```
# Aluno: Carlos Eduardo Fontaneli
# RA: 769949
library(tidyverse)
library(sf)
library(ggplot2)
library(rpart)
library(rpart.plot)

dados <- read_csv('/home/fonta42/Desktop/ICDuR/Datasets/archive/wine.csv')

# Questão 1
# Análise visual
ggplot(data = dados, aes(x = `free sulfur dioxide`, y = `total sulfur
dioxide`)) +
  geom_point() +
  geom_smooth()

# Coeficiente de Pearson ~= 0.67 indicando uma correlação positiva moderada
cor(dados$`free sulfur dioxide`, dados$`total sulfur dioxide`)


# Questão 2
#a)
# Objetivo verificar como os níveis de açúcar residual  e álcool se afetam a
# classificação da qualidade do vinho.

# b)
# Para isso foram utilizados os atributos 'alcohol', 'residual sugar' e
'quality'

# c)
ggplot(data = dados) +
  geom_col(aes(
    x = `residual sugar`,
    y = alcohol,
    group = quality,
    colour = quality,
    fill = quality),
    position = position_dodge()) +
  ggtitle("Qualidade do vinho em relação a níveis de açúcar e álcool", ) +
  xlab("Açúcar residual") +
  ylab("Nível de álcool")

# d)

# A partir da visualização é possível concluir que a maioria dos vinhos
classi-
# ficados como bons possuem um nível de açúcar relativamente baixo e um teor
# álcoolico superior aos vinhos classificados como ruins, salvo certas
exceções.

# Questão 3

# a)
# divisao em conjunto de treino e de teste
```

```
prepare_hold_out <- function(tbl, training_perc) {
  # misturando as observacoes
  tbl_mixed <- tbl[sample(1:nrow(tbl)), ]
  nrow <- nrow(tbl_mixed)

  nrow_train <- ceiling(training_perc * nrow)
  data_trn <- tbl_mixed[1:nrow_train, ]
  data_tst <- tbl_mixed[(1 + nrow_train):(nrow), ]

  # retorna como uma lista nomeada
  list(training = data_trn, test = data_tst)
}

# Configurando uma seed para possibilitar reprodução exata dos resultados
set.seed(12345)

# Dados divididos em conjunto de treino(80%) e teste(20%)
dados_misturados <- prepare_hold_out(dados, 0.8)

# Construindo a árvore
tree <- rpart(quality ~ `fixed acidity`+ `volatile acidity`+ `citric acid`+
`residual sugar`+ chlorides + `free sulfur dioxide`+ `total sulfur dioxide`+
density+ pH+ sulphates+ alcohol,data = dados_misturados$training)

# b)
rpart.plot(tree)

# c)
# O atributo escolhido para o nó raiz foi 'alcohol'

# d)
# predições do modelo
classes_preditas <- predict(tree, dados_misturados$test, type = "class")

# matriz de confusão
confusion_matrix <- table(dados_misturados$test$quality, classes_preditas)
confusion_matrix

# i. O número de classificações corretas para “good”
confusion_matrix[2,2]

# ii. O número de classificações corretas para “bad”
confusion_matrix[1,1]

# iii. O número de falsos positivos
confusion_matrix[1,2]

# iv. O número de falsos negativos
confusion_matrix[2,1]
```