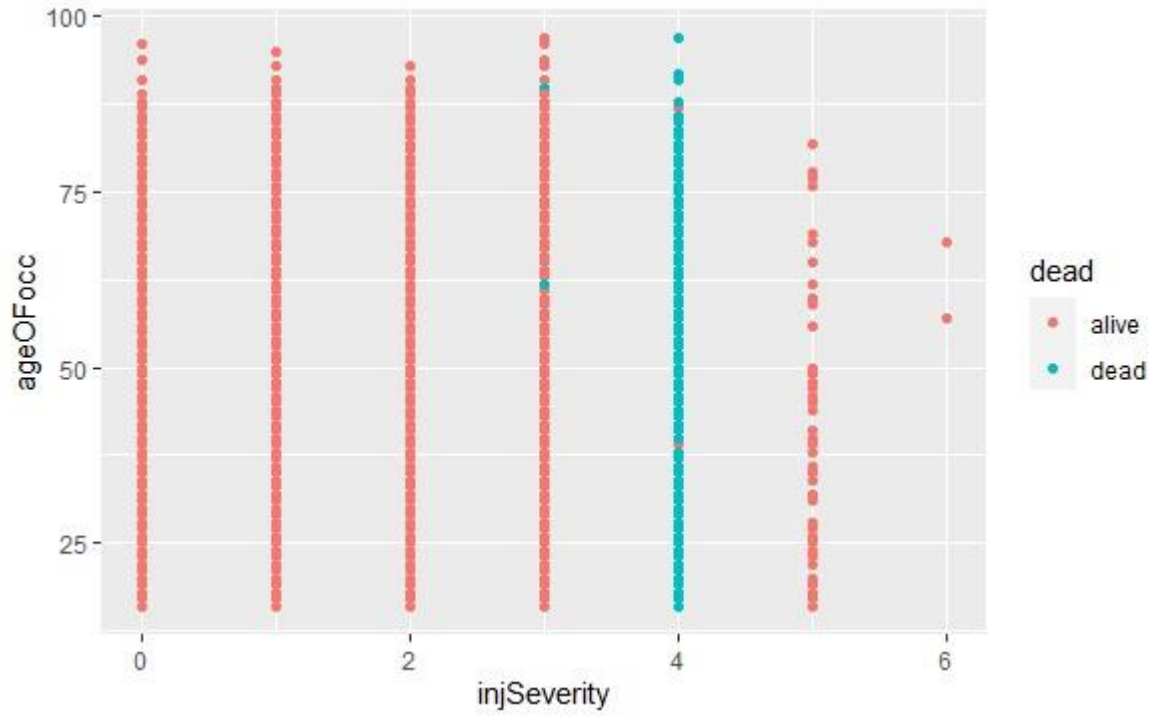
```
ggplot(data,aes(weight, ageOFocc))+geom_point(aes(color=dead))
```



Frm the above scatter plot we can find that the outliers are present

Hide

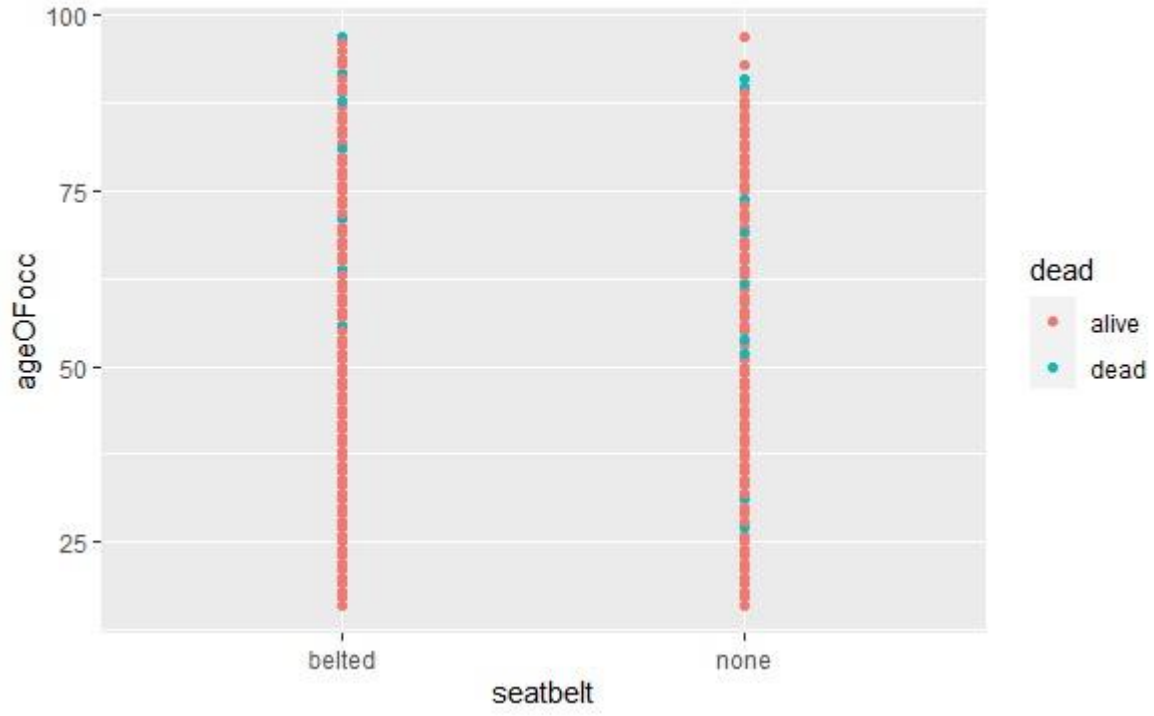
```
ggplot(data1,aes(injSeverity, ageOFocc))+geom_point(aes(color=dead))
```



InjSeverity: 0:None 1:None 2:NO Incapacity 3:Incapacity 4:Killed 5:Unknown 6:Prior Death So most of the people who are dead are in category 4.

Hide

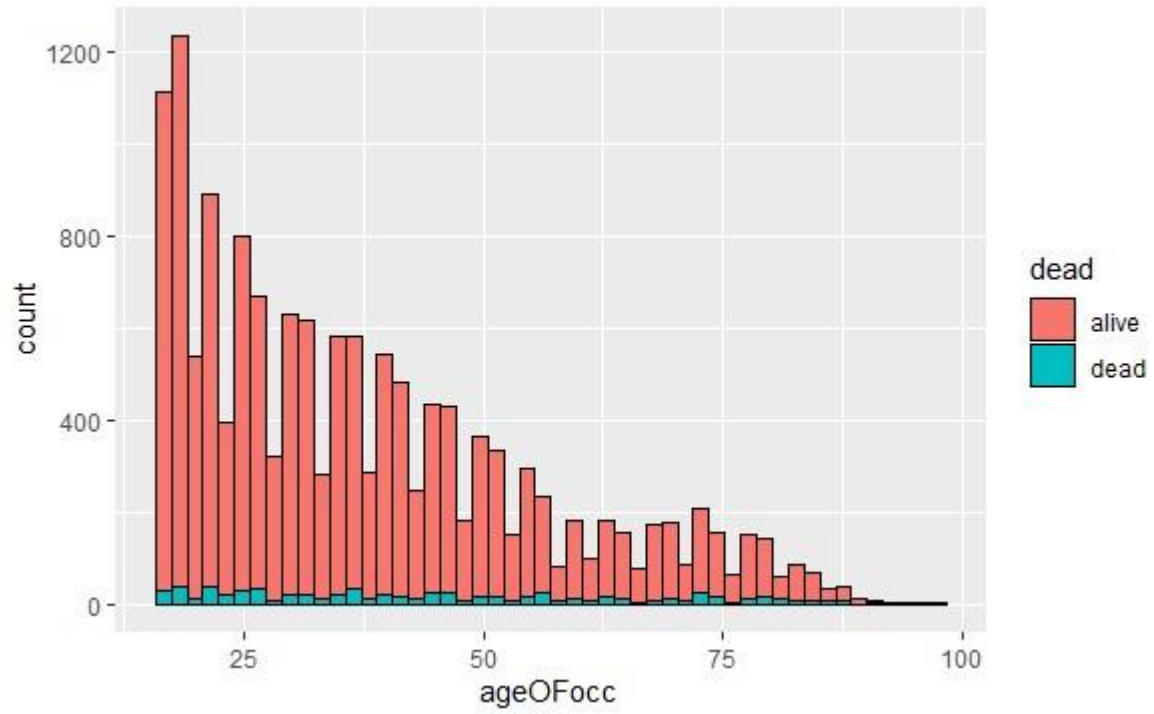
```
ggplot(data1,aes(seatbelt, ageOFocc))+geom_point(aes(color=dead))
```



Most of the alive people wore seatbelt.

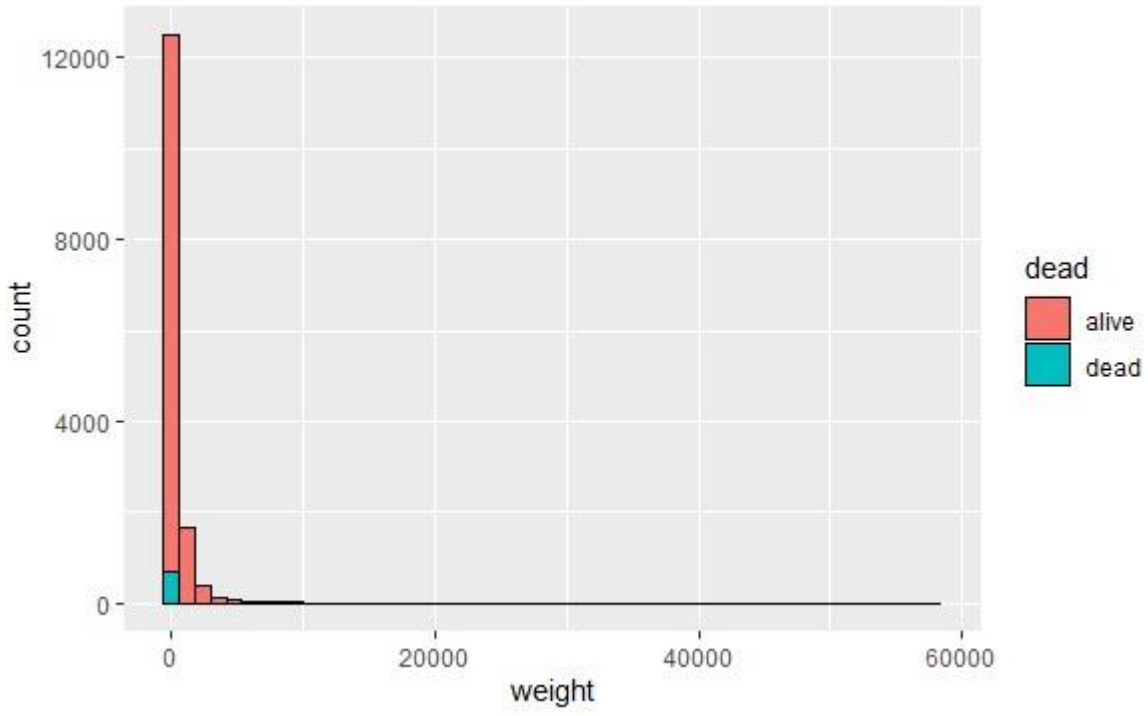
Hide

```
ggplot(data1,aes(ageOFocc))+geom_histogram(aes(fill=dead),color =   'black',bins = 50)
```



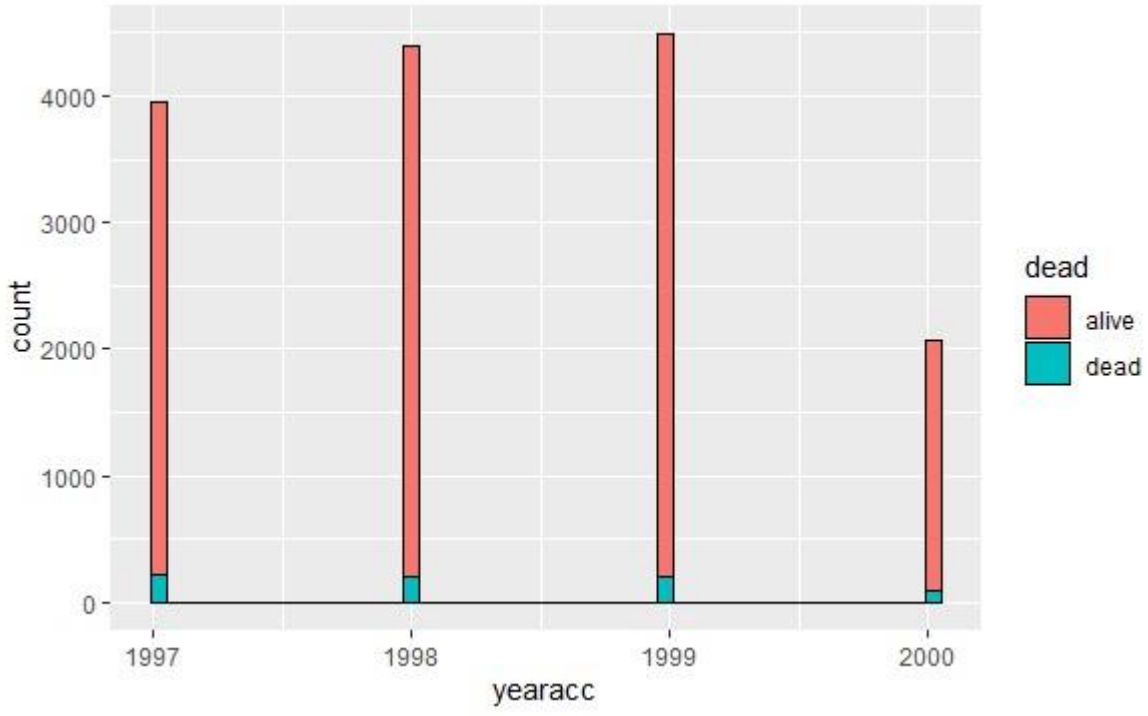
Hide

```
ggplot(data1,aes(weight))+geom_histogram(aes(fill=dead),color =   'black',bins = 50)
```



Hide

```
ggplot(data1,aes(yearacc))+geom_histogram(aes(fill=dead),color =   'black',bins = 50)
```



USing the CATOOLS lib we split our data to train and test our model

Hide

```
library(caTools)
set.seed(100)
sample = sample.split(data1$dead, SplitRatio =  0.70)
train = subset(data1, sample ==  TRUE)
test = subset(data1, sample ==  FALSE)
```

Now we train and test our model using lib RPART

Hide

```
library(rpart)
library(rpart.plot)
tree <- rpart(dead ~., method =  'class', data = train)
```

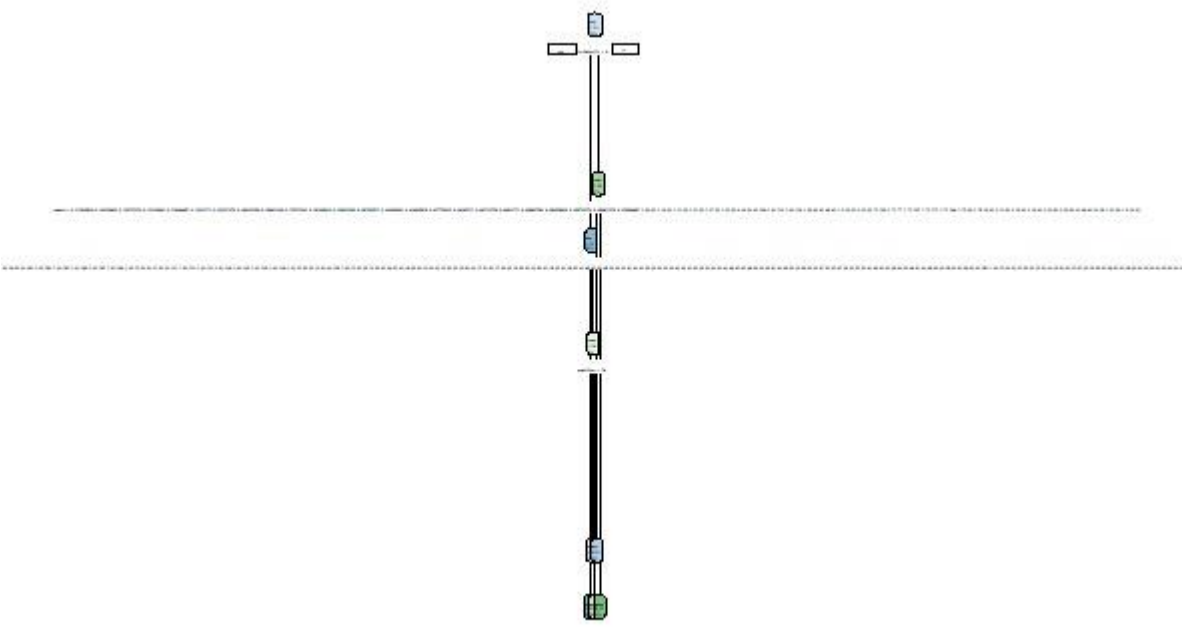
Hide

```
tree.preds <- predict(tree, test)
head(tree.preds)
```

```
      alive        dead
7  0.998987 0.001012966
13 0.998987 0.001012966
17 0.998987 0.001012966
24 0.998987 0.001012966
26 0.998987 0.001012966
29 0.998987 0.001012966
```

Hide

```
rpart.plot(tree, box.palette =  "auto")
```



Hide

```
tree.preds <- as.data.frame(tree.preds)
joiner <- function(x){
  if (x >0.5){
    return("dead")
  }else{
    return("alive")
  }
}
```

If the values in X is > 0.5, then Alive, else dead

Hide

```
tree.preds
```

	alive <dbl>	dead <dbl>
7	0.9989870	0.001012966
13	0.9989870	0.001012966
17	0.9989870	0.001012966
24	0.9989870	0.001012966
26	0.9989870	0.001012966
29	0.9989870	0.001012966
30	0.9989870	0.001012966
34	0.9989870	0.001012966
36	0.9989870	0.001012966
37	0.9989870	0.001012966

1-10 of 4,477 rows Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
tree.preds$dead <- sapply(tree.preds$`dead`, joiner)
head(tree.preds)
```

	alive <dbl>	dead <chr>
7	0.998987	alive
13	0.998987	alive
17	0.998987	alive
24	0.998987	alive
26	0.998987	alive
29	0.998987	alive

6 rows

We can understand our result better through the confusion Matrix

Hide

```
library(caret)
cf<-table(tree.preds$dead, test$dead)
confusionMatrix(cf, positive =  "dead")
```

```
Confusion Matrix and Statistics


       alive dead
  alive  4245    9
  dead     17  206

               Accuracy : 0.9942
                 95% CI : (0.9915, 0.9962)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9376

 Mcnemar's Test P-Value : 0.1698

            Sensitivity : 0.95814
            Specificity : 0.99601
         Pos Pred Value : 0.92377
         Neg Pred Value : 0.99788
             Prevalence : 0.04802
         Detection Rate : 0.04601
   Detection Prevalence : 0.04981
      Balanced Accuracy : 0.97708

       'Positive' Class : dead
```

Accuracy is 99.42%. Kappa Value is 93.76%.