R Commander: Usando menus para análise de dados no R

Felipe Rafael Ribeiro Melo

Departamento de Métodos Quantitativos

21ª Semana de Integração Acadêmica da UNIRIO 18 de outubro de 2023

O que é R?

- R é um *software* livre voltado principalmente para o tratamento e análise de dados.
- De maneira mais estendida, pode ser visto como uma linguagem de programação especializada na gestão e análise de dados.
- Começou a ser desenvolvido pelos pesquisadores Robert Gentleman e Ross Ihaka, do Departamento de Estatística da Universidade de Auckland, na Nova Zelândia, em meados dos anos 90.

Por que usar o R?

- Gratuito, potente e estável.
- Pode ser copiado e distribuído sem nenhum problema.
- É um programa que tem uma longa história, com mais de 20 anos de desenvolvimento.
- É apoiado por uma grande equipe de desenvolvedores em todo o mundo.
- Diversos fóruns na Internet para solucionar dúvidas.
- Amplamente utilizado no meio acadêmico em diversas áreas de conhecimento.
- Disponível para Windows, Linux e Mac.

Desvantagens do R

- A principal desvantagem do R reside na sua interface pouco "amigável", apta a receber apenas linhas de comando.
- Cabe ressaltar que há um *software* que fornece um ambiente de desenvolvimento integrado (IDE) para o R chamado RStudio programa gratuito e "queridinho" da maioria dos usuários de R.



• Todavia, no RStudio, a maior parte das funcionalidades do R ainda necessita da compilação de linhas de comando.

O pacote Rcmdr

- Uma das características marcantes do *software* R é a grande quantidade de *pacotes* disponíveis, os quais fornecem funcionalidades específicas.
- Rcmdr (uma forma abreviada de R Commander) é um pacote disponível para o *software* R que possibilita a execução de várias funcionalidades voltadas ao tratamento e à análise de conjuntos de dados por meio de menus, sem a necessidade de digitar linhas de comando, fornecendo uma interface "point & click".

OK! Mas...

- 1. Onde baixar o software R?
- 2. Sua instalação é simples?
- 3. Uma vez que o R esteja instalado, como instalar o pacote Rcmdr nele?
- 4. Como chegar nesta desejada interface "point & click" fornecida pelo pacote Rcmdr?

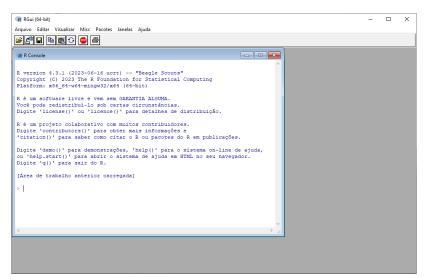
Download e instalação do R (no Windows)

Acesse www.r-project.org e clique sequencialmente em:

- CRAN (lado esquerdo da tela);
- um dos *mirrors* disponíveis (sugestão: algum *mirror* do Brasil);
- Download R for Windows;
- *base*;
- no link em destaque (Download R-4.3.1 for Windows).

O processo de instalação é simples: duplo clique no arquivo executável baixado e, por comodidade, clique em OK e sequencialmente em "Avançar".

Ao abrirmos o R, visualizamos a já mencionada "interface pouco amigável". Vamos portanto instalar o pacote Rcmdr!



Instalação do pacote Rcmdr

- No R, acesse o menu $Pacotes \rightarrow Instalar\ pacote(s)$.
- Escolha algum mirror (sugestão: θ -Cloud) e dê OK.
- Na janela seguinte, para instalar **apenas** o pacote Rcmdr: selecione Rcmdr e dê OK.
- Após isto, caso sejam abertas duas pequenas janelas, uma após a outra, com opções Sim e $N\tilde{a}o$, selecione Sim em ambas.
- A conclusão do processo de instalação se dá quando o prompt de comando (sinal > na cor vermelha) surgir novamente no canto inferior esquerdo da janela R Console.

• Caso você deseje instalar o pacote Rcmdr e seus *plugins* simultaneamente: selecione todos os pacotes que começam com "Rcmdr" (incluindo o próprio pacote Rcmdr) e dê OK. Este processo de instalação é mais demorado.

Observação 1

A instalação de qualquer pacote também pode ser realizada via linha de comando. O caminho de instalar apenas o Rcmdr (sem seus plugins) exposto no slide anterior pode ser substituído simplesmente por digitar a linha de comando

install.packages("Rcmdr")

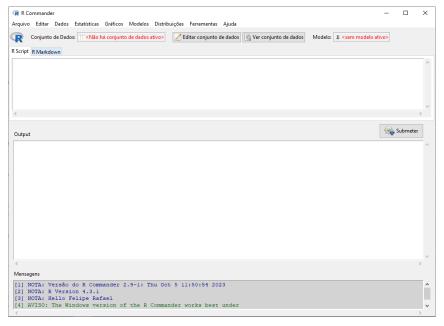
e apertar Enter em seguida.

Carregando o pacote Rcmdr

- A instalação do pacote Remdr, assim como a instalação do R, <u>só precisa ser feita uma vez</u>.
- Para ter acesso a interface que o pacote Rcmdr fornece (isto é, a janela *R Commander*), é necessário *carregar* este pacote. Para tal, basta abrir o R, digitar <u>uma</u> das duas linhas abaixo e apertar <u>Enter</u> em seguida:

require(Rcmdr)
library(Rcmdr)

• Na primeira vez que o pacote Rcmdr for carregado, pode surgir uma pequena janela com opções Sim e $N\~ao$. Clique em Sim e, na janela seguinte, em OK.



Estrutura base da janela R Commander

A janela R Commander é dividida em três janelas:

- R Script: Nesta janela são impressas as linhas de comando de tudo que é executado via menus. 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 e mensagens de erro.

E se esta janela *R Commander* for acidentalmente fechada?

<u>Não</u> será possível reabri-la na atual sessão do R, ou seja, carregar novamente o pacote Rcmdr não surtirá efeito. Portanto, será necessário abrir uma nova sessão do R e carregar novamente o pacote Rcmdr, por meio das linhas de comando require(Rcmdr) ou library(Rcmdr).

Os três principais menus para quem está começando no R Commander:

- Dados, para tudo que se refere à geração, importação, carregamento e/ou manipulação de conjunto de dados;
- *Gráficos*, para a geração de visualizações gráficas;
- Estatísticas, para a geração de resultados que não sejam gráficos, como distribuições de frequências, tabelas de contingência, resumos numéricos, intervalos de confiança e testes de hipóteses.

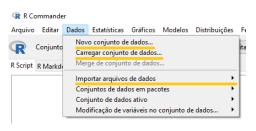
Referências para o "Rcmdr básico"

- Acesse gae.uniriotec.br/7/material.html (ou github.com/feliperafaelrm/SIA2023_Minicurso).
- Apostila introdutória "Introdução ao R Commander Notas de Aula" e apostila complementar "R Commander: um pouco além dos menus gráficos", além de conjuntos de dados utilizados em exemplos destas duas apostilas.
- Apostilas revisadas pela última vez em fevereiro de 2019 (versão do R: 3.5.2 / Versão do Rcmdr: 2.5-1).
- Em essência, elas podem ser bem aproveitadas, ainda que atualizações do pacote Rcmdr tenham acrescentado novas possibilidades não abordadas nas apostilas (e talvez removido outras possibilidades).

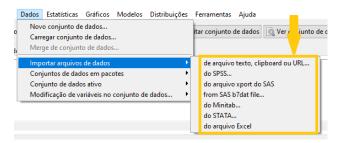
Inserindo conjunto de dados

Há três formas de inserir conjunto de dados via R Commander.

- **1** Novo conjunto de dados: digitar diretamente no R.
- Carregar conjunto de dados: quando o conjunto de dados de interesse está no formato .RData.
- **1** Importar arquivo de dados: quando o conjunto de dados de interesse está em outros formatos.



• Como é comum lidarmos com dados secundários em formatos diferentes de .RData, frequentemente os dados são inseridos via *importação*.



• Após a importação, o conjunto de dados pode ser facilmente salvo no formato .RData e, como isso, ele pode ser recuperado em qualquer sessão futura por meio da opção de *carregamento*, sem a necessidade de importálo novamente.

Importando conjunto de dados em .xls/.xlsx

- Usaremos aqui o conjunto de dados presente no arquivo dataset_minicurso.xlsx, disponível em http://gae. uniriotec.br/7/educaser/dataset_minicurso.xlsx.
- Caso o link acima n\u00e3o funcione, baixe o conjunto de dados em https://github.com/feliperafaelrm/SIA2023_ Minicurso/blob/main/dataset_minicurso.xlsx.
- Consiste num conjunto de dados fictício de alunos de uma turma do final do ensino fundamental, com variáveis intrínsecas aos alunos e outras variáveis relacionadas ao número de faltas e notas numa determinada disciplina, em cada um dos 4 bimestres do ano letivo.

	Α	В	С	D	Е	F	G
1	NOME	SEXO	IDADE	REGIAO_RESID	MAIOR_INSTR_PAIS	FALTAS_BIM1	NOTA_BIM1
2	Ana Clara	1	14	Zona Oeste	Ensino Médio	1	5,5
3	Andressa	1	13	Baixada Fluminense	Ensino Fundamental	0	8,3
4	Arthur	2	12	Zona Norte	Ensino Médio	0	8,0
5	Átila	2	13	Zona Oeste	Ensino Fundamental	3	4,2
6	Bianca	1	14	Zona Oeste	Ensino Médio	0	9,5
7	Breno	2	14	Zona Oeste	Ensino Médio	0	8,0
8	Bruna	1	13	Zona Norte	Ensino Superior	0	9,1
9	Cauan	2	13	Zona Norte	Ensino Médio	2	5,4
10	Daniel	2	12	Zona Oeste	Ensino Médio	1	8,3
11	Ellen	1	12	Zona Norte	Sem instrução	1	7,6
12	Fabrício	2	13	Baixada Fluminense	Ensino Fundamental	0	6,7
13	Fernanda	1	13	Zona Norte	Ensino Médio	0	8,5
14	Gabriela	1	14	Zona Norte	Ensino Médio	0	9,0

• Menu para a importação de um conjunto de dados em "arquivo Excel" (formatos .xls ou .xlsx) no R Commander:

 $Dados \rightarrow Importar \ arquivos \ de \ dados \rightarrow \ do \ arquivo \ Excel$

- Dê um nome ao conjunto de dados <u>sem usar espaços</u> (sugestão: Educa), substitua *<casela vazia>* por *NA* e clique em OK.
- Busque pelo arquivo dataset_minicurso.xlsx e, após duplo-clique, não esqueça de clicar em *Dados* na pequena janela que surge, pois é esse o nome da planilha onde está o conjunto de dados dentro do arquivo selecionado na importação.

O nome digitado para o conjunto de dados importado aparece escrito, na cor azul, ao lado de "Conjunto de dados:", onde estava inicialmente escrito, na cor vermelha: "<Não há conjunto de dados ativo>".

• Antes da importação:



• Depois da importação:



- O conjunto de dados cujo nome está escrito na cor azul é o chamado *conjunto de dados ativo*.
- Clicando na caixinha "Ver conjunto de dados", o conjunto de dados ativo será exibido numa janela à parte, à título de visualização. Ou seja, fechar esta janela <u>não</u> remove o conjunto de dados ativo.



• IMPORTANTE: Se a janela *R Commander* estiver maximizada, a janela à parte que mostra o conjunto de dados <u>não</u> ficará visível. Portanto, antes de clicar em "Ver conjunto de dados", restaure o tamanho da janela *R Commander*, caso esta esteja maximizada.

Importando conjuntos de dados em mais formatos

O menu $Dados \rightarrow Importar arquivos de dados possibilita a importação de conjuntos de dados em outros formatos diferentes de .xls ou .xlsx:$

Formatos	Submenu
*.txt; *.dat; *.csv	de arquivo texto, clipboard ou URL
*.por ; *.sav	do~SPSS
*.xpt; *.xport	do arquivo xport do SAS
*.sas7bdat	from SAS b7dat file
*.mtp	$do\ Minitab$
*.dta	$do\ STATA$

Salvando e carregando conjunto de dados

• Para salvar o conjunto de dados (que está elencado como o conjunto de dados ativo): menu

$$\begin{array}{c} Dados \rightarrow \ Conjunto \ de \ dados \ ativo \\ \rightarrow \ Salvar \ conjunto \ de \ dados \ ativo \end{array}$$

 Feito isto, não é mais necessário importar o conjunto de dados, e após a atual sessão do R Commander ser fechada
 acidentalmente ou não - este conjunto de dados pode ser facilmente recuperado via menu de carregamento:

$Dados \rightarrow Carregar \ conjunto \ de \ dados$

• Ainda, é aconselhável salvar o conjunto de dados a cada modificação que for feita nele, de forma a não precisar refazê-las em sessões futuras.

Variável categórica x variável numérica

- A primeira coluna do conjunto de dados que importamos expressa a identificação dos alunos.
- Todas as outras colunas expressam *variáveis*, que são interpretadas como categóricas ou como numéricas.
 - Variáveis categóricas: Região de residência e Maior nível de instrução dos pais.
 - Variáveis numéricas: Sexo, Idade, Faltas (em cada um dos bimestres) e Nota (em cada um dos bimestres).
- No R Commander, há funcionalidades que estão disponíveis apenas para variáveis categóricas e outras que estão disponíveis apenas para variáveis numéricas.

Funciona apenas para variável categórica

Resultado	Menu	
Gráfico de Barras	$Gr\'{a}ficos ightarrow Gr\'{a}fico de Barras$	
Gráfico de Pizza	Gráficos o Gráfico de Pizza	
Distribuição de Frequências	$ Estatísticas \rightarrow Resumos \\ \rightarrow Distribuições de frequência $	
Tabela de dupla entrada (tabela de contingência com 2 variáveis)	Estatísticas → Tabelas de Contingência → Tabela de dupla entrada	

Funciona apenas para variável numérica

Resultado	Menu		
Histograma	$Gr\'{a}ficos ightarrow Histograma$		
Gráfico de "hastes"	$Gr\'{a}ficos \rightarrow Plot\ discrete\ numeric\ variable$		
Boxplot	$Gr\'{a}ficos o Boxplot$		
Densidades	$Gr\'{a}ficos ightarrow Estimativa de densidade$		
Medidas descritivas (média, quartis, etc.)	$ Estatísticas \rightarrow Resumos \\ \rightarrow Resumos \ num\'ericos $		
Coeficiente(s) de correlação	$ Estatísticas \rightarrow Resumos \\ \rightarrow Matriz \ de \ correlação $		

Três questões importantes a serem levantadas:

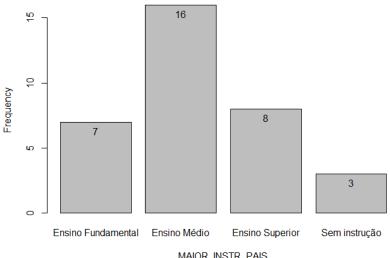
- Q1. Como fazer o R entender a ordenação natural de variável categórica do tipo ordinal?
- Q2. Como decodificar variável categórica expressa de forma numérica, fazendo com que o R a "entenda" como categórica?
- Q3. Como obter distribuição de frequências de uma variável numérica?

Q1. Reordenando categorias

• No conjunto de dados *Educa*, o "maior nível de instrução dos pais" (MAIOR_INSTR_PAIS) é uma variável categórica ordinal. Uma forma de verificar como seus atributos estão ordenados se dá por meio de um gráfico de barras, disponível no menu

$Gráficos \rightarrow Gráfico de Barras$

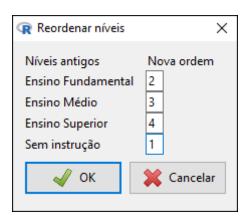
- Neste menu, apenas clique em MAIOR_INSTR_PAIS e, em seguida, em OK.
- Gráficos gerados pelos menus do R Commander $\underline{\tilde{nao}}$ são exibidos na janela R Commander, mas sim na janela do R (R Gui), numa sub-janela denominada R Graphics.



 Para que os atributos da variável MAIOR_INSTR_PAIS passem a seguir a ordenação "Sem instrução" / "Ensino fundamental" / "Ensino médio" / "Ensino superior" em vez da ordem alfabética, acesse o menu

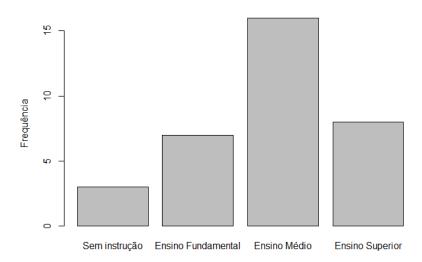
Dados → Modificação de variáveis no conjunto de dados → Reordenar níveis dos fatores

- Escolha a variável MAIOR_INSTR_PAIS, clique em OK e confirme em seguida o fato de que a "nova" variável irá sobrescrever a antiga.
- Renumere os campos conforme a nova ordenação desejada e clique em OK.



- Uma forma de verificar que a reordenação foi, de fato, executada, é refazer o gráfico de barras de MAIOR_INSTR_PAIS. Aproveitemos para "caprichar" um pouco mais nele.
- Acesse o meu *Gráficos* → *Gráfico de Barras*, selecione MAIOR_INSTR_PAIS e, antes de clicar em OK, clique na aba *Opções* e faça as seguintes edições:
 - rótulo do eixo-x:
 - rótulo do eixo-y: Frequência
 - Título do gráfico: Maior nível de instrução dos pais
 - desmarque a caixinha Show counts or percentages in bars.

Maior nível de instrução dos pais



Salvando gráficos

- Note que este último gráfico gerado (o mais "caprichado") eliminou o anterior.
- Portanto, para não perder um gráfico gerado, é necessário salvá-lo (como um arquivo de imagem ou um arquivo pdf).
- Para tal, basta clicar sobre o gráfico e, <u>ainda na</u> <u>janela *R Gui*</u>, acessar o menu

$Arquivo \rightarrow Salvar\ como$

e escolher o formato desejado (sugestão: Png).

Q2. Variável de natureza categórica codificada por números

- Note que a variável SEXO do nosso conjunto de dados está rotulada por números (1 para feminino e 2 para masculino). Desta forma, o R Commander a vê como uma variável numérica.
- Ao acessar qualquer menu de funcionalidades restritas a variáveis categóricas (como gráfico de barras, gráfico de pizza, distribuições de frequências), a variável SEXO <u>não</u> estará listada.
- Para que o R entenda tal variável como categórica, será necessário recorrer ao menu

Dados → Modificação de variáveis no conjunto de dados → Converter variável numérica para fator

e gerar uma nova coluna onde 1 se torne "feminino" e 2 se torne "masculino" (ou ainda sobrescrever a variável SEXO original deste conjunto de dados).

 No menu acima citado, selecione a variável SEXO, dê OK e confirme que a "nova" variável sobrescreverá a original. Em seguida, preencha conforme abaixo e clique em OK.

R Nomes dos níveis para SEXO X	
Valor numérico Nome do nível	
1	Feminino
2	Masculino
 ✓ OK	Cancelar

- Clique em "Ver conjunto de dados" (sem a janela *R Commander* estar maximizada) para verificar a alteração na coluna SEXO, que eliminou a coluna numérica original, conforme optamos.
- Após o procedimento acima, ao acessar qualquer menu de funcionalidades restrito a variáveis categóricas, a variável SEXO estará listada. Um destes menus, que ainda não exploramos, é o que gera gráficos de pizza:

Gráficos o Gráfico de Pizza

• Neste menu, selecione SEXO e, em *Título do gráfico*, escreva seu título de preferência.

Q3. Distribuição de frequências para variáveis numéricas

- Uma vez que o menu Estatísticas → Resumos → Distribuições de frequência funciona apenas para variáveis categóricas, como fazer, portanto, a distribuição de frequência das idades dos alunos, uma vez que a variável IDADE tem, de fato, natureza numérica?
- Alternativa: usar o mesmo menu de conversão pelo qual substituímos 1 e 2 por Feminino e Masculino na variável SEXO.
 - ⇒ A diferença é que, dessa vez, <u>não</u> queremos "perder" a coluna IDADE na sua forma original, pois ela expressa, de fato, uma variável de natureza numérica.

- Acessar o menu Dados → Modificação de variáveis no conjunto de dados → Converter variável numérica para fator, escolher variável IDADE, marcar Use números e dar um nome diferente de IDADE no campo Novo nome de variável (sugestão: IDADE_categ).
- Uma nova coluna (com o nome escolhido acima) é gerada no conjunto de dados, "espelhando" os valores da variável IDADE, mas com o R interpretando esses caracteres não como números, mas sim como meros "rótulos".
- Fazer a distribuição de frequências da variável IDADE_categ no menu $Estatísticas \rightarrow Resumos \rightarrow Distribuições de frequência.$

Apesar da alternativa acima ser razoável para a geração de distribuições de frequências, o mesmo não se pode dizer para gráficos de barras. No conjunto de dados que dispomos, há alunos com 12, 13, 14 e 16 anos, mas não há alunos com 15 anos. O gráfico de barras de IDADE_categ não entende que a magnitude de salto de 14 para 16 é diferente da magnitude de salto de 12 para 13 e de 13 para 14, uma vez que interpreta estes números apenas como "rótulos".

Solução: usar gráficos propícios para variáveis numéricas, tais como gráfico de pontos (menu $Gráficos \rightarrow Gráfico \ de \ pontos$) ou gráfico de "hastes" (menu $Gráficos \rightarrow Plot \ discrete \ numeric \ variable$).

Explorando a janela R Script

- Linhas de comando são escritas sequencialmente na janela R Script a cada ação "finalizada" nos menus do R Commander.
- Todo início de linha de comando na janela *R Script* começa "colado" à margem.
- Linhas de comando muito compridas podem ocupar mais de uma linha na janela R Script. Quando uma linha <u>não</u> começa "colada" na margem, ela é continuação da linha acima - ou seja, ainda é a mesma linha de comando da linha acima.

- Para ilustrar, gere um gráfico de barras da variável REGIAO_RESID por meio do menu Gráficos → Gráfico de Barras (com título pertinente).
- Considere que o último comando executado antes do gráfico sugerido acima foi o gráfico de "hastes" das idades. As três linhas na figura abaixo representam o final da janela R Script, onde são ilustradas apenas duas linhas de comando, uma vez que a terceira linha é a continuação da segunda linha, formando uma única linha de comando.

```
With (Educa, discretePlot(IDADE, scale="frequency"))
with (Educa, Barplot(REGIAO_RESID, xlab="", ylab="Frequência",
main="Região de residência", label.bars=FALSE))
```

- Uma das vantagens da disponibilização de linhas de comando na janela *R Script* é a possibilidade de usar a "estrutura base" destas para realizar modificações pertinentes que <u>não</u> podem ser feitas pelos menus do R Commander.
- Marcando toda a linha de comando de interesse (como se fóssemos copiá-la) e clicando em seguida no botão Submeter (atalho: Ctrl + R), tal linha de comando modificada ou não é executada.
- Vamos ilustrar esta ideia modificando a cor das barras do último gráfico de barras gerado.

Antes dos dois "fecha parênteses" no final da linha de comando associada a este gráfico de barras na janela R Script, digite

de forma que fique algo do tipo

```
with(Educa, Barplot(REGIAO_RESID, xlab="", ylab="Frequência",
    main="Região de residência", label.bars=FALSE,col="red"))
```

Marque toda a linha de comando como se fosse copiá-la, mas em vez de $\lceil Ctr \rceil + \lceil C \rceil$, use $\lceil Ctr \rceil + \lceil R \rceil$ ou clique em Submeter. Por fim, confira na janela R Graphics (dentro da janela R Gui) o gráfico de barras com barras vermelhas.

E para deixar cada uma das três barras com uma cor diferente?

- No lugar de "red", usar um vetor de três cores.
- Por exemplo, no lugar de \[\, \text{col="red"} \], usar

```
,col=c("blue","yellow","green")
```

Para ver a lista com os nomes das 657 cores que o R disponibiliza, digite colors() na janela *R Script* e submeta-a, ou então digite-a na mesma janela que você digitou require(Rcmdr) ou library(Rcmdr) e aperte Enter.

Referências Bibliográficas

- Fox, J. (2022). RcmdrMisc: R Commander Miscellaneous Functions. R package version 2.7-2.
 - Fox, J., & Bouchet-Valat, M. (2022). Rcmdr: R Commander. R package version 2.8-0.
- Ihaka, R. (1998). R: Past and Future History. A Draft of a Paper for Interface '98. Statistics Department, The University of Auckland, Auckland, New Zealand. Disponível em https://www.stat.auckland.ac.nz/~ihaka/downloads/Interface98.pdf.
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Ribeiro Melo, F.R. (2019). Introdução ao R Commander Notas de Aula. Disponível em http://gae.uniriotec.br/7/Notas_de_aula_Rcmdr.pdf.
- Ribeiro Melo, F.R. (2019). R Commander: um pouco além dos menus gráficos. Disponível em http://gae.uniriotec.br/7/Rcmdr_alem_dos_menus.pdf.
 - Terrón, A., Cabellero, P. & Alcaraz, L. (2011). Estadística Básica con R-Commander. Disponível em http://www.bubok.es/libros/203887/Estadística-basica-con-RCommander.

MUITO OBRIGADO!!!







