PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
COLETA PREPARAÇÃO E ANÁLISE DE DADOS

JOÃO GABRIEL DOURADO CERVO, ISRAEL SEGALIN E PEDRO VAZ LOREA

RELATÓRIO TRABALHO FINAL

Porto Alegre

2023

INTRODUÇÃO:

O propósito deste projeto consiste em aplicar na prática os conhecimentos aprendidos em sala de aula. O escopo do trabalho envolve a utilização do processo ETL (Extract, Transform, Load) para extrair informações de uma base de dados, realizar transformações caso necessárias e, em seguida, transferir esses dados para uma planilha no Excel. Logo, com a intenção de utilizar a plataforma Power BI para transformar esses dados em gráficos significativos, capazes de fornecer respostas a questões específicas de interesse. Sendo assim, se deu a criação de uma visualização muito mais intuitiva, interessante e informativa por meio do Power BI.

DESENVOLVIMENTO

- **Base de dados escolhida**

A base de dados escolhida foi a de Mortalidade desde 1996 pela CID-10 (mortalidade geral **e** infantil) do datasus. Dessa base, se utilizou os dados dos últimos 5 anos disponíveis (2017-2021)

- **Técnicas de Extração e Transformação utilizadas**

Em primeiro momento se realizou a coleta dos dados de forma manual no site do datasus.

Para isso, se realizou o acesso ao site do tabnet e selecionados os arquivos com os seguintes filtros – obrigatórios - diretamente no site:

- Linha: Categoria CID 10
- Categoria: Unidade da Federação

Foi realizado o download dos arquivos '.csv' do site dos últimos cinco anos – individualmente -, sendo um '.csv' para cada arquivo.

Essa extração se deu tanto para a mortalidade geral quanto para a mortalidade infantil.

Então, foi criado um algoritmo – em Python – para o processamento e limpeza dos dados. Esse algoritmo realiza a leitura dos arquivos '*.csv' extraidos na etapa anterior utilizando a biblioteca 'pandas'. Em seguida realiza as seguintes etapas:

1. Limpeza: São removidos do dataframe dados 'indesejados'. Isso é, removendo dados mal formatados e de mortes que não são de câncer (CIDs não pertencente a grupos de câncer)
2. Agrupamento: Como os dados brutos vem com mortes por estado em diferentes anos, se dá a necessidade de agrupar essa estatística. Os dados então são agrupados de acordo com seu CID.
3. Formatação: No fim do processo, os dados são formatados de modo que tenham o mesmo padrão e possam ser utilizados no Power BI. Esses dados são exportados para dois arquivos: 'dados_infantis.csv' e 'dados_adultos.csv', que contam com as seguintes colunas:

"CID", "RO", "AC", "AM", "RR", "PA", "AP", "TO", "MA", "PI", "CE", "RN", "PB", "PE", "AL", "SE", "BA", "MG", "ES", "RJ", "SP", "PR", "SC", "RS", "MS", "MT", "GO", "DF", "TOTAL"

A primeira coluna contém o CID da doença, a última contém o total de mortes daquele tipo de câncer e o restante tem como valor a quantia de mortes em determinado estado.

Por fim, se utilizou o Power BI para visualização desses dados tratados a fim de responder duas perguntas chave: 1) Qual o tipo de câncer que mais causa óbitos infantis e; 2) Qual a porcentagem de óbitos infantis por câncer que acontecem no Rio Grande do Sul? E em comparação com óbitos por câncer em adultos?

As respostas para essas perguntas podem ser vistas nas Figuras 1 e 2. Se utilizaram gráficos em pizza e barras para melhor visualização.

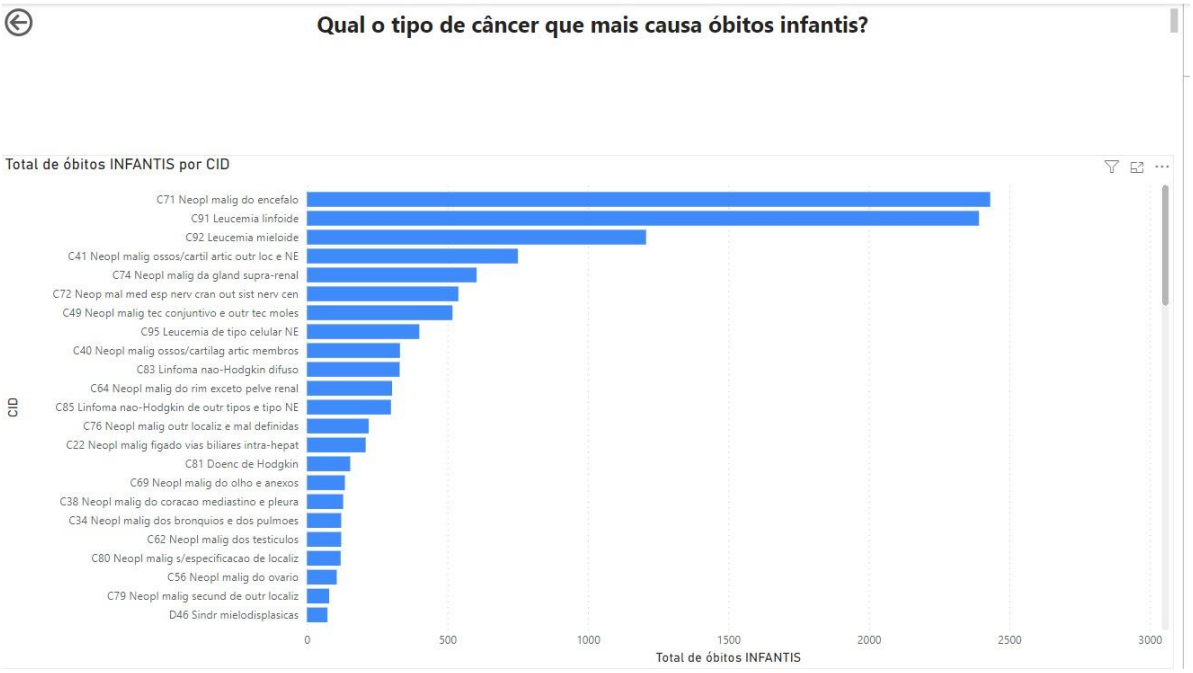

Figura 1 - Análise da pergunta "Qual o tipo de câncer que mais causa óbitos infantis?"

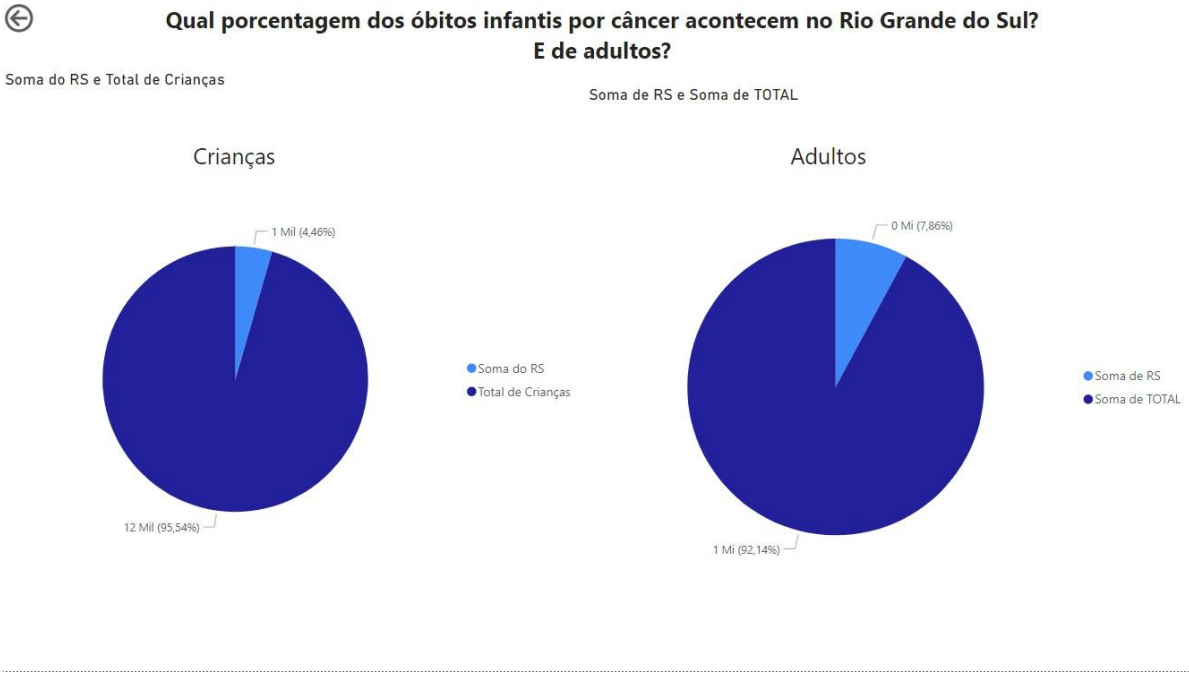

Figura 2 - Análise da pergunta "Qual a porcentagem de óbitos infantis por câncer que acontecem no Rio Grande do Sul? E em comparação com óbitos por câncer em adultos?**"**

- **Link GitHub**

O link para o repositório do projeto pode ser encontrado abaixo:

https://github.com/Gabriel-Cervo/t3-dados

CONCLUSÃO

O desenvolvimento deste projeto evidencia a aplicação prática dos conhecimentos adquiridos em sala de aula, destacando a importância do processo ETL.

A abordagem de limpeza, agrupamento e formatação permitiu criar conjuntos de dados coesos e prontos para serem explorados visualmente. Através desses dados foi possível realizar uma análise através da ferramenta Power BI, que foi essencial para a melhor visualização e entendimento dos dados obtidos.

Esse projeto como um todo ressaltou a importância do uso integrado de ferramentas e técnicas para extrair informações significativas a partir de conjuntos complexos de dados.

REFERÊNCIAS

Fonte dos dados: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10uf.def