**Tratamento e Análise de Dados Agrícolas Públicos (SIDRA/IBGE)**

Relatório Técnico-Científico

Trabalho desenvolvido em parceria entre a Becomex e a Escola DNC

Grupo 4 – Turma 2025  
Nome dos participantes:

Data de entrega: 27/06/25

Trabalho apresentado e aprovado em 30 de junho de 2025.

**Prof. Guilhermino dos Santos  
 Orientador – Escola DNC**

**Denys Cabral | Gustavo Felizardo  
 Representante Técnico – Becomex**

**SUMÁRIO**

**1. Business Understanding**

**1.1 Objetivo do Negócio**

**1.2 Avaliação da Situação**

**1.3 Metas de Mineração de Dados**

**1.4 Plano do Projeto**

**2. Data Understanding**

**2.1 Coleta Inicial de Dados**

**2.2 Descrição dos Dados**

**2.3 Explorando os Dados**

**2.4 Qualidade dos Dados**

**3. Data Preparation**

**3.1 Selecionar Dados**

**3.2 Limpando os Dados**

**3.3 Construindo os Dados**

**3.4 Integrando os dados**

**3.5 Formatando os Dados**

**4. Modeling**

**4.1 Selecionando Técnica de Modelagem**

**4.2 Gerar Design de Teste**

**4.2.1 Modelo de Construção**

**4.3 Avaliar o Modelo**

**5. Evaluation**

**5.1 Avaliação dos Resultados**

**5.2 Revisão do Processo**

**5.3 Determinação dos Próximos Passos**

**6. Deployment**

**6.1 Plano de Implementação**

**6.2 Plano de Monitoramento e Manutenção**

**6.3 Relatório Final**

**6.4 Revisão do Projeto**

**7. Referências e Recursos Utilizados**

**1. Business Understanding**

**1.1 Objetivo do Negócio**

O objetivo principal deste projeto é estruturar, tratar e organizar os dados públicos do agronegócio brasileiro, disponibilizados pelo SIDRA/IBGE, a fim de torná-los acessíveis, compreensíveis e práticos para análises e tomadas de decisão.

A proposta é permitir que usuários, sejam eles stakeholders internos ou externos, possam realizar consultas como:

* "Onde estão localizadas as plantações de café no Brasil?"
* "Qual município ou estado possui maior produção de soja?"
* "Como se distribuem as lavouras permanentes e temporárias no território brasileiro?"

Para isso, será construída uma base de dados limpa, estruturada e pronta para ser consumida, além de um dashboard interativo que permitirá a visualização dos principais indicadores e KPIs do setor agrícola.

O projeto tem como finalidade principal democratizar o acesso a dados públicos do setor, transformando-os em informações estratégicas para apoio à tomada de decisão, desenvolvimento de pesquisas, apoio comercial, mapeamento de mercado e geração de insights econômicos.

**1.2 Avaliação da Situação**

O projeto foi desenvolvido considerando os seguintes recursos e restrições:

* **Fontes de Dados:** Dados públicos disponibilizados pelo SIDRA/IBGE, extraídos via API utilizando o pacote sidrapy em Python.
* **Equipe Técnica:** Alunos da Escola DNC responsáveis pela coleta, tratamento, análise e desenvolvimento do dashboard.
* **Ferramentas:** Python (Pandas e sidrapy,), Google Colab para desenvolvimento e Power BI para construção do dashboard final.
* **Restrições:** Limitações associadas ao formato dos dados fornecidos pelo SIDRA, que muitas vezes vêm em estruturas pouco amigáveis, demandando elevado esforço na etapa de limpeza e transformação.
* **Infraestrutura:** O projeto foi desenvolvido em ambiente cloud (Google Colab) e ferramentas de BI desktop/cloud (Power BI).
* **Riscos:** Qualidade dos dados, ausência de documentação clara em alguns datasets do IBGE, atualizações periódicas dos dados que exigem processos de atualização e manutenção recorrentes.
* **Benefícios Esperados:** Entregar uma base estruturada e um dashboard capaz de fornecer respostas rápidas, confiáveis e visualmente intuitivas, reduzindo tempo de análise e aumentando a capacidade de tomada de decisão.

**1.3 Metas de Mineração de Dados**

As metas técnicas do projeto são:

* Coletar dados agrícolas do SIDRA/IBGE de forma automatizada via API.
* Estruturar e higienizar os dados, removendo inconsistências, valores ausentes, códigos não tratados e possíveis duplicatas.
* Criar novos atributos derivados que aumentem o valor analítico da base, como total de produção por estado, área plantada consolidada, entre outros.
* Gerar uma base de dados pronta para análise exploratória, construção de dashboards e desenvolvimento futuro de modelos preditivos, caso necessário.
* Implementar visualizações que permitam identificar padrões espaciais e temporais no agronegócio brasileiro.

**1.4 Plano do Projeto**

**Fases do Projeto:**

1. **Coleta dos Dados:**
   * Via API SIDRA utilizando o pacote sidrapy.
2. **Análise e Entendimento dos Dados:**
   * Análise da estrutura das tabelas, chaves, hierarquias e padrões dos dados brutos.
3. **Limpeza e Transformação dos Dados:**
   * Padronização de nomenclaturas, tratamento de valores ausentes, criação de variáveis utilizando python e DAX futuramente no Power Bi.
4. **Criação de Base Analítica:**
   * Integração dos dados e formatação no modelo final.
5. **Desenvolvimento do Dashboard:**
   * Power BI para visualização dos principais KPIs.
6. **Documentação Técnica:**
   * Relatório detalhado, documentação do processo e dicionário de dados.

**Ferramentas:** Python (sidrapy, pandas,), Google Colab, Microsoft Excel e Power BI

**2. Data Understanding**

**2.1 Coleta Inicial de Dados**

A coleta de dados teve como origem o banco oficial do SIDRA/IBGE, com foco em tabelas relacionadas ao setor agropecuário, como a Tabela 1613(CULTURAS PERMANENTES) e a Tabela 1612 (CULTURAS TEMPORARIAS) entre outras relevantes para a análise da produção agrícola nacional. A escolha dessas tabelas se deu pela profundidade estatística, cobertura geográfica abrangente e periodicidade histórica, características que as tornam ideais para análises comparativas e extração de insights estratégicos.

A extração dos dados foi realizada utilizando a biblioteca Python sidrapy, o que possibilitou uma coleta automatizada, padronizada e documentada. Durante esse processo, foi adotada uma estrutura de versionamento dos arquivos, com separação por tipo de indicador (quantitativo absoluto, percentual, variações anuais etc.), bem como organização por nível geográfico (Brasil, estados e municípios).

O tratamento de dados contemplou a padronização de colunas, conversão de unidades de medida, limpeza de valores nulos e inconsistentes, além da criação de campos derivados para facilitar análises temporais e espaciais. Também foram consideradas as legendas de tabelas do IBGE, com especial atenção às simbologias utilizadas nas tabelas, como “..” (dado inexistente), “X” (não aplicável) e “0” (valor zero), garantindo maior rigor na interpretação dos dados.

Como parte do processo de gestão e controle do projeto, foi utilizada a ferramenta Trello para organização das tarefas em fluxo contínuo, mapeando as etapas de coleta, tratamento, validação e modelagem dos dados. Essa estrutura contribuiu para o alinhamento das entregas e visibilidade do progresso em tempo real.

Ao final, os dados tratados foram utilizados para a construção de dashboards interativos no Power BI, possibilitando uma visualização clara e dinâmica dos principais indicadores do setor agropecuário. Além disso, está em desenvolvimento uma API em Python para consumo dos dados estruturados, facilitando o acesso automatizado e integrado por parte da equipe.

**2.2 Descrição dos Dados**

* **Origem:** Tabelas do SIDRA relacionadas à produção agrícola.
* **Formato:** Dados tabulares, entregues em CSV, XLSX ou JSON via API.
* **Principais Colunas Identificadas:**
  + Município
  + Estado (UF)
  + Produto (Ex.: Soja, Café, Milho, etc.)
  + Tipo de lavoura (temporária ou permanente)
  + Área plantada (hectares)
  + Área colhida (hectares)
  + Produção (toneladas ou unidade específica)
  + Ano de referência
* **Volume de Dados:** Aproximadamente milhares de linhas por ano, dependendo do produto e da granularidade (municipal, estadual ou nacional).

**2.3 Exploração dos Dados**

Na etapa de exploração dos dados, foram identificadas informações estratégicas relacionadas à distribuição da atividade agrícola no Brasil. Essa análise preliminar possibilitou:

* Mapear as principais áreas produtoras de cada cultura analisada;
* Realizar comparações entre os dados dos anos de 2021, 2022 e 2023, observando variações relevantes;
* Medir o rendimento médio por hectare das culturas selecionadas;
* Investigar como a dimensão das áreas cultivadas se relaciona com os valores financeiros gerados.

A partir dessas análises iniciais, surgiram questões-chave que direcionaram o desenvolvimento dos dashboards e a estruturação dos dados, como:

* Quais culturas são economicamente mais relevantes em termos de valor de produção?
* Em quais localidades se concentram os maiores índices de produtividade?
* Existe correlação entre o tamanho territorial das áreas produtivas e o retorno econômico obtido?

Essas perguntas nortearam a definição dos filtros e indicadores utilizados nos painéis interativos, contribuindo para uma visualização eficiente das informações e facilitando a extração de insights voltados à tomada de decisão e análise de desempenho regional.

**2.4 Qualidade dos Dados**

A etapa de qualificação dos dados foi tratada com especial cuidado, considerando a complexidade das bases do SIDRA/IBGE. Diversas ações foram implementadas paragarantir a integridade e usabilidade das informações.

* **Problemas Encontrados:**
  + Valores nulos em algumas variáveis específicas.
  + Divergência de unidades em algumas tabelas (Kg, Tonelada, etc.).
* **Soluções Aplicadas:**
  + Criação de dicionários para decodificação.
  + Conversão de unidades onde aplicável.
  + Remoção de registros inconsistentes e preenchimento de nulos quando possível.

**3. Data Preparation**

**3.1 Seleção dos Dados**

A seleção das bases de dados teve como ponto de partida uma análise criteriosa das Tabelas 1612 e 1613 do SIDRA/IBGE, que apresentam informações detalhadas sobre a área plantada, área colhida, quantidade produzida e valor da produção agrícola. Considerando o escopo do projeto e o volume de dados disponíveis, foi necessário estabelecer um recorte estratégico, guiado por três principais critérios:

* **Importância econômica das culturas agrícolas**, tanto no mercado interno quanto no cenário de exportações, priorizando produtos com maior impacto financeiro e representatividade nacional;
* **Viabilidade técnica para o tratamento dos dados**, levando em conta a estrutura das tabelas, a padronização das variáveis e a possibilidade de automação via script Python;
* **Capacidade analítica das variáveis contidas nas tabelas**, com foco em indicadores que permitissem gerar KPIs relevantes e visualizações estratégicas, como produtividade por hectare e valor gerado por região.

Essa seleção foi essencial para garantir um equilíbrio entre profundidade analítica e eficiência no processamento, assegurando que as informações extraídas fossem relevantes para a geração de insights e aplicáveis no contexto de apoio à decisão da equipe da Becomex.

**3.2 Limpeza dos Dados**

* A etapa de limpeza dos dados foi fundamental para garantir a consistência e a qualidade da base coletada junto ao SIDRA/IBGE. Como os arquivos apresentavam símbolos e estruturas heterogêneas, especialmente nas Tabelas 1612 e 1613, foram adotadas as seguintes ações, inicialmente conduzidas com scripts exploratórios em Python:
* **Remoção de colunas e linhas totalmente vazias**, que não agregavam valor à análise e comprometiam a estrutura do dataset;
* **Substituição e exclusão de símbolos utilizados pelo IBGE**, como:
* '..': dado não disponível;
* 'X': não aplicável;
* '\*': informação suprimida por sigilo estatístico;
* **Conversão de tipos de dados**, ajustando colunas de texto (object) para tipos numéricos (float ou int) conforme o contexto da variável;
* **Padronização dos nomes de produtos e variáveis**, garantindo uniformidade e facilitando agrupamentos, filtros e visualizações no Power BI;
* **Organização das variáveis por tipo de indicador**, separando claramente informações sobre área plantada, área colhida, produção e valor da produção;
* **Tratamento das unidades de medida**, com base nas descrições metodológicas do IBGE, assegurando coerência entre anos e culturas analisadas.

Essas etapas tornaram possível a construção de uma base limpa, estruturada e confiável, adequada para análises comparativas, geração de KPIs e visualização em dashboards.

Além disso, todo o processo de transformação e carregamento dos dados foi centralizado no **Power Query**, dentro do próprio ambiente do Power BI onde mais algumas colunas e tabelas auxiliares foram criadas utilizando DAX conforme demandas operacionais para confecções do dashboard.

**3.3 Construção dos Dados**

Com os dados devidamente limpos e padronizados, foi possível estruturar o modelo analítico no formato **estrela (Star Schema)**, amplamente recomendado para projetos de Business Intelligence devido à sua simplicidade, desempenho e clareza na organização dos dados.

A estrutura do modelo seguiu os seguintes princípios:

* **Normalização dos dados em uma tabela fato e dimensões auxiliares**, separando indicadores quantitativos de atributos descritivos;
* **Centralização dos dados produtivos na tabela fato “Principal”**, contendo colunas como área colhida, área plantada, valor da produção e identificação dos municípios e anos;
* **Criação de tabelas dimensão específicas para Produto, Região e Ano**, permitindo segmentações eficientes e relacionamentos diretos com a tabela fato.

A modelagem foi orientada pela necessidade de **relacionamentos unidirecionais** e chaves únicas nas dimensões, como exige o Power BI para um bom desempenho e integridade nos filtros cruzados. A imagem do modelo confirma a existência das seguintes tabelas:

* **Principal**: tabela fato contendo os dados quantitativos (área, produção, valor), além de identificadores como id\_municipio, Ano e Lavoura;
* **RegioesBrasil**: dimensão geográfica criada a partir de dados oficiais do IBGE, contendo os campos de UF e Região, permitindo agregações por estado e macrorregião;
* **Grupos\_Categorias\_Produtos**: dimensão categórica que organiza os produtos agrícolas em níveis hierárquicos (Categoria, Grupo, Produto), facilitando análises por tipo de cultura (para futuras alterações).

Essa estrutura favorece a **explorabilidade do modelo**, permitindo a criação de **dashboards interativos com filtros por localização geográfica, produto e período**. A modelagem em estrela também facilita a manutenção futura e permite que usuários com diferentes níveis de conhecimento técnico consigam interpretar e utilizar os dados com facilidade.

Além disso, ao concentrar o processo de ETL no Power Query, garantiu-se maior **rastreabilidade e replicabilidade**, permitindo atualizações automatizadas e contínuas conforme novas extrações de dados forem realizadas junto ao IBGE.

Interface gráfica do usuário, Diagrama

O conteúdo gerado por IA pode estar incorreto.

Figura 1 – Star Schema do projeto

**3.4 Integração dos Dados**

A incorporação dos dados ao ambiente analítico teve início com uma abordagem programática e exploratória, com o uso de bibliotecas Python como pandas para exploração e tratamento preliminar. Em seguida, os dados organizados foram integrados ao Power BI por meio do **Power Query**.

Essa abordagem traz benefícios como:

* Melhor desempenho em consultas no Power BI;
* Clareza nas relações entre as tabelas;
* Facilidade de manutenção e atualização;
* Possibilidade de escalar e expandir o modelo à medida que novas versões das tabelas forem disponibilizadas.

O modelo entregue está, portanto, preparado para atender tanto a demandas analíticas imediatas quanto a evoluções futuras do projeto, com flexibilidade, rastreabilidade e alto grau de reutilização.

**3.5 Formatação dos Dados**

* Conversão final dos dados para CSV e XLSX para uso no Power BI.
* Ajustes de formatos numéricos, datas e categorias.
* Estrutura final em formato tabular, pronto para análise e visualização.

**4. Modeling**

**4.1 Selecionando Técnica de Modelagem**

Considerando a natureza relacional e estatística dos dados coletados no SIDRA/IBGE, a técnica de modelagem escolhida foi o **modelo dimensional em estrela (Star Schema)**. Essa abordagem é amplamente adotada em projetos de Business Intelligence e se mostrou a mais adequada para a necessidade de cruzamento entre variáveis quantitativas (área, produção, valor) e atributos descritivos (produto, localidade, ano). A escolha teve como base critérios de desempenho, clareza estrutural, facilidade de atualização e compatibilidade com ferramentas de visualização como o Power BI.

**4.2 Gerar Design de Teste**

A validação do modelo foi pensada desde o início da construção da estrutura de dados. O design de teste foi baseado em verificações de integridade, relacionamento e consistência analítica. A seguir, detalha-se a construção do modelo e sua validação.

**4.2.1 Modelo de Construção**

O modelo foi construído a partir de uma tabela fato central, contendo os indicadores numéricos da produção agrícola, e tabelas dimensão para descrever os eixos analíticos: tempo, produto e território. O relacionamento entre as tabelas foi feito de forma unidirecional, garantindo performance e consistência.

As tabelas criadas foram:

* **Principal** (fato): contém métricas como área plantada, área colhida, valor da produção e produção física;
* **Grupos\_Categorias\_Produtos** (dimensão de produto): estrutura hierárquica com Categoria, Grupo e Produto;
* **RegioesBrasil** (dimensão geográfica): permite agregações por município, estado e região;
* As informações de tempo foram mantidas na tabela fato como campo referencial Ano.1.

Todos os dados foram previamente limpos, tipados e padronizados, o que permitiu a integração direta no Power BI com mínima necessidade de transformação adicional no Power Query.

**4.3 Avaliar o Modelo**

A avaliação do modelo considerou três pilares:

* **Estrutural**: Todos os relacionamentos entre tabelas foram validados e testados com filtros cruzados e segmentações, confirmando a integridade do modelo.
* **Analítico**: Foram criados KPIs e visualizações de teste no Power BI para assegurar que os indicadores (como produtividade, concentração regional e valor por cultura) estavam sendo calculados corretamente.
* **Operacional**: O processo de atualização foi simulado com substituição de arquivos .xlsx, confirmando que o modelo reage corretamente sem necessidade de reprocessamento manual, o que valida sua escalabilidade e reutilização.

Essa avaliação demonstrou que o modelo é robusto, bem estruturado e adequado para suportar análises estratégicas do setor agropecuário com base nos dados do SIDRA/IBGE.

**5. Evaluation**

**5.1 Avaliação dos Resultados**

Os resultados obtidos com o modelo desenvolvido atenderam plenamente aos objetivos propostos. A estruturação dos dados em um modelo dimensional permitiu análises claras e consistentes sobre a produção agrícola brasileira, com filtros por período, produto e localização. O painel no Power BI apresentou ótima performance mesmo com grande volume de dados, oferecendo respostas rápidas e visualizações intuitivas. A criação de KPIs e gráficos facilitou a identificação de tendências, gargalos e oportunidades, validando a eficáciado modelo como ferramenta de suporte à decisão.

**5.2 Revisão do Processo**

Ao longo do projeto, todas as etapas do processo — da coleta à visualização — foram conduzidas com atenção à qualidade, reprodutibilidade e escalabilidade. A limpeza dos dados foi feita de forma criteriosa, documentada e replicável, assegurando transparência no tratamento. A modelagem seguiu boas práticas de BI, com a definição clara de fatos e dimensões. A escolha por centralizar o ETL no Power Query mostrou-se acertada, garantindo facilidade de manutenção e atualização do modelo.

Além disso, a condução do projeto demonstrou maturidade técnica na utilização de ferramentas como Python, Power BI e recursos do próprio SIDRA/IBGE. A combinação entre análise exploratória inicial, entendimento das variáveis e organização relacional dos dados reforça o compromisso com a entrega de valor real e aplicável.

**5.3 Determinação dos Próximos Passos**

Com a base estruturada e o modelo dimensional consolidado, o projeto está pronto para evoluir em diferentes frentes. Entre os próximos passos, destacam-se:

* Automatização da coleta de dados: por meio do uso avançado da biblioteca sidrapy e scripts Python agendados (via cron ou serviços em nuvem), é possível automatizar a extração e atualização contínua dos dados diretamente do SIDRA, eliminando a necessidade de extrações manuais e garantindo dados sempre atualizados nos dashboards.
* Criação de uma API própria: o modelo pode ser estendido com o desenvolvimento de uma API em Python (usando frameworks como FastAPI ou Flask), permitindo que outros sistemas ou usuários consultem os dados de forma dinâmica e integrada.
* Expansão da base analítica: novas tabelas do IBGE podem ser incorporadas ao projeto, como aquelas ligadas a uso da terra, clima ou dados econômicos, ampliando a abrangência analítica.
* Publicação e compartilhamento: há potencial para publicar o dashboard em ambientes de BI corporativo ou plataformas web (como Power BI Service), permitindo o uso por áreas estratégicas, instituições públicas ou mesmo para fins acadêmicos.
* Inclusão de análises preditivas: com a base consolidada, abrem-se oportunidades para aplicação de modelos estatísticos ou de machine learning que possam prever tendências de produção ou flutuações regionais.

Essas perspectivas demonstram que o trabalho realizado vai além de um exercício técnico — trata-se de uma solução concreta, com potencial de impacto real em ambientes corporativos, institucionais e educacionais.

**6. Deployment**

A entrega deste projeto foi cuidadosamente estruturada para garantir que o resultado final seja uma solução analítica integrada, eficiente e duradoura, evitando a dispersão de arquivos. Para isso, estabelecemos diretrizes claras de organização de diretórios, controle de versões, documentação detalhada e registro sistemático das decisões tomadas durante o desenvolvimento.

Todo o material foi organizado e centralizado no Google Drive da equipe, com nomenclaturas padronizadas e uma hierarquia de pastas segmentada conforme as fases do projeto: Extração, Transformação, Visualização e Documentação. Essa abordagem facilita a manutenção e atualização contínua do modelo, além de assegurar transparência e agilidade para utilização interna ou compartilhamento com futuros parceiros da Becomex.

Com a modelagem concluída e validada, iniciou-se a etapa de implantação do projeto. Neste momento, o foco está em entregar o modelo funcional, garantir que o usuário final compreenda seu uso e definir estratégias para atualização e monitoramento da solução. O modelo está pronto para ser utilizado por diferentes perfis de usuários e adaptado para aplicações futuras, inclusive com potencial de automação e integração a sistemas mais amplos.

**6.1 Plano de Implementação**

A implantação do modelo foi desenhada para garantir simplicidade operacional, rastreabilidade e escalabilidade. A base limpa, modelada e validada foi incorporada ao Power BI com um modelo dimensional bem estruturado, facilitando sua reutilização e interpretação.

**Etapas principais do plano de implementação:**

* Entrega do dashboard e base de dados estruturada: O modelo já está operando com filtros, segmentações e KPIs ativos no Power BI.
* Documentação técnica e de uso: Foi elaborada uma documentação clara, com descrições dos campos, relacionamentos e orientações para manutenção da base.
* Atualização manual assistida: O processo atual permite atualizar os dados substituindo o arquivo .xlsx com novos dados extraídos do SIDRA, mantendo a estrutura e visualizações intactas.
* Capacitação de usuários: O material entregue permite que perfis não técnicos entendam e utilizem o modelo com facilidade.

**Possibilidades de expansão e planos alternativos:**

* Automatização total do processo de coleta e atualização: Por meio de scripts Python utilizando a biblioteca sidrapy, é possível agendar coletas automáticas e integrar diretamente os dados ao Power BI com ferramentas como Power Automate, Azure ou serviços de agendamento.
* Criação de API RESTful: Para permitir consumo externo dos dados estruturados e promover integração com sistemas internos da organização.
* Publicação em Power BI Service: Facilitando o acesso remoto, controle de permissões e atualização automática por parte de stakeholders.

**Riscos mapeados e ações mitigadoras:**

* Erros na atualização manual da base podem ser evitados com um checklist de validação antes da substituição do arquivo.
* Mudanças futuras na estrutura do SIDRA exigirão revisão nos scripts ou no modelo, o que está previsto na documentação.

**6.2 Plano de Monitoramento e Manutenção**

A estratégia de manutenção prevê monitoramento contínuo e ciclos de atualização programados, com foco na estabilidade e evolução da solução.

Ações de monitoramento:

* Verificação periódica da estrutura das tabelas no SIDRA para identificar possíveis alterações;
* Checagem de indicadores-chave (produção, valor, área) após atualizações;
* Validação de relacionamento entre tabelas após cada substituição de dados.

**Critérios para reavaliação do modelo:**

* Inclusão de novos anos ou variáveis relevantes no SIDRA;
* Redução de precisão nos indicadores ou inconsistências detectadas nas visualizações;
* Mudança nas necessidades do usuário final (ex: foco em exportações, clima ou produtividade regional).

**Planos de atualização:**

* Atualizações manuais com substituição de arquivos seguirão como padrão no curto prazo;
* Em médio prazo, a automatização da coleta via API do SIDRA será avaliada;
* Um novo ciclo de mineração poderá ser iniciado caso o domínio de aplicação mude significativamente.

**6.3 Relatório Final**

Este relatório resume todo o projeto, desde a compreensão do problema até a entrega final da solução, com reflexões sobre os aprendizados, resultados e possibilidades futuras.

**Resumo do projeto:**

* Contexto de negócio: Tornar os dados públicos do setor agro, oriundos do SIDRA/IBGE, acessíveis, organizados e úteis para análise estratégica;
* Objetivo: Criar uma base tratada, confiável e visualmente explorável que permita responder a perguntas sobre produção agrícola nacional;
* Mineração de dados: Coleta automatizada e manual, limpeza criteriosa, estruturação em modelo dimensional e construção de dashboards no Power BI;
* Resultados: Modelo robusto, segmentações funcionais, KPIs claros e estrutura pronta para expansão e integração;
* Avaliação: Indicadores funcionando corretamente, estrutura validada tecnicamente, base escalável e replicável;
* Implantação: Processo simples e eficiente, com possibilidades reais de automação e publicação;
* Análise de custo-benefício: Baixo custo tecnológico (uso de ferramentas gratuitas e dados públicos), alto valor agregado para análises estratégicas e operacionais.

**Conclusões para o negócio:**

* A solução facilita a tomada de decisão em contextos públicos e privados, promove o uso de dados abertos e contribui com a democratização da inteligência analítica no setor agro.

**Conclusões para mineração de dados futura:**

* A experiência demonstrou que projetos com dados públicos podem ser altamente valiosos quando tratados com rigor técnico, e que a combinação de Python + Power BI oferece uma base sólida para replicação em outros domínios.

**6.4 Revisão do Projeto**

Ao final do projeto, foram registradas diversas lições importantes que devem ser consideradas em iniciativas futuras semelhantes.

**Pontos fortes do projeto:**

* Clareza na definição do escopo e objetivos;
* Escolha técnica acertada (modelo dimensional e ferramentas utilizadas);
* Documentação contínua e estruturada;
* Flexibilidade do modelo para atender diferentes perfis de usuários.

**Desafios enfrentados:**

* Padronização dos dados do SIDRA exigiu atenção às notas metodológicas e tratamentos manuais;
* Coleta manual inicial, embora necessária, apresentou risco de inconsistência e será substituída por processo automatizado.

**Lições aprendidas:**

* A leitura das simbologias do IBGE e o entendimento da estrutura das tabelas são essenciais para qualquer projeto com dados públicos;
* O uso do Power Query como hub do ETL simplificou o fluxo e aumentou a acessibilidade da solução.

**Melhorias futuras:**

* Automatizar completamente a coleta e atualização;
* Criar indicadores preditivos com base em séries históricas;
* Incorporar novas fontes públicas para análises complementares.

As inteligências artificiais como ChatGPT (OpenAI), Claude (Anthropic) e Gamma (Nvidia) foram utilizadas desde o início do projeto com o intuito de apoiar na estruturação e produtividade. Além disso, essas IAS proporcionaram maior agilidade na análise de dados e contribuíram para a geração de insights mais precisos, elevando a qualidade e a eficiência do desenvolvimento do projeto.

A avaliação final do projeto evidencia que dados públicos, quando adequadamente tratados com metodologia estruturada e foco analítico, transformam-se em poderosas ferramentas para a tomada de decisão. Esses insights são valiosos não só para o setor privado, mas também para governos, entidades sociais e instituições acadêmicas.

**7. Referências e Recursos Utilizados**

**Bases de Dados e Plataformas Oficiais**

* **SIDRA/IBGE – Sistema IBGE de Recuperação Automática**<https://sidra.ibge.gov.br>
* **Becomex – Consultoria Tributária e Aduaneira (Sponsor do Projeto)**<https://becomex.com.br>
* **IBGE - Divisão Territorial Brasileira**<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/divisao-regional/23701-divisao-territorial-brasileira.html>

**Repositórios do Projeto**

* **GitHub – Repositório do projeto -**[MarcioLuizBR/Projeto\_Becomex\_AnaliseDados\_IBGE: Projeto de analise de Dados sobre tabelas disponibilizadas pelo IBGE do Brasil sobre as lavouras agriculas do periodo de 2021 à 2023](https://github.com/MarcioLuizBR/Projeto_Becomex_AnaliseDados_IBGE/tree/main)
* **Google Drive – Diretório do Projeto (Documentos e Bases)**<https://drive.google.com/drive/folders/1FPoNatRQ0vG1216wUOK88w0yFwnFKofD?usp=sharing>
* **Google Colab – Notebooks de ETL, Limpeza e Modelagem**<https://colab.research.google.com/drive/1WrE-Ji9j1fsmHWJRNwrJfYF-JEdlXMQ8>

**Metodologia**

* **CRISP-DM – Metodologia de Mineração de Dados**<https://www.ibm.com/docs/pt-br/SS3RA7_18.5.0/nl/pt/BR/pdf/ModelerCRISPDM.pdf> [https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf](https://web.archive.org/web/20220401041957/https:/www.the-modeling-agency.com/crisp-dm.pdf)

**Inteligência Artificial**

**ChatGPT – OpenAI Copilot para Suporte Técnico e Redação Acadêmica** Utilizado para suporte na engenharia de prompts, revisão textual, organização de etapas, estruturação narrativa, formatação ABNT

**Claude – Anthropic Copilot para Suporte Técnico**

Utilizado para suporte na testes de consistência lógica dos dados e conferencias dos insights gerados

**Gamma – Nvidia para Suporte na criação de apresentações**

Utilizado para suporte em criação de apresentações