**UNIVERSIDADE FEDERAL DE VIÇOSA**

**CAMPUS FLORESTAL**

Introdução à Ciência dos Dados

# RELATÓRIO DO TRABALHO PRÁTICO

| | |
|---|---|
| Júlia Costa de Faria | 4238 |
| Melissa Araújo | 4244 |
| Paulo Henrique Carneiro SIlva | 4221 |

FLORESTAL

2022

## Sumário

# 1 - Introdução

Este trabalho, conforme informado em sua especificação, visa possibilitar que os estudantes avancem seus conhecimentos dentro da ciência de dados por meio da manipulação de dados obtidos em reais. Na etapa em questão, no caso, a etapa 2, devemos realizar a preparação dos dados e, após isso, realizar a análise exploratória sobre eles. A limpeza ocorre pois é provável que os conjuntos de dados possivelmente possuam ruídos, erros de digitação, valores duplicados, dentre outros, enquanto a análise exploratória é extremamente útil em fornecer uma visão geral dos dados . No contexto da equipe em questão, os processos descritos anteriormente foram aplicados sobre os conjuntos de dados referente à inscritos no FIES obtido em uma plataforma do governo. Cabe salientar que nesta etapa inicial, foram considerados os inscritos no segundo semestre do FIES nos anos de 2019 e de 2020.

# 2 - Organização do Projeto

O projeto em questão utilizou, conforme mencionado previamente, dois conjuntos de dados acerca de inscritos no FIES. Todavia, devido ao tamanho apresentado por tais arquivos, visto que cada um possui, ao menos, 200 mil registros iniciais, optou-se pela criação de um repositório público no GitHub[1] exclusivo para o armazenamento de tais dados após sua limpeza.

Uma grande vantagem de adicionar o arquivo no GitHub, é a possibilidade de referenciá-lo por meio do “wget”, o qual permite o acesso remoto deste, sem a necessidade de baixá-lo e adicioná-lo numa pasta de arquivos, por exemplo.

Além disso, a equipe optou por criar arquivos diferentes para a limpeza de dados de 2019 e de 2020, bem como para a criação de arquivos, de forma a auxiliar na organização do código. Os conjuntos de dados obtidos após a realização da limpeza, conforme mencionado anteriormente, foram armazenados em tal repositório de armazenamento de conjunto de dados no formato CSV.

[1] https://github.com/juliacfaria/arquivosTPDados

# 3 - Limpeza dos Dados

No decorrer do processo de limpeza de dados vários critérios foram utilizados, visto que o conjunto de dados original contava com 57 colunas, onde nem todas são coerentes com a proposta do trabalho da equipe. Além disso, o tratamento de colunas também foi necessário corrigir tipos de dados incorretos, remover linhas incoerentes e duplicadas, dentre outros, conforme descrito a seguir.

## 3.1 - Remoção de colunas

Primeiramente, conforme aprendido em sala de aula, foi realizada uma visão geral sobre os dados, por meio do comando “dtypes” e “info”. A partir destes comandos, foi possível verificar inicialmente que duas das colunas presentes no dataset do ano de 2019 estavam completamente vazias, de modo que não apresentavam utilidade ao projeto, sendo elas “'Município da IES” e “UF da IES”. No conjunto de dados de 2020, por sua vez, embora os campos em questão não estivessem vazios em tal conjunto, por questões de padronização e coerência, estes foram removidos também.

Além disso, analisando o conjunto de dados de 2019, pôde-se observar que grande parte dos valores para os campos “Percentual de financiamento” e “Qtde semestre financiado” apresentaram valores nulos. Com base nisso, como nestas colunas haviam mais de 200 mil valores nulos em um conjunto de dados com 271856 valores, estas também foram desconsideradas da análise. De modo semelhante, tais colunas também foram removidas do conjunto de 2020.

Após isso, a equipe discutiu coluna por coluna das remanescentes nos conjuntos de dados em questão, com o intuito de avaliar quais efetivamente seriam utilizadas no processo. Feito isso, foi possível remover 21 colunas que não são coerentes com a proposta do trabalho, mantendo apenas aquelas que posteriormente possibilitarão a resposta das perguntas elaboradas à princípio.

Por fim, cabe salientar que variáveis que não possuem variação, ou seja, seu “unique” apresenta um único valor, foram removidas da base de dados por não contribuírem para a proposta do projeto.

## 3.2 - Limpeza de valores inteiros

O conjunto de dados identificou alguns tipos de valores como inteiros, os quais foram analisados primeiramente. Conforme era de se esperar, por terem sido reconhecidos como inteiro, a maioria destes valores já estava coerente e funcionando adequadamente.

Entretanto, um campo específico gerou dúvidas até que sua compreensão fosse realizada, sendo este denominado “ID do estudante”. Como o identificador é, por padrão, um valor único, a equipe esperava que o número de IDs fosse equivalente ao número de linhas do conjunto de dados, o que se provou não ser verdade. Na verdade, o que ocorria era ter um mesmo ID sendo referenciado em até mais de 3 linhas.

Inicialmente, a equipe imaginou a possibilidade de um determinado ID se repetir para estados diferentes. Devido a isso, foi exibido um “value_counts” para o par do id do estudante e sua respectiva unidade federativa, o que logo refutou a hipótese gerada, visto que todos os IDs pertenciam ao mesmo estado. Após isso, a equipe imaginou que poderiam ser linhas duplicadas no conjunto de dados em questão, o que motivou a escolha de um usuário específico e comparação de seus campos.

Após realizar tal comparação pode-se verificar que, de fato, havia algumas linhas duplicadas. Todavia, é coerente que um mesmo usuário apareça até 3 vezes diferentes no conjunto de dados em questão, visto que o mesmo pode se inscrever em até 3 opções de curso diferentes. Assim, feita a remoção das duplicatas, este campo estava, então, coerente.

## 3.2 - Limpeza de valores float

Assim como esperado, os campos “float” foram inicialmente identificados pelo conjunto de dados como objetos. Isso ocorre, pois, inicialmente, as suas casas decimais estavam sendo separadas por vírgulas e não por pontos, o que impossibilita o reconhecimento do campo como um valor numérico.

Desta forma, para estes campos primeiramente foi necessária a remoção de valores nulos, seguida da substituição das vírgulas por pontos, o que ocorreu por meio do comando “replace”. Feito isso, já é possível converter tal campo para numérico, por meio do comando “to_numeric” da biblioteca pandas. Com isso, os campos em questão serão reconhecidos como “float” pelo *dataframe* em questão. Exemplos de campos que utilizaram este recurso são aqueles relacionados a renda e notas, com exceção da nota da redação do enem.

## 3.2 - Limpeza de valores object

Nos campos do tipo objeto, grande parte das colunas não apresentaram muitos problemas, visto que muitas possuíam as suas opções bem definidas, fazendo com que o tratamento destas fosse desnecessário. Todavia, alguns campos com um conjunto grande de valores, tais como o “Nome da IES”, por exemplo, possuíam erros de acentuação.

Para a correção deste tipo de erro, a equipe optou também pelo uso do comando “replace” substituindo um a um para os caracteres que estão gerando erros. Além disso, conforme mencionado anteriormente, alguns dos valores de tipo objeto deveriam ser do tipo float, mas este caso especial já foi abordado no tópico anterior.

# 4 - Análise Exploratória dos Dados

A seção da Análise Exploratória é composta pela parte de realização de estatísticas descritivas, tabelas e gráficos voltados para as perguntas realizadas no início do projeto.

## 4.1 - Estatística Descritiva

Levando em consideração as perguntas elaboradas pelo grupo, desenvolveu-se algumas análises por meio de estatística descritiva a fim de obter algum conhecimento acerca dos dados utilizados. Sendo assim, primeiramente observou-se o número de inscritos para cada Unidade Federativa do local de oferta em 2019 e obteve-se que São Paulo é o estado com maior número de inscritos e

Roraima com o menor. No caso, São Paulo há um total de 32896 inscritos através do FIES e Roraima 374, fazendo a diferença entre as valores, obtém-se um resultado de 32522, o que indica uma grande diferença entre esses valores. Outros dados obtidos foram que a média de inscritos por estado foi de 8608,6, a mediana 6737, e os percentis de 25%, 75% e 95% foram 2838,50, 10659 e 25700,50, respectivamente. Observa-se que a distribuição está mais concentrada em alguns estados já que a dispersão é visível, o que pode ser confirmado devido ao desvio padrão ser 8181,5.

Em seguida, analisou-se os dados referente à renda mensal per capita do inscritos do FIES, buscando encontrar uma faixa predominante de valores. Primeiramente, calculou-se a média que equivale a 938,77, a mediana, igual a 760 e o desvio padrão, que resultou em 623,36. Através destes dados percebe-se uma distribuição dispersa, considerando a diferença entre a média e a mediana e o alto desvio padrão. Para observar melhor este fato, realizou-se o cálculo dos quartis, sendo que 25%, 75% e 95% foram 499, 1200 e 2254,57, respectivamente. Observa-se uma grande diferença entre os percentis, porém o valor do quarto percentil não é relativamente muito maior se comparado ao valor máximo de renda, que é 9831,36. Dessa forma, entende-se que há uma discrepância entre os dados porém também há diversos outliers que divergem de uma faixa de renda existente.

Em comparação aos dados de 2020, tem-se resultados similares, ambas as distribuições têm uma dispersão parecida, com a diferença de que os valores de 2020 são inferiores para o número de inscritos por estado, o que é coerente, considerando que 2020 teve menos inscritos no FIES, em geral. Por exemplo, a média de renda é 977,88 , a mediana 786,67 e o desvio padrão 635,38, o que implica em uma distribuição muito parecida com aquela observada nos dados de 2019. Além disso, para a Unidade Federativa do Local de Oferta, a média foi 6756,48, a mediana 5340 e o desvio padrão 6239,28, que são valores menores mas têm a distribuição similar à de 2019.

## 4.2 - Tabelas

Quanto à criação de tabelas, a equipe considerou interessante a utilização deste recurso para observar, de modo exato, a relação entre estado e número de

inscrições, tendo em vista que o tamanho de estados é bem definido. Assim, é possível observar com clareza todos os valores da tabela, o que não seria possível nos demais casos por serem geradas tabelas de tamanho demasiadamente grande. A tabela para os valores em questão no conjunto de dados de 2019 é apresentada a seguir, na Tabela 1, enquanto a sua correspondente para os dados de 2020 é disponibilizada logo na sequência, na Tabela 2.

| Índice | Estado | Número de Inscrições |
|---|---|---|
| 0 | PR | 8637 |
| 1 | PE | 13043 |
| 2 | RS | 7017 |
| 3 | MG | 28753 |
| 4 | SP | 32896 |
| 5 | RJ | 16378 |
| 6 | BA | 18578 |
| 7 | MS | 1666 |
| 8 | ES | 2904 |
| 9 | SC | 2204 |
| 10 | GO | 7121 |
| 11 | PB | 9515 |
| 12 | SE | 3109 |
| 13 | AL | 3866 |
| 14 | DF | 8724 |
| 15 | TO | 2704 |
| 16 | CE | 17465 |
| 17 | AM | 6737 |
| 18 | RN | 3804 |
| 19 | PI | 6718 |
| 20 | PA | 11803 |
| 21 | MA | 9367 |
| 22 | AP | 1946 |
| 23 | RO | 3075 |
| 24 | MT | 2773 |

| | | |
|---|---|---|
| 25 | RR | 374 |
| 26 | AC | 1255 |

**Tabela 1: Distribuição de inscrição por estado, ano 2019.**

| Índice | Estado | Número de inscrições |
|---|---|---|
| 0 | PR | 7794 |
| 1 | PE | 10148 |
| 2 | RS | 4471 |
| 3 | SP | 24534 |
| 4 | MG | 20830 |
| 5 | BA | 15569 |
| 6 | RJ | 15072 |
| 7 | MS | 2037 |
| 8 | ES | 2663 |
| 9 | SC | 1986 |
| 10 | GO | 6499 |
| 11 | PB | 6901 |
| 12 | SE | 2396 |
| 13 | AL | 3200 |
| 14 | DF | 6091 |
| 15 | TO | 2804 |
| 16 | CE | 13639 |
| 17 | AM | 5738 |
| 18 | RN | 2140 |
| 19 | PI | 5340 |
| 20 | PA | 8596 |
| 21 | AP | 1367 |
| 22 | RO | 1898 |
| 23 | MT | 2008 |
| 24 | MA | 7272 |
| 25 | RR | 342 |
| 26 | AC | 1090 |

**Tabela 2: Distribuição de inscrição por estado, ano 2020.**

## 4.3 - Gráficos

Quanto aos gráficos, sabe-se que estes foram gerados visando possibilitar uma melhor visão de parte das perguntas geradas no começo do projeto. Cabe ressaltar que os gráficos foram gerados para cada conjunto de dados, ou seja, gerou-se cada tipo de gráfico a seguir para 2019 e 2020.

O primeiro gráfico gerado visou possibilitar uma melhor visualização acerca do cenário da faixa de renda per capita predominante nos inscritos do FIES. Para a verificação desta, foi criado um gráfico do tipo boxplot, para que fique evidente a distribuição dos valores. Nos dados de 2019, assim como visível na Figura 1, a faixa de renda predominante prevaleceu no intervalo de zero a dois mil, com outliers para valores maiores que dois mil. Já para os dados de 2020, da Figura 2, a faixa de renda predominante se assemelhou aos valores de 2019. Tendo como diferenciação os valores dos outliers obtidos, que no ano de 2020 se mantiveram em grande parte na faixa de 2000 a 4000, enquanto no ano de 2019 chegaram a faixa de 5000.

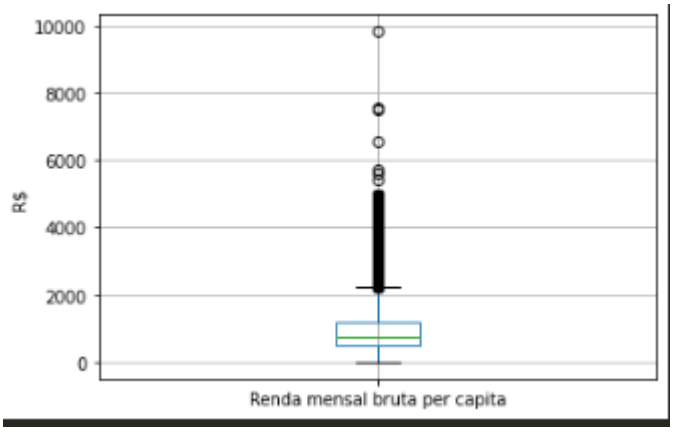


Figura 1 - Gráfico da faixa de renda per capita predominante nos dados de 2019.

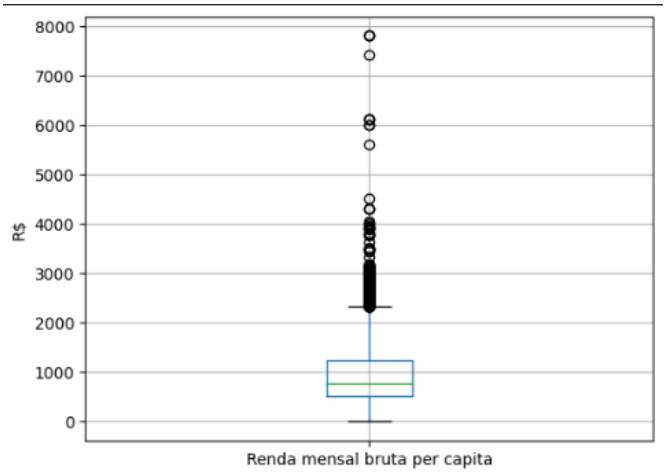


Figura 2 - Gráfico da faixa de renda per capita predominante nos dados de 2020.

Feito isso, buscou-se visualizar a dispersão dos números de inscrição de acordo com os estados. Para isso, foram criados dois tipos de gráficos diferentes. O primeiro deles, foi o boxplot, para que fosse possível analisar com confiança a dispersão dos valores em questão. Posteriormente a isso, foi criado um gráfico de barras para analisar a quantidade de inscritos por estado. No conjunto de dados de 2019, da Figura 3, observou-se que o estado com o maior número de inscritos é São Paulo, enquanto Roraima possui o menor número de inscrições. Já para os dados de 2020, da Figura 4, os estados de maior e menor número de inscrição continuaram os mesmos, mas tendo uma diminuição geral no número de inscrições.

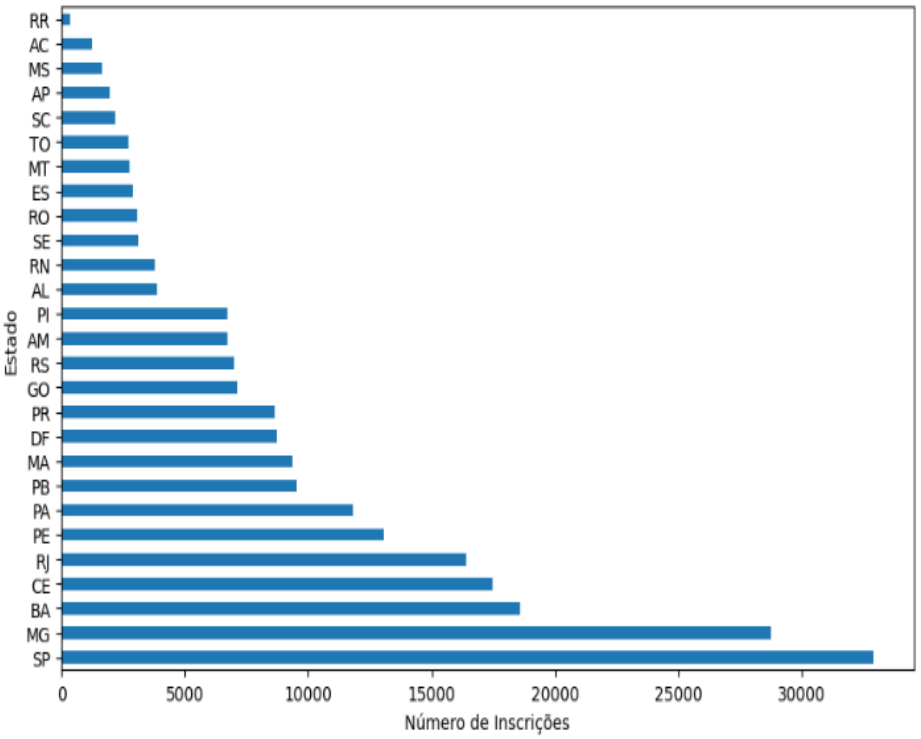


Figura 3 - Gráfico da distribuição de inscrições por estados nos dados de 2019.

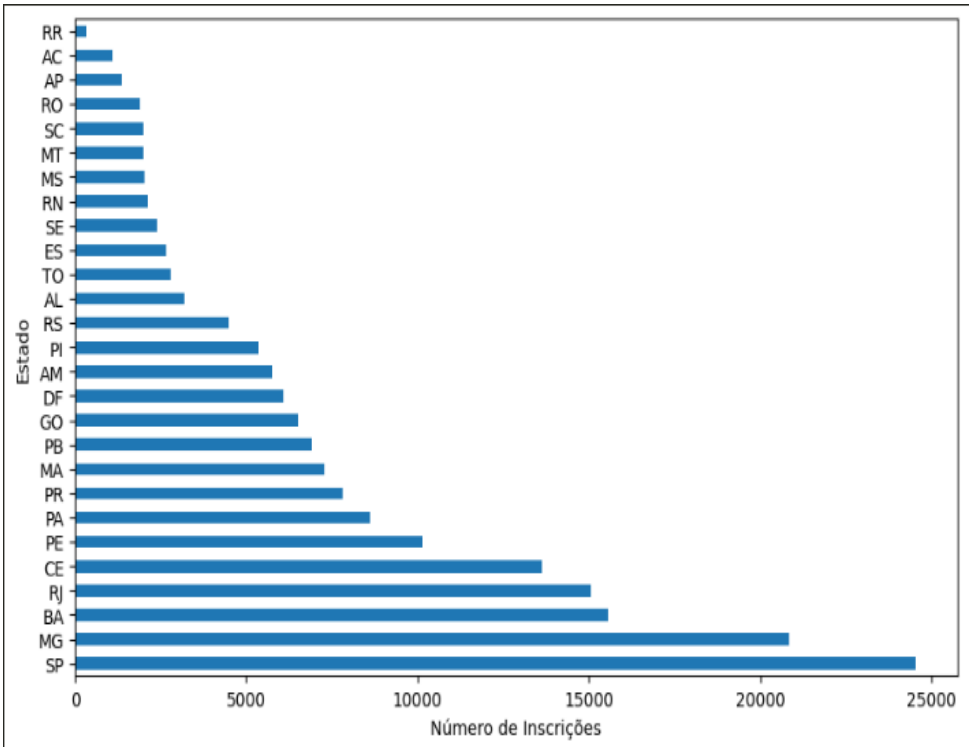


Figura 4 - Gráfico da distribuição de inscrições por estados nos dados de 2020.

Assim, partiu-se para a análise da faixa predominante de espera entre a conclusão do ensino médio e a inscrição na faculdade por meio do FIES. Para facilitar a análise destes dados, criou-se uma nova coluna para armazenar o intervalo buscado em questão e, então, realizar análises sobre elas. Para visualizar os dados em questão, foi criado um boxplot da dispersão dos valores em questão, bem como um histograma em escala logarítmica, conforme visível nas Figuras 5 e 6.

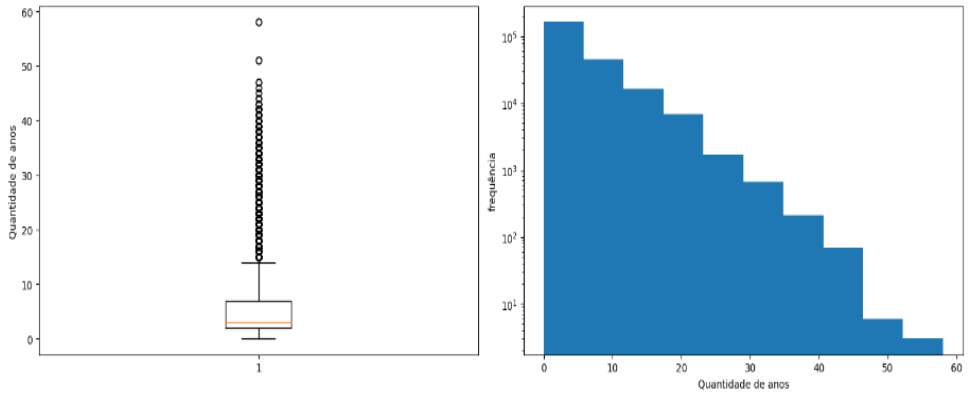


Figura 5 - Gráficos da faixa predominante de espera entre a conclusão do ensino médio e a inscrição na faculdade nos dados de 2019.

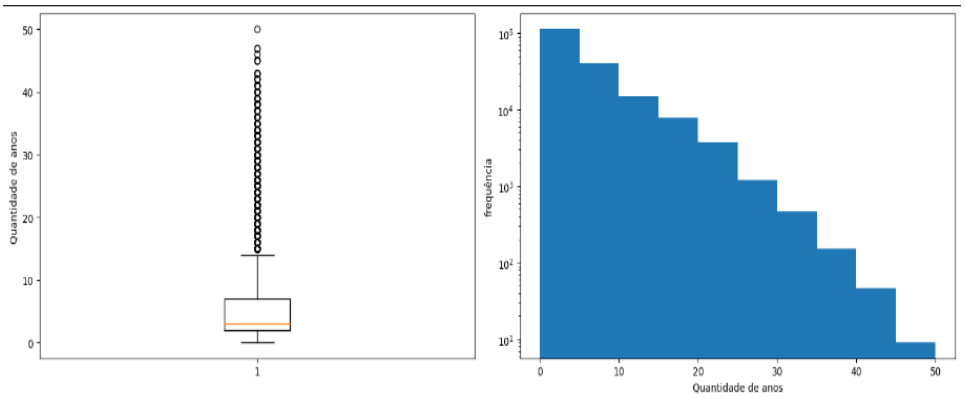


Figura 6 - Gráfico da faixa predominante de espera entre a conclusão do ensino médio e a inscrição na faculdade nos dados de 2020.

Por fim, o último gráfico gerado refere-se à hipótese de que pessoas deficientes estão menos inseridas dentre os aceitos no FIES. Para isso, plotou-se dois gráficos, um em relação à situação da inscrição das pessoas sem deficiência e outro com a situação das pessoas deficientes. A partir dos valores dispostos no

gráfico, não foi possível observar diferenças gritantes, conforme visível nas Figura 7 e 8 para o conjunto de dados de 2019, e 9 e 10 para o conjunto de dados de 2020.

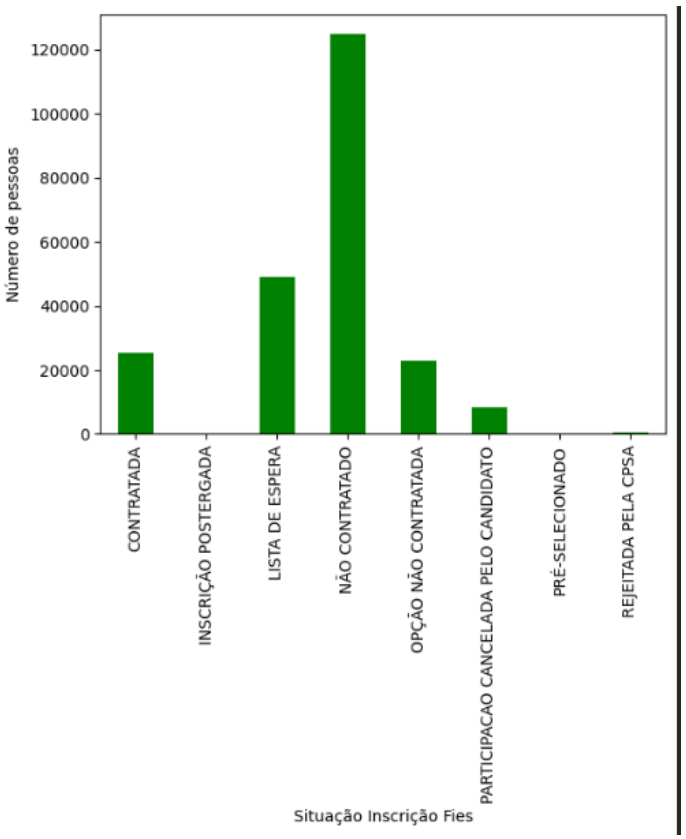


Figura 7 - Gráfico da situação de inscrição para pessoas sem deficiência nos dados de 2019.

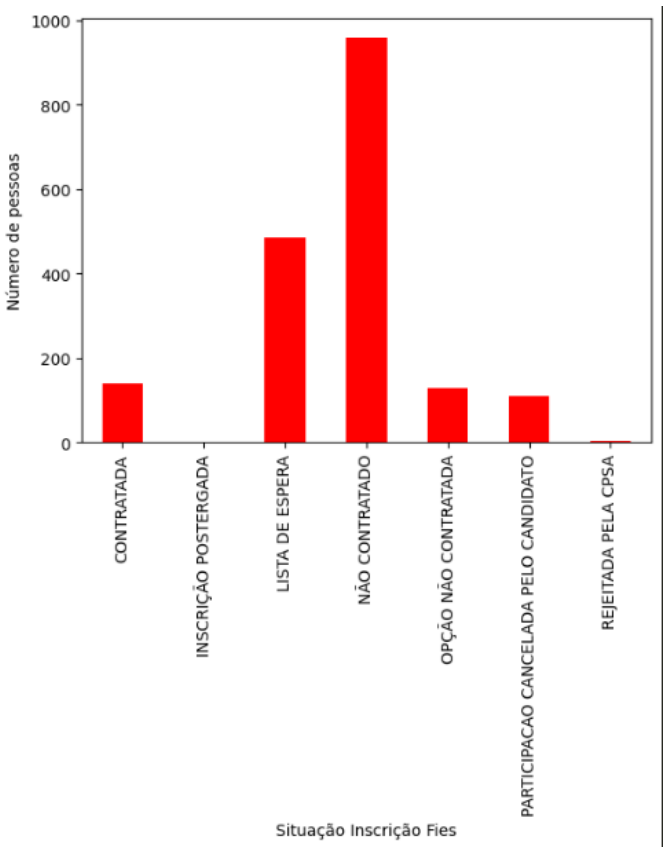


Figura 8 - Gráfico da situação de inscrição para pessoas com deficiência nos dados de 2019.

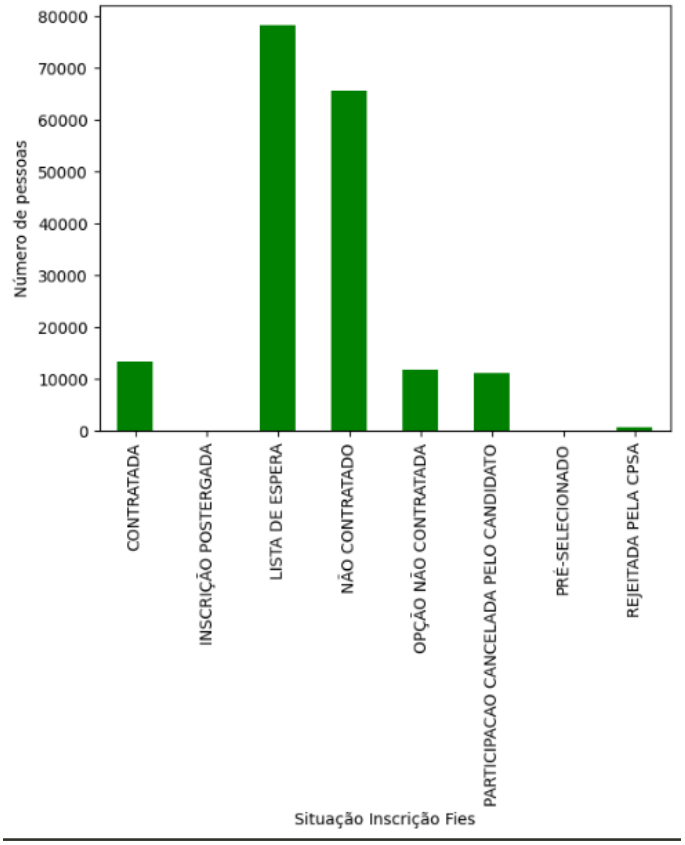


Figura 9 - Gráfico da situação de inscrição para pessoas sem deficiência nos dados de 2020.

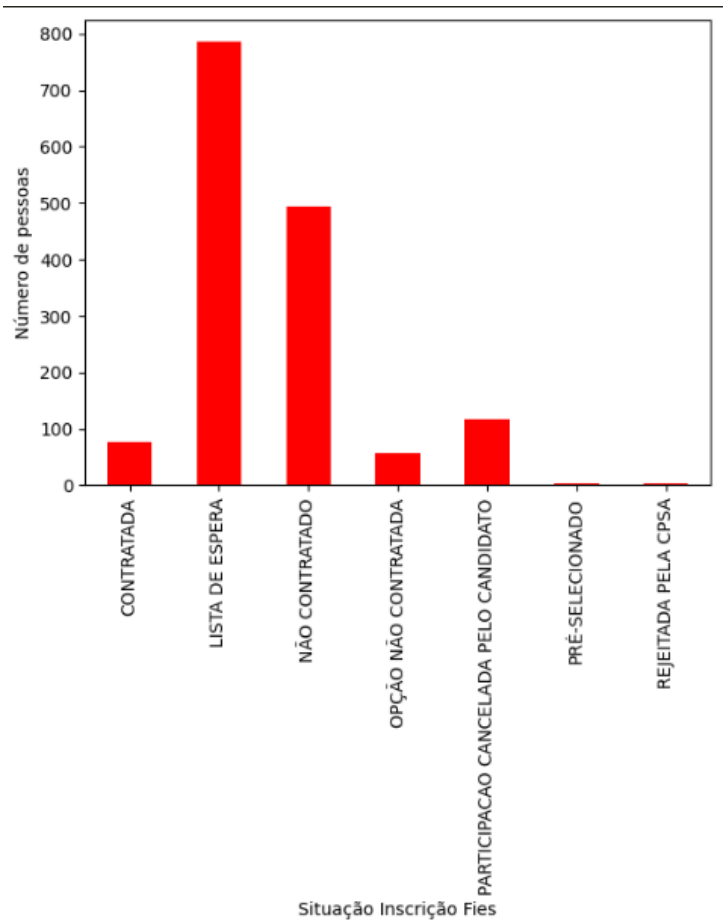


Figura 10 - Gráfico da situação de inscrição para pessoas com deficiência nos dados de 2020.

# 5 - Conclusão

No presente trabalho, foi proposta a limpeza e análise exploratória dos dados utilizados pela equipe no trabalho, no caso, referentes às inscrições no FIES no segundo semestre de 2019 e 2020. Por meio disso, foi possível compreender melhor os fundamentos da realização de tais processos, assim como identificar boas práticas na realização destes.

Além do âmbito do aprendizado, com a finalidade de manter uma boa organização entre o grupo e, consequentemente, um bom desenvolvimento do trabalho, foram utilizadas algumas ferramentas, as quais serão descritas a seguir. Primeiramente, utilizou-se o GitHub para possibilitar o controle de versões do código fonte e a ferramenta de desenvolvimento colaborativo Google Colab, para permitir

uma codificação conjunta nos momentos síncronos de reunião. Além disso, o Google Meet foi a plataforma escolhida para os encontros síncronos do grupo.

Por fim, a partir dos esforços do grupo, foi possível completar a proposta do projeto e alcançar o objetivo principal. O transcorrer do trabalho se deu com complicações que foram resolvidas durante a elaboração do sistema.