Análise de padrões regionais e temporais de mortalidade no Brasil apoiado por um *data warehouse*

Débora Oliveira Santana

Trabalho de conclusão de curso apresentado ao
Instituto de Matemática e Estatística da
Universidade Federal da Bahia
para obtenção do título de
Especialista em Ciência de Dados e Big Data

Orientador: Prof. Esp. Juracy Araújo de Almeida Junior

Salvador, 02 de Dezembro de 2022

Análise de padrões regionais e temporais de mortalidade no Brasil apoiado por um *data warehouse*

Débora Oliveira Santana[1]
Juracy Araújo de Almeida Junior[2]

Resumo

Nos últimos anos observou-se no Brasil e no mundo um processo de transição demográfica e epidemiológica caracterizado por uma pandemia que proporcionou o aparecimento de altos níveis de mortalidade, aumentando as demandas de saúde e representando um grande desafio para a saúde pública. As informações sobre mortalidade durante os anos de 2016 a 2020 somam mais de 6 milhões de óbitos registrados no Sistema de Informação sobre e Mortalidade (SIM). A construção de uma estrutura de *data warehouse* se faz necessária para consolidar os dados, manter a qualidade e a precisão, além de facilitar as transações das informações. Essa estrutura unida a uma técnica de clusterização de dados é um conjunto eficiente de ferramentas tecnológicas para a identificação de padrões em grandes massas de dados. A implementação dessas ferramentas junto ao grande volume de informações do SIM pode ajudar na compreensão do panorama amplo da mortalidade no país, subsidiando o desenvolvimento de políticas de saúde pública que possibilitem uma gestão mais efetiva do Sistema Único de Saúde. O objetivo desta pesquisa é desenvolver uma base de dados multidimensional e realizar uma mineração de dados para reconhecer alguns padrões de mortalidade brasileira no período entre 2016 e 2020 utilizando dados do SIM e das Estimativas Populacionais do IBGE. A construção foi realizada em cinco etapas. Na primeira foram identificados os dados da base do Sistema de Informação sobre Mortalidade dos anos de 2016 a 2020. Na segunda foi construída a arquitetura de *data warehouse*. Na terceira foram elaborados os *dashboards* para visualização e análise dos dados. Na quarta foi desenvolvido o algoritmo de clusterização. Na quinta foram identificadas as respostas a partir de perguntas levantadas com o objetivo de analisar padrões de mortalidade durante esses cinco anos. Os resultados mostram que os padrões se mantiveram ao longo dos anos em relação a regiões, CID, grupo etário e raça/cor, com exceção do período da pandemia de Covid-19, que impactou fortemente na estrutura etária da população.

Palavras chaves: Mortalidade, *Data warehouse*, Clusterização de dados.

1 Introdução

Nos últimos anos, o Brasil sofreu uma modificação significativa no perfil de mortalidade com a doença causada pelo novo coronavírus e, historicamente, os padrões de mortalidade no Brasil são marcados por desigualdades regionais. Em relatórios passados, a Organização Pan Americana de Saúde PAHO (2020) evidencia a necessidade de redução das desigualdades na integração das questões transversais (equidade, gênero, etnia e direitos humanos) pois

[1]Discente, *debbora.os@gmail.com*
[2]Docente, *juracyajr@gmail.com*

isso influencia diretamente na saúde pública mundial e, consequentemente, nos índices de mortalidade.

Os dados abertos de saúde são de suma importância para a construção de uma sociedade mais transparente que possibilite o desenvolvimento de análises que apoiam as tomadas de decisões dos órgãos competentes para criar políticas públicas eficazes e capazes de solucionar os problemas de saúde da população como, por exemplo, a redução da mortalidade.

Com o objetivo de construir uma base de dados dados quantitativos e qualitativos sobre os óbitos ocorridos no Brasil, o Sistema de Informação Sobre Mortalidade (SIM), desenvolvido pelo Ministério da Saúde em 1975, é considerado uma ferramenta importante para a gestão na área da saúde que auxilia a tomada de decisão nas múltiplas áreas da assistência à saúde. A base de dados possui variáveis que permitem, a partir da causa de morte, construir indicadores e processar análises epidemiológicas que contribuam para a eficiência da gestão em saúde (DASNT, 2022).

A PAHO (2020) acredita que algumas ações reduzirá explicitamente as desigualdades nos índices de mortalidade mundial como a) deter a propagação do Covid-19 e diminuir o seu impacto; b) promover e avançar para a saúde universal com base na atenção primária à saúde; c) avançar na prevenção, controle e eliminação das doenças transmissíveis; d) reforçar a preparação e resposta às ameaças à segurança humana; e) focar no fortalecimento das intervenções ao longo da vida; f) adotar abordagens inovadoras e abrangentes para prevenção e controle de DCNT's, saúde mental e condições neurológicas; g) passar para a transformação digital e sistemas de informação dinâmicos para a saúde e uso eficaz da informação; h) combater os determinantes sociais e outros da saúde, proteger as populações vulneráveis e responder às suas necessidades.

Com isso, o objetivo desta pesquisa consiste em desenvolver uma base de dados multidimensional se utilizando da arquitetura proposta por Kimball and Ross (2011) e realizar uma mineração de dados para reconhecer alguns padrões de mortalidade brasileira. A ideia fundamental da modelagem dimensional baseia-se no fato de que pode ser representado como uma espécie de cubo de dados, onde as células do cubo contêm os valores medidos e os lados do cubo definem as dimensões naturais dos dados (Harrison, 1998). Já a mineração de dados utiliza algumas técnicas como, por exemplo, a de reconhecimento de padrões para conseguir extrair informações de grandes bases de dados.

Essa base de dados multidimensional e a mineração desses dados têm como objetivo oferecer uma visão do panorama de mortalidade no Brasil e da influência do contexto demográfico e temporal nos padrões e tendências de mortalidade para apoiar na identificação e decisão das ações necessárias para reduzir a mortalidade nas regiões brasileiras. Para isso, buscamos as seguintes respostas:

1. Qual a evolução da mortalidade no Brasil ao longo de 5 anos?

2. Qual a taxa de mortalidade (total de óbitos de residentes / população residente) nas regiões brasileiras?

3. Quais os cinco CID's com maior incidência de mortalidade no Brasil? Essas posições se mantêm ao longo dos anos da série histórica? Quanto representam do total?

4. O que é possível inferir sobre o comportamento dos dados da série histórica nas regiões brasileiras?

5. Qual o comportamento dos dados com relação às características do indivíduo (gênero, raça/cor, escolaridade, ocupação, grupo etário)?

6. Quais os locais de ocorrência em que há maior incidência da mortalidade?

7. Quais os principais grupos de fatores (CID, região, grupo etário, raça/cor) que influenciam diretamente no óbito de um indivíduo?

Essa pesquisa está estruturada em 5 capítulos. Além desse capítulo de introdução, o capítulo 2 apresenta a fundamentação teórica sobre a mortalidade no Brasil, o sistema de informação sobre mortalidade e a modelagem de *data warehouse*. O capítulo 3 corresponde à metodologia utilizada. Encontram-se descritos no capítulo 4 os resultados e discussões e, por fim, o capítulo 5 relata as conclusões.

2 Fundamentação Teórica

A primeira etapa deste trabalho foi dedicada ao levantamento bibliográfico sobre o panorama da mortalidade no Brasil, o sistema de informação sobre mortalidade, o que é um *data warehouse* e o que é a mineração de dados, mais especificamente, os algoritmos de clusterização.

2.1 Mortalidade no Brasil

No Brasil, os perfis de morbidade (conjunto dos indivíduos que adquirem doenças num dado intervalo de tempo em uma determinada população) e mortalidade (conjunto dos indivíduos que morreram num dado intervalo do tempo) sofreram alterações ao longo dos anos. Fatores como o crescimento da morbidade das doenças crônicas não transmissíveis, o difícil controle de doenças transmissíveis e as causas externas como a alta ocorrência de acidentes e violências se concentram, principalmente, nos grupos mais vulneráveis do ponto de vista social e econômico visto a desigualdade social.

Uma das principais características do processo de transição epidemiológica é o aumento da predominância das doenças crônicas não transmissíveis que vem se espalhando rapidamente pelo Brasil desde a década de 1960. Algumas doenças são ainda mais frequentes a partir dos 60 anos, destacando-se as doenças osteoarticulares, hipertensão arterial sistêmica, doenças cardiovasculares, diabetes mellitus e doenças respiratórias crônicas (Pereira *et al.*, 2015). Já as doenças transmissíveis impactaram significativamente o perfil de morbimortalidade desde 1980 pois são capazes de produzir emergências de saúde pública com elevado custo sanitário, social e econômico.

Segundo Martins *et al.* (2021), a transição epidemiológica brasileira continua caracterizada por uma tripla carga de doenças, em que altas taxas de morbimortalidade por doenças crônicas não transmissíveis, coexistem com uma elevada incidência e prevalência de doenças infecto-parasitárias e de causas externas, com destaque para os homicídios. Tais peculiaridades da transição epidemiológica brasileira são um sério desafio à melhoria das condições de vida e de saúde da população do país e precisam ser levadas em consideração no desenho de políticas públicas de saúde.

Os processos acelerados de urbanização e de envelhecimento da população, exigem que o Sistema Único de Saúde (SUS) realize análises e planejamentos para identificar as melhores estratégias que garantam o acesso efetivo da população ao sistema e incidam diretamente nos determinantes sociais e econômicos.

Com o objetivo de identificar ações e estratégias que colaborem com a diminuição das taxas de morbimortalidade no Brasil, faz-se necessário conhecer os padrões da transição

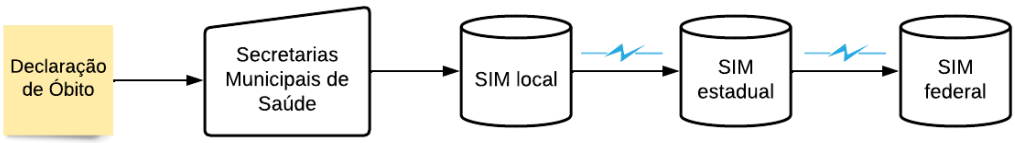
epidemiológica e suas tendências nos últimos anos. Além de não poder-se abster das desigualdades sociais e regionais que influenciam diretamente nesses fatores, é necessário identificar quais as epidemias recorrentes e novas, a negligência da população com as doenças crônicas não transmissíveis, o controle de doenças transmissíveis e a dimensão dos acidentes e da violência são de suma importância para identificar as ações necessárias para o controle epidemiológico brasileiro.

2.2 Sistema de Informação sobre Mortalidade

Em 1975 o Ministério da Saúde desenvolveu o Sistema de Informação sobre Mortalidade (SIM) com a finalidade de reunir dados quantitativos e qualitativos sobre óbitos ocorridos no Brasil. O SIM foi informatizado em 1979 e hoje é considerado uma ferramenta importante para a gestão na área da saúde que subsidia a tomada de decisão em diversas áreas da assistência à saúde.

A ingestão de dados no SIM é realizada a partir do documento de Declaração de Óbito (DO) que é o documento básico e essencial para a coleta de dados sobre mortalidade no Brasil. As DO's são preenchidas pelas unidades notificantes do óbito (habitualmente no local de ocorrência do óbito) e recolhidas pelas Secretarias Municipais de Saúde. Nas Secretarias Municipais de Saúde (SMS), as DO's são digitadas, processadas, criticadas e consolidadas no SIM local. Em seguida, os dados informados pelos municípios sobre mortalidade no nível local são transferidos para base de dados do nível estadual que os agrega e envia-os ao nível federal. Tais transferências são realizadas via web (internet) e ocorrem, simultaneamente, nos três níveis de gestão (DASNT, 2022).

Figura 1: Processo SIM



Fonte: Elaborada pela autora

Em 2005 a Organização Mundial de Saúde (OMS), após analisar sistemas de mortalidade de vários países, avaliou o SIM como um sistema de qualidade intermediária. Esse nível de qualidade também foi avaliado para os sistemas da França, Itália, Bélgica, Alemanha, Dinamarca, Rússia, Holanda, Suíça, entre outros que constituem o bloco dos países ricos. Após a avaliação, com o objetivo de aumentar essa qualidade, o Ministério da Saúde adotou iniciativas como redução da proporção de óbitos com causas mal definidas, desenvolvimento de novos aplicativos informatizados e administração de cursos de formação e de capacitação para codificadores de causas básicas.

Com os dados contidos no SIM sobre os óbitos registrados desde 1979 no Brasil é possível identificar indicadores epidemiológicos como instrumentos estratégicos de suporte ao planejamento das ações, atividades e programas voltados à gestão em saúde além de subsidiar a tomada de decisão em diversas áreas da assistência à saúde como, por exemplo, a redução da mortalidade por causas preveníveis ou evitáveis.

A Rede Interagencial para a Informação em Saúde (RIPSA) utiliza a base de dados do SIM para definir os Indicadores e Dados Básicos de Saúde (IDB) e associa-a também a outras

fontes como, por exemplo, ao Sistema de Informação Hospitalar. Segundo RIPSA (2008), esses indicadores básicos de saúde são medidas-síntese que contêm informação relevante sobre determinados atributos e dimensões do estado de saúde, bem como do desempenho do sistema de saúde. Vistos em conjunto, devem refletir a situação sanitária de uma população e servir para a vigilância das condições de saúde.

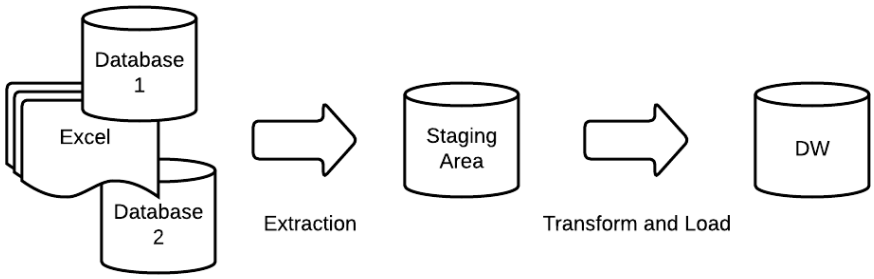
2.3 Data Warehouse

Um *data warehouse* consiste em um grande volume de informações provenientes de uma ou mais bases de dados que são tratadas, formatadas e consolidadas em uma única base de dados. Sua estrutura é desenvolvida de forma a melhorar o desempenho da consulta e análise desses dados para auxiliar na tomada de decisão. Segundo Inmon (2000), um DW pode ser definido como um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar decisões gerenciais.

Para a construção de um DW é necessário que realize-se o processo de ETL onde ocorre a extração, transformação e carregamento desses dados. Na extração é onde ocorre a comunicação com as bases de dados de origem (bancos de dados, API's, planilhas CSV, arquivos txt, dentre outros) para consumir os dados que serão inseridos na área temporária, chamada de *staging area*. Na transformação ocorrem as etapas de padronização, limpeza e qualidade desses dados que podem ou não ser oriundos de diferentes fontes. A etapa final é a de carga, nela os dados são lidos da *staging area* e carregados no DW.

A Figura 2 representa o funcionamento do processo de ETL. Com esse processo faz-se possível que os dados fiquem facilmente acessíveis, de forma consistente, com fácil adaptação às mudanças, acesso em tempo hábil, confiável e, normalmente, com atualização incremental de dados.

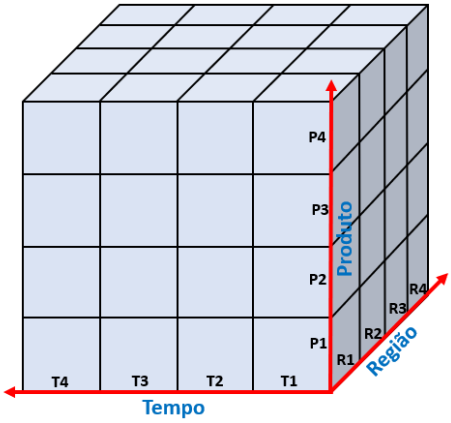
Figura 2: Processo ETL



Fonte: Elaborada pela autora

A modelagem multidimensional é a técnica de projeto lógico de banco de dados mais utilizada na concepção de um *data warehouse*. Diferente da modelagem relacional, ela busca simplificar a compreensão do usuário, melhorar o desempenho de consulta e facilitar as mudanças. Tem como representação uma espécie de cubo pois corresponde a matrizes multidimensionais, onde as células do cubo contêm os valores medidos e os lados do cubo definem as dimensões naturais dos dados, conforme exemplificado na Figura 3.

Figura 3: Cubo multidimensional

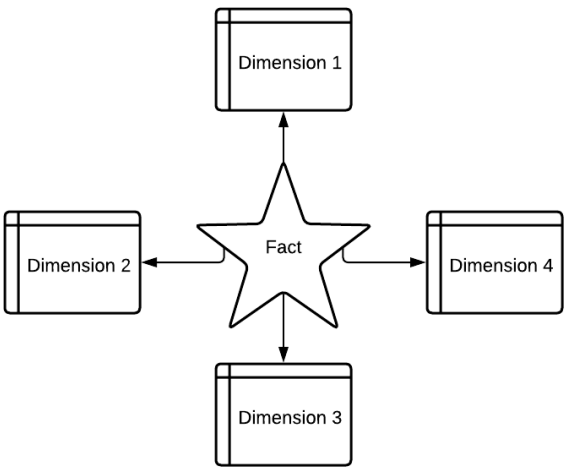


Fonte: estrategiaconcursos.com.br/blog/o-que-e-olap/

Um modelo multidimensional é composto por fatos e dimensões. As tabelas fatos armazenam as células do cubo, ou seja, as medidas de desempenho resultantes dos eventos do processo de negócios de uma organização. As tabelas de dimensão são complementares a uma tabela de fatos e contêm o contexto textual associado a um evento de medição de processo de negócios. Elas descrevem "quem, o quê, onde, quando, como e porquê"associados ao evento (Kimball and Ross, 2011). Existe também as dimensões degeneradas que possui como característica ser uma dimensão regular para a qual os dados de dimensão são armazenados nas tabelas de fatos. As dimensões degeneradas eliminam a necessidade de uma junção a mais com uma dimensão e reduz o espaço de armazenamento.

A estrutura mais utilizada em uma modelagem multidimensional é o esquema estrela. Sua simetria é simples e eficaz, trazendo benefícios de desempenho pois há menos junções. Essa estrutura, que pode ser vista na Figura 4, consiste em tabelas fatos cercadas por tabelas de dimensão.

Figura 4: Esquema estrela

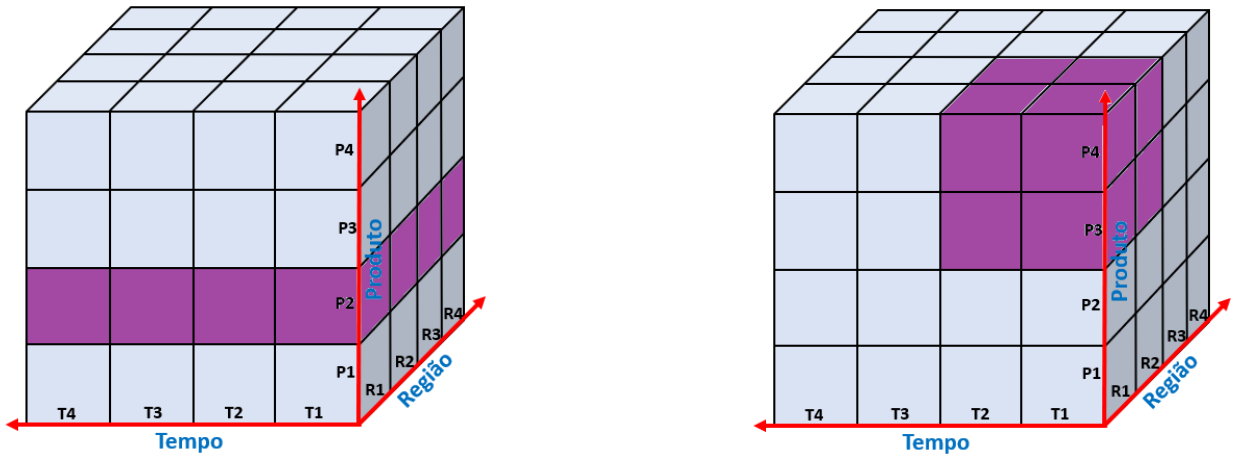


Fonte: Elaborada pela autora

A modelagem multidimensional nos permite realizar algumas operações como, por exemplo, *drill down* e *drill up* que estão relacionadas com a granularidade das informações. O

drill down diminui o nível de granularidade (ex.: ano - mês - dia) e o *drill up* aumenta o nível de granularidade (ex.: dia - mês - ano). Já o *slice* obtém uma fatia do cubo de dados (Figura 5a) e o *dice* obtém um subcubo de informações, ou seja, é a seleção de dois ou mais valores das dimensões (Figura 5b).

Figura 5: Representação de slice e dice



(a) Slice

(b) Dice

Fonte: estrategiaconcursos.com.br/blog/o-que-e-olap/

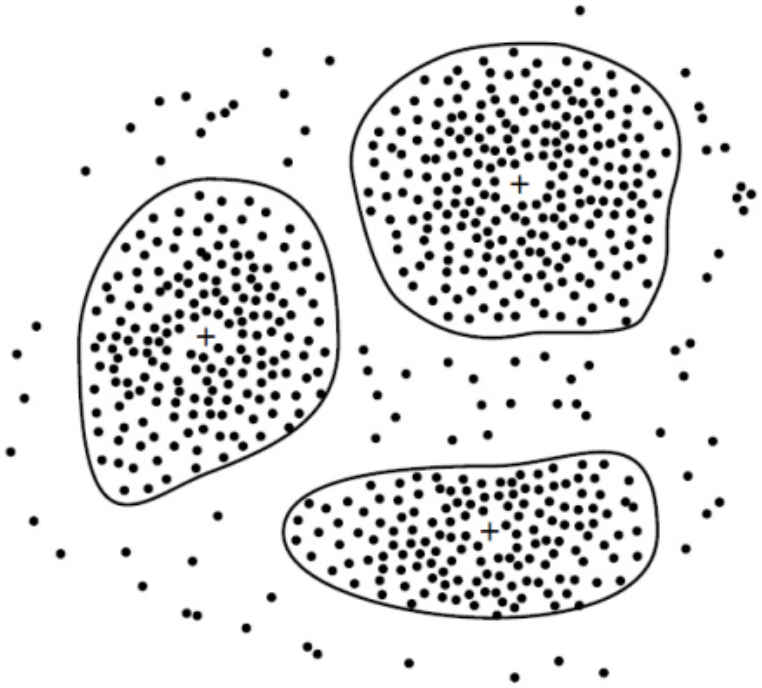
Para exibição das informações armazenadas em um DW são utilizados os *dashboards* que são exibições visuais de dados usados para monitorar os indicadores relevantes, facilitar o entendimento desses indicadores e auxiliar a tomada de decisão orientada a dados. Os *dashboards* podem apresentar as tendências de uma informação específica ou os relacionamentos de várias informações conectadas. Contudo, para que isso seja realizado com sucesso, faz-se necessária a construção de uma visualização de dados eficaz.

2.4 Mineração de Dados

A mineração de dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização para conseguir extrair informações de grandes bases de dados (Cabena *et al.*, 1998).

Há algumas formas de classificar os algoritmos de mineração de dados e a clusterização (ou agrupamento) é uma delas. Cassiano (2014) define a clusterização como uma técnica de mineração de dados multivariados que através de métodos numéricos e informações das variáveis de cada caso, tem por objetivo agrupar por aprendizado não supervisionado (com base em observação e descoberta) os "n"casos da base de dados em "k"grupos, geralmente denominados clusters ou agrupamentos. Ou seja, é um algoritmo que agrupa dados e classifica-os em conjuntos (clusters) que se assemelham de alguma forma. Pode-se dizer que os clusters interessantes são aqueles que se isolam do resto e possuem uma alta densidade, o que significa que seus elementos são mais próximos.

Figura 6: Registros agrupados em três clusters



Fonte: Camilo, C. O., e Silva, J. C. D. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. UGG

O k-means é um algoritmo de clusterização que tem como objetivo encontrar "k"clusters diferentes no conjunto de dados. A tarefa do algoritmo é encontrar, por meio de alguma métrica de distância, o centro de cada cluster (centróide) mais próximo e atribuir o ponto encontrado a esse cluster. Após este passo, os centróides são atualizados sempre tomando o valor médio de todos os pontos naquele cluster. Para este método são necessários valores numéricos para o cálculo da distância, os valores nominais então podem ser mapeados em valores binários para o mesmo cálculo. Em caso de sucesso, os dados são separados organicamente podendo assim ser rotulados e os centróides viram referência para classificar novos dados (Honda, 2017).

O primeiro passo do k-means é definir um "k", ou seja, um número de clusters. Depois define aleatoriamente um centróide para cada cluster. O próximo passo é calcular, para cada ponto, o centróide de menor distância. Cada ponto pertencerá ao centróide mais próximo. Após reposicionar o centróide e fazer com que a nova posição do centróide seja a média da posição de todos os pontos do cluster, os dois últimos passos são repetidos, iterativamente, até obter a posição ideal dos centróides.

Contudo, o k-means usa apenas distâncias numéricas. Para agrupar os dados categóricos há, por exemplo, o algoritmo k-modes que é uma extensão do k-means. Em vez de distâncias, ele usa dissimilaridades (ou seja, quantificação do total de desencontros entre os objetos: quanto menor esse número, mais semelhantes são os os objetos). E em vez de meios, utiliza modos. Um modo é um vetor de elementos que minimiza as diferenças entre o próprio vetor e cada objeto dos dados. Teremos tantos modos quanto o número de clusters necessários, pois eles atuam como centróides. Já para os dados numéricos e categóricos, existe outra extensão desses algoritmos que combinam o k-means e k-modes e é chamado de k-protótipos.

3 Trabalhos Correlatos

Para desenvolvimento do objetivo da pesquisa, fez-se necessária uma busca inicial a fim de obter respostas para a seguintes questões:

1. Quais os padrões de mortalidade já identificados?
2. Como esses padrões se comportaram durante e após a pandemia de Covid-19?

Alves *et al.* (2017) teve como objetivo analisar padrões regionais e temporais da mortalidade no Brasil no período entre 1979 e 2013 por diversos grupos de causa e causas específicas. Para a explicação da questão 1, o autor observou tendências regionais onde a mortalidade por doenças infecciosas decaíram em todas as regiões e mortalidade por neoplasias e doenças do aparelho circulatório aumentaram no Norte e Nordeste e decaíram no Sul e Sudeste. Já o comportamento da mortalidade por causas externas sobressaiu em algumas unidades da federação do Centro-Oeste, Norte e Nordeste onde as desigualdades sociais são mais intensas. Nessa pesquisa não foi possível identificar explicação para a questão 2 pois foi realizada no período anterior à pandemia de Covid-19.

Lasmar and Siviero (2018) tiveram como objetivo analisar o diferencial nos níveis e padrões da mortalidade entre 2000 e 2010 das macrorregiões brasileiras comparadas isoladamente com o país, separada por sexo e grupo etário. Para a explicação da questão 1, os autores obtiveram como resultado que o Brasil apresentava uma queda contínua da mortalidade, porém esse declínio não impactou diretamente nos padrões. Ou seja, a mortalidade entre os homens jovens continuou apresentando picos elevados devido aos óbitos por causas externas. Já a mortalidade feminina aumentou gradativamente no decorrer da sua faixa etária. Comparando as macrorregiões brasileiras, os autores identificaram que o Norte e Nordeste mantiveram abaixo da expectativa de vida ao nascer do país para ambos os sexos e o Sudeste e Sul com expectativas de vida ao nascer superiores aos observados no Brasil. Nessa pesquisa também não foi possível identificar explicação para a questão 2 visto que foi realizada no período anterior à pandemia de Covid-19.

Demenech *et al.* (2020) tiveram como objetivo avaliar, com base em análises espaço temporais, se existe relação entre desigualdade econômica e infecção e morte por Covid-19 nas Unidades Federativas do Brasil. Para a explicação da questão 1, os autores utilizaram de literaturas anteriores para justificar que existem diferenças relacionadas à desigualdade social tanto em outros países com as epidemias de H1N1, SARS e ebola, quanto no Brasil com dengue, tuberculose e HIV. Para a explicação da questão 2, a partir das bases de dados públicas, os autores conseguiram concluir que estados mais desiguais apresentaram progressão mais acentuada na taxa de mortalidade por Covid-19, enquanto entre os menos desiguais ocorreram aumentos sutis. Além disso, a pesquisa evidencia que o risco de morrer por Covid-19 pode ser até 10 vezes maior entre indivíduos residentes de bairros mais vulneráveis da mesma cidade e que pessoas que se identificam como negras possuem maior chance de serem vítimas do vírus.

4 Metodologia

Como produto final desta pesquisa, temos o conjunto de dados, as análises realizadas e os *dashboards* necessários para a visualização das informações. Para o desenvolvimento desta etapa foi realizada a extração manual dos dados na base do openDataSUS que disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde.

Após a extração, foi utilizada a ferramenta Pentaho Data Integration para o processo de transformação e carga. Já para a análise e visualização dos dados foi utilizada a ferramenta Power BI e, por fim, para identificar os grupos de fatores que influenciam diretamente no óbito de um indivíduo, utilizou-se a linguagem Python para implementar a clusterização que é uma técnica de aprendizado de máquina não supervisionado que visa agrupar os dados em determinados conjuntos distintos entre si, conforme Figura 7.

Figura 7: Arquitetura da Pesquisa

Fonte: Elaborada pela autora

4.1 Conjunto de Dados

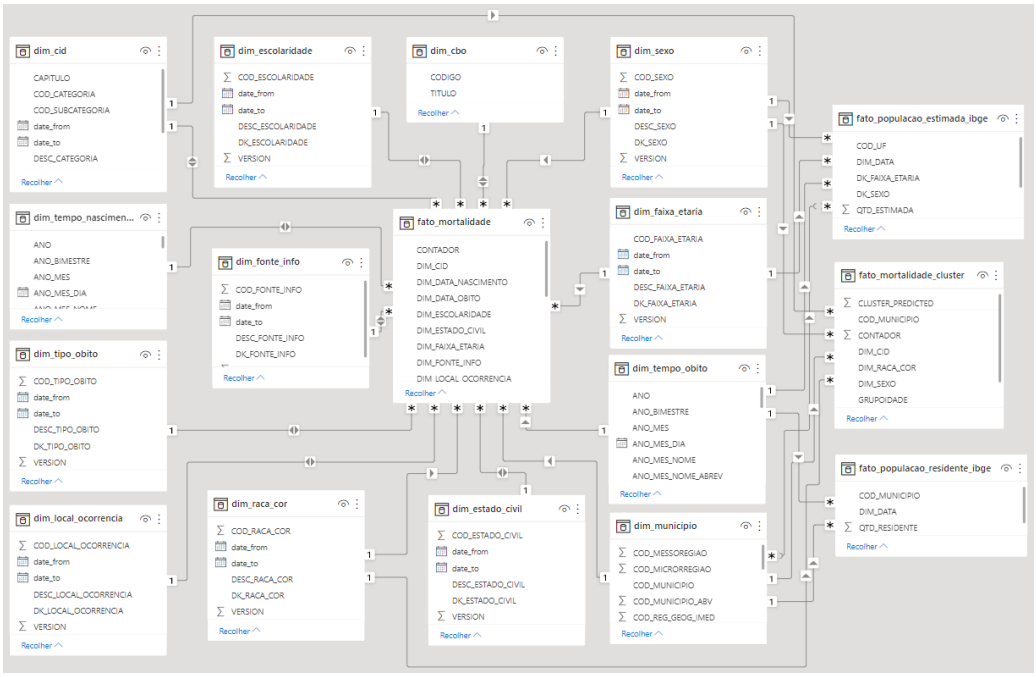
A estrutura de modelagem proposta neste trabalho é a baseada no esquema estrela. Para definir os fatos para a construção do *data warehouse* foi feita uma análise dos principais atributos da estrutura de dados do SIM que nos auxilia nas respostas das questões elencadas durante o período correspondente aos anos de 2016 a 2020. As dimensões foram definidas de maneira a aumentar a performance das consultas. Através dessa análise foi possível extrair uma tabela fato principal que determina o esquema proposto e ela foi nomeada de mortalidade.

Na Figura 8 é possível observar a tabela fato mortalidade, suas dimensões degeneradas e suas relações com as tabelas de dimensões. Os dados que compõem a tabela fato de mortalidade são provenientes do SIM, que registra as informações de mortalidade no Brasil.

As dimensões degeneradas estão relacionadas com algumas informações pessoais do falecido e características específicas do óbito. No caso desses campos, não seria prático ter uma lista *drop-down* de cada um para encontrar dados normalmente encontrados na base do SIM. Por conseguinte, estes itens foram colocados na tabela de fato para usos do tipo *slice* e *dice*, conforme a necessidade da análise.

As tabelas de dimensões possuem características que informam o Código Internacional de Doença (CID) registrado na Declaração de Óbito (DO), escolaridade do falecido, a ocupação do falecido conforme Classificação Brasileira de Ocupações (CBO), a fonte de informação utilizada para o preenchimento dos dados, o município em que ocorreu o óbito, estado civil do falecido, o local de ocorrência do óbito, o tipo do óbito, a data em que ocorreu o óbito, a data de nascimento do falecido e a raça/cor informada do indivíduo falecido.

Figura 8: Modelo de data warehouse



Fonte: Elaborada pela autora

A tabela fato de população estimada tem como fonte a tabela de projeção da população por sexo e idade do IBGE (SIDRA, 2018). Já a tabela fato de população residente tem como fonte a tabela de estimativas de população residente dos municípios do IBGE (SIDRA, 2020). Ambas as fontes foram necessárias para buscar os dados do total da população para que fosse possível realizar o cálculo da taxa de mortalidade que corresponde a quantidade de óbitos dividido pela quantidade residente da população.

A tabela fato de mortalidade cluster contém os dados que foram gerados a partir da aplicação do algoritmo de clusterização na base de dados do SIM.

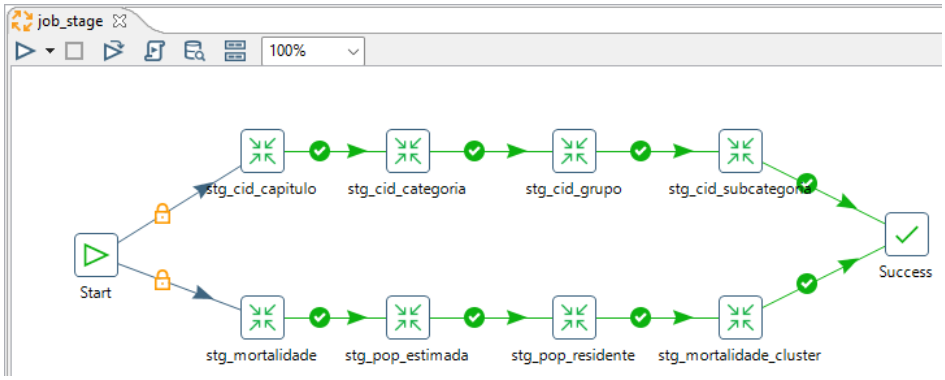
4.2 ETL no Pentaho Data Integration

Para a construção e manipulação da arquitetura de dados que carrega e atualiza o conjunto de dados desenvolvido, foi utilizada a ferramenta Pentaho Data Integration (PDI) que é o componente da suíte Pentaho responsável por criar processos de extração, transformação e carga que alimentam o banco de dados.

No Pentaho Data Integration é possível criar transformações e *jobs*. As transformações operam sobre os dados das tabelas, elas são o registro do passo-a-passo de como a extração ou leitura de uma fonte de informação é realizada. Uma transformação pode conter vários artefatos, são eles: leitura de dados, seleção de campos específicos, concatenação de valores de dois campos distintos, divisão de valores contidos em um único campo gerando dois ou mais novos campos ou linhas, *merge* de dados contidos em bancos de dados diferentes, *merge* de dados originados de diversas extensões de arquivo, aplicação de expressões regulares em texto para limpeza, dentre outros. Já um *job* é uma sequência de operações e transformações com o objetivo de automatizar alguma tarefa.

Após a criação do repositório que será utilizado para armazenar os arquivos de *jobs* e transformações, foram criadas as transformações necessárias para o processo de criação das tabelas na área temporária ou *staging area* (Figura 9).

Figura 9: *Job* de *staging area* no Pentaho

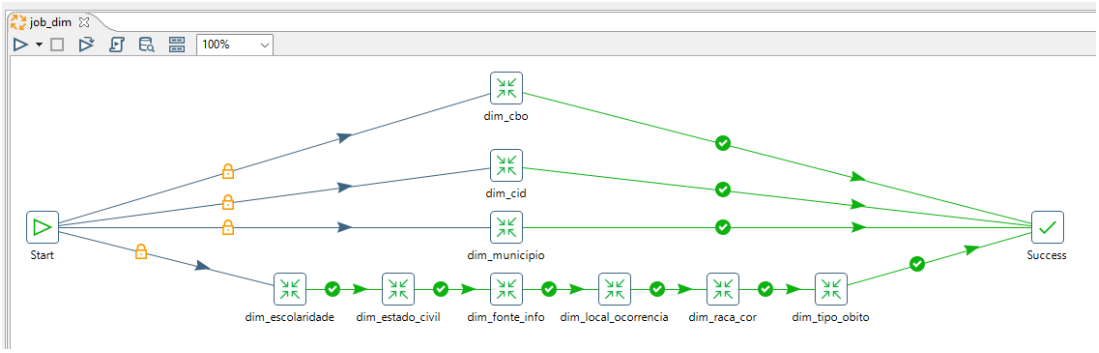


Fonte: Elaborada pela autora

Na base de dados do SIM, o CID é um campo que corresponde a um código único para cada óbito conforme padronizado pela OMS. O CID é dividido em 22 capítulos (que agrupam doenças com características semelhantes), categorias (representadas por uma letra e dois dígitos) e subcategorias (representadas por um número de 0 a 9) e, portanto, fez-se necessária a busca dessa divisão na fonte da OMS. Uma vez que as informações complementares de CID e as informações da base de mortalidade são oriundas de fontes distintas, foi possível a estruturação de uma carga paralela que auxilia na rapidez do carregamento das informações.

O *job* criado para carregar as dimensões, representado na Figura 10, possui algumas transformações que correspondem às dimensões que foram identificadas a necessidade de controle das mudanças que ocorrem nos dados gravados no banco à medida que elas vão acontecendo. O CID, por exemplo, é uma informação que com o passar do tempo podem ocorrer algumas mudanças em sua classificação e, portanto, necessita que haja um registro histórico do atributo original com horários de início e fim daquele registro, indicando o período que o registro anterior era válido e a ativação do início do registro atual pois, ao ocorrer uma mudança nessa classificação, os óbitos registrados precisam ser identificados com a classificação CID que estava vigente naquele período.

Figura 10: *Job* de dimensão no Pentaho

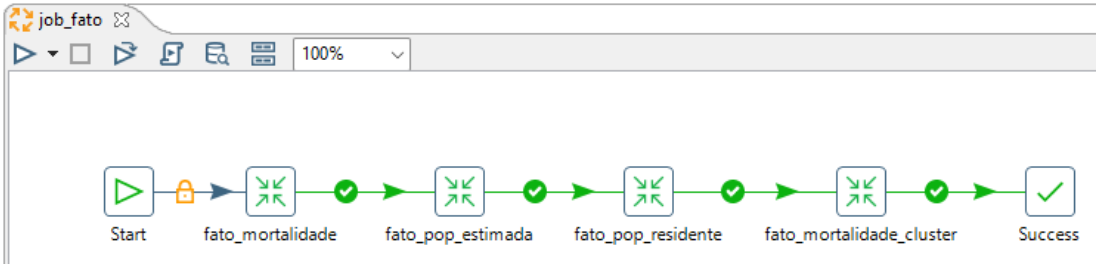


Fonte: Elaborada pela autora

As dimensões que possuem dados oriundos do SIM também possuem o componente de versionamento do Pentaho, pois foi identificado que pode-se incluir e excluir informações ao decorrer dos anos. Nas demais, esse histórico não é armazenado e quando ocorre alguma atualização dos dados, há uma sobreposição das informações.

O *job* criado para carregar a tabela fato (Figura 11) possui a transformação referente ao carregamento da tabela principal e os tratamentos necessários. As dimensões degeneradas partem da tabela de origem identificadas com um código e é nessa transformação que realiza-se o tratamento para identificá-las com uma descrição que será utilizada futuramente no *dashboard* construído. Além disso, é nela que ocorrem também as junções com as chaves das dimensões para construção da relação entre as tabelas do DW.

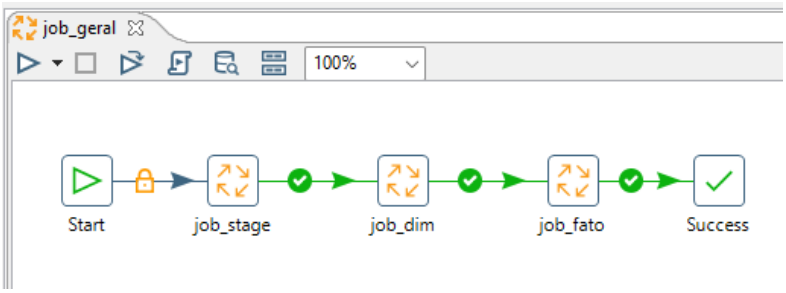
Figura 11: *Job* da tabela fato no Pentaho



Fonte: Elaborada pela autora

Por fim, o *job* geral (Figura 12) é o que inicia a execução (manual ou via batch) e possui a chamada dos demais *jobs* respeitando a ordenação das informações que precisam ser carregadas.

Figura 12: *Job* geral no Pentaho



Fonte: Elaborada pela autora

Após essa modelagem construída no Pentaho, utilizando os conceitos descritos nesta pesquisa, é possível obter um ganho de qualidade na modelagem desse projeto de *data warehouse* e que reflete em ganhos de performance e qualidade dos dados nas análises e na criação dos *dashboards*.

4.3 Visualizações no Power BI

Para a visualização e análise dos dados, foi utilizada a ferramenta Power BI que é uma ferramenta da Microsoft com o propósito de criar visualizações de dados avançadas e interativas por meio de várias fontes de dados e compartilhar importantes *insights* de negócios.

É de suma importância que se desenvolva *dashboards* que possuam um alto impacto visual e que seja fácil para o usuário tomar decisões. Knaflic (2019) recomenda que o público alvo seja identificado, que o desenvolvedor saiba qual o objetivo de comunicação desejado, quais os problemas busca-se resolver, quais as representações visuais que conseguem comunicar precisamente a mensagem ao público alvo, além de como focalizar a atenção do público alvo

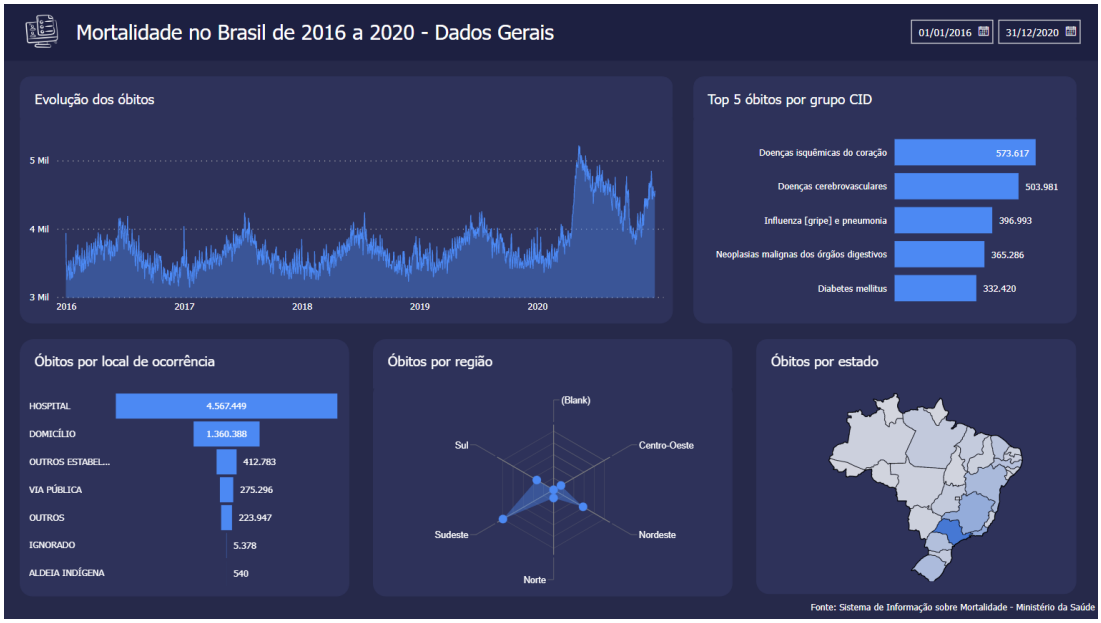
para o que é mais relevante eliminando a carga cognitiva excessiva e utilizando os atributos pré-atentivos (tamanho, cor e posição) para alcançar o objetivo de melhorar a comunicação visual.

Com o objetivo de responder às questões da presente pesquisa e aplicar as melhores práticas de visualização de dados, algumas análises e visualizações foram construídas, conforme visto Figura 13, que corresponde ao principal *dashboard* desenvolvido.

Foi utilizado o gráfico de área para visualizar as tendências e movimentos dos óbitos ao longo do tempo, pois facilita o entendimento e faz-se possível a observação imediata das tendências. Para as informações de óbito por região, foi utilizado o gráfico de radar pois eles se assemelham muito a como nossos cérebros produzem e processam ideias de maneira não linear facilitando assim a identificação rápida de semelhanças, diferenças e discrepâncias das regiões.

Para visualizar os óbitos por estado, nada melhor que apresentá-los em seus respectivos locais cartográficos. Já para identificar as diferenças quantitativas e percentuais dos locais de ocorrência dos óbitos, foi utilizado o gráfico de funil pois facilita a revelação dos afunilamentos das informações. E, por fim, nesse *dashboard* principal foi utilizado o gráfico de barras para apresentar as informações dos cinco maiores quantitativos de óbitos por grupo CID, pois facilita a comparação das cinco categorias.

Figura 13: *Dashboard* principal desenvolvido



Fonte: Elaborada pela autora

Os demais *dashboards* desenvolvidos foram construídos utilizando os mesmos conceitos de melhores práticas de visualização de dados.

4.4 Clusterização em Python

Para identificar padrões nos dados de mortalidade brasileira, foi utilizada a linguagem Python que é uma linguagem de programação de alto nível no aplicativo *open-source* Jupyter Notebook que oferece um ambiente para desenvolvimento online.

Alguns módulos foram utilizados para realizar o pré-processamento de dados, a exploração de dados com análise de dados explicativa e o algoritmo de agrupamento k-modes.

O módulo pandas foi utilizado para manipular os dados, o módulo numpy para cálculos de álgebra linear, os módulos plotnine e matplotlib para visualização de dados e, por fim, o módulo k-modes para a aplicação do algoritmo de agrupamento de variáveis categóricas.

Foram realizadas algumas etapas para aplicação do k-modes na base de dados. A primeira delas foi referente à limpeza das variáveis que não eram o objetivo da pesquisa para a criação dos clusters. As variáveis alvo foram CID, município, data de nascimento, data de óbito, raça e sexo. Conforme característica do k-modes, fez-se necessária a identificação das variáveis que ainda não eram do tipo categórica para convertê-las. Nesse caso, apenas a variável de município foi alterada para o tipo categórica.

A idade do indivíduo que veio a óbito também era uma variável alvo da criação dos clusters e, por esse motivo, foi realizado o cálculo da idade a partir dos campos de data de nascimento e data de óbito e, após o cálculo, foi realizada a exclusão desses dois campos no *dataframe*. Já as variáveis sexo e raça foram categorizadas conforme dicionário de dados da base e a variável idade foi categorizada conforme Figura 14.

Figura 14: Código no Jupyter Notebook da classificação etária

```
In [22]: grupo_idade = []

for i in df['IDADE']:
    if i < 13:
        grupo_idade.append('Crianca')
    elif i >= 13 and i < 19:
        grupo_idade.append('Adolescente')
    elif i >= 19 and i < 60:
        grupo_idade.append('Adulto')
    else:
        grupo_idade.append('Idoso')

df['GRUPOIDADE'] = grupo_idade
df
```

Fonte: Elaborada pela autora

Para melhorar a etapa exploratória dos dados, foi realizada a identificação dos dados ausentes e, como eram apenas 4% do total (3% na coluna raça e 1% na coluna idade - Figura 15), a decisão realizada foi de exclusão desses dados pois não era possível realizar uma correlação dos mesmos e o impacto seria insignificante.

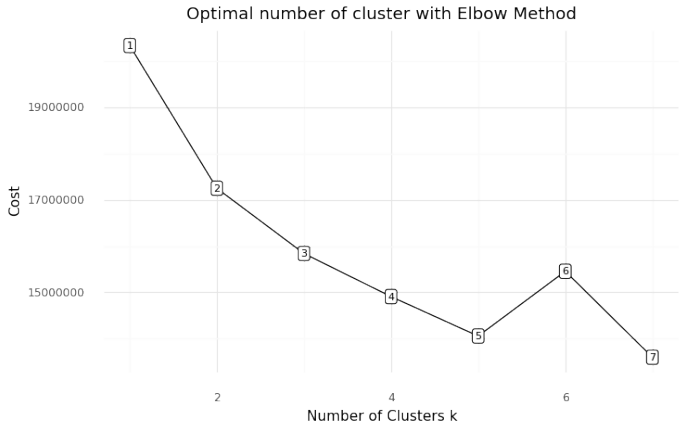
Figura 15: Código no Jupyter Notebook da análise de dados ausentes

```
In [15]: round(df.isnull().mean() * 100,2)

Out[15]: CAUSABAS      0.00
         CODMUNOCOR    0.00
         RACACOR       3.00
         SEXO          0.00
         IDADE         1.32
         dtype: float64

In [16]: print('Total de idivíduos no dataset {}'.format(df.shape[0]))
         print('Total de indivídios com todos os registros preenchidos (linhas) {}'.format(df.dropna().shape[0]))
         print('Percentual de dados com 100% do preenchimento dos dados {}'.format(round(df.dropna().shape[0] / len(df)*100,2)))

         Total de idivíduos no dataset 6845781
         Total de indivídios com todos os registros preenchidos (linhas) 6556646
         Percentual de dados com 100% do preenchimento dos dados 95.78
```

Fonte: Elaborada pela autora

Após as etapas de exploração e limpeza de dados, o *dataframe* foi convertido em matriz para o processamento do k-modes. Para a identificação do "k"ótimo foi utilizado o método do cotovelo (elbow) que plota a variação em função do número de clusters e o ponto da curva que tem a forma de um cotovelo é o "k"ótimo (Figura 16).

Figura 16: Número ótimo de clusters com Método Elbow



Fonte: Elaborada pela autora

A etapa de identificar o "k"ótimo demorou, em média, duas horas para definir a melhor quantidade de clusters que podem ser encontrados nessa base de dados com 6.845.781 registros e, após esse tempo, identificou-se que o "k"ótimo foi o número cinco, ou seja, foram criados cinco clusters. E, por fim, foi realizada a criação do *dataframe* que contém os dados dos clusters identificados e a combinação dos clusters com o *dataframe* original a partir dos centróides (Figura 17) para fazer as análises e interpretações dos clusters.

Figura 17: Combinação dos clusters com o dataframe original

```
In [56]: combinedDf
```

Out[56]:

	CAUSABAS	CODMUNOCOR	RACACOR	SEXO	GRUPOIDADE	cluster_predicted
0	R99	120010	Parda	Masculino	Idoso	4
1	K869	120010	Parda	Feminino	Idoso	1
2	K721	120010	Parda	Feminino	Adulto	1
3	B342	120010	Parda	Feminino	Idoso	1
4	J969	120010	Parda	Feminino	Idoso	1
...	...	...	...	...	...	...
6556641	I64	172100	Preta	Masculino	Idoso	0
6556642	E149	172100	Parda	Feminino	Idoso	1
6556643	I694	171610	Parda	Feminino	Idoso	1
6556644	X954	170210	Parda	Masculino	Adulto	4
6556645	I219	171700	Preta	Feminino	Adulto	2

6556646 rows × 6 columns

Fonte: Elaborada pela autora

5 Resultados

Com o objetivo de identificar os padrões de mortalidade brasileira, realizou-se nos capítulos anteriores o tratamento dos dados e construção das análises a fim de responder às perguntas listadas neste capítulo da pesquisa.

5.1 Evolução da mortalidade no Brasil ao longo de cinco anos

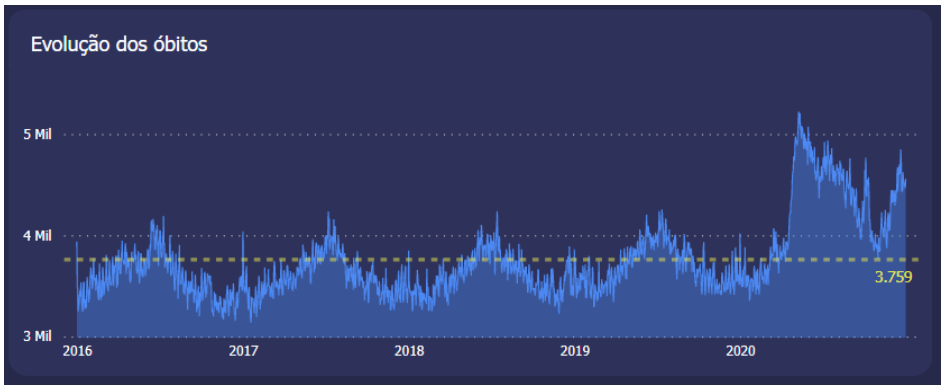
A análise da evolução da mortalidade brasileira possibilita identificar as mudanças no perfil epidemiológico da população brasileira através dos acontecimentos na história. Ao longo

de cinco anos, a mortalidade no Brasil teve como média 3.759 óbitos. Pode-se observar o aumento significativo dos óbitos em março de 2020, logo após a primeira morte anunciada de Covid-19 e, até o final desse mesmo ano, esse quantitativo ainda não tinha se normalizado (Figura 18).

Ao realizar o recorte dos dados apenas do período de março de 2020 até dezembro desse mesmo ano (Figura 19), é possível analisar que nesse período pandêmico, a média de óbitos aumentou em 17.5% se comparado ao mesmo período do ano de 2019 e corresponde a quase 15% a mais desse mesmo período nos anos de 2018 e 2019.

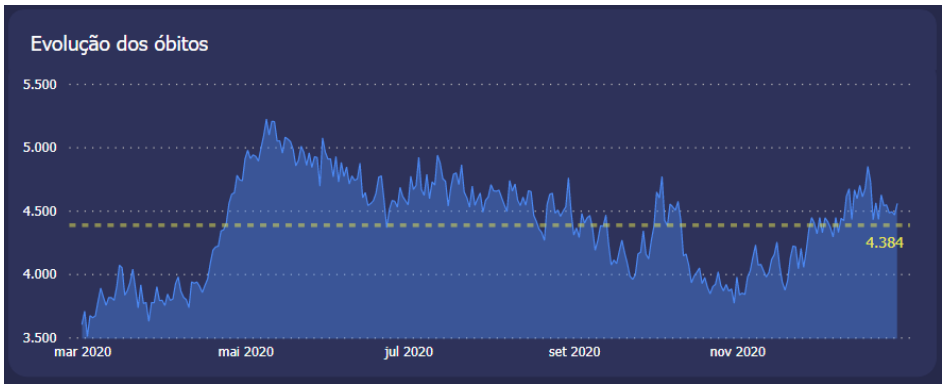
Ao longo desses cinco anos foi possível observar que, no período anterior à pandemia do Covid-19, a quantidade de óbitos estava com um aumento de, em média, 1.52% a cada ano. Após o período pandêmico do Covid-19, segundo a PAHO (2020), acredita-se que a intensificação dos programas de distribuição de renda e a melhoria das condições de vida poderão auxiliar na queda da mortalidade e no aumento da expectativa de vida da população brasileira.

Figura 18: Evolução dos óbitos de 2016 a 2020



Fonte: Elaborada pela autora

Figura 19: Evolução dos óbitos de março a dezembro de 2020



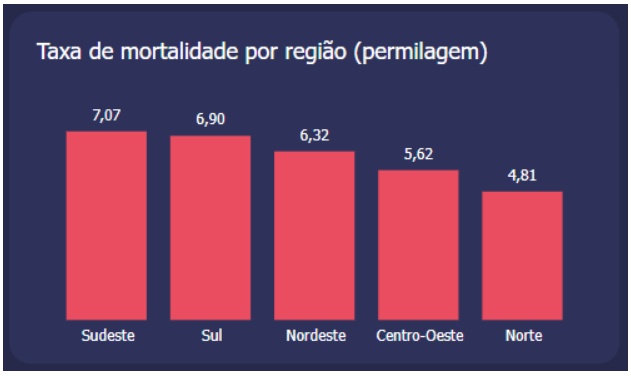
Fonte: Elaborada pela autora

5.2 Taxa de mortalidade nas regiões brasileiras

A taxa de mortalidade foi calculada a partir dos dados do IBGE sobre a população residente no Brasil da seguinte forma: (quantidade de óbitos / população residente) * 1000.

Nas regiões brasileiras, o Sudeste e o Sul aparecem como os líderes das maiores taxas de mortalidade ao longo de cinco anos e se mantiveram acima da média em todos os períodos. Logo atrás está a região Nordeste que se manteve nesse período com valores próximos à média. O Centro-Oeste e o Norte são os que possuem menores taxas de mortalidade e conseguiram se manter bem abaixo da média ao longo desses cinco anos.

Figura 20: Taxa de mortalidade por região



Fonte: Elaborada pela autora

Contudo, conforme os dados de sub-registros de óbitos disponibilizados pelo IBGE (2019) que corresponde ao conjunto dos eventos vitais não registrados no prazo legal previsto, foi possível identificar que o Norte e o Nordeste possuem as maiores taxas de sub-registros (Tabela 1). Isso pode ser argumentado por alguns fatores como as vulnerabilidades sociais e econômicas nessas regiões e as grandes distâncias entre as comunidades locais e os Cartórios de Registro Civil de Pessoas Naturais que resulta na dificuldade do acesso de alguns segmentos populacionais aos serviços.

Ano	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
2019	12,56%	8,00%	1,19%	0,91%	3,04%
2018	13,13%	8,87%	0,91%	1,16%	3,20%
2017	13,27%	9,06%	0,83%	1,23%	3,28%
2016	13,90%	9,47%	1,10%	1,19%	4,29%

Tabela 1: Percentual de sub-registros dos óbitos do IBGE por região

Portanto, quanto à cobertura e qualidade do preenchimento dos dados SIM analisados, destacam-se as regiões Norte e Nordeste que apresentaram os indicadores mais frágeis ao longo dos cinco anos. Essa qualidade depende tanto da organização do serviço de vigilância em saúde quanto do acesso da população aos serviços de assistência à saúde.

5.3 Cinco CID's com maior incidência de mortalidade no Brasil

Dentre o ranking de CID's com maior incidência no Brasil, têm-se as doenças isquêmicas do coração, as doenças cerebrovasculares, a influenza e pneumonia, as neoplasias malignas dos órgãos digestivos e a diabetes mellitus, conforme Figura 21. O somatório desse ranking corresponde a, aproximadamente, 32% dos óbitos totais ao longo de cinco anos.

Figura 21: Ranking dos óbitos por CID

Fonte: Elaborada pela autora

As doenças cardiovasculares mais prevalentes como causa de morte em todo o mundo são as doenças isquêmicas do coração (DIC). A falta de implementação dos programas de hipertensão de maneira eficiente e econômica no nível de atenção primária para identificar os indivíduos com maior risco de doença cardíaca e garantir intervenções personalizadas, acaba resultando em níveis elevados de doenças cardíacas e AVC.

A influenza e a pneumonia sempre foram preocupações para a saúde pública não só no Brasil, mas também no mundo. Elas manifestam-se mais fortemente nas crianças com menos de cinco anos, em idosos e há uma grande chance de levá-los a óbito. Os desafios de seu combate são o alto custo do tratamento e a elevada quantidade de internações decorrentes da doença.

As neoplasias malignas dos órgãos digestivos está no ranking dos CID's com maior incidência no Brasil devido ao fato de que o diagnóstico dessa doença geralmente é tardio pois a disfagia, que é o principal sintoma, manifesta-se apenas quando já há comprometimento de dois terços da luz do órgão (MARQUES *et al.*, 2014). Além disso, independente da terapêutica aplicada, a sobrevida média do indivíduo é de 4 a 6 meses.

Já a diabetes mellitus, que também lidera esse ranking, pode ser justificada devido o aumento das taxas de sobrepeso e obesidade associado às modificações no consumo alimentar da população brasileira - baixa frequência de alimentos ricos em fibras, aumento da proporção de gorduras saturadas e açúcares da dieta - associadas a um estilo de vida sedentário pois compõem os principais fatores etiológicos da diabetes mellitus.

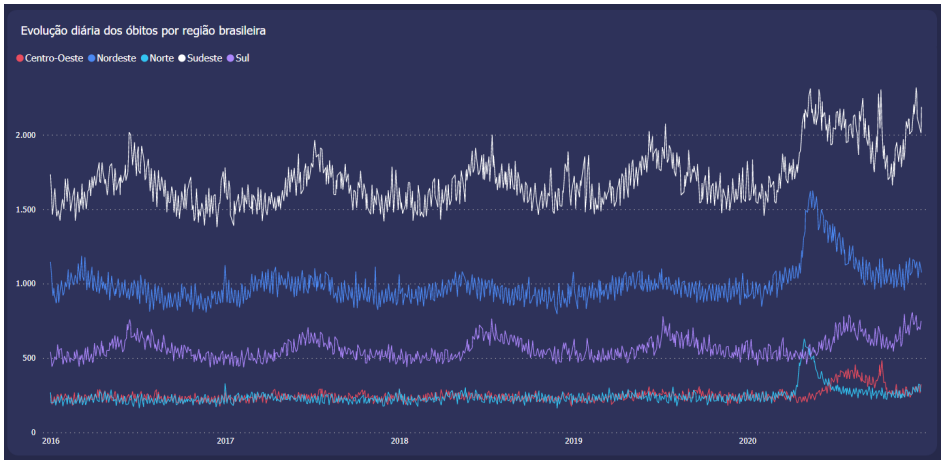
As doenças isquêmicas do coração, as doenças cerebrovasculares, a influenza e pneumonia e as neoplasias malignas dos órgãos digestivos se mantiveram no top quatro em quase todos os anos, exceto 2020 em que perderam a classificação para outras doenças por vírus, o que pode ser justificado pela classificação da pandemia de Covid-19. Já a diabetes mellitus não esteve no top cinco apenas em 2016 e 2017, que perdeu a classificação para outras formas de doença do coração e agressões, respectivamente. As agressões podem ser justificadas pelo fato de que em 2017 houve a maior taxa de homicídio no país, segundo IPEA (2021).

Pode-se também identificar que o ano de 2020 foi muito atípico para o cenário de mortalidade brasileira, a neoplasia perdeu o lugar no ranking também para as classificações associadas ao Covid-19.

5.4 Comportamento dos dados da série histórica nas regiões brasileiras

Nas regiões brasileiras, a maior média diária ao longo dos cinco anos foi do Sudeste, com o valor de 1.708. Na sequência vem o Nordeste e o Sul com 994 e 567, respectivamente. Já o Norte e o Centro-Oeste ficam quase que empatados com valores médios diários logo abaixo dos 250 óbitos diários.

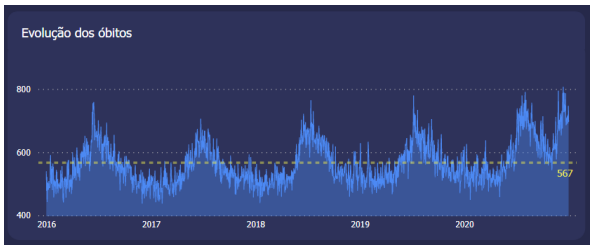
Figura 22: Evolução diária dos óbitos por região brasileira



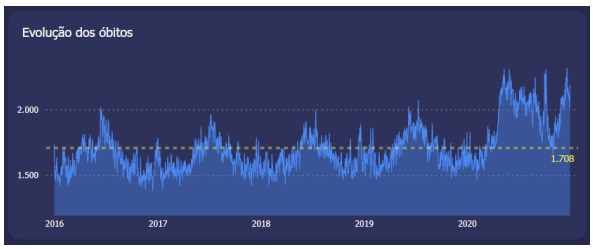
Fonte: Elaborada pela autora

As agressões entram no ranking dos cinco motivos que ocorreram mais óbitos nas regiões de Norte, Nordeste e Centro-Oeste, ficando como o maior motivo na região Norte. Essas três regiões, segundo IPEA (2021), são as que possuem maior taxa de homicídios no país. Além disso, foram registrados óbitos em aldeias indígenas apenas na região Norte e Centro-Oeste pois são as regiões onde se concentra a maioria das localidades indígena.

Figura 23: Histórico Sul e Sudeste



(a) Histórico Sul



(b) Histórico Sudeste

Fonte: Elaborada pela autora

O Sul (Figura 23a) e o Sudeste (Figura 23b) possuem alguns padrões bem específicos que ficam acima da média em épocas bem próximas todos os anos (entre os meses de junho e julho). Os motivos desses padrões não foram identificados.

Conforme identificado na seção 5.2, esse indicador também mostrou-se frágil devido a qualidade do preenchimento dos dados nas regiões Norte e Nordeste.

5.5 Comportamento dos dados com relação às características dos indivíduos

Dentro das características dos indivíduos, foi possível observar que a quantidade de óbitos da população que possuía idade mais avançada estava decrescendo devido ao aumento da expectativa de vida. Contudo, durante a pandemia Covid-19 o quantitativo dessa faixa etária teve um salto muito grande, onde a quantidade de óbitos dos indivíduos com idade acima de 70 anos subiu em 16%.

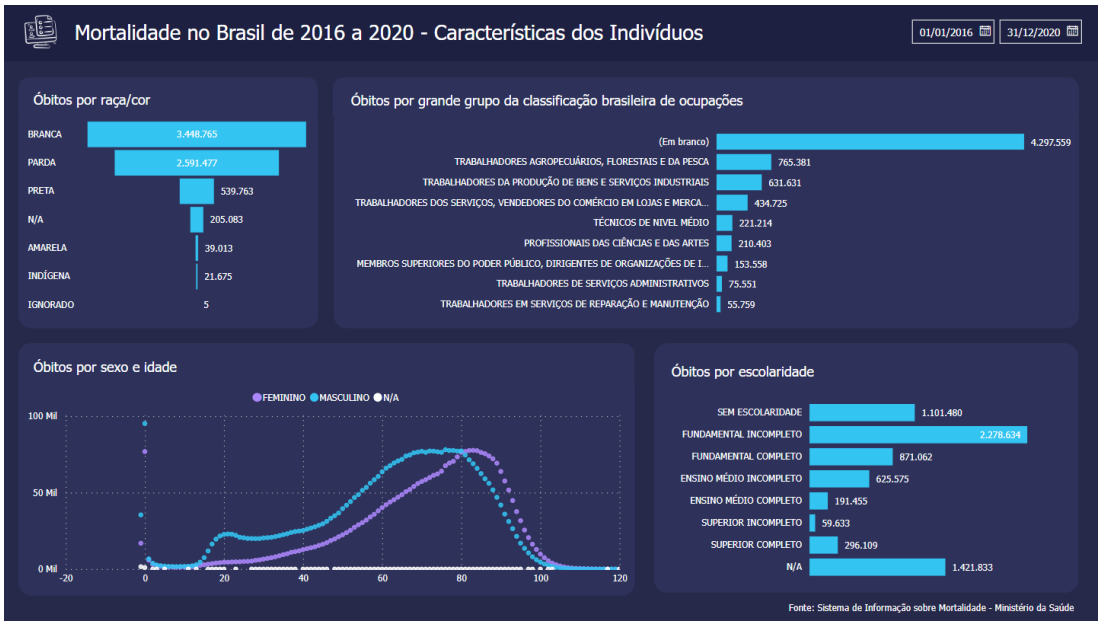
Outro ponto a se destacar é em relação ao aumento dos óbitos masculinos com idade entre 15 e 23 anos e as causas violentas são o fator determinante dessa discriminação. Segundo o IPEA (2021), é um fato global que homens adolescentes e jovens entre 15 e 29 anos são os que mais apresentam risco de serem vítimas de homicídios.

Nas mulheres, o ranking de óbito é referente às doenças cerebrovasculares que pode ser justificado pelos diversos picos hormonais ao longo da vida das mulheres e o uso de estrogênio seja no anticoncepcional ou na reposição hormonal pós-menopausa que aumentam a chance de acidentes vasculares cerebrais.

Foi possível observar também o preenchimento incorreto nos dados de ocupação do indivíduo que veio a óbito que dificulta, mais uma vez, uma análise mais completa dessa informação. Entretanto, diante do sinalizador presente na base de dados referente ao óbito ter sido oriundo de um acidente de trabalho ou não, foi possível identificar que dos acidentes de trabalho que ocasionaram óbitos, os trabalhadores da produção de bens e serviços industriais são os que lideram o ranking dos óbitos ao longo dos cinco anos.

Já nos óbitos por escolaridade, 71% dos indivíduos que vieram a óbito não possuem o ensino médio completo. Em contrapartida, 21% desses indivíduos não possuem escolaridade informada e apenas os 8% restantes possuem ensino médio completo e/ou superior completo.

Figura 24: *Dashboard* da mortalidade a partir das características dos indivíduos



Fonte: Elaborada pela autora

Os óbitos da população identificada como branca está na liderança da estratificação por raça/cor devido ao fato da quantidade populacional dessa característica no Brasil comparado às demais. Contudo, ao analisar o quantitativo de óbitos com a quantidade total da

população estimada pelo IBGE por raça/cor da população brasileira, foi possível identificar que a taxa de mortalidade das pessoas identificadas como pretas só difere em 20% das pessoas identificadas como brancas.

Contudo, pode-se considerar o fato de que também há sub-registros dentro dos dados da população identificada como preta que, historicamente, sofre com maior desigualdade social e, consequentemente, encontra-se com menos acesso aos serviços de saúde. Além disso, segundo o IPEA (2021), entre os anos de 2009 e 2019 as regiões Norte e Nordeste do país apresentaram aumento nas taxas de homicídios de negros e que, nessa pesquisa, pode também ter sido afetado pelo fato dos sub-registros nessas regiões.

5.6 Locais de ocorrência em que há maior incidência da mortalidade

Diante da relação do alto quantitativo de óbitos com as doenças crônicas não transmissíveis e também ao Covid-19, consequentemente os óbitos ocorrem na sua maior quantidade em hospitais ou em domicílios, que corresponde a 86% ao longo dos cinco anos.

Um ponto a se destacar são os óbitos ocorridos nas vias públicas que há um destaque nos acidentes de trânsito no ranking das causas de óbito, mas o número mais relevante são das agressões que correspondem a 43% do total nesse local de ocorrência. Já nas aldeias indígenas, no ranking das maiores causas de óbito aparecem as lesões autoprovocadas intencionalmente.

Pode-se ponderar também que o quantitativo de óbitos é maior nos hospitais devido ao fato dos registros serem mais efetivos nos serviços de saúde uma vez que o preenchimento da declaração de óbito é um ato médico obrigatório em todo o país.

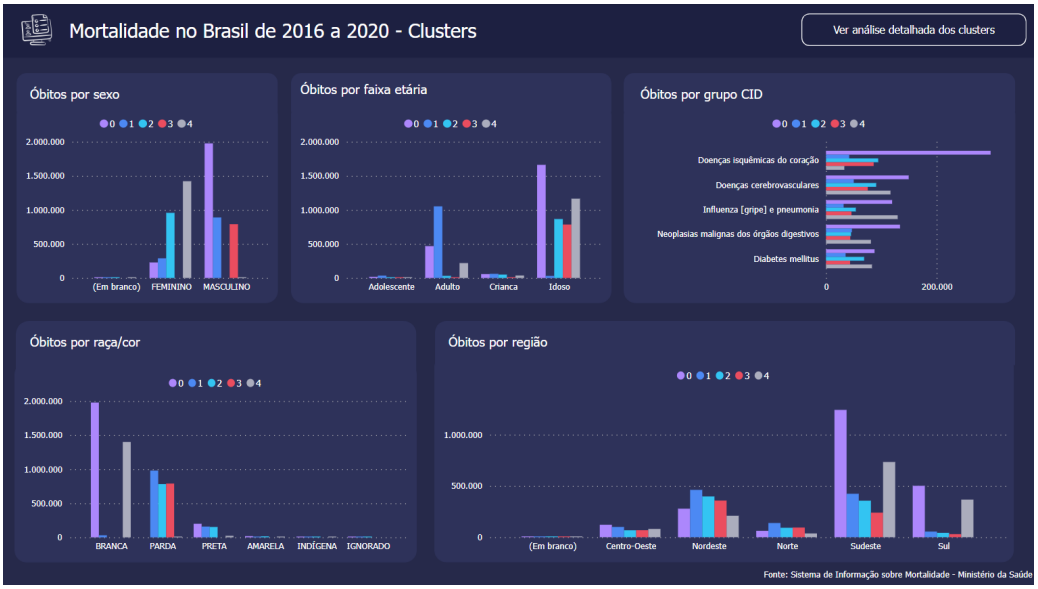
Figura 25: Quantitativo dos óbitos por local de ocorrência

Fonte: Elaborada pela autora

5.7 Grupos de fatores que influenciam diretamente no óbito de um indivíduo

A partir da execução do algoritmo k-modes, cinco clusters foram identificados para o agrupamento de fatores que influenciam nos padrões do óbito de um indivíduo. Os *dashboards* (Figura 26) foram construídos para analisar esses clusters e verificar as proximidades das características dos indivíduos.

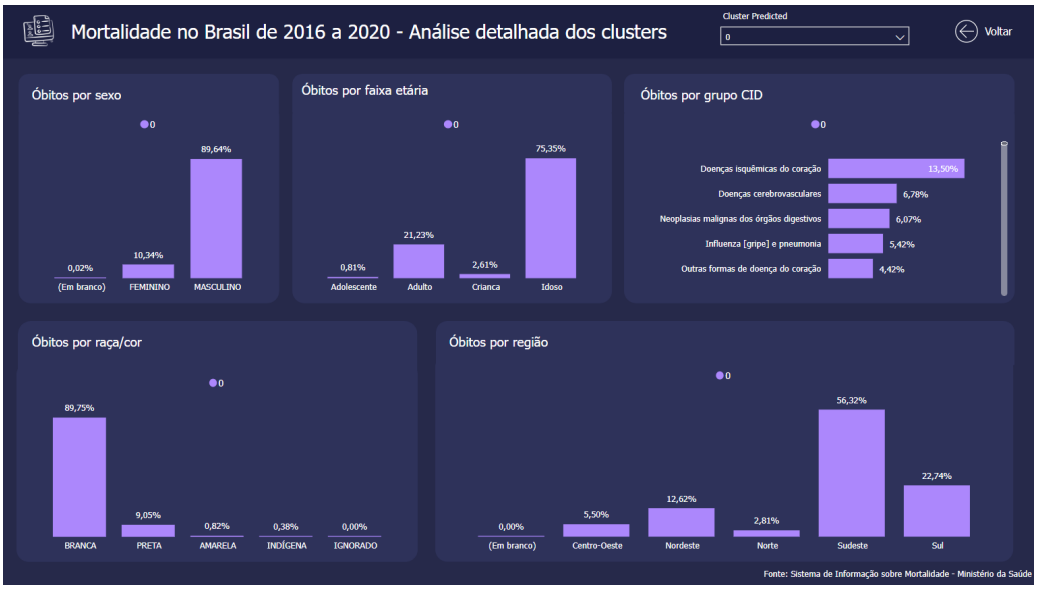
Figura 26: *Dashboard* dos resultados dos clusters



Fonte: Elaborada pela autora

No cluster 0 (Figura 27), foi possível analisar que idosos, identificados como brancos do sexo masculino nascidos na região Sudeste possuem maiores proximidades de características com os indivíduos que vieram a óbito por doenças isquêmicas do coração.

Figura 27: Análise detalhada do cluster 0



Fonte: Elaborada pela autora

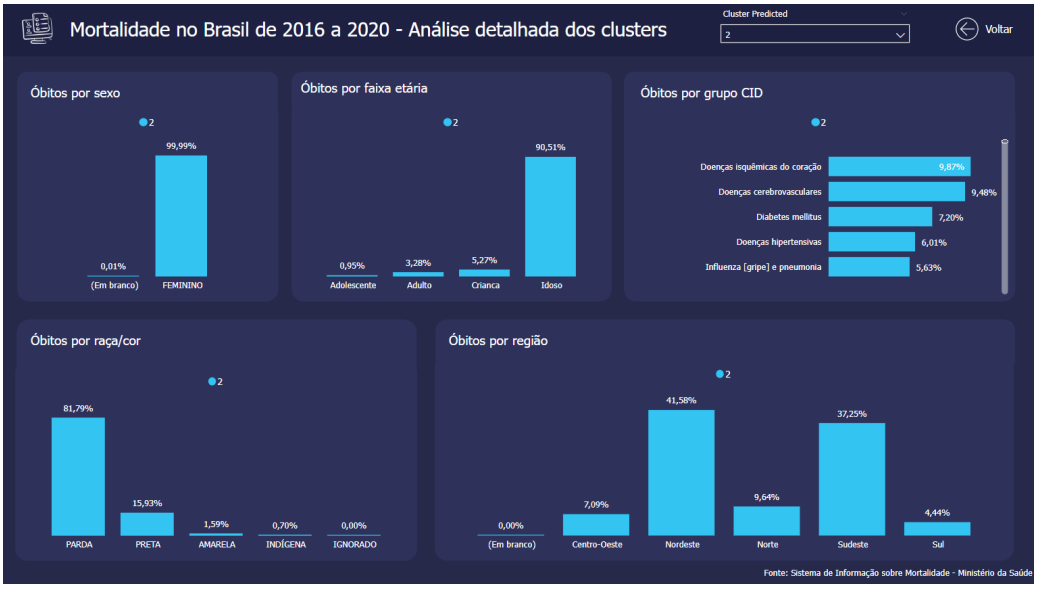
No cluster 1 (Figura 28), foi possível analisar que adultos, identificados como pardos do sexo masculino nascidos nas regiões Sudeste e Nordeste possuem maiores proximidades de características com os indivíduos que vieram a óbito por agressões.

Figura 28: Análise detalhada do cluster 1

Fonte: Elaborada pela autora

No cluster 2 (Figura 29), foi possível analisar que idosos, identificados como pardos do sexo feminino nascidos nas regiões Sudeste e Nordeste possuem maiores proximidades de características com os indivíduos que vieram a óbito por doenças isquêmicas do coração e doenças cerebrovasculares.

Figura 29: Análise detalhada do cluster 2



Fonte: Elaborada pela autora

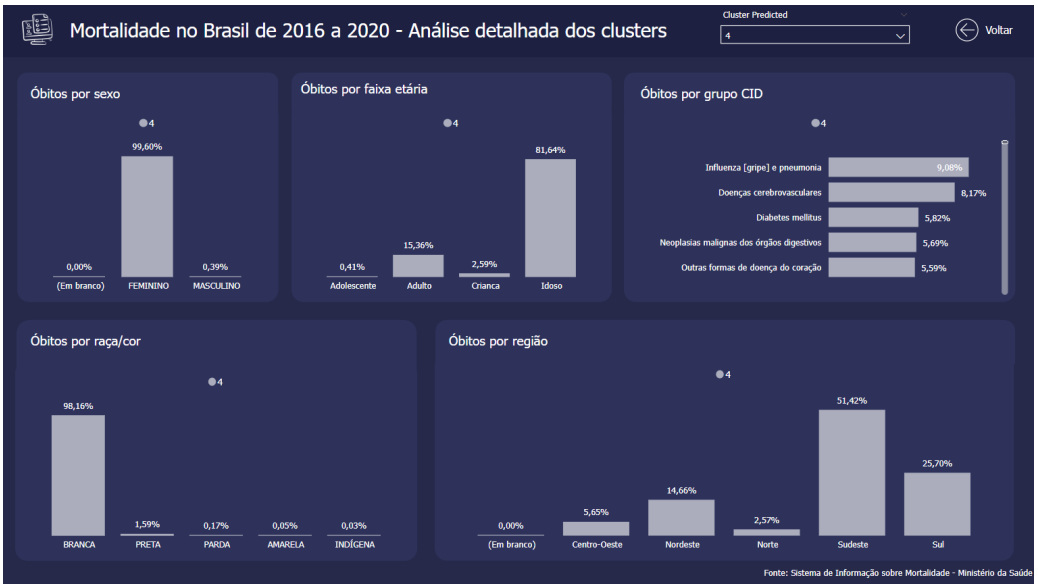
No cluster 3 (Figura 30), foi possível analisar que idosos, identificados como pardos do sexo masculino nascidos na região Nordeste possuem maiores proximidades de características com os indivíduos que vieram a óbito por doenças isquêmicas do coração.

Figura 30: Análise detalhada do cluster 3

Fonte: Elaborada pela autora

No cluster 4 (Figura 31), foi possível analisar que idosos, identificados como brancos do sexo feminino nascidos na região Sudeste possuem maiores proximidades de características com os indivíduos que vieram a óbito por influenza [gripe] e pneumonia.

Figura 31: Análise detalhada do cluster 4



Fonte: Elaborada pela autora

Dentro de cada cluster fez-se possível a classificação e agrupamento das características dos indivíduos que vieram a óbito para que facilite sua identificação preliminar a fim de realizar uma assistência mais assertiva para a população brasileira.

6 Conclusões

Essa pesquisa teve como objetivo desenvolver uma base de dados multidimensional e realizar uma mineração de dados para reconhecer alguns padrões de mortalidade brasileira. Através da conclusão dos objetivos específicos: (i) Identificar os dados da base do Sistema de Informação sobre Mortalidade dos anos de 2016 a 2020; (ii) Construir a arquitetura de *data warehouse*; (iii) Elaborar *dashboards* para visualização e análise de dados; (iv) Aplicar o algoritmo de clusterização e (v) Identificar as respostas a partir das perguntas levantadas para análise dos padrões de mortalidade, o objetivo do trabalho foi alcançado.

Com base na análise dos dados e das respostas alcançadas, pôde-se identificar que os padrões de mortalidade se mantiveram ao longo dos cinco anos, com exceção do período pandêmico da Covid-19 que impactou fortemente na estrutura etária da população e consequentemente na expectativa de vida. Identifica-se que a grande discrepância dos indicadores observados entre as regiões e as características dos indivíduos possui relação com o padrão da desigualdade em saúde no Brasil ao relacionar com a mortalidade e com a organização dos serviços de saúde não só entre as regiões brasileiras, mas também entre os estados intra-regionais.

Um ponto de limitação a ser destacado diz respeito ao problema já identificado em pesquisas anteriores em relação aos sub-registros na base do SIM o que leva a questionar se parte dessa população pode estar desassistida pelos órgãos de gestão da saúde. Além disso, foi possível identificar uma negligência no preenchimento dos campos relacionados às características dos indivíduos.

É importante ressaltar o potencial de utilização desse grande volume de dados organizados na estrutura de um *data warehouse* para subsidiar o planejamento da saúde pública no Brasil e a clusterização dos dados como um segmentador do conjunto de características dos indivíduos que necessitam de uma melhor assistência no sistema único de saúde.

Como trabalhos futuros, pode-se sugerir a evolução da mineração de dados aplicando outros algoritmos de aprendizagem supervisionada com o objetivo de prever os óbitos ou de árvore de decisão a fim de auxiliar o processo de tomada de decisão nos serviços de saúde.

Referências

Alves, D. d. S. B. *et al.* (2017). *Mineração de dados na identificação de padrões de mortalidade no Brasil de 1979 a 2013*. Ph.D. thesis, Fiocruz.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cassiano, K. M. (2014). Análise de séries temporais usando análise espectral singular (ssa) e clusterização de suas componentes baseada em densidade. *Pontifícia Universidade Católica do Rio de Janeiro*.

DASNT (2022). *Sistema de Informação sobre Mortalidade*. Secretaria de Vigilância em Saúde, Brasil. Available at <http://svs.aids.gov.br/dantps/cgiae/sim/apresentacao/>. Visited in July, 2022.

Demenech, L. M., Dumith, S. d. C., Vieira, M. E. C. D., and Neiva-Silva, L. (2020). Desigualdade econômica e risco de infecção e morte por covid-19 no brasil. *Revista Brasileira de Epidemiologia*, **23**.

Harrison, T. H. (1998). *Intranet Data Warehouse: ferramentas e técnicas para a utilização do data warehouse na intranet*. Berkerley/ABDR.

Honda, H. (2017). *Introdução Básica à Clusterização*. Brasil. Available at <https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao/>. Visited in October, 2022.

IBGE (2019). *Sistema de Estatísticas Vitais*. Instituto Brasileiro de Geografia e Estatística, Brasil. Available at <https://www.ibge.gov.br/estatisticas/sociais/populacao/26176-estimativa-do-sub-registro.html?edicao=32265t=sobre/>. Visited in October, 2022.

Inmon, W. H. (2000). Building the data warehouse: Getting started. *Recuperado de: http://www. academia. edu/3081161/Building_the_data_warehouse*.

IPEA (2021). *Atlas da Violência 2021*. Fórum Brasileiro de Segurança Pública, Brasil. Available at <https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/5141-atlasdaviolencia2021completo.pdf/>. Visited in October, 2022.

Kimball, R. and Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Knaflic, C. N. (2019). *Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios*. Alta Books.

Lasmar, M. P. F. and Siviero, P. C. L. (2018). Níveis e padrões da mortalidade brasileira e suas macrorregiões: uma análise com base em indicadores demográficos, 2000 e 2010. *Revista debate econômico*, **6**(1), 100–118.

MARQUES, M. N. *et al.* (2014). Câncer gastrointestinal: dificuldades para o acesso ao diagnóstico e tratamento.

Martins, T. C. d. F., Silva, J. H. C. M. d., Máximo, G. d. C., and Guimarães, R. M. (2021). Transição da morbimortalidade no brasil: um desafio aos 30 anos de sus. *Ciência & Saúde Coletiva*, **26**, 4483–4496.

PAHO (2020). Annual report of the director of the pan american sanitary bureau 2020. saving lives and improving health and well-being. Technical report, Pan American Health Organization, Washington, D.C.

Pereira, R. A., Alves-Souza, R. A., and Vale, J. S. (2015). O processo de transição epidemiológica no brasil: uma revisão de literatura. *Revista Científica da Faculdade de Educação e Meio Ambiente*, **6**(1), 99–108.

RIPSA (2008). *Indicadores básicos para a saúde no Brasil: conceitos e aplicações*. Brasília, Brasil.

SIDRA (2018). *Tabela 7358 - População, por sexo e idade*. Instituto Brasileiro de Geografia e Estatística, Brasil. Available at <https://sidra.ibge.gov.br/tabela/7358/>. Visited in September, 2022.

SIDRA (2020). *Tabela 6579 - População residente estimada*. Instituto Brasileiro de Geografia e Estatística, Brasil. Available at <https://sidra.ibge.gov.br/tabela/6579/>. Visited in September, 2022.