Trabalho Final – Módulo 5

A atividade final do módulo de introdução a engenharia de dados se dará utilizando os conhecimentos ministrados ao decorrer do módulo. O objetivo será pedir que cada aluno entregue um projeto com **temática diferente**, simulando uma demanda de um profissional da engenharia de dados, em menor escala.

O objetivo é que os alunos extraiam e preparem uma base de dados para um pesquisador da área da saúde, para isso precisarão extrair os dados utilizando técnicas de webscrapping, unir duas bases de dados, manipularem esses dados, fazer o tratamento e limpeza e entregar metadados para esta base. Tudo isso será avaliado através de um projeto do GitHub.

Portanto, as etapas avaliadas serão:

1. Webscrapping de dados de saúde (Processo utilizando alguma das bibliotecas ministradas no curso).

2. Projeto no GitHub (Com documentação e README).

3. Desenvolvimento de um pipeline, detalhando as etapas envolvidas desde o webscrapping até a entrega de uma base pronta para análise (ferramentas, sistema, linguagem utilizados). Pode ser em formato de texto usando fluxo de tarefas, com imagens usando um fluxograma e texto para complementar com descrição de cada tarefa.

4. Manipulação de dados e transformação (avaliados através de scripts de manipulação usando técnicas similares as ministradas em aula).

5. Agregação de dados e junção (merge) de bases de dados (também avaliado através dos scripts).

6. Descritiva simples de pelo menos 10 variáveis da base final (se for numérica, avaliar média, mediana e fazer um boxplot, se for categórica mostrar percentuais de preenchimento em cada categoria, se for data, avaliar mínimo e máximo).

7. Entregar um dicionário das principais variáveis na base de dados (entre 10-15 variáveis), contendo o nome da variável, uma descrição breve, o tipo, e a completude na base de dados.

As problemáticas disponíveis para seleção, e, portanto, o problema no qual vocês irão trabalhar estarão dispostas abaixo. Além disso, na etapa de webscrapping, precisaremos de dados das bases de saúde de **pelo menos 5 anos** (os 5 mais atuais disponíveis), e para os dados socioeconômicos se possível também dos últimos 5 anos (se for um índice único, utilizar apenas o que está disponível).

Problema 1:

O pesquisador gostaria de avaliar o impacto do bolsa família na evolução do peso das crianças cadastradas o Cadastro Único, acompanhadas através do SISVAN (Sistema de Vigilância Alimentar e Nutricional).

Ele precisa dos dados agregados por município do bolsa família e informações sobre idade, altura, peso, IMC para todas as idades mensuradas para serem agregadas por município. Para os dados de bolsa família queremos a quantidade de famílias beneficiadas, o valor repassado e o valor médio dos benefícios por município.

Dados:

- SISVAN: https://dados.gov.br/dados/conjuntos-dados/sistema-de-vigilancia-alimentar-e-nutricional---sisvan
- Bolsa família: https://dados.gov.br/dados/conjuntos-dados/bolsa-familia
- População no município: -> Tabelas em anexo -> População coletada e população imputada, por município. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/22827-censo-demografico-2022.html?edicao=37225&t=resultados

Problema 2:

O pesquisador quer avaliar as condições de nascimentos usando alguns dados socioeconômicos, sugerimos usar os dados do índice do IPS também agregado por município.

Para avaliar condições de nascimento precisamos de dados dos nascimentos, peso ao nascer, anomalias congênitas identificadas, assistência no nascimento, apgar 1 e 5 e local de nascimento. No IPS queremos o índice principal, o de Necessidades Humanas Básicas, o PIB per capita e a população do município.

Dados:

- SINASC: https://dados.gov.br/dados/conjuntos-dados/sistema-de-informacao-sobre-nascidos-vivos-sinasc-1996-a-20201
- IPS: https://ipsbrasil.org.br/pt/explore/dados

Problema 3:

O pesquisador quer avaliar as condições de óbitos usando alguns dados socioeconômicos, observando se existe alguma relação entre municípios com maior IVS (Índice de Vulnerabilidade Social) ou pior IED (Índice de Equidade e Dimensionamento) em relação à quantidade de óbitos.

Para isso ele precisará das informações sobre óbito, se o evento está relacionado ao processo de trabalho, se houve correção ou alteração da causa do óbito após investigação, o(s) CID(s) informados no atestado de óbito, a causa básica da declaração do óbito (CID 10), o código do município de residência, a data de nascimento e data de óbito, também tem interesse em saber o nível de escolaridade do indivíduo e escolaridade da mãe (se disponível) e informações sobre raça/cor do indivíduo. E para informações socioeconômicas queremos os valores de IED e IVS, além do número de habitantes e a faixa de porte populacional segundo o IBGE.

Dados:

- SIM: https://dados.gov.br/dados/conjuntos-dados/sim-1979-2019
- IED e IVS - https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-3.493-de-10-de-abril-de-2024-553573811

Dados gerais: Quaisquer dúvidas, estamos a disposição para orientá-los.