

Trabalho Final

Coleta, Preparação e Análise de Dados



Prof. Luan Garcia

Introdução

Este trabalho extensionista consiste em coleta, preparação e análise de dados governamentais abertos disponíveis através do portal DATASUS. A nota deste trabalho representa 40% da nota final na disciplina.

Esta atividade deve ser realizada em grupos de 3 até 5 pessoas. Não serão permitidas entregas individuais.

Entrega e Apresentação

Por ser uma atividade extensionista, toda produção do trabalho deve ser disponibilizada de forma aberta. Para isto, cada grupo deve ter um repositório no Github contendo seu código e/ou notebooks utilizados para o pré-processamento e análise de dados.

O que deve ser entregue:

1. Link para Github aberto contendo código e notebooks utilizados no trabalho.
2. Dashboards.
3. Relatório textual descrevendo as hipóteses de pesquisa, as bases de dados escolhidas, técnicas de integração, limpeza e transformação utilizadas, conclusões obtidas a partir das análises realizadas, além de referências utilizadas.

Data de entrega: 16/06/2024 (Moodle). Atrasos na entrega acarretarão em descontos na nota.

Data de apresentação: 17/06/2024 e 19/06/2024. A ordem de apresentação ocorrerá por sorteio.

Descrição do Trabalho

Você deve escolher uma ou mais bases de dados abertos do DATASUS, pré processá-las, e criar uma visualização para esses dados em formato de *dashboard* que dê suporte à análise de alguma questão de pesquisa a ser respondida com os dados. O portal do DATASUS pode ser acessado no link a seguir:

<https://datasus.saude.gov.br/>

Todas as bases de dados disponíveis podem ser encontradas no portal, e seu download pode ser realizado através do link abaixo, grande maioria em formato DBC:

<https://datasus.saude.gov.br/transferencia-de-arquivos/>

Também é possível acessar as bases através de um protocolo FTP:

<ftp://ftp.datasus.gov.br/dissemin/publicos/>

Ou utilizando TABNET, via web, que entrega os dados em formato tabular:

Trabalho Final

Coleta, Preparação e Análise de Dados



Prof. Luan Garcia

<https://datasus.saude.gov.br/informacoes-de-saude-tabnet/>

Além disso, existem algumas bases disponibilizadas no repositório openDataSus:

<https://opendatasus.saude.gov.br/dataset/>

Há também algumas bibliotecas disponíveis na internet que criam uma interface para download e pré-processamento dos dados, como o PySUS, uma biblioteca em Python para lidar com dados do DATASUS. Neste caso, não são todas as bases que estão disponíveis para acesso. Abaixo segue uma lista dos sistemas os quais podem ser acessados pelo PySUS:

- **SINAN**: Sistema de Informação de Agravos de Notificação.
- **SINASC**: Sistema de Informação sobre Nascidos Vivos.
- **SIM**: Sistema de Informação sobre Mortalidade.
- **SIH**: Sistema de Informações Hospitalares
- **SIA**: Sistema de Informações Ambulatoriais

Mais informações sobre o PySUS podem ser encontradas no seu repositório:

<https://github.com/AlertaDengue/PySUS>

O uso do PySUS não é obrigatório, todo trabalho pode ser realizado realizando o download das bases diretamente do portal DATASUS, via o portal de transferências ou FTP.

Atenção: Os dados crus em sua grande maioria estão disponibilizados em formato .DBC, um formato próprio do Ministério da Saúde. Este não é um formato que pode ser lido de maneira fácil e em geral é preciso converter para formatos mais amigáveis, como .DBF, .CSV ou Parquet. O próprio ministério da saúde disponibiliza um sistema para tabulação dos dados .DBC e posterior conversão para .DBF. Este sistema chama-se TabWin e é de uso gratuito.

Além disso, também existem algumas bibliotecas para Python e para R que convertem arquivos .DBC em outros formatos, como por exemplo *pyreaddbc* (utilizada pelo próprio PySUS), *datasus-dbc-py*, e *microdatasus* e *read.dbc* (ambos para linguagem R).

Ferramentas Sugeridas

Sugere-se o uso das ferramentas a seguir:

- Bibliotecas pandas, scikit-learn e seaborn para o processamento das bases em Python.
- PySUS para acesso a algumas das bases diretamente em Python.
- pyreaddbc para bases de dados não disponibilizadas diretamente pelo PySUS.
- Tableau ou PowerBI para os dashboards.

Trabalho Final

Coleta, Preparação e Análise de Dados



Prof. Luan Garcia

Sugestões de Trabalhos

Seu grupo é livre para escolher questões de pesquisa e bases de interesse no DATASUS. Entretanto, segue abaixo algumas sugestões de trabalhos.

- Com a base SINAN, investigar quais regiões do Brasil possuem mais casos de picadas por animais peçonhentos.
- Com a base SINAN, investigar a evolução temporal de casos de dengue nas diferentes regiões do país.
- Com a base SINASC e dados do CENSO, investigar se existe alguma relação entre tipos de partos (natural ou cesariana) e renda familiar, escolaridade, região do país, etc.
- Com a base SIM e base E-SUS, investigar se existe alguma modificação no número de mortes antes, durante e depois da pandemia, e se existe alguma relação com a vacinação para COVID.

Recursos Interessantes

Abaixo, alguns recursos complementares interessantes.

- Documentação PySUS:
<https://pysus.readthedocs.io/en/latest/>
- Curso de PySUS:
<https://github.com/AlertaDengue/curso-pysus>
- Pyreaddbc, biblioteca para converter arquivos DBC em python:
<https://github.com/osl-incubator/pyreaddbc>
- Datasus-dbc-py, biblioteca para converter arquivos DBC em python.
<https://github.com/mymatsubara/datasus-dbc-py>
- Artigo sobre como obter e tratar Dados do Datasus:
<https://dadosabertos.social/t/como-obter-e-tratar-dados-do-datasus>
- Artigo no médium sobre download e pré-processamento de dados sobre violência do SINAN:
<https://medium.com/psicodata/download-e-pr%C3%A9-processamento-de-dados-do-sinan-datasus-sobre-viol%C3%Aancia-com-python-ddf46ff9a3f5>
- Artigo no médium sobre investigações de suicídios no Brasil utilizando a base SIM:
<https://medium.com/psicodata/baixando-e-processando-dados-do-datasus-sobre-suic%C3%ADdio-com-python-656afa17f6ad>
- Repositório com diversas análises em notebooks sobre COVID-19:
https://github.com/ceciliasilvads/analises_covid19
- Microdatasus: pacote em R para interface com bases do DATASUS.
<https://github.com/rfsaldanha/microdatasus>
- read.dbc: pacote em R para leitura de arquivos .DBC
<https://github.com/danicat/read.dbc>