Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Matemática
Mestrado Profissional em Matemática

Fernando Gualberto Silva Soares

ANÁLISE ESTATÍSTICA MULTIVARIADA DE DADOS EDUCACIONAIS

Uma abordagem para evasão e abandono escolar

Teresina - 2023

PROFMAT

Fernando Gualberto Silva Soares

Dissertação de Mestrado:

ANÁLISE ESTATÍSTICA MULTIVARIADA DE DADOS EDUCACIONAIS: Uma abordagem para evasão e abandono escolar

Dissertação submetida à Coordenação do Programa de Mestrado Profissional em Matemática - PROFMAT, da Universidade Federal do Piauí, como requisito parcial para obtenção do grau de Mestre em Matemática na modalidade profissional.

Orientadora:
Prof. Dra. Jackelya Araujo da Silva

Teresina - 2023

Universidade Federal do Piauí - UFPI, Centro de Ciência da Natureza - CCN, Programa de Pós-Graduação em Matemática, Mestrado Profissional em Matemática. Cep 64049-550 - Teresina, PI.

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

S676a Soares, Fernando Gualberto Silva.
Análise estatística multivariada de dados educacionais: uma abordagem para evasão e abandono escolar / Fernando Gualberto Silva Soares. – 2023.
70 f.

Dissertação (Mestrado Profissional) - Universidade Federal do Piauí, Programa de Pós-Graduação em Matemática, Teresina, 2023.
"Orientadora: Profa. Dra. Jackelya Araújo da Silva."

1. Análise estatística multivariada. 2. Dados Educacionais - BNCC. 3. Educação. I. Silva, Jackelya Araújo da. II. Titulo.

CDD 519.535

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Fernando Gualberto Silva Soares

ANÁLISE ESTATÍSTICA MULTIVARIADA DE DADOS EDUCACIONAIS: Uma abordagem para evasão e abandono escolar

Dissertação submetida à banca examinadora abaixo discriminada em defesa pública e aprovada em 30/08/2023.

BANCA EXAMINADORA

Jackelya Araujo da Silva.

Prof. Dra. Jackelya Araujo da Silva (Orientadora)

Universidade Federal do Piauí

Cleide Mayra Menezes Lima

Prof. Dra. Cleide Mayra Menezes Lima

Universidade Federal do Piauí

(Coordenação de Bach. em Estatística)

Documento assinado digitalmente
gov.br
VALMARIA ROCHA DA SILVA FERRAZ
Data: 12/09/2023 20:28:10-0300
Verifique em https://validar.iti.gov.br

Prof. Dra. Valmária Rocha Da Silva Ferraz

Universidade Federal do Piauí

Teresina - 2023

Dedico esta dissertação à minha família: Maria de Jesus (minha mãe), Kariny Amorim (minha esposa), Maria Clara (minha filha), Mathilde Soares (minha irmã).

Agradecimentos

Agradeço primeiramente a Deus, por me conceder o dom da vida, me mostrando o tempo todo o caminho certo a seguir. Agradeço à minha mãe, Maria de Jesus a quem tenho um amor incondicional ela é responsavel por eu ser o homem que sou hoje e que durante minha jornada no PROFMAT esteve sempre ao meu lado.

À minha esposa Kariny Amorim pelo apoio incondicional durante esta etapa da minha vida.

À minha filha, Maria Clara pelo sorriso de todos dias e por ser o presente mais grandioso que Deus poderia me dar.

À minha irmã, Mathilde Soares por todo suporte e conselhos.

À minha orientadora Jackelya Araujo da Silva pela paciência infinita e principalmente pela disponibilidade e didática agradável na condução do trabalho.

Aos membros da banca examinadora pelas contribuições.

À todos os funcionários da EMEF JOSÉ SARNEY em especial minha querida amiga Francy Mara pela amizade e compreensão.

Aos meus amigos e familiares que me apoiaram durante todo o PROFMAT.

"*Deus nos concede, a cada dia, uma página de vida nova no livro do tempo. Aquilo que colocarmos nela, corre por nossa conta.*".

Chico Xavier.

Resumo

A proposta da Base Nacional Comum Curricular(BNCC) é estabelecer um conjunto de aprendizagens fundamentais que devem ser desenvolvidas por todos os estudantes, independentemente da região do país ou da rede de ensino em que estejam matriculados. Isso visa assegurar uma educação de qualidade, equitativa e que atenda às necessidades da sociedade contemporânea. É fundamental que as unidades de ensino do país capitaniadas pelos seus gestores tenham acesso a dados que possam ajudar a monitorar a aplicação da BNCC. Como são geradas muitas informações educacionais, uma análise univariada contemplando somente aspectos quantitativos, pode não ser o suficiente para descrever o panorama educacional e verificar a completude da execução das propostas apresentadas na BNCC. Dessa forma, o presente estudo tem por objetivo apresentar, de maneira didática, uma análise multivariada de grandes conjuntos de dados educacionais. No estudo a metodologia de agrupamento $K-$means possibilitou descrever que a taxa de evasão apresentou menor variação para o ano de 2018, que o estado de Pernambuco apresentou maior taxa de mé dia de horas aula para rede pública de ensino, evidenciou que os ensinos da rede pública e privada possuem aspectos diferenciados, que o ensino público no Estado do Ceará para ano de 2022, apresentou similaridade ao ensino privado do Estado São Paulo. A metodologia auxiliou na identificação das variáveis que contribuíram para que o grupo dos estados do Rio Grande do Norte, Pará e Bahia apresentassem alta dissimilaridade em relação aos demais estados da unidade federativa do Brasil.

Palavras-chave: Educação, BNCC, Dados Educacionais, Matemática, Análise Multivariada.

Abstract

The proposal of the National Common Curricular Base (BNCC) is to establish a set of fundamental learning that must be developed by all students, regardless of the region of the country or the education network in which they are enrolled. This aims to ensure a quality, equitable education that meets the needs of contemporary society. It is fundamental that the teaching units in the country headed by their managers have access to data that can help monitor the application of the BNCC. As a lot of educational information is generated, a univariate analysis contemplating only quantitative aspects, may not be enough to describe the educational panorama and verify the completeness of the execution of the proposals presented in the BNCC. Thus, the present study aims to present, in a didactic way, a multivariate analysis of large sets of educational data. In the study, the $K-$means grouping methodology made it possible to describe that the dropout rate showed a smaller variation for the year of 2018, that the state of Pernambuco had a higher rate of average class hours for the public school system, showed that Public and private education have different aspects, with public education in the State of Ceará for the year 2022. similar to private education in the State of São Paulo. The methodology helped to identify the variables that contributed to the group of the states of Rio Grande do Norte, Pará and Bahia presenting high dissimilarity in relation to the other states of the federative unit of Brazil.

Key words: Education, BNCC, Educational Data, Mathematics, Multivariate Analysis..

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

Temáticas relativas às análises de dados educacionais são inúmeras. A Educação no Brasil passou por diversos processos de mudanças com o propósito de adequar-se as normativas de atualização no ensino. Para os estudos relacionados ao ensino de matemática adota-se como elementos fundamentais a formação social e intelectual do aluno, possibilitando o desenvolvimento congnitivo para capacitar um cidadão apto a realizar tarefas que auxiliem em mudanças na sociedade.

Segundo a Base Nacional Comum Curricular(BNCC), no Ensino Fundamental, a matemática, por meio da articulação de seus diversos campos – Aritmética, Álgebra, Geometria, Estatística e Probabilidade , precisa garantir que os alunos relacionem observações empíricas do mundo real a representações (tabelas, figuras e esquemas) e associem essas representações a uma atividade matemática (conceitos e propriedades), fazendo induções e conjecturas. Assim, espera-se que os alunos desenvolvam a capacidade de identificar oportunidades de utilização da matemática para resolver problemas, aplicando conceitos, procedimentos e resultados para obter soluções e interpretá-las segundo os contextos das situações.

As mudanças atuais no curriculo no ensino de matemática requer um docente mais preparado para desenvolver atividades que vão além dos aspectos algébricos apresentados em sala de aula. Diante do amplo acesso à rede de computadores, faz-se necessário que os professores desenvolvam habilidades de ensinar matemática usando computadores e outras ferramentas.

A proposta do desenvolvimento das habilidades dos professores devem contemplar uma diversidade de temáticas nas áreas científicas, que englobam o saber dos conteúdos a serem ensinados e, na área tecnológica, em que compete o uso assertivo das tecnológias disponíveis para ensinar matemática. O aspecto teórico-prático em que se baseia a atual abordagem de ensino no Brasil busca fomentar a criação de dinâmicas diferenciadas para que a aprendizagem seja desenvolvida em um ambiente que favoreça o entendimento dos conteúdos a serem estudados.

Contudo, muitos professores desenvolvem além de atividades educacionais, funções administrativas escolares, como por exemplo a função de gestor escolar.

Os gestores escolares são os responsáveis por coordenar atividades acadêmicas, são

responsáveis pelas práticas investigativas em diferentes perspectivas didático-pedagógicas, é aquele que em conjunto com a direção escolar, tem a responsabilidade de realizar levantamentos numéricos sobre o desenvolvimento acadêmico dos alunos e dentro de suas competências propor mudanças para melhoria no ambiente escolar.

O Programa Nacional Escola de Gestores da Educação Básica Pública faz parte das ações do Plano de Desenvolvimento da Educação (PDE) e surgiu da necessidade de se construir processos de gestão escolar compatíveis com a proposta e a concepção da qualidade social da educação, baseada nos princípios da moderna administração pública e de modelos avançados de gerenciamento de instituições públicas de ensino, buscando assim, qualificar os gestores das escolas da educação básica pública, a partir do oferecimento de cursos de formação a distância. A formação dos gestores é feita por uma rede de universidades públicas, parceiras do MEC. O Programa Nacional Escola de Gestores da Educação Básica Pública tem como objetivos gerais: formar, em nível de especialização (lato sensu), gestores educacionais efetivos das escolas públicas da educação básica, incluídos aqueles de educação de jovens e adultos, de educação especial e de educação profissional; contribuir com a qualificação do gestor escolar na perspectiva da gestão democrática e da efetivação do direito à educação escolar com qualidade social (MEC, 2022).

De maneira geral, a estatística pode contruibuir para a formação dos gestores no sentido de auxiliar na tomada de decisão. A estatística está inserida em diversas áreas do conhecimento e é bastante utilizada para análise de dados educacionais. As análises numéricas educacionais fornecem um panorama da educação no Brasil, e podem contribuir para o acompanhamento dos aspectos educacionais dos alunos, bem como tornar visível, através de apresentações gráficas, o desempenho acadêmico escolar da unidade educacional da qual o gestor faz parte. A estatística, pode ainda, em suas muitas metodologias guiar o gestor escolar na identificação de problemas relativos à evasão e baixo rendimento acadêmico dos alunos. E dessa forma, poderá tomar medidas preventivas para promover a diminuição do quadro de déficit educacional da sua unidade escolar.

Dentro da proposta de oferta de eduação de boa qualidade em que se baseia a BNCC, vale notar que a Base não mudará os aspectos regionais e culturais observados nas diferentes regiões brasileiras. O Brasil é constituído de 26 estados e o Distrito Federal de modo que cada região possui peculiaridades diferentes. A BNCC, enfatiza que a aprendizagem de qualidade é uma meta que o Brasil deve perseguir incansavelmente.

Com a BNCC, apesar das diferenças regionais existentes no Brasil, o objetivo é garantir o conjunto de aprendizagens essenciais e comuns aos estudantes brasileiros, bem como promover mudanças além do curriculo e influenciar a formação inicial e continuada dos educadores. Contudo, espera-se que as diferentes regiões brasileiras possam sanar as déficits acadêmicos de aprendizagem, repetência e abandono.

Nesse contexto, anualmente, as unidades escolares realizam um agrupamento de informações para que o Ministério da Educação possa obter, através das análies estatísticas,

um panorama da situação educacional no Brasil. E dessa forma, exercer um acompanhamento dos cumprimentos das metas apresentadas no BNCC.

Contudo, dada a grande massa de dados educacionais gerada anualmente, a estatística se estabelece como uma poderosa ferramenta no auxílio à interpretação dos diversos cenários educacionais apresentados no Brasil pelas unidades federativas. Todavia, uma análise univariada para a grande quantidade de dados que são geradas pode camuflar informações e interpretações que não são perceptíveis frente a uma análise multivariada de dados.

Nessa perspectiva, este trabalho tem como objetivo apresentar uma análise multivaridada a partir de uma grande massa de dados educacionais, bem como disponibilizar o procedimento metodológico para a reprodução da metodologia de agrupamento utilizada neste trabalho.

Motivação

Como introduzido anteriormente, a BNCC preconiza um ensino de qualidade e tem como um dos objetivos melhorar os índices escolares com a oferta de um currículo unificado para o Brasil. Alguns dos índices escolares que precisam ser melhorados são: evasão e abandono escolar, repetência, promoção entre outros.

Contudo, várias questões motivaram a construção deste trabalho: será que as taxas de abandono ou evasão escolar são mais fortemente características regionais? Poderia alguma Unidade Federativa do Brasil apresentar destaque quanto aos índices educacionais? Considerando uma análise em que engloba de forma conjunta (aspectos multivariados) os índíces educacionais, haveria semelhanças entre as diferentes regiões do Brasil? Que grupos de estados da Unidade Federativa do Brasil apresentam o mesmo patamar educacional? Se pudéssemos agrupar os estados brasileiros que apresentam características semelhantes quanto aos índices educacionais, como seria o agrupamento? Poderiam, por exemplo, os estados das regiões norte e sul estarem no mesmo grupo? As redes pública e privada de ensino apresentam índices educacionais semelhantes em todas as regiões do Brasil? Poderia haver destaque, quanto aos melhores índices educacionais, para algum estado cujo ensino seja da rede pública?

Dessa forma, na perspectiva de responder as essas e outras questões, este trabalho motivou-se pela obtenção de maiores informações sobre o panorama geral da educação no Brasil sob uma visão multivariada na análise de dados educacionais.

Objetivos

Este estudo tem como objetivo geral apresentar a descrição do panorama educacional do Brasil considerando uma análise estatística multivariada dos dados educacionais em que as taxas de evasão e abandono escolar estão inseridas no contexto das Unidades Federativas do Brasil. Além disso, os objetivos específicos deste estudo se concentram em:

(i) Realizar análise descritiva dos conjuntos de dados com demais variáveis;

(ii) Analisar e descrever os conjuntos de dados educacionais sob o aspecto multivaridado;

(iii) Apresentar os aspectos em que a metodologia proposta neste trabalho poderá ser útil em demais casos;

(iv) Proceder com apresentação da rotina no Software R para reprodutibilidade desde estudo;

(v) Apontar as dificuldades e impactos em que a metodologia utilizada neste trabalho poderá auxiliar em outros aspectos os dados educacionais;

Estrutura do Trabalho

A partir do objetivo principal proposto neste estudo, de analisar dados educacionais de forma multivariada no que se refere as taxas de evasão, abandono entre outras, este trabalho está dividido em quatro capítulos. O primeiro capítulo apresenta os aspectos relativos ao referencial teórico básico sobre as variáveis que serão estudadas, bem como o contexto geral de análises educacionais e tecnologias.

O segundo apresenta a metodologia e os materiais utilizados no estudo em questão, temos uma breve apresentação dos conjuntos de dados, bem como das variaveis utilizadas. Aqui, também, é apresentado o software utilizado para a realização das análises estatísticas e, como forma de facilitar a reprodução deste trabalho, fornecemos o passo a passo de como proceder com a instalação e uso do software R.

No terceiro capítulo, de forma didática, apresenta-se as análises multivariadas com aplicação da metodologia $K-$means para os conjuntos de dados referentes ao ano de 2018 e com brevidade a análise para o ano de 2022.

E por fim, no Apêndice, encontram-se as rotinas computacionais realizadas no software R para que este trabalho e a metodologia possam ser reproduzidos.

Capítulo 1

Fundamentação Teórica

Neste capítulo serão apresentados os aspectos teóricos que fundamentam este estudo. Serão abordados tópicos essenciais e inerentes à metodologia, conteúdos relativos ao sistema educacional brasileiro, alguns referenciais teóricos sobre tecnologia e educação, bem como estudos metodológicos da apresentação de análises estatísticas básicas usuais que dificilmente captam os aspectos de análises conjunta dos dados educacionais na perspectiva de identificação de formação de grupos.

1.1 Aspectos dos Ensinos Fundamental e Médio

O sistema educacional brasileiro passou por diversas reformas ao longo dos anos, incluindo mudanças significativas nas estruturas dos Ensinos Fundamental e Médio. Como diretrizes básicas foi instituída a Base Nacional Comum Curricular (BNCC) que é um marco na educação brasileira, em que estabelece normas e diretrizes para o currículo comum brasileiro das etapas dos Ensinos Fundamental e Médio. A BNCC tem como proposta unificar os objetivos de aprendizagem em todo o país, garantindo uma educação de qualidade, equidade e alinhada com as necessidades contemporâneas (EDUCAÇÃO|BNCC, 2022).

Ainda de acordo com Educação|BNCC (2022), o Ensino Fundamental constitui a etapa inicial da educação formal, na qual os estudantes desenvolvem as bases de conhecimento, competências e valores essenciais para sua formação integral. No Brasil, a estrutura do Ensino Fundamental tem evoluído ao longo dos anos em resposta a desafios educacionais e sociais, visando garantir uma educação de qualidade e equidade para todos.

O Ensino Fundamental Brasileiro é organizado em nove anos obrigatórios, destinados a estudantes na faixa etária dos 6 aos 14 anos. A estrutura curricular abrange diferentes áreas de conhecimento, como Língua Portuguesa, Matemática, Ciências Naturais, Ciências Sociais, Arte e Educação Física. A legislação educacional também assegura a obrigatoriedade do ensino de História e Cultura Afro-Brasileira e Indígena, bem como o

estímulo à inclusão de temas transversais, como ética, cidadania e meio ambiente (FERREIRA; ABREU, 2021).

Apesar dos avanços, o Ensino Fundamental brasileiro enfrenta desafios relevantes. A qualidade do ensino pode ser afetada por muitos fatores: adequação da formação acadêmica dos professores, gestão unilateral da unidade escolar, desigualdades regionais, entre outros. Além disso, a evasão escolar e a dificuldade de aprendizagem em algumas áreas do conhecimento continuam sendo obstáculos a serem superados.

Diversas iniciativas de reforma têm sido implementadas para aprimorar a estrutura e a qualidade do Ensino Fundamental. A adoção de programas de formação continuada para professores, a revisão dos currículos para incluir abordagens mais interdisciplinares e contextualizadas, e o investimento em tecnologias educacionais são exemplos de esforços para enfrentar esses desafios.

Segundo Favero et al. (2017) uma educação de qualidade no Ensino Fundamental é essencial para o desenvolvimento socioeconômico do Brasil. Além de contribuir para a formação de cidadãos críticos e participativos, ela prepara os estudantes para os desafios futuros, promove a igualdade de oportunidades e fortalece a competitividade nacional. Portanto, investir na estrutura e na qualidade do Ensino Fundamental é essencial para construir uma sociedade mais justa e preparada para enfrentar os desafios globais.

Dessa forma, percebe-se que a estrutura do Ensino Fundamental no Brasil desempenha um papel fundamental na formação dos indivíduos e no desenvolvimento da sociedade como um todo. Apesar dos desafios enfrentados, as reformas em curso buscam aprimorar a qualidade e a pertinência dessa etapa educacional. O comprometimento contínuo com a melhoria da estrutura, formação de professores e recursos educacionais é essencial para assegurar uma educação de qualidade e equidade, preparando os estudantes para um futuro promissor.

A estrutura do Ensino Médio no Brasil é geralmente dividida em três anos, com uma carga horária mínima de 3.000 horas, distribuídas em disciplinas obrigatórias e optativas. Disciplinas como Matemática, Português, Ciências e História são comuns a todos os estudantes, enquanto as disciplinas optativas podem variar de acordo com a escola e a ênfase escolhida pelo aluno. Além disso, a Lei de Diretrizes e Bases da Educação Nacional (LDB) estabelece a necessidade de promover a interdisciplinaridade e a contextualização dos conteúdos, buscando uma educação mais integrada e relevante para os estudantes (PIOLLI; SALA, 2021).

Para Costa e Silva (2019), sem dúvidas, uma das reformas mais notáveis no Ensino Médio brasileiro é a implementação da Base Nacional Comum Curricular (BNCC), que busca estabelecer diretrizes para os currículos em todas as escolas do país. A BNCC introduziu um novo modelo de organização curricular, baseado em itinerários formativos, que permitem aos alunos escolherem aprofundamentos em áreas de interesse, como

Linguagens, Ciências da Natureza, Matemática e Ciências Humanas.

As reformas, incluindo a implementação da BNCC, têm o potencial de promover uma educação mais flexível e alinhada com as necessidades dos alunos. Ao permitir que os estudantes escolham seus itinerários formativos, a estrutura do Ensino Médio pode se tornar mais atrativa e relevante, aumentando a motivação e a participação dos alunos. Além disso, uma educação mais contextualizada e interdisciplinar pode preparar melhor os alunos para os desafios do século XXI, promovendo habilidades como pensamento crítico, resolução de problemas e colaboração (FERRETTI, 2018).

No que se refere ao ensino de matemática, ensinar e aprender são conceitos distintos, mas interligados na prática da sala de aula. Isso leva à reflexão sobre criar ambientes de aprendizagem e desenvolver competências no professor, como domínio de conceitos matemáticos e estratégias didáticas. Na Educação Matemática, a resolução de problemas é reconhecida como recurso de aprendizagem, apontado por especialistas em Educação e psicologia (SPINILLO et al., 2017).

Fidelis et al. (2022), enfatiza que a busca por soluções e a reflexão sobre problemas levam à geração do conhecimento. Ao enfrentar desafios apresentados pelos problemas, os alunos tomam decisões, preveem resultados e criam novas abordagens de resolução. Acrescentam ainda, que a aprendizagem não ocorre apenas pela resolução de problemas, mas também pela reflexão sobre eles e suas soluções, destacando que os problemas são o cerne para trabalhar a Matemática em todos os níveis escolares.

A Matemática é uma disciplina fundamental no currículo escolar, contribuindo para o desenvolvimento de habilidades cognitivas, lógicas e analíticas (SPINILLO et al., 2017).

Segundo Fidelis et al. (2022), Spinillo et al. (2017) a ênfase em resoluções de problemas é um ponto de partida para o ensino. Essa abordagem permite aos estudantes usar conhecimentos prévios e desenvolver habilidades para gerenciar informações, permitindo assim a compreensão de conceitos, procedimentos e atitudes matemáticas.

Portanto, a Base Nacional Comum Curricular introduziu uma abordagem inovadora e desafiadora para o ensino da Matemática no Brasil. Ao colocar a compreensão, a aplicação e a resolução de problemas no centro do ensino, ela busca formar estudantes mais preparados para enfrentar as demandas do século XXI. Porém, a implementação bem-sucedida requer a colaboração de educadores, formuladores de políticas e demais agentes educacionais, visando à construção de uma educação matemática mais enriquecedora e eficaz.

1.2 Aspectos da evasão e abandono escolar

Um fator preocupante que afeta o sistema escolar, que é de conhecimento público, é a evasão escolar. Esse problema ocorre quando os estudantes abandonam a escola antes de completar seus estudos, resultando em sérias consequências pessoais, sociais e econômicas.

A evasão escolar tem sido um dos problemas que tem atravessado décadas, necessitando que o governo e a sociedade decidam em conjunto quais serão as medidas necessárias para que haja superação desse quadro.

Segundo Branco et al. (2020) a evasão escolar é um fenômeno complexo e multifacetado, resultado da interação de diversos fatores e entre as principais causas, destacam-se os aspectos socioeconômicos, problemas familiares, desmotivação, problemas de saúde, bullying e violência, infraestrutura escolar precária, necessidades de formação inicial e continuada dos professores, possíveis desajustes na prática didático-metodológica, gestão autoritária, falta de identidade do aluno com a escola, entre outros. Os autores retratam que o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira-INEP faz uma distinção entre evasão e abandono escolar. O termo "abandono"é usado quando um aluno se desliga da escola temporariamente, mas retorna no ano seguinte. Já a "evasão"é quando o aluno sai da escola e não retorna.

De acordo Filho e Araújo (2017) a evasão escolar tem impactos negativos significativos, tanto para os indivíduos quanto para a sociedade como um todo, sendo eles: desvantagem educacional, onde os estudantes que abandonam a escola têm oportunidades de emprego e crescimento profissional limitadas; ciclo de pobreza, uma vez que a falta de educação formal dificulta a obtenção de empregos melhores e bem remunerados; risco de delinquência, jovens sem educação e perspectivas futuras estão mais propensos a se envolverem em comportamentos delinquentes e criminosos e; impacto na economia, evasão escolar representa uma perda de capital humano para a sociedade e pode afetar negativamente o crescimento econômico de um país.

A prevenção da evasão escolar é também uma questão complexa que requer esforços coordenados entre escolas, famílias, comunidades e governos. Algumas estratégias são sugeridas por diversos autores incluem: programas de intervenção como apoio psicossocial e emocional para estudantes em situação de risco; inclusão e acolhimento torna o ambiente escolar acolhedor, inclusivo e seguro pode melhorar o engajamento e o desempenho acadêmico; monitoramento do desempenho, acompanhar o desempenho dos alunos de forma contínua pode identificar precocemente aqueles em risco de evasão, permitindo intervenções oportunas; parcerias com a comunidade, estabelecer parcerias com organizações locais e comunitárias pode ampliar as oportunidades de suporte aos estudantes; programas de reinserção, desenvolver ações que busquem reinserir jovens que abandonaram a escola, gera motivação para que eles retomem seus estudos (BRANCO et al., 2020).

Portanto, um dos grandes desafios da sociedade, escolas e educadores é assegurar o direito à educação, o acesso e a permanência dos alunos, além de proporcionar uma educação de qualidade que atenda às necessidades dos estudantes e promova identidade e pertencimento entre eles. Para reduzir significativamente os índices de evasão e reprovação, é fundamental promover inovação no processo de ensino e aprendizagem, bem como ações do poder público. Ao investir no combate à evasão escolar, é possível construir um futuro mais promissor para nossos jovens e para a sociedade.

Um aspecto importante é observar o fato de que é necessário que a escola desenvolva metodologias que incentivem os alunos a permanecerem na escola auxiliando para a minimização frequente das razões de evasão mais apontadas, como por exemplo a falta de recursos financeiros e o compromissos do aluno com a escola. Segundo Demir e Karabeyoglu (2015) compromissos dos alunos com a escola, controle dos pais e avaliação do ambiente escolar juntos explicam 22% das variações de evasão ou abandono.

De acordo com o Ministerio da Educação (MEC) a definição de abandono escolar ocorre quando o aluno deixa de frequentar as aulas durante o ano letivo . De acordo com Júnior, Santos e Maciel (2016) o estudante é matriculado na escola e ao longo do ano se afasta e deixa de frequentar, concluindo o ano sem aproveitamento mínimo satisfatório.E para alem disso o comprometimento dos alunos com a escola é o preditor mais importante de evasão.

A relação entre a evasão e o abandono escolar está bem estabelecida, pelo fato de que as intervenções de evasão são por vezes consideradas como medidas de prevenção do abandono (ALEXANDER; ENTWISLE; HORSEY, 1997).

As metodologias de intervenções na escola como por exemplo, gestão de contingencia, reorganização escolar, parcerias com a comunidade e atividades que aproximam a familia da escola podem auxiliar na diminuição nas taxas de abandono (SUTPHEN; FORD; FLAHERTY, 2010).

Como foi visto, logo acima, a falta de recursos financeiros não se justifica por si só, mas em geral o desligamento também ocorre pela incompatibilidade entre horário de estudo e trabalho. O problema da evasão escolar preocupa a escola e sua direção ao constatar que há alunos com pouca vontade de estudar.

O desempenho escolar é um aspecto decisivo do sistema educacional e está intrinsecamente ligado ao sucesso acadêmico e desenvolvimento dos estudantes. É uma medida que avalia a eficácia do processo de ensino-aprendizagem e o alcance dos objetivos educacionais estabelecidos (BASSETTO, 2019).

Segundo Rosa, Fernandes e Lemos (2020), o desenvolvimento acadêmico é uma preocupação social, em que, como estudantes, equipe docente, pais e comunidade, têm papéis importantes na construção de uma base sólida de conhecimento, possibilitando que os alunos desenvolvam habilidades sociais e pensamento crítico em relação ao mundo ao

seu redor.

Para os autores, o mau desempenho escolar pode ser resultado de diversas causas pessoais, familiares, emocionais, pedagógicas e sociais, que explicam o fracasso do estudante. Alguns desses fatores incluem a motivação para aprender, o ambiente de aprendizagem, as características do professor, os recursos disponíveis no ambiente familiar, a capacidade do aluno em enfrentar desafios e os aspectos socioeconômicos e socioculturais presentes.

De acordo com Fernandes et al. (2018), a avaliação do desempenho escolar pode ser realizada de diversas formas com base em indicadores e métodos de mensuração, entre eles: notas e avaliações, testes padronizados, avaliações formativas e somativas. Além desses critérios, são utilizados ainda, indicadores qualitativos que abrangem aspectos como habilidades de resolução de problemas, criatividade, pensamento crítico e habilidades sociais. Assim, o desempenho escolar é uma medida fundamental para compreender o progresso dos estudantes e a eficácia do sistema educacional como um todo e este reflete a interação de diversos fatores, desde o ambiente familiar até a qualidade do corpo docente e os recursos educacionais disponíveis.

Portanto, observa-se que a análise cuidadosa do desempenho escolar pode auxiliar na identificação de desafios e necessidades específicas dos alunos, bem como no aprimoramento do ensino e das políticas educacionais, visando proporcionar uma educação mais eficiente e inclusiva. Entre esses desafios é possível avaliar o cenário de abandono e evasão escolar de forma quantitativa, ou seja, através de índices e taxas de aprovação e reprovação, bem como qualitativa, levantando os fatores que levam o aluno a deixar o ambiente escolar.

1.3 Educação e tecnologia

As interferencias relativas às mudanças ocorridas nas instituições sociais e nas relações de experiência dentro da aldeia global digital tem provocado alterações importantes nos conteúdos, nas formas e nos códigos, nos processos de socialização das pessoas, principalmente das novas gerações, o que leva à criação de demandas e exigências educacionais nas instituições de ensino (GÓMEZ, 2015).

O Plano Nacional de Educação (PNE) prever a formação, em nível de pós-graduação, 50% (cinquenta por cento) dos professores da educação básica, até o último ano de vigência deste PNE, e garantir a todos (as) os (as) profissionais da educação básica formação continuada em sua área de atuação, considerando as necessidades, demandas e contextualizações dos sistemas de ensino, bem como universalizar o ensino fundamental de 9 (nove) anos para toda a população de 6 (seis) a 14 (quatorze) anos e garantir que pelo menos 95% (noventa e cinco por cento) dos alunos concluam essa etapa na idade recomendada, até o último ano de vigência do PNE, em 2024 (BRASIL, 2014).

Assim, os estudos voltados para as análises de taxa de evasão, abandono escolar entre outros, podem fornecer um panorama da execução das estratégias previstas no PNE, no que diz respeito a meta (3) da estratégia (3.2), que está relacionada a garantir formação básica comum nas Unidades Federativas do Brasil.

No que se refere as competencias digitais na formação inicial de professores em Portugal, tendo como base o cruzamento entre os parâmetros propostos pela União Europeia e pela Organização para a Cooperação e Desenvolvimento Econômico e as pesquisas científicas realizadas no país, o estudo sobre as "Reflexões teóricas sobre o lugar e o papel das tecnologias digitais na formação inicial de professores em Portugal" pode concluir políticas públicas impulsionadas pelas diretrizes estabelecidas pela União Europeia fortalecem a aquisição destas competências tão essenciais na sociedade em que vivemos hoje (SILVA; COSTA, 2022).

Em sentido contrário ao abordado no estudo de Silva e Costa (2022), o trabalho sobre a prática de formação desenvolvida por meio de tecnologias de Fonseca (2020) apresenta um estudo em que busca discutir a prática de formação desenvolvida por meio da utilização das tecnologias de informação e comunicação na formação inicial.

O estudo aponta para as dificuldades em oferecer oportunidades para que os futuros professores possam observar e experimentar exemplos de usos pedagógicos com as tecnologias digitais. É certo que as mudanças na área tecnológica são constantes, imprimindo um ritmo muito acelerado onde muito rapidamente as tecnologias ficam obsoletas, o que requer uma atitude de constante aprendizagem por seus usuários.

Assim, no contexto desta nova realidade que se coloca, faz-se necessário ressignificar as competências digitais para os ambientes em rede, colaborativos e interativos.

Existem muitas tecnologias que auxiliam no processo de ensino e aprendizagem de conceitos matemáticos, por exemplo softwares como *LOGO, Cabri Géomètre, Simcalc, Funcion Probe, Matlab, Geogebra* entre outros.

A utilização dos softwares proporciona aos estudantes uma facilitaçao na assimilação dos conteudos, o professor consegui desenvolver uma aula mais dinamica e propiciar assim uma participação mais efetivas dos alunos nas atividades propostas em sala de aula. E partindo do pressuposto das relaçoes de tecnologias e aprendizagem de conceitos matemáticos através da cultura digital e dentro de uma linha sociocultural da aprendizagem, mostraram como a exploração de ferramentas da web e de dispositivos móveis, por exemplo, contribuem para o desenvolvimento de atividades em que os estudantes constroem conceitos, resolvem problemas e socializam soluções de forma conjunta (CASTRO; FREIREB; CASTRO, 2017).

Capítulo 2

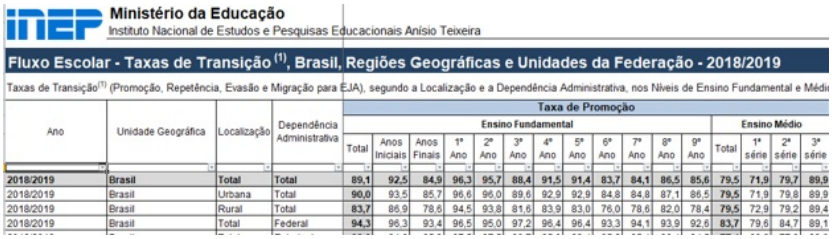
Metodologia e Materiais do Estudo

2.1 Apresentação dos conjuntos de dados

Para este trabalho foram utilizados quatro conjuntos de microdados referentes aos anos escolares 2018 e 2022. Para o primeiro conjunto de dados, ano escolar 2018 considerou-se dois novos conjuntos de microdados, a saber: dados relativos às taxas educacionais e dados cujas informações eram sobre média de horas aulas para todas os estados brasileiros para os ensinos fundamental e médio. De forma similar foi realizado para o ano letivo 2022. Os conjuntos de microdados foram obtidos a partir da plataforma de acesso a informação de dados abertos do governo do Brasil (INEP, 2022).

A Figura 2.1 apresenta parcialmente dois dos conjuntos de microdados utilizados no estudo.

Figura 2.1: Prints dos conjuntos de dados originais - ano 2018

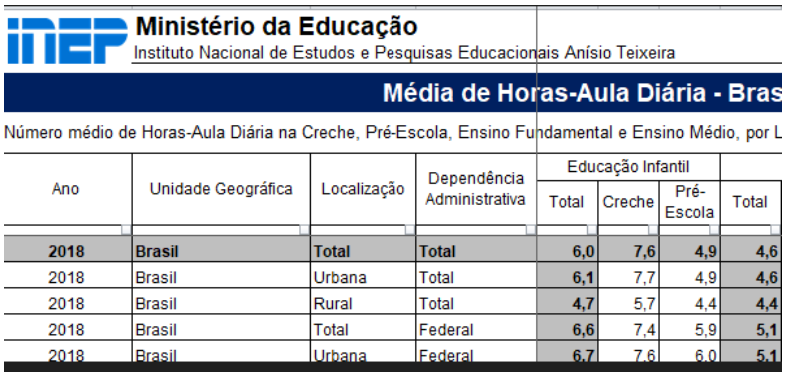


Ministério da Educação
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Fluxo Escolar - Taxas de Transição (1), Brasil, Regiões Geográficas e Unidades da Federação - 2018/2019

Taxas de Transição(1) (Promoção, Repetência, Evasão e Migração para EJA), segundo a Localização e a Dependência Administrativa, nos Níveis de Ensino Fundamental e Médi

Ano	Unidade Geográfica	Localização	Dependência Administrativa	Taxa de Promoção															
				Ensino Fundamental												Ensino Médio			
				Total	Anos Iniciais	Anos Finais	1º Ano	2º Ano	3º Ano	4º Ano	5º Ano	6º Ano	7º Ano	8º Ano	9º Ano	Total	1ª série	2ª série	3ª série
2018/2019	**Brasil**	**Total**	**Total**	**89,1**	**92,5**	**84,9**	**96,3**	**95,7**	**88,4**	**91,5**	**91,4**	**83,7**	**84,1**	**86,5**	**85,6**	**79,5**	**71,9**	**79,7**	**89,9**
2018/2019	Brasil	Urbana	Total	**90,0**	93,5	85,7	96,6	96,0	89,6	92,9	92,9	84,8	84,8	87,1	86,5	**79,5**	71,9	79,8	89,9
2018/2019	Brasil	Rural	Total	**83,7**	86,9	78,6	94,5	93,8	81,6	83,9	83,0	76,0	78,6	82,0	78,4	**79,5**	72,9	79,2	89,4
2018/2019	Brasil	Total	Federal	**94,3**	96,3	93,4	96,5	95,0	97,2	96,4	96,4	93,3	94,1	93,9	92,6	**83,7**	79,6	84,7	89,1



Ministério da Educação
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Média de Horas-Aula Diária - Bras

Número médio de Horas-Aula Diária na Creche, Pré-Escola, Ensino Fundamental e Ensino Médio, por L

Ano	Unidade Geográfica	Localização	Dependência Administrativa	Educação Infantil			
				Total	Creche	Pré-Escola	Total
2018	**Brasil**	**Total**	**Total**	**6,0**	**7,6**	**4,9**	**4,6**
2018	Brasil	Urbana	Total	**6,1**	7,7	4,9	**4,6**
2018	Brasil	Rural	Total	**4,7**	5,7	4,4	**4,4**
2018	Brasil	Total	Federal	**6,6**	7,4	5,9	**5,1**
2018	Brasil	Urbana	Federal	**6,7**	7,6	6,0	**5,1**

Para melhorar a estrutura dos conjuntos de microdados, realizou-se os seguintes tratamentos nos dados: seleção das variáveis a serem consideradas para o estudo, construção e unificação dos conjuntos de dados, segundo os anos escolares, para aplicação da metodologia, modificações (codificação das variáveis) na nomenclatura das variáveis para facilitar a implementação do algoritmo computacional.

As variáveis de estudo e suas respectivas codificações são apresentadas na Tabela 2.1 a seguir abaixo:

Tabela 2.1: Variáveis para o estudo e codificações para os conjuntos de dados

Anos escolares	Nomenclaturas originais das variáveis	Codificação
2018	Taxa de promoção	TXP
	Taxa de repetência	TXR
	Taxa de evasão	TXE
2022	Taxa de reprovação	TRE
	Taxa de aprovação	TAP
	Taxa de abandono	TAB
2018 e 2022	Média de horas aula	MHA
	Ensino Fundamental 1	EF1
	Ensino Fundamental 2	EF2
	Ensino Médio	EM
	Localidade (urbana, rural e total que faz referência a todo o estado)	LOCAL
	Dependência Administrativa (pública e privada)	DA
	Unidade Federativa	UF

Fonte: Elaborado pelo Autor

De acordo com a Tabela 2.1 a codificação, por exemplo Ensino Fundamental 1, taxa de promocão será $EF1\_TXP$. E, por exemplo, para o ano de 2022 a codificação para Ensino Fundamental 2, taxa de abandono será $EF2\_TAB$.

Para o ano escolar 2022 a variável taxa de evasão escolar não foi informada na plataforma de acesso a informação de dados abertos, porém apresenta as taxas de abandono para o ensinos fundamental e médio. Segundo a estrutura da educação Educação|BNCC (2022), o ensino fundamental é composto de dois níveis: Ensino Fundamental 1 que compõe os anos iniciais e o Ensino Fundamental 2 que faz referência aos anos finais da formação educacional do estudante no ensino fundamental.

A seguir a Figura 2.2 apresenta, parcialmente, uma das novas configurações dos conjuntos de dados.

Figura 2.2: Print parcial do conjunto de dados codificados

UFNum	EF1_TXP	EF2_TXP	EM_TXP	EF1_TXR	EF2_TXR	EM_TXR	EF1_TXE	EF2_TXE	EM_TXE
RO	91.3	86.5	78.7	7.5	6.5	6.1	1.1	4.5	9.6
AC	87.8	86.4	78.5	9.2	5.9	9	2.8	5.4	8.7
AM	89.4	84.8	80.5	8.2	7.7	8.6	2	4.5	9.6
RR	90.6	85.8	77	7.2	9.1	10.6	2.2	3.8	8.1
PA	84.9	77.9	75.3	12.5	12.3	10	2.3	6	12.7
AP	87.1	78.2	70.4	10.4	13.3	13.1	2	4	11.2
TO	92.7	84.3	80	6.4	9.9	8.4	0.8	4.4	8.3
MA	89.5	80.1	78.8	7.8	10.9	7.4	2.4	6	12
PI	87.8	80	78.3	9.7	11.1	8.8	2	4.7	9
CE	95.8	89.6	87	3.1	5	3.7	1	3.6	7.8
RN	89.2	75.6	72.2	8.9	15.4	13.3	1.4	4.2	11.4
PB	88.5	77.6	74.2	9.1	13.3	10	2.1	5.6	11.6
PE	90	84.5	85.8	7.6	7.5	4.8	2	4.5	6.3
AL	90	79.3	77	7.4	10.1	8.3	2.3	6.2	10.6
SE	86.8	74	77.8	11.1	16.4	10	1.7	4.6	10.5
BA	86	74.3	72.6	10.8	15.6	12.8	2.6	5.4	11.8
MG	96.7	85.1	76.2	2.5	9.5	8.9	0.8	4.2	11.8
ES	92.7	81.4	77.3	6.1	10.9	9.3	1.1	4.5	8
RJ	89.5	83.1	77.1	8.4	10.4	10.8	1.9	4	8.9
SP	97.4	93.6	86.5	2	3.5	5.4	0.6	2.2	6.5

A primeira coluna da Figura 2.2 representa os estados e as demais colunas são as variáveis em estudo.

2.2 O software estatístico R

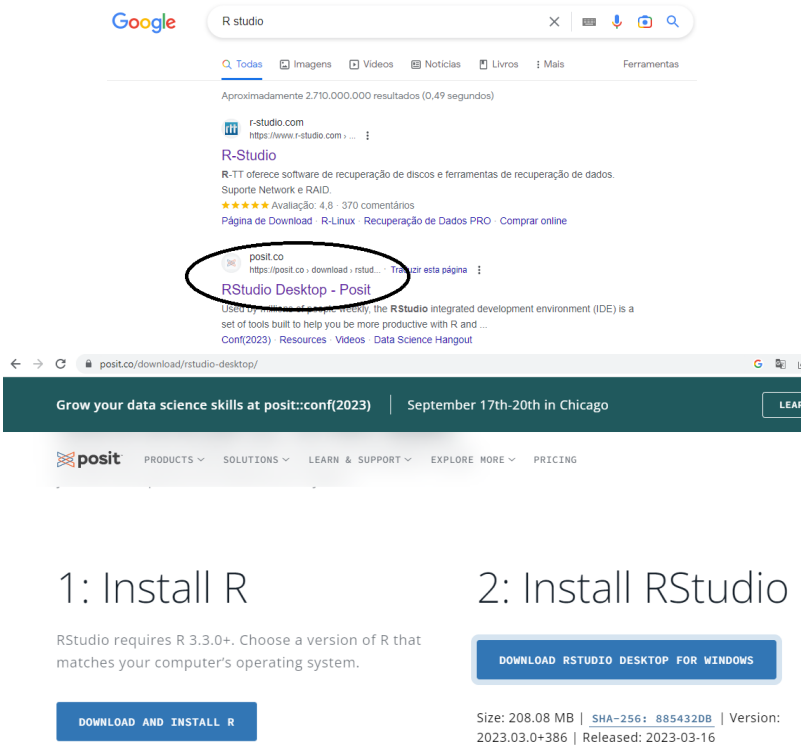
As análises dos dados educacionais propostos neste estudo foram realizadas no software R, versão 4.2.2. O RStudio é um software livre de ambiente de desenvolvimento integrado para a linguagem de programação, gráficos e cálculos estatísticos. É um software que é bastante utilizado por pesquisadores de diversas áreas e por ser um software livre e de fácil acesso, ainda conta com desenvolvedores de programas ou pacotes que auxiliam no avanço das pesquisas. Vejamos, como proceder para instalar o RStudio para um sistema operacional Microsoft Windows.

2.2.1 Instalação

Todas as análises estatísticas gráficas e aplicação da metodologia foram realizadas no software R (R Core Team, 2022). A instalação do RStudio é realizada quase que de forma automática. Abaixo segue o passo a passo para a instalação:

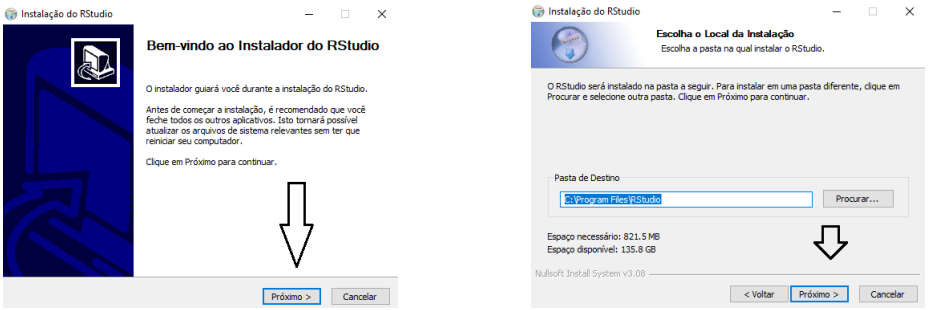
1. Digite em um site de busca: Rstudio. Depois escolha a opção de instalação do software segundo o sistema operacional do seu computador. O processo de instalação abaixo é para a segunda opção, isto é, versão Microsft Windows.

Figura 2.3: Localizando Software Rstudio



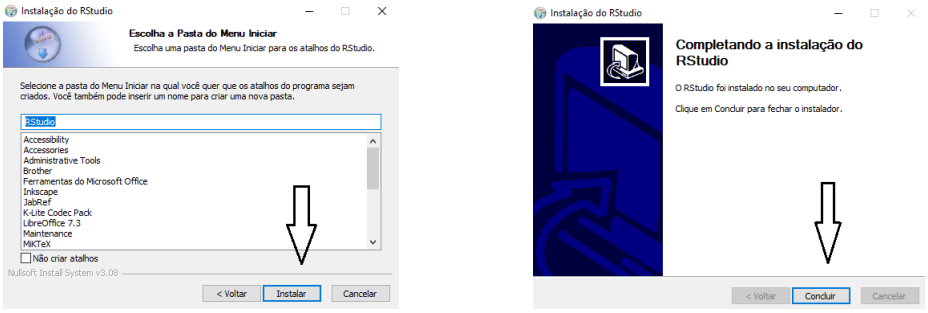
2. Após a escolha por **Install RStudio**, aguarde o dowload ser realizado e comece o processo de instalação realizando dois cliques duplos no arquivo executável.

Figura 2.4: Instalando o Software Rstudio



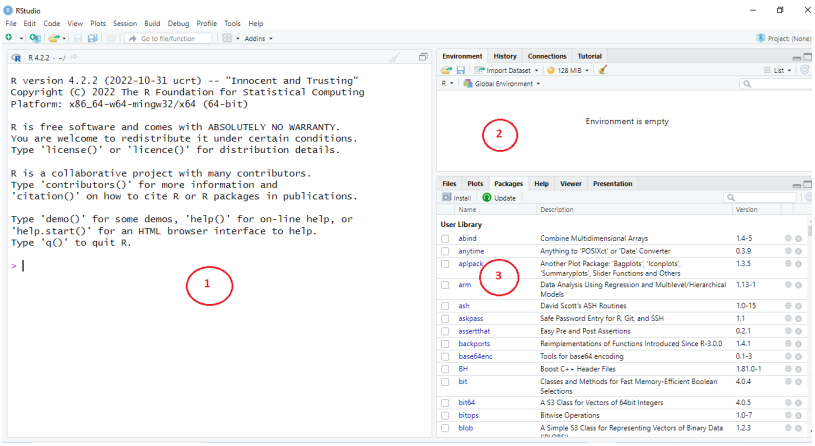
3. Geralmente o software é instalado segundo a opção dada pelo seu computador. Depois é só clicar nos ícones ativos para seguir e concluir os passos de instalação.

Figura 2.5: Concluindo a instalação Software Rstudio



4. Após instalado, localize o Rstudio e vejamos uma visão geral do software. Na tela inicial encontraremos 3 (três) janelas, conforme apresentado abaixo:

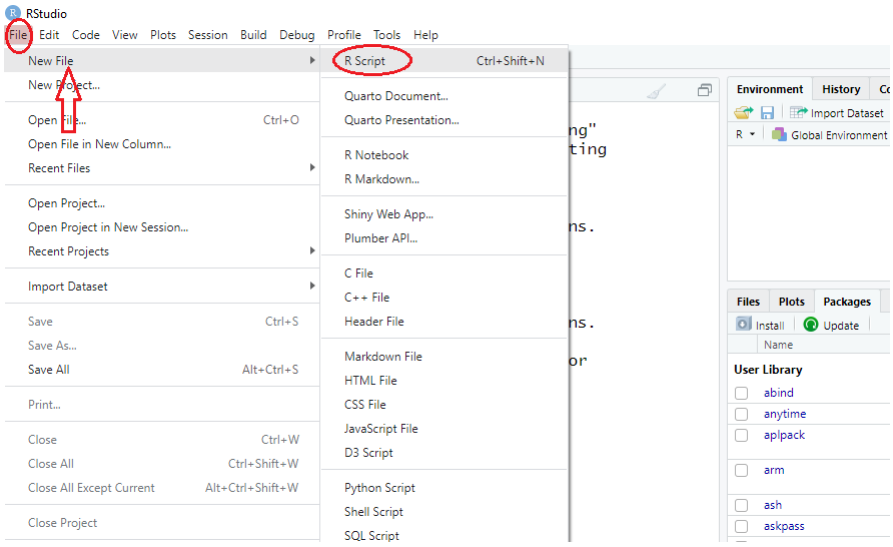
Figura 2.6: Visão geral das janelas inicias do Rstudio



5. Cada janela possui sua funcionalidade. Na primeira janela encontramos os ícones para operacionalização do software, por exemplo: abrir arquivo, visualizar códigos, realizar mudanças na interface do software entre outros. A segunda janela armazena e apresenta os resultados obtidos a partir da compilação da sua rotina. A terceira janela apresenta os pacotes utilizados, locais dos arquivos para fácil acesso, icone de ajuda do sofware entre outros.

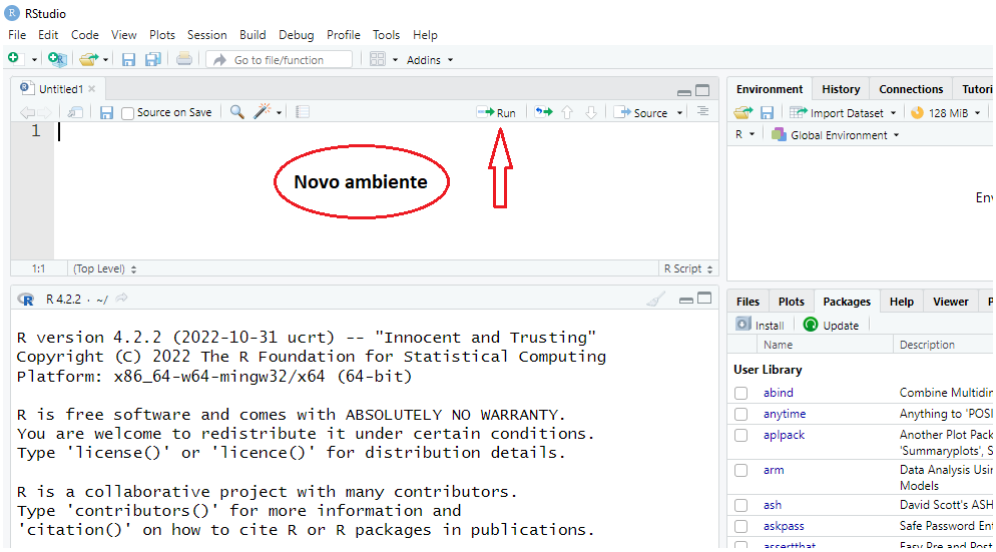
Para gerar um primeiro arquivo em R, devemos abrir um novo arquivo da seguinte forma: File - New File - R Script como apresentado abaixo

Figura 2.7: Novo arquivo no Rstudio



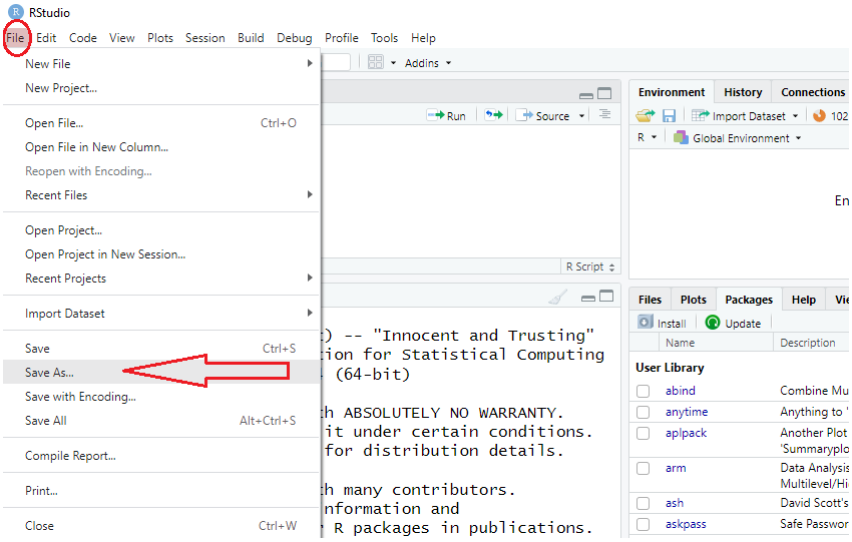
6. Teremos um novo ambiente. É nesse novo espaço que as rotinas computacionais são construídas e para realizar a compilação das rotinas e assim gerar os arquivos em R, basta clicar no ícone como apresentado na figura abaixo:

Figura 2.8: Como compilar rotina no Rstudio



7. Executando File - Salve As as rotinas poderão ser salvas e consultadas a qualquer momento.

Figura 2.9: Como salvar arquivo no Rstudio



2.3 Análise de agrupamento em Estatística

A análise de agrupamento pode ser entendida como um processo que permite descobrir relações existentes entre exemplares de um conjunto de dados descritos por várias características.

No geral as técnicas de agrupamentos se baseiam em similaridades ou diferenças entre os exemplares, a partir das características em análise. Os exemplares que possuem menor "distância" são mais similares e maiores "distância" são ditos dissimilares. Ao final da execução do algoritmo de agrupamento, os grupos são formados de maneira que exemplares similares são agrupados de forma que a similaridade intergrupos tenha sido minimizada e a relação intragrupos tenha sido maximizada.

A tarefa da construção ou da análise de agrupamento tem o interesse no problema de minimizar as medidas de "distância" dentro dos grupos e maximizar a "distância" entre os grupos.

Após a obtenção dos grupos, o analista dos dados conhecerá o que deve ser útil para proceder com análise e entendimento das relações ou grupos encontrados.

Dessa forma, faz-se necessário que o pesquisador proceda com a busca das explicações que ajudem os usuários do modelo a entender os motivos pelos quais os exemplares são similiares/dissimilares.

Segundo Hastie, Tibshirani e Walther (2001), o termo "grupo" deve ser usado quando não existe quaisquer informações sobre como é a organização dos dados. Assim, comumente denomina-se agrupamento o processo pelo qual se estuda as relações de similaridade entre os exemplares, determinando como estão organizados em grupos. As estratégias de agrupamento podem ser dividas em:

- Estratégias hierárquicas: o processo se inicia colocando os exemplares em um único grupo e de forma iterativa, dividi-se os grupos em subgrupos menores e isso ocorre até que todos os exemplares sejam alocados em grupos;
- Estratégias por partição: são criadas partições do grupo. Iterativamente, os exemplares são realocados entre as partições de modo que o modelo de grupo mude e se ajuste melhor ao objetivo da maximização da dissimilaridade entre os grupos;
- Estratégias baseadas em densidade: inicia-se a partição com um grupo pequeno de exemplares e iterativamente, recebem mais exemplares (exemplares que se mostram mais similares aos exemplares do grupo) até que um limiar seja atingindo.

Neste trabalho será utilizada a estratégia por partição. Para encontrar as partições no conjunto de dados de forma que k grupos disjuntos de exemplares sejam descobertos.

O agrupamento $K-$means MacQueen (1967) é o algoritmo de aprendizado de máquina não supervisionado mais comumente usado para particionar um determinado conjunto de dados em um conjunto de k grupos (ou seja, k grupos), em que k representa o número de grupos pré-especificados pelo analista, porém não é qualquer k, é necessário realizar análise para especificar o valor de k.

Ele classifica objetos em múltiplos grupos, de modo que os objetos/exemplares dentro do mesmo grupo sejam tão semelhantes quanto possível (ou seja, alta similaridade intraclasse/intragrupo), enquanto objetos/exemplares de diferentes grupos sejam tão diferentes quanto possível (ou seja, baixa similaridade de interclasse/intergrupo).

No agrupamento $K-$means, cada grupo é representado por seu centro (isto é, centróide) que corresponde à média dos pontos atribuídos ao grupo.

A ideia básica por trás do *clustering* $K-means$ consiste em definir grupos de forma que a variação total intracluster (conhecida como variação total dentro do cluster) seja minimizada.

2.3.1 O algoritmo $K-$means

Existem vários algoritmos $K-$means disponíveis. O algoritmo padrão é o algoritmo Hartigan, Wong et al. (1979), que define a variação total dentro do grupo como a soma das distâncias quadradas das distâncias euclidianas entre os itens e o centróide correspondente:

$$W(G_k) = \sum_{x_i \in G_k} (x_i - \mu_k)^2 \tag{2.1}$$

em que x_i $i-$ésimo exemplar pertencente ao $k-$ésimo grupo; μ_k é o $k-$ésimo ponto médio do $k-$ésimo grupo.

Cada observação (x_i) é atribuída a um determinado grupo de modo que a distância da soma dos quadrados (SQ) da observação ao centróide (μ_k) seja a mínima.

Definimos a variação total dentro do grupo (iremos nomear como ocorre no Software R - WSS) da seguinte forma:

$$WSS = \sum_{k=1}^{K} W(G_k) = \sum_{k=1}^{K} \sum_{x_i \in G_k} (x_i - \mu_k)^2 \tag{2.2}$$

A soma de quadrado total dentro do grupo (WSS) mede a compacidade (ou seja, qualidade) do agrupamento e queremos que seja o menor possível, de forma que menor variação dentro do grupo, maior a compacidade.

O primeiro passo ao usar o *clustering* $K-means$ é indicar o número de grupo (k) que serão gerados na solução final. O algoritmo começa selecionando aleatoriamente k

objetos do conjunto de dados para servir como centros iniciais para os clusters. Os objetos selecionados também são conhecidos como médias de cluster ou centróides. Em seguida, cada um dos objetos restantes é atribuído ao centróide mais próximo, em que o mais próximo é definido usando a distância euclidiana entre o objeto e a média do cluster, definido por:

$$d_{euc}(x, \mu_k) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_k)^2} \tag{2.3}$$

Esta etapa é chamada de "etapa de atribuição de grupo". Observe que, para usar a distância de correlação, os dados são inseridos como escores z, ou seja padronizados (centrados na média e com variância um).

Após a etapa de atribuição, o algoritmo calcula o novo valor médio de cada grupo. O termo "atualização do centróide" do grupo é usado para projetar esta etapa.

Agora que os centros foram recalculados, cada observação é verificada novamente para verificar se pode estar mais próxima de um grupo diferente.

Todos os objetos são reatribuídos novamente usando o cluster atualizado.As etapas de atribuição do grupo e atualização do centróide são repetidas iterativamente até que as atribuições do grupo parem de mudar (ou seja, até que a convergência seja alcançada). Assim, os *clusters* formados na iteração atual são os mesmos obtidos na iteração anterior. O algoritmo $K-$means pode ser resumido da seguinte forma:

1. Utilizando informações do conjunto de dados e uma metodologia para seleação de K, o pesquisador especifica o número de grupo (K) a serem criados;

2. Selecione aleatoriamente k objetos/exemplares do conjunto de dados como os centros ou meios iniciais do grupo;

3. Atribui cada observação ao centróide mais próximo, com base na distância euclidiana entre o objeto e o centróide;

4. Para cada um dos k grupos,atualize o centróide do grupo calculando os novos valores médios de todos os pontos de dados no grupo. O centróide de um $k-$ ésimo agrupamento é um vetor de comprimento p contendo as médias de todas as variáveis para as observações no $k-$ ésimo agrupamento; p é o número de variáveis;

5. Minimize, iterativamente, a soma de quadrado total dentro do grupo. Ou seja, repita as etapas 3 e 4 até que as atribuições do grupo parem de mudar ou o número máximo de iterações seja atingido. Por padrão, o software R usa 10 como valor padrão para o número máximo de iterações. Mas, é possível inserir mais iterações.

2.3.2 Seleção do número ótimo de grupo

Para a seleção do número ótimo de grupo, há na literatura algumas opções. Os métodos disponíveis no software R são:

- Método Elbow: é bastante usual e pode ser usado para o pesquisador ter uma indicação do número de grupos. Consiste na interpretação de um gráfico de linhas com formato de "cotovelo."O número de grupos é onde o cotovelo se dobra, apresentando a menor soma de quadrados totais dentro do grupo(WSS), indicando menor variação dentro do grupo. O eixo horizontal do gráfico corresponde ao número de grupos e o eixo vertical é a soma dos quadrados dentro dos grupos para cada número de grupos (SYAKUR et al., 2018).

- Método Silhouete: A análise de silhueta gráfica pode ser usada para estudar a distância de separação entre os grupos resultantes. O gráfico de silhueta do conjunto de dados exibe uma medida do quão próximo cada ponto (objeto) em um grupo está dos pontos (objetos) dos grupos vizinhos e, portanto, fornece uma maneira de avaliar o número de clusters visualmente. Esta medida tem um intervalo de $[-1, 1]$. Coeficientes de silhueta (como esses valores são chamados) próximos a $+1$ indicam que a amostra está distante dos conglomerados vizinhos. Um valor igual a zero indica que a amostra está no ou muito próximo do limite de decisão entre dois grupos vizinhos e valores negativos indicam que essas amostras podem ter sido atribuídas ao grupo errado (ROUSSEEUW, 1987).

- Método da Estatística Gap: A abordagem pode ser aplicada a qualquer método de agrupamento. A estatística Gap compara a variação intragrupo total para diferentes valores de k grupos com seus valores esperados sob distribuição de referência dos dados (isto é, uma distribuição sem agrupamento óbvio). O conjunto de dados de referência é gerado usando processos de amostragens através de simulações computacionais (TIBSHIRANI; WALTHER; HASTIE, 2001).

Como o $K-$means tem por objetivo particionar o conjunto em grupos de forma que a variação dentro de cada grupo seja a mínima, ou seja, que a soma de quadrado dentro do grupo (WSS) seja a mínima e que a variação entre os grupos seja maximizada, dessa forma a ideia é calcular o agrupamento $k-$means usando diferentes valores de clusters k e investigar que número de grupos produzirá melhor explicação para o conjunto de dados.

2.3.3 Visualização dos grupos k-means

A visualização gráfica da análise de agrupamento facilitará a interpretação numérica dos resultados do grupos. Os gráficos são utilizados para avaliar a escolha do número de grupos, bem como comparar duas análises de agrupamentos diferentes.

Note que os conjuntos de dados, possuem mais de duas variáveis em estudo. Realizar uma análise gráfica bidimensional tradicional, considerando todas as variáveis, torna-se impossível. Dessa forma, em análise multivariada, existem técnicas gráficas para execução gráfica nesses casos.

Para visualização dos dados multivariados, que contém mais de duas variáveis, em um gráfico de dispersão (plano cartesiano) deve-se escolher as variáveis duas a duas para construção gráfica. Mas, a proposta deste trabalho está em apresentar soluções de análises estatísticas considerando de maneira geral e unificada, todas as informações apresentadas no conjunto de dados.

Uma solução para a apresentação gráfica, considerando todas as variáveis, é reduzir o número de dimensões do conjunto de dados. As dimensões em multivariada está associada a explicação da variabilidade do conjunto de dados. Por exemplo, duas novas dimensões podem explicar o que ocorre com a variabilidade dos dados em mais de 70%.

Em outras palavras, se tivermos um conjunto de dados multidimensional, uma solução é obter a redução da dimensionalidade do conjunto de dados e plotar os pontos de acordo com as dimensões que maior explicam a variabilidade dos dados.

Dessa forma, como maiores detalhes, na seção 3.1.1 serão apresentadas as análise gráficas multivariadas utilizadas neste trabalho.

2.3.4 Validação estatística dos grupos

O termo validação dos grupos é usado para verificar a qualidade do ajuste dos grupos construídos pelos algoritmos de agrupamento, ou seja, para o pesquisador verificar o melhor agrupamento que pode ser realizado com a metodologia $k-$means. Isso é importante para evitar encontrar padrões em dados aleatórios, bem como na situação em que se deseja comparar dois algoritmos.

Geralmente, as estatísticas de validação de agrupamento podem ser categorizadas em 3 classes, segundo (THEODORIDIS; KOUTROUMBAS, 2006; CHARRAD et al., 2014; BROCK et al., 2008):

1. Validação de agrupamento interno, que usa as informações internas do processo de agrupamento para avaliar a qualidade de uma estrutura de agrupamento sem referência a informações externas. Também pode ser usado para estimar o número de grupos e o algoritmo de agrupamento apropriado sem nenhum dado externo.

2. Validação de agrupamento externo, que consiste em comparar os resultados de uma análise de agrupamento com um resultado conhecido externamente, como rótulos de classe fornecidos externamente. Ele mede até que ponto os rótulos de grupos correspondem aos rótulos de classe fornecidos externamente. Identificando o número

de grupos, essa abordagem é usada principalmente para selecionar o algoritmo de agrupamento correto para um conjunto de dados específico.

3. Validação de agrupamento relativo, que avalia a estrutura de agrupamento variando diferentes valores de parâmetros para o mesmo algoritmo (por exemplo: variando o número de agrupamentos k). Geralmente é usado para determinar o número ótimo de grupos.

As medidas de validação interna geralmente refletem a compacidade, a conectividade e a separação das partições do grupo. Esse critério tem por objetivo encontrar grupos com um subconjunto de objetos/exemplares que maximize a similaridade intragrupos, e uma forma bastante simples de medir a compacidade é usar a medida estatística de variância, de forma que, quanto menor a variância dos exemplares dentro de um grupo, maior a sua compacidade.

- Compacidade ou coesão do grupo: mede o quão próximos estão os objetos dentro do mesmo cluster.Uma variação menor dentro do cluster é um indicador de uma boa compactação (ou seja, um bom agrupamento). Os diferentes índices para avaliar a compactação de grupos são baseados em medidas de distância, como o as distâncias dentro dos grupos entre as observações.

- Separação: Mede o quão bem separado um grupo está de outros grupos. Os índices usados como medidas de separação incluem: distâncias entre centros de grupos e as distâncias mínimas emparelhadas entre objetos em diferentes grupos.

- Conectividade: corresponde à medida em que os itens são colocados no mesmo grupos que seus vizinhos mais próximos no espaço de dados. A conectividade tem um valor entre 0 e infinito e deve ser minimizada.

A validação de agrupamento baseado em critérios externos depende da existência de algum conhecimento sobre o agrupamento do conjunto de dados sob análise.

Geralmente, a maioria dos índices usados para validação de agrupamento interno combina medidas de compacidade e separação como segue:

$$I = \frac{(\alpha \times \text{separação})}{(\beta \times \text{compacidade})}$$

em que α e β são pesos atribuídos à separação e compacidade.

Índice de Dunn

O índice de Dunn (introduzido por JC Dunn em 1974), é uma medida de validação de agrupamento interno que pode ser calculada seguindo os seguintes passos:

1. Para cada grupo, calcule a distância entre cada um dos objetos no grupo e os objetos nos outros grupos;

2. Use o mínimo dessa distância de pares como a separação entre grupo;

3. Para cada grupo, calcule a distância entre os objetos no mesmo grupo;

4. Use a distância intragrupo máxima (ou seja, diâmetro máximo) como a compacidade intragrupo;

5. Calcule o índice de Dunn (D) da seguinte forma:

$$I_{Dunn} = \frac{\text{menor separação}}{\text{diâmetro máximo}}$$

Se o conjunto de dados contiver grupos compactos e bem separados, espera-se que o diâmetro dos grupos seja pequeno e que a distância entre os grupos seja grande. Assim, o índice de Dunn deve ser maximizado.

Dessa forma, quanto maior o índice, melhor será a separabilidade entre os grupos e a compacidade de grupos.

Índice Silhouette

A análise de silhueta do conjuntos de dados mede o quão bem uma observação é agrupada e estima a distância média entre os agrupamentos. O gráfico de silhueta exibe uma medida de quão próximo cada ponto no grupo está dos pontos nos grupos vizinhos. Para cada observação i, a largura da silhueta s_i é calculada da seguinte forma (ROUSSEEUW, 1987):

1. Para cada observação i, é calculada a média de dissimilaridade a_i entre i e todos os outros pontos do grupo ao qual o objeto/exemplar i pertence;

2. Para todos os grupos C, em que i não faz parte, calcule-se a média de dissimilaridade $d(i, C)$ de i, para todas as observações de C. A menor $d(i, C)$ é definida como $b_i = min_C d(i, C)$.

   O valor de b_i pode ser visto como a dissimilaridade entre i e os vizinhos do grupo, isto é, o ponto mais próximo ao qual não pertence ao grupo.

3. Finalmente a largura da silhueta da observação i é definida por:

$$S_i = I_{Sl} = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

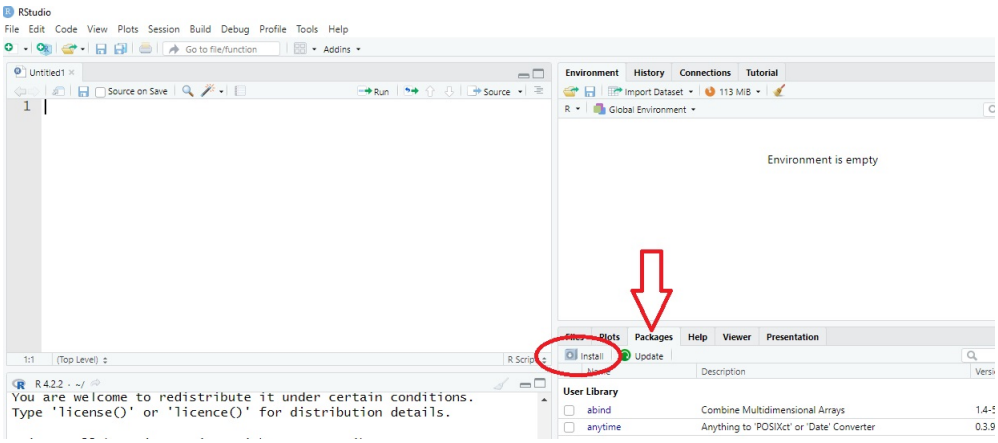
Podemos interpretar o índice silhueta da seguinte forma:

- I_{sl} grande (quase 1) indica que estão bem definidos os agrupamentos.
- Um I_{sl} pequeno (em torno de 0) significa que a observação está entre dois grupos.
- As observações com um I_{sl} negativo são objetos, provavelmente inseridos no grupo errado. E valores próximo de -1 indicam que a organização do grupo não está boa.

2.3.5 Comandos do R

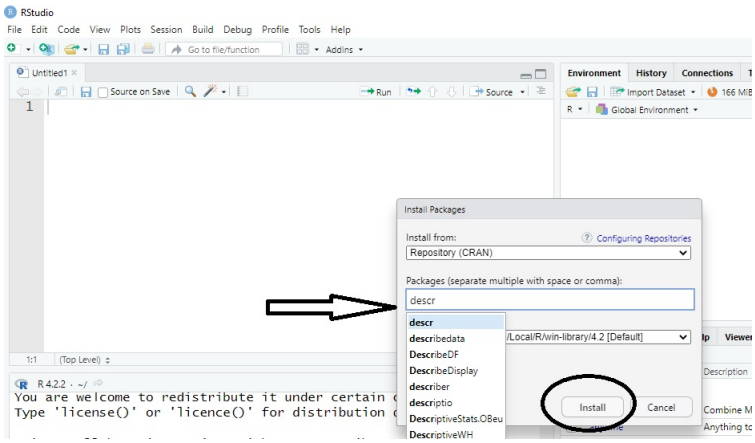
Para a aplicação da metodologia e construções gráficas será necessária a instalação dos seguintes pacotes: **descr, cluster, FactoMineR, fpc, factoextra e tidyverse**. Para a instalação dos pacotes deve-se seguir os passos descritos nas Figuras (2.10) e (2.11).

Figura 2.10: Instalando pacotes para o uso da metodologia



Após solicitar a instalação do pacote no ícone "install" uma nova janela irá se abrir. Digite o nome do pacote desejado e depois clique em "install" como apresentado na Figura (2.11).

Figura 2.11: Selecionando o pacote para instalação



Capítulo 3

Análise dos Dados e Discussão dos Resultados

3.1 Análise descritiva

Considerando os microdados referentes ao ano escolar 2018, será discutido a proposta da metodologia $K-$means, bem como as discussões dos resultados. Para uma proposta mais atualizada, será apresentada ao final dessa seção, uma análise multivariada para o ano escolar 2022. Vale ressaltar, que só foi possível realizar comparações mínimas entre os dois anos, 2018 e 2022. Isso ocorreu devido ao fato da mudança do cálculo do número de média de horas aulas ocorrido no último ano, bem como a ausência de algumas taxas encontradas no ano de 2018 que não estavam disponíveis no ano de 2022.

Segundo o censo escolar, em 2018 houve um acréscimo no número de matrículas em tempo integral e no ensino médio. O Censo Escolar de 2018 registrou $48,5$ milhões de matrículas nas $181,9$ mil escolas de educação básica brasileiras. São $1,3$ milhão estudantes a menos que em 2014, o que representa uma redução de $2,6\%$ em cinco anos. Só no ensino médio o número total de matrículas reduziu $7,1\%$. E Segundo Carlos Eduardo Moreno Sampaio, diretor de Estatística Educacionais do Inep, o total de matrículas do ensino médio segue tendência de queda nos últimos anos. "Isso se deve tanto a componentes demográficos, quanto à melhoria no fluxo no ensino médio, no qual a taxa de aprovação subiu três pontos percentuais de 2013 a 2017. A queda também pode ser explicada pelas altas taxas de evasão e da migração de alunos para a Educação de Jovens e Adultos (EJA)", explica e conforme encontra-se na reportagem (INEP, 2019).

Ainda de acordo com o censo escolar de 2018, distorção idade-série se torna mais intensa a partir do terceiro ano do ensino fundamental e se acentua também no sexto ano do ensino fundamental e na primeira série do ensino médio. A taxa de distorção idade-série alcançou $11,2\%$ das matrículas nos anos iniciais do ensino fundamental, $24,7\%$ nos anos finais e $28,2\%$ no ensino médio. Além disso, a taxa de distorção do sexo masculino

é maior que a do sexo feminino em todas as etapas de ensino.

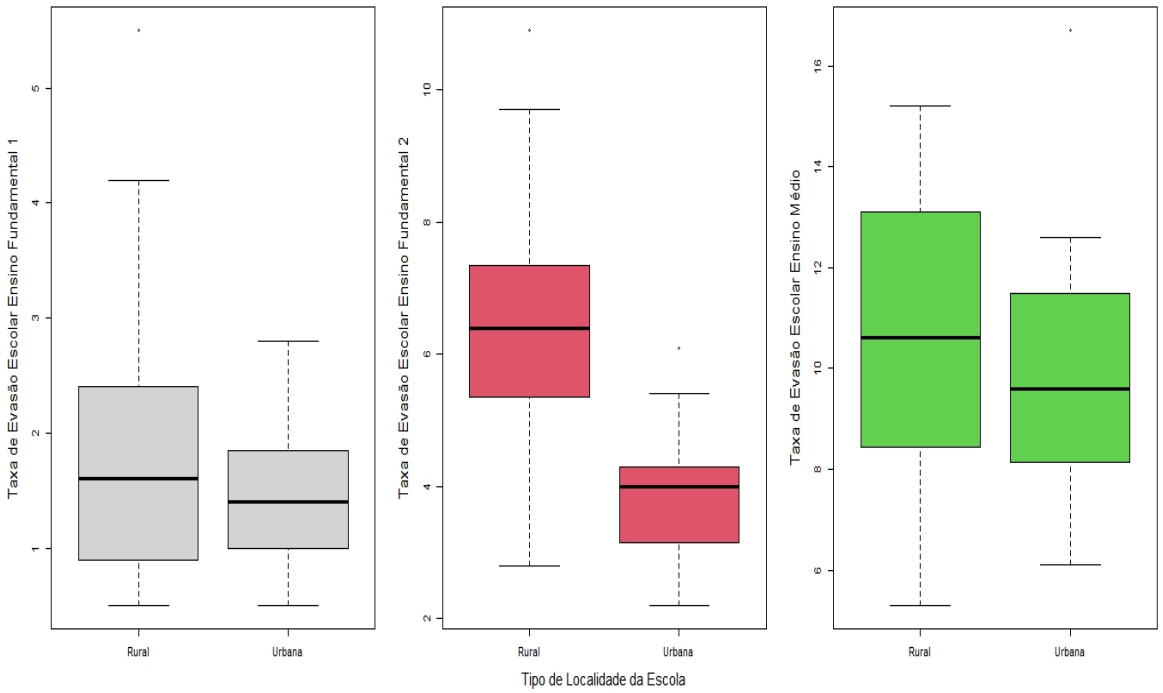
A Tabela 3.1 apresentar algumas estatística descritivas numéricas do conjunto de dados para o ano escolar 2018 para todos os estados brasileiros. Nota-se que a média da taxa de promoção é decrescente quando comparada com os ensinos fundamental e médio e que as taxas de evasões escolares para o ensino fundamental das séries finais é aproximadamente quatro vezes às taxas para o fundamental das séries iniciais.

Tabela 3.1: Estatísticas descritivas - Conjunto de dados ano 2018

Código (Descrição)	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo	Desvio Padrão
EF1_TXP (taxa de promoção - ensino fundamental 1)	78,50	87,30	90,35	90,03	93,28	97,50	4,8108
EF2_TXP (taxa de promoção - ensino fundamental 2)	69,0	77,78	81,30	82,09	87,38	93,70	6,1554
EM_TXP (taxa de promoção - ensino médio)	67,70	75,03	78,10	78,46	81,90	91,80	5,0211
EF1_TXR (taxa de repetência - ensino fundamental 1)	2,0	5,825	7,350	8,094	9,600	17,900	3.9453
EF2_TXR (taxa de repetência - ensino fundamental 2)	2,7	6,775	10,100	10,113	12.350	19,500	4.0275
EM_TXR (taxa de repetência - ensino médio)	2,5	6,850	8,950	8.709	10,450	14.400	2.7145
EF1_TXE (taxa de evasão - ensino fundamental 1)	0,500	1,00	1,500	1,665	2,075	5,500	0.9882
EF2_TXE (taxa de evasão - ensino fundamental 2)	2,200	3,925	4,800	5,170	6,350	10,900	1.9408
EM_TXE (taxa de evasão - ensino médio)	5,300	8,125	10,400	10,287	12,050	16,700	2.6088
EF1_MHA (média de horas aula - ensino fundamental 1)	4,100	4,200	4,350	4,409	4,500	5,300	0.2679
EF2_MHA (média de horas aula - ensino fundamental 2)	4,100	4,300	4,500	4,587	4,775	5,500	0.3566
EM_MHA (média de horas aulas - ensino médio)	3.800	4,800	5,100	5,217	5,600	7,400	0.6837

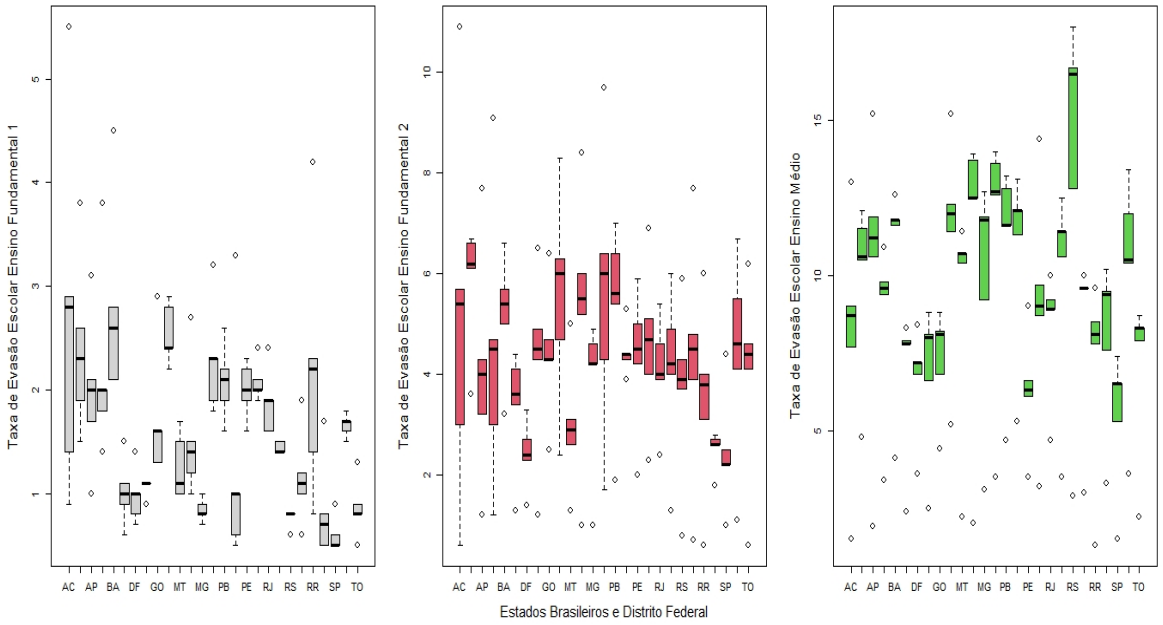
Para melhor visualizar as taxas de evasão escolar dos ensinos fundamental e médio a Figura 3.1 apresenta as estatísticas descritivas que auxiliam no entendimento da variabilidade do conjunto de dados.

Figura 3.1: Boxplot das taxas de evasão escolar do ensino fundamental e médio



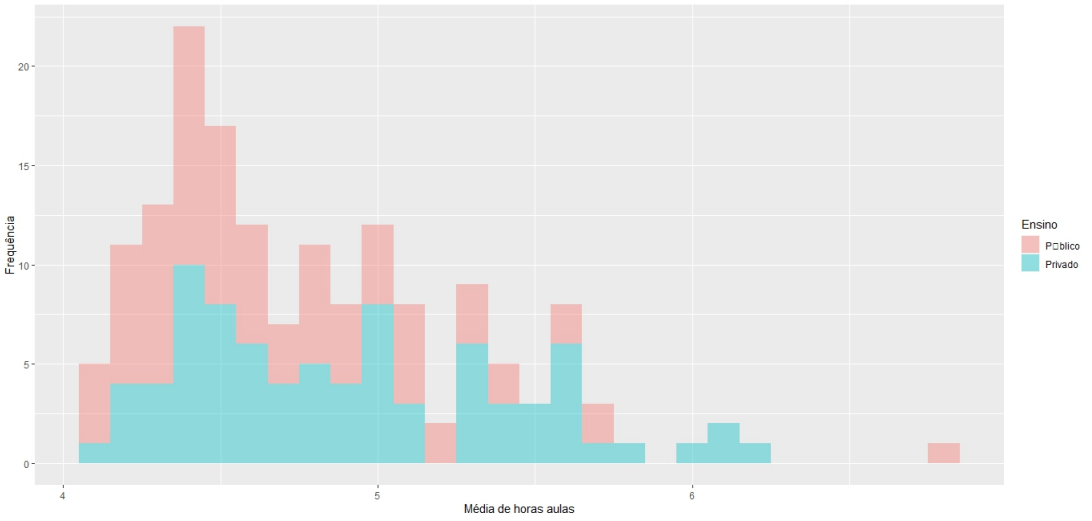
Segundo a Figura 3.1 nota-se que nos anos iniciais a variabiliade referente a taxa de evasão é maior na zona rural e que de forma semelhante a mesma pode ser observada no ensino médio. Para os anos finais do ensino fundamental a mediana das taxas de evasao escolar na zona rural apresenta-se maior que na zona urbana.

Figura 3.2: Boxplot das taxas de evasão escolar nos estados por níveis de ensino



Pela Figura 3.2 observa-se que a taxa de evasão escolar nos estados é crescente quando comparamos com os niveis de ensino. Observe que o grafico referente ao ensino médio apresenta, de forma geral, medianas maiores com relação ao ensino fundamental.

Figura 3.3: Distribuição de frequência de média de horas aula



Na Figura 3.3 que representa a quantidade de estados e o número de horas aulas, observa-se, por exemplo, que na maioria dos estados, no que se refere a rede publica, a quantidade de horas aula é menor do que na rede privada. Observa-se também que há um estado que apresenta o maior número de média de horas com oferta de ensino público, o estado de Pernambuco com 6,8 horas aulas.

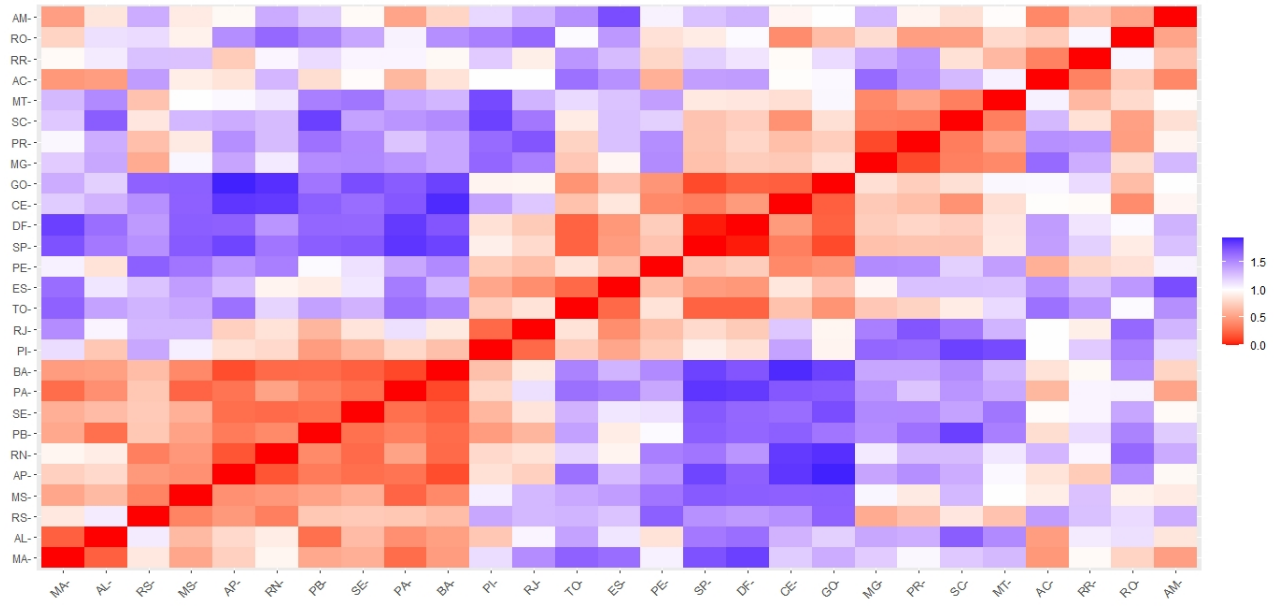
3.1.1 Análise multivariada: aplicação da metodologia

Como definido na metodologia $k-$means a representação das similaridades dos estados são apresentadas através da matriz de similiridade, ou seja, uma matriz que representa as correlações entre os estados brasileiros de acordo com as variáveis descritas na Tabela 3.1.

Alta similaridade é representada na cor vermelha e baixa similaridade na cor azul. Para os elementos na diagonal secundária da matriz tem-se altas similaridades, pois representam as correlaçoes entre os próprios estados.

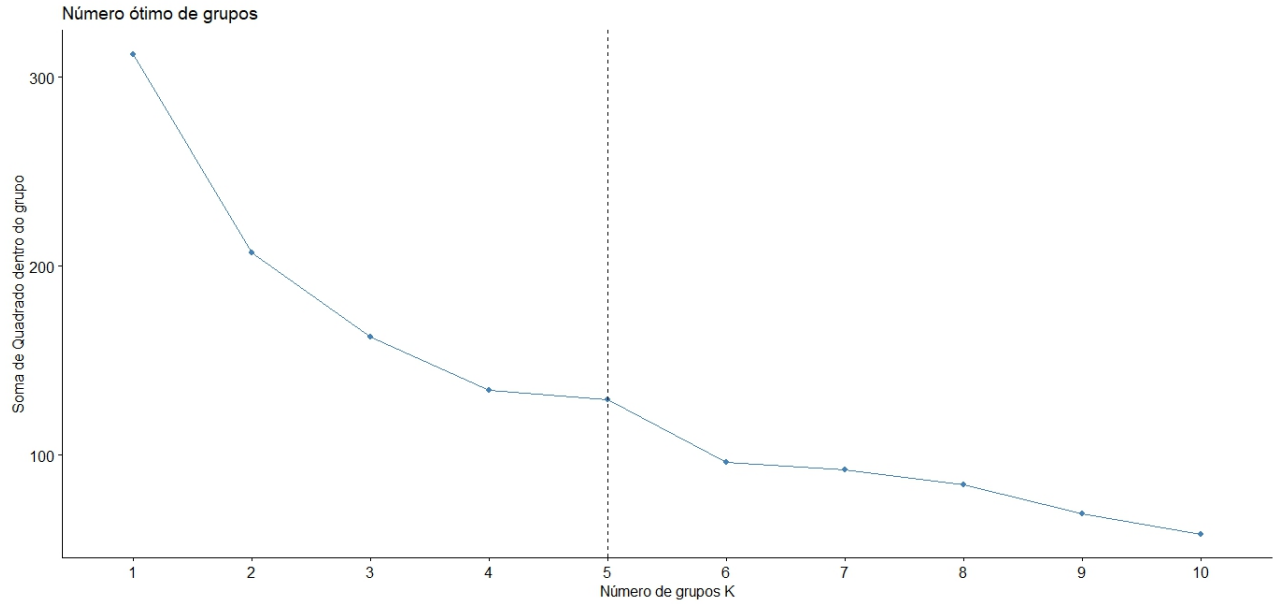
Na Figura 3.4 podemos observar grupos de estados que apresentam uma alta similaridade em relação as variáveis estudadas. Temos como exemplo os estados de (MG, PR, SC, MT), (BA, PA, SE, PB). Já com baixa similaridade observa-se, o grupo de estados (GO, CE, DF, SP).

Figura 3.4: Gráfico da matriz de similaridades dos dados educacionais



Para a seleção do número ótimo de grupos realizou-se a escolha gráfica pelo Método Elbow. O número ótimo de grupos é escolhido levando em consideração que a soma de quadrados dentro do grupo deve ser minimizada, o valor a ser encontrado é observado no eixo horizontal como podemos ver na Figura 3.5.

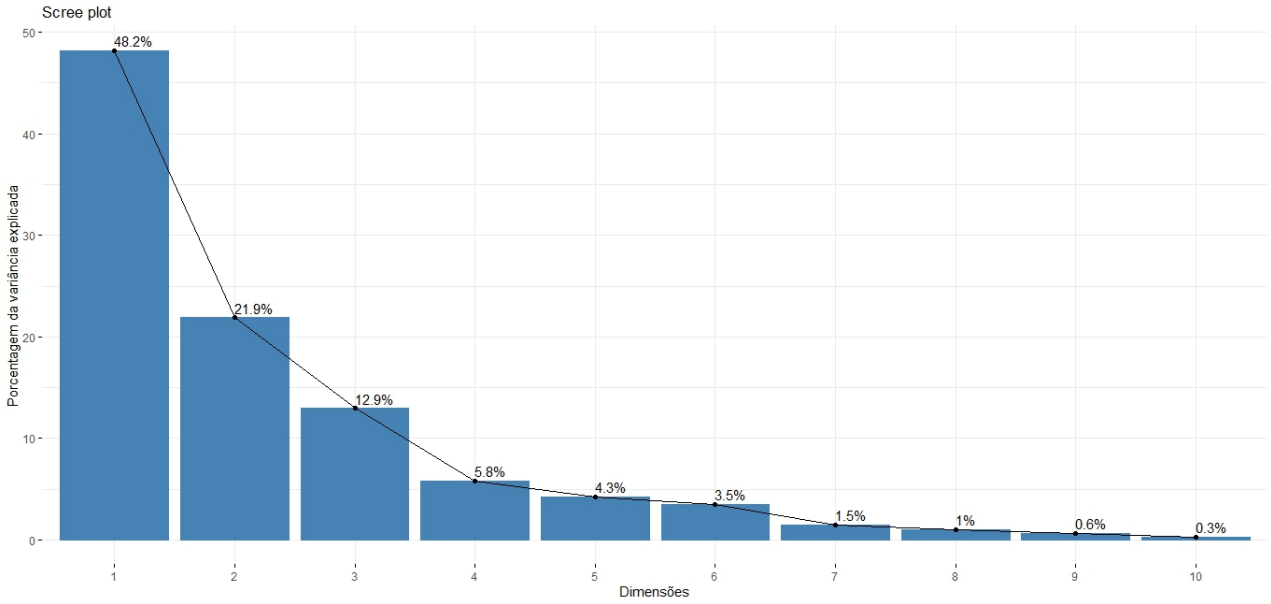
Figura 3.5: Escolha do número de grupos - Método Elbow



Dessa forma, optou-se por realizar a análise com 5 grupos, pois observou-se que nessa configuração tem-se $63,9\%$ da explicação total do conjunto de dados. Esse percentual refere-se ao quociente entre a soma quadrados dentro dos grupos e a soma de quadrados totais obtidas após a construção dos grupos.

O conjunto de dados possui 12 variáveis, dessa forma para uma aprensentação gráfica bidimensional, faz-se necessário a redução da dimensão do conjunto de dados. Assim, a Figura 3.6 possibilita visualizar as escolhas pelo número de dimensões.

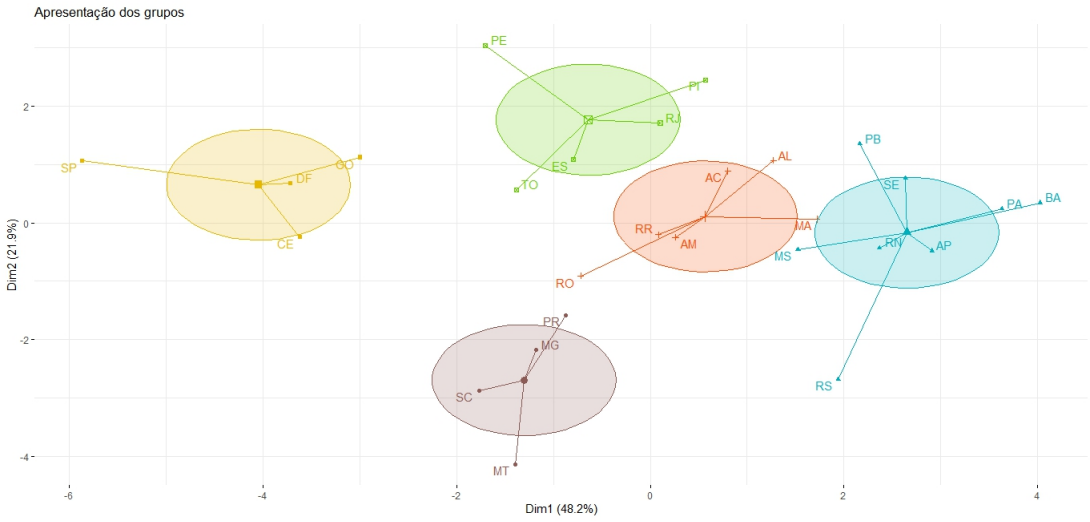
Figura 3.6: Gráfico do percentual de variabilidade explicada



O gráfico de agrupamento apresenta a disposição dos grupos. Pode-se observar que as duas primeiras dimensões explicam $70,1\%$ da variabilidade total do conjunto de dados, conforme Figura 3.6.

Na Figura 3.7 nota-se, por exemplo que o Estado Piauí agrupou-se aos estados do Rio Janeiro, Espírito Santo, Tocantis e Pernambuco.

Figura 3.7: Disposição do agrupamento dos estados

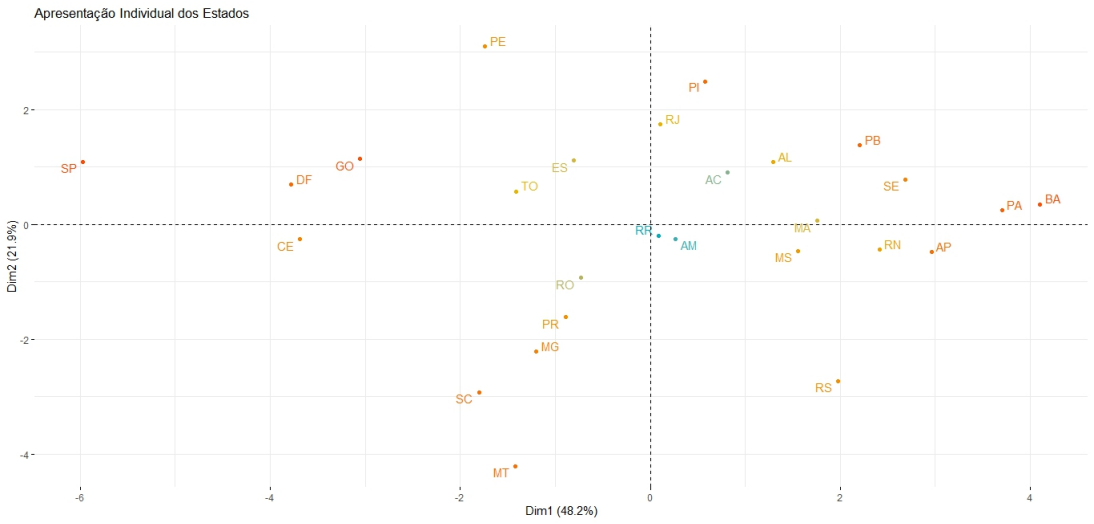


Vale notar que os referidos estados pertencem a regiões geográficas diferentes. Isso nos faz crer que dada as variáveis selecionadas para esse estudo, as questões relativas as

taxas de evasão, promoção, aprovação, reprovação e média de horas aulas não são aspectos unicamente de localização geográfica.

De forma semelhante à matriz de similaridade, observou-se na Figura 3.8 que os estados do Amazonas e Roraima são os mais dissimilares quando comparados aos demais estados da federação brasileira.

Figura 3.8: Apresentação individual dos estados



No gráfico da Figura 3.9 apresenta-se de forma conjunta os estados e as variáveis do estudo. O gráfico biplot auxilia na análise exploratória da relação entre os estados e as variáveis.

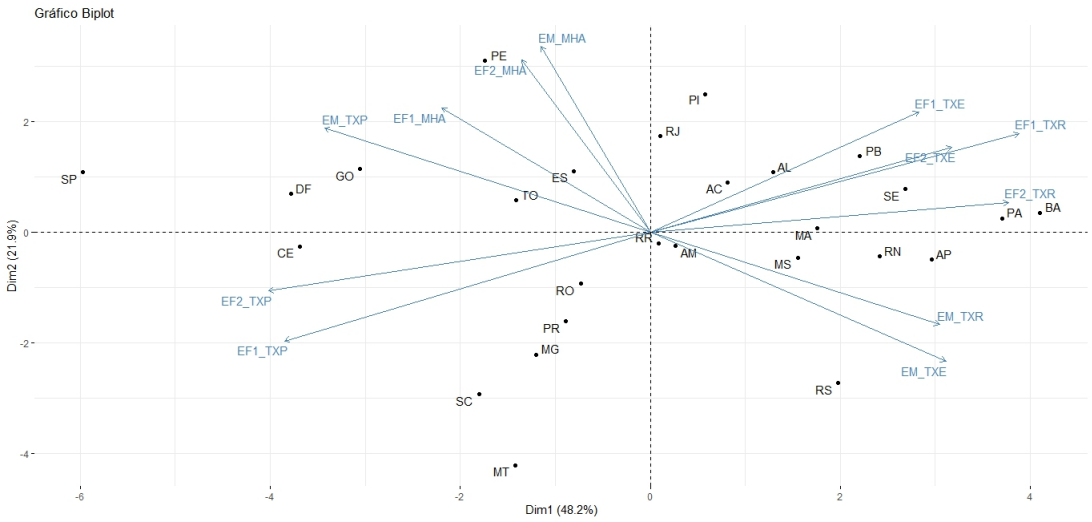
Analisando o comprimento das linhas dos vetores, que representam as variáveis, nota-se que o nível de variação das variáveis são semelhantes. Com relação a representação dos ângulos entre as variáveis taxa de evasão ($EF1\_TXE$ e $EF2\_TXE$) e repetência ($EF1\_TXR$ e $EF2\_TXR$) observa-se que o ângulo formado entre as duas linhas de vetores é menor que 90 graus, portanto as referidas variáveis possuem correlações positivas.

No sentido contrário quando observamos as taxas de promoção do ensino fundamental 1 e 2 com relação as taxas de evasão nota-se que as variáveis são correlacionadas negativamente, pois o ângulo entre elas é maior do que 90 graus.

Observando no gráfico biplot a relação de proximidade entre os estados e as variáveis, temos por exemplo que a variável que mais contribuiu para alocar os estados do Pará e Bahia no mesmo grupo foi a taxa de repetência.

No grafico biplot da Figura 3.9 as variáveis que contribuíram para a similaridade entre os estados Paraiba (PB), Sergipe (SE) e Bahia (BA) foram as taxas de evasão e repetência. Para os estados de Pernambuco (PE) e Espírito Santo (ES) o que melhor descreve a variação é a média de horas aulas, assim como a taxa de promoção descreve melhor a variação para o estado do Ceará (CE).

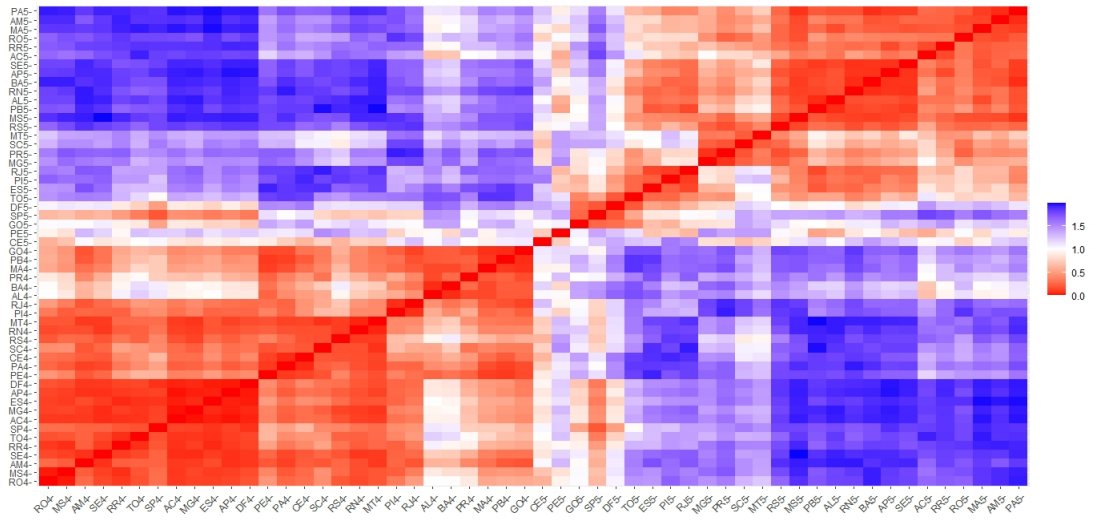
Figura 3.9: Gráfico Biplot para o estudo das relações entre os estados e as variáveis consideradas no estudo.



A proposta da BNCC é unificar com equidade o ensino em todo territorio nacional. Dada a existência de diferenças regiões entre a oferta e ensino na rede pública e privada no Brasil, resolveu-se proceder com as análises do conjunto de dados referentes a rede pública e privada.

3.2 Analisando o conjunto de dados para rede pública e privada

Figura 3.10: Matriz de silmilaridade para rede pública e privada



Para proceder com a análise realizou-se a seguinte codificação: Nome do estado e o número 4 representam rede privada (por exemplo, para o estado do Pará rede privada - **PA4**) e para Nome do Estado e o número 5 será rede pública(por exemplo, para o estado do Pará rede pública - **PA5**).

No grafico da Figura 3.10 os estados Rondônia (RO), Mato Grosso do Sul (MS), Amazonas (AM), Sergipe (SE), Roraima (RR), Tocantins (TO), São Paulo (SP), Acre (AC), Minas Gerais (MG), Espírito Santo (ES), Amapá (AP) no que diz respeito a rede privada caracterizada pelo número 4, eles apresentam alta similaridade, ou seja, a oferta do ensino privado nesses estados, segundo as variáveis em estudo, são semelhantes.

No caso da rede pública de ensino, observou-se que os estados Rio Grande do Sul (RS), Mato Grosso do Sul (MS), Paraíba (PB), Alagoas (AL), Rio Grande do Norte (RN) e Bahia (BA) apresentam oferta de ensino da rede pública semelhante segundo as variáveis em estudo.

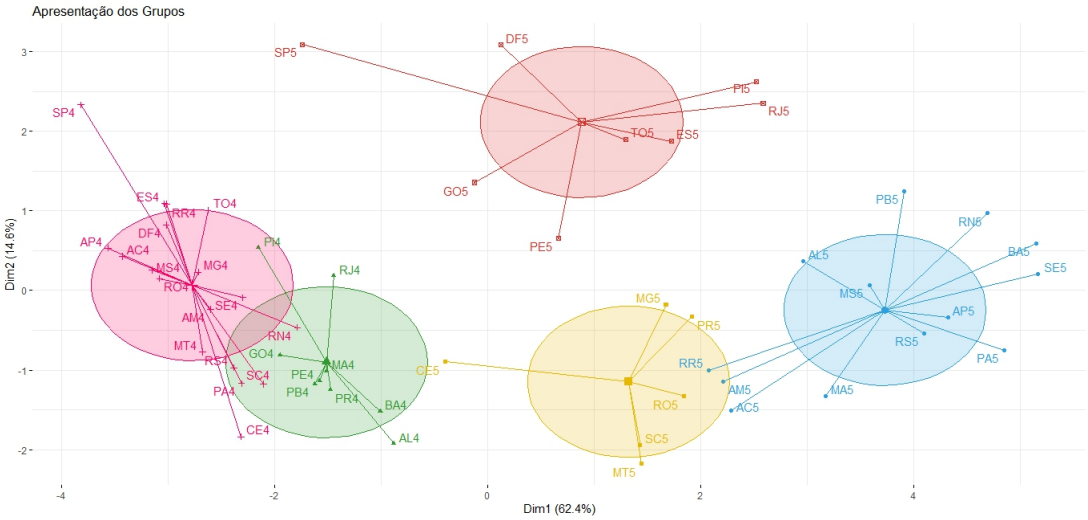
Pode-se observar, também, que o gráfico da Figura 3.10 apresenta uma divisão entre a oferta de ensino público e privado caracterizada pela apresentação de quatro quadrantes bem definidos.

Utilizando o Método Elbow para seleção do número de grupos, escolheu-se realizar a análise com 5 grupos de forma que a explicação total do conjuntos de dados foi de $72,6\%$.

No gráfico da Figura 3.11 os estados de SP, DF, GO, PE, TO, ES, RJ e PI com oferta de ensino de rede pública formaram um grupo bem definido. Nota-se também que os estados RO, AM, e AC são similares ao grupo representado na cor laranja, porém não o bastante a ponto pertencer a ele.

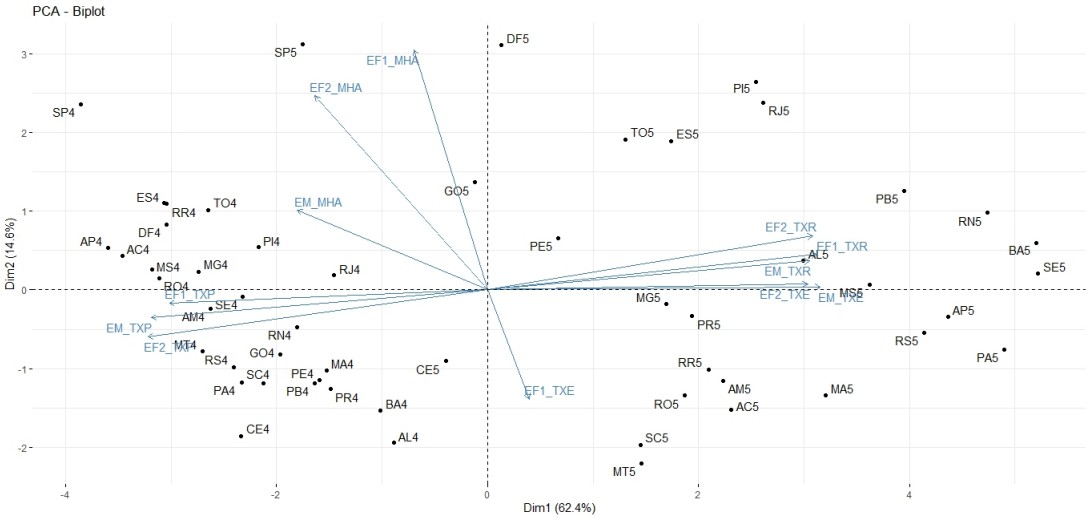
Nota-se, ainda, que houve uma separação em dois grandes grupos: grupos da rede pública, localizado no lado direito do gráfico e os grupos da rede privada que estão do lado esquerdo.

Figura 3.11: Apresentação dos estados nos grupos para os ensinos público e privado



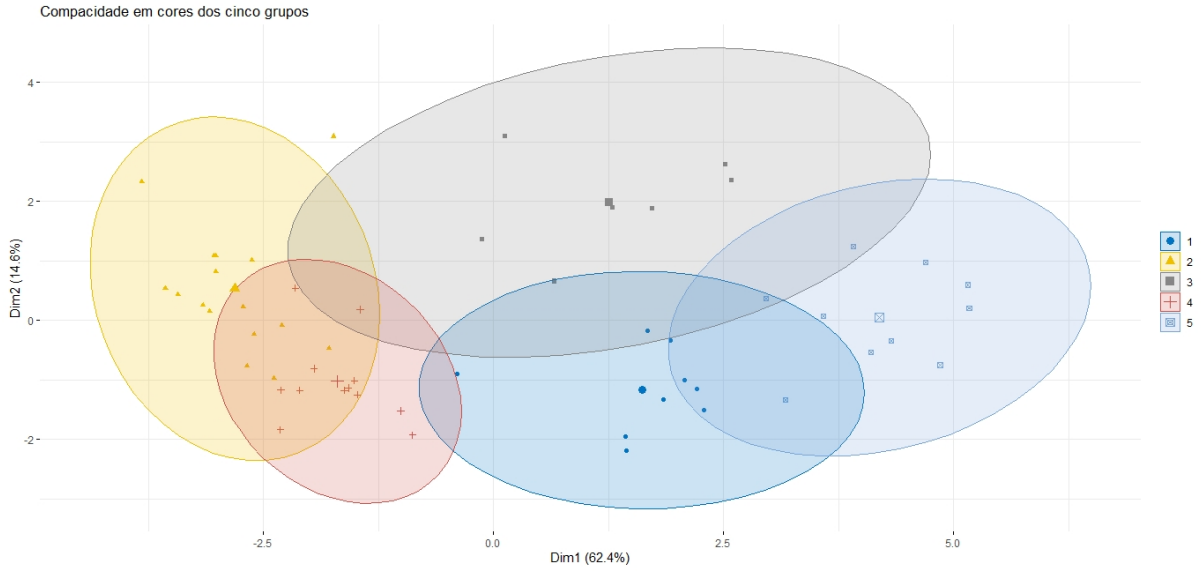
Podemos observar no gráfico da Figura 3.12 que os comprimentos das linhas dos vetores das variáveis taxa de evasão para o ensino fundamental 1 e média de horas aula para o ensino médio possuem menor variação.

Figura 3.12: Gráfico Biplot para os ensinos das redes pública e privada



No sentido contrário, as demais taxas apresentam maiores variações. Contudo, as taxas de repetência para os ensinos fundamental 1, 2 e médio, e taxas de evasão escolar dos ensinos fundamental 2 e ensino médio apresentaram maior valor para os estados de Alagoas, Mato Grosso do Sul, Minas Gerais. A taxa de promoção explica bem os estados de AM, GO, MT, RS no que diz respeito a rede privada.

Figura 3.13: Apresentação gráfica dos grupos para as redes de ensino pública e privada



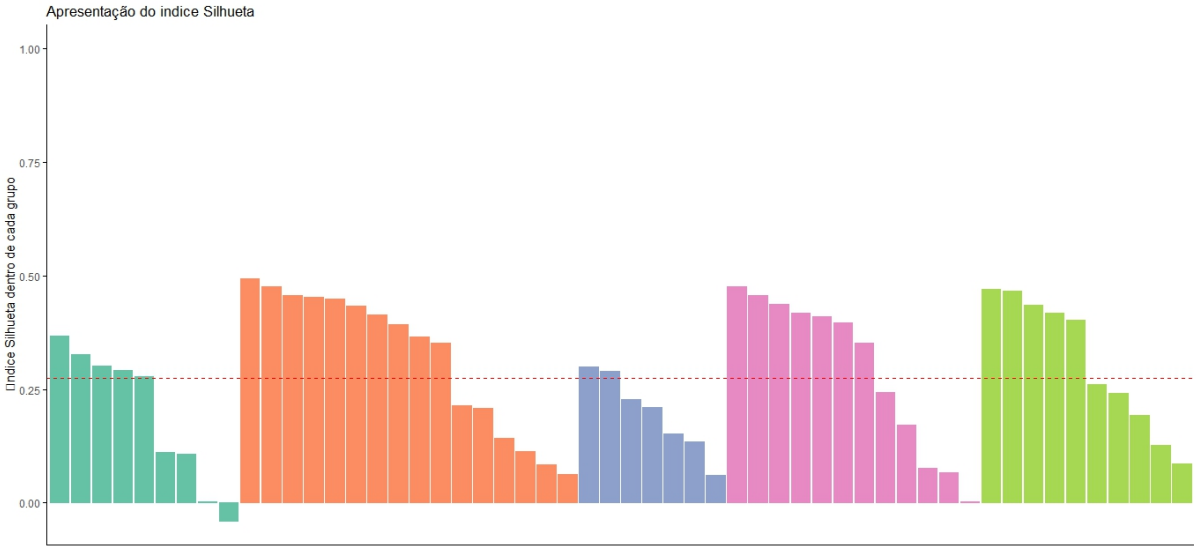
O Grafico da Figura 3.13 o agrupamento dos estados e os centros de cada grupo (centróides), que são vetores de médias das variáveis considerando os estados e rede de ensino, eles são representados pelas figuras geométricas identificadas ao lado do gráfico. Mostra, também, que existem estados que apesar de estarem em grupos diferentes possuem alguma similaridade, mas a similiridade não a suficiente para que estejam agrupados distintamente.

3.2.1 Validação dos grupos para os ensinos das redes pública e privada

Para verificação do ajuste do agrupamento, utilizou-se o índice de silhueta apresentado na seção 2.3.4. O gráfico de silhueta é usado para avaliar a qualidade dos agrupamentos produzidos pelo algoritmos $K-$means. Ele mede a semelhança os pontos de um grupo em relação a outro. Quanto mais alta for a pontuação da silhueta, melhor os dados estão agrupados.

A Figura 3.14 gerado observou-se que os percentuais dos grupos estão em torno de 25% a 50%. Isso significa que a distribuição da disposição dos estados dentro dos grupos é razoavel, uma vez que não apresentou índice de silhueta menor do que zero, exceto para o primeiro grupo, indicando que houve um elemento não deveria pertencer a esse grupo. Quanto ao índice de Dunn, apresentado na 2.3.4, obteve-se o valor de $0,3699$.

Figura 3.14: Índice de silhueta para os grupos: ensinos da rede pública e privada



De forma geral, se realizássemos um maior número de grupos poderíamos encontrar um maior valor para o índice de silhueta dos grupos. Contudo, após realizar a análise do conjunto de dados com números de grupos maiores notou-se que a contribuição era mínima para o índice.

3.3 Análise Multivariada para o ano de 2022

De acordo com o censo escolar 2022, INEP (2022) foram contabilizadas $47,4$ milhões de matrículas nas $178,3$ mil escolas de educação básica no Brasil, cerca de 714 mil matrículas a mais em comparação com o ano de 2021, o que corresponde a um aumento de $1,5\%$ no período.

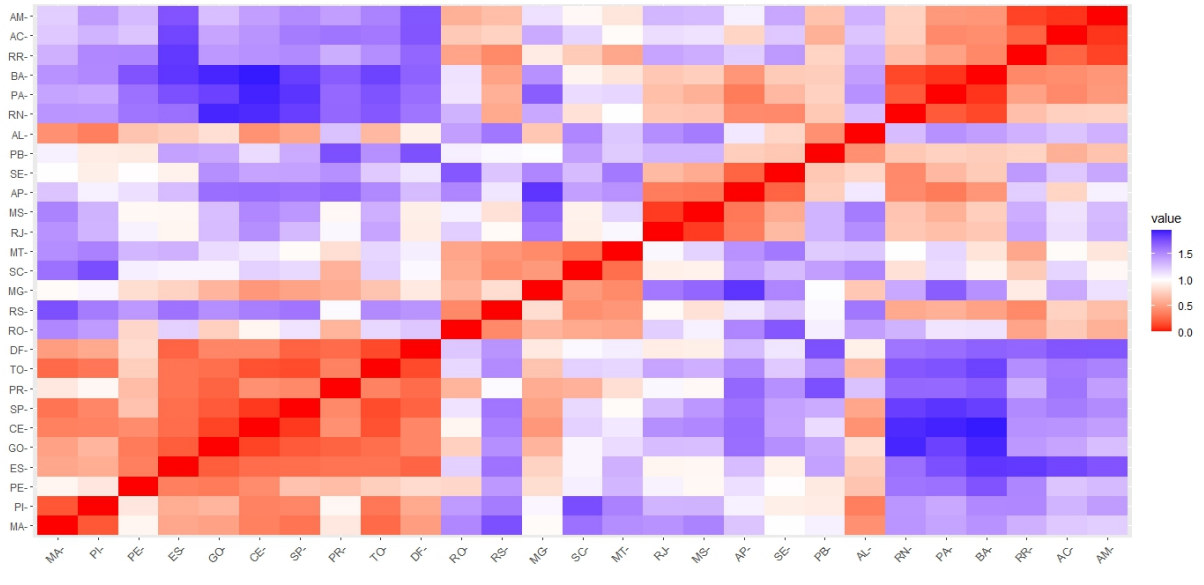
Entre os anos de 2021 e 2022 a rede privada de ensino expandiu $10,6\%$, chegando próximo ao nível observado em 2019, antes da pandemia.

Ainda nesse periodo, segundo o censo de 2022 quase a metade dos alunos matriculados são atendidos pelos municípios brasileiros $49,0\%$. Em 2022, a rede privada teve uma participação de $19,0\%$. Na educação básica, a União tem uma participação inferior a 1%.

Foram registradas $7,9$ milhões de matrículas no ensino médio em 2022, aumentando $1,2\%$ no último ano. De acordo com o censo esse crescimento estabelece uma tendência de aumento nas matrículas observada desde 2019 aumento de $5,4\%$.

Segundo a matriz de correlação, representada pela a Figura 3.15 observa-se que os estados Piauí e Maranhão apresentam alta similiaridade, indicado que para o grupo de variáveis estuda, os estado possuem oferta de ensino similar. E para os estados do Rio Grande do Norte, Pará e Bahia a figura mostra alta dissimilaridade com estado do Ceará, mesmo havendo semelhanças regionais geográficas entre alguns estados.

Figura 3.15: Matriz de similaridade dos estados - ano 2022

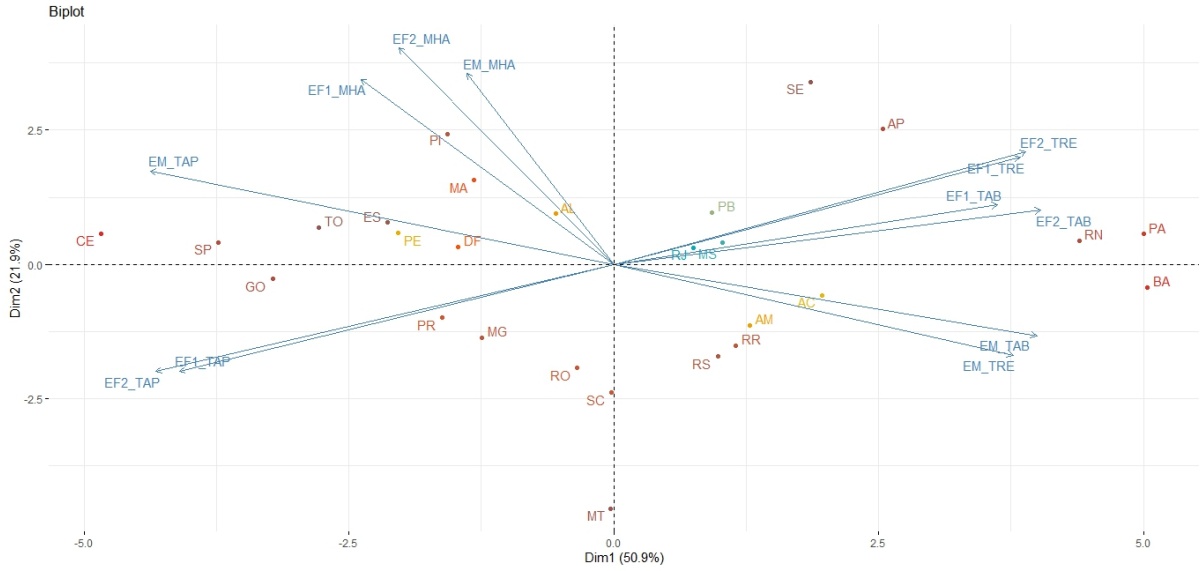


Para o agrupamento dos estados no ano de 2022 houve uma realocação dos estados do Rio Grande do Norte, Pará e Bahia em relação ao ano de 2018. De acordo o gráfico Biplot para os estados e as variáveis, Figura 3.16, nota-se que as variáveis que mais contribuíram para essa nova configuração foram as taxas de reprovação ($EF1\_TRE$, $EF2\_TRE$), abandono ($EF1\_TAB$, $EF2\_TAB$) para os ensinos fundamental 1 e 2.

Para nova realocação do estado do Piauí e Maranhão para o ano de 2022 as variáveis que obtiveram o maior valor para os referidos estados foram a média de horas aula nos ensinos fundamental e médio. Em 2018 observou-se que essas variáveis apresentaram maior valor para o estado de Pernambuco.

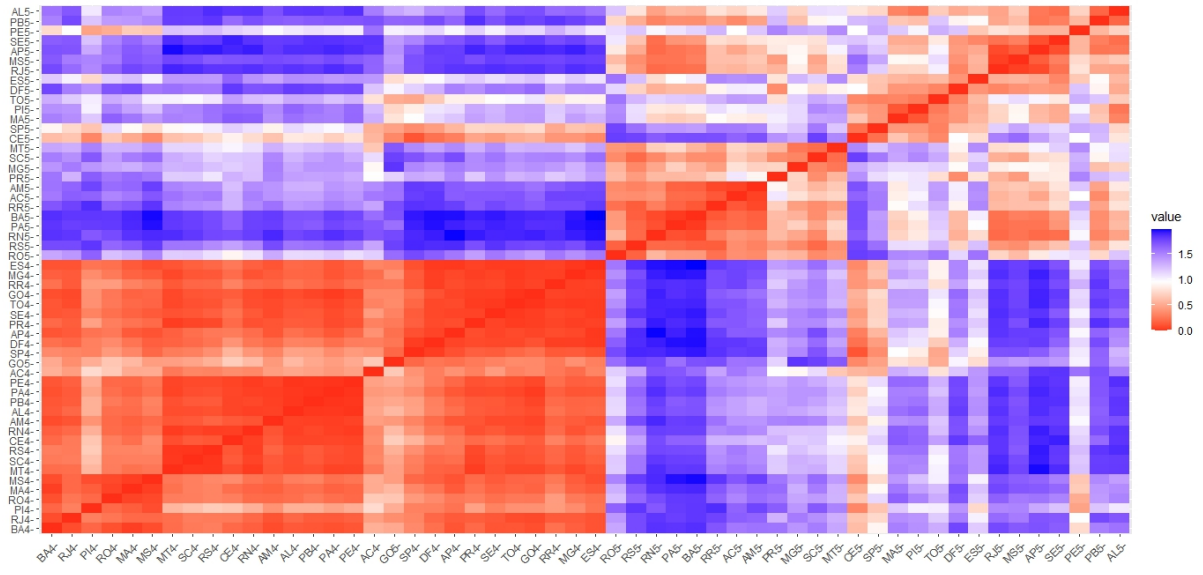
Podemos observar pela Figura 3.16 que a variação das variáveis em estudo são similares, diferentemente, do que ocorreu no ano de 2018 em que a variável taxa de evasão apresentou menor variação em relação a todas as variáveis em estudo para aquele ano.

Figura 3.16: Gráfico biplot para estados e as variáveis consideradas no estudo.



Concluimos a análise para os dados de 2022 com a construção da matriz de similaridades entre os estados e levando em consideração as redes pública e privada de ensino. Na Figura 3.17 observa-se aspectos diferentes do apontado para a matriz de similiaridade das redes pública e privada do ano de 2018, Figura 3.10.

Figura 3.17: Matriz de similaridades rede pública e privada - ano 2022



Em 2018 a matriz de similaridade apresentou quatro quadrantes de cores bem definido, separando a rede de ensino público e privado. E para rede pública as similaridades entres os estados altas, fato este que não ocorre em 2022.

Vale ressaltar que deve-se ter em mente que algumas variáveis estudadas em 2018 não são as mesmas observadas no estudo para 2022. Apesar disso, nota-se que o ensino público apresentou graficamente diferenças de similaridades para o ano de 2022.

Note que o padrão gráfico da Figura 3.17 é bastante diferente do padrão da matriz de similaridade para ano 2018, representada na Figura 3.10. E o ensino público do Estado do Ceará apresentou-se similariade ao ensino privado do Estado de São Paulo. A princípio não se sabe que mudanças ou políticas públicas contribuíram para que os estados de São Paulo para rede privada e Ceará, rede pública se mostrassem tão equiparados em termos de ensino para o ano de 2022.

Considerações Finais

A Base Nacional Comum Curricular (BNCC) no que se refere a proposta de tornar o ensino unificado e com equidade, foi uma das motivações para construção deste trabalho. A BNCC foi instituida em 2017 e com a análise de dados educacionais referentes aos anos de 2018 e 2022, evidenciou-se que as taxas escolares de evasão, abandono, repetência, aprovação não são características regionais geográficas.

O estudo apontou que a variável taxa de evasão para o ensino fundamental 1 apresentou menor variação com relação as demais variáveis para o ano de 2018 quando a análise estatística é realizada para os ensinos das redes pública e privada. Para o ano de 2022 a metodologia auxilou na identificação da alta variação das variáveis em estudo.

Analisamos como os índices escolares, evasão, abandono, taxas de promocão, aprovação e repetência se distribuem de forma conjunta nos estados e também as relação dessas variáveis com as redes de ensino publico e privado.

De maneira geral, para o ano de 2022 a metodologia mostrou que as variáveis taxas de reprovação e abandono para os ensinos fundamental 1 e 2 tiveram maiores valores para o agrupamento dos estados do Rio Grande do Norte, Pará, Bahia.

A proposta da metodologia pode ser aplicada para auxilar o gestor escolar na identificação de grupos de alunos que possuem desempenho acadêmico diferenciado, na perspectiva de diminuir a evasão na unidade escolar na qual o gestor tem a responsabilidade de realizar busca ativa no sentido de minimizar danos acadêmicos futuros.

De maneira similar ao apresentado nas análises dos ensinos de rede pública e privada, em houve dois grupos distinto quando apesentou-se dois grupos distintos, a saber, estados com oferta de ensino de rede pública e privado, o gestor poderá utilizar a metodologia para encontrar agrupamento de alunos que precisam de atendimento especial da unidade escolar.

Como trabalho futuro pretende-se desenvolver a metodologia para auxiliar na tomada de decisão enquanto gestor escolar e também divulgar o estudo para as demais unidades escolares do município de Timon no estado do Maranhão.

Referências Bibliográficas

ALEXANDER, K. L.; ENTWISLE, D. R.; HORSEY, C. S. From first grade forward: Early foundations of high school dropout. **Sociology of education**, JSTOR, p. 87–107, 1997.

BASSETTO, C. F. Background familiar e desempenho escolar: uma abordagem com variáveis binárias a partir dos resultados do saresp. **Revista Brasileira de Estudos de População**, SciELO Brasil, v. 36, 2019.

BRANCO, E. P. et al. Evasão escolar: desafios para permanência dos estudantes na educação básica. **Revista Contemporânea de Educação**, v. 15, n. 34, p. 133–155, 2020.

BRASIL. **Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação-PNE e dá outras providências**. 2014. Disponível em: <http://www.planalto.gov.br/ccivil03/ato2011-2014/2014/lei/l13005.htm>.

BROCK, G. et al. clvalid: An r package for cluster validation. **Journal of statistical Software**, v. 25, p. 1–22, 2008.

CASTRO, J. A.; FREIREB, R. S.; CASTRO, J. B. D. Tecnologia e aprendizagem de conceitos matemáticos. **Jornal Internacional de Estudos em Educação Matemática**, Universidade Bandeirante de São Paulo, v. 10, n. 2, p. 93–98, 2017.

CHARRAD, M. et al. Nbclust: an r package for determining the relevant number of clusters in a data set. **Journal of statistical software**, v. 61, p. 1–36, 2014.

COSTA, M. d. O.; SILVA, L. A. d. Educação e democracia: Base nacional comum curricular e novo ensino médio sob a ótica de entidades acadêmicas da área educacional. **Revista Brasileira de Educação**, SciELO Brasil, v. 24, 2019.

DEMIR, K.; KARABEYOGLU, Y. A. Factors associated with absenteeism in high schools. **Eurasian Journal of Educational Research**, v. 16, n. 62, 2015.

EDUCAÇÃO|BNCC, M. da. **Base Nacional Comum Curricular - Educação é a Base**. 2022. Disponível em: <http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf>.

FAVERO, E. et al. O primeiro ano do ensino fundamental de nove anos: uma revisão teórica. **Psicologia Escolar e Educacional**, SciELO Brasil, v. 21, p. 397–406, 2017.

FERNANDES, L. d. M. et al. Preditores do desempenho escolar ao final do ensino fundamental: histórico de reprovação, habilidades sociais e apoio social. **Trends in Psychology**, SciELO Brasil, v. 26, p. 215–228, 2018.

FERREIRA, L. G.; ABREU, R. M. de A. Características e desafios dos/nos anos iniciais do ensino fundamental: vozes de estagiários. **Revista de Estudos em Educação e Diversidade-REED**, v. 2, n. 5, p. 1–31, 2021.

FERRETTI, C. J. A reforma do ensino médio e sua questionável concepção de qualidade da educação. **Estudos avançados**, SciELO Brasil, v. 32, p. 25–42, 2018.

FIDELIS, J. M. et al. Relações entre raciocínio quantitativo e resolução de problemas matemáticos: um estudo sobre as estratégias de um grupo de estudantes de 3° e 4° anos do ensino fundamental. **Bolema: Boletim de Educação Matemática**, SciELO Brasil, v. 35, p. 1658–1677, 2022.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. **Educação por escrito**, v. 8, n. 1, p. 35–48, 2017.

FONSECA, G. R. As tic na formação inicial de professores–práticas de formação de formadores. **Da Investigação às Práticas: Estudos de Natureza Educacional**, v. 10, n. 2, p. 4–25, 2020.

GÓMEZ, Á. I. P. **Educação na era digital: a escola educativa**. [S.l.]: Penso Editora, 2015.

HARTIGAN, J. A.; WONG, M. A. et al. A k-means clustering algorithm. **Applied statistics**, USA, v. 28, n. 1, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; WALTHER, G. Estimating the number of data clusters via the gap statistic. **J Roy Stat Soc B**, v. 63, p. 411–423, 2001.

INEP. **Censo escolar 2022**. 2022. <https://download.inep.gov.br/censo_escolar/resultados/2022/apresentacao_coletiva.pdf>.

INEP, I. N. de Estudos e P. E. A. T. . **Censo Escolar 2018 revela crescimento de 18% nas matrículas em tempo integral no ensino médio**. 2019. <https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/censo-escolar-2018>.

INEP, I. N. de Estudos e P. E. A. T. . **Virtualizacao - VMWare e Xen**. 2022. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>.

JÚNIOR, F. T.; SANTOS, J. R. D.; MACIEL, M. de S. Análise da evasão no sistema educacional brasileiro. **Pesquisa E Debate Em Educação**, v. 6, n. 1, p. 73–92, 2016.

MACQUEEN, J. Classification and analysis of multivariate observations. In: UNIVERSITY OF CALIFORNIA LOS ANGELES LA USA. **5th Berkeley Symp. Math. Statist. Probability**. [S.l.], 1967. p. 281–297.

MEC. **Escola de Gestores da Educação Básica**. 2022. <https://portal.mec.gov.br/escola-de-gestores-da-educacao-basica>.

PIOLLI, E.; SALA, M. A reforma do ensino médio e a educação profissional: da lei de diretrizes e bases (ldb) às diretrizes curriculares nacionais para o ensino médio e para a educação profissional. **Revista Exitus**, Universidade Federal do Oeste do Pará, v. 11, 2021.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

ROSA, A. R.; FERNANDES, G. N. A.; LEMOS, S. M. A. Desempenho escolar e comportamentos sociais em adolescentes. **Audiology-Communication Research**, SciELO Brasil, v. 25, 2020.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.

SILVA, W. A. D.; COSTA, F. A. Reflexões teóricas sobre o lugar e o papel das tecnologias digitais na formação inicial de professores em portugal. **Pesquisa e Debate em Educação**, v. 12, n. 1, p. 1–e35328, 2022.

SPINILLO, A. G. et al. Formulação de problemas matemáticos de estrutura multiplicativa por professores do ensino fundamental. **Bolema: Boletim de Educação Matemática**, SciELO Brasil, v. 31, p. 928–946, 2017.

SUTPHEN, R. D.; FORD, J. P.; FLAHERTY, C. Truancy interventions: A review of the research literature. **Research on social work practice**, Sage Publications Sage CA: Los Angeles, CA, v. 20, n. 2, p. 161–171, 2010.

SYAKUR, M. et al. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In: IOP PUBLISHING. **IOP conference series: materials science and engineering**. [S.l.], 2018. v. 336, p. 012017.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. [S.l.]: Elsevier, 2006.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 2, p. 411–423, 2001.

Apêndice A

Rotinas computacionais: Análise descritiva

```
#Conjuntos de dados: 1. Taxas Escolares(2018/2019)
#                    2. Media de Horas aulas 2018
#
#Fonte:
# https://www.gov.br/inep/pt-br/acesso-a-informacao/
#         dados-abertos/microdados
#
#Observações: 1. Foram modificadas a estrutura dos
#                  conjuntos de dados
#             2. dados referentes aos Ensinos
#                 Fundamental I, II e Médio.
#
#Descrição dos conjuntos de dados:
#
#1. TXP: Taxa de promoção(o aluno em 2018 estava
#                         na série K e em 2019 foi
#   para a série seguinte)
#2. TXR: Taxa de repetência(o aluno em 2018 estava
#                          na série K e em 2019
#   permaneceu na mesma série)
#3. TXE: Taxa de evasão(o aluno em 2018 estava na
#                       série K e em 2019 não
# encontra-se matriculado)
#4. MHA: Média de horas aula no ano de 2018.
#
```

```
#5. EF1: Anos iniciais do ensino Fundamental
#
#6. EF2: Anos finais do ensino Fundamental
#
#7. EM: Ensino médio
#
#8. UF: Unidade Federativa
#
#9. Local: Tipo de localidade - Total/Urbana/Rural
#
#10. DA: Dependência Administrativa - Total/Publica/Privada
#
##################################


#   *************
# Pacotes necessários
#   ***********


require(tidyverse)

require(descr)

require(cluster)

require(FactoMineR)

require(fpc)

require(factoextra)

            ##   **************
            # Realizando a Leitura dos conjuntos de dados
            #  Ano analisado: 2018
            #   ******************

setwd("C:...")

dados1 <- read.table("TaxasEscolares.txt", h = TRUE,
```

```
                    dec = ".")

dados2 <- read.table("MediaHorasAulas2018.txt", h = TRUE,
                    dec = ".")

## Agrupando os dados1 e dados2

dados <- dados1 %>% inner_join(dados2)

## Análise quanto a localidade das escolas: Urbana e Rural ###

Escola.local <- subset(dados.estado,
                       dados.estado$Local != "Total")

Escola.local$Local <- as.factor(Escola.local$Local)


##Figura 3.1:
#Boxplot das taxas de evasão escolar dos ensinos
#fundamental e médio

par(mfrow=c(1,3))

plot(EF1_TXE ~ Local, Escola.local, xlab =
      "Tipo de Localidade da Escola",
ylab = "Taxa de Evasão Escolar Ensino Fundamental 1")

plot(EF2_TXE ~ Local, Escola.local, xlab =
       "Tipo de Localidade da Escola",
ylab = "Taxa de Evasão Escolar Ensino Fundamental 2", col = 2)

plot(EM_TXE ~ Local, Escola.local, xlab =
       "Tipo de Localidade da Escola",
ylab = "Taxa de Evasão Escolar Ensino Médio", col = 3)


    # ********************************
    ## Analisando a taxa de Evasão Escolar nos Estados
    #************************
```

```
dados.estado$UF <- as.factor(dados.estado$UF)

levels(dados.estado$UF) <- c("AC", "AL", "AP", "AM", "BA", "CE",
                             "DF", "ES", "GO", "MA", "MT", "MS",
                             "MG", "PA", "PB", "PR", "PE", "PI",
                             "RJ", "RN", "RS", "RO", "RR", "SC",
                             "SP", "SE", "TO")

##Figura 3.2:
# Boxplot das taxas de evasão escolar nos estados
# por níveis de ensino


par(mfrow=c(1,3))

plot(EF1_TXE ~ UF, dados.estado, xlab = " ",
ylab = "Taxa de Evasão Escolar Ensino Fundamental 1",
cex.lab = 1.4, cex = 1.2, cex.sub = 1.5)

plot(EF2_TXE ~ UF, dados.estado, xlab =
     "Estados Brasileiros e Distrito Federal",
ylab = "Taxa de Evasão Escolar Ensino Fundamental 2", col = 2,
cex.lab = 1.4, cex = 1.2, cex.sub = 1.5)

plot(EM_TXE ~ UF, dados.estado, xlab = " ",
ylab = "Taxa de Evasão Escolar Ensino Médio", col = 3,
cex.lab = 1.4, cex = 1.2, cex.sub = 1.5)


##  *****************
## Apresentação gráfica de um histograma para a média de
# de horas aulas para o ensino público e privado
##  ****************


Escola.ensino <- subset(dados.estado,
                        dados.estado$DA != "Total")

Publica <- subset(Escola.ensino,
```

```
                  Escola.ensino$DA == "Publica" )

Privada <- subset(Escola.ensino,
                  Escola.ensino$DA == "Privada" )

## Para o Ensino Público

Pb.EF1 <- Publica$EF1_MHA  #Esc. Pública Ensino Fund. 1

Pb.EF2 <- Publica$EF2_MHA  #Esc. Pública Ensino Fund. 2

Pb.EM <- Publica$EM_MHA    #Esc. Pública Ensino Médio


## Para o Ensino Privado


Pr.EF1 <- Privada$EF1_MHA  #Esc. privada Ensino Fund. 1

Pr.EF2 <- Privada$EF2_MHA  #Esc. privada Ensino Fund. 2

Pr.EM <- Privada$EM_MHA    #Esc. privada Ensino Médio


## Parte Gráfica para os tipos de ensinos

privada <- c(Pr.EF1, Pr.EF2, Pr.EM)

publica <- c(Pb.EF1, Pb.EF2, Pb.EM)

Media.horas <- c(privada, publica)

Ensino <- rep(c("Privado", "Público"), each = 81)

data <- data.frame(Media.horas, Ensino)


## Figura 3.3:
## Distribuição de frequência de média de horas aula
```

```
ggplot(data, aes(x = Media.horas, fill = Ensino)) +
geom_histogram(binwidth = 0.1, alpha = 0.4, bins = 50)+
labs(x = "Média de horas aulas",y = "Frequência")
```

Apêndice B

Rotinas computacionais: Metodologia $K-$means

```
dados.estado <- subset(dados, dados$UF != "Brasil" &
                           dados$UF != "Norte" &
                           dados$UF != "Nordeste" &
                           dados$UF != "CentroOeste" &
                           dados$UF != "Sul" &
                           dados$UF != "Sudeste")

Dados <- read.table("NovosDados2018.txt", h = TRUE,
                    dec = ".", row.names = 1)
data <- scale(Dados)

DadosEstado <- read.table("DadosEstado2018.txt", h = TRUE,
                    dec = ".", row.names = 1)
data2 <- scale(DadosEstado)

dist.cor2 <-  get_dist(data2, method = "pearson")


    #Figura 3.4:
#Gráfico da matriz de similaridade dos
    #dados educacionais


fviz_dist(dist.cor2) +
theme(legend.title = element_blank())
```

```
# Método Elbow
#Figura 3.5:
#Escolha do número de grupos - Método Elbow

fviz_nbclust(data2, kmeans, method = "wss") +
geom_vline(xintercept = " ", linetype = 2)+
labs(title = "Número ótimo de grupos",
 x = "Número de grupos K",y =
     "Soma de Quadrado dentro do grupo")

km.res <- kmeans(data2, centers = 5,
          iter.max = 20, nstart = 5 )

data <- DadosEstado

res.pca <- PCA(data, graph = FALSE)


#Figura 3.6:
#Gráfico do percentual de variabilidade explicada

fviz_screeplot(res.pca, addlabels = TRUE)+
labs(x = "Dimensões", y =
     "Porcentagem da variância explicada")


#Figura 3.7:
# Disposição do agrupamento dos estados

fviz_cluster(km.res, data = DadosEstado,
             palette = c("#8D5B55", "#00AFBB", "#E7B800",
                         "#FC4E07", "#66CF00"),
             ellipse.type = "euclid",
             star.plot = TRUE, repel = TRUE,
             ggtheme = theme_minimal()
)+
labs(title = "Apresentação em cores dos cinco grupos")+
theme(legend.title = element_blank())
```

```
#Figura 3.8:
# Apresentação individual dos estados

fviz_pca_ind(res.pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB",
            "#E7B800", "#FC4E07"),
             repel = TRUE)+
labs(title = "Apresentação Individual
    dos Estados para cinco grupos")+
theme(legend.position = "none")

#Figura 3.9:
# Gráfico Biplot para o estudo das relações
# entre os estados e variáveis

fviz_pca_biplot(res.pca, repel = TRUE)+
labs(title = "Gráfico Biplot")
```

B.1 Análise para os ensinos das redes pública e privada

```
#Ajustando o conjunto de dados:

DA <- read.table("DadosDA2018.txt", h = TRUE,
                      dec = ".", row.names = 1)

data <- scale(DA)   #realiza mudança de escala

dist.cor <-  get_dist(data, method = "pearson")


#Figura 3.10:
# Matriz de similaridade para rede pública e privada

fviz_dist(dist.cor) + theme(legend.title = element_blank())
```

```
# Figura 3.11:
# Apresentação dos estados nos grupos para os ensinos
# públicos e privado

km.res <- kmeans(data, centers = 5, iter.max = 20,
          nstart = 5 )

fviz_cluster(km.res, data = DA,
             palette = c("#2E9FDF", "#339933",
                         "#E7B800", "#FF0066","#DE2D26"),
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)+ labs(title = "Apresentação dos Grupos") +
theme(legend.position = "none")

data <- DA

res.pca <- PCA(data, graph = FALSE)


#Figura3.12
# Gráfico Biplot para os ensinos das redes
# pública e privada

fviz_pca_biplot(res.pca, repel = TRUE)+
labs(title = "Biplot")

DA <- read.table("DadosDA2018.txt", h = TRUE,
                     dec = ".", row.names = 1)

data <- scale(DA)

km.res <- eclust(data, "kmeans", k = 5, nstart = 60,
                 graph = FALSE)

# Figura 3.13:
# Apresentação gráfica dos grupos para as redes
# de ensino pública e privada
```

```
fviz_cluster(km.res, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())+
labs(title = "Compacidade em cores dos cinco grupos")+
theme(legend.title = element_blank())


# Figura 3.14:
# Apresentação do Índice de silhueta para os ensinos
# da rede pública e privada


fviz_silhouette(km.res, palette = "Set2",
                ggtheme = theme_classic())+
labs(title = "Apresentação do índice Silhueta",
y = "Índice Silhueta dentro de cada grupo")+
theme(legend.position = "none")
```