Análise da base US accidents

Ana Beatriz da Silva Maques João Victor Pietchaki
Leonardo Eizo Sakai Vitor Pignatari

2024-12-15

A principal atividade de um estatístico e cientista de dados é a análise de dados, quaisquer sejam as complexidades envolvidas. Tal consiste em um estudo pormenorizado de um conjunto de informações a fim de gerar conhecimento, instigar curiosidade, produzir ideias, detectar padrões a partir do que antes não tinha significado nenhum. Este presente artigo publica um conjunto de análises estatísticas sobre a base de dados US Accidents. Foi elaborado por estudantes do curso de Bacharelado em Estatística e Ciência de Dados da UFPR para a disciplina "Elementos de Programação para Estatística", a fim de que aprimorem suas habilidades como analistas. As análises foram feitas utilizando a linguagem de programação R. Com ela, dados foram manipulados, análises exploratórias e estatísticas realizadas, gráficos elaborados, este presente relatório feito, e o dashboard desenvolvido. Dessa forma, comunicaremos informações de 7 milhões de linhas de forma intuitiva e clara, evidenciando tendências e detectando padrões entre a ocorrência de acidentes e as variáveis que os acompanham.

Índices

0.1 Introdução

A análise de dados é uma atividade central para estatísticos e cientistas de dados, permitindo transformar informações aparentemente desorganizadas em conhecimento útil, detectar padrões e gerar insights. Este artigo apresenta um conjunto de análises estatísticas sobre a base de dados US Accidents, elaborado por estudantes do curso de Bacharelado em Estatística e Ciência de Dados da UFPR para a disciplina “Elementos de Programação para Estatística”. As análises foram conduzidas utilizando a linguagem de programação R, abrangendo etapas de manipulação de dados, análise exploratória, construção de gráficos, desenvolvimento de um dashboard interativo em Shiny e a produção deste relatório técnico em formato de artigo científico. Com base nos 7 milhões de registros da base, foram identificadas tendências e padrões na ocorrência de acidentes e suas variáveis associadas, comunicados de maneira clara e intuitiva. Este trabalho visa aprimorar as habilidades analíticas dos estudantes e contribuir para a compreensão dos dados relacionados a acidentes nos Estados Unidos.

0.2 Materiais e Métodos

0.2.1 Fonte de Dados

Para este projeto, utilizamos o linguagem de programação R para analisar a base de dados US Accidents, a partir de um aplicativo Shiny para a criação de um dashboard interativo, permitindo tanto uma visão abrangente dos dados, quanto informações específicas sobre o comportamento das variáveis inter-relacionada e autonomamente. A base de dados pode ser encontrada em https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents, e foi coletada utilizando APIs de trânsito em tempo real, nos Estados Unidos entre os anos de 2016 e 2023

(Moosavi, Samavatian, Parthasarathy, & Ramnath, 2019), (Moosavi, Samavatian, Parthasarathy, Teodorescu, et al., 2019).

0.2.2 A Base de Dados US_accidents

A base de dados consiste em 46 variáveis e aproximadamente 7,8 milhões de observações sobre acidentes de trânsito nos Estados Unidos entre fevereiro de 2016 e março de 2023, com informações sobre a severidade do acidente, posição geográfica, data e hora do ocorrido, descrição, endereço, condições climáticas, infraestrutura do local e condições de luminosidade no momento do ocorrido. O programa R foi utilizado com os pacotes tidyverse para a importação e limpeza dos dados e análise exploratória, ggplot para a criação de gráficos, leaflet para a criação de visualizações em mapas, e shiny para a criação do dashboard. Em conjunto com Shiny, foi utilizado o pacote bslibs, que usa versões mais atualizadas do Bootstrap, permitindo a criação de layouts mais modernos e flexíveis.

A fim de começar a entender e analisar os acidentes nos EUA, iniciamos lendo o banco de dados `US_Accidents_March23.csv`. Os dados foram coletados por Sobhan Moosavi entre 01/01/2016 e 04/01/2023. Para otimizar o tempo de carregamento, salvamos o arquivo original dos dados em um de extensão `.Rdata`.

```
# Carrega os dados originais
# load("../../99_dados/acidentes_US.RData")
load("../../99_dados/acidentes_US_sample.RData")
# load("../../99_dados/acidentes_US_date.RData")
load("../../99_dados/acidentes_US_date_sample.RData")
st_read("../../99_dados/Estados_US/tl_2024_us_state.shp")
acidentes_US_sample %>% class()
```

A fim de começarmos a entender esse conjunto de dados, vamos verificar a estrutura dos elementos presentes no objeto `acidentes_US`.

```
acidentes_US_sample %>% glimpse()
```

O conjunto de dados dispõe de 7728394 observações e 46 variáveis. Cada coluna é uma variável que descreve um aspecto do acidente, ao passo que cada linha corresponde única e exclusivamente a uma observação de acidente.

```
acidentes_US_sample %>% head(10)
```

0.2.2.1 Grupos de Variáveis

Olhando em um primeiro momento, todas essas informações nos confundem. Para contornar esse problema, identificamos grupos de variáveis que podem ser analisadas em conjunto e excluímos outras que observamos não contribuir significativamente.

> Escolhemos não trabalhar com as colunas `ID`, `Source`, `Timzone`, `Airport_code`, `Weather_Timestamp`, `Sunrise_Sunset`, `Nautical_Twilight`, `Astronomical_Twilight`, `zipcode`, `End_Lat`,`End_Lng`, `Street`, `County`, `Wind_Direction`, `Weather_Timestamp`, `Description`.

Para otimizar nossas análises, então, definimos 5 grandes grupos de variáveis: tempo, localização, condições climáticas, condições de infraestrutura e sinalização, descrição humana do acidente.

A coluna de `severity` não foi agrupada e será analisada interconjuntamente.

As variáveis que compõem cada grupo são:

- Tempo:
  - `Start_Time`, `End_Time`, `Civil_Twilight`;
- Localização:
  - `Start_Lat`,`Start_Lng`, `End_Lat`, `End_Lng`, `Distance(mi)`, `Street`, `City`, `County`, `State`, `Zipcode`, `Country`;
- Condiçoes Climaticas:
  - `Temperature(F)`, `Wind_Chill(F)`, `Humidity(%)`, `Pressure(in)`, `Visibility(mi)`, `Wind_Speed(mph)`, `Precipitation(in)`, `Weather_Condition`;
- Condiçoes de Infraestrutura e Sinalizaçao:
  - `Bump`, `Crossing`, `Give_Way`, `Junction`, `No_Exit`, `Railway`, `Roundabout`, `Station`, `Stop`, `Traffic_Calming`, `Traffic_Signal`, `Turning_Loop`; ...
- Descrição Humana do Acidente:
  - `Description`.

0.2.3 Manipulação de dados

Definimos que cada uma das cinco abas seria de cada um dos grupos de variáveis. Para isso, tivemos de fazer um pré-processamento dos dados para:

- otimizar o tempo de processamento de scripts;
- filtrar colunas que não seriam utilizadas;

- tratar valores faltantes;
- transformar variáveis para formatos mais adequados;
- criar novas variáveis;
- sumarizar os dados.

0.2.3.1 Pré-processando as variáveis do grupo 'Tempo'

O primeiro passo foi filtar as colunas que fazem parte do grupo 'Tempo' e criar uma nova tabela com essas variáveis.

```
# Seleciona as variaveis de 'Tempo'
acidentes_US_date <- acidentes_US_sample[, .(ID, Severity, Start_Time, End_Time, Civil_Twili

acidentes_US_date <- acidentes_US_sample[, `:=`(Start_Time = Start_Time,
                                                End_Time = End_Time,
                                                Civil_Twilight = Civil_Twilight)]

acidentes_US_date <- acidentes_US_date[, .(ID, Severity, Start_Time, End_Time, Civil_Twilight
```

Com apenas as colunas de interesse, a tabela `acidentes_US_date` foi criada. A partir dela, criamos novas variáveis para enriquecer a análise dos dados.

Esse data.table `acidentes_US_date` foi salvo em um arquivo `. RData` para ser utilizado posteriormente. Ele será a tabela base para todas as análises e gráficos realizados para o grupo 'Tempo'.

0.2.3.2 Pré-Processando as variáveis do grupo posição geográfica

Para a visualização com o mapa, foi consultada uma base externa no site, sobre os *shapefiles* dos estados. Assim, tendo os limites dos estados, foi possível criar uma paleta de cores de acordo com os valores da quantidade de acidentes que cada estado teve.

Para a plotagem do mapa de calor na visualização geográfica, foi calculada a localização aproximada do acidente, arredondando a latitude e longitude para que a rfenderização do código fosse mais rápida.

```
# Alterando o nome, para não ter que alterar todas as variáveis após
acidentes_arredondado <- acidentes_US_sample

# Shapefiles
estados <- st_read("../../99_dados/Estados_US/tl_2024_us_state.shp")

# Arredondando latitudes e longitudes
```

```r
acidentes_arredondado$Start_Lat <- round(acidentes_arredondado$Start_Lat,2)
acidentes_arredondado$Start_Lng <- round(acidentes_arredondado$Start_Lng,2)

# Alterando a base shp para facilitar o entendimento
nomeestados <- as.data.frame(estados) %>% select(STUSPS,NAME)
names(nomeestados) <- c("STUSPS","NOMEESTADO")

# Juntando as informações já tratadas
acidentes <- left_join(acidentes_arredondado,nomeestados, by = c("State"="STUSPS"))

acidentes_p_estado <- acidentes %>%
  group_by(State, NOMEESTADO) %>%
  summarise(n = n(), .groups = "drop")


# Juntando com a geometria dos estados
acidentes_p_estado <- left_join(acidentes_p_estado, estados, by = c("State" = "STUSPS"))
acidentes_p_estado <- st_as_sf(acidentes_p_estado)

# Substituindo NA por 0 para estados sem acidentes
acidentes_p_estado$n[is.na(acidentes_p_estado$n)] <- 0

mapa_de_calor <- acidentes %>%
  group_by(Start_Lat, Start_Lng) %>%
  summarise(n = n(), .groups = "drop")

cores <- c("white", "red", "darkred")
palestado <- colorNumeric(palette = cores, domain = c(0, max(acidentes_p_estado$n, na.rm = T
```

0.2.3.3 Pré-processando as variáveis do grupo 'Condições Climáticas'

A variável 'Weather Condition' possuía muitos possíveis valores, o que a deixava difícil analisar. O primeiro passo foi agrupar os valores parecidos.

```r
#Agrupando valores parecidos em 'Weather Condition'.

acidentes_US_sample <- acidentes_US_sample %>%
  mutate(Weather_Cluster = case_when(
    grepl("Rain|Drizzle|Shower|Showers|Light Rain|Heavy Rain|Rain Showers|Light Rain Showers
    grepl("Snow|Blowing Snow|Light Snow|Heavy Snow|Snow Grains|Snow Showers|Snow and Thunder
    grepl("Thunder|Tornado|Thunderstorm|Light Thunderstorm|Heavy Thunderstorms and Rain|Thun
    grepl("Wind|Windy|Blowing Dust|Blowing Snow|Duststorm|Dust Whirls|Blowing Dust / Windy|W
```

```
    grepl("Fog|Haze|Shallow Fog|Patches of Fog|Mist|Light Fog|Fog / Windy|Partial Fog", Weath
    grepl("Clear|Fair|Mostly Cloudy|Partly Cloudy|Overcast|Cloudy", Weather_Condition, ignore
    grepl("Smoke|Volcanic Ash|Haze", Weather_Condition, ignore.case = TRUE) ~ "Hazardous",
    grepl("Sleet|Ice Pellets|Freezing Rain|Light Freezing Rain|Freezing Drizzle|Light Freezi
    grepl("Squalls|Hail|Small Hail|Thunder / Wintry Mix|Hail", Weather_Condition, ignore.case
    grepl("Mist|Blowing Snow Nearby|Sand|Blowing Snow", Weather_Condition, ignore.case = TRU
    grepl("Volcanic Ash|Blowing Sand|Sand", Weather_Condition, ignore.case = TRUE) ~ "Dust St
    grepl("Fair / Windy|Fair", Weather_Condition, ignore.case = TRUE) ~ "Fair / Windy",
    TRUE ~ "Other"
  ))
```

Depois, foi criada uma nova variável para identificar se choveu choveu ou não.

```
acidentes_US_sample <- acidentes_US_sample %>%
  mutate(Choveu = ifelse(`Precipitation(in)` > 0, "Sim", "Não"))
```

0.2.3.4 Pré-processando as variáveis do grupo 'Condiçoes de Infraestrutura e Sinalizaçao'

```
# Seleciona as variáveis de interesse
poi <- acidentes_US_sample[, .(Severity, Start_Time, City, State, `Temperature(F)`,
                           Amenity, Bump, Crossing, Give_Way, Junction, No_Exit,
                           Railway, Roundabout, Station, Stop, Traffic_Calming,
                           Traffic_Signal, Turning_Loop)]

# Cria duas seleções de dados, contando a quantidade de acidentes com e sem cada ponto de int

poi_true <- poi %>%
  select(Amenity:Turning_Loop) %>%
  pivot_longer(Amenity:Turning_Loop, names_to = "PoI") %>%
  filter(value == 1) %>%
  group_by(PoI) %>%
  summarise(t = n()) %>%
  arrange(desc(t))

poi_false <- poi %>%
  select(Amenity:Turning_Loop) %>%
  pivot_longer(Amenity:Turning_Loop, names_to = "PoI") %>%
  filter(value == FALSE) %>%
  group_by(PoI) %>%
  summarise(f = n()) %>%
```

```
  arrange(f)

# Cria uma nova coluna com a soma dos pontos de interesse
poi_sum <- poi[, PoI_Sum := rowSums(.SD), .SDcols = c("Amenity", "Bump","Crossing", "Give_Way
                                                      "Stop", "Traffic_Calming",
                                                      "Traffic_Signal", "Turning_Loop")]
```

0.2.3.5 Pré-processando as variáveis do grupo "Duração do acidente".

Primeiramente, foi preciso criar uma nova variável chamada "Duration_minutes" e calcular a duração do acidente em minutos à partir das duas variáveis: Start Date e End Date.

```
acidentes_US_sample <- acidentes_US_sample %>%
  mutate(Duration_minutes = as.numeric(difftime(End_Time, Start_Time, units = "mins")))
```

0.2.4 Técnicas Estatísticas Aplicadas

Como nossas análises foram baseadas em grupos de variáveis que compartilham contextos específicos, as técnicas estatísticas utilizadas variaram de uma para a outra de acordo com os respectivos objetivos.

0.2.4.1 Técnicas de Análise para as Variáveis Temporais

Para as variáveis temporais, foram considerados os diferentes níveis de granularidade das informações, como ano, mês, dia, hora, dia da semana e faixa de horário. A análise foi feita com base numa análise exploratória de dados, estatística descritiva, análise de dados categóricos, visualização de dados aplicada.

Um dos objetivos foi analisar a distribuição das frequências de acidentes entre os diferentes níveis de granularidade de tempo (ano, mês, mês-dia, dia), e verificar se existem padrões sazonais. Para isso, foram utilizados gráficos de barras e de pontos, bem como tabelas de frequência que foram geradas com o auxílio dos pacotes `tidyverse` e `ggplot2`.

```
acidentes_ano_barplot <- acidentes_US_sample_date[Civil_Twilight == "Day" | Civil_Twilight ==
                                                  .(.N,
                                                    Distinct_date = uniqueN(Date),
                                                    Media_diaria = round(.N/uniqueN(Date),0)),
                                                  by = Year][
                                                    order(Year)][
                                                      , Percent_variation := round((Media_diaria -
  setorder(Year) %>% # order data set increasingly by year
```

```
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Year, y = N,
      text = paste("Ano:", Year,
                   "<br>Qtd. Acidentes:", round(N/1000, 1), "k",
                   "<br>Disa Obs. no Ano:", Distinct_date,
                   "<br>Acidentes por Dia:", round(Media_diaria/1000, 1), "k")) +
  geom_bar(stat = "identity", fill = "darkred") +
  geom_text(aes(label = percent(Percent_variation)), size = 3) +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  labs(title = "Qtd. de Acidentes por Ano", x = "Ano", y = "Qtd. Acidentes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  plotly()
acidentes_ano_barplot %>% ggplotly(tooltip = "text")
```

Juntamente, para acrescentar mais informações à análise, foi feita uma sumarização dessas distribuições de frequências de acidentes por crepúsculo civil, a fim de verificar se existe uma relação entre a quantidade de acidentes e a luminosidade do dia.

```
acidentes_ano_barplot_CT <- acidentes_US_sample_date[Civil_Twilight == "Day" | Civil_Twilight
  , N_year := sum(N), by = Year][
    , Percent := round((N/N_year), 2)] %>%
  setorder(Year, Civil_Twilight) %>% # order data set increasingly by year
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Year, y = N,
      text = paste("Ano:", Year,
                   "<br>Qtd. Acidentes:", round(N/1000, 1), "k",
                   "<br>% Total Acidentes:", round(100*Percent, 0), "%")) +
  geom_bar(aes(fill = Civil_Twilight), stat = "identity") +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  scale_fill_manual(values = c('Day' = '#D9A404', 'Night' = '#0F3BBF')) +
  labs(title = "Qtd. de Acidentes por Ano", x = "Ano", y = "Qtd. Acidentes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
acidentes_ano_barplot_CT %>% ggplotly(tooltip = "text") %>%
  layout(
```

```
    legend = list(
      title = list(text = 'Tipo de Dia'),  # Título da legenda
      font = list(size = 14),  # Tamanho da fonte
      itemsizing = 'constant',  # Define o tamanho fixo dos itens da legenda
      traceorder = 'normal',  # Ordem das legendas
      marker = list(
        size = 10
      )
    )
  )
```

Isso foi feito para os seguintes níveis de granularidade de tempo: ano, mês e dia. Para este último, foi traçada uma curva de ajuste com o auxílio da função geom_smooth do pacote ggplot2. Obtivemos resultados interessantes e os discutiremos na próxima seção.

Alternativamente, foi feita uma análise entre as variáveis categóricas, dia da semana e faixa horária, a fim de explorar a relação entre elas. Para isso, foram utilizadas tabelas de contingência bem como matriz de calor, sempre com `ggplot2`. Essa técnica foi aplicada para verificar se existem horários e dias da semana com maior incidência de acidentes ou com variação na severidade dos acidentes.

```
# Cria uma tabela de contingencia de acidentes por dia da semana e faixa horaria
acidentes_US_sample_date[, .N, by = .(Weekday, Time_Range)] %>%
  dcast(Time_Range ~ Weekday, value.var = "N", fill = 0) %>% setorder(-Time_Range)
```

```
# Cria uma matriz de calor para as variaveis dias da semana e faixa horaria
acidentes_data_matriz_AWKTR <- acidentes_US_sample_date[Civil_Twilight == "Day" | Civil_Twili
  setorder(Weekday, -Time_Range) %>% # order data set increasingly by weekday and time range
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Weekday, y = Time_Range, fill = N) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "darkred") +
  labs(title = "Agenda Diaria de Acidentes", x = "Dia da Semana", y = "Horario") +
  theme_minimal() +
  scale_x_discrete(position = "top") +
  theme(
    axis.title.x.top = element_text(),
    axis.text.x.top = element_text(),
    axis.line.x.top = element_line(),
    axis.ticks.x.top = element_line()

```

```
  )
acidentes_data_matriz_AWKTR
```

Ademais, a fim de investigar a severidade média dos acidentes ao longo do tempo, foi realizada uma análise de tendência da média de severidade dos acidentes por ano. Para isso, foi feito um gráfico de pontos associado a um ajuste de curva (geom_smooth).

De modo geral, para cada análise foi feita uma sumarização de dados, seguida de uma visualização dos resultados obtidos. Os códigos acima são apenas exemplos e, na seção de resultados e discussão, mais detalhes serão mostrados.

0.2.4.2 Técnicas de Análise para as Variáveis de Localização

```
leaflet() %>%
      addTiles() %>%
      addPolygons(
        data = acidentes_p_estado,
        fillColor = ~palestado(n),
        weight = 2,
        color = "black",
        fillOpacity = 0.4,
        label = ~NAME,
        popup = ~glue("<b>Estado: </b> {NAME}<br><b>Quantidade de acidentes: </b> {n}"),
        highlightOptions = highlightOptions(
          weight = 3,
          color = "#666",
          fillOpacity = 0.6,
          bringToFront = TRUE
        ),
        group = "Nível estadual"
      ) %>%
      addLegend(
        "bottomright",
        pal = palestado,
        values = acidentes_p_estado$n,
        title = "Número de Acidentes",
        opacity = 1,
        group = "Nível estadual"
      ) %>%
      addHeatmap(
        data = mapa_de_calor,
        lng = ~`Start Lng`,
```

```
        lat = ~`Start Lat`,
        intensity = ~n,
        blur = 20,
        radius = 15,
        group = "Mapa de calor"
      ) %>%
      addLayersControl(
        baseGroups = c("Nível estadual", "Mapa de calor"),
        position = "topright"
      )
```

0.2.4.3 Técnicas de Análise para as Variáveis de Condições Climáticas

Para as variáveis de condições climáticas, foram realizadas análises de frequência de acidente por condição climática.

```
condicoes_count <- acidentes_US_sample %>%
  group_by(Weather_Cluster) %>%
  tally(name = "Count")
ggplot(condicoes_count, aes(x = reorder(Weather_Cluster, -Count), y = Count, fill = Weather_C
  geom_bar(stat = "identity") +
  labs(title = "Número de Acidentes por Condição Climática", x = "Condição Climática", y = "N
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Este gráfico mostra que a maioria dos acidentes ocorrem quando o céu está limpo.

Depois, foi analisado a correlação entre as variáveis climáticas.

```
climatic_data <- acidentes_US_sample %>%
  select(`Temperature(F)`, `Humidity(%)`, `Pressure(in)`, `Wind_Speed(mph)`,`Visibility(mi)`
  na.omit()

cor_matrix <- cor(climatic_data)
corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black")
```

Foi analisado que, algumas delas possuem correlação positiva e outras, negativa.

No próximo passo, o objetivo era analisar o número de acidentes por Condição Climática. Para isso, foi preciso agrupar por condição climática, contar o número de acidentes e fazer um gráfico de barras.

```r
# Número de acidentes por condição climática
condicoes_count <- acidentes_US_sample %>%
  group_by(Weather_Cluster) %>%
  tally(name = "Count")
ggplot(condicoes_count, aes(x = reorder(Weather_Cluster, -Count), y = Count, fill = Weather_C
  geom_bar(stat = "identity") +
  labs(title = "Número de Acidentes por Condição Climática", x = "Condição Climática", y = "N
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Depois, foram analisadas as distribuições de precipitação e visibilidade.

```r
ggplot(acidentes_US_sample, aes(x = `Precipitation(in)`)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black") +
  labs(title = "Distribuição de Precipitação", x = "Precipitação (in.)", y = "Contagem")


ggplot(acidentes_US_sample, aes(x = `Visibility(mi)`)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black") +
  labs(title = "Distribuição de Visibilidade", x = "Visibilidade (milhas)", y = "Contagem")
```

Para analisar o número de acidentes por condição de chuva no dia foi feito um gráfico de barras.

```r
accident_chuva <- acidentes_US_sample %>%
  group_by(Choveu) %>%
  summarise(Accident_Count = n())


ggplot(accident_chuva, aes(x = Choveu, y = Accident_Count, fill = Choveu)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Acidentes em Cidades com Chuva vs. Sem Chuva",
    x = "Choveu?",
    y = "Número de Acidentes"
  ) +
  theme_minimal()
```

Aqui podemos analisar que há um número muito maior de acidentes quando não chove do que quando chove.

Por último, foram realizadas análises da relação entre a precipitação(chuva) e a severidade do acidente.

```
  ggplot(acidentes_US_sample, aes(x = `Precipitation(in)`, y = Severity)) +
    geom_jitter(width = 0.2, alpha = 0.5) +
    labs(
      title = "Relação entre Precipitação e Severidade",
      x = "Precipitação (mm ou in)",
      y = "Severidade"
    ) +
    theme_minimal()
```

0.2.4.4 Técnicas de Análise para as Variáveis de Sinalização e Infraestrutura

Para as variáveis de infraestrutura e sinalização, foram realizadas análises de frequência de acidentes com e sem cada ponto de interesse

```
barplot <- acidentes_US_sample %>% select(Amenity:Turning_Loop) %>%
      pivot_longer(everything(), names_to = "poi", values_to = "value") %>%
      ggplot(aes(x = poi, fill = value)) +
      geom_bar() +
      labs(title = "Pontos de Interesse",
           x = NULL,
           y = NULL,
           fill = "Presença de poi",
           legend.position = "bottom") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
      scale_y_continuous(labels = comma) +
      scale_fill_manual(values = c("darkgreen", "darkred"),
                        labels = c("Não", "Sim"))

    ggplotly(barplot)
```

Este gráfico torna evidente que a grande maioria dos acidentes não ocorre próximo a qualquer ponto de interesse, porém é notável que há um aumento de acidentes em locais onde fluxos diferentes se encontram, como cruzamentos e junções.

0.2.4.5 Técnicas de Análise para as Variáveis de Duração do acidente.

Para analisar a duração dos acidentes, foi construído um boxplot para analisar a distribuição de acidentes por severidade e por duração, para verificar se tinha uma relação entre essas duas variáveis.

```
ggplot(acidentes_US_sample, aes(x = as.factor(Severity), y = Duration_minutes)) +
  geom_boxplot(outlier.shape = NA) +
  labs(
    title = "Duração dos Acidentes por Severidade",
    x = "Severidade",
    y = "Duração (minutos)"
  ) +
    scale_y_continuous(limits = quantile(acidentes_US_sample$Duration_minutes, c(0.0, 0.9))
  theme_minimal()
```

Como possuía muitos outliers e isso dificultava a visualização, optamos por retirá-los e calcular os limites a partir da densidade dos dados.

0.3 Resultados e Discussão

Com os dados processados e metologias estatisticas aplicadas, vamos agora de fato analisar os dados e extrair informações relevantes. Far-lo-emos para cada grupo de variáveis, apresentando os resultados obtidos e discutindo-os.

0.3.1 Resultados para as Variáveis Temporais

Começando pelo básico, vamos observar a distribuição de acidentes ao longo dos anos.

```
acidentes_ano_barplot <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight == "Night
                                           .(.N,
                                             Distinct_date = uniqueN(Date),
                                             Media_diaria = round(.N/uniqueN(Date),0)),
                                           by = Year][
                                             order(Year)][
                                               , Percent_variation := round((Media_diaria -
  setorder(Year) %>% # order data set increasingly by year
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Year, y = N,
      text = paste("Ano:", Year,
                   "<br>Qtd. Acidentes:", round(N/1000, 1), "k",
                   "<br>Disa Obs. no Ano:", Distinct_date,
```

```
                    "<br>Acidentes por Dia:", round(Media_diaria/1000, 1), "k")) +
  geom_bar(stat = "identity", fill = "darkred") +
  geom_text(aes(label = percent(Percent_variation)), size = 3) +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  labs(title = "Qtd. de Acidentes por Ano", x = "Ano", y = "Qtd. Acidentes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  plotly()
acidentes_ano_barplot %>% ggplotly(tooltip = "text")
```

Não nos resta dúvida de que houve um aumento significativo na quantidade de acidentes ao longo dos anos, com aumentos percentuais de 56, 24, 7, 23, 33, 16 e −33 por cento entre cada um dos anos de 2016 até 2023. Observa-se, entretanto, que houve apenas 73 dias observados no último ano, o que pode ter influenciado na queda de 33 por cento na quantidade de acidentes, não nos permitindo tirar conclusões definitivas.

Agora, realizando a mesma investigação, porém observando apenas ao nível mês da granularidade do tempo:

```
acidentes_mes_barplot <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight == "Night
                                           .(.N,
                                             Distinct_date = uniqueN(Date),
                                             Media_diaria = round(.N/uniqueN(Date),0)),
                                           by = Month][
                                             order(Month)][
                                               , Percent_variation := round((Media_diaria -
  setorder(Month) %>% # order data set increasingly by month
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Month, y = N,
      text = paste("Mes:", Month,
                   "<br>Qtd. Acidentes:", round(N/1000, 1), "k",
                   "<br>Dias Obs. no Mes:", Distinct_date,
                   "<br>Acidentes por Dia:", round(Media_diaria/1000, 1), "k")) +
  geom_text(aes(label = percent(Percent_variation)), size = 3) +
  geom_bar(stat = "identity", fill = "darkred") +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  labs(title = "Qtd. de Acidentes por Mes", x = "Mes", y = "Qtd. Acidentes") +
```

```
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
acidentes_mes_barplot %>% ggplotly(tooltip = "text")
```

Aqui, observamos que a quantidade de acidentes é maior nos meses de outubro, novembro e dezembro, com uma queda significativa a partir de janeiro até julho. Isso nos permite inferir um certo grau de sazonalidade entre os meses de um ano, com pico sempre no mês mais festivo: dezembro.

Por finalizar as análises de distribuição de frequências, analisamos os acidentes dia a dia:

```
data_timeseries <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight == "Night", .(
  setorder(Date) %>% # order data set increasingly by date
  as.data.frame() %>%  # Transforming data.table into data.frame
  ggplot() +
  aes(x = Date, y = N) +
  geom_bar(stat = "identity", alpha = 0.4, fill = "darkred", alpha = 0.6) +
  geom_point(size = 0.5, alpha = 0.5, color = "darkred",
             aes(text = paste("Data:", Date, "<br>Qtd. Acidentes:", N/1000, "k"))) +
  geom_smooth(method = "gam", color = "darkred", size = 0.5) +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  labs(title = "Qtd. de Acidentes por Data", x = "Data", y = "Qtd. Acidentes") +
  theme_minimal()
data_timeseries %>% ggplotly(tooltip = "text")
```

A partir deste gráfico de pontos e colunas (frequência de acidentes por data), podemos observar tanto a tendência crescente anual bem como a sazonalidade mensal. Além disso, e de forma mais interessante, é nítida a existência visual de uma banda superior e inferior: isto nos indica uma variabilidade muito grande na ocorrência de acidentes em datas muito próximas. Utilizando o zoom do gráfico, podemos investigar mais a fundo essas variabilidades: os dias com menos acidentes ocorrem na sua imensa maioria dois a dois, e os com mais, de cinco em cinco. Ou seja, a variabilidade pode ser explicada pelo fato de que a quantidade de acidentes é muito maior entre os dias da semana do que em finais de semana!

Além disso, neste gráfico fica claro uma queda abrupta dos acidentes durante a pandemia, seguida de um aumento expressivo em quantidade e variabilidade.

Okay, e quanto à faixa horária? Como os acidentes se distribuem? Vamos investigar com a agenda diária de acidentes:

```r
acidentes_data_matriz_AWKTR <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight ==
  setorder(Weekday, -Time_Range) %>% # order data set increasingly by weekday and time range
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Weekday, y = Time_Range, fill = N) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "darkred") +
  labs(title = "Agenda Diaria de Acidentes", x = "Dia da Semana", y = "Horario") +
  theme_minimal() +
  scale_x_discrete(position = "top") +
  theme(
    axis.title.x.top = element_text(),
    axis.text.x.top = element_text(),
    axis.line.x.top = element_line(),
    axis.ticks.x.top = element_line()

  )
acidentes_data_matriz_AWKTR
```

Aqui, podemos observar que a quantidade de acidentes é maior nos horários de pico, entre 7hrs e 9hrs e entre 16hrs e 18hrs, e que a quantidade de acidentes é maior durante a semana do que nos finais de semana. Isso nos permite inferir que a quantidade de acidentes é maior nos horários de pico e durante a semana.

Em uma de minhas intuitivas explorações, decidi investigar a relação entre a quantidade de acidentes e a luminosidade do dia (crepúsculo civil). Para isso, novamente realizei um gráfico de pontos e barras de acidentes por data, porém agora separando pelas categorias do crepúsculo civil:

```r
acidentes_data_timeseries_CT <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight ==
  setorder(Date, Civil_Twilight) %>% # order data set increasingly by date
  as.data.frame() %>%  # Transforming data.table into data.frame
  ggplot() +
  aes(x = Date, y = N,
      text = paste("Data:", Date, "<br>Qtd. Acidentes:", N/1000, "k")) +
  geom_bar(stat = "identity", alpha = 0.4) +
  geom_point(aes(color = Civil_Twilight), alpha = 0.4, size = 0.4) +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufix
  ) +
  scale_color_manual(values = c('Day' = '#D9A404', 'Night' = '#0F3BBF')) +
  labs(title = "Qtd. de Acidentes por Data",
```

```
       x = "Data", y = "Qtd. Acidentes") +
  theme_minimal()
acidentes_data_timeseries_CT %>% ggplotly(tooltip = "text") %>%
  layout(
    legend = list(
      title = list(text = 'Crepusculo Civil'),  # Título da legenda
      font = list(size = 14),  # Tamanho da fonte
      itemsizing = 'constant',  # Define o tamanho fixo dos itens da legenda
      traceorder = 'normal',  # Ordem das legendas
      marker = list(
        size = 10
      )
    )
  )
```

A partir deste gráfico, conseguimos observar todos os resultados já vistos até aqui e mais um pouco: agora conseguimos observar que o crepúsculo civil apresenta uma sazonalidade mensal muito expressiva. Em geral, de outubro a dezembro, os acidentes noturnos atingem o seu pico e os acidentes diurnos os seus vales.

Entretanto, ao mesmo tempo, não podemos afirmar que essa sazonalidade é devida ao crepúsculo civil, devido ao fato de que as estações do ano influenciam significativamente a luminosidade do dia no inverno e verão. Para contornar essa dúvida, investigamos e colocamos o mesmo gráfico, entretanto, ao invés de categorizar pelo crepúsculo civil, categorizamos por faixa horária do dia: acidentes que aconteceram das 06:00 às 18:00 e das 18:00 às 06:00.

```
acidentes_data_timeseries_6_18 <- acidentes_US_date[Day_Period == "06:00:00 -> 18:00:00" | Da
  setorder(Date, Day_Period) %>% # order data set increasingly by date
  as.data.frame() %>%  # Transforming data.table into data.frame
  ggplot() +
  aes(x = Date, y = N,
      text = paste("Data:", Date, "<br>Qtd. Acidentes:", N/1000, "k")) +
  geom_bar(stat = "identity", alpha = 0.4) +
  geom_point(aes(color = Day_Period), alpha = 0.4, size = 0.4) +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  scale_color_manual(values = c('06:00:00 -> 18:00:00' = '#D9A404', '18:00:00 -> 06:00:00' =
  labs(title = "Qtd. de Acidentes por Data", x = "Data", y = "Qtd. Acidentes") +
  theme_minimal()
acidentes_data_timeseries_6_18 %>% ggplotly(tooltip = "text") %>%
  layout(
```

```
    legend = list(
      title = list(text = 'Periodo do Dia'),  # Título da legenda
      font = list(size = 14),  # Tamanho da fonte
      itemsizing = 'constant',  # Define o tamanho fixo dos itens da legenda
      traceorder = 'normal',  # Ordem das legendas
      marker = list(
        size = 10
      )
    )
  )
```

A partir deste gráfico, conseguimos observar que a expressividade da sazonalidade mensal nao existe e, na verdade, é sutil em ambas as faixas horárias. Isso nos permite inferir que a sazonalidade observada no gráfico anterior é devida à luminosidade do dia e não às estações do ano. Entretanto, conseguimos perceber que diferença da quantidade de acidentes entre dias da semana e finais de semana permanecem em ambas as faixas horárias.

E como se comporta a severidade dos acidentes ao longo do tempo? Vamos investigar a tendência da média de severidade dos acidentes por data:

```
severity_data_points <- acidentes_US_date[(Civil_Twilight == "Day" | Civil_Twilight == "Night
                                           .(Severidade_Media = mean(Severity)),
                                           by = .(Date)][order(Date)] %>%
  ggplot() +
  aes(x = Date, y = Severidade_Media) +
  geom_point(size = 0.5, alpha = 0.1, color = "darkred",
             aes(text = paste("Data:", Date,
                              "<br>Severidade Média:", round(Severidade_Media, 1)))) +
  geom_smooth(method = "lm", size = 0.5, se = FALSE, color = "darkred") +
  ylim(0, 4) +
  labs(title = "Severidade Média por Data", x = "Data", y = "Severidade") +
  theme_minimal()
severity_data_points %>% ggplotly(tooltip = "text")
```

Essa tendência linear decrescente pode ser explicada pela variação da quantidade de acidentes ao longo do tempo para cada uma das severidades. Observemo-las:

```
acidentes_ano_barplot <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight == "Night
                                           .(.N,
                                             Distinct_date = uniqueN(Date),
                                             Media_diaria = round(.N/uniqueN(Date),0)),
                                           by = .(Year, Severity)][
```

```
                                                order(Year,Severity)] %>%
  setorder(Year) %>% # order data set increasingly by year
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Year, y = N,
      text = paste("Ano:", Year,
                   "<br>Qtd. Acidentes:", round(N/1000, 1), "k",
                   "<br>Disa Obs. no Ano:", Distinct_date,
                   "<br>Acidentes por Dia:", round(Media_diaria/1000, 1), "k")) +
  geom_bar(stat = "identity", fill = "darkred") +
  scale_y_continuous(
    labels = label_number(scale = 1/1000, suffix = "k") # Divide por 1000 e adiciona o sufixo
  ) +
  labs(title = "Qtd. de Acidentes por Ano", x = "Ano", y = "Qtd. Acidentes") +
  facet_wrap(~Severity, scales = "free_y") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  plotly()
acidentes_ano_barplot %>% ggplotly(tooltip = "text")
```

Aqui, podemos observar que a quantidade de acidentes é maior para a severidade 2, seguida pela 3, 4 e 1, respectivamente. Além disso, a quantidade de acidentes para a severidade 1 e 2 são as que mais crescem, ao passo que a quantidade de acidentes para a severidade 3 é a que mais decresce. Os acidentes de severidade 4, por outro lado, apresentam uma tendência de crescimento ao longo dos anos proporcionalmente menor. A explicação das causas da diminuição da severidade média não está no escopo desta análise, entretanto, podemos destacar que algumas hipóteses podem ser levantadas, como a melhoria da infraestrutura viária, a conscientização dos motoristas, a melhoria dos sistemas de segurança dos veículos, alteração de legislações de trânsito, entre outros.

Para fechar as análises de variáveis temporais, investigamos o padrao da média de severidade dos acidentes entre as faixas horárias e dias da semana, bem como ao longo dos anos:

```
acidentes_data_matriz_SWKTR <- acidentes_US_date[Civil_Twilight == "Day" | Civil_Twilight ==
  setorder(Weekday, -Time_Range) %>% # order data set increasingly by weekday and time range
  as.data.frame() %>% # Transforming data.table into data.frame
  ggplot() +
  aes(x = Weekday, y = Time_Range, fill = Severity) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "darkred") +
  labs(title = "Agenda Diaria da Severidade Média de Acidentes", x = "Dia da Semana", y = "H
  theme_minimal() +
```

```
  scale_x_discrete(position = "top") +
  theme(
    axis.title.x.top = element_text(),
    axis.text.x.top = element_text(),
    axis.line.x.top = element_line(),
    axis.ticks.x.top = element_line()

  )
acidentes_data_matriz_SWKTR
```

Destes, conseguimos observar que, mesmo que os acidentes ocorram menos durante os finais de semana, são eles os mais severos. Além disso, a severidade média dos acidentes tem diminuído ao longo dos anos!

0.3.2 Resultados para as Variáveis de Localização

A maioria dos acidentes ocorre em cidades maiores e mais populosas, bem como em estados com maior densidade demográfica. Isso pode ser atribuído ao maior volume de tráfego, maior número de veículos em circulação, e maior complexidade das vias urbanas nessas regiões. Além disso, áreas urbanas costumam apresentar condições mais propensas a acidentes, como interseções movimentadas, trânsito intenso e uma maior frequência de eventos climáticos adversos que afetam a visibilidade e a dirigibilidade. Fatores socioeconômicos também podem desempenhar um papel, já que cidades maiores tendem a concentrar atividades econômicas e culturais, atraindo um número significativo de pessoas e aumentando o risco de incidentes

```
leaflet() %>%
      addTiles() %>%
      addPolygons(
        data = acidentes_p_estado,
        fillColor = ~palestado(n),
        weight = 2,
        color = "black",
        fillOpacity = 0.4,
        label = ~NAME,
        popup = ~glue("<b>Estado: </b> {NAME}<br><b>Quantidade de acidentes: </b> {n}"),
        highlightOptions = highlightOptions(
          weight = 3,
          color = "#666",
          fillOpacity = 0.6,
          bringToFront = TRUE
        ),
```

```
        group = "Nível estadual"
      ) %>%
      addLegend(
        "bottomright",
        pal = palestado,
        values = acidentes_p_estado$n,
        title = "Número de Acidentes",
        opacity = 1,
        group = "Nível estadual"
      ) %>%
      addHeatmap(
        data = mapa_de_calor,
        lng = ~`Start Lng`,
        lat = ~`Start Lat`,
        intensity = ~n,
        blur = 20,
        radius = 15,
        group = "Mapa de calor"
      ) %>%
      addLayersControl(
        baseGroups = c("Nível estadual", "Mapa de calor"),
        position = "topright"
      )
```

0.3.3 Resultados para as Variáveis de Condições Climáticas

A análise das condições climáticas revelou que a maioria dos acidentes ocorreu em dias com céu limpo. A correlação entre variáveis climáticas, como temperatura, umidade e pressão, mostrou interações tanto positivas quanto negativas entre elas. A frequência de acidentes variou conforme as condições climáticas, com maior incidência em dias claros.

A distribuição de precipitação indicou que a maioria dos acidentes ocorreu sem chuva, e a visibilidade também foi um fator importante. Quando analisada a relação entre precipitação e severidade do acidente, não foi observada uma correlação clara, sugerindo que outros fatores influenciam a gravidade dos acidentes.

0.3.4 Resultados para as Variáveis de Sinalização e Infraestrutura

```
  poi_Severity <- acidentes_US_sample[,c(3,30:42)]

```

```r
    poi_sum <- poi_Severity[, Total := rowSums(.SD),
                            .SDcols = c("Amenity", "Bump", "Crossing", "Give_Way", "Junction"
                                        "No_Exit", "Railway", "Roundabout", "Station", "Stop"
                                        "Traffic_Calming","Traffic_Signal", "Turning_Loop")]

    severityPoi_sum_table <- poi_sum[, .(Severity, Total)] %>% table
    severityPoi_sum_df <- as.data.frame(severityPoi_sum_table)

    severity_sumPoI <- severityPoi_sum_df %>% summarise(.by = Severity, sum_PoI = sum(Freq))
    severityPoi_sum_df <- severityPoi_sum_df %>%
      left_join(severity_sumPoI) %>% mutate(Rel_Freq = Freq/sum_PoI)


   absolute_severity_bars <-  severityPoi_sum_df %>% as.data.frame %>%
      ggplot(aes(x = Severity, y=Freq, fill = Total)) +
      geom_bar(stat = "identity", position = "dodge") +
      scale_y_continuous(labels = comma) +
      labs(title = "Severidade de acidentes por ponto de interesse",
           x = NULL,
           y = NULL) +
      theme_minimal()
   absolute_severity_bars
```

O gráfico de frequência absoluta da severidade por soma de pontos de interesse nos mostra que a maior quantidade de pontos ao redor de um acidente é 7, e também que a maioria dos acidentes ocorre em locais sem pontos de interesse, e ocorrem com frequência reduzida quanto mais pontos de interesse existem ao redor, porém pode apenas indicar que existem poucos lugares com muitos pontos de interesse. Também é que evidente que acidentes com grau 2 de severidade são extremamente mais comuns que os outros.

0.3.5 Resultados para as Variáveis de Duração do acidente

Os dados possuem muitos outliers e acidentes que duraram até 2 milhões de minutos, o que dificultou a nossa análise. Porém, de acordo com o gráfico mostrado nas técnicas estatísticas analisadas, podemos observar que os acidentes de severidade 2 e 4 têm uma variabilidade muito maior em questão do tempo. Podemos observar também, que a maior parte dos acidentes de severidade 1 e 3, costumam demorar entre 0 e 100 minutos.

Referências

Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). *A Countrywide Traffic Accident Dataset*. https://arxiv.org/abs/1906.05409

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 33–42. https://doi.org/10.1145/3347146.3359078