EXPLORATARY DATA ANALYSIS

23CSEG28

```r
#Required ibraries
library(ggplot2)
library(lattice)
library(reshape2)
library(dplyr)

##
## 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

#Loading the dataset
data(Orange)
#Structure of the data
str(Orange)

## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.f
rame':   35 obs. of  3 variables:
##  $ Tree         : Ord.factor w/ 5 levels "3"<"1"<"5"<"2"<..: 2 2 2 2
 2 2 2 4 4 4 ...
##  $ age          : num  118 484 664 1004 1231 ...
##  $ circumference: num  30 58 87 115 120 142 145 33 69 111 ...
##  - attr(*, "formula")=Class 'formula'  language circumference ~ age
| Tree
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "labels")=List of 2
##   ..$ x: chr "Time since December 31, 1968"
##   ..$ y: chr "Trunk circumference"
##  - attr(*, "units")=List of 2
##   ..$ x: chr "(days)"
##   ..$ y: chr "(mm)"

#Summary of the data
summary(Orange)

##  Tree       age         circumference
##  3:7   Min.   : 118.0   Min.   : 30.0
##  1:7   1st Qu.: 484.0   1st Qu.: 65.5
##  5:7   Median :1004.0   Median :115.0
```

```
##  2:7   Mean   : 922.1   Mean   :115.9
##  4:7   3rd Qu.:1372.0   3rd Qu.:161.5
##        Max.   :1582.0   Max.   :214.0

#Dimension of the data
dim(Orange)

## [1] 35  3

#Checking missing values
sum(is.na(Orange))

## [1] 0

#Univariet analysis
#Distribution of Age of Trees
histogram(~age,data=Orange,main="Distribution of Age of Trees",xlab="Ag
e of tree",col="black",fill="age")
```

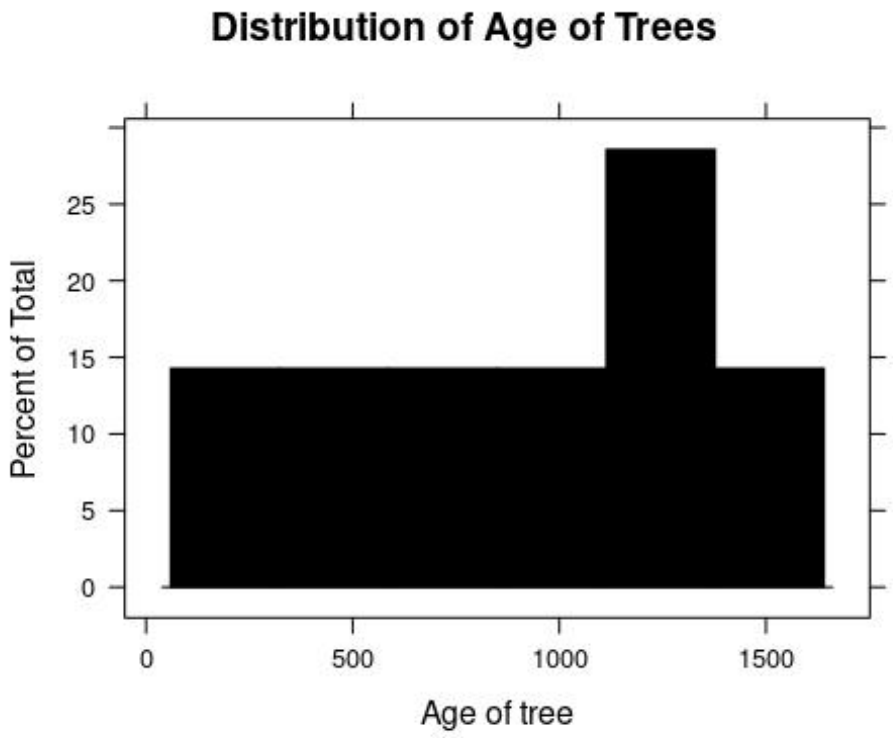

Fig 1.1

```
#Distribution of circumference of the trees
histogram(~circumference,data=Orange,main="Distribution of circumferenc
e of the trees",xlab="circumference of trees",col="black",fill="circumf
erence")
```

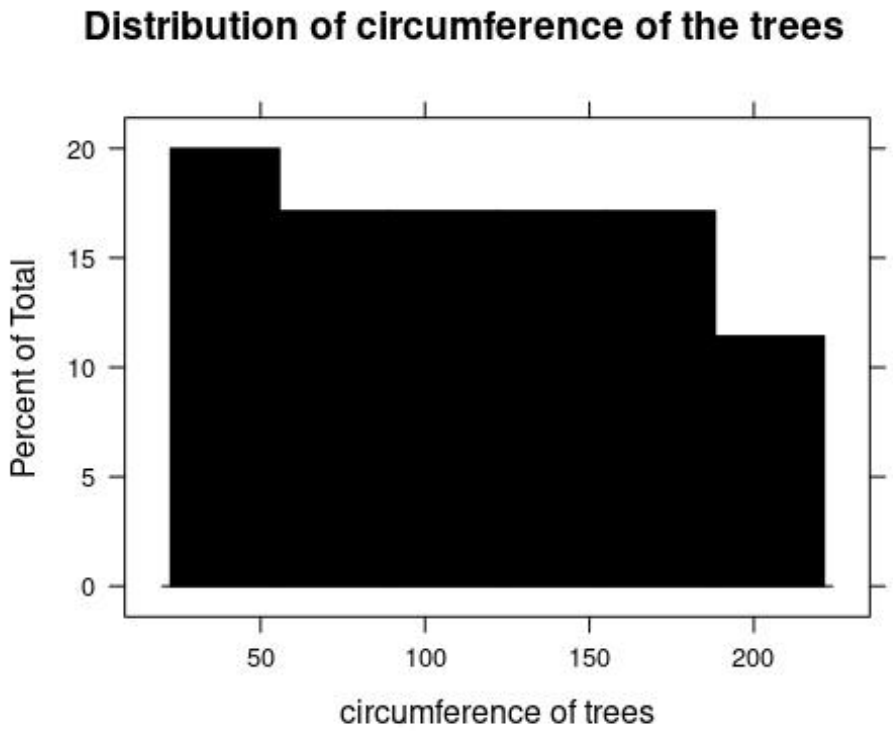

Fig 1.2

```
Orange%>%
  group_by(Tree)%>%
  summarise(mean(age))
```

```
## # A tibble: 5 × 2
##   Tree  `mean(age)`
##   <ord>       <dbl>
## 1 3            922.
## 2 1            922.
## 3 5            922.
## 4 2            922.
## 5 4            922.
```

```
#Distribution of trees and their age
histogram(~age|factor(Tree),data=Orange,main="Distribution of trees and
 their age")
```

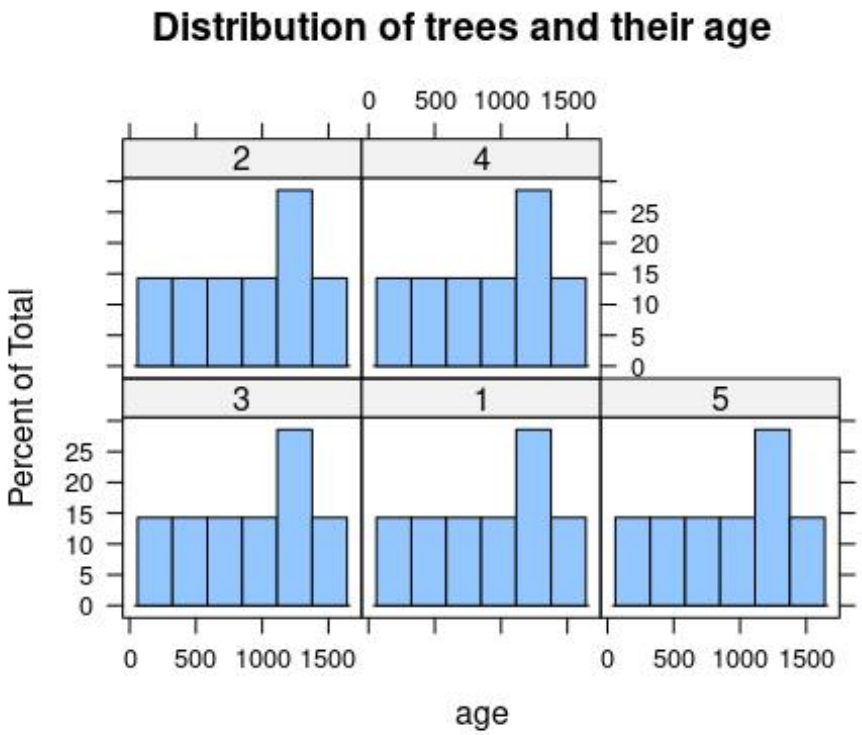

Fig 1.3

```r
#Circumference by Tree
boxplot(circumference ~ Tree, data = Orange, main = "Circumference by T
ree")
```

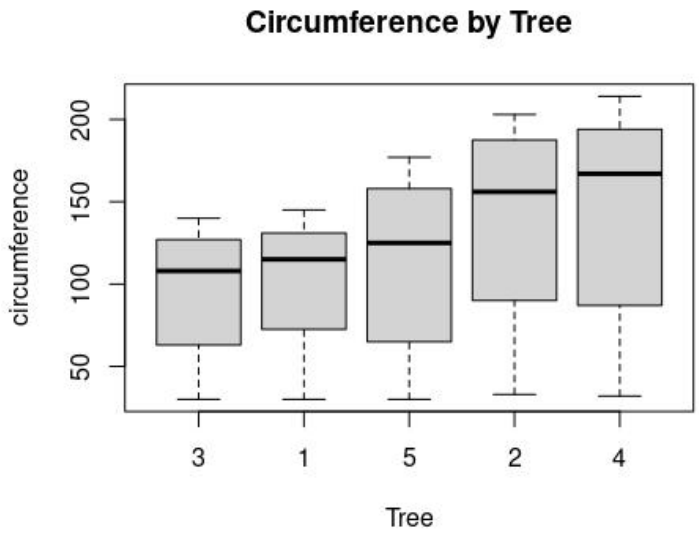

Fig 1.4

```r
Orange%>%
  group_by(Tree)%>%
  summarise(val=mean(circumference))%>%
  arrange(desc(val))
```

```
## # A tibble: 5 × 2
##   Tree    val
##   <ord> <dbl>
## 1 4     139.
## 2 2     135.
## 3 5     111.
## 4 1      99.6
## 5 3      94
```

```r
#Bivariet analysis
#Age vs Circumference
xyplot(circumference~age,data=Orange,main="Age vs Circumference",xlab =
"Age",ylab="Circumference",col="black")
```

Fig 1.5

```
#Tree Vs Circumference
ggplot(Orange, aes(x = as.factor(Tree), y = circumference)) +geom_point
(alpha = 0.5) +facet_wrap(~ age) +
  labs(x = "Tree", y = "Circumference", title = "Tree Vs Circumference")
```

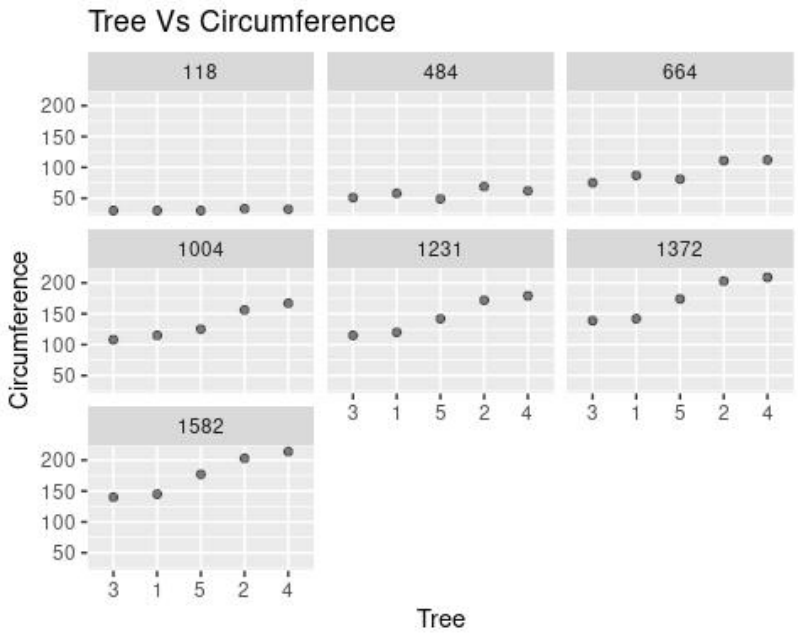

Fig 1.6

```r
#Multivariet analysis
#Relationship between all the numerical attribute
data=cor(Orange[sapply(Orange, is.numeric)])
data1= melt(data)
#Relationship between all the numerical attribute
ggplot(data1, aes(x = Var1, y = Var2, fill = value)) +geom_tile() +labs
(title = "Relationship between all the numerical attribute",x="Numerica
l attributes",y="numerical attributes")
```

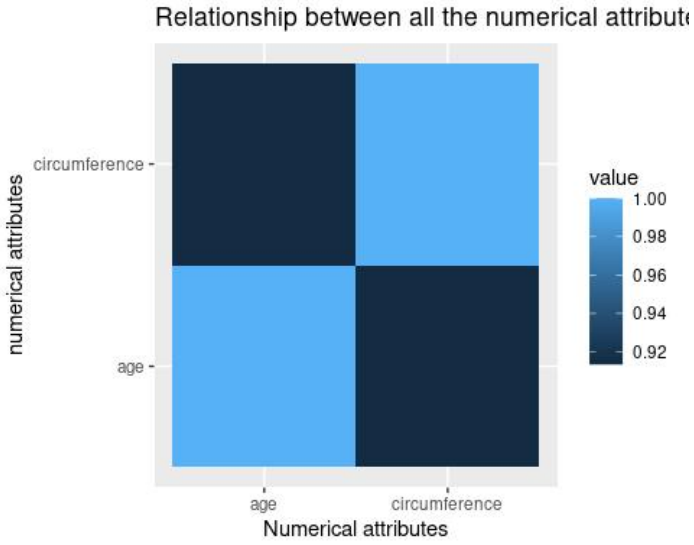

Fig 1.7

```r
#Relationship of Circumference by Age and Tree
levelplot(circumference ~ age * Tree, data = Orange, col.regions = colo
rRampPalette(c("white", "black")), xlab = "Age", ylab = "Tree", main =
"Relationship of Circumference by Age and Tree")
```

Fig 1.8