# **Data exploration and preparation**

**Scenario**

You have just started working as a data miner/analyst in the Analytics Unit of a company. The Head of the Analytics Unit has brought you a dataset [a welcome present;-))]. The dataset includes two files: a description of the attributes and a table with the actual values of these attributes. The Head of the Analytics Unit has mentioned to you that this is some sort of marketing data that a potential client has provided for analysis. The Head of the Analytics Unit would like to have a report with some insights about the data, that he/she could deliver to the client. Your tasks include:

* understanding the specifics of the dataset;
* extracting information about each of the attributes, possible associations between them and other specifics of the dataset.

The tasks in the assignment are specified below.

**Tasks**

**1A. Initial data exploration**

1. Identify the attribute type of each **attribute** in your dataset {BATHRM, HF\_BATHRM, HEAT, HEAT\_D…….} (nominal, ordinal, interval or ratio). If it's not clear, you may need to justify why you chose the type.
2. Identify the values of the summarising properties for the **attributes**, including frequency, location and spread (e.g. value ranges of the attributes, frequency of values, distributions, medians, means, variances, percentiles, etc. - the statistics that have been covered in the lectures and materials given). Note that not all of these summary statistics will make sense for all the attribute types, so use your judgement! Where necessary, use proper visualisations for the corresponding statistics.
3. Using KNIME or other tools, explore your dataset and identify any outliers, clusters of similar instances, "interesting" attributes and specific values of those attributes. Note that you may need to 'temporarily' recode attributes to numeric or from numeric to nominal. The report includes the corresponding snapshots from the tools and an explanation of what has been identified there.

Present your findings in the assignment report.

**1B. Data preprocessing**

Perform each of the following data preparation tasks (each task applies to the original data) using your choice of tool:

1. Use the following **binning** techniques to smooth the values of the "**PRICE**” attribute:

* Equi-width binning
* Equi-depth binning.

In the assignment report, for each of these techniques, you need to illustrate your steps. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet. Use your judgement in choosing the appropriate number of bins - and justify this in the report.

2. Use the following techniques to **normalise** the attribute "**PRICE**":

* min-max normalization to transform the values onto the range [0.0-1.0].
* z-score normalization to transform the values.

The assignment report provides an explanation of each of the applied techniques. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

3. **Discretise** the "**PRICE**" attribute into the following categories: Low, Medium, High, and Expensive (e.g.: Low=0-50k; Medium=50k-100k; High=100k-1000k; Expensive= 1000k+ ). Provide the frequency of each category in your dataset.

The assignment report provides an explanation of each of the applied techniques. In your Excel workbook file place the results in a separate column in the corresponding spreadsheet.

4. **Binarise** the "**STRUCT\_D**"variable [with values "0" or "1"].

The assignment report provides an explanation of the applied binarisation technique. In your Excel workbook file place the results in separate columns in the corresponding spreadsheet.

**1C. Summary**

At the end of the report include a summary section in which you summarise your findings. The summary **is not** a narrative of what you have done, but a condensed informative section of **what you have found** about the data that you should report to the Head of the Analytics Unit. The summary may include the most important findings (specific characteristics (or values) of some attributes, important information about the distributions, some clusters identified visually that you propose to examine, associations found that should be investigated more rigorously, etc.).

**Datasets**

For this dataset, you only have the attribute headings and a brief of what they mean, Each student is assigned an individual dataset with the actual values of these attributes.

**Deliverables**

The deliveries are:

* A report, for which the structure should follow the tasks of the assignment, and
* An Excel workbook file with individual spreadsheets for each task (spreadsheets should be labelled according to the task names, for example, "1B"). Each of the results of parts (1) through (4) in task 1B should be presented in a separate spreadsheet (and respectively table in the assignment report).

In the report, include a section (starting with a section title) for each of the tasks in the assignment.

**Length**

The task requires a report of about 20 pages with the information and sections described above. Use 11 or 12 point Times or Arial fonts.

**Submission format**

Submit a report in Adobe PDF (preferable) or MS Word Doc with Filename ida\_a2\_xxx.pdf or ida\_a2\_xxx.doc and an Excel workbook with ida\_a2\_xxx.xls, where xxx is your student id.