Assignment 02 Loading, saving and describing data

Erica Peng

2023-09-13

## Section 1 Describe the dataset you are using

(what is this data measuring? how was it collected? what kinds of research questions are you hoping to use it to answer?) and in terms of its format (what type of file is it saved in? what if it is in a flat file, is it fixed width or delimited? if it is delimited, what is the delimiter? if it is binary, what is the program that would normally be used to open it?).

#Resoucre comes from: <https://fivethirtyeight.com/features/dear-mona-followup-where-do-people-drink-the-most-beer-wine-and-spirits/> The data is measuring the average of each alcohol and bervage of serving sizes per person, which was been collected by World Health Organization (WHO). The file has been saved in a flat file with a “Fixed-Width Format”, which we can see in our dataset, each column of data has a specific width, and each record in the file represents data for a specific location, or consumption. For example, the country of Albania recorded the beer consumption of 88 servings; 132 servings of spirit consumption, and 54 servings of wine. This means that the data values are aligned at specific positions within each line and row.

#this makes a new data.frame called text\_tbl with three columns, Names and Description  
text\_tbl <- data.frame(Names = c("beer\_servings","spirit\_servings","wine\_servings"), Description = c("The data shows the average serving sizes of beer per person","The data shows the average serving sizes of spirit per person","The data shows the average serving sizes of wine per person")  
)  
  
  
#prints the table  
text\_tbl

## Names Description  
## 1 beer\_servings The data shows the average serving sizes of beer per person  
## 2 spirit\_servings The data shows the average serving sizes of spirit per person  
## 3 wine\_servings The data shows the average serving sizes of wine per person

## Section 2 Reading the data into R

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/alcohol-consumption/drinks.csv"  
data <- read\_csv(url)

## Rows: 193 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): country  
## dbl (4): beer\_servings, spirit\_servings, wine\_servings, total\_litres\_of\_pure...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Section 3 Clean the data

# Load the dplyr library if not already loaded  
# install.packages("dplyr") # Uncomment and run if dplyr is not installed  
library(dplyr)  
library(tidyverse)  
  
# Calculate the average of a specific column in a dataframe  
average\_value\_b <- mean(data$beer\_servings, na.rm = TRUE)  
average\_value\_b

## [1] 106.1606

average\_value\_s <- mean(data$spirit\_servings, na.rm = TRUE)  
average\_value\_s

## [1] 80.99482

# Keep rows where both 'beed\_serving' is greater than 106 and 'spirit\_serving' is more than 80  
filtered\_data <- filter(data, beer\_servings > 106, spirit\_servings > 80)  
filtered\_data

## # A tibble: 52 × 5  
## country beer\_servings spirit\_servings wine\_servings total\_litres\_of\_pure\_a…¹  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Andorra 245 138 312 12.4  
## 2 Bahamas 122 176 51 6.3  
## 3 Barbados 143 173 36 6.3  
## 4 Belarus 142 373 42 14.4  
## 5 Belgium 295 84 212 10.5  
## 6 Belize 263 114 8 6.8  
## 7 Brazil 245 145 16 7.2  
## 8 Bulgaria 231 252 94 10.3  
## 9 Canada 240 122 100 8.2  
## 10 Chile 130 124 172 7.6  
## # ℹ 42 more rows  
## # ℹ abbreviated name: ¹​total\_litres\_of\_pure\_alcohol

## Section 4 Characteristics of the data

# Write inline code

This data set has 193 country and 5 filtered\_data.

## Section 5 Subset and Summary (Subset your dataset)

# picking three columns to use summary function:  
data\_pick3 <- select(data, beer\_servings, spirit\_servings, wine\_servings)  
  
data\_pick3

## # A tibble: 193 × 3  
## beer\_servings spirit\_servings wine\_servings  
## <dbl> <dbl> <dbl>  
## 1 0 0 0  
## 2 89 132 54  
## 3 25 0 14  
## 4 245 138 312  
## 5 217 57 45  
## 6 102 128 45  
## 7 193 25 221  
## 8 21 179 11  
## 9 261 72 212  
## 10 279 75 191  
## # ℹ 183 more rows

## Section 5 Subset and Summary (Produce a summary of the subset)

#creates the summary  
Summarytable<-summary(data\_pick3)   
  
#prints the summary in your output  
Summarytable

## beer\_servings spirit\_servings wine\_servings   
## Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 20.0 1st Qu.: 4.00 1st Qu.: 1.00   
## Median : 76.0 Median : 56.00 Median : 8.00   
## Mean :106.2 Mean : 80.99 Mean : 49.45   
## 3rd Qu.:188.0 3rd Qu.:128.00 3rd Qu.: 59.00   
## Max. :376.0 Max. :438.00 Max. :370.00

#or you can do this to print:  
print(Summarytable)

## beer\_servings spirit\_servings wine\_servings   
## Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 20.0 1st Qu.: 4.00 1st Qu.: 1.00   
## Median : 76.0 Median : 56.00 Median : 8.00   
## Mean :106.2 Mean : 80.99 Mean : 49.45   
## 3rd Qu.:188.0 3rd Qu.:128.00 3rd Qu.: 59.00   
## Max. :376.0 Max. :438.00 Max. :370.00

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.