Roteiro geral de analises, modelagem e dashboards

> Documento complementar aos guias existentes em `docs/`. Objetivo: orientar colegas sobre a sequencia minima para atualizar dados, executar as analises/modelos e abastecer dashboards sem se perder na estrutura Bronze/Silver/Gold.

1. Visao rapida do projeto

- Escopo: seis municipios da RMB (Belem, Ananindeua, Marituba, Benevides, Santa Barbara do Para, Santa Izabel do Para) no periodo 2018-2025.
- Hipotese central: melhor servico (cobertura, eficiencia), boa qualidade da agua e financas equilibradas reduzem internacoes por agravos hidricos, moduladas por clima.
- Entregaveis: base Gold anual (`data/gold/gold_features_ano.*`), notebooks de EDA/modelagem (pasta `notebooks/`), dashboard no Looker Studio com artefatos exportados em `dashboard/`.

2. Estrutura de dados atualizada

Camada	Uso	Diretorios principais	Formato	Observacoes
Bronze	Dados brutos como baixados	`data/bronze/sih`, `data/bronze/sih/csv`, `data/bronze/sisagua`, `data/bronze/siops`, `data/bronze/inmet`, `data/bronze/ibge`, `data/bronze/snis`	DBC, CSV, ZIP	Manter apenas anos/fonte relevantes.
Silver	Dados padronizados (chaves, tipos)	`data/silver/sih`, `data/silver/sisagua`, `data/silver/siops`, `data/silver/inmet`, `data/silver/ibge_populacao`	Parquet	Colunas chaves: `cod_mun`, `municipio`, `ano`.
Gold	Conjuntos prontos para analise/dash	`data/gold/gold_features_ano.*`, `data/gold/snis_rmb_indicadores_v2.*`, `data/gold/gold_qualidade_agua.*`	CSV, Parquet	Fonte oficial para notebooks, modelos e dashboard.
Dashboards	Entregáveis finais e artefatos de apresentação	`dashboard/`	PDFs, imagens, links	Guardar exports do Looker/PowerPoint para referência do grupo.

> Referencia cruzada: `README.md` (secao 4) traz detalhes extras e deve ser atualizado se novos arquivos surgirem.

3. Passo a passo essencial

3.1 Preparacao do ambiente

1. Ativar a virtualenv (`source .venv/bin/activate`).
2. Instalar dependencias se necessario: `pip install -r requirements.txt`.
3. Garantir que PyArrow e Pandas estao presentes (requeridos para Parquet).

3.2 Confirmar camada Bronze

- `data/bronze/sih`: arquivos `RDPAyymm.dbc` por mes (2018-2025).
- `data/bronze/sih/csv`: CSVs convertidos a partir dos DBCs (um por mes) quando precisar recalcular Silver.
- `data/bronze/sisagua`: pacotes `.csv.zip` dos controles mensais 2018-2025.
- `data/bronze/inmet`: pastas por ano com CSVs `INMET_N_PA_*` (estacoes A201, A202, A227).
- `data/bronze/siops`: arquivos consolidados `siops_indicadores_rmb_*.csv`.
- `data/bronze/ibge`: `sidra6579_pop_aaaa.csv`.
- `data/bronze/snis`: planilhas anuais 2018-2023.

3.3 Gerar camada Silver (quando houver atualizacao)

Fonte	Script	Comando base	Saida
SNIS (servico)	`scripts/etl_snis_indicadores_rmb.py`	`python scripts/etl_snis_indicadores_rmb.py`	CSV/Parquet direto em Gold (ja aplicado).
SISAGUA	`scripts/bronze_to_silver_sisagua_parquet.py`	`python scripts/bronze_to_silver_sisagua_parquet.py --input-dir data/bronze/sisagua`	Parquet particionado em `data/silver/sisagua`.
SIH	`scripts/bronze_to_silver_sih_parquet.py`	`python scripts/bronze_to_silver_sih_parquet.py --csv-dir data/bronze/sih/csv`	Parquet particionado em `data/silver/sih` (ano/mes).
SIOPS	`scripts/bronze_to_silver_siops_parquet.py`	`python scripts/bronze_to_silver_siops_parquet.py --input data/bronze/siops/siops_indicadores_rmb_2018_2025.csv`	`data/silver/siops/indicadores.parquet`.
INMET	`scripts/inmet_to_parquet.py`	`python scripts/inmet_to_parquet.py --input-dir data/bronze/inmet`	Dataset particionado em `data/silver/inmet`.
IBGE (pop)	`scripts/bronze_to_silver_ibge_pop_parquet.py`	`python scripts/bronze_to_silver_ibge_pop_parquet.py --input-dir data/bronze/ibge`	`data/silver/ibge_populacao/populacao.parquet`.

3.4 Montar camada Gold

1. SNIS v2: `scripts/fix_snis_csv.py` (gera CSV/Parquet corrigidos em `data/gold/`).
2. Qualidade da agua: usar notebook ou script auxiliar (ver `docs/Guia de Preparacao de Dados – ...`) para agregacoes e salvar `data/gold/gold_qualidade_agua.*`.
3. Merge Silver -> Gold final: `python scripts/silver_to_gold_features.py` (usa `config/rmb_municipios.csv` e salva `data/gold/gold_features_ano.*`).
4. Se precisar revisitar versao anterior do ETL SNIS, `scripts/fix_snis_scale_and_ibge.py` continua disponivel.

4. Orientacoes de uso dos dados

- **Chaves padrao**: sempre trabalhar com `cod_mun` (7 digitos), `municipio` (uppercase) e `ano`. Conferir se ha 6 municipios x n anos.
- **Normalizacao populacional**: utilizar `data/silver/ibge_populacao/populacao.parquet` para gerar indicadores per capita ou por 10k hab.
- **Qualidade da agua**: `gold_qualidade_agua` traz tanto contagem de amostras quanto percentuais por parametro (`pct_conformes_*`). Para dashboards, priorizar percentuais; para modelos, usar contagens como peso.
- **SNIS v2**: ja corrigido (percentuais >100 convertidos) e possui colunas `idx_*` padronizadas. Usar essa versao, nao a raw.
- **SIOPS**: indicadores em percentual estao em escala 0-100; despesas per capita em R$ constantes (ano corrente). Validar se ha anos faltantes ao cruzar com Gold.
- **Clima**: dados agregados por estacao sao replicados aos municipios conforme mapeamento (`A201` -> Belem + entorno, `A202` -> Castanhal + Sta Izabel + Sta Barbara, `A227` -> proxy Barcarena). Ajustar se novos municipios forem incluidos.
- **Saude (SIH)**: `internacoes_total` e `internacoes_hidricas` devem ser agregados antes de calcular taxas (`internacoes_hidricas_10k`). Sempre verificar se CID filtrado continua coerente.

5. Roteiro de analises e modelagem

1. **EDA servico e qualidade**: notebook `01_eda_servicos.ipynb` utilizando `data/gold/snis_rmb_indicadores_v2.*` e `data/gold/gold_qualidade_agua.*`.
2. **EDA saude e clima**: notebook `02_eda_saude_clima.ipynb` combinando `data/gold/gold_features_ano.*` (foco em clima + internacoes).
3. **Modelagem (target `internacoes_10k_hab`)**: notebook `03_modelagem.ipynb` carrega `gold_features_ano`.
   - Modelos: Regressao Linear, Lasso, RandomForestRegressor.
   - Metricas: MAE, RMSE, R2 (k-fold). Guardar resultados em `reports/` se atualizar.
   - Explicabilidade: permutation importance + SHAP (opcional).
4. **Classificacao opcional**: binarizar nao conformidade (p.ex. `pct_conformes_global < 95`) e testar Logistic vs RandomForestClassifier.

5.1 Cobertura das perguntas de negocio

Pergunta (doc MVP)	Entregavel/Dado que responde	Acoes previstas
1. Quais acoes reduzem mais as internacoes?	Notebook `03_modelagem.ipynb` (feature importance e analise de elasticidade) + aba "Priorizar & Simular" do dashboard	Rodar modelos explicaveis, calcular efeitos marginais (`what-if`) e documentar insights na narrativa do dashboard.
2. Qual retorno por municipio/sistema?	`gold_features_ano` + calculo "ganho por ponto" nas notebooks e tabelas do dashboard	Criar tabela ranqueando municipios pelo impacto projetado; publicar no dashboard e no relatório final.
3. Como a chuva modula o risco?	Notebook `02_eda_saude_clima.ipynb` (series A00–A09 x chuva com defasagens) + cards no dashboard	Incluir analises de correlação/lag e visualizações sazonais (linha + heatmap).
4. Qual ranking de priorizacao (12 meses)?	`03_modelagem.ipynb` + dashboard (painel Priorizar)	Gerar score composto e ordenação; exportar versão para `dashboard/`.
5. Como medir ODS 6/3/11?	ODS Tracker no dashboard + indicadores em `gold_features_ano`	Garantir seção dedicada no dashboard com metas e status; anexar ao relatório.

6. Dashboards (Looker Studio)

- **Fonte principal**: `data/gold/gold_features_ano.csv`.
- **Complementos**: `data/gold/gold_qualidade_agua.csv` para visoes mensais, `data/gold/snis_rmb_indicadores_v2.csv` para detalhamento de indicadores.
- **Paginas sugeridas**:
  1. Visao RMB (KPI cards + ranking + series temporais).
  2. Visao municipal (cartoes por eixo, barras de nao conformidade, explicabilidade do modelo).
  3. Painel de prioridades (tabela com ranking + simulador simples).
- **Boas praticas**: fixar filtros para anos e municipios; destacar meta (ex. `pct_conformes_global >= 95`).
- **Entrega**: salvar versões exportadas (PDF, PNG, links compartilháveis) dentro da pasta `dashboard/`.

6.1 Simulador de cenarios e priorizacao (secao 12 do MVP)

- **Dados de entrada:** `gold_features_ano.csv` (indicadores anuais) + elasticidades estimadas em `03_modelagem.ipynb` para `idx_atend_agua_total`, `idx_tratamento_esgoto`, `idx_perdas_distribuicao`, `pct_conformes_global` e outras variáveis relevantes.
- **Mecanismo:** utilizar os coeficientes/regressões para calcular (\Delta) previsto em `internacoes_10k_hab` quando o usuário ajusta sliders (% coleta, % tratado, perdas, conformidade).
- **Saídas:** painel com ranking atualizado (municipios/sistemas) e comparação com baseline, exportado para `dashboard/`.
- **Documentacao:** registrar metodologia e supostos nos notebooks e replicar resumo na página "Priorizar & Simular" do dashboard.

7. Checklist rapido antes de compartilhar

- [ ] Conferir contagem de linhas das bases Gold (6 municipios x anos vigentes).
- [ ] Validar escalas (percentuais 0-100, valores monetarios em R$).
- [ ] Atualizar `README.md` se novas colunas/arquivos surgirem.
- [ ] Executar notebooks limpando o kernel para garantir reprodutibilidade.
- [ ] Exportar dashboard atualizando a fonte de dados, se necessario, e arquivar em `dashboard/`.
- [ ] Validar que as perguntas de negocio (doc MVP, secao 7) estao respondidas em notebooks/dashboards e documentar conclusoes.

8. Recursos adicionais

- `docs/Guia de Preparacao de Dados – ... .md`: detalhes de cada fonte e ETL.
- `docs/Catalogo De Dados – RMB ...`: glossario completo (versao DOC/PDF).
- `README.md`: visao geral do repositorio e dicionario resumido.
- Contato: registrar duvidas via issue no repositorio ou canal Teams do grupo.

Mantemos este roteiro como documento vivo. Ao adicionar novas fontes, modelos ou visualizacoes, registrar a mudanca aqui, no README e nos guias correspondentes para garantir alinhamento da equipe.