**TITLE : Predicting the age of abalone**

**ROLL NO: 1)17P106**

**2)17E105**

**NAME: 1)ARUNPRASATH P**

**2)ARUN KUMAR**

**DEPT: 1)CSE**

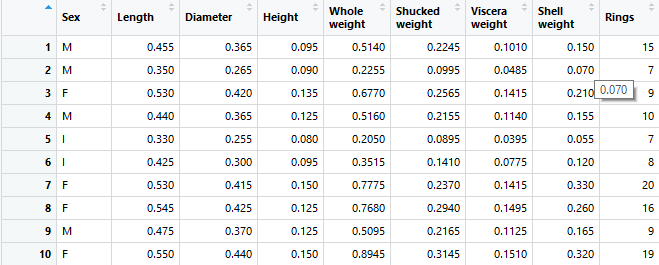
**2)EEE**

LOAD THE DATASET

dataset <-read.delim("abalone.data",sep=",",header = F)

colnames(dataset)=c("Sex","Length","Diameter","Height","Whole weight","Shucked weight","Viscera weight","Shell weight","Rings")

View(dataset)



EXPLORATORY DATA ANALYSIS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| library(dplyr)   |  | | --- | | > str(dataset)  'data.frame': 4177 obs. of 9 variables:  $ Sex : Factor w/ 3 levels "F","I","M": 3 3 1 3 2 2 1 1 3 1 ...  $ Length : num 0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...  $ Diameter : num 0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...  $ Height : num 0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...  $ Whole weight : num 0.514 0.226 0.677 0.516 0.205 ...  $ Shucked weight: num 0.2245 0.0995 0.2565 0.2155 0.0895 ...  $ Viscera weight: num 0.101 0.0485 0.1415 0.114 0.0395 ...  $ Shell weight : num 0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...  $ Rings : int 15 7 9 10 7 8 20 16 9 19 ... | |  | | |  | | --- | |  | |   > summary(dataset)  Sex Length Diameter Height Whole weight Shucked weight  F:1307 Min. :0.075 Min. :0.0550 Min. :0.0000 Min. :0.0020 Min. :0.0010  I:1342 1st Qu.:0.450 1st Qu.:0.3500 1st Qu.:0.1150 1st Qu.:0.4415 1st Qu.:0.1860  M:1528 Median :0.545 Median :0.4250 Median :0.1400 Median :0.7995 Median :0.3360  Mean :0.524 Mean :0.4079 Mean :0.1395 Mean :0.8287 Mean :0.3594  3rd Qu.:0.615 3rd Qu.:0.4800 3rd Qu.:0.1650 3rd Qu.:1.1530 3rd Qu.:0.5020  Max. :0.815 Max. :0.6500 Max. :1.1300 Max. :2.8255 Max. :1.4880  Viscera weight Shell weight Rings  Min. :0.0005 Min. :0.0015 Min. : 1.000  1st Qu.:0.0935 1st Qu.:0.1300 1st Qu.: 8.000  Median :0.1710 Median :0.2340 Median : 9.000  Mean :0.1806 Mean :0.2388 Mean : 9.934  3rd Qu.:0.2530 3rd Qu.:0.3290 3rd Qu.:11.000  Max. :0.7600 Max. :1.0050 Max. :29.000 |
|  |

quantile(dataset$Rings)

> quantile(dataset$Rings)

0% 25% 50% 75% 100%

1 8 9 11 29

**TARGET ATTRIBUTES**

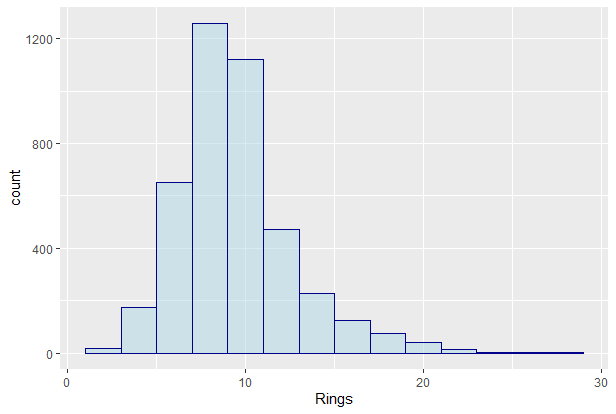
Target attribute is Rings

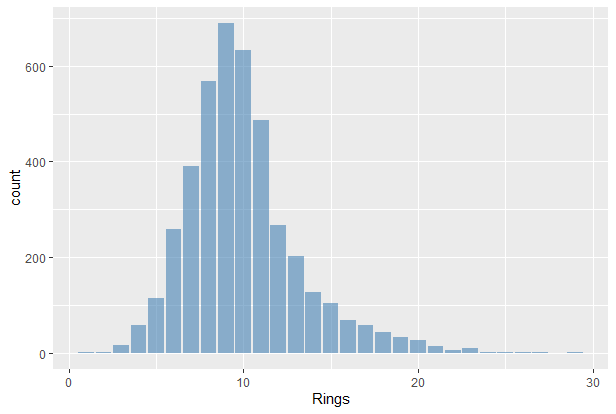
FREQUENCY BASED ON RINGS

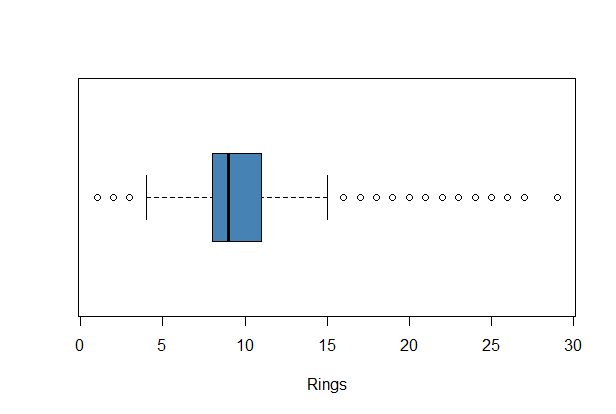
dataset %>% ggplot(aes(x=Rings))+geom\_histogram(fill="lightblue",color="Darkblue",binwidth = 2,alpha=0.5)

dataset %>% ggplot(aes(x=Rings))+geom\_bar(fill="steelblue",alpha=0.6)

boxplot(dataset$Rings,col = "steelblue",horizontal = TRUE,border="black",xlab=("Rings"))







Thus the most frequent values of the attribute Rings are highly concentrated around the median of the distribution value of 9.

**SEX ATTRIBUTE**

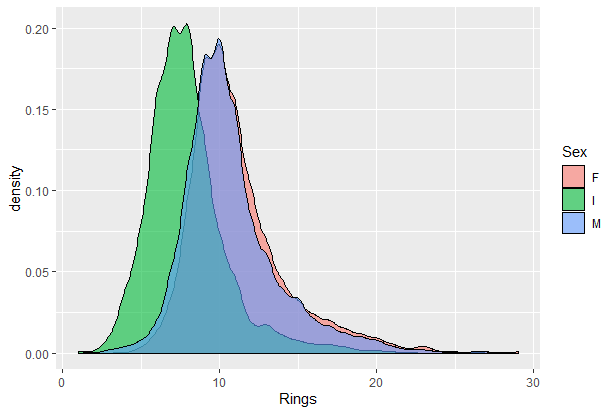
> table(dataset$Sex)

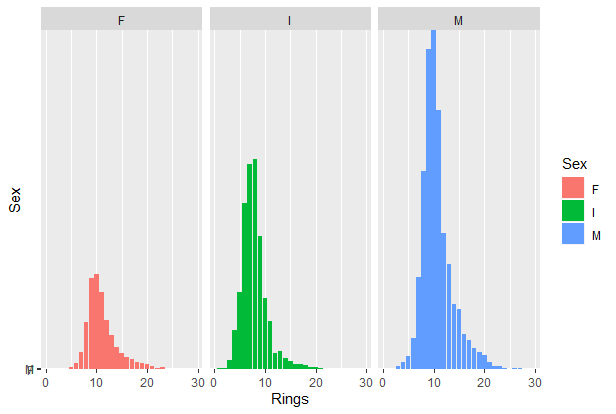
F I M

1307 1342 1528

ggplot(dataset) + aes(Rings, fill = Sex) + geom\_density(alpha=0.6)

ggplot(dataset,aes(x=Rings,y=Sex,fill=Sex))+geom\_col()++facet\_grid(~Sex)





From the above plots we know that the female abalone has longer lifespan than others but the count of male abalone is maximum than others.

**WEIGHT ATTRIBUTE**

**Whole weight**

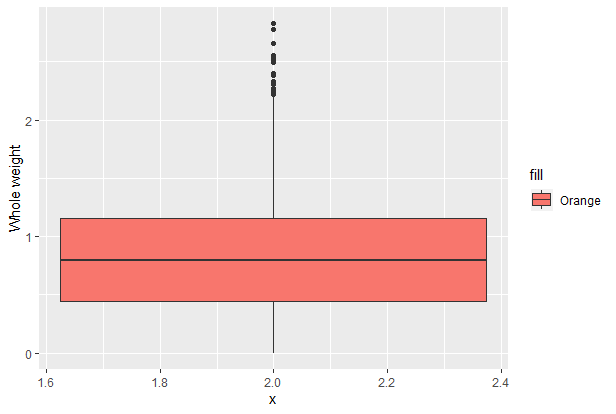
> summary(dataset$`Whole weight`)

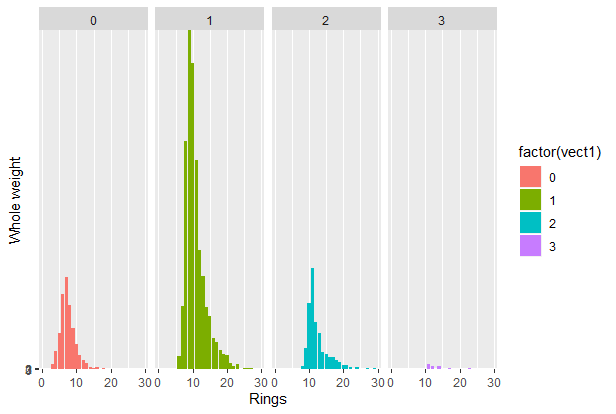
Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0020 0.4415 0.7995 0.8287 1.1530 2.8255

ggplot(dataset,aes(y=`Whole weight`,x=2,fill="Orange"))+geom\_boxplot()

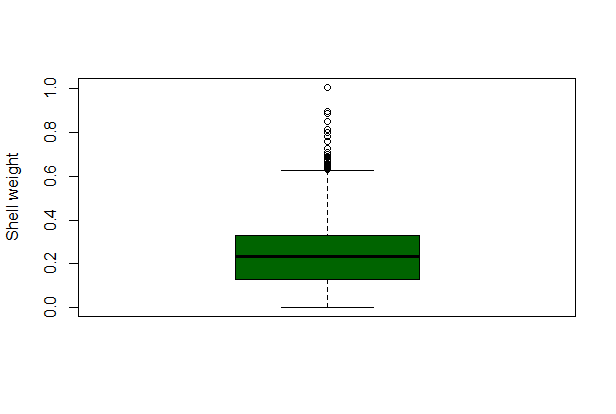
ggplot(dataset,aes(y=factor(vect1),x=Rings,fill=factor(vect1)))+geom\_col()+facet\_grid(~factor(vect1))+ylab("Whole weight")

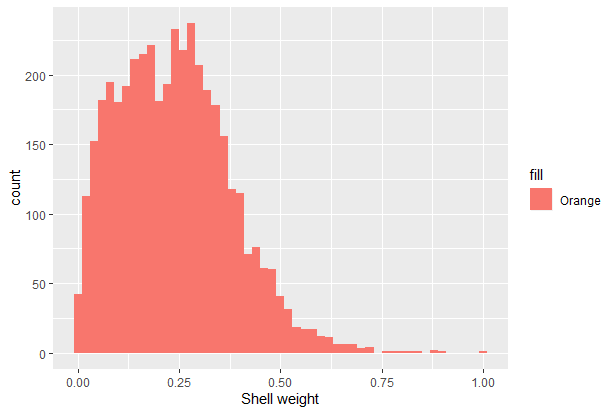




**Shell weight**

boxplot(dataset$`Shell weight`,col = "darkgreen",border="black",ylab="Shell weight")

ggplot(dataset,aes(x=`Shell weight`,fill="Orange"))+geom\_bar(binwidth = 0.02)



From the above plots we know that the weight of abalone (1.5 to 2.4)gm has longer lifespan than others but the weight of abalone (0.5 to 1.4)gm has maximum count than others.

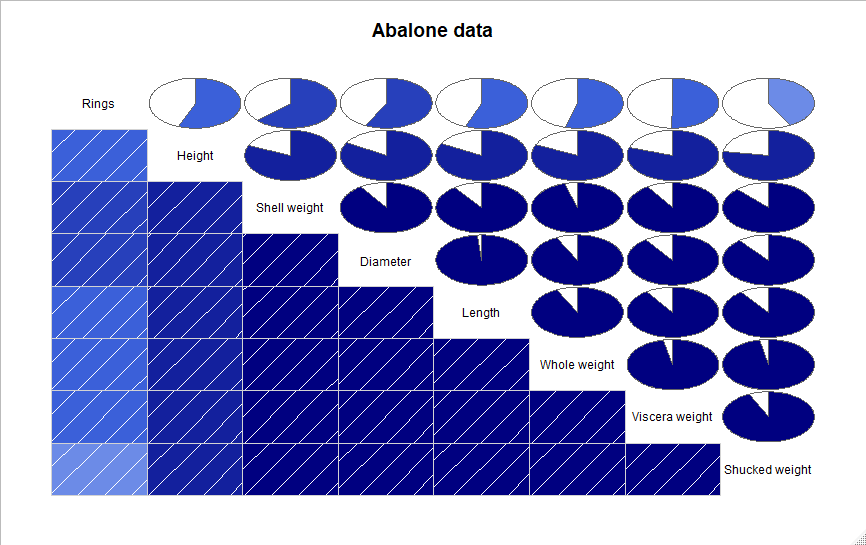
**CORELATION BETWEEN THE DATA**

library(corrplot)

corrgram(dataset,order=T,lower.panel=panel.shade,

upper.panel=panel.pie, text.panel=panel.txt,

main="Abalone data")



**PREDICTING THE AGE OF ABALONE**

**Correlation analysis of other variables and rings**

cor(dataset$Rings,dataset$Length)

> cor(dataset$Rings,dataset$Length)

[1] 0.5567196

cor(dataset$Rings,dataset$Diameter)

> cor(dataset$Rings,dataset$Diameter)

[1] 0.5746599

cor(dataset$Rings,dataset$`Whole weight`)

> cor(dataset$Rings,dataset$`Whole weight`)

[1] 0.5403897

cor(dataset$Rings,dataset$`Shell weight`)

> cor(dataset$Rings,dataset$`Shell weight`)

[1] 0.627574

clean\_dataset<-dataset %>% mutate(age=case\_when(Rings %in% 1:5~"young",Rings %in% 6:13~"adult",Rings %in% 14:30~"old"))

clean\_dataset<-clean\_dataset %>% select(c(2,3,5,8,10)) %>% na.omit()

clean\_dataset

library(caTools)

set.seed(123)

sample=sample.split(clean\_dataset,SplitRatio = 0.8)

traindata<-subset(clean\_dataset,sample==T)

traindata

testdata<-subset(clean\_dataset,sample==F)

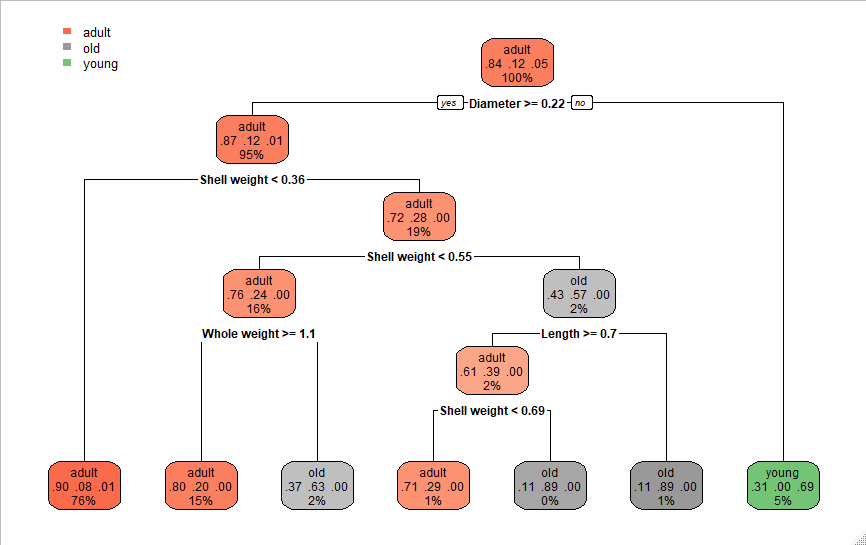
library(rpart)

model <-rpart(formula=age~.,data=traindata,method="class")

summary(model)

library(rpart.plot)

rpart.plot(model)



predicted\_model<-predict(object=model,newdata=testdata,type="class")

confusion.matrix<-table(testdata$age,predicted\_model)

confusion.matrix

> confusion.matrix

predicted\_model

adult old young

adult 675 6 17

old 89 11 0

young 8 0 29

accuracy<-sum(diag(confusion.matrix))/sum(confusion.matrix)

accuracy

> accuracy

[1] 0.8562874

Our predicted model satisfies the 85% accuracy of solution

**CONCLUSION**

Thus the age of abalone cannot be find by the appearances.