# **Relatório Final: Análise de Fomento à Pesquisa no Ensino Superior Brasileiro**

## **1. Introdução e Objetivo**

Este documento apresenta os resultados de um projeto completo de análise de dados, cujo objetivo foi investigar os fatores associados ao fomento à pesquisa no ensino superior do Brasil. Utilizando microdados públicos do Censo da Educação Superior (INEP) e do portal de dados abertos do CNPq, ambos referentes ao ano de 2023, o projeto buscou identificar padrões de distribuição de bolsas, características institucionais relacionadas ao sucesso na captação de recursos e, por fim, gerar insights que possam subsidiar a tomada de decisão para políticas públicas e gestão acadêmica.

## **2. Documentação Detalhada e Abordagem Metodológica**

A execução do projeto seguiu uma metodologia estruturada em quatro fases principais: Engenharia de Dados (ETL), Análise Exploratória (EDA), Análise Estatística e Relatório de Conclusões.

### **2.1. Engenharia de Dados (ETL - Extração, Transformação e Carga)**

Esta fase foi crucial para garantir a qualidade e a integridade dos dados para a análise subsequente.

* **Extração:** Os dados foram carregados a partir de três fontes CSV distintas:
  + **INEP:** MICRODADOS\_ED\_SUP\_IES\_2023.CSV (Dados cadastrais das Instituições de Ensino Superior).
  + **INEP:** MICRODADOS\_CADASTRO\_CURSOS\_2023.CSV (Dados sobre os cursos de graduação).
  + CNPq: Relatorio\_de\_dados\_abertos\_CNPq (1º SEM 2023)(snICJ).csv (Dados sobre bolsas de fomento concedidas).  
    A abordagem inicial de download automatizado foi substituída pelo upload manual dos arquivos no ambiente Google Colab devido à instabilidade dos servidores de origem.
* **Tratamento e Padronização:**
  1. **Normalização de Nomes de Colunas:** Uma função foi criada para padronizar todos os cabeçalhos para o formato snake\_case (minúsculas, sem acentos, com underscores), eliminando inconsistências e facilitando o acesso programático.
  2. **Tratamento de Codificação:** O arquivo do CNPq apresentou problemas de formatação de caracteres. A leitura foi ajustada para a codificação utf-8-sig, que corrigiu a exibição de acentos e removeu o caractere de marca de ordem de byte (BOM) do início do arquivo.
  3. **Limpeza de Dados:** Colunas irrelevantes, como Unnamed:, geradas por possíveis erros de exportação nos arquivos originais, foram sistematicamente removidas.
* **Criação da Base Analítica Unificada:** O passo mais complexo foi a criação de um único DataFrame analítico.
  1. **Agregação do INEP:** Os dados do arquivo de Cursos foram agregados para criar uma nova variável: qtd\_cursos\_graduacao para cada instituição (co\_ies), servindo como um *proxy* para o tamanho e a diversidade da IES.
  2. **Agregação do CNPq:** Os dados de bolsas foram agregados para obter a variável total\_bolsas para cada instituição. A coluna instituicao\_destino foi identificada como a chave para o nome da IES.
  3. **Junção (Merge):** Para unificar as bases, foi essencial normalizar os nomes das instituições em ambos os conjuntos de dados (convertendo para maiúsculas e removendo acentos). Isso criou uma chave de junção comum (nome\_normalizado) que permitiu o cruzamento bem-sucedido das informações.

### **2.2. Análise Exploratória (EDA) e Visualização**

Com o dataframe unificado, a análise buscou responder às perguntas do desafio:

* **Análise de Distribuição:** Foram gerados gráficos de barras para identificar as **Top 20 instituições** e as **Top 15 áreas do conhecimento** com maior número de bolsas.
* **Análise Geográfica:** Um **mapa coroplético** foi criado para visualizar a distribuição de bolsas entre os estados brasileiros, destacando as concentrações regionais.
* **Análise Comparativa:** A disparidade entre **instituições públicas e privadas** foi analisada com um **boxplot**, que ilustra de forma eficaz a distribuição, a mediana e a amplitude dos dados para cada categoria.

### **2.3. Análise Estatística**

Para validar as observações visuais, foram aplicados testes estatísticos:

* **Teste T de Welch:** Utilizado para comparar a média de bolsas entre instituições públicas e privadas. O teste confirmou que a diferença observada é **estatisticamente significativa**.
* **Correlação de Pearson:** Aplicada para medir a força e a direção da relação linear entre o qtd\_cursos\_graduacao e o total\_bolsas. O resultado indicou uma correlação positiva e significativa.

## **3. Principais Desafios Enfrentados**

1. **Disponibilidade dos Dados:** O maior obstáculo técnico foi a **instabilidade dos servidores de origem**, que impediu a automação do processo de download e exigiu a adoção de um fluxo de trabalho com upload manual.
2. **Inconsistência dos Dados:** Foram encontrados problemas clássicos de dados públicos, como codificação de caracteres incorreta, nomes de colunas não padronizados e variações nos nomes das instituições, que exigiram um robusto processo de limpeza e normalização.
3. **Limitação de Dados:** A **ausência do arquivo de dados de Docentes** (MICRODADOS\_CADASTRO\_DOCENTES\_2023.CSV) foi a principal limitação analítica. Sem ele, não foi possível testar a hipótese central de que a titulação dos professores (ex: número de doutores) é um fator chave para a captação de fomento. A análise foi, portanto, adaptada de forma inteligente para utilizar as variáveis disponíveis.

## **4. Descobertas Mais Relevantes**

* **Hegemonia do Setor Público:** A análise estatística confirmou que **instituições públicas recebem, em média, um volume de bolsas significativamente maior** que as instituições privadas.
* **Concentração Geográfica:** A pesquisa no Brasil é altamente concentrada. O mapa e os gráficos mostram que a grande maioria das bolsas se destina a instituições da região **Sudeste**, com algumas universidades específicas (USP, UNICAMP, UFRJ, UNESP) liderando com folga.
* **Relação entre Tamanho/Diversidade e Fomento:** A análise de correlação demonstrou uma associação **positiva e estatisticamente significativa** entre o **número de cursos de graduação** de uma IES e o volume de bolsas que ela atrai. Isso sugere que instituições maiores e mais diversificadas tendem a ter ecossistemas de pesquisa mais estabelecidos.
* **Áreas de Destaque:** As grandes áreas de **Ciências da Saúde**, **Ciências Biológicas** e **Engenharias** foram as que receberam o maior número de bolsas no período analisado, indicando os principais focos de pesquisa ou as maiores comunidades de pesquisadores do país.

## **5. Recomendações Baseadas nos Dados**

1. **Para Órgãos de Fomento (CAPES/CNPq):** Para promover um desenvolvimento científico mais equitativo, recomenda-se a criação de **políticas de fomento com foco na descentralização**. Editais específicos e programas de incentivo para instituições localizadas nas regiões Norte, Nordeste e Centro-Oeste podem ser estratégias eficazes para diminuir a acentuada assimetria geográfica.
2. **Para Gestão de IES:** O crescimento e a diversificação do portfólio de cursos de uma instituição parecem estar associados a uma maior capacidade de atrair fomento. Isso pode ser um indicador de maturidade institucional que favorece a pesquisa. Instituições, especialmente as privadas, devem avaliar como a expansão estratégica da oferta de cursos pode alavancar a captação de recursos para pesquisa.
3. **Para Análises Futuras:** A inclusão dos dados de docentes é fundamental. A próxima etapa natural deste projeto seria cruzar os dados atuais com as informações de titulação (qtd\_doutores) e regime de trabalho para construir um modelo preditivo mais completo sobre os fatores que impulsionam a produção científica no Brasil.