Pontifícia Universidade Católica do Rio Grande do Sul
Introdução a Ciência de Dados

Relatório do Trabalho Final - Análise de Dados do Sistema de Informações Hospitalares (SIH)

Porto Alegre, 08 de julho de 2024

1 Resumo

Este trabalho extensionista consiste em coleta, preparação e análise de dados governamentais abertos disponíveis através do DATASUS. A base de dados selecionada foi a do Sistema de Informações Hospitalares (SIH).

2 Introdução

Este relatório documenta o processo de manipulação e análise dos dados provenientes do Sistema de Informações Hospitalares (SIH). A partir do *dashboard* é possível se responderem às seguintes perguntas:

- Qual o tempo médio de internação hospitalares por doença;
- Quais as doenças que causam mais internações;
- Que doenças possuem maior duração média;
- Que doenças possuem maior valor médio de internação;
- Qual foi a variação de gastos de internação dos municípios ao longo dos meses;

A partir de tais perguntas, é possível, por exemplo, se verificar se existe uma relação entre o aumento e diminuição de gastos com internações e determinadas épocas do ano.

Para isto os dados foram preparados e organizados em um data *warehouse* no modelo estrela criando uma tabela fatos e as tabelas dimensões com o objetivo de facilitar as análises futuras.

3 Contexto do Trabalho em Grupo

Este trabalho foi realizado em grupo, para isso foi criado um repositório no Github, com o qual cada integrante pode contribuir. Para organizar o desenvolvimento cada integrante criou sua branch e desenvolvia uma etapa.

Ao final um integrante compilou e preparou as tabelas e *dashboards* utilizando a ferramenta BI.

4 Desenvolvimento

Os dados do SIH incluem informações sobre internações hospitalares, como número da internação, identificação do paciente, diagnóstico principal, valores relacionados à internação e outros atributos relevantes. A partir destes dados para o desenvolvimento foram realizadas as etapas que seguem.

4.1 Etapa 1 – Definição do tema e criação do repositório de desenvolvimento.

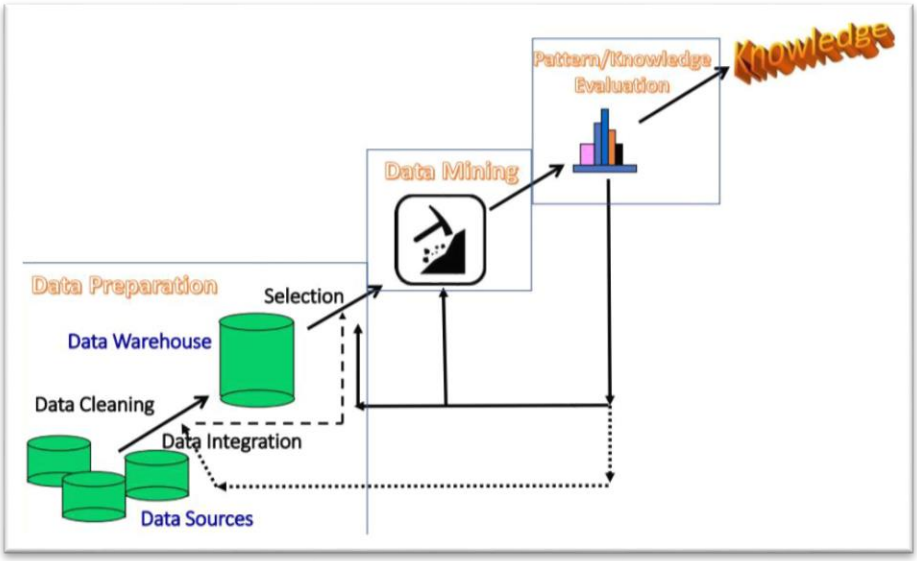
Nesta etapa foi definido a base de dados e as questões a serem respondidas. A base de dados selecionada foi do Sistema de Informações Hospitalares (SIH) e as questões a serem respondidas foram as mencionadas anteriormente.

Nesta etapa também foi criado um repositório no GitHub para o desenvolvimento do trabalho em grupo.

GitHub: https://github.com/fcfrison/t3_colet_prep_dados.git

4.2 Etapa 2 coleta e preparação dos dados:

O desenvolvimento foi realizado de acordo com o material estudado durante o semestre seguindo o modelo de processo KDD.



4.2.1 Modelo do Data Warehouse

Nesta etapa foi definido a criação e o modelo do data warehouse conforme o modelo do Anexo I.

4.2.2 Extração dos Dados

Para realizar a extração de dados do SIH, foi utilizada a biblioteca *pysus* escrita em *Python.* Foram baixados os dados dos anos de 2015 e 2016, para o Estado do Rio Grande do Sul, na modalidade RD-Reduzida. A tabela do SIH contém 113 colunas, sendo que o dicionário de dados pôde ser obtido a partir do seguinte link:

https://repositorio.ipea.gov.br/bitstream/11058/9409/1/Uma_analise_da_base_de_dados_do_sistema_de_informacao_hospitalar.pdf

Foram também utilizadas como fontes de dados:

- Tabela contendo as subcategorias CID-10: dados obtidos a partir de https://github.com/cleytonferrari/CidDataSus/tree/master/CIDImport/Repositorio/Resources

- Tabela contendo a relação de estados com seus nomes e siglas: dados obtidos a partir de

https://github.com/leogermani/estados-e-municipios-ibge/blob/master/estados.csv

- Tabela contendo a latitude e longitude dos municípios brasileiros: dados obtidos a partir de https://github.com/kelvins/municipios-brasileiros/blob/main/csv/municipios.csv

4.2.3 Transformação dos dados

Após os dados do SIH terem sido importados, foi necessário se escolherem as colunas a serem trabalhadas, sendo que as colunas escolhidas foram as seguintes:

'N_AIH', 'ANO_CMPT', 'MES_CMPT', 'DIAG_PRINC', 'MUNIC_RES', 'NASC', 'SEXO', 'QT_DIARIAS','VAL_TOT'

Após isso, foram realizadas diversas transformações. Abaixo seguem algumas delas:

Transformações
Criada coluna 'TIME_KEY' concatenando 'ANO_CMPT' e 'MES_CMPT'.
Criada tabela única de pacientes com colunas renomeadas e duplicatas removidas.
Convertida coluna 'NASC' para o formato 'dd/mm/yyyy'.
Criada tabela de municípios a partir de um arquivo CSV do IBGE.
Criada tabela de estados a partir de um arquivo CSV do IBGE.
Criada tabela 'CID_10_KEY' a partir de 'DIAG_PRINC' (cortada para 3 primeiros dígitos).
Criada tabela CID a partir de um arquivo CSV.
Removidos espaços e converte colunas 'QT_DIARIAS' e 'VALOR_INTERNACAO' para tipos numéricos.
Insere colunas 'COD_IBGE', 'lat' e 'long' com dados de outro DataFrame.

4.2.4 Criação das Tabelas Dimensões

A partir das transformações acima descritas, foram desenvolvidas as seguintes tabelas de dimensões:

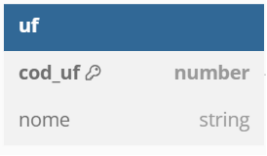
- Tabela de Dimensão “paciente”

A tabela de dimensão paciente foi criada a partir das colunas NASC (data de nascimento) e SEXO, com um ID único gerado para cada paciente.

- Tabela de Dimensão “municipio”

A tabela de dimensão “municipio” contém a seguinte estrutura:

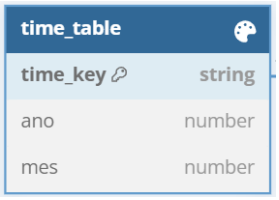
- Tabela de Dimensão “uf”



- Tabela de Dimensão “cid_10”

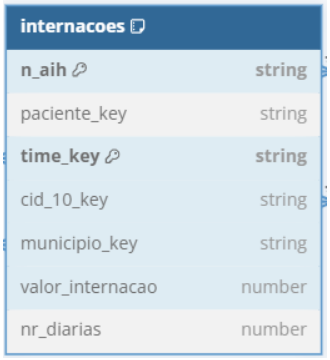
A tabela de dimensão CID foi criada a partir do código de diagnóstico principal e sua descrição:

- Tabela de dimensão “time_table”



- Tabela Fatos

A tabela fatos foi constituída pelas chaves estrangeiros das tabelas dimensões além dos fatos a serem analisados como número de diárias e valores das internações.



4.3 Apresentação dos dados ferramenta BI

Foram desenvolvidos dois *dashboards* para a apresentação das informações extraídas a partir do SIH: (i) Quadro de doenças e (ii) Quadro de gastos. Em (i), o registro das diversas doenças constantes na tabela CID-10 é priorizado. Já em (ii), os gastos dos municípios do estado do Rio Grande do Sul são enfocados.

5 Conclusão do trabalho

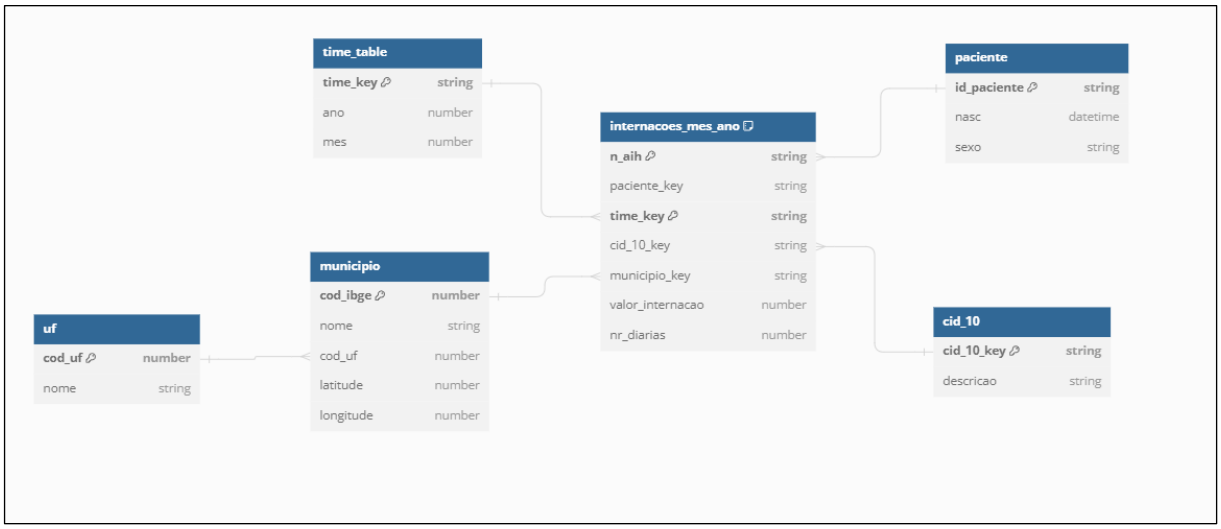
Este relatório documentou o processo de manipulação e análise dos dados do SIH para modelar um data warehouse. As tabelas de fatos e dimensões foram criadas para

facilitar análises futuras, proporcionando uma estrutura organizada e eficiente para consultas e relatórios.

6 Futuras Extensões

Automatização: Desenvolver scripts para automatizar o processo de atualização dos dados e geração de relatórios.

ANEXO I


time_table
time_key string
ano number
mes number
internacoes_mes_ano
n_aih string
paciente_key string
time_key string
cid_10_key string
municipio_key string
valor_internacao number
nr_diarias number
paciente
id_paciente string
nasc datetime
sexo string
municipio
cod_ibge number
nome string
cod_uf number
latitude number
longitude number
uf
cod_uf number
nome string
cid_10
cid_10_key string
descricao string