**1st Iteration Pre-processing: Data Cleaning: Outlier Detection and Handling**

An outlier is a value that is significantly outside the range of the other values. While the explorative data analysis has show that there are outliers in most of the variables of our dataset, in the data transformation phase these outliers should be detected and handled.

In the first iteration our team has decided to only handle the outliers detected by the means of the simple one-dimensional outlier detection method. In this iteration we have started with the simple boxplot outlier detection method experimenting with the heights of the antennas of the boxplot, finally used the robust z-score method as a final method for the outlier detection in this iteration. For the reasons of the simplification we have handled the outliers detected by the named methods for every numerical variable.

## Boxplot method

The distance between the first and the third quartiles (interquartile distance -IQR) of a boxplot is measure of variation of the variable around the median. Whiskers of a boxplot are commonly calculated using the factor of 1,5 or 3 and multiplying it with the IQR. It is a common praxis to consider the values that lie outside the boarders of the whiskers as outliers.

In order to find the extreme outliers for the numerical variables we have calculated the first and the second quartiles of the variable and the IQR. We have then calculated the whiskers as 3\*IQR. The values that were outside the limits set by the whiskers were considered as outliers. We handled the outliers by setting their value to the value of its closest whisker, which allowed us to get rid of the observation that are significantly out of the range of the values of the variable without influencing the values of the median and the quartiles of the variable. The effect of the method is shown in the Figure 1 on example of the variable “rev\_Mean” (mean monthly revenue). On its left-hand side the figure shows a boxplot of the variable “rev\_Mean” before the outlier handling, and a boxplot of the same variable after the outlier handling on its right-hand side (cf. Buttler, n.d., pp. 7f).



Figure 1: BoxPlots oft he variable „rev\_Mean“ with range of 3 before(left) and after(right) the outlier handling

## Z-score method

After having discussed the z-score in the “Business Analytics and Data Science” lecture we have decided to implement this method for the outlier detection in the first iteration. Being similar to the boxplot outlier detection method described above, z-score method appeared more robust to us because it calculates a score for each values based on a standard deviation.

We used the following z-score formula ( ) to calculate the distance of the observation from the population’s mean measured in standard deviation. Every observation that appeared to have the z-score higher that 3 (of lower than -3) was considered to be an outlier. Every outlier was than set to the value the population’s mean ( added the three standard deviations (3\* (cf. Lessmann, 2015, Weinberg/Abramowitz, 2002 pp. 105f).

In its core idea and implementation, the z-score method is very similar to the simple boxplot method described above – the outliers are set to the value of the whiskers, whose length is equal to three times standard deviation of the population, neither the value of the mean nor (in most cases) the values of the quartiles are influenced. The result of the z-score outlier handling is presented exemplary on the variable “rev\_Mean” in the Figure 2.



Figure 2: Boxplots oft he variable „rev\_Mean“ with range of 3 before(left) and after(rright) the outlier handling with z-score method

Sources:

Buttler, G. (n.d.). Ein einfaches Verfahren zur Identifikation von Ausreißern bei multivariaten Daten. [online] Available at: http://www.statistik.wiso.uni-erlangen.de/forschung/d0009.pdf [Accessed 28 Dec. 2015].

Lessmann, S. (2015). *Business Analytics and Data Science: Chapter 9 Data Preparation*.

Weinberg, S. and Abramowitz, S. (2002). *Data analysis for the behavioral sciences using SPSS*. Cambridge, UK: Cambridge University Press.