Universidade Presbiteriana Mackenzie

CURSO: TECNOLOGIA EM CIÊNCIA DE DADOS
SEMESTRE: 2° SEMESTRE
COMPONENTE CURRICULAR / TEMA: PROJETO APLICADO I
NOME DO PROFESSOR: EVERTON KNIHS

Universidade Presbiteriana Mackenzie

UNIVERSIDADE PRESBITERIANA MACKENZIE

FACULDADE DE COMPUTAÇÃO E INFORMÁTICA

Análise Exploratória das Relações entre Taxa de Mortalidade e Fatores Socioeconômicos no Brasil

São Paulo

2023

SUMÁRIO

1 TÍTULO ESCOLHIDO 4
2 MEMBROS..... 4
3 CONTEXTO DO ESTUDO..... 5
3.1 PREMISSAS DO PROJETO.....6
3.2 OBJETIVOS E METAS.....7
3.3 METODOLOGIA.....7
4 CRONOGRAMA DE ATIVIDADES 8
5 REFERÊNCIAS DE AQUISIÇÃO DO DATASET 9
5.1 PERÍODO DE COLETA.....10
6 DESCRIÇÃO DO DATASET 10
7 EXPLORAÇÃO DE CORRELAÇÕES RELEVANTES 10
8 PROBLEMAS E FENÔMENOS ENCONTRADOS 11
8.1 PENSAMENTO COMPUTACIONAL NO CONFRONTO DOS PROBLEMAS.....12
9 METADADOS.....14
10 ANÁLISE EXPLORATÓRIA DE DADOS.....14
10.1 TRABALHO INICIAL.....14
10.2 OBJETIVO DA ANÁLISE EXPLORATÓRIA DE DADOS.....16
11 ESBOÇO DO STORY TELLING.....19
12 LINK DO GITHUB 29
13 13 L LINK DO VÍDEO NO YOUTUBE.....29

REFERÊNCIAS 30

1. TÍTULO ESCOLHIDO:

Análise Exploratória das Relações entre Taxa de Mortalidade e Fatores Socioeconômicos no Brasil

2. MEMBROS:

Nome: Adriano Mamoru Takeshita

TIA: 23022647

E-mail: 10923022647@mackenzista.com.br

Nome: Camila Vieira

TIA: 23005432

E-mail: camila.vieira1@mackenzista.com.br / 10923005432@mackenzista.com.br

Nome: Gabriel Schonenberger de Campos

TIA: 23011165

E-mail: 10923011165@mackenzista.com.br

Nome: Gustavo Santiago Zarpelão

TIA: 23002824

E-mail: 10923002824@mackenzista.com.br

Nome: Luís Eduardo Alves de Moura da Silva

TIA: 23009470

E-mail: 10923009470@mackenzista.com.br

3. CONTEXTO DO ESTUDO:

Brasil como Foco:

O Brasil, como um dos países mais populosos e diversificados do mundo, apresenta um cenário sociodemográfico complexo e dinâmico, tornando-se um contexto extremamente relevante para a análise de dados, especialmente sobre a taxa de mortalidade. A rica heterogeneidade regional, as desigualdades socioeconômicas marcantes e os desafios em saúde pública proporcionam uma oportunidade única para investigar as interações entre fatores sociais, econômicos e demográficos e seu impacto na taxa de mortalidade.

Instituição escolhida: Instituto Brasileiro de Geografia e Estatística (IBGE)

Missão: Produzir informações geográficas e estatísticas de qualidade, subsidiando o conhecimento da realidade brasileira e a formulação de políticas públicas.

Visão: Ser reconhecido nacional e internacionalmente como referência em produção e disseminação de informações estatísticas, geográficas e cartográficas.

Valores: Transparência, ética, qualidade, responsabilidade social, inovação e valorização dos colaboradores.

Segmento de atuação do IBGE: O IBGE atua no segmento de produção, análise e disseminação de informações estatísticas, geográficas e cartográficas no Brasil, sendo a principal fonte de referência para dados relacionados à população, economia, geografia e demais aspectos sociodemográficos do país.

Posicionamento no mercado e Colaboradores:

A instituição governamental é amplamente reconhecida e respeitada, empregando uma equipe substancial de colaboradores, incluindo estatísticos, geógrafos, cartógrafos, pesquisadores, analistas e técnicos em diversas áreas.

Iniciativas na área de Data Science: O IBGE tem investido cada vez mais em iniciativas relacionadas à Data Science, modernizando métodos de coleta e análise de dados, utilizando técnicas avançadas de processamento de informações, como big data e aprendizado de máquina, e desenvolvendo ferramentas tecnológicas para melhorar a acessibilidade e a utilidade dos dados produzidos.

Trabalhos em Destaque: A instituição é conhecida por trabalhos importantes, como o Censo Demográfico, realizado a cada década, que fornece informações detalhadas sobre a população brasileira. Além disso, realiza pesquisas contínuas sobre diversos aspectos socioeconômicos do país, incluindo o mercado de trabalho, inflação, Produto Interno Bruto (PIB) e outros indicadores essenciais. Esses trabalhos impactam significativamente a formulação de políticas públicas, planejamento estratégico e tomada de decisões no Brasil.

Ao unir a diversidade do Brasil com a qualidade dos dados do IBGE, este projeto busca fornecer insights importantes para o entendimento das relações complexas entre variáveis sociais, econômicas e de saúde, contribuindo para o desenvolvimento de estratégias de melhoria e bem-estar da população brasileira.

3.1 Premissas do Projeto:

1. Utilização dos dados fornecidos pelo IBGE como fonte confiável e abrangente de informações demográficas e socioeconômicas.

2. Implementação de técnicas de ciência de dados para analisar e extrair insights das informações relacionadas à taxa de mortalidade.

3. Consideração das variáveis sociais, econômicas e demográficas para uma análise mais completa e precisa.

4. Através da Análise Exploratória, foco na identificação de padrões, correlações e tendências que possam ajudar a compreender os fatores associados à taxa de mortalidade.

5. Fonte de Dados Confiável: Utilização de dados do IBGE como fonte confiável de informações socioeconômicas e demográficas.

6. Foco na Taxa de Mortalidade: Ênfase na análise da taxa de mortalidade no Brasil e suas relações com fatores sociais e econômicos.

7. Pensamento Computacional: Aplicação das premissas de decomposição, reconhecimento de padrões, abstração e algoritmo na análise dos dados.

8. Contribuição para Políticas Públicas: Objetivo de oferecer insights relevantes que possam ser possíveis fontes de informação voltadas ao bem-estar da população.

3.2 Objetivos e Metas:

Objetivo Principal: Analisar as relações entre a taxa de mortalidade e diversos fatores sociais, econômicos e demográficos no Brasil.

Metas: Identificar correlações significativas entre a taxa de mortalidade e outras variáveis; compreender como fatores como renda, acesso a serviços básicos, segurança e saúde estão relacionados à taxa de mortalidade.

3.3 Metodologia:

Neste projeto, adotou-se uma abordagem baseada no pensamento computacional e na análise exploratória de dados, utilizando as premissas de decomposição, reconhecimento de padrões, abstração e algoritmo.

O objetivo principal é desvendar insights relevantes sobre as relações entre a taxa de mortalidade no Brasil e os fatores socioeconômicos, aproveitando ao máximo os dados fornecidos pelo IBGE.

Decomposição:

Iniciaremos decompondo o problema em partes menores e mais gerenciáveis. Dividiremos a análise em etapas, considerando diferentes variáveis, como renda, educação, acesso a serviços básicos e

segurança. Isso nos permitirá focar em aspectos específicos e compreender melhor como eles contribuem para a taxa de mortalidade.

Reconhecimento de Padrões:

Utilizaremos técnicas de reconhecimento de padrões para identificar correlações preliminares entre variáveis. Também utilizaremos outros estudos que já existem. Através de gráficos, tabelas e métricas descritivas, buscaremos padrões visuais e tendências iniciais nos dados, destacando possíveis conexões entre os fatores sociais e a taxa de mortalidade.

Abstração:

A abstração nos permitirá simplificar a complexidade dos dados, concentrando-nos nos aspectos mais relevantes. Agruparemos variáveis relacionadas e criaremos indicadores compostos que representam características socioeconômicas mais abrangentes. Essa abordagem simplificada facilitará a análise e a interpretação dos resultados.

Algoritmo:

Implementaremos algoritmos de análise exploratória, como cálculos de médias, desvios-padrão e coeficientes de correlação. Isso nos ajudará a quantificar as relações entre as variáveis e a medir a força das associações encontradas. Representaremos visualmente as tendências identificadas através de R e Python.

Através da combinação dessas premissas do pensamento computacional e técnicas de análise exploratória de dados, exploraremos as nuances das relações entre a taxa de mortalidade e fatores sociais no Brasil. Essa abordagem nos permitirá desenvolver uma compreensão mais profunda das interações subjacentes e fornecer insights relevantes que contribuam para uma análise sólida e informada.

4. CRONOGRAMA DE ATIVIDADES

1ª ETAPA

1. **1º encontro: Brainstorm (12/08/2023) –** Adriano, Camila, Gabriel, Luis.
2. **Pesquisa de indicadores e possíveis relações (Entre 12/08/2023 e 18/08/2023) –** Adriano, Camila, Gabriel, Luis.

3. **2º encontro: Discussão sobre os indicadores e Busca de Dados (18/08/2023):** Coleta de informações demográficas, socioeconômicas e de saúde do IBGE; Discussão e definição do escopo - Adriano, Camila, Gabriel, Luis.
4. **Downloads das tabelas de indicadores definidos (Semana do dia 20/08) –** Gabriel e Luis.
5. **Criação de Conta no GitHub (Semana do dia 20/08)**
6. **Desenvolvimento do documento da 1ª etapa (De 13/08 a 21/08) –** Camila.
7. **Revisão documento da 1ª etapa (Até 22/08) -** Todos.
8. **Entrega da 1ª etapa (Até 28/08).**

2ª, 3ª e 4ª ETAPA – PREVISÃO DO CRONOGRAMA

1. **Pré-processamento e Limpeza de Dados (Entre 02/09 e 08/09):** Tratamento de dados faltantes, duplicados e inconsistências através de Python e R.
2. **Análise Exploratória (Até 24/09): Exploração inicial dos dados para identificar padrões e insights iniciais -** Todos**.**
3. **Entrega da 2ª etapa (25/09).**
4. **Esboço do Storytelling (Entre 26/09 a 05/10) –** Todos.
5. **Modelagem dos dados e Scripts (Entre 05/10 a 12/10) –** Todos.
6. **Visualização de Dados:** Criação de gráficos e visualizações para apresentar os resultados de forma clara e compreensível (R e Python) **(12/10 a 20/10) -** Todos.
7. **Revisão e entrega no GitHub (Até 30/10) -** Todos.
8. **3ª entrega: Relatório Final (De 13/08 a 30/10):** Documentação completa das análises realizadas e dos insights obtidos - todos.
9. **Gravação da apresentação (Até 15/11) – Todos.**
10. **Entrega da 4ª etapa (20/11).**

5. REFERÊNCIAS DE AQUISIÇÃO DO DATASET

O macro dos dados utilizados neste projeto fora obtidos do Instituto Brasileiro de Geografia e Estatística (IBGE), uma fonte confiável de informações demográficas, socioeconômicas e geográficas no Brasil.

Também explorou-se a Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua. Este conjunto de dados oferece uma visão abrangente das condições socioeconômicas da população brasileira, incluindo indicadores como renda, educação, emprego e saúde.

Censo Demográfico 2020: O Censo Demográfico fornece um panorama detalhado da população, habitação e características socioeconômicas do Brasil.

5.1 Período de Coleta:

Os dados utilizados neste projeto foram coletados entre 2014 e 2019 considerando os períodos específicos de realização das pesquisas da PNAD, Censo Demográfico 2020.

6. DESCRIÇÃO DO DATASET

O dataset utilizado neste projeto explora informações relevantes sobre a taxa de mortalidade no Brasil, além de dados socioeconômicos que podem influenciar ou serem influenciados por essa métrica crucial de saúde pública. A proposta deste conjunto de dados é investigar as relações complexas entre a taxa de mortalidade e diversos fatores socioeconômicos, permitindo uma compreensão mais profunda dos fenômenos que impactam a saúde e o bem-estar da população brasileira.

Conteúdo do Dataset:

O dataset engloba as variáveis demográficas, econômicas e sociais:

- Indicador 1.2.1 - Proporção da população vivendo abaixo da linha de pobreza nacional, por sexo
- Indicador 1.2.1 - Proporção da população abaixo da linha de pobreza nacional, por grupos de idade
- Indicador 3.9.2 - Taxa de mortalidade atribuída a fontes de água inseguras, saneamento inseguro e falta de higiene, por sexo e grupo de idade
- Indicador 16.1.4 - Proporção da população de 15 anos ou mais de idade que se sente segura quando caminha sozinha na área onde vive durante a noite, por cor ou raça
- Indicador 16.1.1 - Número de vítimas de homicídios intencionais por 100 mil habitantes, por sexo
- Indicador 3.5.2 - Consumo de álcool em litros de álcool puro per capita, por pessoas de 15 anos ou mais de idade

- Indicador 3.6.1 - Taxa de mortalidade por acidentes de trânsito, por sexo e grupos de idade

7. EXPLORAÇÃO INICIAL DE CORRELAÇÕES RELEVANTES

Ao constatar os dados, é possível que haja correlações entre esses indicadores:

- Taxa de Mortalidade por Álcool & Proporção da População Abaixo da Linha de Pobreza e escolaridade
- Taxa de Mortalidade por Doenças Crônicas &. Proporção da População Abaixo da Linha de Pobreza
- Taxa de Homicídios & Sensação de Segurança Noturna
- Taxa de Homicídios & Proporção de Pessoas Abaixo da Linha de Pobreza
- Taxa de homicídio & renda por região

- Taxa de homicídios (apenas sexo feminino) & Sensação de segurança noturna
- Taxa de homicídios (filtrar pretos e pardos) & Proporção da população abaixo da linha da pobreza
- Taxa de mortalidade por falta de saneamento & Proporção da população abaixo da linha da pobreza

8. PROBLEMAS E FENÔMENOS ENCONTRADOS

No desenvolvimento deste estudo, deparamo-nos com alguns desafios relacionados à qualidade e ao processamento dos dados. Estes desafios desempenham um papel fundamental na garantia da confiabilidade e validade das análises conduzidas, e, portanto, requerem uma abordagem cuidadosa.

Despadronização dos Dados:

Uma das questões mais notáveis que encontramos foi a falta de padronização dos dados em nossas fontes. Isso incluiu a mistura de informações em colunas, como ano e faixa etária, o que dificultou a análise e a interpretação dos dados. Para resolver este problema, dedicamos um

esforço significativo à padronização das planilhas. Isso envolveu a reorganização das informações para garantir que todas sigam a mesma estrutura de dados e formatação.

Dados representados de forma Inconsistente:

Identificamos casos de dados incorretos, como faixas etárias apresentadas de formas inconsistentes. Esses erros podem comprometer a qualidade de nossas análises. Para lidar com essa questão, realizamos uma revisão detalhada dos dados e corrigimos as informações inconsistentes sempre que possível. Quando os dados incorretos não puderam ser corrigidos, optamos por excluí-los, garantindo que nossas análises se baseassem em informações confiáveis.

Lacunas nos Dados:

Ao explorar nossos indicadores, notamos a presença de lacunas em alguns dados relevantes. Estas lacunas podem limitar a capacidade de responder completamente às nossas perguntas de pesquisa. Para superar essa limitação, estamos considerando a coleta de dados adicionais ou a busca de fontes alternativas de informações para preencher as lacunas (caso seja de importância que esses dados estejam no estudo, senão, optaremos por excluí-los).

Necessidade de Análise de Padrões:

Reconhecemos a importância de identificar padrões nos dados para responder às perguntas de pesquisa e encontrar insights valiosos. A análise de padrões pode revelar correlações, tendências temporais e diferenças regionais que são cruciais para nosso estudo. Para isso, estamos aplicando técnicas de análise exploratória de dados e criando visualizações para facilitar a identificação desses padrões.

Fomentando uma Abordagem Crítica:

É igualmente fundamental adotar uma abordagem crítica ao longo de todo o processo de análise. Isso implica estar aberto a resultados que possam desafiar nossas hipóteses iniciais e estar disposto a questionar afirmações. A prática da análise crítica pode levar a descobertas inesperadas e enriquecer nossas conclusões.

8.1 PENSAMENTO COMPUTACIONAL NO CONFRONTO DOS PROBLEMAS ENCONTRADOS

Para enfrentar esses desafios, a abordagem baseada no pensamento computacional vem sendo usada durante todo o estudo. Primeiramente, dividimos nosso problema em etapas menores, incluindo a limpeza de dados, análise exploratória e modelagem estatística.

Decomposição:

Decidimos dividir os indicadores e priorizar a análise das taxas de mortalidade em nossa primeira etapa. Isso nos permitirá focar em um conjunto específico de indicadores, simplificando a complexidade do projeto.

Para enfrentar a despadronização dos dados, optamos por organizar as informações em planilhas separadas por ano (como citado anteriormente, os dados vieram mesclados), abrangendo o período de 2015 a 2019. Dessa forma, conseguimos uma estrutura mais clara e consistente que facilita a análise e a comparação dos indicadores ao longo do tempo.

Reconhecimento de Padrões e Abstração:

Identificamos elementos comuns em todas as tabelas, como faixa etária, sexo e região. No caso das regiões, embora os dados estivessem inicialmente mesclados, observamos que poderíamos simplificar a análise, abstraindo as informações para o nível nacional (Brasil). Isso nos permitirá trabalhar com uma visão geral dos indicadores, sem a necessidade de separar detalhadamente as diferentes regiões do Brasil.

Algoritmos:

Para iniciar as primeiras análises estatísticas, recorremos ao R e desenvolvemos scripts que podem ser encontrados em nosso diretório do GitHub. Com base nas divisões de dados realizadas, calculamos medidas de posição, como média, mediana e quartis, para entender tendências centrais em nossos indicadores. Além disso, aplicamos medidas de dispersão, como variância e desvio padrão, para avaliar a variabilidade dos dados.

Ao adotar essa abordagem de pensamento computacional, estamos capacitados para enfrentar os desafios em nossos dados de maneira estruturada e sistemática, garantindo que nossas análises sejam conduzidas de maneira eficaz e que possamos extrair insights significativos para contribuir com nosso estudo sobre as relações entre os indicadores sociais, econômicos e de saúde no contexto brasileiro.

9. METADADOS

Tipo de Arquivo: O IBGE disponibilizou os arquivos em tabelas Excel (xls), mas posteriormente transferimos esses dados para o ambiente R.

Origem dos Dados: Os dados utilizados são de fonte aberta, disponíveis publicamente para consulta e análise.

Sensibilidade dos Dados: Os dados do IBGE são de natureza estatística e não possuem informações sensíveis.

Validade dos Dados: No contexto de dados estatísticos, como os do IBGE, esses dados não têm um prazo de validade definido, pois representam medições em momentos específicos.

Proprietário dos Dados: O Instituto Brasileiro de Geografia e Estatística (IBGE) é o órgão responsável por essas informações.

Restrições de Uso: Verificamos se há restrições legais ou regulatórias que afetam o uso dos dados, como a Lei Geral de Proteção de Dados (LGPD). No entanto, em geral, os dados do IBGE estão disponíveis para uso público.

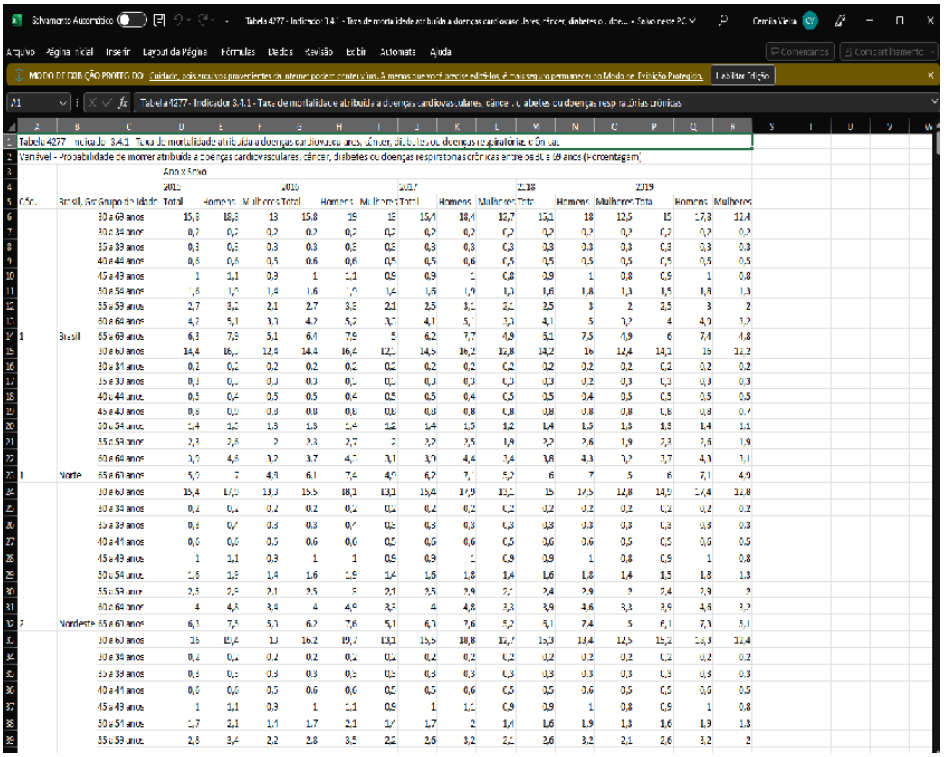
Descrição dos Atributos: Para uma análise eficaz, fornecemos descrições detalhadas dos atributos presentes em nossos datasets. Isso inclui definições claras, tipos de dados (como numéricos, categóricos ou datas) e, se aplicável, os possíveis valores ou categorias para cada atributo. Essa descrição detalhada dos atributos é fundamental para compreender a natureza dos dados e prepará-los adequadamente para análises.

10. ANÁLISE EXPLORATÓRIA DE DADOS

10.1 Trabalho inicial:

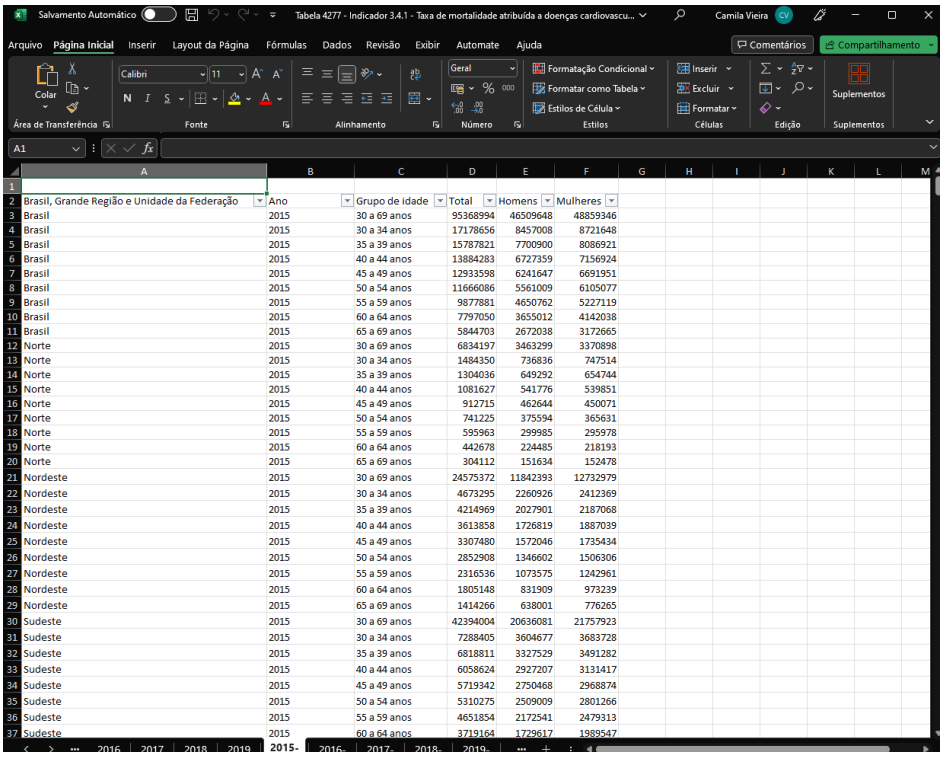
Após baixar os dados, como descrito no Capítulo 8, divergências foram encontradas e precisou-se aplicar algumas alterações, sendo elas:

Ao baixar os arquivos em Excel, primeiramente adequamos os dados mesclados (linhas/colunas, dados strings e numéricos), separando-os:



1	Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas																						
2	Variável - Probabilidade de morrer atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas entre 30 e 69 anos (Porcentagem)																						
3			Ano x Sexo																				
4			2012			2016			2017			2018			2019								
5	Cód.	Brasil, Grande Grupo de Idade	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres
6		30 a 69 anos	15,8	18,8	13	15,8	19	13	15,4	18,4	12,7	15,1	18	12,5	15	17,8	12,4						
7		30 a 34 anos	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2						
8		35 a 39 anos	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3						
9		40 a 44 anos	0,6	0,6	0,5	0,6	0,6	0,5	0,5	0,6	0,5	0,5	0,5	0,5	0,5	0,5	0,5						
10		45 a 49 anos	1	1,1	0,9	1	1,1	0,9	0,9	1	0,8	0,9	1	0,8	0,9	1	0,8						
11		50 a 54 anos	1,6	1,7	1,4	1,6	1,7	1,4	1,6	1,9	1,3	1,6	1,8	1,3	1,5	1,8	1,3						
12		55 a 59 anos	2,7	3,2	2,1	2,7	3,3	2,1	2,5	3,1	2,1	2,5	3	2	2,5	3	2						
13		60 a 64 anos	4,7	5,1	3,7	4,7	5,2	3,7	4,1	5,	3,3	4,1	5	3,2	4	4,9	3,2						
14	1	Brasil	65 a 69 anos	6,3	7,5	5,1	6,4	7,5	5	6,2	7,7	4,9	5,1	7,5	4,9	6	7,4	4,8					
15		30 a 60 anos	14,4	16,	12,4	14,4	16,4	12,	14,5	16,2	12,8	14,2	16	12,4	14,1	16	12,2						
16		30 a 34 anos	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2						
17		35 a 39 anos	0,3	0,	0,3	0,3	0,	0,	0,3	0,3	0,3	0,3	0,2	0,3	0,3	0,3	0,3						
18		40 a 44 anos	0,5	0,4	0,5	0,5	0,4	0,5	0,5	0,4	0,5	0,5	0,4	0,5	0,5	0,5	0,5						
19		45 a 49 anos	0,8	0,	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,7						
20		50 a 54 anos	1,4	1,2	1,3	1,3	1,4	1,2	1,4	1,5	1,2	1,4	1,5	1,3	1,3	1,4	1,1						
21		55 a 59 anos	2,3	2,8	2	2,3	2,7	2	2,2	2,5	1,9	2,2	2,6	1,9	2,3	2,6	1,9						
22		60 a 64 anos	3,7	4,6	3,2	3,7	4,	3,1	3,3	4,4	3,4	3,8	4,3	3,2	3,7	4,3	3,1						
23	1	Norte	65 a 69 anos	5,2	7	4,3	6,1	7,4	4,9	6,2	7,	5,2	6	7	5	6	7,1	4,9					
24		30 a 60 anos	15,4	17,2	13,3	15,5	18,1	13,1	15,4	17,9	13,1	15	17,5	12,8	14,9	17,4	12,8						
25		30 a 34 anos	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2						
26		35 a 39 anos	0,3	0,7	0,3	0,3	0,7	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3						
27		40 a 44 anos	0,6	0,6	0,5	0,6	0,6	0,5	0,6	0,6	0,5	0,6	0,6	0,5	0,5	0,6	0,5						
28		45 a 49 anos	1	1,1	0,9	1	1	0,9	0,9	1	0,9	0,9	1	0,8	0,9	1	0,8						
29		50 a 54 anos	1,6	1,8	1,4	1,6	1,9	1,4	1,6	1,8	1,4	1,6	1,8	1,4	1,5	1,8	1,3						
30		55 a 59 anos	2,5	2,9	2,1	2,5	3	2,1	2,5	2,9	2,1	2,4	2,9	2	2,4	2,9	2						
31		60 a 64 anos	4	4,8	3,4	4	4,9	3,3	4	4,8	3,3	3,9	4,6	3,3	3,9	4,6	3,2						
32	2	Nordeste	65 a 69 anos	6,1	7,5	5,1	6,2	7,6	5,1	6,1	7,6	5,2	5,1	7,4	5	6,1	7,3	5,1					
33		30 a 60 anos	16	18,4	13	16,2	19,	13,1	15,5	18,8	12,7	15,3	18,4	12,5	15,2	18,3	12,4						
34		30 a 34 anos	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2						
35		35 a 39 anos	0,3	0,2	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,3						
36		40 a 44 anos	0,6	0,6	0,5	0,6	0,6	0,5	0,5	0,6	0,5	0,5	0,6	0,5	0,5	0,6	0,5						
37		45 a 49 anos	1	1,1	0,9	1	1,1	0,9	1	1,1	0,9	0,9	1	0,8	0,9	1	0,8						
38		50 a 54 anos	1,7	2,1	1,4	1,7	2,1	1,4	1,7	2	1,4	1,6	1,9	1,3	1,6	1,9	1,3						
39		55 a 59 anos	2,8	3,4	2,2	2,8	3,3	2,2	2,6	3,2	2,1	2,6	3,2	2,1	2,6	3,2	2						

Figura 1: Tabela de Taxa de Mortalidade sem tratamento



Brasil, Grande Região e Unidade da Federação	Ano	Grupo de idade	Total	Homens	Mulheres
Brasil	2015	30 a 69 anos	95368994	46509648	48859346
Brasil	2015	30 a 34 anos	17178656	8457008	8721648
Brasil	2015	35 a 39 anos	15787821	7700900	8086921
Brasil	2015	40 a 44 anos	13884283	6727359	7156924
Brasil	2015	45 a 49 anos	12933598	6241647	6691951
Brasil	2015	50 a 54 anos	11666086	5561009	6105077
Brasil	2015	55 a 59 anos	9877881	4650762	5227119
Brasil	2015	60 a 64 anos	7797050	3655012	4142038
Brasil	2015	65 a 69 anos	5844703	2672038	3172665
Norte	2015	30 a 69 anos	6834197	3463299	3370898
Norte	2015	30 a 34 anos	1484350	736836	747514
Norte	2015	35 a 39 anos	1304036	649292	654744
Norte	2015	40 a 44 anos	1081627	541776	539851
Norte	2015	45 a 49 anos	912715	462644	450071
Norte	2015	50 a 54 anos	741225	375594	365631
Norte	2015	55 a 59 anos	595963	299985	295978
Norte	2015	60 a 64 anos	442678	224485	218193
Norte	2015	65 a 69 anos	304112	151634	152478
Nordeste	2015	30 a 69 anos	24575372	11842393	12732979
Nordeste	2015	30 a 34 anos	4673295	2260926	2412369
Nordeste	2015	35 a 39 anos	4214969	2027901	2187068
Nordeste	2015	40 a 44 anos	3613858	1726819	1887039
Nordeste	2015	45 a 49 anos	3307480	1572046	1735434
Nordeste	2015	50 a 54 anos	2852908	1346602	1506306
Nordeste	2015	55 a 59 anos	2316536	1073575	1242961
Nordeste	2015	60 a 64 anos	1805148	831909	973239
Nordeste	2015	65 a 69 anos	1414266	638001	776265
Sudeste	2015	30 a 69 anos	42394004	20636081	21757923
Sudeste	2015	30 a 34 anos	7288405	3604677	3683728
Sudeste	2015	35 a 39 anos	6818811	3327529	3491282
Sudeste	2015	40 a 44 anos	6058624	2927207	3131417
Sudeste	2015	45 a 49 anos	5719342	2750468	2968874
Sudeste	2015	50 a 54 anos	5310275	2509009	2801266
Sudeste	2015	55 a 59 anos	4651854	2172541	2479313
Sudeste	2015	60 a 64 anos	3719164	1729617	1989547

Figura 2: Tabela tratada e separada por anos e classificação dos dados

Outras transformações, considerando outros padrões foram necessárias:

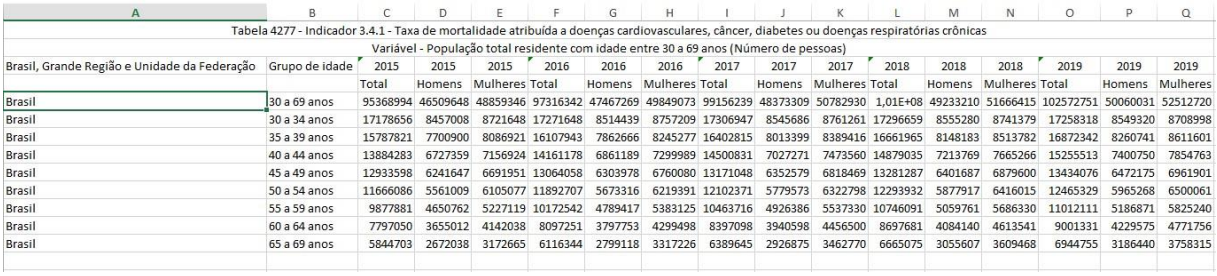


Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas																
Variável - População total residente com idade entre 30 a 69 anos (Número de pessoas)																
Brasil, Grande Região e Unidade da Federação	Grupo de idade	2015	2015	2015	2016	2016	2016	2017	2017	2017	2018	2018	2018	2019	2019	2019
		Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres	Total	Homens	Mulheres
Brasil	30 a 69 anos	95368994	46509648	48859346	97316342	47467269	49849073	99156239	48373309	50782930	1,01E+08	49233210	51666415	102572751	50060031	52512720
Brasil	30 a 34 anos	17178656	8457008	8721648	17271648	8514439	8757209	17306947	8545686	8761261	17296659	8555280	8741379	17258318	8549320	8708998
Brasil	35 a 39 anos	15787821	7700900	8086921	16107943	7862666	8245277	16402815	8013399	8389416	16661965	8148183	8513782	16872342	8260741	8611601
Brasil	40 a 44 anos	13884283	6727359	7156924	14161178	6861189	7299989	14500831	7027271	7473560	14879035	7213769	7665266	15255513	7400750	7854763
Brasil	45 a 49 anos	12933598	6241647	6691951	13064058	6303978	6760080	13171048	6352579	6818469	13281287	6401687	6879600	13434076	6472175	6961901
Brasil	50 a 54 anos	11666086	5561009	6105077	11892707	5673316	6219391	12102371	5779573	6322798	12293932	5877917	6416015	12465329	5965268	6500061
Brasil	55 a 59 anos	9877881	4650762	5227119	10172542	4789417	5383125	10463716	4926386	5537330	10746091	5059761	5686330	11012111	5186871	5825240
Brasil	60 a 64 anos	7797050	3655012	4142038	8097251	3797753	4299498	8397098	3940598	4456500	8697681	4084140	4613541	9001331	4229575	4771756
Brasil	65 a 69 anos	5844703	2672038	3172665	6116344	2799118	3317226	6389645	2926875	3462770	6665075	3055607	3609468	6944755	3186440	3758315

Figura 3: Tabela de indicador de taxa de mortalidade atribuída a doenças antes das alterações

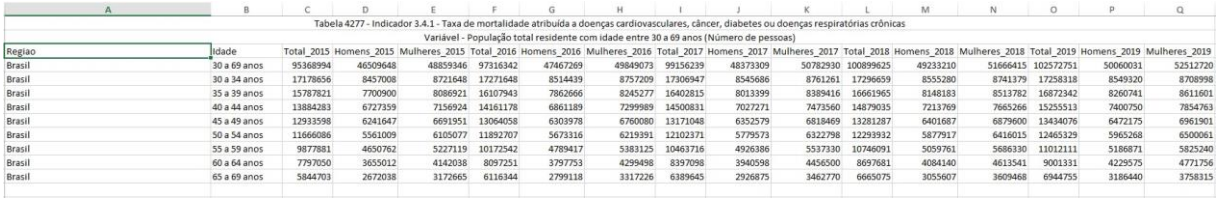


Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas																
Variável - População total residente com idade entre 30 a 69 anos (Número de pessoas)																
Regiao	Idade	Total_2015	Homens_2015	Mulheres_2015	Total_2016	Homens_2016	Mulheres_2016	Total_2017	Homens_2017	Mulheres_2017	Total_2018	Homens_2018	Mulheres_2018	Total_2019	Homens_2019	Mulheres_2019
Brasil	30 a 69 anos	95368994	46509648	48859346	97316342	47467269	49849073	99156239	48373309	50782930	100899625	49233210	51666415	102572751	50060031	52512720
Brasil	30 a 34 anos	17178656	8457008	8721648	17271648	8514439	8757209	17306947	8545686	8761261	17296659	8555280	8741379	17258318	8549320	8708998
Brasil	35 a 39 anos	15787821	7700900	8086921	16107943	7862666	8245277	16402815	8013399	8389416	16661965	8148183	8513782	16872342	8260741	8611601
Brasil	40 a 44 anos	13884283	6727359	7156924	14161178	6861189	7299989	14500831	7027271	7473560	14879035	7213769	7665266	15255513	7400750	7854763
Brasil	45 a 49 anos	12933598	6241647	6691951	13064058	6303978	6760080	13171048	6352579	6818469	13281287	6401687	6879600	13434076	6472175	6961901
Brasil	50 a 54 anos	11666086	5561009	6105077	11892707	5673316	6219391	12102371	5779573	6322798	12293932	5877917	6416015	12465329	5965268	6500061
Brasil	55 a 59 anos	9877881	4650762	5227119	10172542	4789417	5383125	10463716	4926386	5537330	10746091	5059761	5686330	11012111	5186871	5825240
Brasil	60 a 64 anos	7797050	3655012	4142038	8097251	3797753	4299498	8397098	3940598	4456500	8697681	4084140	4613541	9001331	4229575	4771756
Brasil	65 a 69 anos	5844703	2672038	3172665	6116344	2799118	3317226	6389645	2926875	3462770	6665075	3055607	3609468	6944755	3186440	3758315

Figura 4: Tabela de indicador de taxa de mortalidade atribuída a doenças após as alterações.

O processo se repetiu entre esses dois padrões, em todas as tabelas adquiridas.

10.2 Objetivo da Análise Exploratória de Dados

Durante esta fase, nosso principal objetivo foi caracterizar e registrar os datasets que temos à nossa disposição, bem como realizar uma exploração inicial das principais características desses conjuntos de dados.

Número de Exemplares (Linhas) e Dimensões (Colunas): Para cada dataset, verificamos o número de observações (linhas) e variáveis (colunas) presentes. Isso nos proporcionou uma visão inicial da extensão e complexidade dos dados. Os resultados desse levantamento estão resumidos abaixo:

Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas (Probabilidade em %) - Este dataset tem 297 observações e 17 variáveis

Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas (número de pessoas com doenças) - Este dataset tem 297 observações e 17 variáveis

Tabela 4277 - Indicador 3.4.1 - Taxa de mortalidade atribuída a doenças cardiovasculares, câncer, diabetes ou doenças respiratórias crônicas (número de mortes) - Este dataset tem 297 observações e 16 variáveis

Tabela 4408 - Indicador 3.6.1 - Taxa de mortalidade por acidentes de trânsito, por sexo e grupos de idade - Este dataset tem 430 observações e 15 variáveis

Tabela 5845 - Indicador 1.2.1 - Proporção da população vivendo abaixo da linha de pobreza nacional, por cor ou raça - Este dataset tem uma observação e 15 variáveis

Tabela 6825 - Indicador 1.2.1 - Proporção da população vivendo abaixo da linha de pobreza nacional, por sexo - Este dataset possui uma observação e 15 variáveis

Tabela 5878 - Indicador 3.5.2 - Consumo de álcool em litros de álcool puro per capita, por pessoas de 15 anos ou mais de idade - Este dataset possui uma observação e 5 variáveis

Tabela 7876 - Indicador 16.1.1 - Número de vítimas de homicídios intencionais por 100 mil habitantes, por sexo e grupo de idade - Este dataset possui 430 observações e 11 variáveis

Tabela 8191 - Indicador 3.9.2 - Taxa de mortalidade atribuída a fontes de água inseguras, saneamento inseguro e falta de higiene, por sexo e grupo de idade - Este dataset possui 430 observações e 16 variáveis

Tipos de Dados: Identificamos que os datasets contêm uma variedade de tipos de dados, incluindo valores numéricos e categorias.

Medidas de Posição e Dispersão: Calculamos medidas estatísticas descritivas, incluindo média, mediana, quartis, variância e desvio padrão para algumas variáveis-chave em cada dataset. Essas medidas nos forneceram insights sobre as tendências centrais e a variabilidade dos indicadores.

Distribuição e Frequência: Utilizaremos na próxima etapa gráficos e histogramas para visualizar a distribuição dos dados em algumas variáveis. Isso nos ajudará a identificar padrões e assimetrias nos indicadores.

Correlações: Começamos a explorar possíveis correlações entre alguns indicadores, buscando entender as relações entre eles. Esse processo nos permitirá identificar associações que podem ser exploradas em análises posteriores.

Valores Perdidos ou Incorretos: Durante a análise, identificamos a presença de valores ausentes e inconsistências em alguns dos datasets. Essas irregularidades foram tratadas:

Indicador 1.2.1 - Proporção da população vivendo abaixo da linha de pobreza nacional, por sexo: Foi realizada uma modificação para selecionar apenas os registros referentes ao Brasil como país de origem, garantindo a consistência dos dados.

Indicador 1.2.1 - Proporção da população abaixo da linha de pobreza nacional, por grupos de idade: Realizamos uma modificação no campo "Ano", que estava apresentando divergências com o ambiente R. Optamos por mesclar esse campo com os campos "Total", "Homens" e "Mulheres" para garantir a integridade dos dados.

Indicador 1.2.1 - Proporção da população abaixo da linha de pobreza nacional, por grupos de idade: Para solucionar problemas relacionados ao cálculo de média, mediana e outras funções estatísticas, procedemos com a remoção do campo "Idade: 30 a 69 anos", que estava gerando divergências nos resultados.

Indicador 16.1.1 - Número de vítimas de homicídios intencionais por 100 mil habitantes, por sexo: Aplicamos a mesma alteração realizada no conjunto de dados "Indicador 1.2.1 - Proporção da população vivendo abaixo da linha de pobreza nacional, por sexo" para manter a consistência nas informações.

Scripts:

Através de Indicadores de Taxa de Mortalidade x Número de óbitos, desenvolveu-se

- Média

- Mediana

- Quartil

- Variância

- Desvio

11. ESBOÇO DO STORYTELLING

Introdução

A diversidade populacional, heterogeneidade regional, desigualdades socioeconômicas e desafios em saúde pública criam um **contexto relevante** para a **análise da taxa de mortalidade no Brasil**, revelando a interação entre fatores sociais, econômicos e demográficos.

O **IBGE**, uma instituição estatística renomada, foi usado como fonte de dados e desempenha um papel crucial na coleta e divulgação de dados geográficos, contribuindo para o entendimento da realidade brasileira e o desenvolvimento de políticas públicas.

Premissas do Projeto

Para garantir a qualidade e eficácia da nossa análise, estabelecemos algumas premissas-chave:

Fonte de dados confiável | **Foco nas taxas de mortalidade** | **Pensamento Computacional** | **Contribuição para políticas públicas**

Objetivos e Metas

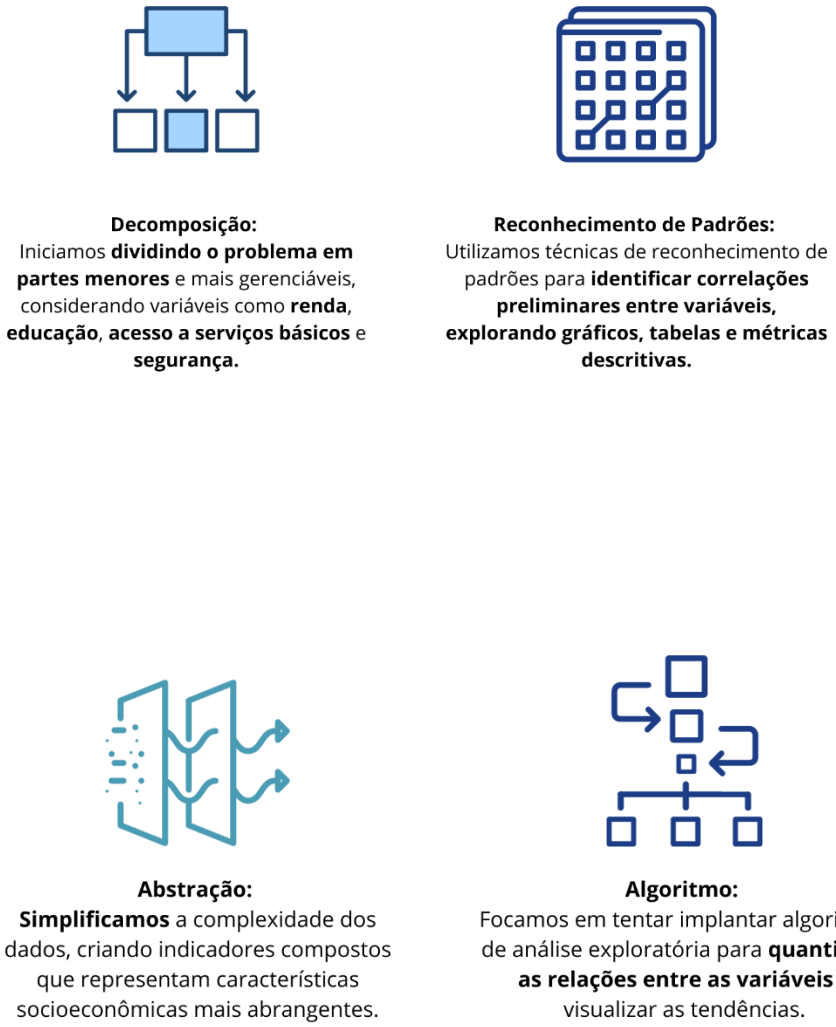
Objetivo principal:
Analisar as relações entre a taxa de mortalidade e diversos fatores sociais, econômicos e demográficos no Brasil.

Metas:
Identificar correlações significativas entre a taxa de mortalidade e outras variáveis; compreender como fatores como renda, acesso a serviços básicos, segurança e saúde estão relacionados à taxa de mortalidade.

Metodologia

Neste projeto, adotamos uma abordagem baseada no pensamento computacional e na análise exploratória de dados, com a finalidade de desvendar insights relevantes sobre as relações entre a

taxa de mortalidade no Brasil e os fatores socioeconômicos, aproveitando ao máximo os dados fornecidos pelo IBGE.



Desafios Enfrentados

Durante o desenvolvimento deste estudo, enfrentamos desafios relacionados à qualidade e ao processamento dos dados. Esses desafios desempenham um papel fundamental na garantia da confiabilidade e validade das análises.

- **Despadronização dos Dados**
- **Dados apresentados de forma inconsistente**
- **Lacuna nos dados**
- **Necessidade de Análise de padrões**

Dados infográficos

Extensão territorial: 8,5 milhões de quilômetros quadrados

PIB: US$ 1,4 trilhão

Diversidade geográfica: 5 regiões (Norte, Nordeste, Centro-Oeste, Sudeste, Sul)

Diversidade cultural: Mais de **200 grupos étnicos**

Biodiversidade: Estima-se que o Brasil abrigue **10-20% de todas as espécies do planeta**

Recursos naturais: Posição líder mundial em reservas de água doce

População total brasileira (2020)

211 milhões

Quantidade de pessoas do sexo feminino
107 milhões

Quantidade de pessoas do sexo masculino
104 milhões

Coeficiente de Gini:
O coeficiente de Gini do Brasil está em torno de 53 em 2020, indicando uma alta desigualdade de renda. Quanto mais próximo de 100, maior a desigualdade..

Concentração de Riqueza:
Cerca de 28% da renda total do Brasil est nas mãos dos 1% mais ricos da população.

Pobreza Multidimensional:
Cerca de 25% da população brasileira enfrenta algum tipo de privação em saúde, educação e padrão de vida.

Taxa de Alfabetização (2020):
93%
O que indica que aproximadamente 93% da população com 15 anos ou mais sabia ler e escrever.

Desafios de Saúde Pública:
O SUS é o sistema de saúde pública do Brasil e oferece atendimento gratuito a toda a população, mas enfrenta desafios relacionados à capacidade e financiamento.

Análise Exploratória de Dados

OBJETIVO INICIAL

Nosso principal objetivo foi caracterizar e registrar os datasets que temos à nossa disposição, bem como realizar uma exploração inicial das principais características desses conjuntos de dados.

PRIMEIROS TRATAMENTOS

Caracterização dos DataSets

Identificação dos tipos de dados

Adequação de dados mesclados e exclusão de dados perdidos/incorretos

Cálculo de medidas de dispersão e posição e exame de distribuição dos dados

Compreensão do cenário dos dados

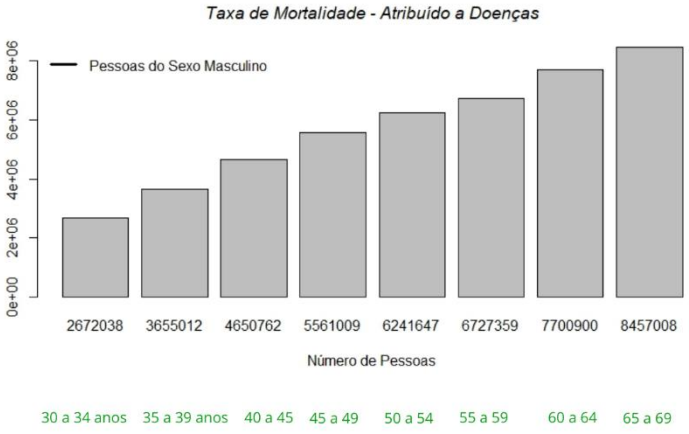
SCRIPTS

Utilizamos o software R para analisar os dados.

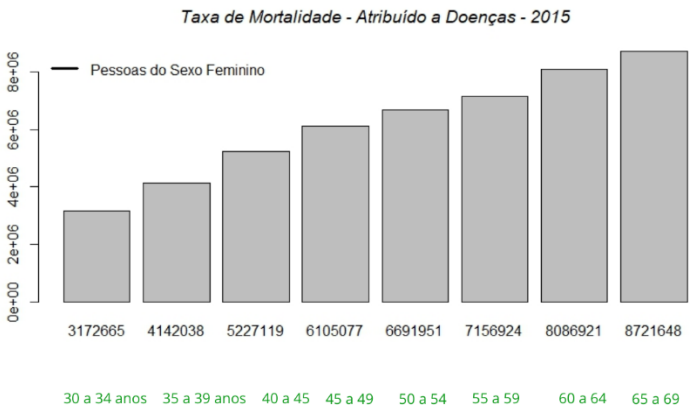
Isso incluiu cálculos de **média, mediana, quartil, variância e desvio padrão.**

```
> mean(npessoas_doenca$Homens_2015)
[1] 5708217
> mean(npessoas_doenca$Mulheres_2015)
[1] 6163043
> mean(obito_doenca$Homens_2015)
[1] 20484.5
> mean(obito_doenca$Mulheres_2015)
[1] 15869.88
> median(npessoas_doenca$Homens_2015)
[1] 5901328
> median(npessoas_doenca$Mulheres_2015)
[1] 6398514
> median(obito_doenca$Homens_2015)
[1] 17595.5
> median(obito_doenca$Mulheres_2015)
[1] 14187
> quantile(npessoas_doenca$Homens_2015)
     0%     25%     50%     75%    100%
2672038 4401825 5901328 6970744 8457008
> quantile(npessoas_doenca$Mulheres_2015)
     0%     25%     50%     75%    100%
3172665 4955849 6398514 7389423 8721648
> quantile(obito_doenca$Homens_2015)
     0%     25%     50%     75%    100%
 3247.00  7216.00 17595.50 32437.75 43673.00
> quantile(obito_doenca$Mulheres_2015)
     0%     25%     50%     75%    100%
 2982.00  6778.75 14187.00 23875.00 32882.00
> var(npessoas_doenca$Homens_2015)
[1] 3.917679e+12
> var(npessoas_doenca$Mulheres_2015)
[1] 3.631635e+12
> var(obito_doenca$Homens_2015)
[1] 243342268
> var(obito_doenca$Mulheres_2015)
[1] 122251824
> # Usando a Função "sd" para encontrar o Desvio Padrão de maneira individual
> sd(npessoas_doenca$Homens_2015)
[1] 1979313
> sd(npessoas_doenca$Mulheres_2015)
[1] 1905685
> sd(obito_doenca$Homens_2015)
[1] 15599.43
> sd(obito_doenca$Mulheres_2015)
[1] 11056.75
>
```

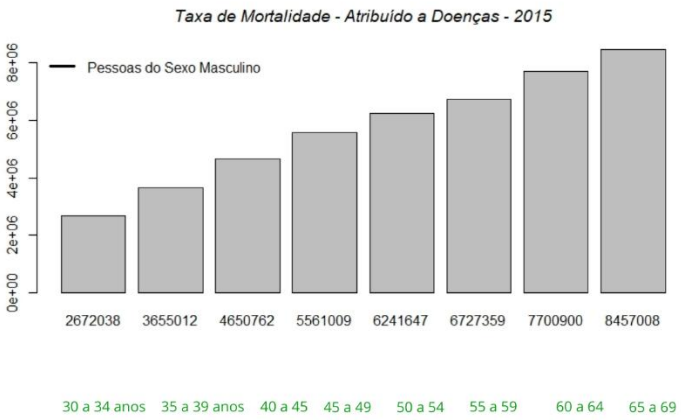
RESULTADOS DA ANÁLISE EXPLORATÓRIA



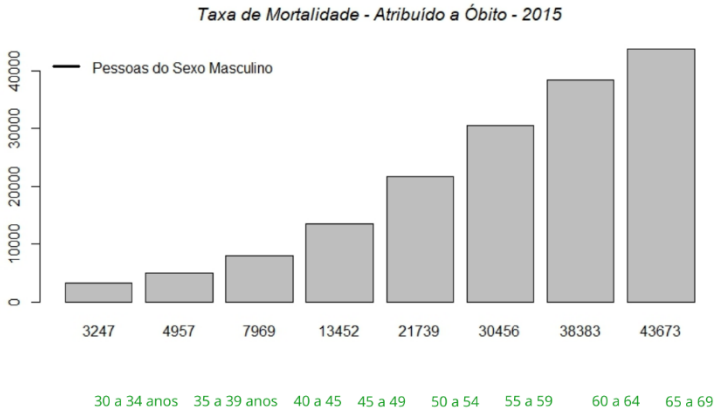
RESULTADOS DA ANÁLISE EXPLORATÓRIA

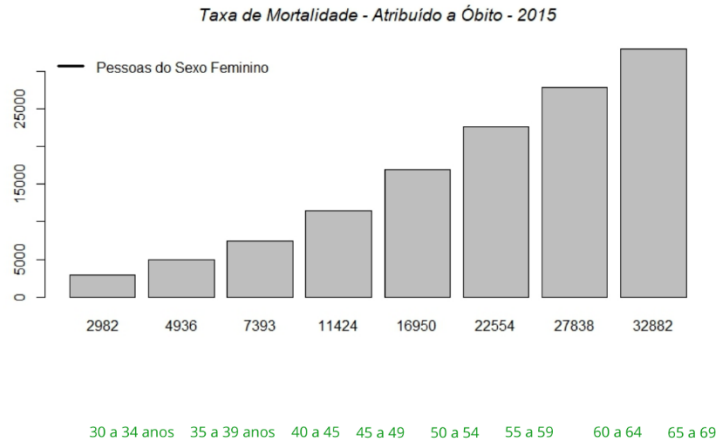


RESULTADOS DA ANÁLISE EXPLORATÓRIA

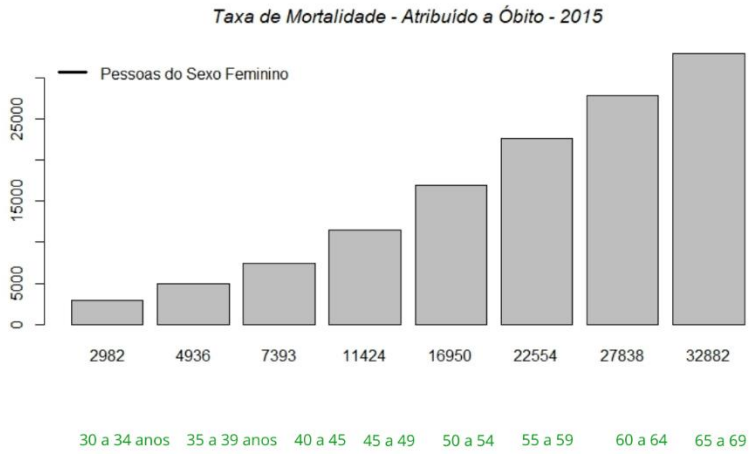


RESULTADOS DA ANÁLISE EXPLORATÓRIA

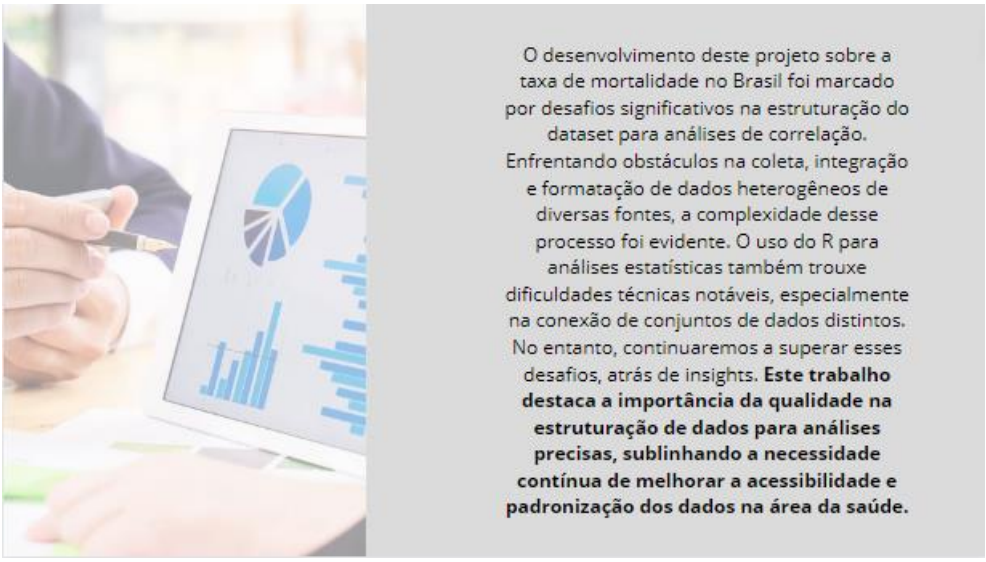




RESULTADOS DA ANÁLISE EXPLORATÓRIA



Conclusão



12. LINK DO GITHUB

https://github.com/Projeto-Aplicado-I-Mackenzie

13. LINK DO VÍDEO NO YOUTUBE

https://www.youtube.com/watch?v=MhJgkPjV0O4

Referências Bibliográficas

Instituto Brasileiro de Geografia e Estatística (IBGE)

Instituto Brasileiro de Geografia e Estatística. Dados Demográficos e Socioeconômicos do Brasil. Disponível em: <https://sidra.ibge.gov.br/acervo#/S/C2/T/QL>. Acesso em: 18 de agosto de 2023.

Departamento de Informática do Sistema Único de Saúde (DATASUS)

Departamento de Informática do Sistema Único de Saúde. Dados de Saúde e Indicadores Epidemiológicos. Disponível em: <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/obt10uf.def>. Acesso em: 24 de agosto de 2023.

Pesquisa Nacional por Amostra de Domicílios (PNAD)

Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional por Amostra de Domicílios Contínua. Disponível em: < https://www.ibge.gov.br/estatisticas/sociais/populacao/9127-pesquisa-nacional-por-amostra-de-domicilios.html>. Acesso em: Acesso em: 24 de agosto de 2023.

Censo Demográfico:

Instituto Brasileiro de Geografia e Estatística. Censo Demográfico 2020. Disponível em: < https://www.ibge.gov.br/estatisticas/sociais/trabalho/22827-censo-demografico-2022.html>. Acesso em: Acesso em: 24 de agosto de 2023.