Mini project: Diamonds dataset

Aim: This project is an exploratory study on the amount of money people are willing to spend on diamonds. The aim is to analyze the aspects that characterize diamond prices.

Dataset description

The dataset "dataset_diamonds_NS.csv" contains the prices and other attributes of $1,000$ diamonds that have been purchased in the US in 2008. The variables available are the following (see Figure 1 for the visualization of features related to the diamond structure):

- `price`: price in US dollars (range $326 - 18,823$);
- `carat`: diamond's weight (range $0.2 - 5.01$);
- `cut`: quality of the cut (i.e., Fair, Good, Very Good, Premium, Ideal);
- `color`: diamond color, from J (worst) to D (best);
- `clarity`: measur of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best));
- `x`: length in mm (range $0 - 10.74$);
- `y`: width in mm (range $0 - 58.9$);
- `z`: depth in mm (range $0 - 31.8$);
- `depth`: total depth percentage $= \frac{z}{\text{mean}(x,y)} = \frac{2*z}{(x+y)}$ (range $43 - 79$);
- `table`: width of the top of the diamond relative to its widest point (range $43 - 95$).

Workflow

1. From your working environment `nicole_miniproject` import the dataset ("dataset_diamonds_NS.csv and change variable types, if necessary.
2. Add a column to you dataframe (or datatable) in which the *price per carat* is stored.

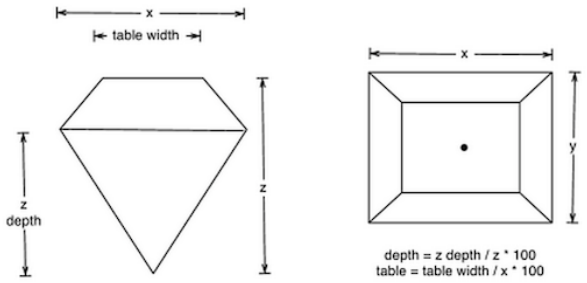

Figure 1: Diamond features.

- What does it change if we consider this new variable in the study? How might the answer to the research question change? State what you decide to do and why, step by step. [*hint: start your workflow without thinking too much to these questions. If you are not satisfied with your results you can go back to the beginning and change your point of view. Answer to the questions at the end, everything will be clearer... You can do all the changes you want.*]

3. Present and inspect the data (e.g. type of variables, sample size...). Provide a meaningful visual representation of the data with respect to the aspects you are going to investigate. Report descriptive statistics (do not forget categorical variables and their possible interactions).

4. What are the variables that influence diamond prices? What do you observe that might be unusual? What can you conclude on the diamond purchase habits?

5. Filter your data in order to obtain results more in line with what you thought it would influence diamond prices at the beginning (i.e. before analyzing the data). What can you conclude?

6. Which are the characteristics (physical and of the cost) the most purchased diamonds have? Are there any differences with the physical characteristics of the entire set of diamonds?

7. Write a function which takes as input a dataset and returns the indexes by columns only of the numeric variables stored in the dataset. Apply the function to the diamonds dataset and report the result.

Github

Follow the slides at

`https://www.slideshare.net/dadepo/introduction-to-git-and-github-13513987`

and set up a personal github account. Create a Public `diamond_miniproject` github folder, commit and push there your R code, these instructions, and your report. Edit the `.gitignore` file so as not to push to the remote folder the dataset itself
(i.e., "dataset_diamonds_NS.csv").