Atv11

1 - Estatística Descritiva Univariada:

Baixamos as bibliotecas que serão necessarias durante a atividade:

```
library(readr)
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

Atribuimos o arquivo desejado a variavel "barley" e usamos o comando de visualização.

```
barley <- read.table("~/Ufmg/2022-1/Pacotes/Atividades/08-07-22/barley.txt", header=TRUE, sep = ",")
View(barley)
```

Fazemos a seleção de linhas especificas da tabela para analise que possa ser feita a analise das variaveis desejadas

```
barley[c(1,7,45,87),]
```

```
##    row.names    yield   variety year            site
## 1          1 27.00000 Manchuria 1931 University Farm
## 7          7 43.06666   Glabron 1931 University Farm
## 45        45 29.86667  Peatland 1931          Morris
## 87        87 46.63333     Trebi 1932          Morris
```

Atraves da função summary podemos ter alguns dados para analise rapidamente.

```
summary(barley)
```

```
##    row.names          yield         variety               year
##  Min.   :  1.00   Min.   :14.43   Length:120         Min.   :1931
##  1st Qu.: 30.75   1st Qu.:26.88   Class :character   1st Qu.:1931
##  Median : 60.50   Median :32.87   Mode  :character   Median :1932
##  Mean   : 60.50   Mean   :34.42                      Mean   :1932
##  3rd Qu.: 90.25   3rd Qu.:41.40                      3rd Qu.:1932
##  Max.   :120.00   Max.   :65.77                      Max.   :1932
##      site
##  Length:120
##  Class :character
##  Mode  :character
##
##
##
```

Também podemos especificar os itens que queremos analisar.

```
summary(barley[c(1,7,45,87),])
```

```
##    row.names        yield         variety               year
##  Min.   : 1.0   Min.   :27.00   Length:4           Min.   :1931
##  1st Qu.: 5.5   1st Qu.:29.15   Class :character   1st Qu.:1931
##  Median :26.0   Median :36.47   Mode  :character   Median :1931
##  Mean   :35.0   Mean   :36.64                      Mean   :1931
##  3rd Qu.:55.5   3rd Qu.:43.96                      3rd Qu.:1931
##  Max.   :87.0   Max.   :46.63                      Max.   :1932
##      site
##  Length:4
##  Class :character
##  Mode  :character
##
##
##
```

Transformamos os itens da lista em "objetos" para facilitar a manipulação dos dados.

```
attach(barley)
```

Geramos um grafico para uma variavel numerica e outra categorica, podemos perceber que funciona somente para variaveis numericas.

```
#Numerica:
stem(yield)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
```

```
##
##   1 | 4
##   1 | 579
##   2 | 0011122233333
##   2 | 55566666666777777889999999
##   3 | 00000011122222233333344444
##   3 | 5555667777888999
##   4 | 000112223334444
##   4 | 567777779999
##   5 | 00
##   5 | 5889
##   6 | 4
##   6 | 6
```

```
#Categorica(Programa retorna erro):
#stem(variety)
```

Podemos verificar a assimetria dos dados atraves da função abaixo.

```
quantile(yield, seq(0.1, 0.9, by=0.1))
```

```
##      10%      20%      30%      40%      50%      60%      70%      80%
## 22.49667 26.08000 28.09000 29.94667 32.86667 35.13333 38.97333 43.32000
##      90%
## 47.45666
```

Podemos selecionar quais dados queremos analisar baseado em outra variavel definida:

```
summary(yield[year==1931])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.70   29.09   34.20   37.08   43.85   65.77
```

Para descobrir o ano mais produtivo podemos utiliazar a função abaixo, que nos retorna a probabilidade de se encontrar

```
quantile(yield[year==1931], 0.9)
```

```
##      90%
## 49.90334
```

```
quantile(yield[year==1932], 0.9)
```

```
##   90%
## 44.28
```

Podemos usar a função "by" para facilitar a analise dos dados

```
by(barley, year, summary)
```

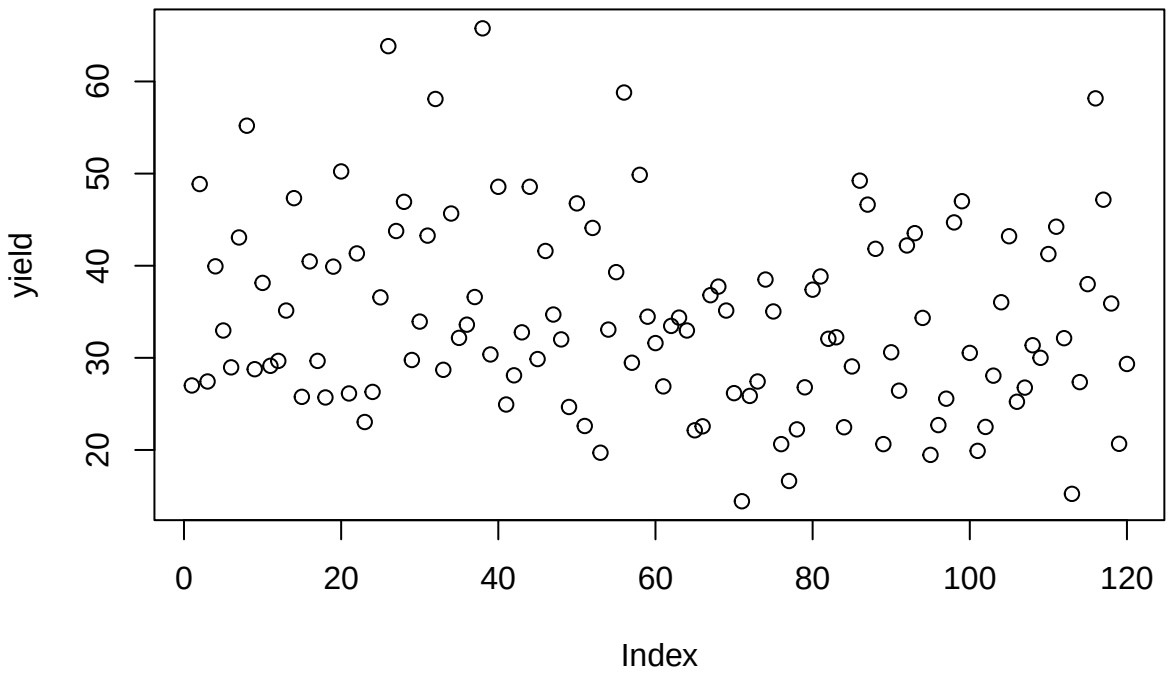
```
## year: 1931
##    row.names         yield         variety               year
##  Min.   : 1.00   Min.   :19.70   Length:60          Min.   :1931
##  1st Qu.:15.75   1st Qu.:29.09   Class :character   1st Qu.:1931
##  Median :30.50   Median :34.20   Mode  :character   Median :1931
##  Mean   :30.50   Mean   :37.08                      Mean   :1931
##  3rd Qu.:45.25   3rd Qu.:43.85                      3rd Qu.:1931
##  Max.   :60.00   Max.   :65.77                      Max.   :1931
##      site
##  Length:60
##  Class :character
##  Mode  :character
##
##
##
## ------------------------------------------------------------
## year: 1932
##    row.names          yield         variety               year
##  Min.   : 61.00   Min.   :14.43   Length:60          Min.   :1932
##  1st Qu.: 75.75   1st Qu.:25.48   Class :character   1st Qu.:1932
##  Median : 90.50   Median :30.98   Mode  :character   Median :1932
##  Mean   : 90.50   Mean   :31.76                      Mean   :1932
##  3rd Qu.:105.25   3rd Qu.:37.80                      3rd Qu.:1932
##  Max.   :120.00   Max.   :58.17                      Max.   :1932
##      site
##  Length:60
##  Class :character
##  Mode  :character
##
##
##
```

2 - Análise Exploratória Gráfica Univariada:

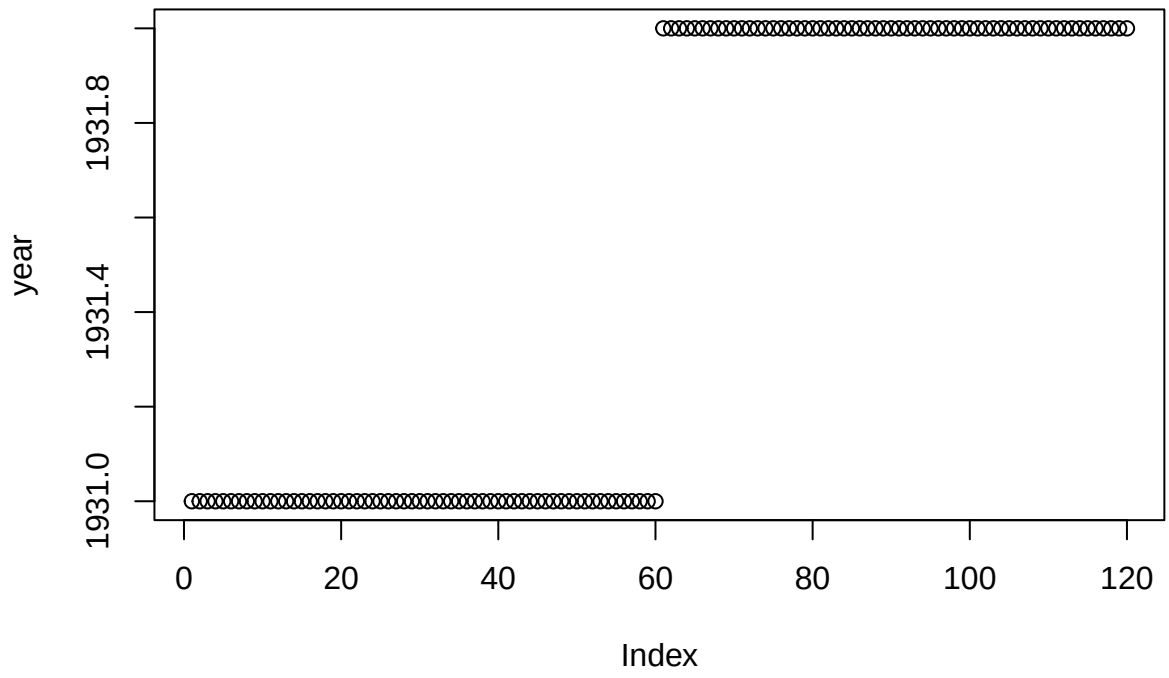
Para conseguirmos vizualizar todos os graficos de uma vez:

```
par(mfrow=c(1,1)) # divide tela gráfica em matriz 2 x 1

plot(yield)
```




```
plot(year)
```



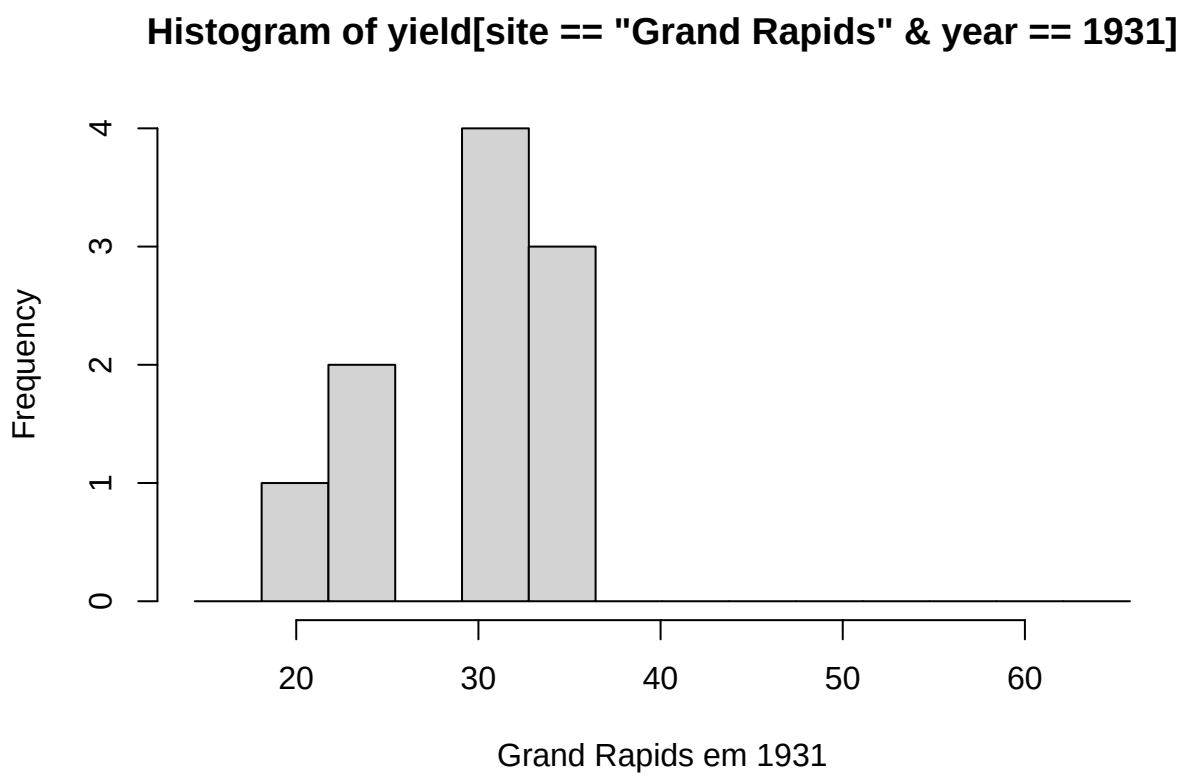
Os graficos abaixo retornam a frequência de ocorrencias de objetos selecionados:

Definindo as variáveis que serão usadas:

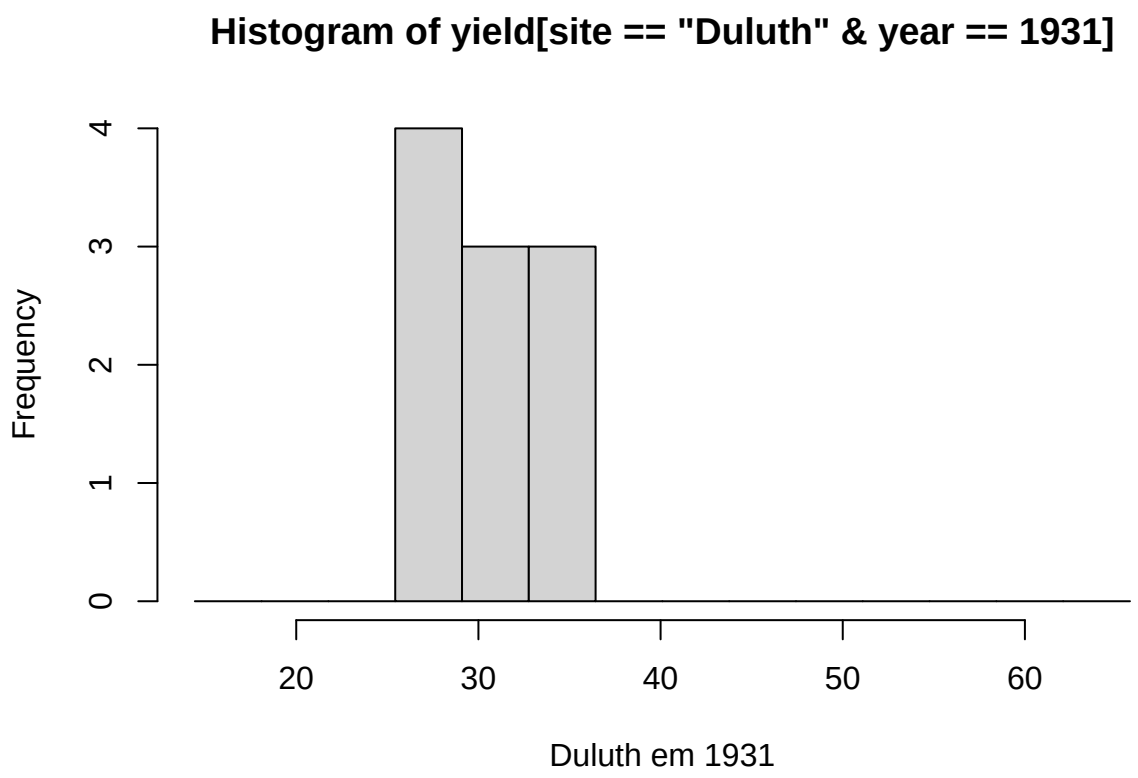
```
yrange <- range(yield)
limits <- seq(min(yield), max(yield), length=15)
```

Ano e localização.

```
hist(yield[site=="Grand Rapids"& year==1931],
     xlim=yrange, breaks=limits, xlab="Grand Rapids em 1931")
```

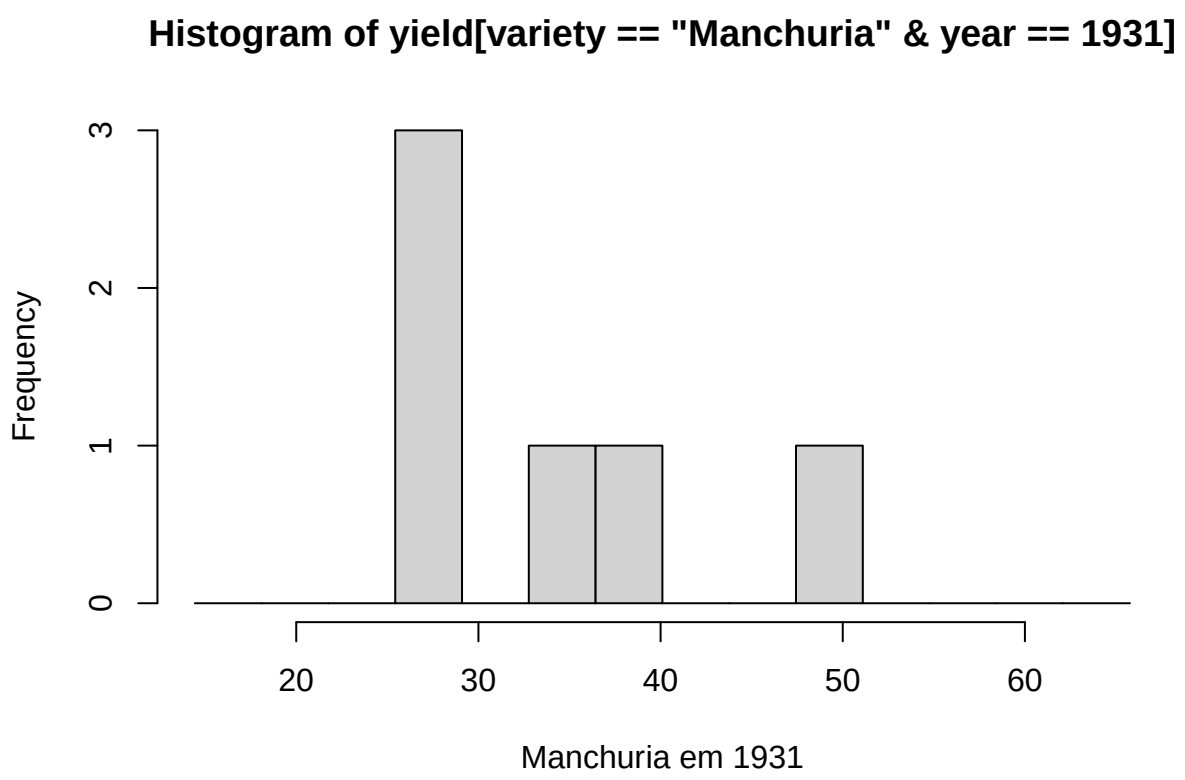



```
hist(yield[site=="Duluth"& year==1931],
     xlim=yrange, breaks=limits, xlab="Duluth em 1931")
```

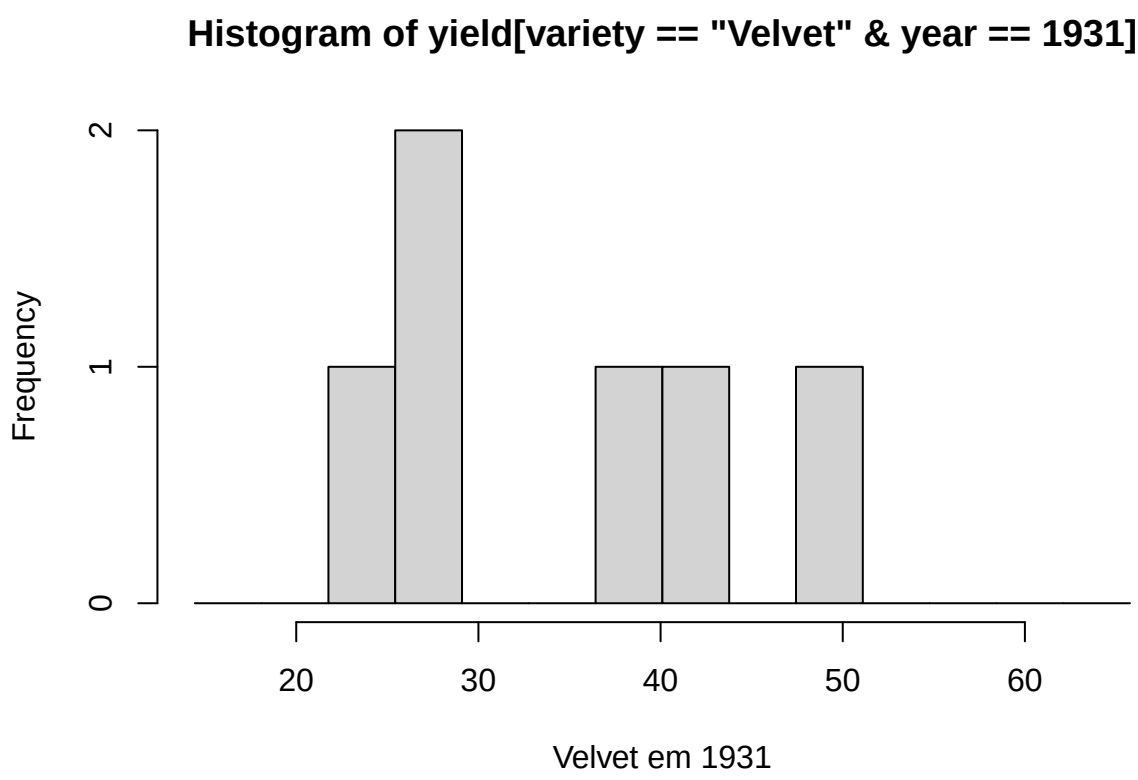


Ano e Variedade:

```
hist(yield[variety=="Manchuria"& year==1931],
     xlim=yrange, breaks=limits, xlab="Manchuria em 1931")
```

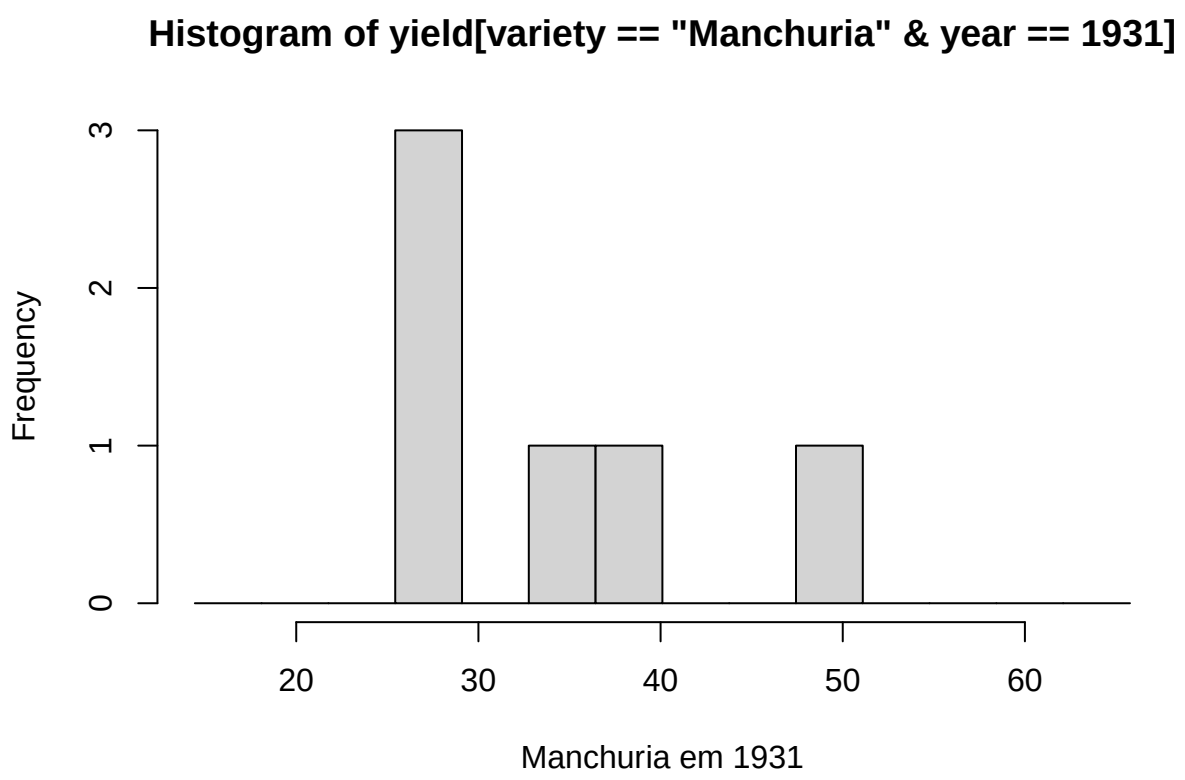



```
hist(yield[variety=="Velvet"& year==1931],
     xlim=yrange, breaks=limits, xlab="Velvet em 1931")
```

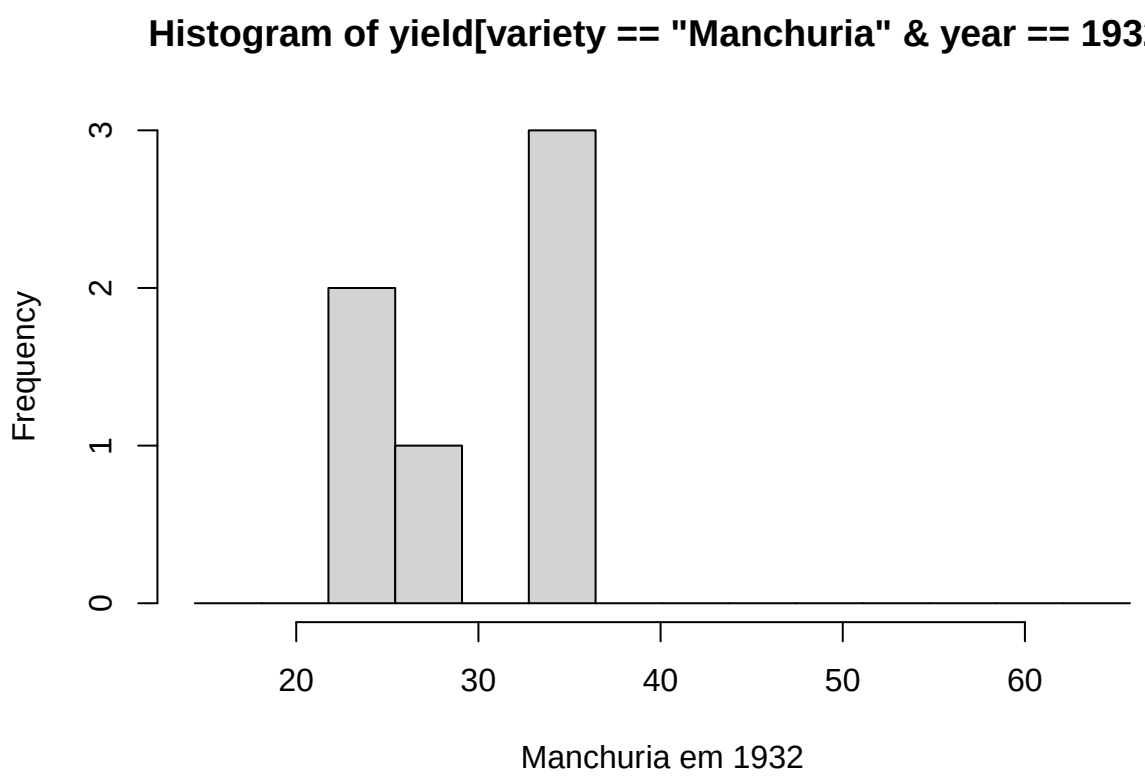


Comparando mesma variedade em anos diferentes:

```
hist(yield[variety=="Manchuria"& year==1931],
     xlim=yrange, breaks=limits, xlab="Manchuria em 1931")
```




```
hist(yield[variety=="Manchuria"& year==1932],
     xlim=yrange, breaks=limits, xlab="Manchuria em 1932")
```



3 - Estatística Descritiva Multivariada:

Podemos verificar a frequência de cada "site" em cada ano usando o seguinte comando:

```
table(year, site)
```

```
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931        10     10           10     10              10     10
##   1932        10     10           10     10              10     10
```

Podemos verificar quantas vezes cada "variety" aconteceu em cada "site" por "year" como comando:

```
table(year, site, variety)
```

```
## , , variety = Glabron
##
##       site
```

```
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Manchuria
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = No. 457
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = No. 462
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = No. 475
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Peatland
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Svansota
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Trebi
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Velvet
```

```
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
##
## , , variety = Wisconsin No. 38
##
##       site
## year   Crookston Duluth Grand Rapids Morris University Farm Waseca
##   1931         1      1            1      1               1      1
##   1932         1      1            1      1               1      1
```

A função abaixo retorna a "tabela" feita para dados quantitativos de forma que não é gerado uma variavel para cada valor diferente:

```
h2d <- hist2d(yield, yield, show=FALSE, nbins=c(5,5))
h2d
```

```
##
## ----------------------------
## 2-D Histogram Object
## ----------------------------
##
## Call: hist2d(x = yield, y = yield, nbins = c(5, 5), show = FALSE)
##
## Number of data points:  120
## Number of grid bins:  5 x 5
## X range: ( 14.43333 , 65.7667 )
## Y range: ( 14.43333 , 65.7667 )
```

Podemos arredondar os valores para o inteiro mais próximo:

```
yield.round <- round(yield, 0)
yield.round
```

```
##   [1] 27 49 27 40 33 29 43 55 29 38 29 30 35 47 26 40 30 26 40 50 26 41 23 26 37
##  [26] 64 44 47 30 34 43 58 29 46 32 34 37 66 30 49 25 28 33 49 30 42 35 32 25 47
##  [51] 23 44 20 33 39 59 29 50 34 32 27 33 34 33 22 23 37 38 35 26 14 26 27 38 35
##  [76] 21 17 22 27 37 39 32 32 22 29 49 47 42 21 31 26 42 44 34 19 23 26 45 47 31
## [101] 20 22 28 36 43 25 27 31 30 41 44 32 15 27 38 58 47 36 21 29
```

Também poodemos arredondar o valor para a DEZENA mais próxima

```
yield.round <- round(yield, -1)
yield.round
```

```
##   [1] 30 50 30 40 30 30 40 60 30 40 30 30 40 50 30 40 30 30 40 50 30 40 20 30 40
##  [26] 60 40 50 30 30 40 60 30 50 30 30 40 70 30 50 20 30 30 50 30 40 30 30 20 50
##  [51] 20 40 20 30 40 60 30 50 30 30 30 30 30 30 20 20 40 40 40 30 10 30 30 40 40
##  [76] 20 20 20 30 40 40 30 30 20 30 50 50 40 20 30 30 40 40 30 20 20 30 40 50 30
## [101] 20 20 30 40 40 30 30 30 30 40 40 30 20 30 40 60 50 40 20 30
```

Podemos gerar uma tabela com a quantidade de valores que pertencem a cada dezena:

```
table(yield.round)
```

```
## yield.round
## 10 20 30 40 50 60 70
##  1 18 51 31 13  5  1
```

Tabela com os valores das dezenas de "yeld" sub divididos baseados nas variedades:

```
table(variety, yield.round)
```

```
##                   yield.round
## variety            10 20 30 40 50 60 70
##   Glabron           1  0  5  5  0  1  0
##   Manchuria         0  2  8  1  1  0  0
##   No. 457           0  2  5  3  1  1  0
##   No. 462           0  3  4  2  2  0  1
##   No. 475           0  4  4  3  1  0  0
##   Peatland          0  0  8  3  1  0  0
##   Svansota          0  3  4  4  1  0  0
##   Trebi             0  1  4  3  3  1  0
##   Velvet            0  2  5  4  1  0  0
##   Wisconsin No. 38  0  1  4  3  2  2  0
```

Função que retorna dados mais detalhados de "yield" baseado na "variety" para analise mais profunda:

```
by(yield, variety, summary)
```

```
## variety: Glabron
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.43   28.12   32.40   33.34   37.83   55.20
## ------------------------------------------------------------
## variety: Manchuria
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.13   26.98   30.97   31.46   33.69   48.87
## ------------------------------------------------------------
```

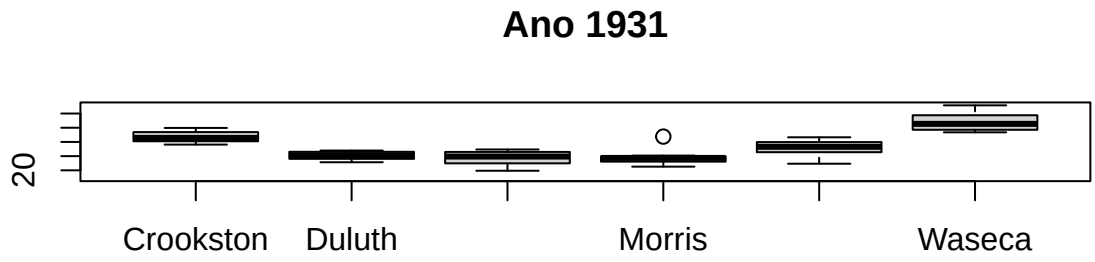
```
## variety: No. 457
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.47   28.13   33.97   35.85   43.33   58.10
## ------------------------------------------------------------
## variety: No. 462
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.90   25.41   30.45   35.38   45.27   65.77
## ------------------------------------------------------------
## variety: No. 475
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.23   24.15   31.07   31.76   41.98   46.77
## ------------------------------------------------------------
## variety: Peatland
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.23   29.42   32.38   34.18   37.42   48.57
## ------------------------------------------------------------
## variety: Svansota
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.63   24.83   28.55   30.38   35.98   47.33
## ------------------------------------------------------------
## variety: Trebi
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.63   30.39   39.20   39.40   46.71   63.83
## ------------------------------------------------------------
## variety: Velvet
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.47   26.26   32.15   33.06   39.10   50.23
## ------------------------------------------------------------
## variety: Wisconsin No. 38
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.67   31.07   36.95   39.39   47.84   58.80
```

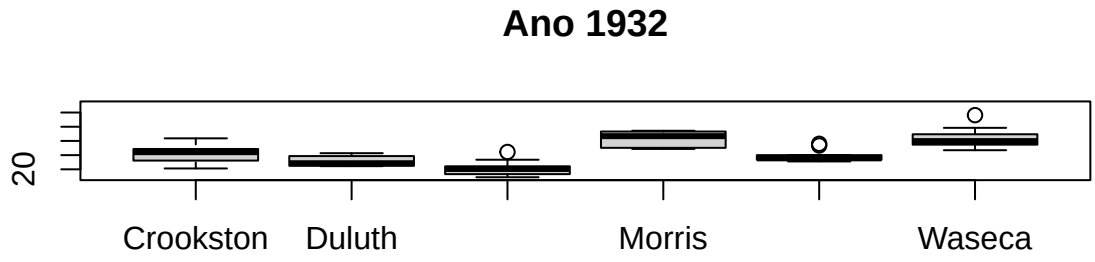
Fica evidente a superioridade produtiva das variedades Wisc No. 38 e Trebi.

4 - Análise Exploratória Gráfica Multivariada:

```
par(mfrow=c(2,1)) # divide tela gráfica em matriz 2 x 1
is.1931 <- year==1931 # True ou False, dependendo do ano 1931/1932
data.split.31 <- split(yield[is.1931], site[is.1931])
boxplot(data.split.31, main="Ano 1931", ylim=range(yield))

data.split.32 <- split(yield[!is.1931], site[!is.1931])
boxplot(data.split.32, main="Ano 1932", ylim=range(yield))
```






```
par(mfrow=c(3,2)) # divide tela gráfica em matriz 3 x 2
is.Crookston <- site == "Crookston"
data.split.1 <- split(yield[is.Crookston], year[is.Crookston])
boxplot(data.split.1, main = "Crookston", ylim=range(yield))

is.Morris <-site == "Morris"
data.split.2 <- split(yield[is.Morris], year[is.Morris])
boxplot(data.split.2, main = "Morris", ylim=range(yield))

is.Waseca <- site == "Waseca"
data.split.3 <- split(yield[is.Waseca ], year[is.Waseca ])
boxplot(data.split.3, main = "Waseca", ylim=range(yield))

is.University.Farm <- site == "University Farm"
data.split.4 <- split(yield[is.University.Farm], year[is.University.Farm])
boxplot(data.split.4, main = "University Farm", ylim=range(yield))

is.Grand.Rapids <- site == "Grand Rapids"
data.split.5 <- split(yield[is.Grand.Rapids], year[is.Grand.Rapids])
boxplot(data.split.5, main = "Grand Rapids", ylim=range(yield))

is.Duluth <- site == "Duluth"
data.split.6 <- split(yield[is.Duluth], year[is.Duluth])
boxplot(data.split.6, main = "Duluth", ylim=range(yield))
```

Pela analise dos graficos podemos concluir que “Morris” foi a unica cidade a produzir mais em 1932 do que em 1931