

Visualização de Dados com o R - Versão Online

Aplicações para Finanças e Economia

Marcelo S. Perlin (marcelo.perlin@ufrgs.br)

2022-11-23

Visualização de Dados com o R

Aplicações para Finanças e Economia

por **Marcelo Scherer Perlin**

© 2022 Marcelo S. Perlin. Todos Direitos Reservados.

Publicação independente e impressa por demanda.

Versão online disponível em <https://www.msperlin.com/vdr/>

Capa: Rubens Lima - <https://capista.com.br>
ISBN (paperback): 9798849109626
ISBN (hardcover): 9798849170893
ISBN (ebook): -

Histórico de Edições:

2022-10-15 Primeira Edição

Embora o autor tenha boa fé para garantir que as instruções e o código contidos neste trabalho sejam precisos, ele se exime de toda responsabilidade por erros ou omissões, incluindo, sem limitação, a responsabilidade por danos resultantes do uso ou da confiança neste trabalho e em seus resultados. O uso das informações contidas neste trabalho é por sua conta e risco. Se qualquer código deste livro estiver sujeito a licenças de código aberto ou direitos de propriedade intelectual de terceiros, o cumprimento desses direitos e licenças é inteiramente de sua responsabilidade como usuário.

SUMÁRIO

Prefácio	9
Como ler o livro	10
Instalação do R e RStudio	11
Pacote vdr	12
Recursos na Internet	13
Exercícios	13
Convenções	14
Agradecimentos	17
1 Introdução	19
1.1 O Gráfico do Jeremy Siegel	20
2 Princípios de Visualização de Dados	25
2.1 Por que criamos uma figura?	26
2.1.1 Transmissão de informação	26
2.1.2 Manipulação da atenção	27
2.1.3 Independência do elemento gráfico	29
2.1.4 Herança e reprodutibilidade	29
2.2 Visualizando a inflação para o Brasil	31
2.3 Exercícios	35
3 Componentes de Gráficos	37

3.1	Elementos textuais	38
3.2	Elementos gráficos	41
3.3	A escolha dos canais	45
3.3.1	Uso de linhas	45
3.3.2	Uso de formas	45
3.3.3	Uso de barras	47
3.4	Cores	50
3.4.1	Paletas sequenciais	50
3.4.2	Paletas divergentes	56
3.4.3	Paletas qualitativas	58
3.5	Tamanho	59
3.6	Exercícios	61
4	Visualização de Dados com o R	65
4.1	Sobre o R	65
4.2	Pacote ggplot2	66
4.2.1	Alternativas ao ggplot2	66
4.3	O ciclo de criação de figuras	66
4.4	Dados de entrada	66
4.4.1	O formato dos dados	66
4.4.2	Tipos de dados	67
4.5	O sistema de camadas do ggplot2	67
4.6	Mapeamento de canais com aes()	67
4.6.1	Função geom_line()	67
4.6.2	Função geom_point()	68
4.6.3	Função geom_col()	68
4.6.4	Função geom_text()	68
4.7	Uma aplicação com dados reais	68
4.8	Paletas de cores	68
4.8.1	Aplicando uma paleta a um gráfico	69
4.9	Temas	69
4.9.1	Aplicando um tema a um gráfico	69
4.9.2	Módulos adicionais de temas	69
4.10	Painéis e facetas	70
4.11	Exportando figuras	70
4.12	Exercícios	70
5	Gráficos Estatísticos	73

5.1	Visualizando uma distribuição	73
5.1.1	Histogramas	73
5.1.2	Densidades	74
5.1.3	Quartis	74
5.2	Visualizando múltiplas distribuições	74
5.2.1	Boxplots	74
5.2.2	Uso de painéis/facetas	75
5.2.3	Pacote ggribges	75
5.3	Visualizando correlações	75
5.3.1	Pacote corrplot	75
5.4	Visualizando incerteza	75
5.5	Exercícios	76
6	Mapas	79
6.1	Pacote geobr	79
6.1.1	Mapa do Brasil	79
6.1.2	Mapa do Brasil com cidades e capitais	80
6.2	Pacote ggmap	80
6.2.1	Registrando um projeto	80
6.2.2	Utilizando ggmap	80
6.2.3	Codificação geográfica	81
6.3	Exercícios	81
7	Programando com o ggplot2	83
7.1	Criando figuras via funções	83
7.2	Testando temas	84
7.3	Testando paletas	84
7.4	Testando temas e paletas	84
7.5	Exercícios	84
8	Conclusão	89

LISTA DE FIGURAS

1.1	O gráfico de Jeremy Siegel (Siegel, 2015)	21
2.1	Gráfico de exemplo para a inflação Brasileira	31
2.2	Gráfico melhorado para a inflação Brasileira	33
2.3	Gráfico da inflação Brasileira para diferentes períodos	34
2.4	Publicação do reddit/dataisuglu em 27/09/2022	36
3.1	Esqueleto de um gráfico com elementos textuais	38
3.2	Esqueleto melhorado de um gráfico com elementos textuais	40
3.3	Os canais de visualização mais utilizados: posição, formato, tamanho, cores (Wilke, 2019)	42
3.4	Gráfico de linhas e pontos para os preços da PETR3	43
3.5	Gráfico melhorado de linhas e pontos para os preços da PETR3	44
3.6	Exemplo de gráfico de linhas	46
3.7	Exemplo de gráfico de formas	48
3.8	Exemplo de gráfico de barras	49
3.9	Exemplo de cinco paletas de cores sequenciais em tom único do pacote colorspace	51
3.10	Exemplo de paleta sequencial com tom único para um gráfico de barras	53

3.11 Exemplo de cinco paletas de cores sequenciais em tons múltiplos do pacote colorspace	54
3.12 Exemplo de paleta sequencial com tons múltiplos para um gráfico de barras	55
3.13 Exemplo de cinco paletas de cores divergentes do pacote colorspace	56
3.14 Exemplo de uso de paleta divergente em gráfico de dispersão	57
3.15 Exemplo de quatro paletas de cores divergentes do pacote colorspace	58
3.16 Exemplo de uso de paletas qualitativas	59
3.17 Exemplo de uso de tamanho para um gráfico de dispersão	60
3.18 Exemplo de uso de tamanho para diferenciar pontos	61
3.19 Postagem do LatinoMetrics em 22/09/2022	62
3.20 Postagem do Statspanda em 19/09/2022	63
3.21 Postagem do Reddit/dataisbeautiful	64

PREFÁCIO

Quem trabalha com análise de dados sabe a importância e o potencial de comunicação que um gráfico possui. **Figuras e ilustrações facilitam a transmissão de uma mensagem e o convencimento de um público.** Em um mundo cada vez mais digitalizado e visual, o resultado de uma análise pode ser compartilhado com o mundo inteiro em questões de segundos, ao clique de um botão.

Em contraste, raros são os cursos de universidades e instituições de pesquisa que oferecem algum tipo de treinamento formal na construção de gráficos baseados em dados. Por tentativa e erro, professores e alunos aprendem “na marra”, sem uma clara noção de estrutura, qualidade e aplicabilidade. Curiosamente, discussões sobre uso de *software* de análise de dados se assemelham a discussões sobre futebol em uma mesa de bar: muita paixão e pouca razão. Em uma triste inércia, grupos de pesquisa se utilizam de *softwares* defasados, implicando a proliferação de práticas de pesquisa individualizadas e pouco eficientes.

A versão inicial deste livro surgiu como uma extensão do capítulo 10 da minha outra obra, **Análise de Dados Financeiros e Econômicos com o R** (Perlin, 2021). Em uma leitura pós-publicação, meu sentimento foi que, enquanto o material anterior

se justificava no sentido de mostrar a mecânica de criação de figuras específicas da área de finanças e economia, faltava um conteúdo mais abrangente e completo sobre o tópico.

Neste livro, **Visualização de Dados com o R**, escrevo sobre tudo aquilo que aprendi em mais de doze anos de carreira como professor, cientista de dados e consultor de empresas. Primeiro, iremos estudar os fundamentos e princípios a serem seguidos na criação de figuras baseadas em dados. Em seguida estudaremos a forma de trabalho com o R. Ao longo do caminho, teremos diversos exemplos de visualização de dados reais.

Como ler o livro

O livro foca na didática e foi escrito para o público em geral. O formato foi escolhido para que o leitor consiga aplicar as técnicas no seu dia a dia. Todos os exemplos de código são reproduzíveis em seu próprio computador. O material, porém, exige algum conhecimento sobre programação com o R. Sempre que possível, tentei simplificar e destilar o funcionamento do código no texto.

A melhor maneira de utilizar o material do livro é replicando os exemplos apresentados em cada capítulo. Como falei antes, cada pedaço de código é auto-suficiente e foi criado para que rode no seu próprio computador.

Como sugestão, uma maneira didática de aprender as técnicas do livro é:

- 1) Procure primeiro entender o que o código está fazendo antes de utilizá-lo. Esse é o passo mais importante pois é assim que se aprende programação, construindo conexões entre o gráfico na sua cabeça, e aquele produzido pelo código no computador.
- 2) Replique o código original do exemplo no seu próprio computador, de preferência escrevendo o código e não copiando e colando via *Control+C*. A digitação vai facilitar a memorização de comandos e estruturas de procedimentos. Quando se sentir confortável em digitar os mesmos comandos, podes

atalhar o processo via cópia do código e colagem. Após a execução, confira se a figura obtida no seu próprio computador está idêntica a aquela apresentada no livro.

- 3) Modifique o código para que a figura tenha algo diferente (título, nome dos eixos, fontes e cores, ..). Assim memorizará onde cada elemento gráfico e textual pode ser modificado. Em outras palavras, a modificação de código já existente facilitará a criação de conexões entre o código e o gráfico de saída.
- 4) Utilize outras fontes de dados com o mesmo código.

A replicação de código vai ajudar a memorização da sintaxe do R e do **ggplot2** (Wickham et al., 2022a). No começo certamente será difícil, mas não se desmotive. Tudo que vale a pena é difícil. Com o tempo, irás pegar o jeito e absorver a necessária fluência para produzir figuras de forma autônoma.

Instalação do R e RStudio

O primeiro passo para aproveitar o material é instalar e configurar a sua máquina para reproduzir os códigos do livro. Abaixo destaco os programas necessários, ordem de instalação, e *links* para download de arquivos de instalação:

1. **R**: motor de programação, cuja instalação já inclui módulos básicos. Disponível gratuitamente em <https://www.r-project.org/>.
2. **R-Tools** (apenas usuários do Windows): ferramentas para a compilação de pacotes do R. Disponível em <https://cran.r-project.org/bin/windows/Rtools/>.
3. **RStudio** (opcional): Interface de programação voltada ao R. Apesar de não ser estritamente necessário, o RStudio facilita muito o trabalho, e recomendo a todos a instalação e o uso do programa. Disponível em <https://www.rstudio.com/products/rstudio/>.

Caso tiver algum problema na instalação, podes encontrar um tutorial completo no Capítulo 1¹ de (Perlin, 2021).

Pacote **vdr**

Todo o material construído neste livro é aberto ao público e reproduzível. Para facilitar a importação dos dados, construí um módulo do R chamado **vdr** (Perlin, 2022b), o qual pode ser instalado de forma prática e fácil. Com ele será possível importar dados para a sua sessão do R e rodar todos os exemplos no seu próprio computador.

Assumindo que já tens o R e RStudio instalados, vamos prosseguir com a instalação do pacote. Para isto, inicie o RStudio e execute as seguintes linhas de comando no *prompt*:

```
# package remotes is needed..  
install.packages('remotes')  
  
# install vdr from github  
remotes::install_github(  
  'msperlin/vdr',  
  dependencies = TRUE  
)
```

Pronto. Após a instalação, todos requisitos estarão disponíveis no seu computador. Reforço que o comando anterior já instala todos os módulos necessários para reproduzir os exemplos do livro.

Após a instalação, teste o carregamento do módulo com o seguintes comando:

```
library(vdr)
```

Pronto! Se tudo der certo, verás uma mensagem anterior informando que o pacote **vdr** foi carregado corretamente. Caso tenha tido uma mensagem de erro, refaça as etapas até que os comandos anteriores rodem sem problema.

¹<https://www.msperlin.com/adfeR/introducao.html#instalando-o-rstudio>

Recursos na Internet

Na internet é possível encontrar uma variedade de conteúdo sobre visualização de dados. Esta é uma eficiente maneira de se manter atualizado nas melhores práticas e buscar inspiração para futuros trabalhos. Abaixo destaco alguns links com material de alta qualidade e que merecem a sua atenção:

- Instagram – LatinoMetrics <https://www.instagram.com/latinometrics/?hl=en> – Grupo especializado em visualizações de dados da América Latina. Apresenta gráficos muito interessantes sobre o Brasil e outros países do continente americano.
- Instagram – Statspanda – Página especializada em visualizações de dados em diferentes tópicos. O conteúdo é direto e fácil de entender.
- Reddit – Dataisbeautiful – Fórum online do Reddit, voltado a agregação de conteúdo relativo a visualização de dados. Aqui é possível encontrar uma variada gama de gráficos e fontes relacionados a dados.

Exercícios

O final de cada capítulo apresenta exercícios a serem resolvidos pelo leitor. O gabarito dos exercícios, incluindo saída de código, está disponível através do pacote **vdr** e função **vdr::exercises_compile_solution()** :

```
dir_solution <- "~"
vdr::exercises_compile_solution(dir_output = dir_solution)
```

A execução do código anterior criará um arquivo *.html* na pasta “Documentos” (atalho de ~), com todas as soluções dos exercícios. Para abrir o arquivo, localize o mesmo na referida pasta e clique duas vezes. Deves abrir o arquivo com seu navegador de internet de preferência, tal como o *Chrome* ou *Firefox*. Alternativamente, caso não conseguir compilar o arquivo de soluções localmente, po-

des baixá-lo diretamente do meu site².

Convenções

No decorrer do livro usarei código ao longo do texto, como o apresentado a seguir:

```
# Code example
my_object <- 1:10

# print object
print(my_object)
```

```
R> [1] 1 2 3 4 5 6 7 8 9 10
```

O código anterior está mostrando na tela o conteúdo do objeto `my_object`, o qual contém uma sequência entre 1 e 10. A saída do código é registrada no texto com o símbolo `R>`, como em `R> [1] 1`. Verás este mesmo padrão por todo o livro.

Além disso, note o uso de comentários com o *hashtag* (`#`). Estes não são compilados pelo interpretador do R e servem para escrever texto no código. Note também que usarei a língua inglesa para definir comentários e código em si. Esta decisão pode não ser muito intuitiva, porém todo código deve respeitar padrões internacionais e o uso de uma língua comum, neste caso o inglês. No papel de instrutor, entendo que é importante aprender a maneira correta desde o início e, por isso, minha escolha por escrever códigos em inglês. Mas, não se assuste, o vocabulário necessário é reduzido e acessível.

Ao longo do livro usaremos diversas funções, de diferentes módulos. Para organizar o texto e facilitar o entendimento da origem das funções, estas serão indicadas com o formato nativo do R. Por exemplo, função `geom_line` do pacote `ggplot2` será representada como **`ggplot2::geom_line()`**. Isto é, o pacote de origem é indicado primeiro, seguido pelo símbolo de separação (`::`), seguido pelo nome da função. Para evitar repetições desnecessárias, o

²<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

nome do pacote aparece somente na primeira menção da função no texto.

Espero que goste e aproveite o conteúdo do livro. Esta e demais obras são parte de um projeto pessoal na montagem de um conteúdo prático sobre o uso do R e, assim, popularizar o uso da plataforma na produção de ciência. Caso tiver alguma dúvida sobre o conteúdo, sugestões ou reclamações, sinta-se convidado a entrar em contato comigo pelo email *marceloperlin@gmail.com*.

Boa leitura.

Marcelo S. Perlin, 23/11/2022.

AGRADECIMENTOS

Visualização de dados sempre foi uma das minhas paixões em pesquisa científica. Registro aqui minha satisfação pessoal em poder entregar esse projeto para um público maior. Escrever um livro técnico exige uma grande dedicação e esforço individual do autor.

Este livro não seria possível sem a devida autonomia do meu cargo de professor. Assim, deixo aqui o meu agradecimento a UFRGS (Universidade Federal do Rio Grande do Sul), por possibilitar e incentivar este empreendimento no mercado literário. Agradeço também todos os meus colegas professores, os quais sempre me incentivaram a escrever livros.

Adicionalmente, não posso também deixar de agradecer a toda a comunidade do R. Em especial, agradeço os autores dos pacotes **bookdown** (Xie, 2022a) e **ggplot2** (Wickham et al., 2022a), sem os quais não seria possível compilar este livro. Adicionalmente, abaixo destaco os respectivos pacotes disponíveis no CRAN utilizados e suas devidas referências. A lista foi gerada automaticamente e está em ordem alfabética.

artyfarty (Smeets, 2022), **base** (R Core Team, 2022), **bbplot** (Stylianou et al., 2022), **bookdown** (Xie, 2022a), **colorblindr** (McWhite and Wilke, 2022), **colorspace** (Ihaka et al., 2022), **conflicted** (Wickham, 2021), **corrplot** (Wei and Simko, 2021),

cowplot (Wilke, 2020), **DiagrammeR** (Iannone, 2022), **dplyr** (Wickham et al., 2022b), **dviz.supp** (Wilke, 2022a), **forcats** (Wickham, 2022a), **fs** (Hester et al., 2021), **geobr** (Pereira and Goncalves, 2022), **ggmap** (Kahle et al., 2022), **ggplot2** (Wickham et al., 2022a), **ggpubr** (Kassambara, 2022), **ggrepel** (Slowikowski, 2022), **ggridges** (Wilke, 2022b), **ggtech** (Bion, 2018), **ggtext** (Wilke and Wiernik, 2022), **ggthemes** (Arnold, 2021), **gt** (Iannone et al., 2022), **hrbrthemes** (Rudis, 2020), **httr** (Wickham, 2022b), **knitr** (Xie, 2022b), **lubridate** (Spinu et al., 2022), **paletteer** (file., 2022), **patchwork** (Pedersen, 2022), **plotly** (Sievert et al., 2022), **purrr** (Henry and Wickham, 2022), **renv** (Ushey, 2022), **rmarkdown** (Allaire et al., 2022), **scales** (Wickham and Seidel, 2022), **sf** (Pebesma, 2022), **stringr** (Wickham, 2022c), **tibble** (Müller and Wickham, 2022), **tidyr** (Wickham and Girlich, 2022), **tidyverse** (Wickham, 2022d), **vdr** (Perlin, 2022b), **yfR** (Perlin, 2022a)

CAPÍTULO 1

INTRODUÇÃO

Visualização de dados é a arte de transformar medidas em padrões visuais. O objetivo é incrivelmente simples, **apresentar evidências de um evento ou causalidade do mundo real**. E, como qualquer arte, é fácil de descrever e difícil de executar. Na prática, construir figuras baseadas em dados exige *expertise* da área de conhecimento, noções de *design* e conhecimento das ferramentas corretas de programação. O retorno para o autor, porém, é imenso.

Impressões visuais nunca foram tão importantes quanto hoje. A popularidade de sites¹ como *Instagram*, *Twitter* e *Facebook* mostram o potencial digital que uma figura pode apresentar. Assim, o seu conteúdo de visualização de dados em um relatório técnico ou trabalho acadêmico pode ser facilmente distribuído, com uma chance de viralizar e lhe promover em sua área de trabalho.

Ao mesmo tempo, como uma faca de dois gumes, gráficos também podem ser utilizados para a manipulação e distorção da realidade.

¹Já peço desculpas antecipadas se estás lendo este livro e os sites citados não existem mais. Quem lembra do Orkut?

Como um analista de dados, seu trabalho também será identificar tais casos. Ao aprender a criar figuras baseadas em dados, entenderás as escolhas realizadas na produção do gráfico e, inevitavelmente, um possível despreparo ou malícia do autor.

No lado do mercado de trabalho, ser a **pessoa que saber montar gráficos** é uma maneira direta e factível de se diferenciar como profissional. Muitos sabem criar um gráfico em planilhas eletrônicas, por exemplo, mas poucos conseguem ir além e apresentar ilustrações de dados que geram um real impacto ao olho, com a apresentação de informações visuais que facilitam a entrega de uma mensagem. A especialização nesta arte, com o correto uso das ferramentas de programação, certamente lhe diferenciárá.

1.1 O Gráfico do Jeremy Siegel

Para entender o impacto de uma figura baseada em dados, vamos olhar o caso do Jeremy Siegel – professor americano especializado no tópico de investimentos. Em seu livro “Investindo em Ações no Longo Prazo: O Guia Indispensável do Investidor do Mercado Financeiro” (Siegel, 2015), Jeremy apresenta nas primeiras páginas uma figura que impactou o mundo dos investimentos. Veja o gráfico em questão na Figura 1.1. Tome alguns bons segundos na análise das informações disponíveis ali. De antemão, perceba que cada configuração do gráfico foi uma decisão do autor.

Ao longo do livro, o autor defende e justifica o investimento financeiro em empresas como sendo o mais rentável no longo prazo. A principal mensagem do livro, e também da figura, é a superioridade de performance financeira quando uma pessoa investe em empresas no mercado de ações. Em resumo, quando olhamos um investimento no longíssimo prazo, mais de 10 anos, as empresas oferecem o maior retorno financeiro entre as alternativas.

Voltando ao gráfico em si, Figura 1.1, as seguintes classes de investimentos são apresentadas:

Stocks (ações) investimento em empresas através das ações negociadas em bolsa. Utiliza-se do índice SP500 como refe-

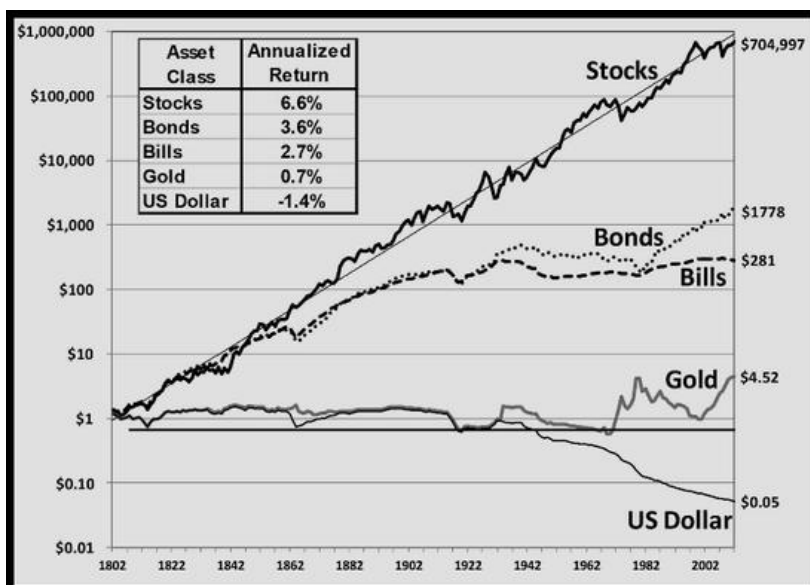


Figura 1.1: O gráfico de Jeremy Siegel (Siegel, 2015)

rência para investimentos em renda variável. Em outras palavras, o investidor que investe no índice irá receber uma média de desempenho das 500 ações que o compõem. O investimento equivalente para o Brasil seria o índice Ibovespa.

Bonds (renda fixa de longo prazo) investimento em dívida governamental com prazo maior do que um ano. Traduzindo para o Brasil, seria o equivalente a um investimento no Tesouro Direto, isto é, compra de dívida do governo federal.

Bills (renda fixa de curto prazo) dívidas governamentais com duração menor do que um ano. Aqui não temos uma tradução direta, mas pode-se entender como equivalente a nossa conta poupança, um veículo para investimentos de curto prazo, com bastante liquidez.

US Dollar (dólares americanos) Compra de dólares físicos. A equivalência para o Brasil seria a compra de notas físicas de Reais.

O gráfico funciona pois demonstra, de forma clara, a mensagem do livro: **investir em ações é a alternativa mais lucrativa no longo prazo**. Já descontando a inflação, alguém que tivesse investido 1 dólar no índice de ações em 1802, teria observado o seu

patrimônio crescer para mais de 704,997 dólares em 2004.

Destrinchando a construção do gráfico, nota-se os seguintes pontos e decisões dos autores:

- 1) **Todos os diferentes investimentos – dólar, ações, dívida pública americana — começam com um dólar.** Esta é uma manipulação simples de dados, com o cálculo do retorno acumulado para cada investimento. A escolha padroniza a análise entre as alternativas, e facilita a interpretação. Para verificar a diferença de performance entre cada possibilidade de investimento, basta observar o valor final obtido. Assim, para alguém simular um investimento inicial diferente, é suficiente multiplicar o valor final do investimento pelo montante investido. Por exemplo, 100 dólares investidos em ações em 1802, se tornariam 704 milhões de dólares em 2004.
- 2) **O retorno acumulado no investimento em ações é ressaltado no gráfico, com o uso de uma linha com uma cor mais intensa.** Claramente é uma estratégia visual para passar a mensagem implícita da dominância de ações sobre os demais investimentos. Na primeira batida de olho no gráfico, a linha de ações chama a atenção e já sugere o resultado da análise. Note que, se fosse o contrário, com a linha de ações menos espessa que as demais, a mensagem não seria tão óbvia. O autor mesclou, portanto, o aspecto financeiro de investimento com um padrão gráfico que suporta o mesmo argumento: “o investimento em ações resulta em algo maior do que as alternativas”.
- 3) **A escala do eixo vertical é não-linear.** Veja que os intervalos entre os pontos não são os mesmos. O primeiro ponto do eixo é 1, depois 10, depois 100, até chegar em 1.000.000. A razão é simples, quando colocado em um espaço de gráfico não-linear, os investimentos em ações se assemelham a uma linha reta. Veja, porém, que a linha reta é um resultado ilusório. Se a escala fosse espacialmente correta, os investimentos em renda fixa e dólar não apareceriam no gráfico devido ao alto valor no resultado para ações. E isso distorce-

ria a análise, pois seria impossível enxergar ambas as linhas no mesmo espaço do gráfico.

- 4) **O valor final do investimento é apresentado como texto na direita do gráfico.** Isso facilita a interpretação para o leitor pois não precisa olhar para o eixo da esquerda para identificar o valor final de cada alternativa.
- 5) A tabela no lado esquerdo superior do gráfico mostra o **retorno anualizado de cada investimento**, isto é, o retorno equivalente em um ano que resulta no desempenho observado no gráfico. Em comparação, as ações dominam os demais investimentos, com 6,6% de retorno real anualizado.

Estes aspectos do gráfico são de total responsabilidade do autor. Podes ter certeza que o mesmo ponderou cada decisão de configuração do gráfico até chegar no estado atual. Diversas versões devem ter sido criadas, e o que funciona foi mantido, e o que não funciona foi eliminado.

Para se ter uma ideia do impacto deste gráfico, o nome do autor é praticamente sinônimo do gráfico. Ao procurar imagens relacionadas a busca “Jeremy Siegel” no *Google Images*, o resultado mostra esta mesma imagem nas mais diversas páginas da internet. Indo além, ao realizar uma busca reversa, isto é, usando a própria imagem como ponto de partida da busca na internet, o Google retorna 3,720,000,000 resultados, incluindo sites de notícias, blogs, bancos e corretoras de investimentos.

Resumindo, o gráfico do livro do Siegel tomou vida própria e pode ser considerado um dos materiais mais impactantes do autor, talvez até mais que o livro em si! Não tenho dúvida que, mesmo desatualizada – o dados utilizados acabam em 2004 – a imagem continuará sendo utilizada nas próximas décadas como um exemplo da mensagem do livro. O grande sucesso do autor foi compilar uma única figura que transmitisse, de forma simples e direta, a mensagem principal de sua obra.

Meu objetivo principal quando escrevi os capítulos do livro *Visualização de Dados com o R* é mostrar **como atingir o mesmo impacto dentro de sua área de trabalho**. Talvez não consiga a

capilaridade histórica do gráfico do professor Siegel, mas terás conhecimento para criar visualizações de dados que impressionam e transmitem a informação de forma clara. Assim, nossa primeira etapa é entender o que torna os gráficos visualmente atrativos e impactantes, o que funciona e o que não funciona, assim como as ferramentas para construí-los no seu próprio computador. Para começar, no próximo capítulo vamos primeiro buscar entender os princípios por trás da criação de figuras.

CAPÍTULO 2

PRINCÍPIOS DE VISUALIZAÇÃO DE DADOS

O que torna a figura de Jeremy Siegel tão impactante? Qual foi o processo de criação da figura? Neste capítulo vamos responder estas questões através da discussão de princípios e componentes visuais de um gráfico de dados. Discutiremos elementos centrais e independentes da plataforma. Ou seja, podes aplicar os mesmos princípios para gráficos produzidos em outras linguagens de programação.

Este é um tema complexo e existem ótimas referências a serem consultadas. Aqui focaremos na prática de criação de figuras, deixando a teoria de lado. Como referência complementar, recomendando os seguintes livros:

- **Fundamentals of Data Visualization** (Wilke, 2019), disponível gratuitamente em <https://clauswilke.com/dataviz/>.
- **Data Visualization - A practical introduction** (Healy, 2018), disponível gratuitamente em <https://socviz.co/>.

2.1 Por que criamos uma figura?

Primeiro, e mais importante, é preciso reforçar que **a razão principal da análise de dados é a comunicação de ideias**. De forma simplificada, é seu trabalho analisar informações retiradas de um banco de dados e sugerir possíveis efeitos no mundo real. Um relatório técnico é nada mais do que uma opinião profissional sobre determinado problema. O mesmo é verdadeiro para um trabalho acadêmico, onde se testa uma hipótese baseada nos dados, ou o estudo de uma decisão empresarial, tal como maneiras de se obter maior impacto em campanhas de marketing.

Assim sendo, um sólido parecer técnico pode ter seu impacto ofuscado por uma baixa capacidade de comunicação (Schwabish, 2014). O poder de um argumento é multiplicado pela sua capilaridade, a facilidade com que o mesmo pode ser transmitido para o seu público. Veja que, de nada adianta realizar um excelente trabalho técnico na análise de dados, se a comunicação escrita e visual não transmite a mensagem desejada. De fato, um dos erros frequentes que encontro na revisão de trabalhos acadêmicos é o foco na técnica, deixando de lado a *venda do peixe*. Enquanto a parte técnica e científica é certamente importante, a comunicação do material potencializa o impacto do conteúdo para com seu público.

2.1.1 Transmissão de informação

O primeiro princípio na criação de gráficos é que **uma figura deve justificar a sua existência e conversar com a sua audiência**. Um erro muito comum por parte de alunos é tentar criar os mais variados gráficos sem se perguntar se os mesmos adicionam informações novas na análise, ou se o público irá entendê-los. Só porque você pode fazer um gráfico, não significa que você deve mostrá-lo. O valor de um conteúdo está diretamente relacionado às novas informações que ele traz na análise, e a capacidade do público em entendê-las.

Ao decidir sobre os elementos de um gráfico, atenha-se àqueles que ajudam a transmitir a sua mensagem. Não hesite em cortar

material. Em um documento escrito, sempre que você encontrar uma figura que não seja sendo discutida em pelo menos dois parágrafos do texto, não tenha receio em retirá-la do documento. Se não consegue escrever mais do que dois parágrafos sobre uma figura, provavelmente não é importante.

Uma figura conversa com o seu público quando todos os elementos do gráfico convergem para a mensagem. Assim, utilize todas as ferramentas disponíveis – cores, formas, tamanhos e transparência – para transmitir uma mensagem clara e objetiva. Na prática, sugiro as seguintes ações:

- Use cores e formas para distinguir grupos dentro dos dados. Se um grupo é mais importante do que outros, diferencie o mesmo usando uma cor ou forma específica;
- Use cores para transmitir sentimentos e adjetivos tais como bom/ruim, alto/baixo, lucro/prejuízo;
- Use tamanhos, cores e a transparência de linhas e formas para indicar graus de importância nos resultados. Por exemplo, se estiver comparando a inflação de países diferentes em relação ao Brasil, realce o caso brasileiro no gráfico com uma cor ou densidade mais forte;

2.1.2 Manipulação da atenção

Use e abuse de formas visuais de chamar a atenção. Em outras palavras, facilite e direcione a análise para o seu público, sempre verificando se os gráficos produzidos indicam uma mensagem clara e direta. Destaque nos gráficos o que o público deve procurar e como lê-lo. Como regra de bolso, **o melhor gráfico é aquele que não precisa ser explicado.**

Assim, entendas o limite do conhecimento e expectativas do seu público. Ao comunicar uma figura, não espere que todos tenham o mesmo conhecimento técnico que você. O que é óbvio para você, que está semanas (ou anos) trabalhando no mesmo problema, pode ser confuso para outros. Entenda o que seu público espera e qual a motivação para ler o seu conteúdo. Por exemplo, não

apresente para um grupo de executivos o mesmo material que apresenta para o seu orientador acadêmico. Cada qual possuem diferentes demandas e características, e irão avaliar o trabalho de acordo com suas expectativas. Não é incomum um gráfico ser elogiado pelo público em geral, e rechaçado pelo público técnico.

A forma de digestão visual de um gráfico é importante. No momento em que um gráfico é analisado, os olhos do leitor irão percorrer o material sistematicamente. Enquanto as pessoas podem processar dados visuais de forma diferentes, a sequência de avaliação tende a seguir a forma como lemos, de cima pra baixo. Assim, os seguintes elementos são avaliados em sequência:

- 1) título;
- 2) subtítulo;
- 3) miolo da figura (linhas e formas), incluindo cores, formas e transparências;
- 4) textos dos eixos horizontais e verticais;
- 5) outros elementos do gráfico.

Assim, use a sequência dos elementos em uma narrativa sobre a mensagem. O título e subtítulo são elementos extremamente importantes pois definem a primeira mensagem do gráfico. Assim, use textos claros e sugestivos no título e subtítulo do gráfico, os quais devem indicar como analisar o miolo da figura.



Um erro comum de professores que trabalham em consultoria é tratar o público corporativo da mesma forma que o universitário. Minha experiência sugere que o público corporativo, focado em problemas empresariais, está mais interessado no **impacto real**, isto é, como o que está sendo apresentado afetará a prática da empresa, em linhas simples e fáceis de entender. Para este tipo de público, o foco deve ser em clareza das informações e sugestão de ações reais. Evite, assim, justificativas técnicas e científicas que o público provavelmente não entenderá e/ou necessitará de muito tempo para explicar. Lembre que o público corporativo não está interessado em questionar as suas decisões técnicas, mas sim solucionar um problema.

2.1.3 Independência do elemento gráfico

Todas as informações técnicas, como origem e período de dados, devem ser claramente indicadas no título, subtítulo ou legenda do gráfico. Se o público precisar buscar informações adicionais para entender o gráfico, então existe espaço para melhoria. É mais fácil dizer do que fazer, mas tente comunicar o máximo de informações possíveis, desde que não polua o gráfico. Lembre-se de que existe um sensível equilíbrio entre uma estética elegante, e os detalhes técnicos que o seu público espera. Assim sendo, use e abuse de elementos textuais tais como títulos e subtítulos para acrescentar informações relevantes ao entendimento do problema. Em resumo, busque criar gráficos que podem ser entendidos sem explicação adicional, onde os elementos textuais – títulos, subtítulos, eixos e legendas – fornecem todas as informações necessárias para se entender a figura.

2.1.4 Herança e reprodutibilidade

A ciência e a análise de dados evoluem na forma de blocos de construção, um encima do outro. Sempre verifique os gráficos produzidos em suas referências de texto. Eles guiarão sobre o que seu público espera. Da mesma forma, você pode até usar

figuras de artigos anteriores para comparar seus resultados, o que é especialmente conveniente quando o mesmo conjunto de dados são usados. Por isso, sempre informe qual a origem dos dados utilizados no gráfico, facilitando que outra pessoa, ou até mesmo você mesmo, replique o gráfico no futuro.

Ao escrever um código de criação de figuras, tenha a certeza de que o mesmo é facilmente reproduzível no futuro. A melhor versão de um código de visualização de dados é aquela que, com mínimas modificações, executa corretamente daqui a cinco anos. Assim, faça as minhas sugestões:

- Evite qualquer trabalho manual para a criação de gráficos. Plataformas de programação e análise de dados, tal como o R e o Python, facilitam todo o processo;
- Automatize o processo de importação e manipulação de dados para que, daqui a alguns anos, possa replicar o gráfico com dados atualizados;
- Escreva códigos que se adaptem aos dados, e nunca o contrário. Se alguma modificação nos dados de entrada exige modificação no código, então vale a pena revisar o trabalho. Ao manter um código que se adapta aos dados, fica mais fácil manter a reprodutibilidade da análise ao longo do tempo.



Uma maneira eficiente de transitar o mesmo material entre diferentes públicos é primeiro criar uma versão básica do gráfico, e depois adaptar o conteúdo às demandas de cada público. Assim, a partir do mesmo modelo de gráfico, podes modificar os elementos facilmente em cada cenário. Convenientemente, o uso do R e pacote ggplot2 facilita este processo pois se utiliza de um sistema de camadas, onde a primeira camada é a versão básica e, conforme necessidade, adicionamos e removemos camadas de acordo com o público.

2.2 Visualizando a inflação para o Brasil

Agora que já entendemos a teoria, vamos para a prática. No gráfico a seguir, Figura 2.1, é apresentado uma série de **inflação histórica** para o Brasil, retirada do sistema de séries temporais do Banco Central do Brasil.

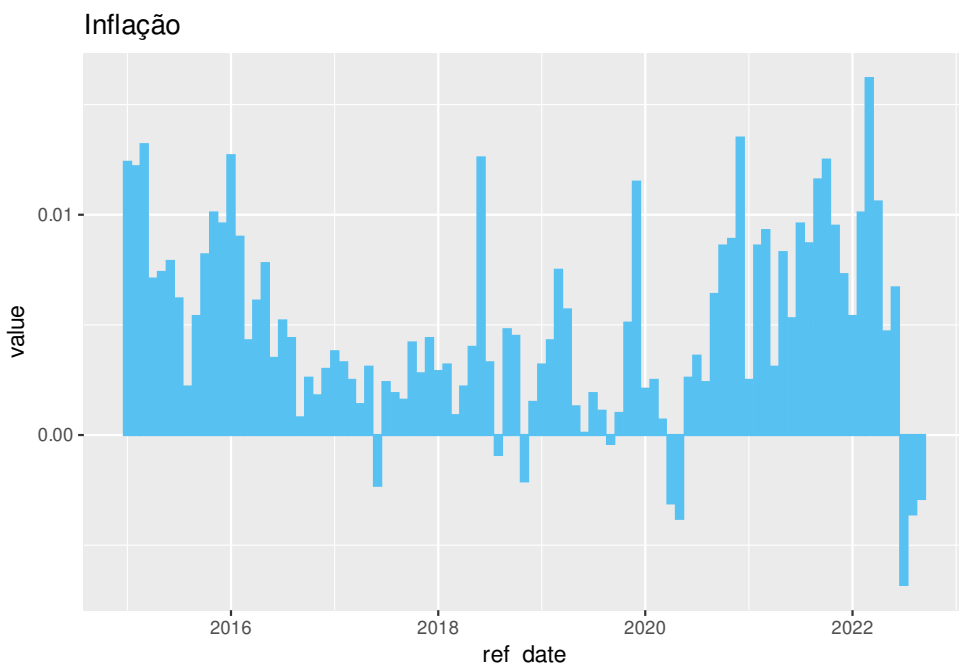


Figura 2.1: Gráfico de exemplo para a inflação Brasileira



Sobre os Dados

A inflação refere-se ao aumento sistemático de preços em um país e é geralmente definida como o aumento de preço em uma cesta de produtos. Observar e entender a inflação é tarefa de todo economista: a inflação tem efeitos negativos em toda a economia, corroendo o poder de compra de consumidores e desincentivando o consumo. No Brasil, o índice de inflação mais popular é o IPCA, calculado e comunicado pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

O gráfico anterior mostra o IPCA (*Índice de Preço ao Consumidor Amplo*) mensal entre 2015 e 2022. A olho nu, ler este gráfico e retirar alguma conclusão é tarefa quase impossível:

- Qual o objetivo do gráfico? Por que o mesmo está sendo mostrado?
- O que os valores do eixo vertical representam exatamente?
- De onde os dados saíram e como o gráfico pode ser replicado?
- Mais importante, qual a mensagem transmitida?

Entender o gráfico anterior requer tempo e esforço cerebral. Como regra, toda vez que um gráfico forçar a audiência a pensar demais, existe um alto risco do material ser irrelevante para o mundo real. Na dificuldade do entendimento da figura em menos de cinco segundos, provavelmente o leitor irá ignorar a imagem e prosseguir para o próximo elemento do texto. Aposto que você, ao observar o gráfico, fez a mesma coisa. Ou seja, o gráfico rompe dois princípios básicos de visualização de dados, a justificativa de existência e manipulação da atenção. Uma versão melhorada é apresentada a seguir, Figura 2.2.



Um erro muito comum de iniciantes é utilizar o *template* original do gráfico, isto é, exportar o primeiro gráfico criado, sem muito esforço na construção de algo atrativo. Saiba que um olho treinado e experiente conhece muito bem o *template* básico e isso vai pesar em qualquer avaliação. Além de mostrar certa preguiça por parte do criador, o qual simplesmente aceitou a primeira versão do gráfico, também reforça a falta de capacidade na utilização das ferramentas de programação.

Comparando a primeira com a segunda versão, vemos as seguintes diferenças:

- 1) **Elementos textuais** – o título e o subtítulo em 2.2 já indicam o que estamos analisando no gráfico e qual a mensagem, neste caso a variabilidade da inflação mensal. Em outras palavras, estamos analisando como a inflação varia ao longo do

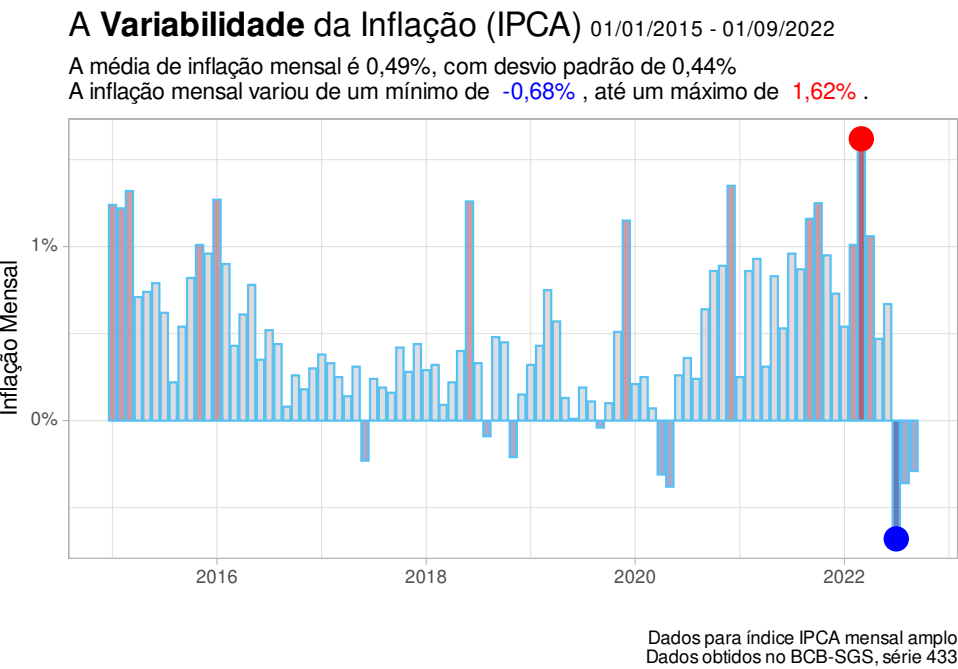


Figura 2.2: Gráfico melhorado para a inflação Brasileira

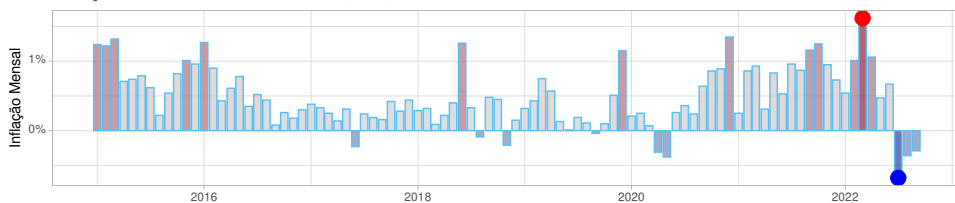
tempo, com seus altos e baixos. Note também a variação do tom da letra do título com o uso do negrito e tamanho da fonte para reforçar o objetivo do gráfico. Abaixo, na direita, vemos a origem e a identificação dos dados, neste caso o sistema de séries temporais do Banco Central do Brasil (BCB-SGS), série de número 433. Dado que a origem dos dados é pública, qualquer pessoa pode acessar e reproduzir o mesmo gráfico.

- 2) **Elementos gráficos** – Uso de cores para distinguir os períodos de alta inflação dos períodos de baixa, reforçando e salientando a variação da inflação entre os meses. Círculo azuis e vermelhos mostram os pontos máximos e mínimos da inflação mensal no período analisado.
- 3) **reprodutibilidade** – Uma mudança nos dados não exige mudança no código pois todos os elementos são baseados nos dados de entrada. Para demonstrar a reprodutibilidade do gráfico, a seguir, Figura 2.3, apresentamos o mesmo gráfico

para três períodos diferentes.

v01 A Variabilidade da Inflação (IPCA) 01/01/2015- 01/09/2022

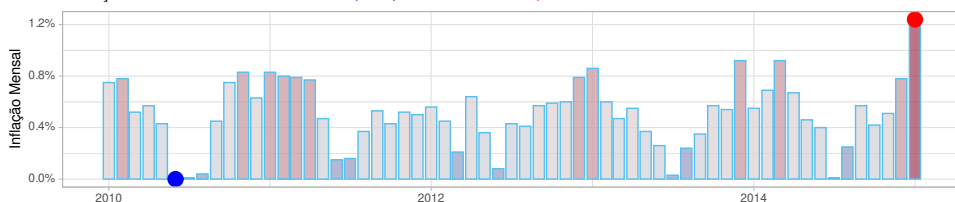
A média de inflação mensal é 0,49%, com desvio padrão de 0,44%
A inflação mensal variou de um mínimo de -0,68% , até um máximo de 1,62% .



Dados para índice IPCA mensal amplo
Dados obtidos no BCB-SGS, série 433

v02 A Variabilidade da Inflação (IPCA) 01/01/2010 01/01/2015

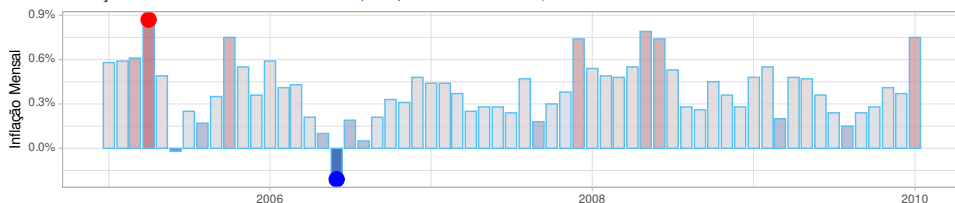
A média de inflação mensal é 0,51%, com desvio padrão de 0,26%
A inflação mensal variou de um mínimo de 0,00% , até um máximo de 1,24% .



Dados para índice IPCA mensal amplo
Dados obtidos no BCB-SGS, série 433

v03 A Variabilidade da Inflação (IPCA) 01/01/2005 01/01/2010

A média de inflação mensal é 0,39%, com desvio padrão de 0,20%
A inflação mensal variou de um mínimo de -0,21% , até um máximo de 0,87% .



Dados para índice IPCA mensal amplo
Dados obtidos no BCB-SGS, série 433

Figura 2.3: Gráfico da inflação Brasileira para diferentes períodos

Note as semelhanças e as diferenças entre os painéis de 2.3. Além dos elementos de barra e cores, todas as datas e os valores do texto são calculados a partir dos dados. Mais importante, depois de montado o código, a criação dos três gráficos é muito fácil. Reforçando, o esforço para construir um gráfico com reprodutibilidade se concentra na concepção inicial. Depois do código pronto, a reutilização é eficiente e instantânea.

2.3 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, podes baixar uma versão do arquivo compilado no site do livro¹.

01 - O grupo LatinoMetrics produz e distribui um conteúdo muito interessante de visualização de dados econômicos para a América Latina. Observando o material do Instagram, visualize as seis últimas imagens disponibilizadas na página principal. Observando as figuras como um todo, destaque os elementos comuns na criação das imagens. Isto, é, destaque os elementos visuais que foram repetidos entre uma figura e outra.

02 - No Reddit é possível encontrar o grupo `r/dataisugly`, o qual contém inúmeros posts sobre visualizações de dados realizadas da forma errada. Na data de 27/09/2022 foi publicado a seguinte mensagem no fórum:

Analise o gráfico e, sem buscar a resposta no fórum, indique qual o problema com o gráfico.

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

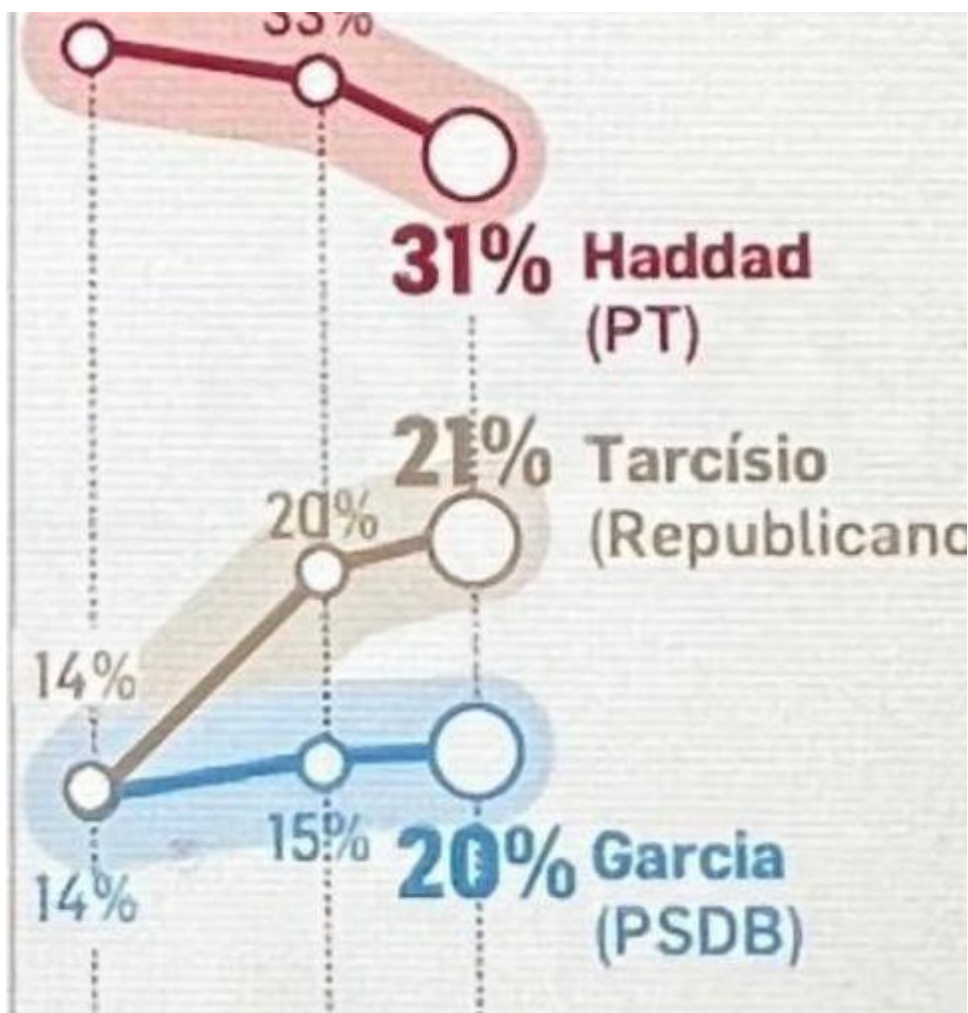


Figura 2.4: Publicação do reddit/dataisuglu em 27/09/2022

CAPÍTULO 3

COMPONENTES DE GRÁFICOS

Os componentes de uma figura baseada em dados separam-se entre textuais e gráficos. Os **elementos textuais** referem-se a textos explícitos que auxiliam o entendimento da mensagem. Pense neles como o esqueleto do gráfico, suportando toda a estrutura visual controlada pelos dados, tal como pontos e linhas. Isto inclui textos dos eixos, legendas, títulos e subtítulos.

Os **elementos gráficos** são aqueles que representam formas geométricas relacionadas com os dados em si. Isto é, quando novos dados são inseridos, os elementos gráficos mudam. Estes são canais de visualização, os quais traduzem medidas numéricas ou categóricas em elementos gráficos na forma de linhas, pontos, formas e cores.

Com esta separação entre elementos textuais e gráficos, vamos montar uma figura sequencialmente, etapa por etapa, explicando cada elemento no processo.

3.1 Elementos textuais

Figura 3.1 apresenta os componentes textuais de um gráfico, incluindo títulos, subtítulos e texto do eixo horizontal e vertical. Este é um típico exemplo de **gráfico de dispersão**, o qual mostra a relação entre duas variáveis, x (coordenada horizontal) e y (coordenada vertical).

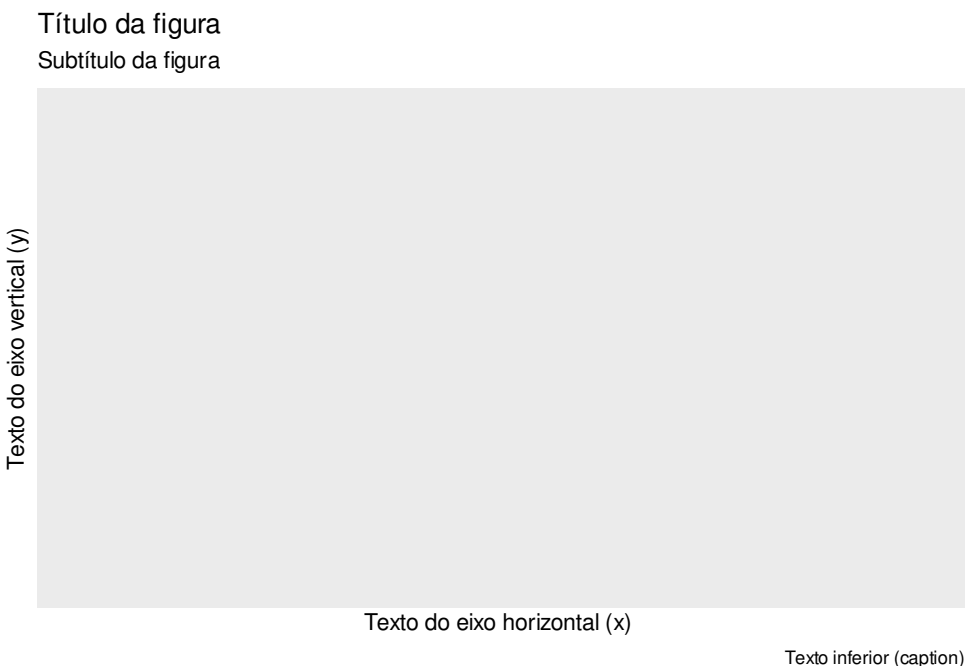


Figura 3.1: Esqueleto de um gráfico com elementos textuais

Os componentes da Figura 3.1 são:

Título Texto que inicia o gráfico e provavelmente será o **primeiro elemento a ser lido pelo leitor**. Busque usar poucas e informativas palavras, sem detalhes técnicos e focando em um público ingênuo. Ao mesmo tempo, use o título para introduzir o tema e a mensagem do gráfico. Por exemplo, o título “Performance Financeira das Ações da Petrobrás” é melhor e mais fácil de entender do que “Preços ajustados a dividendos e outros eventos corporativos, entre 2010 e 2020”. Todo conteúdo técnico e específico sobre os dados utilizados deve ser colocado no subtítulo ou texto inferior (*caption*).

Subtítulo da Figura O subtítulo é um dos pontos mais importantes sobre um gráfico, e onde um olhar técnico irá focar. O subtítulo pode, por exemplo, oferecer descrições do tamanho e forma de coleta da amostra de dados. Saiba que o subtítulo é um ponto muito indicativo da capacidade técnica do autor. Por exemplo, a informação de como os dados foram calculados, tal como “Dados ajustados a dividendos e outros eventos corporativos”, pode ser colocada no subtítulo. Como sugestão, procure não repetir informações já disponíveis em outros locais e busque sempre oferecer o máximo de informações para entender o gráfico.

Texto do eixo vertical (y) Texto explicativo no eixo vertical, definindo a principal variável de interesse. Relembre que, em um gráfico de dispersão, lê-se como “variável y é afetada por variável x”. Portanto, y é a variável que merece maior atenção nos componentes textuais, tal como título e subtítulo.

Texto do eixo horizontal (x) Texto correspondente ao eixo horizontal. Geralmente utiliza-se algo como *tempo* ou outra variável de interesse. Em um gráfico de barras, por exemplo, o eixo x pode ser algum tipo de grupo existente nos dados (exemplo: masculino/feminino). Reforço que a variável do eixo horizontal é secundária em relação a variável do eixo vertical.

Texto inferior (caption) É o texto de menor fonte, geralmente indicado no lado direito e abaixo do gráfico, com informações secundárias sobre a origem dos dados brutos ou sobre o autor, incluindo link para verificação. Por exemplo, “Dados obtidos em <https://www.transparencia.rs.gov.br/>”.



Quando formatando os elementos textuais de um gráfico, um exercício que costumo fazer é tentar prever quais questões sobre o gráfico serão perguntadas pelo público técnico. Se alguma informação não está clara e for fácil de inserir, não hesite em fazê-lo. Da mesma forma, antes de publicar um gráfico, apresente o mesmo para diferentes pessoas e preste atenção nas respostas, posteriormente ajustando o código onde necessário.

Um exemplo trabalhado de elementos textuais de um gráfico, porém ainda sem incluir elementos dinâmicos tal como pontos ou linhas, é apresentado a seguir, Figura 3.2.



Figura 3.2: Esqueleto melhorado de um gráfico com elementos textuais

Note que, mesmo sem adicionar os dados em si, o esqueleto já diz quais informações podemos esperar no gráfico: a performance da ação da Petrobrás ao longo do tempo, obtidos do Yahoo Finance. Os elementos textuais são formatados de acordo com os dados de entrada. Saiba que os anos do título – 2015 a 2022 – são obtidos diretamente dos dados. O efeito que procuramos é que, a cada nova entrada de dados, os elementos textuais mudem automaticamente.



Sobre os Dados

As ações de uma empresa, tal qual Petrobrás, são contratos financeiros que permitem a participação no controle e resultado (lucro) de uma empresa. Em resumo, ao comprar uma ação, passas a ser um dos donos da empresa, com direito a recebimento de participação proporcional ao lucro da mesma. Estes contratos são negociados diariamente no mercado financeiro, com forte variação de preços. Todas ações são identificadas por um símbolo chamado *ticker*, composto por quatro letras e números. A ações do tipo ordinária da Petrobrás, por exemplo, são identificadas por PETR3, enquanto ações do tipo preferencial são representadas pelo ticker PETR4.

3.2 Elementos gráficos

Os componentes gráficos – miolo da figura – são os canais visuais vinculadas ao dados, incluindo:

- Linhas (sólida, tracejada, ..);
- Formas (círculos, triângulos, quadrados, ..);
- Cores (azul, vermelho, verde, ..);
- Tamanhos (1mm, 5mm, ..).

Os mais comuns, e fáceis de lidar, são os gráficos com linhas e formas, tal como um círculo ou quadrado. Ao adicionarmos mais camadas ao gráfico, mais complexo ele fica. Na prática, usamos a interação entre os canais para reforçar uma mensagem. Por exemplo, se temos grupos dentro dos dados, podemos usar linhas diferentes (tracejada, sólida, etc) para separar cada grupo. **Os gráficos mais impactantes são aqueles em que usamos o conhecimento da área para construir a interação dos canais visuais.**

Figura 3.3 mostra os principais canais de visualização de dados. A primeira, **posição**, usa o próprio espaço para indicar relações entre duas variáveis, x e y , e é tipicamente utilizada em gráficos de dispersão. A segunda, **formato**, representa a forma geométrica a

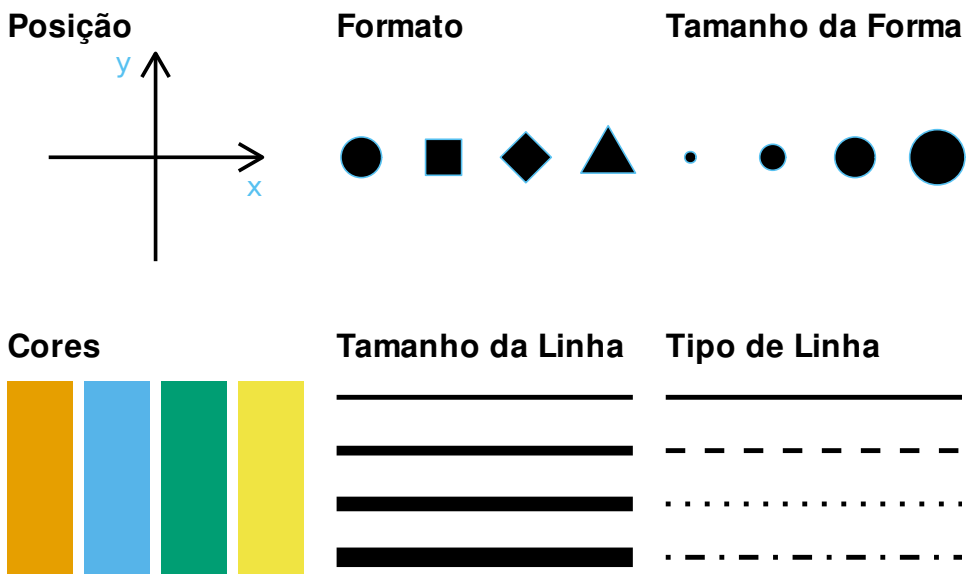


Figura 3.3: Os canais de visualização mais utilizados: posição, formato, tamanho, cores (Wilke, 2019)

ser inserida nas coordenadas, seja um círculo, quadrado ou outra. Podemos também usar o canal **tamanho**, o qual controla a escala espacial da forma geométrica utilizada.

Na criação de figuras, raramente utilizaremos apenas um canal de visualização. O mais comum é mesclarmos o uso de posição, formato e cores para salientar um resultado. Assim, usamos diferentes canais para representar as diferentes dimensões dos dados analisados.

Como um primeiro exemplo, vamos reconstruir o gráfico anterior, Figura 3.2, adicionando uma camada gráfica com linhas e pontos no gráfico:

O gráfico em 3.4 se utiliza dos canais visuais posição, linhas e pontos para construir a figura. Os pontos são dados pelas coordenadas x e y, enquanto as linhas conectam os pontos.

Agora, com a adição de um objetivo de comunicação ou mensagem do gráfico, podemos usar outros elementos visuais para salientar informações a respeito dos dados. Vamos assumir que o objetivo do gráfico é analisar a rentabilidade das ações da Petrobrás

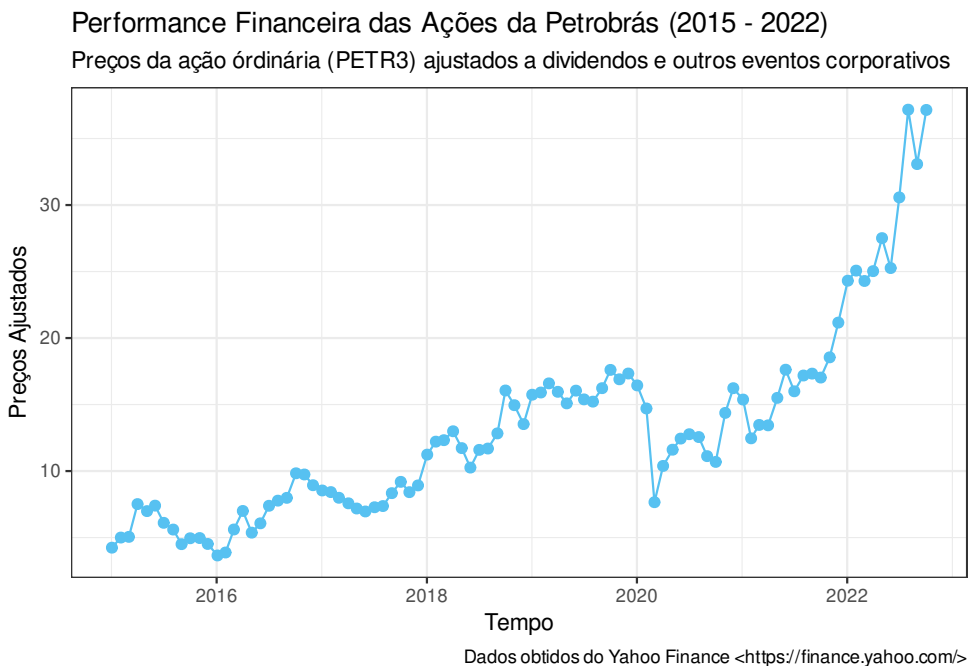


Figura 3.4: Gráfico de linhas e pontos para os preços da PETR3

(PETR3) no período. Para isso, é intuitivo e tecnicamente correto comparar o retorno acumulado da ação contra o índice Ibovespa, um *benchmark* para investimentos em empresas. O que procuramos entender é se as ações da Petrobrás tiveram performance superior a uma média do mercado como um todo.

Note que, apesar de serem exatamente os mesmos dados, as escolhas dos canais impactam diretamente o gráfico e a percepção do leitor. Observe os seguintes pontos adicionais:

- Ao contrário de se utilizar valores nominais de preço da Petrobrás e valores do índice Ibovespa, calculou-se o retorno acumulado partindo de 100% (linha cinza horizontal). Isso permite a comparação visual entre o comportamento de ambos investimentos.
- A linha **escura e densa** dos preços da ação PETR3 se destaca em relação a linha **tracejada e menos densa** do Ibovespa. Esta escolha é para atrair a atenção do leitor e, implicitamente, mandar a mensagem de que “Petrobrás é maior e

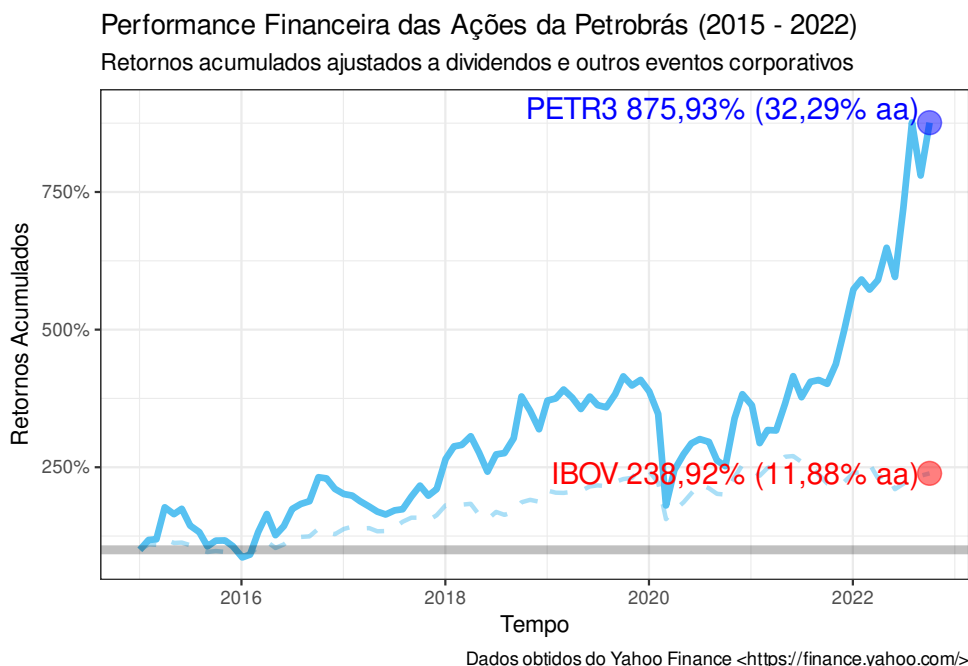


Figura 3.5: Gráfico melhorado de linhas e pontos para os preços da PETR3

mais forte do que o índice Ibovespa”.

- O texto com o retorno acumulado total e o equivalente anual foi adicionado ao gráfico. A cor azul para PETR3 destaca a melhor performance das ações da Petrobrás. A escolha das cores reflete princípios de contabilidade onde, a cor azul é relacionada com lucro, enquanto a vermelha é prejuízo. Ou seja, azul indica algo financeiramente positivo, e vermelho algo negativo.

Em resumo, os elementos em 3.5 foram escolhidos para reforçar a mensagem do gráfico: as ações da Petrobrás tiveram, no período analisado, melhor performance financeira do que o índice Ibovespa. Enquanto as ações da Petrobrás retornaram 775,93% de lucro para o investidor, o índice retornou apenas 138,92%. Este é um exemplo simples e direto de como podemos melhorar uma visualização de dados ao adicionar informações precisas e claras para o leitor com a seleção de tamanho, tipo de linha, cores, transparência e adição de textos.

3.3 A escolha dos canais

Partindo do caso mais simples, a primeira decisão na construção de um gráfico baseado em dados é qual a forma de traduzir números em imagens. Podemos usar linhas, pontos ou barras. Apesar de existirem regras gerais, alguns gráficos fazem mais sentido com linhas, enquanto outros com pontos ou colunas. O uso de cores no gráfico também é discutível: enquanto um pouco de cor pode ajudar o leitor, o uso de muitas cores pode atrapalhar.

3.3.1 Uso de linhas

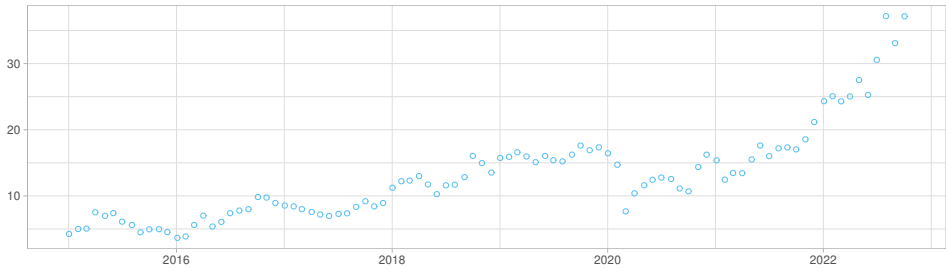
O uso de **linhas** faz sentido quando os dados adjacentes tem conexão entre si, tal como o próprio tempo. Por exemplo, no caso anterior, visualizamos os preços da ação da Petrobrás com linhas. Faz sentido conectar os pontos dos preços das ações pois o resultado final representa o caminho do preços ao longo dos dias, onde pontos adjacentes são conectados pelo tempo.

Veja o exemplo na Figura 3.6, onde apresentamos o gráfico de preços da EGIE3 – empresa do setor elétrico – em três formatos diferentes. O primeiro gráfico, usando pontos, é razoável. O segundo é somente estranho e mais parece um código de barras, sem muita relação com o problema em si. Como regra, o uso de barras sempre força o eixo vertical a partir do valor zero, o que pode resultar em gráficos pouco atrativos. Como esperado, o terceiro gráfico, painel C, é o que tem forma mais intuitiva: linhas para preços de ações, representando a dependência temporal entre um preço e outro.

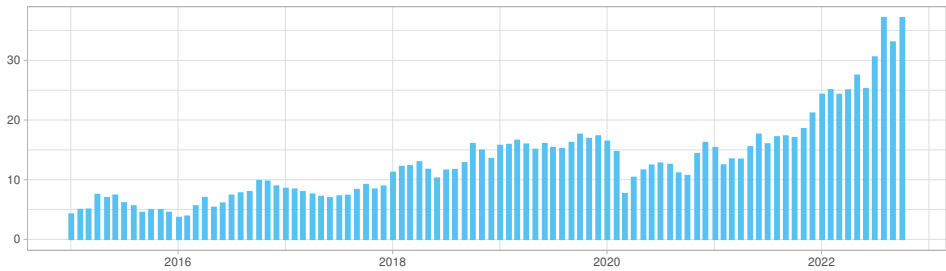
3.3.2 Uso de formas

O uso de **pontos e formas** fazem mais sentido em gráficos onde cada ponto pode ser entendido como **independente** dos demais. Este é o caso clássico de **gráficos de dispersão**, onde buscamos explicar uma variável numérica com base em outra. Por exemplo, considere analisar a relação entre sobra de salário (*disposable income*) e poupança (*savings ratio*) para diferentes países, Figura 3.7. Cada ponto do gráfico é um país diferente. Aqui, os dados

A Usando Pontos



B Usando Barras



C Usando Linhas

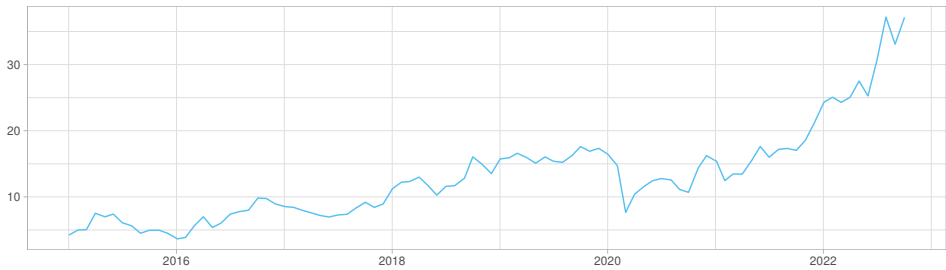


Figura 3.6: Exemplo de gráfico de linhas

do Brasil, por exemplo, não tem relação direta com os dados da Espanha. Assim, não faz sentido ligar os dados com linhas, mas sim usar pontos independentes.



Sobre os Dados

Os dados de razão de poupança (*savings ratio*) referem-se a divisão do valor poupado e o salário. O mesmo é calculado por país, tirando uma média regional de salário e poupança em determinado período. Em geral, quanto maior a razão de poupança para determinado país, melhor. Isso indica uma superior capacidade da população em poupar os seus rendimentos para o consumo futuro, aumentando a estabilidade financeira da região.

Comparando as imagens da Figura 3.7, note que o gráfico de barras (painel B) e linhas (painel C) resultam em figuras sem muita intuição e são difíceis de entender. Isso ocorre por que os canais de visualização com barras e linhas não funcionam para o problema. O uso de linhas, por exemplo, reforça uma dependência inexistente entre os pontos adjacentes e confunde a análise. Todavia, painel A – uso de pontos – deixa claro na visualização de que pessoas com renda maior tem uma proporção de poupança menor do que pessoas com renda menor.

3.3.3 Uso de barras

Os **gráficos de barras** funcionam muito bem quando a variável explicativa é uma categoria. Por exemplo, utilizando os mesmos dados anteriores, considere comparar a capacidade de poupança dos cinco países com maior valor, Figura 3.8.

Painéis A e B apresentam resultados razoáveis e intuitivos. Particularmente, o ordenamento dos dados faz com o que painel B tenha o formato de uma escada (ou pódio de premiação), o que facilita a mensagem do gráfico de que país com identificador 1 tem maior poder de poupança. Note também que o uso de barras (painel B) força o eixo y a iniciar em zero. Painel C é difícil de entender pois transmite a mensagem de interligação entre os paí-

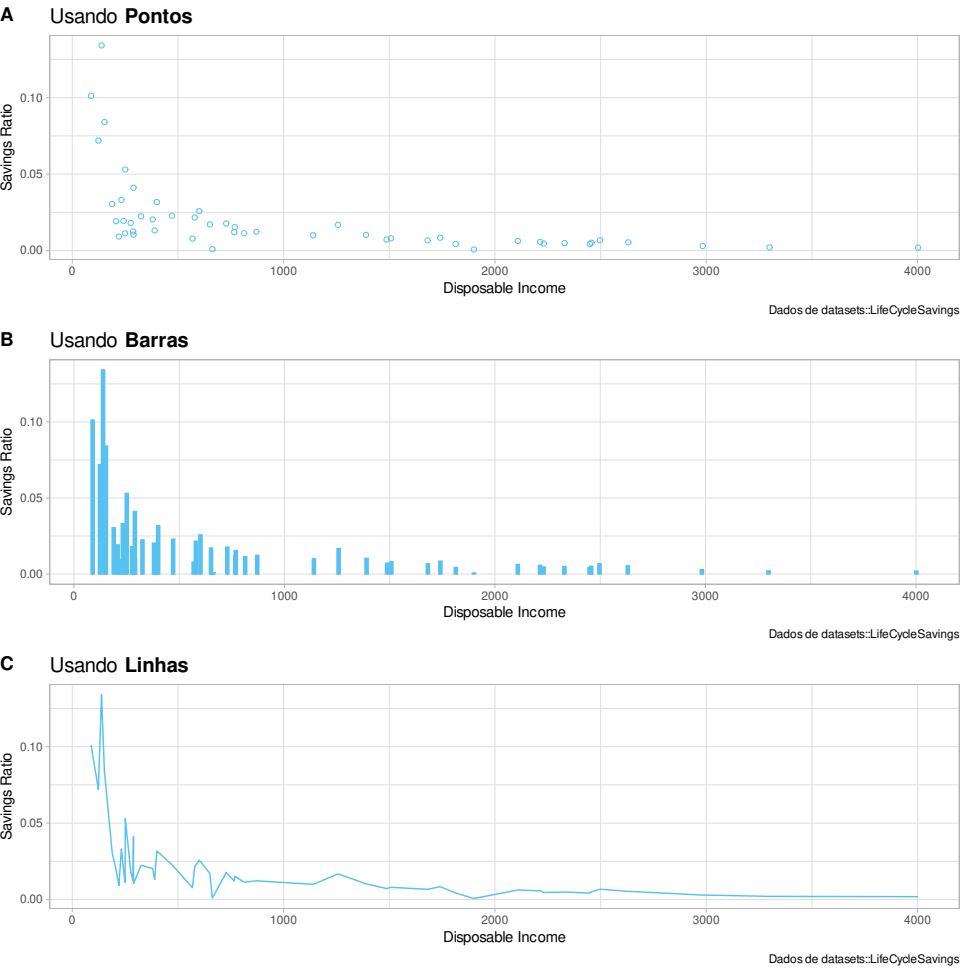


Figura 3.7: Exemplo de gráfico de formas

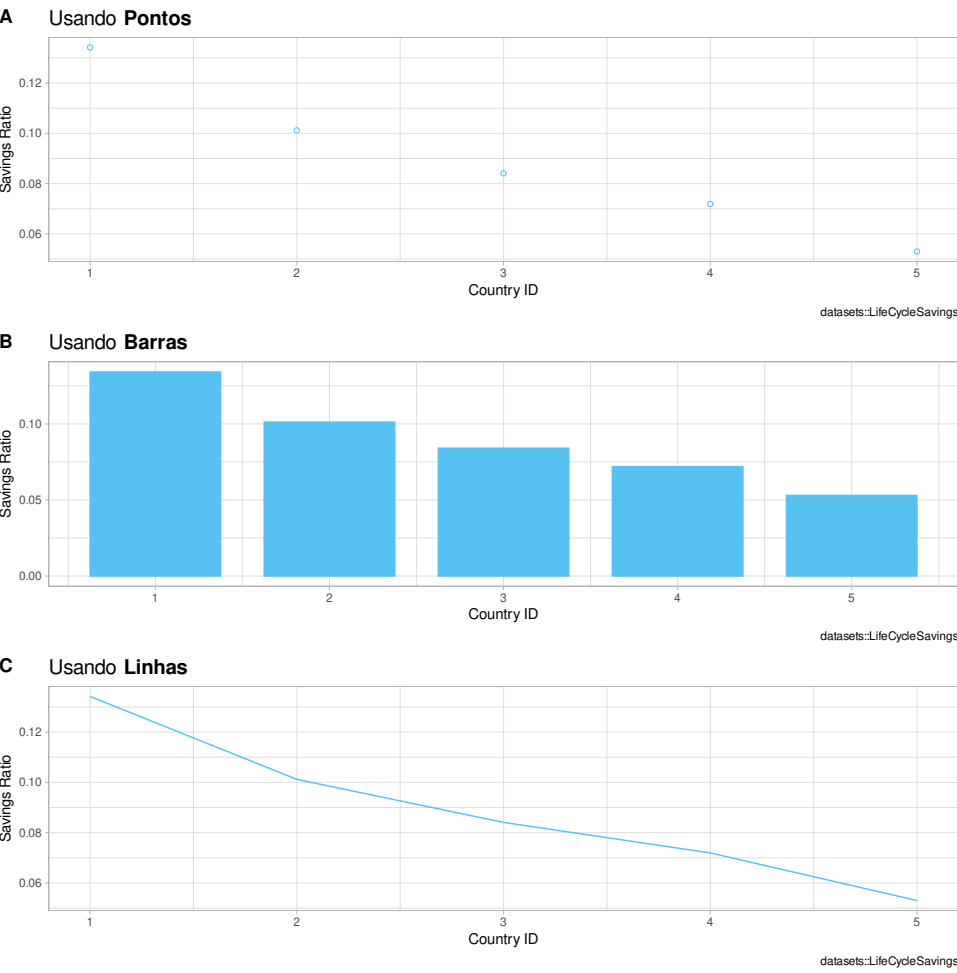


Figura 3.8: Exemplo de gráfico de barras

ses (eixo x), o que, dada a natureza dos dados, não é visualmente intuitivo.

3.4 Cores

Muito utilizado pela indústria de marketing, as **cores** direcionam a atenção e incitam algum tipo de sentimento pelo leitor. Na construção de figuras baseadas em dados, o uso de cores serve para distinguir grupos ou intensidades. Quando usado com parcimônia, as cores funcionam muito bem e facilitam o entendimento da mensagem do gráfico. Cuidado, porém, com excessos.

É indiscutível que o uso de cores funciona. Um exemplo clássico sobre o efeito do uso de cores em finanças é o artigo *In the red: The effects of color on investment behavior* (Bazley et al., 2017). Os autores conduziram experimentos para verificar se o uso da cor vermelha em relatórios de investimentos mudaria o comportamento futuro dos investidores. Os resultados mostraram que, quando a cor vermelha é utilizada para destacar prejuízos, os investidores se tornam mais avessos ao risco e diminuem as suas expectativas de retorno futuro. Perceba como uma pequena decisão nas cores pode causar um grande impacto na audiência.

No dia a dia, a melhor forma de trabalhar com cores no R é se utilizar de paletas – conjunto de cores pré-estabelecidas – e deixar que o próprio *software* escolha as cores específicas do gráfico. O uso de paletas automatiza o processo de construção de um gráfico pois as cores são predeterminadas em uma fila e selecionadas a medida que são necessárias. Adicionalmente, cada paleta de cores possui um tema, tal como quente/frio, claro/escuro, facilitando a escolha de acordo com o objetivo da mensagem. Para melhor entender o uso de paletas, vamos explorar os três tipos de paletas de cores: sequenciais, divergentes e qualitativas.

3.4.1 Paletas sequenciais

As paletas de cores sequenciais servem para indicar diferentes intensidades para um mesmo efeito. Pode-se usar apenas um tom ou

múltiplos para criar a paleta sequencial. Por exemplo, podemos usar vermelho escuro para representar dados em uma extremidade da distribuição, e vermelho claro para a outra. Como regra, as cores desse tipo de paleta tem o mesmo tom, com luminosidade decrescente ao longo das cores.

3.4.1.1 Paletas sequenciais de tom único

A seguir, Figura 3.9, apresenta-se cinco paletas sequenciais de tom único, proveniente do pacote **colorspace** .

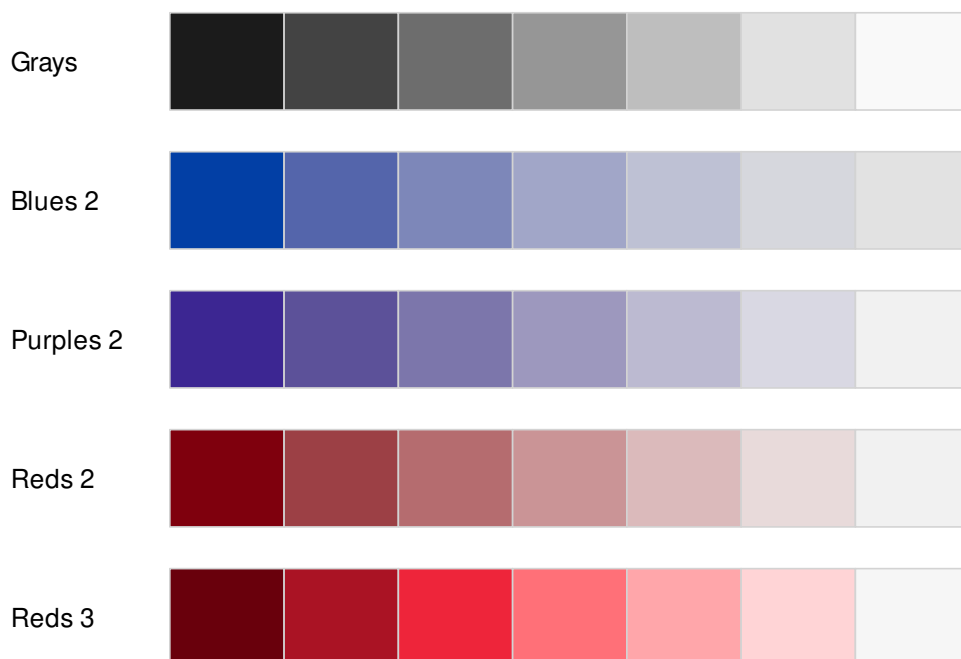


Figura 3.9: Exemplo de cinco paletas de cores sequenciais em tom **único** do pacote **colorspace** .



Entender a acessibilidade e a mídia é importante! Muito cuidado com a escolha das cores para uma audiência maior. Pessoas com incapacidade de diferenciar cores podem ter problema em analisar os seus gráficos. Se possível, tenha em mente uma maneira alternativa – usando formas de pontos, por exemplo – para formatar a sua figura. Adicionalmente, busque saber como o material será consumido pelo público. Por exemplo, livros são geralmente impressos em preto e branco e, portanto, qualquer uso de cores sequer será notado.

Todas as paletas em 3.9 se utilizam do mesmo tom (*single hue*), diminuindo a intensidade sequencialmente, da esquerda para a direita. Os nomes (“Grays”, “Blues 2”, ..) são pré-definidos pelo módulo de paletas. Futuramente, seção 4.8 do livro, iremos mais a fundo no uso de paletas do pacote **colorspace**.

Na Figura 3.10 apresenta-se um exemplo de aplicação de paleta sequencial com tom único para um gráfico de barras. Note como os países com maiores razões de poupança são salientados pela maior intensidade das cores.

3.4.1.2 Paletas sequenciais de tons múltiplos

Em 3.11 apresenta-se paletas sequenciais com tons múltiplos (*multi-hue*). Neste caso, além da mudança da intensidade, muda-se também as cores nos extremos.

O uso de paletas sequenciais com tons múltiplos resulta em gráficos mais elegantes do que para tons únicos. Veja a seguir, Figura 3.12, um exemplo de aplicação de tons múltiplos para o mesmo gráfico de barras anterior.

O resultado em 3.12 é elegante para qualquer paleta de cores. Particularmente, a paleta “Blues 2” me parece mais apropriada por utilizar a cor azul, a qual é geralmente relacionada com coisas positivas. Assim, a mensagem reforçada é de que países com maior razão de poupança são mais positivos.

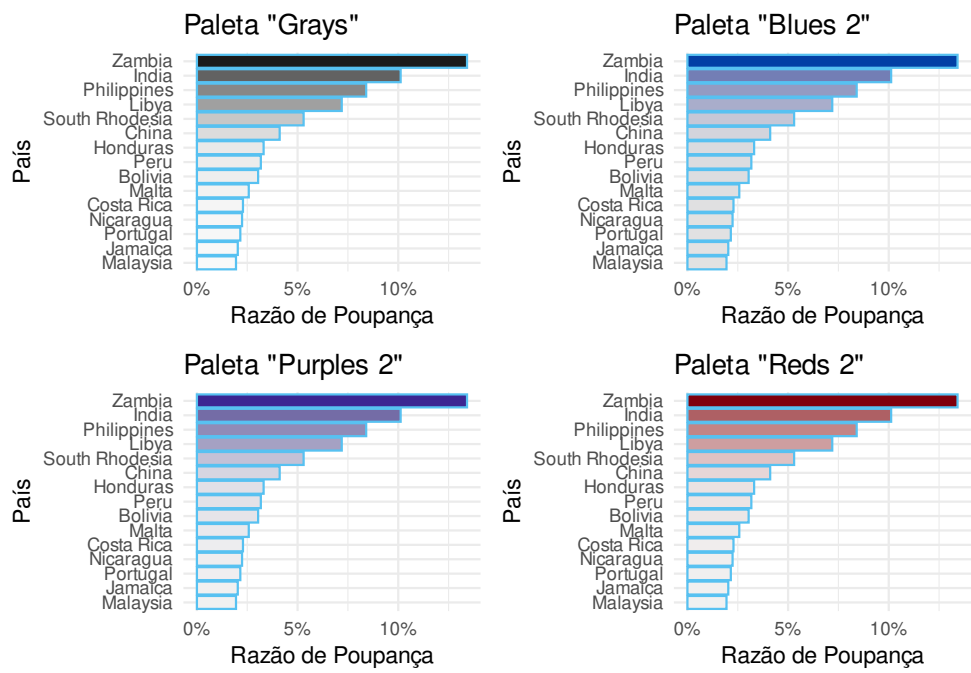


Figura 3.10: Exemplo de paleta sequencial com tom único para um gráfico de barras

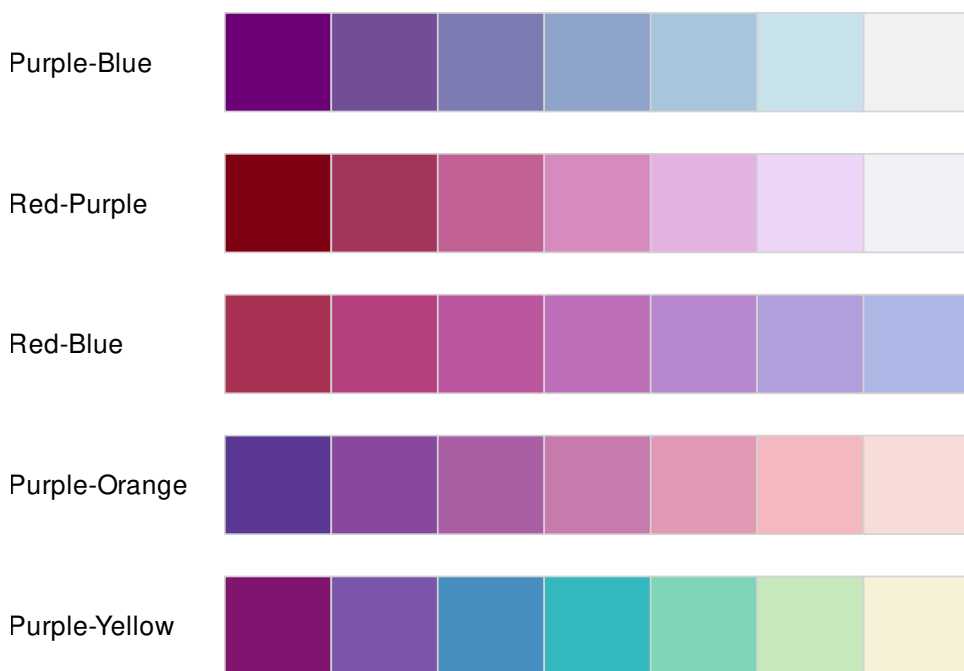


Figura 3.11: Exemplo de cinco paletas de cores sequenciais em tons **múltiplos** do pacote **colorspace** .

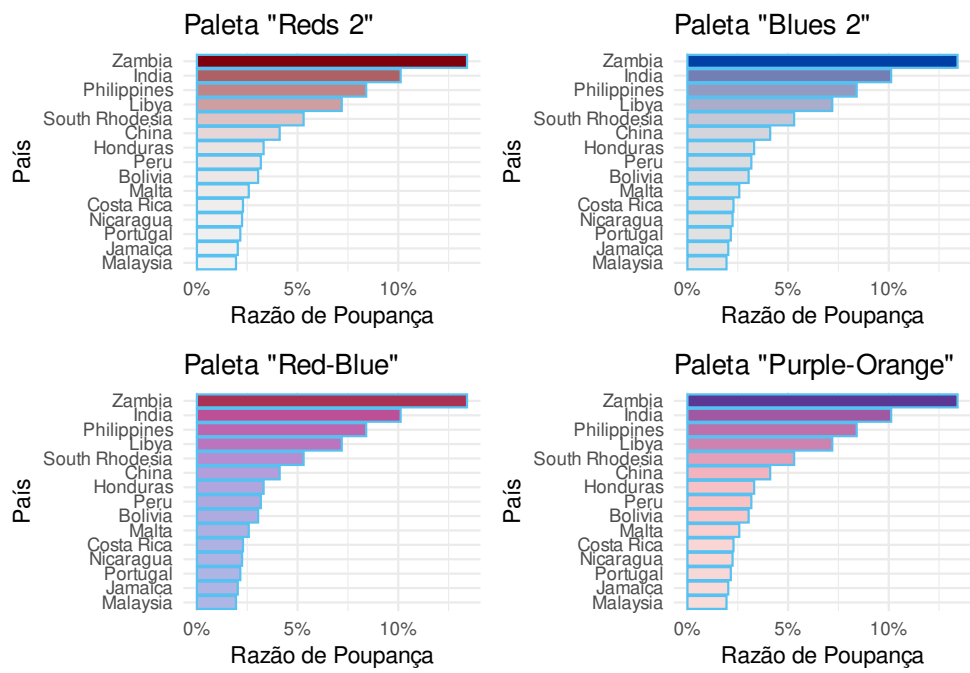


Figura 3.12: Exemplo de paleta sequencial com tons múltiplos para um gráfico de barras

3.4.2 Paletas divergentes

Paletas de cores divergentes realçam a diferença entre os extremos e servem para indicar intensidades. Definimos uma cor para o mínimo, outra para o máximo e todas as cores intermediárias são escolhidas como um ponto intermediário das cores nos extremos. A finalidade das paletas divergentes é destacar as diferenças entre os valores extremos dos dados. Veja na Figura 3.13 um exemplo de cinco paletas de cores divergentes do pacote **colorspace** .

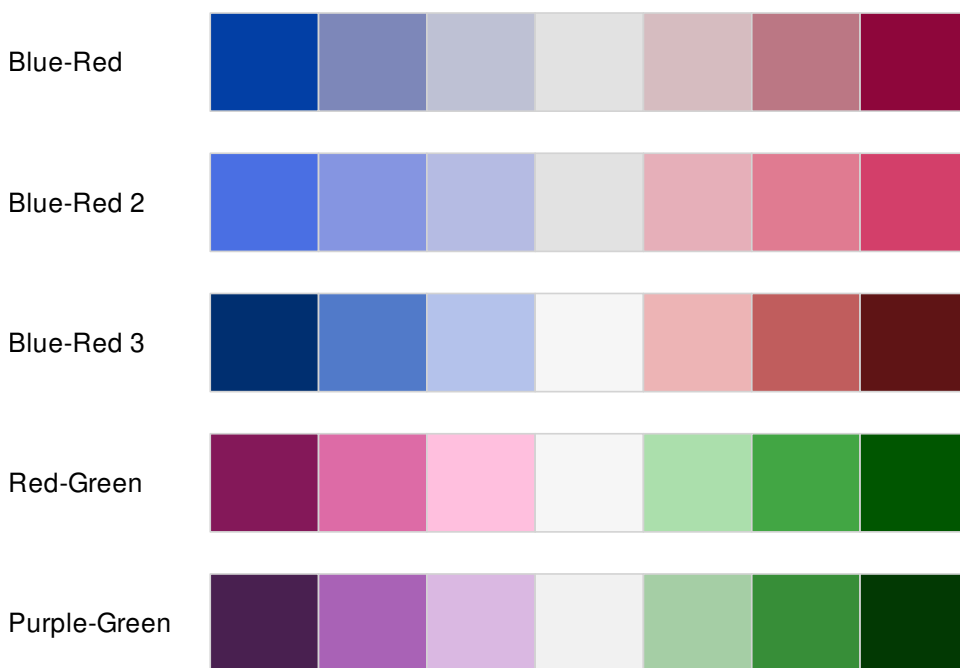


Figura 3.13: Exemplo de cinco paletas de cores divergentes do pacote **colorspace** .

Um exemplo de uso de paletas divergentes é o gráfico de dispersão dos dados de poder de poupança, Figura 3.14, onde utilizamos a média da razão de poupança como a cor do centro. Assim, as diferenças da média definem os extremos utilizados pela paleta escolhida.

Para a Figura 3.14, os pontos próximos da média possuem tom cinzento em todos os gráficos. A medida que os valores mudam de um extremo para o outro, as cores se distinguem de acordo com

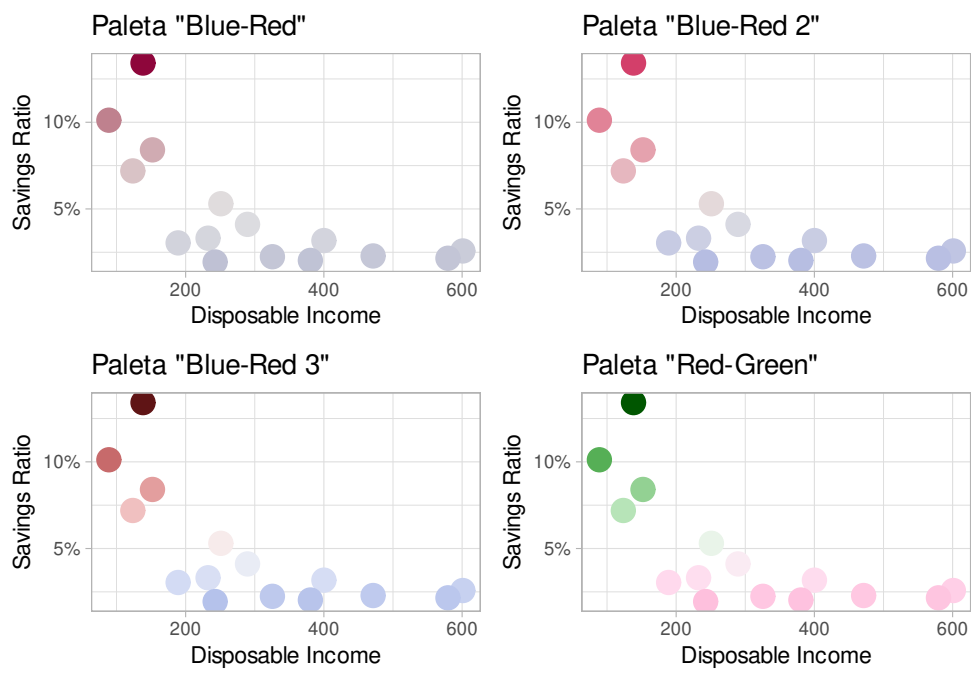


Figura 3.14: Exemplo de uso de paleta divergente em gráfico de dispersão

a paleta escolhida. Esta é uma ótima técnica para salientar escalas numéricas. Na mesma figura, note como os pontos abaixo de 4,75% – média de *savings ratio* – ficam com cores semelhantes, enquanto aqueles acima possuem cores mais fortes e diferenciadas, destacando os países com maiores razões de poupança (*savings ratio*).

3.4.3 Paletas qualitativas

Paletas de cores qualitativas servem para indicar grupos heterogêneos nos dados. Isto é, queremos distinguir pontos no gráfico, porém não indicando intensidade. Um exemplo de cinco paletas qualitativas do pacote **colorspace** são apresentadas na Figura 3.15.

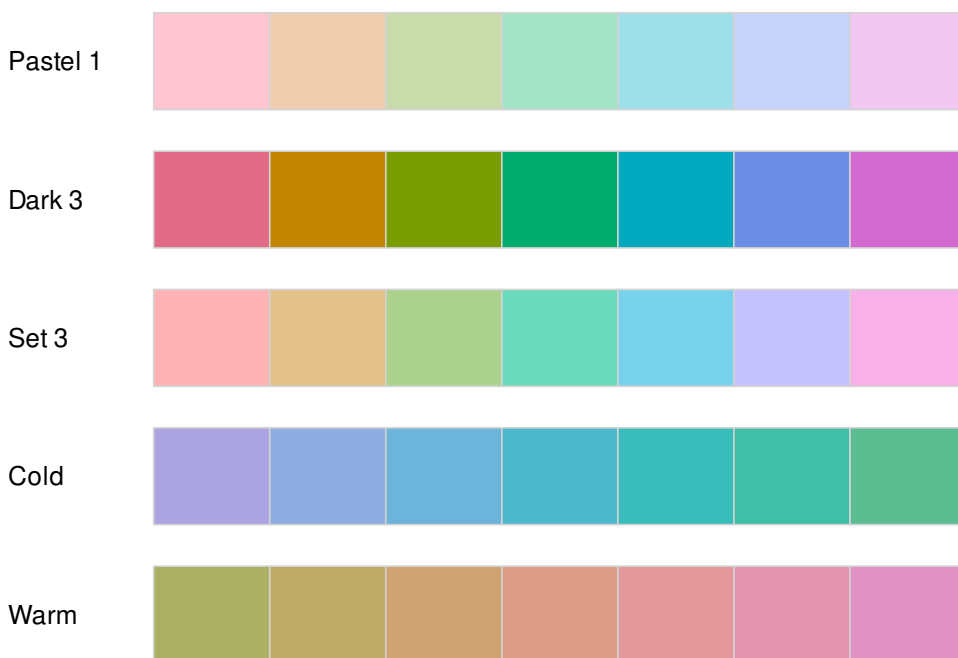


Figura 3.15: Exemplo de quatro paletas de cores divergentes do pacote **colorspace** .

Veja que as cores em 3.15 não indicam sequência, e sim contraste. Um exemplo de uso de paletas qualitativas é o gráfico de dispersão dos dados de poder de poupança, Figura 3.16, onde utilizamos uma paleta de cores para distinguir os diferentes países.

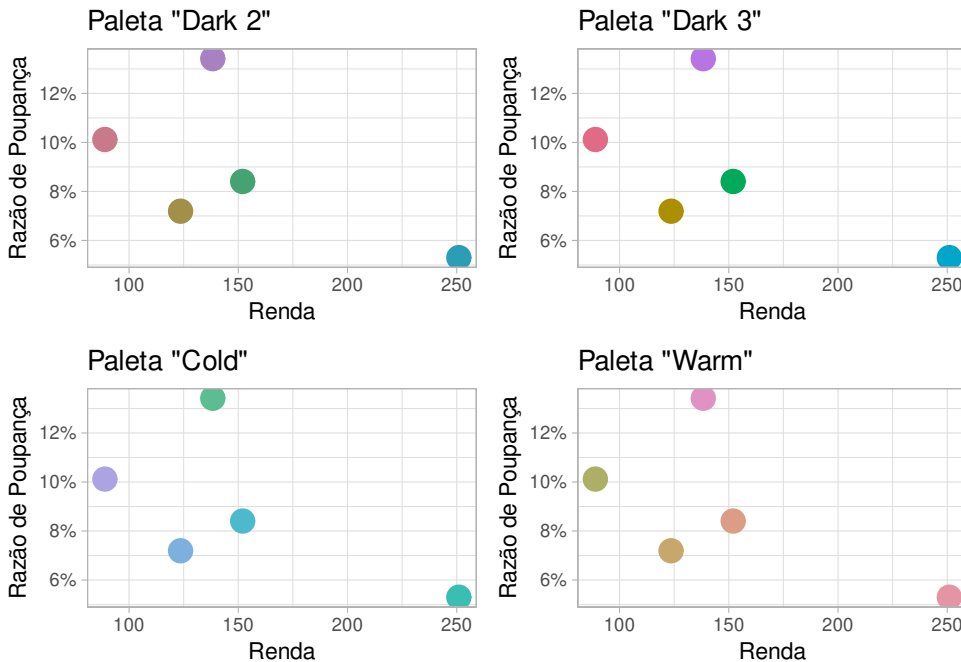


Figura 3.16: Exemplo de uso de paletas qualitativas

3.5 Tamanho

O canal tamanho refere-se ao espaço tomado por determinada forma como linha, círculos ou triângulos. O uso é simples, aumentamos ou diminuimos a forma para salientar escalas numéricas. Em outras palavras, vinculamos a escala geométrica a escala numérica dos dados. Veja na Figura 3.17 um exemplo simples de uso do tamanho dos pontos.

Note no gráfico anterior como os círculos são maiores de acordo com o valor de *Savings Ratio*, eixo vertical, com uma escala numérica disponível na legenda do gráfico. Podemos também utilizar o canal de tamanho juntamente com o canal de cor, Figura 3.18.

Cuidado, porém, com excessos. O gráfico em 3.18 contém bastante informação ao mesclar o uso do canal de cor e tamanho. Como regra, busque construir gráficos simples e intuitivos: não espere que o leitor gaste muita energia mental para entender o gráfico. Caso precisar, construa mais de uma figura para passar a sua mensagem.

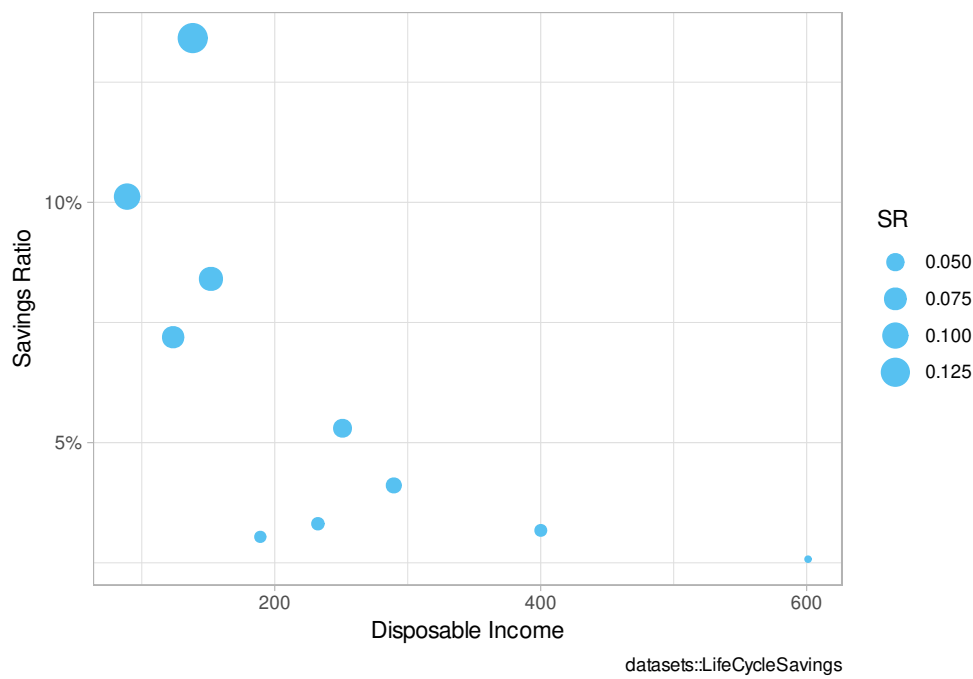


Figura 3.17: Exemplo de uso de tamanho para um gráfico de dispersão

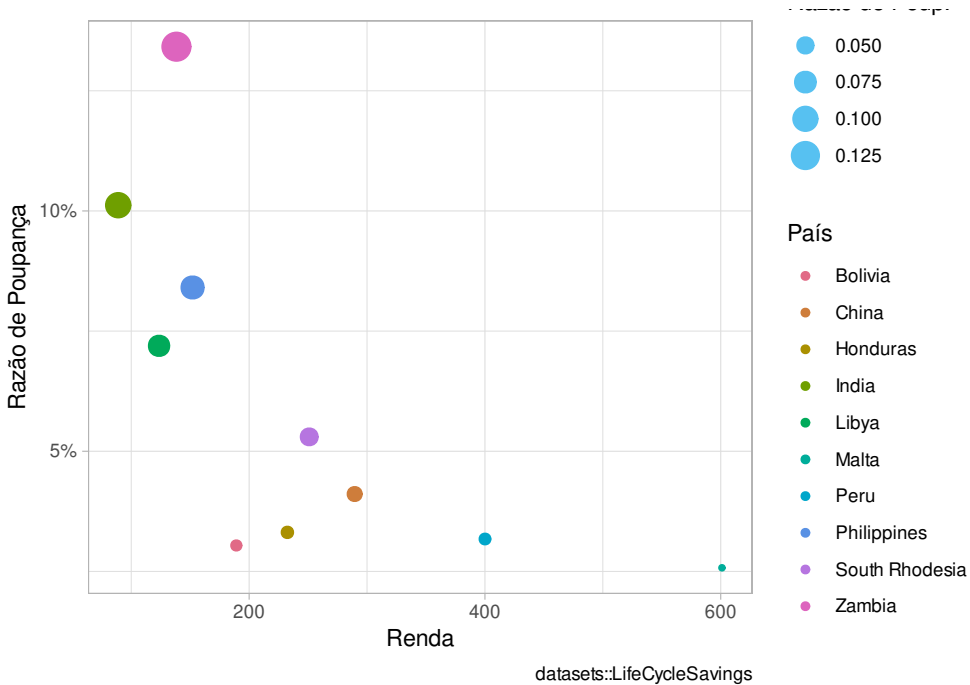


Figura 3.18: Exemplo de uso de tamanho para diferenciar pontos

3.6 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, podes baixar uma versão do arquivo compilado no site do livro¹.

01 - Em 22/09/2022, o grupo LatinoMetrics publicou uma visualização de dados a respeito da impressão de confiança da população brasileira em relação as diferens mídias jornalísticas. O conteúdo pode ser acessado no Instagram, e é disponibilizado abaixo:

Observando a figura, destaque os elementos **textuais** e **gráficos** utilizados pelo autor.

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

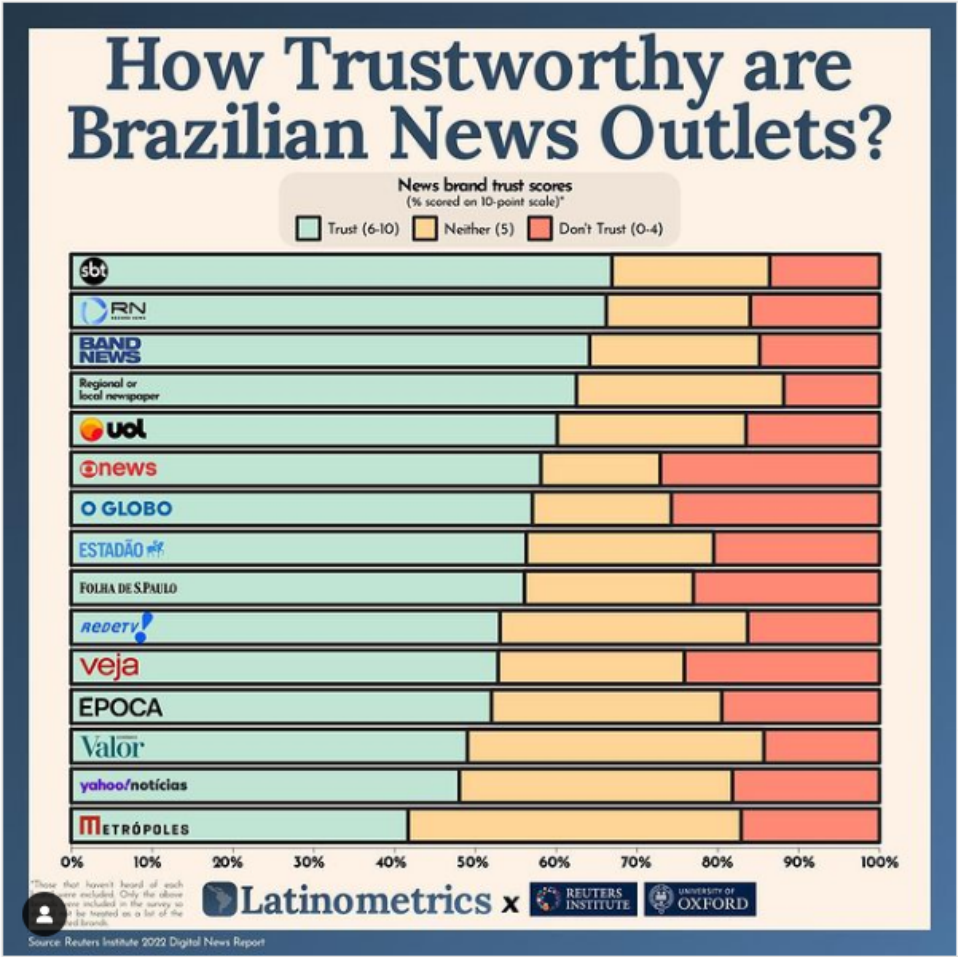


Figura 3.19: Postagem do LatinoMetrics em 22/09/2022

02 - Statspanda é outro grupo especializado em produção de conteúdo relacionado a visualização de dados, porém com assuntos muito mais abrangentes que o LatinoMetrics. Em 17/09/2022, o grupo publicou a seguinte figura no Instagram:

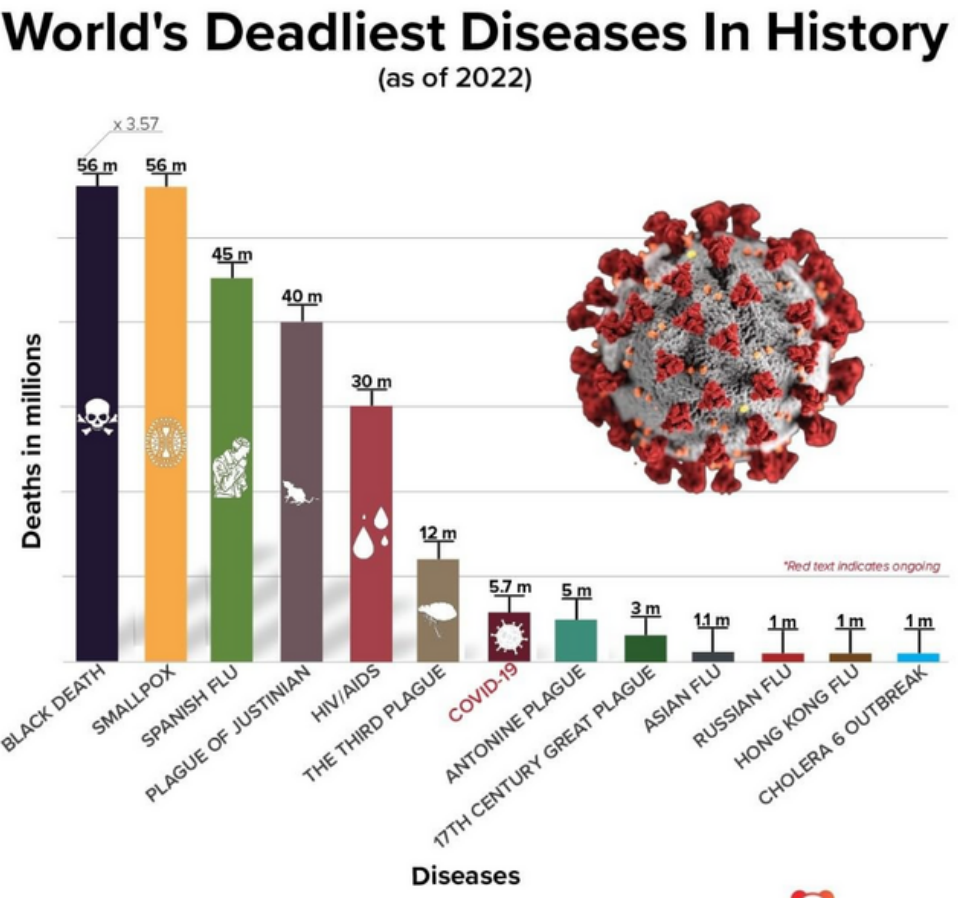


Figura 3.20: Postagem do Statspanda em 19/09/2022

Neste caso, quais foram os elementos **gráficos** utilizados pelo autor e como os mesmos se associam a mensagem da figura.

03 - No Reddit/dataisbeautiful é possível encontrar a visualização da receita de todos os filmes da franchise *Star Wars*.

Apesar de ser esteticamente interessante, o mesmo poderia ser melhorado. Com base no que aprendeu neste capítulo do livro,

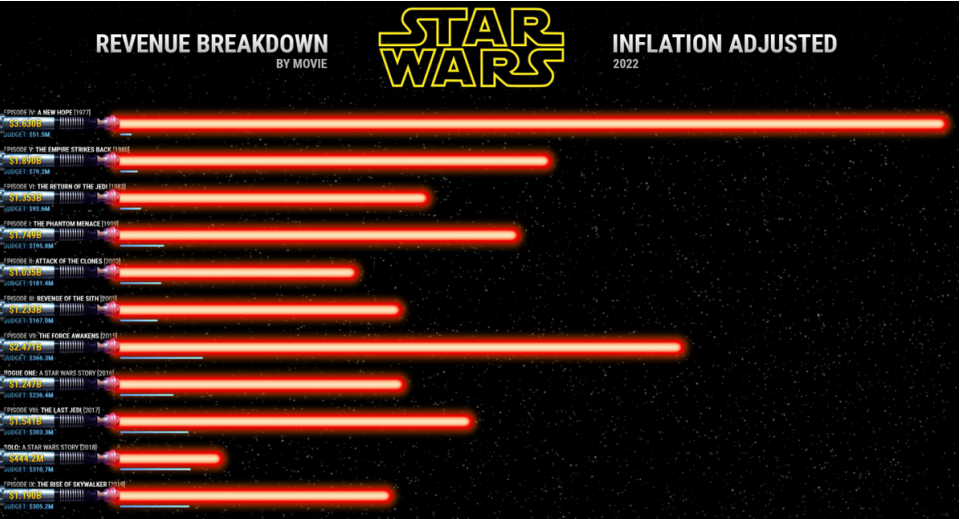


Figura 3.21: Postagem do Reddit/dataisbeautiful

analise o gráfico e faça recomendações para sua melhoria buscando sempre maior clareza e simplicidade.

CAPÍTULO 4

VISUALIZAÇÃO DE DADOS COM O R

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.1 Sobre o R

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.2 Pacote ggplot2

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.2.1 Alternativas ao ggplot2

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.3 O ciclo de criação de figuras

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.4 Dados de entrada

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.4.1 O formato dos dados

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do

livro é uma ótima maneira de suportar este e outros projetos do autor.

4.4.2 Tipos de dados

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.5 O sistema de camadas do ggplot2

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.6 Mapeamento de canais com aes()

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.6.1 Função geom_line()

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.6.2 Função `geom_point()`

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.6.3 Função `geom_col()`

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.6.4 Função `geom_text()`

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.7 Uma aplicação com dados reais

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.8 Paletas de cores

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do

livro é uma ótima maneira de suportar este e outros projetos do autor.

4.8.1 Aplicando uma paleta a um gráfico

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.9 Temas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.9.1 Aplicando um tema a um gráfico

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.9.2 Módulos adicionais de temas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.10 Painéis e facetas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.11 Exportando figuras

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

4.12 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, podes baixar uma versão do arquivo compilado no site do livro¹.

01 - Em um novo script do R, crie um vetor de valores aleatórios da distribuição Normal com o comando `rnorm(N)`, onde `N` é igual a 100. Agora, crie um gráfico de pontos onde o eixo `y` é representado pela série anterior, e o eixo `x` é simplesmente a contagem dos valores (1..100). Para este gráfico, utilize o *template* básico do `ggplot2`, isto é, não precisas modificar nenhum elemento textual do gráfico, por enquanto.

02 - Para o gráfico anterior, adicione os seguintes elementos textuais e gráficos:

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

- título, subtítulo;
 - *caption* com a data e tempo de compilação do gráfico;
 - textos nos eixos *x* e *y*;
 - aplique o tema `theme_light`
-

03 - Para o mesmo gráfico anterior, adicione uma nova coluna chamada `type` no dataframe, a qual pode tomar o valor "A" ou "B". Para isto, podes usar o comando `sample(c("A", "B"), size = N, replace = TRUE)`. Note que o valor de `N` foi definido anteriormente.

Com base no novo dataframe, crie um gráfico de linhas com cores diferentes para cada valor em `type`.

04 - Para o mesmo gráfico anterior, adicione uma camada de linhas no gráfico.

05 - Agora, use o gráfico anterior, mas modifique o formato dos pontos, definindo que o argumento `shape` será mapeado de acordo com coluna `type`. Também aumente o tamanho de todos os pontos do gráfico para 3.

06 - Agora, modifique o código para que o tamanho dos pontos seja relativo a coluna `x` do dataframe de entrada. Isto é, o efeito desejado é que os pontos aumentem de tamanho ao longo do eixo horizontal.

07 - Neste exercício, remova a camada de pontos utilizada no gráfico e adicione o texto disponível na coluna `type` (A ou B) no mesmo local onde os pontos se situavam.

08 - Para o gráfico anterior, use as funções do pacote `colorspace`

para implementar a paleta de cores “Harmonic” do canal color (dados qualitativos).

09 - Exporte a figura anterior com as seguintes informações:

1. arquivo de exportação chamado “fig-eoc-vdr.png” e localizado na pasta padrão “Documentos” (atalho com ~);
2. use um tamanho 10 X 10 em centímetros

Verifique se o arquivo foi criado corretamente na pasta desejada.

10 - Com base na função `yfR::yf_collection_get`, baixe os dados de preços de ações para a composição atual do índice Ibovespa, com início a cinco anos atrás e término como sendo a data atual. Com base nos dados importados, siga os seguintes passos:

1. filtre os dados para manter apenas as 5 ações com maior rentabilidade acumulado na período, e as 5 com menor.
2. construa uma figura com os retornos acumulados das 10 ações selecionadas anteriormente, onde o eixo horizontal representa as datas.
3. Implemente as seguintes modificações no gráfico:
 - Adicione título, subtítulo e caption e também o texto dos eixo horizontal e vertical;
 - modifique a escala do eixo horizontal para percentagens com comando `scale_y_continuous(labels = scales::percent)`;
 - use o tema `theme_light`;
4. exporte a figura resultante para um arquivo de tamanho tamanho 10 cm (height) X 15 cm (width) chamado “fig-ibov-10-ações.png”, e localizado na pasta padrão “Documentos” (atalho com ~);

CAPÍTULO 5

GRÁFICOS ESTATÍSTICOS

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.1 Visualizando uma distribuição

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.1.1 Histogramas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do

livro é uma ótima maneira de suportar este e outros projetos do autor.

5.1.2 Densidades

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.1.3 Quartis

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.2 Visualizando múltiplas distribuições

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.2.1 Boxplots

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.2.2 Uso de painéis/facetas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.2.3 Pacote ggridges

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.3 Visualizando correlações

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.3.1 Pacote corrplot

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

5.4 Visualizando incerteza

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do

livro é uma ótima maneira de suportar este e outros projetos do autor.

5.5 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, pode baixar uma versão do arquivo compilado no site do livro¹.

01 - Utilizando o pacote `yfR`, baixe os dados dos preços da Grendene, `ticker = GRND3.SA`, desde 2015 até o data atual. Com base nos dados importados, crie um histograma dos retornos diários dos preços ajustados da ação (coluna `ret_adjusted_prices`).

02 - Utilizando o pacote `yfR` e argumento `freq_data = 'monthly'` em `yf_get`, baixe os dados dos preços **mensais** da Petrobras, `ticker = PETR3.SA`, de 2015 até o dia atual. Com base nos dados importados, crie um histograma dos retornos mensais da ação, adicionando os seguintes itens ao gráfico:

- título e subtítulo;
- texto no eixo x e y;
- caption;
- tema `theme_light`

03 - Pacote `microdatasus`, criado por Raphael Saldanha (`rf-saldanha`), importa e limpa dados do Sistema de Saúde Único (SUS). No pacote do livro é possível encontrar dados de mortalidade importados pelo `microdatasus` e filtrados para manter apenas algumas colunas da base completa. O arquivo é chamado `Exercises-sus-mortality-data.csv` e pode ser importado com a função `vdr::data_import`.

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

Com base nos dados da coluna `age`, crie um gráfico de **densidade** para a distribuição de idades de falecimento, incluindo os seguintes itens no gráfico:

- título;
 - texto no eixo x e y;
 - caption;
 - tema `theme_light`
-

04 - Para os mesmos dados de mortalidade retirados do SUS, crie um gráfico QQ com a distribuição de idades de falecimento (coluna `age`).

05 - Considerando novamente os dados de mortalidade do SUS, crie um gráfico de densidade de distribuição, porém separe as linhas do gráfico de acordo com o gênero (Masculino/Feminino). Isto é, queremos uma linha de densidade para o sexo masculino, e outra para o feminino. Dica: remova os casos de `SEXO = NA (Not Available)` antes de criar o gráfico.

06 - Repita o exercício anterior, porém use facetas para mostrar as diferentes densidades de mortalidade entre os gêneros masculino e feminino. Dentro da função `facet_wrap`, use argumento `nrow = 2` para empilhar os painéis verticalmente.

07 - Considerando novamente os dados do SUS, use o comando `facet_grid` para mostrar as densidades de idade de falecimento quando separando os dados por gênero (coluna `SEXO`) e escolaridade (coluna `ESC`). DICA: remova os casos de `NA (Not Available)` encontrados em ambas colunas, `SEXO` e `ESC`.

08 - Crie um gráfico de diagrama (*boxplot*), onde o eixo horizontal é a escolaridade (coluna `ESC`) e o eixo vertical representa a

distribuição dos dados de idade de falecimento, coluna (age).

CAPÍTULO 6

MAPAS

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.1 Pacote geobr

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.1.1 Mapa do Brasil

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do

livro é uma ótima maneira de suportar este e outros projetos do autor.

6.1.2 Mapa do Brasil com cidades e capitais

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.2 Pacote ggmap

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.2.1 Registrando um projeto

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.2.2 Utilizando ggmap

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.2.3 Codificação geográfica

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

6.3 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, pode baixar uma versão do arquivo compilado no site do livro¹.

01 - Utilizando função `geobr::read_country`, baixe os dados do território do Brasil e mostre a área em um novo gráfico.

02 - Utilizando função `geobr::read_indigenous_land`, baixe os dados do território indígena no Brasil e adicione esta camada ao gráfico do Brasil criado anteriormente, com as fronteiras da área indígena representadas na cor vermelha.

03 - Utilizando função `geobr::read_state()`, baixe os dados de todos os estados do Brasil e, em um mapa do país, mostre os limites de cada estado na cor azul.

04 - Utilizando função `geobr::read_state()`, baixe os dados do estado do Rio Grande do Sul e, com os dados disponíveis no arquivo “Chapter07-latlong-cities-brazil.csv” do pacote `vdr`:

1. mostre as cidades gaúchas como pontos azuis no mapa, ajustando o tamanho e transparência dos pontos para melhor vi-

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

- sualização do mapa;
2. mostre a capital, Porto alegre, como um ponto vermelho com tamanho e intensidade maior.
 3. adicione o texto “PORTO ALEGRE” em preto, acima do ponto de localização da capital.
 4. adicione título e caption para o gráfico

CAPÍTULO 7

PROGRAMANDO COM O GGPLOT2

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

7.1 Criando figuras via funções

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

7.2 Testando temas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

7.3 Testando paletas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

7.4 Testando temas e paletas

Você chegou ao fim da versão online do livro **Visualização de Dados com o R**. O conteúdo integral da obra pode ser adquirido na loja da Amazon como ebook ou livro impresso. A compra do livro é uma ótima maneira de suportar este e outros projetos do autor.

7.5 Exercícios

A solução dos exercícios de final de capítulo pode ser compilada localmente com o comando `vdr::exercises_compile_solution()`. Alternativamente, pode baixar uma versão do arquivo compilado no site do livro¹.

01 - O código abaixo irá criar um gráfico simples com os preços da ação da empresa Engie (*EGIE.SA*).

¹<https://www.msperlin.com/vdr/vdr-eoc-solutions.html>

```
library(ggplot2)

df_yf <- yfR::yf_get("EGIE3.SA",
                     first_date = "2018-01-01")

p <- ggplot(
  df_yf,
  aes(x = ref_date, y = price_adjusted)
) + geom_line()
```

Após a execução do código acima, use as ferramentas de programação do R para testar os seguintes temas da figura:

- `ggplot2::theme_light()`
- `ggplot2::theme_bw()`
- `ggplot2::theme_grey()`
- `ggplot2::theme_minimal()`

e lembre de adicionar a informação do tema no título da figura.

02 - Considere o gráfico *template* gerado pelo código a seguir:

```
library(ggplot2)

p <- ggplot(
  yfR::yf_get("USIM5.SA", first_date = "2018-01-01"),
  aes(x = ref_date, y = price_adjusted)
) +
  geom_line() +
  labs(subtitle = "SUBTITLE",
       caption = "CAPTION")
```

Pacote `hrbrthemes` oferece uma grande variedade de temas adicionais a serem utilizados em um gráfico do `ggplot2`. Todas funções de tema possuem um padrão bem distinto, onde o nome da função inicia com o texto `theme_`. Baseado nesta informação e também no uso da função `ls`, podemos descobrir o nome de todas as funções de tema do `hrbrthemes` com o código abaixo:

```
library(hrbrthemes)

all_fcts <- ls("package:hrbrthemes")

idx <- all_fcts |>
  stringr::str_which("theme_")

theme_fcts <- all_fcts[idx]
```

Com base no código anterior:

1. use as ferramentas de programação do R para, de forma automática, aplicar todos os temas anteriores para o gráfico gerado acima.
2. agregue todos os gráfico em um único arquivo com função `cowplot::plot_grid`
3. salve a figura resultante como `hrbrthemes-demo.png`, pasta documentos (atalho com ~). Defina o tamanho da figura como `height = 20cm` e `width = 30cm`. DICA: Caso a figura gravada sair com fundo transparente, abra-a em um navegador da web, tal como o Chrome ou firefox.

03 - Considere o gráfico criado de barras criado com o código abaixo:

```
library(ggplot2)

tickers <- c("PETR3.SA", "ITSA3.SA",
             "GRND3.SA", "GGBR3.SA",
             "EGIE3.SA", "USIM5.SA")

df_yf <- yfR::yf_get(tickers,
                    "2020-01-01")

df_tab <- df_yf |>
  dplyr::group_by(ticker) |>
  dplyr::summarise(
    total_ret = dplyr::last(cumret_adjusted_prices) - 1
```

```

    ) |>
    dplyr::arrange(total_ret)

p <- ggplot(df_tab, aes(x = reorder(ticker, total_ret),
                             y = total_ret,
                             fill = total_ret)) +
  geom_col(color = "white") +
  theme_minimal() +
  coord_flip() +
  labs(title = stringr::str_glue(
    "Performance de {length(tickers)} Ações Selecionadas"
  ),
        y = "Retorno Total",
        x = "Ticker")

```

Com base no código anterior, use as ferramentas de programação do R para, de forma automática, aplicar todas as possíveis combinações entre os temas:

- `ggplot2::theme_light`
- `hrbrthemes::theme_modern_rc`
- `ggtech::theme_airbnb_fancy()`

e as paletas do `colorspace`

- Red-Green
- Green-Brown
- Tropic

Adicionalmente:

1. agregue todos os gráfico em um único arquivo com função `cowplot::plot_grid`
2. salve a figura resultante como `themes-and-palettes-demo.png`, pasta documentos (atalho com `~`). Defina o tamanho da figura como `height = 30cm` e `width = 60cm`. DICA: Caso a figura gravada sair com fundo transparente, abra-a em um navegador da web, tal como o Chrome ou firefox.

CAPÍTULO 8

CONCLUSÃO

E assim chegamos ao fim do livro! Espero que tenha gostado do conteúdo e aprendido alguns novos truques na manipulação e visualização de dados com o R. Truques este que, possivelmente, lhe rendam uma maior reputação no trabalho ou então na conquista de seu título acadêmico.

Aprender a utilizar o R na criação de gráficos é uma jornada sem fim. Apesar de minha maior experiência com a plataforma – uso o R tanto para produzir conteúdo de aula quanto para realizar pesquisas científicas –, ainda me surpreendo por estar em constante evolução. Portanto, não pare com este livro. Aqui tentei fornecer uma introdução ao tema de visualização de dados com o R. O universo do **ggplot2** é imenso e valioso de se explorar.

Boa jornada e bons estudos!

Marcelo S. Perlin, 23/11/2022.

REFERÊNCIAS BIBLIOGRÁFICAS

- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2022). *rmarkdown: Dynamic Documents for R*. R package version 2.18.
- Arnold, J. B. (2021). *ggthemes: Extra Themes, Scales and Geoms for ggplot2*. R package version 4.2.4.
- Bazley, W. J., Cronqvist, H., and Mormann, M. M. (2017). In the red: The effects of color on investment behavior. *Swedish House of Finance research paper*, 17:16.
- Bion, R. (2018). *ggtech: ggplot2 tech themes and scales*. R package version 0.1.1.
- file., S. A. (2022). *paletteer: Comprehensive Collection of Color Palettes*. R package version 1.5.0.
- Healy, K. (2018). *Data visualization: a practical introduction*. Princeton University Press.
- Henry, L. and Wickham, H. (2022). *purrr: Functional Programming Tools*. R package version 0.3.5.

- Hester, J., Wickham, H., and Csárdi, G. (2021). *fs: Cross-Platform File System Operations Based on libuv*. R package version 1.5.2.
- Iannone, R. (2022). *DiagrammeR: Graph/Network Visualization*. R package version 1.0.9.
- Iannone, R., Cheng, J., Schloerke, B., Hughes, E., and Seo, J. (2022). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.8.0.
- Ihaka, R., Murrell, P., Hornik, K., Fisher, J. C., Stauffer, R., Wilke, C. O., McWhite, C. D., and Zeileis, A. (2022). *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes*. R package version 2.1-0/r646.
- Kahle, D., Wickham, H., and Jackson, S. (2022). *ggmap: Spatial Visualization with ggplot2*. R package version 3.0.1.
- Kassambara, A. (2022). *ggpubr: ggplot2 Based Publication Ready Plots*. R package version 0.5.0.
- McWhite, C. D. and Wilke, C. O. (2022). *colorblindr: Simulate colorblindness in R figures*. R package version 0.1.0.
- Müller, K. and Wickham, H. (2022). *tibble: Simple Data Frames*. R package version 3.1.8.
- Pebesma, E. (2022). *sf: Simple Features for R*. R package version 1.0-9.
- Pedersen, T. L. (2022). *patchwork: The Composer of Plots*. R package version 1.1.2.
- Pereira, R. H. M. and Goncalves, C. N. (2022). *geobr: Download Official Spatial Data Sets of Brazil*. R package version 1.7.0.
- Perlin, M. (2022a). *yfR: Downloads and Organizes Financial Data from Yahoo Finance*. <https://github.com/ropensci/yfR>.
- Perlin, M. S. (2021). *Análise de Dados Financeiros e Econômicos com o R*. publicação independente, terceira edição edition.
- Perlin, M. S. (2022b). *vdr: Data and Functions for Book Visualizacao de Dados com o R*. R package version 0.5.0.

- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rudis, B. (2020). *hrbrthemes: Additional Themes, Theme Components and Utilities for ggplot2*. R package version 0.8.0.
- Schwabish, J. A. (2014). An economist's guide to visualizing data. *Journal of Economic Perspectives*, 28(1):209–34.
- Siegel, J. J. (2015). *Investindo em Ações no Longo Prazo-5*. AMGH Editora.
- Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2022). *plotly: Create Interactive Web Graphics via plotly.js*. R package version 4.10.1.
- Slowikowski, K. (2022). *ggrepel: Automatically Position Non-Overlapping Text Labels with ggplot2*. R package version 0.9.2.
- Smeets, B. (2022). *artyfarty: Themes for ggplot2*. R package version 0.0.1.
- Spinu, V., Grolemond, G., and Wickham, H. (2022). *lubridate: Make Dealing with Dates a Little Easier*. R package version 1.9.0.
- Stylianou, N., william.dahlgreen@bbc.co.uk Dahlgreen, W., robert.cuffe@bbc.co.uk Cuffe, R., Calver, T., and Mpini, R. (2022). *bbplot: MAKING GGPLOT GRAPHICS IN BBC NEWS STYLE*. R package version 0.2.
- Ushey, K. (2022). *renv: Project Environments*. R package version 0.16.0.
- Wei, T. and Simko, V. (2021). *corrplot: Visualization of a Correlation Matrix*. R package version 0.92.
- Wickham, H. (2021). *conflicted: An Alternative Conflict Resolution Strategy*. R package version 1.1.0.
- Wickham, H. (2022a). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.2.

- Wickham, H. (2022b). *httr: Tools for Working with URLs and HTTP*. R package version 1.4.4.
- Wickham, H. (2022c). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.1.
- Wickham, H. (2022d). *tidyverse: Easily Install and Load the Tidyverse*. R package version 1.3.2.
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., and Dunnington, D. (2022a). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.4.0.
- Wickham, H., François, R., Henry, L., and Müller, K. (2022b). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.
- Wickham, H. and Girlich, M. (2022). *tidyr: Tidy Messy Data*. R package version 1.2.1.
- Wickham, H. and Seidel, D. (2022). *scales: Scale Functions for Visualization*. R package version 1.2.1.
- Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.
- Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2*. R package version 1.1.1.
- Wilke, C. O. (2022a). *dviz.supp: Supporting materials for Claus Wilke's data visualization book*. R package version 0.1.0.
- Wilke, C. O. (2022b). *ggridges: Ridgeline Plots in ggplot2*. R package version 0.5.4.
- Wilke, C. O. and Wiernik, B. M. (2022). *ggtext: Improved Text Rendering Support for ggplot2*. R package version 0.1.2.
- Xie, Y. (2022a). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.30.
- Xie, Y. (2022b). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.41.

ÍNDICE REMISSIVO

bookdown, 17

colorspace, 51, 52, 54, 56, 58

ggplot2, 11, 17, 89

vdr, 12, 13

exerci-

ses_compile_solution,

13