Pontifícia Universidade Católica do Rio de Janeiro
Departamento de Informática - Pós-Graduação em Data Science & Analytics
Disciplina: Engenharia de Dados
Autoavaliação - MVP
Aluno: João Felipe Maia Barbosa
Data: Dezembro/2025

Ao longo do desenvolvimento deste MVP, pude aplicar de forma prática e integrada os principais conceitos estudados no módulo de Engenharia de Dados, percorrendo todas as etapas de um pipeline completo (da ingestão bruta ao tratamento, modelagem, validação e análise orientada ao negócio). A execução do trabalho permitiu consolidar conhecimentos técnicos fundamentais e fortalecer minha capacidade de conduzir projetos de dados de ponta a ponta.

Inicialmente, realizei a ingestão dos datasets do Inside Airbnb e a criação da camada Bronze. Essa etapa me ajudou a entender, na prática, a importância de preservar dados na forma mais próxima possível de sua origem, garantindo rastreabilidade e segurança para as transformações subsequentes. Em seguida, apliquei limpeza, padronização e enriquecimento dos dados na camada Silver, eliminando inconsistências e preparando uma base sólida e coerente para modelagem analítica. Foi nessa fase que exercitei, com maior profundidade, boas práticas de qualidade de dados e decisões de normalização. Além disso, destaco o desafio técnico de contornar limitações do ambiente Databricks Community (Serverless) e a conversão de dados geoespaciais complexos (GeoJSON) para formatos tabulares, o que exigiu o desenvolvimento de scripts de pré-processamento específicos**.**

A modelagem dimensional permitiu consolidar todo o conteúdo da disciplina. A escolha de um esquema estrela com tabelas fato e dimensão, juntamente com a definição clara dos grãos e das chaves estrangeiras, mostrou-se adequada ao tipo de análise prevista. Esse processo exigiu validação rigorosa da integridade referencial, verificação de granularidade e reflexão sobre as métricas mais adequadas para representar corretamente o fenômeno estudado. Concluída a modelagem, construí a camada Gold, que serviu como base confiável e estruturada para as análises exploratórias.

Na etapa de validação da qualidade dos dados, pude comprovar a consistência das tabelas Gold. Verifiquei completude, coerência lógica e conformidade das regras de negócio, identificando limitações inerentes ao dataset do Inside Airbnb (como a ausência de preço diário para determinadas datas e a presença de valores estimados para ocupação annual) mas confirmando que, apesar disso, a base estava adequada para análise e tomada de decisão.

A fase final envolveu análises aprofundadas orientadas por perguntas reais de negócio. Aqui apliquei técnicas estatísticas, segmentações por percentis, comparações entre cidades, análises de correlação, decomposição de comportamento por tipo de imóvel, capacidade, reputação, profissionalização dos hosts e localização geográfica. O processo permitiu extrair insights concretos sobre o funcionamento do mercado de hospedagem em Rio de Janeiro e New York City, revelando padrões de performance, competitividade e dinâmica espacial. Foi também

a etapa em que desenvolvi visualizações mais avançadas, incluindo mapas geoespaciais baseados em polígonos de bairro.

Do ponto de vista técnico, o MVP demonstrou que fui capaz de implementar um pipeline completo e robusto, documentando cada decisão e mantendo consistência metodológica. Do ponto de vista analítico, consegui interpretar corretamente os resultados, conectando as evidências dos dados às hipóteses de negócio. Ao mesmo tempo, tive a oportunidade de refletir sobre limitações da base, como estimativas de ocupação e receita, concentração de outliers e lacunas em colunas específicas. Essas reflexões fortalecem a maturidade necessária para lidar com dados reais, que raramente são perfeitos.

De forma geral, considero que o MVP foi bem executado. Consegui aplicar o conteúdo aprendido, escrever código limpo e replicável, construir um modelo dimensional coerente, validar a qualidade dos dados e responder às perguntas de negócio de maneira estruturada e fundamentada. Essa experiência reforçou meu domínio sobre ETL, SQL, PySpark, modelagem dimensional e análise exploratória, além de aprimorar minha capacidade de comunicação analítica ao longo do relatório. Finalizo esta etapa confiante de que evoluí tecnicamente e academicamente, demonstrando clareza, rigor metodológico e visão de negócio na execução do projeto.