

Computational Risk Management

David L. Olson  
Georg Lauhoff

# Descriptive Data Mining

*Second Edition*



Springer

# **Computational Risk Management**

## **Editors-in-Chief**

Desheng Dash Wu, RiskLab, University of Toronto, Toronto, ON, Canada

David L. Olson, Department of Supply Chain Management and Analytics,  
University of Nebraska-Lincoln, Lincoln, NE, USA

John Birge, University of Chicago Booth School of Business, Chicago, IL, USA

Risks exist in every aspect of our lives and risk management has always been a vital topic. Most computational techniques and tools have been used for optimizing risk management and the risk management tools benefit from computational approaches. Computational intelligence models such as neural networks and support vector machines have been widely used for early warning of company bankruptcy and credit risk rating. Operational research approaches such as VaR (value at risk) optimization have been standardized in managing markets and credit risk, agent-based theories are employed in supply chain risk management and various simulation techniques are employed by researchers working on problems of environmental risk management and disaster risk management. Investigation of computational tools in risk management is beneficial to both practitioners and researchers. The Computational Risk Management series is a high-quality research book series with an emphasis on computational aspects of risk management and analysis. In this series, research monographs as well as conference proceedings are published.

More information about this series at <http://www.springer.com/series/8827>

David L. Olson · Georg Lauhoff

# Descriptive Data Mining

Second Edition



Springer

David L. Olson  
College of Business  
University of Nebraska–Lincoln  
Lincoln, NE, USA

Georg Lauhoff  
San Jose, CA, USA

ISSN 2191-1436  
Computational Risk Management  
ISBN 978-981-13-7180-6  
<https://doi.org/10.1007/978-981-13-7181-3>

ISSN 2191-1444 (electronic)  
ISBN 978-981-13-7181-3 (eBook)

Library of Congress Control Number: 2019934798

© Springer Nature Singapore Pte Ltd. 2017, 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,  
Singapore

# Preface

Knowledge management involves the application of human knowledge (epistemology) with the technological advances of our current society (computer systems) and big data, both in terms of collecting data and in analyzing it. We see three types of analytic tools. **Descriptive** analytics focus on the reports of what has happened. **Predictive** analytics extend statistical and/or artificial intelligence to provide forecasting capability. It also includes classification modeling. **Diagnostic** analytics can apply analysis to sensor input to direct control systems automatically. **Prescriptive** analytics applies quantitative models to optimize systems, or at least to identify improved systems. Data mining includes descriptive and predictive modeling. Operations research includes all the three. This book focuses on descriptive analytics.

Lincoln, USA  
San Jose, USA

David L. Olson  
Georg Lauhoff

# **Book Concept**

The book seeks to provide simple explanations and demonstration of some descriptive tools. This second edition provides more examples of big data impact, updates the content on visualization, clarifies some points, and expands coverage of association rules and cluster analysis. Chapter 1 gives an overview of the context of knowledge management. Chapter 2 discusses some basic software support to data visualization. Chapter 3 covers fundamentals of market basket analysis, and Chap. 4 provides a demonstration of RFM modeling, a basic marketing data mining tool. Chapter 5 demonstrates association rule mining. Chapter 6 has more in-depth coverage of cluster analysis. Chapter 7 discusses link analysis.

Models are demonstrated using business-related data. The style of the book is intended to be descriptive, seeking to explain how methods work, with some citations, but without deep scholarly references. The data sets and software are all selected for widespread availability and access by any reader with computer links.

# Contents

<b>1 Knowledge Management . . . . .</b>	1
Computer Support Systems . . . . .	2
Examples of Knowledge Management . . . . .	4
Data Mining Descriptive Applications . . . . .	7
Summary . . . . .	8
References . . . . .	8
<b>2 Data Visualization . . . . .</b>	11
Data Visualization . . . . .	11
R Software . . . . .	12
Loan Data . . . . .	13
Energy Data . . . . .	20
Basic Visualization of Time Series . . . . .	21
Conclusion . . . . .	28
References . . . . .	30
<b>3 Market Basket Analysis . . . . .</b>	31
Definitions . . . . .	32
Co-occurrence . . . . .	33
Demonstration . . . . .	37
Fit . . . . .	38
Profit . . . . .	38
Lift . . . . .	41
Market Basket Limitations . . . . .	43
References . . . . .	44
<b>4 Recency Frequency and Monetary Analysis . . . . .</b>	45
Dataset 1 . . . . .	46
Balancing Cells . . . . .	50
Lift . . . . .	52
Value Function . . . . .	53

Data Mining Classification Models . . . . .	58
Logistic Regression . . . . .	58
Decision Tree . . . . .	59
Neural Networks . . . . .	59
Dataset 2 . . . . .	59
Conclusions . . . . .	63
References . . . . .	65
<b>5 Association Rules . . . . .</b>	67
Methodology . . . . .	68
The Apriori Algorithm . . . . .	69
Association Rules from Software . . . . .	71
Non-negative Matric Factorization . . . . .	75
Conclusion . . . . .	76
References . . . . .	76
<b>6 Cluster Analysis . . . . .</b>	77
K-Means Clustering . . . . .	78
A Clustering Algorithm . . . . .	78
Loan Data . . . . .	79
Clustering Methods Used in Software . . . . .	81
Software . . . . .	82
R (Rattle) K-Means Clustering . . . . .	82
Other R Clustering Algorithms . . . . .	88
KNIME . . . . .	96
WEKA . . . . .	98
Summary . . . . .	105
References . . . . .	106
<b>7 Link Analysis . . . . .</b>	107
Link Analysis Terms . . . . .	107
Basic Network Graphics with NodeXL . . . . .	114
Network Analysis of Facebook Network or Other Networks . . . . .	118
Link Analysis of Your Emails . . . . .	124
Link Analysis Application with PolyAnalyst (Olson and Shi 2007) . . . . .	125
Summary . . . . .	128
References . . . . .	128
<b>8 Descriptive Data Mining . . . . .</b>	129

# About the Authors

**David L. Olson** is the James & H.K. Stuart Chancellor's Distinguished Chair and Full Professor at the University of Nebraska. He has published research in over 150 refereed journal articles, primarily on the topic of multiple objective decision-making, information technology, supply chain risk management, and data mining. He teaches in the management information systems, management science, and operations management areas. He has authored over 20 books. He is Member of the Decision Sciences Institute, the Institute for Operations Research and Management Sciences, and the Multiple Criteria Decision Making Society. He was a Lowry Mays endowed Professor at Texas A&M University from 1999 to 2001. He was named the Raymond E. Miles Distinguished Scholar award for 2002, and was a James C. and Rhonda Seacrest Fellow from 2005 to 2006. He was named Best Enterprise Information Systems Educator by IFIP in 2006. He is a Fellow of the Decision Sciences Institute.

**Georg Lauhoff** is Technologist at Western Digital Corporation and carries out R&D in materials science and its application in data storage devices and uses the techniques described in this book for his work. He co-authored 38 refereed journal articles and over 30 conference presentations, primarily on the topic of materials science, data storage materials, and magnetic thin films. He was awarded scholarships and research grants in the UK and Japan. He was the Clerk Maxwell Scholar from 1995 to 1998 and is a Fellow of the Cambridge Philosophical Society. He studied physics at Aachen (Diplom) and Cambridge University (Master and Ph.D.) specializing in the field of materials science and magnetic thin films and sensors. After graduating, he moved to Japan and held a faculty position in Materials Science and Engineering at the Toyota Technological Institute and then carried out research in the sequencing of DNA using magnetic sensors at Cambridge University before moving in 2005 to the recording industry in the Bay area.

# Chapter 1

## Knowledge Management



We live in an era of ubiquitous information, with masses of data available concerning practically every aspect of life. Our daily lives can be supported with fitbits to monitor health-related data, with sensor systems to monitor our homes and screen those who come to our front door, to monitor our automobiles at the same level that used to be applied to space shuttles, and to guide our driving paths to avoid traffic stops. In business, farming can be guided by GPS, using advanced genetics for seeds, fertilizers, and plant disease control, much as Wal-Mart monitors their inventories at a micro-level and banks optimize their marketing materials. Heaven only knows what governments around the world are doing to keep us safe, or conversely, to put us at risk.

All of this is made possible by applying statistical analysis with artificial intelligence to process all of this data into something useful. Knowledge management is an overarching term referring to the ability to identify, store, and retrieve knowledge. **Identification** requires gathering the information needed and to analyze available data to make effective decisions regarding whatever the organization does. This include research, digging through records, or gathering data from wherever it can be found. **Storage** and **retrieval** of data involves database management, using many tools developed by computer science. Thus knowledge management involves understanding what knowledge is important to the organization, understanding systems important to organizational decision making, database management, and analytic tools of data mining.

The era of big data is here. Davenport (2014) defines big data as:

- Data too big to fit on a single server;
- Too unstructured to fit in a row-and-column database;
- Flowing too continuously to fit into a static data warehouse;
- Having the characteristic of lacking structure

Knowledge management (KM) needs to cope with big data by identifying and managing knowledge assets within organizations. KM is process oriented, thinking in terms of how knowledge can be acquired, as well as tools to aid decision making.

Rothberg and Erickson (2005) give a framework defining data as **observation**, which when put into context becomes **information**, which when processed by human understanding becomes **knowledge**. The point of big data is to analyze, converting data into insights, innovation, and business value. It can add value by providing real-time measures of performance, provide more timely analyses based on more complete data, and lead to sounder decisions (Manyika et al. 2011).

Waller and Fawcett (2013) describe big data in terms of **volume**, **velocity**, and **variety**.

- Volume is clearly massive when considering scientific endeavors such as weather forecasting. Satellites fire streams of data to the National Oceanic and Atmospheric Administration computers to digest and feed into multiple forecasting software services. A similar scale of data is found in retail organizations such as Wal-Mart, monitoring sales and inventories throughout their supply chains to provide the lowest cost—(always?).
- With respect to velocity, sales data can be real-time, as well as aggregated to hourly, daily, weekly, and monthly form to support marketing decisions. Inventory data obtained in real-time can be aggregated to hourly or monthly updates. Location and time information can be organized to manage the supply chain.
- Variety is magnified in this context by sales events from cash registers in brick-and-mortar locations, along with Internet sales, wholesale activity, international activity, and activity by competitors. All of this information can be combined with social media monitoring to better profile customers. Inventory activity can be monitored by type of outlet as well as by vendor. In the healthcare realm, variety includes textual records as well as x-rays, MRIs, photographs, and practically every form of data type.

This volume, velocity, and variety can only be coped with through use of computer software. Sensor-obtained data can be traced by workers involved, paths used, and locations. Artificial intelligence is widely applied to cope with this flood of big data.

## Computer Support Systems

Computer systems have been applied to support business decision making for decades (Olson and Courtney 1992). When personal computers came out, they were used to provide analytic tools for specific problems (**decision support systems**) (Sprague and Carlson 1982). Commercial software firms (such as Execucom and Comshare) extended this idea to dedicated systems to serve executives by providing them the key data they were expected to be interested in at their fingertips (**executive support systems**). Another commercial application was **on-line analytic processing**, developing database spreadsheet software capable of providing reports on any of a number of available dimensions.

In a parallel universe, statisticians and students of artificial intelligence revolutionized the field of statistics to develop data mining, which when combined with database capabilities evolving on the computer side led to **business intelligence**. The quantitative side of this development is **business analytics**, focusing on providing better answers to business decisions based on access to massive quantities of information ideally in real-time (**big data**).

Davenport (2018) reviewed four eras of analytics (see Table 1.1). The first era involved business intelligence, with focus on computer systems to support human decision making (for instance, use of models and focused data on dedicated computer systems in the form of decision support systems). This reliance upon computers to support humans was limited temporally by human limitations. The second era saw use of big data, through internet and social media generation of masses of data. Search and recommendation were very useful tools. Davenport sees a third era in a data-enriched environment where on-line real-time analysis can be conducted by firms in every industry. This is accomplished through new tools, using Hadoop clusters and NoSQL databases to enable data discovery, applying embedded analytics supporting cross-disciplinary data teams.

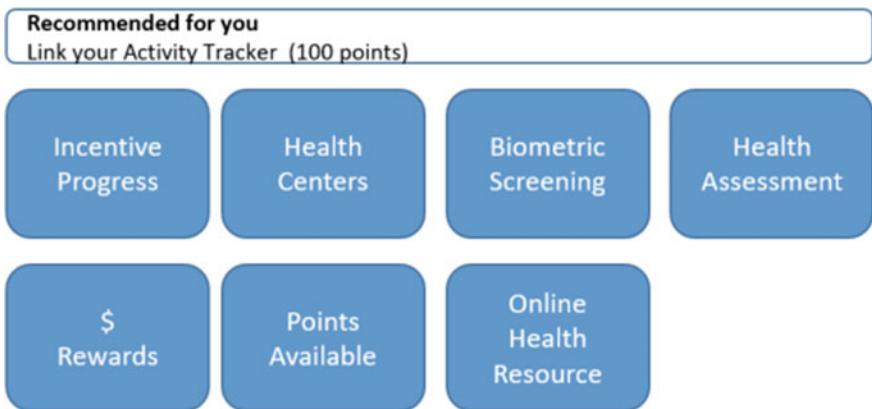
The most recent developments have seen the use of artificial intelligence to cope with the masses of fast-flowing data, beyond the capability of human analysts. Massive data processing has been applied to generate embedded learning algorithms to automate many processes.

One source of all of this data is the Internet of Things. Not only do people send messages now cars, phones, and machines communicate with each other (Kellmereit and Obodovski 2013). This enables much closer monitoring of patient health, to include little wristbands to monitor the wearer's pulse, temperature, blood pressure forwarded on to the patient's physician. How people ever survived until 2010 is truly a wonder. But it does indicate the tons of data in which a minuscule bit of important data exists.

To elaborate on the information provided by personal health devices, they not only monitor the patient health, but support or enable **management and education** of the patient or employee to achieve a healthier life style. Companies in the US now frequently offer workplace health programs that provide financial awards for a healthier life style. The employee may link a fitness tracker by an app to the company wellness program and can earn financial rewards based on how many

**Table 1.1** Eras of analytical activity

	Era (Davenport)	Specific meaning
Analytics 1.0 artisanal descriptive	1970–1985	Data analysis to support decision making
Analytics 2.0 big data	1990–2000	Search capabilities Recommendation
Analytics 3.0 data economy	2000–2010	Automated data mining
Analytics 4.0 artificial intelligence	2010–now	Large, unstructured, fast-moving data



**Fig. 1.1** Example of a wellness program

steps the employee walked in a day (see Fig. 1.1). Such programs might encourage employees to better adhere to specific medication or self-managed care guidelines. The lifestyles of people in the workforce are important both for the sake of their own health and for the sake of their employer's productivity. Companies often subsidize these programs in the hope that they will save companies money in the long run by improving health, morale and productivity of its employees. Such programs are “reward programs” but are on the other side excluding employees from financial benefits, who may not want to share their daily activity, monitored by a fitbit, with their employer!

Monitors in homes can reduce electricity use, thus saving the globe from excessive warming. Cars can send signals to dealers about engine problems, so that they might send a tow truck to the location provided by the car's GPS. Insurance companies already advertise their ability to attach devices to cars to identify good drivers, a euphemism for detection of bad driving so that they can cancel policies more likely to call for claims.

## Examples of Knowledge Management

There have been impressive accomplishments using big data. Google detected the SARS epidemic much sooner than the US health system (Brynjolfsson and McAfee 2014). The Harvard Medical School found Tweets to be as accurate as official reports in tracking cholera after the 2010 Haitian earthquake, and two weeks faster. Movie firms have found Tweets to be good at predicting box-office revenues.

Sports have long been a major source of interest, going back to gladiators in Rome if not earlier. A great deal of effort has been expended by those interested in the relative speed of horses, activity that we might consider rudimentary data

mining. Just as card playing was a major factor in the development of the science of statistics, the interest in sports continues, with casinos offering odds on sporting events, and the Web offering social venues for those interested in Fantasy Football. Baseball has always been considered one of the most measured of sporting activities. The Web has seen sites such as baseball.reference.com, providing access of detailed statistics to all. Sabermetricians have added all kinds of new statistics to baseball, and a new industry in applying statistics such as On-Base Percentage, Slugging Percentage, and Wins Against Replacement (WAR) have replaced traditional metrics like batting averages, home runs, and wins. We are told that each major league baseball team now employs a staff of dozens to measure performance of players and prospects (Law 2017). Basketball also has seen much greater emphasis on new statistics, such as plus/minus tracking. Sports tracking could also be expanded to support (replace) officials. Television provides better views of play than soccer referees have, but they can't use replays (it would really slow the game down). A great deal of replay is used in football and basketball. The technology exists to call balls and strikes in baseball.

Wu et al. (2014) provided a knowledge management framework for the product lifecycle, to include classification of knowledge types:

- Customer knowledge—CRM focus in data mining terms;
- Development knowledge—product design involving engineering expertise;
- Production knowledge—knowledge of production processes;
- Delivery & Service knowledge—knowledge of the processes needed to serve customers.

Knowledge of customers is a classical customer profiling matter. The other three bullets are classical business process reengineering matters, often involving tacit knowledge which organizations generate in the form of their employees' expertise. Management of these forms of knowledge require:

- A mechanism to identify and access knowledge;
- A method for collaboration to identify who, how, and where knowledge is;
- A method to integrate knowledge for effectively making specific decisions.

Data can be found in statistics of production measures, which accounting provides and which industrial engineers (and supply chain managers) analyze for decision making. Knowledge also exists in the experience, intuition, and insight found in employees (tacit information). This tacit knowledge includes organizational value systems. Thus expression of such knowledge is only available through collaboration within organizations. With respect to knowledge management, it means that the factual data found in accounting records needs to be supplemented by expertise, and a knowledge management system is closely tied to the idea of business process mapping. Business process mapping in turn is usually expressed in the form of a flowchart of what decisions need to be made, where knowledge can be found, and the approval authority in the organizations control system.

Kellmereit and Obodovski (2013) viewed this brave new world as a platform for new industries, around intelligent buildings, long-distance data transmission, and expansion of services in industries such as health care and utilities. Humans and machines are contended to work best in tandem, with machines gathering data, providing analytics, and applying algorithms to optimize or at least improve systems while humans provide creativity. (On the other hand, computer scientists such as Ray Kurzweil (2000) expect machines to develop learning capabilities circa 2040 in the Great Singularity). Retail organizations (like Wal-Mart) analyze millions of data sets, some fed by RFID signals, to lower costs and thus serve customers better.

Use of all of this data requires increased data storage, the next link in knowledge management. It also is supported by a new data environment, allowing release from the old statistical reliance on sampling, because masses of data usually preclude the need for sampling. This also leads to a change in emphasis from hypothesis generation and testing to more reliance on pattern recognition supported by machine learning. A prime example of what this can accomplish is **customer relationship management**, where every detail of company interaction with each customer can be stored and recalled to analyze for likely interest in other company products, or management of their credit, all designed to optimize company revenue from every customer.

Knowledge is defined in dictionaries as the expertise obtained through experience or education leading to understanding of a subject. Knowledge acquisition refers to the processes of perception, learning, and reasoning to capture, structure, and represent knowledge from all sources for the purpose of storing, sharing, and implementing this knowledge. Our current age has seen a view of a knowledge being used to improve society.

Knowledge discovery involves the process of obtaining knowledge, which of course can be accomplished in many ways. Some learn by observing, others by theorizing, yet others by listening to authority. Almost all of us learn in different combinations of these methods, synthesizing different, often conflicting bits of data to develop our own view of the world. Knowledge management takes knowledge no matter how it is discovered and provides a system to provide support to organizational decision making.

In a more specific sense, knowledge discovery involves finding interesting patterns from data stored in large databases through use of computer analysis. In this context, the term **interesting** implies non-trivial, implicit, previously unknown, easily understood, useful and actionable knowledge. **Information** is defined as the patterns, correlations, rules, or relationships in data providing knowledge useful in decision making.

We live in an age swamped with data. As if satellite feeds of weather data, military intelligence, or transmission of satellite radio and television signals weren't enough, the devices of the modern generation including Twitter, Facebook, and their many competitors flood us with information. We sympathize with the idea that parents can more closely monitor their children, although whether these children

will develop normally without this close supervision remains to be seen. But one can't help wonder how many signals containing useless information clutter up the lives of those with all of these devices.

## Data Mining Descriptive Applications

Knowledge management consists of the overall field of human knowledge (epistemology) as well as means to record and recall it (computer systems) and quantitative analysis to understand it (in business contexts, business analytics). There are many applications of quantitative analysis, falling within the overall framework of the term business analytics. Analytics has been around since statistics became widespread. With the emergence of computers, we see three types of analytic tools. **Descriptive** analytics focus on reports of what has happened. Statistics are a big part of that. Descriptive models are an example of unsupervised learning, where the algorithm identifies relationships without user direction. They don't predict some target value, but rather try to provide clues to data structure, relationships, and connectedness. **Predictive** analytics extend statistical and/or artificial intelligence to provide forecasting capability. They are directed in the sense that a target is defined. This can be a continuous variable to forecast. It also includes categorical output, especially classification modeling that applies models to suggest better ways of doing things, to include identification of the most likely customer profiles to send marketing materials, or to flag suspicious insurance claims, or many other applications. **Diagnostic** analytics can apply analysis to sensor input to direct control systems automatically. This is especially useful in mechanical or chemical environments where speed and safety considerations make it attractive to replace human monitors with automated systems as much as possible. It can lead to some problems, such as bringing stock markets to their knees for short periods (until humans can regain control). **Prescriptive** analytics applies quantitative models to optimize systems, or at least to identify improved systems. Data mining includes descriptive and predictive modeling. Operations research includes all three. This book focuses on the forecasting component of predictive modeling, with the classification portion of prescriptive analytics demonstrated.

Predictive modeling is well-studied, starting with linear regression but extending through autoregressive integrated moving-average (ARIMA) to generalized autoregressive conditional heteroscedasticity (GARCH) models and many variants. The crux of data mining modeling is classification, accomplished by logistic regression, neural network, and decision tree prescriptive models. They are covered in other books. Prescriptive analytics is an emerging field, with many interesting developments to come. This book presents begins with discussion of visualization, a simple but important step in data mining. Business analytics starts with getting data, which can come from many sources, internal or external to organizations. We do not delve into the even more important aspect of data transformation and cleansing, where the bulk of data mining work occurs. We do provide simple

explanations of some descriptive tools that are more machine learning, supporting the evolving field of pattern recognition. Cluster analysis is well-known from statistical theory to identify groupings within datasets. We will then look at some methods evolving from marketing, to include recency, frequency, and monetary expenditures (RFM) models, an early and simplified means of classification. Link analysis deals with making sense of social networks.

## Summary

The primary purpose of knowledge management is to wade through all of this noise to pick out useful patterns. That is data mining in a nutshell. Thus we view knowledge management as:

- Gathering appropriate data
  - Filtering out noise
- Storing data (DATABASE MANAGEMENT)
- Interpret data and model (DATA MINING)
  - Generate reports for repetitive operations
  - Provide data as inputs for special studies

Descriptive modeling are usually applied to initial data analysis, where the intent is to gain initial understanding of the data, or to special kinds of data involving relationships or links between objects.

## References

- Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Co, New York
- Davenport TH (2014) Big data at work. Harvard Business Review Press, Boston
- Davenport TH (2018) From analytics to artificial intelligence. J Bus Analytics 1:1
- Kellmereit D, Obodovski D (2013) The silent intelligence: the internet of things. DnD Ventures, San Francisco
- Kurzweil R (2000) The age of spiritual machines: when computers exceed human intelligence. Penguin Books, New York
- Law K (2017) Smart baseball: the story behind the old stats that are ruining the game, the new ones that are running it, and the right way to think about baseball. William Morrow, New York
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers HA (2011) Big data: the next frontier for innovation, competition and productivity. McKinsey Global Institute, New York
- Olson DL, Courtney JF Jr (1992) Decision support models and expert systems. MacMillan Publishing Co, New York
- Rothberg HN, Erickson GS (2005) From knowledge to intelligence: creating competitive advantage in the next economy. Elsevier Butterworth-Heinemann, Woburn, MA

- Sprague RH, Carlson ED (1982) Building effective decision support systems. Prentice-Hall, Englewood Cliffs, NJ
- Waller MA, Fawcett SE (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *J Bus Logistics* 34(2):77–84
- Wu ZY, Ming XG, Wang YL, Wang L (2014) Technology solutions for product lifecycle knowledge management: framework and a case study. *Int J Prod Res* 52(21):6314–6334

# Chapter 2

## Data Visualization



Data and information are important resources to be managed in modern organizations. Business analytics refers to the skills, technologies, applications and practices for exploration and investigation of past business performance to gain insight and aid business planning. The focus is on developing new insights and understanding based on data and statistical analysis. The emphasis is on fact-based management to drive decision making.

Data visualization is an important aspect of decision maker and/or business analyst learning. There are many useful visualization tools offered by geographic information systems, which quickly plot data by map, usually by county. These are highly useful in politics, as well as in other forms of marketing, to include tracking where sales of different products might occur. They are also useful for law enforcement, seeking to identify hot areas of particular problems. This chapter will review the visualization tools offered by the open source data mining software R, and will demonstrate simple Excel models of time series data. These are meant as representative demonstrations—there clearly are many visualization tools offered by many different software products. All support the very important process of initial understanding of data relationships.

## Data Visualization

There are many excellent commercial data mining software products, although these tend to be expensive. These include SAS Enterprise Miner and IBM's Intelligent Miner, as well as many more recent variants and new products appearing regularly. Two sources of information that are useful are [www.kdnuggets.com](http://www.kdnuggets.com) under “software” and <https://www.predictiveanalyticstoday.com/> which contains thorough coverage of contemporary business analytics software. Some of these are free. The most popular software by [rdstats.com/articles/popularity](http://rdstats.com/articles/popularity) (February 2016) by product are shown in Table 2.1.

**Table 2.1** Data mining software by popularity (rdstats.com)

Rank		
1	R	Open source
2	SAS	Commercial
3	SPSS	Commercial
4	WEKA	Open source
5	Statistica	Commercial
5	Rapid Miner	Commercial

Rattle is a GUI system for R (also open source), and is also highly recommended. WEKA is a great system but we have found issues with reading test data making it a bit troublesome.

## R Software

Almost every data mining software provides some support in the form of visualization of data. We can use R as a case in point.

To install R, visit <https://cran.rstudio.com/>

Open a folder for R

Select Download R for windows

To install Rattle:

Open the R Desktop icon (32 bit or 64 bit) and enter the following command at the R prompt. R will ask for a CRAN mirror. Choose a nearby location.

```
> install.packages("rattle")
```

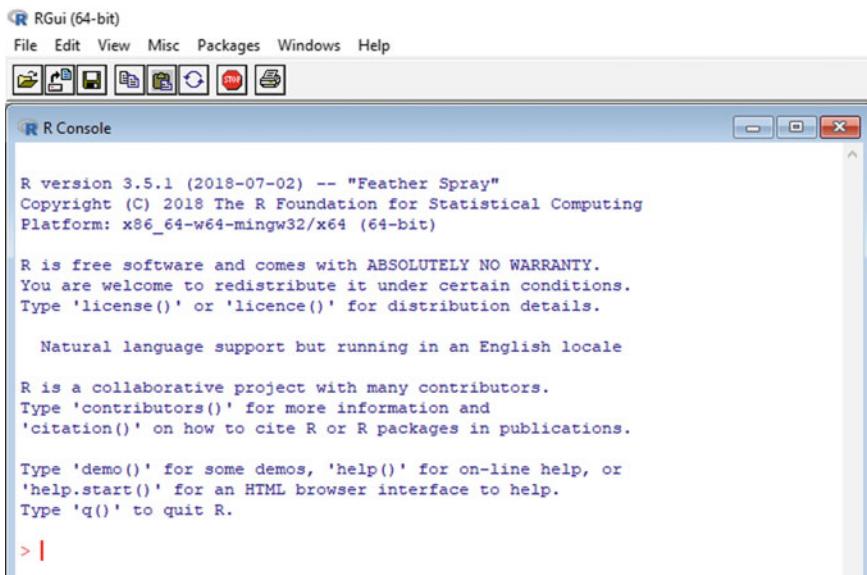
Enter the following two commands at the R prompt. This loads the Rattle package into the library and then starts up Rattle.

```
> library(rattle)
```

```
> rattle()
```

If the RGtk2 package has yet to be installed, there will be an error popup indicating that libatk-1.0-0.dll is missing from your computer. Click on the OK and then you will be asked if you would like to install GTK+. Click OK to do so. This then downloads and installs the appropriate GTK+ libraries for your computer. After this has finished, do exit from R and restart it so that it can find the newly installed libraries.

When running Rattle a number of other packages will be downloaded and installed as needed, with Rattle asking for the user's permission before doing so. They only need to be downloaded once.



**Fig. 2.1** R console

The installation has been tested to work on Microsoft Windows, 32bit and 64bit, XP, Vista and 7 with R 3.1.1, Rattle 3.1.0 and RGtk2 2.20.31. If you are missing something, you will get a message from R asking you to install a package. I read nominal data (string), and was prompted that I needed “stringr”. On the R console (see Fig. 2.1), click on the “Packages” word on the top line.

Give the command “Install packages” which will direct you to HTTPS CRAN mirror. Select one of the sites (like “USA(TX) [https]”) and find “stringr” and click on it. Then upload that package. You may have to restart R.

## ***Loan Data***

This data set consists of information on applicants for appliance loans. The full data set, taken from a previous text (Olson and Shi 2007), involves 650 past observations. Applicant information on age, income, assets, debts, and credit rating (from a credit bureau, with red for bad credit, yellow for some credit problems, and green for clean credit record) is assumed available from loan applications. Variable Want is the amount requested in the appliance loan application. For past observations, variable On-Time is 1 if all payments were received on time, and 0 if not (Late or Default). The majority of past loans were paid on time. Asset, debt, and loan amount (variable Want) are used by rule to generate categorical variable risk. Risk was categorized as high if debts exceeded assets, as low if assets exceeded the sum of debts plus the borrowing amount requested, and average in between.

An extract of Loan Data is shown in Table 2.2.

In Fig. 2.2, 8 variables are identified from the file LoanRaw.csv. By default, Rattle will hold out 30% of the data points for testing or other purposes by default. That would leave 448 observations. We could include all if we wished, but we proceed with this training set. In Fig. 2.2 we specify 70% training, 15% validation, and 15% testing. Variables 3, 4 and 5 are used to calculate variable credit, so they are duplications. By clicking the “Ignore” radio button, these variables are deleted from analysis. The outcome variable is variable 8, “On-time”, so the user should make sure that the Target radio button is highlighted. When the user is satisfied with the variables for analysis, the **Execute** button on the top ribbon can be selected.

Next the **Explore** tab can be selected, which provides Fig. 2.3, where the user can select various visualization displays.

Figure 2.3 provides basic statistics for continuous variables, and the number of categories for categorical data. It has a number of visualization tools to examine each variable. In Fig. 2.3, we selected box plots for Age and Income. Selecting **Execute** yields Fig. 2.4.

Figure 2.4 gives an idea for the range and distribution of Age and Income. Each variable’s box plot is displayed using All data points (454 in this case), as well as by outcome variable result. We can see from Fig. 2.4 that borrowers that repaid on time have older observations, while problem borrowers have age distribution that is younger.

We can also explore categorical variables through bar plots in Fig. 2.5:

**Table 2.2** Extract of loan data

Age	Income	Assets	Debts	Want	Risk	Credit	Result
20	17,152	11,090	20,455	400	high	Green	On-time
23	25,862	24,756	30,083	2300	high	Green	On-time
28	26,169	47,355	49,341	3100	high	Yellow	Late
23	21,117	21,242	30,278	300	high	Red	Default
22	7127	23,903	17,231	900	low	Yellow	On-time
26	42,083	35,726	41,421	300	high	Red	Late
24	55,557	27,040	48,191	1500	high	Green	On-time
27	34,843	0	21,031	2100	high	Red	On-time
29	74,295	88,827	100,599	100	high	Yellow	On-time
23	38,887	6260	33,635	9400	low	Green	On-time
28	31,758	58,492	49,268	1000	low	Green	On-time
25	80,180	31,696	69,529	1000	high	Green	Late
33	40,921	91,111	90,076	2900	average	Yellow	Late
36	63,124	164,631	144,697	300	low	Green	On-time
39	59,006	195,759	161,750	600	low	Green	On-time
39	125,713	382,180	315,396	5200	low	Yellow	On-time
55	80,149	511,937	21,923	1000	low	Green	On-time
62	101,291	783,164	23,052	1800	low	Green	On-time
71	81,723	776,344	20,277	900	low	Green	On-time
63	99,522	783,491	24,643	200	low	Green	On-time

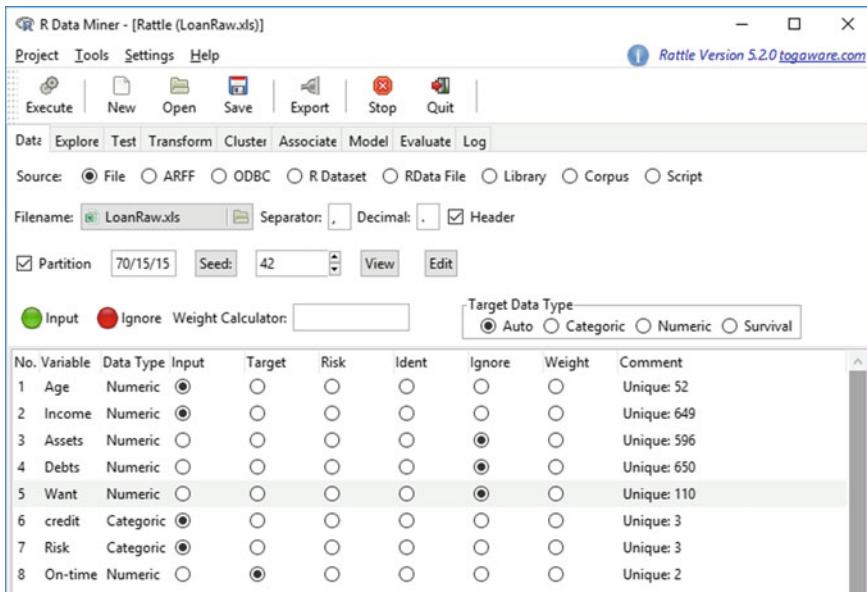


Fig. 2.2 Loading data file in R

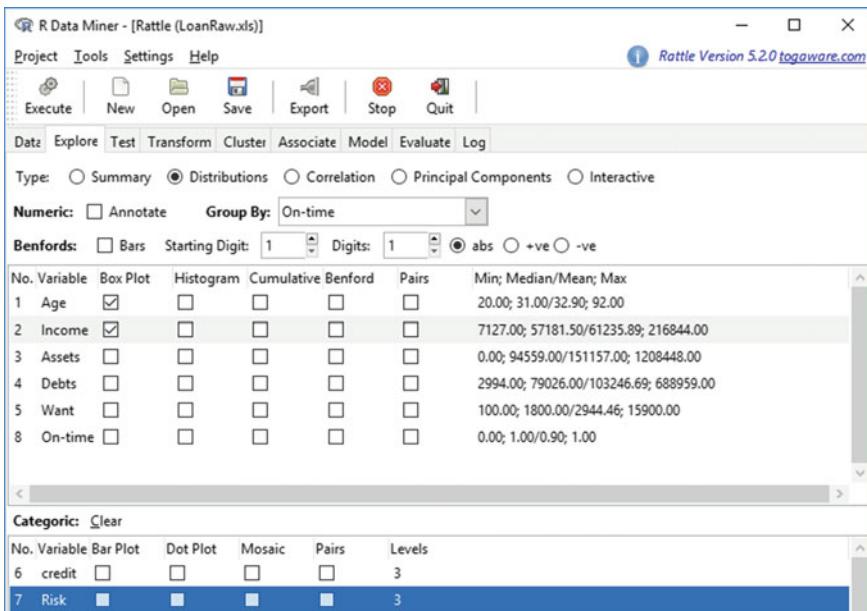


Fig. 2.3 Initial data visualization in R

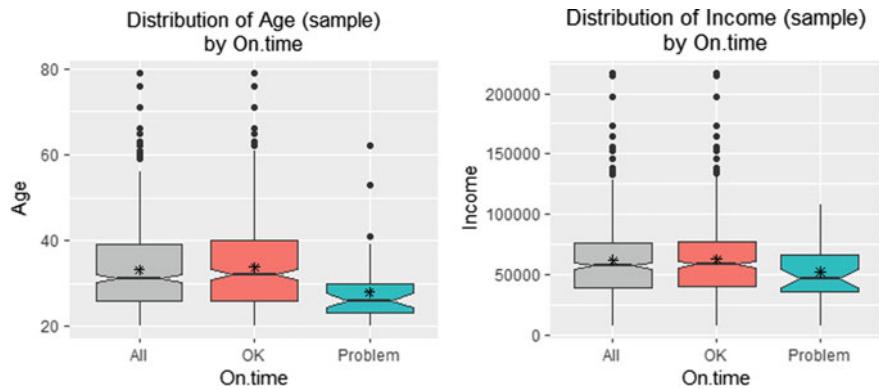


Fig. 2.4 Distribution visualization for continuous input variables

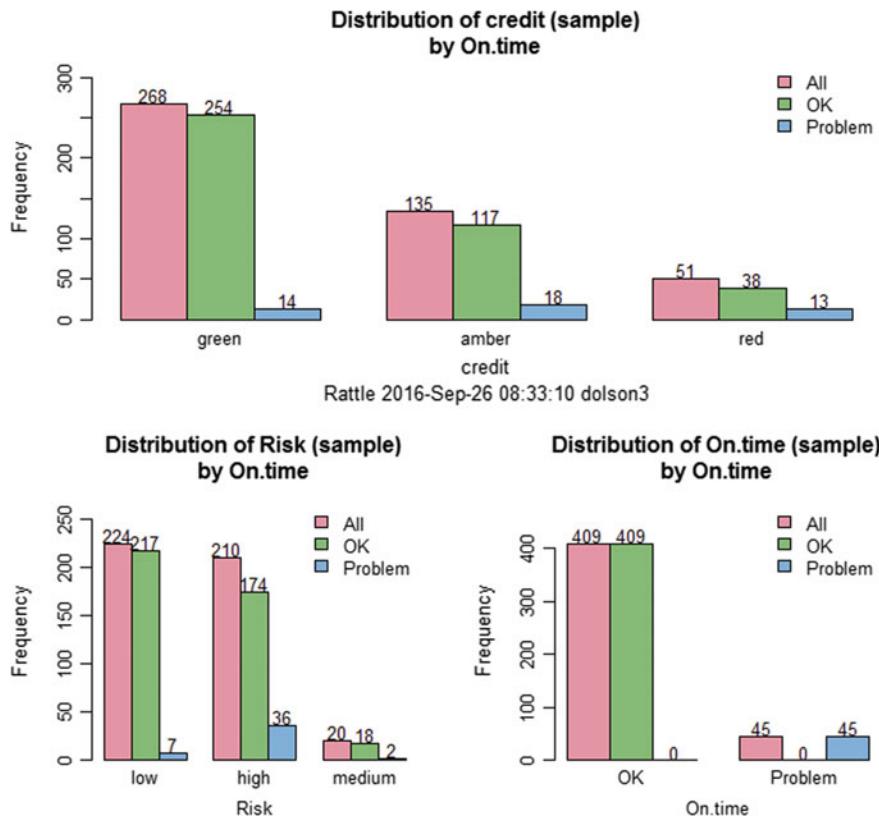


Fig. 2.5 Categorical data visualization in R

Figure 2.5 shows the distributions of variables Credit, Risk, and On-time by categories. The On-time display is tautological, as the 409 OK outcomes by definition were OK. But the displays for variables Credit and Risk show the difference in outcome by each category. With respect to credit, there is a higher proportion of problem loans for category “red” than for either “amber” or “green”. For variable Risk, the proportion of problems is clearly worse for “high” risk than for “medium” or “low” (as would be expected).

Another visualization option is **Mosaic**. Figure 2.6 shows the display using this tool for Credit and for Risk.

It says the same thing as Fig. 2.5, but in a different format.

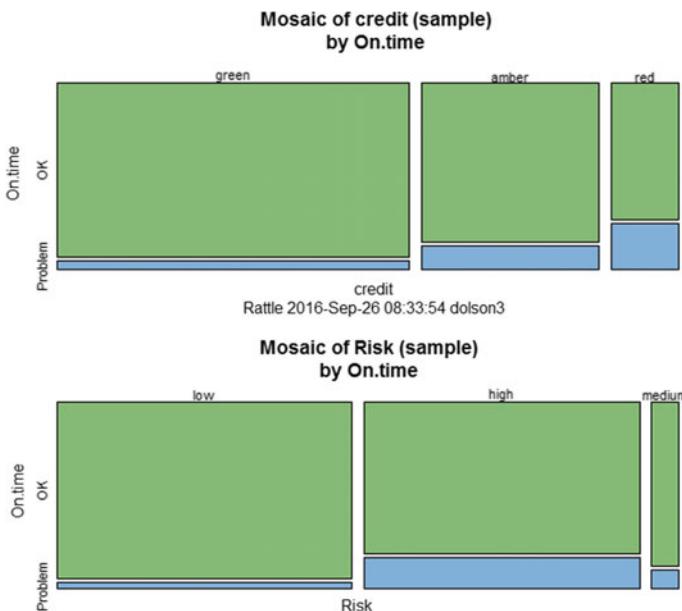
**Histograms** is another nice feature from Rattle. Figure 2.7 displays histograms for Age and Income, as well as a joint scatter plot and correlation obtained by checking **Pairs**.

This output provides the same information content as Figs. 2.5 and 2.6, but provides another way to show users data relationships.

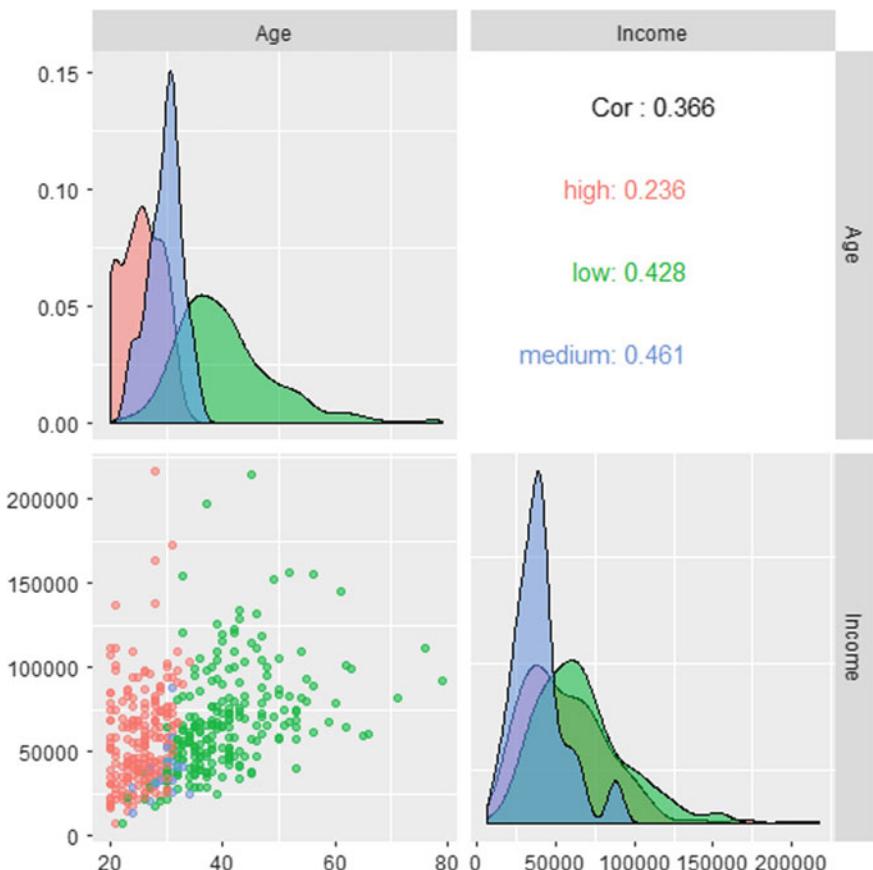
Rattle also provides Principal Components display. This is called up as shown in Fig. 2.8.

The output obtained after selecting **Execute** is shown in Fig. 2.9.

The output is a mass of observations on the two principal component vectors obtained from the algorithm. Input variables Age and Income provide a bit of frame



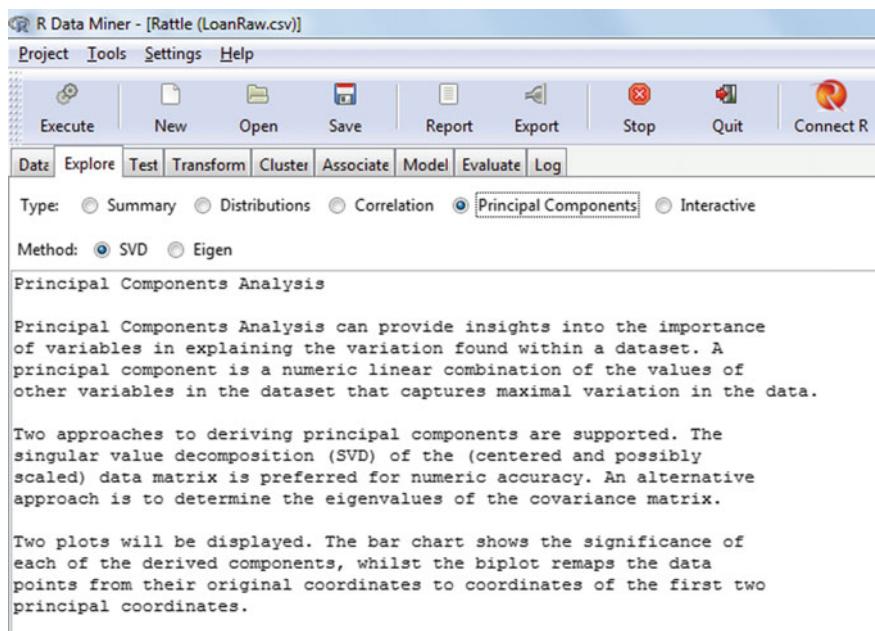
**Fig. 2.6** Mosaic plot from R



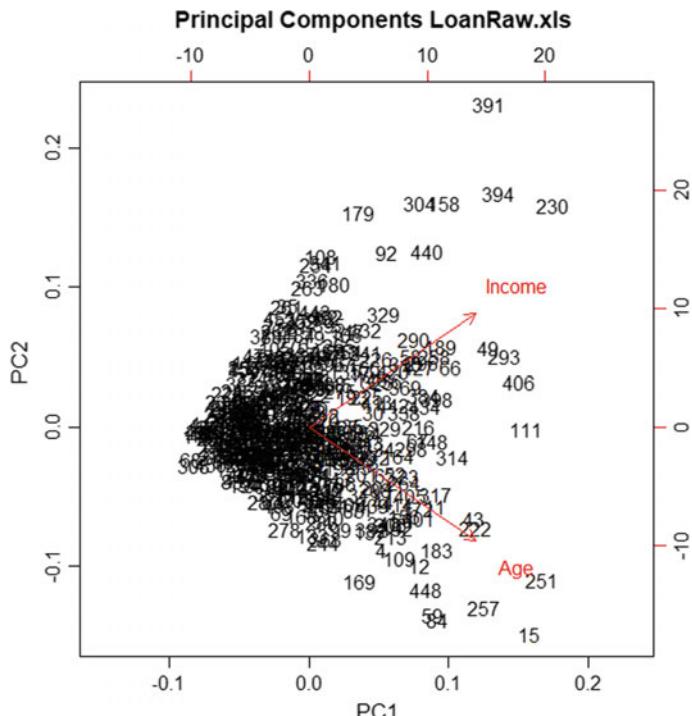
**Fig. 2.7** Histogram and Pairs visualization in R

of reference, showing those outlying observations that are extreme. For instance, observation 104 has high income but average age, while observation 295 is the reverse. In the data set, observation 104 the age was 45, and income 215,033, while for observation 295 the age was 79 and income 92,010. From Fig. 2.3, we see that mean age was 32.90 and mean income 61,236. This provides some frame of reference for the Principal Components output.

There are other tools to aid data visualization. One of the most important is correlation, which will be covered in depth in Chap. 6, cluster analysis. Here we will redirect attention to basic data display obtained from simpler software tools. We will use Excel tools to obtain a visualization of energy data.



**Fig. 2.8** Principal components description from rattle



**Fig. 2.9** Principal components plot for loan data in R

## Energy Data

Energy is one of the major issues facing our society. The United States has prospered by utilizing a complex system of energy. Early US development relied on mills driven by hydro-power. Late in the 19th Century Edison and Tesla brought commercial grade electricity to cities, and ultimately to rural communities. John D. Rockefeller organized the oil industry, providing complex derivatives that have proven useful in many contexts. This includes a transportation system based on gasoline and diesel driven vehicles.

There is much disagreement about energy policy. First, there was a strong opinion that the world's supply of crude oil was limited, and we were about to run out ([Deffeyes 2001](#)). But as oil prices soared and then in early summer 2008 plummeted, the pendulum returned with fracking, new prospects from Canada and Brazil, and the Bakken field in North Dakota coming on line. Furthermore, while some analysts argued that Saudis were running out of oil, Saudis kept quiet and appear to have lots of oil.

The second point relates to global warming. Clearly there is a rise in sea levels, as glaciers and small islands in the Pacific disappear, and we actually now can ship goods over the Northwest Passage of North America. This is caused in part by carbon emissions. One solution is to retain the current infrastructure where we can all drive our own automobiles with cleaner petroleum products, and run our coal electric plants with cleaner coal. Others feel that we need to scrap our current culture and eliminate carbon energy sources. This is a political issue that will not go away, and will drive election outcomes around the world into the foreseeable future. It is a fact of nature that people will disagree. Unfortunately, armaments often are involved more and more. Thus it is very important to understand the energy situation.

Those opposed to all carbon emissions propose wind and solar power. The US Government provided support to enterprises such as Solyndra to create viable solar power generation. This, however, has encountered problems, and Solyndra filed for bankruptcy in August 2011 ([Meiners et al. 2011](#)). There appear to be problems in translating what is physically possible with economic viability. Wind energy also involves some issues. If you fly into northern Europe, you can often see fields of 100 very large wind turbines just off-shore. They also exist off Brazil. There are more and more wind farms in the US as well. But they have annoying features for nearby residents, kill birds, and like solar energy, are not capable of continuous generation of power.

Another non-carbon based energy source is nuclear. In the 1950s there was strong movement toward penny-cheap electricity generated by turning the sword of nuclear weapons into the plowshare of nuclear power plants. Nearly 100 such plants were built in the United States. But problems were encountered in 1979 at Three Mile Island in Pennsylvania ([Levi 2013](#)), and then the catastrophe at Chernobyl in the Ukraine in 1986. People don't want nuclear plants anymore, and what was a cheap source of power became expensive after the Federal Government insisted of

retrofitting plants to provide very high levels of protection. There also is the issue of waste disposal that has become a major political issue, especially in Nevada.

Thus there are a number of important issues related to generation of energy in the United States, as well as in the world. The US Department of Energy provides monthly reports that provide an interesting picture that we might visualize with simple Excel tools. The source of this data over time is: <http://www.eia.gov/totalenergy/data/monthly/>.

One of the obvious ways to visualize time series data is to plot it over time. Table 2.3 displays a recap of US energy production, displaying the emergence of geothermal energy in 1970, and solar and wind energy in 1990. Nuclear power didn't come on line until the late 1950s, and has slowly grown over the years. Natural gas plant liquids (NGPL) was small in 1950, but has exceeded hydropower. Crude oil has been highly variable. The volumes for hydroelectricity has been fairly steady, and relatively small compared to natural gas, crude oil, and coal.

## Basic Visualization of Time Series

Figure 2.1 displays the growth in carbon versus non-carbon production. The data comes from Table 2.3, reorganized by adding columns Coal through NGPL for Carbon, columns Nuclear through Biomass for Non-Carbon. Using the Year column of Table 2.3 as the X-axis, Fig. 2.10 displays an Excel plot for the Carbon based total versus the Non-carbon based total by year.

Figure 2.10 shows that there has been more variance in Carbon-based production of energy. There was a peak around 1965, with a drop in 1970 as the economy slowed. Growth resumed in 1975, but there was a decline due to massive increases in oil price due to OPEC and the Iranian crisis. Carbon-based production actually languished and even dropped around 2000. At that time there was a predominant thought that we had reached a “peak oil” point, where the amount of reserves available was finite and would soon run out (Simmons 2005). However, Saudi’s continued to contend that they had plenty of reserves, and in the US, a new major field in North Dakota was brought into production, while fracking increased output from old oil fields. The US then switched from a net oil importer to a net oil exporter. Figure 2.10 shows that oil production is increasing. Table 2.3 shows that coal production is dropping, but natural gas and crude production are growing significantly. This information is displayed graphically in Fig. 2.11.

As to alternative energy, Fig. 2.12 shows the small level of US energy production from wind and solar energy, the primary alternative energy sources.

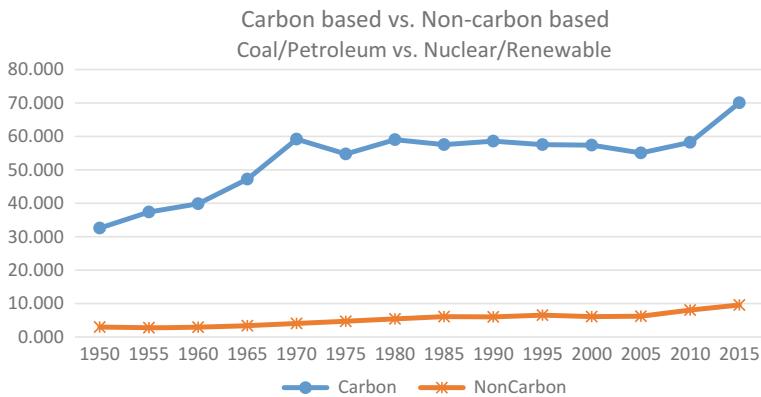
Table 2.4 gives US Annual consumption in trillion BTUs by sector.

Figure 2.13 displays US energy use by sector graphically.

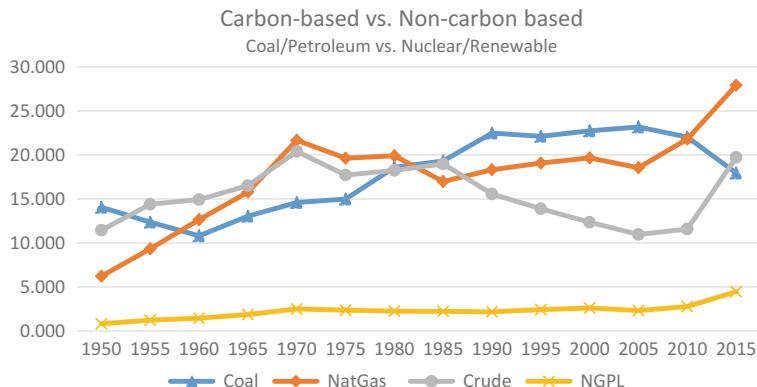
Looking at Fig. 2.13 we can see that all sectors have grown, with a bit of a dip around the 2008 economic crisis. The most volatile is the largest sector, industrial, which suffered downturns around the 1973 OPEC emergence (which raised the price of oil substantially, bringing on intensive inflation for the rest of the 1970s),

Table 2.3 US energy production

Year	Coal	NatGas	Crude	NGPL	Nuclear	Hydro	Geotherm	Solar	Wind	Biomass	Total
1950	14.06	6.23	11.45	0.82	0.00	1.42	0.00	0.00	0.00	1.56	35.54
1955	12.37	9.34	14.41	1.24	0.00	1.36	0.00	0.00	0.00	1.42	40.15
1960	10.82	12.66	14.93	1.46	0.01	1.61	0.00	0.00	0.00	1.32	42.80
1965	13.06	15.78	16.52	1.88	0.04	2.06	0.00	0.00	0.00	1.33	50.67
1970	14.61	21.67	20.40	2.51	0.24	2.63	0.01	0.00	0.00	1.43	63.50
1975	14.99	19.64	17.73	2.37	1.90	3.15	0.03	0.00	0.00	1.50	61.32
1980	18.60	19.91	18.25	2.25	2.74	2.90	0.05	0.00	0.00	2.48	67.18
1985	19.33	16.98	18.99	2.24	4.08	2.97	0.10	0.00	0.00	3.02	67.70
1990	22.49	18.33	15.57	2.17	6.10	3.05	0.17	0.06	0.03	2.74	70.70
1995	22.13	19.08	13.89	2.44	7.08	3.21	0.15	0.07	0.03	3.10	71.17
2000	22.74	19.66	12.36	2.61	7.86	2.81	0.16	0.06	0.06	3.01	71.33
2005	23.19	18.56	10.97	2.33	8.16	2.70	0.18	0.06	0.18	3.10	69.43
2010	22.04	21.81	11.59	2.78	8.43	2.54	0.21	0.09	0.92	4.32	74.72
2015	17.95	27.93	19.72	4.47	8.34	2.39	0.22	0.43	1.82	4.72	87.99



**Fig. 2.10** Carbon versus non-carbon US production



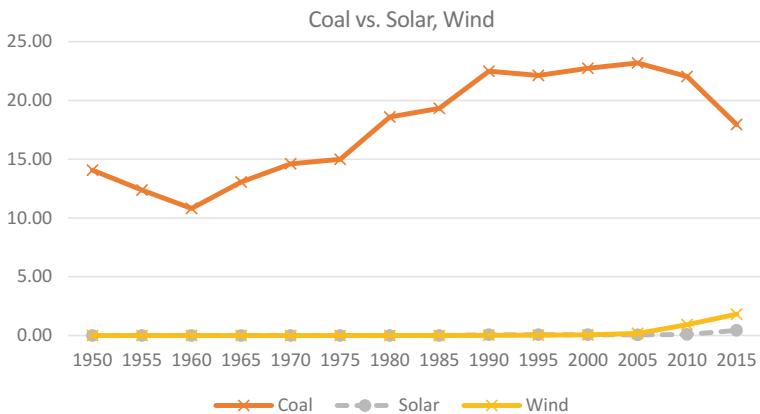
**Fig. 2.11** US production of carbon-based energy

and around 1985 when the Iranian crisis led to another major increase in the price of crude oil. There has been a noticeable slowing in industrial consumption since 2000. Residential consumption also has dipped a bit between 2010 and 2015. Transportation increased every period except around the 2008 global financial crisis.

Figure 2.14 shows another type of data display. In this case, the Department of Energy monitors detailed flows of production sources to consumption sectors.

This excellent graphic gives a picture of the complex flows in aggregate terms. We have taken the individual annual graphs (available since 2001) and extracted values for major inputs and outputs in Table 2.5.

This data for the source side of the equation (the left side of Fig. 2.14) is graphed in Excel to provide Fig. 2.15.



**Fig. 2.12** Comparison of coal, solar, and wind energy—US

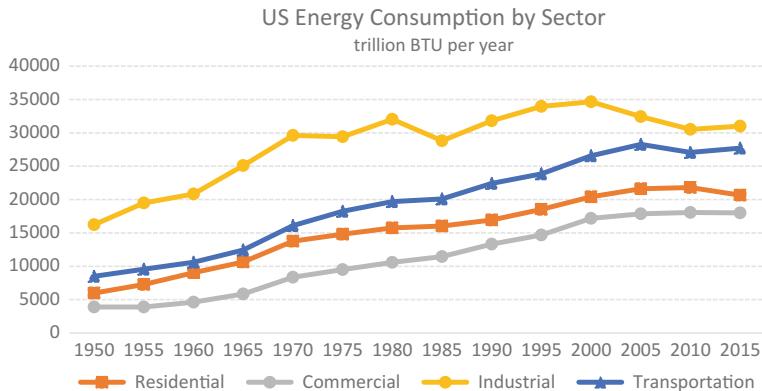
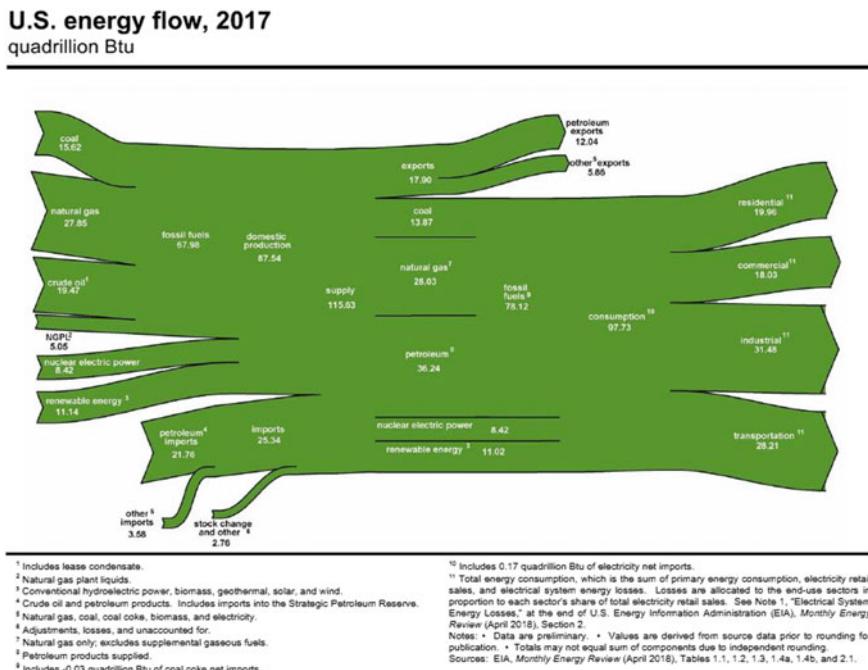
**Table 2.4** US trillion BTUs/year

Year	Residential	Commercial	Industrial	Transportation
1950	5989	3893	16,241	8492
1955	7278	3895	19,485	9550
1960	9039	4609	20,842	10,596
1965	10,639	5845	25,098	12,432
1970	13,766	8346	29,628	16,098
1975	14,813	9492	29,413	18,245
1980	15,753	10,578	32,039	19,697
1985	16,041	11,451	28,816	20,088
1990	16,944	13,320	31,810	22,420
1995	18,517	14,690	33,970	23,851
2000	20,421	17,175	34,662	26,555
2005	21,612	17,853	32,441	28,280
2010	21,793	18,057	30,525	27,059
2015	20,651	17,993	31,011	27,706

It can be seen from this graph that imports were the primary source of US energy, and were bothersome to the public, which feared that oil reserves had peaked. However, around 2011 fracking increased crude production, as did North Dakota oil. Imports correspondingly declined, and the United States now finds itself the leading oil producer in the world. This same information can be displayed for any given year by a pie chart, as in Fig. 2.16.

Figure 2.17 plots the right side of Fig. 2.14.

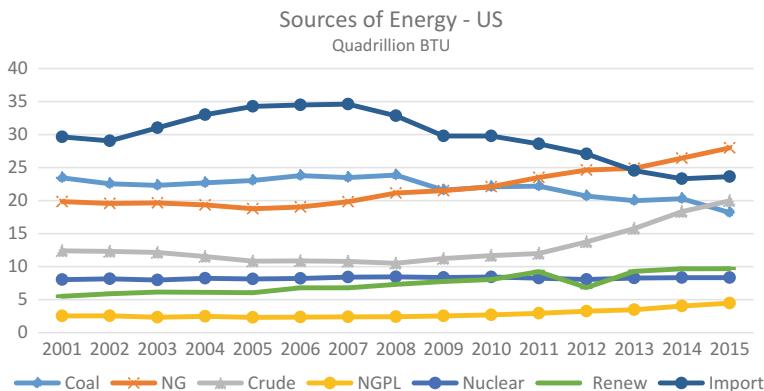
This same information for the year 2015 in pie chart form is shown in Fig. 2.18.

**Fig. 2.13** Plot of US energy consumption by sector**Fig. 2.14** US DOE display of energy flows in 2017

Consumption can be seen to be quite stable, with a drop in industrial consumption in 2009 (reflecting response to the 2008 financial crisis). Exports were quite low until 2007, after which time they have had steady increase.

**Table 2.5** US energy by source and sector

	Coal	NG	Crude	NGPL	Nuclear	Renew	Import	Resident	Comm	Indust	Trans	Export
2001	23.44	19.84	12.39	2.54	8.03	5.52	29.65	20.16	17.44	32.6	26.75	3.92
2002	22.55	19.56	12.31	2.56	8.15	5.9	29.04	20.94	17.4	32.49	26.52	3.65
2003	22.31	19.64	12.15	2.34	7.97	6.15	31.02	21.23	17.55	32.52	26.86	4.05
2004	22.69	19.34	11.53	2.47	8.23	6.12	33	21.18	17.52	33.25	27.79	4.43
2005	23.05	18.76	10.84	2.32	8.13	6.06	34.26	21.87	17.97	31.98	28.06	4.64
2006	23.79	19.02	10.87	2.35	8.21	6.79	34.49	21.05	18	32.43	28.4	4.93
2007	23.48	19.82	10.8	2.4	8.41	6.8	34.6	21.75	18.43	32.32	29.1	5.36
2008	23.86	21.15	10.52	2.41	8.46	7.32	32.84	21.64	18.54	31.21	27.92	7.06
2009	21.58	21.5	11.24	2.54	8.35	7.76	29.78	21.21	18.15	28.2	27.03	6.93
2010	22.08	22.1	11.67	2.69	8.44	8.06	29.79	22.15	18.21	30.14	27.51	8.17
2011	22.18	23.51	11.99	2.93	8.26	9.24	28.59	21.62	18.02	30.59	27.08	10.35
2012	20.68	24.63	13.76	3.25	8.06	6.84	27.07	20.08	17.41	30.77	26.75	11.36
2013	19.99	24.89	15.77	3.47	8.27	9.3	24.54	21.13	17.93	31.46	27.01	11.8
2014	20.28	26.43	18.32	4.03	8.33	9.68	23.31	21.53	18.34	31.33	27.12	12.22
2015	18.18	27.99	19.96	4.47	8.34	9.69	23.62	20.87	18.01	31.07	27.72	13.11



**Fig. 2.15** Plot of US energy sources

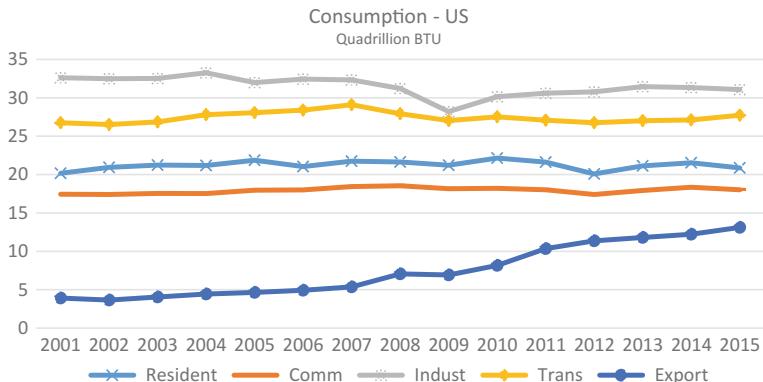
**Fig. 2.16** Pie chart of US energy sources in 2015



Figure 2.19 displays another Department of Energy graphic, in this case providing details between each major source and each major consumption sector.

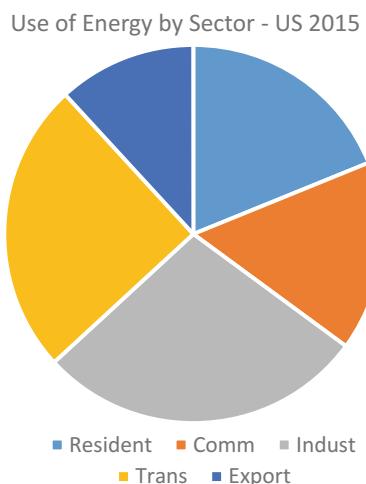
This data was extracted in tabular form in Table 2.6 by multiplying the given percentages by given quantities. Rounding leads to some imprecision.

Electrical power comes mostly from coal (about 34% in 2017), but this is dropping. This is due to Federal policy, seeking to shift power generation away from coal. That policy seeks to have growth occur in the renewable sector, but at the moment, that source is only 17% of electrical power generation (up from 13% in 2015). Transportation energy comes predominately from petroleum. Electric vehicles are being emphasized by current government policy, but we are far away



**Fig. 2.17** Plot of US energy consumption

**Fig. 2.18** US Energy consumption by sector in 2015

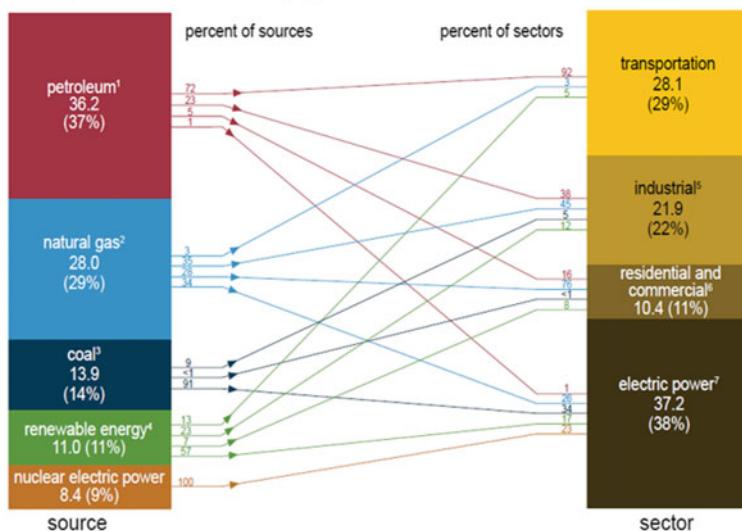


from a sustainable system of electrically powered vehicles. Furthermore, they would add to the need for electricity. Natural gas was discouraged by government policy in the 1970s, but has come back strong, and become a growing source of power that is flexible enough to deliver to a number of energy users.

## Conclusion

This chapter has brushed the tip of a very big ice berg relative to data visualization. Data visualization is important as it provides humans an initial understanding of data, which in our contemporary culture is overwhelming. We have demonstrated

**U.S. primary energy consumption by source and sector, 2017**  
 Total = 97.7 quadrillion British thermal units (Btu)



<sup>1</sup> Does not include biofuels that have been blended with petroleum—biofuels are included in “Renewable Energy.”

<sup>2</sup> Excludes supplemental gaseous fuels.

<sup>3</sup> Includes -0.03 quadrillion Btu of coal coke net imports.

<sup>4</sup> Commercial hydroelectric power, geothermal, solar, wind, and biomass.

<sup>5</sup> Includes industrial combined-heat-and-power (CHP) and industrial electricity-only plants.

<sup>6</sup> Includes commercial combined-heat-and-power (CHP) and commercial electricity-only plants.

<sup>7</sup> Electricity-only and combined-heat-and-power (CHP) plants whose primary business is to sell electricity, or electricity and heat, to the public. Includes 0.17 quadrillion Btu of electricity net imports not shown under “source.”

Notes: • Primary energy is energy in the form that it is accounted for in a statistical energy balance, before any transformation to secondary or tertiary forms of energy occurs (for example, coal is used to generate electricity). • The source total may not equal the sector total because of differences in the heat contents of total, end-use, and electric power sector consumption of natural gas. • Data are preliminary. • Values are derived from source data prior to rounding. • Sum of components may not equal total due to independent rounding.

Sources: U.S. Energy Information Administration, *Monthly Energy Review* (April 2018), Tables 1.3, 1.4a, 1.4b, and 2.1-2.6.



**Fig. 2.19** US DOE display of US energy flow in 2017

**Table 2.6** US energy flow—2017

2015	Qbillbtu	To TRANS	To IND	To RES/Com	To Elect
Petro	36.2	26.064	8.326	1.810	0.362
NG	28.0	0.840	9.800	7.840	9.520
Coal	13.9		1.251	0	12.649
Renew	11.0	1.430	2.530	0.770	6.270
Nuc	8.4				8.4
totals	97.5	28.334	21.907	10.420	37.201

two types of data visualization. Data mining software provides tools as were shown from R. Excel also provides many tools that are within the reach of the millions who use Microsoft products. Neither are monopolists, and there are other products that can provide more and better visualization support. These sources have the benefit of being affordable.

We spent a great deal of attention to a specific field of data, obtained by US Government publication. Most of the work of data mining involves getting data,

and then identifying which specific data is needed for the particular issue at hand. We have explored visualization of time series that can be informative relative to a number of important energy issue questions.

## References

- Deffeyes KS (2001) Hubbert's peak: the impending world oil shortage. Princeton University Press, Princeton, NJ
- Levi M (2013) The power surge: energy, opportunity, and the battle for america's future. Oxford University Press, Oxford
- Meiners RE, Morrise A, Bogart WT, Dorchak A (2011) The false promise of green energy. Cato Institute, Washington, DC
- Olson DL, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, New York
- Simmons MR (2005) Twilight in the desert: the coming Saudi oil shock and the world economy. Wiley, Hoboken, NJ

# Chapter 3

## Market Basket Analysis



Knowledge discovery is the effort to find information from data. In contemporary terms, it is the application of tools (from statistics and from artificial intelligence) to extract interesting patterns from data stored in large databases. Here **interesting** means non-trivial, implicit, previously unknown, and easily understood and described knowledge that can be used (**actionable**). One of the original applications of data mining to generate interesting knowledge for business was market basket analysis (Agrawal et al. 1993).

**Market-basket analysis** refers to methodologies studying the composition of a shopping basket of products purchased during a single shopping event. This technique has been widely applied to grocery store operations (as well as other retailing operations, to include restaurants). Market basket data in its rawest form would be the transactional list of purchases by customer, indicating only the items purchased together (with their prices). This data is challenging because of a number of aspects:

- A very large number of records (often millions of transactions per day)
- Sparseness (each market basket contains only a small portion of items carried)
- Heterogeneity (those with different tastes tend to purchase a specific subset of items).

The aim of market-basket analysis is to identify what products tend to be purchased together. Analyzing transaction-level data can identify purchase patterns, such as which frozen vegetables and side dishes are purchased with steak during barbecue season. This information can be used in determining where to place products in the store, as well as aid inventory management. Product presentations and staffing can be more intelligently planned for specific times of day, days of the week, or holidays. Another commercial application is electronic couponing, tailoring coupon face value and distribution timing using information obtained from market-baskets. There have been many applications beyond retail store analysis of customers. In the business realm, market basket analysis can aid decision making

relative to employee benefits, dysfunctional employee behavior, or identification of entrepreneurial talent. It has also been used in bioinformatics, pharmaceuticals, geophysics, and nuclear science (Aguinis et al. 2013).

## Definitions

Market-basket analysis examines the tendencies of customers to purchase items together. This can include buying products at the same time, such as milk and cookies, or bread, butter, and jam. It also can be sequential relationships, such as purchasing a house followed by purchases of furniture, or purchasing a car one year and purchasing new tires two years later. Knowledge of customer tendencies can be very valuable to retail organizations. Information about purchase timing can be useful. Monday night football purchases can be expected to motivate Monday afternoon sales, for instance, so stores may be wise to ensure ample supplies of beer and potato chips. Other information may not be useful, such as hypothesized relationships between new hardware store openings and toilet ring sales (Berry and Linoff 1997). This information has no actionable content. While hardware stores may wish to make sure that they have toilet rings on hand when they open, market basket information has no real value unless it contains information that can be explained (conclusions should make sense).

Market-basket analysis (along with clustering) is an undirected data mining operation, seeking patterns that were previously unknown. This makes it a form of knowledge discovery. Market-basket analysis begins with categorizing customer purchase behavior. The next step is to identify actionable information improving profit by purchase profile. Once profitability by purchase profile is known, retailers have factual data that can be used for key decision making. Laying out retail stores by purchase categories is referred to as **affinity positioning**. An example is to include coffee and coffee makers in the office products area. Affinity is measured a number of different ways, the simplest of which is correlation. If product positioning matters, a store choice model assumes that change in the mix of customers due to marketing activities lead to correlations. If product positioning doesn't matter, global utility models view cross—category dependence due to consumer choice. Either view finds value in knowing which products tend to be sold together. **Cross-selling** refers to the propensity for the purchaser of a specific item to purchase a different item. Retail outlets can maximize cross-selling by locating those products that tend to be purchased by the same consumer in places where both products can be seen. A good example of cross-selling is orange juice, cold medicine, and tissues, which are all attractive to consumers with colds. Market-basket analysis leads to identification of which products are being purchased together. Such information can be useful in more effectively laying out stores, or catalogs, as well as in selecting products for promotion. This information is typically used for advertising and promotion planning, space allocation, product placement, and personal customer relations.

Market-basket analysis can be vitally important in effective selection of promotions and merchandising strategies. Analysis can uncover buried consumer spending patterns, and identify lucrative opportunities to promote products together. Schmidt reported Italian entrees, pizza, bakery pies, Oriental entrees, and orange juice as products most sensitive to price promotion. Analysis has revealed a high correlation of orange juice and waffle sales. Data mining analysis includes the entire process of not only identifying such correlations, but also deriving a reason for such relationships, and more important, devising ways to increase overall profit.

## Co-occurrence

Simple market-basket analysis can be applied beginning with a **co-occurrence table**, listing the number of incidents out of a given sample size in which products are purchased together. For instance, six customers may have products in their grocery market baskets as given in Table 3.1.

This information considering five products is displayed in a binary format in Table 3.2.

The co-occurrence for these five selected products would be as shown in Table 3.3.

Table 3.3 can be explained by manual counting of Table 3.2. For example, beer and potato chips were found in purchases #1 and #6, giving a co-occurrence of 2. This can be done in Excel using the COUNTIFS function as shown in Fig. 3.1. Co-occurrence will be further addressed in Chap. 4.

**Table 3.1** Possible grocery market baskets

Customer #1	Beer, pretzels, potato chips, aspirin
Customer #2	diapers, baby lotion, grapefruit juice, baby food, milk
Customer #3	soda, potato chips, milk
Customer #4	soup, beer, milk, ice cream
Customer #5	soda, coffee, milk, bread
Customer #6	beer, potato chips

**Table 3.2** Possible grocery market baskets as displayed in binary format

Purchase	Beer	Potato chips	Milk	Diapers	Soda
1	1	1	0	0	0
2	0	0	1	1	0
3	0	1	1	0	1
4	1	0	1	0	0
5	0	0	1	0	1
6	1	1	0	0	0

**Table 3.3** Co-occurrence table

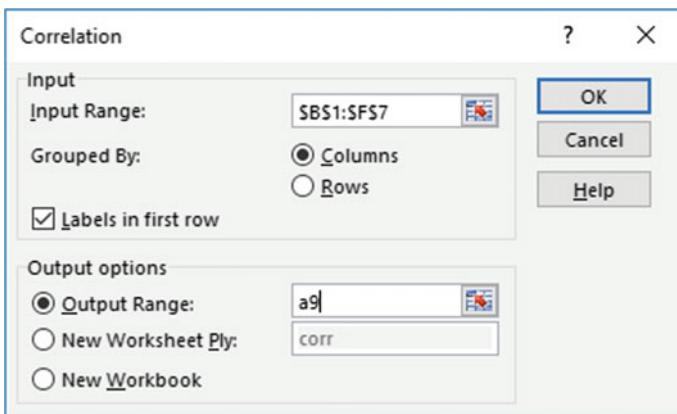
	Beer	Potato chips	Milk	Diapers	Soda
Beer	3	2	1	0	0
Potato chips	2	3	1	0	1
Milk	1	1	4	1	2
Diapers	0	0	1	1	0
Soda	0	1	2	0	2

The screenshot shows an Excel spreadsheet with two main sections. The top section, labeled 'purchase' (A1:A7), contains binary values (0 or 1) representing whether items were purchased together. The bottom section, labeled 'Co-occurrence' (B9:C12), uses the COUNTIFS function to calculate the count of purchases where both items are present. The formula in C11 is =COUNTIFS(B2:B7,1,B2:B7,1).

A	B	C	
1	purchase	Beer	Potato Chips
2	1	1	1
3	2	0	0
4	3	0	1
5	4	1	0
6	5	0	0
7	6	1	1

9	Beer	Potato Chips
10	=COUNTIFS(B2:B7,1,B2:B7,1)	=COUNTIFS(B2:B7,1,C2:C7,1)
11		
12		=COUNTIFS(C2:C7,1,C2:C7,1)

**Fig. 3.1** Excel COUNTIFS function to obtain the co-occurrence table**Fig. 3.2** Correlation matrix for co-occurrence table

Correlation can be obtained from Excel's Data Analysis Toolpak shown in Fig. 3.1, using the cell references given in Fig. 3.2. In this example, it the strongest correlation is between beer and potato chips. This is an expected relationship, because the two products go well together.

The screenshot shows an Excel spreadsheet with the Data tab selected. The top ribbon includes File, Home, Insert, Page Layout, Formulas, Data, and Review tabs. Under the Data tab, there are buttons for Get External Data, New Query, Refresh All, Connections, Properties, Edit Links, and Sort.

The spreadsheet contains two tables. The first table (rows 1-7) shows binary purchase data:

	A	B	C	D	E	F
1	<b>purchase</b>	Beer	Potato Chips	Milk	Diapers	Soda
2	1	1	1	0	0	0
3	2	0	0	1	1	0
4	3	0	1	1	0	1
5	4	1	0	1	0	0
6	5	0	0	1	0	1
7	6	1	1	0	0	0

The second table (rows 9-14) shows correlation coefficients:

	Beer	Potato Chips	Milk	Diapers	Soda	
10	Beer	1				
11	Potato Ch	0.333333	1			
12	Milk	-0.70711	-0.707106781	1		
13	Diapers	-0.44721	-0.447213595	0.316228	1	
14	Soda	-0.70711	-1.96262E-17	0.5	-0.31623	1

Fig. 3.3 Excel output for demonstration data

This yields the output shown in Fig. 3.3.

While the correlation between diapers and milk is low, every instance of diaper purchase in this small sample included milk (and for that matter, other baby products), things that would be expected. It appears that it would not make sense to seek cross-sales of milk or soda to beer drinkers. This information could be actionable in a negative sense to advertisers (or maybe beer drinkers just haven't thought of the value of milk in combination with beer). On the other hand, it would make sense to push potato chips to beer drinkers. Some combinations, such as between milk and soda, may not include any actionable content, even though the statistical relationship may be quite strong.

There are basic ways to identify which products in a market basket go together. Correlation was just demonstrated. However, correlation is not particularly good when dealing with binary data (and raw market basket data is binary). The product-moment correlation coefficient standardizes for mean and standard deviation. A second method, Jaccard's coefficient, is very simple but effective. The Jaccard coefficient is the ratio of the number of cases where two products were purchased together to the total number of cases where each product was purchased. The formula is:

$$\text{Jaccard Coefficient} = \frac{\text{Support(joint)}}{\text{Support(Antecedent)} + \text{Support(Consequent)} - \text{Support(Joint)}}$$

To demonstrate using the example given above, Table 3.3 informs us that beer and potato chips were purchased together twice. The total number of beer purchases was 3, and the total number of potato chip purchases was also 3. Support(joint) is another expression of co-occurrence, which we see is 2 from Table 3.3. Thus the Jaccard coefficient for beer and potato chips would be  $\{2/(3 + 3 - 2)\} = 0.500$ . The Jaccard coefficients for the other products relative to beer are 0.143 for milk, 0 for diapers, and 0 for soda. Table 3.4 gives the Jaccard coefficients corresponding to the correlation coefficients given in Table 3.4.

Comparing the two results, a similar ranking is observed. The strongest correlations are negative. Those purchasing Milk don't buy much Beer or Potato Chips, and those buying Soda don't buy much Beer. The strongest positive relationship is Milk and Soda. The strongest Jaccard relationships are between Beer and Potato Chips and that of Milk and Soda. Of course this is for demonstration, based on an extremely small sample size, and was made up anyway. The Correlation coefficients infer relationships of cross-references, while the Jaccard coefficients focus on the pairs of items directly. Both measures are easily obtained (correlation coefficients from widely available software; Jaccard coefficients from simple formulas). The primary issue is the relative accuracy in identifying relationships. The simpler Jaccard coefficient could well be more accurate, but requires a significant amount of representative data before results are reliable. (For instance, in our simple example, there were many zeros, which are not reliable on a large scale. However, in real retail establishments with a great deal of turnover, such data would be widely available.)

**Table 3.4** Jaccard coefficients for co-occurrence table

	Beer	Potato chips	Milk	Diapers	Soda
Beer					
Potato chips	0.500				
Milk	0.167	0.167			
Diapers	0	0	0.250		
Soda	0	0.333	0.500	0	

Setting up market-basket analysis requires significant investment and effort. Retail organizations can take over 18 months to implement this type of analysis. However, it provides a tool to compete in an increasingly competitive environment.

## Demonstration

We can demonstrate concepts by looking at Amazon.com's list of product categories (taken from their website). The first step was to associate products by category. The thirty seven **purchase profiles** are given in Table 3.5.

We generated a set of pseudo-Amazon data selecting twelve of these categories, while expanding books into e-books, hard back books, and paperback books. The next step is to determine what percentage of each of these categories was in each market basket. Here we make up numbers to demonstrate. Market baskets are assigned to each profile based on greatest dollar value. These profiles were viewed as capturing the reason the shopper was at the site. Market-basket analysis reveals that customers do not shop based on product groupings, but rather on personal needs. The consumer orientation allows understanding combinations of product purchases. For instance, hardback books, paperbacks, and e-books are different products, but all are part of the books purchase profile. Particular customers could buy a market basket in one profile (say books) during one visit, and another profile (say Movies and TV) later. The focus is on the market basket, not customers as individuals.

**Table 3.5** Purchase profiles for market-basket analysis

Apps and games	Cell phones and accessories	Gift cards	Pet supplies
Arts, crafts and sewing	Clothing shoes and jewelry-women	Handmade	Software
Automotive	Clothing shoes and jewelry-Men	Industrial and Scientific	Sports and outdoors
Baby	Clothing shoes and jewelry-girls	Luggage and travel gear	Tools and home improvement
Beauty	Clothing shoes and jewelry-boys	Luxury beauty	Toys and games
Books	Clothing shoes and jewelry-babies	Magazine subscriptions	Video games
CDs and vinyl	Collectibles and fine art	Movies and TV	Wine
Computers	Grocery and gourmet food	Musical instruments	
Digital music	Health and personal care	Office products	
Electronics	Home and business services	Patio, lawn and garden	

## ***Fit***

Correlation works best for this type of data (due to the large number of combinations). Data has to be numerical for correlation, but we can obtain Table 3.6 from Excel showing an extract of this data:

This data then is used to generate the correlations in Table 3.7.

The combinations with correlations over 0.3 in absolute value are:

Hard cover and paperback books	+0.724
Ebooks and paperback books	+0.705
Ebooks and hard cover books	+0.683
Baby and toys	+0.617
Ebooks and movies	-0.447
Paperback books and movies	-0.358
Ebooks and software	-0.348
Hard cover books and movies	-0.320

One might question the data with respect to movies and books, but maybe people that go to movies don't have time to read books. Otherwise these numbers seem plausible.

Jaccard coefficients can be calculated with Excel as well, although this involves a lot of manipulation. Results are shown in Table 3.8.

The results have similarities for combinations with positive correlation, but Jaccard results don't pick up combinations with negative correlation. Those pairs with Jaccard coefficients greater than 0.1 are shown in Table 3.9.

We see that the Jaccard results don't reflect negative relationships, while correlation does. There also is a different order of relationship strength obtained, although in general they are somewhat consistent.

## ***Profit***

Retailers typically determine the profitability of each purchase profile. The returns for some of the Amazon purchase profiles (generated subjectively here) n dollars

**Table 3.6** Pseudo-Amazon data for correlation

Auto	Baby	EBooks	Hard	Paper	Music	Elect	Health	GiftC	Luggage	Mag	Movies
0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	0	0	0

**Table 3.7** Pseudo-Amazon correlations

	Auto	Baby	EBooks	Hard	Paper	Music	Elect	Health	GiftC	Luggage	Mag	Movies	Software	Toys
Auto	1													
Baby	-0.029	1												
EBooks	-0.161	-0.212	1											
Hard	-0.105	-0.204	0.683	1										
Paper	-0.123	-0.170	0.705	0.724	1									
Music	-0.045	-0.072	-0.194	-0.243	-0.295	1								
Elect	-0.025	-0.002	-0.234	-0.099	-0.152	0.226	1							
Health	0.134	-0.007	-0.058	-0.057	-0.067	0.042	0.029	1						
GiftC	-0.030	-0.058	-0.061	-0.113	-0.124	0.057	0.016	-0.030	1					
Luggage	-0.008	-0.015	-0.050	-0.031	-0.031	-0.023	-0.013	-0.017	-0.016	1				
Mag	-0.021	0.017	-0.076	-0.090	-0.079	0.016	-0.003	0.031	0.065	-0.011	1			
Movies	-0.048	-0.078	-0.447	-0.320	-0.358	0.041	0.131	0.077	0.172	0.116	0.117	1		
Software	-0.044	-0.040	-0.348	-0.227	-0.243	-0.011	0.237	-0.042	-0.059	-0.022	-0.020	-0.034	1	
Toys	-0.044	0.617	-0.285	-0.249	-0.220	-0.022	-0.024	0.009	-0.033	-0.023	0.000	-0.073	-0.075	1
Wine	-0.030	0.020	-0.249	-0.173	-0.184	-0.023	-0.028	-0.065	0.033	-0.016	0.012	0.044	0.050	-0.019

**Table 3.8** Jaccard results for pseudo-Amazon data

**Table 3.9** Jaccard and correlation comparison

Variable pairs	Jaccard coefficients	Correlation coefficients
Ebooks and paperback books	0.763	+0.705
Hard cover books and paperback books	0.755	+0.724
Ebooks and hard cover books	0.746	+0.683
Baby and toys	0.426	+0.617
Electronics and software	0.144	+0.237
Music and electronics	0.137	+0.226
Gift cards and movies	0.125	+0.172

**Table 3.10** Purchase profile returns

	Profit per customer in \$/Year	Prob of response	Product	Expected profit/year	Number	Profile Profit/Year
Books	56	0.50	28.00	8.00	80,000	640,000
Baby and toy	90	0.30	27.00	7.00	10,000	70,000
Auto	60	0.10	6.00	-14.00	500	-7000
Music	24	0.27	6.48	-13.52	16,000	-216,320
Elect and SW	64	0.20	12.80	-7.20	8000	-57,600
Health	172	0.22	37.84	17.84	12,000	214,080
Gift	16	0.31	4.96	-15.04	15,000	-225,600
Luggage	26	0.08	2.08	-17.92	400	-7168
Magazine	14	0.15	2.10	-17.90	2000	-35,800
Movies	72	0.33	23.76	3.76	17,000	63,920
Wine	56	0.21	11.76	-8.24	800	-6592

(profit) per market basket (mythical numbers) are shown in Table 3.10 (numbers generated subjectively by the author).

This information can assist in promotion decision-making. Often retailers find that their past advertising has been for products in purchase profiles with low profitability returns. The same approach can be applied to other promotional efforts, although promotions often cross profiles. The effect of promotions can be estimated by measuring the profitability by market basket.

## ***Lift***

We can divide the data into groups as fine as we want. These groups have some identifiable feature, such as zip code, income level, etc. (a profile). We can then sample and identify the portion of sales for each group, some way to obtain

expected profit per customer in a given profile. The idea behind lift is to send promotional material (which has a unit cost) to those groups that have the greatest probability of positive response first. We can visualize lift by plotting responses against the proportion of the total population of potential customers as shown in Table 3.11. Note that segments are listed in Table 3.10 sorted by expected customer response.

Both the cumulative responses and the cumulative proportion of the population are plotted to identify lift. Lift is the difference between the two lines in Fig. 3.4.

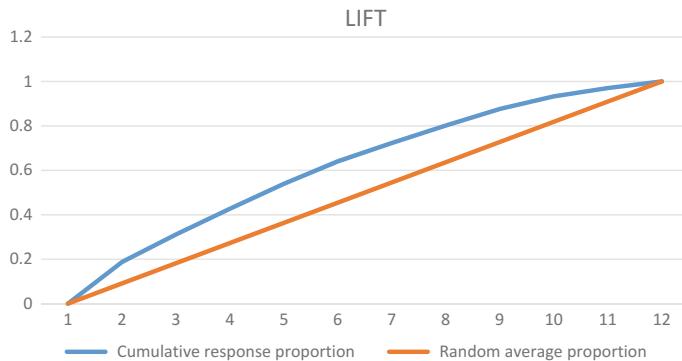
The purpose of lift analysis is to identify the most responsive segments. Here the greatest lift is obtained from the first segment, with the next five contributing above average response. We are probably more interested in profit, however. We can identify the most profitable policy. What needs to be done is to identify the portion of the population to send promotional materials to. For instance, if a promotion (such as a coupon) costing \$20 is proposed, we need information available to calculate an average profit per profile. Table 3.12 shows these calculations.

The profit function reaches its maximum with the fourth segment. The implication is that in this case, promotional materials should be sent to these segments. If there was a promotional budget, it would be applied to as many segments as the budget would support, in order of expected response rate, up to the fifth segment.

It is possible to focus on the wrong measure. The basic objective of lift analysis in marketing is to identify those customers whose decisions will be influenced by marketing in a positive way. In short, the methodology described above identifies those segments of the customer base who would be expected to purchase. This may or may not have been due to the marketing campaign effort. The same methodology

**Table 3.11** Lift calculation

	Expected customer response	Proportion (expected responses)	Cumulative response proportion	Random average proportion	LIFT
origin	0	0	0	0	0
Books	0.500	0.187	0.187	0.091	0.096
Movies	0.330	0.124	0.311	0.182	0.129
Gift	0.310	0.116	0.427	0.273	0.154
Baby and toy	0.300	0.112	0.539	0.364	0.176
Music	0.270	0.101	0.640	0.455	0.186
Health	0.220	0.082	0.723	0.545	0.177
Wine	0.210	0.079	0.801	0.636	0.165
Elect and SW	0.200	0.075	0.876	0.727	0.149
Magazine	0.150	0.056	0.933	0.818	0.114
Auto	0.100	0.037	0.970	0.909	0.061
Luggage	0.080	0.030	1	1	0



**Fig. 3.4** Lift for pseudo-amazon

**Table 3.12** Calculation of expected payoff

	Profit per customer in \$/year	Prob of response	Product	Expected profit/year	Potential customers	Responses year	Profile profit year
Books	56	0.5	28	8	80,000	40,000	640,000
Health	172	0.22	37.84	17.84	12,000	2640	214,080
Baby and toy	90	0.3	27	7	10,000	3000	70,000
Movies	72	0.33	23.76	3.76	17,000	5610	63,920
Wine	56	0.21	11.76	-8.24	800	168	-6592
Auto	60	0.1	6	-14	500	50	-7000
Luggage	26	0.08	2.08	-17.92	400	32	-7168
Magazine	14	0.15	2.1	-17.9	2000	300	-35,800
Elect and SW	64	0.2	12.8	-7.2	8000	1600	-57,600
Music	24	0.27	6.48	-13.52	16,000	4320	-216,320
Gift	16	0.31	4.96	-15.04	15,000	4650	-225,600

can be applied, but more detailed data is needed to identify those whose decisions would have been changed by the marketing campaign rather than simply those who would purchase.

## Market Basket Limitations

Identifying relationships among product sales is no good unless it is used. Measurement of effects is critical to sound data mining. One of the most commonly cited examples of market-basket analysis was the hypothesized tendency for men to

purchase beer and diapers together. If this were true, the knowledge would only be valuable if action could be taken that would increase sales. For instance, one theory is that retailers should locate beer and diapers close to each other, encouraging sales of both. An alternative theory would be to locate the two products as far apart as possible, forcing customers to see the maximum number of products available in the store. In either case, measurement of impact is required to gain useful knowledge. Market-basket analysis, an exploratory algorithm, can only generate hypotheses. Once hypotheses are generated, they need to be tested.

Market-basket analysis is often an initial study intended to identify patterns. Once a pattern is detected, it can be explored more thoroughly through other methods, such as neural networks, regression, or decision trees. The same analytic approaches can be used in applications outside of retail. Telecommunications companies and banks often bundle products together, and market-basket analysis has been applied in those fields. The insurance industry has been a major user of link analysis to detect fraud rings. In the medical field, combinations of symptoms can be analyzed to gain deeper understanding of a patient's condition.

Market basket analysis has limits of course. However, knowing which products are associated doesn't answer all retailer questions. Cross-sales strategies assume that products that appear together are complements. Some research has found that market basket analysis may identify as many substitutes as it does complements (Vindvogel et al. 2005).

The overall strengths of market-basket analysis include clear results through simple computations. Market-basket analysis can be undirected, in that hypothetical relationships do not need to be specified prior to analysis. Different data forms can be used as well. Weaknesses of the method are (1) that the complexity of the analysis grows exponentially with the volume of products considered, and (2) that it is difficult to identify an appropriate number of product groupings. The thirty groupings demonstrated above is a good workable, actionable number. Too few product groupings provide no benefit, while too many makes it impossible to make sense of the analysis. Market-basket analysis is a technique that is good for undirected, or unstructured, problems with well-defined items. It is very suitable for cash register data.

## References

- Agrawal R, Imielinski T, Swami A (1993) Database mining: a performance perspective. *IEEE Trans Knowl Data Eng* 5(6):914–925
- Aguinis H, Forcum LE, Joo H (2013) Using market basket analysis in management research. *J Manag* 39(7):1799–1824
- Berry MJA, Linoff G (1997) Data mining techniques. Wiley, New York
- Vindvogel B, Van den Poel D, Wets G (2005) Why promotion strategies based on market basket analysis do not work. *Expert Syst Appl* 28:583–590

# Chapter 4

## Recency Frequency and Monetary Analysis



Recency, Frequency, and Monetary (RFM) analysis seeks to identify customers who are more likely to respond to new offers. While lift looks at the static measure of response to a particular campaign, RFM keeps track of customer transactions by time, by frequency, and by amount.

1. **Recency:** time since the customer made his/her most recent purchase
2. **Frequency:** number of purchases this customer made within a designated time period
3. **Monetary:** average purchase amount.

Time is important, as some customers may not have responded to the last campaign, but might now be ready to purchase the product being marketed. Customers can also be sorted by frequency of responses, and by dollar amount of sales. The Recency, Frequency, and Monetary (RFM) approach is a method to identify customers who are more likely to respond to new offers. Subjects are coded on each of the three dimensions. A common approach is to have five cells for each of the three measures, yielding a total of 125 combinations, each of which can be associated with a probability of a positive response to the marketing campaign.

RFM has been found to work relatively well if expected response rate is high. The original RFM model can be extended either by considering additional variables (e.g., socio-demographics) or by combining with other response techniques. Other variables that may be important include customer income, customer lifestyle, customer age, product variation, and so on. That would make traditional data mining tools such as logistic regression more attractive. The three variables tend to be correlated, especially F and M. Because of the high correlation between F and M, Yang (2004) offered a version of RFM model collapsing the data to a single variable “Value” = M/R. To overcome the problem of data skewed in RFM cells, Olson et al. (2009) proposed an approach to balance observations in each of the 125 RFM cells.

## Dataset 1

We demonstrate RFM with two retail datasets. This research design includes data obtained from the Direct Marketing Educational Foundation. The first dataset we present included 101,532 individual purchases from 1982 to 1992 in catalog sales. A test set of 20,000 observations was held out for testing, leaving a training set of 81,532. The last four months (Aug–Dec) of the data was used as the target period: Aug–Dec 1992 for Dataset 1. The average response rate was 0.096. The raw data contained customer behavior represented by account, order (or donation) date, order (donation) dollars, and many other variables. We followed the general coding scheme to compute R, F, and M. Various data preparation techniques (e.g., filtering, transforming) were used during this process. The order date of last purchase (or the date of last donation) was used to compute R (R1, R2, R3, R4, R5). The data set contained order (or donation) history and order dollars (or donation amounts) per each customer (or donor), which were used for F (F1, F2, F3, F4, F5) and M (M1, M2, M3, M4, M5). We also included one response variable (Yes or No) to the direct marketing promotion or campaign. An initial correlation analysis was conducted, showing that there was some correlation among these variables, as shown in Table 4.1.

All three variables were significant at the 0.01 level. The relationship between R and response is negative, as expected. In contrast, F and M are positively associated with customer response. R and F are stronger predictors for customer response.

Dividing the data into 125 cells, designated by 5 categories for each of the three groups, the most attractive group would be **555**, or Group 5 for each of the 3 variables. Here RFM was accomplished in Excel. Table 4.2 shows boundaries. Group 5 was assigned the most attractive group, which for R was the minimum, and for F and M the maximum.

Note the skewness of the data for F, which is often encountered. Here the smaller values dominate that metric. Table 4.3 displays the counts obtained for these 125 cells.

The correlation across F and M (0.631 in Table 4.1) can be seen in Table 4.3, looking at the R = 5 categories. In the M = 1 column of Table 4.3, F entries are 0 for every F5 category, usually increasing through M = 2 through M = 5 columns. When F = 5, the heaviest density tends to be in the column where M = 5. This skewness is often recognized as one of the problems with RFM. Our approach to this issue was through more equal density (size-coding) to obtain data entries for all RFM cells. We accomplished this by setting cell limits by count within the training set for each variable.

**Table 4.1** Variable correlations

	R	F	M	Ordered
R	1			
F	-0.192**	1		
M	-0.136**	0.631**	1	
Ordered	-0.235**	0.241**	0.150**	1

\*\*Correlation is significant at the 0.01 level (2-tailed)

**Table 4.2** RFM boundaries

Factor	Min	Max	Group 1	Group 2	Group 3	Group 4	Group 5
R	12	3810	1944+	1291–1943	688–1290	306–687	12–305
Count			<b>16,297</b>	<b>16,323</b>	<b>16,290</b>	<b>16,351</b>	<b>16,271</b>
F	1	39	1	2	3	4–5	6+
Count			<b>43,715</b>	<b>18,274</b>	<b>8206</b>	<b>6693</b>	<b>4644</b>
M	0	4640	0–20	21–38	39–65	66–122	123+
Count			<b>16,623</b>	<b>16,984</b>	<b>15,361</b>	<b>16,497</b>	<b>16,067</b>

**Table 4.3** Count by RFM cell—training set

RF	R	F	M1	M2	M3	M4	M5
55	R 12-305	F 6+	0	0	16	151	1761
54		F 4-5	2	18	118	577	1157
53		F 3	9	94	363	756	671
52		F 2	142	616	1012	1135	559
51		F 1	2425	1978	1386	938	387
45	R 306-687	F 6+	0	1	11	101	1018
44		F 4-5	0	16	87	510	927
43		F 3	6	88	316	699	636
42		F 2	150	707	1046	1140	616
41		F 1	2755	2339	1699	1067	416
35	R 688-1290	F 6+	0	1	5	70	799
34		F 4-5	1	16	122	420	832
33		F 3	9	88	319	706	589
32		F 2	163	697	1002	1128	645
31		F 1	2951	2567	1645	1078	437
25	R 1291-1943	F 6+	0	0	9	56	459
24		F 4-5	0	22	72	372	688
23		F 3	9	95	290	678	501
22		F 2	211	749	1096	1128	561
21		F 1	3377	2704	1660	1108	478
15	R 1944+	F 6+	0	0	3	22	170
14		F 4-5	1	11	74	243	409
13		F 3	9	122	261	511	380
12		F 2	268	878	1108	995	522
11		F 1	4145	3177	1641	908	449
	Totals		16,623	16,984	15,361	16,497	16,067

The proportion of responses (future order placed) for the data is given in Table 4.4. Bold numbers indicate cells with positive response (confidence) of at least 0.1 and minimum support of 50. Italicized numbers indicate those cells with support below 50.

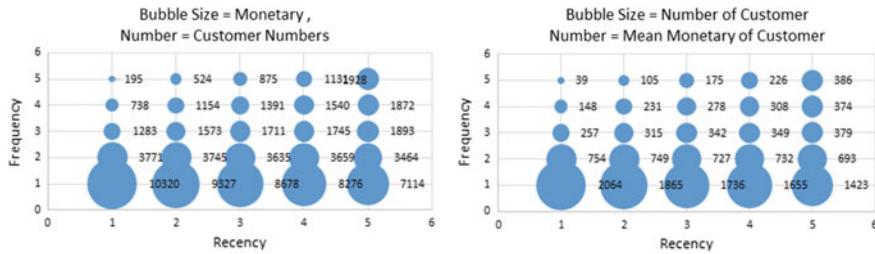
The data shown in Table 4.3 has been visualized in Fig. 4.1.

The bubble size in the left figure represents the total monetary value of the market segment, while on the right hand side the bubble size represents the number of customers. One could represent the numbers by a color: For example many customers could be represented by red (hot) and few customers by blue (cold). This enables quick visualization identifying that the largest market segment has lower frequency both in terms of number of customers and the mean monetary value of each customer.

We can further display the data. For example a shop like Costco might have this type of customers, which don't shop very frequently but spend large amounts when they do, while high frequency customers might only buy a piece of pizza, but

**Table 4.4** Response ratios by cell

RF	R	F	M1	M2	M3	M4	M5
55	R 12-305	F 6+	–	–	0.687	<b>0.563</b>	<b>0.558</b>
54		F 4-5	0	0.500	<b>0.415</b>	<b>0.426</b>	<b>0.384</b>
53		F 3	0.111	<b>0.426</b>	<b>0.342</b>	<b>0.381</b>	<b>0.368</b>
52		F 2	<b>0.296</b>	<b>0.289</b>	<b>0.281</b>	<b>0.283</b>	<b>0.256</b>
51		F 1	<b>0.173</b>	<b>0.196</b>	<b>0.201</b>	<b>0.158</b>	<b>0.152</b>
45	R 306-687	F 6+	–	0	<b>0.273</b>	<b>0.238</b>	<b>0.193</b>
44		F 4-5	–	0.125	0.092	<b>0.112</b>	<b>0.123</b>
43		F 3	0	0.091	0.082	0.089	0.101
42		F 2	0.060	0.075	0.069	0.081	0.078
41		F 1	0.047	0.049	0.052	0.053	0.041
35	R 688-1290	F 6+	–	1.000	0	<b>0.100</b>	<b>0.125</b>
34		F 4-5	0	0.063	<b>0.107</b>	<b>0.107</b>	<b>0.103</b>
33		F 3	0.111	0.023	0.066	0.059	0.075
32		F 2	0.049	0.047	0.061	0.063	0.060
31		F 1	0.030	0.031	0.029	0.026	0.021
25	R 1291-1943	F 6+	–	–	<b>0.111</b>	0.054	0.078
24		F 4-5	–	<b>0.091</b>	0.028	0.065	0.060
23		F 3	0	0.053	0.048	0.049	0.064
22		F 2	0.043	0.020	0.039	0.041	0.039
21		F 1	0.018	0.021	0.018	0.020	0.019
15	R 1944+	F 6+	–	–	<b>0.000</b>	0.045	0.041
14		F 4-5	0	<b>0.091</b>	0.024	0.025	0.039
13		F 3	0.111	0.041	0.050	0.033	0.053
12		F 2	0.019	0.046	0.036	0.031	0.044
11		F 1	0.021	0.015	0.016	0.020	0.016

**Fig. 4.1** Bubble graph from excel

spending nothing else. Discovering such market segments, Costco could actually control its customer base through the membership fee. Only customers who are willing to spend an annual membership of \$50 are permitted to enter the shop. This will then result in fewer low volume customers and increasing the average monetary value spent by each customer. However the membership might reduce desirable customers. Having divided our customers into 125 groups it is easy to get confused and draw the wrong conclusions.

The data shown in Table 4.3 may be subdivided into fewer than the 125 groups. In the Fig. 4.2 the division of the 125 groups into groups based on Regency and Frequency is shown together with the % of customers. For each customer segment a different strategy might be developed to optimize return. RFM analysis is therefore less of a tool to fit data but a way to categorize and build a strategy for each segment. In this case Regency, Frequency and Monetary is chosen as the three key input parameters. However the same principle can be applied to other problems with different key input parameters.

In the training set, 10 of 125 possible cells were empty, even with over 80,000 data points. The cutoff for profitability would depend upon cost of promotion compared to average revenue and rate of profit. For example, if cost of promotion were \$50, average revenue per order \$2000, and average profit rate \$0.25 per dollar of revenue, the profitability cutoff would be 0.1. In Table 4.4, those cells with return ratios greater than 0.1 are shown in bold. Those cells with ratios at 0.1 or higher with support (number of observations) below 50 are indicated in italics. They are of interest because their high ratio may be spurious. The implication is fairly self-evident—seek to apply promotion to those cases in bold without italics.

**Fig. 4.2** Graphic of RFM segments

**Table 4.5** Basic RFM models by cutoff

Cutoff	R	F	M
0.1	R = 5	Any	Any
	R = 4	F = 5	M = 3, 4, or 5
		F = 4	M = 4 or 5
		F = 3	M = 5
	R = 3	F = 4 or 5	M = 3, 4, or 5
0.2	R = 5	F = 2, 3, 4, or 5	Any
	R = 4	F = 5	M = 3, 4, or 5
0.3	R = 5	F = 3, 4, or 5	M = 2, 3, 4, or 5
0.4	R = 5	F = 4 or 5	M = 2, 3, 4 or 5
0.5	R = 5	F = 5	M = 3, 4, or 5

The idea of dominance can also be applied. The combinations of predicted success for different training cell proportions are given in Table 4.5.

The RFM model from the Excel spreadsheet model was correct ( $13,961 + 1337 = 15,298$ ) times out of 20,000, for a correct classification rate of 0.765. The error was highly skewed, dominated by the model predicting 4113 observations to be 0 that turned out to respond. An alternative model would be degenerate—simply predict all observations to be 0. This would have yielded better performance, with 18,074 correct responses out of 20,000, for a correct classification rate of 0.904. This value could be considered a par predictive performance.

Increasing the test cutoff rate leads to improved models. We used increasing cutoffs of 0.2, 0.3, 0.4, and 0.5, yielding correct classification rates of 0.866 for minimum confidence of 0.2, 0.897 for minimum confidence of 0.3, 0.903 for minimum confidence of 0.5, and 0.907 for minimum confidence of 0.5. Only the model with a cutoff rate of 0.5 resulted in a better classification rate than the degenerate model. In practice, the best cutoff rate would be determined by financial impact analysis, reflecting the costs of both types of errors. Here we simply use classification accuracy overall, as we have no dollar values to use.

## Balancing Cells

One of the problems with RFM is skewness in cell densities. Our data set is small, and obviously it would be better to have millions of observations, which would increase cell counts. However, sometimes data is not available, and our purpose is to demonstrate RFM. Thus a second approach might be to try to obtain cells with more equal density (size-coding). We can accomplish this by setting cell limits by count of the building set. We cannot obtain the desired counts for each of the 125 combined cells because we are dealing with three scales. But we can come closer, as in Table 4.6. Difficulties arose primarily due to F having integer values. Table 4.6 limits were generated sequentially, starting by dividing R into 5 roughly

equal groups. Within each group, F was then sorted into groups based on integer values, and then within those 25 groups, M divided into roughly equally sized groups.

The unevenness of cell densities is due to uneven numbers in the few integers available for the F category. The proportion of positive responses in the training set is given in Table 4.7, with correct classification rates of 0.1 or higher in bold. All cells in Table 4.7 had minimum support of at least 149 as shown in Table 4.6.

If M = 5, this model predicts above average response. There is a dominance relationship imposed, so that cells 542 and better, 532 and better, 522 and better, 512 and better, 452 and better, 442 and better, and 433 and better are predicting above average response. Cells 422, 414, and 353 have above average training response, but cells with superior R or F ratings have below average response, so these three cells were dropped from the above average response model. The prediction accuracy ((13,897 + 734)/20,000) for this model was 0.732 (see the Balance on 0.1 row in the Appendix). In this case, balancing cells did not provide added accuracy over the basic RFM model with unbalanced cells. Using the cutoff rate of

**Table 4.6** Balanced group cell densities—training set

RF	M1	M2	M3	M4	M5
55	186	185	149	223	187
54	185	186	185	185	186
53	187	185	188	186	187
52	184	184	185	184	185
51	186	187	186	187	186
45	268	265	270	289	246
44	269	269	268	274	264
43	272	267	280	251	296
42	263	263	265	245	283
41	268	261	261	259	277
35	331	330	349	316	330
34	324	325	322	325	324
33	332	331	329	332	335
32	330	330	330	331	330
31	323	324	323	326	324
25	733	730	735	737	733
24	735	736	735	737	734
23	747	746	751	749	748
22	705	704	707	704	707
21	731	733	730	735	732
15	1742	1746	1739	1740	1744
14	1718	1715	1713	1713	1716
13	1561	1809	1689	1675	1684
12	1768	1775	1771	1779	1762
11	1830	1831	1832	1824	1839

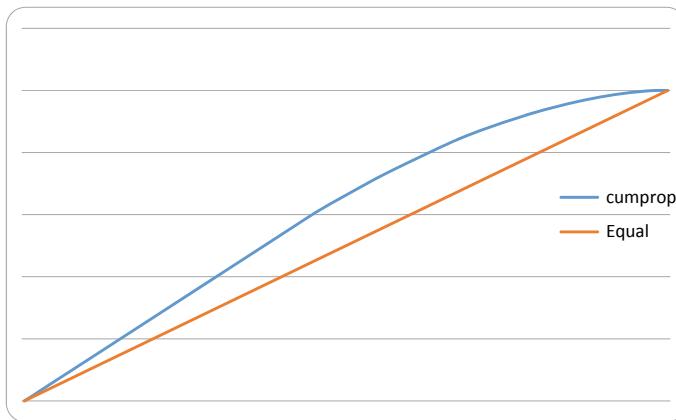
**Table 4.7** Training set proportion of responses by cell

RF	M1	M2	M3	M4	M5
55	<b>0.129</b>	<b>0.178</b>	<b>0.101</b>	<b>0.673</b>	<b>0.818</b>
54	0.059	<b>0.118</b>	<b>0.189</b>	<b>0.541</b>	<b>0.629</b>
53	0.064	<b>0.130</b>	<b>0.287</b>	<b>0.392</b>	<b>0.647</b>
52	0.076	<b>0.103</b>	<b>0.200</b>	<b>0.424</b>	<b>0.605</b>
51	0.054	<b>0.102</b>	<b>0.274</b>	<b>0.406</b>	<b>0.527</b>
45	0.037	<b>0.109</b>	<b>0.141</b>	<b>0.211</b>	<b>0.378</b>
44	0.041	<b>0.108</b>	<b>0.116</b>	<b>0.281</b>	<b>0.417</b>
43	0.033	0.052	<b>0.125</b>	<b>0.072</b>	<b>0.483</b>
42	0.049	0.118	0.098	0.073	<b>0.544</b>
41	0.045	0.038	0.092	0.116	<b>0.531</b>
35	0.045	0.067	0.138	0.060	<b>0.458</b>
34	0.052	0.043	0.059	0.080	<b>0.448</b>
33	0.042	0.048	0.058	0.093	<b>0.433</b>
32	0.027	0.045	0.058	0.097	<b>0.379</b>
31	0.050	0.040	0.062	0.080	<b>0.414</b>
25	0.037	0.051	0.056	0.084	<b>0.254</b>
24	0.024	0.046	0.052	0.076	<b>0.309</b>
23	0.051	0.047	0.055	0.080	<b>0.273</b>
22	0.027	0.040	0.055	0.068	<b>0.246</b>
21	0.027	0.038	0.048	0.076	<b>0.242</b>
15	0.017	0.021	0.025	0.051	<b>0.146</b>
14	0.016	0.017	0.033	0.054	<b>0.167</b>
13	0.010	0.019	0.034	0.052	<b>0.156</b>
12	0.018	0.021	0.036	0.043	<b>0.137</b>
11	0.016	0.022	0.014	0.044	<b>0.154</b>

0.5, the model is equivalent to predict the combination of R = 5, F = 4 or 5, and M = 4 or 5 as responding and all others not. This model had a correct classification rate of 0.894, which was inferior to the degenerate case. For this set of data, balancing cells accomplished better statistical properties per cell, but was not a better predictor.

## Lift

Lift is the marginal difference in a segment's proportion of response to a promotion and the average rate of response. The methodology sorts segments by probability of response, and compares cumulative response curve with average response. The basic objective of lift analysis in marketing is to identify those customers whose decisions will be influenced by marketing in a positive way. Target customers are identified as the small subset of people with marginally higher probability of purchasing.



**Fig. 4.3** Lift for equalized data groups

We are probably more interested in profit, however. Lift itself does not consider profitability. We can identify the most profitable policy, but what really needs to be done is identify the portion of the population to send promotional materials to. For our purposes, we demonstrate without dollar values (which are not available), noting that the relative cost of marketing and expected profitability per segment will determine the optimal number of segments to market. The lift chart for this data is given in Fig. 4.3.

Lift maxes out at Group 554, the 73rd of 125 cells. This cell had a response rate of 0.75, slightly above the training set data average of 0.739. Of course, the point is not to maximize lift, but to maximize profitability, which requires knowing expected profit rate for revenue, and cost of marketing. The test results for coding the data with effort to balance cell size yielded overall correct classification was relatively better at 0.792.

## Value Function

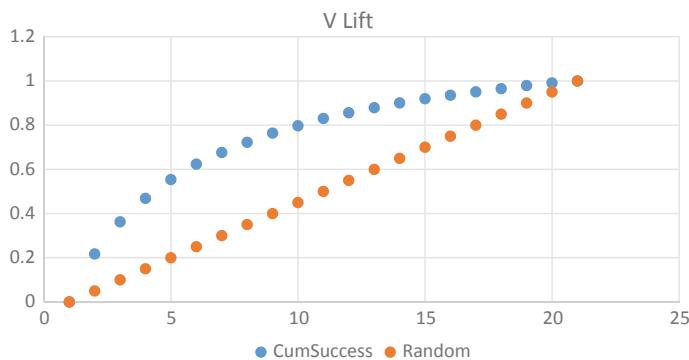
The value function compresses the RFM data into one variable— $V = M/R$ . Since F is highly correlated with M (0.631 in Table 4.1), the analysis is simplified to one dimension. Dividing the training set into groups of 5%, sorted on V, generates Table 4.8.

Figure 4.4 shows lift as the difference between cumulative success and random for data sorted by value ratio.

In Fig. 4.4, the most responsive segment has an expected return of slightly over 40%. The lift line is the cumulative average response as segments are added (in order of response rate).

**Table 4.8** V values by cell

Cell	Min V	UL	Hits	N	Success
1	0.0000	4077	91	4076	0.0223
2	0.0063	8154	69	4077	0.0169
3	0.0097	12,231	116	4077	0.0285
4	0.0133	16,308	109	4077	0.0267
5	0.0171	20,385	120	4077	0.0294
6	0.0214	24,462	119	4077	0.0292
7	0.0263	28,539	151	4077	0.0370
8	0.0320	32,616	174	4077	0.0427
9	0.0388	36,693	168	4077	0.0412
10	0.0472	40,770	205	4077	0.0503
11	0.0568	44,847	258	4077	0.0633
12	0.0684	48,924	256	4077	0.0628
13	0.0829	53,001	325	4077	0.0797
14	0.1022	57,078	360	4077	0.0883
15	0.1269	61,155	408	4077	0.1001
16	0.1621	65,232	542	4077	0.1329
17	0.2145	69,309	663	4077	0.1626
18	0.2955	73,386	827	4077	0.2028
19	0.4434	77,463	1134	4077	0.2781
20	0.7885	81,540	1686	4070	0.4143
Total/Avg			7781	81,532	0.0954

**Fig. 4.4** Lift by value ratio cell

### Software Demonstration of RFM Analysis by Excel

There are different software tools that can be used to carry out a RFM Analysis. The computations can be done using Excel. For the RFM analysis, each parameter is divided into five categories. Therefore we must find a way to determine these borders. There are several approaches as previously discussed in this chapter.

**Table 4.9** Summary statistics from excel

	Recency	Frequency	Monetary
Sum	301,385	6629	746,096
Average	301	7	746
Max	1540	56	8579
Min	1	1	5
Range	1539	55	8575
5	385	14	2144

A higher frequency results in shorter time period between each period. In other words a high frequency results in shorter Regency time as also observed in the histogram. In this histogram the entire data range is divided up equally. To identify the borders, determine the minimum and maximum and other parameters of the data which is done in Excel with the function = MIN(range) and = MAX(range) with output shown in Table 4.9.

Having obtained the maximum and minimum, we obtain the desired step size by dividing the data range by five as in this case as we would like to create five categories. Dividing these would give the segment ranges: Recency (1): between 0 and 308; (2) between 309 and 616; (3) between 617 and 924; (4) between 925 and 1232; and (5) between 1233 and 1540.

We can then obtain the assignment by using a simple equation:

Regency Assignment = 1 + INT((Recency-Recency Minimum)/385) using appropriate Excel cell references. Next we count the number of cases in each category, which can be done with the COUNTIF function. For example the equation below will count how many of the cells in H25–H1024 meet the condition as specified in cell M17

$$= \text{COUNTIF}(H25 : H1024, M17)$$

We carry out this analysis and obtain the data as shown in Table 4.10.

We see that equally spaced ranges give large values for a few cells and very low (even empty) counts for others. After we assigned for example a customer to R = 1, F = 1 and M = 1, we would assign the RF score = 11 for this customer. If the values desired are in cells H25 and I25 the function would be:

$$= H25\&I25$$

Now we would like to count how many customer meet these conditions or obtain the sum or mean customer value. This can be done in Excel with the SUMIFS function and the COUNTIF function:

$$= \text{IF}(\$V25 < > 0, (\text{SUMIFS}(C\$25 : C\$1024, \$L\$25 : \$L\$1024, \$Q25)), \text{""})$$

Using these functions we obtain the values in Fig. 4.5.

We can obtain a visual as shown in Fig. 4.6.

**Table 4.10** Counts from excel

	Recency	Frequency	Monetary
1	756	918	930
2	98	71	60
3	52	10	7
4	93	0	2
5	1	1	1
Count	1000	1000	1000

Sum Monetary		Number of Customers		Monetary per Customer	
Monetary	Frequency	Customers	Frequency	Monetary	Frequency
Recency	1 2 3 4 5	Recency	1 2 3 4 5	Recency	1 2 3 4 5
1	468304 165331 38533 8070	1	674 71 10 1	1	695 2329 3853 8070
2	37312	2	98	2	381
3	16963	3	52	3	326
4	11527	4	93	4	124
5	55	5	1	5	55

Monetary	Frequency	Customers	Frequency	Monetary	Frequency
Recency	1 2 3 4 5	Recency	1 2 3 4 5	Recency	1 2 3 4 5
1	63% 22% 5% 1%	1	67% 7% 1%	1	93% 312% 516% 1082%
2	5%	2	10%	2	51%
3	2%	3	5%	3	44%
4	2%	4	9%	4	17%
5	0%	5	0%	5	7%

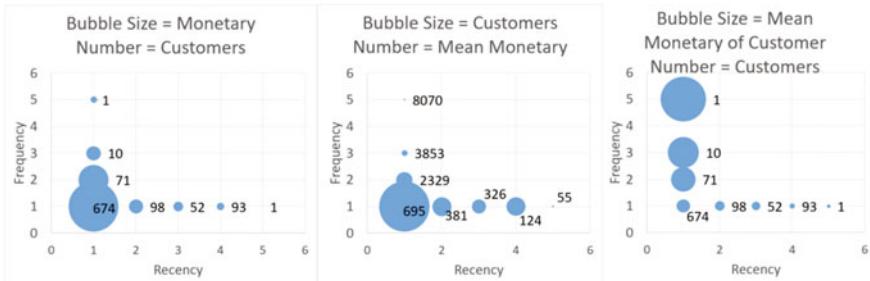
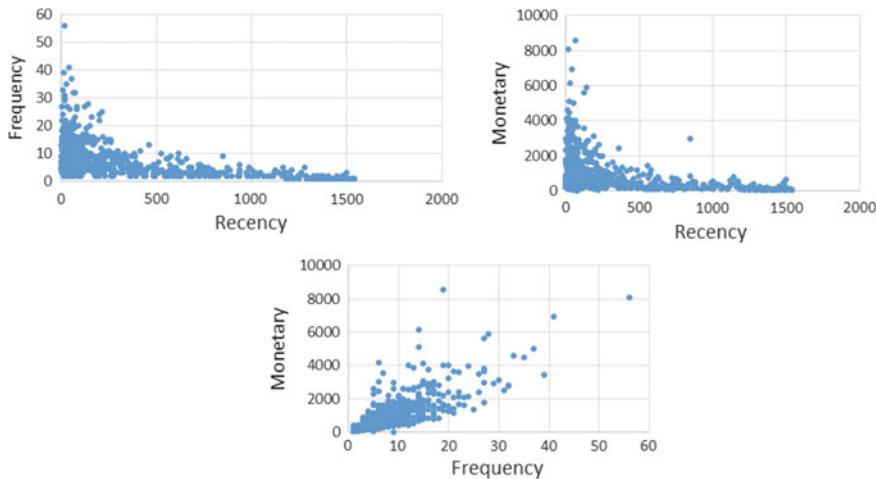
**Fig. 4.5** Counts in example**Fig. 4.6** Plots from excel

Figure 4.6 includes bubble sizes to better see show density of points. A similar approach to visualize the data was proposed by Kohavi and Parekh (2004). This can be elaborated through scatter plots as in Fig. 4.7.

### Calculating Percentiles

The above analysis was based on standard deviations. However in the RFM analysis, the borders are created by percentiles. This better describes the data and enables a strategy for different customers. The Italian economist Vilfredo Pareto was a avid gardener and he noticed that 20% of the pea pods in his garden



**Fig. 4.7** Scatter plots of RFM data

contained a whopping 80% of the overall peas. He then applied this interesting finding to his economics work and discovered that about 20% of the people in Italy owned about 80% of the land. The same type of observation we like to carry out in the RFM analysis. But like Pareto, who looked at is Pea plants and may have tried to control the water, compost and light supply to the peas to improve the yield, we need to have an open mind. RFM analysis helps us to see who our better customers are, enabling us to treat them well.

Finding these percentiles can be done in Excel. For example the equation below determines the value for the 20% percentile of the values in cells B15 to B1014:

$$= \text{PERCENTILE}(B15 : B1014, 20\%)$$

We can count to obtain more balanced cells. Most interesting is how the values are set. For example in the case of recency there is a small step size for the lower percentile and larger step size for the higher percentile (Table 4.11).

**Table 4.11** Counts based on percentiles

Percentile		Recency			Frequency			Monetary		
		Count	Min	Grade	Count	Min	Grade	Count	Min	Grade
0										
20%	1	190	1	5	200	1	1	200	5	1
40%	2	210	24	4	141	3	2	200	169	2
60%	3	199	57	3	215	4	3	200	334	3
80%	4	201	166	2	196	6	4	200	583	4
100%	5	200	555	1	248	9	5	200	1111	5

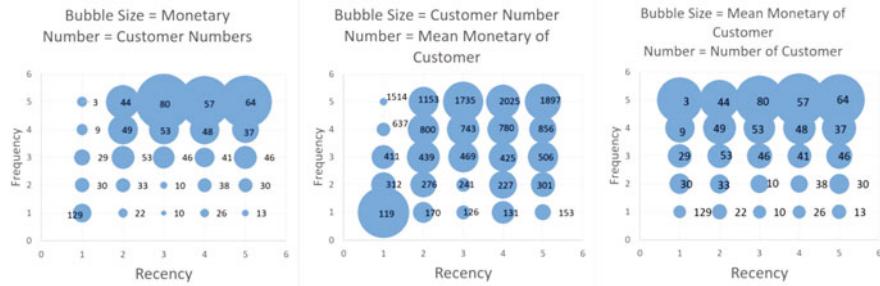


Fig. 4.8 Plots of more equally distributed cells

Now we can plot again the data as in Fig. 4.8.

## Data Mining Classification Models

There are three fundamental data mining classification algorithms: Logistic regression, decision trees, and neural network models.

### Logistic Regression

The purpose of logistic regression is to classify cases into the most likely category. Logistic regression provides a set of  $\beta$  parameters for the intercept (or intercepts in the case of ordinal data with more than two categories) and independent variables, which can be applied to a logistic function to estimate the probability of belonging to a specified output class. Logistic regression is among the most popular data mining techniques in marketing DSS and response modeling. A logistic regression model was run on RFM variables. The model results were as shown in Table 4.12.

M was not significant in this model. Application to test data yielded an overall correct classification rate of 0.907. Logistic regression has the ability to include other variables. The only external variable to RFM available was promotion. Including promotion improved the model fit by the smallest of margins.

Table 4.12 Regression betas for logistic regression

Variable	Beta	Significance
Constant	-1.5462	0.05
R	-0.0015	<0.05
F	0.2077	<0.05
M	-0.0002	

## ***Decision Tree***

Decision trees in the context of data mining refer to the tree structure of rules. They have been applied by many in the analysis of direct marketing data. The data mining decision tree process involves collecting those variables that the analyst thinks might bear on the decision at issue, and analyzing these variables for their ability to predict outcome. Decision trees are useful to gain further insight into customer behavior, as well as lead to ways to profitably act on results. One of a number of algorithms automatically determines which variables are most important, based on their ability to sort the data into the correct output category. The method has relative advantage over neural network in that a reusable set of rules are provided, thus explaining model conclusions.

For Dataset 1, we used J48, one of the most popular decision tree algorithms. The J48 decision tree algorithm was applied to the test set of 20,000. The resultant decision tree was as shown in Table 4.13.

Note that F was not included at all. This is explainable by the high correlation between M and F, and the dominance of R in obtaining better fit. This model did very well on the test data, with a correct classification rate of 0.984.

## ***Neural Networks***

Neural networks are the third classical data mining tool found in most commercial data mining software products, and have been applied to direct marketing applications. NN are known for their ability to train quickly on sparse data sets. PNN separates data into a specified number of output categories. NN are three layer networks wherein the training patterns are presented to the input layer and the output layer has one neuron for each possible category.

There are many other neural network models with a number of parameters that can be selected. Running a number of these, the best fit was obtained with a probabilistic neural network (PNN), yielding a correct classification rate of 0.911.

Table 4.14 gives comparative performance of the models applied to Dataset 1.

## ***Dataset 2***

The second dataset (also from the Direct Marketing Education Foundation) is based on the data of 1,099,009 individual donors' contributions to a non-profit organization collected between 1991 and 2006. Average response rate was 0.062. The purchase orders (or donations) included ordering (or donation) date and ordering amount. The last four months (Aug–Dec 2006) was used as test data for Dataset 2. The analysis process consisted of model building using each data mining technique and model assessment. An initial correlation analysis was conducted, showing that there was significant correlation among these variables, as shown in Table 4.15.

**Table 4.13** J48 decision tree

R	M	Yes	Total	P(Yes)	P(No)	Conclusion	Error
0–36		1	1	1.000		Yes	
37–152		41	619	0.066	0.934	No	41
153		605	606	0.998	0.002	Yes	1
154–257		53	1072	0.049	0.951	No	53
258–260		449	500	0.898	0.102	Yes	51
261–516		0	2227	0.000	1.000	No	
517–519		119	144	0.826	0.174	Yes	25
520–624		0	1219	0.000	1.000	No	
625		206	227	0.907	0.093	Yes	21
626–883		0	2047	0.000	1.000	No	
884		51	68	0.750	0.250	Yes	17
885–989		0	1116	0.000	1.000	No	
990		135	160	0.844	0.156	Yes	25
991–1248		0	1773	0.000	1.000	No	
1249		31	37	0.838	0.162	Yes	6
1250–1354		0	985	0.000	1.000	No	
1355		85	108	0.787	0.213	Yes	23
1356–1612		0	1290	0.000	1.000	No	
1613–1614		17	28	0.607	0.393	Yes	11
1615–1720		0	786	0.000	1.000	No	
1721		36	36	1.000	0.000	Yes	
1722–2084		14	1679	0.008	0.992	No	14
2085–2086		18	18	1.000	0.000	Yes	
2087–2343		0	831	0.000	1.000	No	
2344–2345		7	7	1.000	0.000	Yes	
2346–2448		0	404	0.000	1.000	No	
2449–2451	M > 44	21	24	0.875	0.125	Yes	3
	M ≤ 44	8	12	0.667	0.333	No	8
2452–2707		0	665	0.000	1.000	No	
2708–2710		3	5	0.600	0.400	Yes	2
2711+		26	1306	0.020	0.980	No	26
Total		1926	20,000	0.096	0.904		327

F and M appear to have a strong correlation. R and F appear to be strong predictors for customer response. Table 4.16 shows RFM limits for this dataset, and cell counts.

We built an RFM model by following the same procedures described in Dataset 1. An RFM model using a cutoff rate of 0.1 was built on half of the dataset, and tested on the other half. This yielded a model with a correct classification rate of 0.662. This was far worse than any of the other models tested.

**Table 4.14** Comparative model results—dataset 1

Model	Actual no response, model response	Actual response, model no response	Correct response	Overall correct classification
Degenerate	<b>0</b>	<b>1926</b>	18,074	<b>0.904</b>
Basic RFM on 0.1	<b>4113</b>	<b>589</b>	15,295	0.765
Basic RFM on 0.2	<b>1673</b>	<b>999</b>	17,328	0.866
Basic RFM on 0.3	<b>739</b>	<b>1321</b>	17,940	0.897
Basic RFM on 0.4	<b>482</b>	<b>1460</b>	18,058	0.903
Basic RFM on 0.5	<b>211</b>	<b>1643</b>	18,146	<b>0.907</b>
Balance using 0.5	<b>1749</b>	<b>379</b>	17,872	0.894
Value function	<b>623</b>	<b>4951</b>	14,426	0.721
Logistic regression	<b>1772</b>	<b>91</b>	18,137	<b>0.907</b>
Neural network	<b>119</b>	<b>1661</b>	18,220	<b>0.911</b>
Decision tree	<b>185</b>	<b>142</b>	19,673	<b>0.984</b>

**Table 4.15** Variable correlations

	R	F	M	Response
R	1			
F	-0.237**	1		
M	-0.125**	0.340**	1	
Response	-0.266**	0.236**	0.090**	1

\*\*Correlation is significant at the 0.01 level (2-tailed)

**Table 4.16** RFM boundaries

Factor	Min	Max	Group 1	Group 2	Group 3	Group 4	Group 5
R	1	4950	2811+	1932– 2811	935– 1932	257– 935	1–257
Count			220,229	219,411	220,212	219,503	219,654
F	1	1027	1	2	3	4	5+
Count			599,637	190,995	95,721	57,499	155,157
M	0	100,000	0–9	10–24	25–39	40–89	90+
Count			248,639	343,811	77,465	209,837	219,257

Difficulties arose in balancing cells due to F being only a few integer values (1, 2, 3, 4, 5+) and highly skewed, letting a majority of the data assigned into F group1.

Figure 4.9 displays the lift chart for the V models. The lift chart shows that the 5% of cases with the most likely response is much more likely to respond than the least responsive 50%. The proportion of responses in the test set for the 5% highest training set V scores had a response ratio of 0.311, compared to less than 0.010 for the worst 50%. We applied different V levels (0.05 and up; 0.10 and up; 0.15 and up; 0.20 and up; 0.25 and up; and 0.30 and up). These six models had very consistent results as shown in the appendix, just slightly inferior to the degenerate

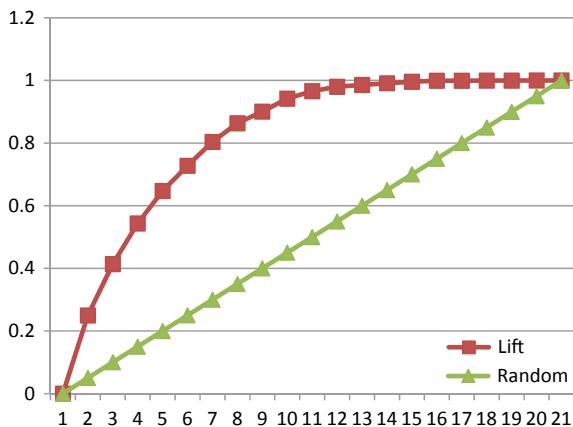


Fig. 4.9 Lift chart for dataset 2

Table 4.17 Comparative model results—second dataset

Model	Actual no response, model response	Actual response, model no response	Correct response	Overall correct classification
Degenerate	<b>0</b>	<b>34,598</b>	515,123	<b>0.9371</b>
Basic RFM	<b>4174</b>	<b>181,357</b>	364,190	0.6625
Value function > 5	<b>6212</b>	<b>30,418</b>	513,091	0.9334
Value function > 10	<b>3344</b>	<b>31,830</b>	514,547	0.9360
Value function > 15	<b>2296</b>	<b>32,475</b>	514,950	0.9367
Value function > 20	<b>1712</b>	<b>32,867</b>	515,142	0.9371
Value function > 25	<b>1400</b>	<b>33,136</b>	515,185	0.9372
Value function > 30	<b>1153</b>	<b>33,330</b>	515,238	0.9373
Logistic regression	<b>821</b>	<b>32,985</b>	515,915	<b>0.9385</b>
Neural network	<b>876</b>	<b>32,888</b>	515,957	<b>0.9386</b>
Decision tree	<b>393</b>	<b>33,373</b>	515,955	<b>0.9386</b>

model. When datasets are highly skewed as this is, with roughly only 5% responding, the degenerate model becomes very hard to beat.

All three predictive data mining models (DT, LR, NN) were built as we did in Dataset 1. The result is that those three models are performed equally in terms of accuracy (0.938), as shown in Table 4.17.

We performed gain analysis reported in Table 4.18. The predictive models using decision tree, logistic regression, and neural networks outperformed the RFM Score model. The performance gap is more significant when a small sample size (e.g., 20%) is chosen for donor solicitation.

**Table 4.18** Gains

	10%	20%	30%	40%	50%
RFM score	40.38	62.39	84.66	95.63	97.90
LR	43.24	66.22	86.10	95.75	99.75
DT	44.68	70.75	87.41	96.63	97.96
NN	43.64	67.58	86.12	95.75	99.77

## Conclusions

The results for all of the data mining classification models are quite similar, and considering that we only have two data sets tested, none can be identified as clearly better. In fact, that is typical, and it is common practice to try logistic regression, neural networks, and decision trees simultaneously to fit a given dataset. The decision tree model provides a clear descriptive reasoning with its rules. Logistic models have some interpretability for those with statistical backgrounds, but are not as clear as linear regression models. Neural network models very often fit complex data well, but have limited explanatory power.

Basic RFM analysis is based on the simplest data. It involves significant work to sort the data into cells. One traditional approach is to divide the data into 125 cells (five equally scaled divisions for recency, frequency, and monetary value). But this approach leads to highly unequal cell observations. Analysts could apply the analysis to number of customers or to dollar values purchased. Since dollar volume is usually of greater interest than simple customer count, it is expected to usually provide better decision making information. This was what we observed in our data.

The worst performance on all three dimensions typically has much higher density than other cells (Miglautsch 2002). But these customers may offer a lot of potential opportunities for growth. We demonstrated how to balance cell sizes, which involves quite a bit more data manipulation. However, it gave a much better predictive model when applied to our dataset.

Another drawback is that these three variables (R, F, and M) are not independent. Frequency and monetary value tend to be highly correlated. Yang's value function simplifies the data, focusing on one ratio. In our analysis, this data reduction led to improvements over the basic RFM model.

We then applied three classic data mining classification algorithms, all of which performed better than the RFM variants, all three giving roughly equivalent results (the PPN neural network model gave a slightly better fit, but that was the best of five neural network models applied). These models differ in their portability, however. Logistic regression provides a well-known formula of beta coefficients (although logistic output is more difficult to interpret by users than ordinary linear squares regression output). Decision trees provide the simplest to use output, as long as you are able to keep the number of rules generated to a small number (here we only had two rules, but models can involve far more rules).

Overall, the classification of customers to identify the most likely prospects in terms of future sales is very important (see Table 4.19). We have reviewed and demonstrated a number of techniques developed. We also have evaluated the relative benefits and drawbacks of each method, exhibiting a rough idea of relative accuracy based on the sample data used.

Marketing professionals have found RFM to be quite useful, primarily because the data is usually at hand and the technique is relatively easy to use. However, previous research suggests that it is easy to obtain a stronger predictive customer response model with other data mining algorithms (Olson and Chae 2012). RFM has consistently been reported to be less accurate than other forms of data mining models, but that is to be expected, as the original RFM model segmenting.

**Table 4.19** Comparison of methods

Model	Relative advantages	Relative disadvantages	Inferences
Degenerate	Tends to have high accuracy when outcome highly skewed	Mindless Simply says no Provides no marginal value	If cost of missing good responses is low, don't do anything
Basic RFM	Widely used Data readily available Software obtainable	Predictive accuracy consistently weak	Can do better using conventional data mining (RFM implicitly a special case)
RFM with balanced data	Better statistical practice	May not actually improve accuracy	Not worth the trouble
Value function	Easy to apply (uses 2 of the 3 RFM variables, so data readily available) Focuses on uncorrelated variables	Not necessarily more accurate	Value function is superior to RFM
Logistic regression	Can get better fit Can include many variables Model statistically interpretable	Logistic output harder to interpret than OLS for managers	Decision trees easier to interpret
Neural network	Can get better fit Can include many variables	Output not conducive to interpretation Can't apply model outside of software used to build model	Decision trees easier to interpret
Decision trees	Can get better fit Can include many variables Output easily understandable by managers	Model may involve an excessive number of rules	Best option, if can control the number of rules obtained (through minimum required response parameter)

Balancing cell sizes by adjusting the limits for the three RFM variables is sound statistically, but did not lead to improved accuracy in our tests. In both Dataset 1 and Dataset 2, the basic RFM model significantly underperformed other predictive models, except the V function model in Dataset 1. These results indicate that balancing cells might help improve fit, but involves significant data manipulation for very little predictive improvement in the data set we examined.

Using the V ratio is an improvement to RFM that is useful in theory, but in our tests the results are mixed. In Dataset 1, the technique did not provide better predictive accuracy. In Dataset 2, it did yield an improved classification rate but underperformed the degeneracy model. Thus, this technique deserves a further inquiry. Overall, the results above indicate that some suggested alternatives to the traditional RFM have limitations in prediction.

The primary conclusion of our study, as was expected, is that classical data mining algorithms outperformed RFM models in terms of both prediction accuracy and cumulative gains. This is primarily because decision tree, logistic regression, and neural networks are often considered the benchmark “predictive” modeling techniques.

While we used predication accuracy along with cumulative gains for model comparison, in practice the type of error can be considered in terms of relative costs, thus enabling influence on profit. For example, our study shows that increasing the cutoff level between predicting response or not can improve correct classification. However, a more precise means to assess this would be to apply the traditional cost function reflecting the cost of the two types of error. This is to be a consideration in evaluating other predictive models as well. Thus, specific models should be used in light of these relative costs.

The good performance of those data mining methods (particularly decision tree), in terms of prediction accuracy and cumulative gains, indicates that three variables (R, F, and M) alone can be useful for building a reliable customer response model. This echoes the importance of RFM variables in understanding customer purchase behavior and developing response models for marketing decisions. Including non-RFM attributes (e.g., income) is likely to slightly improve the model performance. However, a sophisticated model with too many variables is not very effective for marketing practitioners and reducing variables is important for practical use of predictive models. Marketers should be aware of this tradeoff between a simple model (with fewer variables) and a sophisticated model (with a large number of variables) and develop a well-balanced model using their market and product knowledge.

## References

- Kohavi R, Parekh R (2004) Visualizing RFM segmentation. In: Proceedings of the 2004 SIAM international conference on data mining, pp 391–399  
Miglautsch J (2002) Application of RFM principles: what to do with 1-1-1 customer? J Database Mark 9(4):319–324

- Olson DL, Chae B (2012) Direct marketing decision support through predictive customer response modeling. *Decis Support Syst* 54(1):443–451
- Olson DL, Cao Q, Gu C, Lee D-H (2009) Comparison of customer response models. *Serv Bus* 3 (2):117–130
- Yang ZX (2004) How to develop new approaches to RFM segmentation. *J Target Measur Anal Mark* 13(1):50–60

# Chapter 5

## Association Rules



Association rules seek to identify combinations of things that frequently occur together (**affinity analysis**). This is also the basis of market basket analysis, which we discussed in terms of correlation and Jaccard ratios. Association rules take things a step further by applying a form of machine learning, the most common of which is the apriori algorithm.

Association rules can provide information can be useful to retailers in a number of ways:

- Identify products that could be placed together when customers interested in one are likely to be interested in the other
- Targeting customers through campaigns (coupons, mailings, e-mailings, etc.) seeking to get them to expand the products purchased
- In on-line marketing, drive recommendation engines.

Outside of retailing, there are other uses for association rules. There are many uses of association rules. Classically, they were applied to retail transaction analysis, akin to market basket analysis. With the emergence of big data, the ability to apply association rules to streams of real-time data is highly useful, enabling a great deal of Web mining for many applications, including e-business retail sales. Association rule mining is one of the most widely used data mining techniques. This can be applied to target marketing, by customer profile, space allocation strategy within stores, but can also be extended to business applications such as international trade and stock market prediction. In the sciences, remotely sensed imagery data has been analyzed to aid precision agriculture and resource discovery (to include oil). It has been used in manufacturing to analyze yield in semiconductor manufacturing. It has been used to improve efficiency of packet routing over computer networks. In medicine it has been used for diagnosis of diseases. They also could be used in human resources management and other places where pairing behavior with results are of interest (Aguinis et al. 2013).

## Methodology

Association rules deal with **items**, which are the objects of interest. In the case of the pseudo-Amazon data used in the prior chapter, these would be the products marketed. Association rules group items into sets representing groups of items tending to occur together (an example being a transaction). Rules have the form of an item set on the left (antecedent) with a consequence on the right. For instance, if a customer bought an ebook, correlation indicated a strong likelihood of buying a paperback book.

A limitation of association rule analysis is the enormous number of combinations. Furthermore, the data includes many null entries. Software deals with this quite well, but it leads to burying interesting rules within many meaningless negative relationships. Our correlation and Jaccard analysis from the Market Basket chapter found that the pseudo-Amazon dataset is dominated by ebooks, hardbacks, and paperbacks. Table 5.1 provides the counts of the transactions for eight combinations of these three book products.

Totals are 619 ebook sales 493 hardbacks, and 497 paperbacks. There are 12 pairs of these three variables plus 8 triplets. There are 48 rules from only one antecedent, plus another 24 with two antecedents, yielding 72 possible rules. An example rule might be:

IF{ebook} THEN {paperback}

This can be extended to multiple conditions:

IF{ebook & hardback} THEN {paperback}

There are measures of interest for rules. The **support** of an item or item set is the proportion of transactions containing that item set. In the pseudo-Amazon dataset, there are 619 cases out of 1000 where an ebook was purchased. Thus its support is 0.619. Rules have **confidence** measures, defined as the consequent occurs if the antecedent is present. Thus confidence gives the probability of paperbacks being bought should ebooks have been bought in the example above. Of the 619

**Table 5.1** Pseudo-amazon cases

Ebooks	Hardbacks	Paperbacks	Count
Yes	Yes	Yes	419
Yes	Yes	No	56
Yes	No	Yes	64
Yes	No	No	80
No	Yes	Yes	7
No	Yes	No	11
No	No	Yes	7
No	No	No	356

customers who purchased one or more ebooks, 483 also purchased one or more paperbacks. Thus confidence in this case is  $483/619 = 0.780$ . The **lift** of a rule is really the same as described in the market basket chapter, although the formulas given for the context of association rules looks a bit different at first glance. The conventional formula is support{antecedent & consequent} divided by the support of the antecedent times the support of the consequent. This is equivalent to the confidence of the rule divided by (support of the consequent). For the ebook and paperback rule above, this would be 0.780 divided by the average propensity for customers who purchased paperbacks ( $497/1000$ ), yielding lift of 1.57. Most sources give lift as support for the rule divided by (the independent support of the antecedent times the independent support of the consequent), or in this case  $(483/1000)/[(619/1000 \times 497/1000)]$  also equal to 1.57. Cursory inspection of course explains this as the support for the rule is  $483/619 = 0.780$ . Algorithm users are allowed to set minimum support and confidence levels. **Frequent sets** are those for which support for the antecedent is at least as great as the minimum support level. **Strong sets** are frequent and have confidence at least as great as the minimum confidence level.

## The Apriori Algorithm

The apriori algorithm is credited to Agrawal et al. (1993) who applied it to market basket data to generate association rules. Association rules are usually applied to binary data, which fits the context where customers either purchase or don't purchase particular products. The apriori algorithm operates by systematically considering combinations of variables, and ranking them on either support, confidence, or lift at the user's discretion.

The apriori algorithm operates by finding all rules satisfying minimum confidence and support specifications. First, the set of frequent 1-itemsets is identified by scanning the database to count each item. Next, 2-itemsets are identified, gaining some efficiency by using the fact that if a 1-itemset is not frequent, it can't be part of a frequent itemset of larger dimension. This continues to larger-dimensioned itemsets until they become null. The magnitude of effort required is indicated by the fact that each dimension of itemsets requires a full scan of the database. The algorithm is:

To identify the candidate itemset  $C_k$  of size  $k$

1. Identify frequent items  $L_1$

For  $k = 1$  generate all itemsets with support  $\geq \text{Support}_{\min}$   
If itemsets null, STOP

Increment  $k$  by 1

For itemsets of size  $k$  identify all with support  $\geq \text{Support}_{\min}$

END

2. Return list of frequent itemsets
3. Identify rules in the form of antecedents and consequents from the frequent items
4. Check confidence of these rules.

If confidence of a rule meets  $\text{Confidence}_{\min}$  mark this rule as strong.

The output of the apriori algorithm can be used as the basis for recommending rules, considering factors such as correlation, or analysis from other techniques, from a training set of data. This information may be used in many ways, including in retail where if a rule is identified indicating that purchase of the antecedent occurred without that customer purchasing the consequent, then it might be attractive to suggest purchase of the consequent.

The apriori algorithm can generate many frequent itemsets. Association rules can be generated by only looking at frequent itemsets that are strong, in the sense that they meet or exceed both minimum support and minimum confidence levels. It must be noted that this does not necessarily mean such a rule is useful, that it means high correlation, nor that it has any proof of causality. However, a good feature is that you can let computers loose to identify them (an example of machine learning).

To demonstrate using data from Table 5.1, establish  $\text{Support}_{\min} = 0.4$  and  $\text{Confidence}_{\min} = 0.5$ :

1.  $L_1 = \text{Ebooks}$  (support 0.619),  $\text{Paperbacks}$  (support 0.497), and  $\text{Hardbacks}$  (support 0.493);  $\text{noHardbacks}$  (support 0.507),  $\text{no Paperbacks}$  (support 0.503). The item  $\text{noEbooks}$  fails because its support of 0.381 is below  $\text{Support}_{\min}$ .
2.  $L_2 = \text{Ebooks} \& \text{ Hardbacks}$  (support 0.475),  $\text{Ebooks} \& \text{ Paperbacks}$  (support 0.483),  $\text{Hardbacks} \& \text{ Paperbacks}$  (support 0.426), and  $\text{noHardbacks} \& \text{ noPaperbacks}$  (support 0.436). Itemsets  $\text{Ebooks} \& \text{ noHardbacks}$  fail with support of 0.144,  $\text{Ebooks} \& \text{ noPaperbacks}$  with support of 0.136,  $\text{Hardbacks} \& \text{ noPaperbacks}$  with support of 0.067,  $\text{Paperbacks} \& \text{ noEbooks}$  with support of 0.014,  $\text{Paperbacks} \& \text{ noHardbacks}$  with support of 0.071,  $\text{noEbooks} \& \text{ noPaperbacks}$  with support of 0.367.
3.  $L_3 = \text{Ebooks} \& \text{ Hardbacks} \& \text{ Paperbacks}$  (support 0.419). Itemsets  $\text{Ebooks} \& \text{ Hardbacks} \& \text{ noPaperbacks}$  fail with support of 0.056,  $\text{Ebooks} \& \text{ Paperbacks} \& \text{ noHardbacks}$  with support of 0.064, and  $\text{Hardbacks} \& \text{ Paperbacks} \& \text{ Ebooks}$  with support of 0.007.
4. There aren't four items, so  $L_4$  is null.
5. Identify rules from frequent items:

$\text{Ebooks} \rightarrow \text{Hardbacks}$	Confidence 0.767
$\text{Ebooks} \rightarrow \text{Paperbacks}$	Confidence 0.780
$\text{Hardbacks} \rightarrow \text{Ebooks}$	Confidence 0.963
$\text{Hardbacks} \rightarrow \text{Paperbacks}$	Confidence 0.850
$\text{Paperbacks} \rightarrow \text{Ebooks}$	Confidence 0.972
$\text{Paperbacks} \rightarrow \text{Hardbacks}$	Confidence 0.843

noHardbacks → noPaperbacks	Confidence 0.860
noPaperbacks → noHardbacks	Confidence 0.867
Ebooks & Hardbacks → Paperbacks	Confidence 0.882
Ebooks & Paperbacks → Hardbacks	Confidence 0.886
Hardbacks & Paperbacks → Ebooks	Confidence 0.984

All other combinations of frequent itemsets in  $L_3$  failed the minimum support test.

These rules now would need to be evaluated, possibly subjectively by the users, for interestingness. Here the focus is on cases where a customer who buys one type of book might be likely according to this data to buy the other type of books. Another indication is that if a customer never bought a paperback, they are not likely to buy a hardback, and vice versa.

## Association Rules from Software

R allows setting support and confidence levels, as well as the minimum rule length. It has other options as well. We will set support and confidence (as well as lift, which is an option for sorting output) below. Our pseudo-Amazon database has 1000 customer entries (which we treat as transactions). The data needs to be put into a form the software will read. In Rattle, that requires data be categorical rather than numerical. The rules generated will be positive cases (IF you buy diapers THEN you are likely to buy baby powder) and negative cases are ignored (IF you **didn't** buy diapers THEN you are likely to do whatever). If you wish to study the negative cases, you would need to convert the blank cases to No. Here we will demonstrate the positive case.

Association rule mining seeks all rules satisfying specified minimum levels. Association rules in R and WEKA require nominal data, an extract of this is shown in Table 5.2.

R's Association Rule screen is shown in Fig. 5.1.

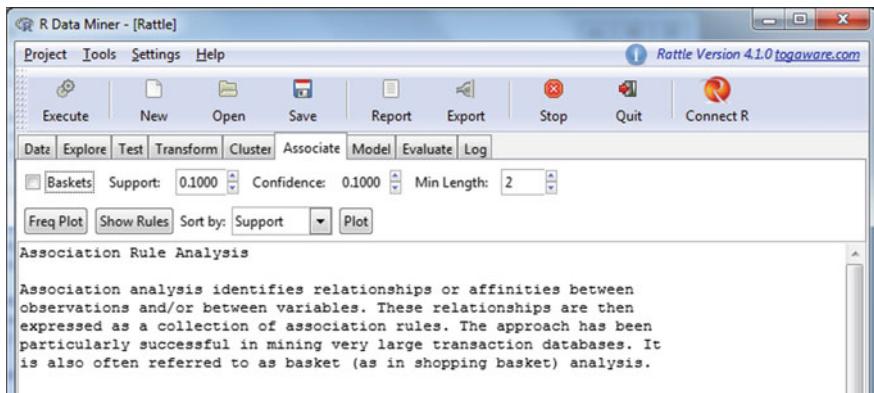
**Table 5.2** Extract of full pseudo-amazon dataset

Auto	Baby	Ebooks	Hard	Paper	Music	Elect	Health	GiftC
	Yes							
	Yes							
Yes								
	Yes		Yes					
	Yes				Yes			
	Yes							

Selecting the options given in Fig. 5.1 yields nine rules after Execute (see Fig. 5.2).

This shows that nine rules are generated (see Fig. 5.4). Minimum support and confidence control the number of rules. Since the minimum support in the dataset is 0.419 and minimum confidence 0.7552, one should obtain nine rules up to those levels, which is the case (see Fig. 5.3).

The nine rules generated are displayed in Fig. 5.4.



**Fig. 5.1** Screen of R's association rule tab

```

  Execute | New | Open | Save | Export | Stop | Quit |
Data | Explore | Test | Transform | Cluster | Associate | Model | Evaluate | Log |
 Baskets | Support: 0.1000 | Confidence: 0.1000 | Min Length: 2 |
Freq Plot | Show Rules | Sort by: Support | Plot |

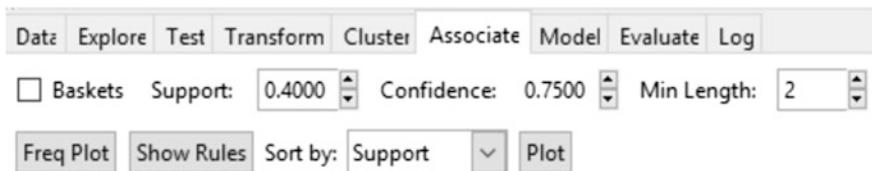
Summary of the Transactions:
Length          Class          Mode
1000 transactions          S4

Summary of the Apriori Association Rules:
Number of Rules: 9

Summary of the Measures of Interestingness:
support      confidence      lift      count
Min. :0.4190  Min. :0.7552  Min. :1.532  Min. :419.0

```

**Fig. 5.2** Rattle association screen

**Fig. 5.3** Setting support and confidence

```
All Rules

lhs                      rhs          support  confidence  lift      count
[1] {Paper=Yes}           => {EBooks=Yes} 0.483    0.9718310  1.545041 483
[2] {EBooks=Yes}           => {Paper=Yes}  0.483    0.7678855  1.545041 483
[3] {Hard=Yes}             => {EBooks=Yes} 0.475    0.9634888  1.531779 475
[4] {EBooks=Yes}           => {Hard=Yes}   0.475    0.7551669  1.531779 475
[5] {Paper=Yes}            => {Hard=Yes}   0.426    0.8571429  1.738626 426
[6] {Hard=Yes}             => {Paper=Yes}  0.426    0.8640974  1.738626 426
[7] {Hard=Yes, Paper=Yes}  => {EBooks=Yes} 0.419    0.9835681  1.563701 419
[8] {EBooks=Yes, Paper=Yes}=> {Hard=Yes}   0.419    0.8674948  1.759624 419
[9] {EBooks=Yes, Hard=Yes} => {Paper=Yes}  0.419    0.8821053  1.774860 419
```

**Fig. 5.4** Rattle association rules

Running association rules on either WEKA or R involves having the computer do enormous numbers of combinatorial calculations. They are quite good at that. Software allows users to specify minimum support and confidence (and in WEKA's case even specify the maximum number of rules). The impact of different support levels is indicated in Table 5.3, where mean values for support, confidence, and lift are given, along with the number of rules obtained.

To demonstrate what is going on, we focus on only three of these variables: Ebooks, Hardbacks, and Paperbacks. Table 5.4 shows the association rules generated by R sorted by lift.

What are the implications? For this data, grouping EBooks, Hard Cover Books, and Paperback Books together yields the greatest lift.

**Table 5.3** Rules obtained by support level specified

Specified support	Specified confidence	Min length	Support obtained	Confidence obtained	Lift
0.1	0.1	1	0.4186	0.7552	1.532
0.1	0.1	2	0.4186	0.7552	1.532
0.4	0.75	2	0.4186	0.7552	1.532
0.46	0.75	2	0.475	0.7552	1.532
0.48	0.75	2	0.483	0.77	1.545
0.48	0.77				

**Table 5.4** R association rules

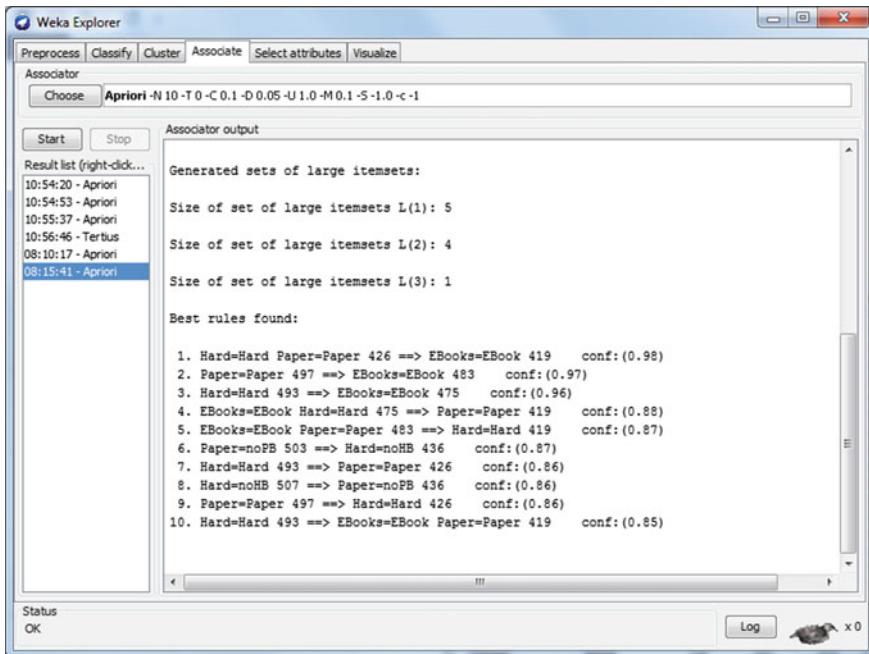
ID	Antecedent	Consequent	Support	Confidence	Lift	Count
1	EBooks, Hard	Paper	0.42	0.88	1.77	419
2	EBooks, Paper	Hard	0.42	0.87	1.76	419
3	Hard	Paper	0.43	0.86	1.74	426
4	Paper	Hard	0.43	0.86	1.74	426
5	Hard, Paper	EBooks	0.42	0.98	1.56	419
6	Paper	EBooks	0.48	0.97	1.55	483
7	EBooks	Paper	0.48	0.77	1.55	483
8	EBooks	Hard	0.48	0.76	1.53	475
9	Hard	EBooks	0.48	0.96	1.53	475
10	{}	Software	0.11	0.11	1	111
11	{}	Toys	0.11	0.11	1	112
12	{}	Movies	0.13	0.13	1	132
13	{}	Music	0.12	0.12	1	118
14	{}	Paper	0.50	0.50	1	497
15	{}	Hard	0.49	0.49	1	493
16	{}	EBooks	0.63	0.63	1	629

WEKA (and almost all other data mining software) also supports association rule mining. WEKA allows a number of metrics by which to evaluate association rules. Confidence is the proportion of the examples covered by the premise that are also covered by the consequence (Class association rules can only be mined using confidence). Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence. Conviction is given by:

$$[1 - \text{Support}(\text{consequent})] / [1 - \text{Confidence}(\text{IF antecedent THEN consequent})]$$

This is the ratio of the probability of the antecedent occurring without the consequent divided by the observed frequency of incorrect cases. The WEKA screen is shown in Fig. 5.5.

Correlation found the strongest relationship between purchase of hard cover books and paperbacks. R ranked the rules combining these two options number 11, 12, 13 and 14 based on lift. WEKA ranked these combinations 6, 7, 8 and 9 in Fig. 5.2 based on confidence. The strongest Jaccard coefficient was found between Ebooks and paperbacks. R (based on lift) ranked these 1, 2, 12 and 13, while WEKA (based on confidence) ranked one combination 2 in Fig. 5.2. The point is that there are lots of options in measuring what things go together, and different metrics will yield different results.



**Fig. 5.5** WEKA association rule output for top 10 by confidence

## Non-negative Matric Factorization

There are advanced methods applied to association rule generation. Non-negative matrix factorization (NMF) was proposed by Lee and Seung (1999) as a means to distinguish parts of data for facial recognition as well as text analysis. Principal components analysis and vector quantization learn holistically rather than breaking down data into parts. These methods construct factorizations of the data. For instance, if there is a set of customers  $N$  and a set of products  $M$  a matrix  $V$  can be formed where each row of  $V$  represents a market basket with one customer purchasing products. This can be measured in units or in dollars. Association rules seek to identify ratio rules identifying the most common pairings. Association rule methods, be they principal components analysis or other forms of vector quantization, minimize dissimilarity between vector elements. Principal components allows for negative associations, which in the context of market baskets does not make sense. NMF imposes non-negativity constraints into such algorithms.

## Conclusion

Association rules are very useful in that they provide a machine-learning mechanism to deal with the explosion of big data. This can be for good or bad, as in any data mining application. Real-time automatic trading algorithms have caused damage in stock markets, for instance. However, they provide great value not only to retail analysis (to serve customers better), but also in the medical field to aid in diagnosis, in agriculture and manufacturing to suggest greater efficient operations, and in science to establish expected relationships in complex environments.

Implementing association rules is usually done through the apriori algorithm, although refinements have been produced. This requires software for implementation, although that is available in most data mining tools, commercial or open source. The biggest problem with association rules seems to be sorting through the output to find interesting results.

## References

- Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S (eds) Proceedings of the 1993 ACM SIGMOD international conference on management of data, 207–216. Association for Computing Machinery, New York
- Aguinis H, Forcum LE, Joo H (2013) Using market basket analysis in management research. *J Manag* 39(7):1799–1824
- Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791

# Chapter 6

## Cluster Analysis



This chapter covers a number of aspects of cluster analysis. Initially, it presents clustering manually, using standardized data. This is to show how basic algorithms work. The second section shows how software works on this standardized data. The third section will demonstrate software with original data not requiring standardization. If you don't care what computers are doing, you can proceed to this section.

Cluster analysis is usually used as an initial analytic tool, enabling data mining analysts the ability to identify general groupings in the data. It often follows initial graphical display of data, and provides a numeric means to describe underlying patterns. This can include pattern identification. Cluster analysis is thus a type of model, but one that is usually applied in the process of data understanding.

Clustering involves identification of groups of observations measured over variables. Here we distinguish between discriminant analysis (where the groups are given as part of the data, and the point is to predict group membership) and cluster analysis (where the clusters are based on the data, and thus not pre-determined, and the point is to find items that belong together rather than to predict group membership). Cluster analysis is an unsupervised technique, where data is examined without reference to a response variable. (You can include outcome variables in cluster analysis, but they are treated just as any other variable, and don't play a role in directing the search.) It thus is an example of machine learning, where the value of clustering models is their ability to capture interesting data groupings. The technique requires a large enough dataset to establish statistical significance, but conversely suffers from the curse of dimensionality in that the more variables and values these variables can take on, the more difficult the computational task. A typical use is to initially apply clustering models to identify segments of the data that are used in subsequent predictive analyses. There are a number of techniques used for cluster analysis.

## K-Means Clustering

The most general form of clustering analysis allows the algorithm to determine the number of clusters. At the other extreme, the number of clusters may be pre-specified. Partitioning is used to define new categorical variables that divide the data into a fixed number of regions (k-means clustering, for example). A common practice is to apply factor analysis as a pre-processing technique to get a reasonable idea of the number of clusters, as well as to give managers a view of which types of items go together. Given a number ( $k$ ) of centers, data observations are assigned to that center with the minimum distance to the observation. A variety of distance measures are available, although conventionally the centroid (a centroid has the average value—mean, median, etc.—for each variable) of each cluster is used as the center, and squared distance (or other metric) is minimized. This is the most widely used form of cluster analysis in data mining.

Cluster analysis has been used by data miners to segment customers, allowing customer service representatives to apply unique treatment to each segment. Data needs to be numerical for clustering algorithms to work. We will demonstrate methods with standardized data (ranging from 0 to 1) because metrics will be warped by different measurement scales. This isn't necessary with software, as most datamining software does this for you in the algorithm.

### *A Clustering Algorithm*

The following is a simple k-means algorithm (Johnson and Wichern 1998):

1. Select the desired number of clusters  $k$  (or iterate from 2 to the maximum number of clusters desired).
2. Select  $k$  initial observations as seeds (could be arbitrary, but the algorithm would work better if these seed values were as far apart as possible).
3. Calculate average cluster values over each variable (for the initial iteration, this will simply be the initial seed observations).
4. Assign each of the other training observations to the closest cluster, as measured by squared distance (other metrics could be used, but squared distance is conventional).
5. Recalculate cluster averages based on the assignments from Step 4.
6. Iterate between steps 4 and 5 until the same set of assignments are obtained twice in a row.

Note that this algorithm does not guarantee the same result no matter what the initial seeds are. However, it is a relatively straightforward procedure. The problem of how to determine  $k$  can be dealt with by applying the procedure for 2 clusters, then for 3, and so forth until the maximum desired number of clusters is reached.

Selecting from among these alternatives may be relatively obvious in some cases, but can be a source of uncertainty in others.

There are some drawbacks to k-means clustering. The data needs to be put in standardized form to get rid of the differences in scale. However, even this approach assumes that all variables are equally important. If there are some variables more important than others, weights can be used in the distance calculation, but determining these weights is another source of uncertainty.

## ***Loan Data***

This data set consists of information on applicants for appliance loans, and was used in Chap. 2 to demonstrate visualization. We will use the loan application dataset to demonstrate clustering software. The business purpose here is to identify the type of loan applicants least likely to have repayment problems. In the dataset, an outcome of On-Time is good, and Late is bad. Distance metrics are an important aspect of cluster analysis, as they drive algorithms, and different scales for variable values will lead to different results. Thus we will transform data for demonstration of how clustering works. We will use 400 of the observations for cluster analysis. Transformation of data to standardized form (between 0 and 1) is accomplished as follows:

Age	<20	0
	20–50	(Age-20)/30
	50–80	1 – (age-50)/30
	>80	0
Income	<0	0
	0 to \$100,000	Income/100,000
	>\$100,000	1
Risk	Max 1, min 0	Assets/(debts + want) (higher is better)
Credit	Green	1
	Amber	0.3
	Red	0

The standardized values for the data given in Table 2.2 of Chap. 2 is shown in Table 6.1.

Dividing the data into 400 for a training set and retaining 250 for testing, simply identifying average attribute values for each given cluster, we begin with sorting the 400 training cases into on-time and late categories, and identifying the average performance by variable for each group.

These averages are shown in Table 6.2.

**Table 6.1** Standardized loan data

Age	Income	Risk	Credit	On-time
0	0.17152	0.531767	1	1
0.1	0.25862	0.764475	1	1
0.266667	0.26169	0.903015	0.3	0
0.1	0.21117	0.694682	0	0
0.066667	0.07127	1	0.3	1
0.2	0.42083	0.856307	0	0
0.133333	0.55557	0.544163	1	1
0.233333	0.34843	0	0	1
0.3	0.74295	0.882104	0.3	1
0.1	0.38887	0.145463	1	1
0.266667	0.31758	1	1	1
0.166667	0.8018	0.449404	1	0
0.433333	0.40921	0.979941	0.3	0
0.533333	0.63124	1	1	1
0.633333	0.59006	1	1	1
0.633333	1	1	0.3	1
0.833333	0.80149	1	1	1
0.6	1	1	1	1
0.3	0.81723	1	1	1
0.566667	0.99522	1	1	1

**Table 6.2** Group standard score averages for loan application data

Cluster	On-time	Age	Income	Risk	Credit
C1 (355 cases)	1	0.223	0.512	0.834	0.690
C2 (45 cases)	0	0.403	0.599	0.602	0.333

Cluster 1 included members that tended to be younger with better risk measures and credit ratings. Income tended to be the same for both, although Cluster 1 members had slightly lower incomes.

Step 3 of the k-means algorithm calculates the ordinary least squares distance to these cluster averages. This calculation for test case 1 to Cluster 1, using Age, Income, Risk and Credit standardized scores, would be:

$$(0.223 - 0.967)^2 + (0.512 - 0.753)^2 + (0.834 - 1)^2 + (0.690 - 0)^2 = 1.115$$

The distance to Cluster 2 is:

$$(0.403 - 0.967)^2 + (0.599 - 0.753)^2 + (0.602 - 1)^2 + (0.333 - 0)^2 = 0.611$$

Because the distance to Cluster 2 (0.611) is closer than to Cluster 1 (1.115), the algorithm would assign Case 1 to Cluster 2.

If there were a reason to think that some variables were more important than others, you could apply weights to each variable in the distance calculations. In this case, there clearly are two interesting classes, so we might stop at two clusters. However, in principle, you can have as many clusters as you want. The basic distance calculation is the same. Analyzing the differences in clusters accurately depends upon knowledge of the underlying data.

## Clustering Methods Used in Software

The most widely used clustering methods are hierarchical clustering, Bayesian clustering, K-means clustering, and self-organizing maps. Hierarchical clustering algorithms do not require specification of the number of clusters prior to analysis. However, they only consider local neighbors at each stage, and cannot always separate overlapping clusters. The two-step method is a form of hierarchical clustering. Two-step clustering first compresses data into subclusters, and then applying a statistical clustering method to merge subclusters into larger clusters until the desire number of clusters is reached. Thus the optimal number of clusters for the training set will be obtained. Bayesian clustering also is statistically based. Bayesian clustering is based on probabilities. Bayesian networks are constructed with nodes representing outcomes, and decision trees constructed at each node. K-means clustering involves increasing the number of clusters as demonstrated earlier. Software allows you to specify the number of clusters in K-means. Self-organizing maps use neural networks to convert many dimensions into a small number (like two), which has the benefit of eliminating possible data flaws such as noise (spurious relationships), outliers, or missing values. K-means methods have been combined with self-organizing maps as well as with genetic algorithms to improve clustering performance.

K-means algorithms work by defining a fixed number of clusters, and iteratively assigning records to clusters. In each iteration, the cluster centers are redefined. The reassignment and recalculation of cluster centers continues until any changes are below a specified threshold. The methods demonstrated earlier in this chapter fall into this class of algorithm.

Kohonen self-organizing maps (SOM, or Kohonen networks) are neural network applications to clustering. Input observations are connected to a set of output layers, with each connection having a strength (weight). A general four-step process is applied (Kohonen 1997):

1. **Initialize map:** A map with initialized reference vectors is created, and algorithm parameters such as neighborhood size and learning rate are set.
2. **Determine winning node:** For each input observation, select the best matching node by minimizing distance to an input vector. The Euclidean norm is usually used.
3. **Update reference vectors:** Reference vectors and its neighborhood nodes are updated based upon the learning rule.
4. **Iterate:** Return to step 2 until the selected number of epochs is reached, adjusting neighborhood size.

Small maps (a few hundred nodes or less) are recommended. Large neighborhood sizes and learning rates are recommended initially, but can be decreased. With small maps, these parameters have not been found to be that important. There are a number of variants, to include self-organizing tree maps (Astudillo and Oommen 2011) and self-organizing time maps (Sarlin 2013). Self-organizing maps are a useful tool for machine learning as applied to cluster analysis.

## Software

Chapter 2 introduced Rattle, a GUI interface for the open software R. We will demonstrate clustering with the three datasets using software. R has four algorithms—Kmeans, Ewkm (entropy weighted k-means), Hierarchical, and BiCluster. KNIME has eight options, to include K-means, K-medoids (medians), hierarchical, self-organizing maps, and fuzzy c-means. WEKA has eleven algorithms for continuous data, to include simple K-means and Farthest-First algorithms. K-means clustering assigns records to a specified number of clusters through iteratively adjusting the cluster centers (much as described earlier in this chapter). We will compare K-means (with 2 clusters) for these three softwares. We use the loan dataset for demonstration.

### **R (Rattle) K-Means Clustering**

We use an expanded loan dataset, adding a FICO score. Assets, Debts, and Want were used to generate the variable credit. Raw data prior to standardization is used (Rattle standardizes internally to run clustering). The data screen for the dataset LoanFICOcluster.csv (650 observations) is shown in Fig. 6.1. Here there are six input variables (risk is called Safety to reflect higher is better, and On-Time is made an input variable to enable interpretation). All are numeric (required for clustering).

Moving to the Explore page (Fig. 6.2) we get descriptions of data type and summary statistics to include minima, maxima, and quartile values as well as means.

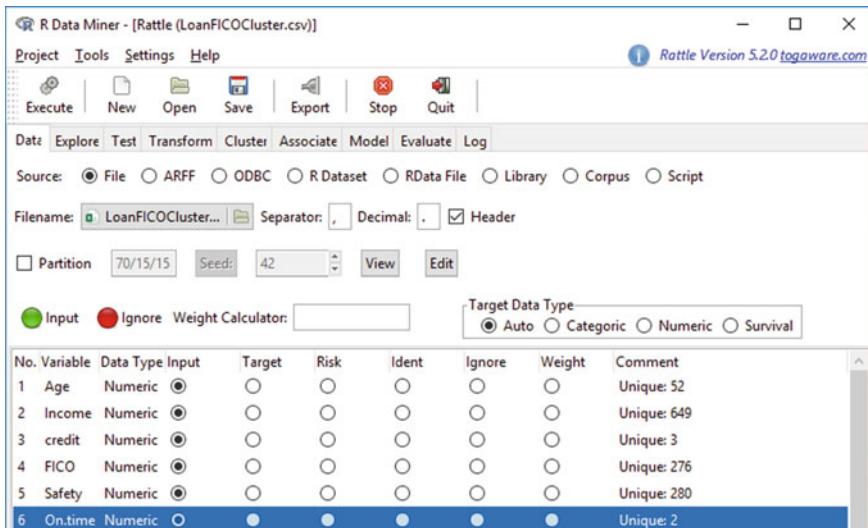


Fig. 6.1 Rattle display of loaded clustering data

Asking for Distributions and selecting Box Plots for each variable yields Fig. 6.3.

These are box-and-whisker plots displaying the means (asterisks), quartiles (box ends) and outliers (outlying dots). Correlation can be obtained by selecting the **Correlation** radio button, yielding Fig. 6.4.

The graphic shows Credit, Risk, and Age to have strong relationships to On-Time payment, with Income having a much weaker relationships. Figure 6.5 displays this information in tabular form.

Figure 6.5 shows that the relationship between On-Time to each variable is low, but positive (and each above 0.1). There is a strong relationship between FICO and Credit (after all they are seeking to measure the same thing) as well as Safety and Age (older people have safer financial situations). Correlation shows cross-relationships among inputs. Income is related to Age, and better Safety measures are strongly related to Age.

Cluster analysis was applied using the K-Means algorithm in Rattle. The user specifies the number of clusters. Rattle's cluster screen using KMeans with 2 clusters is shown in Fig. 6.6.

Select the **Execute** button, we obtained the output given in Fig. 6.7.

Note that the Re-Scale box is checked. This reports the clusters in standardized data. Here the results indicate both clusters had relatively high on time performance. The biggest difference is in credit rating, with FICO (highly correlated with credit) also clearly different. The rest of the cluster measures are relatively similar. The implication drawn is that cluster 1 contains 277 cases (see Cluster sizes) with lower credit scores but only slightly weaker on-time performance to the 373 in cluster 2.

The screenshot shows the R Data Miner application window titled "R Data Miner - [Rattle (LoanFICOCluster.csv)]". The menu bar includes Project, Tools, Settings, Help, and a Rattle Version 5.2.0 link. Below the menu is a toolbar with Execute, New, Open, Save, Export, Stop, and Quit buttons. A navigation bar below the toolbar has tabs for Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, and Log, with Data selected. Under the Data tab, there are checkboxes for Type (Summary, Distributions, Correlation, Principal Components, Interactive), Summary (checked), and other descriptive options like Describe, Basics, Kurtosis, Skewness, Show Missing, and Cross Tab. The main pane displays the following text:

```

Below we summarise the dataset.

Data frame:crs$dataset[, c(crs$input, crs$risk, crs$target)]    650 observations and 6 var

  Storage
Age   integer
Income  integer
credit  double
FICO   integer
Safety  double
On.time integer

For the simple distribution tables below the 1st and 3rd Qu. refer to the first and third quartiles, indicating that 25% of the observations have values of that variable which are less than or greater than (respectively) the value listed.

  Age      Income      credit      FICO
Min. :20.0    Min. : 7127    Min. :0.0000    Min. :402.0
1st Qu.:26.0   1st Qu.: 39360   1st Qu.:0.3000   1st Qu.:645.2
Median :31.0   Median : 57182   Median :1.0000   Median :723.0
Mean   :32.9   Mean   : 61236   Mean   :0.6662   Mean   :706.8
3rd Qu.:38.0   3rd Qu.: 76247   3rd Qu.:1.0000   3rd Qu.:786.5
Max.   :92.0    Max.   :216844   Max.   :1.0000   Max.   :849.0
  Safety      On.time
Min. :0.0000    Min. :0.0
1st Qu.:0.6601   1st Qu.:1.0
Median :0.9801   Median :1.0
Mean   :0.7849   Mean   :0.9
3rd Qu.:1.0000   3rd Qu.:1.0
Max.   :1.0000   Max.   :1.0

```

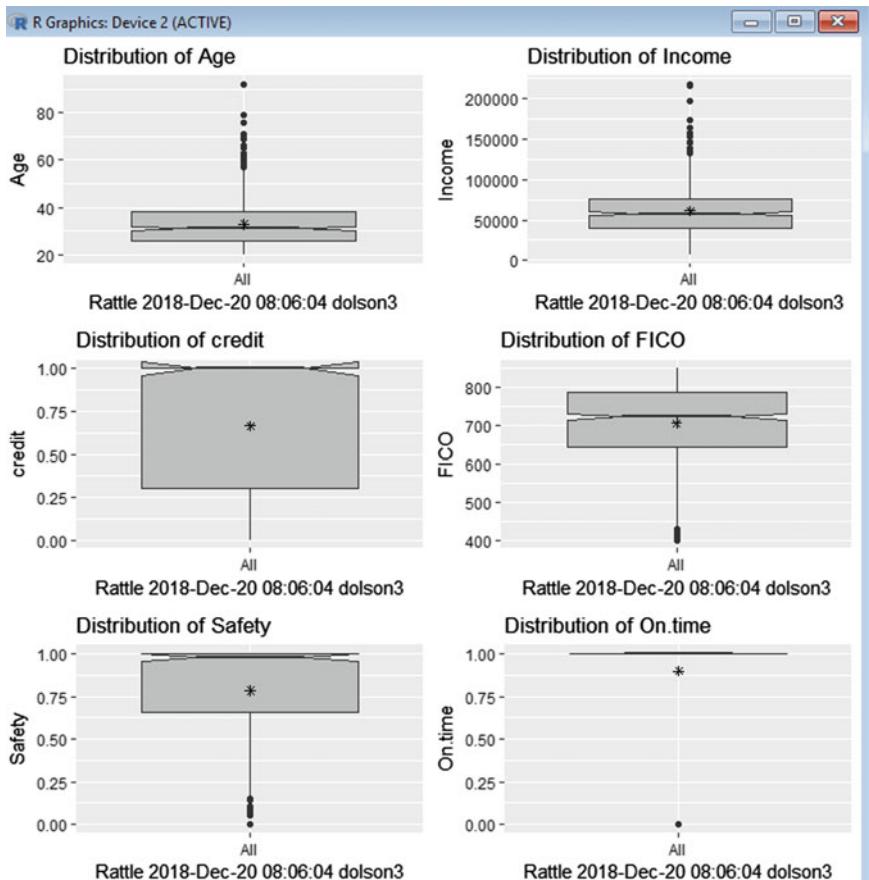
At the bottom of the main pane is a "Find" input field and "Find", "Next", and "Previous" buttons.

**Fig. 6.2** Initial summary output

We can see numbers in more human terms by unchecking the Re-Scale box and obtaining the clusters in Fig. 6.8.

Note that now the numbers are in terms of the data. The clusters are different, although similar. Here cluster 1 is the slightly better on-time group (226 instead of 373), and here Income has a much greater difference, while credit, FICO, and Safety don't. So Re-Scaling clearly makes a difference in the clusters obtained. Cluster results tend to be unpredictable, with any change often making quite a difference in output.

There are other tools provided by Rattle clustering. The Stats button gives detailed statistics which don't usually have much interest. The Data button, however, provides a view of how each variable is grouped by cluster, as shown in Fig. 6.9.

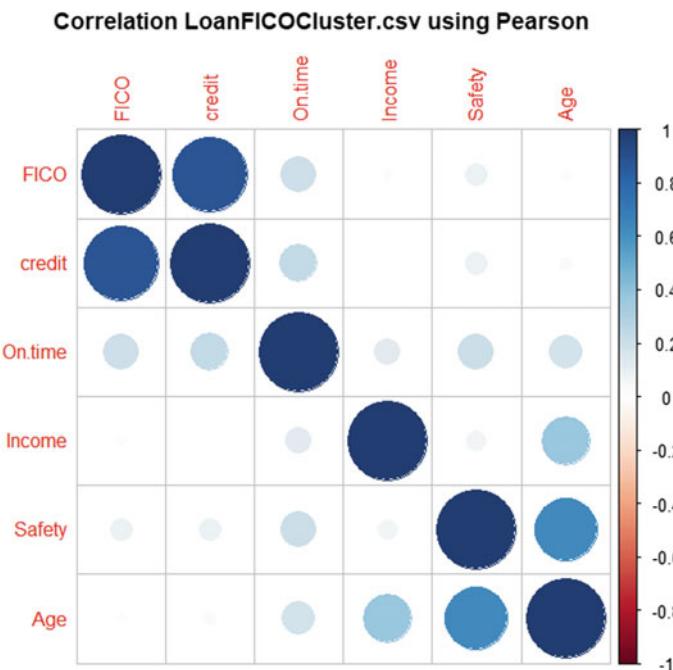


**Fig. 6.3** Distributions output from rattle

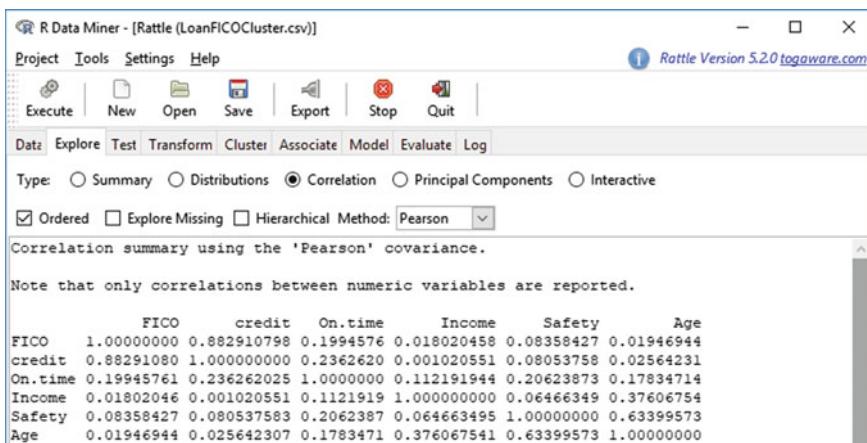
Rattle warns us that over five variables are cluttered, and only displays the first five. But we can see from FICO that the higher values are displayed in red, indicating that red must be cluster 2. Then we can see that Age, income, and safety are spread out a lot. Credit had three ratings—0 for poor, 0.3 for not that good, and 1.0 for OK. The OK rated observations are all red (cluster 2), the others black (cluster 1).

The Discriminant button provides a plot of the first two eigen vectors (see Fig. 6.10).

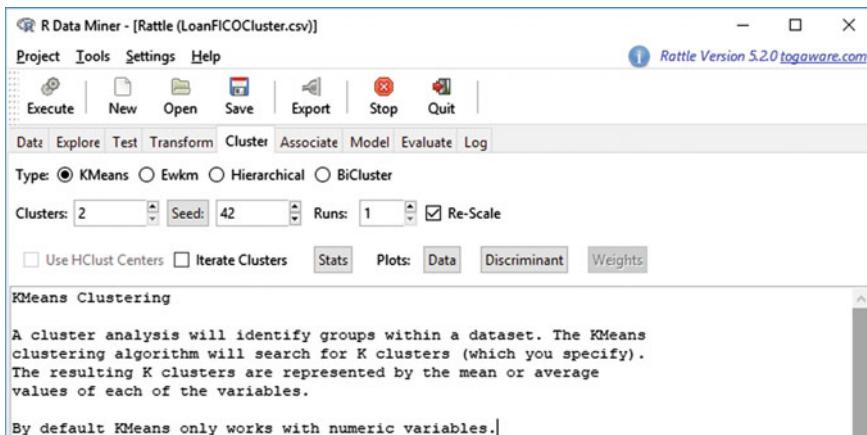
The Components here are assigned by Rattle based on discriminant analysis, essentially compressing the six variables into two. Circles are for cluster 1 and triangles for cluster 2, and here there is clear difference, indicated by the ovals.



**Fig. 6.4** Rattle correlation graphic



**Fig. 6.5** Rattle correlation table



**Fig. 6.6** Rattle clustering menu

The screenshot shows the R Data Miner - Rattle interface with the clustering output displayed in the main pane. The menu bar, toolbar, and sub-menu for "KMeans" are identical to Fig. 6.6. The output pane displays the following results:

```

Cluster sizes:
[1] "277 373"

Data means:
    Age   Income   credit      FICO   Safety   On.time
0.1791453 0.2580091 0.6661538 0.6819549 0.7849075 0.9000000

Cluster centers:
    Age   Income   credit      FICO   Safety   On.time
1 0.1750903 0.2566286 0.2166065 0.4716481 0.7589093 0.8231047
2 0.1821567 0.2590342 1.0000000 0.8381345 0.8042145 0.9571046

Within cluster sum of squares:
[1] 94.65254 69.92575

```

**Fig. 6.7** Rattle clustering output for K = 2

The screenshot shows the R Data Miner - Rattle interface. The menu bar includes Project, Tools, Settings, Help, and a version notice 'Rattle Version 5.2.0 togaware.com'. The toolbar has icons for Execute, New, Open, Save, Export, Stop, and Quit. Below the toolbar is a navigation bar with tabs: Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, and Log. The 'Cluster' tab is selected. A 'Type' section has a radio button for 'KMeans' (selected) and other options like Ewkm, Hierarchical, and BiCluster. Below it are fields for 'Clusters' (set to 2), 'Seed' (set to 42), and 'Runs' (set to 1). There is also a checkbox for 'Re-Scale'. Under the navigation bar, there are checkboxes for 'Use HClust Centers' and 'Iterate Clusters', and buttons for Stats, Plots, Data, Discriminant, and Weights. The main area displays R console output:

```

Cluster sizes:
[1] "226 424"

Data means:
      Age     Income    credit      FICO      Safety
 32.8984615 61235.8892308  0.6661538 706.8338462  0.7849075
  On.time
  0.9000000

Cluster centers:
      Age     Income    credit      FICO      Safety  On.time
1 36.93363 92550.93 0.6641593 709.3717 0.7811608 0.9380531
2 30.74764 44544.38 0.6672170 705.4811 0.7869046 0.8797170

Within cluster sum of squares:
[1] 135277684987 85947065074

```

**Fig. 6.8** Rattle clustering output for  $K = 2$  without re-scale

### Other R Clustering Algorithms

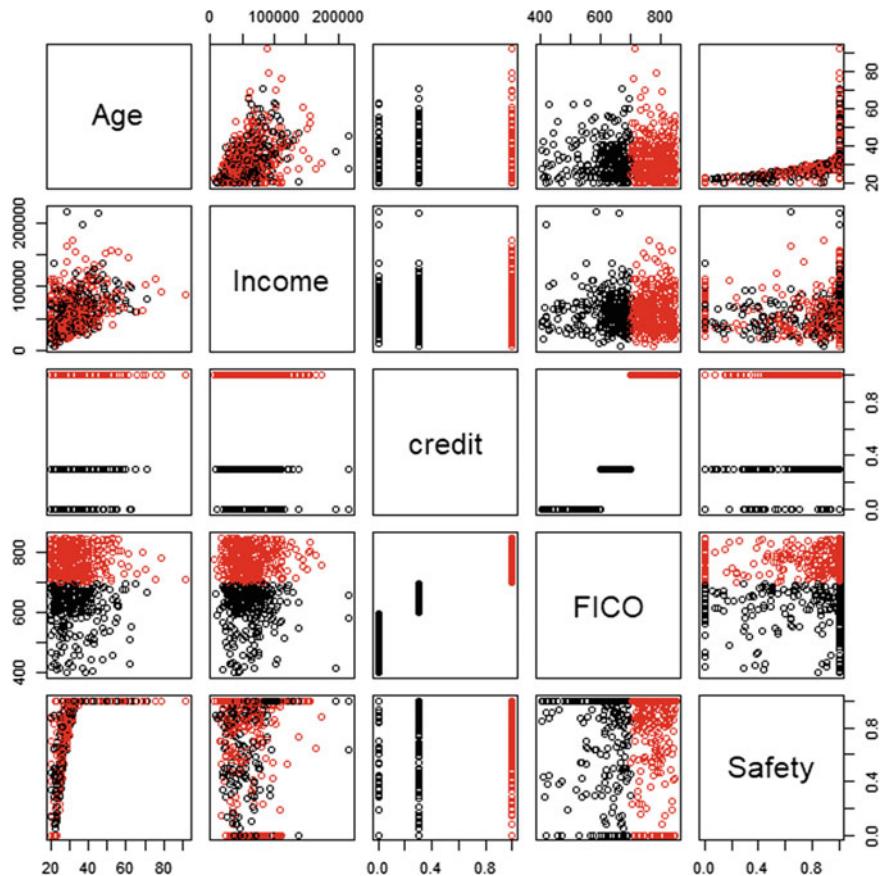
The data was rerun with the EWKM algorithm, yielding Fig. 6.11.

In this model, the second cluster had a slightly higher on-time performance. Cluster 1(285 observations) is younger with lower Safety and slightly lower income than Cluster 2 (365 observations). The discriminant plot here shows a lot of overlap (Fig. 6.12).

A hierarchical model was also run with Clusters set to 2, yielding Fig. 6.13 (from the Stats button).

Here cluster 1 has a slightly lower on-time performance, with the only real differences being cluster 1 is younger and has lower income. Figure 6.14 shows a great deal of overlap between the 353 observations in cluster 1 (you have to scroll down the Stats output to find) and the 297 in cluster 2.

We again emphasize that different models yield different clusters, and different settings within the same algorithm yield different clusters. In the spirit of data mining, the general practice is to run multiple models and analyze them in the context of the problem.



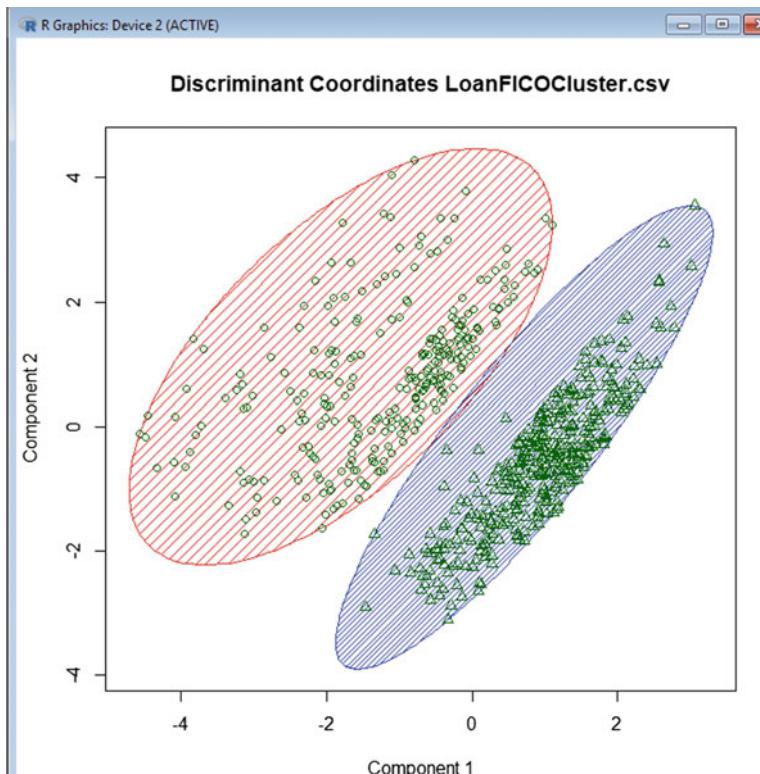
**Fig. 6.9** Data output from rattle for re-scaled clusters

A last tool we will look at is the Evaluation tab (Fig. 6.15). This enables the user to find the assigned cluster for any given model for either the original data (Full button), data split into training/validation/test sets (each with a button if the data was partitioned), or new cases. New cases have to be loaded through a file compatible with the input data, using the CSV File button.

We can compare multiple clustering across algorithms. Using data that is not re-scaled (so that the numbers make sense), we obtain the following tables of clusters. Table 6.3 displays K-means results.

Note that Table 6.3 sorted data on On-time outcome so that we might more easily see what happens across cluster sizes. Here we see that each cluster set is completely different. When we try EWKM, we obtain Table 6.4.

It is interesting that for K of 2, 3, and 5, exactly the same clusters were obtained as with K-means. Clusters for K of 4 are different. EWKM is K-means with



**Fig. 6.10** Discriminant plot for re-scaled clusters

different weighting. The defaults sometimes repeat K-means results. Table 6.5 gives results for Hierarchical clustering.

Here the clusters differ from both K-means and EWKM. But there is interesting performance across cluster sizes within the Hierarchical clusters. For K of 3, cluster 2 for K of 2 is split (new cluster 3 has slightly higher income, but otherwise not much difference). For K of 4, clusters 4 and 3 are identical to clusters 3 and 2 for K of 3, while cluster 1 for K of 3 has been split into clusters 1 and 2 for K of 4. For K of 5, cluster 5 is the same as cluster 4 for K of 4, cluster 3 is the same as cluster 3 of K of 4, cluster 2 is the same as cluster 2 for K of 4, and cluster 1 for K of 4 is split into clusters 1 and 4 for K of 5. We can clearly see the hierarchical splitting occur.

We may further visualize the difference in the clusters by displaying the values as a % of the means. The data shown in Table 6.5 are given in Table 6.6 in this form.

```

R Data Miner - [Rattle (LoanFICOCluster.csv)]
Project Tools Settings Help
Rattle Version 5.2.0 togaware.com
Execute New Open Save Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type:  KMeans  Ewkm  Hierarchical  BiCluster
Clusters: 2 Seed: 42 Runs: 1  Re-Scale
 Use HClust Centers  Iterate Clusters Stats Plots Data Discriminant Weights
EWKM: 2 clusters, 7 iterations, 0 restarts, 8 total iterations.
Cluster sizes:
[1] "285 365"
Data means:
    Age   Income   credit     FICO   Safety   On.time 
0.1791453 0.2580091 0.6661538 0.6819549 0.7849075 0.9000000
Cluster centers:
    Age   Income   credit     FICO   Safety   On.time 
1 0.06856725 0.2256132 0.6435088 0.6699792 0.5159631 0.8280702
2 0.26548706 0.2833045 0.6838356 0.6913058 0.9949052 0.9561644
Cluster weights:
    Age   Income   credit     FICO   Safety   On.time 
1 0.98 0.01 0 0 0 0 
2 0.00 0.00 0 0 1 0
Within cluster sum of squares:
[1] 0 0
< >
The KMeans model has been built. Time taken: 0.00 secs

```

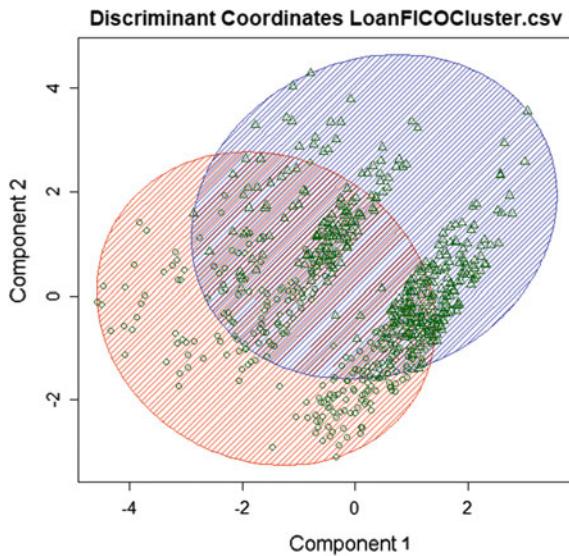
**Fig. 6.11** EWKM clustering from R

We can then display these data using for example in Excel using as radar chart as shown in Fig. 6.16. This enables us to see the relationships how the features of the clusters change without looking at the exact numbers. For  $K = 2$  only income shows a large difference. For larger values of  $K$ , income is still the biggest differentiator, but age starts to play a bigger role in differentiation. This is verified by correlation coefficients shown in Table 6.7.

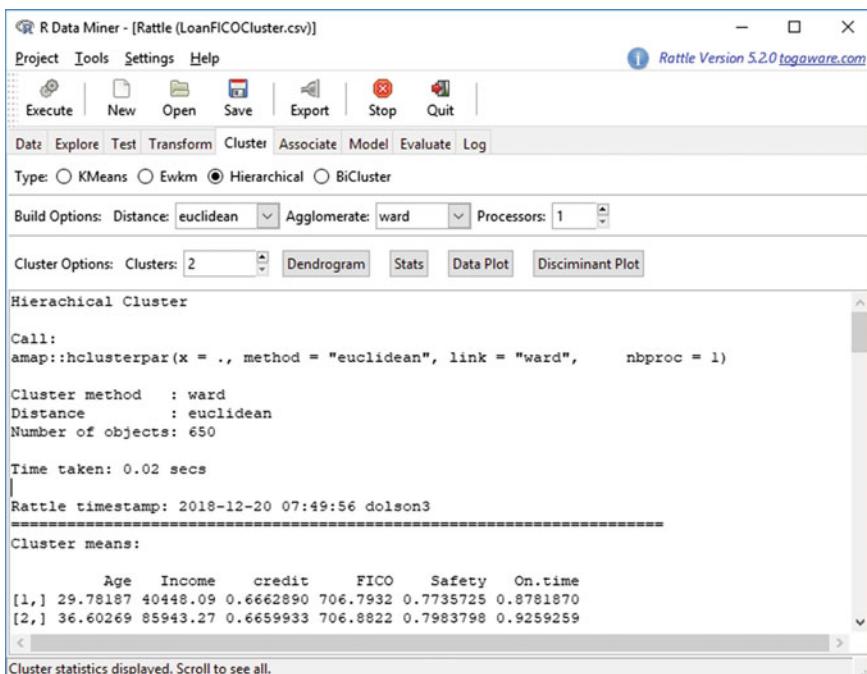
As observed in Table 6.7, income of the cluster centroids is highly correlated with age and on-time payments. The high income clusters contain the older customers which pay on time, while other clusters contain lower income customers that are younger and have worse on-time payments.

We repeated this set of runs omitting On-time. The results were completely different, indicating that adding data will yield different results. We also repeated the runs with Re-Scaling. Again, results were completely different. We infer that clustering is highly volatile.

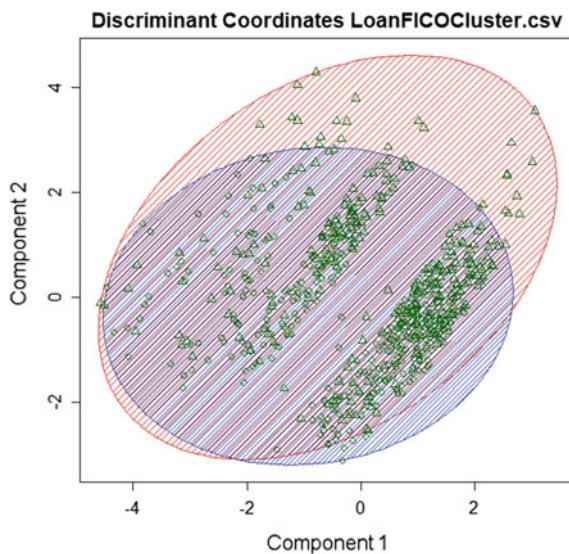
We now review two other open source data mining software cluster models.



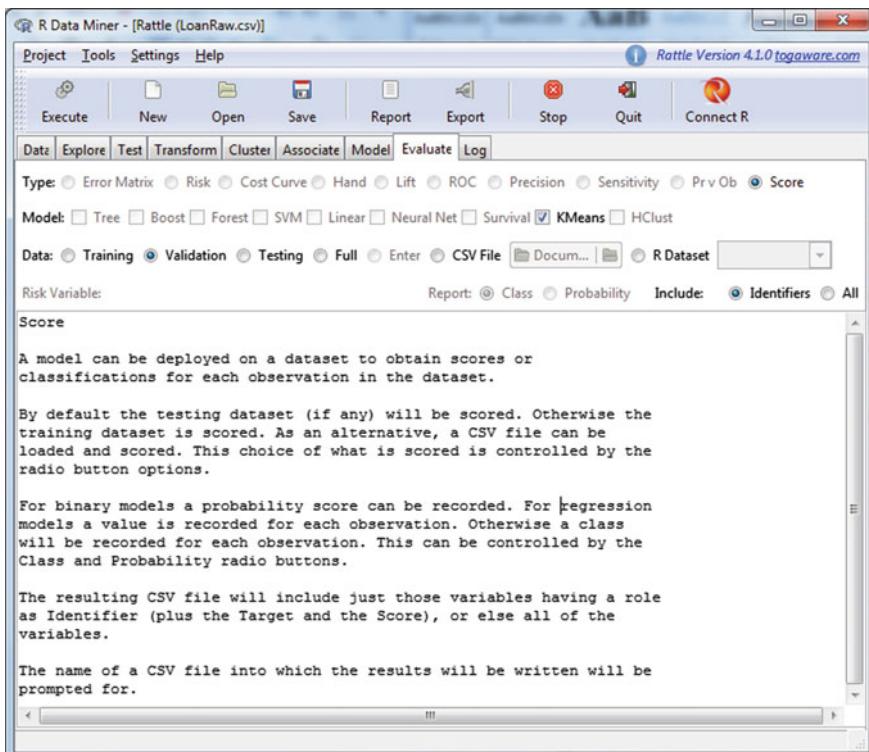
**Fig. 6.12** EWKM discriminant plot



**Fig. 6.13** Hierarchical stats output for K = 2



**Fig. 6.14** Hierarchical cluster discriminant plot



**Fig. 6.15** Rattle's evaluation tab

**Table 6.3** K-means clusters for loan data

Algorithm	K	Cluster	N	Age	Income	Credit	FICO	Safety	On-time
K-means	2	1	226	36.9	92,551	0.664	709	0.781	0.938
		2	424	30.7	44,544	0.667	705	0.787	0.880
	3	3	89	38.6	114,942	0.681	718	0.800	0.966
		1	274	35.2	69,466	0.657	703	0.812	0.905
		2	287	28.9	36,724	0.670	707	0.755	0.875
	4	3	17	40.4	157,751	0.700	712	0.877	1.000
		1	117	38.2	97,479	0.670	718	0.783	0.957
		4	259	34.2	64,557	0.656	700	0.809	0.900
		2	257	28.6	35,006	0.672	708	0.756	0.868
	5	3	16	40.2	159,560	0.725	714	0.869	1.000
		1	86	39.3	102,852	0.688	721	0.809	0.953
		4	188	32.2	50,515	0.668	708	0.809	0.910
		5	188	35.1	72,867	0.653	700	0.781	0.904
		2	172	27.4	30,286	0.663	705	0.743	0.849

**Table 6.4** EWKM clusters for loan data

Algorithm	K	Cluster	N	Age	Income	Credit	FICO	Safety	On-time
EWKM	2	1	226	36.9	92,551	0.664	709	0.781	0.938
		2	424	30.7	44,544	0.667	705	0.787	0.880
	3	3	90	39.0	114,688	0.684	718	0.803	0.967
		1	275	35.1	69,264	0.656	703	0.811	0.905
		2	285	28.9	36,610	0.671	707	0.754	0.874
	4	3	74	38.9	119,084	0.714	723	0.813	0.973
		4	200	35.8	76,388	0.641	701	0.780	0.910
		1	195	32.5	51,978	0.683	710	0.816	0.908
		2	181	27.7	30,817	0.657	704	0.746	0.851
	5	5	16	40.2	159,560	0.725	714	0.869	1.000
		3	86	39.3	102,852	0.688	721	0.809	0.953
		1	188	32.2	50,515	0.668	708	0.809	0.910
		4	188	35.1	72,867	0.653	700	0.781	0.904
		2	172	27.4	30,286	0.663	705	0.743	0.849

**Table 6.5** Hierarchical clusters for loan data

Algorithm	K	Cluster	N	Age	Income	Credit	FICO	Safety	On-time
Hierarchical	2	2	297	36.6	85,943	0.666	707	0.798	0.926
		1	353	29.8	40,448	0.666	707	0.774	0.878
	3	3	88	38.6	115,199	0.677	717	0.798	0.966
		2	209	35.8	73,625	0.661	703	0.798	0.909
		1	353	29.8	40,448	0.666	707	0.774	0.878
	4	4	88	38.6	115,198	0.677	717	0.798	0.966
		3	209	35.8	73,625	0.661	703	0.798	0.909
		1	228	31.5	47,645	0.661	706	0.792	0.890
		2	125	26.6	27,322	0.676	709	0.741	0.856
	5	5	88	38.9	115,199	0.677	717	0.798	0.966
		3	209	35.8	73,625	0.661	703	0.798	0.909
		1	103	32.8	54,614	0.654	706	0.825	0.903
		4	125	30.5	41,902	0.666	706	0.764	0.880
		2	125	26.6	27,322	0.676	709	0.741	0.856

**Table 6.6** K-means and normalized hierarchical clusters for loan data

Algorithm	K	Cluster	Cluster size	Age	Income	Credit	FICO	Safety	On-time
Means			1	32.8	61,235.9	0.66	706.8	0.78	0.9
Hierarchical	2	2	46%	111%	140%	101%	100%	102%	103%
		1	54%	91%	66%	101%	100%	99%	98%
	3	3	14%	117%	188%	102%	101%	102%	107%
		2	32%	109%	120%	100%	99%	102%	101%
		1	54%	91%	66%	101%	100%	99%	98%
	4	4	14%	117%	188%	102%	101%	102%	107%
		3	32%	109%	120%	100%	99%	102%	101%
		1	35%	96%	78%	100%	100%	101%	99%
		2	19%	81%	45%	102%	100%	94%	95%
	5	5	14%	118%	188%	102%	101%	102%	107%
		3	32%	109%	120%	100%	99%	102%	101%
		1	16%	100%	89%	99%	100%	105%	100%
		4	19%	93%	68%	101%	100%	97%	98%
		2	19%	81%	45%	102%	100%	94%	95%



**Fig. 6.16** Hierarchical cluster radar plots (displayed using excel) for the different number of clusters as shown in Table 6.6

**Table 6.7** Correlation coefficient for the 5 clusters displayed in Table 6.6

Cluster size		Age	Income	Credit	FICO	Safety	On-time
Cluster size	1						
Age	0.0	1					
Income	-0.1	1.0	1				
Credit	-0.3	-0.1	0.2	1			
FICO	-0.7	0.4	0.6	0.7	1		
Safety	0.0	0.7	0.6	-0.7	0.0	1	
On-time	-0.2	1.0	1.0	0.1	0.6	0.6	1

## KNIME

KNIME is a workflow process system. First load data by selecting input **IO** (for input/output), select **File Reader** and drag to the workflow, selecting the same LoanClusterStd.csv file used with R. Click on the File Reader icon, and select **Configure**, followed by **Execute and Open Views**. This yields Fig. 6.17.

Next select **Analytics**, **Mining**, and **Clustering**, and drag the **K-Means** icon into the workflow. Drag from the triangle on the right side of the File Reader icon to the input triangle on the K-means icon. Right-click on the K-Means icon and **Configure**. **Execute and Open Views** yields Fig. 6.18.

Note that the default is to exclude the output variable. This is fine, but you can add it back if you choose to include On-Time (as we did with R and as we do here).

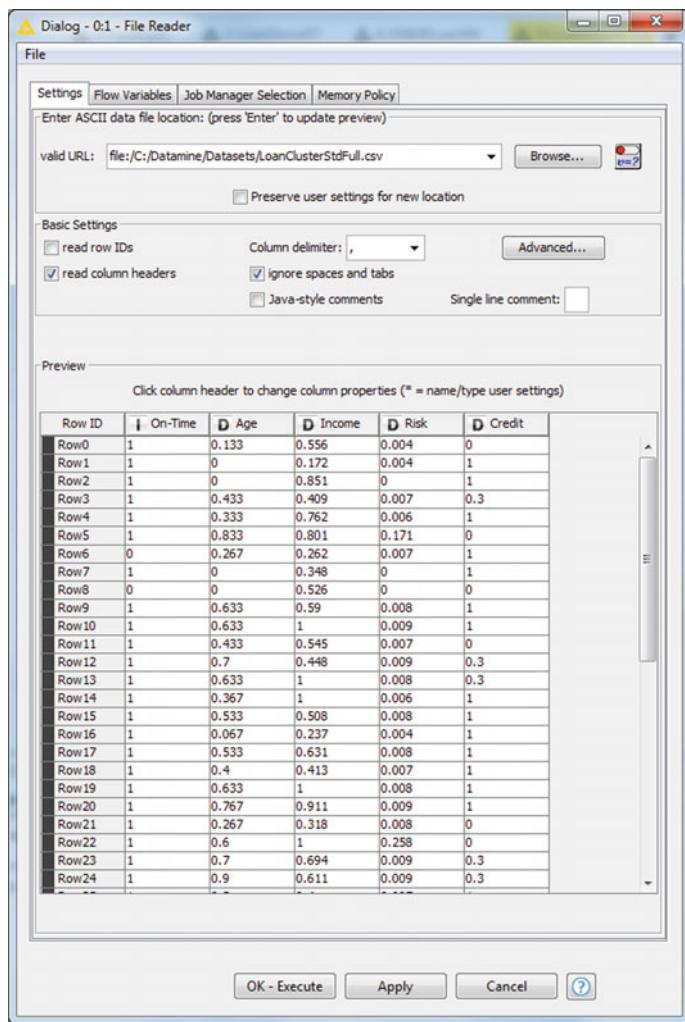
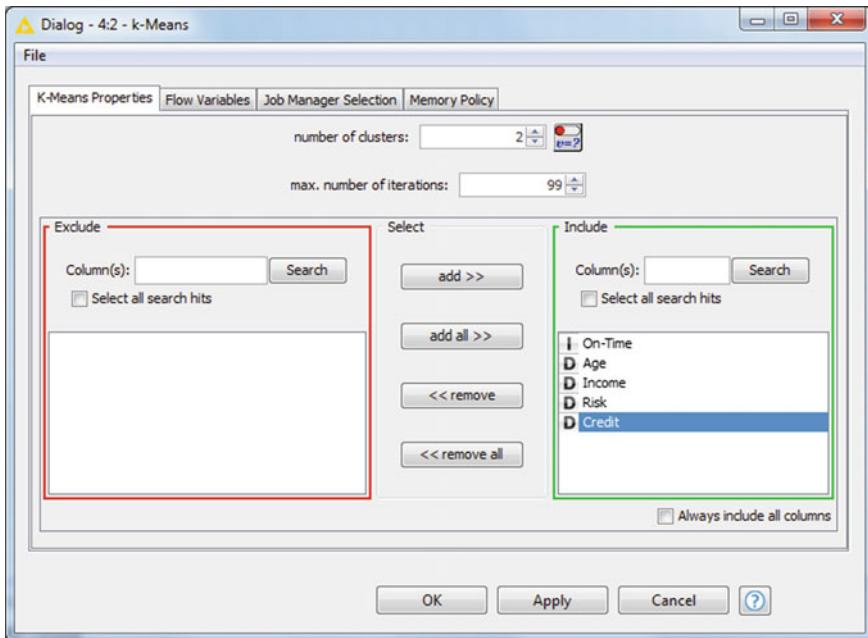


Fig. 6.17 KNIME file reader result

R also defaults to exclude output variables. Clustering again treats all included variables the same, seeking clusters of similar observation vectors.

Next we need to apply a K-Means model. Drag a **Cluster Assigner** icon onto the workflow, and connect the triangle icon on the left of the Cluster Assigner to the triangle on the File Reader, and the Cluster Assigner's box icon on its left to the box icon on the right of the K-Means icon. Click on the Cluster Assigner, **Configure**, and **Execute and Open Views**. Now go to **Views** and drag an **Interactive Table** to the workflow. Connect the output triangle from the Cluster Assigner to the Interactive Table's triangle. Right click on the Interactive Table, **Configure**, and **Execute and Open Views**. This yields a cluster assignment shown in Table 6.8.



**Fig. 6.18** KNIME K-means control

**Table 6.8** KNIME K-means clustering output

Cluster	On-time	Age	Income	Risk	Credit
C1 (277 cases)	0.822	0.386	0.588	0.786	0.217
C2 (373 cases)	0.957	0.397	0.592	0.831	1.0

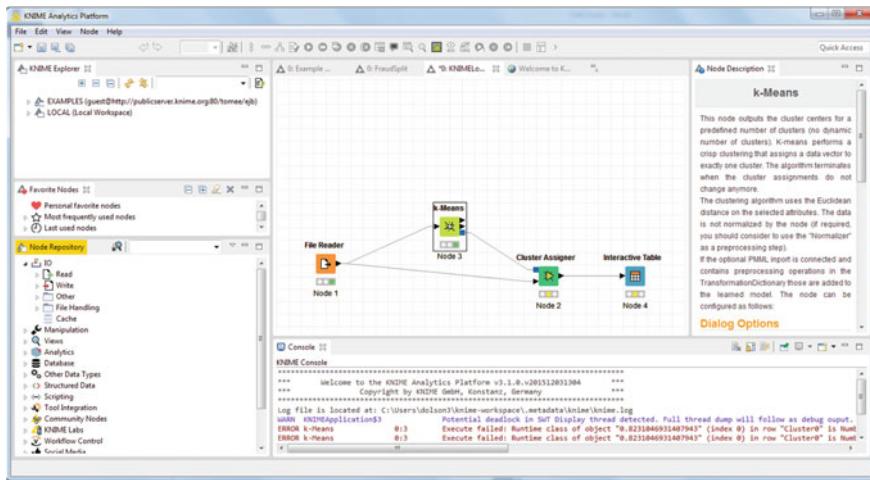
This is very close to the R clustering output for k-means.

The workflow and icon palates are shown in Fig. 6.19. It is trivial to change datasets with the KNIME system—simply redo the File Reader icon, select the new file, and go through the sequence of Configure and Execute for each icon. That takes less time than it takes to write this sentence.

## WEKA

You can download WEKA off the Web at <http://www.cs.waikato.ac.nz/ml/weka/>. The download comes with documentation. Open WEKA, and select **Explorer**, obtaining Fig. 6.20.

Select **Explorer**, yielding Fig. 6.21.



**Fig. 6.19** KNIME workflow for K-means

**Fig. 6.20** WEKA opening screen

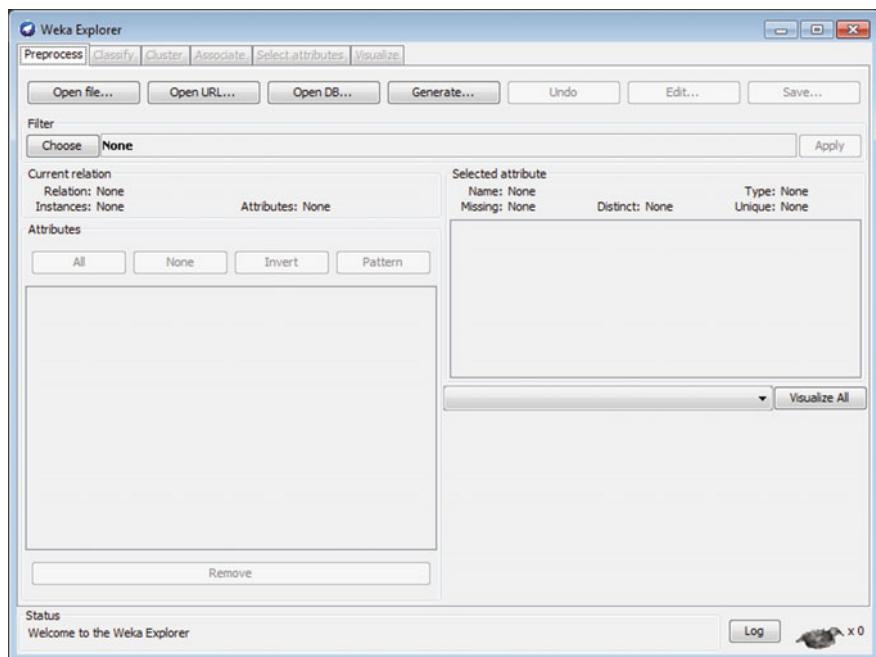


Clustering with WEKA starts with opening a file, using the same LoanClusterStd.csv file we used for R and KNIME. The WEKA screen opens as shown in Fig. 6.22, displaying a histogram of any variable selected (in this case On-Time was first, with a 0–1 distribution of 65 zeros and 585 ones).

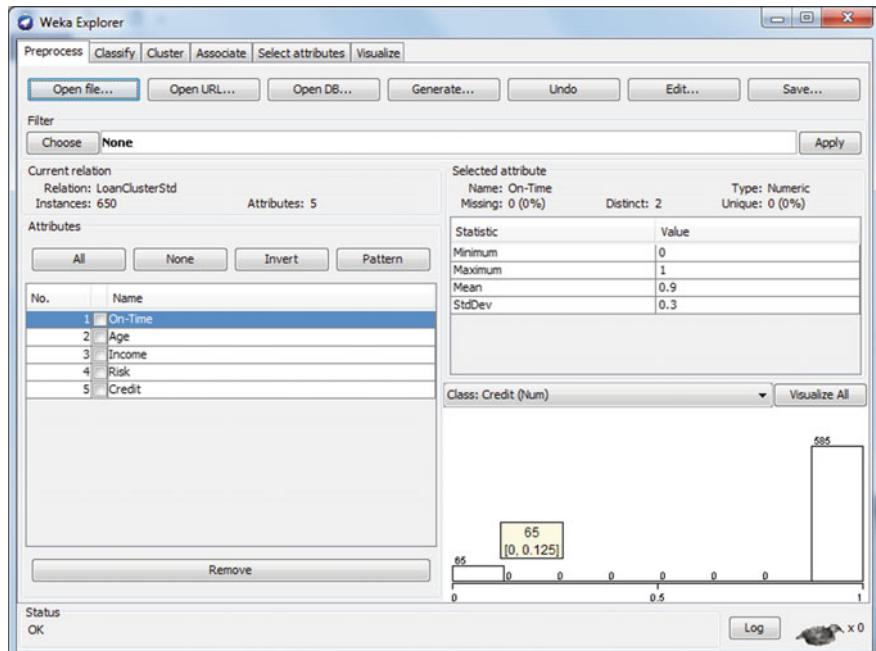
Use WEKA and select **Open file...** and pick file from your hard drive. In Fig. 6.23 we pick **LoanRaw.csv**

You can play around with Visualize to see how the data acts. We select Assets, Debts and Want and **Remove** them. We can then select the **Cluster** tab, as shown in Fig. 6.24. The menu of algorithms for clustering for continuous input data includes:

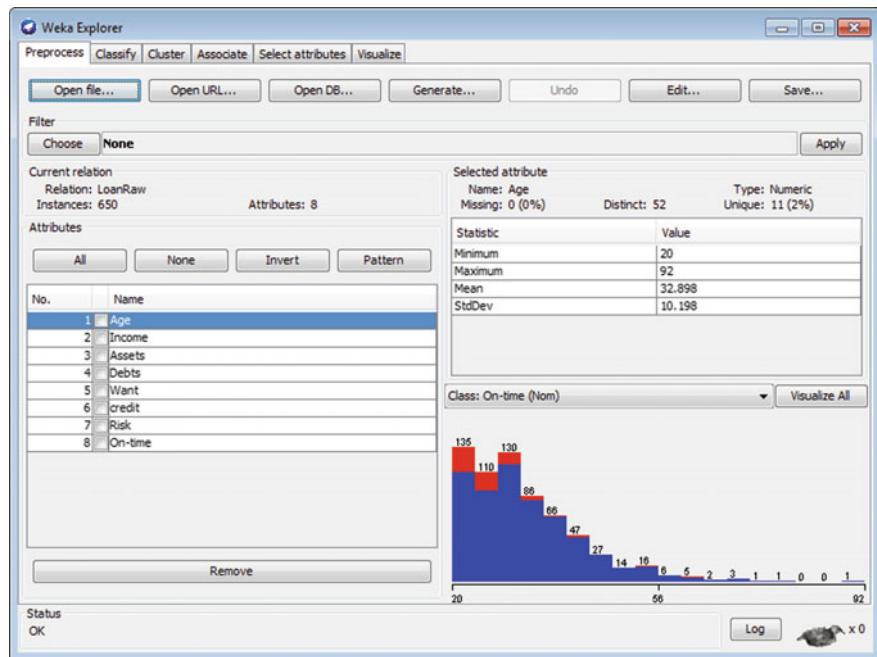
- Cobweb
- DBSCAN
- EM



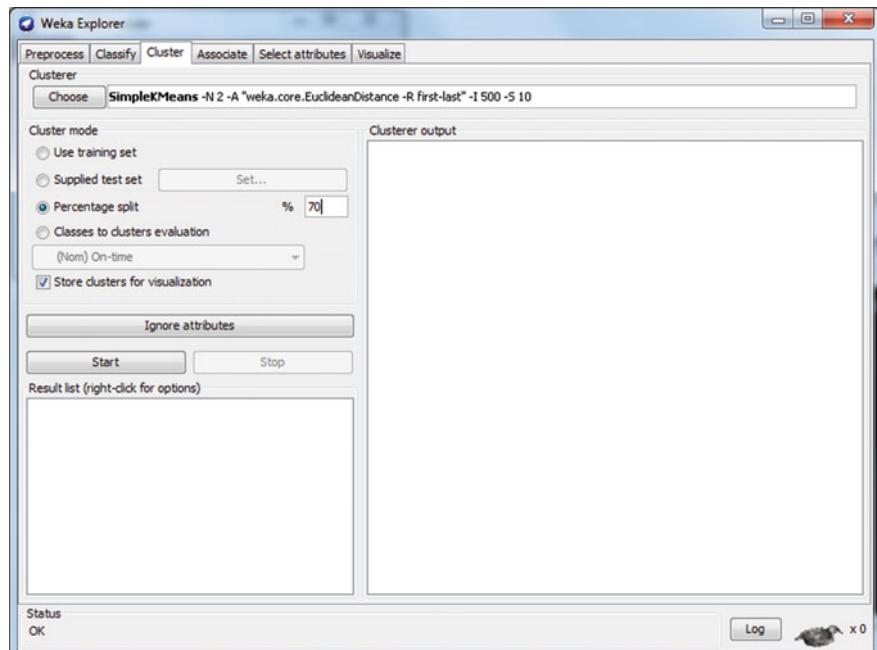
**Fig. 6.21** WEKA explorer screen



**Fig. 6.22** WEKA file opening screen



**Fig. 6.23** WEKA screen for LoanRaw.csv



**Fig. 6.24** WEKA cluster screen

- Farthest First
- Filtered Clusterer
- Hierarchical Clusterer
- MakeDensityBasedClusterer
- OPTICS
- SimpleKMeans.

Most of these are more involved than we need. SimpleKMeans applies the same algorithm as K-Means used by R.

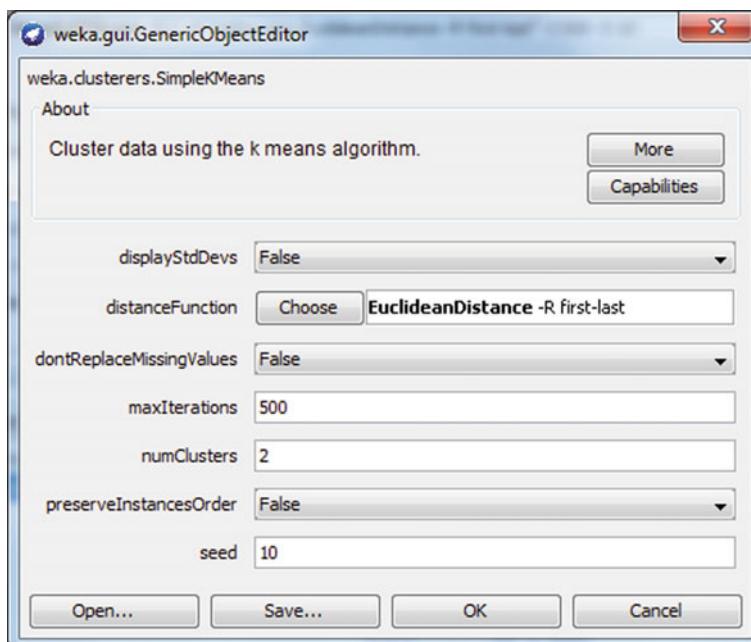
Select **Choose** and you can change parameters, to include the number of clusters (“N”). Figure 6.25 shows the menu.

The default is to use EuclideanDistance (minimize sum of squared distance). Other options are Chebycheff (minimize maximum distance) or Manhattan (minimize sum of linear distance). Manhattan is equivalent to the Farthest First clustering algorithm. The Euclidean Distance algorithm yields the model shown in Fig. 6.26.

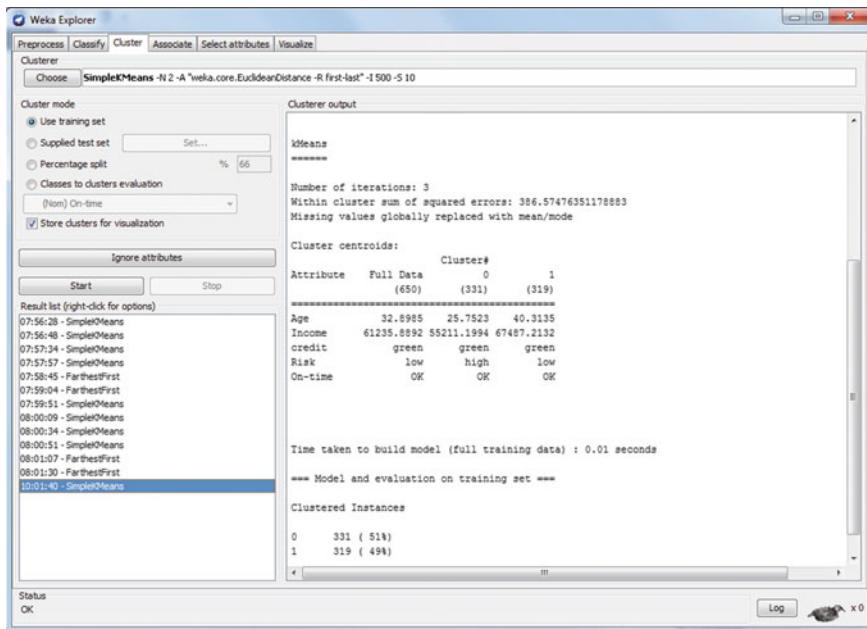
We can run the Manhattan model and the Farthest First clustering algorithm, obtaining different results as compared in Table 6.9.

Table 6.9 demonstrates that different software yields different clusters (WEKA considered the categorical variables, R did not), and the metric can yield different results as well (here the Euclidean and Manhattan results were similar, but slightly different ages and incomes) (Fig. 6.27 shows the K-means screenshot in WEKA).

Click on the **Cluster** tab and then select **Choose**. There are eleven optional clustering models in WEKA for this type of data. Simple K-Means allows use of



**Fig. 6.25** K means cluster menu in WEKA



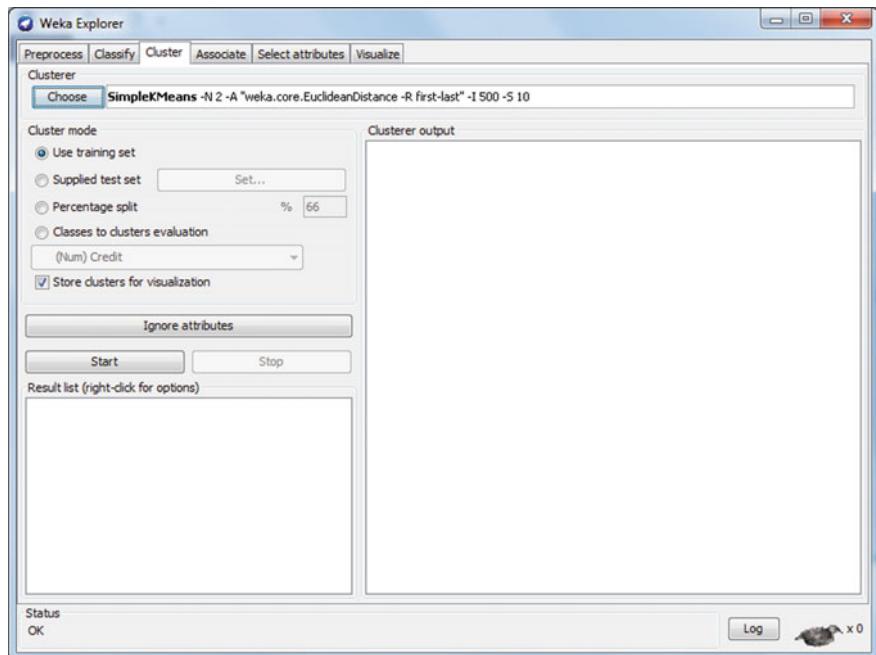
**Fig. 6.26** WEKA clusters with K = 2 Euclidean

**Table 6.9** Comparison of Clusters with K = 2

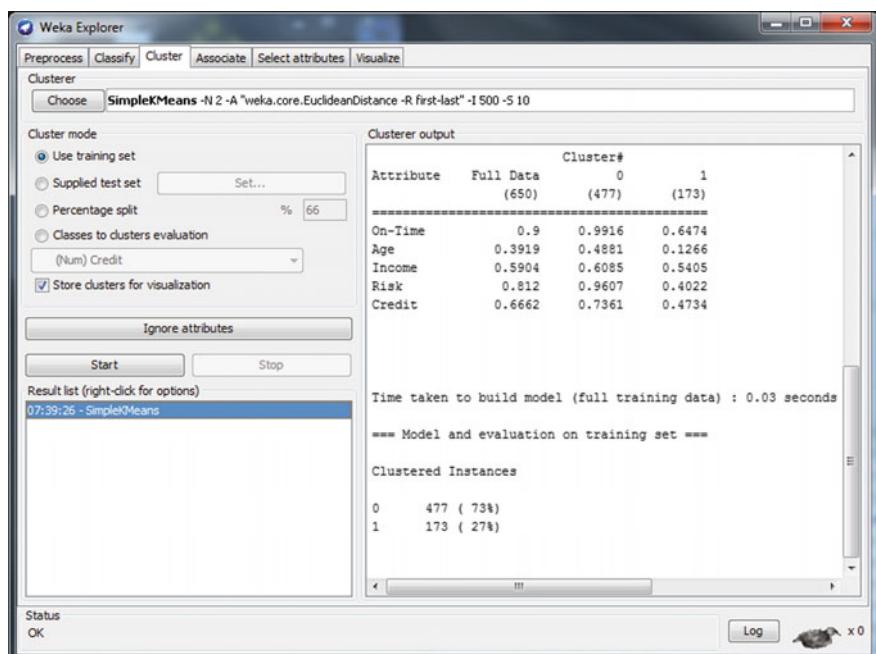
Metric		Age	Income	Credit	Risk	On-time	N
R Euclidean	C1	30.748	44,544				424
	C2	36.934	92,551				226
WEKA Euclidean	C1	25.461	56,670	Green	High	OK	331
	C2	40.313	67,487	Green	Low	OK	319
WEKA Manhattan	C1	26	50,456	Green	High	OK	331
	C2	38	63,124	Green	Low	OK	319
WEKA FarthestFirst	C1	32	42,681	Amber	Medium	OK	536
	C2	62	64,188	Red	Low	Problem	114

Manhattan (minimizing sum of absolute distances), Euclidean (minimizing sum of squared distances), Chebycheff (minimizing maximum distance), and other options. The default is Euclidean. Other algorithms available include Farthest First, which should match the Chebycheff metric for K-Means.

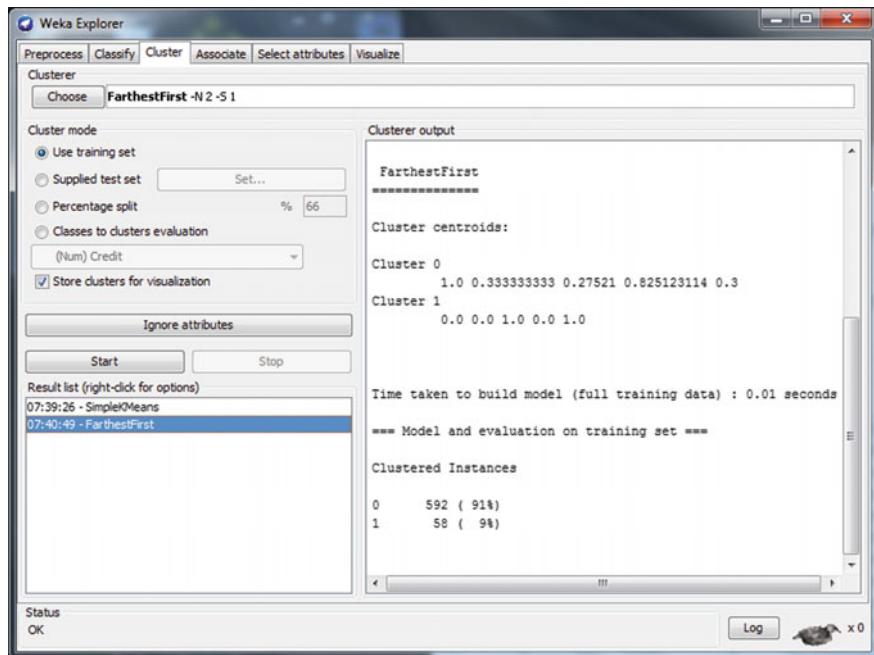
You have options with WEKA about evaluative schemes. The default is **Use training set** to test the model. It is better form to build the model with part of the data and test it on other data. This can be done by a **Supplied test set** (which honestly doesn't usually work with WEKA), a **Percentage Split**, or other options. Using the training data we select **Start** which yields the model given in Fig. 6.28.



**Fig. 6.27** WEKA K-means screen



**Fig. 6.28** WEKA K = 2 clustering output



**Fig. 6.29** WEKA farthest first cluster

Here the two clusters vary most on all variables. The 477 in cluster 1 have higher on-time payment, are older, have more income, better risk ratings, and higher credit compared to the 173 in cluster 2. Using the Farthest First algorithm on WEKA yielded a different model, shown in Fig. 6.29.

This is a much more distinctly different set of clusters, with Cluster 0 having a much better on-time record despite a much lower credit rating and lower income. Cluster 0 had an older average. There were 592 instances out of 650 in cluster 0, and 58 in Cluster 1.

## Summary

Cluster analysis is a very attractive initial data examination tool. Once different clusters are identified, other methods are often used to discover rules and patterns. Outcome variables can be included, but can also be dropped. They are treated just like any other variable in clustering, but it is convenient to include them to make interpretation of the differences in clusters easier. The intent of clustering, however, is to identify patterns rather than to predict.

Sometimes the median has been used rather than the mean as the basis for cluster centers, as the first real example did. That decision was made expecting the median

to be more stable than the mean. This is because outlier observations (radically different from the norm) do not affect the median, but do influence the mean. It is very simple to implement the median rather than the mean in Excel (although not in packaged data mining software). In Excel all one does is use the formula “=MEDIAN(range)” rather than “AVERAGE(range)”.

Some problems may not have an obvious set of clusters. There are a number of options for determining the number of clusters. Agglomeration is an approach where you start with the maximum number of clusters, and then merge clusters iteratively until there is only one cluster left. Then the cluster value that fits best (by whatever metric is selected, and based upon the need for correct prediction—fewer clusters are better, along with the need for discrimination of difference—more clusters are better) is chosen. Commercial tools have a number of different parameters and methods. Some in fact use probability density rather than distance measures, which tends to work better when clusters overlap.

## References

- Astudillo CA, Oommen BJ (2011) Imposing tree-based topologies onto self organizing maps. *Inf Sci* 181:2798–3815
- Johnson RA, Wichern DW (1998) Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River, NJ
- Kohonen T (1997) Self-organizing maps. Springer, Berlin
- Sarlin p (2013) Self-organizing time map: an abstraction of temporal multivariate patterns. *Neurocomputing* 99:496–508

# Chapter 7

## Link Analysis



Link analysis considers the relationship between entities in a network. They are interesting in many contexts, to include social network analysis (Knoke and Yang 2008) which has been used to measure social relationships, to include social media and collaboration networks. People (or customers) can be represented as nodes and the relationships between them can be links in a graph. In biological science they have been applied to analyze protein interactions. They also have been applied to law enforcement and terrorism. In business, they are of interest to marketing, especially related to product recommendation analysis. Amazon is famous for its recommendation engine, supported by link analysis of customer selections. Clickstream analysis provides Web business sites with the ability to predict where customers are going, potentially allowing systems to interact with these customers to increase the probability of their purchasing the vendor's goods.

### Link Analysis Terms

Link analysis in general provides tools for analysis of large scale graphs. The effectiveness of such tools is bounded by graph dimensionality, yet large scale network analysis has been applied to Twitter, human neural networks, and other web graphs. Graph analysis is challenging in terms of efficient computing, making scalability a problem.

There are a number of important concepts in graph representation of networks. These include **degree**, which is the number of connections a node (**vertex**) has to other nodes. These connections are also referred to as **edges**. Examples of networks commonly found in businesses of various types are shown in Table 7.1, taken from: <https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/>.

A current example, where Link Analysis might be useful in gaining understanding is Brexit, the UK leaving the EU. A network might demonstrate

**Table 7.1** Example business networks

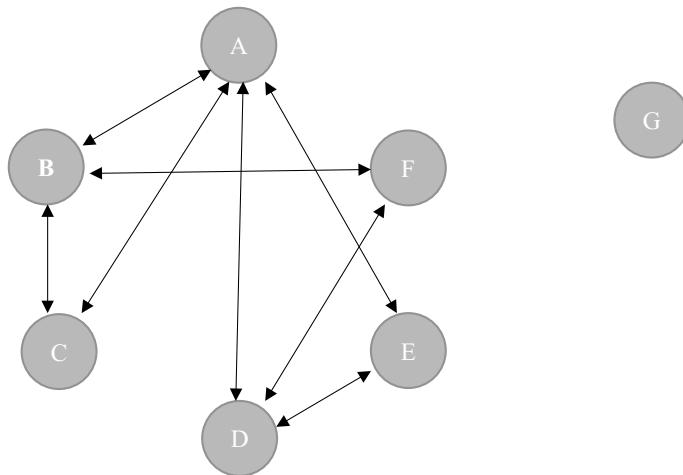
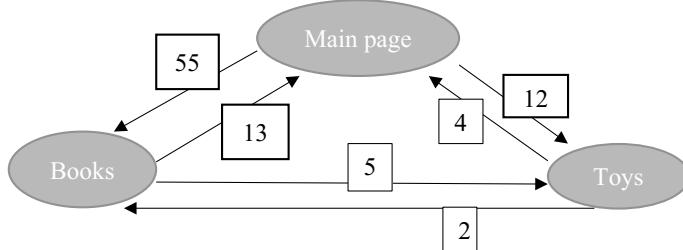
Network	Vertices	Vertex attributes	Edges	Edge attributes
Airlines	Airports	Terminal capacity, Type of plane able to land, city population	Airplanes, routes	# Passenger, plane type, custom and immigration regulations
Banking network	Account holders	Name, demographics, banking regulations?	Transactions	Type, amount, time, location device
Social network	Users	User, connections	Interactions	Medium, type of content, topic
Physician network	Doctors	Demographics, speciality	Patients	Demographics, medical history, insurance
Supply chain network	Warehouses	Location, access (roads, rail, ...), custom regulations	Trucks, trains, planes	Capacity, utilization, speed, cost

implications in the aspects described above. British Airlines flying to the EU or within the EU may lose such rights due to different regulations. Banking is expected to be most severely affected. While not participating in the EURO currency, London has been serving as a financial center for EU transactions. After Brexit, banks in the UK probably will lose rights to continue selling their products and services throughout the European Union. With a change in immigration rules, real changes for the affected families will occur. Changes are expected even on Social Networks. The EU and UK may have different privacy rules, which may affect, how and where information may be stored. The Supply Chain Network is the most obvious one to worry about. Trucks waiting in Customs at the English Channel will affect customers down the supply chain, while in Northern Ireland the question is how to keep the border open when separating these regulatory markets. A Network Analysis really would be of great benefit to all involved.

Graphs can represent a number of situations. A graph of social acquaintances would probably involve undirected links, where relationships ran both ways. Consider a network of seven people as in Table 7.2, one of whom (George) doesn't deal with any of the others. Figure 7.1 displays this network graphically in bidirectional form.

**Table 7.2** Social network

	Albert	Betty	Charles	Daisy	Edward	Fern	George
Albert	0	1	1	1	1	0	0
Betty	1	0	1	0	0	1	0
Charles	1	1	0	0	0	0	0
Daisy	1	0	0	0	1	1	0
Edward	1	0	0	1	0	0	0
Fern	0	1	0	1	0	0	0
George	0	0	0	0	0	0	0

**Fig. 7.1** Social graph**Fig. 7.2** Valued graph

You also can place values on graphs. For instance, on e-business web sites, the nodes could be pages and it would be important to know the volume of traffic on each link. This can be used for purposes of pricing, as well as many kinds of analysis. Figure 7.2 shows a valued graph. Note that the value in terms of traffic can depend on the direction of the link.

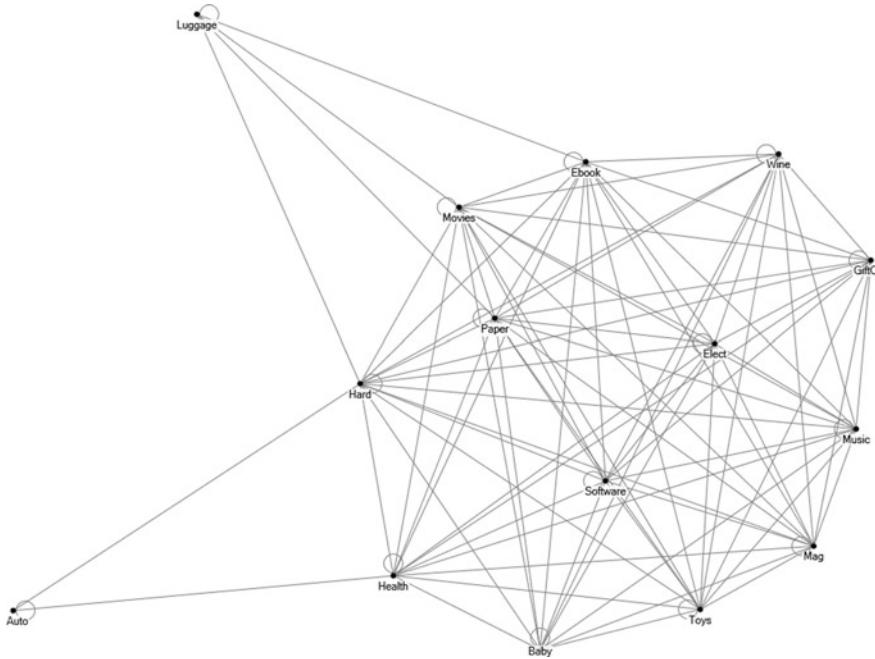
Figure 7.1 is vastly simplified for demonstration purposes, but shows how the main flow of traffic on the site is to books, but the site allows movement from each node to all others. Thus Fig. 7.1 demonstrates both direction as well as value.

To demonstrate additional measures for graphs, we generated a set of pseudo-Amazon data for 15 product categories in the Market Basket Analysis chapter. That data here is aggregated into the counts (degrees) for each pair of products (nodes) as in Table 7.3.

Figure 7.3 gives a network graph of the data in Table 7.3, generated by NodeXL software that will be described later in this chapter. (Note that this table is symmetric.)

**Table 7.3** Pseudo-Amazon node degrees

	Auto	Baby	Ebook	Hard	Paper	Music	Elect	Health	GiftC	Luggage	Mag	Movies	Software	Toys	Wine
Auto	15	0	0	1	0	0	5	0	0	0	0	0	0	0	0
Baby	0	52	10	3	7	1	2	3	0	0	2	1	3	49	4
Ebook	0	10	619	475	483	44	3	34	29	1	11	10	17	27	8
Hard	1	3	475	493	426	19	10	25	15	1	6	11	19	16	8
Paper	0	7	483	426	497	10	4	24	14	1	7	5	16	21	7
Music	0	1	44	19	10	118	19	11	11	0	4	20	12	11	5
Elect	0	2	3	10	4	19	40	4	3	0	1	14	19	3	1
Health	5	3	34	25	24	11	4	65	2	0	3	14	4	8	0
GiftC	0	0	29	15	14	11	3	2	57	0	4	21	2	4	5
Luggage	0	0	1	1	0	0	0	0	4	0	3	0	0	0	0
Mag	0	2	11	6	7	4	1	3	4	0	27	10	2	3	2
Movies	0	1	10	11	5	20	14	14	21	3	10	132	11	7	11
Software	0	3	17	19	16	12	19	4	2	0	2	11	111	5	10
Toys	0	49	27	16	21	11	3	8	4	0	3	7	5	112	5
Wine	0	4	8	8	7	5	1	0	5	0	2	11	10	5	57



**Fig. 7.3** Network graph of pseudo-Amazon data

Here there are fifteen nodes. The degree of connectivity for automotive products (Auto) is to two other products, hard-back books (Hard) and health products (Health). Degree can be directional. For instance, here we don't have a direction inherently present, so the graph direction could go from Auto to Health or Health to Auto. However if person A trusted person B, but person B didn't trust person A, trust may have a direction. The bidirectional degree for Auto is thus 4. Baby products have connections to 11 other products, bidirectional, yielding a degree of 22.

**Density** is the number of connections divided by the number of possible connections (disregarding loops that would create duplicates). The density for Auto in Table 7.3 is 4/28, or 0.143. Conversely, the density for Baby is 22/28, or 0.786. The maximum density is 1.000 for hard-back books (Hard), the only product whose purchasers (493 out of 1000) purchased every other product category.

**Connectedness** refers to the degree of ability of the overall graph to reach all nodes. The Krackhardt connectedness score sums each node's unidirectional degree divided by {the number of nodes ( $n$ ) times ( $n - 1$ )}. This measure has a maximum of 1.0, indicating all nodes connect to all other nodes, and a minimum of 0 where all nos are completely isolated (there are no node connections). In the pseudo-Amazon data given in Table 7.3, Table 7.4 shows the degree (one-way in this case) for each node.

**Table 7.4** Unidirectional degrees for pseudo-Amazon data

	Degree
Auto	4
Baby	13
Ebook	15
Hard	16
Paper	15
Music	14
Elect	14
Health	14
GiftC	13
Luggage	6
Mag	14
Movies	15
Software	14
Toys	14
Wine	13

These fifteen degrees add to 194. The Krackhardt connectedness score would thus be:

$$194/(15 \times 14) = 0.92$$

This is a high degree of connectedness.

We can calculate density for each of the seven participants given in Table 7.2 as shown in Table 7.5.

The connectedness of this graph is equal to the sum of connections ( $16 -$  the sum of the numerators in the second column in Table 7.5) divided by  $42 (n \times n - 1)$  or 0.381.

**Geodesic distance** is the length of the shortest path between two nodes. This would required calculating all paths between the two nodes in question, finding the lowest value of each path, and dividing this lowest value by the length of the path, then choosing the highest value among these outcomes. Table 7.6 gives geodesic distances (path lengths) for the network in Fig. 7.1.

**Table 7.5** Densities for social network (unidirectional)

	Degree	Connections	Density
Albert	4	4/6	0.067
Betty	3	3/6	0.500
Charles	2	2/6	0.333
Daisy	3	3/6	0.500
Edward	2	2/6	0.333
Fern	2	2/6	0.333
George	0	0/6	0

**Table 7.6** Geodesic distances

	A	B	C	D	E	F	G
A	0	1	1	1	1	2	$\infty$
B	1	0	1	2	2	1	$\infty$
C	1	1	0	2	2	2	$\infty$
D	1	2	2	0	1	1	$\infty$
E	1	2	2	1	0	2	$\infty$
F	2	1	2	1	2	0	$\infty$
G	$\infty$						

**Table 7.7** Betweenness calculations

Node A	Node B	Node C	Node D	Node E	Node F
B–C no	A–C no	A–B no	A–B no	A–B no	A–B no
B–D yes	A–D no	A–D no	A–C no	A–C no	A–C no
B–E yes	A–E no	A–E no	A–E no	A–D no	A–D no
B–F no	A–F yes	A–F no	A–F yes	A–F no	A–E no
C–D yes	C–D no	B–D no	B–C no	B–C no	B–C no
C–E yes	C–E no	B–E no	B–E no	B–D no	B–D yes
C–F no	C–F yes	B–F no	B–F no	B–F no	B–E no
D–E no	D–E no	D–E no	C–E no	C–D no	C–D no
D–F no	D–F no	D–F no	C–F no	C–F no	C–E no
E–F no	E–F no	E–F no	E–F yes	D–F no	D–E no

**Betweenness centrality** is a measure of the degree to which a particular node lies on the shortest paths between other nodes in the graph. Thus it can reflect how other nodes control or mediate relations between pairs of other nodes that are not directly connected. The formula for Betweenness Centrality  $C_b$  is the sum of shortest paths (geodesics) from this node to all other nodes divided by the number of distinct geodesics in the graph system. Considering the linked set of nodes A through F in Fig. 7.1, Table 7.7 shows this calculation.

Thus Node A lies on 4 geodesics out of 15, for a betweenness centrality of 0.267, B 2/15 = 0.133, C 0/15 = 0.0, D 2/15 = 0.133, E 0/15 = 0/0, and F 1/15 = 0.067. G connects with nothing, and thus is not a member of any geodesic, so has a betweenness centrality of 0.0. If direction were considered both numerator and denominator would simply double, yielding the same betweenness calculations since Fig. 7.1 was bidirectional. If it were unidirectional, it would affect the number of “yes” values in Table 7.7.

**Closeness** reflects how near a node is to the other nodes in a network. This can indicate how quickly nodes can interact, with longer closeness implying the need to go through more intermediaries. **Closeness centrality** is the inverse of the sum of geodesic distances between a node and the other nodes. If you add the non-infinite

**Table 7.8** Betweenness centrality for Fig. 7.1

Node	V-1	Sum of distances	Closeness centrality
A	6	6	1.000
B	6	7	0.857
C	6	8	0.750
D	6	7	0.857
E	6	8	0.750
F	6	8	0.750
G	6	$\infty$	0

values in Table 7.6, you get the sum of distances between all nodes (in this case 22). The closeness centrality measure is:

$$\text{Closeness Centrality} = \{\text{Number of nodes} - 1\} / \{\text{sum of distances}\}$$

In the case of Fig. 7.1, these values are as shown in Table 7.8.

This measure decreases if the number of nodes reachable from the node in question decreases, or if the distances between nodes increases. Here A has the most “closeness”, while G has the least.

## Basic Network Graphics with NodeXL

NodeXL (Network Overview, Discovery and Exploration for Excel) is a template for Microsoft Excel that provides the ability to explore network graphs. NodeXL Excel Template opens from the Windows Start Menu, creating a new workbook customized for network graph data. Figure 7.4 displays the basic layout, with the NodeXL ribbon tab at the top, the Edges worksheet on the left, and a graph pane to display networks on the right. At the bottom are tabs for the five worksheets (Edges, Vertices, Groups, Group Vertices, and a hidden fifth worksheet—Overall Metrics). Note that most metrics require NodeXL Pro version or higher.

These five worksheets (taken from NodeXL web documentation) are described in Table 7.9.

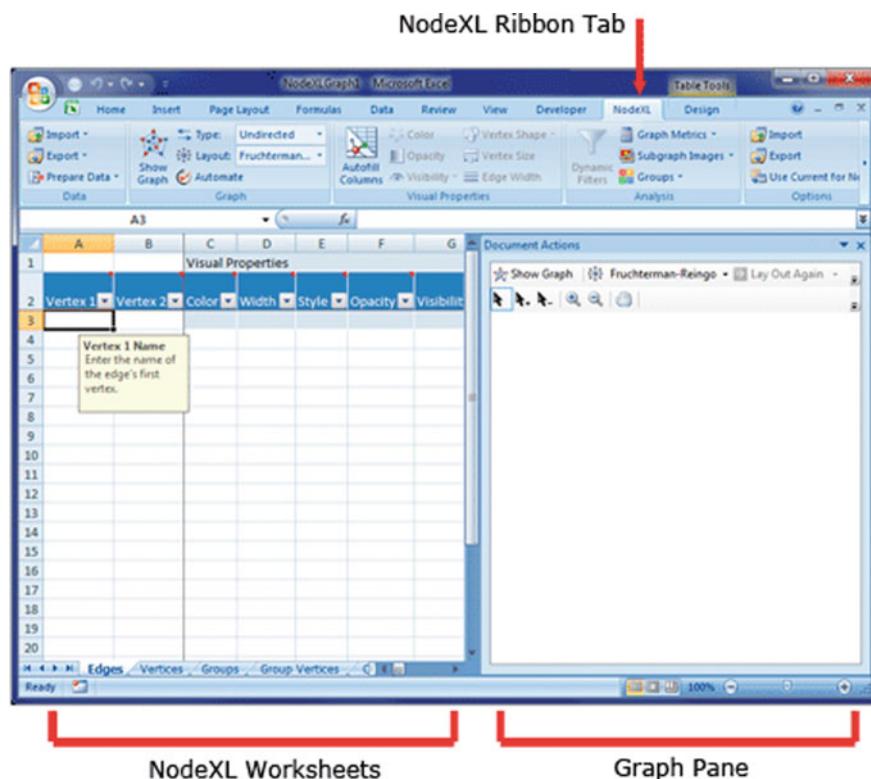
Figure 7.5 shows the input of the social network example given in Table 7.2 into NodeXL.

This is accomplished as follows.

First open a blank NodeXL template. Next open the matrix workbook in the same instance of Excel (just drag it into the title bar of the current NodeXL Excel instance) with the table shown in Table 7.2 (see Fig. 7.6).

To import this matrix, switch back to NodeXL and select the Data > Import > From Open Matrix Workbook as shown in Fig. 7.7.

Here a simple asymmetric matrix with equal weights for tie strengths is used. The graph pane for the input shown in Fig. 7.6 is displayed in Fig. 7.8.



**Fig. 7.4** Initial NodeXL screen

**Table 7.9** NodeXL worksheets

Worksheet	Description
Edges	Vertex 1 and Vertex 2 columns are used to define edges, with additional columns available to set properties
Vertices	NodeXL automatically creates vertex rows after the user selects <u>show the graph</u> . If there are isolated vertices, they can be entered as additional rows on the Vertices page
Groups	You can define edge relationships belonging to groups you define
Group vertices	Names of graph group vertices are entered
Overall metrics	Overall metrics are displayed. Versions higher than NodeXL Basic are required
Group edges	Worksheet created when group metrics are calculated

The screenshot shows a Microsoft Excel spreadsheet with two tabs visible at the top: 'Edge' and 'Vertices'. The 'Edge' tab is active, displaying a table with columns 'Vertex 1', 'Vertex 2', and 'Weight'. The 'Vertices' tab is also visible, showing a table with columns for visual properties like color, width, and style, along with columns for labels, graph metrics, and other columns.

Vertex 1	Vertex 2	Edge
Weight		
Daisy	Fern	1
Betty	Fern	1
Daisy	Edward	1
Albert	Edward	1
Albert	Daisy	1
Betty	Charles	1
Albert	Charles	1
Albert	Betty	1

A	B	C	D	E	F	G	H	I	J	K	N	O
	Visual Properties						Labels	Label Text	Label Font	Graph Metrics	Other Columns	
1	Vertex 1	Vertex 2	Color	Width	Style	Opacity	Visibility	Label	Color	Size	Reciprocated?	Add Your Own
2	Albert	Betty						AB				2
3	Albert	Charles						AC				2
4	Albert	Daisy						AD				2
5	Albert	Edward						AE				2
6	Betty	Albert						BA				2
7	Betty	Charles						BC				2
8	Betty	Fern						BF				2
9	Charles	Albert						CA				2
10	Charles	Betty						CB				2
11	Daisy	Albert						DA				2
12	Daisy	Edward						DE				2
13	Daisy	Fern						DF				2
14	Edward	Albert						EA				2
15	Edward	Daisy						ED				2
16	Fern	Betty						FB				2
17	Fern	Daisy						FD				2

**Fig. 7.5** NodeX input for social network example

Figure 7.9 gives vertices output. Note that not all of this output can be obtained with the free version of NodeXL.

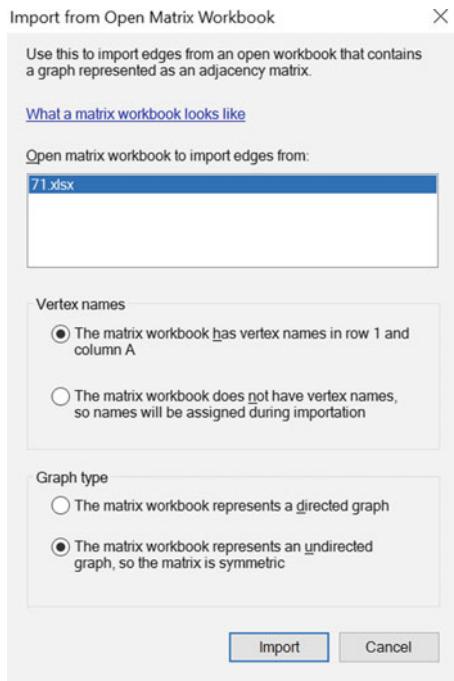
Figure 7.10 displays statistics provided by NodeX for the social network example.

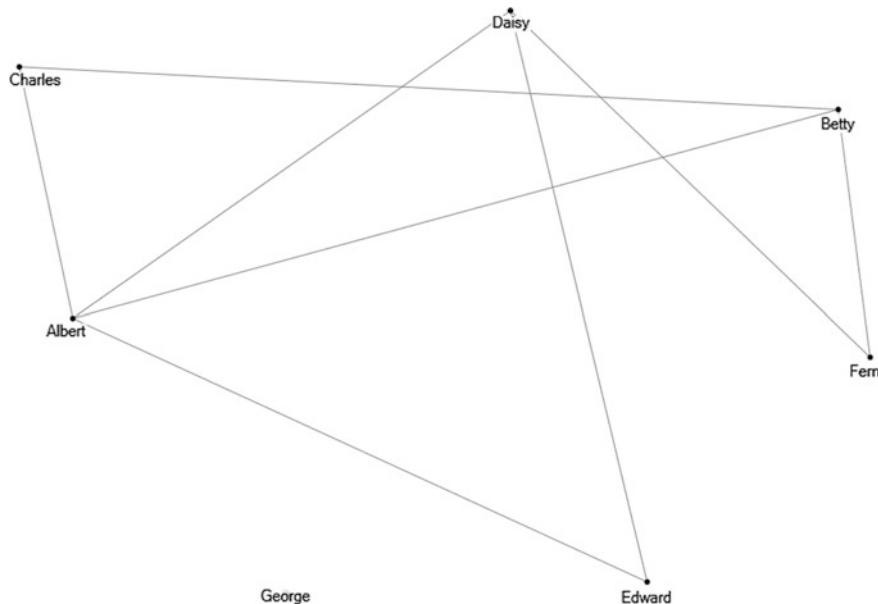
Figure 7.11 displays degree relationships for inputs and outputs.

Figure 7.12 shows Betweenness Centrality. Displays for Closeness Centrality and Eigenvector Centrality are also displayed within NodeXL.

Figure 7.13 displays clustering coefficients.

	A	B	C	D	E	F	G	H
1		Albert	Betty	Charles	Daisy	Edward	Fern	George
2	Albert	0	1	1	1	1	0	0
3	Betty	1	0	1	0	0	1	0
4	Charles	1	1	0	0	0	0	0
5	Daisy	1	0	0	0	1	1	0
6	Edward	1	0	0	1	0	0	0
7	Fern	0	1	0	1	0	0	0
8	George	0	0	0	0	0	0	0

**Fig. 7.6** NodeXL imported table**Fig. 7.7** NodeXL import



**Fig. 7.8** NodeXL graph display for simple social network

	A	I	J	K	L	M	N	O	P
1	rics	In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank	Clustering Coefficient	Reciprocated Vertex Pair Ratio
2	Vertex	4	4	7.000	0.167	0.229	1.439	0.333	1.000
3	A	4	4	7.000	0.167	0.229	1.439	0.333	1.000
4	B	Vertex Name Enter the name of the vertex.	3	3.000	0.143	0.178	1.117	0.333	1.000
5	C		2	0.000	0.125	0.145	0.772	1.000	1.000
6	D		3	3.000	0.143	0.178	1.117	0.333	1.000
7	E		2	0.000	0.125	0.145	0.772	1.000	1.000
8	F		2	1.000	0.125	0.126	0.783	0.000	1.000
9									

**Fig. 7.9** NodeX vertices output

## Network Analysis of Facebook Network or Other Networks

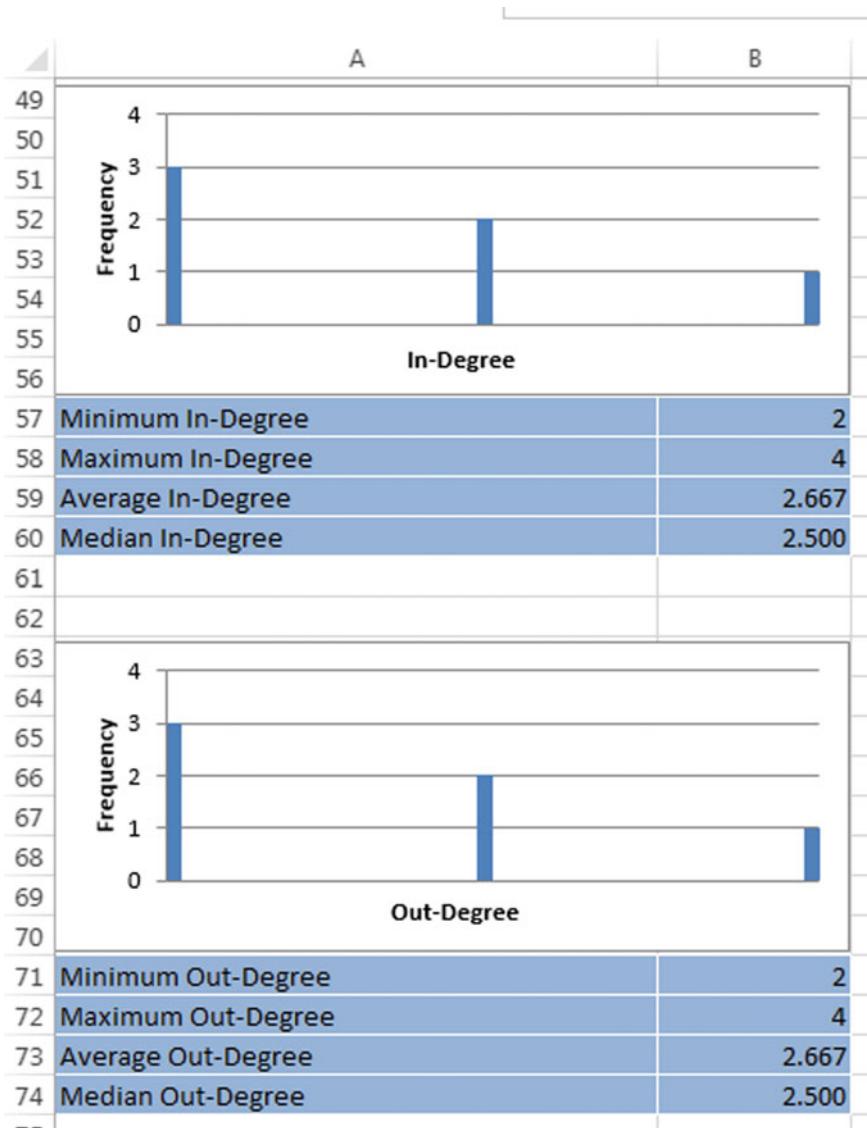
Many of you are probably members of several Social Networks such as LinkedIn, Facebook, and Twitter. Therefore it might be interesting to carry out a Network Analysis on such a Network. We demonstrate with a Chrome Extension, which can be used to collect, visualize and download network data. Even though the Extension accesses the Face-book network, the data are only locally stored on the user's computer.

Graph Metric	Value
Graph Type	Undirected
Vertices	7
Unique Edges	8
Edges With Duplicates	0
Total Edges	8
Self-Loops	0
Reciprocated Vertex Pair Ratio	Not Applicable
Reciprocated Edge Ratio	Not Applicable
Connected Components	2
Single-Vertex Connected Components	1
Maximum Vertices in a Connected Component	6
Maximum Edges in a Connected Component	8
Maximum Geodesic Distance (Diameter)	2
Average Geodesic Distance	1.222222
Graph Density	0.380952381
Modularity	Not Applicable
NodeXL Version	1.0.1.381

**Fig. 7.10** NodeXL overall metrics for simple social network example

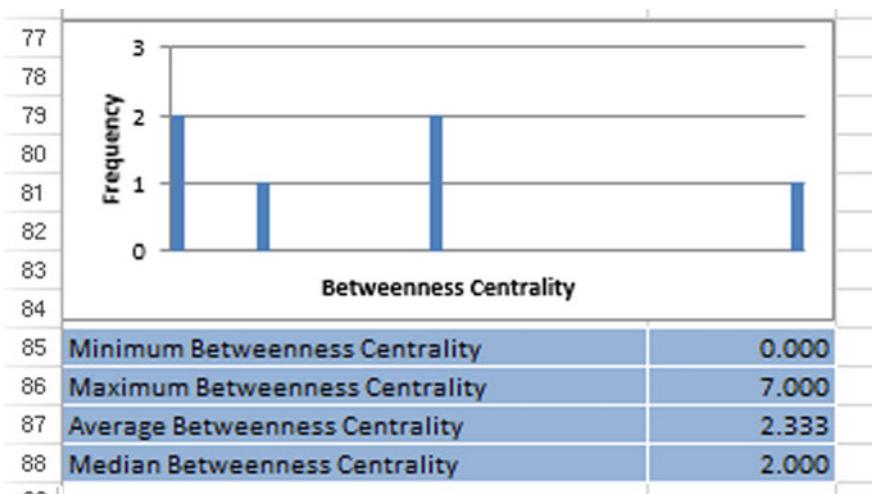
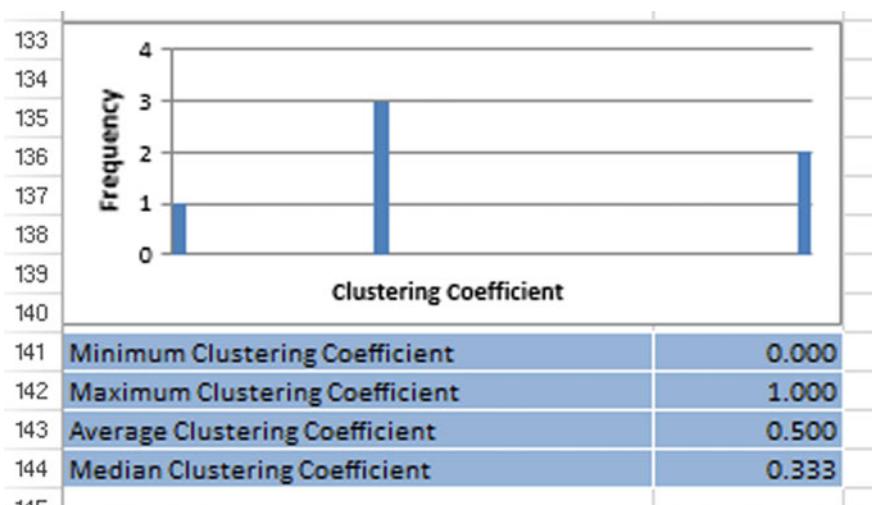
The first step to have the Chrome Browser installed and then add the Extension by following the link from <https://lostcircles.com> from the Chrome Webstore. Then one needs to be logged into Facebook inside Chrome Press “Start loading” from the Lost Circles menu.

One can then use the “Visualize” Button to see network. It is very interesting to see, how the network changes, while data are acquired.



**Fig. 7.11** In-degree and out-degree displays from NodeXL

In order to further analyze these data, as described in the previous section, we can download the data. We have so far described NodeXL for Network Analysis. The full version is expensive for a student or private user, while the free version has limited options. An excellent alternative is the software Gephi which is an interactive visualization and exploration platform for all kinds of networks with the goal to explore and understand graphs.

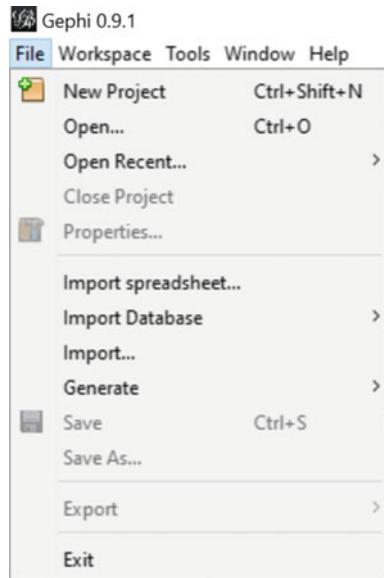
**Fig. 7.12** Betweenness centrality display**Fig. 7.13** Clustering coefficient display

<https://github.com/gephi/gephi/wiki>  
<https://gephi.org/>

We recommend version 0.9.1 even though this is not the latest version and which can be downloaded under <https://github.com/gephi/gephi/releases/tag/v0.9.1>.

After installing Gephi, we now open the file previously generated with Lost Circles (see Fig. 7.14).

**Fig. 7.14** Gephi opening screen



After selecting “open” we next see data table, graph, or Statistics by choosing the desired options (see Fig. 7.15).

The graph displayed may not be as are used to. If we now want to use NodeXL to again display the graph, we can do this. Since the free version of NodeXL can only open CSV files, we export our Facebook data table within Gephi. We need to export both the Nodge data table and the Edge data table. In Figs. 7.16 and 7.17 our Facebook Social Network has been displayed using NodeXL using two different styles.

We observe a very dense network, which could be a student’s school friends, and smaller networks such as family relatives, university friends, or other social interactions, which may have little interaction with the student’s initial school friends. We will later describe how to obtain such data yourself. However looking at friendship graphs makes the connections of a person visible. One could add additional dimensions by observing how the network may change over time which could be achieved by filtering the graph by a time range. Or one could add a weight to each connection by how frequently messages are exchanged across connections. In Chap. 4 the recency, frequency and monetary model was introduced. One could apply similar logic to social network friends using the recency of interaction, and the frequency and length of message to discover the “best friends” or “former associates.”

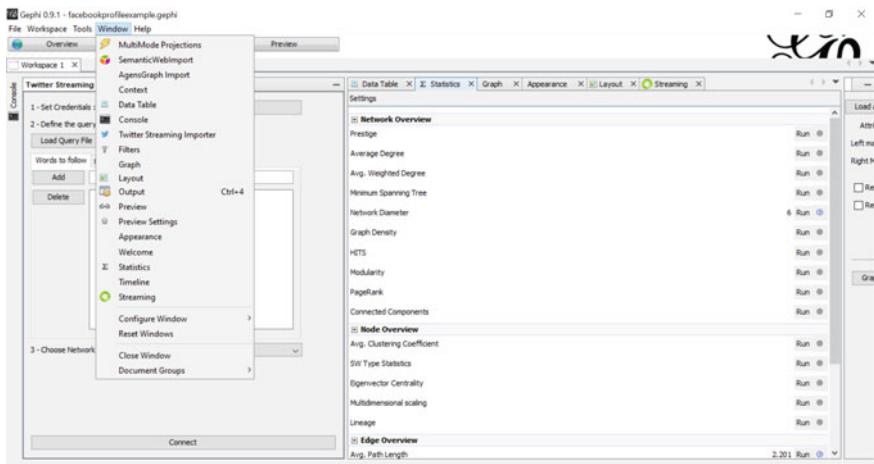


Fig. 7.15 Gephi screen

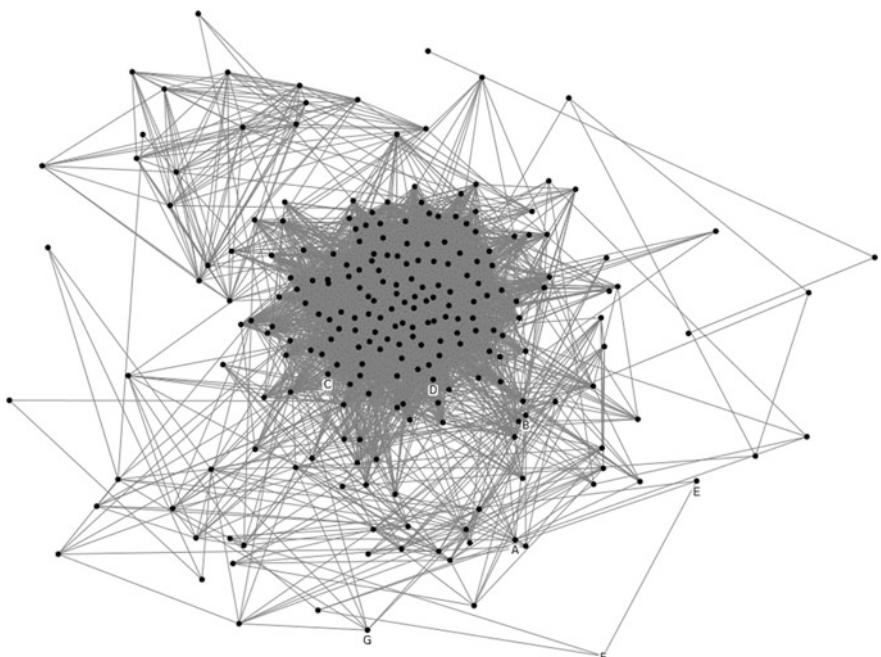
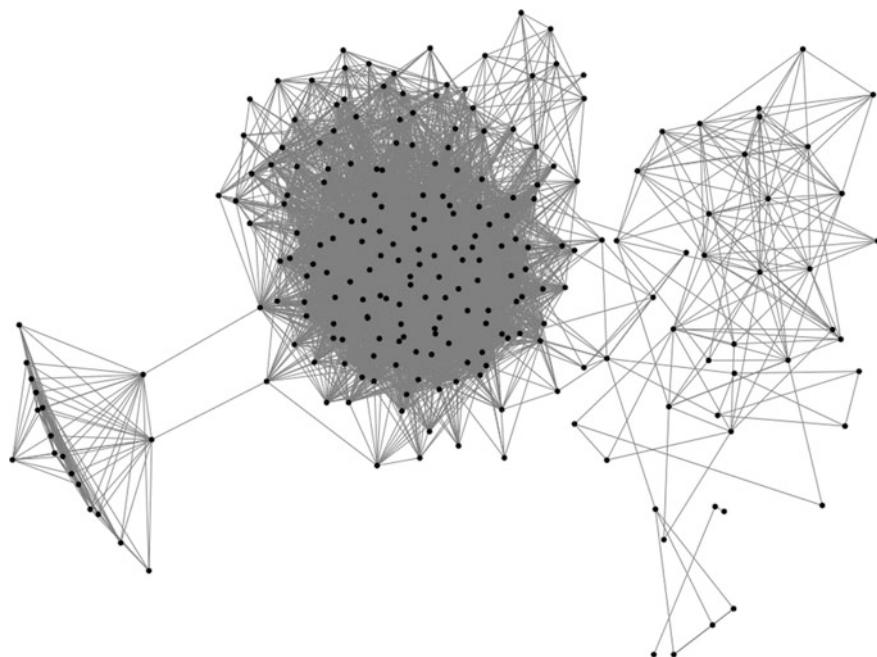


Fig. 7.16 NodeXL social network graph in Fruchterman Reingold style



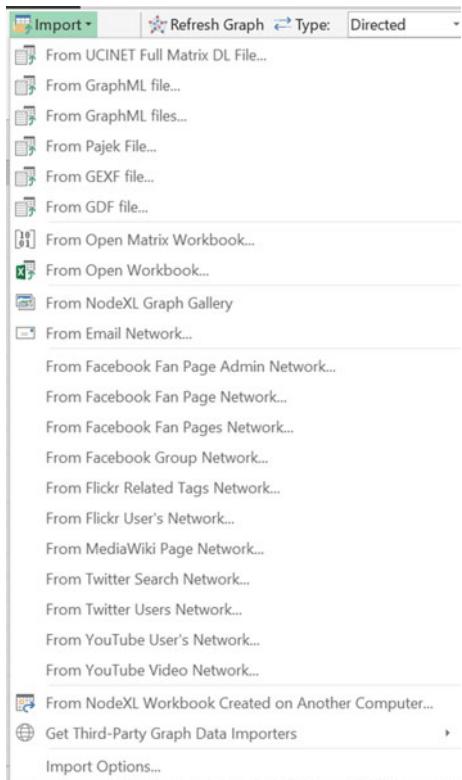
**Fig. 7.17** NodeXL social network graph in Haren-Koren Style

## Link Analysis of Your Emails

Not all of us may have a Facebook account or are members of a similar social network. However most of us communicate by e-mail. Who do we communicate with? Who are our friends. The value of such collection is demonstrated by the secret collection of such information by intelligence agencies. If it is useful to others, it might be at least interesting for us to see.

NodeXL offers a free version with limited features and as well as a full version. For Students, the full version is still very affordable—currently \$39 per year. We recommend starting with the free version. If the free features are not sufficient, you may be able to obtain a 30 day trial version by contacting NodeXL by e-mail. The NodeXL Pro version allows analysis of networks from different sources. The link is [NodeXL -> Import -> From Email Network](#) (see Fig. 7.18).

**Fig. 7.18** NodeXL will then import the information about to whom you send or receive e-mails



## Link Analysis Application with PolyAnalyst (Olson and Shi 2007)

NodeXL provides a simple way to work with link analysis, applicable in many contexts. It is a bit touchy to load, and some features require investing nominal amounts of money. We now discuss a well-developed system for applying link analysis to text data mining.

Text mining can be used for many applications. One such application is to analyze a series of text messages, such as customer complaints. Let us assume a distributor of printers deals with three products: inkjet black and white printers, inkjet color printers, and laser black and white printers. Table 7.10 gives five such messages, providing product, a quality rating on a 1 (bad) to 5 (best) scale, the name of the representative involved, and comments.

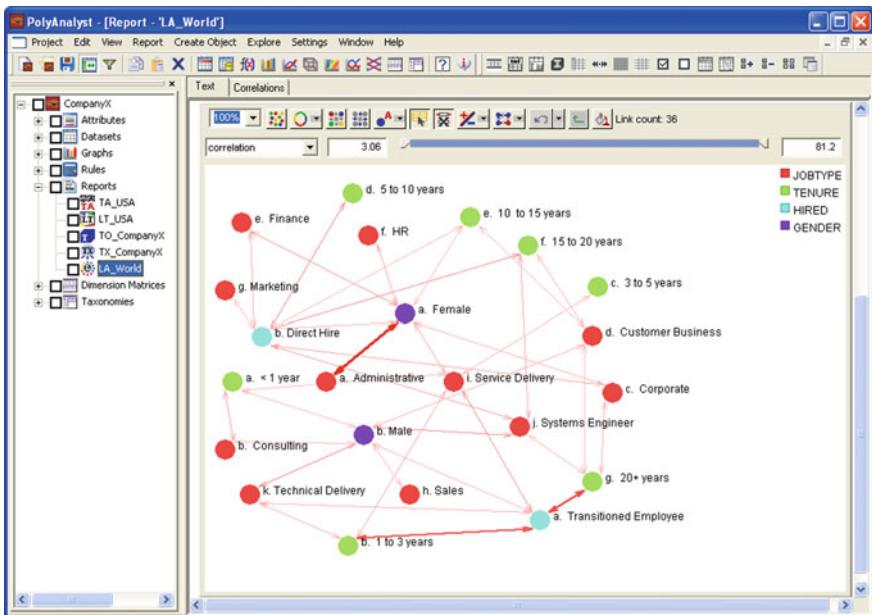
This set of comments involves some variety of comments. In order to apply text mining, the first step is to generate a set of key words. Some software products, such as PolyAnalyst by Megaputer, provide Text Analyst capabilities that focus on words that appear at a high rate. The user of the system can delete words which are not pertinent, and a file developed that will include the sentences where these

**Table 7.10** Printer complaints

Printer product	Quality rating	Rep.	Comments
Inkjet	1	Ben	This printer is a piece of <b>junk</b> . It is so <b>cheap</b> that it constantly <b>clogs</b> . In futile attempts to fix it, it <b>breaks</b> regularly
Color	3	Abner	I wish to commend your representative, who was very <b>understanding</b> of my lack of <b>knowledge</b> , and <b>kindly</b> provided every guidance to help me learn about this printer
Inkjet	2	Chuck	The printer works most of the time, but jams <b>paper</b> and involves expensive <b>service</b> . Furthermore, your representative was highly <b>abusive</b> on some occasions, and I won't be bothering you with future business once I settle this difficult matter
Inkjet	2	Abner	I am not happy with your printer. It <b>smears paper</b> when it is in <b>service</b> , which is not all that often
Laser	4	Dennis	Your representative was very <b>knowledgeable</b> about your product, and enabled me to get it to function

selected key words appear. One use of such files is to condense large documents to some percentage of their original size, focusing on sentences containing words of interest to the user. Another use is to identify those terms of interest to the user. For instance, in Table 7.10, words such as those in bold could be selected for link analysis. Note the ability to pick up different variants of the key terms (knowledge and knowledgeable, break and breaks (or broken), kind and kindly). Then a data file can be constructed using each of the key terms as a binary (yes/no, Boolean, 0/1) variable reflecting the presence or absence of the key term in each message. The file containing these key-term variables can include any other variables, such as the categorical variables for printer product and representative, as well as numeric data for quality ratings. Figure 7.19 shows the screen from PolyAnalyst for a small data set of 100 complaints (from which those in Table 7.10 were selected). This is an initial output, where all variables are displayed where any of the variables had common entries.

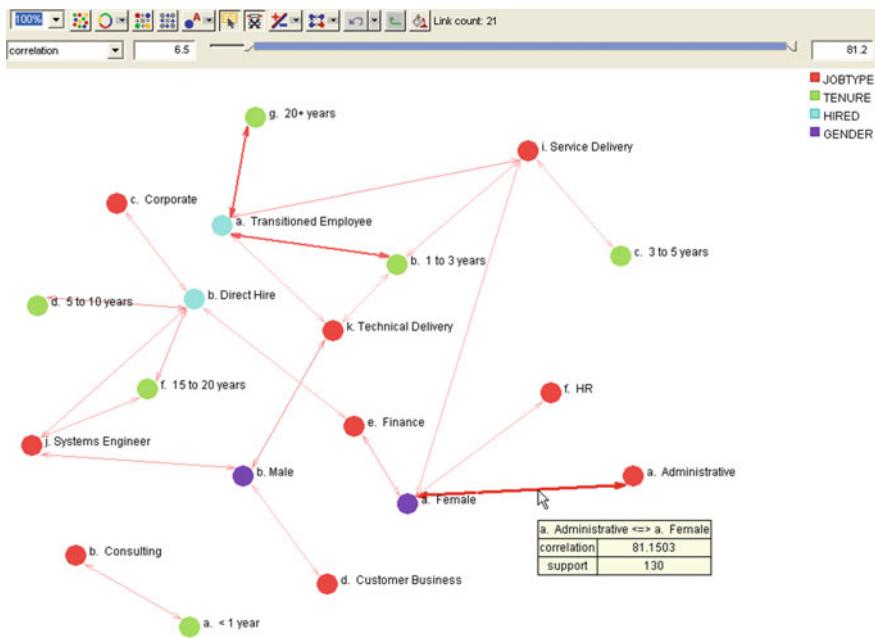
The link analysis utilizes all categorical and binary data (not the numeric quality rating). While it doesn't show in Fig. 7.19, positive and negative correlations are color coded. The user has the option of deleting either. The system uses a minimum value to include relationships. One factor for this data set is that it was highly skewed, with 70 inkjet printers, 20 color inkjets, and only 10 laser printers. Therefore, the links shown between laser printers and representatives is based upon a very small sample (and representatives like Emil had no chance to appear). This may be appropriately useful in some cases (maybe Emil doesn't sell laser printers), but since the sample size is small, there is no linkage shown of Emil's connection with laser printers.



**Fig. 7.19** PolyAnalyst initial link analysis output

Since Fig. 7.19 was very cluttered, the analyst has the ability to reset the minimum count necessary to show in the link analysis display. Figure 7.20 gives such an output for a minimum setting of 3 occurrences.

This makes relationships much clearer. It appears that Chuck deals with a number of inkjet printers, but his customers tend to think that they are junk. All that shows for Dennis is that he often sells laser printers. (This could indicate that Chuck concentrates on the low end of the business, and Dennis on the high end.) Color printers seem to have trouble with cartridges. There are no other systematic patterns at the minimum setting of 3 between products and key words. There are relationships among key words, in that those who complain about cheap printers that clog encounter abusive representatives. Those who complement representatives as being understanding also tend to include the word kind. Those who use the word service often use the word smear. The identification of relationships through link analysis can show some patterns or trends that might be useful in identifying communication flaws of representatives, or product defects.



**Fig. 7.20** PolyAnalyst initial link analysis output at minimum setting of 3

## Summary

Link analysis is valuable in many contexts. The growth of social networks has proven an especially rich domain for its application. This chapter began with what is hoped was a very simple demonstration of basic link analysis (social network) terms and measures. NodeXL was reviewed as a widely available tool that is very good at generating graphs. PolyAnalyst was also reviewed, a more powerful tool providing a means to obtain and apply link analysis to real problems defined in terms of text mining.

## References

- Knoke D, Yang S (2008) Social network analysis, 2nd edn. Sage Publications, Thousand Oaks, CA  
 Olson DL, Shi Y (2007) Introduction to business data mining. Irwin/McGraw-Hill, NY

# 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

**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.