Using R for data analysis SSA

Boerhaave Nascholing Leiden

May 17th, 2022

Important note

The primary goal of the self-study assignment (SSA) is to write an R Markdown document containing **R code** answering the questions below. Using the course and other on-line 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.
3. *When Knitting is possible:* Use `Knit` button regularly to check that your code generates the html report without any errors.

Diamonds dataset

You will analyse the `diamonds` dataset which contains various attributes of sold diamonds (see also `?diamonds`). The dataset comes with the `tidyverse` package. After you load the `tidyverse` library you will 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
```

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

- `price`: Price in US dollars.
- `carat`: Weight of the diamond (in carat units: 1 carat = 0.2g).
- `cut`: Quality of the cut (`Fair`, `Good`, `Very Good`, `Premium`, `Ideal`).
- `color`: Diamond colour, 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 / 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 `cut`.

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

Q3. [1p] Print the last 7 rows of the `diamonds` table.

Q4. [1p] Calculate the median (number!) of the `depth` column from the `diamonds` table.

Q5. [1p] Calculate and print the number of levels of the factor in the `cut` column.

Q6. [3p] Make a `list` with two elements calculated as follows from the `diamonds` table. Name the first list element `maxPrice` and set it to the maximum diamond price. Name the second list element `colors` and set it to the levels of the column `color`. Obviously, the first element should be a number and the second a character vector.

Q7. Frequencies and cross table.

a) [1p] Count all the combinations of the value pairs in columns `cut` and `color`. This table should have three columns: `cut`, `color` and the number of occurrences.

b) [2p] Print a crosstable of `cut` and `color`, with `cut` categories given in columns.

Q8. [3p] Group the `diamonds` table by `cut`. Summarise the mean `price` and the mean `carat` in each group.

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 a new variable `diamonds_volume`.

b) [2p] Use the data from `diamonds_volume` variable and plot the `volume` (vertical axis) against the `carat` (horizontal axis) in a scatterplot. Colour points by `cut`. Make points transparent (0.5).

c) [1p] Replot the scatterplot in Q9.b but now with rows where $volume > 0$ and $volume \leq 800$.

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.