Metodologia

1. Fonte de Dados

A análise foi realizada com base em dados públicos extraídos de diferentes fontes governamentais e acadêmicas, garantindo uma abordagem abrangente e confiável. As principais bases de dados utilizadas foram:

- Dados de acidentes em rodovias federais: Informações detalhadas sobre acidentes, vítimas, circunstâncias e características dos veículos envolvidos, disponibilizadas pela Polícia Rodoviária Federal (PRF) (link).
- Dados populacionais por estado: Informações sobre a população residente em cada unidade federativa, extraídas da Base dos Dados (link).
- Dados da frota de veículos: Estatísticas sobre o número de veículos registrados por estado, disponíveis na Base dos Dados (link).
- Dados dos indicadores sociais: Métricas referentes aos índices como IDHM, IVS etc. (link)
- Referência acadêmica: O estudo "A letalidade dos acidentes de trânsito nas rodovias federais brasileiras em 2016", publicado na Revista Brasileira de Estudos de População (RBEP), serviu como base metodológica para a segmentação dos dados e cálculos estatísticos (link).
- Referência acadêmica: O estudo "Acidentes de trânsito com vítimas: sub-registro, caracterização e letalidade", Cadernos de Saúde Pública, v. 19, n. 4, p. 979-986, 2003, aponta uma maior frequência aos finais de semana (link).
- Referência acadêmica: O estudo "Serviço de Atendimento Móvel de Urgência: Um observatório dos acidentes de transportes terrestre em nível local", Revista Brasileira de Epidemiologia, v. 14, n. 1, p. 3-14, 2011., aponta uma maior frequência de atendimentos móveis hospitalares nos finais de semana, entre 18h00 e 5h59 (noite e madrugada). (link).
- Referência acadêmica: O estudo "ÍNDICE DE DESENVOLVIMENTO HUMANO MUNICIPAL (IDHM) NO BRASIL: UM RELATO DA DESIGUALDADE ENTRE EDUCAÇÃO, RENDA E LONGEVIDADE" aponta que os Estados do Nordeste e Norte apresentam consistentemente os piores índices de IDHM do país (link)
- Referência acadêmica: O estudo "Vulnerabilidade Social, Fome e Pobreza Nas Regiões Norte e Nordeste Do Brasil" indica que as regiões do Nordeste e Norte possuem a população mais carente e socialmente vulnerável do país (link)

2. Seleção e Limpeza dos Dados

Para garantir a qualidade da análise, foram realizadas etapas de pré-processamento dos dados:

- Exclusão de indivíduos não expostos ao risco de morte: Seguindo a lógica metodológica do estudo acadêmico referenciado, foram removidos os registros de indivíduos classificados como ilesos, focando apenas naqueles que sofreram algum impacto no acidente.
- Tratamento de valores ausentes e inconsistentes: Foram eliminadas ou corrigidas informações incompletas, garantindo maior confiabilidade nas estatísticas.
- Conversão e padronização de colunas: Ajuste dos formatos de data, categorização de tipos de veículos e agrupamento de variáveis para facilitar as análises.

3. Indicadores e Técnicas Estatísticas

A análise exploratória envolveu a criação de métricas-chave e aplicação de testes estatísticos para entender padrões e correlações. Entre as principais abordagens utilizadas:

- Percentual de fatalidade: Cálculo da proporção de vítimas fatais em relação ao total de vítimas expostas ao risco.
- Análise de tendências: Construção de gráficos temporais para observar a evolução dos acidentes e óbitos entre 2007 e 2023.
- Segmentação geográfica: Comparação entre estados e regiões para identificar locais com maior incidência de acidentes e mortes.

4. Visualização e Comunicação dos Resultados

Os resultados foram representados por meio de:

- Gráficos interativos e dashboards, desenvolvidos para facilitar a interpretação dos dados.
- Tabelas comparativas, destacando variações entre períodos e categorias de acidentes.

Dessa forma, a metodologia adotada permitiu uma análise detalhada e fundamentada dos padrões de acidentes em rodovias federais brasileiras, auxiliando na formulação de insights para políticas públicas e estratégias de mitigação de riscos.

Normalização dos Dados

Estado Físico das Vítimas:
Afim de normalizar os dados, agrupamos os casos em que a vítima possui feridas leves e os casos em que possui feridas graves para um estado físico em comum: "SOBREVIVENTE". Após a normalização, os casos onde o estado físico é ileso correspondem a **2.779.606 registros** (60% do total do conjunto de dados), os casos onde o estado físico é sobrevivente correspondem a 1.567.635 registros (34%), os casos de óbito a 121.376 registros (3%) e os casos não identificados "N.I" a 189.507 registros (4%). Para a análise descritiva são desconsiderados os campos com "N.I" e para o cálculo do percentual de fatalidade são excluídos os estados físicos "ILESOS" mantendo somente os casos onde houve risco real à vida (vítimas com algum tipo de ferimento – sejam feridas leves ou graves);

Normalização da Variável Sexo:
- Afim de garantir a limpeza e normalização dos dados na variável sexo, aplicamos um conjunto de substituições. Após esse processo, o sexo masculino corresponde a 3.598.900 registros (77% do total do conjunto de dados), o sexo feminino corresponde a 857.697 registros (18%) e os casos sem a identificação "N.I" do sexo a 201.527 registros (4%). Para a análise descritiva são desconsiderados os campos com "N.I";

Tipo do Veículo:
Normalizamos a coluna de tipo do veículo pois o conjunto original possui uma grande variedade de veículos e com a maior parte com pouco volume de dados e baixa proporção do todo. As variações resultantes são: AUTOMÓVEL, CAMINHÃO, MOTOCICLETA, ÔNIBUS, N.I (Não Identificado) e OUTROS. Para a análise descritiva são desconsiderados os campos com "N.I" e "OUTROS".

Criação da coluna Período:
Afim de garantir a normalização dos dados na variável horário, criamos segmentos representando faixas de tempo. De 00:00:01 até 06:00:00 temos a "MADRUGADA", de 06:00:01 até 12:00:00 a "MANHÃ", de 12:00:01 até 18:00:00 a "TARDE" e de 18:00:01 até 23:59:59 a "NOITE".

Criação da coluna Região:
Afim de garantir a normalização e segmentação dos dados, criamos a variável região que representa a região geográfica do país. Com ela realizamos análise espacial e a identificação de outliers.

Normalização da coluna Uso do Solo:
Afim de garantir a normalidade dos dados segmentamos a variável uso do solo com duas opções: "URBANO" para áreas urbanizada e "RURAL" para áreas rurais.