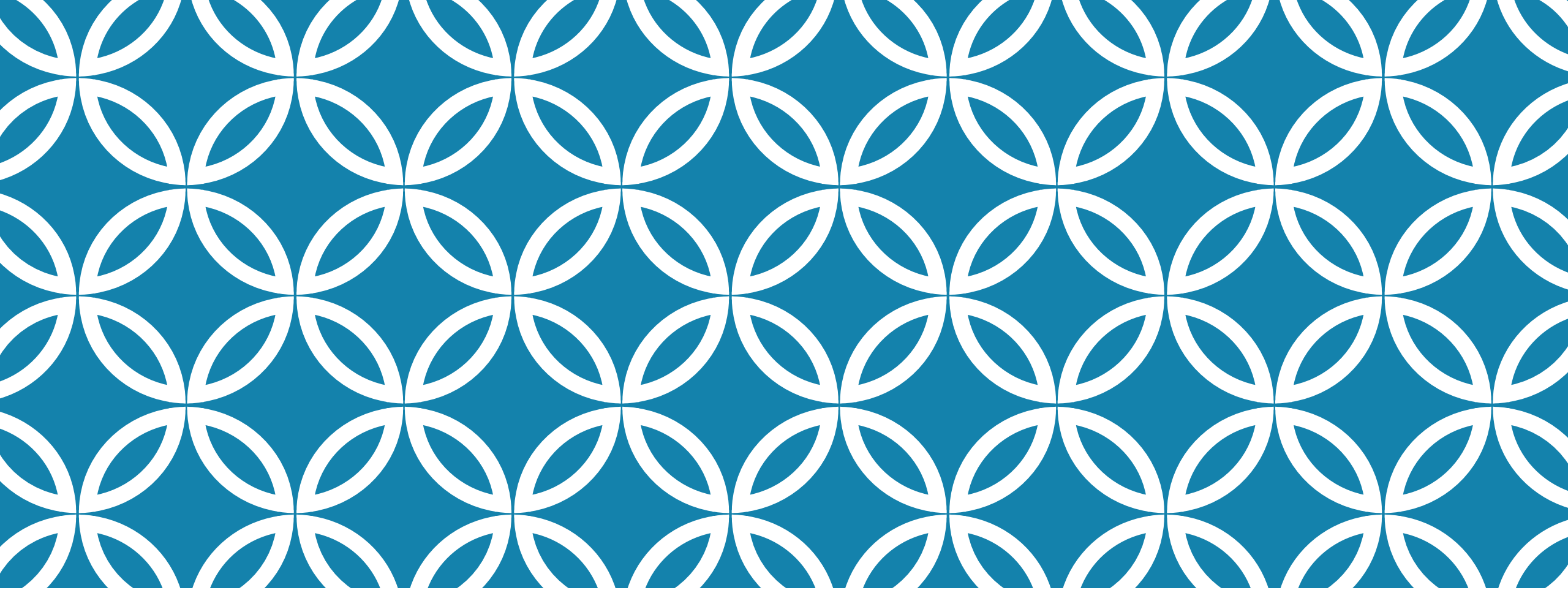


LAB 2 – ANÁLISE E VISUALIZAÇÃO DOS DADOS

André Gustavo Adami
Daniel Luis Notari

ANÁLISE DOS DADOS

A análise dos dados tem por objetivo compreender melhor as suas características com o objetivo de permitir a aplicação de métodos apropriados

Uma análise permite descobrir possibilidades, insights, padrões, desafios e até erros

Diversas ferramentas de estatística e visualização são empregadas para a realização desta etapa

Nesta etapa, assumimos que o conjunto de dados está “limpo” (i.e., não possui valores faltantes)

PACOTE

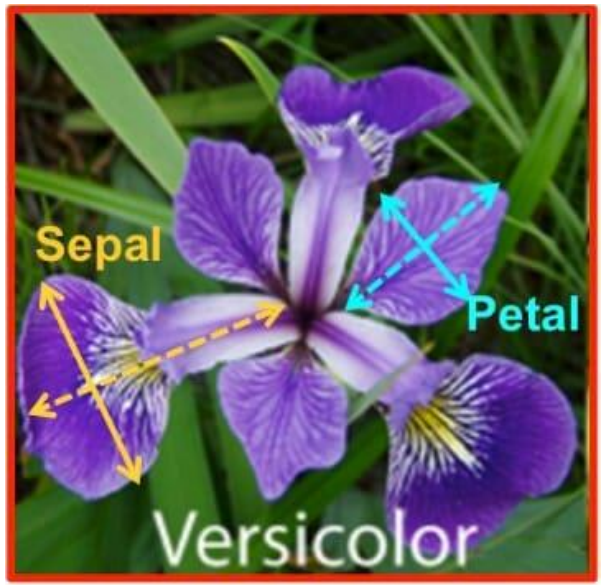
Para este laboratório vamos precisar dos seguintes pacotes para a criação de alguns gráficos

- ggplot2: criação de gráficos (gramática de gráficos - camadas) - https://cran.r-project.org/web/packages/ggplot2
- gridExtra: plotar múltiplos gráficos em uma única figura - https://cran.r-project.org/web/packages/gridExtra
- Ggally: extensão do pacote ggplot2, com uma série de novas funções gráficas - https://cran.r-project.org/web/packages/GGally

```
install.packages("ggplot2", dependencies = TRUE)
install.packages("gridExtra",dependencies = TRUE)
install.packages("GGally", dependencies = TRUE)
```

ANÁLISE DOS DADOS

A base de dados íris consiste de 150 amostras de três espécies da flor íris: setosa, virginica e versicolor. Cada amostra possui medidas do comprimento e a largura das sépalas e pétalas, em centímetros.



ANÁLISE DOS DADOS

Em uma primeira análise, é possível verificar que existem 150 amostras (observações) e 5 variáveis (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width e Species)

```
> data("iris")
> str(iris)
'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

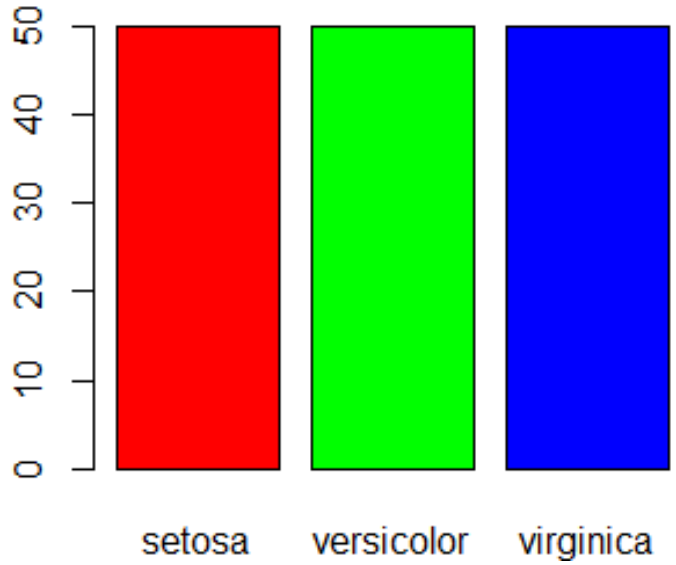
As variáveis *Sepal.Length*, *Sepal.Width*, *Petal.Length*, *Petal.Width* são da classe *numeric* (double)

A variável *Species* é do tipo Factor com os rótulos de cada espécie da flor Íris

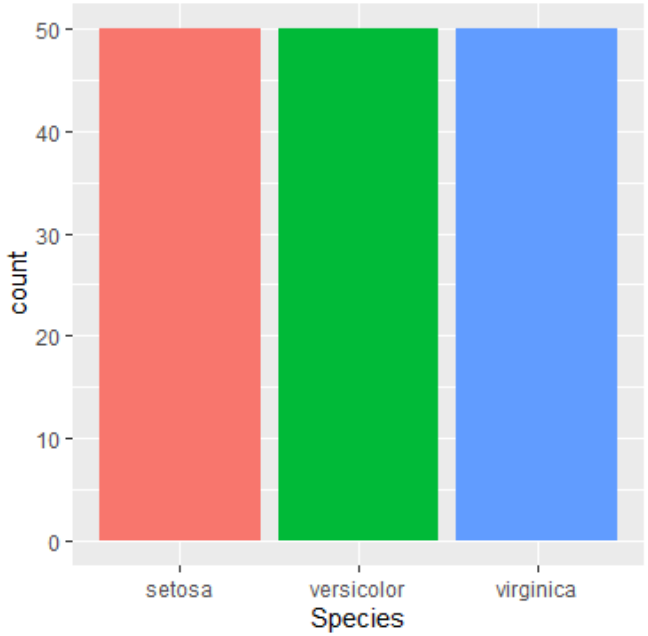
ANÁLISE DOS DADOS

A quantidade de amostras por classes pode ser verificada

```
> summary(iris$Species)
    setosa versicolor  virginica
        50         50         50
barplot(summary(iris$Species),
            col=c("red","green","blue"))
```




```
library(ggplot2)
ggplot(iris,aes(Species,fill=Species)) +
  geom_bar()+theme(legend.position="none")
```



Existem 50 amostras por classe (a quantidade é balanceada entre as classes)

ANÁLISE ESTATÍSTICA

Em uma primeira análise dos dados, podemos estimar diversas medidas estatísticas básicas

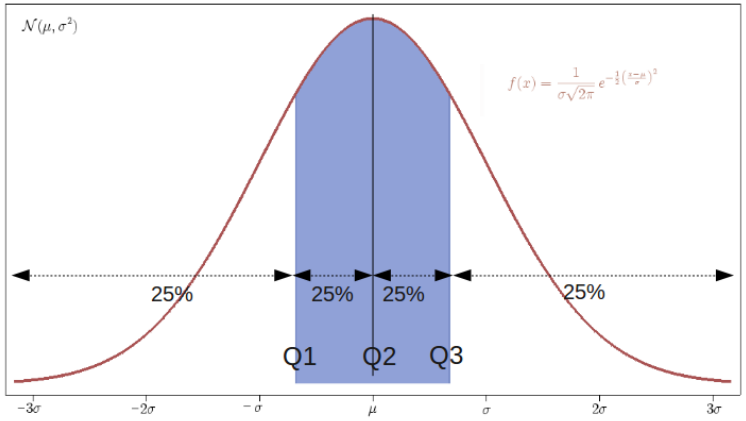
A função `summary()` retorna o mínimo, máximo, os 1° e 3° quartis, a média e a mediana de cada variável

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

ANÁLISE ESTATÍSTICA - QUARTIS

Os Quartis (Q1, Q2 e Q3) são os valores de uma população, ordenada de forma crescente, que dividem a distribuição dos valores de uma variável em 4 grupos iguais

- Mediana é Q2, i.e., divide a população no meio



O intervalo interquartis é a dispersão dos dados em torno do centro da distribuição (distância entre Q1 e Q3)

- Concentração de 50% da distribuição

Permite avaliar que tipo de distribuição é mais apropriada

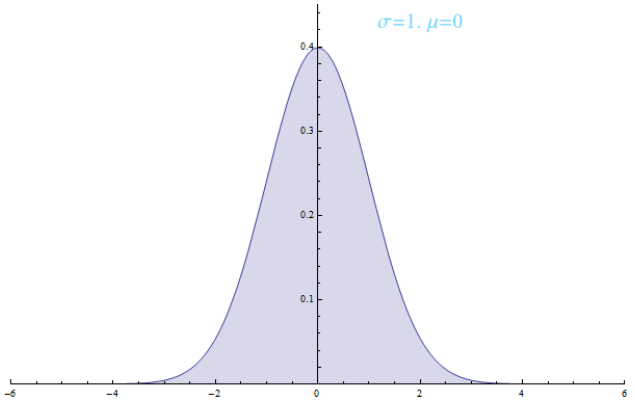
ANÁLISE ESTATÍSTICA

A média é a medida da tendência central de qualquer distribuição de dados

```
> mean(iris$Sepal.Length)
[1] 5.843333
```

A média não é um descritor de dados suficiente, pois a dispersão dos dados em torno da média pode ser diferente. Por isso, utiliza-se a medida da variância σ^2 (quadrado do desvio padrão σ)

```
> var(iris$Sepal.Length)
[1] 0.6856935
> sd(iris$Sepal.Length)
[1] 0.8280661
```



ANÁLISE ESTATÍSTICA

A mediana é mais robusta que a média, pois valores atípicos (outliers) não afetam a estimação desta medida

```
> median(iris$Sepal.Length)
[1] 5.8
```

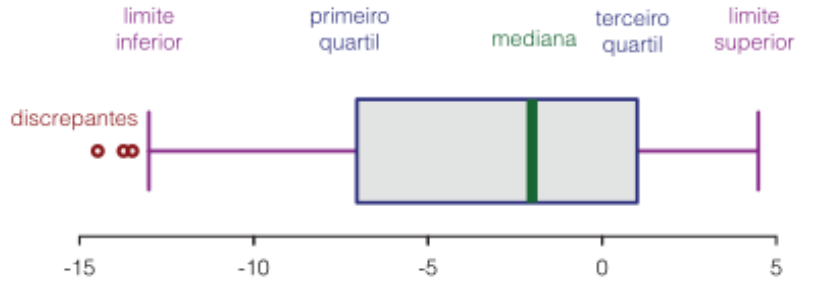
Quando a média e a mediana são iguais, isto quer dizer que a distribuição é simétrica

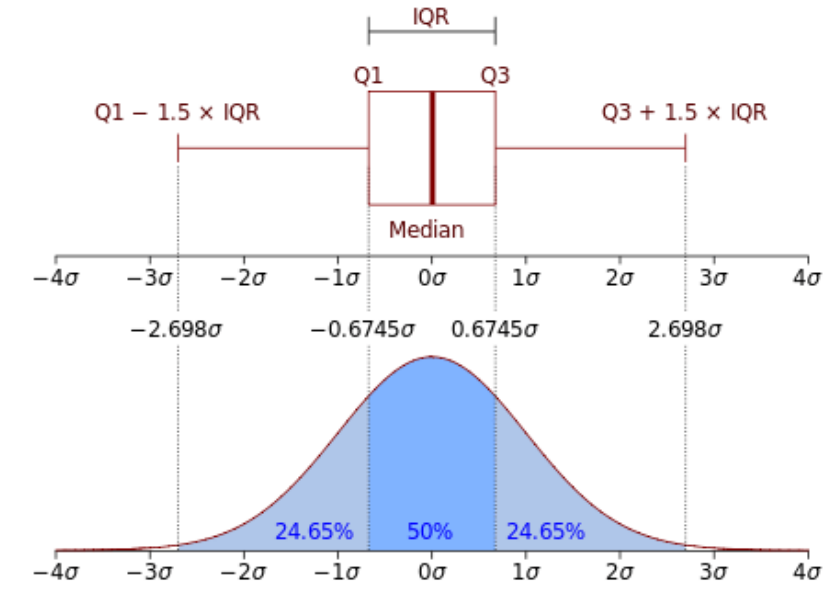
VISUALIZAÇÃO - BOXPLOT

Uma outra maneira de visualizar as mesmas estatísticas, pode-se utilizar o gráfico boxplot

O boxplot mostra a distribuição dos dados com base em 5 medidas

- Mediana (Q2)
- Primeiro (Q1) e terceiro (Q3) quartis
  - Intervalo interquartil (IQR = Q3-Q1)
- Mínimo: Q3 + 1.5*IQR
- Máximo: Q1 -1.5*IQR

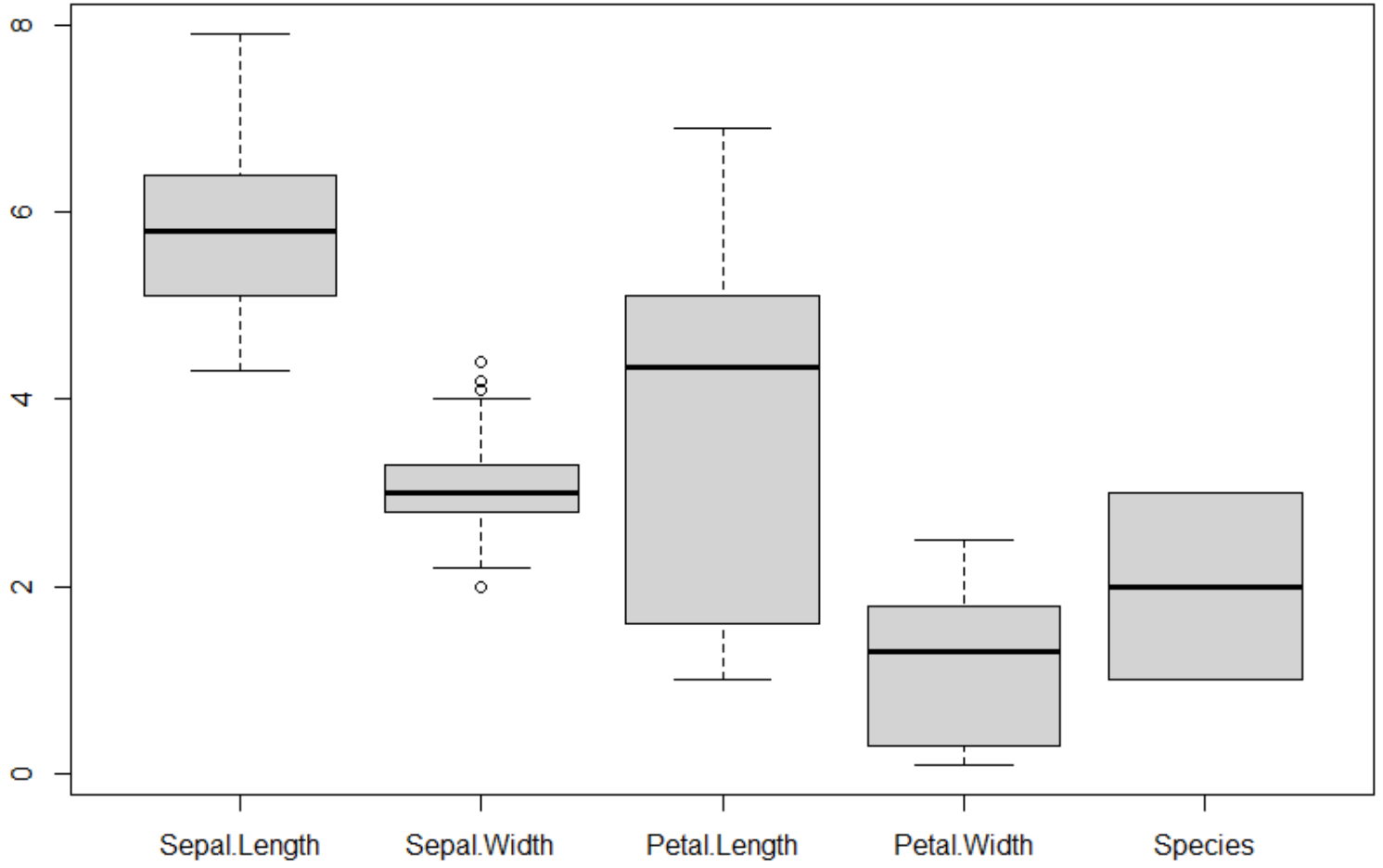




https://commons.wikimedia.org/w/index.php?curid=1452428 5

VISUALIZAÇÃO - BOXPLOT

```
> boxplot(iris)
```



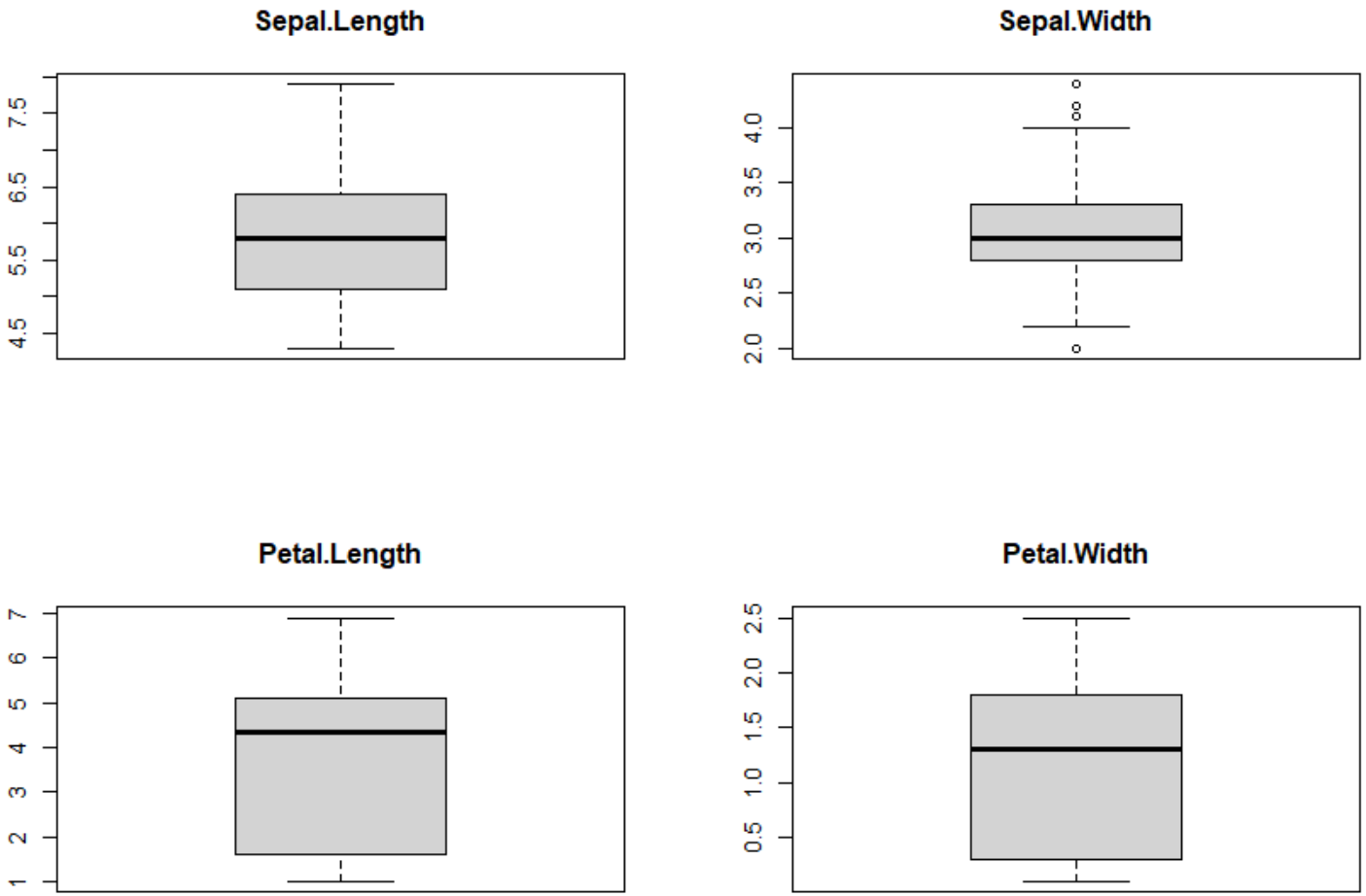
VISUALIZAÇÃO - BOXPLOT

No caso de intervalos de valores diferentes, deve-se construir gráficos separados

```
par(mfrow=c(2,2))
for (i in 1:4) {
  boxplot(iris[,i],
          main=names(iris)[i])
}
```

Define parâmetros gráficos

- mfrow=(nr,nc) define que os gráficos sejam desenhados em uma matriz nr por nc, linha por linha
- mfcol=(nr,nc) realiza a mesma configuração com a diferença que o preenchimento é por coluna

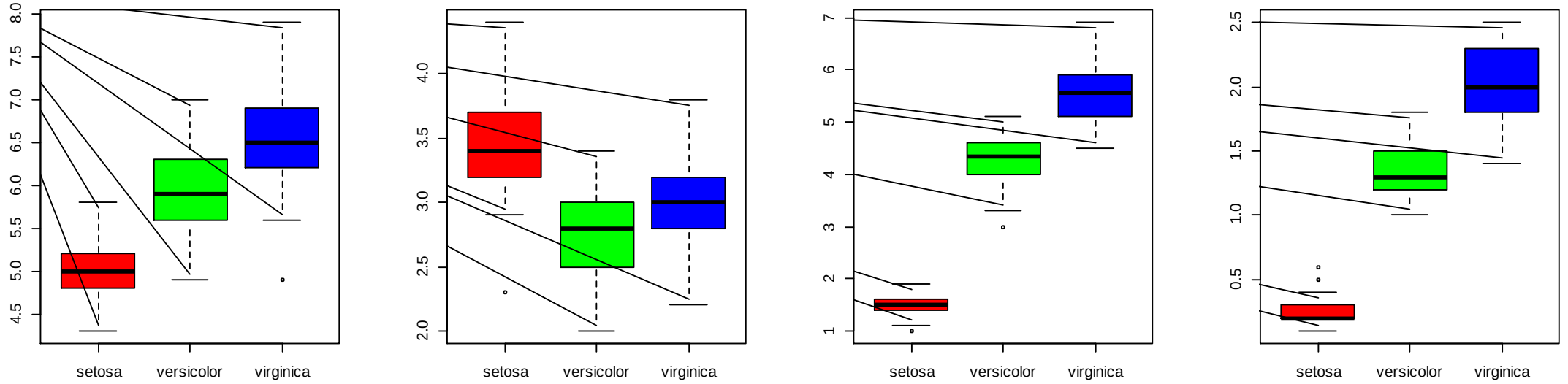


VISUALIZAÇÃO - BOXPLOT

No caso de problemas de classificação, os gráficos são construídos por classes (separação das distribuições)

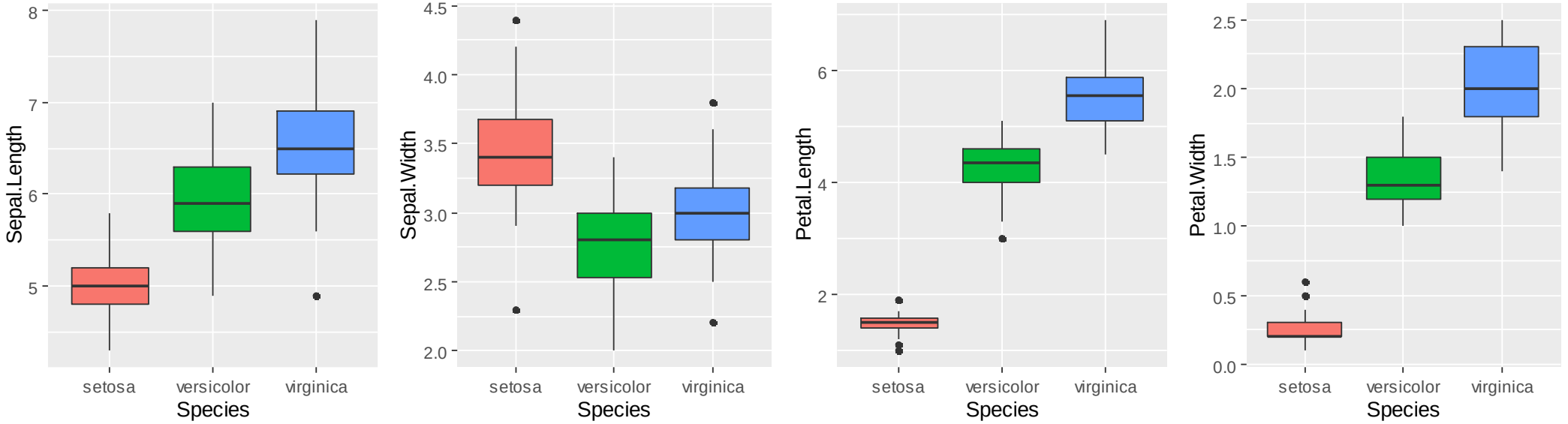
```
cores = c("red","green","blue")
par(mfrow=c(1,4))
for (i in 1:4) {
  boxplot(iris[,i] ~ iris$Species, col=cores, xlab="Species",
                                   ylab="", main=names(iris)[i])
}
```

A fórmula representa que os dados iris[,i] será dividido em grupos de acordo com a variável iris$Species



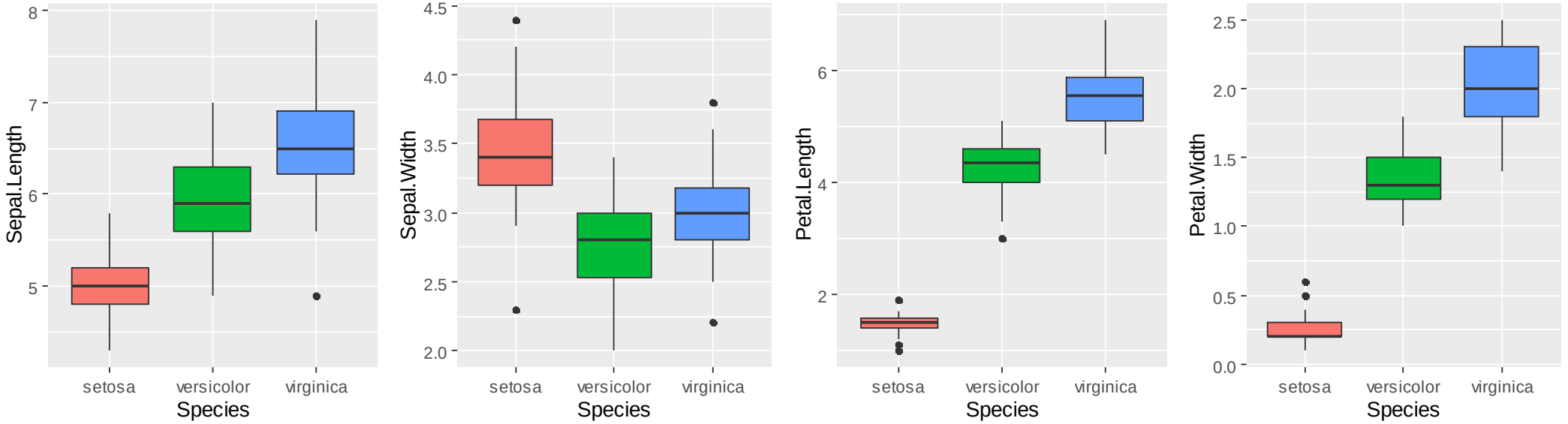
VISUALIZAÇÃO - BOXPLOT

```
library(gridExtra)
library(ggplot2)
p1 = ggplot(iris, aes(x=Species, y=Sepal.Length, fill=Species)) + geom_boxplot() + theme(legend.position="none")
p2 = ggplot(iris, aes(x=Species, y=Sepal.Width,  fill=Species)) + geom_boxplot() + theme(legend.position="none")
p3 = ggplot(iris, aes(x=Species, y=Petal.Length, fill=Species)) + geom_boxplot() + theme(legend.position="none")
p4 = ggplot(iris, aes(x=Species, y=Petal.Width,  fill=Species)) + geom_boxplot() + theme(legend.position="none")
grid.arrange(p1,p2,p3,p4,ncol=4)
```



VISUALIZAÇÃO - BOXPLOT

```
p = list()
for(i in 1:4){
  p[[i]] = ggplot(iris, aes_string(x="Species", y=names(iris)[i],fill="Species")) +
                geom_boxplot() + theme(legend.position="none")
}
do.call(grid.arrange,c(p,ncol=4))
```

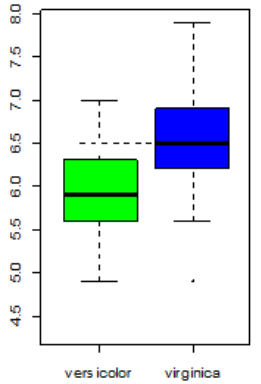


VISUALIZAÇÃO - BOXPLOT

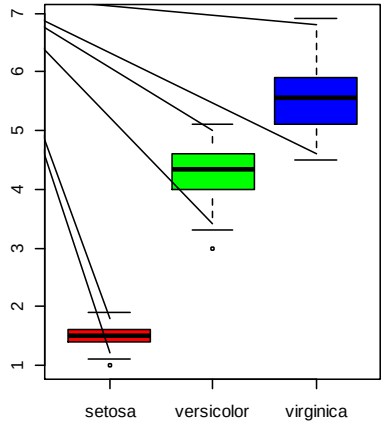
No caso de classificação, o interesse é que a distribuição das classes sejam diferentes para permitir a discriminação entre elas

Algumas pistas

- Quando a mediana de uma classe estiver fora da caixa (Q1 a Q3) da outra classe, então existe uma probabilidade de que as distribuições são diferentes



- Quando as caixas não se sobreporem, existe uma probabilidade de que as distribuições são diferentes com uma confiança de 95%



VISUALIZAÇÃO - HISTOGRAMA

O histograma representa as frequências dos valores divididos em intervalos

- Método simples para examinar a distribuição dos dados

```
par(mfrow=c(2,2))
for (i in 1:4) {
  minimo= min(iris[,i])
  maximo =max(iris[,i])
  hist(iris[iris$Species=="virginica",i], 10, xlim=c(minimo,maximo), col="blue",
         main="", xlab=names(iris)[i])
  hist(iris[iris$Species=="versicolor",i], 10, xlim=c(minimo,maximo), col="green",
         main="", xlab=names(iris)[i],add=T)
  hist(iris[iris$Species=="setosa",i], 10, xlim=c(minimo,maximo), col="red",
         main="", xlab=names(iris)[i],add=T)
}
legend("topright", legend = levels(iris$Species), col=c("red","green","blue"),
        pt.cex=2, pch=15)
```

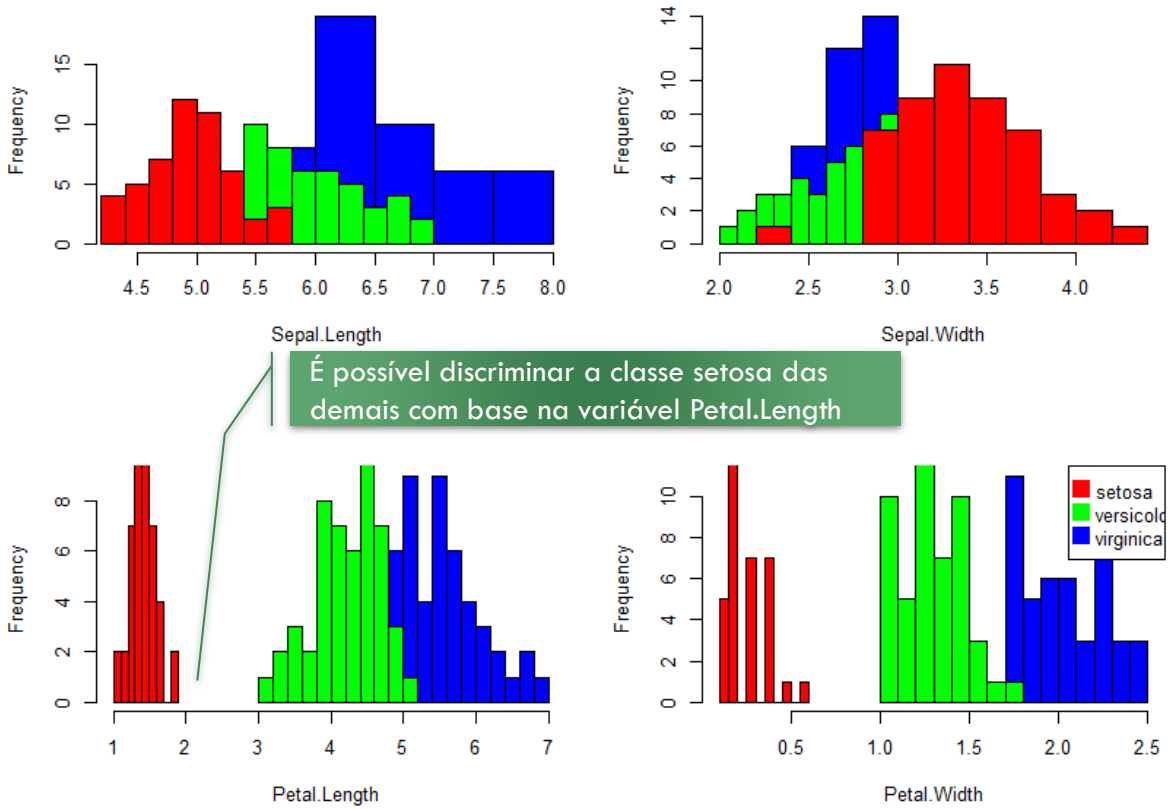
Número de intervalos (breaks=)

Intervalo do eixo x (padronizar todos os histogramas)

VISUALIZAÇÃO - HISTOGRAMA

O histograma por classe permite visualizar se ocorre sobreposição das classes (baixa discriminação)

- Note que esta é uma visualização por variável



Classes podem ser facilmente discriminadas em diferentes espaços de variáveis (combinação de duas ou mais variáveis)

- O aumento da dimensionalidade é uma das abordagens de alguns métodos para uma melhor discriminação

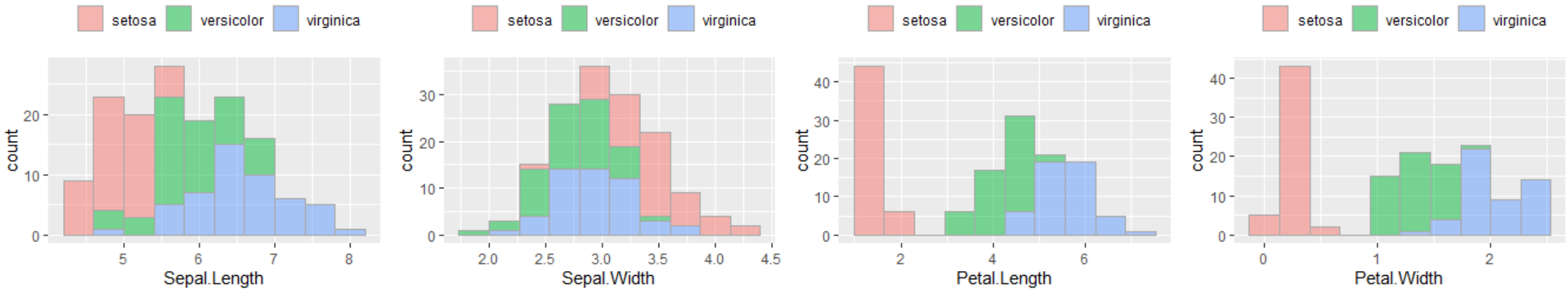
VISUALIZAÇÃO - HISTOGRAMA

Base de dados

Variável

Rótulo

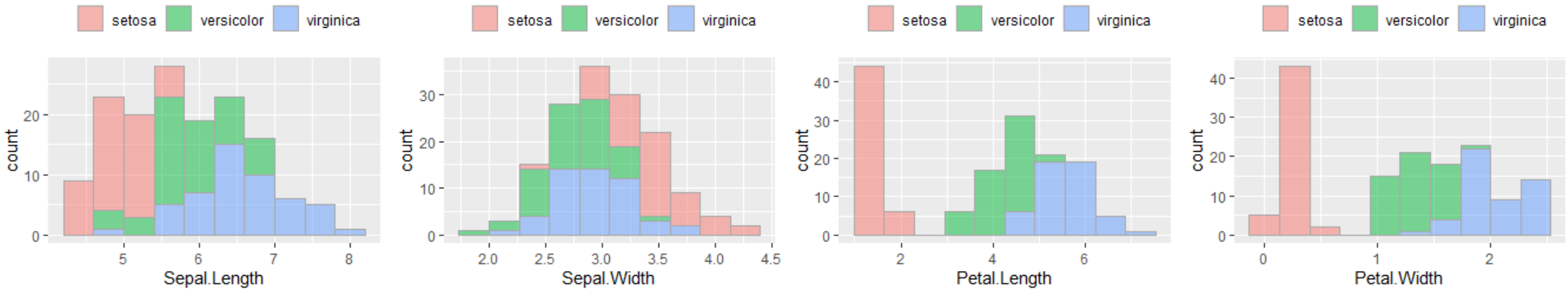
```
library(gridExtra)
library(ggplot2)
p1 = ggplot(iris,aes(x=Sepal.Length, fill=Species)) + geom_histogram(alpha=0.5,bins=10,color="darkgray") +
  theme(legend.position="top", legend.title = element_blank())
p2 = ggplot(iris,aes(x=Sepal.Width, fill=Species)) + geom_histogram(alpha=0.5,bins=10,color="darkgray") +
  theme(legend.position="top", legend.title = element_blank())
p3 = ggplot(iris,aes(x=Petal.Length, fill=Species)) + geom_histogram(alpha=0.5,bins=10,color="darkgray") +
  theme(legend.position="top", legend.title = element_blank())
p4 = ggplot(iris,aes(x=Petal.Width, fill=Species)) + geom_histogram(alpha=0.5,bins=10,color="darkgray") +
  theme(legend.position="top", legend.title = element_blank())
grid.arrange(p1,p2,p3,p4,ncol=4)
```



VISUALIZAÇÃO - HISTOGRAMA

```
library(gridExtra)
library(ggplot2)
p = list()
for(i in 1:4){
  p[[i]] = ggplot(iris, aes_string(x=names(iris)[i],fill="Species")) +
           geom_histogram(alpha=0.5,bins=10,color="darkgray") +
           theme(legend.position="top", legend.title = element_blank())
}
do.call(grid.arrange,c(p,ncol=4))
```

Seleciona a variável pelo nome

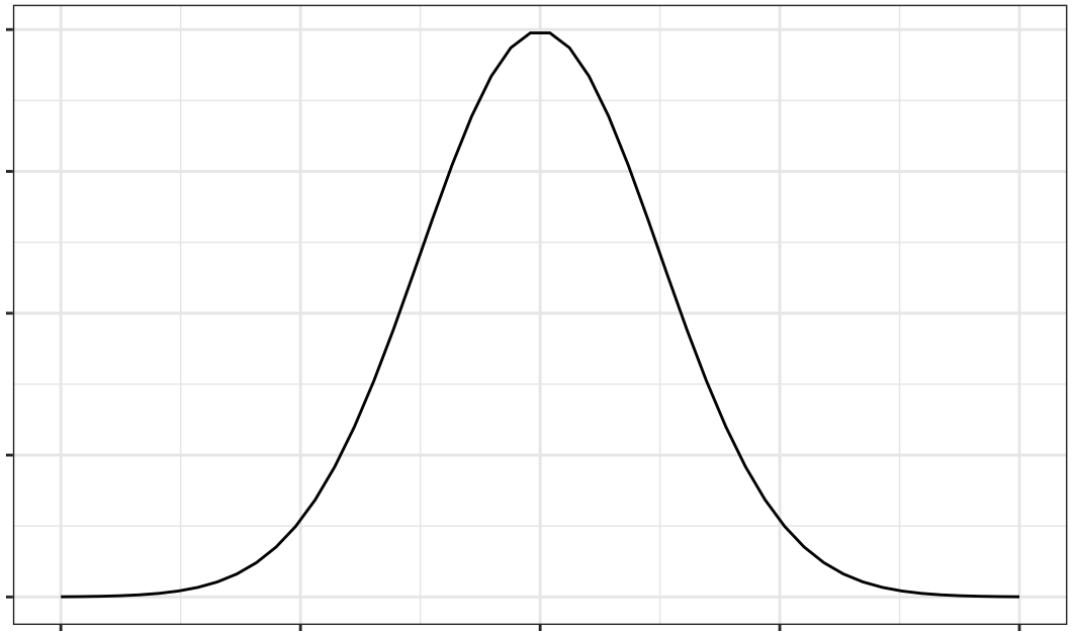


VISUALIZAÇÃO - DENSIDADE

Os gráficos de densidade, também conhecidos como curvas de densidade suavizadas, são esteticamente mais atraentes que os histogramas

- Facilita a comparação entre 2 distribuições

Como a maioria dos métodos de discriminação probabilísticos utilizam a distribuição normal para a modelagem das classes, busca-se encontrar uma distribuição normal dos dados

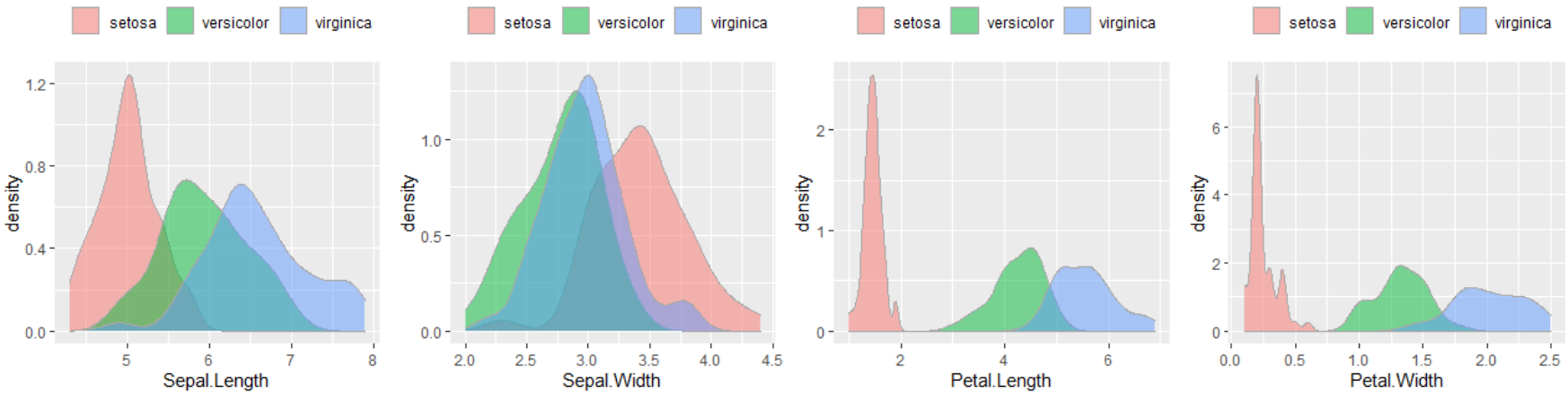


VISUALIZAÇÃO - DENSIDADE

A densidade produz um gráfico da distribuição de uma variável mais suave

```
library(gridExtra)
library(ggplot2)
p = list()
for(i in 1:4){
  p[[i]] = ggplot(iris, aes_string(x=names(iris)[i],fill="Species")) +
           geom_density(alpha=0.5,color="darkgray") +
           theme(legend.position="top", legend.title = element_blank())
}
do.call(grid.arrange,c(p,ncol=4))
```

Única diferença em relação ao histograma



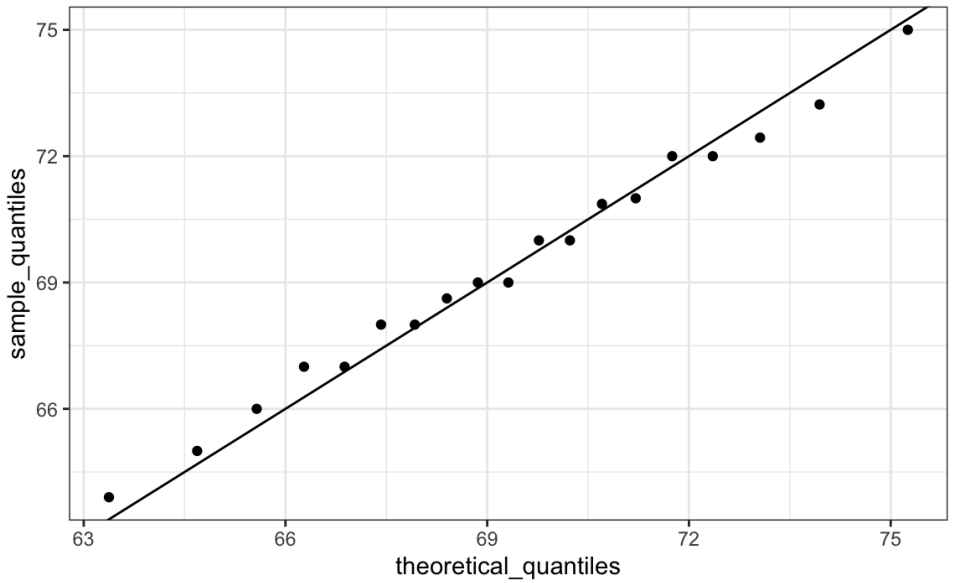
VISUALIZAÇÃO - NORMALIDADE

Um gráfico muito utilizado para verificar visualmente se uma distribuição é normal é o quartil-quartil (*quantile-quantile plot, qq-plot*)

O gráfico quartil-quartil permite verificar se as proporções observadas e previstas correspondem

Em geral, a ideia básica é a de calcular o valor teoricamente esperado para cada ponto de dados com base na distribuição em questão

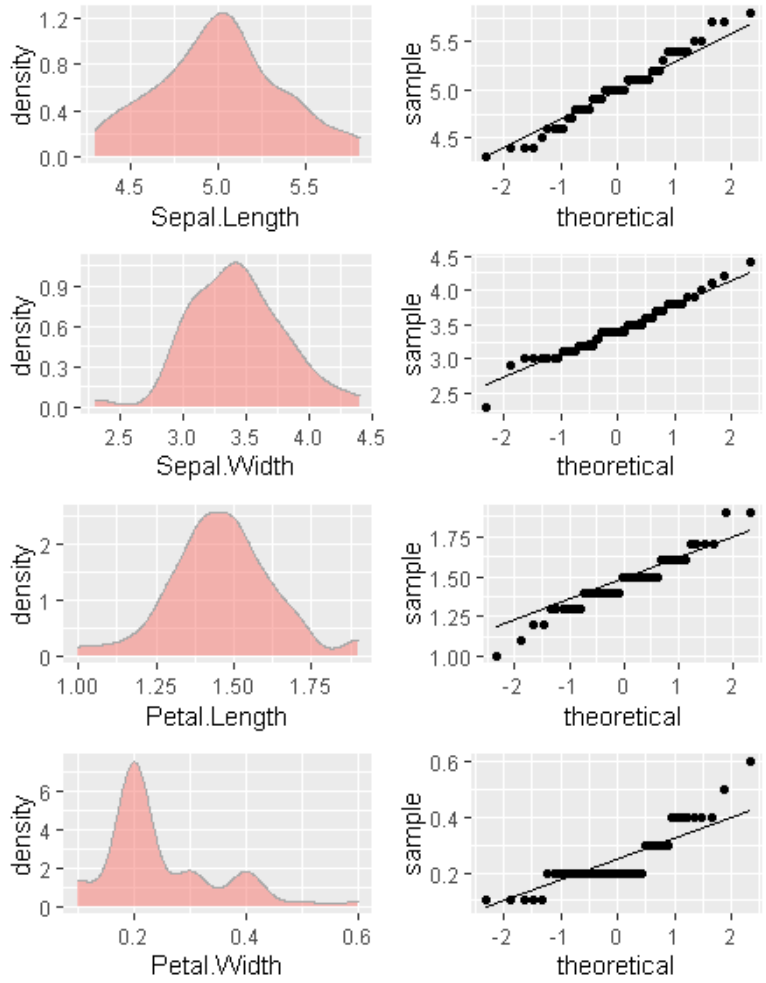
Se os dados de fato seguirem a distribuição assumida os pontos deste gráfico formarão aproximadamente uma linha reta



VISUALIZAÇÃO - NORMALIDADE

Vamos analisar a classe Setosa

```
library(gridExtra)
library(ggplot2)
p = list()
for(i in 1:4){
  p[[(i-1)*2+1]] = ggplot(iris[iris$Species=="setosa",],
           aes_string(x=names(iris)[i],fill="Species")) +
          geom_density(alpha=0.5,color="darkgray") +
          theme(legend.position="none")

  p[[i*2]]= ggplot(iris[iris$Species=="setosa",],
aes_string(sample=names(iris)[i]),color="Species")+
    stat_qq()+stat_qq_line()
}

do.call(grid.arrange,c(p,ncol=2))
```



VISUALIZAÇÃO - NORMALIDADE

Como a inspeção visual nem sempre é confiável, é possível utilizar um teste de significância para verificar se os dados desviam da normalidade

O teste Shapiro-Wilk (baseado na correlação) é recomendado para testes de normalidade

A hipótese (nula) deste teste é que os dados são distribuídos normalmente

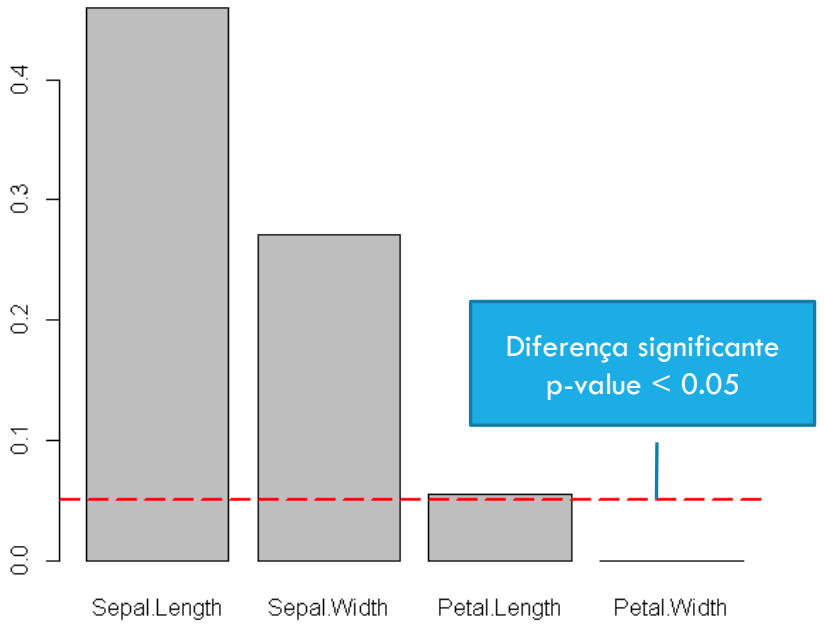
- O valor-p (p-value) deve ser maior que 0.05 para rejeitar a hipótese alternativa, isto é, a distribuição não é normal
- O valor-p mede a força do teste (quanto menor ele for, mais forte é a evidência contra a hipótese nula)

VISUALIZAÇÃO - NORMALIDADE

Vamos aplicar o teste nas variáveis pra a classe setosa

```
pvalue = c()
for(i in 1:4){
   pvalue[i] = shapiro.test(iris[iris$Species=="setosa",i])$p.value
}
barplot(pvalue,names.arg = names(iris)[1:4])
abline(h=0.05, col = "Red", lty = 5, lwd = 2)
```

Como a variável Petal.Width possui um $p - value < 0.05$, podemos concluir que ela não possui uma distribuição normal

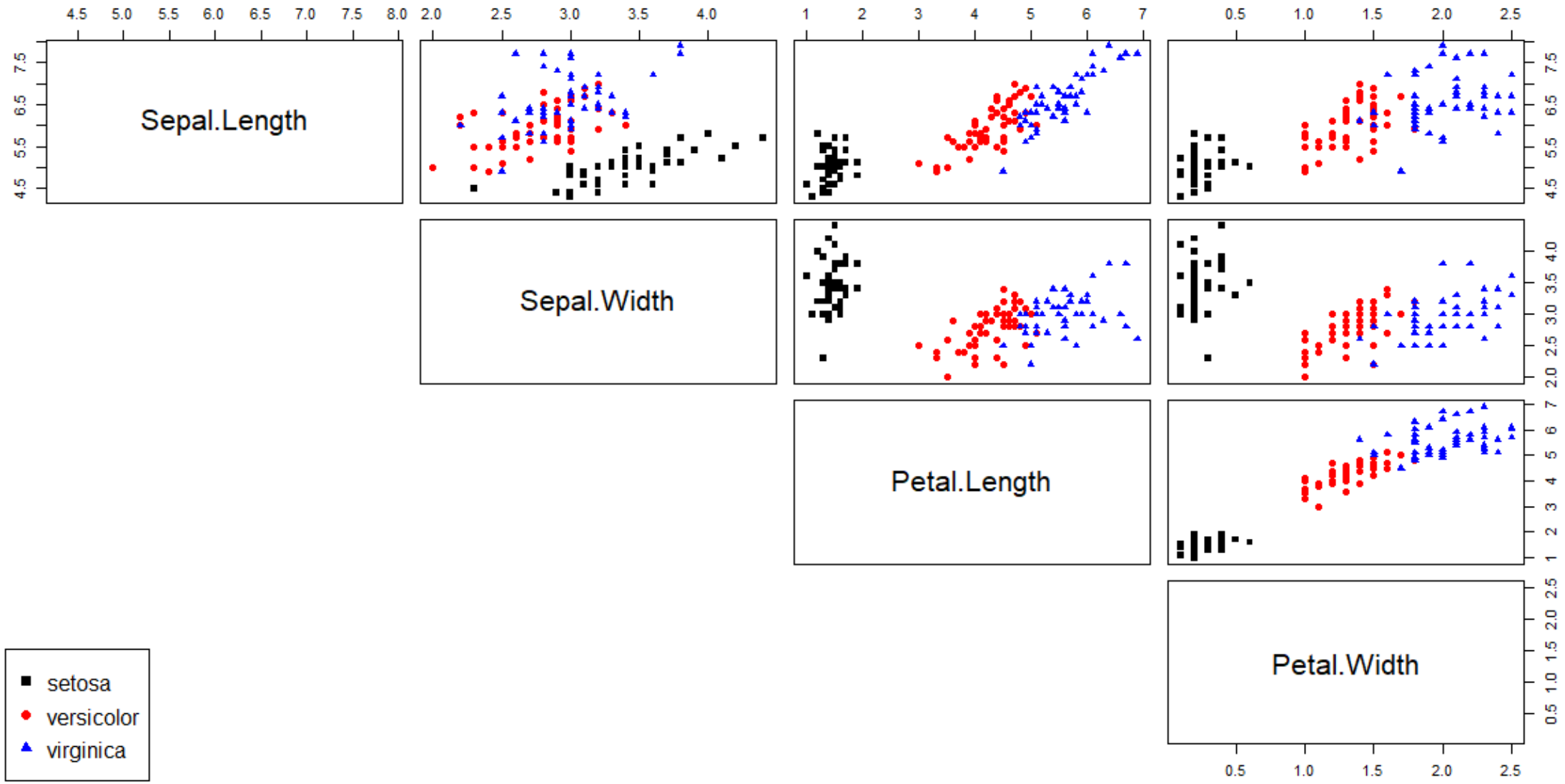


VISUALIZAÇÃO - DISPERSÃO

Um gráfico de dispersão auxilia na identificação de relacionamentos entre 2 variáveis contínuas

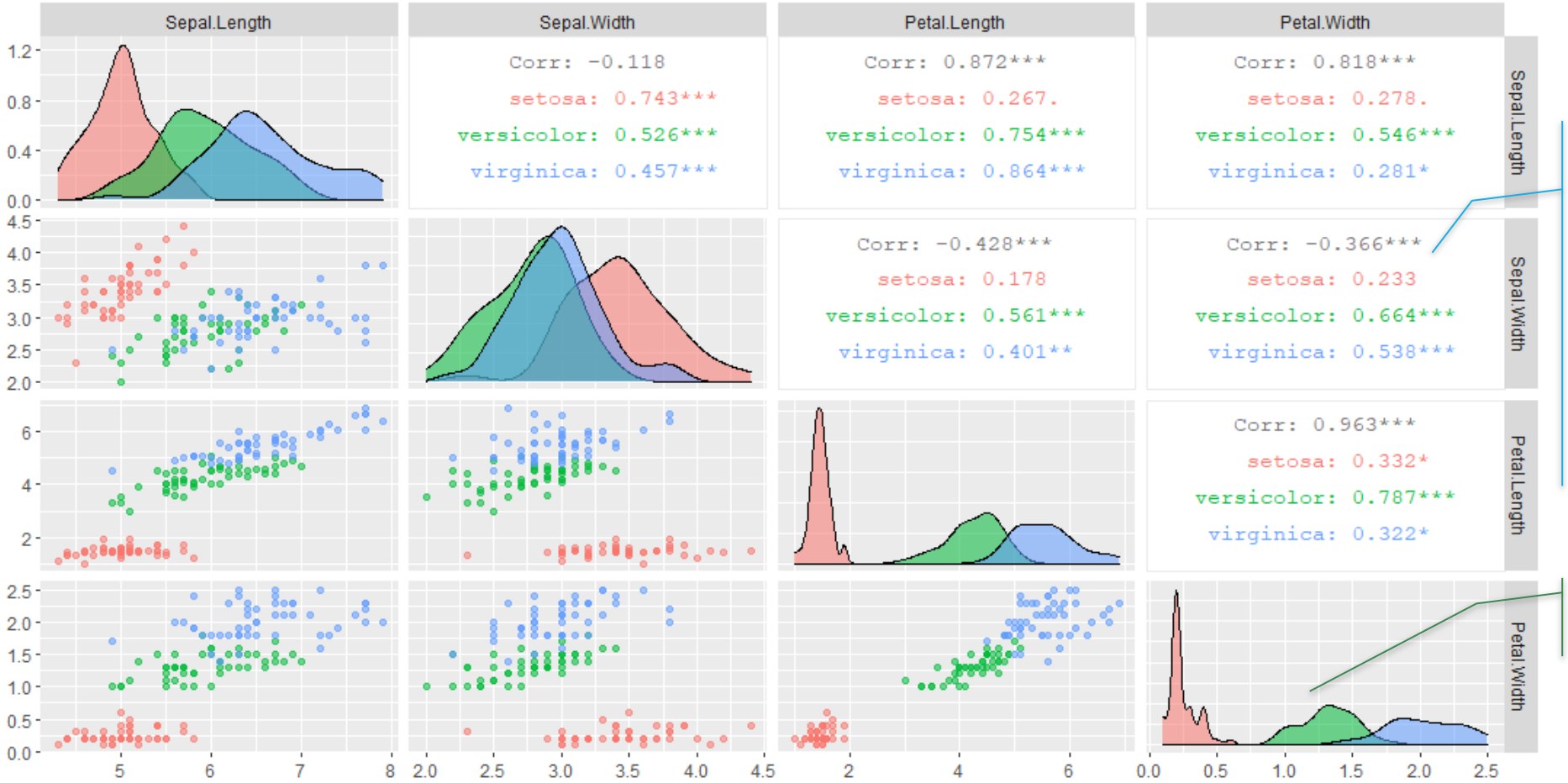
```
pairs(iris[1:4], pch=(15:17)[iris$Species],
      col=c("black", "red", "blue")[iris$Species],
      lower.panel=NULL)

par(xpd = TRUE)

legend("bottomleft", legend=levels(iris$Species),
      col=c("black", "red", "blue"), pch=15:17)
```

VISUALIZAÇÃO - DISPERSÃO



PRÁTICA: VISUALIZAÇÃO DOS DADOS - DISPERSÃO

```
library(GGally)
ggpairs(iris[,1:4], aes(colour = iris$Species, alpha = 0.4))
```



Correlação

*** p-value<0.001

** p-value<0.01

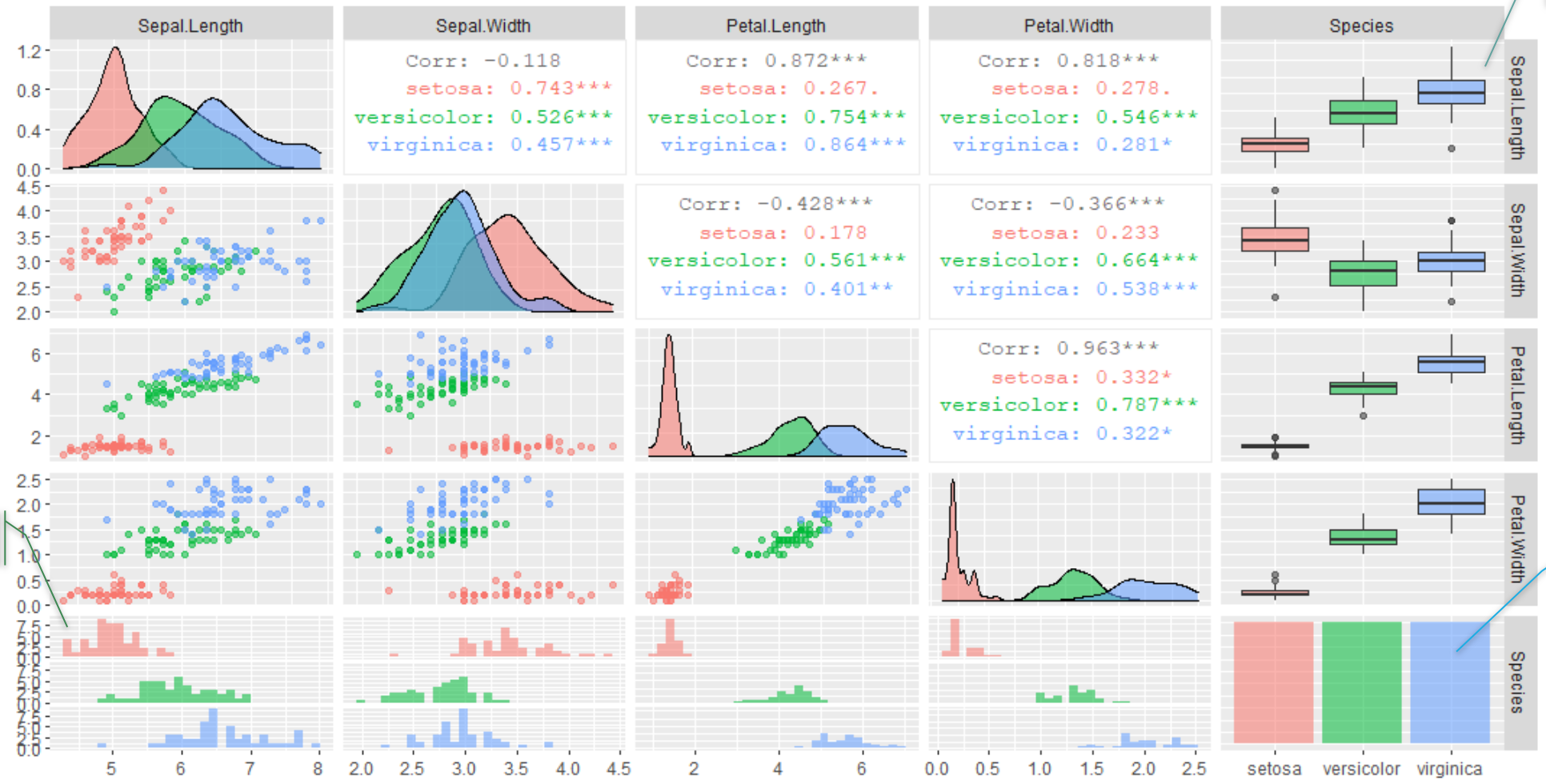
* p-value<0.05

. p-value<0.10

Densidade

VISUALIZAÇÃO - DISPERSÃO

ggpairs(iris, aes(colour = iris$Species, alpha = 0.4))



Boxplot

Histograma

Barras

CONCLUSÃO

Poder estimar informações e visualizar os dados de um determinado problema permite avaliar que estratégias e algoritmos podem ser utilizados no tratamento de tais dados

É importante entender que todas as ferramentas e análises exploram de uma a duas dimensões dos dados

Podem existir outra relações além das duas dimensões