Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Um estudo do número de inscrições e dos inscritos no ENEM no período de 2013 a 2022

Gabriele de Oliveira Arantes

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Um estudo do número de inscrições e dos inscritos no ENEM no período de 2013 a 2022

Gabriele de Oliveira Arantes
Orientador: Pedro Ferreira Filho

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Janeiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

A study of the number of registrations and participants in the ENEM from 2013 to 2022

Gabriele de Oliveira Arantes
Advisor: Pedro Ferreira Filho

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
January 2024

Gabriele de Oliveira Arantes

Um estudo do número de inscrições e dos inscritos no ENEM no período de 2013 a 2022

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Gabriele de Oliveira Arantes e aprovado pela banca examinadora.

Aprovado em 16 de janeiro de 2024

Banca Examinadora:

- Prof. Pedro Ferreira Filho
- Prof. Dr. Francisco Antonio Rojas Rojas
- Ms. Gretta Rossi Ferreira

Aos meus familiares, especialmente às mulheres de minha família, e amigos por acreditarem em mim e celebrarem minhas conquistas.

Agradecimentos

Ao meu orientador, Professor Pedro Ferreira Filho, expresso minha mais profunda gratidão pela inestimável orientação, apoio e auxílio fornecidos ao longo da execução deste trabalho.

Aos membros da banca examinadora, Ms. Gretta Rossi Ferreira e Professor Dr. Francisco Antonio Rojas Rojas, agradeço pelas valiosas sugestões, críticas construtivas e correções precisas que aprimoraram significativamente a qualidade deste trabalho.

À minha irmã Haryanna e à minha mãe Gerciana, dedico este trabalho com imensa gratidão pelo amor incondicional, incentivo constante e apoio inabalável em todos os momentos. Vocês são minhas maiores fontes de inspiração e força, e sem vocês, este sonho não teria se tornado realidade.

A todos que, direta ou indiretamente, me apoiaram durante o desenvolvimento deste trabalho, agradeço pelo incentivo e colaboração.

"Educar é semear a esperança."

(Paulo Freire)

Resumo

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é uma autarquia federal vinculada ao Ministério da Educação (MEC) e tem como objetivo promover ações educacionais no Brasil. Ele desempenha um papel fundamental na criação de medidas relacionadas à educação que são essenciais para o progresso social e econômico do país.

Desde o ano de 1998 o INEP passou a organizar e executar todos os procedimentos do então criado Exame Nacional do Ensino Médio (ENEM), destinado a avaliar alunos egressos do ensino médio. A partir de 2009 o ENEM é objeto de um amplo processos de reformulação e passa a ser utilizado como instrumento para seleção de alunos de instituições de ensino superior pública e privada tornando-se um dos principais instrumentos para a democratização do acesso à educação superior. O novo ENEM começou num processo de crescimento do número de inscritos, porém os últimos anos registram uma grande redução da quantidade de inscritos. Observa-se que poucos trabalhos têm sido realizados na perspectiva de identificar os fatores responsáveis por essa evasão no número de participantes no exame.

Nesse sentido, este trabalho teve como objetivo realizar um estudo visando identificar os possíveis fatores que têm contribuído para a redução do número de participantes do ENEM. Resultados obtidos indicam que idade, tempo de conclusão do ensino médio e nível de escolaridade dos pais foram identificados como fatores que contribuem para a redução do número de inscritos. Com apoio de diferentes métodos estatísticos foi também analisado o aproveitamento dos participantes nas provas. Resultados alcançados indicam um padrão de aproveitamento das provas no período de 2013 a 2016 e, outro padrão com algumas diferenças nos anos de 2017 a 2022. Além disso, análises mostram que existe associação entre grupos de participantes com características específicas e seu desempenho médio alcançado nas provas ao longo dos anos, em particular daqueles em que foi observada uma redução da participação no ENEM.

Palavras-chave: *ENEM, evasão, aproveitamento de provas.*

Abstract

The Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) is a federal agency linked to the Ministério da Educação (MEC) and aims to promote educational actions in Brazil. It plays a fundamental role in creating education-related measures that are essential for the country's social and economic progress.

Since 1998, INEP has been responsible for organizing and executing all procedures of the then-created Exame Nacional do Ensino Médio (ENEM), designed to assess high school graduates. From 2009, the ENEM underwent a comprehensive reformulation process and began to be used as a tool for selecting students for public and private higher education institutions, becoming one of the main instruments for democratizing access to higher education. The new ENEM initially experienced growth in the number of registrants, but recent years have seen a significant reduction in the number of participants. It is observed that few studies have been carried out to identify the factors responsible for this decline in the number of exam participants.

Therefore, this study aimed to identify the possible factors contributing to the reduction in the number of ENEM participants. The results obtained indicate that age, time of high school completion, and parents' level of education were identified as factors contributing to the decrease in the number of registrants. With the support of different statistical methods, the performance of participants in the tests was also analyzed. The achieved results indicate a pattern of test performance from 2013 to 2016 and another pattern with some differences in the years from 2017 to 2022. Additionally, analyses show an association between groups of participants with specific characteristics and their average performance in the tests over the years, particularly among those where a reduction in ENEM participation was observed.

Keywords: *ENEM, non-participation, test performance.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é uma autarquia federal vinculada ao Ministério da Educação (MEC) que desempenha ações educacionais que acontecem no Brasil, criando medidas relacionadas à educação essenciais para o desenvolvimento social e econômico do país. O instituto tem como finalidade incentivar a realização de pesquisas e avaliações periódicas acerca do sistema educacional no Brasil, visando fornecer informações que auxiliam na criação e execução de políticas públicas na área da educação.

Conforme (Saviani, 2012) o Instituto Nacional de Estudos Pedagógicos como era nomeado quando fundado, foi criado em julho de 1938, mas só passou a ser autarquia federal após a aprovação da (Medida Provisória, 1997) em março de 1997 redefinindo assim suas finalidades, ao invés de ser o responsável pela organização de documentação e disseminação de informações educacionais desempenhando um papel como órgão executor e promotor de pesquisa, funções a ele atribuídas desde sua fundação até a promulgação dessa lei, se tornou um órgão responsável pela avaliação da educação brasileira em todos os níveis e modalidades. Sua principal característica passou a ser subsidiar a formulação de políticas educativas, ou seja, o instituto deixou de ser um órgão de pesquisa para se tornar uma agência de avaliação.

Em 1998 foi instituído pelo (Inep, 2023) o Exame Nacional do Ensino Médio (ENEM), com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2009, o ENEM passou por um grande processo de mudança. O exame teve toda a sua metodologia reformulada incluindo novos procedimentos em particular a Teoria de Resposta ao Item (Andrade *et al.*, 2000). A partir deste novo modelo o ENEM se transforma de um procedimento de avaliação do ensino médio, ao principal

mecanismo para seleção de alunos na maior parte das instituições de ensino superior públicas e privadas dos país.

Como mencionado por Neto *et al.* (2014) o ENEM é um dos principais instrumentos para a democratização do acesso à educação superior, o exame possui a competência para diminuir as desigualdades de acesso ao ensino superior, possibilitando o ingresso em diferentes programas sociais e em diferentes cursos e universidades de todo o país.

Ao mesmo tempo, em 2012 foi aprovada a (Lei de Cotas, 2012) como parte de um programa mais amplo de políticas de ações afirmativas (PAA). De acordo com (Gaspar e Barbosa, 2013) as políticas de ações afirmativas visam à garantia de direitos historicamente negados a grupos minoritários, garantindo a obrigatoriedade da reserva de 50% das vagas nas instituições federais de ensino para estudantes oriundos de escolas públicas, com renda per capita inferior a um salário mínimo e meio, pessoas com deficiência desde a (Alteração da Lei de Cotas, 2016), e autodeclarados pretos, pardos ou indígenas.

Apesar da sua amplitude e importância, observa-se ao longo dos anos uma diminuição do número de inscritos no ENEM. De acordo com (Estado de Minas, 2022) a última edição do exame, sucedida no ano de 2022 teve 3,4 milhões de inscritos, representando um dos menores números em mais de uma década, com uma taxa de abstenção de 32,4%, uma das maiores da história. Ainda vale observar que em 2020, o primeiro ano da pandemia, obteve-se uma taxa de de abstenção equivalente a 55,5% dos inscritos.

Vários motivos podem ser apontados como aqueles que contribuíram para esta redução no número de inscritos. Podem ser mencionados, a pandemia acompanhado do ensino remoto o que diminuiu a expectativa dos estudantes de prestar o exame, a crise econômica e o corte de verbas de bolsas de auxílio estudantil do Programa Nacional de Assistência Estudantil (Pnaes), além da total falta de políticas públicas de incentivo a participação no ENEM. Como consequência mais visível da quantidade reduzida de inscritos, temos um significativo aumento no número de vagas ociosas, não preenchidas, nos cursos de graduação das universidades e institutos federais e também nas instituições privadas de ensino superior.

Ao longo deste período, o Inep divulga anualmente, desde a sua reformulação em 1998, os dados referentes ao ENEM. Os microdados (Microdados, 2023) do ENEM contêm informação detalhada sobre todas as aplicações do exame, atendendo à necessidade de informações específicas ao disponibilizar características econômicas e sociais dos candidatos, provas, gabaritos, notas e fluxo dos participantes durante as datas das avaliações.

Este estudo teve como interesse analisar as informações presentes nos microdados do ENEM durante o período dos últimos 10 anos, dando início no ano seguinte em que a política de cotas foi promulgada (2013) até o último ano da aplicação do exame (2022). Tal período foi marcado por diversas mudanças no contexto socioeconômico e educacional do país, como políticas públicas na área da educação, crises econômicas, pandemia de COVID-19, avanços tecnológicos, entre outros fatores. Esses elementos podem ter influenciado diretamente a participação e o desempenho dos candidatos no ENEM, sendo importante analisar como esses aspectos impactaram a evasão e os resultados das provas do exame ao longo do tempo.

Dessa forma, nesta década buscou-se analisar o comportamento dos participantes a cada ano e se essas características sofreram alteração ao decorrer do tempo, na perspectiva de identificar os fatores que contribuíram na evasão do número de inscritos bem como o desempenho dos candidatos nas áreas de conhecimento do exame.

Este trabalho está organizado em 4 capítulos, de forma que no Capítulo 2 são apresentados os materiais e métodos a serem utilizados, no Capítulo 3 os resultados das diferentes análises realizadas e, por fim, no Capítulo 4 são apresentadas as conclusões deste estudo.

Além dos autores já citados, serão mencionados no decorrer do estudo: (Johnson, 2008), (Inep, 2023), (Balanço, 2022), (Mingoti, 2005), (Fávero *et al.*, 2009), (Morettin e Singer, 2022), (Ferreira Filho *et al.*, 1998), (Decicino, 2012), (MEC, 2023b), (MEC, 2023a), (G1, 2023) e (dos Reis Silva *et al.*, 2019).

Capítulo 2

Material e métodos

2.1 Material

Desde 1998, e ainda mais a partir de 2009, o INEP divulga anualmente os dados do Exame Nacional do Ensino Médio (ENEM), um importante meio para promover e contribuir na democratização do acesso ao ensino superior. A implementação de políticas de ações afirmativas, como a Lei de Cotas, produziu num primeiro momento um aumento de inscrições no ENEM. Porém, a partir de 2017 tem sido observada uma significativa redução no número de inscritos no exame. No entanto, observa-se que poucos estudos tem sido realizados com o objetivo de identificar e avaliar os fatores que têm contribuído para essa diminuição no número de participantes no exame.

Nesta perspectiva, este estudo teve como propósito analisar as informações contidas nos microdados do ENEM ao longo dos últimos 10 anos, começando no ano seguinte à promulgação da política de cotas (2013) até o último ano de aplicação do exame (2022). Durante essa década, foi investigado o comportamento dos participantes a cada ano a fim de verificar se essas características sofreram alterações ao longo do tempo. Além disso, foi procurado identificar os fatores que contribuíram para a diminuição no número de inscritos e analisar o desempenho dos participantes nas diferentes áreas de conhecimento do exame.

A base de dados utilizada neste trabalho está disponível em (Inep, 2023). Estão à disposição as informações de todos os candidatos inscritos no Enem desde a suas primeiras realizações. Neste trabalho foram considerados os dados a partir de 2013 dado que foi no ano de 2012 que foi estabelecido o programa de ações afirmativas nas instituições publicas federais de ensino superior. Estas bases de dados contém informações pessoais

dos inscritos bem como os resultados obtidos nas diferentes provas do ENEM.

Num primeiro momento foram identificadas nas bases as informações presentes em todos os anos em estudo de forma a ser possível avaliar as características dos inscritos ao longo da década fixada visando identificar possíveis fatores que possam ter contribuído para a diminuição do número de inscritos, em particular nos últimos anos.

Foram consideradas também as notas obtidas pelos inscritos dos respectivos anos presentes no estudo. Esta análise de desempenho dos participantes nas provas foi realizada considerando aqueles candidatos inscritos que estiveram presentes nos dois dias que foram aplicadas as avaliações. Desta forma procura-se avaliar o resultado nestas provas e verificar a existência ou não, de um padrão de aproveitamento ao longo dos anos. Finalmente é interesse relacionar o aproveitamento nas provas, aos possíveis fatores considerados como responsáveis pela redução de inscritos.

Inicialmente, foram selecionadas as variáveis de interesse para o estudo, que foram aquelas consideradas com maior potencial para influenciar no número de inscritos, incluindo também, aquelas que têm como finalidade avaliar o comportamento do aproveitamento nas provas do ENEM.

Foi identificado ao longo dos anos destacados para a análise que algumas variáveis mudaram a forma de sua categorização. Dado isso, foi realizada uma padronização das categorias de todas as variáveis selecionadas, ou seja, as mesmas foram categorizadas do mesmo modo em todos os anos avaliados, para que assim seus resultados pudessem ser comparados ao longo do período do estudo. O critério de categorização utilizado foi o mesmo utilizado pelo INEP em seus relatórios (Balanço, 2022) dos resultados do ENEM.

Ainda, para fins de análise e comparação dos resultados ao longo dos anos estudados, foram criadas 3 variáveis em relação a presença do participante nas provas, indicando presença no primeiro dia, no segundo dia e em ambos os dias das provas. Foram classificados como ausentes aqueles que tenham faltado pelo menos em um dos dias da realização das provas, ou que tenham sido eliminados em função dos critérios de desempenho mínimo, e considerados presentes caso contrário. Foi elaborada também uma variável que diz respeito ao tempo de formação do indivíduo, que descreve há quantos anos o sujeito concluiu o ensino médio.

Sendo assim, as variáveis selecionadas para a análise e suas respectivas categorizações encontram-se na Tabela A.1 no Apêndice A.

2.2 Métodos

Inicialmente o trabalho teve por objetivo investigar o comportamento de um conjunto de variáveis de interesse, selecionadas na base de dados do INEP, ao longo dos anos fixados para o estudo. Desta forma, no caso destas variáveis, o estudo consistiu em uma análise bivariada (Morettin e Singer, 2022) considerando o ano e cada uma das variáveis presentes no estudo. Cada análise foi baseada em uma tabela de contingência com uma respectiva representação gráfica, conforme apresentado por (Decicino, 2012). No caso do estudo do aproveitamento nas provas do ENEM, a comparação dos diferentes anos foi realizada com base em medidas resumo de posição e dispersão, média e desvio padrão, respectivamente, com representação gráfica a partir do uso do boxplot (Morettin e Singer, 2022).

Em um segundo momento foi realizada uma análise conjunta das variáveis presentes no estudo. Com o objetivo de identificar o efeito conjunto destas variáveis na redução no número de inscritos no ENEM e, por se tratarem de variáveis categóricas, uma análise fatorial de correspondências múltiplas (AFCM) (Mingoti, 2005) foi realizada para cada um dos anos com o objetivo de verificar a existência ou não de um mesmo padrão de comportamento conjunto ao longo dos anos. Desta forma foi possível identificar um possível subconjunto de variáveis que tiveram a sua participação reduzida ao longo do tempo.

Finalmente, uma análise de componentes principais (ACP) (Johnson, 2008) foi realizada para avaliar o comportamento conjunto das notas das diferentes provas. Tal como no caso da AFCM, foi realizada uma ACP para cada ano na busca de identificar um padrão de relacionamento conjunto das notas ao longo dos anos. Identificado este padrão foi estabelecido uma relação entre o os resultados da AFCM e da ACP, ou seja se aquele padrão de aluno que deixou de participar do ENEM, apresentou menor desempenho nas provas.

2.3 Análise Fatorial de Correspondências Múltiplas

Segundo (Johnson, 2008), a Análise de Correspondência (AFC) é uma técnica multivariada, que tem como objetivo descrever dados qualitativos, analisando informações contidas em tabelas de contingência e representando graficamente a estrutura das mesmas. Quando se trata de uma única tabela de contingência (duas variáveis em estudo), utiliza-se a Análise de Correspondência Simples (AFCS). Nos casos que envolvem múltiplas tabe-

las, ou análise conjunta de três ou mais variáveis, a Análise Fatorial de Correspondência Múltipla (AFCM) é utilizada.

Para o estudo da associação entre duas variáveis (X e Y), os dados são organizados em uma tabela de contingência $I \times J$, em que I representa o número de linhas e, J o número de colunas, como representado na Tabela 2.1, ainda, n_{ij} representa o número de elementos pertencentes à categoria i da variável X e à categoria j da variável Y (Mingoti, 2005).

Tabela 2.1: Representação da tabela de contingência.

	Variável Y				
Variável X	1	2	$\cdots$	J	Total
1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	$n_{..}$=N

A Análise Fatorial de Correspondência Simples (AFCS) é uma técnica estatística multivariada que permite estudar a associação entre duas variáveis categóricas. Para isso, a AFCS calcula a distância entre as frequências observadas e esperadas das categorias das variáveis em uma tabela de contingência. Quanto menor a distância, maior a associação entre as variáveis.

Para verificar se a associação é estatisticamente significativa, a AFCS utiliza o teste qui-quadrado. A estatística χ^2 é dada pela razão entre os resíduos (diferença entre as frequências observadas e esperadas da tabela de contingência) ao quadrado e a frequência esperada. Se o valor da estatística de teste qui-quadrado for maior que um valor crítico, rejeita-se a hipótese de independência entre as variáveis.

Na análise de correspondência, seu desenvolvimento algébrico começa com a construção de uma matriz de proporções, também conhecida como matriz de correspondência, que é representada por $\mathbf{P}=\{p_{ij}\}$. Os elementos da matriz de proporções são dados pela razão entre as frequências observadas e o total de observações, ou seja, $p_{ij} = \frac{n_{ij}}{N}$.

A seguir, são definidas as matrizes de perfil linha ($\mathbf{D}_r$) e coluna ($\mathbf{D}_c$). Essas matrizes são diagonais, ou seja, possuem apenas elementos não-nulos na diagonal principal. Os elementos das matrizes de perfil são dados pelas frequências marginais das variáveis, ou seja, $\mathbf{r}' = \left(\frac{n_{1.}}{N}, \frac{n_{2.}}{N}, \cdots, \frac{n_{I.}}{N}\right)$ e $\mathbf{c}' = \left(\frac{n_{.1}}{N}, \frac{n_{.2}}{N}, \cdots, \frac{n_{.J}}{N}\right)$.

Com isso, a matriz $\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}'$, pode ser decomposta em autovalores e autovetores,

isto é, em valores singulares (Mingoti, 2005), conforme Expressão (2.1).

$$\tilde{\mathbf{P}}_{I\times J} = \mathbf{A}\mathbf{\Lambda}\mathbf{B}', \tag{2.1}$$

Onde $\mathbf{A} = \mathbf{D}_r^{1/2}\mathbf{U}_{I\times k}$ e $\mathbf{B} = \mathbf{D}_c^{1/2}\mathbf{V}_{J\times k}$. $\mathbf{A}$ e $\mathbf{B}$ são matrizes de transformação que mapeiam as linhas e colunas da matriz de proporções $\mathbf{P}$ para uma nova base. As matrizes $\mathbf{U}$ e $\mathbf{V}$ são matrizes ortogonais, ou seja, os seus autovetores são ortogonais entre si, contendo os autovetores da matriz $\tilde{\mathbf{P}}\tilde{\mathbf{P}}'$ e $\tilde{\mathbf{P}}'\tilde{\mathbf{P}}$, respectivamente. A matriz $\mathbf{\Lambda}$ é uma matriz diagonal com dimensão $k \times k$, cujos elementos são os autovalores da matriz $\mathbf{P}$ ordenados de forma decrescente.

As coordenadas principais das linhas e colunas da matriz $\tilde{\mathbf{P}}$ são descritas de acordo com as Equações (2.2) e (2.3) a seguir:

$$\mathbf{Y}_{I\times k} = \mathbf{D}_r^{-1}\mathbf{A}_{I\times k}\mathbf{\Lambda}_{k\times k}\,, \tag{2.2}$$

$$\mathbf{Z}_{J\times k} = \mathbf{D}_c^{-1}\mathbf{B}_{J\times k}\mathbf{\Lambda}_{k\times k}\,. \tag{2.3}$$

Portanto, a matriz $\tilde{\mathbf{P}}$ pode ser expressa em função dos autovalores e das coordenadas principais, como segue a Equação (2.4).

$$\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}' = \sum_{i=1}^{k} \hat{\lambda}_i\, \tilde{a_i}\, \tilde{b_i}', \tag{2.4}$$

Sendo $\tilde{a_i}$ e $\tilde{b_i}$ a i-ésima coluna da matriz $\mathbf{A}$ e $\mathbf{B}$, respectivamente, e k= posto$(\tilde{P})$ = $\min(I-1, J-1)$.

As coordenadas principais obtidas podem ser representadas em um gráfico chamado mapa perceptual, que permite avaliar a associação entre as linhas e colunas da tabela. Esse gráfico identifica possíveis relações entre as categorias das variáveis. Para identificar como determinada linha ou coluna contribui para a construção de cada eixo do mapa perceptual, é necessário utilizar os componentes da variação total existentes no sistema, chamada de inércia total. A inércia total é expressa na Equação (2.5).

$$\sum_{i=1}^{k} \lambda_i^2, \tag{2.5}$$

Os autovalores não-nulos da diagonal da matriz $\mathbf{\Lambda}$ são representados por λ_i, para $i = 1, 2, ..., k$. Ainda, a Equação (2.5) pode ser expressa em termos da estatística qui-quadrado, da seguinte forma: $\sum_{i=1}^{k} \lambda_i^2 = \frac{\chi^2}{N}$.

A proporção da contribuição da i-ésima coordenada principal para a inércia total é apresentada na Equação (2.6).

$$\frac{\lambda_i^2}{\sum_{i=1}^{k} \lambda_i^2} \tag{2.6}$$

Como os autovalores são definidos em ordem decrescente, as primeiras coordenadas principais capturam a maior parte da variação total. Portanto, o mapa perceptual das primeiras coordenadas principais identifica, nos casos em que a estatística qui-quadrado é significativa, as categorias das variáveis em estudo que apresentam os maiores desvios da hipótese de independência entre as variáveis e, consequentemente, que ocorrem juntas com maior frequência do que o esperado.

A Análise Fatorial de Correspondência Múltipla é uma extensão da Análise Fatorial de Correspondência Simples. A AFCS estuda a associação entre duas variáveis categóricas, enquanto a AFCM estuda a associação entre mais de duas variáveis categóricas. Na AFCS, a estatística χ^2 é utilizada para verificar a associação entre as duas variáveis. No entanto, na AFCM, não é possível utilizar a χ^2, pois ela é calculada com base em uma tabela de contingência bidimensional (Fávero *et al.*, 2009), não sendo possível ser aplicada para verificar a existência da associação de três ou mais categorias simultaneamente. Na AFCM, os autovalores são calculados para cada categoria das variáveis. As coordenadas das categorias são definidas com base nos autovalores. Essas coordenadas são utilizadas para representar as categorias no mapa perceptual. O mapa perceptual permite visualizar como as categorias estão distribuídas.

Como não é possível representar todas as categorias em uma tabela bidimensional, é necessário encontrar uma forma de representar conjuntamente as Q variáveis com suas respectivas categorias. Para isso, um primeiro procedimento da análise de correspondência

múltipla é construir uma tabela lógica. Nessa tabela, para cada categoria de cada variável, atribui-se 1 se uma determinada observação apresenta a característica e 0, caso contrário. A tabela lógica é uma matriz binária $\mathbf{X}$, na qual nas linhas estão as unidades de observação e nas colunas as diferentes categorias de cada uma das q variáveis presentes na análise (Ferreira Filho *et al.*, 1998).

Após a construção da tabela lógica, a análise de correspondência múltipla segue com o cálculo da matriz de Burt, representada pela equação $\mathbf{B} = \mathbf{X}'\mathbf{X}$. A matriz de Burt é uma matriz quadrada e simétrica, composta por matrizes menores. As matrizes do bloco diagonal da matriz de Burt representam a distribuição de frequência de cada variável. As demais matrizes, localizadas fora do bloco diagonal, representam todas as possíveis tabelas duas a duas, das variáveis presentes no estudo.

A análise de correspondência múltipla pode ser realizada a partir da tabela lógica ou da tabela de Burt. Embora sejam duas tabelas diferentes, ambas possuem características específicas. Os procedimentos utilizados são análogos aos da análise de correspondências simples, com o objetivo de obter as coordenadas de cada categoria das variáveis para serem representadas no mapa perceptual. Em um estudo com Q variáveis, sendo que a q-ésima variável possui J_q categorias, a inércia principal total de $\mathbf{X}$ é obtida pela Equação (2.7).

$$I_T = \frac{\sum_{q=1}^{Q}(J_q - 1)}{Q} = \frac{J - Q}{Q}, \text{ em que } J = \sum_{q=1}^{Q} J_q \,. \tag{2.7}$$

A inércia é uma medida da variabilidade dos dados. Ela é calculada a partir do número de variáveis e suas respectivas categorias. Como demonstrado na Equação (2.7), quanto menor for a frequência de uma determinada categoria, maior será sua contribuição para a inércia. Desta forma, categorias com baixa frequência acabam se destacando no mapa perceptual, muitas vezes dificultando a identificação das relações entre as categorias das demais variáveis.

Para evitar esse problema, recomenda-se que, se possível, categorias com baixa frequência sejam agrupadas a outras, ou então excluídas da análise e estudadas separadamente. Eliminados esses casos de baixa frequência, o mapa perceptual pode ser interpretado de forma análoga ao caso da análise de correspondência simples.

As primeiras coordenadas principais capturam a maior parte da variabilidade dos

dados. Portanto, a interpretação das mesmas será suficiente para identificar as principais associações simultâneas entre categorias das diferentes variáveis.

2.4 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) é uma técnica de análise multivariada que visa examinar a estrutura de interdependência entre um conjunto de variáveis observadas em um conjunto de dados. Esse método explora a estrutura de variâncias e covariâncias das variáveis, procurando identificar combinações lineares (componentes principais) dessas variáveis. Essa abordagem permite a redução da complexidade do problema em estudo, simplificando a análise e facilitando a interpretação das interdependências entre elas (Johnson, 2008).

Em síntese, a ACP tem como objetivos principais a descrição e compreensão da estrutura de dependência entre as variáveis, a redução da complexidade do problema (ao diminuir a dimensionalidade) e a obtenção de novas variáveis, representadas por combinações lineares das variáveis originais, as quais sejam mais facilmente interpretáveis. Essa técnica visa não apenas identificar as relações entre os dados, mas também simplificar a representação dessas relações de maneira compreensível e prática.

O objetivo primordial é encontrar uma representação dos indivíduos e variáveis em um espaço reduzido, simplificando a análise e interpretação dos dados em estudo. Dentro dessa perspectiva, o método dos componentes principais se propõe a definir um novo espaço que seja função de todas as unidades e variáveis observadas, capturando o máximo possível da variabilidade dos dados.

O propósito da ACP é simplificar a disposição dos dados, visando a representação gráfica dos indivíduos em um espaço R_k (onde k representa o número de variáveis) e das variáveis em um espaço R_n (onde n representa o número de indivíduos) da maneira mais simplificada possível. Em outras palavras, a ACP busca novos referenciais tanto para a nuvem de pontos formada pelos indivíduos quanto para a nuvem de pontos formada pelas variáveis. Estes dados podem ser organizados em uma tabela, representada geometricamente de duas maneiras distintas: no espaço dos indivíduos e no espaço das variáveis. Neste contexto, os indivíduos são vetores, com as observações das k variáveis como coordenadas, dispostas nas linhas da tabela. Analogamente, as k variáveis são representadas por vetores, onde as medidas referentes às características dos n indivíduos são coordenadas

dispostas nas colunas da tabela.

Assim, na representação geométrica, torna-se viável observar as distâncias entre quaisquer duas observações. Quanto menor a distância entre as duas observações no gráfico, maior será a similaridade entre esses dois indivíduos. Para analisar a semelhança entre os indivíduos i e j quaisquer, pode-se empregar a distância euclidiana, calculada pela Equação (2.8):

$$d^2(i,j) = \sum_{k=1}^{k} (x_{ik} - x_{jk})^2 \tag{2.8}$$

Para analisar a relação entre duas variáveis k e p quaisquer, avalia-se por meio do coeficiente de correlação linear, expresso pela Equação (2.9):

$$r(k,p) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_{ik} - \bar{x}_k}{S_{xk}} \right) \left(\frac{x_{ip} - \bar{x}_p}{S_{xp}} \right) \tag{2.9}$$

Comumente, as variáveis observadas possuem diferentes unidades de medida, o que pode resultar em valores de variância bastante discrepantes entre elas. Dado que a Análise de Componentes Principais (ACP) se baseia na decomposição da variância total das variáveis (soma das variâncias individuais), o fato de uma variável específica apresentar uma variância significativamente maior do que as outras terá um impacto direto nos componentes obtidos pela ACP. Por exemplo, uma única variável com variância muito acima das demais pode dominar um componente principal, prejudicando desta forma a proposta de redução da dimensão dos dados.

Para mitigar essa questão, é recomendável centralizar (padronizar) todas as variáveis, evitando assim esse problema. Padronizados os dados e calculando a matriz de variância e covariância dos mesmos, verifica-se que a mesma é a matriz de correlações dos dados originais (Johnson, 2008).

Os componentes principais na Análise de Componentes Principais são combinações lineares ortogonais das variáveis em estudo, obtidas através da decomposição da matriz de variâncias e covariâncias ou, no caso de variáveis centralizadas, da matriz de correlações na escala original. Os coeficientes dessas combinações são determinados pelos autovetores associados aos autovalores resultantes da decomposição espectral da matriz de variâncias

e covariâncias (Johnson, 2008). Os coeficientes da primeira componente principal são os autovetores do maior autovalor, seguidos pelos autovetores dos autovalores subsequentes para as outras componentes principais.

Na representação gráfica da Análise de Componentes Principais, os componentes são colocados nos eixos, permitindo a criação de gráficos que combinam diferentes pares de componentes principais. Nestes gráficos, os coeficientes dos componentes principais para cada variável no estudo são exibidos. Por exemplo, ao visualizar o gráfico do Componente Principal 1 em relação ao Componente Principal 2, cada ponto do gráfico é determinado pelos coeficientes de cada variável em seu respectivo componente.

Ao analisarmos os segmentos de reta que vão da origem até o ponto observado para cada variável, podemos identificar que, quanto maior for o comprimento desses segmentos de reta, maior será a contribuição da variável no componente em estudo e, quanto menor for o ângulo formado pelos segmentos de reta associados a duas variáveis, maior será a correlação entre elas.

Ainda, existe uma outra forma de representação gráfica que é realizada após o cálculo do escore para cada indivíduo (ou unidade observada) em cada componente principal. Gráficos são elaborados da mesma maneira que no caso das variáveis, porém, desta vez, cada ponto no gráfico representa o valor do escore nas componentes em questão. Essa representação permite a identificação de indivíduos que se destacam nos componentes utilizados. Além disso, ao analisar simultaneamente os dois gráficos (de variáveis e de indivíduos) para o mesmo par de componentes principais, é possível identificar as características mais relevantes de cada indivíduo.

Capítulo 3

Resultados

3.1 Análise Descritiva

Inicialmente, foram realizadas Análises Descritivas univariada e bivariada, referente às variáveis de cada ano de estudo, ou seja, para o período de 2013 a 2022. Estas análises foram desenvolvidas com o objetivo de examinar as características de interesse a cada ano a fim de identificar a existência de um comportamento das mesmas ao longo do tempo e, possíveis relações a serem investigadas posteriormente de forma conjunta.

Para este estudo preliminar não foram selecionadas todas as variáveis para a análise descritiva, mas sim aquelas consideradas com maior possibilidade de causar impacto na diminuição do número de inscritos no ENEM, foram elas: faixa etária, sexo, há quanto tempo se formou no ensino médio, tipo de escola do ensino médio, tipo de instituição que concluiu ou concluirá o ensino médio, sigla da unidade da federação da escola, grau de escolaridade do pai e da mãe, notas das 5 áreas do conhecimento e número de inscritos presentes e ausentes nos dias de provas.

A Tabela 3.1 apresenta as frequências absoluta e percentual dos indivíduos presentes e ausentes em relação ao número de inscritos no ENEM no período estabelecido do estudo. Lembrando que foram considerados presentes aqueles que compareceram em ambos os dias da aplicação da prova.

Verifica-se pela Tabela 3.1 que o número de inscritos no ENEM sofreu uma drástica diminuição com o decorrer do tempo. Exceção é observada no ano de 2020, primeiro ano da COVID-19, no qual o número de inscritos voltou a crescer, porém foi também o ano com menor proporção de indivíduos presentes (44%) nos dois dias de provas. Para os demais anos a proporção de presentes e ausentes nos dias de provas ficaram em torno de

Tabela 3.1: Número de inscritos, presentes e ausentes nas provas.

Ano	**Presentes**		**Ausentes**		**Inscritos**
	F_i	%	F_i	%	
2013	5007934	69.81	2165629	30.19	7173563
2014	5947909	68.19	2774339	31.81	8722248
2015	5604905	72.35	2141522	27.65	7746427
2016	5818264	67.44	2808915	32.56	8627179
2017	4426692	65.76	2304586	34.24	6731278
2018	3893729	70.62	1620004	29.38	5513733
2019	3701910	72.66	1393261	27.34	5095171
2020	2588681	44.76	3194428	55.24	5783109
2021	2238107	66.02	1151725	33.98	3389832
2022	2344823	67.46	1131282	32.54	3476105

65 a 70% e, 30 a 35% respectivamente.

Ainda, sobre o último exame do ENEM sucedido no ano de 2023, obteve-se um número de inscritos equivalente a 3,9 milhões, destes 2,8 milhões (71.9%) participaram do exame, segundo o Inep (2023). Com isso, pode-se dizer que a quantidade de inscritos e participantes tem aumentado gradativamente a cada ano após o período da pandemia de COVID-19.

A Figura 3.1 diz respeito aos dados apresentados na Tabela 3.1. Observa-se que as linhas possuem um comportamento semelhante, pois a medida que o número de inscritos (representado pela linha azul) aumenta o número de presentes (representado pela linha verde) também aumenta, e o mesmo ocorre quando há um decrescimento, exceto o ano de 2020, como anteriormente destacado, em que o número de inscritos cresce, contudo a quantidade de presentes diminui.

A Tabela 3.2 apresenta as frequências absoluta e percentual das faixas etárias de todos os indivíduos inscritos no ENEM no período estabelecido do estudo.

Tabela 3.2: Frequência absoluta e percentual das faixas etárias.

Ano	**Menor de 18 anos**		**De 18 a 30 anos**		**De 31 a 60 anos**		**Maior de 60 anos**		**Total**
	F_i	%	F_i	%	F_i	%	F_i	%	
2013	1642687	22.90	4435699	61.83	1086396	15.14	8781	0.12	7173563
2014	1974006	22.63	5322250	61.02	1413617	16.21	12375	0.14	8722248
2015	1744845	22.52	4860848	62.75	1130054	14.59	10680	0.14	7746427
2016	1956358	22.68	5394004	62.52	1263799	14.65	13018	0.15	8627179
2017	1370096	20.35	4404181	65.43	947383	14.07	9618	0.14	6731278
2018	1268746	23.01	3546106	64.31	690177	12.52	8704	0.16	5513733
2019	1247945	24.49	3249792	63.78	589175	11.56	8259	0.16	5095171
2020	1112209	19.23	3782501	65.41	876631	15.16	11768	0.20	5783109
2021	895102	26.41	2138977	63.10	349749	10.32	6004	0.18	3389832
2022	1014883	29.20	2170349	62.44	284973	8.20	5900	0.17	3476105

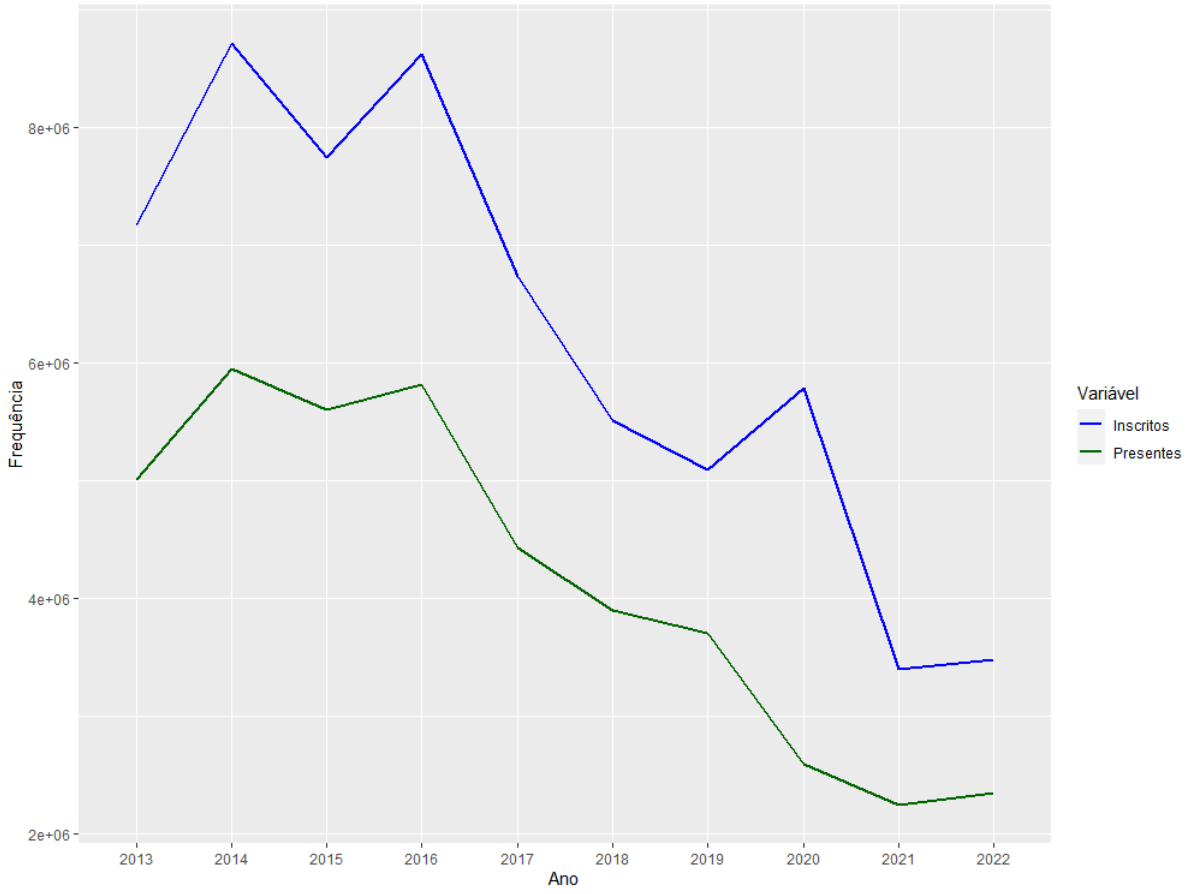

Figura 3.1: Número de indivíduos inscritos e presentes em ambos os dias das provas.

Verifica-se na Tabela 3.2 que a faixa etária de 18 a 30 anos é a que possui maior número de indivíduos, seguido da faixa etária menor de 18 anos, de 31 a 60 anos e por fim, maior de 60 anos. Observa-se que esse comportamento se repete em todos os anos analisados.

É possível analisar ainda que a faixa etária menor de 18 anos sofre um aumento nos últimos dois anos, e o oposto ocorre na faixa etária de 31 a 60 anos, em que nos últimos dois anos sua frequência tem se reduzido.

Com respeito ao sexo dos indivíduos, observa-se pelas Tabelas 3.3 e 3.4 que todos os anos possuem um comportamento muito similar, em que na Tabela 3.3 os indivíduos estão mais concentrados na categoria Feminino com uma proporção sempre próxima de 60%, e consequentemente a categoria Masculino com uma proporção próxima de 40%.

Apesar da variável sexo não ser um possível efeito a ser analisado, pois apresenta um comportamento uniforme ao longo dos anos, foi realizado um mosaico de sua frequência para fins de auxílio visual e comparativo, relacionando assim o seu comportamento com o de uma variável que seja um possível efeito a ser analisado, representado pela Figura 3.2

Tabela 3.3: Frequência absoluta e percentual dos gêneros masculino e feminino.

Ano	**Masculino**		**Feminino**		**Total**
	F_i	%	F_i	%	
2013	2988209	41.66	4185354	58.34	7173563
2014	3652734	41.88	5069514	58.12	8722248
2015	3285979	42.42	4460448	57.58	7746427
2016	3644640	42.25	4982539	57.75	8627179
2017	2784564	41.37	3946714	58.63	6731278
2018	2256035	40.92	3257698	59.08	5513733
2019	2063411	40.50	3031760	59.50	5095171
2020	2314304	40.02	3468805	59.98	5783109
2021	1299306	38.33	2090526	61.67	3389832
2022	1355586	39.00	2120519	61.00	3476105

indicando a proporção de indivíduos do gênero masculino e feminino no período do estudo.

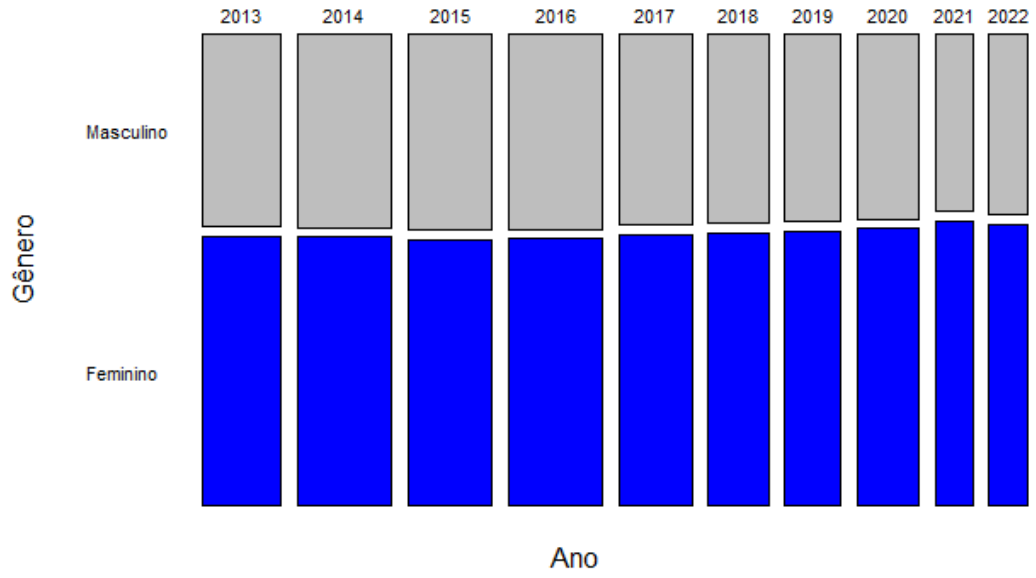

Figura 3.2: Proporção dos gêneros masculino e feminino.

Percebe-se pela Figura 3.2 que os anos que possuem maior número de indivíduos têm colunas mais largas, enquanto aqueles que possuem menos observações apresentam colunas mais estreitas. Ainda, verifica-se que ao longo dos anos as proporções são muito similares, uma vez que as categorias exibem frequências muito parecidas, ou seja, as partes cinzas e azuis (feminino e masculino, respectivamente) são semelhantes em todos os anos.

Quanto ao tipo de escola cursada no ensino médio, verifica-se na Tabela 3.4 que os

indivíduos estão mais concentrados na categoria Pública com uma proporção sempre próxima de 80%, e a categoria Privada com uma proporção próxima de 20%.

Tabela 3.4: Frequência absoluta e percentual dos tipos de escola no período do ensino médio.

Ano	**Pública**		**Privada**		**Total**
	F_i	%	F_i	%	
2013	1315458	80.86	311422	19.14	1626880
2014	5706368	84.75	1026485	15.25	6732853
2015	1344863	81.54	304372	18.46	1649235
2016	1561837	83.02	319395	16.98	1881232
2017	1488632	83.53	293449	16.47	1782081
2018	1137488	81.89	251613	18.11	1389101
2019	1247234	85.09	218627	14.91	1465861
2020	1194496	85.58	201331	14.42	1395827
2021	958611	83.30	192244	16.70	1150855
2022	1105355	83.89	212205	16.11	1317560

Tabela 3.5: Frequência absoluta e percentual dos tipos de ensino no período do ensino médio.

Ano	**Ensino regular**		**Educação especial-Modalidade substitutiva**		**Total**
	F_i	%	F_i	%	
2013	5040277	88.89	630033	11.11	5670310
2014	5958969	88.51	773887	11.49	6732856
2015	1470357	89.15	178878	10.85	1649235
2016	1652430	87.84	228824	12.16	1881254
2017	1583723	88.68	202149	11.32	1785872
2018	3129504	89.85	353571	10.15	3483075
2019	2864336	99.52	13799	0.48	2878135
2020	1294245	99.29	9201	0.71	1303446
2021	1089923	99.37	6905	0.63	1096828
2022	1255177	99.40	7567	0.60	1262744

A partir da Tabela 3.5 é possível perceber que a frequência referente a categoria Ensino regular tende a aumentar conforme o decorrer dos anos, resultando em quase 100% dos indivíduos nos últimos anos do estudo, e consequentemente, a categoria Educação especial diminui com o passar dos anos.

A partir dos dados apresentados na Tabela 3.5 foi criada a Figura 3.3 que ilustra a proporção do tipo de ensino a cada ano. Verifica-se que a proporção da categoria ensino especial diminui com o decorrer dos anos, enquanto a categoria ensino regular tende a aumentar. Além do mais, vê-se que os primeiros dois anos possuem maior número de

observações da variável tipo de ensino, apresentando colunas mais largas, e os três últimos são os que possuem menos observações comparados aos demais.

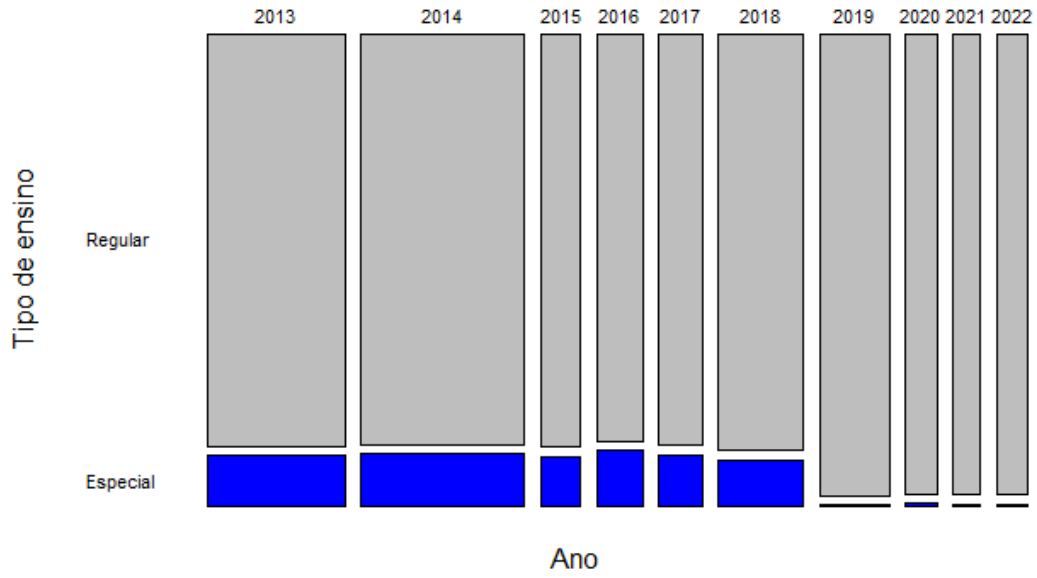

Figura 3.3: Proporção do tipo de ensino.

Tabela 3.6: Frequência absoluta e percentual do tempo de formação do ensino médio.

Ano	Tempo de formação												Total
	Ano atual		Há 1 ano		Há 2 anos		Há 3 anos		Há 4 anos		Há mais de 5 anos		
	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	
2013	1626936	28.69	814344	14.36	540446	9.53	441645	7.79	347138	6.12	1899792	33.50	5670301
2014	1748632	25.97	953127	14.16	670723	9.96	493481	7.33	436474	6.48	2430444	36.10	6732881
2015	1716938	27.95	857487	13.96	636731	10.37	475326	7.74	355969	5.80	2099453	34.18	6141904
2016	1882219	27.64	966814	14.20	699973	10.28	527301	7.74	416449	6.11	2317612	34.03	6810368
2017	1786686	29.49	857157	14.15	604361	9.97	455347	7.52	352365	5.82	2002962	33.06	6058878
2018	1640094	34.17	698806	14.56	454552	9.47	331276	6.90	254651	5.30	1420863	29.60	4800242
2019	1465862	33.64	660946	15.17	447318	10.27	307220	7.05	232858	5.34	1243473	28.54	4357677
2020	1395827	29.99	646471	13.89	437387	9.40	338872	7.28	252946	5.44	1582443	34.00	4653946
2021	1150857	41.74	390967	14.18	247087	8.96	168323	6.10	125733	4.56	674195	24.45	2757162
2022	1317560	47.26	424513	15.23	201403	7.22	163986	5.88	115505	4.14	565015	20.27	2787982

Observa-se pela Tabela 3.6 que a categoria dos indivíduos que se formam no ensino médio no ano atual tende a crescer com o passar dos anos analisados, apresentando uma concentração de quase metade dos inscritos no último ano. A categoria de formação há 1 ano mantém uma frequência semelhante ao longo do período estudado, já as demais categorias apresentam uma queda com o decorrer do tempo, com ênfase no tempo de formação há mais de 5 anos. Tendo visto isso, observa-se que com o passar dos anos

as inscrições de indivíduos que possuem um contato mais recente com o ensino médio aumenta, enquanto aqueles que se formaram há mais tempo tendem a diminuir. Este comportamento é afirmado pela Figura 3.4, em que a proporção de indivíduos que se formaram há mais de cinco anos decresce enquanto a de inscritos que se formam no ano atual aumenta.

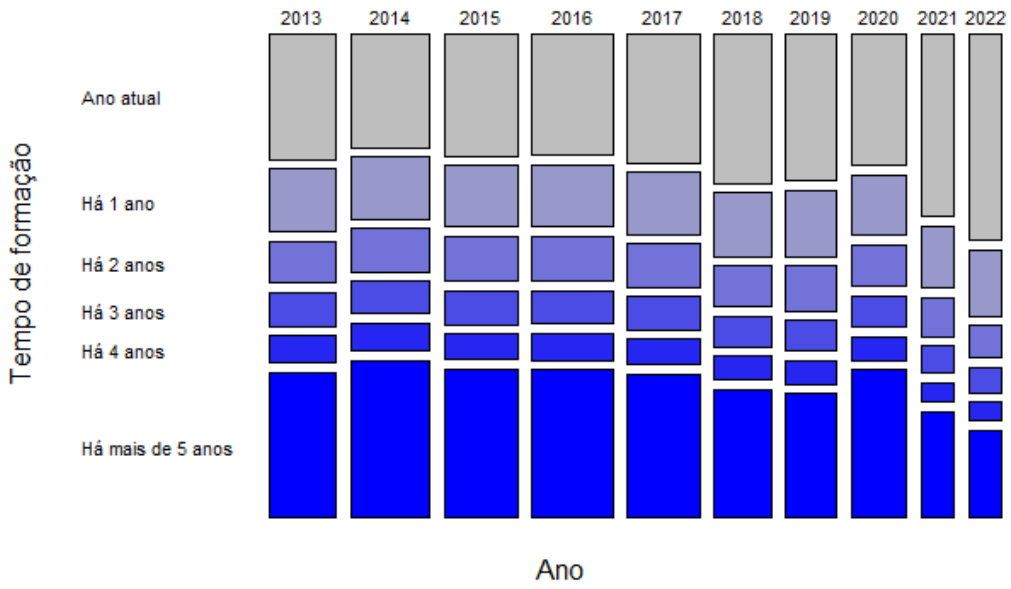

Figura 3.4: Proporção de inscritos por tempo de formação do ensino médio.

Para a análise da distribuição quanto a sua origem segundo as unidades da federação (26 estados do Brasil mais o distrito federal) foram divididas em 5, equivalentes as 5 regiões do Brasil: norte, nordeste, centro oeste, sudeste e sul, para uma melhor interpretação do comportamento dos dados. A frequência dessas 5 categorias está representada na Tabela 3.7.

A Tabela 3.7 explicita que a região Nordeste aumenta sua frequência relativa enquanto a Sudeste decai com o passar dos anos. Já para as demais regiões, suas frequências são similares ao longo do período estudado. Este comportamento também pode ser verificado pela Figura 3.5, em que as proporções da região Nordeste aumentam e da Sudeste decrescem.

A fim de explorar as regiões Nordeste e Sudeste, que apresentaram um comportamento com diferenças das demais, foram analisados separadamente os estados que compõem cada

Tabela 3.7: Frequência absoluta e percentual de inscritos nas 5 regiões do Brasil.

Ano	Norte		Nordeste		Centro Oeste		Sudeste		Sul		Total
	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	
2013	150819	9.22	449747	27.49	136864	8.37	671138	41.03	227280	13.89	1635848
2014	158747	9.03	469512	26.70	202256	11.50	691939	39.35	235897	13.42	1758351
2015	147029	8.91	453311	27.49	142916	8.67	688801	41.76	217201	13.17	1649258
2016	182275	9.69	520315	27.66	159035	8.45	758859	40.34	260739	13.86	1881223
2017	181969	10.19	507930	28.44	147610	8.27	732735	41.03	215637	12.07	1785881
2018	139180	9.61	452951	31.26	124519	8.59	570518	39.38	161648	11.16	1448816
2019	107380	9.36	390115	34.00	90181	7.86	426104	37.14	133607	11.64	1147387
2020	91143	10.08	300582	33.23	86134	9.52	319799	35.35	106911	11.82	904569
2021	82139	10.09	265849	32.67	85251	10.48	276005	33.92	104562	12.85	813806
2022	90326	9.49	299663	31.48	96893	10.18	341597	35.88	123465	12.97	951944

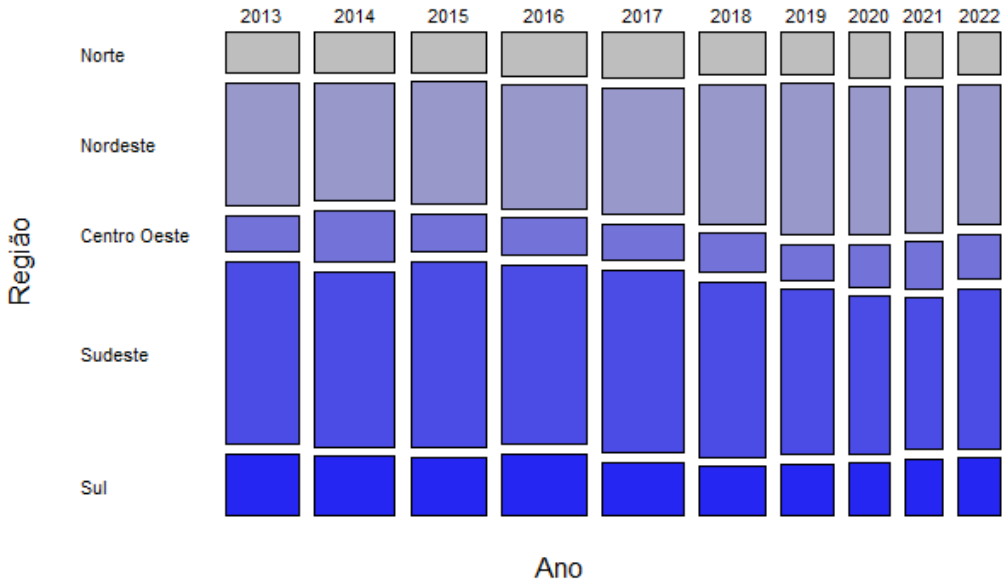

Figura 3.5: Proporção de inscritos nas 5 regiões do Brasil.

uma das duas regiões.

A Tabela 3.8 e Figura 3.6 se referem a região Nordeste. Verifica-se que o estado que se destaca com um comportamento diferente dos demais é o Ceará (CE), apresentando uma frequência crescente com o decorrer dos anos, enquanto os demais estados possuem uma distribuição similar no período analisado. Esse crescimento provavelmente se deve ao bom desempenho do estado do Ceará nos processos avaliativos do ensino médio (G1, 2023).

A Tabela 3.9 e Figura 3.7 dizem respeito a região Sudeste. Verifica-se que o estado

Tabela 3.8: Frequência absoluta e percentual de inscritos na região do Nordeste.

Ano	Sigla referente ao estado																		Total
	AL		BA		CE		MA		PB		PE		PI		RN		SE		
	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	
2013	20456	1.25	91237	5.58	116632	7.13	41735	2.55	31184	1.91	70712	4.32	30814	1.88	28657	1.75	18320	1.12	1635848
2014	21879	1.24	93742	5.33	123550	7.03	48805	2.78	31440	1.79	73931	4.20	31055	1.77	27886	1.59	17224	0.98	1758351
2015	22378	1.36	90675	5.50	118295	7.17	46796	2.84	30971	1.88	74127	4.49	29420	1.78	25316	1.53	15333	0.93	1649258
2016	25596	1.36	108269	5.76	123416	6.56	62508	3.32	33240	1.77	84272	4.48	35088	1.87	28616	1.52	19310	1.03	1881223
2017	25389	1.42	103477	5.79	120701	6.76	64647	3.62	31801	1.78	80121	4.49	34453	1.93	28163	1.58	19178	1.07	1785881
2018	21072	1.45	82825	5.72	119982	8.28	49890	3.44	28402	1.96	82370	5.69	29120	2.01	22728	1.57	16562	1.14	1448816
2019	17551	1.53	70323	6.13	112313	9.79	43496	3.79	26938	2.35	61506	5.36	24145	2.10	20349	1.77	13494	1.18	1147387
2020	13976	1.55	47235	5.22	104981	11.61	26104	2.89	19760	2.18	47583	5.26	15738	1.74	14327	1.58	10878	1.20	904569
2021	10686	1.31	48040	5.90	93159	11.45	21140	2.60	17647	2.17	40227	4.94	13531	1.66	12939	1.59	8480	1.04	813806
2022	15685	1.65	44483	4.67	102972	10.82	27679	2.91	21483	2.26	43562	4.58	16296	1.71	16701	1.75	10802	1.13	951944

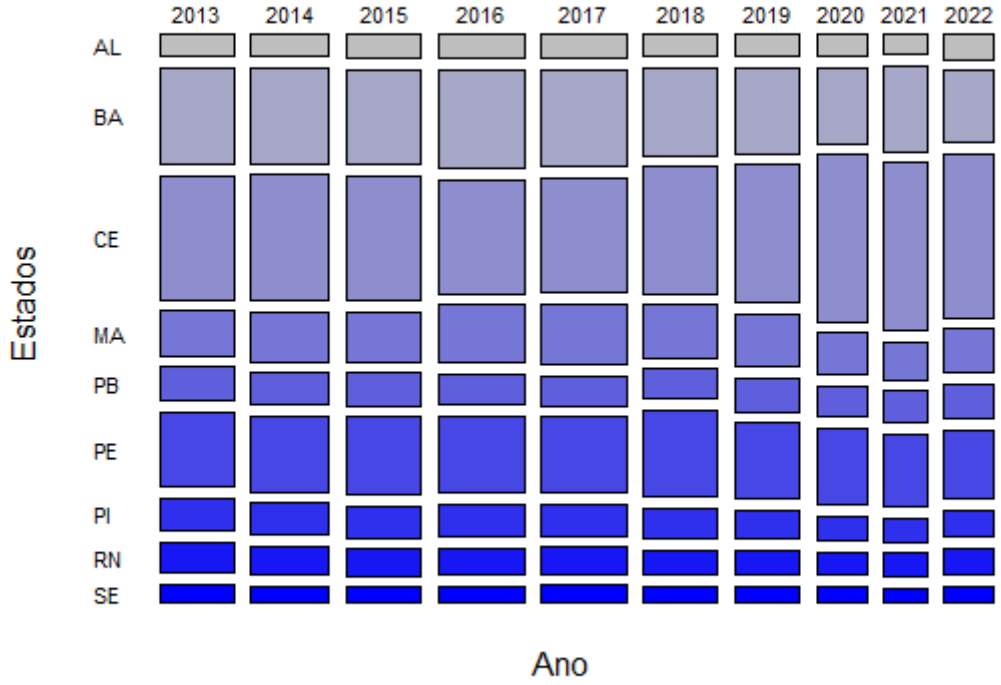

Figura 3.6: Proporção de inscritos na região do Nordeste.

que se destaca com um comportamento diferente dos demais é o de Minas Gerais (MG), apresentando uma frequência decrescente com o decorrer dos anos, enquanto os demais estados possuem uma distribuição similar no período analisado. A priori não há conhecimento de uma possível causa dessa redução do número de inscritos no estado de Minas Gerais.

Nas Tabelas 3.10 e 3.11 e Figuras 3.8 e 3.9 são apresentadas as categorias do grau de escolaridade do pai e da mãe, respectivamente. Verifica-se que ambas possuem o mesmo

Tabela 3.9: Frequência absoluta e percentual de inscritos na região do Sudeste.

Ano	Sigla referente ao estado								Total
	ES		MG		RJ		SP		
	F_i	%	F_i	%	F_i	%	F_i	%	
2013	38304	2.34	167942	10.27	127264	7.78	337628	20.64	1635848
2014	37930	2.16	172018	9.78	129649	7.37	352342	20.04	1758351
2015	38356	2.33	158814	9.63	133135	8.07	358496	21.74	1649258
2016	39414	2.10	181995	9.67	128395	6.83	409055	21.74	1881223
2017	38349	2.15	177552	9.94	130120	7.29	386714	21.65	1785881
2018	31814	2.20	137963	9.52	102500	7.07	298241	20.59	1448816
2019	25422	2.22	108173	9.43	73085	6.37	219424	19.12	1147387
2020	18419	2.04	71614	7.92	62583	6.92	167183	18.48	904569
2021	17657	2.17	64236	7.89	54639	6.71	139473	17.14	813806
2022	19998	2.10	68869	7.23	68435	7.19	184295	19.36	951944

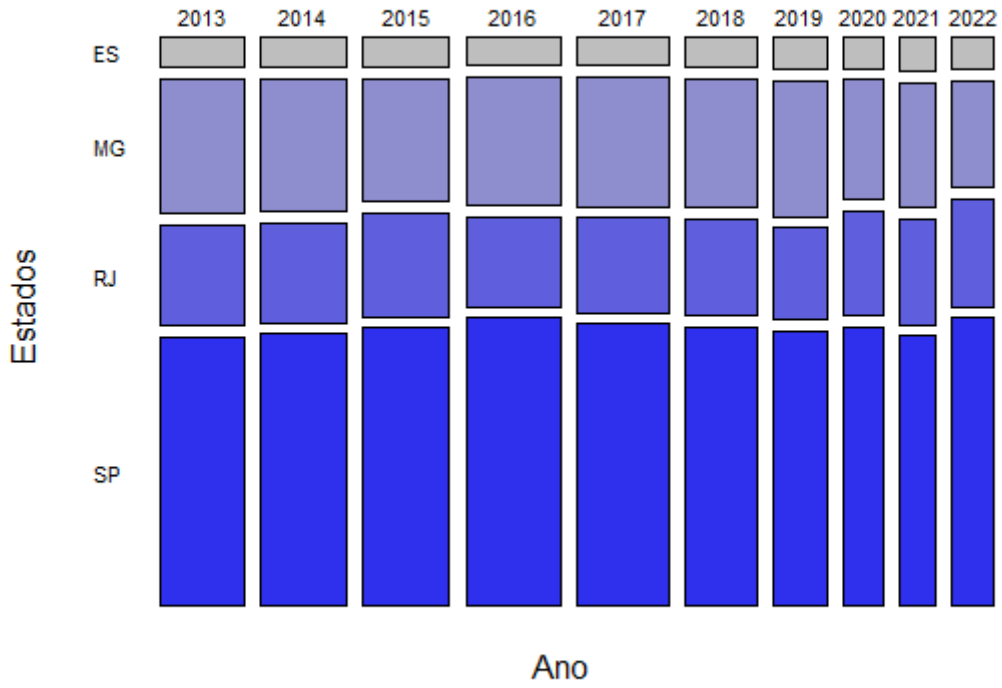

Figura 3.7: Proporção de inscritos na região do Sudeste.

comportamento da distribuição dos dados, em que com o decorrer dos anos a concentração de indivíduos que possuem pai ou mãe com o menor grau de escolaridade, Nunca estudou, tende a diminuir, o mesmo ocorre com a categoria Fundamental 2. Já para as classes

Ensino médio, Ensino superior e Pós graduação acontece o oposto, havendo um aumento de sua frequência ao longo do tempo. Para a categoria Fundamental 1 percebe-se que há um crescimento de sua concentração entre os períodos de 2015 a 2020, e nos últimos anos seu comportamento tende a decair novamente, com uma proporção semelhante aos dois primeiros anos analisados. Tendo visto isso, vê-se que com o passar dos anos as inscrições de indivíduos de pais com maior grau de escolaridade aumenta, enquanto aqueles que possuem pais com menor grau de escolaridade tendem a diminuir.

Tabela 3.10: Frequência absoluta e percentual do grau de escolaridade do pai.

Ano	**Nunca estudou**		**Fundamental 1**		**Fundamental 2**		**Ensino médio**		**Ensino superior**		**Pós graduação**		**Total**
	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	
2013	532591	8.18	2165244	33.25	1631507	25.06	1557925	23.93	451721	6.94	172185	2.64	6511173
2014	664023	8.44	2600281	33.05	1988573	25.28	1880933	23.91	530325	6.74	203245	2.58	7867380
2015	482477	6.90	3032914	43.35	959339	13.71	1819907	26.01	453032	6.47	249033	3.56	6996702
2016	556841	7.15	3396096	43.63	1073176	13.79	1999779	25.69	490010	6.29	268701	3.45	7784603
2017	421767	6.93	2607942	42.82	814764	13.38	1602873	26.32	411862	6.76	231033	3.79	6090241
2018	325041	6.47	2061612	41.06	641006	12.77	1384875	27.58	383845	7.64	224606	4.47	5020985
2019	290612	6.25	1859304	39.98	586610	12.61	1308970	28.15	375707	8.08	229177	4.93	4650380
2020	336216	6.49	2163207	41.74	615284	11.87	1444832	27.88	382851	7.39	239720	4.63	5182110
2021	148389	4.79	1091134	35.24	373620	12.07	955155	30.85	308151	9.95	219969	7.10	3096418
2022	139728	4.45	1058055	33.67	388901	12.38	1002676	31.91	316880	10.08	236316	7.52	3142556

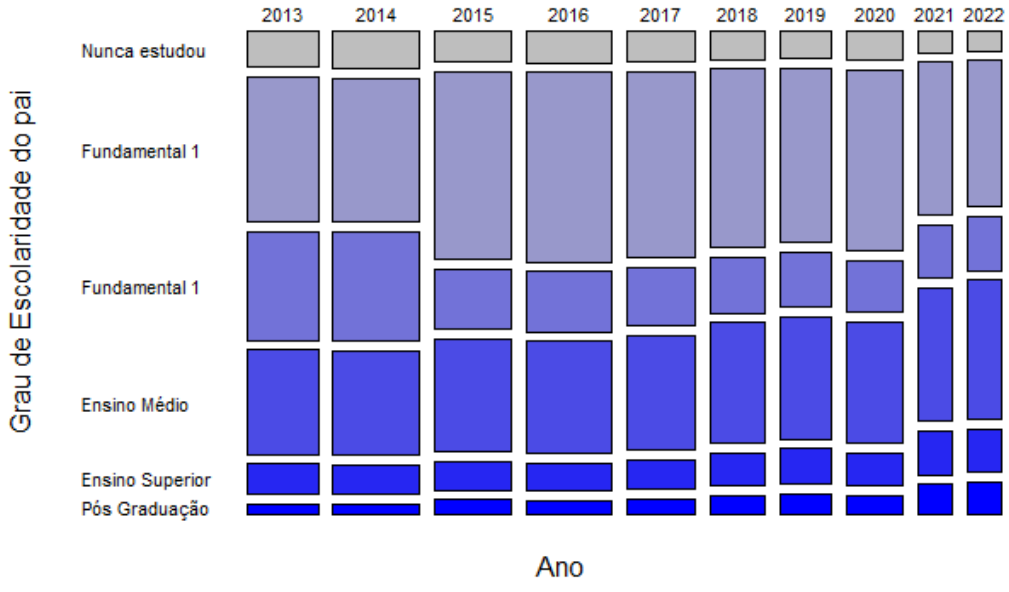

Figura 3.8: Proporção do grau de escolaridade do pai.

Com relação ao aproveitamento dos candidatos nas diferentes áreas do ENEM, observa-se na Tabela 3.12 os valores das médias e desvios padrões das 5 áreas do conhecimento:

Tabela 3.11: Frequência absoluta e percentual do grau de escolaridade da mãe.

Ano	Nunca estudou		Fundamental 1		Fundamental 2		Ensino médio		Ensino superior		Pós graduação		Total
	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	F_i	%	
2013	404035	5.84	1856998	26.85	1825608	26.40	1910626	27.63	589649	8.53	329288	4.76	6916204
2014	502973	6.01	2220062	26.53	2224319	26.58	2329119	27.83	700611	8.37	390701	4.67	8367785
2015	374691	5.04	2695035	36.25	1093654	14.71	2229080	29.99	602518	8.11	438649	5.90	7433627
2016	424006	5.14	2991112	36.24	1232497	14.93	2479072	30.03	658213	7.97	469489	5.69	8254389
2017	312351	4.82	2261190	34.89	940208	14.51	1998623	30.84	559109	8.63	408993	6.31	6480474
2018	228127	4.27	1743949	32.65	735178	13.77	1721766	32.24	520864	9.75	390979	7.32	5340863
2019	194620	3.93	1541001	31.13	667400	13.48	1633666	33.00	513696	10.38	399459	8.07	4949842
2020	239085	4.32	1832525	33.13	697557	12.61	1796085	32.47	532008	9.62	434437	7.85	5531697
2021	99094	3.00	861670	26.11	399095	12.09	1148575	34.81	411426	12.47	380126	11.52	3299986
2022	89040	2.65	817948	24.30	415361	12.34	1217025	36.15	424771	12.62	402007	11.94	3366152

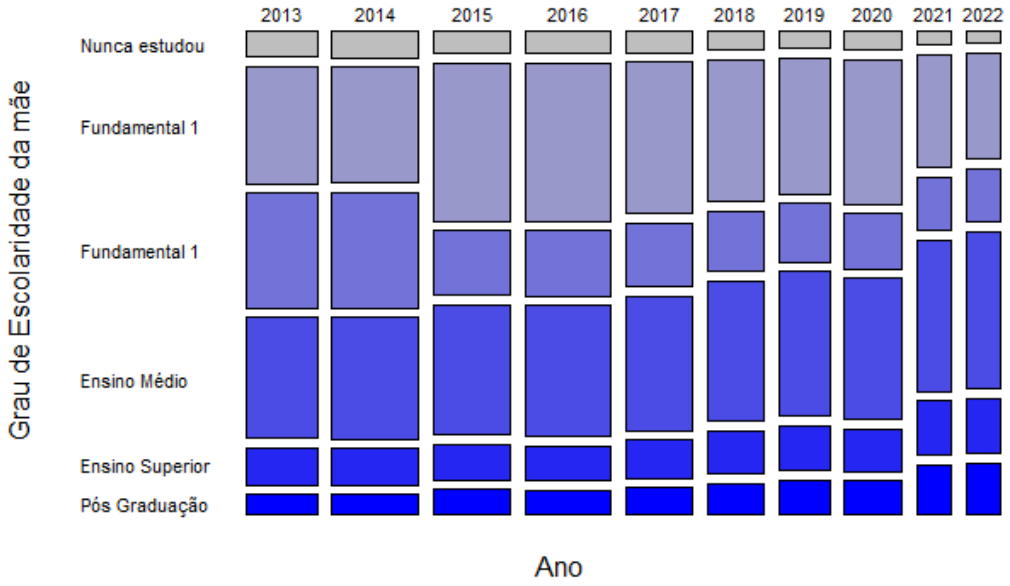

Figura 3.9: Proporção do grau de escolaridade da mãe.

linguagens e códigos, ciências humanas, ciências da natureza, matemática e redação em cada um dos anos estudados.

É possível registrar que para todos os anos existe um comportamento muito semelhante da média e desvio padrão nas 5 áreas do conhecimento, pois todos os anos possuem valores parecidos de suas medidas.

As Figuras 3.10, 3.11, 3.12, 3.13 e 3.14 apresentam o comportamento dos dados referente as notas das 5 áreas do conhecimento dos indivíduos que participaram de ambos os dias das provas.

Percebe-se pela Figura 3.10 que a distribuição dos dados das notas de Linguagens e códigos é semelhante em todos os anos, os valores da mediana, 1º e 3º quartil são bem

Tabela 3.12: Média e desvio padrão das 5 área de conhecimento.

Ano	Linguagens e códigos		Ciências humanas		Ciências da natureza		Matemática		Redação	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
2013	515.74	70.97	541.20	80.69	486.58	75.56	506.17	111.01	564.63	153.94
2014	515.54	70.94	540.78	80.72	486.26	75.63	506.03	111.01	563.64	153.88
2015	515.71	71.05	541.05	81.00	486.43	75.92	506.13	110.91	564.31	154.30
2016	515.56	71.00	540.97	80.70	486.24	75.59	505.93	110.95	563.62	154.21
2017	515.67	70.99	540.94	80.72	486.28	75.59	505.88	111.03	563.94	154.11
2018	515.83	70.79	541.34	80.67	486.47	75.74	506.25	111.10	564.05	154.37
2019	515.60	70.84	540.97	80.57	486.29	75.50	505.95	111.15	563.88	154.11
2020	515.57	70.98	540.84	80.65	486.41	75.50	505.88	110.95	563.98	154.40
2021	515.72	70.88	541.03	80.63	486.26	75.70	506.19	111.14	564.24	154.35
2022	515.69	70.88	541.03	80.79	486.40	75.52	505.90	110.79	564.02	154.00

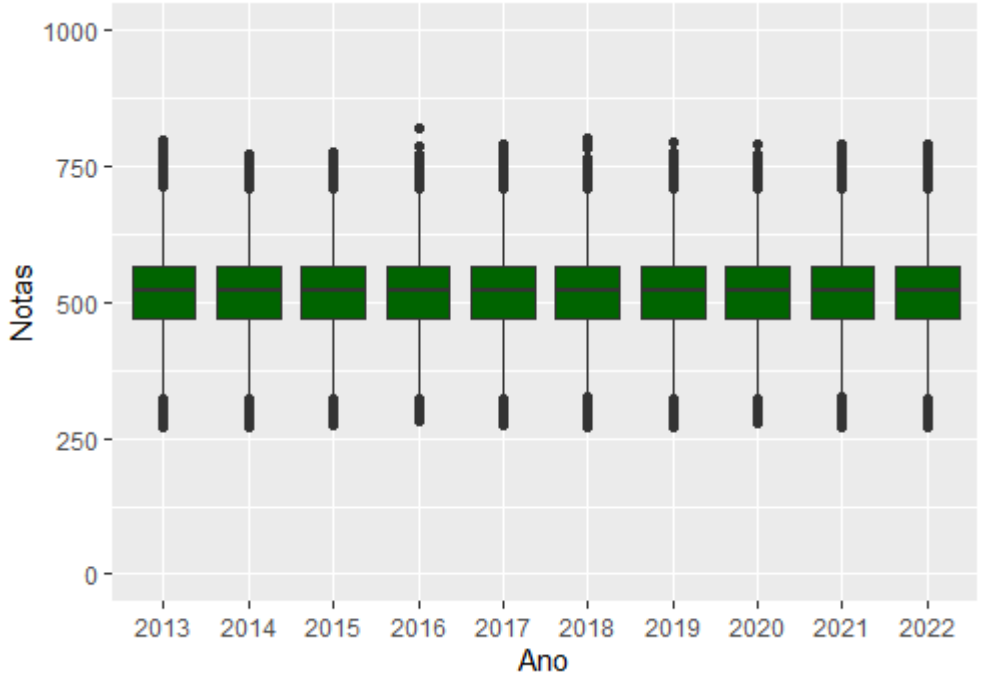

Figura 3.10: Nota dos indivíduos na área de conhecimento em Linguagens e Códigos.

próximos, e todos os anos possuem *outliers* acima do limite superior até aproximadamente 900, e abaixo do limite inferior até próximos de 250. Um comportamento semelhante ocorre em relação a Figura 3.12, que diz respeito a área de Ciências humanas, contudo, vê-se que seus valores se deslocam para valores um pouco maiores comparados a Linguagens e códigos.

A partir da Figura 3.12 observa-se que a distribuição dos dados das notas de Ciências da natureza é semelhante em todos os anos, os valores da mediana, 1º e 3º quartil são bem próximos, e todos os anos possuem *outliers* apenas acima do limite superior até aproximadamente 900. Um comportamento semelhante ocorre em relação a Figura 3.13, que

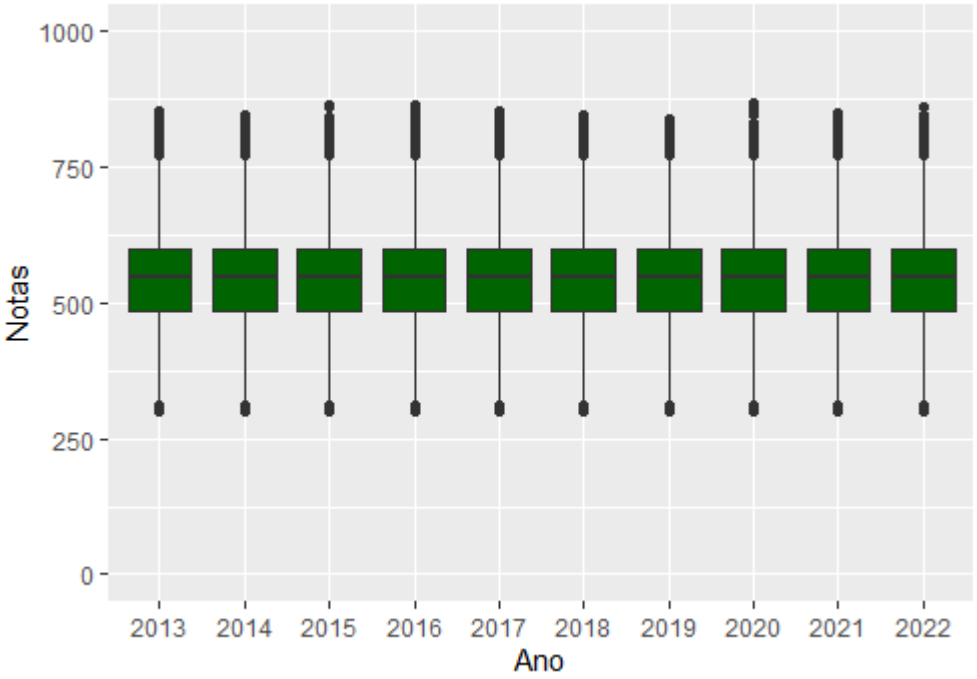

Figura 3.11: Nota dos indivíduos na área de conhecimento em Ciências Humanas.

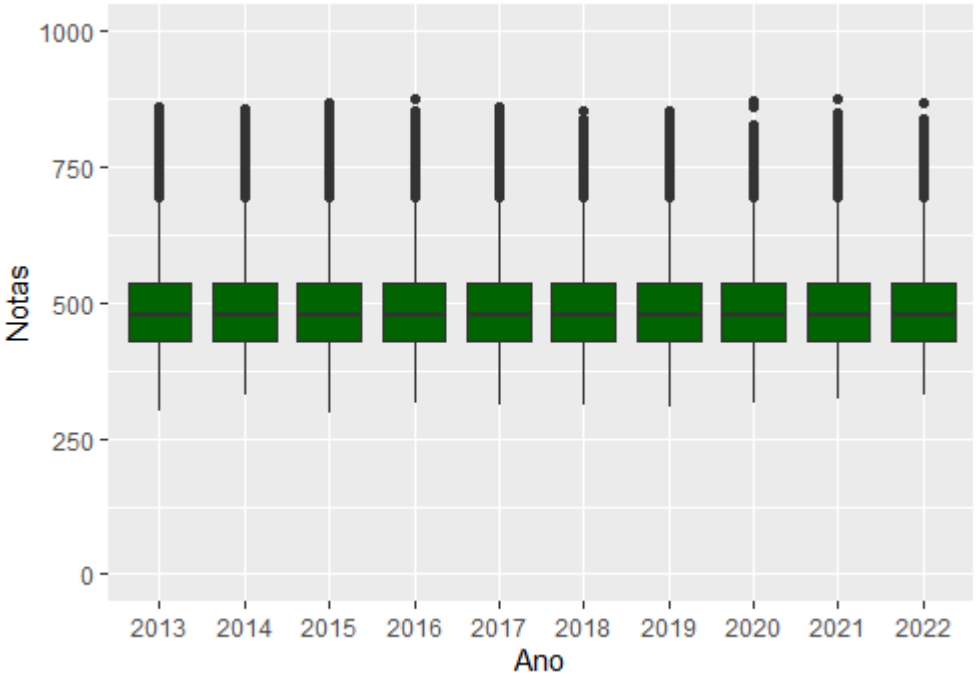

Figura 3.12: Nota dos indivíduos na área de conhecimento em Ciências da Natureza.

diz respeito a área de Matemática, contudo, vê-se que suas observações são mais dispersas, uma vez que os valores discrepantes se aproximam de 1000, e seus valores máximos ultrapassam a pontuação 750, o que indica que sua variabilidade é maior comparada a Ciências da natureza.

Vê-se pela Figura 3.14 as notas das redações no período de estudo estabelecido,

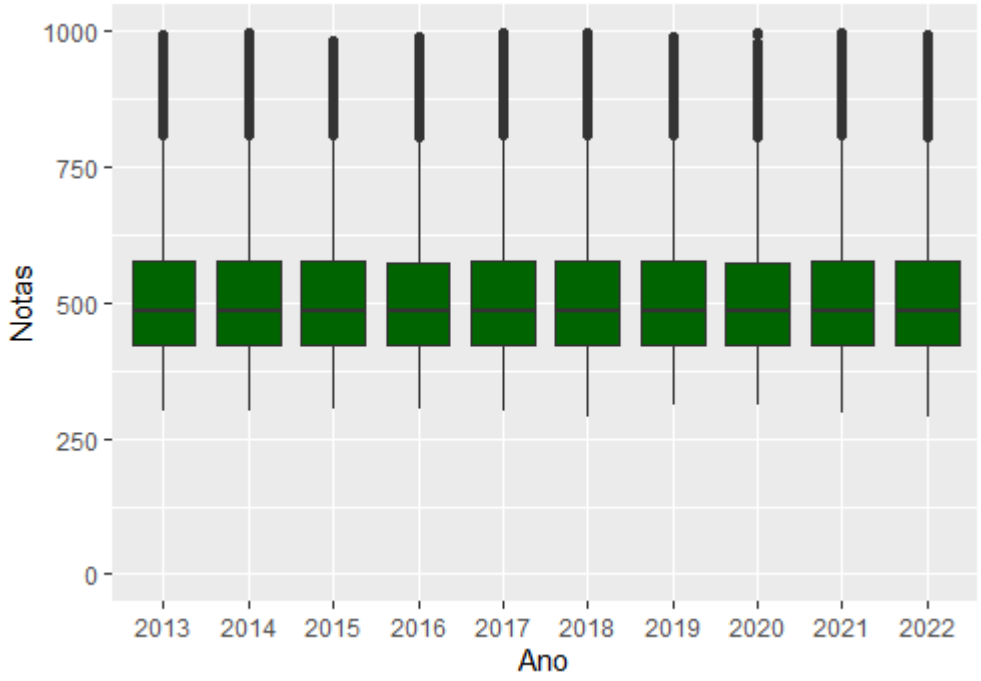

Figura 3.13: Nota dos indivíduos na área de conhecimento em Matemática.

identifica-se que seus valores da mediana e 1º quartil são muito próximos em todos os anos analisados, ainda repara-se que os valores do 3º quartil e limites inferior e superior sofrem uma pequena oscilação com o passar do tempo, se estabilizando nos 3 últimos anos. Ainda é possível perceber que existe uma grande variabilidade nos dados, pois há uma notável dispersão nos valores das notas.

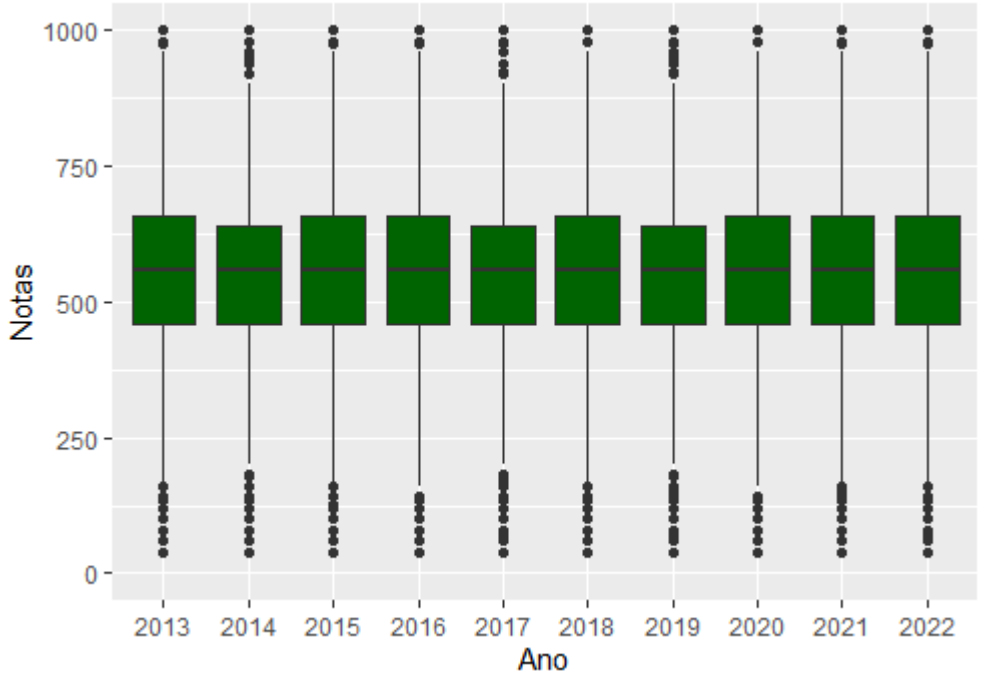

Figura 3.14: Nota dos indivíduos na área de conhecimento em Redação.

A partir dos resultados obtidos na análise descritiva dos dados, foram identificadas as principais características dos inscritos no ENEM ao longo da década investigada. A partir desta caracterização é possível apontar prováveis fatores que podem estar contribuindo para a redução do número de inscritos no ENEM.

As variáveis examinadas que apresentaram um comportamento dos dados semelhante nos anos analisados como a faixa etária, gênero, tipo de escola e sigla da unidade da federação da escola, foram descartadas como possíveis efeitos na diminuição do número de inscritos do ponto de vista univariado, mas serão utilizadas em análises conjuntas com outras variáveis. Foram também mantidas para análises conjuntas aquelas cujo comportamento dos dados sofreu alteração com o decorrer dos anos, característica desejada para a variável ser considerada um possível fator na evasão de inscritos, que foram: tipo de ensino, tempo de formação e grau de escolaridade do pai e da mãe.

Com respeito ao aproveitamento nas provas das 5 áreas do conhecimento foi verificado um mesmo padrão ao longo do período estudado, dado que o desempenho dos inscritos foi próximo ao longo do tempo. A partir deste fato, as notas foram utilizadas também para análises conjuntas, procurando num primeiro momento identificar relações entre as mesmas e após, relacioná-las aos possíveis fatores considerados como responsáveis pela redução do número de inscritos, com o intuito de analisar se existe alguma relação entre as características e as notas obtidas pelos participantes.

3.2 Análise Fatorial de Correspondência Múltipla

Na seção anterior o conjunto de variáveis selecionadas no estudo foi estudada de forma univariada, ou seja, cada uma delas de forma isolada das demais. Sabemos porém que, frequentemente, o impacto das variáveis ocorre de forma conjunta, ou seja, a combinação das categorias de diferentes variáveis podem causar impacto no evento de interesse no estudo, neste caso, a redução de inscritos no ENEM. Com o propósito de identificar possíveis relações conjuntas entre as diferentes características dos inscritos selecionadas para o estudo foi utilizada a Análise Fatorial de Correspondência Múltipla (Mingoti, 2005).

Vale ressaltar que, para uma melhor visualização dos pontos nos gráficos, os nomes das variáveis foram recodificados, como apresentado na Tabela 3.13. Também nesse sentido, algumas categorias das variáveis foram agrupadas, de forma a evitar categorias com baixa frequência. Um exemplo desse agrupamento foi a variável Faixa Etária, em que as duas últimas categorias, De 31 a 60 anos e Maior que 60 anos foram unidas, devido as suas baixas frequências comparadas as demais classes. Além disso, variáveis como Tipo de escola, Tipo de ensino e, Sigla da unidade de federação (Regiões do Brasil) foram descartadas da análise por apresentarem classes muito desbalanceadas entre si e por não ter a possibilidade de agrupamento das mesmas. Esses agrupamentos e seleção de variáveis foram feitos para evitar seu impacto na Análise de Correspondência, pois classes muito desproporcionais podem prejudicar a precisão dos resultados da análise.

Tabela 3.13: Codificação das variáveis qualitativas.

Categoria	**Codificação**
Faixa Etária	
Menor de 18 anos	FE1
18 a 30 anos	FE2
31 anos ou mais	FE3
Sexo/Gênero	
Feminino	F
Masculino	M
Estado Civil	
Solteiro(a)	EC1

Continua na próxima página

Tabela 3.13 – *Continuação da tabela*

Categoria	**Codificação**
Casado(a)/Mora com um(a) companheiro(a)	EC2
Divorciado(a)/Desquitado(a)/Separado(a)/Viúvo(a)	EC3
Cor/Raça	
Branca	RA1
Preta	RA2
Parda	RA3
Amarela	RA4
Indígena	RA5
Ano de Formação do Ensino Médio	
Formação no ano atual	FMA
Formação há 1 ano	FM
Formação há 2 anos	FM2
Formação há 3 anos	FM3
Formação há 4 anos	FM4
Formação há mais de 5 anos	FM5
Nível de Escolaridade do Pai	
Nunca estudou	EP1
Fundamental 1 completo	EP2
Fundamental 2 completo	EP3
Ensino Médio completo	EP4
Ensino Superior	EP5
Pós-graduação	EP6
Nível de Escolaridade da Mãe	
Nunca estudou	EM1
Fundamental 1 completo	EM2
Fundamental 2 completo	EM3
Ensino Médio completo	EM4
Ensino Superior	EM5
Pós-graduação	EM6

Fim da tabela

As Figuras 3.15, 3.16, 3.17 e 3.18 dizem respeito a análise dos anos 2013, 2016, 2019 e 2022, os gráficos relacionados aos demais anos, que apresentam um comportamento similar a estes, encontram-se no Apêndice B.

Para cada um dos anos acima citados duas representações gráficas são apresentadas: A primeira com a representação das duas primeiras dimensões resultantes do uso da AFCM e a segunda com a contribuição das categorias no primeiro plano fatorial formado pelas duas dimensões iniciais.

Verifica-se que o primeiro gráfico que diz respeito ao comportamento conjunto das categorias das diferentes características, apresenta basicamente o mesmo padrão em todos os anos analisados. É possível observar uma proximidade das características Menor de 18 anos, Ano de formação do ensino médio no ano atual, e nível de escolaridade do pai e da mãe mais elevados como ensino superior ou pós-graduação, ou seja, indivíduos mais novos que acabaram de concluir o ensino médio e possuem pais com grau de escolaridade mais elevado, indicando que tais características são as mais comuns de ocorrerem conjuntamente.

Ainda, é possível identificar uma proximidade entre as características 31 anos ou mais, Casado ou Divorciado, Ano de formação do ensino médio há mais de 5 anos, e nível de escolaridade do pai e da mãe mais baixos em Nunca estudou, ou seja, indivíduos mais velhos, sejam eles casados ou divorciados, que concluíram o ensino médio há mais tempo e possuem pais com grau de escolaridade mais baixo, indicando que tais características ocorrerem conjuntamente com maior frequência.

Estes grupos de relações conjuntas mostram que quanto mais jovem é o indivíduo, menor é o ano de formação do ensino médio e, maior é o nível de escolaridade de seus pais, e quanto mais velho é o indivíduo, maior é o ano de formação do ensino médio e, menor é o nível de escolaridade de seus pais. As observações mostram uma inversão na relação entre a faixa etária e o nível de escolaridade dos pais. Isso significa que a idade dos indivíduos está associada não só ao seu próprio nível educacional (ano de formação do ensino médio), mas também ao nível educacional de seus pais, de maneira inversa.

Essa tendência se mantém consistente ao longo dos anos analisados. A relação entre a faixa etária, ano de formação e o nível de escolaridade dos pais persiste e segue uma lógica hierárquica.

Através das Figuras 3.15, 3.16, 3.17 e 3.18, nota-se pelo primeiro gráfico que a proporção da inércia explicada no primeiro plano fatorial corresponde a aproximadamente

30%, este valor é semelhante em todos os anos. O que significa que os 2 primeiros componentes são capazes de capturar 30% das principais tendências dos dados, medindo o quanto os dados são explicados por tais componentes.

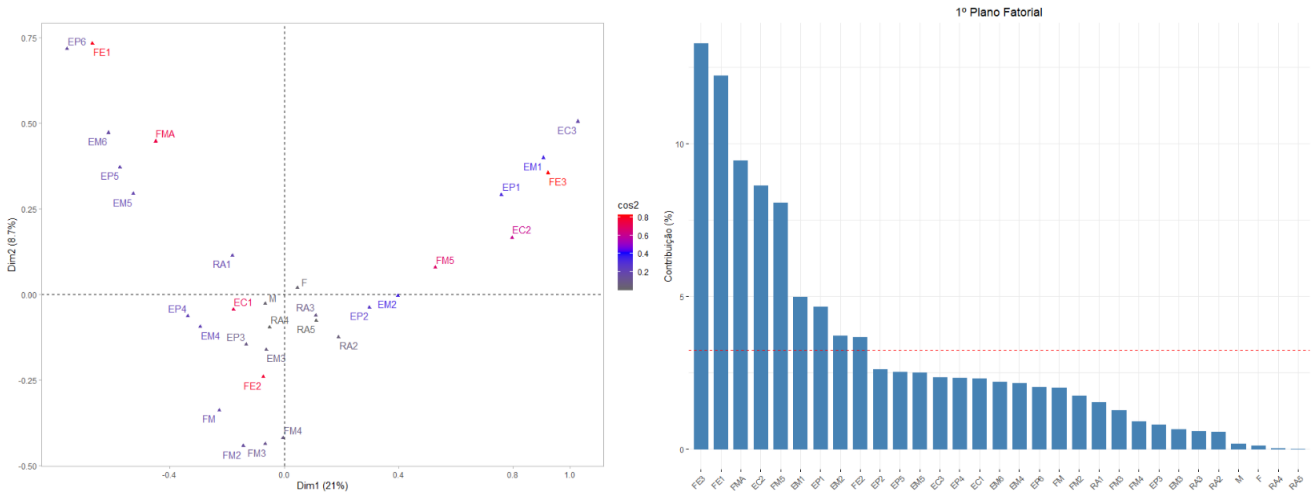

Figura 3.15: AFCM do perfil dos inscritos no ano de 2013.

No segundo gráfico das Figuras 3.15, 3.16, 3.17 e 3.18, é apresentado a contribuição de cada categoria no primeiro plano fatorial formado pelas duas primeiras dimensões da AFCM. Quanto maior o valor observado, maior a importância da categoria no primeiro plano. Verifica-se também neste caso o mesmo padrão de comportamento nos diferentes anos. Com relação às variáveis referentes a gênero e raça, nota-se que as mesmas não apresentam uma alta contribuição para a explicação da inércia no primeiro plano fatorial. Para as demais variáveis, pelo menos uma categoria apresenta contribuição acima da média. Em resumo, a baixa contribuição das variáveis de gênero e raça no primeiro plano fatorial sugere que, em relação aos padrões identificados pela análise de correspondência, essas variáveis não estão tão fortemente ligadas às outras variáveis analisadas.

Um fato que deve ser destacado é que contribuição é inversamente proporcional a frequência da classe, ou seja, quanto menor a frequência de uma característica, maior será sua contribuição nas primeiras dimensões e consequentemente no primeiro plano fatorial.

Identificado o comportamento conjunto das categorias das diferentes características, é importante relacioná-lo com a redução do número de inscritos no ENEM. Para isso dois perfis de indivíduos foram criados. O Perfil 1 é composto pelos indivíduos mais novos que acabaram de concluir o ensino médio e possuem pais com grau de escolaridade mais elevado e, o Perfil 2 por indivíduos mais velhos, casados ou divorciados, que concluíram o ensino médio há mais tempo e possuem pais com grau de escolaridade mais baixo. Para

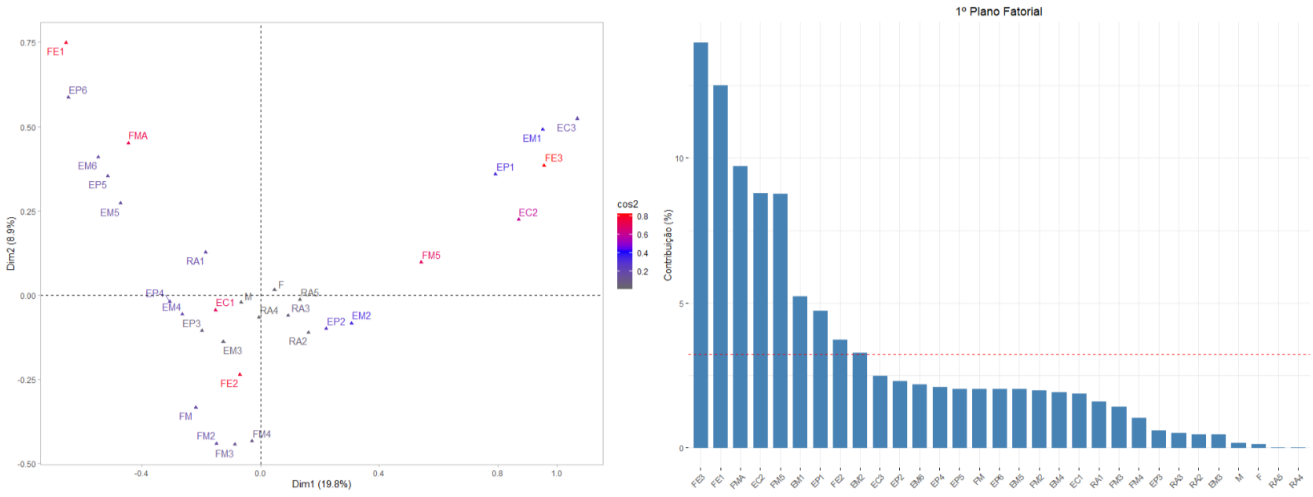

Figura 3.16: AFCM do perfil dos inscritos no ano de 2016.

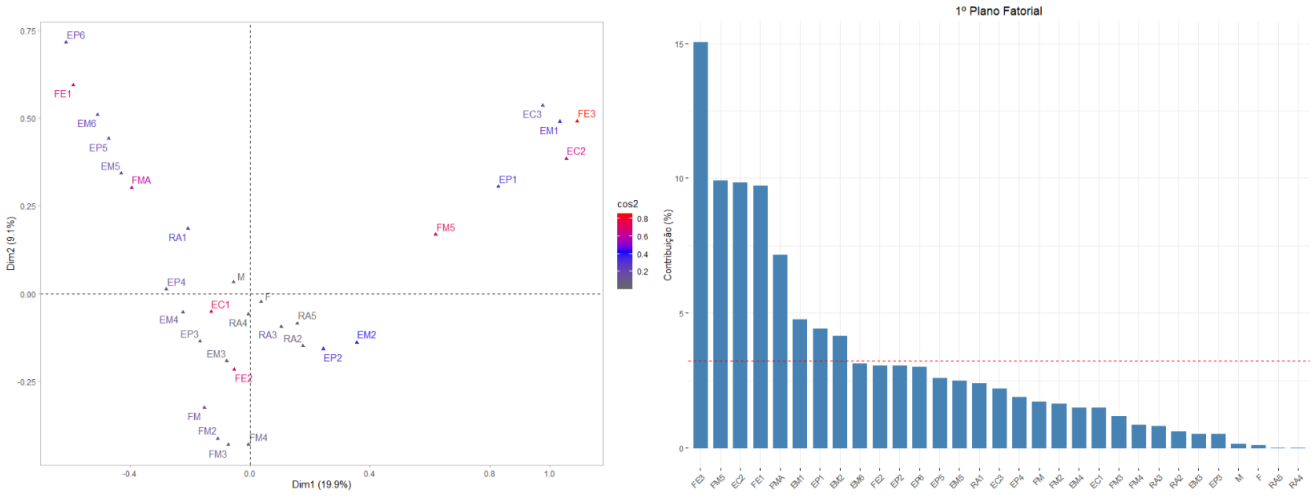

Figura 3.17: AFCM do perfil dos inscritos no ano de 2019.

cada ano do estudo foram selecionados os indivíduos de cada um dos perfis acima e a Tabela 3.14 apresenta a frequência ao longo dos anos dos dois perfis.

Verifica-se que o número absoluto de respondentes, ou seja, indivíduos que apresentaram resposta para todas as variáveis e não possuíam valores faltantes em nenhuma das características qualitativas analisadas na AFCM, representado pela coluna Total, diminui com o passar dos anos, como já observado pela Capítulo 3. Porém o número absoluto de participantes do Perfil 1 permanece com valores próximos em todos os anos, enquanto para aqueles do Perfil 2 esse valor tende a diminuir. De forma lenta a proporção de participantes no Perfil 1 cresce enquanto a de participantes no Perfil 2 se reduz. Tal resultado, indica que candidatos com as características presentes no perfil 2 vêm reduzindo a sua participação no ENEM ao longo dos anos.

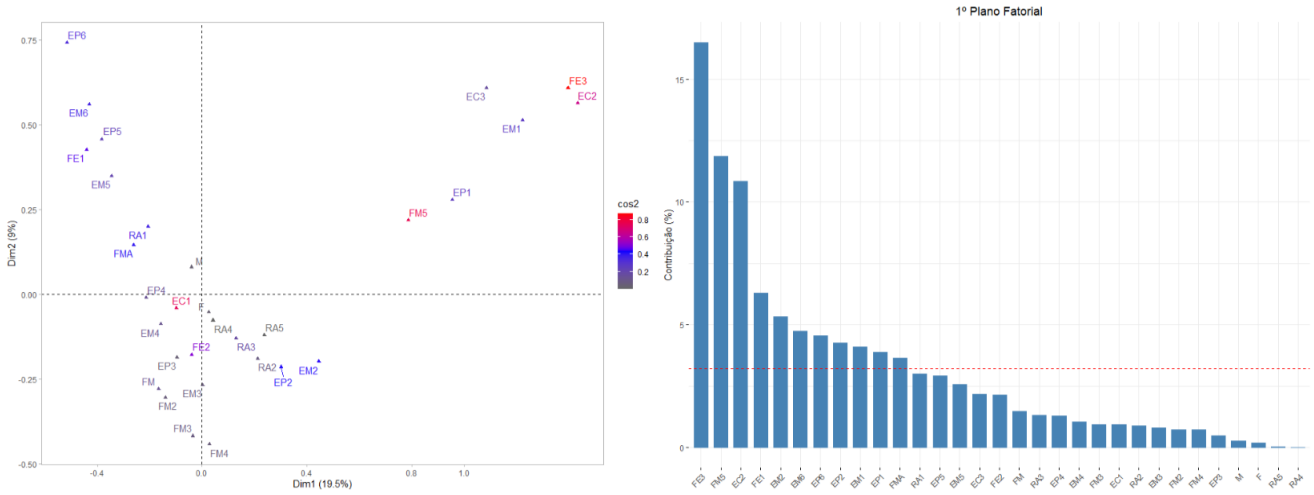

Figura 3.18: AFCM do perfil dos inscritos no ano de 2022.

Tabela 3.14: Número de inscritos nos perfis 1 e 2 a cada ano.

Ano	**Perfil 1**		**Perfil 2**		**Total**
	F_i	%	F_i	%	
2013	71413	1.41	37092	0.73	5064316
2014	74606	1.25	49617	0.83	5989202
2015	72054	1.38	30107	0.58	5207638
2016	76151	1.32	34351	0.6	5751947
2017	74667	1.46	29113	0.57	5110397
2018	67235	1.65	20166	0.5	4072355
2019	68933	1.86	17165	0.46	3706685
2020	70001	1.8	17996	0.46	3887840
2021	68275	2.9	6807	0.29	2353575
2022	71964	3.05	5613	0.24	2360058

Em 2017 o ENEM passou por um grande processo de reformulação (MEC, 2023a), não sendo mais utilizado para certificação de conclusão do ensino médio. Tal fato pode ser destacado como sendo um fator que auxiliou na diminuição da participação de alguns candidatos no Exame, dado que muitas pessoas prestavam a prova com o intuito de concluírem o ensino médio de forma mais rápida.

3.3 Análise de Componentes Principais

Nesta seção foi feita a análise conjunta das variáveis quantitativas presentes no estudo, ou seja, as notas das diferentes provas do ENEM, para isso foi realizada uma Análise de Componentes Principais (ACP). O objetivo neste caso é o de verificar a existência ou não de correlações entre as notas das diferentes provas. Foram obtidos os componentes principais das cinco provas ao longo dos anos de forma a identificar ou não, um padrão de correlação entre as notas das diferentes provas. Nesta seção não utilizaremos a expressão "e suas Tecnologias"que faz parte do nome das provas, exceto a de Redação.

As Figuras 3.19, 3.20, 3.21, 3.22, 3.23, 3.24 e 3.18 dizem respeito a análise dos anos 2013, 2014, 2015, 2016, 2017, 2018 e 2022, os gráficos relacionados aos demais anos, encontram-se no Apêndice B

Nas Figuras 3.19, 3.20, 3.21, 3.22, 3.23, 3.24 e 3.18 são apresentadas a representação gráfica da ACP no primeiro plano fatorial e de seus respectivos scree plots para os anos de 2013 a 2018 e também 2022. A representação para estes anos é feita pelo fato de que é possível identificar um padrão de relação entre as variáveis no período de 2013 a 2016 e um outro padrão, com algumas diferenças para o período de 2017 a 2022 (os anos de 2019 a 2021 não são apresentados por terem o mesmo padrão dos demais anos do período de 2017 a 2022). O scree plot por sua vez tem o mesmo comportamento para todos os anos corroborando o fato de que os dois primeiros componentes são suficientes para uma boa análise das relações entre as notas das cinco provas.

Para todos os anos foi observado que os dois primeiros componentes principais representavam pelo menos 75% da variabilidade conjunta das variáveis. Desta forma todas as cinco provas foram bem representadas no primeiro plano fatorial formado pelas duas primeiras componentes principais.

Para o período de 2013 a 2016, representado pelas Figuras 3.19, 3.20, 3.21 e 3.22, observa-se que o primeiro componente pode ser visto como um escore geral das provas enquanto o segundo pode ser visto como um contraste entre a nota da prova de redação e as notas das provas de Matemática e Ciências da Natureza. Pode ainda ser identificado que as notas de Matemática e Ciências da Natureza são bem correlacionadas assim como as notas de Linguagens e Códigos e de Ciências Humanas. Estes dois blocos apresentam também uma correlação moderada entre eles. A prova de redação por sua vez apresenta-se isolada das demais com uma moderada correlação com o bloco de Linguagens e Códigos

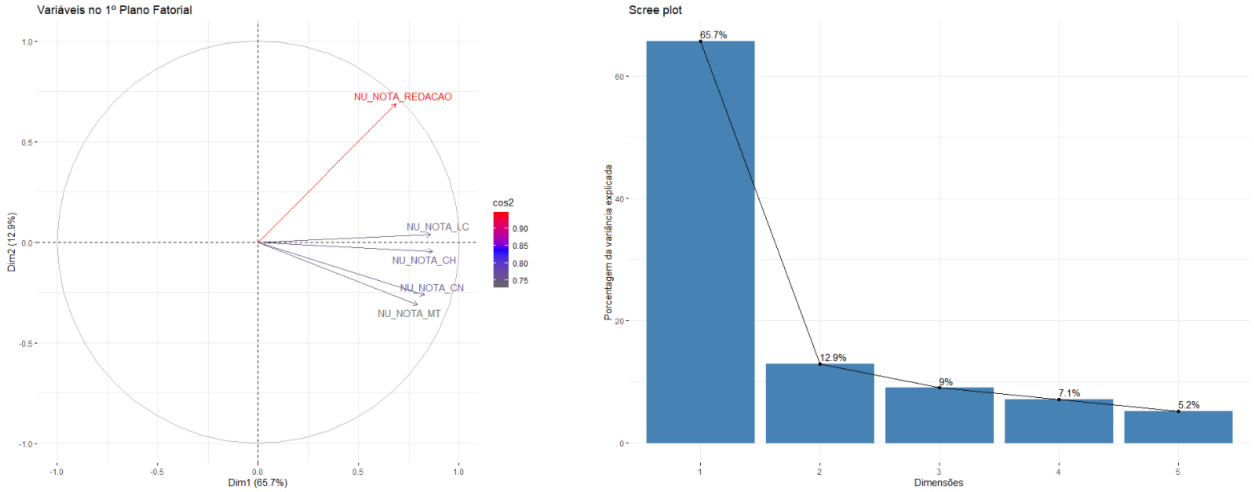

Figura 3.19: ACP das notas dos inscritos no ano de 2013.

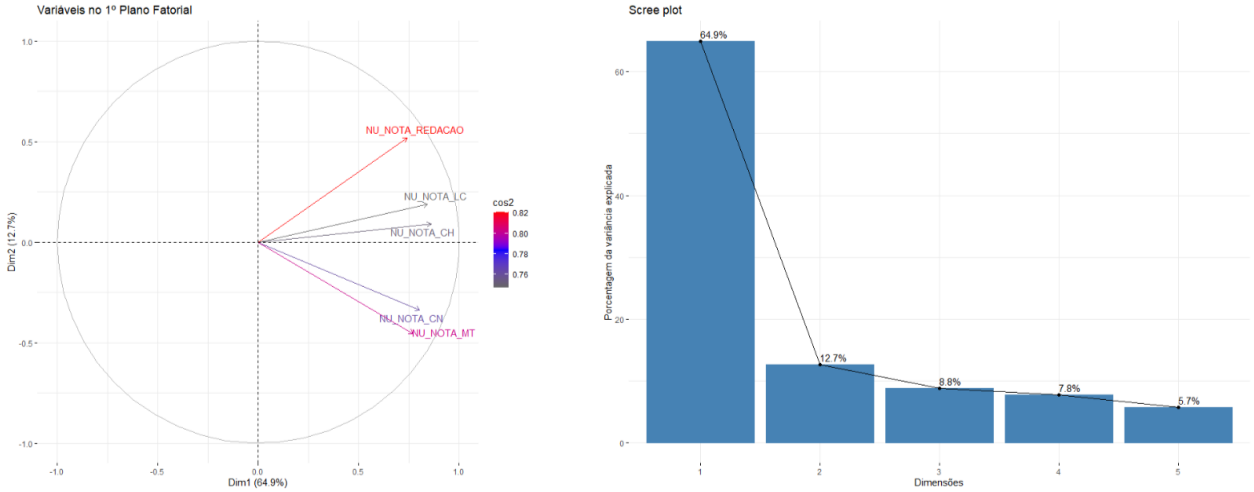

Figura 3.20: ACP das notas dos inscritos no ano de 2014.

e de Ciências Humanas e uma menor correlação com o bloco de Matemática e Ciências da Natureza.

Com relação ao período de 2017 a 2022, representado pela Figuras 3.23, 3.24 e 3.25, vê-se que o primeiro componente também pode ser visto como um escore geral das provas enquanto o segundo pode ser visto como representado principalmente pela nota de redação, que é a prova de destaque para esse eixo, enquanto as demais se mantém próximas do valor 0. Pode ainda ser identificado que as notas referentes as 4 provas, Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática são bem correlacionadas. Já a prova de redação por sua vez apresenta-se isolada das demais com uma correlação fraca com o bloco das 4 provas mencionadas.

Além disso, repara-se que os dois blocos citados anteriormente: de Linguagens e

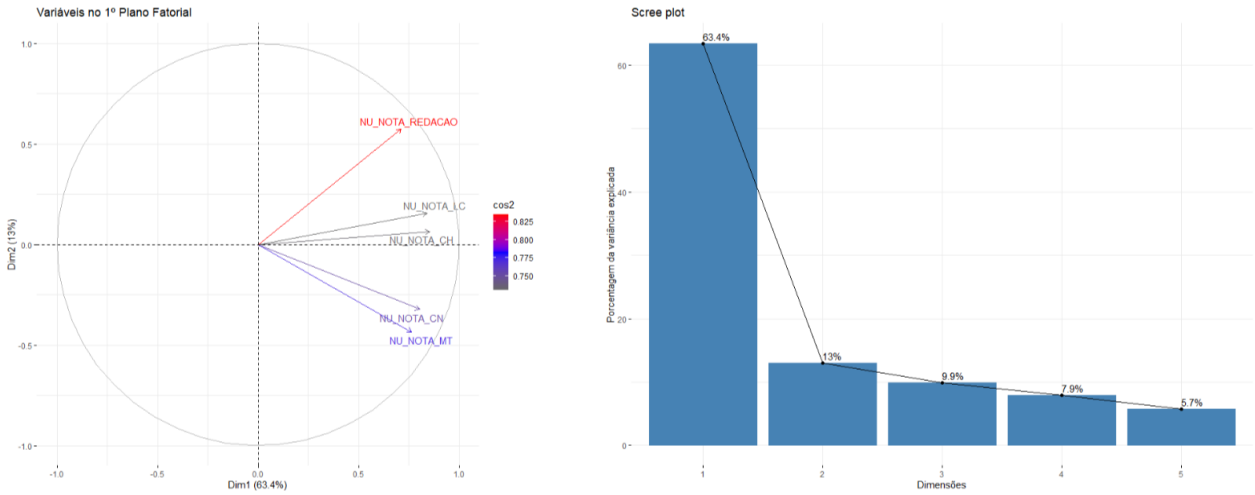

Figura 3.21: ACP das notas dos inscritos no ano de 2015.

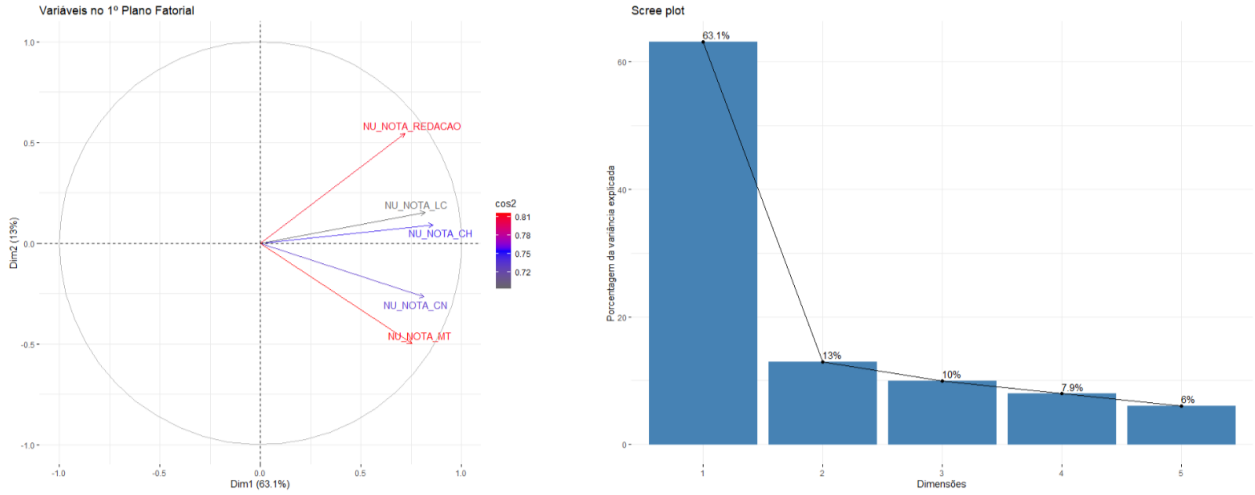

Figura 3.22: ACP das notas dos inscritos no ano de 2016.

Códigos e Ciências Humanas e, o bloco de Matemática e Ciências da Natureza. Apresentados nos anos de 2013 a 2016, se aproximam nos anos de 2017 a 2022, formando apenas um bloco com as 4 provas, o que significa que elas aumentam a sua correlação ao decorrer dos anos, diminuindo ainda a sua correlação com a nota de redação. A diminuição da correlação entre as notas das 4 áreas e a redação, sugere que tais áreas estão se tornando mais relacionadas entre si enquanto se afastam, em termos de correlação, da nota de redação.

A partir do ano de 2017 houve uma grande mudança nos dias de provas do ENEM (MEC, 2023b), que antes eram aplicadas em dias consecutivos, sábado e domingo, passaram a ser com uma semana de diferença, em dois domingos seguidos. A ordem de aplicação das provas também foi invertida, até 2016 no 1° dia eram aplicadas as provas

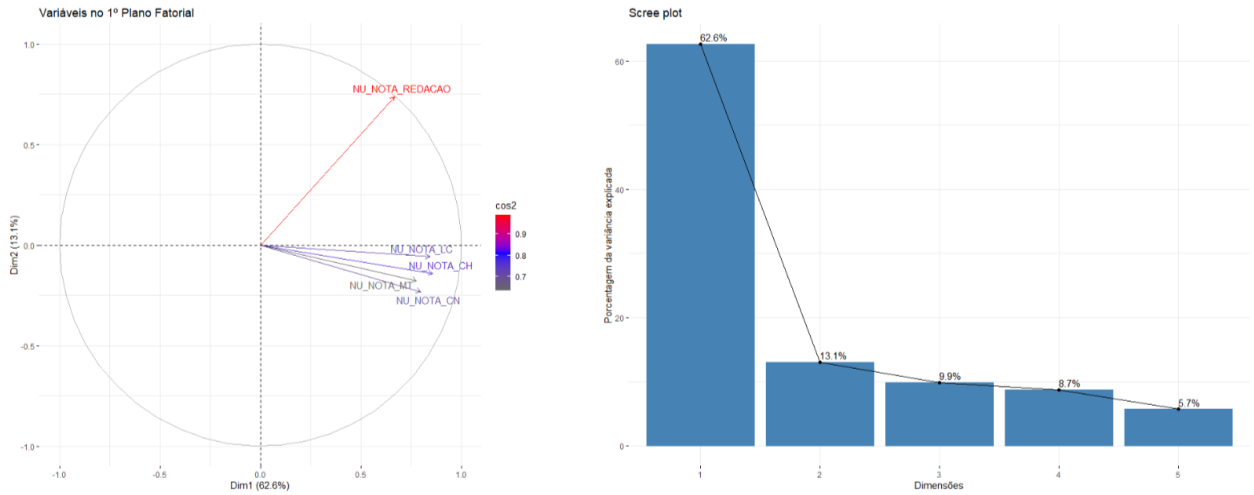

Figura 3.23: ACP das notas dos inscritos no ano de 2017.

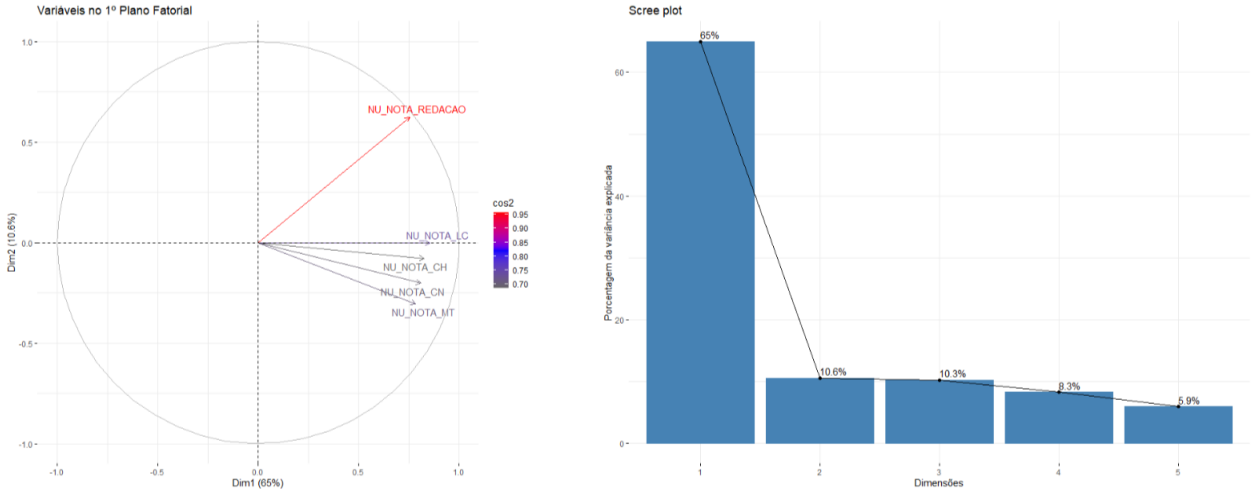

Figura 3.24: ACP das notas dos inscritos no ano de 2018.

de Ciências Humanas e Ciências da Natureza, e no 2° dia de Linguagens e Códigos, Matemática e Redação, já em 2017 passaram a ser aplicadas as provas de Ciências Humanas, Linguagens e Códigos e Redação no 1° dia e, Matemática e Ciências da Natureza no 2° dia. Essas alterações na logística de aplicação das provas do ENEM justificam a mudança de correlação entre as provas, principalmente Redação que passou a ser aplicada no primeiro dia do exame.

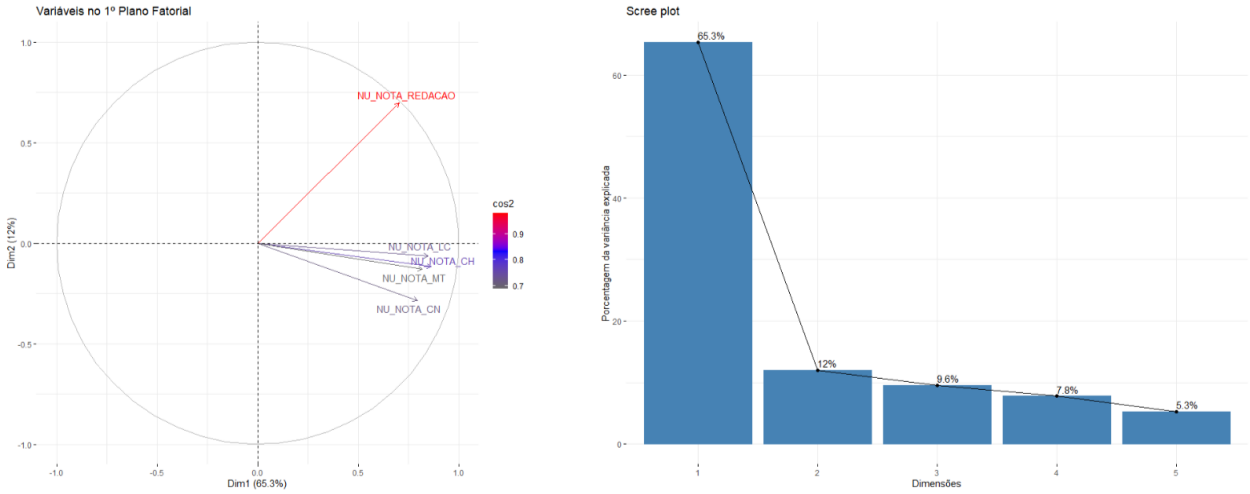

Figura 3.25: ACP das notas dos inscritos no ano de 2022.

3.4 Análise dos Perfis da AFCM combinado com os Escores da ACP

Nesta seção, são apresentados os resultados da análise dos perfis 1 e 2 dos indivíduos obtidos a partir da AFCM na Seção 3.2, com base nos resultados obtidos na ACP da Seção 3.3.

Para isto ser feito foram inicialmente obtidos os escores fatoriais para o primeiro e segundo componentes principais para todos os indivíduos em cada um dos anos. Foram então identificados os indivíduos pertencentes ao perfil 1 e 2 de cada ano e então calculado um escore médio para cada um dos perfis em cada um dos anos. Este par de escores médios dos componentes 1 e 2 para cada ano, referente a cada perfil é apresentado na Figura 3.26, obtida através dos dados apresentados na Tabela 3.15.

Verifica-se que, como observado na Seção 3.3, lembrando ainda que foi utilizado a matriz de correlações para realização da ACP, o primeiro componente representa um escore geral das provas, desta forma quando o valor do primeiro componente é positivo houve um desempenho acima da média geral, e consequentemente, no caso de valor negativo houve um desempenho abaixo da média geral da provas.

Em relação ao segundo componente, também como observado na Seção 3.3 ele pode ser interpretado como um contraste entre a nota da prova de redação e as notas das provas de Matemática e Ciências da Natureza entre os anos de 2013 a 2016, enquanto que nos anos de 2017 a 2022 ele é devido principalmente a nota de redação. Neste caso

Tabela 3.15: Escores médios dos perfis 1 e 2 para os dois primeiros componentes principais de cada ano.

Ano	Perfil 1		Perfil 2	
	CP 1	CP 2	CP 1	CP 2
2013	4.4575	-0.01121	-0.8493	0.51695
2014	4.45169	-0.09798	-0.8776	0.31762
2015	4.28847	-0.12259	-1.02708	0.44215
2016	4.22981	-0.17166	-1.65804	-0.05924
2017	3.75624	-0.01841	-1.81297	-0.22602
2018	3.81792	-0.02055	-2.1008	-0.10287
2019	3.84541	0.00753	-2.21882	-0.11444
2020	3.44742	0.0507	-2.26923	-0.19341
2021	2.82162	0.05519	-2.19444	-0.18195
2022	2.78003	0.02854	-2.16319	-0.20986

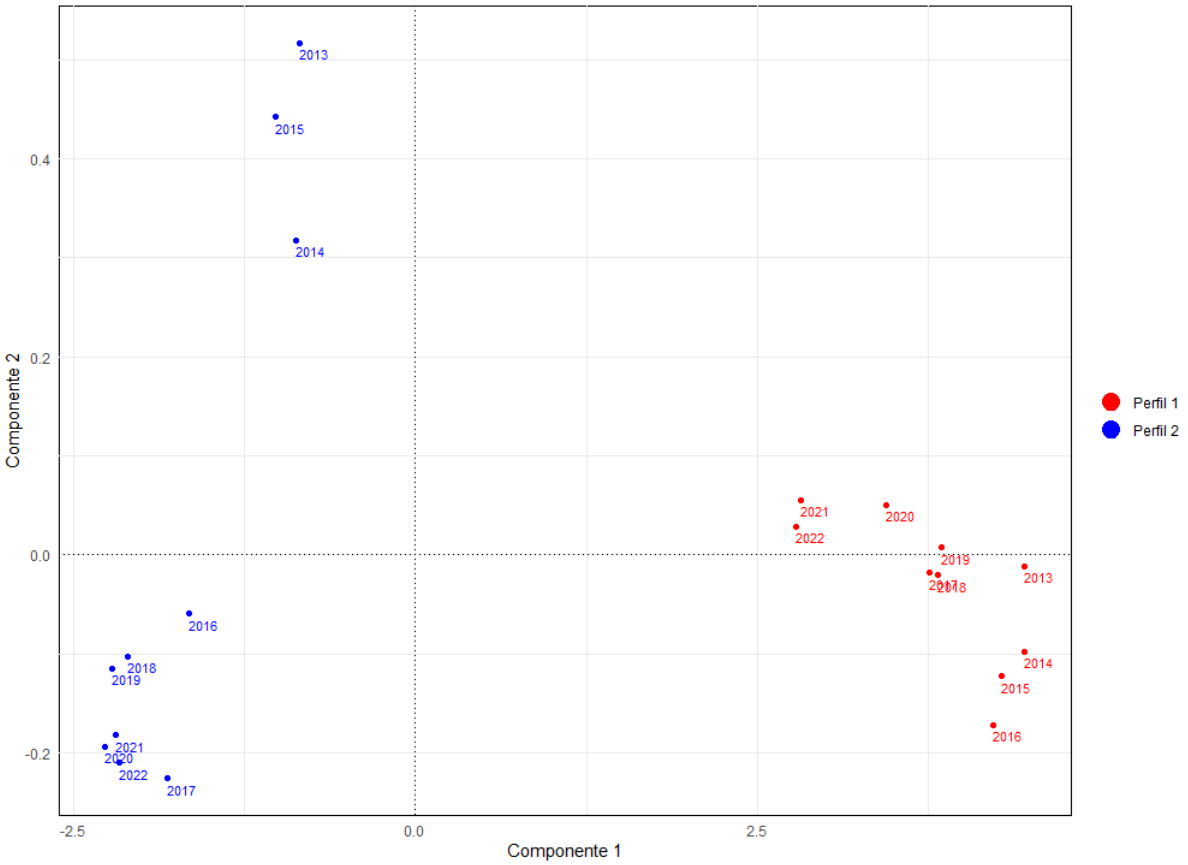

Figura 3.26: Escores médios dos perfis 1 e 2 para as duas primeiras componentes principais a cada ano.

temos que, para todos os anos de uma forma geral, é possível afirmar que valores positivos indicam melhor desempenho na prova de redação e valores negativos melhor desempenho nas demais provas.

Vale lembrar que dado o uso da matriz de correlações no cálculo da ACP, os dados

padronizados foram utilizados e consequentemente todos componentes possuem média zero, desta forma, quanto maior a distância positiva dos componente da origem melhor é o desempenho médio do perfil analisado e, quanto maior a distância negativa da origem menor é o desempenho médio.

Em função da interpretação dos componentes acima apresentada, analisando a Figura 3.26, observa-se que para os indivíduos de perfil 1, todos os anos analisados possuem desempenho nas 5 provas acima da média no componente 1, indicando que estes indivíduos obtiverem nota acima da média ao longo do anos. Ainda que nos últimos 4 anos (2019, 2020, 2021 e 2022) este desempenho acima da média tenha sido menor e, que nos anos de 2014, 2015 e 2016 um melhor desempenho foi observado nas provas de Matemática e Ciências da Natureza para o componente 2.

No caso dos indivíduos do perfil 2 observa-se um desempenho geral (componente 1) abaixo da média em todos os anos com a diferença que no segundo componente, nos anos iniciais (2013 a 2015) foi obtido um melhor resultado na prova de redação quando comparado com os demais anos (2016 a 2022) nas demais provas.

Os resultados acima obtidos demonstram a relação entre aqueles obtidos nas seções 3.2 e 3.3. O grupo de indivíduos com perfil 1, que apresentaram maior participação no total dos inscritos no ENEM ao longo dos anos, são aqueles que também apresentam um desempenho acima da médias na provas em todos os anos. Por outro lado, os indivíduos de perfil 2, que tem proporção de inscritos no ENEM reduzida ao longo dos anos, apresentam desempenho abaixo da média na provas, com melhor desempenho na prova de redação nos anos iniciais em que sua participação era maior, e com melhores resultados nas demais provas nos anos onde sua participação tem maior redução (período de 2016 a 2022). Em relação ao desempenho apresentado nas provas pelos indivíduos do perfil 2 é compreensível ter sido inferior à média, uma vez que esse grupo é constituído, dentre outros fatores, por pessoas que concluíram o ensino médio há mais tempo.

Capítulo 4

Conclusões

A partir dos resultados obtidos no Capítulo 3, foram identificadas as principais características dos inscritos no Enem ao longo da década investigada. Observou-se, que em grande parte das variáveis estudadas, sob o ponto de vista univariado, existe um comportamento padrão ao longo dos dos anos. Porém, esta análise inicial permitiu apontar prováveis fatores que podem estar contribuindo para a redução do número de inscritos no ENEM, identificando as variáveis cujo comportamento sofreu alteração com o decorrer dos anos, característica desejada para a variável ser considerada um possível fator na evasão de inscritos, que foram: tipo de ensino, tempo de formação e grau de escolaridade do pai e da mãe.

Após realizado o estudo das variáveis de maneira univariada, examinando cada uma individualmente, foram analisadas as possíveis relações conjuntas entre as diferentes características dos inscritos, identificando assim, as combinações das categorias de diferentes variáveis responsáveis pelo impacto da redução de inscritos no ENEM, conforme apresentado na Seção 3.2.

A Análise Fatorial de Correspondência Múltipla, permitiu identificar dois perfis de participantes do exame que se relacionam quanto a redução do número de inscritos no ENEM, como apresentado na Tabela 3.14: Perfil 1 composto pelos indivíduos mais novos que acabaram de concluir o ensino médio e possuem pais com grau de escolaridade mais elevado e, o Perfil 2 por indivíduos mais velhos, casados ou divorciados, que concluíram o ensino médio há mais tempo e possuem pais com grau de escolaridade mais baixo. Ainda, foi destacado que de forma lenta a proporção de inscritos no Perfil 1 cresce enquanto a de inscritos no Perfil 2 se reduz. Indicando que candidatos com as características presentes no perfil 2 vêm decrescendo a sua participação no ENEM ao longo dos anos.

Por outro lado, a Análise de Componentes Principais, apresentada na Seção 3.3, identificou para o período de 2013 a 2016, que as notas de Matemática e Ciências da Natureza são bem correlacionadas, identificando um primeiro bloco, assim como as notas de Linguagens e Códigos e de Ciências Humanas, identificando um segundo bloco. Estes dois blocos apresentam também uma correlação moderada entre eles. Ainda, sobre a prova de redação, foi percebido que ela se apresenta isolada das demais com uma moderada correlação com o bloco de Linguagens e Códigos e de Ciências Humanas e uma menor correlação com o bloco de Matemática e Ciências da Natureza. Em relação ao período de 2017 a 2022, foi identificado que as notas referentes as 4 provas, Linguagens e Códigos, Ciências Humanas, Ciências da Natureza e Matemática são bem correlacionadas. Já a prova de redação por sua vez apresenta-se isolada das demais com uma correlação fraca com o bloco das 4 provas mencionas. Portanto, verificou-se que, com a redução do número de inscritos no ENEM e a diminuição da participação de indivíduos com perfil 2, o aproveitamento nas diferentes provas do ENEM, exceto a de Redação, apresentaram uma correlação maior, ou seja, indivíduos com um melhor desempenho em uma das provas, também apresentou melhor desempenho nas demais, valendo o mesmo para o caso de menor aproveitamento.

De forma mais específica, comparando os períodos de 2013 a 2016 e, 2017 a 2022, observou-se uma mudança nos blocos mencionados anteriormente: Linguagens e Códigos e Ciências Humanas, assim como o bloco de Matemática e Ciências da Natureza. Enquanto nos anos de 2013 a 2016 eles se apresentavam separadamente, nos anos de 2017 a 2022, esses blocos se aproximaram, formando um único conjunto com as quatro provas. Isso indica um aumento na correlação entre essas disciplinas ao longo dos anos, ao mesmo tempo em que diminui a correlação das mesmas com a nota da redação. Modificações implementadas na logística de aplicação das provas do ENEM a partir de 2017 (MEC, 2023b) justificam a mudança de correlação entre as provas nos períodos citados.

Na Seção 3.4, relacionando o perfil dos participantes com as notas obtidas nas 5 provas, pode-se dizer que, o grupo de indivíduos do perfil 1, que registrou maior participação no ENEM ao longo da década, também apresentou desempenho acima da média nas provas em todos os anos. Por outro lado, os indivíduos do perfil 2, cuja proporção de inscritos no ENEM diminuiu ao longo do tempo, apresentaram um desempenho abaixo da média nas provas, com melhor desempenho na prova de redação nos primeiros anos em que sua participação era mais expressiva, e nos demais anos de 2016 a 2022, demonstraram

melhores resultados nas demais provas nos anos em que sua participação diminuiu significativamente. Os participantes do perfil 2, por, dentre outros fatores, terem concluído o ensino médio há mais tempo, tiveram um desempenho médio inferior nas provas.

No contexto educacional atual, compreender as mudanças nos perfis dos inscritos no ENEM ao longo do tempo é crucial para diversas razões. Análises apresentadas são essenciais para a formulação e implementação de políticas públicas voltadas para o aprimoramento do sistema de ensino no Brasil. Além disso, essa compreensão permite identificar possíveis desafios enfrentados por diferentes grupos de indivíduos, possibilitando ações direcionadas para melhorar a qualidade e equidade da educação no país.

Nesse sentido, o presente estudo buscou não apenas caracterizar os diferentes perfis de participantes, mas também inspirar o desenvolvimento de pesquisas futuras cujo interesse seja investigar mais a fundo os motivos por trás das diferenças de desempenho entre os perfis identificados, buscando possíveis fatores que influenciam tal discrepância, o que destaca a relevância dos resultados obtidos.

Finalmente, não é possível deixar de destacar que os resultados aqui obtidos estão relacionados com a realidade política, econômica, social e principalmente educacional vivida nos últimos anos no país. Em particular, a partir de 2016, quando se inicia uma maior redução do número de inscritos no ENEM, as políticas governamentais, pouco incentivavam os alunos a prosseguirem seus estudos no ensino superior (dos Reis Silva *et al.*, 2019), consolidado pela redução de verbas para manutenção de indivíduos de baixa renda nas instituições públicas e pela redução de financiamentos para ingresso nas instituições particulares. Este fato fica mais evidente no fato que em 2023, com as mudanças significativas nas políticas para a educação, incluindo-se, aquelas relacionadas ao ensino superior, o número de inscritos no ENEM voltou a crescer, com a expectativa que isto continue a ocorrer nos próximos anos. Um último ponto a ser destacado, e que deve ser melhor observado nos próximos anos, é o impacto do período de ensino remoto em função da COVID-19, particularmente com respeito ao aproveitamento nas provas do ENEM.

Referências Bibliográficas

Alteração da Lei de Cotas (2016). Lei nº 13.409. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/Lei/L13409.htm. Acesso em: 03 de junho 2023.

Andrade, D. F., Tavares, H. R. e da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

Balanço (2022). Divulgado balanço dos resultados do enem 2021. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/enem/divulgado-balanco-dos-resultados-do-enem-2021. Acesso em: 03 de junho 2023.

Decicino, D. V. (2012). *Procedimentos Gráficos na Análise de Tabelas de Contingência*. Trabalho de Conclusão de Curso- DES UFSCar.

dos Reis Silva, R. H., Machado, R. e da Silva, R. N. (2019). Golpe de 2016 e a educação no brasil: implicações nas políticas de educação especial na perspectiva da educação inclusiva. *Revista HISTEDBR On-line*.

Estado de Minas (2022). Enem 2022: o que explica a fuga de alunos e o esvaziamento do exame? Disponível em: https://www.em.com.br/app/noticia/gerais/2022/12/04/interna_gerais,1429214/enem-2022-o-que-explica-a-fuga-de-alunos-e-o-esvaziamento-do-exame.shtml. Acesso em: 03 de junho 2023.

Fávero, L. P. L., Belfiore, P. P., Silva, F. L. d. e Chan, B. L. (2009). Análise de dados: modelagem multivariada para tomada de decisões.

Ferreira Filho, P., Bereta, E. M. P. e Ribeiro, F. B. (1998). Tabela de burt. Relatório Técnico 04 - Série C - Notas Didáticas, DEs-UFSCar, São Carlos, SP.

G1 (2023). Ceará tem 18 municípios entre os 20 com maiores notas do brasil em índice sobre educação. Acesso em: 03 de junho 2023.

Gaspar, L. e Barbosa, V. (2013). Ações afirmativas e políticas de cotas no brasil: Uma bibliografia 1999–2012. *Ministério da Educação e Cultura,(Fundação Joaquim Nabuco), Recife*.

Inep (2023). Exame nacional do ensino médio (enem). Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem. Acesso em: 03 de junho 2023.

Inep (2023). Ministério da educação. Acesso em: 03 de junho 2023.

Johnson, RA e Wichern, D. (2008). *Análise estatística multivariada aplicada (Vol. 6º)*. Pearson.

Lei de Cotas (2012). Decreto nº 7.824, de 11 de outubro de 2012. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/decreto/d7824.htm. Acesso em: 03 de junho 2023.

MEC (2023a). Exame não será mais utilizado para certificar o ensino médio. Acesso em: 03 de junho 2023.

MEC (2023b). Enem passa a ser realizado em dois domingos seguidos. Acesso em: 03 de junho 2023.

Medida Provisória (1997). Medida provisÓria nº 1.568. Disponível em: https://legislacao.presidencia.gov.br/atos/?tipo=MPV&numero=1568&ano=1997&ato=e73QTUq5kMJpWT413. Acesso em: 03 de junho 2023.

Microdados (2023). Enem. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 03 de junho 2023.

Mingoti, S. A. (2005). *Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada*. Editora UFMG.

Morettin, P. A. e Singer, J. d. M. (2022). Estatística e ciência de dados.

Neto, R. d. D. M., Medeiros, H. A. V., da Silva Paiva, F. e Simões, J. L. (2014). O impacto do enem nas políticas de democratização do acesso ao ensino superior brasileiro. *Comunicações*, **21**(3), 109–123.

Saviani, D. (2012). O inep, o diagnóstico da educação brasileira e a rbep. *Revista brasileira de estudos pedagógicos*, **93**(234), 291–322.

Apêndice A

Dicionário

Tabela A.1: Dicionário dos dados utilizados.

Variável	**Categorias**
Faixa etária	Menor de 18 anos; De 18 a 30 anos; De 31 a 60 anos; Maior de 60 anos.
Sexo	Masculino; Feminino.
Estado Civil	Solteiro(a); Casado(a)/Mora com um(a) companheiro(a); Divorciado(a)/Desquitado(a)/ Separado(a); Viúvo(a).
Raça	Branca; Preta; Parda; Amarela; Indígena.

Continua na próxima página

Tabela A.1 – *Continuação da tabela*

Variável	**Categorias**
Situação de conclusão do Ensino Médio	Já concluí o Ensino Médio; Estou cursando e concluirei o Ensino Médio no ano atual; Estou cursando e concluirei o Ensino Médio após este ano; Não concluí e não estou cursando o Ensino Médio.
Ano de Conclusão do Ensino Médio	Ano em que o ensino médio foi concluído.
Há quanto tempo se formou no ensino médio	Ano atual; Há um ano; Há dois anos; Há 3 anos; Há 4 ano; Há 5 anos ou mais.
Tipo de escola do Ensino Médio	Pública; Privada.
Tipo de instituição que concluiu ou concluirá o Ensino Médio	Ensino Regular; Educação Especial - Modalidade Substitutiva.
Nome do município da escola	Nome específico do município em que a escola que o indivíduo concluiu o ensino médio está situada.
Sigla da Unidade da Federação da escola	Abreviação que identifica a unidade da federação na qual a escola que o indivíduo concluiu o ensino médio está localizada.

Continua na próxima página

Tabela A.1 – *Continuação da tabela*

Variável	**Categorias**
Localização (Escola)	Urbana; Rural.
Presença na prova objetiva de Ciências da Natureza	Faltou à prova; Presente na prova; Eliminado na prova.
Presença na prova objetiva de Ciências Humanas	Faltou à prova; Presente na prova; Eliminado na prova.
Presença na prova objetiva de Linguagens e Códigos	Faltou à prova; Presente na prova; Eliminado na prova.
Presença na prova objetiva de Matemática	Faltou à prova; Presente na prova; Eliminado na prova.
Nota da prova de Ciências da Natureza	Pontuação de 0 a 1000.
Nota da prova de Ciências Humanas	Pontuação de 0 a 1000.
Nota da prova de Linguagens e Códigos	Pontuação de 0 a 1000.
Nota da prova de Matemática	Pontuação de 0 a 1000.
Nota da prova de redação	Pontuação de 0 a 1000.
Presença no primeiro dia de prova	Não compareceu ao 1º dia; Compareceu ao 1º dia.
Presença no segundo dia de prova	Não compareceu ao 2º dia; Compareceu ao 2º dia.
Presença em ambos os dias	Não compareceu em ambos os dias; Compareceu em ambos os dias.

Continua na próxima página

Tabela A.1 – *Continuação da tabela*

Variável	**Categorias**
Grau de escolaridade do pai	Nunca estudou; Fundamental 1 completo; Fundamental 2 completo; Ensino Médio completo; Ensino Superior; Pós-graduação.
Grau de escolaridade da mãe	Nunca estudou; Fundamental 1 completo; Fundamental 2 completo; Ensino Médio completo; Ensino Superior; Pós-graduação.
Possui computador em sua residência	Sim; Não.
Possui internet em sua residência	Sim; Não.

Fim da tabela

Apêndice B

Figuras

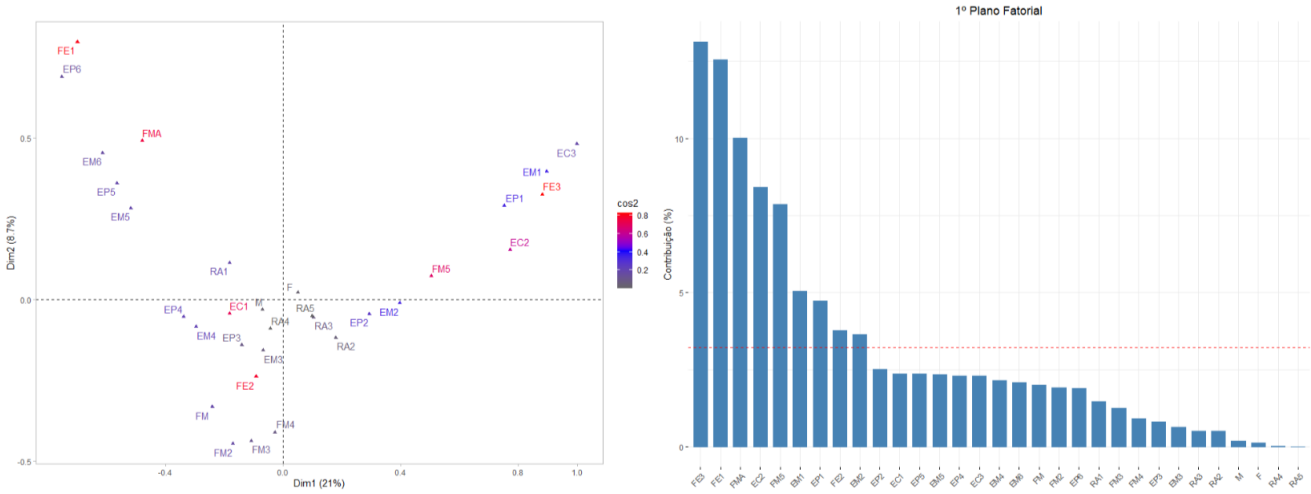

Figura B.1: AFCM do perfil dos inscritos no ano de 2014.

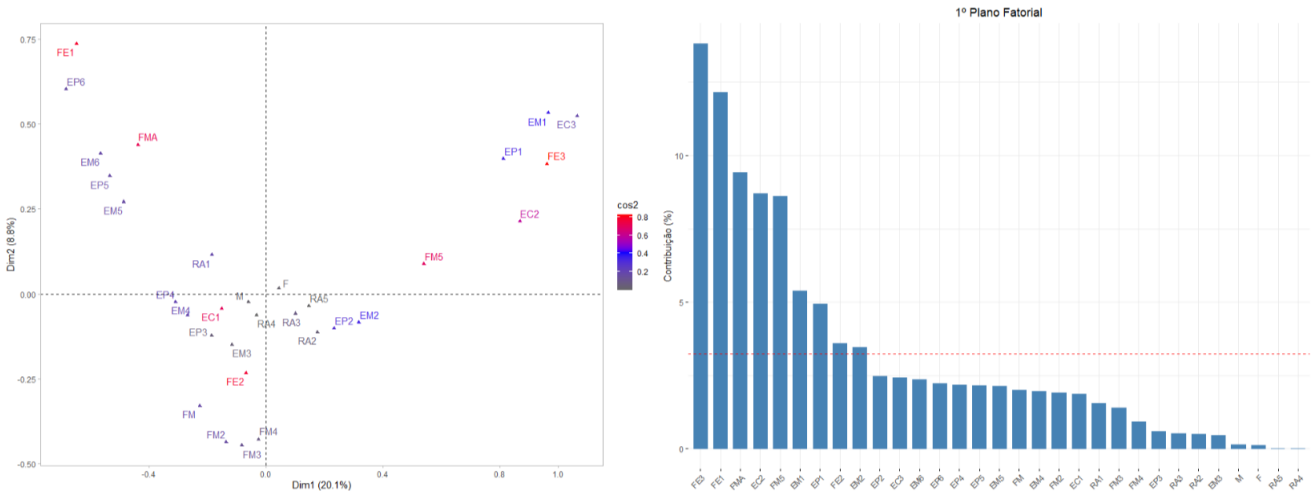

Figura B.2: AFCM do perfil dos inscritos no ano de 2015.

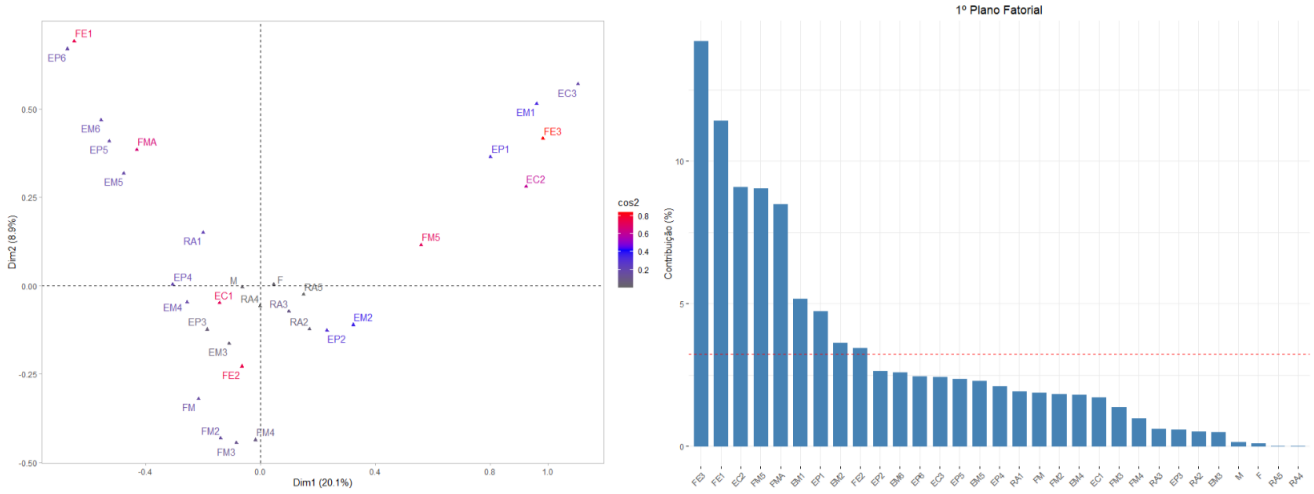

Figura B.3: AFCM do perfil dos inscritos no ano de 2017.

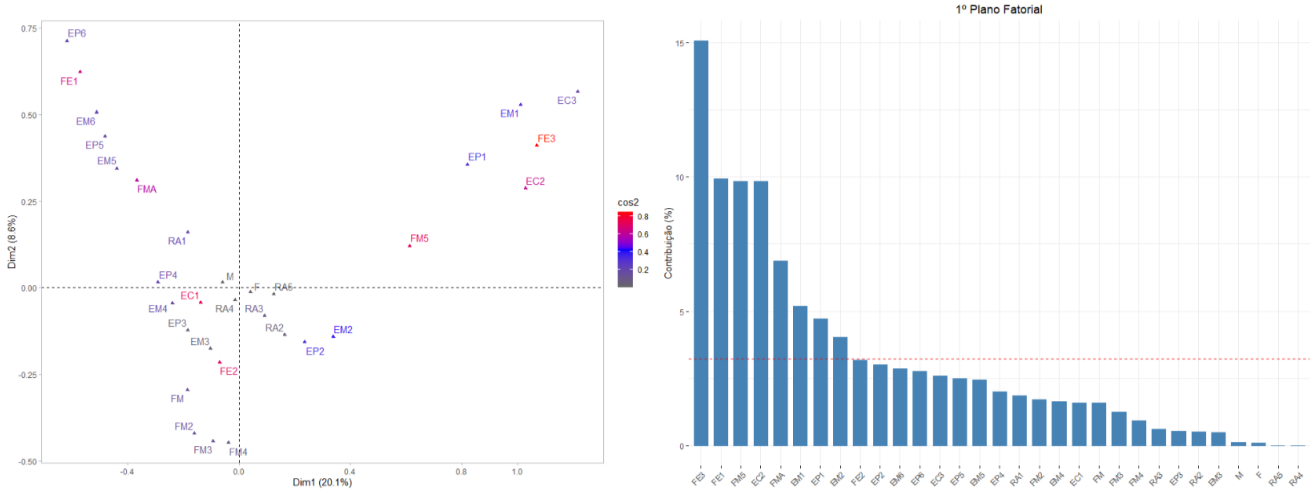

Figura B.4: AFCM do perfil dos inscritos no ano de 2018.

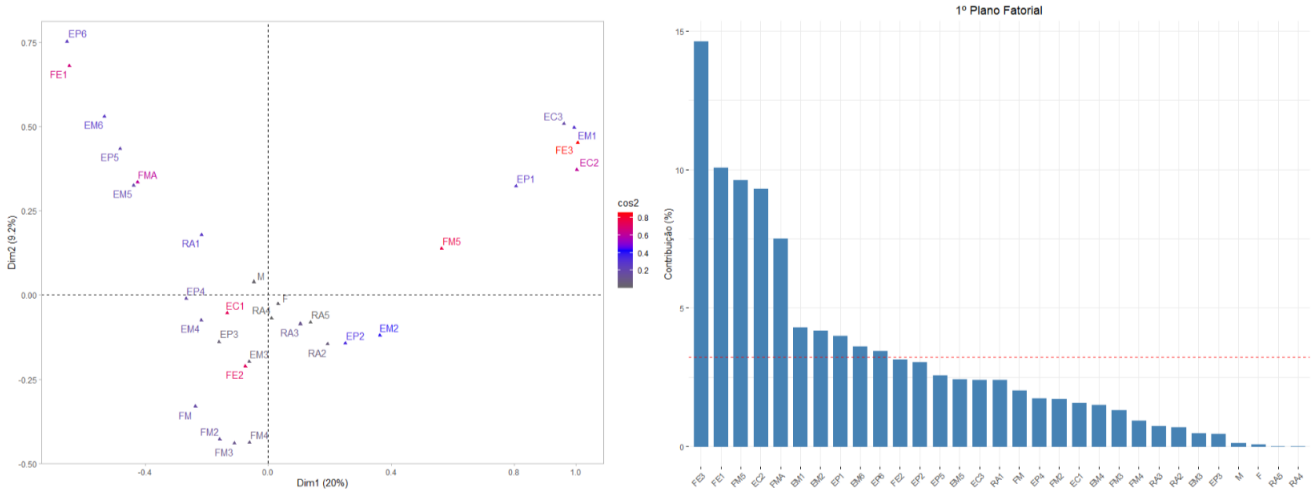

Figura B.5: AFCM do perfil dos inscritos no ano de 2020.

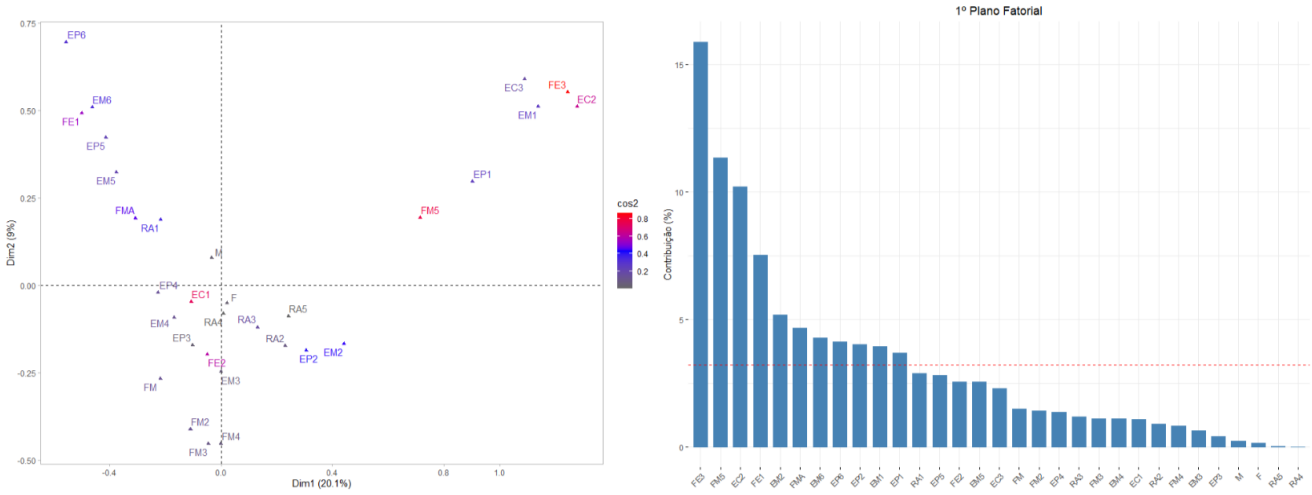

Figura B.6: AFCM do perfil dos inscritos no ano de 2021.

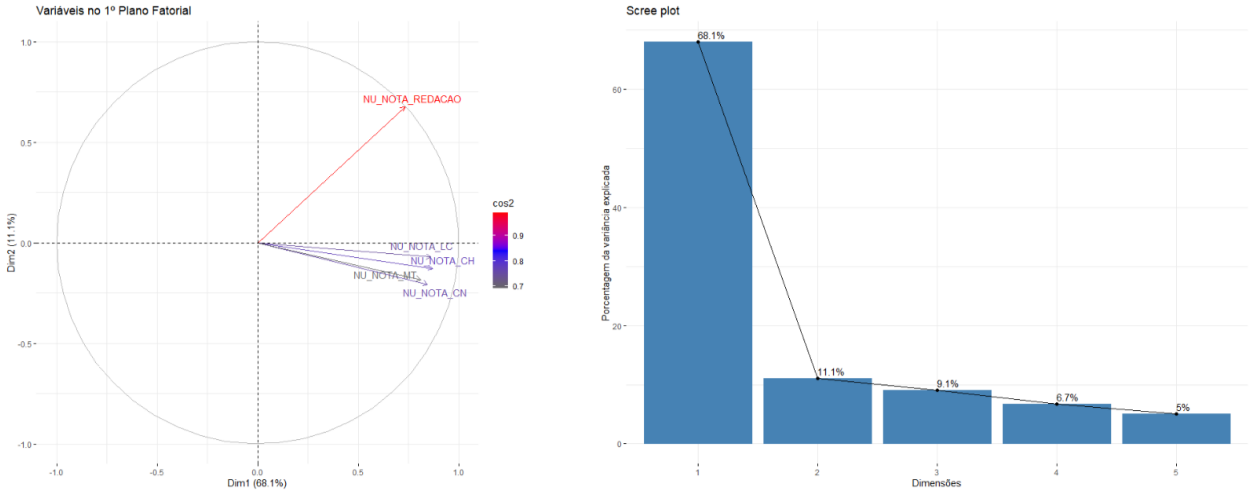

Figura B.7: ACP das notas dos inscritos no ano de 2019.

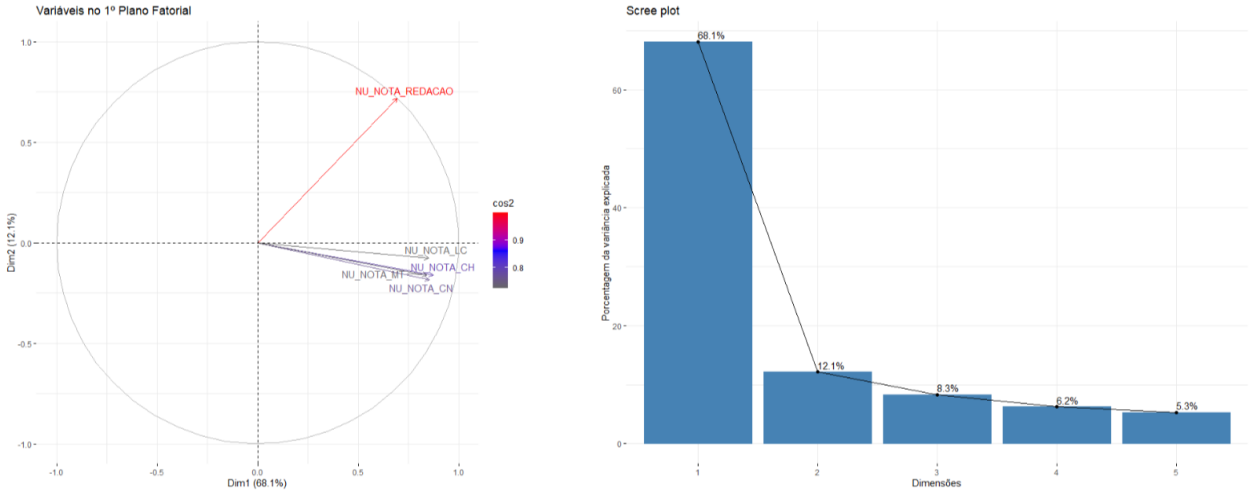

Figura B.8: ACP das notas dos inscritos no ano de 2020.

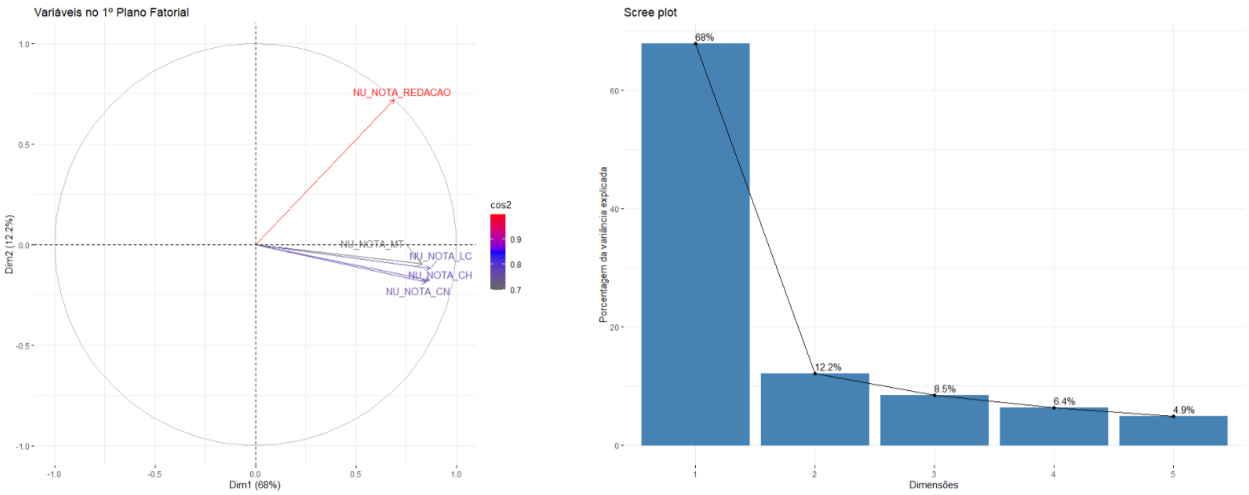

Figura B.9: ACP das notas dos inscritos no ano de 2021.

Apêndice C

Códigos utilizados

```
#########################
#### Gráficos ADED ####

library(tidyr)
library(ggplot2)
library(scales)
library(plyr)
library(vcd)
library(mosaic)
library(ggmosaic)

tab_evasao <- read.csv("tab_evasao.csv")

#Combinando as colunas Inscritos e Presença em uma única coluna chamada'Variável'
tab_evasao_teste <- tidyr::gather(tab_evasao, key = "Variável", value = "Valor",
                                  -Ano,-Ausentes,-Per_pres,-Per_aus)

ggplot(tab_evasao_teste, aes(x=as.factor(Ano),y = Valor, fill = Variável)) +
  geom_col(position = "dodge", color = 'black') +
  labs(#title = "Inscritos x Presença",
       x = "Ano",
       y = 'Frequência') +
  scale_fill_manual(values = c("blue", "gray")) +
  theme_gray()

ggplot(tab_evasao_teste, aes(x = as.factor(Ano), y = Valor, color = Variável,
                             group = Variável)) +
  geom_line(linewidth = 1) +
  labs(#title = "Inscritos x Presença",
```

```
       x = "Ano",
       y = "Frequência") +
  scale_color_manual(values = c("blue", "darkgreen")) +
  theme_gray()


ggplot(tab_evasao, aes(x=as.factor(Ano),y = Ausentes)) +
  geom_bar(stat = "identity", col = 'black', fill = "blue") +
  labs(#title = "Evasão de inscritos",
       x = "Ano",
       y = 'Evasão') +
  #scale_y_continuous(labels = percent, limits = c(0, 1)) +
  theme_gray()

############ box plot Linguagens e Códigos ###################

tab_notas_LC <- read.csv("tab_notas_LC.csv")

dados_long_LC <- tidyr::gather(tab_notas_LC, key = "Coluna", value = "Valor")

# Criar o boxplot com base nos dados no formato "long"
ggplot(data = dados_long_LC, aes(x = Coluna, y = Valor)) +
  geom_boxplot(fill = 'darkgreen') +
  labs(x = "Ano", y = "Notas") +
  scale_y_continuous(limits=c(0,1000)) +
  scale_x_discrete(labels = c("2013","2014","2015","2016","2017",
                              "2018","2019","2020","2021","2022")) #+
  #ggtitle("Linguagens e Códigos")

############ box plot Ciências Humanas ###################

tab_notas_CH <- read.csv("tab_notas_CH.csv")

dados_long_CH <- tidyr::gather(tab_notas_CH, key = "Coluna", value = "Valor")

# Criar o boxplot com base nos dados no formato "long"
ggplot(data = dados_long_CH, aes(x = Coluna, y = Valor)) +
  geom_boxplot(fill = 'darkgreen') +
  labs(x = "Ano", y = "Notas") +
  scale_y_continuous(limits=c(0,1000)) +
  scale_x_discrete(labels = c("2013","2014","2015","2016","2017",
```

```
                                "2018","2019","2020","2021","2022")) #+
  #ggtitle("Ciências Humanas")

############ box plot Redação ###################

tab_notas_REDACAO <- read.csv("tab_notas_REDACAO.csv")

dados_long_REDACAO <- tidyr::gather(tab_notas_REDACAO, key = "Coluna",
                                    value = "Valor")

# Criar o boxplot com base nos dados no formato "long"
ggplot(data = dados_long_REDACAO, aes(x = Coluna, y = Valor)) +
  geom_boxplot(fill = 'darkgreen') +
  labs(x = "Ano", y = "Notas") +
  scale_y_continuous(limits=c(0,1000)) +
  scale_x_discrete(labels = c("2013","2014","2015","2016","2017",
                              "2018","2019","2020","2021","2022")) #+
  #ggtitle("Redação")

############ box plot Ciências da Natureza ###################

tab_notas_CN <- read.csv("tab_notas_CN.csv")

dados_long_CN <- tidyr::gather(tab_notas_CN, key = "Coluna", value = "Valor")

# Criar o boxplot com base nos dados no formato "long"
ggplot(data = dados_long_CN, aes(x = Coluna, y = Valor)) +
  geom_boxplot(fill = 'darkgreen') +
  labs(x = "Ano", y = "Notas") +
  scale_y_continuous(limits=c(0,1000)) +
  scale_x_discrete(labels = c("2013","2014","2015","2016","2017",
                              "2018","2019","2020","2021","2022")) #+
  #ggtitle("Ciências da Natureza")

############ box plot Matemática ###################

tab_notas_MT <- read.csv("tab_notas_MT.csv")

dados_long_MT <- tidyr::gather(tab_notas_MT, key = "Coluna", value = "Valor")

# Criar o boxplot com base nos dados no formato "long"
```

```
ggplot(data = dados_long_MT, aes(x = Coluna, y = Valor)) +
  geom_boxplot(fill = 'darkgreen') +
  labs(x = "Ano", y = "Notas") +
  scale_y_continuous(limits=c(0,1000)) +
  scale_x_discrete(labels = c("2013","2014","2015","2016","2017",
                              "2018","2019","2020","2021","2022")) #+
  #ggtitle("Matemática")

############## GRÁFICOS VARIÁVEIS DE EFEITO ###############
############################################################

##### Tipo de ensino #####

tab_ensino <- read.csv("tab_ensino.csv")
# Combinando as colunas Regular e Especial em uma única coluna chamada'Variável'
tab_ensino_teste <- tidyr::gather(tab_ensino, key = "Variável", value = "Valor",
                                  -Ano,-per_ER,-per_EE,-Total)

ggplot(tab_ensino_teste, aes(x=as.factor(Ano),y = Valor, fill = Variável)) +
  geom_col(position = "dodge", color = 'black') +
  labs(#title = "Regular x Especial ",
    x = "Ano",
    y = 'Frequência') +
  scale_fill_manual(values = c("blue", "gray"),
                    labels = c("Regular", "Especial")) +
  theme_gray()

ggplot(tab_ensino_teste, aes(x = as.factor(Ano), y = Valor, fill = Variável)) +
  geom_bar(stat = "identity", position = "stack",color = 'black') +
  labs(x = "Ano", y = "Frequência", fill = "Tipo de Ensino") +
  scale_fill_manual(values = c("blue", "gray"),
                    labels = c("Regular", "Especial")) +
  theme_minimal()

# Calculando as proporções para cada ano
tab_ensino_perc <- ddply(tab_ensino_teste, .(Ano), transform,
                         Porcentagem = Valor / sum(Valor) * 100)

# Criação do gráfico de barras empilhadas com position = "fill"
ggplot(tab_ensino_perc, aes(x = as.factor(Ano), y = Porcentagem,
                            fill = Variável)) +
```

```
  geom_bar(stat = "identity", position = "fill",color = 'black') +
  labs(x = "Ano", y = "Porcentagem (%)", fill = "Tipo de Ensino") +
  scale_fill_manual(values = c("blue", "gray"),
                    labels = c("Regular", "Especial")) +
  theme_minimal()



# creating a random dataset
data_ensino <- matrix(c(tab_ensino$Ensino.Regular,
                        tab_ensino$Educação.Especial...Modalidade.Substitutiva),
                      nrow= 10, ncol = 2, byrow= FALSE)

# creating dataset with above values
mosaic_ensino <- as.table(matrix(data_ensino,nrow = 10,
                        byrow = FALSE,
                        dimnames = list(
      Ano = c('2013','2014','2015', '2016', '2017',
                      '2018','2019','2020','2021','2022'),
      Ensino = c('Regular', 'Especial'))))
## mosaico tipo de ensino
mosaicplot(mosaic_ensino,
           main = '',
           xlab = "Ano",
           ylab = "Tipo de ensino",
           las = 1,
           color =  colorRampPalette(c('gray','blue'))(2),
           border = "Black")
##############################
##### Tempo de formação #####
tab_formacao <- read.csv("tab_formacao.csv")
# creating a random dataset
data_formacao <- matrix(c(tab_formacao$Formação.no.ano.atual,
                          tab_formacao$Formação.há.1.ano,
                          tab_formacao$Formação.há.2.anos,
                          tab_formacao$Formação.há.3.anos,
                          tab_formacao$Formação.há.4.anos,
                          tab_formacao$Formação.há.mais.de.5.anos),
                        nrow= 10, ncol = 6, byrow= FALSE)

# creating dataset with above values
```

```
mosaic_formacao <- as.table(matrix(data_formacao,nrow = 10,
                                   byrow = FALSE,
                                   dimnames = list(
                                     Ano = c('2013','2014','2015', '2016', '2017',
                                             '2018','2019','2020','2021','2022'),
                                     Formação = c('Ano atual', 'Há 1 ano',
                                                  'Há 2 anos','Há 3 anos',
                                                  'Há 4 anos',
                                                  'Há mais de 5 anos'))))
## mosaico
mosaicplot(mosaic_formacao,
           main = '',
           xlab = "Ano",
           ylab = "Tempo de formação",
           las = 1,
           color =  colorRampPalette(c('gray','blue'))(6),
           border = "Black")
##############################
##### Questão Q1 #####
tab_q1 <- read.csv("tab_q1.csv")
# creating a random dataset
data_q1 <- matrix(c(tab_q1$Nunca.estudou,tab_q1$Fundamental.1.completo,
                    tab_q1$Fundamental.2.completo,tab_q1$Ensino.Médio.completo,
                    tab_q1$Ensino.Superior,
                    tab_q1$Pós.graduação),
                        nrow= 10, ncol = 6, byrow= FALSE)

# creating dataset with above values
mosaic_q1 <- as.table(matrix(data_q1,nrow = 10,
                                   byrow = FALSE,
                                   dimnames = list(
                                     Ano = c('2013','2014','2015','2016','2017',
                                             '2018','2019','2020',
                                             '2021','2022'),
                                     Grau_de_Escolaridade =
                                       c('Nunca estudou', 'Fundamental 1',
                                         'Fundamental 1', 'Ensino Médio',
                                         'Ensino Superior', 'Pós Graduação'))))
## mosaico
mosaicplot(mosaic_q1,
           main = '',
```

```
            xlab = "Ano",
            ylab = "Grau de Escolaridade do pai",
            las = 1,
            color =  colorRampPalette(c('gray','blue'))(6),
            border = "Black")
##############################
##### Questão q2 #####
tab_q2 <- read.csv("tab_q2.csv")
# creating a random dataset
data_q2 <- matrix(c(tab_q2$Nunca.estudou,tab_q2$Fundamental.1.completo,
                    tab_q2$Fundamental.2.completo,
                    tab_q2$Ensino.Médio.completo,tab_q2$Ensino.Superior,
                    tab_q2$Pós.graduação),
                        nrow= 10, ncol = 6, byrow= FALSE)

# creating dataset with above values
mosaic_q2 <- as.table(matrix(data_q2,nrow = 10,
                                    byrow = FALSE,
                                    dimnames = list(
                                      Ano = c('2013','2014','2015','2016','2017',
                                              '2018','2019','2020',
                                              '2021','2022'),
                                      Grau_de_Escolaridade =
                                        c('Nunca estudou', 'Fundamental 1',
                                          'Fundamental 1', 'Ensino Médio',
                                          'Ensino Superior', 'Pós Graduação'))))
## mosaico
mosaicplot(mosaic_q2,
            main = '',
            xlab = "Ano",
            ylab = "Grau de Escolaridade da mãe",
            las = 1,
            color =  colorRampPalette(c('gray','blue'))(6),
            border = "Black")
##############################
##### Tabela região escola #####
tab_regiao <- read.csv("tab_regiao.csv")
# creating a random dataset
data_regiao <- matrix(c(tab_regiao$NORTE,tab_regiao$NORDESTE,
                        tab_regiao$CENTRO_OESTE,tab_regiao$SUDESTE,
                        tab_regiao$SUL),
```

```
                  nrow= 10, ncol = 5, byrow= FALSE)

# creating dataset with above values
mosaic_regiao <- as.table(matrix(data_regiao,nrow = 10,
                                 byrow = FALSE,
                                 dimnames = list(
                                   Ano = c('2013','2014','2015', '2016', '2017',
                                           '2018','2019','2020','2021','2022'),
                                   Regiao =
                                     c('Norte', 'Nordeste',
                                       'Centro Oeste', 'Sudeste','Sul'))))
## mosaico
mosaicplot(mosaic_regiao,
           main = '',
           xlab = "Ano",
           ylab = "Região",
           las = 1,
           color =  colorRampPalette(c('gray','blue'))(6),
           border = "Black")
##############################
##### Tabela região nordeste #####
tab_nordeste <- read.csv("tab_nordeste.csv")
# creating a random dataset
data_nordeste <- matrix(c(tab_nordeste$AL,tab_nordeste$BA,tab_nordeste$CE,
                          tab_nordeste$MA,tab_nordeste$PB,tab_nordeste$PE,
                          tab_nordeste$PI,tab_nordeste$RN,tab_nordeste$SE),
                        nrow= 10, ncol = 9, byrow= FALSE)

# creating dataset with above values
mosaic_nordeste <- as.table(matrix(data_nordeste,nrow = 10,
                                   byrow = FALSE,
                                   dimnames = list(
                                     Ano = c('2013','2014','2015', '2016', '2017',
                                             '2018','2019','2020','2021','2022'),
                                     Estados =
                                       c('AL', 'BA','CE','MA','PB','PE','PI',
                                         'RN', 'SE'))))
## mosaico
mosaicplot(mosaic_nordeste,
           main = '',
           xlab = "Ano",
```

```
            ylab = "Estados",
            las = 1,
            color =  colorRampPalette(c('gray','blue'))(9),
            border = "Black")
##################################
##### Tabela região sudeste #####
tab_sudeste <- read.csv("tab_sudeste.csv")
# creating a random dataset
data_sudeste <- matrix(c(tab_sudeste$ES,tab_sudeste$MG,
                         tab_sudeste$RJ,tab_sudeste$SP),
                       nrow= 10, ncol = 4, byrow= FALSE)

# creating dataset with above values
mosaic_sudeste <- as.table(matrix(data_sudeste,nrow = 10,
                                  byrow = FALSE,
                                  dimnames = list(
                                    Ano = c('2013','2014','2015','2016','2017',
                                            '2018','2019','2020',
                                            '2021','2022'),
                                    Estados =
                                      c('ES', 'MG','RJ','SP'))))
## mosaico
mosaicplot(mosaic_sudeste,
            main = '',
            xlab = "Ano",
            ylab = "Estados",
            las = 1,
            color =  colorRampPalette(c('gray','blue'))(5),
            border = "Black")
##################################
##### Tabela sexo #####
tab_sexo <- read.csv("tab_sexo.csv")
# creating a random dataset
data_sexo <- matrix(c(tab_sexo$Masculino,tab_sexo$Feminino),
                    nrow= 10, ncol = 2, byrow= FALSE)

# creating dataset with above values
mosaic_sexo <- as.table(matrix(data_sexo,nrow = 10,
                                byrow = FALSE,
                                dimnames = list(
                                  Ano = c('2013','2014','2015', '2016','2017',
```

```
                                                '2018','2019','2020','2021','2022'),
                                      Sexo =
                                        c('Masculino','Feminino'))))
## mosaico
mosaicplot(mosaic_sexo,
           main = '',
           xlab = "Ano",
           ylab = "Gênero",
           las = 1,
           color =  colorRampPalette(c('gray','blue'))(2),
           border = "Black")

#################################
#### AFCM ########################
library(FactoMineR)
library(factoextra)
library (corrplot)

quali_2020 <- na.omit(quali_2020[,-c(1,9:12)])

acm_2020<- MCA(quali_2020, method="Burt", graph = TRUE)


fviz_mca_var(acm_2020,repel = TRUE ,labelsize=4 ,col.var = "cos2",
             gradient.cols=c("grey40","blue","red"))+ labs(title = "")+
  theme_light()+
  theme(panel.grid = element_blank())+
  scale_x_continuous(breaks = c(-0.4,0,0.4,0.8,1.0))


#primeiro plano
fviz_contrib(acm_2020, choice = "var",axes = 1:2 )+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "1º Plano Fatorial", y="Contribuição (%)")

## Scree plot ##

fviz_screeplot(acm_2020, addlabels = TRUE, ylim = c(0, 95),
               geom = c("bar", "line"),
               barfill = "cadetblue") +
  theme_bw() +
```

```
  labs(y = "Porcentagem da variância explicada", x = "Dimensões",
       title = "Gráfico de Cotovelo") +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 15, family ="serif"))

# Scree Plot
fviz_eig(acm_2020, addlabels=TRUE, xlab = 'Dimensões',
         ylab = 'Porcentagem da variância explicada')

acm_2020$eig
acm_2020$var

##################################
#### PCA ########################

acp_2020=PCA(quantt_2020, scale.unit=TRUE, graph=TRUE, quali.sup=NULL)

# Scree Plot
fviz_eig(acp_2020, addlabels=TRUE, xlab = 'Dimensões',
         ylab = 'Porcentagem da variância explicada')

fviz_pca_var(acp_2020, col.var="cos2", gradient.cols= c("grey40","blue","red"),
             repel= TRUE, title= "Variáveis no 1º Plano Fatorial")
# Contribuição das Variáveis no 1o CP
fviz_contrib(acp_2020, choice="var", axes=1:2)+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "1º Plano Fatorial", y="Contribuição (%)")



# Contribuição das Variáveis no 2o CP
fviz_contrib(acp_2020, choice="var", axes=2)

# Contribuição das Variáveis para os CP

corrplot(acp_2020$var$cos2, is.corr=FALSE, col = 'darkblue')
fviz_cos2(acp_2020, choice = "var", axes = 2:3, title = "Grafico para Cos2")

fviz_screeplot(acp_2020, addlabels = TRUE, ylim = c(0, 95),
               geom = c("bar", "line"),
```

```
                barfill = "cadetblue") +
  theme_bw() +
  labs(y = "Porcentagem da variância explicada", x = "Dimensões",
       title = "Gráfico de Cotovelo") +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 15, family ="serif"))

# Autovalores

acp_2020$eig

# Autovetores
acp_2020$var

########################################
# SCORES POR PERFIL

### proporção perfil 1
p01 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE1' & dados_2013$TP_FORMACAO == 'FMA'
            & dados_2013$Q001 == 'EP6' & dados_2013$Q002 == 'EM6'))/nrow(dados_2013)
p02 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE1' & dados_2013$TP_FORMACAO == 'FMA'
           & dados_2013$Q001 == 'EP5' & dados_2013$Q002 == 'EM6'))/nrow(dados_2013)
p03 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE1' & dados_2013$TP_FORMACAO == 'FMA'
           & dados_2013$Q001 == 'EP6' & dados_2013$Q002 == 'EM5'))/nrow(dados_2013)
p04 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE1' & dados_2013$TP_FORMACAO == 'FMA'
           & dados_2013$Q001 == 'EP5' & dados_2013$Q002 == 'EM5'))/nrow(dados_2013)
prop01<- p01 + p02 + p03 + p04
round(prop01, 4)
### proporção perfil 2
p11 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE3' & dados_2013$TP_FORMACAO == 'FM5'
            & dados_2013$TP_ESTADO_CIVIL == 'EC2'
           & dados_2013$Q001 == 'EP1' & dados_2013$Q002 == 'EM1'))/nrow(dados_2013)
p12 <- (sum(dados_2013$TP_FAIXA_ETARIA == 'FE3' & dados_2013$TP_FORMACAO == 'FM5'
            & dados_2013$TP_ESTADO_CIVIL == 'EC3'
              & dados_2013$Q001 == 'EP1' & dados_2013$Q002 == 'EM1'))/nrow(dados_2013)
prop11<- p11 + p12
round(prop11, 4)

## calculando scores dos componentes 1 e 2 ##
```

```
############################
############ 2013

# padronizar
quantt_2013 <- quantt_2013[,-1]
quantt_2013$NU_NOTA_CN <- scale(quantt_2013$NU_NOTA_CN, center = TRUE,
                                scale = TRUE)
quantt_2013$NU_NOTA_CH <- scale(quantt_2013$NU_NOTA_CH, center = TRUE,
                                scale = TRUE)
quantt_2013$NU_NOTA_LC <- scale(quantt_2013$NU_NOTA_LC, center = TRUE,
                                scale = TRUE)
quantt_2013$NU_NOTA_MT <- scale(quantt_2013$NU_NOTA_MT, center = TRUE,
                                scale = TRUE)
quantt_2013$NU_NOTA_REDACAO <- scale(quantt_2013$NU_NOTA_REDACAO, center = TRUE,
                                     scale = TRUE)
quali_2013 <- quali_2013[,-1]

library(dplyr)

# Junte os dois data frames
dados_2013 <- cbind(quali_2013, quantt_2013)

# localizando as pessoas do perfil 1

subconjunto1<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE1' &
                        TP_FORMACAO == 'FMA' &
                        Q001 == 'EP6' & Q002 == 'EM6')

subconjunto2<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE1' &
                        TP_FORMACAO == 'FMA' &
                        Q001 == 'EP5' & Q002 == 'EM6')

subconjunto3<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE1' &
                        TP_FORMACAO == 'FMA' &
                        Q001 == 'EP6' & Q002 == 'EM5')

subconjunto4<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE1' &
                        TP_FORMACAO == 'FMA' &
                        Q001 == 'EP5' & Q002 == 'EM5')
```

```
matriz_perfil1 <- rbind(as.matrix(na.omit(subconjunto1[,c(12:16)])),
                        as.matrix(na.omit(subconjunto2[,c(12:16)])),
                        as.matrix(na.omit(subconjunto3[,c(12:16)])),
                        as.matrix(na.omit(subconjunto4[,c(12:16)])))

autov_a<- matrix(c(0.8285421,0.8702085,0.8592757,0.7949059,0.6869268),
                 nrow = 5,ncol = 1)

autov_b<- matrix(c(-0.26421111,-0.04585635,0.03896060,-0.31179342,
                   0.68884071),
                 nrow = 5,ncol = 1)

# calculando cada compoente
componente_a1 <- matriz_perfil1 %*% autov_a
round(mean(componente_a1),5)

componente_b1 <- matriz_perfil1 %*% autov_b
round(mean(componente_b1),5)

# localizando as pessoas do perfil 2

subconjunto5<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE3' & TP_FORMACAO == 'FM5'
                      & TP_ESTADO_CIVIL == 'EC2'& Q001 == 'EP1' & Q002 == 'EM1')

subconjunto6<- subset(dados_2013,TP_FAIXA_ETARIA == 'FE3' & TP_FORMACAO == 'FM5'
                      & TP_ESTADO_CIVIL == 'EC3'& Q001 == 'EP1' & Q002 == 'EM1')

matriz_perfil2 <- rbind(as.matrix(na.omit(subconjunto5[,c(12:16)])),
                        as.matrix(na.omit(subconjunto6[,c(12:16)])))
# calculando cada compoente
componente_a2 <- matriz_perfil2 %*% autov_a
round(mean(componente_a2),5)

componente_b2 <- matriz_perfil2 %*% autov_b
round(mean(componente_b2),5)



# Combina os dados dos dois perfis
combined_data <- rbind(
  transform(scores_perfis, Perfil = "Perfil 1"),
```

```
  transform(scores_perfis, Perfil = "Perfil 2", P1_CP1 = P2_CP1, P1_CP2 = P2_CP2)
)

ggplot(combined_data, aes(x = -P1_CP1, y = -P1_CP2, color = Perfil)) +
  geom_point() +
  geom_text(aes(label = Ano), vjust = 1.5, hjust = 0, size = 3) +
  ggtitle("") +
  xlab("Componente 1") +
  ylab("Componente 2") +
  scale_color_manual(values = c("Perfil 1" = "red", "Perfil 2" = "blue")) +
  theme_minimal() +
  labs(color = "Perfil") +
  guides(
    color = guide_legend(
      title = NULL,
      override.aes = list(shape = c(16, 16), size = 5)
    )
  )+
  geom_vline(xintercept = 0, linetype = "dotted") +  # Vertical line at x = 0
  geom_hline(yintercept = 0, linetype = "dotted") +  # Horizontal line at y = 0
  theme(panel.background = element_rect(fill = "white"))  # Gray background
```