Universidade Federal Fluminense
Programa de Pós-Graduação em Computação
Disciplina: Visualização de Dados

Relatório

Sistema de Visualização Interativa para os dados do Enem 2019

Docente: Marcos de O. L. Ferreira
Discentes: Marcos V. dos P. Oliveira
Henrique do P. Linhares

Niterói - RJ
2020

INTRODUÇÃO

No Brasil, o Exame Nacional do Ensino Médio (Enem) tem como o objetivo avaliar o desempenho do estudante ao fim da escolaridade básica e servir como meio de acesso ao ensino superior em quase todas as instituições de ensino superior do território nacional e em convênios com instituições de Portugal (BARROS, 2014; SILVEIRA; BARBOSA; SILVA, 2015). O Enem obtém anualmente dados sobre o conhecimento técnico de seus participantes, bem como informações socioeconômicas e culturais (DE CASTRO; TIEZZI, 2004).

Dada a importância do exame, neste trabalho selecionamos o conjunto de dados referente ao Exame Nacional do Ensino Médio (Enem) do ano de 2019, o qual está disponível no Portal Brasileiro de Dados Abertos[1]. Os dados são de fácil acesso, possuem ampla documentação, e são transferidos em um arquivo no formato CSV. Este arquivo contém 5.095.270 linhas e 136 colunas, onde cada linha representa uma inscrição no ENEM 2019 e cada coluna representa um atributo do candidato. Os atributos armazenam dados dos participantes, um questionário socioeconômico respondido pelo participante e dados dos pedidos de atendimento especializado e os dados da prova ("Microdados do Exame Nacional do Ensino Médio - Enem - Portal Brasileiro de Dados Abertos", [s.d.]).

O objetivo deste trabalho é apresentar um sistema de visualização interativa para os dados do Enem 2019. Nossa **motivação** é permitir que usuários utilizem o sistema para realizar estudos e análises sobre os dados do Enem 2019, assim como formular hipóteses, coletar evidências e realizar descobertas. O conhecimento extraído do sistema pode ampliar a compreensão sobre a situação da educação no Brasil, não só através das hipóteses que formulamos e discutimos, mas também pelas hipóteses que podem ser formuladas e discutidas por qualquer usuário do sistema.

DESAFIOS TÉCNICOS

Encontramos diversos desafios técnicos durante o desenvolvimento deste trabalho. O arquivo que transferimos do Portal Brasileiro de Dados Abertos pesa aproximadamente 3GB, e contém mais de 5 milhões de registros. Por conta de limitações de memória RAM, não foi possível carregar o conjunto inteiro na memória do computador para realizar as análises. Para contornar este problema, nós realizamos uma etapa de pré-processamento onde selecionamos apenas as colunas que iríamos usar no sistema de visualização. Além disso, aproveitamos a etapa de pré-processamento para criar atributos que armazenam resultados de contas (por exemplo, criamos o atributo média geral que armazena a média aritmética entre as notas obtidas por cada candidato). Dessa forma, otimizamos o tempo de geração dos gráficos.

Enfrentamos outros desafios relacionados com o alto número de registros do conjunto de dados. Não conseguimos criar visualizações que representem diretamente todos os indivíduos do conjunto, com um elemento do gráfico para cada indivíduo. Além disso, muitos gráficos do nosso sistema permitem o uso de filtros para selecionar os dados que serão exibidos. Em alguns casos, a aplicação pode demorar alguns segundos para filtrar os dados e construir o gráfico. Sabemos que o ideal é o usuário obter uma atualização do gráfico o mais rápido possível. Um caminho para atingir este objetivo poderia ser uma etapa de pré-processamento mais elaborada, em que houvesse uma

[1] https://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem

separação do conjunto de dados em subconjuntos específicos. Essa separação poderia reduzir o esforço computacional causado pelos filtros.

METODOLOGIA

O sistema de visualização foi projetado em três módulos independentes, cada um com suas próprias características, objetivos e funcionalidades. Os três módulos são denominados **caracterização**, **desempenho** e **presença**.

Caracterização

O primeiro módulo, caracterização, tem como objetivo fornecer uma visão geral dos dados, para que o usuário compreenda as características gerais dos indivíduos inscritos no Enem 2019. Na caracterização não há um foco específico em analisar hipóteses. O módulo de caracterização tem como objetivo responder as seguintes perguntas:

- (C1) Qual é a proporção entre homens e mulheres no Enem 2019?
- (C2) Qual é a distribuição de idades dos inscritos no Enem 2019?
- (C3) Qual é a proporção entre presentes e ausentes nas provas do Enem 2019?
- (C4) Qual é o estado civil dos participantes no Enem 2019?
- (C5) Qual é a raça / cor dos participantes do Enem 2019?

Desempenho

O módulo de análise de desempenho tem como objetivo analisar atributos que podem estar relacionados com o desempenho do participante nas provas do Enem 2019. O objetivo deste módulo é responder as seguintes perguntas:

- (D1) Qual é a relação entre a renda do candidato e o desempenho do candidato nas provas do Enem 2019?
- (D2) Qual é a relação entre a localização geográfica do candidato e o desempenho do candidato nas provas do Enem 2019?
- (D3) Qual a relação entre a escolaridade dos pais e o desempenho do candidato nas provas do Enem 2019?
- (D4) Existem outros atributos que possuem correlação com o desempenho do candidato no Enem 2019?

Para responder essas perguntas o módulo de análise de desempenho separa os participantes em grupos de acordo com as características de cada pergunta. Por exemplo, na análise da pergunta D1 os participantes são agrupados de acordo com grupos de renda familiar mensal, e então dados estatísticos (como média, mediana e quartis) são extraídos desses grupos e sintetizados em um gráfico.

Presença

Este módulo tem como objetivo analisar quais os atributos que possuem relação com a presença ou ausência dos inscritos nas provas do Enem 2019. Pretendemos responder as seguintes perguntas:

- (P1) Existe relação entre idade e presença nas provas?
- (P2) Existe relação entre o candidato possuir ou não um veículo e a presença do candidato nas provas?

- (P3) Existe relação entre o estado civil do candidato e a presença nas provas?

RESULTADOS

Caracterização

C1: Qual a proporção entre homens e mulheres no Enem 2019?

A Figura 1 apresenta a proporção entre homens e mulheres no Enem 2019. A ferramenta permite que o usuário aplique três filtros, podendo selecionar o intervalo de idade que deseja visualizar, podendo filtrar por todos os inscritos ou apenas pelos presentes nas provas, podendo selecionar indivíduos de todos os estados brasileiros, ou apenas de determinados estados.

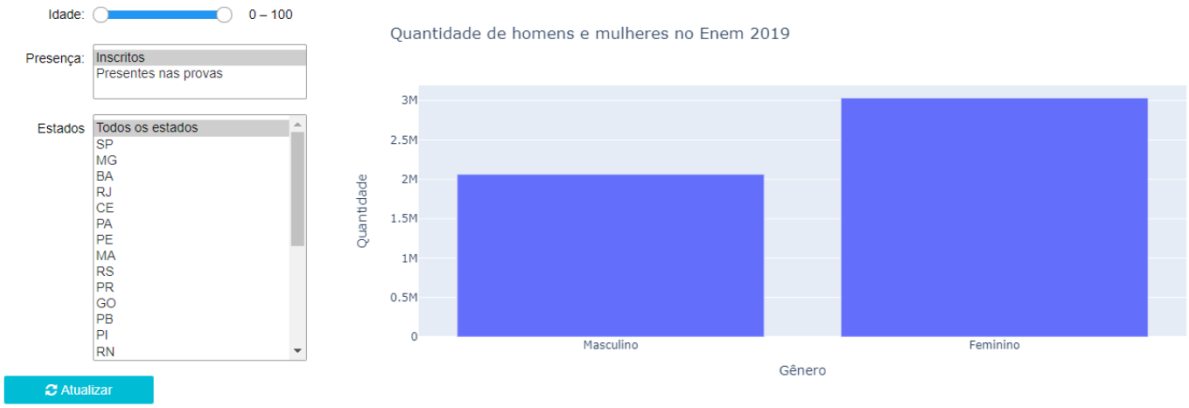

Figura 1. Proporção entre indivíduos do gênero masculino e feminino no ENEM 2019.

Ao analisar o gráfico gerado podemos observar uma quantidade maior de inscritos do gênero feminino quando comparado ao número de inscritos masculinos. Este padrão se repete em diferentes intervalos de idade, em diferentes estados, e tanto no conjunto de inscritos quanto no conjunto de presentes nas provas. Estes resultados nos trazem algumas dúvidas que não serão respondidas no escopo deste trabalho, mas podem ser analisadas em trabalhos futuros: (1) Qual é a proporção de homens e mulheres na população brasileira? (2) Como a proporção de homens e mulheres na população brasileira se compara com a proporção de homens e mulheres no ENEM 2019?

C2: Qual a proporção entre presentes e ausentes nas provas do Enem 2019?

A Figura 2 apresenta o resultado da distribuição de idade dos inscritos no Enem 2019. Podemos observar que as duas maiores frequências de idade são de 17 anos e 18 anos.

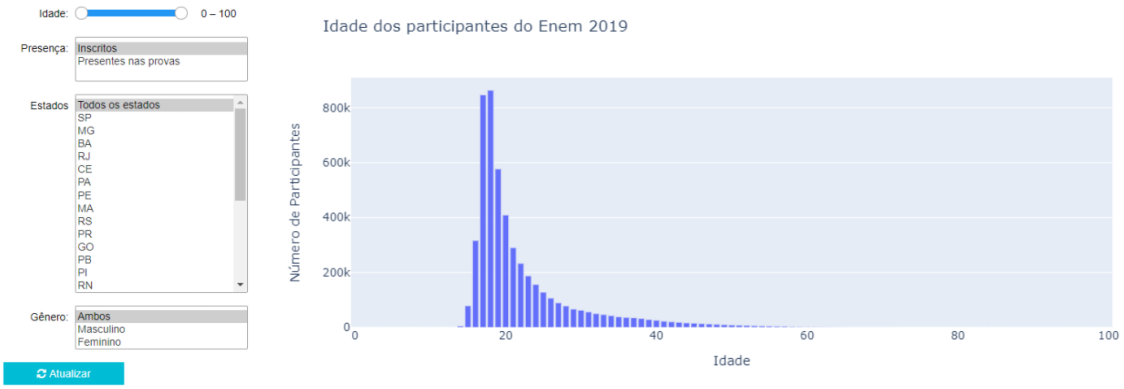

Figura 2. Distribuição de idade dos participantes do Enem 2019

Após 18 anos as frequências vão caindo progressivamente. Assim como no gráfico de caracterização por gênero, o gráfico de distribuição de idades também permite filtros por idade, presença, estado brasileiro. O gráfico apresentado na Figura 2 também suporta o filtro por gênero.

C3: Qual a proporção entre presentes e ausentes no Enem 2019?

A Figura 3 apresenta a proporção entre presentes e ausentes nas provas do Enem 2019. Vale notar que o ENEM 2019 foi composto por provas que foram aplicadas em duas datas distintas: As provas de Linguagens e Códigos e Ciências Humanas foram aplicadas no dia 3 de novembro de 2019, enquanto as provas de Ciências da Natureza e Matemática foram aplicadas no dia 10 de novembro. Podemos observar que o número de presentes nas provas que foram aplicadas no dia 3 (Linguagens e Códigos e Ciências Humanas) foi maior do que o número de presentes nas provas aplicadas no dia 10.

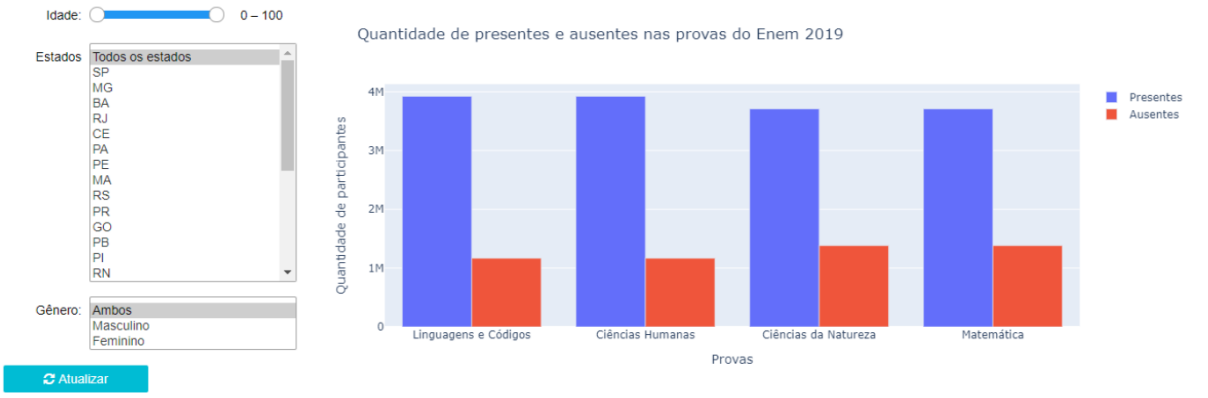

Figura 3. Proporção entre presentes e ausentes nas provas do Enem 2019

C4: Qual é o estado civil dos participantes do Enem 2019?

Para responder essa pergunta nós desenvolvemos uma ferramenta que permite a criação de gráficos customizados. Esta ferramenta permite que o usuário do sistema de visualização escolha qual atributo ele deseja visualizar, assim como os filtros que ele deseja aplicar, e a aplicação gera o gráfico de acordo com os parâmetros fornecidos, como ilustra a Figura 4.

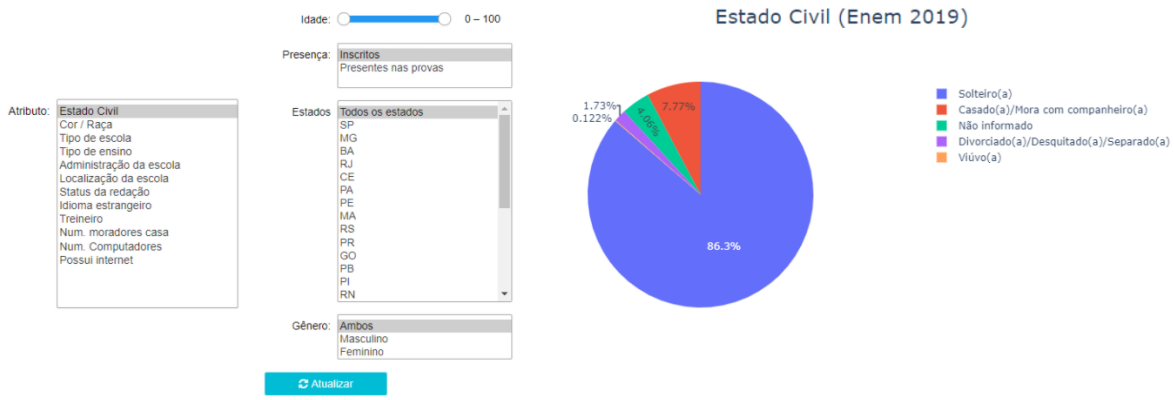

Figura 4. Estado Civil dos participantes do Enem 2019

Podemos observar que a maior parte dos inscritos no Enem 2019 são solteiros. É interessante notar que essa proporção de modifica significativamente de acordo com a idade selecionada.

C5: Qual é a cor / raça dos participantes do Enem 2019?

Nesta visualização também utilizamos a ferramenta de criação customizada de gráficos. Neste caso utilizamos um gráfico de barras para visualizar a distribuição de cor / raça dos participantes (vale lembrar que o atributo cor / raça é definido por autodeclaração), como mostra a Figura 5.

Figura 5. Distribuição de cor / raça dos participantes do Enem 2019

Desempenho

D1: Qual é a relação entre a renda do candidato e o desempenho do candidato nas provas do Enem 2019?

Nesta análise agrupamos os candidatos de acordo com a resposta da renda familiar mensal (respondida no questionário socioeconômico), e para cada grupo calculamos a média aritmética, o primeiro quartil (Q1, 25%) a mediana (Q2, 50%) e o terceiro quartil (Q3, 75%). A Figura 6 apresenta o resultado desta análise.

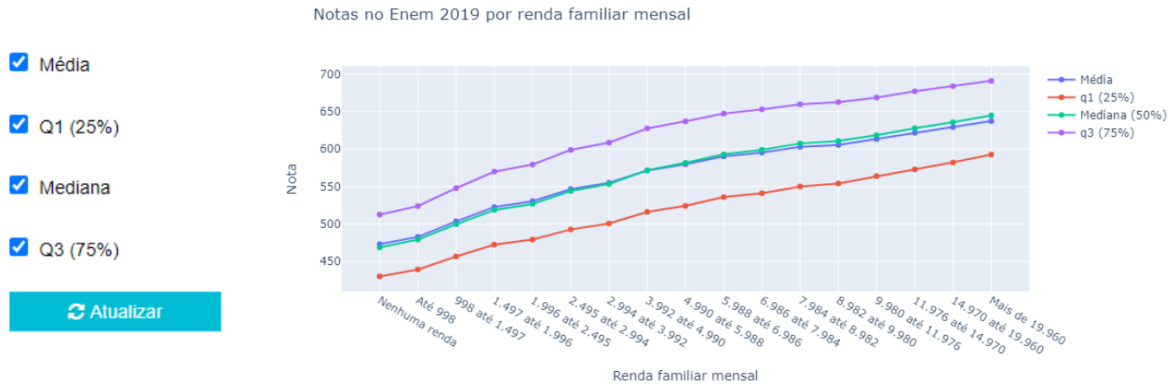

Figura 6. Relação entre nota (média/mediana) e renda familiar mensal dos participantes do Enem 2019

Os resultados trazem evidências de que existe uma correlação entre o desempenho do aluno na prova e a renda familiar. Conforme a renda familiar aumenta, o desempenho do participante na prova também aumenta.

Análises que podem ser feitas em trabalhos futuros são: (1) Analisar a renda mensal por indivíduo. Os participantes responderam no questionário o número de pessoas que moram na mesma residência que ele. Ao dividir a renda mensal familiar pelo número de moradores, poderíamos ter a renda mensal do participante, e construir uma análise para este atributo. (2) Analisar a relação entre renda mensal familiar e localização geográfica dos participantes. Considerando que que existem candidatos com diferentes rendas familiares tanto em regiões urbanas quanto em regiões rurais, será que existe relação entre a região e o desempenho do candidato? (3) Será que candidatos que possuem renda familiar elevada e residem em regiões urbanas vão apresentar o mesmo desempenho que candidatos que possuem renda familiar elevada e residem em regiões rurais? (4) E será que candidatos que possuem renda familiar baixa e residem em regiões urbanas vão apresentar o mesmo desempenho que candidatos que possuem renda familiar baixa e residem em regiões rurais?

D2: Qual a relação entre a localização geográfica do candidato e o desempenho do candidato nas provas do Enem 2019?

A Figura 7 ilustra a relação entre as notas médias dos candidatos e sua localização geográfica. Percebe-se que as regiões Sul e Sudeste apresentam as melhores médias, em contra partida a região norte tem o pior desempenho. A partir das constatações, algumas questões podem ser levantadas, tais como: (1) O desenvolvimento de uma região determina tais desempenhos? (2) As regiões como melhores médias têm maiores investimento em educação? (3) As políticas educacionais do sul e sudeste são mais bem elaboradas, por isso as regiões obtêm os melhores resultado? Inúmeras questões podem ser levantadas e discutidas com base nesse gráfico afim de fundamentar conclusões sobre tais dados.

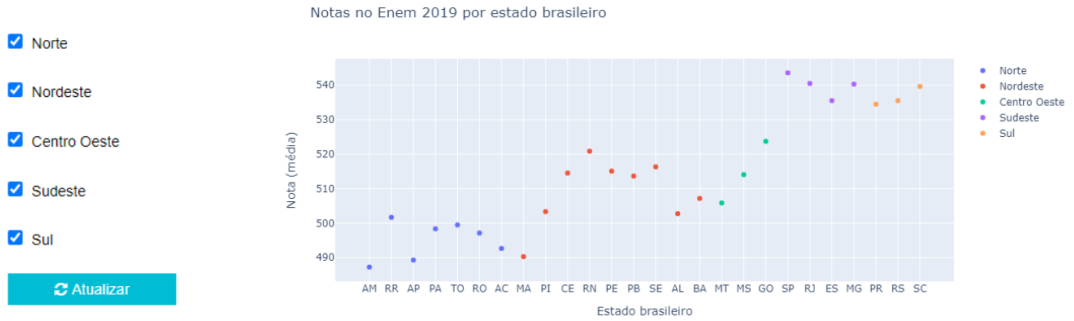

Figura 7. Distribuição das notas médias dos candidatos em função dos estados brasileiros

D3: Qual a relação entre a escolaridade dos pais e o desempenho do candidato nas provas do Enem 2019?

A Figura 8 apresenta a relação entre escolaridade dos pais e desempenho do candidato nas provas do Enem 2019. O usuário pode filtrar se deseja selecionar a escolaridade do pai ou da mãe.

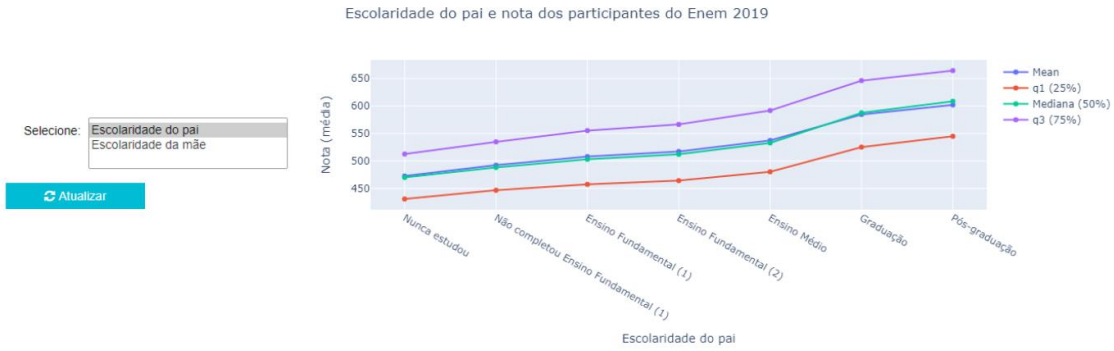

Figura 8. Relação entre escolaridade dos pais e desempenho dos participantes do Enem 2019

Os resultados trazem evidências de que existe uma correlação entre o desempenho do aluno na prova e a escolaridade dos pais. Conforme a escolaridade dos pais aumenta, o desempenho do participante na prova também aumenta. Não observamos diferenças significantes entre analisar a escolaridade do pai ou da mãe, pois em ambos os casos a correlação se apresenta da mesma forma.

D4: Existem outros atributos que possuem correlação com o desempenho do candidato no Enem 2019?

Além definir os principais fatores que se correlacionam com o desempenho de um candidato, é também de grande importância tentar verificar outros atributos que podem contribuir para o bom ou mau desempenho médio de um determinado candidato. Como ilustra a ferramenta da Figura 9.

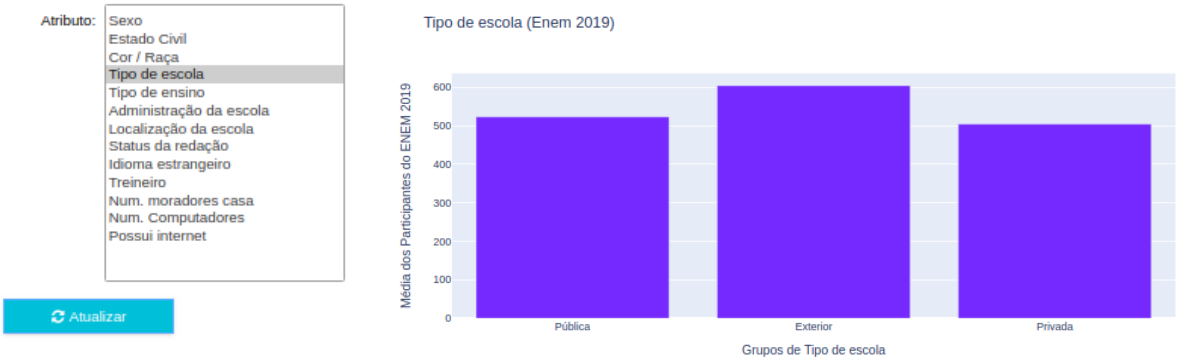

Figura 9 Comparativo de desempenho entre Homens e Mulheres com diferentes perfis

A partir dessa ferramenta é possível elencar e verificar o quanto outros atributos como o sexo, o estado civil, tipo de escola e outros se correlacionam com o desempenho do candidato. Assim, algumas análises podem ser feitas baseadas em hipótese como: (1) Mulheres têm melhor resultados que os homens? (2) Os candidatos de escola federal ou estadual apresentam os melhores resultado? (3) Os candidatos vim do exterior apresentam o melhor resultado? Entre outras questões que podem ser levantadas e discutidas.

Presença

Por fim, a Figura 10 ilustra uma ferramenta voltada para análises sobre as características dos candidatos ausentes na prova do Enem de 2019. Ela possibilita verificação de algumas hipóteses, tais como: (P1) Existe relação entre idade e presença nas provas? (P2) Existe relação entre o candidato possuir ou não um veículo e a presença do candidato nas provas? (P3) Existe relação entre o estado civil do candidato e a presença nas provas? Os filtros dessa ferramenta possibilitam elencar diferentes perfis desses candidatos visando identificar os candidatos com forte tendência a não comparecer no exame. Em média mais de 2 milhões de inscritos deixam de fazer a prova todos os anos.

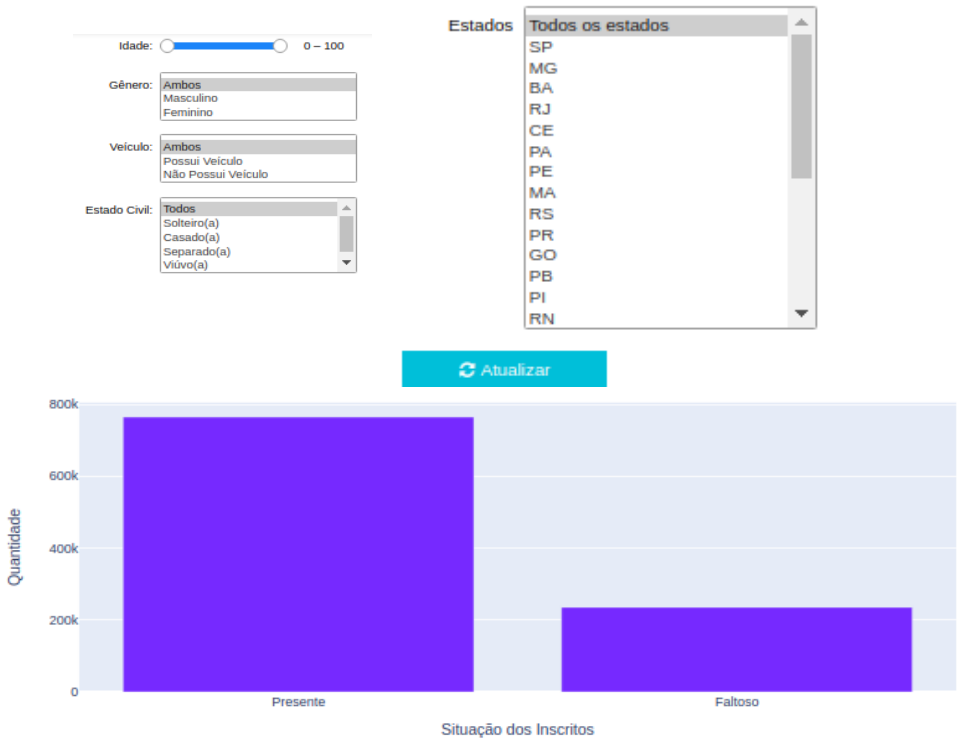

Figura 10. Caracterização dos candidatos presentes e ausentes no Enem 2019

Diversas hipóteses podem ser levantadas e discutidas a partir do levantamento de perfis dos candidatos do Enem, demonstrando o potencial de análise e discussões que essas ferramentas possibilitam.

REFERÊNCIAS

BARROS, A. DA S. X. Vestibular e Enem: um debate contemporâneo. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 22, n. 85, p. 1057–1090, 2014.

DE CASTRO, M. H. G.; TIEZZI, S. A reforma do ensino médio e a implantação do Enem no Brasil. **Desafios**, v. 65, n. 11, p. 46–115, 2004.

Microdados do Exame Nacional do Ensino Médio - Enem - Portal Brasileiro de Dados Abertos. Disponível em: <https://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem>. Acesso em: 29 nov. 2020.

SILVEIRA, F. L. DA; BARBOSA, M. C. B.; SILVA, R. DA. Exame Nacional do Ensino Médio (ENEM): uma análise crítica. **Revista Brasileira de Ensino de Física**, v. 37, n. 1, p. 1101, 2015.