1. Understanding and Analyzing Data

In this lab, you will work with a dataset about costumers of digital devices. You can get the data file from the course web page[1]. Please, note that all tasks that need computation can be done using the predefined Excel functions and the following statistical tool `http://www.alcula.com/calculators/statistics/`. So, no programming is required.

1. Analysis

The dataset contains information about costumers profiles and the items they have purchased from a store of digital devices.

1. How many attributes are contained in the dataset? and of which types they are?
2. Which attribute needs to be disregarded when we want to find similar costumers?
3. If we need to discover patterns of costumers buying only new devices, should we consider the binary attribute *second hand* as symmetric or asymmetric?
4. Is the attribute *weight* interval-scaled or ratio-scaled? Explain why.
5. Compute the mode, the mean, and the median of the attribute *expenses*. What do you conclude?
6. Represent the values of the attribute *budget* using a Boxplot. How many outliers does the data contain?
7. Compare the median and mean of the *budget* attribute. What do you observe? explain which measure represents better the budget values and why?

[1] http://www.inf.unibz.it/ mkacimi/teaching.shtml

8. To analyze the correlation between attributes, create a scatter plot of:
   - attributes *age* and *budget*. What do you observe? does the budget depend on the age of the costumer?
   - attributes *budget* and expenses . What do you observe?

2. Preprocessing

1. Find the most similar costumer to costumer 101, in terms of budget and expenses, using a Manhattan or an Euclidian distance. What do you observe?
2. Normalize the attributes *budget* and *expenses* and then find the most similar costumer to costumer 101. What do you observe?
3. Standardize the attributes *budget* and *expenses* and then find the most similar costumer to costumer 101. What do you observe?
4. What difference do you observe between normalization and standardization? What is the impact of outliers on both methods?
5. How would you compute the distance between two costumers taking into account all types of attributes? How would the formula change if we target only costumers of second hand devices?