Lab #03
Uso de Registros Administrativos e Sistemas Públicos de Informação

Ricardo Ceneviva
Métodos Computacionais Aplicados em Políticas Públicas
Universidade Federal do ABC

6 de julho de 2025

Instruções gerais:

- Trabalhar **em duplas**. Escolham *R* ou *Python* e mantenham a mesma linguagem em todos os exercícios.
- Entreguem um arquivo `.Rmd` ou `.ipynb` com todo o código, respostas e comentários.
- Comentem o código: expliquem cada passo, decisões de limpeza e interpretações de resultados.
- Prazo de entrega: *até as 23h59 do dia seguinte à aula de laboratório*. Submeter via repositório Git da disciplina.

Objetivos didáticos

- Compreender, em ambiente controlado, conceitos de limpeza, *record linkage*, anonimização e avaliação de qualidade.
- Adquirir prática de acesso reproduzível a bases públicas por **API** (Base dos Dados) e construção de endpoints simples.
- Discutir trade-offs entre qualidade dos dados, privacidade e utilidade analítica.

1 Aquecimento com dados simulados

Exercício 1.1 – Limpeza e padronização

1.1 Gere dois data frames simulados chamados `alunos_escolaA` e `alunos_escolaB`, cada um com 2 000 registros e as variáveis abaixo:

- `cpf` (11 dígitos, 5% com dígitos aleatórios trocados).
- `nome` (introduzir variações de acentuação e caixa).
- `data_nasc` (formatos mistos: dd/mm/aaaa e aaaa-mm-dd).

- `cep` (formato com ou sem hífen; 10% faltante).

1.2 Padronize todas as colunas: retire acentos, uniformize caixa, converta datas ao padrão ISO e formate CEP como `12345678`. Registre cada transformação num objeto `log_transformacao` (data frame ou lista).

Exercício 1.2 – Linkage determinístico vs. probabilístico

1.3 Realize *merge* determinístico usando `cpf` e compute verdadeiros–positivos, falsos–positivos, falsos–negativos.

1.4 Execute linkage probabilístico (Fellegi–Sunter) com `nome`, `data_nasc` e `sexo`. Compare *precision*, *recall* e *F1* entre os dois métodos.

Exercício 1.3 – Privacy-Preserving Record Linkage (PPRL)

1.5 Crie o hash SHA-256 dos CPFs com *salt*. Repita o linkage determinístico apenas com os hashes.

1.6 Discuta o impacto sobre acurácia e riscos de ataque de força bruta.

Exercício 1.4 – Dados sintéticos

1.7 Gere uma versão sintética do data frame consolidado usando `synthpop` (R) ou `sdv` (Python).

1.8 Compare distribuições de sexo, idade e taxa de reprovação entre real e sintético.

1.9 Reflita sobre utilidade estatística versus proteção de privacidade.

2 Exercício com API e dados reais – Base dos Dados

2.1 Preparação

- Criar conta gratuita em Base dos Dados e gerar chave de acesso.
- Instalar pacotes sugeridos conforme a linguagem escolhida.

2.2 Consulta programática ao Censo Escolar

2.1 Desenvolva função que receba o código IBGE de um município e retorne, para 2018–2022, o número de matrículas no 5.º e no 6.º ano, bem como a taxa de transição 5→6.

2.2 Automatize a função para todos os municípios de um estado à escolha; identifique outliers usando a regra do IQR.

2.3 Visualize a série temporal da taxa de transição de três municípios (`ggplot2` ou `matplotlib`).

2.3 Mini-API REST

2.4 Construa endpoint `/fluxo-escolar/{ano}/{id_municipio}` devolvendo JSON com taxa de transição e evasão.

2.5 Utilize `plumber` (R) ou `fastapi` (Python); carregue a chave BD via variável de ambiente.

2.6 Teste localmente com `curl` ou Postman e documente no README.

2.4 Extensão opcional

Crie endpoint para dados do TSE (abstenção) ou SUS (internações por dengue) seguindo o mesmo padrão.

3 Discussão em Plenária

- O que mudou de percepção ao migrar de dados simulados para reais?
- Como APIs podem democratizar o acesso e a reprodutibilidade?
- Quais limitações ou gargalos técnicos foram observados?

Cronograma sugerido (2 h 30 min)

- 00:00–00:20 – Introdução conceitual
- 00:20–01:50 – Exercícios 1.1–1.4 (dados simulados)
- 01:50–02:20 – Exercício 2 (API + dados reais)
- 02:20–02:30 – Debate e entrega dos notebooks

Bons estudos e mãos à obra!