
Relatório de Análise: Projetos de Investimento no DF (ObrasGov)

Data: 17 de Outubro de 2025

1. Introdução e Objetivo

Este relatório apresenta os resultados de uma análise exploratória dos dados de projetos de investimento público no Distrito Federal (DF), obtidos através da API da plataforma ObrasGov.br. O objetivo principal foi extrair, tratar e visualizar os dados para identificar padrões de investimento, entender a distribuição de recursos por setor e órgão executor, avaliar o status atual dos projetos e comunicar os insights de forma clara.

2. Fonte dos Dados

Os dados foram coletados programaticamente via API pública do ObrasGov.br, utilizando o endpoint /projeto-investimento com filtro para a Unidade Federativa uf=DF. O processo de extração exigiu a implementação de lógicas para lidar com a paginação da API e limites de taxa de requisição (erros 429). Os dados brutos extraídos (834 registros) foram salvos em ../data/raw/.

3. Metodologia e Tratamento dos Dados

O processo de análise seguiu um fluxo rigoroso de engenharia e tratamento de dados, documentado no notebook analise_obrasgov_df.ipynb:

1. **Aquisição e Carga:** Extração via API e carregamento inicial dos dados brutos.
2. **Limpeza Fundamental:** Criação de uma cópia de trabalho (df_tratado) e remoção imediata de registros duplicados baseados no id_unico (resultando em 667 registros únicos).
3. **Diagnóstico Pós-Deduplicação:** Análise detalhada da estrutura, tipos de dados, valores ausentes (Nulos) e estatísticas descritivas sobre os dados únicos.
4. **Engenharia de Features (Flattening):** Expansão das colunas com dados estruturados (strings de listas/dicionários como executores, eixos, fontes_de_recurso, etc.) em novas colunas "achatadas" (ex: executor_nome, eixo_descricao, origem_recurso).
5. **Cálculo de Métricas:** Criação da coluna valor_total_previsto pela soma validada dos valores em fontes_de_recurso.
6. **Tipagem Adequada:** Conversão das colunas para os tipos semânticos corretos e otimizados (datetime64[ns], float64, boolean, string, category), com validação da escolha baseada na cardinalidade.
7. **Identificação e Registro de Anomalias:** Documentação formal dos problemas de qualidade encontrados (detalhados na seção abaixo).
8. **Otimização e Limpeza Final:** Remoção das colunas originais redundantes do DataFrame master (df_tratado).
9. **Armazenamento:** Salvamento do df_tratado na tabela projetos_investimento do banco SQLite (../db/obras_df.db).

10. **Preparação para Análise:** Carregamento dos dados do SQLite, criação do df_analise focado na visualização, e filtragem final de 110 registros com valores simbólicos (valor_total_previsto <= R$ 1). Salvamento do df_analise (557 registros) em ../data/processed/.

4. Anomalias e Qualidade dos Dados Encontrados

Durante a exploração e tratamento, diversas anomalias e problemas de qualidade foram identificados, impactando a análise e exigindo tratamento específico:

1. **Nomes de Projetos como Identificadores:** Um número significativo de projetos na coluna nome continha códigos de processo, contrato ou outros identificadores (ex: 26.782.3006.7XT1.0053, 00 00680/2023, 202111-22-Ronald 1), dificultando a interpretação direta em algumas visualizações. Os dados foram mantidos conforme a origem.

2. **Registros Duplicados e Inconsistentes:** Foram encontrados 158 registros com id_unico duplicado. Desses, 99 eram cópias exatas, mas 59 apresentavam dados diferentes em outras colunas, indicando inconsistências na fonte. Adotou-se a estratégia de manter apenas a primeira ocorrência (keep='first').

3. **Valores Simbólicos:** Identificou-se 110 projetos (aprox. 16.5% do total único) com valor_total_previsto igual a R$ 0.00 ou R$ 0.01. Estes foram considerados *placeholders* ou erros de cadastro e foram **filtrados** do DataFrame df_analise para não distorcerem as estatísticas e visualizações de valor.

4. **Alta Porcentagem de Valores Nulos:** Colunas relevantes para análise de impacto, como qdt_empregos_gerados (80.51% nulos pós-deduplicação) e populacao_beneficiada (80.81% nulos pós-deduplicação), apresentaram preenchimento muito baixo, limitando a profundidade da análise nesses aspectos.

5. Principais Descobertas e Insights

As análises visuais e estatísticas sobre os 557 projetos com valores realistas revelaram os seguintes padrões:

1. **Distribuição de Valores:** A maioria dos projetos se concentra na faixa de **R$ 1 milhão a R$ 10 milhões**. A presença de poucos projetos de altíssimo custo (outliers) eleva a média (R$ 15,07M) muito acima da mediana (R$ 2,10M).

2. **Evolução Temporal:** Houve um pico expressivo de cadastros (quantidade e valor) em **2021**, seguido por uma estabilização em patamares mais baixos, porém consistentes, nos anos subsequentes (~100 projetos/ano, ~R$ 1 Bi/ano).

3. **Distribuição Setorial:** O setor **Econômico** domina os investimentos (46.74% do valor), seguido pelo **Social** (25.46%), **Administrativo** (~15%) e **Militar** (~11%).

4. **Principais Executores (por Quantidade):** Órgãos federais lideram, com **DNIT** (70 projetos) à frente, seguido por **IFB** (48) e **FUB (UnB)** (45), destacando o foco em transporte e educação/pesquisa.

5. **Situação dos Projetos:** A grande maioria (**70.38%**) dos projetos está apenas "Cadastrada", indicando um volume significativo de planejamento ainda não iniciado. Apenas ~15% estão "Em execução" e ~10% "Concluídos".

(Os insights detalhados e as explicações para cada análise encontram-se nas células de Markdown correspondentes dentro do notebook.)

6. Conclusão

A análise dos dados do ObrasGov para o Distrito Federal oferece um panorama valioso sobre o planejamento e a execução de investimentos públicos. Revelou-se uma concentração de projetos de médio porte, um pico de planejamento em 2021, a dominância do setor econômico e de executores federais, e um grande volume de projetos ainda na fase inicial de cadastro. As anomalias encontradas ressaltam a importância de um tratamento cuidadoso dos dados antes da análise.