UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

RAFAEL TOLEDO COSTA DE ALMEIDA

Análise de fatores associados ao desempenho dos participantes do ENEM 2018 utilizando o `sparklyr`

Salvador
2021

RAFAEL TOLEDO COSTA DE ALMEIDA

Análise de fatores associados ao desempenho dos participantes do ENEM 2018 utilizando o `sparklyr`

Trabalho de Conclusão de Curso apresentado ao Curso de graduação em Estatística, Instituto de Matemática e Estatística, Universidade Federal da Bahia, como requisito para aprovação na disciplina de Trabalho de Conclusão de Curso II.

Orientadora: Profa. Dra. Edleide de Brito
Co-orientadora: Profa. Dra. Gecynalda Soares da Silva Gomes

Salvador
2021

SUMÁRIO

1 INTRODUÇÃO

A criação de oportunidades, a disseminação do conhecimento e fomentação do ensino e da visão crítica auxiliam na busca por melhores condições socioeconômicas na sociedade como um todo. A educação é um dos principais fatores que permitem transformar a sociedade e os seus indivíduos.

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1988 com a finalidade de avaliar o desempenho dos alunos concluintes do ensino médio. Inicialmente, no referido ano, o resultado no ENEM foi utilizado para ingresso por duas instituições de ensino superior. No ano seguinte, o exame registrou um aumento expressivo na participação de Instituições de Ensino Superior (IES), totalizando 93 instituições usando o resultado do ENEM como forma de ingresso (INEP, 2019).

A partir do ano 2000, o ENEM dá um grande salto relacionado à acessibilidade, promovendo a equidade por meio de assistência às pessoas que necessitam de atendimento especializado, permitindo-as disputar uma vaga para acesso as IES.

O ENEM tornou-se o maior exame nacional brasileiro por conta das melhorias relacionadas a tecnologia e internet. Essas melhorias estão atreladas as inovações do processo de inscrição que obteve uma crescente demanda de inscrições e novos locais de provas para atender grande parte das localidades do Brasil. Em 2003, além da revisão do questionário de inscrição sobre a situação de conclusão do participante, também foi inserido o questionário socioeconômico para traçar o perfil dos inscritos. A partir de 2004, foi criado o Programa Universidade para Todos (ProUni), possuindo a nota do ENEM como critério para validação da participação do candidato. Mediante isso, as possibilidades foram ampliadas para ingresso nas IES por meio do ENEM, o que evidencia a relevância desse exame para o cenário educacional brasileiro (INEP, 2019).

Após uma década de criação, inovações e atualizações foram feitas buscando garantir a adaptação e acessibilidade a todos. Essas inovações auxiliaram na inclusão de participantes portadores de alguma deficiência e possibilitaram a criação de cotas para diferentes categorias para uma equidade na concorrência das vagas. Cada vez mais o ENEM passa a ser considerado como o principal pré-requisito em vários processos e programas de ingresso em IES. Um exemplo disso é a sua condição de pré-requisito de participação ao programa Fies (Fundo de Financiamento Estudantil - programa este que o participante pode estudar e financiar os estudos nas IES de vínculo privado).

A partir de 2013, o ENEM é formatado como o exame que conhecemos atualmente. Todas as IES estão presentes no processo e as implementações a respeito dos aspectos socioeconômicos advindos do questionário socioeconômico estão inseridas no processo de inscrição. Além disso, os aspectos referentes ao processo de inclusão e acessibilidade por meio do atendimento especial e da adequação da realização das provas para esse grupo

estão definidas e em prática (INEP, 2019).

Diante da estrutura atual, o ENEM é imprescindível para o futuro do país e para a educação brasileira porque possibilita que todos os indivíduos tenham a oportunidade de ingressar nas IES.

O INEP possui o Planejamento de Dados Abertos (PDA) que detalha as informações produzidas pelo INEP, como estatísticas, indicadores educacionais e resultados de avaliação educacional dos diferentes níveis de ensino, que servem à produção de estudos e pesquisas para acompanhamento e controle social (INEP, 2019). Diante disso, disponibiliza os microdados originados das avaliações, pesquisas e exames referentes aos principais programas do instituto, como o ENEM, o ENADE, o Saeb, Censo Escolar, o Censo da Educação Superior, entre outros.

As informações alusivas a esses programas são disponibilizadas em formato de compressão específico (.zip) e podem ser obtidas via *download* no *website* do INEP, contendo os *inputs* (canais de entrada) para que possa ser feita a leitura através dos *softwares* SAS, SPSS, R entre outros. Vale salientar que para a descompressão dos arquivos é necessário a utilização de um programa descompactador.

A proposta do presente trabalho é processar e analisar os microdados do ENEM do ano de 2018 através do *framework Spark* e do pacote `sparklyr`, respectivamente, para verificar os fatores que estão associados ao desempenho dos participantes no referido ano por meio da análise multivariada. Além disso, apresentar as vantagens e desvantagens de trabalhar com *Big Data* em uma máquina com especificações intermediária (ver Seção 4) usando o *sparklyr*.

No presente trabalho, além da Seção Introdução, outras cinco seções compõem o referido trabalho. Na Seção 2, é apresentado um panorama do uso dos microdados do ENEM que são comumente trabalhados com propostas voltadas a análise subdivididas desses dados, sendo determinado um subconjunto por estado, região ou até mesmo um conjunto menor de variáveis. Além disso, apresenta-se a importância do trabalho pela proposta da análise dos microdados do ENEM 2018 em sua totalidade, ou seja, analisando os dados do ENEM 2018 de todo o país. A Seção 3, designa a questão do referencial teórico, no qual é evidenciado alguns trabalhos que utilizaram os microdados do ENEM com outras metodologias e análises. É visto que os microdados do ENEM são comumente usados para análise descritiva e exploratória desses dados. Além do uso de análises voltadas a Teoria de Resposta ao Item (TRI) para análise da interação dos itens, inscritos e das áreas das provas estudadas, também é evidenciado o uso da análise de correspondência para dados voltados a área de ensino.

Na Seção 4, é apresentada as informações referentes ao conjunto de dados como a quantidade e descrição das variáveis, assim como a explicação da filtragem dos dados e o motivo de se trabalhar com as referidas variáveis. Em seguida, é apresentada a definição do termo *Big Data* de acordo com os cinco "V's" e a importância da análise desses volumosos

conjuntos de dados.

A Subseção 4.3 é referente ao pré-processamento dos microdados do ENEM. Evidencia-se o uso do *framework Spark* e o pacote `sparklyr` no *software* R para o processamento dessa grande quantidade de dados em uma máquina com especificações intermediárias. Diante disso, é descrito os desafios de se trabalhar com essa ferramenta de pré-processamento de *Big Data*, assim como é inserido as vantagens e desvantagens do `sparklyr`. Já na Subseção 4.5, encontra-se a fundamentação teórica sobre análise multivariada, dando ênfase na análise de correspondência múltipla e os conceitos que norteiam esta abordagem.

Na seção 5, encontra-se a visualização dos resultados da análise exploratória dos dados, como por exemplo o perfil dos inscritos do ENEM. Apresenta-se o desempenho dos participantes nas provas objetivas e na prova discursiva, e a confecção dos mapas de distribuição dos inscritos segundo esse desempenho no cenário por estado. Ademais, na Subseção 5.2 são apresentados os resultados da análise de correspondência múltipla, evidenciando os fatores que estão associados ao desempenho dos participantes no ENEM de 2018 através dos mapas de correspondência.

2 JUSTIFICATIVA

O uso dos microdados do ENEM são comumente trabalhados por propostas que visam analisar esses dados de maneira subdividida. Nesse sentido, é considerado um subconjunto desses dados, podendo ser a escolha de determinado estado, região, bem como um conjunto menor de variáveis. De acordo com o referencial teórico, as análises realizadas nos dados do ENEM estão associadas a análises descritivas, bem como o uso de técnicas multivariadas levando em consideração o subconjunto de dados de interesse do estudo.

Nesse contexto, o presente trabalho traz uma proposta diferenciada de análise desses microdados, pois serão utilizados os microdados do ENEM no ano de 2018 em sua totalidade, ou seja, a nível nacional. Considerando os desafios da análise de *Big Data* e do uso da recente ferramenta de pré-processamento *Spark*, o objetivo principal do presente trabalho é analisar os fatores associados ao desempenho de todos os participantes no ENEM 2018 por meio da análise de correspondência múltipla, descobrindo novas maneiras de aplicar esses métodos juntamente ao *framework Spark*. Além disso, o objetivo específico deste trabalho é a criação de um painel para ilustrar os diferentes resultados encontrados e facilitar a compreensão das informações geradas. Dessa forma, será criado uma *Dashboard* em *Shiny R*.

3 REVISÃO DA LITERATURA

O ENEM tem sido tema de propostas para diversas análises, por ser considerado o maior exame nacional, pela sua importância como forma de acesso as IES e por possuir informações que podem responder perguntas no âmbito educacional, social e econômico de todo o Brasil. Os trabalhos referentes aos microdados do ENEM, utilizam como principal tipo de análise, a descritiva. É perceptível a necessidade de se conhecer o comportamento desses dados, principalmente quando se pretende traçar outros tipos de análises. Entretanto, a questão de utilizar, na maioria das vezes, apenas a análise descritiva evidencia que não há tantos trabalhos que utilizam os dados do ENEM em outros tipos de análises mais modernas (LIMA, 2019).

Análises envolvendo o desempenho e rendimento dos inscritos por escola ou por disciplina são mais usadas, e têm como finalidade verificar quais condições estão atreladas a esses resultados. No trabalho de Álvaro Almeida (2014), o objeto do estudo é avaliar os resultados dos estudantes dos Institutos Federais da região Nordeste durante a validade do Plano Nacional de Educação (PNE) entre os anos de 2001 e 2010, após a ampliação das matrículas no referido instituto para o nível técnico profissionalizante integrado ao ensino médio. Foi mensurado os resultados de 52 escolas com o intuito de obter respostas relacionadas a qualidade do ensino de vínculo público e também de fatores intraescolares, como por exemplo o resultado escolar e fluxo escolar (ALMEIDA, 2014).

A Teoria de Resposta ao Item (TRI) é bastante abordada quando o tema são os dados do ENEM. O INEP desempenha a análise da dimensionalidade (dimensão de itens, como por exemplo, a interpretação de texto ligado as provas de Linguagens e Ciências Humanas/Naturais ou raciocínio lógico ligado a prova de Matemática) das provas objetivas do ENEM, e para o modelo TRI proposto, as provas devem ser essencialmente unidimensionais. Essa abordagem foi trabalhada em Vieira (2016) para identificar modelos da TRI que melhor se ajustam a uma prova integrada de quatro áreas do ENEM, analisando a interação dos itens, dos inscritos e das áreas estudadas. Ainda neste trabalho, o autor utiliza a análise fatorial exploratória com o objetivo de analisar os possíveis traços latentes medidos pela prova agregada. Já a análise fatorial confirmatória foi utilizada para verificar se cada área da referida prova pode ser considerada uma dimensão. Os principais resultados apontam que devido a prova ser considerada basicamente unidimensional, infere-se que os alunos com alta cognição dominam todas as áreas do conhecimento presentes no ENEM. Entretanto, por meio dos resultados advindos das análises de modelos mais complexos houve uma distinção de duas proficiências, tendo como exemplo raciocínio lógico e leitura e interpretação de textos, no qual sugeriram o modelo bidimensional como sendo o mais adequado para gerar as proficiências dos alunos (VIEIRA, N., 2016).

A análise fatorial é uma técnica multivariada que busca simplificar um estudo com-

plexo, reduzindo o número de variáveis correlacionadas observadas para uma quantidade menor de fatores, que podem ser vistas como variáveis não observadas. Tem como finalidade investigar a variabilidade entre essas variáveis correlacionadas (JOHNSON, R.; WICHERN, D.W., 1998). O uso na análise fatorial de informação completa (BOCK E AITKIN, 1981) no trabalho de Sousa (2015), possibilitou analisar a dimensionalidade da prova do ENEM no ano de 2001 e relacionar os itens da TRI com os relativos fatores e com as competências de cada item. A análise fatorial de informação completa não necessita do cálculo dos coeficientes de correlação inter-itens, porque o método esclarece as dificuldades enfrentadas durante a análise, a título de exemplo os problemas de acerto ao acaso e de itens não apresentados. No trabalho de Sousa (2015), ocorre a comparação dos alunos que obtiveram a mesma nota de acordo com os escores estimados dos modelos multidimensionais criado segundo os itens da TRI. Além disso, evidencia a importância da modelagem da TRI com o intuito de avaliar os inscritos que responderam o exame de natureza multidimensional, não por uma nota geral e sim por uma determinada dimensão e o que a mesma representa para o contexto do exame, no qual pode selecionar o candidato com maior proficiência.

Em outra vertente, existe o uso dos dados do ENEM na prática da área de mineração de dados, pois importante ressaltar, os microdados do ENEM são considerados *Big Data* e esse massivo conjunto de dados é composto por informações importantes para diversas áreas, principalmente educacional. O uso de Mineração de Dados Educacionais (EDM) trata da criação de procedimentos para minerar dados na área educacional (ROMERO, C.; VENTURA, S., 2010).

Em 2014, a proposta do trabalho de Silva (2014) foi o uso da mineração de dados do ENEM 2010 para aplicação do método *Knowledge Discovery from Databases* (KDD), em português, "Descoberta de Conhecimento de Bases de dados". O intuito desta aplicação é justamente o pré-processamento dos dados do ENEM referentes ao desempenho e ao questionário socioeconômico, minerando os dados pertinentes das capitais da região Sudeste. Do questionário socioeconômico foi utilizado as quatro primeiras perguntas/questões que são alusivas a quantidade de pessoas que moram com o inscrito, o nível de escolaridade da mãe, tipo de escola que frequenta/frequentou e a renda familiar mensal. A utilização da técnica de associação de dados com o algoritmo *a priori* para avaliar a causa e o efeito referente ao desempenho na referida prova possibilitou observar que os aspectos socioeconômicos como a renda familiar baixa, a escolaridade dos responsáveis de nível primário e o número alto de pessoas que moram com o estudante são atributos que diminuem o desempenho do aluno.

Na perspectiva da análise de correspondência, o trabalho de Nascimento, Cavalcanti e Ostermann (2017) intitulado como "Análise de Correspondência Aplicada à Pesquisa em Ensino de Ciências" teve como objetivo abordar o uso da análise de correspondência na esfera educacional com aplicabilidade nos microdados do ENEM no ano de 2014. Os

autores escolheram esse conjunto de dados por apresentar uma considerável quantidade de variáveis categóricas que são adequadas para a aplicação da análise em questão. Interligado a esta metodologia, outro objetivo do referido trabalho é explicar o perfil de candidatos bem sucedidos e mal sucedidos no exame educacional no ano de 2014. As variáveis utilizadas nesse trabalho estão relacionadas com a dependência administrativa da escola no qual o participante obteve a conclusão do Ensino Médio, a etnia e o desempenho do mesmo. Os resultados desse trabalho evidenciaram uma grave desigualdade social no sistema de ensino brasileiro, bem como que a técnica multivariada usada conduziu a resultados consistentes nas respectivas variáveis.

4 METODOLOGIA

4.1 DESCRIÇÃO DOS DADOS

Os microdados do ENEM podem ser obtidos no *website* do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Neste projeto iremos trabalhar com os microdados do ENEM 2018, caracterizados pela compactação dos dados coletados por meio das inscrições *online*, das avaliações realizadas nas datas definidas segundo o calendário do respectivo ano.

A base de dados é composta por 137 variáveis distribuídas em diferentes aspectos, como por exemplo, aspectos educacionais e socioeconômicos dos inscritos. Essas variáveis remetem aos dados dos participantes, das escolas que frequentaram, dos pedidos de atendimento especializado ou específico (no caso dos inscritos que declaram possuírem algum tipo de deficiência e que necessitam de determinados atendimentos para a realização das provas), do local de aplicação da prova, da prova objetiva e da redação, bem como dados referentes ao questionário socioeconômico composto por 27 perguntas relacionadas às condições do participante e de sua família. Algumas informações são codificadas para que o participante não seja identificado, por exemplo, o número de inscrição do participante não está completo na base de dados.

Através do pré-processamento dos dados, utilizando o *framework Spark* e o pacote `spaklyr` no R, foi decidido por utilizar as seguintes características: número de inscrição, código do município de residência, sexo, idade, estado civil, nacionalidade, cor/raça, situação de conclusão, ano de conclusão, tipo de escola e tipo de ensino do Ensino Médio, indicador de inscrito treineiro, indicador de inscrito idoso, código do município e localização da escola, indicador de deficiência física, deficiência mental, presença nas provas por área (ciências da natureza, ciências humanas, linguagens e códigos, matemática e suas tecnologias), notas das respectivas provas por área e na redação, linguagem estrangeira escolhida e as 27 questões do questionário socioeconômico. Vale salientar que, o público alvo do presente projeto são os inscritos que estavam presentes em todos os dias da realização das provas do referido ano.

A escolha das variáveis citadas anteriormente se dá por conta do objetivo deste trabalho, no qual é de interesse analisar os fatores que estão associados ao desempenho dos inscritos no ENEM no referido ano. Nesse sentido, utilizar as variáveis de controle do participante, da escola, das provas objetivas e do questionário socioeconômico agrega valor para à análise.

Além disso, as variáveis que não foram utilizadas são referentes ao aspecto da prova (como a cor do caderno de provas) e também ao gabarito e itens da prova, que são voltados a Teoria de Resposta ao Item (TRI) e não corroboram para o objetivo deste trabalho.

Todas as análises foram feitas no *software* R (versão 4.0.0) usando a *interface* RStudio (versão 1.2.5042) com usabilidade do *pacote* `sparklyr` (versão 3.0) do *framework Spark* (versão 3.0) em uma máquina com especificação de 8GB de memória RAM, processador *Intel(R) Core(TM) i5-8265U CPU* com *1.60 GHz* e *1.80 GHz* e *1TB* de armazenamento.

4.2 *BIG DATA*

Com o surgimento das inovações no ramo da tecnologia e maior facilidade de acesso a dispositivos digitais, a era dos dados estabeleceu-se no século XX como uma grande expansão de tecnologias e dados em diversas áreas. Esse rápido crescimento tecnológico resultou em uma quantidade massiva de dados advindos de órgãos governamentais, da mídia e das redes sociais.

Atualmente, vivemos rodeados por tecnologias capazes de coletar informações a todo momento. Tecnologias desenvolvidas para captação de informações climáticas, navegação em sites, registro de operações de venda e compra, GPS de telefones celulares, informações de busca e visualizações de conteúdos de interesse nas redes sociais são exemplos que evidenciam o tamanho da circulação desses dados. E como a natureza das informações está mudando pela disseminação de serviços oferecidos pelos aplicativos de *smartphones*, grande parte desses conjuntos de dados são produzidos em tempo real.

Proveniente dessa nova realidade mundial, as grandes empresas e instituições governamentais perceberam a necessidade de possuir tecnologias para armazenamento, processamento e transformação desses dados em informações que possam auxiliar em decisões fundamentais para a sociedade em diversas áreas, como por exemplo: ciência, saúde, educação, economia, segurança e desenvolvimento. Nesse sentido e capacidade que o termo *Big Data* tornou-se um dos mais importantes nos últimos anos, pois dita a forma como essa massa de dados será traduzida em informações e, posteriormente, convertida em resultados.

Big Data refere-se a conjuntos de dados cujo tamanho está longe da capacidade do uso das ferramentas de coleta, armazenamento, processamento e análise de base de dados mais convencionais estruturados em planilhas/tabelas (MANYIKA, 2011). O termo *Big Data* é caracterizado pelos cinco "V's", que são volume, variedade, velocidade, veracidade e valor.

O volume explica bem a definição de *Big Data* pelo fato do surgimento de grandes quantidades de dados diariamente e pelo questionamento de como lidar com esses dados quanto ao armazenamento e processamento desses dados. Por meio disso, a variedade desses dados ocorre por causa das diferentes maneiras de coleta, bem como pelos diversos tipos de dados existentes no mundo. Esses dados são gerados de maneira muita rápida, em pouco tempo, uma massiva quantidade de dados é gerada. A veracidade desses dados é uma característica importante, pois na era da informação em que vivemos há uma grande

quantidade de informações que precisam ser verificadas e validadas a ponto de se tornarem dados confiáveis para possíveis análises. O vínculo ou valor, possibilita o cruzamento dos dados por meio das ferramentas de *Big Data* com a finalidade de extrair informações (MAYER-SCHÕNBERGER, V.; CUKIER, K., 2013).

É importante especificar que a definição de *Big Data* não está apenas associada ao seu tamanho em *gigabytes*, *terabytes* ou até mesmo *petabytes*, mas também na forma como esses dados crescem conforme as inovações tecnológicas surgem e na impossibilidade de alguns *softwares* e máquinas convencionais em tratar esses dados (MANYIKA, 2011).

Em consequência do *Big Data*, novas tecnologias foram desenvolvidas para armazenamento, transferência e processamento de dados, bem como a implementação de novas técnicas para manipulação, análise e visualização desses dados. A Estatística é indispensável para o sucesso da coleta, organização e interpretação desses dados por meio das técnicas estatísticas. O R é um dos *softwares* que têm importância nas etapas dos processos descritos anteriormente. Entretanto, por mais que a criação de novo método tenha sido feita para lidar com os novos desafios, nem todos (indivíduos, empresas de tamanho micro, pequeno ou médio porte) possuem condições de trabalhar com essas inovações, pois a depender da base de dados e do projeto, é necessário alto investimento em máquinas, servidores para um alto poder computacional.

Outra questão relevante é como o *Big Data* é compreendido e utilizado no âmbito acadêmico ou científico e no âmbito empresarial ou governamental nas áreas de conhecimento (SCHROEDER, 2018). É possível entender que na primeira situação, o *Big Data* é utilizado para desenvolver o conhecimento estatístico e computacional. Em contrapartida, no âmbito das instituições de vínculo privado ou público, o trabalho com o *Big Data* se dá não só pela questão de processamento e análise, mas bem como na visualização dos dados e na obtenção de resultados voltados aos comportamentos sociais dos indivíduos na sociedade. Alguns exemplos são a análise do perfil dos clientes para oferecimento de novos serviços em um determinado banco, a distribuição e a quantidade de indivíduos contaminados pelo vírus COVID-19, a realização de avaliações de discentes e docentes, a distribuição do perfil dos inscritos no ENEM em determinado ano.

Com o direcionamento da inovação técnico-científico, a usabilidade do *Big Data* estará mais difundida na sociedade, tanto na esfera acadêmica quanto na esfera governamental. Esse assunto ainda precisa ser mais expandido para que possa ser evidenciado o seu poder e a sua importância para definir o rumo da sociedade. Define tendências, gera volume de informações ao mesmo tempo que otimiza tempo e diminui custos. O *Big Data* não é mais um termo do futuro, e sim, um termo já presente no dia a dia e cada vez mais em ascensão juntamente com a tecnologia.

4.3 PRÉ-PROCESSAMENTO DOS DADOS

Com a base de dados definida, surgiu-se um novo desafio a ser entendido e superado: trabalhar com uma grande base de dados em uma máquina com especificações intermediárias voltada para o uso de atividades que não exijam muito do processamento (memória RAM) da máquina. Diante disso, a ideia de buscarmos uma forma de pré-processar os dados tornou-se imprescindível para que pudéssemos realizar atividades relevantes para a metodologia do presente projeto como a leitura e organização da base de dados, análise descritiva e exploratória e análise multivariada.

Nesse sentido, a utilização do *Spark Apache* torna-se viável e necessária na etapa de pré-processamento de dados de grande escala (*Big Data*). Uma breve descrição do *Spark* é apresentada a seguir.

4.3.1 *Spark*

O *Spark* é um projeto que foi desenvolvido em 2009 pela organização *Apache Software Foundation*, como proposta de ser gratuito e de código aberto (*open source*) para ser utilizado como uma importante ferramenta de processamento de *Big Data*. O desempenho do *Spark* se dá em evitar o consumo total da memória, pois o seu funcionamento consiste em usar a memória e o disco de maneira particionada.

Com isso, guarda-se alguns resultados temporários na memória RAM e na memória em disco. Mesmo que haja mais memória em disco nos computadores, essa memória é mais lenta. Diante disso, a junção do uso dessas duas memórias possibilita o processamento dos dados e o surgimento das análises de forma mais rápida e prática se comparado ao processamento sem o uso de forma particionada.

Além do *Spark*, outro *framework open source* da *Apache Foundation* bastante utilizado é o *Hadoop*, o qual é caracterizado pelo armazenamento de dados e execução de aplicações por meio de agrupamento dos dados, proporcionando acesso e segurança nas aplicações desses dados. Com isso, o *Hadoop* permite grande armazenamento de dados para processamento e realização de futuras análises (CETAX, 2020). Em comparação com o *Hadoop*, a execução por *Spark* é mais rápida, pois utiliza melhor a memória de acesso rápido (*cache*), fazendo com que o tempo de acesso a dados armazenados seja o mais breve possível. Desse modo, toda vez que o *software* solicita ou realiza alguma atividade, os dados que estão guardados são prontamente utilizados e o tempo de processamento encurtado. O *Spark* possui o recorde de classificação de nuvem, que o caracteriza como a solução mais econômica e rápida para classificação de grandes conjuntos de dados na nuvem (LURASCHI, 2019).

4.3.2 Sparklyr

O pacote `sparklyr` foi criado para ser uma extensão do *framework Spark* para o *software* R. O conceito é usar funções do *Spark* dentro do ambiente do R. A instalação do `sparklyr` é igual a qualquer outro pacote do R, trazendo familiaridade aos programadores intermediários e avançados no *software* que desejam trabalhar com análise de *Big Data*.

O *Spark* é uma ferramenta moderna e poderosa. O `sparklyr` por ser uma extensão do *Spark*, traduz as implementações de R para a linguagem de programação *Scala* (pelo fato que o *Spark* roda em *Scala*) e então realiza o processamento de grandes bases de dados.

Dessa forma, possui grande relevância porque expande a usabilidade no âmbito estatístico para diversos meios - conforme a área do conjunto de dados - e gera resoluções e implementações novas. Além disso, todo esse procedimento pode ocorrer em máquinas que não possuem requisitos suficientes para suportar o processamento de *Big Data*.

A busca pelo aprendizado referente a aplicação do pré-processamento do `sparklyr` tornou-se um desafio por ser uma nova maneira de se trabalhar com importação de dados, juntamente com a questão de aprender a lidar com outro pacote importante para manipulação de dados. A instalação do pacote `dplyr` do R tornou-se fundamental, pois a base de dados é convertida em *data frame* pelo *Spark*, e o `dplyr` converte as codificações em linguagem SQL (*Standard Query Language*) do *Spark*. Dessa forma, o pacote `dplyr` auxilia na seleção das variáveis de interesse, na filtragem das informações de interesse e, posteriormente, nas análises.

Os comandos do `dplyr` são escritos por meio do operador `‘‘pipe’’ (%>%)`. Exemplificando o uso do operador `pipe` para calcular a média aritmética de uma variável em uma base de dados, temos:

```
dados %>%
select(variavel1) %>%
summarise(Media = mean(variavel1, na.rm=TRUE)
```

Outro comando que geralmente é utilizado é o `collect()`, o qual permite que após o processamento dos dados, o usuário possa coletar esses dados e armazenar em uma nova base de dados para realização de análises e visualização de dados no R utilizando a memória RAM da máquina.

Basicamente, o processamento de funções mais básicas - por exemplo medidas de tendência central - ocorre da memória para o disco. Entretanto, no caso de atividades que demandam mais processamento da máquina, principalmente pela grande quantidade de observações, o uso da memória RAM ao invés do uso no disco torna possível a realização da análise descritiva e de técnicas exploratórias no campo multivariado.

4.3.2.1 Vantagens do Sparklyr

Devido a situação do desafio em lidar com *Big Data*, o uso do *Spark* por meio do pacote `sparklyr` de acordo com o progresso do presente projeto evidenciou algumas percepções quanto as vantagens em utilizar o `sparklyr` para essa etapa do projeto.

Por ser de fácil instalação no *software* R e por ser distribuído em CRAN, o `sparklyr` é uma importante ferramenta para pré-processamento de *Big Data*, sendo uma extensão do *Spark*. Além disso, por não exigir tanto da memória RAM de máquinas básicas ou intermediárias, o `sparklyr` surge como alternativa para realizar conexão local (no próprio computador) e cluster remoto (servidor remoto) para justamente equilibrar esse uso e não forçar diretamente a máquina que está realizando o processo.

Outra vantagem do `sparklyr` é ser extensível, ou seja, permite ser utilizado com outros pacotes do R para tarefas como manipulação dos dados para estruturação, organização e visualização desses dados. O pacote `dplyr` auxilia na questão da manipulação dos dados, pois pela sua compatibilidade com o `sparklyr`, traduz a codificação do R em SQL para o *Spark* e trabalhando na manipulação dos dados conforme a utilização das funções `select()`, `filter()`, `summarise()`, `mutate()`, `arrange()` e `copy_to()`.

Para a visualização dos dados, o uso do `sparklyr` mediante o pacote `ggplot2`, promoveu a construção dos gráficos de distribuição do perfil dos inscritos segundo alguns fatores sociais como sexo, raça/cor, tipo de escola, escolaridade do pai e da mãe ou responsável, bem como a visualização do desempenho médio dos inscritos no ENEM 2018 por áreas de ensino por estados. Vale salientar que sem o uso do `sparklyr` não seria possível obter esses gráficos usando um notebook com especificações iguais ou menores a 8GB de memória RAM e processador *Intel(R) Core(TM) i5-8265U CPU com 1.60GHz e 1.80 Ghz*.

4.3.2.2 Desvantagens do Sparklyr

Por mais que o `sparklyr` apresente vantagens relevantes na alçada de pré-processar uma massiva quantidade de dados, alguns pontos negativos também são reconhecidos e tornaram o desafio em lidar com esse pacote ainda maior.

Umas das desvantagens do `sparkly` é que, por ser recente, existem poucos materiais e artigos científicos que abordam o seu uso de forma mais expandida para o *software* R, principalmente no idioma Português. Faltam livros do tema, e há uma quantidade pequena de artigos que abordam o uso computacional do `sparklyr`. A referência utilizada para o estudo e desenvolvimento do `sparklyr` no R, foi a obra de Javier Luraschi, Kuo Javier e Ruiz Edgar intitulada como *"Mastering Spark with R"*, publicada em 2019, totalmente em inglês.

Outro ponto, é a limitação do uso da função `copy_to()` para criar uma nova base de dados *Spark*, pois a função não guarda/preserva as colunas/variáveis formatadas. Nesse

sentido, foi observado que a função `collect()` também necessita de uma quantidade limitada de dados para que possa ser coletada e gerada uma nova base de dados, com a finalidade de usar 100% da memória RAM da máquina e não mais a conexão com o *Spark*. Entretanto, caso essa quantidade ainda seja grande, após a coleta da nova base, as funcionalidades que necessitam do poder da máquina irão ocasionar no reinício forçado do R e o travamento do processamento da máquina.

Algumas funções não funcionaram de forma direta no `sparklyr`, ocasionando na necessidade da busca por resoluções em fóruns e artigos. Isso seria algo relativamente comum em um estudo ou em uma resolução de um problema na esfera computacional. Entretanto, como foi citado antes, algumas implementações utilizando R e o pacote `sparklyr` são relativamente escassas. Por isso algumas implementações foram realizadas e testadas para que atividades como categorização de variáveis pudessem ser feitas para a continuidade das análises. Nem sempre as codificações utilizadas somente no R - com ou sem o pacote `dplyr` - são funcionais quando se trabalha com `sparklyr`, bem como não há dicionários a respeito dos comandos que são úteis ao uso conjunto do R mais `sparklyr`.

4.4 ANÁLISE MULTIVARIADA

A análise multivariada é composta por um conjunto de métodos estatísticos voltados a análise simultânea de múltiplas medidas sobre cada observação. Essa análise simultânea ocorre em mais de duas características para cada observação de um determinado conjunto de dados. Identificar que método multivariado utilizar é onde encontra-se o desafio, pois é necessário analisar o conjunto de dados em estudo e estudar os pressupostos de cada técnica, além das vantagens e desvantagens.

O objetivo da análise multivariada é utilizar técnicas exploratórias para reduzir e sintetizar a estrutura de variabilidade dos dados, ou seja, reduz as variáveis originais sem que haja a perda de informações importantes para o estudo, tornando assim mais simplificadas as interpretações dos resultados. Ainda, esta análise é bastante conhecida por usar técnicas que visam a classificação e discriminação por meio da criação de grupos de itens ou variáveis de acordo com a condição de similaridade conforme as características em estudo. Outra finalidade é o estudo da relação entre as variáveis, no qual explora se há associação entre as variáveis (JOHNSON, R. A.; WICHERN, D. W., 1998).

Em relação a esta finalidade, neste trabalho o uso da análise multivariada é importante para investigar a relação das variáveis de controle dos participantes, da escola e do questionário socioeconômico com o desempenho no ENEM no ano de 2018. Como foi dito anteriormente, é necessário saber qual técnica é adequada para que se possa obter resultados consistentes sem que haja perda de informações e viés no trabalho. Nesse sentido, decidiu-se por utilizar a Análise de Correspondência Múltipla, pois é uma técnica que melhor se encaixa na finalidade de estudar as associações dos fatores e o desempenho

dos participantes. A seguir, temos a decrição da Análise de Correspondência, bem como a Análise de Correspondência Múltipla.

4.4.1 Análise de Correspondência

A Análise de Correspondência é uma técnica estatística multivariada válida para o estudo de dados categóricos que possibilita avaliar graficamente as associações após a redução da dimensionalidade em um determinado conjunto de dados. Essencialmente, esta técnica é caracterizada por tabelas de contingência, seja de ordem 2 ou superior para verificar a associação entre duas ou mais variáveis categóricas de acordo com as linhas e colunas das mesmas.

Por ser capaz de investigar a associação entre as variáveis, essa técnica é importante nas Ciências Sociais por possibilitar verificar a relação das variáveis voltadas aos aspectos socioeconômicos. Além disso, por ser uma análise de característica multivariada, esse método difere das outras análises exploratórias por ser possível verificar relações que talvez não seriam tão facilmente vistas por meio dos pares de variáveis (NASCIMENTO, M. M.; CAVALCANTI, C.; OSTERMANN, F., 2017).

Um dos pressupostos que caracteriza a Análise de Correspondência é que os dados que serão analisados precisam ser positivos e formatados como uma tabela retangular. Conforme a quantidade de variáveis dispostas nesta tabela, temos que esta análise pode ser dividida em dois tipos de análises: análise de correspondência simples (CA, do inglês *Correspondence Analysis*) e análise de correspondência múltipla (MCA, do inglês *Multiple Correspondence Analysis*).

Quando os dados são apresentados em tabelas de ordem 2, usamos CA e podemos verificar a associação das variáveis por meio da utilização do mapa de correspondência. Se temos um conjunto de pontos I correspondentes às linhas e um conjunto de pontos $K = 2$ correspondentes às colunas. No mapa da CA as posições dos pontos refletem associações. Os pontos das linhas que se aproximam indicam linhas que têm perfis semelhantes através das colunas. Os pontos de coluna que estão próximos uns dos outros indicam colunas com perfis semelhantes ao longo das linhas. Os pontos das linhas que estão próximos dos pontos das colunas representam combinações que ocorrem com mais frequência do que seria de esperar de um modelo de independência (as categorias das linhas não estão relacionadas com as categorias das colunas). O resultado habitual de uma análise de correspondência inclui a "melhor" representação bidimensional dos dados, juntamente com as coordenadas dos pontos traçados, e uma medida (chamada inércia) da quantidade de informação retida em cada dimensão (JOHNSON, R. A.; WICHERN, D. W., 1998).

A MCA é caracterizada pela utilização de tabelas com dimensões superiores a dois ($K > 2$). Com isso, esta análise evidencia o uso de pelo menos três variáveis categóricas e suas respectivas associações.

4.5 ANÁLISE MULTIVARIADA

A análise multivariada é composta por um conjunto de métodos estatísticos voltados a análise simultânea de múltiplas medidas sobre cada observação. Essa análise simultânea ocorre em mais de duas características para cada observação de um determinado conjunto de dados. Identificar que método multivariado utilizar é onde encontra-se o desafio, pois é necessário analisar o conjunto de dados em estudo e estudar os pressupostos de cada técnica, além das vantagens e desvantagens.

O objetivo da análise multivariada é utilizar técnicas exploratórias para reduzir e sintetizar a estrutura de variabilidade dos dados, ou seja, reduz as variáveis originais sem que haja a perda de informações importantes para o estudo, tornando assim mais simplificadas as interpretações dos resultados. Ainda, esta análise é bastante conhecida por usar técnicas que visam a classificação e discriminação por meio da criação de grupos de itens ou variáveis de acordo com a condição de similaridade conforme as características em estudo. Outra finalidade é o estudo da relação entre as variáveis, no qual explora se há associação entre as variáveis (JOHNSON, R. A.; WICHERN, D. W., 1998).

Em relação a esta finalidade, neste trabalho o uso da análise multivariada é importante para investigar a relação das variáveis de controle dos participantes, da escola e do questionário socioeconômico com o desempenho no ENEM no ano de 2018. Como foi dito anteriormente, é necessário saber qual técnica é adequada para que se possa obter resultados consistentes sem que haja perda de informações e viés no trabalho. Nesse sentido, decidiu-se por utilizar a Análise de Correspondência Múltipla, pois é uma técnica que melhor se encaixa na finalidade de estudar as associações dos fatores e o desempenho dos participantes. A seguir, temos a decrição da Análise de Correspondência, bem como a Análise de Correspondência Múltipla.

4.5.1 Análise de Correspondência

A Análise de Correspondência é uma técnica estatística multivariada válida para o estudo de dados categóricos que possibilita avaliar graficamente as associações após a redução da dimensionalidade em um determinado conjunto de dados. Essencialmente, esta técnica é caracterizada por tabelas de contingência, seja de ordem 2 ou superior para verificar a associação entre duas ou mais variáveis categóricas de acordo com as linhas e colunas das mesmas.

Por ser capaz de investigar a associação entre as variáveis, essa técnica é importante nas Ciências Sociais por possibilitar verificar a relação das variáveis voltadas aos aspectos socioeconômicos. Além disso, por ser uma análise de característica multivariada, esse método difere das outras análises exploratórias por ser possível verificar relações que talvez não seriam tão facilmente vistas por meio dos pares de variáveis (NASCIMENTO, M. M.; CAVALCANTI, C.; OSTERMANN, F., 2017).

Um dos pressupostos que caracteriza a Análise de Correspondência é que os dados que serão analisados precisam ser positivos e formatados como uma tabela retangular. Conforme a quantidade de variáveis dispostas nesta tabela, temos que esta análise pode ser dividida em dois tipos de análises: análise de correspondência simples (CA, do inglês *Correspondence Analysis*) e análise de correspondência múltipla (MCA, do inglês *Multiple Correspondence Analysis*).

Quando os dados são apresentados em tabelas de ordem 2, usamos CA e podemos verificar a associação das variáveis por meio da utilização do mapa de correspondência. Se temos um conjunto de pontos I correspondentes às linhas e um conjunto de pontos $K = 2$ correspondentes às colunas. No mapa da CA as posições dos pontos refletem associações. Os pontos das linhas que se aproximam indicam linhas que têm perfis semelhantes através das colunas. Os pontos de coluna que estão próximos uns dos outros indicam colunas com perfis semelhantes ao longo das linhas. Os pontos das linhas que estão próximos dos pontos das colunas representam combinações que ocorrem com mais frequência do que seria de esperar de um modelo de independência (as categorias das linhas não estão relacionadas com as categorias das colunas). O resultado habitual de uma análise de correspondência inclui a "melhor" representação bidimensional dos dados, juntamente com as coordenadas dos pontos traçados, e uma medida (chamada inércia) da quantidade de informação retida em cada dimensão (JOHNSON, R. A.; WICHERN, D. W., 1998).

A MCA é caracterizada pela utilização de tabelas com dimensões superiores a dois ($K > 2$). Com isso, esta análise evidencia o uso de pelo menos três variáveis categóricas e suas respectivas associações.

4.5.2 Análise de Correspondência Múltipla

A MCA é uma generalização da CA, pois um número maior de variáveis ($K > 2$) estão disponíveis para os objetos ou indivíduos, logo tem-se um maior número de dimensões. Dessa forma, o uso de tabelas multidimensionais é imprescindível para organizar os objetos observados nas linhas e as categorias das variáveis nas colunas.

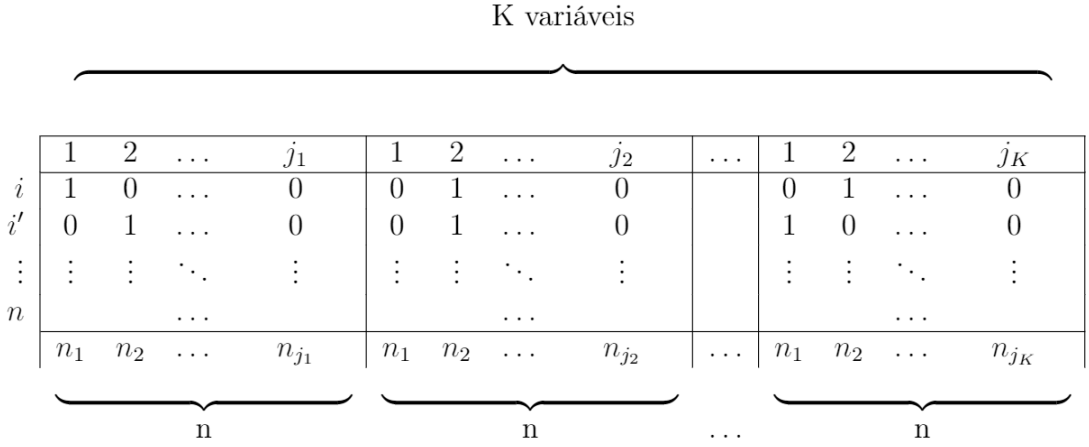
Na MCA existem duas formas comumente definidas e utilizadas para composição de uma matriz de dados para a realização da referida análise. Uma das formas é a matriz indicadora e a outra a matriz de Burt. É possível obter a matriz de Burt a partir da matriz indicadora. Segundo Naito (2007), essas duas formas possuem resultados equivalentes, portanto é possível a utilização de ambos os meios. Vale salientar que neste trabalho, a matriz de dados foi composta por meio da primeira abordagem. O uso da matriz indicadora $\mathbf{G}$ auxilia na obtenção da representação gráfica bidimensional da informação de uma tabela de contingência multidirecional (RENCHER, A. C., 2012).

A matriz indicadora é uma forma de representar os itens e categorias das variáveis que serão analisadas durante o processo de análise de correspondência múltipla, sendo que os objetos são alocados nas linhas e as categorias nas colunas. Esta matriz possui

uma linha para cada item, bem como o número de linhas é o total de itens do conjunto de dados. Para o número de colunas, temos o total de categorias considerando todas as variáveis.

Dada esta definição, temos que essa matriz é composta de 1's e 0's, ou seja, o elemento será igual a 1 se o item pertencer à categoria correspondente e será 0 se mesmo não pertencer (variável *dummy*). Portanto, temos que na matriz indicadora, o numero de 1's em cada linha é referente a quantidade de variáveis. Na Figura 1 é possível exemplificar a matriz indicadora com K variáveis.

Figura 1 – Matriz indicadora **G**.



Definindo **X** como uma matriz com dados referentes as variáveis de interesse do trabalho, temos que esta matriz é composta por x_{ij} elementos em uma tabela de contingência multidimensional com as frequências ou contagens. As linhas e colunas de **X** correspondem aos objetos observados e as diferentes categorias das variáveis. Se n é o total de frequências na matriz de dados **X**, a matriz de proporções **P** é definida como $\boldsymbol{P} = p_{ij}$, no qual divide-se cada elemento de **X** por n. Desta forma, temos que $p_{ij} = \frac{x_{ij}}{n}$, com $i = 1, 2, ..., I$ e $j = 1, 2, ..., K$.

O conceito de perfil em análise de correspondência é dado como o conjunto de frequências dividido pelo seu total. Além disso, temos definido que a soma dos vetores das linhas e colunas são, respectivamente, r_i e c_j dados por:

$$r_i = \sum_{j=1}^{K} p_{ij} = \sum_{j=1}^{K} \frac{x_{ij}}{n}$$

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}$$

em que $i = 1, 2, ..., I$ e $j = 1, 2, ..., K$.

Observe que r_i também é caracterizado como o perfil da i-ésima linha de um vetor no espaço K-dimensional, $r_i = [r_{i1}, r_{i2}, ..., r_{iK}]$. Cada r_i é afetado por um peso p_{i+}, com

$i = 1, 2, ..., I$, sendo que é proporcional aos respectivos totais de linha do conjunto de dados. No caso de c_j, tem-se que é caracterizado como o perfil da j-ésima coluna de um vetor no espaço I-dimensional, no qual $c_j = [c_{1j}, c_{2j}, ..., c_{Ij}]^T$. Cada c_j é afetado por um peso p_{+j}, com $j = 1, 2, ..., K$, sendo proporcional aos respectivos totais de coluna do dados.

Esses pesos são atrelados aos perfis - tanto linha quanto coluna - e são caracterizados como massas. As massas são proporcionais a soma dos elementos das linhas da tabela de contingência multidimensional. É definido assim para ser utilizado como termo alternativo para peso, principalmente para diferenciar a ponderação utilizada na CA em relação a ponderação geométrica (GREENACRE, M., 2008).

É possível considerar cada perfil como uma centróide, sendo caracterizada como a média ponderada dos perfis, ou seja, perfil médio de linha e perfil médio de coluna. Além disso, é viável obter o perfil médio linha como o centróide dos perfis linhas, realizando a ponderação de cada perfil com suas respectiva massa. É necessário frisar, que analogamente, o mesmo pode ser feito para os perfis colunas. As centróides de linhas e de coluna são definidas, respectivamente:

$$r_0 : \left[\frac{n_{+1}}{n_{++}}, \frac{n_{+2}}{n_{++}}, ..., \frac{n_{+K}}{n_{++}}\right]$$

$$c_0 : \left[\frac{n_{1+}}{n_{++}}, \frac{n_{2+}}{n_{++}}, ..., \frac{n_{I+}}{n_{++}}\right]^T$$

em que $n_{i+} = \sum_{j=1}^{K} x_{ij}$, $n_{+j} = \sum_{i=1}^{I} x_{ij}$ e $n_{+j} = \sum_{i=1}^{I} n_{ij}$ e n_{++} é a soma de todas as células da tabela multidimensional.

Os elementos de uma centróide apresentam a importância relativa da categoria em comparação às demais. Quando as linhas ou colunas têm perfis iguais, é dito que são distribucionalmente equivalentes e são representadas unicamente na análise gráfica.

Na questão da dispersão dos pontos, temos a introdução do conceito de inércia (total). A inércia mede as distâncias entre os perfis por meio da distância qui-quadrado, que é a distância euclidiana ponderada entre os perfis linha/coluna e os respectivos centróides. Também é caracterizada como uma média ponderada dessas distâncias entre os perfis linha e seu respectivo perfil médio. Similarmente, tem-se que o mesmo ocorre para os perfis coluna e sua média.

A inércia mede a dispersão desses perfis no espaço de perfis. Caso exista pequenas diferenças entre os perfil e sua respectiva média, a inércia possui valor próximo a zero. Além disso, proporciona a identificação dos pontos que mais contribuem para a definição das dimensões principais, permitindo agregar valor na representação dos pontos.

No âmbito da análise gráfica, temos que o *biplot* é uma representação gráfica contendo informação de uma matriz de dados. O termo refere-se aos dois tipos de informação contidos numa matriz de dados: nas linhas consta a informação pertencente as observa-

ções e nas colunas a informação pertencente as variáveis (JOHNSON, R. A.; WICHERN, D. W., 1998).

Pode ser construída para as variáveis, para as categorias, para os indivíduos e de forma conjunta para as categorias e os indivíduos. Nessa representação gráfica a dispersão dos indivíduos é dada por meio da similaridade dos indivíduos de acordo com as características análogas, e com isso, estarão mais próximas no gráfico (ROUX; ROUANET, 2004).

Neste trabalho, o uso da MCA tem por finalidade a análise multivariada dos fatores que estão associados ao desempenho dos inscritos do ENEM 2018. Dessa maneira, as variáveis de controle do participante (sexo, raça/cor), as variáveis de controle da escola (tipo de escola), as variáveis de controle das provas (notas das provas objetivas e discursiva) foram analisadas com as variáveis pertencentes ao questionário socioeconômico de inscrição para investigar a associação desses fatores quanto ao desempenho nas provas.

4.6 MCA e sparklyr

A dificuldade de se trabalhar com um grande volume de dados, ocasionou em novas estratégias para a aplicação da MCA, pois diante da limitação quanto a especificações técnicas da máquina, o desafio de analisar simultaneamente diversas variáveis trouxe à tona o grande desafio que é lidar com análise multivariada com os microdados do ENEM.

O uso dos pacotes `sparklyr` e do `dplyr` - como visto na Seção 4.3 - possibilitou o uso da memória de forma particionada, a manipulação e estruturação do conjunto de dados de forma eficiente. A realização da manipulação de dados foi útil para a construção de novos conjuntos de dados com as variáveis referentes ao perfil dos participantes, ao questionário socioeconômico e ao desempenho dos mesmos nas provas. Diante disso, essa manipulação dos dados também foi utilizada na recategorização de algumas variáveis com a finalidade de utilizar a MCA.

Para a variável renda, foi utilizado a divisão da renda familiar em classes, como por exemplo, “até 2 salários mínimos”, “de 2 a 4”, de “4 a 10” e “acima de 10” salários mínimos. Foi levado em consideração o valor do salário mínimo no referido ano, para que a divisão fosse realizada de forma correta.

Para a escolaridade do pai e da mãe ou responsável do participante, obteve-se as categorias “não estudou”, “ensino fundamental”, “ensino médio”, “ensino superior/pós graduação” e “não sei”.

No caso das notas, inicialmente foram analisados os quartis para cada tipo de prova, e com base nisso decidiu-se categorizar as notas em “sem rendimento”, “desempenho baixo”, “desempenho médio”, “desempenho alto”. O desempenho “sem rendimento” é referente aos participantes que obtiveram nota zero. A categoria “desempenho baixo” é referente aos participantes que obtiveram as 25% menores pontuações (abaixo do primeiro quartil),

enquanto que a categoria "desempenho alto" são os participantes que obtiveram as 25% maiores pontuações (acima do terceiro quartil). Em contrapartida, a categoria "desempenho médio" estão os demais participantes (entre o primeiro e terceiro quartil). Vale frisar que a categorização para cada nota foi realizada conforme os resultados dos quartis referentes a cada prova. As categorias receberam nomenclatura "CN", "CH", "LC", "MT" e "RE", sendo que correspondem, respectivamente, as provas de Ciências Naturais, Ciências Humanas, Linguagens, Matemática e Redação. O intuito da utilização destas nomenclaturas nas categorias referentes ao desempenho é para diferenciar as categorias de cada prova e para que pudesse ser evidenciado a distinção no mapa de correspondência.

Após essa primeira categorização, verificou-se que os participantes que obtiveram nota zero, estavam influenciando de forma negativa a visualização das associações no mapa de correspondência, bem como a questão da explicabilidade da variabilidade. Desta forma, decidiu-se por retirar as notas zeros para realização da análise de correspondência múltipla porque foi verificado que a inclusão das notas zero não estava contribuindo para a definição das dimensões, bem como na representação dos fatores na análise gráfica. Portanto, as categorias para a variável desempenho foram recategorizadas sem as notas zero, e com isso houve uma diminuição na quantidade de categorias para essa variável, passando de quatro para três categorias.

Após a recategorização das variáveis, o novo subconjunto de dados é composto por 11 variáveis. As variáveis são: sexo do participante, autodeclaração de raça/cor, tipo de escola, renda familiar em salários mínimos, escolaridade do pai e da mãe ou responsável, e as variáveis de desempenho das provas objetivas e discursiva.

Dado essa nova estratégia, o subconjunto de dados foi coletado e o processo da análise de correspondência múltipla foi feito utilizando a memória RAM. A decisão de se trabalhar com esta etapa da análise desta maneira se dá pelo fato de que a maioria dos trabalhos envolvendo análise de correspondência com *Spark*, utilizavam a análise de correspondência simples e não a múltipla. Outro motivo para esta decisão é a criação da *Dashboard* para os resultados. Neste caso, rodar os dados na máquina foi necessário, pois a manipulação e estruturação dos dados separadamente facilitou o desenvolvimento e processamento dos resultados na *Dashboard*, bem como demandou menos tempo do que se comparado ao uso do `sparklyr` diretamente no código da *Dashboard*.

A compreensão da utilização das novas estratégias e do novo desafio de realizar a análise de correspondência múltipla para uma base de dados com mais de 3 milhões de observações e 11 variáveis tornou o trabalho mais desafiador e interessante. O desafio de se trabalhar com os microdados do ENEM demanda conhecimento e tempo para processar cada etapa da análise de correspondência múltipla. Desta forma, foi possível obter os resultados da qualidade de representação dos fatores, da contribuição das variáveis, bem como a análise gráfica.

5 RESULTADOS

5.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A base de dados do ENEM 2018 possui mais de $5,5$ milhões de linhas e 137 colunas, tendo o respectivo arquivo tamanho maior que $3,3$ milhões de *kilobytes*, correspondendo $3,35$ *gigabytes*. Após realizar o processamento dos dados e a filtragem das variáveis de interesse, o número de linhas diminuiu para $3,38$ milhões, e o número de colunas reduziu para 55. Entretanto, a nova base de dados não foi coletada e armazenada pela função `copy_to()` do pacote `dplyr` por permanecer sendo uma massiva quantidade de observações, inviabilizando as análises sem o uso do *framework Spark*.

Após o pré-processamento da base de dados efetuou-se a análise descritiva e exploratória com a finalidade de verificar o perfil dos inscritos no ENEM 2018. Os resultados dessa análise serão apresentados a seguir.

Figura 2 – Perfil dos participantes no ENEM 2018.

De acordo com a Figura 2, podemos observar o perfil dos participantes no ENEM 2018. A maioria dos participantes no exame são do sexo feminino. Temos que apenas 12% dos participantes são treineiros - candidatos inscritos que realizam a prova com o objetivo de treinar e obter experiência para os próximos anos - e referente aos 12%, a maioria dos treineiros são do sexo masculino $(7,5\%)$.

Em relação a idade dos participantes, observa-se que em média possuem $22,4$ anos. Além disso, o participante mais novo tem idade igual a 10 anos, enquanto que o participante mais velho tem idade igual a 98 anos. Vale salientar que os valores abaixo de 10

anos e acima de 100 anos não são considerados na base de dados disponível no *website* do INEP.

Considerando a raça/cor dos participantes, temos que grande parte dos participantes no ENEM 2018 se autodeclaram pardos. Analisando o tipo de escola, a maioria dos participantes estudam/estudaram em escolas públicas de ensino. Vale frisar também que apenas $0,04\%$ responderam que não frequentam a escola.

Um dos temas que compõem o questionário socioeconômico presente na inscrição é o de acesso a internet na respectiva residência do inscrito. É importante analisar essa situação, pois a falta ao acesso a internet e a tecnologia como um todo é uma condição de desigualdade socioeconômica e pode estar atrelada a outras condições como ensino/estudo e possivelmente desempenho. Diante disso, temos que aproximadamente 25% dos participantes não possuem acesso a internet, no qual evidencia que uma parcela dos inscritos ainda não utilizam a internet como ferramenta de estudo/pesquisa, sendo uma realidade da desigualdade socioeconômica do país.

Com o intuito de analisar a distribuição dos participantes no ENEM no referido ano em relação ao nível de escolaridade do pai e da mãe (ou respectivos responsáveis), temos que tanto a mãe quanto o pai completaram o ensino médio (EM), mas não completaram a faculdade. Além disso, observa-se que as mães possuem um nível educacional maior do que o dos pais. No entanto, pode-se frisar que $2,83\%$ e $8,36\%$ responderam que não sabem o nível de escolaridade das mães e dos pais, respectivamente.

Tabela 1 – Medidas resumo das notas por provas.

Provas	**Média**	**Máximo**	**Desvio Padrão**	**CV (%)**
C. Naturais	494	870	74,1	15,0
C. Humanas	571	850	79,2	13,9
Linguagens	529	817	72,3	13,7
Matemática	535	996	103	19,3
Redação	516	1000	184	35,7

Na Tabela 1, observa-se que na prova de Ciências Naturais a nota média foi igual a 494, sendo que a nota máxima na referida prova é igual a 800. No caso da prova de Ciências Humanas, a nota máxima dos alunos é 850, tendo o aluno médio obtido nota igual a 571. Vale salientar que a média da prova de Ciências Humanas é a maior em relação as demais provas.

Na prova de Linguagens, o aluno médio atingiu nota igual a 529, enquanto que a nota máxima é igual a 817. Já na prova de Matemática, a nota máxima obtida pelos participantes é igual a 996, sendo a maior nota se comparado as demais provas, exceto a prova de Redação.

Agora analisando o desempenho na prova de Redação, observa-se que os participantes cumpriram todas as competências exigidas da referida prova e, consequentemente,

obtiveram nota máxima. O aluno médio logrou nota igual a 516.

Todas as provas tiveram participantes sem rendimento, ou seja, participantes que obtiveram nota zero. Para completar esses resultados temos os percentuais de notas zero nas referidas provas: Ciências Naturais (0,015%), Ciências Humanas (0,096%), Linguagens (0,037%), Matemática (0,019%) e Redação (2,069%). Então, o percentual de zero nas provas objetivas e discursiva corresponde a 2,24% dos participantes, sendo que um mesmo participante pode ter ficado sem rendimento em mais de uma prova.

Quanto a variação das notas em relação à média, observa-se que as notas de todas as provas - exceto a prova de Redação - possuem coeficientes de variação baixos, ou seja, essas notas são mais homogêneas em torno da média. No caso da prova de Redação, o coeficiente de variação é maior do que as demais provas e essas notas variam bastante em torno da média.

Tabela 2 – Quartis das notas das provas do ENEM 2018 segundo o sexo.

Provas	**Sexo**	$\boldsymbol{Q_1}$	$\boldsymbol{Q_2}$	$\boldsymbol{Q_3}$
Ciências Naturais	Feminino	433	478	536
	Masculino	441	491	531
Ciências Humanas	Feminino	510	578	624
	Masculino	518	595	638
Linguagens	Feminino	475	530	578
	Masculino	483	539	585
Matemática	Feminino	446	501	573
	Masculino	474	544	640
Redação	Feminino	360	520	640
	Masculino	360	520	620

*Q_1, Q_2 e Q_3: 1^{o}, 2^{o} e 3^{o} quartis.

Na Tabela 2 estão apresentados os quartis das notas das provas segundo o sexo dos participantes do ENEM 2018. Os resultados desta tabela também serão úteis para avaliar os *boxplots* das notas de acordo com o sexo dos participantes. Diante desta tabela, observa-se que 50% das alunas obtiveram nota até 478 na prova de Ciências Naturais, enquanto que entre os alunos esse valor foi até 491. Com isso, percebe-se que o desempenho dos alunos é 13 pontos maior do que o desempenho das alunas.

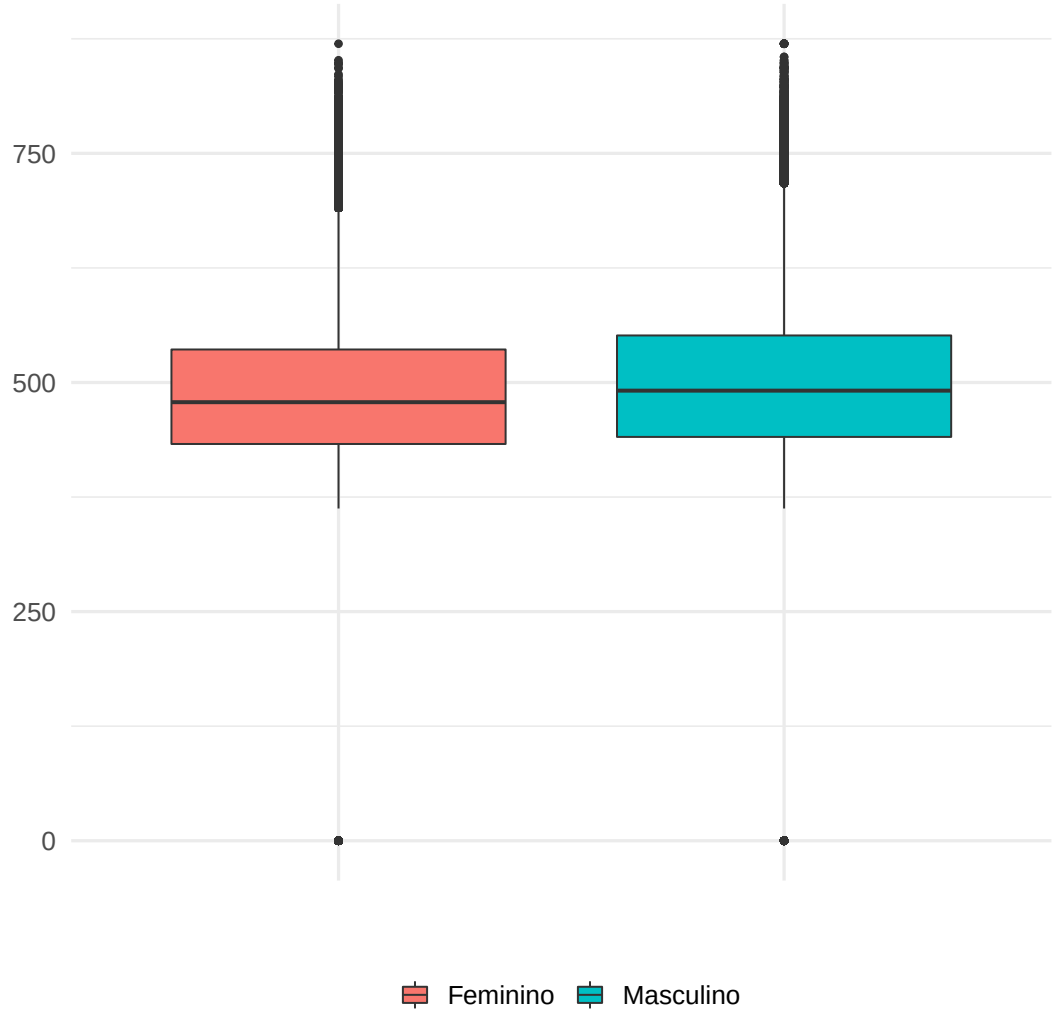
Observa-se também que o mesmo ocorre nas provas de Ciências Humanas, Linguagens e Matemática, pois 50% dos alunos obtiveram notas até 595, 539 e 544, respectivamente. Com isso, tem-se que os participantes do sexo masculino obtiveram 17 pontos de diferença para a prova de Ciências Humanas, enquanto que para a prova de Linguagens,

a diferença é de 9 pontos. No caso da prova de Matemática observa-se uma diferença de 43 pontos, sendo a maior diferença entre os participantes do sexo masculino e feminino nas provas objetivas. No caso dos demais quartis, nota-se que os participantes do sexo masculino também obtiveram notas melhores nas provas objetivas em comparação os participantes do sexo feminino.

Em contraste, na prova de Redação, percebe-se que 50% dos participantes tanto do sexo feminino quanto do sexo masculino obtiveram notas até 520 na prova de redação no ENEM 2018. O mesmo ocorre para o primeiro quartil, ou seja, 25% dos participantes de ambos os sexos obtiveram notas até 360 na prova discursiva.

É de referir que estamos considerando como "melhores alunos", os alunos que obtiveram notas acima do percentil 75. Desta forma, considera-se que 25% dos participantes lograram melhor desempenho nas provas em relação aos demais nas respectivas análises desta seção.

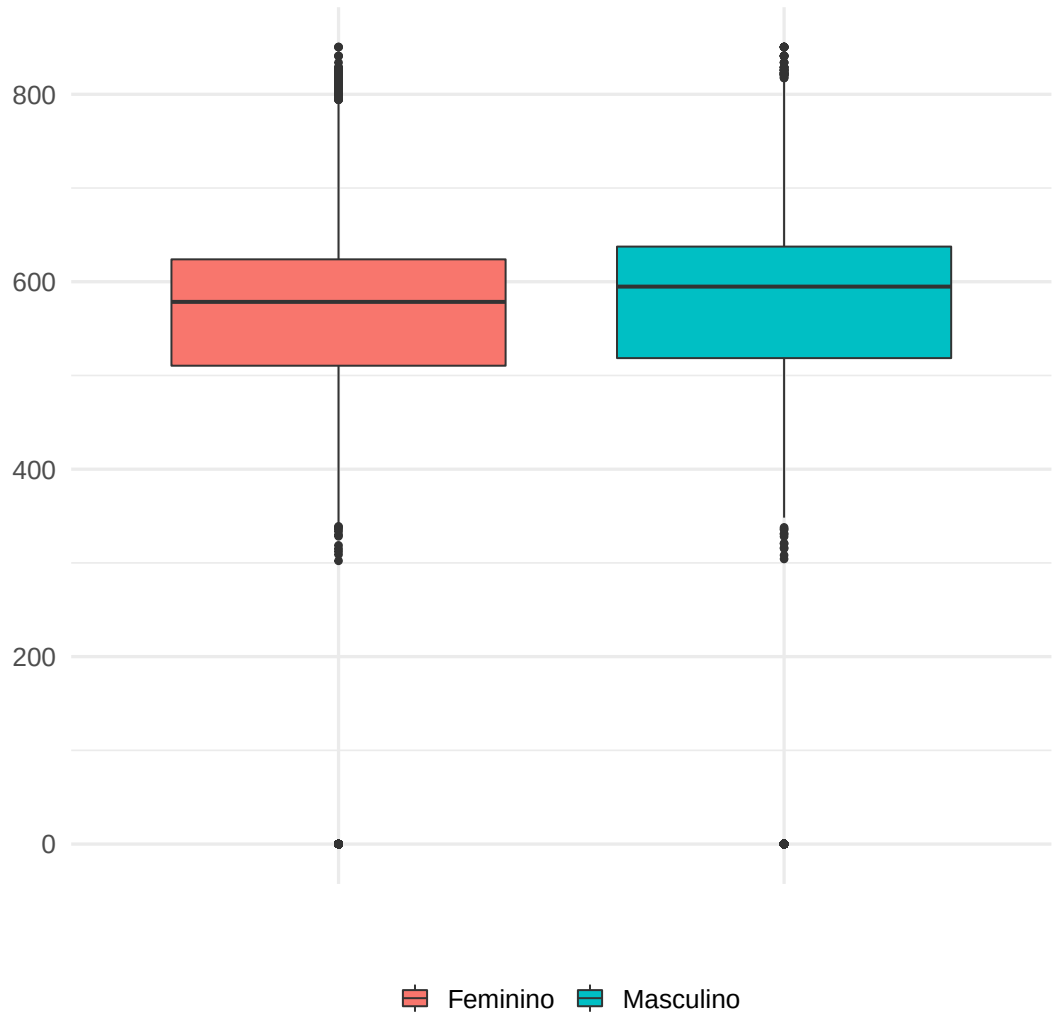
Figura 3 – Boxplot das notas de Ciências Naturais dos inscritos no ENEM 2018 segundo o sexo.



Na Figura 3, observa-se que as notas na prova de Ciências Naturais são assimétricas negativas, pois 50% dos participantes obtiveram notas próximas das 25% menores pontuações (1^{o} quartil). Quanto a variabilidade das notas, observa-se que a variabilidade das notas dos participantes do sexo masculino é levemente maior do que dos participantes do sexo feminino.

As notas acima de 700 são valores discrepantes, as quais destacam os alunos com desempenho próximos das notas máximas de ambos os sexos. Além disso, os melhores alunos obtiveram desempenho acima de 536, no caso dos participantes do sexo masculino, e acima de 531 para os participantes do sexo feminino. Vale notar que a nota zero representa a nota mínima na referida prova, bem como representa os participantes sem rendimento para os dois grupos.

Figura 4 – Boxplot das notas de Ciências Humanas dos participantes no ENEM 2018 segundo o sexo.

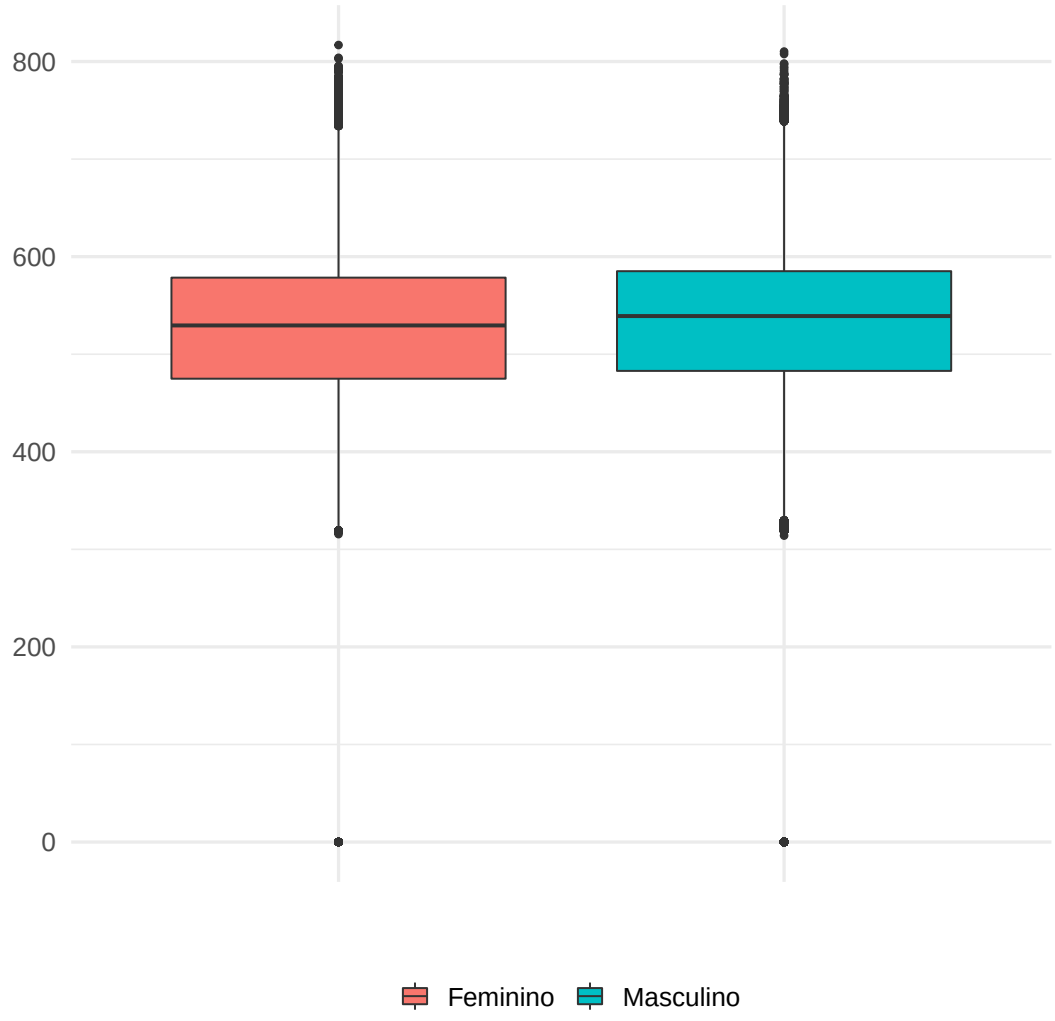


De acordo com a Figura 4, observa-se que a simetria das notas na prova de Ciências

Humanas é assimétrica positiva, ou seja, 50% dos participantes obtiveram notas próximas das 75% maiores pontuações (3º quartil). Além disso, observa-se que não há diferença expressiva nas notas em ambos os grupos. No caso da variabilidade, aparentemente há uma maior variabilidade das notas para os participantes do sexo masculino.

Quanto as notas discrepantes, observa-se que as notas encontram-se acima de 800, representando os alunos com notas próximas da nota máxima na prova de Linguagens. Outra situação é que temos notas discrepantes abaixo de 400, sendo que 25% dos participantes obtiveram desempenho até 510 para os alunos, e 518 para as alunas. Percebe-se então que essas notas abaixo de 400 evidenciam um desempenho baixo na respectiva prova.

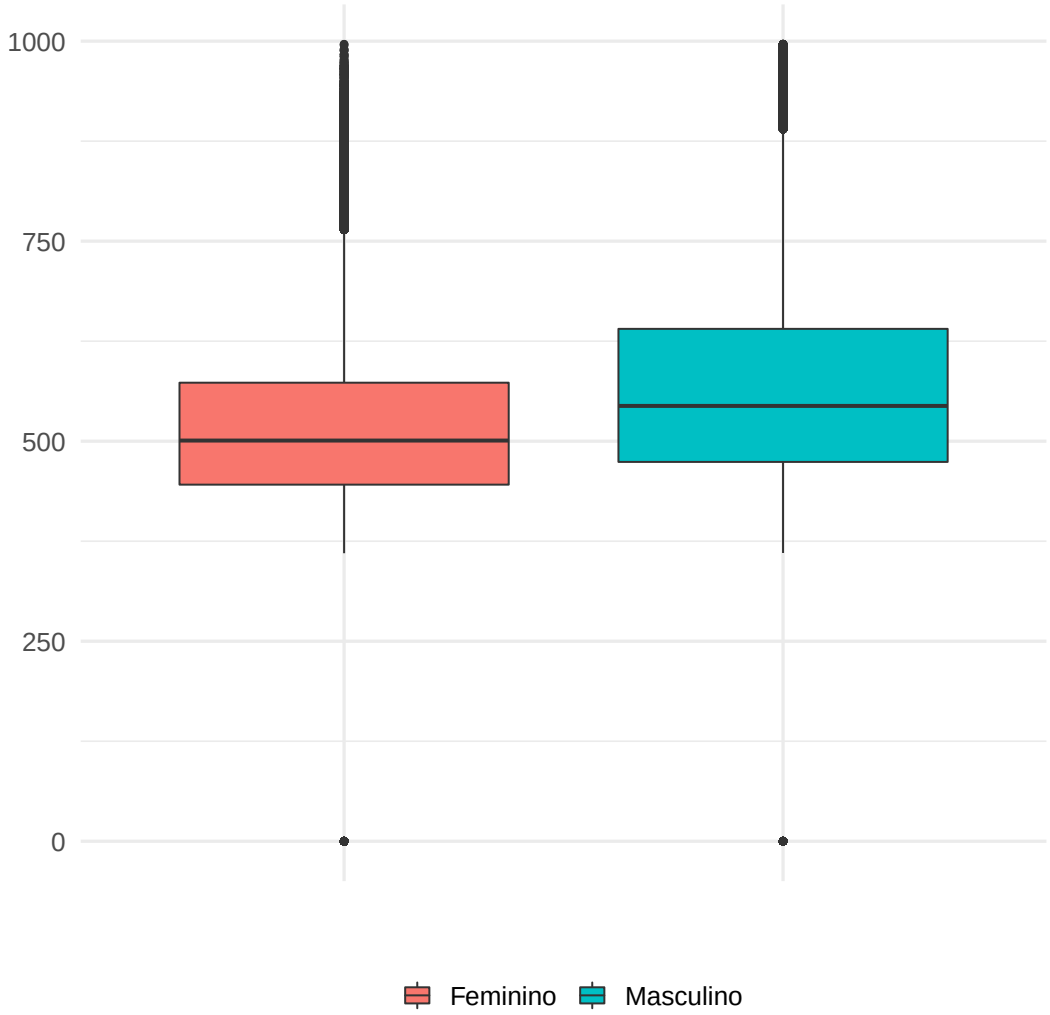
Figura 5 – Boxplot das notas de Linguagens dos participantes no ENEM 2018 segundo o sexo.



No caso das notas da prova de Linguagens - ver Figura 5 - observa-se que a simetria das notas é assimétrica negativa, enquanto que aparentemente não há uma diferença

expressiva quanto a variabilidade entre as notas dos participantes do sexo masculino e do sexo feminino. As notas discrepantes encontram-se acima da nota 700, nos quais são os participantes com notas próximas da nota máxima 817.

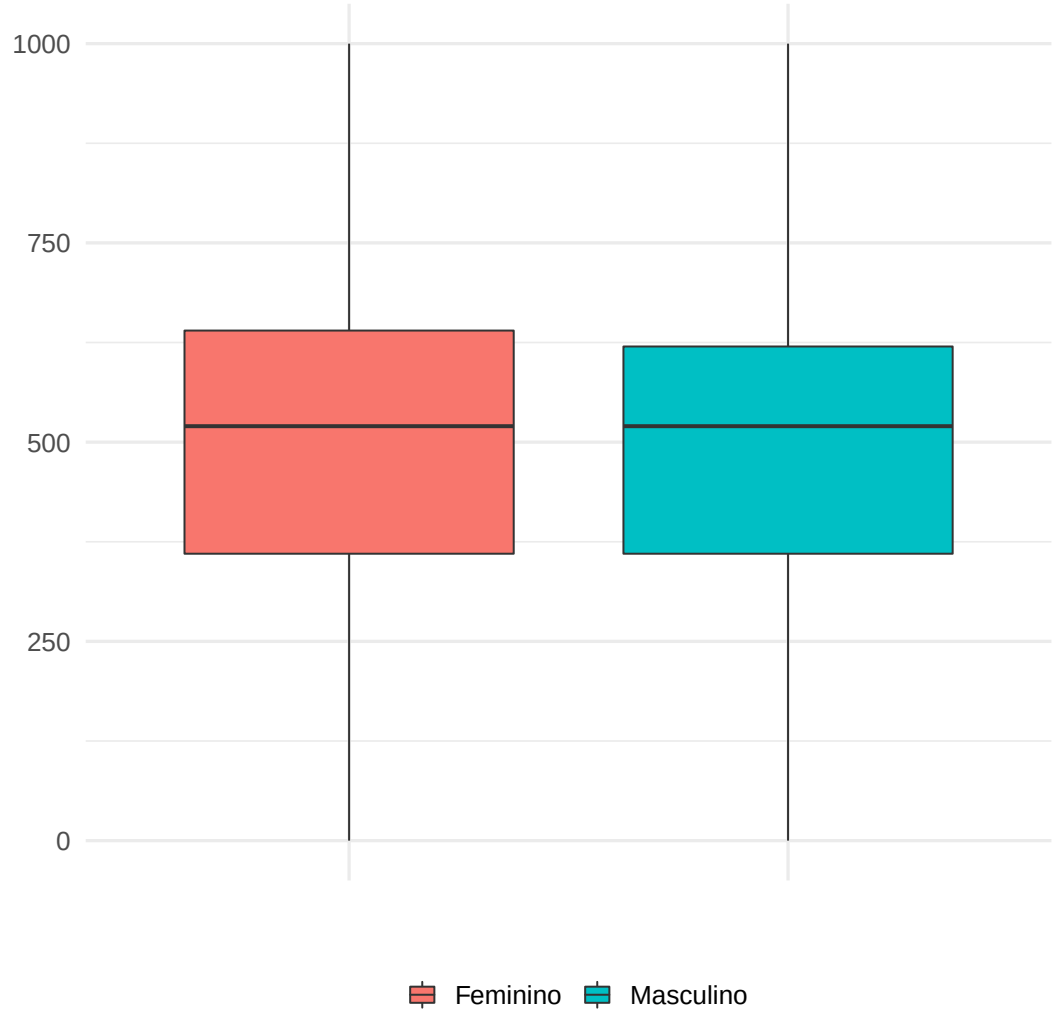
Figura 6 – Boxplot das notas de Matemática dos participantes no ENEM 2018 segundo o sexo.



Na prova de Matemática - ver Figura 6 - a simetria das notas dos participantes é assimétrica negativa, pois 50% das notas encontram-se mais próximas das 25% menores pontuações (primeiro quartil). Observa-se que há uma maior variabilidade das notas nos participantes do sexo masculino em relação ao do sexo feminino.

Vale frisar que os melhores alunos obtiveram notas acima 573 para as alunas e acima de 640 para os alunos. Ao comparar esses resultados, percebe-se que as notas dos alunos são destoantes das alunas. Os valores discrepantes foram próximos da nota 966, que consequentemente são dos participantes que obtiveram desempenho próximo da nota máxima.

Figura 7 – Boxplot das notas de Redação dos participantes no ENEM 2018 segundo o sexo.



De acordo com a Figura 7, nota-se que a simetria das notas dos participantes na prova de Redação do ENEM 2018 é assimétrica negativa.

Diferente das provas objetivas, a prova de Redação apresenta uma variabilidade das notas para os participantes do sexo feminino. Outra observação é que a variabilidade na referida prova se comparado com as demais provas. Além disso, 50% das notas dos participantes tanto do sexo feminino quanto do sexo masculino apresentam valores muito próximos. Os melhores alunos obtiveram notas acima da nota 640, bem como não há notas discrepantes para a referida prova.

Na Tabela 3, estão apresentados os quartis das notas segundo a raça/cor dos participantes do ENEM 2018. As informações contidas nesta tabela serão úteis para avaliar os *boxplots* das notas segundo a raça/cor dos participantes.

Tabela 3 – Quartis das notas das provas do ENEM 2018 segundo a raça/cor.

Provas	**Raça/Cor**	Q_1	Q_2	Q_3
Ciências Naturais	Parda	429	472	527
	Preta	429	470	522
	Branca	450	505	566
	Amarela	436	484	545
	Indígena	418	455	504
	Não declarado	441	496	562
Ciências Humanas	Parda	500	570	619
	Preta	500	571	618
	Branca	542	606	645
	Amarela	510	581	627
	Indígena	475	540	598
	Não declarado	524	600	647
Linguagens	Parda	467	519	568
	Preta	468	520	567
	Branca	501	556	599
	Amarela	478	531	578
	Indígena	445	494	544
	Não declarado	485	548	599
Matemática	Parda	448	503	576
	Preta	444	496	564
	Branca	474	546	646
	Amarela	456	518	606
	Indígena	435	482	542
	Não declarado	462	530	632
Redação	Parda	360	500	600
	Preta	360	500	600
	Branca	400	560	680
	Amarela	360	520	620
	Indígena	340	440	560
	Não declarado	380	540	640

*Q_1, Q_2 e Q_3: 1º, 2º e 3º quartis.

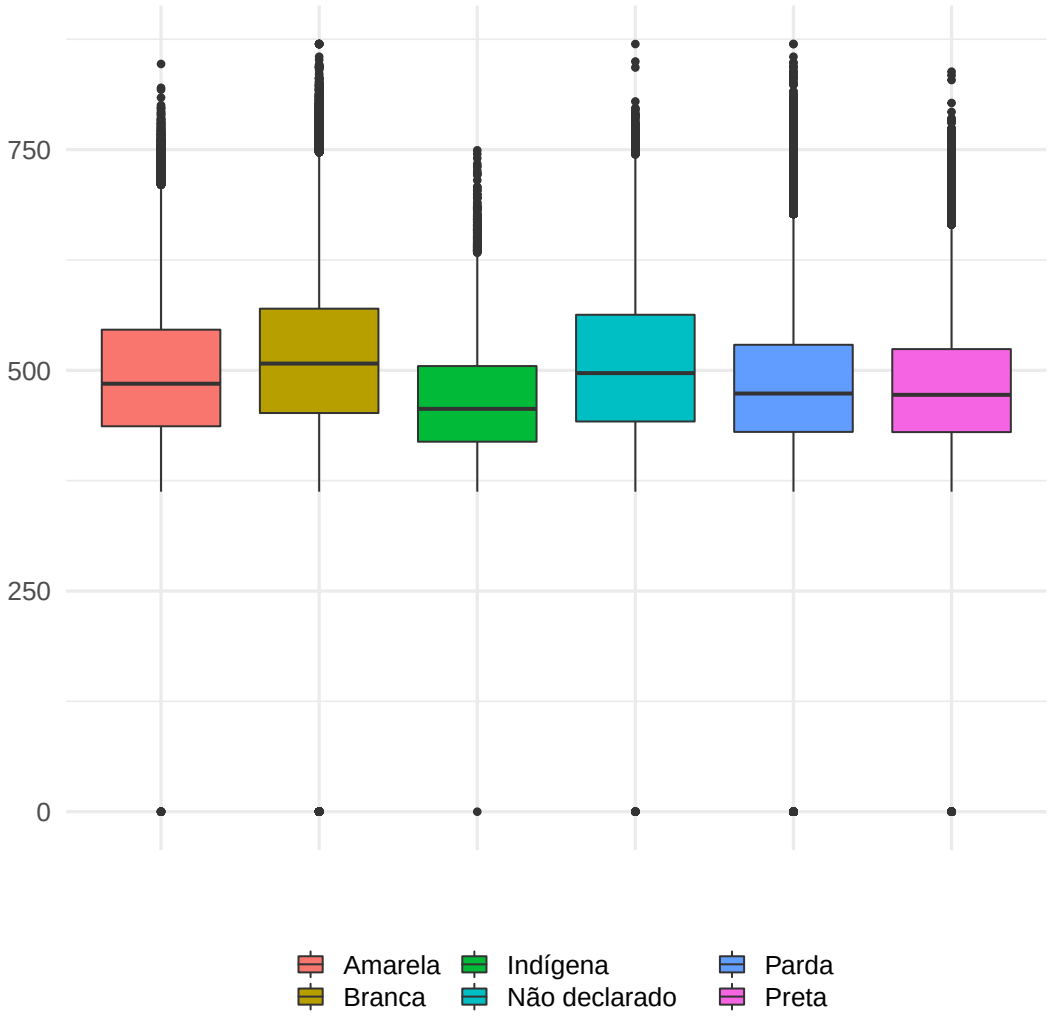
Nota-se que 25% dos participantes autodeclarados negros e autodeclarados pardos obtiveram nota até 500 na prova de Ciências Humanas, sendo esse um desempenho 25 pontos maior do que o menor desempenho que é dos participantes autodeclarados indígenas (475) na referida prova. Além disso, observa-se que ambos os grupos possuem notas com pontuações muito próximas ou iguais. Observe no caso da prova de Ciências Naturais, Ciências Humanas e Linguagens, a diferença entre os dois grupos é de no máximo dois

pontos. Para a prova de Redação, 50% dos participantes de ambos os grupos obtiveram notas até 500.

No caso dos participantes autodeclarados brancos, observa-se que 50% das notas desses participantes são maiores do que as notas dos demais grupos em todas as provas. O mesmo ocorre para os demais quartis, o que evidencia um melhor desempenho desse grupo se comparado aos demais. Vale frisar que os participantes que não declararam sua respectiva raca/cor, obtiveram desempenho próximo aos participantes autodeclarados brancos.

Para os participantes autodeclarados indígenas, há uma diferença expressiva em relação aos quartis das notas. Observa-se que em todos os quartis, as notas desses participantes são menores que os demais grupos. Por exemplo, na prova de Linguagens, 50% dos participantes obtiveram notas até 494, sendo que em relação aos demais participantes há uma diferença de até 62 pontos.

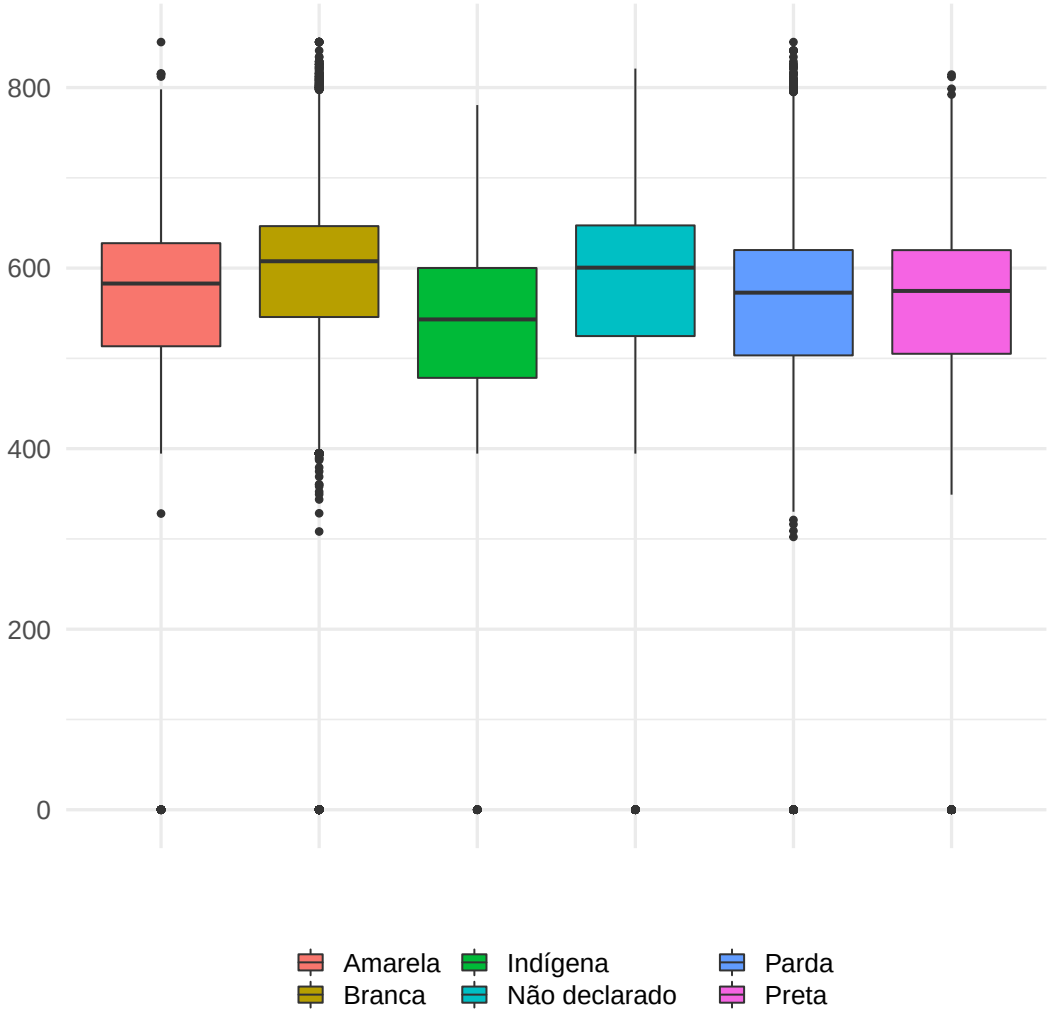
Figura 8 – Boxplot das notas de Ciências Naturais dos participantes no ENEM 2018 segundo a raça/cor.



Quanto a simetria das notas dos participantes dos grupos segundo a autodeclaração da raça/cor, ver Figura 8, observa-se que as notas na prova de Ciências Naturais são assimétricas positivas, ou seja, as notas estão mais próximas das 25% menores notas (primeiro quartil). Além disso, há uma maior variabilidade das notas para os participantes autodeclarados brancos e os não autodeclarados.

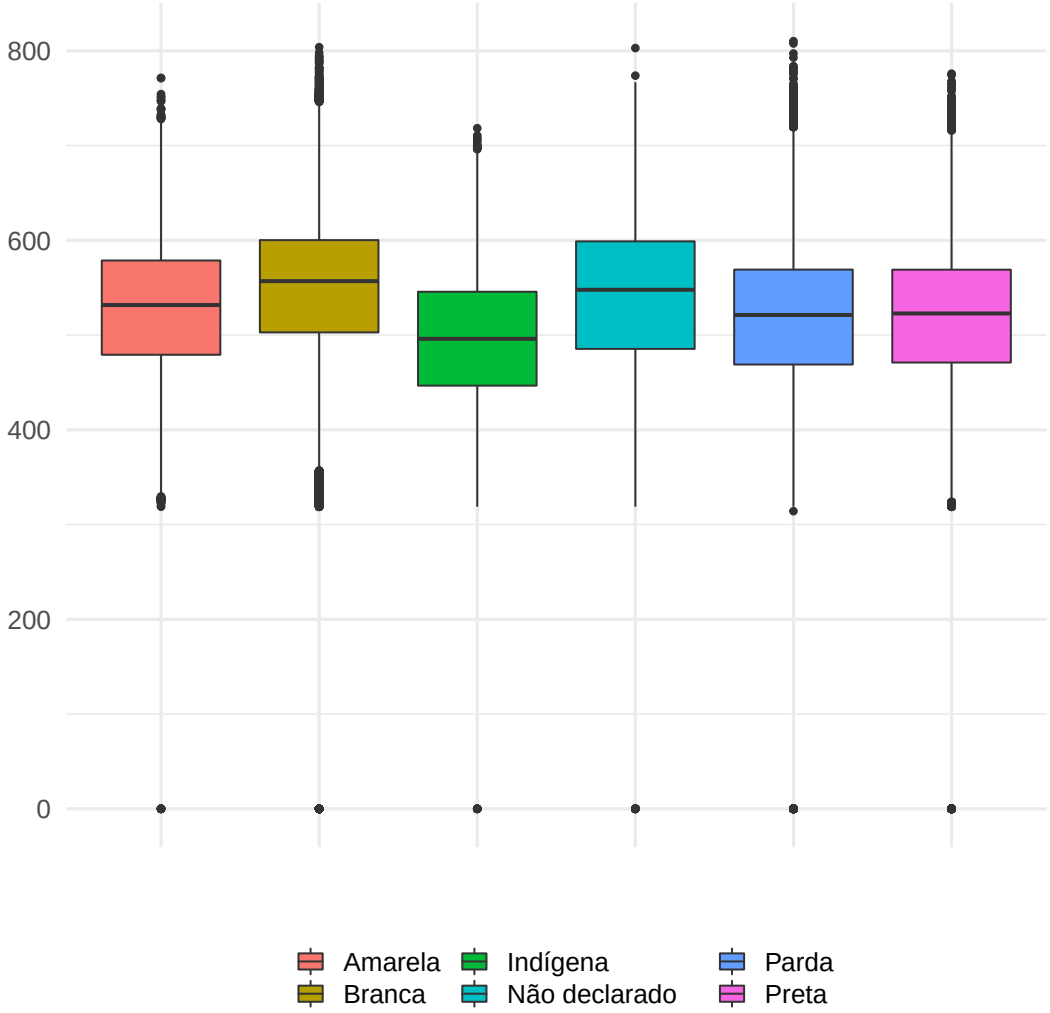
Em todos os grupos há participantes com desempenho sem rendimento, ou seja, obtiveram nota zero na prova.

Figura 9 – Boxplot das notas de Ciências Humanas dos participantes no ENEM 2018 segundo a raça/cor.



Na Figura 9, observa-se que a simetria das notas em todos os grupos é assimétrica negativa. Com isso, as notas estão mais próximas das 25% maiores notas (terceiro quartil). No caso da variabilidade, as notas variam mais nos grupos autodeclarados brancos e não declarados.

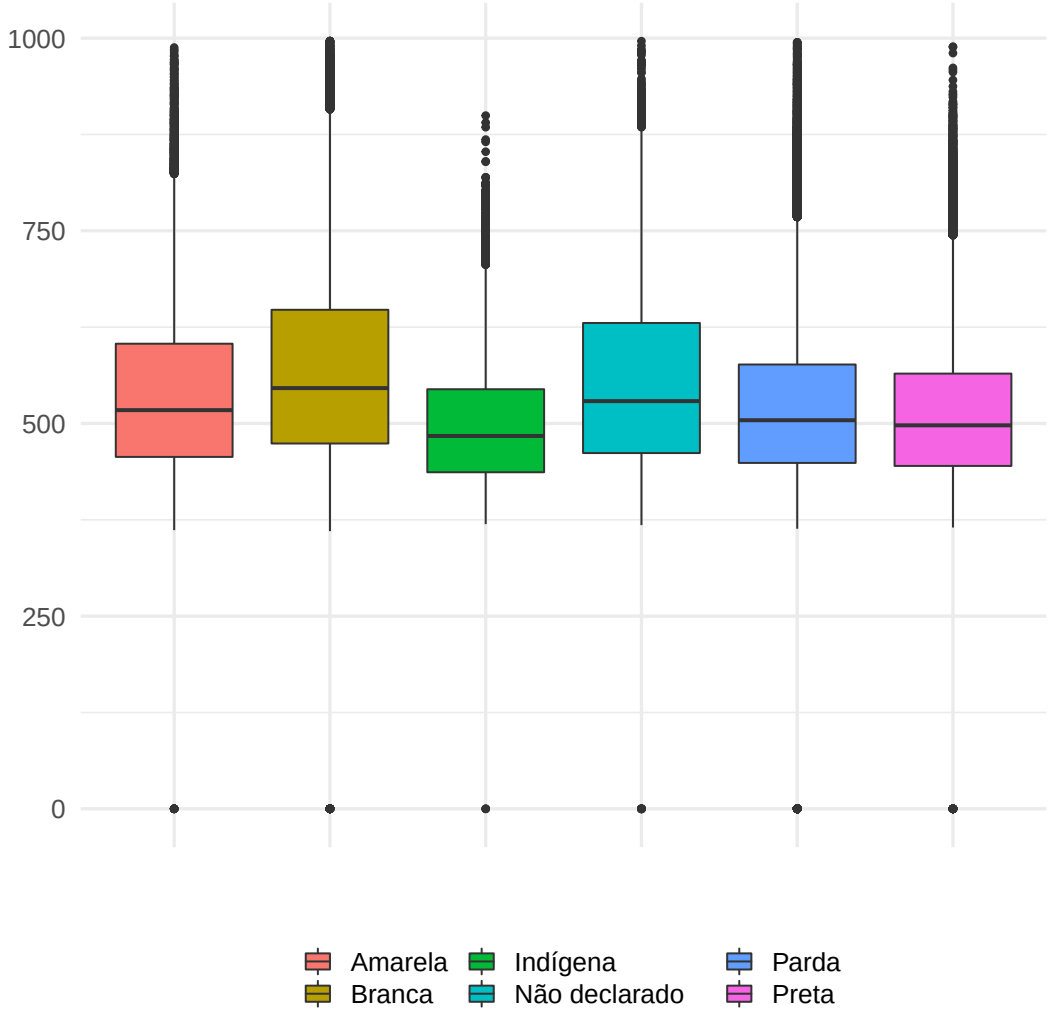
Figura 10 – Boxplot das notas de Linguagens dos participantes no ENEM 2018 segundo a raça/cor.



De acordo com a Figura 10, a simetria das notas da prova de Linguagens para os grupos autodeclarados pretos ou pardos aparenta ser simétrica, enquanto que os demais grupos apresentam assimetria negativa. Considerando a simetria para os participantes autodeclarados negros e autodeclarados pardos, observa-se uma diferença da prova de Linguagens em comparação as provas de Ciências Naturais e Ciências Humanas. Em contrapartida, o grupo que possui maior variabilidade nas notas é o grupo referente aos participantes não autodeclarados segundo a raça/cor.

Com base na Figura 11, as notas na prova de Matemática são assimétricas positivas, e vale notar que há notas discrepantes concentradas acima da nota 650. Vale salientar que essas notas extremas estão próximas da nota máxima 996, evidenciando um melhor aproveitamento da prova e consequentemente um melhor desempenho. A maior variabilidade das notas na referida prova está nos grupos autodeclarados brancos, autodeclarados amarelos e não autodeclarados.

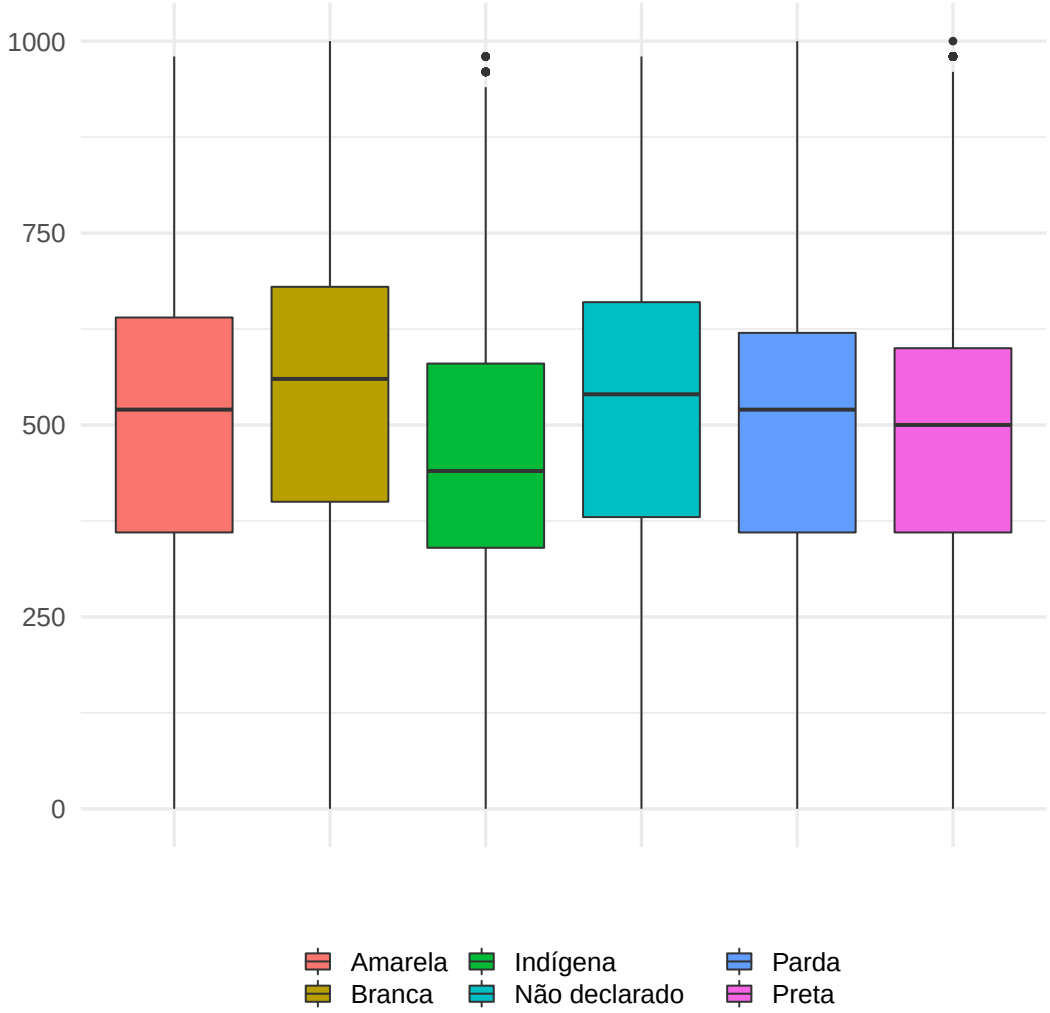
Figura 11 – Boxplot das notas de Matemática dos participantes no ENEM 2018 segundo a raça/cor.



A Figura 12 apresenta o *boxplot* das notas da prova de Redação dos participantes autodeclarados segundo a raça/cor. As notas da prova de Redação são assimétricas negativas, pois estão mais próximas das 25% das maiores notas (terceiro quartil). Como a prova de redação possui uma grande variabilidade de notas, cada grupo possui uma grande diferença de pontos entre si. Por exemplo, 75% dos participantes que não declararam sua respectiva raça/cor tem desempenho inferior 40 pontos se comparado aos participantes considerados brancos (ver Figura 3).

Esta variabilidade das notas entre os grupos na prova de Redação pode ser um resultado que esteja ligado às questões socioeconômicas - como por exemplo: tipo de escola, escolaridade do pai e da mãe e renda familiar - atreladas aos grupos pretos, pardos e indígenas.

Figura 12 – Boxplot das notas de Redação dos participantes no ENEM 2018 segundo a raça/cor.



Nesta parte da análise, temos os mapas de distribuição dos participantes de acordo com o desempenho nas provas objetivas e na redação no ENEM 2018 por estado. Cada mapa é apresentado conforme uma cor específica e a intensidade da tonalidade da cor no mapa representa o maior desempenho dos participantes do estado. Em contraste a isso, os estados que apresentam menos intensidade da tonalidade da cor representam menor desempenho dos participantes de cada estado.

Na Figura 13, observa-se que os estados de São Paulo e Santa Catarina ilustram uma maior intensidade da cor verde no mapa. Dessa forma, temos que os participantes desses estados obtiveram maior desempenho médio na referida prova. Há alguns resultados a serem notados além das regiões Sul-Sudeste. Pode-se observar que o estado que destaca-se na região Norte é Roraima, enquanto que Goiás e Mato Grosso do Sul destacam-se na região Centro-Oeste. Já na região Nordeste, o estado que destaca-se é o Rio Grande do Norte.

Figura 13 – Mapa da distribuição dos participantes de acordo com o desempenho na prova de Ciências Naturais do ENEM, Estados, 2018.

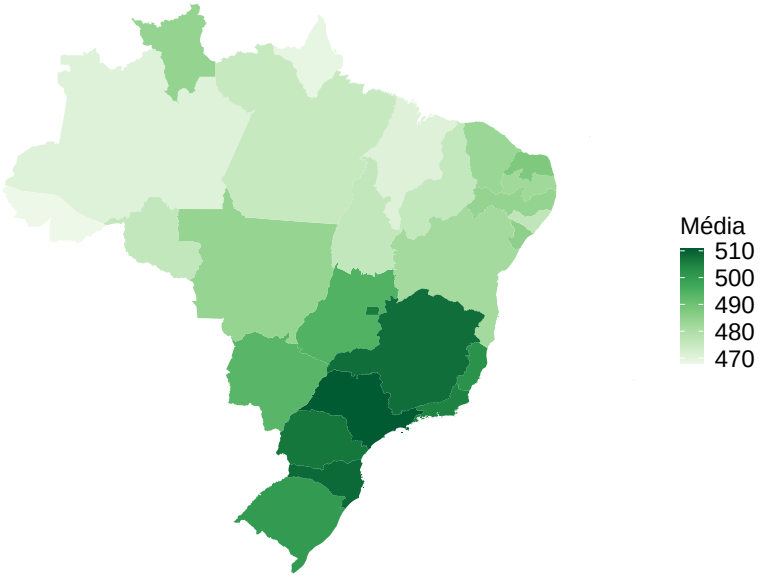
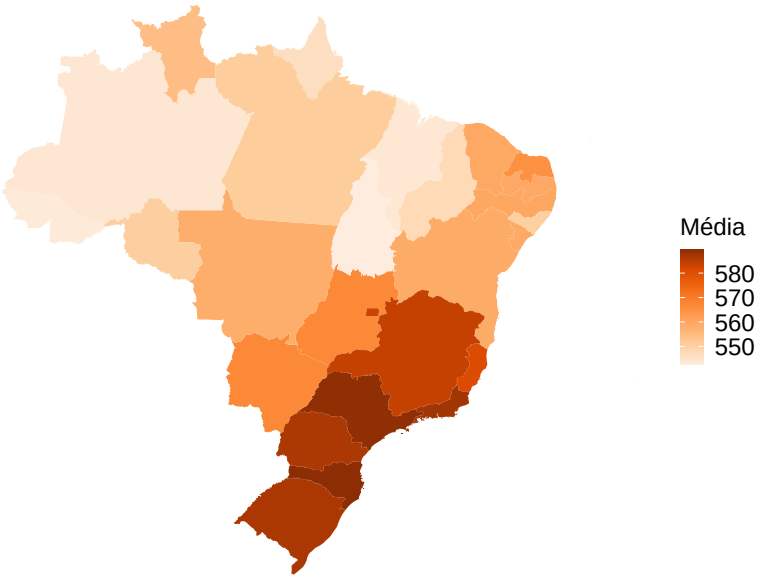


Figura 14 – Mapa da distribuição dos participantes de acordo com o desempenho na prova de Ciências Humanas do ENEM, Estados, 2018.

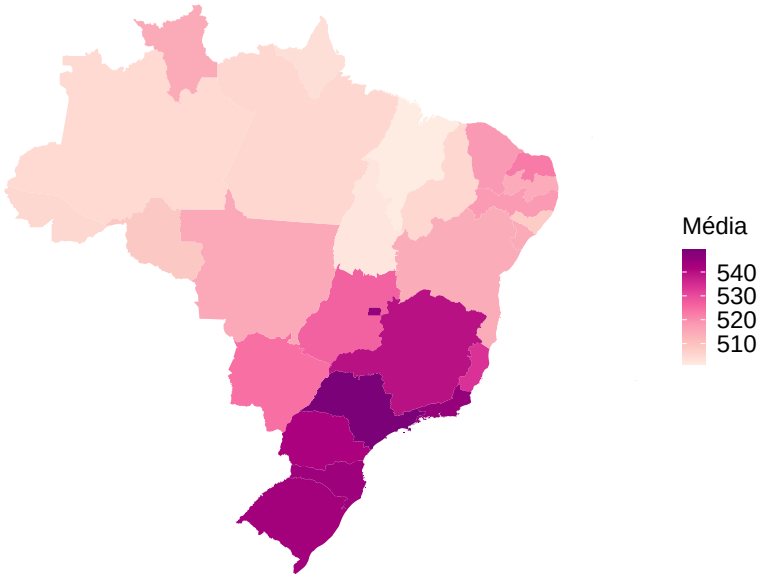


Na Figura 14, observamos o desempenho médio na prova de Ciências Humanas no exame, no qual os estados de São Paulo e Santa Catarina destacam-se conforme ilustra a intensidade da cor laranja no mapa. Os participantes desses estados obtiveram as maiores médias na referida prova.

Podemos perceber que os estados que compõem a região Sul estão com maior intensidade de cor, indicando maiores médias se comparado aos outros estados do país. Em contrapartida, os estados da região Norte possuem o menor desempenho na prova de Ciências Humanas. Vale salientar que o estado de Roraima possui maior intensidade de cor - maior desempenho - se comparado aos dos demais estados da região Norte.

Outro resultado interessante é a quantidade de participantes que obtiveram nota máxima por estado na provas de Ciências Naturais e Humanas. Na prova de Ciências Naturais, seis participantes do estado de São Paulo obtiveram nota máxima. Já na prova de Ciências Humanas, os estados de Minas Gerais e São Paulo empatam cada um com três participantes com nota máxima.

Figura 15 – Mapa da distribuição dos participantes de acordo com o desempenho na prova de Linguagens do ENEM, Estados, 2018.



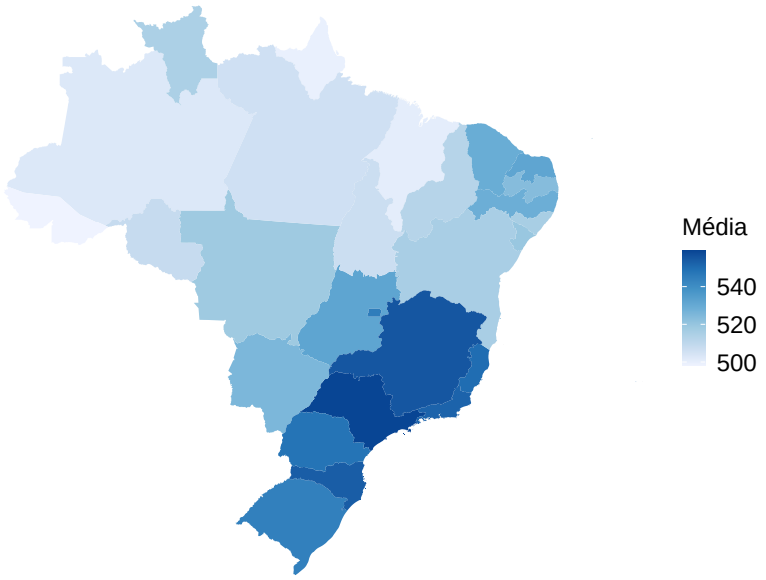
Conforme a Figura 15, observa-se o mapa de distribuição dos participantes de acordo com o desempenho na prova de Linguagens. Percebe-se que há uma maior intensidade/destaque da cor roxa nos estados de São Paulo e Rio de Janeiro, bem como no Distrito Federal.

Mesmo que os estados de Minas Gerais e Espírito Santo tenham médias um pouco menores se comparadas a São Paulo e Rio de Janeiro, o desempenho na referida prova ainda é maior do que o desempenho dos participantes dos estados da Região Nordeste.

Na região Norte, o estado de Roraima possui maior intensidade da cor roxa em relação aos demais estados da região, evidenciando um maior desempenho dos participantes desse estado.

Uma outra consideração é que apenas o estado de Pernambuco teve participante com nota máxima na prova de Linguagens.

Figura 16 – Distribuição dos participantes de acordo com o desempenho na prova de Matemática do ENEM, Estados, 2018.



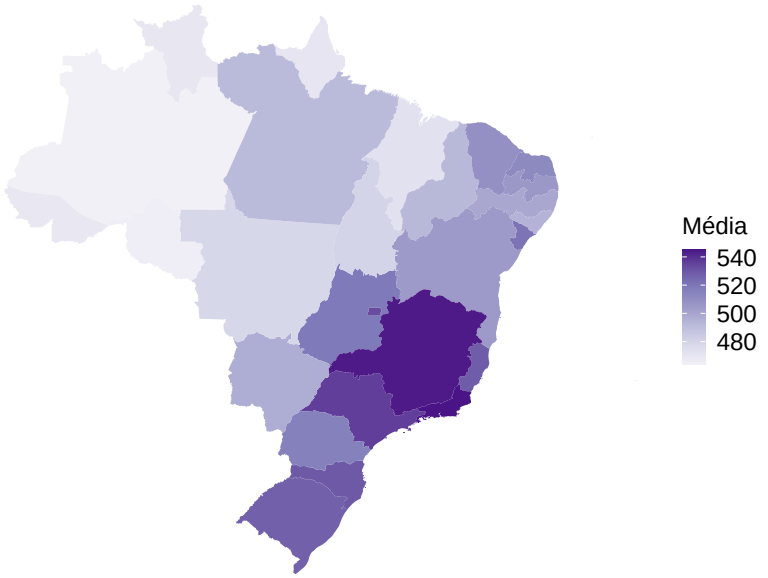
De acordo com a Figura 16, observa-se o mapa da distribuição dos participantes segundo o desempenho na prova de matemática. Diante disso, pode-se observar que os estados com maior intensidade da cor azul são os da região Sudeste, principalmente São Paulo e Minas Gerais. Assim, tem-se que os participantes com os melhores desempenhos estão concentrados nos estados de São Paulo e Minas Gerais.

Em outra perspectiva, na região Norte concentra-se a parte com menos intensidade da cor azul, evidenciando a distribuição dos participantes com os menores desempenhos na prova. Mais uma vez, o estado de Roraima apresenta maior intensidade na cor azul na região Norte. Então, temos que os participantes desse estado apresentam maior desempenho em relação aos demais estados da referida região.

Nota-se também que os estados Ceará, Rio Grande do Norte e Pernambuco evidenciam um pequeno destaque da intensidade da cor na região Nordeste, ou seja, os participantes desses estados obtiveram melhor desempenho na prova de Matemática entre os participantes da região Nordeste.

Os estados Ceará, Pernambuco e São Paulo contaram com o maior número de alunos que obtiveram nota máxima na prova de matemática, com cada estado contendo quatro alunos com nota próxima de 1000.

Figura 17 – Distribuição dos participantes de acordo com o desempenho na prova de Redação do ENEM, Estados, 2018.



Quanto ao desempenho da prova de redação no ENEM 2018, ver Figura 17, os estados de Minas Gerais e Rio de Janeiro, destacam-se com maior intensidade da cor roxa no mapa. Entretanto, vale notar que o estado de São Paulo também possui uma maior intensidade de coloração, indicando também um ótimo desempenho médio.

Podemos citar os estados de Santa Catarina e Rio Grande de Sul, pois destacam-se quanto a intensidade da cor roxa na região Sudeste. Nota-se também que o estado de Sergipe possui maior intensidade da referida cor na região Nordeste. Portanto, os alunos desses estados possuem melhor desempenho se comparado aos outros alunos dos estados pertencentes das regiões Sudeste e Nordeste, respectivamente.

Em todos os outros mapas das provas objetivas, o estado de Roraima demonstrou maior intensidade de cor e portanto melhor desempenho do que os demais estados da região Norte. Entretanto, na prova discursiva o estado que revela maior intensidade de

cor é o Pará, e com isso tem-se que os alunos desse estado obtiveram desempenho um pouco melhor do que os demais participantes da região Norte.

Há um destaque para os estados de Minas Gerais e Rio de Janeiro, pois ambos constam com 14 alunos que obtiveram nota máxima na prova de redação. Os estados São Paulo e Ceará estão logo atrás, com 4 e 5 alunos com nota 1000 nessa prova. Outro destaque interessante é que pelo menos um aluno de cada região obteve nota máxima na prova de redação, fato este que não ocorreu nas provas objetivas.

5.2 ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA (MCA)

Nesta seção estão dispostos os resultados da análise de correspondência múltipla para as variáveis referentes ao perfil socioeconômico dos participantes e o desempenho nas provas. O conjunto de dados é composto por 11 variáveis sendo: sexo do participante, autodeclaração da raça/cor, tipo de escola, renda familiar em salários mínimos, escolaridade do pai e da mãe ou responsável, desempenho na prova de ciências naturais, ciências humanas, linguagens, matemática e redação. O intuito é apresentar os fatores que estão associados ao desempenho dos participantes nas provas objetivas e discursiva no ENEM 2018.

Esta seção inicia-se com a visualização do percentual de inércia (variância) explicado por cada dimensão. Observa-se na Figura 18 que as duas primeiras dimensões juntas explicam $20,9\%$ da variabilidade. Note que como temos esse resultado, é necessário o uso de uma dimensão superior a 2 para que tenhamos um valor maior para a inércia. Por exemplo, com as 10 dimensões visualizadas na referida figura tem-se $53,3\%$ de variância explicada. Esta situação pode ser ocasionado por conta da quantidade de categorias de cada variável, entretanto, este assunto será discutido na Seção 6.

A Figura 19 ajuda a identificar as variáveis mais correlacionadas com cada dimensão. As variáveis sexo, raça/cor e tipo de escola são mais correlacionadas com a dimensão 1. Na dimensão 2, observa-se que as variáveis referentes a escolaridade do pai e da mãe ou responsável são as mais correlacionadas com esta dimensão.

Note que as duas dimensões representam $20,9\%$, valor este que representa o quanto as duas dimensões estão explicando a variabilidade dos dados.

A Figura 20 apresenta a análise de correspondência múltipla para as categorias das variáveis de controle do participante, de controle da escola, de controle das provas, do questionário socioeconômico e do desempenho nas provas. Observe que há associação entre as categorias de alto desempenho nas provas e a categoria escola privada, além de que essas categorias estão associadas as categorias de escolaridade de ensino superior/pós-graduação do pai e da mãe ou responsável. Com isso, espera-se que os participantes que estudam/estudaram em escola privada e que os pais possuem escolaridade de nível superior/pós-graduação apresentam desempenho alto (acima do 3^{o} quartil) nas provas

objetivas e discursiva.

Observa-se que há uma associação entre as categorias escola pública, autodeclaração preta/parda, amarela, sexo feminino, escolaridade ensino fundamental do pai e da mãe ou responsável, renda familiar até dois salários mínimos com o desempenho médio nas provas. Então, espera-se que os participantes do sexo feminino autodeclarados pretos/pardos ou amarelos, no qual o pai e mãe possuem escolaridade nível fundamental e a renda familiar equivalente até dois salários mínimos, apresentam desempenho médio (entre o 1º e 3º quartis) nas provas objetivas e discursiva.

Além disso, observa-se que as categorias de desempenho baixo nas referidas provas estão associadas a autodeclaração indígena e que o participante não sabe a escolaridade da mãe ou responsável.

Figura 18 – Gráfico da variância explicada por dimensão.

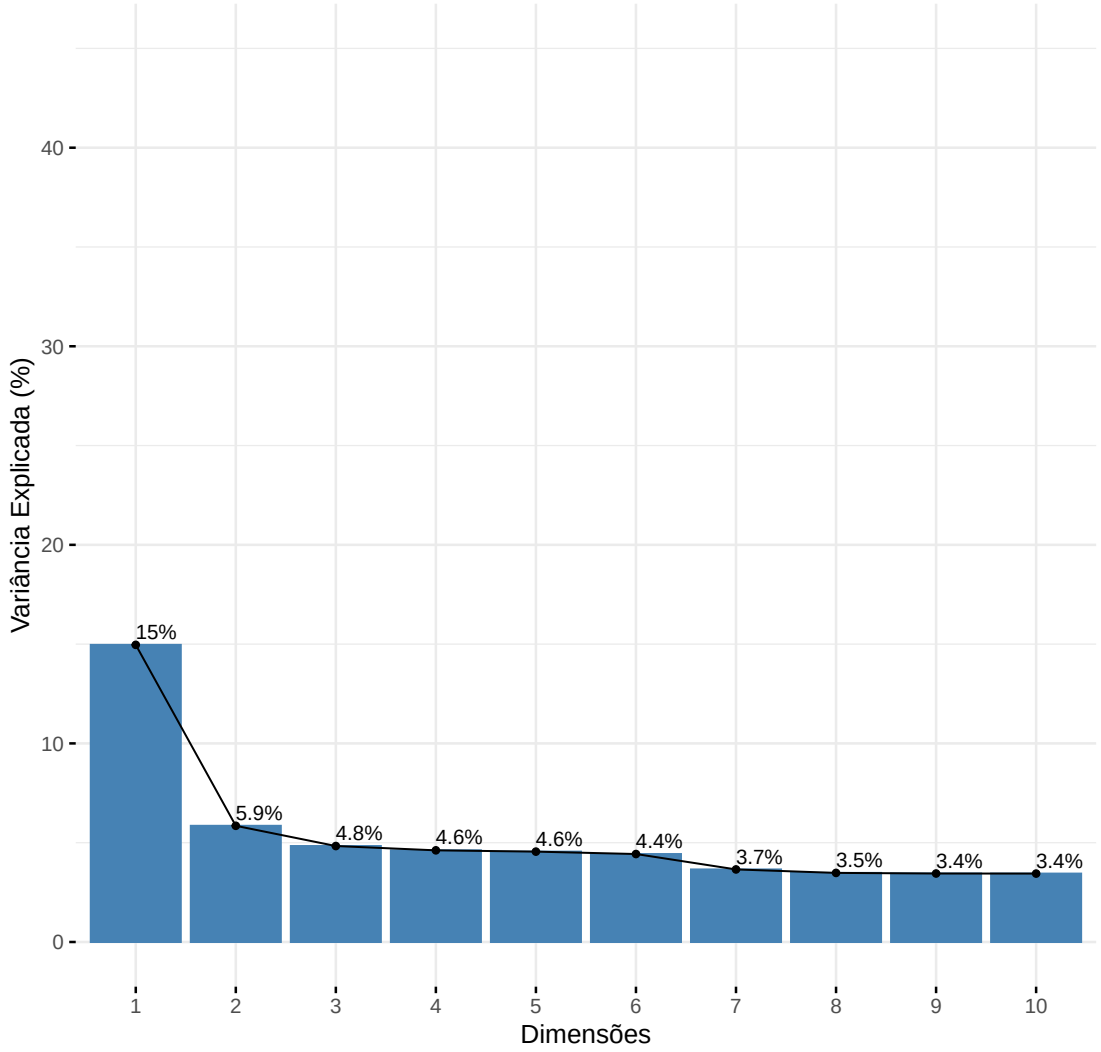


Figura 19 – Mapa de correspondência segundo a correlação das variáveis referentes ao perfil socioeconômico dos participantes segundo o desempenho nas provas do ENEM 2018.

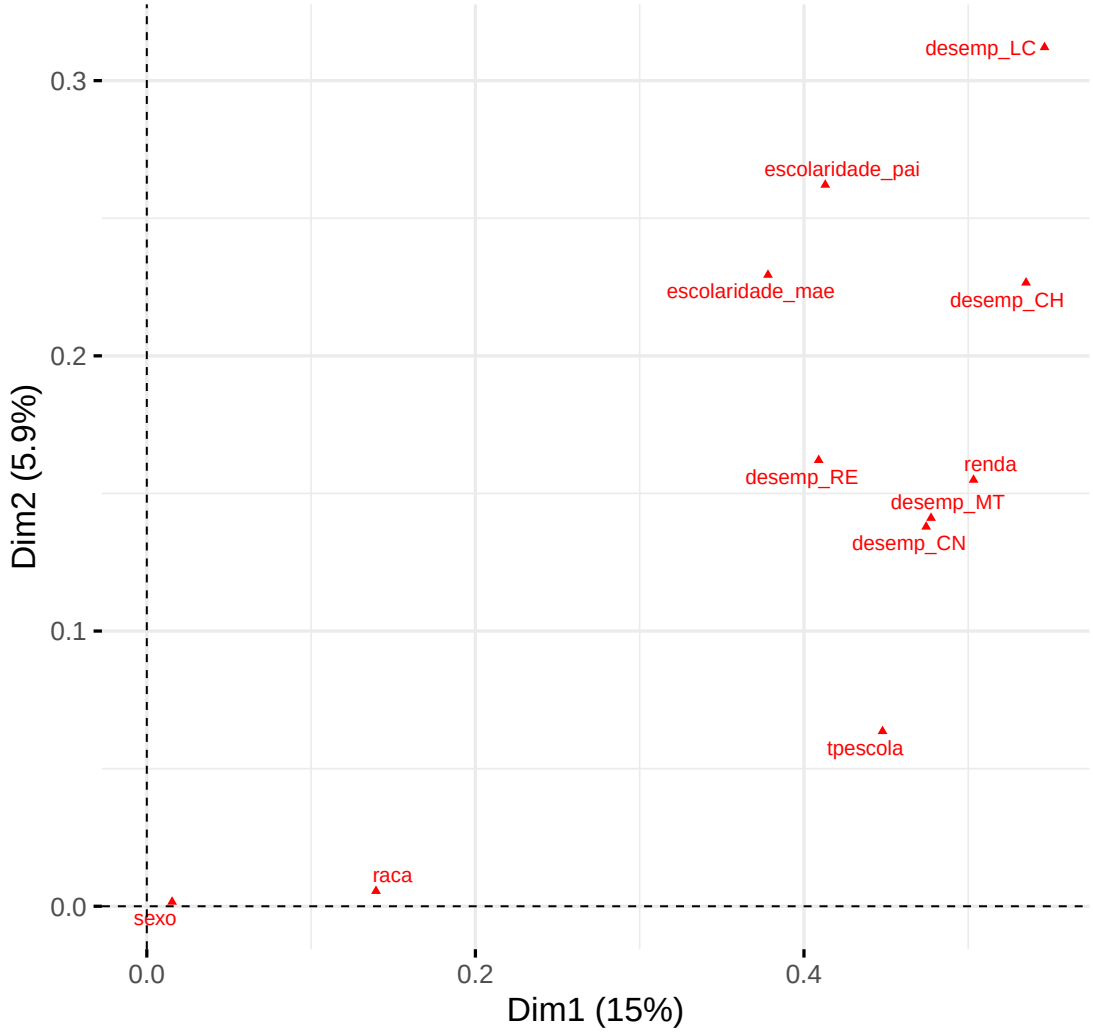
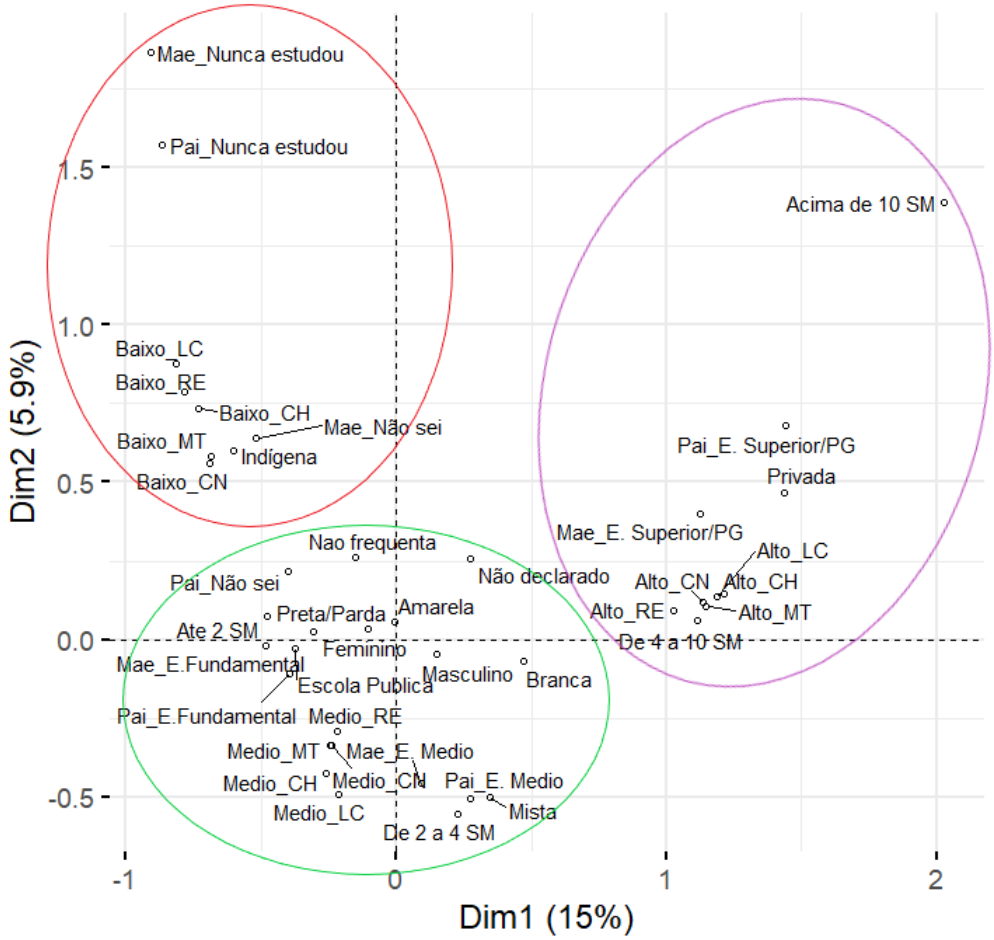


Figura 20 – Mapa de correspondência segundo a associação das variáveis/categorias referentes ao perfil socioeconômico dos participantes segundo o desempenho na provas do ENEM 2018.



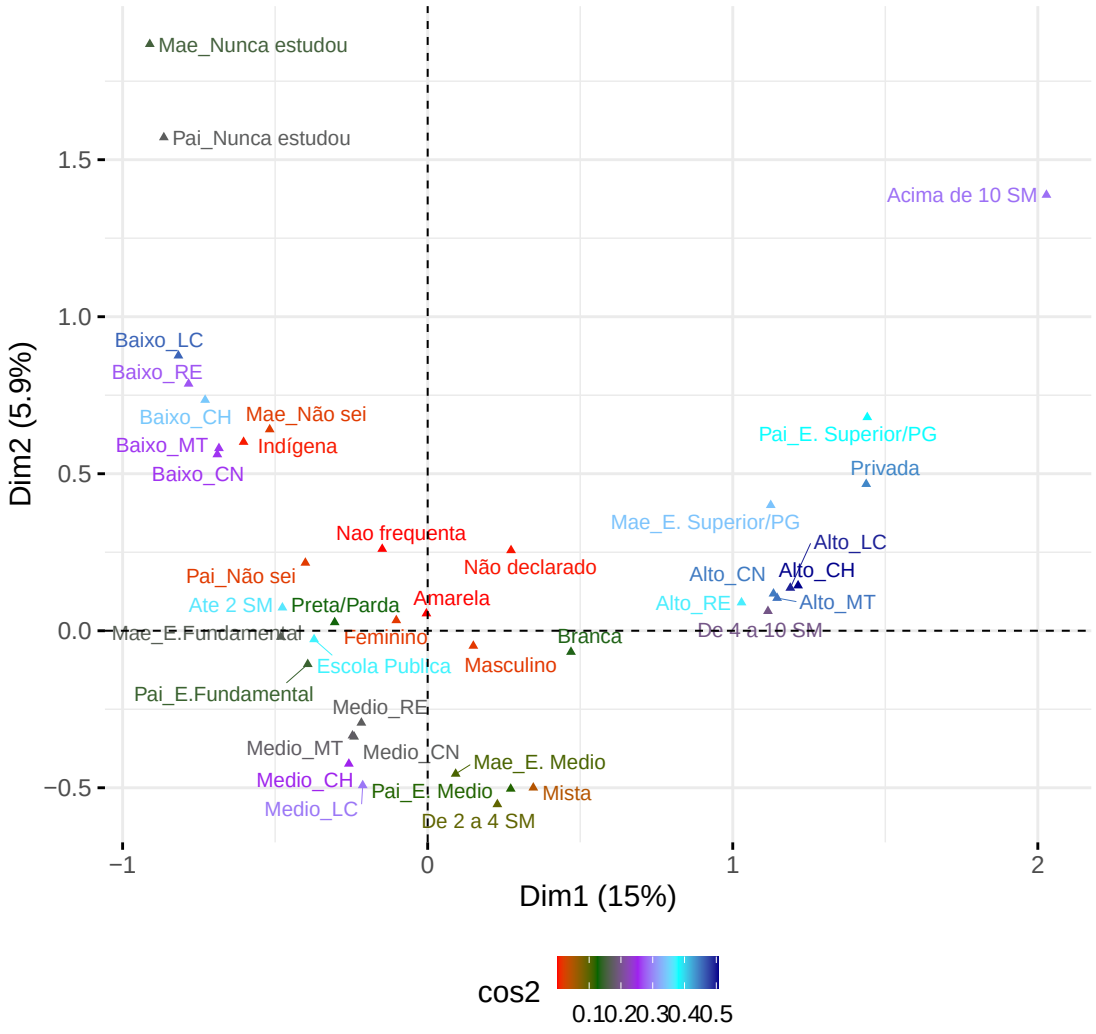
Com base na Figura 21, observa-se a representação gráfica das variáveis e categorias de acordo com a qualidade dos fatores no mapa de correspondência. O "cos 2" (cosseno ao quadrado - $\cos^2$) mede o grau de associação entre as categorias das variáveis em um determinado eixo. A categoria de uma variável é considerada bem representada por duas dimensões se a soma do cos 2 é próxima de 1.

Observa-se que as categorias de alto desempenho nas provas objetivas e discursiva, do tipo de escola sendo privada e pública, a renda familiar até dois salários mínimos e a escolaridade do pai e da mãe sendo ensino superior/pós graduação estão sendo bem representados pelas duas primeiras dimensões no mapa de correspondência.

Em contrapartida, as categorias feminino e masculino, autodeclaração amarela, não frequenta escola, autodeclaração não declarada, autodeclaração indígena não são muito bem representadas pelas duas primeiras dimensões. Isso implica que a posição dos pontos no gráfico de correspondência deve ser interpretada com algum cuidado para que não haja erros na percepção desses resultados. Diante disso, observa-se a necessidade de utilizar

mais dimensões.

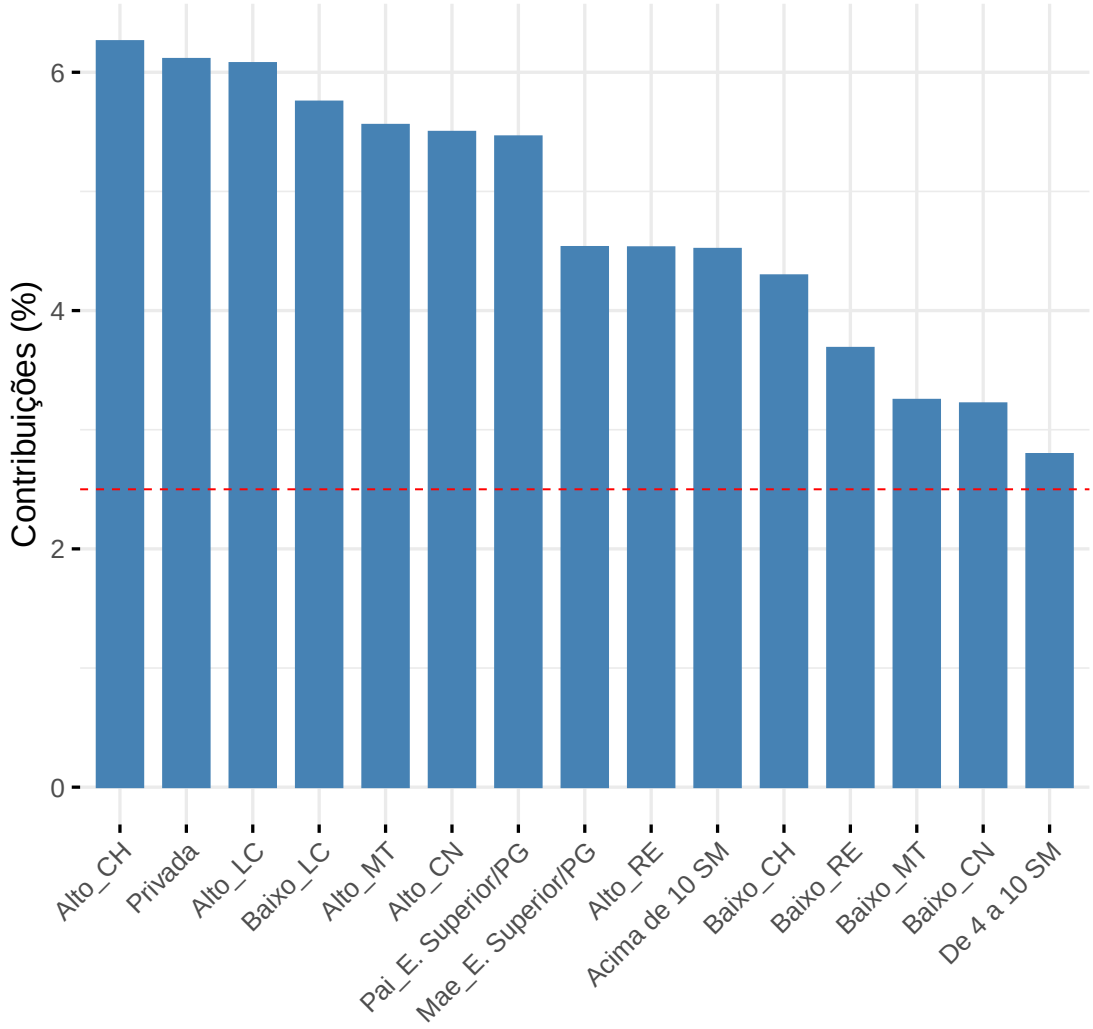
Figura 21 – Mapa de correspondência segundo a qualidade de representação das categorias das variáveis referentes ao perfil socioeconômico dos participantes segundo o desempenho nas provas do ENEM 2018.



A Figura 22 evidencia o top 15 das categorias que mais contribuem para a definição das duas primeiras dimensões. Observe que as categorias alto desempenho nas provas de Ciências Humanas, Linguagens, Matemática e Ciências Naturais estão destacando-se na contribuição nas dimensões. Vale salientar que as categorias de baixo desempenho também contribuem para a definição das duas primeiras dimensões, entretanto, menos que as categorias referentes ao alto desempenho dos participantes. Outra observação é que as categorias escola privada e escolaridade ensino superior/pós graduação do pai e da mãe dos participantes destacam-se das demais.

Vale frisar que a linha tracejada vermelha na Figura 22 indica o valor médio esperado se as contribuições fossem uniformes.

Figura 22 – Gráfico do top 15 das categorias segundo a contribuição das duas primeiras dimensões.



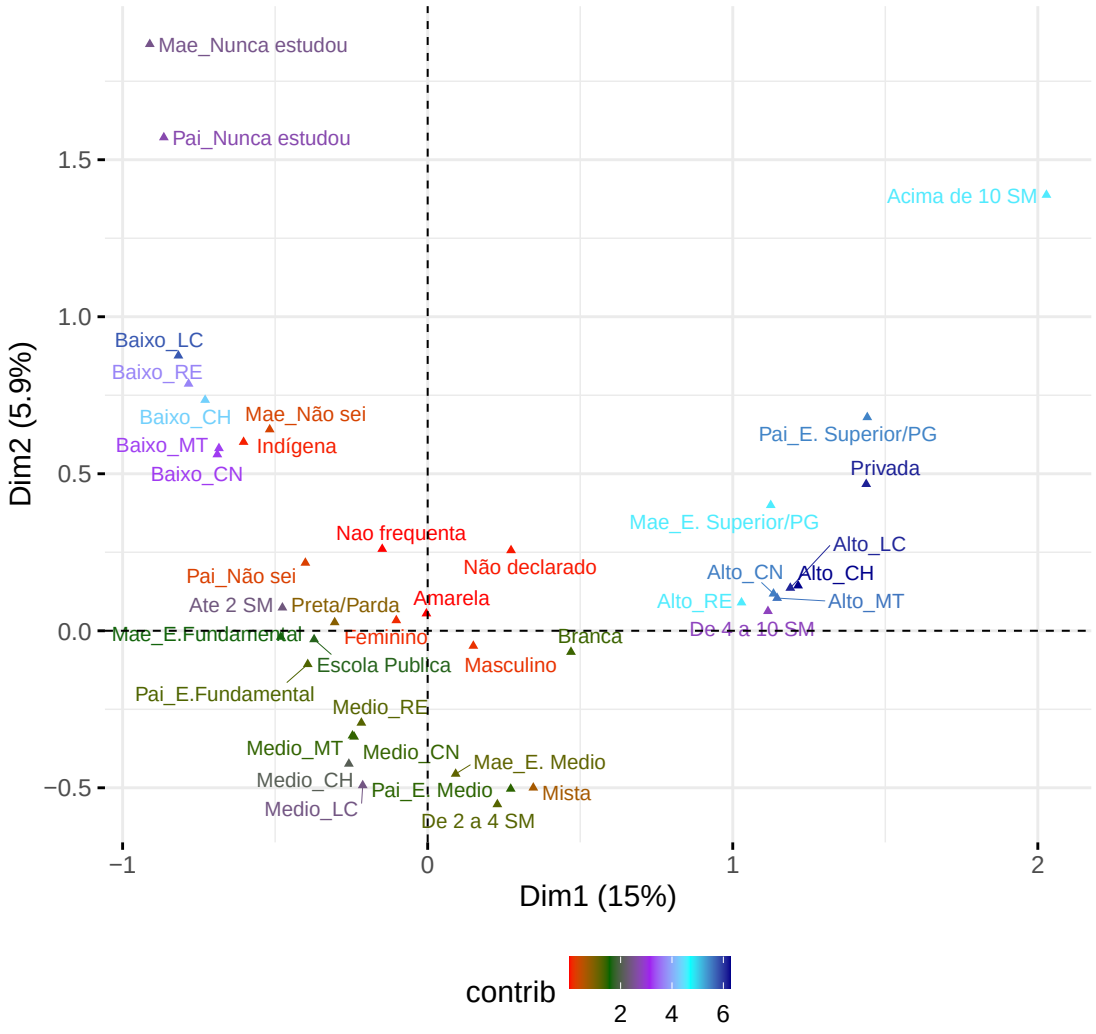
Após a análise do top 15 de contribuição das categorias, a Figura 23, apresenta de maneira mais ampla a contribuição de todas as categorias em estudo. Observa-se o mapa de correspondência das categorias segundo os valores da contribuição para a definição das dimensões. As categorias de variáveis com um perfil semelhante/similar são agrupadas. As categorias de variáveis negativamente correlacionadas são posicionadas em lados opostos da origem do gráfico (quadrantes opostos). E a distância entre os pontos da categoria e a origem mede a qualidade da categoria da variável no mapa.

Mediante isso, pode-se interpretar que as categorias privada, ensino superior/pós graduação do pai e da mãe, escolaridade do pai e da mãe indicando nunca ter estudado e a renda familiar acima de dez salários mínimos são as categorias mais importantes na definição da primeira e segunda dimensão, ou seja, maior contribuição para a definição das dimensões e a explicação da variabilidade dos dados.

Por outro lado, observa-se que as categorias indígena, amarela, raça/cor não decla-

rado, masculino e feminino são as categorias que menos contribuem para a definição das duas primeiras dimensões.

Figura 23 – Mapa de correspondência segundo a contribuição das categorias referentes ao perfil socioeconômico e o desempenho na prova de Ciências Naturais dos participantes do ENEM 2018.



As categorias das variáveis de controle do participante como sexo e autodeclaração da raça/cor, as categorias da variável de controle da escola (tipo de escola, escolaridade do pai e da mãe ou responsável do participante) e as categorias das variáveis do questionário socioeconômico (renda familiar) estão muito próximas no mapa de correspondência segundo a contribuição. Com isso, observa-se que há uma relação entre escola pública e as autodeclarações preto, pardo e amarela, além de estarem associadas ao sexo, a renda familiar de até dois salários mínimos e indicativo de não ter tido alto desempenho na prova de ciências naturais. Entretanto, um contraste a isso é o fato de que observa-se também a associação entre a escola privada, escolaridade nível superior/pós-graduação

do pai e da mãe, renda familiar acima de dez salários mínimos com a questão de ter alto desempenho na referida prova.

Levando em consideração os aspectos sociais do país, este resultado evidencia a questão da desigualdade socioeconômica advindos da má distribuição de renda associada a raça/cor da população, além da diferença entre o ensino de vínculo público e o ensino de vínculo privado.

5.3 SHINY DASHBOARD ENEM

O `Shiny` é um pacote do *software* R que auxilia na criação de aplicativos *web* com interatividade sendo desenvolvidos em linguagem R em comunicação com outros tipos de linguagem como HTML, CSS e JavaScript. Por meio deste pacote é possível criar aplicativos na *web* ou até mesmo *Dashboards*.

A *Dashboard* referente a este trabalho foi denominada de "*Dashboard* ENEM". Nesta *Dashboard* consta os resultados desenvolvidos neste trabalho com a finalidade de divulgação para quem estiver interessado em conhecer mais sobre os resultados do perfil dos participantes, do desempenho nas provas objetivas e discursiva, além da análise de correspondência múltipla dessas variáveis e do questionário socioeconômico associado ao desempenho.

A *Dashboard* desenvolvida está organizada da seguinte forma:

- Início;
- Dados;
- Pré-processamento;
- Brasil:
  - Perfil dos participantes;
  - Desempenho;
  - Análise gráfica;
  - Mapas;
  - Análise de Correspondência.
- Regiões:
  - Centro-Oeste;
  - Nordeste;
  - Norte;
  - Sudeste;
  - Sul.

O objetivo da *Dashboard* é divulgar os resultados dos microdados do ENEM 2018 a nível nacional, sumarizando os resultados apresentados neste trabalho para que possa gerar interesse e agregar conhecimento a respeito do ENEM 2018. Este exame possui grande importância no aspecto educacional, e a divulgação dos resultados referentes a análise dos fatores que estão associados ao desempenho dos participantes pode ser uma forma de evidenciar os fatores que precisam de mais atenção dos órgãos governamentais ou até mesmo de colaboradores e pesquisadores que trabalham na área da educação ou que se interessam pela mesma.

Diante disso, foi realizado a criação da Seção "Regiões" com o intuito de adicionar mais informações para a *Dashboard* para que possa atender a um maior número de pessoas.

No apêndice - ver seção A - são apresentadas algumas imagens alusivas a apresentação da estrutura da *Dashboard* bem como suas respectivas seções e subseções, tanto a nível nacional quanto por regiões.

A publicação da *Dashboard* ocorrerá após os resultados deste trabalho serem publicados em formato de artigo científico, para que se possa preservar todo o processo desenvolvido durante o trabalho.

6 CONSIDERAÇÕES FINAIS

O desafio de processar uma base de dados de mais de três *gigabytes* em uma máquina de especificações intermediárias tornou a análise dos microdados do ENEM uma oportunidade ímpar. A busca por conhecimento em processamento em paralelismo, bem como manipulação e estruturação de *Big Data* por meio do *framework Spark* e o pacote `sparklyr` no ambiente R foi estimada por este desafio.

Por mais que o objetivo inicial do ENEM tenha sido avaliar o desempenho dos estudantes concluintes nos anos de realização da prova, nota-se que nos dias atuais, este é o principal exame educacional do Brasil e, juntamente com o Sistema de Seleção Unificada (SISU), viabiliza o acesso às Instituições de Ensino Superior(IES) do país.

Os resultados da análise descritiva e exploratória dos dados trazem à tona o desempenho dos participantes por estado. Os participantes dos estados das regiões Sul e Sudeste apresentaram melhores notas em comparação aos participantes de outros estados e regiões. Observa-se um contraste na educação e consequentemente no desempenho dos participantes nos estados que estão mais distantes do centro econômico do Brasil. Este fator pode está relacionado aos aspectos sociais e econômicos encontrados em cada estado.

Diante disso, o uso da técnica de análise de correspondência múltipla (MCA) auxiliou no estudo dos fatores que mais estão associados ao desempenho dos participantes do ENEM 2018.

Foi observado nos resultados que a inércia total não apresentou valores tão expressivos. As variáveis utilizadas no presente trabalho, possuem uma grande quantidade de categorias, e isto pode ser um fator que está influenciando na quantidade de variância explicada. A inclusão ou remoção de alguma variável pode ser um outro fator que pode ser pensado, e que talvez possa está intervindo na inércia. Além disso, apesar das duas primeiras dimensões explicarem menos de 30% da variabilidade dos dados, os resultados encontrados na análise de correspondência múltipla apontou que há associação entre as variáveis de controle do participante, da escola, do questionário socioeconômico e o desempenho nas provas objetivas e discursiva. Esta relação entre os aspectos socioeconômicos e o desempenho dos participantes do ENEM 2018, é uma situação que evidencia uma desigualdade social, econômica e cultural que o Brasil ainda enfrenta e que influencia outros aspectos, sendo um deles a educação.

A publicação da *Dashboard* criada para divulgação dos resultados será feita após os resultados do referido trabalho serem formatados e produzidos em forma de artigo científico a fim de publicação em revista científica. Esta é uma decisão voltada para a preservação do processo desenvolvido durante o trabalho.

Para trabalhos futuros, a possibilidade de atualização e expansão da *Dashboard* com os resultados dos microdados do ENEM nos demais anos pode complementar no objetivo

da *Dashboard* de difundir o conhecimento. Além disso, a possibilidade de analisar a influência das categorias na definição das dimensões, por meio da recategorização das variáveis, bem como a adição e remoção de mais variáveis e suas respectivas categorias por meio da MCA pode apresentar resultados que irão agregar aos apresentados neste trabalho.

REFERÊNCIAS

LURASCHI, JAVIER; KUO, KEVIN; RUIZ, EDGAR. **Mastering Spark with R**., 1.ed., O'Reilly, 2019. 209 p.

SCHROEDER, RALPH. **Big Data: shaping knowledge, shaping everyday life**., University of Oxford, Oxford Internet Institute. Oxford, Reino Unido. 2018.

MANYIKA, JAMES; CHUI, MICHAEL; BROWN, BRAD; BUGHIN, JACQUES; DOBB, RICHARD; ROXBURGH, CHARLES; BYERS, ANGELA. H. **Big Data: The next frontier for innovation, competition, and productivity**., McKinsey & Company, McKinsey Global Institute. 2011.

MAYER-SCHÕNBERGER, V.; CUKIER. **Big Data: A Revolution That Will Transform How We Live, Work, and Think**., Eamon Dolan Book, Houghton Mifflin Harcourt. Boston, New York. 2013.

ALMEIDA, ÁLVARO C. A. F.; **Modelo de Mensuração do Desempenho dos Institutos Federais: Uma análise a partir de microdados**. Universidade Federal da Paraíba, Centro de Educação - Centro de Ciências Sociais Aplicadas. João Pessoa. 2014.

VIEIRA, NARA N.; **As provas das quatro áreas do ENEM vistas como prova única na ótica de modelos da Teoria da Resposta ao Item Uni e Multidimensional**. Universidade Federal de Santa Catarina, Florianópolis. 2016.

ROMERO, CRISTÓBAL; VENTURA, SEBASTIÁN; **Educational Data Mining: A Review of the State-of-the-Art**., IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews. 2010.

SILVA, LEANDO A.; MORINO, ANDERSON H.; SATO, THIAGO M. C.; **Prática de Mineração de Dados no Exame Nacional do Ensino Médio**., Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie, São Paulo. 2014.

JOHNSON, RICHARD A.; WICHERN, DEAN W.; **Applied multivariate statistical analysis**., 4th ed. Upper Saddler River, NJ: Prentice - Hall. 1998.

LIMA, PRISCILA DA SILVA NEVES; **Análise de dados do Enade e Enem: uma revisão sistemática da literatura**., Avaliação (Campinas), Sorocaba, v. 24, n. 1, p. 89-1074, Maio 2019.

MONTEIRO, M. N.; CAVALCANTI, C.; OSTERMANN, F. **Análise de Correspondência Aplicada à Pesquisa em Ensino de Ciências**. Universidade Federal do Rio Grande do Sul. 2017.

JOHNSON, RICHARD A.; WICHERN, DEAN W. **Applied multivariate statistical analysis**. 4th ed. Upper Saddler River, NJ: Prentice - Hall, c1998. 816 p.

GREENACRE, M. J.; **Theory and Applications of Correspondence Analysis**. London: Academic Press. 1984.

GREENACRE, M. J.; **Correspondence Analysis in Practice**. London: Academic Press. 1993.

LE ROUX, B., ROUANET,H.; **Multiple Correspondence Analysis**. Londres: SAGE, 2010.

APÊNDICE A – *Dashboard* ENEM 2018

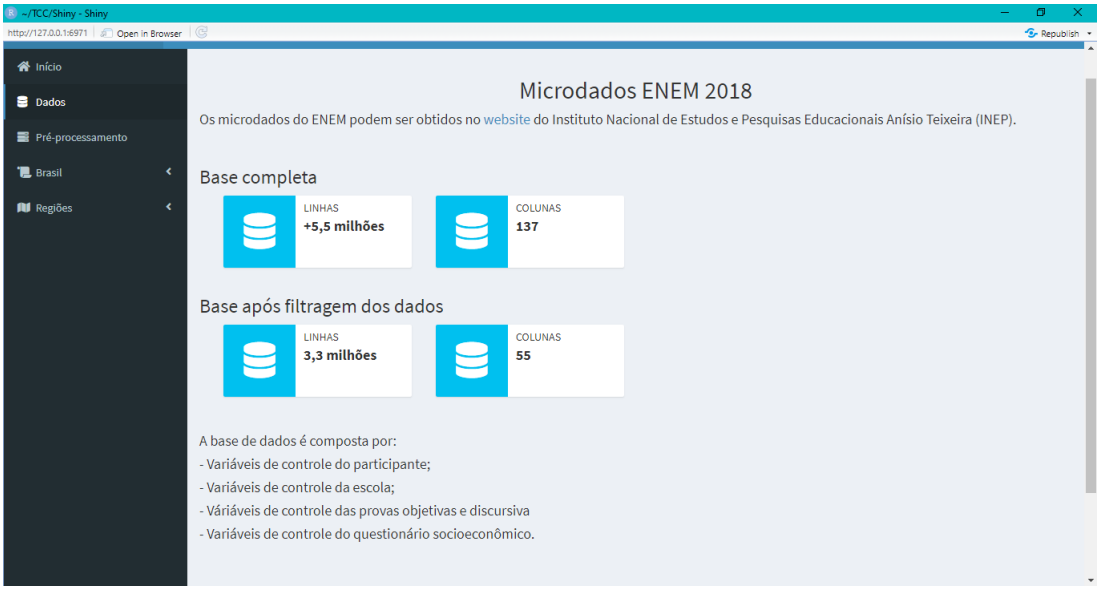
Nesta seção são apresentadas as imagens com as principais partes da *Dashboard ENEM*. Essas imagens apresentam a estrutura dos resultados desenvolvidos neste trabalho. A seguir são explicadas as seções que compõem a estrutura da *Dashboard*.

Na aba "Início" - Figura 24 - são apresentados as informações referentes ao objetivo e motivação do trabalho, a finalidade da *Dashboard*; *hyperlinks* que direcionam ao *GitHub* e *Linkedin* do autor principal do trabalho. No *Github* está disponibilizados os códigos e links dados utilizados no trabalho.

Figura 24 – Página inicial da *Dashboard* ENEM.

Na aba "Dados" estão as informações sobre a base de dados utilizada e o link de acesso ao *website* do INEP onde os microdados do ENEM podem ser encontrados. Atualmente os microdados de 1998 a 2019 estão disponíveis no *website* do INEP. Além disso, constam as informações sobre a base de dados completa e após a filtragem das variáveis de interesse.

Figura 25 – Seção Dados da *Dashboard* ENEM.



A aba "Pré-processamento" informa sobre o *framework Spark* e o pacote `sparklyr`, bem como os links do *R documentation* para incrementar no conhecimento de quem se interessar em descobrir mais sobre o processamento em paralelismo.

Figura 26 – Seção Pré-processamento da *Dashboard* ENEM.

A aba denominada "Brasil" é referente aos resultados dos microdados do ENEM 2018 a nível nacional, em sua totalidade. Primeiro temos a Subseção "Perfil dos participantes" - ver Figura 27 - no qual são apresentados os resultados sobre os inscritos, como sexo, treineiro, idade, tipo de escola, raça/cor, acesso a internet, nacionalidade, tipo de escola, escolaridade do pai e da mãe ou responsável. Após isso, tem-se a Subseção "Desempenho" - ver Figura 28 - que é apresentado as medidas de tendência central das notas de acordo com as provas objetivas e discursiva.

Figura 27 – Subseção Perfil dos Participantes da Seção Brasil.

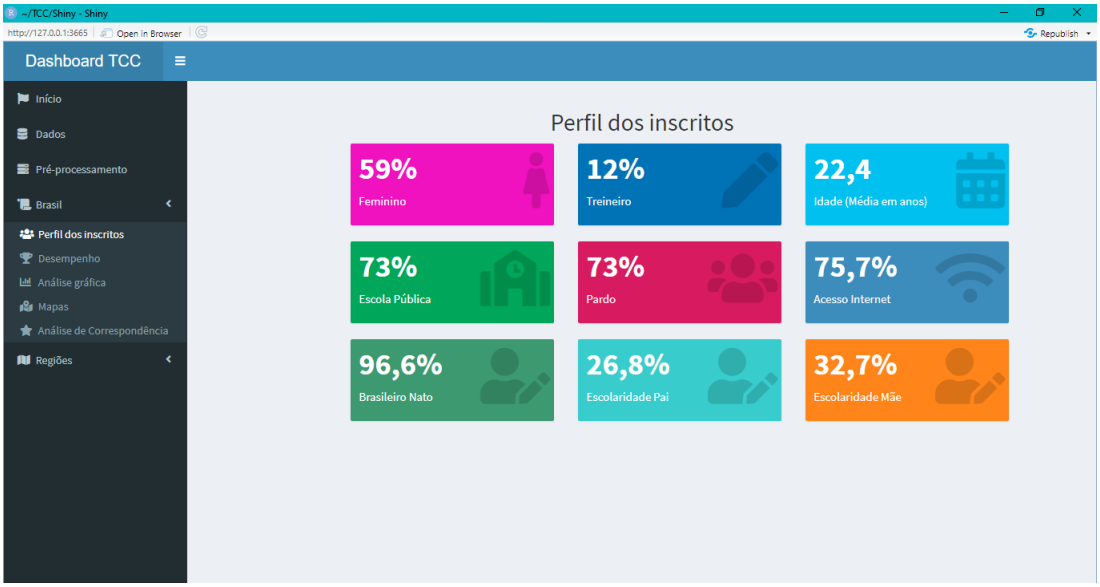
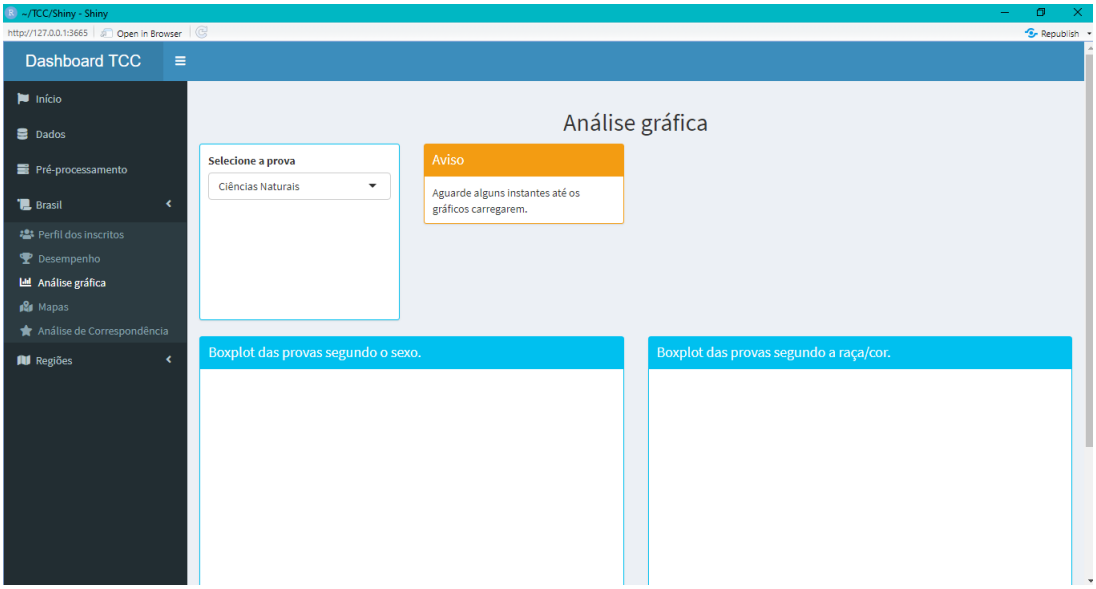


Figura 28 – Subseção Desempenho dos Participantes da Seção Brasil.

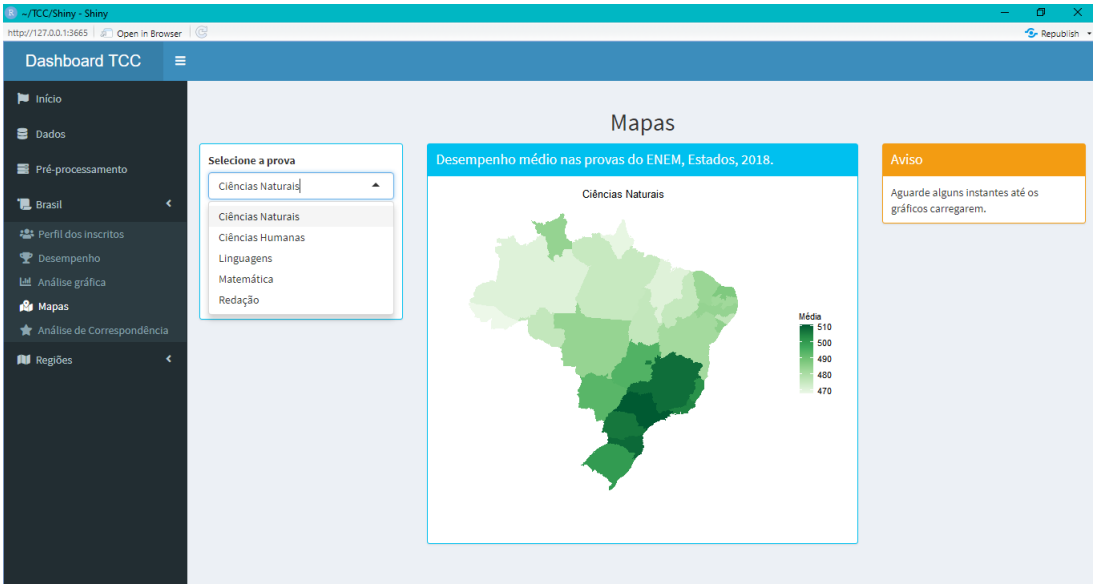
Já na Subseção "Análise gráfica" - ver Figura 29 - estão os *boxplots* do desempenho dos inscritos segundo o sexo e a raça/cor, sendo que há uma caixa de seleção no qual é possível selecionar qual prova deseja-se conhecer os resultados. Vale frisar que há uma caixa de aviso para informar que a geração dos gráficos pode demorar alguns segundos.

Figura 29 – Subseção Análise gráfica da Seção Brasil.



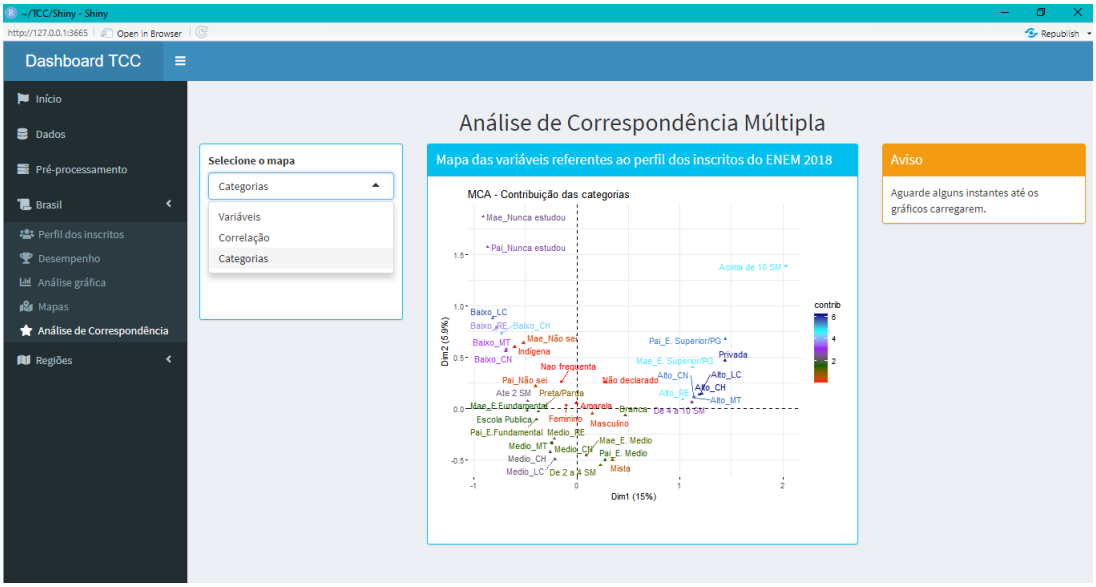
Na Subseção "Mapas" - Figura 30 - encontra-se os mapas de distribuição dos participantes de acordo com o desempenho por estado no ENEM 2018. Nesta subseção também é possível a seleção por prova e o gráfico aparece em alguns instantes.

Figura 30 – Subseção Mapas da Seção Brasil.



No fim da aba de resultados, é apresentado a análise de correspondência múltipla (ver Figura 31). É possível selecionar entre o mapa de correlação das variáveis e o mapa de contribuição entre as categorias das variáveis segundo o desempenho dos participantes.

Figura 31 – Subseção Análise de Correspondência da Seção Brasil.



Um fato interessante é que na *Dashboard* foi criado uma seção contendo informações sobre o perfil dos participantes e a análise gráfica por região. É possível selecionar a prova que deseja-se visualizar os resultados para cada região (ver Figura 32).

Figura 32 – Seção Regiões da *Dashboard* ENEM.