111023 Statistical Learning

Mattia G.

2023-10-11

R commands notes 2

```
weight <- c(80, 70, 82, 76, 90)
height <- c(170, 198, 176, 181, 180)
smoker <- c("yes", "yes", "no", "no", "yes")
survey <- data.frame(weight, height, smoker)

survey
```

```
##   weight height smoker
## 1     80    170    yes
## 2     70    198    yes
## 3     82    176     no
## 4     76    181     no
## 5     90    180    yes
```

```
M <- matrix(1:9, ncol = 3)

#### list
# same as vectors but can contain any type

Lst <- list("Fred", 3, c(4, 7, 9), M, survey)

class(Lst)
```

```
## [1] "list"
```

```
is.list(Lst)
```

```
## [1] TRUE
```

```
Lst[1:3]
```

```
## [[1]]
## [1] "Fred"
##
## [[2]]
## [1] 3
##
## [[3]]
## [1] 4 7 9
```

```
SubLst <- Lst[4]
SubLst
```

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

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
is.matrix(SubLst)
```

```
## [1] FALSE
```

```
A <- Lst[[4]]
is.list(A)
```

```
## [1] FALSE
```

```
is.matrix(A)
```

```
## [1] TRUE
```

```
Lst <- list(name = "Fred", n.child = 3, child.age = c(4, 7, 9), my.matrix = M,
            my.data = survey)

Lst$n.child
```

```
## [1] 3
```

```
A <- Lst$my.matrix
A
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
A <- Lst[[4]]
A
```

```
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
```

```
#### function data()

data() # it can be utilized for uploading data

# if I input data() and press return I will get a list of data sets included in R
```

```
# Diameter, Height and Volume for Black Cherry Trees

data("trees")
View(trees) #to view the dataset


#### data cars

library(MASS) # package
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
```

```
##
##     survey
```

```
data(package = "MASS")

data("Cars93")
View(Cars93)

names(Cars93)
```

```
##  [1] "Manufacturer"       "Model"              "Type"
##  [4] "Min.Price"          "Price"              "Max.Price"
##  [7] "MPG.city"           "MPG.highway"        "AirBags"
## [10] "DriveTrain"         "Cylinders"          "EngineSize"
## [13] "Horsepower"         "RPM"                "Rev.per.mile"
## [16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
## [19] "Length"             "Wheelbase"          "Width"
## [22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
## [25] "Weight"             "Origin"             "Make"
```

```
Cars93[1:5, 3]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```
Cars93[1:5,"Type"]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```
library(MASS) # package


data(package = "MASS")

data("Cars93")
View(Cars93)

names(Cars93)
```

```
##  [1] "Manufacturer"       "Model"              "Type"
##  [4] "Min.Price"          "Price"              "Max.Price"
##  [7] "MPG.city"           "MPG.highway"        "AirBags"
## [10] "DriveTrain"         "Cylinders"          "EngineSize"
## [13] "Horsepower"         "RPM"                "Rev.per.mile"
## [16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
## [19] "Length"             "Wheelbase"          "Width"
## [22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
## [25] "Weight"             "Origin"             "Make"
```

```
Cars93[1:5, 3]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```
Cars93[1:5,"Type"]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```r
Cars93$Type[1:5]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```r
#Type[1:5]  Does not work because the ds hasn't been attached

attach(Cars93) # attaching the ds
Type[1:5]
```

```
## [1] Small   Midsize Compact Midsize Midsize
## Levels: Compact Large Midsize Small Sporty Van
```

```r
detach(Cars93) # detaching ds
```

End of introduction

Begininng of R instructions from the slides

```r
attach(Cars93)
table(Type)
```

```
## Type
## Compact   Large Midsize   Small  Sporty     Van
##      16      11      22      21      14       9
```

```r
freq.tb.Type <- table(Type)
freq.tb.Type
```

```
## Type
## Compact   Large Midsize   Small  Sporty     Van
##      16      11      22      21      14       9
```

```r
freq.tb.Type["Large"]
```

```
## Large
##    11
```

```r
rel.freq.tb.Type <- freq.tb.Type/sum(freq.tb.Type)
rel.freq.tb.Type
```

```
## Type
##    Compact      Large    Midsize      Small     Sporty        Van
## 0.17204301 0.11827957 0.23655914 0.22580645 0.15053763 0.09677419
```

```r
round(rel.freq.tb.Type, digits = 2)
```

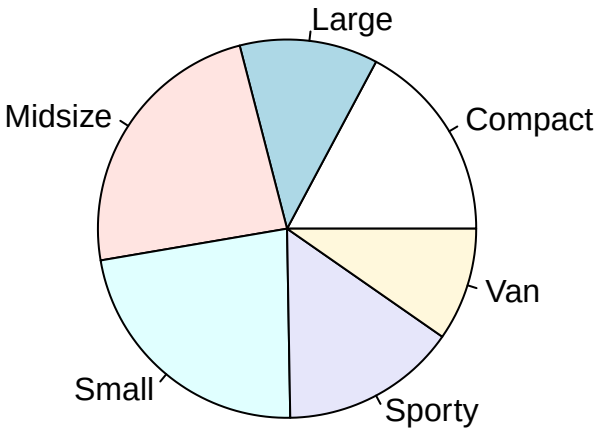
```
## Type
## Compact   Large Midsize   Small  Sporty     Van
##    0.17    0.12    0.24    0.23    0.15    0.10
```

```r
library(xtable) #needs to be installed first
xtable(freq.tb.Type)
```

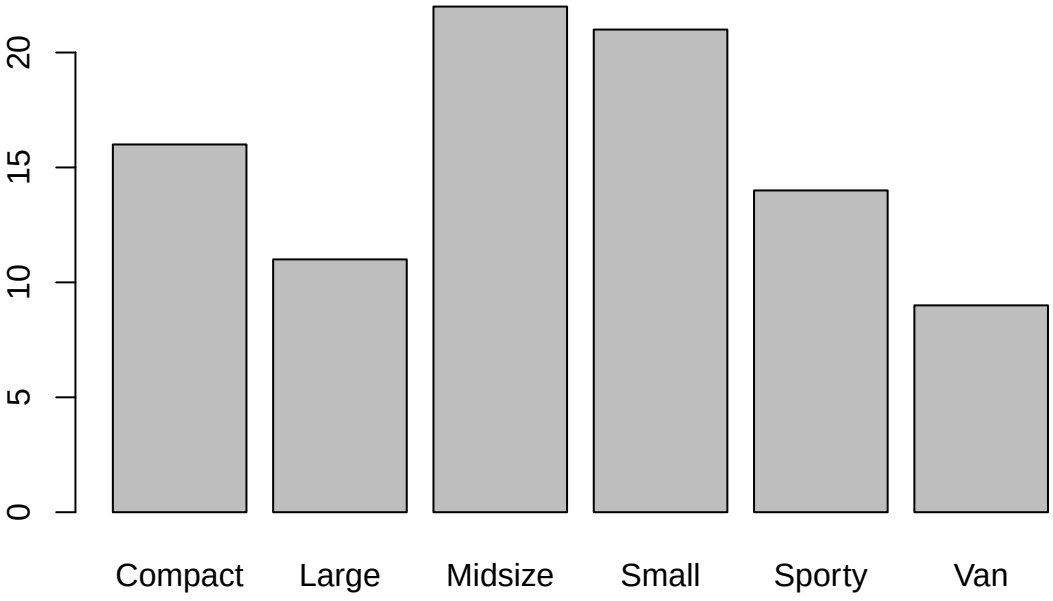
```
## % latex table generated in R 4.1.2 by xtable 1.8-4 package
## % Wed Oct 11 11:53:21 2023
## \begin{table}[ht]
## \centering
```

```
## \begin{tabular}{rr}
##   \hline
##  & Type \\
##   \hline
## Compact &  16 \\
##   Large &  11 \\
##   Midsize &  22 \\
##   Small &  21 \\
##   Sporty &  14 \\
##   Van &   9 \\
##    \hline
## \end{tabular}
## \end{table}
```

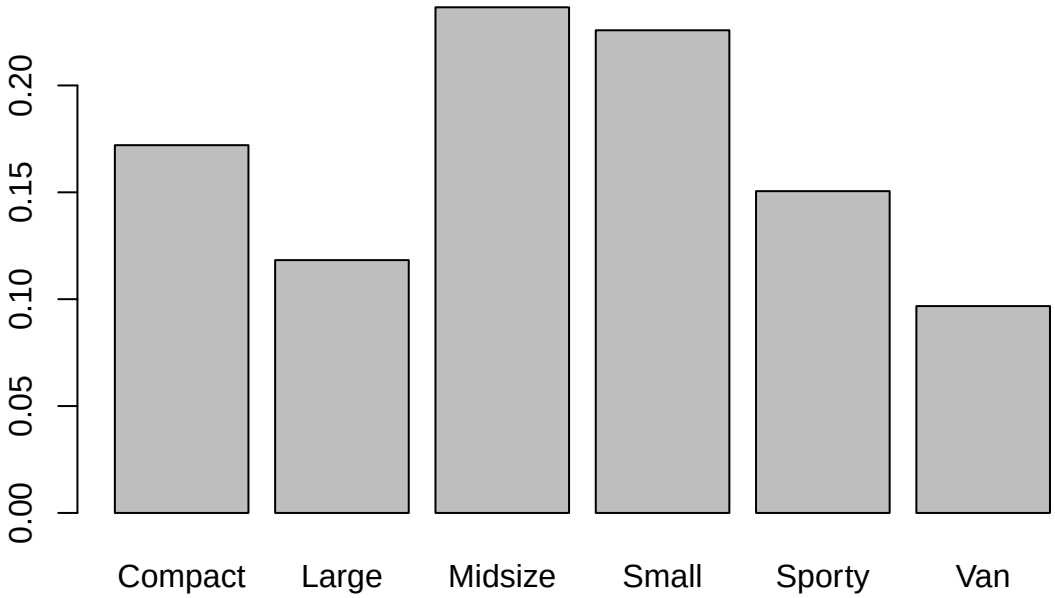
```
#### pie chart

pie(freq.tb.Type)
```




```
help("pie")

barplot(freq.tb.Type)
```



```
barplot(rel.freq.tb.Type)
```



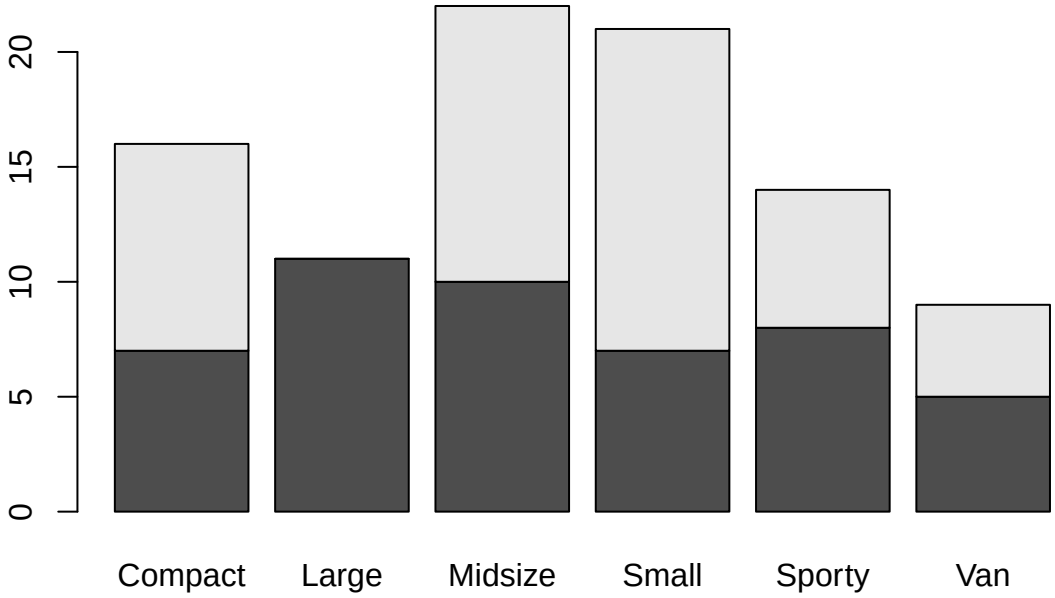

```
#### plotting more variables

tb.Origin.Type <- table(Origin, Type)
tb.Origin.Type
```

```
##          Type
## Origin    Compact Large Midsize Small Sporty Van
##   USA           7    11      10     7      8   5
##   non-USA       9     0      12    14      6   4
```

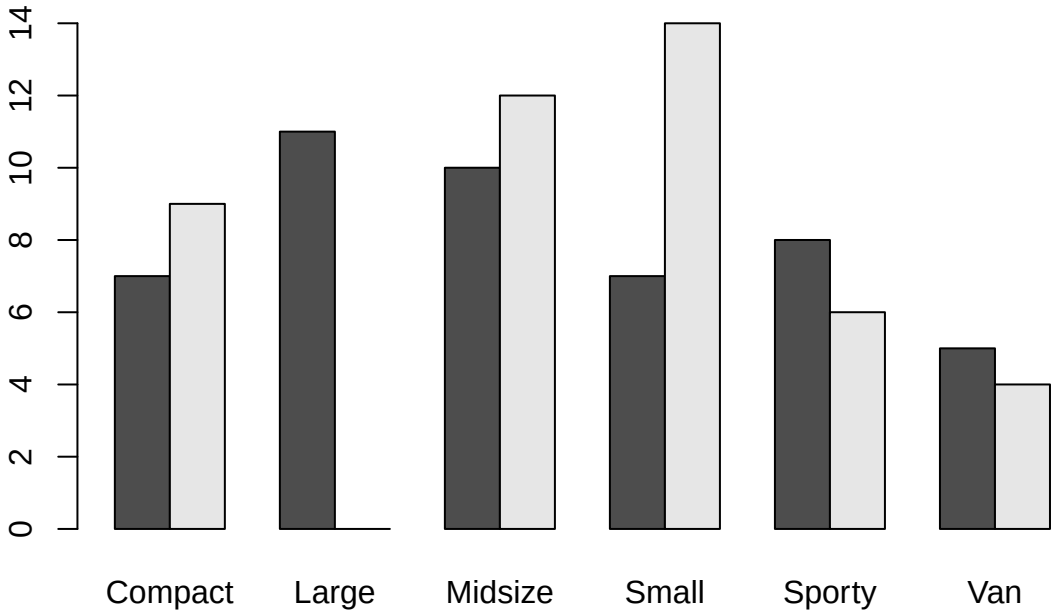
```
tb.Origin.Type["USA", "Large"]
```

```
## [1] 11
```

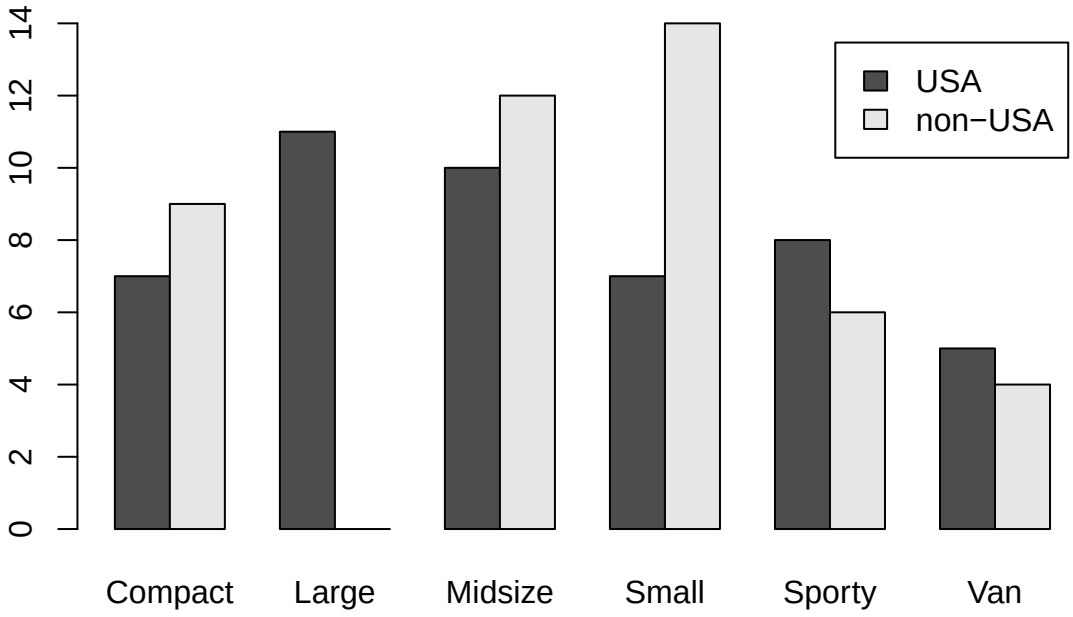
```
barplot(tb.Origin.Type)
```



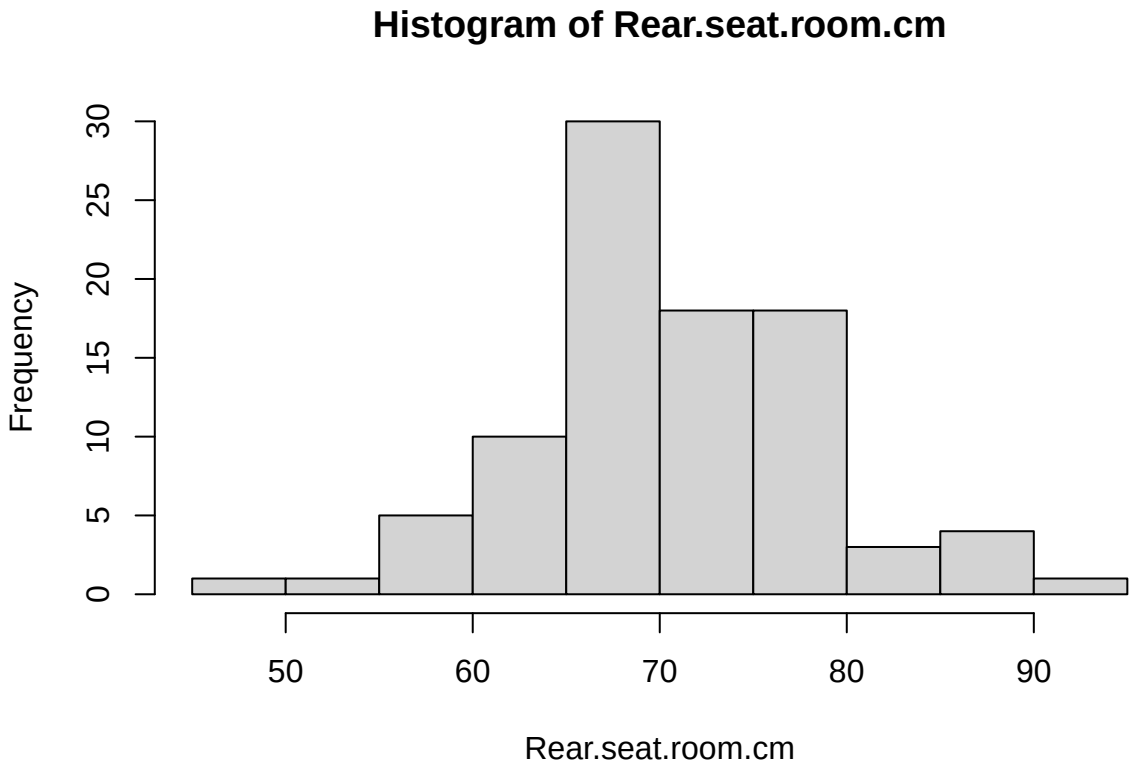
```
barplot(tb.Origin.Type, beside = TRUE)
```



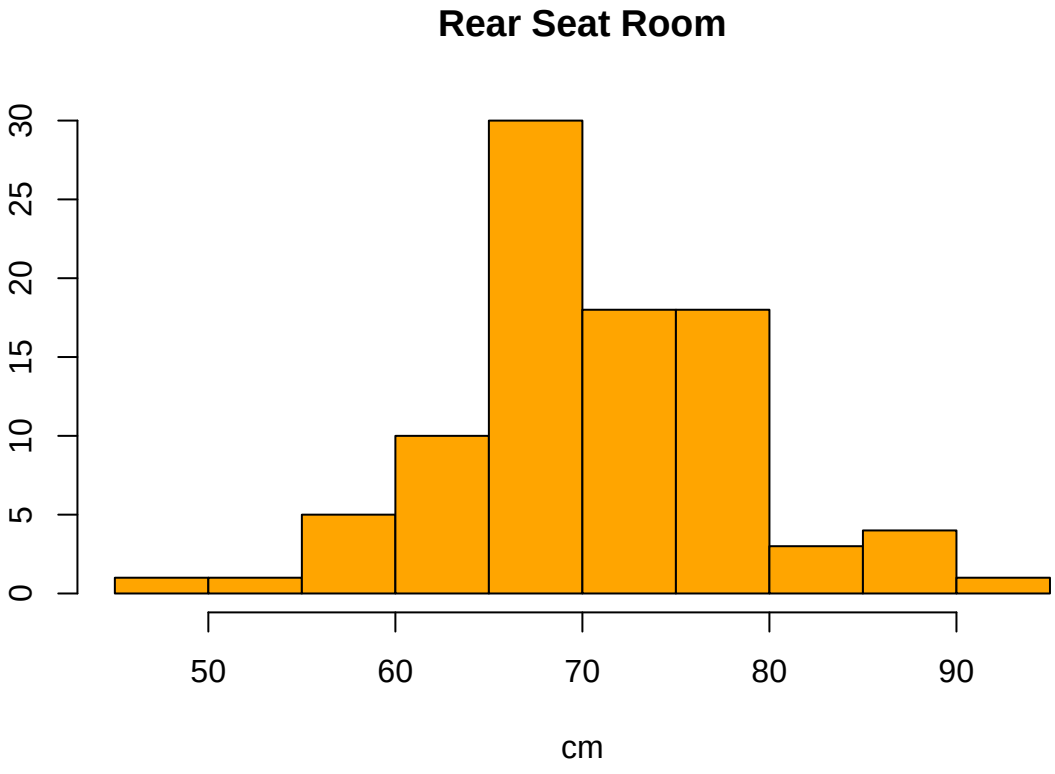

```
#### add legend

barplot(tb.Origin.Type, beside = TRUE, legend = TRUE)
```



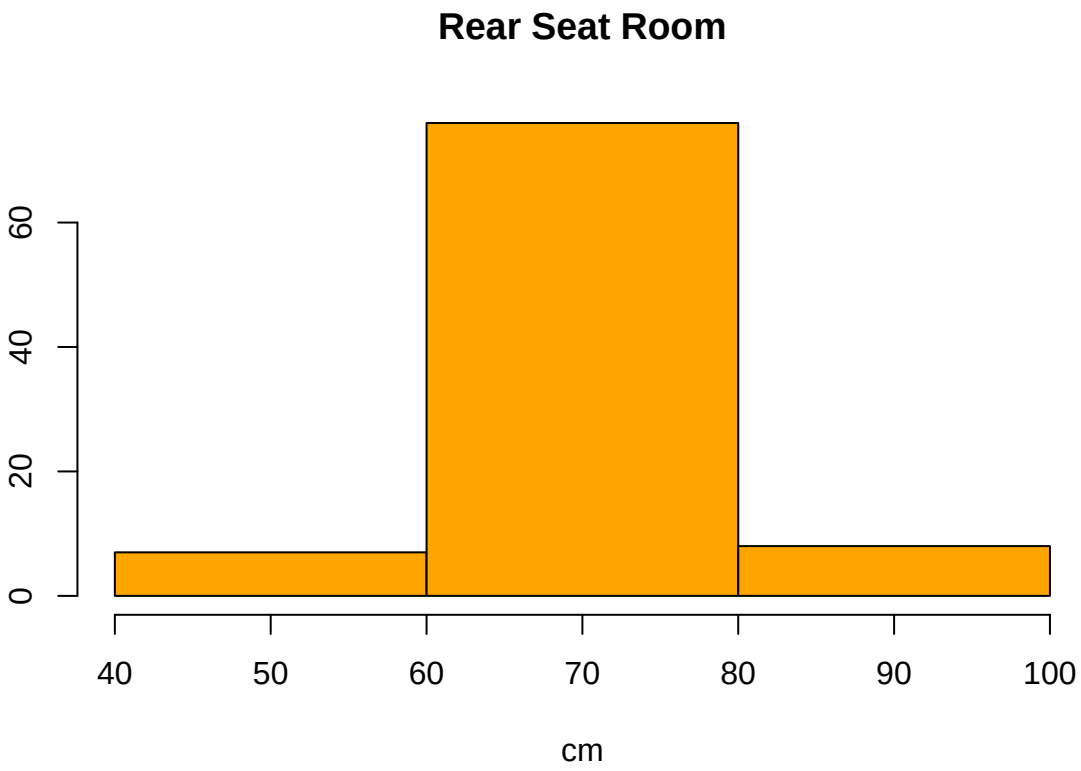

```
help("barplot")

#### histograms

Rear.seat.room.cm <- Rear.seat.room * 2.54
hist(Rear.seat.room.cm)
```



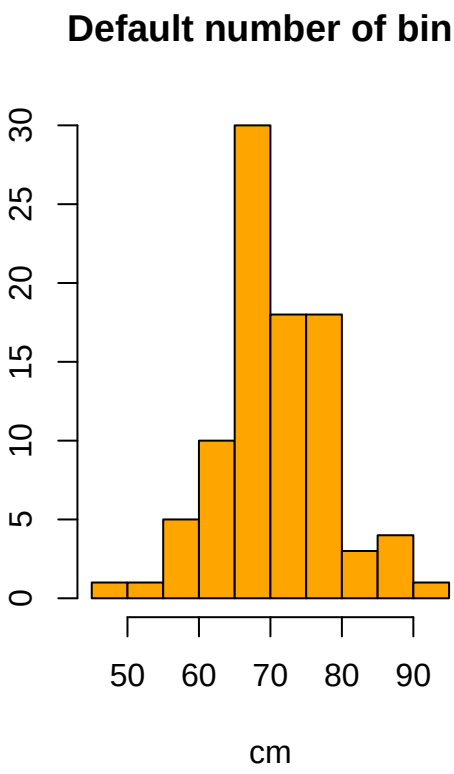

```
#### making it look nicer (why not)

hist(Rear.seat.room.cm, xlab = "cm", ylab = "",
     main = "Rear Seat Room", col = "orange")
```

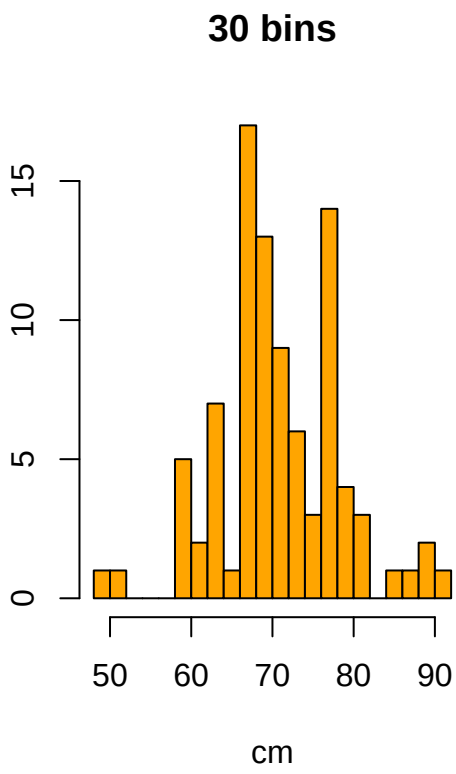


```r
#### change number of bins

hist(Rear.seat.room.cm, breaks = 3, xlab = "cm", ylab = "",
     main = "Rear Seat Room", col = "orange")
```




```r
hist(Rear.seat.room.cm, xlab = "cm", ylab = "",
     main = "Default number of bins", col = "orange")

par(mfrow = c(1, 2))

hist(Rear.seat.room.cm, xlab = "cm", ylab = "",
     main = "Default number of bins", col = "orange")

hist(Rear.seat.room.cm, breaks = 30, xlab = "cm", ylab = "",
     main = "30 bins", col = "orange")
```






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

range(Rear.seat.room.cm)
```

```
## [1] NA NA
```

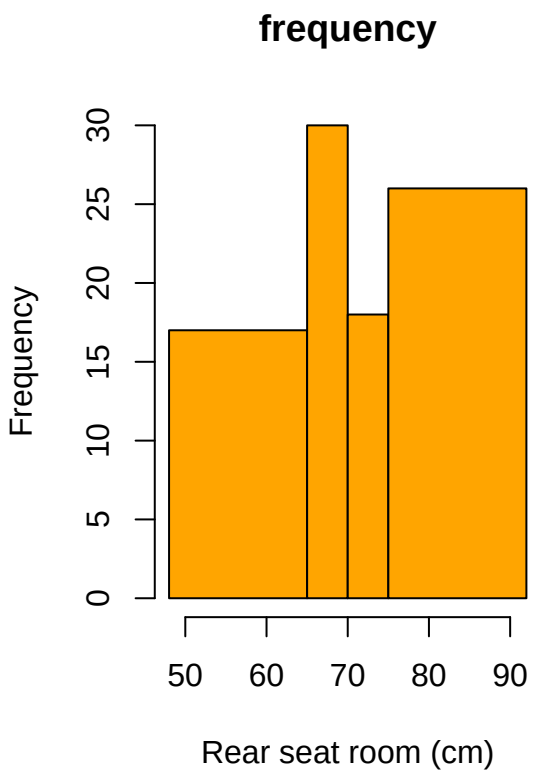
```
range(Rear.seat.room.cm, na.rm = TRUE)
```

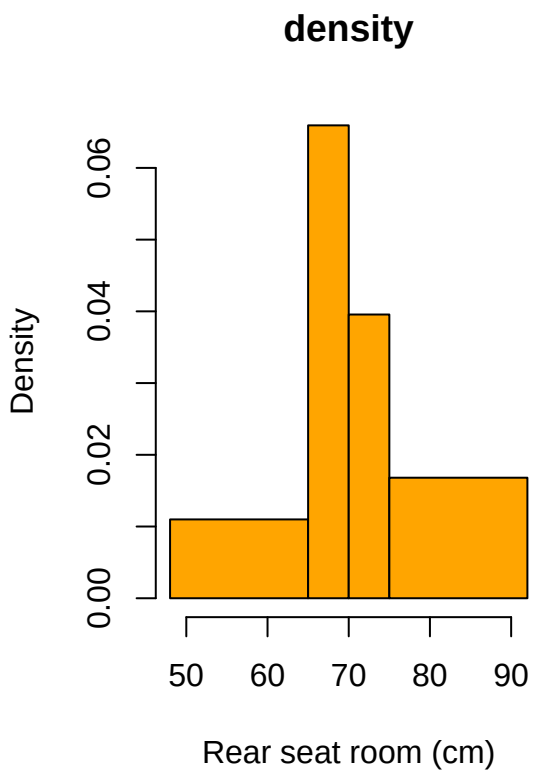
```
## [1] 48.26 91.44
```

```
bins <- c(48, 65, 70, 75, 92)

par(mfrow = c(1, 2))

hist(Rear.seat.room.cm, freq=TRUE, xlab="Rear seat room (cm)",
     main="frequency", breaks =bins, col="orange")
```

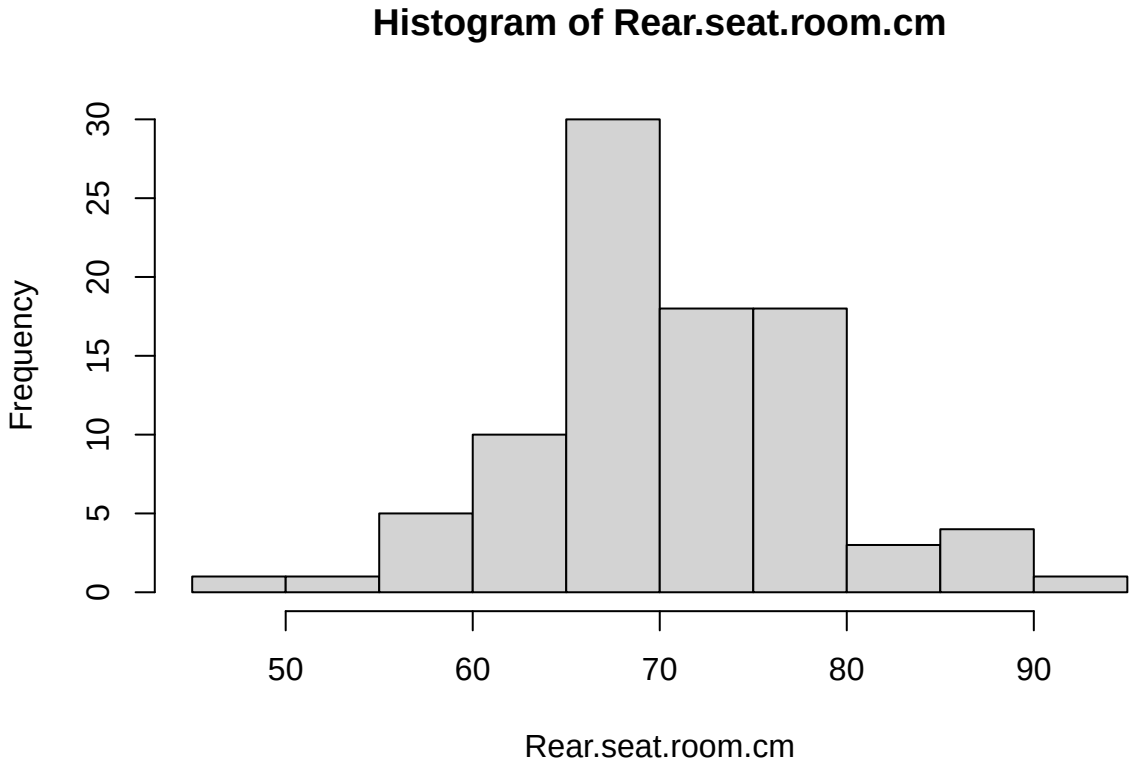
```
## Warning in plot.histogram(r, freq = freq1, col = col, border = border, angle =
## angle, : the AREAS in the plot are wrong -- rather use 'freq = FALSE'
```

```
hist(Rear.seat.room.cm, freq=FALSE,  xlab="Rear seat room (cm)",
     main="density", breaks=bins, col="orange")
```





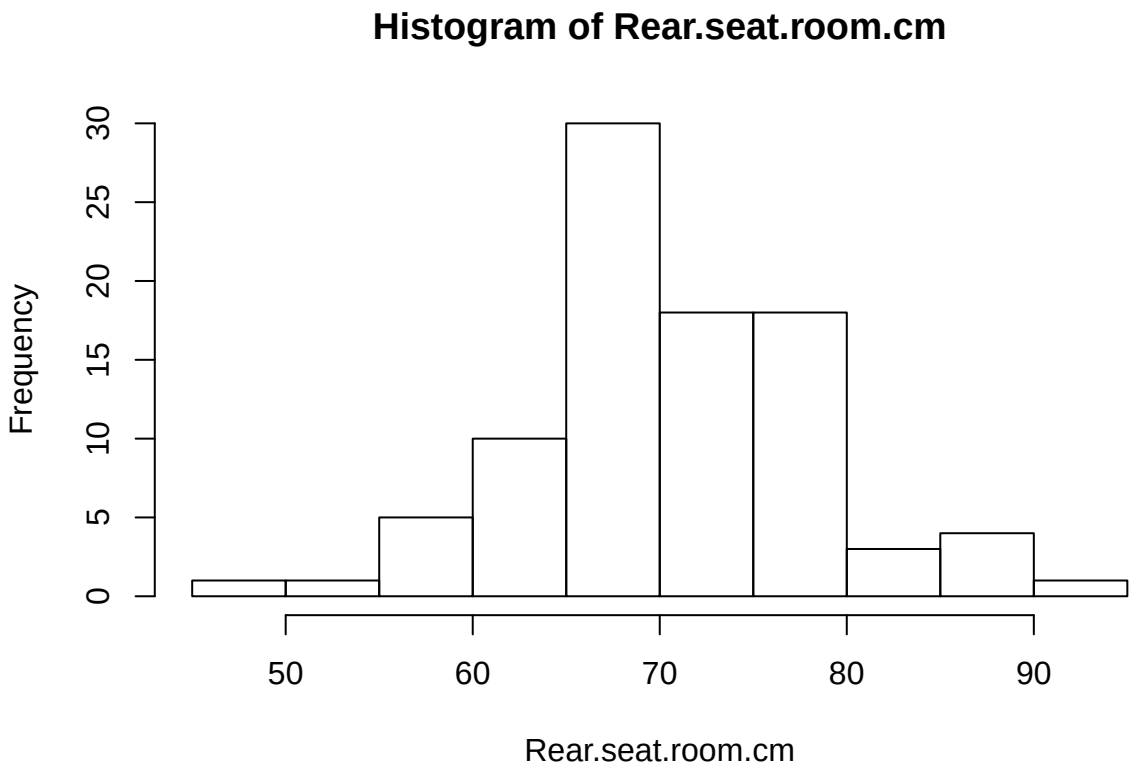

```
par(mfrow=c(1,1))


#### output of the function hist()

hist(Rear.seat.room.cm)

h <- hist(Rear.seat.room.cm)
```



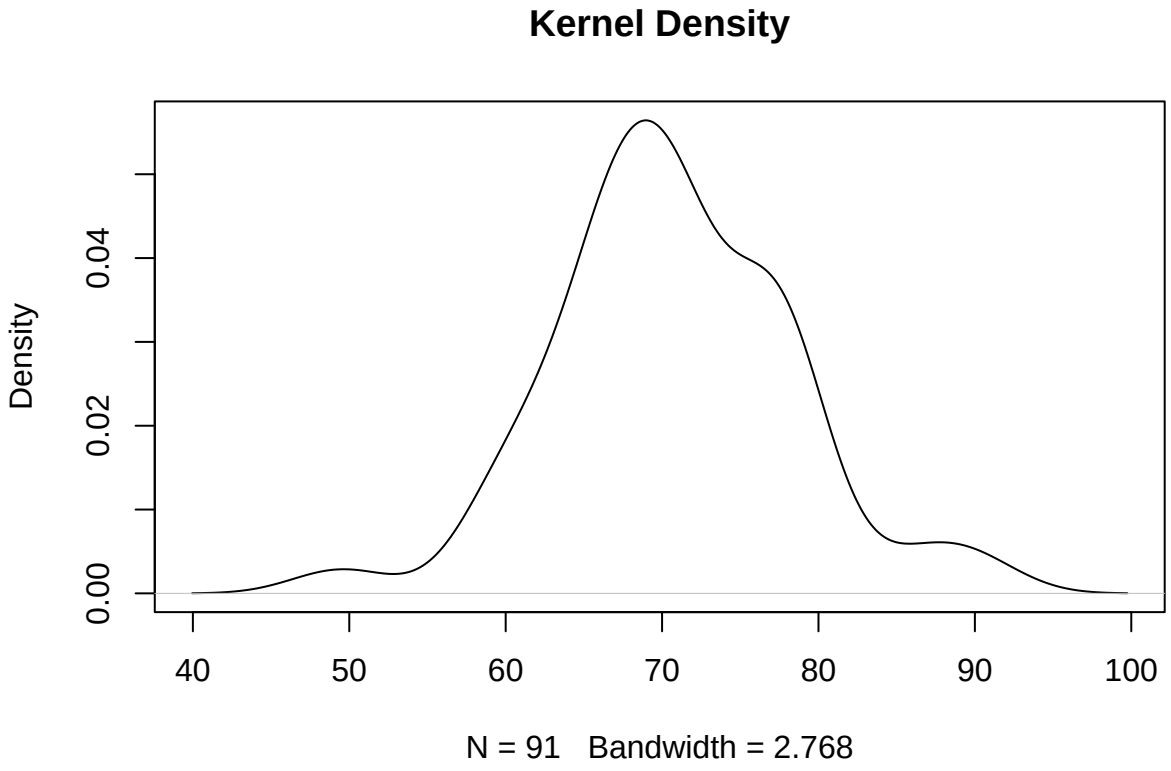

```
h
```

```
## $breaks
##  [1] 45 50 55 60 65 70 75 80 85 90 95
##
## $counts
##  [1]  1  1  5 10 30 18 18  3  4  1
##
## $density
##  [1] 0.002197802 0.002197802 0.010989011 0.021978022 0.065934066 0.039560440
##  [7] 0.039560440 0.006593407 0.008791209 0.002197802
##
## $mids
##  [1] 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5
##
## $xname
## [1] "Rear.seat.room.cm"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

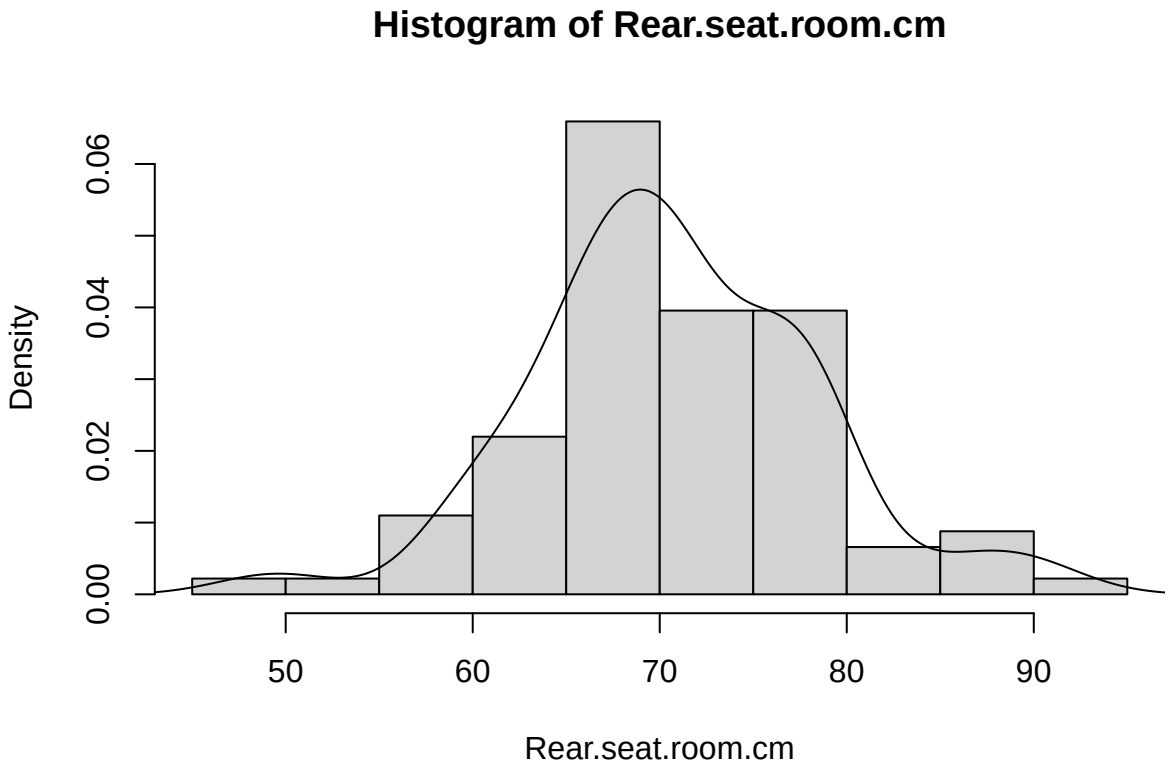
```
plot(h)
```



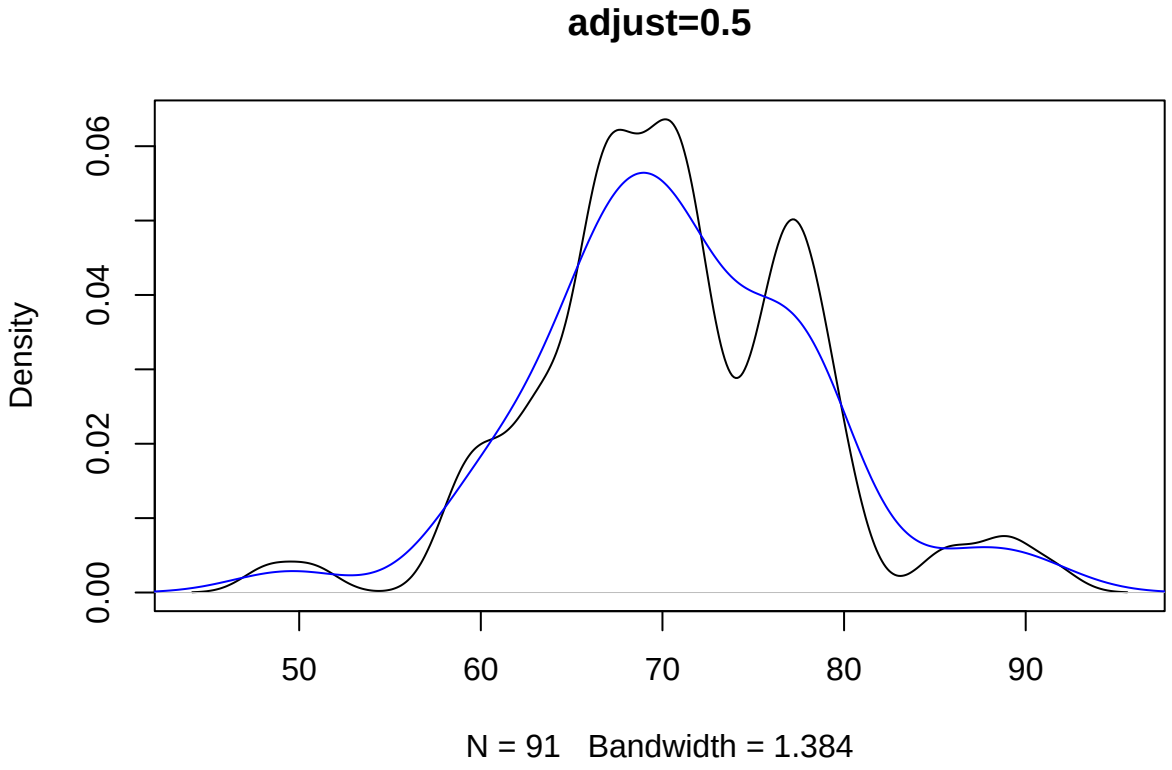

```
#### density plot

dens <- density(Rear.seat.room.cm, na.rm=TRUE)
plot(dens, main="Kernel Density")
```



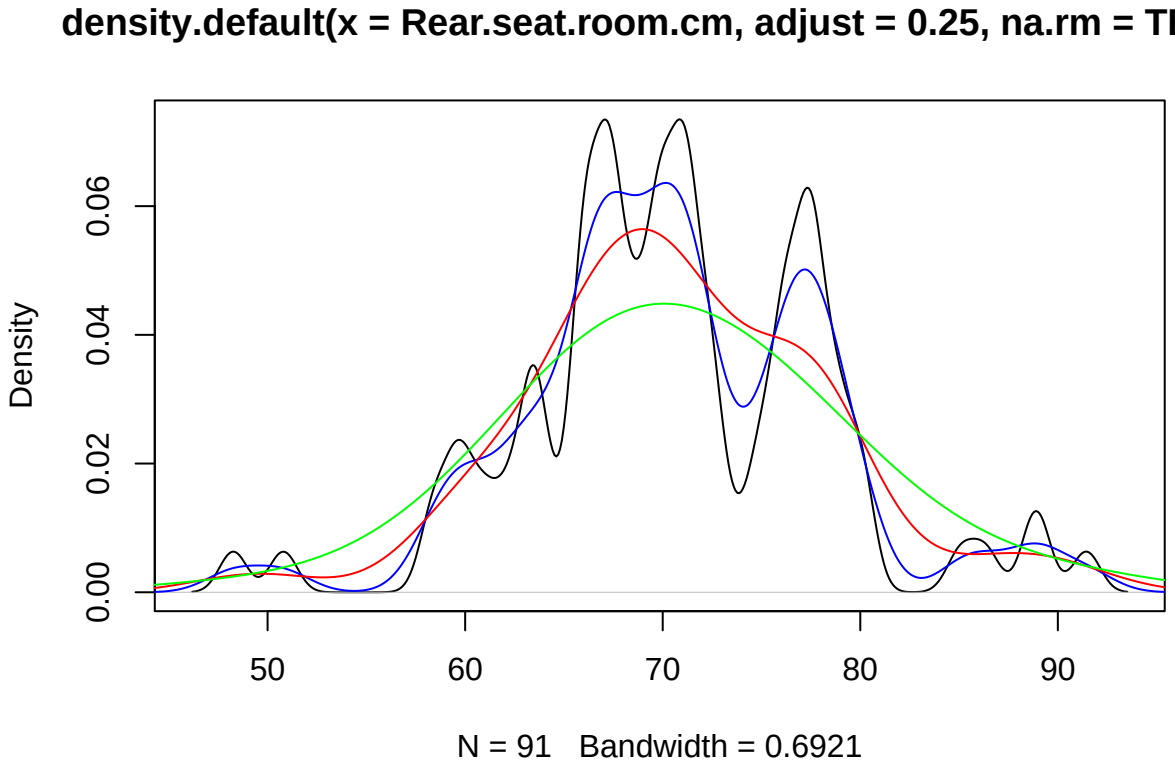
```
#### compare density plot with histogram
hist(Rear.seat.room.cm, freq=FALSE)
lines(dens)
```




```
#### change the bandwidth

dens0.5 <- density(Rear.seat.room.cm, adjust=0.5, na.rm=TRUE)


plot(dens0.5, main="adjust=0.5")
lines(dens, col="blue")
```




```
#### compare different bandwidths

dens0.25 <- density(Rear.seat.room.cm, adjust=0.25, na.rm=TRUE)
dens2 <- density(Rear.seat.room.cm, adjust=2, na.rm=TRUE)

plot(dens0.25)
lines(dens0.5, col="blue")
lines(dens, col="red")
lines(dens2, col="green")
```




```
Rear.seat.room.cm <- Rear.seat.room * 2.54


#### mean and median

mean(Rear.seat.room.cm)
```

```
## [1] NA
```

```
mean(Rear.seat.room.cm, na.rm=TRUE)
```

```
## [1] 70.68736
```

```
median(Rear.seat.room.cm, na.rm=TRUE)
```

```
## [1] 69.85
```

```
sort(Rear.seat.room.cm)[46]
```

```
## [1] 69.85
```

```
#### removing NA's

# using the function is.na()
is.na(Rear.seat.room.cm)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
x <- Rear.seat.room.cm[!is.na(Rear.seat.room.cm)]

# using the function na.omit()
x <- na.omit(Rear.seat.room.cm)

#### variance and standard deviation

s2 <- sum((x-mean(x))^2)/(length(x)-1)
var(x)
```

```
## [1] 57.64217
```

```
sd(x)
```

```
## [1] 7.592244
```

```
sqrt(s2)
```

```
## [1] 7.592244
```

```
#### quantiles

quantile(Rear.seat.room.cm, 0.3, na.rm=TRUE)
```

```
##   30%
## 67.31
```

```
quantile(Rear.seat.room.cm, c(0.3, 0.6), na.rm=TRUE)
```

```
##   30%   60%
## 67.31 71.12
```

```
quantile(Rear.seat.room.cm, na.rm=TRUE)
```

```
##    0%   25%   50%   75%  100%
## 48.26 66.04 69.85 76.20 91.44
```

```
IQR(Rear.seat.room.cm, na.rm=TRUE)
```

```
## [1] 10.16
```