

PROGRAMAÇÃO

4. Visualização de Dados

Rita P. Ribeiro

Mestrado em Informática Médica

Curso de Especialização em Informática da Saúde

2016/2017



FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO



FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Programa

1. Conceitos Básicos do R
2. Manipulação de Dados
3. Sumarização de Dados
4. Visualização de Dados
5. Geração de Relatórios

Motivação

- ▶ Os seres humanos são excepcionais na deteção de padrões visuais
- ▶ Os métodos de visualização tentam tirar partido deste facto
- ▶ Apesar das suas vantagens os métodos de visualização de dados também têm várias limitações, particularmente com conjuntos de dados de elevada dimensão

Package `graphics`

Gráficos Standard do R

- ▶ Os gráficos standard, disponibilizados pela package **`graphics`**, incluem várias funções que produzem gráficos clássicos, como:
 - ▶ Scatterplots
 - ▶ Boxplots
 - ▶ Piecharts
 - ▶ Barplots
 - ▶ etc.
- ▶ Estes gráficos são obtidos tipicamente por uma chamada a uma função
 - ▶ Um exemplo de um scatterplot

```
plot(1:10,sin(1:10))
```

- ▶ Estes gráficos podem ser facilmente estendidos adicionando outros elementos aos mesmos (linhas, texto, etc.).

Package graphics

Graphics Devices

- ▶ As funções de gráficos do R produzem um output que depende do *device* gráfico que está ativo
- ▶ Por omissão este *device* é o écran
- ▶ O R possui muitos outros *devices* gráficos, como o *device* para **pdf**, o *device* para **jpeg**, etc.
- ▶ O utilizador unicamente tem que abrir (e no fim fechar) o *device* para ter o gráfico no tipo de output que pretende. O R trata de produzir o **tipo de output que é exigido pelo device**
- ▶ Isto significa que obter um certo gráfico no écran ou como um ficheiro GIF é exatamente igual. Só temos que inicialmente abrir o *device* GIF e no fim fechá-lo!
- ▶ Podemos ter vários *devices* abertos ao mesmo tempo mas só um pode estar ativo

Package graphics

Alguns exemplos

Um scatterplot

```
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))
```

O mesmo mas gravado num ficheiro jpeg

```
jpeg('exp.jpg')  
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))  
dev.off()
```

E agora como pdf

```
pdf('exp.pdf',width=6,height=6)  
plot(seq(0,4*pi,0.1),sin(seq(0,4*pi,0.1)))  
dev.off()
```

Package `ggplot2`

- ▶ A package `ggplot2` implementa as ideias de Wilkinson (2005) sobre uma gramática para gráficos
- ▶ Esta gramática é o resultado de um estudo teórico sobre os diferentes de gráficos estatísticos que existem
- ▶ `ggplot2` é baseada nesta teoria implementando-a pela noção de gráficos em camadas (Wickham, 2009)
- ▶ A gramática define um gráfico estatístico como:
 - ▶ um mapeamento dos dados para um conjunto de **propriedades estéticas** (côr, forma, tamanho, etc.) de um conjunto de **objetos gráficos** (pontos, linhas, barras, etc.)

L. Wilkinson (2005). The Grammar of Graphics. Springer.

H. Wickham (2009). A layered grammar of graphics. Journal of Computational and Graphical Statistics.

Package ggplot2

Os Conceitos Básicos da Gramática de Gráficos

- ▶ Ideia base: especificar, de forma independente, as camadas que compõem o gráfico.
- ▶ Através do sinal + vamos adicionando as camadas.
- ▶ Elementos chave de um gráfico estatístico:
 - ▶ dados
 - ▶ mapeamentos estéticos
 - ▶ objetos gráficos
 - ▶ transformações estatísticas
 - ▶ escalas
 - ▶ sistemas de coordenadas
 - ▶ facetas

Package ggplot2

Mapeamentos Estéticos

- ▶ Aquilo que conseguimos ver.
- ▶ Definem a relação entre as variáveis dos dados e as variáveis nos gráficos.
 - ▶ mapear a *Temperatura* num conjunto de dados para a coordenada do eixo dos x num scatter plot
 - ▶ mapear o *Tamanho* de um conjunto de produtos para a *côr* de pontos de um gráfico
 - ▶ etc.
- ▶ Exemplos incluem:
 - ▶ position: `aes(x=...,y=...)`
 - ▶ color: `aes(color=...)`
 - ▶ fill: `aes(fill=...)`
 - ▶ shape: `aes(shape=...)`
 - ▶ linetype: `aes(linetype=...)`
 - ▶ size: `aes(size=...)`
 - ▶ etc.

Package `ggplot2`

Objetos Geométricos

- ▶ Definem o que é visto nos gráficos
 - ▶ mostrar cada observação de um conjunto de dados através de um ponto usando um mapeamento estético que usa como valor para o eixo do x do ponto o valor de uma variável da observação, e como valor para o y outra variável
 - ▶ etc.
- ▶ Exemplos incluem:
 - ▶ pontos: `geom_point()`
 - ▶ linhas: `geom_line()`
 - ▶ boxplot: `geom_boxplot()`
 - ▶ histogramas: `geom_histogram()`
 - ▶ gráficos de barras: `geom_bar()`
 - ▶ etc.

Package ggplot2

Transformações Estatísticas

- ▶ Permitem calcular e fazer análise estatística sobre os dados no gráfico
 - ▶ aproximar os dados (representados por ponto) por uma reta
 - ▶ contar as ocorrências de certos valores nos dados e mostrar essas contagens
 - ▶ etc.
- ▶ Exemplos incluem:
 - ▶ `stat_count(geom="bar") / geom_bar(stat="count")`
 - ▶ `..density..`, `..count..`: variáveis criadas pela estatística
 - ▶ `stat_unique()`
 - ▶ etc.

Package `ggplot2`

Escalas

- Mapeam os valores dos dados nos valores de um certo sistema de coordenadas do device

scale	Tipos	Exemplos
<code>scale_color_</code>	identity	<code>scale_color_discrete()</code>
<code>scale_fill_</code>	manual	<code>scale_fill_continuous()</code>
<code>scale_size_</code>	continuous	<code>scale_size_manual()</code>
	discrete	<code>scale_size_discrete()</code>
<code>scale_shape_()</code>	discrete	<code>scale_shape_discrete()</code>
<code>scale_linetype_()</code>	identity	<code>scale_shape_manual()</code>
	manual	<code>scale_linetype_discrete()</code>
<code>scale_x_()</code>	continuous	<code>scale_x_continuous()</code>
<code>scale_y_()</code>	discrete	<code>scale_y_discrete()</code>
	reverse	<code>scale_x_reverse()</code>
	log	<code>scale_y_log()</code>
	date	<code>scale_x_date()</code>
	datetime	<code>scale_y_datetime()</code>

Package `ggplot2`

Sistema de Coordenadas

- ▶ O sistema de coordenadas usado para desenhar os dados
- ▶ Exemplos incluem:
 - ▶ Cartesiano: `coord_cartesian()`
 - ▶ Polar: `coord_polar()`
 - ▶ etc.

Package `ggplot2`

Facetas

- ▶ Dividir os dados em sub-grupos e desenhar gráficos para cada sub-grupo para melhor contrastar os valores
- ▶ Exemplos incluem:
 - ▶ `facet_wrap()`: define sub-grupos em função das categorias de uma variável
 - ▶ `facet_grid()`: define sub-grupos em função do cruzamento das categorias de duas variáveis

Package ggplot2

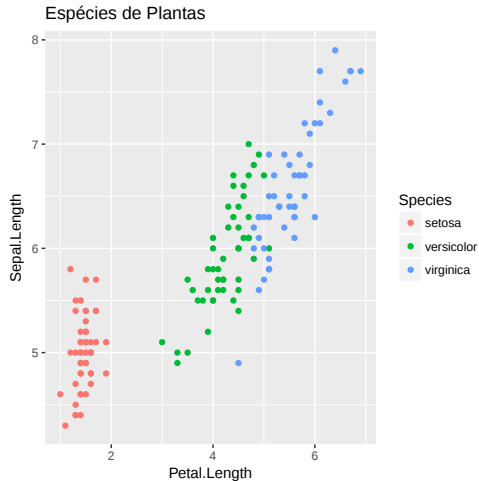
Etiquetas

- ▶ Etiquetar os eixos dos x e dos y e o título do gráfico.
- ▶ Exemplos incluem:
 - ▶ `ggtitle()`
 - ▶ `xlab()`
 - ▶ `ylab()`
 - ▶ `labs(title=...,x=...,y=...)`

Package ggplot2

Um Exemplo Simples

```
data(iris)
library(ggplot2)
ggplot(iris,
       aes(x=Petal.Length,
           y=Sepal.Length,
           color=Species)) +
  geom_point() +
  ggtitle("Espécies de Plantas")
ggsave("plot.png",width=5,height=5)
```



Gráficos Univariados

Variável Discreta - Gráfico de Barras

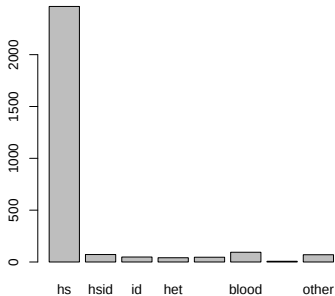
- ▶ O Gráfico de Barras mostra um conjunto de valores como alturas de um conjunto de barras
- ▶ Este gráfico pode ser usado para mostrar a frequência de ocorrência de cada um dos valores de uma variável discreta/nominal da seguinte forma:
 - ▶ começar por obter o número de ocorrências de cada valor
 - ▶ em seguida usar o gráfico de barras para mostrar estes números

Gráficos Univariados

Variável Discreta - Gráfico de Barras no `graphics`

```
data(Aids2, package="MASS")  
barplot(table(Aids2$T.categ), main='Distribuição do Tipo de Transmissão do Vírus')
```

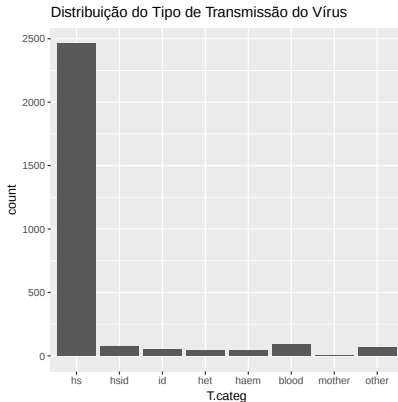
Distribuição do Tipo de Transmissão do Vírus



Gráficos Univariados

Variável Discreta - Gráfico de Barras no *ggplot2*

```
ggplot(Aids2, aes(x=T.categ)) + geom_bar() +  
  ggtitle('Distribuição do Tipo de Transmissão do Vírus')
```



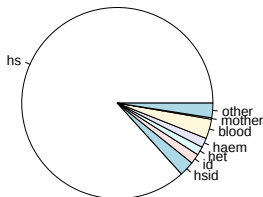
Gráficos Univariados

Variável Discreta - Gráfico Circular

Os gráficos circulares servem o mesmo propósito dos gráficos de barras mas mostram a informação sobre a forma de fatias de uma pizza.

```
pie(table(Aids2$T.categ),main='Distribuição do Tipo de Transmissão do Vírus')
```

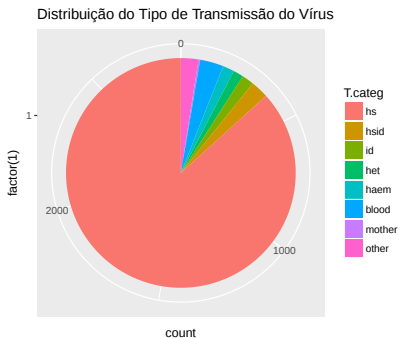
Distribuição do Tipo de Transmissão do Vírus



Gráficos Univariados

Variável Discreta - Gráfico Circular no *ggplot2*

```
ggplot(Aids2,aes(x=factor(1),fill=T.categ)) + geom_bar(width=1) +
  ggtitle('Distribuição do Tipo de Transmissão do Vírus') +
  coord_polar(theta="y")
```



Gráficos Univariados

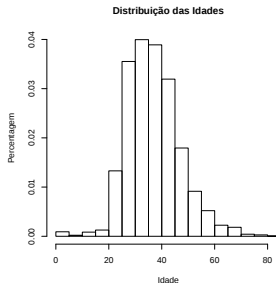
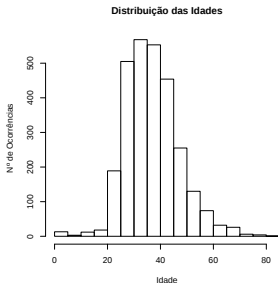
Variável Contínua - Histograma

- ▶ O Histograma mostra como os valores de uma variável contínua se distribuem
- ▶ Pode ser obtido da seguinte forma:
 - ▶ em primeiro lugar, a gama de valores da variável é dividida num conjunto de intervalos de valores
 - ▶ em seguida, o número de ocorrências de valores em cada intervalo é contado
 - ▶ finalmente, estes números são mostrados como barras

Gráficos Univariados

Variável Contínua - Histograma no *graphics*

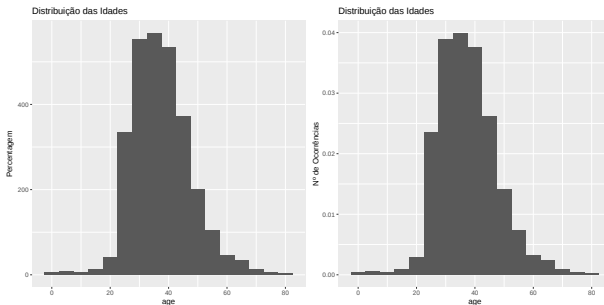
```
hist(Aids2$age,main='Distribuição das Idades',
     xlab='Idade',ylab='Nº de Ocorrências')
hist(Aids2$age,main='Distribuição das Idades',
     xlab='Idade',ylab='Percentagem',prob=TRUE)
```



Gráficos Univariados

Variável Contínua - Histograma no ggplot2

```
ggplot(Aids2,aes(x=age)) + geom_histogram(binwidth=5) +
  ggtitle("Distribuição das Idades") + ylab("Percentagem")
ggplot(Aids2,aes(x=age)) + geom_histogram(binwidth=5,aes(y=..density..)) +
  ggtitle("Distribuição das Idades") + ylab("Nº de Ocorrências")
```



Gráficos Univariados

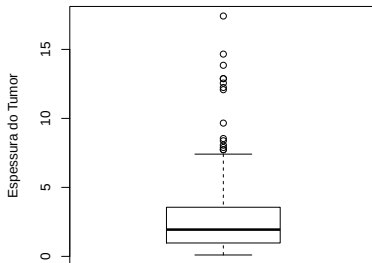
Variável Contínua - Box Plot

- ▶ Os *box plots* fornecem um sumário interessante da distribuição de uma variável contínua
- ▶ Por exemplo, informam-nos sobre o intervalo inter-quartil, bem como da existência de eventuais outliers

Gráficos Univariados

Variável Contínua - Box Plot no `graphics`

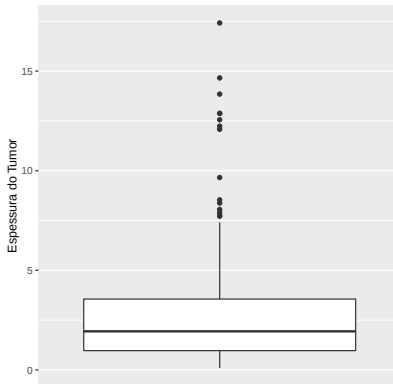
```
data(Melanoma, package="MASS")  
boxplot(Melanoma$thickness, ylab="Espessura do Tumor")
```



Gráficos Univariados

Variável Contínua - Box Plot no *ggplot2*

```
ggplot(Melanoma, aes(x=factor(0), y=thickness)) + geom_boxplot() +  
  ylab("Espessura do Tumor") + xlab("") + scale_x_discrete(breaks=NULL)
```



Gráficos Condicionados

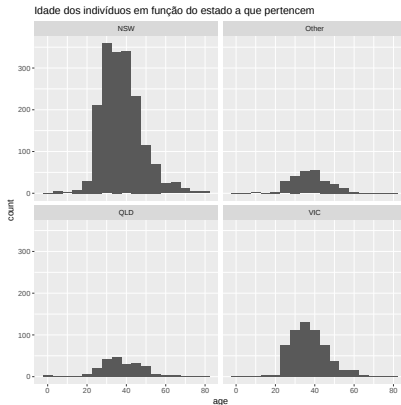
- ▶ Os conjuntos de dados possuem frequentemente variáveis nominais cujos valores podem ser usados para criar **sub-grupos** desses dados
 - ▶ p.ex. o sub-grupo de indivíduos infectados por transfusão de sangue.
- ▶ Alguns dos gráficos obtidos anteriormente podem ser obtidos para cada um dos sub-grupos
- ▶ O objetivo dos gráficos condicionados é a **apresentação simultânea dos gráficos destes sub-grupos** para permitir **encontrar eventuais diferenças entre eles**
- ▶ Os gráficos base não têm condicionamento mas os gráficos `ggplot2` possuem essa capacidade através da noção de facetas

Gráficos Condicionados

Histogramas Condicionados

Objetivo: Contrastar a distribuição dos sub-grupos

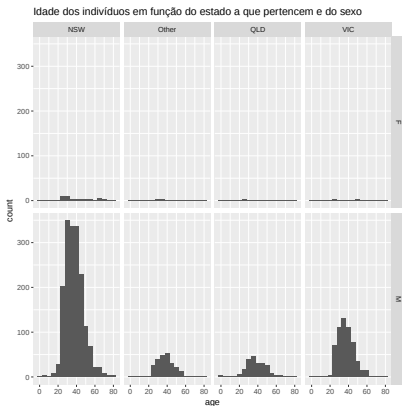
```
ggplot(Aids2,aes(x=age)) + geom_histogram(binwidth=5) + facet_wrap(~ state) +  
  ggtitle('Idade dos indivíduos em função do estado a que pertencem')
```



Gráficos Condicionados

Histogramas Condicionados - 2

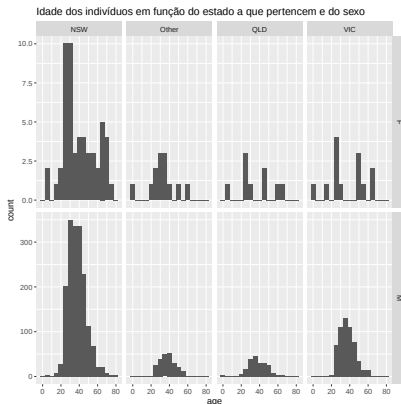
```
ggplot(Aids2, aes(x=age)) + geom_histogram(binwidth=5) + facet_grid(sex ~ state) +  
  ggtitle('Idade dos indivíduos em função do estado a que pertencem e do sexo')
```



Gráficos Condicionados

Histogramas Condicionados - 3

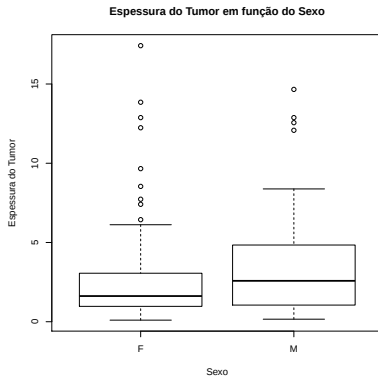
```
ggplot(Aids2,aes(x=age)) + geom_histogram(binwidth=5) +  
  facet_grid(sex ~ state,scales = "free_y") +  
  ggtitle('Idade dos indivíduos em função do estado a que pertencem e do sexo')
```



Gráficos Condicionados

Box Plots Condicionados no `graphics`

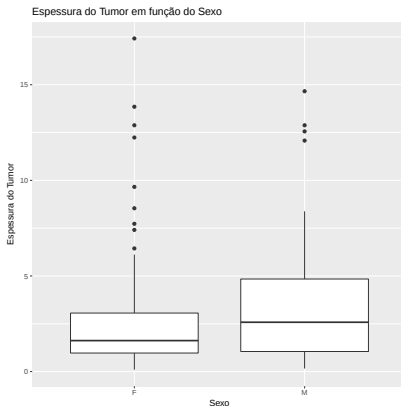
```
Melanoma$sex <- factor(Melanoma$sex, labels=c("F", "M"))  
boxplot(thickness ~ sex, Melanoma,  
        main='Espessura do Tumor em função do Sexo',  
        xlab='Sexo', ylab='Espessura do Tumor')
```



Gráficos Condicionados

Box Plots Condicionados no *ggplot2*

```
ggplot(Melanoma, aes(x=sex, y=thickness)) + geom_boxplot() +  
  xlab("Sexo") + ylab("Espessura do Tumor") +  
  ggtitle('Espessura do Tumor em função do Sexo')
```

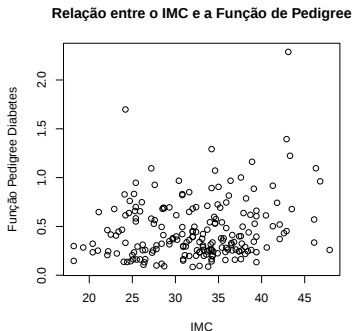


Gráficos Multivariados

Scatterplots no graphics

- O *Scatterplot* (Gráfico de Dispersão) é o gráfico mais usado para mostrar a relação entre duas variáveis

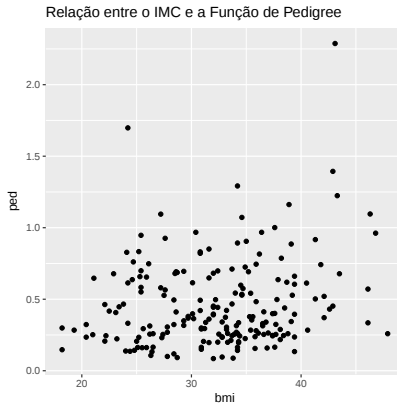
```
plot(Pima.tr$bmi,Pima.tr$ped,  
     main='Relação entre o IMC e a Função de Pedigree',  
     xlab='IMC',ylab='Função Pedigree Diabetes')
```



Gráficos Multivariados

Scatterplots no *ggplot2*

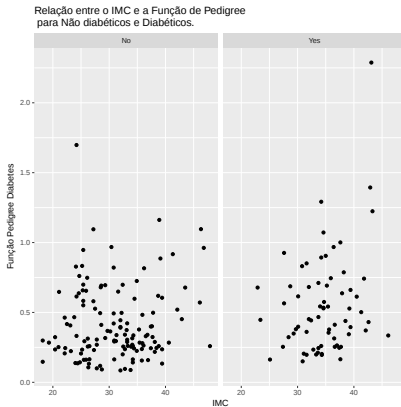
```
ggplot(Pima.tr, aes(x=bmi, y=ped)) + geom_point()+  
  ggtitle('Relação entre o IMC e a Função de Pedigree')
```



Gráficos Multivariados

Scatterplots Condicionados no *ggplot2*

```
ggplot(Pima.tr, aes(x=bmi, y=ped)) + geom_point() + facet_wrap(~ type) +  
  xlab("IMC") + ylab("Função Pedigree Diabetes") +  
  ggtitle(paste('Relação entre o IMC e a Função de Pedigree',  
                '\n para Não diabéticos e Diabéticos.'))
```

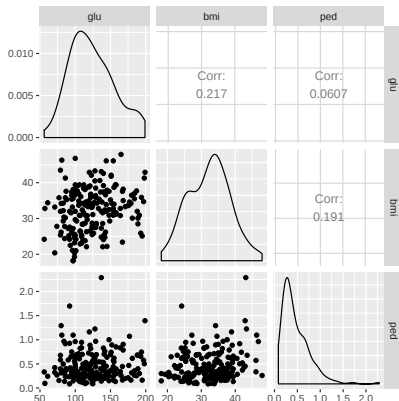


Gráficos Multivariados

Matrizes de Scatterplots no GGally

- Tentam apresentar todos os pares de scatterplots num conjunto de dados. Não são realizáveis para muitas variáveis

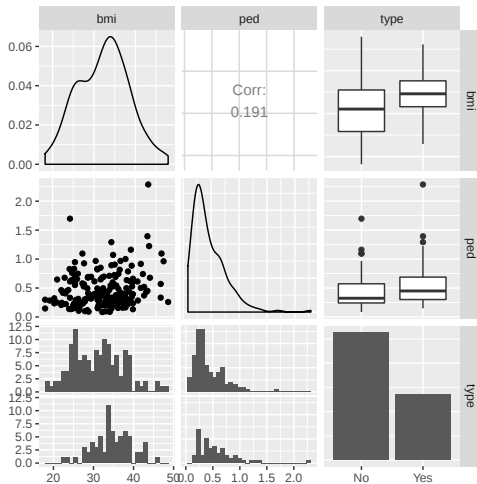
```
library(GGally)
ggpairs(Pima.tr, columns=c("glu", "bmi", "ped"),
        diag=list(continuous="densityDiag", discrete="barDiag"), axisLabels="show")
```



Gráficos Multivariados

Matrizes de Scatterplots envolvendo variáveis nominais

```
ggpairs(Pima.tr, columns=c("bmi", "ped", "type"),
        diag=list(continuous="densityDiag", discrete="barDiag"), axisLabels="show")
```



Gráficos Multivariados

Parallel Plots

- Os gráficos *Parallel* fornecem formas interessantes de olhar para um conjunto de dados

```
ggparcoord(Pima.tr, columns=c(2,5,6), groupColumn="type")
```

