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 at the following address: "".
Please, note that all tasks that need computation can be done using the predefined Excel functions. No programing 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?
   answer: there are 8 attributes: id, age, student, budget, expenses, devices, weight, and second hand. The attributes id and devices are nominal, age, budget, expenses, and weight are numeric, student and second hand are binary.

2. Which attribute needs to be disregarded when we want to find similar costumers?
   answer: the attribute id

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?
   answer: asymetric

4. Is the attribute *weight* interval-scaled or ratio-scaled? Explain why.
   answer: ratio scaled because the weight in grams has a true 0 value.

5. Compute the mode, the mean, and the median of the attribute *expenses*. What do you conclude?
   answer: mode =180, median=205, and mean=262.40. Thus, the distribution of expenses values is positively skewed.

6. Represent the values of the attribute *budget* using a Boxplot. How many outliers does the data contain?
   answer: Q1=2350, median= 3790, Q3= 5760, IQR= 5115, min=0 (we do not go below 0), max=10875. There are 2 values that go beyond the Whiskers, which are 18000 and 34200. So there are 2 outliers

7. Compare the median and mean of the *budget* attribute. What do you observe? explain which measure represents better the budget values and why?
   answer: mean= 7339.6 and median= 3792. The median represents the budget in a better way because it is not influenced by extreme values as the mean.

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?
     answer: age and budget are positively correlated. The higher the age, the higher budget is. See Excel file "CostumerDigitalSolution" for the scatter plot
   - attributes *budget* and expenses . What do you observe?
     answer: budget and expenses are not correlated. See Excel file "CostumerDigitalSolution" for the scatter plot

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?
   answer: There are two costumers with the same distance to costumer 101 which are the closest: 108 and 109. In practice 108 should be the closest because he has similar expenses to 101 than 109. The difference in USD should be considered more important than the difference in CNY. This is exactly the problem of attributes using different scales.

2. Normalize the attributes *budget* and *expenses* and then find the most similar costumer to costumer 101. How does that compare the first result?
   answer: there is only one costumer with the closest distance to costumer 101 which is 108 (See Excel file "CostumerDigitalSolution" for the scatter plot for normalization). The values of budget are between 0 and 1 but most

of them are very small because of dividing by the maximum number which is an outlier.

3. Standardize the attributes *budget* and *expenses* and then find the most similar costumer to costumer 101. How does that compare the first result?
   answer: there is only one costumer with the closest distance to costumer 101 which is 108 (See Excel file "CostumerDigitalSolution" for the scatter plot for standardization). The values are comparable to the expenses values and less influenced by outliers

4. What difference do you observe between normalization and standardization? What is the impact of outliers on both methods?
   answer: standardization is less sensitive to outliers.

5. How would you compute the distance between two costumers taking into account all types of attributes?
   answer: see slides and textbook.
   How would the formula change if we target only costumers of second hand devices?
   answer: see slides and textbook-Data Mining Techniques and Concepts.