Análise de correlação

As medidas de posição e de dispersão se constituem como ferramentas de análise exploratória de um único conjunto de valores.

A análise de correlação , por outro lado, permite estudar a relação entre dois conjuntos de valores.
Nesse tipo de análise, quantifica-se o quanto um conjunto de valores (ou uma variável) está relacionado com outro, no sentido de determinar a intensidade e a direção dessa relação.

Em outras palavras, a correlação indica se, e com que intensidade, os valores de uma variável aumentam (ou diminuem) enquanto os valores da outra variável aumentam (ou diminuem).

A quantificação da correlação é feita por meio de um coeficiente, sendo que o mais conhecido é o coeficiente de correlação de Pearson (r).

O cálculo de r para dois conjuntos de valores (ou variáveis) v_1 e v_2, com n valores cada, é:

$$r_{v_1,v_2} = \frac{\sum(v_{1i} - media(v_1))(v_{2i} - media(v_2))}{\sqrt{\sum(v_{1i} - media(v_1))} \times \sqrt{\sum(v_{2i} - media(v_2))}}$$

em que v_{1i} e v_{2i} são elementos dos conjuntos de valores v_1 e v_2.

O resultado do coeficiente de correlação assumirá valores entre −1 e +1.

O sinal do resultado indica a direção, se a **correlação é positiva ou negativa**, e o valor indica a intensidade da correlação. Valores de correlação acima de $|0,70|$ indicam **forte correlação**, sendo que $|\cdot|$ indica a operação de módulo.

Como exemplo, imagine que se deseja comparar a quantidade de vendas do prato feijoada (Quantidade I) e da bebida caipirinha (Quantidade II), no decorrer dos meses de novembro a julho, que ocorrem conforme apresentado na Tabela abaixo:

Mês	**nov**	**dez**	**jan**	**fev**	**mar**	**abr**	**mai**	**jun**	**jul**
Quantidade I (feijoada)	24	27	29	30	32	58	64	64	65
Quantidade II (caipirinha)	10	25	20	36	28	38	50	60	69

Para facilitar os cálculos, considere que Quantidade I será representada pela variável v_1, e a Quantidade II, pela variável v_2.

Todos os cálculos necessários para encontrar o coeficiente de correlação de Pearson estão resumidos na Tabela abiaxo.

A correlação r_{v+1,v_2} é de **0,89** (uma **correlação alta e positiva**), indicando que o comportamento das vendas dos dois itens é similar em relação à tendência de aumento (ou queda) de vendas.

Quando o prato feijoada tem mais procura no restaurante, a bebida caipirinha também tem o comportamento de venda alterado positivamente. Se cai a venda da feijoada, cai a venda da caipirinha.

	v1	v2	v1−média(v1)	v2−média(v2)	A*B	(v1−média(v1))^2	(v2−média(v2))^2
	24	10	-19,67	-27,33	537,56	386,78	747,11
	27	25	-16,67	-12,33	205,56	277,78	152,11
	29	20	-14,67	-17,33	254,22	215,11	300,44
	30	36	-13,67	-1,33	18,22	186,78	1,78
	32	28	-11,67	-9,33	108,89	136,11	87,11
	58	38	14,33	0,67	9,56	205,44	0,44
	64	50	20,33	12,67	257,56	413,44	160,44
	64	60	20,33	22,67	460,89	413,44	513,78
	65	69	21,33	31,67	675,56	455,11	1002,78
soma	**393**	**336**				2690	2966
médias	**43,67**	**37,33**			**aux1 = 2528**	**aux2 = 51,87**	**aux3 = 54,46**

Complementado os calulos temos:

$$r_{v_1,v_2} = \frac{aux1}{aux2 \times aux3} = \frac{2558}{51,87 \times 54,46} = 0,89$$

Exemplo com R

Um restaurante está interessado em avaliar se **há uma relação entre o consumo de feijoada e de caipirnha** para isso ele fez um levantamento durante alguns meses da quantidade total de caiprinahs vendidas e de feijoadas vendidas. O que você acha?

In [62]:
```r
# Vetor de valores v1 (Feijoada)
v1 <- c(24,27,29,30,32,58,64,64,65)
```

In [63]:
```r
# Vetor de valores v1 (Feijoada)
v2 <- c(10,25,20,36,28,38,50,60,69)
```

In [64]:
```r
# Calculado a correlação
cor(v1,v2)
```

0.894984175042704

Parace que temos uma correlação forte, ou seja se aumenta o consumo de feijoada também aumenta o consumo de caipirinha.

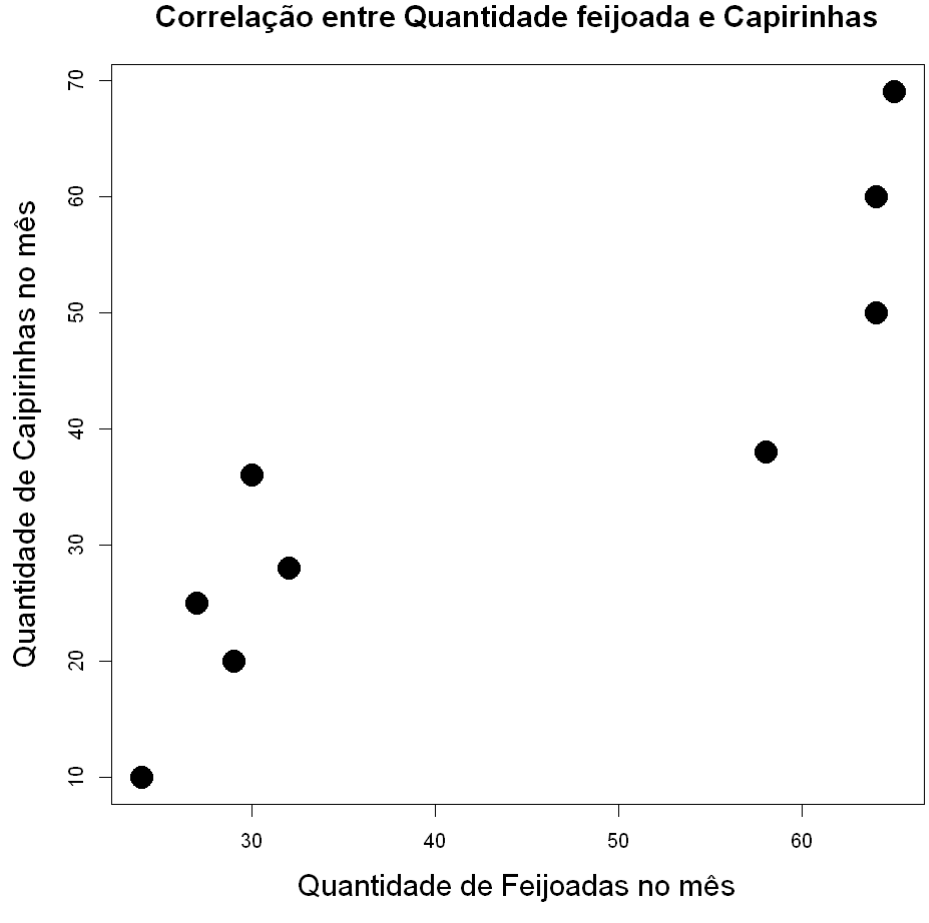

THE
BUSINESS

Visualização da Correlação no R

Agora que entendemos um pouco mais a idéia de correlação vamos visualiza-lá

In [65]:

```
plot(v1,v2, cex=2.5,
     main = 'Correlação entre Quantidade feijoada e Capirinhas',
     pch=19,
     ylab='Quantidade de Caipirinhas no mês',
     xlab='Quantidade de Feijoadas no mês',
     cex.lab=1.5,
     cex.main=1.5)
```



Aconselhado por um especilaista em R. O Restaurante decidiu antes de tomar outra decisão entender um pouco melhor a correlação. As recomendações foram:

- Olhar para um período mais longo, ou seja coletar mais dados
- Comparar com outros produtos
- Analisar os dados e buscar *insigths*


iNSiGHT

Transformando os dados coletados em vetores

In [66]:
```
# Quantidade de feijoadas vendidas entre nov/18 e out/20
v1<-c(30,47,25,37,43,65,38,60,35,39,36,60,24,27,29,30,32,58,64,64,65,26,64,51)
# Quantidade de caipirinhas vendidas entre nov/18 e out/20
v2<-c(63,48,16,47,16,41,29,30,32,12,13,30,10,25,20,36,28,38,50,60,69,23,48,49)
# Quantidade de refrigerantes vendidas entre nov/18 e out/20
v3<-c(87,92,54,58,116,114,85,118,51,64,75,110,45,82,84,90,43,128,123,148,120,55,114,79)
# Quantidade de refrigerantes cervejas entre nov/18 e out/20
v4<-c(71,101,60,75,76,145,68,111,75,86,83,120,33,64,62,51,60,129,143,118,126,50,115,113)
```

Correlação entre Feijoada e Caipirinhas

In [67]:
```
# Correlação entre Feijoada e Caipirinahs
cor(v1,v2)
```

0.575868609446574

Parece que a correlação entre feijoada e caipirinha foo forte somente no perído estudado incialmente. Para surpresa geral essas correlção não s emostrou tão forte num prazo mais extenso

Correlação entre Feijoada e Refrigerantes

In [68]:
```
# Correlação entre Feijoada e Refrigerantes
cor(v1,v3)
```

0.828530582708968

Já a correlação entre feijoada e refirgerantes é forte nesse período. Mas não era essa aimpressão que se tinha antes.

Correlação entre Feijoada e Cervejas

In [44]:
```
# Correlação entre Feijoada e Cervejas
cor(v1,v4)
```

0.952882265531835

Correlação entre Feijoada e Cervejas é bastante forte, ou seja se aumenta o consumo de feijaoda também aumenta o consumo de cerveja

Agora sabemos que o aumento do consumo de feijoada impacta no aumento de cervejas e refrigerantes do que no de caipirinhas.

Como que o dono do restaurante pode utilizar esta informação?

Esse processo que estamos fazendo é baseado na idéia de pirâmide do conhecimento que explica como um dado bruto pode se tranformar em sabedoria

1° Pilar: Dados

O item que está na base da pirâmide tem ligação com números, imagens e outros fatores que nos permitem compreender melhor algum tema. Observe que nessa fase temos apenas dados que ainda não nos dizem muito, mas que visualmente já nos trazem alguma informação que será decodificada mais à frente'm'm

2° Pilar: Informação

Tendo os dados da base da Pirâmide do Conhecimento, podemos relacioná-los de maneira a obter as informações contidas neles. Este item somente é possível a partir do anterior, basicamente nessa ferramenta vamos interligando as camadas para que elas adicionem mais conteúdo à seguinte, se complementem e façam sentido.

3° Pilar: Conhecimento

Nessa etapa é possível tirar conclusões a respeito das informações que foram obtidas na camada anterior por meio dos dados. O conhecimento que se tem a respeito de um tema é o que nos torna aptos a trabalhá-lo com mais assertividade e a conectar com os demais saberes que acumulamos.

4º Pilar: Sabedoria

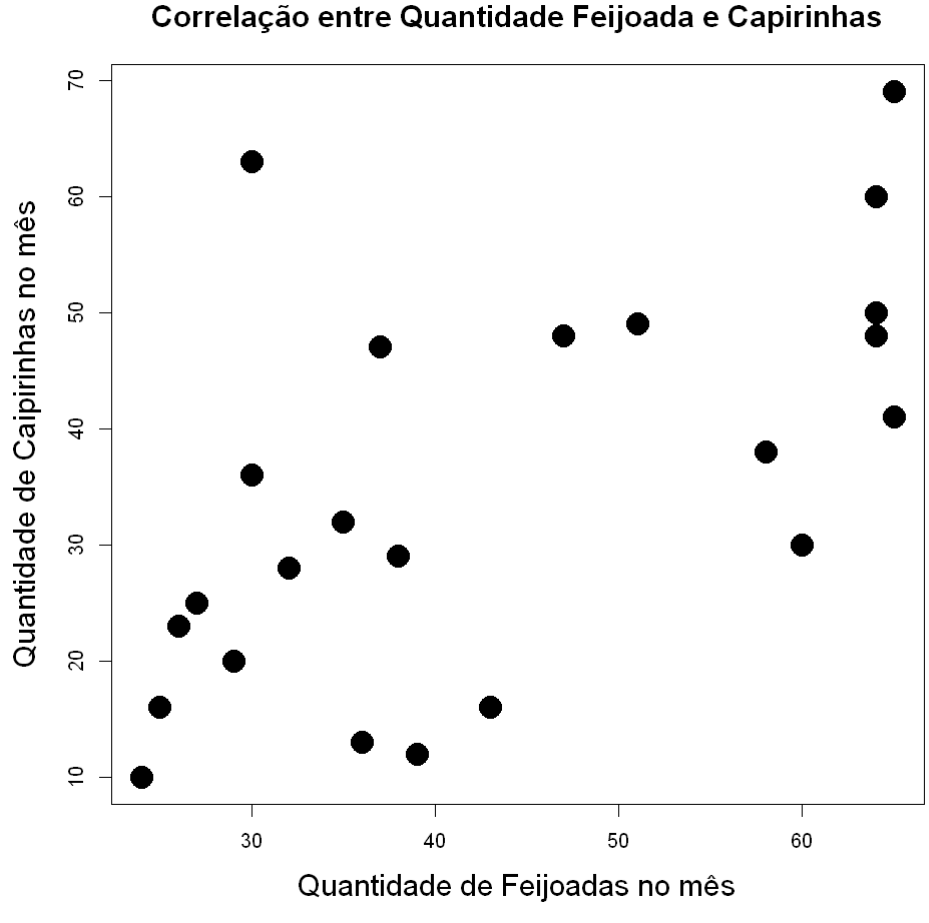
Aqueles que têm o conhecimento acerca de algum tema adquirem, então, a sabedoria de entender quando usar essa riqueza em seu favor. Ser sábio é saber qual o melhor momento de usar as suas informações, algo que se torna possível a partir dos dados que foram coletados e de sua interpretação ao longo da pirâmide do conhecimento e, claro, da sua maturidade para compreender o timing certo.

Agora sabemos que o aumento do consumod e feijoada impacat no aumento de cervejas e refrigerantes do que no de caipirinhas. Como que o dono do restaurante pode utilizar esta informação?

Vamos visualizar um pouco no R

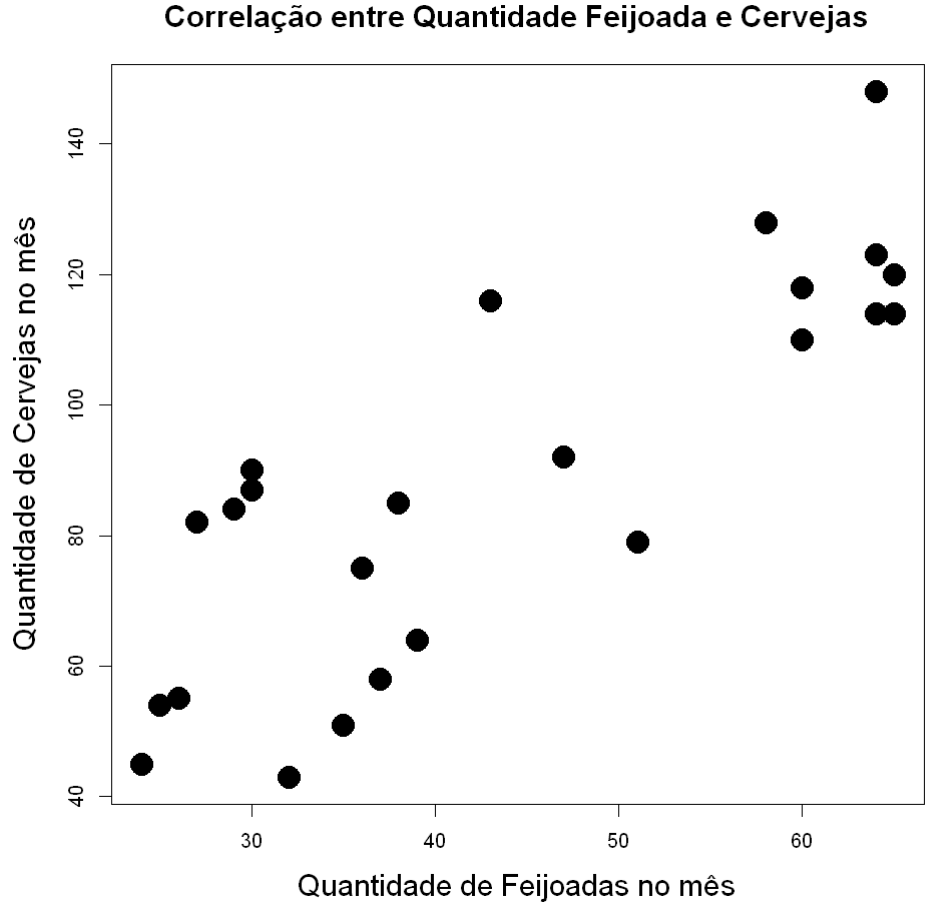
In [69]:

```
plot(v1,v2, cex=2.5,
     main = 'Correlação entre Quantidade Feijoada e Capirinhas',
     pch=19,
     ylab='Quantidade de Caipirinhas no mês',
     xlab='Quantidade de Feijoadas no mês',
     cex.lab=1.5,
     cex.main=1.5)
```



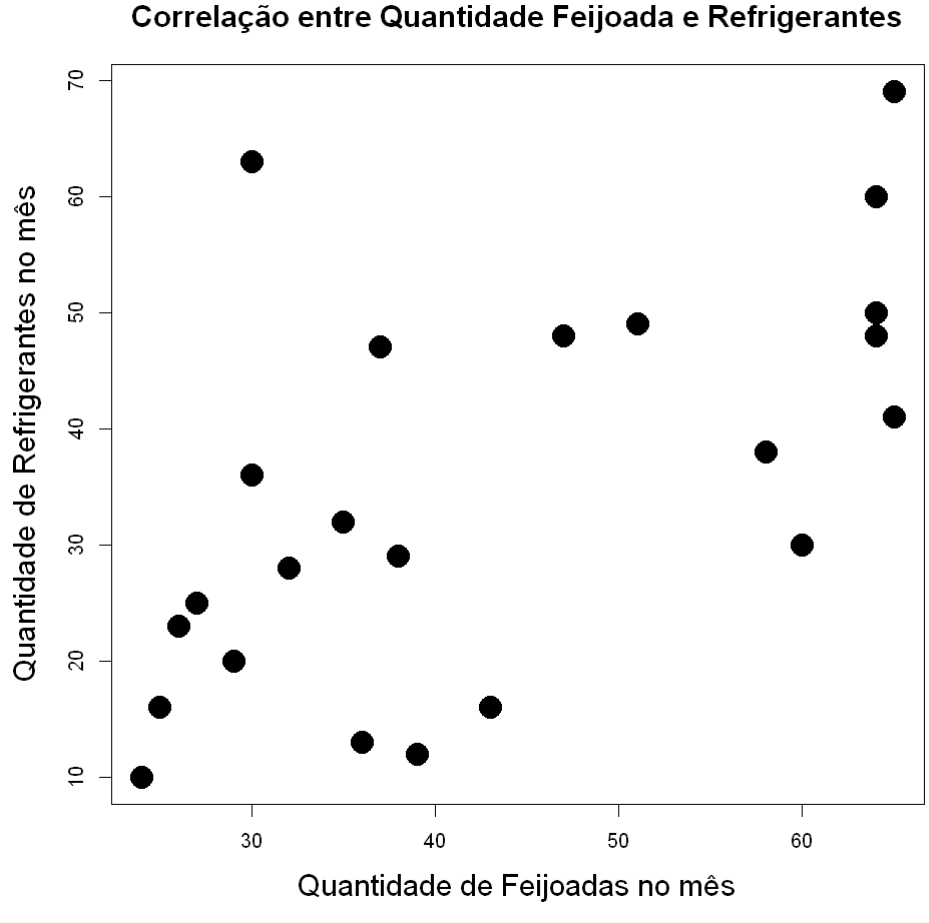
In [70]:

```
plot(v1,v3, cex=2.5,
     main = 'Correlação entre Quantidade Feijoada e Cervejas',
     pch=19,
     ylab='Quantidade de Cervejas no mês',
     xlab='Quantidade de Feijoadas no mês',
     cex.lab=1.5,
     cex.main=1.5)
```



In [71]:

```
plot(v1,v2, cex=2.5,
     main = 'Correlação entre Quantidade Feijoada e Refrigerantes',
     pch=19,
     ylab='Quantidade de Refrigerantes no mês',
     xlab='Quantidade de Feijoadas no mês',
     cex.lab=1.5,
     cex.main=1.5)
```



Vamos agorar criar um dataframe (lembram??) para facilitar a manipulação dos dados.

In [72]:
```
cons <- data.frame(v1, v2, v3,v4)
names(cons)=c('Feijoada','Caip','Cerv','Refr')
head(cons)
```

A data.frame: 6 × 4

	Feijoada	Caip	Cerv	Refr
	<dbl>	<dbl>	<dbl>	<dbl>
1	30	63	87	71
2	47	48	92	101
3	25	16	54	60
4	37	47	58	75
5	43	16	116	76
6	65	41	114	145

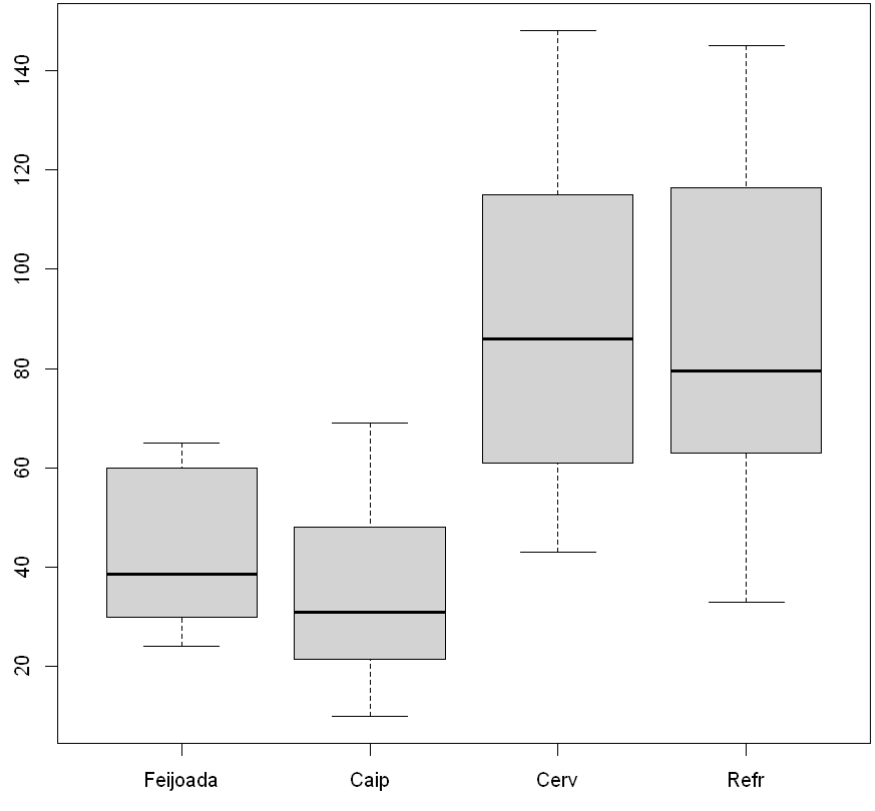
Vamos criar um sumário dos dados

In [73]:
```
summary(cons)
```

```
   Feijoada          Caip            Cerv            Refr       
Min.   :24.00   Min.   :10.00   Min.   : 43.00   Min.   : 33.00  
1st Qu.:30.00   1st Qu.:22.25   1st Qu.: 62.50   1st Qu.: 63.50  
Median :38.50   Median :31.00   Median : 86.00   Median : 79.50  
Mean   :43.71   Mean   :34.71   Mean   : 88.96   Mean   : 88.96  
3rd Qu.:60.00   3rd Qu.:48.00   3rd Qu.:114.50   3rd Qu.:115.75  
Max.   :65.00   Max.   :69.00   Max.   :148.00   Max.   :145.00  
```

Vamos criar um boxplot dos dados

In [74]:
```
boxplot(cons)
```



Vamos criar uma tabela de correlações

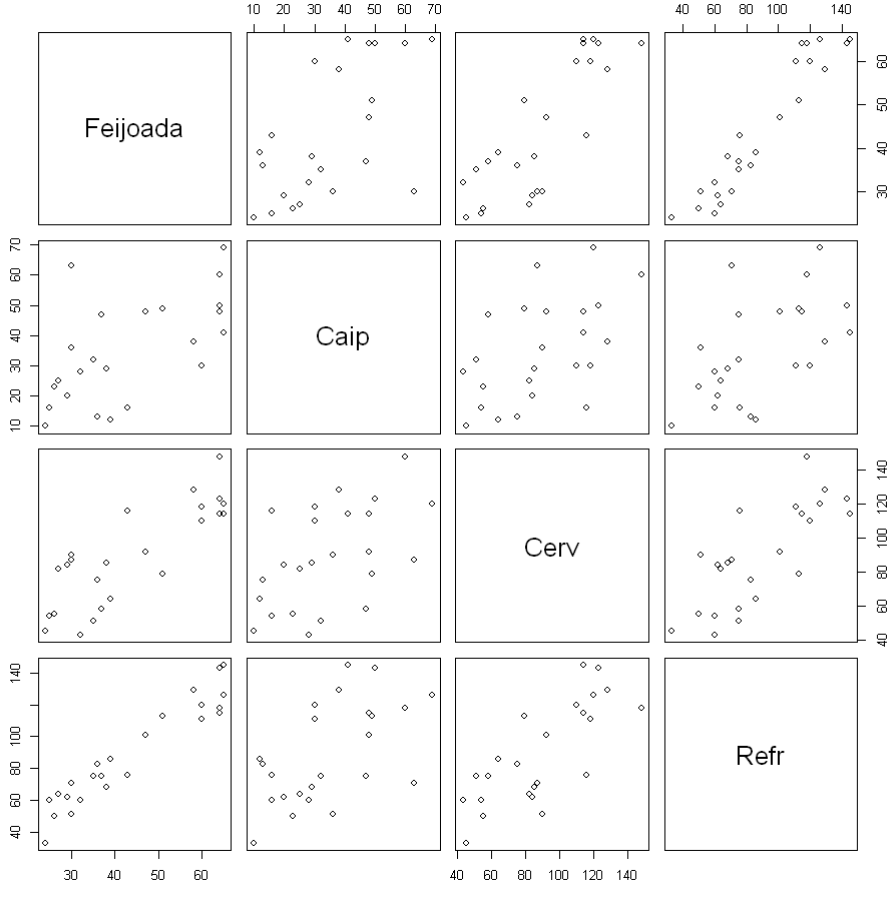
In [53]:
```
cor(cons)
```

A matrix: 4 × 4 of type dbl

	Feijoada	Caip	Cerv	Refr
Feijoada	1.0000000	0.5758686	0.8285306	0.9528823
Caip	0.5758686	1.0000000	0.5237354	0.5702872
Cerv	0.8285306	0.5237354	1.0000000	0.7701508
Refr	0.9528823	0.5702872	0.7701508	1.0000000

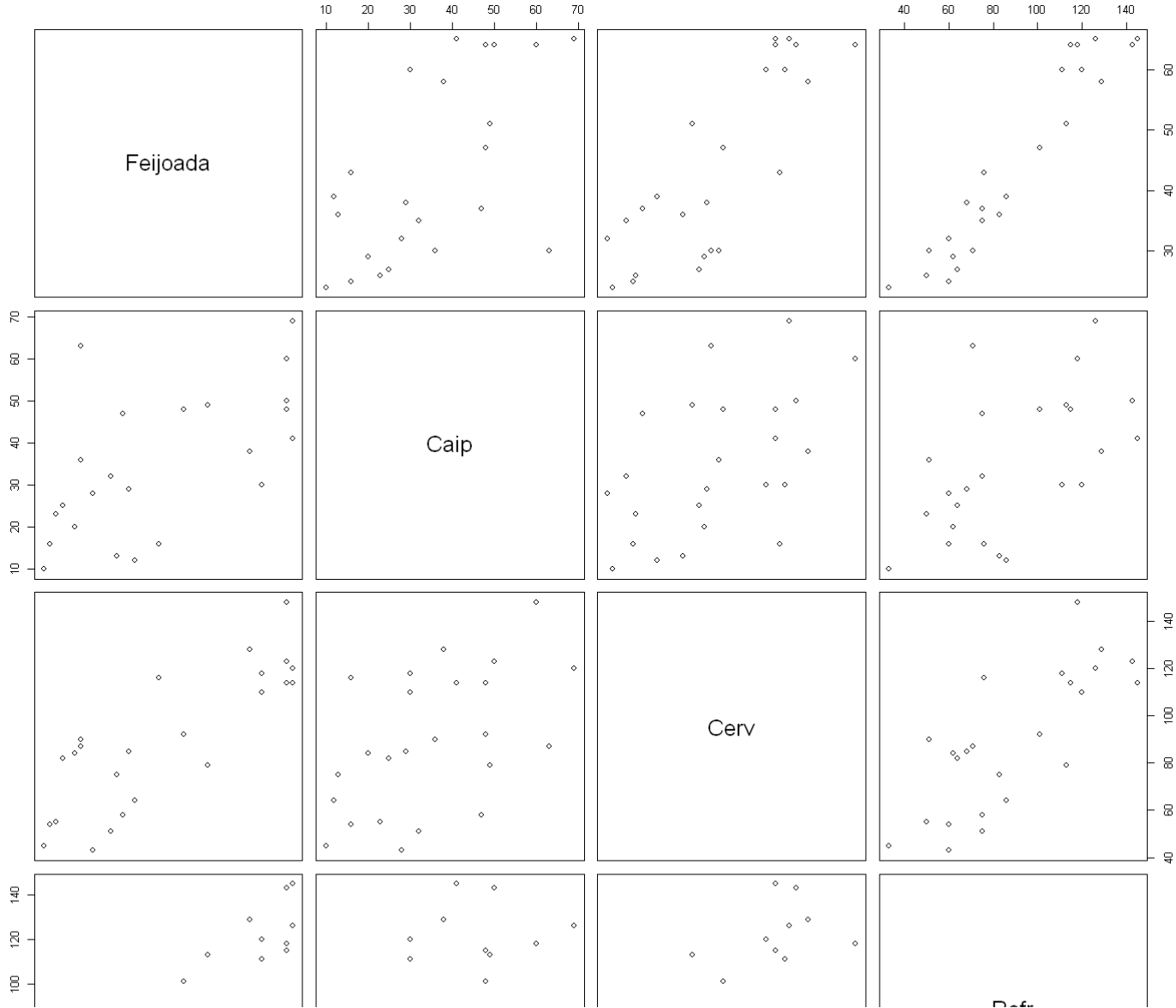
Vamos criar uma matriz de gráfcios para visaulizar todos os dados

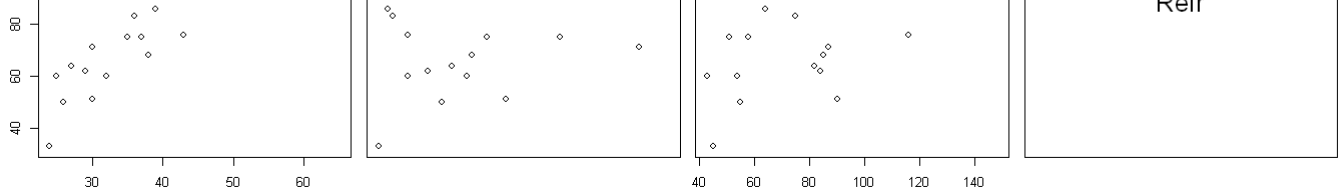
In [54]:
```
pairs(cons)
```



Melhorando a visaulização

In [55]:
```
library(repr)
options(repr.plot.width=12, repr.plot.height=12)
pairs(cons)
```

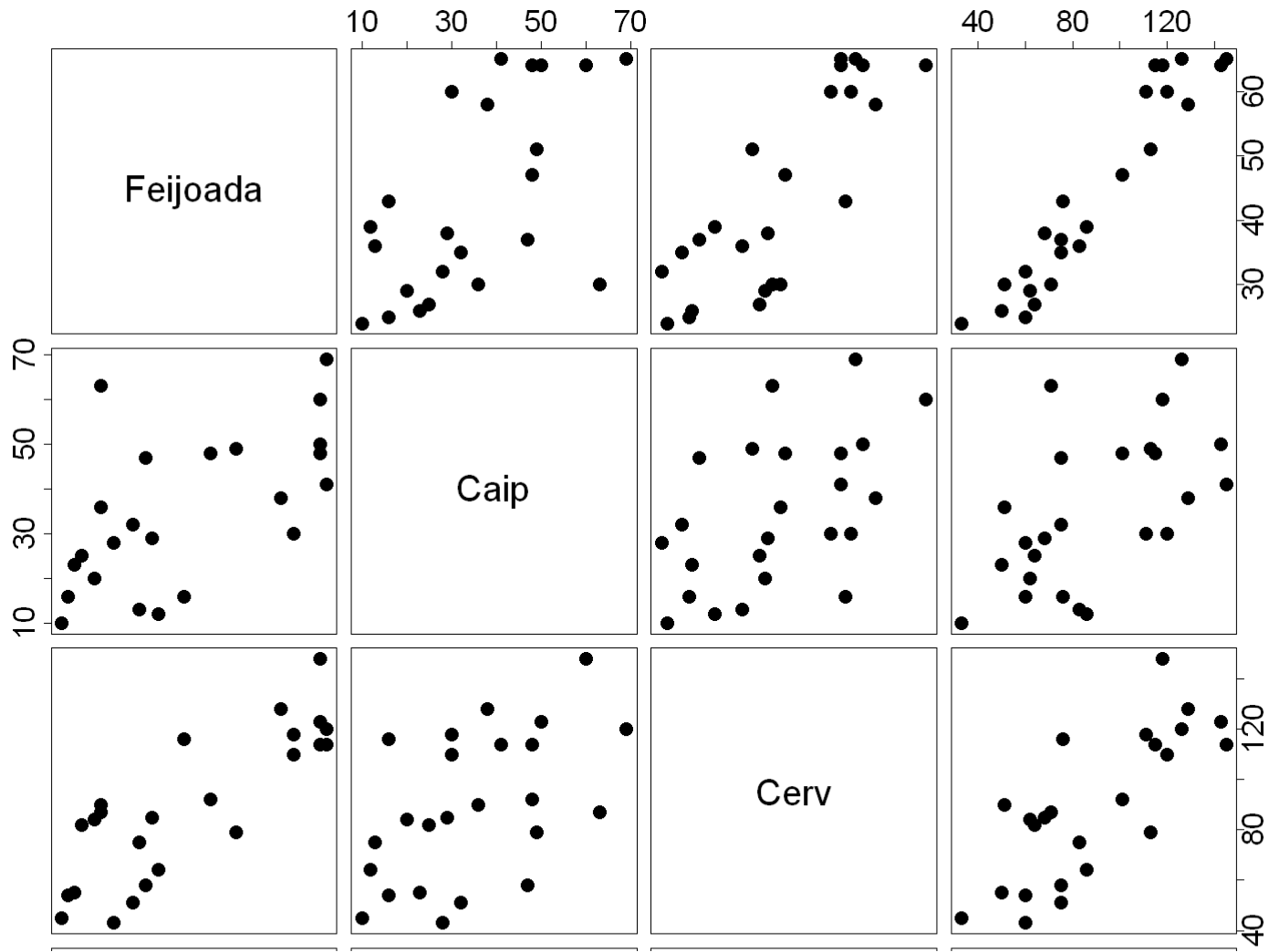


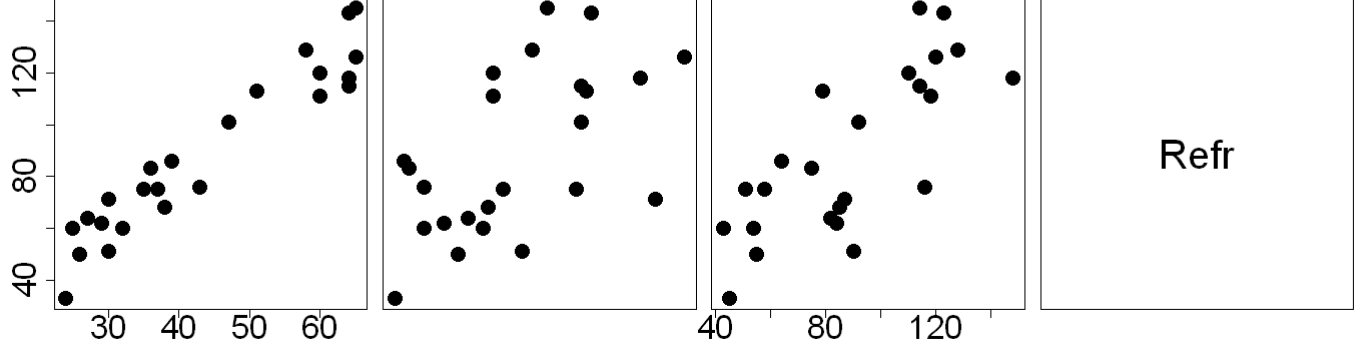

40
60
80
30
40
50
60
40
60
80
100
120
140
Reif

Melhorando a visaulização

In [56]:

```
library(repr)
options(repr.plot.width=12, repr.plot.height=12)
pairs(cons,
      cex=2.5,
      pch=19,
      cex.axis=2.5,
      cex.labels=3.0)
```

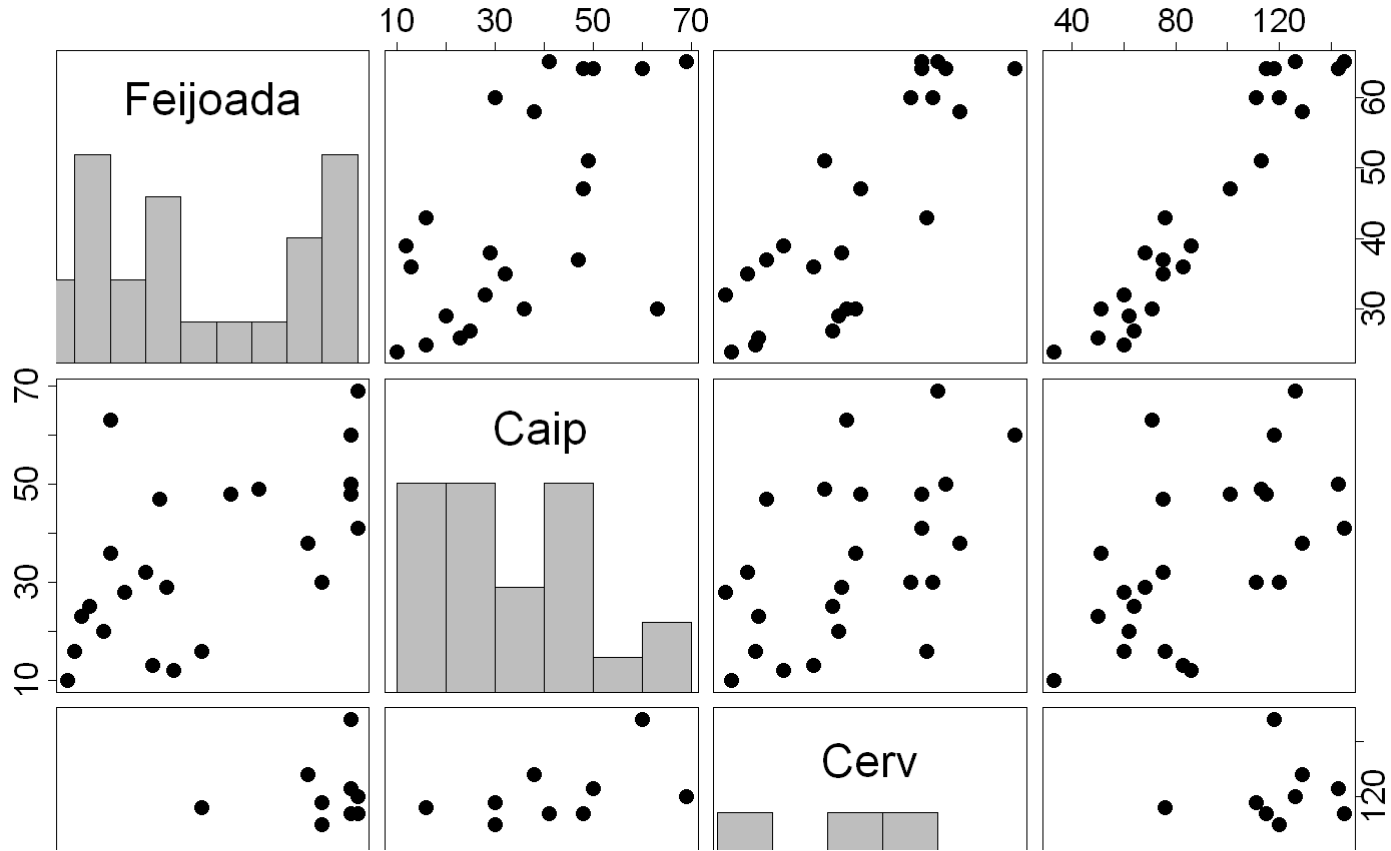


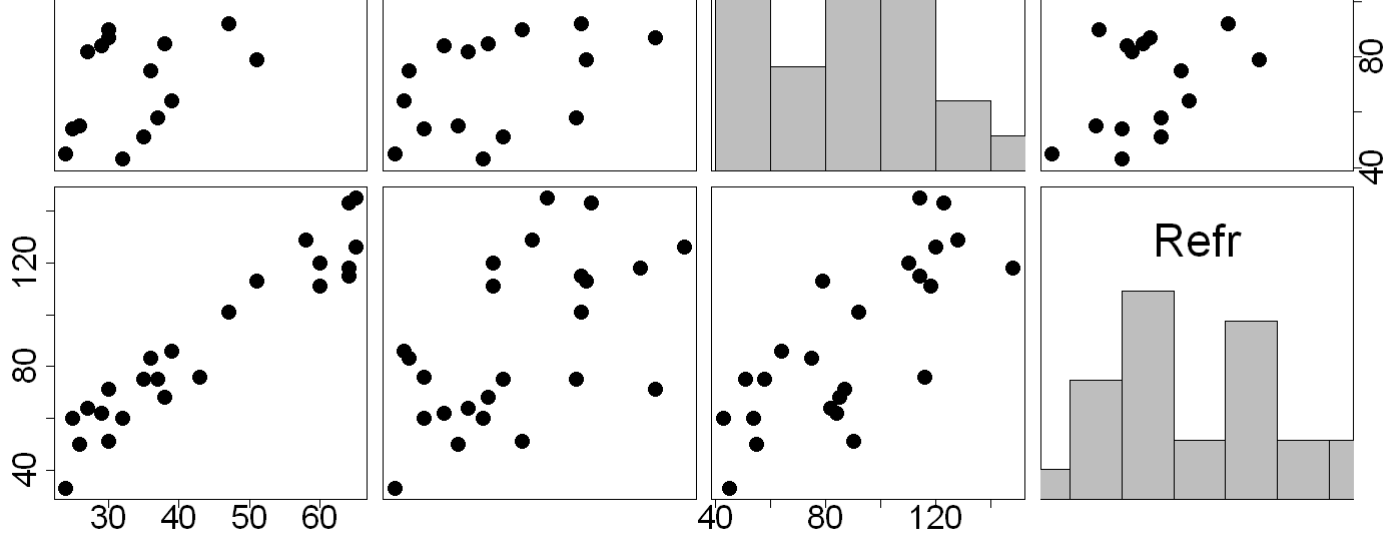

40
80
120
30
40
50
60
40
80
120
Refr

Melhorando a visaulização

In [57]:

```
#função retirada do help(pairs)
panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "gray", ...)
}
pairs(cons, diag.panel = panel.hist, cex=2.5, pch = 19, cex.lab=2.5, cex.axis=2.5, cex.m
ain=2.5, cex.labels = 3.5)
```

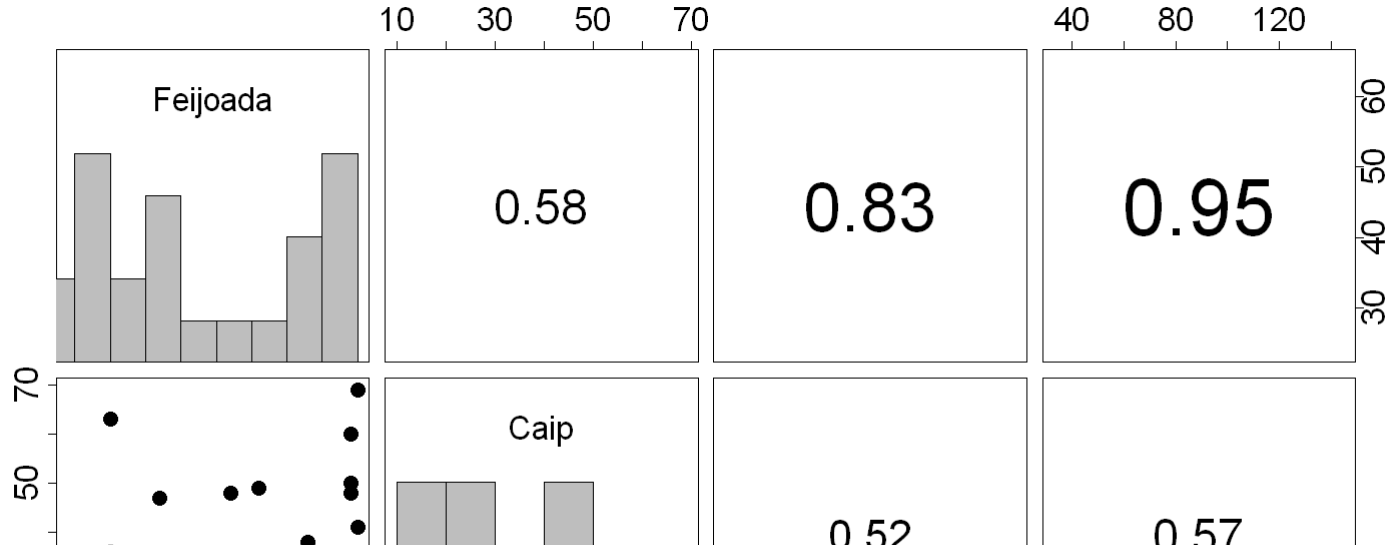


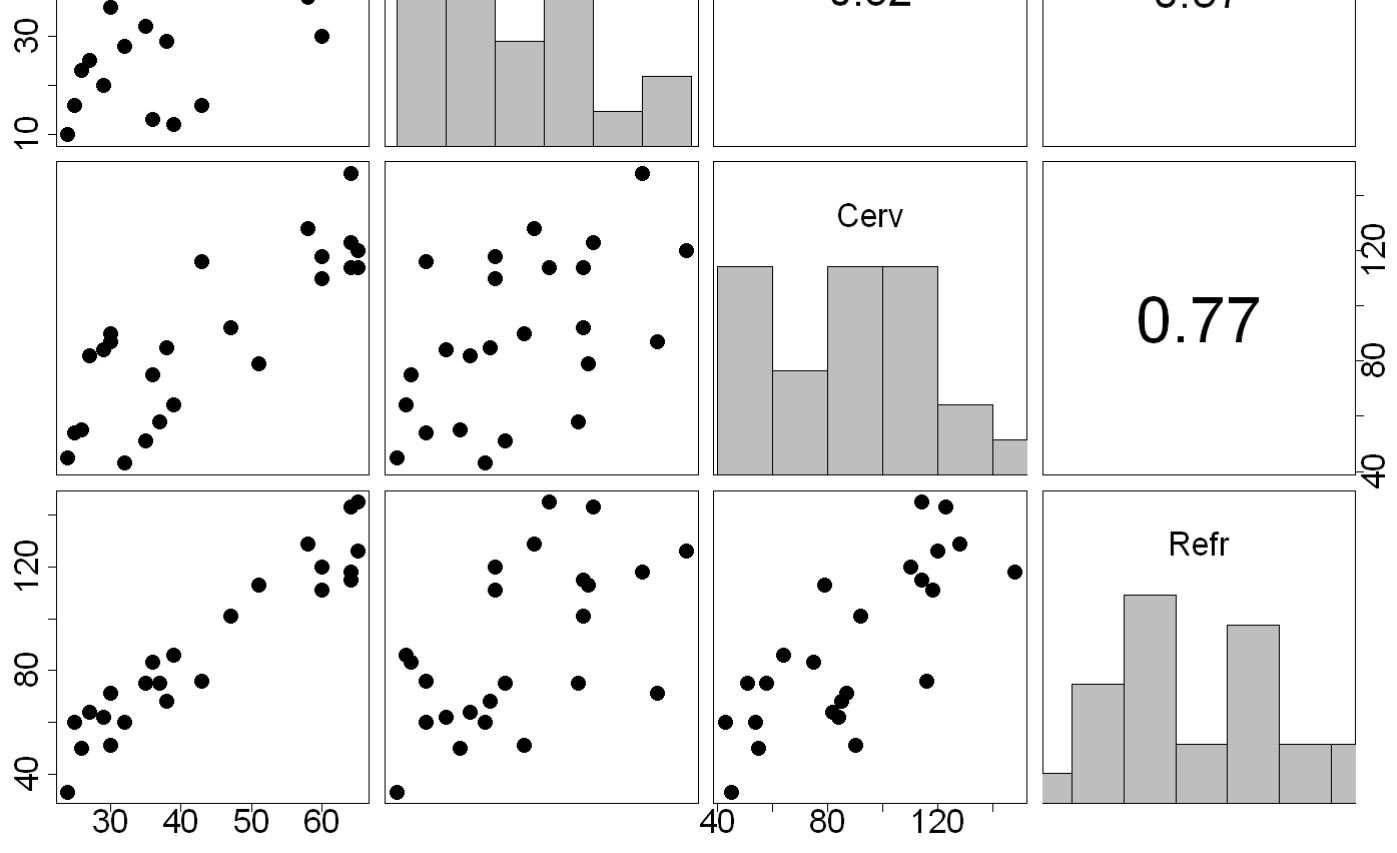

Refr
40
80
120
30
40
50
60
40
80
120
40
80

Melhorando a visaulização

In [58]:

```r
#função retirada do help(pairs)
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r*2.5)
}
pairs(cons,
      diag.panel = panel.hist,
      upper.panel = panel.cor,
      cex=2.5,
      pch = 19,
      cex.lab=2.5,
      cex.axis=2.5,
      cex.main=2.5,
      cex.labels = 2.5)
```

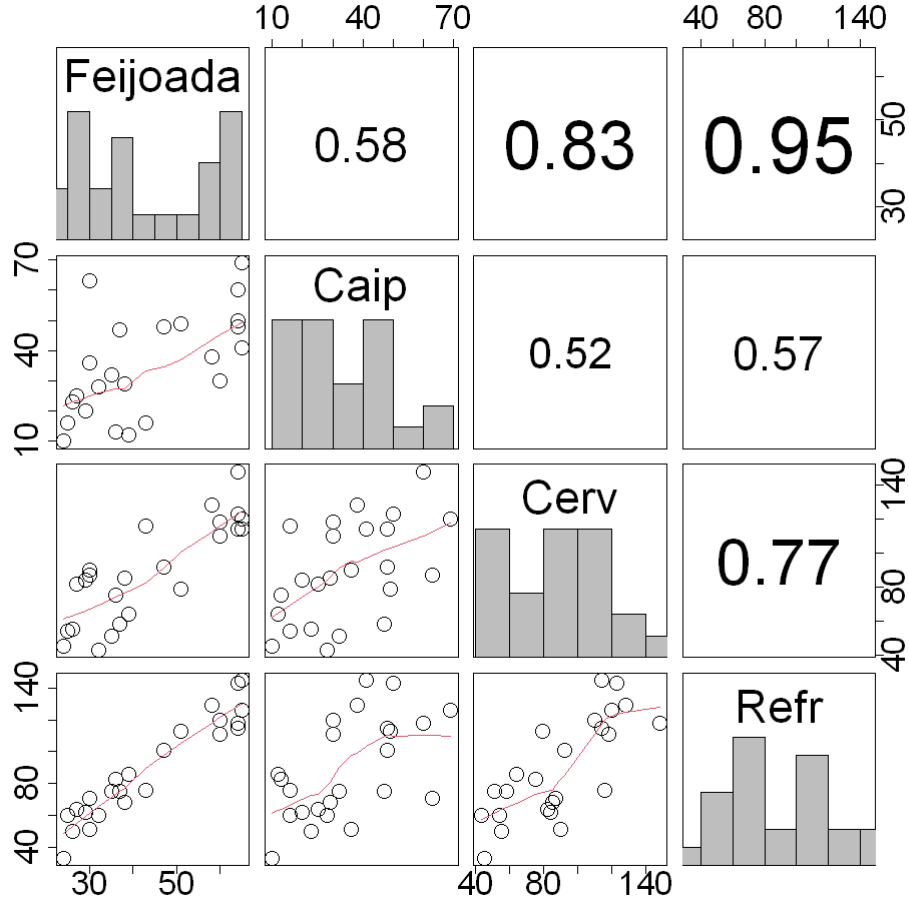



Cerv
0.77
Refr
10
30
40
80
120
30
40
50
60
40
80
120

Melhorando a visaulização

In [75]:

```
#panel.smooth já existe como função no R
pairs(cons, diag.panel = panel.hist, upper.panel = panel.cor,
      lower.panel = panel.smooth, cex=2.5, pch = 1, cex.lab=2.5, cex.axis=2.5, cex.main=
2.5, cex.labels = 3.5)
```



Instalando outro pacote

In [60]:
```
install.packages("GGally")
```

```
Warning message:
"package 'GGally' is in use and will not be installed"
```

In [61]:

```
library(repr)
library(repr)
options(repr.plot.width=8, repr.plot.height=8)
library("GGally")
ggcorr(cons, label=T, size=8, label_size = 12)
```