

Estatística para Análise de Dados na Administração Pública

Fundação Escola Nacional de Administração Pública

Diretoria de Desenvolvimento Profissional

Conteudista

Fernando Barbalho (conteudista, 2023).

enap

Enap, 2023

Fundação Escola Nacional de Administração Pública

Diretoria de Desenvolvimento Profissional

SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Módulo 1: Princípios da Estatística Básica	6
Unidade 1: Introdução aos Campos da Estatística Básica	6
1.1 Introdução ao Campo de Estudo da Estatística	6
1.2 Características da Estatística Descritiva	10
1.3 Características da Estatística Inferencial	15
1.4 Probabilidade	17
Referências	19
Módulo 2: Introdução à Linguagem R.....	20
Unidade 1: Funções do Ambiente RStudio	20
1.1 Como o R Pode Facilitar a sua Vida (Facilidades e Benefícios)	20
1.2 Como Instalar o Ambiente de RStudio.....	28
1.3 Conhecendo as Funcionalidades do Ambiente RStudio.....	28
1.4 Fazendo o Primeiro <i>Script</i> em R	29
Referências	32
Unidade 2: Utilizando a Estatística para Simplificar um Conjunto de Informações por Distribuição de Frequências	33
2.1 Tipos de Variáveis	33
2.2 Distribuição de Frequências	39
Referências	44
Unidade 3: Utilizando a Estatística para Simplificar um Conjunto de Informações por Medidas Resumo	46
3.1 Medidas de Centralidade	46
3.2 Medidas de Dispersão	49
3.3 Medidas de Centralidade e Dispersão Usando R	53
Referências	57
Módulo 3: Modelando Dados e Gráficos.....	59
Unidade 1: Manipulação de Dados	59
1.1 Filtros de Dados	59
1.2 Construindo Agrupamentos	64
1.3 Alterando a Estrutura de Dados	67
1.4 Manipulando Dados com Pacote Tidyverse	70
Referências	76

Unidade 2: Criando Gráficos Estatísticos com o Pacote Ggplot2	78
2.1 Biblioteca Ggplot2: uma Gramática para Desenho de Gráficos	78
2.2 Como utilizar o Pacote Ggplot2	85
Referências	91
Unidade 3: Transformando Dados Usando R.....	93
3.1 Transformação Logarítmica	93
3.2 Elaborando um Histograma com Escala Logarítmica	98
3.3 Box-Plot e Escala Logarítmica	104
3.4 Análise de Distribuição de Dados com o Histograma e Box-plot usando Ggplot2	109
Referências	113
Unidade 4 - Análise Bivariada na Linguagem R	115
4.1 Associação entre Variáveis Quantitativas: Correlação	115
4.2 Usando R para Realizar Análise de Correlação	119
Referências	122
Módulo 4: Análise de Dados na Administração Pública	124
Unidade 1: Dados Abertos Governamentais	124
1.1 O que São Dados Abertos?	124
1.2 Trabalhando com Portais de Dados Abertos	127
Referências	130
Unidade 2 - Análise de Dados de Finanças Públicas	131
2.1 O Campo de Finanças Públicas e a Análise de Dados	131
2.2 Análises de Dados sobre Receitas e Despesas Primárias do Governo Central Brasileiro através do Pacote RTN	134
2.3 Pacote Rcofog - Consumo de Dados sobre Funções Típicas de Governo	136
Referências	139
Unidade 3: Análise de Dados de Saúde Pública	141
3.1 O Campo de Saúde Pública e a Análise de Dados	141
3.2 Pacote Microdatasus (Consumo de Dados Relacionados ao Sistema Único de Saúde).....	145
Referências	147
Unidade 4: Análise de Dados de Educação Pública	148
4.1 O Campo de Educação Pública e a Análise de Dados.....	148
4.2 Analisando Dados de Censo Escolar	153
Referências	155

Apresentação e Boas-vindas

Olá! Seja bem-vindo(a) ao curso de Estatística para Análise de Dados na Administração Pública.

O objetivo deste curso é capacitar você a reconhecer os princípios da estatística descritiva em linguagem R para que consiga analisar dados na Administração Pública. Visando organizar seus estudos, os temas foram estruturados em quatro módulos, com tópicos e subtópicos.

No primeiro módulo “**Princípios da Estatística Básica**” serão abordados os aspectos conceituais de estatística e a aplicação dessa disciplina para análise de dados.

No segundo módulo “**Introdução à Linguagem R**” será feita uma apresentação da linguagem R e do ambiente RStudio demonstrando o uso da linguagem nos principais temas de estatística descritiva.

No terceiro módulo “**Modelando Dados e Gráficos**” você verá os usos de R com modelagem de dados e construção de gráficos.

O quarto módulo “**Análise de Dados na Administração Pública**” apresentará aplicações práticas de análise de dados em setores importantes da Administração Pública.

Para iniciar, assista a videoaula Apresentação do Curso.



Videoaula: [Apresentação do Curso](#)

Sucesso em sua jornada de conhecimento! Mão à obra!

Módulo

1 Princípios da Estatística Básica

Nesse módulo, composto por uma unidade, você verá conceitos importantes que ajudarão a compreender o papel da estatística na análise de dados. Além disso, serão descritas as características dos três principais campos de estatística, com ênfase na estatística descritiva que é o foco desse curso.

Unidade 1: Introdução aos Campos da Estatística Básica

Objetivo de aprendizagem

Ao final desta unidade você será capaz de reconhecer os principais conceitos da estatística básica tendo em vista a análise de dados.

1.1 Introdução ao Campo de Estudo da Estatística

Não é uma coincidência haver um curso sobre análise de dados na Administração Pública usando o ferramental da estatística. A origem desse campo de estudo está fortemente atrelada à administração do Estado. Na verdade, o próprio nome estatística em sua etimologia é derivado do neolatim *statisticum collegium* ("conselho de estado") que levou ao termo do italiano *statista* ("estadista" ou "político") (WIKIMEDIA FOUNDATION, 2022).

“ A definição mais habitual do campo indica que se trata de uma área do conhecimento que utiliza teorias probabilísticas para explicação de eventos, estudos e experimentos. A missão da estatística passa por “obter, organizar e analisar dados, determinar suas correlações, tirar delas suas consequências para descrição, explicar o que passou e fazer previsões” (WALPOLE, 2009). ”

Essas características do conceito de estatística levam à divisão do campo em três subcampos:

- ① Estatística Descritiva, mais interessada em descrever e sumarizar os achados dos dados;
- ② Estatística Inferencial, que se interessa em fazer inferências sobre características de uma população a partir de um número de casos estudados; e
- ③ Probabilidade, que se encarrega de realizar os cálculos de probabilidade de eventos específicos ocorrerem.

Se no início a estatística associava-se às possibilidades de análises dos dados de censo e resultados econômicos das ações do Estado, hoje há uma enorme diversidade de aplicações que vão da Astronomia à Biologia, da Engenharia ao Esporte e da História à Psicologia. Em todas essas aplicações, percebe-se uma ou mais manifestações da missão da estatística.

Mas e na Administração Pública? Bem, especificamente na Administração Pública, a estatística é usada para descrever resultados de políticas públicas, como o alcance de programas de assistência social, números de campanhas de vacinação, cálculo de índice de inflação, consolidação de resultados fiscais e avaliação de ranking de escolas públicas, entre outras aplicações.



Estatística

1. Descrever resultado de políticas públicas
2. Resultado de campanhas de vacinação
3. Cálculo de índice de inflação,
4. Consolidação de resultados fiscais
5. avaliação de ranking de escolas públicas

Estatística na Administração Pública.

Fonte: Freepik (2023). Elaboração: CEPED/UFSC (2023).

Agora que tal refletir um pouco? É muito fácil perceber todas essas aplicações no nosso dia a dia, concorda? Na verdade basta abrir as páginas físicas ou virtuais de jornais e revistas para verificar isso. Como exemplo, veja um recorte de um artigo do site Tilt UOL a seguir:

“ Observar estrelas e galáxias pode ser mais fácil - mas, ainda assim, os cientistas precisam recorrer às estatísticas. Para começar o cálculo, os cientistas escolhem uma área específica do Universo para observar. Com imagens obtidas por telescópios comuns, eles contam o número de estrelas ou galáxias encontradas nesta área. A partir daí, entra a estatística para extrapolar o número contado na amostra e encontrar dados que representem o volume total. Há ao menos um consenso quando cientistas contam galáxias: eles concordam com o princípio de que o Universo tem a mesma aparência em qualquer direção que se observa. Isso quer dizer que um observador, em qualquer em qualquer parte do Universo, vai sempre observar as mesmas propriedades, independentemente da geometria geral dele tilt uol (2022) ”

Você percebe que o texto de Tilt UOL deixa muito claro que os cientistas usam o ferramental da estatística para calcular o número de estrelas a partir de imagens de áreas do universo observável? Especificamente, eles descrevem no recorte do artigo o uso de técnicas de amostragem e extração.

E se mudarmos de assunto para Biologia, o que se pode encontrar? Veja um recorte da [matéria](#) do Globo Ciência (2013) que apresenta um teste estatístico que apoia uma metodologia laboratorial para inclusão ou exclusão de paternidade por testes de DNA. Acompanhe!



A pesquisadora aponta que após a coleta do material biológico, há a extração do DNA do núcleo da célula, e testes genéticos são realizados para que seja possível estabelecer a individualidade de cada pessoa. Em geral, avalia-se primeiro o filho e verifica-se a contribuição genética que veio da mãe. A partir daí, é possível saber a contribuição genética que teria que vir do pai biológico.

"Esses materiais são comparados com o material genético do possível pai. Se ele apresentar todos os elementos que o pai biológico precisaria ter, há a inclusão de paternidade. Para a inclusão ou exclusão de paternidade é realizado um teste estatístico em conjunto com a metodologia laboratorial", explica a pesquisadora.

Elizeu Fagundes, coordenador do Laboratório de Diagnóstico por DNA (LabDNA), da Universidade do Estado do Rio de Janeiro (UERJ), explica que o DNA nas células humanas está organizado em 23 pares de cromossomos. Um dos componentes do par foi herdado da mãe e outro do pai por meio dos gametas, que se uniram para formar a célula ovo, ou zigoto, quando da concepção de um novo indivíduo. "O que se busca verificar é se o suposto pai e o suposto filho apresentam 23 cromossomos absolutamente idênticos, o que caracteriza a paternidade. A investigação de paternidade por DNA, ou de maternidade, tem eficiência de 100%", ressalta o coordenador. Globo Ciência (2014).

Teste estatístico de paternidade.

Fonte: Freepik (2023).

A matéria em recorte indica que o teste estatístico apoia uma metodologia laboratorial para inclusão ou exclusão de paternidade por testes de DNA. Não é difícil imaginar a aplicação da estatística, mais precisamente da probabilidade, nesse procedimento.

Veja o que explica sobre isso [Marcos Noé](#), da equipe da Brasil Escola:

“a Genética é outra área que utiliza as teorias da probabilidade, pois os acontecimentos nesse ramo da Biologia envolvem eventos aleatórios, como o encontro dos gametas masculinos e femininos com determinados genes na fecundação". NOÉ (s.d.).**”**

Se você se interessou por esses exemplos, vale a pena procurar na Internet outras manifestações da estatística, tais como:

- gráfico das temperaturas medidas ao longo dos anos e séculos, demonstrando a evolução das mudanças climáticas;
- probabilidade de alguma equipe da NBA ou do Brasileirão ser campeã ao final do torneio;
- datação de objetos encontrados em sítios arqueológicos;
- gráfico da cotação do dólar ao longo do ano; ou
- gráfico de evolução da média móvel de infecções e óbitos pela COVID.

E caso você se pergunte, “certo, mas onde estão os exemplos da Administração Pública?”, fique atento(a)! Esses exemplos serão apresentados ao longo de todo o curso, incluindo alguns iniciais nos tópicos dessa unidade conforme você segue em frente!

1.2 Características da Estatística Descritiva



A estatística descritiva provavelmente já foi apresentada a você nos primeiros anos na escola. Basta lembrar que para conseguir passar de ano era necessário atingir uma média mínima no conjunto de avaliações a que você era submetido.

Talvez você não soubesse na época, ou mesmo até logo antes de ler esse parágrafo, mas a média é um dos melhores exemplos de quantificação que pode indicar o uso de estatística descritiva.

Você se lembra dos primeiros anos de escola?

Fonte: Freepik (2023).

O conceito de estatística descritiva está relacionado ao processo de análise para descrever e sumarizar dados. No caso do exemplo da escola, a média sumariza um conjunto de resultados alcançados pelos alunos, ao serem avaliados, que são medidos de forma quantitativa. Você deve se lembrar também dos seus colegas do ensino fundamental ou médio que alcançavam sempre as melhores notas. Informar a nota máxima obtida também é uma maneira de sumarizar o conjunto de dados.

A prática do uso da sumarização de dados tem uma história já antiga que começou principalmente pela tabulação de populações e dados econômicos. Mais recentemente, principalmente com o avanço dos recursos computacionais, ocorre com mais força o crescimento da chamada Análise Exploratória de Dados (AED). Sobre isso, Bussab e Morettin (2010) apontam que:

“

A análise descritiva de dados “limita-se a calcular algumas medidas de posição e variabilidade, como a média e variância, por exemplo” (BUSSAN; MORETTIN, 2010). Com o fortalecimento da AED passou-se a investir mais em técnicas gráficas em contraposição ao mero resumo numérico. Dessa nova abordagem surgiu uma série de novas formas de representar a distribuição de valores usando gráficos estatísticos e métodos de visualização de dados.

Os autores destacam que os gráficos facilitam a compreensão de uma mensagem de modo mais efetivo do que a leitura de tabelas e sumários numéricos. Neste curso, por exemplo, é feito uso intenso de gráficos por conta dessa influência da corrente da AED. Isso é facilitado pelos vários recursos de construção de visualização de dados disponibilizados pela linguagem R.

”



DESTAQUE

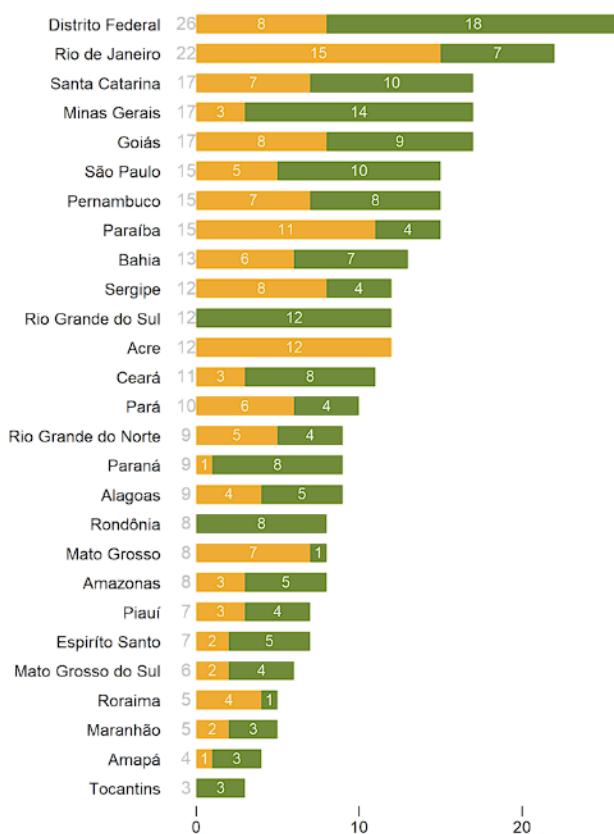
A linguagem R é uma das mais importantes opções de análise de dados disponíveis atualmente. É uma ferramenta de software livre, portanto gratuita, que ao longo dos anos incorporou o esforço de centenas de programadores, estatísticos, matemáticos, biólogos, cientistas políticos, artistas, entre diversas outras profissões que encontraram no R uma ferramenta de análise de dados e expressão de seus resultados. Mais adiante você estudará mais sobre essa linguagem. Voltemos a falar de AED!

Algumas figuras podem ajudar a entender o ponto que a abordagem de AED defende. A imagem a seguir é um gráfico de barras em formato de ranking.

Analise essa figura demonstrando o gráfico de ranking em formato de colunas horizontais.

Quantidade de empresas por Estado

Dependentes Não Dependentes



Ranking em formato de colunas horizontais.

Fonte: Brasil (2022a).

Observe que o gráfico mostra claramente uma ordem em que os rótulos, no caso unidades da federação (UF), são colocados em ordem decrescente de acordo com o valor de interesse, que nesse caso é o número de empresas estatais de cada UF. O gráfico permite você ver também, com clareza, os padrões que se formam. No caso, há alguns estados que apresentam aproximadamente o mesmo número de empresas dependentes e não dependentes. Vamos para o próximo exemplo. Dessa vez um gráfico de linha.

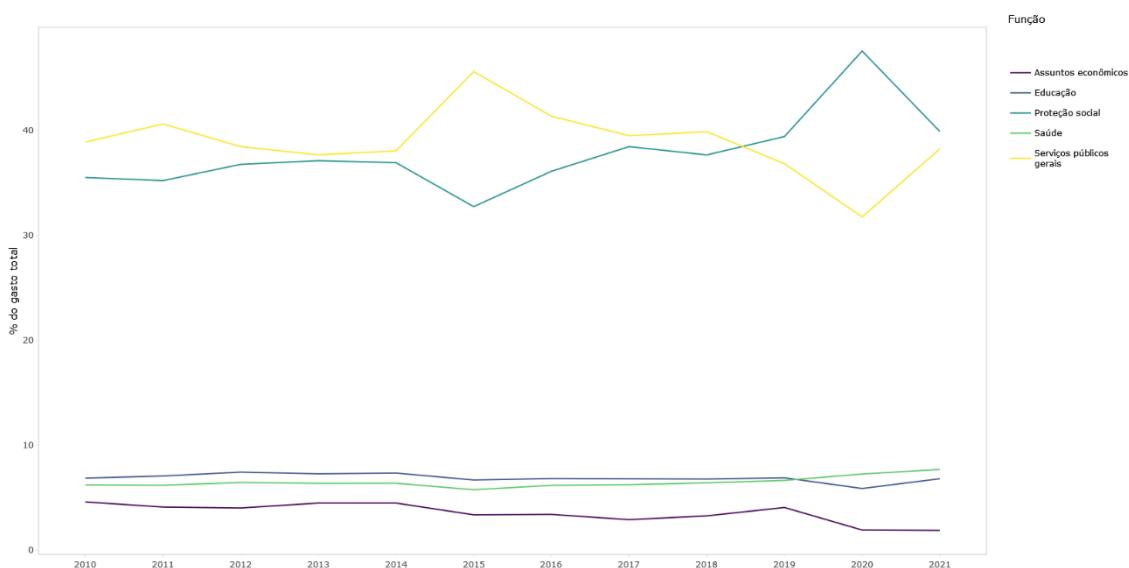


Gráfico de linhas.

Fonte: Brasil (2022b).

O gráfico de linha é bastante útil para se acompanhar a evolução de valores ao longo do tempo. Neste exemplo, note como as diversas despesas por função de governo evoluíram desde 2010. Aqui também surgem padrões. Na verdade, uma das grandes utilidades de análise exploratória de dados é justamente a descoberta de padrões. Vê-se pela figura que a linha amarela que se refere a serviços públicos gerais, principalmente pagamento de juros, foi a principal despesa até 2018 e no ano seguinte foi superada por proteção social. Observe ainda que o pagamento de juros volta a tomar fôlego em 2021, reduzindo a participação da proteção social. Temos então o padrão que envolve a evolução dos gastos com juros e proteção social que lideram o ranking de gastos para todos os anos apresentados no eixo horizontal do gráfico.

Veja novamente o gráfico! Você consegue reconhecer os efeitos da pandemia sobre a priorização de gastos? O que traz de informações sobre o padrão revelado no contraste entre os gastos com saúde, educação e proteção social. Veja que apesar

de percentuais de gastos similares, educação teve fatias de gasto maiores do que saúde até 2019. Já em 2020, houve um maior percentual para a saúde em relação à educação, que permaneceu em 2021. Veja também que a despesa com proteção social subiu fortemente em 2020, principalmente em decorrência do pagamento de assistência emergencial.

É fácil perceber então que o gráfico traz pistas importantes sobre o comportamento dos fenômenos que se quer analisar através dos números. Cumpre-se então, nesse caso, o papel da AED.

Analise agora o próximo gráfico, dessa vez um pouco menos usual, o *box-plot*.

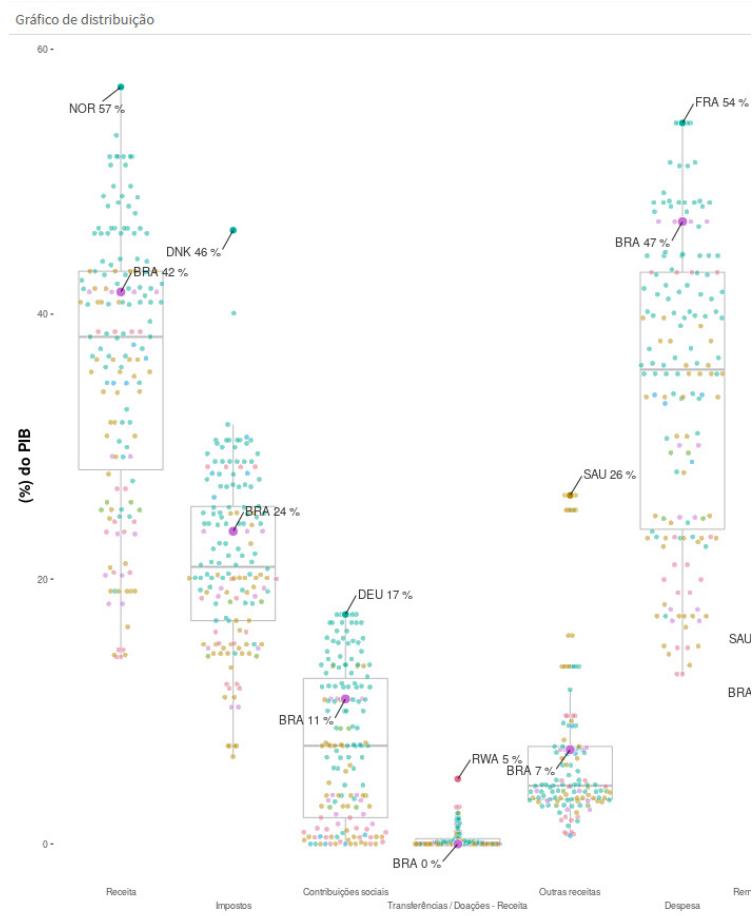


Gráfico box-plot.

Fonte: GFS Internacional.

É interessante explicar que o gráfico de *box-plot* permite verificar como se distribuem os valores de uma variável que se quer analisar. Na figura, temos interesse em entender como se distribuem as receitas e despesas de governo de diversos países. Observe que para todas as variáveis há uma caixa com três retas horizontais.

Essa caixa indica onde se concentram 75% dos valores das variáveis numa faixa intermediária da distribuição. Na figura, o *box-plot* recebe o auxílio de um gráfico de pontos. Cada ponto representa um país. Veja os pontos onde fica o Brasil e compare com outros países em destaque (NOR-Noruega, DNK-Dinamarca, DEU-Alemanha, RWA-Ruanda, SAL-Arábia Saudita).

Observe que em relação às receitas, o Brasil fica dentro da caixa se comparando com os países intermediários. Já em relação à variável da despesa, o Brasil se afasta do caixa central e se aproxima dos países com maiores valores. É fácil comparar a diferença de posição entre o Brasil e a França que tem maior despesa total. Não siga adiante sem antes fazer essa análise, ok?

1.3 Características da Estatística Inferencial

A estatística inferencial é aquela que permite chegar-se à conclusão sobre uma população a partir de uma amostra que a represente. Bussab e Morettin (2010, p. 1) ressaltam que

“Um aspecto importante da modelagem dos dados é fazer previsões, a partir das quais se podem tomar decisões”.

Só para alinharmos os conceitos, a amostra se refere a um subconjunto de uma população que se pretende analisar. Essa amostra, para ser válida precisa ter sido formada a partir de procedimentos reconhecidos dentro dos parâmetros metodológicos apropriados normalmente aceitos pela literatura acadêmica, tais como:

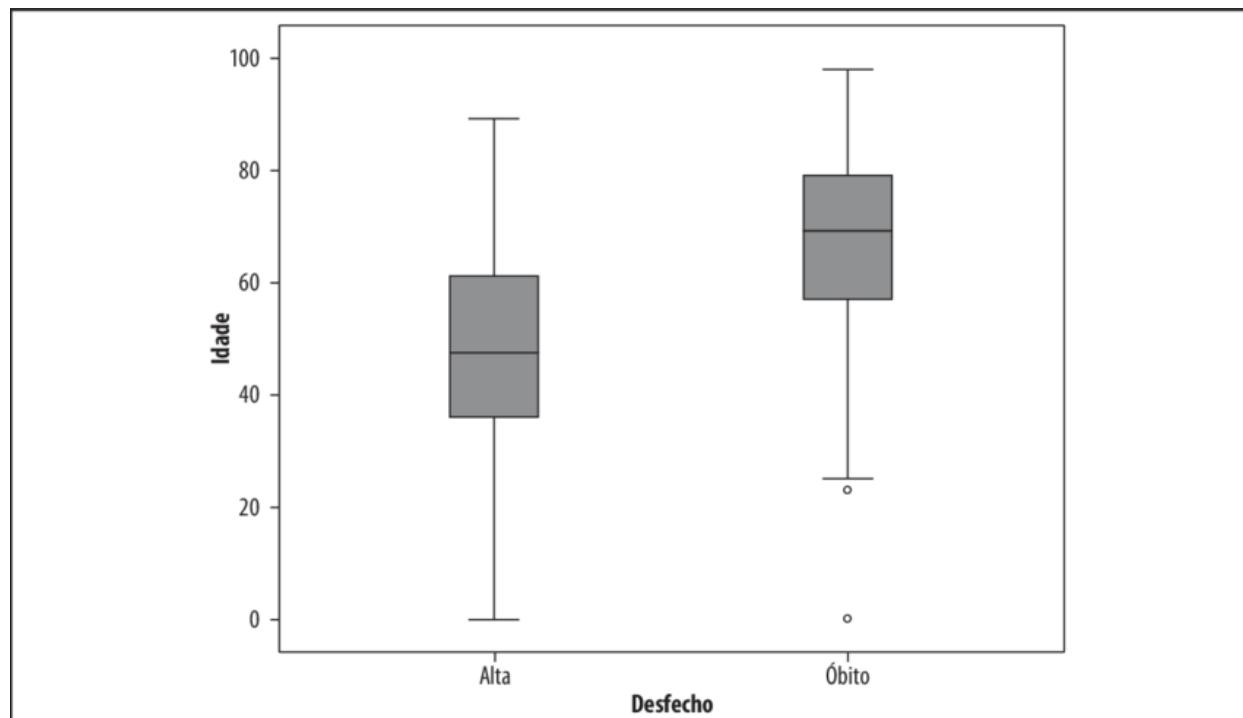
- Tamanho da amostra;
- Aleatoriedade; e
- Heterogeneidade, entre outros.

São exemplos de amostra:

- Conjunto de eleitores que são entrevistados em pesquisa de intenção de votos;
- Conjunto de estudantes que se submetem a avaliações de conhecimentos;
- ou Conjunto de pacientes que se submetem a tratamentos.

Normalmente, num processo de análise de dados é feita uma exploração com os gráficos e as métricas da estatística descritiva para identificar padrões que possam apontar para algum(ns) modelo(s) que são usados em inferência estatística.

Veja esse gráfico a seguir que demonstra um *box-plot* comparando dois grupos.



O gráfico mostra a comparação entre dois grupos de resultados de internação em decorrência de COVID-19. O grupo mais à esquerda mostra a distribuição das idades para os pacientes que tiveram alta e o grupo mais à direita mostra a distribuição das idades para os pacientes que faleceram. Esses dois grupos foram formados a partir de amostras e não de toda a população de pessoas infectadas pela doença e que tiveram que se internar. O analista ao se deparar com esse gráfico vai perceber que os pacientes que tiveram alta possuem idade inferior aos que morreram. Com essa informação, o analista testará modelos que possam gerar uma inferência para a população em modo geral, com o objetivo de medir a diferença entre as médias de idade entre os dois grupos.

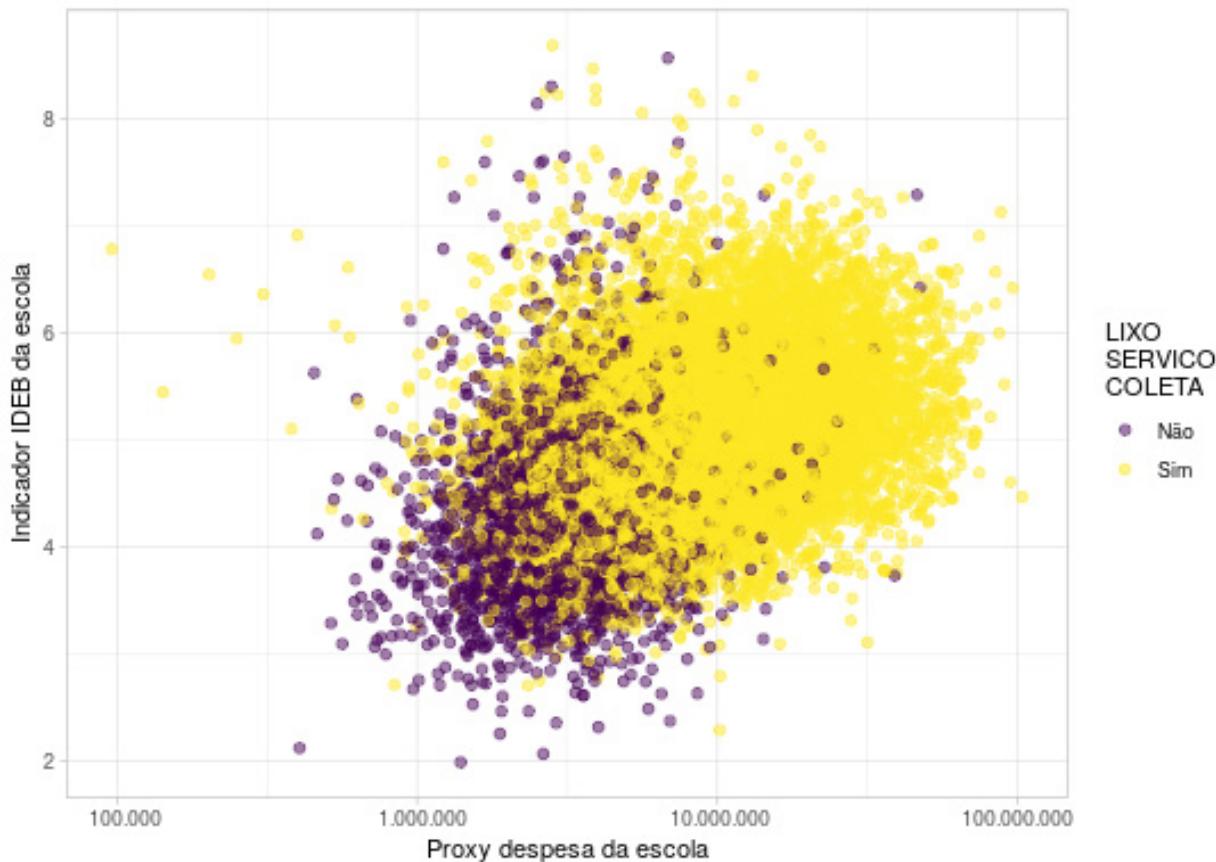


Gráfico com diagrama de dispersão.

Fonte: Brasil (2022c).

A figura apresenta o resultado de uma amostra de escolas públicas no exame do Índice de Desenvolvimento da Educação Básica (IDEB) no Brasil. O analista poderia se basear no gráfico para tentar entender a dimensão da influência que tanto a despesa de uma escola quanto a presença de coleta de lixo naquela escola tem para o resultado do IDEB.

É importante frisar que todos os gráficos mostrados até aqui foram produzidos utilizando o R.

1.4 Probabilidade

A probabilidade está associada ao contraste entre um número de situações favoráveis a um evento e o número total de situações possíveis. O exemplo mais clássico é o do lançamento de um dado. Os livros textos de estatística sempre levantam o problema de calcular a probabilidade de um dado apresentar uma face qualquer após o lançamento do objeto.

Você vai ver nos livros que o cálculo dessa probabilidade resulta em **1/6**, já que 1 corresponde ao número de situações favoráveis para que o dado mostre um valor desejado e 6 é o número total de situações possíveis já que o número de faces de um dado é seis.

Na prática isso significa que se você quiser apostar no número 5 e jogar um dado não viciado um grande número de vezes, 600 vezes, por exemplo, o dado vai apresentar o valor que você apostou em torno de 600/6 ocorrências.



Lançamento de um dado.

Fonte: Freepik (2023).

O estudo de probabilidade é fundamental para o desenvolvimento da estatística inferencial já que muitos dos modelos precisam ter como referência o comportamento probabilístico do fenômeno estudado. O estudo mais aprofundado de probabilidade também faz uso dos conhecimentos de estatística descritiva, principalmente o conceito de média e a compreensão de como devem ser tratadas as distâncias entre valores medidos e as médias calculadas.

Os campos de estudo da probabilidade e estatística inferencial não serão aprofundados, mas já fica claro neste módulo de introdução que a estatística descritiva é fundamental para avanços posteriores sobre esses temas mais complexos. Portanto fica aqui o convite para que você se dedique ao que vem adiante tendo em mente que a estatística descritiva serve não só para apresentar melhor a síntese de achados de exploração de dados como também é ferramental básico para o uso em aplicações mais complexas.

Agora é a hora de você testar seus conhecimentos. Para isso, acesse o exercício avaliativo disponível no ambiente virtual. Bons estudos!

Referências

- BRASIL. Secretaria do Tesouro Nacional. **Análises sobre SICONFI**: Despesas com educação x IDEB. 2022c. Disponível em: https://siconfi-ideb-2019.tesouro.gov.br/dashboard_ideb_2019.Rmd#section-análise-por-outras-variáveis. Acesso em: 8 mar. 2023.
- BRASIL. Secretaria do Tesouro Nacional. **Raio-x das empresas dos Estados brasileiros em 2021**. Brasília, DF: STN, 2022a. Disponível em: <https://tesouro.github.io/empresas-estados/>. Acesso em: 8 mar. 2023.
- BRASIL. Secretaria do Tesouro Nacional. **Série temporal de função de governo**. [Painel COFOG]. 2022b. Disponível em: https://painel-cofog.tesouro.gov.br/dashboard_cofog.Rmd#section-séries-temporais. Acesso em: 8 mar. 2023.
- BUSSAB, W; MORETTIN, O. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2010.
- FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.
- GLOBO CIÊNCIA. **Criada em 1985, identificação por DNA permitiu exames de paternidade**. 2014. Disponível em: <http://redeglobo.globo.com/globociencia/noticia/2013/06/criada-em-1985-identificacao-por-dna-permitiu-exames-de-paternidade.html>. Acesso em: 7 mar. 2023.
- IHAKA, Ross. **The R Project**: a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.
- MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo**. [Epidemiol]. Brasília, DF: ServSaúde, 2020
- MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.
- UNIVERSO ONLINE (UOL). **Quantos planetas e estrelas existem no espaço? Entenda a conta polêmica**. [Tilt UOL]. 2022. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2022/10/15/numero-astronomico-quantos-planetas-e-estrelas-existem-no-espaco.htm>. Acesso em: 7 mar. 2023.
- WALPOLE, R. E. et al. **Probabilidade e estatística para engenharia e ciências**. 8. ed. Tradução: Luciane F. Pauleti Vianna. São Paulo: Pearson Prentice Hall, 2009.
- WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

Módulo

2 Introdução à Linguagem R

Esse módulo abordará os primeiros passos de convivência com a linguagem R e o RStudio, que é o ambiente de desenvolvimento que apoia a construção de bons códigos e análises de dados usando essa linguagem. Você verá que a apresentação da linguagem será feita de forma prática e dirigida às necessidades de análise de dados.

Na primeira unidade será apresentado, rapidamente, o ambiente do RStudio; na segunda unidade você aprenderá os conceitos fundamentais de estatística descritiva com aplicações diretas no R e, por fim, será abordada a organização de dados com o apoio de pacotes específicos da linguagem R. Então finalmente você colocará a mão na massa com os códigos! Avante!

Unidade 1: Funções do Ambiente RStudio

Objetivo de aprendizagem

Ao final desta unidade você será capaz de esclarecer as principais funcionalidades do ambiente Rstudio.

1.1 Como o R Pode Facilitar a sua Vida (Facilidades e Benefícios)

A linguagem R é uma opção de análise de dados em software livre e uma das mais importantes da atualidade.



Linguagem R.

Fonte: Adaptado de Freepik (2023). Elaboração: CEPED/UFSC (2023).

A origem da linguagem está associada à necessidade de dois professores do departamento de estatística na Universidade de Auckland chamados Ross Ihaka e Robert Gentleman que precisavam oferecer suas aulas de introdução à análise de dados com um ferramental mais apropriado às necessidades da estatística (WIKIMEDIA FOUNDATION, 2021).

As linguagens de programação existentes à época, (ano de 1991) não permitiam uma abordagem em que códigos de programação fossem executados parcialmente. Ou se executava todo um programa ou não se executava nada.

Isso é bem diferente das necessidades de um analista de dados que precisa perceber os resultados de seus códigos de forma mais gradual. O que os criadores do R desenvolveram à época como uma resposta a essa demanda foi "uma estrutura básica na qual as pessoas poderiam começar a ligar as coisas" (MIDDLETON, 2009). Assim o estatístico poderia começar com a obtenção do dado de alguma fonte qualquer, depois testar se o dado tem qualidade, em seguida fazer análises de estatística descritiva para observar os primeiros padrões e por aí vai. O usuário teria assim a oportunidade de ir aprendendo com os dados de forma cumulativa, ligando os seus diversos achados.



Linguagem R (dados cumulativos).

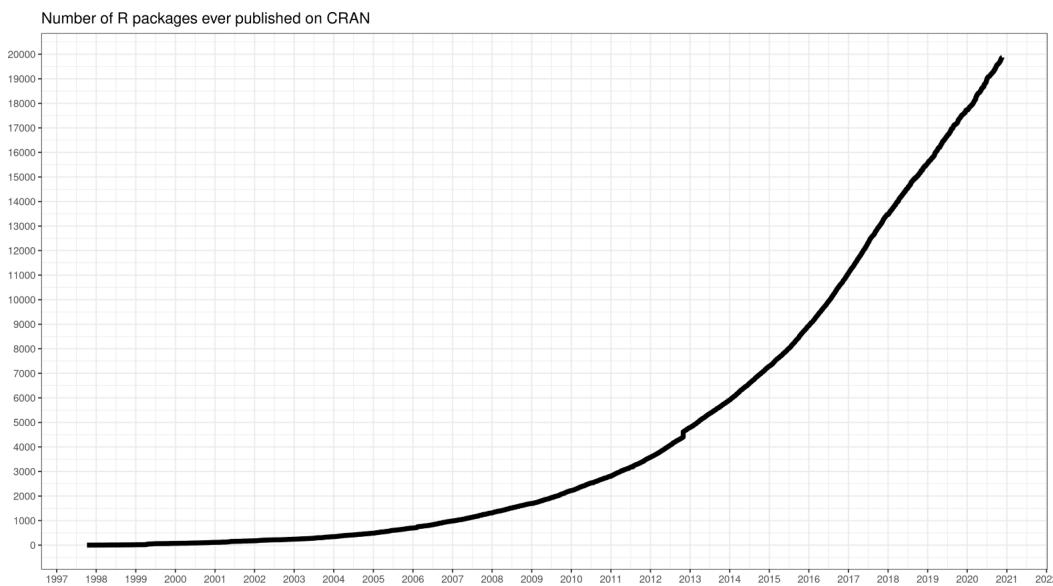
Fonte: Freepik (2023).

Na adoção dessa abordagem cumulativa não seria de se estranhar que os usuários passassem a vislumbrar novas formas de resolver os problemas que estavam enfrentando para além do que a linguagem já oferecia. Os usuários poderiam então desenvolver seus próprios métodos de cálculos e incorporá-los ao seu dia a dia e estender as funcionalidades originais da ferramenta.

A dinâmica da contribuição de terceiros, amplamente favorecida pela opção de tornar a linguagem um software livre, permitindo que as pessoas se sentissem à vontade em devolver para a comunidade as suas extensões de funcionalidade. Normalmente essas extensões são feitas a partir de pacotes, que são conjuntos de códigos que resolvem situações específicas relacionadas aos mais diversos aspectos de tratamento, cálculos e visualização de dados.

Desde abril de 1997 há um esforço de concentrar a maior parte dos pacotes desenvolvidos pela comunidade num único ponto focal, o Comprehensive R Archive Network (CRAN, na sigla em inglês). Isso permite que se localize mais facilmente os pacotes disponíveis e se garanta padronização e uma certa qualidade mínima dos códigos.

Ao longo dos anos o número de pacotes tem crescido substancialmente, denotando o sucesso não só da linguagem como também da opção por uma construção de ecossistema amplamente colaborativa e participativa. O gráfico a seguir demonstra esse crescimento do número de pacotes disponíveis no CRAN.



Crescimento do número de pacotes disponíveis no CRAN.

Fonte: Daróczi (20--).

Como se vê no gráfico, o crescimento do número de pacotes disponibilizados na plataforma CRAN segue acelerado. Ao final de 2021 já havia mais de 20000 pacotes disponibilizados na página do site [CRAN](#).

No uso cotidiano do R você vai notar que existem pacotes para as mais diversas possibilidades de uso. Olha só esse exemplo de um projeto real na figura abaixo.

```
library(tidyverse)
library("basedosdados")
library(cluster)
library(colorspace)
library(geobr)
library(caret)
```

Print de código.

Fonte: Barbalho (2022).

A figura acima demonstra uma lista de seis pacotes mostrados em um texto que trata de análise sobre dados de censos demográficos brasileiros. Cada um desses pacotes tem um conjunto de atribuições bem específicas para as análises que foram feitas. Veja os detalhes.

tidyverse

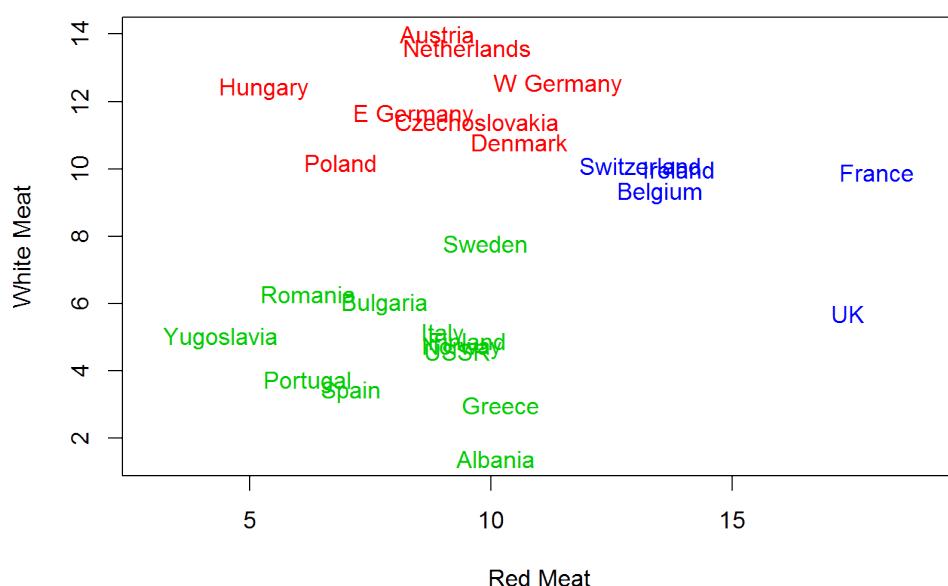
O primeiro pacote, {tidyverse}, é na verdade uma coleção de pacotes. É bastante utilizado atualmente quando se deseja gerar gráficos e manejos de dados tais como filtros, agrupamentos e transformação de dados. Entende-se que o desenvolvimento e disponibilização desse pacote revolucionou o mundo R, ampliando e facilitando o uso da linguagem para públicos sem maior experiência com programação.

basedosdados

O {basedosdados} é uma das ótimas contribuições das organizações da sociedade civil para o conjunto de pessoas e organizações que trabalham com dados abertos no Brasil. Trata-se de um pacote que facilita a consulta de grande conjunto de dados que são disponibilizados principalmente pelo governo federal brasileiro.

clusters

O terceiro pacote, {clusters}, permite operações estatísticas que tratam de agrupamentos, também chamados de clusters. Podemos pensar, por exemplo, que queremos dividir os países a partir do consumo de carne branca e vermelha. Com o uso dessa técnica descobrimos que os países europeus se dividem entre aqueles que consomem ambos os tipos de carne pouco. Além disso, também se pode descobrir os países que preferem consumir somente carne branca e os países que consomem muito somente carne vermelha, como exemplificado no gráfico abaixo.



Consumo de carne vermelha e carne branca nos países europeus.

Fonte: Martos (2014).

colorspace

O pacote {colorspace} disponibiliza paletas de cores apropriadas para os mais diversos tipos de representação gráfica, o que valoriza os relatórios dos achados de análise de dados.

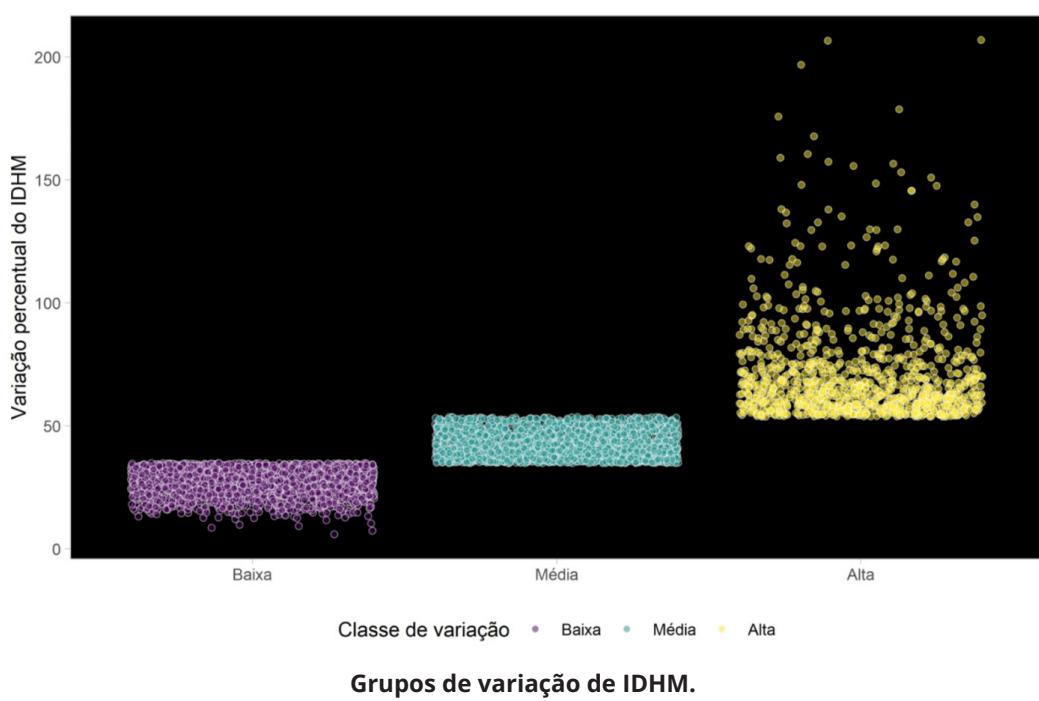
geobr

Já o {geobr} traz as geometrias de diversos temas da cartografia do Brasil. Entre outras aplicações, pode-se pegar as áreas de municípios e estados ou a localização da sede dos municípios e representar esses elementos em forma de mapas.

caret

Por fim, o pacote {caret} permite a operacionalização de uma série de procedimentos típicos de aprendizagem de máquina. São aqueles algoritmos que ajudam a máquina a tomar decisões a partir do que já foi observado de dados anteriores.

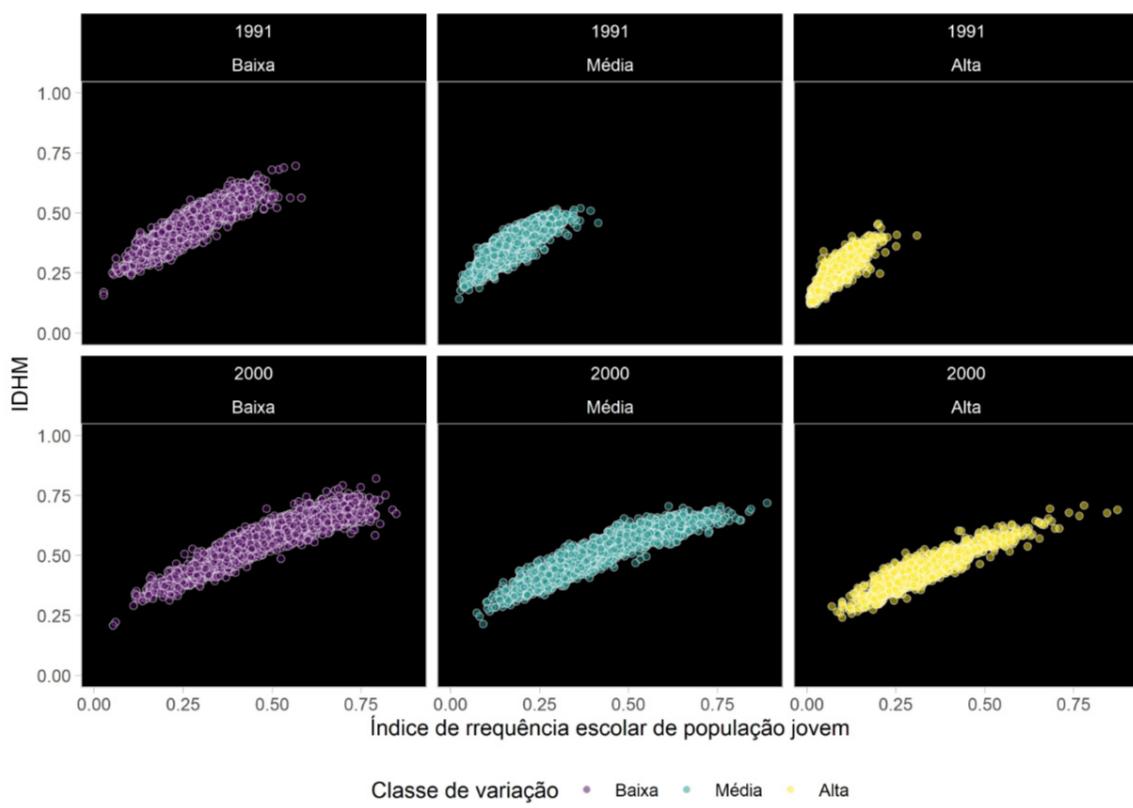
E quando são unidos todos esses pacotes, o que pode sair? Veja adiante algumas imagens que ilustram o estudo que mostra a importância da educação para as subsequentes melhorias do IDH dos municípios ao longo de duas décadas. Todas as figuras foram geradas usando R com o suporte dos pacotes listados nos parágrafos anteriores. Observe!



Na figura, note como os municípios brasileiros se agrupam de acordo com o critério de variação de mudanças de Índice de Desenvolvimento Humano Municipal (IDHM). No caso, existem três grupos de variação: baixa, média e alta. Cada ponto no gráfico é um município.

Aqui se destaca o uso do pacote {cluster} para gerar os três grupos. O pacote {tidyverse} foi usado para fazer as manipulações dos dados e gerar os gráficos, já o {basedosdados} foi utilizado para buscar os dados de IDH de todos os municípios brasileiros para os anos de 1991, 2000 e 2010. Por fim, as cores foram sugeridas pelo pacote {colorspace}.

Esta figura demonstra a conclusão de que a educação foi fundamental na definição dos grupos de municípios.

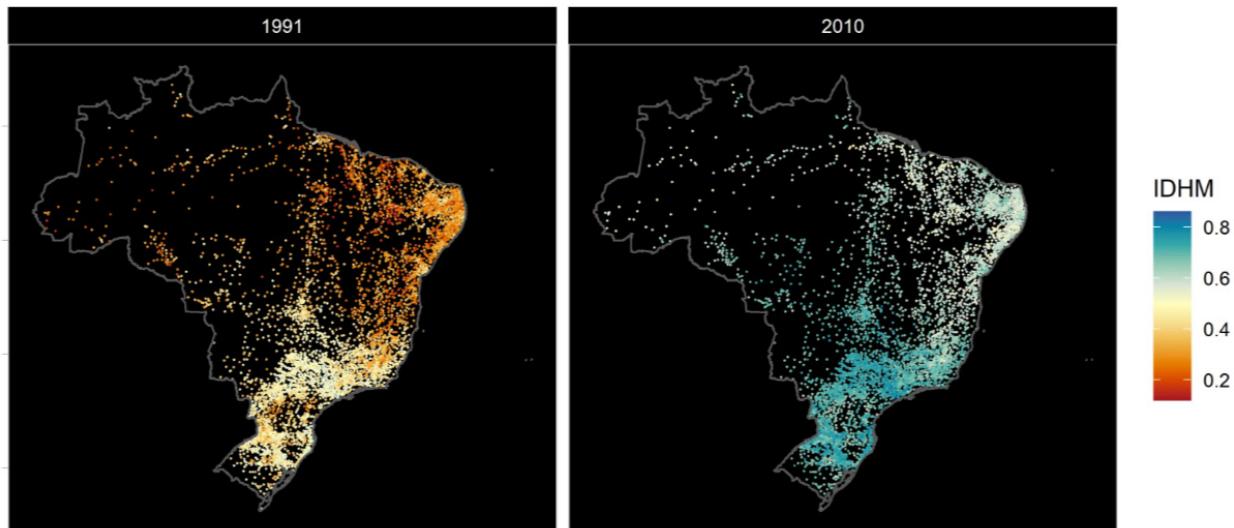


A importância da educação na determinação dos grupos.

Fonte: Barbalho (2022).

Observa-se que na comparação entre os anos de 2000 e de 2010 os IDHs dos municípios aumentaram consideravelmente em função do aumento da frequência da população jovem nas escolas. Observa-se principalmente que esse aumento foi mais importante para o grupo de alto crescimento do IDH. Para se chegar a essa conclusão utilizou-se o pacote {caret}.

Por fim, para finalizar as comparações a figura a seguir compara dois mapas do Brasil. Veja!



Mapa da mudança de IDH entre 1991 e 2010.

Fonte: Barbalho (2022).

O mapa da esquerda mostra como era o Brasil em 1991. Cada ponto no mapa é uma sede de município. As cores em tons vermelhos indicam que a maior parte do país apresentava IDH baixo de acordo com o critério da [ONU](#). Já em 2010 tudo, literalmente, muda de figura. Percebe-se que os municípios estão quase todos com tonalidades azuis, demonstrando que conseguiram alcançar índices satisfatórios de IDH.

Para esta imagem o destaque é o pacote `{geobr}`, que permite trazer o formato do mapa do Brasil e a localização das sedes de cada um dos municípios do país.

Você viu até aqui as várias possibilidades de uso do R em análise de dados e aprendeu um pouco sobre o que motivou os criadores, como o R fomenta uma rede de colaboração e um exemplo real de como essa rede de colaboração acaba por contribuir numa análise de dados abertos que se relaciona com o resultado de políticas públicas no Brasil. Agora conheça um pouco mais sobre o RStudio, o ambiente de desenvolvimento que permite a análise de dados utilizando o R como ferramenta de programação.

1.2 Como Instalar o Ambiente de RStudio

Com a linguagem R instalada em sua máquina você já pode trabalhar com código conforme as funcionalidades da ferramenta. No entanto, o uso profissional ou acadêmico costuma exigir recursos de produtividade que normalmente não são encontrados nessa instalação da linguagem de programação. É para isso que se recomenda o uso de interfaces próprias para desenvolvimento de programas quando se trabalha com análises de dados usando R. Há várias opções disponíveis no mercado, no entanto o enfoque será no ambiente RStudio.



DESTAKE

Atualmente há duas opções interessantes para uso do RSutdio: RStudio Desktop e RStudio Cloud. Para este curso, optou-se pelo RStudio Desktop. Nesse caso é necessário a instalação do software em sua máquina, precedida da instalação da própria linguagem R. Veja a seguir um passo-a-passo de como se faz a instalação do RStudio.

Atualmente há duas opções interessantes para uso do RSutdio: RStudio Desktop e RStudio Cloud. Para este curso, optou-se pelo RStudio Desktop. Nesse caso é necessário a instalação do software em sua máquina, precedida da instalação da própria linguagem R. **Nesse caso é necessário a instalação do software em sua máquina, precedida da instalação da própria linguagem R.**

Para instalar o RStudio em sua máquina [clique aqui](#) e siga o passo a passo apresentado.

Agora que você já tem o R e o RStudio instalado em sua máquina, prossiga para a apresentação do ambiente.

1.3 Conhecendo as Funcionalidades do Ambiente RStudio

O ambiente do RStudio é bastante rico e possibilita uma série de operações que facilitam a programação e análise de dados. Aqui foi adotada a opção de te apresentar visualmente a ferramenta através de um vídeo. Preste atenção no que vai ser exposto e faça as comparações com o seu ambiente de RStudio para fixar bem o entendimento. Acompanhe!



Videoaula: [Apresentando o Ambiente RStudio](#)

1.4 Fazendo o Primeiro *Script* em R

Na videoaula em que foi apresentado o ambiente do RStudio, mostrou-se um primeiro exemplo de *script* em R. Que tal copiar o código e colar para testar em sua máquina? O código é este a seguir:

```
#isso aqui é um comentário. Todo comentário começa com o símbolo #
#O R despreza tudo que estiver numa linha que vier depois de #
#comentários são úteis para documentar o seu código

#Na linha abaixo executa-se um cálculo simples.
(5*3)/2

#No R usamos variáveis ou objetos para armazenarmos valores

#Na linha abaixo atribuímos o valor 5 à variável base_triangulo
#O operador de atribuição de valor a uma variável mais usado é "<-
base_triangulo <- 5

#Na linha abaixo atribuímos o valor 3 à variável altura_triangulo
altura_triangulo<- 3

#Na linha abaixo arribuímos o resultado de (base_triangulo * altura_
triangulo)/2
#à variável area_triangulo

area_triangulo<- (base_triangulo * altura_triangulo)/2

#para ver o valor de area_triangulo execute a linha abaixo. O
resultado aparecerá no console ou na aba Environment

area_triangulo
```

Exemplo de *script* em R/ Código.

Fonte: Barbalho (2020).

Experimente copiar, colar e executar. Lembre-se: se tiver dúvidas sobre como se faz, reveja a videoaula, mas, espera-se que tenha funcionado perfeitamente. Então é hora de praticar! Realize mais alguns exercícios.

Exercício 1

Faça um código que tenha uma variável “lado” e a partir dessa variável seja possível calcular a área de um quadrado de “lado =5”. Lembre-se que a área de um quadrado é dada pela multiplicação: “lado” x “lado”. Nesse caso, o resultado do seu código deve ser 25. Vamos lá?

Exercício 2

Escreva um *script* que mostre o resultado de uma empresa que teve receita de 25 milhões e despesa de 13 milhões. Nesse *script* você vai ter três variáveis: “receita”, “despesa” e “resultado”. O resultado vai ser o valor de receita - despesa. Certifique-se que ao final o “resultado” foi de 12 milhões. Avante!

Exercício 3

Aproveite o *script* anterior e troque os valores de receita com a despesa e calcule o novo resultado. Feito?

Exercício 4

Sabendo que a nota final de um aluno ao final do curso é dada pela soma das duas notas parciais dividido por dois, calcule a nota final de um aluno que tirou **8** na primeira nota parcial e **5** na segunda nota parcial. Lembre-se de criar três variáveis “Nota1”, “Nota2” e “NotaFinal”. Certifique-se que ao concluir a execução do seu *script* o valor da variável “NotaFinal” será **6.5**.

Antes de encerrar essa unidade, é importante saber também acerca de algumas observações sobre objetos do R:

*Os nomes dos objetos são reconhecidos de formas diferentes se você usar letra maiúscula ou minúscula. Observe o exemplo a seguir.

```
base_triangulo<- 5 #0 nome do objeto começa com b minúsculo  
Base_triangulo<- 7 #0 nome do objeto começa com B maiúsculo  
  
base_triangulo  
Base_triangulo
```

*Se você executar as quatro linhas acima, o resultado que sairá no console será:

```
> base_triangulo<- 5  
> Base_triangulo<- 7  
> base_triangulo  
[1] 5  
> Base_triangulo  
[1] 7
```

*Até aqui você trabalhou com objetos do tipo numérico. É importante dizer que também é possível trabalhar com objetos do tipo texto. Para isso, os valores atribuídos aos objetos devem estar entre aspas. Veja o exemplo abaixo:

```
 objeto_tipo_texto<- "Isso é um objeto de tipo texto"  
 objeto_tipo_texto
```

*Se executar as duas linhas acima você verá o seguinte resultado no console:

```
> objeto_tipo_texto<- "Isso é um objeto de tipo texto"  
> objeto_tipo_texto  
[1] "Isso é um objeto de tipo texto"
```

Até aqui você reconheceu as principais características do ambiente RStudio e experimentou os primeiros códigos. Parabéns por seu comprometimento!

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, F. **Education as the driver of human development in Brazil**: An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN**. GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

MARTOS, G. **Cluster analysis with R**. RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

POSIT SOFTWARE. **RStudio Desktop**. S.d. Disponível em: <https://posit.co/download/rstudio-desktop/>. Acesso em: 14 mar. 2023.

WIKIMEDIA FOUNDATION. **R(linguagem de programação)**. [Wikipédia]. 2021. Disponível: [https://pt.wikipedia.org/wiki/R_\(linguagem_de_programa%C3%A7%C3%A3o\)](https://pt.wikipedia.org/wiki/R_(linguagem_de_programa%C3%A7%C3%A3o)). Acesso em: 24 mar. 2023.

Unidade 2: Utilizando a Estatística para Simplificar um Conjunto de Informações por Distribuição de Frequências

Objetivo de aprendizagem

Ao final desta unidade você será capaz de reconhecer o processo de operacionalização e sumarização de dados através de tabelas de distribuição de frequências.

2.1 Tipos de Variáveis

Ao trabalhar com estatística, as características que se deseja medir são operacionalizadas nas chamadas variáveis. Se quisermos, por exemplo, verificar algumas características dos estados do Nordeste do Brasil, pode -ser que haja interesse em identificar a sigla dos estados, suas capitais, suas áreas e suas populações. Cada uma dessas características são variáveis para um estudo sobre o Nordeste. Nesse caso é possível pensar na construção de uma estrutura de dados como a que está organizada no infográfico que você deve analisar a seguir.

estado	capital	área	população
MA	São Luiz	331937,4	7114598
PI	Teresina	251577,7	3264531
CE	Fortaleza	146348,3	9132078
RN	Natal	52809,6	3534165
PB	João Pessoa	56467,2	4039277
PE	Recife	98149,1	9616621
AL	Maceió	27848,1	3337357
SE	Aracaju	21915,1	2278308
BA	Salvador	564733,1	14930634

Estados do Nordeste.

Fonte: Wikimedia Foundation (2022).

Tudo analisado?! Antes de prosseguir nas caracterizações dos tipos de variáveis, é importante ter clareza sobre os aspectos conceituais de uma tabela, como a disponível acima. Tabelas são estruturas de dados que se caracterizam fundamentalmente pela organização dos dados em linhas e colunas. Veja como essa definição se reflete na tabela sobre os estados do Nordeste:

Valores da primeira linha:

MA	São Luiz	331937,4	7114598
----	----------	----------	---------

Valores da segunda linha:

PI	Teresina	251577,7	3264531
----	----------	----------	---------

Valores da primeira coluna:

estado
MA
PI
CE
RN
PB
PE
AL
SE
BA

Valores da quarta coluna:

população
7114598
3264531
9132078
3534165
4039277
9616621
3337357
2278308
14930634



IMPORTANTE

Supõe-se que com a descrição anterior você já consiga identificar quais são os valores que estão associados às outras linhas e colunas. Antes de prosseguir com a leitura tire um tempo e analise a tabela: quais características dos estados do Nordeste estão presentes na segunda e terceira colunas?

Observe que na tabela os nomes área e população estão com ortografia não compatível com a língua portuguesa. Mais à frente justificaremos essa opção.

Feito? A partir da sua análise você deve ter percebido que a segunda e terceira colunas dizem respeito respectivamente a **capital** e **área**. Perceba que as variáveis em estruturas de tabelas estão sempre dispostas em colunas. Nesse caso têm-se quatro colunas, logo estamos interessados em quatro variáveis: estado, capital, área e população.

Se continuar suas observações sobre a tabela, você verá que algumas colunas apresentam textos e outras apresentam apenas números. Essa diferenciação é importante para as possibilidades analíticas que envolvem cada uma das variáveis. Vá em frente para explorar essa distinção!

Variáveis Categóricas

As variáveis que envolvem apenas textos ou combinação de textos com números são conhecidas como categóricas, ou alfanuméricas. No caso da tabela dos estados do Nordeste, encontram-se duas variáveis que se caracterizam como categóricas: estado e capital.

Uma das principais implicações das caracterizações das variáveis é a restrição das possibilidades de uso em operações de análise de dados. Como é de se esperar, não é possível fazer operações matemáticas usando variáveis categóricas, por exemplo. Por outro lado, pode-se fazer ordenações, filtros e agrupamentos a partir dessas variáveis. A seguir está exemplificada como fica a tabela original reordenada alfabeticamente a partir da variável categórica estado.

estado	capital	área	população
AL	Maceió	27848,1	3337357
BA	Salvador	564733,1	14930634
CE	Fortaleza	146348,3	9132078
MA	São Luiz	331937,4	7114598
PB	João Pessoa	56467,2	4039277
PE	Recife	98149,1	9616621
PI	Teresina	251577,7	3264531
RN	Natal	52809,6	3534165
SE	Aracaju	21915,1	2278308

Estados do Nordeste em ordem alfabética pela coluna estado.

Fonte: Wikimedia Foundation (2022).

Volte à tabela original e observe que nela não havia nenhuma ordem aparente, ao contrário dessa nova construção em que as linhas são organizadas pela ordem alfabética da sigla dos estados. Dessa forma a linha referente ao estado Alagoas (sigla AL), aparece agora antes da Bahia (sigla BA) e assim sucessivamente até chegar a Sergipe (sigla SE).

As variáveis “estado” e “capital” são conhecidas como categorias **nominais**. Esse tipo de variável é apresentada em formato de texto e não carrega consigo um valor implícito de ordem. Quando uma variável alfanumérica tem significados de ordem nos valores que podem assumir, elas são conhecidas como **variáveis ordinais**.

Pode-se, por exemplo, trabalhar com dados sobre alunos de uma escola e uma das variáveis referir-se à série do aluno. Nesse caso os valores possíveis seriam: “primeiro ano”, “segundo ano”, “terceiro ano” e “quarto ano”. Há implicitamente uma ordem entre esses valores que costuma ter muito mais potencial analítico do que uma ordem alfabética. Observe que se aplicássemos a ordem alfabética a esses valores a ordem seria “primeiro ano”, “quarto ano”, “segundo ano” e “terceiro ano”.

Variáveis Quantitativas

Quando se fala sobre as variáveis numéricas ou quantitativas o conjunto de possibilidades de uso se diferencia bastante. Com essas variáveis é possível aplicar uma infinidade de operações matemáticas que vão de simples somas a cálculos complexos envolvendo integrais e derivadas. Além, é claro, da ordenação de valores, tal como é possível nas variáveis categóricas. Veja abaixo alguma das possibilidades.

estado	capital	área	população
BA	Salvador	564733,1	14930634
PE	Recife	98149,1	9616621
CE	Fortaleza	146348,3	9132078
MA	São Luiz	331937,4	7114598
PB	João Pessoa	56467,2	4039277
RN	Natal	52809,6	3534165
AL	Maceió	27848,1	3337357
PI	Teresina	251577,7	3264531
SE	Aracaju	21915,1	2278308

Ordenação da tabela pela população em forma descendente.

Fonte: Wikimedia Foundation (2022).

Veja o impacto sobre a tabela quando se reordena os valores pela variável quantitativa “população”. Agora as três primeiras linhas se referem a Bahia, Pernambuco e Ceará. Veja que nessa ordem a população da Bahia é maior que a de Pernambuco, que é maior do que a do Ceará. Observe ainda que a última linha continua sendo ocupada por Sergipe, já que esse estado possui a menor população entre os nove do Nordeste.

Soma das populações dos estados do Nordeste: **57.247.569**.

Média da área ocupada pelos estados do Nordeste, calculada a partir da soma das áreas dos estados dividido pelo número de estados: **172.420,62**.

Observe que foram usadas duas operações matemáticas distintas, soma e média, uma operação para cada variável numérica disponível na tabela.

Observe ainda que a variável **área** apresenta valores com casas decimais, duas unidades após a vírgula, enquanto a variável **população** apresenta os valores inteiros. Essa característica gera uma distinção para as variáveis quantitativas que podem ser divididas em:

variáveis contínuas que se relacionam com valores que fazem sentido serem medidos na escala real, tais como área, temperatura, cotação do dólar, peso e altura; e

variáveis discretas que se relacionam com contagens como população de um estado, quantidade de alunos em sala de aula, quantidade de leitos em hospitais, quantidade de beneficiários de política pública.

Até aqui você viu os conceitos de variáveis, os seus tipos e sub-tipos. O próximo passo é assistir a uma videoaula que demonstra como se trabalha com variáveis e tabelas no R. Depois você terá contato com uma lista de sugestões de práticas usando o R.



Videoaula: [Criação Dataframe](#)

Agora é hora de pôr as mãos no código! Abaixo está o script que foi apresentado na videoaula. Copie e cole para seu RStudio.

```

#A função c cria um objeto que se associa a um conjunto de valores
de um mesmo tipo

#na linha abaixo criamos um conjunto chamado nome_escola composto
por três elementos tipo texto
nome_escola<- c("Elefante Branco", "Liceu", "Pedro II")
nome_escola

#na linha abaixo criamos um conjunto chamado quantidade_alunos_
inscritos composto por três elementos tipo numérico discreto
quantidade_alunos_inscritos<- c(200, 250, 300)
quantidade_alunos_inscritos

#na linha abaixo criamos um conjunto chamado media_escola composto
por três elementos tipo numérico contínuo
media_escola<- c(400.34 , 453.27, 425.32)
media_escola

#No R um dos tipos de tabela que usamos é o dataframe.
#Para criar um data.frame usamos a função data.frame.
#Na função data.frame, para cada coluna você deve indicar o nome da
coluna seguido pelo símbolo "="
#e pelo conjunto de dados referente a essa coluna

#na linha abaixo criamos um data.frame chamado df_enem que é formada
pela combinação dos três conjuntos criados anteriormente
df_enem<- data.frame(nome_escola=nome_escola, #nome da coluna =
conjunto de dados da coluna

quantidade_alunos_inscritos=quantidade_alunos_inscritos, #nome da
coluna = conjunto de dados da coluna
media_escola=media_escola) #nome da coluna =
conjunto de dados da coluna

df_enem

```

Exemplo de script em R/ Código.

Fonte: Barbalho (2020).

Agora faça as adaptações no código que permitam responder ao que se pede. Altere o conjunto de dados sobre média de forma que este passe a possuir os seguintes valores: 404.45, 487.2 e 452.23. Em seguida execute novamente o script e certifique-se de que o objeto df_enem tenha ao final a configuração indicada abaixo:

```

> df_enem
      nome_escola quantidade_alunos_inscritos media_escola
1 Elefante Branco                      200        404.45
2           Liceu                      250        487.20
3       Pedro II                      300        452.23

```

Exemplo de script em R/ Código.

Fonte: Barbalho (2020).

2.2 Distribuição de Frequências

Uma das formas mais práticas e com resultados mais eficientes para se organizar variáveis categóricas é a disponibilização dos dados através de um tipo de tabela especial conhecida como tabela de frequência.

A frequência na linguagem estatística corresponde ao número de vezes que uma determinada classe de uma variável categórica se manifesta. Explore a tabela a seguir para verificar essa aplicação. Ela apresenta, hipoteticamente, uma coleta de dados sobre o departamento de pessoal de um órgão de governo.

Id_servidor	sexo	formação	Iotação	Remuneração
1	Masculino	Administração	RH	8000
2	Masculino	Engenharia	Financeiro	7500
3	Masculino	Economia	Financeiro	8200
4	Feminino	Administração	Atendimento	8300
5	Feminino	Economia	Planejamento	8200
6	Feminino	Medicina	Perícia Médica	8300
7	Feminino	Psicologia	RH	8250

Frequência na linguagem estatística.

Fonte: Barbalho (2020).

Note que a tabela possui sete linhas com cinco colunas. Três dessas colunas são categóricas e podem ser resumidas com certa facilidade em uma tabela de frequência. Aqui queremos contar quantas ocorrências existem de cada um dos dois sexos: masculino e feminino. Veja como fica esse resumo na primeira tabela de frequência.

sexo	Frequência absoluta
Masculino	3
Feminino	4

Podemos também experimentar checar a frequência por formação.

formação	Frequência absoluta
Administração	2
Economia	2
Engenharia	1
Medicina	1
Psicologia	1

Resumo de frequência na linguagem estatística.

Fonte: Barbalho (2020).

Claro que é possível fazer a mesma coisa com a lotação, e isso fica como sugestão de exercício pra você, combinado?

Uma variação importante da distribuição de frequência é a que inclui a informação da frequência relativa, ou seja, o percentual alcançado por cada classe. Veja como ficam esses dois exemplos que você acabou de visualizar com essa informação adicionada.

sexo	Frequência absoluta	Frequência relativa
Masculino	3	42,8%
Feminino	4	57,2%

formação	Frequência absoluta	Frequência relativa
Administração	2	28,6%
Economia	2	28,6%
Engenharia	1	14,3%
Medicina	1	14,3%
Psicologia	1	14,3%

Exemplos de frequência na linguagem estatística.

Fonte: Barbalho (2020).

Pode ser interessante calcular a frequência acumulada observando uma ordem, crescente ou decrescente dos valores de frequência. Veja como fica.

formação	Frequência simples		Frequência acumuladas	
	Absoluta	Relativa	Absoluta	Relativa
Administração	2	28,6%	2	28,6%
Economia	2	28,6%	4	57,2%
Engenharia	1	14,3%	5	71,5%
Medicina	1	14,3%	6	85,8%
Psicologia	1	14,3%	7	100%

Frequência acumulada.

Fonte: Barbalho (2020).

Apesar dos exemplos indicados serem todos relacionados a variáveis categóricas, você vai explorar, em seguida, como faz sentido fazer esse tipo de resumo de dados para variáveis quantitativas.

Agora que você aprendeu o conceito, o próximo passo é entender como a distribuição de frequência pode ser feita usando o R. Assista a videoaula a seguir que aborda este assunto.



Videoaula: [Construção de Distribuição de Frequências Usando o R](#)

E como você já deve imaginar, depois de assistir é a sua vez de praticar. Adiante você tem acesso ao código utilizado na videoaula. Copie, cole e execute o *script* no seu ambiente RStudio, de preferência em um arquivo R *script* novo. Caso tenha alguma dúvida, assista a videoaula novamente para entender o que foi feito.

```

#Para instalar um novo pacote no R usamos a função install.package
#A linha abaixo instala o pacote questionr.
install.packages("questionr")

#a linha abaixo carrega o pacote na memória
library(questionr)

#Cria dois conjuntos de dados: nome_hospital, rede_hospital
nome_hospital<- c("São Sebastião","João Paulo II","Trindade","São Clemente","Hospital Regional", "Hospital Geral")
rede_hospitalar<-
c("municipal","municipal","federal","estadual"," estadual","municipal")
#cria o data.frame hospitais com as colunas nome_hospital e rede_hospital
hospitais<- data.frame(nome_hospital= nome_hospital, rede_hospitalar=
rede_hospitalar)

#para se verifica o conteúdo de uma coluna específica de um data.frame use
#a notação <data.frame>$<nome_coluna>

#Na linha abaixo você vai ver o conteúdo da coluna nome_hospital
hospitais$nome_hospital

#Na linha abaixo você vai ver o conteúdo da coluna rede_hospitalar
hospitais$rede_hospitalar

#A função freq presente no pacote questionr possibilita gerar uma tabela
de frequência
#A função freq pede que você informe alguns parâmetros, os principais
são:
#x: o conjunto de dados que você deseja fazer a tabela de frequência
#cum: indicação se você quer que a tabela de frequência apareça ou não
os valores acumulados (TRUE/FALSE)
#sort: indicação se você quer que os valores sejam ordenados em ordem
crescente ("inc") ou decrescente ("dec")

#Na linha abaixo geramos uma tabela de frequência sobre os dados de rede
hospitalar, com frequência acumulada em ordem decrescente
questionr::freq(x= hospitais$rede_hospitalar, cum = TRUE, sort = "dec")

```

Exemplo de R/ Código.

Fonte: Barbalho (2020).

Agora reescreva o código para gerar uma tabela de frequência em ordem crescente. O resultado final deve ser o que está logo abaixo.

	n	%	val%	%cum	val%cum
federal	1	16.7	16.7	16.7	16.7
estadual	2	33.3	33.3	50.0	50.0
municipal	3	50.0	50.0	100.0	100.0

Tabela de frequência em ordem crescente.

Fonte: Barbalho (2020).

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN**. GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

IHAKA, Ross. **The R Project**: a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo**. [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R**. RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

Unidade 3: Utilizando a Estatística para Simplificar um Conjunto de Informações por Medidas Resumo

Objetivo de aprendizagem

Ao final desta unidade você será capaz de reconhecer como simplificar um conjunto de informações por meio de medidas de centralidade e dispersão.

3.1 Medidas de Centralidade

“ Conforme Bussab e Morettin (2010) as medidas de centralidade (ou medidas de posição central) são aquelas que resumem os elementos de um conjunto de dados em um único valor representativo. As três medidas mais importantes são: **moda, média aritmética e mediana.** ”

Moda

A **moda** deve ser entendida como a ocorrência mais frequente de um conjunto de dados. Vejamos a tabela abaixo com as idades dos funcionários fictícios do órgão de governo que estamos analisando.

Funcionário	Idade
1	40
2	38
3	39
4	39
5	39
6	43
7	45

Se fizermos uma tabela de frequências para as idades teremos:

Idade	Frequência absoluta
38	1
39	3
40	1
43	1
45	1

Pela tabela de frequência logo acima fica fácil perceber que a ocorrência mais frequente é da idade de 39 anos, com a frequência de 3 observações.

Média Aritmética

A **média aritmética** é dada pela soma dos valores dividida pelo número de observações. Trata-se do conceito familiar que você provavelmente já usou no colégio para saber a nota que precisava para “passar de ano”. Lembra-se disso?

Naquela época, para saber a média final, você tinha de somar as notas em cada prova e dividir pelo número de provas. Se quiser aplicar esse conceito para as idades dos funcionários, temos:

$$\text{Média de idade} = (40 + 38 + 39 + 39 + 39 + 43 + 45)/7 = 40,4 \text{ anos}$$

Mediana

Finalmente, a **mediana** refere-se ao valor ocupado pela posição central de um conjunto de dados depois de ordenado em ordem crescente. Para o caso de um número ímpar de elementos é exatamente o ponto que divide o conjunto em dois subconjuntos, onde um deles corresponde aos números iguais ou menores que o valor observado no elemento que ocupa a posição central e o outro corresponde aos números iguais ou maiores que o valor do elemento da posição central. Use a fórmula abaixo para calcular a posição central:

$$\text{Posição central} = (n + 1)/2$$

Se o número de elementos for par, “usa-se como mediana a média aritmética das duas observações centrais” (BUSSAB; MORETTIN, 2010)

Veja o caso das idades dos funcionários. Observe a tabela abaixo já reordenada em ordem crescente das idades.

Linha	Funcionário	Idade
1	1	40
2	2	38
3	3	39
4	4	39
5	5	39
6	6	43
7	7	45

Como a tabela tem sete funcionários, vamos usar a regra do número ímpar de elementos, ou seja, observar a posição central, que no caso é a quarta linha. Cabe lembrar que essa posição central é dada pelo resultado da fórmula $(n + 1)/2$, nesse caso $(7 + 1)/2$. Observe que essa posição é ocupada pelo funcionário 5. A idade desse funcionário é 39 anos, portanto a mediana é 39.

Vamos imaginar que um oitavo funcionário passou a fazer parte da organização e esse funcionário tem 44 anos. A nova tabela será essa.

Linha	Funcionário	Idade
1	2	40
2	3	38
3	4	39
4	5	39
5	6	39
6	7	43
7	8	45

Nesse caso a mediana passa a ser calculada pela regra do número par de elementos. As posições centrais são as das linhas 4 e 5, ocupadas pelos funcionários 5 e 1, cujas idades são 39 e 40 anos. Dessa forma a mediana passa a ser: $(39 + 40)/2 = 39,5$ anos.

A opção entre a média ou a mediana deve levar em conta qual das medidas melhor caracteriza o conjunto de dados. Veja a videoaula do professor Alexandre Patriota sobre média e mediana para entender alguns pontos importantes para a seleção de medida de centralidade ([clique aqui](#)).

Antes de fechar esse tema é importante ter em mente que as medidas de centralidade são subconjuntos das medidas de posição mencionadas pelo professor. Para além dessas medidas de centralidade, existe também a utilização do conceito de quantis que permite abordar outros pontos da distribuição dos dados. Para que você tome contato com esse conceito, assista a essa outra videoaula do professor Alexandre ([clique aqui](#)).

3.2 Medidas de Dispersão

“ Bussab e Morettin (2010, p. 37) indicam que “o resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações”. ”

Imagine que foram coletadas as notas obtidas por três grupos de alunos de escolas distintas num exame de proficiência em matemática. Essas notas estão indicadas na tabela.

Grupo 1	Grupo 2	Grupo 3
6,6	5,4	6,2
5,8	5,3	4,7
5,3	4,4	3,5
7	7	7,1
6,2	5,9	5,5
6,5	6,2	5,9
8	8,5	9
7,5	7,8	8
7,2	7,2	7,3
7,1	7,2	7,2

As médias dos três grupos são respectivamente: 6,7; 6,6; e 6,4. Veja que os valores da média não estão tão distantes entre si, mas observe os intervalos das notas dos grupos. No grupo 1 as notas variam entre 5,3 e 8. No grupo 2 essa variação é entre 4,4 e 8,5. Já no grupo 3 os extremos são ainda mais distantes: 3,5 e 9. Pensando nessas situações é que foram desenvolvidas as medidas de dispersão que procuram, principalmente, medir o quanto os dados se afastam das médias.

Amplitude

A primeira medida de dispersão é bem simples de ser observada. Trata-se da **amplitude**. Essa medida é dada pela subtração do valor máximo pelo valor mínimo de cada conjunto de dados. Dessa forma, tem-se as amplitudes a seguir.

Amplitude do Grupo 1

Valor máximo do grupo 1 – valor mínimo do grupo 1 = $8 - 5,3 = 2,7$

Amplitude do Grupo 2

Valor máximo do grupo 2 – valor mínimo do grupo 2 = $8,5 - 4,4 = 4,1$

Amplitude do Grupo 3

Valor máximo do grupo 3 – valor mínimo do grupo 3 = $9 - 3,5 = 5,5$

Como se vê, a amplitude explicita em números aquela primeira constatação que os extremos dos três grupos variavam bastante, mesmo a média sendo bem parecida.

Por mais que a amplitude seja uma medida interessante e que já permita ao analista perceber facilmente as discrepâncias dentro de cada grupo e compará-las entre os grupos, em grande parte das vezes é preferível uma medida que indique o quanto os valores se afastam da média.

A Variância

A **variância** ou **VAR** é uma das medidas que indicam o quanto o conjunto de dados se afasta da média e existe uma fórmula para calcular essa medida. Para facilitar a compreensão do cálculo vamos apresentar essa fórmula através de um exemplo prático usando tabelas com os valores dos grupos de notas.

Antes de começar, é interessante frisar um elemento importante da estatística que é o tamanho dos conjuntos de dados, normalmente representado pela letra **N**. No caso dos três grupos, o tamanho é 10, logo N = 10.

Começando com o grupo 1.

Nota	Média	(nota - média)	$(nota-média) ^ 2$
6,6	6,7	-0,1	0,1
5,8	6,7	-0,9	0,81
5,3	6,7	-1,4	1,96
7	6,7	0,3	0,09
6,2	6,7	-0,5	0,25
6,5	6,7	0,2	0,04
8	6,7	1,3	1,69
7,5	6,7	0,8	0,64
7,2	6,7	0,5	0,25
7,1	6,7	0,4	0,16
Soma de $(nota-média) ^ 2$			5,9
Soma de $(nota-média) ^ 2 / N$			0,59

A última linha da tabela acima é a variância das notas do grupo 1. Eis os passos para se chegar a esse valor:

- ① na primeira coluna incluímos todas as notas alcançadas pelos alunos;
- ② na segunda coluna informamos a média do grupo, portanto 6,7;
- ③ na terceira coluna subtraímos cada nota pela média do grupo;
- ④ na quarta coluna elevamos ao quadrado o valor encontrado para cada elemento da terceira coluna;
- ⑤ na décima primeira (penúltima) linha calculamos a soma dos valores encontrados na quarta coluna;
- ⑥ na décima segunda (última) linha chegamos ao resultado da variância dividindo o valor encontrado na penúltima linha pelo tamanho do conjunto de dados (que é 10).

Repete-se o mesmo procedimento para o grupo 2...

Nota	Média	(nota - média)	$(nota-média)^2$
6,4	6,6	-0,2	0,04
5,3	6,6	-1,3	1,69
4,4	6,6	-2,2	4,84
7	6,6	0,4	0,16
5,9	6,6	-0,7	0,49
6,2	6,6	-0,4	0,16
8,5	6,6	1,9	3,61
7,8	6,6	1,2	1,44
7,2	6,6	0,6	0,36
7,2		0,6	0,36
Soma de $(nota-média)^2$			13,15
Soma de $(nota-média)^2 / N$			1,315

... e também para o grupo 3

Nota	Média	(nota - média)	$(nota-média)^2$
6,2	6,4	-0,2	0,004
4,7	6,4	-1,7	2,89
3,5	6,4	-2,9	8,41
7,1	6,4	0,7	0,49
5,5	6,4	-0,9	0,81
5,9	6,4	-0,5	0,25
9	6,4	2,6	6,76
8	6,4	1,6	2,56
7,3	6,4	0,9	0,81
7,2	6,4	0,8	0,64
Soma de $(nota-média)^2$			23,66
Soma de $(nota-média)^2 / N$			2,366

Com as três variâncias calculadas é possível perceber que o maior valor dessa medida é para o grupo 3, com 2,35. De certa forma isso confirma a dispersão mais elevada que já havia sido identificada com a amplitude.

O Desvio Padrão ou DP

O **desvio padrão**, ou **DP**, é a terceira medida de dispersão que vamos trabalhar. Conforme Bussab e Morettin (2010, p. 28) “sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm²), pode causar problemas de interpretação. Costuma-se usar, então, o *desvio padrão*, que é definido como a raiz quadrada positiva da variância”.

Com essa definição, pode-se calcular o desvio padrão de cada um dos grupos:

$$\text{Desvio padrão do Grupo 1} = \sqrt{\text{VAR}} = \sqrt{0,59} = 0,77$$

$$\text{Desvio padrão do Grupo 2} = \sqrt{\text{VAR}} = \sqrt{1,32} = 1,15$$

$$\text{Desvio padrão do Grupo 3} = \sqrt{\text{VAR}} = \sqrt{2,36} = 1,54$$

3.3 Medidas de Centralidade e Dispersão Usando R

Imagino que agora você já imagine o que vem em seguida. Isso mesmo, veja a videoaula de como aplicar os conceitos de média e mediana usando a linguagem R.



O *script* usado na videoaula é o que está logo abaixo:

```
#podemos criar conjuntos de valores consecutivos usando a notação
valor_inicial:valor_final.
#no exemplo abaixo criamos um conjunto de 100 números variando de
1 a 100
1:100
```

```

#Quando trabalhamos com estatística não é raro que precisemos gerar
valores aleatórios.
#Para que valores aleatórios sejam repetidos em execuções consecutivas
utilizamos a função set.seed
#A função set.seed pede que o analista informe um número inteiro
qualquer para marcar qual é a semente da geração do valor aleatório
set.seed(1972)

#A função sample gera números aleatórios. São necessários pelo
menos os seguintes parâmetros para se gerar esse números
#x: um conjunto de valores que os números podem assumir
#size: o tamanho do conjunto de números aleatórios que desejamos
criar

#gera um conjunto com 50 números aleatórios entre 1 e 100
valores_aleatorios<- sample(x=1:100, size = 50)

valores_aleatorios

#A função sum soma todos os valores de um conjunto. Na linha abaixo,
somamos so valores entre 1 e 3
sum(1:3)

#A função NROW indica o número de linhas de um conjunto ou o número
de linhas de um datafrme.
NROW(1:3)

#Podemos usar as funções sum e NROW para calcular a média dos
números no objeto valores_aleatorios
sum(valores_aleatorios)/NROW(valores_aleatorios)

#Mas o R disponibiliza a função mean que faz esse cálculo de forma
mais prática.
#Basta informar no parâmetro x qual o conjunto que se quer calcular
a média
mean(x= valores_aleatorios)

#Já para a mediana utiliza-se a função median
median(x= valores_aleatorios)

#As medidas de dispersão também são simples no R

```

```
#O valor mínimo é dado pela função min  
min(valores_aleatorios)
```

```
#O valor máximo é dado pela função max  
max(valores_aleatorios)
```

```
#E a amplitude você calcula pela valor máximo menos o valor mínimo  
max(valores_aleatorios) - min(valores_aleatorios)
```

```
#Já o desvio padrão é dado pela função sd  
sd(valores_aleatorios)
```

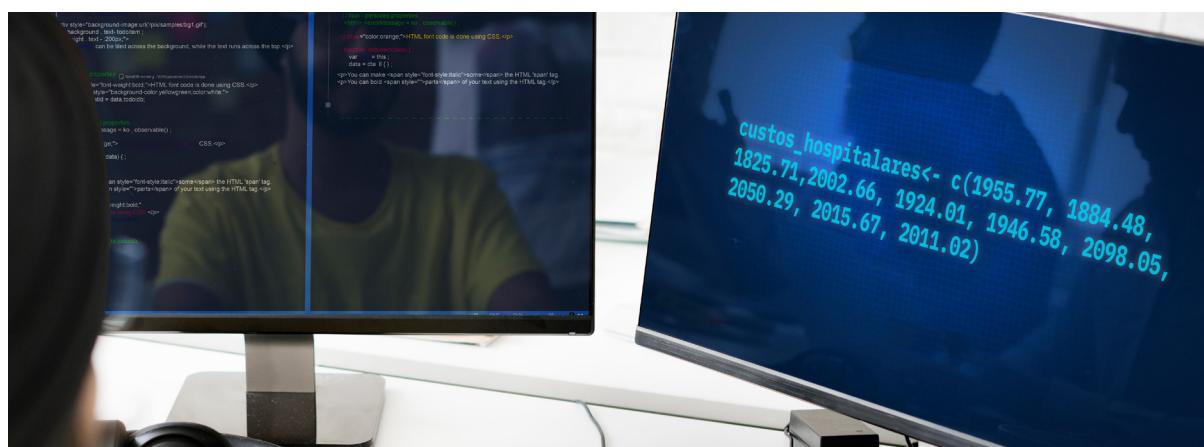
Script usado para média e mediana.

Fonte: Barbalho (2020).

Como sempre, copie e cole o código acima num novo arquivo de script R. Execute linha a linha procurando entender o que você está fazendo. Use os comentários como referência e se ainda houver dúvidas reveja a videoaula.

Agora é com você! Pratique um pouco as funções do R para medidas de centralidade e dispersão. Imagine que você tem em mãos os dados de custos hospitalares de um determinado hospital da rede pública. O seu chefe pede que você indique: média, mediana e desvio padrão dos custos. Considere que os custos sejam os indicados abaixo:

```
custos_hospitalares<- c(1955.77, 1884.48, 1825.71, 2002.66, 1924.01,  
1946.58, 2098.05, 2050.29, 2015.67, 2011.02)
```



Dados hospitalares.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Feito?! Verifique se os valores de média, mediana e desvio padrão são respectivamente: 1971.424, 1979.215 e 80.73098.

Agora é a hora de você testar seus conhecimentos. Para isso, acesse o exercício avaliativo disponível no ambiente virtual. Bons estudos!

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN**. GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

IHAKA, Ross. **The R Project**: a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo**. [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R**. RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

Módulo

3 Modelando Dados e Gráficos

Esse módulo apresentará as principais práticas para modelagem de dados e construção de gráficos.

Na unidade 1 você aprenderá quais são as técnicas mais comuns de manipulação de dados e quais as opções que o R oferece para trabalhar essas técnicas a partir do uso do pacote {tidyverse}. Na unidade 2 o foco será nos gráficos!

Será apresentado a você os principais componentes de uma gramática de visualização de dados e em seguida você verá como usar o pacote {ggplot2} para a construção das imagens.

Já na unidade 3, a missão é fazer você resgatar na memória o conceito de logaritmo que aprendeu ainda no ensino médio (ou segundo grau para os mais velhos) e em seguida, verificar como é possível aproximar as populações de São Paulo e Borá. Na unidade 4, você vai fechar o módulo analisando simultaneamente duas variáveis. Respire aí, tome uma água e vamos aos dados!

Unidade 1: Manipulação de Dados

Objetivo de aprendizagem

Ao final desta unidade você será capaz de esclarecer as possibilidades de manipulação de dados usando o pacote {tidyverse}.

1.1 Filtros de Dados

Quando se está trabalhando com dados é muito comum encontrar situações em que não é preciso todo o conjunto de dados que se tem à disposição.

Nesses casos, o que mais se deseja é que os dados sejam manipulados de forma a gerar um subconjunto com apenas as informações necessárias para a análise. Esse tipo de seleção é conhecido como filtro, e pode ser feito tanto sobre as linhas disponíveis como também sobre as colunas. Observe a tabela a seguir.

Município	Pop	desp_alun	UF	Escola	Nota	Matrículas	SEC_EDU	SEG_PUB	SEC_SAUDE
Três Cachoeiras	10989	36743,27381	RS	EMEF FERNANDO FERRARI	5,553167	192	1	0	0
Mostardas	12760	22660,61764	RS	EMF DR DINARTE SILVEIRA MARTINS	4,963667	264	1	0	0
Mostardas	12760	22660,61764	RS	E M F NOSSA SENHORA APARECIDA	5,839833	284	1	0	0
Lagoa Bonita do Sul	2884	30597,80854	RS	ESCOLA MUNICIPAL DE EDUCACAO BASICA RAINHA DOS APOSTOLOS	6,066333	143	1	0	0
Pirapetinga	10731	21994,84054	MG	CENTRO EDUCACIONAL MUNICIPAL DE PIRAPETINGA	5,711333	202	1	0	0
Pirapetinga	10731	21994,84054	MG	EM CEL RIBEIRO DOS REIS	5,142667	563	1	0	0
Independência	6228	39107,21986	RS	EMEF PRESIDENTE GETULIO VARGAS	5,479667	503	1	0	0
Esperança do Sul	2969	37226,56322	RS	ESCOLA MUNICIPAL ENSINO FUNDAMENTAL ESPERANCA	5,103667	294	1	0	0
Alagoa	2683	19402,8741	MG	EM CORONEL PORFIRIO MENDES PINTO	5,6995	297	1	0	0
Butiá	20929	30110,91	RS	ESCOLA MUNICIPAL DE ENSINO FUNDAMENTAL PROFESSORA MARIA ALZIRA	5,205	426	1	0	0
Pontão	3908	33734,03	RS	EMEF ALBERTO TORRES	6,087333	268	1	0	0
Roca Sales	11300	28512,16	RS	ESC MUN ENS FUN PERPETUO SOCORRO	5,101167	201	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF ESCOLA DO HOJE	4,745	304	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF BENONI ROSADO	5,4185	278	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF DOM JOSE GOMES	5,800667	233	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF ANTONINO XAVIER	5,636	280	1	0	0
Passo Fundo	201767	27966,61	RS	EMEF URBANO RIBAS	5,412333	458	1	0	0

Notas IDEB com despesas por aluno matriculado.

Fonte: Brasil (2021).

A tabela é um extrato de um banco de dados real com 13880 linhas e 188 colunas. É isso mesmo! Você pode encontrar tabelas que atingem esses valores. E essas dimensões na verdade ainda são baixas. Para poder ilustrar os pontos sobre manipulação de dados, escolheu-se um recorte das 18 primeiras linhas e apenas 10 colunas que são: "município", "população do município", "gasto em educação por aluno no município", "UF", "Nome da escola", "média no IDEB", "número de alunos matriculados", "indicador de que a escola é da secretaria de educação", "indicador de que a escola é da secretaria de segurança pública" e "indicador de que a escola é da secretaria de saúde".

Conforme indicado acima, pode-se entender que a tabela já passou por dois tipos de filtro para ser exibida neste nosso texto. O primeiro foi um filtro de linhas, já que de 13880 linhas optou-se por exibir apenas 19. O segundo filtro foi o de colunas, de 188 colunas selecionou-se apenas 10.

É possível ainda implementar mais filtros. Veja a tabela a seguir.

Município	Pop	desp_alun	UF	Escola	Nota	Matrículas
Três Cachoeiras	10989	36743,27381	RS	EMEF FERNANDO FERRARI	5,553167	192
Mostardas	12760	22660,61764	RS	EMF DR DINARTE SILVEIRA MARTINS	4,963667	264
Mostardas	12760	22660,61764	RS	E M F NOSSA SENHORA APARECIDA	5,839833	284
Lagoa Bonita do Sul	2884	30597,80854	RS	ESCOLA MUNICIPAL DE EDUCACAO BASICA RAINHA DOS APOSTOLOS	6,066333	143
Pirapetinga	10731	21994,84054	MG	CENTRO EDUCACIONAL MUNICIPAL DE PIRAPETINGA	5,711333	202
Pirapetinga	10731	21994,84054	MG	EM CEL RIBEIRO DOS REIS	5,142667	563
Independência	6228	39107,21986	RS	EMEF PRESIDENTE GETULIO VARGAS	5,479667	503
Esperança do Sul	2969	37226,56322	RS	ESCOLA MUNICIPAL ENSINO FUNDAMENTAL ESPERANCA	5,103667	294
Alagoa	2683	19402,8741	MG	EM CORONEL PORFIRIO MENDES PINTO	5,6995	297
Butiá	20929	30110,91	RS	ESCOLA MUNICIPAL DE ENSINO FUNDAMENTAL PROFESSORA MARIA ALZIRA	5,205	426
Pontão	3908	33734,03	RS	EMEF ALBERTO TORRES	6,087333	268
Roca Sales	11300	28512,16	RS	ESC MUN ENS FUN PERPETUO SOCORRO	5,101167	201

Passo Fundo	201767	27966,61	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421
Passo Fundo	201767	27966,61	RS	EMEF ESCOLA DO HOJE	4,745	304
Passo Fundo	201767	27966,61	RS	EMEF BENONI ROSADO	5,4185	278
Passo Fundo	201767	27966,61	RS	EMEF DOM JOSE GOMES	5,800667	233
Passo Fundo	201767	27966,61	RS	EMEF ANTONINO XAVIER	5,636	280
Passo Fundo	201767	27966,61	RS	EMEF URBANO RIBAS	5,412333	458

Filtro de coluna.

Fonte: Brasil (2021).

Veja que agora trabalha-se com apenas sete colunas. Eventualmente podemos achar que a informação sobre a qual secretaria a escola está vinculada não importa para uma análise específica sobre esses dados. Observe ainda que esse pequeno extrato apresenta dados sobre 12 municípios diferentes. Podemos estar interessados apenas nas escolas do município de Passo Fundo, que no caso da tabela acima são as seis últimas. Nesse caso, a tabela filtrada ficaria assim:

Município	Pop	desp_alun	UF	Escola	Nota	Matrículas
Passo Fundo	201767	27966,61	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421
Passo Fundo	201767	27966,61	RS	EMEF ESCOLA DO HOJE	4,745	304
Passo Fundo	201767	27966,61	RS	EMEF BENONI ROSADO	5,4185	278
Passo Fundo	201767	27966,61	RS	EMEF DOM JOSE GOMES	5,800667	233
Passo Fundo	201767	27966,61	RS	EMEF ANTONINO XAVIER	5,636	280
Passo Fundo	201767	27966,61	RS	EMEF URBANO RIBAS	5,412333	458

Filtros de linha (município).

Fonte: Brasil (2021).

Você observou que todas as escolas de todos os outros municípios não estão mais presentes na tabela acima? Só restaram mesmo as escolas de Passo Fundo. Nesse caso fica muito claro que foi feito um filtro de linha. Inicialmente você via 18 linhas e agora vê apenas 6.

É possível fazer filtros de linha tendo a quantidade como referência. Se voltarmos à tabela com as 18 linhas veja que o número de alunos matriculados varia entre 143 e 563. Pode ser que agora você esteja interessado em todas as escolas que possuem mais de 300 alunos, independentemente de qual município. Veja como fica.

Município	Pop	desp_alun	UF	Escola	Nota	Matrículas
Pirapetinga	10731	21994,84054	MG	EM CEL RIBEIRO DOS REIS	5,142667	563
Independência	6228	39107,21986	RS	EMEF PRESIDENTE GETULIO VARGAS	5,479667	503
Butiá	20929	30110,91	RS	ESCOLA MUNICIPAL DE ENSINO FUNDAMENTAL PROFESSORA MARIA ALZIRA	5,205	426
Pontão	3908	33734,03	RS	EMEF ALBERTO TORRES	6,087333	268
Passo Fundo	201767	27966,61	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421
Passo Fundo	201767	27966,61	RS	EMEF ESCOLA DO HOJE	4,745	304
Passo Fundo	201767	27966,61	RS	EMEF URBANO RIBAS	5,412333	458

Filtros de linha (quantidade).

Fonte: Brasil (2021).

Observe que da tabela original com 18 linhas existem agora apenas 6, já que apenas 6 escolas possuem mais de 300 alunos.

E se quiser combinar filtros? Você acha que isso é possível? Certamente que sim. Você pode, por exemplo, recuperar o filtro das escolas de Passo Fundo com as escolas com mais de 300 alunos. Veja o resultado.

Município	Pop	desp_alun	UF	Escola	Nota	Matrículas
Passo Fundo	201767	27966,61	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421
Passo Fundo	201767	27966,61	RS	EMEF ESCOLA DO HOJE	4,745	304
Passo Fundo	201767	27966,61	RS	EMEF URBANO RIBAS	5,412333	458

Combinar filtros.

Fonte: Brasil (2021).

Veja só, agora quando foram combinados os dois filtros de linha, apenas 3 escolas atendem à dupla condição de estarem em Passo Fundo e terem mais de 300 matrícululas. De 18 linhas iniciais, o filtro nos devolveu 3 linhas.

1.2 Construindo Agrupamentos

Muitas vezes existe a necessidade de resumir os dados de acordo com as classes de variáveis categóricas e nesse resumo se procura extrair algumas estatísticas descritivas tais como, quantidade de municípios por estado ou quantidade de escola por municípios. Podemos ainda investigar a relação de variáveis categóricas com variáveis quantitativas, como por exemplo a média de notas por estado e a soma de matrículas por municípios.

Veja alguns exemplos.

Estado	Contagem de Escola
MG	3
RS	15

Na tabela acima, veja o agrupamento da tabela original feito pela variável categórica “estado”, em que interessa contar as escolas presentes nas 18 linhas originalmente identificadas. Vê-se então que na tabela original apareceram apenas dois estados: Minas Gerais (MG) e Rio Grande do Sul (RS), sendo 3 escolas de MG e 15 do RS.

Um outro exemplo:

Município	Contagem de Escolas
Alagoa	1
Butiá	1
Esperança do Sul	1
Independência	1
Lagoa Bonita do Sul	1
Mostardas	2
Passo Fundo	6
Pirapetinga	2
Pontão	1
Roca Sales	1
Três Cachoeiras	1
Total Geral	18

Agora agrupamos as linhas por municípios e contamos o número de escolas para cada um dos municípios.

Estado	Soma de matrículas
MG	1062
RS	4549

Na tabela acima, foram agrupadas as 18 linhas originais por estado, ficando, portanto, com apenas duas linhas que correspondem aos estados que aparecem na tabela original conforme mostrado anteriormente. Dessa vez utilizou-se o agrupamento para somar o número de matrículas. E aí chegamos à conclusão de que entre as 18 linhas que estamos trabalhando nesta tabela, MG possui 1062 alunos matriculados e o RS 4549 alunos.

Outra possibilidade de cálculo com essas mesmas variáveis: “estado” e “matrícula” é a média de alunos matriculados por escola.

Estado	Média de matrículas
MG	354,0
RS	303,3

Observe que MG possui uma média de 354 alunos matriculados por escola enquanto cada escola do RS fica com 303,3 alunos matriculados em média.

Pode-se fazer um agrupamento que busque a média das notas do IDEB por município.

Rótulos de Linha	Média de Nota
Alagoa	5,70
Butiá	5,21
Esperança do Sul	5,10
Independência	5,48
Lagoa Bonita do Sul	6,07
Mostardas	5,40
Passo Fundo	5,49
Pirapetinga	5,43
Pontão	6,09
Roca Sales	5,10
Três Cachoeiras	5,55

Para melhorar a visualização e conclusão sobre os dados, vale lembrar que é possível ordenar as variáveis quantitativas em ordem crescente ou decrescente.

Rótulos de Linha	Média de Nota
Pontão	6,09
Lagoa Bonita do Sul	6,07
Alagoa	5,70
Três Cachoeiras	5,55
Passo Fundo	5,49
Independência	5,48
Pirapetinga	5,43
Mostardas	5,40
Butiá	5,21
Esperança do Sul	5,10
Roca Sales	5,10

Pronto, agora fica melhor a nossa análise de dados. Veja que as médias das notas no IDEB dos municípios nessa tabela que está analisando varia entre 5,10 e 6,09. O município de maior média de nota é Pontão e o de menor média é Roca Sales.

1.3 Alterando a Estrutura de Dados

Para muitas análises é importante que se maneje os dados de forma a acrescentar colunas que sejam resultados de valores de outras colunas. Suponhamos que seja interessante entender o percentual da população total do município atendida em cada escola. Para isso é necessário verificar a relação entre o número de matriculados para cada escola e o total da população do município.

Município	Pop	UF	Escola	Nota	Matrículas	% população
Alagoa	2683	MG	EM CORONEL PORFIRIO MENDES PINTO	5,6995	297	11,07
Butiá	20929	RS	ESCOLA MUNICIPAL DE ENSINO FUNDAMENTAL PROFESSORA MARIA ALZIRA	5,205	426	2,04
Esperança do Sul	2969	RS	ESCOLA MUNICIPAL ENSINO FUNDAMENTAL ESPERANCA	5,103667	294	9,9
Independência	6228	RS	EMEF PRESIDENTE GETULIO VARGAS	5,479667	503	8,08
Lagoa Bonita do Sul	2884	RS	ESCOLA MUNICIPAL DE EDUCACAO BASICA RAINHA DOS APOSTOLOS	6,066333	143	4,96
Mostardas	12760	RS	EMF DR DINARTE SILVEIRA MARTINS	4,963667	264	2,07
Mostardas	12760	RS	E M F NOSSA SENHORA APARECIDA	5,839833	284	2,23
Passo Fundo	201767	RS	EMEF ELOY PINHEIRO MACHADO	5,946	421	0,21
Passo Fundo	201767	RS	EMEF ESCOLA DO HOJE	4,745	304	0,15
Passo Fundo	201767	RS	EMEF BENONI ROSADO	5,4185	278	0,14
Passo Fundo	201767	RS	EMEF DOM JOSE GOMES	5,800667	233	0,12
Passo Fundo	201767	RS	EMEF ANTONINO XAVIER	5,636	280	0,14
Passo Fundo	201767	RS	EMEF URBANO RIBAS	5,412333	458	0,23
Pirapetinga	10731	MG	CENTRO EDUCACIONAL MUNICIPAL DE PIRAPETINGA	5,711333	202	1,88
Pirapetinga	10731	MG	EM CEL RIBEIRO DOS REIS	5,142667	563	5,25
Pontão	3908	RS	EMEF ALBERTO TORRES	6,087333	268	6,86
Roca Sales	11300	RS	ESC MUN ENS FUN PERPETUO SOCORRO	5,101167	201	1,78
Três Cachoeiras	10989	RS	EMEF FERNANDO FERRARI	5,553167	192	1,75

Percentual da população total do município atendido em cada escola.

Fonte: Brasil (2021).

Acrescenta-se a coluna “% População” cujo valor é derivado das colunas “Número de matrículas” e “População”. Com essa operação é possível saber que em Alagoa a escola **EM CORONEL PORFIRIO MENDES PINTO** atende a 11,07% da população do município!!!

Pensou em modificar uma coluna já preexistente? Então veja! Digamos que para efeito de comparações internacionais as notas precisem ser multiplicadas por 100. Não há problema! Ferramentas como o R fazem isso com muita tranquilidade. O resultado de uma alteração como essa seria o da tabela a seguir.

Município	Pop	UF	Escola	Nota	Matrículas	% população
Alagoa	2683	MG	EM CORONEL PORFIRIO MENDES PINTO	569,95	297	11,07
Butiá	20929	RS	ESCOLA MUNICIPAL DE ENSINO FUNDAMENTAL PROFESSORA MARIA ALZIRA	520,5	426	2,04
Esperança do Sul	2969	RS	ESCOLA MUNICIPAL ENSINO FUNDAMENTAL ESPERANCA	510,3667	294	9,9
Independência	6228	RS	EMEF PRESIDENTE GETULIO VARGAS	547,9667	503	8,08
Lagoa Bonita do Sul	2884	RS	ESCOLA MUNICIPAL DE EDUCACAO BASICA RAINHA DOS APOSTOLOS	606,6333	143	4,96
Mostardas	12760	RS	EMF DR DINARTE SILVEIRA MARTINS	496,3667	264	2,07
Mostardas	12760	RS	E M F NOSSA SENHORA APARECIDA	583,9833	284	2,23
Passo Fundo	201767	RS	EMEF ELOY PINHEIRO MACHADO	594,6	421	0,21
Passo Fundo	201767	RS	EMEF ESCOLA DO HOJE	474,5	304	0,15
Passo Fundo	201767	RS	EMEF BENONI ROSADO	541,85	278	0,14
Passo Fundo	201767	RS	EMEF DOM JOSE GOMES	580,0667	233	0,12
Passo Fundo	201767	RS	EMEF ANTONINO XAVIER	563,6	280	0,14
Passo Fundo	201767	RS	EMEF URBANO RIBAS	541,2333	458	0,23
Pirapetinga	10731	MG	CENTRO EDUCACIONAL MUNICIPAL DE PIRAPETINGA	571,1333	202	1,88
Pirapetinga	10731	MG	EM CEL RIBEIRO DOS REIS	514,2667	563	5,25
Pontão	3908	RS	EMEF ALBERTO TORRES	608,7333	268	6,86
Roca Sales	11300	RS	ESC MUN ENS FUN PERPETUO SOCORRO	510,1167	201	1,78
Três Cachoeiras	10989	RS	EMEF FERNANDO FERRARI	555,3167	192	1,75

Modificando uma coluna preexistente.

Fonte: Brasil (2021).

Observe que todas as notas foram multiplicadas por 100, o que já poderia permitir a comparação internacional.

Tudo o que você viu sobre filtro, agrupamento, ordenação e mudança de estrutura é bastante fácil de ser implementado utilizando a linguagem R, principalmente com o apoio do pacote {tidyverse}. O próximo tema vai mostrar como aplicar esses conceitos com práticas de codificação.

1.4 Manipulando Dados com Pacote Tidyverse

O pacote {tidyverse} é na verdade uma coleção de pacotes que se dedicam a facilitar a vida do analista de dados principalmente nas tarefas relacionadas à manipulação e visualização de dados.

Veja a videoaula a seguir e saiba como usar o {tidyverse} para importar arquivos.



Videoaula: [Uso do Tidyverse para Importar Arquivos](#)

Agora assista a videoaula abaixo para aprender a fazer filtros.



Videoaula: [Uso do Tidyverse para Fazer Filtros](#)

Por fim, assista à próxima videoaula e aprenda a usar o {tidyverse} para importar, fazer agrupamentos, ordenação e mudança de estrutura de dados.



Videoaula: [Uso do Tidyverse \(Agrupamentos, Ordenação e Mudança de Estrutura de Dados\)](#)

Agora que você já assistiu aos vídeos, tome contato com o script que reúne todos os códigos apresentados nessa unidade. Faça o já tradicional copiar-colar para o seu RStudio. Execute tudo linha a linha e se tiver dúvidas, volte a ver as videoaulas!

```

#Você deve instalar o pacote tidyverse para executar o script
#Lembre-se que a instalação precisa ser feita apenas uma vez.
#Se você for executar esse script mais de uma vez, insira um # antes
da linha logo abaixo a partir da segunda execução.
install.packages("tidyverse")

#Carrega o pacote tidyverse para a memória
library(tidyverse)

#O objeto arquivo abaixo indica onde está na internet a tabela com
os dados sobre municípios brasileiros
arquivo<-      "https://raw.githubusercontent.com/fernandobarbalho/
enap_auto_instucional/main/data/dados_municipios.csv"

#A função read_csv permite a leitura de um arquivo e a transformação
do arquivo em um data.frame
#É necessário pelo menos a indicação do seguinte parâmetro para se
usar a função read_scv:
#file: caminho completo para se chegar a um arquivo. O caminho deve
terminar com o nome do arquivo
#Deve-se observar que file pode conter um caminho para um endereço
na internet ou para uma pasta

#gera o data.frame dados_municipios a partir do que está presente
no endereço representado pelo objeto arquivo
dados_municipios<- read_csv(file= arquivo)

#Mostra o conteúdo de dados municipios
dados_municipios

glimpse(dados_municipios)

#A função filter permite fazer filtros sobre um dataframe.
#Deve-se informar pelo menos os seguintes parâmetros:
#.data: data frame com os dados
#expressão lógica: expressão que indica o filtro que será feito

#Na linha abaixo, vamos filtrar o data frame dados_municipios
permanecendo apenas os municípios do estado de Pernambuco
#Nesse caso, a expressão lógica é sigla_uf=="PE"

filter(.data= dados_municipios, sigla_uf=="PE" )

#Se quisermos os municípios do Ceará

```

```

filter(.data= dados_municpios, sigla_uf=="CE" )

#O pacote tidyverse permite o uso do operador %>%, conhecido como
pipe (cano, em inglês), que facilita a construção de uma sequência
de operações
#O operador %>% leva o conteúdo de um dataframe para o primeiro
parâmetro de uma função que vem logo após o %>%
#Isso elimina a necessidade de informar o valor do primeiro parâmetro
da função

#No exemplo abaixo, o %>% leva o conteúdo de dados_municpios para
o primeiro argumento da função filter, que é .data

dados_municipios %>%
  filter(sigla_uf=="CE")

#Isso permite fazer uma sequência de comandos sobre um dado original.
#No exemplo abaixo, aplicamos um novo filtro ao filtro original
#Agora além de filtrar por estado, vamos filtrar também por população,
permanecendo apenas os municípios com mais
#de 100000 habitantes

dados_municipios %>%
  filter(sigla_uf=="CE") %>%
  filter(populacao > 100000)

#Já vimos que a tabela contém 26 colunas para saber o nome das
colunas usamos a função name
#Passamos como parâmetro para a função name o nome do dataframe

names(dados_municipios)

#Vamos imaginar que agora queremos trabalhar apenas com as colunas
sigla_uf, nome e populacao.
#Para isso acrescentamos à sequência do comandos concatenadas por
%>% a função select. Veja abaixo

dados_municipios%>%
  filter(sigla_uf == "CE") %>%
  filter(populacao > 100000) %>%
  select(sigla_uf,nome, populacao) #indica as colunas que devem
aparecer no resultado da sequência de comandos

#Talvez seja interessante ordenar esses municípios. Para isso
acrescentamos à sequência de comandos a função arrange

```

#Na função arrange devemos indicar qual a coluna que será feita a ordenação

```
dados_municpios %>%
  filter(sigla_uf == "CE") %>%
  filter(populacao > 100000) %>%
  select(sigla_uf, nome, populacao) %>% #indica as colunas que devem aparecer no resultado da sequência de comandos
  arrange(populacao)
```

#Para que a ordenação seja feita em ordem decrescente use a função desc. Veja abaixo

```
dados_municpios %>%
  filter(sigla_uf == "CE") %>%
  filter(populacao > 100000) %>%
  select(sigla_uf, nome, populacao) %>% #indica as colunas que devem aparecer no resultado da sequência de comandos
  arrange(desc(populacao)) #ordenação em ordem decrescente de população
```

#Para fazer agrupamentos utilizando o tidyverse utilizamos duas funções: group_by e summarise

#Na função group_by informa-se as variáveis que queremos agrupar
#Na função summarise informa-se operações que se deseja fazer com as variáveis agrupadas

#No exemplo abaixo, para cada sigla_uf, somamos a população de todos os municípios

```
dados_municipios %>%
  group_by(sigla_uf) %>% #Agrupa pela variável sigla_uf
  summarise(
    populacao_estado = sum(populacao) #soma a populacao de todos os municípios
  )
```

#Agora o resultado da sequência de comandos anterior em ordem decrescente de população do estado

```
dados_municipios %>%
  group_by(sigla_uf) %>% #Agrupa pela variável sigla_uf
  summarise(
    populacao_estado = sum(populacao) #soma a populacao de todos os municípios
```

```

) %>%
arrange(desc(populacao_estado)) #faz a ordenação descrescente por
populacao_estado

#Podemos também contar o número de municípios em cada estado usando
a função n

dados_municipios %>%
  group_by(sigla_uf) %>% #Agrupa pela variável sigla_uf
  summarise(
    quantidade_municipios = n() #soma a populacao de todos os
municípios
  ) %>%
  arrange(desc(quantidade_municipios)) #faz a ordenação descrescente
por populacao_estado

#E combinar num mesmo resultado as informações de população do
estado, quantidade de municípios, média de habitantes por município
e mediana de habitantes

dados_municipios %>%
  group_by(sigla_uf) %>% #Agrupa pela variável sigla_uf
  summarise(
    quantidade_municipios = n(),
    populacao_estado = sum(populacao),
    media_populacao = mean(populacao),
    mediana_populacao = median(populacao)
  ) %>%
  arrange(desc(quantidade_municipios)) #faz a ordenação descrescente
por populacao_estado

#Podemos alterar a estrutura de uma tabela usando a função mutate
#No caso abaixo vamos criar uma nova coluna a partir da combinação
das variáveis nome, sigla_uf e nome_regiao_saude
#vamos usar a função paste para fazer a concatenação de variáveis
tipo texto. Vamos separar as variáveis com "-"

dados_municipios %>%
  mutate(nome_regiao_saude_uf = paste(nome, nome_regiao_saude,
sigla_uf, sep = "-")) %>%
  select(nome_regiao_saude_uf, populacao)

```

Script dos códigos da unidade.

Fonte: Barbalho (2020).

Que tal um pouco mais de prática? Procure resolver o que segue.

Como você viu, usou-se:

```
filter(.data= dados_municípios, sigla_uf=="CE" )
```

Para filtrar o *dataframe* “dados_municípios” para permanecer apenas as linhas que se refiram ao Ceará, experimente substituir o == por != e veja qual o efeito sobre o resultado do filtro.

Execute uma série de comandos que gere uma tabela com valores agrupados pela coluna “nome_regiao”, mostrando o número de municípios e a população total para cada região. Tudo isso em ordem decrescente de população total. O resultado final deverá ser parecido com o que está logo abaixo.

	nome_regiao	quantidade_municípios	populacao_regiao
	<chr>	<int>	<dbl>
1	Sudeste	1668	89012240
2	Nordeste	1794	57374243
3	Sul	1191	30192315
4	Norte	450	18672591
5	Centro-Oeste	467	16504303

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica: o caso de política pública sobre dados abertos.** 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN.** GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

IHAKA, Ross. **The R Project:** a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo.** [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R.** RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdadosus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

Unidade 2: Criando Gráficos Estatísticos com o Pacote Ggplot2

Objetivo de aprendizagem

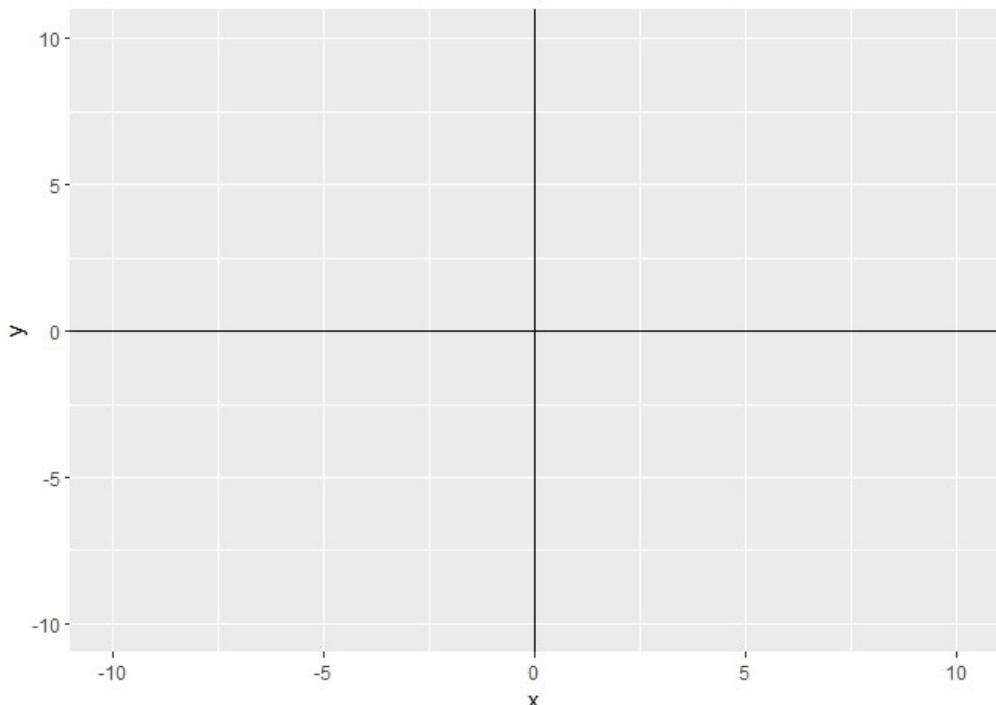
Nesta unidade você vai identificar como criar gráficos estatísticos utilizando as funcionalidades do pacote `{ggplot2}`.

2.1 Biblioteca Ggplot2: uma Gramática para Desenho de Gráficos

A biblioteca `{ggplot2}` trabalha com a ideia de que existe uma gramática que ajuda a identificar os principais elementos que se relacionam com a construção de gráficos. Nesse tópico você vai explorar alguns dos principais elementos dessa gramática.

A primeira informação que precisa ter em mente é que, de um modo geral, os gráficos se organizam num plano cartesiano definido por dois eixos: horizontal e vertical.

Normalmente o eixo horizontal é denominado de eixo **x** e o vertical de eixo **y**. A figura que você vê é um exemplo desse plano com os dois eixos.

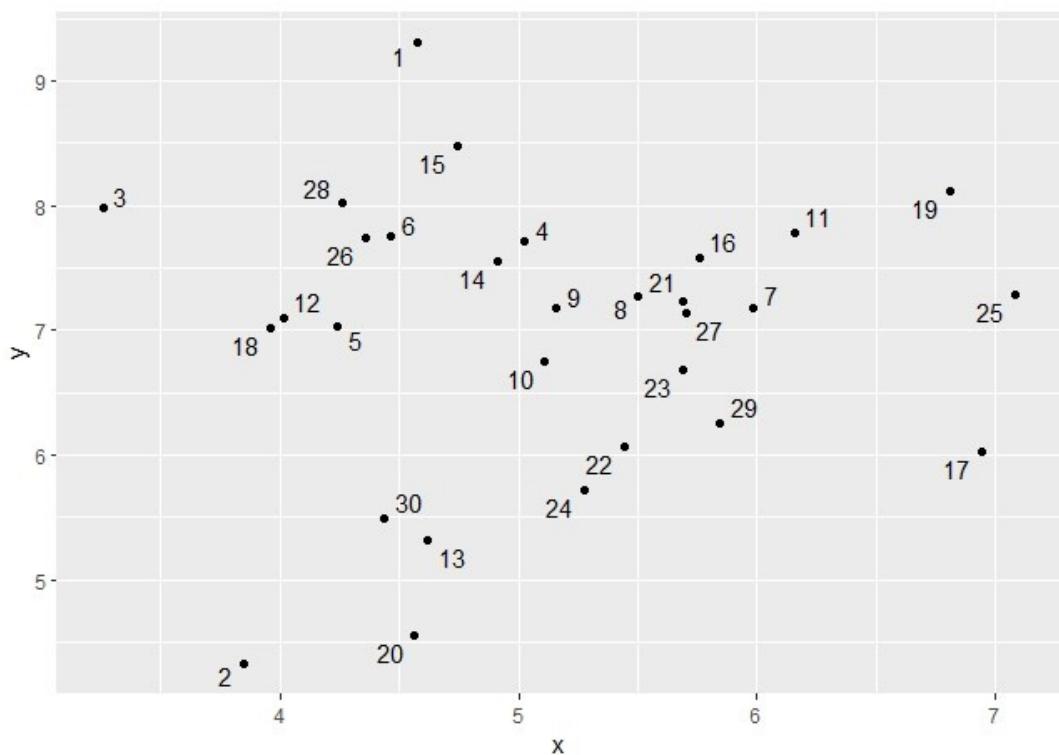


Eixos horizontal e vertical.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Ainda nessa figura, vê-se que o eixo x se estende na horizontal e assume valores que variam entre -10 e 10. O eixo vertical y também varia entre -10 e 10. A indicação do conjunto de valores possíveis para os eixos x e y faz parte da definição da **estética** de um gráfico de acordo com a gramática do {ggplot}.

O próximo elemento importante para a gramática do {ggplot} é a forma geométrica de um gráfico. Aqui é que se define qual vai ser a geometria de um gráfico, se de linhas, pontos, barras, entre outros desenhos. Veja algumas dessas possibilidades.



Forma geométrica de um gráfico.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

O gráfico que você viu está associado à geometria ponto. A ideia é exibir cada ponto da nossa base de dados no eixo cartesiano.

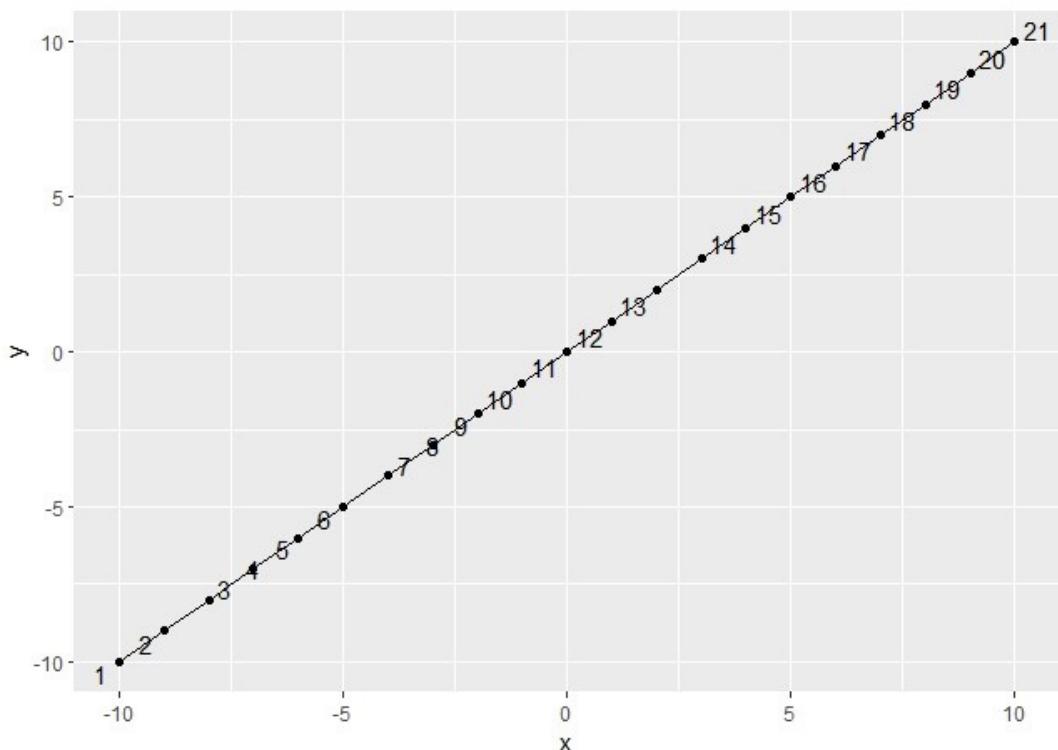
Veja a seguir os dados que foram utilizados para gerar o gráfico.

ponto	x	y
1	4,577683	9,302174
2	3,844798	4,327104
3	3,257079	7,976179
4	5,026635	7,717275
5	4,240126	7,024384
6	4,465826	7,749615
7	5,980487	7,183005
8	5,502892	7,271822
9	5,156656	7,176476
10	5,110177	6,752545
11	6,154802	7,775746
12	4,012977	7,097288
13	4,619916	5,320437
14	4,909219	7,549327
15	4,742403	8,479989
16	5,756593	7,580235
17	6,939885	6,028956
18	3,959698	7,010619
19	6,807548	8,109871
20	4,563802	4,556953
21	5,689434	7,234765
22	5,443821	5,443821
23	5,687899	6,686312
24	5,27498	5,720128
25	7,081019	7,279701
26	4,357921	7,745802
27	5,701247	7,140011
28	4,261474	8,014817
29	5,845618	6,247937
30	4,434005	5,483695

Dados usados para gerar o gráfico.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Experimente comparar os números que identificam os pontos que aparecem no gráfico com os que estão na tabela acima. Mais adiante você verá que esse tipo de gráfico é muito importante para verificar possíveis associações entre duas variáveis. Seguindo para um próximo tipo de geometria, observe a figura:



Geometria (ponto/linha).

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Na figura, além da geometria **ponto**, exibe-se também a geometria **linha**. Uma linha procura ligar pontos consecutivos de um conjunto de dados.

Veja a tabela abaixo e compare com o gráfico novamente. Você percebe que a posição do ponto 1 se liga ao ponto 2? Que por sua vez se liga ao ponto 3 e assim sucessivamente.

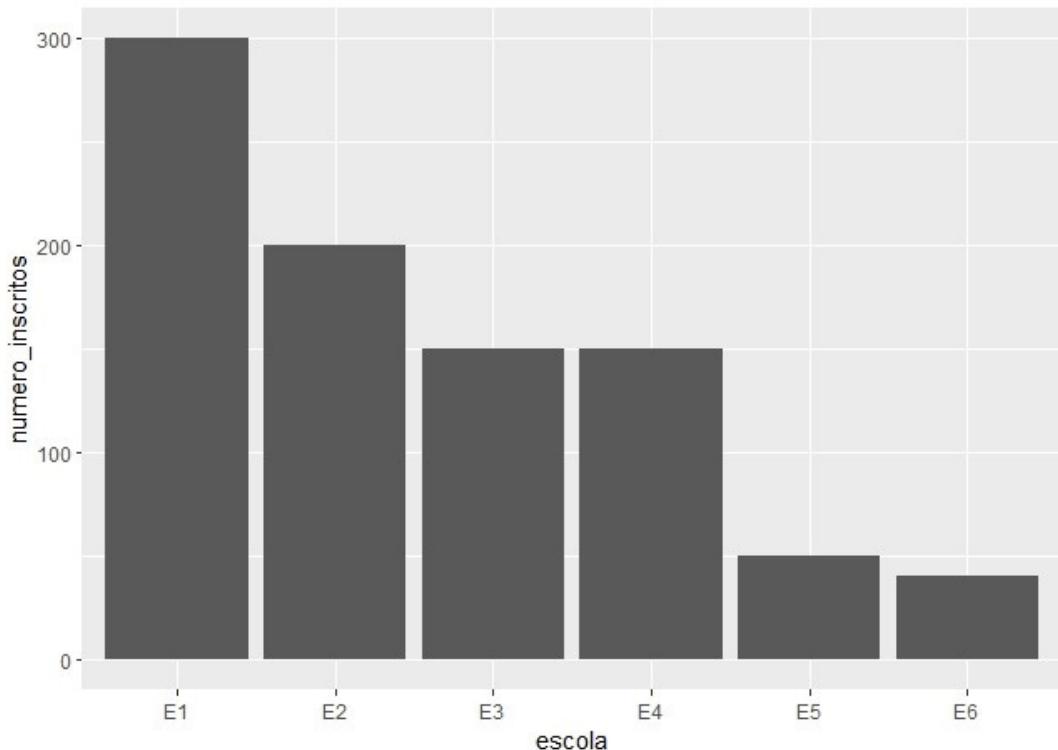
ponto	x	y
1	-10	-10
2	-9	-9
3	-8	-8
4	-7	-7
5	-6	-6
6	-5	-5
7	-4	-4
8	-3	-3
9	-2	-2
10	-1	-1
11	0	0
12	1	1
13	2	2
14	3	3
15	4	4
16	5	5
17	6	6
18	7	7
19	8	8
20	9	9
21	10	10

Comparação entre gráfico e tabela.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Esse tipo de comportamento do gráfico de linha associado à ligação de pontos é bem interessante para analisar como valores de variáveis se modificam ao longo do tempo. É também útil para gerar retas que simulem a relação entre duas variáveis distintas.

Veja mais uma geometria, dessa vez os gráficos de colunas.



Gráficos de colunas.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Os gráficos de colunas são úteis para comparar valores numéricos entre classes de uma variável categórica. No exemplo acima você pode presumir que estamos tratando de inscrições para o exame ENEM. Colocamos no eixo horizontal **x**, seis escolas para serem comparadas. Já no eixo vertical **y** fica a informação do número de inscritos por escola. Observe que fica muito fácil perceber que a escola **E1** se distancia das demais, que as escolas **E3** e **E4** possuem o mesmo número de inscritos e que as escolas **E5** e **E6** têm número bem menor de inscritos que as demais escolas.

Agora, compare o gráfico com os dados da tabela.

escola	numero_inscritos
E1	300
E2	200
E3	150
E4	150
E5	50
E6	40

Comparando gráfico com os dados da tabela.

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Existem vários outros tipos de geometria na gramática do {ggplot2}. Alguns desses tipos serão trabalhados de forma mais aprofundada na próxima unidade, que explorará histogramas e *box-plots*.

Voltando à estética na gramática do {ggplot2}. Dessa vez a ideia é usar recursos de cores para destacar informações. Veja a figura.

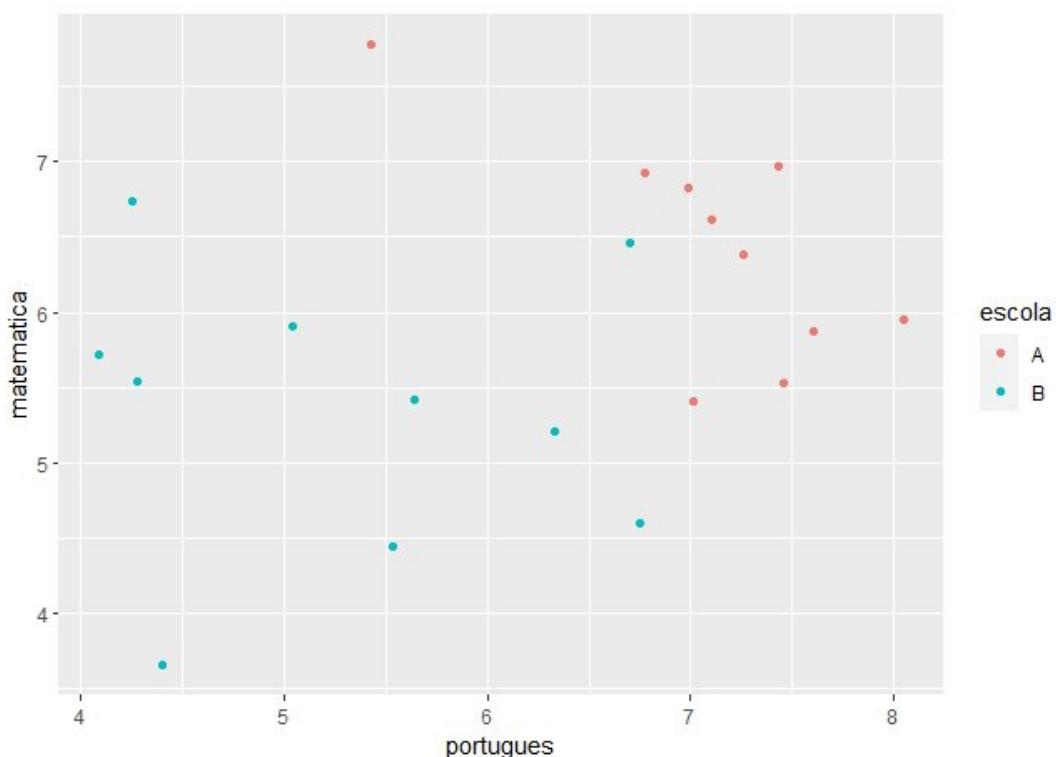


Gráfico com cores. Fonte: Barbalho (2020).

Fonte: Barbalho (2020). Elaboração: CEPED/UFSC (2023).

Essa figura descreve uma comparação de desempenho entre duas escolas hipotéticas A e B. Os alunos são comparados pelas notas nas disciplinas de **matemática** e de **português**. Observe que as notas de português ficam no eixo horizontal **x** e as de matemática no eixo vertical **y**. As cores diferenciam as notas obtidas pelos alunos das duas escolas. Neste caso a escola **A** está com os pontos pintados em vermelho e a escola **B** está com os pontos pintados em azul. Com esse código de cores fica fácil perceber pelo gráfico que os alunos da escola **A** obtêm quase sempre notas melhores nas duas disciplinas com exceção de um aluno que obteve a melhor nota em matemática mas não conseguiu um bom desempenho em português, ficando com notas abaixo das que foram alcançadas por grande parte dos alunos da escola **B**.

Esse é um exemplo de um dos usos de cores para destacar informações em um gráfico. São possíveis diversos outros, certo?!

Veja bem! O que você viu nesse tópico é apenas uma pequena introdução da gramática de gráficos. No tópico seguinte você vai mergulhar nos códigos para aplicar o que aprendeu até aqui. Lá, serão dadas novas indicações para aprofundar o aprendizado em gráficos. Fique firme!

2.2 Como utilizar o Pacote Ggplot2

No tópico anterior foram apresentados alguns dos principais conteúdos iniciais da gramática do {ggplot2}. Agora chegou o momento de ir ao RStudio e praticar com o R o desenvolvimento de gráficos comumente usados em análise de dados. Veja a videoaula sobre o tema.



Videoaula: [Gráfico Colunas](#)

Agora a videoaula a seguir servirá para aprender a fazer gráfico de pontos e linhas. Acompanhe!



Videoaula: [Gráfico de Pontos e Linhas](#)

Os códigos usados nas videoaulas são esses que você verá a seguir.

Agora copie o código para o seu ambiente de RStudio, depois procure executar e entender o que foi feito. Mais do que isso, procure analisar o que os gráficos estão mostrando pra você. Que conclusões analíticas você consegue tirar ao olhar esses gráficos?

```
library(tidyverse)

arquivo<-      "https://raw.githubusercontent.com/fernandobarbalho/
enap_auto_instucional/main/data/dados_municipios.csv"

dados_municipios<- read_csv(file= arquivo)

#ggplot é uma biblioteca do pacote tidyverse que permite trabalhar
com gráficos
#Para se trabalhar com gráfico usando ggplot2 usa-se uma combinação
de funções
```

```

#uma obrigatória é a ggplot. É ela quem indica ao R que o que vem
em seguir é um conjunto de instruções sobre gráficos
#é obrigatório indicar qual a geometria que vai ser usada
#para se trabalhar com colunas ou barras, usa-se a função geom_col
#a função geom_col exige que se informe algumas informações estéticas
a partir da função aes
#dois elementos fundamentais de estética são as variáveis que ficarão
nos eixos x e y
#A tabela de referência para uso do ggplot pode ser informado a
partir de uma sequência de comandos concatenados por %>%
#As instruções de gráficos são separadas pelo operador +

```

#Na sequência de instruções abaixo geramos um gráfico de colunas com as populações dos municípios pernambucanos com mais de 100 mil habitantes

```

# no eixo x (horizontal) aparecem as populações e no eixo y (vertical)
o nome das cidades
dados_municpios %>%
  filter(sigla_uf == "PE") %>%
  filter(populacao>100000) %>%
  ggplot()+
  geom_col(aes(x=populacao, y=nome))

```

#A função slice_max retorna as n linhas que possuem os maiores valores para uma dada variável

#Veja como podemos usar a função slice_max para mostrar os 10 municípios mais populosos do Brasil

```

dados_municipios %>%
  slice_max(populacao, n=10) %>%
  ggplot()+
  geom_col(aes(x=populacao, y=nome))

```

#Observe que o gráfico mostra as cidades em uma ordem aleatória. É conveniente deixarmos o gráfico em forma de ranking

#Usando a função mutate reordenamos a variável nome para que ela passe a aparecer no gráfico a partir da variável população

```

dados_municipios %>%
  slice_max(populacao, n=10) %>%
  mutate(nome=reorder(nome, populacao)) %>% #reordena a variável
  nome a partir do valor da variável população
  ggplot()+
  geom_col(aes(x=populacao, y=nome))

```

```
#podemos preencher as barras com cores que mudam de acordo com o
valor de uma variável
#fazemos isso associando o parâmetro fill da função aes.
#indicamos para esse parâmetro qual a variável que servirá de
referência para as cores.
#No caso abaixo as 10 cidades mais populosas terão as cores de suas
barras preenchidas de acordo com a variável nome_região
```

```
dados_municipios %>%
  slice_max(populacao, n=10) %>%
  mutate(nome=reorder(nome, populacao)) %>%
  ggplot() +
  geom_col(aes(x=populacao, y=nome, fill=nome_regiao))
```

Script da videoaula.

Fonte: Barbalho (2020).

```
#Instale o pacote dados
install.packages("dados")

library(dados)
library(tidyverse)

#vamos agora fazer algumas experiências com gráficos de pontos
#usamos para isso a função geom_point
#No gráfico abaixo identificamos a relação entre pib_per_capita no
eixo x e expectativa de vida no eixo y
#fazemos o filtro para os países das Américas no ano de 2007

dados_gapminder %>%
  filter(continente=="Américas") %>%
  filter(ano==2007) %>%
  ggplot() +
  geom_point(aes(x=pib_per_capita, y=expectativa_de_vida))

#Podemos incluir mais um continente no filtro para isso vamos usar
o operador %in% que checa se um determinado valor está presente num
conjunto

conjunto_continentes <- c("Américas", "África") #Cria um objeto com
o conjunto de continentes a ser usado no filtro
```

```

dados_gapminder %>%
  filter(continente %in% conjunto_continentes) %>%
  filter(ano==2007) %>%
  ggplot()+
  geom_point(aes(x=pib_per_capita, y=expectativa_de_vida))

#O gráfico não permite identificar quais pontos são de países da
África e quais são das Américas.
#Vamos colorir os pontos de acordo com o continente usando o
parâmetro color dentro da função aes

dados_gapminder %>%
  filter(continente %in% conjunto_continentes) %>%
  filter(ano==2007) %>%
  ggplot()+
  geom_point(aes(x=pib_per_capita, y=expectativa_de_vida, color=
continente))

#Os dados do dataframe dados_gapminder referem-se a uma série de
indicadores coletados ao longo de vários anos
#Para saber como um determinado indicador evoluiu uma opção
interessante é usar o gráfico de linha
#Faremos uso da função geom_line

#O gráfico abaixo mostra a evolução da expectativa de vida no Brasil
desde o início da série histórica
dados_gapminder %>%
  filter(pais == "Brasil") %>%
  ggplot()+
  geom_line(aes(x=ano, y=expectativa_de_vida))

#Talvez o uso de pontos ajude a identificar os anos em que ocorreram
as coletas de dados.
#Nesse caso vamos trabalhar com a combinação de duas geometrias:
geom_line e geom_point. Veja abaixo

dados_gapminder %>%
  filter(pais == "Brasil") %>%
  ggplot()+
  geom_line(aes(x=ano, y=expectativa_de_vida)) +
  geom_point(aes(x=ano, y=expectativa_de_vida))

```

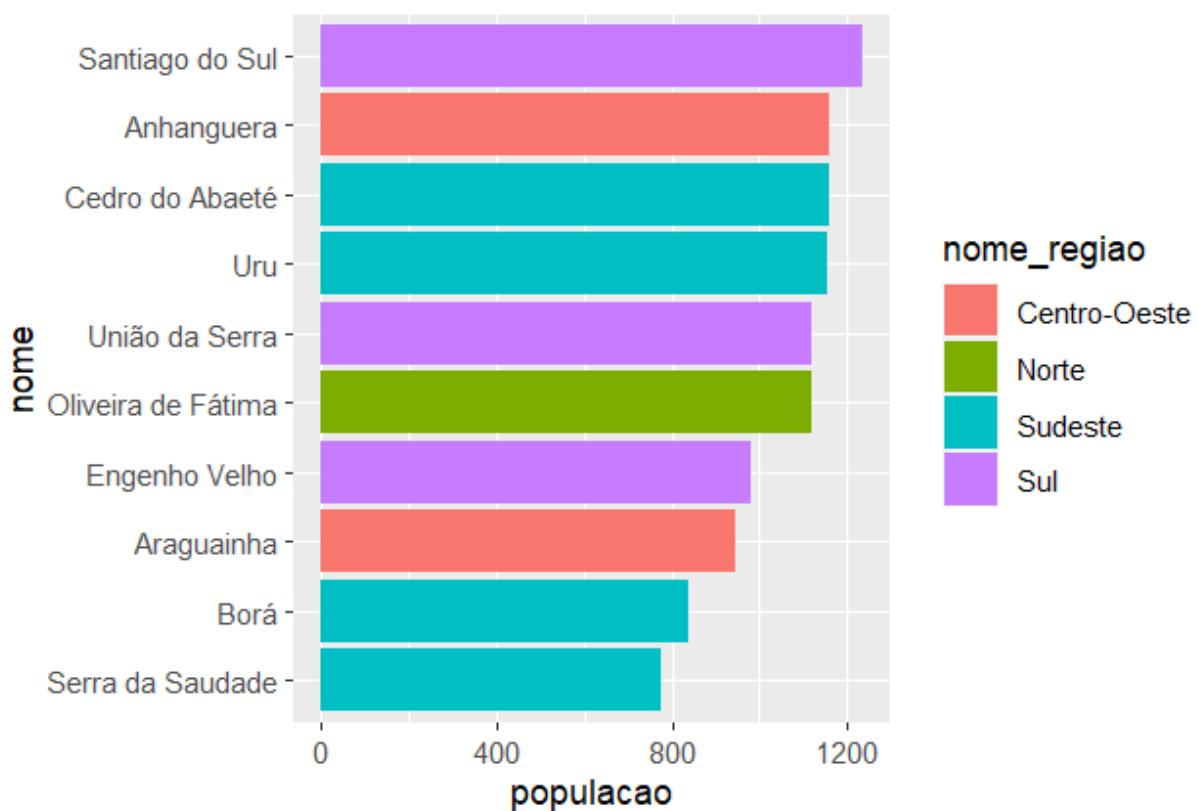
Script da videoaula.

Fonte: Barbalho (2020).

Agora uma lista de práticas sugeridas.

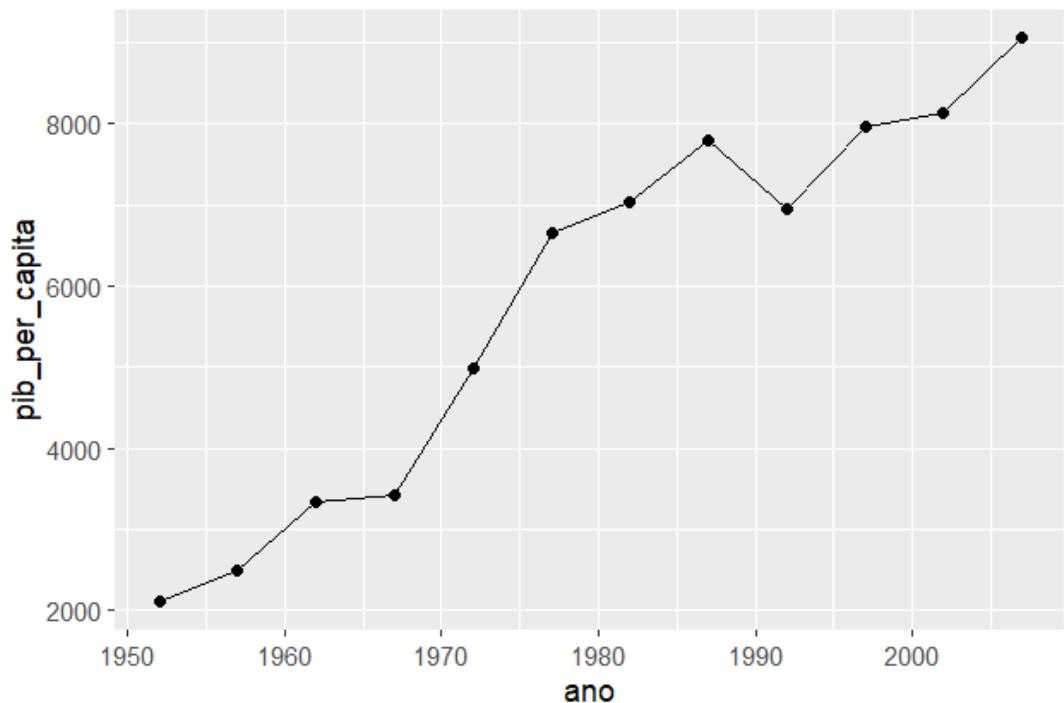
- No *script* das videoaulas, faça um gráfico com os 10 municípios de menor população do Brasil usando a função “slice_min” no lugar da que já utilizamos anteriormente, “slice_max”. Preencha as colunas com as cores das regiões. Procure analisar o gráfico e veja quais são as regiões que mais aparecem.

Veja as populações dos municípios e compare com o gráfico dos dez maiores municípios. Tire suas conclusões. Ah, o gráfico que você vai gerar é semelhante a esse aqui.



Refaça o gráfico de linha + ponto em relação ao script do nível 2, agora acompanhando a evolução de “pib_per_capita” ao longo dos anos. Procure entender mais uma vez o que o gráfico está mostrando, por exemplo: há um crescimento constante do PIB *per capita*? Há uma estabilização dessa variável? Tente comparar o gráfico que você fez com a evolução da expectativa de vida. Será que é possível afirmar que tanto a expectativa de vida como o PIB *per capita* tem comportamento de crescimento ou retração semelhantes?

O gráfico **ano x PIB per capita** que você vai gerar deve ser semelhante a esse:



Exemplo de gráfico (ano x PIB per capita) a ser gerado.

Fonte: Barbalho (2020).

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica: o caso de política pública sobre dados abertos.** 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN.** GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

IHAKA, Ross. **The R Project:** a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo.** [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R.** RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdadosus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

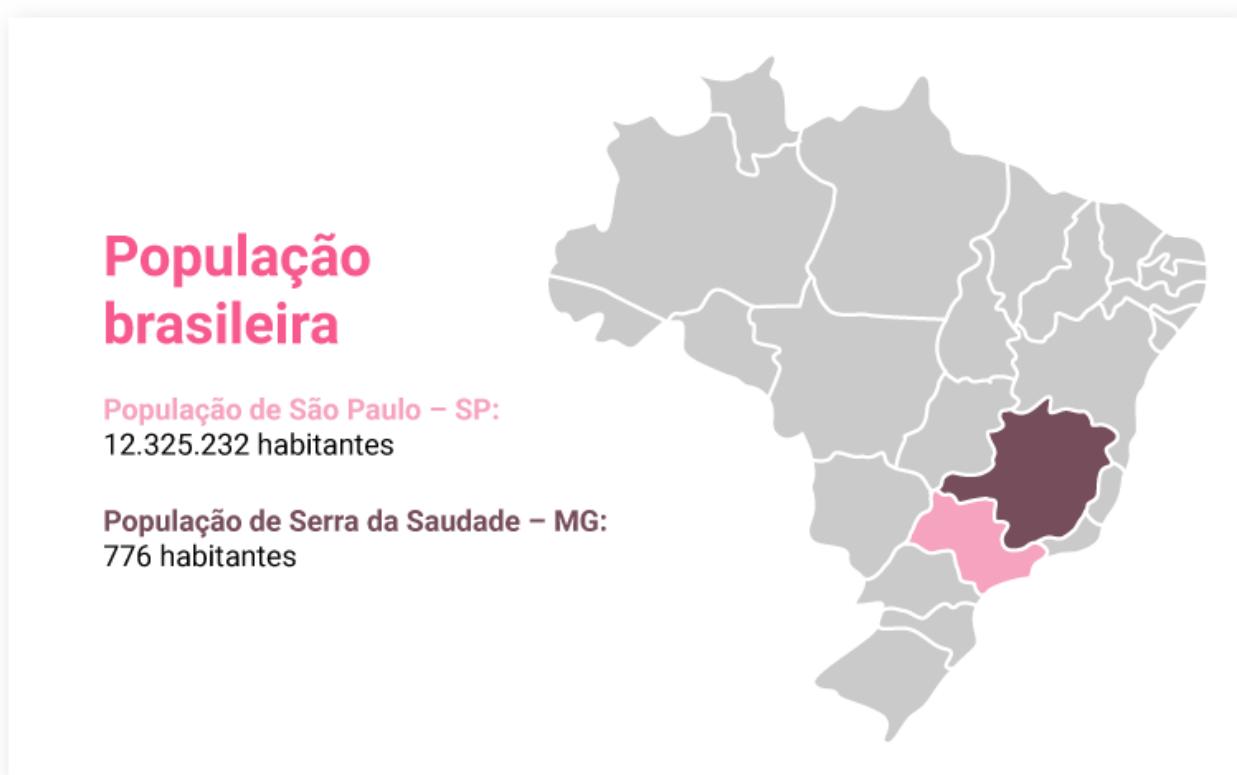
Unidade 3: Transformando Dados Usando R

Objetivo de aprendizagem

Ao final desta unidade você será capaz de identificar situações que requerem transformação de dados em R.

3.1 Transformação Logarítmica

O Brasil é um país imenso e muito diverso. Essa afirmação quase clichê se revela em várias situações, inclusive nas análises de dados. Analise, por exemplo, o caso da população das cidades brasileiras na estimativa de 2020. Observe esses dois extremos:



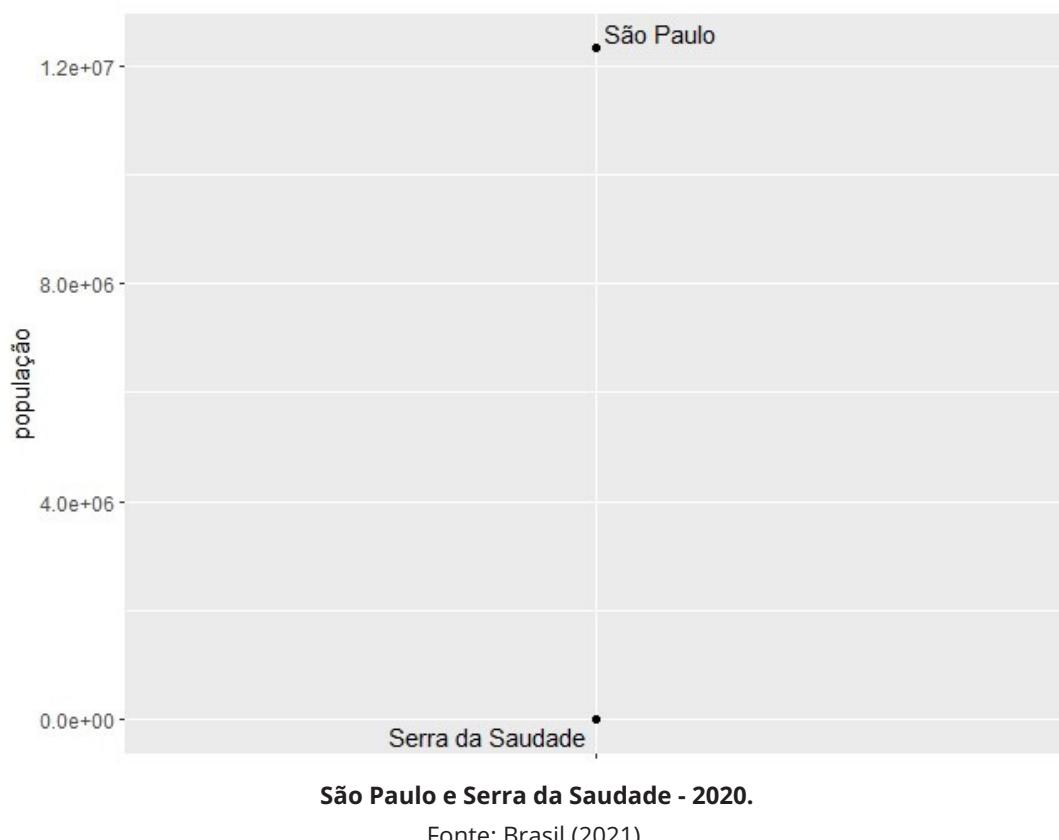
População brasileira.

Fonte: Freepik (2023). Elaboração: CEPED/UFSC (2023).

É impressionante a diferença entre esses dois dados, concorda? Se por algum motivo um contingente populacional do tamanho da população de Serra da Saudade resolvesse sair da capital paulista, praticamente não se perceberia qualquer alteração na vida da grande metrópole nacional. É bem provável que o número de paulistanos que desce a Serra do Mar para aproveitar um feriado em Santos seja muito maior do que toda a população da cidade menos populosa do Brasil.

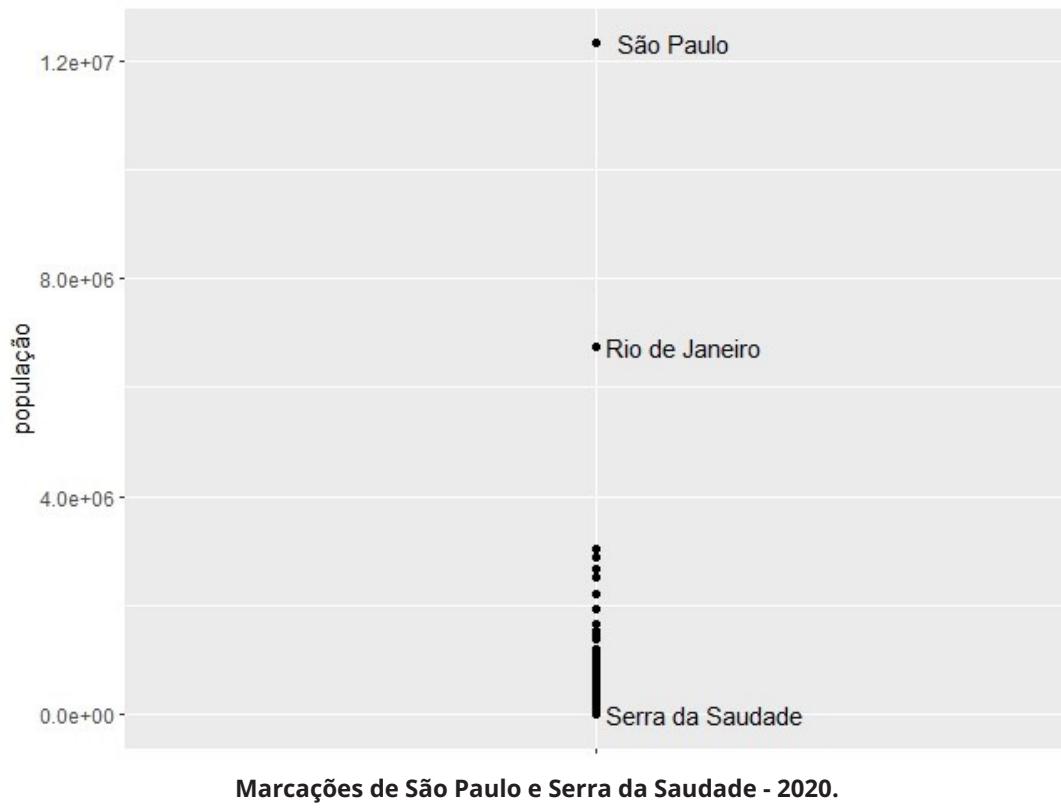
Agora, veja o impacto desses extremos quando se analisa o conjunto dos municípios brasileiros.

Analise um gráfico que mostra, inicialmente, apenas São Paulo e Serra da Saudade.



Observe a enorme distância entre as duas cidades na grandeza de quantidade de população. Caso se queira representar a população de todos os outros municípios brasileiros, os pontos que marcam as populações devem ficar entre as marcações de São Paulo e Serra da Saudade.

Veja como fica essa representação neste gráfico.



Com a inclusão da população de todos os outros 5570 municípios fica fácil observar que os pontos estão mais próximos de Serra da Saudade do que de São Paulo. Observe que há apenas um ponto isolado, referente ao município do Rio de Janeiro que fica a meio caminho entre os dois extremos.

Em outras palavras, há mais de 5.500 municípios representados em uma pequena porção do gráfico acima. Esse tipo de situação pode levar a conclusões equivocadas ao pensar, por exemplo, sobre a população que melhor caracteriza os municípios brasileiros.

Conforme já aprendido por você, as medidas de centralidade, dentre as quais média e mediana são as principais, têm como missão caracterizar em um único valor o que está representado num conjunto de dados. Dito isso, veja os valores da média e da mediana da população brasileira em 2020 calculada com as funções "mean" e "median" do RStudio.

Média = 38.017,18 habitantes

Mediana = 11.665,50 habitantes

Observe que há uma enorme diferença entre os dois valores das medidas centrais. Ao dividir os valores você pode ver que a média é 3,2 vezes maior que a mediana. Compreendendo que a mediana indica o ponto que separa um conjunto de dados ordenado em dois conjuntos de mesmo tamanho, pode-se depreender que 50% dos municípios brasileiros possuem até 11.665 habitantes. Portanto não parece lógico que a média, que apresenta um valor 3 vezes maior do que a mediana, represente a medida que melhor caracteriza o conjunto da população dos municípios brasileiros.



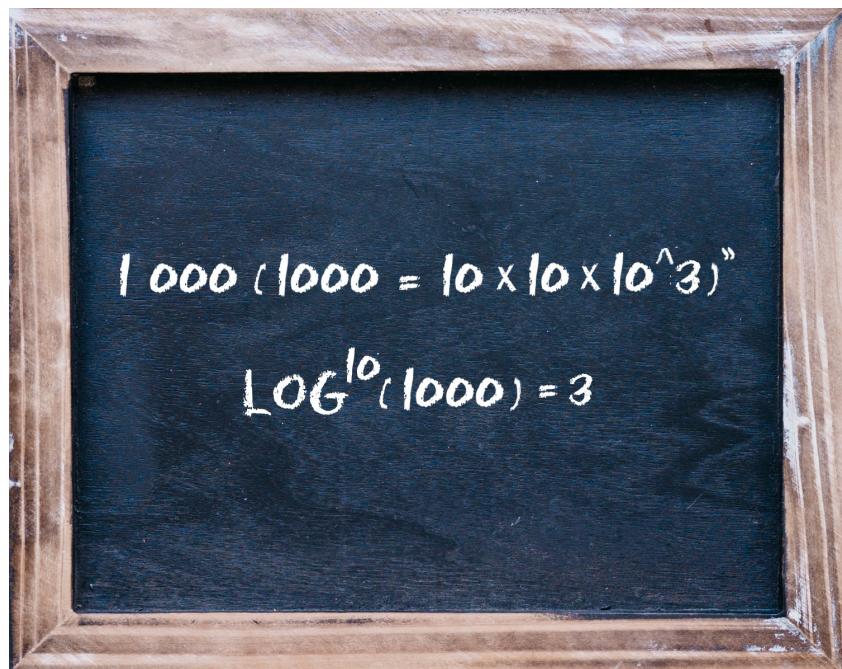
DESTAQUE

Com o que foi apresentado acima fica fácil indicar que a mediana é a medida de centralidade mais adequada para caracterizar a quantidade de habitantes dos municípios brasileiros. A mediana é bem menos influenciada por valores extremos como os números da população de São Paulo, Rio de Janeiro, ou mesmo de outras grandes metrópoles como Salvador, Fortaleza e Belo Horizonte. Nesse sentido, toda vez que houver esse tipo de situação a mediana passa a ser medida preferencial de centralidade.

Porém esse assunto não se encerra com a constatação de qual medida é melhor para descrever o ponto central de um conjunto de dados. Frequentemente, a média precisa ser usada para resolução de problemas de estatística mais avançados. Além disso, o gráfico que mostra os pontos de quase todos os municípios acumulados em uma pequena região da figura não nos permite fazer análises mais detalhadas, principalmente sobre a dispersão da população. Nessas situações se faz necessário encontrar mecanismos que aproximem a medida de população de São Paulo da de Serra da Saudade. E é aí onde entra o logaritmo.

Conforme encontrado na Wikipedia, o “logaritmo de um número é o expoente a que outro valor fixo, a base, deve ser elevado para produzir este número. Por exemplo, o logaritmo de 1000 na base 10 é 3 porque 10 elevado ao cubo é 1 000 ($1000 = 10 \times 10 \times 10 = 10^3$)” (WIKIMEDIA FOUNDATION, 2022). Em outras palavras, temos que:

$$\log_{10}(1000) = 3.$$



Logaritmo de 1000 na base 10 é 3.

Fonte: Freepik (2023). Elaboração: CEPED/UFSC (2023).

Com essa definição, calcule o logaritmo na base 10 das populações de São Paulo e de Serra da Saudade. Ah! Essa operação inclusive será denominada de transformação logarítmica.

$$\log_{10}(12325232) = 7,090795$$

$$\log_{10}(776) = 2,889862$$

Agora subtraia os resultados dos logaritmos calculados:

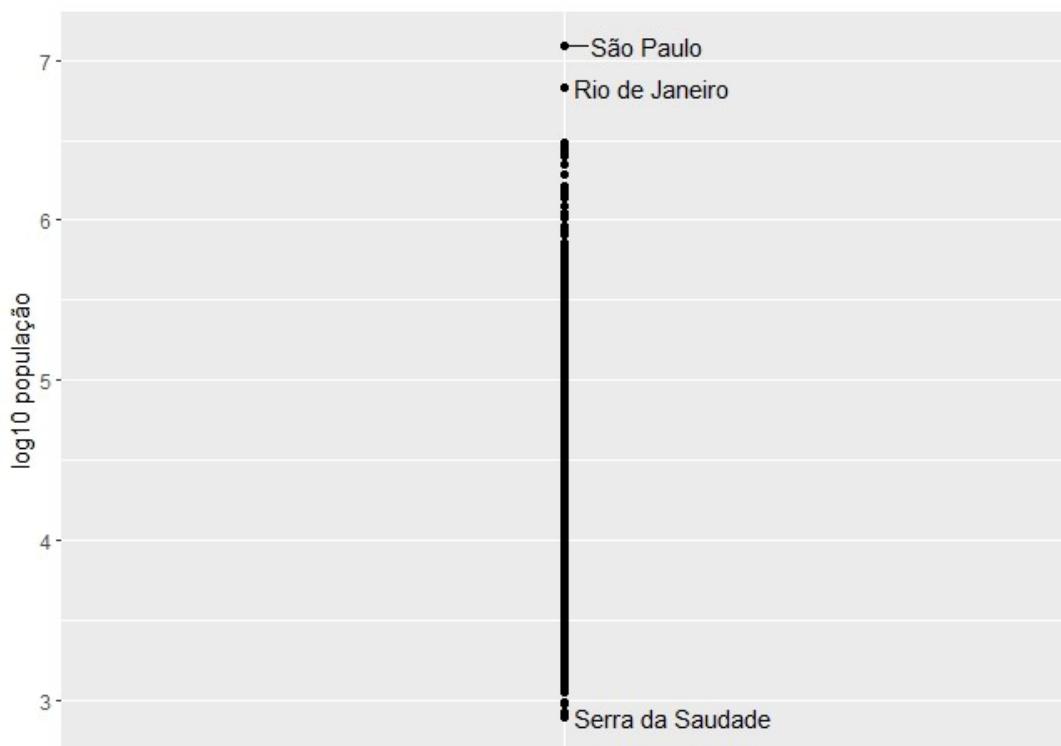
$$7,090795 - 2,889862 = 4.200933$$

Você percebeu o que aconteceu? A distância que separava São Paulo de Serra da Saudade que era mais de 12 milhões virou apenas 4,2. Veja o efeito dessa transformação na média e na mediana.

Média do logaritmo das populações dos municípios: 4,11

Mediana do logaritmo das populações dos municípios: 4,06

Isso mesmo, a média e a mediana apresentam quase os mesmos valores. A diferença passa a ser muito pequena. E agora veja o que ocorre com aquele gráfico de pontos com todas as cidades quando aplicada a transformação logarítmica para as populações de todos os municípios.



Transformação logarítmica para as populações de todos os municípios.

Fonte: Brasil (2021).

Fica fácil perceber agora que os pontos das populações dos municípios se distribuem de forma mais uniforme ao longo da linha vertical. A transformação em logaritmo permite assim uma análise melhor de medidas de dispersão. Mais adiante serão apresentados dois tipos de gráficos úteis para avaliar melhor a dispersão dos dados: **histograma** e **box-plot**. Na utilização desses gráficos ficará ainda mais claro o efeito da transformação logarítmica.

3.2 Elaborando um Histograma com Escala Logarítmica

O histograma é o tipo de gráfico utilizado para verificar como a quantidade de ocorrências dos diversos valores ou intervalos de valores que aparecem em um conjunto de dados, se distribuem. Imagine os dados da tabela de uma hipotética coleta de dados sobre o peso de vinte crianças com 10 anos de idade em uma escola rural do Brasil.

peso
28
29
29
29
29
29
30
30
30
30
30
30
30
31
31
31
31
32
32

Queremos organizar os dados de forma que se possa contar quantas crianças estão enquadradas em cada um dos pesos que são exibidos na tabela que você acabou de ver. E para ter uma interpretação mais rápida dos dados, queremos uma visualização gráfica. É justamente nesse ponto que o histograma aparece como opção. Veja como fica um típico gráfico de histograma para essa situação.

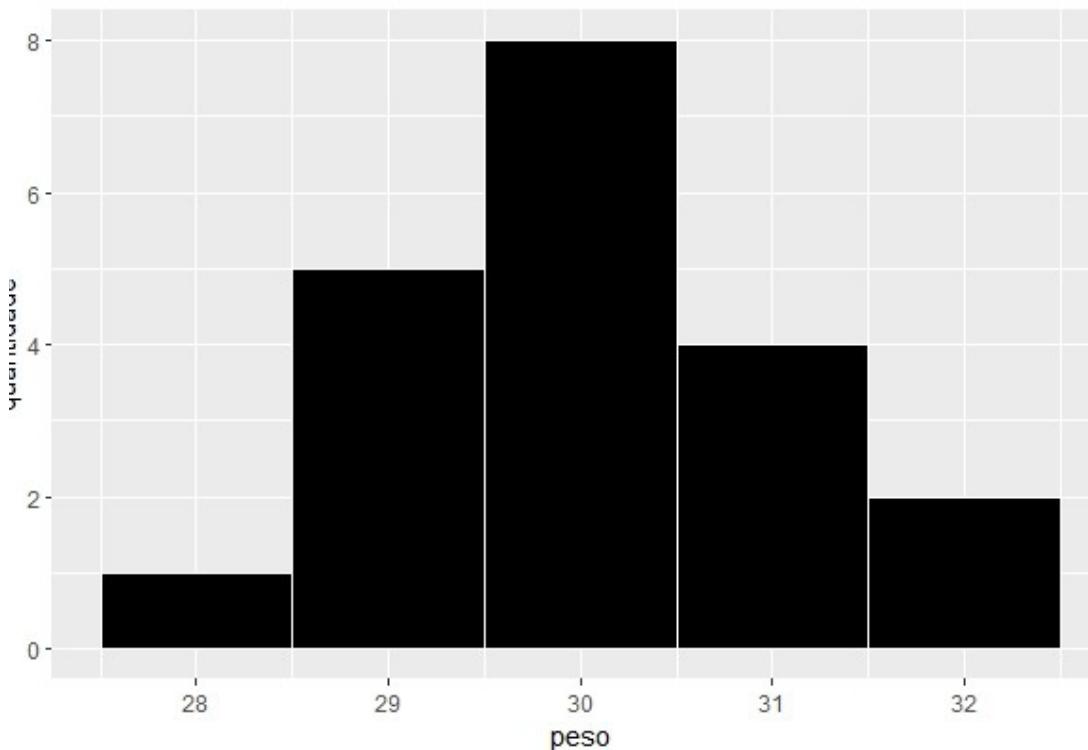


Gráfico de histograma.

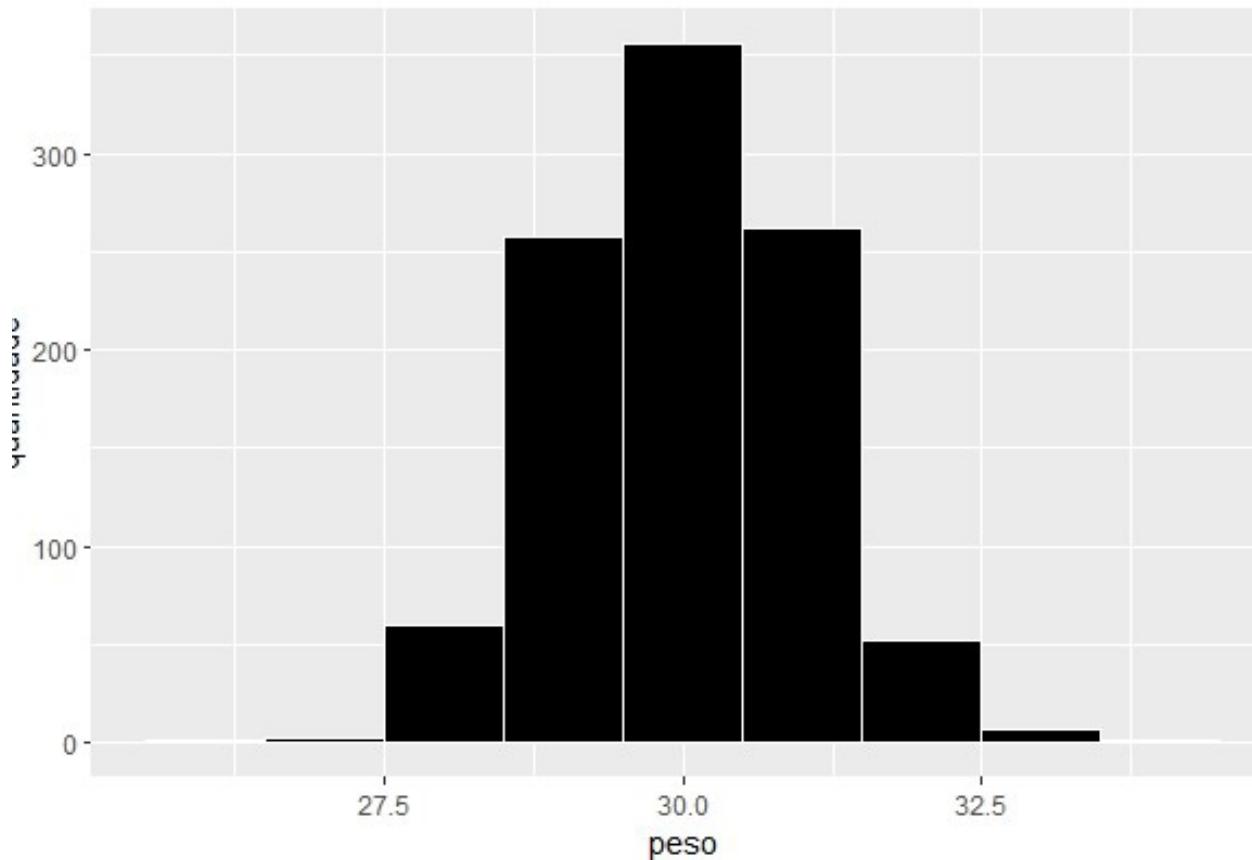
Elaboração: CEPED/UFSC (2023).

Volte sua atenção para o gráfico e veja que no eixo horizontal **x** estão exibidos os pesos dos alunos que aparecem na tabela. Como a tabela tem apenas 20 linhas é simples perceber que o gráfico mostra, no eixo vertical, a quantidade de ocorrências de alunos com o peso indicado no eixo **x**.

Por exemplo, existe apenas um aluno com peso 28, portanto o gráfico apresenta a quantidade 1 junto ao peso 28. Já o peso 30 é o mais frequente e isso pode ser visto na altura da coluna do gráfico. Temos marcada a quantidade 8 que é justamente a quantidade que pode ser contada na tabela. O histograma é, portanto, uma forma rápida de visualizar como os dados se distribuem.

Há um crescimento contínuo entre o peso 28 e o peso 30, e depois há uma queda também contínua até se chegar ao peso 32. E esse tipo de representação certamente se torna muito mais útil quando o número de ocorrências cresce.

Veja outro exemplo em que também são simulados os pesos de alunos com 10 anos de idade, agora com 1000 ocorrências.

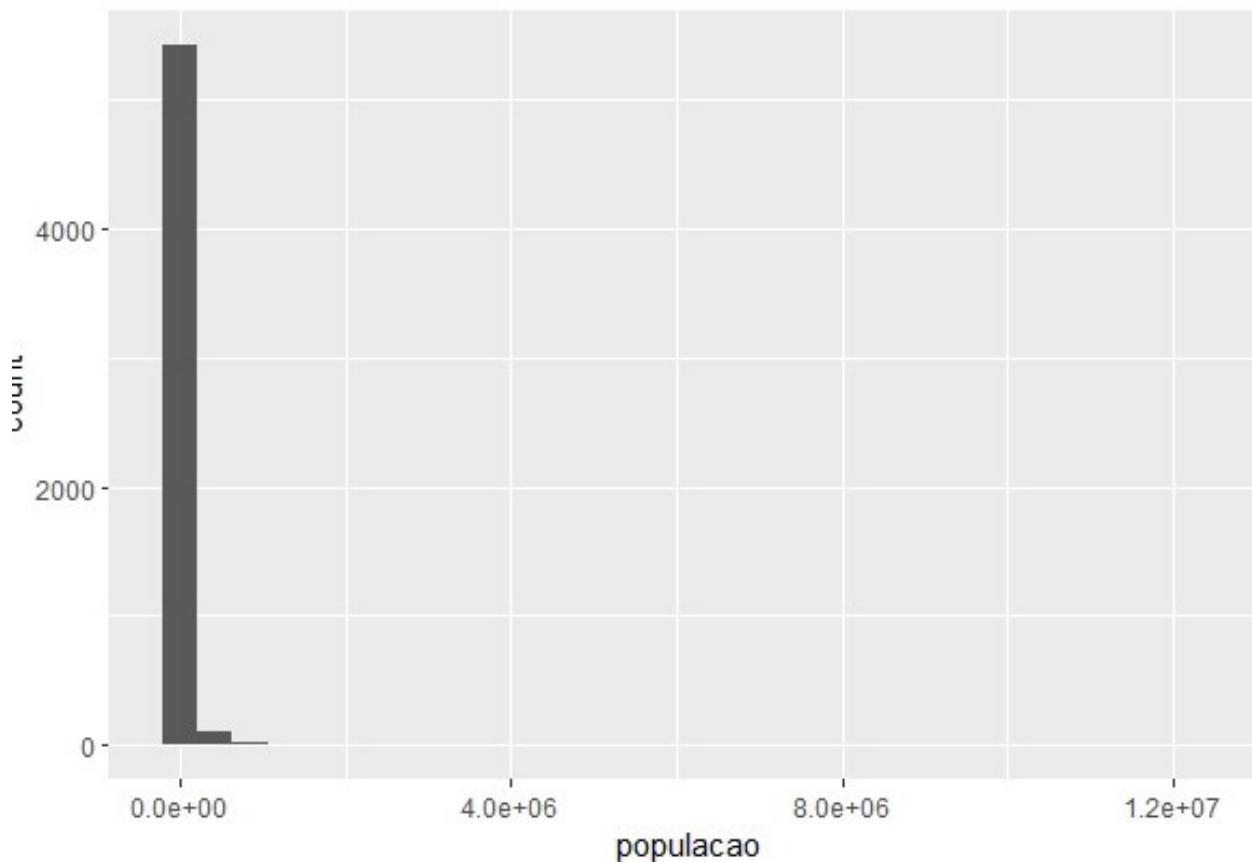


Simulação de pesos de alunos com 10 anos de idade.

Elaboração: CEPED/UFSC (2023).

No gráfico, novamente o que se vê é a predominância do peso 30, só que agora são aproximadamente 300 ocorrências para esse peso. A quantidade de ocorrências de alunos para o peso assinalado no eixo horizontal vai diminuindo de forma relativamente simétrica, tanto à esquerda como à direita, conforme os pesos se afastam da posição central.

Vamos voltar para o caso dos municípios brasileiros e montar um histograma para que você visualize a distribuição da população dos municípios.

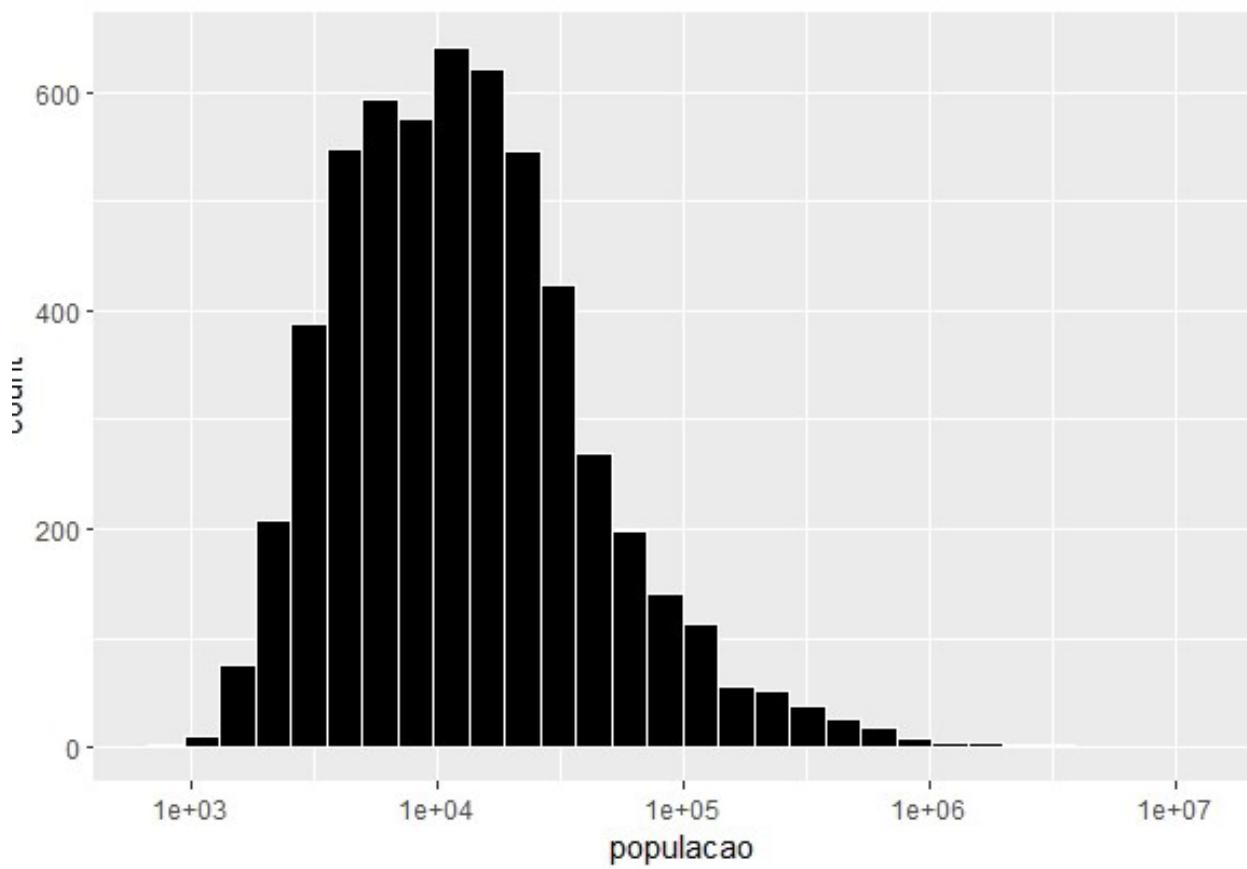


Distribuição da população dos municípios (histograma).

Elaboração: CEPED/UFSC (2023).

Observe no gráfico que não é possível ter a menor ideia da distribuição dos municípios. Isso se deve ao fato de que a maior parte dos municípios tem uma população muito pequena, ao mesmo tempo em que algumas poucas cidades ultrapassam 1 milhão de habitantes e, especificamente, há a população de São Paulo com mais de 12 milhões, o que gera a expansão horizontal do gráfico.

Você já viu que é necessário fazer a transformação da variável “população” usando logaritmo para que possa ver melhor a distribuição. Quando se trabalha com gráficos é possível usar a escala logarítmica que representa os valores no eixo horizontal ou vertical usando essa transformação. Veja o resultado.



Escala logarítmica.

Fonte: Brasil (2021).

Usando a escala logarítmica no eixo horizontal x é possível agora ver uma distribuição. Entenda que há um crescimento das quantidades de municípios até chegar a aproximadamente 10.000 ($1e+04$) habitantes. Nesse ponto pode-se entender que a quantidade de municípios com aproximadamente 10.000 habitantes é um pouco mais do que 600.

A contagem de municípios se reduz gradualmente à medida que aumenta a população. Quando se atinge o marco de 100.000 habitantes ($1e+05$) já se vê que apenas um pouco mais de 100 municípios atingem essa população. E à medida que mais se afasta à direita, as quantidades de municípios assinaladas são menores.



DESTAQUE

O que o analista deve sempre lembrar aqui é que não está sendo usada uma escala linear onde o intervalo entre dois pontos é sempre igual. Na escala logarítmica cada posição corresponde a um aumento no número do expoente que eleva o valor da base 10. Na primeira marcação, esse expoente é 3, o que leva a $10^3 = 1000$, na segunda marcação o expoente é 4, implicando no valor $10^4 = 10.000$ e assim sucessivamente.

Então é isso! Você aprendeu aqui sobre o uso do gráfico de histograma. Você conseguiu perceber a sua utilidade para verificar a distribuição do dado? Viu também como agir caso os dados estejam muito concentrados por conta da presença de alguns valores extremos. No próximo tópico conheça o *box-plot* e sua utilidade para identificar alguns pontos importantes da análise estatística.

3.3 Box-Plot e Escala Logarítmica

Três medidas resumo importantes que não foram apresentadas até o momento serão apresentadas no gráfico *box-plot*. São elas: **primeiro-quartil, segundo-quartil ou mediana e terceiro-quartil**. Os valores dessas medidas são descobertos quando se ordena de forma crescente a variável que se deseja investigar e divide o conjunto de dados em quatro subconjuntos de mesmo tamanho, daí o nome quartil. Vamos ilustrar com o exemplo abaixo.

Conjunto de dados original:

{3, 1, 2, 20, 5, 4, 6, 7, 14, 10, 11, 9, 12, 8, 13, 15, 18, 16, 17, 19}

Conjunto de dados original ordenado em ordem crescente:

{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}

Quatro subconjuntos derivados do conjunto de dados original ordenado:

{1, 2, 3, 4, 5}

{6, 7, 8, 9, 10}

{11, 12, 13, 14, 15}

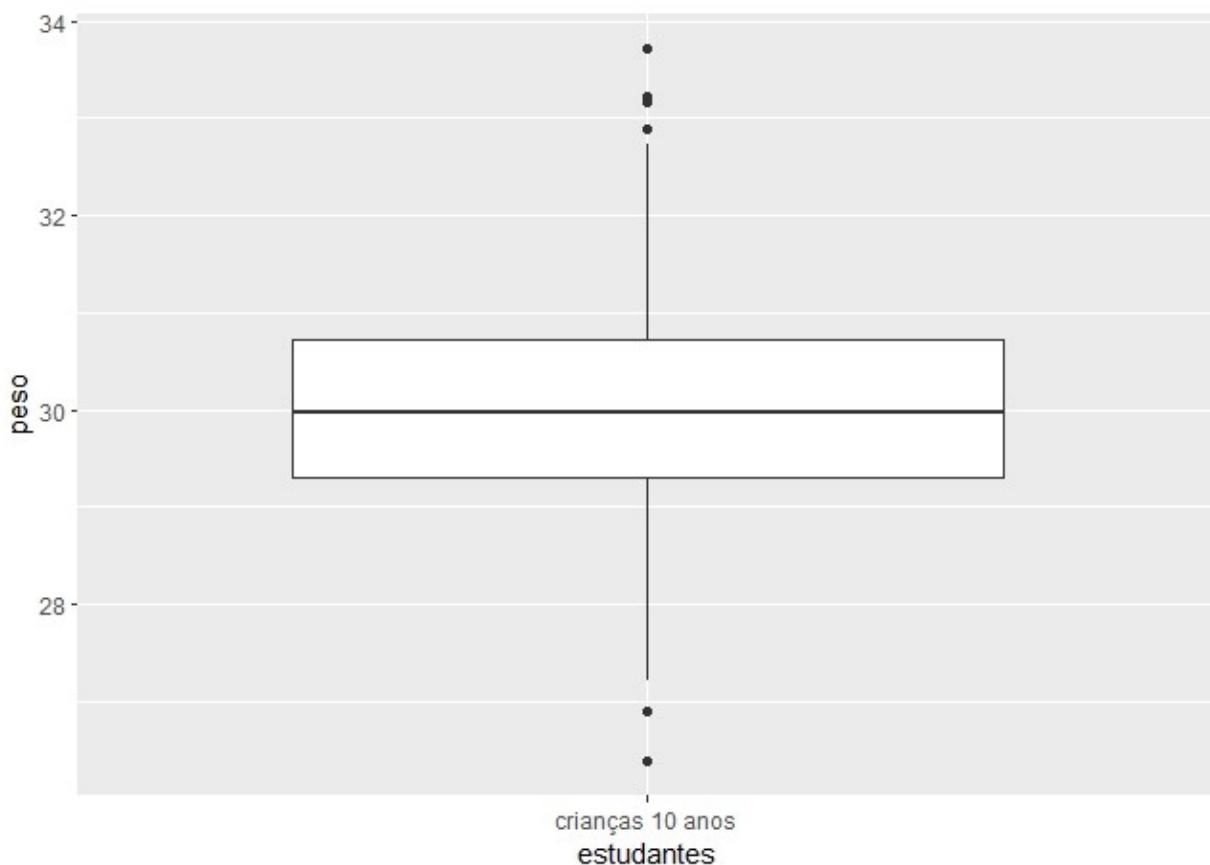
{16, 17, 18, 19, 20}

O primeiro quartil se refere ao valor mais alto do primeiro subconjunto de dados. Isso significa que o valor que está no primeiro quartil é maior do que 25% de todos os valores do conjunto de dados original da variável. No caso do exemplo simples que se está trabalhando, o primeiro quartil é **5**.

O segundo quartil segue o mesmo procedimento de identificação da mediana e corresponde ao valor mais alto do segundo sub-conjunto de dados, portanto é maior do que 50% de todos os valores do conjunto. Para o nosso exemplo, o segundo quartil é **10**.

O terceiro quartil corresponde ao valor que está no final do terceiro subconjunto, que é maior do que 75% de todos os valores. No nosso caso simples, é **15**.

Com essa introdução sobre os quartis já é possível apresentar o *box-plot* por meio do exemplo dos pesos das mil crianças já trabalhados no tópico anterior.



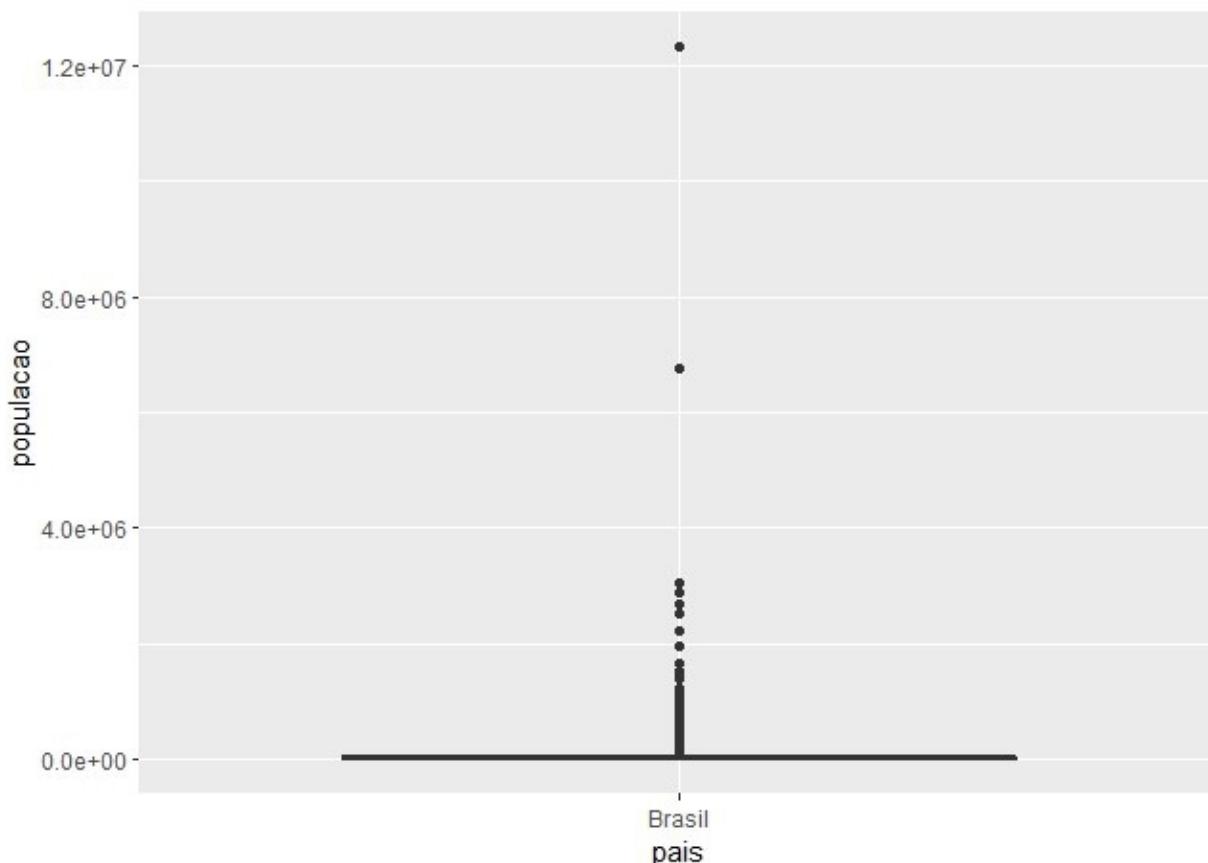
Box-plot (pesos das mil crianças).

Elaboração: CEPED/UFSC (2023).

No gráfico se pode ver uma caixa formada por três retas horizontais. A reta horizontal mais baixa corresponde ao primeiro quartil, a reta intermediária indica o segundo quartil e, por fim, a terceira reta horizontal, mostra o terceiro quartil. Os valores entre a primeira reta horizontal e a terceira reta horizontal é conhecido como intervalo interquartil, ou seja, o valor que corresponde à diferença entre o terceiro quartil e o primeiro quartil.

Note ainda uma reta vertical. Essa reta corresponde aos valores que são aceitos como valores dentro do esperado de acordo com a distribuição do dado. Por fim veja alguns pontos isolados tanto acima quanto abaixo da reta vertical. Esses pontos são conhecidos como pontos extremos. São aqueles que estão fora de um intervalo que se poderia esperar da distribuição dos dados. Temos então pontos extremos superiores, os que estão acima da reta vertical e pontos extremos inferiores, os que estão abaixo da reta vertical.

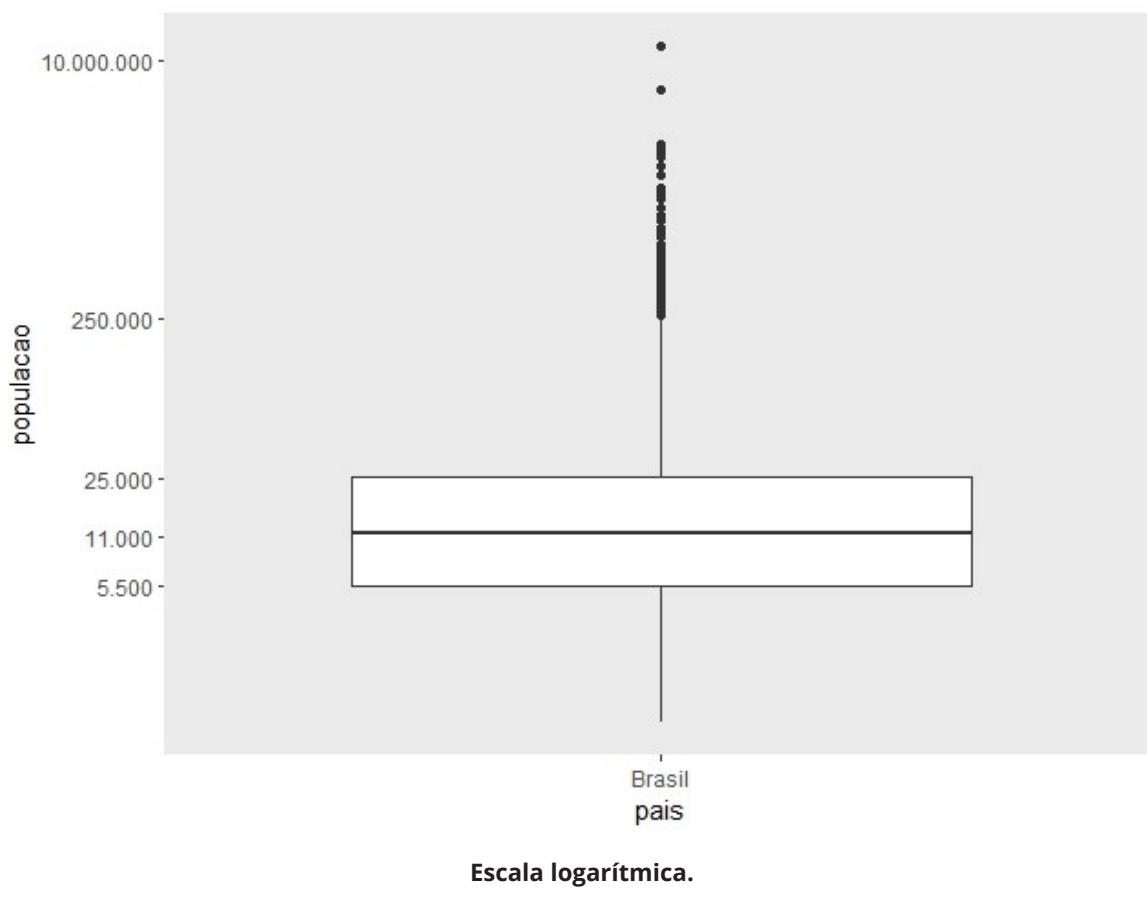
Experimente agora um *box-plot* para o caso das populações dos municípios brasileiros.



Box-plot para o caso das populações dos municípios brasileiros.

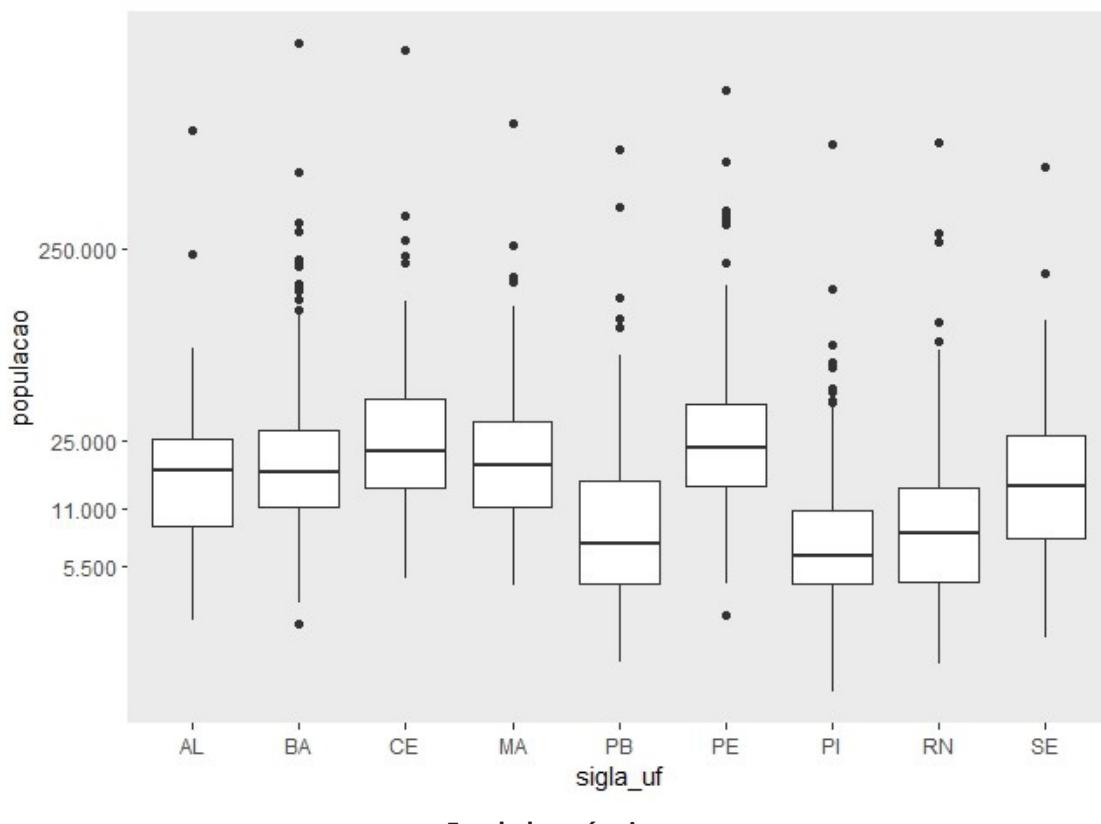
Elaboração: CEPED/UFSC (2023).

O *box-plot* com a escala original dos dados de população de municípios sofre exatamente dos mesmos problemas de concentração de valores que já foram mencionados anteriormente. Observe que na figura não é possível ver a caixa que marca o intervalo interquartil. Só se consegue distinguir vários pontos que correspondem a populações extremas de alguns municípios. É necessário portanto apelar para a escala logarítmica para que se possa fazer uma leitura dos dados. Veja isso na próxima imagem.



Agora sim, com a escala logarítmica é possível ver todos os elementos importantes para uma análise com box-plot. Descobre-se assim que o primeiro quartil que corresponde à primeira reta horizontal é aproximadamente 5500. Ou seja, 25% dos municípios brasileiros possuem menos que 5500 habitantes. A mediana, representada pela segunda reta horizontal, é 11.000 habitantes, e o terceiro quartil é 25000, o que indica que 75% dos municípios brasileiros têm menos de 25000 habitantes. Vale ainda destacar que os pontos extremos começam a partir da população 250.000. É possível ver uma série desses pontos que culminam com uma marcação acima de 10.000.000 de habitantes que corresponde ao município de São Paulo.

Uma análise interessante que pode ser feita por você com *box-plot* é o corte por variável categórica. Imagine, por exemplo, que você deseja verificar como se distribuem as populações dos municípios para cada um dos estados do Nordeste. O resultado é o que verá agora.



Pelo gráfico fica fácil perceber vários achados. Olha só essa lista com alguns:

- Ceará e Pernambuco possuem as duas maiores medianas;
- Ceará e Pernambuco também possuem os maiores terceiros quartis;
- Bahia e Pernambuco são os únicos estados que possuem valores extremos inferiores;
- Piauí é o estado de menor mediana; e
- Alagoas e Sergipe, com duas ocorrências cada, são os estados que têm a menor quantidade de valores extremos superiores.

Você consegue ver outros achados? Dê uma olhada com calma e pense o que mais pode ser descoberto a partir do *box-plot*.

Todos esses gráficos acima foram produzidos usando o {ggplot2}. No próximo tópico você verá uma videoaula que demonstra como foram construídos os histogramas e *box-plot* de população dos municípios brasileiros. Comece a preparar o seu RStudio.

3.4 Análise de Distribuição de Dados com o Histograma e Box-plot usando Ggplot2

Você já conheceu alguns dos principais elementos da gramática do {ggplot2} e já sabe como se lê e para que servem os gráficos de histograma e box-plot. Agora chegou o momento de aprender como juntar tudo isso. Veja a videoaula explicativa.



Videoaula: [Box-plot e Histograma Usando Ggplot2](#)

Agora está na hora de copiar e colar o código para você repetir o que foi feito na videoaula que acabou de assistir.

```
library(tidyverse)

arquivo<-
"https://raw.githubusercontent.com/fernandobarbalho/enap_auto_
instucional/main/data/dados_municipios.csv"

dados_municipios<- read_csv(file= arquivo)

#para fazer gráficos box-plot usamos a geom_boxplot
dados_municipios%>%
  filter(nome_regiao=="Nordeste") %>%
  ggplot()+
  geom_boxplot(aes(x=sigla_uf, y=populacao))

#é necessário usar a escala logaritmica no eixo y para tornar
viável a visualização
#usa-se a função scale_y_log10 pra representação em escala logaritmica
```

```

dados_municpios%>%
  filter(nome_regiao=="Nordeste") %>%
  ggplot()+
  geom_boxplot(aes(x=sigla_uf, y=populacao)) +
  scale_y_log10()

#para fazer gráficos de histograma usamos a função geom_histogram
#no caso do histograma precisamos indicar apenas o eixo x.
#O R se encarrega de fazer a contagem necessária para gerar os
valores do eixo y

dados_municpios%>%
  ggplot()+
  geom_histogram(aes(x=populacao))

#Aqui também se faz necessário transformações logaritmicas, porém
sobre o eixo x
#Nesse caso usamos a função scale_x_log10

dados_municpios%>%
  ggplot()+
  geom_histogram(aes(x=populacao), color = "white") +
  scale_x_log10()

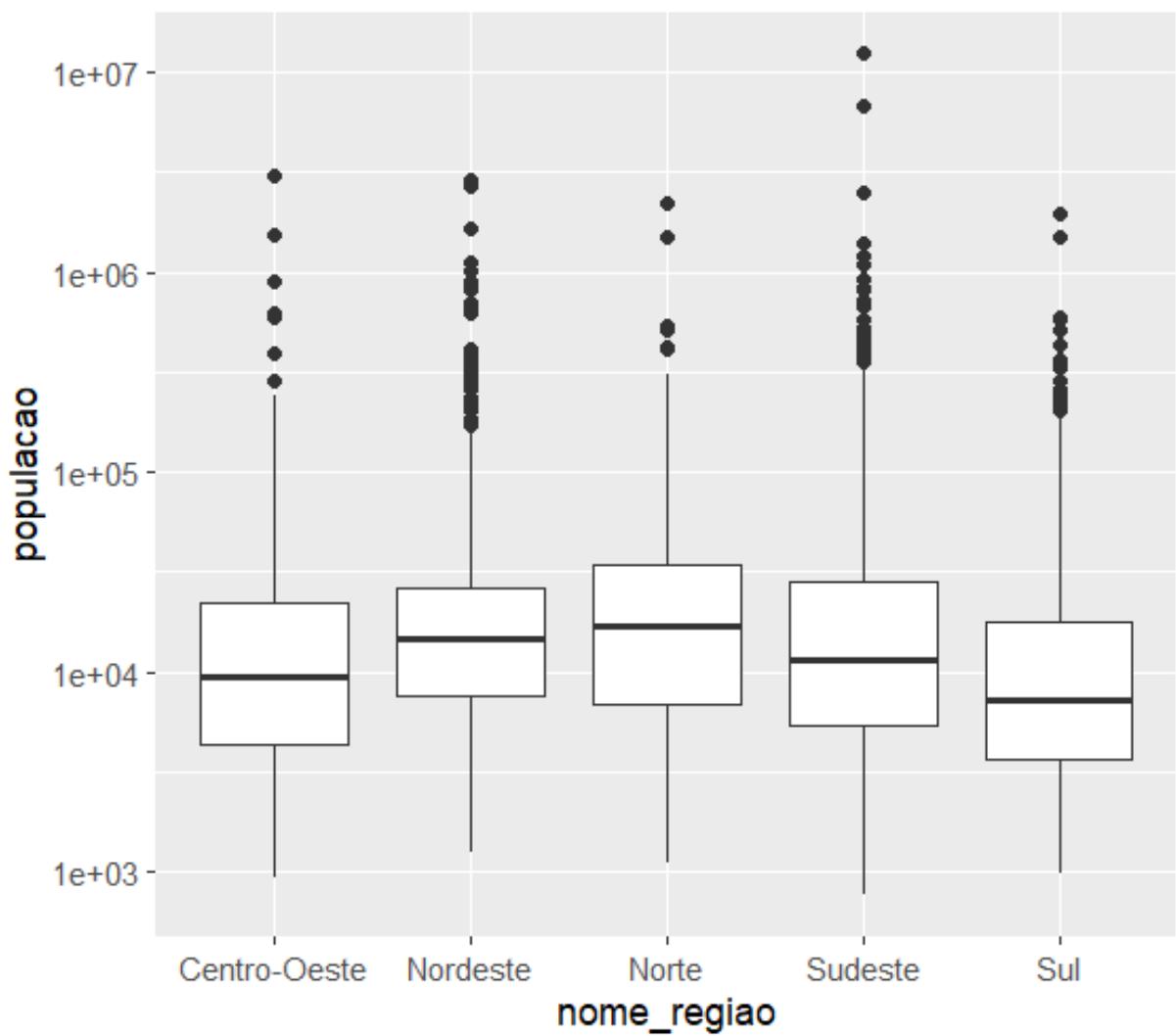
```

Código.

Fonte: Barbalho (2020).

E mais uma lista de sugestão de práticas para você:

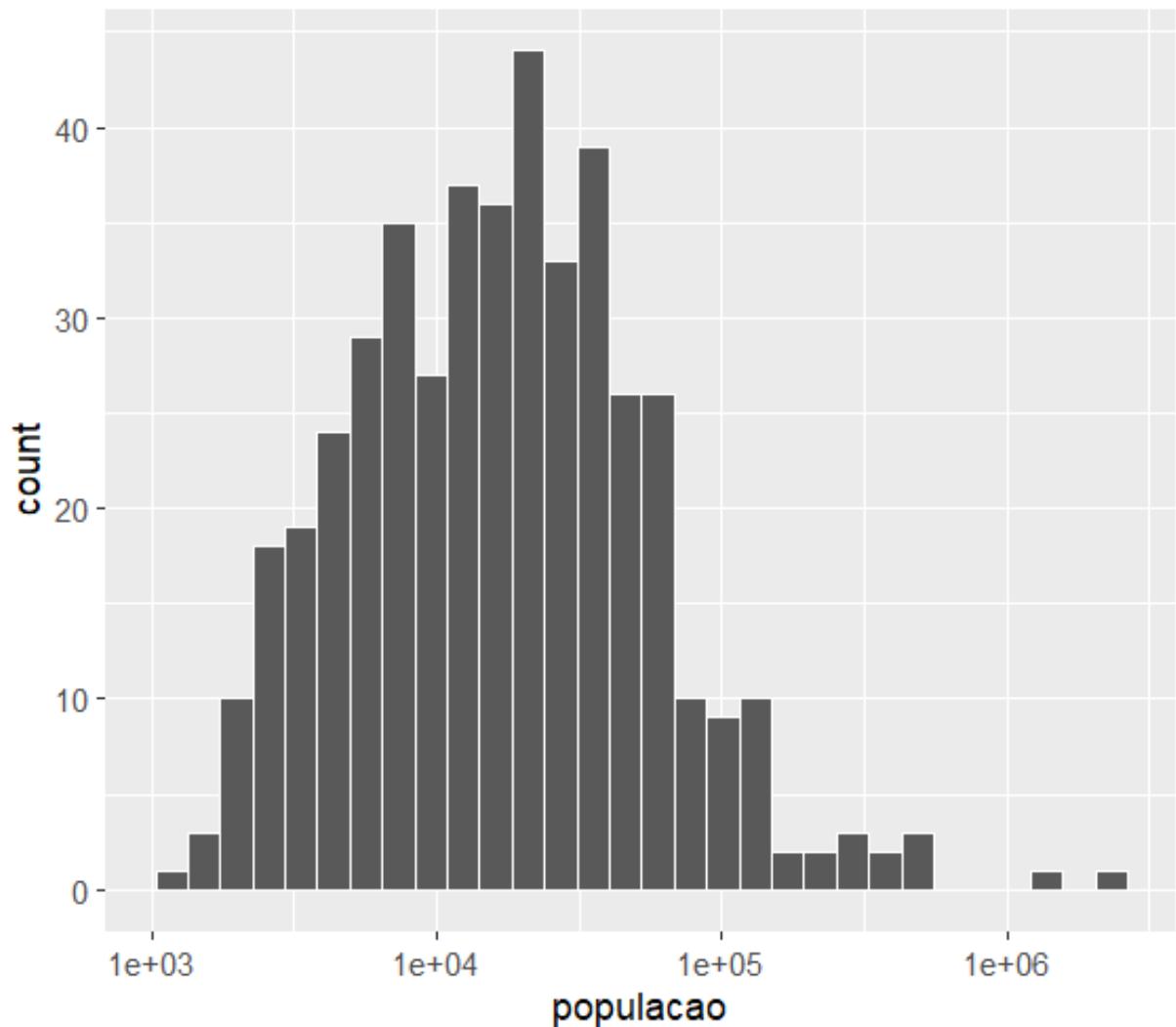
Faça o *box-plot* de população considerando a distribuição por região do Brasil. Analise o gráfico resultante, buscando entender as principais medidas: mediana, primeiro quartil e terceiro quartil, além de identificar ocorrências de pontos extremos. O gráfico deve ser gerado como o que você vê agora.



Box-plot de população.

Elaboração: CEPED/UFSC (2023).

Faça um novo histograma considerando apenas os municípios da Região Norte. Analise o resultado procurando identificar em qual faixa de população se concentra a maioria dos municípios. O gráfico gerado deve ser semelhante ao que você vê agora:



Faixa de população que concentra a maioria dos municípios.

Elaboração: CEPED/UFSC (2023).

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). **Cidades e Estados.** 2021. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>. Acesso em: 20 mar. 2023.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN.** GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

IHAKA, Ross. **The R Project:** a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo.** [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R.** RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status.** [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdadosus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Estatística.** [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Logaritmo.** [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Logaritmo>. Acesso em: 20 mar. 2023.

Unidade 4 - Análise Bivariada na Linguagem R

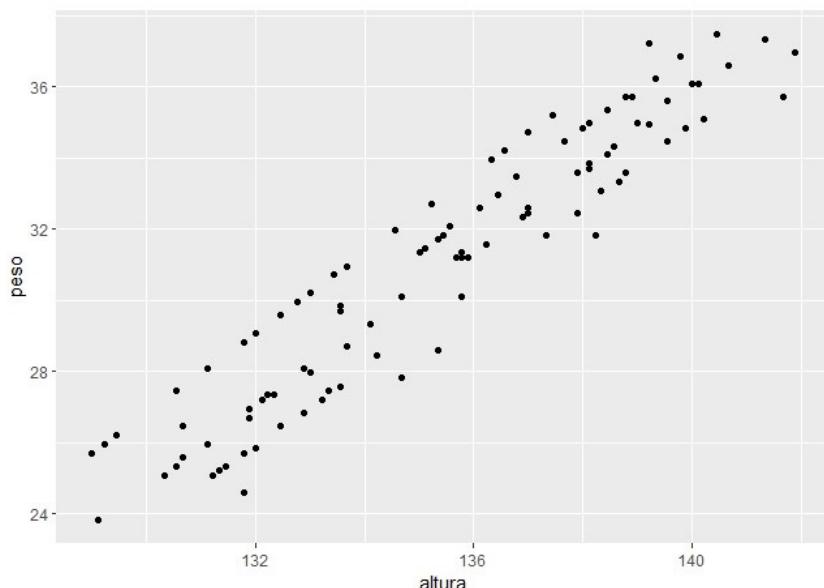
Objetivo de aprendizagem

Ao final desta unidade você será capaz de utilizar as técnicas mais comuns de análises bivariadas através da linguagem R.

4.1 Associação entre Variáveis Quantitativas: Correlação

“Conforme Bussab e Morettin (2010) a correlação entre duas variáveis é indicada por um coeficiente que mede o grau de associação entre essas duas variáveis e a proximidade dos dados a uma reta que represente essa associação.”

Veja a seguir um gráfico com dados hipotéticos que busca representar a associação entre as variáveis “altura e peso” para 100 crianças de 10 anos de idade.



Representação das variáveis altura e peso para 100 crianças de 10 anos de idade.

Elaboração: CEPED/UFSC (2023).

O gráfico usa a geometria ponto. Quando existe esse objetivo de identificar visualmente a associação entre duas variáveis, costuma-se chamar essa figura como gráfico de dispersão.

Observe no gráfico que à medida que a altura aumenta, na maioria das vezes também aumenta o peso. Isso indica que há uma forte associação entre essas duas variáveis que você já viu e ela é conhecida como correlação.



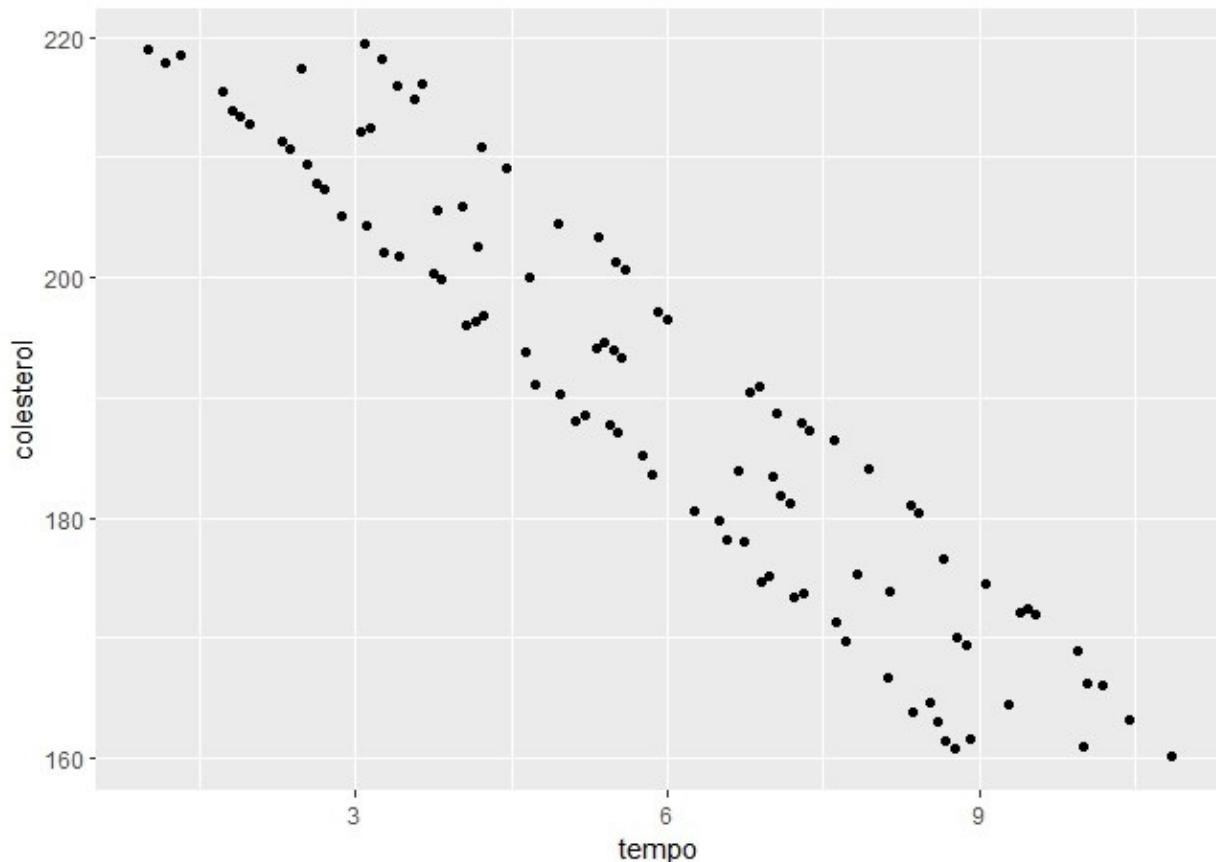
DESTAQUE

O cálculo da medida de correlação é um tanto complexo. Optou-se por não detalhar o cálculo e focar na interpretação do resultado. É importante saber que na linguagem R uma função de fácil uso faz todo o cálculo, restando ao analista a tarefa de entender o que o resultado implica para os seus objetivos.

Voltando ao último gráfico, o coeficiente de correção calculado pelo R é **0,946678**. Os valores podem variar entre -1 e 1. Como o coeficiente que se chegou é muito próximo de 1, é possível afirmar que há uma forte correlação entre peso e altura.

Vamos supor agora que se está trabalhando com políticas públicas que busquem desestimular o sedentarismo. Imagine que se quer identificar alguma correlação entre horas semanais de atividade física e o valor do colesterol total.

“Corremos atrás” dos dados fictícios de 100 atletas e montamos o gráfico que você vê agora.



Horas semanais de atividade física e o valor do colesterol total.

Elaboração: CEPED/UFSC (2023).

Observe que a forma do gráfico é exatamente o inverso do outro que mostra a associação entre altura e peso. Percebe-se claramente que quanto maior o número de horas semanais de atividade física, menor o colesterol total. O valor da correlação calculada para o gráfico é de **-0,9375099**. O valor negativo e muito próximo de -1 indica que se trata uma correlação negativa muito forte entre as duas variáveis trabalhadas.

Vamos para uma terceira situação. Agora estamos preocupados com políticas públicas sobre alimentação saudável. O que se quer verificar é se existe associação entre o consumo de alimentos naturais, medido pelo número de porções diárias, e o consumo de alimentos processados, também medido pelo número de porções diárias. Suponha que você almoçou com outras 100 pessoas e gerou o gráfico visualizado.

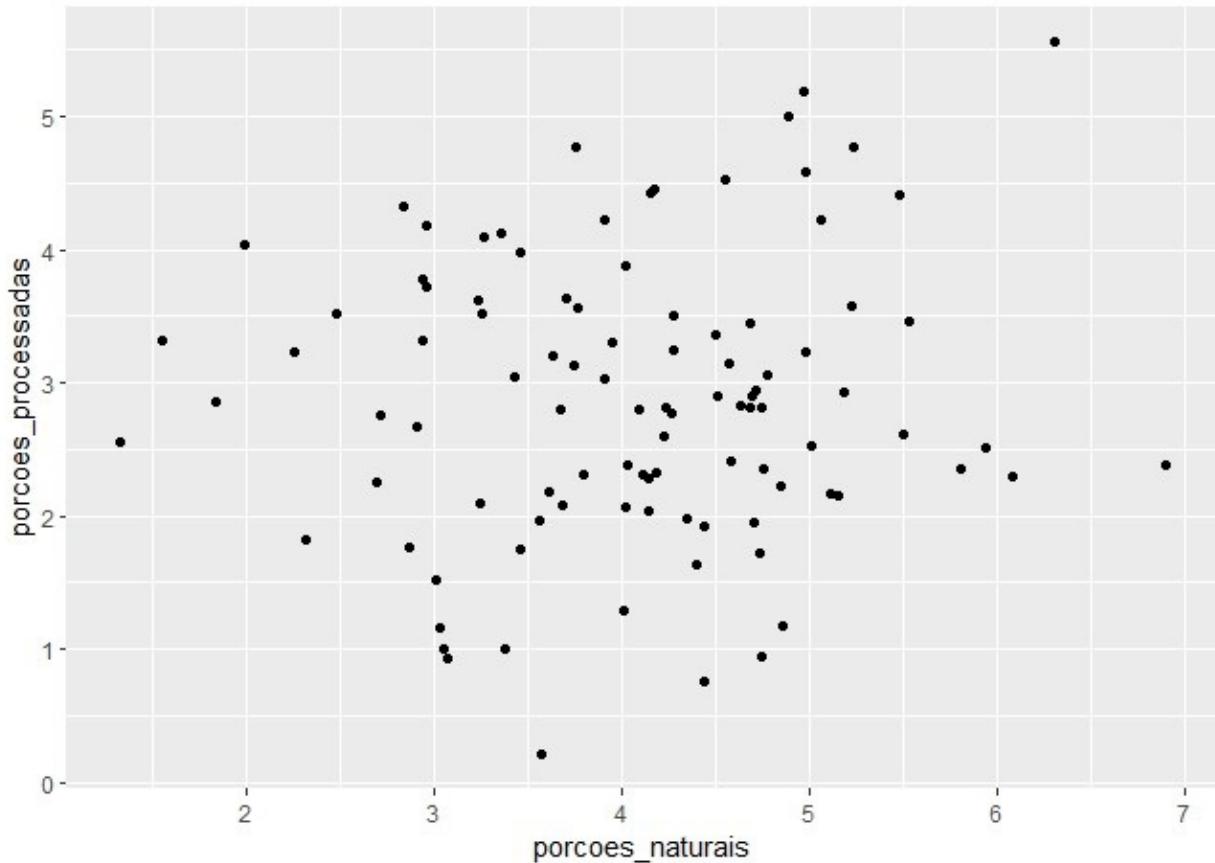


Gráfico relativo ao consumo de alimentos naturais e processados.

Elaboração: CEPED/UFSC (2023).

Você percebeu que no gráfico não há nenhum padrão de inclinação positivo ou negativo? O que se percebe é que não há nada que indique que quanto mais porções naturais se consome, menor o número de alimentos processados. Também não se consegue perceber a associação inversa que pudesse apontar que quanto maior o número de porções naturais, maior também o número de porções de alimentos processados.

O coeficiente de correlação que está associado ao gráfico é de 0,08454356, um valor muito próximo a zero. Isso descarta a possibilidade de haver qualquer associação entre as duas variáveis testadas.



DESTAQUE

Conforme Mukaka (2012 apud WIKIMEDIA FOUNDATION, 2022), existe a seguinte regra para classificar o nível de associação entre duas variáveis quantitativas distintas a depender do coeficiente de correlação:

- *0.9 para mais ou para menos indica uma correlação muito forte.*
- *0.7 a 0.9 positivo ou negativo indica uma correlação forte.*
- *0.5 a 0.7 positivo ou negativo indica uma correlação moderada.*
- *0.3 a 0.5 positivo ou negativo indica uma correlação fraca.*
- *0 a 0.3 positivo ou negativo indica uma correlação desprezível.*

Wikimedia Foundation (2022).

4.2 Usando R para Realizar Análise de Correlação

A linguagem R possibilita identificar rapidamente os valores de correlação entre duas variáveis distintas. Além disso, sempre é bom contar com os recursos do {ggplot2} para desenhar os gráficos de dispersão que ilustram melhor as análises. Acompanhe a videoaula deste tema!



Videoaula: [Correlação](#)

Os códigos para copiar, colar e testar em seu ambiente depois da videoaula são estes:

```
library(tidyverse)

#O endereço indicado logo abaixo contém um arquivo com dados sobre
#municípios, incluindo informações sobre gastos com saúde
arquivo<-      "https://raw.githubusercontent.com/fernandobarbalho/
enap_auto_instucional/main/data/dados_saude_municipio.csv"

dados_saude_municipio<-
  read_csv(file = arquivo)
```

```

#Os gráficos de correlação são feitos a partir de gráficos de pontos
#Nos eixos x e y colocamos as variáveis que queremos verificar
visualmente se há correlação
#No caso do gráfico abaixo queremos testar se há correlação entre
população e percentual gasto com saúde
dados_saude_municipio%>%
  ggplot()+
  geom_point(aes(x=populacao, y=perc))

#Como há pontos extremos muito elevados, é possível que a representação
do eixo x em escala logarítmica ajude a melhorar o gráfico

dados_saude_municipio%>%
  ggplot()+
  geom_point(aes(x=populacao, y=perc))+
  scale_x_log10()

#Para se calcular a correlação entre as duas variáveis usamos a
função cor
#Na função indicamos as duas variáveis que queremos checar a
correlação
#No gráfico fizemos a transformação logarítmica no eixo x que
corresponde à variável população
#O mais indicado é testar a correlação transformando os dados
de população usando logarítmico. Para isso usamos a função
logcor(log10(dados_saude_municipio$populacao),
dados_saude_municipio$perc)
cor(log10(dados_saude_municipio$populacao),
dados_saude_municipio$perc)

#Podemos também testar a correlação entre população e o valor
absoluto gasto em saúde

dados_saude_municipio%>%
  ggplot()+
  geom_point(aes(x=populacao, y=valor))

#O gráfico mostra a possibilidade de termos pontos muito extremos
tanto no eixo x como no y
#Precisaremos usar escala logarítmica nos dois eixos

```

```

dados_saude_municipio%>%
  ggplot()+
  geom_point(aes(x=populacao, y=valor))+
  scale_x_log10() +
  scale_y_log10()

#O cálculo da correlação admite a transformação em logaritmo para
as duas variáveis
cor(log10(dados_saude_municipio$populacao),
log10(dados_saude_municipio$valor))

```

Valores de correlação entre duas variáveis distintas.

Fonte: Barbalho (2020).

E algumas sugestões de prática que não poderiam ficar de fora.

Execute o *script* abaixo e verifique se a correlação entre população e percentual gasto em saúde nos municípios com menos de 20.000 habitantes é aproximadamente a mesma do cálculo que foi feito no script original quando se considerou todos os municípios do Brasil. Faça adaptações para usar escala logarítmica, caso julgue necessário.

```

municipios_pequenos<-
dados_saude_municipio%>%
  filter(populacao<20000)

municipios_pequenos%>%
  ggplot()+
  geom_point(aes(x=populacao, y=perc))

cor(municipios_pequenos$populacao, municipios_pequenos$perc)

```

Script.

Fonte: Barbalho (2020).

Agora é a hora de você testar seus conhecimentos. Para isso, acesse o exercício avaliativo disponível no ambiente virtual. Bons estudos!

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). **Cidades e Estados.** 2021. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>. Acesso em: 20 mar. 2023.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN.** GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

IHAKA, Ross. **The R Project:** a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo.** [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R.** RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status.** [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdadosus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Coeficiente de correlação de Pearson.** [Wikipédia]. 2022. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Estatística.** [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Logaritmo.** [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Logaritmo>. Acesso em: 20 mar. 2023.

Módulo

4 Análise de Dados na Administração Pública

Você se manteve firme e finalmente chegou ao último módulo. Agora conhecerá casos práticos de análise de dados da Administração Pública.

Na unidade 1 você entenderá o conceito de dados abertos governamentais. Na unidade 2, terá contato com as possibilidades de análise de dados de finanças públicas. Já na unidade 3, conhecerá um pouco do funcionamento do campo da saúde pública no Brasil. Por fim, na unidade 4, reconhecerá as possibilidades de análise de dados em educação pública. Então vamos adiante!

Unidade 1: Dados Abertos Governamentais

Objetivo de aprendizagem

Nessa unidade você irá reconhecer o conceito e as formas de acesso a dados abertos governamentais.

1.1 O que São Dados Abertos?

“ Dados abertos referem-se a dados livremente disponíveis que podem ser reutilizados sem imposição de restrições (BRAUNSCHWEIG et al., 2012; BRAZILIAN et al., 2012). Entre as fontes desses dados estão processos organizacionais, além de notícias, informações de mercado e dados de competidores (O'RIAN et al., 2012). Tratando especificamente de um governo, entende-se que dados abertos governamentais são “[...] qualquer dado produzido pelo setor público para qualquer pessoa para ser usado em qualquer propósito” (CRAVEIRO et al., 2013). ”

Oito princípios devem balizar as iniciativas de dados abertos governamentais (BRAUNSCHWEIG et al., 2012; CRAVEIRO et al., 2013). São eles:

- 1  Dados devem estar completos.
- 2  Dados devem ser coletados na fonte como maior nível possível de granularidade.
- 3  Dados devem estar disponíveis no menor tempo possível após a sua geração.
- 4  Dados devem estar disponíveis ao maior número possível de usuários.
- 5  Dados devem ser reconhecíveis e processáveis por máquinas.
- 6  Dados devem ser disponíveis a qualquer um sem a necessidade de registro.
- 7  Dados disponíveis em formato que nenhuma entidade possa arguir controle exclusivo.
- 8  Dados não podem estar submetidos a restrições trazidas por direito de cópia, patente ou práticas semelhantes.

Princípios de dados abertos governamentais.

Fonte: Freepik (2023). Elaboração: CEPED/UFSC (2023).

“Dados abertos são uma decorrência do movimento aberto que promove pautas como: acesso aberto, código aberto, software aberto, inovação aberta, modelagem aberta, conhecimento aberto, crowd sourcing, governo aberto, entre outros. Caracteriza-se pela utilização de recursos tecnológicos que viabilizam redes de colaboração envolvendo, entre outros atores, a sociedade civil organizada (BARBALHO, 2014). **”**

Vale destacar ainda que o campo de dados abertos no Brasil tem sido amplamente favorecido com o estabelecimento e amadurecimento da Lei de Acesso à Informação (LAI). Essa lei se aplica aos órgãos públicos de todos os poderes e de todas as esferas.

Especificamente para o caso de dados abertos vale destacar o seu artigo 8º que apresenta essa redação:



Art. 8º É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas.

(...)

§ 2º Para cumprimento do disposto no caput, os órgãos e entidades públicas deverão utilizar todos os meios e instrumentos legítimos de que dispuserem, sendo obrigatória a divulgação em sítios oficiais da rede mundial de computadores (internet).

§ 3º Os sítios de que trata o § 2º deverão, na forma de regulamento, atender, entre outros, aos seguintes requisitos:

(...)

II – possibilitar a gravação de relatórios em diversos formatos eletrônicos, inclusive abertos e não proprietários, tais como planilhas e texto, de modo a facilitar a análise das informações;

III – possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina. (BRASIL, 2011).



SAIBA MAIS

Para entender o impacto da LAI sobre dados abertos, vale a pena a leitura da wikiLAI, produzida pela ONG “Fiquem Sabendo”. Especificamente, recomenda-se a leitura dos seguintes verbetes indicados abaixo. Clique no verbete para seguir o link:

- [Acidentes aéreos](#)
- [Agenda de autoridades](#)
- [Armas de fogo](#)
- [Auxílio emergencial](#)
- [Cargos em comissão](#)
- [Pensões](#)
- [Preço da gasolina](#)
- [Salário de servidores públicos](#)

Agora que você já conhece o conceito de dados abertos e já tem uma primeira ideia de como as pessoas utilizam esse recurso, que tal identificar algumas fontes de dados abertos no Brasil?

1.2 Trabalhando com Portais de Dados Abertos

Há uma grande diversidade de portais de Internet que se especializam em divulgar dados abertos governamentais no Brasil. Conheça na videoaula a seguir os potenciais de um conjunto de dados disponibilizado pelo dados.gov.br.



Videoaula: [Portais de Dados Abertos no Brasil](#)

Logo abaixo está o código que foi apresentado a você na videoaula. Faça o tradicional copiar-colar e execute em seu ambiente RStudio.

```

library(tidyverse)

##Dados abertos de dados.gov.br

#url do dado
url_habitacao<- "http://sishab.mdr.gov.br/dados_abertos/_contratacoes_pcmv_pcva.csv"

#download do dado a partir do endereço da url
download.file(url = url_habitacao, destfile = "contratacoes_pcmv_pcva.csv", mode="wb")

#gera um dataframe a partir do arquivo csv baixado
contratacoes_pcmv_pcva <- read_delim("contratacoes_pcmv_pcva.csv",
delim = "|", escape_double = FALSE, locale = locale(encoding =
"LATIN1"),
trim_ws = TRUE) #sugestão: procure ler sobre encoding

contratacoes_pcmv_pcva%>%
group_by(txt_uf) %>%
summarise(
contratadas= sum(qtd uh_contratadas), #soma das quantidades de
unidades habitacionais contratadas
entregues = sum(qtd uh_entregues), #soma das quantidades de
unidades habitacionais entregues
perc_entregues = (entregues/contratadas)*100 #de unidades
habitacionais entregues
) %>%
arrange(desc(perc_entregues))

```

Código da videoaula.

Fonte: Barbalho (2023).

Agora experimente fazer o que segue:

Acrescente as duas linhas demonstradas no *script* que você vê aqui e procure identificar os pontos mais importantes sobre a distribuição dos contratos por estado. Procure checar, por exemplo, qual o estado que marca os percentuais acumulados acima de 50% e qual a posição desse estado.

Compare o percentual isolado do primeiro estado com o do último estado. Quais regiões do Brasil estão representadas nos cinco primeiros estados do ranking? Aqui as linhas a serem acrescentadas:

```
library(questionr)
questionr::freq(contratacoes_pcmv_pcva$txt_uf, cum=TRUE, sort = "dec")
```

Script.

Fonte: Barbalho (2020).

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BRASIL. **Lei nº 12.527, de 18 de novembro de 2011.** Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília, DF: Presidência da República, 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 22 mar. 2023.

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...].** New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...],** Wilmington: IARIA Press, 2013.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

Unidade 2 - Análise de Dados de Finanças Públicas

Objetivo de aprendizagem

Nessa unidade você reconhecerá as possibilidades de análise de dados de finanças públicas.

2.1 O Campo de Finanças Públicas e a Análise de Dados

As finanças públicas se referem às atividades de governo relacionadas ao cuidado com orçamento, contabilidade pública, política monetária e equilíbrio fiscal, entre outras atribuições. Destacam-se nesse campo instituições como:

- Secretaria de Orçamento Federal (SOF), responsável principalmente pelo planejamento orçamentário do governo federal;
- Secretaria do Tesouro Nacional (STN), órgão que cuida principalmente do equilíbrio entre receitas e despesas e da contabilidade pública;
- Banco Central (BACEN), instituição que lidera as políticas monetárias no Brasil;
- Receita Federal do Brasil (RFB), órgão que é responsável pela maior parte do sistema arrecadatório do governo federal.



SAIBA MAIS

No âmbito das esferas estaduais e municipais se encontram principalmente as Secretarias de Fazenda e de Planejamento, que cuidam das partes de gestão orçamentária e equilíbrio fiscal que cabem aos entes subnacionais.

Você encontra dados abertos importantes e muito bem catalogados em algumas dessas instituições. Veja abaixo uma lista de portais onde se pode consumir dados de finanças públicas.

- [Tesouro Transparente](#): Portal da STN que oferece diversas soluções de transparência relacionadas aos temas trabalhados pela instituição. Entre essas soluções está o repositório de dados abertos.

- **Dados abertos do BACEN:** Catálogo de dados abertos do Banco Central.
- **Receita Data:** Portal de dados abertos da Receita Federal do Brasil.

Tal como o dados.gov.br, todos esses portais são estruturados de forma a facilitar que sejam baixados dados para posterior tratamento em ferramentas analíticas.

A seguir, veja um exemplo que você pode copiar e colar para o seu ambiente RStudio. Aproveite que o código está muito comentado para buscar entender o que ocorre em cada linha.

Uma informação importante: os dados vêm do portal [Tesouro Transparente](https://www.tesourotransparente.gov.br/).

```
#Dados abertos do tesouro transparente
install.packages("janitor") #instale o pacote para melhorar a
legibilidade dos nomes da tabela

library(tidyverse)
library(janitor)

#url do dado
url_estoque_dpf<- "https://www.tesourotransparente.gov.br/ckan/
dataset/0998f610-bc25-4ce3-b32c-a873447500c2/resource/b6280ed3-
ef7e-4569-954a-bded97c2c8a1/download/EstoqueDPF.csv"

#lê o arquivo diretamente da url, observe que aqui não precisamos
fazer o download do dado
estoque_dpf<- read_csv2(url_estoque_dpf)

#veja os nomes das colunas do dataframe
names(estoque_dpf)

#A função clean_names do pacote janitor deixa o nome das variáveis
mais legíveis
estoque_dpf<- janitor::clean_names(estoque_dpf)

#veja como ficam os nomes das colunas do dataframe após a limpeza
names(estoque_dpf)
```

```

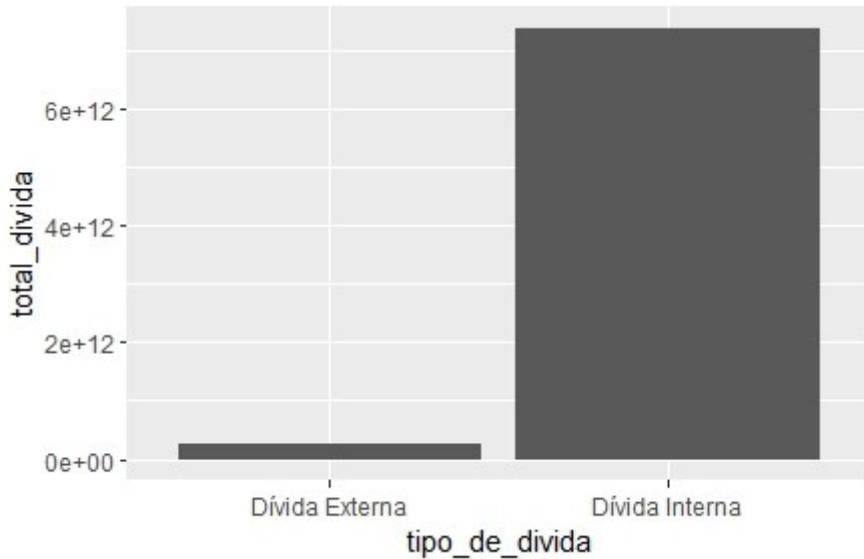
#gráfico a partir do agrupamento por tipo de dívida
estoque_dpf %>%
  filter(mes_do_estoque=="12/2021") %>% #filtra para permanecer apenas
dados de 12/2021
  group_by(tipo_de_divida) %>% #agrupa por tipo de dívida (interna/
externa)
  summarise(
    total_divida = sum(valor_do_estoque) #calcula o total da dívida
por tipo
  ) %>%
  ggplot()+
  geom_col(aes(x=tipo_de_divida,y=total_divida))

```

Código.

Fonte: Barbalho (2023).

O código deste exemplo trabalha com o estoque, que é o valor em reais da dívida brasileira. Se der tudo certo na execução do seu código, você verá uma figura que compara o valor da dívida interna com o da dívida externa em dezembro de 2021. O gráfico deve ser esse logo abaixo.



Estoque, valor em reais, da dívida brasileira.

Fonte: Brasil (2023a).

Como pode ser visto, em dezembro de 2021 a dívida interna brasileira era muito maior do que a dívida externa.



SAIBA MAIS

Além dos links já indicados, acesse outras sugestões de dados e temas sobre finanças públicas a partir dos seguintes verbetes da Wikilai. Clique nos links para ler os conteúdos.

- [LRF – lei de Responsabilidade Fiscal](#);
- [Dívidas do Fies](#); e
- [Dívidas de políticos](#).

2.2 Análises de Dados sobre Receitas e Despesas Primárias do Governo Central Brasileiro através do Pacote RTN

Você já sabe que uma das principais atribuições da STN é buscar o equilíbrio entre as receitas e despesas do governo federal brasileiro. Uma das medidas que indicam o sucesso desse equilíbrio é o chamado resultado primário. Esse indicador é calculado pela diferença entre as receitas primárias, aquelas que não sejam derivadas de recebimentos de juros, e as despesas primárias, aquelas que não estejam relacionadas a pagamentos de dívidas, incluindo os juros. Se as receitas primárias forem maiores do que as despesas primárias, teremos *superávit* primário. Se ocorrer o contrário, será registrado um *déficit* primário.

A STN desenvolveu um pacote que permite manipular com facilidade os dados relacionados ao Relatório do Tesouro Nacional (RTN). O RTN é o instrumento de transparência que mostra as receitas primárias, as despesas primárias e o resultado primário, possuindo uma série histórica de dados que começa em janeiro de 1997 e é atualizada mês a mês. Para compreender melhor o relatório e os seus conceitos, e mais do que isso, ter uma ideia dos dados que podem ser analisados, é recomendável a leitura do texto dinâmico indicado [aqui](#).

E agora, aprenda como trabalhar com esses dados a partir do uso do pacote {rtn}. Uma videoaula vai ajudar você nessa missão.



Videoaula: [Uso do Pacote RTN](#)

Que tal praticar com o pacote {rtn}? Abaixo está o código usado na videoaula para você levar para o seu RStudio e fazer experiências. Sugestão: veja novamente a videoaula para buscar inspirações para novos gráficos de séries temporais.

Antes de executar o código abaixo, será necessário instalar o executável Rtools. Para isso, clique no link [disponível aqui](#).

O navegador vai fazer o download do arquivo. Após o download completo, clique no arquivo e em seguida faça a instalação, sempre pressionando o botão next.

```
#A instalação do pacote RTN requer a execução das duas linhas
seguintes
install.packages("devtools")
devtools::install_github("tchiluanda/rtn")

#Ao executar a linha acima é possível que o console espere uma
resposta sua perguntando se deseja instalar outros pacotes que podem
estar desatualizados. Nesse momento, pressione enter no console
para que a instalação do pacote continue.

library(rtn)
library(tidyverse)

#Busca todas as contas
rtn::get_full_account_name()

#conta de despesa de benefícios previdenciários

despesa_beneficios_previdenciarios<- "4.1 Benefícios Previdenciários"

###Atenção é possível que a linha acima precise ser modificada caso
você perceba que a execução do
#código não retornou o que foi mostrado no vídeo. Uma das explicações
para essa situação é que o nome da conta pode
#eventualmente estar um pouco diferente, às vezes basta um espaço
a mais para a execução ser comprometida.
#Na execução da linha 5 aparece todos os nomes de todas as contas,
você pode
#copiar e colar o nome da conta benefícios previdenciários e
subsituir na linha 10.
```

```

#Busca quanto se gastou em benefícios previdenciários nos meses de
dezembro, mês 12
get_account_data_by_month(despesa_beneficios_previdenciarios,month
= 12)

#Busca quanto se gastou em benefícios previdenciários em todos os
meses: 1:12
dados_previdenciarios<-get_account_data_by_month(despesa_
beneficos_previdenciarios,month = 1:12)

#gera gráfico da série temporal de dados de benefícios previdenciários
dados_previdenciarios%>%
  plot_rtn_series()

#ou ainda
get_account_data_by_month(despesa_beneficos_previdenciarios,month
= 1:12) %>%
  plot_rtn_series()

#Busca novamente todas as contas
get_full_account_name()

#Códigos de contas associadas a despesas obrigatórias com controle
de fluxo
despesas_obrigatorias<-           c("4.4.1.1",           "4.4.1.2",
"4.4.1.3","4.4.1.4","4.4.1.5" )

#gera valores acumulados em 12 meses para as contas selecionadas
get_12_month_accumulated_account_data_by_month(despesas_
obrigatorias,
                                               month = c(1:12),
#indica que quer o valor acumulado para todos os 12 meses do ano
                                               match_required
= FALSE)%>% #indica que vai fazer o match sem precisar do nome
completo da conta
  plot_rtn_series() #gera o gráfico de linha

```

Código pacote {rtn}.

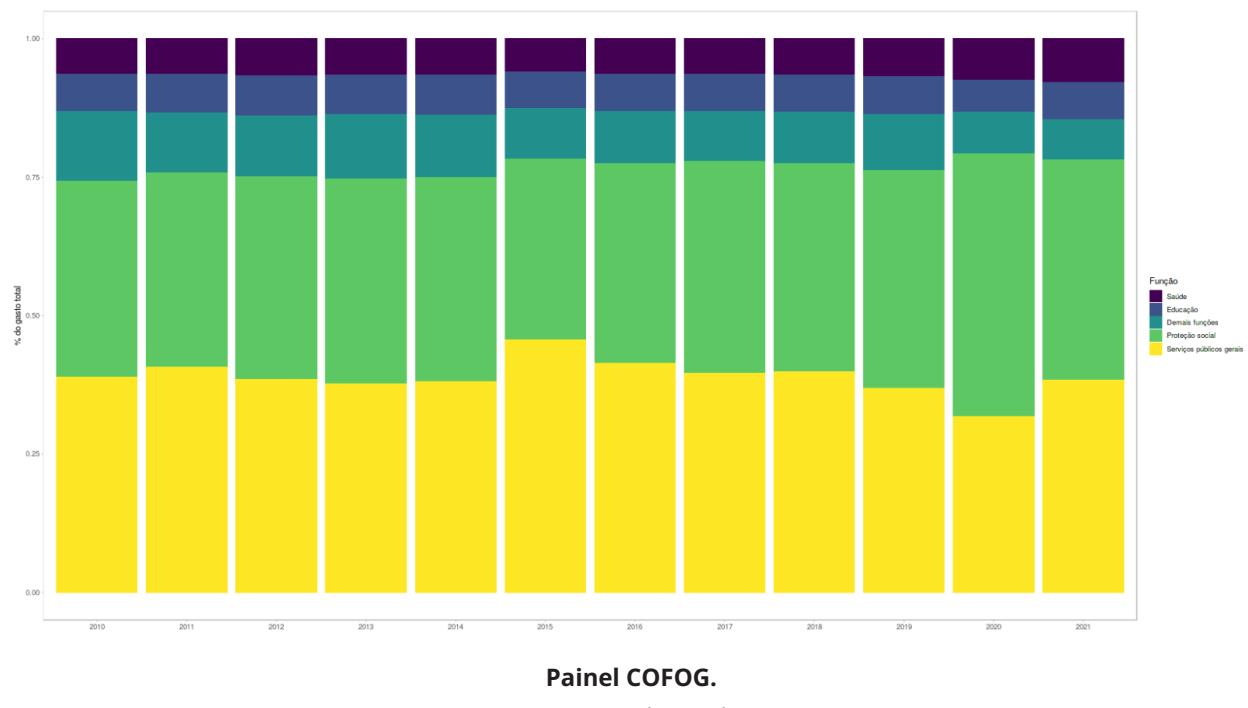
Fonte: Barbalho (2023).

É possível que ao executar o código acima você encontre problema ao instalar o pacote.

2.3 Pacote Rcofog - Consumo de Dados sobre Funções Típicas de Governo

O governo executa seu orçamento com gastos que buscam cumprir com suas funções. Dessa forma, muitas vezes é importante saber quanto se gasta em temas importantes como educação, saúde, justiça e transporte. Existem alguns protocolos que indicam como esses gastos são medidos. Um dos mais importantes é o [COFOG](#) que é proposto por OCDE e ONU.

No Brasil, a SOF e a STN são as instituições responsáveis pela elaboração do demonstrativo e da abertura dos dados relacionados ao COFOG. Além dos dados abertos disponíveis [aqui](#), a STN também disponibiliza um pacote R que permite emular as principais funcionalidades de um dashboard que acompanha a evolução das despesas a partir de 2010. Uma ilustração do que pode ser analisado a partir do COFOG pode ser observada na figura abaixo.



No gráfico, é usado o Painel do COFOG para analisar a evolução da distribuição dos gastos entre as principais funções de governo a partir de 2010.



SAIBA MAIS

Para conhecer algumas possibilidades de análise usando os dados do COFOG e o pacote {rcofog}, é importante a leitura do texto, disponível [aqui](#), que apresenta o painel do COFOG.

Agora, assista à videoaula que explica como usar o pacote {rcofog}.



Videoaula: [Pacote Rcofog](#)

A seguir vem o código usado na videoaula.

```
#Instala o pacote Rcofog
devtools::install_github("tchiluanda/Rcofog")

library(Rcofog)
library(tidyverse)

#Gera um gráfico de fluxo entre funções e sub-funções de governo para
#o ano de 2020
Rcofog::dataExpenseFlow(year=2020)%>%
  Rcofog::graphExpenseFlow()

#Gera um gráfico de série temporal para comparar saúde, educação e
#defesa
funcoes<- c("Saúde", "Educação", "Defesa")

#Antes vamos ver apenas os dados
Rcofog::dataTimeSeries(sel_function = funcoes)

#Agora vamos ver o gráfico também
Rcofog::dataTimeSeries(sel_function = funcoes) %>%
  Rcofog::graphTimeSeries()
```

Código da videoaula.

Fonte: Barbalho (2023).

Você chegou ao final desta unidade de estudo. Caso ainda tenha dúvidas, reveja o conteúdo e se aprofunde nos temas propostos.

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. **Estoque da Dívida Pública Federal.** Portal Tesouro Nacional Transparente. 2023a. Disponível em: <https://www.tesourotransparente.gov.br/ckan/dataset/estoque-da-dívida-publica-federal>. Acesso em: 22 mar. 2023.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). **Cidades e Estados.** 2021. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>. Acesso em: 20 mar. 2023.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRASIL. Secretaria do Tesouro Nacional (STN). **Análises de séries temporais do COFOG.** [Painel COFOG]. 2023b. Disponível em: https://painel-cofog.tesouro.gov.br/dashboard_cofog.Rmd. Acesso em: 24 mar. 2023.

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...].** New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...],** Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN.** GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

IHAKA, Ross. **The R Project**: a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stats.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo**. [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R**. RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

WIKIMEDIA FOUNDATION. **Coeficiente de correlação de Pearson**. [Wikipédia]. 2022. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Logaritmo**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Logaritmo>. Acesso em: 20 mar. 2023.

Unidade 3: Análise de Dados de Saúde Pública

Objetivo de aprendizagem

Nessa unidade você reconhecerá as possibilidades de análise de dados de saúde pública.

3.1 O Campo de Saúde Pública e a Análise de Dados



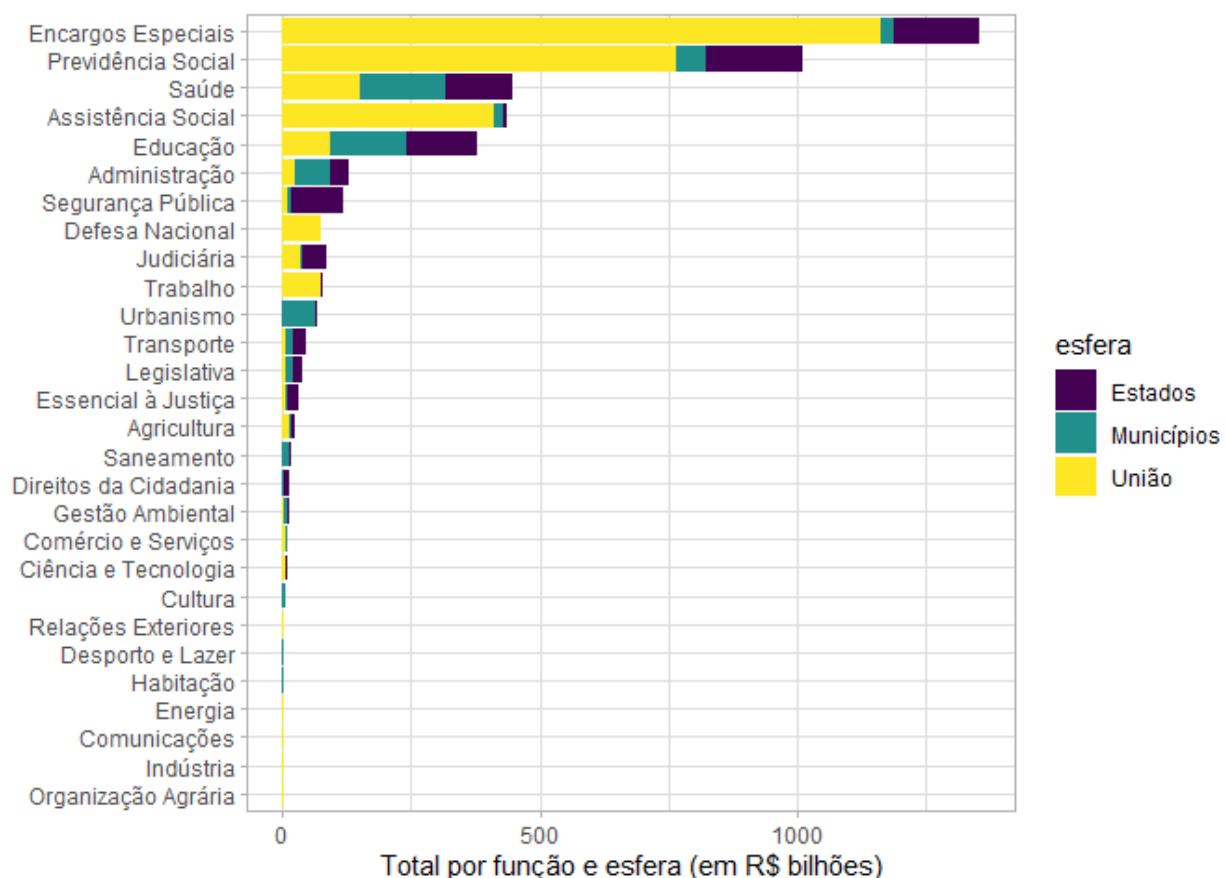
Saúde pública e a análise de dados.

Fonte: Freepik (2023).

A Constituição Federal traz direcionamentos importantes sobre a organização dos serviços em saúde pública no Brasil. De acordo com a carta magna, além de ser uma competência comum de todos os entes, o serviço de atendimento à saúde da população compete aos municípios, com a cooperação técnica e financeira da União e do Estado. Na esfera legal, a lei nº 8.080/90 regula o SUS (Sistema Único de Saúde) e atribui à União as funções de gestão e organização do sistema, além da possibilidade, em casos de calamidade grave, de executar diretamente ações de saúde (BRASIL, 1990). Aos estados cabe prestar apoio técnico e financeiro aos municípios, onde os serviços de saúde são de fato executados.

Um outro fator importante que está na constituição e que reflete sobre os serviços de saúde é a destinação mínima de recursos para a educação nos três níveis federativos: União, Estado e Município.

Isso ajuda a garantir que os gastos com saúde estejam sempre entre os principais destinos dos orçamentos de governo. Veja a figura a seguir com informações sobre o ano de 2020.

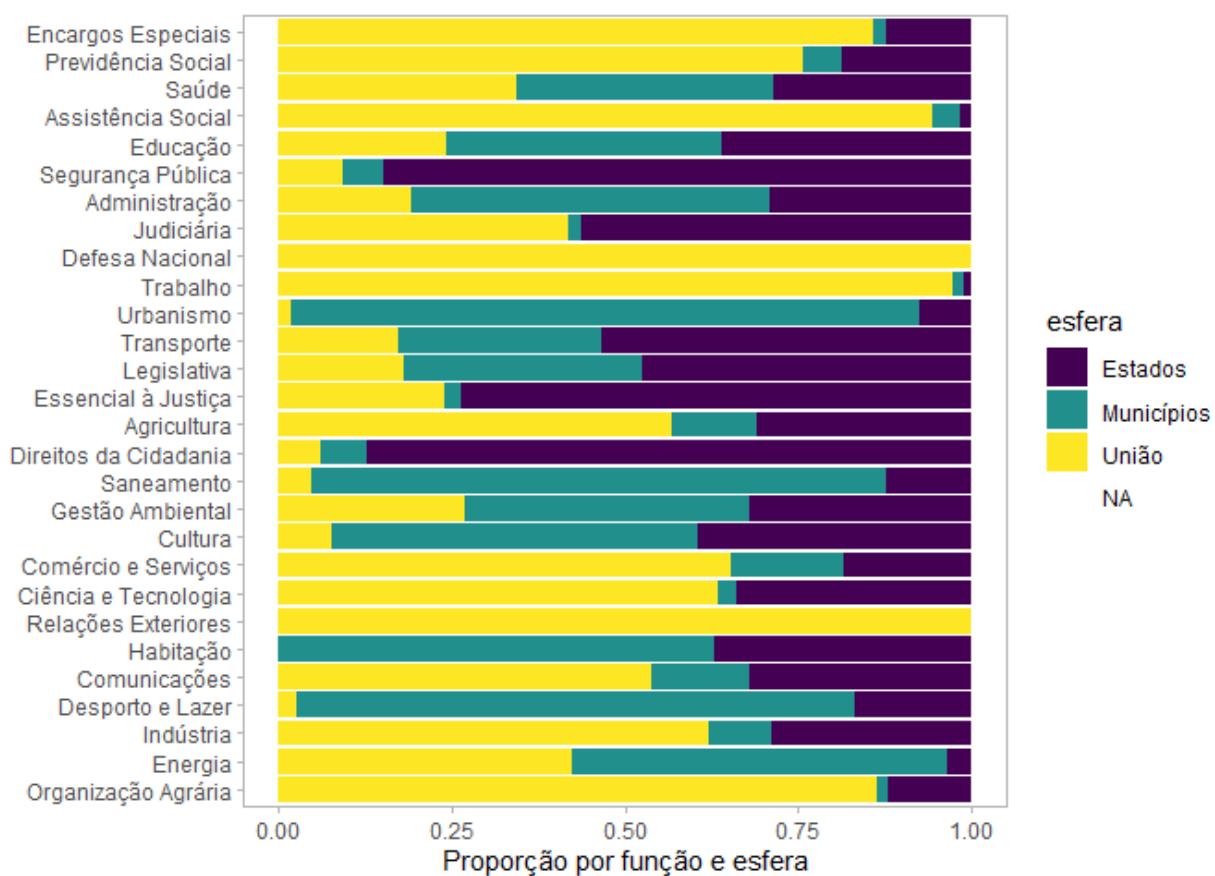


Análise de dados do governo (gráfico 1).

Fonte: Brasil (2020). Elaboração: Barbalho (2020).

Pelo gráfico você consegue perceber que os gastos com Saúde assumiram a terceira posição em 2020, ficando atrás apenas de Encargos Sociais e Previdência social? É possível perceber também que há uma aparente distribuição equilibrada dos gastos entre as três esferas de governo.

Agora, neste outro, o gráfico consegue revelar essa situação de uma forma mais clara. Observe!

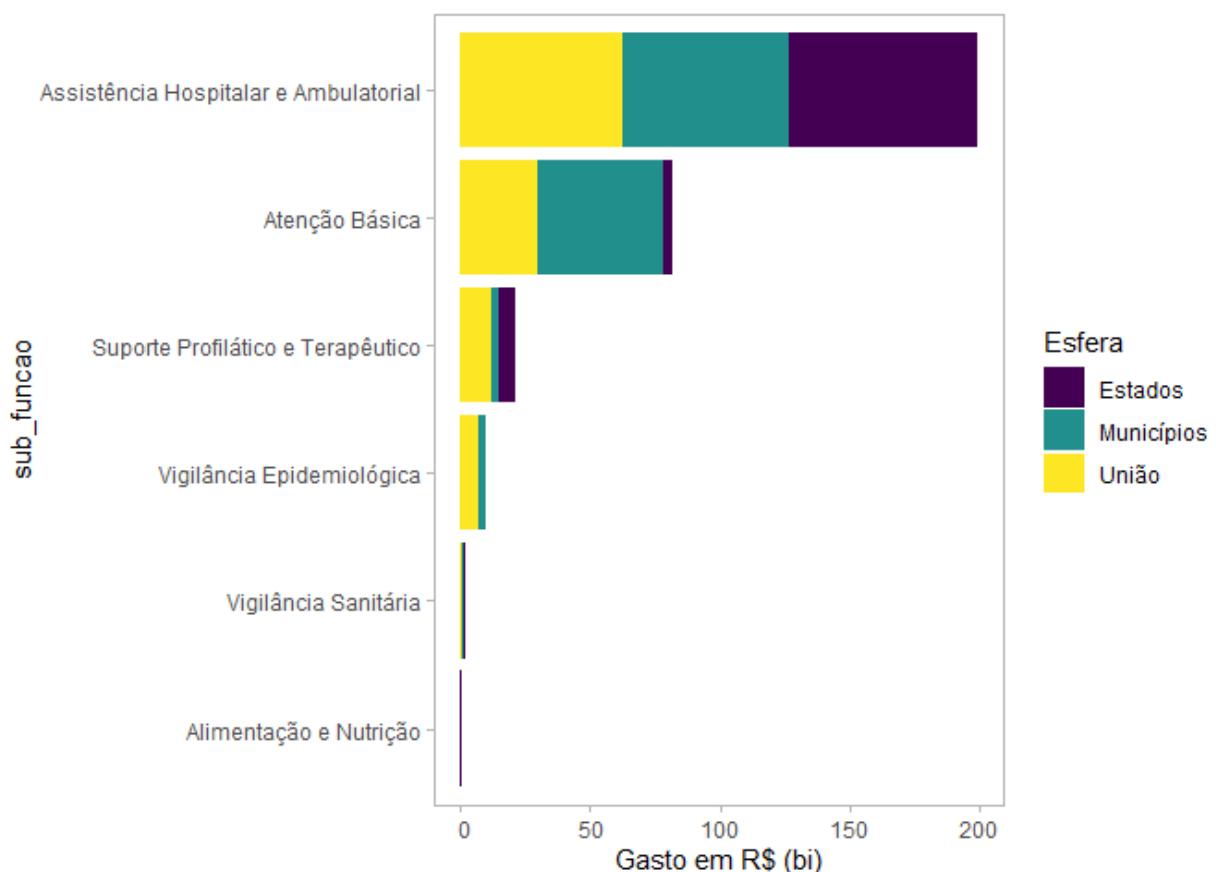


Análise de dados do governo (gráfico 2).

Fonte: Brasil (2020). Elaboração: Barbalho (2020).

No gráfico em que o foco é na distribuição proporcional, percebe-se que a Saúde junto com a Educação são as duas funções de governo mais equilibradas entre as três esferas. Percebe-se facilmente que em várias das outras funções há uma concentração em apenas uma esfera, notadamente Relações Exteriores, Defesa Nacional e Urbanismo, enquanto em outras situações uma das esferas tem participação bastante reduzida, como é o caso de Habitação e Judiciária.

É possível ainda entender como se organizam os gastos de saúde a partir das subfunções, ou especialidades, dos gastos. Veja no gráfico.



Análise de dados do governo (gráfico 3).

Fonte: Brasil (2020). Elaboração: Barbalho (2020).

Como pode ser percebido, a Assistência Hospitalar e Ambulatorial reúne o maior gasto por subfunção quando analisado o conjunto dos entes federativos. Percebe-se ainda que há uma distribuição igualitária desses gastos entre os três níveis.

A Atenção Básica é a subfunção que vem em seguida. Aqui já se vê que os estados são responsáveis, proporcionalmente, por uma parcela bem menor e que os municípios respondem por uma porção maior do que a União. Em tempos de pandemia os gastos com Vigilância Epidemiológica ficaram em quarto lugar no ranking, sendo o gasto dominado pela União.



SAIBA MAIS

Que tal conhecer mais possibilidades de análises de dados sobre saúde? Então, leia o material disponível com o contexto de alguns temas a partir dos verbetes da WikiLai:

- [Saúde indígena](#)
- [Covid-19 nas escolas](#)

Para fechar esse tour pelo campo da saúde pública é importante conhecer como se estrutura o Sistema Único de Saúde. Para tanto, acesse o endereço abaixo que leva à página do Ministério da Saúde que traz explicações sobre o tema.

- [Sistema Único de Saúde \(SUS\): estrutura, princípios e como funciona](#)

No próximo tópico você entenderá como a estrutura do SUS e os gastos federativos na saúde vão se refletir, na prática, nos dados sobre estabelecimentos de saúde, atendimentos hospitalares, equipamentos hospitalares e até mesmo óbitos. Avante!

3.2 Pacote Microdatasus (Consumo de Dados Relacionados ao Sistema Único de Saúde)

Os serviços de saúde pública são realizados nos hospitais e outras instalações de saúde através de atendimentos, consultas, exames, vacinações e internações. Os resultados são as altas hospitalares, os custos dos procedimentos e, infelizmente, os óbitos.

Dentro da estrutura do SUS existe o Datasus, que é o Departamento de Informática do sistema. Entre os produtos desenvolvidos pelo órgão se destacam os conjuntos de dados que serão explorados aqui e que trazem o diagnóstico do SUS.

Existem diversas maneiras de se trabalhar com os dados do Datasus. Para o analista que usa R, a forma mais fácil é usar o pacote {microdatasus} desenvolvido por Saldanha et al. (2019). A videoaula a seguir mostra algumas formas de usar o pacote possibilitando fazer análises importantes sobre a evolução do quadro da saúde pública no Brasil. Acompanhe!



Videoaula: [Pacote Microdatasus](#)

Veja, copie e cole o script a seguir com as experiências mostradas na videoaula. Depois de executar linha a linha, procure fazer variações no código em busca de outras análises com os dados disponíveis. Vamos lá?!

```

install.packages("remotes")
remotes::install_github("rfsaldanha/microdatasus")

#Ao executar a linha acima é possível que o console espere uma
#resposta sua perguntando se deseja instalar outros pacotes
#que podem estar desatualizados. Nesse momento pressione enter
#no console para que a instalação do pacote continue.

library(microdatasus)
library(tidyverse)

ano_inicio<- 2020
ano_fim <- 2020
mes_inicio<-12
mes_fim<-12
estado<- "GO"

#Traz dados sobre internações hospitalares: SIH-RD
resultado<-microdatasus::fetch_datasus(year_start = ano_inicio,
year_end = ano_fim,
uf = estado,
month_start = mes_inicio,
month_end = mes_fim,
information_system = "SIH-RD")

dados_internacoes<- microdatasus::process_sihs(resultado)

dados_internacoes%>%
filter(COD_IDADE=="Anos")%>% #filtrar para excluir bebês que ainda
não completaram um ano
mutate(IDADE = as.numeric(IDADE))%>% #converte idade de variável
categórica para numérica
ggplot() +
geom_boxplot(aes(x=SEXO, y=IDADE)) +
scale_y_continuous(breaks = seq(0,100,10))

```



SAIBA MAIS

A maioria dos pacotes utilizados no curso seguem o roteiro de instalação via a função `install.packages()`. Essa regra geral acontece também no mundo real. A maior parte dos pacotes mais importantes estarão disponíveis no repositório cran que é o que a função `intall.packages` referência quando é acionada.

No nosso curso o pacote `{rtn}` foi instalado usando a função `install_github` que está presente no pacote `{remotes}`. Nós fazemos isso porque o pacote `{rtn}` não está disponível no cran.

Já o pacote `{microdatasus}` tem uma peculiaridade. No vídeo é possível ver que instalamos o pacote usando a função `install.packages`, porém no script fazemos a instalação usando a função `install_github`. Isso ocorreu porque no intervalo entre a confecção do vídeo e a liberação do curso, o pacote `{microdatasus}` saiu do cran. Dessa forma tivemos que rever o script do material em texto e alteramos a instalação para o uso da função `install_github`. Essas situações podem ocorrer com alguma frequência. Inclusive alguns pacotes podem sofrer alterações que impeçam a plena reproduzibilidade dos exemplos dados no curso. Esses fatos não devem ser considerados como impeditivos para o aprendizado e sim etapas a mais que eventualmente precisam ser consideradas nos processos de análises de dados.

Referências

BRASIL. **Lei nº 8.080, de 19 de setembro de 1990.** Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências. Brasília, DF: Presidência da República, 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8080.htm. Acesso em: 22 mar. 2023.

BRASIL. Secretaria do Tesouro Nacional (STN). **Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro (Siconfi).** 2020. Disponível em: <https://siconfi.tesouro.gov.br/siconfi/index.jsf>. Acesso em: 24 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdadosus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

Unidade 4: Análise de Dados de Educação Pública

Objetivo de aprendizagem

Nessa unidade você reconhecerá as possibilidades de análise de dados de educação pública.

4.1 O Campo de Educação Pública e a Análise de Dados

Assim como em relação a políticas públicas de saúde, a Constituição Federal também procurar trazer direcionamentos importantes sobre a estrutura da educação pública no Brasil. A seção I do capítulo III do texto constitucional trata especificamente sobre a Educação e ali se pactua a educação como direito de todos e indica a responsabilização de autoridade competente que não ofertar ensino obrigatório dentro do que é previsto no artigo 208 (BRASIL, 1988).

Sobre as competências de cada esfera da federação, a fundação Todos pela Educação destaca que a constituição traz algumas orientações, porém não faz isso de forma muito clara “A carta define que o município cuida da Educação Infantil e também do Ensino Fundamental 1; o Ensino Médio é prioridade do governo estadual e do Distrito Federal, mas eles também gerem o Ensino Fundamental 2. A União, por sua vez, fica com função de coordenação financeira e técnica dessa orquestra, ao mesmo tempo em que conduz as universidades federais.” ([TODOS PELA EDUCAÇÃO](#), 2018).

A figura que você verá a seguir ajuda a evidenciar o papel da União nos gastos com educação.

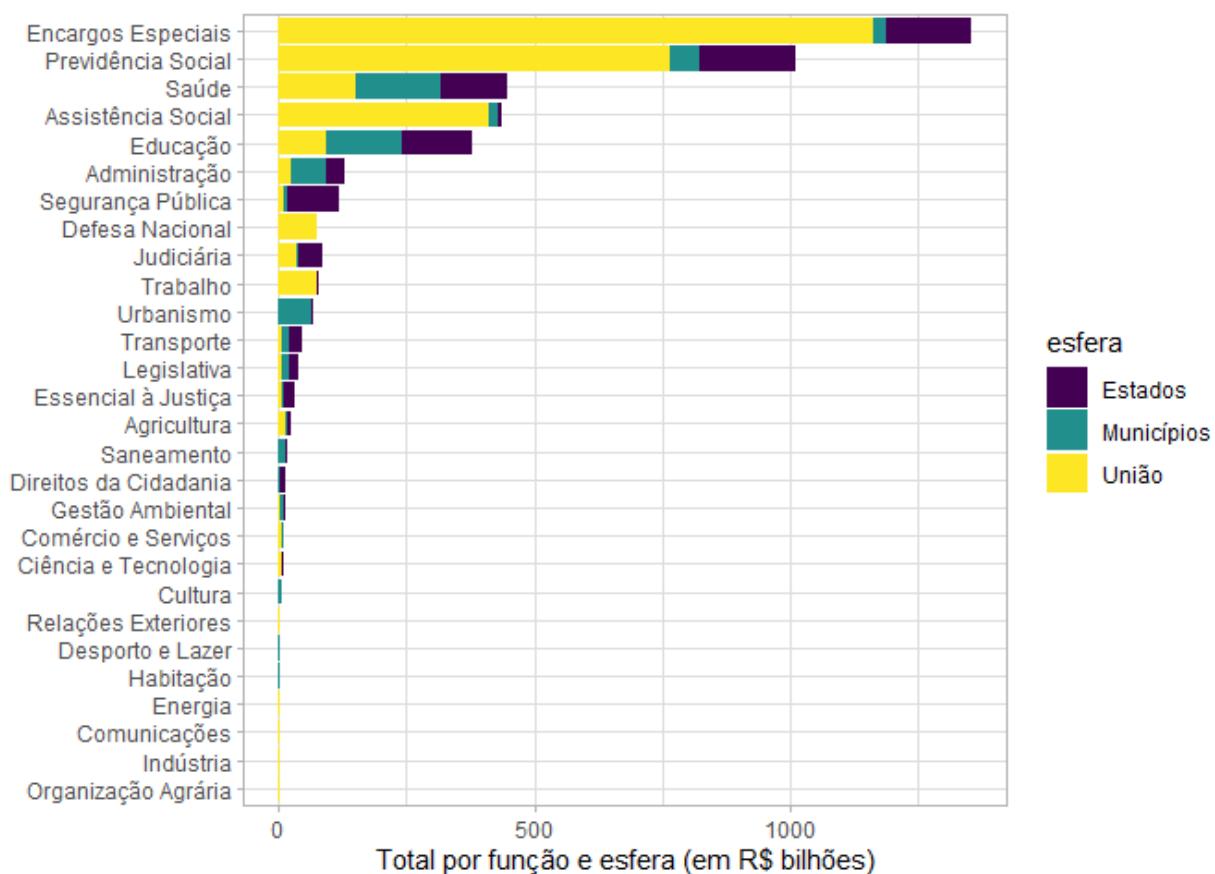


O papel da União nos gastos com educação.

Fonte: Painel COFOG (2018).

Perceba que os gastos que a União tem com **Educação Infantil e Ensino Fundamental I**, bem como os gastos em **Ensino Fundamental II** e Médio, são principalmente transferências. Ou seja, quem efetivamente executa os gastos na ponta das escolas são outras esferas, os estados ou municípios. O papel da União para esses níveis escolares é basicamente de coordenação financeira e/ou técnica. Já quando se observa a remuneração de empregados, vê-se que o Ensino Superior representa a maior parte dos gastos. Isso é um indício da responsabilização da condução de universidades e outras instituições do Ensino Superior.

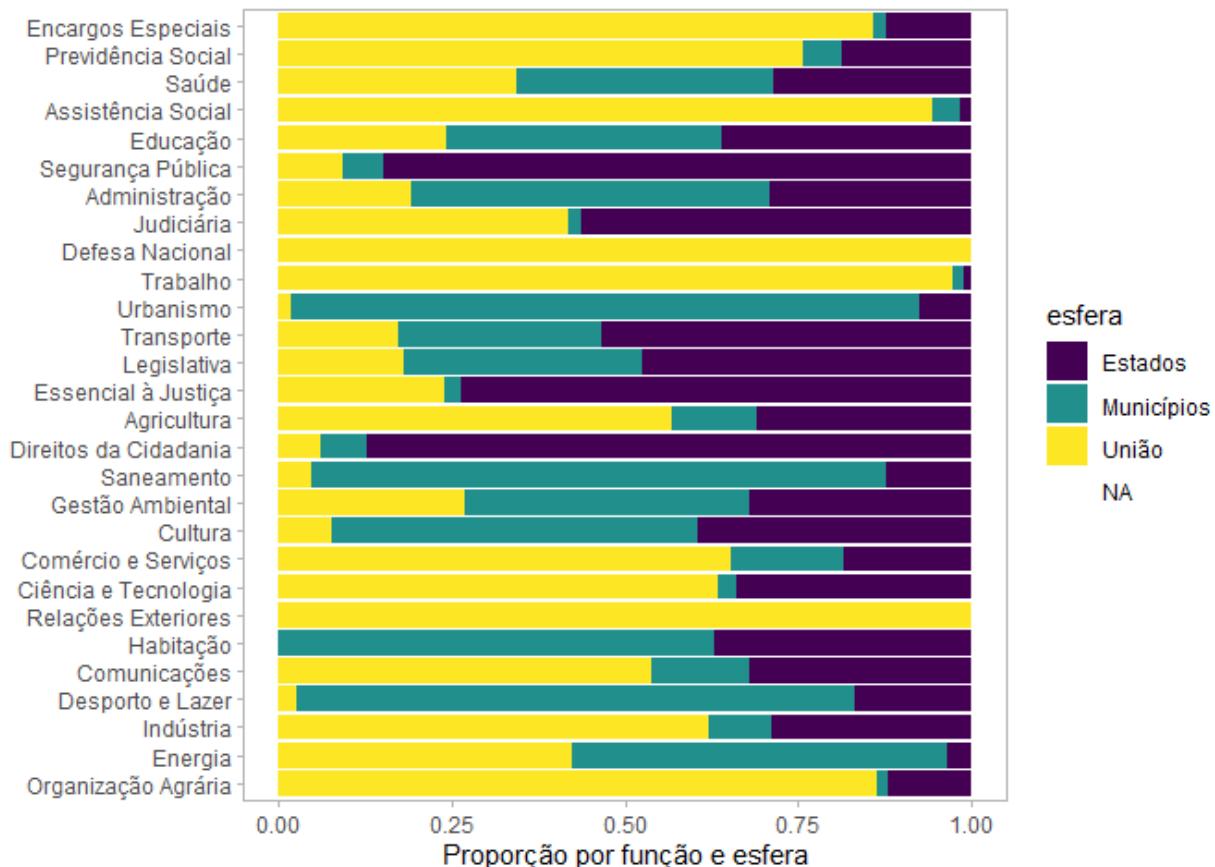
É importante ressaltar ainda que, tal como determinado para a saúde pública, há também mínimos constitucionais de gastos destinados à educação. E novamente recorremos às figuras que mostram as divisões dos gastos entre as funções de governo para evidenciar que há uma equidade nos gastos entre as três esferas e também há um destaque da Educação entre as funções mais financiadas pelo orçamento público. Veja!



Divisões dos gastos entre as funções de governo.

Fonte: Brasil (2020). Elaboração: Barbalho (2020).

Nesta figura é possível ver que a Educação foi a quinta função em gastos totais de todas as esferas de governo para o ano de 2020.



Comparando a proporção dos gastos.

Fonte: Brasil (2020). Elaboração: Barbalho (2020).

Quando se compara a proporção de gastos, vê-se que há um certo equilíbrio entre as três esferas, com uma participação um pouco menor da União no total dos gastos.



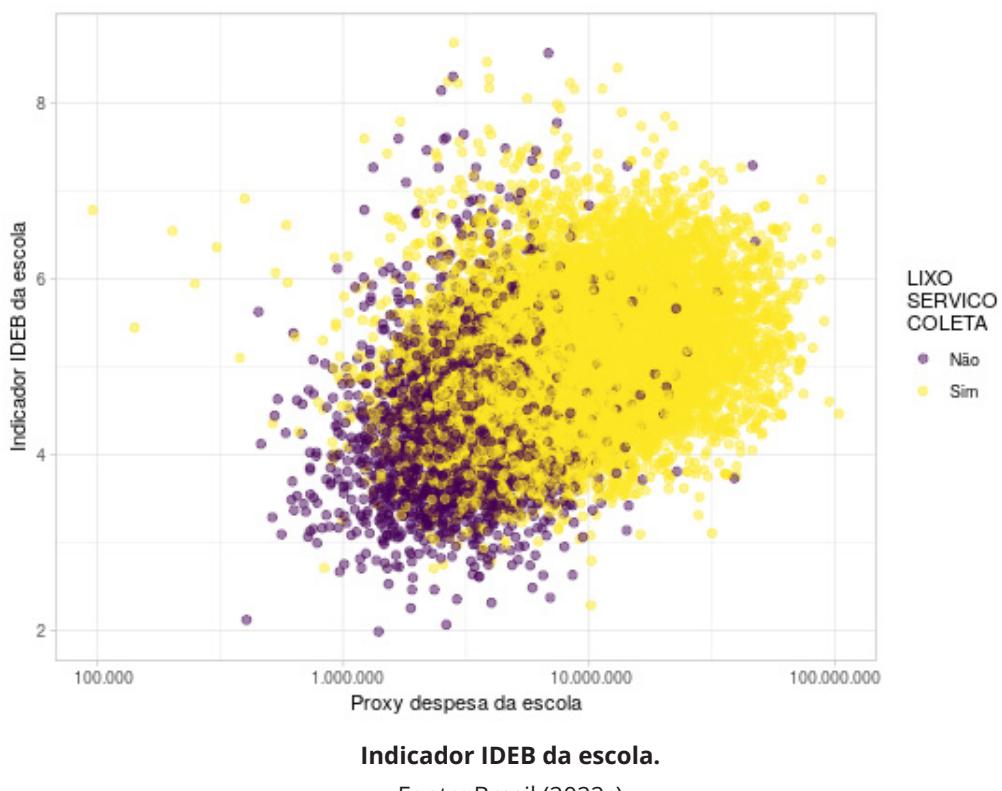
SAIBA MAIS

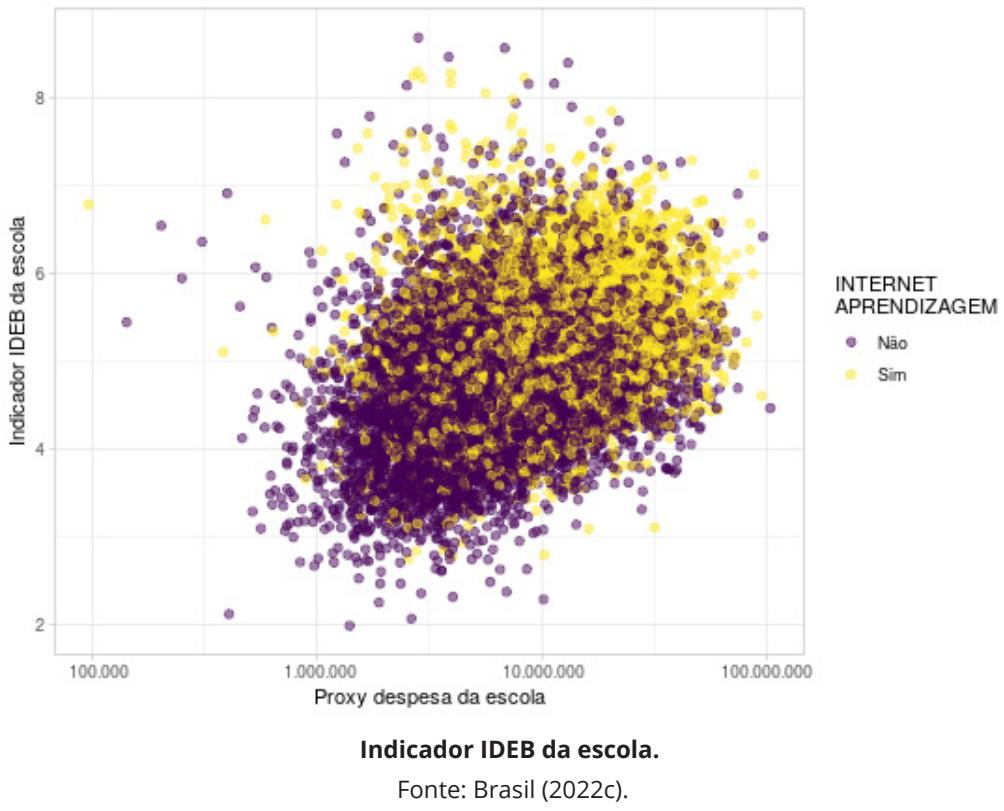
Para conhecer melhor os regramentos sobre a educação no Brasil é importante tomar conhecimentos sobre a Lei de Diretrizes e Bases (LDB). Veja aqui: [Leis de diretrizes e bases da educação - Comentários](#)

Tal como nos outros tópicos, é possível identificar diversas oportunidades de análises de dados sobre educação a partir dos verbetes da WikiLai. Clique nos links:

- [Censo escolar](#)
- [Merenda escolar](#)
- [Transporte escolar](#)

Dentre os três verbetes indicados no Saiba Mais, vale destacar uma análise desenvolvida tendo como referência o Censo Escolar. Entre os diversos dados disponibilizados nesse censo, existem os que se referem à infraestrutura das escolas. É possível, a partir daí, buscar inferir se existe associação entre equipamentos escolares e desempenho escolar. Veja dois gráficos que ilustram essa possibilidade.





Os dois gráficos buscam associar as notas do IDEB a variáveis de infraestrutura de escolas. No caso, o que está em foco são as notas obtidas pelos alunos de Ensino Fundamental II que estudavam em escolas públicas municipais em 2019.

No primeiro gráfico percebe-se que grande parte das escolas possui serviço de coleta de lixo. São os pontos pintados em amarelo. É fácil identificar que essas escolas obtêm melhores notas do que as que não têm disponibilidade de coleta de lixo.

O segundo gráfico indica que poucas escolas usam Internet para aprendizagem, conforme pode ser visto pelos poucos pontos amarelos. Nota-se ainda que essas escolas estão sempre associadas a melhores notas no IDEB.

Os dados de Censo Escolar e de notas do IDEB são possíveis de serem analisados de forma prática usando R.

No próximo tópico você verá algumas possibilidades de análises do Censo Escolar.

4.2 Analisando Dados de Censo Escolar

Os dados de Censo Escolar estão disponíveis a partir do portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Essa instituição é a responsável pela execução anual do censo. Na sua [página](#) estão disponíveis os dados num grande nível de detalhes, os chamados microdados, desde a edição de 1996.

Como você já entendeu, é possível trabalhar com os dados do Censo Escolar de forma bem fácil usando R. Veja como o conjunto de conhecimentos que viu até aqui, principalmente as práticas de manipulação de dados, ajudarão a trazer achados importantes sobre a situação do Ensino Básico no Brasil.



Estes são os códigos usados na videoaula. Teste você mesmo no seu ambiente.

```
library(tidyverse)
#Endereço onde está o arquivo com dados de censo escolar
url_dados_2021<- "https://download.inep.gov.br/
dados_abertos/microdados_censo_escolar_2021.zip"
#Faz o download do arquivo
download.file(url= url_dados_2021, destfile =
"censo_escolar_2021.zip", mode="wb")
#Descompacta o arquivo zip
unzip("censo_escolar_2021.zip", files = "microdados_ed_
basica_2021/dados/microdados_ed_basica_2021.csv")

#lê o arquivo csv que estava no arquivo zip e alimenta o dataframe
microdados_ed_basica_2021 <- read_delim("microdados_ed_
basica_2021/dados/microdados_ed_basica_2021.csv",
              delim = ";", escape_
double = FALSE, locale = locale(encoding = "LATIN1"),
              trim_ws = TRUE)
#verificar uma amostra dos dados presentes no dataframe
glimpse(microdados_ed_basica_2021)
#Gera gráfico de ranking no número de escolas
microdados_ed_basica_2021 %>%
  group_by(SG_UF) %>%
  summarise(
```

```
quantidade = n()
) %>%
mutate(SG_UF = reorder(SG_UF, quantidade)) %>%
ungroup() %>%
ggplot() +
geom_col(aes(x=quantidade, y=SG_UF)) +
scale_x_continuous(breaks = seq(0,35000,5000))
```

Script da videoaula.

Fonte: Barbalho (2020).

Por aqui você está quase encerrando essa capacitação. Espera-se que tenha aprendido os princípios da estatística descritiva em linguagem R para que consiga analisar dados na Administração Pública.

Agora é a hora de você testar seus conhecimentos. Para isso, acesse o exercício avaliativo disponível no ambiente virtual. Bons estudos!

Referências

BARBALHO, Fernando Almeida. **Emergência de um campo de ação estratégica:** o caso de política pública sobre dados abertos. 2014. 254 f., il. Tese (Doutorado em Administração) — Universidade de Brasília (UnB), Brasília, DF, 2014.

BARBALHO, F. **Education as the driver of human development in Brazil:** An analysis of Brazilian census data. [Towards Data Science]. Medium, 2022. Disponível em: <https://towardsdatascience.com/education-as-the-driver-of-human-development-in-brazil-e95f9f0124fe>. Acesso em: 27 nov. 2022.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil.** Brasília, DF: Senado Federal, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 22 mar. 2023.

BRASIL. **Estoque da Dívida Pública Federal.** Portal Tesouro Nacional Transparente. 2023a. Disponível em: <https://www.tesourotransparente.gov.br/ckan/dataset/estoque-da-dívida-pública-federal>. Acesso em: 22 mar. 2023.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). **Cidades e Estados.** 2021. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>. Acesso em: 20 mar. 2023.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **IDEB Resultados 2019.** 2021. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>. Acesso em: 20 mar. 2023

BRASIL. Secretaria do Tesouro Nacional. **Análises sobre SICONFI:** Despesas com educação x IDEB. 2022c. Disponível em: https://siconfi-ideb-2019.tesouro.gov.br/dashboard_IDEB_2019.Rmd#section-análise-por-outras-variáveis. Acesso em: 8 mar. 2023.

BRAUNSCHWEIG, K.; EBERIES, J.; THIELE, M.; LEHNER, W. The state of open data. In: World Wide Web Conference, 21st, 2012, Lyon. **Anais [...]**. New York: Web Science Track, 2012.

BRAZILIAN, M. et al. Open source software and crowdsourcing for energy analysis. **Energy Policy**, v. 49, p. 149–153, 2012

BUSSAB, W; MORETTIN, O. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010.

CRAVEIRO, G.; SANTANA, M.; ALBUQUERQUE, J. Assessing Open Government Budgetary Data in Brazil. In: International Conference on Digital Society (ICDS), 7th, 2013, Nice. **Anais [...]**, Wilmington: IARIA Press, 2013.

DARÓCZI, Gergely. **Number of R packages submitted to CRAN**. GitHub Gist. [20--]. Disponível em: <https://gist.github.com/daroczig/3cf06d6db4be2bbe3368>. Acesso em: 13 mar. 2023.

FREEPIK COMPANY. [Banco de Imagens]. **Freepik**, Málaga, 2023. Disponível em: <https://br.freepik.com>. Acesso em: 7 mar. 2023.

IHAKA, Ross. **The R Project**: a brief history and thoughts about the future. Auckland: The University of Auckland, S.d.. 34 slides, color. Disponível em: <https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf>. Acesso em: 8 mar. 2023.

MACIEL, E. et al. **Fatores associados ao óbito hospitalar por COVID-19 no Espírito Santo**. [Epidemiol]. Brasília, DF: ServSaúde, 2020

MARTOS, G. **Cluster analysis with R**. RPubs. 2014. Disponível em: <https://rpubs.com/gabrielmartos/ClusterAnalysis>. Acesso em 27 nov. 2022.

MIDDLETON, Juliet. **Academic unfazed by rock star status**. [NZ Herald]. 2009. Disponível em: <https://www.nzherald.co.nz/nz/academic-unfazed-by-rock-star-status/LMU5EIMP7QCMZFCBM3UNW4C7CI/>. Acesso em: 3 out. 2022.

O'BRIEN, S.; CURRY, E.; HARTH, A. XBRL and open data for global financial ecosystems: a linked data approach. **International Journal of Accounting Information Systems**, v. 13, n. 2, p. 141–162, 2012.

SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdados: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cad. Saúde Pública**, Rio de Janeiro, v. 35, n. 9, 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 3 ago. 2023.

TODOS PELA EDUCAÇÃO. **Qual é o papel da união, dos estados e dos municípios na educação?**. 2018. Disponível em: <https://todospelaeducacao.org.br/noticias/qual-e-o-papel-da-uniao-dos-estados-e-dos-municipios-na-educacao/>. Acesso em: 22 mar. 2023.

WIKIMEDIA FOUNDATION. **Coeficiente de correlação de Pearson**. [Wikipédia]. 2022. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Estatística**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Estat%C3%ADstica>. Acesso em: 3 out. 2022.

WIKIMEDIA FOUNDATION. **Logaritmo**. [Wikipédia]. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Logaritmo>. Acesso em: 20 mar. 2023.