Correlograma

O que é uma matriz de correlação?

É uma tabela que mostra o nível de correlação (Pearson, Spearman, Kendall, etc.) entre variáveis (quantitativas/ordinais). Veja um exemplo a seguir:

Tabela 1: Correlação linear de Pearson entre variáveis do dataset `mtcars`

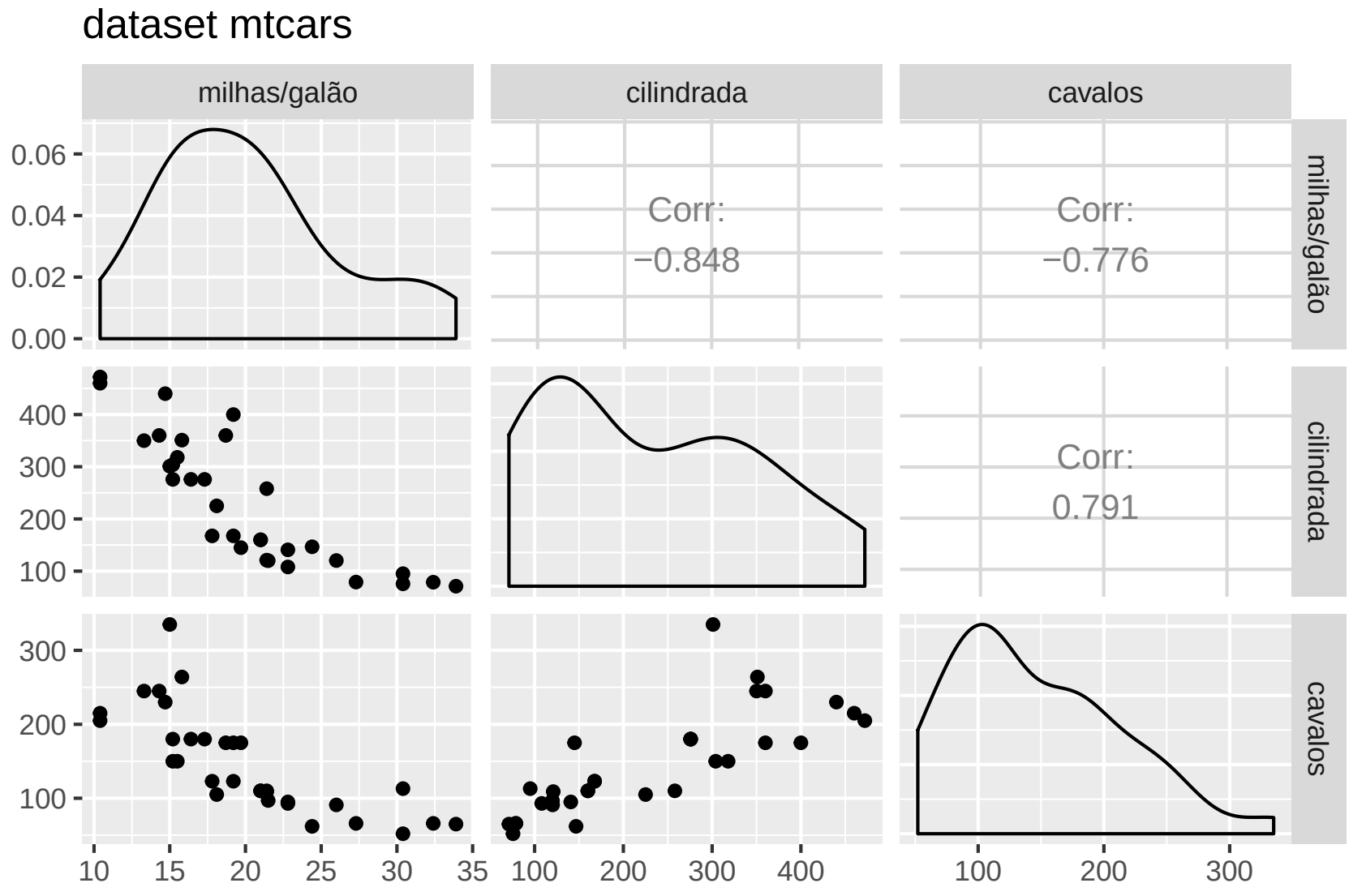
	mpg	disp	hp
mpg	1.0000000	-0.8475514	-0.7761684
disp	-0.8475514	1.0000000	0.7909486
hp	-0.7761684	0.7909486	1.0000000

Esse tipo de representação pode ser muito útil, mas com o aumento do número de variáveis envolvidas surge a dificuldade em identificar as correlações mais relevantes. Para facilitar essa análise, temos como alternativa o correlograma.

O que é o correlograma?

O correlograma facilita a visualização de uma matriz de correlação e ainda permite explorar alguns outros aspectos dos dados, incorporando gráficos e cores, além das medidas de correlação em si.

Correlograma com o pacote GGally



Correlograma com o pacote GGally

```
library(GGally)

ggpairs(data = mtcars,
        columns = c("mpg", "disp", "hp"),
        columnLabels = c("milhas/galão",
                         "cilindrada",
                         "cavalos"),
        title = "dataset mtcars")
```

data: data frame contendo os dados;

columns: indica quais variáveis serão selecionadas para o gráfico;

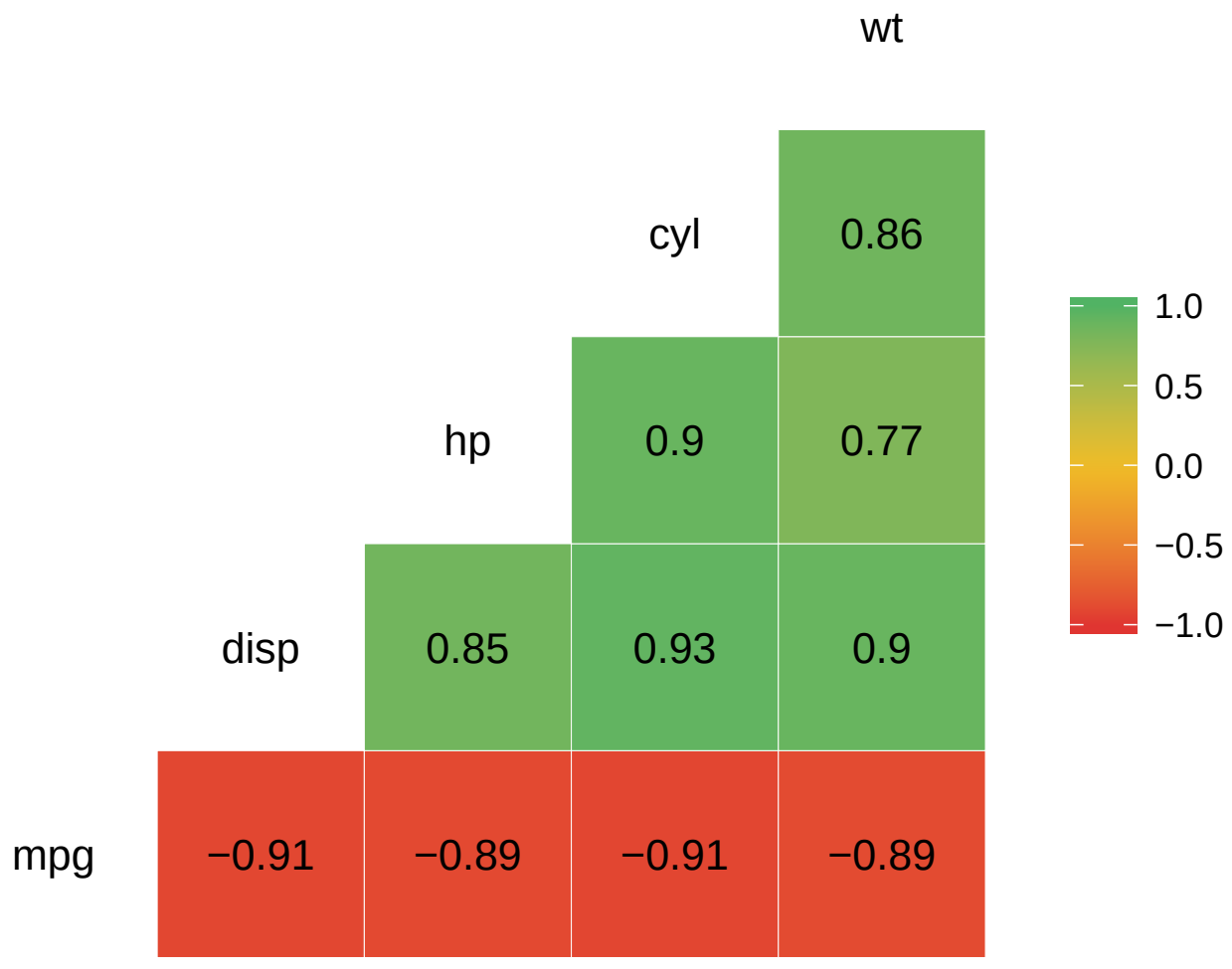
columnLabels: rótulos para as respectivas variáveis;

title: título do gráfico.

Obs: Com este pacote é ainda possível escolher os tipos de gráficos mostrados no correlograma (densidade/histograma/barras) de acordo com o tipo de variável (contínua/discreta), além das medidas de correlação, tanto para a parte acima quanto abaixo da diagonal principal.

Em seguida veremos uma outra forma de correlograma, feito pelo mesmo pacote. Este permite explorar um número muito maior de variáveis, resumindo a matriz de correlação em uma grade multicolorida.

Correlograma com o pacote GGally



Como implementar?

```r
library(GGally); library(dplyr)

mtcars %>%
  select(mpg, disp, hp, cyl, wt) %>%
  ggcorr(data = ., method = c("everything", "spearman"),
         label=TRUE, label_round=2,
         low="#e03531", mid="#f0bd27", high="#51b364")

# Obs: escolher as variáveis previamente
```

data: data frame contendo os dados;

method: vetor contendo o tipo de tratamento aos dados faltantes e a medida de correlação que será usada;

label: ativa/desativa o aparecimento do valor das correlações;

label_round: número de casas decimais no `label`;

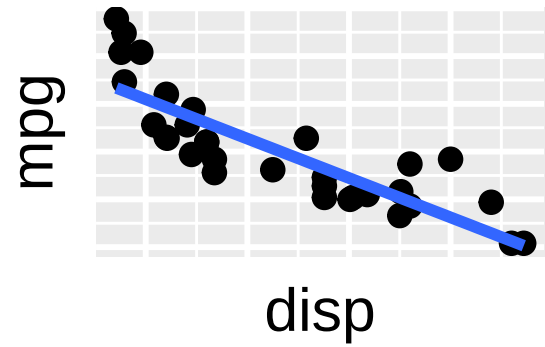
low, **mid**, **high**: cores inferiores, intermediárias e superiores da escala de cores, respectivamente.

Precauções

- Variáveis com nível de mensuração pelo menos ordinal;
- Escolha adequada das medidas de correlação;
- Tratar dados faltantes (NAs);
- Rótulos curtos para as variáveis.

Alguns dos usos do correlograma

- Encontrar relações entre variáveis a serem exploradas



- Estabelecer hipóteses e inferir sobre correlações

$\mathbf{H_0}$: $\text{corr}(disp, mpg) = 0$ $\mathbf{H_1}$: $\text{corr}(disp, mpg) < 0$

- Identificar possíveis relações causais

Ex: `veículos com mais cilindros consomem mais combustível?`