Diamonds analysis

Oriade Simpson (s172084) Pietro Lombardo (s231756)

2023-09-05

Contribution Table

Task	Oriade	Pietro
Student ID	s172084	s231756
Question 1	x	
Question 2	x	
Question 3		x
Question 4		x
Exam Prob 1		x
Exam Prob 2	x	
Exam Prob 3	x	
Exam Prob 4		x
Exam Prob 5		x
Exam Prob 6	x	

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 behavior of the diamonds' price.

The PCA requires the standardization of the attributes so that the variability of each of them is comparable. But such operation cannot be carried out for non-numeric values, like the three ordinal variables *cut*, *color* and *clarity*. For these reason they are converted into ordinal numbers (from 1 to the upper level number) according to their ranking.

After these operations, the dataset appears as below:

Table 2: Head of the diamonds dataset arranged for the PCA

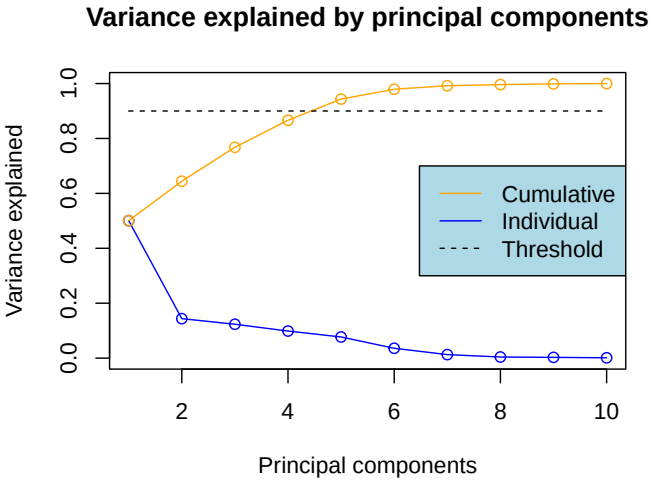
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	5	2	2	61.5	55	326	3.95	3.98	2.43
0.21	4	2	3	59.8	61	326	3.89	3.84	2.31
0.23	2	2	5	56.9	65	327	4.05	4.07	2.31
0.29	4	6	4	62.4	58	334	4.20	4.23	2.63
0.31	2	7	2	63.3	58	335	4.34	4.35	2.75
0.24	3	7	6	62.8	57	336	3.94	3.96	2.48

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.

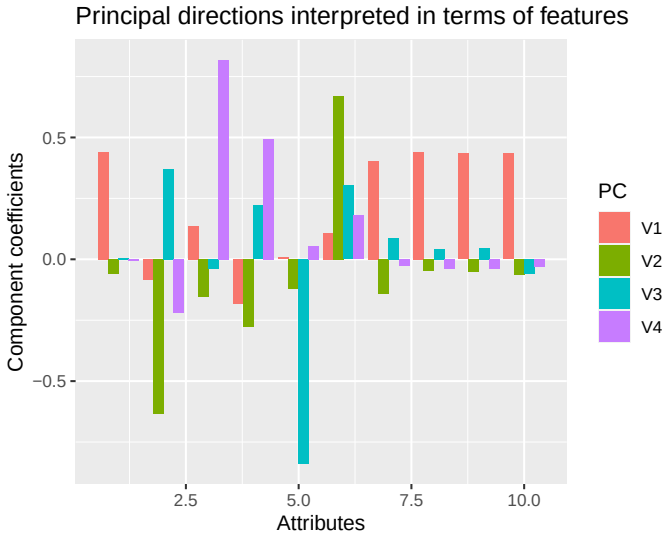
Below the standard deviations of the variables are shown:

```
##   carat     cut   color clarity   depth   table   price       x       y       z
##    0.47    1.12    1.70    1.65    1.43    2.23 3989.44    1.12    1.14    0.71
```

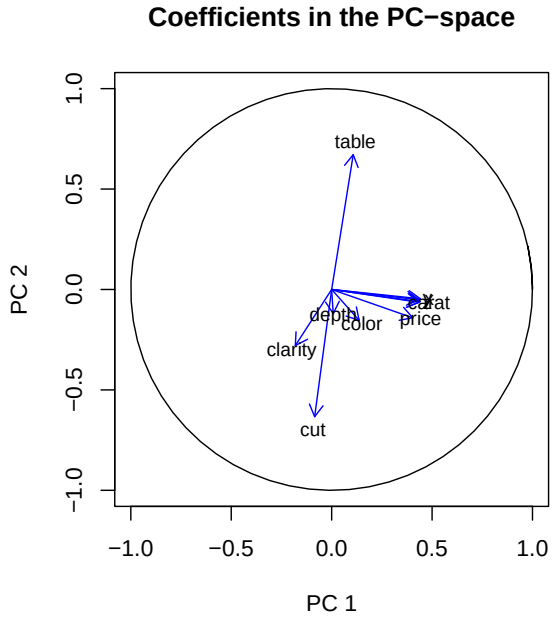
Since standard deviation of *price* is some orders of magnitude larger than the others, the dataset has to be standardized by subtracting the mean and dividing by the standard deviation of the whole set of observations of that variable.



The new standardized dataset can be now used to perform the Singular Value Decomposition (SVD), which gives rise to three different matrices: U, Σ and V^T. By the extraction of the diagonal of the matrix Σ, it can be seen how much variance is explained by each principal component. The cumulative explained variance should reach the percentage of 90% in order to describe properly the main features of the dataset. The figure on the left shows that the first 4 principal components explain little less than the 90% of variance



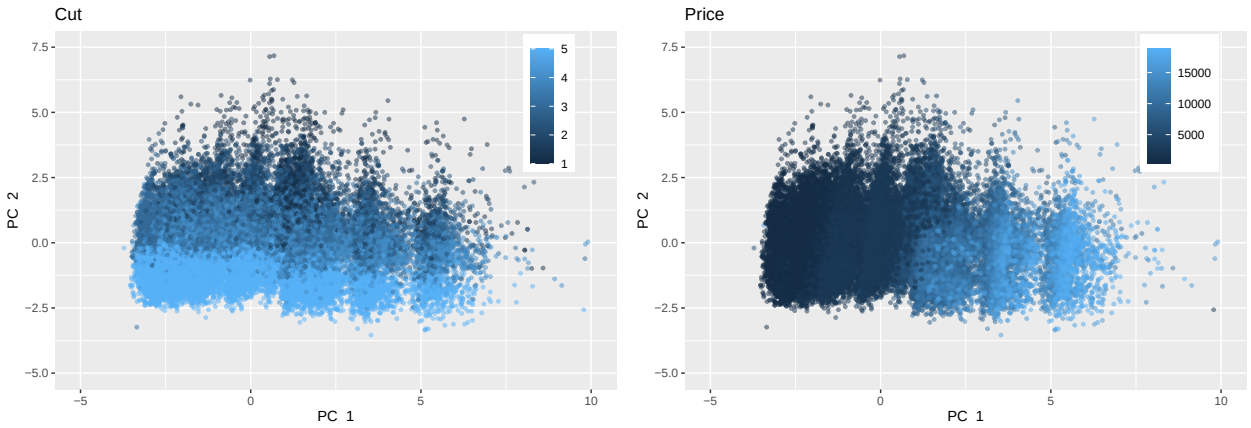
The matrix V^T contains the ten decadimensional vectors defining the principal components (PCs). Focusing on the first four PCs, they contains the weights associated to each of original component. The figure on the left shows the valuea of the weights. It can be noticed that the first PC mainly describes the dimensional quantities of the diamonds (*carat*, size in x, y, z) and the *price*. It seems that these five characteristics alone explain half of the variability of the diamonds. As for the second PC, it focuses more on the quality characteristics (*cut*, *color*, *clarity*) and on the *table*.

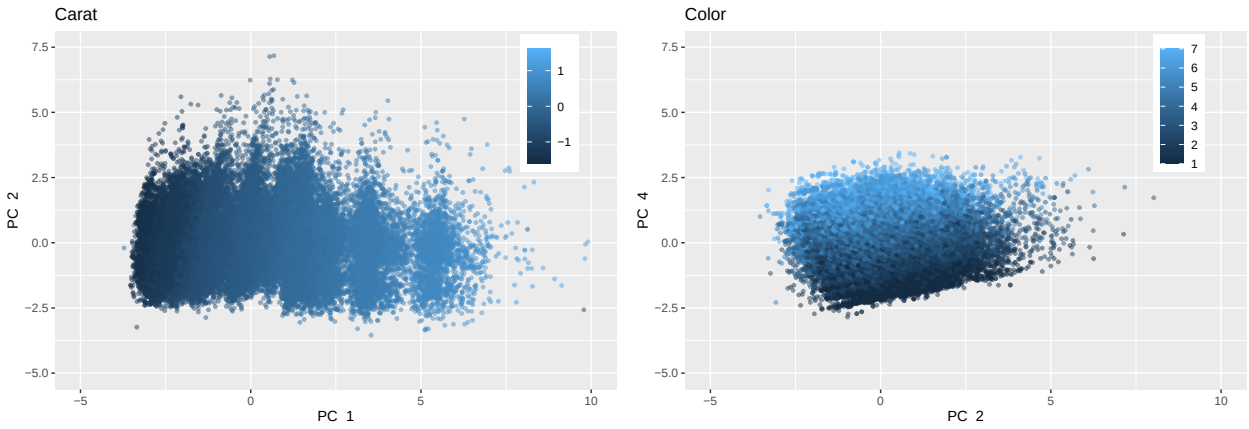


Unfortunately, it is impossible to represent data in a four-dimensional space, so data are projected onto the bi-dimensional space defined by the first two PCs. Coefficients can be projected in this space as well, showing the directions followed by original attributes. It can clearly be seen that the first PC is eastward oriented, so that it collects very well the eastward attributes. Probably, the second PC is southward directed so that it collects the southward attributes (with the plus sign) and the northward *table* attribute (with the minus sign)

Data can be projected in all the possible bi-dimensional spaces defined by each combination of two of the PCs. By projecting them, some insteresting features can be observed, as shown by the four graphs below.

- Variable *cut* can assume five possible levels, so it is the easiest one to be visualized. The graph shows that *cut* varies mostly in the direction of the second PC, so the *cut* of diamonds strongly depends on their quality features (the better are *color* and *clarity*, the better will be the cut), and is less affected by the size (x, y and z) and the *carat* (weight);
- variable *price* is strongly asymmethrical towards low values, and it is very evident also in the graph where cold blue is predominant with respect to light blue. *Price* strongly varies with the first PC, so it depends on quantity factors of diamonds (the bigger and heavier is the diamond, the more expensive will be) more than on quality ones;
- variable *carat* is very concentrated between 0 and 1 and widespread beyond 1. So, in order to have a better visualization, the graph below represent the logarithm of the *carat*. It can be seen that it is correlated to the dimensional quantities explained by the first PC (the more expensive and bigger is the diamond, the heavier will be) more than quality ones;
- variable *color* has a better representation if data are projected onto the plan defined by the second and forth PCs. It can be seen that *color* varies with the forth PC, so it is correlated with *clarity*, *cut* and *table*.





4) Results: what data have shown