

Using R for data analysis

Boerhaave Nascholing

June 11th, 2021

Important note

The primary goal of this self-study assignment is to write an R Markdown document containing **the code** which calculates the answers to the questions below. Using the course and online materials is permitted.

The steps you need to take:

1. Create a new R Markdown file.
2. Develop the code with your answers in the R Markdown file. Put each question in a separate section, with the question number in the section title, just the question number would suffice.
3. Use **Knit** button regularly to check that your code does not produce errors. A knittable R Markdown file is the final result.

R Markdown header:

Replace the header of your R Markdown file with what is shown below and put your name as **author**:

```
---
title: "Using R for data analysis"
subtitle: "Boerhaave Nascholing"
date : "June 11th, 2021"
author: "name"
output: html_document
---
```

Diamonds dataset

For this self-study you will be using the **diamonds** dataset. It contains various attributes of sold diamonds (see also `?diamonds`). The dataset comes with the **tidyverse** package. Once you load the **tidyverse** library you'll have access to the dataset in the **diamonds** variable. Make sure you put `library(tidyverse)` in the R chunk at the top of your R Markdown file.

```
library( tidyverse )
diamonds
```

```
# A tibble: 53,940 x 10
```

	carat	cut	color	clarity	depth	table	price	x	y	z
	<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47

```

 8  0.26 Very Good H      SI1      61.9    55   337  4.07  4.11  2.53
 9  0.22 Fair          E       VS2      65.1    61   337  3.87  3.78  2.49
10  0.23 Very Good H      VS1      59.4    61   338  4      4.05  2.39
# ... with 53,930 more rows

```

Diamonds dataset tibble

Each row of the `diamonds` tibble describes one sold diamond. There are the following variables (columns):

- **price**: Price (in US dollars).
- **carat**: Weight (in carat units: 1 carat = 0.2g).
- **cut**: Quality of the cut (**Fair**, **Good**, **Very Good**, **Premium**, **Ideal**).
- **color**: Color, from J (worst) to D (best).
- **clarity**: How clear the diamond is (**I1** (worst), **SI2**, **SI1**, **VS2**, **VS1**, **VVS2**, **VVS1**, **IF** (best)).
- **x**, **y**, **z**: Length, width, depth (each in mm).
- **depth**: Total depth percentage = $z / \text{mean}(x, y) = 2 * z / (x + y)$.
- **table**: Width of top of diamond relative to widest point.

Questions

Q1. [0.5p] Show the type/class of the `diamonds` table. [0.5p] Show the type of the column `clarity`.

Q2. [1p] Show the structure of the `diamonds` table.

Q3. [1p] Print the rows 12-15 (hint: combine `head` and `tail`).

Q4. [1p] Calculate the mean of the `price` column.

Q5. [1p] Give the **number** of levels of the factor in the `clarity` column.

Q6. [3p] Make a **list** with two elements calculated as follows from the `diamonds` table. Name the first list element `medianDepth` and set it to the median diamond depth. Name the second list element `clarities` and set it to the **vector of** levels of the column `clarity`.

Q7. Frequencies and cross table.

a) [1p] Count all the combinations of the value pairs in columns `cut` and `clarity`.

b) [2p] Print a **cross table** of `cut` and `clarity`, with `cut` categories given in columns.

Q8. [3p] Group the `diamonds` table by `color`. In each group calculate min, max, median and mean `price`.

Q9. Diamond volume in a scatter plot.

a) [1p] Add a new column `volume` representing diamond's volume in cubic millimetres given the dimensions `x`, `y` and `z`. Store the tibble with the added column in the variable `diamonds_volume`.

b) [2p] Use `diamonds_volume` tibble to plot the `volume` (vertical axis) against the `price` (horizontal axis) in a scatterplot. Colour points by `clarity`. Make points 0.5 transparent.

c) [1p] Replot the scatterplot in Q9.b but now with rows where $volume > 0$ and $volume \leq 600$.

Q10. Read/write CSV files.

a) [1p] Write the table `diamonds_volume` to a *comma-separated values* (CSV) file. Give the following name to the file: `diamonds_volume.csv`

b) [1p] Read the file `diamonds_volume.csv` back into variable `d` and show it.