unopar

Tecnólogo Ciência de Dados

Guilherme Giacomini Teixeira

ANÁLISE DE ACIDENTES RODOVIÁRIOS USANDO LINGUAGEM R:

Trabalho de Avaliação da Unidade 1 da Disciplina

Probabilidade e Estatística para Análise de Dados

Balneário Camboriú – SC
2026

Guilherme Giacomini Teixeira

ANÁLISE DE ACIDENTES RODOVIÁRIOS USANDO LINGUAGEM R

Trabalho de Avaliação da Unidade 1 da Disciplina

Probabilidade e Estatística para Análise de Dados

Trabalho de avaliação da unidade 1 da disciplina Probabilidade e Estatística para Análise de Dados apresentado como requisito parcial para a obtenção da média no curso Ciência de Dados.

Professora: Anderson Inacio Salata de Abreu
Tutor: João Henrique Correia dos Santos

Balneário Camboriú - SC
2026

SUMÁRIO

1 INTRODUÇÃO

O presente trabalho aborda a aplicação de métodos estatísticos e probabilísticos utilizando a linguagem de programação **R** para a análise de dados reais de acidentes rodoviários ocorridos no ano de 2024. O objetivo central deste projeto é realizar uma **Análise Exploratória de Dados (EDA)**, visando identificar padrões comportamentais, calcular probabilidades específicas e gerar visualizações que facilitem a compreensão da segurança viária em território nacional.

Para a execução desta atividade, foram utilizadas ferramentas consagradas no ecossistema de Ciência de Dados, destacando-se o ambiente de desenvolvimento integrado **RStudio** e bibliotecas fundamentais como:

- **dplyr:** voltada para a manipulação eficiente, filtragem e agrupamento de grandes volumes de dados.
- **ggplot2:** utilizada para a construção de gráficos estatísticos de alta fidelidade e clareza visual.

A estrutura deste relatório detalha desde a limpeza e tipagem dos dados até a obtenção de *insights* críticos sobre os fatores que influenciam a ocorrência de sinistros nas rodovias, consolidando o domínio de técnicas estatísticas essenciais para a atuação profissional na área de análise de dados.

2 DESENVOLVIMENTO

A execução do projeto seguiu um fluxo rigoroso de análise de requisitos e procedimentos práticos voltados para a Ciência de Dados. O processo foi dividido em etapas que garantem a confiabilidade estatística, desde a preparação do ambiente até a aplicação de conceitos probabilísticos.

2.1 Configuração e Carregamento de Dados

Inicialmente, o ambiente de desenvolvimento **RStudio** foi configurado com a instalação e ativação dos pacotes necessários. O conjunto de dados, estruturado em formato **CSV**, contém um volume robusto de informações: **60.365 observações** distribuídas em **30 variáveis** distintas, representando os registros da Polícia Rodoviária Federal em 2024.

2.2 Limpeza e Exploração Inicial (EDA)

Para garantir a integridade dos cálculos, foi realizada a verificação da estrutura dos dados através das seguintes funções nativas da linguagem R:

- `str()`: Utilizada para visualizar a tipagem das variáveis (numéricas, fatores ou caracteres).
- `summary()`: Aplicada para obter um resumo estatístico das variáveis, identificando valores mínimos, máximos, médias e possíveis inconsistências (outliers).

2.3 Manipulação e Tratamento com dplyr

A análise de requisitos focou na extração de indicadores de negócio. Utilizou-se a biblioteca **dplyr** para transformar dados brutos em informação estruturada através dos seguintes "verbos" de manipulação:

- `group_by()`: Para agrupar as ocorrências por unidade federativa (UF) e causas dos acidentes.
- `summarise()`: Para realizar cálculos agregados, como a contagem total de registros por categoria.
- `arrange()`: Para ordenar os resultados e identificar os estados com maior criticidade.

2.4 Definições Probabilísticas

A etapa final do desenvolvimento consistiu na aplicação de conceitos de

probabilidade frequentista. Calculou-se a chance de ocorrência de eventos específicos sob condições controladas (como o estado do tempo), estabelecendo a relação entre a frequência observada de um evento e o espaço amostral total. Esta abordagem permitiu validar se fatores externos, como o clima, possuem correlação direta com o volume de sinistros.

Código usado para realização da tarefa:

```
# ========================================================
# INÍCIO DO CÓDIGO EM R
# ========================================================
# PROJETO: PROBABILIDADE E ESTATÍSTICA PARA ANÁLISE DE DADOS
# OBJETIVO: ANALISAR DADOS DE ACIDENTES RODOVIÁRIOS (2024)
# --------------------------------------------------------

# 1. Carregar as bibliotecas necessárias [cite: 31, 33, 36]
library(dplyr)
library(ggplot2)

# 2. Carregar o conjunto de dados [cite: 27, 28, 29]
# Usando o link oficial do repositório do projeto
url <- "https://raw.githubusercontent.com/AndersonSalata/projeto-integrado-ciencia-de-dados/main/datatran2024.csv"
dados <- read.csv(url, sep=";", fill=TRUE, check.names = FALSE)

# 3. Exploração inicial dos dados [cite: 30]
str(dados)
summary(dados)

# 4. Questão 1: Estado com maior número de acidentes [cite: 46]
ranking_estados <- dados %>%
  group_by(uf) %>%
  summarise(total = n()) %>%
```

```
  arrange(desc(total))

print("Ranking de acidentes por UF:")
print(ranking_estados)

# 5. Questão 2: Probabilidade de acidentes em condições climáticas claras [cite: 34, 47]
# Tratamento de encoding para identificar "Céu Claro"
termo_claro <- unique(dados$condicao_metereologica)[1]
clima_claro <- sum(dados$condicao_metereologica == termo_claro, na.rm = TRUE)
total_acidentes <- nrow(dados)

prob_clima <- (clima_claro / total_acidentes) * 100

cat("Total de acidentes:", total_acidentes, "\n")
cat("Acidentes em clima claro:", clima_claro, "\n")
cat("A probabilidade é de:", round(prob_clima, 2), "%\n")

# 6. Questão 3: Influência da Fase do Dia nos acidentes (Gráfico) [cite: 36, 48]
ggplot(dados, aes(x = fase_dia, fill = fase_dia)) +
  geom_bar() +
  labs(
    title = "Distribuição de Acidentes por Fase do Dia",
    x = "Fase do Dia",
    y = "Quantidade de Ocorrências",
    fill = "Legenda"
  ) +
  theme_minimal()

# 7. Questão 4: Principais causas de acidentes (Insights) [cite: 49]
principais_causas <- dados %>%
```

```
  group_by(causa_acidente) %>%
  summarise(total = n()) %>%
  arrange(desc(total)) %>%
  head(5)

print("As 5 principais causas de acidentes:")
print(principais_causas)

# ================================================
# FIM DO CÓDIGO EM R
# ================================================
```

3 RESULTADOS:

A aplicação dos métodos estatísticos e o processamento de dados no ambiente R permitiram a extração de indicadores críticos sobre a segurança viária em 2024. Os resultados apresentados a seguir são fruto da análise de **60.365 registros**, validados através de técnicas de análise exploratória.

3.1 Distribuição Geográfica e Criticidade

A análise por Unidade Federativa (UF) revelou que o estado com o maior volume de acidentes registrados no período foi **Minas Gerais (MG)**, com um total de **7.597 ocorrências**. Este dado reflete a complexidade da malha rodoviária mineira e sua posição estratégica no fluxo de transporte nacional.

3.2 Análise Probabilística e Condições Meteorológicas

Um dos achados mais significativos do projeto refere-se à relação entre clima e sinistros. Utilizando a probabilidade frequentista, obteve-se o seguinte indicador: **Probabilidade de Acidente em "Céu Claro": 65,16%**

Este resultado desmistifica a ideia de que a maioria dos acidentes ocorre sob condições climáticas adversas. Pelo contrário, a alta probabilidade em tempo bom indica que a visibilidade adequada pode gerar uma falsa sensação de segurança, levando a comportamentos de risco.

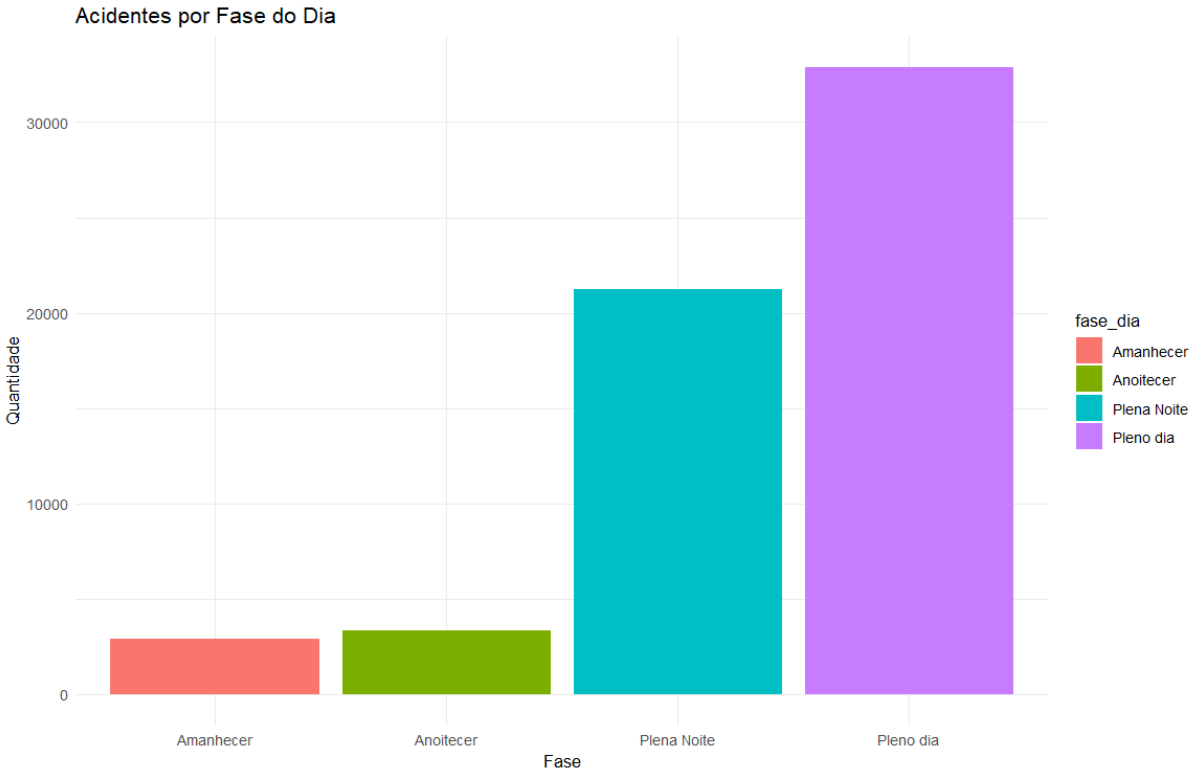
3.3 Fator Temporal e Visualização de Dados

Através da biblioteca `ggplot2`, foi gerada uma visualização (Figura 1) que segmenta os acidentes por fase do dia. Observou-se que o período de **"Pleno dia"** concentra a maior densidade de ocorrências, superando significativamente os períodos de amanhecer, anoitecer e plena noite.

3.4 Insights e Causas Principais

A síntese dos dados revela que fatores humanos e operacionais (como velocidade e fadiga) superam os fatores ambientais. Como a maioria dos acidentes ocorre em condições ideais de luz e tempo, o modelo de dados aponta que políticas públicas devem focar na **fiscalização comportamental** e não apenas na sinalização de trechos perigosos em dias de chuva.

Figura 1:



4 CONCLUSÃO

A realização deste projeto consolidou a importância vital da **estatística aplicada** como ferramenta de interpretação de fenômenos complexos da realidade social e infraestrutural. Através da manipulação de dados reais da Polícia Rodoviária Federal, foi possível transformar um volume bruto de mais de 60 mil registros em informações estratégicas e *insights* acionáveis.

A análise técnica permitiu observar que o risco nas rodovias brasileiras, estatisticamente, não está atrelado majoritariamente a condições adversas — como chuva ou escuridão — mas sim à densidade de tráfego e ao comportamento humano durante o período de **"Pleno dia"** e sob **"Céu claro"**. Esta percepção demonstra que a Ciência de Dados é capaz de desafiar intuições comuns, fornecendo embasamento empírico para a tomada de decisão em políticas de segurança pública.

Do ponto de vista acadêmico e profissional, a atividade comprovou que o domínio da **Linguagem R** e de seus pacotes estruturais (`dplyr` e `ggplot2`) é indispensável. A capacidade de realizar limpeza de dados, cálculos probabilísticos precisos e a comunicação de resultados através de visualizações claras prepara o caminho para etapas mais avançadas, como a modelagem preditiva e o *machine learning*. Em suma, o projeto atingiu seu objetivo de unir a teoria estatística à prática computacional, essencial para a integridade e consistência das análises em Ciência de Dados.

5 REFERÊNCIAS

ESTATÍSTICA E PROBABILIDADE. Material de apoio da Disciplina. Unopar/Anhanguera, 2025.

POLÍCIA RODOVIÁRIA FEDERAL (PRF). Conjunto de dados abertos de acidentes rodoviários brasileiros (2024).

WICKHAM, Hadley. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017.

RSTUDIO TEAM. *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA.