# Case study

*Julie Josse, Erwan Scornet*

*Introduction to Machine Learning 2017*

## Easy Cost

A common approach to determine the cost of products is the **should cost** method. It consists in estimating what a product should cost based on materials, labor, overhead, and profit margin. Although this strategy is very accurate, it has the drawback of being tedious and it requires expert knowledge of industrial technologies and processes. To get a quick estimation, it is possible to build a statistical model to predict the price of products given their characteristics. With such a model, it would no longer be necessary to be an expert or to wait several days to assess the impact of a design modification, a change in supplier or a change in production site. Before builing a model, it is important to explore the data which is the aim of this case study.

This study was commissioned by a cosmetics company that wants to estimate the price of Screw Caps (bouchon Ã  vis) of shampoo bottles:


1) The data `ScrewCap.csv` contains 195 lots of screw caps described by 11 variables. Diameter, weight, length are the physical characteristics of the cap; nb.of.pieces corresponds to the number of elements of the cap (the picture above corresponds to a cap with 2 pieces: the valve (clapet) is made of a different material); Mature.volume corresponds to the number of caps ordered and bought by the compagny (number in the lot).

```
##        Supplier      Diameter          weight       nb.of.pieces
##  Supplier A: 31   Min.   :0.4458   Min.   :0.610   Min.   : 2.000
##  Supplier B:150   1st Qu.:0.7785   1st Qu.:1.083   1st Qu.: 3.000
##  Supplier C: 14   Median :1.0120   Median :1.400   Median : 4.000
##                   Mean   :1.2843   Mean   :1.701   Mean   : 4.113
##                   3rd Qu.:1.2886   3rd Qu.:1.704   3rd Qu.: 5.000
##                   Max.   :5.3950   Max.   :7.112   Max.   :10.000
##      Shape     Impermeability        Finishing   Mature.Volume
##  Shape 1:134   Type 1:172     Hot Printing: 62   Min.   :  1000
##  Shape 2: 45   Type 2: 23     Lacquering  :133   1st Qu.: 15000
##  Shape 3:  8                                     Median : 45000
##  Shape 4:  8                                     Mean   : 96930
##                                                  3rd Qu.:115000
##                                                  Max.   :800000
##  Raw.Material     Price            Length
##  ABS: 21      Min.   : 6.477   Min.   : 3.369
##  PP :148      1st Qu.:11.807   1st Qu.: 6.161
##  PS : 26      Median :14.384   Median : 8.086
##               Mean   :16.444   Mean   :10.247
##               3rd Qu.:18.902   3rd Qu.:10.340
##               Max.   :46.610   Max.   :43.359
```

2) We start with univariate and bivariate descriptive statistics. Using appropriate plot(s) or summaries answer the following questions.

- How is the distribution of the Price? Comment your plot with respect to the quartiles of the Price.
- Does the Price depend on the Length? weight?
- Does the Price depend on the Impermeability? Shape?
- Which is the less expensive Supplier?

3) One important point in exploratory data analysis consists in identifying potential outliers. Could you give points which are suspect regarding the Mature.Volume variable? Give the characteristics (other features) of the observations that seem suspect.

4) Perform a PCA on the dataset ScrewCap, explain briefly what are the aims of PCA and how categorical variables are handled?

5) Compute the correlation matrix between the variables and comment it with respect to the correlation circle.

6) On what kind of relationship PCA focuses? Is it a problem?

7) Comment the PCA outputs.

- Comment the position of the categories Impermeability=type 2 and Raw.Material=PS.
- Comment the percentage of inertia

8) Give the the R object with the two principal components which are the synthetic variables the most correlated to all the variables.

9) PCA is often used as a pre-processing step before applying a clustering algorithm, explain the rationale of this approach and how many components $k$ you keep.

10) Perfoms a kmeans algorithm on the selected $k$ principal components of PCA. How many cluster are you keeping? Justify.

11) Performs an AHC on the selected $k$ principal components of PCA.

12) Comments the results and describe precisely one cluster.

13) If someone ask you why you have selected $k$ components and not $k+1$ or $k-1$, what is your answer? (could you suggest a strategy to assess the stability of the approach? - are there many differences between the clustering obtained on $k$ components or on the initial data)

14) The methodology that you have used to describe clusters can also be used to describe a categorical variable, for instance the supplier. Use the function catdes(data, num.var=1) and explain how this information can be useful for the compagny.

15) To simultaneously take into account quantitative and categorical variables in the clustering you should use the HCPC function on the results of the FAMD ones. FAMD stands for Factorial Analysis of Mixed Data and is a PCA dedicated to mixed data. Explain what will be the impacts of such an analysis on the results?

16) Perform a model to predict the Price. Explain how the previous analysis can help you to interpret the results.

17) If someone ask you why you did one global model and not one model per supplier, what is your answer?

18) These data contained missing values. One representative in the compagny suggests either to put 0 in the missing cells or to impute with the median of the variables. Comment. For the categorical variables with missing values, it is decided to create a new category "missing". Comment.