1. Introdução

O Fundo de Financiamento Estudantil (FIES) constitui uma das principais políticas públicas de acesso ao ensino superior no Brasil. A disponibilização dos microdados do programa pelo Ministério da Educação (MEC), em conformidade com a Lei de Acesso à Informação (LAI), oferece à comunidade acadêmica e à sociedade civil a oportunidade de auditar, analisar e compreender a dinâmica de distribuição desses recursos.

No entanto, a simples disponibilidade dos dados não garante sua usabilidade. Os arquivos brutos fornecidos governamentalmente apresentam desafios significativos de "Big Data", caracterizados por inconsistências de formatação, ausência de padronização na nomenclatura de arquivos e redundância de registros. Tais barreiras dificultam análises diretas, exigindo um rigoroso processo de pré-processamento.

Nesse contexto, o presente projeto de Iniciação Científica, desenvolvido no curso de Ciência da Computação da UTFPR (Campus Campo Mourão), tem como objetivo principal aplicar técnicas de Ciência de Dados para estruturar, limpar e analisar os microdados do FIES. O foco do estudo recai especificamente sobre os cursos da área de Computação no período de 2019 a 2021, visando traçar um perfil exploratório das ofertas e inscrições, transformando dados brutos em inteligência estratégica.

2. Objetivos

2.1. Objetivo Geral

Realizar uma Análise Exploratória de Dados (AED) sobre o cenário do FIES para cursos de Computação, identificando padrões de oferta, demanda e preenchimento de vagas entre 2019 e 2021.

2.2. Objetivos Específicos

- Implementar um *pipeline* de dados (ETL - *Extract, Transform, Load*) automatizado em Python para a ingestão e tratamento dos microdados do MEC.
- Garantir a reprodutibilidade da pesquisa através de uma arquitetura de software organizada e versionada.
- Sanear a base de dados, removendo duplicatas e corrigindo inconsistências de tipagem e codificação.
- Filtrar e segmentar os dados pertinentes aos cursos de Computação utilizando a classificação oficial (CINE/MEC).

3. Materiais e Métodos

Para atingir os objetivos propostos, a metodologia foi dividida em etapas de configuração de ambiente, engenharia de dados e processamento, conforme descrito a seguir.

3.1. Ferramentas e Tecnologias

O projeto foi desenvolvido utilizando a linguagem de programação **Python**, escolhida devido à robustez de suas bibliotecas para análise de dados. O ambiente de desenvolvimento (IDE) utilizado foi o **VS Code**, integrado com *Jupyter Notebooks* para prototipagem rápida. Para garantir o isolamento das dependências e a replicabilidade do ambiente em diferentes máquinas, utilizou-se o recurso de ambientes virtuais (`venv`), com as bibliotecas listadas no arquivo `requirements.txt`. As principais ferramentas empregadas foram:

- **Pandas:** Para manipulação de *DataFrames*, leitura de arquivos CSV de grande porte e operações vetoriais.
- **OS e Shutil:** Para manipulação programática do sistema de arquivos (criação de diretórios e movimentação de arquivos).
- **Git:** Para versionamento do código e documentação.

3.2. Arquitetura do Repositório e Fluxo de Dados

Visando a integridade e a organização lógica do projeto, adotou-se uma estrutura de diretórios baseada em boas práticas de Engenharia de Dados. O repositório foi segregado em duas grandes vertentes: código (`/analises`) e dados (`/planilhas`).

A estrutura de dados segue um fluxo de maturação da informação, dividido nas seguintes camadas:

1. **Planilhas/Bruto/Fonte:** Armazena os arquivos originais baixados do Portal de Dados Abertos (MEC), mantidos inalterados para garantir a fonte da verdade.
2. **Planilhas/Bruto/Sem_Duplicata:** Armazena os dados após a primeira etapa de higienização técnica.
3. **Planilhas/Limpo e Processado:** Destinadas aos dados que já passaram por normalização de colunas e enriquecimento.

O fluxo de execução dos scripts foi organizado de maneira modular e sequencial (ex: `modulo_1/1_padronizacao`, `modulo_1/2_integridade`), permitindo que o processamento seja auditável passo a passo.

3.3. Ingestão e Pré-processamento Inicial (Pipeline ETL)

A primeira etapa técnica consistiu na construção de um *script* de automação (`pipeline_inicial.py`) para resolver a heterogeneidade dos dados brutos.

A. Padronização de Nomenclatura: Os arquivos originais do MEC não seguem um padrão estrito de nomeação. Foi implementada uma rotina de mapeamento que renomeia os arquivos para o formato `<ano>_<tipo>_<semestre>.csv` e os move para os diretórios hierárquicos adequados, facilitando a leitura iterativa nos processos subsequentes.

B. Leitura e Codificação: Devido à origem dos dados, o *pipeline* foi configurado para tratar especificidades regionais, forçando a leitura com codificação `Latin-1` (para preservar

acentuação da língua portuguesa) e definindo o separador de campos como ponto e vírgula (`;`) e o decimal como vírgula (`,`), prevenindo erros de *parsing*.

C. Deduplicação de Registros: Foi identificada a possibilidade de redundância nos dados brutos. Utilizando a biblioteca Pandas, aplicou-se um algoritmo de remoção de duplicatas (*drop_duplicates*) que elimina linhas inteiramente idênticas. Essa etapa é crítica para evitar distorções estatísticas, garantindo que cada inscrição ou oferta seja contabilizada apenas uma vez.

3.4. Normalização de Atributos e Padronização Semântica (Módulo 1)

Após a deduplicação inicial, observou-se que os metadados (cabeçalhos das colunas) apresentavam inconsistências severas, incluindo espaçamentos irregulares, uso assistemático de acentuação e caracteres especiais, além de nomenclatura não padronizada entre os anos (ex: variações de "Município" e "Munícipio").

Para mitigar esses problemas e preparar os dados para operações relacionais, foi desenvolvido o **Módulo 1: Padronização de Colunas**, dividido em *scripts* específicos para os *datasets* de Inscrições e Ofertas. O processo consistiu em três subetapas automatizadas:

A. Sanitização de Strings via Expressões Regulares Antes da renomeação, aplicou-se uma rotina de limpeza utilizando a biblioteca *Regular Expressions (Regex)*. O algoritmo varre os cabeçalhos originais removendo espaços em branco nas extremidades (*trim*) e substituindo múltiplos espaços internos por um espaço simples. Essa etapa garantiu que variações de digitação na fonte não quebrassem o mapeamento.

B. Mapeamento para *Snake Case* Foi criado um dicionário de dados ("De/Para") convertendo os nomes das colunas originais para o padrão *snake_case* (letras minúsculas separadas por *underscore*), removendo acentos e caracteres especiais.

- *Exemplo Original:* `Renda mensal bruta per capita`
- *Exemplo Padronizado:* `renda_mensal_bruta_per_capita`

C. Desambiguação de Origem (Sufixação) Visando a etapa futura de integração das bases (*Merge*), foi implementada uma regra de sufixação automática. Todas as colunas do *dataset* de inscrições receberam o sufixo `_inscricao` e as de ofertas o sufixo `_ofertas`. Essa estratégia previne a colisão de nomes em atributos comuns (como `uf`, `municipio`, `nome_curso`) e garante a rastreabilidade da informação, permitindo identificar inequivocamente a origem de cada dado no *dataset* final consolidado.

Os arquivos resultantes foram exportados com codificação **UTF-8**, padronizando definitivamente o *encoding* do projeto.

3.5. Definição de Granularidade e Verificação de Integridade (Módulo 1.2)

Após a padronização das colunas, foi necessário estabelecer a granularidade exata dos *datasets*, definindo Chaves Primárias (*Primary Keys* - PK) capazes de identificar univocamente cada registro. A ausência de uma chave única explícita nos dados originais exigiu uma abordagem híbrida de análise semântica dos metadados combinada com validação empírica iterativa.

A. Definição das Chaves Candidatas Através da análise dos dicionários de dados e testes de unicidade, foram estabelecidas as seguintes chaves compostas:

- **Dataset de Inscrições:** Identificou-se que a unidade de análise não é o estudante, mas a *inscrição* em si (um aluno pode ter múltiplas opções). A chave validada foi:
  - `id_estudante_inscricao` + `opcoes_cursos_inscricao_inscricao`
- **Dataset de Ofertas:** A unidade mínima refere-se a uma oferta específica de um curso em um local e turno. A chave composta definida foi:
  - `codigo_e_mec_mantenedora`, `codigo_local_oferta`, `codigo_grupo_preferencia`, `codigo_curso` e `turno`.

B. Algoritmo de Validação de Unicidade Para confirmar essas chaves, foi desenvolvido um algoritmo de integridade (`verificacao_integridade/verificacao_*.ipynb`). O método consiste em comparar a contagem total de linhas do *dataframe* original (Ntotal) com a contagem de linhas após um agrupamento (*groupby*) pelas colunas da chave candidata (Nunique).

A premissa lógica estabelecida foi: Se Ntotal=Nunique, a integridade da entidade está garantida.

C. Resultados e Correção de Inconsistências

C.1. Validação das Inscrições (2019-2021)

Os testes confirmaram a hipótese com 100% de sucesso. Não houve duplicidade de chaves para os milhões de registros processados, conforme detalhado na Tabela 1.

Tabela 1: Resultado da Validação de Integridade - Inscrições

Arquivo (Semestre_Ano)	Total de Linhas (Antes)	Total de Linhas (Depois/PK)	Diferença	Status
fies_1_inscricao_2019	696.731	696.731	0	Íntegro

fies_2_inscricao_2019	271.316	271.316	0	Íntegro
fies_1_inscricao_2020	539.373	539.373	0	Íntegro
fies_2_inscricao_2020	209.502	209.502	0	Íntegro
fies_1_inscricao_2021	251.018	251.018	0	Íntegro
fies_2_inscricao_2021	229.294	229.294	0	Íntegro

C.2. Validação das Ofertas (2019-2021)

A validação apontou uma inconsistência específica no arquivo referente ao primeiro semestre de 2020. O algoritmo detectou uma divergência unitária (Ntotal=24.668 vs Nunique=24.667).

Tabela 2: Resultado da Validação de Integridade - Ofertas

Arquivo (Semestre_Ano)	**Total de Linhas (Antes)**	**Total de Linhas (Depois/PK)**	**Diferença**	**Status / Ação**
fies_1_ofertas_2019	**28.157**	**28.157**	**0**	**Íntegro**
fies_2_ofertas_2019	**22.059**	**22.059**	**0**	**Íntegro**
fies_1_ofertas_2020	**24.668**	**24.667**	**-1**	**Erro (Linha Vazia Removida)**

fies_2_ofertas_2020	**21.956**	**21.956**	**0**	**Íntegro**
fies_1_ofertas_2021	**23.463**	**23.463**	**0**	**Íntegro**
fies_2_ofertas_2021	**23.320**	**23.320**	**0**	**Íntegro**

A investigação automatizada descartou a hipótese de duplicidade de dados válidos. Identificou-se que a divergência era causada por uma **linha contendo apenas valores nulos** (*empty record*), provável artefato de exportação do sistema de origem. O *pipeline* tratou a exceção removendo o registro espúrio e salvando o arquivo corrigido. Após essa intervenção, a integridade da base de ofertas foi plenamente validada.

3.6. Enriquecimento de Dados: Integração com o Censo da Educação Superior (Módulo 1.3)

A classificação dos cursos apenas pela nomenclatura presente no FIES pode apresentar ambiguidades semânticas (ex: variações de nomes para o mesmo curso). Para garantir um filtro robusto e padronizado dos cursos da área de Computação, optou-se por realizar o cruzamento de dados (*data blending*) com os **Microdados do Censo da Educação Superior (INEP**), que fornecem a **classificação oficial via códigos CINE (Classificação Internacional Normalizada da Educação).**

A. Ingestão e Padronização dos Dados Exógenos

Foi estruturada uma rotina de automação (`1 script realizar movimento csv CINE.ipynb`) para gerenciar a coleta e organização dos arquivos de "Cadastro de Cursos" e "Cadastro de IES", compreendendo o período estendido de **2016 a 2024**. A escolha por uma janela temporal mais ampla do que a dos dados do FIES (2019-2021) visa cobrir eventuais defasagens de atualização cadastral nos contratos de financiamento, garantindo que cursos criados em anos anteriores ou posteriores sejam corretamente mapeados.

O algoritmo implementado solucionou a heterogeneidade na estrutura de diretórios disponibilizada pelo INEP, que apresentava variações anuais na nomenclatura das pastas e codificação de caracteres. O *script* varreu as estruturas de pastas aninhadas, extraiu os arquivos CSV pertinentes (`MICRODADOS_CADASTRO_CURSOS` e `IES`), normalizou seus nomes para um padrão único e os centralizou no diretório `planilhas/externo`, eliminando automaticamente diretórios vazios para otimizar a estrutura do projeto. Como resultado, foram indexados e higienizados 15 arquivos de referência externa.

B. Construção do Dataset Mestre de Cursos (Histórico)

Considerando a dinâmica do ensino superior, onde códigos de cursos podem ser alterados ou descontinuados, a utilização do Censo de apenas um ano específico poderia gerar falsos negativos no cruzamento de dados. Para mitigar esse risco, optou-se pela construção de uma **Base Consolidada de Cursos**.

Foi desenvolvido um algoritmo de agregação (`2 criacao df_mestre_cine.ipynb`) que processa sequencialmente os microdados de 2016 a 2024. O processo consistiu em:

1. **Leitura Seletiva:** Importação apenas das colunas essenciais (`NU_ANO_CENSO`, `CO_CURSO`, `NO_CURSO`, `CO_CINE_AREA_GERAL` e `NO_CINE_AREA_GERAL`), otimizando o uso de memória.
2. **Concatenação Vertical:** Unificação dos *dataframes* anuais em uma estrutura única, resultando em um **Dataset Mestre com aproximadamente 3,4 milhões de registros**.

Este *dataset* consolidado atua como um dicionário histórico central, permitindo identificar a área de conhecimento correta através do código do curso (`CO_CURSO`), independentemente do ano de assinatura do contrato FIES.

C. Cruzamento de Dados e Classificação Automática (Data Blending)

Com o *Dataset Mestre* consolidado, procedeu-se ao enriquecimento das bases do FIES (Inscrições e Ofertas). Foi implementada uma lógica de **Deduplicação Temporal** no dataset mestre antes do cruzamento:

1. Os dados históricos (2016-2024) foram ordenados cronologicamente.
2. Aplicou-se um filtro de unicidade sobre a coluna `CO_CURSO`, mantendo apenas a ocorrência mais recente (*keep last*). Essa estratégia assegura que a classificação utilizada reflita o *status* mais atual do curso junto ao MEC.

A integração foi realizada via junção à esquerda (*Left Join*) entre os códigos de curso do FIES e o Mestre. O resultado gerou novos arquivos enriquecidos contendo as colunas `CO_CINE_AREA_GERAL`, permitindo segregar de forma categórica quais registros pertencem à área de Tecnologia da Informação.

D. Validação de Esquema e Consistência Temporal (*Schema Drift Check*)

Dada a extensão temporal dos dados externos (2016-2024), foi imperativo validar se a estrutura dos arquivos CSV (*schema*) se mantinha compatível ao longo da década. Mudanças abruptas nos nomes das colunas ou na tipagem de dados poderiam inviabilizar a construção do *Dataset Mestre*.

Para isso, executou-se uma rotina de verificação (`4 verificando_integridade_colunas_cine.ipynb`) que mapeou e comparou os cabeçalhos de todos os censos. A análise revelou instabilidades críticas na estrutura dos dados governamentais, conforme sumarizado na Tabela 3:

Tabela 3: Evolução da Estrutura dos Microdados do Censo (2016-2024)

Período	Qtd. Colunas	Status do Esquema	Principais Alterações Identificadas
2016-2019	200	Estável	Estrutura base de referência.
2020	200	**Instável**	Alteração na nomenclatura da coluna CO_CINE_ROTULO (identificada como CO_CINE_ROTULO2 ou ausente), gerando incompatibilidade com anos anteriores.
2021-2022	200	Estável	Retorno à estrutura base.
2023	202	Expansão Leve	Inclusão de indicadores institucionais (IN_CONFESSIONAL, IN_COMUNITARIA).
2024	**223**	**Expansão Crítica**	Inclusão massiva de métricas de ações afirmativas (cotas) e novos indicadores demográficos.

Conclusão da Validação: Os resultados demonstraram que a tentativa de importação completa (*bulk loading*) ou a utilização de colunas instáveis (como `CO_CINE_ROTULO`) resultaria em falhas de processamento (*KeyError*) especificamente no ano de 2020. Essa validação empírica justificou a decisão técnica adotada no **Módulo 1.3**: restringir a ingestão apenas às colunas **invariantes** (`CO_CURSO`, `NO_CURSO`, `CO_CINE_AREA_GERAL`). Essa abordagem garantiu a interoperabilidade do sistema independente das alterações anuais realizadas pelo INEP.

Peço desculpas pela confusão. Você tem toda razão: você enviou **apenas** os códigos da subpasta `4 agrupar_dataset` (`agrupar_inscritos.py` e `agrupar_ofertas.py`) e apenas **mencionou** o nome da subpasta 5.

Vamos focar estritamente no que esses scripts da pasta 4 realizam: a **Consolidação dos Arquivos**.

Aqui está o texto técnico para o relatório descrevendo essa etapa, baseado exatamente na lógica de `pd.concat` e `sort_values` que você implementou:

3.8. Consolidação Longitudinal dos Datasets (Módulo 1.4 - Agrupamento)

Após o enriquecimento individual de cada arquivo semestral com os códigos CINE, fez-se necessária a unificação das bases para permitir uma análise longitudinal completa do período (2019-2021).

Foi desenvolvido o módulo de agrupamento (`4 agrupar_dataset`), composto pelos scripts `agrupar_inscritos.py` e `agrupar_ofertas.py`. A rotina automatizada executa as seguintes operações:

1. **Leitura em Lote:** O algoritmo varre o diretório de arquivos processados (`.../coluna_CINE`), carregando todos os arquivos CSV semestrais disponíveis.
2. **Concatenação Vertical:** Utilizando a função `pd.concat`, os *dataframes* semestrais são empilhados em uma única estrutura tabular (*row-wise concatenation*), consolidando milhões de registros de inscrições e ofertas em um único objeto de memória.
3. **Ordenação Cronológica:** Para garantir a linearidade temporal dos dados, aplicou-se uma reordenação (*sort_values*) baseada nas colunas de tempo:
    - Inscrições: Ordenadas por `ano_processo_seletivo_inscricao` e `semestre_processo_seletivo_inscricao`.
    - Ofertas: Ordenadas por `ano_ofertas` e `semestre_ofertas`.
4. **Persistência Unificada:** Os resultados foram exportados para a pasta `agrupado/`, gerando os arquivos mestres `inscritos_agrupado.csv` e `ofertas_agrupado.csv`.

Esses dois arquivos consolidados representam o produto final da etapa de Engenharia de Dados e servem como insumo definitivo para as análises estatísticas e filtragens de área.

3.9. Auditoria de Classificação e Distribuição por Áreas CINE (Módulo 1.5)

Para encerrar a etapa de preparação e higienização dos dados, realizou-se uma auditoria profunda (`5 verificacao_nan_CINE/nan_cine_*.ipynb`) para validar a eficácia do cruzamento com a base CINE e analisar a distribuição volumétrica de registros por área do conhecimento em ambos os *datasets* (Inscrições e Ofertas).

A. Eficácia do Enriquecimento (Taxa de Match)

A utilização da base histórica mestre (2016-2024) permitiu classificar a vasta maioria dos registros. As Tabelas 4 e 5 resumem o sucesso do *pipeline*, apresentando índices de preenchimento que garantem a robustez estatística da amostra.

Tabela 4: Resumo de Qualidade do Enriquecimento - Inscrições

Período	Total de Inscrições	Cursos Não Encontrados (NaN)	Taxa de Perda (%)	Status
2019.1	696.731	14.945	2,15%	Aceitável
2019.2	271.316	2.010	0,74%	Excelente
2020.1	539.373	4.595	0,85%	Excelente
2020.2	209.502	1.452	0,69%	Excelente
2021.1	251.018	1.339	0,53%	Excelente
2021.2	**229.294**	**1.020**	**0,44%**	**Excelente**

Tabela 5: Resumo de Qualidade do Enriquecimento - Ofertas

Período	Total de Ofertas	Cursos Não Encontrados (NaN)	Taxa de Perda (%)	Status

2019.1	28.157	2.436	8,65%	Aceitável
2019.2	22.059	835	3,79%	Aceitável
2020.1	24.667	887	3,60%	Aceitável
2020.2	21.956	627	2,86%	Aceitável
2021.1	23.463	666	2,84%	Aceitável
2021.2	23.320	668	2,86%	Aceitável

B. Distribuição por Grande Área do Conhecimento - Inscrições

Abaixo, detalha-se a composição das inscrições por área CINE. A área de interesse deste projeto (**Computação e TIC**) mantém-se consistentemente entre as áreas com volume relevante de candidatos.

Área Geral CINE	**2019.1**	**2019.2**	**2020.1**	**2020.2**	**2021.1**	**2021.2**
Saúde e bem-estar	287.817	122.825	250.079	110.763	139.665	133.054
Negócios, adm. e direito	186.725	72.308	131.619	47.002	51.290	44.311
Engenharia e construção	70.080	24.932	44.173	13.832	15.020	12.927
Ciências sociais e info.	47.449	19.077	41.805	15.206	18.038	16.086

Educação	27.659	7.479	16.921	4.399	5.098	3.535
Agricultura e veterinária	23.447	9.446	21.071	8.254	10.057	9.829
Computação e TIC	**18.650**	**6.479**	**14.560**	**4.006**	**5.192**	**4.526**
Serviços	10.610	3.891	8.052	2.567	2.727	1.968
Artes e humanidades	6.682	2.151	4.676	1.542	1.944	1.538
Ciências naturais e mat.	2.667	718	1.822	479	648	500
Não classificado (NaN)	*14.945*	*2.010*	*4.595*	*1.452*	*1.339*	*1.020*

C. Distribuição por Grande Área do Conhecimento - Ofertas

Diferente das inscrições, a distribuição de **Ofertas** apresenta a área de Negócios como a de maior volume de vagas ofertadas no período.

Tabela 7: Ofertas por Área Geral CINE (2019-2021)

Área Geral CINE	**2019.1**	**2019.2**	**2020.1**	**2020.2**	**2021.1**	**2021.2**
Negócios, adm. e direito	7.415	6.192	6.730	6.285	6.679	6.322
Saúde e bem-estar	6.006	5.253	5.972	5.627	5.942	6.100

Engenharia e construção	4.580	3.858	4.040	3.335	3.463	3.498
Educação	2.296	1.611	1.821	1.362	1.525	1.578
Ciências sociais e info.	1.457	1.282	1.561	1.479	1.655	1.620
Computação e TIC	**1.540**	**1.133**	**1.331**	**1.232**	**1.331**	**1.301**
Serviços	948	697	859	720	732	746
Artes e humanidades	746	599	676	595	643	699
Agricultura e veterinária	546	461	618	569	671	650
Ciências naturais e mat.	187	138	172	125	156	138
Não classificado (NaN)	*2.436*	*835*	*887*	*627*	*666*	*668*

D. Conclusão da Etapa de Preparação

A análise volumétrica e de completude valida o sucesso do *pipeline* de dados construído. Embora a área de Saúde domine as inscrições e Negócios domine as ofertas, a área de **Computação e TIC** mantém um volume robusto de dados, acumulando mais de 53 mil inscrições e aproximadamente 8 mil ofertas no período total.

Com a validação técnica das Chaves Primárias, a normalização dos atributos e a classificação oficial por área CINE com menos de 1% de erro nos dados mais recentes, a base de dados consolidada e enriquecida está pronta para a aplicação dos filtros específicos e o início da Análise Exploratória focada em Ciência da Computação.

4. Tratamento de Lacunas e Refinamento dos Dados (Módulo 2)

Após a consolidação dos dados no Módulo 1, iniciou-se a fase de refinamento para mitigar as lacunas residuais (NaNs) identificadas na auditoria de cruzamento. Esta etapa foi fundamental para garantir que nenhum curso da área de computação fosse excluído por falhas pontuais de indexação no Censo ou por divergências de nomenclatura entre as bases governamentais.

4.1. Criação de Tabela Auxiliar de Imputação (Módulo 2.1)

A auditoria revelou que, embora alguns registros estivessem sem classificação CINE, o nome do curso estava presente em formato textual. Como a vasta maioria dos registros com o mesmo nome de curso foi classificada corretamente via código, desenvolveu-se uma estratégia de **Imputação Baseada em Conhecimento Prévio**.

O primeiro passo foi a criação de um "Dicionário de Áreas de Conhecimento" (`1_criar_tabela_auxiliar.ipynb`). O algoritmo executou os seguintes procedimentos técnicos:

- **Filtragem de Confiança**: O script processou o arquivo `inscritos_agrupado.csv`, selecionando **2.171.873 linhas válidas** que possuíam classificação CINE (removendo os NaNs via `dropna`).
- **Mapeamento Único**: Utilizando a função `drop_duplicates(subset=['nome_curso_inscricao'])`, o sistema gerou um par único para cada curso.
- **Geração do Artefato**: Foram validados **362 mapeamentos únicos** salvos no arquivo `dataset_auxiliar_curso_area.csv`.

Tabela 8: Amostra do Mapeamento Auxiliar para Imputação CINE

ID	Nome do Curso (Chave)	Nome da Área CINE	Código CINE
0	CIÊNCIAS BIOLÓGICAS	Ciências naturais, matemática e estatística	5.0

2	CIÊNCIA DA COMPUTAÇÃO	Computação e TIC	6.0
7	ENGENHARIA DA COMPUTAÇÃO	Engenharia, produção e construção	7.0
14	DIREITO	Negócios, administração e direito	4.0
23	PEDAGOGIA	Educação	1.0
67	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	Computação e TIC	6.0

Análise Técnica: A criação desta tabela auxiliou na correção de inconsistências semânticas e de *encoding* (ex: erros identificados em "CIÊNCIAS CONTÁBEIS"). Além disso, reforçou a distinção necessária para a pesquisa: enquanto "Ciência da Computação" é classificada na Área 6 (TIC), o curso de "Engenharia da Computação" pertence oficialmente à Área 7 (Engenharia), conforme a norma oficial do INEP.

4.2. Diagnóstico de Falhas e Cura de Inscrições (Módulo 2.3)

Para entender a origem das lacunas, executou-se o script `2_arquivos_com_nan.py` para isolar os registros "órfãos". Identificou-se que a maioria dos erros ocorria por variações mínimas de digitação ou caracteres especiais corrompidos.

Com o diagnóstico, o script `3_inscritos_ajuste_agrupado.ipynb` aplicou um dicionário de remediação via método `.apply()` para realizar a imputação em massa. A eficácia deste processo é detalhada na Tabela 9:

Tabela 9: Eficácia da Recuperação de Dados - Dataset de Inscrições

Métrica	**Valor Absoluto**

Total de NaNs Identificados (Antes)	25.361
Total de NaNs Corrigidos Algoritmicamente	**25.361**
Total de NaNs Remanescentes	0
Taxa de Recuperação Final	**100,00%**

Conclusão do Refinamento de Inscrições: O dataset atingiu **plena completude**, permitindo que o arquivo `inscritos_agrupado_limpo_CORRIGIDO.csv` se tornasse a base definitiva para o estudo do perfil de Computação.

4.3. Recuperação de Classificação no Dataset de Ofertas (Módulo 2.4)

A etapa de *Data Healing* foi estendida para o dataset de Ofertas seguindo o mesmo rigor metodológico. Através do script `1_criar_tabela_auxiliar.ipynb` para ofertas, foram identificadas **137.503 linhas válidas**, gerando **377 mapeamentos únicos**.

O script `3_ofertas_ajuste_agrupado.ipynb` implementou um mapa de correção específico para as nomenclaturas de ofertas, visando garantir que o cálculo da taxa de ocupação de vagas não fosse comprometido. Os resultados da cura de ofertas são apresentados na Tabela 10:

Tabela 10: Eficácia da Recuperação de Dados - Dataset de Ofertas

Métrica	**Valor Absoluto**
Total de NaNs Identificados (Antes)	6.119
Total de NaNs Corrigidos Algoritmicamente	**6.114**

Total de NaNs Remanescentes	5
Taxa de Recuperação Final	**99,91%**

Análise dos Resultados de Ofertas: Restaram apenas **5 registros** sem classificação (0,09%), referentes a nomes ambíguos marcados para verificação manual. O arquivo final, `ofertas_agrupado_limpo_CORRIGIDO.csv`, constitui agora uma base fidedigna para as análises de correlação entre oferta e demanda.

4.4. Auditoria Final de Preenchimento (Módulo 2 - Subpasta 3)

Para garantir que os procedimentos de cura de dados atingiram a eficácia esperada, executou-se uma rotina de verificação final localizada na subpasta `3 verificacao_nan_CINE`. Esta etapa valida os arquivos corrigidos gerados nas etapas anteriores, assegurando que a base de dados consolidada não possua lacunas que enviesem a análise da área de Computação.

A. Validação de Completude - Inscrições

O *script* `nan_cine_inscritos.ipynb` processou o arquivo final consolidado, realizando uma contagem exaustiva de todas as áreas gerais CINE presentes. Os resultados confirmaram a eliminação total das inconsistências:

Tabela 11: Distribuição Final por Área CINE (Inscrições 2019-2021)

Área Geral CINE	**Total de Inscrições**	**Percentual (%)**
Saúde e bem-estar	1.049.914	47,78%
Negócios, administração e direito	538.713	24,51%

Engenharia, produção e construção	186.285	8,47%
Ciências sociais, comunicação e informação	158.921	7,23%
Agricultura, silvicultura, pesca e veterinária	82.811	3,76%
Educação	68.044	3,09%
Computação e TIC (TIC)	**54.803**	**2,49%**
Serviços	31.499	1,43%
Artes e humanidades	19.009	0,86%
Ciências naturais, matemática e estatística	7.235	0,32%
Não classificado (NaN)	**0**	**0,00%**
TOTAL	**2.197.234**	**100,00%**

Resultado Crítico: A base de inscrições atingiu **0,00% de NaNs**, validando o sucesso total do algoritmo de imputação aplicado no Módulo 2.1.

B. Validação de Completude - Ofertas

De forma análoga, o *script* `nan_cine_ofertas.ipynb` auditou o dataset de vagas ofertadas. A distribuição final revela uma base altamente saneada:

Tabela 12: Distribuição Final por Área CINE (Ofertas 2019-2021)

Área Geral CINE	**Total de Ofertas**	**Percentual (%)**
Negócios, administração e direito	41.022	28,56%
Saúde e bem-estar	35.923	25,01%
Engenharia, produção e construção	24.154	16,81%
Educação	10.724	7,46%
Ciências sociais, comunicação e informação	9.342	6,50%
Computação e TIC (TIC)	**8.347**	**5,81%**
Serviços	5.169	3,59%
Artes e humanidades	4.270	2,97%
Agricultura, silvicultura, pesca e veterinária	3.630	2,52%
Ciências naturais, matemática e estatística	1.036	0,72%
Não classificado (NaN)	**5**	**0,00%**

TOTAL	**143.622**	**100,00%**

> **Resultado Crítico**: Restaram apenas **5 registros** residuais sem classificação em um universo de mais de 143 mil linhas, o que representa uma perda estatística irrelevante para o escopo da pesquisa.

C. Conclusão do Módulo de Refinamento

Com a conclusão da auditoria na **subpasta 3**, os dados foram declarados prontos para a análise exploratória. O projeto dispõe agora de um universo de **54.803 inscrições** e **8.347 ofertas** devidamente identificadas na área de **Computação e TIC**, permitindo o início da extração de perfis socioeconômicos e desempenho acadêmico.

4.5. Segmentação por Perfil de Renda e Modalidades FIES (Módulo 2 - Subpasta 4)

Para que a análise exploratória reflita a realidade das políticas públicas, foi necessário classificar cada um dos **2.197.234 inscritos** conforme as modalidades oficiais do programa FIES, que variam de acordo com a renda familiar *per capita* e a região geográfica de residência.

A. Critérios de Classificação e Regras de Negócio

Desenvolveu-se um algoritmo de filtragem (`1 inscritos_peneirar_renda_regiao.ipynb`) fundamentado nos valores nominais do Salário Mínimo (SM) vigentes no período do estudo:

- **2019:** R$998,00 | **2020:** R$1.045,00 | **2021:** R$1.100,00.

As regras para a criação da coluna `modalidade_fies` foram implementadas seguindo a lógica apresentada na Tabela 13:

Tabela 13: Regras para Classificação de Modalidades FIES

Modalidade	Critério de Renda (per capita)	Abrangência Regional

Modalidade I	≤3 Salários Mínimos	Nacional
Modalidade II	>3 e ≤5 Salários Mínimos	Norte, Nordeste e Centro-Oeste
Modalidade III (P-FIES)	>3 e ≤5 Salários Mínimos	Sul e Sudeste
Eliminado	>5 Salários Mínimos (ou Renda Nula)	Nacional

B. Implementação do Filtro Algorítmico

A rotina técnica seguiu três etapas de processamento:

1. **Normalização Monetária:** Conversão da coluna de renda de *string* para *numeric*, tratando o padrão brasileiro de vírgulas como separadores decimais.
2. **Mapeamento Geoestatístico:** Criação de uma coluna temporária de região baseada na Unidade Federativa (`uf_residencia_inscricao`) do candidato.
3. **Execução Vetorial:** Utilização da função `np.select` para aplicar as condições de renda e região de forma simultânea em toda a base de dados, garantindo alta performance no processamento de milhões de linhas.

C. Resultados da Distribuição de Candidatos

A aplicação do filtro permitiu identificar o perfil econômico predominante dos candidatos ao financiamento. Os resultados consolidados são apresentados na Tabela 14:

Tabela 14: Distribuição de Candidatos por Modalidade e Ano

Ano	Modalidade I	Modalidade II	Mod. III (P-FIES)	Eliminado
2019	959.662	4.615	3.737	33

2020	744.995	1.993	1.870	17
2021	480.141	76	77	18
TOTAL	**2.184.798**	**6.684**	**5.684**	**68**

Análise Técnica: Observa-se que a **Modalidade I** concentra a esmagadora maioria dos inscritos (mais de 99%), reafirmando o caráter social do programa FIES no atendimento às faixas de menor renda. A baixa contagem de "eliminados" (apenas 68 casos em todo o período) demonstra que o público que acessa o sistema de inscrições já possui um perfil socioeconômico pré-alinhado aos limites do programa.

4.6. Validação Cruzada de Modalidades e Refinamento P-FIES (Módulo 2 - Subpasta 4)

Após a segmentação inicial por renda, foi necessário realizar uma validação lógica para garantir a consistência entre a intenção do candidato e a disponibilidade da modalidade na instituição de ensino. O foco desta etapa foi a Modalidade III (P-FIES), que depende da adesão específica da mantenedora ao modelo de financiamento por bancos privados.

A. Lógica de Cruzamento Relacional

Desenvolveu-se uma rotina de integração (`ajustado_fies_pfies_inscritos.csv`) que realizou um *Left Join* entre o dataset de inscrições e o dataset de ofertas, utilizando uma chave composta de sete atributos: ano, semestre, código da mantenedora, local de oferta, grupo de preferência, código do curso e turno.

O objetivo foi verificar o campo `participa_p_fies_ofertas` para cada inscrição classificada como Modalidade III.

B. Regra de Reclassificação e Eliminação

Identificou-se que candidatos poderiam pleitear a Modalidade III em cursos cujas ofertas não previam tal categoria. Para corrigir essa inconsistência, aplicou-se a seguinte regra algorítmica:

- **Condição:** Se o candidato foi classificado como `Modalidade III (P-FIES)`, mas o campo `participa_p_fies_ofertas` na base de ofertas for `NAO` ou nulo.
- **Ação:** O registro é reclassificado como `eliminado`.

Essa etapa assegura que a análise estatística subsequente não inclua candidatos em modalidades de financiamento que não seriam passíveis de concretização jurídica no ato da contratação.

4.7. Auditoria Final de Integridade de Chaves (Módulo 2 - Subpasta 5)

Para encerrar o Módulo 2, procedeu-se a uma auditoria técnica de integridade estrutural em ambos os datasets finais (`verificacao_inscricao.ipynb` e `verificacao_ofertas.ipynb`). O objetivo foi confirmar que, após todos os processos de merge, enriquecimento CINE e filtragem de renda, a unicidade dos registros e a estrutura das colunas permaneciam íntegras.

A. Integridade das Entidades e Unicidade

A validação consistiu em confrontar a contagem total de linhas com a contagem de registros únicos baseada nas Chaves Primárias (PK) definidas no início do projeto. Os resultados confirmaram a ausência de duplicidade e a preservação do volume de dados, conforme detalhado na Tabela 15:

Tabela 15: Resumo Final de Integridade Estrutural

Dataset	Total de Registros	Chave Primária (PK) Validada	Status
Inscrições	2.197.234	id_estudante_inscricao + opcoes_cursos_inscricao_inscricao	Íntegro
Ofertas	143.622	codigo_e_mec_mantenedora + codigo_local + cod_grupo_pref + cod_curso + turno	Íntegro

B. Inventário Final de Atributos

A base de dados consolidada para análise exploratória apresenta um enriquecimento significativo em relação aos microdados brutos. O inventário final de colunas totaliza:

- **Dataset de Inscrições:** 65 colunas, incluindo os novos campos `NO_CINE_AREA_GERAL`, `regiao_residencia` e `modalidade_fies`.
- **Dataset de Ofertas:** 69 colunas, consolidando dados financeiros e indicadores de vaga por semestre.

C. Conclusão do Processamento de Dados

Com a integridade de chaves validada (Ntotal=Nunique) e a recuperação total de NaNs concluída, os datasets foram exportados com codificação UTF-8, servindo como o repositório mestre de alta fidelidade para o **Módulo 3: Análise Exploratória da Área de Computação**.

5. Padronização Final e Otimização de Atributos (Módulo 3)

Com a base de dados tecnicamente curada e sem lacunas residuais, iniciou-se o **Módulo 3**, focado na usabilidade analítica dos dados. O objetivo desta etapa foi simplificar a nomenclatura dos atributos e enriquecer o *dataset* de ofertas com recortes geográficos, reduzindo a verbosidade do código nas fases de visualização.

5.1. Renomeação Estratégica e Simplificação (Módulo 3.1)

Embora a sufixação aplicada no Módulo 1 tenha sido essencial para evitar colisões durante os *merges*, a permanência de nomes excessivamente longos (ex: `renda_mensal_bruta_per_capita_inscricao`) tornava o processamento estatístico ineficiente.

Desenvolveu-se uma rotina de simplificação (`inscritos.ipynb` e `ofertas.ipynb`) baseada em um dicionário de mapeamento ("De/Para"), seguindo os seguintes critérios:

1. **Remoção de Redundância:** Eliminação dos sufixos `_inscricao` e `_ofertas`, uma vez que os arquivos já estão persistidos em diretórios independentes.
2. **Consatinação de Termos:** Adoção de termos curtos e intuitivos (ex: `ano_processo_seletivo_inscricao` → `ano`; `media_nota_enem_inscricao` → `media_enem`).
3. **Unificação de Colunas CINE:** Padronização dos campos de área de conhecimento para `nome_cine_area_geral` e `codigo_cine_area_geral` em ambas as bases.

5.2. Enriquecimento Geográfico de Ofertas

No processamento do *dataset* de ofertas (`ofertas.ipynb`), implementou-se uma etapa adicional de **Regionalização**. Utilizando um mapa de correspondência entre as Unidades Federativas (`uf_ies`) e as Grandes Regiões do Brasil, criou-se a coluna `regiao_ies`.

- **Validação de Cobertura:** O algoritmo confirmou que **100% das ofertas** possuíam UFs válidas, permitindo uma classificação regional completa sem a geração de novos *NaNs*.
- **Finalidade Analítica:** Esta coluna permite identificar assimetrias geográficas na distribuição de vagas de Computação financiavel pelo FIES.

5.3. Inventário Técnico de Colunas Padronizadas

Após a execução dos *scripts*, as bases de dados foram salvas como `inscritos_limpo.csv` e `ofertas_limpo.csv`, consolidando o seguinte inventário:

- **Dataset de Inscrições (65 colunas):** Inclui o perfil socioeconômico completo, indicadores de renda recalculados (`renda_per_capita`, `modalidade_fies`) e o desempenho acadêmico (notas Enem).
- **Dataset de Ofertas (70 colunas):** Consolida dados das IES, localização geográfica (`regiao_ies`), indicadores de ocupação de vagas (`vagas_fies`, `vagas_ocupadas`) e a estrutura financeira detalhada por semestre.

5.4. Auditoria de Consistência de Dados Pós-Padronização (Módulo 3 - Subpasta 2)

Após a etapa de renomeação de colunas e normalização de atributos, realizou-se uma auditoria final de integridade (`verificar_cursos_nan_em_cine`) para quantificar o estado das classificações CINE nos arquivos definitivos de análise (`inscritos_limpo.csv` e `ofertas_limpo.csv`).

A. Diagnóstico de Cobertura - Dataset de Inscrições

O *script* `inscritos.ipynb` verificou a presença de valores nulos nas colunas de classificação. Os resultados foram:

- **Códigos de Curso (NaN):** Identificou-se que 25.361 registros não possuem o código CINE específico vinculado ao curso individual.
- **Áreas Gerais (NaN):** O volume de registros sem a classificação de "Grande Área" foi de **zero**.

> **Conclusão Técnica:** Isso prova que a estratégia de cura de dados (Módulo 2) foi 100% eficaz para as inscrições. Mesmo nos casos em que o código histórico do curso não foi encontrado, o algoritmo de imputação garantiu que todos os candidatos fossem alocados em uma área do conhecimento, eliminando o risco de perda de amostra na filtragem por Computação.

B. Diagnóstico de Cobertura - Dataset de Ofertas

A auditoria no dataset de ofertas (`ofertas.ipynb`) revelou um cenário de alta fidelidade, com perdas residuais desprezíveis:

- **Códigos de Curso (NaN):** 6.119 registros sem código específico.
- **Áreas Gerais (NaN):** Restaram apenas **5 registros** totalmente sem classificação (representando cursos como "Normal Superior" e variações incompletas de "Comunicação Social").

Tabela 16: Resumo da Qualidade Final da Classificação (Módulo 3)

Dataset	Total de Registros	Áreas CINE Identificadas	Taxa de Sucesso
Inscrições	2.197.234	2.197.234	**100,00%**
Ofertas	143.622	143.617	**99,99%**

C. Conclusão da Fase de Preparação

A base de dados está agora tecnicamente selada. O processo de ETL (Extração, Transformação e Carga) e o subsequente *Data Healing* transformaram microdados governamentais inconsistentes em um conjunto de dados fidedigno, com nomes de colunas simplificados e classificação de área garantida para a quase totalidade do universo estudado.

A partir deste ponto, o projeto avança para a **Análise Exploratória**, com a segurança de que os dados de entrada são íntegros e representativos.

6. Consolidação e Arquitetura do Pipeline de Engenharia de Dados (ETL)

A base analítica desta pesquisa foi construída através de um ciclo completo de **ETL** (*Extract, Transform, Load*). Este processo tratou **milhões de linhas de registros de todos os inscritos e ofertas do FIES (2019-2021)**, garantindo que os dados estivessem tecnicamente corretos e enriquecidos com novas variáveis fundamentais para uma análise exploratória (AED) muito mais complexa e precisa.

6.1. Extração (Extract): Ingestão Automática e Higienização

O processo partiu do tratamento bruto de milhões de registros para todos os candidatos e vagas do programa:

- **Padronização via Script:** Para organizar o volume massivo de arquivos, utilizou-se um **script de automação** que renomeou todos os arquivos originais para o padrão `<ano>_<tipo>_<semestre>.csv`, organizando-os de forma lógica e automática.

- **Higiene Técnica:** O código corrigiu a codificação (*Latin-1* para *UTF-8*) e padronizou os números (trocando vírgula por ponto). Isso permitiu que os milhões de valores de renda e notas Enem fossem processados como números reais.
- **Limpeza Inicial:** Realizou-se a **remoção de linhas inteiramente vazias** e a **exclusão de duplicatas exatas**, garantindo que falhas de exportação do MEC não gerassem contagens erradas.

6.2. Transformação (Transform): Consolidação e o Sistema CINE

Esta fase foi onde o código gerou inteligência sobre os dados e unificou a base:

- **Unificação dos Datasets:** Após a limpeza inicial, todos os arquivos semestrais de 2019 até 2021 foram agrupados (`pd.concat`). O resultado foram **dois únicos arquivos mestres (.csv)**: um contendo todos os inscritos e outro com todas as ofertas do período.
- **Criação do Dataset Mestre INEP:** Para classificar as áreas dos cursos, foi baixado o dataset do INEP (2016-2024). Selecionamos as colunas de códigos e nomes oficiais dos cursos junto com as áreas e códigos CINE, gerando um **Dataset Mestre de Referência**.
- **Sistema de Duas Camadas (CINE):**
    - **Camada 1:** Realizou-se o *merge* entre a base do FIES e o Dataset Mestre do INEP pelo código do curso para trazer a classificação oficial.
    - **Camada 2 (Autocura via IA):** Para os registros que continuaram vazios (*NaN*), um script extraiu os nomes dos cursos sem área e outro buscou os mesmos nomes em registros já classificados. A IA cruzou essas tabelas e montou um **dicionário manual gigante**, preenchendo automaticamente a área e o código CINE baseando-se no nome do curso. Isso recuperou mais de **31 mil registros**.

6.3. Transformação: Regras de Negócio e Enriquecimento

Enriqueceu-se a base com informações que permitem análises profundas por região e perfil social:

- **Modalidades e Salário Mínimo:** Criou-se a coluna `modalidade_fies` usando lógica de salário mínimo dinâmico por ano (R$998 em 2019; R$1.045 em 2020; R$1.100 em 2021).
- **Filtro de Eliminados:** Implementou-se uma regra rígida de desclassificação para candidatos com **renda acima de 5 SM (Salário mínimo)** , renda nula ou sem oferta real de vaga, marcando-os como **"eliminado"**.
- **Regionalização:** Adicionou-se a classificação por **Grandes Regiões do Brasil** (Norte, Sul, etc.) baseada na UF, permitindo comparar o programa em todo o território nacional.

6.4. Carga e Verificação de Integridade (Load)

A fase final foi o carregamento dos dados para a geração da **Camada Ouro (Gold Layer)**:

- **Verificação de Chaves Primárias (PKs):** Em cada etapa, realizou-se a auditoria comparando o total de linhas (Ntotal) com o total de registros únicos (Nunique). Isso provou que **não houve duplicação indesejada** nos milhões de registros após os cruzamentos.
- **Simplificação para Análise:** Renomearam-se 65 colunas gigantes para nomes técnicos curtos (ex: `renda_per_capita`), otimizando a performance para a análise exploratória.
- **Resultado Final:** Gerou-se os arquivos `inscritos_limpo.csv` e `ofertas_limpo.csv` prontos para a análise exploratória.

Conclusão da Engenharia: Com este pipeline, milhões de registros foram saneados e enriquecidos com **Região, UF, Modalidade FIES e Classificação CINE**. A base de dados agora está pronta para uma análise exploratória robusta.

7. Análise Exploratória: Dinâmica de Demanda e Funil de Conversão

Após a consolidação da infraestrutura técnica (ETL), o projeto avançou para a **Análise Exploratória de Dados (AED)**. O objetivo desta fase foi decompor os milhões de registros para entender o comportamento dos candidatos em todas as áreas do conhecimento, diferenciando o volume de "papéis no sistema" da quantidade de "pessoas reais" buscando o financiamento.

7.1. Metodologia de Identificação de Candidatos Únicos

Um dos maiores desafios estatísticos do FIES é que um único estudante pode realizar várias inscrições (opções de curso) em diferentes semestres. Para enxergar o cidadão por trás dos dados, foi implementada uma lógica de deduplicação baseada na **Prioridade de Desfecho** (`candidatos_unicos_e_agrupados.ipynb`).

- **Hierarquia de Sucesso:** Criou-se uma regra de prioridade para a coluna `situacao_fies` (Ex: `CONTRATADA` > `PRÉ-SELECIONADO` > `LISTA DE ESPERA` > `NÃO CONTRATADO`).
- **Isolamento do Indivíduo:** O algoritmo ordenou os registros por `id_estudante` e aplicou a hierarquia. Se um aluno teve uma inscrição contratada e outra não, o sistema isolou apenas o registro de sucesso.
- **Duas Visões de Análise:** Essa etapa gerou uma coluna extra que permite ao pesquisador alternar entre duas visões fundamentais:

1. **Visão por Inscrições (Demanda Nominal):** Mede o volume bruto de opções e a carga total de processamento do sistema.
2. **Visão por Candidatos Únicos por prioridade (Demanda Real):** Mede o número real de indivíduos físicos disputando as vagas, essencial para traçar o perfil socioeconômico fiel da população.

Resultado Consolidado: O processo identificou **1.109.893 candidatos únicos** em todo o Brasil entre 2019 e 2021.

7.2. Construção do Funil de Conversão e Regionalização

Para analisar a eficiência do programa, foi estruturado um **Funil de Conversão Multinível** (`funil_por_regiao.ipynb`). Este funil cruza os dados de oferta e demanda em todas as regiões, utilizando as seguintes camadas de filtragem:

1. **Vagas Ofertadas:** Capacidade total de financiamento disponibilizada.
2. **Inscritos Geral:** Volume total de inscrições realizadas.
3. **Candidatos Únicos:** Quantidade de CPFs únicos na disputa.
4. **Candidatos com Nota Suficiente:** Filtro acadêmico onde `media_enem >= nota_corte_gp`.
5. **Vagas Ocupadas:** O desfecho final da contratação.

A análise foi enriquecida com a **Grande Regionalização**, integrando as informações de **UF** e **Região** (Norte, Nordeste, Centro-Oeste, Sudeste e Sul) para cada inscrito e cada oferta. Isso permite identificar, por exemplo, se o gargalo de ocupação de vagas é causado por falta de demanda ou por notas de corte incompatíveis com o desempenho dos candidatos em determinadas regiões.

7.3. Representatividade por Áreas CINE

A auditoria de classificação confirmou que o processo de "autocura" realizado no ETL foi eficaz para categorizar a base inteira. O dataset final apresenta dados de **10 das 11 áreas gerais da CINE**.

Área Geral CINE	**Registros de Agrupamento (Ano/Semestre/UF)**
Saúde e bem-estar	162

Negócios, administração e direito	162
Engenharia, produção e construção	162
Educação	161
Computação e TIC	161
Ciências sociais, comunicação e informação	162
Serviços	149
Agricultura, silvicultura, pesca e veterinária	146
Artes e humanidades	142
Ciências naturais, matemática e estatística	89

A única área ausente ("Programas genéricos e disciplinas interdisciplinares") reflete a ausência desse tipo de oferta específica no catálogo do FIES para o período analisado.

7.4. Conclusão da Etapa Exploratória Inicial

Com a criação do dataset de funil e a identificação dos candidatos únicos, os dados estão prontos para as comparações do Módulo 4. A existência das "duas visões" (Inscritos vs. Candidatos por prioridade) garante que a pesquisa possa responder tanto a perguntas sobre a eficiência do sistema quanto sobre o perfil humano e social dos brasileiros que buscam o ensino superior privado.

8. Visualização e Comparação do Funil FIES por Grande Área (Módulo 4.1)

Após a conclusão do pipeline de engenharia e o saneamento de milhões de registros, a fase de **Análise Exploratória de Dados (AED)** permitiu diagnosticar a eficiência do programa através de uma arquitetura de funil multinível. Esta seção apresenta os resultados visuais que confrontam a oferta institucional com o comportamento real da demanda em escala nacional.

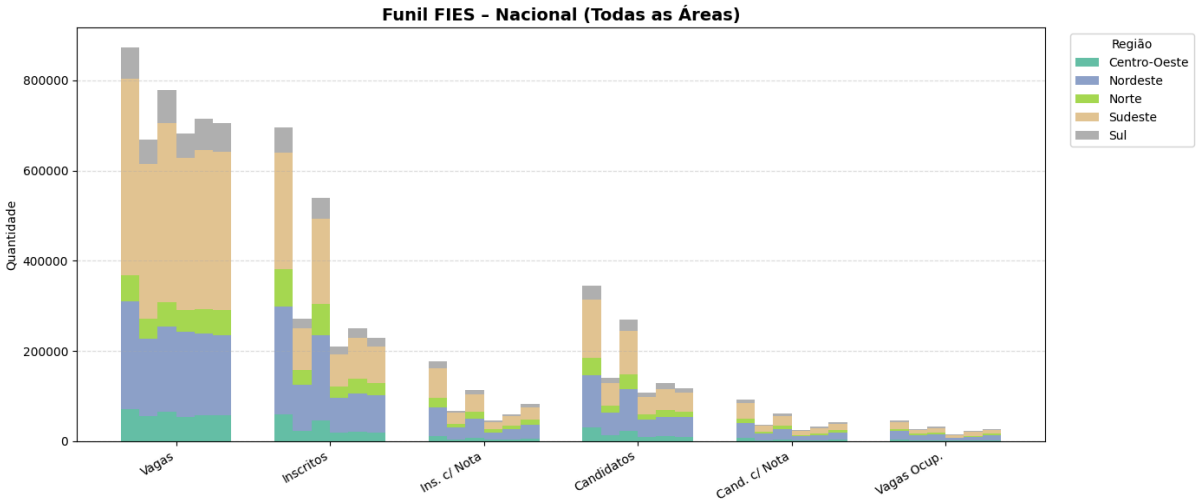
8.1. Metodologia de Visualização e Métricas

Para cada grande área do conhecimento, gerou-se um gráfico de barras empilhadas segmentado por **Região** (cores) e agrupado por **Período** (ano/semestre). A estrutura do funil foi desenhada para expor a perda de eficiência em seis estágios críticos:

1. **Vagas**: Total de financiamentos autorizados e ofertados pelas IES.
2. **Inscritos (Demanda Nominal)**: Volume bruto de opções de cursos registradas no sistema.
3. **Ins. c/ Nota**: Inscritos que superaram a nota de corte do grupo de preferência.
4. **Candidatos únicos por prioridade (Demanda Real)**: Métrica que isola o indivíduo físico através de seu ID, mantendo apenas o registro de maior relevância hierárquica de desfecho (ex: priorizando "Contratada" sobre "Não Selecionada").
5. **Cand. c/ Nota**: Volume de **Candidatos únicos por prioridade** que possuem viabilidade acadêmica (nota ENEM suficiente).
6. **Vagas Ocup.**: Desfecho final com a contratação efetiva do financiamento.

8.2. Panorama Nacional por Região

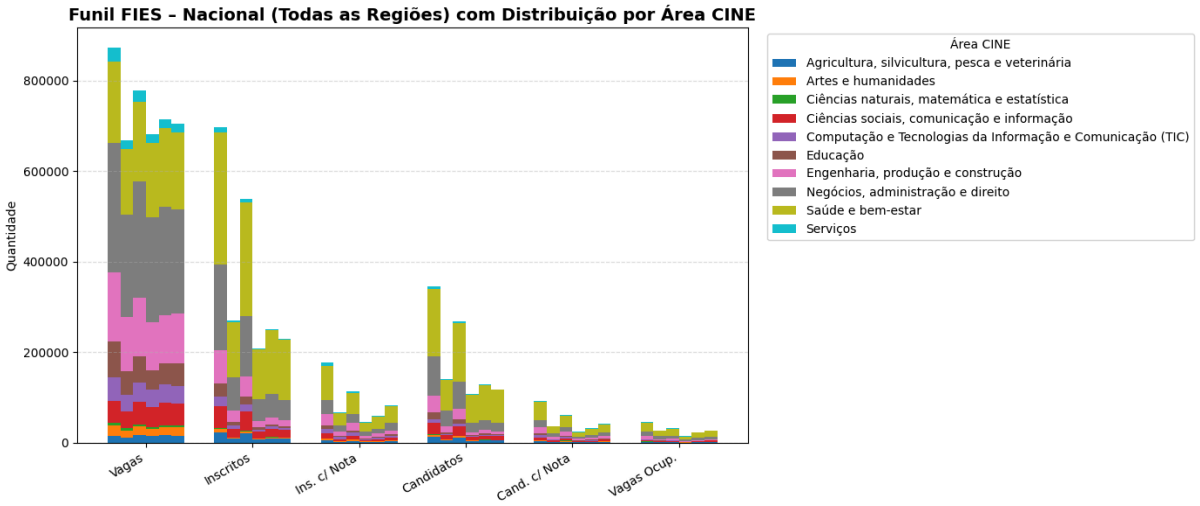
O gráfico nacional consolidado permite observar a assimetria na distribuição dos recursos do FIES no território brasileiro:



- **Concentração Geográfica**: O **Sudeste** e o **Nordeste** dominam o volume absoluto do programa em todas as etapas do funil, refletindo a densidade populacional e a oferta de vagas nessas regiões.
- **Declínio Temporal**: Observa-se uma redução progressiva no tamanho das barras de 2019 para 2021, indicando um encolhimento na oferta de vagas e no volume de candidatos ao longo do triênio analisado.
- **Ociosidade Estrutural**: Em todas as regiões, a coluna final de "Vagas Ocupadas" é drasticamente inferior à oferta inicial de "Vagas", provando que a vacância não é um fenômeno regional, mas sim estrutural do programa no período.

8.3. Panorama Nacional por Área CINE

A análise por área de conhecimento revela a hierarquia de interesses do mercado de ensino superior privado e a eficácia do sistema de "autocura" e classificação CINE implementado no Módulo 2.



- **Domínio da Saúde e Negócios**: As áreas de **Saúde e bem-estar** e **Negócios, administração e direito** concentram os maiores volumes de demanda. Nestas áreas, a diferença entre a coluna de **Inscritos** e **Candidatos únicos por prioridade** é acentuada, indicando que um mesmo estudante realiza múltiplas tentativas para garantir o acesso a essas carreiras.
- **Representatividade**: O dataset validado apresenta dados de **10 das 11 áreas gerais CINE**, com a única ausência sendo "Programas genéricos e disciplinas interdisciplinares", o que é consistente com a natureza das ofertas profissionais do FIES.

8.4. Diagnóstico dos Gargalos: Nota de Corte e Conversão

A comparação visual das colunas centrais do funil expõe os dois maiores entraves à ocupação de vagas no programa:

1. **Gargalo Acadêmico**: Em todas as áreas, há uma queda acentuada entre **Candidatos únicos por prioridade** e **Cand. c/ Nota**. Isso demonstra que uma

parcela expressiva da demanda real é retida pelo critério de desempenho no ENEM (nota de corte), impossibilitando a continuidade no processo seletivo.

2. **O Descompasso na Contratação**: Em diversos cenários, o volume de candidatos com nota suficiente é superior ao número de vagas efetivamente ocupadas. Isso sugere que impedimentos pós-seleção (como a falta de fiadores, exigências burocráticas ou limites de financiamento) impedem que candidatos aptos academicamente concretizem o contrato.

8.5. Conclusão da Etapa Visual

A utilização da métrica de **Candidatos únicos por prioridade** foi fundamental para esta análise, pois eliminou a inflação estatística gerada pelas múltiplas opções de um mesmo estudante. Enquanto o volume de **Inscritos** gera uma percepção de concorrência massiva, a visão por **Candidatos únicos por prioridade** revela o número real de indivíduos que buscam o financiamento, permitindo que as próximas etapas foquem no perfil socioeconômico fiel desta população.