Projeto Integrador de Ciência de Dados e Inteligência Artificial III

Coleta, Tratamento e Análise de Dados do DATASUS

Relatório de Matheus Lima Ribeiro

Orientador:
Prof. Sérgio da Costa Côrtes

2 de maio de 2024

Introduction

Este projeto foi concebido como parte do curso de Projeto Integrador de Ciência de Dados e Inteligência Artificial III no IESB. O principal objetivo era desenvolver um sistema robusto para a coleta, tratamento e análise de dados de internações hospitalares do DATASUS entre 2019 e 2023. Através deste projeto, busquei compreender melhor as tendências de saúde, os padrões de doenças e os resultados de tratamentos em diferentes regiões do Brasil, fornecendo assim, dados valiosos para decisões políticas e gestão de recursos no sistema de saúde.

1 Metodologia

1.1 Captação de Dados

A primeira e segunda etapa do projeto consistiu na coleta de dados através de webscraping, utilizando a linguagem Python com as bibliotecas Selenium para extrair os dados e Pandas para manipulá-los. Os dados foram extraídos do portal DATASUS, especificamente da seção que aborda a Produção Hospitalar do SIH/SUS. Este processo envolveu a programação de scripts que automatizaram a extração de dados de várias páginas web, garantindo a captura de um conjunto de dados abrangente e representativo. `http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/qibr.def`.

1.2 Definição de Variáveis e Construção do Dicionário de Dados

Na fase de definição do nome das variáveis e construção do dicionário de dados, adotei uma abordagem metódica para garantir clareza e consistência em todo o projeto de análise de dados do DATASUS. Primeiramente, estabeleci nomes de variáveis que fossem intuitivos e descritivos, seguindo convenções que facilitassem a identificação e o entendimento de cada campo pelos usuários finais. Por exemplo, nomes como Codigo_Municipio, Quantidade_Aprovada, e Valor_Aprovado foram escolhidos para refletir diretamente o conteúdo de cada variável. Após definir os nomes, construí um dicionário de dados detalhado. Para cada variável, incluí informações como o tipo de dado (por exemplo, inteiro, decimal, texto), uma descrição detalhada explicando o que a variável representa, quaisquer valores possíveis ou restrições, e como os dados são coletados ou calculados. Esse dicionário foi então formatado em uma tabela clara e disponibilizado junto com o dataset, servindo como um guia essencial para todos os envolvidos no projeto, desde analistas de dados até stakeholders, facilitando a compreensão e o uso eficaz dos dados coletados.

1.3 Integração das Bases de Dados de Subgrupo de Procedimentos

Durante o processo de integração das bases de dados de "Subgrupo de Procedimentos", concentrei-me em combinar eficientemente informações detalhadas provenientes de diversas tabelas relacionadas à "Quantidade Aprovada"e "Valor Aprovado"para cada procedimento. Utilizei a biblioteca Pandas em Python para executar essa tarefa. Inicialmente, carreguei cada dataset em DataFrames separados, assegurando que todos os campos-chave, como códigos de procedimento e datas, estivessem formatados de maneira uniforme e padronizada para garantir uma junção precisa. Após a preparação inicial, executei operações de merge usando esses campos-chave para criar um único DataFrame integrado. Esse DataFrame combinado não apenas preservava todas as informações críticas de cada subgrupo, mas também organizava os dados de forma que análises subsequentes pudessem ser realizadas de maneira mais eficiente e intuitiva, maximizando assim a utilidade dos dados integrados para insights estratégicos e operacionais.

1.4 Verificação dos Valores Totais

Após a integração dos dados, procedi com a verificação dos valores totais para assegurar que não houvesse discrepâncias nas somas e contagens. Utilizei consultas SQL complexas para sumarizar os dados e validar contra os totais conhecidos, um passo crucial para a integridade do projeto.

1.5 Integração das Base de dados em uma só

Para integrar as bases de dados de "Quantidade Aprovada"e "Valor Aprovado"em uma única base de dados, utilizei um script em Python que aproveitou as funcionalidades da biblioteca Pandas. Primeiramente, carreguei os dois conjuntos de dados separados em dataframes do Pandas a partir de arquivos CSV. Após a carga inicial, realizei uma inspeção preliminar para entender a estrutura e os tipos de dados de cada coluna, assegurando que as chaves de união, como os códigos dos municípios ou datas, estivessem em formatos consistentes entre os dois conjuntos.

1.6 Integração com a Base de População do Censo Demográfico

Para enriquecer a análise dos dados hospitalares coletados do DATASUS, integrei os mesmos com a Base de População do Censo Demográfico de 2022, que foi fornecida pelo professor. Esta integração foi crucial para contextualizar os dados de saúde dentro das realidades geográficas e demográficas dos municípios brasileiros. Utilizei Python e a biblioteca Pandas para combinar as bases de dados, focando em atribuir a cada registro de saúde as correspondentes população, latitude e longitude dos municípios. Primeiro, garanti que ambas as bases de dados tivessem uma coluna comum de identificação municipal, normalmente o código IBGE do município, que então usei como chave para realizar um merge dos datasets. Este processo não só adicionou camadas de contexto aos dados de saúde, permitindo análises mais profundas sobre acessibilidade e necessidades de saúde por região, mas também facilitou visualizações geográficas dos dados, essenciais para a identificação de padrões espaciais nas tendências de saúde.

1.7 Geração de um Único Arquivo com Todos os Dados e Variáveis

Para consolidar efetivamente as diversas fontes de dados utilizadas no projeto de análise de dados do DATASUS, gerei um único arquivo contendo todas as informações relevantes e variáveis derivadas. Utilizando a linguagem Python e a biblioteca Pandas, executei uma série de operações de integração e transformação de dados. Após garantir que todas as bases de dados — incluindo as informações de procedimentos hospitalares, dados demográficos do censo de 2022 e outros indicadores de saúde — estavam corretamente formatadas e alinhadas por meio das chaves comuns como códigos de municípios, realizei a junção final desses datasets em um único DataFrame. Este DataFrame foi meticulosamente verificado para assegurar a integridade e a precisão dos dados, aplicando filtros e validações adicionais para excluir qualquer inconsistência. Por fim, exportei esse DataFrame integrado para um arquivo CSV, utilizando o método to_csv() do Pandas, resultando em um arquivo compreensivo que contém todas as variáveis necessárias para análises futuras, garantindo assim um recurso valioso e facilmente acessível para pesquisa e tomada de decisão baseada em evidências.

1.8 Verificação da Qualidade dos Dados com Estatísticas Descritivas

Para garantir a confiabilidade e precisão do projeto de análise de dados do DATASUS, realizei uma verificação rigorosa da qualidade dos dados coletados, empregando técnicas estatísticas descritivas. Utilizando Python e a biblioteca Pandas, calculei medidas como média, mediana, desvio padrão e intervalos interquartis para cada variável relevante no conjunto de dados. Essas estatísticas forneceram uma visão clara da distribuição e das tendências centrais dos dados, permitindo-me identificar valores atípicos (outliers) que poderiam distorcer as análises subsequentes. Além disso, implementei funções para detectar e quantificar valores ausentes (missing values), explorando as possíveis razões para tais lacunas e avaliando a necessidade de imputação ou exclusão desses dados. A verificação de anomalias também incluiu a análise de consistência interna dos dados, assegurando que todas as entradas fossem lógicas e aderentes aos padrões esperados. Essa etapa crítica não só melhorou a qualidade do dataset como também assegurou que as inferências e conclusões tiradas nas fases posteriores de análise estivessem baseadas em dados sólidos e confiáveis.

2 Conclusão

Concluindo, o projeto de análise de dados do DATASUS representou uma oportunidade significativa para aplicar técnicas avançadas de ciência de dados no contexto da saúde pública brasileira. Através da coleta, tratamento, integração e análise meticulosas dos dados de internações hospitalares, foi possível oferecer insights sobre as tendências de saúde, eficácia dos tratamentos e alocação de recursos. A integração dos dados hospitalares com informações demográficas proporcionou uma perspectiva mais detalhada, permitindo análises contextualizadas que são essenciais para formuladores de políticas e administradores de saúde pública. Este trabalho não apenas demonstrou a importância crítica da ciência de dados na melhoria dos serviços de saúde, mas também estabeleceu uma base sólida para futuras investigações e intervenções baseadas em dados.

3 Referências

- Documentação oficial do DATASUS.
- Python.org - Documentação oficial das bibliotecas utilizadas.