Gabriel de Mello Cambuy Ferreira RA: 142641
Igor Vinícius Santos Vieira - RA: 141004
Nicole Cristine de Faria Santos - RA: 156636

Análise de acidentes na rodovia Presidente Dutra

São José dos Campos - Brasil
Julho de 2022

Gabriel de Mello Cambuy Ferreira RA: 142641
Igor Vinícius Santos Vieira - RA: 141004
Nicole Cristine de Faria Santos - RA: 156636

Análise de acidentes na rodovia Presidente Dutra

Relatório apresentado à Universidade Federal de São Paulo como parte dos requisitos para aprovação na unidade curricular de Probabilidade e Estatística do primeiro semestre de 2022.

Docente: Prof. Dra. Luzia Pedroso de Oliveira

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia - Campus São José dos Campos

São José dos Campos - Brasil
Julho de 2022

Resumo

O relatório descreve as ocorrências acidentais da BR-116/SP e BR-116/RJ e suas características. Para tal finalidade, a metodologia consiste na obtenção de dados abertos, elaborada pela ANTT (Agência Nacional de Transportes Terrestres), sobre os acidentes ocorridos na rodovia presidente Dutra. A partir disso, através do uso da linguagem de programação R, uma análise de dados foi feita para avaliar a relação entre as variáveis qualitativas e quantitativas e auxiliar na compreensão situacional da estrada. Nesse sentido, a análise dos resultados indica qual o trecho com a maior quantidade de acidentes, estabelece uma relação entre os veículos e o total de pessoas envolvidas e identifica uma redução gradual de ocorrências ao longo dos anos, de modo a auxiliar na compreensão da rodovia e suas perspectivas futuras.

Palavras-chaves: Estatística. Análise de Dados, R, Acidentes, Rodovia Dutra.

Lista de ilustrações

Sumário

1. Introdução

Os acidentes de trânsito, em geral, ocasionam vários impactos sociais em um país. Diante dos dados divulgados pelo Instituto de Pesquisa Econômica Aplicada (IPEA), o Brasil ocupa a quinta posição entre os países com maiores taxas de vítimas de trânsito. Uma das justificativas para esse fato tem relação com os registros diários de 14 mortes e 190 acidentes nas rodovias federais brasileiras.

Nesse viés, de acordo com um painel divulgado pela Confederação Nacional do Transporte (CNT) [3], a BR-116 destaca-se entre as rodovias com mais mortes em acidentes. Por esse motivo, os trechos das BR-116/SP e BR-116/RJ foram escolhidos para o estudo, tendo em vista que são os trechos de maior movimentação e pertencentes à rodovia Presidente Dutra.

Com base no contexto descrito, o seguinte relatório, tem como objetivo revelar a situação da BR-116/SP e BR-116/RJ, através da análise do número de acidentes e suas singularidades, como a quantia e tipos de veículos envolvidos, trechos que mais ocorrem essas situações, tipo do acidente e o nível das fatalidades.

Para atingir esse propósito, foi necessária a utilização dos dados divulgados pelo Portal Brasileiro de Dados Abertos [1], que estabelecem um demonstrativo acerca da situação das rodovias brasileiras e as características de cada acidente, no período entre 2010 até 2021.

Com isso, será possível compreender melhor a situação da estrada e estabelecer possíveis pontos de melhorias para que menos tribulações ocorram, de modo a alcançar os objetivos da análise.

2. Objetivos

2.1 Geral

Analisar estatisticamente o conjunto de dados com todas as ocorrências registradas na Rodovia Presidente Dutra visando encontrar informações como: em qual trecho mais acidentes ocorrem, em qual período do dia mais acidentes acontecem, em qual período do dia mais acontecem acidentes fatais.

2.2 Específico

Aplicar os conceitos teóricos de Estatística e práticos da utilização do *software RStudio* que foram vistos na Unidade Curricular de Probabilidade e Estatística para analisar as diferentes variáveis presentes na base de dados e relacioná-las, de modo a visualizar as ocorrências registradas na rodovia Presidente Dutra.

3. Metodologia

A coleta de dados foi feita através do Portal Brasileiro de Dados Abertos [1], o qual forneceu um arquivo csv com as mais variadas informações referentes aos ocorridos na rodovia presidente Dutra (BR-116/SP e BR-116/RJ). Para filtrar e interpretar melhor essas informações o software RStudio foi utilizado, assim gráficos poderiam ser analisados e tabelas organizadas de acordo com os objetivos estabelecidos.

3.1 Tratamento de Dados

O conjunto de dados original possui 23 colunas contendo as seguintes informações: ‘data’, ‘horário’, ‘número da ocorrência’, ‘tipo de ocorrência’, ‘km’, ‘trecho’, ‘sentido’, ‘tipo de acidente’, ‘automóvel’, ‘bicicleta’, ‘caminhão’, ‘moto’, ‘ônibus’, ‘outros’, ‘tração animal’, ‘transporte de cargas especiais’, ‘trator e máquinas’, ‘utilitários’, ‘ilesos’, ‘levemente feridos’, ‘moderadamente feridos’, ‘gravemente feridos’ e ‘mortos’.

Visando que os dados fossem analisados corretamente, foi feito um tratamento na planilha para excluir as informações inconsistentes e adequar os dados a certos padrões. A planilha original possui 115810 ocorrências reportadas dentro do período de 01/01/2010 até 29/04/2022. Por conta das informações referentes ao ano de 2022 ainda estarem incompletas a análise se restringe ao período de 01/01/2010 até 31/12/2021. Dentro das ocorrências reportadas foi encontrada uma linha com dados referente a um acidente na BR-393/RJ que originalmente não possui relação com o conjunto de dados da rodovia Presidente Dutra e portanto foi necessário remover essa linha. Assim o conjunto considerando reporta um total de 115081 acidentes.

Além de disponibilizar os dados, o Portal Brasileiro de Dados Abertos também disponibiliza um arquivo chamado Dicionário de Dados [4] com uma breve descrição das informações que estão contidas na planilha. Através do Dicionário de

Dados podemos ver que a coluna ‘tipo de ocorrência' informa se houve ou não pelo menos uma vítima fatal no acidente reportado, podemos observar que ao longo dos anos foram utilizados diversos padrões para preencher esse campo variando a escrita e a acentuação como: ‘sem vitima’, ‘sem vítima’, ‘acidente sem vitima’. Para viabilizar as análises utilizando essa coluna foi necessário padronizar a escrita sem perder ou modificar a informação original e para isso foram adotados os padrões ‘sem vitima’ e ‘com vitima’ para descrever as duas possibilidades.

A partir da coluna ‘horário’ que informa o momento em que a ocorrência foi registrada, foi criada uma nova coluna no conjunto de dados que foi chamada de ‘período’. Dentro da coluna ‘período’ foram consideradas 4 possibilidades que incluem todos os horários possíveis para o registro dentro de um dia que são: Madrugada (00h - 05h), Manhã (06h - 11h), Tarde (12h - 17h) e Noite (18h - 23h).

Também foi criada uma nova coluna que determina a quantidade de pessoas envolvidas em uma ocorrência chamada de ‘quantidade de pessoas’ que recebe a soma das seguintes colunas: ‘ilesos’, ‘levemente feridos’, ‘moderadamente feridos’, ‘gravemente feridos’ e ‘mortos’.

4. Resultados e Discussões

4.1 - Análise de uma variável qualitativa

Para analisar a variável quantitativa correspondente ao total de acidentes, foi elaborado um gráfico de setores para segmentar as ocorrências acidentais de acordo com o trecho correspondente, conforme o gráfico abaixo:

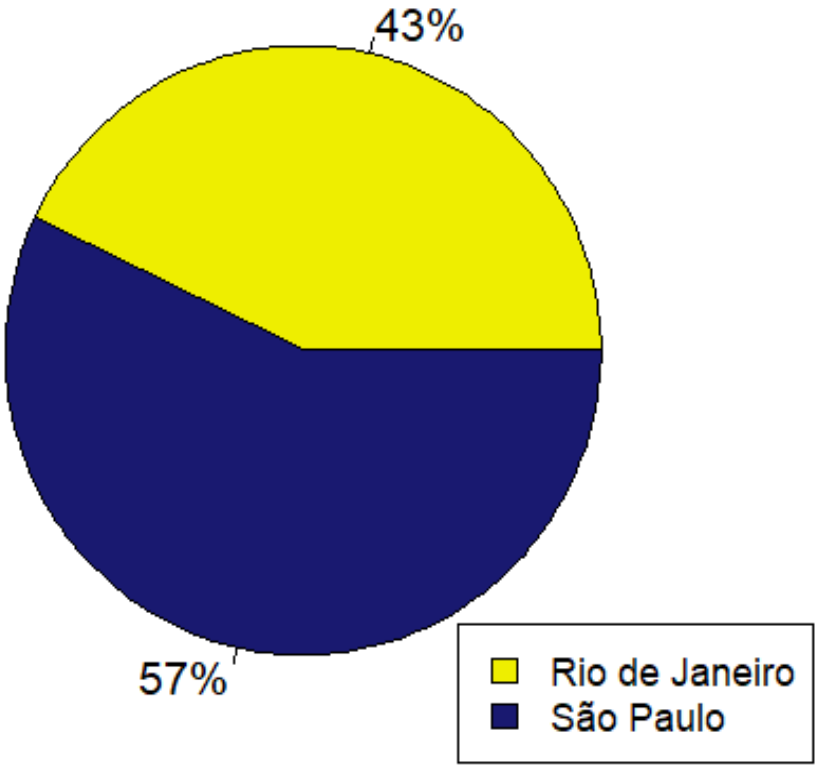

Figura 1 - Gráfico de setores com a porcentagem de acidentes por trecho

Tendo em vista o gráfico da figura 1 é possível perceber a maior quantidade de acidentes localizada no trecho do estado de SP, um dos motivos pode ser devido a extensão do trecho 230 km em relação ao do RJ 170 km [6], dessa forma por ser

maior em comprimento, a parte paulista acaba se destacando no número de acidentes.

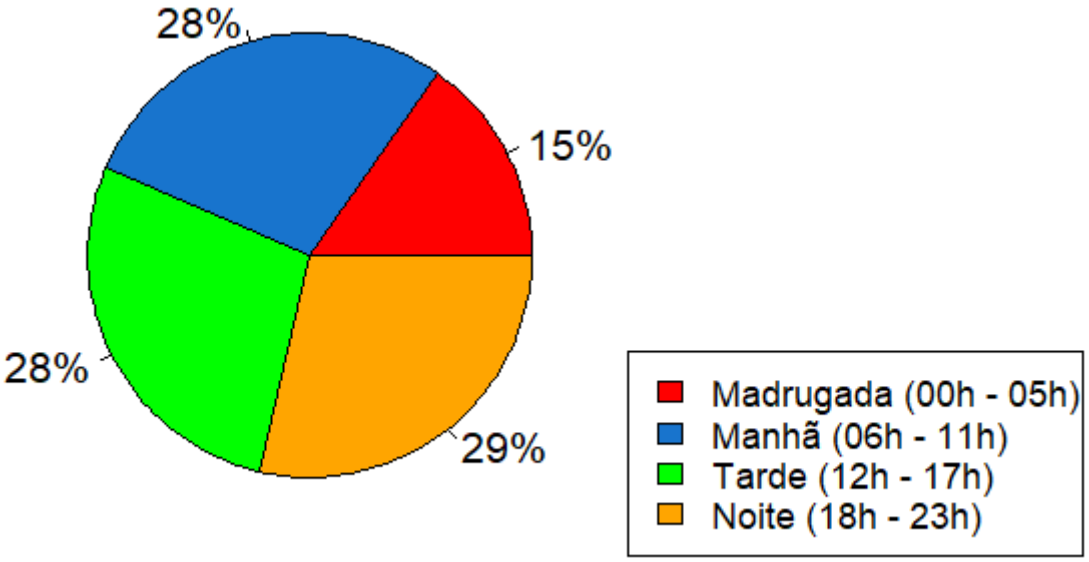

Figura 2 - Gráfico de setores com a porcentagem de ocorrências por período do dia

É bem perceptível no gráfico de setores uma distribuição mais equivalente nos períodos da manhã, tarde e noite. Por outro lado, nos horários da madrugada são os momentos em que menos acidentes ocorrem, ocupando apenas 15% da parcela do gráfico.

4.2 - Análise de uma variável quantitativa

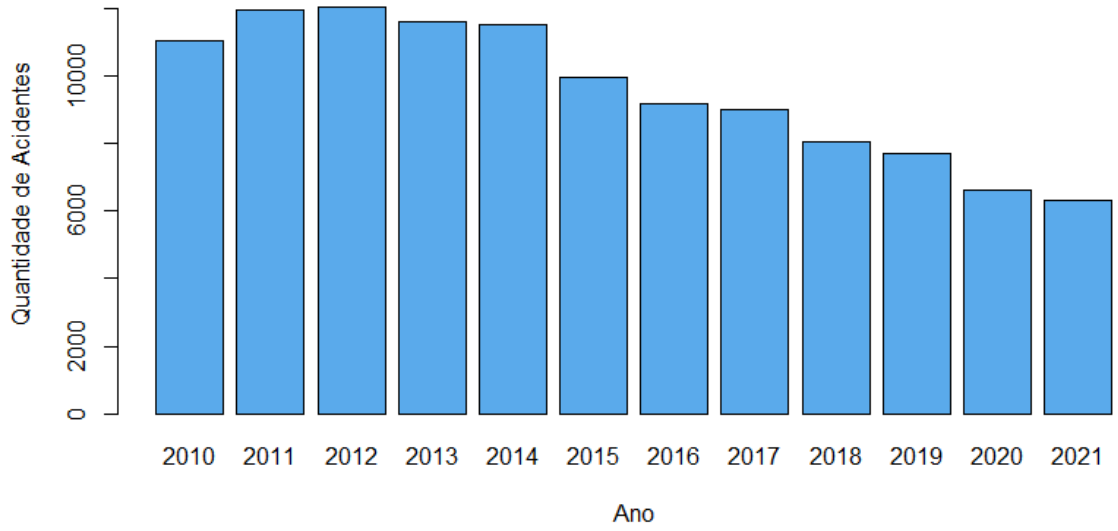

Figura 3 - Gráfico de barras relacionando o ano e a quantidade de ocorrências.

Na figura 2, através do gráfico de barras, é bem perceptível a queda na quantidade de acidentes a partir de 2015, além da diminuição gradativa nos anos seguintes, tal queda pode se justificar devido a algumas melhorias que ocorreram ao longo do tempo, como por exemplo a implantação de emissoras de rádio [2], o qual transmitem constantemente a situação da rodovia aos motoristas, podendo melhorar as condições do tráfego. Um exemplo é a quantidade de acidentes em 2012 em relação ao ano de 2021, nota-se que o número se reduziu quase pela metade.

4.3 - Análise de duas variáveis qualitativas

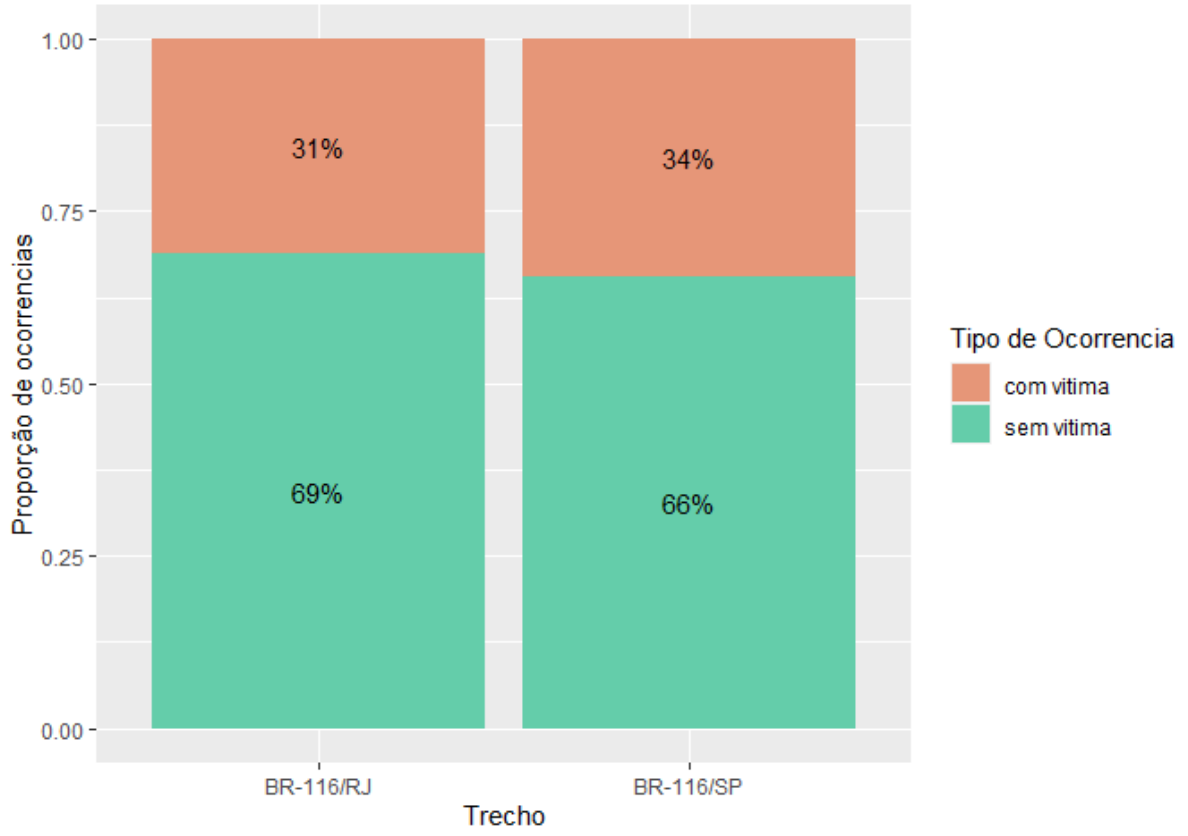

Figura 4 - Gráfico de barras entre duas variáveis qualitativas (tipo de ocorrência e trecho)

Podemos ver na figura 3 que há uma proximidade na proporção de ocorrências com ou sem vítimas em ambos os trechos da Rodovia Presidente Dutra. O trecho de São Paulo possui uma porcentagem um pouco maior de acidentes com pelo menos uma vítima fatal quando comparado ao trecho do Rio de Janeiro.

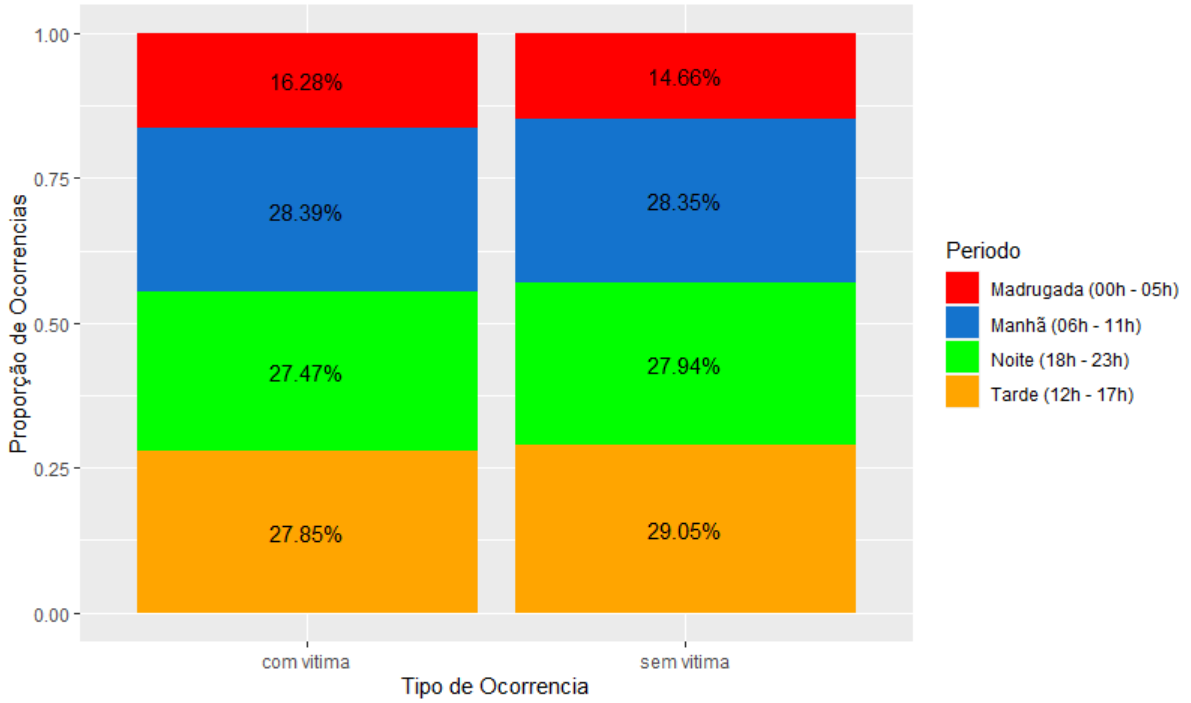

Figura 5 - Gráfico de barras relacionando período e tipo de ocorrência

Observando o gráfico de barras elaborado podemos ver que durante os períodos da Manhã (06h - 11h) e da Noite (18h - 23h) as porcentagens de acidentes com vítimas fatais e sem vítimas fatais são bem próximas enquanto no período da Tarde (12h - 17h) há uma quantia consideravelmente maior de acidentes sem vítima fatal o que pode se dar ao trânsito lento em horários de pico e durante o período da Madrugada (00h - 05h) há uma quantidade consideravelmente maior de acidentes com vítimas fatais o que pode se dar aos frequentes acidentes envolvendo o consumo de álcool e de motoristas de caminhão que acabam caindo no sono enquanto dirigem [7].

4.4 - Análise de uma variável quantitativa

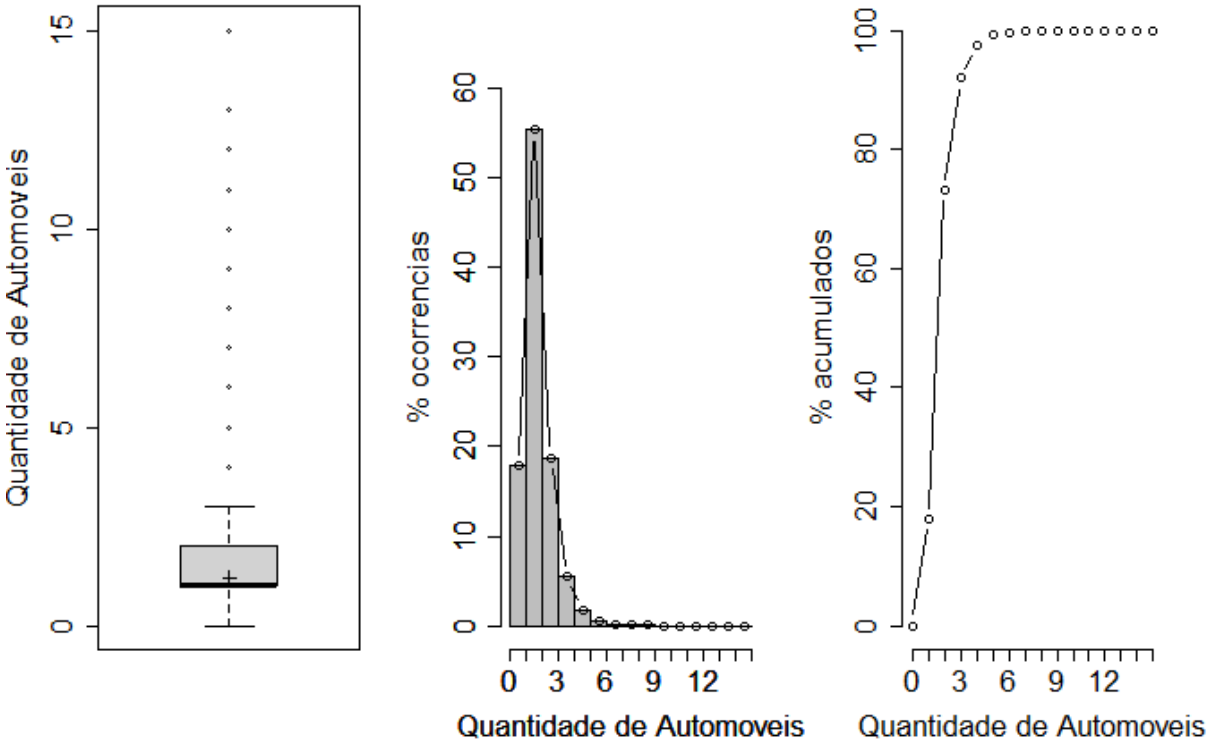

Figura 6 - Diagrama de caixa, histograma com polígono de frequência e gráfico de frequências acumuladas respectivamente, que relacionam os acidentes com a quantidade de automóveis envolvidos.

No gráfico do boxplot é interessante de se notar vários outliers que superam os valores máximos estabelecidos pelo gráfico, ou seja, alguns acidentes ocorrem com uma quantidade maior de veículos que superam o número de 3, porém, como observado no intervalo do primeiro e terceiro quartil, 50% dos ocorridos estão entre 1 e 2 veículos envolvidos, enquanto a mediana se encontra no segundo quartil, sendo o valor pouco maior que 1. O histograma de frequência exibe de forma mais detalhada a porcentagem de acidentes que ocorrem de acordo com a quantidade de automóveis, enquanto o gráfico de frequências acumuladas aponta que cerca de 90% dos acidentes acontecem com 3 veículos ou menos.

```
> mean(quantidadeCarros) #média
[1] 1.114111
> median(quantidadeCarros) #mediana
[1] 1
> sd(quantidadeCarros) #standard deviation = desvio padrão
[1] 0.954161
> (sd(quantidadeCarros)/mean(quantidadeCarros))*100 # coeficiente de variação
[1] 85.64326
```

Figura 7 - Média, Mediana, Desvio Padrão e Coeficiente de Variação da variável quantitativa

Calculando diretamente os valores obtemos que a média da quantidade de carros envolvidos nas ocorrências é de: 1.114111. A mediana obtida após ordenar o vetor quantidade de carros foi de: 1. O desvio padrão obtido foi de 0.954161. O coeficiente de variação foi de 85% o que indica que há uma grande variação na quantidade de carros envolvidos que chegam até 15 carros envolvidos em uma mesma ocorrência.

4.5 - Análise de duas variáveis quantitativas

Através da observação do gráfico de dispersão abaixo, é perceptível que há uma correlação linear positiva entre as variáveis quantitativas. Dessa maneira, nota-se que quando a quantidade de ônibus envolvidos aumenta, a tendência será que o mesmo ocorra com número de pessoas envolvidas.

Ainda que seja notório essa relação, também se observa a presença de outliers, em que, de acordo com o conjunto de dados, uma ocorrência envolvendo pelo menos um ônibus tem o envolvimento de cerca de 500 pessoas.

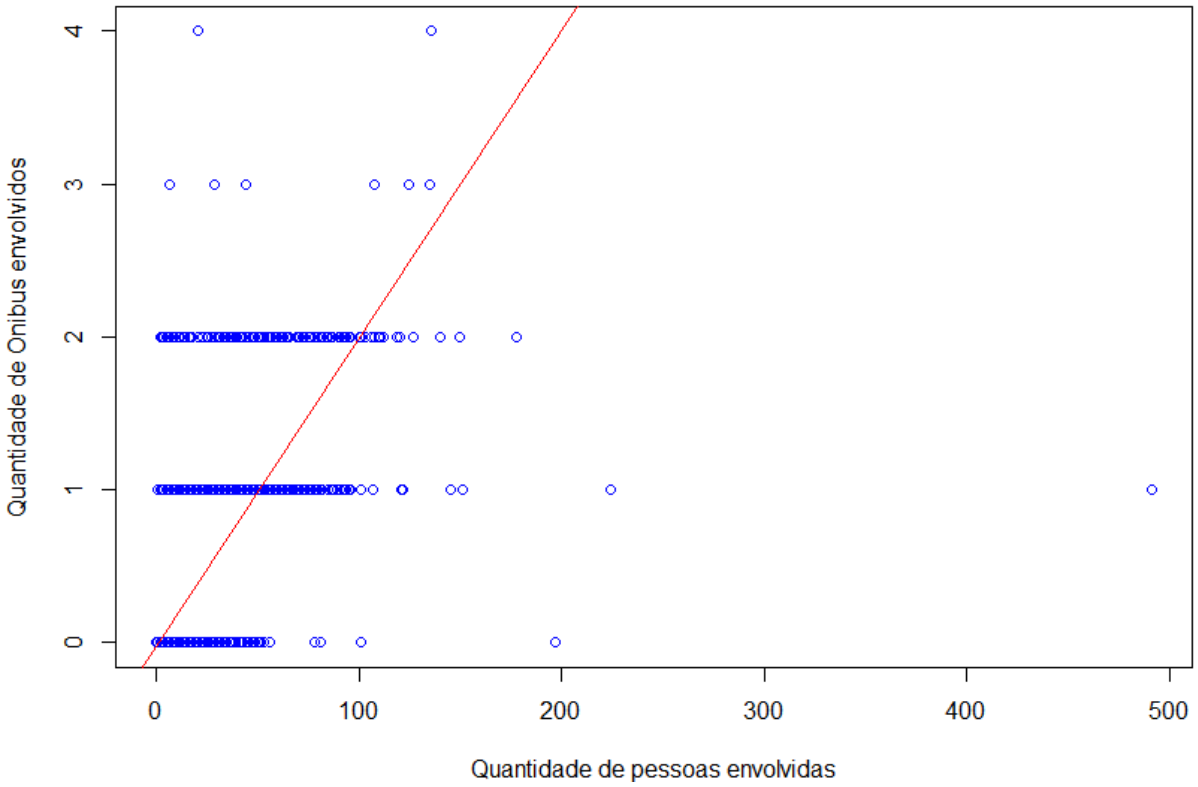

Figura 8 - Gráfico de correlação analisando o número de ônibus e pessoas envolvidas em ocorrências.

Dessa maneira, ao calcular os coeficientes de correlação, foi obtido um valor de, aproximadamente, 0,6429 para o coeficiente de Pearson e 0,2412 para o de Spearman.

Além disso, como o valor do coeficiente de Pearson foi superior ao do coeficiente de Spearman, é possível afirmar que uma das causas para esse fenômeno é a presença de outliers. A representação desses valores está expressa na imagem abaixo:

```
> pearson <- cor(dadosRodovia$onibus, dadosRodovia$qtde_pessoas); pearson
[1] 0.6429018
> spearman <- cor(dadosRodovia$onibus, dadosRodovia$qtde_pessoas, method="spearman"); spearman
[1] 0.2412087
```

Figura 9 - Coeficiente de Pearson e Spearman

4.6 - Análise da coluna ‘tipo de acidente’

A coluna tipo de acidente contém uma breve descrição que classifica o tipo de acidente que foi reportado conforme a imagem abaixo.

H
tipo_de_acidente
Engavetamento
Colisão Traseira
Queda de ribanceira
Colisão Traseira
Colisão Traseira
Queda de moto

Figura 10 - Trecho da coluna tipo de acidente

Observando o conjunto de dados podemos observar que diversos tipos de acidentes diferentes são reportados, o que dificulta a utilização dos dados em gráficos tradicionais. Para a análise dessa coluna foi elaborada uma nuvem de palavras.

Figura 11 - Nuvem de Palavras da coluna tipo de acidente

A nuvem de palavras considera a frequência em que uma palavra aparece no conjunto de dados para determinar o tamanho que ela será representada na nuvem. Podemos observar que choque e colisão são palavras que se destacam. Colisão se refere a um acidente em que ambos os veículos estão em movimento e Choque quando um dos objetos envolvidos está parado podendo ser um veículo ou não e outro está em movimento [5]. Outras palavras que se destacam são traseira e abalroamento longitudinal que descrevem a posição dos veículos envolvidos no acidente no momento do contato. Outro conjunto de palavras que se destaca é Barreira New Jersey que são barreiras de concreto colocadas para evitar que veículos atravessem para o outro lado da pista e haja uma colisão frontal.

4.7 - Análise conjunta de uma variável qualitativa e uma quantitativa

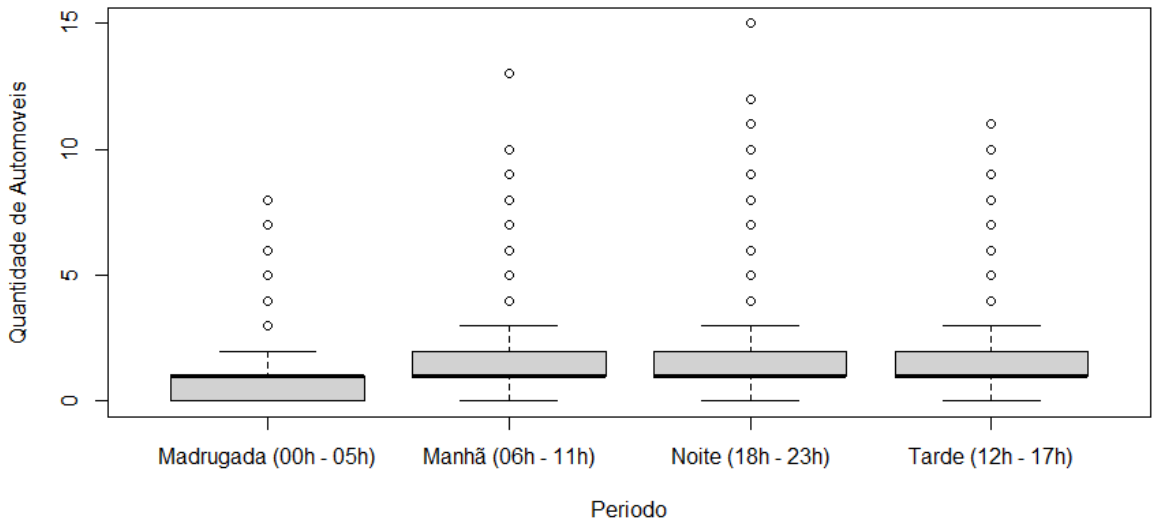

Figura 12 - Análise da quantidade de veículos envolvidos em acidentes e o período do dia em que ocorrem.

Podemos observar que durante os períodos de Manhã, Tarde e Noite são registrados , em média, um maior número de automóveis envolvidos nas

ocorrências do que em comparação com o período da Madrugada. Todos os gráficos plotados possuem outliers que se destacam a partir do envolvimento de 4 ou mais automóveis e podendo chegar até 15 automóveis.

5. Conclusão

Após a obtenção de todos os dados, e a organização dos mesmos em gráficos através do Software RStudio, foi possível compreender muitos fatores sobre a rodovia, principalmente ao separar as variáveis e analisá-las isoladamente e em conjunto. Dentre elas, percebeu-se que o trecho da BR-116/SP possui um número maior de acidentes em relação ao trecho do RJ e também uma porcentagem pouco maior de ocorrências fatais, sendo tais incidentes acontecendo com maior frequência através de choques e colisões, nas posições longitudinal e traseira, sendo a 85% deles acontecendo nos períodos da manhã, tarde e noite, enquanto a parcela da madrugada ocupa apenas 15%. Por outro lado, ao longo dos anos (2010-2021), a quantidade de incidentes se reduziu praticamente pela metade, ao se considerar ambos os trechos (SP e RJ), com cerca de 90% do total, acontecendo com 3 veículos ou menos envolvidos. Outro ponto que se destacou foi em relação a um outlier dentre os dados analisados, no qual um ônibus foi capaz de envolver 500 pessoas num único acidente, sendo esses veículos os responsáveis pelas ocorrências com maior número de envolvidos. Tendo em vista todos esses aspectos, é possível concluir que os objetivos do relatório foram atingidos, ao se compreender todos os pontos estabelecidos previamente.

De acordo com o Painel CNT de Acidentes Rodoviários [3] o número de acidentes em rodovias federais vem caindo no Brasil desde 2014 e o tipo de acidente mais comum é a colisão, o que é semelhante ao padrão encontrado na análise das ocorrências reportadas na Rodovia Presidente Dutra.

Por fim, ainda há muita margem de melhoria para a rodovia, apesar dos números estarem diminuindo, continuam sendo bem altos, tendo uma necessidade de abordagens mais eficientes para a contínua diminuição de acidentes, com ações por parte do poder público, poder privado que administra a Rodovia e também da sociedade que tem a BR-116, uma rodovia que conecta duas grandes cidades como São Paulo e Rio de Janeiro, trecho fundamental para o seu deslocamento diário.

6. Referências Bibliográficas

[1] Acidentes - Conjuntos de Dados. Portal Brasileiro de Dados Abertos. 2022. Disponível em <https://dados.gov.br/dataset/acidentes-rodovias>. Acesso em 20/06/22

[2] Nova Dutra tem emissora de rádio que informa motoristas sobre condições da via. Agência Brasil . Disponível em <https://agenciabrasil.ebc.com.br/geral/noticia/2015-09/nova-dutra-tem-emissora-de-radio-que-informa-motoristas-sobre-condicoes-da> Acesso em 07/07/22

[3] Painel de acidentes. Conselho Nacional do Transporte Disponível em: <https://www.cnt.org.br/painel-acidente> Acesso em:08/07/2022

[4] Dicionário de Dados - Demonstrativo de Acidentes. Disponível em: <https://dados.gov.br/dataset/acidentes-rodovias/resource/47b96524-3bf9-47ef-b7a6-2d02c39d72ac> Acesso em: 20/06/22

[5] ARAÚJO, Marcelo. Conceitos e Definições em Acidentes. Disponível em: <https://www.portaldotransito.com.br/opiniao/conceitos-e-definicoes-em-acidentes-2/#:~:text=Vimos%20que%20na%20defini%C3%A7%C3%A3o%20de,que%20ocorreu%20um%20%E2%80%9Cchoque%E2%80%9D> Acesso em: 07/07/22

[6] Tudo sobre a Rodovia Presidente Dutra [BR-116]. Disponível em: <https://www.rodoviapresidentedutra.com.br/> Acesso em: 08/07/22

[7] CCR NovaDutra faz campanha para alertar caminhoneiros sobre sono no volante. Disponível em: <https://viatrolebus.com.br/2021/08/ccr-novadutra-faz-campanha-para-alertar-caminhoneiros-sobre-sono-no-volante/> Acesso em 16/07/22

7. Anexos

```
install.packages(c("summarytools", "fdth", "worldcloud2", "tm",
"ggplot2", "readxl", "readr", "dplyr", "stringr", "tidyverse"))

require(stringr)
require(summarytools)
require(fdth)
require(ggplot2)
require(readxl)
library(wordcloud2)
library(tm) # biblioteca de text mining para limpar o texto (wordcloud)
library(readr)
library(tidyverse)
library(dplyr)
#renomear o arquivo que foi baixado do portal de dados abertos
brasileiro

dadosRodovia <- demostrativo_acidentes_novadutra

#tratamento dos dados

#como os dados de 2022 estão incompletos (apenas 4 meses) eles serão
desconsiderados na análise
dadosRodovia$data <- as.Date(dadosRodovia$data, format= "%d/%m/%Y")
dadosRodovia <- filter(dadosRodovia, data >= "2010-01-01", data <=
"2021-12-31")

#o dataset original conta com celulas em branco representando 0. para as
analises é necessário substituir todas as celulas em branco por 0
dadosRodovia[is.na(dadosRodovia)] <- 0

#criar uma nova coluna no dataframe com a soma de todos os envolvidos na
ocorrencia = qtde_pessoas
dadosRodovia$qtde_pessoas = rowSums(dadosRodovia[,c("ilesos",
"levemente_feridos", "moderadamente_feridos", "gravemente_feridos",
"mortos")])

#criar uma nova coluna no dataframe usando o horario para determinar em
qual periodo do dia a ocorrencia foi registrada
dadosRodovia$periodo = substring(dadosRodovia$horario, 0, 2)
dadosRodovia['periodo'][dadosRodovia['periodo'] == '00' |
dadosRodovia['periodo'] == '01' | dadosRodovia['periodo'] == '02' |
dadosRodovia['periodo'] == '03' | dadosRodovia['periodo'] == '04' |
```

```
dadosRodovia['periodo'] == '05'] <- 'Madrugada (00h - 05h)'
dadosRodovia['periodo'][dadosRodovia['periodo'] == '06' |
dadosRodovia['periodo'] == '07' | dadosRodovia['periodo'] == '08' |
dadosRodovia['periodo'] == '09' | dadosRodovia['periodo'] == '10' |
dadosRodovia['periodo'] == '11'] <- 'Manhã (06h - 11h)'
dadosRodovia['periodo'][dadosRodovia['periodo'] == '12' |
dadosRodovia['periodo'] == '13' | dadosRodovia['periodo'] == '14' |
dadosRodovia['periodo'] == '15' | dadosRodovia['periodo'] == '16' |
dadosRodovia['periodo'] == '17'] <- 'Tarde (12h - 17h)'
dadosRodovia['periodo'][dadosRodovia['periodo'] == '18' |
dadosRodovia['periodo'] == '19' | dadosRodovia['periodo'] == '20' |
dadosRodovia['periodo'] == '21' | dadosRodovia['periodo'] == '22' |
dadosRodovia['periodo'] == '23'] <- 'Noite (18h - 23h)'

#para a analise é necessário padronizar os dados, a seguir o campo
tipo_de_ocorrencia é padronizado utilizando os parametros 'sem vitima'
ou 'com vitima'
#a planilha original possui acentuação na palavra vítima e por conta
disso é necessário uma correção

dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "sem v'tima", "sem
vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Acidente sem vitima",
"sem vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Acidente sem vítima",
"sem vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Atropelamento sem
morte", "sem vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "AC03 - Acidente sem
VITIMA", "sem vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Sem v'tima", "sem
vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Sem vítima", "sem
vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "sem vítima", "sem
vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "AC04 - Atropelamento",
```

```
"sem vitima")

dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "com vítima", "com
vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "AC01 - Acidente com
VITIMA FATAL", "com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Acidente com morte",
"com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "AC02 - Acidente com
VITIMA", "com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Acidente com vitima",
"com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Acidente com vítima",
"com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "AC05 - Atropelamento
Fatal", "com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Atropelamento com
morte", "com vitima")
dadosRodovia$tipo_de_ocorrencia =
str_replace_all(dadosRodovia$tipo_de_ocorrencia, "Com vítima", "com
vitima")

#elaborar o grafico de setores com a coluna trecho do dataframe
cols <- c("yellow2","midnightblue")
pielabels<-
paste(round(table(dadosRodovia$trecho)/length(dadosRodovia$trecho)*100),
"%", sep="")
pie(round((table(dadosRodovia$trecho)/length(dadosRodovia$trecho)*100),2
),labels=pielabels, cex=1.3, col=cols)
legend("bottomright", c("Rio de Janeiro","São Paulo"), cex = 1.1, fill =
c("yellow2","midnightblue"))

#grafico de setores utilizando o campo periodo para encontrar em qual
periodo mais aconteceram ocorrencias
cols <- c("red", "dodgerblue3", "green", "orange")
pielabels<-
paste(round(table(dadosRodovia$periodo)/length(dadosRodovia$periodo)*100
), "%", sep="")
```

```
pie(round((table(dadosRodovia$periodo)/length(dadosRodovia$periodo)*100)
,2),labels=pielabels, cex=1.3, col=cols)
legend("bottomright", c("Madrugada (00h - 05h)","Manhã (06h - 11h)",
"Tarde (12h - 17h)", "Noite (18h - 23h)"), cex = 1.1, fill = c("red",
"dodgerblue3", "green", "orange"))

#extrair o ano da coluna data que possui ano/mes/dia e elaborar o
grafico de barras com quantidade de acidentes x ano
anosAcidentes = select(dadosRodovia, data) #extrai a coluna data do
dataset
anosAcidentes = format(as.Date(anosAcidentes$data,
format="%d/%m/%Y"),"%Y") # o formato de data é dia/mês/ano, para a
primeira análise foi considerado apenas o ano
barplot(table(anosAcidentes), xlab = "Ano", ylab = "Quantidade de
Acidentes", cex.lab=1.0, cex.names=1.0, cex.axis=1.0, col="steelblue2",
ylim=c(0,13000))

#elaborar o grafico de barras com 2 variaveis qualitativas (tipo de
ocorrencia x trecho)
percentData <- dadosRodovia %>% group_by(`trecho`) %>%
count(`tipo_de_ocorrencia`) %>%  mutate(ratio=scales::percent(n/sum(n)))
ggplot(dadosRodovia,
aes(x=factor(trecho),fill=factor(tipo_de_ocorrencia))) +
  geom_bar(position="fill") +
  geom_text(data=percentData, aes(y=n,label=ratio),
position=position_fill(vjust=0.5))+
  xlab("Trecho") +
  ylab("Proporção de ocorrencias") +
  scale_fill_manual(name="Tipo de Ocorrencia", values = c("darksalmon",
"aquamarine3"))

#janela com vários gráficos de uma variável quantitativa (quantidade de
carros envolvidos nas ocorrencias)
quantidadeCarros <- dadosRodovia$automovel
quantidadeCarros <- as.integer(quantidadeCarros)
quantidadeCarros <- sort(quantidadeCarros)

tab=fdt(quantidadeCarros, start=0,h=1,end=15)
par(mfrow=c(1,3))
boxplot(quantidadeCarros, ylab = "Quantidade de Automoveis",
cex.axis=1.6, cex.lab=1.6)
points(mean(quantidadeCarros), pch=3)
plot(tab, type='rfph', xlab="Quantidade de Automoveis",ylab="%
ocorrencias", cex.axis=1.6, cex.lab=1.6)
par(new=TRUE)
plot(tab,type='rfpp', xlab="Quantidade de Automoveis",ylab="", col =
```

```
"black",cex.axis=1.6, cex.lab=1.6)
plot(tab,type='cfpp', xlab="Quantidade de Automoveis",ylab="%
acumulados", ylim=c(0,100), col = "black", cex.axis=1.6, cex.lab=1.6)

mean(quantidadeCarros) #média
median(quantidadeCarros) #mediana
sd(quantidadeCarros) #standard deviation = desvio padrão
(sd(quantidadeCarros)/mean(quantidadeCarros))*100 # coeficiente de
variação

#diagrama de dispersao, spearman e pearson
par(mar=c(4,4,2,0), oma=c(0,0,0,10))
plot(dadosRodovia$onibus~dadosRodovia$qtde_pessoas, xlab = "Quantidade
de pessoas envolvidas", ylab = "Quantidade de Onibus envolvidos", col =
"blue")
abline(lm(dadosRodovia$onibus~dadosRodovia$qtde_pessoas), col= "red")

pearson <- cor(dadosRodovia$onibus, dadosRodovia$qtde_pessoas); pearson
spearman <- cor(dadosRodovia$onibus, dadosRodovia$qtde_pessoas,
method="spearman"); spearman

#associação entre duas variaveis qualitativas (trecho x periodo)
percentData <- dadosRodovia %>% group_by(tipo_de_ocorrencia) %>%
count(periodo) %>%  mutate(ratio=scales::percent(n/sum(n)))
ggplot(dadosRodovia,
aes(x=factor(tipo_de_ocorrencia),fill=factor(periodo))) +
  geom_bar(position="fill") +
  geom_text(data=percentData, aes(y=n,label=ratio ),
position=position_fill(vjust=0.5))+
  xlab("Tipo de Ocorrencia") +
  ylab("Proporção de Ocorrencias")+
  scale_fill_manual(name="Periodo", values = c("red", "dodgerblue3",
"green", "orange"))

#criação de uma nuvem de palavras utilizando a coluna tipo_de_acidente
#fonte https://www.youtube.com/watch?v=0cToDzeDLRI&ab_channel=Dataslice

medium.corpus = Corpus(VectorSource(dadosRodovia$tipo_de_acidente))
medium.corpus = medium.corpus %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace) %>%
  tm_map(content_transformer(tolower))

tdm = TermDocumentMatrix(medium.corpus) %>%
  as.matrix()
```

```
words = sort(rowSums(tdm), decreasing = TRUE)
df = data.frame(word = names(words), freq = words)
wordcloud2(df, size = 1.0, minSize = 2, rotateRatio = 0)

#associação entre uma variavel quantitativa x uma variavel qualitativa
ggboxplot(dadosRodovia$automovel ~ dadosRodovia$periodo,xlab =
"Periodo", ylab = "Quantidade de Automoveis")
points(1:nlevels(dadosRodovia$periodo), tapply(dadosRodovia$automovel,
dadosRodovia$periodo, mean),  pch=3)
```

Conjunto de Dados