Como a composição de gênero da coorte de ensino médio influencia a escolha de graduação dos estudantes?

Uma abordagem de ciência de dados

Dayanne Cristina Pereira Gomes

Relatório apresentado ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para exame de qualificação de
Mestrado em Ciências

Programa: Ciência da Computação
Orientador: Prof. Dr. Fabio Kon
Colaboradora: Drª. Bruna Borges (EESP-FGV)

São Paulo
Dezembro de 2023

Como a composição de gênero da coorte de ensino médio influencia a escolha de graduação dos estudantes?

Uma abordagem de ciência de dados

Dayanne Cristina Pereira Gomes

Esta é a versão original do texto de qualificação elaborado pela candidata Dayanne Cristina Pereira Gomes, tal como submetido à Comissão Julgadora.

Resumo

Dayanne Cristina Pereira Gomes. **Como a composição de gênero da coorte de ensino médio influencia a escolha de graduação dos estudantes?: *Uma abordagem de ciência de dados***. Exame de Qualificação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Dentro do contexto educacional, existe um momento decisivo na trajetória de jovens estudantes do ensino médio: a escolha do curso de graduação. O ingresso no ensino superior leva em consideração múltiplos fatores determinantes, dentre eles, as questões de gênero. Mulheres enfrentam diferentes desafios, sendo alvos de disparidade de gênero dentro de suas carreiras acadêmicas e profissionais. Alguns estudos existentes na literatura abordam como a exposição a mais pares do gênero feminino pode impactar nas escolhas de meninos e meninas, mas ainda há poucos estudos que abordam o cenário brasileiro. O objetivo desta pesquisa de mestrado é utilizar ciência de dados para investigar como a composição de gênero da coorte do ensino médio influencia a escolha do curso superior de jovens estudantes no Brasil. Para isso, utilizaremos dados do Exame Nacional do Ensino Médio (ENEM), combinando dados do Sistema de Seleção Unificada (SISU), utilizando metodologias econométricas para análise do efeito. Em específico, utilizaremos variações idiossincráticas na composição de gênero entre coortes de uma mesma escola para estimação de um efeito causal. Além disso, visamos caracterizar diferenças de gênero no desempenho no ENEM, bem como caracterizar as escolas de origem dos estudantes e as propensões de jovens do gênero feminino a escolherem cursos com baixa representatividade feminina, especificamente na área de ciências exatas e engenharias. São utilizados dados educacionais públicos do ENEM, SISU e Censo Escolar para alimentar o processo de ciência de dados, empregando uma metodologia de pré-processamento e análise exploratória de dados, a fim de extrair percepções valiosas dos dados disponíveis. Nesta etapa preliminar do trabalho, foi realizada a combinação dos conjuntos de dados do ENEM e SISU através da criação de um identificador único. Isso possibilitou a caracterização do perfil dos candidatos, instituições de ensino superior e cursos de graduação, observando-se diferenças entre os gêneros. Na continuação deste trabalho, será realizada a exploração dos dados das escolas dos estudantes e futuramente a aplicação do método de variações idiossincráticas na composição de gênero.

Palavras-chave: Ciência de dados. Composição de gênero. Escolhas de graduação. Ensino médio. Educação. SISU. ENEM. Censo Escolar.

Abstract

Dayanne Cristina Pereira Gomes. **How does the gender composition of a high school cohort influences students' major choice?: *A data science approach***. Qualifying Exam (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

In the educational context, there is a decisive moment in the trajectory of young high school students: major choices. Admission to higher education takes into account multiple determining factors, including gender issues. Women face different challenges, being targets of gender disparity within their academic and professional careers. Some studies in the literature address how exposure to more female peers can impact the choices of boys and girls, but there are still few studies that address the Brazilian scenario. The objective of this master's research is to use data science to investigate how the gender composition of the high school cohort influences the higher education course choice of young students in Brazil. To do this, we will use data from the National Secondary Education Examination (ENEM), combining data from the Unified Selection System (SISU), using econometric methodologies to analyze the effect. Specifically, we will use idiosyncratic variations in gender composition between cohorts from the same school to estimate a causal effect. Furthermore, we aim to characterize gender differences in ENEM performance, as well as characterize the students' schools and the propensities of young females to choose majors with low female representation, specifically in the STEM field. Public educational data from ENEM, SISU, and School Census are used to feed the data science process, employing a pre-processing and exploratory data analysis methodology to extract valuable insights from the available data. In this preliminary stage of the work, the ENEM and SISU data sets were combined by creating a unique identifier. This made it possible to characterize the profile of candidates, higher education institutions, and majors, observing differences between genders. In the continuation of this work, data from students' schools will be explored and, in the future, the method of idiosyncratic variations in gender composition will be applied.

Keywords: Data science. Gender composition. Major choices. High school. Education. SISU. ENEM. School Census.

Lista de Abreviaturas

IBGE	Instituto Brasileiro de Geografia e Estatística
IME	Instituto de Matemática e Estatística
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IES	Instituição de Ensino Superior
MEC	Ministério da Educação
SISU	Sistema de Seleção Unificada
USP	Universidade de São Paulo

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O momento da escolha de carreira profissional é um acontecimento de grande importância na vida de um indivíduo. Esse processo é determinante na definição de características individuais e coletivas de pessoas que se veem na responsabilidade de tomar uma decisão impactante (Akosah-Twumasi *et al.*, 2018) cujas consequências podem influenciar uma vida inteira. As implicações de optar por um determinado caminho, tanto no âmbito acadêmico como profissional, vão além da preferência por uma área do conhecimento. Elas podem influenciar de forma significativa a qualidade de vida futura, já que sua ocupação tem poder de influenciar personalidade, nível de renda, status social e grupos sociais nos quais os sujeitos se caracterizarão (Shahid Kazi e Akhlaq, 2017).

As mudanças ocasionadas pelas transformações sociais fizeram com que se percebesse a necessidade de entender como a juventude é afetada no mundo contemporâneo (UNESCO, 2006). A responsabilidade de tomar decisões significativas não é alheia a esses jovens. Segundo Gati e Saka (2001), adolescentes estão envolvidos diretamente no processo de escolhas complexas. Preocupações com questões educacionais, como estudos e carreira, mostram um nível de entendimento dos riscos e implicações envolvidos, além das dificuldades que podem surgir.

Esses temas podem ser observados de forma muito expressiva em estudantes do ensino médio. Para aqueles que estão vivendo o término da adolescência e início da vida adulta, a transição envolve não só o amadurecimento etário, mas também a passagem da educação secundária para a superior. Para uma grande maioria desses estudantes, há a aspiração de adentrar algum curso de graduação (Venezia e Jaeger, 2013). A escolha de frequentar a universidade pode começar cedo durante a jornada educacional e, nesse caminho até adentrar o curso desejado, existem diversas etapas (Cabrera e Nasa, 2000). Em todo o mundo, exames educacionais competitivos são aplicados para medir o desempenho de alunos secundaristas. Os resultados desses exames podem ser utilizados para ranquear os melhores candidatos e convocá-los para admissão em uma instituição. Alguns exemplos são ACT e SAT, nos Estados Unidos, Baccalauréat, na França e GCSE, no Reino Unido.

No Brasil, tais exames são conhecidos como vestibular. Os participantes costumam realizá-lo no último ano do ensino médio, quando estão encerrando seus estudos, ou até que consigam a classificação para a vaga pretendida. As instituições têm autonomia para adotar

uma prova específica, cuja aplicação é realizada pela própria instituição. Outra opção é a adoção do ENEM como forma de entrada. O Exame Nacional do Ensino Médio é uma prova anual com abrangência em todo território brasileiro. A nota desse exame, além de avaliar o desempenho de estudantes secundaristas, pode ser utilizada para adentrar algum curso superior através de programas governamentais, como SISU, ProUni e FIES.

Dadas diversas formas de ingresso, ainda é parte da escolha do estudante qual curso de graduação ele deseja seguir. Múltiplos fatores devem ser considerados nesse processo. Para Borchert (2001), existem três grandes perpectivas que afetam a escolha de carreira: oportunidade, personalidade e ambiente. Segundo Abbagnano (2012), ambiente pode ser definido como um complexo conjunto de relações entre mundo natural e ser vivo, que influem na vida e no comportamento do indivíduo. O gênero tem um grande papel nas questões relacionadas ao ambiente no qual os jovens se encontram, incluindo o escolar. No cenário brasileiro, as questões de gênero são relevantes e consideradas no processo de levantamento de indicadores sociais. Um exemplo é o estudo Estatísticas de gênero: indicadores sociais das mulheres no Brasil (IBGE, 2021). Nele, são levantados diferentes aspectos da vida da população, incluindo desigualdades no mercado de trabalho e na educação enfrentadas pelas mulheres. Apesar de haver avanços recentes na igualdade de participação, com mais acesso ao ensino superior, elas ainda sofrem com uma subrepresentatividade em áreas com predominância masculina, como Ciência, Tecnologias, Engenharias e Matemáticas (STEM) (Saavedra *et al.*, 2010). Questões de disparidade de gênero afetam a diversidade das áreas e interferem no processo de escolha de uma educação superior, que podem ser uma barreira enfrentada por jovens mulheres em busca de uma carreira em STEM (Verdugo-Castro *et al.*, 2022). Mesmo sendo maioria nas universidades e tendo uma maior presença no mercado de trabalho, mulheres tendem a seguir carreiras menos prestigiosas, com consequências salariais e nas relações de poder (Senkevics, 2021).

1.1 Motivação e Objetivos

Neste contexto, queremos entender o perfil de estudantes e suas escolas de ensino médio, bem como um fator específico, a exposição aos seus colegas do gênero feminino durante a trajetória escolar, afeta as escolhas de curso de graduação, em particular de cursos com baixa representatividade feminina. Observando os desafios enfrentados pelo gênero feminino no mercado de trabalho, queremos trazer um maior entendimento sobre as decisões de carreira de jovens mulheres. Assim, esperamos auxiliar na orientação de esforços para criação de ambientes mais inclusivos e diversificados na educação superior. Para auxiliar no processo de análise desses pontos, levantamos as seguintes questões de pesquisa:

- **Q1** Há diferença no desempenho de homens e mulheres nas disciplinas do ENEM?
- **Q2** Homens e mulheres que realizam o processo seletivo do SISU fazem escolhas distintas? Em quais cursos estão mais concentrados? Há uma diferença na proporção daqueles que selecionam cursos STEM?
- **Q3** Quais são as características das escolas de ensino médio nas quais os estudantes se formaram?

- **Q4** A composição de gênero da coorte do ensino médio influencia a escolha de graduação das estudantes? Mulheres em escolas com mais colegas do sexo feminino são mais ou menos propensas a selecionaram cursos STEM? Há um efeito sobre a probabilidade das mulheres escolherem o curso de ciência da computação? Há efeitos diferentes entre as regiões do Brasil?

Para alcançar os objetivos dessa pesquisa, utilizaremos uma metodologia econométrica que compara a composição de gênero entre coortes dentro de uma mesma escola para identificação dos efeitos da composição de gênero. Serão aplicadas técnicas de ciência de dados para alcançar os objetivos deste trabalho. A ciência de dados é uma área multidisciplinar que envolve técnicas econométricas, matemáticas e computacionais para resolver problemas de um determinado domínio. Os dados utilizados nesses processo passam uma série de etapas, como tratamento, análise e visualização, a fim de dar suporte à tomada de decisões estratégicas. Trabalharemos com dados educacionais a nível nacional de múltiplas fontes, incluindo ENEM, SISU e Censo Escolar.

Capítulo 2

Fundamental teórico

Neste capítulo, apresentaremos alguns conceitos e definições teóricas que serão abordadas nesta pesquisa. A Seção 2.1 apresenta a definição de estudo de coorte e a Seção 2.2 apresenta o método de variações idiossincráticas na composição de gênero.

2.1 Estudo de coorte

A palavra coorte tem origem no latim *cohors*. Esse termo era utilizado para nomear uma unidade militar do Império Romano, que compunha uma legião. Pode ser considerada equivalente ao conceito moderno de batalhão. No contexto científico, uma coorte pode ser definida como um grupo de pessoas que possuem uma característica ou experiência em comum. Coortes podem ser estabelecidas com vários propósitos. Um deles é a análise de grupos dentro de um determinado domínio, como em estudos econométricos, epidemiológicos e demográficos. Dentro das metodologias utilizadas para explorar uma questão específica estão os estudos observacionais. O que diferencia os estudos de caráter observacional de outros estudos é a realização de intervenções. Nesses estudos, os pesquisadores não interferem nos fenômenos estudados, apenas os observam, fazendo com que a variável considerada não esteja sob controle (Song e Chung, 2010).

Estudos observacionais podem ser divididos de acordo com o período de coleta de dados. Quando as observações são feitas em um momento específico, coletando dados em um curto intervalo de tempo, chamamos de estudo de corte transversal. Eles são úteis para observar o estado e as condições atuais dos participantes, sem que haja um acompanhamento dos indivíduos ao longo do tempo (Zangirolami-Raimundo *et al.*, 2018).

Já quando os indivíduos são acompanhados por longos períodos de tempo, os estudos são classificados como longitudinais. São realizadas coletas de dados contínuas ou repetidas em intervalos regulares, como dias, meses e anos. Como os dados levam em consideração um grupo pré-definido, é possível ajustar os métodos estatísticos para mudanças observadas ao longo do tempo para o grupo como um todo ou para sujeitos específicos (Caruana *et al.*, 2015).

Os estudos de coorte são um tipo de estudo longitudinal que leva em consideração uma segmentação específica da população, que é a coorte. Um exemplo de estudo de coorte é a

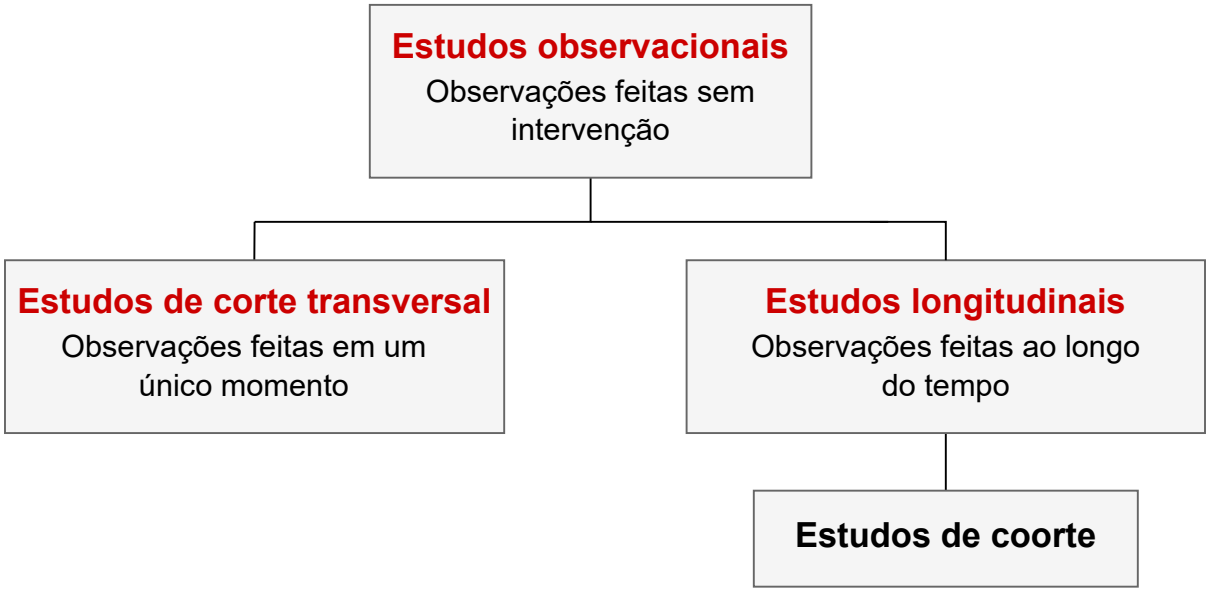


Figura 2.1: *Tipos de estudos observacionais*

observação do desenvolvimento de doenças em uma população. A amostra populacional pode ser dividida em duas coortes: a coorte 1, de expostos à doença e a coorte 2, de não expostos. Assim, é possível comparar a ocorrência da doença entre os grupos.

Na área da educação, estudos de coorte podem ser utilizados para analisar como questões específicas impactam sujeitos inseridos no sistema educacional, como professores, estudantes e outros envolvidos. Björkenstam *et al.* (2011) investigaram a associação do desempenho escolar com taxas de suicídio em estágios posteriores da vida, utilizando uma coorte de nascidos entre 1972 e 1981 na Suécia. Outro exemplo foi o trabalho de Ensminger e Slusarcick (1992), que examinou os caminhos de desenvolvimento de uma coorte de estudantes negros de uma escola de Chicago entre 1966 e 1977.

A realização de estudos de coorte educacionais, em particular os retrospectivos, como é o atual, é facilitada pela disponibilização de dados públicos de diferentes abrangências. No Brasil, há a inicativa do Portal Brasileiro de Dados Abertos, com dados nos níveis federal e local de múltiplas áreas. No sentido desta pesquisa, utilizaremos dados de estudantes que prestaram ENEM em um determinado ano para identificar suas escolas e acompanhá-las ao longo do tempo. Dessa forma, podemos analisar a composição de gênero vis-à-vis outras coortes estudantis.

2.2 Estratégias empíricas: Variações idiossincráticas na composição de gênero

Um desafio na condução de uma pesquisa é isolamento e entendimento do impacto de fatores específicos, como o efeito de pares. O efeito de pares (do inglês *peer effect*) se refere à influência e impactos que indivíduos dentro do círculo próximo (familiar, social ou educacional) têm nas suas decisões e comportamentos. Um trabalho muito relevante no entendimento dos efeitos sociais do grupo que uma pessoa se insere é o de Manski (1993).

Ele aponta a existência de efeitos de pares correlatos, que são comportamentos similares em pessoas do mesmo grupo por conta da semelhança de características individuais ou ambientes institucionais.

Isso é especialmente importante no contexto educacional, já que estudantes do ensino médio compartilham o ambiente escolar durante sua formação. Inseridos em classes diversificadas, eles devem conviver diariamente com outros adolescentes por uma parte relevante de suas vidas. Os colegas de escola podem ser uma importante força social não só no desempenho acadêmico, mas também nas aspirações profissionais e decisões de seguir uma área específica (Tang *et al.*, 2008).

A literatura observa que os pares têm grande relevância em comportamentos, escolhas e resultados educacionais (Sacerdote, 2014; Zimmerman, 2003). Para entender melhor esse efeito, é importante considerar como estão estruturados os grupos de referências de estudantes, já que eles são altamente vulneráveis à influência uns dos outros pela exposição contínua e proximidade. Os efeitos dos pares podem se apresentar tanto de maneira positiva, quanto negativa, que se refletem em pontuações de prova, motivação e hábitos de estudo, por exemplo.

Manski observa uma dificuldade no isolamento desse efeito no comportamento de uma população. Dentro de um grupo, indivíduos tendem a se relacionar com outros que são parecidos com eles, o que pode dificultar na identificação do efeito real dos pares, já que mudanças comportamentais podem ocorrer tanto devido à influência dos pares, quanto pela associação natural de pessoas com comportamentos similares. Esse indivíduo também pode influenciar o par e por ele ser influenciado simultaneamente. Uma forma de isolar essas influências é através da observação da composição de diferentes coortes ao longo do tempo, isto é, como diferentes distribuições dos pares nas salas de aula afetam os alunos.

Um importante precursor dessas observações é o estudo de Hoxby (2000). Ela identificou e mensurou a existência de efeitos dos pares em coortes escolares que diferem na composição de fatores específicos, como o gênero. Nesse sentido, múltiplos trabalhos exploram o problema com uma metodologia similar, denominada de variações idiossincráticas na composição de gênero. Essa estratégia apoia-se na variação aleatória das proporções de estudantes nas coortes escolares. Quando uma característica qualquer é analisada, pode-se ter uma dificuldade na estimativa da regressão por conta da causalidade reversa - variável X influencia váriavel Y -, sem saber qual é a direção da influência. Isso não acontece quando utilizamos a característica de gênero, já que ela é fixa no tempo. Por isso, adotamos a análise do fator de gênero dentro das coortes escolares.

Os modelos econométricos, apesar de semelhantes, diferem por serem estimados utilizando conjuntos de dados educacionais de diferentes países, que se configuram em sistemas de ensino distintos, bem como se apoiam em outros recursos, como registros demográficos. Além disso, podem ser consideradas diferentes etapas da educação básica, como anos iniciais e finais do fundamental e médio.

Como parte da nossa metodologia se propõe a analisar o efeito da composição de gênero nas escolhas de graduação, utilizaremos como referência o trabalho de Silva Borges, 2021, que se volta para coortes de escolas de ensino médio brasileiras, um contexto similar

ao nosso. Para esse propósito, ela emprega dados do Censo Escolar e do vestibular da Universidade Estadual de Campinas (UNICAMP). As definições teóricas apresentadas, assim como a equação definida, foram extraídas do capítulo *Gender peer effects on major choice* (Silva Borges, 2021).

2.2.1 Efeitos de gênero em escolhas de graduação

Variações idiossincráticas podem ser definidas como particularidades e diferenças individuais únicas que podem influenciar resultados ou comportamentos. Esses fatores comuns ou gerais que afetam um grupo de indivíduos também influenciam os sistemas nos quais estes estão inseridos (Meister, 1991). Essas variações podem ser diversas, abrangendo uma ampla gama de características, experiências e situações pessoais. Alguns exemplos de variações são:

- **Contexto pessoal:** Dinâmica familiar, status socioeconômico, crenças, bagagem cultural, valores familiares;
- **Carreira:** Aspirações e objetivos profissionais, experiência prévia de trabalho, habilidades técnicas;
- **Educação:** Qualidade escolar, atividades extracurriculares, exposição prematura a tópicos avançados.

Pais e alunos podem escolher suas turmas potenciais, o que pode ser um problema por conta do viés da auto-seleção (*self-selection bias*). Apesar das escolhas pessoais das famílias, utilizamos como hipótese da estratégia empírica que elas não conseguem prever corretamente as variações coorte a coorte na composição de gênero dos estudantes, que seguem processos aleatórios. Assim, explorando a variação idiossincrática de coortes na proporção de alunas, temos um modelo de regressão que estima o impacto de alguns fatores nas variáveis dependentes (y_{iect}). No seguinte modelo econométrico, i indexa estudante, e escolas, c coorte e t ano do ENEM [1]:

$$y_{iect} = \gamma_0 + \beta_1 Feminino_i + \beta_2 Prop\_FemEM_{iec} + \beta_3 Feminino_i \times Prop\_FemEM_{iec} + \alpha_e + \alpha_t + X_i\omega + Z_{ec}\delta + \gamma_{et} + \epsilon_{iect}$$

A equação é composta por:

- $Feminino_i$, variável binária indicadora de gênero;
- $Prop\_FemEM_{iec}$, proporção de colegas do gênero feminino na escola e no ano de conclusão do ensino médio c;
- α_e, efeitos fixos da escola;
- α_t, efeitos fixos do ano do ENEM;
- X_i, vetor de características individuais do aluno;

[1] Nesta etapa preliminar, temos dados de apenas um ano (2016). Acrescentaremos outros anos nas etapas subsequentes.

- Z_{ec}, variáveis da coorte escolar: tamanho da turma, proporção de alunos que frequentam aulas diurnas, idade média dos colegas;
- γ_{ec}, tendência linear de tempo específica da escola;
- β_1, coeficiente de diferenças de gênero nas variáveis de resultados;
- β_2, coeficiente de efeitos de pares de gênero aplicáveis tanto a homens quanto a mulheres;
- β_3, coeficiente do impacto diferencial de mulheres terem uma proporção maior de colegas do sexo feminino;
- ϵ_{iect}, clusterização dos erros-padrão.

Em todas as estimativas, é realizado um ajuste para a potencial correlação de erros dentro da mesma escola. O conjunto de variáveis de resultado (y_{iect}) está relacionado ao objetivo do estudo, que é analisar se a exposição a maiores proporções de colegas do gênero feminino durante o ensino médio afeta as escolhas do curso de graduação. São criadas variáveis binárias para indicar escolha de curso da área STEM e cursos de Tecnologia da Informação (Ciência da Computação e afins).

Utilizaremos metodologia similar à utilizada por Silva Borges (2021) para analisar os efeitos da composição de gênero das coortes de ensino médio nas opções de carreira dos estudantes, dessa vez numa perspectiva mais abrangente, que é proporcionada pelo conjunto de dados do SISU. No próximo capítulo, apresentaremos os trabalhos relacionados, com enfoque naqueles que utilizaram metodologias semelhantes para analisar o efeito da composição de gênero e que se voltam para o cenário brasileiro.

Capítulo 3

Trabalhos relacionados

Um estudo de grande relevância para o entendimento de como os pares afetam os colegas na sala de aula é o de Hoxby (2000). Nessa pesquisa, a economista Caroline Hoxby explora como a composição de uma turma escolar pode influenciar a experiência de aprendizado dos estudantes e suas conquistas acadêmicas. São utilizados dados de alunos do 4° ao 7° ano do ensino fundamental de escolas públicas do Texas, Estados Unidos na década de 1990. A autora aborda duas fontes de variação idiossincrática, sendo elas as mudanças na composição de gênero e raça de uma turma escolar em anos adjacentes. Duas estratégias empíricas discutem essas variações: na primeira, são avaliados os efeitos de ter um grupo com maioria feminina e diferente composição racial; na segunda, são avaliados os efeitos das conquistas dos pares em grupos masculinos e femininos. Os resultados de Hoxby sugerem que ter uma maioria de colegas do gênero feminino aumenta a pontuação de meninos e meninas em leitura e matemática, que pode ser devido a fenômenos como a menor tendência de mulheres causarem disrupções nas salas de aula.

De forma similar, outros trabalhos estimam os efeitos da composição de gênero através de variações idiossincráticas. Schøne *et al.* (2019) visaram entender como a composição de gênero afeta escolhas educacionais. Para isso, foram utilizados dados de alunos dos anos finais do ensino fundamental da Noruega, entre os anos de 2003 e 2008. Nesse cenário, são considerados os efeitos da parcela feminina de pares nas escolhas de áreas e disciplinas do ensino médio, mais especificamente naqueles orientados a STEM. Os autores observam que um aumento na proporção de meninas torna tanto meninos quanto meninas mais propensos a escolher mais disciplinas da área STEM e menos disciplinas de linguagens no ensino médio. Também sugere-se que uma maioria feminina tem um efeito positivo no desempenho de garotas em matemática, que pode impactar na probabilidade de escolher mais cursos de STEM e diminuir a discriminação de gênero sentida por elas.

Lavy e Schlosser (2011) investigam os efeitos da composição de gênero em diferentes estágios da trajetória escolar, além do desempenho acadêmico de meninos e meninas. Para isso, são utilizados dados de estudantes de Israel do ensino fundamental, entre os anos de 2002 a 2005, e médio, de 1993 a 2000. São utilizados como resultados as notas de disciplinas e performance em exames de entrada. Alguns dos mecanismos destacados são a disrupção e violência na sala de aula, interação entre estudantes, relacionamento aluno-professor e senso de fatiga de professores na sua profissão. Evidências apontam que o aumento de

proporções de garotas, em especial quando elas são maioria na coorte, levam a melhores ambientes escolares e aprendizagem. Os benefícios são ainda maiores para aqueles em contexto socioeconômico vulnerável, com grandes impactos entre estudantes imigrantes e que tenham pais com baixa escolaridade.

Schneeweis e Zweimüller (2012) identificam o impacto causal da composição de gênero na escolha de um campo acadêmico. Os dados cobrem os anos de 1988 e 2006 da cidade de Linz, Áustria, do ensino fundamental, com foco no último ano. No contexto austríaco, os estudantes podem escolher uma área dentro de uma escola vocacional, que prepara para uma vaga de trabalho, ou seguir estudos superiores na universidade. Além de levar em consideração o interesse vocacional, também é estimado o impacto de estudar em turmas com mais colegas do gênero feminino, que pode levar a escolhas de áreas mais técnicas, a depender da composição. Os resultados desse estudo mostram que estudantes do gênero feminino são menos prováveis de escolher áreas com maioria feminina e mais prováveis de seguir áreas técnicas quando expostas a maiores proporções femininas.

No trabalho de Brenøe e Zölitz (2020) é observado o efeito dos pares a longo prazo. A utilização de dados de registros estudantis que ingressam na área de matemática no ensino médio entre os anos de 1980 e 1994 possibilitou o acompanhamento dessa população pelo período de 20 anos. Assim, a investigação dos efeitos da composição de gênero na participação na área STEM na Dinamarca pode acompanhar não só as escolhas educacionais, mas as consequências diretas e retardadas ao longo do tempo em turmas de ensino médio. São observadas probabilidade de entrada e finalização de um curso STEM, bem como os ganhos salariais e propensão de homens e mulheres terem filhos em diferentes etapas da vida. Brenøe e Zölitz observam que maiores proporções de colegas do gênero feminino no ensino médio tornam as escolhas são mais esteriotipadas, ou seja, pessoas de um gênero escolhem áreas mais associadas ao seu gênero. Nesse cenário, mulheres escolhem mais áreas da saúde, ao passo que homens escolhem mais áreas STEM. Além disso, observações a longo prazo indicam uma menor probabilidade de mulheres trabalharem em cargos STEM, além de terem uma menor remuneração. Elas também se caracterizam de forma diferente aos homens com relação à decisão de ter filhos.

Ainda sobre estudos que observam variações idiossincráticas na composição de gênero, temos um recente trabalho que considera o contexto nacional. O trabalho de Silva Borges (2021) utiliza essa metodologia para avaliar se a composição de gênero de coortes do ensino médio influencia a escolha de curso de graduação de estudantes, em especial para mulheres. Essa análise foi realizada com dados do vestibular entre os anos 2000 a 2008 de uma universidade pública, a UNICAMP, que foram relacionados aos dados do Censo Escolar. Borges observou que a porcentagem de pares femininos impacta na escolha de curso, diferindo para homens e mulheres. Uma maior proporção de colegas do gênero feminino reduz a probabilidade da escolha de cursos focados em matemática ou ciência para ambos os gêneros. A autora também verifica que, para mulheres, essa exposição aumenta a escolha de cursos competitivos, com maiores notas de corte. Candidatas expostas a menores proporções de pares femininos ($\leq 33\%$) escolhem cursos menos competitivos.

Outros estudos também utilizaram dados educacionais brasileiros para analisar diferentes fatores. Machado e Szerman (2021) utilizaram dados entre os anos de 2010 e 2017 do SISU e do Censo Escolar para investigar os impactos de sistemas de admissão centralizados

na composição de estudantes. As autoras observaram características dos estudantes como gênero, idade, etnia e migração, além de características das escolas para mensurar os efeitos do SISU na atração de candidatos de diversos perfis. Mello (2022) analisa como reformas educacionais que expandiram o acesso à educação superior impactaram a admissão de estudantes de baixa renda. Essas políticas incluem a expansão da centralização de aplicações com o SISU e mais oferta de cotas de ações afirmativas. São utilizados dados do Censo da Educação Superior dos anos de 2010 a 2015 e do ENEM dos anos de 2009 a 2014. Otero *et al.* (2021) também conduziram um estudo sobre as consequências de ações afirmativas no contexto da admissão em universidades brasileiras. Eles exploraram questões como escolhas de área, frequência e persistência na universidade e rendimentos projetados. Para tal, são utilizados dados do ENEM de 2009 a 2015, SISU de 2016, Censo da Educação Superior entre 2009 e 2019 e Relação Anual de Informações Sociais de 2017. Carvalhaes *et al.* (2022) investigaram a interseção entre renda e raça na estruturação do acesso ao ensino superior. São utilizados dados de estudantes concluintes do ensino médio entre 2012 e 2017 para estimar a probabilidade de pessoas de diferentes contextos ingressarem no ensino superior, bem como a decomposição de efeitos diretos e indiretos de renda e raça.

Existem diversos trabalhos na literatura que abordam como múltiplos fatores influenciam jovens inseridos no contexto educacional, sendo um deles o efeito da composição de gênero. Dentre aqueles que considerassem o contexto brasileiro, porém, foram encontrados poucos estudos que focassem nesse fator específico. Entre os trabalhos apresentados, o trabalho de Silva Borges (2021) é o que mais se assemelha ao que estamos fazendo, porém seus dados se restringem a uma universidade em particular. Nossa motivação é basear-se na metodologia de variações idiossincráticas na composição de gênero entre coortes de uma mesma escola para expandir a análise ao nível nacional, utilizando dados do ENEM, SISU e Censo Escolar para observar como estudantes realizam suas escolhas de curso superior em todo o país e como elas são influenciadas por seus pares no ensino médio.

Capítulo 4

Metodologia do trabalho já desenvolvido

Neste capítulo, definimos a metodologia empregada para responder as questões levantada na pesquisa. As bases de dados utilizadas estão descritas na Seção 4.1. Já as etapas do processo de ciência de dados, que explora os dados obtidos, estão descritas na seções subsequentes. O pré-processamento está descrito na Seção 4.2 e a análise exploratória de dados está descrita na Seção 4.3.

4.1 Bases de dados

Nesta seção, descrevemos as bases de dados utilizadas nesta etapa da pesquisa. Para realizar a análise das escolhas de graduação, utilizamos dados educacionais dos estudantes. Uma das fontes é o Exame Nacional do Ensino Médio (ENEM). Através dele, conseguimos obter informações sobre alunos concludentes e que já concluíram o ensino médio, bem como de suas escolas. A outra fonte utilizada é o Sistema de Seleção Unificada (SISU). Com ela, obtemos informações relativas à inscrição dos alunos em cursos de nível superior, além de detalhes das instituições e cursos ofertados.

4.1.1 ENEM

O Exame Nacional do Ensino Médio (ENEM) é um exame realizado pelo Ministério da Educação cujo objetivo é avaliar o desempenho escolar no final da educação básica. Desde 2009, ele passou a ser utilizado como mecanismo de ingresso à educação superior, cujas notas podem ser aproveitadas no Sistema de Seleção Unificada (SISU) e Programa Universidade para Todos (ProUni). O exame também possibilita o pleito de certificação do ensino médio. Os participantes realizam provas em quatro áreas de conhecimento: linguagens, ciências humanas, ciências da natureza e matemática. Além disso, eles devem desenvolver um texto dissertativo-argumentativo dada uma situação-problema, conhecido como redação (Inep, 2023).

Os dados do Exame Nacional do Ensino Médio são disponibilizados publicamente através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Os dados abertos do INEP incluem os microdados do ENEM, que reúnem um conjunto de informações relativas ao exame. Os microdados são o menor nível de desagregação de dados recolhidos, sendo disponibilizados dados dos anos de 1998 a 2022. Nesta etapa preliminar da pesquisa, utilizamos os dados do ano de 2016. Além dos microdados relativos às edições anuais, também são disponibilizados outros arquivos relevantes, como dicionário de dados, documentos técnicos, provas, gabaritos e programa para leitura da base.

Com o passar dos anos, os microdados foram se diferenciando à medida que eram incluídas ou retiradas determinadas variáveis, mas pode-se observar uma estrutura comum entre as edições. Em 2016, os dados estão divididos nas categorias:

- Dados do participante;
- Dados da escola;
- Dados dos pedidos de atendimento especializado;
- Dados dos pedidos de atendimento específico;
- Dados dos pedidos de recursos especializados e específicos para realização das provas;
- Dados dos pedidos de certificação do ensino médio;
- Dados do local de aplicação da prova;
- Dados da prova objetiva;
- Dados da redação;
- Dados do questionário socioeconômico.

Como parte da política adotada pela Lei Geral de Proteção de Dados Pessoais (LGPD), os dados passam por um tratamento antes de serem publicados. Isso significa que dados cadastrais e sensíveis, como nome, endereço, RG, etc, não são disponibilizados ou passam por uma máscara para anonimizá-lo, como é o caso do número de inscrição. Os arquivos de microdados são disponibilizados no formato .csv (valores separados por vírgulas). Os dados de 2016 estão contidos em apenas um arquivo, com uma tabela. Cada linha da tabela representa a inscrição de um candidato de forma individual, bem como as colunas são as variáveis definidas anteriormente, que caracterizam o participante.

4.1.2 SISU

O Sistema de Seleção Unificada (SISU) é um sistema eletrônico do Ministério da Educação, no qual instituições públicas de ensino superior de todo o Brasil oferecem vagas para estudantes participantes do Exame Nacional do Ensino Médio. Durante o período da oferta de vagas, os alunos são ranqueados de acordo com as notas no exame e, aqueles com melhor classificação, são selecionados. Em cada processo seletivo do SISU, que tem duas aberturas anuais, o candidato pode escolher até duas opções de curso. É possível verificar informações sobre as vagas oferecidas, como cursos, instituições e localizações, turnos e modalidade de concorrência (MEC, 2023).

No ato da inscrição, o sistema recupera as notas da edição mais recente anterior do ENEM. Por exemplo, o SISU 2023 leva em consideração a edição do ENEM 2022. Apenas aqueles que obtiveram nota superior a zero na redação e não têm o status de treineiro no ENEM podem se inscrever. O processo é totalmente digital e gratuito, sendo o estudante o responsável por acompanhar o status da sua inscrição. Quando não há a aprovação em uma das duas opções selecionadas, conhecido por chamada regular, ainda é possível a disputa por vaga através da lista de espera.

Os dados do Sistema de Seleção Unificada foram obtidos através do Portal de Dados Abertos do MEC. O portal é uma plataforma que disponibiliza dados e informações públicas do Ministério da Educação, que podem ser usadas no desenvolvimento de aplicativos e ações. Além do SISU, é possível observar conjuntos de dados de outros programas como FIES, ProUni e PRONATEC. São disponibilizados dados relacionados às inscrições realizadas nos processos seletivos dos anos de 2017 a 2022. Nesta etapa preliminar da pesquisa, utilizamos os dados do ano de 2017.

São fornecidas informações detalhadas sobre o participante, como dados pessoais e desempenho nas provas do ENEM, a vaga para qual ele se inscreve, além da classificação e aprovação. Diferente do ENEM, não há especificação de categoria dos dados no dicionário fornecido. Também há a aplicação de máscara para anonimizar dados sensíveis, como CPF e número de inscrição. Os arquivos também são disponibilizados no formato .csv. Desta vez, são constituídos por múltiplas tabelas. Cada tabela representa uma etapa do processo de convocação dos candidatos, sendo divididas entre chamada regular e lista de espera. Ocorre duas chamadas regulares ao longo do ano, uma em cada semestre. Já a quantidade de listas de espera varia, conforme o preenchimento de vagas nas etapas anteriores. Nesta etapa da pesquisa, utilizamos apenas os dados das chamadas regulares. Cada linha da tabela representa uma inscrição de um candidato, sendo possível que um candidato tenha múltiplas inscrições por conta das duas aberturas do processo ao longo do ano e por poder se inscrever em mais de um curso.

4.1.3 Censo Escolar

O Censo Escolar é um levantamento de informações da educação básica brasileira em escolas e instituições de ensino por todo o país. Essa ferramenta demográfica realiza coletas anuais em colaboração entre o Inep e as secretarias estaduais e municipais de educação, contando com a participação de todas as escolas públicas (federais, estaduais e municipais) e privadas da rede de ensino. O Censo abrange diferentes etapas e modalidades de ensino da educação básica e profissional. Ele permite a obtenção de dados individualizados, em diversos aspectos, de estudantes, professores, turmas e escolas. A pesquisa é realizada em duas etapas: a primeira coleta informações sobre os estabelecimentos de ensino, gestores, turmas, alunos e profissionais escolares em sala de aula; a segunda, informações sobre o movimento e o rendimento escolar dos alunos. Os dados do Censo Escolar, de forma semelhante aos anteriores, são disponibilizados ao público pelo INEP no formato .csv.

4.2 Pré-processamento

Na etapa da metodologia de pré-processamento, os dados são tratados a fim de adaptá-los às necessidades do projeto. Com isso, otimizamos as etapas posteriores através de obtenção de um conjunto de dados que seja mais relevante para a pesquisa, facilitando o processo de análise. As técnicas aplicadas podem ser agrupadas em redução, integração, limpeza e transformação de dados. Utilizamos como referência os trabalhos de Jafari (2022) e García *et al.* (2016), que provêm definições e exemplos práticos de como realizar esses processos.

É importante frisar que o processo de ciência de dados não é rígido e estático, mas sim um processo que se flexibiliza à medida que novas necessidades vão surgindo durante o projeto. Assim, as técnicas aplicadas no pré-processamento também são utilizadas além da etapa inicial. Para isso, utilizamos a linguagem Python, com as bibliotecas pandas e NumPy para análise e manipulação de dados.

Os datasets originais possuem um grande volume de dados, tanto pela quantidade de participantes inscritos, quanto pela quantidade de informações armazenadas sobre eles, expressas pelas colunas (ou *features*). Uma consequência disso é uma maior utilização de recursos computacionais para processá-los, seja nos processos de leitura e escrita de arquivo quanto no armazenamento. Isso é particularmente importante pela inclusão de múltiplas fontes de dados. Assim, visamos diminuir a quantidade de dados pouco relevantes para a pesquisa, gerando dados menos volumosos e mais representativos.

Uma técnica aplicada na redução foi a seleção de *features*. Na seleção de *features*, visando diminuir a dimensionalidade do conjunto de dados, é gerado um subconjunto das *features* originais através da identificação e remoção daquelas pouco relevantes ou redundantes. Isso foi realizado no *dataset* do ENEM, que possui uma grande quantidade de colunas, e algumas informações, como as referentes à aplicação da prova, não são importantes na análise a ser realizada. Outra técnica aplicada foi a tipagem explícita de dados. Uma particularidade da biblioteca utilizada, pandas, é que a inferência de tipos nem sempre é a mais eficiente, o que pode levar a uma utilização de memória maior que o esperada, além de dificultar operações específicas, como manipulação de *strings* e realização de cálculos matemáticos. Para contornar esse problema, fizemos uma análise dos valores, a fim de mapeá-los para tipos de dados mais precisos, que pudessem melhorar os resultados da análise.

Outra característica dos *datasets* é que eles são imperfeitos. Apesar da presença de documentação auxiliar que define como os dados se comportam, na prática, os dados têm um estado diferente, incompletos e com "sujeiras". Como a qualidade dos dados interfere nos resultados obtidos, foi necessário observar características dos dados para definir uma abordagem adequada. Um dos problemas notados foi a ausência de valores em alguns campos. Por exemplo, foram encontrados registros de inscrições do SISU que não possuíam o curso de graduação do candidato. Optamos por não descartar esses registros que tivessem valores faltando, já que poderia gerar uma escolha não-aleatória e a inserção de potenciais viéses.

Visando solucionar essa questão, criamos novas classes de valores para representar uma instância de informação ausente e, quando possível, utilizamos funções de probabilidade

para inserir valores inferidos. Também tornamos os dados consistentes entre si, já que para a combinação das bases de dados, é necessário que os valores sejam do mesmo tipo e estejam uniformes em cada uma elas. Ao realizarmos uma análise de anomalias para detectar valores problemáticos, observamos uma situação particular na qualidade dos dados do SISU. Por conter mais de um *dataset* da chamada regular, era necessário agrupá-los para obter o cenário do ano como um todo, mas um deles estava parcialmente corrompido. A utilização inadvertida de dados não-estruturados pode levar a interpretações incorretas e falsas conclusões no processo de análise. Assim, realizamos um filtro dessas anomalias, relacionadas aos nomes das instituições de ensino, seus campi e cursos de graduação, que foram identificadas e tratadas para refletir o comportamento esperado.

4.3 Análise exploratória de dados

A partir da obtenção dos conjuntos de dados pré-processados, pudemos realizar uma análise exploratória de dados. Nesta etapa, iremos utilizar técnicas de manipulação de dados e ferramentas estatísticas para investigar os dados. Nela, conseguimos ter um entendimento melhor do cenário geral a ser explorado, como os dados estão distribuídos, quais são as variáveis, como elas se relacionam entre si e como podemos utilizá-las para responder as questões de pesquisa.

Por definição, o *dataset* do ENEM 2016 possui os dados do exame de estudantes de todo o país, que totalizam 8.627.367 registros. Cada registro corresponde à inscrição de um único estudante. No ato de realização do exame, esses participantes estão em diferentes situações em relação ao ensino médio, podendo já ter concluído, estar cursando ou não ter concluído e não estar cursando. Essa informação está expressa em duas variáveis: TP_ST_CONCLUSAO, Situação de conclusão do Ensino Médio, e Q046, que é a resposta do questionário socioeconômico à pergunta "Você já concluiu ou está concluindo o Ensino Médio?", visualizadas na Tabela 4.1. Em ambas, mais da metade dos estudantes já concluíu o ensino médio, seguidos daqueles que estão cursando e irão concluir em 2016.

Situação	TP_ST_CONCLUSAO		Q046	
	Quantidade	Percentual	Quantidade	Percentual
Já concluí o Ensino Médio	4.928.251	57,12%	4.947.935	57,35%
Estou cursando e concluirei o Ensino Médio em 2016	1.882.278	21,82%	1.872.570	21,70%
Estou cursando e concluirei o Ensino Médio após 2016	1.344.085	15,58%	1.331.073	15,43%
Não concluí e não estou cursando o Ensino Médio	472.753	5,48%	475.785	5,51%

Tabela 4.1: *Situação de conclusão do ensino médio dos candidatos do ENEM*

Outros dados relevantes são o ano e o tipo de ensino em que o participante concluiu o ensino médio. A primeira informação está expressa na variável TP_ANO_CONCLUIU, que explicita dos anos anteriores ao exame até 2007, agrupando o restante em anterior a 2007 ou não informado, que pode ser visualizado na Tabela 4.2. No caso de anos não informados, isso acontece tanto por haver valores ausentes da fonte, quanto pelos alunos

concluíntes não terem um ano de conclusão especificado (ainda a concluir). A segunda está expressa em TP_ENSINO, que descreve o tipo de ensino da instituição na qual o aluno concluiu o ensino médio, que pode ser visualizada na Tabela 4.3; nem todos os alunos possuem essa informação. Pode-se observar que um valor significativo dos estudantes não têm o ano de conclusão informado (42,88%), e, dentre os que informaram, a maior parte que concluiu no ensino regular (19,15%). Uma variável importante para essa análise da composição das turmas de ensino médio é CO_ESCOLA. Ela representa um código que identifica de maneira única a instituição junto ao Ministério da Educação. É através dela que poderemos associar a qual escola esse estudante pertence dentro da base do Censo Escolar. De todos os estudantes, apenas 21.81% possuem esse código no conjunto de dados de 2016.

Ano	Quantidade	Percentual
2015	966.842	11,21%
2014	699.987	8,11%
2013	527.310	6,11%
2012	416.454	4,83%
2011	317.364	3,68%
2010	294.214	3,41%
2009	244.461	2,83%
2008	199.619	2,31%
2007	178.091	2,06%
Anterior a 2007	1.083.909	12,56%
Não informado	3.699.116	42,88%

Tabela 4.2: *Ano de conclusão de ensino médio dos candidatos do ENEM*

Tipo de ensino	Quantidade	Percentual
Ensino Regular	1.652.485	19,15%
Educação Especial - Modalidade Substitutiva	10.295	0,12%
Educação de Jovens e Adultos	218.532	2,53%
Não informado	6.746.055	78,20%

Tabela 4.3: *Tipo de ensino da escola dos candidatos do ENEM*

Como nem todos os registros presentes no conjunto contêm todos os dados necessários para a pesquisa, estabelecemos alguns critérios para a aplicação de filtros. Dessa forma, apenas os seguintes registros foram mantidos em nossa base de trabalho:

- Registros que possuam informação de gênero, já que sem ela não podemos analisar o efeito diferenciado para homens e mulheres;
- Registros de estudantes que já concluíram o ensino médio ou que irão concluir em 2016, excluindo aqueles que concluirão após 2016 ou não estão cursando e não concluíram;
- Registros de estudantes que tenham concluído o ensino médio após 2007, já que não há especificação do ano anterior a 2007;

- Registros de estudantes que realizaram ensino médio no Brasil, excluindo aqueles que estudaram no exterior por não haver informação dessas escolas;
- Registros de estudantes que não estão realizando o exame como treineiros, já que treineiros não estão aptos a utilizar a nota do ENEM para ingresso em universidade;
- Registros que possuam informação da escola de ensino médio, já que sem ela não é possível identificar a escola em outra base;
- Registros que possuam informação de todas as notas obtidas no ENEM.

A Tabela 4.4 descreve as observações durante a aplicação dos filtros. Cada filtro foi aplicado sequencialmente, sendo que os valores mostrados correspondem ao total do conjunto original subtraindo os total de registros excluídos até o momento. A queda no percentual entre o sétimo e oitavo filtros ocorreu por conta de apenas aqueles que concluirão em 2016 possuem informação da escola de ensino médio. Após a aplicação de todos os filtros, restaram-se 1.388.044 registros, que correspondem a 16,09% do conjunto original. Esse subconjunto de dados filtrados do original é o que passa a ser utilizado nas análises subsequentes.

Filtro	Quantidade	Percentual
Conjunto original	8.627.367	100,00%
Com informação de gênero	8.627.367	100,00%
Já concluíram ou concluirão EM em 2016	6.810.529	78,94%
Concluíram após 2007	5.726.620	66,38%
Concluíram EM no Brasil	5.725.635	66,37%
Concluíram EM no ensino regular	5.496.816	63,71%
Não são treineiros	5.496.816	63,71%
Com informação presente da escola de EM	1.652.471	19,15%
Com informação presente de notas	1.388.044	16,09%

Tabela 4.4: *Filtros aplicados no conjunto de dados do ENEM*

Já no *dataset* do SISU 2017, os dados estão estruturados de forma diferente, com uma listagem (ou abertura) por semestre na realização do processo seletivo. Cada registro corresponde a uma inscrição de um estudante em um curso, totalizando 6.665.892 registros. Na Tabela 4.5, podemos observar a quantidade de inscrições e participantes inscritos, divididos por listagem. Em ambos os casos, a primeira listagem concentra a maioria dos registros.

Listagem	Inscrições		Inscritos	
	Quantidade	Percentual	Quantidade	Percentual
1	4.868.545	73,03%	2.494.173	72,72%
2	1.797.347	26,97%	935.538	27,28%
Total	6.665.892	100%	3.429.711	100%

Tabela 4.5: *Inscrições e inscritos no SISU, divididos por listagem*

Quando segmentado pela opção de cursos nos quais o candidato se inscreveu, que

pode ser de apenas 1 ou 2 cursos, pode-se perceber que 94,3% dos participantes opta por se inscrever nas duas opções disponíveis, como mostra a Tabela 4.6.

Opção de curso	Quantidade	Percentual
Apenas 1 curso	193.530	5,65%
2 cursos	3.236.181	94,35%
Total	3.429.711	100%

Tabela 4.6: *Quantidade de inscritos por cursos optados no SISU*

É possível que um candidato realize inscrição em somente uma das aberturas ou em ambas, observado na Tabela 4.7. Mais da metade (64,4%) optou por se inscrever apenas na primeira abertura.

Caso de inscrição	Quantidade	Percentual
Apenas na listagem 1	1.689.691	64,36%
Apenas na listagem 2	131.056	5%
Ambas as listagens	804.482	30,64%
Total	2.625.229	100%

Tabela 4.7: *Quantidade de inscritos por caso de inscrição em listagem no SISU*

Também é possível que o candidato tenha até 4 inscrições em cursos ao longo do ano, com o máximo de 2 por semestre. A Tabela 4.8 mostra a quantidade de inscritos pelo total de inscrições. A maior parte dos estudantes realizou 2 inscrições (65,08%), com o segundo maior total sendo de 4 inscrições (27.90%).

Total de inscrições	Quantidade	Percentual
1	116.802	4,45%
2	1.708.599	65,09%
3	67.420	2,56%
4	732.408	27,90%
Total	2.625.229	100%

Tabela 4.8: *Quantidade de inscritos pelo total de inscrições no SISU*

De forma semelhante ao caso do ENEM, foram realizados alguns filtros no *dataset* do SISU, sendo eles a exclusão de registros que não possuam informação de gênero e de registros que não possuam a informação de todas as notas. Não houve retirada de nenhum registro após a aplicação dos filtros, já que essas informações estavam disponíveis em todos eles. Também foi aplicada uma técnica de transformação de dados, de modo a agrupar todas as inscrições de um candidato em um único registro. Assim, houve uma redução de 60,62% no *dataset*, passando de 6.665.892 a 2.625.230 registros. Agora, o registro deixa de ser a inscrição de um candidato em um único curso para ser a inscrição no SISU como um todo.

Para identificar cada participante unicamente nas bases de dados, é necessário atribuir um campo como chave identificadora. Boas chaves candidatas para esse propósito seriam

o CPF ou número de inscrição. Diferente de outros estudos da literatura, estamos lidando com dados disponibilizados publicamente. Isso significa que informações sensíveis, tais como essas chaves, não foram fornecidas. Assim, foi necessário desenvolver uma nova chave a partir dos campos existentes. Para isso, realizamos uma investigação nas notas das provas, que foram agrupadas para observar se é possível utilizá-las como identificador único.

No *dataset* do ENEM, as notas agrupadas foram suficientes para identificar cada um dos alunos. Isso porque não há valores duplicados para essas notas, ou seja, cada aluno possui notas únicas. No SISU, isso não foi possível de ser observado, já que 22 notas estão repetidas. Essas notas pertencem a 75 inscritos. As notas repetidas se enquadram em dois casos: 4 notas objetivas zeradas, com exceção da redação (ex.: 0 - 0 - 0 - 0 - 520) ou todas as notas maiores que zero (ex.: 570,4 - 619,6 - 433,5 - 546,1 - 520). Não há caso de candidatos com todas as notas iguais a zero, pois o SISU impede a inscrição quando a nota da redação é zerada. Então, utilizamos outros campos para compor o identificador. O identificador final constitui-se das notas agrupadas, idade, unidade federativa de residência e gênero. Como esses campos estão disponíveis em ambas as bases, é possível identificar o mesmo candidato nas bases com a mesma chave identificadora.

Através do identificador criado, foi realizada a combinação das bases de dados. O resultado foi uma base única que associa todas as informações dos candidatos presentes nos *datasets* do ENEM e SISU. Nessa base, são encontrados 712.526 registros, que representam participantes encontrados tanto no SISU, quanto no ENEM. 27,14% dos participantes do ENEM foram encontrados no SISU, ao passo que 51,33% dos participantes do SISU foram encontrados no ENEM. Vale notar que, para se inscrever no SISU, é obrigatório a realização do ENEM, mas nesse estudo, estamos utilizando apenas uma parte dos dados do exame, por isso a correspondência de participantes do SISU no ENEM não é de 100%.

O conjunto de dados obtido será utilizado na combinação com uma terceira fonte, a do Censo Escolar. Através do Censo, poderemos caracterizar as escolas de ensino médio nas quais os estudantes se formaram. Os dados dessa fonte passarão por pré-processamento e análise exploratória, semelhante ao já realizado. Também com esses dados, realizaremos a aplicação da metodologia de variações idiossincráticas na composição de gênero através das coortes escolares [1]. Isso será realizado em etapas futuras da pesquisa, que serão descritas no Capítulo 6. O próximo capítulo apresentará os resultados da análise descritiva dos dados levantados até então. Essa análise está segmentada por gênero e por aprovação nos cursos.

[1] Para variável indicadora de curso, os cursos de graduação foram agrupados de acordo com a área do conhecimento, utilizando a classificação ISCED. As áreas do conhecimento foram organizadas nos seguintes grupos: Educação; Humanidades e Artes, Ciências sociais, Negócios e Direito; Engenharia, Manufatura e construção; Ciência; Agricultura; Saúde e bem-estar; Serviços; Não informado.

Capítulo 5

Resultados preliminares

Neste capítulo, apresentaremos os resultados preliminares obtidos através de uma análise descritiva. Os resultados estão segmentados por gênero e situação de aprovação nos cursos inscritos.

5.1 Perfil dos candidatos, instituições de ensino e cursos de graduação

Para descrever o perfil dos participantes e cursos de graduação, realizamos um levantamento através de algumas variáveis específicas. No conjunto de dados do ENEM, a análise foi segmentada por gênero, observado na Tabela 5.1. As mulheres constituem 58% do total de candidatos, ao passo que homens constituem 42%. Escolas públicas de nível estadual concentram a maioria dos estudantes, com mulheres tendo percentual superior em escolas estaduais (77,33%) e homens em escolas federais (2,99%) e privadas (22,32%). Em relação à escolaridade, a maior parte das mães dos candidatos possui ensino médio completo, enquanto a maioria dos pais não possui ensino médio completo. A escolaridade de pais de candidatos do gênero masculino é maior, 55,69% das mães e 44,72% dos pais possuem ensino médio ou superior. A faixa etária com mais estudantes é de 16 a 20 anos, sendo que a idade que possui mais candidatos é a de 17 anos. Dentro dessa faixa etária, mulheres são maioria nos 16 (4,81%) e 17 anos (53,82%), homens nos 18 (31,04%) e 19 anos (9,68%). Quanto à identificação racial, brancos e pardos concentram a maioria do total. Homens têm mais candidatos brancos e pretos, ao passo que as mulheres são maioria entre os pardos.

Já no SISU, a segmentação é realizada por situação de aprovação, observada na Tabela 5.2. Aprovados são 4,4% do total e reprovados 95,6%. Cursos de grau bacharelado e licenciatura e de turno integral e noturno são os que possuem mais inscrições. Cursos de licenciatura têm mais aprovados (26,08%) e bacharelado e licenciatura mais não aprovados (66,39% e 9,52%). Observamos que homens possuem mais aprovações (57,94%), ao passo que mulheres são a maioria dos não aprovados (53,51%). Ações afirmativas beneficiam mais da metade dos participantes, em ambos os segmentos.

Também é levantado o desempenho dos estudantes nas provas do exame. No ENEM,

	Gênero	Feminino	Masculino
Variável	**Descrição**		
Dependência administrativa da escola	Estadual	77,33%	73,80%
	Federal	2,21%	2,99%
	Municipal	0,92%	0,88%
	Privada	19,53%	22,32%
Escolaridade mãe	Com EM completo	50,26%	55,69%
	Sem EM completo	49,74%	44,31%
Escolaridade pai	Com EM completo	39,01%	44,72%
	Sem EM completo	60,99%	55,28%
Idade	15 anos ou menos	0,16%	0,11%
	16 anos	4,81%	3,65%
	17 anos	53,82%	49,28%
	18 anos	28,93%	31,04%
	19 anos	7,33%	9,68%
	20 anos	2,25%	3,33%
	21 a 30 anos	2,18%	2,66%
	31 a 40 anos	0,37%	0,17%
	41 a 50 anos	0,12%	0,06%
	50 anos ou mais	0,03%	0,02%
Raça	Amarela	2,45%	1,86%
	Branca	40,53%	41,57%
	Indígena	0,54%	0,61%
	Não declarado	1,44%	1,83%
	Parda	44,30%	42,40%
	Preta	10,73%	11,72%

Tabela 5.1: *Estatísticas descritivas dos participantes do ENEM por gênero*

a observação das notas em intervalos pode ser vista na Tabela 5.3. Em todas as provas, para ambos os gêneros, o intervalo de notas que possui a maior observação é o de (400, 600]. A distribuição em gráfico de densidade na Figura 5.1, divididas por gênero. As curvas de redação são similares para homens e mulheres, com notas que variam de 0 a 1000. Já as provas objetivas, as configurações mudam: as notas se concentram entre 250 e 800 em todas as provas, exceto em matemática, que tende a pouco mais de 900. Mulheres têm maiores picos em todas as provas. No SISU, isso pode ser visto na Tabela 5.4 e na Figura 5.2, divididas por situação de aprovação. Aprovados tendem a se concentrar nos intervalos de (600, 800], ao passo que não aprovados se concentram em (400, 600]. As curvas do gráfico são diferentes em todas as provas. Aprovados têm maiores picos em todas as provas, exceto matemática. As notas de aprovados tendem a se localizar em intervalos de notas maiores.

As estatísticas descritivas de notas e idades do ENEM estão disponíveis na Tabela 5.5 e do SISU na Tabela 5.6. Homens possuem maiores médias de notas objetivas e mulheres possuem maior média de nota em redação. Aprovados possuem maiores médias em todas as provas. Em ambos os conjuntos, realizamos um teste-t bicaudal para comparar as médias

	Situação	Aprovado	Não aprovado
Variável	**Descrição**		
Grau	Bacharelado	62,82%	66,39%
	Licenciatura	26,08%	21,62%
	Tecnológico	7,48%	9,52%
	Área Básica de Ingresso (ABI)	3,63%	2,46%
Gênero	Feminino	46,49%	57,94%
	Masculino	53,51%	42,06%
Modalidade	Ampla concorrência	45,78%	45,76%
	Ações afirmativas	54,22%	54,24%
Turno	Integral	44,27%	45,34%
	Matutino	12,34%	12,28%
	Noturno	35,45%	34,67%
	Vespertino	7,94%	7,70%

Tabela 5.2: *Estatísticas descritivas de cursos e participantes do SISU por situação de aprovação*

de homens e mulheres e aprovados e não aprovados. Considerando um α de 0,01, para todas as variáveis, o p-valor foi menor que o nível de significância ($p < 0,01$), rejeitando a hipótese nula de que as médias são iguais e sugerindo que a diferença nas médias é estatisticamente significativa.

A Tabela 5.7 mostra a classificação das áreas dos cursos por gênero. As áreas mais ocupadas por mulheres são Saúde e bem-estar e Ciências sociais, negócios e direito, enquanto as áreas mais ocupadas por homens são Ciência e Ciências sociais, negócios e direito. Com relação à proporção de homens e mulheres por área, homens são maioria nas áreas de Ciência (55,4%) e Engenharia, manufatura e construção (60,25%). Algumas áreas possuem uma diferença relevante entre os gêneros. Educação e Saúde e Bem-estar possuem mais que o dobro de candidatos do gênero feminino. Algo similar pode ser observado com o gênero masculino nas áreas de Ciência e Engenharia, manufatura e construção.

A Tabela 5.8 mostra a distribuição das universidades com mais inscrições. As 3 instituições com mais inscritos são as universidades federais do Maranhão, Rio de Janeiro e Fluminense. A Figura 5.3 mostra a distribuição das unidades federativas no SISU, divididas pelas inscrições em instituições de ensino e residência dos candidatos. Os estados que mais possuem inscrições em IES são Minas Gerais, Rio de Janeiro e Bahia, ao passo que os estados de residência com mais candidatos são Minas Gerais, São Paulo e Rio de Janeiro.

Ao final desta análise, podemos perceber diferenças entre os perfis dos candidatos quando segmentados por gênero e situação de aprovação. Homens e mulheres possuem perfis similares, apesar de terem desempenhos diferentes nas provas do ENEM. Aprovados e não aprovados também têm desempenhos diferentes, sendo que aprovados tendem a ter notas maiores em todas as provas. Além disso, mulheres fazem escolhas de cursos diferentes dos homens no SISU, com uma menor concentração em cursos STEM.

Intervalo		(0, 200]	(200, 400]	(400, 600]	(600, 800]	(800, 1000]
Prova	**Gênero**					
Ciências Humanas	Feminino	0,00%	4,06%	78,43%	17,51%	0,01%
	Masculino	0,00%	3,90%	71,81%	24,28%	0,02%
Ciências da Natureza	Feminino	0,00%	12,75%	80,99%	6,25%	0,00%
	Masculino	0,00%	9,14%	80,03%	10,80%	0,03%
Linguagens	Feminino	0,00%	4,18%	83,41%	12,42%	0,00%
	Masculino	0,00%	4,51%	81,05%	14,44%	0,00%
Matemática	Feminino	0,00%	22,11%	66,32%	11,02%	0,55%
	Masculino	0,00%	13,67%	63,26%	21,16%	1,90%
Redação	Feminino	0,24%	11,00%	57,36%	26,08%	5,32%
	Masculino	0,43%	15,48%	58,48%	21,72%	3,88%

Tabela 5.3: *Observações de notas dos participantes do ENEM por gênero*

Intervalo		(0, 200]	(200, 400]	(400, 600]	(600, 800]	(800, 1000]
Prova	**Situação**					
Ciências Humanas	Aprovado	0,00%	0,02%	20,60%	79,19%	0,18%
	Não aprovado	0,06%	1,95%	69,81%	28,16%	0,02%
Ciências da Natureza	Aprovado	0,00%	0,30%	53,09%	46,42%	0,19%
	Não aprovado	0,06%	7,80%	80,93%	11,19%	0,02%
Linguagens	Aprovado	0,00%	0,15%	42,26%	57,59%	0,00%
	Não aprovado	0,01%	2,44%	79,82%	17,73%	0,00%
Matemática	Aprovado	0,00%	1,47%	32,22%	57,44%	8,87%
	Não aprovado	0,01%	14,04%	64,61%	19,88%	1,46%
Redação	Aprovado	0,24%	0,22%	15,26%	53,90%	30,62%
	Não aprovado	0,12%	8,33%	55,70%	29,53%	6,32%

Tabela 5.4: *Observações de notas dos participantes do SISU por situação de aprovação*

Gênero	Feminino		Masculino	
Estatística	Média	Desvio padrão	Média	Desvio padrão
Variável				
Idade	17,69	1,96	17,77	1,64
Nota Ciências Humanas	532,12	72,75	544,01	76,72
Nota Ciências da Natureza	474,44	71,18	495,02	78,94
Nota Linguagens	522,25	68,01	524,28	71,02
Nota Matemática	476,82	96,86	518,36	115,32
Nota Redação	559,52	150,15	531,96	155,58

Tabela 5.5: *Estatísticas descritivas das notas e idade do ENEM por gênero*

Gênero	Aprovado		Não aprovado	
Estatística	Média	Desvio padrão	Média	Desvio padrão
Variável				
Idade	20,73	5,77	22,38	7,11
Nota Ciências Humanas	638,67	50,47	557,41	72,75
Nota Ciências da Natureza	593,83	69,85	497,64	78,74
Nota Linguagens	605,11	48,45	538,94	65,55
Nota Matemática	646,52	114,98	513,66	110,82
Nota Redação	746,67	120,40	585,07	133,80

Tabela 5.6: *Estatísticas descritivas das notas e idade do SISU por situação de aprovação*

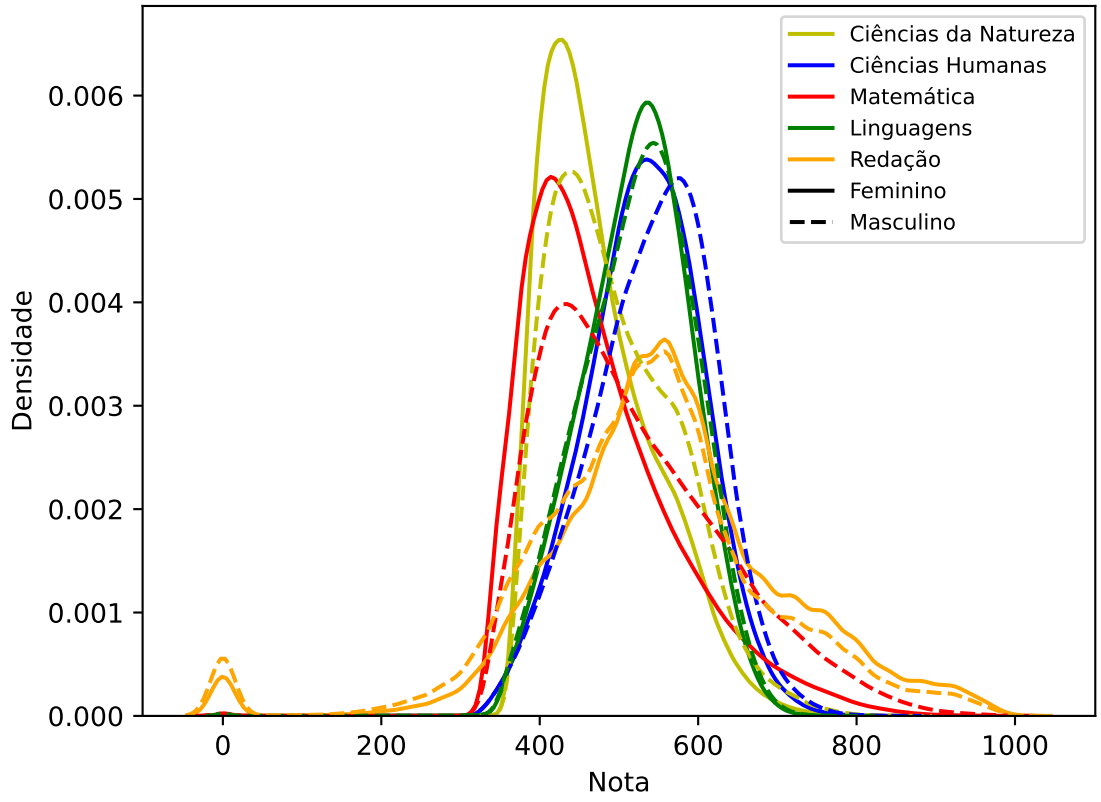


Figura 5.1: *Distribuição das notas do ENEM por gênero*

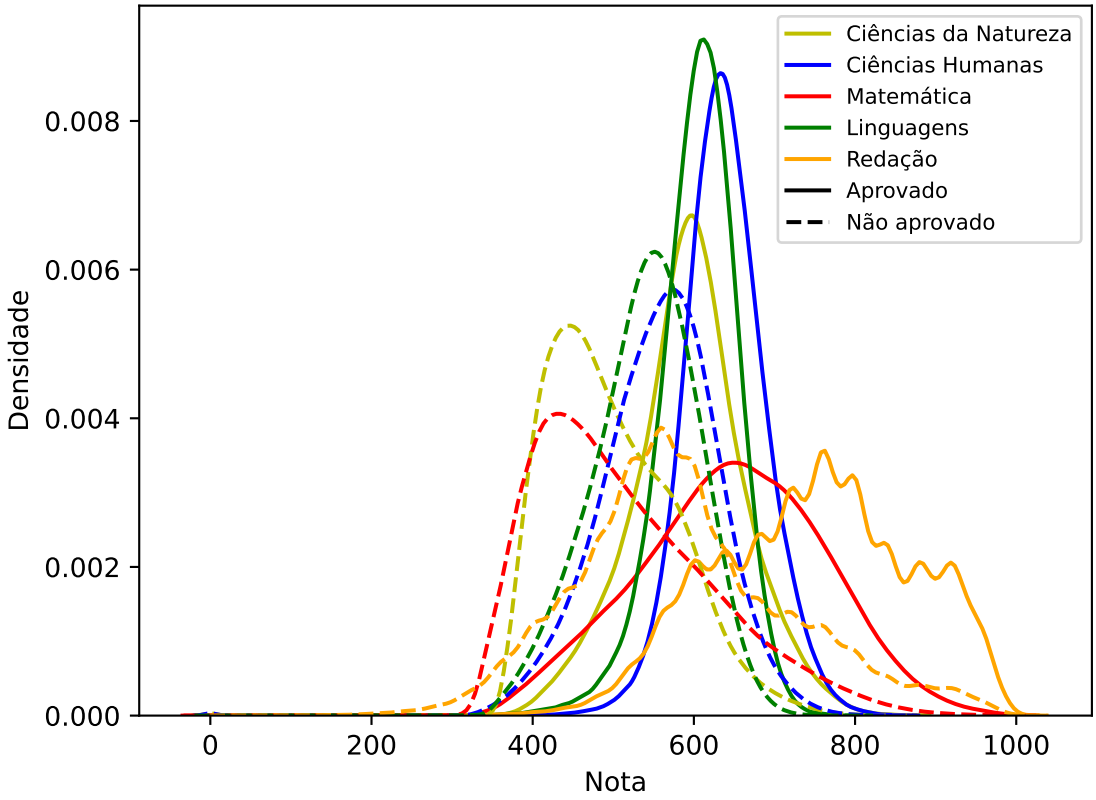


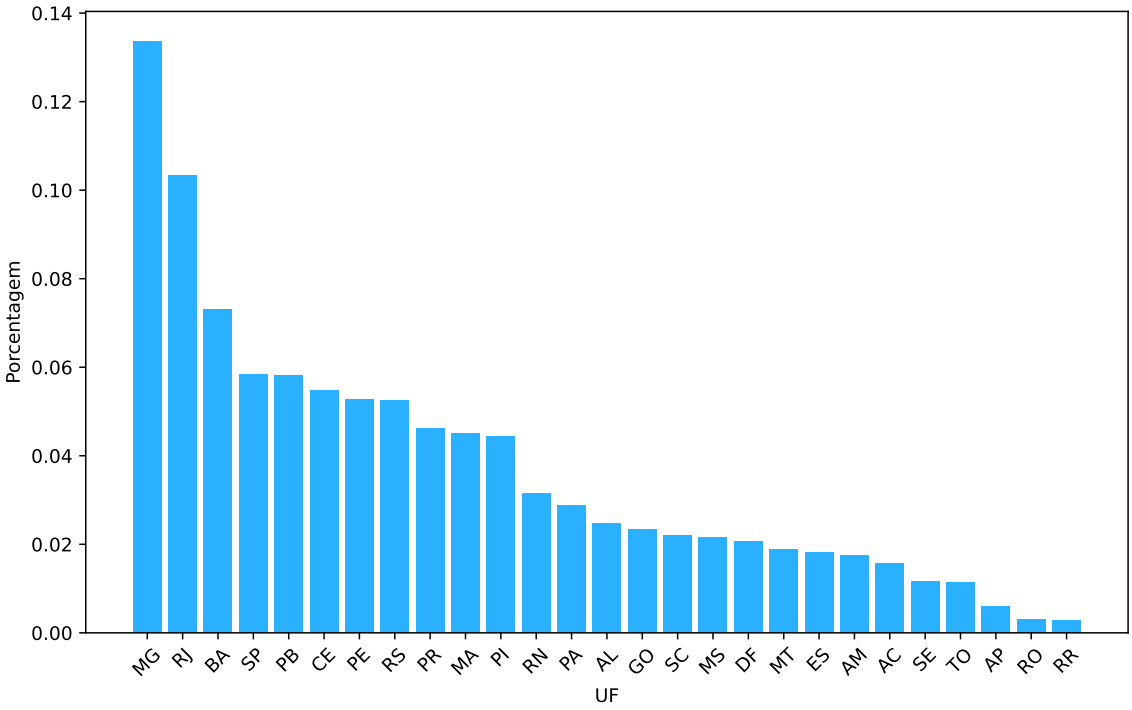
Figura 5.2: *Distribuição das notas do SISU por situação de aprovação*

Classificação	Feminino	Masculino
Educação	6,29%	1,81%
Humanidades e artes	8,94%	7,69%
Ciências sociais, negócios e direito	22,24%	20,85%
Ciência	13,01%	21,85%
Engenharia, manufatura e construção	9,76%	19,96%
Agricultura	5,19%	4,90%
Saúde e bem-estar	22,95%	10,42%
Serviços	6,53%	7,76%
Não informado	5,10%	7,76%

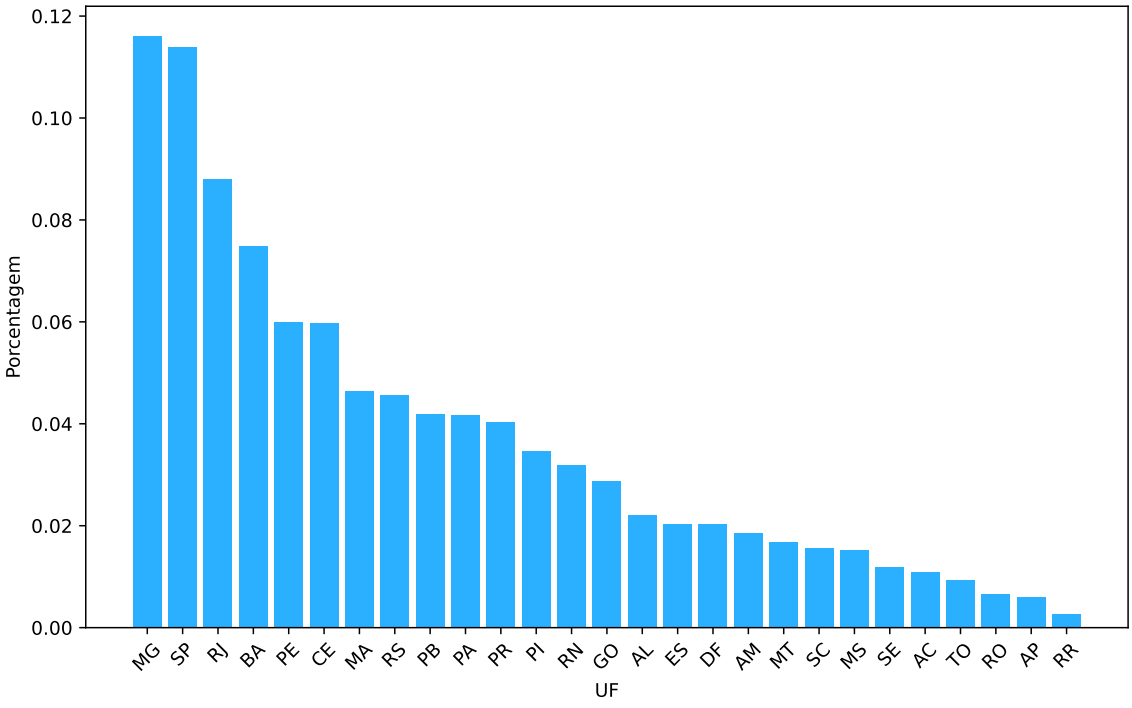
Tabela 5.7: *Distribuição da classificação dos cursos do SISU por gênero*

Instituição de Ensino	Porcentagem
Universidade Federal do Maranhão	3,79%
Universidade Federal do Rio de Janeiro	3,43%
Universidade Federal Fluminense	3,13%
Universidade Federal da Bahia	3,01%
Instituto Federal de Educação, Ciência e Tecnologia ds São Paulo	2,63%
Universidade Federal do Piauí	2,60%
Universidade Federal da Paraíba	2,60%
Universidade Federal de Minas Gerais	2,58%
Universidade Tecnológica Federal do Paraná	2,29%
Universidade Federal de Pernambuco	2,17%

Tabela 5.8: *Distribuição das 10 universidades com mais inscrições no SISU*



(a) *Inscrições em instituições de ensino*



(b) *Residência dos candidatos*

Figura 5.3: *Distribuição das unidades federativas no SISU*

Capítulo 6

Plano de trabalho

Nos capítulos anteriores, exploramos dois conjuntos de dados educacionais brasileiros, do ENEM e SISU, para entender como os estudantes estão caracterizados. Alguns aspectos levantados desse processo estão diretamente ligados com os objetivos desse trabalho, como as notas do exame, que refletem o desempenho dos participantes, e escolhas de curso e instituição de ensino, que variam de acordo com a segmentação. Para trabalhar com ambos os conjuntos numa visão única, desenvolvemos um novo identificador, que permite associar os registros e agregar todas as informações disponibilizadas apenas com campos públicos.

Apesar disso, não foi possível obter um panorama detalhado das escolas de ensino médio desses jovens, já que essas informações estão disponíveis no conjunto de dados do Censo Escolar, que ainda não foi explorado. Sem esses dados, também não é possível a aplicação do método de variações idiossincráticas na composição de gênero, conforme apresentado no Seção 2.2, já que ele precisa das informações das coortes de ensino médio. Então, para que possamos dar continuidade à pesquisa, levantamos as seguintes atividades:

- Levantamento das escolas dos estudantes presentes no conjunto ENEM/SISU através do código INEP;
- Realização de pré-processamento e análise exploratória dos dados do Censo Escolar, a fim de apresentar as características das escolas;
- Aplicação do método de variações idiossincráticas na composição de gênero, a fim de entender o efeito dos pares na escolha de graduação dos estudantes.

Realizaremos um refinamento dos resultados levantados até então, finalizando a exploração dos dados do ENEM e SISU, conforme levantado pelas duas primeiras questões de pesquisa (Q1 e Q2). Após a conclusão desta etapa, serão executadas as tarefas relacionadas aos dados do Censo Escolar para realização do estudo do efeito da composição de gênero, com o intuito de responder as questões de pesquisa restantes (Q3 e Q4). A Tabela 6.1 apresenta o cronograma das atividades restantes. Pretende-se realizar a defesa da dissertação em Agosto de 2024.

Atividades	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago
Refinamento de resultados e levantamento de escolas	X							
Pré-processamento e análise exploratória dos dados do Censo Escolar		X	X					
Aplicação do método de variações idiossincráticas na composição de gênero				X	X	X		
Escrita da dissertação					X	X	X	X
Defesa da dissertação								X

Tabela 6.1: *Cronograma de atividades (Jan 2024 - Ago 2024)*

Referências

[Abbagnano 2012] Nicola Abbagnano. *Dicionário de filosofia*. Mar. de 2012 (citado na pg. 2).

[MEC 2023] Portal Único de Acesso ao Ensino Superior. *SISU*. 2023. URL: https://acessounico.mec.gov.br/sisu (acesso em 13/10/2023) (citado na pg. 16).

[Akosah-Twumasi *et al.* 2018] Peter Akosah-Twumasi, Theophilus I. Emeto, Daniel Lindsay, Komla Tsey e Bunmi S. Malau-Aduli. "A systematic review of factors that influence youths career choices—the role of culture". Em: *Frontiers in Education* 3 (jul. de 2018). DOI: 10.3389/feduc.2018.00058. URL: https://doi.org/10.3389/feduc.2018.00058 (citado na pg. 1).

[Björkenstam *et al.* 2011] Charlotte Björkenstam *et al.* "School grades, parental education and suicide—a national register-based cohort study". Em: *Journal of Epidemiology & Community Health* 65.11 (2011), pgs. 993–998. ISSN: 0143-005X. DOI: 10.1136/jech.2010.117226. eprint: https://jech.bmj.com/content/65/11/993.full.pdf. URL: https://jech.bmj.com/content/65/11/993 (citado na pg. 6).

[Borchert 2001] Michael Borchert. "Career choice factors of high school students". Em: *Career choice factors* (nov. de 2001). URL: https://minds.wisconsin.edu/bitstream/handle/1793/40311/2002borchertm.pdf?sequence=1 (citado na pg. 2).

[Brenøe e Zölitz 2020] Anne Ardila Brenøe e Ulf Zölitz. "Exposure to more female peers widens the gender gap in STEM participation". Em: *Journal of Labor Economics* 38.4 (out. de 2020), pgs. 1009–1054. DOI: 10.1086/706646. URL: https://doi.org/10.1086/706646 (citado na pg. 12).

[Cabrera e Nasa 2000] Alberto F. Cabrera e Steven M. La Nasa. "Understanding the college-choice process". Em: *New Directions for Institutional Research* 2000.107 (2000), pgs. 5–22. DOI: 10.1002/ir.10701. URL: https://doi.org/10.1002/ir.10701 (citado na pg. 1).

[Caruana *et al.* 2015] Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez e Piergiorgio Solli. "Longitudinal studies". Em: *Journal of Thoracic Disease* 7.11 (2015). ISSN: 2077-6624. URL: https://jtd.amegroups.org/article/view/5822 (citado na pg. 5).

[Carvalhaes *et al.* 2022] Flavio Carvalhaes, Adriano S. Senkevics e Carlos A. Costa Ribeiro. "The intersection of family income, race, and academic performance in access to higher education in brazil". Em: *Higher Education* 86.3 (ago. de 2022), pgs. 591–616. issn: 1573-174X. doi: 10.1007/s10734-022-00916-7. url: http://dx.doi.org/10.1007/s10734-022-00916-7 (citado na pg. 13).

[Ensminger e Slusarcick 1992] Margaret E. Ensminger e Anita L. Slusarcick. "Paths to high school graduation or dropout: a longitudinal study of a first-grade cohort". Em: *Sociology of Education* 65.2 (abr. de 1992), pg. 95. doi: 10.2307/2112677. url: https://doi.org/10.2307%2F2112677 (citado na pg. 6).

[Inep 2023] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Exame Nacional do Ensino Médio (Enem)*. 2023. url: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem (acesso em 13/10/2023) (citado na pg. 15).

[García *et al.* 2016] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez e Francisco Herrera. "Big data preprocessing: methods and prospects". Em: *Big Data Analytics* 1.1 (nov. de 2016). doi: 10.1186/s41044-016-0014-0. url: https://doi.org/10.1186/s41044-016-0014-0 (citado na pg. 18).

[Gati e Saka 2001] Itamar Gati e Noa Saka. "High school students' career-related decision-making difficulties". Em: *Journal of Counseling & Development* 79.3 (jul. de 2001), pgs. 331–340. doi: 10.1002/j.1556-6676.2001.tb01978.x. url: https://doi.org/10.1002/j.1556-6676.2001.tb01978.x (citado na pg. 1).

[IBGE 2021] Instituto Brasileiro de Geografia e Estatística. *Estatísticas de Gênero - Indicadores sociais das mulheres no Brasil*. 2021. url: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101784 (acesso em 13/10/2023) (citado na pg. 2).

[Hoxby 2000] Caroline Hoxby. *Peer Effects in the Classroom: Learning from Gender and Race Variation*. Rel. técn. Ago. de 2000. doi: 10.3386/w7867. url: https://doi.org/10.3386%2Fw7867 (citado nas pgs. 7, 11).

[Jafari 2022] Roy Jafari. *Hands-On Data Preprocessing in Python*. en. Birmingham, England: Packt Publishing, mai. de 2022 (citado na pg. 18).

[Lavy e Schlosser 2011] Victor Lavy e Analía Schlosser. "Mechanisms and impacts of gender peer effects at school". Em: *American Economic Journal: Applied Economics* 3.2 (abr. de 2011), pgs. 1–33. doi: 10.1257/app.3.2.1. url: https://doi.org/10.1257/app.3.2.1 (citado na pg. 11).

[Machado e Szerman 2021] Cecilia Machado e Christiane Szerman. "Centralized college admissions and student composition". Em: *Economics of Education Review* 85 (2021), pg. 102184. issn: 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2021.102184. url: https://www.sciencedirect.com/science/article/pii/S027277572100100X (citado na pg. 12).

[Manski 1993] Charles F. Manski. "Identification of endogenous social effects: the reflection problem". Em: *The Review of Economic Studies* 60.3 (1993), pgs. 531–542. ISSN: 00346527, 1467937X. URL: http://www.jstor.org/stable/2298123 (acesso em 24/10/2023) (citado na pg. 6).

[Meister 1991] "Chapter 6 - idiosyncratic variables". Em: *Psychology of System Design*. Ed. por David Meister. Vol. 17. Advances in Human Factors/Ergonomics. Elsevier, 1991, pgs. 245–265. DOI: https://doi.org/10.1016/B978-0-444-88378-0.50011-4. URL: https://www.sciencedirect.com/science/article/pii/B9780444883780500114 (citado na pg. 8).

[Mello 2022] Ursula Mello. "Centralized admissions, affirmative action, and access of low-income students to higher education". Em: *American Economic Journal: Economic Policy* 14.3 (ago. de 2022), pgs. 166–197. DOI: 10.1257/pol.20190639. URL: https://doi.org/10.1257/pol.20190639 (citado na pg. 13).

[Otero *et al.* 2021] Sebastián Otero, Nano Barahona e Cauê Dobbin. "Affirmative action in centralized college admission systems: evidence from brazil". Em: (nov. de 2021). URL: https://siepr.stanford.edu/publications/working-paper/affirmative-action-centralized-college-admission-systems-evidence-brazil (citado na pg. 13).

[Saavedra *et al.* 2010] Luísa Saavedra, Maria do Céu Taveira e Ana Daniela Silva. "A subrepresentatividade das mulheres em áreas tipicamente masculinas: factores explicativos e pistas para a intervenção". Em: *Revista Brasileira de Orientação Profissional* 11.1 (2010), pgs. 49–59. ISSN: 1984-7270. URL: http://pepsic.bvsalud.org/pdf/rbop/v11n1/v11n1a06.pdf (citado na pg. 2).

[Sacerdote 2014] Bruce Sacerdote. "Experimental and quasi-experimental analysis of peer effects: two steps forward?" Em: *Annual Review of Economics* 6.1 (ago. de 2014), pgs. 253–272. DOI: 10.1146/annurev-economics-071813-104217. URL: https://doi.org/10.1146/annurev-economics-071813-104217 (citado na pg. 7).

[Schneeweis e Zweimüller 2012] Nicole Schneeweis e Martina Zweimüller. "Girls, girls, girls: gender composition and female school choice". Em: *Economics of Education Review* 31.4 (ago. de 2012), pgs. 482–500. DOI: 10.1016/j.econedurev.2011.11.002. URL: https://doi.org/10.1016/j.econedurev.2011.11.002 (citado na pg. 12).

[Schøne *et al.* 2019] Pål Schøne, Kristine von Simson e Marte Strøm. "Peer gender and educational choices". Em: *Empirical Economics* 59.4 (abr. de 2019), pgs. 1763–1797. DOI: 10.1007/s00181-019-01697-2. URL: https://doi.org/10.1007/s00181-019-01697-2 (citado na pg. 11).

[Senkevics 2021] Adriano Souza Senkevics. "O acesso, ao inverso: desigualdades à sombra da expansão do ensino superior brasileiro, 1991-2020". Tese de dout. Universidade de São Paulo, Programa de Pós-Graduação em Educação e Ciências Sociais: Desigualdades e Diferenças, 2021. DOI: 10.11606/t.48.2021.tde-11012022-103758. URL: https://www.teses.usp.br/teses/disponiveis/48/48137/tde-11012022-103758/pt-br.php (citado na pg. 2).

[Shahid Kazi e Akhlaq 2017] Asma Shahid Kazi e Abeeda Akhlaq. "Factors affecting students' career choice". Em: *Journal of Research and Reflections in Education* 11 (dez. de 2017), pgs. 187–196 (citado na pg. 1).

[Silva Borges 2021] Bruna Pugialli da Silva Borges. "Gender in higher education". Tese de dout. Universidade de São Paulo, Programa de Pós-Graduação em Economia, 2021. doi: 10.11606/t.12.2021.tde-27052021-215611. url: https://www.teses.usp.br/teses/disponiveis/12/12138/tde-27052021-215611/pt-br.php (citado nas pgs. 7–9, 12, 13).

[Song e Chung 2010] Jae W. Song e Kevin C. Chung. "Observational studies: cohort and case-control studies". Em: *Plastic and Reconstructive Surgery* 126.6 (dez. de 2010), pgs. 2234–2242. doi: 10.1097/prs.0b013e3181f44abc. url: https://doi.org/10.1097%2Fprs.0b013e3181f44abc (citado na pg. 5).

[Tang *et al.* 2008] Mei Tang, Wei Pan e Mark D. Newmeyer. "Factors influencing high school students' career aspirations". Em: *Professional School Counseling* 11.5 (jun. de 2008), pg. 2156759X0801100. doi: 10.1177/2156759x0801100502. url: https://doi.org/10.1177/2156759x0801100502 (citado na pg. 7).

[UNESCO 2006] UNESCO. *Juventude e contemporaneidade: possibilidades e limites*. pt. 2006. url: https://unesdoc.unesco.org/ark:/48223/pf0000154569 (citado na pg. 1).

[Venezia e Jaeger 2013] Andrea Venezia e Laura Jaeger. "Transitions from high school to college". Em: *The Future of Children* 23.1 (2013), pgs. 117–136. issn: 10548289, 15501558. url: http://www.jstor.org/stable/23409491 (acesso em 20/10/2023) (citado na pg. 1).

[Verdugo-Castro *et al.* 2022] Sonia Verdugo-Castro, Alicia García-Holgado e Mª Cruz Sánchez-Gómez. "The gender gap in higher stem studies: a systematic literature review". Em: *Heliyon* 8.8 (ago. de 2022), e10300. issn: 2405-8440. doi: 10.1016/j.heliyon.2022.e10300. url: http://dx.doi.org/10.1016/j.heliyon.2022.e10300 (citado na pg. 2).

[Zangirolami-Raimundo *et al.* 2018] Juliana Zangirolami-Raimundo, Jorge De Oliveira Echeimberg e Claudio Leone. "Research methodology topics: cross-sectional studies". Em: *Journal of Human Growth and Development* 28.3 (nov. de 2018), pgs. 356–360. doi: 10.7322/jhgd.152198. url: https://doi.org/10.7322%2Fjhgd.152198 (citado na pg. 5).

[Zimmerman 2003] David J. Zimmerman. "Peer effects in academic outcomes: evidence from a natural experiment". Em: *Review of Economics and Statistics* 85.1 (fev. de 2003), pgs. 9–23. doi: 10.1162/003465303762687677. url: https://doi.org/10.1162/003465303762687677 (citado na pg. 7).