

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.

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.

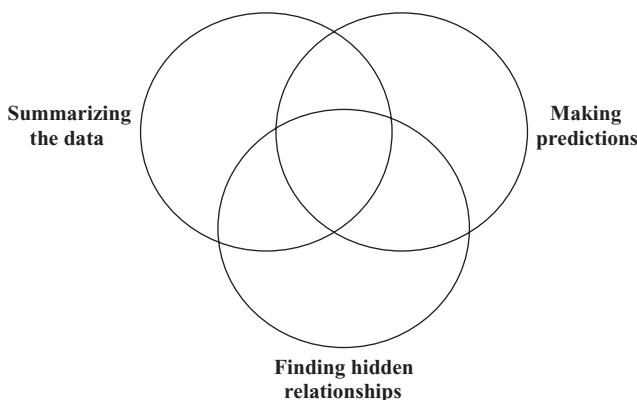


Figure 1.1. Data analysis tasks

- **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

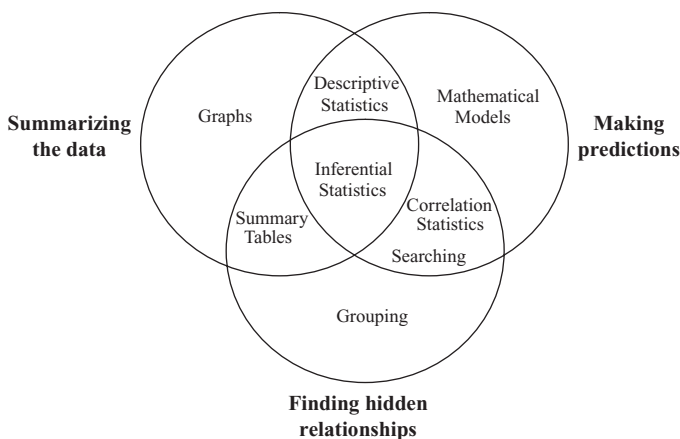


Figure 1.2. Data analysis tasks and methods

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

series of methods for grouping data including clustering, associative rules, and decision trees. Chapter 7 outlines the process and methods to be used in building predictive models. In addition, the chapter covers a series of methods including simple regression, k-nearest neighbors, classification and regression trees, and neural networks.

4. Deployment of results: Chapter 8 reviews some of the issues around deploying any results from data analysis and data mining projects including planning and executing deployment, measuring and monitoring the solution’s performance, and reviewing the entire project. A series of common deployment scenarios are presented. Chapter 9 concludes the book with a review of the whole process, a case study, and a discussion of data analysis

Table 1.1. Summary of project steps

Steps	Description
1. Problem definition	Define ● Objectives ● Deliverables ● Roles and responsibilities ● Current situation ● Timeline ● Costs and benefits
2. Data preparation	Prepare and become familiar with the data: ● Pull together data table ● Categorize the data ● Clean the data ● Remove unnecessary data ● Transform the data ● Partition the data
3. Implementation of the analysis	Three major tasks are ● Summarizing the data ● Finding hidden relationships ● Making prediction Select appropriate methods and design multiple experiments to optimize the results. Methods include ● Summary tables ● Graphs ● Descriptive statistics ● Inferential statistics ● Correlation statistics ● Searching ● Grouping ● Mathematical models
4. Deployment	● Plan and execute deployment based on the definition in step 1 ● Measure and monitor performance ● Review the project

and data mining issues associated with common applications. Exercises are included at the end of selected chapters to assist in understanding the material.

This book uses a series of data sets to illustrate the concepts from Newman (1998). The Auto-Mpg Database is used throughout to compare how the different approaches view the same data set. In addition, the following data sets are used in the book: Abalone Database, Adult Database, and the Pima Indians Diabetes Database.

1.7 SUMMARY

The four steps in any data analysis or data mining project are summarized in Table 1.1.

1.8 FURTHER READING

The CRISP-DM project (CRoss Industry Standard Process for Data Mining) has published a data mining process and describes details concerning data mining stages and relationships between the stages. It is available on the web at: <http://www.crisp-dm.org/>

SEMMA (Sample, Explore, Modify, Model, Assess) describes a series of core tasks for model development in the SAS[®] Enterprise Miner[™] software and a description can be found at: <http://www.sas.com/technologies/analytics/datamining/miner/semma.html>