# Pre-processing

Having received the real world data for the task, our team has been faced to the problem of the data being huge in size and having high number of variables. It therefore appeared of to be impossible to analyse the existence of possible amoralities in the data manually (inconsistencies, incompleteness, noisiness). In the pre-processing phase of the KDD Process our team has therefore applied several techniques of the Explorative Data Analysis in order to gain more detailed insights into the given dataset, detect the outliers and amoralities, uncover the underling data structure and to understand the given variables.

## Explorative Data Analysis

We have applied the following techniques before the firs iteration:

* histogram creation
* box plot
* scatter pots
* descriptive numerical summaries (mean, median, max, min etc.)
* correlation tables

### Histogram of numerical variables

We plotted and analysed all 138 numerical variables (!only numerical) of our dataset. Figure … shows an example of the 5 variables’ histograms.



Figure 1

Viewing all the 138 plots we could discover that our variables were mostly (not in all cases) distributed normally or following some kind of exponential density function. It became evident that many numerical variables had a huge amount of observations valued with “0”, which appeared to be a standard value in many cases. Thus for example the variables “age1” (blue) and “age2“ (red) – the age of the first and second adult in the household I are plotted in the Figure 2.



Figure 2

It is evident that in both variables many observations are valued with “0”, which is not a valid value for these variables. “=” appeared to be a standard value for these variables. A standard values, “0” would falsify the results in the Data Mining phase, and will have to be treated as missing values in the transformation phase of the KDD process in the first and second iterations. Figure 3 shows the plot of the variables “age1” and “age2” without the missing “0” values.

From the plot is evident that most of the first household members are between 40 and 45 years old. Furthermore, the plot does not exhibit any noticeable problems concerning the outliers or missing values.

We took a look on the histograms of several variables in order to find the possible outliers. Such, the Figure … shows the distribution of the variable “adjrev” - billing adjusted total revenue over the life of the customer, for all values (left) and only for the values greater than 8000 (right).

As can be seen from the left histogram of the Figure 4 due to the he amount of observations, the outliers cannot be seen in the histogram when showing the whole data. In contrary, while only plotting the data starting from a higher value (e.g. the median or third quantile) makes it possible to visually identify the existence of the outliers. For the better assessment of the existence of the outliers, we decided to use boxplots.



Figure 3



Figure 4

### 2. Box-plotting the data

Box plots are commonly used for the visual outlier identification. For all numerical variables a boxplot was created. Figure 5 shows an example of a boxplot with the range value of 3 for the variable “afjrev” (billing adjusted total revenue over the life of the customer). From the box plot it is evident that this variable has outliers. Depending on the range value, and thus on the height of the upper antenna, different number of values can be considered as outliers. The three single values in the range 15000 – 30000 can be definitely considered as outliers.

The outlier handling will be described in more detail in the following chapter in the individual iterations. The boxplots have given a good impression about the existence of the outliers and the distribution of the values of the variables. It is important to say, the outlier analysis should be considered after the missing value handling, since the handling of missing/default values (such as default “0”) will influence the distribution of a variable and so the outlier detection.



Figure 5

### 3. Scatterplots

In order to create an impression about the dependency structure of several interrelated variables several scatter plots have been created. Several scatter plot matrixes have been created, similar to the one presented in the Figure 6. From the presented scatterplot matrix for example can be seen, that the total revenue (totrev) tends to grow with the growing mean total monthly recurring charge (totmrc\_Mean). Furthermore, it is evident that there are many outliers in each of the scatterplot – the multidimensional outlier handling might be useful in this dataset.

### 4. Descriptive numerical summaries

Numerical summaries of variables are useful in many aspects while trying to gain more detailed insights into the data. Thus, for example, by estimating the median of a variable and evaluating the range between the maximal and the minimal value of a variable can be understood if a variable has outliers. Using the “summary” function of R we have calculated the following characteristic values of every numeric variable: mean, median, 1st and 3rd quartiles, as well as maximum and minimum values.

Thereafter, we have divided the variables among the group members and scanned the numerical summary for the conspicuous values and errors. Thus, we have discovered a negative minimum value of the variables “REV\_MEAN” and “TOTMRC\_MEAN” that appeared to be an invalid outlier, since this variable cannot be negative as they represent revenue and the monthly recurring charges.

In general, the numerical summary appeared to be a good identifier for the detection of variables with outliers. Thus, for example, variables who’s mean/average deviation appeared to be big, can be in most cases successfully further investigated on subject of outlier detection.



Figure 6

### 5. Correlation

In order to visualize the correlation between the variables oft he dataset, the correlation plot was used that is presented in the Figure 7 . The dark blue colour indicates a strong correlation between variables. From the plot it is evident that there are several variables that correlate strongly (between 0,8 and 1). Consequently it will make sense to filter out the strongly correlated variables in the transformation phase of the KDD process.



Figure