Working with Data Frames in R

Prachi Sardana

## Question 1

### To load data csv file into a dataframe df.wines

df.wines <- read.csv(url("https://s3.us-east-2.amazonaws.com/artificium.us/datasets/whitewines.csv"),header = TRUE,stringsAsFactors = FALSE)

## Question 2

### Inspect the dataframe and determine it’s structure

head(df.wines,5)

## fixed.acidity volatile.acidity citric.acid residual.sugar chlorides  
## 1 7.0 0.27 0.36 20.7 0.045  
## 2 6.3 0.30 0.34 1.6 0.049  
## 3 8.1 0.28 0.40 6.9 0.050  
## 4 7.2 0.23 0.32 8.5 0.058  
## 5 7.2 0.23 0.32 8.5 0.058  
## free.sulfur.dioxide total.sulfur.dioxide density pH sulphates alcohol  
## 1 45 170 1.0010 3.00 0.45 8.8  
## 2 14 132 0.9940 3.30 0.49 9.5  
## 3 30 97 0.9951 3.26 0.44 10.1  
## 4 47 186 0.9956 3.19 0.40 9.9  
## 5 47 186 0.9956 3.19 0.40 9.9  
## quality  
## 1 6  
## 2 6  
## 3 6  
## 4 6  
## 5 6

str(df.wines)

## 'data.frame': 4898 obs. of 12 variables:  
## $ fixed.acidity : num 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...  
## $ volatile.acidity : num 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...  
## $ citric.acid : num 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...  
## $ residual.sugar : num 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...  
## $ chlorides : num 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...  
## $ free.sulfur.dioxide : num 45 14 30 47 47 30 30 45 14 28 ...  
## $ total.sulfur.dioxide: num 170 132 97 186 186 97 136 170 132 129 ...  
## $ density : num 1.001 0.994 0.995 0.996 0.996 ...  
## $ pH : num 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...  
## $ sulphates : num 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...  
## $ alcohol : num 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...  
## $ quality : int 6 6 6 6 6 6 6 6 6 6 ...

## Question 3

### How many wines have a residual sugar level above 0.5 and an alcohol level above 7?

# Subseting df.wines to determine how many wines have residual sugal above 0.5 and alcohol level above 7  
filter\_df.wines <- subset(df.wines, df.wines$residual.sugar > 0.5 & df.wines$alcohol > 7)  
  
# To count the number of rows   
count\_rows <- nrow(filter\_df.wines)  
  
# Using cat function to print the count of wines  
cat("Number of wines with Residual Sugar > 0.5 and Alcohol > 7:", count\_rows, "\n")

## Number of wines with Residual Sugar > 0.5 and Alcohol > 7: 4898

## Question 4

### How many wines have an alcohol content between 9.5 and 11.5 (exclusive) and a quality rating below 7?

filter\_df.wines <- subset(df.wines, df.wines$alcohol >= 9.5 & df.wines$alcohol <= 11.5 & df.wines$quality < 7)  
  
count\_rows <- nrow(filter\_df.wines)  
  
# Using cat function to print the count of wines  
cat("Number of wines with alcohol content between 9.5 and 11.5 (exclusive) and a quality rating below 7:", count\_rows, "\n")

## Number of wines with alcohol content between 9.5 and 11.5 (exclusive) and a quality rating below 7: 2201

## Question 5

### Are there any wines with an alcohol content above 14?

filter\_df.wines <- subset(df.wines, df.wines$alcohol > 14)  
count\_rows <- nrow(filter\_df.wines)  
  
# Using cat function to print the count of wines  
cat("Number of wines with alcohol content above 14:", count\_rows, "\n")

## Number of wines with alcohol content above 14: 2

## Question 6

### What are the median and mean alcohol content of all wines?

mean\_alcohol <- mean(df.wines$alcohol)  
print(mean\_alcohol)

## [1] 10.51427

median\_alcohol <- median(df.wines$alcohol)  
print(median\_alcohol)

## [1] 10.4

## Question 7

### Add a new column to the data frame df.wines called swill\_index that is calculated by dividing the alcohol content by the the quality and multiplying it by the residual sugar content.

df.wines$swill\_index <- (df.wines$alcohol/df.wines$quality)\* df.wines$residual.sugar

## Question 8

### Add a new column to the data frame df.wines called alcohol.z that calculates the z-score for alcohol. The z-score for a feature is calculated as the number of standard deviations that the value is from the mean. So, calculate the mean and the standard deviation for alcohol and then set the value of the alcohol.z column to the value to: mean(alcohol) - alcohol] / sd(alcohol).

mean\_alcohol <- mean(df.wines$alcohol)  
sd\_alcohol <- sd(df.wines$alcohol)  
  
# Calculating the z score and adding the new coloumn to the dataframe  
df.wines$alcohol.z <- (mean\_alcohol - df.wines$alcohol)/sd\_alcohol