# PACDII: QUIZ I

Grupo 7

21 de setembro, 2023

## Elementos do grupo:


- Allan Kardec da Silva Rodrigues, nº 103380
- André Plancha Fernandes, nº 105289
- Diogo Alexandre Alonso de Freitas, nº 104841
- João Francisco Marques Gonçalves da Silva Botas, nº 104782
- Marco Delgado Esperança, nº 110451


## Nota:

Deve efetuar todos os Save com "Save with encoding UTF-8" de modo a manter palavras acentuadas e caracteres especiais**

### Base de dados:condutores.csv

```
# Remover tudo!
rm(list = ls())
# Incluir as libraries de que necessita (instala caso não tenha -->
p_load)
pacman::p_load(VIM, tidyverse, conflicted, skimr, ggplot2, lsr,
lubridate, nycflights13, tidyverse, dplyr)
```

## Questão 1 [5 valores]

### Leitura dos dados condutores.csv.

```
Data <-read.csv("condutores.csv", header=TRUE, stringsAsFactors = T,
sep=",",dec=",",check.names=F,na.strings=c("NA","NÃO
DEFINIDO"),fileEncoding = "utf-8")
```

### Averigue a existência de valores omissos e, caso existam, identifique as respetivas variáveis. Realize a imputação dos valores omissos da variável "Tempo.Condução.Continuada" considerando as variáveis Tipo.Veiculo, Tipo.Serviço e Distrito. Faça upload do sumário de Tempo.Condução.Continuada após a imputação.

```
# Visualizar os nulos em cada uma das colunas
summary(is.na(Data))
```

```
##  Id. Acidente     Datahora          Sexo         Lesões a 30 dias
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:40209     FALSE:40209     FALSE:39514     FALSE:40209
##                                  TRUE :695
##  Licença Condução Teste Alcool    Acções Condutores
##  Mode :logical    Mode :logical   Mode :logical
##  FALSE:39285      FALSE:40172     FALSE:39764
##  TRUE :924        TRUE :37        TRUE :445
##  Inf. Comp. a Acções e Manobras Nomeoutrosfactores Tempo Condução
Continuada
##  Mode :logical                  Mode :logical      Mode :logical
##  FALSE:40099                    FALSE:40209        FALSE:39072
##  TRUE :110                                         TRUE :1137
##  Acessórios Condutores Categoria Veículos Tipo Veiculo    Tipo Serviço
##  Mode :logical         Mode :logical      Mode :logical   Mode
:logical
##  FALSE:39090           FALSE:40155        FALSE:29756     FALSE:40106
##  TRUE :1119            TRUE :54           TRUE :10453     TRUE :103
##  Veiculo Especial Ano matricula   Inspecção Periódica Certificado Adr
##  Mode :logical    Mode :logical   Mode :logical       Mode :logical
##  FALSE:232        FALSE:38065     FALSE:40014         FALSE:19
##  TRUE :39977      TRUE :2144      TRUE :195           TRUE :40190
##  Carga Lotação     Pneus          Seguros          Distrito
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:39719     FALSE:39773     FALSE:40116     FALSE:40209
##  TRUE :490       TRUE :436       TRUE :93
##   Concelho       Condutor Gr.Etario(<=5) SUM Condutor Gr.Etario(6-9)
SUM
##  Mode :logical   Mode :logical               Mode :logical
##  FALSE:40209     FALSE:40209                 FALSE:40209
##
##  Condutor Gr.Etario(10-14) SUM Condutor Gr.Etario(15-17) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(18-20) SUM Condutor Gr.Etario(21-24) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(25-29) SUM Condutor Gr.Etario(30-34) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(35-39) SUM Condutor Gr.Etario(40-44) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(45-49) SUM Condutor Gr.Etario(50-54) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
```

```
##
##  Condutor Gr.Etario(55-59) SUM Condutor Gr.Etario(65-69) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(70-74) SUM Condutor Gr.Etario(>=75) SUM
##  Mode :logical                 Mode :logical
##  FALSE:40209                   FALSE:40209
##
##  Condutor Gr.Etario(Não Def.) SUM
##  Mode :logical
##  FALSE:40209
##
```

```
# Colunas com omissos
(colunas_com_nulos <- names(Data)[colSums(is.na(Data)) > 0])
```

```
##  [1] "Sexo"                           "Licença Condução"
##  [3] "Teste Alcool"                   "Acções Condutores"
##  [5] "Inf. Comp. a Acções e Manobras" "Tempo Condução Continuada"
##  [7] "Acessórios Condutores"          "Categoria Veículos"
##  [9] "Tipo Veiculo"                   "Tipo Serviço"
## [11] "Veiculo Especial"               "Ano matricula"
## [13] "Inspecção Periódica"            "Certificado Adr"
## [15] "Carga Lotação"                  "Pneus"
## [17] "Seguros"
```

```
# Imputação dos dados sei lá
k_escolha = round(sqrt(nrow(Data)))
dfn <- kNN(Data[c(10, 13, 14, 22)], variable = c("Tempo Condução
Continuada"), k = k_escolha)
Data$`Tempo Condução Continuada` <- dfn$`Tempo Condução Continuada`

# Sumário da variável Tempo.Condução.Continuada após a imputação
summary(Data$`Tempo Condução Continuada`)
```

```
##  De 1 a 3 horas  De 3 a 5 horas        Ignorada Mais de 5 horas Menos
de 1 hora
##           2338             102           16224             123
21422
```

- De 1 a 3 horas: 2338
- De 3 a 5 horas: 102

- Ignorada: 16224
- Mais de 5 horas: 123
- Menos de 1 hora: 21422

## Questão 2 [5 valores] Efetue a análise descritiva da variável Ano.matricula, de modo a completar:

```
#Efetue a análise descritiva da variável Ano.matricula, de modo a
completar:

#O veículo mais antigo é do ano _1914_ e a percentagem (cumulativa) de
veículos com ano de matrícula inferior ou igual a 2007 é
_51.5_%(percentagem com uma casa decimal); os veículos com ano de
matrícula inferior a _1977_ (indique o ano) podem ser considerados
outliers. Considerando o ano de 2020 para o cálculo da idade dos
veículos, observa-se que 25% dos veículos têm menos de _5_ anos.
```

*Resolução 1*

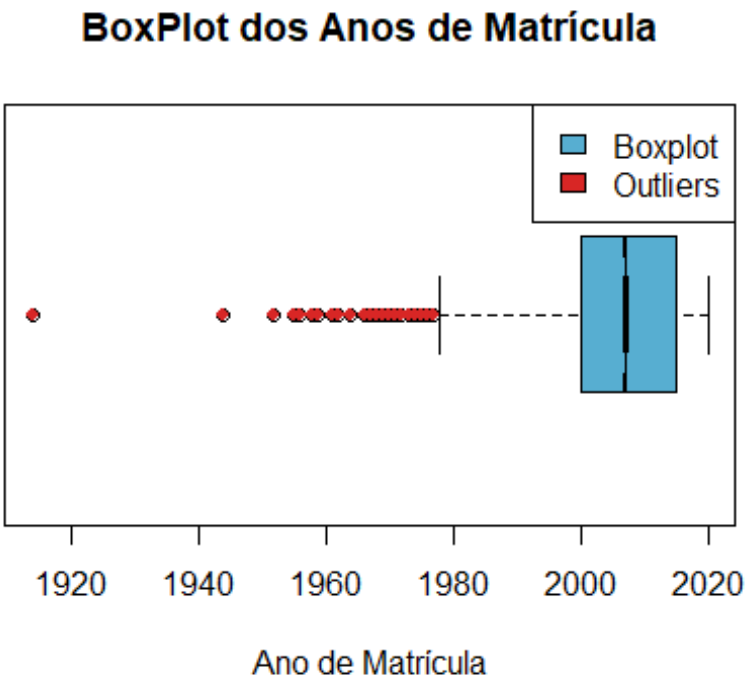
```
min(Data$`Ano matricula`, na.rm = TRUE)

## [1] 1914
```

*Resolução 2*

```
denominador <- 1/(nrow(Data) - nrow(dplyr::filter(Data, is.na(`Ano
matricula`))))
Data %>% dplyr::filter(`Ano matricula` <= 2007) %>% nrow() %*%
denominador %*% 100 %>% round(1)

##      [,1]
## [1,] 51.5
```

*Resolução 3*

```
bp <- boxplot(Data$`Ano matricula`, horizontal = TRUE, outline = TRUE,
col = "#57aed1", border = "black", notch = TRUE)
title(main = "BoxPlot dos Anos de Matrícula", xlab = "Ano de Matrícula")
points(bp$out, bp$group, col = "#d82625", pch = 19, cex = .8)
legend("topright", legend = c("Boxplot", "Outliers"), fill = c("#57aed1",
"#d82625"))
```




```
max(bp$out)
```

```
## [1] 1977
```

```
floor(lower_t <- quantile(Data$`Ano matricula`, probs = 0.25, na.rm = T)
- 1.5*IQR(Data$`Ano matricula`, na.rm = T))
```

```
##  25%
## 1977
```

*Resolução 4*

```
quantile(Data$`Ano matricula`,prob=.75, na.rm = T)
```

```
##  75%
## 2015
```

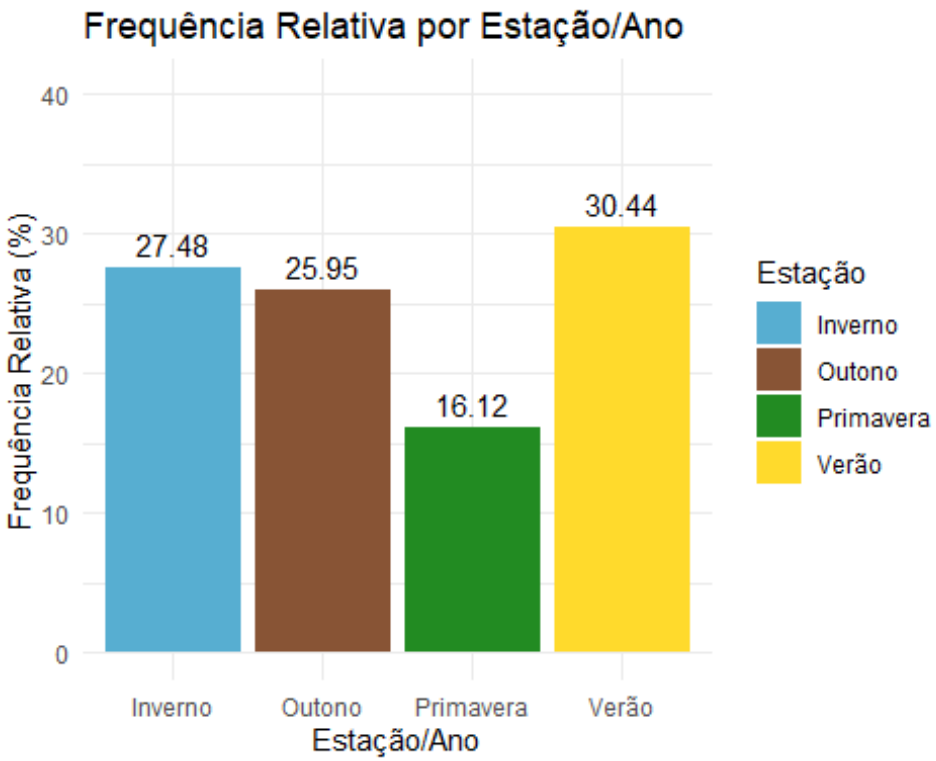
O veículo mais antigo é do ano **1914** e a percentagem (cumulativa) de veículos com ano de matrícula inferior ou igual a 2007 é **51.5%**(percentagem com uma casa decimal); os veículos com ano de matrícula inferior a **1977** (indique o ano) podem ser considerados outliers. Considerando o ano de 2020 para o cálculo da idade dos veículos, observa-se que 25% dos veículos têm menos de **5** anos.

## Questão 3 [5 valores] Crie a variável nominal Estação_Ano com as classes "Inverno", "Primavera", "Verão" e "Outono" e faça upload do correspondente gráfico de barras apresentando as frequências relativas.

- Inverno: 22/12 -> 20/03

- Primavera: 20/03 -> 21/06
- Verão: 21/06 -> 23/09
- Outono: 23/09 -> 22/12

```
Data <-  Data %>%
  mutate(Datahora = as.Date(Datahora, format = "%Y:%m:%d")) %>%
  mutate(Estação_Ano = case_when(
    (month(Datahora) == 12 & day(Datahora) >= 22) |
    (month(Datahora) == 1 | month(Datahora) == 2) |
    (month(Datahora) == 3 & day(Datahora) < 20) ~ "Inverno",

    (month(Datahora) >= 3 & month(Datahora) <= 5) |
    (month(Datahora) == 6 & day(Datahora) < 21) ~ "Primavera",

    (month(Datahora) >= 6 & month(Datahora) <= 8) |
    (month(Datahora) == 9 & day(Datahora) < 23) ~ "Verão",

    (month(Datahora) >= 9 & month(Datahora) <= 11) |
    (month(Datahora) == 12 & day(Datahora) < 22) ~ "Outono",

    TRUE ~ "Inverno"
  ))

Data$Estação_Ano <- as.factor(Data$Estação_Ano)
freq_table <- table(Data$Estação_Ano)
freq_relativa_estacoes <- round((freq_table / sum(freq_table)) * 100, 2)

# Inverno - azul, verão- amarelo, outono - castanho, primavera - verde
season_colors <- c("#57aed1", "#885435", "#228b22", "#ffda2c")

plot_data <- data.frame(freq_relativa_estacoes)
colnames(plot_data)[1] <- "Estação"

# Create the ggplot barplot
ggplot(plot_data, aes(x = Estação, y = Freq, fill = Estação)) +
   geom_bar(stat = "identity") +
   ylim(0, max(freq_relativa_estacoes) + 10) +
   geom_text(aes(label = Freq), vjust = -0.5, size = 4) +
   scale_fill_manual(values = season_colors) +
   labs(title = "Frequência Relativa por Estação/Ano") +
   xlab("Estação/Ano") +
   ylab("Frequência Relativa (%)") +
   theme_minimal()
```




**Questão 4 [5 valores] Com base no cálculo de uma medida de associação, estude a relação existente entre a variável Lesoes.a.30.dias e cada uma das seguintes variáveis: Estacao_ano e Ano.matricula. Faça o upload dos valores das medidas de associação e dos gráficos que ilustram as mesmas associações.**

```
# Visualização das colunas pretendidas
summary(Data[c(4,42,16)])
```

```
##      Lesões a 30 dias    Estação_Ano    Ano matricula
##  Ferido grave: 1162   Inverno  :11050   Min.   :1914
##  Ferido leve :20514   Outono   :10436   1st Qu.:2000
##  Ileso       :18215   Primavera: 6482   Median :2007
##  Morto       :  318   Verão    :12241   Mean   :2007
##                                         3rd Qu.:2015
##                                         Max.   :2020
##                                         NA's   :2144
```

```
# Associação/Relação entre variáveis nominais
(VC <- cramersV(Data$`Lesões a 30 dias`, Data$Estação_Ano))
```

```
## [1] 0.03244379
```

```
# Associação/Relação entre uma variável nominal e métrica
anova_eta <- aov(Data$`Ano matricula` ~ Data$`Lesões a 30 dias`) %>%
etaSquared()
anova_eta[1,1]
```

```
## [1] 0.002390516
```

```
eta <- sqrt(anova_eta[1,1])
eta
```

```
## [1] 0.0488929
```

- V de Cramer entre “Lesões a 30 dias” e “Estação_Ano”: **0.0324438**;
- Eta entre “Lesões a 30 dias” e “Ano Matrícula”: **0.0488929**;

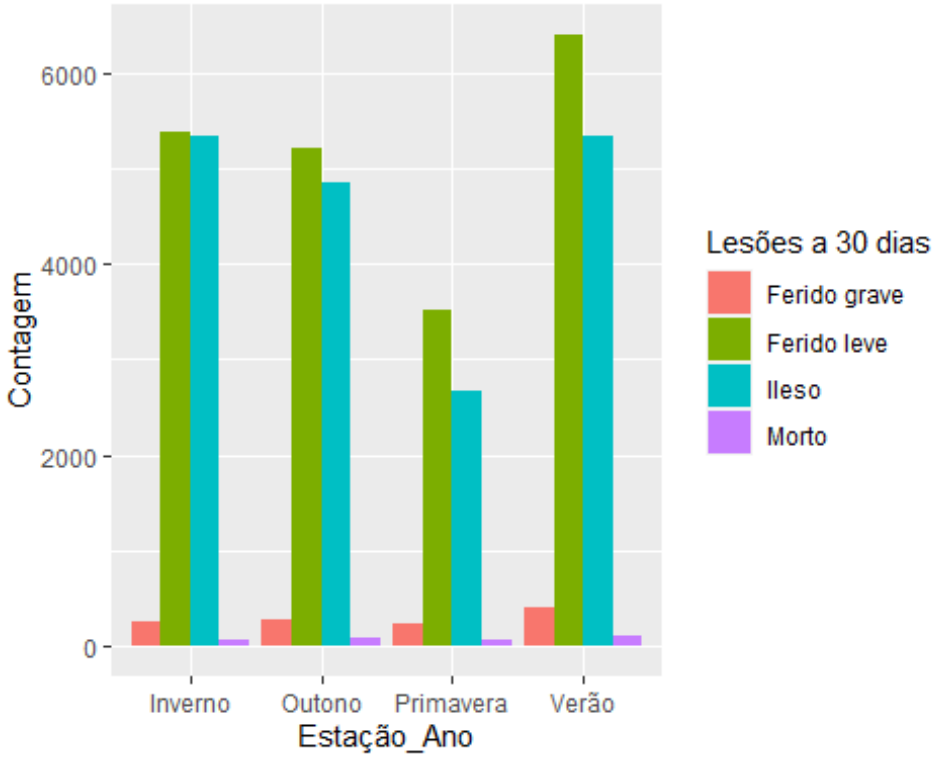
```
# apenas visualização
head(Data[, c('Lesões a 30 dias', 'Estação_Ano', 'Ano matricula')])
```

```
##   Lesões a 30 dias Estação_Ano Ano matricula
## 1      Ferido leve     Inverno          2014
## 2            Ileso     Inverno          2016
## 3            Ileso     Inverno          2015
## 4      Ferido leve     Inverno          2018
## 5      Ferido leve     Inverno          1996
## 6            Ileso     Inverno          1999
```

```
dados_preparados <- Data %>%
  group_by(Estação_Ano, `Lesões a 30 dias`) %>%
  summarize(Contagem = n())
```

```
## `summarise()` has grouped output by 'Estação_Ano'. You can override
using the
## `.groups` argument.
```

```
dados_preparados %>% ggplot(aes(x = Estação_Ano,y=Contagem, fill =
`Lesões a 30 dias`)) + geom_bar(stat = "identity", position = "dodge")
```




```
dados_preparados <- Data %>%
  group_by(`Ano matricula`, `Lesões a 30 dias`) %>%
  summarize(Contagem = n())
```

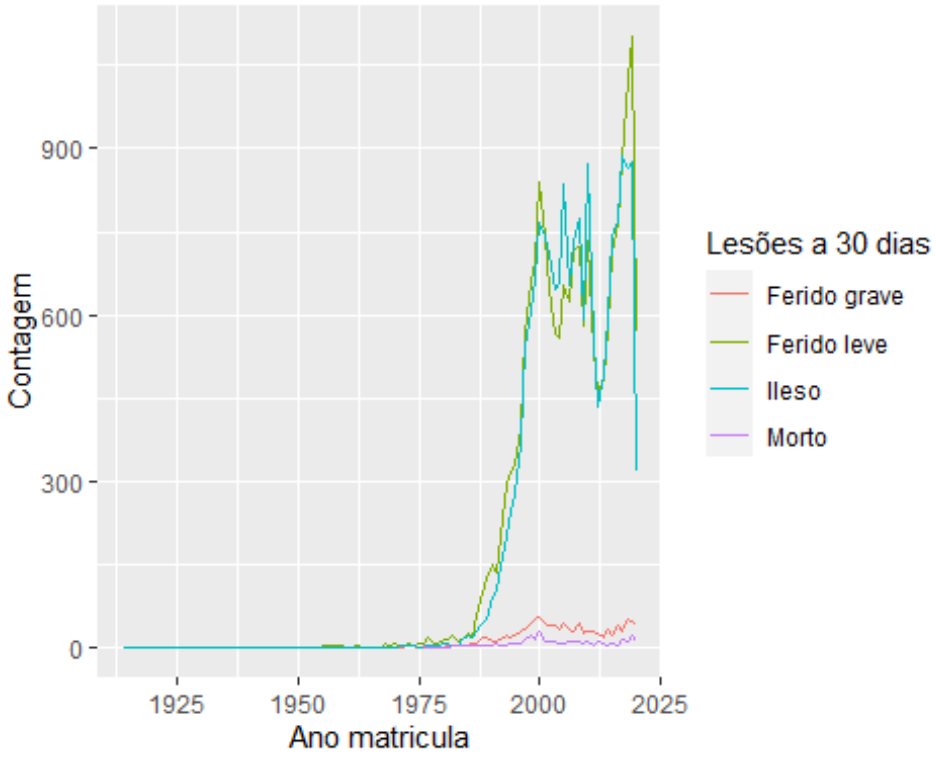
```
## `summarise()` has grouped output by 'Ano matricula'. You can override
using the
## `.groups` argument.
```

```
dados_preparados <- na.omit(dados_preparados)
dados_preparados
```

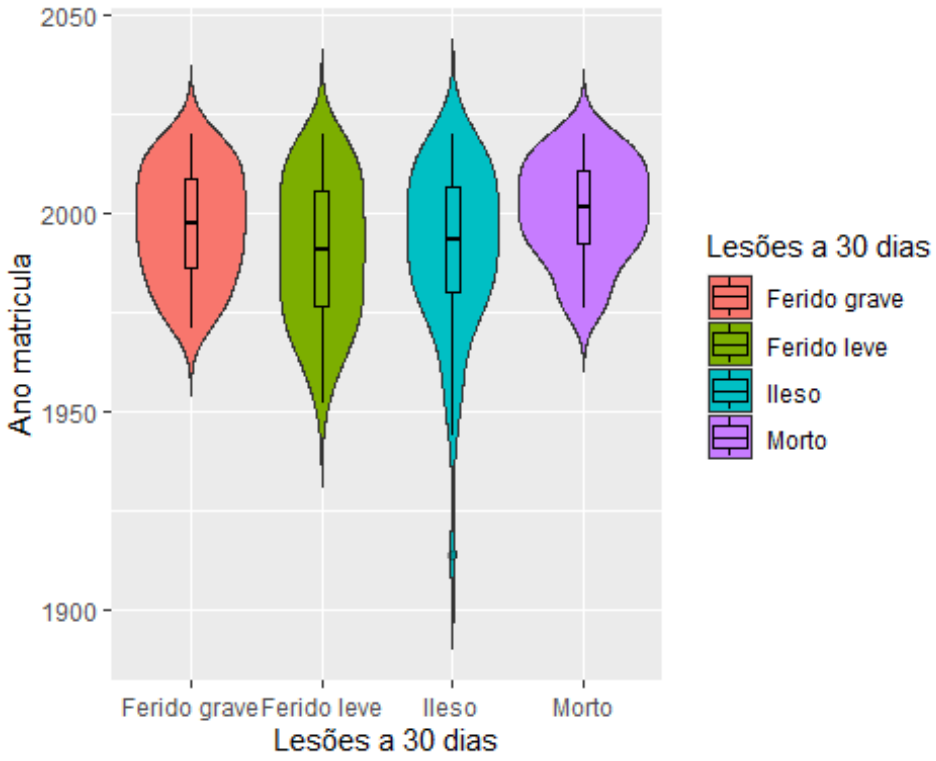
```
## # A tibble: 197 × 3
## # Groups:   Ano matricula [65]
##    `Ano matricula` `Lesões a 30 dias` Contagem
##              <int> <fct>                  <int>
##  1            1914 Ileso                      1
##  2            1944 Ileso                      1
##  3            1952 Ferido leve                1
##  4            1955 Ileso                      1
##  5            1956 Ileso                      1
##  6            1958 Ferido leve                2
##  7            1959 Ileso                      1
##  8            1961 Ferido leve                1
##  9            1962 Ferido leve                2
## 10            1964 Ferido leve                1
## # ℹ 187 more rows
```

```
dados_preparados %>%
  ggplot() + geom_line(aes(x = `Ano matricula`, y = Contagem, group =
`Lesões a 30 dias`, col = `Lesões a 30 dias`), size = .7) + ylim(0,
max(dados_preparados$Contagem))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2
3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was
## generated.
```




```
ggplot(dados_preparados, aes(x = `Lesões a 30 dias`, y = `Ano matricula`,
fill = `Lesões a 30 dias`)) +
  geom_violin(width=1,trim=FALSE) +
  geom_boxplot(width=0.1, color="black",alpha=0.2)
```




**Tarefa final: Submeta, no Moddle, um ficheiro pdf resultado da compilação do TEMPLATE_QUIZ1.**

**Caso os resultados apresentados não sejam coerentes com as respostas dadas, a classificação será penalizada.**