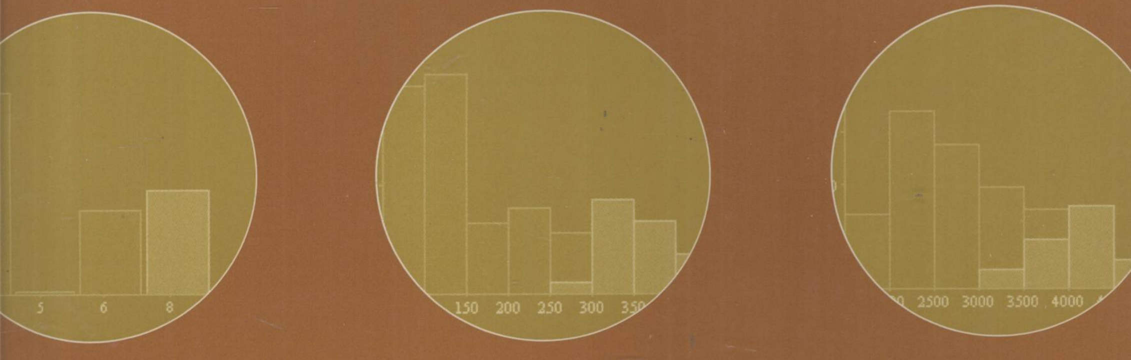
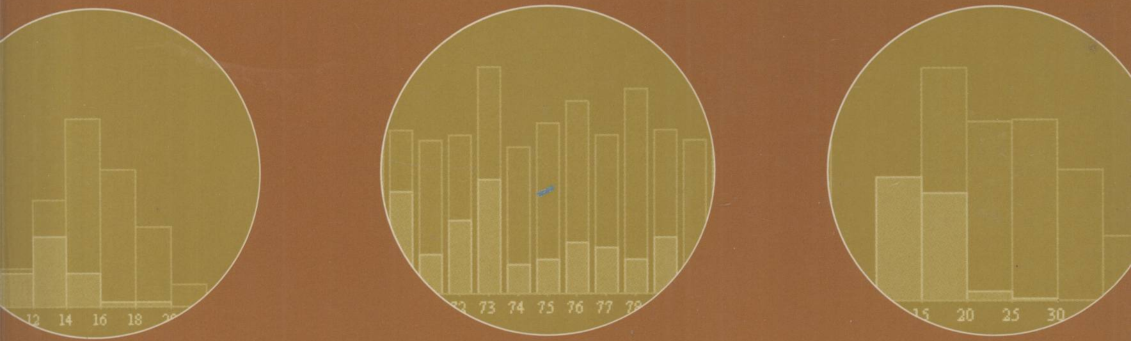
WWW.
LINK AVAILABLE

WILEY



MAKING SENSE OF DATA

A Practical Guide to Exploratory Data Analysis and Data Mining



GLENN J. MYATT

Making Sense of Data

A Practical Guide to Exploratory Data Analysis and Data Mining

Glenn J. Myatt

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

ISBN-13: 978-0-470-07471-8
ISBN-10: 0-470-07471-X

Printed in the United States of America
10 9 8 7 6 5 4 3

Making Sense of Data

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

Preface

Almost every field of study is generating an unprecedented amount of data. Retail companies collect data on every sales transaction, organizations log each click made on their web sites, and biologists generate millions of pieces of information related to genes daily. The volume of data being generated is leading to *information overload* and the ability to make sense of all this data is becoming increasingly important. It requires an understanding of exploratory data analysis and data mining as well as an appreciation of the subject matter, business processes, software deployment, project management methods, change management issues, and so on.

The purpose of this book is to describe a practical approach for making sense out of data. A step-by-step process is introduced that is designed to help you avoid some of the common pitfalls associated with complex data analysis or data mining projects. It covers some of the more common tasks relating to the analysis of data including (1) how to summarize and interpret the data, (2) how to identify nontrivial facts, patterns, and relationships in the data, and (3) how to make predictions from the data.

The process starts by understanding what business problems you are trying to solve, what data will be used and how, who will use the information generated and how will it be delivered to them. A plan should be developed that includes this problem definition and outlines how the project is to be implemented. Specific and measurable success criteria should be defined and the project evaluated against them.

The relevance and the quality of the data will directly impact the accuracy of the results. In an ideal situation, the data has been carefully collected to answer the specific questions defined at the start of the project. Practically, you are often dealing with data generated for an entirely different purpose. In this situation, it will be necessary to prepare the data to answer the new questions. This is often one of the most time-consuming parts of the data mining process, and numerous issues need to be thought through.

Once the data has been collected and prepared, it is now ready for analysis. What methods you use to analyze the data are dependent on many factors including the problem definition and the type of data that has been collected. There may be many methods that could potentially solve your problem and you may not know which one works best until you have experimented with the different alternatives. Throughout the technical sections, issues relating to when you would apply the different methods along with how you could optimize the results are discussed.

Once you have performed an analysis, it now needs to be delivered to your target audience. This could be as simple as issuing a report. Alternatively, the delivery may involve implementing and deploying new software. In addition to any technical challenges, the solution could change the way its intended audience operates on a daily basis, which may need to be managed. It will be important to understand how well the solution implemented in the field actually solves the original business problem.

Any project is ideally implemented by an interdisciplinary team, involving subject matter experts, business analysts, statisticians, IT professionals, project managers, and data mining experts. This book is aimed at the entire interdisciplinary team and addresses issues and technical solutions relating to data analysis or data mining projects. The book could also serve as an introductory textbook for students of any discipline, both undergraduate and graduate, who wish to understand exploratory data analysis and data mining processes and methods.

The book covers a series of topics relating to the process of making sense of data, including

- Problem definitions
- Data preparation
- Data visualization
- Statistics
- Grouping methods
- Predictive modeling
- Deployment issues
- Applications

The book is focused on practical approaches and contains information on how the techniques operate as well as suggestions for when and how to use the different methods. Each chapter includes a further reading section that highlights additional books and online resources that provide background and other information. At the end of selected chapters are a set of exercises designed to help in understanding the respective chapter's materials.

Accompanying this book is a web site (http://www.makingsenseofdata.com/) containing additional resources including software, data sets, and tutorials to help in understanding how to implement the topics covered in this book.

In putting this book together, I would like to thank the following individuals for their considerable help: Paul Blower, Vinod Chandnani, Wayne Johnson, and Jon Spokes. I would also like to thank all those involved in the review process for the book. Finally, I would like to thank the staff at John Wiley & Sons, particularly Susanne Steitz, for all their help and support throughout the entire project.

Contents

Chapter 1

Introduction

1.1 OVERVIEW

Disciplines as diverse as biology, economics, engineering, and marketing measure, gather and store data primarily in electronic databases. For example, retail companies store information on sales transactions, insurance companies keep track of insurance claims, and meteorological organizations measure and collect data concerning weather conditions. Timely and well-founded decisions need to be made using the information collected. These decisions will be used to maximize sales, improve research and development projects and trim costs. Retail companies must be able to understand what products in which stores are performing well, insurance companies need to identify activities that lead to fraudulent claims, and meteorological organizations attempt to predict future weather conditions. The process of taking the raw data and converting it into meaningful information necessary to make decisions is the focus of this book.

It is practically impossible to make sense out of data sets containing more than a handful of data points without the help of computer programs. Many free and commercial software programs exist to sift through data, such as spreadsheets, data visualization software, statistical packages, OLAP (On-Line Analytical Processing) applications, and data mining tools. Deciding what software to use is just one of the questions that must be answered. In fact, there are many issues that should be thought through in any exploratory data analysis/data mining project. Following a predefined process will ensure that issues are addressed and appropriate steps are taken.

Any exploratory data analysis/data mining project should include the following steps:

1. **Problem definition:** The problem to be solved along with the projected deliverables should be clearly defined, an appropriate team should be put together, and a plan generated for executing the analysis.
2. **Data preparation:** Prior to starting any data analysis or data mining project, the data should be collected, characterized, cleaned, transformed, and partitioned into an appropriate form for processing further.

Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining,
By Glenn J. Myatt

3. **Implementation of the analysis:** On the basis of the information from steps 1 and 2, appropriate analysis techniques should be selected, and often these methods need to be optimized.
4. **Deployment of results:** The results from step 3 should be communicated and/or deployed into a preexisting process.

Although it is usual to follow the order described, there will be some interactions between the different steps. For example, it may be necessary to return to the data preparation step while implementing the data analysis in order to make modifications based on what is being learnt. The remainder of this chapter summarizes these steps and the rest of the book outlines how to execute each of these steps.

1.2 PROBLEM DEFINITION

The first step is to define the business or scientific problem to be solved and to understand how it will be addressed by the data analysis/data mining project. This step is essential because it will create a focused plan to execute, it will ensure that issues important to the final solution are taken into account, and it will set correct expectations for those both working on the project and having a stake in the project's results. A project will often need the input of many individuals including a specialist in data analysis/data mining, an expert with knowledge of the business problems or subject matter, information technology (IT) support as well as users of the results. The plan should define a timetable for the project as well as providing a comparison of the cost of the project against the potential benefits of a successful deployment.

1.3 DATA PREPARATION

In many projects, getting the data ready for analysis is the most time-consuming step in the process. Pulling the data together from potentially many different sources can introduce difficulties. In situations where the data has been collected for a different purpose, the data will need to be transformed into an appropriate form for analysis. During this part of the project, a thorough familiarity with the data should be established.

1.4 IMPLEMENTATION OF THE ANALYSIS

Any task that involves making decisions from data almost always falls into one of the following categories:

- **Summarizing the data:** *Summarization is a process in which the data is reduced for interpretation without sacrificing any important information.* Summaries can be developed for the data as a whole or any portion of the data. For example, a retail company that collected data on its transactions could develop summaries of the total sales transactions. In addition, the company could also generate summaries of transactions by products or stores.

- **Finding hidden relationships:** *This refers to the identification of important facts, relationships, anomalies or trends in the data, which are not obvious from a summary alone.* To discover this information will involve looking at the data from many angles. For example, a retail company may want to understand customer profiles and other facts that lead to the purchase of certain product lines.
- **Making predictions:** *Prediction is the process where an estimate is calculated for something that is unknown.* For example, a retail company may want to predict, using historical data, the sort of products that specific consumers may be interested in.

There is a great deal of interplay between these three tasks. For example, it is important to summarize the data before making predictions or finding hidden relationships. Understanding any hidden relationships between different items in the data can help in generating predictions. Summaries of the data can also be useful in presenting prediction results or understanding hidden relationships identified. This overlap between the different tasks is highlighted in the Venn diagram in Figure 1.1.

Exploratory data analysis and data mining covers a broad set of techniques for summarizing the data, finding hidden relationships, and making predictions. Some of the methods commonly used include

- **Summary tables:** The raw information can be summarized in multiple ways and presented in tables.
- **Graphs:** Presenting the data graphically allows the eye to visually identify trends and relationships.

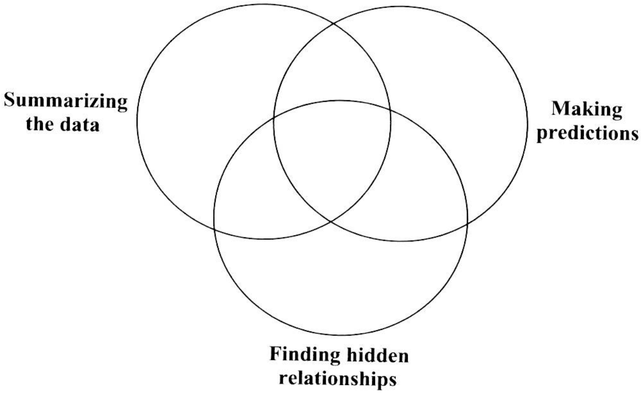


Figure 1.1. Data analysis tasks

- **Descriptive statistics:** These are descriptions that summarize information about a particular data column, such as the average value or the extreme values.
- **Inferential statistics:** Methods that allow claims to be made concerning the data with confidence.
- **Correlation statistics:** Statistics that quantify relationships within the data.
- **Searching:** Asking specific questions concerning the data can be useful if you understand the conclusion you are trying to reach or if you wish to quantify any conclusion with more information.
- **Grouping:** Methods for organizing a data set into smaller groups that potentially answer questions.
- **Mathematical models:** A mathematical equation or process that can make predictions.

The three tasks outlined at the start of this section (summarizing the data, finding hidden relationships, and making predictions) are shown in Figure 1.2 with a circle for each task. The different methods for accomplishing these tasks are also positioned on the Venn diagram. The diagram illustrates the overlap between the various tasks and the methods that can be used to accomplish them. The position of the methods is related to how they are often used to address the various tasks.

Graphs, summary tables, descriptive statistics, and inferential statistics are the main methods used to summarize data. They offer multiple ways of describing the data and help us to understand the relative importance of different portions of the data. These methods are also useful for characterizing the data prior to developing predictive models or finding hidden relationships. Grouping observations can be useful in teasing out hidden trends or anomalies in the data. It is also useful for characterizing the data prior to building predictive models. Statistics are used

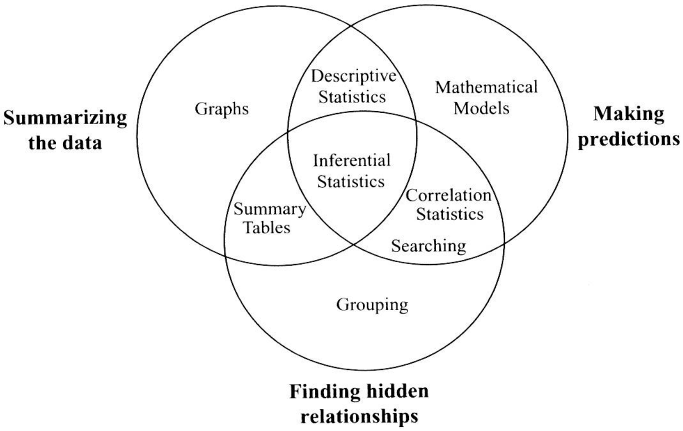


Figure 1.2. Data analysis tasks and methods

throughout, for example, correlation statistics can be used to prioritize what data to use in building a mathematical model and inferential statistics can be useful when validating trends identified from grouping the data. Creating mathematical models underpins the task of prediction; however, other techniques such as grouping can help in preparing the data set for modeling as well as helping to explain why certain predictions were made.

All methods outlined in this section have multiple uses in any data analysis or data mining project, and they all have strengths and weaknesses. On the basis of issues important to the project as well as other practical considerations, it is necessary to select a set of methods to apply to the problem under consideration. Once selected, these methods should be appropriately optimized to improve the quality of the results generated.

1.5 DEPLOYMENT OF THE RESULTS

There are many ways to deploy the results of a data analysis or data mining project. Having analyzed the data, a static report to management or to the customer of the analysis is one option. Where the project resulted in the generation of predictive models to use on an ongoing basis, these models could be deployed as standalone applications or integrated with other softwares such as spreadsheets or web pages. It is in the deployment step that the analysis is translated into a benefit to the business, and hence this step should be carefully planned.

1.6 BOOK OUTLINE

This book follows the four steps outlined in this chapter:

1. **Problem definition:** A discussion of the definition step is provided in Chapter 2 along with a case study outlining a hypothetical project plan. The chapter outlines the following steps: (1) define the objectives, (2) define the deliverables, (3) define roles and responsibilities, (4) assess the current situation, (5) define the timetable, and (6) perform a cost/benefit analysis.
2. **Data preparation:** Chapter 3 outlines many issues and methods for preparing the data prior to analysis. It describes the different sources of data. The chapter outlines the following steps: (1) create the data tables, (2) characterize the data, (3) clean the data, (4) remove unnecessary data, (5) transform the data, and (6) divide the data into portions when needed.
3. **Implementation of the analysis:** Chapter 4 provides a discussion of how summary tables and graphs can be used for communicating information about the data. Chapter 5 reviews a series of useful statistical approaches to summarizing the data and relationships within the data as well as making statements about the data with confidence. It covers the following topics: descriptive statistics, confidence intervals, hypothesis tests, the chi-square test, one-way analysis of variance, and correlation analysis. Chapter 6 describes a