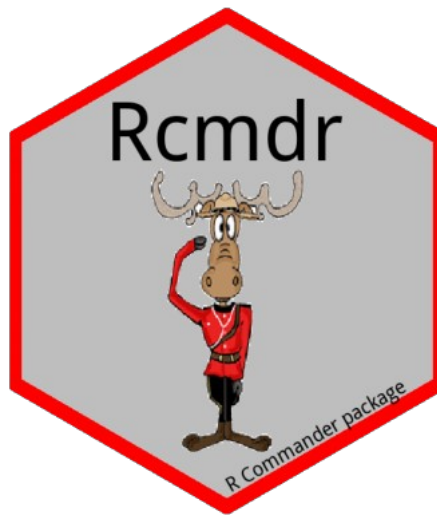


# Introdução ao R Commander

---

Notas de Aula



Felipe Rafael Ribeiro Melo

Departamento de Métodos Quantitativos

Universidade Federal do Estado do Rio de Janeiro - UNIRIO

Rio de Janeiro, Brasil

Setembro de 2024

# Sumário

<b>1</b>	<b>Introdução</b>	<b>6</b>
1.1	<i>Download</i> e instalação do <i>software</i> R . . . . .	6
1.2	Instalando o pacote <b>Rcmdr</b> . . . . .	7
1.3	Carregando o pacote <b>Rcmdr</b> . . . . .	8
<b>2</b>	<b>Conjuntos de dados</b>	<b>10</b>
2.1	Criando um conjunto de dados . . . . .	10
2.2	Importando conjunto de dados com formatos <i>.xls</i> ou <i>.xlsx</i> . . . . .	14
2.3	Importando conjunto de dados com formato <i>.txt</i> ou <i>.csv</i> . . . . .	14
2.4	Exportando um conjunto de dados do R Commander . . . . .	16
2.5	Conjuntos de dados nativos do R Commander . . . . .	17
<b>3</b>	<b>O menu <i>Dados</i></b>	<b>17</b>
3.1	Salvando conjuntos de dados . . . . .	18
3.2	Carregando conjuntos de dados (no formato <i>.RData</i> ) . . . . .	18
3.3	Agrupar dois conjuntos de dados . . . . .	18
3.4	Nomes das variáveis no conjunto de dados ativo . . . . .	19
3.5	Definir nomes das observações em cada linha . . . . .	19
3.6	Subconjunto do conjunto de dados ativo . . . . .	20
3.7	Ordenar linhas do conjunto de dados . . . . .	21
3.8	Modificação de variáveis no conjunto de dados . . . . .	24
3.8.1	Converter variável quantitativa (numérica) para variável qualitativa	24
3.8.2	Reordenar níveis dos fatores . . . . .	25
3.8.3	Agrupar em classes uma variável quantitativa . . . . .	26
3.8.4	Computar nova variável . . . . .	27
3.8.5	Renomear e apagar variáveis . . . . .	28
3.8.6	Recodificar variáveis . . . . .	29
<b>4</b>	<b>O menu <i>Gráficos</i></b>	<b>31</b>
4.1	Gradiente de cores . . . . .	31
4.2	Gráfico de setores (gráfico de pizza) . . . . .	32
4.3	Gráfico de barras . . . . .	33

4.4	Gráfico de barras múltiplas . . . . .	36
4.5	Gráfico de pontos . . . . .	39
4.6	Gráfico de hastes ( <i>Plot discrete numeric variable</i> ) . . . . .	41
4.7	Histograma . . . . .	43
4.8	Boxplot . . . . .	45
4.9	Diagrama de dispersão . . . . .	47
4.10	Salvando gráficos . . . . .	50
<b>5</b>	<b>O menu <i>Estatísticas</i></b>	<b>50</b>
5.1	Resumos numéricos de todas as variáveis . . . . .	50
5.2	Resumo numérico de uma variável (quantitativa) . . . . .	51
5.3	Contando dados faltantes . . . . .	52
5.4	Distribuições de frequências . . . . .	53
5.5	Matriz de correlação . . . . .	54
5.6	Tabelas de contingência . . . . .	55
<b>6</b>	<b>Considerações finais</b>	<b>57</b>
	<b>Referências</b>	<b>57</b>

# Lista de Figuras

1	Apresentando o R Commander. . . . .	9
2	Planilha para criação de conjunto de dados no R Commander. . . . .	11
3	Apresentando as janelas <i>R Script</i> , <i>Output</i> e <i>Mensagens</i> . . . . .	13
4	Visualizando conjunto de dados no R Commander. . . . .	13
5	Importação de conjunto de dados em formato <i>.txt</i> ou <i>.csv</i> . . . . .	15
6	Exportando conjunto de dados. . . . .	17
7	Entradas da coluna <i>Nome</i> como nomes dos casos no conjunto de dados <i>Turma_1</i> . . . . .	20
8	Definindo subconjunto do conjunto de dados <i>Turmas_2_e_3</i> apenas com as meninas. . . . .	21
9	Ordenando linhas do Conjunto de dados <i>Turma_1</i> pela nota dos alunos (da maior par a menor). . . . .	22
10	Ordenando linhas do conjunto de dados <i>Turmas_2_e_3</i> pela nota dos alunos e, em seguida, pelo número de faltas dos alunos, ambos em ordem decrecente. . . . .	23
11	Convertendo a variável de rótulos numéricos <i>Casado</i> para a variável qualitativa <i>Casado_fator</i> no conjunto de dados <i>Bussab</i> . . . . .	25
12	Agrupando variável quantitativa em classes. . . . .	27
13	Computando nova variável no conjunto de dados <i>Bussab</i> : <i>Renda_em_reais</i> . . . . .	28
14	Criando a variável <i>Superior_completo</i> no conjunto de dados <i>Bussab</i> por meio de recodificação da variável <i>Instrução</i> . . . . .	29
15	Criando a variável <i>Situação</i> no conjunto de dados <i>Turmas_2_e_3</i> por meio de recodificação da variável <i>Notas</i> . . . . .	30
16	Gradiente de cores original do R Commander. . . . .	31
17	Gráfico de setores da variável <i>Casado_fator</i> . . . . .	33
18	Construindo gráfico de barras para a variável <i>Proced</i> . . . . .	34
19	Gráfico de barras da variável <i>Proced</i> . . . . .	34
20	Gráfico de barras da variável <i>Filhos_fator</i> . . . . .	35
21	Construindo gráfico de barras múltiplas com as variáveis <i>Proced</i> e <i>Instrução</i> . . . . .	37
22	Gráfico de barras múltiplas com as variáveis <i>Proced</i> e <i>Instrução</i> . . . . .	37
23	Perfil percentual do nível de instrução em cada região de procedência. . . . .	38

24	Gráfico de pontos da variável <i>Filhos</i> . . . . .	40
25	Gráficos de pontos da variável <i>Idade</i> para os casados e para os não casados. . . . .	40
26	Gráfico de hastes da variável <i>Filhos</i> . . . . .	41
27	Gráficos de hastes da variável <i>Idade</i> para os casados e para os não casados. . . . .	43
28	Histograma da variável <i>Renda</i> . . . . .	44
29	Histogramas da variável <i>Renda</i> para os casados e para os não casados. . . . .	45
30	Boxplot da variável <i>Idade</i> . . . . .	46
31	Boxplots da variável <i>Renda</i> agrupados pelo nível de instrução. . . . .	47
32	Construindo diagrama de dispersão. . . . .	48
33	Diagrama de dispersão entre as variáveis <i>Idade</i> e <i>Renda</i> . . . . .	49
34	Diagramas de dispersão sobrepostos entre as variáveis <i>Idade</i> e <i>Renda</i> dos indivíduos casados (na cor azul) e dos indivíduos que não são casados (na cor magenta). . . . .	49
35	Resumo numérico da variável <i>Renda</i> . . . . .	51
36	Resumos numéricos por nível de instrução da variável <i>Renda</i> . . . . .	52
37	Distribuição de frequências por classes solicitada no Exercício 11. . . . .	53
38	Distribuição de frequências da variável <i>Instrução</i> . . . . .	54
39	Matriz de correlação com as variáveis <i>Idade</i> e <i>Renda</i> . . . . .	54
40	Tabela de contingência com as variáveis <i>Casado_fator</i> e <i>Instrução</i> , seguida do teste qui-quadrado de independência de Pearson entre estas variáveis. . . . .	55
41	Perfil percentual da variável <i>Instrução</i> (escolhida como variável coluna) em cada atributo da variável <i>Casado_fator</i> (escolhida como variável linha). . . . .	56
42	Perfil percentual da variável <i>Casado_fator</i> (escolhida como variável linha) em cada atributo da variável <i>Instrução</i> (escolhida como variável coluna). . . . .	56
43	Percentual de cada cruzamento dos atributos das variáveis <i>Casado_fator</i> e <i>Instrução</i> em relação ao total de observações. . . . .	56

# 1 Introdução

Em disciplinas de Estatística, Bioestatística e afins, o uso *software* estatístico R (R Core Team, 2024) traz diversas vantagens, como a geração de gráficos e de entradas de tabelas, além de resolver rapidamente cálculos trabalhosos. Contudo, tal *software* apresenta uma interface sob a qual é necessário escrever linhas de comando para a geração ou importação de dados, bem como para a aplicação de qualquer ferramenta estatística. Felizmente, o R possui uma grande quantidade de **pacotes**, os quais fornecem funcionalidades específicas. Este material explora, de forma introdutória, a interface “amigável” R Commander, proveniente de um pacote chamado Rcmdr (Fox; Bouchet-Valat, 2024), com menus que propiciam várias funcionalidades do R sem a necessidade de escrever linhas de comando, tais como:

- digitação de conjuntos de dados;
- importação de conjuntos de dados construídos em outros *softwares* (Excel, Bloco de notas, *etc.*);
- análises de dados por meio de gráficos;
- análises de dados por meio de tabelas.

A seguir, uma breve explicação de como baixar e instalar o *software* R, e como instalar e carregar o pacote Rcmdr.<sup>1</sup>

## 1.1 Download e instalação do *software* R

O *software* R é gratuito e pode ser obtido em [www.r-project.org](http://www.r-project.org). Acesse essa página no seu navegador de preferência e siga os passos abaixo:

1. Clique em **CRAN** (lado esquerdo da tela do seu navegador);
2. Opte por um dos *mirrors* disponíveis (aconselhável optar um *mirror* de instituição brasileira);
3. Escolha o seu sistema operacional (*Linux*, *macOS* ou *Windows*) no campo *Download and Install R* (os passos seguintes derivam da escolha do sistema operacional *Windows*);

---

<sup>1</sup>Este material foi escrito utilizando o *software* R versão 4.4.1 e pacote Rcmdr versão 2.9-2.

4. Clique em **base**;
5. Por fim, clique no *link* destacado (algo do tipo **Download R 4.4.1 for Windows**) para baixar o programa.

**Observação 1.** Qualquer um dos *links* abaixo leva diretamente ao último dos passos acima para *download* do R no sistema operacional *Windows*:

- <https://cran-r.c3sl.ufpr.br/bin/windows/base/>;
- <https://vps.fmvz.usp.br/CRAN/bin/windows/base/>;
- <https://brieger.esalq.usp.br/CRAN/bin/windows/base/>.

Uma vez concluído o *download*, o processo de instalação do programa é simples, bastando apenas dar um duplo clique no arquivo executável baixado, clicar em OK na tela do idioma do instalador e clicar sequencialmente em *Avançar* até a conclusão do processo de instalação. Assim, o R está instalado e pronto para ser utilizado.



## 1.2 Instalando o pacote Rcmdr

Ao abrirmos o R, visualizamos uma janela grande (chamada *R Gui*) contendo uma janela chamada *R Console*. Nesta última, vemos uma série de créditos na cor azul e, em seguida, o sinal **>** na cor vermelha. Este símbolo é chamado ***prompt de comando***, e significa que o R está apto a receber um **comando** nesta linha. Ou seja, na interface apresentada, é necessário escrever linhas de comando para a realização de qualquer operação. Entretanto, o R possui uma grande quantidade de **pacotes**, os quais fornecem funcionalidades específicas.

Usamos nesta disciplina um pacote chamado **Rcmdr** (uma forma abreviada de **R Commander**), que fornece ao usuário do R **uma interface mais “amigável”**, com menus que propiciam várias funcionalidades do R ***sem a necessidade de escrever linhas de comando***.

Para a instalação do pacote Rcmdr, clique em *Pacotes* na parte superior da janela *R Gui* e, em seguida, *Instalar pacote(s)*. Uma janela chamada *Secure CRAN mirrors* é aberta, e nela, selecione o espelho CRAN de sua preferência (sugestões: *0-Cloud*, *Brazil (PR)*,


*Brazil (SP 1)* ou *Brazil (SP 2)*). Dê OK e uma lista com todos os pacotes disponíveis para instalação (no espelho CRAN escolhido) é aberta. Os pacotes estão listados em ordem alfabética. Procure<sup>2</sup> por *Rcmdr*, clique nele uma única vez e siga uma das duas opções abaixo – a primeira delas é mais simples e suficiente para o conteúdo deste material, ao passo que a segunda opção demanda um tempo de instalação maior.

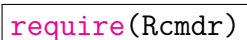
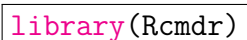
- Para instalar **apenas** o pacote *Rcmdr*, clique em OK.
- Para uma instalação **mais completa**, o *Rcmdr* pode ser instalado junto aos seus *plugins*. Para isto, após clicar em *Rcmdr*, mantenha a tecla  apertada e, com a seta  do seu teclado, marque todos os pacotes que começam com a expressão “*Rcmdr*”. Com todos estes pacotes (*Rcmdr* até *RcmdrPlugin.WorldFlora*) já marcados com fundo azul, clique em OK para instalá-los.


Caso sejam abertas duas pequenas janelas, uma após a outra, com opções *Sim* e *Não*, selecione *Sim* em ambas. A conclusão do processo de instalação se dá quando o *prompt* de comando (sinal *>* na cor vermelha) aparecer novamente no canto esquerdo inferior da janela *R Console*.

**Observação 2.** É necessária conexão com a Internet para a instalação de pacotes do R por meio do menu *Pacotes* → *Instalar pacote(s)*.

### 1.3 Carregando o pacote *Rcmdr*

Uma vez **instalado** o pacote *Rcmdr* no seu computador, **não** há mais necessidade de repetir o processo descrito na Subseção 1.2. Quando quisermos trabalhar com o R Commander, será necessário apenas **carregá-lo**. Para isto, digite no R (janela *R Console*) um dos dois comandos abaixo e aperte .

 ou .

Note que, em cada um dos comandos acima, apenas uma das letras é maiúscula. Ao digitar qualquer um destes dois comandos na janela *R Console* e apertar  em seguida, uma nova janela é aberta: a janela *R Commander* (Figura 1), na qual vamos trabalhar.

<sup>2</sup>Caso o pacote *Rcmdr* não esteja nesta lista, clique em *Cancelar*, escolha outro espelho CRAN no menu *Pacotes* → *Escolher espelho CRAN* e acesse novamente o menu *Pacotes* → *Instalar pacote(s)*.



Ela é dividida em três janelas: *R Script*, *Output* e *Mensagens*. Na Subseção 2.1, é debatido o propósito de cada uma delas.

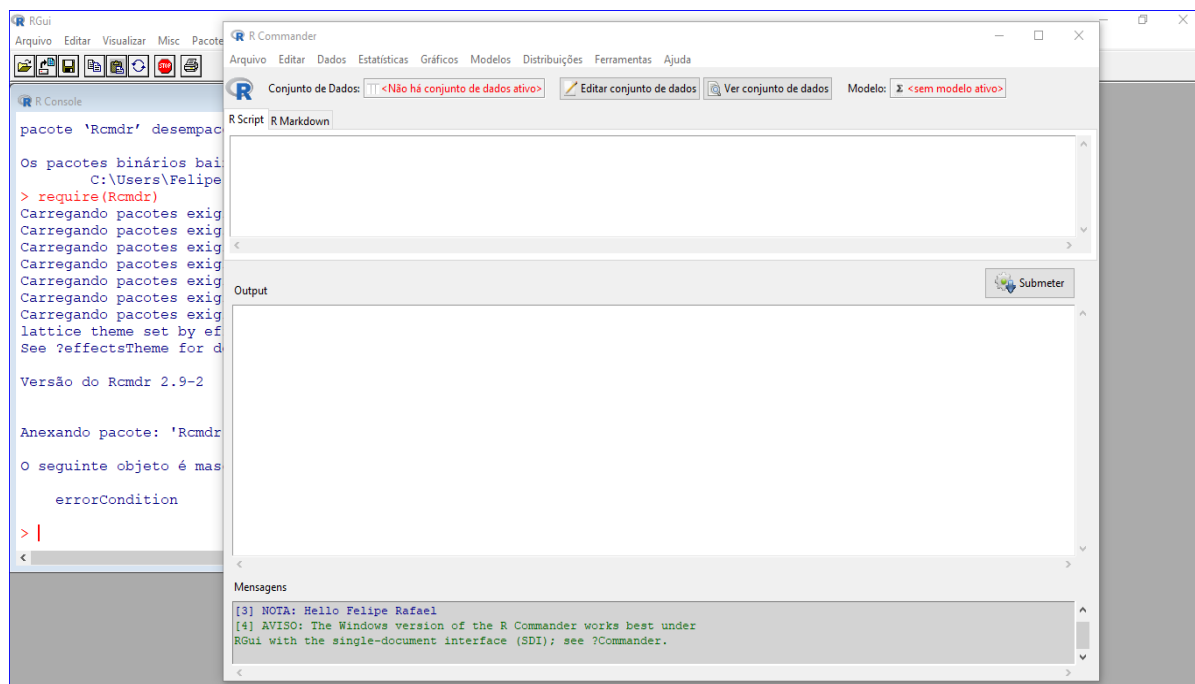


Figura 1: Apresentando o R Commander.

**Observação 3.** Na primeira vez que o R Commander for carregado, pode surgir uma pequena janela com opções *Sim* e *Não*. Clique em *Sim* e, na janela seguinte, em OK.

**Observação 4.** Se, ao carregar o pacote `Rcmdr`, a janela *R Commander* não abrir, verifique a mensagem de erro que foi gerada na janela *R Console*. É possível que seja ausência de algum outro pacote necessário. Por exemplo, uma mensagem de erro com os trechos

```
não há nenhum pacote chamado 'zoo'
```

e

```
Failed with error: pacote 'sandwich' não pôde ser carregado
```

acusa a necessidade da instalação de pelo menos um desses pacotes (`zoo` ou `sandwich`), o que pode ser feito no menu *Pacotes* → *Instalar pacote(s)* da janela *R Console*. Após a(s) instalação(ões) faltante(s), carregue novamente o pacote `Rcmdr`. Se, ainda assim, surgir uma nova mensagem de erro em vez da janela *R Commander*, instale o(s) novo(s) pacote(s) mencionado(s) nesta nova mensagem de erro antes de uma nova tentativa de carregar o pacote `Rcmdr`.

**Observação 5.** Carregar o R Commander em uma janela *R Console* na qual ele já foi anteriormente carregado não surte efeito. Portanto, caso você feche a janela *R Commander* e queira abri-la novamente, é necessário fechar o R (janela *R Gui*) e abri-lo novamente.

Nas próximas seções, seguem detalhadas e explicadas algumas das muitas funcionalidades possíveis de serem implementadas na interface R Commander.

## 2 Conjuntos de dados

Já temos um *software* capaz de fazer análises estatísticas, bem como um pacote que torna viável executar tais análises para quem desconhece as linhas de comando do R. Interessa-nos agora aplicar as análises desejadas a um **conjunto de dados** de interesse.

Em geral, conjuntos de dados não estão em formatos que o R reconhece imediatamente. É comum, por exemplo, um conjunto de dados ser escrito numa planilha Excel (formato *.xls* ou *.xlsx*), ou até mesmo no Bloco de Notas (formato *.txt*). Para que o R e, consequentemente, o R Commander consigam “entender” um conjunto de dados criado fora deles, é necessário **importar** este arquivo (onde está o conjunto de dados) com o R Commander. Aqui, tratamos apenas da importação de conjuntos de dados em arquivos nos formatos: *.xls*, *.xlsx*, *.txt* e *.csv*. Mas, primeiramente, vamos ver como criar um conjunto de dados diretamente no R Commander via digitação.

### 2.1 Criando um conjunto de dados

A título de ilustração, vamos escrever este pequeno conjunto de dados abaixo.

Nome	Sexo	Nota	Faltas
André	M	8.5	0
Carla	F	7.0	3
Fernando	M	4.5	4
Larissa	F	9.0	2

Na barra de ferramentas da janela *R Commander*, vá em **Dados → Novo conjunto de dados**. Defina um nome para o conjunto de dados a ser digitado, sem utilizar espaço. Vamos usar o nome *Turma\_1*. Ao dar OK, uma pequena janela chamada *Editor de dados* (ou *Data Editor*) é aberta por cima da janela *R Commander*.

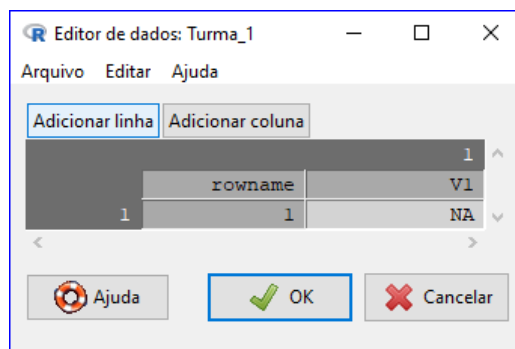



Figura 2: Planilha para criação de conjunto de dados no R Commander.

A cada clique no botão *Adicionar linha*, uma linha é adicionada, e a cada clique no botão *Adicionar coluna*, uma coluna é adicionada. Lembrando que o conjunto de dados que vamos digitar possui quatro colunas (*Nome*, *Sexo*, *Nota* e *Faltas*) e quatro observações (alunos). Ignore a coluna *rowname* e faça o seguinte:

- clique no botão *Adicionar coluna* três vezes para a planilha ficar com quatro colunas (nomeadas genericamente de *V1*, *V2*, *V3* e *V4*);
- clique no botão *Adicionar linha* três vezes para a planilha ficar com quatro linhas;
- clique na célula onde está escrito *V1* e escreva *Nome*;
- clique na célula onde está escrito *V2* e escreva *Sexo*;
- clique na célula onde está escrito *V3* e escreva *Nota*;
- clique na célula onde está escrito *V4* e escreva *Faltas*.

Agora preencha a planilha com nome, sexo, nota e faltas de cada aluno nas células das respectivas colunas (não se preocupe em apagar os NA's; basta "escrever por cima"). **Não** use  para mudar de célula! Use o *mouse* ou as setas do teclado. Ainda, ao digitar as notas, use ponto ao invés de vírgula como separador de casa decimal (o R e seus pacotes sempre trabalham com ponto como separador de casa decimal). Ao terminar, clique em OK.

**Observação 6.** Para apagar linhas adicionadas acidentalmente: clique com o botão esquerdo do *mouse* em qualquer célula dessa linha; clique com o botão direito do *mouse*;

e escolha a opção *Apagar linha corrente*. Procedimento análogo para apagar colunas adicionadas acidentalmente, porém clicando em *Apagar coluna corrente*.

Observe a Figura 3. Em seu canto superior esquerdo, está escrito *Turma\_1* na cor azul (ao lado de “Conjunto de dados:”), que é o conjunto de dados que acabamos de criar no R Commander. Isto quer dizer que este é, atualmente, o **conjunto de dados ativo**, ou seja, o conjunto de dados selecionado para a análise de dados. Ao lado direito do nome do conjunto de dados ativo, temos dois botões úteis:

- **Editar conjunto de dados:** Uma janela com o conjunto de dados ativo é aberta para a realização de modificações em qualquer célula desejada. Feitas as modificações (caso necessário), não se esqueça de clicar em OK.
- **Ver conjunto de dados.** Uma janela com o conjunto de dados ativo é aberta e, para visualizá-la, é necessário que a janela *R Commander* não esteja maximizada. Esta janela é apenas para visualização, de modo que, ao fechá-la, o conjunto de dados não é perdido. Observe a Figura 4. A coluna *rowname*, a qual ignoramos na digitação, é a coluna com fundo cinza à esquerda do conjunto de dados, com os números inteiros de 1 até 4 indicando 1ª linha, 2ª linha, 3ª linha e 4ª linha.

Ainda visualizando a Figura 3, estamos em posição de discutir as janelas *R Script*, *Output* e *Mensagens* da janela *R Commander*.

- **R Script:** Nesta janela são impressas as linhas de comando que foram executadas. Isso mesmo, linhas de comando! Quase tudo que é feito por meio de menus no R Commander, como, por exemplo, importação ou edição de um conjunto de dados, gera na janela *R Script* a respectiva linha de comando que é utilizada no R para obter a mesma saída. A última linha de comando nesta janela se refere ao último comando executado, a penúltima linha de comando se refere ao penúltimo comando executado, e assim sucessivamente.
- **Output:** Esta é a janela de saída, na qual são exibidos os resultados de alguns comandos executados.
- **Mensagens:** Informações que o sistema julga relevante compartilhar com o usuário, tais como avisos (na cor verde), notas (na cor azul) e mensagens de erro (na cor vermelha).

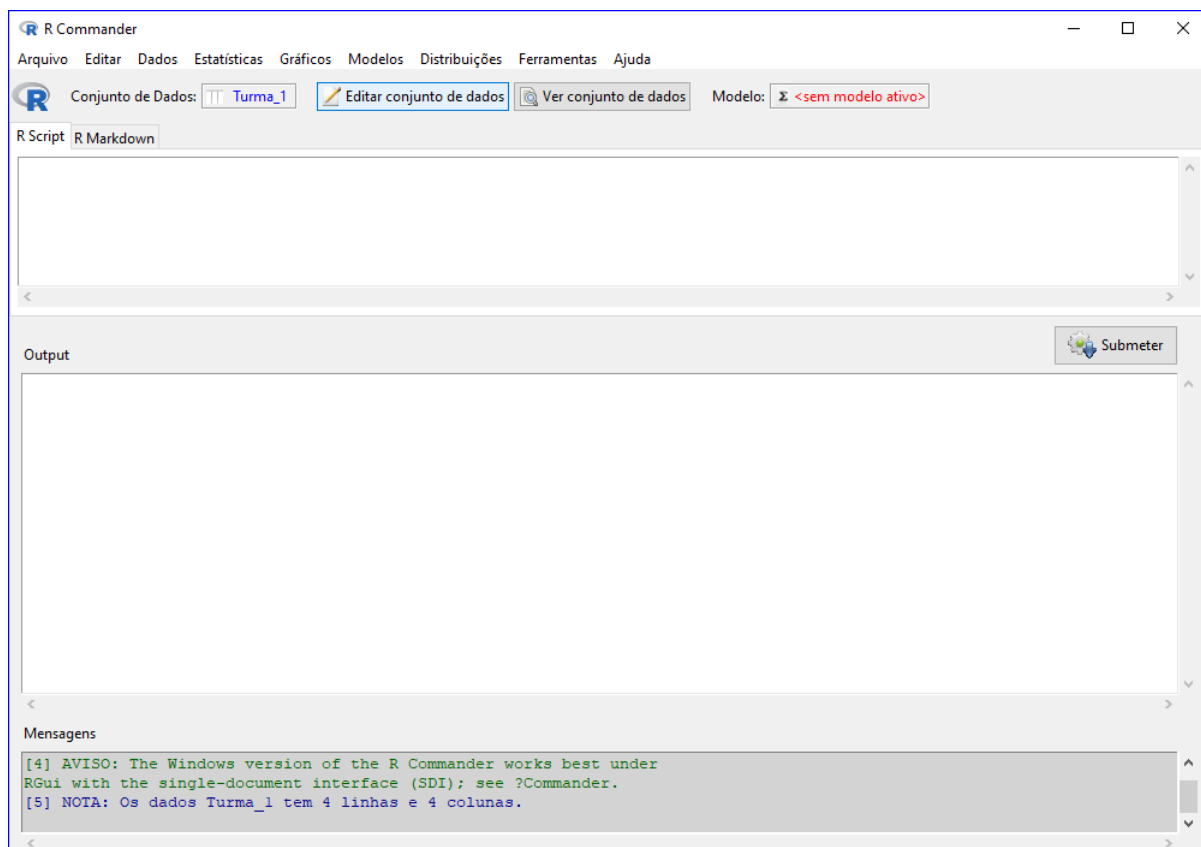





Figura 3: Apresentando as janelas *R Script*, *Output* e *Mensagens*.

	Nome	Sexo	Nota	Faltas
1	André	M	8.5	0
2	Carla	F	7	3
3	Fernando	M	4.5	4
4	Larissa	F	9	2

Figura 4: Visualizando conjunto de dados no R Commander.

**Observação 7.** Uma linha de comando na janela *R Script* sempre começa colada à margem esquerda. Quando uma linha na janela *R Script* não começa colada à margem esquerda, isto significa que ela é continuação da linha imediatamente acima, e não uma nova linha de comando. Ainda, é possível alterar e executar (ou apenas executar) qualquer linha de comando exibida na janela *R Script*: basta selecionar toda a linha de comando desejada com o *mouse* (ou com as setas do teclado, mantendo a tecla  pressionada) e clicar em *Submeter* (ou utilizar o atalho do teclado  + ). As Subseções 3.7 e 4.6 utilizam este artifício para contornar limitações dos menus do R Commander.

## 2.2 Importando conjunto de dados com formatos *.xls* ou *.xlsx*

É comum trabalharmos com conjuntos de dados digitados em uma planilha em um arquivo Excel (formato *.xls* ou *.xlsx*). Para que o R consiga “entender” este conjunto de dados, é necessário importá-lo com o R Commander. Para tal, façamos o seguinte:

1. Na barra de ferramentas da janela *R Commander*, vá em **Dados → Importar arquivos de dados → do arquivo Excel**.
2. Dê um nome para o seu conjunto de dados (sem utilizar espaço). No campo *Símbolo p/ dados faltantes*, é aconselhável substituir `<casela vazia>` por `NA`. Dê OK.
3. Procure o arquivo onde está o conjunto de dados e dê duplo clique nele.
4. Uma pequena janela com o nome das planilhas presentes no arquivo é aberta. Selecione o nome da planilha na qual está o conjunto de dados e clique em OK.

**Exercício 1.** Clique em [github.com/feliperafaelrm/Rcmdr/raw/main/Turma2.xls](https://github.com/feliperafaelrm/Rcmdr/raw/main/Turma2.xls) para baixar o arquivo *Turma2.xls* e, no R Commander, importe o conjunto de dados presente na planilha *Plan1* deste arquivo. Dê ao conjunto de dados importado o nome *Turma\_2*.

**Observação 8.** Feito o Exercício 1, passamos a ter dois conjuntos de dados carregados: *Turma\_1* e *Turma\_2*, este último tomando a posição de conjunto de dados ativo. Para alternar entre conjuntos de dados já carregados, basta clicar em cima do nome do conjunto de dados em azul (logo abaixo da barra de ferramentas) e escolher o conjunto de dados com o qual se deseja trabalhar (ou seja, qual desejamos tornar o conjunto de dados ativo).

**Observação 9.** A edição de dados também pode ser feita em conjuntos de dados importados (e não apenas com conjuntos criados diretamente no R Commander), bem como a sua visualização.

## 2.3 Importando conjunto de dados com formato *.txt* ou *.csv*

Não é raro encontrarmos conjuntos de dados salvos no formato *.txt* ou no formato *.csv*. Apesar de programas como o MS Excel terem suporte para abrir arquivos com estas extensões, a forma de importação para o R é diferente daquela explicada na Subseção 2.2. Segue o passo a passo.

1. Na barra de ferramentas da janela *R Commander*, vá em **Dados → Importar arquivos de dados → de arquivo texto, clipboard ou URL**. Uma janela tal qual a Figura 5 é aberta.
2. Dê um nome para o seu conjunto de dados (não utilize espaço).
3. Recomendável manter marcadas as caixinhas *Nome das variáveis no arquivo* e *Convert character variables to factors*.
4. Símbolo para dados faltantes: mantenha NA (recomendado).
5. Localização do Arquivo de dados: mantenha em *Sistema de Arquivos Local*.
6. Para preencher corretamente os campos *Separador de campos* e *Separador de decimais*, abra o arquivo que contém o conjunto de dados com o Bloco de Notas e verifique qual o símbolo utilizado para separar os campos e qual símbolo utilizado para separador de casa decimal. Após a verificação, feche-o.
7. Dê OK e procure o arquivo para concluir a importação.

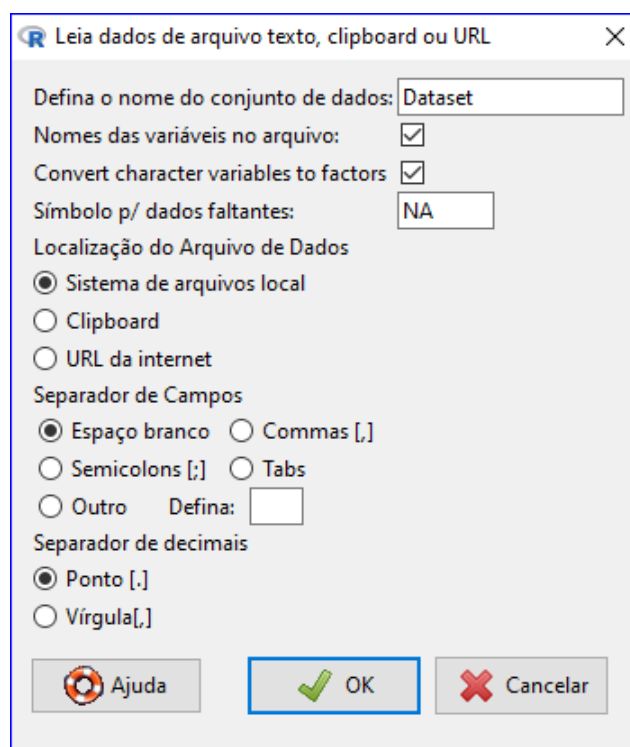



Figura 5: Importação de conjunto de dados em formato *.txt* ou *.csv*.

**Exercício 2.** Clique em [github.com/feliperafaelrm/Rcmdr/raw/main/Turma3.txt](https://github.com/feliperafaelrm/Rcmdr/raw/main/Turma3.txt) para baixar o arquivo *Turma3.txt* e, no R Commander, importe o conjunto de dados presente neste arquivo. Dê ao conjunto de dados importado o nome *Turma\_3*. Neste arquivo em particular, é utilizado ponto e vírgula (*Semicolons*) como separador de campos e vírgula como separador de casas decimais.

Ao concluir o Exercício 2, clique em *Ver conjunto de dados*. Caso os nomes acentuados (Fábio e Vítor) estejam com formatação alterada (por exemplo: F\xe1bio e V\xedtor), digite, na janela *R Console*,

```
options(encoding="latin1")
```

e aperte . Em seguida, retorne à janela *R Commander* e refaça o Exercício 2, sobrescrevendo o conjunto importado com os nomes alterados.

**Observação 10.** Uma forma de importar o conjunto de dados no Exercício 2 sem a necessidade de baixar previamente o arquivo *Turma3.txt* é marcar, no Passo 5, *URL da internet* em vez de *Sistema de Arquivos Local*. Desse modo, o Passo 7 requer a digitação da URL <https://github.com/feliperafaelrm/Rcmdr/raw/main/Turma3.txt> em vez da procura pelo arquivo baixado.

## 2.4 Exportando um conjunto de dados do R Commander

Assim como a importação de conjuntos de dados, o R Commander também possibilita a exportação de um conjunto de dados (para o formato *.txt*). O caminho é *Dados* → *Conjuntos de dados ativo* → *Exportar conjunto de dados ativo*. Feito isto, uma janela tal qual a Figura 6 é aberta. A primeira caixinha refere-se à inclusão do nome das variáveis (colunas) no arquivo exportado, e a segunda caixinha refere-se à inclusão do nome das linhas (coluna cinza à esquerda do conjunto de dados) no arquivo exportado. A terceira caixinha menciona a inclusão de aspas em torno dos atributos das variáveis qualitativas (o que pode ser útil ou não, a depender dos nossos propósitos).

**Exercício 3.** Selecione o conjunto de dados *Turma\_1* como conjunto de dados ativo e exporte-o, deixando marcada apenas a caixinha referente à inclusão do nome das variáveis. Sugestão: em *Separador de Campos*, selecione *Semicolons*.



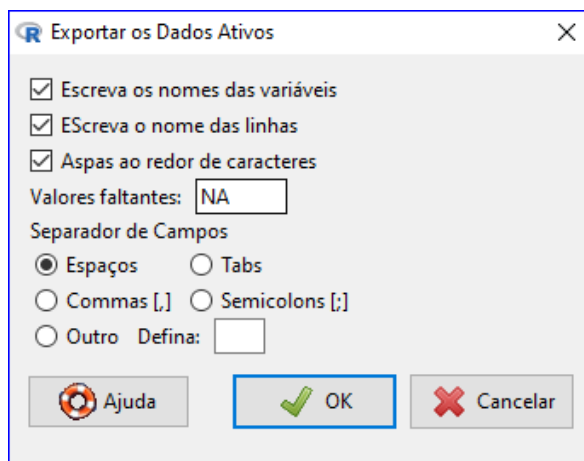


Figura 6: Exportando conjunto de dados.

## 2.5 Conjuntos de dados nativos do R Commander

Existem alguns conjuntos de dados que são disponibilizados ao carregarmos o pacote `Rcmdr`. Para acessá-los, vá em **Dados → Conjuntos de dados em pacotes → Ler dados de pacote “attachado”**. A título de ilustração, acesse este menu, dê um duplo clique no pacote `carData` e depois um duplo clique no conjunto de dados `Davis`. Clicando em OK<sup>3</sup>, ele é carregado. Visualize-o clicando em *Ver conjunto de dados*.

## 3 O menu *Dados*

Todos os procedimentos expostos na Seção 2 (digitação, importação e exportação de conjuntos de dados, bem como o carregamento de conjuntos de dados nativos) se deram por meio do menu **Dados**. Em particular, há mais duas funcionalidades abordadas na Seção 2 que também poderiam ser obtidas pelo por este menu:

- **Dados → Conjunto de dados ativo → View Data**: alternativa ao clique em *Ver conjunto de dados*, com a possibilidade de escolher apenas colunas específicas.
- **Dados → Conjunto de dados ativo → Selecionar conjunto de dados ativo**: alternativa à troca do atual conjunto de dados ativo por um outro já digitado, importado ou carregado.

---

<sup>3</sup>Antes de clicar em OK, pode ser interessante clicar na caixa *Ajuda no conjunto de dados selecionado*, pois dessa forma é aberta (no navegador de internet) uma breve explicação do conjunto de dados selecionado, incluindo o significado de suas variáveis, as quais são comumente escritas de forma abreviada.


Nesta seção, exploramos outros comandos interessantes do menu **Dados**.

### 3.1 Salvando conjuntos de dados

Para evitar a necessidade de digitar ou importar novamente um conjunto de dados, podemos salvá-lo no formato que o R reconhece (*.RData*). Primeiramente, clique no nome do conjunto de dados em azul (abaixo da barra de ferramentas) e escolha, como conjunto de dados ativo, aquele que se deseja salvar. Feito isto, basta acessar **Dados** → **Conjunto de dados ativo** → **Salvar conjunto de dados ativo** e salvá-lo com o nome de preferência para o arquivo (é recomendável manter o mesmo nome do conjunto de dados).

### 3.2 Carregando conjuntos de dados (no formato *.RData*)

Para carregar um conjunto de dados salvo no formato *.RData*, o caminho é **Dados** → **Carregar conjunto de dados ativo**.

Caso não estejamos com o R aberto e o conjunto de dados salvo (no formato *.RData*) estiver iconizado por uma letra R, é possível carregá-lo dando um duplo clique neste arquivo salvo. Dessa forma, uma sessão do R é aberta, e basta carregar o R Commander (digitando `require(Rcmdr)` ou `library(Rcmdr)` e apertando  logo após). O R Commander expressa, na cor vermelha, que “*Não há conjunto de dados ativo*”, mas basta clicar sobre esta expressão em vermelho para selecionar o conjunto de dados carregado como conjunto de dados ativo.

### 3.3 Agrupar dois conjuntos de dados

O R Commander possibilita que agrupemos dois conjuntos de dados por meio do menu **Dados** → **Merge de conjunto de dados**. Esse agrupamento pode ser por linhas (*Merge de linhas*) ou por colunas (*Merge de colunas*). O agrupamento por linhas é indicado quando temos dois conjuntos de dados com as mesmas variáveis e dispostas na mesma ordem, tal como nos conjuntos de dados *Turma\_1*, *Turma\_2* e *Turma\_3*. Já o agrupamento por colunas é indicado quando as observações são as mesmas e dispostas na mesma ordem em dois conjuntos de dados, porém em cada um deles são avaliadas variáveis diferentes.

**Exercício 4.** Agrupe o conjunto de dados *Turma\_2* com o conjunto de dados *Turma\_3* (sugestão de nome: *Turmas\_2\_e\_3*). Logo após: clique em *Ver conjunto de dados* para visualizar o conjunto de dados gerado por este procedimento; e acesse o menu **Dados → Conjunto de dados ativo → Salvar conjunto de dados ativo** para salvar este conjunto de dados.

### 3.4 Nomes das variáveis no conjunto de dados ativo

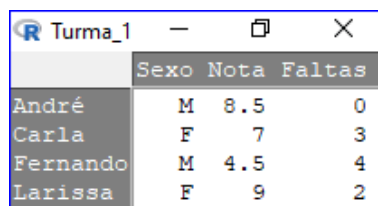
Ao criarmos, importarmos ou carregarmos um conjunto de dados, a janela *Mensagens* (localizada no rodapé da janela *R Commander*) relata o número de linhas e de colunas deste conjunto de dados. No formato “padrão” de conjunto de dados, cada coluna representa uma variável.<sup>4</sup> Para verificar o nome dessas variáveis (ou seja, o nome atribuído a cada coluna), vá em **Dados → Conjunto de dados ativo → Variáveis no conjunto de dados ativo**. Os nomes das variáveis aparecem na janela *Output* do R Commander.

### 3.5 Definir nomes das observações em cada linha

Nos conjuntos de dados que dispomos (*Turma\_1*, *Turma\_2*, *Turma\_3* e *Turmas\_2\_e\_3*), a primeira coluna não representa uma variável, mas sim uma identificação das observações (os alunos, identificados pelo primeiro nome) em cada linha. Todavia, o R Commander entende que toda coluna é uma variável (o que, a rigor, não chega a ser um grande problema). Uma forma da coluna *Nome* (em qualquer um destes conjuntos de dados) ser de fato compreendida pelo R Commander como identificação das observações em cada linha é acessar o menu **Dados → Conjunto de dados ativo → Definir nomes dos casos**, escolher *Nome* e clicar em OK. Feito isto (por exemplo, com *Turma\_1* como conjunto de dados ativo), clique no botão *Ver conjunto de dados* para verificar a diferença: os nomes dos alunos foram movidos para a coluna com fundo cinza à esquerda do conjunto de dados, que passa a ser vista apenas como identificadora de cada linha, ou seja, de cada aluno (compare as Figuras 4 e 7).

**Exercício 5.** Repita o procedimento acima com os outros três conjuntos de dados.

<sup>4</sup>A primeira coluna de um conjunto de dados pode estar associada simplesmente à identificação de cada observação, não se tratando portanto de uma variável propriamente dita.



	Sexo	Nota	Faltas
André	M	8.5	0
Carla	F	7	3
Fernando	M	4.5	4
Larissa	F	9	2

Figura 7: Entradas da coluna *Nome* como nomes dos casos no conjunto de dados *Turma\_1*.

**Observação 11.** Na Subseção 2.1, ignoramos a coluna *rowname* na digitação do conjunto de dados *Turma\_1*. Porém, se adicionássemos duas colunas ao invés de três colunas e digitássemos os nomes dos alunos na coluna *rowname* (e, nas colunas *V1*, *V2* e *V3*, as variáveis *Sexo*, *Nota* e *Faltas*, respectivamente), teríamos como saída o conjunto de dados exibido na Figura 7.

### 3.6 Subconjunto do conjunto de dados ativo

Quando queremos gerar um novo conjunto de dados apenas com as linhas que obedecem certa(s) condição(ões) em determinado conjunto de dados (por exemplo, no conjunto de dados *Turmas\_2\_e\_3*, considerar apenas indivíduos do sexo feminino), devemos acessar o menu **Dados** → **Conjunto de dados ativo** → **Definir sub-conjunto de dados ativo**. Com *Turmas\_2\_e\_3* selecionado como conjunto de dados ativo, acesse tal menu e, no campo *Expressão (subset expression)*, escreva `Sexo=="F"`. No campo *Nome para o novo conjunto de dados*, substitua a expressão padrão por um novo nome (sugestão: *Turmas\_2\_e\_3\_fem*) e clique em OK. O novo conjunto de dados gerado é um subconjunto de *Turmas\_2\_e\_3* contendo apenas as informações das meninas das Turmas 2 e 3.

**Exercício 6.** Considere que um prêmio foi oferecido para cada aluno(a) do conjunto de dados *Turmas\_2\_e\_3* com nota maior ou igual a 9,0 e, no máximo, 2 faltas. Defina um subconjunto de *Turmas\_2\_e\_3* que contenha apenas as informações dos alunos premiados, ou seja, repita o procedimento acima, usando a expressão `Nota >= 9 & Faltas <= 2` em vez de `Sexo=="F"` (sugestão de nome para este novo conjunto de dados: *Turmas\_2\_e\_3\_premiados*). Após clicar em OK, clique em *Ver conjunto de dados* para visualizar quais foram os alunos premiados.

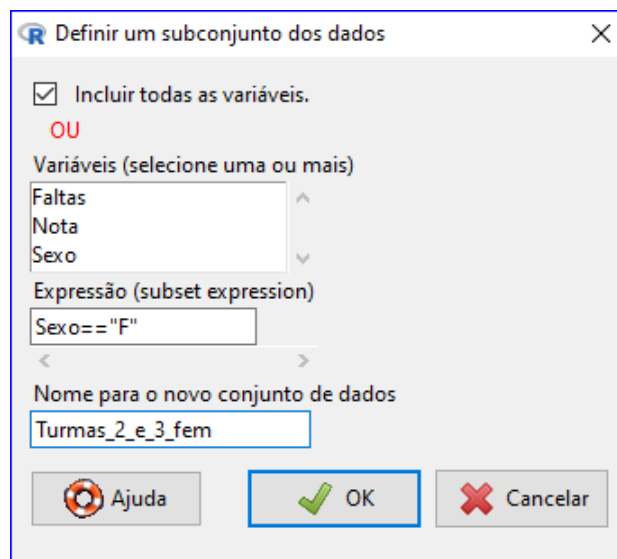


Figura 8: Definindo subconjunto do conjunto de dados *Turmas\_2\_e\_3* apenas com as meninas.

**Observação 12.** Para qualquer subconjunto de dados obtido pelo menu *Dados* → *Conjunto de dados ativo* → *Definir sub-conjunto de dados ativo*, é recomendável acessar o menu *Dados* → *Modificação de variáveis no conjunto de dados* → *Abandonar fatores não usados* (com o subconjunto em questão selecionado como conjunto de dados ativo), marcar a caixinha *todos os fatores* e dar OK. Para o conjunto de dados *Turmas\_2\_e\_3\_fem*, por exemplo, não executar este procedimento faz com que o sexo masculino siga presente no gráfico de barras e na distribuição de frequências da variável *Sexo*, explicitando uma desnecessária frequência igual a zero.

### 3.7 Ordenar linhas do conjunto de dados

Em alguns conjuntos de dados, pode ser interessante organizar as linhas em ordem crescente ou decrescente de acordo com uma coluna (variável) de referência. Por exemplo, podemos ordenar as linhas de qualquer um dos conjuntos de dados que dispomos conforme a nota do(a) aluno(a) (da maior nota para a menor nota, ou vice-versa). Para isto, primeiramente defina o conjunto de dados de interesse como conjunto de dados ativo (a título de ilustração, escolha o conjunto de dados *Turma\_1*) e, logo após, acesse o menu *Dados* → *Conjunto de dados ativo* → *Sort active dataset*. Clique em *Nota* e marque a opção *Decreasing* para que as linhas sejam organizadas por ordem decrescente

de nota, tal como ilustra a Figura 9. No campo *Nome para o novo conjunto de dados*, é possível digitar outro nome para este “novo” conjunto de dados com a ordenação desejada, de forma que a ordem original do conjunto de dados *Turma\_1* não seja alterada. Por comodidade, não altere este campo e clique em OK. Na próxima janela, confirme que deseja sobrescrever o conjunto de dados original clicando em *Sim*. Logo após, clique no botão *Ver conjunto de dados* e observe que as linhas de *Turma\_1* estão organizadas por ordem decrescente de notas, e não mais na ordem alfabética dos nomes dos alunos, conforme digitado na Subseção 2.1.

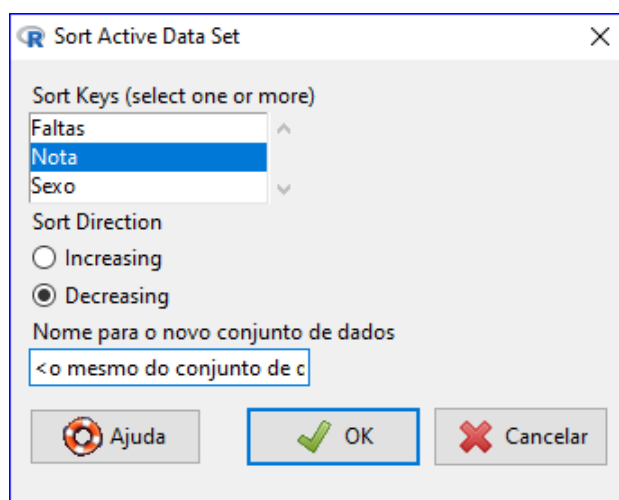


Figura 9: Ordenando linhas do Conjunto de dados *Turma\_1* pela nota dos alunos (da maior par a menor).

No caso de duas notas semelhantes (o que não ocorre no conjunto de dados *Turma\_1*, mas ocorre no conjunto de dados *Turmas\_2\_e\_3*), poderíamos estar interessados em usar o número de faltas como um “critério de desempate” para a ordenação. Isto também é possível por meio do mesmo menu acima: note que, na janela exibida na Figura 9, é possível escolher uma ou mais variáveis. No caso da escolha de pelo menos duas variáveis, uma nova janela é exibida após clicarmos em OK, para definir qual o primeiro critério de ordenação (indicado pelo número 1), qual o segundo critério de ordenação (indicado pelo número 2), e assim sucessivamente.

Para ilustrar a situação acima, selecione o conjunto de dados *Turmas\_2\_e\_3* como conjunto de dados ativo e acesse o menu **Dados → Conjunto de dados ativo → Sort active dataset**. Marque as variáveis *Nota* e *Faltas*, bem como a opção *Decreasing* e dê OK (novamente clique em *Sim* na janela com o aviso de sobrescrever conjunto de

dados). Na janela seguinte, digite 1 para *Nota* e 2 para *Faltas*, como mostra a Figura 10, e dê OK. Clicando em *Ver conjunto de dados*, observe que há dois alunos com nota 6,5, porém primeiramente é listado(a) o(a) aluno(a) com maior número de faltas (lembre que foi marcada a opção *Decreasing*, ou seja, ordem decrescente).

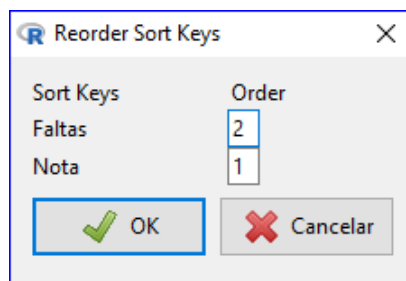


Figura 10: Ordenando linhas do conjunto de dados *Turmas\_2\_e\_3* pela nota dos alunos e, em seguida, pelo número de faltas dos alunos, ambos em ordem decrescente.

**Observação 13.** Um ponto negativo deste menu é que todos os critérios são ordenados no mesmo sentido (todos crescentes ou todos decrescentes). Há uma forma de contornar isto por meio de edição da linha de comando gerada por esta ordenação: por exemplo, para que o segundo critério (*Faltas*) obedeça ordem crescente, substitua `decreasing=TRUE` por `decreasing=c(TRUE,FALSE)` na janela *R Script*, de modo que fique desta forma:

```
Turmas_2_e_3 <- with(Turmas_2_e_3, Turmas_2_e_3[order(Nota, Faltas,
  decreasing=c(TRUE,FALSE)), ])
```

Marque toda esta linha de comando na janela *R Script* com o *mouse* e clique no botão *Submeter*, que fica entre as janelas *R Script* e *Output*, no lado direito.

**Observação 14.** Ainda considerando possibilidades de ordenação das linhas de um conjunto de dados via edição de linha de comando, é possível ordenar as linhas de um conjunto de dados pelos nomes dos casos. Por exemplo, para ordenar as linhas do conjunto de dados *Turmas\_2\_e\_3* conforme os nomes dos alunos (em ordem alfabética), a linha de comando deve ser do tipo

```
Turmas_2_e_3 <- with(Turmas_2_e_3,
  Turmas_2_e_3[order(rownames(Turmas_2_e_3)), ])
```

Marcando toda esta linha de comando na janela *R Script* com o *mouse* e clicando em seguida em *Submeter*, o resultado é alcançado.

### 3.8 Modificação de variáveis no conjunto de dados

Esta subseção é dedicada às principais funcionalidades fornecidas pelo menu **Dados** → **Modificação de variáveis no conjunto de dados**. Primeiramente, clique em [github.com/feliperafaelrm/Rcmdr/raw/main/Bussab.xlsx](https://github.com/feliperafaelrm/Rcmdr/raw/main/Bussab.xlsx) para baixar o arquivo **Bussab.xlsx** e, em seguida, importar o conjunto de dados presente na planilha **Dados** deste arquivo Excel, tal como orientado na Subseção 2.2. O nome desse arquivo é uma alusão a (Bussab; Morettin, 2017), uma vez que o conteúdo da planilha **Dados** foi extraído do Capítulo 2 deste livro. Este conjunto de dados fictícios traz 36 linhas, cada uma associada a um empregado de uma seção de orçamentos, e registra, para cada um deles: grau de instrução; renda; se é ou não casado; idade, em anos completos; região de procedência; e número de filhos (dos indivíduos casados). Sugerimos que esse conjunto de dados importado, ao qual nos referenciamos como **Bussab**, seja salvo ao final de cada uma das próximas subseções da Seção 3, no menu **Dados** → **Conjunto de dados ativo** → **Salvar conjunto de dados ativo**, uma vez que cada uma destas subseções sugere modificar ou adicionar variáveis no conjunto de dados **Bussab**.

#### 3.8.1 Converter variável quantitativa (numérica) para variável qualitativa

No conjunto de dados **Bussab**, note que as variáveis **Casado** e **Instr** estão representadas por rótulos em forma de números, apesar da natureza qualitativa de ambas. Para que estas variáveis sejam vistas pelo R Commander como variáveis qualitativas, é necessário converter estes números em **categorias** (atributos), ou seja, transformar uma variável (vista pelo R como) quantitativa em qualitativa.

Acesse o menu **Dados** → **Modificação de variáveis no conjunto de dados** → **Converter variável numérica para fator**. Selecione a variável de interesse (vamos começar pela variável **Casado**). Em **Níveis dos fatores**, selecione **Defina nome dos níveis**, e dê um nome para esta nova variável (sugestão: **Casado\_fator**, ver Figura 11). Ao clicar em OK, uma janela é aberta para que seja escrito o nome do atributo para cada rótulo numérico. Escreva: *Não* ao lado de 0; e *Sim* ao lado de 1. Clique em OK para concluir esta conversão. A nova variável criada (**Casado\_fator**) pode ser visualizada clicando em **Ver conjunto de dados**, alocada na última coluna do conjunto de dados.

**Observação 15.** É possível manter os rótulos como números (optando por *Use números* em vez de *Defina nome dos níveis* no campo **Níveis dos fatores**). A princípio, podemos



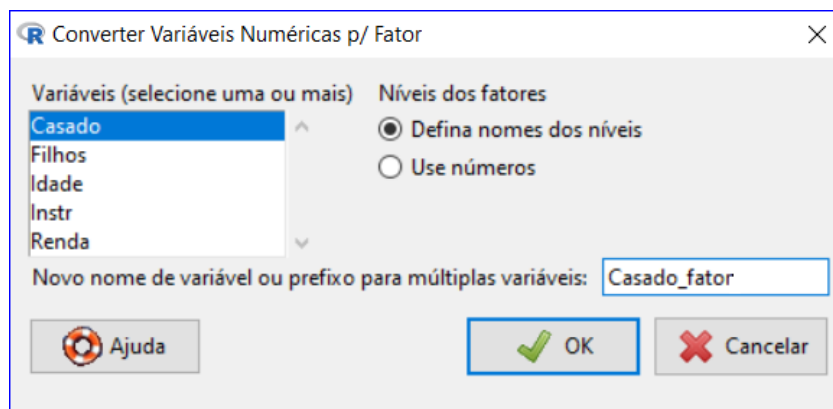


Figura 11: Convertendo a variável de rótulos numéricos *Casado* para a variável qualitativa *Casado\_fator* no conjunto de dados *Bussab*.

pensar que isto não faz diferença. Mas, a rigor, a variável convertida passa a se comportar como variável qualitativa (e não mais como quantitativa), apesar dos rótulos numéricos.

**Observação 16.** Neste processo de conversão, podemos manter o nome da variável convertida como o mesmo da variável original. Entretanto, essa escolha faz com que a “nova” variável criada sobrescreva a variável original, e desta forma **não conseguimos recuperar a variável na sua forma original** (a não ser que o conjunto de dados seja novamente digitado, importado ou carregado).

**Exercício 7.** Crie uma nova variável no conjunto de dados *Bussab* chamada *Instrução* que converta os rótulos numéricos 0, 1 e 2 da variável *Instr* em seus respectivos significados: Fundamental, Médio e Superior.

### 3.8.2 Reordenar níveis dos fatores

Quando o R lida com variáveis naturalmente qualitativas (isto é, variáveis qualitativas que não foram originadas por outra variável), a ordenação dos atributos é feita em ordem alfabética. Por exemplo, um gráfico de barras da variável qualitativa *Proced* posiciona as barras associadas às frequências das resposta *Capital*, *Interior* e *Outra* exatamente nesta ordem, da esquerda para a direita.

No caso da variável *Casado\_fator*, temos um exemplo de variável qualitativa que foi criada baseada em outra variável: *Casado*. Como esta última é tomada pelo R como quantitativa (numérica) e assume apenas os valores 0 e 1, o atributo de *Casado\_fator*

equivalente ao 0 é seu primeiro atributo, enquanto o atributo de *Casado\_fator* equivalente ao 1 é seu segundo atributo. Portanto, a ordenação dos atributos de *Casado\_fator* é: *Não* e *Sim*. Contudo, é natural que tabelas e gráficos desta variável adotem a ordem inversa, ou seja, posicionem a resposta *Sim* antes/acima/à esquerda da resposta *Não*. Sendo assim, é conveniente reordenar estes atributos antes de qualquer análise gráfica ou tabular.

Para fazer a reordenação dos níveis de uma variável qualitativa (seja nominal ou ordinal), acesse o menu **Dados → Modificação de variáveis no conjunto de dados → Reordenar níveis dos fatores**. Feito isto, selecione a variável *Casado\_fator* e clique em OK. Como o campo *Nome do fator* não foi alterado, uma pequena janela com as opções *Sim* e *Não* é aberta. Clique em *Sim* para confirmar que a “nova” coluna com a ordenação desejada sobrescreva a coluna com a ordenação original. Em seguida, digite 1 para *Sim* e 2 para *Não* no campo *Nova ordem*, invertendo a ordenação original, e clique em OK. Aparentemente, nada mudou; porém, a partir dessa modificação, a ordem dos atributos de *Casado\_fator* está configurada como *Sim* para primeiro atributo e *Não* para segundo atributo em qualquer análise que vier a ser feita para esta variável.

**Observação 17.** A variável *Instrução* criada no Exercício 7, tal como a variável *Casado\_fator*, foi criada baseada em uma variável tomada pelo R como quantitativa. Os atributos *Fundamental*, *Médio* e *Superior* de *Instrução* são equivalentes aos valores 0, 1 e 2 da variável *Instr*, respectivamente. Portanto, a ordenação tomada pelo R para os atributos da variável *Instrução* é: *Fundamental*, *Médio* e *Superior* (coincidentemente, estão em ordem alfabética).

### 3.8.3 Agrupar em classes uma variável quantitativa

Quando lidamos com uma variável quantitativa contínua (ou uma variável quantitativa discreta com muitos valores distintos), pode ser conveniente resumir este conjunto de valores em **classes**. Para fazer isto pelo do R Commander, vá em **Dados → Modificação de variáveis no conjunto de dados → Agrupar em classes uma variável numérica (para criar fator)**. Uma janela conforme exibida na Figura 12 é aberta (lembrando que estamos utilizando *Bussab* como conjunto de dados ativo).

A título de ilustração, acesse o menu acima destacado e escolha a variável *Renda*. Dê um nome para a “variável agrupada” a ser criada no campo *Novo nome de variável* (sugestão: *Renda\_classes*). Defina o número de classes (sugestão: 4 classes) e opte pela

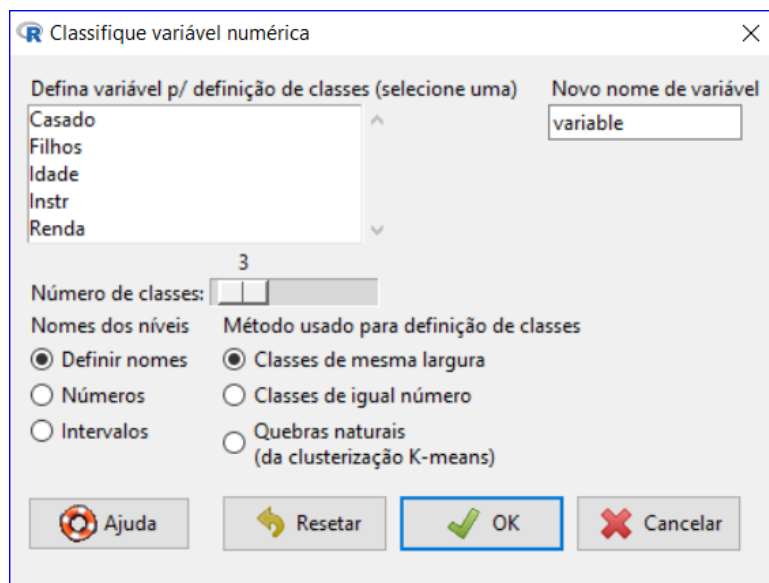


Figura 12: Agrupando variável quantitativa em classes.

forma com a qual cada classe é rotulada em *Nome dos níveis*. Esses rótulos podem ser nomes, números ou intervalos (sugestão: marcar *Intervalos*). No campo *Método usado para definição de classes*, mantenha em *Classes de mesma largura*. Clique em OK para concluir este agrupamento por classes e clique em *Ver conjunto de dados* para visualizar esta nova variável criada na última coluna do conjunto de dados. Apesar de termos optado por classes de mesma largura, as classes geradas possuem amplitudes ligeiramente diferentes.

**Exercício 8.** Repita o agrupamento da variável *Renda* feito acima, porém selecionando *Definir nomes* no campo *Nomes dos níveis* (pode sobrescrever a variável criada acima, isto é, continuar usando o nome *Renda\_classes*). Sugestões de nomes para as classes: *Baixa*, *Média*, *Alta* e *Muito alta*.

### 3.8.4 Computar nova variável

No conjunto de dados *Bussab*, a variável *Renda* está expressa em salários mínimos. Uma vez que se conheça o valor de um salário mínimo, pode ser interessante adicionar uma coluna neste conjunto de dados que expresse a renda em reais. Tomando por base o valor do salário mínimo no ano de 2024 (R\$1412), temos

$$\text{Renda em reais} = \text{Renda em salários mínimos} \times 1412.$$

O R Commander possibilita fazer funções de variáveis, tal como acima, por meio do menu *Dados → Modificação de variáveis no conjunto de dados → Computar nova variável*. Preencha a janela *Compute Nova Variável* conforme a Figura 13 (duplo clique no nome da variável faz surgir seu nome no campo *Expressão p/ computar*). Use asterisco para sinal de multiplicação. Após dar OK, clique em *Ver conjunto de dados* para visualizar as rendas em reais na última coluna do conjunto de dados (parece que a sugestão de nomes para as classes no Exercício 8 não está muito realista).

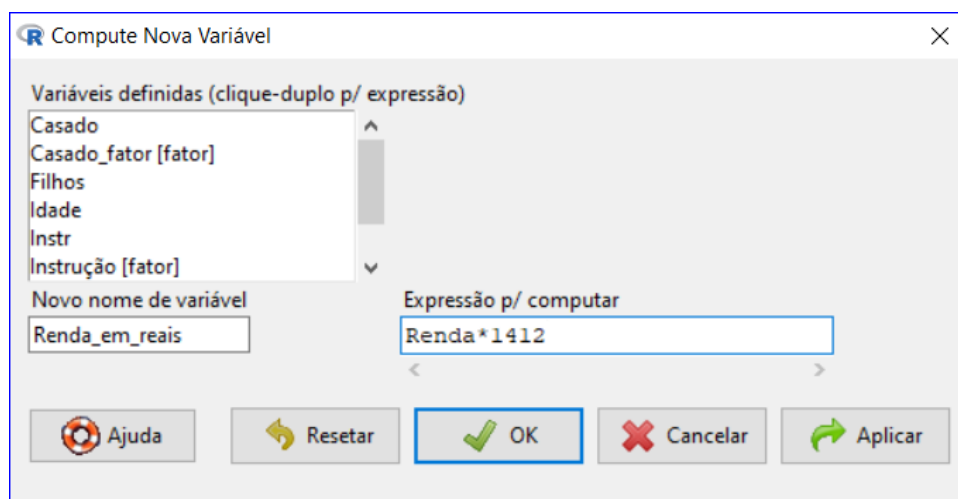




Figura 13: Computando nova variável no conjunto de dados *Bussab*: *Renda\_em\_reais*.

### 3.8.5 Renomear e apagar variáveis

Para **renomear** uma variável do conjunto de dados ativo, vá em *Dados → Modificação de variáveis no conjunto de dados → Renomear variáveis*. É possível selecionar mais de uma variável de uma só vez (basta segurar a tecla  e clicar nos nomes das variáveis que se deseja renomear).

Para **apagar** uma variável (ou seja, eliminar uma coluna) do conjunto de dados ativo, vá em *Dados → Modificação de variáveis no conjunto de dados → Apagar variáveis de um conjunto de dados*. É possível selecionar mais de uma variável de uma só vez (basta segurar a tecla  e clicar nos nomes das variáveis que se deseja remover do conjunto de dados).

**Exercício 9.** Apague a variável “equivocada” *Renda\_classes* do conjunto de dados *Bussab*.

### 3.8.6 Recodificar variáveis

Recodificar variável consiste na criação de uma nova variável no conjunto de dados baseada nos resultados da variável a ser recodificada. Nas Subseções 3.8.1 até 3.8.4, foram realizados casos particulares de recodificação de variáveis. Entretanto, o menu

*Dados → Modificação de variáveis no conjunto de dados  
→ Recodificar variáveis*

permite uma forma geral de recodificação. Suponha que queiramos incluir, no conjunto de dados *Bussab*, uma variável (qualitativa) chamada *Superior\_completo*, com atributos: *Sim*, se o indivíduo tem ensino superior; e *Não*, se o indivíduo não tem ensino superior. Para inserir esta variável, acesse o menu de recodificação de variáveis, preencha a janela conforme a Figura 14 e dê OK. As aspas em volta de cada nome faz com que o R de fato entenda-os como rótulos (atributos). Clique em *Ver conjunto de dados* para visualizar a nova variável criada na última coluna do conjunto de dados.

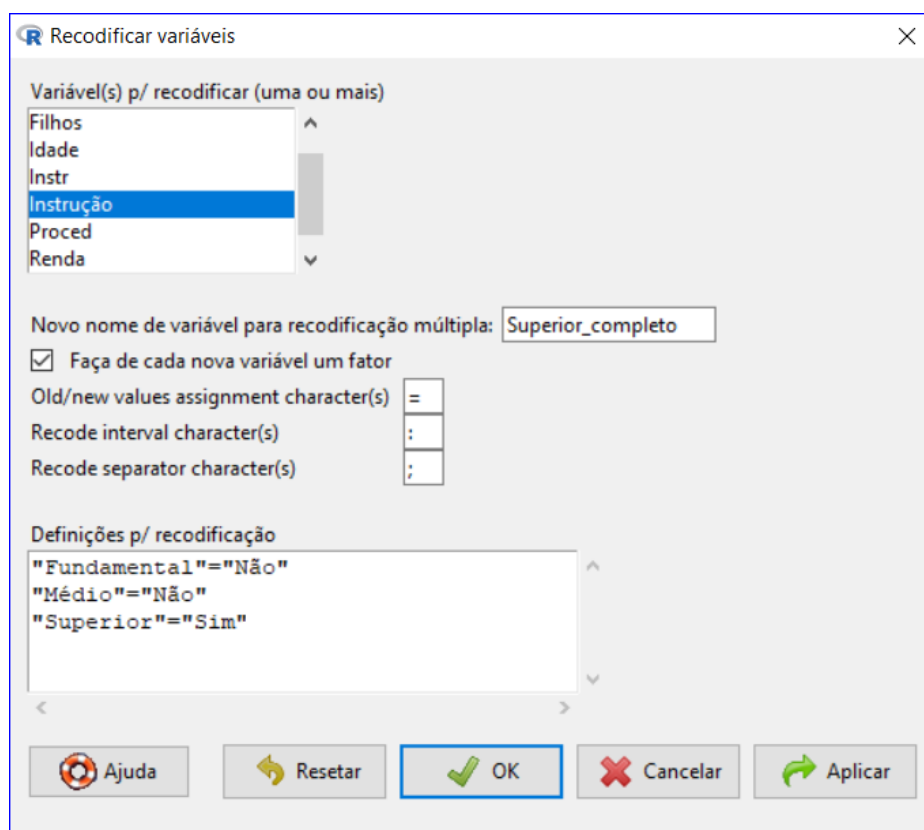


Figura 14: Criando a variável *Superior\_completo* no conjunto de dados *Bussab* por meio de recodificação da variável *Instrução*.

O menu de recodificação de variáveis não se restringe apenas a variáveis qualitativas. Também podemos recodificar variáveis quantitativas, bem como ter uma variável quantitativa como resultado de recodificação. A título de ilustração: selecione o conjunto de dados *Turmas\_2\_e\_3* como conjunto de dados ativo, acesse o menu de recodificação de variáveis e preencha a janela conforme a Figura 15. Os termos *lo* e *hi* significam *low* (inferior, abaixo de) e *high* (superior, acima de). Após clicar em OK, clique na caixa *Ver conjunto de dados* para visualizar a nova variável criada na última coluna do conjunto de dados.

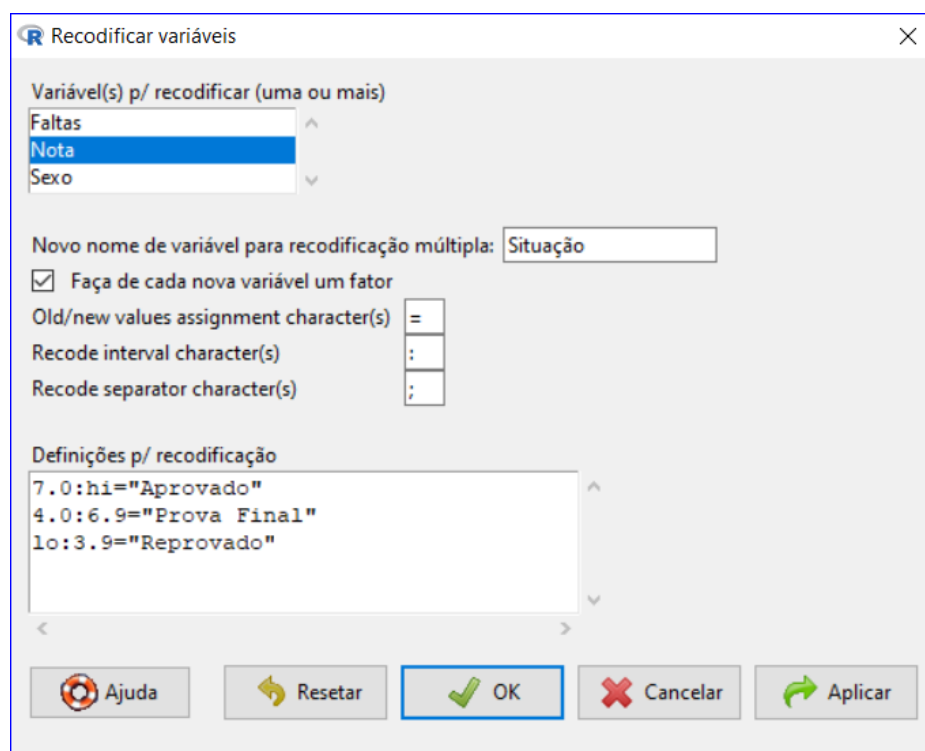


Figura 15: Criando a variável *Situação* no conjunto de dados *Turmas\_2\_e\_3* por meio de recodificação da variável *Notas*.

**Observação 18.** Quando o resultado da recodificação for de natureza quantitativa (ou seja, quando, no campo *Definições p/ recodificação*, são postos números ao invés de palavras à direita dos sinais de igual), é **imprescindível desmarcar a caixinha *Faça de cada nova variável um fator***. Além disso, tanto entradas numéricas como saídas numéricas não devem ser postas entre aspas, diferentemente do que ocorre com atributos de variáveis qualitativas. Observe que, tanto na Figura 14 como na Figura 15, as expressões à direita do sinal de igual estão entre aspas.

## 4 O menu *Gráficos*

Esta seção é dedicada à geração de gráficos de variáveis do conjunto de dados ativo. De posse desses gráficos, é possível fazer (ou começar a fazer) uma análise estatística das variáveis de interesse. Outros tipos de resultados, como tabelas de distribuição de frequências, tabelas de contingência e medidas resumo, seguem abordados na Seção 5.

Abordamos neste material apenas os gráficos mais interessantes para disciplinas básicas de Estatística, e adotamos o conjunto de dados *Bussab* como conjunto de dados ativo no decorrer de toda esta seção. É importante ressaltar que os gráficos feitos pelo R Commander surgem na janela *R Gui* (ao lado da janela *R Console*, em uma janela chamada *R Graphics*). Além disto, gráficos gerados em sequência sobrescrevem os anteriores. Ou seja, caso executemos dois gráficos em sequência, o segundo apaga o primeiro. Portanto, é importante **salvar** gráficos para não perdê-los. Tal procedimento é simples e detalhado na Subseção 4.10.

### 4.1 Gradiente de cores

Antes de abordarmos os gráficos propriamente ditos, vamos começar explorando o menu *Gráficos* → *Gradiente de cores (color palette)*. Ao clicar sobre qualquer um dos retângulos coloridos da janela oriunda deste menu (Figura 16), é possível configurar uma nova cor para tal retângulo, alterando assim a paleta de cores original ao clicar em OK. Mas para o que isto serve?

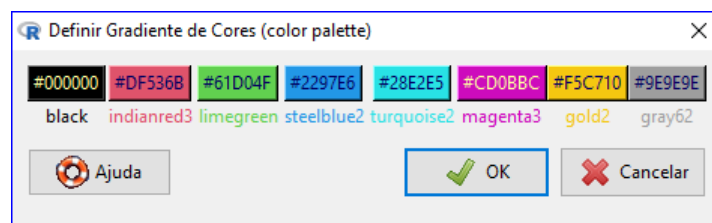


Figura 16: Gradiente de cores original do R Commander.

Alguns gráficos apresentados nesta seção podem ser coloridos segundo esta paleta de cores, esteja ela na sua forma padrão ou modificada. O primeiro retângulo, originalmente na cor preta, não é considerado. Logo, gráficos com apenas uma cor (como gráficos de barras simples) são coloridos na cor definida para o segundo retângulo, gráficos com duas

cores (como em gráficos de pizza para uma variável com apenas dois atributos) recebem as cores definidas para o segundo e o terceiro retângulo, e assim por diante.

## 4.2 Gráfico de setores (gráfico de pizza)

Uma representação gráfica frequentemente utilizada para variáveis qualitativas, sobretudo qualitativas nominais, é o gráfico de setores, popularmente conhecido como gráfico de pizza. Para gerá-lo no R Commander, o caminho é **Gráficos → Gráfico de Pizza**. Com o conjunto de dados *Bussab* selecionado como conjunto de dados ativo, acesse o menu acima. A janela de construção do gráfico de pizza é simples: além da escolha da variável, é possível editar o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico. Também é possível optar pela paleta de cores debatida na Subseção 4.1 selecionando *From color palette* no campo *Color selection*. Para gráficos de pizza, em particular, os rótulos dos eixos devem ficar em branco (isto ocorre mantendo a expressão <auto> em ambos). O gráfico de pizza da variável *Casado\_fator* ilustrado na Figura 17 foi obtido digitando

```
Indivíduos casados\nConjunto de dados Bussab
```

no lugar de <auto> no campo de *Título do gráfico* (a expressão \n implica quebra de linha), e mantendo o campo *Color selection* em *Default* (o R Commander usa, por padrão, uma paleta de cores especial para gráficos de pizza, oriunda do pacote *colorspace*). Ao lado de *Sim* e de *Não*, são exibidos, entre parênteses, os percentuais aproximados destas respostas para a variável *Casado\_fator*. Isto se deve ao fato da manutenção da opção *Percentages* no campo *Include in Segment Labels* na janela de construção do gráfico. Também é possível expressar as frequências absolutas em vez das porcentagens ou não exibir nenhuma destas informações, optando por *Contagens de frequências* ou *Neither* neste campo, respectivamente.

**Observação 19.** Ao repetir a geração do gráfico acima, porém optando por *From color palette* no campo *Color selection*, a cor associada ao *Sim* é a cor definida para o segundo retângulo da paleta de cores da Figura 16, e a cor associada ao *Não* é a cor definida para o terceiro retângulo desta paleta. Lembrando que, nesta variável, *Sim* vem “antes” de *Não* em virtude da reordenação realizada na Subseção 3.8.2.



**Indivíduos casados**  
**Conjunto de dados Bussab**

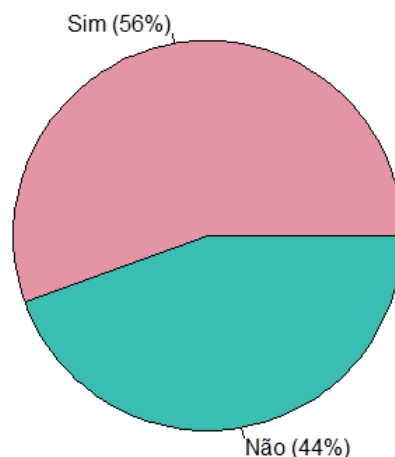


Figura 17: Gráfico de setores da variável *Casado\_fator*.

### 4.3 Gráfico de barras

O gráfico de barras é útil para variáveis qualitativas e também pode ser interessante para variáveis quantitativas discretas, desde que tenhamos poucos valores distintos desta variável no conjunto de dados em questão. O caminho para fazer este tipo de gráfico é **Gráficos → Gráfico de Barras**. A janela de construção possui duas abas: *Dados* e *Opções*. Na aba *Dados*, selecione a variável *Proced* (região de procedência). Na aba *Opções*, apenas os campos *Escala do eixo*, *Color selection* e *Legendas* são de interesse para gráficos de barras simples. Preenchendo esta aba conforme a Figura 18 e clicando em OK, o gráfico de barras gerado segue na Figura 19. O campo *rótulo do eixo-x* na Figura 18 está propositalmente em branco, pois caso contrário teríamos uma repetição desnecessária de informação (o título do gráfico já informa que a variável em questão é a região de procedência). Os números nas barras representam a frequência de cada atributo e se devem à caixinha *Show counts or percentages in bars* estar marcada ao clicar em OK na janela de construção do gráfico.

**Observação 20.** Caso optássemos por *From color palette* no campo *Color selection*, todas as barras seriam coloridas na cor definida para o segundo retângulo da paleta de cores do menu **Gráficos → Gradiente de cores (color palette)** (por padrão, uma tonalidade de vermelho indiano). Já o campo *Escala do eixo* permite gerar gráficos de

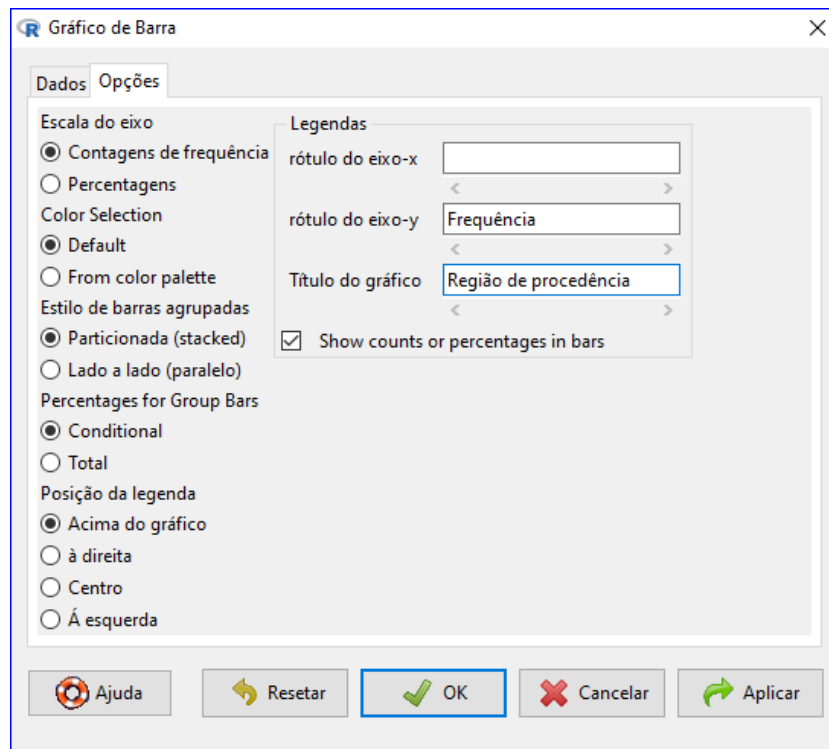


Figura 18: Construindo gráfico de barras para a variável *Proced.*

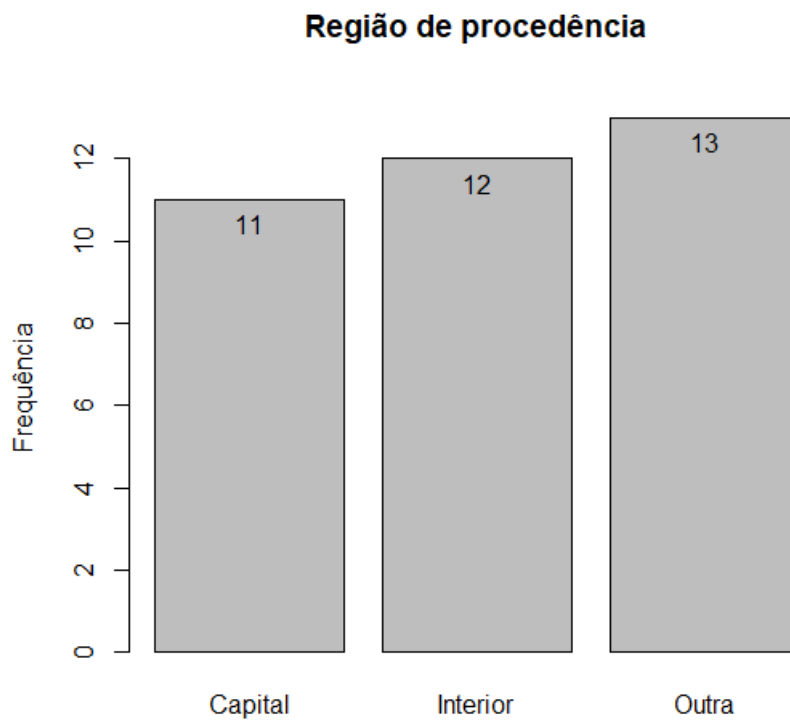


Figura 19: Gráfico de barras da variável *Proced.*

barras com porcentagens no lugar das frequências, e caso seja feita esta alteração, não faz sentido digitarmos *Frequência* para o rótulo do eixo-y; devemos usar, no seu lugar, termos como *Porcentagem*, *Percentual*, etc. Ainda, deixar marcada a caixinha *Show counts or percentages in bars* exibe percentuais em vez de frequências no interior de cada barra.

Note que o R Commander permite o traçado de gráfico de barras apenas para variáveis qualitativas. Se tivermos, no conjunto de dados de interesse, uma variável quantitativa discreta com poucos valores distintos e desejarmos fazer um gráfico de barras para ela, é necessário converter seus valores numéricos para atributos, tal como mostrado na Subseção 3.8.1 (mais precisamente, na Observação 15). Por exemplo, para gerarmos um gráfico de barras do número de filhos, é necessário criar uma nova variável no conjunto de dados *Bussab* por meio do menu **Dados → Modificação de variáveis no conjunto de dados → Converter variável numérica para fator**, optando pela variável *Filhos* e, por comodidade, em *Use números* no campo *Níveis dos fatores* (sugestão de nome para a nova variável: *Filhos\_fator*). O gráfico em questão segue na Figura 20.

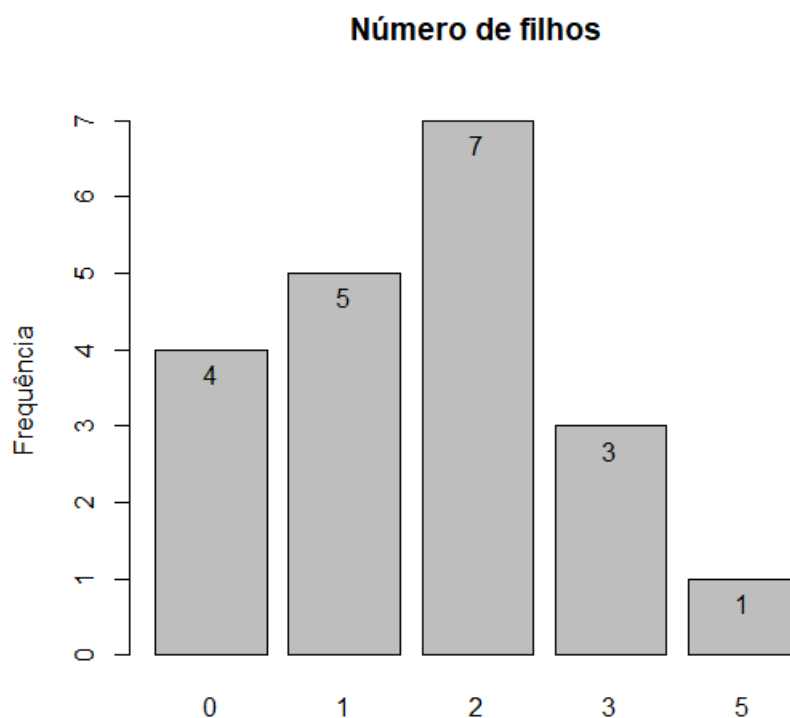


Figura 20: Gráfico de barras da variável *Filhos\_fator*.

Em alguns casos, representar variável quantitativa por um gráfico de barras seguindo este procedimento pode ser “visualmente traçoeiro”, pois os números são tratados como

meros rótulos. Perceba que, neste conjunto de dados em particular, nenhum indivíduo possui exatamente 4 filhos, porém a distância entre as barras de 3 filhos e de 5 filhos não difere da distância entre os outros pares de barras vizinhas, cuja diferença é de 1 filho, e não 2 filhos. Alternativas mais adequadas – e menos trabalhosas – ao gráfico de barras para variáveis quantitativas discretas com poucos valores distintos são o **gráfico de pontos** e o **gráfico de hastes**, abordados nas Subseções 4.5 e 4.6 deste material.

## 4.4 Gráfico de barras múltiplas

Além de gráficos de barras simples, também é possível gerar gráficos de barras múltiplas no menu *Gráficos* → *Gráfico de Barras*. Por exemplo, para “quebrarmos” cada barra do gráfico da Figura 19 (correspondentes a cada região de procedência) conforme o nível de instrução, retorne ao menu de construção de gráficos de barras, escolha a variável *Proced* e, logo abaixo, clique em *Gráfico por grupos*. Escolha a variável *Instrução* como variável de grupo e dê OK. Por fim, preencha a aba *Opções* da janela de construção do gráfico conforme a Figura 21 e clique em OK para gerar o gráfico de barras múltiplas exibido na Figura 22. A legenda está à esquerda, conforme definido no campo *Posição da legenda* na janela de construção do gráfico. Porém, para não correr o risco da legenda sobrepor uma ou mais barras de um gráfico de barras múltiplas, é recomendável selecionar a opção *Acima do gráfico* no campo *Posição da legenda*. Tal como nos gráficos de pizza, a opção *Default* no campo *Color Selection* utiliza a paleta de cores do pacote *colorspace*.

**Observação 21.** O R Commander trata o gráfico que fizemos acima como um gráfico de barras (múltiplas) da variável *Proced*, com a variável *Instrução* escolhida como variável de grupo. Entretanto, pode ser mais fácil entender o gráfico de barras múltiplas na Figura 22 como três gráficos de barras (simples) da variável *Instrução*: o primeiro (à esquerda) considera apenas os indivíduos cuja região de procedência é *Capital*; o segundo (no centro) considera apenas os indivíduos cuja região de procedência é *Interior*; e o terceiro (à direita) considera apenas os indivíduos cuja região de procedência é *Outra*.

Um gráfico de barras múltiplas permite visualizar graficamente a distribuição de frequências de uma variável qualitativa em diferentes grupos, fornecidos por outra variável qualitativa. No gráfico da Figura 22, é possível visualizar graficamente a distribuição de frequências do nível de instrução dos indivíduos, separadamente em cada região de

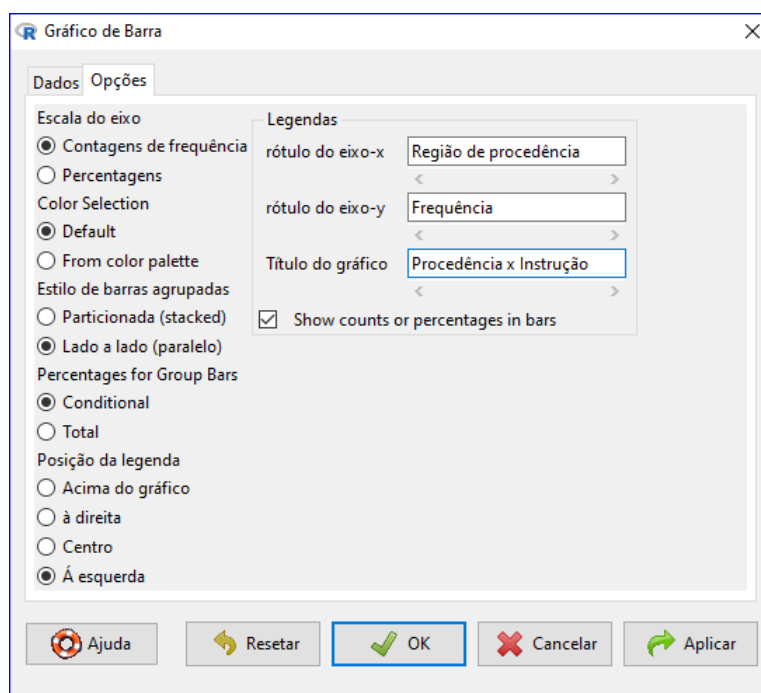


Figura 21: Construindo gráfico de barras múltiplas com as variáveis *Proced* e *Instrução*.

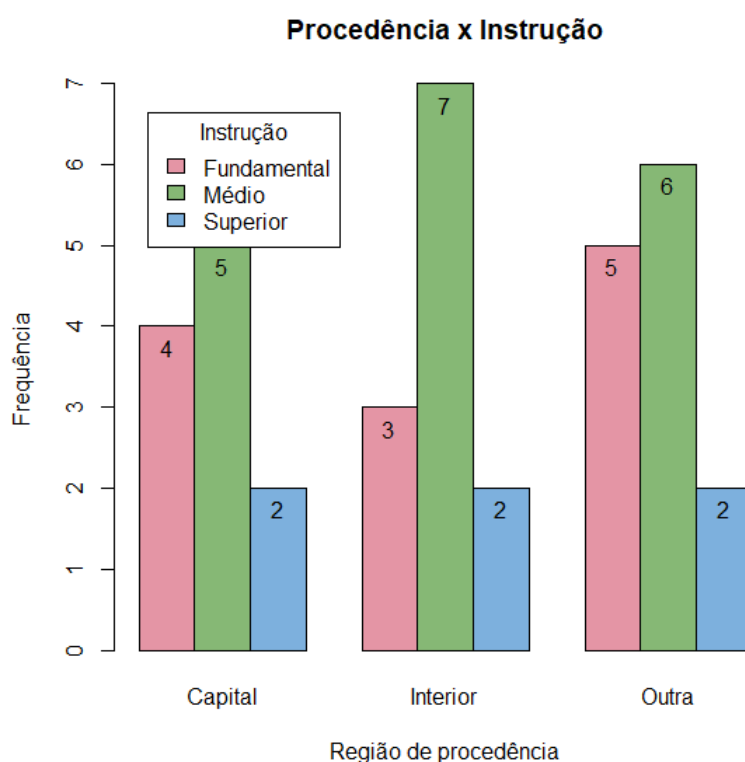


Figura 22: Gráfico de barras múltiplas com as variáveis *Proced* e *Instrução*.

procedência destes indivíduos. Entretanto, o número de indivíduos procedentes de cada região (*Capital*, *Interior* e *Outra*) é diferente, e assim torna-se razoável tomar o perfil percentual do nível de instrução em cada região de procedência quando o objetivo é comparar como se distribui o nível de instrução nas diferentes regiões de procedência dentro do conjunto de dados *Bussab*. Para realizar tal gráfico com porcentagens no lugar das frequências, tomando por base novamente a Figura 21, basta selecionar *Percentagens* no campo *Escala do eixo* (e para que o gráfico faça sentido, digite *Porcentagem* no lugar de *Frequência* no campo *rótulo do eixo-y*). O gráfico exibido na Figura 23 surge como fruto das alterações acima e também das alterações nos campos *Color selection* e *Posição da legenda* (de *À esquerda* para *Acima do gráfico*). Como as frequências nas regiões de procedência são próximas (11, 12 e 13), os gráficos ilustrados nas Figuras 22 e 23 apresentam formatos bem parecidos. A soma das porcentagens exatas em cada um dos três grupos é igual a 100%, uma vez que a opção *Conditional* foi mantida no campo *Percentages for Group Bars*. Porém, as porcentagens exibidas em cada barra são aproximadas. Logo, é possível que a soma das probabilidades aproximadas em algum(ns) grupo(s) retorne 99% ou 101%.

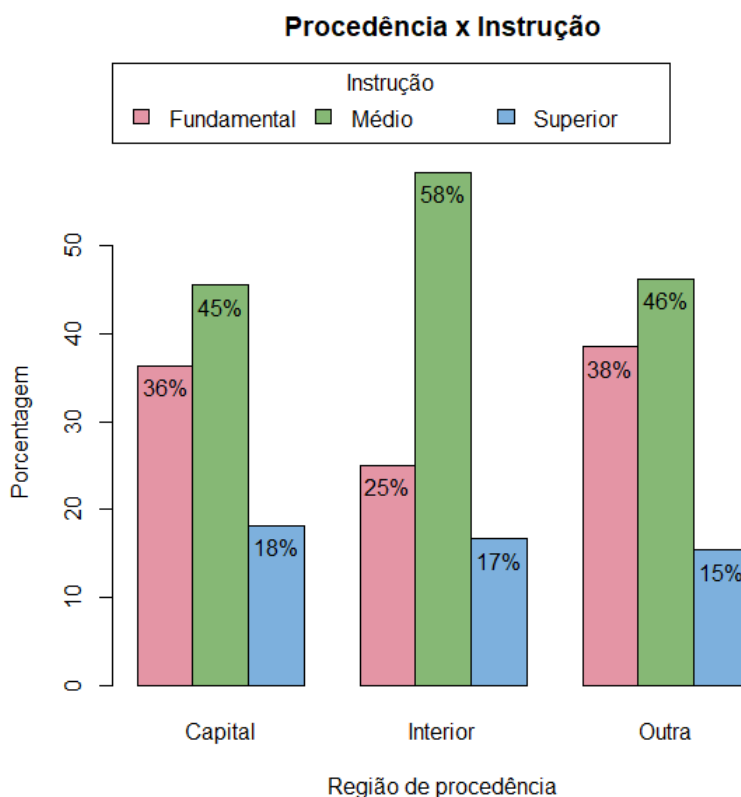


Figura 23: Perfil percentual do nível de instrução em cada região de procedência.

**Observação 22.** Caso optássemos por *Total* ao invés de *Conditional* no campo *Percentages for Group Bars*, a altura de cada barra retornaria o percentual aproximado do respectivo cruzamento de nível de instrução com região de procedência em todo o conjunto de dados, e não apenas dentro de um grupo específico. Consequentemente, a soma das porcentagens exatas nos nove cruzamentos retornaria 100%. Esta escolha não é adequada quando o intuito é comparar as distribuições percentuais entre grupos.

**Observação 23.** Outra modificação que pode ser feita no menu de construção de um gráfico de barras múltiplas é optar por *Particionada (stacked)* ao invés de *Lado a lado (paralelo)* no campo *Estilo de barras agrupadas*. Com esta escolha, as barras em cada grupo ficam “empilhadas”, o que pode ser visualmente interessante na comparação de perfis percentuais entre dois ou mais grupos.

**Observação 24.** Para gerar um gráfico de barras simples após gerar gráficos de barras múltiplas, é necessário clicar no botão *Resetar* na janela de construção de gráficos de barras.

## 4.5 Gráfico de pontos

Conforme comentado na Subseção 4.3, o gráfico de pontos é uma alternativa ao gráfico de barras para variáveis quantitativas discretas, e está disponível no menu **Gráficos** → **Gráfico de pontos**. A variável de interesse deve ser selecionada na aba *Dados* da janela de construção deste gráfico (ignore por enquanto a caixa *Gráfico por grupos*). Já na aba *Opções*, é possível editar o rótulo do eixo-x (nome desejado para a variável no gráfico). A Figura 24 exhibe o gráfico de pontos para a variável *Filhos* do conjunto de dados *Bussab*: na aba *Opções* da janela de construção do gráfico, basta escrever *Número de filhos* no campo *rótulo do eixo-x* e manter a caixinha *Variável de particionamento* desmarcada. Diferentemente do gráfico na Figura 20, a natureza numérica da variável foi respeitada.

É possível gerar gráficos de pontos (de uma mesma variável quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa deste mesmo conjunto de dados. Isto é interessante para compararmos o comportamento de uma mesma variável (quantitativa) em diferentes grupos dentro do conjunto de dados sob análise. Por exemplo, podemos gerar um gráfico de pontos das idades dos indivíduos casados e um gráfico de pontos das idades dos indivíduos que não são casados. Para tal, acesse o menu

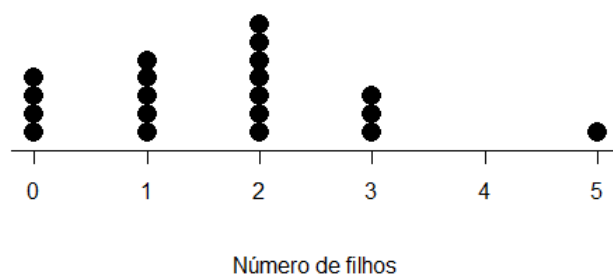


Figura 24: Gráfico de pontos da variável *Filhos*.

de construção do gráfico de pontos, escolha a variável *Idade* e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado\_fator* e dê OK nesta janela. Na aba *Opções*, escreva *Idade (em anos completos)* como rótulo do eixo-x e clique em OK. O gráfico gerado segue na Figura 25.

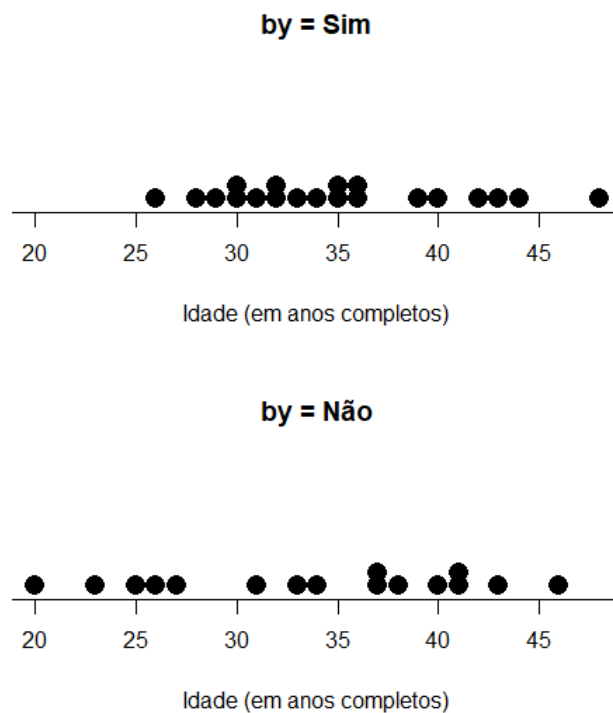


Figura 25: Gráficos de pontos da variável *Idade* para os casados e para os não casados.

**Observação 25.** Para gerar um gráfico de pontos sem divisão por grupos após gerar gráficos de pontos por grupos, é necessário clicar no botão *Resetar* na janela de construção deste gráfico.



## 4.6 Gráfico de hastes (*Plot discrete numeric variable*)

Alternativa mais elegante ao gráfico de pontos, o gráfico de hastes assemelha-se ao gráfico de barras, porém com hastes no lugar das barras. No R Commander, é possível gerar este tipo de gráfico no menu **Gráficos** → **Plot discrete numeric variable**. A variável de interesse deve ser selecionada na aba *Dados* da janela de construção deste gráfico (ignore por enquanto a caixa *Gráfico por grupos*). Já na aba *Opções*, é possível editar o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico, além do campo *Escala do eixo*, com opções de frequência ou de porcentagem para a altura das hastes. O gráfico de hastes da variável *Filhos* na Figura 26 foi construído deixando o rótulo do eixo-x em branco e escrevendo *Frequência* e *Número de filhos* no rótulo do eixo-y e em *Título do gráfico*, respectivamente. Assim como no gráfico de pontos, a natureza numérica da variável quantitativa no gráfico de hastes também é respeitada.

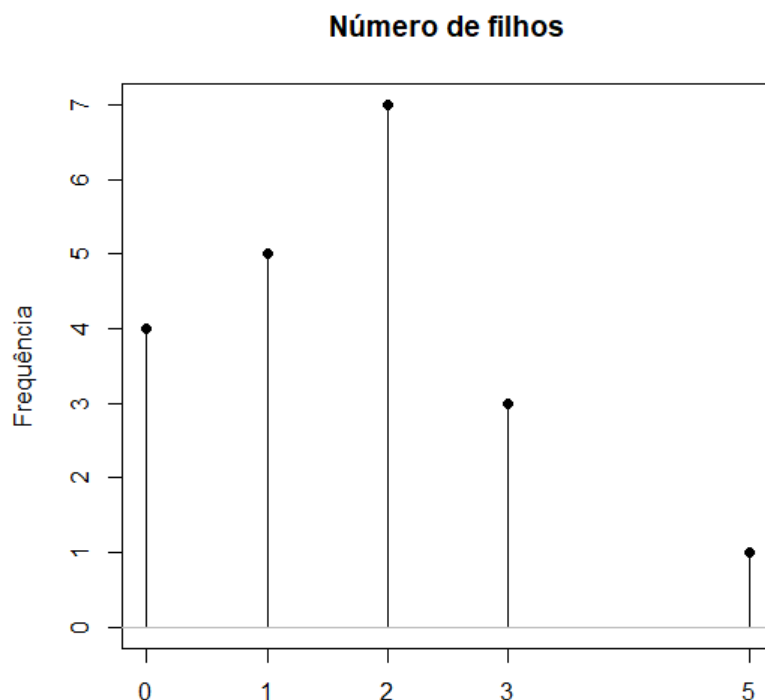


Figura 26: Gráfico de hastes da variável *Filhos*.

Assim como fizemos com gráficos de pontos, é possível gerar gráficos de hastes (de uma mesma variável quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa do mesmo conjunto de dados. Como este tipo de gráfico

permite trabalhar com escala percentual, ilustremos aqui a geração de um gráfico de hastes das idades dos indivíduos casados e um gráfico de hastes das idades dos indivíduos que não são casados, ambos na escala percentual. **Porém, temos um problema aqui!** Ao abrirmos o menu *Gráficos → Plot discrete numeric variable*, não vemos a variável *Idade* como opção de escolha, pois este menu não lista variáveis quantitativas com muitos valores distintos – a saber, os 36 indivíduos do conjunto de dados *Bussab* tem 24 idades diferentes. Uma forma de contornar isto é:

1. Fazer um gráfico de hastes escolhendo outra variável qualquer, preenchendo a aba *Opções* como se estivesse escolhendo a variável desejada;
2. Na linha de comando gerada na janela *R Script* pelo passo acima, substituir o nome da variável escolhida por *Idade*;
3. Selecionar toda esta linha de comando com o *mouse* e clicar em *Submeter*.

Portanto, acesse o menu de construção do gráfico de hastes, escolha uma das variáveis disponíveis (sugestão: *Instr*) e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado\_fator* e dê OK nesta janela. Já na aba *Opções*, selecione *Percentagens* no campo *Escala do eixo* e, no rótulo do eixo-y, escreva *Porcentagem*. A linha de comando gerada na janela *R Script* é da forma

```
with(Bussab, discretePlot(Instr, by=Casado_fator, scale="percent",
  xlab="", ylab="Porcentagem", main="Idade (em anos completos)"))
```

a qual deve ser modificada substituindo *Instr* por *Idade*, como segue abaixo:

```
with(Bussab, discretePlot(Idade, by=Casado_fator, scale="percent",
  xlab="", ylab="Porcentagem", main="Idade (em anos completos)"))
```

Marcando toda esta linha de comando na janela *R Script* com o *mouse* e, em seguida, clicando em *Submeter*, o gráfico desejado está pronto, conforme segue na Figura 27.

**Observação 26.** Para gerar um gráfico de hastes sem divisão por grupos após gerar gráficos de hastes por grupos, é necessário clicar no botão *Resetar* na janela de construção deste gráfico.

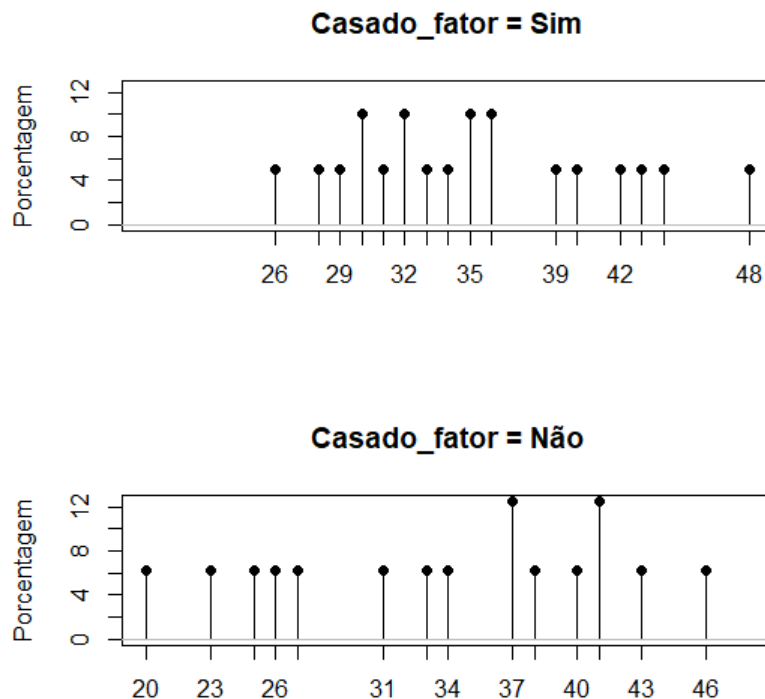


Figura 27: Gráficos de hastes da variável *Idade* para os casados e para os não casados.

## 4.7 Histograma

O histograma é um dos gráficos mais utilizados para verificar o comportamento de variáveis quantitativas (sobretudo quantitativas contínuas). Ao acessarmos o menu **Gráficos** → **Histograma**, somos direcionados a uma janela com duas abas (*Dados* e *Opções*), tal qual as janelas de construção de gráficos de barras, pontos e hastes.

Na aba *Dados*, escolha a variável cujo comportamento queremos verificar por meio de um histograma (ignore por enquanto a caixa *Gráfico por grupos*). Na aba *Opções*, podemos editar: o rótulo do eixo-x, o rótulo do eixo-y e o título do gráfico; o número de classes; e a escala de altura das barras como *Contagens de frequência*, *Percentagens* ou *Densidades*.

**Exercício 10.** Reproduza o histograma da Figura 28. Mantenha o número de classes em `<auto>`, e lembre que a expressão `\n` implica quebra de linha.

**Observação 27.** A Subseção 5.2 aborda uma forma de verificar precisamente em quais valores começa e termina cada classe no histograma, bem como suas respectivas frequências e percentagens.

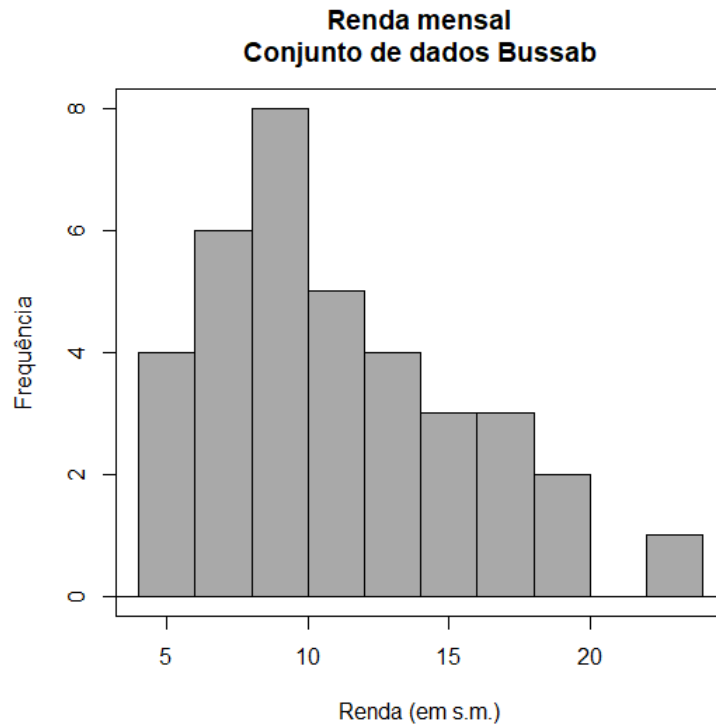


Figura 28: Histograma da variável *Renda*.

Na construção de histogramas pelo R Commander, é possível gerar histogramas de uma variável (quantitativa) de diferentes grupos do conjunto de dados de acordo com alguma variável qualitativa deste mesmo conjunto de dados, o que nos permite comparar o comportamento de uma mesma variável quantitativa em diferentes grupos dentro do conjunto de dados sob análise. Por exemplo, podemos gerar um histograma da renda dos indivíduos casados e um histograma da renda dos indivíduos que não são casados, preferencialmente em escala percentual, pois a frequência de indivíduos que são casados e a frequência de indivíduos que não são casados não são iguais. Para tal, acesse o menu de construção de histogramas, escolha a variável *Renda* e clique na caixa *Gráfico por grupos*. No campo *Variável de grupo*, selecione *Casado\_fator* e dê OK nesta janela. Na aba *Opções*, selecione *Percentagens* no campo *Escala do eixo*, digite *Porcentagem* (ou algo do tipo) em *rótulo do eixo-y* e, em *Título do gráfico*, deixe em branco ou digite algum título sem quebra de linhas (ou seja, sem “\n”), em virtude do espaço reduzido no topo da janela *R Graphics* - diferente do que ocorre na geração de gráficos de hastes por grupos, o título do gráfico em histogramas por grupos não é omitido. Clique em OK para obter o gráfico exibido na Figura 29.

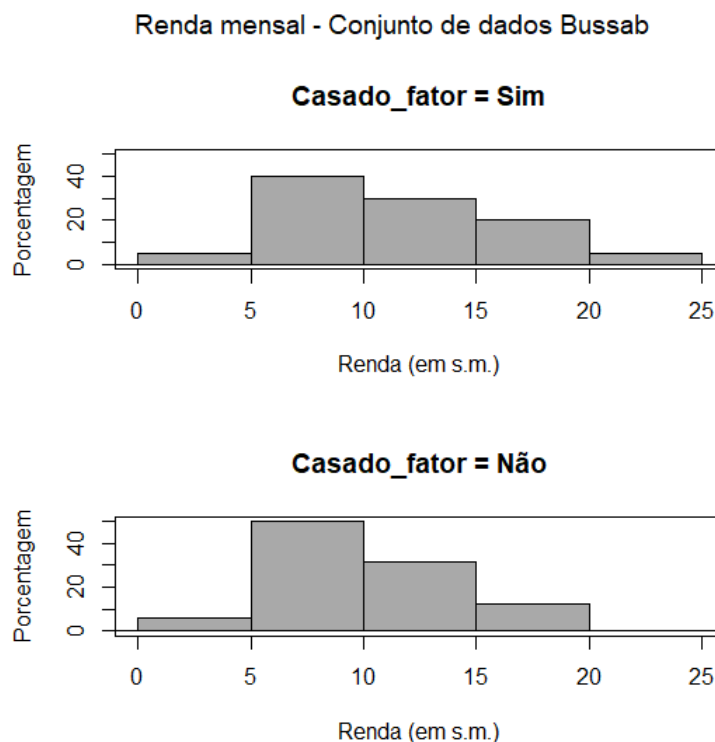


Figura 29: Histogramas da variável *Renda* para os casados e para os não casados.

**Observação 28.** Para fazer um histograma sem divisão por grupos (como o exibido na Figura 28) após fazer histogramas por grupos (como os exibidos na Figura 29), é necessário clicar no botão *Resetar* na janela de construção de histogramas.

## 4.8 Boxplot

Além do histograma, outro gráfico bastante utilizado para analisar graficamente uma variável quantitativa é o boxplot. Para obtê-lo, acesse o menu **Gráficos → Boxplot**. A janela de construção de boxplots se assemelha bastante à janela de construção de histogramas. A diferença é que, no lugar do número de classes e da escala de altura das barras, temos o campo *Identificar “outliers”*, que remete à identificação de cada observação que é considerada valor discrepante na variável em análise. Todavia, tal identificação pode deixar o gráfico visualmente poluído, pois os números (ou nomes) de identificação das observações podem ficar “um por cima do outro” quando ocorrem valores discrepantes iguais ou muito próximos. Para um gráfico mais “limpo”, marque a opção *Não* em *Identificar “outliers”*, ou a opção *Com o mouse*. Neste último caso, os valores

discrepantes são identificados quando clicamos sobre eles no gráfico (clicar com o botão direito sobre a janela *R Graphics* abre a opção de encerramento da marcação dos valores discrepantes). Para a geração do boxplot das idades dos indivíduos do conjunto de dados *Bussab* ilustrado na Figura 30, a única modificação feita na aba *Opções* se deu no título do gráfico. Como este boxplot não apresenta valores discrepantes, o campo *Identificar “outliers”* é irrelevante.

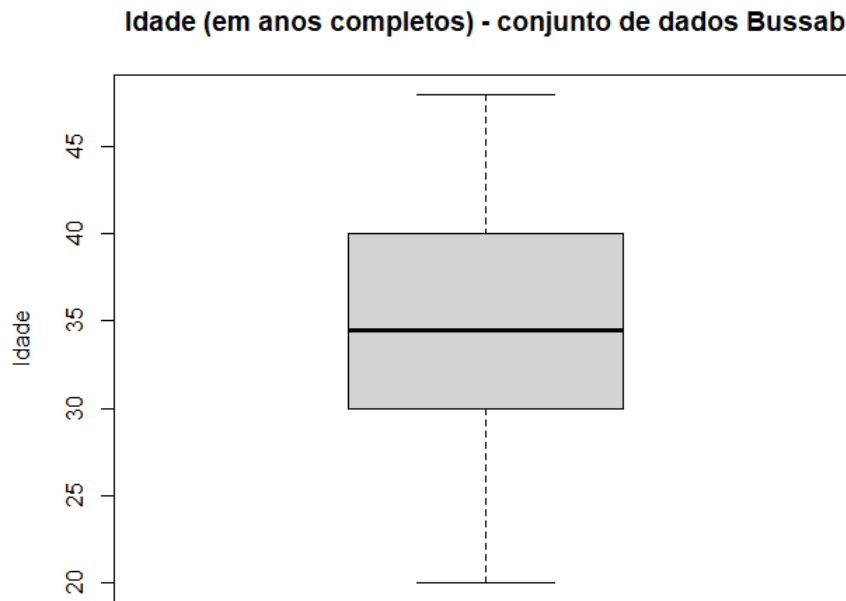


Figura 30: Boxplot da variável *Idade*.

Podemos fazer boxplots por grupos no R Commander, tal como fora feito para histogramas, gráficos de pontos e gráficos de hastes. O procedimento é análogo aos apresentados anteriormente: clique na caixa *Gráfico por grupos* no menu de construção do gráfico, selecione a variável de grupo e clique em OK. Na Figura 31, seguem boxplots da variável *Renda* do conjunto de dados *Bussab*, agrupados pelo nível de instrução (variável *Instrução* escolhida como variável de grupo). Tal como na Figura 30, não há valores discrepantes nestes boxplots, e as edições pertinentes na aba *Opções* da janela de construção destes gráficos se dão apenas nos rótulos dos eixos e no título do gráfico.

**Observação 29.** Para gerar um boxplot simples (sem divisão por grupos) após ter feito boxplots por grupos, clique em *Resetar* na janela de construção de boxplots.

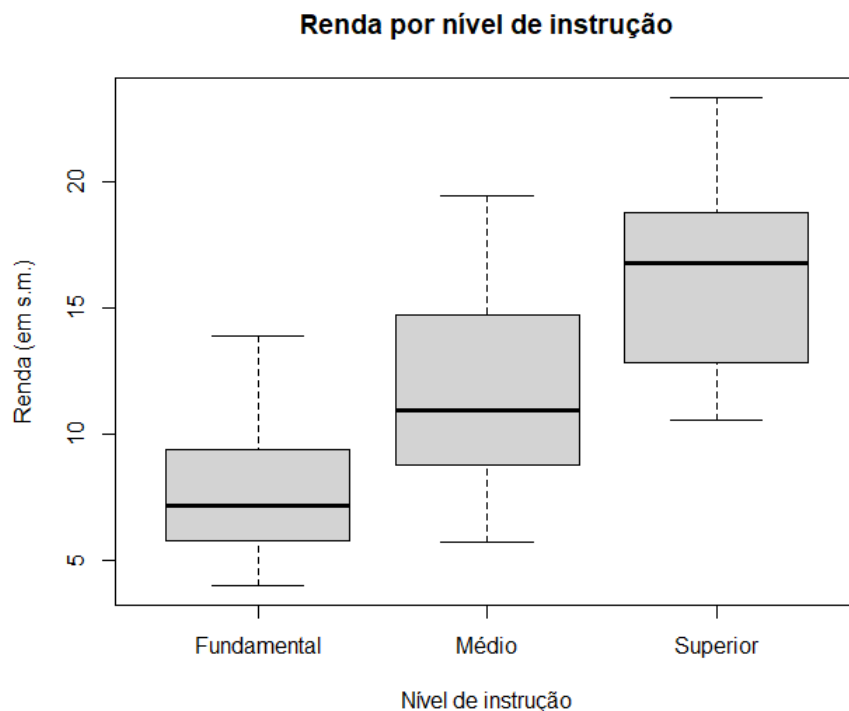


Figura 31: Boxplots da variável *Renda* agrupados pelo nível de instrução.

## 4.9 Diagrama de dispersão

O diagrama de dispersão é um gráfico que permite visualizar associação (ou falta de associação) entre duas variáveis quantitativas em um conjunto de dados e pode ser gerado por meio do menu **Gráficos → Diagrama de dispersão**. Por exemplo, para visualizarmos como estão associadas as idades e as rendas dos indivíduos do conjunto de dados *Bussab*, acesse este menu e, na aba *Dados*, escolha *Idade* para variável-x e *Renda* para variável-y. Em seguida, preencha a aba *Opções* tal qual mostra a Figura 32 e dê OK. A Figura 33 mostra o diagrama de dispersão obtido, onde a linha azul representa a reta de regressão (referenciada no menu como *Linha de quadrados mínimos*).

Após gerar este gráfico, acesse novamente a janela de construção de diagramas de dispersão, mas agora clique na caixinha *Gráfico por grupos*, selecione *Casado\_fator* como variável de grupo e clique em OK. Clicando em OK na janela de construção do gráfico, são gerados dois diagramas de dispersão sobrepostos, conforme exibe a Figura 34: um diagrama de dispersão das variáveis *Idade* e *Renda* dos indivíduos casados, e um diagrama de dispersão destas mesmas variáveis, porém apenas entre os indivíduos que não são

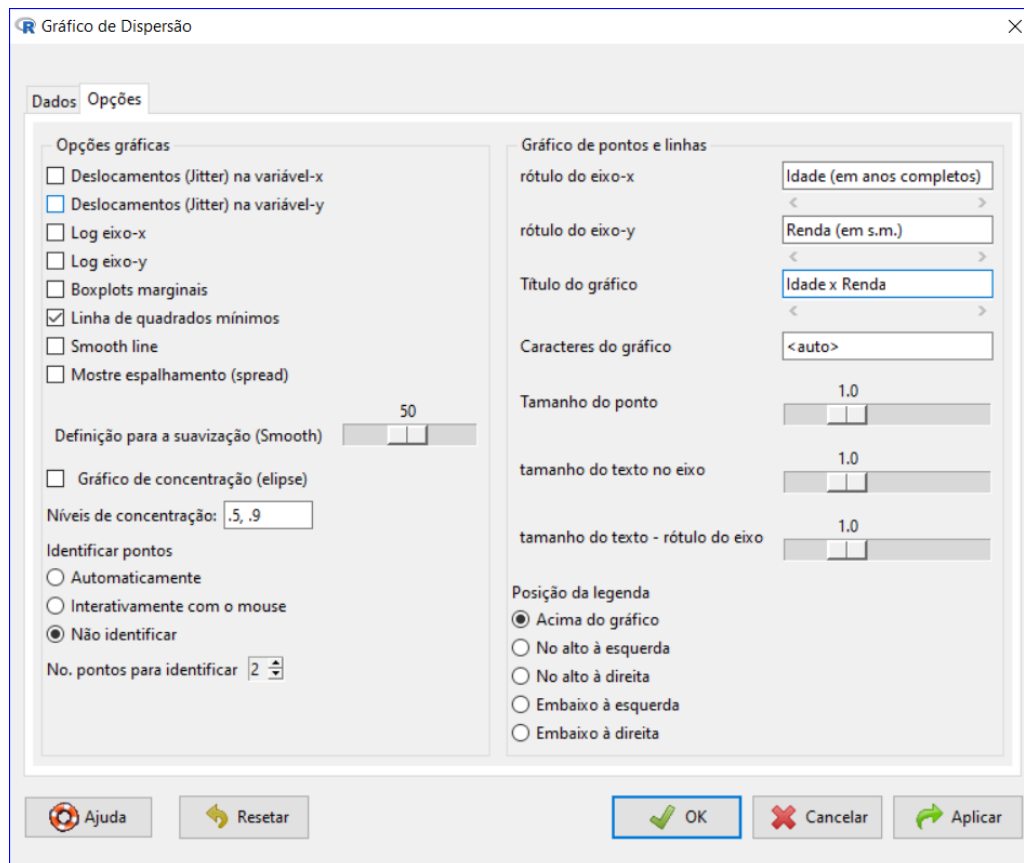



Figura 32: Construindo diagrama de dispersão.

casados, ambos com sua respectiva reta de regressão<sup>5</sup>. As cores associadas aos diagramas de dispersão provém de uma paleta de cores específica do pacote `car`. Para mais detalhes, digite `carPalette() [-1]` na janela *R Console* e aperte .

**Observação 30.** Para construir um diagrama de dispersão sem divisão por grupos (como na Figura 33) após executar diagramas de dispersão por grupos (como na Figura 34), é necessário clicar no botão *Resetar* na janela de construção de diagramas de dispersão.

**Observação 31.** Na construção de diagramas de dispersão (Figura 32), podemos escrever um número inteiro de 1 até 25 no lugar de `<auto>` no campo *Caracteres do gráfico*. Cada número retorna um símbolo diferente para os pontos do diagrama de dispersão. O número 19, por exemplo, resulta em círculos preenchidos, e o número 17 em triângulos preenchidos. No caso de diagramas de dispersão por grupos, é necessário escrever um número para cada atributo da variável de grupo, separados por vírgulas.

<sup>5</sup>Desmarcar a caixinha *Gráfico de linhas por grupo* (na janela de escolha da variável de grupo) faz com que seja traçada a reta de regressão de todas as observações ao invés das retas de regressão de cada grupo.



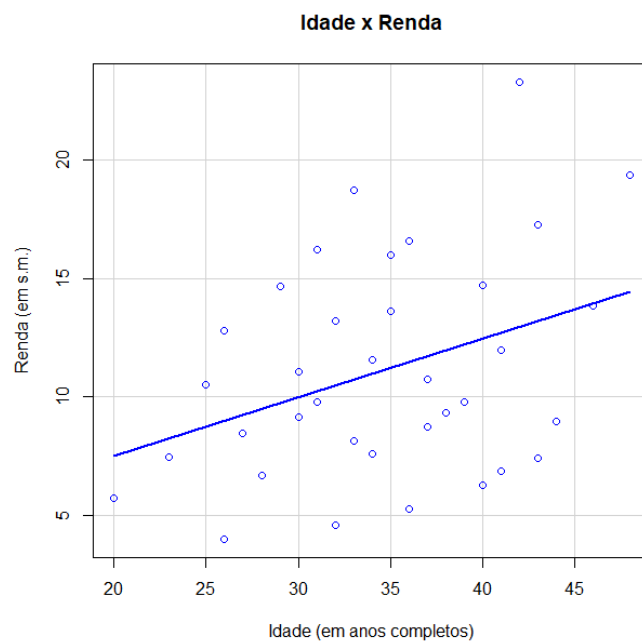


Figura 33: Diagrama de dispersão entre as variáveis *Idade* e *Renda*.

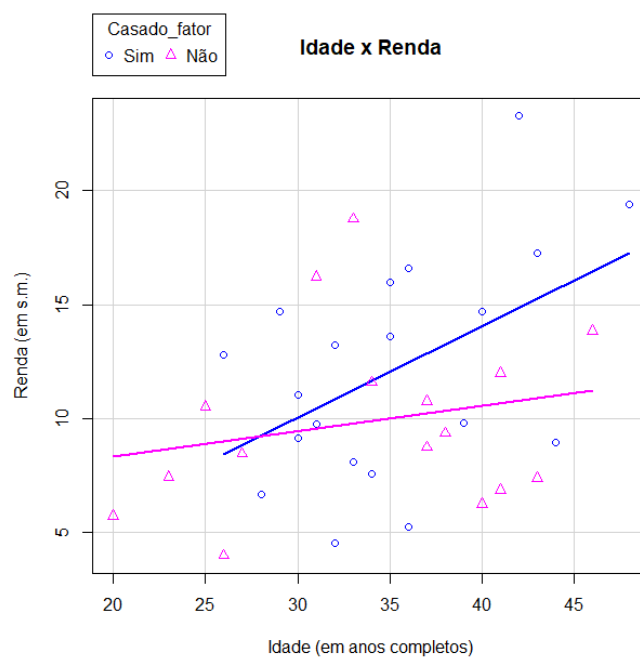


Figura 34: Diagramas de dispersão sobrepostos entre as variáveis *Idade* e *Renda* dos indivíduos casados (na cor azul) e dos indivíduos que não são casados (na cor magenta).

## 4.10 Salvando gráficos

Para **salvar um gráfico pela janela *R Gui***, clique na janela do gráfico (*R Graphics*), dimensione-a na forma que desejar e vá em **Arquivo → Salvar como**. Escolha o formato do arquivo de imagem (sugestão: Png).

Também há como **salvar um gráfico pela janela *R Commander***, por meio do menu **Gráficos → Salvar gráfico em arquivo**. Porém, o resultado é quase sempre frustrante. Além disso, uma desvantagem de salvar figuras pelo *R Commander* é que não conseguimos visualizar de antemão a dimensão da figura que estamos salvando. Já pelo *R Gui*, a janela *R Graphics* pode ser dimensionada com auxílio do *mouse*, e o arquivo salvo exhibe fielmente o gráfico da forma que ele está disposto nesta janela.

## 5 O menu *Estatísticas*

Em geral, a análise de gráficos consiste apenas de parte de uma análise estatística. Nesta seção, é apresentada parte do ferramental estatístico disponível no menu *Estatísticas*, o que torna nossas análises mais completas. **Os resultados destes comandos são exibidos na janela *Output* do *R Commander*, na cor azul.**

Apresentamos aqui apenas as funcionalidades mais básicas dos dois primeiros submenus do menu *Estatísticas*: ***Estatísticas → Resumos*** e ***Estatísticas → Tabelas de Contingência***. Resultados de estatística inferencial, como intervalos de confiança e testes de hipóteses, também podem ser obtidos por meio do menu *Estatísticas* em vários de seus submenus, e modelos estatísticos (como modelos de regressão linear e modelos lineares generalizados) podem ser ajustados no menu ***Estatísticas → Ajuste de modelos***.

Tal como na Seção 4, adotamos o conjunto de dados *Bussab* como conjunto de dados ativo no decorrer de toda a Seção 5.

### 5.1 Resumos numéricos de todas as variáveis

Este comando gera um resumo numérico de cada variável do conjunto de dados ativo. Para variáveis qualitativas, é gerada a distribuição de frequências dos seus atributos. E para variáveis quantitativas, é feito, de fato, um resumo numérico (isto é, um conjunto de medidas descritivas) contendo: menor valor, 1º quartil, mediana, média, 3º quartil e maior valor. Também são computados os dados faltantes (NA's) em cada variável, caso existam.

O caminho para gerar estes resumos é *Estatísticas* → *Resumos* → *Conjunto de dados ativo*.

## 5.2 Resumo numérico de uma variável (quantitativa)

Para gerar um resumo numérico de uma variável quantitativa em particular, vá em *Estatísticas* → *Resumos* → *Resumos numéricos*. Na aba *Dados*, escolha a(s) variável(is) de interesse. Na aba *Estatísticas*, é possível escolher quais medidas descritivas serão geradas pelo comando, entre elas: média, mediana, 1º quartil, 3º quartil, desvio padrão, distância interquartil e coeficiente de variação.

Deixando a aba *Estatísticas* na sua forma original, o resumo numérico da variável *Renda* (do conjunto de dados *Bussab*) segue na Figura 35, no qual temos: média (*mean*), desvio padrão (*sd*, de *standard deviation*), distância interquartil (*IQR*), menor valor (0%), 1º quartil (25%), mediana (50%), 3º quartil (75%), maior valor (100%) e total de observações (*n*). Quando a variável apresenta dados faltantes, também é exibido no resumo numérico quantos deles temos, isto é, o total de NA's para a variável em questão no conjunto de dados ativo.

mean	sd	IQR	0%	25%	50%	75%	100%	n
11.12222	4.587458	6.5075	4	7.5525	10.165	14.06	23.3	36

Figura 35: Resumo numérico da variável *Renda*.

É possível fazer resumos numéricos por grupos (de forma análoga a exposta na Seção 4 para gráficos de pontos, gráficos de hastes, histogramas, boxplots e diagramas de dispersão) clicando em *Resuma por grupos* na aba *Dados* do menu *Estatísticas* → *Resumos* → *Resumos numéricos*. Na Figura 36, segue o resumo numérico da variável *Renda*, com a variável *Instrução* escolhida como variável de grupo, o que gera três resumos numéricos de rendas (em salários mínimos) dispostos em três linhas. A primeira linha corresponde ao resumo numérico das rendas dos indivíduos do conjunto de dados *Bussab* cujo grau de instrução é ensino fundamental. Analogamente, a segunda e a terceira linha correspondem ao resumo numérico das rendas dos indivíduos com ensino médio e com ensino superior como nível de instrução, respectivamente.

	mean	sd	IQR	0%	25%	50%	75%	100%	Renda:n
Fundamental	7.836667	2.956464	3.155	4.00	6.0075	7.125	9.1625	13.85	12
Médio	11.528333	3.715144	5.580	5.73	8.8375	10.910	14.4175	19.40	18
Superior	16.475000	4.502438	4.730	10.53	13.6475	16.740	18.3775	23.30	6

Figura 36: Resumos numéricos por nível de instrução da variável *Renda*.

**Observação 32.** Para gerar um resumo numérico simples (sem divisão por grupos, conforme ilustrado na Figura 35) após ter sido feito um resumo numérico por grupos (conforme ilustrado na Figura 36), clique em *Resetar* ao acessar o menu de geração de resumos numéricos.

No menu de geração de resumos numéricos, perceba que, na aba *Estatísticas*, há uma caixinha (por padrão, desmarcada) chamada *Frequency Counts*. Marcando esta caixinha, o resumo numérico gerado vem acompanhado da distribuição de frequências da variável escolhida, caso ela tenha poucos valores distintos, ou de uma distribuição de frequências por classes desta variável, caso ela tenha muitos valores distintos. Isto pode ser muito útil para verificar precisamente as alturas das hastes do gráfico (de hastes) ou das barras do histograma desta mesma variável, uma vez que o procedimento padrão utilizado para geração de classes é o mesmo.

**Exercício 11.** Refaça o resumo numérico da variável *Renda* sem divisão por grupos (ou seja, clique em *Resetar* assim que acessar o menu), porém marque a caixinha *Frequency Counts*. Compare a frequência de cada classe com a altura da respectiva barra do histograma ilustrado na Figura 28.

### 5.3 Contando dados faltantes

Apesar dos resumos numéricos debatidos nas Subseções 5.1 e 5.2 computarem dados faltantes (comumente representados por *NA*), há um menu no R Commander específico para isto: *Estatísticas* → *Resumos* → *Contar observações faltantes*. Ao acessar este menu, o R Commander exibe (na janela *Output*) o total de dados faltantes de todas as variáveis do conjunto de dados ativo. Considerando apenas as variáveis originais do conjunto de dados *Bussab*, apenas a variável *Filhos* apresenta *NA*'s, e estes dados faltantes se devem ao fato da informação do número de filhos ter sido restrita aos indivíduos casados.

Binned distribution of Renda		
	Count	Percent
[4, 6]	4	11.11
(6, 8]	6	16.67
(8, 10]	8	22.22
(10, 12]	5	13.89
(12, 14]	4	11.11
(14, 16]	3	8.33
(16, 18]	3	8.33
(18, 20]	2	5.56
(20, 22]	0	0.00
(22, 24]	1	2.78
Total	36	100.00

Figura 37: Distribuição de frequências por classes solicitada no Exercício 11.

**Observação 33.** Há situações nas quais pode ser interessante remover linha(s) do conjunto de dados na(s) qual(is) uma ou mais variáveis apresentem dados faltantes. Isto pode ser feito rapidamente no R Commander por meio do menu *Dados → Conjunto de dados ativo → Remover observações com dados faltantes*. Por segurança, é recomendável dar um outro nome para o conjunto de dados sem as “linhas indesejáveis”, de forma a preservar o conjunto de dados com todas as observações.

## 5.4 Distribuições de frequências

Para obtermos a distribuição de frequências de uma variável no R Commander, basta acessarmos o menu *Estatísticas → Resumos → Distribuições de frequência*, escolhermos a(s) variável(is) de interesse e clicarmos em OK. Este comando produz duas tabelas de distribuição de frequências para cada variável escolhida: uma com contagens (frequências absolutas) e outra com percentuais (frequências relativas  $\times 100$ ).


Note que as variáveis disponíveis para escolha são apenas as variáveis qualitativas<sup>6</sup> do conjunto de dados ativo. Porém, quando temos uma variável quantitativa discreta com poucos valores distintos, também é interessante obter sua distribuição de frequências. Isto pode ser contornado convertendo esta variável quantitativa em rótulos numéricos (ver Subseção 3.8.1), criando assim uma nova coluna no conjunto de dados que é interpretada pelo R Commander como uma variável qualitativa e, portanto, passível da aplicação de distribuição de frequências no menu supracitado.

<sup>6</sup>A rigor, variáveis que o R Commander “enxerga” como qualitativas, isto é, variáveis “não-numéricas”.

counts:		
Instrução		
Fundamental	Médio	Superior
12	18	6
percentages:		
Instrução		
Fundamental	Médio	Superior
33.33	50.00	16.67

Figura 38: Distribuição de frequências da variável *Instrução*.

## 5.5 Matriz de correlação

Na Subseção 4.9, mostramos como obter um diagrama de dispersão (com a reta de regressão sobreposta ao gráfico). Quando fazemos esta representação gráfica, é importante também relatar o coeficiente de correlação entre as variáveis analisadas. Por meio da matriz de correlação, podemos verificar o(s) coeficiente(s) de correlação de um ou mais pares de variáveis quantitativas do conjunto de dados ativo. Tal matriz é obtida no R Commander no menu *Estatísticas* → *Resumos* → *Matriz de Correlação*. Uma vez acessado este menu, escolha duas ou mais variáveis (segurando a tecla  enquanto as seleciona) e clique em OK. Podemos verificar na Figura 39 que o coeficiente de correlação (de Pearson) entre as variáveis *Idade* e *Renda* do conjunto de dados *Bussab* é aproximadamente 0,363, que configura uma correlação positiva moderada (quase fraca).

	Idade	Renda
Idade	1.0000000	0.3633622
Renda	0.3633622	1.0000000

Figura 39: Matriz de correlação com as variáveis *Idade* e *Renda*.

**Observação 34.** Na janela de construção da matriz de correlação, marcar a caixinha *P-valores pareados* antes de clicar em OK gera, além da matriz de correlação conforme ilustrada na Figura 39, uma matriz de p-valores de testes de correlação conforme a opção escolhida no campo *Tipos de Correlações*. Em particular, esses p-valores também podem ser obtidos no menu *Estatísticas* → *Resumos* → *Teste de Correlação*.

## 5.6 Tabelas de contingência

Para um conjunto de dados com pelo menos duas variáveis qualitativas, é possível construir tabelas de contingência. Aqui, abordamos apenas a construção de tabelas de dupla entrada, isto é, tabelas de contingência compostas por um par de variáveis. Para tal, o caminho é *Estatísticas* → *Tabelas de Contingência* → *Tabela de dupla entrada*.

Na janela de construção da tabela, ainda na aba *Dados*, selecione *Casado\_fator* como variável linha e *Instrução* como variável coluna. Na aba *Estatísticas*, temos opções associadas a testes de hipóteses baseados na tabela de contingência (a única caixinha marcada por padrão é a caixinha do teste qui-quadrado de independência). Deixando esta aba do jeito que está, obtemos uma saída conforme ilustra a Figura 40: acima da expressão *Pearson Chi-squared test*, encontra-se, de fato, a tabela de contingência solicitada; e, abaixo desta expressão, são listadas as saídas provenientes do teste qui-quadrado de independência de Pearson.

```
Frequency table:
      Instrução
Casado_fator Fundamental Médio Superior
      Sim           5      12       3
      Não           7       6       3

      Pearson's Chi-squared test

data: .Table
X-squared = 1.9125, df = 2, p-value = 0.3843
```

Figura 40: Tabela de contingência com as variáveis *Casado\_fator* e *Instrução*, seguida do teste qui-quadrado de independência de Pearson entre estas variáveis.

Além do campo *Testes de hipóteses*, a aba *Estatísticas* do menu de construção de tabelas de dupla entrada apresenta o campo *Computar Percentagens*. A opção padrão *Sem percentual* expressa apenas uma tabela de dupla entrada, conforme exibida na Figura 40. As outras três opções deste campo (*Percentual nas linhas*, *Percentual nas colunas* e *Percentagens do total*) exibem, além da tabela de dupla entrada com frequências absolutas, uma tabela de dupla entrada cujas entradas são todas na escala percentual (exceto na linha ou coluna *Count*), conforme detalhado a seguir.

- **Percentual nas linhas:** Apresenta o perfil percentual da variável escolhida como variável coluna em cada linha, como mostra a Figura 41. As entradas na coluna *Count* retornam o total de observações que satisfazem o atributo da respectiva linha.

Row percentages:					
	Instrução				
Casado_fator	Fundamental	Médio	Superior	Total	Count
Sim	25.0	60.0	15.0	100.0	20
Não	43.8	37.5	18.8	100.1	16

Figura 41: Perfil percentual da variável *Instrução* (escolhida como variável coluna) em cada atributo da variável *Casado\_fator* (escolhida como variável linha).

- **Percentual nas colunas:** Apresenta o perfil percentual da variável escolhida como variável linha em cada coluna, como ilustra a Figura 42. As entradas na linha *Count* consistem do total de observações que satisfazem o atributo da respectiva coluna.

Column percentages:				
	Instrução			
Casado_fator	Fundamental	Médio	Superior	
Sim	41.7	66.7	50	
Não	58.3	33.3	50	
Total	100.0	100.0	100	
Count	12.0	18.0	6	

Figura 42: Perfil percentual da variável *Casado\_fator* (escolhida como variável linha) em cada atributo da variável *Instrução* (escolhida como variável coluna).

- **Percentagens do total:** Apresenta o percentual de cada cruzamento de atributos das duas variáveis em relação ao total de observações, como ilustra a Figura 43.

Total percentages:				
	Fundamental	Médio	Superior	Total
Sim	13.9	33.3	8.3	55.6
Não	19.4	16.7	8.3	44.4
Total	33.3	50.0	16.7	100.0

Figura 43: Percentual de cada cruzamento dos atributos das variáveis *Casado\_fator* e *Instrução* em relação ao total de observações.



**Observação 35.** A coluna *Total* de uma tabela de percentuais nas linhas e a linha *Total* de uma tabela de percentuais nas colunas representam as somas das percentagens na respectiva linha ou coluna. Matematicamente, estas somas resultam em 100%, caso os percentuais somados sejam os percentuais exatos. Todavia, alguns dos percentuais vistos nas três figuras acima estão aproximados por apenas uma casa decimal, e a soma destes percentuais aproximados pode retornar algum valor ligeiramente diferente de 100%. Por exemplo, na Figura 41, a entrada na coluna *Total* e linha *Não* retorna 100,1%. Este percentual é, na verdade, uma soma de percentuais que passaram por aproximação de apenas uma casa decimal: as entradas 43.8 e 18.8 (em escala percentual) são aproximações de  $7/16 = 43,75\%$  e  $3/16 = 18,75\%$ . De fato,  $43,75\% + 37,5\% + 18,75\% = 100\%$ .

## 6 Considerações finais

Para estudantes de graduação de diversos cursos da UNIRIO, cursar disciplinas de Estatística, Bioestatística e afins com suporte do *software* R é uma realidade, que se consolidou graças ao uso de um “R com menus”. O principal objetivo deste material é reunir, em uma única apostila, do passo a passo para *download* do R às principais funcionalidades propiciadas pelo R Commander para análise exploratória de dados, de forma que mesmo quem nunca utilizou o *software* R consiga acompanhá-la e reproduzir as tarefas indicadas a cada seção/subseção.

Por ser um material introdutório, esta apostila explora apenas os três menus principais do R Commander: *Dados*, *Gráficos* e *Estatísticas* (e, nenhum deles, de forma integral). Ferramentas de estatística indutiva, tais como intervalos de confiança e testes de hipóteses, não são abordadas neste material, mas há muitas delas no menu *Estatísticas*. Este menu também possibilita a implementação de modelos estatísticos, e o menu *Modelos* é dedicado à exploração de modelos previamente implementados. Distribuições de probabilidade podem ser exploradas no menu *Distribuições*, e pelo menu *Ferramentas*, é possível carregar *plugins* pré-instalados do *Rcmdr* e acessar o menu de opções da janela *R Commander*.

Ao unir os tópicos relatados no parágrafo acima com tudo o que foi abordado nas Seções 2 até 5, fica claro que a interface R Commander contempla uma gama enorme de possibilidades em seus menus para quem deseja fazer uma boa análise de dados.

## Referências

- Bussab, W.; Morettin, P. (2017). *Estatística Básica*. São Paulo: Saraiva, 9<sup>a</sup> edição.
- Fox, J. (2005). The R Commander: A Basic-Statistics Graphical User Interface to R. *Journal of Statistical Software*, **14**(9), 1–42. <https://doi.org/10.18637/jss.v014.i09>.
- Fox, J.; Weisberg, S. (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. Disponível em: <https://www.john-fox.ca/Companion/index.html>.
- Fox, J.; Weisberg, S.; Price, B. (2022). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-5. Disponível em: <https://CRAN.R-project.org/package=carData>.
- Fox, J. (2022) *Getting started with the R Commander - Version 2.8-0*. Disponível em: <https://cran.r-project.org/web/packages/Rcmdr/vignettes/Getting-Started-with-the-Rcmdr.pdf>.
- Fox, J.; Bouchet-Valat, M. (2024). *Rcmdr: R Commander*. R package version 2.9-2. Disponível em: <https://cran.r-project.org/web/packages/Rcmdr/index.html>.
- Mignozzetti, U.G. (2009). *Introdução ao R Commander*. Disponível em: [https://www.bessegato.com.br/UFJF/est000\\_edu/aprcmdr.pdf](https://www.bessegato.com.br/UFJF/est000_edu/aprcmdr.pdf).
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Viena, Áustria. Disponível em: <https://www.R-project.org/>
- Zeileis, A.; Hornik, K.; Murrell, P. (2009). Escaping RGBland: Selecting Colors for Statistical Graphics. *Computational Statistics & Data Analysis*, **53**(9), 3259-3270. <https://doi.org/10.1016/j.csda.2008.11.033>.
- Zeileis, A.; Fisher, J.C.; Hornik, K.; Ihaka, R.; McWhite, C.D.; Murrell, P.; Stauffer, R.; Wilke, C.O. (2020). colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes. *Journal of Statistical Software*, **96**(1), 1-49. <https://doi.org/10.18637/jss.v096.i01>.