Diamonds analysis

Oriade Simpson (s172084), Pietro Lombardo (s231756)

2023-09-05

1) Description of the dataset

- Explain what your data is about. I.e. what is the overall problem of interest?
- Provide a reference to where you obtained the data.
- Summarize previous analysis of the data. (i.e. go through one or two of the original source papers and read what they did to the data and summarize their results).
- You will be asked to apply (1) classification and (2) regression on your data in the next report. For now, we want you to consider how this should be done. Therefore: Explain, in the context of your problem of interest, what you hope to accomplish/learn from the data using these techniques?. Explain which attribute you wish to predict in the regression based on which other attributes? Which class label will you predict based on which other attributes in the classification task? If you need to transform the data in order to carry out these tasks, explain roughly how you plan to do this.

One of these tasks (1)–(5) is likely more relevant than the rest and will be denoted the main machine learning aim in the following. The purpose of the following questions, which asks you to describe/visualize the data, is to allow you to reflect on the feasibility of this task.

2) Explanation of the attributes of the data

- Describe if the attributes are discrete/continuous, Nominal/Ordinal/Interval/ Ratio,
- Give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so.
- Include basic summary statistics of the attributes.

If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense).

3) Data visualization

Principal component analysis

The aim of the Principal Component Analysis ("PCA" from now on), is to simplify the problem dimension by reducing the number of variables which explains the behaviour of the diamonds' price.

As written before, the goal of the analysis is the regression of the variable *price* based on the other attributes of the dataset. For this reason, the dataset has to be deprived of the target variable *price*, which has to be explained by the other variables.

The PCA requires the dataset to be composed by numeric attributes. The reason is that each variable has to be standardized by subtracting the mean and dividing by the standard deviation of the whole set of observations of that variable, and such operations cannot be carried out for non-numeric values. Such requirement leads to trasform the three ordinal variables *cut*, *color* and *clarity* into ordinal numbers (from 1 to the upper level number).

After these operations, the dataset appears as below:

Table 1: Head of the diamonds dataset arranged for the PCA

carat	cut	color	clarity	depth	table	x	y	z
0.23	5	2	2	61.5	55	3.95	3.98	2.43
0.21	4	2	3	59.8	61	3.89	3.84	2.31
0.23	2	2	5	56.9	65	4.05	4.07	2.31
0.29	4	6	4	62.4	58	4.20	4.23	2.63
0.31	2	7	2	63.3	58	4.34	4.35	2.75
0.24	3	7	6	62.8	57	3.94	3.96	2.48

Now the dataset contains all numeric variables, so the next step is to compare the standard deviations of each variable and check whether they are different. In case they are different, the dataset has to be standardized so that all variables have the same order of magnitude in their values.

The figure below shows the standard deviations of the variables:

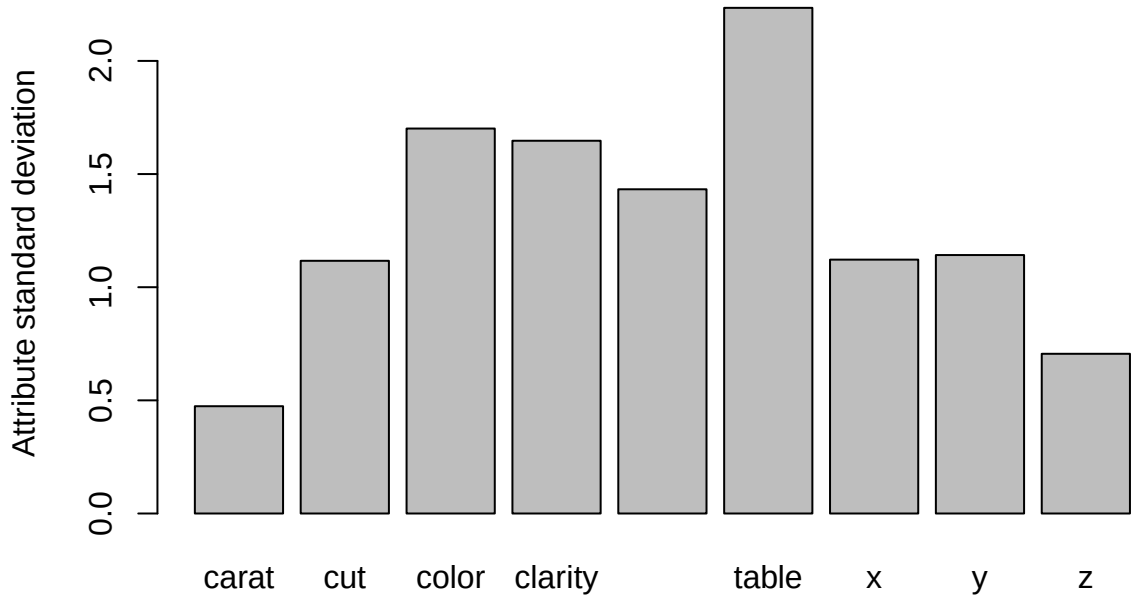

Figure 1: Standard deviations of variables

The variables have different standard deviations, so they have to be standardized.

Now the dataset is ready for the PCA

4) Results: what data have shown