Chapter 8
Descriptive Data Mining

This book addresses descriptive analytics, an initial aspect of data mining. As stated in the preface, it looks at various forms of statistics to gain understanding of what has happened in whatever field is being studied. The book begins with a chapter on knowledge management, seeking to provide a context of analytics in the overall framework of information management. It begins reviewing computer information systems, a source of much data of importance, and its storage and retrieval to aid decision making. The impact of big data on this environment has been dramatic, requiring greater reliance on artificial intelligence and automated processing of data. Thus knowledge management needs to identify useful patterns by collecting data, storing it, retrieving as needed for modeling and interpreting results to gain useful, actionable information.

Chapter 2 focuses on the general topic of visualization. Of the many ways visualization is implemented to inform humans of what statistics can reveal, we look at data mining software visualization tools, as well as simple spreadsheet graphs enabling understanding of various kinds of data. US energy data is used to demonstrate rich opportunities for students to further study important societal issues.

Chapter 3 describes basic cash register information by sale that has been used by retail organizations to infer understanding of what items tend to be purchased together. This can be useful to support product positioning in stores, as well as other business applications. Market basket analysis is among the most primitive forms of descriptive data mining. The chapter looks at basic tools of co-occurrence, lift, and correlation.

Chapter 4 addresses a basic marketing tool that has been around for decades. Retailers have found that identifying how recently a customer has made a purchase is important in gauging their value to the firm, as well as how often they have made purchases, and the amount purchased. Recency, Frequency and Monetary (RFM) analysis provides a quick and relatively easy to implement methodology to categorize customers. There are better ways of analysis, and there is a lot of data

D. L. Olson and G. Lauhoff, *Descriptive Data Mining*, Computational Risk Management, https://doi.org/10.1007/978-981-13-7181-3_8

Table 8.1 Descriptive data mining methods

Method	Descriptive process	Basis	Software
Visualization	Initial exploration	Graphical statistics	Spreadsheet
Market basket analysis	Retail cart analysis	Correlation	Spreadsheet
Recency/Frequency/ Monetary	Sales analysis	Volume	Spreadsheet manipulation
Association rules	Grouping	Correlation	APriori, others
Cluster analysis	Grouping	Statistics	Data mining (R, WEKA)
Link analysis	Display	Graphics	PolyAnalyst, NodeXL

transformation work involved, but this methodology helps understand how descriptive data can be used to support retail businesses.

Chapter 5 deals with the first real data mining tool—generation of association rules by computer algorithm. The basic a priori algorithm is described, and R software support demonstrated. A hypothetical representation of e-commerce sales is used for demonstration. Fundamental concepts of support, confidence, and lift are demonstrated, both for manual calculation for understanding as well as Rattle computation.

Chapter 6 presents basic algorithms used in cluster analysis, followed by analysis of typical bank loan data by three forms of open source data mining software. Rattle provides K-means, Entropy Weighted K-Means, and Hierarchical algorithms. KNIME and WEKA software are also briefly demonstrated. More powerful tools such as self-organizing maps are briefly discussed.

Finally, the use of link analysis is shown with two forms of software in Chap. 7. First, basic social network metrics are presented. An open source version of NodeXL is demonstrated. It is not powerful, and cannot do what the relatively inexpensive proprietary version can do. The output of the commercial software PolyAnalyst is used to demonstrate some valuable applications of link analysis.

These methods can be compared, as in Table 8.1.

The methods in Table 8.1 often require extensive data manipulation. Market basket analysis and RFM may call for extensive manipulation of spreadsheet data. There are commercial software products that can support these applications. Such products tend to come and go, so a search of the Web is appropriate should you wish to find software. Regardless, keep in mind that almost all data mining applications require extensive data manipulation and cleansing.

Descriptive analysis involves many different problem types, and is supported by a number of software tools. With the explosion of big data, initial data analysis by description is useful to begin the process of data mining.