1

INTRODUCTION

1.1 WHAT IS BUSINESS ANALYTICS?

Business analytics (BA) is the practice and art of bringing quantitative data to bear on decision-making. The term means different things to different organizations.

Consider the role of analytics in helping newspapers survive the transition to a digital world. One tabloid newspaper with a working-class readership in Britain had launched a web version of the paper, and did tests on its home page to determine which images produced more hits: cats, dogs, or monkeys. This simple application, for this company, was considered analytics. By contrast, the *Washington Post* has a highly influential audience that is of interest to big defense contractors: it is perhaps the only newspaper where you routinely see advertisements for aircraft carriers. In the digital environment, the *Post* can track readers by time of day, location, and user subscription information. In this fashion the display of the aircraft carrier advertisement in the online paper may be focused on a very small group of individuals—say, the members of the House and Senate Armed Services Committees who will be voting on the Pentagon's budget.

Business analytics, or more generically, *analytics*, includes a range of data analysis methods.

Many powerful applications involve little more than counting, rule checking, and basic arithmetic. For some organizations, this is what is meant by analytics.

The next level of business analytics, now termed *business intelligence* (BI), refers to the use of data visualization and reporting for becoming aware and understanding "what happened and what is happening." This is done by use of charts, tables, and dashboards to display, examine, and explore data. Business intelligence, which earlier consisted mainly of generating static reports, has evolved into more user-friendly and effective tools and practices, such as creating interactive dashboards that allow the user not only to access real-time data, but also to directly interact with it. Effective dashboards are those that tie directly to company data, and give managers a tool to see quickly what might not readily be apparent in a large complex database. One such tool for industrial operations managers

Machine Learning for Business Analytics: Concepts, Techniques, and Applications with JMP Pro®,
Second Edition. Galit Shmueli, Peter C. Bruce, Mia L. Stephens, Muralidhara A, and Nitin R. Patel.

displays customer orders in one two-dimensional display using color and bubble size as added variables. The resulting 2 by 2 matrix shows customer name, type of product, size of order, and length of time to produce.

Business analytics now typically includes BI as well as sophisticated data analysis methods, such as statistical models and machine learning algorithms used for exploring data, quantifying and explaining relationships between measurements, and predicting new records. Methods like regression models are used to describe and quantify "on average" relationships (e.g., between advertising and sales), to predict new records (e.g., whether a new patient will react positively to a medication), and to forecast future values (e.g., next week's web traffic).

Readers familiar with the earlier edition of this book might have noticed that the book title changed from *Data Mining for Business Analytics* to *Machine Learning for Business Analytics*. The change reflects the more recent term BA, which overtook the earlier term BI to denote advanced analytics. Today, BI is used to refer to data visualization and reporting. The change from *data mining* to *machine learning* reflects today's common use of *machine learning* to refer to algorithms that learn from data. This book uses primarily the term *machine learning*.

WHO USES PREDICTIVE ANALYTICS?

The widespread adoption of predictive analytics, coupled with the accelerating availability of data, has increased organizations' capabilities throughout the economy. A few examples:

Credit scoring: One long-established use of predictive modeling techniques for business prediction is credit scoring. A credit score is not some arbitrary judgement of creditworthiness; it is based mainly on a predictive model that uses prior data to predict repayment behavior.

Future purchases: A more recent (and controversial) example is Target's use of predictive modeling to classify sales prospects as "pregnant" or "not-pregnant." Those classified as pregnant could then be sent sales promotions at an early stage of pregnancy, giving Target a head start on a significant purchase stream.

Tax evasion: The US Internal Revenue Service found it was 25 times more likely to find tax evasion when enforcement activity was based on predictive models, allowing agents to focus on the most likely tax cheats (Siegel, 2013).

The business analytics toolkit also includes statistical experiments, the most common of which is known to marketers as A/B testing. These are often used for pricing decisions:

- Orbitz, the travel site, has found that it could price hotel options higher for Mac users than Windows users.
- Staples online store found that it could charge more for staplers if a customer lived far from a Staples store.

Beware the organizational setting where analytics is a solution in search of a problem: a manager, knowing that business analytics and machine learning are hot areas, decides that her organization must deploy them too, to capture that hidden value that must be lurking somewhere. Successful use of analytics and machine learning requires both an understanding of the business context where value is to be captured and an understanding of exactly what the machine learning methods do.

1.2 WHAT IS MACHINE LEARNING?

In this book, *machine learning* or *data mining* refers to business analytics methods that go beyond counts, descriptive techniques, reporting, and methods based on business rules. While we do introduce data visualization, which is commonly the first step into more advanced analytics, the book focuses mostly on the more advanced data analytics tools. Specifically, it includes statistical and machine learning methods that inform decision-making, often in automated fashion. Prediction is typically an important component, often at the individual level. Rather than "what is the relationship between advertising and sales?" we might be interested in "what specific advertisement, or recommended product, should be shown to a given online shopper at this moment?" Or we might be interested in clustering customers into different "personas" that receive different marketing treatment, then assigning each new prospect to one of these personas.

The era of Big Data has accelerated the use of machine learning. Machine learning methods, with their power and automaticity, have the ability to cope with huge amounts of data and extract value.

1.3 MACHINE LEARNING, AI, AND RELATED TERMS

The field of analytics is growing rapidly, both in terms of the breadth of applications, and in terms of the number of organizations using advanced analytics. As a result, there is considerable overlap and inconsistency in terms of definitions. Terms have also changed over time.

The older term *data mining* means different things to different people. To the general public, it may have a general, somewhat hazy and pejorative meaning of digging through vast stores of (often personal) data in search of something interesting. *Data mining*, as it refers to analytic techniques, has largely been superseded by the term *machine learning.*

Other terms that organizations use are *predictive analytics*, *predictive modeling*, and most recently *machine learning* and *artificial intelligence (AI).*

Many practitioners, particularly those from the IT and computer science communities, use the term AI to refer to all the methods discussed in this book. AI originally referred to the general capability of a machine to act like a human, and, in its earlier days, existed mainly in the realm of science fiction and the unrealized ambitions of computer scientists. More recently, it has come to encompass the methods of statistical and machine learning discussed in this book, as the primary enablers of that grand vision, and sometimes the term is used loosely to mean the same thing as *machine learning*. More broadly, it includes generative capabilities such as the creation of images, audio, and video.

Statistical Modeling vs. Machine Learning

A variety of techniques for exploring data and building models have been around for a long time in the world of statistics: linear regression, logistic regression, discriminant analysis, and principal components analysis, for example. But the core tenets of classical statistics—computing is difficult and data are scarce—do not apply in machine learning applications where both data and computing power are plentiful.

This is what gives rise to Daryl Pregibon's description of "data mining" (in the sense of machine learning) as "statistics at scale and speed" (Pregibon, 1999). Another major difference between the fields of statistics and machine learning is the focus in statistics on inference from a sample to the population regarding an "average effect"—for example, "a $1 price increase will reduce average demand by 2 boxes." In contrast, the focus in machine learning is on predicting individual records—"the predicted demand for person i given a $1 price increase is 1 box, while for person j it is 3 boxes." The emphasis that classical statistics places on inference (determining whether a pattern or interesting result might have happened by chance in our sample) is absent from machine learning. Note also that the term *inference* is often used in the machine learning community to refer to the process of using a model to make predictions for new data, also called *scoring*, in contrast to its meaning in the statistical community.

In comparison to statistics, machine learning deals with large datasets in an open-ended fashion, making it impossible to put the strict limits around the question being addressed that classical statistical inference would require. As a result, the general approach to machine learning is vulnerable to the danger of *overfitting*, where a model is fit so closely to the available sample of data that it describes not merely structural characteristics of the data, but random peculiarities as well. In engineering terms, the model is fitting the noise, not just the signal.

In this book, we use the term *machine learning algorithms* to refer to methods that learn directly from data, especially local data, often in layered or iterative fashion. In contrast, we use *statistical models* to refer to methods that apply global structure to the data that can be written as a simple mathematical equation. A simple example is a linear regression model (statistical) vs. a k-nearest neighbors algorithm (machine learning). A given record would be treated by linear regression in accord with an overall linear equation that applies to *all* the records. In k-nearest neighbors, that record would be classified in accord with the values of a small number of nearby records.

1.4 BIG DATA

Machine learning and Big Data go hand in hand. *Big Data* is a relative term—data today are big by reference to the past and to the methods and devices available to deal with them. The challenge Big Data presents is often characterized by the four V's—volume, velocity, variety, and veracity. *Volume* refers to the amount of data. *Velocity* refers to the flow rate—the speed at which it is being generated and changed. *Variety* refers to the different types of data being generated (time stamps, location, numbers, text, images, etc.). *Veracity* refers to the fact that data is being generated by organic distributed processes (e.g., millions of people signing up for services or free downloads) and not subject to the controls or quality checks that apply to data collected for a study.

Most large organizations face both the challenge and the opportunity of Big Data because most routine data processes now generate data that can be stored and, possibly, analyzed.

The scale can be visualized by comparing the data in a traditional statistical analysis on the large size (e.g., 15 variables and 5000 records) to the Walmart database. If you consider the traditional statistical study to be the size of a period at the end of a sentence, then the Walmart database is the size of a football field. And that probably does not include other data associated with Walmart—social media data, for example, which comes in the form of unstructured text.

If the analytical challenge is substantial, so can be the reward:

- OKCupid, the dating site, uses statistical models with their data to predict what forms of message content are most likely to produce a response.
- Telenor, a Norwegian mobile phone service company, was able to reduce subscriber turnover 37% by using models to predict which customers were most likely to leave and then lavishing attention on them.
- Allstate, the insurance company, tripled the accuracy of predicting injury liability in auto claims by incorporating more information about vehicle type.

The examples above are from Eric Siegel's *Predictive Analytics* (2013, Wiley).

Some extremely valuable tasks were not even feasible before the era of Big Data. Consider web searches, the technology on which Google was built. In early days, a search for "Ricky Ricardo Little Red Riding Hood" would have yielded various links to the "I Love Lucy" show, other links to Ricardo's career as a band leader, and links to the children's story of Little Red Riding Hood. Only once the Google database had accumulated sufficient data (including records of what users clicked on) would the search yield, in the top position, links to the specific *I Love Lucy* episode in which Ricky enacts, in a comic mixture of Spanish and English, Little Red Riding Hood for his infant son.

1.5 DATA SCIENCE

The ubiquity, size, value, and importance of Big Data has given rise to a new profession: the *data scientist*. *Data science* is a mix of skills in the areas of statistics, machine learning, math, programming, business, and IT. The term itself is thus broader than the other concepts we discussed above, and it is a rare individual who combines deep skills in all the constituent areas. Harlan Harris, in *Analyzing the Analyzers* (with Sean Murphy and Marck Vaisman, O'Reilly 2013) describes the skill sets of most data scientists as resembling a "T"—deep in one area (the vertical bar of the T), and shallower in other areas (the top of the T).

At a large data science conference session (Strata-Hadoop World, October 2014) most attendees felt that programming was an essential skill, though there was a sizable minority who felt otherwise. And, although Big Data is the motivating power behind the growth of data science, most data scientists do not actually spend most of their time working with terabyte-size or larger data.

Data of the terabyte or larger size would be involved at the deployment stage of a model. There are manifold challenges at that stage, most of them IT and programming issues related to data handling and tying together different components of a system. Much work must precede that phase. It is that earlier piloting and prototyping phase on which this book focuses—developing the statistical and machine learning models that will eventually be plugged into a deployed system. What methods do you use with what sorts of data and problems? How do the methods work? What are their requirements, their strengths, their weaknesses? How do you assess their performances?

1.6 WHY ARE THERE SO MANY DIFFERENT METHODS?

As can be seen in this book or any other resource on machine learning, there are many different methods for prediction and classification. You might ask yourself why they coexist and whether some are better than others. The answer is that each method has advantages and disadvantages. The usefulness of a method can depend on factors such as the size of the dataset, the types of patterns that exist in the data, whether the data meet some underlying assumptions of the method, how noisy the data are, and the particular goal of the analysis. A small illustration is shown in Figure 1.1, where the goal is to find a combination of *household income level* and *household lot size* that separate owners (blue markers) from non-owners (orange markers) of riding mowers. The first method (top panel) looks only for horizontal lines to separate owners from non-owners, whereas the second method (bottom panel) looks for a single diagonal line.

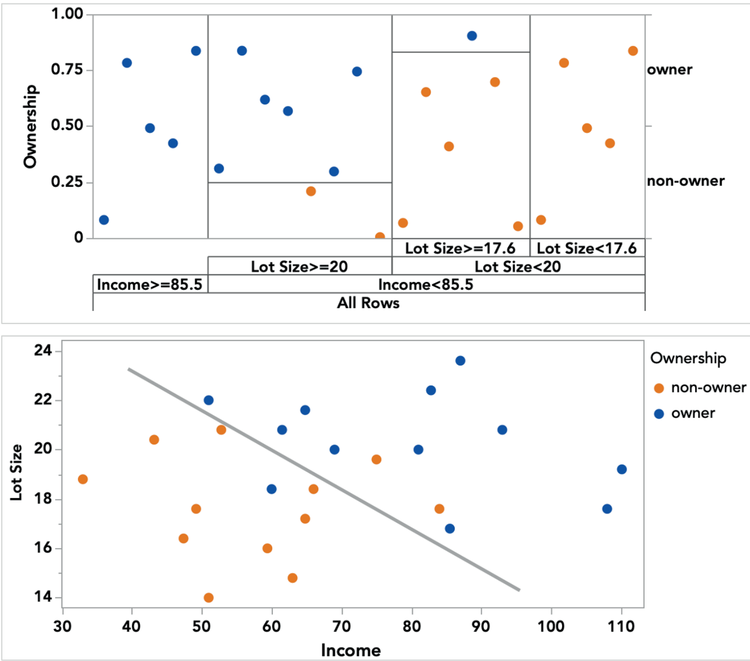

FIGURE 1.1 Two methods for separating owners from non-owners

Different methods can lead to different results, and their performances can vary. It is therefore customary in machine learning to apply several different methods and perhaps their combination and select the one that appears most useful for the goal at hand.

1.7 TERMINOLOGY AND NOTATION

Because of the hybrid origins of data science, its practitioners often use multiple terms to refer to the same thing. For example, in the machine learning (artificial intelligence) field, the variable being predicted is the *output variable* or *target variable*. A categorical output variable is often called a *label*. To a statistician or social scientist, the variable being predicted is the *dependent variable* or the *response*. Here is a summary of terms used:

Algorithm Refers to a specific procedure used to implement a particular machine learning technique: classification tree, discriminant analysis, and the like.

Attribute see **Predictor**.

Binary variable A variable that takes on two discrete values (e.g., fraud and non-fraud transactions); also called *binominal variable*.

Case see **Observation**.

Confidence A performance measure in association rules of the type "IF *A* and *B* are purchased, THEN *C* is also purchased." Confidence is the conditional probability that *C* will be purchased IF *A* and *B* are purchased.

Dependent variable see **Response**.

Estimation see **Prediction**.

Feature see **Predictor**. The term feature is also used in the context of selecting variables (feature selection) or generating new variables (feature generation) through some mechanism. More broadly, this process is called feature engineering.

Holdout sample (or **Holdout set**) A sample of data not used in fitting a model, used to assess the performance of that model; this book uses the term *validation set* or, if one is used in the problem, *test set* instead of *holdout sample*.

Inference In statistics, the process of accounting for chance variation when making estimates or drawing conclusions based on samples; in machine learning, the term often refers to the process of using a model to make predictions for new data (see Score).

Input variable see **Predictor**.

Label A nominal (categorical) attribute being predicted in supervised learning.

Model An algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).

Nominal variable A variable that takes on one of several fixed values, for example, a flight could be on-time, delayed, or canceled; also called *categorical variable* or *polynominal variable.*

Numerical variable A variable that takes on numerical (integer or real) values.

Observation The unit of analysis on which the measurements are taken (a customer, a transaction, etc.); also called *instance*, *sample*, *example*, *case*, *record*, *pattern*, or *row*. (Each row typically represents a record; each column, a variable. Note that the use of the term "sample" here is different from its usual meaning in statistics, where it refers to all the data sampled from a larger data source, not simply one record.)

Outcome variable see **Response**.

Output variable see **Response**.

$P(A|B)$ The conditional probability of event A occurring given that event B has occurred. Read as "the probability that A will occur given that B has occurred."

Profile The set of measurements on an observation (e.g., the height, weight, and age of a person).

Positive class The class of interest in a binary outcome variable (e.g., purchasers in the outcome variable *purchase/no purchase*); the positive class need not be favorable.

Prediction The prediction of the value of a continuous output variable; also called *estimation.*

Predictor A variable, usually denoted by X, used as an input into a predictive model; also called a *feature*, *input variable*, *independent variable*, or, from a database perspective, a *field*.

Record see **Observation**.

Response A variable, usually denoted by Y, which is the variable being predicted in supervised learning; also called *dependent variable*, *output variable*, *target variable*, or *outcome variable*.

Score A predicted value or class. *Scoring new data* means to use a model developed with training data to predict output values in new data.

Success class see **Positive class**

Supervised learning The process of providing an algorithm (logistic regression, classification tree, etc.) with records in which an output variable of interest is known, and the algorithm "learns" how to predict this value with new records where the output is unknown.

Target see **Response**.

Test data (or **Test set**) The portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on additional data.

Training data (or **Training set**) The portion of data used to fit a model.

Unsupervised learning An analysis in which one attempts to learn something about the data other than predicting an output value of interest (e.g., whether it falls into clusters).

Validation data (or **Validation set**) The portion of the data used to assess how well the model fits, to adjust some models, and to select the best model from among those that have been tried.

Variable Any measurement on the records, including both the input (X) variables and the output (Y) variable.

1.8 ROAD MAPS TO THIS BOOK

The book covers many of the widely used predictive and classification methods as well as other machine learning tools. Figure 1.2 outlines machine learning from a process perspective and where the topics in this book fit in. Chapter numbers are indicated beside the topic. Table 1.1 provides a different perspective: it organizes supervised and unsupervised machine learning procedures according to the type and structure of the data.

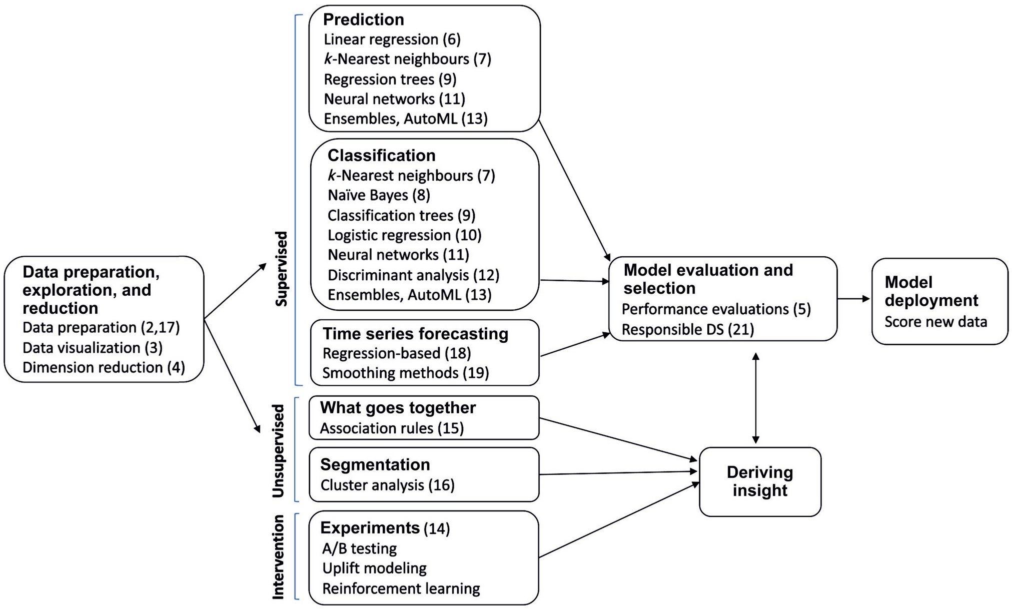

FIGURE 1.2 Machine learning from a process perspective. Numbers in parentheses indicate chapter numbers

TABLE 1.1 Organization of machine learning methods in this book, According to the nature of the data[a]

	Supervised		Unsupervised
	Continuous response	**Categorical response**	**No response**
Continuous predictors	Linear regression (6)	Logistic regression (10)	Principal components (4)
	Neural nets (11)	Neural nets (11)	Cluster analysis (16)
	k-Nearest neighbors (7)	Discriminant analysis (12)	Collaborative filtering (15)
	Ensembles (13)	*k*-Nearest neighbors (7)	
		Ensembles (13)	
Categorical predictors	Linear regression (6)	Neural nets (11)	Association rules (15)
	Neural nets (11)	Classification trees (9)	Collaborative filtering (15)
	Regression trees (9)	Logistic regression (10)	
	Ensembles (13)	Naive Bayes (8)	
		Ensembles (13)	

[a]Numbers in parentheses indicate the chapter number.

Order of Topics

The book is divided into nine parts: Part I (Chapters 1 and 2) gives a general overview of machine learning and its components. Part II (Chapters 3 and 4) focuses on the early stages of data exploration and dimension reduction.

Part III (Chapter 5) discusses performance evaluation. Although it contains only one chapter, we discuss a variety of topics, from predictive performance metrics to misclassification costs. The principles covered in this part are crucial for the proper evaluation and comparison of supervised learning methods.

Part IV includes eight chapters (Chapters 6–13), covering a variety of popular supervised learning methods (for classification and/or prediction). Within this part, the topics are generally organized according to the level of sophistication of the algorithms, their popularity, and ease of understanding. The final chapter introduces ensembles and combinations of methods.

Part V (Chapter 14) introduces the notions of experiments, intervention, and user feedback. This single chapter starts with A/B testing, then its use in uplift modeling, and finally expands into reinforcement learning, explaining the basic ideas and formulations that utilize user feedback for learning best treatment assignments.

Part VI focuses on unsupervised learning of relationships. It presents association rules and collaborative filtering (Chapter 15) and cluster analysis (Chapter 16).

Part VII includes three chapters (Chapters 17–19), with the focus on forecasting time series. The first chapter covers general issues related to handling and understanding time series. The next two chapters present two popular forecasting approaches: regression-based forecasting and smoothing methods.

Part VIII presents the data analytics topic called text mining (Chapter 20). This method applies machine learning methods to text data. The final chapter on responsible data science (Chapter 21) introduces key issues to consider for when carrying out a machine learning project in a responsible way. Finally, Part IX includes a set of cases.

Although the topics in the book can be covered in the order of the chapters, each chapter stands alone. We advise, however, to read Parts I–III before proceeding to chapters in Parts IV–VI. Similarly, Chapter 17 should precede other chapters in Part VII.

USING JMP AND JMP Pro

To facilitate the learning experience, this book uses JMP Pro (Figure 1.3), a desktop statistical package for Mac OS and Windows operating systems from JMP Statistical Discovery LLC (see `jmp.com/system` for complete system requirements).

JMP comes in two primary flavors: JMP (the standard version) and JMP Pro (the professional version). The standard version of JMP is dynamic and interactive, and it offers a variety of built-in tools for graphing and analyzing data. JMP has extensive tools for data visualization and data preparation, along with statistical and machine learning techniques for classification, prediction, and forecasting. It offers a variety of supervised machine learning tools, including neural nets, classification and regression trees, linear and logistic regression, and discriminant analysis. It also offers unsupervised algorithms, such as principal component analysis, *k*-means clustering, and hierarchical clustering.

JMP Pro has all of the functionality of JMP, but adds advanced tools for predictive modeling, including K-Nearest neighbor, penalized regression, naive Bayes, text analytics, association analysis, uplift modeling, advanced trees, and additional options for creating neural networks and ensemble models. Importantly, it also provides a number of modeling utilities for preparing data for modeling and includes built-in tools for data partitioning and model validation, model comparison, model selection, and generating model deployment code.

For these reasons, we use JMP Pro throughout this book. JMP Pro is required for the built-in tools for model cross-validation and comparison that are not available in the standard version of JMP. However, the standard version of JMP can be used for creating many of the graphs, summaries, analyses, and models presented in this book.

While we provide JMP instructions throughout this book, there are many resources available for new users. For tips on getting started with JMP, go to `jmp.com/gettingstarted`. An in-depth introduction to JMP, *Discovering JMP*, is available online or in JMP (under *Help > JMP Documentation Library*). Additional resources on specific topics, along with short videos, can be found in the JMP Learning Library at `jmp.com/learn`. For additional details on the features available only in JMP Pro, see `jmp.com/pro`.

JMP Pro is available through academic licenses at most colleges and universities and through site licenses in many organizations—see your software IT administrator for availability and download information. For academic licensing information, or to request an evaluation of JMP Pro, write to *academic@jmp.com*. If you do not qualify for an academic license, an evaluation copy of JMP Pro can be requested at `jmp.com/proeval`.

JMP runs natively on both the Windows and Mac OS. On Windows, the JMP Home window (shown top, Figure 1.3) and a Tip of the Day window appear when you open JMP. On Mac, a JMP Starter window (Figure 1.4) appears instead of the JMP Home window. These windows can be accessed from the *View* and *Windows* menus in JMP. Use the *File > New* menu to open a new data table (bottom Figure 1.3), or *File > Open* to open an existing JMP data table or data in another file type.

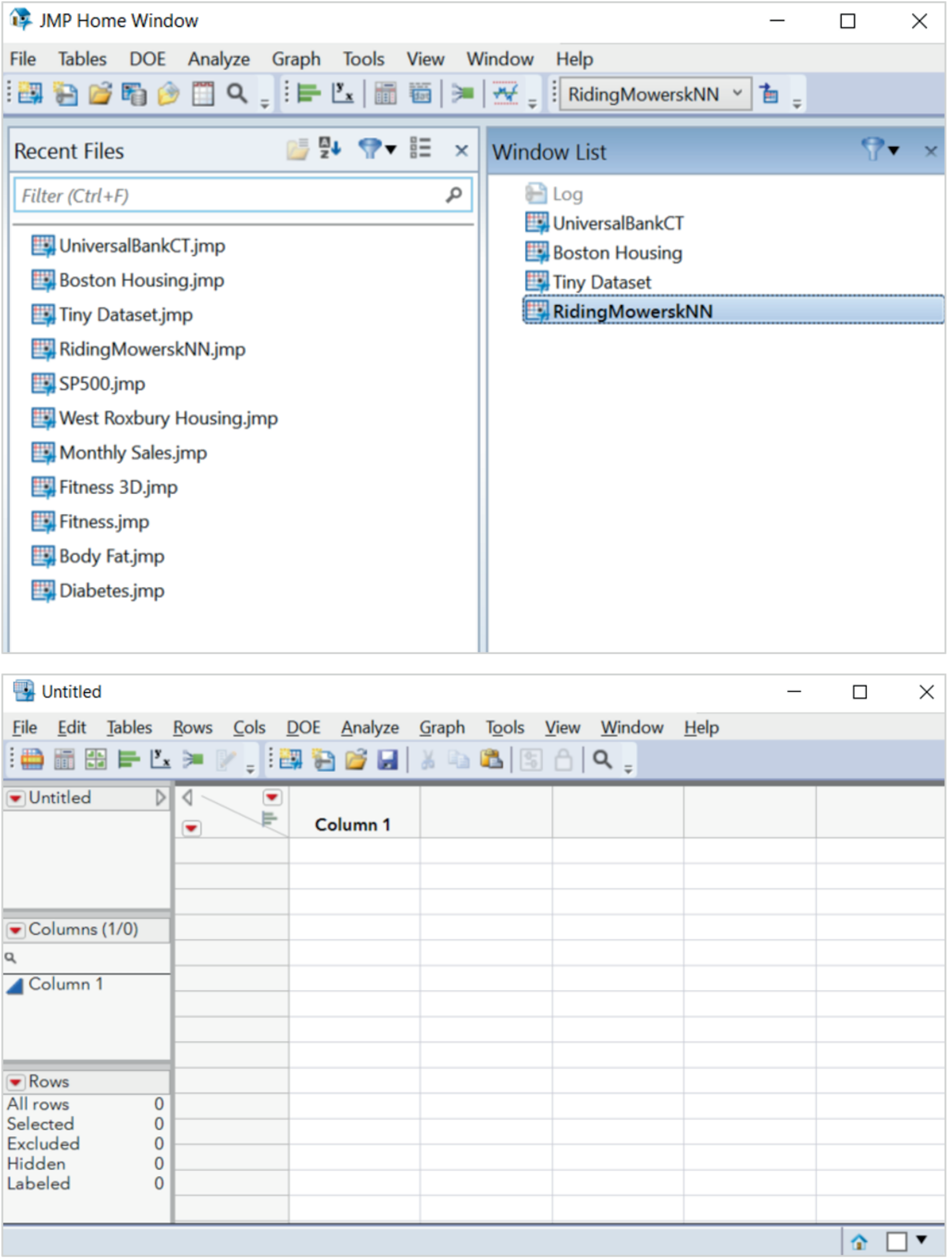

FIGURE 1.3 JMP home window (top) and data table (bottom) (on a Windows Pc)

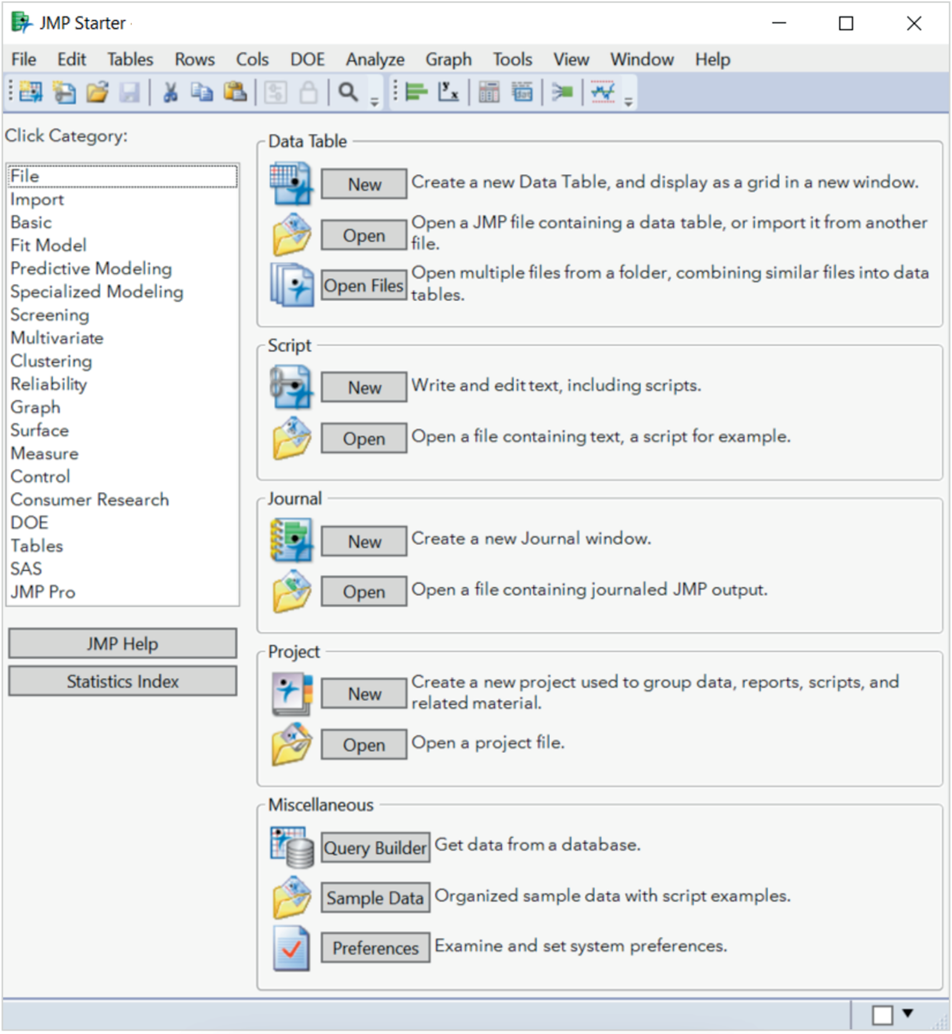

FIGURE 1.4 JMP starter window (on a Windows Pc)