Retail Market Basket Data Set

Tom Brijs

Research Group Data Analysis and Modeling
Limburgs Universitair Centrum
Universitaire Campus, B-3590 Diepenbeek, BELGIUM
email:tom.brijs@luc.ac.be

Abstract

This document describes the retail market basket data set supplied by a anonymous Belgian retail supermarket store. The document describes the contents of the data, the period over which the data were collected, some characteristics of the data and legal issues with respect to the use of this data set.

Keywords: Interestingness, Association Rules

1 General Description of the Data

The data are collected over three non-consecutive periods. The first period runs from half December 1999 to half January 2000. The second period runs from 2000 to the beginning of June 2000. The third and final period runs from the end of August 2000 to the end of November 2000. In between these periods, no data is available, unfortunately. This results in approximately 5 months of data. The total amount of receipts being collected equals 88,163.

Each record in the data set contains information about the date of purchase (variable 'date'), the receipt number (variable 'receipt_nr'), the article number (variable 'article_nr'), the number of items purchased (variable 'amount'), the article price in Belgian Francs (variable 'price' with 1 Euro = 40.3399 BEF) and the customer number (variable 'customer_nr'). Note that the article price in the data set equals the unit price of the article times the amount of items purchased.

Over the entire data collection period, the supermarket store carries 16,470 unique SKU's, but some of them only on a seasonal basis, e.g. Christmas items. Although most of the products are identified by a unique barcode (i.e. the barcode), some article numbers in the data set represent a group of products rather than an individual product item. For instance, this is the case with fruits, vegetables, meat, and a few others.

In total, 5,133 customers have purchased at least one product in the supermarket during the data collection period.

2 Data Statistics

This paragraph presents some general statistics about the data set that may be used to put the data mining results into the right context.

Figure 1 shows the average number of distinct items purchased per shopping visit. The average number of distinct items (i.e. different products) purchased per shopping visit equals 13 and most customers buy between 7 and 11 items per shopping visit.

Figure 1: Average number of distinct items purchased per visit

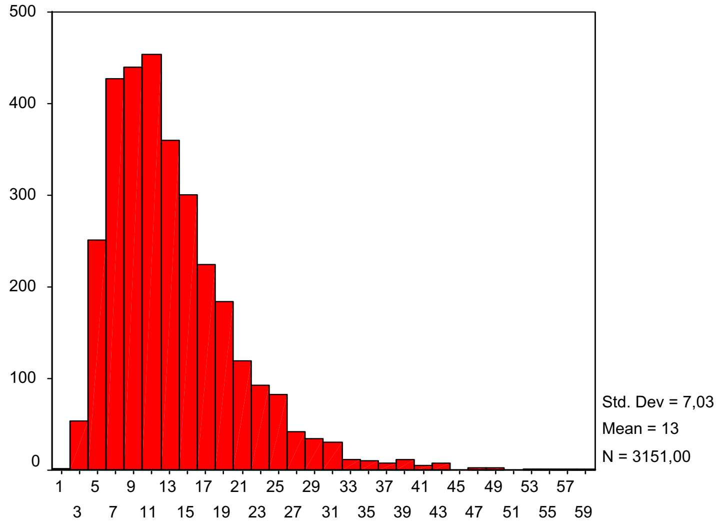


Figure 2 shows the distribution of the average amount spent (in Belgian francs) per shopping visit. The average amount spent, over all households, equals 1276 BEF (or 31.63 Euro).

Figure 2: Average amount spent (in BEF) per shopping visit

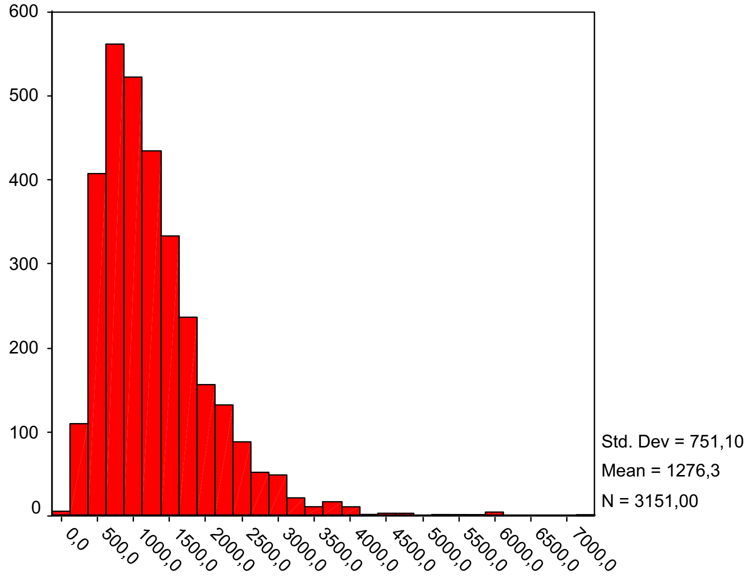
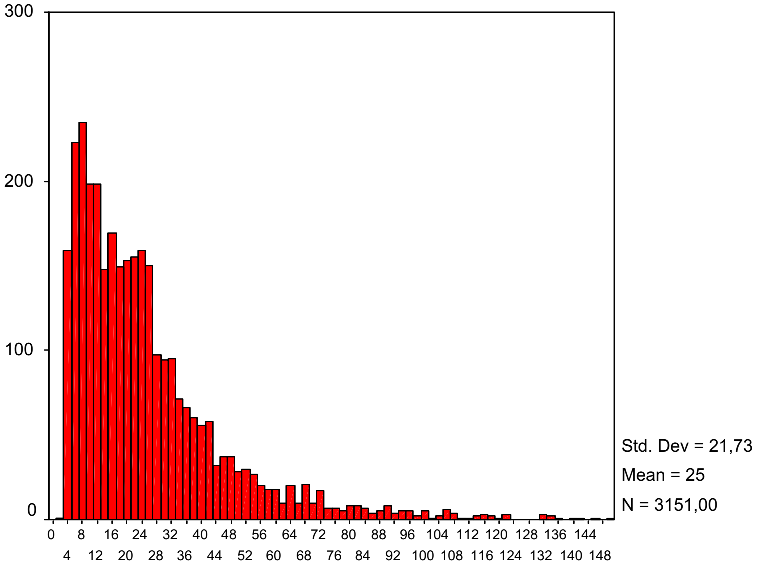


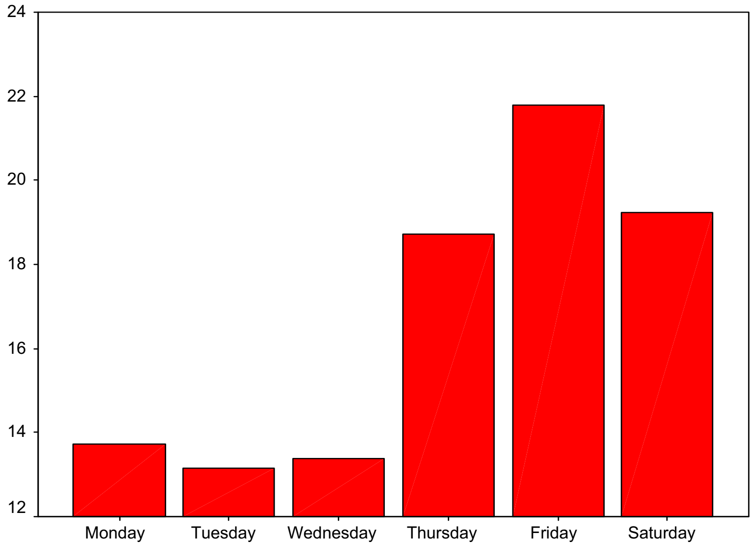
Figure 3 shows the distribution of the total number of visits over the period of data collection. Although most customers have visited the store from 4 to 24 times over the entire period (24 weeks), the average number of visits to the store equals 25, which corresponds to about once per week.

Figure 3: Total number of visits over 24 weeks



Finally, figure 4 shows the distribution of the daily visits to the store. From this figure, it is clear that most of the visits to the store take place on Thursday, Friday and Saturday.

Figure 4: Distribution of visits per day of the week



Furthermore, it is worth mentioning that further analysis showed that 89% the items in the assortment are slow moving, i.e. sold on average less than once per day.

3 Legal Issues

The data are provided ’as is’. Basically, any use of the data is allowed as long as the proper acknowledgment is provided and a copy of the work is provided to Tom Brijs (see details below). For papers with a bibliographic section, reference should be made to the following paper (which is available for download at http://citeseer.nj.nec.com/brijs99using.html) where parts of this dataset were used and described:

Brijs T., Swinnen G., Vanhoof K., and Wets G. (1999), The use of association rules for product assortment decisions: a case study, in: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), August 15-18, pp. 254-260. ISBN: 1-58113-143-7.

The bibtex entry is:

```
@inproceedings brijs99using,
    author = ”Tom Brijs and Gilbert Swinnen and Koen Vanhoof and Geert Wets”,
    title = ”Using Association Rules for Product Assortment Decisions: A Case Study”,
    booktitle = ”Knowledge Discovery and Data Mining”,
    pages = ”254-260”,
    year = ”1999”}
```

The first submission and final text of any written work utilizing this Retail market basket data set must be sent to the Research Group Data Analysis and Modelling along with the date and title of the publication where such work will appear. E-mail copies are preferred and should be sent to tom.brijs@luc.ac.be. The address for mail correspondence is:

Tom Brijs, Ph.D.
Limburgs Universitair Centrum
Research Group Data Analysis and Modelling
Universitaire Campus - gebouw D
B-3590 Diepenbeek, Belgium

Note that this mail or e-mail is for tracking purposes. No approval is required as long as the above conditions are met.