**Documentação do Projeto**

**HousePriceProject**

**1. Escopo do Projeto**

Definir escopo do problema: modelagem em dados em painel ou em séries temporais?

Inicia-se com a proposta de modelar em painel. Ampliaremos para séries temporais apenas caso seja possível coletar e construir séries históricas dentro do prazo de entrega.

Projeto de Referência: <https://www.agenteimovel.com.br/mercado-imobiliario/a-venda/df/brasilia/>

**2. Alinhamentos**

* Vamos predizer apenas o preço de vendas ou também o preço do aluguel?

Ambos.

* Definir quais variáveis coletar:

O mínimo de variáveis está definido no Dfimóveis (abrindo espaço para coleta de outras variáveis para discussão)

* Definir de quais sites raspar os dados (ou melhor, em quais aplicar web scraping nas variáveis definidas anteriormete) :

Dfimoveis: <https://www.dfimoveis.com.br/venda/df/todos/imoveis>

QuintoAndar: <https://www.quintoandar.com.br/alugar/imovel/sao-paulo-sp-brasil>

Wimoveis: <https://www.wimoveis.com.br/venda/imoveis/df/brasilia>

Vivareal: <https://www.vivareal.com.br/aluguel/distrito-federal/brasilia/>

ZapImoveis: <https://www.zapimoveis.com.br/>

**3. Pré-requisitos de Infraestrutura e Engenharia de Dados**

Primeira parte do pipeline é a obtenção dos dados via web scraping: extrair, organizar e guarda em um xlsx. Ou seja, o ETL. Note que não é um processo automatizado real-time, pois o script precisa ser acionado manualmente e os dados não estão sendo armazenados em um Sistema Gerenciador de Banco de Dados (SGBD) via SQL (padrão tabular). Mas sim localmente em um arquivo excel.

**Precisa ser desenvolvido um pipe automatizado:** o processo do pipeline deve ser extrair via web scraping, armazenar em um SGBD, automatizar o script de coleta com algum orquestrador – possivelmente o Apache Airflow (deixando o script rodar 1 vez ao mês, por exemplo. Assim, atualiza os dados evitando defasagem, perda de informações sazonais ou viés amostral). Por fim, há o pré-processamento de dados: organização, limpeza, filtragem e compatibilização entre as diversas fontes de coleta.

**4. Atribuições**

**4.1. Membros responsáveis por cada scraping:**

DfImoveis → Luiz Paulo

QuintoAndar → (trabalho em equipe)

Wimoveis → Fabio

Vivareal → Felipe Tavares

ZapImoveis → Dudu

Há duas regras para o web scraping:

Web Scraping precisam ser automatizadas;

Padronização dos dicionários (por enquanto salvar em xlsx).

**4.2. Membros responsáveis pela engenhria de dados:**

**Responsável pela engenharia de dados: Caetano Fleury**

(Web Scraping → SGBD [NoSQL] → pipeline automático [Apache Airflow]).

→ MongoDB

→ Padronização do dicionário derivado do scraping (Definir posteriormente)

→ Desafios futuras com a cloud (deploy).

**→ Decidir se os dados serão pré-processados antes ou não.**