Trabalho Prático 2

Manipulação e análise de conjuntos de dados públicos

Autores: Bernardo Alves, Cauã Magalhães, Pedro Moreira, Davi Couto, Lucas Albuquerque

Universidade Federal de Minas Gerais (UFMG)

Belo Horizonte – MG – Brasil

1 – Escolha dos conjuntos de dados

Foram escolhidos dois conjuntos disponibilizados pelo MEC no site dados.gov.br. Esses datasets contém as matrículas feitas em instituições de ensino superior (IES) usando o PROUNI e o conjunto de chamadas/matrículas feitas por estudantes pelo SISU no período de um ano (escolhemos 2018).

2 – Carregamento e combinação dos bancos de dados

Para carregar os dados em um SGBD com PostgreSQL, utilizamos a Neon Console. Neon é um serviço de PostgreSQL totalmente gerenciado e serverless que oferece uma gama de recursos destinados a aumentar a produtividade dos desenvolvedores e reduzir os custos operacionais. A escolha foi feita pela possibilidade de disponibilizar o banco de dados na nuvem, de forma simplificada e gratuita.

3 – Análise Exploratória dos dados

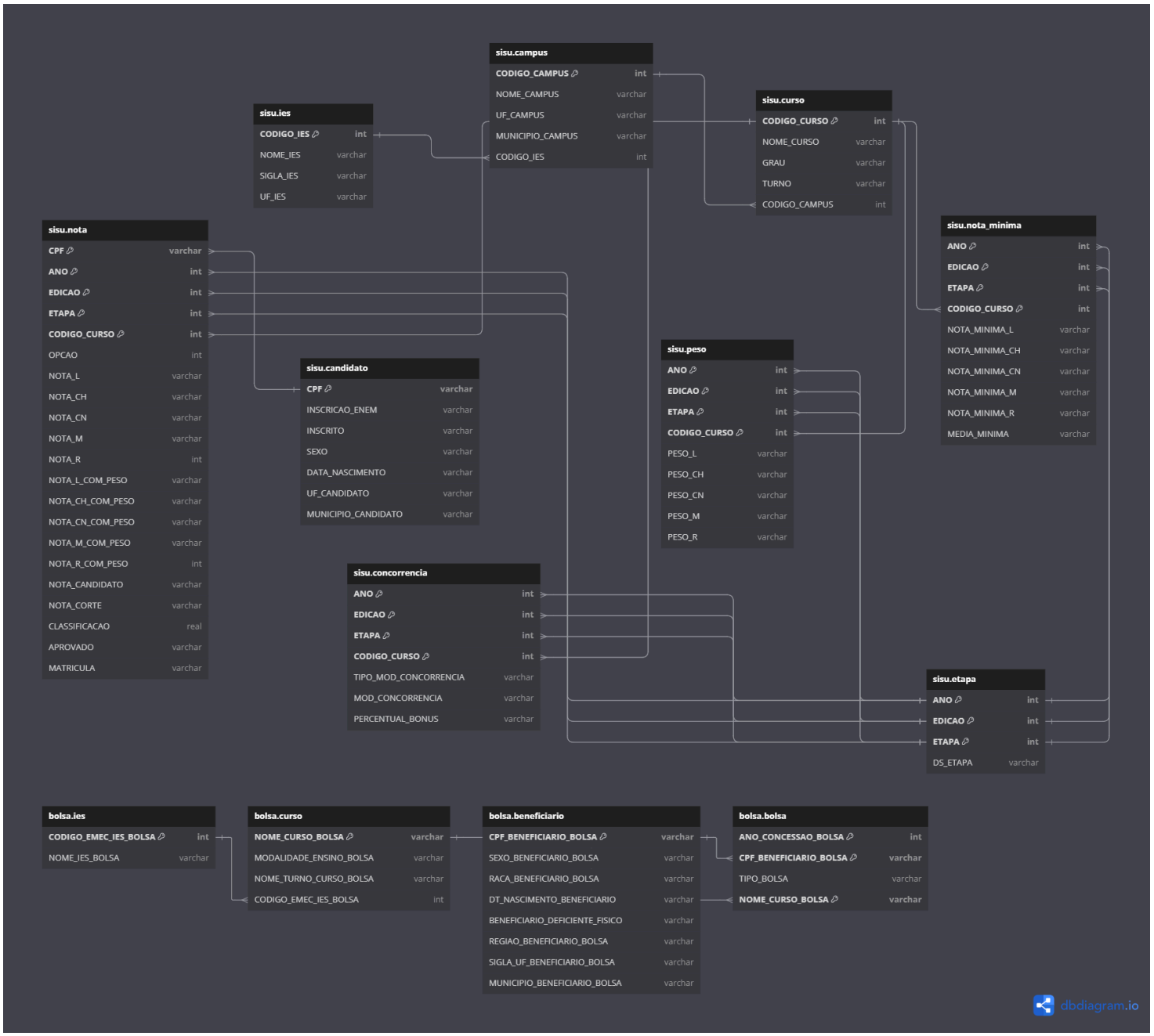
3.0 - Conjunto de metadados

Os dados foram obtidos dia 16/06/2024, e de acordo com o site citado, foram atualizados pela última vez em 22/11/2022. O conjunto não possui limitações conhecidas, os dados cobrem todo o território nacional e estavam dispostos em .csv

Repositório Utilizado

DATABASE_URL=postgresql://SISU_owner:EGIDVa13ytWm@ep-wild-tooth-a51fx8mc.us-east-2.aws.neon.tech/SISU?sslmode=require

Link para o dicionário



3.1 - Preparação dos dados

Os dados obtidos a partir dos bancos disponibilizados no site dados.gov.br, não apresentavam grandes inconsistências. Foi feita uma análise minuciosa nos conjuntos de dados (do SISU e do PROUNI) após a importação dos arquivos .csv para um SGBD, com base nisso, não detectamos valores faltantes, dados duplicados ou algo que pudesse prejudicar significativamente nossa análise. Vale ressaltar que resolvemos alguns problemas de formatação, mas nada muito grave.

3.2 - Objetivos da análise

Traçar o perfil dos estudantes que optam pelo PROUNI e o perfil daqueles que escolhem o SISU. Isso foi feito após análise de algumas estatísticas básicas descritas abaixo (3.3). Além disso, buscamos testar hipóteses com relação a influência de questões socioeconômicas no processo de ingresso no ensino superior, e analisar algumas preferências de ensino de pessoas com deficiência.

3.3 Análise descritiva

A análise foi feita com base nas consultas feitas no neon e nos seus resultados.

1- Preferência de período de estudos - PROUNI:

Segundo levantamento a partir do banco de dados, levando em consideração todas as modalidades de concorrência aos alunos do PROUNI , houve forte predileção pelo período noturno de estudos, fato este que pode ter relação com o aspecto rotineiro dos estudantes, como a necessidade de trabalhar ou estudar durante o dia. Em seguida, cursos a distância(EAD), foram a segunda opção mais escolhida, fato este predileto por pessoas com algum tipo de deficiência e por pessoas que preferem o estudo remoto ao presencial pelo conforto ou facilidade. Segundo levantamento, 44% dos estudantes estudam de forma integral, logo, o período integral foi o menos predileto pelos estudantes, o que denota certo descontentamento com o horário/turno vigente em que estão inseridos.

2- Preferência de período de estudos - SISU:

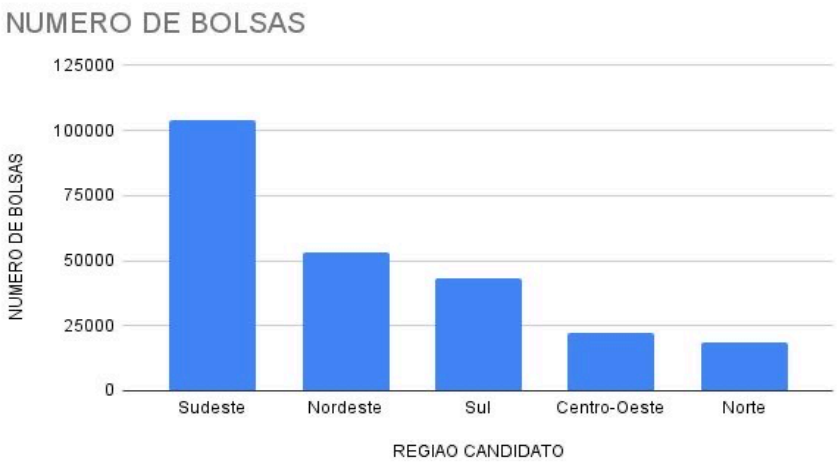
Para o levantamento dos alunos inscritos através do SISU, a predileção é diferente em alguns aspectos. O período noturno perde a preferência para o integral, tendo em vista que a opção de EAD deixa de estar disponível, os alunos optam por períodos em que suas necessidades sejam bem atendidas, levando em conta que a média de idade dos inscritos é 26.9 anos, o período integral se encaixa melhor para pessoas mais jovens. Além disso, os cursos mais escolhidos são para Medicina e Direito, logo a opção pelo período integral está relacionado com o grau de dificuldade desses cursos. Neste levantamento, as instituições de ensino com mais inscritos são a UFMG e a UFC, o que leva ao discernimento do alto nível de comprometimento com os estudos, logo a predileção pelo período integral.

3 - Preferência de período por pessoas com deficiência:

Já para alunos que apresentam algum tipo de deficiência o cenário não se altera por completo, ou seja, a predileção não é tão diferente. Ao que se nota na preferência também pelo período noturno de estudos, porém entre os alunos com algum tipo de deficiência existe uma predileção maior pelos cursos a distância, pelo fato de facilitar o acesso à educação por estas pessoas.

4 - Número de bolsas por região:

Levantamento sobre o número de bolsas recebidas por estados brasileiros. Dentre as bolsas oferecidas, a que se destaca é a bolsa para o curso de Direito e Administração, com mais de 17 mil bolsas de cada, distribuídas. Quanto à modalidade das bolsas, 76,5% são para cursos presenciais e apenas 23,5% para vagas EAD. No geral, a região mais beneficiada é a Sudeste com mais de 100 mil bolsas distribuídas em seguida a região Nordeste e Sul na faixa de 50 mil bolsas distribuídas.



5- Divisão de Bolsas - PROUNI:

A divisão das bolsas distribuídas aos alunos do PROUNI, dois terços são bolsas totais e o restante são parciais(50%). Ao todo são mais de 400 mil bolsas distribuídas pelo sistema, mais de 100 mil são contempladas pelo estado de São Paulo e em segundo o estado de Minas Gerais com 39 mil bolsas do PROUNI distribuídas. As bolsas parciais são para pessoas ter um resultado maior que um salário mínimo e meio e menor ou igual a três salários mínimos.

6 - Distribuição de Bolsistas por Raça - PROUNI:

A distribuição de bolsistas por raça é beneficiada pelo sistemas de cotas incluso no PROUNI, assim como o sistema do SISU que estabelece cotas econômicas e raciais, o sistema de cotas do PROUNI é apenas social, tendo em vista que o quesito econômico já está incluso no programa de aceitação, as bolsas são em parte reservadas às cotas sociais, para candidatos pretos, pardos e indígenas(PPI) e para candidatos com deficiência.

3.4 - Valores discrepantes

No conjunto de dados do SISU, em algumas regiões do país, os estudantes locais recebem um aumento percentual da nota geral de até 20% (medida tomada pelo MEC para que as IES sejam ocupadas pelas pessoas da região). Numa possível análise feita com as notas, esse fator pode gerar discrepâncias (notas extremamente altas) e deve ser levado em consideração.

4 - Análise crítica das fontes de dados utilizadas

1 – Dificuldades na Obtenção dos Dados:

O site dados.gov.br, utilizado para a obtenção das informações, apresentou diversas inconsistências no download de vários bancos de dados, impossibilitando o acesso a muitos deles. Além disso, o sistema de filtragem da plataforma deixou muito a desejar, pois as pesquisas frequentemente não retornavam os resultados corretos. Essa ineficiência prejudicou significativamente o processo de coleta de dados, impactando a qualidade e a integridade dos dados obtidos.

2 – Dificuldades no Carregamento dos Dados:

Ao carregar os dados na plataforma escolhida (Neon) utilizando o PostgreSQL 14, enfrentamos vários desafios. As limitações de armazenamento, devido ao tamanho do conjunto de dados, desaceleraram o processo de upload e ocasionaram algumas falhas graves. Além disso, problemas de formatação nos arquivos originais exigiram que os dados fossem tratados antes de serem carregados como um dataframe e, posteriormente, carregados na plataforma. Esse processo adicional de tratamento aumentou a complexidade e o tempo necessário para a conclusão do upload, uma vez que o banco do SISU teve que ser particionado para a realização do trabalho.

3 – Questões de Privacidade:

As notas do ENEM de todos os participantes do primeiro SISU de 2018 estavam presentes no conjunto de dados, vinculadas a nomes, endereços e partes dos CPFs. Isso constitui uma violação da LGPD (Lei Geral de Proteção de Dados), que exige que as notas do ENEM sejam mantidas em sigilo. A exposição desses dados pessoais sensíveis levanta sérias preocupações sobre a privacidade e a proteção dos indivíduos, exigindo medidas rigorosas para assegurar a conformidade com a legislação vigente e proteger os direitos dos titulares dos dados.

Além disso, a divulgação de informações sobre as universidades para as quais cada participante aplicou intensifica o risco de identificação pessoal e pode acarretar em danos irreparáveis à privacidade dos indivíduos afetados. Assim, é imperativo que o MEC adote práticas de anonimização e pseudonimização de dados, além de implementar políticas de transparência e segurança robustas

para prevenir futuras violações e garantir a confiança pública no tratamento de dados educacionais.

Embora os dados tenham sido coletados antes da vigência da LGPD, a sua disponibilização contínua representa uma não conformidade com a legislação atual. Portanto, é crucial que medidas imediatas sejam tomadas para corrigir essa situação e assegurar a proteção adequada dos dados pessoais.

5 – Análise da correlação entre os dados

Uma análise de correlação não é adequada neste cenário, pois a maioria dos dados não são numéricos, mas sim representam classificações em categorias como período, raça, e região. Portanto, não há dados que possam substituir outros. A análise integrada dos dados do PROUNI e do SISU revela insights importantes sobre a distribuição dos participantes, a escolha dos cursos e o perfil socioeconômico dos candidatos. Essa integração é crucial para compreender melhor o impacto e a eficácia dos programas de acesso ao ensino superior e pode orientar políticas futuras para aprimorar a inclusão e a equidade no sistema educacional brasileiro.

6 - Conclusões

A análise detalhada dos dados do Prouni e do Sisu de 2018, obtidos do site dados.gov.br, proporcionou uma visão abrangente e precisa sobre os perfis dos estudantes que optam por esses programas e suas preferências educacionais. A ausência de inconsistências significativas nos dados permitiu que a investigação fosse realizada de maneira eficiente, garantindo resultados confiáveis.

Os dados mostraram que estudantes com deficiência tendem a preferir cursos a distância, uma escolha que facilita o acesso à educação para esse grupo. A preferência pelo período noturno também é evidente, indicando uma necessidade de flexibilidade adicional. Essa informação é crucial para o desenvolvimento de políticas educacionais inclusivas que atendam melhor às necessidades específicas desses estudantes.

A análise revelou diferenças notáveis entre os estudantes do Prouni e do Sisu. Os dados do Prouni mostraram uma forte predileção pelo período noturno, possivelmente devido à necessidade de conciliar estudos com trabalho ou outras atividades diurnas. Além disso, a modalidade de ensino a distância (EAD) foi a segunda opção mais escolhida, destacando-se especialmente entre estudantes com deficiência e aqueles que preferem a flexibilidade e o conforto do estudo remoto. Em contraste, o período integral foi o menos preferido entre os estudantes do Prouni, refletindo uma possível insatisfação com essa modalidade.

No caso do Sisu, o período integral foi a preferência predominante, especialmente entre os cursos mais concorridos como Medicina e Direito. A média de idade dos inscritos (26,9 anos) sugere que estudantes mais jovens, possivelmente com menos responsabilidades familiares ou profissionais, estão mais inclinados a escolher o período integral. As instituições com maior número de inscritos, como a UFMG e a UFC, indicam um alto nível de comprometimento acadêmico, que pode estar associado à escolha do período integral.

A análise das influências socioeconômicas no ingresso ao ensino superior mostrou que ambos os programas, Prouni e Sisu, são eficazes em oferecer oportunidades a estudantes de diversas faixas de renda. No entanto, as preferências de período de estudo e modalidade refletem a necessidade de adaptar o ensino às circunstâncias pessoais e socioeconômicas dos estudantes. Por exemplo, a preferência por cursos noturnos e EAD entre os beneficiários do Prouni pode indicar uma maior necessidade de conciliar estudos com outras atividades, possivelmente devido a condições econômicas que exigem trabalho durante o dia.

Dessa forma, através do cruzamento de dados dos programas Prouni e Sisu, foi possível traçar um perfil detalhado dos estudantes e suas preferências, além de entender melhor como questões socioeconômicas influenciam suas escolhas. Os resultados obtidos podem guiar futuras políticas educacionais, promovendo uma distribuição mais equitativa de oportunidades e garantindo que as necessidades de todos os estudantes sejam atendidas de maneira eficaz. O trabalho contínuo de análise e aperfeiçoamento desses dados é fundamental para assegurar a evolução e a melhoria contínua dos programas de acesso ao ensino superior no Brasil.