



<b>Disciplina:</b>	Introdução a Ciência de Dados
<b>Aluno:</b>	Angela Favretto Pastorello - 1196243 Bruna Oenning Amador - 2241463 Daniel de Viveiros Inácio - 1923820 Matheus Biscaya Gutierrez - 1657470
<b>Equipe:</b>	Dados Educação

## ANÁLISE EXPLORATÓRIA

### 1. Tema

O tema do projeto é 'Analisar o impacto da pandemia na educação através da análise das desistências do ENEM', sendo o objetivo principal é classificar o quanto a pandemia impactou nas desistências do ENEM e buscar a relação dos resultados com os aspectos socioeconômicos, identificando o que obteve maior influência.

### 2. Obtenção e processamento de dados

Como foram utilizadas diferentes fontes para a realização do projeto, as próximas seções descrevem tais fontes e os procedimentos de limpeza e processamento realizados para cada uma delas, incluindo etapas de integração de diferentes fontes.

#### 2.1 Enem

Foram utilizados os dados do Enem que podem ser acessados pelo site do INEP<sup>1</sup>. Por ser uma prova que é realizada anualmente, é possível fazer uma comparação com edições anteriores e analisar como os aspectos socioeconômicos influenciaram nas desistências durante o período de pandemia.

Os dados do ENEM de 2020 possuem dados de participantes, da escola e do local de aplicação da prova de cada participante, dados da prova objetiva e da redação e, por fim, dados do questionário socioeconômico. As edições anteriores podem ter algumas diferenças com relação ao que foi disponibilizado, dessa forma, serão utilizados somente os dados citados acima.

Para as análises, foram consideradas somente pessoas que estão terminando o Ensino Médio no ano em que a prova foi aplicada, ou seja, que se pronunciaram como 'Estou cursando e concluirei o Ensino Médio em 2020'. Também, os dados numéricos foram alterados ou sinalizados de acordo com seus respectivos significados, para facilitar a visualização e compreensão da análise.

<sup>1</sup> <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>



## 2.2 Produto Interno Bruto dos Municípios

Para analisar a relação das desistências do ENEM de 2020 com aspectos econômicos por município e estado, foi utilizado dados do Produto Interno Bruto dos Municípios (PIB), que podem ser acessados pelo site do IBGE<sup>2</sup>. Como não foram disponibilizados dados atuais, a equipe utilizou os dados de 2019 para a análise.

Dentre os dados disponibilizados, foram utilizados somente o PIB per capita e os nomes dos municípios e estados para a análise. Tendo isso em vista, foi necessário alterar o tipo de dado dos valores do PIB per capita para um campo numérico e realizar a junção com a tabela dos dados do ENEM, descrita na Seção 2.1. Para a análise de acordo com o município, foi necessária a junção dos códigos dos locais de prova em que participantes realizaram o ENEM com os respectivos códigos dos municípios presentes na tabela de Códigos de Municípios, descrita na Seção 2.3.

## 2.3 Tabela de Códigos de Municípios

A Tabela de Códigos de Municípios do IBGE apresenta a lista dos municípios brasileiros associados a um código composto de 7 dígitos, sendo os dois primeiros referentes ao código da Unidade da Federação. Ela foi utilizada para obter o nome dos municípios de acordo com os códigos disponibilizados na tabela de dados do ENEM.

## 2.4 Dados sobre a pandemia

Os dados do Ministério da Saúde<sup>3</sup> detalham a situação da pandemia durante cada dia até a data da consulta, e inclui ambas as mudanças diárias dos casos e das mortes e a situação das regiões até uma determinada data. Os dados estão distribuídos por município, estado e macrorregião. Juntamente com as estimativas populacionais do Tribunal de Contas da União (TCU) de 2019, foram determinadas a quantidade de casos e mortes per capita para cada município.

## 2.5 Índice de Desenvolvimento Humano Municipal (IDHM)

O Índice de Desenvolvimento Humano Municipal (IDHM) é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda, sendo que o índice varia de 0 a 1 e quanto mais próximo de 1, maior o desenvolvimento humano.

Para a análise, foram utilizados os dados de 2010, considerando somente os dados das três dimensões apresentadas acima e do próprio IDHM. Também, foi necessário utilizar

---

2

<https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html>

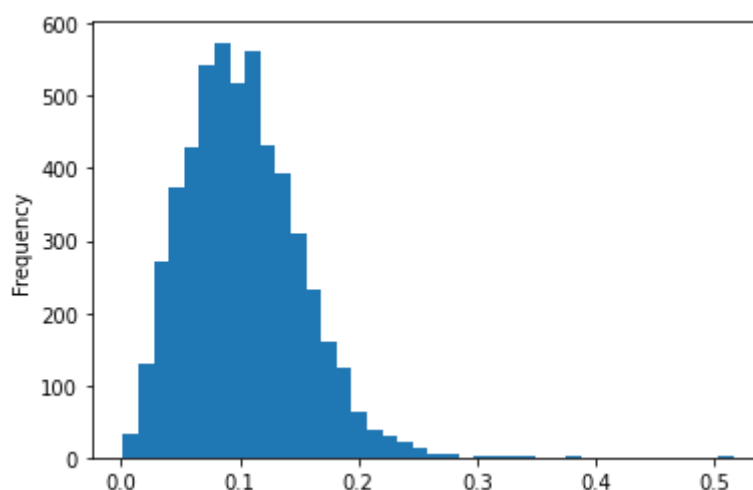
<sup>3</sup> <https://www.gov.br/saude/pt-br>



a junção dos dados do ENEM com os códigos das cidades, para então poder fazer a correlação dos dados de abstenção do ENEM com os dados do IDHM.

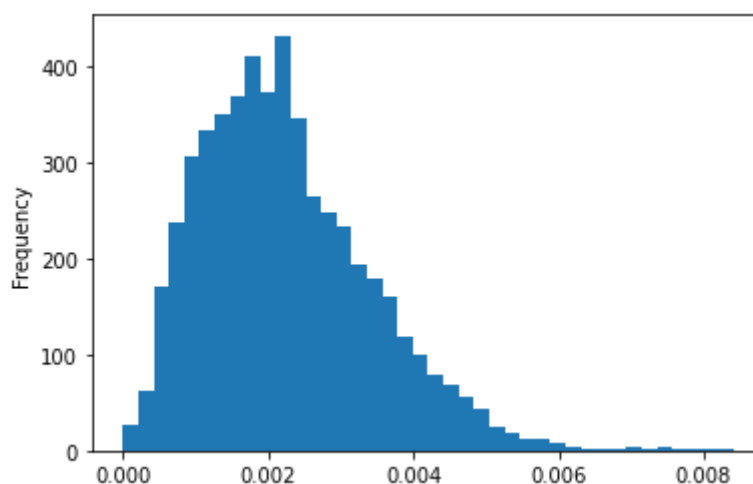
### 3 Cobertura e distribuição dos dados

A situação da pandemia até a data do ENEM mostra uma distribuição de casos que varia entre 0.0 e 0.2 casos per capita, onde o valor médio é 0.102 e o valor mediano é 0.098, com alguns poucos outliers, como pode ser visualizado no Histograma 1.



**Histograma 1 - Casos per capita por município**

Para as mortes per capita, conforme apresentado no Histograma 2, a distribuição varia entre 0.0 e 0.004 (0,4%), com alguns poucos outliers.



**Histograma 2 - Mortes per capita por município**

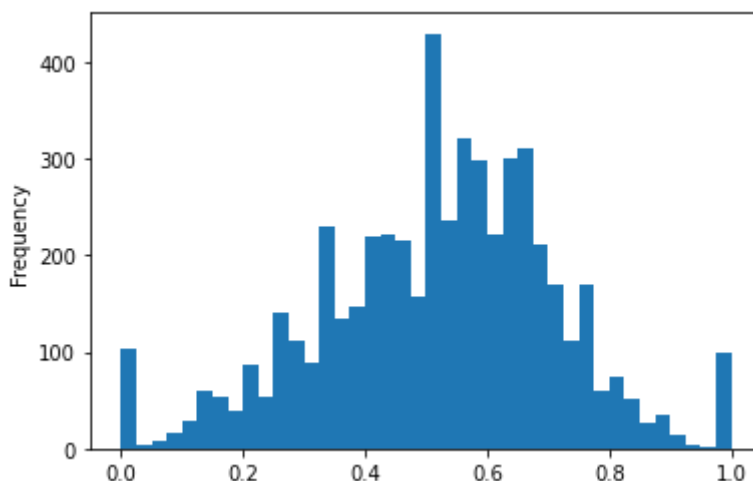


No que tange a situação durante a pandemia, foram analisados no total os casos para 5541 municípios, e as estimativas populacionais do TCU de 2019 foram utilizados para determinar os casos e mortes per capita.

Para a presença durante o ENEM, foi considerada a relação entre o total de alunos inscritos e o total de alunos que compareceram para todas as provas, onde um único dia de falta implica em desistência. Todos os alunos inscritos foram considerados para esta análise.

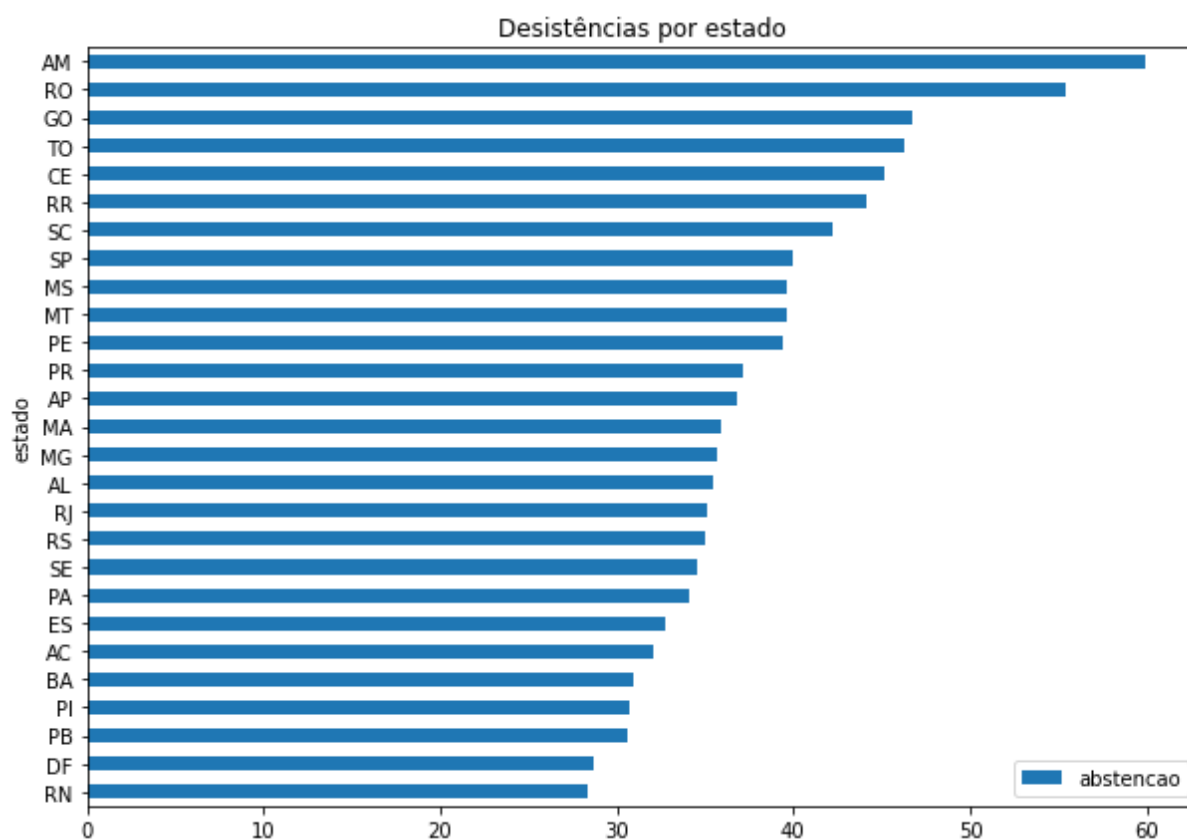
Foi observado que a proporção de desistência está distribuída igualmente entre os extremos, com uma proporção de presença média de 0.519 e mediana de 0.533.

No Histograma 3 estão representados municípios pouco populosos, onde existem poucos alunos inscritos no ENEM, e para esses casos existem situações onde os poucos alunos inscritos acabaram registrando presença ou falta total, o que explica a grande quantidade de casos presentes nos extremos.



**Histograma 3 - Relação de presença por total de alunos cadastrados, por município**

Como análise dos dados do ENEM, foi verificado inicialmente a percentagem de desistências por estado, sendo visível que há maior abstenção na região Norte.

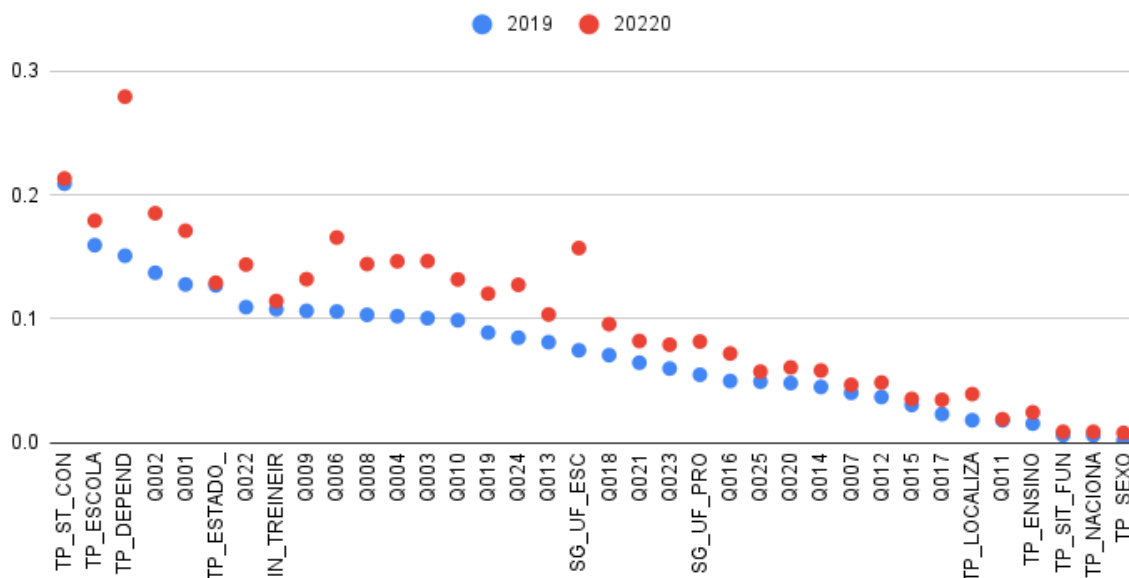


**Gráfico 1 - Desistências por estado**

Também, foi analisada a correlação entre as possíveis variáveis explicativas entre os dois anos, conforme apresentado no Gráfico 2.

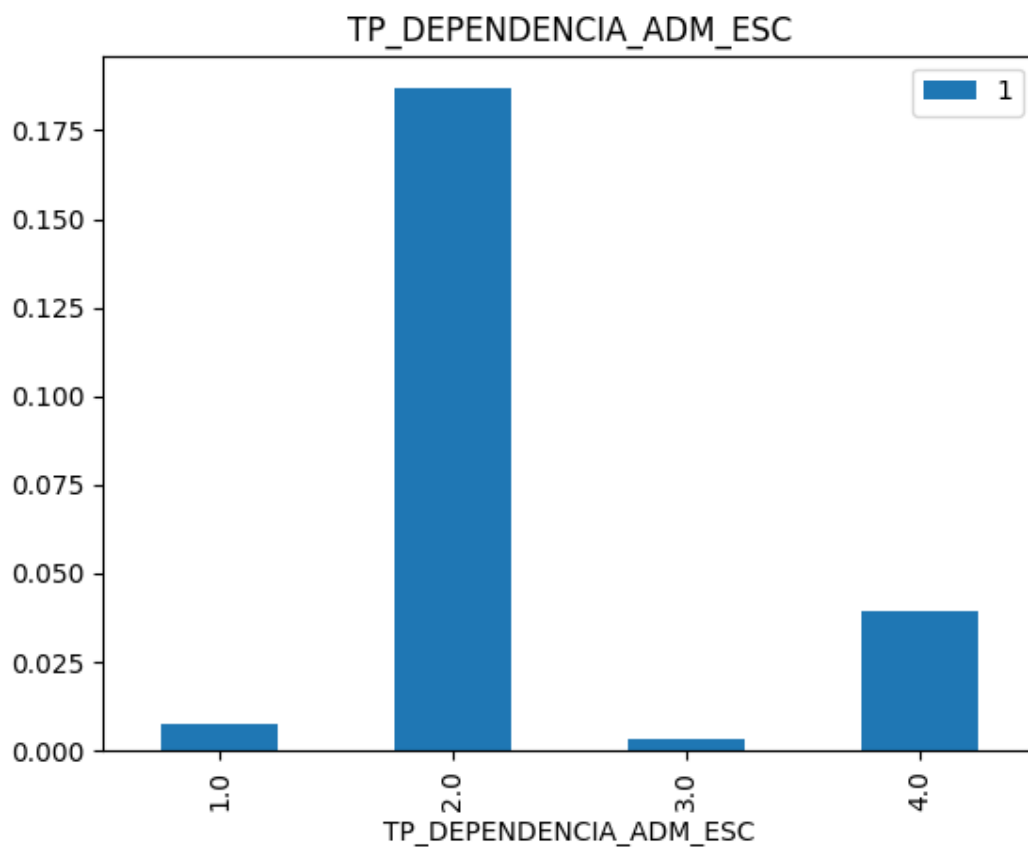


## 2019 y 2020



**Gráfico 2 - Correlação entre as possíveis variáveis explicativas entre os dois anos**

É visível que as variáveis que tiveram um aumento significativo são Dependência administrativa (Escola), renda mensal e o estado da escola do aluno, a seguir veremos alguns gráficos específicos dessas variáveis

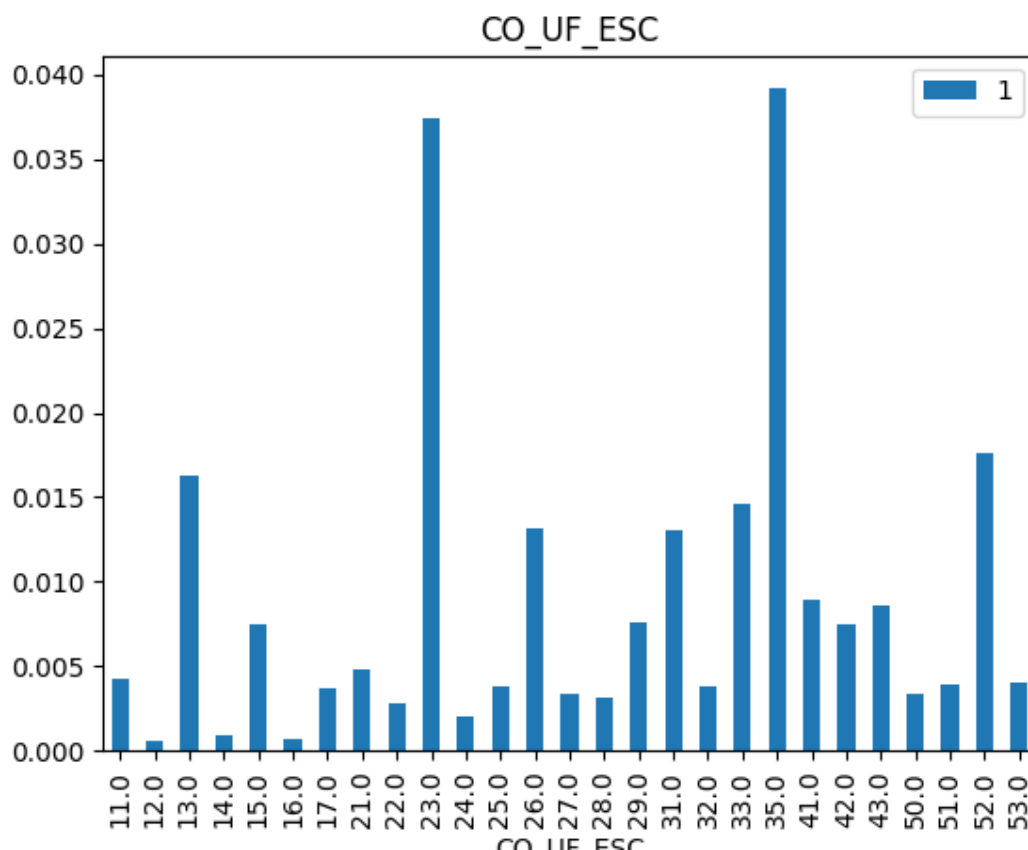


**Gráfico 3 - Tipo de dependência administrativa**

Houve uma porcentagem de aumento de desistência de 2019 para 2020, como mostra o Gráfico 2, tendo o maior aumento estudantes de escolas estaduais.

Para os Gráficos 4 e 5 são utilizados os seguintes códigos no primeiro dígito:

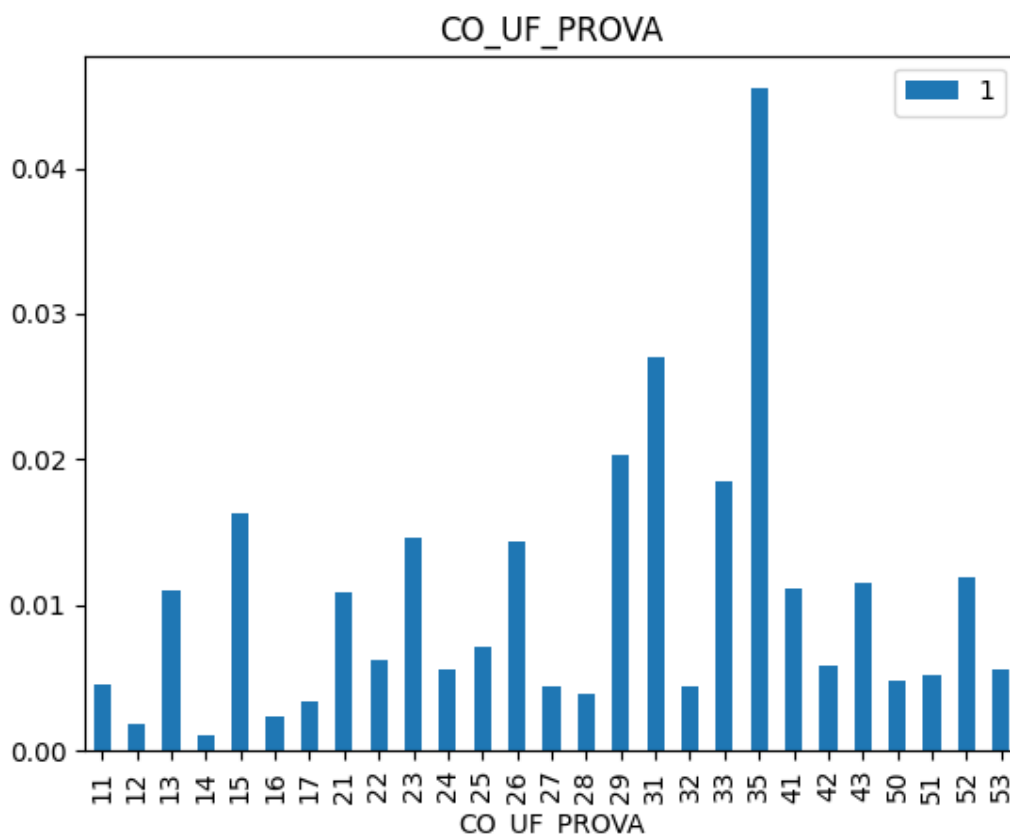
1x	norte
2x	nordeste
3x	sudeste
4x	sul
5x	centro-oeste



**Gráfico 4 - Estado da escola do aluno**

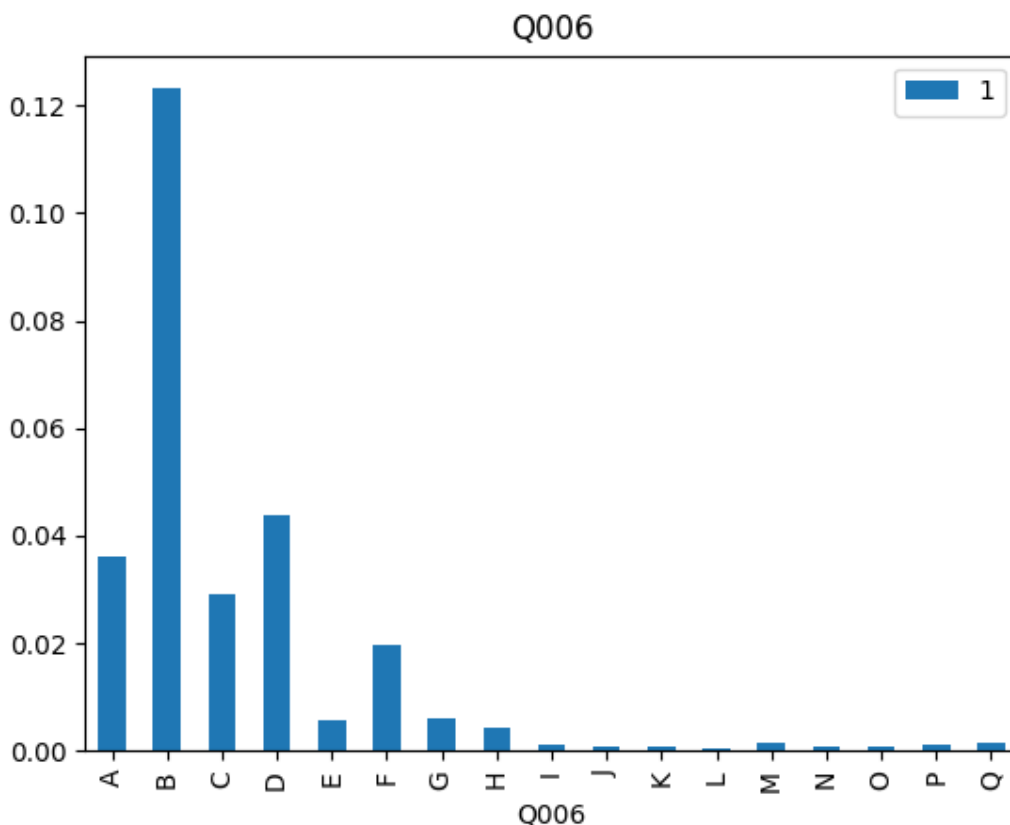
Em relação aos estados da escola do aluno, apresentado no Gráfico 4, alguns estados tiveram um aumento considerável de desistência dentro de sua região, sendo destacados o Ceará (23) e Goiás(52) e a Região Sul(4x) que tiveram um aumento uniforme entre os estados.





**Gráfico 5 - Estado da escola da prova**

Em relação a escola em que o aluno realizou a prova não teve estados com grande aumentos dentro de sua região comparado com o gráfico de anterior mas está claro que a Região Sudeste(3x) teve um aumento considerável de desistências quando comparado com outras regiões, novamente a Região Sul(4x) teve uma distribuição uniforme junto com a Região Centro-Oeste(5x).



**Gráfico 6 - Renda mensal familiar**

Por fim, o maior aumento, com base no Gráfico 6, foi das pessoas com renda mensal de até um salário mínimo.

#### **4 Perguntas de pesquisa e explorações iniciais**

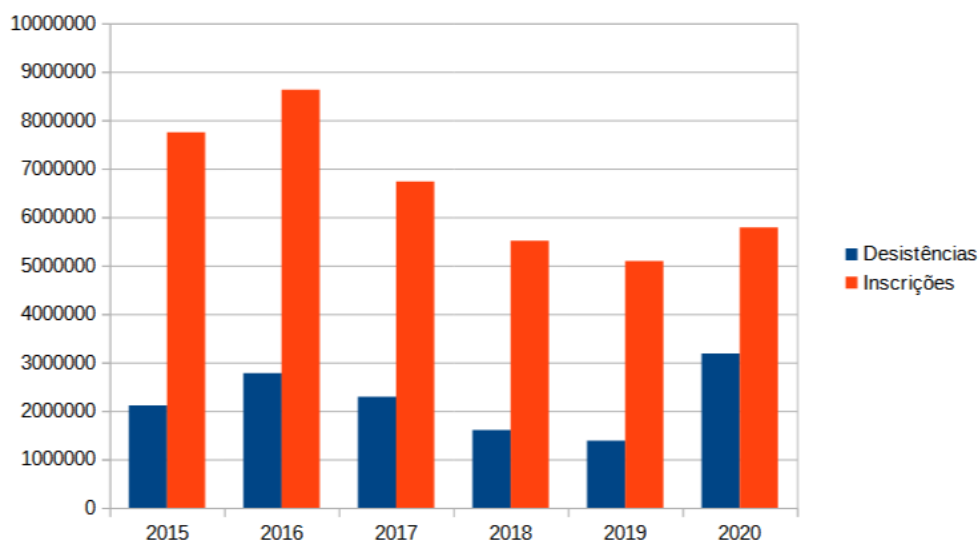
Para facilitar a análise, as perguntas foram divididas em grupos de interesse, sendo eles desistências ao longo dos anos, Questionário Socioeconômico, Produto Interno Bruto (PIB) e COVID.

Com relação às desistências ao longo dos anos, as perguntas iniciais foram:

- Qual o ano que obteve o maior número de desistências?
- Existe uma diferença significativa em relação à aplicação durante a pandemia comparada com os anos anteriores?
- Qual a cidade com o maior número de desistências em 2020?

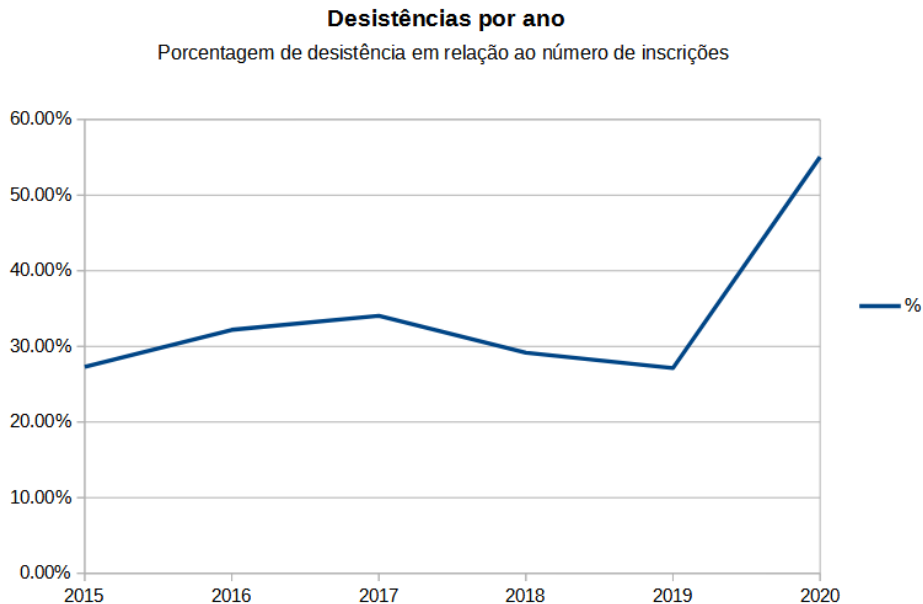
Sobre as questões acima, foi possível apresentar algumas conclusões com base na análise exploratória dos dados do ENEM de 2020, comparado com os dados de anos anteriores à pandemia (2015 até 2019)<sup>4</sup>.

<sup>4</sup> Para a comparação com outros anos, foram utilizados os dados de todas as pessoas participantes e foi considerada desistência somente com relação aos dados do segundo dia de aplicação.



**Gráfico 7 - Inscrições e desistências por ano**

O Gráfico 7 demonstra lado a lado o número absoluto de inscrições dos anos de 2015 até 2020 e o número de desistências de cada ano.



**Gráfico 8 - Desistências por ano**

Enquanto o Gráfico 8 apresenta a comparação entre a porcentagem de desistência em relação ao número de inscrições ao longo dos anos de 2015 até 2020. Dadas as informações ilustradas no gráfico, é notável que houve um aumento na quantidade de estudantes que desistiram da prova no ano de pandemia (2020).

Com relação ao Questionário Socioeconômico, as perguntas iniciais foram:

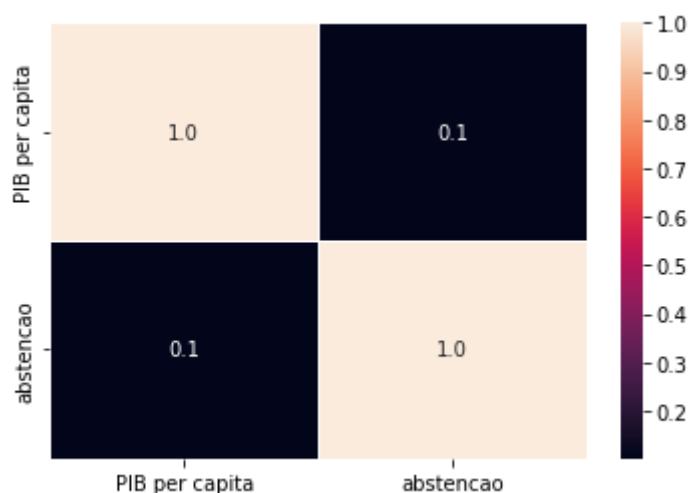


- Como o ambiente doméstico influenciou nas desistências (aspectos gerais)?
- Quais foram as dificuldades do ensino remoto que possivelmente levaram estudantes a desistirem da prova?
- Qual a influência dos sobre acesso à internet, computadores, etc. sobre as desistências?

A análise da correlação apresentada no Gráfico 2 mostra que algumas das informações presentes nos questionários socioeconômicos de 2019 e 2020 não tiveram uma diferença significativa, não tendo tanta influência no aumento das distâncias de 2020.

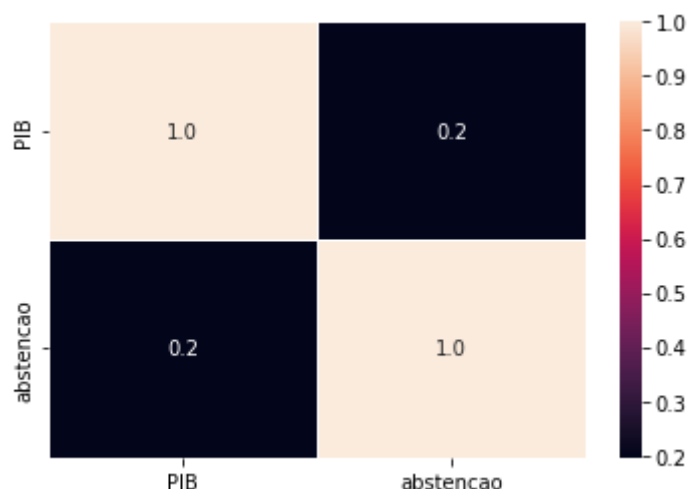
Na próxima etapa, levando em consideração o PIB, a pergunta inicial foi: 'Como a riqueza da cidade/estado influenciou nas desistências?'.

Para auxiliar na resposta da pergunta, foi realizada a análise da correlação entre as variáveis PIB per capita de 2019 e taxa de abstenção do ENEM no ano de 2020, por estado, apresentada no Gráfico 9 e é possível verificar que a correlação é extremamente fraca.



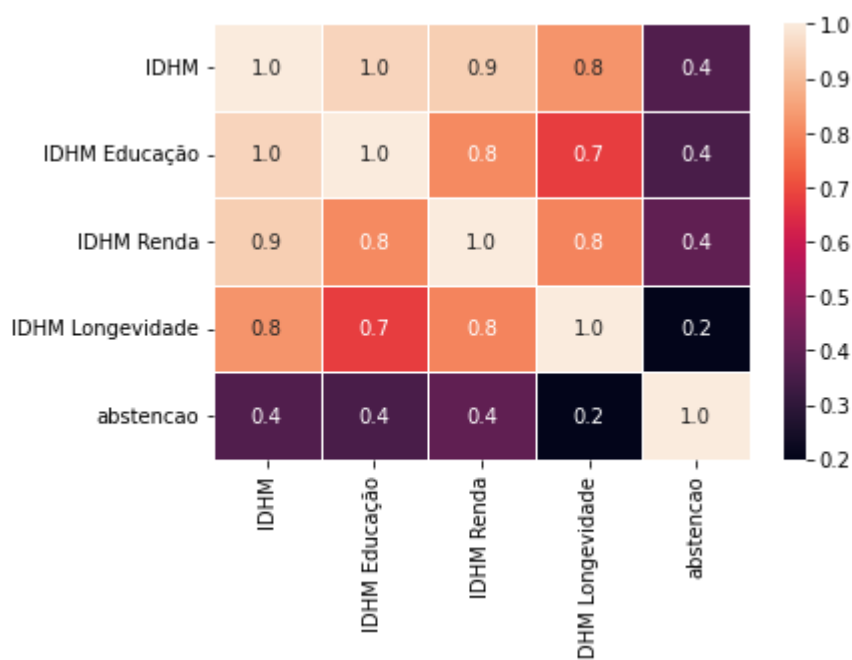
**Gráfico 9 - correlação entre as variáveis PIB per Capita e o número de abstenção por estado**

Nota-se também que, verificando a correlação por cidade, as variáveis possuem uma correlação fraca, conforme apresentado no Gráfico 10.



**Gráfico 10 - correlação entre as variáveis PIB per Capita e o número de abstenção por município**

Com relação ao IDHM, a pergunta inicial foi: 'Como o IDHM influenciou nas desistências?'. Foi possível, através da análise do Gráfico 11, perceber que a correlação entre o número de abstenção por município e o IDHM Longevidade é fraca. Contudo, as variáveis IDHM Educação e Renda possuem correlação moderada, sendo que o mesmo ocorre com o IDHM.



**Gráfico 11 - correlação entre variáveis do IDHM e o número de abstenção por município**

Por fim, para a análise dos aspectos do COVID, foram dispostas as seguintes perguntas iniciais:



- Qual a cidade foi mais afetada pela pandemia?
- A taxa de mortalidade de cada cidade tem relação com o número de desistências?
- O número de contaminação por cidade tem relação com o número de desistências?

	Casos_per_capita	Mortes_per_capita	Relacao_Presenca
Casos_per_capita	1.000000	-0.063433	0.003918
Mortes_per_capita	-0.063433	1.000000	-0.025521
Relacao_Presenca	0.003918	-0.025521	1.000000

**Gráfico 12 - Correlação entre variáveis da pandemia**

Ao analisar a situação da pandemia de cada cidade até o dia da prova, não foi encontrada uma relação significativa entre os casos e mortes acumulados e a relação de presença de alunos durante o ENEM.

## 5 Discussão e próximos passos

De acordo com as análises iniciais apresentadas na Seção 4, a equipe irá buscar selecionar as questões que aparentam ser mais relevantes para aprofundar as discussões sobre qual o aspecto socioeconômico que obteve maior influência.

Também, a equipe vai buscar trabalhar nas dificuldades, visando explorar aspectos não analisados, tais como índices de desigualdade, de desemprego e densidade populacional.

Outro próximo passo é analisar a situação da pandemia durante um determinado período de tempo antes da realização das provas, podendo variar entre alguns dias e algumas semanas. Os datasets utilizados possuem dados de variância diária tanto para novos casos e novos óbitos.