Data Cleaning

In this section we give a brief overview of our investigation of potential data quality issues and how we handle them. We perform the following sanity checks:

Suspicious 0's and NA values

There are no NA values. The data set as it was given to us is complete. There are 619 0's for 'NETSPEND'. These are legitimate as they refer to purchases with a promotion '3F2' (i.e. 3-for-2). The third pack is recorded with a value of 0 for 'NETSPEND'.

Identical rows

37616 observations are not unique but repeat (i.e. there is at least one more row which is identical for all of its column-values). All of these are assumed to refer to purchases of several packs in scope of the same shop visit.

Quasi-identical rows (rows which only differ in their column-value for 'NETSPEND')

During the explanatory analysis we came across the fact that the value for 'NETSPEND' seems to be erroneous in some cases. There are observations for which all column values except 'NETSPEND' are identical. In part this is explained by the type of promotion (e.g. the 3F2 promotion as described above). However, some of the values are flawed. There is no apparent difference in the 2 rows, no promotion applies, but the 'NETSPEND' amount differs nevertheless. This may be due to several reasons. In the majority of cases, however, the error is economically very insignificant. Also, given the total size of the data set, these observations do not influence any part of the analysis meaningfully.

Conclusion

In addition to filtering for retailers and brands no further data cleaning is implemented. The inconsistencies that have been identified are minor and do not influence the analysis significantly.