UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Um Data Mart para Análise Comparativa Entre Dados Oficiais e Não Oficiais de Óbitos por COVID-19 no Brasil

Por

Samuel Saturnino Junior

Serra Talhada,
Outubro/2022

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SAMUEL SATURNINO JUNIOR

Um Data Mart para Análise Comparativa Entre Dados Oficiais e Não Oficiais de Óbitos por COVID-19 no Brasil

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Serra Talhada da Universidade Federal Rural de Pernambuco como requisito parcial à obtenção do grau de Bacharel.

Orientadora: Ellen Polliana R. Souza

Serra Talhada,
Outubro/2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S254d Junior, Samuel
Um Data Mart para Análise Comparativa Entre Dados Oficiais e Não Oficiais de Óbitos por COVID-19 no Brasil / Samuel Junior. - 2022.
34 f. : il.

Orientadora: Ellen Polliana Ramos Souza.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Serra Talhada, 2022.

1. Data Mart. 2. Covid-19. 3. Dados Abertos. I. Souza, Ellen Polliana Ramos, orient. II. Título

CDD 004

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

SAMUEL SATURNINO JUNIOR

Um Data Mart para Análise Comparativa Entre Dados Oficiais e Não Oficiais de Óbitos por COVID-19 no Brasil

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em 05/10/2022 pela banca examinadora.

Banca Examinadora:

Ellen Polliana R. Souza
Orientador(a)
Universidade Federal Rural de Pernambuco

Prof. Luiz Claudio Ribeiro Machado
Universidade Federal Rural de Pernambuco

Prof. Paulo Mello da Silva
Universidade Federal Rural de Pernambuco

Dedico este trabalho aos meus pais,

porque deram tudo de si para eu estar aqui.

AGRADECIMENTOS

Quero agradecer primeiramente a Deus, por ter sido generoso comigo, permitindo que este trabalho pudesse ser concluído. Quero agradecer também a minha querida Mãe a Profª. Maria de Fátima Feitosa, que tem sido uma grande incentivadora para o meu progresso. E sem ela eu não seria capaz de levar adiante muitas coisas na minha vida. Obrigado por tudo.

Quero agradecer ainda aos meus amigos por terem me apoiado em muitos momentos da minha vida.

Por fim, gostaria de agradecer a minha orientadora e amiga, a Dra. Ellen Polliana Ramos Souza, pela paciência na orientação e incentivo que me ajudou a encontrar apoio e ajuda para conseguir resolver muitos problemas e ainda alcançar conhecimento necessário para a conclusão do trabalho, a ela possuo uma eterna gratidão e admiração.

“Por mais inteligente que alguém possa ser, se não for humilde, o seu melhor se perde na arrogância. A humildade ainda é a parte mais bela da sabedoria”

(Autor: desconhecido.)

RESUMO

A situação epidemiológica da COVID-19 no Brasil teve um alto impacto na população, e diferentes autores discutem esse impacto através de análises de indicadores provenientes de fontes oficiais e não oficiais. Neste sentido, este trabalho tem como objetivo disponibilizar um Data Mart para análise comparativa entre dados oficiais e não oficiais de óbitos por COVID-19 no Brasil, referente aos anos de 2019-2022, sendo construído a partir da integração de múltiplas fontes de dados abertos e disponibilizados no formato aberto por meio de uma ferramenta OLAP que permite uma melhor visualização dos dados.

Palavras-chave: Data Mart, Covid-19, Dados Abertos.

ABSTRACT

The epidemiological situation of COVID-19 in Brazil had a high impact on the population, and different authors discuss this impact through analysis of indicators from official and unofficial sources. In this sense, this work aims to provide a Data Mart for comparative analysis between official and unofficial data on deaths by COVID-19 in Brazil, referring to the years 2019-2022, being built from the integration of multiple sources of open data. and made available in an open format through an OLAP tool that allows a better visualization of the data.

Keywords: Data Mart, Covid-19, Open Data.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

BDs	Bases de Dados
COVID-19	Coronavírus
DATASUS	Sistema de Informática do Sistema Único de Saúde
CDA	Covid Data Analytics
DM	Data Mart
DW	Data Warehouse
ETC	Extração, Transformação e Carga
IBGE	Instituto Brasileiro de Geografia e Estatística
OLAP	Online Analytical Processing
OMS	Organização Mundial da Saúde
PNADs	Pesquisa Nacional por Amostra de Domicílios
SARS-CoV	Síndrome Respiratoria Aguda Grave - Coronavírus

SUMÁRIO

1 Introdução

A pandemia declarada em 11 de março de 2020 pela Organização Mundial da Saúde (OMS) trouxe grandes desafios, por se tratar de uma doença respiratória que possui um alto nível de contágio a partir da proximidade social. Atualmente, a América Latina é uma das regiões com maior número de ocorrência da doença e o Brasil é um dos países mais afetados pela pandemia, com mais de 660 mil mortos (OMS, 2022).

A Covid-19 foi identificada pela primeira vez em seres humanos, em dezembro de 2019, na cidade de Wuhan na China. Sua transmissão se dá principalmente por contato social, podendo causar desde um resfriado comum a doenças mais graves, como a Síndrome Respiratória Aguda Grave (SARS-CoV), sendo, em 30 de janeiro de 2020, o surto da doença causada pelo novo coronavírus (COVID-19) classificado como uma emergência de saúde pública de importância internacional e, em 11 de março de 2020, a COVID-19 foi caracterizada como uma pandemia (OMS, 2022).

Assim, a OMS tem trabalhado para aprender sobre o vírus, através de algumas recomendações, feitas a todos os países onde deve-se coletar e disponibilizar de maneira precisa, informações, sempre que possível, contendo: definições de caso, resultados laboratoriais, fonte e tipos de risco, número de casos e de óbitos, condições de propagação da doença, medidas de saúde que estão sendo empregadas, dificuldades enfrentadas e solicitações de apoio caso necessário. Dessa forma nunca se viu tantos dados disponíveis em tão pouco tempo, pois além de conter informações de como o vírus afeta as pessoas e as regiões, os dados informam também as medidas de saúde que estão sendo tomadas por cada país.

Com isso, surge a necessidade de compreender os impactos deixados pela pandemia de COVID-19 na sociedade brasileira. Para isso, necessita-se da extração de informações e conhecimentos provenientes de diversas fontes, esses sendo os dados brutos, que são disponibilizados publicamente por órgãos governamentais se tornando assim dados abertos. Dados esses que, ao serem analisados e expostos ao processo de unificação, acabam gerando algumas dificuldades, como por exemplo, as informações de óbitos registradas, por serem coletadas a partir de múltiplas fontes, nem sempre são associadas a real situação de casos existente, seja de COVID-19 ou de outra doença respiratória.

Neste sentido, se torna útil o uso de um Data Mart, sendo ele uma estrutura que consolida dados de diversas fontes facilitando as análises e tendo como objetivo principal oferecer uma correlação entre dados, mesmo que originem de diferentes fontes, assim dando suporte na realização de consultas e análises.

Desta forma, este trabalho tem como objetivo geral disponibilizar um Data Mart para a análise comparativa entre dados oficiais e não oficiais de óbitos por COVID-19 no Brasil. Para tanto, foram elencados os seguintes objetivos específicos:

1. Mapear os indicadores de óbitos relacionados à COVID-19;
2. Construir um repositório de dados provenientes de múltiplas fontes sobre óbitos relacionados à COVID-19 no Brasil e
3. Disponibilizar uma aplicação OLAP para análise comparativa das curvas de óbitos por COVID-19 no Brasil.

Com isso, as análises disponibilizadas poderão contribuir para as práticas que avaliam o impacto da pandemia. Utilizando os dados abertos referentes a óbitos por COVID-19 da base do Brasil.io (JUSTEN, 2019) e da base do Registro Civil (BRASIL, 2012), busca-se responder, dentre outras, as seguintes questões de pesquisa;

QP1. Qual a quantidade anual de óbitos por COVID-19 nos estados brasileiros;

QP2. Qual o percentual de óbitos por estado e por ano;

QP3. Qual o ranking entre estados com maior ocorrência de óbitos;

QP4. Qual o ranking entre regiões com maior ocorrência de óbitos;

QP5. Qual gênero foi mais impactado por algum tipo de doença respiratória;

QP6. Qual a quantidade de óbitos em cada faixa etária;

QP7. Qual a evolução anual de óbitos por doença respiratória e;

QP8. Qual tipo da localidade com maior ocorrência de óbitos.

Este trabalho está organizado da seguinte maneira: Na Seção 2, são apresentados os trabalhos relacionados, bem como suas contribuições para o presente estudo. A Seção 3, apresenta o método utilizado nesse projeto. A Seção 4 expõe os resultados e a Seção 5 apresenta as conclusões.

2 Referencial Teórico

2.1 Data Mart

Um Data Warehouse (DW) trata-se de um grande repositório de dados, orientado por assunto, integrado, variante no tempo, e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão (INMON, 2003). A maioria dos dados que um DW utiliza são agrupados com a ajuda de um software especializado que efetua a extração os dados provindos de Bases de Dados (BDs) sem nenhum tratamento prévio, sendo os dados processados, sintetizados e agregados, para que sejam armazenados no DW com o intuito de aumentar a sua eficiência de exploração.

Um Data Mart (DM), também conhecido como Warehouse Departamental, é uma abordagem descentralizada do conceito de DW, por se trata da divisão de um DW, em subconjuntos delineados por assuntos específicos e que possui uma série de procedimentos para seu desenvolvimento, responsáveis desde os levantamentos de requisitos até ao uso do usuário final, sendo ele capaz de disponibilizar dados de uma forma simplificada e clara (KIMBALL; ROSS, 2011).

Um DM representa uma parte menor tanto de arquitetura como de foco, sendo excelente para projetos que possuam dados originários de múltiplas fontes e que tenham uma maior complexidade sobre as relações deles. As tecnologias utilizadas tanto nos DW quanto nos DM são as mesmas, o que se diferencia, é o volume de dados, abrangência da arquitetura e o foco para qual tal estrutura foi criada, ou seja, os DM são voltados somente para uma determinada área referenciando um escopo menor, sendo uma unidade de um sistema que forma um DW.

Segundo (KIMBALL; ROSS, 2011), no ambiente de um DW (Figura 2.1), existem quatro componentes distintos sendo eles: sistemas operacionais de origem (sistemas que capturam as transações da empresa), data stanging área (área de armazenamento de dados e de conjunto de processos que preparam os dados de origem para serem utilizados), área de apresentação de dados (local onde os dados ficam armazenados e disponíveis ao usuário final) e, ferramentas de acesso a dados (ferramentas OLAP e de mineração de dados que permitem aos usuários utilizar os dados de uma maneira rápida e fácil para executar análises mensuráveis).

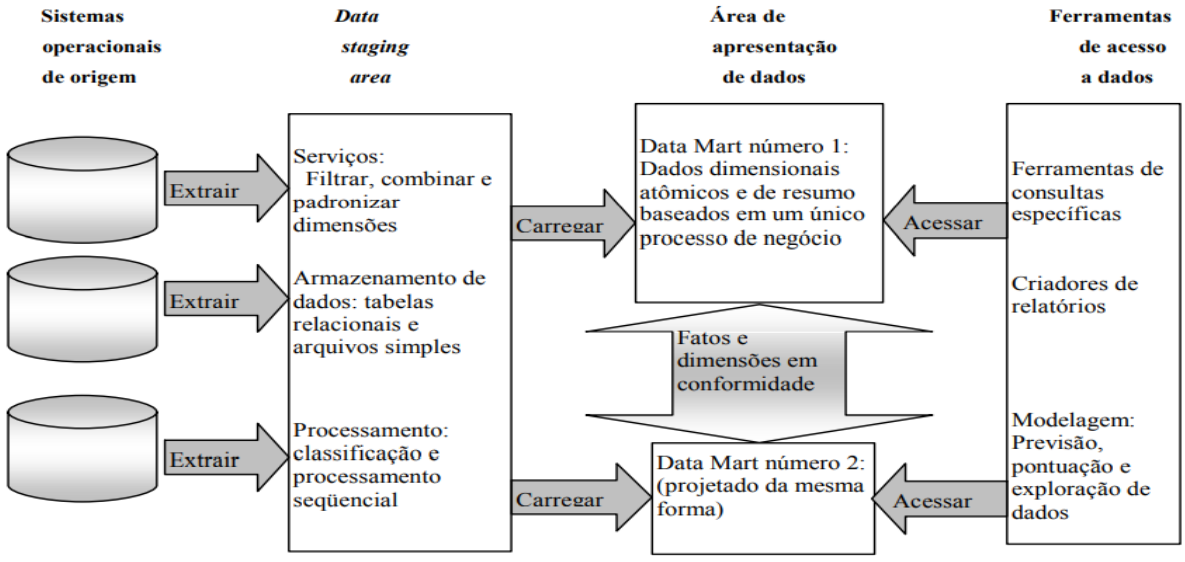
2.2 O modelo Kimball

Ralph Kimball desenvolveu sua metodologia em 1996, (KIMBALL, 1996), recomendando uma arquitectura de múltiplos BDs e DMs, organizadas por áreas de negócio, sendo o DW definido como a soma dos vários DMs.

Para seu desenvolvimento, é recomenda uma metodologia inversa à de (**??**), uma abordagem *bottom-up*, que parte da análise dos vários sistemas individuais terminando com a agregação deles num grande DW.

Além disso, (KIMBALL; ROSS, 2011) propuseram a utilização de um modelo chamado *Star Schema* com o objetivo de simplificar as visualizações de um DW, facilitando a distinção entre tabelas de dimensões e tabelas de fato, ou seja, a fato é uma métrica (algo que pode ser medido ou quantificado) resultantes de um evento do processo de negócio ou relação entre dimensões, que aqui representam o contexto para análise de uma fato, podendo ser vista pelo usuário comum como possíveis filtros que determinam uma tabela de fato.

Figura 2.1 – Elementos Básicos do Data Warehouse.



Fonte:(KIMBALL; ROSS, 2011).

2.3 Trabalhos Relacionados

Em publicações acadêmicas, é possível encontrar diversos trabalhos voltados para análises epidemiológicas. Nesta pesquisa, foram incluídos estudos que visam analisar indicadores relacionados a COVID-19, suas causas e seus impactos para a sociedade, através da coleta de informações disponibilizadas por alguma fonte externa, geralmente dados governamentais abertos.

2.3.1 Covid Data Analytics: Repositório de Dados Provenientes de Múultiplas Fontes sobre a Pandemia de COVID-19 no Brasil

O trabalho desenvolvido por (MOREIRA et al., 2021) apresenta a construção e publicação de um repositório de dados utilizados no âmbito do projeto Covid Data Analytics (CDA), feito pelo Departamento de Ciências da Computação da UFMG. O projeto faz um monitoramento de aspectos referentes a situação social, econômica e epidemiológica da COVID-19 no Brasil, a partir da análise de dados provenientes de fontes oficiais e não oficiais, de redes sociais online e da web em geral.

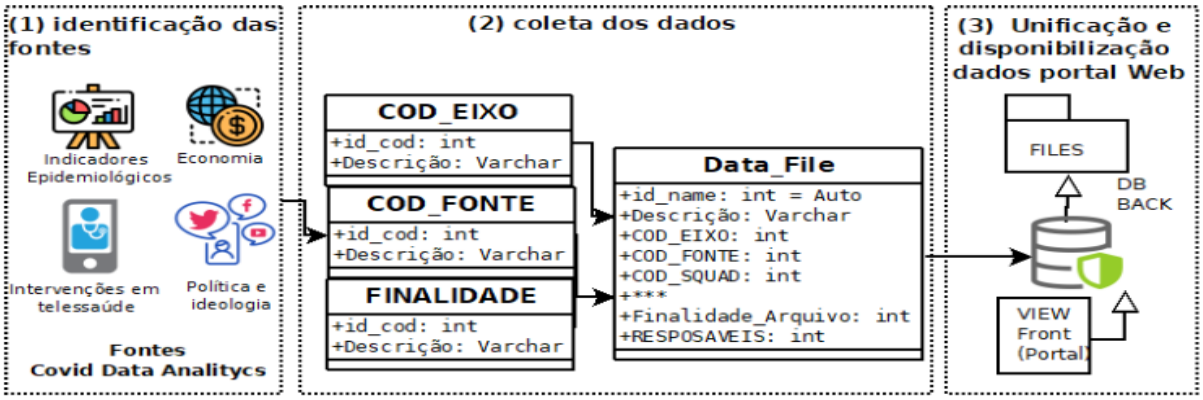
O projeto contém 18 atributos e 1086 registros coletados, enriquecidos e disponibilizados por meio de uma ferramenta de busca desenvolvida exclusivamente para ele. Sua execução se deu entre os anos de 2020 e 2021, e sua equipe de pesquisa se dividiu em quatro linhas de pesquisa: (i) “Análise do comportamento da economia brasileira”; (ii) “Estratégias de intervenções de telessaúde na pandemia de COVID-19”; (iii) “Indicadores epidemiológicos e comportamento na web”; e (iv) “Política, ideologia e informações médicas nas redes”. Dessa forma, cada linha de pesquisa promoveu a coleta de dados de diferentes naturezas, com o objetivo de subsidiar as análises a serem realizadas.

Para a construção do repositório, os autores dividiram a sua metodologia, mostrada na Figura 2.2, em três etapas: a primeira foi a de identificação das fontes que utilizou a aplicação de questionários para destacar os dados com maior relevância, sejam eles de fonte oficiais (IBGE, PNADs, DATASUS), não oficiais (Brasil.IO), dados de redes sociais online (Twitter e YouTube) ou dados da web (Google Trends); a segunda etapa foi a coleta dos dados que compreendeu as fontes selecionadas na primeira etapa; por fim, a terceira etapa foi a unificação dos dados (bases,

tabelas, gráficos, mapas, relatórios, artigos etc.), visando a apresentação dos mesmos via portal web, no qual os arquivos foram enriquecidos com metadados para a implementação do banco de dados que utilizou o MySQL junto a uma interface de busca.

O trabalho resultou na disponibilização das informações mais relevantes sobre COVID-19 em uma interface, que contém todos os materiais coletados e desenvolvidos ao longo do projeto, ficando disponíveis publicamente no portal web chamado Zenodo[1]. Os autores concluíram que o projeto pode ser utilizado em várias análises sobre o impacto da COVID-19 no Brasil, uma vez que a pandemia ainda segue em curso.

Figura 2.2 – Metodologia para Construção do Repositório Covid Data Analytics.



Fonte:(MOREIRA et al., 2021).

2.3.2 Subnotificação de Mortalidade COVID-19 no Brasil: Análise de Dados de Portais da Internet do Governo

O trabalho desenvolvido por (SILVA et al., 2020) faz uma investigação das subnotificações de casos e óbitos relacionados ao COVID-19, visando identificar o real impacto da pandemia, nas seis capitais que tiveram o maiores números de óbitos registrados em portais oficiais do governo brasileiro o DATASUS (Departamento de Informática do Sistema Único de Saúde) e o Portal da Transparência Brasileira do Registro Civil (Registro Civil), sendo elas: Belém (capital do Pará), Fortaleza (capital do Ceará), Manaus (capital do Amazonas), Recife (capital de Pernambuco), Rio de Janeiro (capital do Rio de Janeiro) e São Paulo (capital paulista).

Para isso, foram utilizados dados de óbitos históricos por problemas respiratórios e outras causas naturais, as bases do DATASUS, referente aos anos de 2010-2018, e do Registro

[1] https://zenodo.org/

Civil, referente aos anos de 2019-2020, mostrada na Figura 2.3. Esses dados foram utilizados para a construção de um modelo que utilizou regressão modular para tentar prever os padrões de mortalidade esperados para 2020. As previsões foram utilizadas para estimar o possível número de óbitos registrados incorretamente durante a pandemia e publicados em portais de internet do governo nas cidades mais afetadas do país.

Dessa forma, foi criando um modelo de série temporal para estimar os números reportados incorretamente pelas organizações governamentais. Sua metodologia foi dividida em 4 etapas sendo: (1) Extração de dados - coleta os dados a partir das fontes governamentais; (2) Processamento de dados - pré-processa os dados removendo informações ausentes e duplicadas; (3) Marchine Learning – efetua um treinamento baseado no modelo de regressão modular Fb-Prophet para prever o número esperado de óbitos para 2020; e (4) Interpretação e Validação de Dados - utiliza dados sobre as mortes relacionadas ao COVID-19 das seis capitais com maior número de óbitos, prevendo assim as tendências da doença e determinando a diferença entre o número de casos esperados para 2020 e os casos registrados em 2020.

O modelo proposto destacou taxas significativas de subnotificações de óbitos analisadas, demonstrando que os números oficiais divulgados, são muito inferiores aos números reais, impossibilitando que as autoridades implementem uma resposta pandêmica mais eficaz. Com base em análises realizadas, utilizando diferentes taxas de letalidade, pôde-se inferir que a epidemia brasileira estava piorando, e o número real de infecciosos já poderia estar entre 1 e 5,4 milhões.

Figura 2.3 – Aumento do número de óbitos por insuficiência respiratória e síndrome respiratória aguda grave (SARS).

Fonte:(SILVA et al., 2020).

2.4 Comparação dos trabalhos relacionados com o presente trabalho

Os trabalhos relacionados, apesar de possuírem informações importantes sobre análises de dados, eles possuem dados desatualizados sobre a real situação epidemiológica de COVID-19, mesmo ambos possuindo uma coleta de dados feita a partir de múltiplas fontes, eles possuem diferenças com o trabalho aqui apresentado, sendo algumas delas destacadas na Figura 2.4.

Sobre o trabalho feito por (MOREIRA et al., 2021), que apresenta a construção de um repositório de dados e o disponibiliza por uma aplicação própria, tem como principal diferença, o fato de que, o presente trabalho disponibiliza não só o repositório de dados, mas uma ferramenta OLAP, que apresenta todos os resultados na forma de painéis, gráficos e tabelas com a possibilidade de aplicar novos filtros de consultas abrangendo assim uma maior quantidade de indicadores.

Já o trabalho feito por (SILVA et al., 2020), que apresenta a coleta dados sobre subnotificações feitas por algumas capitais brasileiras, disponibiliza análises de visualização em sua maioria no formato de tabelas e com indicadores ligados à área da estatística, determinado assim um maior conhecimento para o usuário final, e possuindo então uma maior dificuldade de compreensão. O presente trabalho, por sua vez, utilizou diferentes tipos de consultas, além de comparar os resultados obtidos entre os dados oficiais e não oficiais de todos os municípios brasileiros, obtendo assim uma abrangência bem maior de estudo.

Dessa forma, o presente trabalho atualizou as informações usadas nos trabalhos relacionados ao desenvolver uma análise comparativa entre dados oficiais e não oficiais, sobre a curva de óbitos deixada pela pandemia de COVID-19 no Brasil ao utilizar uma ferramenta OLAP para tal finalidade. Assim o foco do trabalho foram os dados oficiais e não oficiais, que foram submetidos ao processo de extração, transformação e carga, já a ferramenta OLAP foi encarregada de facilitar a comparação dos resultados das consultas obtidas, ao manter o nível hierárquico dos indicadores para ambas as bases, assim conseguindo atender as questões de pesquisa proposta.

Figura 2.4 – Comparação dos Trabalhos Relacionados com o Presente Trabalho.

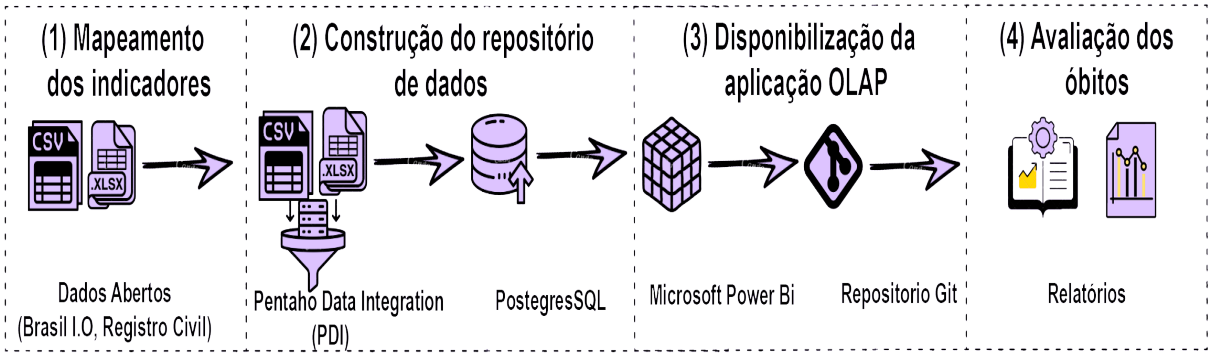
	Silva.	Moreira.	Este Trabalho.
Ano de Abrangência	2019-2020	2020-2021	2019-2022
Disponibiliza o Repositório	Não	Sim	Sim
Disponibiliza Aplicação Para Vizualização dos Dado	Sim	Sim	Sim
Disponibiliza Análise Estadual	Sim	Sim	Sim
Disponibiliza Análise Municipal	Não	Sim	Sim
Disponibiliza Análise Detalhada dos Casos	Sim	Sim	Sim
Disponibiliza Análise Por Gênero	Não	Sim	Sim
Disponibiliza Análise de Casos Diários	Não	Sim	Sim
Disponibiliza Análise por Faixa Etária	Não	Não	Sim
Disponibiliza Análise do Local do Óbito	Não	Não	Sim
Comparar Dados Oficiais com Não Oficiais	Não	Não	Sim
Disponibilizar uma Ferramenta OLAP	Não	Não	Sim

Fonte: Elaborado pelo autor.

3 Método

Nesta Seção, é apresentado o método adotado para construção do Data Mart, adaptado de Kimball, sendo dividido em quatro etapas que abrangem desde o planejamento até a entrega da ferramenta OLAP, conforme apresentado na Figura 3.1. Por se tratar de um Data Mart, sua construção utilizara a abordagem *bottom-up*.

Figura 3.1 – Método para Construção do Data Mart.

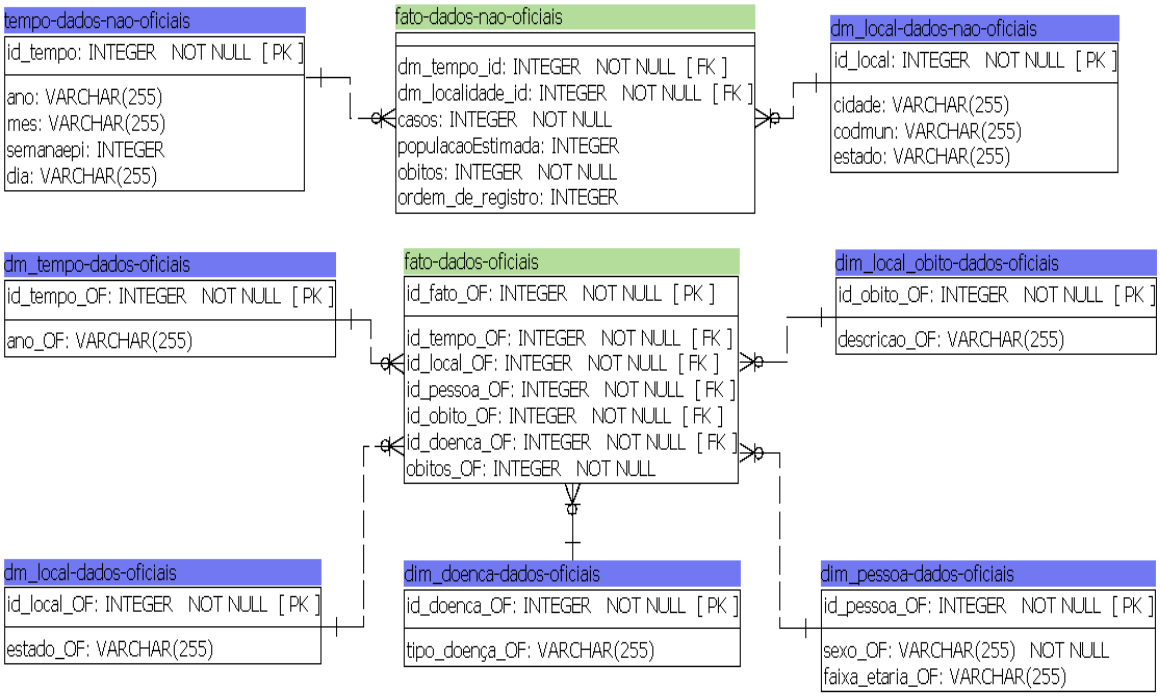


Fonte: Elaborado pelo autor.

O **mapeamento dos indicadores** foi feito com a realização de uma pesquisa bibliográfica, na qual verificou trabalhos já efetuados na área como o de (MOREIRA et al., 2021) e o de (SILVA et al., 2020). Foram utilizados os dados das fontes que atenderam aos objetivos específicos que foram apresentados na Seção 1 e, a partir disso, foi selecionado os dados que continham maior relevância para a pesquisa.

Para a **construção do repositório de dados**, foi criado um modelo dimensional, que utilizou uma abordagem do tipo *"Modelo Estrela"* proposto por (KIMBALL; ROSS, 2011). O modelo contém duas tabelas de fato nomeadas de “fato-dados-nao-oficiais” e “fato-dados-oficiais”, e sete tabelas de dimensões chamadas de “dim-tempo-dados-nao-oficiais”, “dim-tempo-dados-oficiais”, “dim-local-dados-nao-oficiais”, “dim-local-dados-oficiais”, “dim-local-obito-dados-oficiais”,”dim-doenca-dados-oficiais” e “dim-pessoa-dados-oficiais”, como é mostrado na Figura 3.2, a carga das dimensões encontra-se nos anexos desse trabalho (apêndices A1 a A8). O proximo passo foi a construção de um projeto físico, no qual utilizou um banco de dados SGBD relacional *PostgressSQL*.

Figura 3.2 – Modelo Dimensional.



Fonte: Elaborado pelo autor.

Após a construção do modelo dimensional, foi feito o processo de Extração, Transformação e Carga (ETC) por meio da ferramenta *Pentaho Data Integration*, ferramenta essa que efetua o tratamento nos dados brutos, coletados através de arquivos. Os dados foram extraídos de múltiplas fontes, sendo elas; a base oficial de óbitos por doenças respiratórias do portal da transparência do registro civil (BRASIL, 2012) e a base não oficial de boletins informativos e casos do coronavírus do Brasil.io (JUSTEN, 2019), ambas contendo informações referentes sobre a mortalidade causada pelo COVID-19, e abrangendo os últimos anos de pandemia, cada uma contendo suas particularidades nos seus indicadores.

A base do Brasil.io disponibiliza uma coleta diária contendo os seguintes indicadores: Cidade, Código IBGE da Cidade, Data, Semana Epidemiologia, População Estimada, População Estimada em 2019, Última Atualização, Repetido, Último confirmado, Último confirmado por 100k habitantes, Última data de confirmação, Última Taxa de Mortalidade Disponível, Últimas Mortes Disponíveis, Caso Por Lugar, Tipo do Lugar, Estado, Novos Casos Confirmados e Novos Óbitos Confirmados. Já a base do Registro Civil, disponibiliza uma coleta anual contendo os seguintes indicadores: UF, Tipo de doença, Local de óbito, Faixa etária e Sexo.

Os campos escolhidos para serem utilizados no estudo, foram tratados e comparados, com o intuito de agrupar dados relevantes, resultando assim em uma padronização usada

para dar carga no banco de dados criado. Após o tratamento de ETC, foi feito o processo de **disponibilização da aplicação OLAP**, a qual utilizou a ferramenta *Microsoft Power BI* para efetuar as consultas dos indicadores propostos no modelo dimensional acima citado, sendo assim gerado *dashboards* (painéis visuais que apresentam indicadores e suas métricas.) que foram disponibilizados de forma publica através de um repositório no *GitHub*, no qual está hospedando o projeto e disponibilizando seu download. [2]

Por fim, foi feita a **avaliação dos óbitos**, sendo essa etapa responsável por analisar os relatórios das consultas feitas pela ferramenta OLAP, comparando os dados oficiais do Registro Civil (dados coletados e disponibilizados por fontes governamentais) com os dados não oficiais do Brasil.io (dados produzidos pelo poder público, mas organizados e disponibilizados pelo manifesto *Brasil.io*, que tem a missão de tornar acessíveis dados de interesse público para a população).

[2] https://github.com/Samuelssj/Data-Mart-COVID-19-TCC

4 Resultados e Discussão

Nas subseções a seguir, são apresentados os resultados desse trabalho: A Subseção 4.1 apresenta os resultados para as questões de pesquisa, e a Subseção 4.2 apresenta inconsistência e problemas com os dados abertos.

4.1 Resultados Para as Questões de Pesquisa

A aplicação OLAP que foi construída permite diversos tipos de consultas, sobre os indicadores presentes nas bases de dados oficiais e não oficiais. Os seus dados foram validados, comparando o total de óbitos registrados pelas consultas da ferramenta OLAP com o histórico de mortalidades disponibilizado pela portal da OMS, que declara a existência de 34.587.047 casos de COVID-19 no Brasil e 685.376 óbitos, sendo acessado em 20/09/2022. A comparação também foi efetuada com o total de óbitos disponibilizado pelo portal oficial do Registro Civil, como mostra a Figura 4.1.

Figura 4.1 – Óbitos por Doenças Respiratórias (2019-2022).

Fonte: (BRASIL, 2012).

Os resultados das consultas OLAP foram apresentados através do painel inicial da aplicação OLAP, como mostrado na Figura 4.2.

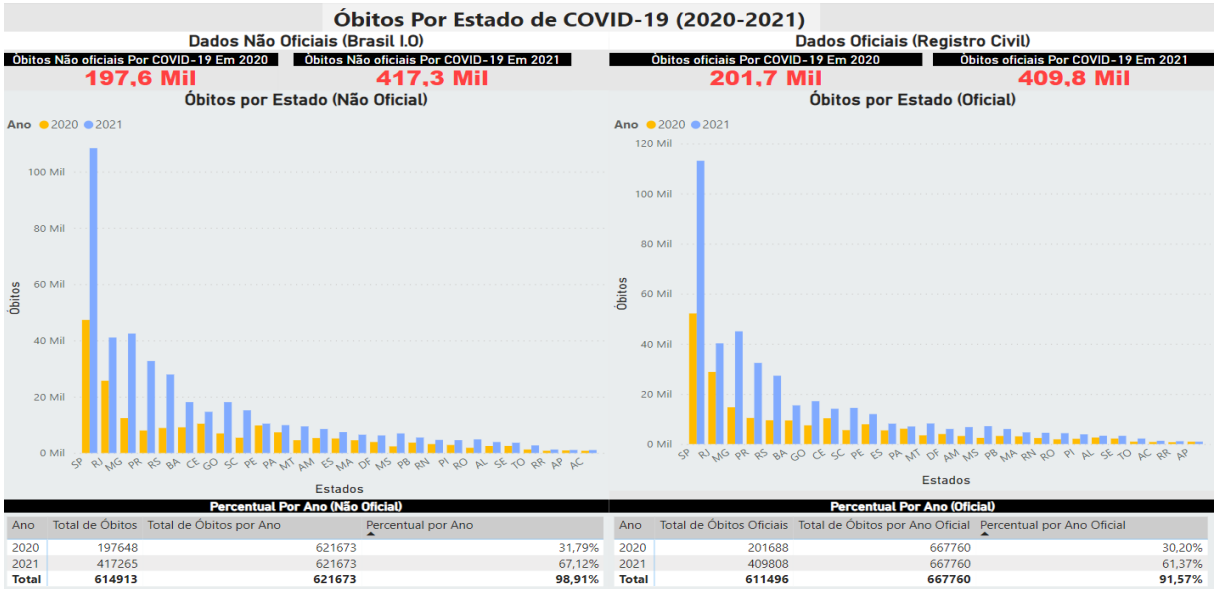
Figura 4.2 – Dashboard Inicial da Ferramenta OLAP.

Fonte: Elaborado pelo autor.

4.1.1 QP1. Qual a quantidade anual de óbitos por COVID-19 nos estados brasileiros?

A Figura 4.3 apresenta a quantidade de óbitos registrados em 2020 e 2021 em todos os estados brasileiros e o quantitativo de cada ano. Apresentando as primeiras divergências entre os dados, já que em 2020 a base oficial possuía valores maiores do que a base não oficial, e em 2021, os valores se inverteram, com a base não oficial apresentando quantidade maior. Assim, o percentual total de óbitos por COVID-19 para a base oficial foi de 667,7 mil casos e para a base não oficial foi de 621,6 mil. A figura ainda apresenta um total de óbitos diferente para alguns estados.

Figura 4.3 – Óbitos Por Estado de COVID-19 (2020-2021).



Fonte: Elaborado pelo autor.

4.1.2 QP2. Qual o percentual de óbitos por estado e por ano?

A Figura 4.4 apresenta o percentual de óbitos por estado de Covid-19, expondo algumas inconsistências nos dados para alguns estados, como é o caso de Goiás e Ceará, que mudam de posição ao terem seu percentual alterado de acordo com a base analisada. Com isso surge a questão sobre qual base apresenta uma maior proximidade da realidade de cada estado.

Figura 4.4 – Percentual de Óbitos Por Estado de COVID-19 (2020-2021).

Dados Não Oficiais (Brasil I.O)

Total de Óbitos Por COVID-19 (Não Oficial)
614,9 Mil

Percentual por Estado (Não Oficial)

Estado	Total de Óbitos	Soma total de Óbitos	Percentual
SP	155656	614913	25,31%
RJ	66742	614913	10,85%
MG	54869	614913	8,92%
PR	40702	614913	6,62%
RS	36809	614913	5,99%
BA	27223	614913	4,43%
CE	25068	614913	4,08%
GO	25046	614913	4,07%
SC	20629	614913	3,35%
PE	20241	614913	3,29%
PA	17273	614913	2,81%
Total	**614913**	**614913**	**100,00%**

Dados Oficiais (Registro Civil)

Total de Óbitos Por COVID-19 (Oficial)
611,5 Mil

Percentual por Estado (Oficial)

Estado	Total de Óbitos Oficiais	Soma total de Óbitos Oficiais	Percentual Oficial
SP	165254	611496	27,02%
RJ	68995	611496	11,28%
MG	59678	611496	9,76%
PR	42800	611496	7,00%
RS	36770	611496	6,01%
BA	24855	611496	4,06%
GO	24580	611496	4,02%
CE	24386	611496	3,99%
SC	19973	611496	3,27%
PE	19856	611496	3,25%
ES	13590	611496	2,22%
Total	**611496**	**611496**	**100,00%**

Fonte: Elaborado pelo autor.

A Figura 4.5 apresenta o percentual de óbitos por ano. Esses totais evidenciam que os dados não oficiais possuem um quantitativo bem maior de óbitos em comparação aos dados oficiais. Assim o percentual total de óbitos para a base oficial foi de 667,7 mil casos, sendo 30,20% desse valor pertencente a 2020 e 61,37% pertencente a 2021, já para a base não oficial o percentual total de óbitos foi de 621,6 mil casos, sendo 31,79% desse valor pertence a 2020 e 67,12% pertencem a 2021.

Figura 4.5 – Percentual de Óbitos Por Ano de COVID-19 (2020-2021).

Dados Não Oficiais (Brasil I.O)

Óbitos Não oficiais Por COVID-19 Em 2020
197,6 Mil

Óbitos Não oficiais Por COVID-19 Em 2021
417,3 Mil

Percentual Por Ano (Não Oficial)

Ano	Total de Óbitos	Total de Óbitos por Ano	Percentual por Ano
2020	197648	621673	31,79%
2021	417265	621673	67,12%
Total	**614913**	**621673**	**98,91%**

Dados Oficiais (Registro Civil)

Óbitos oficiais Por COVID-19 Em 2020
201,7 Mil

Óbitos oficiais Por COVID-19 Em 2021
409,8 Mil

Percentual Por Ano (Oficial)

Ano	Total de Óbitos Oficiais	Total de Óbitos por Ano Oficial	Percentual por Ano Oficial
2020	201688	667760	30,20%
2021	409808	667760	61,37%
Total	**611496**	**667760**	**91,57%**

Fonte: Elaborado pelo autor.

4.1.3 QP3. Qual o ranking entre estados com maior ocorrência de óbitos?

A Figura 4.6 apresenta a comparação entre posições de ranking de cada estado, uma consulta que classifica os estados e as regiões em um posicionamento, ordenando-os de forma decrescente, do maior número de óbitos ao menor, apresentando alterações nos rankings a partir do 7º colocado, alterações essas que podem ser alertas sobre irregularidades nas informações de alguns estados ou na própria base.

Figura 4.6 – Ranking De Óbitos Por COVID-19 (Nível Estadual) 2020-2021.

Dados Não Oficiais (Brasil I.O) **Dados Oficiais (Registro Civil)**

Ranking (Não Oficial) de Óbitos por Estado		
Estado	Total de Óbitos	Ranking
SP	155656	1
RJ	66742	2
MG	54869	3
PR	40702	4
RS	36809	5
BA	27223	6
CE	25068	7
GO	25046	8
SC	20629	9
PE	20241	10
PA	17273	11
MT	14032	12
AM	13857	13
ES	12608	14
MA	11054	15
Total	**614913**	**1**

Ranking (Oficial) de Óbitos por Estado		
Estado	Total de Óbitos Oficiais	Ranking Oficial
SP	165254	1
RJ	68995	2
MG	59678	3
PR	42800	4
RS	36770	5
BA	24855	6
GO	24580	7
CE	24386	8
SC	19973	9
PE	19856	10
ES	13590	11
PA	13127	12
MT	11674	13
DF	10015	14
AM	9984	15
Total	**611496**	**1**

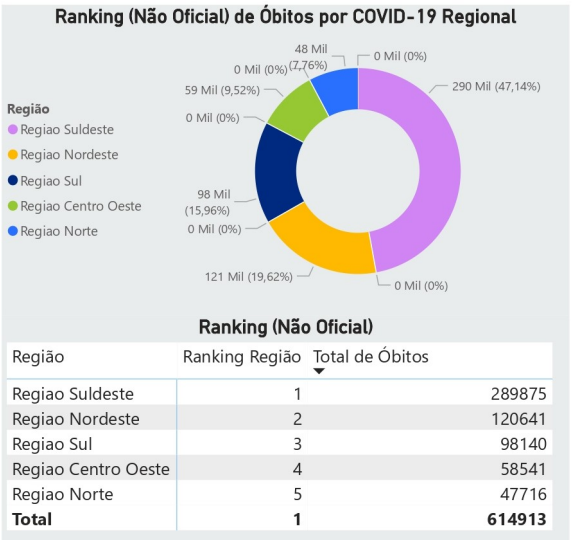
Fonte: Elaborado pelo autor.

4.1.4 QP4. Qual o ranking entre regiões com maior ocorrência de óbitos?

A Figura 4.7 apresenta o ranking das regiões Norte, Nordeste, Centro-Oeste, Sudeste e Sul. O ranking apresentado não teve nenhuma alteração entre as posições das regiões, embora o total de óbitos sim, provando que a densidade demografia de cada região não afetou o resultado das posições. Sendo o quantitativo de óbitos dos dados não oficiais ligeiramente maiores em algumas regiões, e os dados oficiais maiores em outras.

Figura 4.7 – Ranking De Óbitos Por COVID-19 (Nível Regional) 2020-2021.

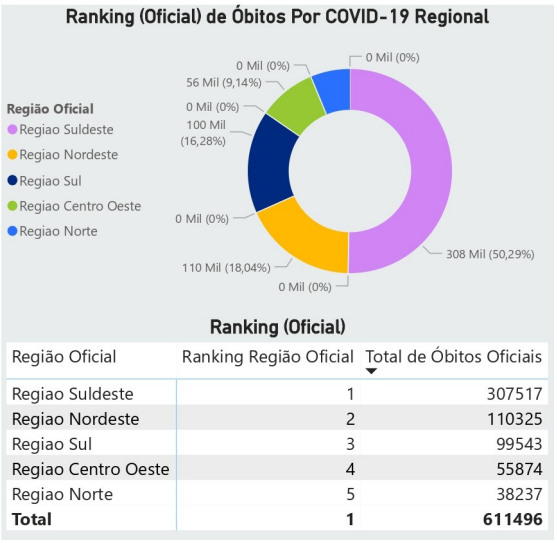
Dados Não Oficiais (Brasil I.O)



Ranking (Não Oficial)

Região	Ranking Região	Total de Óbitos
Regiao Suldeste	1	289875
Regiao Nordeste	2	120641
Regiao Sul	3	98140
Regiao Centro Oeste	4	58541
Regiao Norte	5	47716
Total	**1**	**614913**

Dados Oficiais (Registro Civil)



Ranking (Oficial)

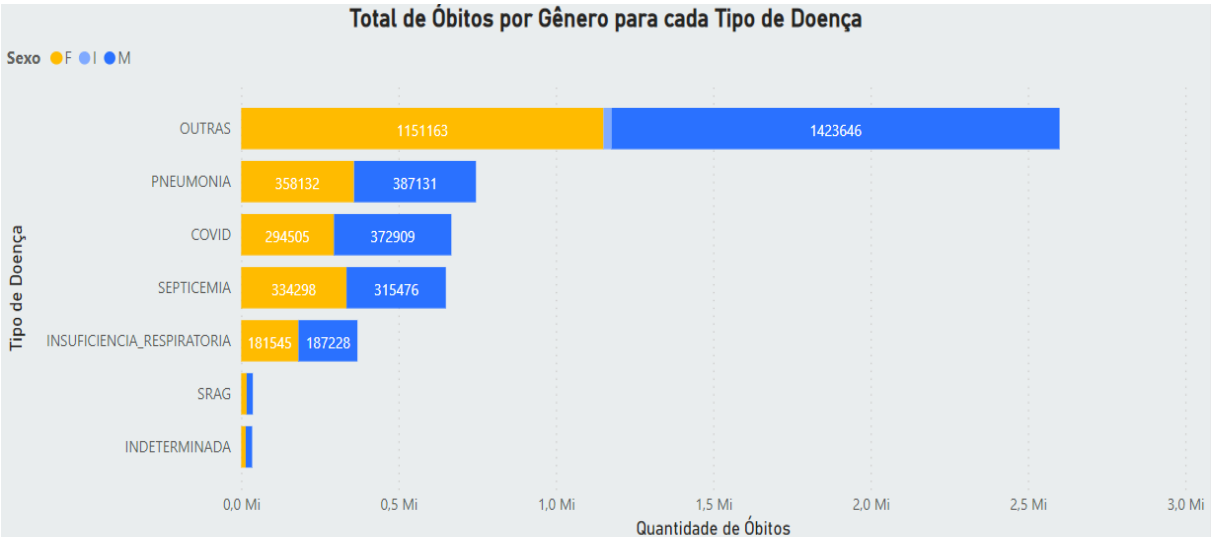
Região Oficial	Ranking Região Oficial	Total de Óbitos Oficiais
Regiao Suldeste	1	307517
Regiao Nordeste	2	110325
Regiao Sul	3	99543
Regiao Centro Oeste	4	55874
Regiao Norte	5	38237
Total	**1**	**611496**

Fonte: Elaborado pelo autor.

4.1.5 QP5. Qual gênero foi mais impactado por algum tipo de doença respiratória?

A partir dessa questão de pesquisa, foi utilizado apenas a base oficial do Registo Civil, pois a base não oficial do Brasil.io não possui os indicadores necessários em seus dados para formular um resultado. A Figura 4.8 apresenta detalhes dos óbitos referentes ao gênero do indivíduo. A consulta revela que o sexo mais afetado por doenças respiratórias entre 2019 e 2022 foi o sexo masculino.

Figura 4.8 – Óbitos por Doenças Respiratórias (Dados Oficiais) 2019-2022.

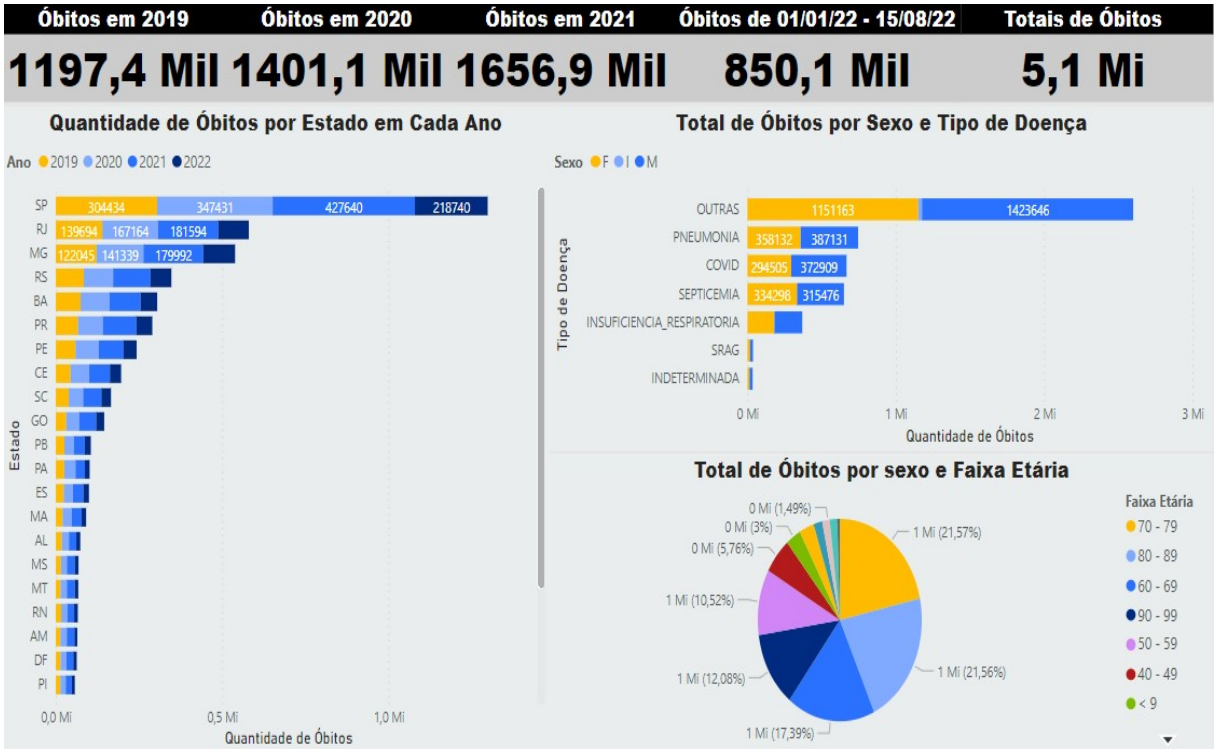


Fonte: Elaborado pelo autor.

4.1.6 QP6. Qual a quantidade de óbitos em cada faixa etária?

A Figura 4.9 apresenta a quantidade de óbitos em cada faixa etária, através de informações do período de 01/01/2019 a 15/08/2022. Outro fato que se destaca nessa consulta e a faixa etária que possui mais casos sendo ela entre 70 e 89 anos.

Figura 4.9 – Óbitos por Doenças Respiratórias (Dados Oficiais) 2020-2021.

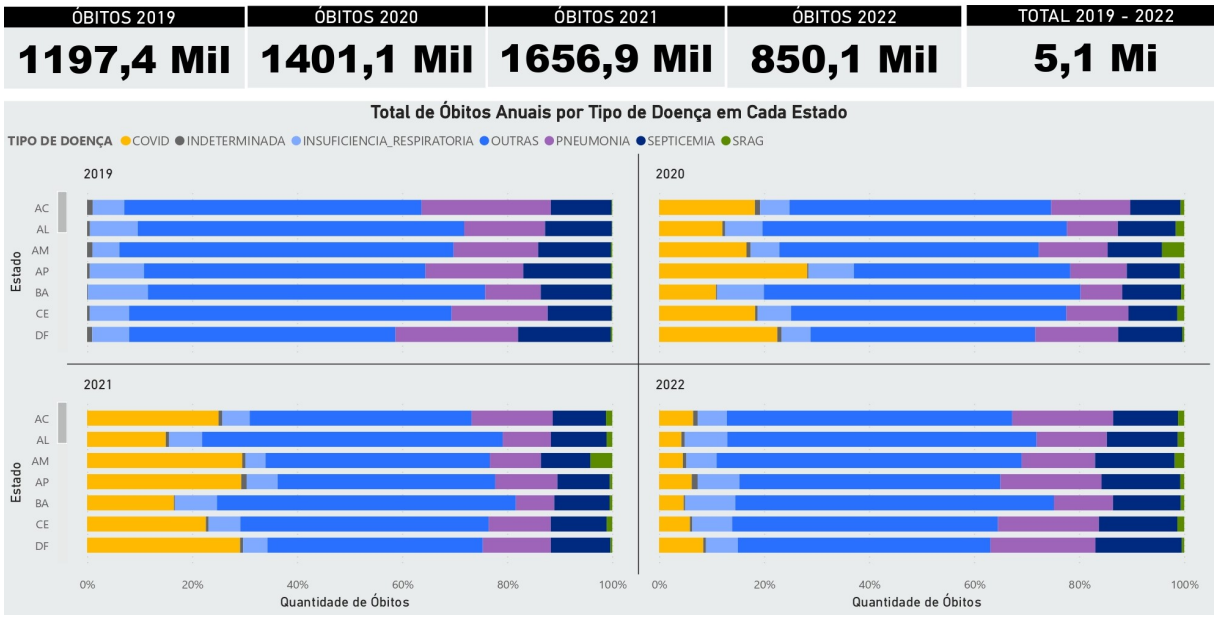


Fonte: Elaborado pelo autor.

4.1.7 QP7. Qual a evolução anual de óbitos por doença respiratória?

A Figura 4.10 apresenta a evolução anual de óbitos por doença respiratória em todos os estados brasileiros, ficando evidenciado o total de óbitos por tipo de doença e destacando-se a pneumonia, por ser uma das doenças que mais matou nos últimos anos.

Figura 4.10 – Total de Óbitos Anuais por Tipo de Doença em Cada Estado.

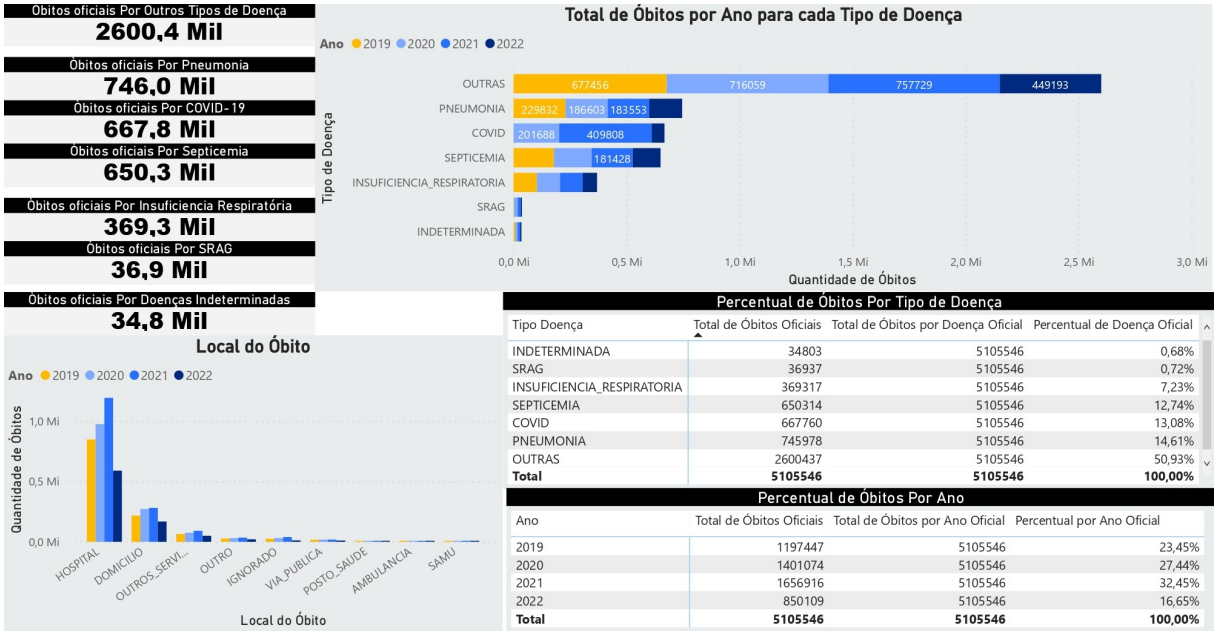


Fonte: Elaborado pelo autor.

4.1.8 QP8. Qual tipo da localidade com maior ocorrência de óbitos?

A Figura 4.11 apresenta as localidades com a maior ocorrência de óbitos, ficando evidente que o maior número de casos ocorreu em zona hospitalar. Dessa forma pode-se destacar que a maioria dos óbitos que ocorreram entre 2019 e 2020, obtiveram acompanhamento de órgãos de Saúde.

Figura 4.11 – Percentual de Doenças Respiratórias 01/01/2019 - 15/08/2022.



Fonte: Elaborado pelo autor.

4.2 Inconsistência e Problemas com os Dados Abertos

Ao longo do desenvolvimento deste trabalho, foram encontradas particularidades nas fontes oficiais e não oficiais, algumas gerando dificuldades para o estudo sendo elas:

Fonte Oficial: a base do Registro Civil possui valores nulos para alguns de seus indicadores e vale a pena ressaltar, que alguns são renomeados sem explicação, como por exemplo o tipo do sexo que é renomeado como indefinido.

Fonte Não Oficial: a base do Brasil.io alerta em seu manual de dados a existência desses possíveis problemas, como o fato de possuir valores negativos, que são usados para corrigir erros de casos contabilizados anteriormente, além de conter campos nulos, branco ou preenchido com um valor padrão. Como exemplo, o nome do município que pode estar em branco, nulo ou preenchido como "Importados/Indefinidos", possuindo também linhas que contêm a contagem total de óbitos pertencentes a algum estado. A base teve sua última atualização feita em 27/03/2022, desse modo impedindo a utilização do ano de 2022 nas comparações.

Outro ponto que vale a pena destacar e o fato de que, cada base consegue responder consultas distintas, ou seja, alguns de seus indicadores pertencem apenas a ela ou possuem uma granularidade específica em suas dimensões, impossibilitando assim uma comparação direta. Algumas das questões de pesquisa foram respondidas apenas utilizando a base do Registro Civil.

5 Conclusão

Foi concluído que, apesar das dificuldades, passadas pelos setores públicos de saúde, os números de divergência entre os dados disponibilizados oficialmente e não oficialmente diferem sim, mas em um grau não tão impactante. Em alguns momentos os dados oficiais apresentaram taxas de óbitos inferiores, mas com o passar do tempo nota-se uma normalização desses valores, podendo assim inferir que as os dados oficiais sobre a COVID-19 no Brasil estão bem próximos dos dados não oficiais divulgados.

Sobre os trabalhos relacionados que foram destacados na Seção 3, esse estudo possibilitou comparar os dados que são disponibilizados por órgãos públicos visando tornar cada vez mais acessíveis informações públicas. Dessa forma, o acesso ao Data Mart desse estudo ficará disponível para acesso público em um repositório do *Git Hub* através do link: *https://github.com/Samuelssj/Data-Mart-COVID-19-TCC*

Por fim, destaca-se que o trabalho cria e implementa um Data Mart para análise comparativa entre as duas fontes de dados, a fim de comparar como as informações estavam sendo distribuídas para a população e se seu conteúdo estava condizente com a real situação, ou seja, se manteve a qualidade dos dados. Assim o trabalho pode ser adaptado a várias áreas de estudo, onde a ferramenta OLAP possa atender trabalhos mais específicos ao utilizar novas bases de dados.

Como trabalhos futuros, nota-se que é possível um melhoramento do Data Mart para a ampliação dos resultados, como, por exemplo, a inclusão de novos dados que permitam outros tipos de análises como dados de vacinação, atendimento hospitalar, equipamentos de saúde etc.

REFERÊNCIAS

BRASIL, C. de Registro Civil do. *Portal de Transparência do Registro Civil*. 2012. https://transparencia.registrocivil.org.br/dados-covid-download. Repósitorio de Dados estatísticos sobre Óbitos.

INMON, W. *Definition of a Data Warehouse. 1999*. 2003. Boletins das Secretarias Estaduais de Saúde (SES).

JUSTEN Álvaro. *Brasil I.O*. 2019. https://brasil.io/dataset/covid19/files/. Boletins das Secretarias Estaduais de Saúde (SES).

KIMBALL, R. *The data warehouse toolkit: practical techniques for building dimensional data warehouses*. [S.l.]: John Wiley & Sons, Inc., 1996.

KIMBALL, R.; ROSS, M. *The data warehouse toolkit: the complete guide to dimensional modeling*. [S.l.]: John Wiley & Sons, 2011.

MOREIRA, P. et al. Covid data analytics: Repositório de dados provenientes de multiplas fontes sobre a pandemia de covid-19 no brasil. p. 107–116, 2021.

OMS. *world health organization*. 2022. https://covid19.who.int/. Acessado em 06 de Abril de 2022.

SILVA, L. V. e et al. Covid-19 mortality underreporting in brazil: analysis of data from government internet portals. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 22, n. 8, p. e21413, 2020.