

# Estística Descritiva

*Max Pereira*

*24/04/2020*

O objetivo é sintetizar os dados de maneira direta, preocupando-se menos com variações e intervalos de confiança dos dados.

## Análise Exploratória

### Conjunto de Dados (Dataset Iris)

```
dados <- iris  
head(dados)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1          5.1          3.5          1.4          0.2   setosa  
## 2          4.9          3.0          1.4          0.2   setosa  
## 3          4.7          3.2          1.3          0.2   setosa  
## 4          4.6          3.1          1.5          0.2   setosa  
## 5          5.0          3.6          1.4          0.2   setosa  
## 6          5.4          3.9          1.7          0.4   setosa
```

### Funções para efetuar estatística descritiva

- média: `mean()`
- desvio padrão: `sd()`
- variância: `var()`
- valor mínimo: `min()`
- valor máximo: `max()`
- mediana: `median()`
- amplitude de valores (min e máximo): `range()`
- quartis: `quantile()`
- função genérica: `summary()`

### Medidas de tendência central

```
mean(dados$Sepal.Length)
```

```
## [1] 5.843333
```

```
median(dados$Sepal.Length)
```

```
## [1] 5.8
```

## Medidas de variabilidade

```
min(dados$Sepal.Length)
```

```
## [1] 4.3
```

```
max(dados$Sepal.Length)
```

```
## [1] 7.9
```

```
range(dados$Sepal.Length)
```

```
## [1] 4.3 7.9
```

```
quantile(dados$Sepal.Length)
```

```
##   0%  25%  50%  75% 100%  
##  4.3  5.1  5.8  6.4  7.9
```

```
var(dados$Sepal.Length)
```

```
## [1] 0.6856935
```

```
sd(dados$Sepal.Length)
```

```
## [1] 0.8280661
```

## Resumo estatístico

```
summary(dados)
```

```
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width  
##   Min.    :4.300   Min.    :2.000   Min.    :1.000   Min.    :0.100  
##   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##   Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##   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  
##           Species  
##   setosa    :50  
##   versicolor:50  
##   virginica  :50  
##  
##  
##
```

## Gráficos de distribuições

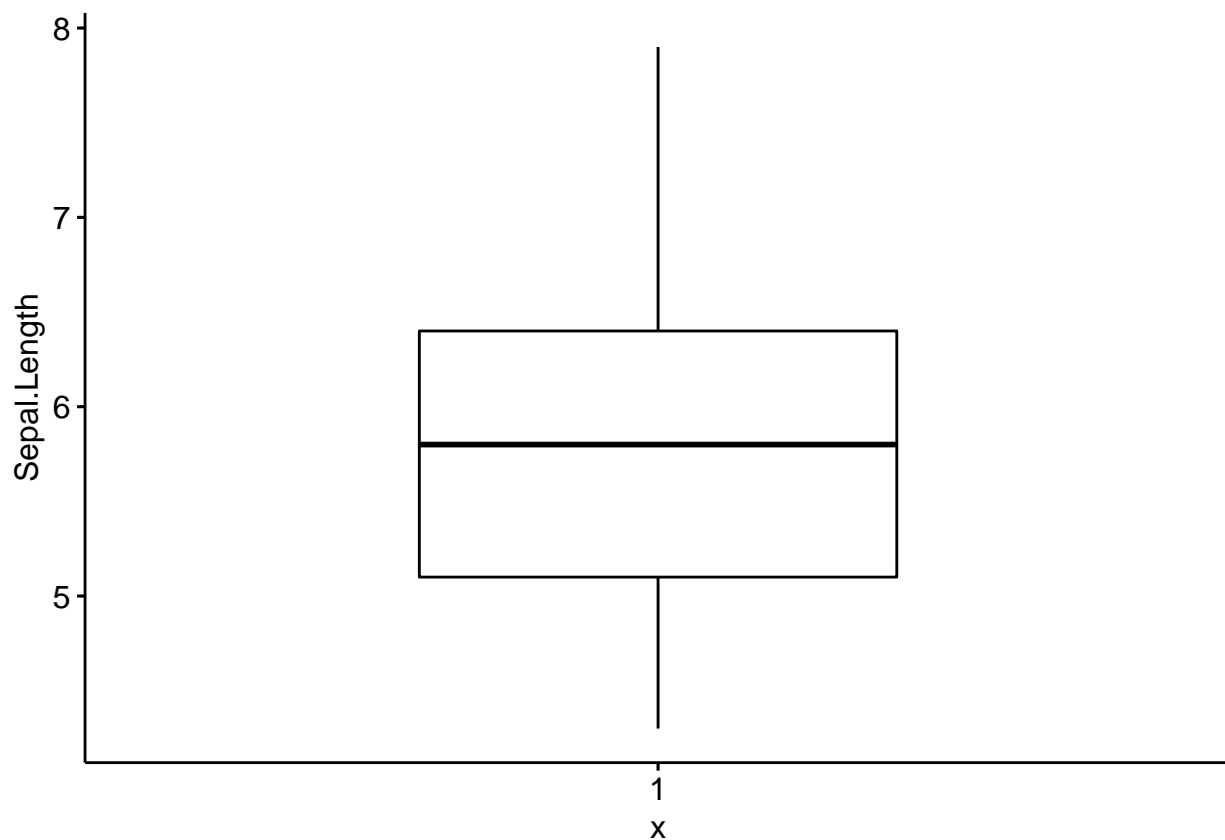
```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

Box plot

```
ggboxplot(dados, y="Sepal.Length", width = 0.5)
```



Analisando o boxplot (outliers)

```
v = c(10.2, 14.1, 14.4, 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4)
v
```

```
## [1] 10.2 14.1 14.4 14.4 14.4 14.5 14.5 14.6 14.7 14.7 14.7 14.9 15.1 15.9
## [15] 16.4
```

```
summary(v)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10.2   14.4   14.6   14.5   14.8   16.4
```

```
IQR(v) # Interquartile range
```

```
## [1] 0.4
```

Se um valor estiver abaixo de  $Q1 - 1.5 \times IQR$  ou acima de  $Q3 + 1.5 \times IQR$  é considerado um outlier

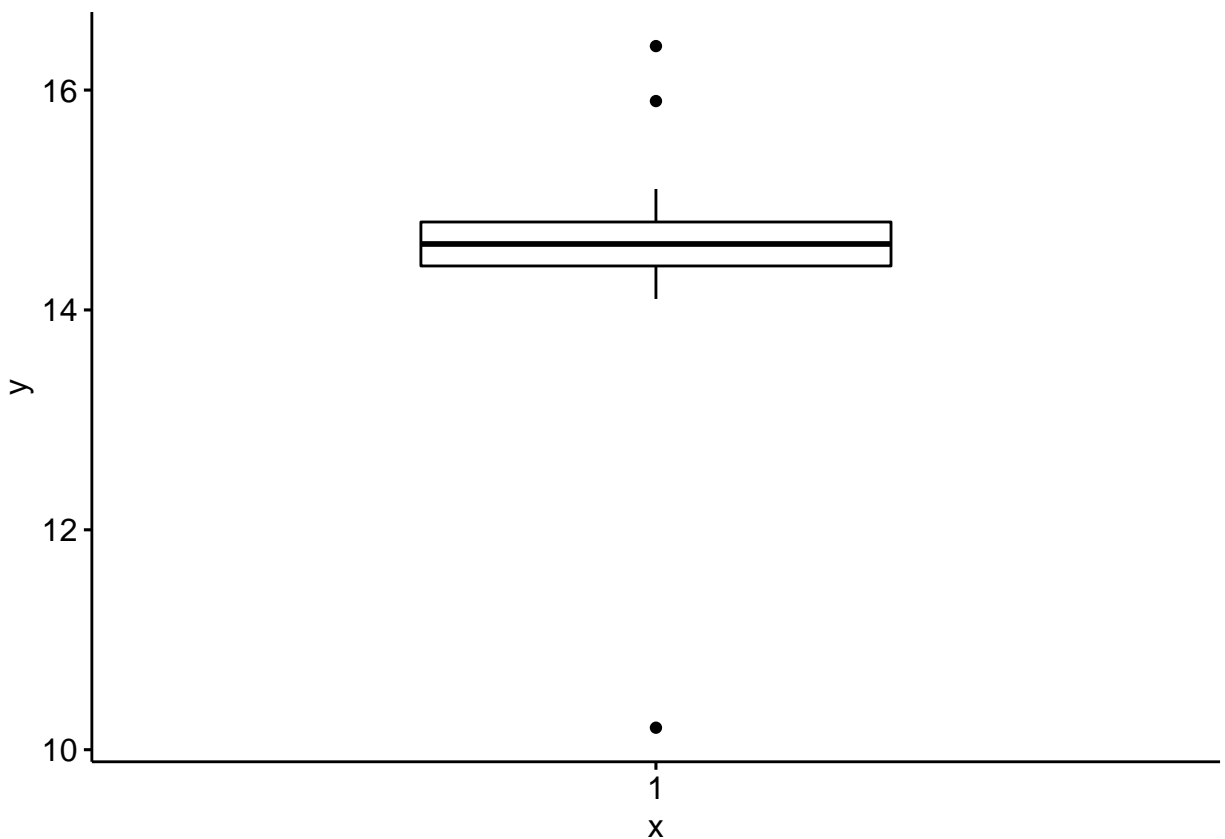
```
q1 = summary(v)[2]
q3 = summary(v)[5]
limite_inferior = q1 - 1.5 * IQR(v)
limite_inferior
```

```
## 1st Qu.
##      13.8
```

```
limite_superior = q3 + 1.5 * IQR(v)
limite_superior
```

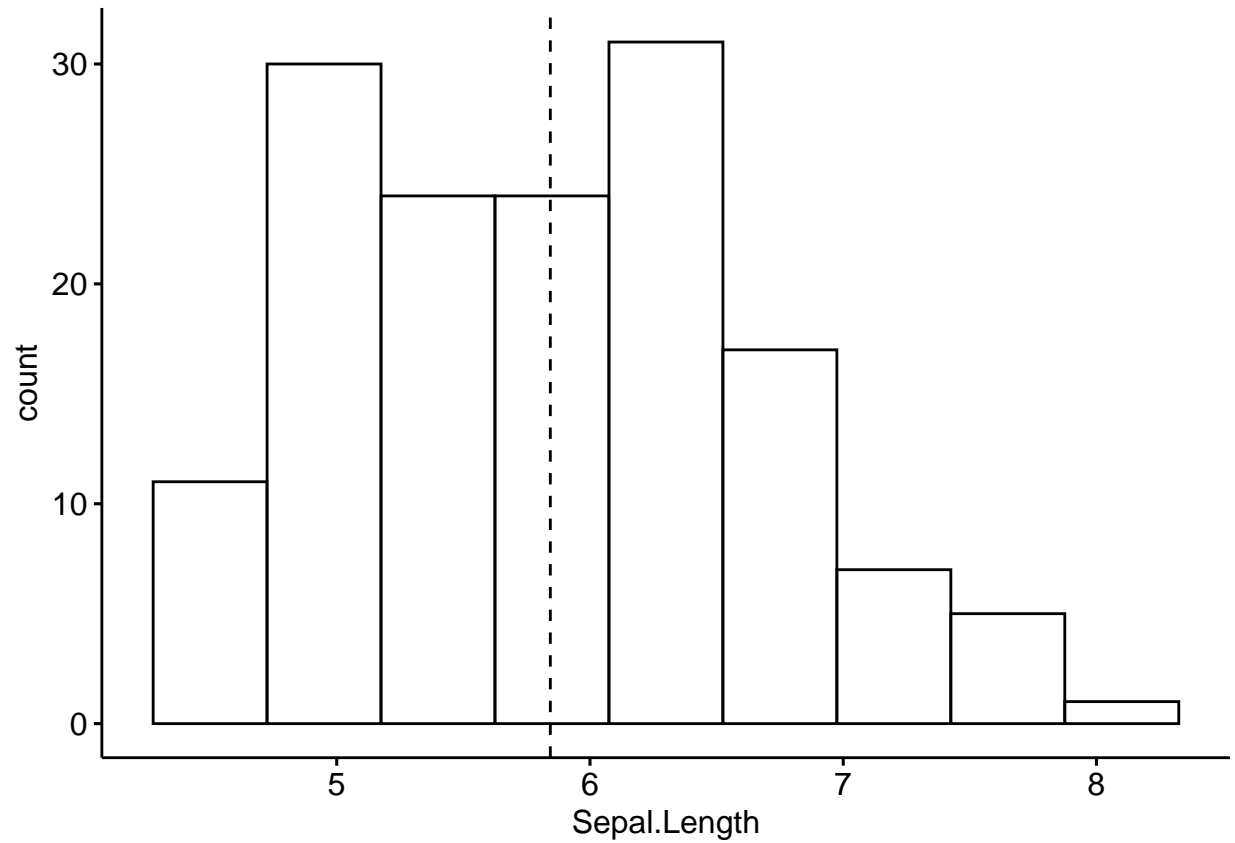
```
## 3rd Qu.
##      15.4
```

```
ggboxplot(v, width = 0.5)
```



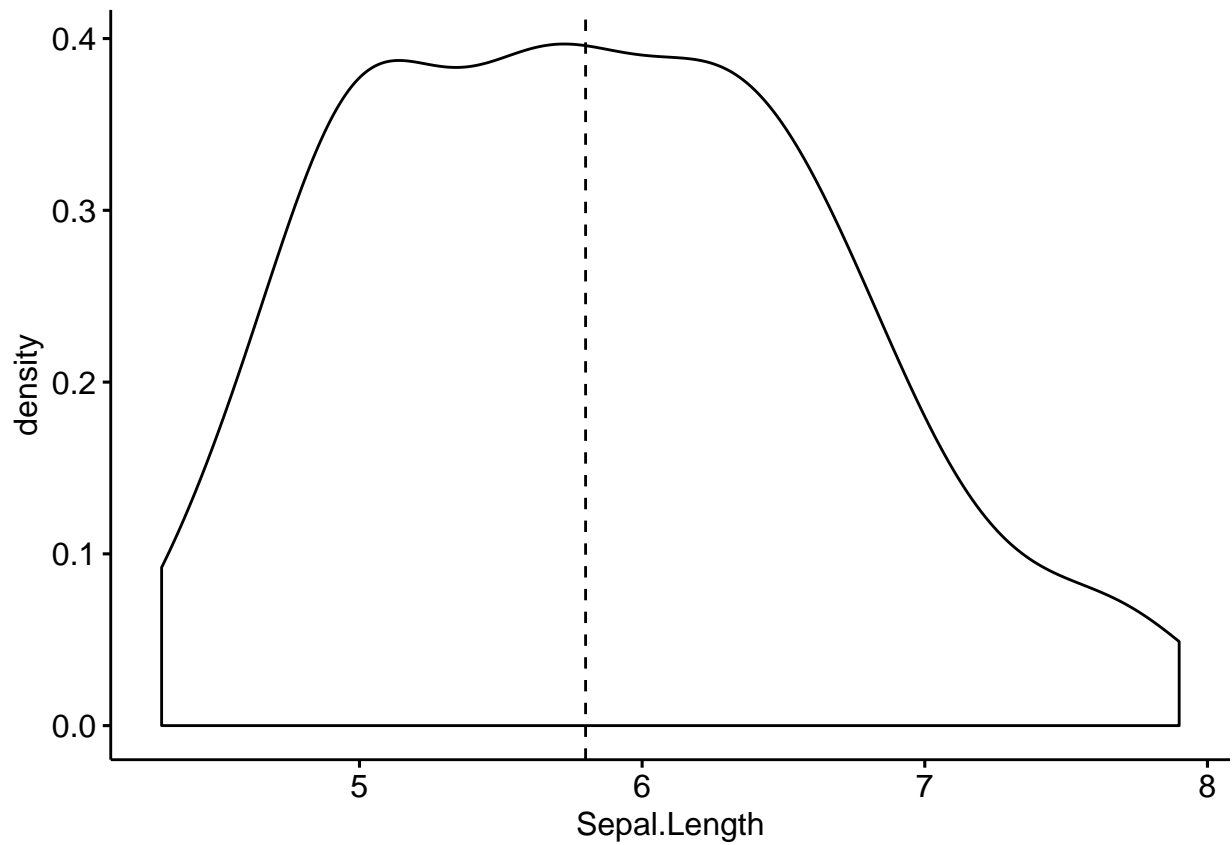
## Histograma

```
gghistogram(dados, x = "Sepal.Length", bins = 9, add = "mean")
```



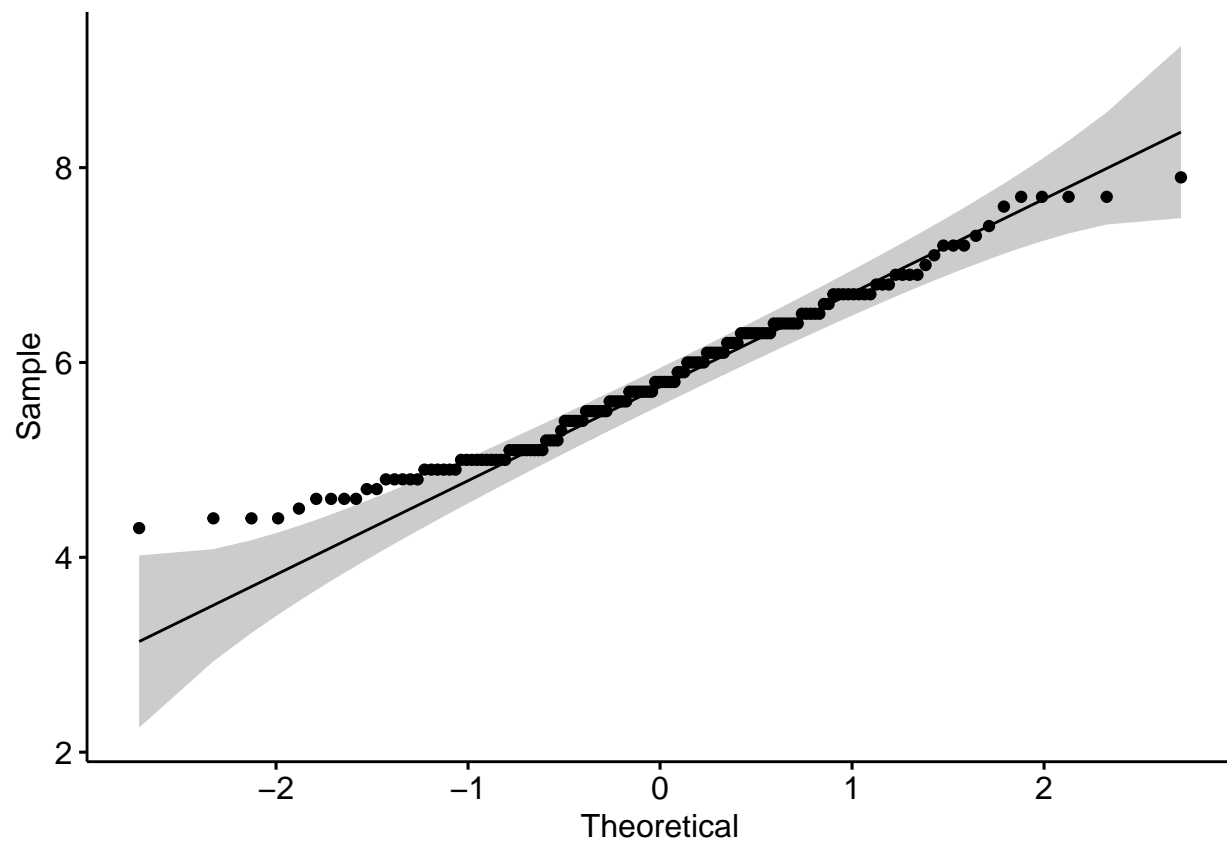
## Gráfico de densidade

```
ggdensity(dados, x = "Sepal.Length", add = "median")
```



Análise de distribuição normal

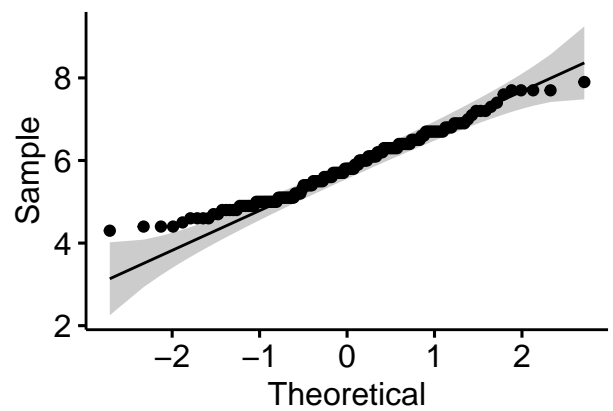
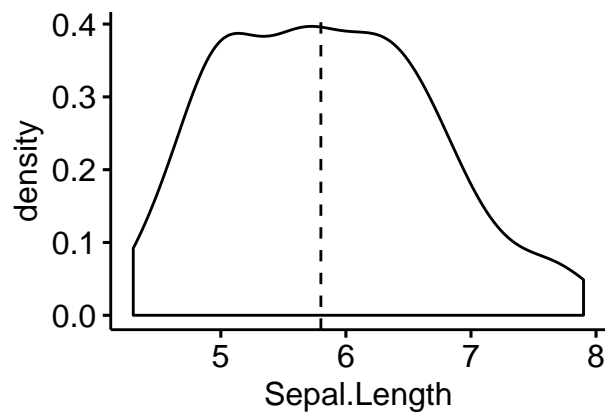
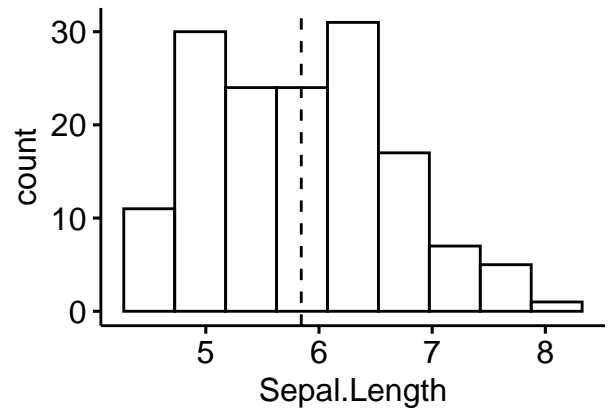
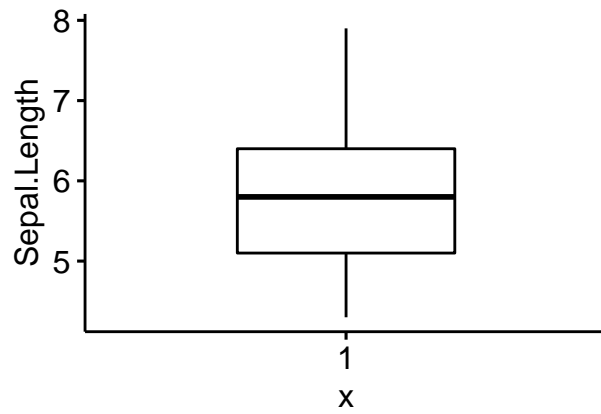
```
ggqqplot(dados, x="Sepal.Length")
```



Combinando os gráficos na mesma área de plotagem

```
bp <- ggboxplot(dados, y="Sepal.Length", width = 0.5)
ht <- gghistogram(dados, x = "Sepal.Length", bins = 9, add = "mean")
ds <- ggdensity(dados, x = "Sepal.Length", add = "median")
dn <- ggqqplot(dados, x="Sepal.Length")

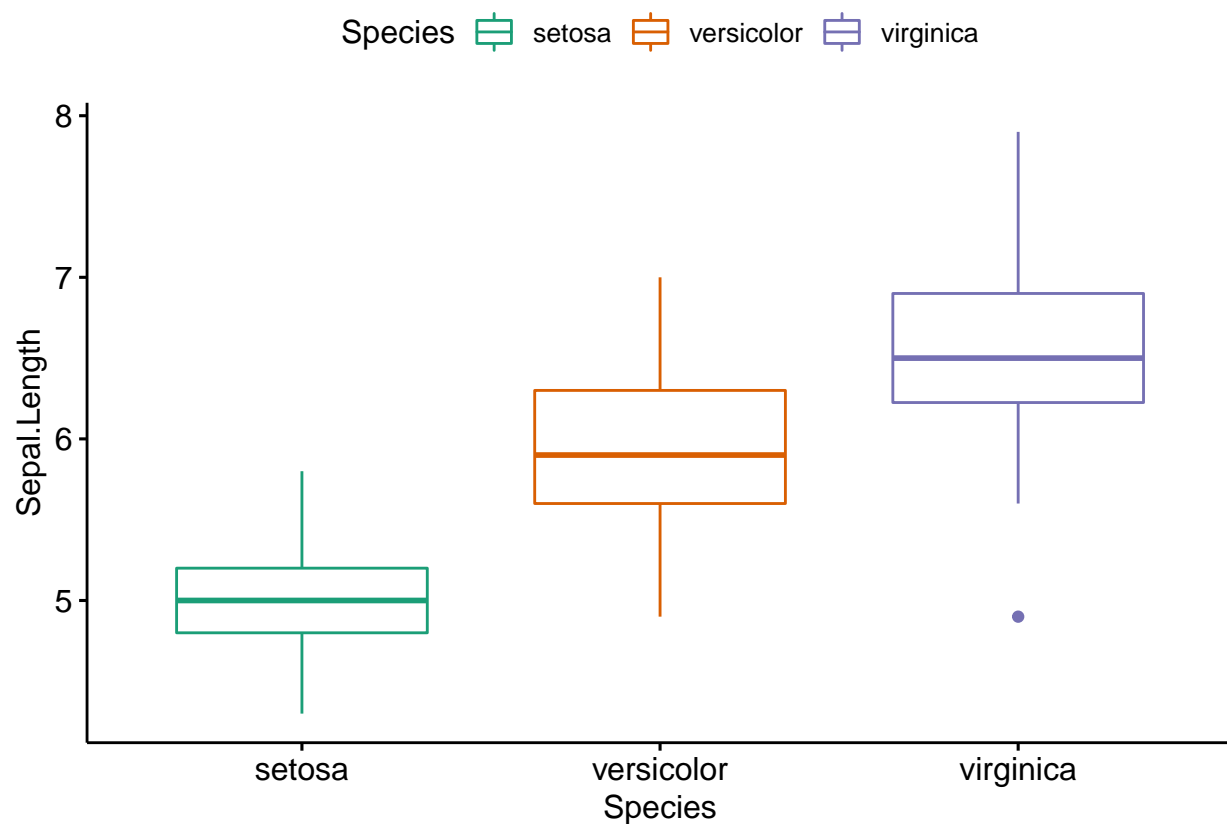
ggarrange(bp, ht, ds, dn)
```



Gráficos para dados agrupados

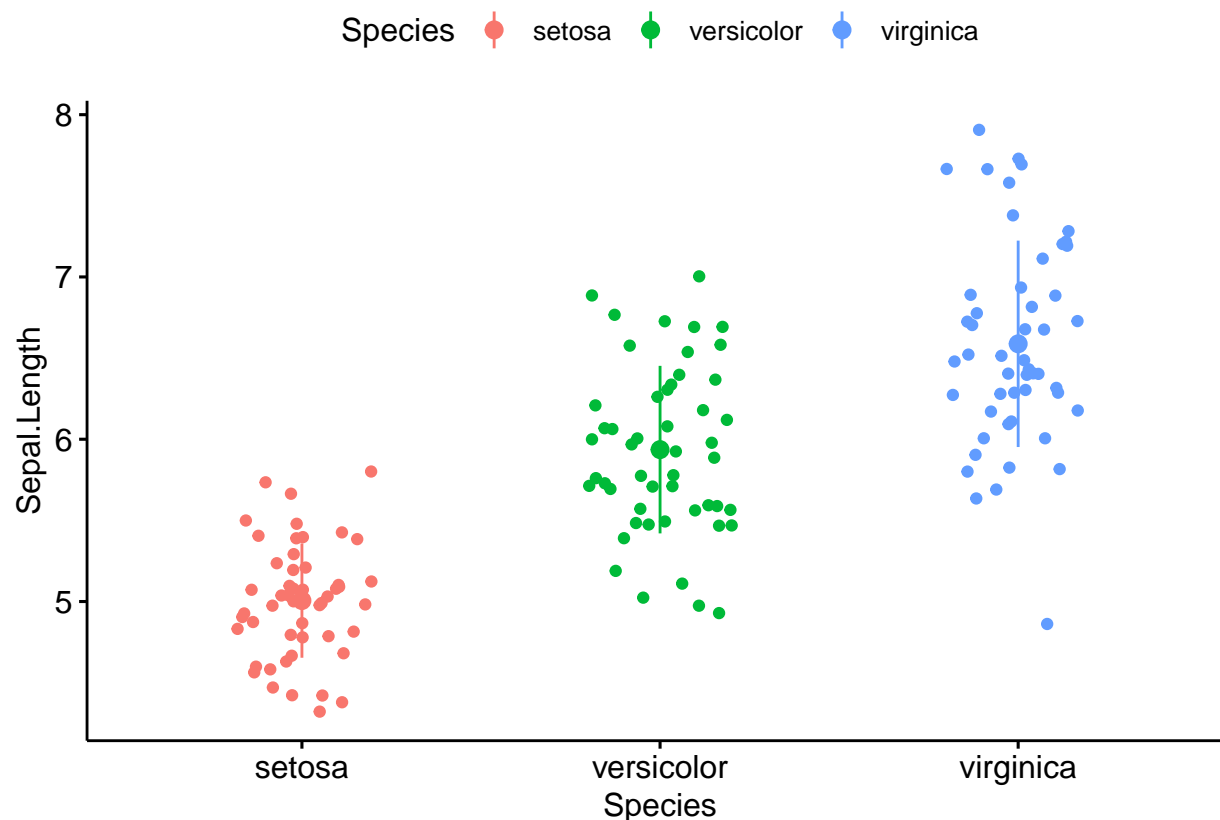
```
ggboxplot(dados, x = "Species", y = "Sepal.Length",
  color = "Species",
  palette = "Dark2")
```





Os valores permitidos para o parâmetro `palette` são: “grey” para paleta de cores cinzas; paletas de cores “brewer”: “RdBu”, “Blues”, “Dark2”; Para ver todas: `RColorBrewer::display.brewer.all()`; paletas customizadas: `c(“blue”, “red”)`; paletas de jornais científicos do pacote “ggsci”: “npg”, “aaas”, “lancet”, “jco”, “ucscgb”, “uchicago”, “simpsons” and “rickandmarty”.

```
ggstripchart(dados, x = "Species", y = "Sepal.Length",  
             color = "Species",  
             palette = "ggsci",  
             add = "mean_sd")
```



Conjunto de Dados (Dataset credito.csv)

```
df <- read.csv("credito.csv")
str(df)
```

```
## 'data.frame':    114 obs. of  12 variables:
## $ Id           : Factor w/ 87 levels "004NZMX60E","017STAOLDV",...: 1 1 2 3 4 5 6 6 6 7 ...
## $ estado       : Factor w/ 34 levels "AK","AL","AR",...: 4 4 23 23 7 22 23 23 26 ...
## $ sexo         : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 2 1 1 1 2 ...
## $ idade        : int   36 36 34 48 32 44 60 60 60 48 ...
## $ raca         : Factor w/ 7 levels "American Indian or Alaska Native",...: 5 5 7 5 7 6 2 2 2 1 ...
## $ estado_civil : Factor w/ 3 levels "Divorced","Married",...: 2 2 2 2 3 3 3 3 2 ...
## $ ocupacao     : Factor w/ 5 levels "Account","Business",...: 5 5 3 1 2 1 4 4 4 1 ...
## $ score_credito : int   710 720 720 670 720 540 840 824 824 490 ...
## $ rendimento   : num   9371 9371 9010 6538 8679 ...
## $ debitos      : num   2000 3014 1000 2099 1000 ...
## $ tipo_financ   : Factor w/ 4 levels "Auto","Credit",...: 4 1 2 3 3 4 4 1 2 4 ...
## $ decisao_financ : Factor w/ 3 levels "Approved","Denied",...: 1 1 1 1 1 2 1 1 1 2 ...
```

## Tabelas de frequência

Usada para descrever variáveis categóricas. Contém as contagens de cada combinação dos níveis dos fatores.

```
tipo.financiamento <- df$tipo_financ
ocupacao <- df$ocupacao
parecer <- df$decisao_financ
```

## Distribuição de frequência:

```
table(tipo.financiamento)
```

```
## tipo.financiamento
##      Auto      Credit      Home Personal
##        40        17        23        34
```

```
table(ocupacao)
```

```
## ocupacao
## Account Business      IT Manager      NYPD
##        18        15        27        26        28
```

```
table(parecer)
```

```
## parecer
## Approved      Denied Withdrawn
##         70         32         12
```

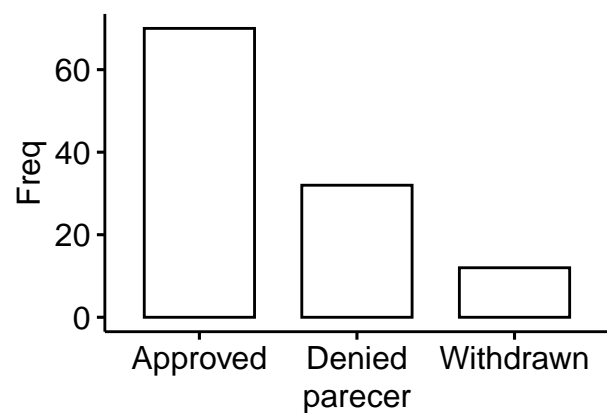
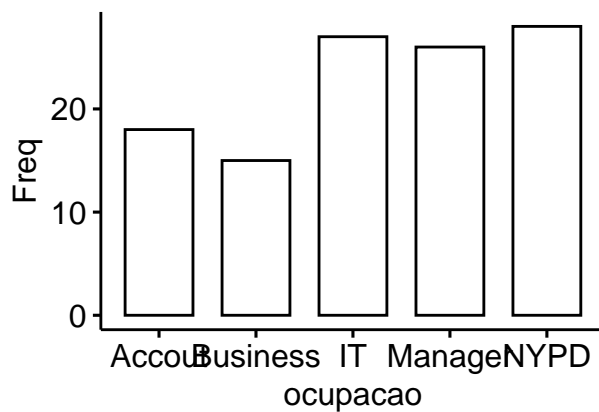
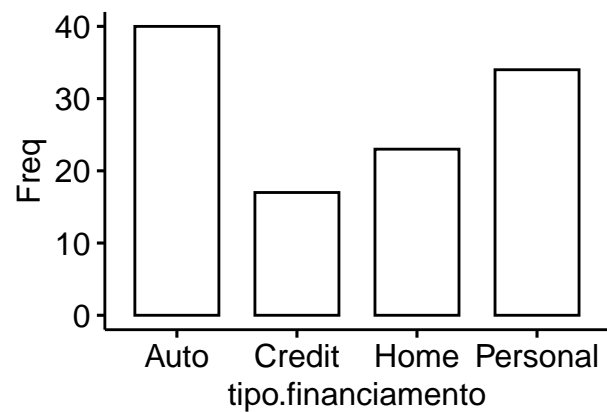
Convertendo as tabelas para dataframes

```
df_fin <- as.data.frame(table(tipo.financiamento))
df_oc <- as.data.frame(table(ocupacao))
df_par <- as.data.frame(table(parecer))
```

## Gráfico de barras

```
fin <- ggbarplot(df_fin, x="tipo.financiamento", y="Freq")
oc <- ggbarplot(df_oc, x="ocupacao", y="Freq")
par <- ggbarplot(df_par, x="parecer", y="Freq")

ggarrange(fin, oc, par)
```



### Gráfico de pizza

```
ggpie(df_par, x="Freq", label = "parecer",
      color = "white",
      fill = "parecer",
      palette = c("blue", "red", "gray"))
```

parecer ■ Approved ■ Denied ■ Withdrawn

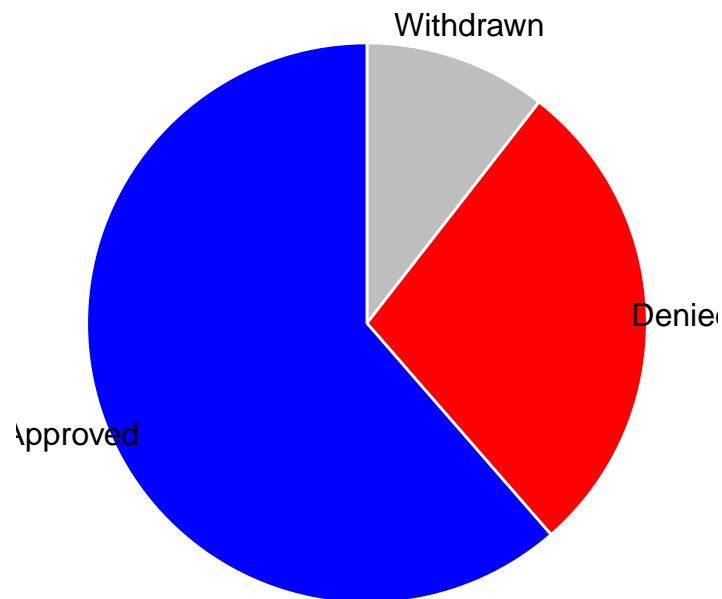


Tabela com duas variáveis categóricas

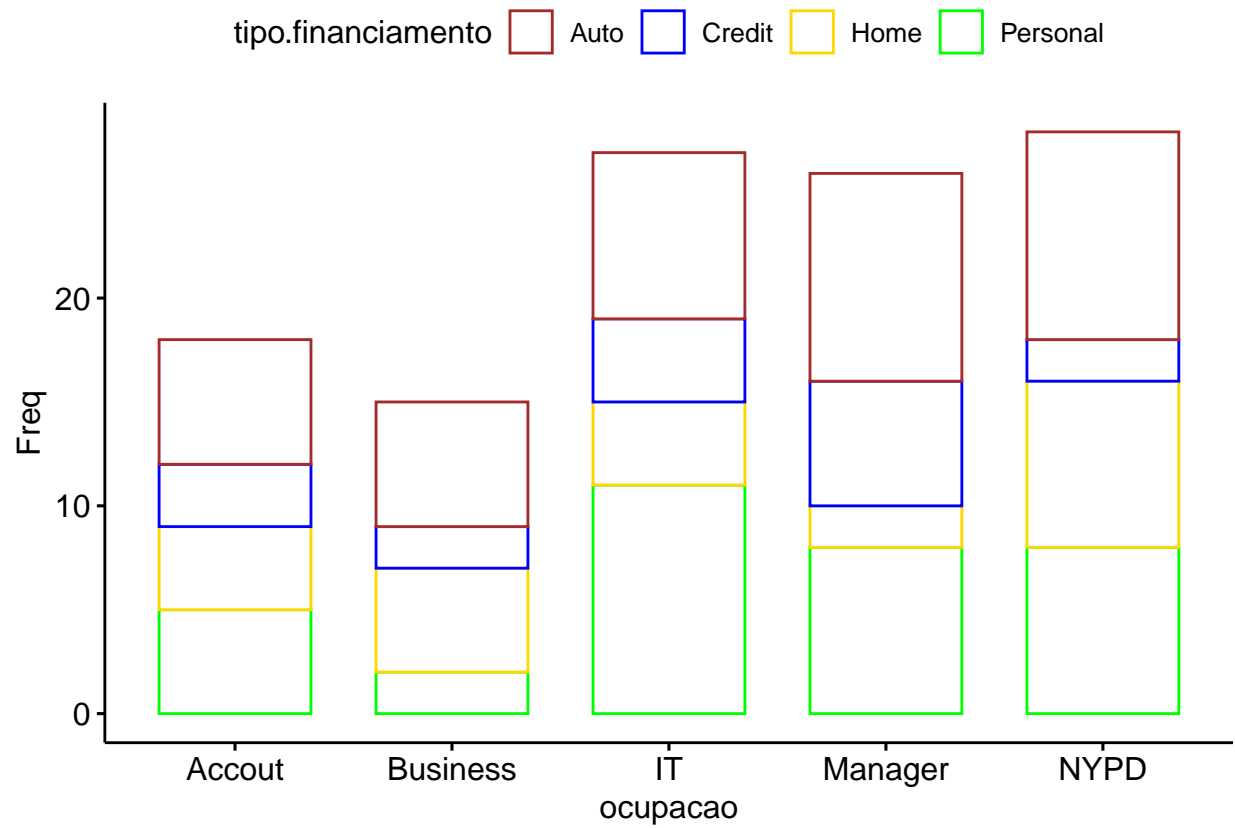
```
tbl2 <- table(tipo.financiamento, ocupacao)
tbl2
```

```
##              ocupacao
## tipo.financiamento Account Business IT Manager NYPD
##           Auto          6          6  8          10  10
##           Credit        3          2  4           6   2
##           Home          4          5  4           2   8
##           Personal       5          2 11           8   8
```

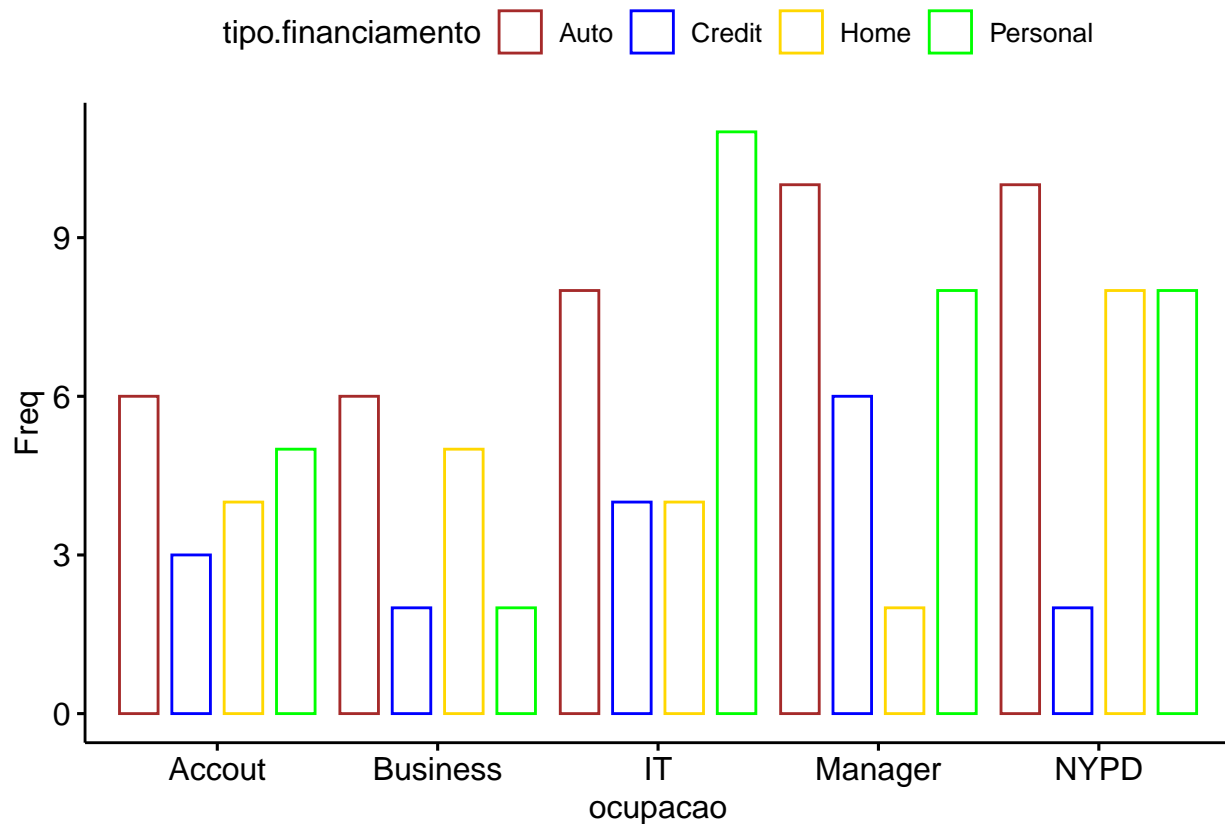
Transformando a tabela em um dataframe para visualizar o gráfico de barras

```
df2 <- as.data.frame(tbl2)

ggbarplot(df2, x="ocupacao", y="Freq",
          color = "tipo.financiamento",
          palette = c("brown", "blue", "gold", "green"))
```



```
ggbarplot(df2, x = "ocupacao", y = "Freq",  
          color = "tipo.financiamento", position = position_dodge(),  
          palette = c("brown", "blue", "gold", "green"))
```



Frequência relativa com a função `prop.table()`

Segundo parâmetro: 1 para linhas e 2 para colunas

```
prop.table(tbl2, 1)
```

```
##                ocupacao
## tipo.financiamento  Accout  Business      IT  Manager      NYPD
##      Auto      0.15000000 0.15000000 0.20000000 0.25000000 0.25000000
##      Credit  0.17647059 0.11764706 0.23529412 0.35294118 0.11764706
##      Home    0.17391304 0.21739130 0.17391304 0.08695652 0.34782609
##      Personal 0.14705882 0.05882353 0.32352941 0.23529412 0.23529412
```

```
prop.table(tbl2, 2)
```

```
##                ocupacao
## tipo.financiamento  Accout  Business      IT  Manager      NYPD
##      Auto      0.33333333 0.40000000 0.29629630 0.38461538 0.35714286
##      Credit  0.16666667 0.13333333 0.14814815 0.23076923 0.07142857
##      Home    0.22222222 0.33333333 0.14814815 0.07692308 0.28571429
##      Personal 0.27777778 0.13333333 0.40740741 0.30769231 0.28571429
```

Percentuais

```
round(prop.table(tbl2, 1), 2)*100
```

```
##                ocupacao
## tipo.financiamento Account Business IT Manager NYPD
##           Auto          15          15 20          25  25
##           Credit         18          12 24          35  12
##           Home          17          22 17           9  35
##           Personal        15           6 32          24  24
```