Centro Federal de Educação Tecnológica de Minas Gerais
PROFMAT - Mestrado Profissional em Matemática em Rede Nacional

Helenice Vasconcelos de Melo Lopes

Estatística Descritiva: *Software* R na análise dos dados da Dengue no Brasil

Belo Horizonte
2024

Helenice Vasconcelos de Melo Lopes

ESTATÍSTICA DESCRITIVA: *SOFTWARE* R NA ANÁLISE DOS DADOS DA DENGUE NO BRASIL

Dissertação apresentada ao Centro Federal de Educação Tecnológica de Minas Gerais como parte das exigências do Programa de Pós-Graduação Mestrado Profissional em Matemática em Rede Nacional, para obter o título de Mestre.

Orientadora
Profa. Dra. Marcela Richele Ferreira

Coorientação
Prof. Dr. Dênis Emanuel da Costa Vargas

Banca Examinadora
Davidson Paulo Azevedo Oliveira
Jahina Fagundes de Assis Hattori
Lívia Maria Dutra

Belo Horizonte
2024

Lopes, Helenice Vasconcelos de Melo
L864e Estatística descritiva: software R na análise dos dados da dengue no Brasil / Helenice Vasconcelos de Melo Lopes. – 2024.
73 f.

Dissertação de mestrado apresentada ao Programa de Mestrado Profissional em Matemática em Rede Nacional.
Orientadora: Marcela Richele Ferreira.
Coorientador: Dênis Emanuel da Costa Vargas.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Estatística – Brasil – Teses. 2. Estatística (Ensino) – Brasil – Teses. 3. R (Software) – Teses. 4. Dengue – Brasil – Teses. 5. DataSUS (Banco de dados) – Teses. I. Ferreira, Marcela Richele. II. Vargas, Dênis Emanuel da Costa. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 519.5081

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6ª 1592 / Cefet/MG

Helenice Vasconcelos de Melo Lopes

ESTATÍSTICA DESCRITIVA: *SOFTWARE* R NA ANÁLISE DOS DADOS DA DENGUE NO BRASIL

Dissertação apresentada ao Centro Federal de Educação Tecnológica de Minas Gerais como parte das exigências do Programa de Pós-Graduação Mestrado Profissional em Matemática em Rede Nacional, para obter o título de Mestre.

Aprovada: 27 de fevereiro de 2024.

Helenice Vasconcelos de Melo Lopes
(Autora)

Marcela Richele Ferreira
(Orientadora)

Belo Horizonte
2024

Dedico esse trabalho aos meus pais. Que valorizam grandemente o conhecimento e sempre me incentivaram e apoiaram a minha caminhada.

Agradecimentos

Muitas pessoas contribuíram para que esse trabalho fosse concluído. Vou citar algumas que são muito importantes e que, sem elas isso não seria possível.

Primeiramente, agradeço a Deus, que me ergueu em momentos que achei que não conseguiria. Aos meus pais que sempre me incentivaram a estudar e a valorizar o conhecimento. E que cuidaram dos meus filhos em vários momentos para que eu pudesse estudar. Agradeço às minha irmãs, Ana Carolina e Amanda pelo apoio incondicional. Obrigada por sempre estarem disponíveis.

Meu marido Samuel. Nem tenho palavras para descrever o que você significa na minha vida. Meu companheiro e maior incentivador. Sem seu apoio (em todos os sentidos), nada disso seria possível. Obrigada por tudo. Agradeço aos meus filhos, Miguel, Helena e Isabel. Vocês são a minha motivação diária!

Sobre meus orientadores. Não há maneira de agradecer a vocês todo empenho em me ajudar. Marcela, você foi a mão de Deus me guiando. Sem você eu não teria conseguido. Você compartilhou muito comigo. Teve paciência não desistiu de mim, quando eu já tinha desistido. Obrigada Dênis. Seu conhecimento foi essencial para que eu conseguisse caminhar. Queria ter aproveitado mais de vocês!!!

Não posso deixar de citar meus colegas e professores, que fizeram parte da minha caminhada no CEFET. Em especial, preciso citar a Luana, minha fiel ouvinte. Em diversos momentos você me ouviu e acalmou. Me ajudou nos momentos tensos e compartilhou das alegrias/alívios que vivemos. Ao Lucas, meu "professor"de cálculo. Você foi muito generoso.

Quero agradecer aos meus colegas de trabalho. Que me ajudaram quando eu precisei me ausentar em diversas situações. Vocês fizeram a diferença na minha vida. Obrigada, Eliana, Ana Cláudia e Aldina.

Aos professores Davidson, Lívia e Jahina, agradeço por aceitarem participar da minha banca e por terem feito tantas contribuições ao meu trabalho.

Enfim, agradeço a todos que contribuíram com o meu trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Resumo

A comunidade escolar tem enfrentado o desafio da lacuna no conhecimento dos estudantes, decorrente da falta de aulas presenciais durante o período de distanciamento social causado pela pandemia da Covid-19. Além disso, há a implementação do novo modelo do Ensino Médio, conforme estabelecido pela Lei nº 13.415/2017, que modifica a Lei de Diretrizes e Bases da Educação Nacional. No contexto desse novo modelo educacional, é importante promover a utilização de tecnologias no ensino. Sabemos que o uso de ferramentas tecnológicas, como os *softwares*, representa uma abordagem eficaz para proporcionar aprendizado prático e direcionado, adaptado às necessidades e interesses dos estudantes. A linguagem R é uma dessas ferramentas tecnológicas que, além de ser gratuita, está entre as mais utilizadas para análise de dados. Além disso, é notória a importância dos conhecimentos de Estatística na formação do cidadão, que pode capacitar os indivíduos a interpretar dados de maneira crítica bem como embasar suas tomadas de decisões. Considerando que a Estatística é uma unidade temática presente na área de Matemática, conforme delineado pela Base Nacional Comum Curricular, optamos por apresentar a Estatística Descritiva a partir da manipulação do um banco de dados referente aos casos de Dengue no Brasil, disponível na página do Ministério da Saúde/DataSUS, utilizando a linguagem R, que possui interpretadores de código aberto disponíveis para diferentes sistemas operacionais ou mesmo online. Assim, proporcionar aos estudantes do ensino médio o acesso a um trabalho com a linguagem R no ensino de estatística com dados reais, com potencial para aprimorar suas experiências de aprendizado em sala de aula e contribuir bastante em sua formação cidadã. Esta dissertação propõe um produto educacional que oferece ao professor um material que o oriente no trabalho com esse tópico.

Palavras-chave: DataSUS; Dengue; Ensino de Estatística; Estatística Descritiva; *Software* R.

Abstract

The school community has faced the challenge of the gap in students' knowledge, resulting from the lack of in-person classes during the period of social distancing caused by the Covid-19 pandemic. Furthermore, there is the implementation of the new High School model, as established by Law No. 13,415/2017, which modifies the National Education Guidelines and Bases Law. In the context of this new educational model, it is important to promote the use of technologies in teaching. We know that the use of technological tools, such as software, represents an effective approach to providing practical and targeted learning, adapted to the needs and interests of students. The R language is one of those technological tools that, in addition to being free, is among the most used for data analysis. Furthermore, the importance of knowledge of Statistics in the training of citizens is well-known, which can enable individuals to interpret data critically as well as support their decision-making. Considering that Statistics is a thematic unit present in the area of Mathematics, as outlined by the National Common Curricular Base, we chose to present Descriptive Statistics based on the manipulation of a database referring to Dengue cases in Brazil, available on the Ministry's website da Saúde/DataSUS, using the R language, which has open-source interpreters available for different operating systems or even online. Thus, providing high school students with access to work with the R language in teaching statistics with real data, with the potential to improve their learning experiences in the classroom and greatly contribute to their civic education. This dissertation proposes an educational product that offers the teacher material that guides him in working with this topic.

Keywords: Descriptive statistics; Software R; Teaching Statistics; DataSUS; Dengue.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Desde os primórdios das civilizações, os governos se interessavam por coletar informações a respeito de suas populações e de suas riquezas. O registro de informações perde-se no tempo. Segundo (MEMÓRIA, 2004) [1], Confúcio, um filósofo chinês que viveu por volta do ano 500 E.C, relatou levantamentos feitos na China, há mais de 2000 anos antes da era cristã. No antigo Egito, os faraós fizeram uso sistemático de informações de caráter estatístico, conforme evidenciaram pesquisas arqueológicas. Desses registros também se utilizaram as civilizações pré-colombianas dos maias, astecas e incas. É conhecido de todos os cristãos o recenseamento dos judeus, ordenado pelo Imperador Augusto. Os balancetes do Império Romano, o inventário das posses de Carlos Magno, o Doomsday Book, registro que Guilherme, o Conquistador, invasor normando da Inglaterra, no século XI, mandou levantar as propriedades rurais dos conquistados anglo-saxões para se inteirar de suas riquezas.

Ainda segundo MEMÓRIA (2004), os romanos foram um dos pioneiros na administração do Estado e usavam de procedimentos para fazer uma gestão eficiente de assuntos públicos. Nesse contexto, a origem da palavra estatística remonta da Roma Antiga.

Sua origem etimológica vem do latim *status*, compreendendo o que se refere ao estado. Os romanos foram pioneiros na administração do Estado e usavam procedimento para fazer uma gestão eficiente de assuntos públicos. (NETO, 2020) [2].

Ainda hoje, o conceito popular de estatística compete dados numéricos apresentados em quadros, tabelas e gráficos, publicados por agências governamentais ou institutos de pesquisas. Entende-se que esses dados, normalmente, são demográficos ou econômicos. Entretanto, hoje o conceito de estatística é muito mais amplo do que essa definição. Admite-se que os processos estatísticos envolvem uma multiplicidade de causas e, por isso usa-se métodos matemáticos para apresentar, analisar e interpretar os dados quantitativos obtidos.

Atualmente, estamos expostos a uma enorme quantidade de informações. Com o

desenvolvimento das tecnologias digitais, a informação é difundida com uma velocidade nunca vista. O acesso à informação e aos dados está facilitado, o que faz com que sempre estejamos expostos a eles. Isso não significa que somos capazes de interpretar as informações ou mesmo acessá-las plenamente.

As grandes empresas, os governos e os veículos de comunicação estão buscando constantemente entender os comportamentos das pessoas. Para isso, pesquisas são feitas com o objetivo de se coletar dados que possibilitarão ter percepções importantes sobre o comportamento de uma determinada sociedade.

Sobre a relevância da Estatística, (BENEVIDES, 2023) [3] afirma:

> Estudar Estatística é importante para que, além da tarefa óbvia de entender o conjunto de dados analisados, possamos também entender os limites da própria Estatística. (...) conhecer a média de um conjunto de valores é uma informação preciosa. Dessa forma, um ponto importante é saber interpretar os dados resumidos para tirar conclusões a partir deles.

A estatística é essencial para a humanidade, pois são os resultados de análises estatísticas que conduzem as ações de governos e empresas. Portanto, é importante que os cidadãos compreendam seus conceitos. As instituições, governos e empresas têm usado de pesquisas para coletar dados e direcionar suas ações. É possível observar situações em que as pesquisas estatísticas são fundamentais nas tomadas de decisões. Como exemplo, pode-se citar as pesquisas de intenção de voto nas eleições, a pandemia da Covid-19, e as análises estatísticas que foram utilizadas para as tomadas de decisões dos governos. Um exemplo ocorrido durante a pandemia da Covid-19, que em 2022, o governo Federal determinou o fim da Emergência em Saúde Pública de Importância Nacional (ESPIN) no Brasil. Situação que instituía, entre outras ações, a obrigatoriedade do uso de máscara. A portaria foi publicada em abril de 2022 e trazia a decisão. Conforme reportagem publicada pelo (MINISTÉRIO DA SAÚDE, 2022) em sua pagina oficial, [4]:

> Para determinar o fim da ESPIN, o Ministério da Saúde considerou a capacidade de resposta do Sistema Único de Saúde (SUS), a melhora no cenário epidemiológico no país e o avanço da campanha de vacinação. O Brasil registra queda de mais de 80% na média móvel de casos e óbitos pela Covid-19, em comparação com o pico de casos originados pela variante Ômicron, no começo deste ano. Os critérios epidemiológicos, com parecer das áreas técnicas da Pasta, indicam que o país não está mais em situação de emergência de saúde pública nacional.

Nesse contexto, faz-se necessário promover o ensino de Estatística com o objetivo de que os estudantes compreendam tais informações, bem como os tornar capazes de avaliar a veracidade delas. É evidente que há uma disparidade no que diz respeito ao acesso às tecnologias e a interpretação de discursos científicos, o que reforça ainda mais a importância de ações no sentido de amenizar tais disparidades.

Sobre a relevância da Estatística para a formação intelectual dos estudantes, (CARZOLA e CASTRO, 2008) [5] defendem em seu trabalho a importância de incluir os conceitos básicos de Estatística e Probabilidades no currículo da Educação Básica, para formar cidadãos capazes de ler e interpretar as informações cotidianas que podem induzir a decisões equivocadas. Como exemplos, as autoras citam que informações estatísticas podem ser manipuladas, distorcidas ou mal compreendidas, tais como as pesquisas eleitorais, as manchetes de jornais e suas representações gráficas.

A dissertação apresentada por (MORAES, 2020) [6] segue a mesma ideia de que a Estatística pode ser usada para manipulação de informações. O trabalho tem o objetivo de questionar a manipulação de dados estatísticos para pesquisas a fim de mostrar a influência da matemática de uma forma diferenciada. A autora apresenta uma análise do livro "Como mentir com estatística" de Darrel Huff e também transcreve e compara reportagens de sites de notícias com diferentes ênfases nos mesmos dados estatísticos.

A Estatística na Educação Básica era tratada de maneira superficial. Nos últimos anos, houve um avanço com o texto da Base Nacional Comum Curricular (BNCC), que "trata a Estatística como um dos campos da Matemática e estabelece que seja trabalhado desde os anos iniciais do Ensino Fundamental. Dando continuidade, no Ensino Médio, ao que fora trabalhado no Ensino Fundamental" [7]. Tais avanços representam um olhar mais atento à importância do ensino de Estatística por parte das autoridades. Essa dissertação apresenta a Estatística na BNCC discutida de maneira mais detalhada no Capítulo 2.

Ainda nesse contexto, a análise estatística vem sendo impactada pelo desenvolvimento tecnológico e a manipulação de dados é amplamente beneficiada por esses avanços. Além disso, o mundo cada vez mais globalizado torna a análise estatística ainda mais essencial. Nesse mesmo cenário, (SOUZA, 2015) [8] afirma que

> cada vez mais o desenvolvimento cognitivo do ser humano está sendo mediado por dispositivos tecnológicos, onde as novas tecnologias de informação e comunicação estão ampliando o potencial humano. Observa-se que a informação se disponibiliza através de tecnologias inovadoras, o

> que demanda novas formas de se pensar, agir, conviver e principalmente aprender com e através dessas tecnologias.

Em sua dissertação, (ARAÚJO, 2020) [9] apresenta sugestões de atividades para o ensino de Estatística por meio do uso de tecnologias no Ensino Médio. Ela fundamenta a escolha desse tema a partir da relevância da Estatística na formação cidadã e da necessidade de acompanhar o progresso tecnológico na educação.

A discussão sobre o uso de tecnologias na educação já vinha ocorrendo nas últimas décadas, mas após o cenário da pandemia da Covid-19, ficou evidente que a modernização do ensino é urgente. Em sua pesquisa, (SOUZA, 2015) [8] defende que novas tecnologias são parceiras no processo de ensino-aprendizagem e que os professores devem compreendê-las e estudá-las para saber a melhor forma de introduzi-las em sala de aula e com elas alcançar melhores resultados entre os alunos. Sua dissertação motiva professores a usar *softwares* de matemática em suas salas de aula, bem como em seus estudos e planejamento. Ainda sobre a relevância dessa inovação no ensino, (GADANIDIS, BORBA e SILVA, 2020) [10] *apud* (BORBA, 2009) [11] afirmam que

> as tecnologias estão mudando a própria noção do que é ser humano. As tecnologias digitais móveis - internet, celular, tablets - estão modificando as normas que vivemos, os valores associados a determinadas ações. Mais uma vez isso acontece em ritmo diferente fora e dentro da escola. Assim o abismo entre práticas que alunos e professores têm fora da escola e dentro da mesma instituição aumenta.

Pensando em tomar uma base de dados relevante, a escolha da análise dos casos de Dengue no Brasil se deu, por entender que há uma Epidemia da doença no Brasil. A Dengue é uma doença infecciosa viral. Os vírus Dengue (DENV) estão classificados cientificamente na família *Flaviviridae* e no gênero *Flavivirus*. Até o momento são conhecidos quatro sorotipos – DENV-1, DENV-2, DENV-3 e DENV-4 –, que apresentam distintos materiais genéticos (genótipos) e linhagens[12]. Transmitida pela fêmea do mosquito *Aedes aegypti* (significa "odioso do Egito"), a Dengue está distribuída pelo país inteiro. O clima do Brasil, bem como problemas relacionados a saneamento básico e tratamento de lixo, facilitam na propagação do mosquito.

A Dengue é considerada uma epidemia, os casos de vêm crescendo a cada ano no Brasil, tornando essencial que as autoridades e a comunidade científica estejam sempre trabalhando para compreender as causas desse aumento e traçar estratégias para prevenir

e erradicar a doença.

Tendo em vista esse cenário, a escolha o banco de dados sendo sobre – os casos de Dengue no Brasil nos últimos dez anos – fez-se relevante. Além disso, a presença dos casos na realidade das famílias torna ainda mais importante falar sobre essa doença.

Acreditando que se faz necessário o uso de tecnologias no contexto escolar e que o ensino de estatística é essencial na formação dos estudantes, a proposta desta dissertação é aliar o ensino de Estatística ao uso de tecnologias no contexto de dados relevantes à população. É proposto, nessa dissertação, uma discussão sobre o ensino da Estatística Descritiva no contexto do Ensino Médio através do *software* R usando os dados referentes aos casos de Dengue no Brasil disponíveis na página do DATASUS. As atividades propostas que compõe esse produto educacional estão descritas detalhadamente nessa dissertação.

Apresento, portanto, no capítulo 2, a fundamentação teórica. Nela, discuto a Estatística na BNCC. Mostro a linguagem R, o *software* R e o *Posit Cloud*[13] (versão *online*), bem como, a maneira de acessar o *software* e sua interface. Ainda nesse capítulo, apresento o DATASUS (TabNet) [14].

O capítulo 3 é dedicado à Estatística. Nele trato das Medidas-Resumo fazendo o tratamento dos dados no *Software* R.

Finalmente, no capítulo 4, os dados são apresentados na forma de tabelas e gráficos.

2 Fundamentação Teórica

Esse capítulo trata da fundamentação teórica desta dissertação, que fornecerá uma base para a compreensão do contexto em que ela se insere. Inicialmente, será mostrado como a estatística é abordada na BNCC e, em seguida, uma breve apresentação do *software* R. E, finalmente, uma seção de acesso aos dados do DATASUS, através do TabNet.

2.1 Estatística na BNCC

A BNCC [7] é um documento normativo que estabelece as competências e habilidades essenciais que todos os estudantes brasileiros devem desenvolver ao longo das etapas e modalidades da Educação Básica. Entende-se por habilidades, as capacidades que devem ser desenvolvidas pelos estudantes ao longo de sua trajetória escolar. Elas envolvem a aplicação prática do conhecimento em situações diversas. Por exemplo, uma habilidade em Matemática pode ser a capacidade de analisar tabelas, gráficos e amostras de pesquisas Estatísticas. As competências, por sua vez, são conjuntos de habilidades que permitem ao aluno impulsionar seus conhecimentos em diferentes contextos e situações. As competências gerais definidas na BNCC são:

- Conhecimento.
- Pensamento científico, crítico e criativo.
- Repertório cultural
- Argumentação
- Autoconhecimento e autocuidado
- Empatia e cooperação
- Responsabilidade e cidadania

Elas geralmente abrangem não apenas habilidades específicas, mas também atitudes, valores e comportamentos que são essenciais para a participação ativa na sociedade. Dentre as Competências específicas de Matemática e suas Tecnologias para o Ensino Médio[7], estão:

- Utilizar estratégias, conceitos e procedimentos matemáticos para interpretar situações em diversos contextos;
- Propor ou participar de ações para investigar desafios do mundo contemporâneo e tomar decisões éticas e socialmente responsáveis;
- Utilizar estratégias, conceitos, definições e procedimentos matemáticos para interpretar, construir modelos e resolver problemas em diversos contextos;
- Compreender e utilizar, com flexibilidade e precisão, diferentes registros de representações matemáticas;
- Investigar e estabelecer conjecturas a respeito de diferentes conceitos e propriedades matemáticas.

A base curricular serve como referência para a elaboração dos currículos escolares de todo o país (BRASIL, 2017) [7]. Elaborada por especialistas de todas as áreas do conhecimento, ela é um documento completo e contemporâneo, que corresponde às demandas do estudante desta época, preparando-o para o futuro. No que se refere à Estatística, o documento estabelece sua inclusão desde os anos iniciais do Ensino Fundamental dando continuidade nos anos finais do Ensino Fundamental e Ensino Médio.

Em seu artigo, (GIORDANO *et al.* 2019)[15] discutem as novas perspectivas para a educação Estatística no Brasil a partir da publicação da BNCC. Defendem que a abordagem por meio de projetos é uma forma de promover o letramento estatístico dos alunos, articulando a análise exploratória de dados com um modelo de investigação científica, que envolve as dimensões cultural, social, educativa e política da Estatística. Destacam as potencialidades, as possibilidades e os desafios dessa proposta didática, que visa desenvolver competências e habilidades dos alunos para lidar com dados em diversos cenários.

A terceira competência específica de matemática para o Ensino Médio da BNCC, a saber, afirma que é importante,

> Utilizar estratégias, conceitos, definições e procedimentos matemáticos para interpretar, construir modelos e resolver problemas em diversos contextos, analisando a plausibilidade dos resultados e a adequação das soluções propostas, de modo a construir argumentação consistente (BRASIL, 2017) [7]

e tem tudo a ver com o estudo de Estatística. Como pode ser visto, ela destaca desenvolvimento de estratégias, conceitos e procedimentos matemáticos em diversas áreas, abrangendo inclusive Probabilidade e Estatística. Isso inclui a capacidade de elaborar modelos matemáticos, interpretar e resolver problemas em diferentes contextos, tudo fundamentado em uma argumentação consistente. Essa abordagem busca não apenas a aplicação de fórmulas e métodos, mas também a compreensão profunda dos princípios matemáticos e a habilidade de argumentar de forma lógica e fundamentada ao enfrentar desafios matemáticos em variados cenários.

Dentre as habilidades de Matemática e suas tecnologias, podemos destacar aqui algumas que se relacionam diretamente ao ensino de estatística no Ensino Médio [7]:

(EM13MAT102) Analisar tabelas, gráficos e amostras de pesquisas Estatísticas apresentadas em relatórios divulgados por diferentes meios de comunicação, identificando, quando for o caso, inadequações que possam induzir a erros de interpretação, como escalas e amostras não apropriadas;

(EM13MAT202) Planejar e executar pesquisa amostral sobre questões relevantes, usando dados coletados diretamente ou em diferentes fontes, e comunicar os resultados por meio de relatório contendo gráficos e interpretação das medidas de tendência central e das medidas de dispersão (amplitude e desvio padrão), utilizando ou não recursos tecnológicos;

(EM13MAT316) Resolver e elaborar problemas, em diferentes contextos, que envolvem cálculo e interpretação das medidas de tendência central (média, moda, mediana) e das medidas de dispersão (amplitude, variância e desvio padrão);

(EM13MAT406) Construir e interpretar tabelas e gráficos de frequências com base em dados obtidos em pesquisas por amostras estatísticas, incluindo ou não o uso de *softwares* que inter-relacionem Estatística, Geometria e Álgebra;

(EM13MAT407) Interpretar e comparar conjuntos de dados estatísticos por meio de diferentes diagramas e gráficos reconhecendo os mais eficientes para sua análise;

É importante destacar a relevância com que a abordagem prática do trabalho com a Estatística é tratada nas habilidades da BNCC. Além disso, o uso de tecnologias digitais

é incentivado em diversos momentos na BNCC, o que valida essa proposta de trabalho apresentada.

2.2 O *software* R

O R é uma linguagem de programação muito utilizada para se trabalhar em ambientes estatísticos, que permite a análise de dados, construção de gráficos e produção de relatórios. Ele foi desenvolvido na década de 1990 na Universidade de Auckland, Nova Zelândia e recebeu esse nome, fazendo referência as iniciais de seus criadores: Ross Ihaka e Robert Gentleman [16]. O *software* oferece uma variedade de métodos (modelagem linear e não linear, testes estatísticos, modelos de séries temporais, classificação, métodos multivariados etc.) e métodos gráficos. Destaca-se no *software* R a facilidade com que gráficos são projetados com alta qualidade, além da possibilidade de se incluir fórmulas matemáticas e símbolos, se necessário. Foi desenvolvido especificamente para análise estatística e é uma das muitas ferramentas utilizadas para a tomada de decisões. O *software* R é um programa de código livre, sendo adaptado aos sistemas operacionais Linux, Mac OS e Windows (BOAS, 2021)[17].

Para fazer o *download* do *software* R basta acessar a página através do *link* https://www.r-project.org/ clicar em "CRAN"e escolher o *link* de *download* [18]. Uma outra opção para download é o *RStudio*, que é um ambiente de desenvolvimento integrado (IDE) para R e *Python*. Inclui um console, editor de realce de sintaxe que suporta execução direta de código e ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho. O *RStudio* está disponível em edições comerciais e de código aberto e roda em desktop (Windows, Mac e Linux).

Há ainda uma opção *online* através do *Posit Cloud* [13], um ambiente de desenvolvimento integrado ao R que apresenta uma interface gráfica mais amigável, com muitos recursos úteis que facilitam a visualização do código, a importação de conjunto de dados e a visualização das imagens. Para acessar essa versão *online*, deve-se entrar no *link* https://posit.cloud/ clicar em *"GET STARTED FOR FREE"* e criar uma conta usando um endereço de e-mail válido. A Figura 2.1 trás um QR code para acesso direto a página do *Posit Cloud*

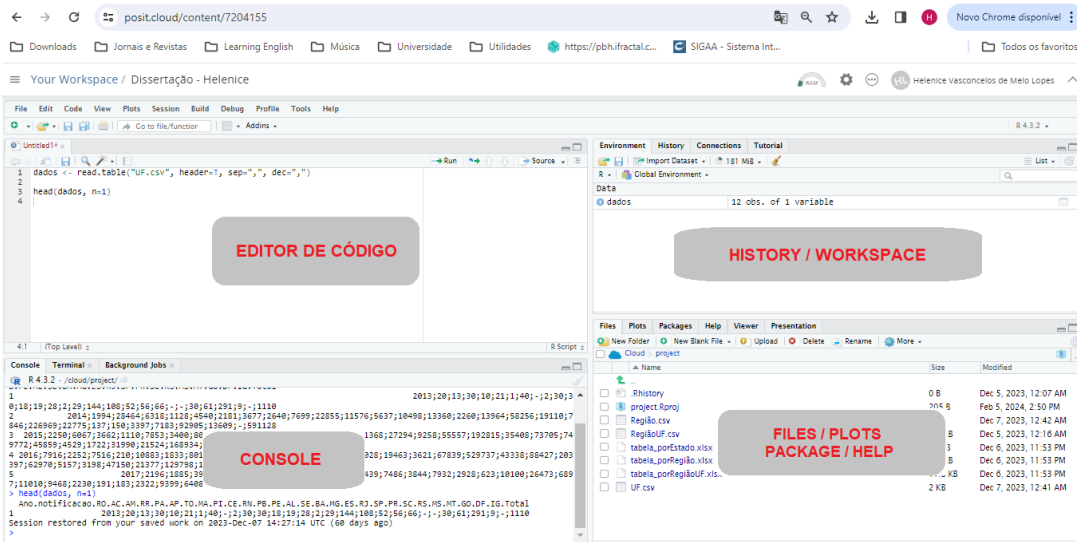
A imagem seguinte (Figura 2.2) representa a interface da plataforma *Posit.Cloud*. Ela mostra o editor de códigos, que é o espaço onde os códigos serão digitados, além

Figura 2.1: QR code para acesso a página do (*Posit Cloud*)

Fonte: Elaborado pela autora (2024).

dos resultados dos códigos que aparecem em *console*. As informações do histórico e das variáveis criadas ficam em *history/workspace* respectivamente. Em *files/plots/package/help* temos o o espaço onde são armazenados arquivos externos, os gráficos plotados, os pacotes e o suporte de ajuda da plataforma.

Figura 2.2: Interface da plataforma (*Posit Cloud*)



Fonte: Elaborado pela autora (2024).

2.3 Acessanndo um Banco de Dados do DATASUS e inserindo no *Posit Cloud*

2.3.1 Acessando os dados no DATASUS (TabNet)

DATASUS [14] é o Departamento de Informática do Sistema Único de Saúde. Surgiu em 1991 com a criação da FUNASA (Fundação Nacional de Saúde). Ele tem como responsabilidade prover os órgãos do SUS de sistemas de informação e suporte de informática, necessários ao processo de planejamento, operação e controle. Em quase 25 anos de atuação, o DATASUS já desenvolveu mais de 200 sistemas que auxiliam diretamente o Ministério da Saúde no processo de construção e fortalecimento do SUS. Atualmente, o Departamento é um grande provedor de soluções de *software* para as secretarias estaduais e municipais de saúde, sempre adaptando seus sistemas às necessidades dos gestores e incorporando novas tecnologias, na medida em que a descentralização da gestão se torna mais concreta. O DATASUS está presente em todas as regiões do país por meio das regionais que executam as atividades de fomento e cooperação técnica em informática nos principais estados brasileiros. As informações apresentadas nessa dissertação foram extraídas da página do TabNet, que é um aplicativo tabulador genérico de domínio público que permite organizar dados de forma rápida, conforme a consulta que se deseja tabular. A Figura 2.3 mostra o *layout* da página do DATASUS/TabNet.

Figura 2.3: Página inicial - DATASUS TabNet

Fonte: Elaborado pela autora (2024).

Ainda na página do DATASUS, consta um tutorial (DATASUS, 2020)[19] com os passos para extrair informações do banco de dados do TabNet. Para acessar as informações que utilizaremos neste trabalho, devemos seguir os seguintes passos:

- Acesse o portal do DATASUS em https://datasus.saude.gov.br/. A Figura 2.4 apresenta o QR code para acesso direto a página do TabNet;

Figura 2.4: QR code para acesso a página do DATASUS - TabNet

Fonte: Elaborado pela autora (2024).

- Acesse o *link* "Acesso à informação", no menu principal, depois role a página para clicar no botão "TabNet", ou
- Na página inicial, role até "Serviços para o Cidadão" e clique no botão "TabNet" como pode ser observado na Figura 2.5.

Na página seguinte, uma tela com várias opções de acesso à origem das informações será apresentada, como ilustrado na Figura 2.5.

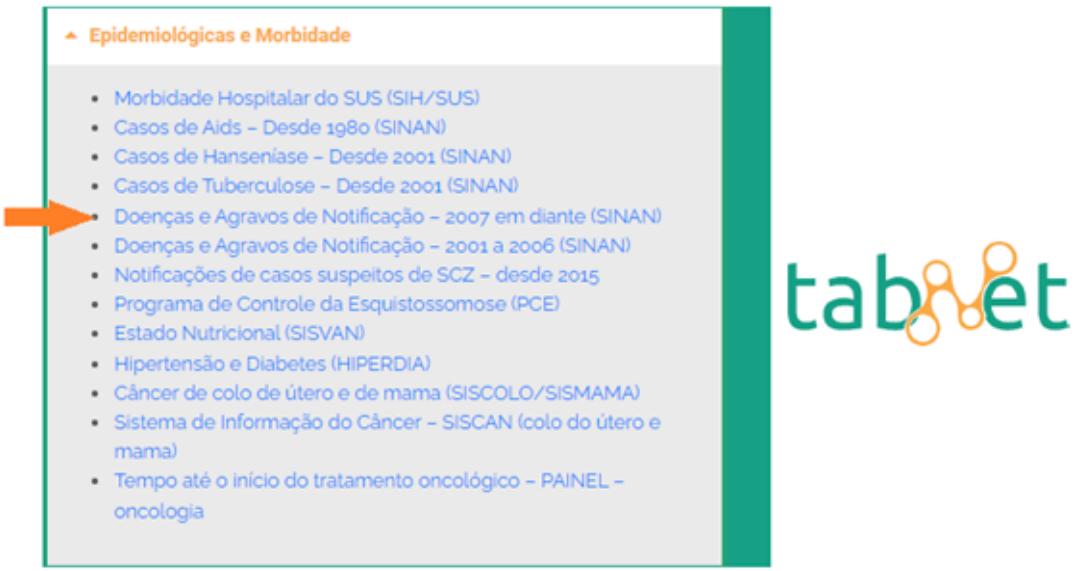
Figura 2.5: DATASUS TabNet

Fonte: (DATASUS, 2020)[19].

Como exemplo utilizado nesta dissertação, os seguintes passos devem ser seguidos:

- Nessa página, acessar "Epidemiológicas e Morbidade" e na sequência clicar em "Doenças e Agravos de Notificação – 2007 em diante (SINAN)", conforme a Figura 2.6 mostra.

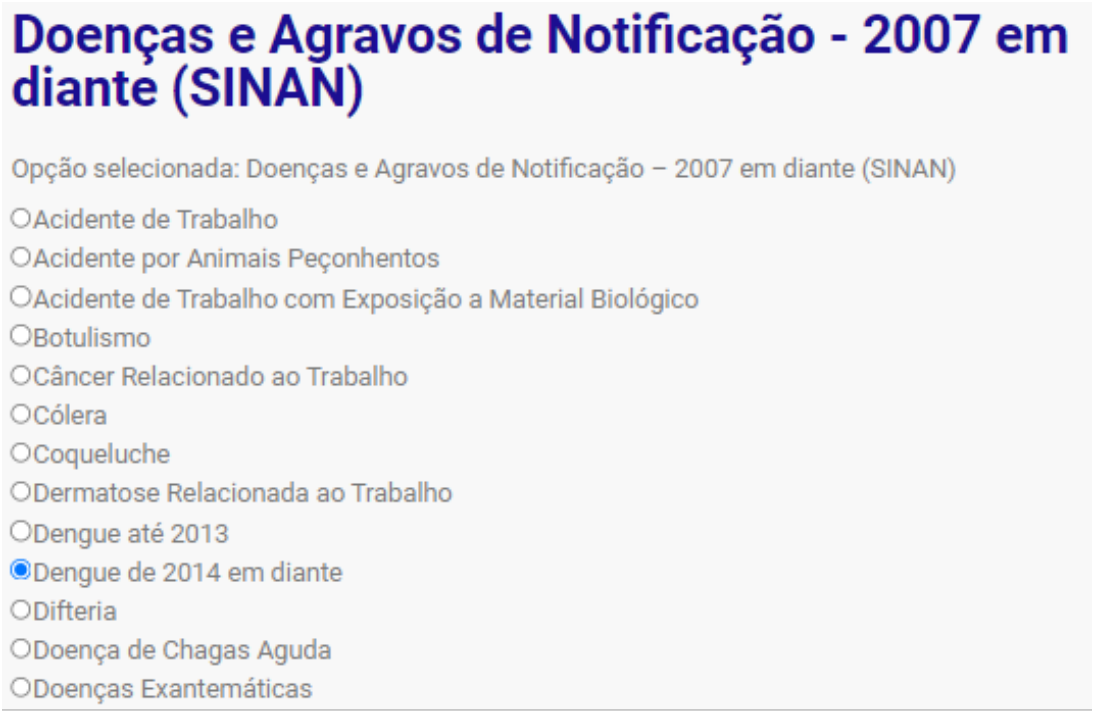
Figura 2.6: Epidemiológicas e Morbidade - TabNet



Fonte: Elaborado pela autora (2024).

- Ao acessar esse *link*, aparecerá uma tela - Figura 2.7 - com várias opções de doenças. Nessa tela clicar em "Dengue de 2014 em diante".

Figura 2.7: Escolha da doença - TabNet

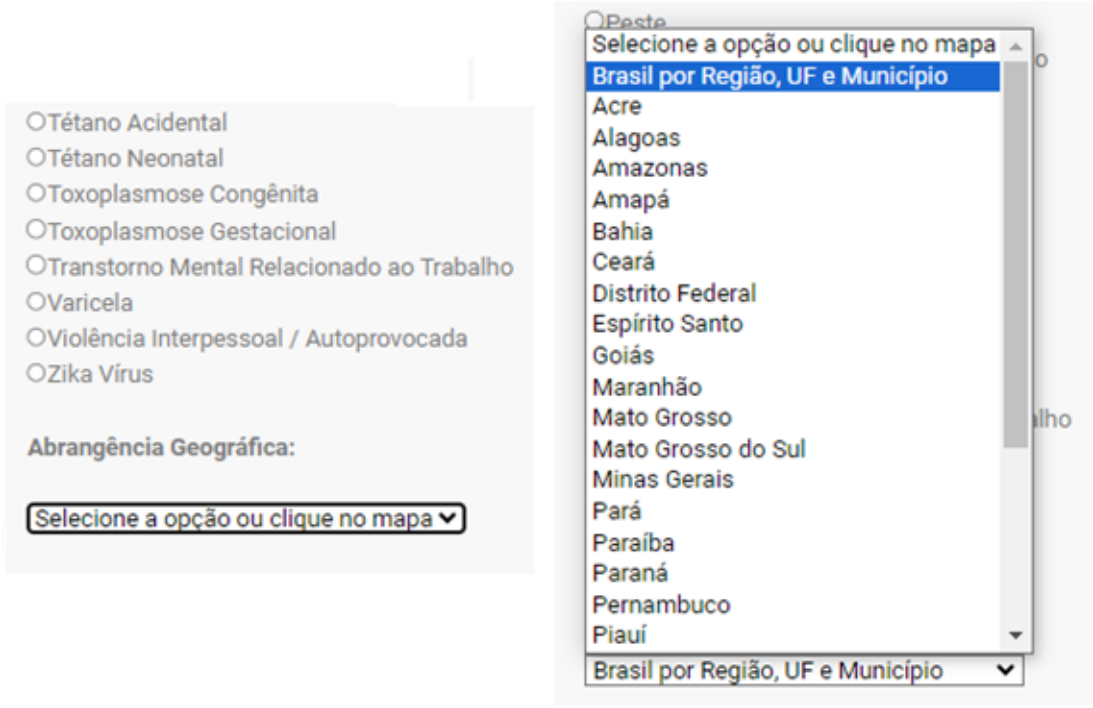


Fonte: Elaborado pela autora (2024).

- No final da página há uma ícone sobre a abrangência geográfica do banco de dados desejado. No caso do banco de dados deste trabalho, a seleção desejada é "Brasil por

Regiões, UF e Municípios".

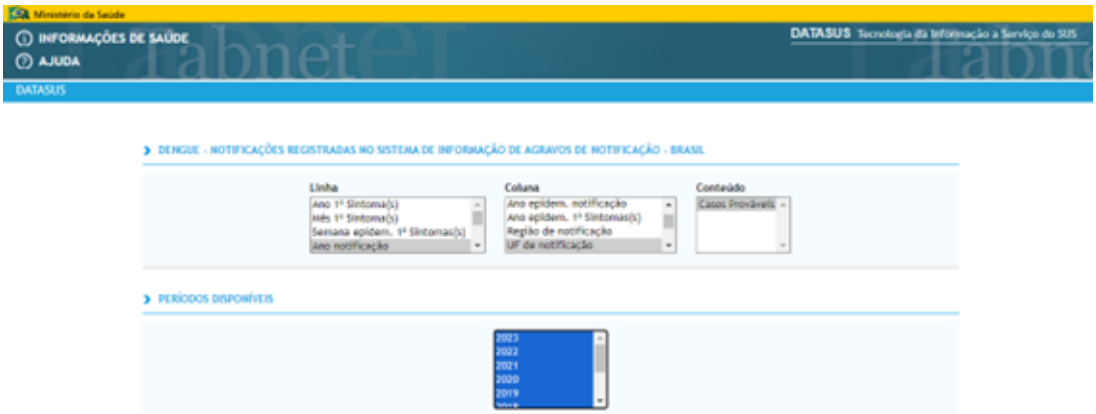
Figura 2.8: Abrangência Geográfica - TabNet



Fonte: Elaborado pela autora (2024).

- Uma página se abrirá e as opções para linhas e colunas serão apresentadas. Vamos selecionar "Ano de notificação" para as linhas e "UF da notificação" para as colunas. Para o período, será selecionado os anos de 2014 a 2023. Como apresentado na Figura 2.9.

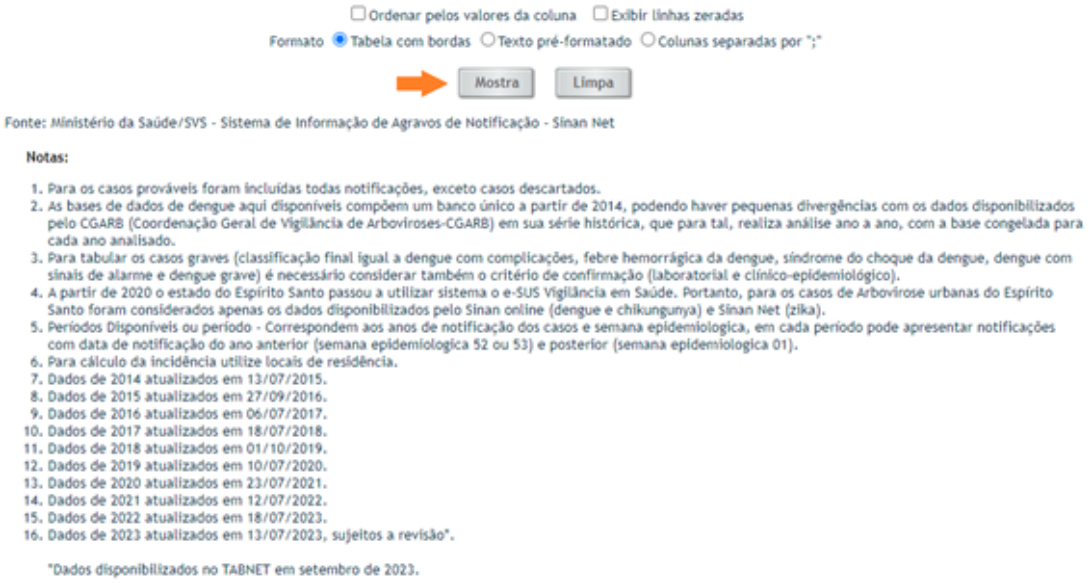
Figura 2.9: Seleção de colunas e linhas da tabela - TabNet



Fonte: Elaborado pela autora (2024).

- Ao final da página, há opções de exibição dos dados. Para exibir a tabela com os dados, basta clicar no ícone "Mostra". Conforme a Figura 2.10.

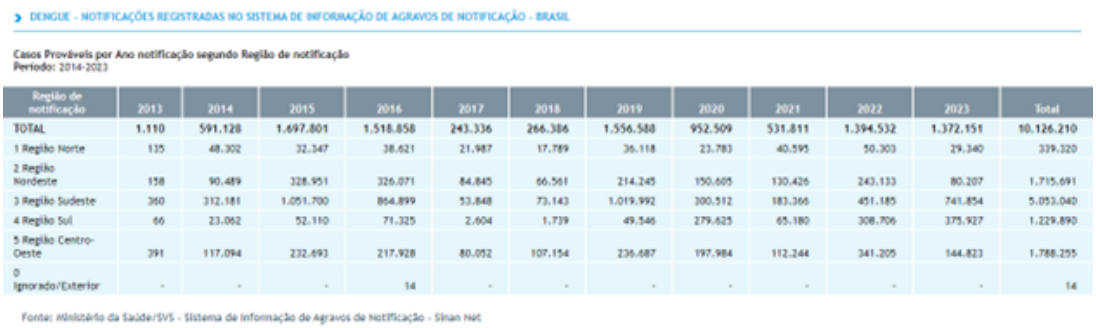
Figura 2.10: Ícone para exibir tabela - TabNet



Fonte: Elaborado pela autora (2024).

- Uma outra aba irá abrir e nela constará uma tabela com os dados exibidos. A tabulação dos dados está pronta. A Figura 2.11 mostra a tabela gerada com as notificações de casos de Dengue por Regiões e os anos.

Figura 2.11: Exemplo de uma tabela gerada no TabNet



DENGUE - NOTIFICAÇÕES REGISTRADAS NO SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO - BRASIL

Casos Prováveis por Ano notificação segundo Região de notificação
Período: 2014-2023

Região de notificação	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
TOTAL	1.110	591.128	1.697.801	1.518.858	243.336	266.386	1.556.588	952.509	531.811	1.394.532	1.372.151	10.126.210
1 Região Norte	135	48.302	32.347	38.621	21.987	17.789	36.118	23.783	40.595	50.303	29.340	339.320
2 Região Nordeste	158	90.489	328.951	326.071	84.845	66.561	214.245	150.605	130.426	243.133	80.207	1.715.691
3 Região Sudeste	360	312.181	1.051.700	864.899	53.848	73.143	1.019.992	300.512	183.366	451.185	741.854	5.053.040
4 Região Sul	66	23.062	52.110	71.325	2.604	1.739	49.546	279.625	65.180	308.706	375.927	1.229.890
5 Região Centro-Oeste	391	117.094	232.693	217.928	80.052	107.154	236.687	197.984	112.244	341.205	144.823	1.788.255
0 Ignorado/Exterior	-	-	-	14	-	-	-	-	-	-	-	14

Fonte: Ministério da Saúde/SVS - Sistema de Informação de Agravos de Notificação - Sinan Net

Fonte: Elaborado pela autora (2024).

- No aplicativo TabNet, além da tabulação de dados, é possível obter os dados já dispostos por gráficos ou mapas. Ao final dessa aba, há três opções para salvar uma cópia da tabela. Neste trabalho iremos utilizar apenas o formato “CSV” (Comma-separated values, em tradução, Valores Separados por Vírgula), conforme mostrado na Figura 2.12 que é um tipo especial de arquivo que você pode criar ou editar no Excel. Em vez de armazenar informações em colunas, os arquivos CSV armazenam informações separadas por vírgulas.

Figura 2.12: Salvando dados no formato "CSV- TabNet

Fonte: Elaborado pela autora (2024).

O banco de dados escolhido se refere aos casos de Dengue notificados no Brasil, no período de 2014 a 2023. A partir de informações provenientes de um banco de dados disponibilizado pelo DATASUS. O Ministério da Saúde disponibiliza, os dados das infecções causadas pelo vírus da Dengue em todo o território nacional, no período de 1975 a 2023. Analisando os dados do sistema, é possível observar que a partir de 2014, houve um aumento significativo nos casos da doença no país.

2.3.2 Carregando um banco de dados no *Posit Cloud*

Nesta dissertação fizemos algumas manipulações do banco de dados relativo aos casos de Dengue no Brasil de 2013 a 2023. Esses dados foram extraídos da página do Ministério da Saúde [14] conforme descrito anteriormente. Com os dados em mãos, temos que inseri-los no *Posit Cloud*, para iniciar as análises.

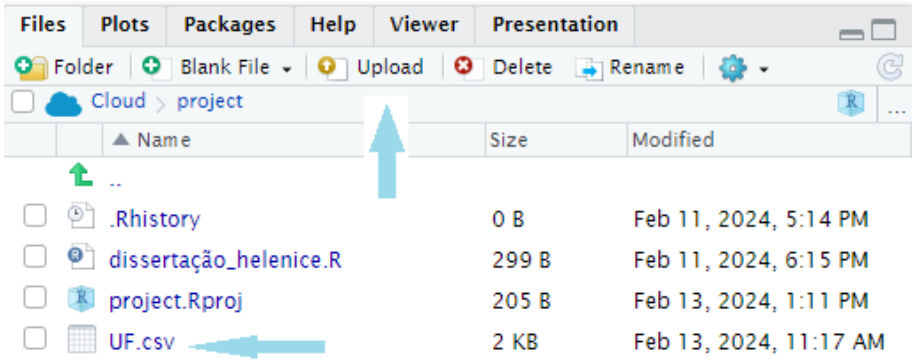
O conjunto de dados escolhido para esta dissertação, está em um arquivo do Excel, no formato CSV. Há outros formatos de arquivos para essa tarefa, mas optamos por essa, que se mostrou mais simples para a linguagem R. O formato CSV é muito útil quando precisamos armazenar dados em forma de tabela para serem manipulados por programas de computador. Nesse tipo de arquivo, cada linha representa uma linha da tabela de dados e os valores em cada uma dessas são separados por vírgulas (,) ou ponto e vírgulas (;).

Com o arquivo, fizemos o *upload* no ambiente do *Posit Cloud*. Na lateral direita da inteface, onde vemos as abas *files/Plots/Packages/Help/Viewer/Presentation*, clicamos em *Upload*, conforme indicado abaixo na Figura 2.13. Em seguida, uma tela irá se abrir e aparecerá a opção de escolher o arquivo - previamente salvo na máquina que está sendo utilizada.

Nesse momento, o arquivo deve ser selecionado. Após executar essa ação, ele irá aparecer na interface, quando selecionamos a aba *files*.

Com o arquivo na base de dados do *Posit Cloud*, é necessário codificar no "Editor de códigos", a leitura do arquivo do tipo CSV. Para isso, usamos a função **read.csv** ("nomedoarquivo", sep = ';',dec = ',').

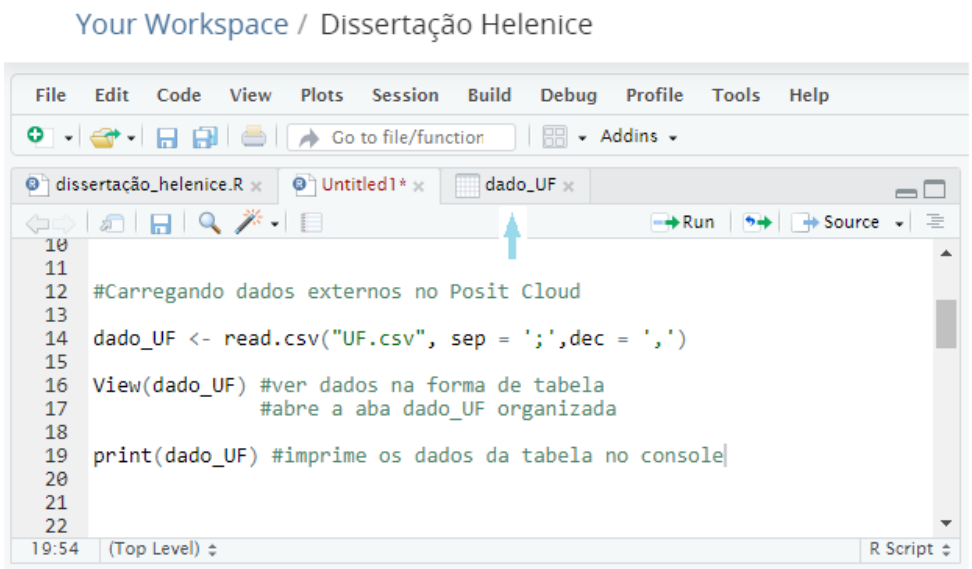
Figura 2.13: Upload de arquivos - *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Nela, é necessário inserir o nome do arquivo entre aspas, o tipo de separador (usamos ponto e vírgula para separar as colunas) e o símbolo para representar decimais (será a vírgula). É importante usar símbolos diferentes para as funções **sep** e **dec**. Por isso, usamos a vírgula e o ponto e vírgula. Podemos ver na Figura 2.14.

Figura 2.14: Arquivo inserido no *Posit Cloud*

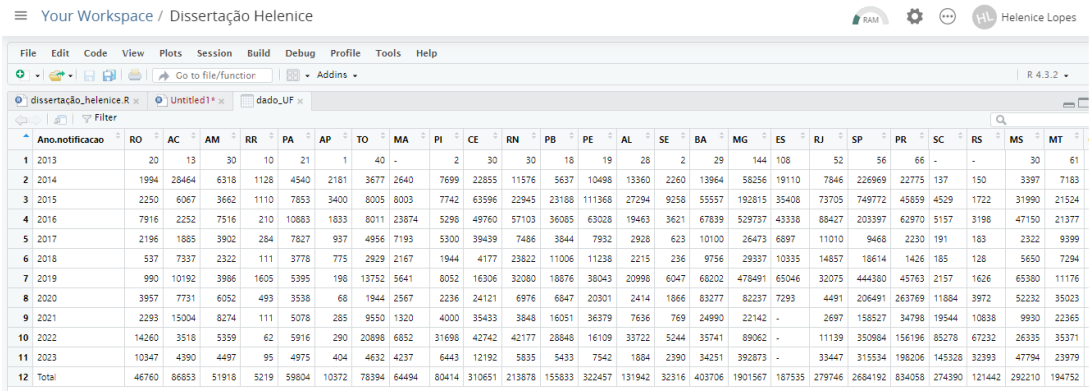


Fonte: Elaborado pela autora (2024).

Damos um nome para a variável, onde os dados são guardados, e com esse nome, usamos a função **View()** (com V maiúsculo) que imprime a tabela criada em uma nova aba da interface, conforme vemos na Figura 2.15. Vemos ainda, como a tabela apresenta os dados bem organizados, como no arquivo do Excel.

Além disso, a função **print()** é usada para imprimir os dados da tabela como vemos

Figura 2.15: Exemplo de uma tabela gerada no *Posit Cloud*



Your Workspace / Dissertação Helenice — Helenice Lopes

File Edit Code View Plots Session Build Debug Profile Tools Help

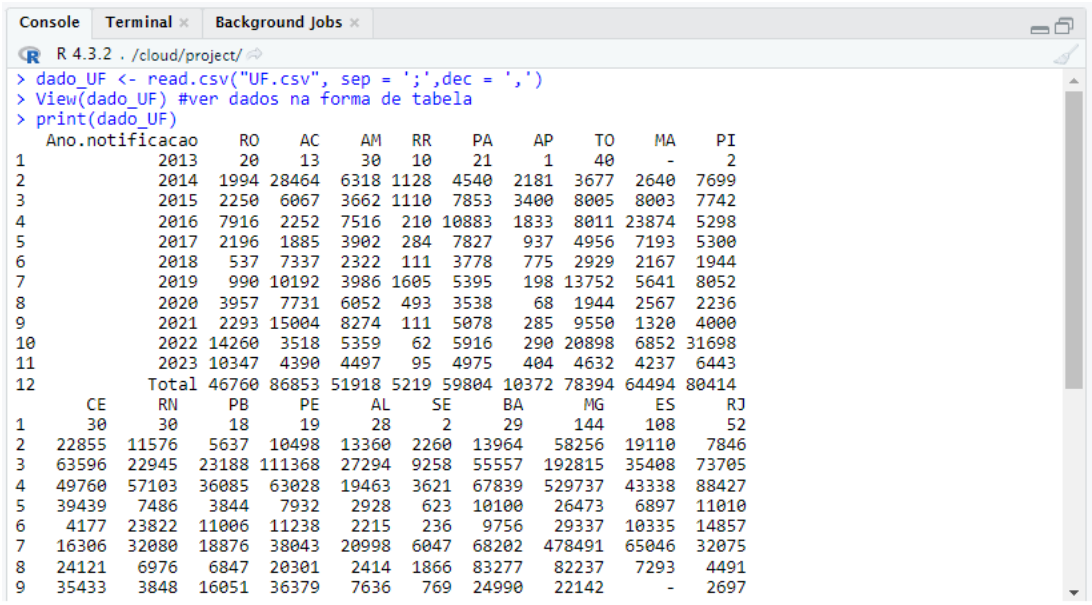
dissertação_helenice.R | Untitled1* | dado_UF

	Ano.notificacao	RO	AC	AM	RR	PA	AP	TO	MA	PI	CE	RN	PB	PE	AL	SE	BA	MG	ES	RJ	SP	PR	SC	RS	MS	MT
1	2013	20	13	30	10	21	1	40	-	2	30	30	18	19	28	2	29	144	108	52	56	66	-	-	30	61
2	2014	1994	28464	6318	1128	4540	2181	3677	2640	7699	22855	11576	5637	10498	13360	2260	13964	58256	19110	7846	226969	22775	137	150	3397	7183
3	2015	2250	6067	3662	1110	7853	3400	8005	8003	7742	63596	22945	23188	111368	27294	9258	55557	192815	35408	73705	749772	45859	4529	1722	31990	21524
4	2016	7916	2252	7516	210	10883	1833	8011	23874	5298	49760	57103	36085	63028	19463	3621	67839	529737	43338	88427	203397	62970	5157	3198	47150	21377
5	2017	2196	1885	3902	284	7827	937	4956	7193	5300	39439	7486	3844	7932	2928	623	10100	26473	6897	11010	9468	2230	191	183	2322	9399
6	2018	537	7337	2322	111	3778	775	2929	2167	1944	4177	23822	11006	11238	2215	236	9756	29337	10335	14857	18614	1426	185	128	5650	7294
7	2019	990	10192	3986	1605	5395	198	13752	5641	8052	16306	32080	18876	38043	20998	6047	68202	478491	65046	32075	444380	45763	2157	1626	65380	11176
8	2020	3957	7731	6052	493	3538	68	1944	2567	2236	24121	6976	6847	20301	2414	1866	83277	82237	7293	4491	206491	263769	11884	3972	52232	35023
9	2021	2293	15004	8274	111	5078	285	9550	1320	4000	35433	3848	16051	36379	7636	769	24990	22142	-	2697	158527	34798	19544	10838	9930	22365
10	2022	14260	3518	5359	62	5916	290	20898	6852	31698	42742	42177	28848	16109	33722	5244	35741	89062	-	11139	350984	156196	85278	67232	26335	35371
11	2023	10347	4390	4497	95	4975	404	4632	4237	6443	12192	5835	5433	7542	1884	2390	34251	392873	-	33447	315534	198206	145328	32393	47794	23979
12	Total	46760	86853	51918	5219	59804	10372	78394	64494	80414	310651	213878	155833	322457	131942	32316	403706	1901567	187535	279746	2684192	834058	274390	121442	292210	194752

Fonte: Elaborado pela autora (2024).

na Figura 2.16 na janela do *Console*.

Figura 2.16: Exemplo de uma tabela apresentada na aba *Console* do *Posit Cloud*



Console | Terminal | Background Jobs

R 4.3.2 · /cloud/project/

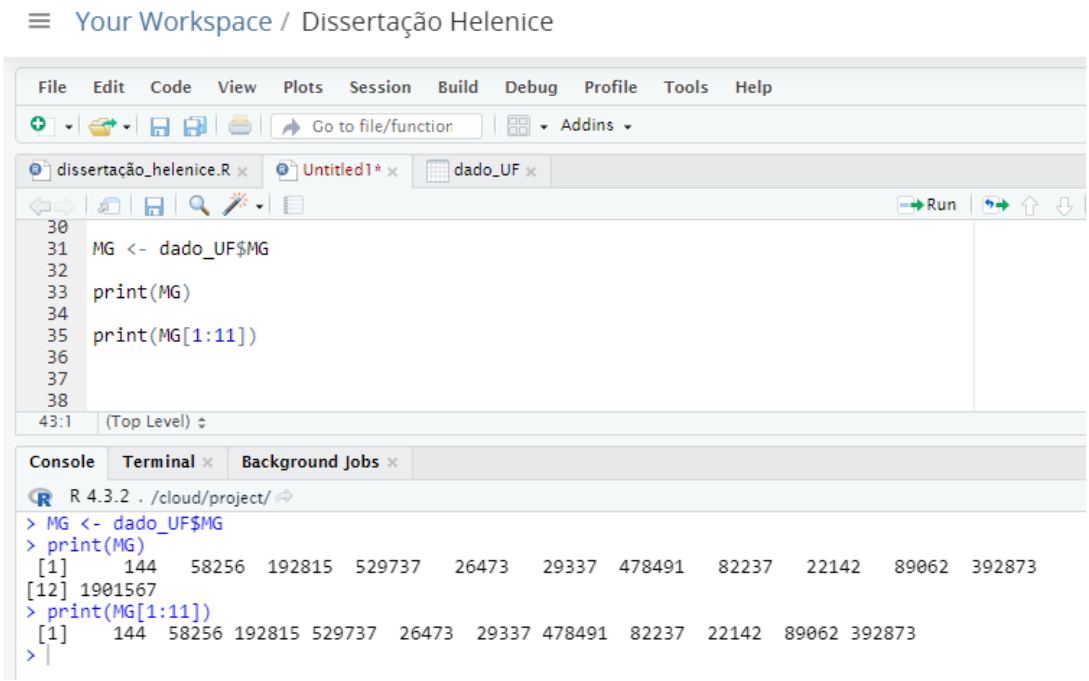
```
> dado_UF <- read.csv("UF.csv", sep = ';',dec = ',')
> View(dado_UF) #ver dados na forma de tabela
> print(dado_UF)
   Ano.notificacao    RO    AC    AM   RR    PA    AP    TO    MA    PI
1             2013    20    13    30   10    21     1    40     -     2
2             2014  1994 28464  6318 1128  4540  2181  3677  2640  7699
3             2015  2250  6067  3662 1110  7853  3400  8005  8003  7742
4             2016  7916  2252  7516  210 10883  1833  8011 23874  5298
5             2017  2196  1885  3902  284  7827   937  4956  7193  5300
6             2018   537  7337  2322  111  3778   775  2929  2167  1944
7             2019   990 10192  3986 1605  5395   198 13752  5641  8052
8             2020  3957  7731  6052  493  3538    68  1944  2567  2236
9             2021  2293 15004  8274  111  5078   285  9550  1320  4000
10            2022 14260  3518  5359   62  5916   290 20898  6852 31698
11            2023 10347  4390  4497   95  4975   404  4632  4237  6443
12           Total 46760 86853 51918 5219 59804 10372 78394 64494 80414
       CE     RN     PB     PE     AL    SE     BA     MG     ES     RJ
1      30     30     18     19     28     2     29    144    108     52
2   22855  11576   5637  10498  13360  2260  13964  58256  19110   7846
3   63596  22945  23188 111368  27294  9258  55557 192815  35408  73705
4   49760  57103  36085  63028  19463  3621  67839 529737  43338  88427
5   39439   7486   3844   7932   2928   623  10100  26473   6897  11010
6    4177  23822  11006  11238   2215   236   9756  29337  10335  14857
7   16306  32080  18876  38043  20998  6047  68202 478491  65046  32075
8   24121   6976   6847  20301   2414  1866  83277  82237   7293   4491
9   35433   3848  16051  36379   7636   769  24990  22142      -   2697
```

Fonte: Elaborado pela autora (2024).

É possível selecionar dentro da tabela gerada, uma coluna específica. Isso é feito usando o seguinte código: **Nomedoconjuntodedados$Nomedacoluna**. Os dados dessa coluna podem ser guardados em uma variável, que pode ser nomeada conforme queiramos. Com esse nome definido podemos manipular os dados. Na Figura 2.17 é possível observar um exemplo de como o código pode ser construído. No exemplo apresentado, tomamos a coluna de casos de Dengue registrados em Minas Gerais dos anos de 2013 a 2023. Pode-se filtrar os anos que queremos analisar usando a função [**númerodaPrimeiravariável:númerodaÚltimavariável**] - a primeira medida desejada e a última medida

desejada. No exemplo da Figura 2.17, MG[1:11] irá retornar apenas os números apenas o número de casos de Dengue da 1ª linha a 11ª linha, ou seja, dos anos 2013 a 2023, excluindo o total, que aparece na 12ª linha do banco de dados.

Figura 2.17: Função **print** no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Nessa dissertação, usamos variadas formas de manipulação do conjunto de dados para exemplificar as definições apresentadas.

3 Estatística Descritiva e o *software* R

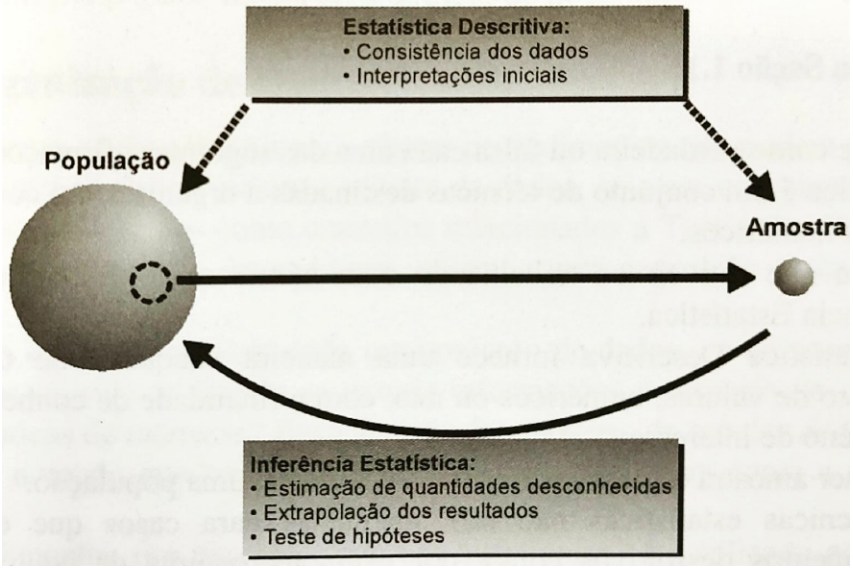
Neste capítulo serão apresentadas as atividades que compõe o produto educacional proposto, que visa proporcionar ao professor uma ferramenta para a integração do *software* R no ensino de Estatística em sala de aula. As primeiras atividades tratam dos comandos do R para se trabalhar com Estatística Descritiva. As demais atividades tratam de como se trabalhar com a apresentação de dados com o R, bem como usar um banco de dados real do DATASUS. A ideia é que o professor tenha à sua disposição um produto educacional com recursos que facilite o trabalho com o R no ensino de estatística em sala de aula utilizando dados reais. Esse produto educacional tem o potencial para auxiliar no desenvolvimento de habilidades e competências previstas na BNCC relativas à estatística e podem contribuir nas práticas dos docentes que o adotar.

3.1 Atividades de Estatística Descritiva com o R

A Estatística é uma parte da Matemática Aplicada que fornece métodos para a coleta, organização, apresentação, descrição, análise e interpretação de dados e para a utilização dos mesmos na tomada de decisões (CRESPO, 2002)[20]. Em outras palavras, Estatística é a ciência que tem como base o estudo de uma população. Esse estudo pode ser feito analisando toda a população ou uma amostra da mesma. A Estatística pode ser aplicada em praticamente todas as áreas do conhecimento humano e em algumas áreas recebe um nome especial. Um exemplo é a Bioestatística, que trata de aplicações da Estatística em Ciências Biológicas e da Saúde. A organização e a descrição dos dados estão a cargo da Estatística Descritiva. Para (BENEVIDES, 2023), "A Estatística Descritiva nos ajuda a entender as propriedades básicas de um conjunto de dados e nos fornece uma visão geral rápida sobre eles, apresentando as informações por meio de tabelas, gráficos, medidas centrais (como a média) e frequências dos dados". [3]

A Figura 3.1 ilustra as etapas da análise estatística.

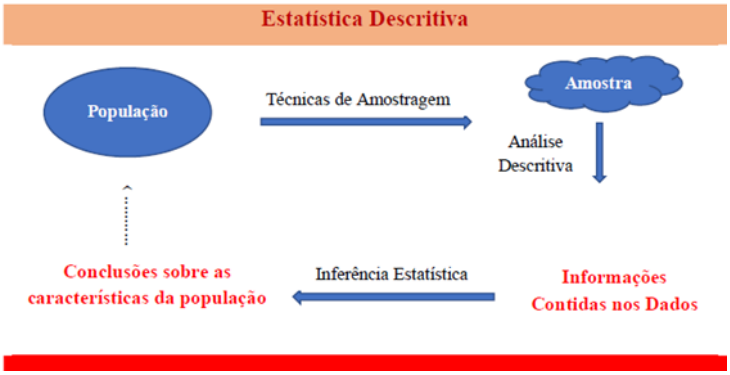
Figura 3.1: Análise Estatística



Fonte: (MAGALHÃES, 2010) [21]

Produzir afirmações sobre uma dada característica de parte da população, na qual estamos interessados, está a cargo da Estatística Descritiva. Ela reúne um conjunto de técnicas para sumarizar os dados e medidas descritivas que permitem tirar muitas informações contidas nos dados. No geral, a Estatística Descritiva é utilizada em momentos em que estamos diante de muitos dados, fazendo-se necessário tornar essas informações manejáveis para poder relacioná-las. Porém, ao simplificar as informações, pode ser introduzido um viés pela redução da informação a um único número. Este viés pode ser minimizado pela utilização, ao mesmo tempo, de medidas de tendência central e dispersão que permitem cruzar a informação e contrapor com outras leituras dos dados resumidos.

Figura 3.2: Estatística Descritiva



Fonte: (MOREIRA *apud* PINHO, 2021) [22][23]

Segundo Triola [24], Estatística é a ciência do planejamento de estudos e experimentos, da obtenção de dados e, em seguida, de sua organização, resumo, apresentação, análise e interpretação e, então, estabelecimento de conclusões com base nesses dados.

Dados são coleções de observações, tais como medidas, gêneros ou respostas de pesquisas e podem se classificados como:

- Dados quantitativos (ou numéricos) que consistem em números que representam contagens ou medidas.
- Dados qualitativos que podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica.

Os dados são a matéria prima da Estatística. Definido o assunto de interesse, os dados são obtidos da medição de determinada característica ou propriedade desse objeto, pessoa ou coisa.

Uma população é a coleção completa de todas as medidas, ou dados, a serem considerados. Um censo (contagem populacional) é a coleção dos dados obtidos de todos os membros da população. Uma amostra é uma subcoleção de elementos extraídos de uma população. O parâmetro é o resumo da característica de interesse da população. Os parâmetros são geralmente desconhecidos, já que calcular essas medidas para toda a população é impraticável na maioria dos casos. Por isso, a estatística é frequentemente usada para estimar os parâmetros [24].

Analisando o banco de dados escolhido para ser discutido nesta dissertação, vamos usar a tabela exibida na Figura 3.3 para exemplificar os conceitos acima. A tabela apresenta informações extraídas do TabNet DATASUS. Nela constam as notificações de Dengue em todos os estados da federação brasileira, uma subdivisão por regiões e o total de casos para o período de 2013 a 2023.

No nosso exemplo, temos:

- População: todos os estados brasileiros, analisados nos anos escolhidos (2013 a 2023);
- Amostra: as notificações referentes à Minas Gerais nos anos de 2013 a 2023;
- Censo: todos os dados apresentados na tabela;

Figura 3.3: Banco de Dados - TabNet

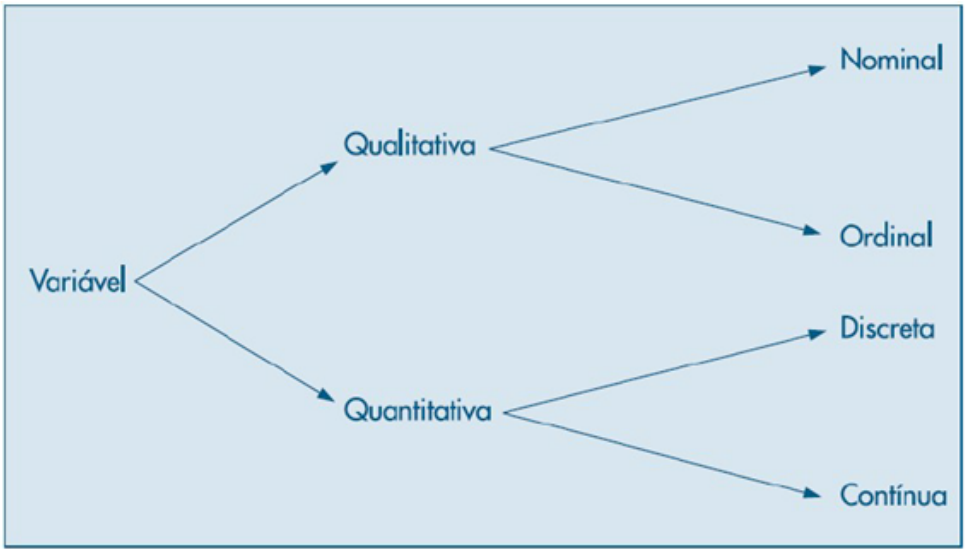
DENGUE - NOTIFICAÇÕES REGISTRADAS NO SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO - BRASIL

Casos Prováveis por Ano notificação segundo Região/UF de notificação
Período: 2014-2023

Região/UF de notificação	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
TOTAL	1.110	591.128	1.697.801	1.518.857	243.336	266.386	1.556.588	952.509	531.811	1.394.532	1.372.151	10.126.209
Região Norte	135	48.302	32.347	38.621	21.987	17.789	36.118	23.783	40.595	50.303	29.340	339.320
.. Rondônia	20	1.994	2.280	7.916	2.196	537	990	3.997	2.293	14.290	10.347	46.760
.. Acre	13	28.464	6.067	2.282	1.885	7.337	10.192	7.731	15.004	3.518	4.390	86.853
.. Amazonas	30	6.318	3.662	7.516	3.902	2.322	3.986	6.052	8.274	5.389	4.497	51.918
.. Roraima	10	1.128	1.110	210	284	111	1.605	493	111	62	95	5.219
.. Pará	21	4.540	7.853	10.883	7.827	3.778	5.395	3.838	5.078	5.916	4.975	59.804
.. Amapá	1	2.181	3.400	1.833	937	775	198	68	285	290	404	10.372
.. Tocantins	40	3.677	8.005	8.011	4.956	2.929	13.752	1.944	9.550	20.898	4.632	78.394
Região Nordeste	158	90.489	328.951	326.071	84.845	66.561	214.245	150.605	130.426	243.133	80.207	1.715.691
.. Maranhão	-	2.540	8.003	23.874	7.193	2.167	5.641	2.567	1.320	6.052	4.237	64.494
.. Piauí	2	7.699	7.742	5.298	5.300	1.944	8.052	2.236	4.000	31.698	6.443	80.414
.. Ceará	30	22.855	63.596	49.760	39.439	4.177	16.306	24.121	35.433	42.742	12.192	310.651
.. Rio Grande do Norte	30	11.576	22.945	57.103	7.486	23.822	32.080	6.976	3.848	42.177	5.835	213.878
.. Paraíba	18	5.637	23.188	36.085	3.844	11.006	18.876	6.847	16.051	28.848	5.433	155.833
.. Pernambuco	19	10.498	111.368	63.028	7.932	11.238	38.043	20.301	36.379	16.109	7.542	322.457
.. Alagoas	28	13.360	27.294	19.463	2.928	2.215	20.998	2.414	7.636	33.722	1.884	131.942
.. Sergipe	2	2.260	9.258	3.621	623	236	6.047	1.866	769	5.244	2.390	32.316
.. Bahia	29	13.964	55.557	67.839	10.100	9.756	68.202	83.277	24.990	36.741	34.251	403.706
Região Sudeste	360	312.181	1.051.700	864.899	53.848	73.143	1.019.992	300.512	183.366	451.185	741.854	5.053.040
.. Minas Gerais	144	58.256	192.815	529.737	26.473	29.337	478.491	82.237	22.142	89.062	392.873	1.901.567
.. Espírito Santo	108	19.110	35.408	43.338	6.897	10.335	65.046	7.293	-	-	-	187.535
.. Rio de Janeiro	52	7.846	73.705	88.427	11.010	14.857	32.075	4.491	2.697	11.139	33.447	279.746
.. São Paulo	56	226.969	749.772	203.397	9.468	18.614	444.380	206.491	156.527	350.984	315.534	2.684.192
Região Sul	66	23.062	52.110	71.325	2.604	1.739	49.546	279.625	65.180	308.706	375.927	1.229.890
.. Paraná	66	22.775	45.859	62.970	2.230	1.426	45.763	262.769	34.798	156.196	198.206	834.058
.. Santa Catarina	-	137	4.529	5.167	191	185	2.157	11.884	19.544	86.278	146.328	274.390
.. Rio Grande do Sul	-	150	1.722	3.198	183	128	1.626	3.972	10.838	67.232	32.393	121.442
Região Centro-Oeste	391	117.094	232.693	217.928	80.052	107.154	236.687	197.984	112.244	341.205	144.823	1.788.255
.. Mato Grosso do Sul	30	3.397	31.990	47.150	2.322	5.650	65.380	52.232	9.930	26.335	47.794	292.210
.. Mato Grosso	61	7.183	21.524	21.377	9.399	7.294	11.176	35.023	22.365	35.371	23.979	194.752
.. Goiás	291	92.905	168.934	129.786	64.080	91.728	120.548	59.873	61.242	206.808	50.391	1.046.598
.. Distrito Federal	9	13.609	10.245	19.603	4.251	2.482	39.583	50.856	18.707	72.691	22.659	254.695
Ignorado/Exterior	-	-	-	13	-	-	-	-	-	-	-	13

Fonte: DATASUS TabNet[14]

Figura 3.4: Tipos de variável



Fonte: Bussab 2017 [25]

As variáveis são as características de interesse do estudo e podem ser classificadas de acordo com a Figura 3.4

Dentre as variáveis qualitativas, temos a variável qualitativa nominal, para a qual não existe nenhuma ordenação nas possíveis respostas, e a variável qualitativa ordinal, para a qual existe uma ordem nos seus resultados. De modo análogo, as variáveis quantitativas são classificadas de duas maneiras: as variáveis quantitativas discretas, cujos possíveis valores formam um conjunto finito ou enumerável de números, e que resultam, frequentemente, de uma contagem; e as variáveis quantitativas contínuas, cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração [25].

3.1.1 Medidas Resumo

As medidas resumo (posição e dispersão) representam uma síntese do conjunto de dados observados. Como as distribuições podem apresentar diferentes formas, é essencial conhecer os diversos tipos de medidas resumo e utilizá-las de forma adequada em cada situação [24]. A fim de ressaltar as características de cada distribuição e compará-las, faz-se necessário obter tais medidas através de números. Assim, temos os seguintes conceitos:

3.1.2 Medidas de Posição

Medidas de posição, como o próprio termo indica, visam a resumir um conjunto de dados em geral numa única medida em algum lugar geométrico entre os extremos observados do conjunto (mínimo e máximo). As medidas de posição mais importantes são as medidas de tendência central, "devido aos dados observados tenderem, em geral, a se agrupar em torno dos valores centrais"(CRESPO,2002 p.79) [20].

Média aritmética Simples

É a soma das observações dividida pela quantidade observada.

Se x_1, x_2,...e x_n são os n valores (distintos ou não) da variável X, a média aritmética, ou simplesmente média, de X pode ser escrita por:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Se os dados são de uma amostra de uma população, a média é representada por $\overline{X}$. Se os dados representam a população inteira, então representamos a média por μ (letra grega minúscula). Estatísticas amostrais são usualmente representadas por letras do alfabeto latino, como $\overline{X}$ e os parâmetros populacionais são, em geral, representados por letras gregas, como μ.

Como exemplo, tomando o banco de dados selecionado no DATASUS - Os casos registrados de Dengue no Brasil nos últimos 11 anos, por Unidade Federativa.

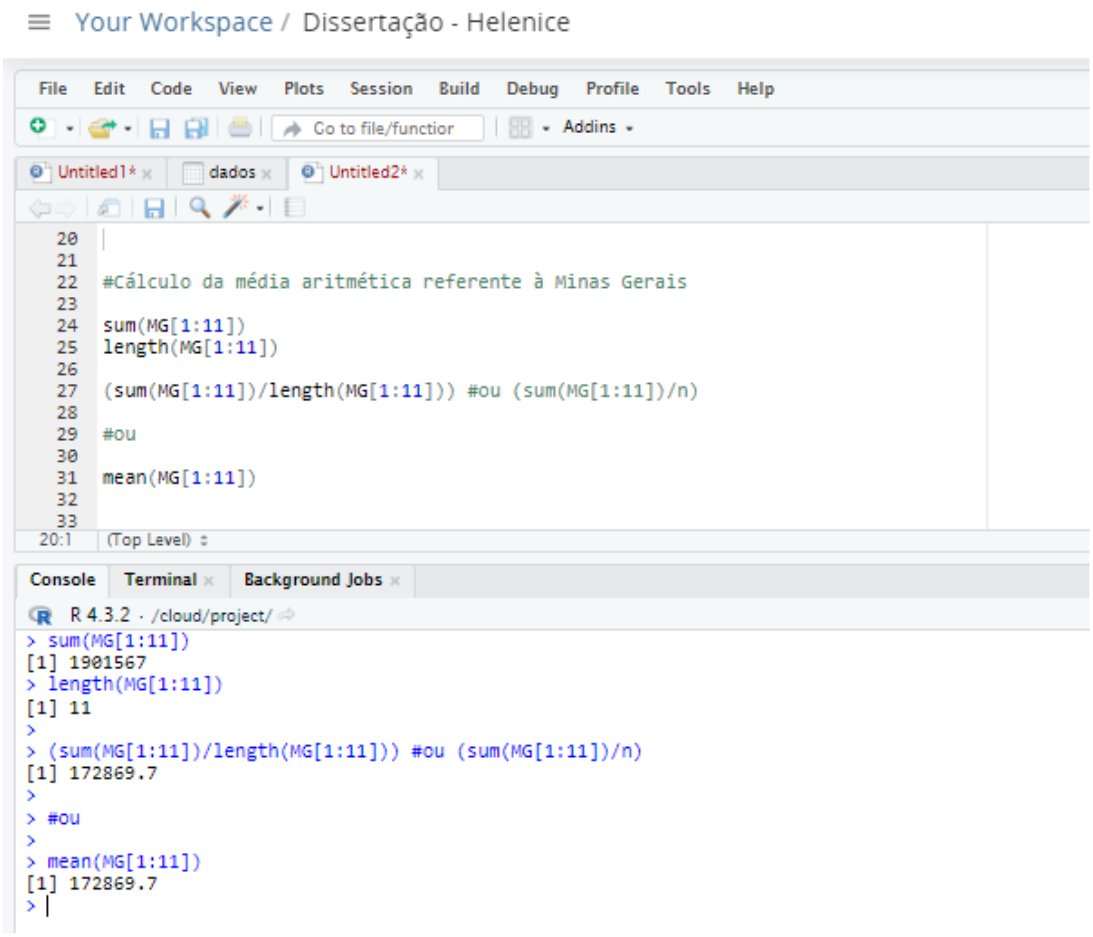
A média aritmética de um conjunto de dados é calculada somando-se todos essas dados e dividindo esse resultado pela quantidade de dados. Na plataforma *Posit Cloud*, podemos fazer esse cálculo de duas maneiras distintas:

- Usando a função **mean()**.

- Usando as funções **sum()** e **length()**. Essas funções retornam, respectivamente, a soma dos dados e a quantidade de dados. Assim, basta escrevermos a fórmula **(sum()/length())** para encontrarmos a média aritmética.

Na manipulação do banco de dados, foi calculada a média dos casos de Dengue referentes ao estado de Minas Gerais nos anos de 2013 a 2023, conforme podemos observar na Figura 3.5.

Figura 3.5: Média aritmética no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

O resultado encontrado - 172869,7 - representa a quantidade média de casos de Dengue no estado de Minas Gerais entre os anos 2013 a 2023.

Média ponderada

Quando se associam pesos diferentes p aos valores dos dados x, podemos calcular uma média ponderada. Sejam x_1,x_2, ..., x_n os n valores da variável X e p_1, p_2, ..., p_n pesos associados a x_1,x_2, ..., x_n, respectivamente. A média aritmética ponderada será

dada por:

$$\overline{X}_p = \frac{x_1.p_1 + x_2.p_2 + \ \ldots + x_n.p_n}{p_1 + p_2 + \ \ldots p_n} = \frac{\sum_{i=1}^{n} x_i.p_i}{\sum_{i=1}^{n} p_i}$$

Quando todos os elementos têm $p = 1$, ou seja, o peso 1 na média ponderada, isso implica que cada elemento contribui igualmente para o cálculo da média. Dessa forma, a média ponderada se torna equivalente à média aritmética simples, onde cada elemento tem o mesmo peso e é somado e dividido pelo total de elementos.

O banco de dados escolhido, possui um conjunto de variáveis e dados com mesmo peso $p = 1$. Nesse caso, é irrelevante pensar na média ponderada, tendo em vista que ela retornará o mesmo resultado encontrado na média aritmética simples, calculada anteriormente.

Para fins de exemplificar o cálculo da média aritmética ponderada, vamos considerar a seguinte situação: Um aluno realizou quatro atividades diferentes em uma disciplina, valendo 10 pontos cada uma. O peso atribuído a cada atividade foi:

- Prova escrita (peso 3): 8 pontos
- Trabalho em grupo (peso 2): 7 pontos
- Projeto individual (peso 4): 6 pontos
- Participação em sala (peso 1): 9 pontos

Para descobrir a nota média desse aluno referente às quatro atividades, vamos calcular a média ponderada. Multiplicamos cada nota pelo seu respectivo peso e somamos os resultados, dividindo pelo total dos pesos. Assim:

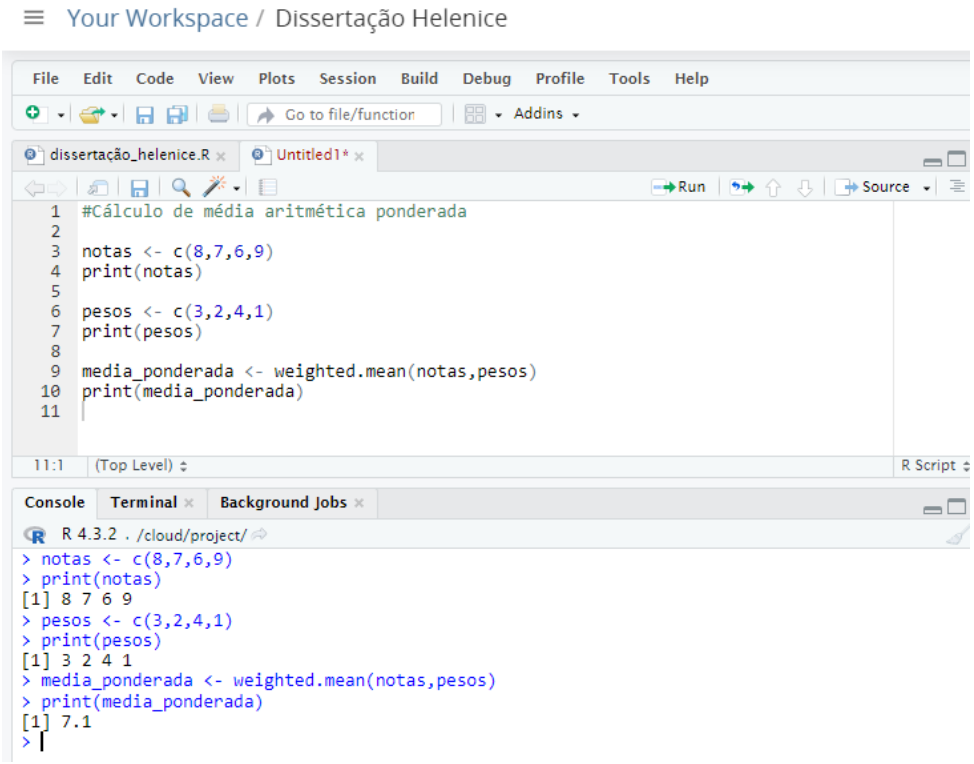
$$\overline{X} = \frac{8.3 + 7.2 + 6.4 + 9.1}{3 + 2 + 4 + 1} = \frac{71}{10} = 7.1$$

A média aritmética ponderada das notas desse aluno é 7.1 em um total de 10 pontos.

Esse cálculo no *Posit Cloud*, será feito conforme a Figura 3.6. Para isso, usamos a função **weighted.mean()**. Ela retorna a média aritmética ponderada usando dois vetores - *valores* e *pesos*. No caso do exemplo trabalhado, os valores são as notas das atividades. Observe que o resultado encontrado foi exatamente o mesmo quando feito utilizando a

fórmula apresentada acima. Para o nosso exemplo, com apenas 4 dados, esse cálculo foi feito de maneira ágil, porém, para o caso de um conjunto de dados muito extenso, torna-se interessante utilizar uma ferramenta que facilite essa operação. Assim, é pertinente a utilização da linguagem *R*.

Figura 3.6: Média Ponderada no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Mediana

A mediana pode ser considerada como um "valor do meio", no sentido de que cerca de metade dos valores no conjunto de dados é menor do que a mediana e a outra metade é maior do que ela. É a medida que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente. Segundo (IEZZI, 2013), [26]

> A média aritmética pode ser muito afetada quando encontramos valores discrepantes em um conjunto de dados, podendo se tornar uma medida de centralidade pouco representativa do resumo de dados. Para contornar questões dessa natureza definiremos uma medida de centralidade mais resistente aos valores discrepantes (...) denominada mediana.

Para encontrar a mediana, primeiro ordene os valores e depois siga um dos dois procedimentos seguintes:

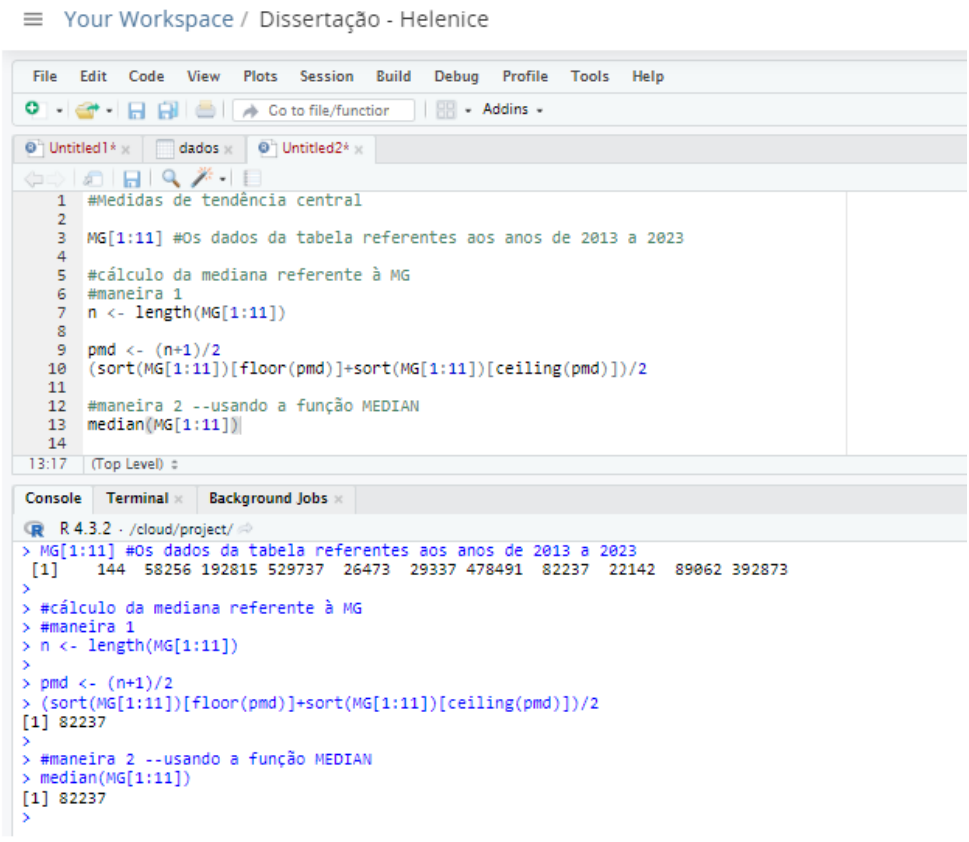
- Se o número de valores for ímpar, ou seja, n é ímpar, a mediana será o número localizado no meio exato da lista.

- Se o número de valores for par, a mediana será encontrada pelo cálculo da média dos dois números do meio.

Assim como ocorreu com o cálculo da média aritmética, temos duas maneiras de encontrar a mediana de um conjunto de dados no *software* R.

- Uma forma de localizar a mediana, é utilizando os passos descritos abaixo, que também podem ser observados na Figura 3.7:

Figura 3.7: Mediana no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Executa-se a função **length()** para encontrar a quantidade de dados e atribui-se um nome para o resultado - chamamos de **n**. Em seguida, encontra-se a posição do dado correspondente à mediana, usando a seguinte função, **pmd <- (n+1)/2**. A função **(sort()[floor(pmd)]+sort()[celling(pmd)])/2** nos fornece a mediana.

- A maneira mais simples é usar a função **median()**. Que já imprime a mediana dos dados.

O banco de dados analisado, consta de 11 dados. Assim, o valor encontrado é a quantidade referente ao ano de 2020. Então, podemos concluir que a mediana dos casos de Dengue nos anos de 2013 a 2023 ocorreu no ano de 2020 e corresponde à 82237 casos registrados.

Moda

A moda é definida como a realização mais frequente do conjunto de valores observados. Um conjunto de dados pode ter uma moda, chamado unimodal, ou mais de uma moda, ou nenhuma moda. Quando dois valores ocorrem com maior frequência, cada um é uma moda, e o conjunto de dados é bimodal. Quando mais de dois valores ocorrem com maior frequência, cada um é uma moda, e o conjunto de dados é multimodal. Quando nenhum valor se repete, dizemos que não há moda.

No *Posit Cloud*, para identificar a moda, é necessário primeiro criar uma tabela de frequência dos resultados da variável. Isso pode ser feito utilizando a função **table()**.

Como podemos observar na Figura 3.8, a função foi utilizada para encontrarmos a moda referente aos dados de Minas Gerais. Ela retorna o número de vezes que cada quantidade de notificações foi observada. Como cada medida foi observada uma única vez, não há moda. O *Posit Cloud*, nos permite ainda, fazer o cruzamento de dados de colunas distintas. Foi feito o cruzamento dos dados dos estados do Sergipe e Piauí, afim de exemplificar os dados, observa-se que o conjunto completo não possui moda.

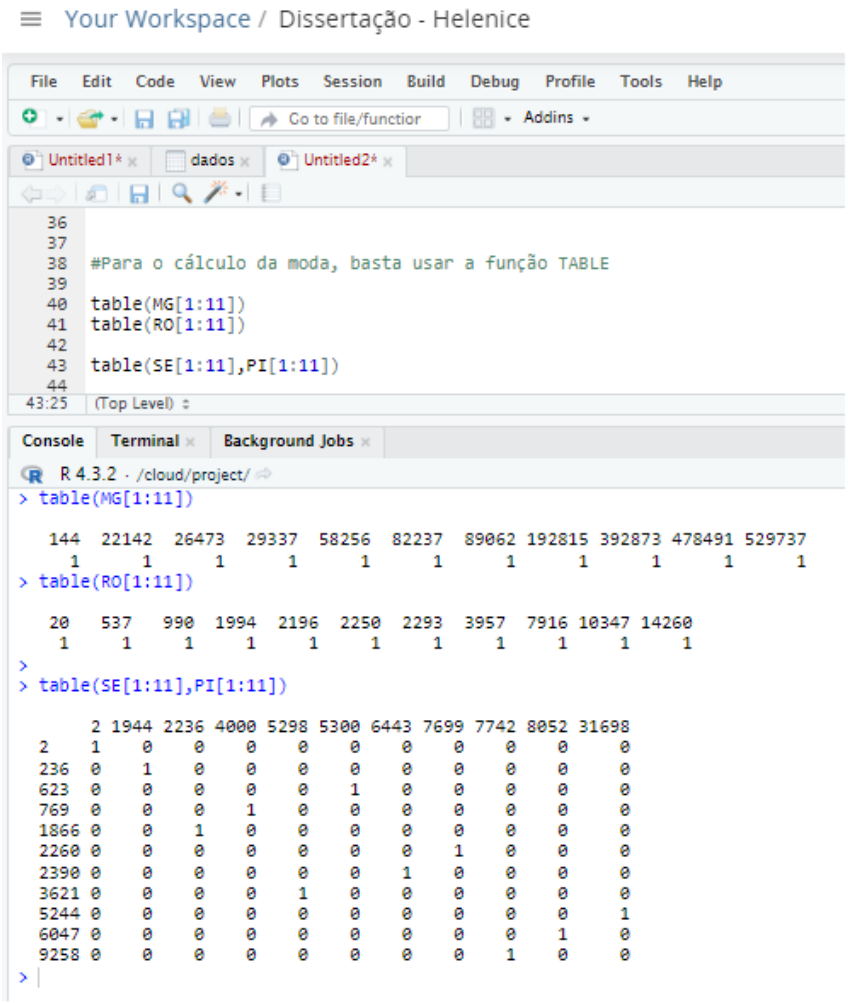
3.1.3 Separatrizes ou medidas de posição relativa

As separatrizes (ou quantis) descrevem a posição relativa, em termos de frequência, de um determinado valor na amostra. Por isso, também são chamadas de medidas de posição relativa. As consideradas mais importantes são os quartis, os decis e os percentis.

Quartil

Os quartis são medidas que dividem a amostra em 4 partes com frequência 1/4 em cada. que são Q_1, Q_2 *e* Q_3, sendo primeiro, segundo e terceiro quartil, respectivamente. O primeiro quartil é o valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 25% do número de observações do conjunto de dados e, consequentemente,

Figura 3.8: Moda no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

acima dele, é 75% do número de observações do conjunto de dados. O segundo quartil é equivalente à mediana, é o valor da distribuição que separa os 50% dos valores ordenados inferiores dos 50% superiores. O terceiro quartil é o valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 75% do número de observações do conjunto de dados. Ele separa os 75% valores ordenados inferiores dos 25% superiores.

Decil

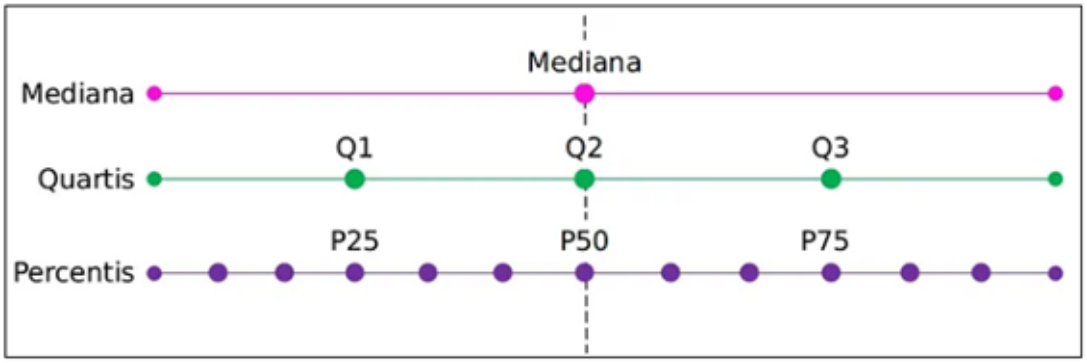
Os decis são medidas que dividem a amostra em 10 partes com frequência 1/10. Estas são denotadas por D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9. O intervalo entre cada decil representa 10% dos dados da amostra. O quinto decil representa a mediana.

Percentil

Percentis são medidas que dividem a amostra em 100 partes iguais, denotadas por P_1, P_2, P_3, ... , P_{99}. Tais partes representam 1% dos valores da amostra em cada um deles. O 50° percentil, denotado por P_{50}, tem cerca de 50% dos valores de dados abaixo dele e cerca de 50% dos valores de dados acima dele. Assim, o 50° percentil é o mesmo que a mediana.

A Figura 3.9 mostra um esquema que relaciona quartis e percentis. Nele fica evidente a mediana como o segundo quartil - Q_2 e o quinquagésimo percentil - P_{50}.

Figura 3.9: Medidas de posição relativa



Fonte: Disponível em: https://medium.com/pyladiesbh/estat [27]

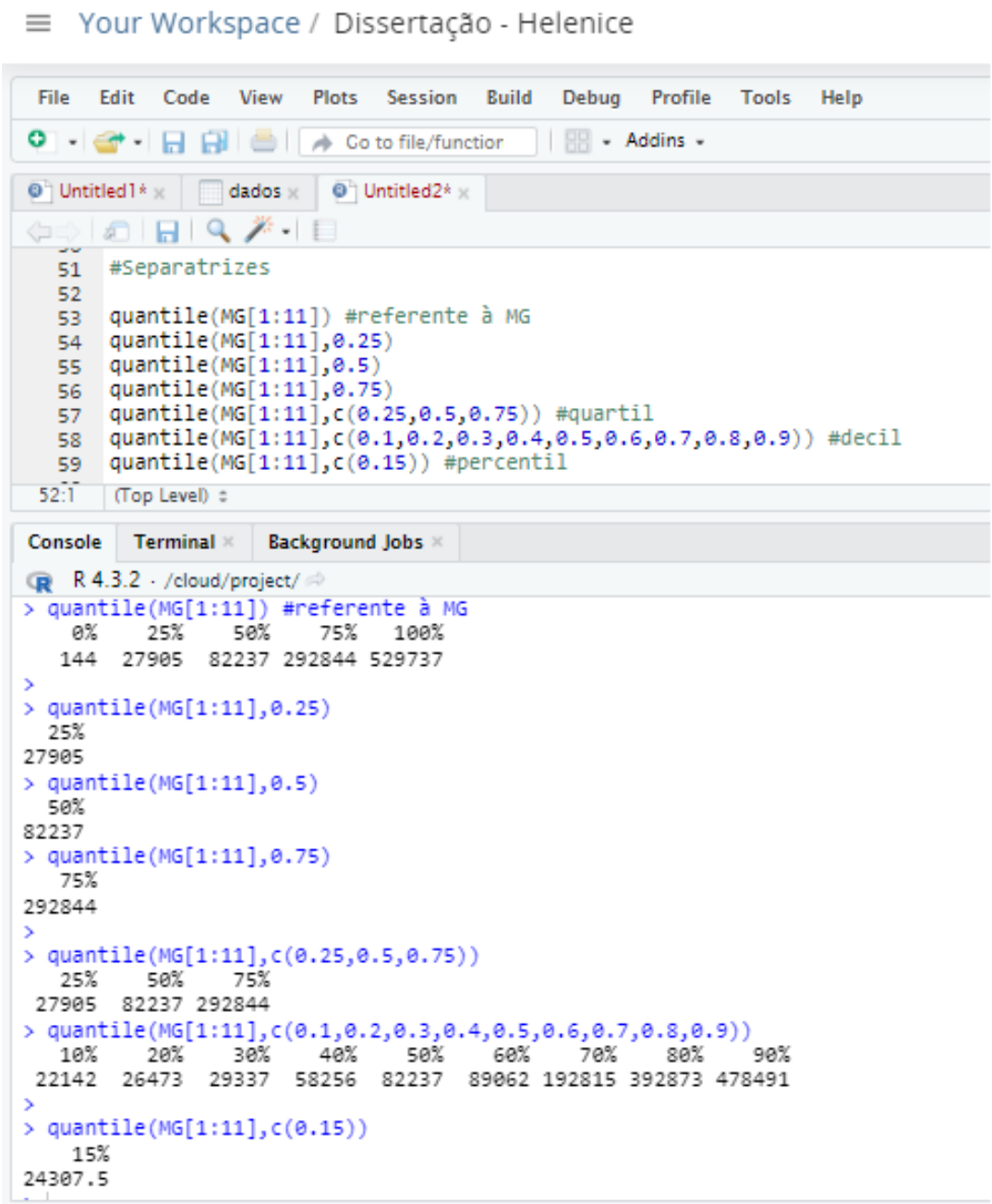
As medidas de posição relativa no *Posit Cloud*

As medidas de posição relativa ou separatrizes, são calculadas no *Posit Cloud* usando a função **quantile()**. Ela é usada inserindo o dado a ser analisado e em seguida coloca-se a partição desejada. Assim,

- Usando apenas a função **quantile()** com os dados desejados, o *software* R imprimirá o menor valor 0%, o maior, 100% e os três quartis Q_1, Q_2 *e* Q_3, respectivamente, 25%, 50% e 75%.

Na Figura 3.10, vemos que é possível encontrar valores e intervalos específicos. Quando usamos a função **quantile(dado, 0.5)** obtemos o Q_1, por exemplo. Podemos ainda, localizar decis e percentis utilizando a mesma função, porém, ao invés de uma separatriz, podemos indicar um vetor com mais posições, ficando a função assim, **quantile(dado, c(0.1,0.2,0.3,0.4,0.5))**. Nesse caso, no segundo argumento da função, é especificado um valor **q**, tal que irá retornar a medida de separatriz em que $\mathbf{q} * 100\%$ das observações são menores ou iguais. O resultado será, então D_1, D_2,D_3,D_4,D_5.

Figura 3.10: Medidas de Posição Relativa no *Posit Cloud* - MG

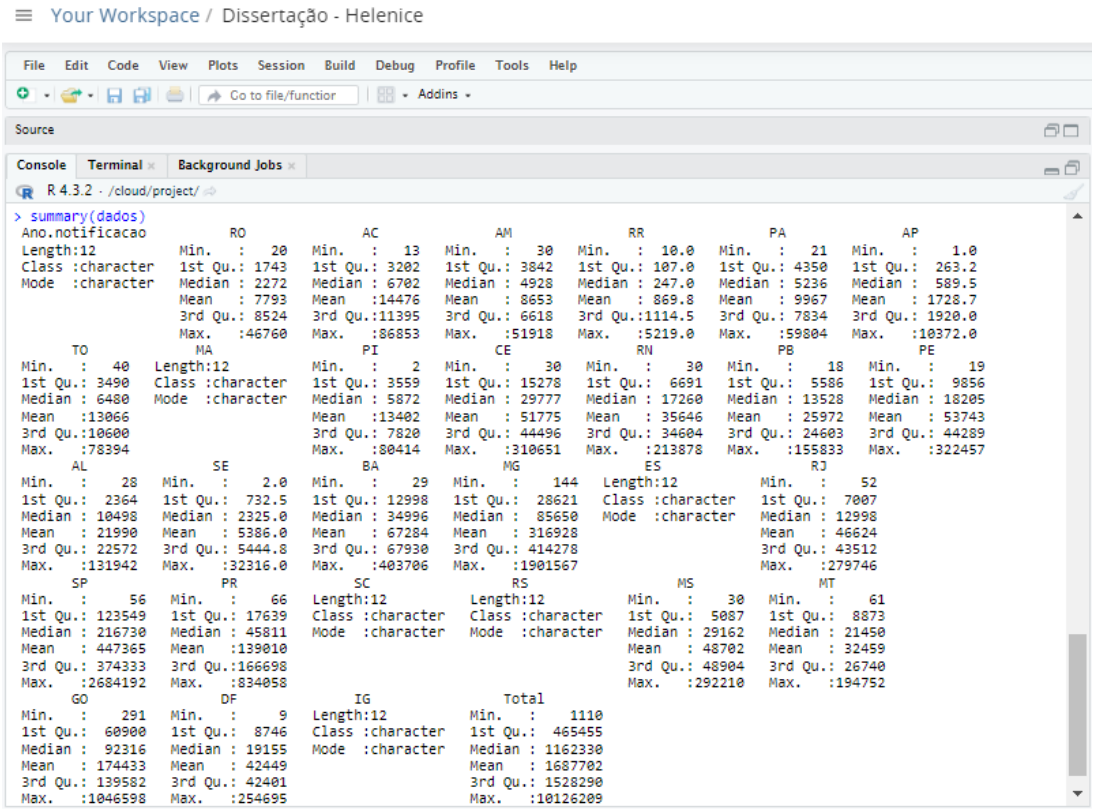


Fonte: Elaborado pela autora (2024).

Se usamos a função **summary()** um detalhamento do banco de dados nos dá mais informações, retorna um conjunto de medidas resumo. A Figura 3.11 representa esse detalhamento para todos os estados brasileiros.

Apresentam-se nessas medidas, *Min.*, que é a menor quantidade de notificações registradas. O *1st Qu.* é o primeiro Quartil. A *Median* e a *Mean* é a mediana e a média aritmética, respectivamente, assim como nas funções **median()** e **mean()** já citadas anteriormente. O terceiro Quartil é *3rd Qu* e *Max.* é a maior quantidade de notificações

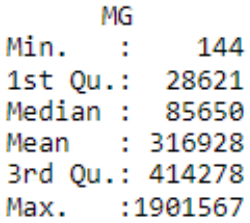
Figura 3.11: Medidas de Posição Relativa na plataforma *Posit Cloud* - Brasil por UF



Fonte: Elaborado pela autora (2024).

registradas. Conforme a imagem abaixo das informações sobre Minas Gerais.

Figura 3.12: Zoom da figura 3.11 - MG



Fonte: Elaborado pela autora (2024)

3.1.4 Medidas de dispersão ou variabilidade

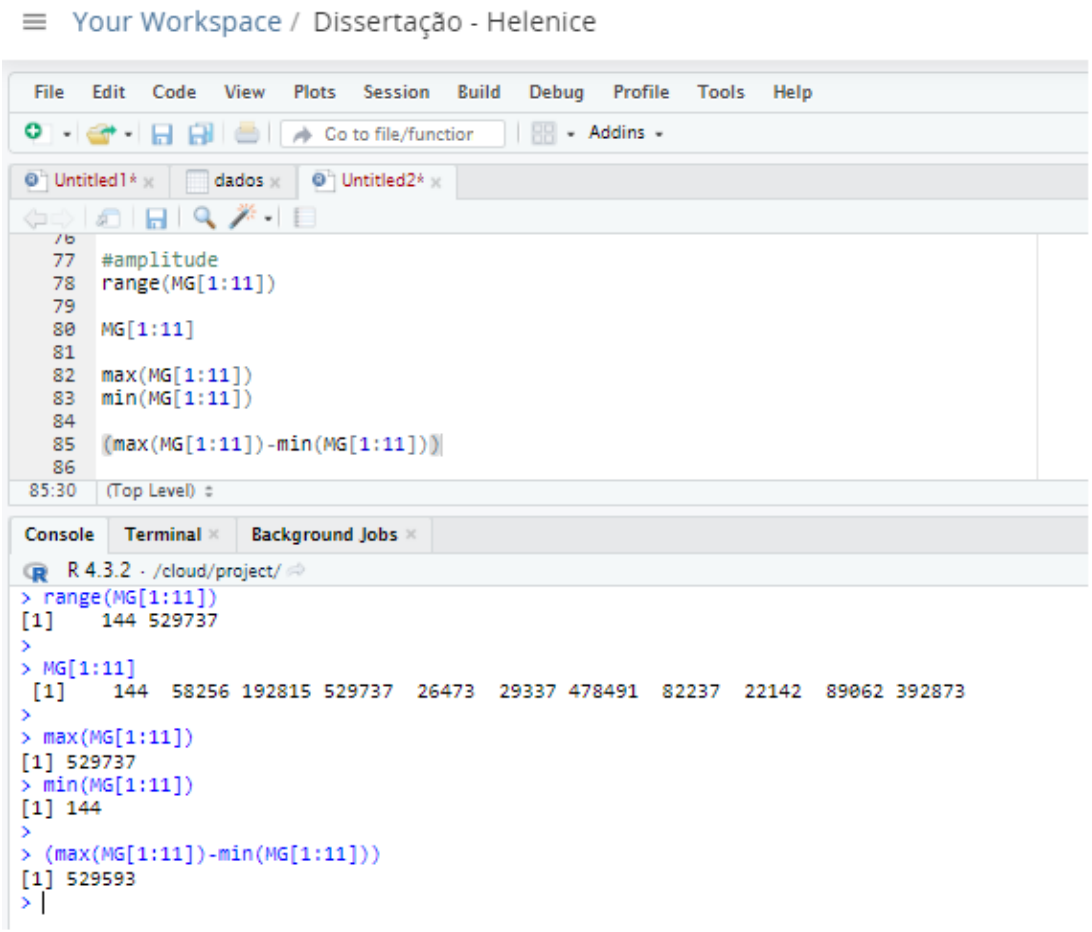
Enquanto as medidas de posição procuram resumir o conjunto de dados em alguns valores situados entre dados coletados, as medidas de dispersão buscam avaliar quão dispersos eles estão. Apresentaremos três medidas importantes de variação: amplitude,

variância e desvio-padrão.

3.1.5 Amplitude

A amplitude de um conjunto de valores de dados é a diferença entre o maior e o menor valor. Na Figura 3.13, vemos a função **range()**. Ela imprime o menor e o maior valor dos dados do conjunto analisado. Fazendo a diferença desses dois valores obtemos a amplitude. Podemos, também, usar as funções **min()**, para encontrar o valor mínimo e **max()**, para encontrar o valor máximo e, em seguida, fazer a diferença **(max()-min())**.

Figura 3.13: Amplitude na plataforma *Posit Cloud*



Fonte: Elaborado pela autora (2024).

3.1.6 Variância

Variância é a medida de variação em relação à média. A unidade básica de variação utilizada no cálculo dessa medida é o desvio de cada observação em relação à média $X_i - \overline{X}$. A variância amostral de um conjunto de valores é uma medida de variação igual

ao quadrado do desvio-padrão. Ela pode ser calculada pela seguinte fórmula $\rho = s^2$. Ou ainda,

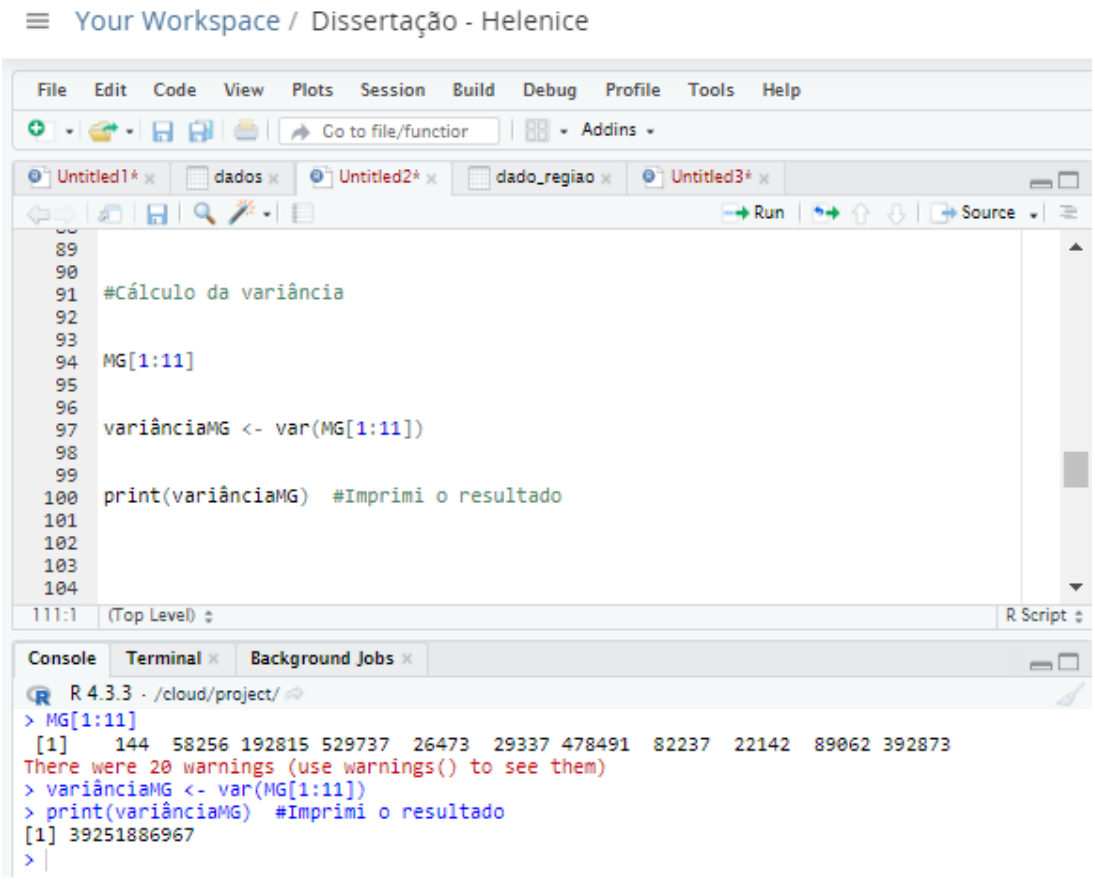
$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

sendo $\overline{X}$ a média aritmética dos dados observados.

A unidade de medida encontrada é o quadrado da unidade de medida de uma variável. Segundo Triola [24],

> Variância é uma estatística usada em alguns métodos estatísticos, mas, para nossos objetivos no momento, a variância tem a séria desvantagem de usar unidades diferentes das unidades dos dados originais. Isso torna difícil entender como ela se relaciona com o conjunto de dados original. Por causa dessa propriedade, enfatizaremos o desvio-padrão ao tentar desenvolver uma compreensão de variação.

Figura 3.14: Cálculo da variância no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

A função usada para calcular a variância no *Posit Cloud* é **var()**. Podemos criar

um vetor de dados ou inserir um conjunto pronto. Como podemos observar na Figura 3.14, definido o conjunto dos dados, basta lançá-los na função. Para as notificações registradas em Minas Gerais, de 2013 a 2023, a variância encontrada foi 39251886967.

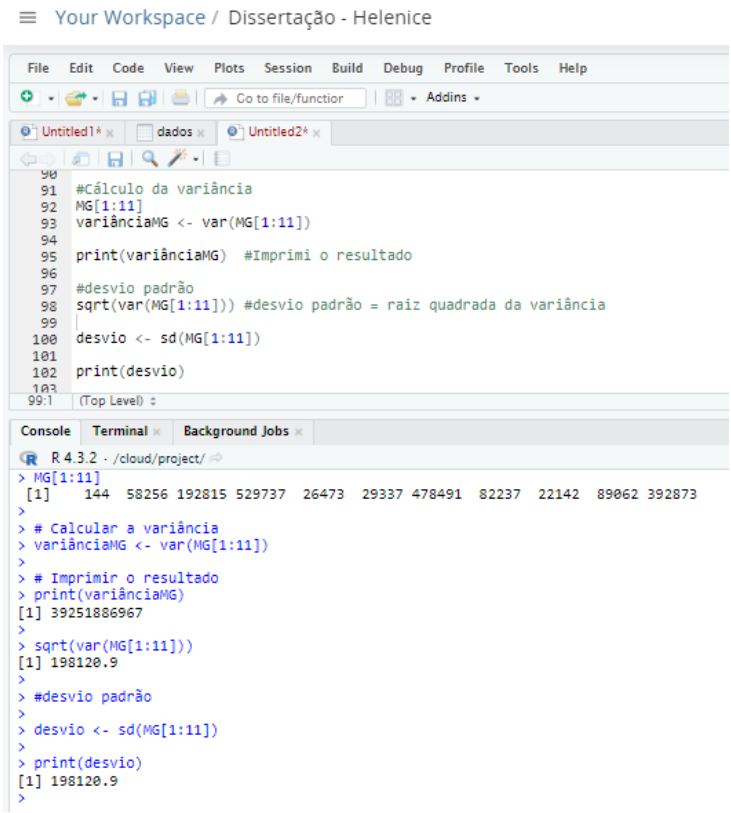
3.1.7 Desvio padrão

O desvio-padrão de um conjunto de valores amostrais, denotado por s, é uma medida de quanto os valores se afastam em média e é calculado pela seguinte fórmula,

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

sendo $\overline{x}$ a média aritmética dos n dados x. Ele é sempre positivo e igual a zero apenas quando todos os valores do conjunto de dados são iguais. Também, maiores valores de s indicam maior variação. O valor do desvio-padrão s pode crescer drasticamente com a inclusão de um ou mais valores atípicos (valores de dados que estão muito afastados dos demais).

Figura 3.15: Cálculo da Variância e Desvio Padrão na plataforma *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Para o cálculo do Desvio Padrão, fazemos a raiz quadrada da Variância. No *Posit Cloud*, as funções **sqrt(var())** e **sd()** imprimem esse resultado, como podemos observar na Figura 3.15. Para as notificações registradas em Minas Gerais, de 2013 a 2023, o desvio padrão encontrado foi 198120.9.

3.1.8 Coeficiente de variação

Nem sempre uma variância pequena (e consequentemente desvio-padrão pequeno) significa pouca dispersão. Tampouco uma variância grande é sempre indicador de alta dispersão. Esses valores podem ser altos ou baixos devido à magnitude (ordem de grandeza) dos dados observados. O coeficiente de variação amostral é uma medida usada para calcular a variação relativa dos dados de um conjunto em torno da média: quanto maior seu valor, maior é a variação relativa em torno da média. Em geral, ele é calculado em termos percentuais da seguinte forma,

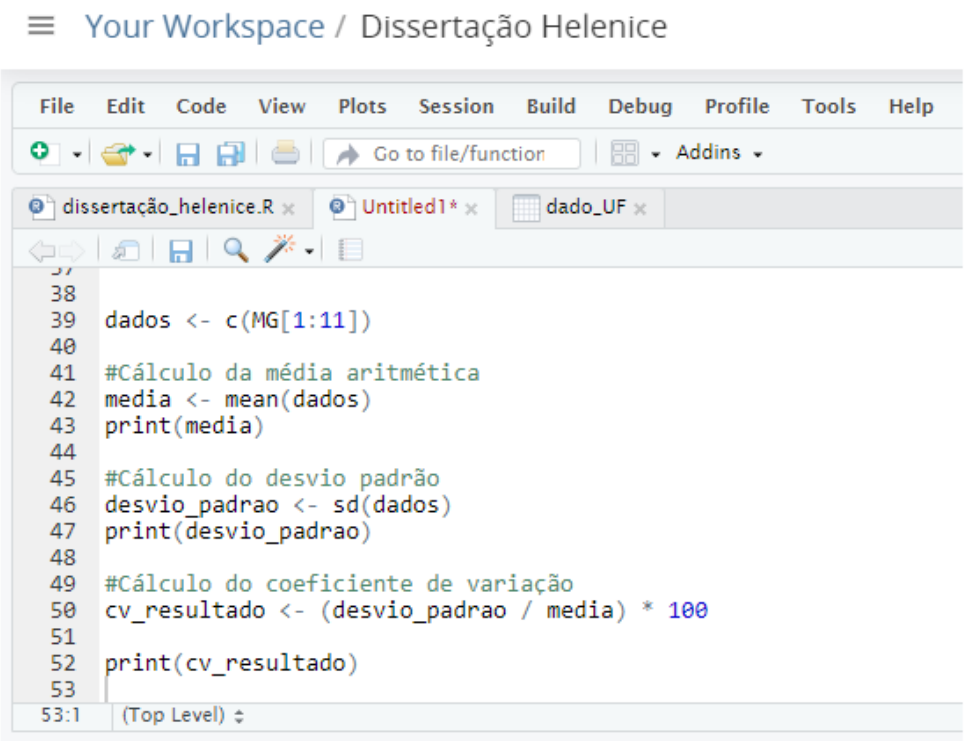
$$CV = \frac{s}{\overline{X}}.100\% \tag{3.1}$$

em que s é o desvio padrão amostral e $\overline{X}$ é a média amostral. Observe que o coeficiente de variação só é definido para conjuntos cuja média é diferente de zero.

No *Posit Cloud*, o Coeficiente de Variação é encontrado fazendo a razão entre a média aritmética e o desvio padrão. A Figura 3.16 representa o código que descreve o Coeficiente de Variação.

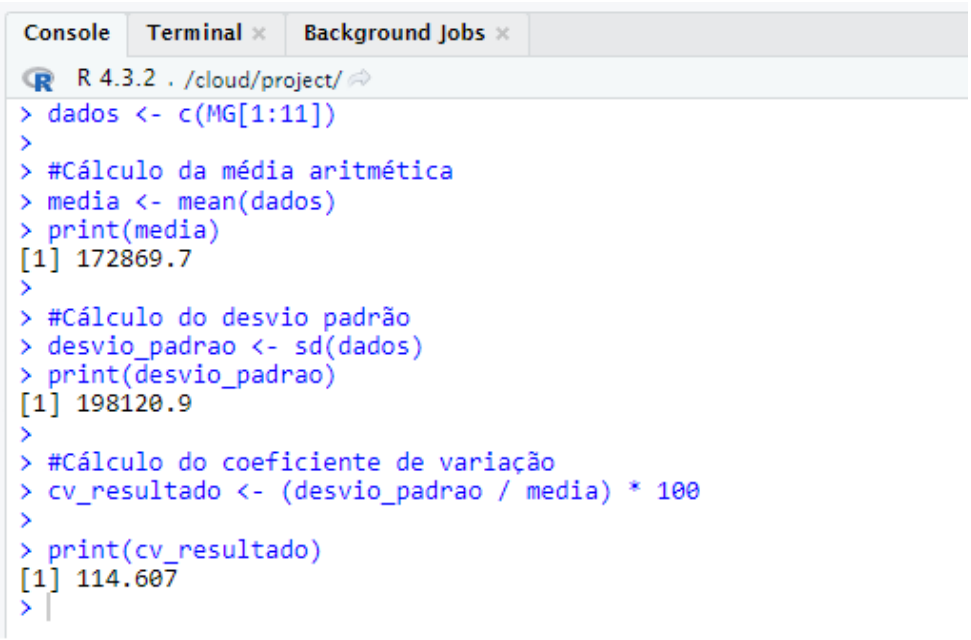
A Figura 3.17 mostra o resultado encontrado quando executamos o código acima.

Figura 3.16: Código - Coeficiente de Variação *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Figura 3.17: Resultado para o código - Coeficiente de Variação *Posit Cloud*



Fonte: Elaborado pela autora (2024).

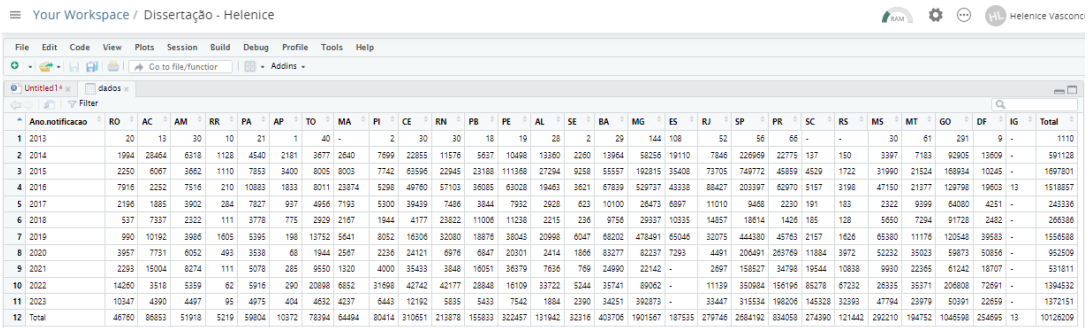
4 Atividades de apresentação de dados com o R

4.1 Tabelas

As tabelas são muito comuns quando estamos tratando da organização de um conjunto de dados. Me arrisco a dizer que essa é a primeira forma que organizamos os dados quando iniciamos uma manipulação estatística. Às vezes chamadas de tabelas de frequências simples, são uma maneira fundamental de organizar e resumir dados em estatística. Elas exibem a contagem ou frequência de cada categoria ou valor em um conjunto de dados. Geralmente, as tabelas simples consistem em duas colunas: uma para os valores possíveis ou categorias e outra para as contagens ou frequências correspondentes. De maneira direta, as tabelas associam variáveis.

No *Posit Cloud*, quando importamos dados externos do **Excel** - o que foi feito com o banco de dados referente à Dengue no Brasil, o *software* já faz uma primeira leitura e gera uma tabela de dados, conforme se vê na Figura 4.1. Tal tabela é gerada no espaço **Editor de código** da interface do *software*.

Figura 4.1: Exemplo de uma tabela simples no *Posit Cloud*



	Ano.notificacao	RO	AC	AM	RR	PA	AP	TO	MA	PI	CE	RN	PB	PE	AL	SE	BA	MG	ES	RJ	SP	PR	SC	RS	MS	MT	GO	DF	IG	Total
1	2013	20	13	30	10	21	1	40	-	2	30	30	18	19	28	2	29	144	108	52	56	66	-	-	30	61	291	9	-	1110
2	2014	1994	28464	6318	1128	4540	2181	3677	2640	7699	22855	11576	5637	10498	13360	2260	13964	58256	19110	7846	226959	22775	137	150	3397	7183	92905	13609	-	591128
3	2015	2250	6067	3662	1110	7853	3400	8005	8003	7742	63596	22945	23188	111368	27294	9258	55557	192815	35408	73705	749772	45859	4529	1722	31990	21524	168934	10245	-	1697801
4	2016	7916	2252	7516	210	10883	1833	8011	23874	5298	49760	57103	36085	63028	19463	3621	67839	529737	43338	88427	203397	62970	5157	3198	47150	21377	129798	19603	13	1518857
5	2017	2196	1885	3902	284	7827	937	4956	7193	5300	39439	7486	3844	7932	2928	623	10100	26473	6897	11010	9468	2230	191	183	2322	9399	64080	4251	-	243336
6	2018	537	7337	2322	111	3778	775	2929	2167	1944	4177	23822	11006	11238	2215	236	9756	29337	10335	14857	18614	1426	185	128	5650	7294	91728	2482	-	266386
7	2019	990	10192	3986	1605	5395	198	13752	5641	8052	16306	32080	18876	38043	20998	6047	68202	478491	65046	32075	444380	45763	2157	1626	65380	11176	120548	39583	-	1556588
8	2020	3957	7731	6052	493	3538	68	1944	2567	2236	24121	6976	6847	20301	2414	1866	83277	82237	7293	4491	206491	263769	11884	3972	52232	35023	59873	50856	-	952509
9	2021	2293	15004	8274	111	5078	285	9550	1320	4000	35433	3848	16051	36379	7636	769	24990	22142	-	2697	158527	34798	19544	10838	9930	22365	61242	18707	-	531811
10	2022	14260	3518	5359	62	5916	290	20898	6852	31698	42742	42177	28848	16109	33722	5244	35741	89062	-	11139	350984	156196	85278	67232	26335	35371	206808	72691	-	1394532
11	2023	10347	4390	4497	95	4975	404	4632	4237	6443	12192	5835	5433	7542	1884	2390	34251	392873	-	33447	315534	198206	145328	32393	47794	23979	50391	22659	-	1372151
12	Total	46760	86853	51918	5219	59804	10372	78394	64494	80414	310651	213878	155833	322457	131942	32316	403706	1901567	187535	279746	2684192	834058	274390	121442	292210	194752	1046598	254695	13	10126209

Fonte: Elaborado pela autora (2024).

4.2 Distribuição de frequências ou Tabela de frequências

Ao trabalhar com grandes conjuntos de dados, as distribuições de frequência (ou tabela de frequências) são frequentemente úteis para organizar e resumir esses dados. Elas nos ajudam a compreender as características de um conjunto de dados. Segundo Triola [24], "uma distribuição de frequência (ou tabela de frequências) mostra como o conjunto de dados é dividido entre várias categorias (ou classes), listando as categorias juntamente com o número (frequências) de valores de dados em cada uma delas."

Para entendermos as distribuições de frequência, é necessário saber que a frequência de uma classe particular é o número dos valores originais que cabem dentro daquela classe.

4.2.1 Distribuição de frequência relativa

Uma variação da distribuição de frequência básica é a distribuição de frequência relativa ou distribuição de frequência percentual, na qual cada frequência de classe é substituída pela frequência relativa (ou proporção) ou uma porcentagem.

4.2.2 Distribuição de frequência acumulada

Outra variação da distribuição de frequência é a distribuição de frequência acumulada, na qual a frequência para cada classe é a soma das frequências daquela classe e de todas as classes anteriores.

4.2.3 Distribuição de frequência acumulada relativa

Essa é a distribuição que mostra a proporção acumulada percentual de observações em uma distribuição. Essa distribuição é útil para entender a distribuição acumulada de valores em um conjunto de dados e como esses valores se comparam em relação ao total.

4.2.4 Comparando frequências

Para compreender melhor as distribuições de frequência, não iremos utilizar o banco de dados extraído do DATASUS/TabNet. Como os valores relativos as notificações de Dengue no Brasil são todos distintos em todo o banco de dados, teríamos uma distribuição de frequência igual a um para todas as notificações. Desse modo, vamos usar o seguinte exemplo.

Suponhamos os seguintes dados brutos sendo as idades dos alunos de uma turma de ginástica: 14, 12, 13, 11, 12, 13, 16, 14, 14, 15, 17, 14, 11, 13, 14, 15, 13, 12, 14, 13, 14, 13, 15, 16 (adaptado de [28]).

Tendo em mãos esses dados, podem facilmente construir a Tabela 4.1 de frequência simples f_i e de frequência relativa simples f_{ir}. Obtemos a frequência relativa fazendo a razão entre a frequência simples f_i e o número total de observações do conjunto de dados. Na Tabela 4.2, podemos observar a frequência Acumulada f_a e a Acumulada Relativa f_{ar}.

Tabela 4.1: Tabela de frequência Simples

Idades dos alunos	Frequência Simples (f_i)	Frequência Relativa Simples (f_{ir})
11	2	$\frac{2}{24} = 0{,}08333$
12	3	$\frac{3}{24} = 0{,}12500$
13	6	$\frac{6}{24} = 0{,}25000$
14	7	$\frac{7}{24} = 0{,}29166$
15	3	$\frac{3}{24} = 0{,}12500$
16	2	$\frac{2}{24} = 0{,}08333$
17	1	$\frac{1}{24} = 0{,}04166$

Fonte: Elaborado pela autora (2024) - adaptado de [28].

No *Posit Cloud*, os dados foram inseridos manualmente como um vetor de dados, e usando a função **table()**, uma tabela simples será gerada no *Console*. Essa será a frequência simples. Para obtermos a frequência relativa simples, usamos a função **prop.table()**. Ela carrega os dados da tabela simples na forma proporcional (apresentados na forma decimal).

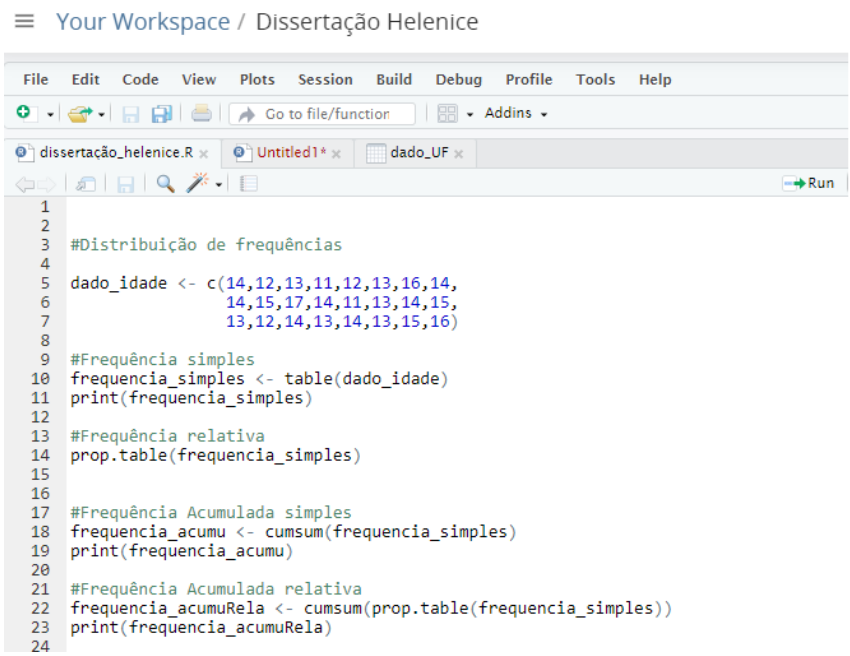
Por sua vez, a tabela de frequência acumulada simples é gerada pela função **cumsum(table())**. E para gerarmos a tabela de frequência acumulada relativa, fazemos a função **cumsum(prop.table())**. Esses códigos estão ilustrados na Figura 4.2 e seus resultados podem ser observados Figura 4.3.

Tabela 4.2: Tabela de Distribuição de Frequências

Idades dos alunos	Frequência Simples (f_i)	Frequência Acumulada (f_a)	Frequência Relativa Simples (f_{ir})	Frequência Relativa Acumulada (f_{ar})
11	2	2	$\frac{2}{24} = 0{,}08333$	$\frac{2}{24} = 0{,}08333$
12	3	5	$\frac{3}{24} = 0{,}12500$	$\frac{5}{24} = 0{,}20833$
13	6	11	$\frac{6}{24} = 0{,}25000$	$\frac{11}{24} = 0{,}45833$
14	7	18	$\frac{7}{24} = 0{,}29166$	$\frac{18}{24} = 0{,}75000$
15	3	21	$\frac{3}{24} = 0{,}12500$	$\frac{21}{24} = 0{,}87500$
16	2	23	$\frac{2}{24} = 0{,}08333$	$\frac{23}{24} = 0{,}95833$
17	1	24	$\frac{1}{24} = 0{,}04166$	$\frac{24}{24} = 1{,}00000$

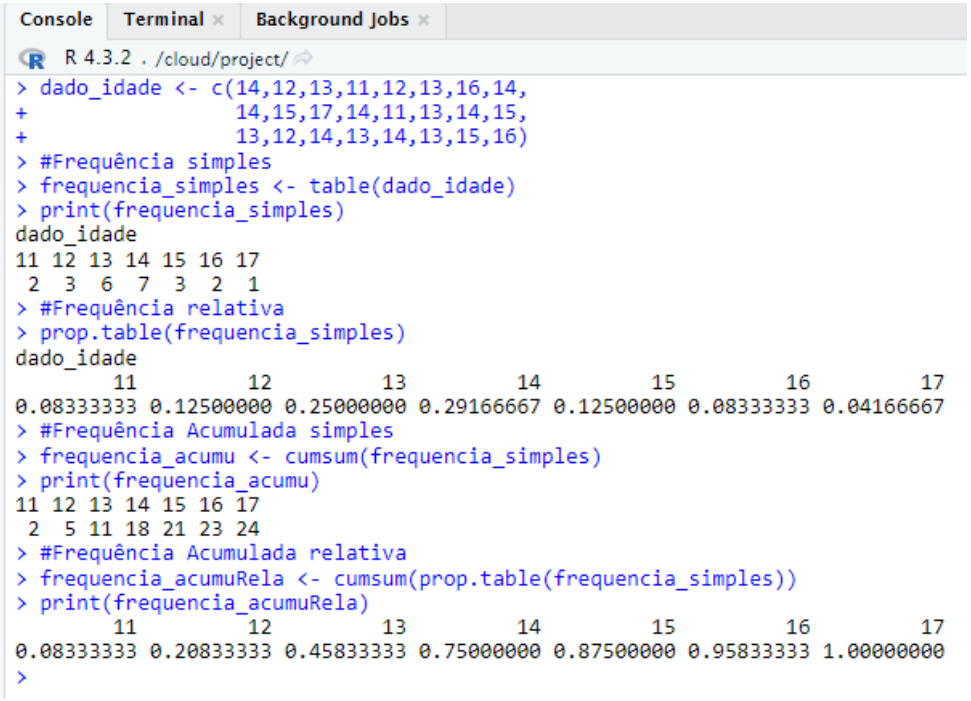
Fonte: Elaborado pela autora (2024) - adaptado de [28].

Figura 4.2: Distribuição de Frequência Simples e Frequência Relativa Simples no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

Figura 4.3: Exemplo de uma tabela de frequência Simples



Fonte: Elaborado pela autora (2024).

4.3 Gráficos

A representação gráfica de uma distribuição de uma variável é mais vantajosa e consistente para mostrar a variação desses dados. Para variáveis quantitativas os gráficos de barras e setores ("pizza") são os mais utilizados. Além desses, o Histograma é uma excelente ferramenta para representar dados quantitativos contínuos.

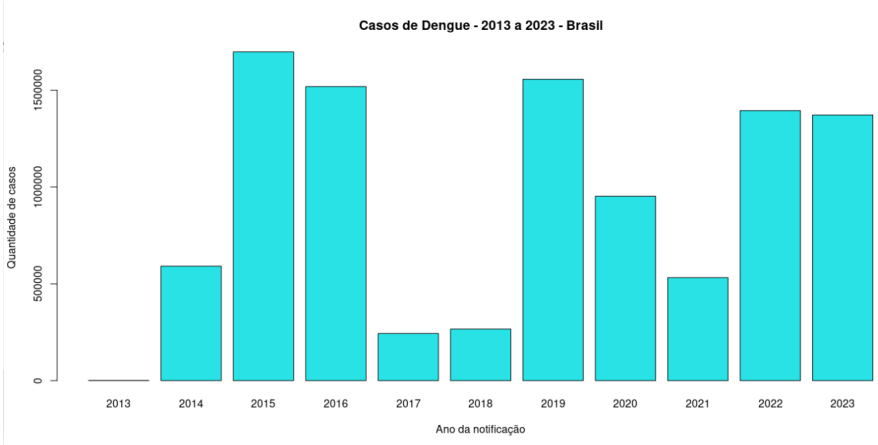
4.3.1 Gráfico de barras

Um gráfico de barra usa barras de igual largura para mostrar as frequências das categorias de dados categóricos (ou qualitativos). A escala vertical representa as frequências ou frequências relativas. A escala horizontal identifica as diferentes categorias dos dados qualitativos. As barras podem, ou não, ser separadas por um pequeno espaço. Um gráfico de barras múltiplas tem dois ou mais conjuntos de barras, e é usado para a comparação de dois ou mais conjuntos de dados.

Como exemplo, vamos utilizar o banco de dados escolhido. No *Posit Cloud*, plotamos os gráficos de Barras dos anos de 2013 a 2023 do cenário nacional, Figura 4.4, e de um estado de cada região do país, de modo a ilustrar o cenário das notificações em cada uma dessas regiões, tendo como referência o estado escolhido.

Os estados escolhidos foram: Minas Gerais, no Sudeste; Paraná, no Sul; Pará, no Norte; Mato Grosso, no Centro Oeste e Bahia, no Nordeste. As Figuras 4.5, 4.6, 4.7, 4.8 e 4.9 representam, respectivamente os gráficos de barras gerados.

Figura 4.4: Gráfico de barras no *Posit Cloud* - Brasil

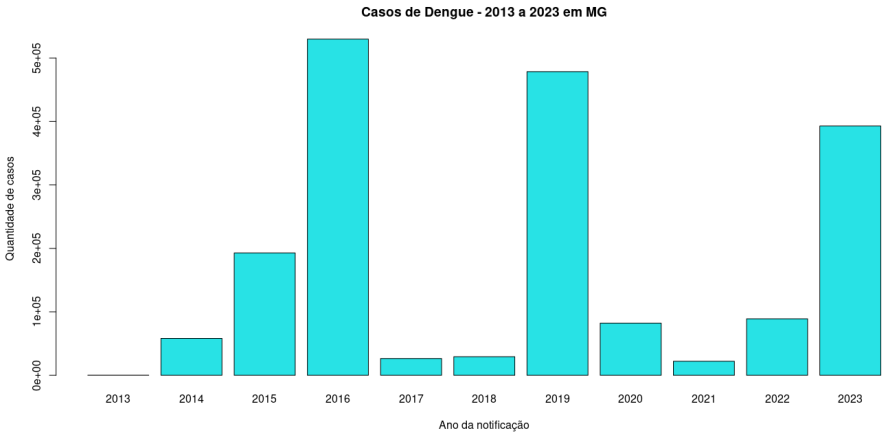


Fonte: Elaborado pela autora (2024).

O gráfico com as notificações de todo país aponta que houve uma maior incidência de casos em 2015 e uma redução significativa das notificações nos anos de 2017 e 2018.

Já em Minas Gerais observamos uma configuração parecida no que se refere ao momento de maior incidência de casos que ocorreu em 2016. Já o período de menor quantidade de casos ocorreu em 2021.

Figura 4.5: Gráfico de barras no *Posit Cloud* - Minas Gerais

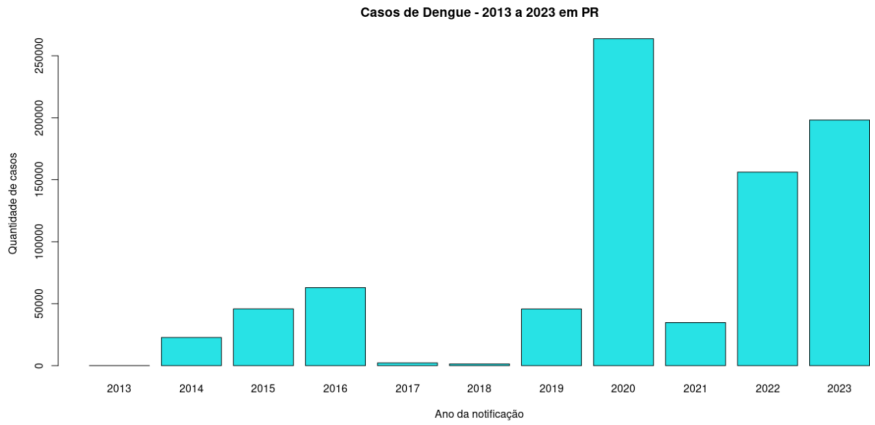


Fonte: Elaborado pela autora (2024).

No gráfico com os registros de Minas Gerais, a representação está na forma $(1e+05) = 1x10^5 = 100000$, que representa potência de 10.

O estado do Paraná teve maior notificação de casos em 2020, enquanto que em 2017 e 2018 aconteceram a menor quantidade de casos. Sendo que nesses dois anos houve estabilidade nos casos.

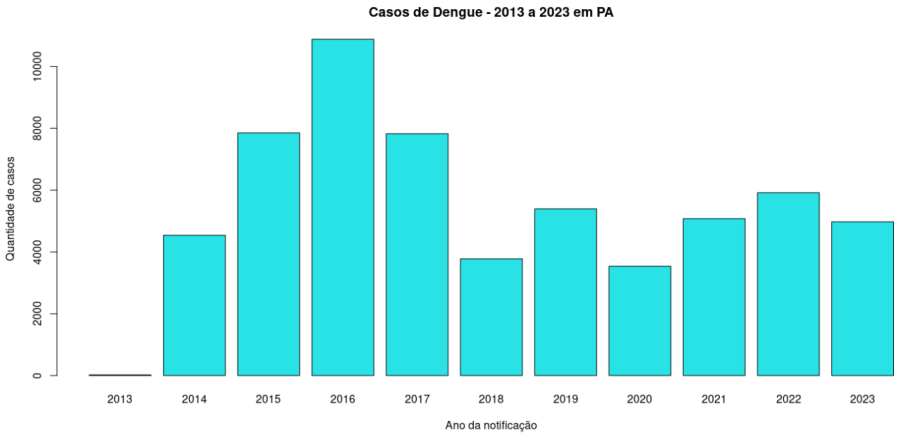
Figura 4.6: Gráfico de barras no *Posit Cloud* - Paraná



Fonte: Elaborado pela autora (2024).

Na região norte, mais especificamente no Pará, o pico nos registros ocorreu me 2016 e em 2020 o menor número de casos foi registrado no estado.

Figura 4.7: Gráfico de barras no *Posit Cloud* - Pará



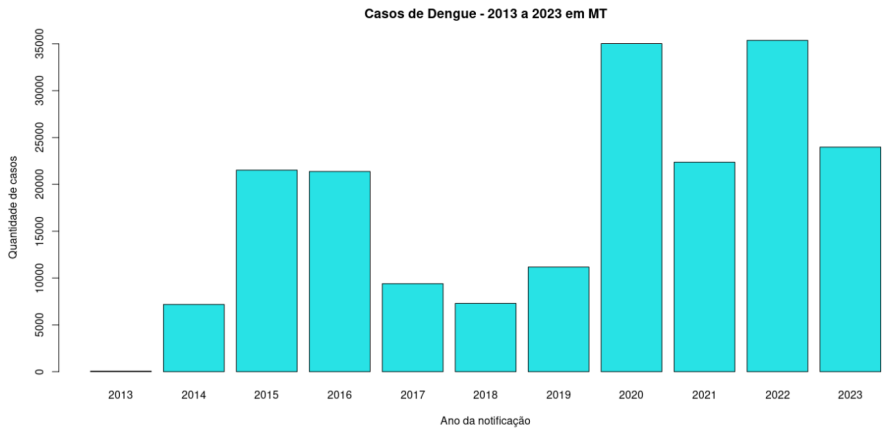
Fonte: Elaborado pela autora (2024).

No Mato Grosso, observamos dois anos em que tivemos uma maior incidência de casos, sendo que o número de notificações foi quase o mesmo. Esses anos foram 2020 e 2022. Em 2018 o registro de casos foi o menor.

A Bahia registrou seu pico de notificações em 2020. E os anos com menor número de casos ocorreu em 2017 e 2018.

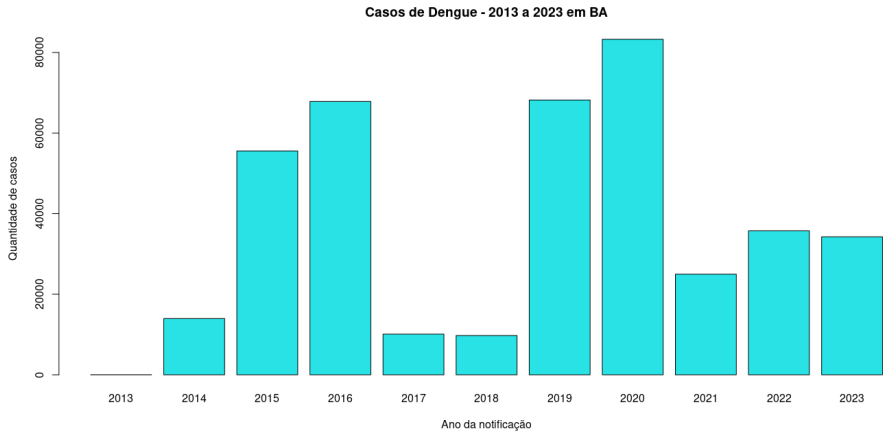
Quando observamos as representações gráficas, podemos fazer análises sobre o

Figura 4.8: Gráfico de barras no *Posit Cloud* - Mato Grosso



Fonte: Elaborado pela autora (2024).

Figura 4.9: Gráfico de barras no *Posit Cloud* - Bahia



Fonte: Elaborado pela autora (2024).

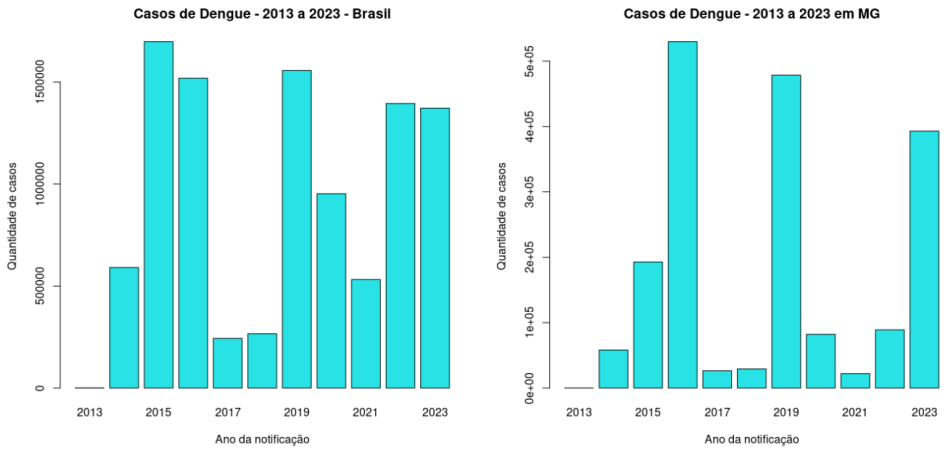
evolução dos casos de Dengue no Brasil em relação aos estados e entre os estados, conforme a necessidade da análise. Se colocamos essas representações lado a lado, conseguimos fazer essas análises mais facilmente. Nas Figuras 4.10 e 4.11 exemplifico a apresentação de dois gráficos. No primeiro, Figura 4.10, estão lado a lado, os gráficos do Brasil e Minas Gerais.

A Figura 4.11, apresenta os gráficos do estado do Mato Grosso e do Pará lado a lado.

Nesta dissertação apresento as funcionalidades e possibilidades para análises. Obviamente, as construções devem seguir alguma justificativa que torne conveniente agrupar gráficos.

No *Posit Cloud*, o Gráfico de Barras é construído usando a função **barplot()**. Podemos editar esse gráfico usando as seguintes funções:

Figura 4.10: Gráfico de barras no *Posit Cloud* - Comparativo Minas Gerais e Brasil



Fonte: Elaborado pela autora (2024).

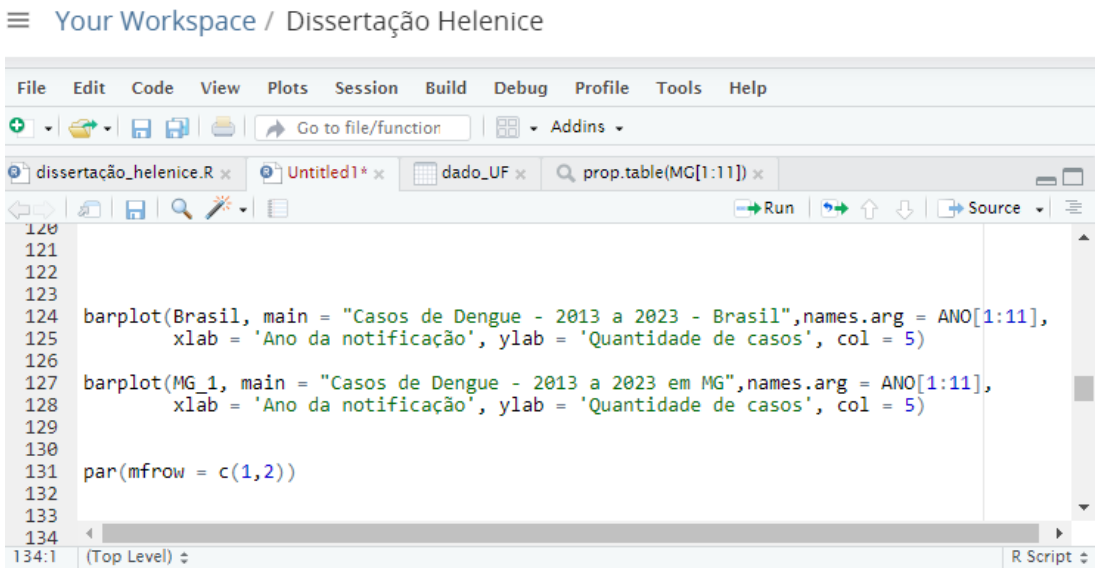
Figura 4.11: Gráfico de barras no *Posit Cloud* - Comparativo Mato Grosso e Pará

Fonte: Elaborado pela autora (2024).

- Inserir título para o gráfico: usamos a função **main = 'título'**;
- Para inserir nomes para as coordenadas: **xlab = 'nome'** e **ylab = 'nome'**;
- Para especificar os argumentos das coordenadas, usamos a função **names.arg = [intervalo]**;
- Podemos mudar a cor das barras usando a função **col = 'nomedacoremIngles'**.

A Figura 4.12 apresenta o código utilizado para gerar os gráficos de barras. Para gerar dois gráficos, um ao lado do outro, a função **par(mfrow(c=(x,y)))**. Essa função é importante para fazer comparações entre gráficos.

Figura 4.12: Código no *Posit Cloud* - Comparativo MG e Brasil



Fonte: Elaborado pela autora (2024).

4.3.2 Histograma

As distribuições de frequência contribuem para o resumo de dados e a investigação da distribuição dos dados. Uma ferramenta ainda melhor é um histograma, que consiste em um gráfico que é de mais fácil interpretação do que uma tabela de números. Segundo (TRIOLA, 2017) [24], "um histograma é, basicamente, um gráfico de uma distribuição de frequência".

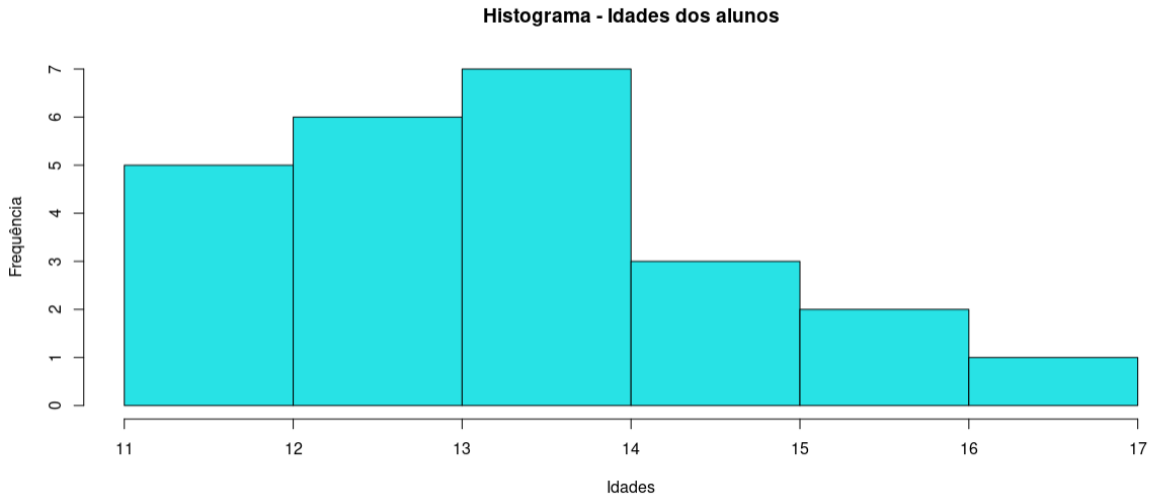
Um histograma é um gráfico que consiste em barras de mesma largura e desenhadas adjacentes umas às outras, exceto quando alguma informação está zerada. A escala horizontal representa classes de valores de dados quantitativos e a escala vertical representa frequências. As alturas das barras correspondem aos valores das frequências.

Sua apresentação é semelhante ao do Gráfico de Barras, porém, existem diferenças significativas entre eles. Enquanto o histograma é usado para visualizar a distribuição de dados contínuos em intervalos, o Gráfico de Barras é usado para comparar quantidades entre diferentes categorias ou rótulos.

Para a apresentação do histograma, vamos inicialmente, usar como exemplo os dados apresentados no seção 4.2.4 sobre as idades dos alunos de uma turma de ginástica. Por se tratar de um conjunto de dados pequeno, fica mais simples a análise do histograma. Na Tabela 4.1, observamos a distribuição dos dados e no histograma abaixo, Figura 4.13, conseguimos ver o comportamento dos dados. Podemos obter informações sobre as medidas

de tendencia central apenas observando o histograma, por exemplo, que a **mediana** e a **moda** das idades da turma está entre 13 e 14 anos.

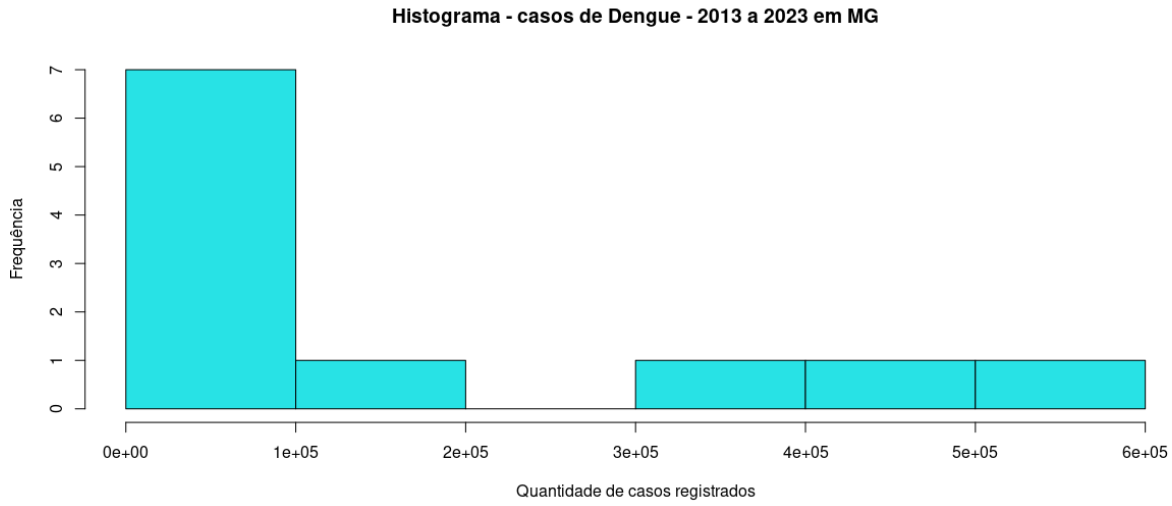
Figura 4.13: Histograma no *Posit Cloud* - Idades



Fonte: Elaborado pela autora (2024).

As próximas analises são sobre os histogramas que representam as notificações de Dengue em Minas Gerais (Figura 4.14), na Bahia 4.15 e no Brasil (Figura 4.16) nos anos 2013 a 2023. Nesse tipo de representação gráfica temos uma noção da frequência das quantidades de casos registrados. Em Minas Gerais, o intervalo mais frequente é de 0 a 10000 casos registrados.

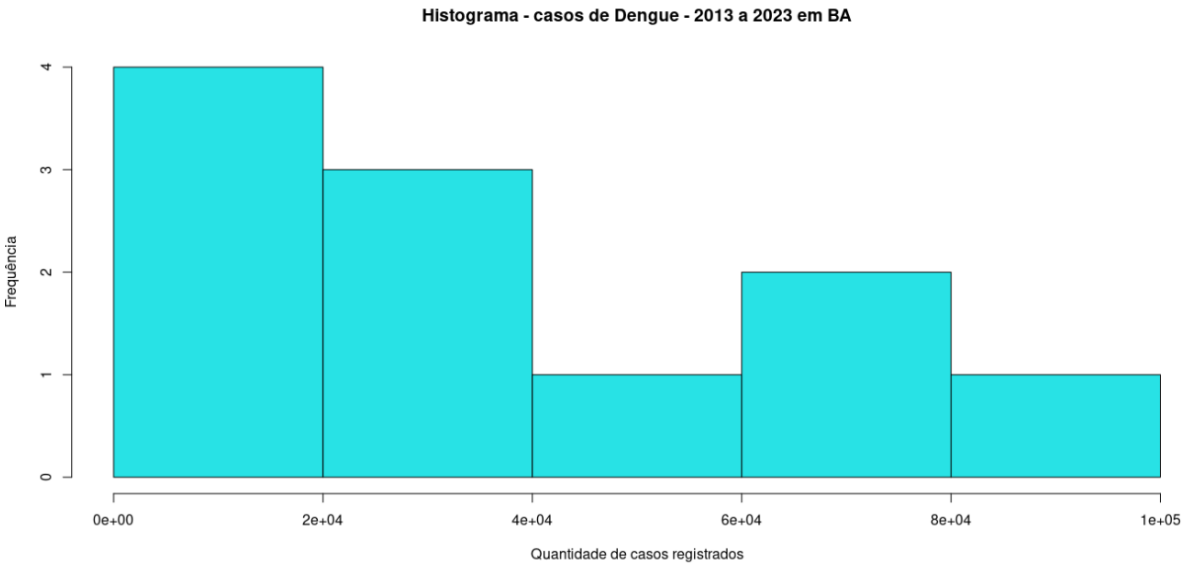
Figura 4.14: Histograma no *Posit Cloud* - MG



Fonte: Elaborado pela autora (2024).

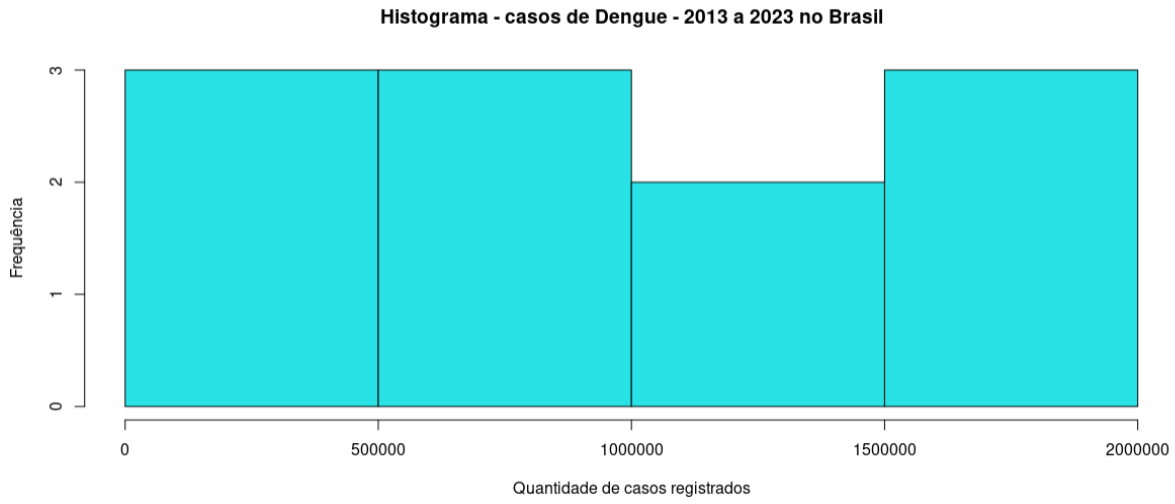
No *Posit Cloud*, a função que imprime o Histograma é **hist()**. As edições são feitas

Figura 4.15: Histograma no *Posit Cloud* - BA



Fonte: Elaborado pela autora (2024).

Figura 4.16: Histograma no *Posit Cloud* - Brasil



Fonte: Elaborado pela autora (2024).

como na função **barplot()**. Essas já foram descritas na seção Gráfico de Barras. A Figura 4.17 apresenta essa elaboração.

Como mentir com Estatística, por Darrell Huff

Em seu livro Como mentir com Estatística [29], Darrell Huff apresenta uma lista das principais falsidades veiculadas pela mídia, políticos e publicitários, muitas vezes respaldadas por manipulações estatísticas. Ele ilustra quão simples é distorcer dados numéricos ou representações gráficas. Os exemplos fornecidos são acompanhados de

Figura 4.17: Código - Histograma no *Posit Cloud*

Fonte: Elaborado pela autora (2024).

explicações detalhadas sobre os métodos estatísticos envolvidos, apresentados de forma didática e em situações relativamente simples. Em cada caso, é possível observar como o método de análise, a parcialidade, o contexto histórico/social e as condições de coleta de dados influenciam a divulgação das estatísticas.

Nesse contexto, vamos observar o comparativo entre os histogramas que representam as notificações em Minas Gerais e no Brasil. A Figura 4.18 mostra os dois histogramas lado a lado e nos induz instantaneamente a comparar os dois cenários. Porém, há um detalhe muito significativo que pode nos levar a conclusões equivocadas: as escalas dos dois histogramas estão diferentes. Isso faz com que a comparação não seja feita corretamente.

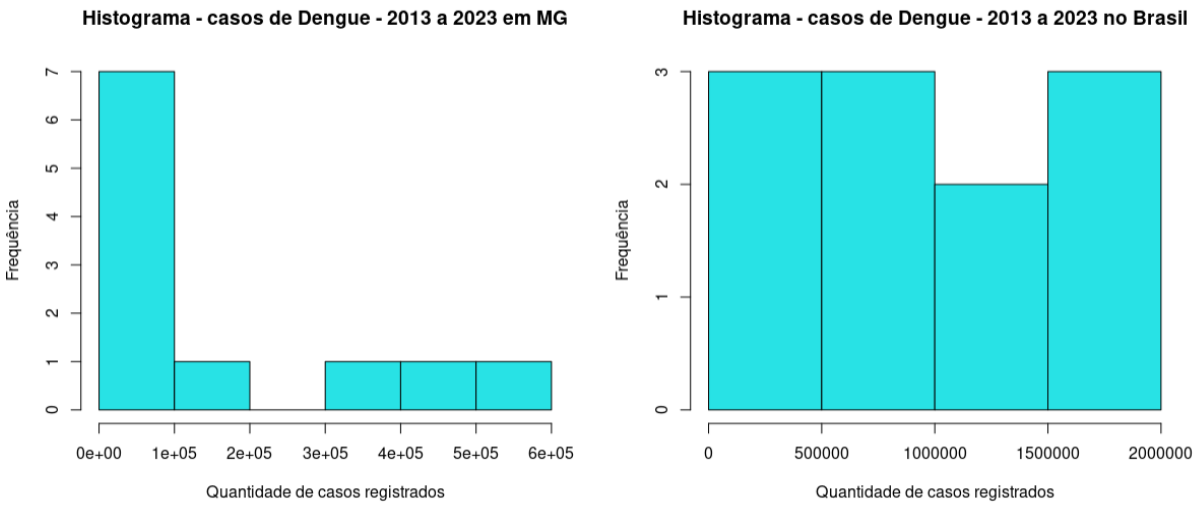
É importante que, não apenas saibamos executar a construção dos gráficos e tabelas, mas que sejamos capazes de fazer análises que considerem os contextos aos quais os dados estão inseridos.

4.3.3 Gráfico de setores

Um gráfico de setores (ou gráfico de pizza) é um gráfico que retrata dados categóricos como setores de um círculo, no qual cada setor é proporcional à contagem de frequência para a categoria.

Assim como nos exemplos dos gráficos de barras, aqui temos a representação dos

Figura 4.18: Histograma no *Posit Cloud* - Comparativo MG e Brasil

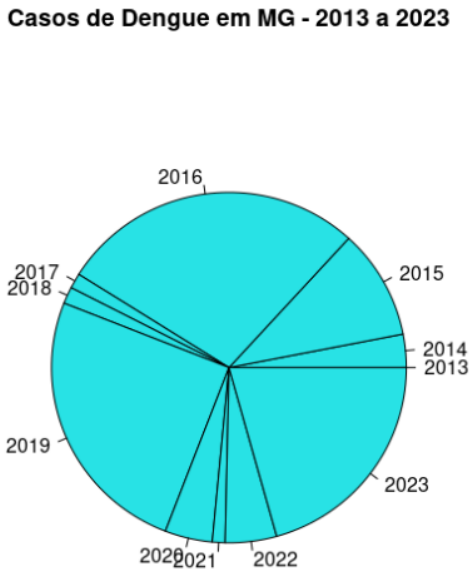


Fonte: Elaborado pela autora (2024).

estados de Minas Gerais, Paraná, Bahia, Mato Grosso e Pará. As Figuras 4.19, 4.23, 4.20, 4.21 e 4.22 indicam essas representações.

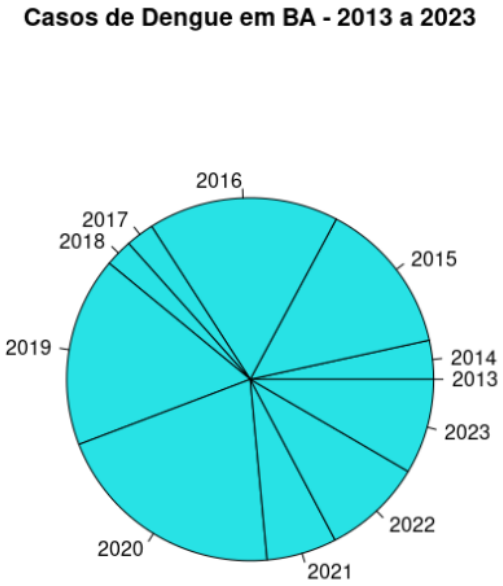
Os gráficos de setores possuem a limitação de nem sempre fornecerem uma escala apropriada. Para o nosso banco de dados, outras representações apresentam análises mais adequadas.

Figura 4.19: Gráfico de setores no *Posit Cloud* - Minas Gerais



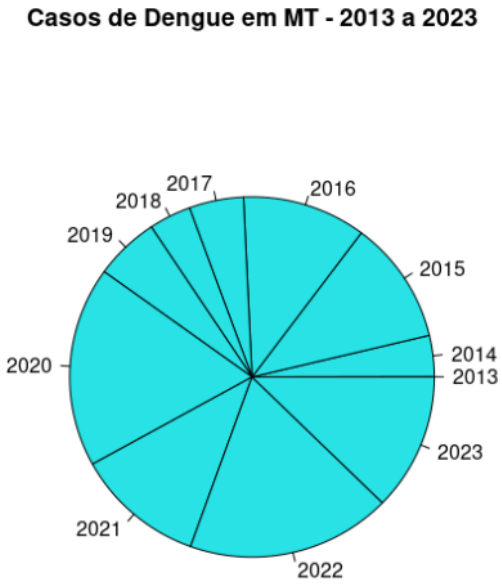
Fonte: Elaborado pela autora (2024).

Figura 4.20: Gráfico de setores no *Posit Cloud* - Bahia



Fonte: Elaborado pela autora (2024).

Figura 4.21: Gráfico de setores no *Posit Cloud* - Mato Grosso

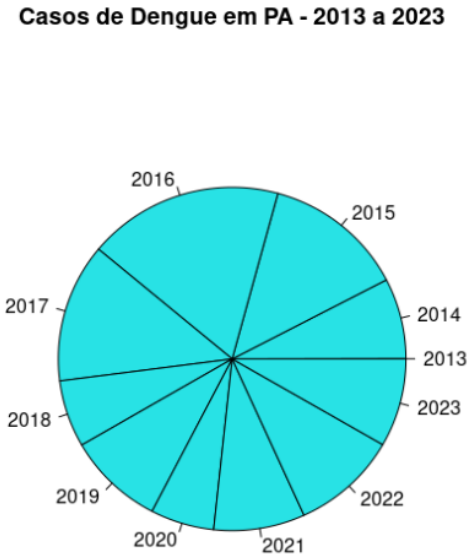


Fonte: Elaborado pela autora (2024).

Para a análise do banco de dados sobre os casos de Dengue, foram feitas duas analises de intervalos de período distintos. Novamente analisamos o estado de Minas Gerais em relação ao cenário nacional.

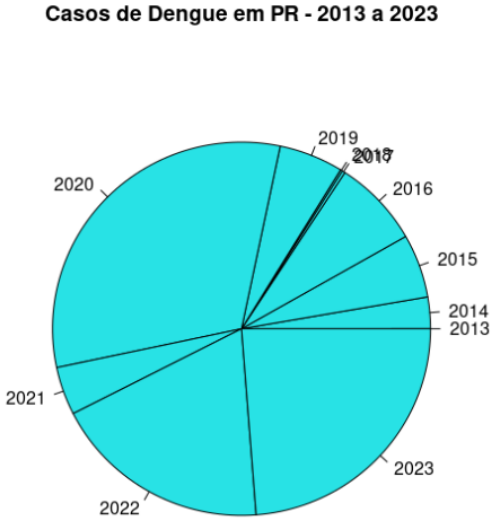
As observações feitas na Figura 4.24 são semelhantes às feitas no gráfico de barras.

Figura 4.22: Gráfico de setores no *Posit Cloud* - Pará



Fonte: Elaborado pela autora (2024).

Figura 4.23: Gráfico de setores no *Posit Cloud* - Paraná

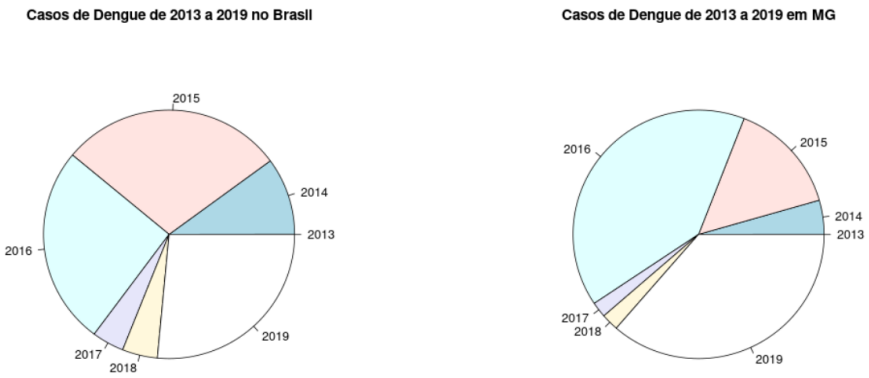


Fonte: Elaborado pela autora (2024).

E, novamente vemos o ano de 2015 discrepante em relação à comparação.

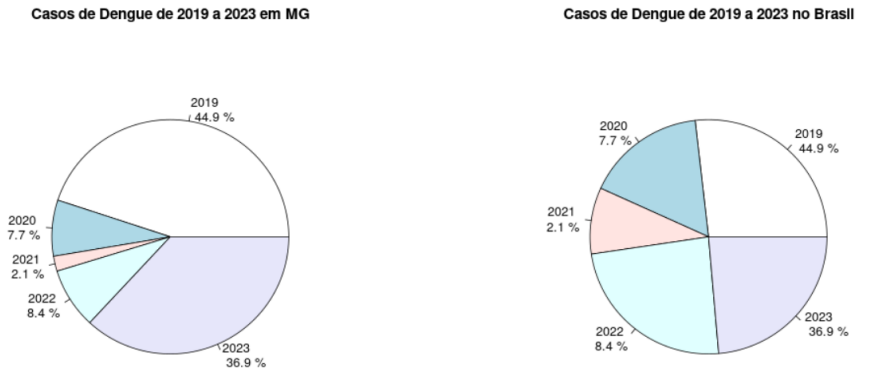
Já na Figura 4.25, em que os dados são impressos em termos percentuais, observamos uma discrepância mais significativa entre as notificações em Minas Gerais e as Nacionais. Essa observação pode ser como objeto de estudo/análise para se entender quais ações motivaram tal diferença.

Figura 4.24: Gráfico de setores no *Posit Cloud* - Comparativo MG e Brasil



Fonte: Elaborado pela autora (2024).

Figura 4.25: Gráfico de setores com percentual no *Posit Cloud* - Comparativo MG e Brasil

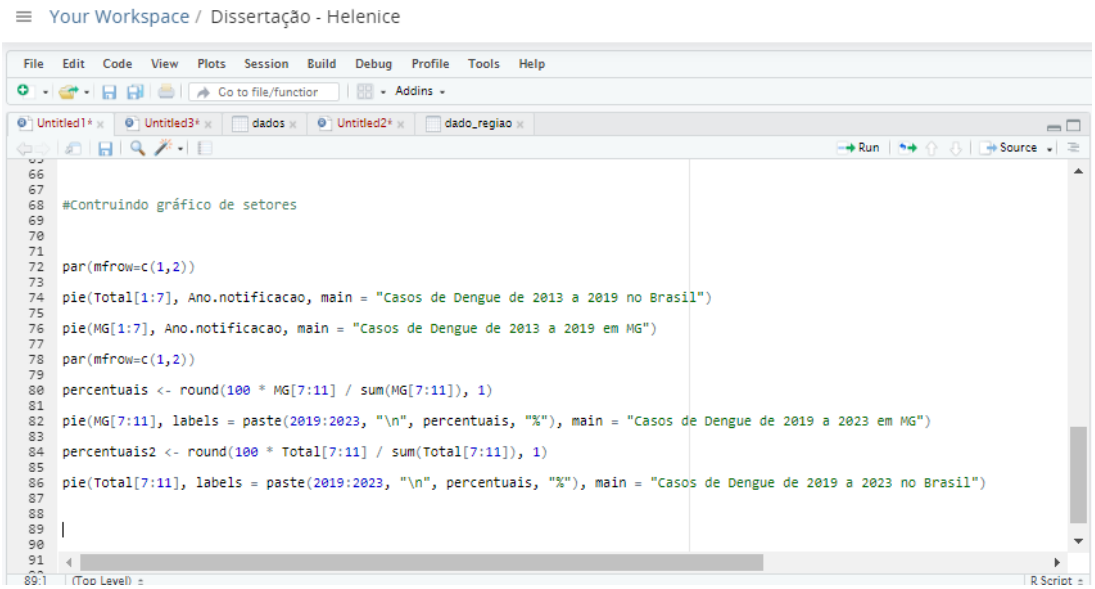


Fonte: Elaborado pela autora (2024).

No *Posit Cloud*, a função que imprime o Gráfico de Setores é a **pie()**. Nessa representação, também é possível inserir o título do gráfico, usando a função **main = ()**, e os nomes das variáveis pode ser inserido criando um vetor de dados. No caso, o vetor foi **Ano.notificacao**, que constava dos anos de 2013 a 2023. Um filtro foi feito para que os anos selecionados fossem os desejados. Para que os dados sejam apresentados em termos percentuais, fazemos a razão entre os dados de cada notificação pela soma das notificações, em seguida usamos a função **round()** para arredondar os resultados da forma desejada. No exemplo apresentado, usamos o arredondamento em uma casa decimal.

A Figura 4.26 apresenta o código para construção dos gráficos de setores no *Posit Cloud*.

Figura 4.26: Código da construção do gráfico de setores no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

4.4 *Boxplot*

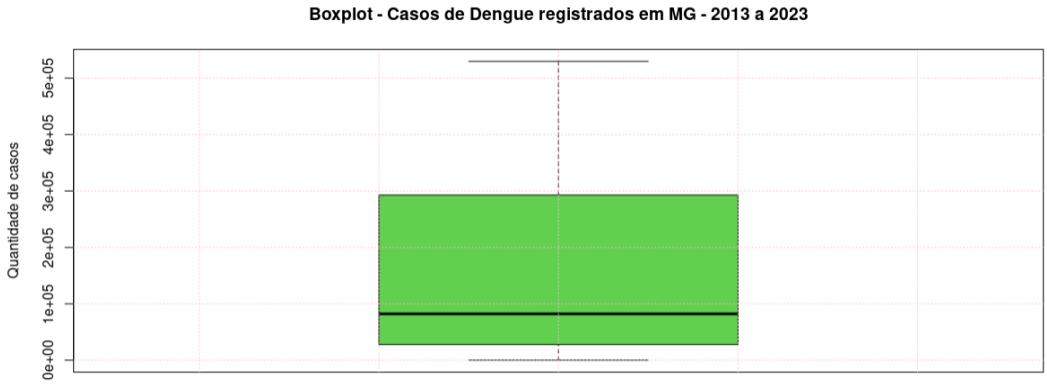
O gráfico *Boxplot* é o tipo de esquema que utiliza cinco medidas resumo. São eles, o valor mínimo, valor máximo, mediana, primeiro e terceiro quartil da variável. Este conjunto de medidas oferece a ideia da posição, dispersão, assimetria e dados discrepantes. A posição central é dada pela mediana e a dispersão pelo desvio interquartílico - amplitude da caixa $(Q_3 - Q_1)$. As posições relativas de Q_1, Q_2 e Q_3 dão uma noção da assimetria da distribuição.

As Figuras 4.27 e 4.28 são os *Boxplots* das notificações de Minas Gerais e do Brasil de 2013 a 2023 respectivamente. Observamos que a ”caixa”da representação mineira, é mais larga que a da representação nacional. Isso significa que há uma maior dispersão dos dados, com uma maior variabilidade entre os valores observados. Vemos ainda, que não há *Outliers* nesse gráfico, ou seja, não há pontos de dados que estão significativamente distante dos outros pontos da distribuição.

Em contrapartida, no *Boxplot* nacional, temos uma ”caixa”estreita e um ponto *Outlier*. Ou seja, há uma menor variabilidade dos dados e há um ponto distante das demais observações.

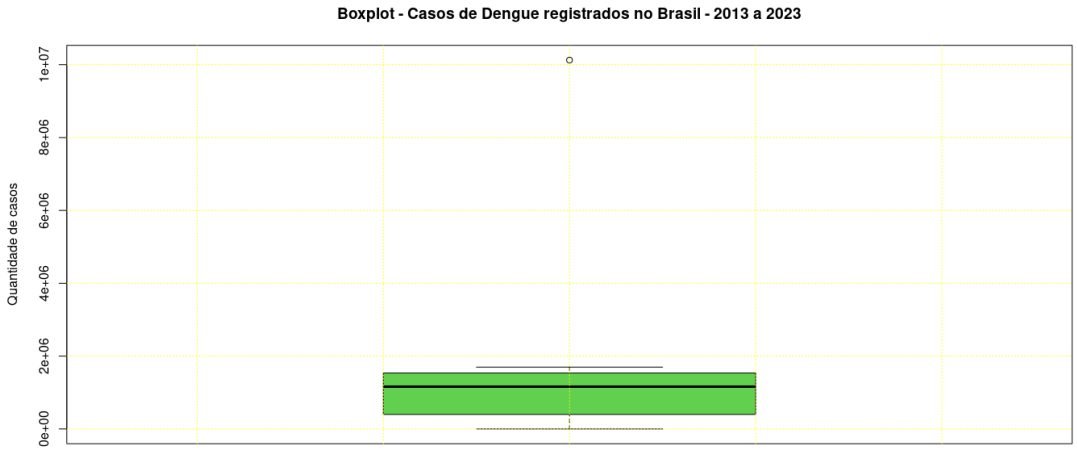
Para se construir um *Boxplot* no *Posit Cloud*, usamos a função **boxplot()**. As edições são feitas do mesmo modo que nos gráficos de barras, histograma e setores.

Figura 4.27: *Boxplot* no *Posit Cloud* - MG



Fonte: Elaborado pela autora (2024).

Figura 4.28: *Boxplot* no *Posit Cloud* - Brasil



Fonte: Elaborado pela autora (2024).

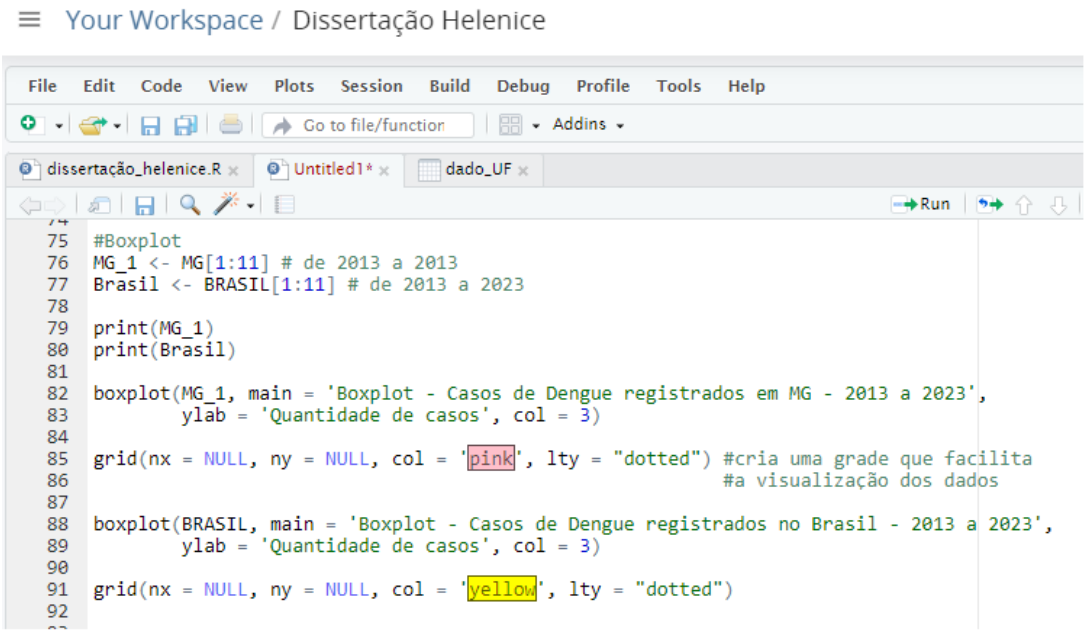
Destacamos um argumento para melhorar a visualização dos dados - **grid()** - ela insere uma grade ao fundo do gráfico. O *Boxplot* pode ser representado na orientação horizontal ou vertical. Caso o interesse seja utilizar a orientação horizontal, usa-se o argumento **horizontal = TRUE** em **boxplot()**. A construção do código está ilustrada na Figura 4.29.

4.4.1 *Outliers*

Um *outlier* ou ponto discrepante é um valor que se localiza distante de quase todos os outros pontos da distribuição.

Valores atípicos ou *outliers* são valores de dados que satisfazem critérios específicos com base nos quartis e na amplitude interquartil. Para encontrar tais valores, devemos seguir os seguintes passos:

Figura 4.29: *Boxplot* no *Posit Cloud*



Fonte: Elaborado pela autora (2024).

- Ache os quartis Q_1, Q_2 *e* Q_3.
- Ache a amplitude interquartil ($AIQ = Q_3 - Q_1$).
- Calcule 1,5 $\times$ AIQ.

Em um diagrama em caixa modificado, um valor observado é um *outliers* se ele está:

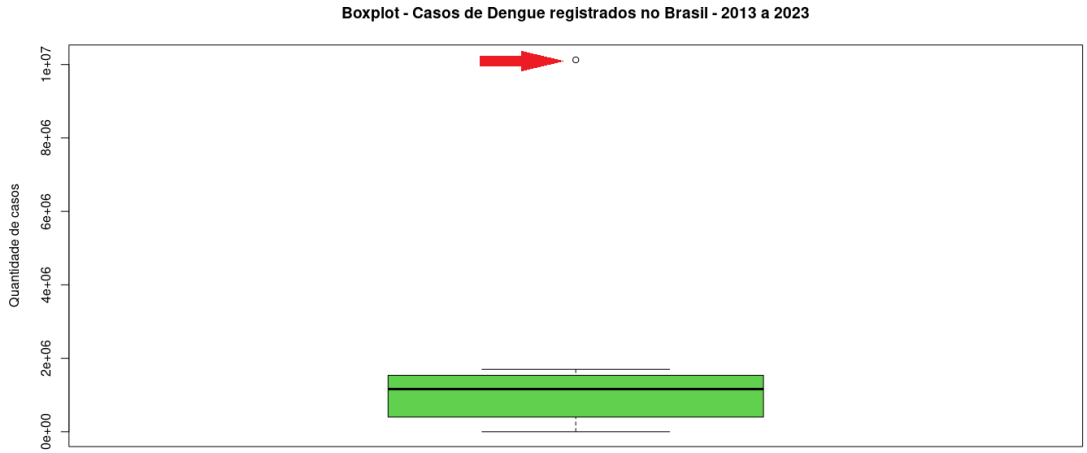
- acima de Q_3 por uma quantidade maior do que 1,5 $\times$ AIQ
- abaixo de Q_1 por uma quantidade maior do que 1,5 $\times$ AIQ

Outliers dos dados - Brasil, Paraná e Pará

Na figura 4.28, em que é apresentado o Boxplot referente aos casos de Dengue no Brasil. Na elaboração desse gráfico, foi utilizado o TOTAL de casos, ou seja, a soma de todos os casos notificados nos anos de 2013 a 2023. Como essa soma - 10126209 notificações - é muito alta, em relação aos outros dados, observamos um *Outliers*, conforme indicado pela seta vermelha na figura 4.30.

Considerando os cinco estados destacados - Minas Gerais, Bahia, Pará, Paraná e Mato Grosso. As únicas observações que possuem *Outliers*, são as notificações do Pará e do Paraná. A Tabela 4.3 apresenta essas notificações dos anos de 2013 a 2023. Nela é

Figura 4.30: *Boxplot - Outliers*



Fonte: Elaborado pela autora (2024).

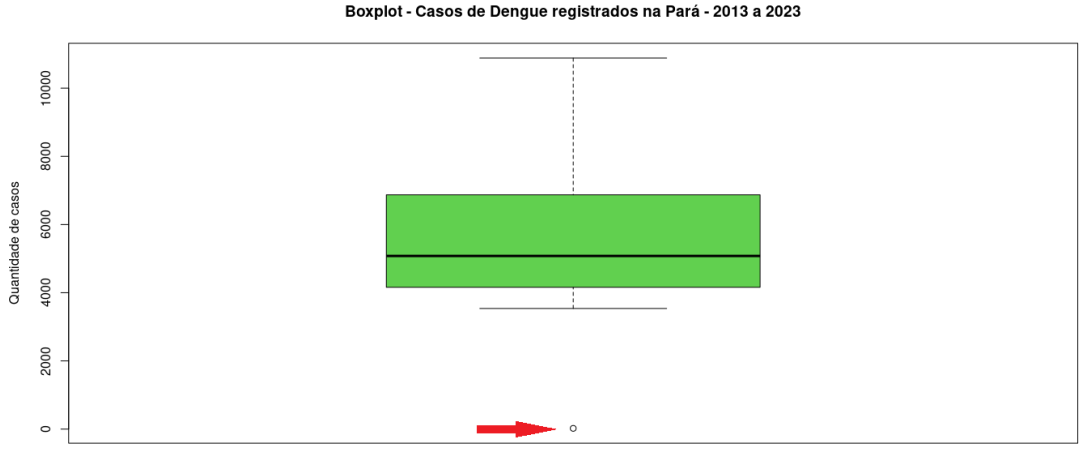
possível observar os valores que correspondem aos *outliers* dos dados obtidos para o estado do Pará e do Paraná. São eles, respectivamente 21 notificações e 263 769 notificações. Tais valores estão representados nos Boxplots das figuras 4.31 e 4.32.

Tabela 4.3: Tabela - dados Pará e Paraná

UF / ANO	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
PARÁ	21	4540	7853	10883	7827	3778	5395	3538	5078	5916	4975
PARANÁ	66	22775	45859	62970	2230	1426	45763	263769	34798	156196	198206

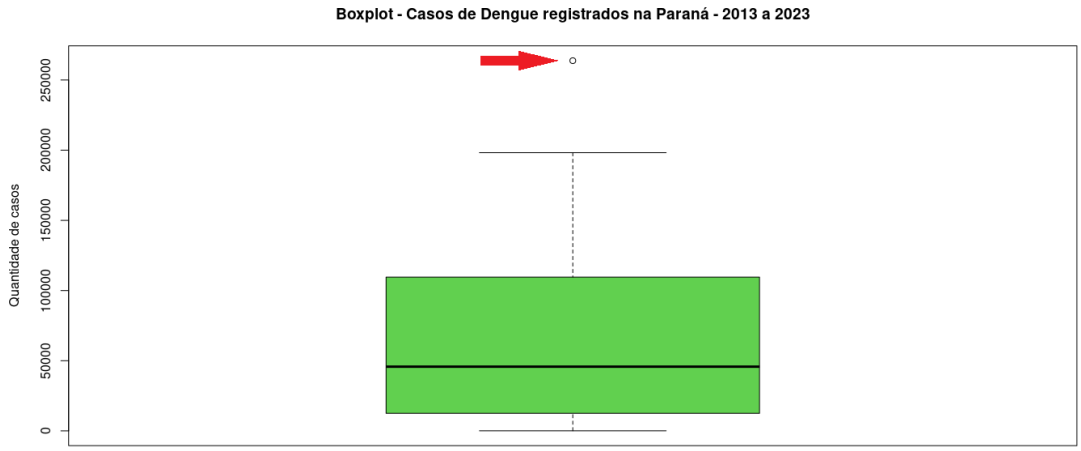
Fonte: Elaborado pela autora (2024) a partir de dados do TabNet.

Figura 4.31: *Boxplot - Outliers* Pará



Fonte: Elaborado pela autora (2024).

Figura 4.32: *Boxplot - Outliers* Paraná



Fonte: Elaborado pela autora (2024).

5 Considerações Finais

A estatística é essencial para a compreensão da realidade humana. De um modo geral, são as pesquisas estatísticas que orientam as ações da sociedade. Portanto, é importante que os cidadãos desenvolvam uma compreensão sólida dos conceitos estatísticos, ainda mais na era da informação que vivemos, uma vez que informações estatísticas podem ser manipuladas, distorcidas ou mal compreendidas. Assim, faz-se necessário promover um ensino de Estatística que preparem os estudantes para lidar e compreender tais informações, bem como os tornar capazes de avaliá-as criticamente.

Essa dissertação propôs um produto educacional constituído por atividades para o ensino da Estatística Descritiva no âmbito do Ensino Médio através da linguagem R, por meio da plataforma online *Posit Cloud*, e utilizando dados reais relacionados aos casos de Dengue no Brasil, acessíveis por meio da plataforma do DATASUS. A linguagem R destaca-se como uma das mais utilizadas para análise de dados, além de ser gratuita e estar disponível para diferentes sistemas operacionais, além de versões *online*. Assim, proporcionar aos estudantes do Ensino Médio o acesso a esta linguagem e a este tópico, intrinsecamente vinculado à era do conhecimento e da *Big Data* que vivemos atualmente, torna-se uma iniciativa crucial para aprimorar suas experiências de aprendizado em sala de aula. Portanto, o produto educacional proposto nessa dissertação assume uma significativa relevância, oferecendo ao professor um material que o oriente em como trabalhar com o R no ensino de estatística em sala de aula utilizando dados reais.

As atividades que compõe o produto educacional tratam dos comandos do R para se trabalhar com Estatística Descritiva e apresentação de dados, bem como usar um banco de dados real do DATASUS. Espera-se que esse trabalho contribua nas práticas dos docentes que o adotarem, uma vez que é notável seu potencial para auxiliar no desenvolvimento de habilidades e competências estabelecidas na BNCC no que diz respeito à estatística.

Referências

1 MEMóRIA, J. M. P. Breve história da estatística. Brasília, DF: Embrapa Informação Tecnológica: Embrapa-Secretaria de Gestão, 2004.

2 NETO, J. W. C. A. e. S. M. d. F. J. F. B. *Dia do Estatístico - 29 de maio*. Recife, CE: [s.n.], 2020. Disponível em: <https://dema.ufc.br/pt/dia-do-estatistico-29-de-outubro/>.

3 BENEVIDES, F. S. *Introdução à Estatística - Portal da Matemática OBMEP*. IMPA, Rio de Janeiro, RJ: [s.n.], 2023. Disponível em: <https://cdnportaldaobmep.impa.br-/portaldaobmep/uploads/material_teorico/a1ydsm2c13404.pdf>.

4 SAúDE, M. da. *Ministério da Saúde declara fim da Emergência em Saúde Pública de Importância Nacional pela Covid-19*. Brasília, DF: Ministério da Saúde: [s.n.], 2022. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2022-/abril/ministerio-da-saude-declara-fim-da-emergencia-em-saude-publica-de-importancia-nacional-pela-covid-19>.

5 CAZORLA, I. M.; CASTRO, F. C. de. O papel da estatística na leitura do mundo: o letramento estatístico. *Publicatio UEPG: Ciências Sociais Aplicadas*, v. 16, n. 1, 2008.

6 MORAES, P. C. A. *Da manipulação estatística do mundo real: utilizando inflação e seus índices*. Dissertação (Programa de Pós-Graduação em Matemática em Rede Nacional - PROFMAT) — Universidade Estadual de Maringá, Maringá, 2020.

7 BRASIL. *Base Nacional Comum Curricular*. Brasília, Distrito Federal: Ministério da Educação, 2017.

8 SOUZA, A. L. *Softwares no ensino de Matemática*. Dissertação (Programa de Pós-Graduação em Matemática em Rede Nacional - PROFMAT) — Universidade Estadual de Santa Cruz, Ilhéus, 2015.

9 ARAUJO, C. D. de. *Estatística no Ensino Médio: Uma proposta de atividades com o uso de tecnologias*. Dissertação (Mestrado Profissional em Matememática em Rede Nacional) — Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2020.

10 GADANIDIS, G.; BORBA, M. de C.; SILVA, R. S. R. da. *Fases das tecnologias digitais em Educação Matemática: sala de aula e internet em movimento*. 3. ed. Belo Horizonte: Autêntica, 2020.

11 BORBA, M. C. Potential scenarios for internet use in the mathematics classroom. *ZDM*, Springer, v. 41, p. 453–465, 2009.

12 SAúDE, M. da. *Dengue*. Janeiro de 2024. Disponível em: <https://www.gov.br/saude-/pt-br/assuntos/saude-de-a-a-z/d/dengue>.

13 Posit software, PBC. *Posit.cloud: Ambiente de Desenvolvimento Integrado para R*. Boston, MA, 2020. Disponível em: <https://posit.cloud/>.

14 SAúDE, M. da. *DataSUS - TabNet*. Brasília, DF: Ministério da Saúde: [s.n.], 2022. Disponível em: <http://datasus.saude.gov.br/>.

15 GIORDANO, C.; ALVES, J.; QUEIROZ, C. de. Educação estatística e a base nacional comum curricular: o incentivo aos projetos. *REVEMAT: Revista Eletrônica de matemática*, Universidade do Extremo Sul Catarinense, v. 14, p. 1–20, 2019.

16 WIKIPÉDIA, a. e. l. *R (LINGUAGEM DE PROGRAMAÇÃO)*. Flórida: Wikimedia Foundation: [s.n.], 2023. Disponível em: <https://pt.wikipedia.org/w/index-.php?title=R_(linguagem_de_programa%C3%A7%C3%A3o)&oldid=66665572>.

17 BOAS, B. R. L. L. R. S. G. M. E. V. *Introdução ao Uso do Software R*. UFMG, Belo Horizonte, MG: [s.n.], 2021. Disponível em: <https://www.est.ufmg.br/~monitoria-/Material/ApostilaR/IntroducaoR1.html>.

18 COMPUTING, T. R. P. for S. *Software R*. 2024. Disponível em: <https://www.r-project.org/>.

19 SAúDE, D. Ministério da. *TUTORIAL - TabNet*. Brasília, DF: Ministério da Saúde: [s.n.], 2020. Disponível em: <https://datasus.saude.gov.br/wp-content/uploads/2020/02-/Tutorial-TABNET-2020.pdf>.

20 CRESPO, A. A. *ESTATÍSTICA FÁCIL*. São Paulo: Editora Saraiva, 2002.

21 LIMA, M. N. M. A. C. P. de. *Noções de Probabilidade e Estatística*. 7. ed. São Paulo, SP: EDUSP, 2010.

22 MOREIRA, T. de Jesus Rocha Vilanova Moreira; Marlei Rosa dos S. A. L. *Estatística Básica para cursos de graduação*. Vol. 1. Terezina, PI: EduESPI, 2021.

23 PINHO, A. G. *Estatística Descritiva: Medidas de Posição e Dispersão*. janeiro de 2024. Disponível em: <https://pt.slideshare.net/AndersonGP/estatstica-descritiva-8427834>.

24 TRIOLA, M. F. *Introdução à Estatística*. 12. ed. Rio de Janeiro: LTC Editora, 2017.

25 MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. São Paulo: Editora Saraiva, 2017.

26 DEGENSZAJN, G. I. S. H. D. *Fundamentos de matemática elementar: matemática comercial, matemática financeira e estatística descritiva*. 2ª. ed. Rio de Janeiro: Atual Editora, 2013.

27 CERQUEIRA, N. *Estatística Descritiva*. 2020. Disponível em: <https://medium.com-/pyladiesbh/estat%C3%ADstica-descritiva-1-ed523dffb99f>. Acesso em: 18 de novembro de 2023.

28 HERONDINO, P. *Descrição e Apresentação dos Dados*. Disponível em: <https://slideplayer.com.br/slide/7308810/>. Acesso em: 20 de novembro de 2023.

29 HUFF, D. *Como mentir com Estatística*. Digital. Rio de Janeiro, RJ: Editora Intrínsica Ltda, 2016.