Prática no R! Roteiro 4 - Correlação

Elaborado por Luis Felipe Bortolatto da Cunha

21 de setembro de 2020

Contents

1. Introdução **1**

2. Diagrama de dispersão **3**

3. Coeficiente de correlação **4**

4. Teste de significância **5**

5. Matriz de correlação **5**

1. Introdução

Você pode baixar este roteiro em formato PDF neste endereço.

Este roteiro tem como objetivo auxiliar na execução de uma **análise de correlação** no software R. Isso inclui a criação de um **diagrama de dispersão**, cálculo do **coeficiente de correlação** e cálculo da estatística teste (**teste de significância**). Adicionalmente, ensina a computar e visualizar uma **matriz de correlação**.

Para executar a análise de correlação vamos utilizar dados demográficos e de consumo de água de 2010, para uma amostra de 4.417 municípios, extraídos do Censo Demográfico (IBGE) e do Sistema Nacional de Informações sobre Saneamento (SNIS), para investigar se **o consumo de água está correlacionado com a renda**, conforme análise apresentada por Carmo et al., 2013.

A base de dados está disponível para download no endereço abaixo:

https://1drv.ms/u/s!AjettDH-3Gbni9kM6Qmtsk8hxOhoFQ?e=ZdHG4a

1.1. Instalação e importação de pacotes

O software R já conta com funções básicas que permitem executar uma análise de correlação. Mas além das funções básicas, vamos usar os pacotes:

- `tidyverse`: para adicionar o operador `%>%` e a função `select()`
- `corrplot`: para visualizar como gráfico uma matriz de correlação

Se você ainda não possui esses pacotes instalados, é necessário executar o comando abaixo para instalar.

```
install.packages("tidyverse")
install.packages("corrplot")
```

Após a instalação, você pode importá-los com o uso da função `library()`.

```
library(tidyverse)
library(corrplot)
```

1.2. Importação da base de dados

A base de dados pode ser importada conforme as instruções do Roteiro 2.

Como ela está hospedada na nuvem, também pode ser importada com o endereço web como argumento, ao invés do endereço local, conforme o exemplo abaixo. É importante lembrar que a importação pelo endereço web exige conexão com a internet!

```
dados <- read.csv2("https://raw.githubusercontent.com/luisfelipebr/mti/master/dados/agua_rede1.csv",
                   encoding="UTF-8")
```

1.3. Análise exploratória

A função `names()` exibe os nomes das variáveis.

```
names(dados)
```

```
##  [1] "ID_IBGE"   "DOMICIL"   "REDE"      "PROPREDE"  "ID_SNIS"   "NOME_MUN"
##  [7] "UF"        "REGIAO"    "PIB"       "RENDAPITA" "GINI"      "IDH"
## [13] "IDH_CLASS" "GE012"     "AG001"     "AG020"     "AG022"     "CONSUMO1"
## [19] "CONSUMO2"
```

Como é possível ver após a execução do comando, os nomes das variáveis estão codificados. Mas a tabela abaixo apresenta uma descrição de cada variável.

Código	Descrição
ID_IBGE	Código IBGE (7 dígitos)
DOMICIL	Quantidade de Domicílios
REDE	Quantidade de Domicílios com Acesso à Rede Geral de Água
PROPREDE	Proporção de Domicílios com com Acesso à Rede Geral de Água (REDE/DOMICIL)
ID_SNIS	Código IBGE (6 dígitos)
NOME_MUN	Nome do Município
UF	Unidade da Federação
REGIAO	Região do País
PIB	Produto Interno Bruto 2010
RENDAPITA	Renda per Capita 2010
GINI	Índice GINI 2010
IDH	Índice de Desenvolvimento Humano 2010
IDH_CLASS	Classificação do Índice de Desenvolvimento Humano 2010: Muito Alto >= 0,9; Alto >= 0,8; Médio >= 0,5; Baixo < 0,5.
GE012	População Total Residente no Município
AG001	População Total Atendida com Abastecimento de Água
AG020	Volume Micromedido nas Economias Residenciais Ativas de Agua - 1.000 m3/ano
AG022	Quantidade de Economias Residenciais Ativas Micromedidas
CONSUMO1	Consumo de Água per capita - População Total - m3/ano (AG020/GE012)
CONSUMO2	Consumo de Água per capita - População Atendida - m3/ano (AG020/AG001)

Vamos usar as variáveis **RENDAPITA** (renda per capita) e **CONSUMO1** (consumo de água per capita) para testar as hipóteses:

- **H0 (hipótese nula):** o consumo de água *não* está correlacionado com a renda
- **H1: (hipótese alternativa)** o consumo de água está correlacionado com a renda

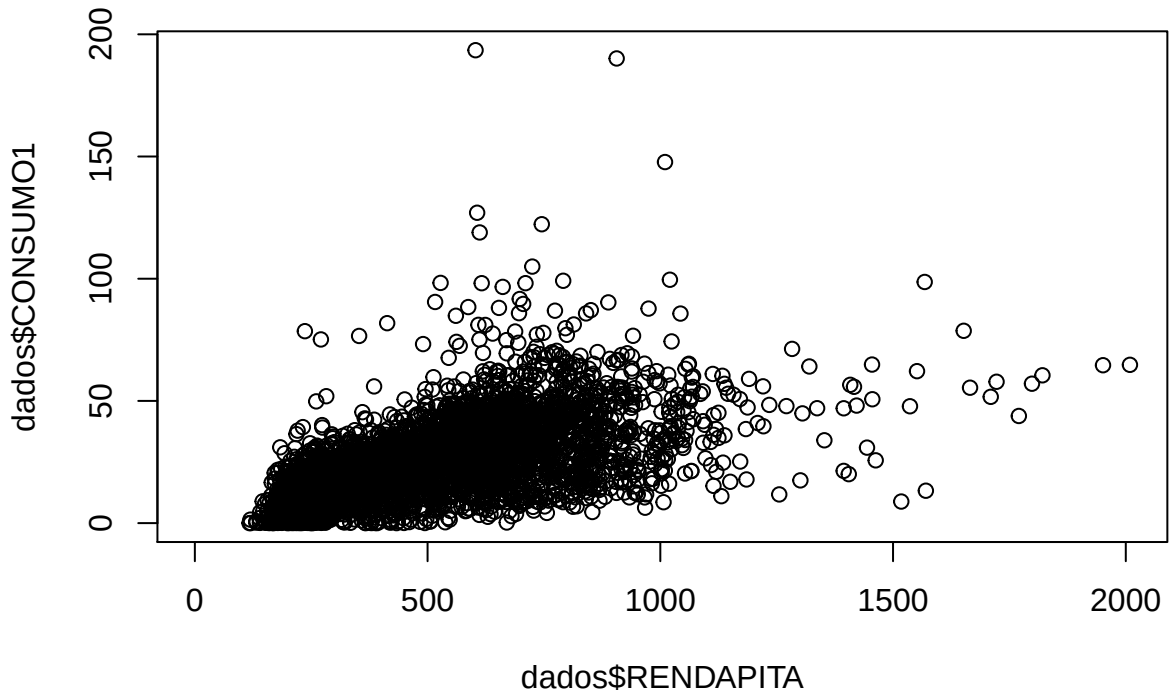
2. Diagrama de dispersão

Com o código das variáveis que desejamos investigar em mãos, já é possível criar um **diagrama de dispersão**.

O diagrama de dispersão é um gráfico que **exibe, para cada observação, o valor de uma variável contra o valor em outra variável**, permitindo a **visualização da correlação entre duas variáveis**.

Para construir um diagrama de dispersão, vamos usar a função `plot()` com dois argumentos (x e y), que são as duas variáveis que desejamos investigar a correlação.

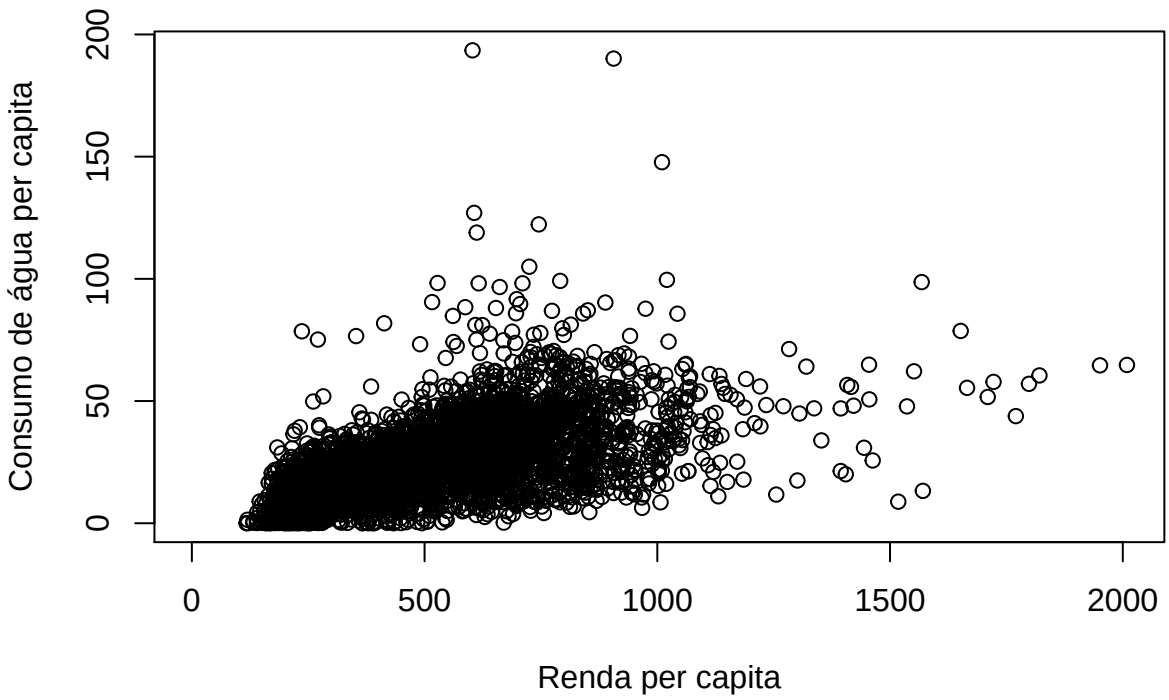
```
plot(x = dados$RENDAPITA,
     y = dados$CONSUMO1)
```



É possível ainda aprimorar o diagrama de dispersão adicionando alguns argumentos:

- `xlab`: um rótulo para o eixo x
- `ylab`: um rótulo para o eixo y
- Para explorar mais opções, acesse a documentação da função `plot` executando o comando `help(plot)`

```
plot(x = dados$RENDAPITA,
     y = dados$CONSUMO1,
     xlab = "Renda per capita",
     ylab = "Consumo de água per capita")
```



O diagrama de dispersão indica que existe uma **correlação positiva** entre as duas variáveis, ou seja, **quanto maior a renda, maior o consumo de água**. Mas vamos verificar essa relação estatísticamente através de um teste de correlação.

3. Coeficiente de correlação

Um **coeficiente de correlação** é uma **medida padronizada do relacionamento entre duas variáveis**. O **coeficiente de correlação de Pearson** apresenta um valor entre -1 e 1, sendo que:

- -1 indica uma correlação negativa perfeita
- 0 indica a ausência de relacionamento linear
- 1 indica uma correlação positiva perfeita

Para calcular o coeficiente de correlação de Pearson entre duas variáveis usaremos a função `cor()`, definindo os argumentos:

- `x`: Renda per capita
- `y`: Consumo de Água per capita
- `method = "pearson"`: para calcular o coeficiente de correlação de Pearson (também é possível calcular coeficiente de correlação de Spearman e o Tau de Kendall)
- `use = "complete.obs"`: para remover os valores faltantes (NA)

```
cor(x = dados$RENDAPITA,
    y = dados$CONSUMO1,
    method = "pearson",
    use = "complete.obs")
```

```
## [1] 0.6012537
```

A correlação entre a Renda e o Consumo de Água é de `0,6`, indicando a mesma **correlação positiva** que o gráfico de dispersão permitiu visualizar.

4. Teste de significância

Ainda é necessário confirmar se a correlação não se deve a um erro amostral, ou seja, ao acaso. Para isso vamos realizar um **teste de significância**.

Para testar a significâcia do coeficiente de correlação de Pearson é necessário executar a função `cor.test()`, especificando os argumentos:

- `x`: Renda per capita
- `y`: Consumo de Água per capita
- `method = "pearson"`: para calcular o coeficiente de correlação de Pearson
- `alternative = "two.sided"`: para teste de hipótese bilateral
- `conf.level = 0.95`: define o nível de significância como 0,05

```
cor.test(x = dados$RENDAPITA,
         y = dados$CONSUMO1,
         method = "pearson",
         alternative = "two.sided",
         conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dados$RENDAPITA and dados$CONSUMO1
## t = 49.997, df = 4415, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5820832 0.6197561
## sample estimates:
##       cor
## 0.6012537
```

Como o p-valor (`0,00000000000000022`) é menor que o nível de significância (`0,05`), **rejeitamos a hipótese nula (H0)**, ou seja, os dados indicam que não é possível falsear a hipótese alternativa (H1), sendo uma evidência de que **pode existir uma correlação entre o Consumo de Água e a Renda**.

5. Matriz de correlação

Quando você estiver explorando a sua hipótese, com os dados do seu trabalho de curso, você pode achar necessário analisar a correlação entre várias duplas de variáveis.

Se esse for o caso, é indicada a construção de uma **matriz de correlação**, que permite a **análise e visualização simultânea da relação entre duas variáveis**.

Para construir uma matriz de correlação, primeiro é necessário selecionar as variáveis que deseja investigar (com a função `select()`), para depois calcular o coeficiente de correlação entre cada dupla de variáveis (com a função `cor()`).

```
dados %>%
  select(RENDAPITA, CONSUMO1, CONSUMO2, PIB, IDH, GINI, REDE, PROPREDE) %>%
  cor(method = "pearson",
    use = "complete.obs")
```

```
##            RENDAPITA   CONSUMO1    CONSUMO2        PIB        IDH        GINI
```

```
## RENDAPITA  1.0000000  0.6012537  0.49269163 0.22623371  0.7415117 -0.21476760
## CONSUMO1   0.6012537  1.0000000  0.85700512 0.11760770  0.5143510 -0.26802459
## CONSUMO2   0.4926916  0.8570051  1.00000000 0.06518107  0.4255431 -0.25814509
## PIB        0.2262337  0.1176077  0.06518107 1.00000000  0.1190256  0.09083846
## IDH        0.7415117  0.5143510  0.42554311 0.11902559  1.0000000 -0.32599859
## GINI      -0.2147676 -0.2680246 -0.25814509 0.09083846 -0.3259986  1.00000000
## REDE       0.2203883  0.1165290  0.05918231 0.95388403  0.1142680  0.10596453
## PROPREDE   0.4012507  0.6117510  0.35458719 0.11643021  0.3731320 -0.17370841
##                 REDE   PROPREDE
## RENDAPITA 0.22038826  0.4012507
## CONSUMO1  0.11652903  0.6117510
## CONSUMO2  0.05918231  0.3545872
## PIB       0.95388403  0.1164302
## IDH       0.11426803  0.3731320
## GINI      0.10596453 -0.1737084
## REDE      1.00000000  0.1343521
## PROPREDE  0.13435207  1.0000000
```

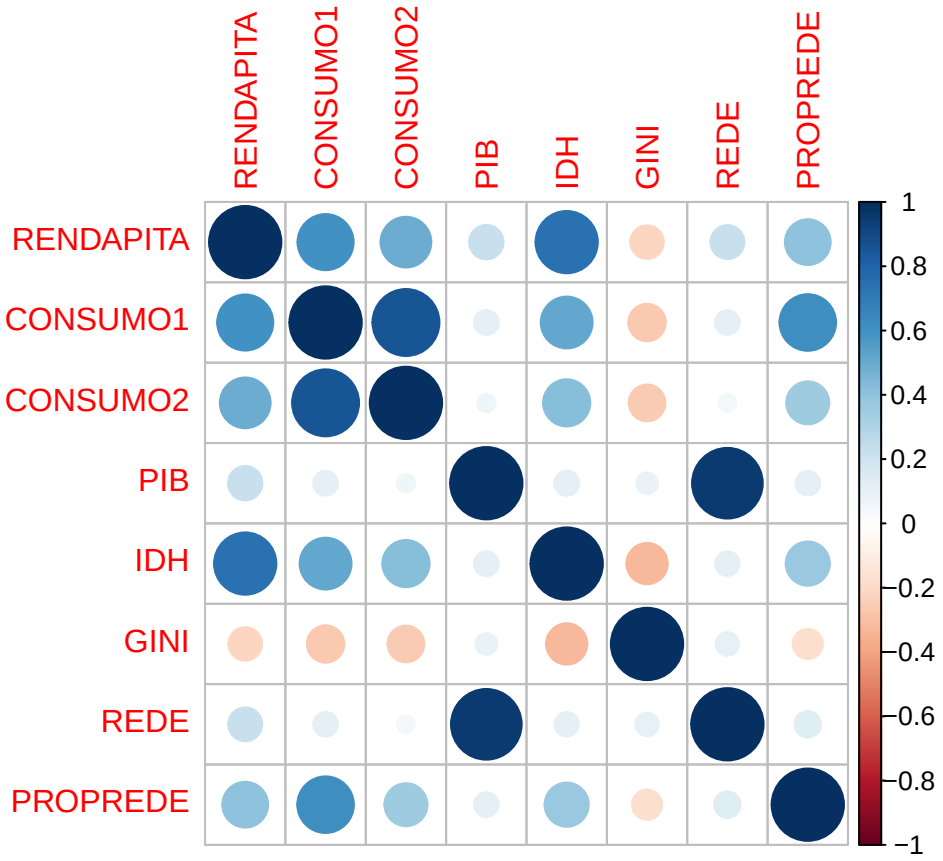
Se você quiser facilitar ainda mais a visualização da matriz de correlação, pode adicionar ainda a função `round()` para arredondar os valores do coeficiente de correlação para duas casas decimais.

```
dados %>%
  select(RENDAPITA, CONSUMO1, CONSUMO2, PIB, IDH, GINI, REDE, PROPREDE) %>%
  cor(method = "pearson",
    use = "complete.obs") %>%
  round(digits = 2)
```

```
##           RENDAPITA CONSUMO1 CONSUMO2  PIB   IDH  GINI REDE PROPREDE
## RENDAPITA      1.00     0.60     0.49 0.23  0.74 -0.21 0.22     0.40
## CONSUMO1       0.60     1.00     0.86 0.12  0.51 -0.27 0.12     0.61
## CONSUMO2       0.49     0.86     1.00 0.07  0.43 -0.26 0.06     0.35
## PIB            0.23     0.12     0.07 1.00  0.12  0.09 0.95     0.12
## IDH            0.74     0.51     0.43 0.12  1.00 -0.33 0.11     0.37
## GINI          -0.21    -0.27    -0.26 0.09 -0.33  1.00 0.11    -0.17
## REDE           0.22     0.12     0.06 0.95  0.11  0.11 1.00     0.13
## PROPREDE       0.40     0.61     0.35 0.12  0.37 -0.17 0.13     1.00
```

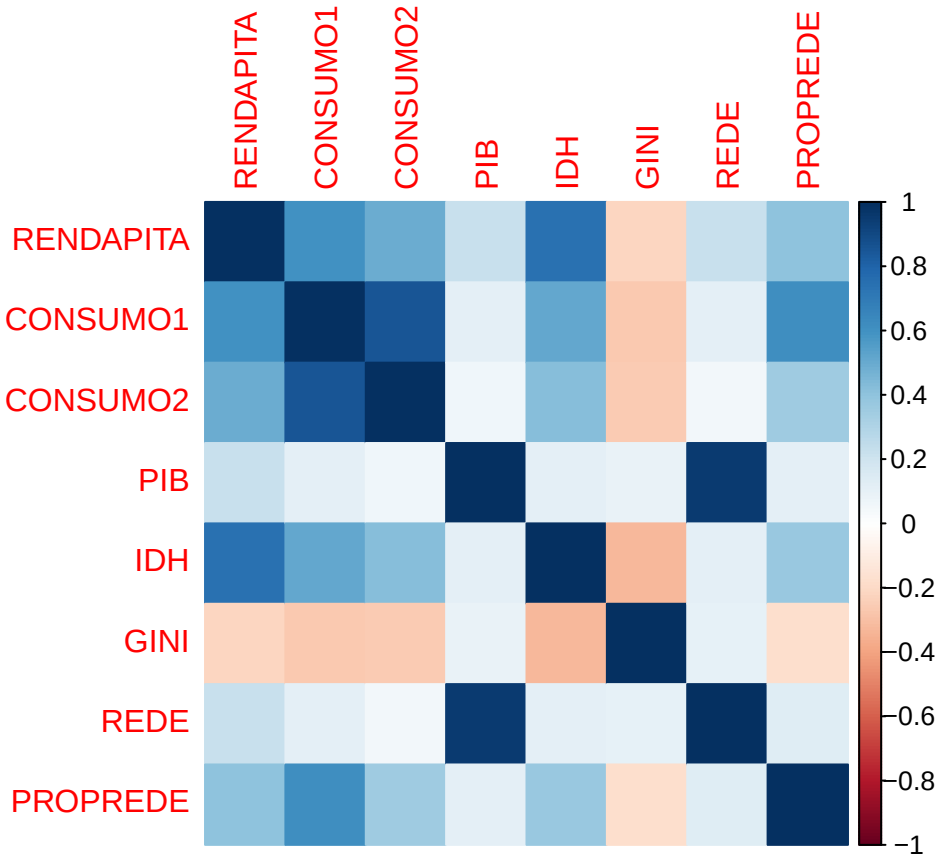
Também é possível visualizar a matriz de correlação construindo um gráfico, usando a função `corrplot()` do pacote `corrplot`.

```
dados %>%
  select(RENDAPITA, CONSUMO1, CONSUMO2, PIB, IDH, GINI, REDE, PROPREDE) %>%
  cor(method = "pearson",
    use = "complete.obs") %>%
  corrplot()
```

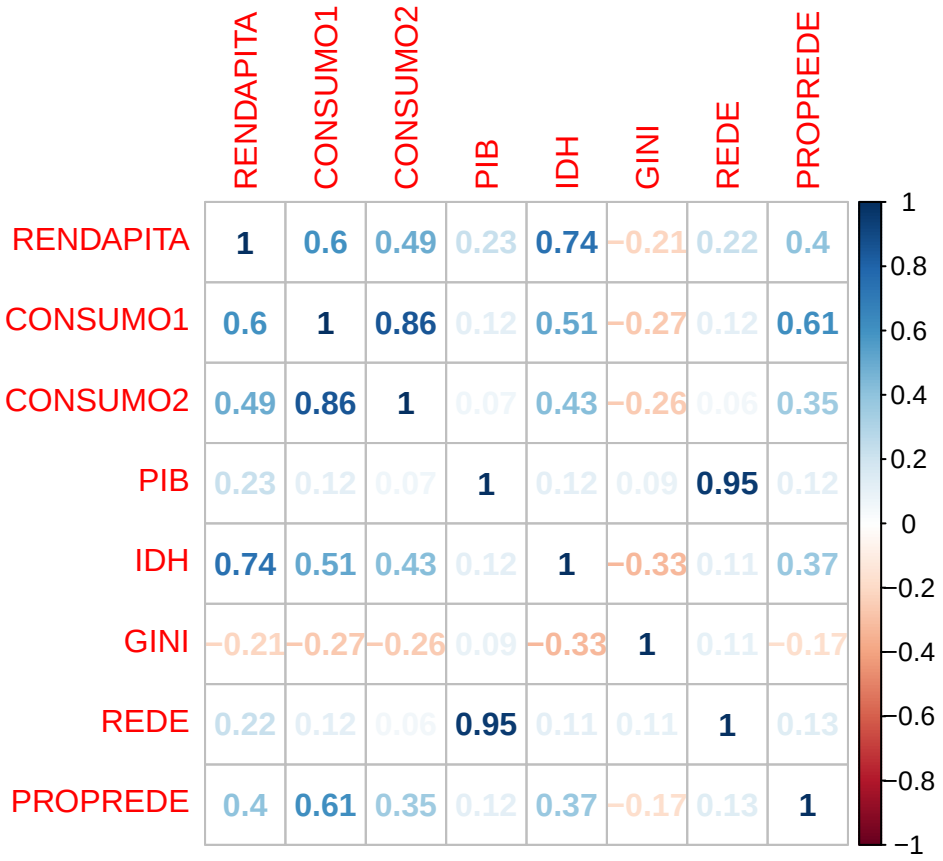


Manipulando o argumento `method`, você pode testar outras visualizações desse gráfico, conforme indica a documentação da função (`help(corrplot)`) e os exemplos abaixo.

```
dados %>%
  select(RENDAPITA, CONSUMO1, CONSUMO2, PIB, IDH, GINI, REDE, PROPREDE) %>%
  cor(method = "pearson",
    use = "complete.obs") %>%
  corrplot(method = "color")
```




```
dados %>%
  select(RENDAPITA, CONSUMO1, CONSUMO2, PIB, IDH, GINI, REDE, PROPREDE) %>%
  cor(method = "pearson",
    use = "complete.obs") %>%
  corrplot(method = "number")
```


RENDAPITA
CONSUMO1
CONSUMO2
PIB
IDH
GINI
REDE
PROPREDE
RENDAPITA 1 0.6 0.49 0.23 0.74 −0.21 0.22 0.4
CONSUMO1 0.6 1 0.86 0.12 0.51 −0.27 0.12 0.61
CONSUMO2 0.49 0.86 1 0.07 0.43 −0.26 0.06 0.35
PIB 0.23 0.12 0.07 1 0.12 0.09 0.95 0.12
IDH 0.74 0.51 0.43 0.12 1 −0.33 0.11 0.37
GINI −0.21 −0.27 −0.26 0.09 −0.33 1 0.11 −0.17
REDE 0.22 0.12 0.06 0.95 0.11 0.11 1 0.13
PROPREDE 0.4 0.61 0.35 0.12 0.37 −0.17 0.13 1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1