PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Edson Ferreira Lopes

ANÁLISE TEMPORAL DOS ACIDENTES DE TRÂNSITO

Belo Horizonte
2021

Edson Ferreira Lopes

ANÁLISE TEMPORAL DOS ACIDENTES DE TRÂNSITO

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução

1.1. Contextualização

De acordo com dados da Organização Mundial da Saúde (OMS) cerca de 1,25 milhão de pessoas morrem por ano no mundo vítimas de acidentes de trânsito, onde aproximadamente 50% das mortes são pedestres, ciclistas e motociclistas.

Um dos objetivos da Agenda para o Desenvolvimento Sustentável 2030 diz respeito a segurança no trânsito, que previa a redução pela metade do número global de mortes e lesões causadas por acidentes de trânsito até 2020.

O trânsito brasileiro é o quarto mais violento do continente americano, segundo dados divulgados pela Organização Mundial da Saúde (OMS). São Paulo é o estado brasileiro com maior número de óbitos no trânsito e dirigir alcoolizado é a segunda maior causa de acidentes. Com o propósito de reduzir o número de acidentes, foi publicada no ano de 2020 a Lei Ordinária 13.546 do Código de Trânsito Brasileiro, aumentando a punição para motoristas que dirigem alcoolizados e que causam mortes no trânsito, ou seja, a pena, que antes era de 2 a 4 anos de detenção, passa para 5 a 8 anos de reclusão.

1.2. O problema proposto

A análise dos dados de acidentes de trânsito no país nos ajuda a entender melhor este problema social e a subsidiar políticas públicas que possam colaborar para diminuição das mortes em nossas estradas.

Os dados a serem observados neste estudo são da Polícia Rodoviária Federal – PRF e se referem aos acidentes e mortes ocorridos nas rodovias federais brasileiras.

O propósito do estudo é fazer uma análise dos dados de acidentes e mortes no trânsito para melhor entendimento do problema social.

As bases que serão analisadas no estudo contém dados do ano de 2010 a 2020 da Polícia Rodoviária Federal – PRF, onde serão estudadas variáveis sobre as características das vítimas (sexo, idade, etc) e dos acidentes (localidade, tipo de pista, condição meteorológica, tipo de veículo envolvido, etc).

2. Coleta de Dados

As bases de dados de acidentes foram extraídas da plataforma on-line de Dados Abertos da Polícia Rodoviária Federal – PRF, compondo as bases anuais de acidentes (agrupadas por pessoa) referente ao período de 2010 a 2020. As bases estão em formato “csv” e foram estruturadas e agregadas no software RSTUDIO:

Dicionário de Dados de Acidentes – Ano 2010 a 2020

Nome da Variável	Descrição	Tipo
Id	Variável com valores numéricos, representando o identificador do acidente.	Numérica
data_inversa	Data da ocorrência no formato dd/mm/aaaa.	Data
dia_semana	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.	Categórica
horario	Horário da ocorrência no formato hh:mm:ss.	Data
uf	Unidade da Federação. Ex.: MG, PE, DF, etc.	Categórica
Regiao	Região: Norte, Nordeste, Sul, Sudeste e Centro-oeste	Categórica
br	Variável com valores numéricos, representando o identificador da BR do acidente.	Numérica
km	Identificação do quilômetro onde ocorreu o acidente, com valor mí imo de 0,1 km e com a casa decimal separada por ponto.	Numérica
municipio	Nome do município de ocorrência do acidente.	Categórica
causa_acidente	Causa presumível do acidente, baseada nos vestígios, indícios e provas colhidas no local do acidente.	Categórica
tipo_acidente	Identificação do tipo de acidente. Ex.: Colisão frontal, Saída de pista, etc.	Categórica
classificação_acidente	Classificação quanto à gravidade do acidente: Sem Vítimas, Com Vítimas Feridas, Com Vítimas Fatais e Ignorado.	Categórica
fase_dia	Fase do dia no momento do acidente. Ex. Amanhecer, Pleno dia, etc.	Categórica
sentido_via	Sentido da via considerando o ponto de colisão: Crescente e decrescente.	Categórica
condição_meteorologica	Condição meteorológica no momento do acidente: Céu claro, chuva, vento, etc.	Categórica
tipo_pista	Tipo da pista considerando a quantidade de faixas: Dupla, simples ou múltipla.	Categórica
tracado_via	Descrição do traçado da via.	Categórica
uso_solo	Descrição sobre as características do local do acidente: Urbano=Sim; Rural=Não.	Categórica
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro. Ex.: Automóvel, Caminhão, Motocicleta, etc.	Categórica

marca	Descrição da marca do veículo.	Categórica
ano_fabricacao_veiculo	Ano de fabricação do veículo, formato aaaa.	Numérica
pesid	Variável com valores numéricos, representando o identificador da pessoa envolvida.	Numérica
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento. Ex.: condutor, passageiro, pedestre, etc.	Categórica
estado_fisico	Condição do envolvido conforme a gravidade das lesões. Ex.: morto, ferido leve, etc.	Categórica
idade	Idade do envolvido. O código "-1" indica que não foi possível coletar tal informação.	Numérica
sexo	Sexo do envolvido. O valor "inválido" indica que não foi possível coletar tal informação.	Categórica
latitude	Latitude do local do acidente em formato geodésico decimal.	Numérica
longitude	Longitude do local do acidente em formato geodésico decimal.	Numérica

Obs.: dicionário com as variáveis comuns as bases anuais de acidentes e de interesse para o estudo.

Link: https://arquivos.prf.gov.br/arquivos/index.php/s/O2B79mMqJr74zoZ#pdfviewer

3. Processamento/Tratamento de Dados

A extração, tratamento e leitura dos dados dos acidentes de trânsito nas rodovias federais foi realizada utilizando o software RSTUDIO. O download das bases de dados referente aos anos de 2010 a 2020 foi extraído do link https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-acidentes conforme descrição:

Bases de Dados de Acidentes Agrupados por Pessoa – PRF

Nome das Bases de Dados	Número de Variáveis	Formato	Tamanho do Arquivo
acidentes2010	28 variáveis	csv	128.088 kb
acidentes2011	28 variáveis	csv	133.382 kb
acidentes2012	28 variáveis	csv	130.175 kb
acidentes2013	28 variáveis	csv	132.157 kb
acidentes2014	28 variáveis	csv	120.218 kb
acidentes2015	28 variáveis	csv	87.887 kb
acidentes2016	28 variáveis	csv	60.773 kb
acidentes2017	35 variáveis	csv	69.455 kb
acidentes2018	35 variáveis	csv	57.079 kb
acidentes2019	35 variáveis	csv	56.502 kb
acidentes2020	35 variáveis	csv	3.009 kb

Algumas consistências nas bases de dados foram realizadas para padronizar as informações relevantes ao nosso estudo. Variáveis desnecessárias e descontinuadas no histórico de acidentes de 2010 a 2020 foram eliminadas (

data_inversa, km, id_veiculo, nacionalidade, naturalidade, ilesos, feridos_leves, feridos_graves, mortos, latitude, longitude, regional, delegacia e uop) e 23 variáveis utilizadas (id, pesid, dia_semana, horário, uf, br, município, causa_acidente, tipo_acidente, classificacao_acidente, fase_dia, sentido_via, condição_meteorologica, tipo_pista, tracado_via, uso_solo, tipo_veiculo, marca, ano_fabricacao_veiculo, tipo_envolvido, estado_fisico, idade e sexo). 10 variáveis irrelevantes para este projeto foram excluídas do banco de acidentes (id_veiculo, nacionalidade, naturalidade, ilesos, feridos leves, feridos graves, mortos, regional, delegacia e uop).

Devido a problemas de padronização de formato nas bases disponibilizadas pela PRF, o script R de leitura dos dados não foi o mesmo em todos os anos:

```
# ACIDENTES - 2010 A 2016

ACIDENTES2010 <- read_csv("C:/TCC/Bases/acidentes2010.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2011 <- read_csv("C:/TCC/Bases/acidentes2011.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2012 <- read_csv("C:/TCC/Bases/acidentes2012.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2013 <- read_csv("C:/TCC/Bases/acidentes2013.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2014 <- read_csv("C:/TCC/Bases/acidentes2014.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2015 <- read_csv("C:/TCC/Bases/acidentes2015.csv", locale = locale(encoding = "WINDOWS-1252"))
ACIDENTES2016 <- read_delim("C:/TCC/Bases/acidentes2016.csv", ";", escape_double = FALSE, locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)


# ACIDENTES - 2017 A 2020

ACIDENTES2017 <- read_delim("C:/TCC/Bases/acidentes2017.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2018 <- read_delim("C:/TCC/Bases/acidentes2018.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2019 <- read_delim("C:/TCC/Bases/acidentes2019.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2020 <- read_delim("C:/TCC/Bases/acidentes2020.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)


ACIDENTES2020 <- read_delim("Bases/acidentes2020.csv", ";", escape_double = FALSE, col_types = cols(km = col_character(), latitude = col_character(), longitude = col_character()), trim_ws = TRUE)
```

A estruturação das variáveis de acidentes foi feita ano a ano e a unificação das bases em dois bancos distintos, o primeiro de acidentes entre 2010 e 2016 e o segundo de acidentes entre 2017 e 2020.

No primeiro banco de dados (acidentes entre 2010 e 2016) foram criadas as variáveis ano do acidente, mês do acidente, latitude e longitude. Ano do acidente foi inserida conforme a identificação do nome das bases extraídas da PRF, mês do acidente foi extraída pela função “substr” aplicada a variável “data_inversa” presente nas bases. As variáveis latitude e longitude não existem para os anos de 2010 a 2016, porém foram criadas “sem registro” para junção com o banco de dados de 2017 a 2020, onde se começou a coletar estes dados.

```
# CRIANDO VARIÁVEL ANO

ACIDENTES2010$Ano <- 2010
ACIDENTES2011$Ano <- 2011
ACIDENTES2012$Ano <- 2012
ACIDENTES2013$Ano <- 2013
ACIDENTES2014$Ano <- 2014
ACIDENTES2015$Ano <- 2015
ACIDENTES2016$Ano <- 2016


# CRIANDO VARIÁVEL MÊS

ACIDENTES2010$Mes <- substr(ACIDENTES2010$data_inversa,4,5)
ACIDENTES2011$Mes <- substr(ACIDENTES2011$data_inversa,4,5)
ACIDENTES2012$Mes <- substr(ACIDENTES2012$data_inversa,4,5)
ACIDENTES2013$Mes <- substr(ACIDENTES2013$data_inversa,4,5)
ACIDENTES2014$Mes <- substr(ACIDENTES2014$data_inversa,4,5)
ACIDENTES2015$Mes <- substr(ACIDENTES2015$data_inversa,4,5)
ACIDENTES2016$Mes <- substr(ACIDENTES2016$data_inversa,4,5)


# AGREGANDO AS BASES DE DADOS DE 2017 A 2020

ACIDENTES2010a2016 <- rbind(ACIDENTES2010, ACIDENTES2011, ACIDENTES2012, ACIDENTES2013, ACIDENTES2014, ACIDENTES2015, ACIDENTES2016)


# CRIANDO AS VARIÁVEIS LATITUDE E LONGITUDE

ACIDENTES2010a2016$latitude  <- ''
ACIDENTES2010a2016$longitude <- ''


# REMOVENDO BASES
remove(ACIDENTES2010,ACIDENTES2011,ACIDENTES2012,ACIDENTES2013,ACIDENTES2014,ACIDENTES2015,ACIDENTES2016)
```

No banco de acidentes entre 2017 e 2020, a leitura original das bases “.csv” não compilou corretamente as variáveis numéricas latitude e longitude. Foi preciso, então, modificar o formato inicial das variáveis numéricas para o formato de caractere, evitando, com isso, perda de informação. O mesmo procedimento foi feito na leitura dos dados de “data_inversa”, pois o formato desta variável não estava no mesmo padrão na série histórica entre 2017 e 2020.

```
# ACIDENTES - 2017 A 2020

ACIDENTES2017 <- read_delim("C:/TCC/Bases/acidentes2017.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2018 <- read_delim("C:/TCC/Bases/acidentes2018.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2019 <- read_delim("C:/TCC/Bases/acidentes2019.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2020 <- read_delim("C:/TCC/Bases/acidentes2020.csv", ";", escape_double = FALSE, col_types = cols(latitude = col_character(), longitude = col_character()), locale = locale(encoding = "WINDOWS-1252"), trim_ws = TRUE)

ACIDENTES2020 <- read_delim("Bases/acidentes2020.csv", ";", escape_double = FALSE, col_types = cols(km = col_character(), latitude = col_character(), longitude = col_character()), trim_ws = TRUE)


ACIDENTES2017$data_inversa <- as.character(ACIDENTES2017$data_inversa)
ACIDENTES2018$data_inversa <- as.character(ACIDENTES2018$data_inversa)
ACIDENTES2019$data_inversa <- as.character(ACIDENTES2019$data_inversa)
ACIDENTES2020$data_inversa <- as.character(ACIDENTES2020$data_inversa)
```

Realizada as alterações no banco de acidentes entre 2017 e 2020 foram criadas as variáveis ano do acidente e mês do acidente, seguindo procedimento semelhante ao banco de 2010 a 2016.

```
# CRIANDO VARIÁVEL ANO

ACIDENTES2017$Ano <- 2017
ACIDENTES2018$Ano <- 2018
ACIDENTES2019$Ano <- 2019
ACIDENTES2020$Ano <- 2020


# CRIANDO VARIÁVEL MÊS

ACIDENTES2017$Mes <- substr(ACIDENTES2017$data_inversa,6,7)
ACIDENTES2018$Mes <- substr(ACIDENTES2018$data_inversa,6,7)
ACIDENTES2019$Mes <- substr(ACIDENTES2019$data_inversa,6,7)
ACIDENTES2020$Mes <- substr(ACIDENTES2020$data_inversa,6,7)


# AGREGANDO AS BASES DE DADOS DE 2017 A 2020

ACIDENTES2017a2020 <- rbind(ACIDENTES2017, ACIDENTES2018, ACIDENTES2019, ACIDENTES2020)


# REMOVENDO BASES

remove(ACIDENTES2017, ACIDENTES2018, ACIDENTES2019, ACIDENTES2020)
```

Um único banco de acidentes de 2010 a 2020 foi criado a partir da junção das duas bases, retirando variáveis distintas e sem utilidade para a análise (id_veiculo, nacionalidade, naturalidade, ilesos, feridos leves, feridos graves, mortos, regional, delegacia e uop).

```
# DELETANDO VARIÁVEIS QUE NÃO SERÃO UTILIZADAS

ACIDENTES2010a2016 <- ACIDENTES2010a2016 %>% select (-id_veiculo, -nacionalidade, -naturalidade)
ACIDENTES2017a2020 <- ACIDENTES2017a2020 %>% select (-id_veiculo, -ilesos, -feridos_leves, -feridos_graves,
                                        -mortos, -regional, -delegacia, -uop)


# JUNÇÃO DOS ACIDENTES EM BASE ÚNICA #

ACIDENTES <- rbind(ACIDENTES2010a2016,ACIDENTES2017a2020)


# REMOVENDO BASES DESNECESSÁRIAS #

remove(ACIDENTES2010a2016,ACIDENTES2017a2020)
```

Uma análise de frequência simples foi realizada para verificar possíveis distorções de valores ou de categorias descritas nas variáveis, resultando na correção e na padronização dos valores observados no banco de dados.

```
# DISTRIBUIÇÃO DE FREQUÊNCIA E ESTRUTURAÇÃO DOS DADOS

table(ACIDENTES$dia_semana)
table(ACIDENTES$uf)
table(ACIDENTES$causa_acidente)
table(ACIDENTES$tipo_acidente)
table(ACIDENTES$classificacao_acidente)
table(ACIDENTES$fase_dia)
table(ACIDENTES$sentido_via)
table(ACIDENTES$condicao_metereologica)
table(ACIDENTES$tipo_pista)
table(ACIDENTES$tracado_via)
table(ACIDENTES$uso_solo)
table(ACIDENTES$tipo_veiculo)
table(ACIDENTES$tipo_envolvido)
table(ACIDENTES$estado_fisico)
table(ACIDENTES$sexo)


# ESTRUTURAÇÃO DA VARIÁVEL IDADE

ACIDENTES$idade2 <- ifelse(ACIDENTES$idade>=0  & ACIDENTES$idade<13  ,"Até 12 anos"     ,
                   ifelse(ACIDENTES$idade>=13 & ACIDENTES$idade<18  ,"13 a 17 anos"    ,
                   ifelse(ACIDENTES$idade>=18 & ACIDENTES$idade<26  ,"18 a 25 anos"    ,
                   ifelse(ACIDENTES$idade>=26 & ACIDENTES$idade<36  ,"26 a 35 anos"    ,
                   ifelse(ACIDENTES$idade>=36 & ACIDENTES$idade<46  ,"36 a 45 anos"    ,
                   ifelse(ACIDENTES$idade>=46 & ACIDENTES$idade<=120,"Acima de 45 anos",
                   "N.I.")))))

ACIDENTES$idade_ord <- ifelse(ACIDENTES$idade2=="Até 12 anos"      ,1,       # ordenador para visualização no PWBI
                       ifelse(ACIDENTES$idade2=="13 a 17 anos"     ,2,
                       ifelse(ACIDENTES$idade2=="18 a 25 anos"     ,3,
                       ifelse(ACIDENTES$idade2=="26 a 35 anos"     ,4,
                       ifelse(ACIDENTES$idade2=="36 a 45 anos"     ,5,
                       ifelse(ACIDENTES$idade2=="Acima de 45 anos" ,6,
                       7))))))
ACIDENTES$br <- as.numeric(ACIDENTES$br)
```

A falta de padronização observada entre as variáveis no banco de dados foi um dos problemas encontrado na estruturação, com isso foi proposto a junção (join) de tabelas “Dimensão” criadas para facilitar a estruturação das variáveis. A tabela

“Dim_BR” foi utilizada para criar uma padronização da variável “br”, a tabela “Dim_DIA_SEMANA” foi utilizada para a padronização da variável “dia_semana”, a “Dim_UF” para a variável “uf”, sendo todas as demais tabelas “dimensões” com a mesmo objetivo. Problemas de espaçamentos nos nomes das variáveis e de valores nulos foram identificados no banco de dados e corrigidos para realizar a junção dos dados (mech code).

```
# ESTRUTURAÇÃO DE TABELAS AUXILIARES PARA CATEGORIZAÇÃO DOS DADOS DE ACIDENTES

Dim_BR          <- read_excel("C:/TCC/Bases Dimensões/Dim_BR.xlsx")
Dim_DIA_SEMANA    <- read_excel("C:/TCC/Bases Dimensões/Dim_DIA_SEMANA.xlsx")
Dim_UF          <- read_excel("C:/TCC/Bases Dimensões/Dim_UF.xlsx")
Dim_CAUSA_ACIDENTE<- read_excel("C:/TCC/Bases Dimensões/Dim_CAUSA_ACIDENTE.xlsx")
Dim_TP_ACIDENTE   <- read_excel("C:/TCC/Bases Dimensões/Dim_TIPO_DE_ACIDENTE.xlsx")
Dim_CLASSIF       <- read_excel("C:/TCC/Bases Dimensões/Dim_CLASSIF_ACIDENTE.xlsx")
Dim_FASE_DO_DIA   <- read_excel("C:/TCC/Bases Dimensões/Dim_FASE_DO_DIA.xlsx")
Dim_SENTIDO_DA_VIA<- read_excel("C:/TCC/Bases Dimensões/Dim_SENTIDO_DA_VIA.xlsx")
Dim_COND_METEOR   <- read_excel("C:/TCC/Bases Dimensões/Dim_COND_METEOR.xlsx")
Dim_TIPO_DE_PISTA <- read_excel("C:/TCC/Bases Dimensões/Dim_TIPO_DE_PISTA.xlsx")
Dim_TRACADO_VIA   <- read_excel("C:/TCC/Bases Dimensões/Dim_TRACADO_VIA.xlsx")
Dim_USO_DO_SOLO   <- read_excel("C:/TCC/Bases Dimensões/Dim_USO_DO_SOLO.xlsx")
Dim_TP_VEICULO    <- read_excel("C:/TCC/Bases Dimensões/Dim_TIPO_VEICULO.xlsx")
Dim_TIPO_ENVOLVIDO<- read_excel("C:/TCC/Bases Dimensões/Dim_TIPO_ENVOLVIDO.xlsx")
Dim_ESTADO_FISICO <- read_excel("C:/TCC/Bases Dimensões/Dim_ESTADO_FISICO.xlsx")
Dim_SEXO          <- read_excel("C:/TCC/Bases Dimensões/Dim_SEXO.xlsx")

Dim_UF <- Dim_UF %>% select(uf, Estado, Regiao, Regiao_ord, uf_ord, Estado_Maps) # variáveis de interesse


# RETIRANDO ESPAÇOS VAZIOS ANTES E NO FINAL DE CADA VARIÁVEL DA BASE DE ACIDENTES


ACIDENTES$br                  <- str_trim(ACIDENTES$br                  )
ACIDENTES$dia_semana          <- str_trim(ACIDENTES$dia_semana          )
ACIDENTES$uf                  <- str_trim(ACIDENTES$uf                  )
ACIDENTES$causa_acidente      <- str_trim(ACIDENTES$causa_acidente      )
ACIDENTES$tipo_acidente       <- str_trim(ACIDENTES$tipo_acidente       )
ACIDENTES$classificacao_acidente <- str_trim(ACIDENTES$classificacao_acidente)
ACIDENTES$fase_dia            <- str_trim(ACIDENTES$fase_dia            )
ACIDENTES$sentido_via         <- str_trim(ACIDENTES$sentido_via         )
ACIDENTES$condicao_metereologica <- str_trim(ACIDENTES$condicao_metereologica)
ACIDENTES$tipo_pista          <- str_trim(ACIDENTES$tipo_pista          )
ACIDENTES$tracado_via         <- str_trim(ACIDENTES$tracado_via         )
ACIDENTES$uso_solo            <- str_trim(ACIDENTES$uso_solo            )
ACIDENTES$tipo_veiculo        <- str_trim(ACIDENTES$tipo_veiculo        )
ACIDENTES$tipo_envolvido      <- str_trim(ACIDENTES$tipo_envolvido      )
ACIDENTES$estado_fisico       <- str_trim(ACIDENTES$estado_fisico       )
ACIDENTES$sexo                <- str_trim(ACIDENTES$sexo                )


# JUNÇÕES DE DIMENSÕES A BASE DE ACIDENTES


ACIDENTES <- merge(ACIDENTES,Dim_BR           , by.x = "br", by.y = "br", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_DIA_SEMANA   , by.x = "dia_semana", by.y = "dia_semana", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_UF           , by.x = "uf", by.y = "uf", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_CAUSA_ACIDENTE , by.X = "causa_acidente", by.y = "causa_acidente", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_TP_ACIDENTE  , by.x = "tipo_acidente", by.y = "tipo_acidente", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_CLASSIF      , by.x = "classificacao_acidente", by.y = "classificacao_acidente",
all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_FASE_DO_DIA  , by.x = "fase_dia", by.y = "fase_dia", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_SENTIDO_DA_VIA , by.x = "sentido_via", by.y = "sentido_via", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_COND_METEOR   , by.x = "condicao_metereologica", by.y =
"condicao_metereologica", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_TIPO_DE_PISTA  , by.x = "tipo_pista", by.y = "tipo_pista", all.x=T)
```

```
ACIDENTES <- merge(ACIDENTES,Dim_TRACADO_VIA    , by.x = "tracado_via", by.y = "tracado_via", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_USO_DO_SOLO    , by.x = "uso_solo", by.y = "uso_solo", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_TP_VEICULO     , by.x = "tipo_veiculo", by.y = "tipo_veiculo", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_TIPO_ENVOLVIDO , by.x = "tipo_envolvido", by.y = "tipo_envolvido", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_ESTADO_FISICO  , by.x = "estado_fisico", by.y = "estado_fisico", all.x=T)
ACIDENTES <- merge(ACIDENTES,Dim_SEXO           , by.x = "sexo", by.y = "sexo", all.x=T)


# CONFERINDO AS VARIÁVEIS DA BASE ESTRUTURADA

str(ACIDENTES)


# RESOLVENDO O PROBLEMA DE CELULAS NULAS IDENTIFICADAS NA BASE DE DADOS #

ACIDENTES$idade2   <- coalesce(ACIDENTES$idade2   ,"N.I.")
ACIDENTES$idade_ord<- coalesce(ACIDENTES$idade_ord,7)

ACIDENTES$dia_semana2   <- coalesce(ACIDENTES$dia_semana2   ,"N.I.")
ACIDENTES$dia_semana_ord<- coalesce(ACIDENTES$dia_semana_ord,8)

ACIDENTES$uf2   <- coalesce(ACIDENTES$uf2   ,"N.I.")
ACIDENTES$uf_ord<- coalesce(ACIDENTES$uf_ord,28)
ACIDENTES$regiao_ord<- coalesce(ACIDENTES$regiao_ord,6)

ACIDENTES$causa_acidente2   <- coalesce(ACIDENTES$causa_acidente2   ,"N.I.")
ACIDENTES$causa_acidente_ord<- coalesce(ACIDENTES$causa_acidente_ord,9)

ACIDENTES$tipo_acidente2 <- coalesce(ACIDENTES$tipo_acidente2   ,"N.I.")
ACIDENTES$tipo_acidente_ord<- coalesce(ACIDENTES$tipo_acidente_ord,9)

ACIDENTES$classificacao_acidente2 <- coalesce(ACIDENTES$classificacao_acidente2   ,"N.I.")
ACIDENTES$classificacao_acidente_ord<- coalesce(ACIDENTES$classificacao_acidente_ord,4)

ACIDENTES$fase_dia2   <- coalesce(ACIDENTES$fase_dia2   ,"N.I.")
ACIDENTES$fase_dia_ord<- coalesce(ACIDENTES$fase_dia_ord,5)

ACIDENTES$sentido_via2  <- coalesce(ACIDENTES$sentido_via2   ,"N.I.")
ACIDENTES$sentido_via_ord<- coalesce(ACIDENTES$sentido_via_ord,4)

ACIDENTES$condicao_metereologica2  <- coalesce(ACIDENTES$condicao_metereologica2   ,"N.I.")
ACIDENTES$condicao_metereologica_ord<- coalesce(ACIDENTES$condicao_metereologica_ord,8)

ACIDENTES$tipo_pista2    <- coalesce(ACIDENTES$tipo_pista2   ,"N.I.")
ACIDENTES$tipo_pista_ord <- coalesce(ACIDENTES$tipo_pista_ord,4)

ACIDENTES$tracado_via2    <- coalesce(ACIDENTES$tracado_via2   ,"N.I.")
ACIDENTES$tracado_via_ord <- coalesce(ACIDENTES$tracado_via_ord,11)

ACIDENTES$uso_solo2    <- coalesce(ACIDENTES$uso_solo2   ,"N.I.")
ACIDENTES$uso_solo_ord <- coalesce(ACIDENTES$uso_solo_ord,3)

ACIDENTES$tipo_veiculo2    <- coalesce(ACIDENTES$tipo_veiculo2   ,"N.I.")
ACIDENTES$tipo_veiculo_ord <- coalesce(ACIDENTES$tipo_veiculo_ord,7)

ACIDENTES$tipo_envolvido2    <- coalesce(ACIDENTES$tipo_envolvido2   ,"N.I.")
ACIDENTES$tipo_envolvido_ord <- coalesce(ACIDENTES$tipo_envolvido_ord,11)

ACIDENTES$sexo2    <- coalesce(ACIDENTES$sexo2   ,"N.I.")
ACIDENTES$sexo_ord <- coalesce(ACIDENTES$sexo_ord,3)

ACIDENTES$sexo2    <- coalesce(ACIDENTES$sexo2   ,"N.I.")
ACIDENTES$sexo_ord <- coalesce(ACIDENTES$sexo_ord,3)
```

O banco de dados estruturado foi exportado em formato “.csv” para visualização dos dados no software Power BI.

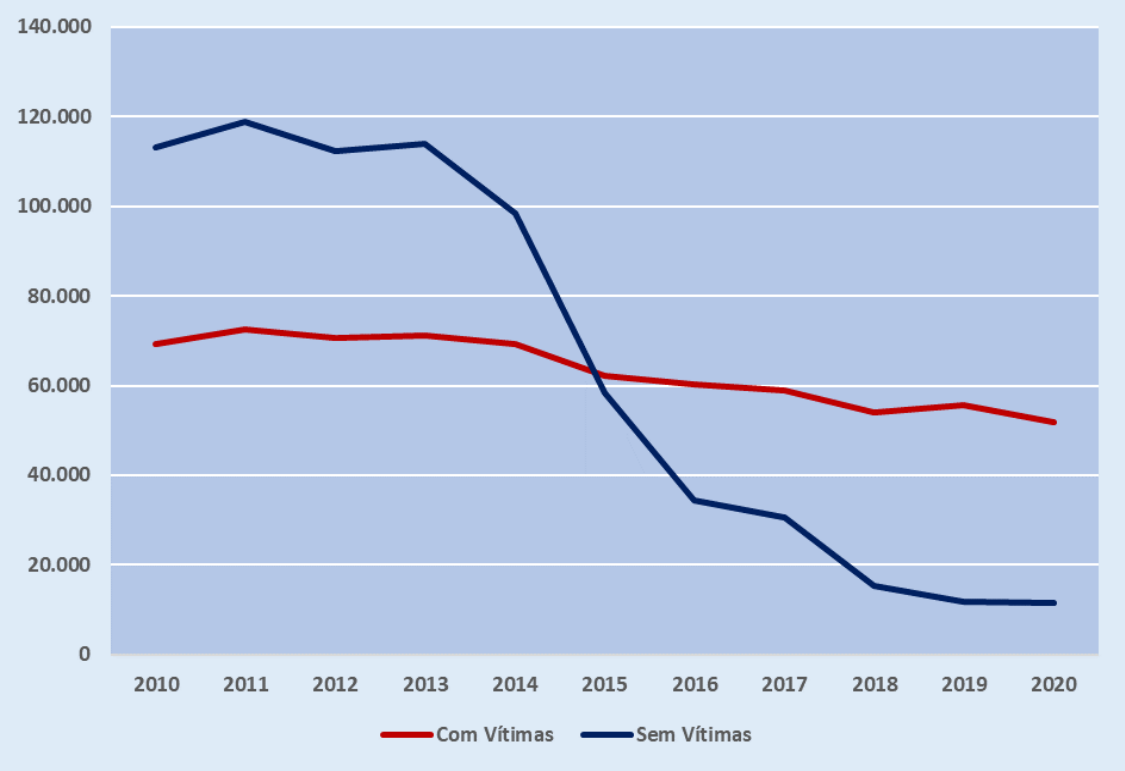
```
# EXPORTANDO OS DADOS PARA VISUALIZAÇÃO EM POWER BI

write.csv2(ACIDENTES, "C:/TCC/Bases/Acidentes_TODOS.csv")
```

4. Análise e Exploração dos Dados

A visualização dos dados de acidentes mostra uma significativa redução no número total de registros de acidentes em rodovias federais brasileiras, principalmente a partir de 2015, quando os registros de ocorrência de acidentes sem vítimas deixaram de ser preenchidos obrigatoriamente pelo boletim de ocorrência elaborado pelos policiais rodoviários e passaram a ser preenchidos por meio da declaração eletrônica de acidente de trânsito (e-DAT) na internet.

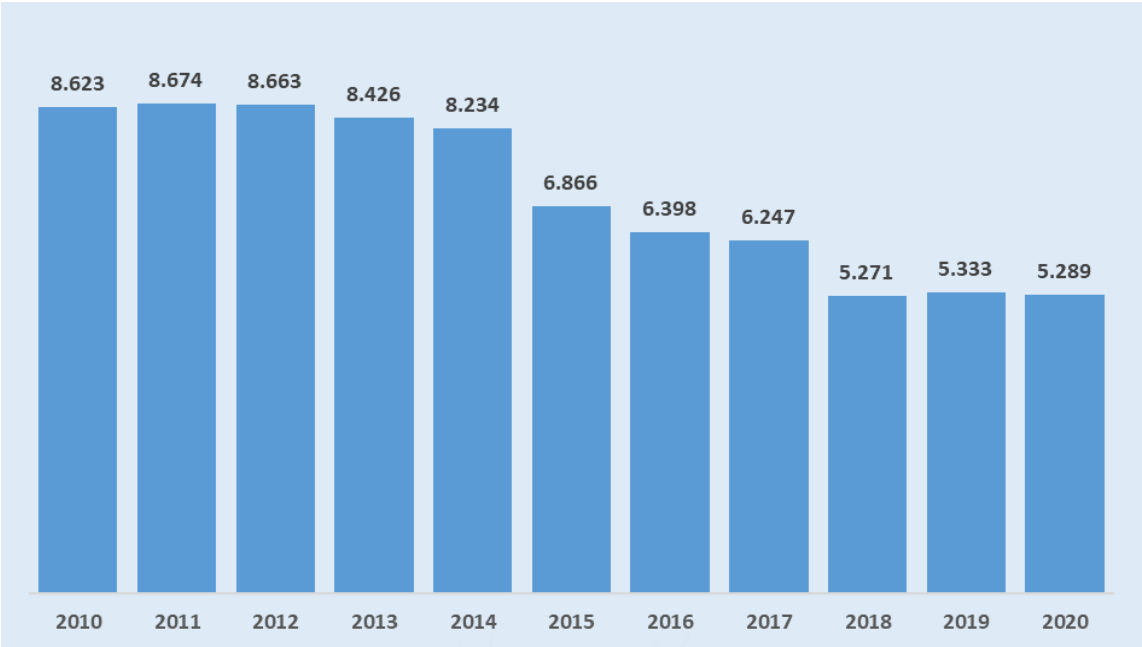
FONTE: PRF – Polícia Rodoviária Federal

A quantidade de acidentes com vítima reduziu 25,1% entre o período de 2010 a 2020, enquanto os acidentes sem vítima reduziram 89,8% no mesmo período.

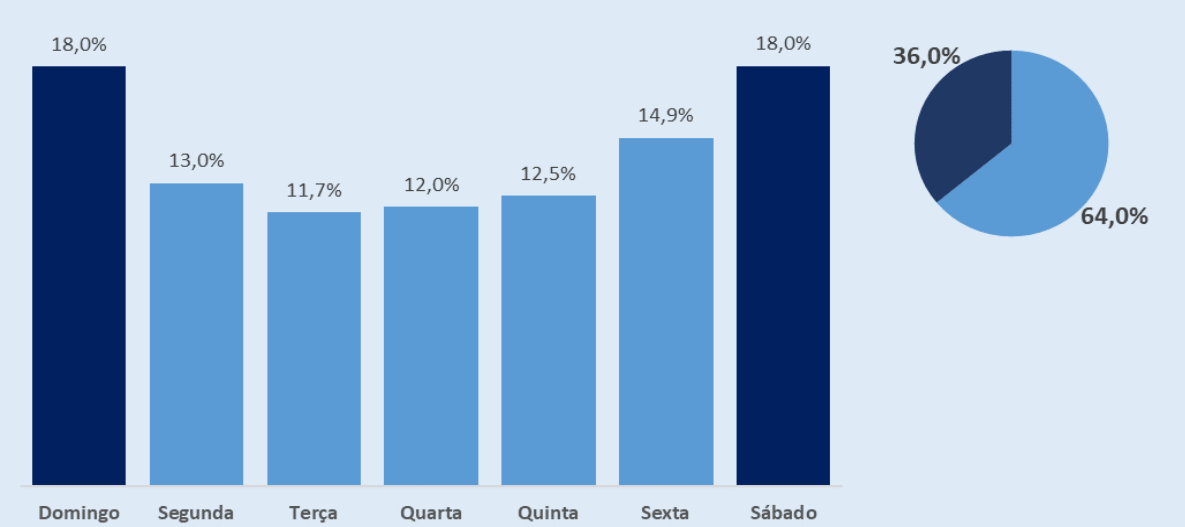


FONTE: PRF – Polícia Rodoviária Federal

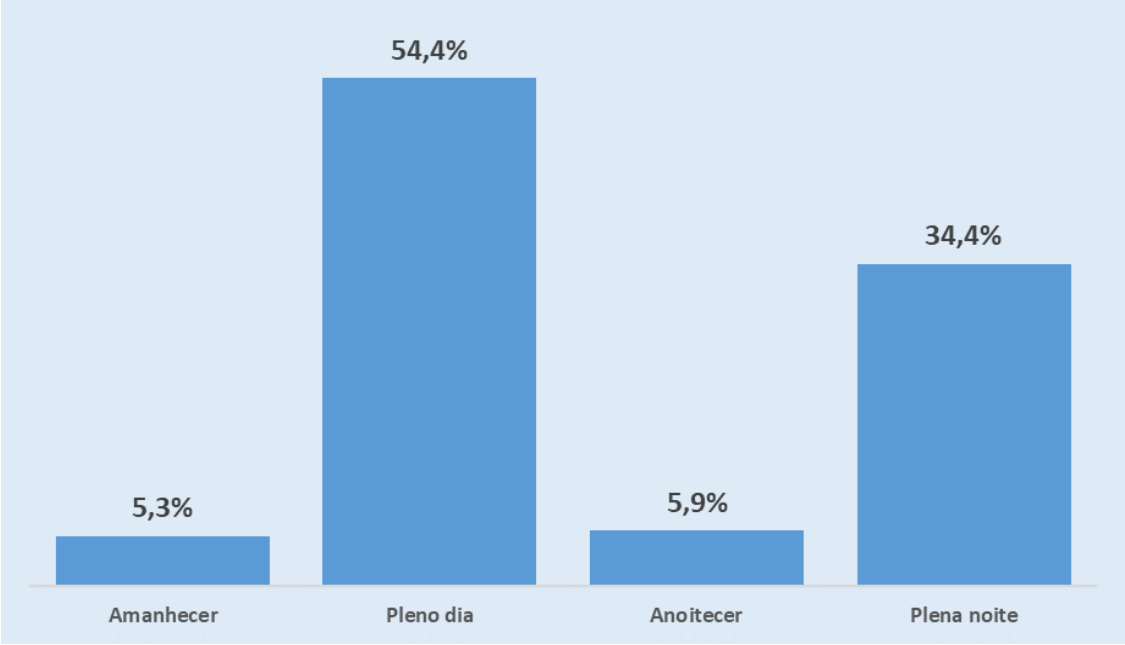
Em 11 anos de série histórica observamos 1.424.653 acidentes ocorridos em rodovias federais brasileiras, que deixaram 78.024 mortos e 1.007.557 feridos, considerando o ano de 365 dias, temos aproximadamente 19 mortes por dia no trânsito em rodovias. A redução de mortes no período foi de 38,7% entre 2010 e 2020.



Os acidentes de trânsito ocorrem com maior frequência nos finais de semana, 36% dos acidentes com vítimas ocorrem somente no sábado e domingo.

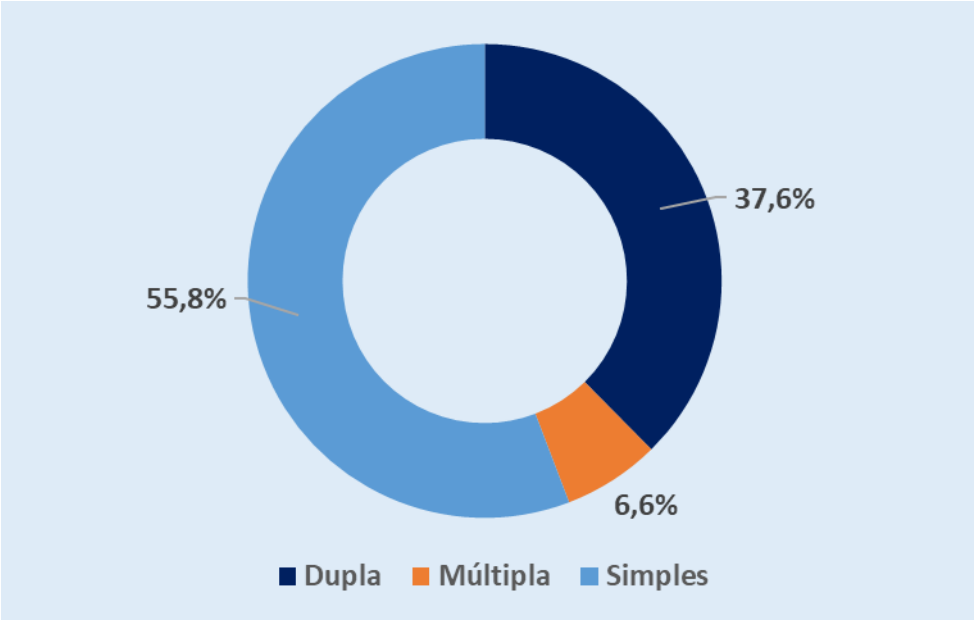


A maioria dos acidentes com vítimas ocorrem em pleno dia (57,9% dos acidentes). A fase do dia onde ocorrem menos acidentes é ao amanhecer (5,1% dos acidentes).

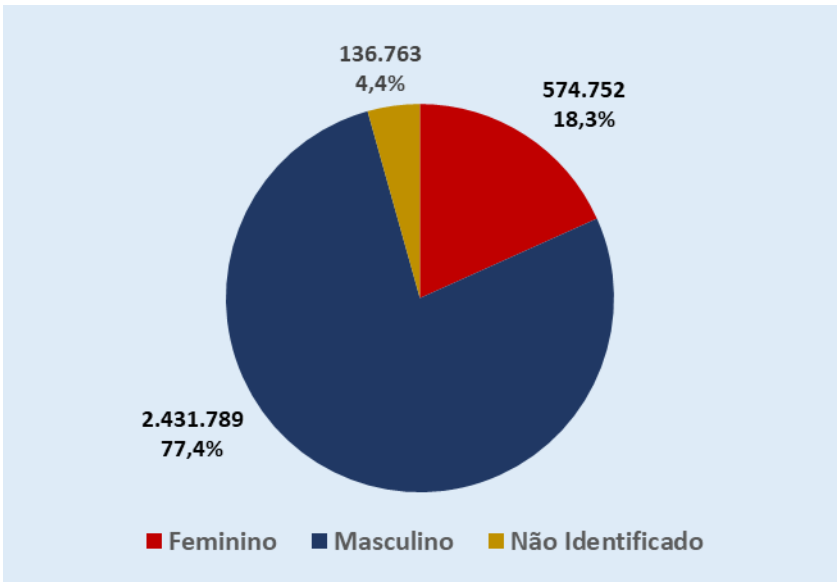


As rodovias de pista simples são onde ocorrem a maioria dos acidentes com vítimas no Brasil (55,8%), possivelmente agravada pela crescente demanda de

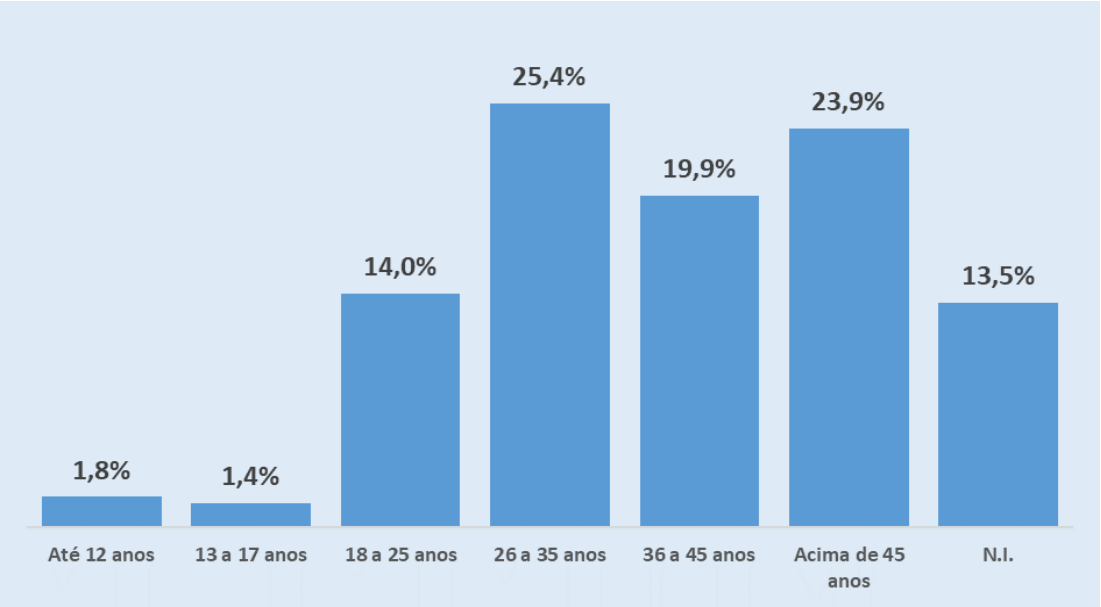
veículos em circulação ao longo dos anos e pelo precário estado de conservação em que se encontram, em sua maioria com buracos e sem acostamento.



Os homens são os que se envolvem mais em acidentes de trânsito, no período de 2010 a 2020 foram 2.431.789 vítimas do sexo masculino envolvidos em acidentes, o que corresponde a 77,4% do total de vítimas registrado no período, e 574.752 vítimas femininas envolvidas (18,3% do total). O registro de 136.763 vítimas não houve identificação de sexo na vítima no acidente (4,4% do total).



Jovens entre 26 a 35 anos são os que mais se envolvem em acidentes, cerca de 25,4% das vítimas envolvidas em acidentes estão nesta faixa etária.



5. Serie Histórica dos Acidentes com Vítimas

Existem vários tipos de modelos preditivos que podem ser utilizados em uma análise de dados, todos eles possuem características distintas e parâmetros próprios que se adequam a cada situação específica de um conjunto de dados. Dentre os modelos mais utilizados podemos citar as Árvores de Decisão, Regressão Logística, Regressão Linear, ARIMA, SARIMA, dentre outros. Toda esta variedade de modelos existe para se ajustar aos distintos objetos de estudo do mundo real.

Avaliando o comportamento do número diário de acidentes com vítimas nas rodovias federais no período de 2007 a 2019 foi proposto um modelo de séries temporais para previsão do número de acidentes com vítimas.

O modelo ARIMA (AutoRegressive Integrated Moving Average) é um dos mais utilizados em previsão de séries temporais, cujo a resposta (y) combina um modelo Auto-regressivo, Integrado e de Média Móvel. “O modelo de séries temporais utiliza dados do passado com o intuito de fazer predições futuras. As variações do modelo ARIMA permitem identificar e considerar a sazonalidade (modelo SARIMA),

podendo ser apenas um modelo autoregressivo (AR), apenas de média móvel (MA), autoregressivo de média móvel (ARMA), dentre outras variações existentes.

A decomposição de uma série temporal pode conter em sua estrutura de cálculo as componentes de autocorrelação, de tendência, de média móvel, de sazonalidade e de aleatoriedade.

A autocorrelação trata-se da correlação existente entre os dados de períodos anteriores com o dado atual, denominado de “lag”.

A tendência refere-se ao comportamento de crescimento e decrescimento da série ao longo do tempo.

A média móvel consiste no cálculo da média em determinados períodos da série temporal com o propósito de suavizar o comportamento da série ao longo do tempo, eliminando pontos extremos.

A sazonalidade refere-se ao padrão de comportamento da série histórica em períodos iguais de tempo, ou seja, em uma série histórica de acidentes com vítimas, temos um comportamento semelhante da série a cada mês.

A aleatoriedade é todo o fator externo que influencia no comportamento da série temporal, porém ele não é explicado matematicamente no modelo.

O modelo ARIMA$(p,d,q)_1(p,d,q)_2[n]$ corresponde ao modelo autoregressivo integrado de média móvel, onde, a letra “p” corresponde a componente de autocorrelação, a letra “d” a componente de integração e a letra “q” o componente de média-móvel, tanto nas componentes da modelagem geral $(p,d,q)_1$, como nas componentes de sazonalidade $(p,d,q)_2$, o “n” representa o período de análise da série, se ela é anual(frequência=1), trimestral(frequência=4), mensal(frequência=12), semanal(frequência=52), etc.

A componente de integração “d” de uma série temporal representa o número de diferenciações que a série precisa realizar para se tornar estacionária, ou seja, para que o comportamento da série histórica tenha variação em torno do valor zero ao longo do tempo.

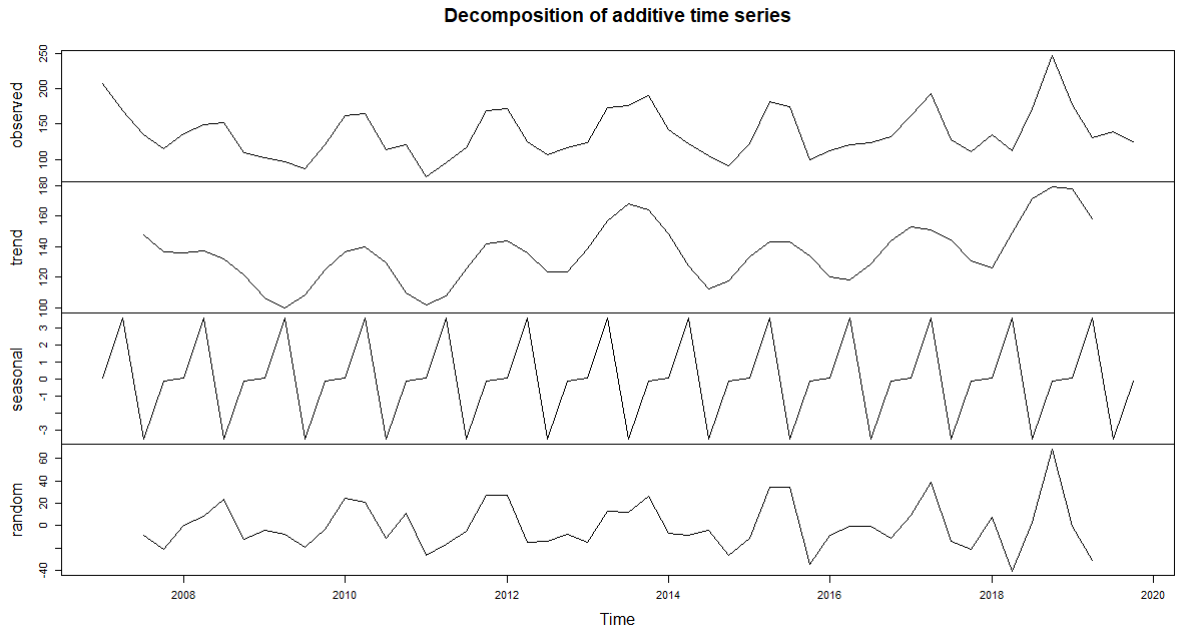
Após testes de consistência com modelos preditivos no software RSTUDIO, obteve-se o ARIMA (0,0,1) (1,0,0) [4] para realizar a previsão do número de acidentes com vítimas nas rodovias federais brasileiras, o modelo obtido para a resposta (y) combina uma modelo média móvel de ordem 1 e autorregressivo sazonal de ordem 1, com sazonalidade observada a cada trimestre (frequência=4).

Testes de consistência para escolha do modelo com menor AIC no RSTUDIO

```
> auto.arima(dados1, trace = TRUE, approximation = FALSE)

 ARIMA(2,0,2)(1,0,1)[4] with non-zero mean : Inf
 ARIMA(0,0,0)           with non-zero mean : 518.9016
 ARIMA(1,0,0)(1,0,0)[4] with non-zero mean : 504.8666
 ARIMA(0,0,1)(0,0,1)[4] with non-zero mean : 500.0763
 ARIMA(0,0,0)           with zero mean     : 664.4969
 ARIMA(0,0,1)           with non-zero mean : 499.8304
 ARIMA(0,0,1)(1,0,0)[4] with non-zero mean : 499.7776
 ARIMA(0,0,1)(2,0,0)[4] with non-zero mean : 501.953
 ARIMA(0,0,1)(1,0,1)[4] with non-zero mean : 502.0208
 ARIMA(0,0,1)(2,0,1)[4] with non-zero mean : Inf
 ARIMA(0,0,0)(1,0,0)[4] with non-zero mean : 514.4958
 ARIMA(1,0,1)(1,0,0)[4] with non-zero mean : 501.51
 ARIMA(0,0,2)(1,0,0)[4] with non-zero mean : 500.934
 ARIMA(1,0,2)(1,0,0)[4] with non-zero mean : 501.451
 ARIMA(0,0,1)(1,0,0)[4] with zero mean     : 541.117

 Best model: ARIMA(0,0,1)(1,0,0)[4] with non-zero mean

Series: dados1
ARIMA(0,0,1)(1,0,0)[4] with non-zero mean

Coefficients:
         ma1     sar1      mean
      0.5895  -0.2328  137.2432
s.e.  0.1002   0.1451    4.8619

sigma^2 = 771.9:  log likelihood = -245.46
AIC=498.93   AICc=499.78   BIC=506.73
```

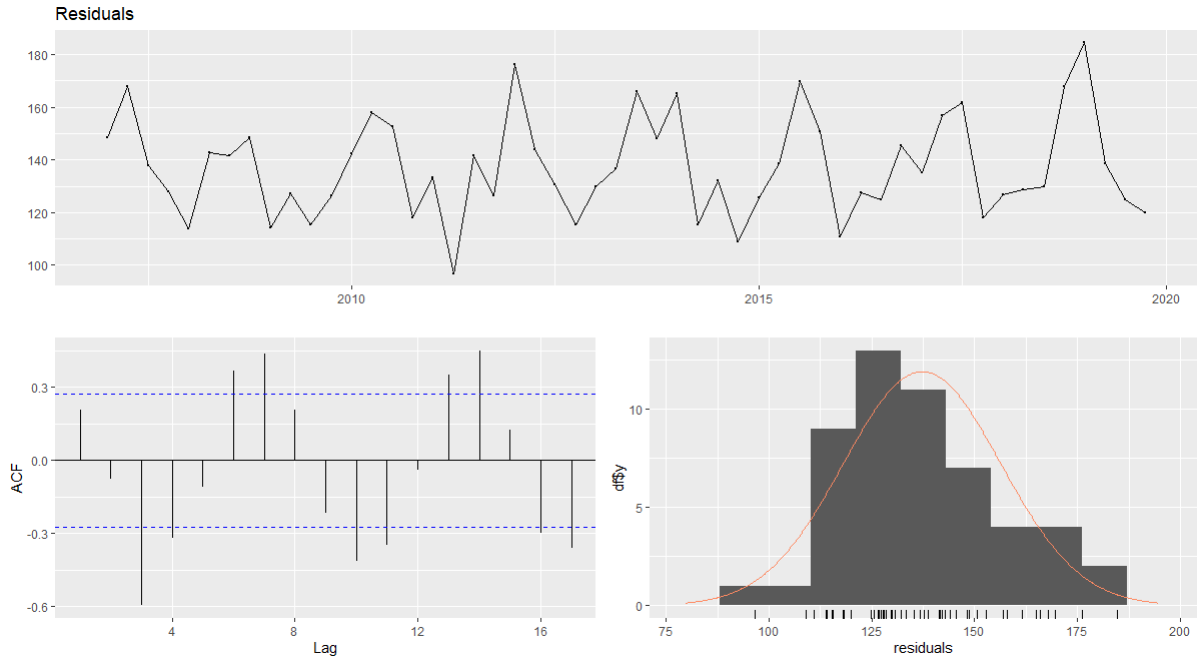
A decomposição da série temporal realizada no software RSTUDIO pode ser descrita em quatros gráficos: o primeiro gráfico (observed), corresponde a série histórica original dos dados de acidentes com vítimas; o segundo gráfico (trend) extrai da série histórica o comportamento de tendência dos dados (se ele é crescente, decrescente ou estacionário ao longo da série), graficamente, não é visível uma tendência de crescimento ou decrescimento, e sim um processo estacionário; o terceiro gráfico (seasonal) se refere ao comportamento sazonal da série, graficamente temos um comportamento sazonal bem definido entre os trimestres de cada ano; o quarto e último gráfico (random), mostra a componente aleatório da série, onde estão os “ruídos” da série e tudo o que não foi “ajustado”, representa o que não pode ser explicado pelo modelo, graficamente ocorre um comportamento levemente aleatório, ou seja, sem nenhum comportamento regular evidente.

Decomposição da série no RSTUDIO



A análise gráfica dos resíduos mostra um padrão aparentemente normal para o modelo adotado, onde a função de autocorrelação (ACF) apresentam valores dentro de um limiar aceitável com alguns “picos” fora do intervalo. Os valores residuais em relação aos valores previstos apresentam um comportamento aparentemente normal, conforme ilustração no histograma.

A análise residual do modelo ARIMA (0,0,1) (0,0,2) [12],



O teste de Shapiro-Wilks foi utilizado no RSTUDIO para verificar se o modelo apresenta distribuição normal significativa na componente aleatória, como resultado verificou-se que o modelo apresenta normalidade significativa (ao nível de 5%). Nem sempre é uma tarefa fácil encontrar aleatoriedade dos erros em modelos de séries temporais, pois “picos” na série aumenta a variância dos dados e compromete o seu grau de ajuste.

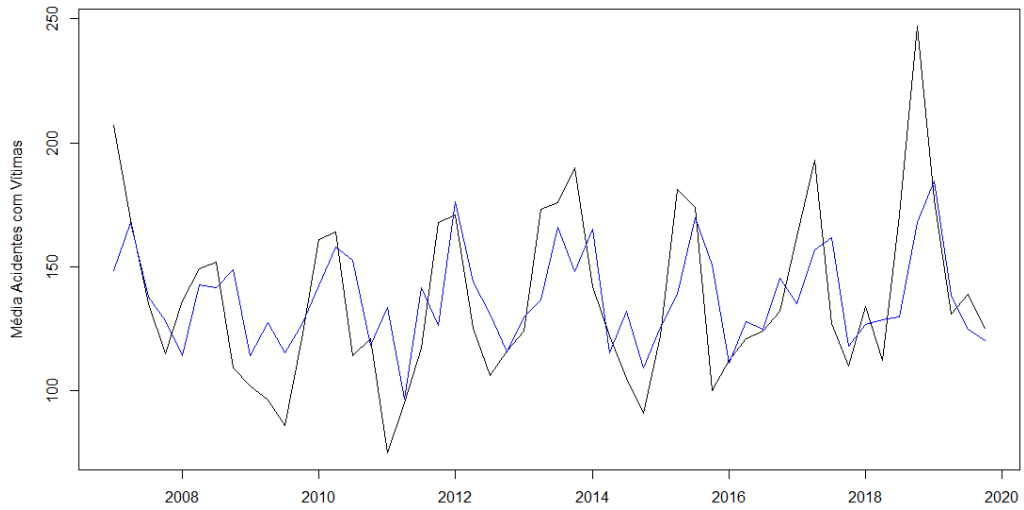
> shapiro.test(componentes$random)

Shapiro-Wilk normality test

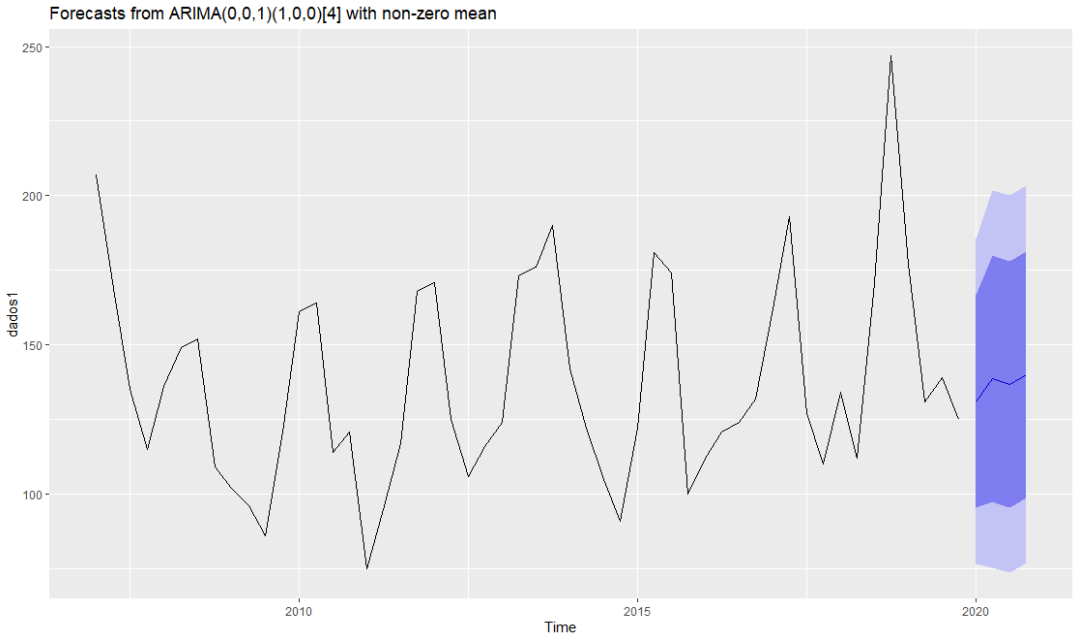
data: componentes$random
W = 0.95405, p-value = 0.05809

O modelo trimestral da média de acidentes com vítimas se ajusta bem aos dados do passado seguindo o comportamento da série histórica de 2007 a 2019.

Valores Observados x Valores Previstos – Ano 2007 a 2019



O gráfico abaixo mostra a previsão do valor médio trimestral de acidentes com vítimas para o ano de 2020, a linha central em azul representa a previsão, a área sombreada com azul escuro representa o intervalo de predição dos valores com intervalo de confiança de 80% e a área sombreada com azul claro representa a predição com 95%.

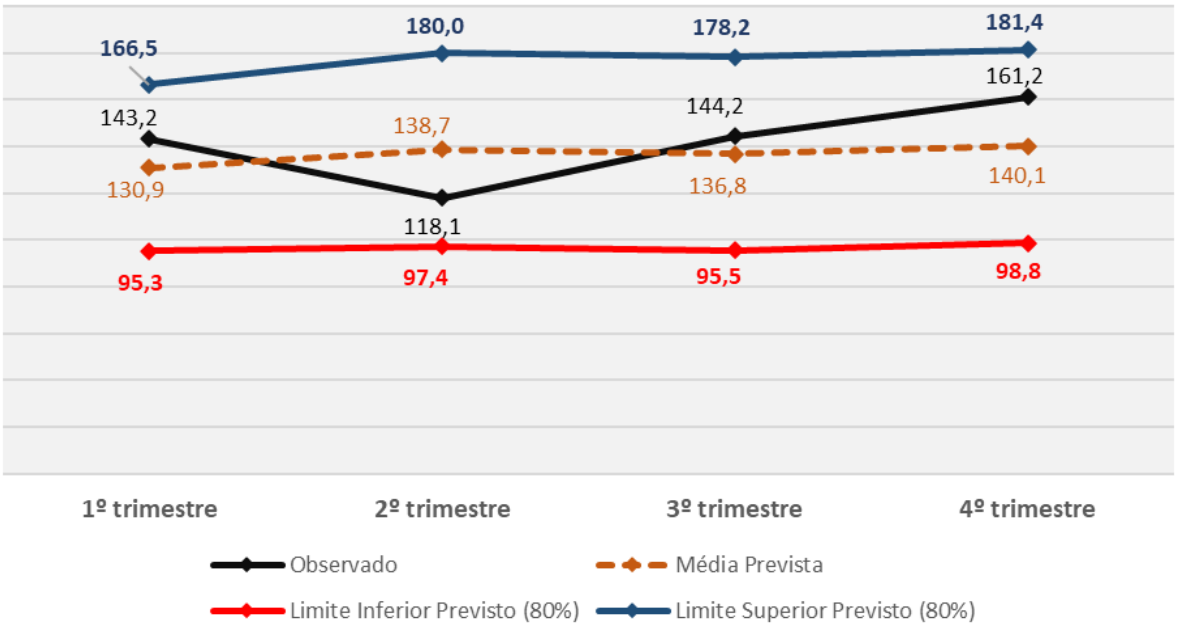


O modelo de séries temporais foi calculado em um período pré-pandemia da COVID-19 (entre 2007 e 2019) com o propósito de eliminar qualquer influência do período extremo da pandemia no Brasil (anos 2020 e 2021) onde ocorreu menos tráfico de veículos de passageiros devido as restrições de locomoção impostas pelos governos estaduais e federais.

As restrições de locomoção no ano de 2020 gerou uma média de acidentes com vítimas abaixo do esperado, principalmente no segundo trimestre do ano, onde o baixo volume de carros nas rodovias federais pode ter sido o principal fator de redução no número de acidentes. Com a política do “fique em casa” houve uma redução significativa na circulação de automóveis nas rodovias federais, principalmente nos dias de feriados nacionais, onde o fluxo de carros costuma ser

mais intenso, em consequência o número de acidentes foi reduzido. Mesmo considerando o comportamento atípico no período de pandemia, o modelo de previsão foi satisfatório para o ano de 2020 dentro do intervalo de predição de 80% e 95%. Comparando a média trimestral de acidentes com à previsão feita pelo modelo no intervalo de predição de 80% ou de 90%, conseguimos prever os valores dentro do intervalo de predição, onde apenas os valor observado no 2º trimestre de 2020 (118,1) ficou abaixo da média prevista (138,7).

Valores Observados x Valores Previstos – ANO 2020



6. Links

Link para o vídeo: youtube.com/...

Link para o repositório: https://github.com/EdsonEST/EdsonEST

REFERÊNCIAS

MORETTIN, Pedro A.;TOLOI, Clélia M. **Análise de Séries Temporais.** São Paulo: Editora Edgard Blucher, 2004.

APÊNDICE

Programação/Scripts

```
### SÉRIE HISTÓRICA DOS ACIDENTES DE TRÂNSITO EM RODOVIAS FEDERAIS #######

bibliotecas = c("forecast", "lmtest", "nortest","dplyr","readr","tidyverse","readxl")
install.packages(bibliotecas)

library(forecast)
library(lmtest)
library(nortest)
library(dplyr)
library(readr)
library(tidyverse)
library(readxl)


### LEITURA DOS DADOS
dados <- read_excel("C:/TCC-PUCMINAS/Acidentes Diários/TS_BASE.xlsx", col_types = c("date", "numeric", "numeric","numeric","numeric",  "numeric", "numeric", "numeric", "numeric"))

### SÉRIE HISTORICA DE JANEIRO DE 2007 A DEZEMBRO DE 2020 - TREINO E TESTE ####
dados <- dados %>% filter(ANO<2021)

### ORDENANDO A SERIE DE DADOS ####
dados <- dados[order(dados$DATA),]

### SÉRIES TEMPORAIS TRIMESTRAL - ACIDENTES COM VÍTIMAS
dados1 <- ts(dados[,5], start = c(2007,1), end = c(2019,4), frequency = 4)
# Obs.: frequency = 1(Ano); frequency = 4(Trimestre); frequency = 12(Mês); frequency = 52(Semanas)

### GRÁFICO DA SÉRIE TEMPORAL
autoplot(dados1, xlab = "Ano", ylab = "Acidentes com Vítimas")

# DECOMPOSIÇÃO DA SÉRIE TEMPORAL
componentes <- stats::decompose(dados1, type=c("additive")) # additive or multiplicative
plot(componentes)

# COMPONENTE ALEATÓRIA DA SÉRIE
shapiro.test(componentes$random)

#TRANSFORMAÇÃO DE BOX COX PARA ESTABILIZAÇÃO DA VARIÂNCIA DA SÉRIE DE DADOS
lambda <- BoxCox.lambda(dados1)
lambda
dados1_bc <- BoxCox(dados1, lambda = lambda)

# NÚMERO DE DIFERENCIAÇÕES NÃO SAZONAIS PARA TRANSFORMA A SÉRIE ESTACIONÁRIA
ndiffs(dados1_bc)  # não apresenta diferenciações

# TESTE COM OS MODELOS ARIMA E SARIMA PARA OBTER O MELHOR MODELO
auto.arima(dados1, trace = TRUE, approximation = FALSE)

# IMPLEMENTAR O ARIMA COM OS PARÂMETROS ENCONTRADOS EM 1
mod <- Arima(dados1, order=c(0,0,1), seasonal = list(order=c(1,0,0),period=4))

# COMPARANDO OS VALORES AJUSTADOS COM A SÉRIE ORIGINAL
plot(dados1, xlab='', ylab='Média Acidentes com Vítimas')
lines(mod$fitted,col='blue', zlab='Predito')

# CHECAR A QUALIDADE DO MODELO
checkresiduals(mod$fitted)
```

```
# MEDIA DE ACIDENTES DIÁRIOS NO TRIMESTRE - 2020 #
dados$trim <- ifelse(dados$MES==1  | dados$MES==2  | dados$MES==3,  1,
                     ifelse(dados$MES==4  | dados$MES==5  | dados$MES==6,  2,
                     ifelse(dados$MES==7  | dados$MES==8  | dados$MES==9 , 3,
                     ifelse(dados$MES==10 | dados$MES==11 | dados$MES==12, 4,0))))
Media <- dados %>% filter(ANO==2020) %>% group_by(ANO,trim) %>%
summarise(media_dia=mean(N_ACIDENTES_VITIMAS,na.rm=TRUE))
as.numeric(Media$media_dia)

# FAZENDO AS PREVISÕES
prev <- forecast(mod, h=4, xlab='', ylab='Média de Acidentes com Vítimas')
autoplot(prev)

# VALORES MÉDIOS PREVISTOS
prev$mean

# VALORES MÉDIOS PREVISTOS DO LIMÍTE INFERIOR
prev$lower

# VALORES MÉDIOS PREVISTOS DO LIMÍTE SUPERIOR
prev$upper
```