

**Introduction to IBM SPSS  
Modeler and Data Mining  
(v16)**  
Student Guide  
**Course Code: 0A005**

Introduction to IBM SPSS Modeler and Data Mining (v16)

0A005

ERC: 1.0

Published April 2014

All files and material for this course, 0A005 Introduction to IBM SPSS Modeler and Data Mining (v16), are IBM copyright property covered by the following copyright notice.

© Copyright IBM Corp. 2010, 2014

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM corp.

IBM, the IBM logo, ibm.com, Cognos and SPSS are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).

Adobe, and the Adobe logo, are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

P-2

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Contents

|   |            |
|---|------------|
| <b>PREFACE.....</b>                           | <b>P-1</b> |
| CONTENTS.....                                 | P-3        |
| COURSE OVERVIEW .....                         | P-9        |
| DOCUMENT CONVENTIONS .....                    | P-11       |
| WORKSHOPS .....                               | P-12       |
| ADDITIONAL TRAINING RESOURCES.....            | P-13       |
| IBM PRODUCT HELP.....                         | P-14       |
| <b>INTRODUCTION TO DATA MINING .....</b>      | <b>1-1</b> |
| OBJECTIVES .....                              | 1-3        |
| DATA-MINING APPLICATIONS .....                | 1-4        |
| A STRATEGY FOR DATA MINING: CRISP-DM .....    | 1-6        |
| STAGES AND TASKS IN CRISP-DM .....            | 1-7        |
| STAGE 1: BUSINESS UNDERSTANDING.....          | 1-8        |
| STAGE 2: DATA UNDERSTANDING .....             | 1-11       |
| STAGE 3: DATA PREPARATION.....                | 1-14       |
| STAGE 4: MODELING.....                        | 1-17       |
| STAGE 5: EVALUATION .....                     | 1-19       |
| STAGE 6: DEPLOYMENT .....                     | 1-21       |
| THE LIFE CYCLE OF A DATA-MINING PROJECT ..... | 1-23       |
| DATA-MINING SUCCESS .....                     | 1-24       |
| DATA-MINING FAILURE.....                      | 1-28       |
| SKILLS NEEDED FOR DATA MINING .....           | 1-32       |
| APPLY YOUR KNOWLEDGE .....                    | 1-36       |
| SUMMARY .....                                 | 1-39       |
| WORKSHOP 1: INTRODUCTION TO DATA MINING ..... | 1-40       |
| <b>WORKING WITH MODELER .....</b>             | <b>2-1</b> |
| OBJECTIVES .....                              | 2-3        |
| INTRODUCING NODES AND STREAMS .....           | 2-4        |
| EXPLORE THE USER-INTERFACE.....               | 2-5        |
| EXPLORE THE USER-INTERFACE: PALETTES .....    | 2-6        |
| EXPLORE THE USER-INTERFACE: PANES .....       | 2-7        |
| CREATING STREAMS: GENERAL RULES .....         | 2-8        |
| CREATING STREAMS: USING THE MOUSE.....        | 2-9        |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

P-3

|   |            |
|---|------------|
| CREATING STREAMS: PLACING NODES.....                | 2-10       |
| CREATING STREAMS: MANAGING NODES .....              | 2-11       |
| CREATING STREAMS: MANAGING CONNECTIONS .....        | 2-12       |
| ENCAPSULATING NODES IN A SUPERNODE.....             | 2-13       |
| GENERATING NODES FROM OUTPUT .....                  | 2-14       |
| RUNNING STREAMS.....                                | 2-15       |
| ONLINE HELP .....                                   | 2-16       |
| DEMO 1: WORKING WITH MODELER.....                   | 2-17       |
| APPLY YOUR KNOWLEDGE .....                          | 2-25       |
| SUMMARY .....                                       | 2-31       |
| WORKSHOP 1: WORKING WITH MODELER.....               | 2-32       |
| <b>A DATA-MINING TOUR .....</b>                     | <b>3-1</b> |
| OBJECTIVES .....                                    | 3-3        |
| THE BASIC FRAMEWORK OF A DATA-MINING PROJECT .....  | 3-4        |
| A BUSINESS CASE .....                               | 3-5        |
| A BUSINESS CASE: HISTORICAL DATA .....              | 3-6        |
| A BUSINESS CASE: INITIAL RESULTS.....               | 3-7        |
| A BUSINESS CASE: INITIAL RESULTS IN A TREE .....    | 3-8        |
| A BUSINESS CASE: FURTHER RESULTS.....               | 3-9        |
| A BUSINESS CASE: FURTHER RESULTS IN A TREE .....    | 3-10       |
| A BUSINESS CASE: A PREDICTIVE MODEL.....            | 3-11       |
| DEPLOYING THE MODEL: SCORING RECORDS .....          | 3-13       |
| A DATA-MINING PROJECT IN MODELER .....              | 3-14       |
| BUILD THE MODEL: OVERVIEW .....                     | 3-15       |
| BUILD THE MODEL: SETTING ROLES IN A TYPE NODE ..... | 3-16       |
| BUILD THE MODEL – SCORE RECORDS .....               | 3-17       |
| TWO HANDY NODES: FILTER AND SORT .....              | 3-18       |
| DEMO 1: A DATA-MINING TOUR .....                    | 3-19       |
| APPLY YOUR KNOWLEDGE.....                           | 3-31       |
| SUMMARY .....                                       | 3-34       |
| WORKSHOP 1: A DATA-MINING TOUR .....                | 3-35       |

|  |            |
|--|------------|
| <b>COLLECTING INITITAL DATA .....</b>                        | <b>4-1</b> |
| OBJECTIVES .....   | 4-3        |
| RECTANGULAR DATA STRUCTURE .....                             | 4-4        |
| THE UNIT OF ANALYSIS .....                                   | 4-5        |
| FIELD STORAGES.....  | 4-6        |
| FIELD MEASUREMENT LEVELS .....                               | 4-7        |
| STORAGE AND FIELD MEASUREMENT LEVEL ILLUSTRATED .....        | 4-8        |
| STORAGE AND MEASUREMENT LEVEL .....                          | 4-9        |
| FIELD INSTANTIATION .....                                    | 4-10       |
| FIELD INSTANTIATION ILLUSTRATED .....                        | 4-11       |
| IMPORTING DATA: THE SOURCES PALETTE.....                     | 4-13       |
| EXPLORE THE SOURCES DIALOG BOXES: THE DATA TAB .....         | 4-15       |
| IMPORTING TEXT FILES .....                                   | 4-17       |
| EXPORTING DATA: THE EXPORT PALETTE .....                     | 4-18       |
| DEMO 1: COLLECTING INITIAL DATA .....                        | 4-19       |
| APPLY YOUR KNOWLEDGE .....                                   | 4-27       |
| SUMMARY .....  | 4-36       |
| WORKSHOP 1: COLLECTING INITIAL DATA .....                    | 4-37       |
| <b>UNDERSTANDING YOUR DATA .....</b>                         | <b>5-1</b> |
| OBJECTIVES .....   | 5-3        |
| DATA AUDIT ILLUSTRATED .....                                 | 5-4        |
| EXPLORE THE DATA AUDIT DIALOG BOX .....                      | 5-6        |
| USE THE STATISTICS NODE AND GRAPHS NODES FOR REPORTING ..... | 5-7        |
| DESCRIBE TYPES OF INVALID VALUES .....                       | 5-8        |
| INVALID VALUES ILLUSTRATED .....                             | 5-9        |
| ACTIONS FOR INVALID VALUES.....                              | 5-10       |
| ACTIONS FOR INVALID VALUES ILLUSTRATED.....                  | 5-11       |
| DEALING WITH MISSING DATA: BLANKS .....                      | 5-12       |
| BLANKS AND ACTIONS ILLUSTRATED .....                         | 5-14       |
| EXPLORE THE TYPE DIALOG BOX.....                             | 5-15       |
| REPORTING BLANKS IN A DATA AUDIT .....                       | 5-16       |
| BLANK VALUES IN ANALYSES .....                               | 5-17       |
| DEMO 1: UNDERSTANDING YOUR DATA .....                        | 5-18       |
| APPLY YOUR KNOWLEDGE .....                                   | 5-27       |
| SUMMARY .....  | 5-31       |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

P-5

|  |             |
|--|-------------|
| <b>WORKSHOP 1: UNDERSTANDING YOUR DATA .....</b>                 | <b>5-32</b> |
| <b>SETTING THE UNIT OF ANALYSIS.....</b>                         | <b>6-1</b>  |
| OBJECTIVES .....   | 6-3         |
| THE REQUIRED UNIT OF ANALYSIS .....                              | 6-4         |
| METHODS TO CREATE DATASETS WITH THE REQUIRED UNIT OF ANALYSIS... | 6-5         |
| DISTINCTING RECORDS.....   | 6-8         |
| EXPLORE THE DISTINCT DIALOG BOX.....                             | 6-9         |
| AGGREGATING RECORDS.....   | 6-11        |
| EXPLORE THE AGGREGATE DIALOG BOX .....                           | 6-12        |
| SETTING TO FLAG FIELDS .....                                     | 6-14        |
| EXPLORE THE SETTOFLAG DIALOG BOX .....                           | 6-15        |
| APPLY YOUR KNOWLEDGE .....                                       | 6-23        |
| SUMMARY .....  | 6-28        |
| WORKSHOP 1: SETTING THE UNIT OF ANALYSIS .....                   | 6-29        |
| <b>INTEGRATING DATA.....</b>                                     | <b>7-1</b>  |
| OBJECTIVES .....   | 7-3         |
| METHODS TO INTEGRATE DATA .....                                  | 7-4         |
| APPENDING RECORDS .....  | 7-5         |
| OPTIONS TO APPEND RECORDS.....                                   | 7-6         |
| APPENDING RECORDS ILLUSTRATED .....                              | 7-7         |
| EXPLORE THE APPEND DIALOG BOX .....                              | 7-8         |
| MERGING FIELDS .....   | 7-9         |
| OPTIONS TO MERGE FIELDS.....                                     | 7-10        |
| MERGING FIELDS ILLUSTRATED .....                                 | 7-11        |
| A ONE-TO-MANY MERGE .....  | 7-13        |
| EXPLORE THE MERGE DIALOG BOX .....                               | 7-14        |
| SAMPLING RECORDS .....   | 7-15        |
| EXPLORE THE SAMPLE DIALOG BOX .....                              | 7-16        |
| CACHING DATA.....  | 7-17        |
| DEMO 1: INTEGRATING DATA.....                                    | 7-18        |
| APPLY YOUR KNOWLEDGE.....  | 7-31        |
| SUMMARY .....  | 7-36        |
| WORKSHOP 1: INTEGRATING DATA .....                               | 7-37        |

|  |            |
|--|------------|
| <b>DERIVING AND RECLASSIFYING FIELDS .....</b>                               | <b>8-1</b> |
| OBJECTIVES .....   | 8-3        |
| METHODS TO CREATE FIELDS.....  | 8-4        |
| INTRODUCING THE CONTROL LANGUAGE FOR EXPRESSION MANIPULATION<br>(CLEM) ..... | 8-5        |
| INTRODUCING CLEM EXPRESSIONS .....   | 8-6        |
| EXPLORE THE EXPRESSION BUILDER DIALOG BOX.....                               | 8-8        |
| DERIVING FIELDS.....   | 8-9        |
| EXPLORE THE DERIVE DIALOG BOX.....   | 8-10       |
| DERIVING FIELDS AND BLANKS.....  | 8-11       |
| RECLASSIFYING FIELDS .....   | 8-12       |
| EXPLORE THE RECLASSIFY DIALOG BOX .....                                      | 8-13       |
| ADD FIELDS TO THE SAME BRANCH .....  | 8-14       |
| CHECKING YOUR RESULTS.....   | 8-15       |
| DEMO 1: DERIVING AND RECLASSIFYING FIELDS.....                               | 8-16       |
| APPLY YOUR KNOWLEDGE .....   | 8-30       |
| SUMMARY .....  | 8-34       |
| WORKSHOP 1: DERIVING AND RECLASSIFYING FIELDS .....                          | 8-35       |
| <b>LOOKING FOR RELATIONSHIPS.....</b>  | <b>9-1</b> |
| OBJECTIVES .....   | 9-3        |
| METHODS TO EXAMINE THE RELATIONSHIP BETWEEN TWO FIELDS .....                 | 9-4        |
| EXPLORE MATRIX OUTPUT.....   | 9-5        |
| EXPLORE THE MATRIX DIALOG BOX.....   | 9-6        |
| EXPLORE DISTRIBUTION OUTPUT .....  | 9-7        |
| EXPLORE THE DISTRIBUTION DIALOG BOX .....                                    | 9-8        |
| EXPLORE MEANS OUTPUT .....   | 9-9        |
| EXPLORE THE MEANS DIALOG BOX .....   | 9-10       |
| EXPLORE HISTOGRAM OUTPUT .....   | 9-11       |
| EXPLORE STATISTICS OUTPUT.....   | 9-12       |
| EXPLORE THE STATISTICS DIALOG BOX .....                                      | 9-13       |
| EXPLORE PLOT OUTPUT.....   | 9-14       |
| DEMO 1: LOOKING FOR RELATIONSHIPS.....                                       | 9-15       |
| APPLY YOUR KNOWLEDGE .....   | 9-23       |
| SUMMARY .....  | 9-29       |
| WORKSHOP 1: LOOKING FOR RELATIONSHIPS .....                                  | 9-30       |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

P-7

|  |             |
|--|-------------|
| <b>INTRODUCTION TO MODELING .....</b>              | <b>10-1</b> |
| OBJECTIVES .....                                   | 10-3        |
| MODELING OBJECTIVES .....                          | 10-4        |
| MODELING OBJECTIVES AND THE MODELING PALETTE ..... | 10-6        |
| OBJECTIVES AND ROLES IN THE TYPE NODE .....        | 10-7        |
| TYPES OF CLASSIFICATION MODELS .....               | 10-8        |
| RULE INDUCTION MODELS .....                        | 10-9        |
| TRADITIONAL STATISTICAL MODELS .....               | 10-10       |
| MACHINE LEARNING MODELS .....                      | 10-11       |
| WHICH CLASSIFICATION MODEL TO USE? .....           | 10-12       |
| RUNNING CLASSIFICATION MODELS .....                | 10-14       |
| MODELING RESULTS: THE MODEL NUGGET .....           | 10-15       |
| EVALUATING CLASSIFICATION MODELS .....             | 10-16       |
| APPLYING CLASSIFICATION MODELS .....               | 10-17       |
| SEGMENTATION MODELS.....                           | 10-18       |
| RUNNING SEGMENTATION MODELS .....                  | 10-20       |
| EXAMINING THE RESULTS: CLUSTER PROFILES.....       | 10-21       |
| DEMO 1: INTRODUCTION TO MODELING.....              | 10-22       |
| APPLY YOUR KNOWLEDGE .....                         | 10-31       |
| SUMMARY .....                                      | 10-35       |
| WORKSHOP 1: INTRODUCTION TO MODELING .....         | 10-37       |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Course Overview

### Course Overview

Introduction to **IBM SPSS Modeler and Data Mining (v16)** is a two day course that provides an overview of data mining and the fundamentals of using IBM SPSS Modeler. The principles and practice of data mining are illustrated using the CRISP-DM methodology. The course structure follows the stages of a typical data mining project, from collecting data, to data exploration, data transformation, and modeling to effective interpretation of the results. The course provides training in the basics of how to read, prepare, and explore data with IBM SPSS Modeler, and introduces the student to modeling.

### Intended Audience

IBM SPSS Modeler Analysts and IBM SPSS Modeler Data Experts who want to become familiar with IBM SPSS Modeler.

Specifically, this is an introductory course for:

- anyone who is new to IBM SPSS Modeler
- anyone considering purchasing IBM SPSS Modeler
- anyone interested in Data Mining

## Topics Covered

- Topics covered in this course include:
- Introduction to Data Mining
- Working with Modeler
- A Data-Mining Tour
- Collecting Initial Data
- Understanding your Data
- Setting the Unit of Analysis
- Integrating Data
- Deriving and Reclassifying Fields
- Looking for Relationships
- Introduction to Modeling

## Course Prerequisites

Participants should have:

- General computer literacy

## Document Conventions

Conventions used in this guide follow Microsoft Windows application standards, where applicable. As well, the following conventions are observed:

### **Bold**

Bold style is used in demo and workshop step-by-step solutions to indicate either:

- actionable items

(Point to **Sort**, and then click **Ascending.**)

- text to type or keys to press

(Type **Sales Report**, and then press **Enter.**)

- UI elements that are the focus of attention

(In the **Format** pane, click **Data**)

### *Italic*

Used to reference book titles.

### CAPITALIZATION

All file names, table names, column names, and folder names appear in this guide exactly as they appear in the application.

To keep capitalization consistent with this guide, type text exactly as shown.

# Workshops

## Workshop Format

Workshops are designed to allow you to work according to your own pace. Content contained in a workshop is not fully scripted out to provide an additional challenge. Refer back to demonstrations if you need assistance with a particular task. The workshops are structured as follows:

### The Business Question Section

This section presents a business-type question followed by a series of tasks. These tasks provide additional information to help guide you through the workshop. Within each task, there may be numbered questions relating to the task. Complete the tasks by using the skills you learned in the module. If you need more assistance, you can refer to the Task and Results section for more detailed instruction.

### The Task and Results Section

This section provides a task based set of instructions that presents the question as a series of numbered tasks to be accomplished. The information in the tasks expands on the business case, providing more details on how to accomplish a task. Screen captures are also provided at the end of some tasks and at the end of the workshop to show the expected results.

## Additional Training Resources

Bookmark [Business Analytics Product Training](#)

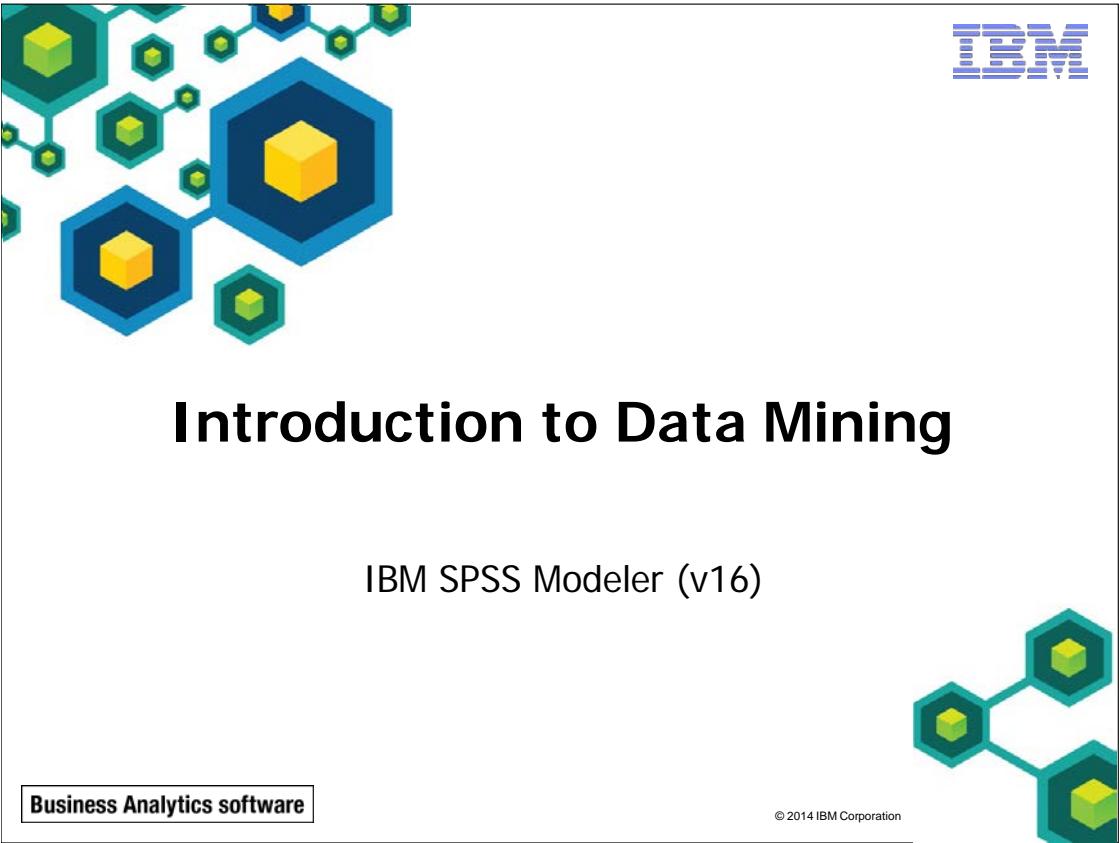
<http://www-01.ibm.com/software/analytics/training-and-certification/> for details on:

- instructor-led training in a classroom or online
- self-paced training that fits your needs and schedule
- comprehensive curricula and training paths that help you identify the courses that are right for you
- IBM Business Analytics Certification program
- other resources that will enhance your success with IBM Business Analytics Software

# IBM Product Help

| Help type                 | When to use  | Location  |
|---------------------------|--|---|
| Task-oriented             | You are working in the product and you need specific task-oriented help.   | <i>IBM Product - Help link</i>  |
| Books for Printing (.pdf) | <p>You want to use search engines to find information. You can then print out selected pages, a section, or the whole book.</p> <p>Use Step-by-Step online books (.pdf) if you want to know how to complete a task but prefer to read about it in a book.</p> <p>The Step-by-Step online books contain the same information as the online help, but the method of presentation is different.</p> | Start/Programs/ <i>IBM Product/Documentation</i>  |
| IBM on the Web            | <p>You want to access any of the following:</p> <ul style="list-style-type: none"> <li>• Training and Certification Web site</li> <li>• Online support</li> <li>• IBM Web site</li> </ul>  | <ul style="list-style-type: none"> <li>• <a href="http://www-01.ibm.com/software/analytics/training-and-certification/">http://www-01.ibm.com/software/analytics/training-and-certification/</a></li> <li>• <a href="http://www-947.ibm.com/support/entry/portal/Overview/Software">http://www-947.ibm.com/support/entry/portal/Overview/Software</a></li> <li>• <a href="http://www.ibm.com">http://www.ibm.com</a></li> </ul> |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



The graphic features a white background with a decorative pattern of blue hexagons containing yellow cubes, some connected by thin teal lines. In the top right corner, the IBM logo is displayed in its signature blue font. Below the graphic, the title "Introduction to Data Mining" is centered in a large, bold, black sans-serif font. Underneath the title, the text "IBM SPSS Modeler (v16)" is centered in a smaller, regular black font. At the bottom left, a small rectangular box contains the text "Business Analytics software". At the bottom right, a copyright notice reads "© 2014 IBM Corporation".

# Introduction to Data Mining

IBM SPSS Modeler (v16)

Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Objectives

- At the end of this module, you should be able to:
  - list two applications of data mining
  - explain the stages of the CRISP-DM process model
  - describe successful data-mining projects and the reasons why projects fail
  - describe the skills needed for data mining

© 2014 IBM Corporation

With increasingly competitive markets and the vast capabilities of computers, many businesses find themselves faced with data and a need to identify useful patterns and actionable relationships, which data mining can be useful for.

A common misconception is that data mining involves passing huge amounts of data through intelligent technologies that alone find patterns and give magical solutions to business problems. Data mining is foremost a process that needs a thorough methodology, which is presented in this module.

No prior knowledge is required for this module.

## Data-Mining Applications (1 of 2)

- Reduce churn (reduce the number of customers who cancel their policies, subscriptions, or accounts)
- Reduce costs by better targeting customers in direct mail campaigns
- Reduce costs in a manufacturing process by preventing machine failures
- Reduce the incidence of a heart attack among those with a cardiac disease

© 2014 IBM Corporation



As an example of a data-mining application, consider a telecommunications firm that is confronted with a huge amount of churners. In a data-mining project, the firm can use modeling techniques on their historical data to find groups of customers with a high churn rate. Next, the firm can apply these models to their current customer database to identify customers at risk. Finally, these customers can be made an interesting offer, so they will hopefully be retained as customers.

Another example of a data-mining application is found in database marketing, where huge volumes of mail are sent out to customers or prospects. Typically, response rates lie around 2%. To cut costs in sending out mail, the database marketing uses their historical data to build models that identify groups with high response rates, so that only these customers will be approached in future campaigns. This will cut mailing costs, while the number of responders (people purchasing the product) will not change significantly. All in all, costs will go down, while revenues stay the same, so the ROI (Return On Investment) will improve.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Data-Mining Applications (2 of 2)

- Better target customers by classifying customers into groups with distinct usage or need patterns
- Reduce costs by preventing fraudulent credit-card activity, or detecting fraud in an earlier stage
- Increase revenues by increasing the number of products sold by cross-selling
- Increase revenues by showing a visitor the best-next-page on a website

© 2014 IBM Corporation



In telecommunications, another common application is to profile customers. For example, by analyzing the customers' usage data (phone calls, text messages, internet usage), a firm can create profiles such as "leaders" and "followers", and approach each group in its own way.

Refer to <http://www-01.ibm.com/software/analytics/spss/products/modeler/downloads.html> for videos and white papers on various data-mining applications.

## A Strategy for Data Mining: CRISP-DM

- A data-mining project can become complicated quickly
- A model is needed that guides you through the critical issues
- Recommendation: use the Cross-Industry Standard Process for Data Mining (CRISP-DM)

© 2014 IBM Corporation

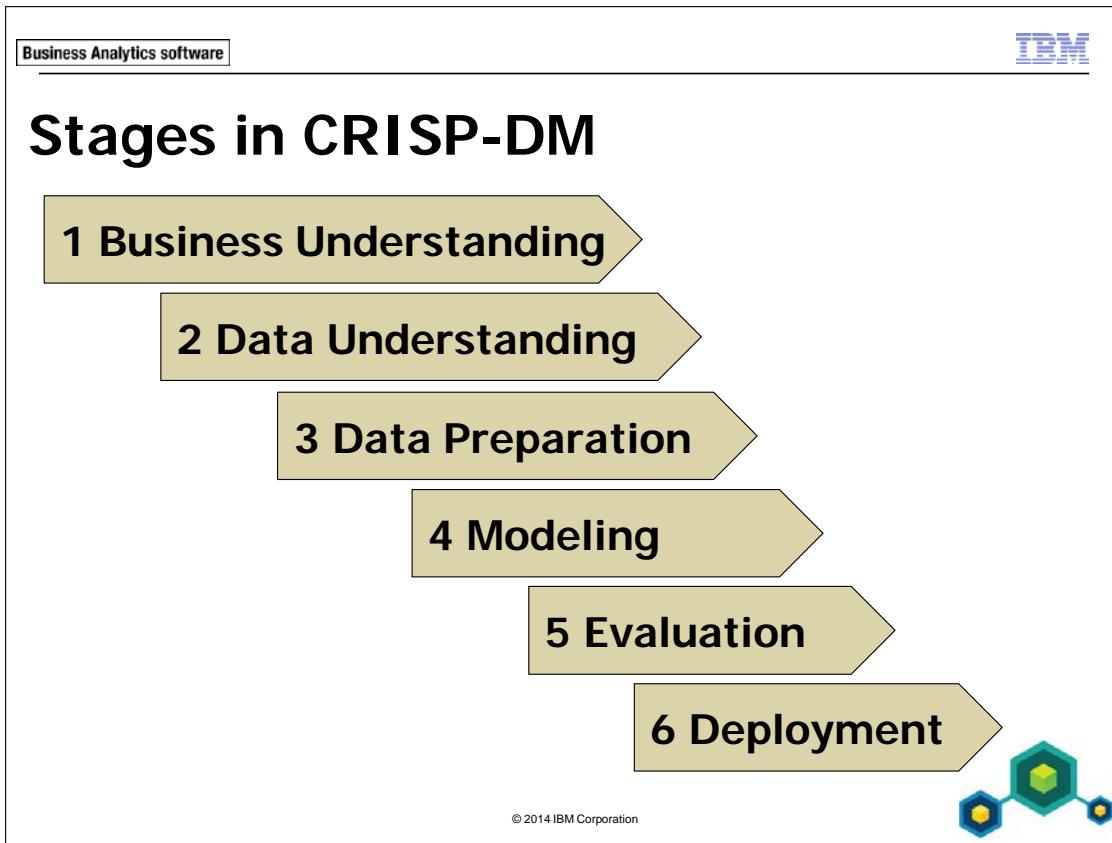


A data-mining project can become complicated very quickly. There is a lot to keep track of: complex business problems, multiple data sources, varying data quality across data sources, an array of data-mining techniques, different ways of measuring data mining success, and so on.

To stay on track, it helps to have an explicitly defined process model for a data-mining project. The process model guides you through the critical issues outlined above and makes sure that the important points are addressed. It serves as a data-mining road map so that you will not lose your way as you dig into the complexities of the data.

The data-mining process model recommended for use with MODELER is the Cross-Industry Standard Process for Data Mining (CRISP-DM). This process model is designed as a general model that can be applied to a wide variety of industries and business problems.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



The general CRISP-DM process model includes six stages that address the main issues in data mining, including how to incorporate data mining into larger business practices.

These stages are listed on this slide, and will be discussed in more detail on the next slides. Also, in the *A Data-Mining Tour* module you will go through the stages, so that you will have a background of what a data-mining project encompasses.

Note: The discussion in this module is not a comprehensive discussion of CRISP-DM. Please refer to MODELER's online Help for an overview of all tasks and sub tasks, or refer to the information readily available on the Web.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software
IBM

## Stage 1: Business Understanding

| Task                             | Sub task 1                   | Sub task 2                                 | Sub task 3                |
|----------------------------------|------------------------------|--|---------------------------|
| Determine business objectives    | Background                   | Business objectives                        | Business success criteria |
| Assess situation                 | Inventory of resources       | Risks and contingencies                    | Terminology               |
| Determine data-mining objectives | Data-mining success criteria |  |                           |
| Produce project plan             | Write a project plan         | Initial assessment of tools and techniques |                           |

© 2014 IBM Corporation



Business Understanding is perhaps the most important phase in a data-mining project. Business objectives and success criteria, resources, constraints, assumptions, risks, costs, and benefits are identified in this stage. Also, specific data-mining goals are set, a project plan is written, and agreed upon.

As an example, consider a telecommunications firm that is confronted with high volumes of churn. The firm can start a project with the objective to reduce churn, and the project could be declared to be a success if churn when reduced by at least 10%.

In this stage, you would ensure that you have all the resources that are needed to complete the project. For example, does the project involve other departments (such as a marketing department that has to make the offer for the customers at risk) or external consultants (with expertise in modeling)?

Also, what are the risks of the project? Reducing churn might be essential to survive in the telecommunications market, so probably there will be a tight deadline. All participants should be informed about the deadline, and about the importance to complete the project on time.

Another risk is the availability of data. Here, important questions are:

- Who are the key persons in accessing the data?
- Will you enrich your data by purchasing demographic data?
- Are there legal restrictions on the use of your data?

When you work together in a project, ensure that the terminology is clear. Define what you mean by "churn"? For example, a customer whose subscription has ended because he did not pay the bill, would you regard him as a chunner? You may want to distinguish between voluntary and involuntary chunners. Or, is it clear to everyone in the team what ARPU means? (ARPU means Average Revenue Per User, and is a popular term in telecommunications firms.)

Translating the business objectives into specific data-mining goals is another task in this stage. A data-mining goal derived from a business objective such as "reduce churn" could be to have a model in place that returns the likelihood of churn. Ask yourself if you want to apply the model to every customer, or if you want to apply it to high-value customers only (preventing churn for low-value customers may cost the company more than letting go of these customers).

Another specific data-mining goal could be that you aim at building a model that identifies 80% of the chunners in 20% of the data.

At this point you may decide to use only a certain class of models. For example, your data-mining goal could be to only use models that give insight in the relationships (which attributes or "fields" are related to churn) and to discard black-box models. This choice will affect the actions that you will take later in the project. For example, if your model tells you that customers with a certain handset show high churn rates, you may offer a new handset to customers with that handset. If you use a black-box model you will not be able to make such a specific offer and you will have to make a more generic offer, such as a discount.

Ensure that you have a project plan listing all tasks and responsibilities. Such a project plan may be written along the lines of the six stages as outlined in the CRISP-DM methodology. What is the time needed to complete each stage? Or: which persons are needed (and/or responsible) in each stage? What are the risks in each stage, and is there a contingency plan?

Last but not least, ensure that you have the tools to complete the project. For example, when you use MODELER, ensure that one or more team members have the skills to use the software. Also ensure that the expiration date of the software lies beyond the end date of the project.

## Stage 2: Data Understanding

| Task                 | Sub task 1                     |
|----------------------|--------------------------------|
| Collect initial data | Initial data-collection report |
| Describe data        | Data-description report        |
| Explore data         | Data-exploration report        |
| Verify data quality  | Data-quality report            |

© 2014 IBM Corporation



Data provides the raw materials of data mining. The stage of Data Understanding addresses the need to understand what your data resources are and the characteristics of those resources. It includes collecting initial data, describing data, exploring data, and verifying data quality.

For example, a telecommunications firm that needs to reduce churn could use the information they have on their customers, such as gender, age, and region. Another dataset could be comprised of call detail records. Also, the firm could use data from the call center. Although this type of data is unstructured, it could well be related to churn (especially if customers call in with a complaint).

For each data source, how many rows (records) and columns (fields) do you have? And, are all records and fields needed to answer your business question? For example, the customer dataset will include the customer's surname, but it is very unlikely that surname is related to churn, so it is better to remove this field. Or, working with call detail records, you will have terabytes of data soon. So, how do want to reduce the amount of data, without losing information?

In your data-description report, explain what the fields in your datasets mean. For example, it may not be obvious for everyone in the project what a field such as HS means, and you might consider renaming fields (for example, HS into HANDSET).

In your data-description report, list the abbreviations that you have encountered in your data, and what they mean. It may be trivial what F and M mean for a field such as gender, but if you have values such as 1, 2, 3 for a field named segment, explain what each code means, to be able to interpret modeling results later.

In your data-description report, also map names of fields in different datasets to each other. In databases, for example, a field such as KEY\_ID in the customer table could be named CUSTOMER\_ID in the product table.

In your data-exploration report, include graphs of the fields of interest, so that you can visually inspect your data. Some modeling techniques work best when fields follow a certain distribution, and you might consider transform fields to conform to that distribution.

In your data-quality report, report inconsistencies or errors in the data. For example, suppose a field such as gender shows values F, FEMALE, Female, or a field such as age shows a value -1. These instances should be reported in the data-exploration report, and you should investigate why these values are in your data. For example, it is not uncommon that a database administrator plugs in a value -1 when a value is unknown.

Report the amount of missing data in your data-quality report, and decide what you want to do with records or fields that have many missing values. Again, investigate the reason why data is missing. For example, a field such as `END_DATE_OF_SUBSCRIPTION` is missing by definition for all current customers, so it is necessary to retain this field, in order to derive a field such as `HAS_CHURNED` from it later.

Two modules in this course relate to this stage of a data-mining project. The *Collecting Initial Data* module presents how you can read data into Modeler. In the *Understanding your Data* module you will be introduced to methods to explore your data, to assess the quality of your data, and to take action on it.

## Stage 3: Data Preparation

| Task                             | Sub task 1                           | Sub task 2     |
|----------------------------------|--------------------------------------|----------------|
| Select data                      | Rational for inclusion and exclusion |                |
| Clean data                       | Data-cleaning report                 |                |
| Construct data                   | Derived attributes                   |                |
| Format data and combine datasets | Set the unit of analysis             | Integrate data |

© 2014 IBM Corporation



After cataloging the data resources you will prepare the data for mining. Preparations include selecting, cleaning, constructing, formatting and integrating data. These tasks can be very time consuming but are critical for the success of the data-mining project.

Think of a telecommunications firm that has initiated a project to reduce churn. Fields such as gender and age are valuable because they can predict churn. But you may also have fields in your dataset that are derived from the churn field itself. Suppose, for example, that churners received a letter of thanks, and so a field such as HAS RECEIVED LETTER may be included in your dataset. This field is derived from churn so it can never be a predictor (if you include this field in a model, you will find that it is perfectly related to churn). This example may sound trivial, but when you work with a database with 34 tables, and hundreds of fields, it may not be so easy to distinguish between fields that are candidate predictors, and fields that are a consequence of the field that you want to predict. It is in this stage of data preparation that you will systematically exclude fields in your dataset.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Apart from removing fields from your dataset, you may want to remove records. For example, if you are only interested in preventing churn of high-value customers, you will not include low-value customers in the dataset on which you build your models. So this group should be removed from the dataset.

In your data-cleaning report, list all the actions you have taken in the process of cleaning the data. For example, you may have reclassified all values F, FEMALE, and Female into a single Female category. Or you may have replaced missing data with the mean value of a field.

In this stage, new fields may be derived. Suppose that you have a field DATE\_OF\_BIRTH, but for the interpretation of the results you prefer a field AGE. You will need to derive the latter field from the first field. Also, consider deriving fields by taking differences or ratios. Think of two customers of a telecommunications firm. One customer has phoned for 5 minutes, and has sent 10 text messages. A second customer has phoned for 50 minutes, and has sent 100 text messages. Although different in absolute value, their patterns of phoning and texting are identical, and it may be that the pattern is related to churn.

In formatting data, think of restructuring data into a form that the analysis requires. For example, a telecommunications firm can have a dataset of call detail records, where every record represents a call. So, if one customer made 4 calls, and another customer made 481 calls, the first customer will have 4 records in the dataset, and the second customer has 481 records. Now suppose that another dataset stores customers, with their gender, age, and a field flagging whether the customer churned. When building a model for churn, we should count a customer's gender, age, and churn only once, and not as many times as he has call detail records. When you want to build models to predict churn, each record should represent a unique customer, so the call detail dataset should be transformed into a dataset where the unit of analysis is a customer.

When the unit of analysis is set for all datasets, the datasets can be integrated, that is, combined into a single dataset.

Three modules in this course relate to this stage of a data-mining project. The *Setting the Unit of Analysis* module presents various ways to reformat your data. The *Integrating Data* module presents methods to combine your datasets. The *Deriving and Reclassifying Fields* module focuses on two methods to cleanse and enrich your data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-16

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Stage 4: Modeling

| Task                       | Sub task 1           | Sub task 2              |
|----------------------------|----------------------|-------------------------|
| Select modeling techniques | Modeling assumptions |                         |
| Generate test design       | Test design          |                         |
| Build model                | Set model parameters | Model descriptions      |
| Assess model               | Model assessment     | Revise model parameters |

© 2014 IBM Corporation



Modeling is the part of data mining where sophisticated analysis methods are used to extract information from the data. This stage involves selecting modeling techniques, generating test designs, and building and assessing models.

In the modeling stage you will probably not consider one model only, but several models. To test the models, it is common practice to apply the model to a test dataset, a dataset that was not used at the time you built the model. To do this you will have to partition your data in a training dataset (on which you will build your models), and a test dataset (on which you will test your models). What then is the percentage of your data that you will assign to each part? Although it is common to use a 70/30 split for training-test, this may leave too few records to build your model, and you may consider boosting your data (duplicating records). These are methods which are not standard in traditional statistical modeling, but which are widely accepted in data mining.

When you build your model you can start with a model's default settings, and fine tune the parameters later. Report which parameters you changed, and how it affected the results, not only in terms of accuracy, but also in execution time. For example, model A may complete within seconds, whereas model B may take hours. Would you then prefer model A, or will you run model B in batch, at night? And if you choose the latter, what are the consequences, for example do you need a colleague to schedule the job for you?

You should also describe the model results in terms of how the model deals with missing data. Some models don't have issues with missing data, while other models delete records with missing data. Decide on which model you prefer. If you use a model that discards records with missing data, how does that affect deploying the model later? If the new dataset to which the model needs to be applied also has a high percentage of missing data, the model cannot be applied to a significant part of that dataset.

Given your findings, rank the models that you have tested according to criteria such as model accuracy, ease of use, interpretation of the results, and ease of deployment. You may also want to rerun models, adjusting model parameters. Or you may select a model that you did not consider before. For example, you did not run black-box models but given the disappointing results of the models that you have examined you would like to try these black-box models. Of course, the question then is how this relates to your data-mining goals, and if those should be revised.

Two modules in this course relate to this stage of a data-mining project. The *Looking for Relationships* module is a first step in exploring relationships in your data. The *Introduction to Modeling* module gives an overview of data mining models, and discusses two of them.

## Stage 5: Evaluation

| Task                 | Sub task 1  | Sub task 2     |
|----------------------|---|----------------|
| Evaluate results     | Assessment of data-mining results with respect to business success criteria | Approve models |
| Review process       | Review of process   |                |
| Determine next steps | List of possible actions  | Decision       |

© 2014 IBM Corporation



In this stage you have built one or more models that appear to have high quality from a data analysis perspective. You now should evaluate how the data-mining results can help to achieve your business objectives.

Continuing with the example of the telecommunications firm, suppose that two candidate models were found. The first model tells you that adolescent men with handset A show the highest churn percentage, and a second model returns the likelihood to churn, but does not give further insight. Suppose that the first model has a lower accuracy than the second model (the first model cannot predict churn as good as the second model). When you apply the first model to your current customers, you can target adolescent men with handset A, and make them an interesting offer for a new handset. When you apply the second model to your current customers, you will use a general strategy, such as offering a discount for those at risk. This is the point to choose one of the models, or maybe to go ahead with both.

Also, review the whole process so far. For example, was the project plan, with the tasks, responsibilities and deadlines for each stage met? If not, what were the reasons for the delay?

Finally, determine the next steps. In the worst case scenario you may have to conclude that your results are unsatisfactory, and you will go back to an earlier stage. For example, it may be that the accuracy of your model is too low, and so you consider to bring in unstructured data from the call center (data that was not used yet in the project). If you decide to use unstructured data, you will probably need specialized software (such as IBM SPSS Modeler Premium - Text Analytics) to analyze this data and you will need someone who has the skills (including business knowledge) to run the analyses.

All in all, you can iterate through the previous stages, and come to a point where you are confident enough to deploy one or more models.

Refer to the *Introduction to Modeling* module for more information on how to measure a model's accuracy and how to compare models.

## Stage 6: Deployment

| Task                 | Sub task 1       | Sub task 2         |
|----------------------|------------------|--------------------|
| Plan deployment      | Deployment plan  |                    |
| Maintenance          | Maintenance plan |                    |
| Produce final report | Final report     | Final presentation |
| Review project       | Documentation    |                    |

© 2014 IBM Corporation



Now that you have invested all of this effort, it is time to reap the benefits. Depending on the requirements, the deployment stage can be as simple as generating a report or as complex as implementing a repeatable data-mining process.

A plan should be developed that, given the models that will be deployed, lists the various actions to take. Continuing with the telecommunications example, suppose that the data miner generates a list of customers at risk of churning. How is this information shared with the marketing department, so that the marketing department can approach these customers? Perhaps the list needs also to be shared with the call center, so that when a customer at risk calls in, that customer can be made an interesting offer. The deployment plan includes these types of actions.

Another task in this stage is to create a maintenance plan. Eventually the model will expire and you will need to start a new project. In the telecommunications example, suppose that your model tells you that adolescent men with handset A are at risk of churning. Now suppose that the marketing department approached this group with an interesting offer for a new handset, and that each person accepted the offer. Then the model will be the victim of its own success, and the model will no longer be applicable. You then will have to return to an earlier stage in the data-mining process.

Also, the maintenance plan should include directives for how you can monitor the model's success. For example, in database marketing it is a common strategy to apply the model to only a part of the customer database. The response rate of this targeted group will then be compared with the response rate of a group of randomly selected customers.

Deployment of the model is not the end of the project. Even if the purpose of the model is to only increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the organization can use for decision-making. Essentially in all projects a final report will need to be produced and distributed.

Finally, review the whole project and document the lessons learned. Also, report the results in terms of the business objectives. In example of the telecommunications firm, the question is if you succeeded in reducing churn by 10%. Also, calculate the ROI (Return On Investment) of the project, by estimating the project costs (personnel, software) and the revenues that you retained because you retained customers who would have churned otherwise.

In the *Introduction to Modeling* module of this course you will learn how to apply the model to new cases, which is the end point of the course. Deployment of the results in the organization is beyond the scope of this course, because it will depend on your business objectives, and the infrastructure of your organization.

## The Life Cycle of a Data-Mining Project

- The stages influence each other in a non-linear way
- Data mining is an ongoing endeavor

© 2014 IBM Corporation



While there is a general tendency for the process to flow through the steps in the order outlined above, there are also a number of places where the phases influence each other in a nonlinear way.

You will rarely, if ever, simply plan a data-mining project, execute it and then pack up the data and go home. Using data mining to address customers' demands is an ongoing iterative endeavor. The knowledge gained from one cycle of data mining will almost invariably lead to new questions, new issues, and new opportunities to identify and meet customers' needs. Those new questions, issues, and opportunities can usually be addressed by mining data once again. This process of mining and identifying new opportunities should become part of the way that you think of the business and a cornerstone of the overall business strategy.

## Data-Mining Success (1 of 4)

- Measures of success:
  - the initial assessment will be directly tied to the predictive accuracy
  - in the long run the success of a data-mining effort is measured by concrete factors

© 2014 IBM Corporation



The CRISP-DM model tells us, in the Evaluation stage, to assess the results with respect to business success, not statistical criteria. And indeed, from the moment you begin to develop a research question, the eventual evaluation of the results should be foremost in mind. The initial assessment will be directly tied to the modeling effort; that is to say that you will be concerned with predictive accuracy (for example, the ability to predict if a customer churns). But in the long run the success of a data-mining effort will be measured by concrete factors such as reduced savings, ROI, profitability, and so forth.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-24

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Data-Mining Success (2 of 4)

- Monitoring:
  - after deployment, collect data to assess the model's success

© 2014 IBM Corporation



To determine success, you must monitor the model after it is deployed. Once a model has been deployed, plans must be put in place to record the data and information that make it possible to assess the model's success. Thus, if a real-time model is being used to supply sales representatives with offers for customers, both the suggested offer and the customer's decision, among other factors, must be retained in a database for future analysis.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

1-25

## Data-Mining Success (3 of 4)

- Cost of errors:
  - there will always be errors, sometimes with high cost
  - if no cost estimates are possible beforehand, then try to gather this information afterwards, for future use

© 2014 IBM Corporation



Do not forget to consider the cost of errors as another measure of success. You tend to focus on success but there will always be errors, and sometimes the cost of making errors can be high. For example, mispredicting which insurance claims are fraudulent may be expensive because of the effort involved to investigate the claim further. Some data-mining tools allow you to take cost into account when estimating the model. Use this feature if it is possible to make even a rough cost estimate. When you cannot use cost in the modeling stage, be sure to think carefully about the costs of errors before deployment. And if no reliable cost estimates are possible beforehand, then try to gather this information after the fact for use in future data-mining projects and as ad hoc evaluation criteria.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-26

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Data-Mining Success (4 of 4)

- Other measures of project successes:
  - seek other measures to determine success from a business perspective
  - bring successes to the attention of colleagues and management early on in the project, so that tracking systems or reports can be developed

© 2014 IBM Corporation



As you develop a model and think about its deployment, consider what other measures can be used to determine how successful and useful it is, from a business or organization perspective. Do not wait to mention these factors until after deployment, but bring them to the attention of colleagues and management early on, so that tracking systems or reports can be developed. In the case of a financial institution using data mining to predict customer retention, there are many other factors to investigate beyond simple retention. Changes in average account balance, account activity, account profitability, the opening of other accounts, and use of other services (ATM card) can be investigated after the model is deployed to see if they are also changing.

## Data-Mining Failure (1 of 4)

- Bad data:

- no data mining algorithm will be able to compensate for large amounts of error in the data
- never scrimp on the time spent on data preparation and cleaning

© 2014 IBM Corporation



Not every data-mining project is successful, or, at the least, not as successful as you might have anticipated. As with any research lots of things can go wrong. In this section some serious problems that can occur are reviewed.

Earlier the need for clean and valid data was stressed as important to the data-mining effort. If the data have large amounts of error, no data-mining technique will be able to compensate for this problem. In the worst case a potentially good set of predictors may fail because of error that masks their effect. Take the time to thoroughly prepare and clean the data and continue to check the data as it is modified during the analysis and afterwards. The time to learn about bad data is before, not after, the report has been written or the model has been deployed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-28

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Data-Mining Failure (2 of 4)

- Organizational resistance:
  - difficulties implementing a solution are still part of the whole data-mining effort
  - to address resistance, educate and convince others about the potential benefits of the solution
  - consider implementation in only a portion of the organization

© 2014 IBM Corporation



Difficulties implementing a solution are still part of the whole data-mining effort. A Health Maintenance Organization (HMO) investigated ways to reduce costs by looking at patterns of treatment and care, and found that there was an optimal length of stay in the hospital for several types of major surgeries. While not requiring doctors to rigidly follow the statistical results (which would be inappropriate for any specific patient), the HMO encouraged doctors to take this information into account. But after a few months, it was clear that length of stay decisions were not changing, that is, that the physicians were sticking to their current practices. When resistance occurs, the best strategy is usually further education on the potential benefits of the solution, or perhaps, implementation in only a portion of the organization. For the HMO, this could mean convincing a few doctors initially to change their release decisions, hoping that eventually more will follow this lead.

## Data-Mining Failure (3 of 4)

- Results that cannot be deployed:
  - factors can be out of the control, or cannot legally be used in marketing or in making decisions

© 2014 IBM Corporation



Sometimes a model cannot be deployed for factors other than organizational opposition. The most common reason is because factors found to be important are out of the control of the organization, or cannot legally be used in marketing or in making decisions. A consumer products company discovered that certain types of promotions were successful and led to repeat business, but could only offer these promotions to customers it could readily identify, which in practice were those who returned a registration card or bought a service contract. Some obstacles can be anticipated, and the data-mining process adjusted accordingly. If a model can be only partially implemented, as with the consumer products firm, it may still be worthwhile to do the analysis when sufficiently good results would justify the effort (this is always a judgment call).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-30

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Data-Mining Failure (4 of 4)

- Cause and effect:
  - you must be certain that inputs/predictors in a model occur before the output

© 2014 IBM Corporation



Research methodology is important for the data-mining effort. One reason is because a carefully formulated study will consider whether there is a cause-and-effect relationship between the predictors and outcome variable. For example, customer satisfaction research often uses attitudes about product/service fields to predict overall satisfaction, willingness to buy again/ to remain a customer, or willingness to recommend a product/service. In terms of cause and effect, all these attitudes about fields and future actions or satisfaction occur at one point in time, that is, when the survey is conducted. It can then be argued that while these attitudes may be correlated, claiming that one attitude causes another is not necessarily correct; instead, the attitudes may be mutually reinforcing. When this is true, the predictions from a model about how changes in attitudes affect the target field may be invalid. The basic point is that you must be certain that predictors in a model occur before the target field.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

1-31

## Skills Needed for Data Mining (1 of 4)

- Understanding the business:
  - asking the right data-mining question requires knowledge of the specific business area and organization
  - evaluating a data-mining solution needs a business perspective

© 2014 IBM Corporation



For a successful data-mining project, several disparate skills are useful, and they rarely reside in a single individual.

Framing the business question to be answered by data mining, evaluating the results in terms of business objectives, and presenting the recommendations all require knowledge of the specific business area and organization. Thus someone who knows the critical issues facing the organization is well suited to pose questions that data mining might address. He or she can also evaluate a data-mining solution in terms of business objectives and whether it makes sense. It should be pointed out that experienced data-mining consultants who focus within an industry could develop a good knowledge of these issues. Without this component, a data-mining project runs the risk of producing a good technical solution to a question unimportant to the business.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-32

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Skills Needed for Data Mining (2 of 4)

- Database knowledge:
  - the database administrator plays an important role:
    - Which data tables or files are available?
    - How are they linked?
    - How are the fields coded?
    - What are reasonable data values?

© 2014 IBM Corporation



A data-mining project cannot succeed without good data. The most sophisticated analytic techniques cannot be expected to overcome inconsistent, incomplete, and irrelevant data. For this reason a database administrator (DBA) is usually a key member of the data-mining project team. Typically, neither the business expert nor the analyst has a sufficiently deep knowledge of the data available on the company's systems to do this. What data tables or files are available? How are they linked? What are reasonable and what are incorrect or outlying data values? What do the fields really mean? Only someone familiar with the corporate data systems can usually answer these and other questions. Without this component, you run the risk of producing an incorrect answer to the right question using the best method, or of failing to find a reachable solution.

## Skills Needed for Data Mining (3 of 4)

- Knowledge of data-mining techniques:
  - best tools for situation
  - fine-tuning techniques
  - assess effects of data on outcome
  - identify anomalies

© 2014 IBM Corporation



Although data-mining tools are available that allow pushbutton ease of running an analysis, as you would expect, knowledge of data-mining techniques is needed. Deciding on the best tools to use for a specific question, knowing how to tweak a technique to its optimum, being able to assess the effects of odd data values or missing data, and recognizing that something does not look right, can all contribute to the success of the project. An analyst skilled (trained or self-taught) in these techniques is needed. Without this component, you may fail to answer or may incorrectly answer an important question, even with the benefit of good data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-34

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Skills Needed for Data Mining (4 of 4)

- Team work combining multiple competencies, such as:
  - business domain knowledge
  - database knowledge
  - data-mining algorithms
  - project management

© 2014 IBM Corporation



The deployment of a model on new data may be done outside of MODELER in the database, or you might use a generated model from MODELER and embed it in another application. Specific skills are needed to implement these types of deployments, and this may call for other team members with programming skills that a data-mining analyst does not possess.

For these reasons, most data-mining projects require teams of individuals who contribute differently to the various steps in the data-mining process. It would be ideal if all the needed skills were to reside in one person, but this is rarely the case.

Occasionally, a team member can serve multiple functions (business and database knowledge, or database and data-mining knowledge), but it is relatively rare that all these skills reside in one individual. Of necessity, this confluence of skills in an individual is more likely to occur in small companies and small projects (those that are resource challenged), and that can be limited in the various types of software employed.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Fill in the blank. There are \_\_\_\_ stages in the CRISP-DM process model.

Question 2: Suppose that you are working as a data miner in a project where you need to access certain tables in the database. You ask your database administrator (DBA) for access to the database and he promises to do so as soon as possible. However, a minute later your company experiences serious problems in the operational systems, which also fall under the DBA's responsibility. It is top priority for your DBA to get the operational systems up and running again, which will cost 2 days at a minimum. In the meantime you are stuck in the project, facing a deadline. Could such a situation occur in a data-mining process where CRISP-DM is followed?

- A. Yes
- B. No

Question 3: Suppose that you are working in a project where the business objective was to reduce churn. After some weeks of data preparation and analyses, you have built a model that identifies 95% of the churners in 10% of the data. You advise your manager to deploy the model on the current customer database, so that 95% of the potential churners can be identified (that is the expectation based on your analyses). However, your manager says that 100% of the potential churners must be identified, no matter what, and so she will not deploy the model that you have built.

Could such a situation happen in a data-mining project where CRISP-DM is followed?

- A. Yes
- B. No

Question 4: Which of the following statements are correct? Refer to the table below, listing the tasks in the Business Understanding stage.

- A. Business objectives should be defined before defining data-mining objectives.
- B. Before defining business success criteria, you should produce a project plan.
- C. Resources should be known before defining data-mining goals.
- D. When you want to ensure that all participants "speak the same language", you need to define the terms used in the project.

| <b>Task</b>                      | <b>Sub tasks</b>   |
|----------------------------------|--|
| 1. Determine business objectives | <ul style="list-style-type: none"> <li>a. Background</li> <li>b. Business objectives</li> <li>c. Business success criteria</li> </ul>  |
| 2. Assess situation              | <ul style="list-style-type: none"> <li>a. Inventory of resources</li> <li>b. Requirements, assumptions and constraints</li> <li>c. Risks and contingencies</li> <li>d. Terminology</li> <li>e. Costs and benefits</li> </ul> |
| 3. Determine data-mining goals   | <ul style="list-style-type: none"> <li>a. Data-mining goals</li> <li>b. Data-mining criteria</li> </ul>  |
| 4. Produce a project plan        | <ul style="list-style-type: none"> <li>a. Project plan</li> <li>b. Initial assessment of tools and technologies</li> </ul>   |

## Answers to questions:

Answer 1: There are 6 stages in the CRISP-DM methodology.

Answer 2: B. No. Problems such as these should be acknowledged in the Business Understanding stage, and properly dealt with at that point.

Answer 3: No. It should be clear upfront that the manager only accepts a solution that identifies 100% of the churners. The data-mining goal should have been addressed in the Business Understanding stage.

Answer 4: A, C, D. Business objectives should be defined before defining data-mining objectives. Resources should be known before defining data-mining goals. You need to define the terms used in the project.

## Summary

- At the end of this module, you should be able to:
  - list two applications of data mining
  - explain the stages of the CRISP-DM process model
  - describe successful data-mining projects and the reasons why projects fail
  - describe the skills needed for data mining

© 2014 IBM Corporation

The key point in this module was that data mining cannot be equated with running advanced algorithms on huge amounts of data. Above all, data mining is a process that needs a thorough methodology to make the project successful. CRISP-DM provides this methodology.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

1-39

# Workshop 1

## Introduction to Data Mining



© 2014 IBM Corporation

No files are used in this workshop.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-40

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Workshop 1:Introduction to Data Mining

You are working in the database marketing department of ACME, a (fictitious) company selling sport products via the Web and via mail campaigns. ACME has 30,000 customers.

It is your job to launch a mail campaign for the long awaited XL Original Orange Baseball Cap. Given the costs of contacting a customer and to not bother customers with unrequested mail you have the following project plan in place:

- You will not address all 30,000 customers right away, but you will randomly select 10,000 customers and send them an e-mail with an offer for the XL Original Orange Baseball Cap (this mailing will be named "test mailing" from now on.)
- For the test mailing, you will record how many customers responded positive (accepted the offer for the XL Original Orange Baseball Cap).
- You will use modeling techniques on your test mailing data to find groups that show high response rates.
- Once you have built a satisfactory model, you will apply that model to the 20,000 customer that were not included in the test mailing. Only groups for which the model predicts high response rates will be contacted by regular mail (this mailing will be named "actual mailing" from now on). For example, if the results on the test mailing tell you that single men up to age 25 show a high response rate, then single men up to age 25 will be selected in the rest of the database and they will be included in the actual mailing.

Given this background, answer the following questions:

1. When you want to apply the results found on the test mailing (by e-mail) to the actual mailing (by regular mail), which assumption are you making?
2. You have a budget for the actual mailing of 10,000. Furthermore, you know that the cost per mail address (cost for contacting a customer by mail) is 2.5.

What is the maximum number of customers that you can contact for the actual mailing?

Two months after that you have sent out the test mailing you know the response for the test mailing: 400 of the 10,000 customers that were included in the test mailing bought the XL Original Orange Baseball Cap, a response percentage of  $(400/10,000) * 100 = 4\%$ .

3. Given the maximum number of customers that can be contacted for the actual mailing (refer to the previous question), how many responders (customers buying the XL Original Orange Baseball Cap) do you expect if you would randomly select customers for the actual mailing (assuming that the results on the test mailing are representative for the actual mailing)?
4. Each XL Original Orange Baseball Cap sold generates 50 revenues. In the scenario of a random mailing and given the expected number of responders in this scenario (refer to the previous question), what is the expected revenues for the mail campaign?
5. Given the costs for the actual mailing (10,000) and the expected revenues for the actual mailing (refer to the previous question), do you think a random selection of customers for the actual mailing will make profit?
6. The entire dataset includes 30,000 customers, of which 10,000 were included in the test mailing. When you build a model to predict whether a customer will buy the XL Original Orange Baseball Cap, will you use the data of all 30,000 customers, or will you select a part of the dataset?
7. The table that follows lists all fields in your dataset. Which fields can be used to predict response? For example, is customer\_id a field that you would use to predict response?

For each of the fields in the table that follows, indicate whether you would include it as a predictor for response (buy or not buy the XL Original Orange Baseball Cap):

| Field                               | Field Description   | Predictor for Response (Y/N)?       |
|-------------------------------------|---|-------------------------------------|
| customer_id                         | the customer's identification number  |                                     |
| gender                              | the customer's gender   |                                     |
| email_address                       | the customer's e-mail address   |                                     |
| postal_code                         | the customer's postal code (20 digits)  |                                     |
| recency 01-01-2011                  | the customer's last order date, before JAN-01-2011  |                                     |
| frequency 01-01-2011                | the customer's number of orders, before JAN-01-2011   |                                     |
| monetary_value 01-01-2011           | the customer's total purchase amount, before JAN-01-2011.   |                                     |
| has received test mailing           | a field that flags whether the customer was in the test mailing, sent out FEB-01-2011. This field is true for all 10,000 customers included in the test mailing, false for all 20,000 customers not included in the test mailing. |                                     |
| response to test mailing 02-01-2011 | for customers in the test mailing, this field flags whether the customer ordered the XL Original Orange Baseball Cap. For customers not in the test mailing, this field is undefined.   | --- (the field you want to predict) |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

1-43

|  |   |  |
|--|---|--|
| order date   | for customers in the test mailing who have ordered the XL Original Orange Baseball Cap, this field gives the date that the XL Original Orange Baseball Cap was ordered. For customers in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined.                    |  |
| number of days between test mailing and order date | gives the time (in days) between the test mailing (FEB-01-2011) and the order date. This field is valid for customers in the test mailing who ordered the XL Original Orange Baseball Cap. For customers in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined. |  |
| ordered within month                               | flags if the order date was within one month after the test mailing went out. This field is valid for those in the test mailing who ordered the XL Original Orange Baseball Cap. For those in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined.               |  |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Workshop 1: Tasks and Results

In this section you will find the answers to the workshop questions.

- When you want to apply the results found on the test mailing (by e-mail) to the actual mailing (by regular mail), which assumption are you making?

*The assumption is that results on the e-mail channel are representative for results on the mail channel. Experience shows that results for e-mail campaigns are a conservative estimate for results on mail campaigns, so it is a safe assumption that response rates on mail campaigns are as good as or even better than response rates for e-mail campaigns.*

- You have a budget for the actual mailing of 10,000. Furthermore, you know that the cost per mail address (cost for contacting a customer by mail) is 2.5.

What is the maximum number of customers that can be contacted for the actual mailing?

*Maximum number of customers that can be contacted is:  $10,000 / 2.5 = 4,000$ .*

- The result for the test mailing was that out of the 10,000 customers, 400 bought the XL Original Orange Baseball Cap, a response percentage of  $(400/10,000) * 100 = 4\%$ .

Given the maximum number of customers that can be contacted for the actual mailing (see the previous question), how many responders (customers buying the XL Original Orange Baseball Cap) do you expect if you would randomly select customers for the actual mailing (assuming that the results on the test mailing are representative for the actual mailing)?

*Expected number of responders in case of a random mailing:  $0.04 * 4,000 = 160$ .*

- Suppose that each XL Original Orange Baseball Cap sold generates 50 revenues. In the scenario of a random mailing and given the expected number of responders in this scenario (see previous question), what is the expected revenues for the mail campaign?

*Expected revenues =  $160 * 50 = 8,000$ .*

5. Given the costs of the actual mailing (10,000) and the expected revenues for the actual mailing (refer to the previous question), do you think that a random selection of customers for the actual mailing will make profit?

*The expected profit in case of a random mailing: revenues - costs = 8,000 - 10,000 = -2,000. So, a random mailing will result in a loss of 2,000. (And that is why you need modeling on the test mailing data to find groups with higher response rates, so you can mail only these groups in the actual mailing).*

6. The entire dataset includes 30,000 customers, of which 10,000 were included in the test mailing. When you build a model to predict whether a customer will buy the XL Original Orange Baseball Cap, will you use the data of all 30,000 customers, or will you select a part of the dataset?

*When you build your model, select the customers who were in the test mailing, because only for these customers you have data on response (whether the customer bought the XL Original Orange Baseball Cap or not).*

7. Which fields in ACME's dataset can be used to predict response? For example, would customer\_id be a field that you would use to predict response?

For each of the fields in the table that follows, indicate whether you would include it as a predictor for response (buy or not buy the XL Original Orange Baseball Cap):

| Field  | Predictor for Response (Y/N)?  |
|--|--|
| customer_id  | No   |
| gender   | Yes  |
| email_address                                      | No   |
| postal_code  | No (too granular; a derived field such as region could be a predictor) |
| recency 01-01-2011                                 | Yes (this field is recorded before the test mailing went out)          |
| frequency 01-01-2011                               | Yes (this field is recorded before the test mailing went out)          |
| monetary_value 01-01-2011                          | Yes (this field is recorded before the test mailing went out)          |
| has received test mailing                          | No (this field is always true for those in the test mailing)           |
| response to test mailing 02-01-2011                | No (the field you want to predict)                                     |
| order date   | No (derived from the target field)                                     |
| number of days between test mailing and order date | No (derived from the target field)                                     |
| ordered within month                               | No (derived from the target field)                                     |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

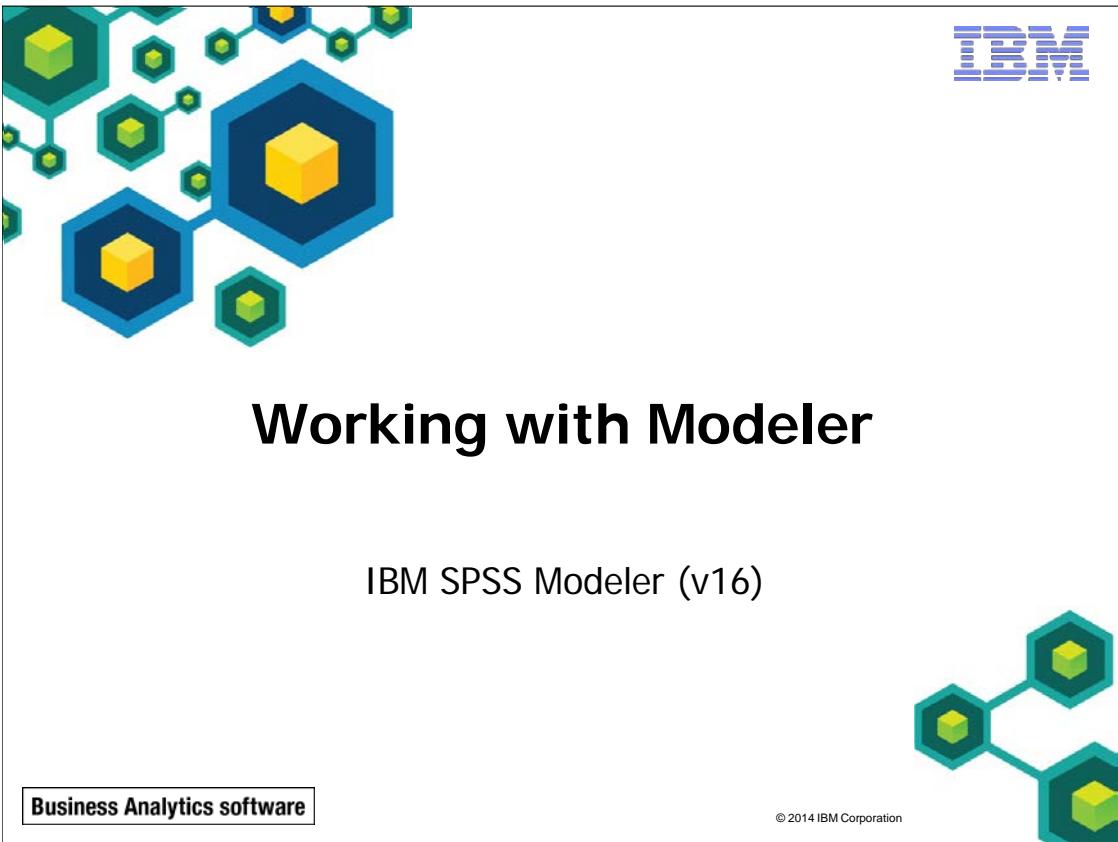
This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

1-47

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

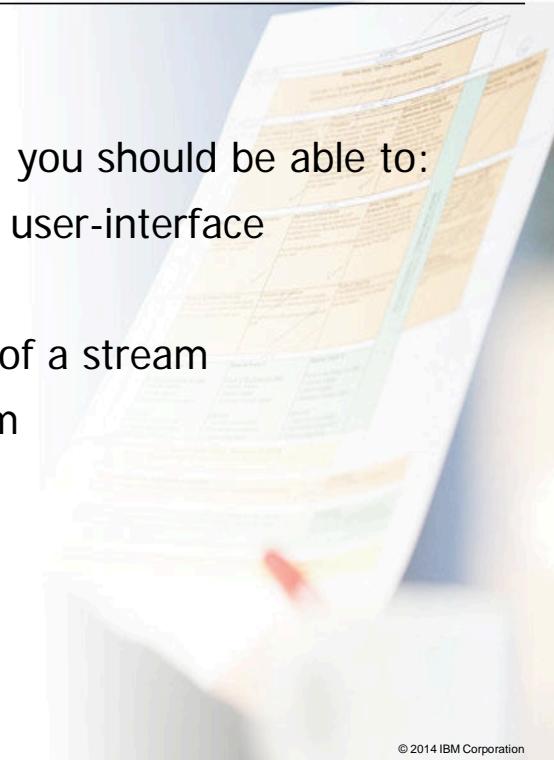
Business Analytics software

IBM

# Objectives

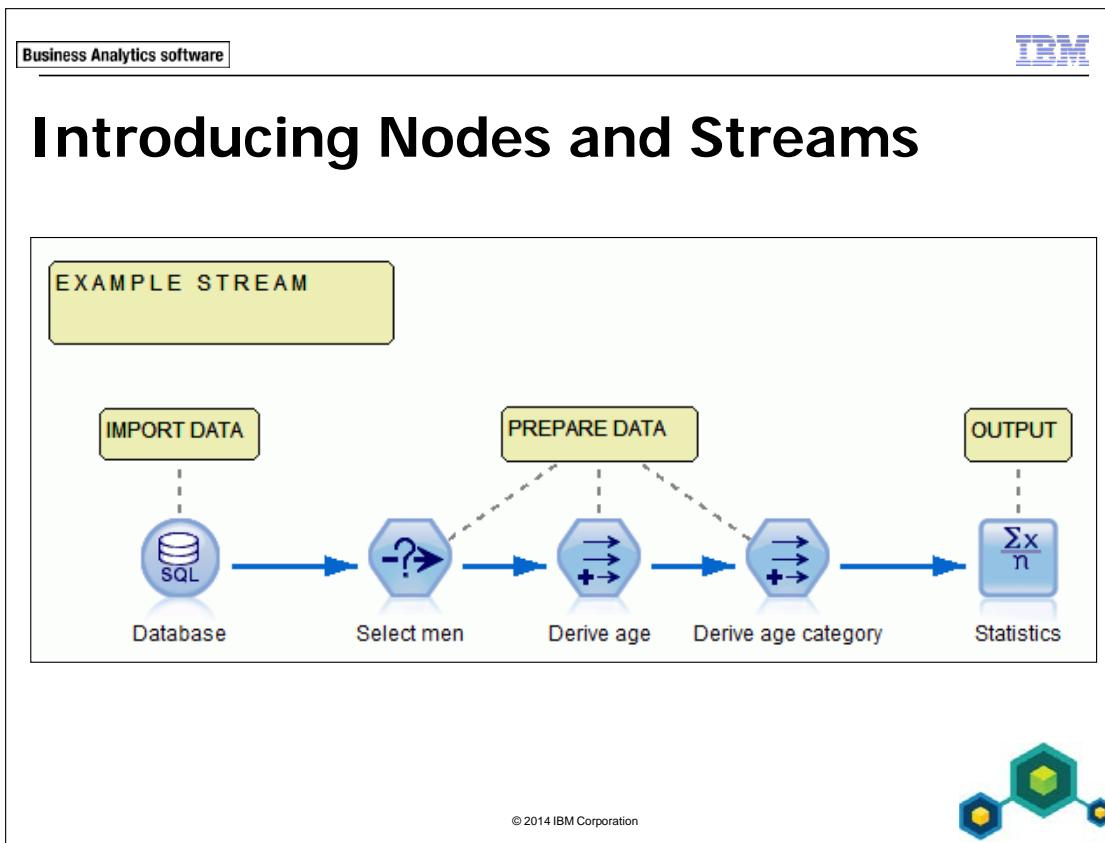
- At the end of this module, you should be able to:
  - describe the MODELER user-interface
  - work with nodes
  - run a stream or a part of a stream
  - open and save a stream
  - use the online Help

© 2014 IBM Corporation



This module introduces you to MODELER. You will become familiar with objects such as streams and nodes, and you will acquire experience with the software.

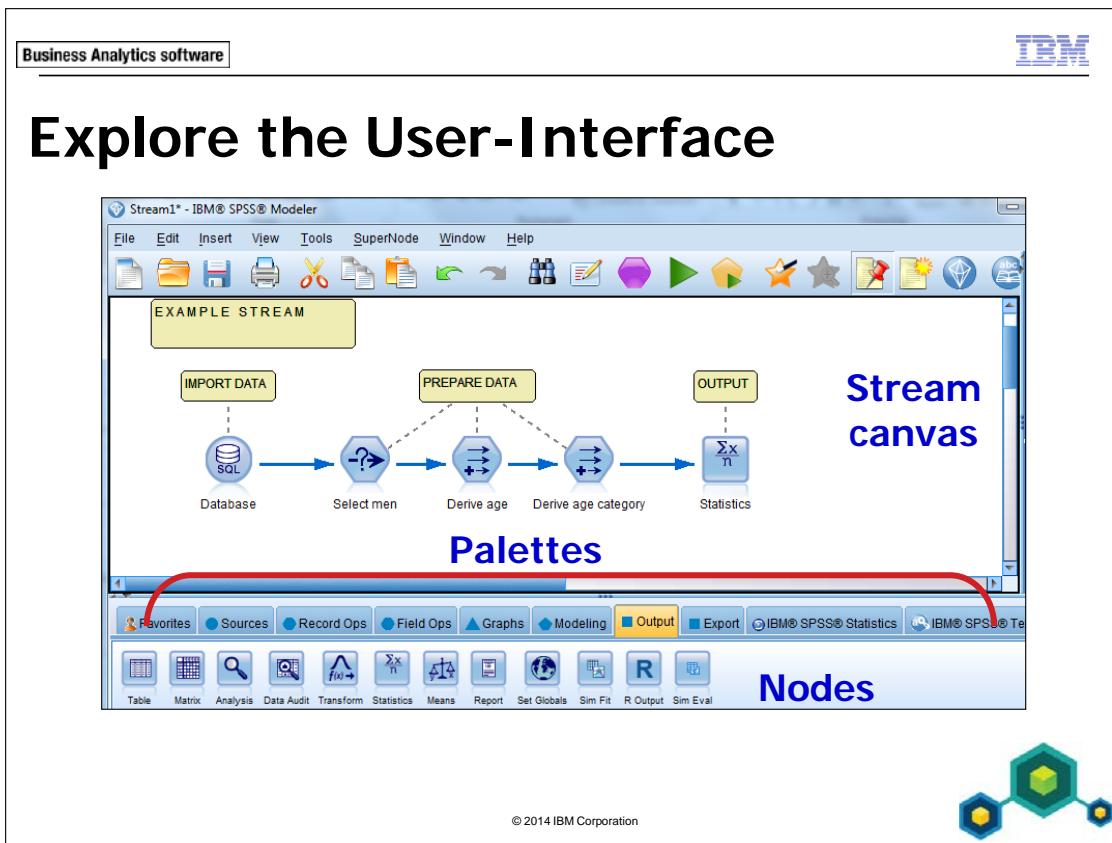
No prior knowledge is required for this module, except general computer literacy.



Before discussing MODELER's user-interface, it is important to introduce two objects: nodes and streams.

Nodes represent operations to be applied to the data. Nodes are linked together to form a stream. A stream represents a flow of data from importing data, through a number of manipulations, to running an analysis. This sequence of operations is known as a stream because the data flows record by record from the data source through each manipulation to the destination, which is some type of output.

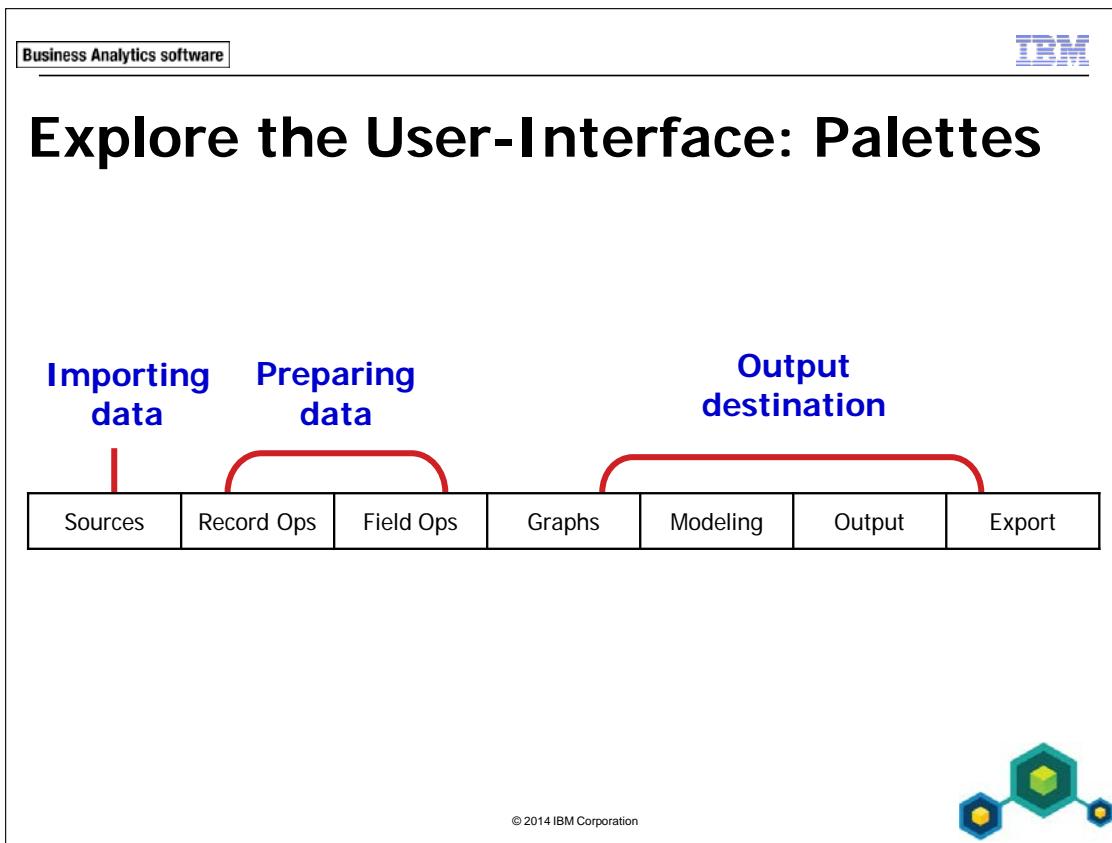
The stream depicted here imports data from a database, prepares the data through a selection of records and the creation of two fields, and then runs a statistical analysis. The stream and nodes are annotated with comments.



At the start of a session MODELER shows a welcome dialog box, asking you what you would like to do. When you cancel this dialog box, you will have the MODELER user-interface, depicted on this slide. (Not shown are two panes, presented later).

The main menu has common entries such as File, Edit, Window, and Help. Specific MODELER menu entries are Insert, View, Tools, and SuperNode. These will be presented later in this course. The toolbar shows a number of standard buttons and buttons specific to MODELER, which will be presented in the course. The buttons that are shown depend on the window that is active. Some extra buttons may appear and some may disappear when another window is open.

The stream canvas is the area where you create your stream. Right-clicking on an empty area on the stream canvas brings up a context menu that enables you to create a new stream, open an existing stream, save a stream, add a comment, set the icon size, zoom in or out, and so forth.



Nodes are contained in palettes. Each palette contains a related group of nodes. The palettes follow the order of the stages in analysis. In the Sources palette you have the nodes to import data from various sources. Two data preparation palettes organize record operations (such as selecting records) and fields operations (such as deriving new fields). Four output destination palettes complete the Sources, Records Ops and Field Ops palettes. The output destination palettes contain nodes for various analysis tasks, such as graphical displays (the Graphs palette), data-mining models (the Modeling palette), reports (the Output palette) and nodes to export your data (the Export palette). Nodes from the output destination palettes appear at the end of a stream and thus are called terminal nodes. No other node can have its input from a terminal node.

Palettes make working with MODELER easy. When you are looking for a certain node, ask yourself if you want to read data, to prepare data or that you want to produce output. If the stage in the analysis is identified, you can zoom in on a particular palette and identify the right node in that palette.

# Explore the User-Interface: Panes

- Manager pane tabs:
  - Streams
  - Output
  - Models
- Project pane tabs:
  - CRISP-DM
  - Classes

© 2014 IBM Corporation



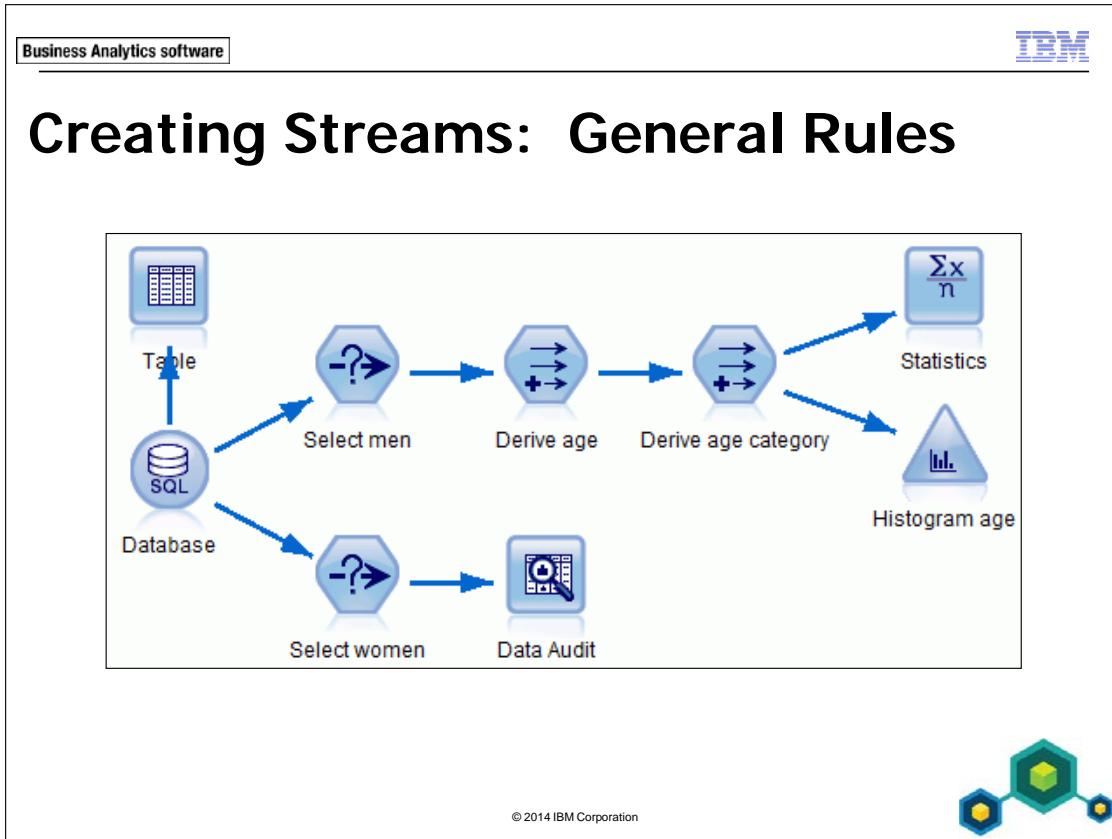
MODELER has two types of panes to manage the work. The Manager pane manages streams, output, and models. For example, on the Output tab you will have output items such as tables and graphs.

The Project pane organizes the work in two possible ways. The CRISP-DM tab helps to organize items according to the stages of the CRISP- DM process model. Even though some items do not involve work in MODELER, the CRISP-DM tab includes all six stages of the CRISP-DM process model so that there is a central location for storing and tracking all materials associated with the project. For example, the Business Understanding stage typically involves documentation to describe the data-mining goals. Such documentation can be stored in the Business Understanding folder, for future reference and inclusion in reports. You also store the work according to the type of the object in the Classes pane. Objects can be added to any of the following categories: Streams, Nodes, Generated Models, Tables, Graphs, Reports, Other (for example documents relevant to the data-mining project).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



Take notice of the following general rules when you build streams:

- A stream starts with one or more nodes from the Sources palette to import the data.
- Most likely, data preparation is the next step, so nodes from the Record Ops and/or the Field Ops palette follow the source node(s).
- Branches are created where needed.
- A terminal node (a node contained in one of the destination palettes, Graphs, Modeling, Output, or Export) will appear at the end of a stream or stream branch.

The stream depicted on this slide branches into three parts from the Database source node. The first branch runs a Table node on the entire dataset. The second branch selects men, derives new fields and runs statistics and a histogram for age. The third branch selects women and runs a data audit.

## Creating Streams: Using the Mouse

| Button  | Use                                 |
|---|-------------------------------------|
| primary mouse button (usually the left mouse button)    | select, place, and position objects |
| secondary mouse button (usually the right mouse button) | invoke a context menu               |
| middle-mouse button                                     | modify connections                  |

© 2014 IBM Corporation



When working with MODELER, the mouse plays an important role in creating streams. Instead of using the mouse, you can use function keys and menus. Throughout this course the mouse will be used.

Note: If your mouse does not have a middle-mouse button, you can simulate this by pressing the Alt key and using the primary mouse button.

## Creating Streams: Placing Nodes

- Identify the appropriate palette
- Identify the appropriate node
- Double-click the node to place it on the canvas (or drag the node to the canvas)

© 2014 IBM Corporation



To build a stream, the appropriate node is picked from the appropriate palette and placed on the stream canvas. Alternatively, use the Insert entry in the main menu to select a palette from there and to drill down to the appropriate node.

A node that you place on the stream canvas will be automatically connected downstream (on the right when you work from left to right) from the node that has focus on the stream canvas. When no node has focus on the stream canvas, the new node will not automatically be connected to a node. In this case, use the middle-mouse button and drag to connect the nodes. If you do not have a middle-mouse button, you can press the Alt key and use the primary mouse button. Alternatively, right-click the node that you want to connect from, select Connect from the context menu, and then click the node you want to connect to.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

2-10

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Creating Streams: Managing Nodes

- Right-click a node, and you can:
  - edit
  - delete, cut, or copy
  - save
  - load
  - annotate or comment
  - disable
  - preview the data

© 2014 IBM Corporation



MODELER provides you with a very user-friendly interface, where objects such as nodes and connections can be managed by using the mouse. This slide and the next slide show some of these operations. A right-click on a node shows the options:

- Edit: Opens a dialog box specific to the node you selected (alternatively, double-click the node).
- Save: Saves the node, so you can reuse the node later (then, load the node).
- Annotate, Comment: Supplies a short (annotation) or longer (comment) description.
- Disable or Enable: Disable a node to keep it in the stream, so that MODELER will ignore it. This is an elegant alternative for deleting a node. If a disabled node needs to be executed, enable the node again.
- Preview: Shows the first records of a dataset (so you can get a feel for the data).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-11

**Business Analytics software**

**IBM**

## Creating Streams: Managing Connections

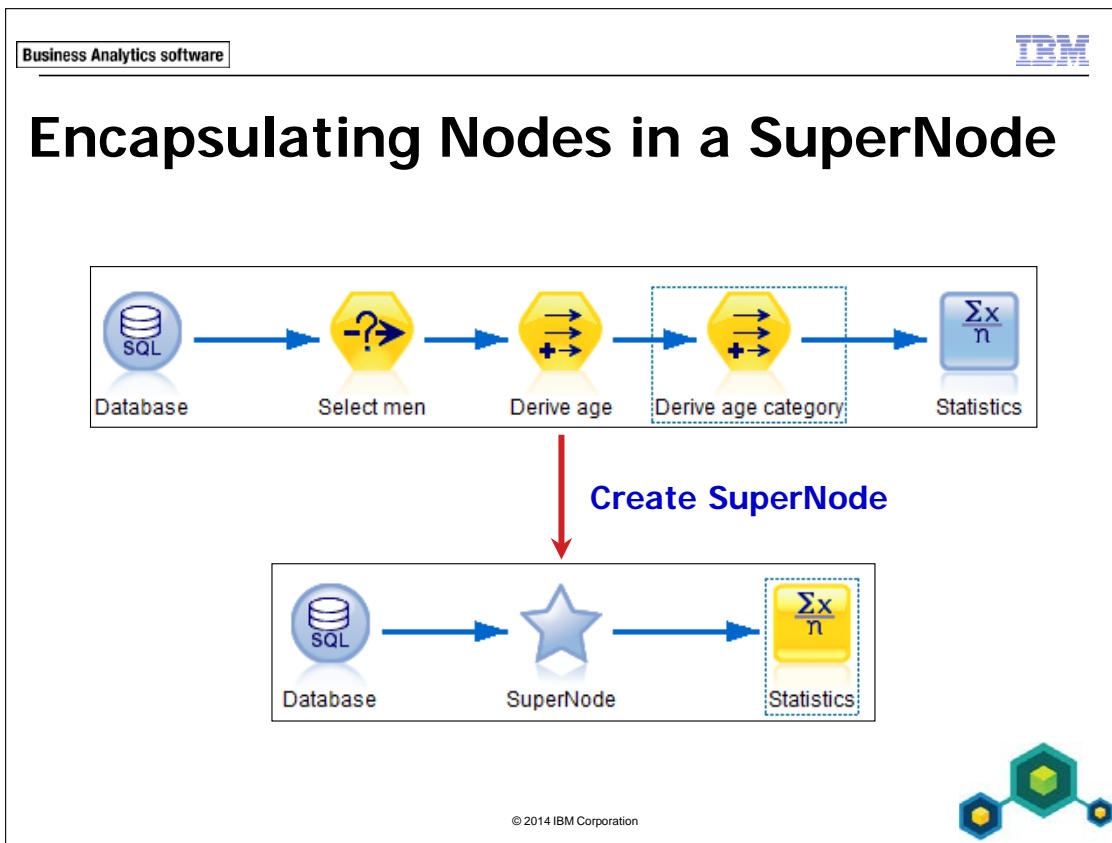
- You can:
  - delete a connection
  - bypassing a node in an existing connection
  - adding a node in an existing connection

© 2014 IBM Corporation

When you want to delete a connection, right-click the connection and choose Delete Connection.

It may happen that you want to bypass a node in an existing connection. For example, node A connects to node B, and node B connects to node C, but you want to take B out. To do so, double-click with the middle mouse button on the node that has to be taken out (or press the Alt key and double-click the primary mouse button when you do not have a middle-mouse button). This will disconnect the node, and it will connect the two nodes that are upstream and downstream from the disconnected node, and then you can delete the selected node.

A very handy feature is shown on this slide. When you want to insert a node, place the new node on the stream canvas, and position it somewhat below the connection that you already have. Then, drag the connection over the node that has to be inserted, and then align the nodes to make the stream neat.



A SuperNode condenses a number of nodes into a single node, which is a useful feature to make a stream neat and manageable. You can also save a SuperNode for future use.

To create a SuperNode, select the nodes that you want to encapsulate in the SuperNode, right-click one of the nodes, and then select Create SuperNode from the context menu (alternatively, select SuperNode in the main menu, and select Create SuperNode).

To view and/or edit the content of the SuperNode, right-click the SuperNode, and then select the Zoom in option from the context menu . The SuperNode window will then open and you can view and edit the nodes. When you want to close the

SuperNode window, click the Zoom out  button. (Do not click the Close  button, because that will end the MODELER session.)

For more information on SuperNodes, refer to the online Help.

The screenshot shows the IBM SPSS Modeler software interface. At the top left is a 'Business Analytics software' logo, and at the top right is the 'IBM' logo. The main title 'Generating Nodes from Output' is displayed prominently. Below the title, a bullet point states: 'Some nodes can be created from an output object'. To the left of the text is a table with three rows and three columns:

| gender | marital_status | response |
|--------|----------------|----------|
| male   | married        | no       |
| female | married        | yes      |

A red arrow labeled 'Generate' points from the table to a vertical list of node options on the right. This list includes:

- Select Node ("Records")
- Select Node ("And")
- Select Node ("Or")
- Derive Node ("Records")
- Derive Node ("And")
- Derive Node ("Or")

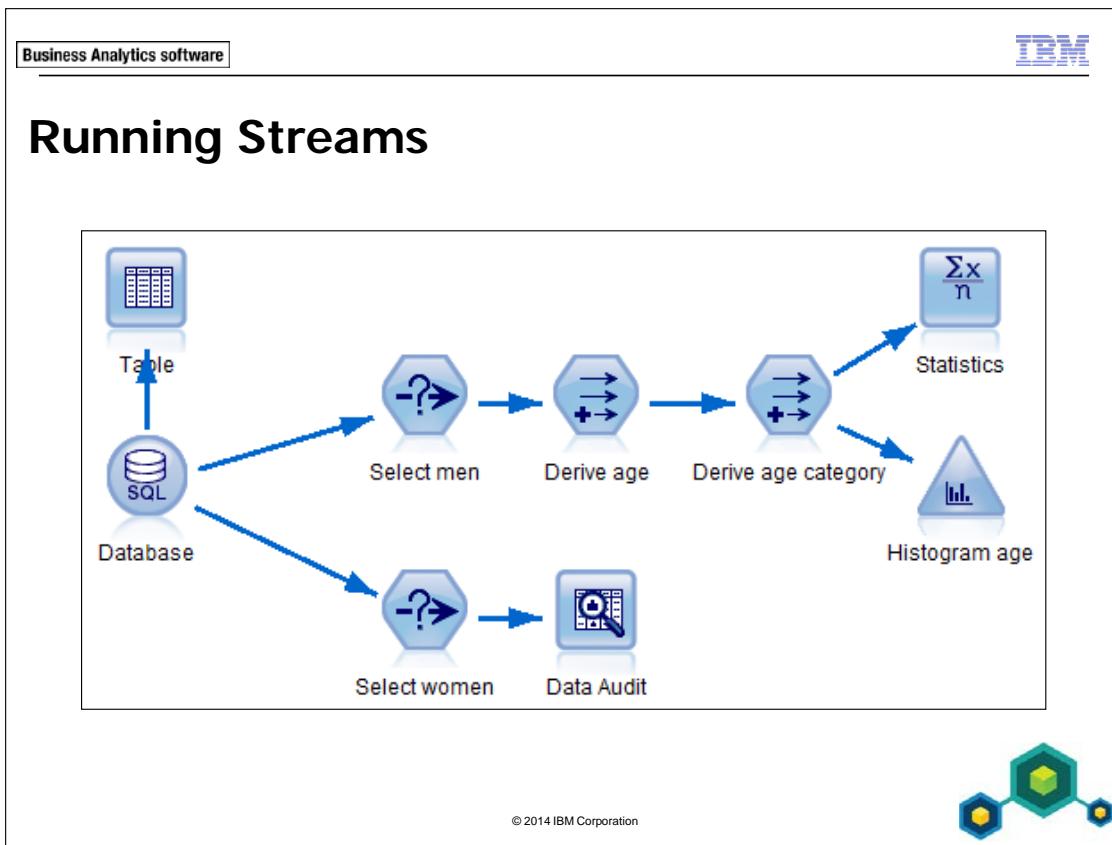
At the bottom right of the interface is a decorative graphic of interconnected hexagons.

You can place nodes on the stream canvas by picking nodes from the appropriate palette. Sometimes there is a more efficient way: you can generate the node from output. For example, you can generate a Select node from the Table output window.

In the example on this slide, the values male and married are selected in the Table output window. Choosing Generate\Select Node ("And") generates a Select node, with the condition: gender = "male" and marital\_status = "married". When you include the generated Select node in your stream, married men will be selected.

The Select Node ("And") option creates a compound condition with each of the parts combined by "and". In the same way, the Select Node ("Or") option creates a compound expression with the parts combined by "or". The Generate\Select Node ("Records") option creates a condition that will select the record that has focus in the Table output window.

The feature to generate a Select node from Table output is very efficient and is illustrated in the demo later in this module.



You have the following options to run a stream:

- Execute all terminal nodes: Click the Run button in the main menu. In the stream shown here, this would run the Table, Statistics, Histogram and Data Audit node.
- Execute selected nodes: Select the nodes you want to execute and click the Run Selection button in the main menu.
- Execute a single node; if it is a terminal node, right-click the node and select the Run option from the context menu (alternatively, edit the node and click the Run button in the dialog box). If it is not a terminal node, right-click the node and select the Run from Here option from the context-menu; all nodes downstream from the node will be executed.

Note: When you choose the Run for Here option, ensure that the branch ends in a terminal node; if not, you will have the message "There are no executable nodes", because MODELER has nothing to do.

**Business Analytics software**

**IBM**

## Online Help

- The Help entry in the main menu
- Context-sensitive help in any dialog box

© 2014 IBM Corporation



MODELER provides online Help in many ways, covering a variety in topics.

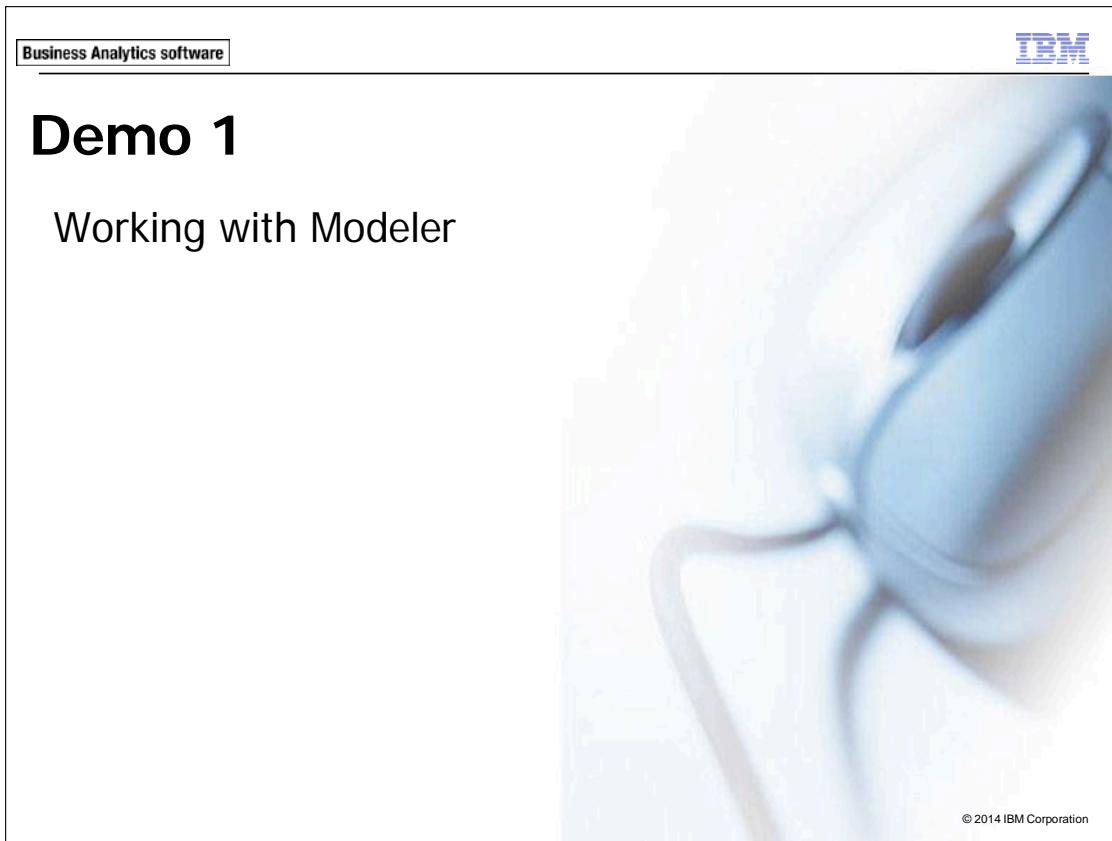
You can also access the help files via Help in the main menu, where you have the following entries:

- Help Topics: Takes the user to the online Help system.
- CRISP-DM Help: Documents the CRISP-DM methodology.
- Application Examples: Leads to a variety of real-life examples of using data-mining.
- Accessibility Help: Informs about keyboard alternatives to the use of the mouse.
- What's This: Changes the cursor into a question mark and provides information about any MODELER item that is selected.

Context sensitive online Help is always available in any open dialog box (click the

Help  button in the dialog box).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



The slide features a large, faint background image of a person wearing a hard hat and safety glasses, looking at a computer screen. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the IBM logo is displayed. At the bottom right of the slide, the text "© 2014 IBM Corporation" is visible.

# Demo 1

## Working with Modeler

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the **C:\Train\0A005** folder and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following files are used in the demo:

- **demo\_data\_working\_with\_modeler.xls** – a Microsoft Excel file from a (fictitious) telecommunications firm
- **02-Working\_with\_Modeler\Start Files\demo\_working\_with\_modeler.str** – a stream that imports the demo data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Demo 1: Working with Modeler

### Purpose:

You want to become familiar with MODELER's user-interface to create streams.

### Task 1. Creating streams.

In this task you can practice your skills in creating streams. No data will be used in this task, so that you can focus on MODELER's user-interface.

The following table outlines the operations in this task, and the appropriate palette/node for the operation.

| Operation  | Palette – Node                          |
|--|---|
| Data from a Microsoft Excel file has to be imported    | Source palette - Excel node             |
| A sample of records has to be drawn                    | Record Operations palette - Sample node |
| A new field has to be derived                          | Field Operations palette - Derive node  |
| A histogram graph for the new field is requested       | Graphs palette - Histogram node         |
| Data have to be exported to a IBM SPSS Statistics file | Export palette - Statistics Export node |
| Add comment to the Histogram                           | Histogram node - context menu – Comment |

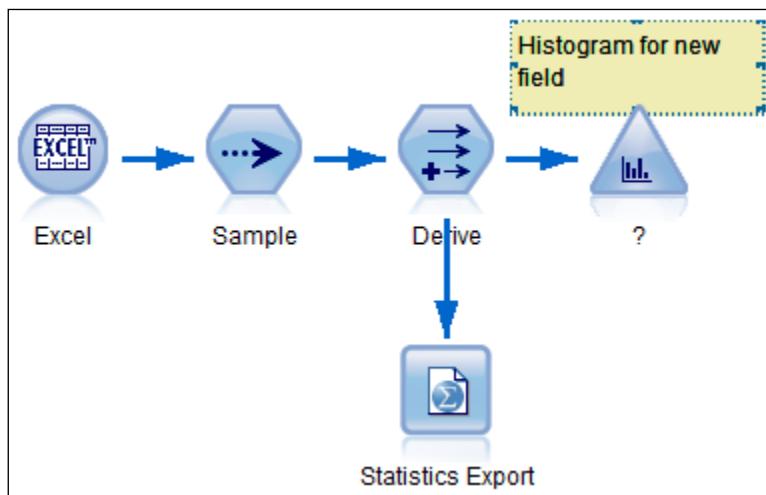
To create a stream that accomplishes this, take the following steps.

1. Click the **Sources** palette, and then double-click the **Excel** node (this places an Excel node on the stream canvas).
2. Select the **Record Ops** palette, and then double-click the **Sample** node (this places a Sample node on the stream canvas, downstream from the Excel node).
3. Click the **Field Ops** palette, and then double-click the **Derive** node (this places a Derive node downstream from the Sample node).
4. Click the **Graphs** palette, and then double-click the **Histogram** node (this places a Histogram node downstream from the Derive node).  
Note: the Histogram asks for a field name. No data is used in this task, so leave this as it is.
5. Click the **Export** palette, and then double-click the **Statistics Export** node (this places a Statistics Export node downstream from the Derive node; the Statistics Export node will not be connected to the Histogram node - recall that nodes from the Graphs, Modeling, Output, and Export palettes are terminal nodes).

You will add a comment to the Histogram node.

6. Right-click the **Histogram** node, select **New Comment** from the context menu, and then type **Histogram for new field** (this places a comment box on the Canvas, with the text just typed).

Your stream appears as follows:



Leave the stream open for the next task.

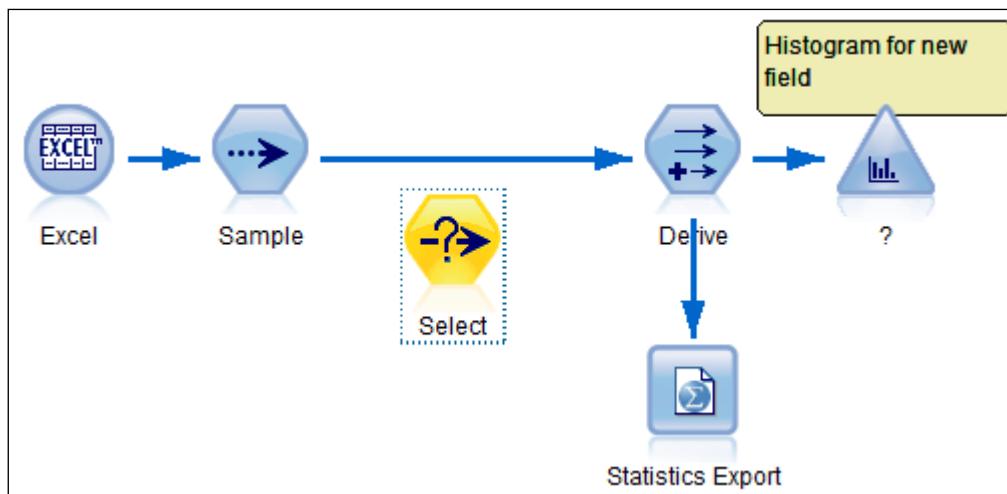
## Task 2. Changing streams.

Suppose that the stream that you just have created must be changed in the following way: after you have drawn a sample you want to select records. Selecting records is a record operation, so the node to use must be found in the Record Ops palette.

You will make room for the node that needs to be inserted.

1. Using the mouse, drag a **rectangle** that selects the **Derive** node, the **Histogram** node, and the **Statistics** node and **move them to the right**.
2. From the **Record Ops** palette, select the **Select** node and place it on the stream canvas, under the connection from the **Sample** node to the **Derive** node.

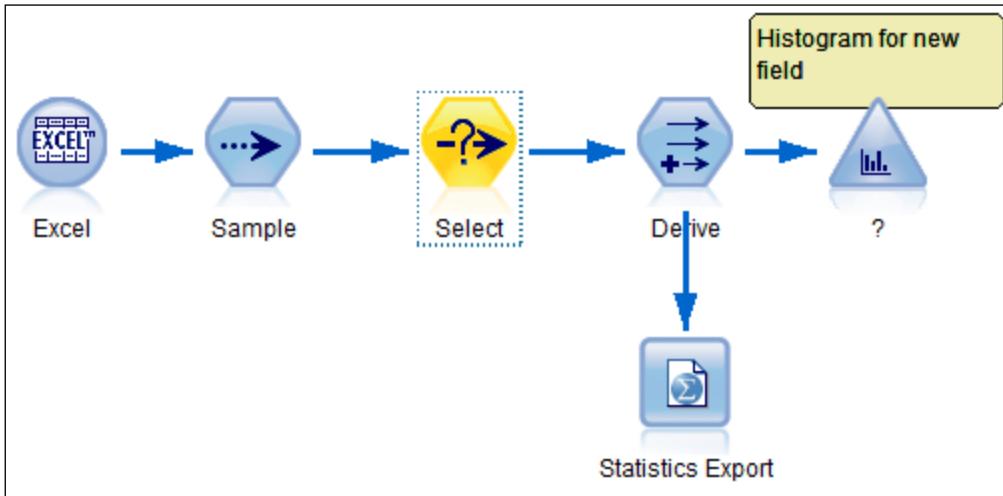
Your stream appears as follows:



3. Select the **connection** from the **Sample node to the Derive node** and drag it over the **Select node**.

- Reposition the **Select** node, so that it is on the same height as the **Sample** and **Derive** node.

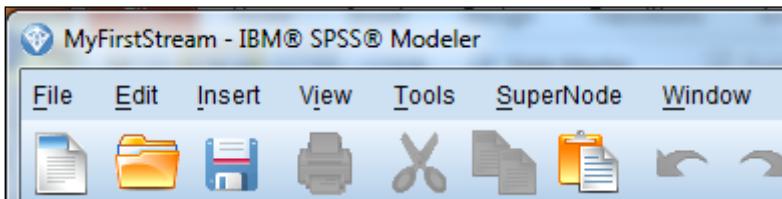
Your stream appears as follows:



Now that you have built a stream, you will save it.

- Select **File\Save Stream** from the main menu; for **File Name**, type **MyFirstStream.str**, and then click **Save**.

The window's title bar shows the file name:



Leave IBM SPSS Modeler open for the next task.

### Task 3. Generate a Select node from Table output.

Nodes can be placed on the stream canvas by selecting them from the appropriate palette. For some nodes there is a handy alternative, as demonstrated in this task.

This task uses a ready-made stream, **demo\_working\_with\_modeler.str** that imports data from Microsoft Excel. You will open this stream.

- Select **File\Open** from the main menu.
- Open **demo\_working\_with\_modeler.str**, located in the **02-Working\_with\_Modeler\Start Files** sub folder.

- To view the data in MODELER, you will run a Table node.
3. Run the **Table** node that is already on the stream canvas.

Running a Table node will show the data. A section of the results appears as follows:

|    | customer_id | data_known | gender   | age    |
|----|-------------|------------|----------|--------|
| 1  | K100150     | yes        | MALE     | 30.... |
| 2  | K100180     | yes        | MALE     | 37.... |
| 3  | K100690     | yes        | MALE     | 30.... |
| 4  | K101420     | yes        | MALE     | 33.... |
| 5  | K104780     | yes        | FEMA...  | 52.... |
| 6  | K105440     | yes        | FEMA...  | 26.... |
| 7  | K108360     | yes        | FEMA...  | 34.... |
| 8  | K110390     | yes        | MALE     | 40.... |
| 9  | Z102390     | no         | \$null\$ | \$n... |
| 10 | Z137850     | no         | \$null\$ | \$n... |

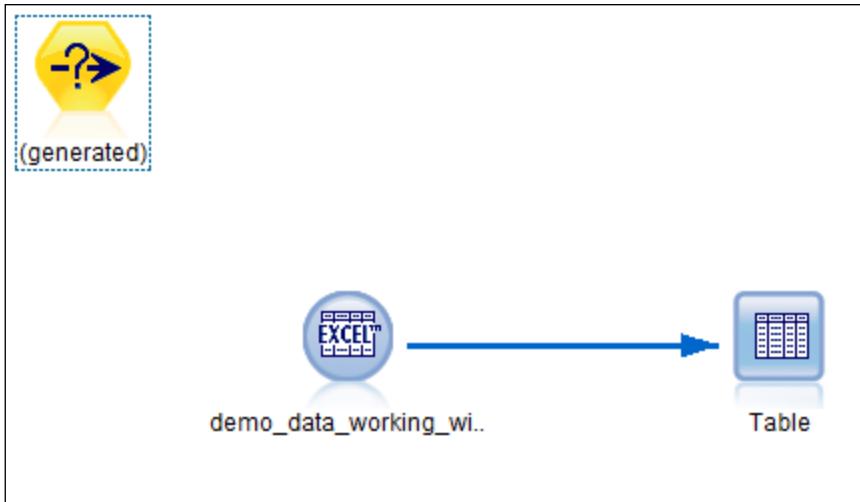
Leave this Table output window open.

Notice, that no data are known for the last two customers, as indicated by the field data\_known (gender through retention are \$null\$, which represents MODELER's undefined value).

Suppose you want to select the records with known data. One method would be to manually add a Select node, edit the node, type the condition, and so on. A quicker and more user-friendly alternative is it to generate the Select node from the Table output window.

4. In the **first row**, in the **data\_known** column, click the value **yes**, so this value is selected.
5. Choose **Generate\Select Node ("And")**.
6. Click **OK** to close the **Table** output window.

Your stream appears as follows:



A Select node is generated with the condition `data_known = "yes"` and placed in the upper left corner of the stream canvas.

At this point, the Select node is not part of the stream, so no data will pass through it. You will include the generated node in the stream.

7. Insert the generated **Select** node between the **Excel** source node and the **Table** node.

You will annotate the node, so it is clear which records are selected.

8. Right-click the **Select** node named **(generated)**, and:

- from the context menu, select **Rename and Annotate**
- replace the text **(generated)** with **data known**
- click **OK** to close the **Select** dialog box

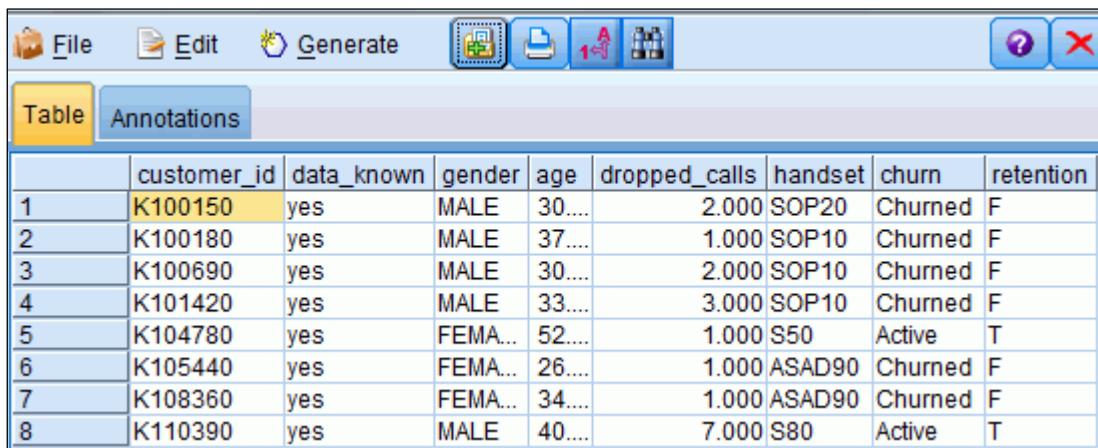
Your stream appears as follows:



To check your results, you will run the Table node.

- Right-click the **Table** node, and select **Run** from the context menu.

The results appear as follows:



|   | customer_id | data_known | gender  | age    | dropped_calls | handset | churn   | retention |
|---|-------------|------------|---------|--------|---------------|---------|---------|-----------|
| 1 | K100150     | yes        | MALE    | 30.... | 2.000         | SOP20   | Churned | F         |
| 2 | K100180     | yes        | MALE    | 37.... | 1.000         | SOP10   | Churned | F         |
| 3 | K100690     | yes        | MALE    | 30.... | 2.000         | SOP10   | Churned | F         |
| 4 | K101420     | yes        | MALE    | 33.... | 3.000         | SOP10   | Churned | F         |
| 5 | K104780     | yes        | FEMA... | 52.... | 1.000         | S50     | Active  | T         |
| 6 | K105440     | yes        | FEMA... | 26.... | 1.000         | ASAD90  | Churned | F         |
| 7 | K108360     | yes        | FEMA... | 34.... | 1.000         | ASAD90  | Churned | F         |
| 8 | K110390     | yes        | MALE    | 40.... | 7.000         | S80     | Active  | T         |

There are only records with known data.

This completes the demo for this module. You will find the solution results in the file **demo\_working\_with\_modeler\_completed.str**, located in the **02-Working\_with\_Modeler\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Is the following statement true or false? To open a stream file you need to select the Stream node from the Sources palette, place the Stream node on the stream canvas, edit it, and select the stream file.

- A. True
- B. False

Question 2: Which of the following is the correct statement?

- A. A stream represents a flow of data from data reading, through a number of manipulations, to running an analysis.
- B. A stream is a data file in the native MODELER format.
- C. A stream is an output object produced by MODELER.
- D. A stream is a data-mining algorithm.

Question 3: Is the following statement true or false? A stream must contain a node from each palette.

- A. True
- B. False

Question 4: Is the following statement true or false? A comment can be added to the stream canvas by right-clicking an empty area on the stream canvas and then selecting New comment from the context menu.

- A. True
- B. False

Question 5: Which of the following is the correct statement?

- A. Suppose that you want to remove duplicate records in a dataset. The appropriate node is found on the Field Ops palette.
- B. Suppose that you want to reclassify a field's values (for example, 7 age categories have to be recoded into 3 age categories). The appropriate node is found on the Field Ops palette.
- C. Suppose that you want to import data from an IBM SPSS Statistics file. The appropriate node is found on the Field Ops palette.
- D. Suppose that you want to sort records (for example, the youngest person must appear first in the dataset, the oldest person must appear last). The appropriate node is found on the Field Ops palette.

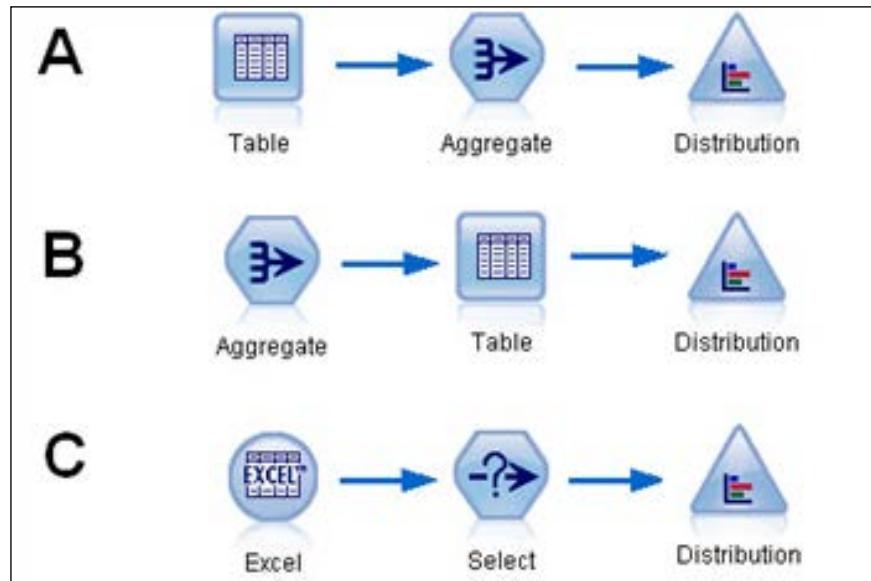
Question 6: Is the following statement true or false? More than one stream can be open at the same time.

- A. True
- B. False

Question 7: Is the following statement true or false? A Select node does not necessarily have to be placed on the stream canvas manually, but can also be generated from a Table output window.

- A. True
- B. False

Question 8: Fill in the blank. Refer to the streams depicted below. Only stream \_\_\_\_\_ (A/B/C) can be created in MODELER.



Question 9: Which of the following statement are correct? A node can be placed on the stream canvas by:

- A. selecting the node from the Insert entry in the main menu
- B. selecting the node from the appropriate palette
- C. selecting the node from the Manager pane on the Outputs tab
- D. clicking the Run button in the toolbar

Question 10: Is the following statement true or false? It is possible to save a node and load that node later.

- A. True
- B. False

Question 11: Which of the following statements are correct?

- A. A SuperNode consists of a number of nodes that have been condensed into a single node.
- B. A SuperNode is represented by a star icon.
- C. It is possible to have a SuperNode within another SuperNode (in other words, to nest SuperNodes).
- D. Modeling nodes can be included in SuperNodes.
- E. It is possible to save a SuperNode.

Question 12: Which of the following is the correct statement? Refer to the figure that follows. In the Table output window, you have selected the value MALE for gender and the value Active for churn. Then, you select Generate\Select Node ("And"). The condition generated is:

- A. gender = "male" and churn = "active"
- B. gender = "MALE" and churn = "Active"
- C. gender = "MALE" or churn = "Active"
- D. This will issue an error message, because only values for the same record can be selected

The screenshot shows the 'Table' tab selected in the output window. The main area displays a table with columns 'customer' and 'ID'. The data rows are numbered 1 to 6, with IDs K100150, K100180, K100690, K101420, K104780, and K105440 respectively. To the right of the table, a context menu is open under the 'Generate' tab. The menu items are: 'Select Node ("Records")', 'Select Node ("And")' (which is highlighted in yellow), 'Select Node ("Or")', 'Derive Node ("Records")', 'Derive Node ("And")', and 'Derive Node ("Or")'. Below the menu, there is a preview table showing 'gender' and 'churn' columns. The first five rows show 'MALE' in the 'gender' column and 'Churned' in the 'churn' column. The last row shows 'FEMALE' in the 'gender' column and 'Active' in the 'churn' column.

|   | customer | ID |
|---|----------|----|
| 1 | K100150  |    |
| 2 | K100180  |    |
| 3 | K100690  |    |
| 4 | K101420  |    |
| 5 | K104780  |    |
| 6 | K105440  |    |

|        | gender  | churn |
|--------|---------|-------|
| MALE   | Churned |       |
| FEMALE | Active  |       |
| FEMALE | Churned |       |

Question 13: Which of the following is the correct statement? Suppose that you have described the objectives of your data-mining project in a text document. Furthermore, suppose you want to store this document in MODELER. What is the appropriate place to store the file?

- A. In the Outputs Manager tab.
- B. In the Project pane, CRISP-DM tab, in the folder Business Understanding.
- C. In a comment added to the stream.
- D. In a comment added to the node that reads the data.

## Answers to questions:

Answer 1: B. False. You open a stream by selecting File\Open Stream from the main menu. The Sources palette only provides nodes to import data files.

Answer 2: A. A stream represents a flow of data from data reading, through a number of manipulations, to running an analysis.

Answer 3: B. False. A stream does not necessarily have a node from each palette.

Answer 4: A. True. A stream comment can be added by a right-click on the stream canvas, then selecting New comment from the context menu.

Answer 5: B. Reclassifying a field's values is a field operation, so the corresponding node can be found in the Field Ops palette.

Answer 6: A. True. More than one stream can be open at the same time. Switch from one stream to another by selecting the stream in the Manager pane - Streams tab. Having multiple streams open makes it easy to copy from one stream to another.

Answer 7: A. True. The best way to add a Select node to your stream is to generate it from the Table output window.

Answer 8: B. Only stream C starts with a data import node, and ends with a terminal node (and a record operation in between).

Answer 9: A, B.

Answer 10: A. True. Nodes can be saved and loaded later.

Answer 11: A, B, C, D, E. All statements are correct.

Answer 12: B.

Answer 13: B. Project information can be saved in the Project pane, CRISP-DM tab.

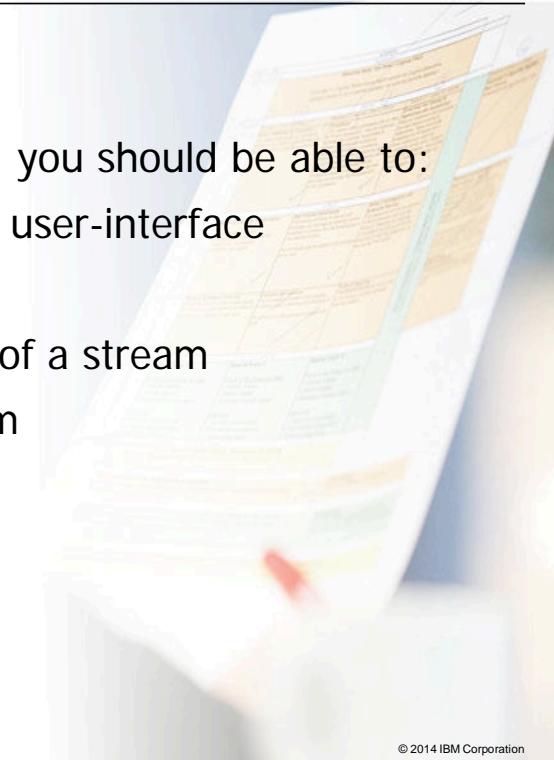
Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - describe the MODELER user-interface
  - work with nodes
  - run a stream or a part of a stream
  - open and save a stream
  - use the online Help

© 2014 IBM Corporation



In this module you were introduced to MODELER. You now should be able to find your way in MODELER.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Workshop 1

## Working with Modeler



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the **C:\Train\0A005** folder and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following files are used in this workshop:

- **workshop\_data\_working\_with\_modeler.xls** – a Microsoft Excel file that contains data from 30,000 customers of ACME
- **02-Working\_with\_Modeler\Start Files**  
**\workshop\_working\_with\_modeler.str** – a MODELER stream, importing the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Workshop 1: Working with Modeler

Practice your skills of building and running streams in this workshop. Initially no data is used, so that you can focus on MODELER's interface. In the last two tasks you will use a dataset taken from a firm named ACME (a fictitious company selling sport products via the Web and mail campaigns).

- Create a stream, that:
  - imports data from an IBM SPSS Statistics (.sav) file (at this moment, no specific data file is used, so do not specify a data file)
  - selects records from the imported IBM SPSS Statistics file
  - sorts records
  - derives a new field
  - requests a Histogram graph for the just derived field
  - derives a second field
  - requests a Distribution graph for the second derived field
  - exports the data to a Microsoft Excel file
- Change the stream just built:
  - remove the second Derive node, but ensure that the stream still flows from the data source to the Excel export node
  - remove the Distribution node that you had on the Derive node that you just removed
  - export the data to an IBM SPSS Statistics file (next to the export to Excel)
  - add a comment to the Derive node
  - add a freestanding comment to the stream, such as your name and the date that you created the stream
- Save the stream; name it **my\_first\_workshop\_stream.str**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- Create a stream in a new window. This stream resembles in a high degree the stream that you have just created. The only difference is that only a sample of records should be exported to the IBM SPSS Statistics file. To create this stream, copy the stream that you already have, paste it into a new window, and change the stream so that only a sample of records is exported to the IBM SPSS Statistics file.
- Make the stream neat by encapsulating the Select and Sort node into a SuperNode. Also, annotate the SuperNode. Zoom in on the SuperNode to verify its content, and when the content is okay, zoom out of the SuperNode.

In the next tasks, select records by generating a Select node from the Table output window. The tasks use **workshop\_working\_with\_modeler.str** (located in the **02-Working\_with\_Modeler\Start Files** sub folder).

- Open **workshop\_working\_with\_modeler.str**, located in the **02-Working\_with\_Modeler\Start Files** sub folder, and then run the Table node to get a feel for the data.

How many records do you have?

Select only those customers that were in the test mailing.

How many records are selected?

- Continue to work with the dataset that includes only those customers that were in the test mailing. Notice that gender is missing (UNKNOWN or \$null\$) for some customers. Select only female and male customers (so, discard customers with missing data for gender).

How many records are selected?

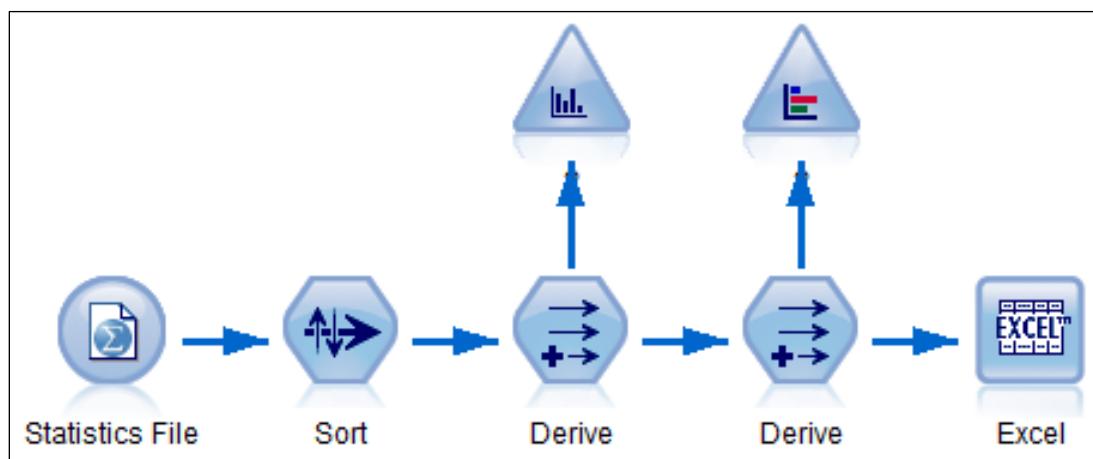
## Workshop 1: Tasks and Results

Task 1. Create a stream that reads data from IBM SPSS Statistics and exports data to Microsoft Excel.

- Place the following nodes on the stream canvas:
  - Statistics File (Sources palette)
  - Select (Record Ops palette)
  - Sort (Record Ops palette)
  - Derive (Field Ops palette)
  - Histogram (Graphs palette)
  - Derive node (Field Ops palette)
  - Distribution (Graphs palette)
  - Excel (Export palette)

Ensure the nodes are connected to form a stream.

The result appears as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

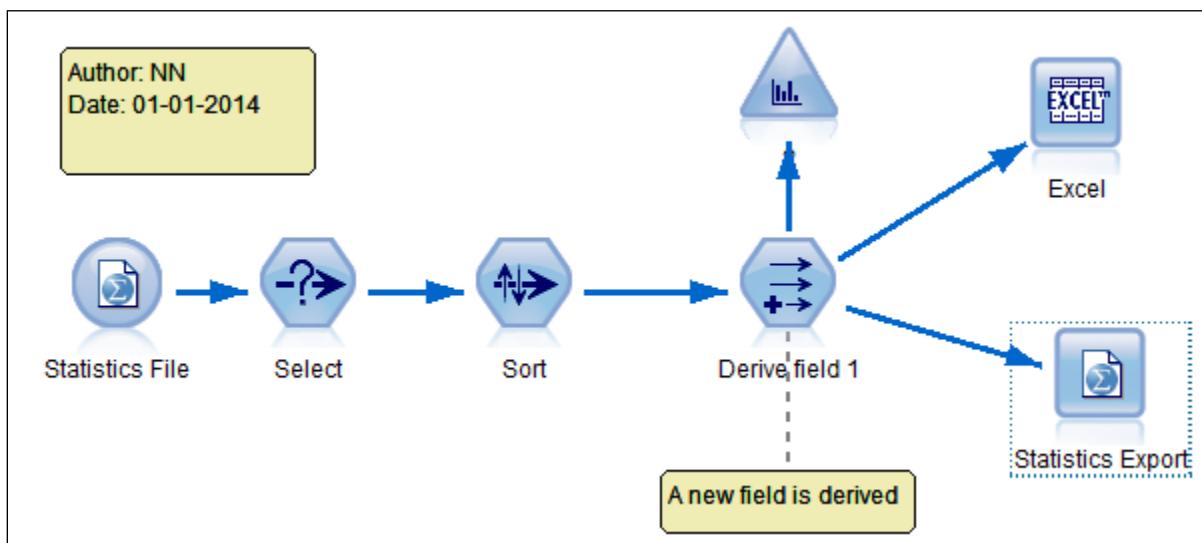
© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Task 2. Change a stream.

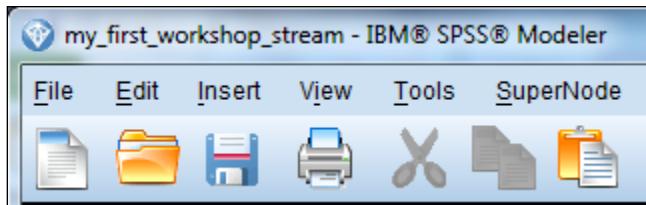
- Delete the second Derive node
- Delete the Distribution node
- Connect the Derive node to the Excel node
- Add a Statistics Export node downstream from the Derive node
- Right-click the Derive node and add comment
- Right-click an empty area on the stream canvas and add comment

Your stream should appear as:



## Task 3. Save a stream.

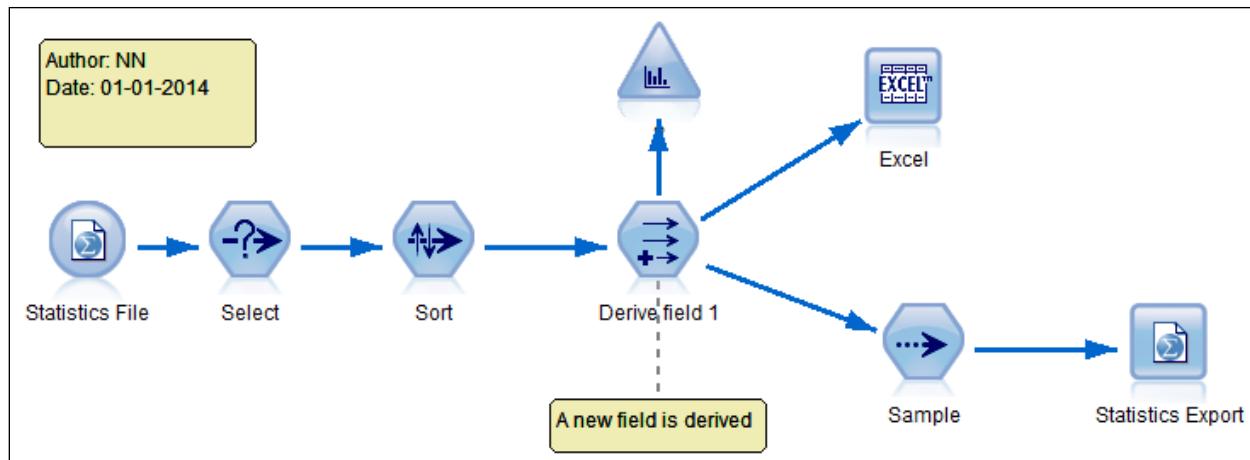
- Select **File\Save Stream** from the main menu, and type **my\_first\_workshop\_stream.str**. The title bar will show the file name.



## Task 4. Create a new stream by copying and pasting from an existing stream.

- Select all nodes (for example, by selecting **Edit\Select All** in the main menu).
- **Copy** the selected nodes.
- Select **File\New Stream** to open a new window.
- **Paste** the content of the clipboard to the stream canvas.
- Insert a **Sample** node upstream from the **Statistics Export** node.

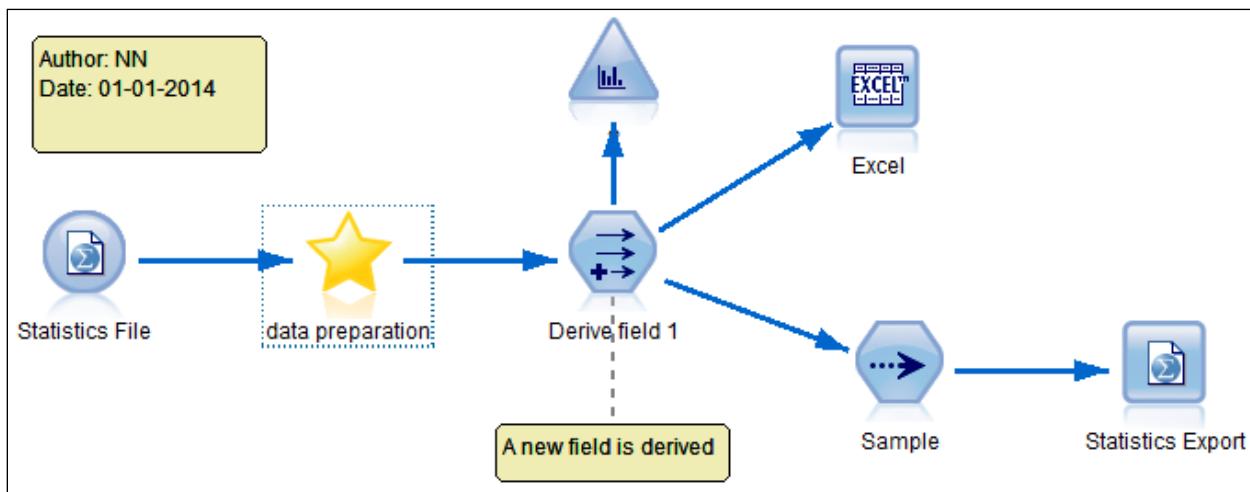
Your stream should appear as:



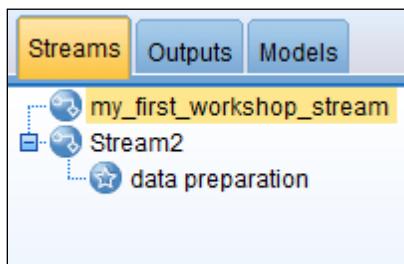
## Task 5. Make the stream neat by using a SuperNode.

- Select the **Select** and **Sort** node (for example, by dragging a rectangle around them).
- **Right-click** one of the selected nodes, and then choose **Create SuperNode** from the context menu.
- **Right-click** the **SuperNode**, select **Rename and Annotate** from the context menu, and then type the text.

Your stream should appear as follows:



Zoom in by double-clicking the **SuperNode**. Then, zoom out by clicking the Zoom out of SuperNode button. Alternatively, use the **Streams** tab to zoom out of the SuperNode:



## Task 6. Generating a Select node from Table output.

- Open **workshop\_working\_with\_modeler.str**, and then run the **Table** node.
- In the **Table** output window, select **yes** for **has\_received\_test\_mailing**, and then choose **Generate\Select Node ("And")** from the main menu.
- Add the generated **Select** node downstream from the **Excel** source node, and then run a **Table**. This will show that 10,000 records are selected.

## Task 7. Further record selections.

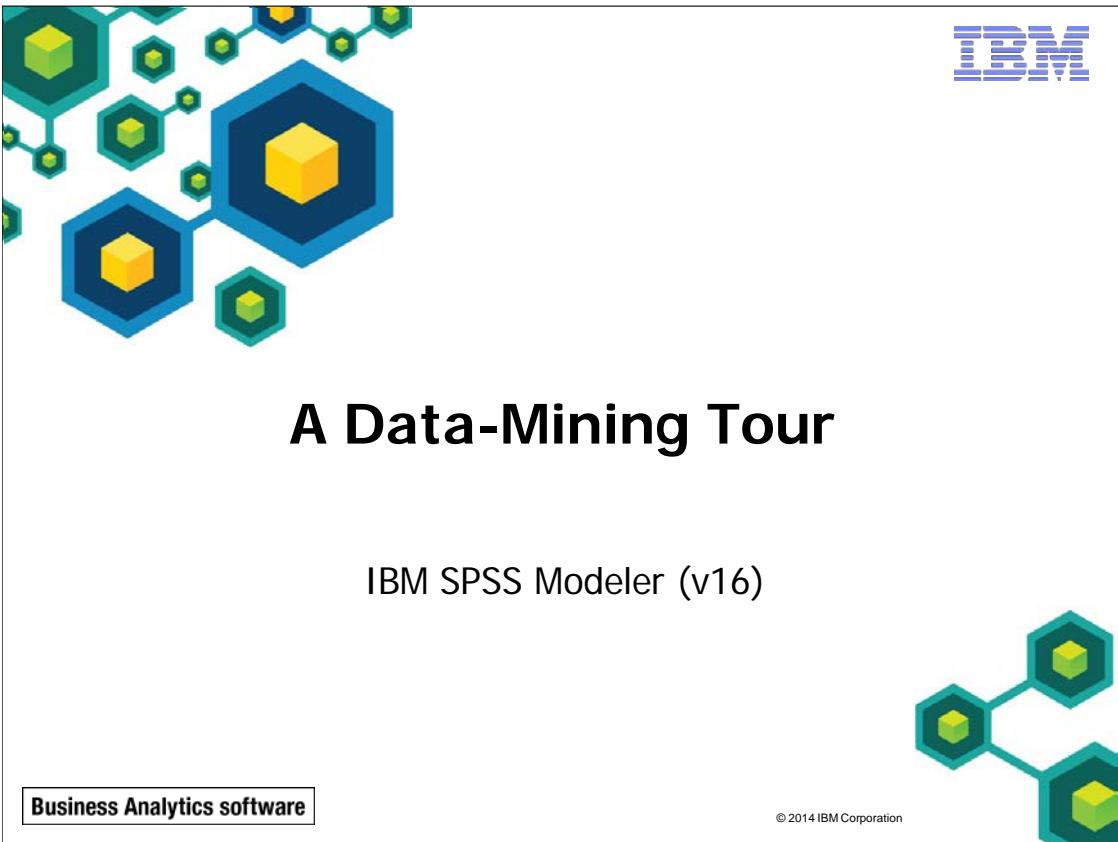
- To select men and women, return to the **Table** output window, and select values **female** and **male** (use the CTRL-key for a multiple selection). Then choose **Generate\Select Node ("Or")**, add the generated **Select** node downstream from the first **Select** node, and then run a **Table** node. This will show that you have 9,980 records.

The stream **workshop\_working\_with\_modeler\_completed.str**, located in the **02-Working\_with\_Modeler\Solution Files** sub folder, provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The banner features a white background with a decorative pattern of blue and green hexagons containing yellow cubes. In the top right corner is the IBM logo. Below the pattern, the text "A Data-Mining Tour" is displayed in a large, bold, black sans-serif font. Underneath that, "IBM SPSS Modeler (v16)" is shown in a smaller, regular black font. At the bottom left, a small rectangular box contains the text "Business Analytics software". On the right side, there is a graphic of three interconnected hexagons, each with a yellow cube inside, connected by teal lines.

IBM

A Data-Mining Tour

IBM SPSS Modeler (v16)

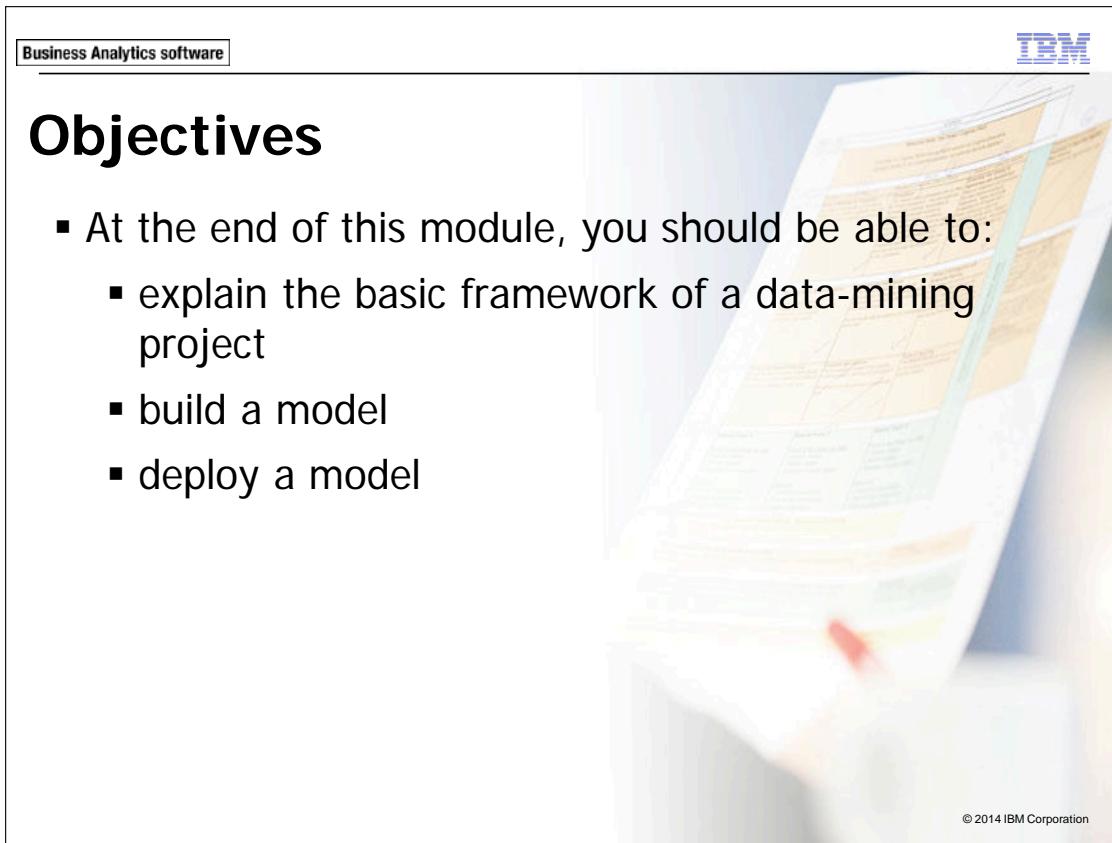
Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



**Business Analytics software**

IBM

# Objectives

- At the end of this module, you should be able to:
  - explain the basic framework of a data-mining project
  - build a model
  - deploy a model

© 2014 IBM Corporation

In the *Introduction to Data Mining* module you were introduced to what data-mining encompasses. In the *Working with Modeler* module you acquired skills to work with MODELER. In this module, these two pieces come together.

The analysis presented in this module happens at the end of a project, and the modules that follow will start from scratch and work towards this point step-by-step. The rationale for starting at the end of the project is that the analysis will provide context and understanding of the earlier stages, so the analysis provides the frame of reference for the next modules in this course and gives direction to decisions to be made in the earlier stages of the analysis.

Note: The objective of this module is not to provide an overview of modeling techniques in MODELER. Refer to the *Introduction to Modeling* module for details.

Before reviewing this module, you should be familiar with the following topics:

- the CRISP-DM process model
- MODELER streams, nodes, and palettes

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## The Basic Framework of a Data-Mining Project

- Predict the behavior of future cases, given the past
  - build a model on historical data
  - apply that model to future cases

© 2014 IBM Corporation



The data-mining goal of any project is to predict the behavior of future cases, given what the past has learned. Therefore, a data-mining project will consist of two parts:

- build a model on historical data (also referred to as analytical data or modeling data)
- apply the model to future cases: the deployment data (also referred to as operational data or scoring data)

The deployment data is leading in the analysis. If a model is to be deployed, a predictor that is used in the model must be available in the deployment data, otherwise it is of no use to include such a field in a model in the first place. For example, suppose that the historical data includes a region field, while this field is not available in the deployment data. You can include this field in a model, but in the end that model cannot be deployed because the region field is missing in the deployment dataset.

# A Business Case

- Scenario:
  - a bank needs to reduce the risk that a loan is not paid back
- Approach:
  - use historical data to build a model for risk
  - apply the model to customers or prospects who apply for a loan

© 2014 IBM Corporation



Suppose that a bank experiences problems with customers who do not pay back their loan, which costs the company a significant amount of money.

To reduce the risk that loans are not paid back, the bank will use modeling techniques on its historical data to find groups of high-risk customers (high risk of not paying back the loan). If a model is found, then the bank will use that model to attach a risk score to those who apply for a loan. When the risk of not paying back the loan is too high, the loan will not be granted.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## A Business Case: Historical Data

| <b>id</b> | <b>gender</b> | <b>howpaid</b> | <b>age_category</b> | <b>has_children</b> | <b>loans_paid_back</b> |
|-----------|---------------|----------------|---------------------|---------------------|------------------------|
| 100001    | female        | monthly        | junior              | no                  | yes                    |
| 100003    | male          | monthly        | junior              | no                  | yes                    |
| 100004    | female        | monthly        | junior              | no                  | yes                    |
| 100005    | male          | monthly        | junior              | no                  | yes                    |
| 100008    | male          | monthly        | junior              | no                  | yes                    |
| 100011    | male          | monthly        | junior              | no                  | yes                    |
| 100012    | female        | monthly        | junior              | no                  | yes                    |
| 100013    | female        | monthly        | junior              | no                  | no                     |

© 2014 IBM Corporation



This slide shows an example of a historical dataset. The dataset includes demographic information and a field that indicates whether the customer has paid back the loan. Most customers (shown on this slide) have paid back their loan. One customer (id 100013) did not pay back her loan.

Typically not all records will be used for modeling, but a sample will be drawn on which models are built, and their accuracy will be compared on the records that were not in the sample. Also in this example it is assumed that the dataset for modeling is a sample of customers.

## A Business Case: Initial Results

| age_category    |          |        |        |        |
|-----------------|----------|--------|--------|--------|
| loans_paid_back |          | junior | senior | Total  |
| no              | Count    | 504    | 55     | 559    |
|                 | Column % | 29.066 | 7.628  | 22.770 |
| yes             | Count    | 1230   | 666    | 1896   |
|                 | Column % | 70.934 | 92.372 | 77.230 |
| Total           | Count    | 1734   | 721    | 2455   |
|                 | Column % | 100    | 100    | 100    |

Cells contain: cross-tabulation of fields (including missing values)

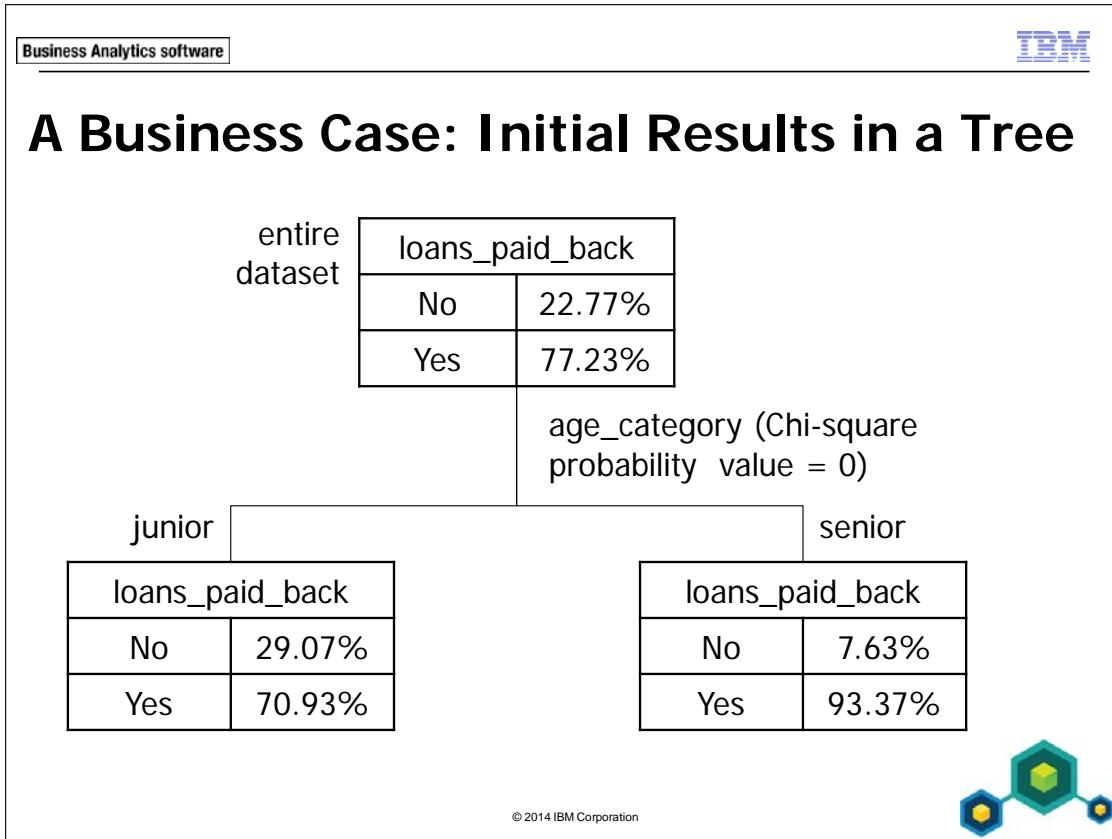
Chi-square = 133.086, df = 1, probability = 0

© 2014 IBM Corporation



Initially the analysis will have an exploratory character, to get a good feel for which fields are important in predicting if loans are paid back. For example, you may want to investigate the relationship between paying back loans and age category.

This slide shows a cross tabulation of the two fields; 29.1% of the juniors in the sample did not pay back their loan, while this percentage was 7.6% for seniors in the sample. Note, that if you work with sample data you will always have a difference between the group percentages. The question is if the difference that you observed in the sample can be attributed to chance (the sampling process) or that the sample difference reflects a difference in the population of all customers. A statistical test such as a Chi-square test answers this question. This test gives the probability that the difference that you observed can be attributed to chance (caused by the sampling process), which in this example would be a probability equal to 0. In other words, it is impossible that the difference that you have observed is caused by chance, and there must be another reason why you have observed a difference: there is a difference between juniors and seniors in the population of all customers.



The result can also be presented in a tree. The root of the tree shows the overall distribution. Here, it tells you that 22.77% of the loans have not been paid back. The split on age category shows that this was 29.1% for juniors, and 7.6% for seniors. Also, the Chi-square test shows that this difference cannot be attributed to chance, because the probability is 0 (as indicated by the so-called Adjusted P-value). This is in agreement with the results in the previous cross tabulation.

## A Business Case: Further Results

| has_children    |          |        |        |        |
|-----------------|----------|--------|--------|--------|
| loans_paid_back |          | no     | yes    | Total  |
| no              | Count    | 76     | 428    | 504    |
|                 | Column % | 17.234 | 33.101 | 29.066 |
| yes             | Count    | 365    | 865    | 1230   |
|                 | Column % | 82.766 | 66.899 | 70.934 |
| Total           | Count    | 441    | 1293   | 1734   |
|                 | Column % | 100    | 100    | 100    |

Cells contain: cross-tabulation of fields (including missing values)

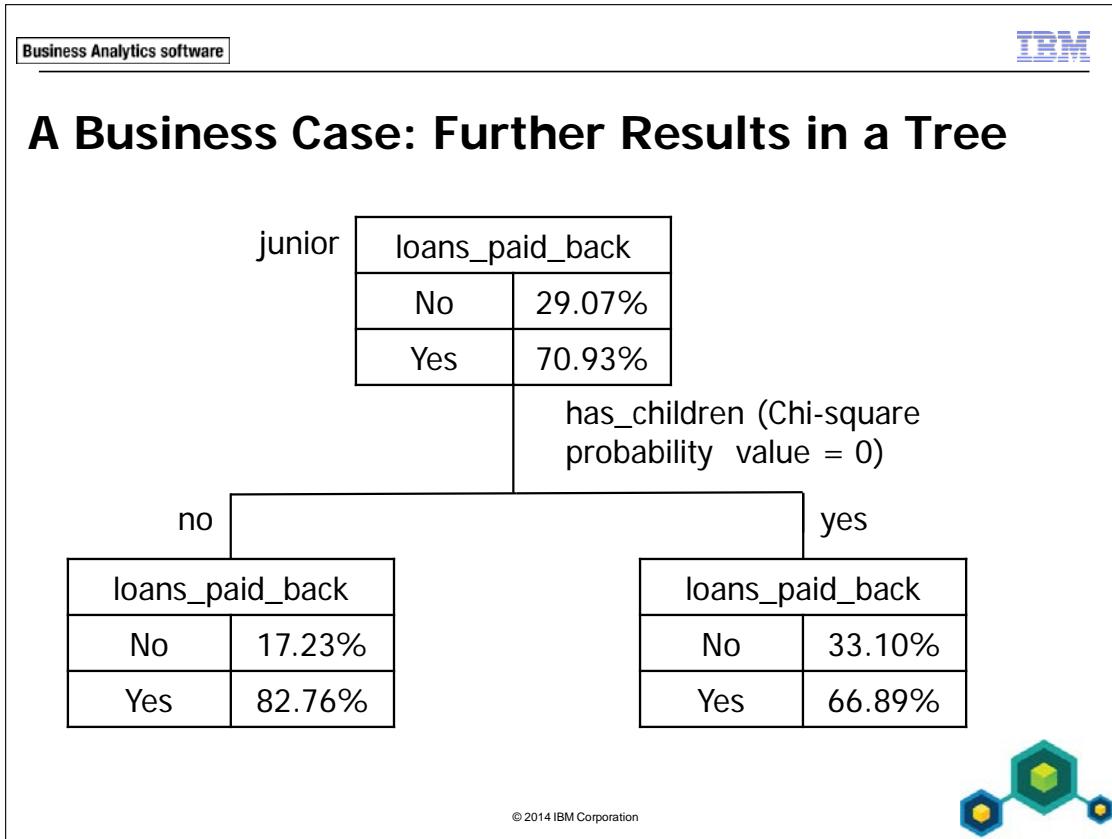
Chi-square = 40.159, df = 1, probability = 0

© 2014 IBM Corporation



Having concluded that juniors have more risk of not paying back loans than seniors, the next step would be to zoom in on one of these groups and to examine if there are further differences within that group. For example, within the group of juniors, do people without children show better or worse pay back rates than people with children?

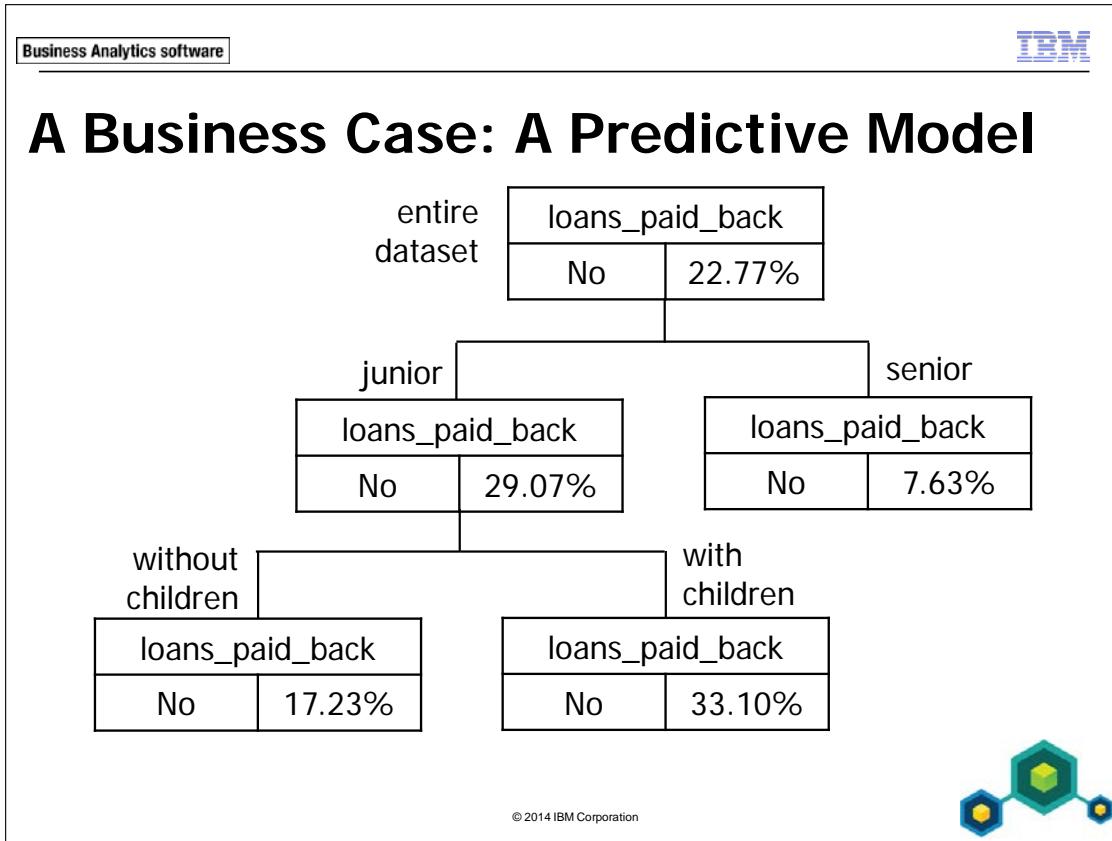
This slide shows the cross tabulation of loans\_paid\_back by has\_children, for the junior group. About 17.2% of the juniors without children did not pay back their loans, against 33.1% for the juniors with children. The Chi-square statistic tells you that this difference cannot be attributed to chance alone (the probability that this difference is a sample result only equals 0). So, within the group of juniors, it matters if one has children or not when it comes to paying back loans.



Again, the result can be presented in a tree, now starting from the junior group.

Further analyses can be run to search for groups at risk of not paying back their loans. For example you can run the same analysis, but then for seniors. Or you can focus on the group of juniors with children and examine if there are further differences in pay back rates between, say, income categories.

Running all of these analyses is a never ending story, and it would be better to have a way to automate the analyses. This is the point where a modeling technique comes in. In fact, the step-by-step analysis that was outlined here is known as a CHAID analysis (the first two letters indicate that it is based on the Chi-square test).



An example of a tree is depicted on this slide (with the focus on the persons that did not pay back their loan). There is no sub tree for the senior group because the CHAID algorithm did not find any other field that made a difference in paying back loans for this group.

In this example, the tree allocates a customer to one of three groups which correspond to the terminal nodes in the tree. Terminal nodes in this context do not refer to nodes in a MODELER stream, but to the end points in the tree. The terminal nodes define three groups:

- Seniors: 7.63% did not pay back their loan
- Juniors without children: 17.23% did not pay back their loan
- Juniors with children: 33.10% did not pay back their loan

These rules make up a predictive model: if you know a customer's age category, and if you know if he or she has children, you know the probability that this customer will not pay back his or her loan. For example, if you know that you are dealing with a junior with children, you know that this person will not pay back his or her loan with a probability of 0.331 (33.1%). Also, you know that this person is more at risk of not paying back the loan than a senior.

# Deploying the Model: Scoring Records

| <b>id</b> | <b>gender</b> | <b>howpaid</b> | <b>age_category</b> | <b>has_children</b> | <b>predicted_category_for_loans_paid_back</b> | <b>confidence_for_the_prediction</b> |
|-----------|---------------|----------------|---------------------|---------------------|---|--------------------------------------|
| 1         | female        | monthly        | junior              | no                  | yes   | .827                                 |
| 2         | male          | monthly        | junior              | yes                 | yes   | .668                                 |
| 3         | male          | monthly        | junior              | yes                 | yes   | .668                                 |
| 4         | female        | monthly        | senior              | yes                 | yes   | .923                                 |

© 2014 IBM Corporation



You can use the model to attach a risk score to each applicant. For example, if a senior comes in, you can predict that the probability that he or she does not pay back the loan is 0.077. Or, equivalently, the probability that he or she will pay back equals .923.

You can also have a decision rule in place to make a yes/no decision about whether an applicant will be granted the loan. Only if the probability of paying back the loan is higher than a certain cut-off value, the applicant will be granted the loan.

This slide shows an example of how MODELER scores records. MODELER uses a cut-off value of 0.5 to classify a customer into a yes (will pay back) or no (will not pay back) category, and adds the predicted category to the dataset. Also, it will add the confidence, the probability that the predicted category is correct, to the dataset. For example, id #4 is a senior and is predicted to pay back her loan with a probability of .923. Notice that the loans\_paid\_back field is missing in the deployment dataset, because these persons are applying for a loan and have no history yet.

## A Data-Mining Project in MODELER

- A data-mining project will show two parts:
  - one branch where you build a model
  - one branch where you will apply the model to customers or prospects

© 2014 IBM Corporation



In this example a predictive model was built with historical data, after which the model was applied to new cases. Your work in MODELER will reflect these two parts. You will have:

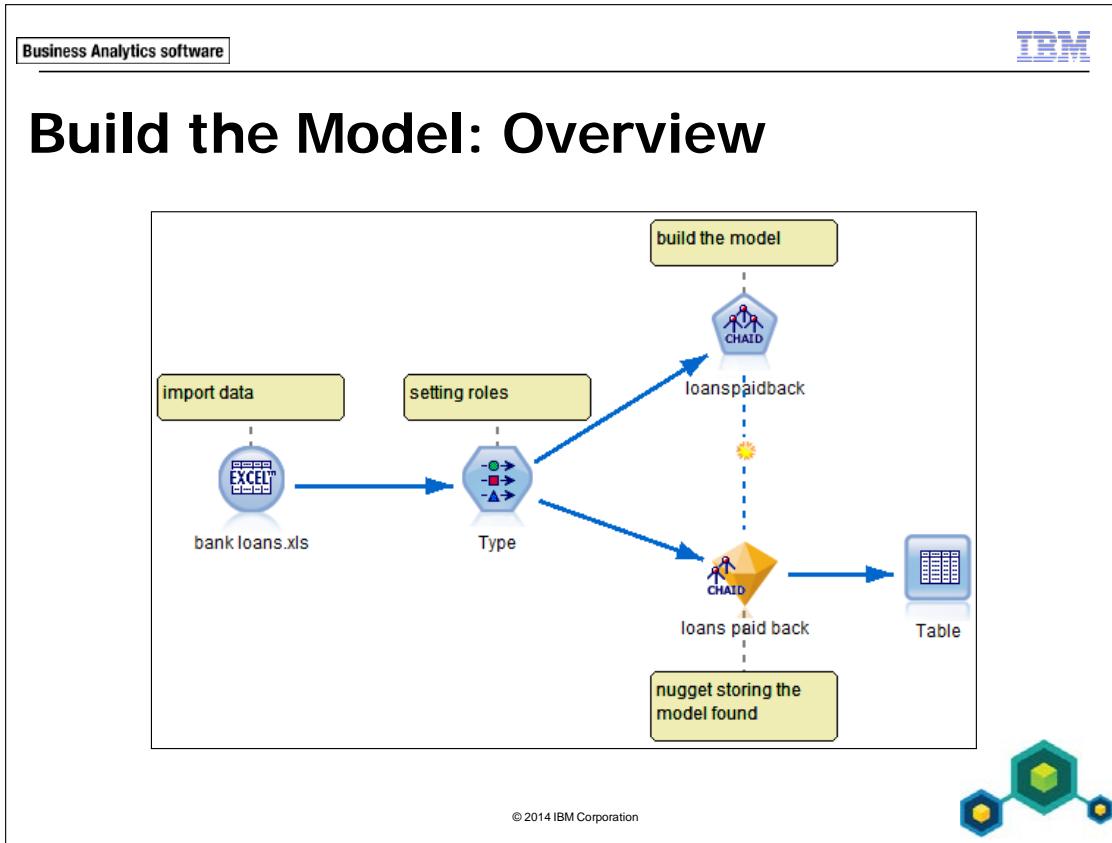
- a stream where the model is built
- a stream where the model is deployed

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

3-14

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



This slide gives a high level overview of a modeling stream (without data preparation or data exploration):

1. A data source node imports the historical data.
2. A Type node sets the fields' roles (predictors are selected, and the target field is specified; refer to the next slide for details).
3. A modeling node is selected (here: CHAID).
4. When the modeling node has been executed, a diamond, or model nugget, is generated. The model nugget stores the model: when you use CHAID, the model nugget stores the rules. When records pass through the model nugget, a predicted value and confidence score is added for each record.
5. A Table node is added downstream from the model nugget, to view the fields that were added by the model nugget.

Business Analytics software

IBM

## Build the Model: Setting Roles in a Type Node

| Field           | Role   |
|-----------------|--------|
| gender          | Input  |
| howpaid         | Input  |
| age category    | Input  |
| has children    | Input  |
| loans paid back | Target |
| # id            | None   |

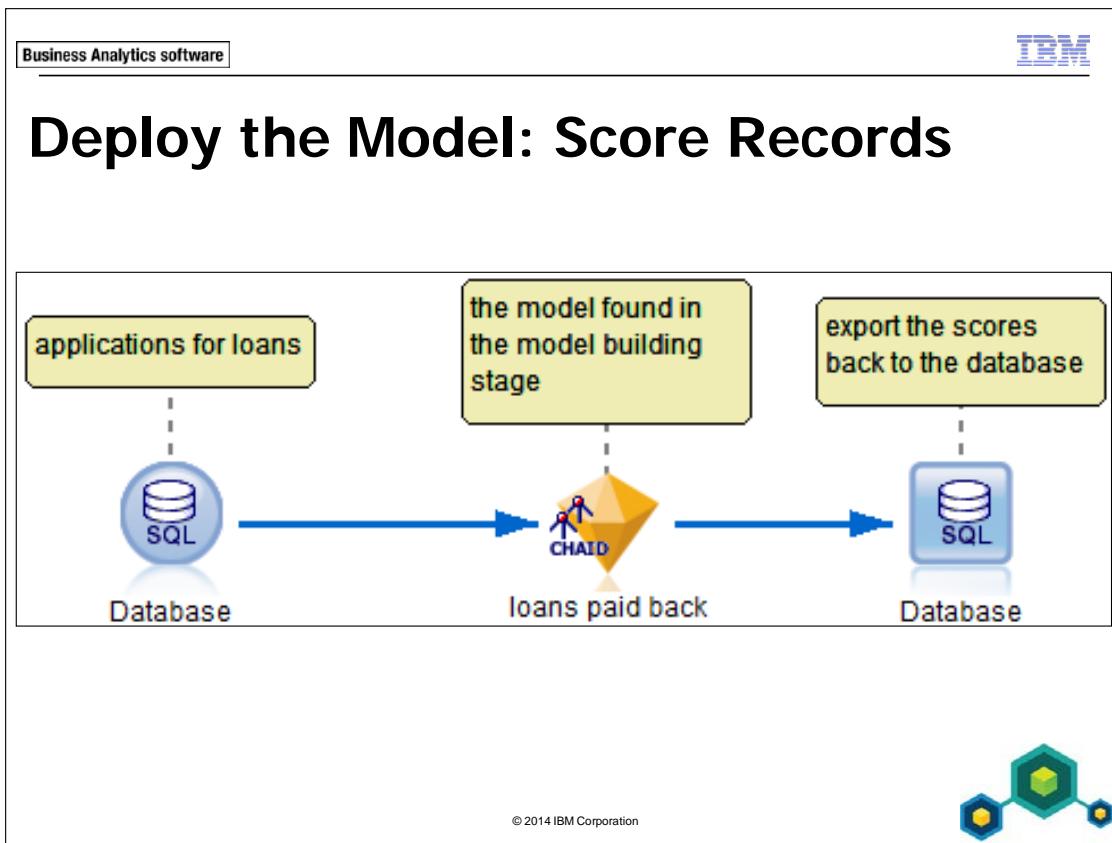
© 2014 IBM Corporation



In the above stream, one of the most important nodes is the Type node. The Type node will be presented in detail in the *Collecting Initial Data* module and in the *Understanding your Data* module in this course. In this module the presentation is limited to the Role column. In the Role column, set the role to:

- Input, for fields that you want to use as predictors
- Target, for the field that you want to predict
- None, for fields that are not used in model building

The Type node is located in the Field Ops palette.



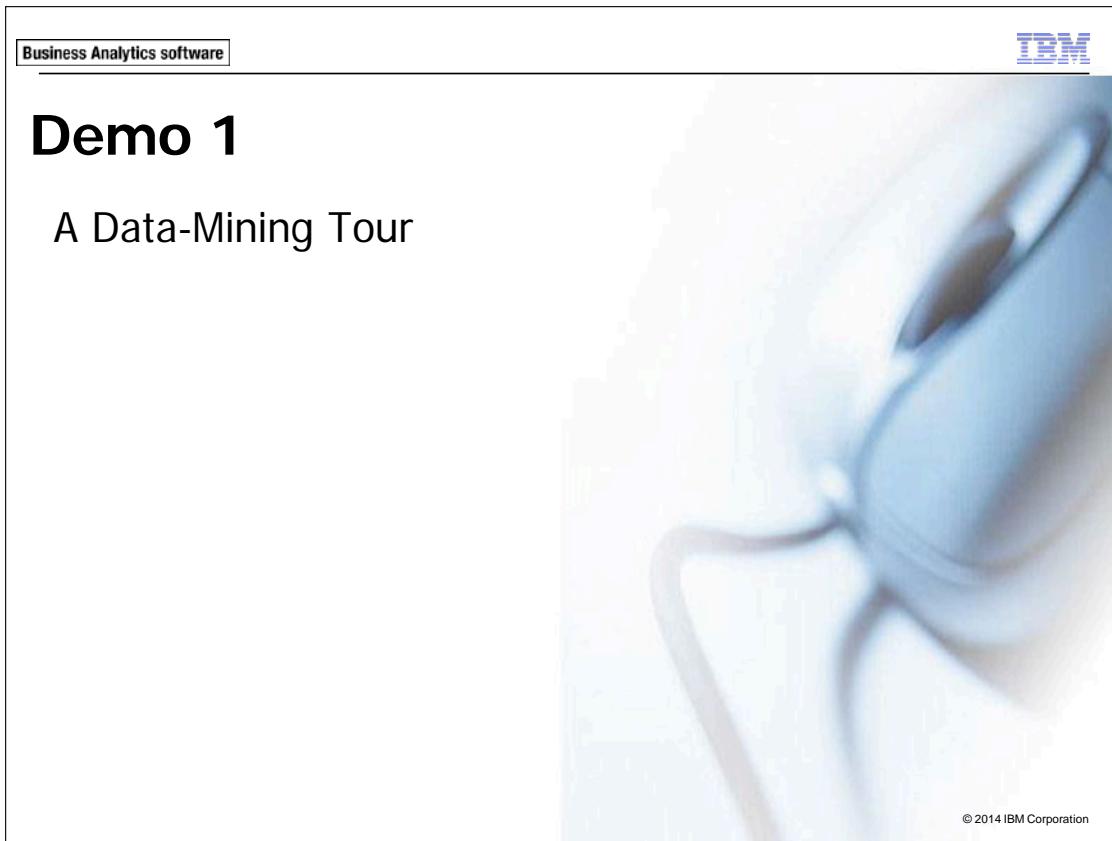
If the model is satisfactory, the model is deployed. This slide shows a typical deployment branch (leaving aside data preparation and data exploration).

1. The deployment data is imported.
2. The model nugget is added to the stream. When the records flow through the model nugget, the model nugget adds the prediction for each record to the dataset.
3. The dataset enriched with the fields added by the model nugget is exported.

When you want to export the scores so that another department can process the modeling results, it may suffice to export only the fields that are added by the model nugget and a field such as customer\_id. Also, you may want to rename the fields so that they have more appealing names than the names that you have in MODELER.

You can remove fields and rename fields in a Filter node , shown on this slide. The Filter node is located in the Field Ops palette (ensure that you select the Filter node, not the Filler node). The left column shows the field names as you have them in your dataset. Specify new names in the right column. You can exclude (or include an excluded field) by clicking the arrow. Clicking the Filter options menu button in the Filter node will provide you with many handy features.

Another easy-to-use node is the Sort node, to sort records. The Sort node is located in the Record Ops palette.



The slide features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The main title "Demo 1" is displayed prominently at the top, followed by the subtitle "A Data-Mining Tour". The background of the slide is a blurred image of a person wearing a bow tie. A small copyright notice "© 2014 IBM Corporation" is located in the bottom right corner of the slide area.

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the **C:\Train\0A005** folder and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following data files are used in this demo:

- **telco x modeling data.xls** – a Microsoft Excel file, storing historical data on customers of a (fictitious) telecommunications firm
- **telco x deployment data.xls** – a Microsoft Excel file, storing the current customers of the telecommunications firm

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

3-19

## Demo 1: A Data-Mining Tour

### Purpose:

**You are working as a data miner for a telecommunications firm. You have to identify customers who are likely to churn.**

Task 1. Build a model using historical data.

The file **telco x modeling data.xls** stores historical data. This dataset will be used to build a model to predict churn.

1. Place an **Excel** node (Sources palette) on the stream canvas.
2. Edit the **Excel** node, select **telco x modeling data.xls** (use default values for the import), and then click the **Preview** button.

A section of the results appears as follows:

| customer_id | data_known | gender | age    | tariff | dropped_calls | handset | peak_mii |
|-------------|------------|--------|--------|--------|---------------|---------|----------|
| K100010     | yes        | Male   | 46.... | CA...  | 1.000         | SOP10   | 36.1     |
| K100020     | yes        | Male   | 27.... | CA...  | 0.000         | SOP10   | 39.4     |
| K100030     | yes        | Male   | 39.... | CA...  | 2.000         | SOP20   | 72.6     |
| K100040     | yes        | Male   | 28.... | CA...  | 2.000         | SOP10   | 72.6     |
| K100050     | yes        | Male   | 47.... | CA...  | 0.000         | SOP10   | 40.6     |
| K100060     | yes        | Male   | 29.... | CA...  | 1.000         | SOP10   | 46.2     |
| K100070     | yes        | Male   | 38.... | CA...  | 1.000         | SOP20   | 56.3     |

The preview gives a first impression of the data. Some fields, such as gender and age, record background information. Other fields, such as dropped\_calls and handset, store information about phone calling.

Also notice the field data\_known, flagging if a customer's data are known.

3. Scroll to the **last fields** in the **Preview** output window.

A section of the results appears as follows:

| <b>gadget_B_revenues</b> | <b>gadget_C_revenues</b> | <b>gadget_D_revenues</b> | <b>churn</b> | <b>retention</b> |
|--------------------------|--------------------------|--------------------------|--------------|------------------|
| 0.000                    | 28.000                   | 0.000                    | Churned      | F                |
| 0.000                    | 0.000                    | 0.000                    | Churned      | F                |
| 23.000                   | 0.000                    | 35.000                   | Churned      | F                |
| 18.000                   | 0.000                    | 41.000                   | Churned      | F                |
| 0.000                    | 0.000                    | 0.000                    | Churned      | F                |
| 0.000                    | 0.000                    | 0.000                    | Churned      | F                |
| 0.000                    | 32.000                   | 0.000                    | Churned      | F                |
| 0.000                    | 0.000                    | 0.000                    | Churned      | F                |

The field of interest is **churn**. This field has the value Churned when the customer has churned, and has the value Active when the customer did not churn.

Notice the **retention** field. This field is F for churners, T for active customers. In other words, this field is the exact counterpart of the **churn** field and this field should be left out when you build a model to predict churn later.

You will view all data to explore the data further.

4. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Excel** dialog box.
5. Add a **Table** node (Output palette) downstream from the **Excel** node.
6. Run the **Table** node.

The title bar in the Table output window reads that there are 31,789 records in the dataset.

7. Scroll to the **last records** in the **Table** output window.

The result appears as follows:

| <b>customer_id</b> | <b>data_known</b> | <b>gender</b> | <b>age</b> | <b>tariff</b> | <b>dropped_calls</b> |
|--------------------|-------------------|---------------|------------|---------------|----------------------|
| Z327540            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |
| Z329100            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |
| Z344350            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |
| Z345840            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |
| Z359810            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |
| Z385060            | no                | \$null\$      | \$n...     | \$n...        | \$null\$             |

When data\_known equals no, all values are \$null\$. The value \$null\$ represents MODELER's undefined value.

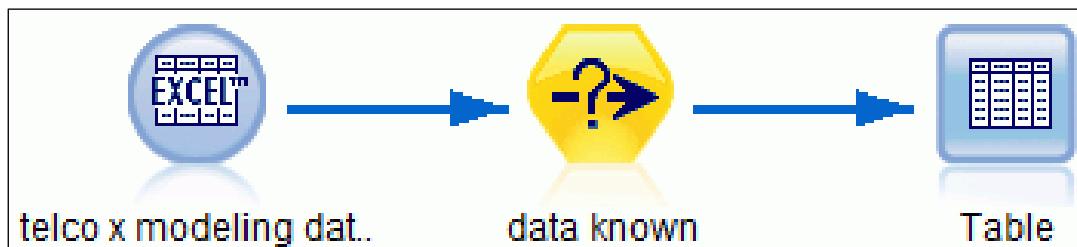
These records are not needed for modeling, so you will select the customers whose data are known. As demonstrated in the *Working with Modeler* module, the fastest way to select records is to generate the Select node from a Table output window. You will use this method to select the relevant records.

8. Scroll to the **beginning** of the **Table** output window, and then:
  - click the value **data\_known="yes"** in the **Table** output window
  - choose **Generate\Select Node ("And")**
  - click **OK** to close the **Table** output window
- A Select node, named (generated), is placed in the upper-left corner on the stream canvas.
9. Insert the generated node between the **Excel** source node and the **Table** node.

10. Edit the generated **Select** node, and then:

- click the **Annotations** tab
- select the **Custom** option
- type the text **data known**
- click **OK** to close the **Select** dialog box

The stream appears as follows:



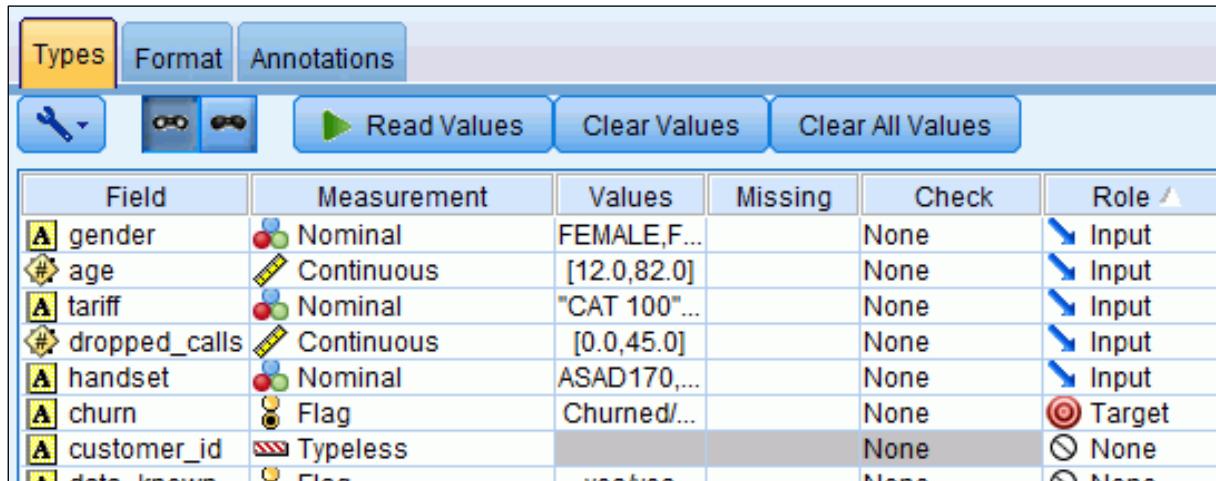
11. Run the **Table** node.

The title bar in the Table output window reads that there are 31,769 records left.

The next step is that you will build a model to predict churn. It is essential for modeling that predictors are selected and that the target field is specified. You will specify the predictors and the target in a Type node, in the Role column:

12. Add a **Type** node (Field Ops palette) downstream from the **Select** node named **data known**.
13. Edit the **Type** node, and then:
  - set the role of **gender**, **age**, **tariff**, **dropped\_calls**, and **handset** to **Input**
  - set the role of **churn** to **Target**
  - set the role for **all other fields** to **None** (fields such as **customer\_id**, **data\_known**, and **retention** are not relevant; other fields are candidate-predictors but are excluded for the sake of a simple model)
  - Click **Read Values** to have MODELER read the data
  - Click in the **Role** column header, so that the fields are sorted according to their role

A section of the Type dialog box appears as follows:



The screenshot shows the 'Types' tab of the Type dialog box. At the top, there are tabs for 'Types', 'Format', and 'Annotations'. Below the tabs are buttons for 'Edit', 'Format', 'Read Values', 'Clear Values', and 'Clear All Values'. The main area is a table with columns: Field, Measurement, Values, Missing, Check, and Role.

| Field         | Measurement | Values       | Missing | Check | Role   |
|---------------|-------------|--------------|---------|-------|--------|
| gender        | Nominal     | FEMALE,F...  |         | None  | Input  |
| age           | Continuous  | [12.0,82.0]  |         | None  | Input  |
| tariff        | Nominal     | "CAT 100"... |         | None  | Input  |
| dropped_calls | Continuous  | [0.0,45.0]   |         | None  | Input  |
| handset       | Nominal     | ASAD170,...  |         | None  | Input  |
| churn         | Flag        | Churned/...  |         | None  | Target |
| customer_id   | Typeless    |              |         | None  | None   |
| data_known    | Flag        |              |         | None  | None   |

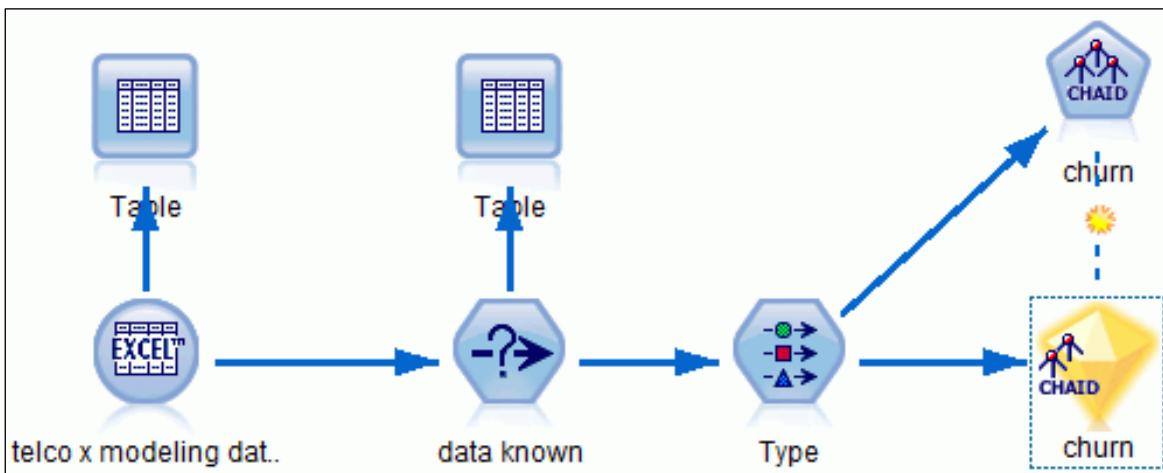
14. Click **OK** to close the **Type** dialog box.

Having specified predictors and target, the next step is to build a model, using one of the modeling nodes. CHAID will be used for modeling.

15. Add a **CHAID** node (Modeling palette) downstream from the **Type** node (the CHAID node will be labeled with the target field, **churn**).

## 16. Run the CHAID node.

The stream canvas appears as follows:

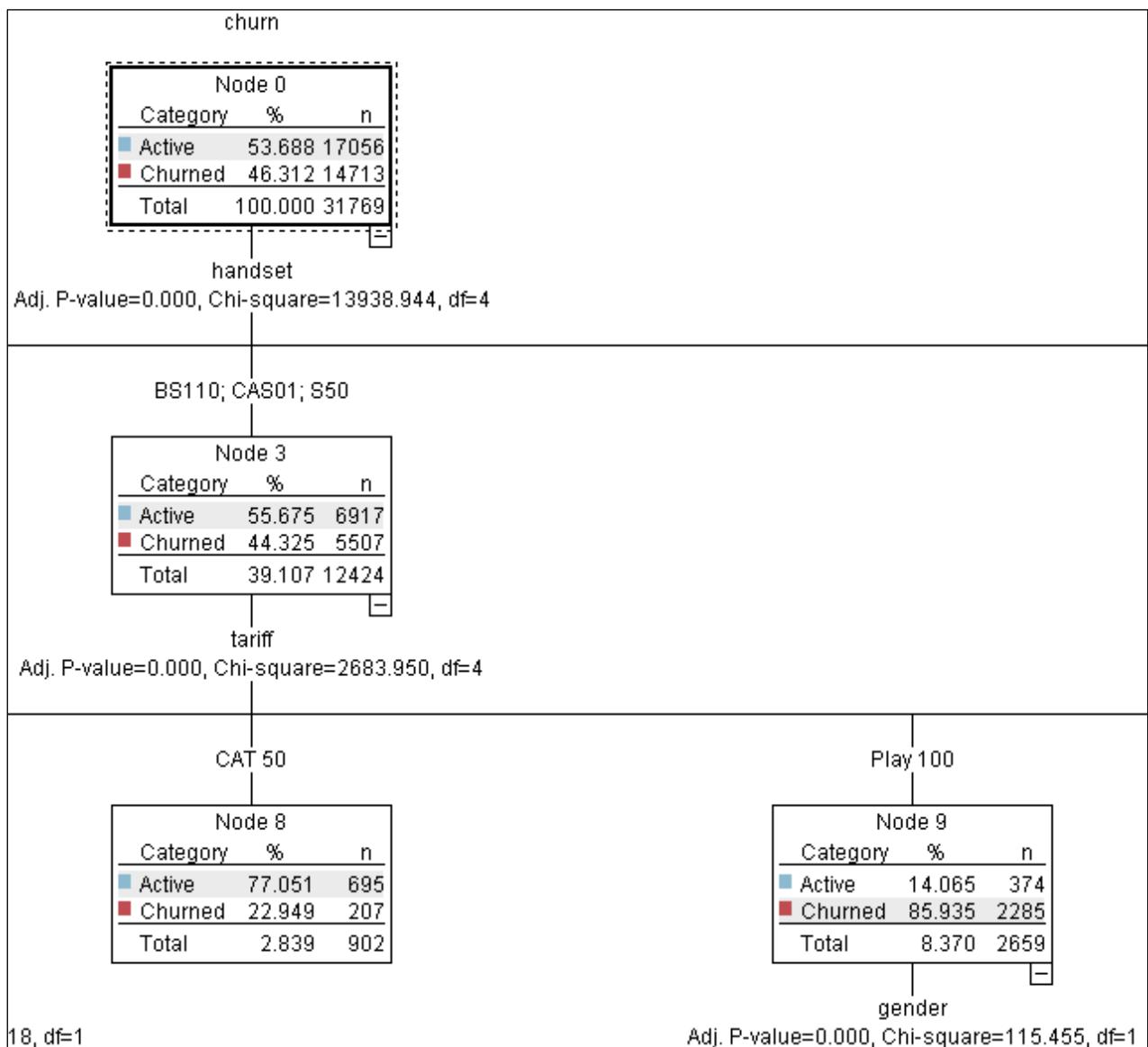


A model nugget is generated, added downstream from the Type node, and linked to the CHAID node. When you re-run the CHAID node (say, with other predictors), the model nugget will be updated automatically because of the link. Notice that the model nugget is also added to the Models Manager pane.

You will view the tree in the tree viewer.

17. Edit the **model nugget** named **churn**, and then click the **Viewer** tab.

A section of the results appears as follows:



14,713 of the 31,769 customers (46.3%) have churned (the root node). The first split was on handset, because that field had the most significant relationship (in terms of the Chi-square statistic) with churn. Some handsets were grouped (for example BS110, CAS01 and S50) by the CHAID algorithm, because they showed a similar percentage of churners. For the group of customers with handsets BS110, CAS01, or S50, there is a further split on tariff. Customers in the Play100 tariff showed the highest churn rates. This group, however, was split on gender. So within the group of customers in the Play100 tariff group having handset BS110, CAS01 or S50, there was a difference between men and women in churn rate.

Scroll through the tree to find groups with a high churn rate.

18. Scroll to the **left** so you have the result for handsets ASAD90, CAS30, SOP10 and SOP20.

The result appears as follows:

| ASAD170; CAS60; WC95 |        |      | ASAD90; CAS30; SOP10; SOP20 |        |      |
|----------------------|--------|------|-----------------------------|--------|------|
| Node 1               |        |      | Node 2                      |        |      |
| Category             | %      | n    | Category                    | %      | n    |
| Active               | 95.389 | 5730 | Active                      | 5.173  | 440  |
| Churned              | 4.611  | 277  | Churned                     | 94.827 | 8065 |
| Total                | 18.908 | 6007 | Total                       | 26.771 | 8505 |

Handsets ASAD90, CAS30, SOP10 and SOP20 were merged into one group by the CHAID algorithm, because they had similar churn rates. The churn percentage within this group was 94.8%.

The model nugget stores the rules, corresponding to the terminal nodes of the tree, and the model nugget will add two fields to the dataset. You will preview the data in the model nugget to examine these fields.

19. Click **Preview**, and scroll to the **last fields**.

A section of the results appears as follows:

| revenues | gadget_D_revenues | churn   | retention | \$R-churn | \$RC-churn |
|----------|-------------------|---------|-----------|-----------|------------|
| 28.000   | 0.000             | Churned | F         | Churned   | 0.948      |
| 0.000    | 0.000             | Churned | F         | Churned   | 0.948      |
| 0.000    | 35.000            | Churned | F         | Churned   | 0.948      |
| 0.000    | 41.000            | Churned | F         | Churned   | 0.948      |
| 0.000    | 0.000             | Churned | F         | Churned   | 0.948      |
| 0.000    | 0.000             | Churned | F         | Churned   | 0.948      |
| 32.000   | 0.000             | Churned | F         | Churned   | 0.948      |

The field \$R-churn stores the predicted category for each customer. The predicted category is the most frequent category in the node to which the customer belongs. For example, a customer with handset ASAD90 belongs to the node ASAD90, CAS30, SOP10, SOP20, and 94.8% of customers in this node has churned, so the predicted category for a customer with handset ASAD90 is Churned.

The field \$RC-churn gives the confidence for the prediction. For example, in the group customers with handsets ASAD90, CAS30, SOP10 or SOP20, 94.8% churned, so the confidence that the prediction ("Churned") is correct equals 0.948. Another interpretation is that of probability: the probability that a customer with one of the handsets ASAD90, CAS30, SOP10 or SOP20 churns equals 0.948.

Notice, that the first customers are predicted correctly (actual churn and predicted churn are the same).

20. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Tree** window.

Leave the stream open for the next task.

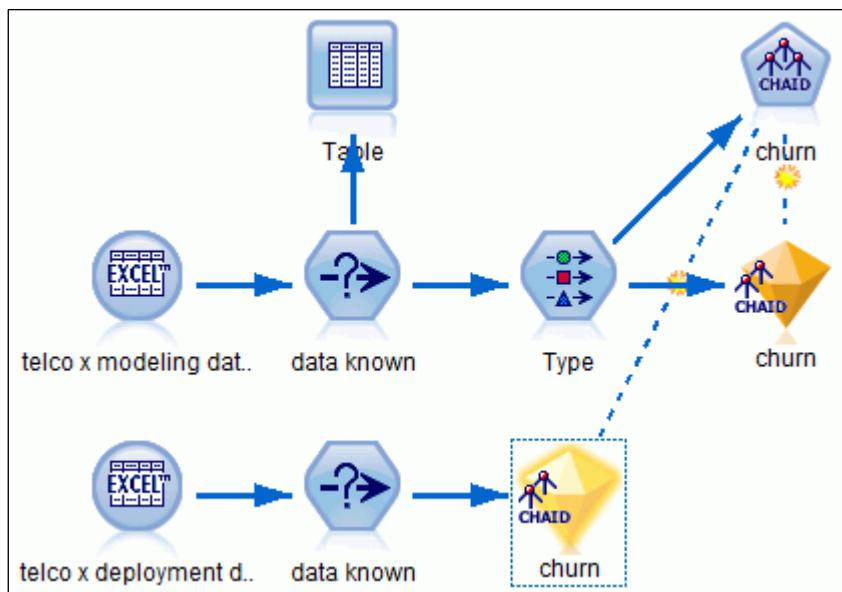
## Task 2. Deploy the model.

Having found a model and assuming that the model is satisfactory, the model can be applied to the current customers (the file **telco x deployment data.xls** stores the data).

In this task, you will build from the previous stream.

1. Add an **Excel** source node (Sources palette) to the stream canvas, edit the **Excel** node, select **telco x deployment data.xls**, and then click **Preview**.  
Data are unknown for some records, as indicated by the **data\_known** field. You will retain only the records with known data.
2. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Excel** dialog box.
3. Right-click the **Select** node named **data known**, select **Copy Node** from the context menu, **Paste** the node, and then add it downstream from the **Excel** node named **telco x deployment data.xls**.
4. Copy and paste the **model nugget**, and then add it downstream from the **Select** node named **data known** in the deployment stream.

The stream canvas appears as follows:



5. Add a **Table** node downstream from the **model nugget** in the deployment stream, and then run the **Table** node.

6. Scroll to the last fields in the **Table** output window.

A section of the results appears as follows:

| <b>t_B_revenues</b> | <b>gadget_C_revenues</b> | <b>gadget_D_revenues</b> | <b>\$R-churn</b> | <b>\$RC-churn</b> |
|---------------------|--------------------------|--------------------------|------------------|-------------------|
| 0.000               | 0.000                    | 0.000                    | Active           | 0.770             |
| 0.000               | 0.000                    | 0.000                    | Churned          | 0.948             |
| 20.000              | 28.000                   | 0.000                    | Active           | 0.954             |
| 24.000              | 0.000                    | 41.000                   | Churned          | 0.640             |
| 23.000              | 33.000                   | 40.000                   | Active           | 0.954             |
| 23.000              | 26.000                   | 43.000                   | Active           | 0.954             |
| 15.000              | 0.000                    | 37.000                   | Churned          | 0.640             |

Two fields are added: the predicted category and the confidence for that prediction. For example, the first record is predicted to stay active, with a probability of 0.77. The second record is predicted to churn, with a probability of at least 0.9.

Finally, you will export the data for the customers at risk to a text file, where at risk means that a customer is predicted to churn with a probability of at least 0.9.

7. In the **Table** output window, select the values **Churned** and **0.948** (use the Control key for a multiple selection), and then choose **Generate\Select Node ("And")** from the main menu.
  8. Click **OK** to close the **Table** output window.
  9. Add the generated **Select** node downstream from the **model nugget**, in the deployment stream.
  10. Edit the **Select** node, and then:
    - on the **Settings** tab, replace the expression '**\$RC-churn**' = **0.948160338544728** with '**\$RC-churn**' >= **0.9**
    - annotate the node with the text **customers@risk**
    - click **OK** to close the **Select** node
- Before exporting the data, rename the fields that are added by the model, and ensure that only these fields together with `customer_id` are exported.
- You will rename and remove fields with a Filter node.
11. Add a **Filter** node (Field Ops palette) downstream from the **Select** node.

12. Edit the **Filter** node, and then:

- remove all fields, except **customer\_id** and the two fields added by the model
- rename **\$R-churn** to **predicted value**
- rename **\$RC-churn** to **probability to churn**

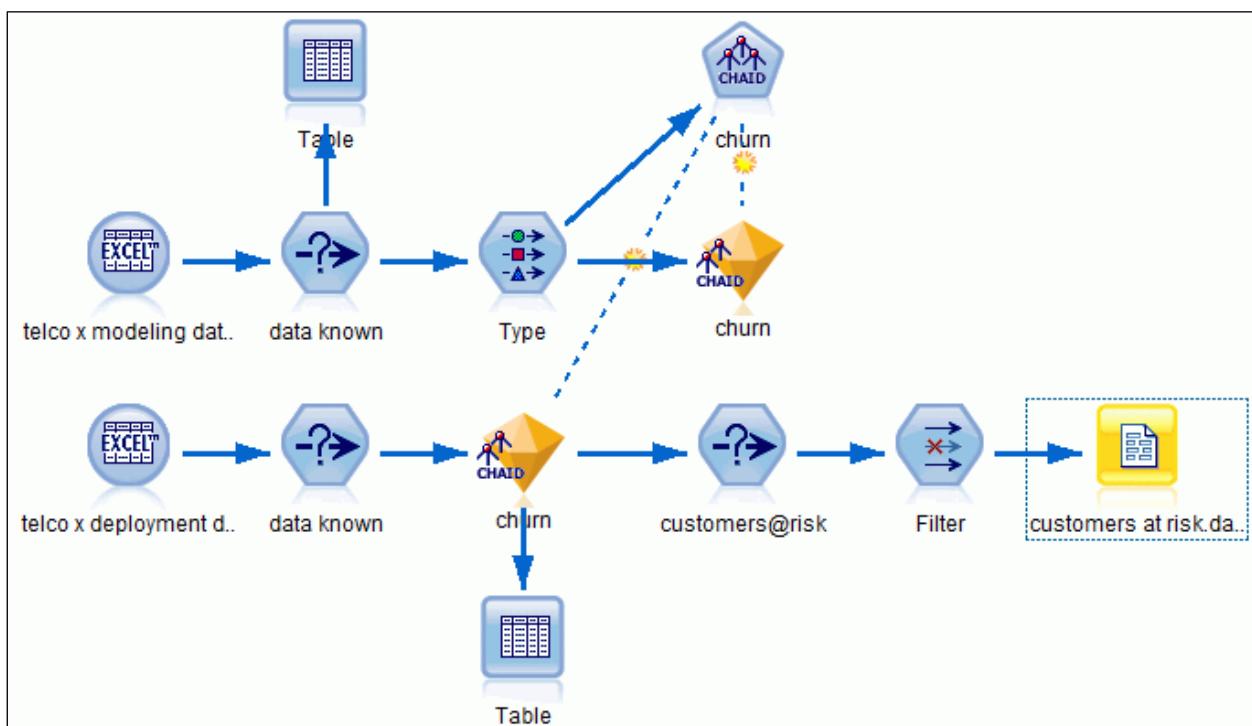
13. Click **OK** to close the **Filter** dialog box.

14. Add a **Flat File** node (Export palette) downstream from the **Filter** node.

15. Edit the **Flat File** node, and then for **Export file**, type **customers at risk.dat**.

16. Click **OK** to close the **Flat File** node.

The stream canvas appears as follows:



17. Run the **Flat File** export node, named **customers at risk.dat**.

This completes the demo for this module. You will find the solution results in the file **demo\_a\_data-mining\_tour\_completed.str**, located in the **03-A\_Data-Mining\_Tour\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Which of the following is the correct statement?

- A. "Deployment data" and "Modeling data" refer to the same type of data.
- B. "Historical data" and "Operational data" refer to the same type of data.
- C. "Historical data" and "Deployment data" refer to the same type of data.
- D. None of the above statements are correct.

Question 2: Which of the following statements are correct?

- A. It is of no use to include a field as a predictor, if that field is not available in the deployment data.
- B. It is of no use to include a field as a predictor, if the field values in the deployment dataset differ from the field values in the historical dataset.
- C. To build a predictive model, one field should have role Target and at least one field should have role Input.

Question 3: Is the following statement true or false? A model nugget adds fields to the dataset.

- A. True
- B. False

Question 4: Which of the following statements is the correct statement? Suppose you have built a model on dataset A, with gender and age used as predictors for churn. You have a model nugget in place that stores this model. Now you want to score records in dataset B, where you have the fields sex and age. When you use the model nugget to score new cases in dataset B MODELER issues the error message *Model refers to undefined field 'gender'*. How can you solve the issue?

- A. Insert a Filter node upstream from the model nugget in dataset B, and rename sex to gender.
- B. Insert a Filter node upstream from the model nugget in dataset B, and rename gender to sex.
- C. Insert a Filter node downstream from the model nugget in dataset B, and rename sex to gender.
- D. None of the above, because the issue cannot be solved.

**Answers to questions:**

Answer 1: D. None of the statements are correct. "Historical data" and "Modeling data" are synonyms, and "Deployment data" and "Operational data" are synonyms.

Answer 2: A, B, C.

Answer 3: A. True. A model nugget will add the model scores (for example the predicted values and the confidences, for a CHAID model nugget).

Answer 4: A. The field name in the model nugget is gender, while it is sex in the deployment dataset. To apply the model, rename sex into gender upstream from the model nugget in a Filter node, in dataset B.

Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - explain the basic framework of a data mining project
  - build a model
  - deploy a model

© 2014 IBM Corporation

In this module MODELER you have stepped through a (fictitious) data-mining project. You have built a model, using one of the modeling nodes (CHAID). Executing the modeling node produced a model nugget, a container for the model. You used the model nugget to score new cases (predicted category and confidence score were added for each record).

Building a model, and then applying that model to future cases is the basic framework of a data-mining project.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

3-34

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Workshop 1

## A Data-Mining Tour



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the **C:\Train\0A005** folder and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following file is used in this workshop:

- **ACME customer and rfm data.sav**: an IBM SPSS Statistics file storing customer data, and fields related to a test mailing that has was sent out.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

3-35

## Workshop 1: A Data-Mining Tour

This workshop is about ACME, a company selling sports products via the Web and via direct mail campaigns. ACME wants to promote a new product, the XL Original Orange Baseball Cap.

ACME has sent out a test mailing to 10,000 randomly selected customers, and recorded the response.

In this workshop, you will build a model using data of the test mailing. This model (hopefully) identifies groups with high response rates. You will then use this model to select the groups with high response rates in the rest of the customer database (only these groups will be included in the actual mailing for the XL Original Orange Baseball Cap).

The data are available in **ACME customer and rfm data.sav** and includes the following fields:

| Field                     | Field Description   |
|---------------------------|---|
| customer_id               | customer's identification number  |
| Gender                    | customer's gender   |
| email_address             | customer's e-mail address   |
| postal_code               | customer's postal code  |
| recency 01-01-2011        | customer's last order date, before JAN-01-2011  |
| frequency 01-01-2011      | customer's number of orders, before JAN-01-2011   |
| monetary_value 01-01-2011 | customer's total purchase amount, before JAN-01-2011  |
| has received test mailing | a field that flags whether the customer was in the test mailing (that was sent out FEB-01-2011) |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

| Field  | Field Description   |
|--|---|
| response   | for customers in the test mailing, this field flags whether the customer ordered the XL Original Orange Baseball Cap; for customers not in the test mailing, this field is undefined  |
| orderdate  | for customers in the test mailing who ordered the XL Original Orange Baseball Cap, this field gives the date that the XL Original Orange Baseball Cap was ordered; for customers in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined                        |
| number of days between test mailing and orderdate: | the number of days between the test mailing (FEB-01-2011) and the order date; this field is only valid for customers in the test mailing who ordered the XL Original Orange Baseball Cap; for customers in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined |
| ordered within month:                              | flags whether the order date was within one month after the test mailing went out; this field is only valid for those in the test mailing who ordered the XL Original Orange Baseball Cap; for those in the test mailing who did not order the XL Original Orange Baseball Cap, and for those not in the test mailing, this field is undefined    |

- Import the data from **ACME customer and rfm data.sav** (an IBM SPSS Statistics file; use the default settings for the import). Run a Table node to get a feel for the data.

How many records do you have in the dataset? And how many fields?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- A model will be built (in a next task) to predict the response on the test mailing. Select all those customers that were in the test mailing. (The field has\_received\_test\_mailing flags if the customer was in the test mailing.)  
How many customers were included in the test mailing?
- Use CHAID to predict response, with gender and the RFM (recency, frequency and monetary value) fields as predictors (recall that you set roles in a Type node, located in the Field Ops palette).
- Examine the model.

Which field is used as the first split field (the field under the root of the tree)?

Which group shows the highest response rate, and what is the probability to respond for this group?

- Which two fields will be added to the data, when records flow through the model nugget? What is their interpretation?
- Apply the model to the rest of the database (the customers that were not in the test mailing). Then, select the customers that are expected to respond positive (predicted value for response is T).

How many customers are selected?

- Export the data for the selected customers (customers expected to respond positive) to a text file (use the Flat File node), but ensure that only customer\_id and the two fields added by the model are exported. Also rename these last two fields to predicted\_category and confidence\_score, respectively (recall, selecting fields and renaming fields is done in a Filter node (ensure you choose the Filter node, and not the Filler node)).

## Workshop 1: Tasks and Results

### Task 1. Explore the data.

- Use the **Statistics File** node (Sources palette) to import data from **ACME customer and rfm data.sav**, add a **Table** node (Output palette) downstream from the **Statistics File** node, and then run the **Table** node.

The dataset has 12 fields and 30,000 records.

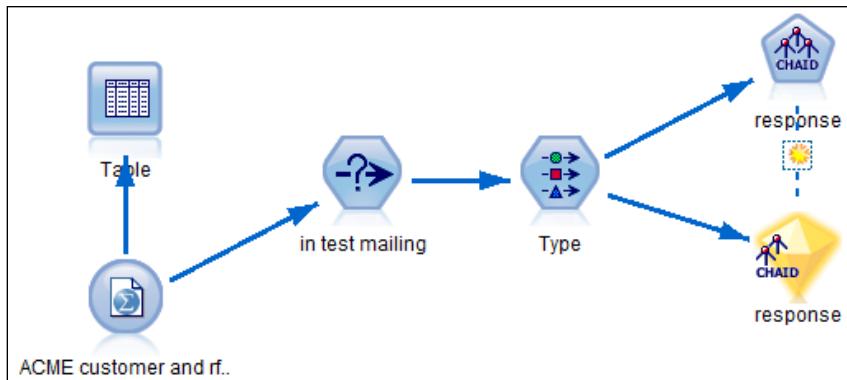
### Task 2. Select modeling data.

- Add a **Select** node that selects only the customers included in the test mailing downstream from the **Statistics File** source node (it is recommended to generate the Select node from a Table output window). Running a **Table** node downstream from the **Select** node shows that 10,000 records were selected.

### Task 3. Build a CHAID model.

- First, add a **Type** node (Field Ops palette) downstream from the **Select** node, edit the **Type** node and then click **Read Values**. Also, set roles:
  - **Inputs:** **gender, monetary\_value\_01\_01\_2011, frequency\_01\_01\_2011, recency\_01\_01\_2011**
  - **Target:** **response**
- Add a **CHAID** node (Modeling palette) downstream from the **Type** node (the CHAID node will be labeled with the name of the target, **response**), and then run the **CHAID** node.

The stream appears as follows:

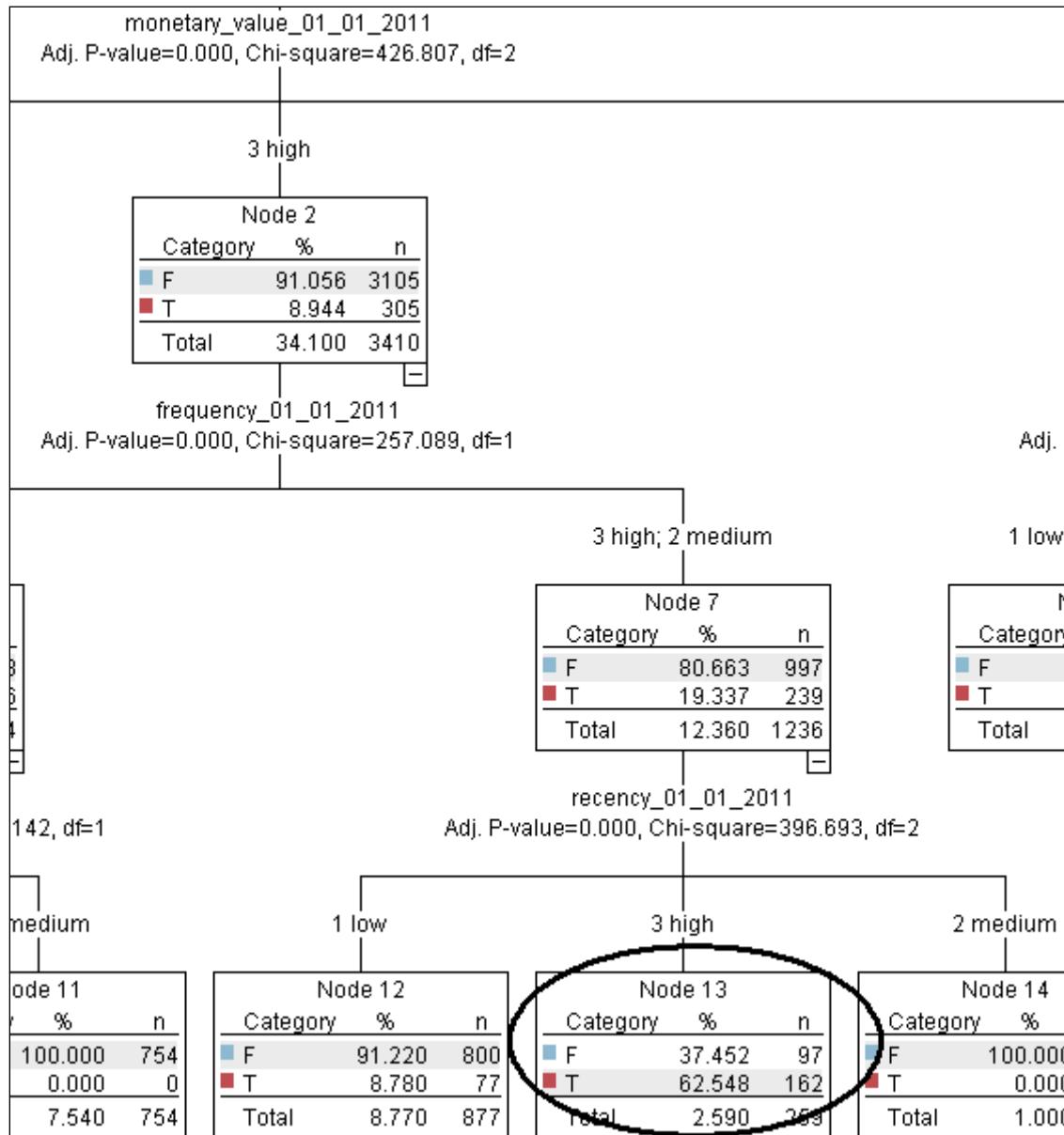


A model nugget was automatically added downstream from the Type node.

#### Task 4. Examine the model.

- To examine the model, edit the **model nugget**, and then click the **Viewer** tab. The first split field is monetary\_value\_01\_01\_2011.

- Scroll through the tree, and locate the group with highest response rate: monetary value high, frequency medium or high, and recency high, refer to the figure below. This group has a probability of 0.625 to buy the product.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Task 5. Interpretation of the fields added by the model nugget.

- Run a **Table** node downstream from the model nugget.

A section of the results appears as follows:

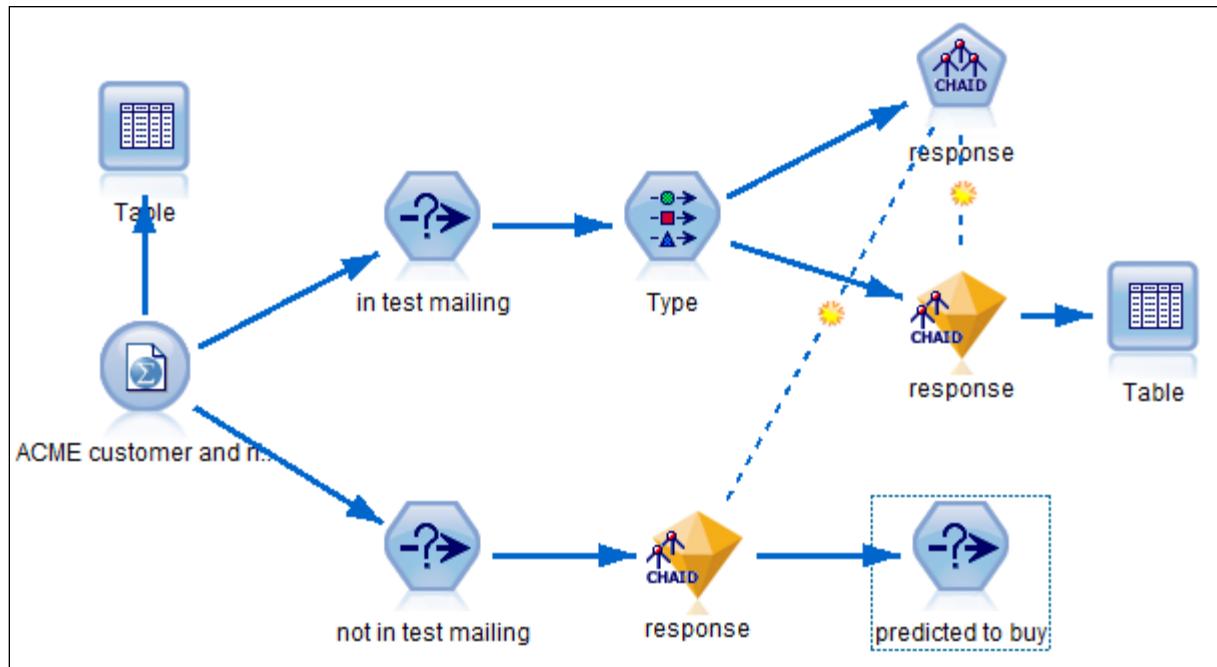
|    | 1   | has_received_test_mailing | response   | orderdate | number_of_days... | ordered_wit... | \$R-response | \$RC-response |
|----|-----|---------------------------|------------|-----------|-------------------|----------------|--------------|---------------|
| 1  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.953         |
| 2  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.993         |
| 3  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.953         |
| 4  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.953         |
| 5  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.993         |
| 6  | yes | T                         | 2011-02-04 |           | 3.000 yes         | F              |              | 0.953         |
| 7  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.993         |
| 8  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.953         |
| 9  | yes | F                         | \$null\$   |           | \$null\$ nap      | F              |              | 0.911         |
| 10 | yes | T                         | 2011-02-15 |           | 14.000 yes        | T              |              | 0.625         |

The field \$R-response gives the predicted response, the field \$RC-response gives the confidence for the prediction. For example, the first record is predicted to not buy the product. The confidence for this prediction equals 0.953. Record #10 is predicted to buy the product. The confidence for the prediction is 0.625 (this customer is in the group with the highest response rate).

## Task 6. Apply the model to the rest of the customers.

- First, select the customers not included in the test mailing (use a Select node).
- To apply the model, copy and paste the model nugget, and connect it downstream from the Select node.
- Next, select the customer predicted to buy the product (\$R-response is T for these customers).

The stream appears as follows:



- Run a **Table** node to get the record count; 254 customers are predicted to buy the product.

## Task 7. Export the results.

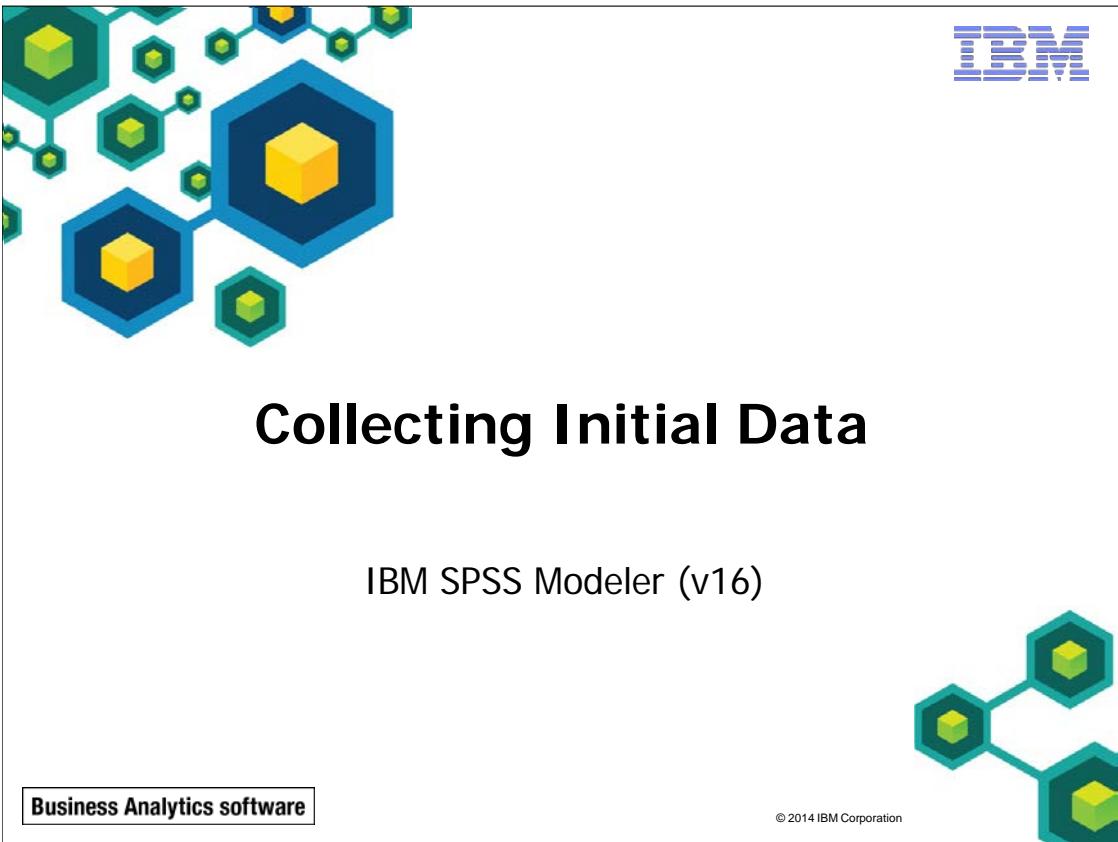
- Add a **Filter** node downstream from the **Select** node named **predicted to buy** (in the screen capture above), de-select all fields except **customer\_id**, **\$R-response**, and **\$RC-response**. Rename the last two fields into **predicted\_category** and **confidence\_score**.
- Add a **Flat File** node (Export palette) downstream from the **Filter** node, edit the node, and then specify the file name. Then run the **Flat File** node to export the data.

The stream **workshop\_a\_data-mining\_tour\_completed.str**, located in the **03-A\_Data-Mining\_Tour\Solution Files** sub folder, provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The advertisement features a white background with a decorative pattern of blue and green hexagons containing yellow cubes in the top left corner. In the top right corner is the IBM logo. Below the pattern, the title "Collecting Initial Data" is displayed in a large, bold, black sans-serif font. Underneath the title, the text "IBM SPSS Modeler (v16)" is shown in a smaller, black sans-serif font. At the bottom left, a small rectangular box contains the text "Business Analytics software". At the bottom right, there is a copyright notice: "© 2014 IBM Corporation".

# Collecting Initial Data

IBM SPSS Modeler (v16)

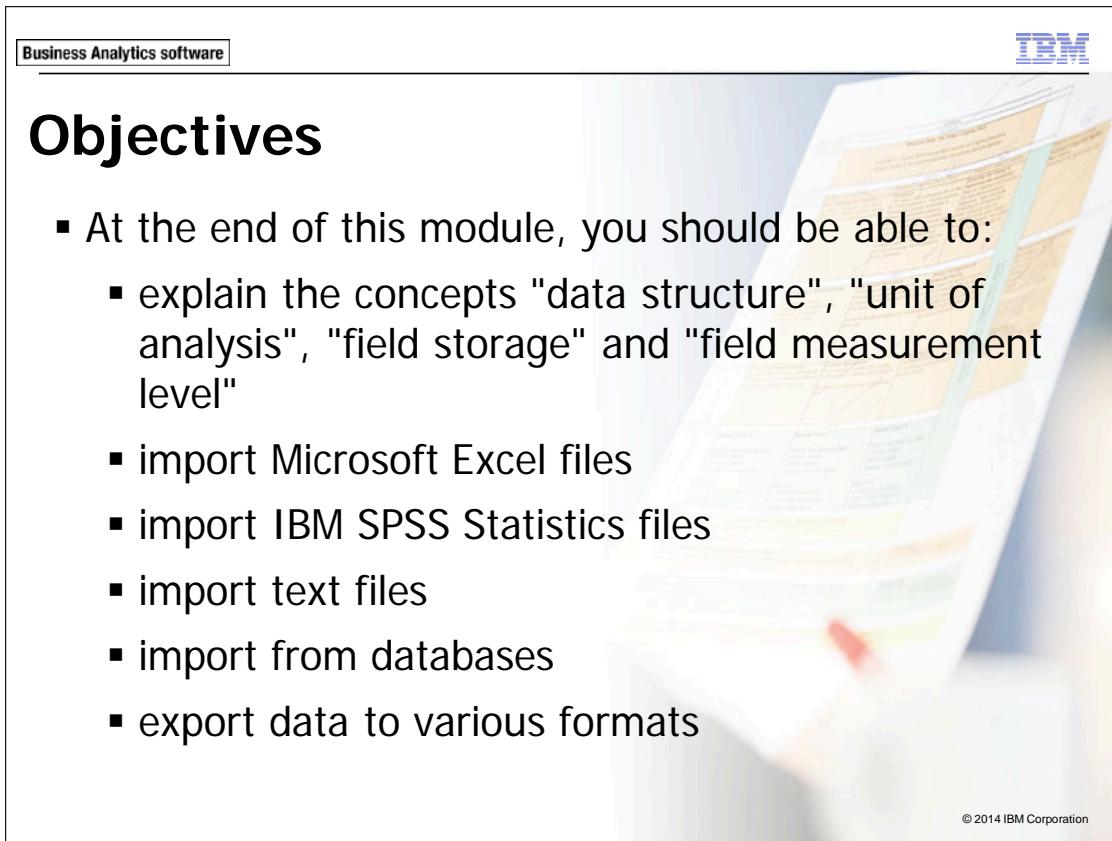
Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



**Business Analytics software**

**IBM**

# Objectives

- At the end of this module, you should be able to:
  - explain the concepts "data structure", "unit of analysis", "field storage" and "field measurement level"
  - import Microsoft Excel files
  - import IBM SPSS Statistics files
  - import text files
  - import from databases
  - export data to various formats

© 2014 IBM Corporation

The focus in this module is on two main tasks in the Data Understanding stage:

- Collect initial data: Getting your data into MODELER is the first step before any analysis can be done.
- Describe data: Describing your data in terms of number of records, the number of fields, the unit of analysis, and fields' measurement levels.

First you will be introduced to terminology such as unit of analysis and measurement level.

Before reviewing this module you should be familiar with the following topics:

- the CRISP-DM process model
- MODELER streams, nodes, and palettes

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Rectangular Data Structure

| Fields  | CUSTOMER_ID | GENDER | REGION | AGE_CATEGORY | INCOME |
|---------|-------------|--------|--------|--------------|--------|
| Records | C1          | Female | 1      | 1            | 1000   |
|         | C2          | Male   | 2      | 3            | 3200   |
|         | C3          | Female | 3      | 2            | 1400   |

© 2014 IBM Corporation



MODELER requires a rectangular data structure when analyses are performed.

A rectangular data structure is the product of records (rows of the data table) and fields (columns of the data table).

- All records contain the same fields.
- All records have a value for all fields, even if a value is undefined (MODELER will then show the \$null\$ value).

Data mining tools typically apply several different techniques and a simple, common data structure facilitates this.

MODELER has no limit on the number of records, nor on the number of fields.

# The Unit of Analysis

- What makes a record unique is called unit of analysis

| CUSTOMER_ID | GENDER | AGE_CATEGORY | GADGET_A | GADGET_B | GADGET_C | HAS_CHURNED |
|-------------|--------|--------------|----------|----------|----------|-------------|
| C1          | Female | 1            | T        | F        | F        | T           |
| C2          | Male   | 3            | F        | T        | T        | F           |
| C3          | Female | 2            | T        | T        | T        | F           |

| CUSTOMER_ID | GENDER | AGE_CATEGORY | GADGET | HAS_CHURNED |
|-------------|--------|--------------|--------|-------------|
| C1          | Female | 1            | A      | T           |
| C2          | Male   | 3            | B      | F           |
| C2          | Male   | 3            | C      | F           |
| C3          | Female | 2            | A      | F           |
| C3          | Female | 2            | B      | F           |
| C3          | Female | 2            | C      | F           |

© 2014 IBM Corporation



The definition of what makes a record unique is called the unit of analysis. A dataset has the same unit of analysis throughout the dataset.

This slide shows two datasets storing the same information. The upper dataset has one record per customer. The unit of analysis is customer in this dataset. In the lower dataset, a customer has as many records as he or she purchased gadgets. Here, it is the combination of customer and gadget that makes a record unique.

To answer a business question such as "What is the percentage of churners?" the number of customers that have a T value for HAS\_CHURNED needs to be divided by the number of customers. This computation is straightforward when each record represents a customer, but it is not when a customer has as many records as he or she has products. Only the upper dataset has the required unit of analysis to answer this business question.

Business Analytics software

IBM

# Field Storages

| Icon | Storage |
|------|---------|
|      | String  |
|      | Integer |
|      | Real    |
|      | Date    |

© 2014 IBM Corporation



Field storage describes how values are stored. MODELER distinguishes between string (or alphanumeric) fields and numeric fields. String fields store text values, as opposed to fields storing numbers. For example, a field GENDER with values MALE and FEMALE is string, while AGE is numeric. A numeric field can be integer or real. For example, when AGE has values such as 20, 34, 54, then AGE is an integer field; when the values are 20.12, 34.49, 54.16 it is a real field.

String and numeric (integer, real) are the most important storage types. Less frequent, but considered as a separate storage type, is the date storage type.

This slide shows the icons that are used for the storage types. Notice that real fields show a hash tag (#) on the icon, while integer fields do not.

Time fields and Timestamp fields are also storage types; they have distinct icons not shown in the table.

# Field Measurement Levels

| Icon  | Measurement Level |
|---|-------------------|
|  | Flag              |
|  | Nominal           |
|  | Ordinal           |
|  | Continuous        |
|  | Typeless          |

Categorical fields

© 2014 IBM Corporation



The storage of a field does not necessarily tell you which analyses are relevant for that field. Think, for example, of two numeric fields, age and postal code. Values for age are meaningful on their own and a statistic such as mean age makes sense. Values for postal code are just codes to represent geographical areas and you would not report its mean.

The concept of measurement level describes what the properties of a field's values are. This slide shows the measurement levels and their icons. Flags are fields with two categories. A nominal field has more than two categories, but there is no order in the categories. An ordinal field also has more than 2 categories, but the categories can be ranked. Flag, nominal and ordinal fields together are referred to as categorical fields because their values represent categories. Computing the mean is not appropriate for categorical fields (regardless their storage). Continuous fields express a quantity, and computing the mean is appropriate.

When a field does not conform to any of the above types, it is typeless.

**Business Analytics software**

**IBM**

## Storage and Measurement Level Illustrated

| CUSTOMER_ID | GENDER | REGION | AGE_CAT | INCOME | HANDSET | HAS_CHURNED |
|-------------|--------|--------|---------|--------|---------|-------------|
| C1          | Female | 1      | 1       | 1000   | ASAD90  | T           |
| C2          | Male   | 2      | 3       | 3200   | CAS30   | F           |
| C3          | Female | 3      | 2       | 1400   | SOP10   | F           |

A red arrow points from the 'REGION' column in the main table to a secondary table below, which details the storage type and measurement level for each field.

| Field       | Storage | Measurement level |
|-------------|---------|-------------------|
| CUSTOMER_ID | String  | Typeless          |
| GENDER      | String  | Flag              |
| REGION      | Integer | Nominal           |
| AGE_CAT     | Integer | Ordinal           |
| INCOME      | Integer | Continuous        |
| HANDSET     | String  | Nominal           |
| HAS_CHURNED | String  | Flag              |

© 2014 IBM Corporation 

In the dataset depicted on this slide, GENDER (with values Male and Female) and HAS\_CHURNED (with values T and F) are string fields with two categories, thus have measurement level flag.

REGION has values 1 thru 3, each representing a geographical area. These values are just arbitrary numbers and it is meaningless to say that region 2 has more region than region 1, or to compute the mean for this field.

HANDSET is a string nominal field.

The values of AGE\_CATEGORY (1 representing young, 2 middle-age, 3 senior) are integer, and indicate an order: you may say that a person in age category 2 is older than someone in age category 2. Computing the mean is not meaningful.

INCOME is an integer continuous field, and a statistic such as mean INCOME is meaningful.

CUSTOMER\_ID is a string typeless field.

## Storage and Measurement Level

- Initially, MODELER assigns a measurement level based on a field's storage:
  - string fields: categorical (  - numeric fields: continuous (
- When all data is read and all values are known, MODELER instantiates categorical to a specific measurement level

© 2014 IBM Corporation



When the first records have been read, MODELER assigns a measurement level to a field based on its storage.

String fields are initially assigned the categorical measurement level. After reading the first records MODELER does not know the specific measurement level because the number of categories is unknown at this point. For example, MODELER knows that a string field such as GENDER is categorical, because the storage of the field is string, but MODELER does not yet know that the field is flag because not all records have been processed. So, the categorical measurement level is temporarily assigned to the field and when all records have been read, MODELER instantiates the categorical measurement level to the flag measurement level (assuming GENDER has two categories).

Integer fields and real fields are assigned the continuous measurement level.

## Field Instantiation

- A field can be:
  - uninstantiated (  )
  - partially instantiated (  for string fields,  
 for numeric fields)
  - fully instantiated (flag, nominal, ordinal, continuous, or typeless icon)

© 2014 IBM Corporation



In MODELER, the process of reading or specifying information such as measurement level and values for a field is called instantiation.

A field that has unknown storage is uninstantiated.

A field is partially instantiated when its storage is known, but there is no information about its measurement level and its values.

When storage, measurement level and values are known, the field is fully instantiated.

# Field Instantiation Illustrated

| Field       | Storage | Measurement level | Values |
|-------------|---------|-------------------|--------|
| CUSTOMER_ID | String  | Categorical       | <Read> |
| GENDER      | String  | Categorical       | <Read> |
| REGION      | Integer | Continuous        | <Read> |
| AGE_CAT     | Integer | Continuous        | <Read> |
| INCOME      | Integer | Continuous        | <Read> |
| HANDSET     | String  | Categorical       | <Read> |
| HAS_CHURNED | String  | Categorical       | <Read> |

Partially  
instantiated  
fields

| Field       | Storage | Measurement level | Values       |
|-------------|---------|-------------------|--------------|
| CUSTOMER_ID | String  | Typeless          |              |
| GENDER      | String  | Flag              | Female/Male  |
| REGION      | Integer | Continuous        | [1, 3]       |
| AGE_CAT     | Integer | Continuous        | [1, 3]       |
| INCOME      | Integer | Continuous        | [1000, 3200] |
| HANDSET     | String  | Nominal           | ASAD90, CAS. |
| HAS_CHURNED | String  | Flag              | T/F          |

Fully instantiated  
fields



© 2014 IBM Corporation

This slide shows the difference between partially instantiated fields and fully instantiated fields. Partially instantiated fields have measurement level categorical or continuous, while fully instantiated fields show a specific measurement level.

The measurement level that MODELER initially assigns is not always correct. REGION has numerical storage and thus is instantiated to measurement level continuous, while it is a nominal field. AGE\_CATEGORY is instantiated to continuous, while the values only represent an order from low to high.

In another dataset it may happen that a field such as HAS\_CHURNED has values 0 and 1, instead of values such as T and F. MODELER will type the field incorrectly as continuous because of its storage.

In these instances you must change the measurement level, not only to prevent that meaningless statistics are reported such as mean REGION, but also because certain modeling techniques only apply to continuous fields, while other modeling techniques only apply to categorical fields. For example, when a field such as HAS\_CHURNED must be predicted, the available models to choose from depends on how its measurement level was defined.

Another moment where you have to change measurement levels is when you have changed a field's values upstream. For example, when you reclassify the values of a nominal field into two values, the field's measurement level has become flag. Upstream data changes, however, will not affect the stored measurement level or its values. To update the values or measurement level you need to re-instantiate the fields by reading values in a separate Type node (located in the Field Ops palette).

All in all, it cannot be emphasized enough that the fields' measurement levels should be correct. If not, the results of any analysis will most likely be worthless.

# Importing Data: The Sources Palette

- MODELER imports from:
  - databases
  - text Files
  - IBM products
  - other
- MODELER can generate data

© 2014 IBM Corporation



MODELER does not have a native data format and imports from:

- Databases: The Database node imports data from databases using the Open DataBase Connectivity (ODBC) protocol. Drivers must be installed and a connection to the database must be configured, which is usually the work of your database administrator.
- Text Files: Use the Var. file Node to import free formatted text files, use the Fixed File node to import fixed formatted text files.

- IBM Products, among which:
  - IBM SPSS Statistics: The Statistics File node imports from IBM SPSS Statistics. You can choose between reading variable names or variable labels, and values or value labels. It is advised to read the labels, especially for the values. MODELER will then treat the field as categorical instead of continuous. An IBM SPSS Statistics file also defines fields' storages and fields' measurement levels, and MODELER can also use this metadata.
  - IBM Cognos: Two nodes enable you to integrate Cognos with the predictive analytics capabilities of MODELER: the Cognos TM1 node and the IBM Cognos BI node.
- Other: Use the Excel node to read data from Microsoft Excel 1997- 2003 (\*.xls) or Microsoft Excel 2007, 2010. When you read data from a Microsoft Excel file, ensure that the file is closed in Microsoft Excel; if it is open in Microsoft Excel you will have the error message *Failed to open/create file*. Use the SAS File node to read SAS files. Also, MODELER supports reading XML files (the XML node).

You can also generate data. The User Input node enables you to type in some synthetic data, which is useful for test purposes. A more sophisticated option is to use the Sim Gen node, which can generate data using statistical distributions.

## Explore the Sources Dialog Boxes

- Three main tabs:
  - Data tab
  - Filter tab
  - Types tab

© 2014 IBM Corporation



All data source nodes have three main tabs in common: Data, Filter, and Types.

On the Data tab you will specify the file name of the file that has to be imported. Also, you will set the specific options for the file type at hand, such as when you import a Microsoft Excel file, if it is a Microsoft Excel 2007 file or a Microsoft 2010 file.

The Var. File and Fixed File node have different specifications on the Data tab. Refer to the *Importing Text Files* section of this module for the details.

The Filter tab enables the user to remove fields and/or to rename fields. The Filter node is also available as a separate node (in the Field Ops palette). Filtering fields in the data source itself, however, has the advantage that the fields will not be imported in the first place.

The Types tab shows, among others, the fields' storages, and their measurements levels. You can change the measurement level in the Measurement column.

To instantiate the data, click the Read Values button. The measurement level categorical will be instantiated to flag if the field has two values, and to nominal if the field has more than two values. If a categorical field has too many unique values, more than 250 by default, its measurement level will be set to typeless. The reason is that a nominal field with more than 250 categories would put too high of a load on the modeling techniques. The default (250) can be changed via Tools\Stream Properties\Options, Options tab, General item, in the Maximum members for nominal fields area).

The screenshot shows the Var. File node interface in IBM SPSS Modeler. The title bar says "Business Analytics software" and the IBM logo is in the top right. The main window has tabs: File, Data (which is selected), Filter, Types, and Annotations. Below the tabs is a table with columns: Field, Override, and Storage. The table lists fields: dropped\_calls, pay method, handset, churn, peak\_calls, peak\_mins, offpeak\_calls, offpeak\_mins, and weekend\_calls. The "dropped\_calls" row has an checked "Override" checkbox and a dropdown menu for "Storage". The dropdown menu is open, showing options: Real, (Unknown), String, Integer (which is selected and highlighted in yellow), Real, Time, Date, and Timestamp. At the bottom left of the interface is a copyright notice: "© 2014 IBM Corporation". At the bottom right is a decorative graphic of three hexagonal nodes connected by lines.

To import text data files, use the Var. File node (for free text format) or the Fixed File node (for fixed text format). Both nodes have an extra tab, File, next to the three main tabs common to all data sources (Data, Filter, and Types). Furthermore, whereas you specified the file name and the options for the import in the Data tab in the other data source nodes, this functionality is in the File tab in the Var. File and Fixed File node.

The Data tab gives the user control over the storage of the fields. Default storage can be overridden by enabling the Override option, and then selecting the appropriate storage in the Storage column.

The XML data source node also enables you to change a field's storage at the moment of data import. The other data source nodes do not provide this option and you need a field operation to change storages for these data sources.

## Exporting Data: The Export Palette

- Most source nodes are also represented on the Export palette
- Fields have to be instantiated when you export to a database, to Microsoft Excel, or to IBM SPSS Statistics
- When you export to IBM SPSS Statistics, MODELER field names have to comply with IBM SPSS Statistics variable name conventions

© 2014 IBM Corporation

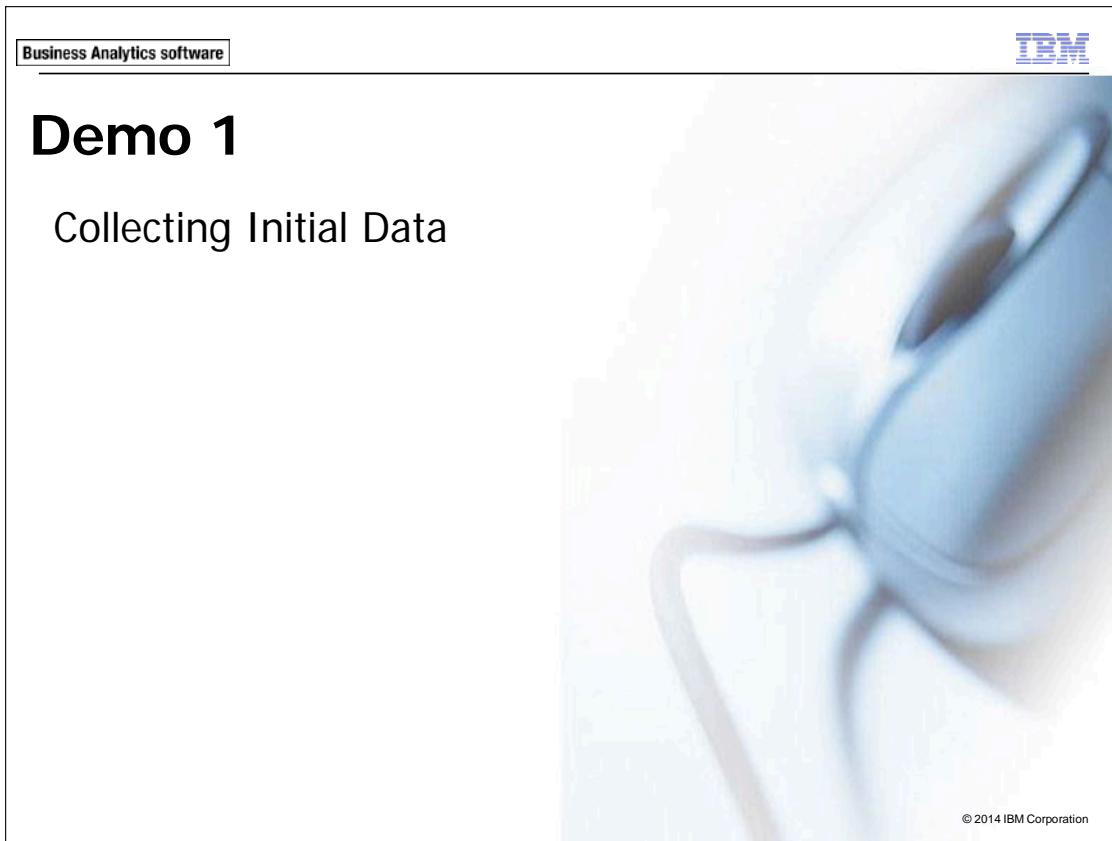


Almost all file types available for import have their counterpart in an export node. The Flat File exports to a delimited text file; exporting to fixed width is not supported.

When you export to a database, to Microsoft Excel, or to IBM SPSS Statistics, take into account that fields have to be instantiated. This means that a Type node will precede the export node, and that values are read in this Type node to instantiate the data.

When you export your data to an IBM SPSS Statistics file, field names have to comply with IBM SPSS Statistics field name conventions. For example, blanks are not allowed in an IBM SPSS Statistics field name. A Filter node upstream the Statistics Export node helps out, because this node has a feature to automatically convert MODELER field names to IBM SPSS Statistics field names.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



The slide is titled "Demo 1" and has a subtitle "Collecting Initial Data". It features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. A faint background image of a person wearing a hard hat and safety glasses is visible. The bottom right corner contains the text "© 2014 IBM Corporation".

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- **telco x customer data.xlsx** - a Microsoft Excel 2007 file, storing customer data
- **telco x products.dat** - a tab-delimited text file, storing data on gadgets that customers have purchased
- **telco x call data q1.sav** - an IBM SPSS Statistics data file storing three months of call detail records

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4-19

## Demo 1: Collecting Initial Data

### Purpose:

You are working in a telecommunications firm as a data miner. You have to import data from various data sources and you have to report on the unit of analysis and the fields' measurement levels.

### Task 1. Importing a Microsoft Excel file.

In this task you will import data stored in a Microsoft Excel 2007 file, **telco x customer data.xlsx**. The Microsoft Excel file has field names in the first row.

1. Place an **Excel** node (Sources palette) on the stream canvas.
2. Edit the **Excel** node, and then:
  - click the **Data** tab
  - for **File type**, select **Excel 2007 (\*.xlsx)**
  - browse to **telco x customer data.xlsx**
  - accept the defaults (ensure the **First row has column names** is enabled)
3. Click **Preview**.

A section of the results appear as follows:

|   | customer_id | gender | age    | postalcode | region | connect_date | end_date   | dropped_ca |
|---|-------------|--------|--------|------------|--------|--------------|------------|------------|
| 1 | K100010     | Male   | 46.... | 6253.000   | 3.000  | 2006-07-18   | 2010-05-10 | 1.0        |
| 2 | K100020     | Male   | 27.... | 4121.000   | 2.000  | 2004-09-18   | 2007-02-07 | 0.0        |
| 3 | K100030     | Male   | 39.... | 3870.000   | 2.000  | 2003-08-23   | 2006-08-13 | 2.0        |
| 4 | K100040     | Male   | 28.... | 8322.000   | 4.000  | 2005-08-18   | 2006-06-04 | 2.0        |
| 5 | K100050     | Male   | 47.... | 2614.000   | 2.000  | 2005-08-11   | 2006-08-13 | 0.0        |
| 6 | K100060     | Male   | 29.... | 1891.000   | 1.000  | 2004-07-06   | 2005-03-13 | 1.0        |
| 7 | K100070     | Male   | 38.... | 1741.000   | 1.000  | 2006-10-14   | 2007-01-01 | 1.0        |

The data import is okay.

4. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Excel** dialog box.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Leave the stream open for the next task.

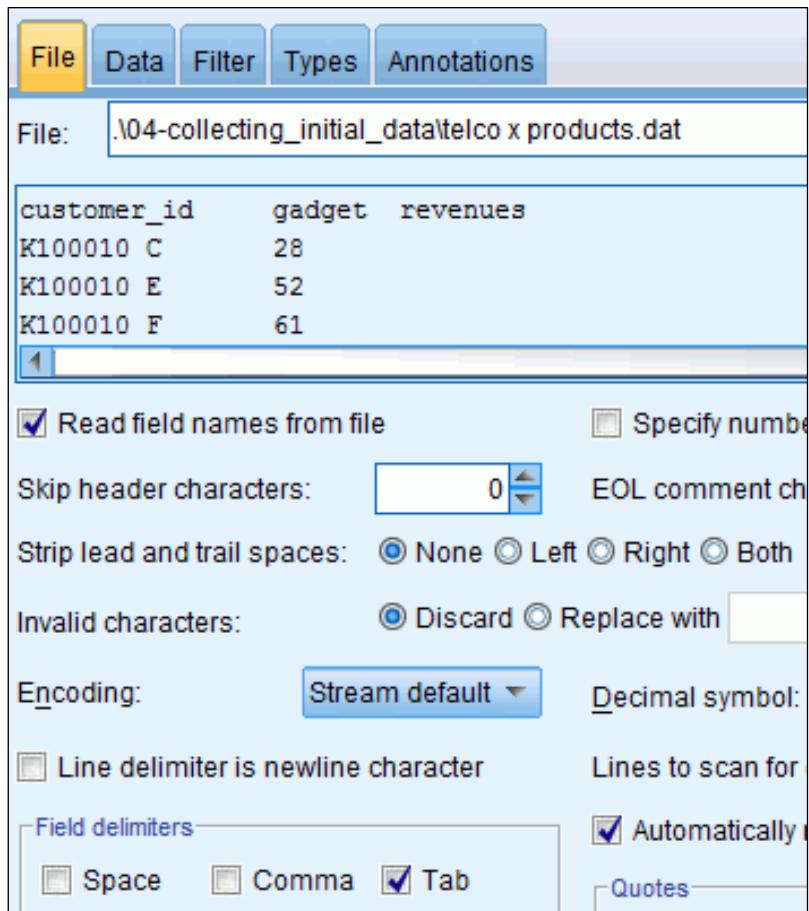
## Task 2. Importing a text file.

Another data file will be imported, **telco x products.dat**. This is a tab-delimited text file with field names in the first row. The file stores information about gadgets bought by the customer. The file has two fields string fields (customer\_id, gadget) and one integer field (revenues).

In this task, you will build from the previous stream.

1. Add a **Var. File** node (Sources palette) to the stream canvas.
2. Edit the **Var. File** node, and then:
  - click the **File** tab
  - for **File**, select **telco x products.dat**
  - ensure that the option **Read field names from file** is enabled
  - clear the **Comma delimiter** box
  - enable the **Tab delimiter** box

A section of the specifications in the Var. File dialog box appear as follows:



### 3. Click **Preview**.

A section of the results appear as follows:

|   | customer_id | gadget | revenues |
|---|-------------|--------|----------|
| 1 | K100010     | C      | 28       |
| 2 | K100010     | E      | 52       |
| 3 | K100010     | F      | 61       |
| 4 | K100010     | K      | 109      |
| 5 | K100020     | A      | 11       |
| 6 | K100020     | F      | 61       |
| 7 | K100020     | G      | 69       |

The data import is okay. Notice that a customer has as many records as he or she has gadgets.

### 4. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Var. File** dialog box.

Leave the stream open for the next task.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

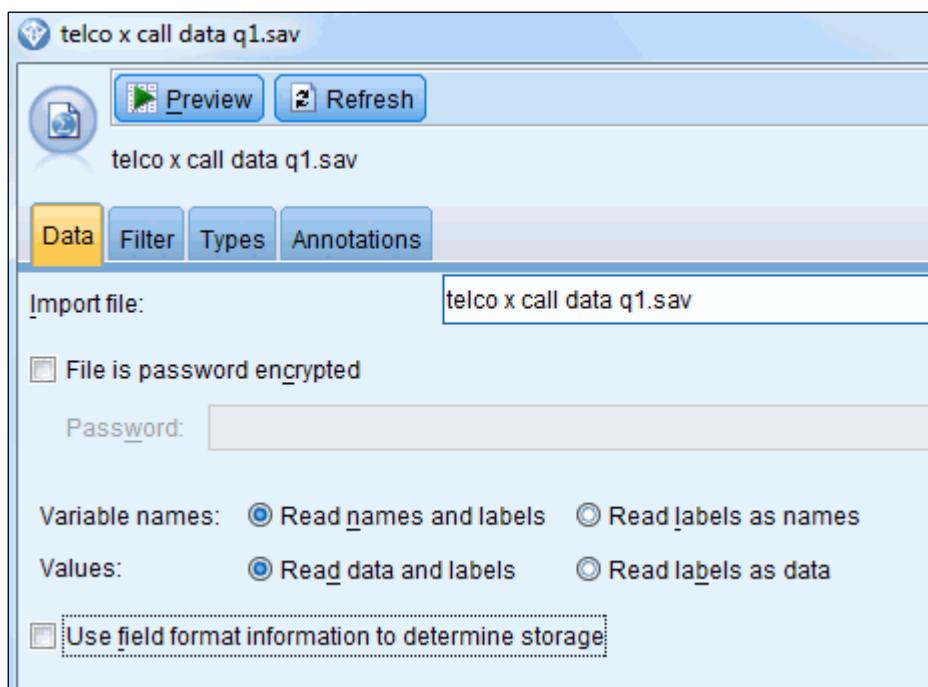
### Task 3. Importing an IBM SPSS Statistics file.

To demonstrate how to read data from IBM SPSS Statistics, consider another data source, **telco x call data q1.sav**. The file stores call detail records (number of peak calls, peak minutes, etc.) for three months. The variables and values in the file do not have labels, so the data can be imported using default values.

In this task, you will build from the previous stream.

1. Add a **Statistics File** node (Sources palette) to the stream canvas.
2. Edit the **Statistics File** node, and then:
  - click the **Data** tab
  - browse to **telco x call data q1.sav**
  - use default values to read variable names and variables values

A section of the specifications in the Statistics File dialog box appear as follows:



**3. Click Preview.**

A section of the results appear as follows:

|   | customer_id | peak_calls | peak_mins | offpeak_calls | offpeak_mins | weekend_calls | weekend_mins |
|---|-------------|------------|-----------|---------------|--------------|---------------|--------------|
| 1 | K100010     | 2.000      | 6.086     | 1.000         | 1.343        | 4.000         | 0.000        |
| 2 | K100010     | 2.000      | 6.060     | 1.000         | 1.337        | 4.000         | 0.000        |
| 3 | K100010     | 2.000      | 5.494     | 1.000         | 1.212        | 3.000         | 0.000        |
| 4 | K100020     | 7.000      | 5.538     | 5.000         | 2.970        | 0.000         | 0.000        |
| 5 | K100020     | 9.000      | 6.875     | 6.000         | 3.688        | 0.000         | 0.000        |
| 6 | K100020     | 8.000      | 6.172     | 5.000         | 3.310        | 0.000         | 0.000        |

Each customer has three records, because each record represents one month of data.

The fields peak\_calls, offpeak\_calls, etc have a real storage in MODELER, while they actually were integer in the IBM SPSS Statistics .sav file (not shown here). To have these fields as integers in MODELER, you will use the dictionary information that is contained in the IBM SPSS Statistics file.

- 4. Click OK to close the Preview output window.**
- 5. Enable the option Use field format information to determine storage.**
- 6. Click Preview.**

A section of the results appear as follows:

|   | customer_id | peak_calls | peak_mins | offpeak_calls | offpeak_mins | weekend_calls | weekend_mins |
|---|-------------|------------|-----------|---------------|--------------|---------------|--------------|
| 1 | K100010     | 2          | 6.086     | 1             | 1.343        | 4             | 0            |
| 2 | K100010     | 2          | 6.060     | 1             | 1.337        | 4             | 0            |
| 3 | K100010     | 2          | 5.494     | 1             | 1.212        | 3             | 0            |
| 4 | K100020     | 7          | 5.538     | 5             | 2.970        | 0             | 0            |
| 5 | K100020     | 9          | 6.875     | 6             | 3.688        | 0             | 0            |
| 6 | K100020     | 8          | 6.172     | 5             | 3.310        | 0             | 0            |

The fields storing calls are integers now.

- 7. Click OK to close the Preview output window, and then click OK to close the Statistics File dialog box.**

Leave the stream open for the next task.

## Task 4. Setting fields' measurement levels.

For the Microsoft Excel file imported in task #1, **telco x customer data.xlsx**, check the measurement levels, and, if needed, change them.

1. Edit the **Excel** source node that imports **telco x customer data.xlsx**, and then click the **Types** tab.

A section of the specification on the Types tab dialog box appear as follows:

| Field        | Measurement | Values |
|--------------|-------------|--------|
| customer_id  | Categorical |        |
| gender       | Categorical |        |
| age          | Continuous  |        |
| postalcode   | Continuous  |        |
| region       | Continuous  |        |
| connect_date | Continuous  |        |

Not all fields are typed correctly: postalcode and region are typed as continuous, because they store numeric values. However, the values are only of a categorical nature, so these fields should be typed as such.

2. In the **Measurement** column:

- click the cell for **postalcode**, and select **Categorical** from the dropdown
- click the cell for **region**, and select **Categorical** from the dropdown

A section of the specifications in the Types tab dialog box appear as follows-

| Field       | Measurement | Values |
|-------------|-------------|--------|
| customer_id | Categorical |        |
| gender      | Categorical |        |
| age         | Continuous  |        |
| postalcode  | Categorical |        |
| region      | Categorical |        |

You will instantiate the data.

3. Click **Read Values**.

A message box pops up, asking you if you want to read values for all fields.

4. Click **OK** to confirm.

A section of the results appear as follows:

| Field        | Measurement | Values                    |
|--------------|-------------|---------------------------|
| customer_id  | Typeless    |                           |
| gender       | Nominal     | FEMALE,Female,MALE,Mal... |
| age          | Continuous  | [-1.0,82.0]               |
| postalcode   | Typeless    |                           |
| region       | Nominal     | 1.0,2.0,3.0,4.0           |
| connect_date | Continuous  | [2003-01-01,2006-12-31]   |

The region field is typed as nominal, with 4 categories. The field postalcode, is instantiated to typeless, because it had more than 250 unique values. The field customer\_id is also instantiated to typeless, for the same reason.

The Values shows the categories for the flag and nominal fields. Notice that gender shows inconsistencies in spelling and that is why its measurement level is set to nominal instead of flag. For continuous fields, minimum and maximum values are displayed, so out-of-range values can be detected easily. In this example age shows negative numbers, which would deserve a closer inspection.

5. Click **OK** to close the **Type** dialog box.

This completes the demo for this module. You will find the solution results in **demo\_collecting\_initial\_data\_completed.str**, located in the **04-Collecting\_Initial\_Data\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Is the following statement true or false? The maximum number of records that MODELER can handle is 1,000,000.

- A. True
- B. False

Question 2: Is the following statement true or false? Fields represent the rows in a dataset.

- A. True
- B. False

Question 3: What is the storage of a field measuring preference for season of the year, with values of 1 (representing Spring), 2 (representing Summer), 3 (representing Fall) and 4 (representing Winter)?

- A. String
- B. Integer
- C. Real
- D. Date

Question 4: ? Refer to the figure below. Which of the following statements are correct?

- A. customer\_id is an integer field.
- B. gender is a string field.
- C. claim\_id is a string field.
- D. claim\_amount is an integer field.

| customer_id | gender | claim_id | claim_amount |
|-------------|--------|----------|--------------|
| 154557      | Female | 69348631 | 23242        |
| 154557      | Female | 69348632 | 12181        |
| 263204      | Male   | 40953049 | 39192        |
| 287476      | Male   | 45780237 | 1622         |
| 441097      | Male   | 89833962 | 481          |
| 441097      | Male   | 89833963 | 172          |
| 441097      | Male   | 89833964 | 435          |
| 524545      | Female | 35782742 | 35250        |
| 608461      | Male   | 7180084  | 2630         |
| 608461      | Male   | 7180085  | 936          |

Question 5: Is the following statement true or false? The dataset depicted below has the right unit of analysis to assess the relationship between gender and claim\_amount.

A. True

B. False

| customer_id | gender | claim_id | claim_amount |
|-------------|--------|----------|--------------|
| 154557      | Female | 69348631 | 23242        |
| 154557      | Female | 69348632 | 12181        |
| 263204      | Male   | 40953049 | 39192        |
| 287476      | Male   | 45780237 | 1622         |
| 441097      | Male   | 89833962 | 481          |
| 441097      | Male   | 89833963 | 172          |
| 441097      | Male   | 89833964 | 435          |
| 524545      | Female | 35782742 | 35250        |
| 608461      | Male   | 7180084  | 2630         |
| 608461      | Male   | 7180085  | 936          |

Question 6: Consider the dataset depicted below, with customers and their claims (a customer can have more than 1 claim, claim\_id is the unique identifier). The unit of analysis in this dataset is defined by:

- A. customer\_id
- B. gender
- C. claim\_id
- D. claim\_amount

| customer_id | gender | claim_id | claim_amount |
|-------------|--------|----------|--------------|
| 154557      | Female | 69348631 | 23242        |
| 154557      | Female | 69348632 | 12181        |
| 263204      | Male   | 40953049 | 39192        |
| 287476      | Male   | 45780237 | 1622         |
| 441097      | Male   | 89833962 | 481          |
| 441097      | Male   | 89833963 | 172          |
| 441097      | Male   | 89833964 | 435          |
| 524545      | Female | 35782742 | 35250        |
| 608461      | Male   | 7180084  | 2630         |
| 608461      | Male   | 7180085  | 936          |

Question 7: Suppose a field region with values 1 through 8 (north, north east, east, south-east, south, south-west, west, north-west) is used as predictor. Also suppose that a model finds the following rules:

- If region < 4.5 then the percentage churners is 12%.
- If region  $\geq 4.5$  and region < 7.5 then the percentage churners is 8%.
- If region  $> 7.5$  then the percentage churners is 1%.

Is the following statement true or false? Region is typed correctly as a continuous field.

- A. True
- B. False

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Question 8: Refer to the figure that follows. Which of the following statements are correct?

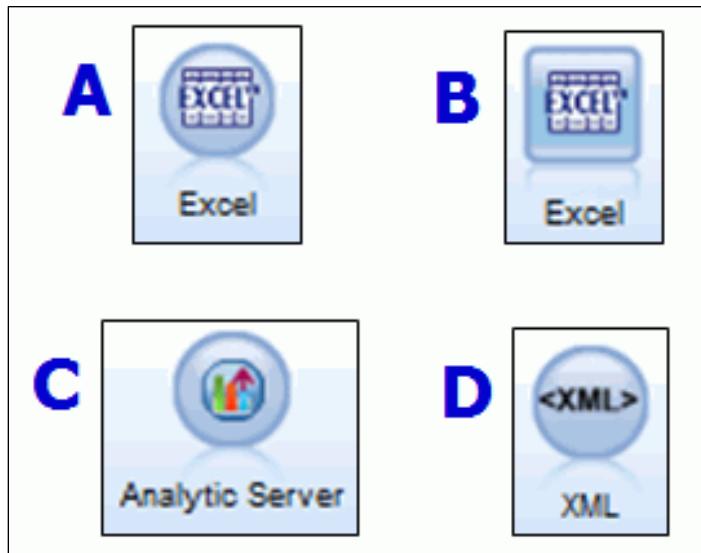
- A. MARITAL\_STATUS is a string nominal field
- B. FRAUD is a string flag field
- C. LENGTH\_OF\_STAY (YEARS) is a real continuous field
- D. CUSTOMER\_ID is an integer continuous field
- E. NUMBER\_OF\_CREDIT\_CARDS is an integer nominal field

| CUSTOMER_ID | MARITAL_STATUS | NUMBER_OF_CREDIT_CARDS | LENGTH_OF_STAY (YEARS) | FRAUD |
|-------------|----------------|------------------------|------------------------|-------|
| 1           | Married        | 1                      | 1.2                    | Yes   |
| 2           | Single         | 6                      | 4.8                    | No    |
| 3           | Divorced       | 2                      | 2.1                    | No    |
| 4           | Single         | 1                      | 0.5                    | Yes   |
| 5           | Married        | 3                      | 3.4                    | no    |

Question 9: Is the following statement true or false? Measurement level of fields is unrelated to the results obtained by modeling.

- A. True
- B. False

Question 10: Fill in the blank. Node \_\_\_\_\_ (A/B/C/D) imports a Microsoft Excel file.



Question 11: Which of the following is the correct statement?

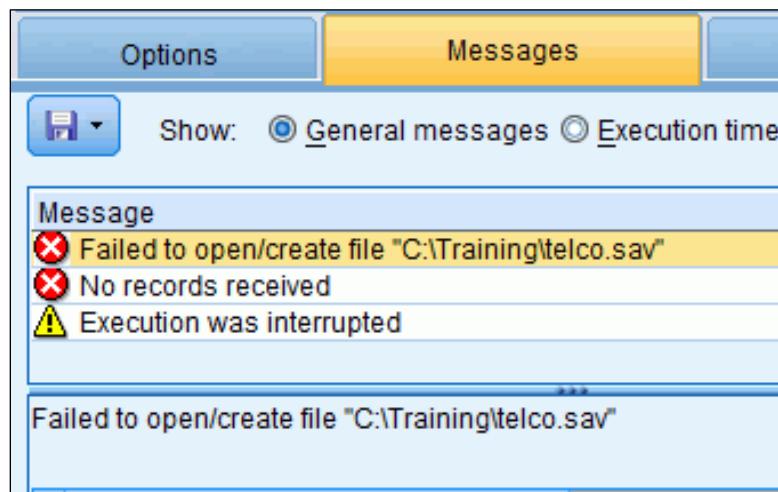
- A. Before you start your analysis with MODELER, you first need to import and save your data as a MODELER data file (extension .mdf).
- B. Data stored in a DB2 database cannot be imported into MODELER, because there is no DB2 Source node.
- C. Data can be imported from an IBM SPSS Statistics (\*.sav) file.
- D. To import data from IBM Cognos TM1, that data must be converted into a text file, and then the text file can be imported in MODELER.
- E. MODELER supports Microsoft Excel .xls files, but not Microsoft Excelxlsx files.
- F. When you read data into MODELER, every data source in the stream must be of the same type (for example, each data source in the stream must be an IBM SPSS Statistics file).

Question 12: Which of the following statements are correct?

- A. When exporting to Microsoft Excel, data have to be instantiated.
- B. When exporting to IBM SPSS Statistics, data have to be instantiated.
- C. When exporting to a text file, data have to be instantiated.
- D. When exporting to IBM SPSS Statistics, MODELER field names have to conform to IBM SPSS Statistics field names.

Question 13: Which of the following is the correct statement? Refer to the figure that follows. What can be a reason for this error message?

- A. The file named telco.sav cannot be found in the folder C:\Training.
- B. The Statistics Export node is used to import an IBM SPSS Statistics file.
- C. The IBM SPSS Statistics file does not have any data in it.
- D. MODELER cannot read IBM SPSS Statistics files.



Question 14: Which of the following is the correct statement? The working folder for MODELER can be set:

- A. by using the Working Folder node, located in the Sources palette
- B. by selecting File\Set Directory from MODELER's main menu
- C. by setting it in the Control Panel of your operation system
- D. all the previous statements are incorrect, because the working folder cannot be set

Question 15: Refer to the figure that follows, that shows a dataset in MODELER. Was the dataset imported correctly?

- A. Yes
- B. No

| C1          | C2     | C3   | C4         | C5     | C6           | C7       | C8            |
|-------------|--------|------|------------|--------|--------------|----------|---------------|
| customer_id | gender | age  | postalcode | region | connect_date | end_date | dropped_calls |
| K100010     | Male   | 46.0 | 6253.0     | 3.0    | 38916.0      | 40308.0  | 1.0           |
| K100020     | Male   | 27.0 | 4121.0     | 2.0    | 38248.0      | 39120.0  | 0.0           |
| K100030     | Male   | 39.0 | 3870.0     | 2.0    | 37856.0      | 38942.0  | 2.0           |
| K100040     | Male   | 28.0 | 8322.0     | 4.0    | 38582.0      | 38872.0  | 2.0           |

## Answers to questions:

Answer 1: B. False. There is no limit to the number of records.

Answer 2: B. False. Rows represent records, not fields.

Answer 3: B. Storage is integer.

Answer 4: A, B, D. C is not correct, because the field is integer, not string.

Answer 5: B. False. To assess the relationship, a dataset is needed with 1 record per customer (in the current dataset, a person's gender counts as many times as he or she has claims).

Answer 6: C. The claim\_id field uniquely identifies a record.

Answer 7: B. False. Region is a nominal field. For example, it could be that customers from region 3 and 7 are much alike in their churn behavior, but because region is typed as continuous, this will not be detected by the modeling technique.

Answer 8: A, B, C. Customer\_id is integer; its measurement level is not continuous but typeless.

Answer 9: B. False. Fields' measurement levels should be defined properly, otherwise modeling results are useless.

Answer 10: A.

Answer 11: C.

Answer 12: A, B, D.

Answer 13: A.

Answer 14: B.

Answer 15: B. False. The field names are imported as data.

## Summary

- At the end of this module, you should be able to:
  - explain concepts "data structure", "unit of analysis", "field storage" and "field measurement level"
  - import Microsoft Excel files
  - import IBM SPSS Statistics files
  - import text files
  - import from databases
  - export data to various formats

© 2014 IBM Corporation



This module introduced you to important concepts in MODELER. Also, you have been introduced to the Sources palette, to import data from a variety of data formats.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

4-36

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Workshop 1

## Collecting Initial Data



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following (synthetic) files used in this workshop are:

- **ACME customers.xls** - information on customers (a Microsoft Excel 2003 file)
- **ACME purchases 1999 - 2004.dat** - purchases made by customers over the period 1999 - 2004 (tab-delimited text)
- **ACME purchases 2005 - 2010.dat** - purchases made by customers over the period 2005 - 2010 (tab-delimited text)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- **ACME orderlines 1999 - 2004.sav** - orders per purchase, over the period 1999 – 2004 (IBM SPSS Statistics)
- **ACME orderlines 2005 - 2010.sav** - orders per purchase, over the period 2005 - 2010 (IBM SPSS Statistics)
- **ACME mailing history.xlsx** - data on test mailings (a Microsoft Excel 2007 file)
- **ACME zip data.csv** - data on zip (postal) codes (comma-separated text)

## Workshop 1: Collecting Initial Data

You are working for ACME, a (fictitious) company selling sport products. ACME has sent out a test mailing for one of their products, and later you will build a model to identify groups with high response rates. At this point you need to import ACME's data files.

- Import data from the following sources (ensure, that each data file below is imported with a corresponding source node, so you will have as many source nodes as data files):
  - **ACME customers.xls**
  - **ACME purchases 1999 - 2004.dat**
  - **ACME purchases 2005 - 2010.dat**
  - **ACME orderlines 1999 - 2004.sav**
  - **ACME orderlines 2005 - 2010.sav**
  - **ACME mailing history.xlsx**
  - **ACME zip data.csv**

For each dataset, determine the number of records, the number of fields, and the unit of analysis (which field, or combination of fields define unique records).

| File                            | # of records | #of fields | unit of analysis defined by |
|---------------------------------|--------------|------------|-----------------------------|
| ACME customers.xls              |              |            |                             |
| ACME purchases 1999 - 2004.dat  |              |            |                             |
| ACME purchases 2005 - 2010.dat  |              |            |                             |
| ACME orderlines 1999 - 2004.sav |              |            |                             |
| ACME orderlines 2005 - 2010.sav |              |            |                             |
| ACME mailing history.xlsx       |              |            |                             |
| ACME zip data.csv               |              |            |                             |

- Examine how the customer dataset relates to the purchases (1999 – 2004 period) dataset, and how the purchases dataset relates to the order lines (1999 – 2004 period) dataset.

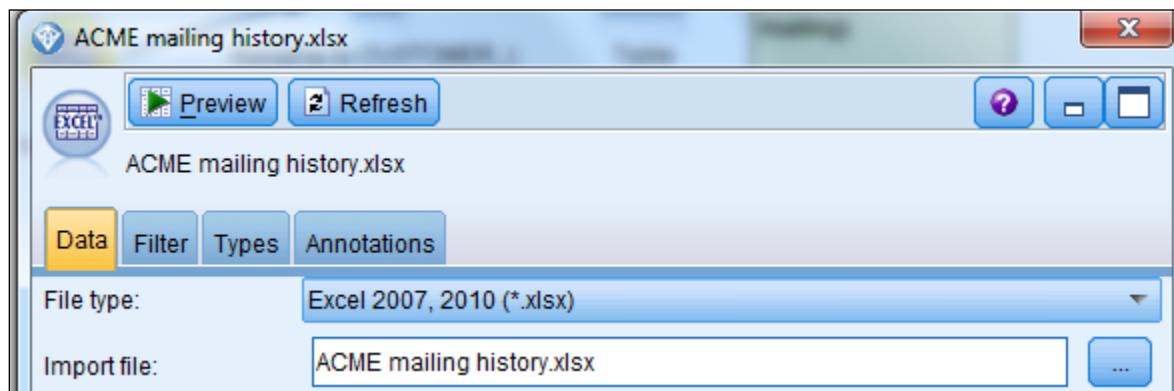
Hint: In the customer dataset, select record with CUSTOMER\_ID = 731. Select the same customer in the purchases (1999 – 2004) dataset. You will notice that this customer has multiple records, one of them being the purchase with PURCHASE\_ID = 6336. Then, in the order lines (1999 - 2004) dataset, select the records where PURCHASE\_ID = 6336.

- The field name that identifies the customer should be CUSTOMER\_ID in all datasets. For your data sources, check if the field name is in upper case, and if not change the field name into upper case.
- For the customer data (data source ACME customers.xls), go through the fields and set the appropriate measurement level.

## Workshop 1: Tasks and Results

### Task 1. Import data.

- **ACME customers.xls**: use an **Excel** node (use default settings for the import).
- **ACME purchases 1999 - 2004.dat, ACME purchases 2005 - 2010.dat**: use a **Var. File** node (ensure that the delimiter is Tab, not Comma).
- **ACME orderlines 1999 - 2004.sav, ACME orderlines 2005 - 2010.sav**: use the **Statistics File** node (use default settings for the import).
- **ACME mailing history.xlsx**: **Excel** node (ensure that the file type is Excel 2007 – 2010 (\*.xlsx)).



- **ACME zip data.csv**: use a **Var. File** node (use default settings for the import).

## Task 2. Determine the unit of analysis.

Use a Table node downstream from each data source to obtain the number of records and the number of fields. The results are shown in the following table:

| File                            | # of records | #of fields | unit of analysis defined by             |
|---------------------------------|--------------|------------|---|
| ACME customers.xls              | 30,000       | 6          | CUSTOMER_ID                             |
| ACME purchases 1999 - 2004.dat  | 4,018        | 3          | PURCHASE_ID                             |
| ACME purchases 2005 - 2010.dat  | 67,109       | 3          | PURCHASE_ID                             |
| ACME orderlines 1999 - 2004.sav | 7,426        | 6          | ITEM_ID                                 |
| ACME orderlines 2005 - 2010.sav | 162,308      | 6          | ITEM_ID                                 |
| ACME mailing history.xlsx       | 21,000       | 2          | customer_id<br>combined with<br>mailing |
| ACME zip data.csv               | 4,983        | 4          | ZIP                                     |

## Task 3. Relationships between datasets.

- One customer can have multiple purchases, and a purchase may consist of multiple items.

For example, CUSTOMER\_ID 731 has 1 record in the customer dataset ACME customers.xls and has 3 records (purchases) in ACME purchases 1999 - 2004.dat. One of these purchases was PURCHASE\_ID 6336. This PURCHASE\_ID has three records in the ACME orderlines 1999 - 2004.sav file.

## Task 4. Ensure that field names match in the various datasets.

- In the mailing history data, the field names are in lower case. Use a **Filter** node (or the Filter tab in the Excel data source node) to rename **customer\_id** into **CUSTOMER\_ID**.

## Task 5. Set measurement levels.

- In the customer dataset, ZODIAC is typed as continuous (because it has real storage). Set its measurement level to nominal in a **Type** node (or in the Types tab of the Excel data source node).

A section of the results appear as follows:

| Field         | Measurement | Values         |
|---------------|-------------|----------------|
| CUSTOMER_ID   | Typeless    |                |
| GENDER        | Nominal     | F,M,f,m        |
| CREDITLIMIT   | Continuous  | [-999999.0...  |
| ZODIAC        | Nominal     | 1.0,2.0,3.0... |
| E-MAIL ADD... | Typeless    |                |
| ZIP           | Typeless    |                |

The stream **workshop\_collecting\_initial\_data\_completed.str**, located in the **04-Collecting\_Initial\_Data\Solution Files** sub folder, provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

The advertisement features a white background with a decorative pattern of teal and yellow hexagons and small cubes in the top left corner. In the top right corner is the blue IBM logo. Below the logo is the title "Understanding Your Data" in a large, bold, black sans-serif font. Underneath the title is the text "IBM SPSS Modeler (v16)" in a smaller, black sans-serif font. At the bottom left is a rectangular box containing the text "Business Analytics software". At the bottom right is a small graphic of three teal hexagons connected by lines, each containing a yellow cube, with the text "© 2014 IBM Corporation" underneath.

**Business Analytics software**

IBM SPSS Modeler (v16)

© 2014 IBM Corporation

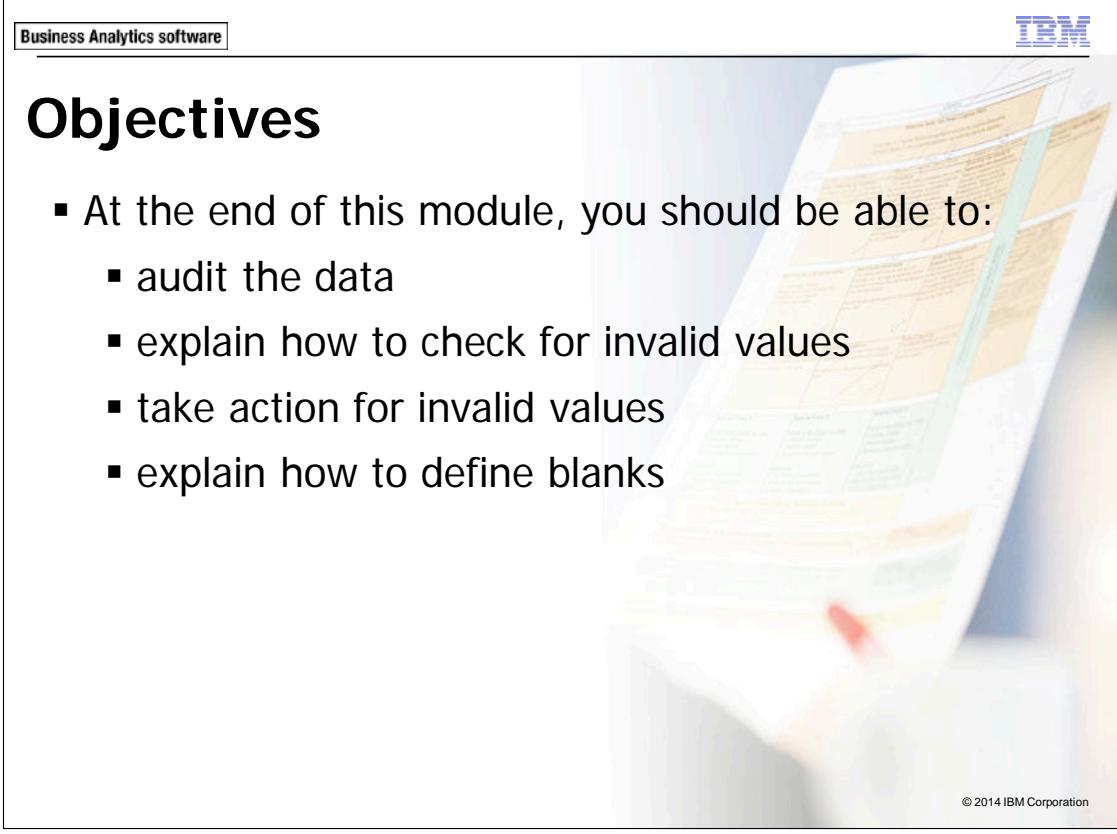
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Objectives

- At the end of this module, you should be able to:
  - audit the data
  - explain how to check for invalid values
  - take action for invalid values
  - explain how to define blanks



© 2014 IBM Corporation

Once data is read into MODELER, the next step is to explore the data and to become thoroughly familiar with its characteristics. The data most likely contains errors and missing information. Therefore, before models can be built, the quality of the data must be assessed. The higher the quality of the data used, the more accurate the predictions and the more useful the results.

The focus in this module is on two tasks in the Data Understanding stage in the CRISP-DM model:

- explore the data by running a data audit
- assess the quality of the data by reporting out-of range values and dealing with missing data

Before reviewing this module participants should be familiar with:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

The screenshot shows a section of the IBM SPSS Modeler interface titled "Data Audit Illustrated". At the top, there are tabs for "Audit" (which is selected), "Quality", and "Annotations". Below the tabs is a table with columns: "Field", "Sample Graph", "Measurement", "Min", "Max", and "Mean".

| Field   | Sample Graph | Measurement | Min   | Max   | Mean      |
|---------|--------------|-------------|-------|-------|-----------|
| AGE     |              | Continuous  | 18    | 50    | 31.820    |
| INCOME  |              | Continuous  | 15005 | 59944 | 25580.212 |
| GENDER  |              | Flag        | --    | --    | --        |
| MARITAL |              | Nominal     | --    | --    | --        |

At the bottom left of the interface, it says "© 2014 IBM Corporation". To the right, there is a decorative graphic of three hexagonal shapes in blue, green, and yellow.

When data is read into MODELER, you will check if the data import was successful. A useful option is to preview the data or to run a Table node, as this will show if field values make sense. For example, a field such as age must show numeric values and not text values.

When your dataset has thousands of records, it is not an option to use Table output to look for suspicious values; instead it is better to run a data audit. This slide shows a section of a data audit report.

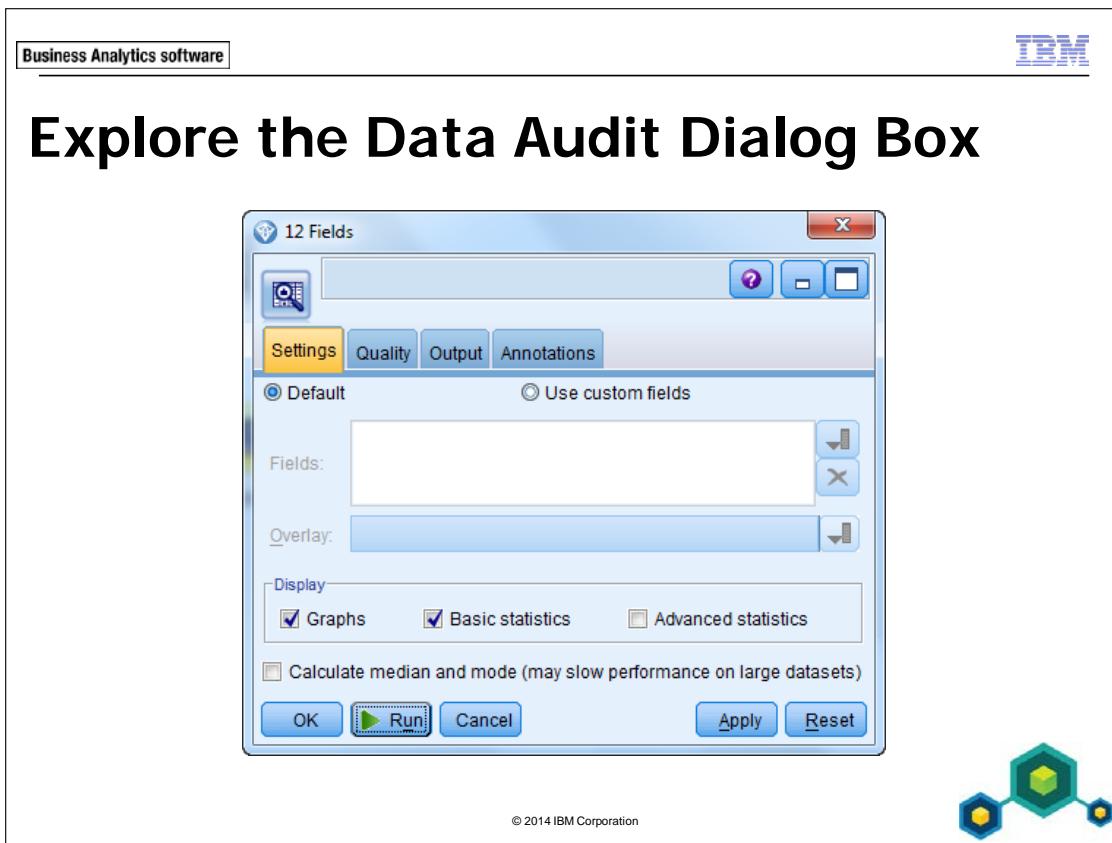
The statistics that are reported depend on the fields' measurement levels. For categorical fields the data audit reports the number of unique values (the number of different categories). For continuous fields the most important statistics are the minimum and the maximum value, as this makes it easy to detect values that are out of range. Also the mean is reported, and more advanced statistics such as standard deviation and skewness. Refer to the online Help for more information.

The data audit node shows different graphs for categorical than for continuous fields. For categorical fields the graph displays the unique values, and the frequency thereof.

For continuous fields it makes no sense to graph all unique values. Think of an income field for 1000 records. A graph displaying all distinct values will produce a graph with 1000 different values, all of frequency 1.

Technically, MODELER will run a distribution graph behind the scenes for categorical fields and a histogram for continuous fields. A histogram bins the values into categories.

The graphs in the data audit report are thumbnails and do not show the full details. If you want more information you can zoom in on a graph by double-clicking it.



The Data Audit node (located in the Output palette) produces a data audit. By default all fields, except typeless fields, will be included in the data audit report. Distribution graphs and histograms will be displayed by default and also statistics. If you want to have more advanced statistics, such as standard errors, enable the Advanced statistics option.

By default the median and the mode are not included in the report, because computation may be time consuming.

The Quality tab gives you control over how you want to detect extreme values. Refer to the online Help or the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for more information.

## Use the Statistics Node and Graphs Nodes for Reporting

- Distribution graphs can also be requested by using the Distribution node (Graphs palette)
- Histograms can also be requested by using the Histogram node (Graphs palette)
- Statistics for continuous fields can also be requested by using the Statistics node (Output palette)

© 2014 IBM Corporation



The Data Audit node produces statistics and graphs. Instead of using the Data Audit node you can use separate nodes to produce this output:

- The Distribution node (Graphs palette) gives a distribution graph, and this is relevant for categorical fields only.
- The Histogram node (Graphs palette) gives a histogram, and this is relevant for continuous fields only.
- The Statistics node (Output palette) reports statistics, and this is relevant for continuous fields only.

The added value of using separate nodes is that they have extra options. Refer to the *Looking for Relationships* module in this course for more details.

## Describe Types of Invalid Values

- Non-allowable values
- Undefined (\$null\$) values
- White space (for string fields)

© 2014 IBM Corporation



MODELER regards the following values as invalid:

- values that are not in the allowed set of values (for categorical fields), or values that are out-of-range (for continuous fields)
- undefined values that are represented as \$null\$ in MODELER
- white space values such as a series of spaces for string values

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

# Invalid Values Illustrated

| ID | AGE      | MARITAL     |
|----|----------|-------------|
| 1  | 18       | married     |
| 2  | 12       | married     |
| 3  | 28       |             |
| 4  | \$null\$ | separated   |
| 5  | 40       | unknown     |
| 6  | 41       | married     |
| 7  | 47       | not married |
| 8  | 51       | not married |
| 9  | -1       | married     |
| 10 | 60       | not married |

© 2014 IBM Corporation



In the example dataset shown on this slide, record #4 and record #9 have an invalid value for AGE (\$null\$ and -1, respectively). The question is whether value 12 is within the range. If the dataset should be comprised of persons aged between 18 and 85, the value 12 would also be out-of-range. For MARITAL, assuming that the allowable values are married, separated and not married, records #3 and #5 have an invalid value (white space and unknown, respectively).

## Actions for Invalid Values

- Nullify
- Coerce
- Discard
- Warn
- Abort

© 2014 IBM Corporation



When an invalid value is found, MODELER's checking process can perform one of many possible actions.

When you choose to nullify values, invalid values are converted to the undefined value (\$null\$). You can also coerce invalid values to valid values. Which value this is depends on the field's measurement level. For flag fields a value other than True or False will be set to False. An invalid value for a nominal field or an ordinal field will be replaced with the first member of the set's values. For continuous fields a value less than the lower bound is set to lower bound of the range and a value greater than the upper bound is set to the upper bound of the range; undefined (\$null\$) values are set to the midpoint of the range.

Discarding records with an invalid value is another option, but you will probably prefer that a warning is issued first, so that you can examine the invalid values. Also, you can stop execution when the first invalid value is encountered (the Abort option).

**Business Analytics software**

**IBM**

## Invalid Values and Actions Illustrated

| ID | AGE      | MARITAL     |
|----|----------|-------------|
| 1  | 18       | married     |
| 2  | 12       | married     |
| 3  | 28       |             |
| 4  | \$null\$ | separated   |
| 5  | 40       | unknown     |
| 6  | 41       | married     |
| 7  | 47       | not married |
| 8  | 51       | not married |
| 9  | -1       | married     |
| 10 | 60       | not married |

→

| ID | AGE | MARITAL     |
|----|-----|-------------|
| 1  | 18  | married     |
| 3  | 28  | \$null\$    |
| 5  | 40  | \$null\$    |
| 6  | 41  | married     |
| 7  | 47  | not married |
| 8  | 51  | not married |
| 10 | 60  | not married |

© 2014 IBM Corporation



This slide shows an example of checking the values for AGE and MARITAL.

For AGE, a [18, 85] range of valid values has been defined. The selected action is to discard records with invalid values. Records #2 (value 12), #4 (the undefined value, which is regarded as invalid by default), and #9 (value -1) have out-of-range values, and thus will be discarded from the dataset.

For MARITAL, the valid values are married, separated and not married. The selected action is to replace invalid values with the undefined (\$null\$) value. Records #3 (white space) and #5 (value unknown) have values outside the allowable set of values and their values for MARITAL will be nullified.

## Dealing with Missing Data: Blanks

- Blanks represent missing values in MODELER
- When a blank value is defined for a certain field, MODELER doesn't treat the value as invalid
- Types of blanks:
  - user-defined values
  - undefined values
  - white space values (string fields)

© 2014 IBM Corporation



Sometimes out-of-range values are out of the range for a reason. For example, your database administrator can plug in the value -1 when a value is unknown. In survey research it is common practice to have a value such as 99 when the respondent refuses to answer a question. Such user-defined missing values are outside the range intentionally and you may want to treat these values different from values that are out of range without a particular reason. By declaring these values as blank values MODELER will not treat the value as invalid, but just as missing information.

The same goes for undefined (\$null\$) values. By default undefined values are regarded as invalid by MODELER and an action such as Discard will remove records with undefined values from the dataset. When this is too rigorous, but you want to Discard records which have out-of-range values, you want MODELER to regard undefined values as valid values. Declaring undefined (\$null\$) values as blanks will accomplish this.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

For string fields there are two other values that can be considered to be missing values:

- empty strings: the string value contains nothing, not even a space
- a series of spaces: the string value consists of spaces such as " "

MODELER refers to both types a white space. In other words, the # of records which have a white space value for field X= # of records which have an empty string for field X+ # of records which have a series of spaces for field X.

All in all, you will have more flexibility in processing data when user-defined missing values, undefined (\$null\$) values and white space values for string fields are declared as blank values.

**Business Analytics software**

**IBM**

## Blanks and Actions Illustrated

| ID | AGE      | MARITAL     |
|----|----------|-------------|
| 1  | 18       | married     |
| 2  | 12       | married     |
| 3  | 28       |             |
| 4  | \$null\$ | separated   |
| 5  | 40       | unknown     |
| 6  | 41       | married     |
| 7  | 47       | not married |
| 8  | 51       | not married |
| 9  | -1       | married     |
| 10 | 60       | not married |



| ID | AGE      | MARITAL     |
|----|----------|-------------|
| 1  | 18       | married     |
| 3  | 28       |             |
| 4  | \$null\$ | separated   |
| 5  | 40       | unknown     |
| 6  | 41       | married     |
| 7  | 47       | not married |
| 8  | 51       | not married |
| 9  | -1       | married     |
| 10 | 60       | not married |

© 2014 IBM Corporation



This slide shows how blank definitions affect the actions. For AGE, the following settings were in place:

- valid range: [18, 85]
- blanks: null and the value -1
- action for invalid values: Discard the record

For MARITAL, the settings were:

- allowable values: married, separated and not married
- blanks: white space and the value unknown
- action for invalid values: Nullify the field

Record #2 is discarded, because AGE is out of range (and 12 is not declared as blank). Records #4 and #9 will not be discarded, because of the blank definitions for AGE. The values for records #3 and #5 for MARITAL will not be replaced with \$null\$, because of the blank definitions for this field.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Business Analytics software

IBM

## Explore the Types Tab Dialog Box

| Field  | Measurement | Values         | Missing | Check   |
|--------|-------------|----------------|---------|---------|
| ID     | Typeless    |                |         | None    |
| AGE    | Continuous  | [18,85]        | *       | Discard |
| INCOME | Continuous  | [15005,599...] |         | None    |
| GENDER | Flag        | m/f            |         | None    |

© 2014 IBM Corporation

Allowable values, blanks and actions can be specified in the Types tab of a data source node, or in the Types tab of a separate Type node. When you want to check the values or take action at the moment of importing the data, it is recommended to set the allowable values and actions in the Types tab of the data source at hand, because data errors will then be spotted immediately. When data must be checked downstream you will use the Types tab in a Type node.

Selecting the field and clicking in the corresponding cell in the Values column will bring up the Values sub dialog box where you specify the set of allowable values (for categorical fields), or the range of allowable values (for continuous fields). You can also specify the blank values, including the null value, in this sub dialog box. For string fields there is an extra option to declare a white space value as blank.

In the Check column, select the action that you want to be taken when an invalid value is encountered.

**Business Analytics software**

**IBM**

## Reporting Blanks in a Data Audit

| ID | AGE      | MARITAL     |
|----|----------|-------------|
| 1  | 18       | married     |
| 3  | 28       |             |
| 4  | \$null\$ | separated   |
| 5  | 40       | unknown     |
| 6  | 41       | married     |
| 7  | 47       | not married |
| 8  | 51       | not married |
| 9  | -1       | married     |
| 10 | 60       | not married |

→

| Field   | Null | Empty string | White space | Blanks |
|---------|------|--------------|-------------|--------|
| ID      | 0    | 0            | 0           | 0      |
| AGE     | 1    | 0            | 0           | 2      |
| MARITAL | 0    | 1            | 1           | 2      |

© 2014 IBM Corporation

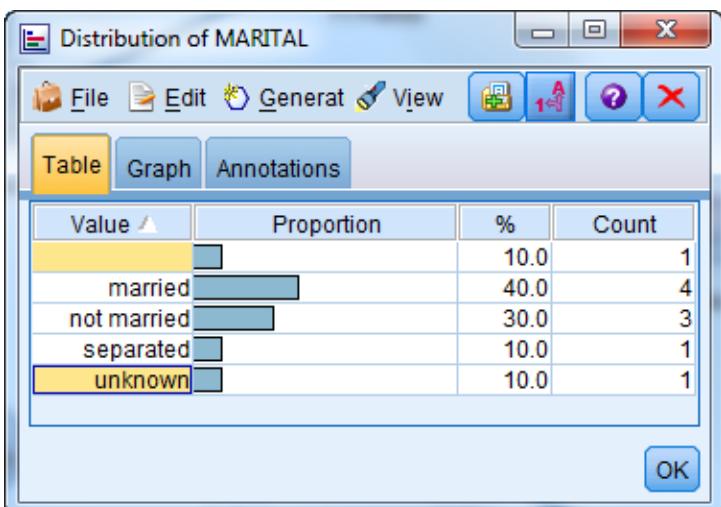


When you have run a Data Audit node, the Data Audit output will include report the blanks on the Quality tab.

For the dataset shown on this slide -1 and null are defined as blanks for AGE, and white space and unknown are defined as blank for MARITAL.

The AGE field has 7 valid records, 2 records have blank values. One of these blank values is the undefined (\$null\$) value (record #4), which leaves one record with a user-defined blank value (record #9, with value -1).

MARITAL also has 7 valid record and 2 blank values. One of these two records, record #3, has an empty string. There are no other records having an empty string for MARITAL. Also, there are no records which have one or more spaces as value for MARITAL. This means that the number of records with a white space value for MARITAL equals  $1 + 0 = 1$ , the number reported in white space column. Record #5 has the value unknown, which was declared as blank.



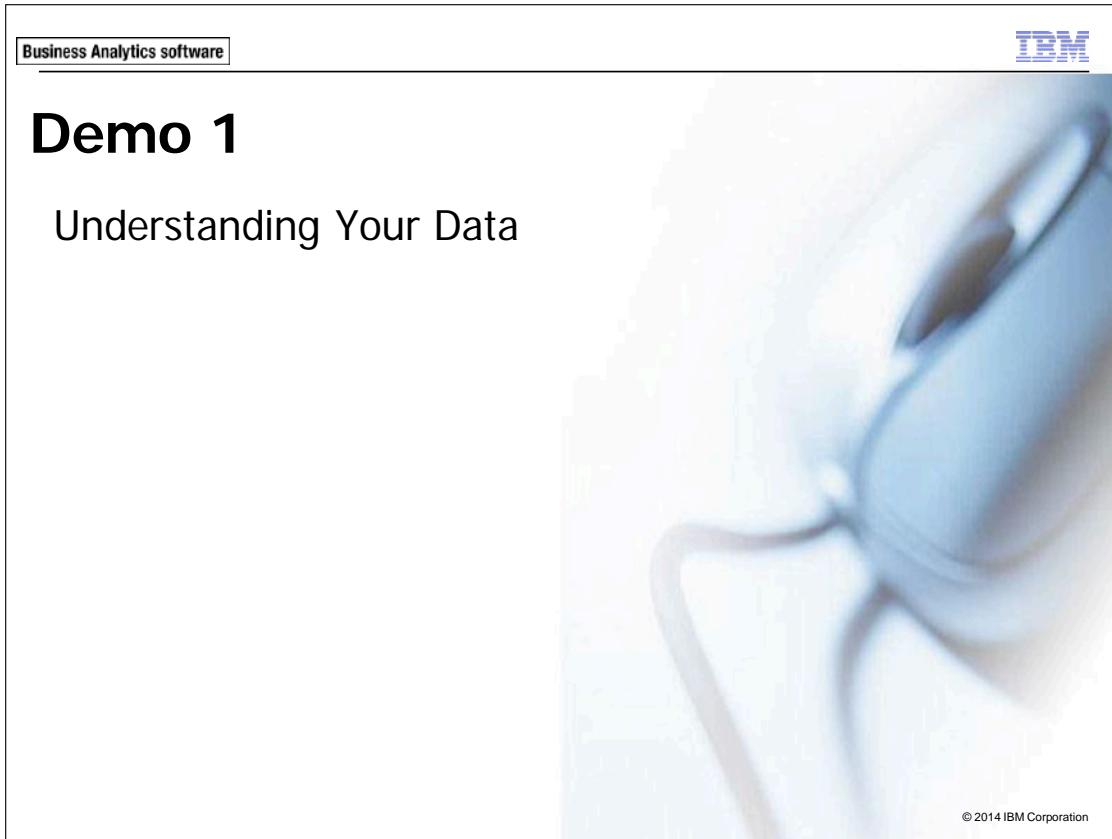
The screenshot shows a dialog box titled "Distribution of MARITAL". The "Table" tab is selected. The table has four columns: Value, Proportion, %, and Count. The data is as follows:

| Value       | Proportion | %    | Count |
|-------------|------------|------|-------|
| married     | 0.40       | 40.0 | 4     |
| not married | 0.30       | 30.0 | 3     |
| separated   | 0.10       | 10.0 | 1     |
| unknown     | 0.10       | 10.0 | 1     |

At the bottom right of the dialog is an "OK" button. Below the dialog, there is a small decorative graphic of three hexagonal nodes connected by lines.

Defining blank values does not necessarily have an effect on how MODELER treats that value for a field in certain analyses. For example, if the white space value and the value unknown are declared as blank, a distribution will display these blank values, which may come as a surprise. MODELER certainly knows that these values are missing, but some nodes, the nodes in the Graphs palette for example, ignore the blank definitions.

The question is also how MODELER will handle blank values when you derive new fields. For example, suppose that you have an INCOME field with -1 defined as blank value and that you derive a field INCOME\_1000 as INCOME/1000. What will then be the result when a record has -1 for INCOME, with -1 is declared as a blank value? You will find the answer in the *Deriving and Reclassifying Fields* module in this course.



Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **demo\_understanding\_your\_data.str**, located in the **05-Understanding\_Your\_Data\Start Files** sub folder.

The following files are used in the demo:

- **telco x customer data.xlsx** – a Microsoft Excel file storing data on customers
- **05-Understanding\_Your\_Data/Start Files/**  
**demo\_understanding\_your\_data.str** - MODELER stream file that imports the data, sets the fields' measurement levels, and instantiates the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

# Demo 1: Understanding Your Data

## Purpose:

You work for a telecommunications firm as a data miner. You have imported data, and you need to assess the quality of the data and define blanks where appropriate.

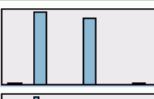
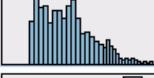
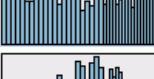
### Task 1. Auditing the data.

1. Add a **Data Audit** node (Output palette) downstream from the **Type** node.

Note: The Type node is already on the stream canvas when you have opened **demo\_understanding\_your\_data.str**, located in the **05-Understanding\_Your\_Data\Start Files** sub folder.

2. Run the **Data Audit** node.

A section of the results appear as follows:

| Field        | Sample Graph  | Measurement | Min        | Max        | Mean   | Std. Dev | Skewness | Unique | Valid |
|--------------|---|-------------|------------|------------|--------|----------|----------|--------|-------|
| gender       |  | Nominal     | --         | --         | --     | --       | --       | 6      | 31781 |
| age          |  | Continuous  | -1.000     | 82.000     | 30.313 | 12.873   | 0.822    | --     | 31781 |
| region       |  | Nominal     | 1.000      | 4.000      | --     | --       | --       | 4      | 31781 |
| connect_date |  | Continuous  | 2003-01-01 | 2006-12-31 | --     | --       | --       | --     | 31781 |
| end_date     |  | Continuous  | 2004-01-01 | 2010-12-29 | --     | --       | --       | --     | 14686 |

The number of valid values for gender, age, region, connect\_date is 31,781, the number of records in the dataset. The end\_date field only has 14,686 valid values. You will examine this field later.

Notice that the minimum value for age is -1. Also notice that the graph for gender shows more than two bars, which is suspect.

3. Double-click the graph for **gender**.

A Distribution output window opens.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4. In the **Distribution** output window, click the **Count** column header twice to sort the values descending on frequency.

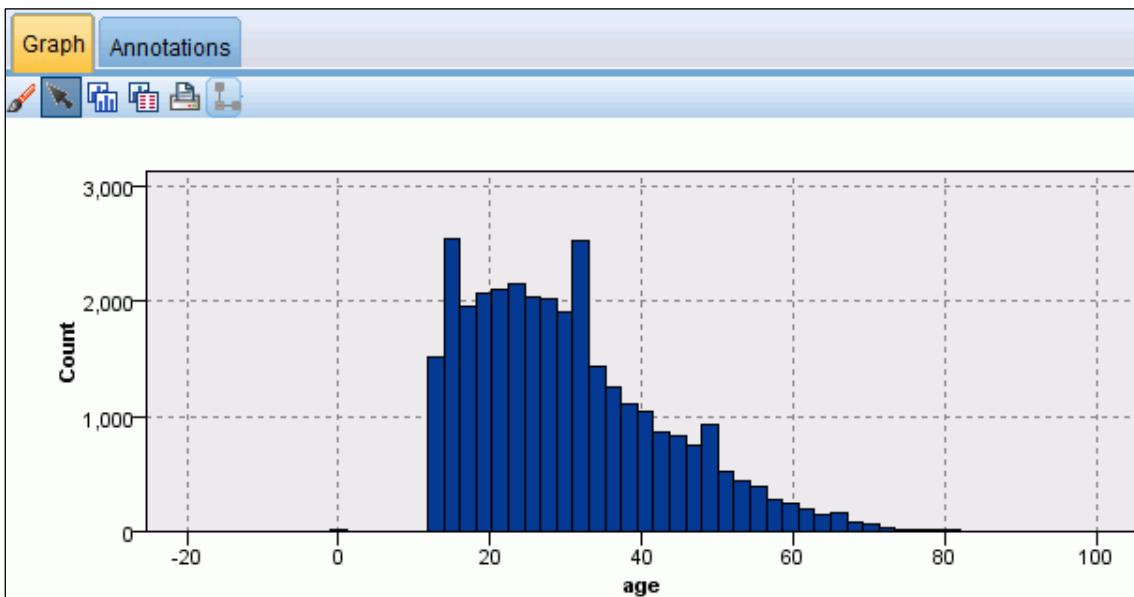
A section of the result appear as follows:

| Value  | Proportion | %     | Count |
|--------|------------|-------|-------|
| Female |            | 49.95 | 15875 |
| Male   |            | 49.88 | 15852 |
| female |            | 0.05  | 16    |
| male   |            | 0.05  | 16    |
| MALE   |            | 0.04  | 13    |
| FEMALE |            | 0.03  | 9     |

The values Female and Male are the values with the highest frequency. Records with spellings other than Female or Male will not be discarded, because that would be a waste of information. The correct approach is to reclassify the values using an appropriate field operations node. The *Deriving and Reclassifying Fields* module in this course will address this issue.

5. Click **OK** to close the **Distribution** output window.  
 6. Double-click the graph for **age**.

A section of the results appear as follows:



Notice the little bar near the value 0 on the x-axis, as it shows that there are a few records with value -1 for age.

7. Click **OK** to close the **Histogram** output window, and then click **OK** to close the **Data Audit** output window.

Leave the stream open for the next task.

## Task 2. Defining valid values and taking action.

In the previous task you have explored the data and you have found that the value -1 was the minimum value for age.

In this task, a message must be issued when an out-of-range value for age is encountered. Hereto, you will define a range of valid values for age.

In this task you will build from the previous stream.

1. Edit the **Type** node, click in the cell for **age** (where it reads [-1.0, 82.0], and then select **Specify** from the menu.

The Values dialog box opens.

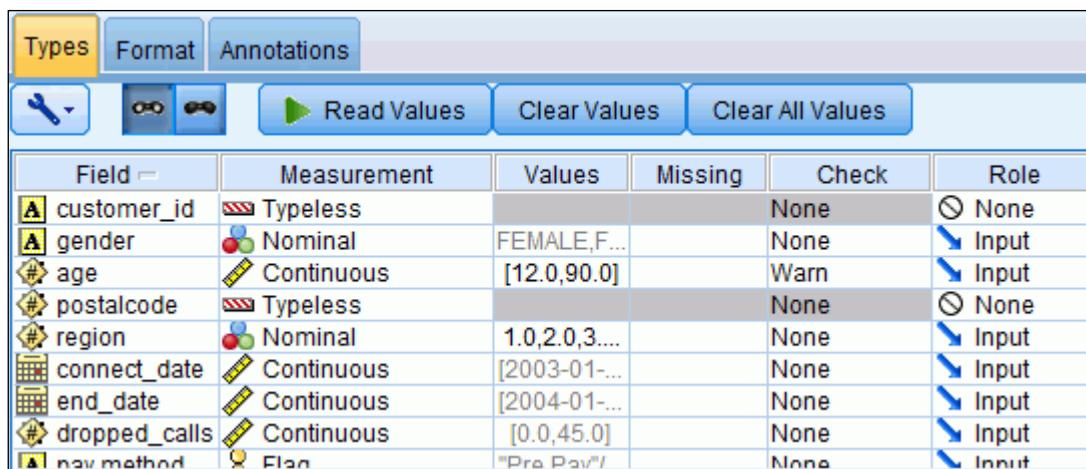
2. In the **Values** dialog box:

- select **Specify values and labels**
- for **Lower**, type **12**
- for **Upper**, type **90**

3. Click **OK** to close the **Values** dialog box.

4. In the **Check** column for **age**, select **Warn**.

A section of the **Types** tab dialog box appear as follows:



The screenshot shows the 'Types' tab of the IBM SPSS Modeler dialog box. The top navigation bar includes tabs for 'Types' (selected), 'Format', and 'Annotations'. Below the tabs are several buttons: a wrench icon for 'Edit Type', a magnifying glass icon for 'Search', a double arrow icon for 'Sync', a play button for 'Read Values', and three buttons for 'Clear Values', 'Clear All Values', and 'Edit Values'. The main area is a table with columns: 'Field', 'Measurement', 'Values', 'Missing', 'Check', and 'Role'. The rows represent different fields with their corresponding types and validation rules:

| Field         | Measurement | Values        | Missing | Check | Role                                   |
|---------------|-------------|---------------|---------|-------|--|
| customer_id   | Typeless    |               |         | None  | <input type="radio"/> None             |
| gender        | Nominal     | FEMALE,F...   |         | None  | <input checked="" type="radio"/> Input |
| age           | Continuous  | [12.0,90.0]   |         | Warn  | <input checked="" type="radio"/> Input |
| postalcode    | Typeless    |               |         | None  | <input type="radio"/> None             |
| region        | Nominal     | 1,0,2,0,3.... |         | None  | <input checked="" type="radio"/> Input |
| connect_date  | Continuous  | [2003-01-...  |         | None  | <input checked="" type="radio"/> Input |
| end_date      | Continuous  | [2004-01-...  |         | None  | <input checked="" type="radio"/> Input |
| dropped_calls | Continuous  | [0.0,45.0]    |         | None  | <input checked="" type="radio"/> Input |
| new_method    | Flag        | "Pre_Pay"/"   |         | None  | <input checked="" type="radio"/> Input |

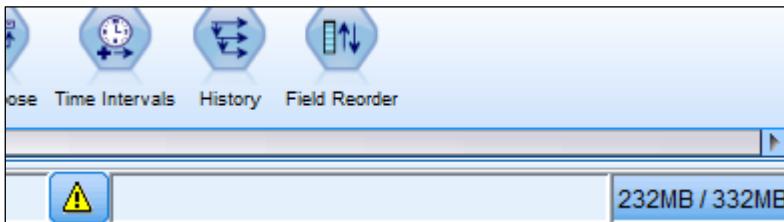
For the effect of the new specifications, you will run the Data Audit node downstream from the Type node.

5. Run the **Data Audit** node.

The value -1 is still reported as minimum value for age, so it seems as if nothing has changed.

6. Click **OK** to close the **Data Audit** output window.

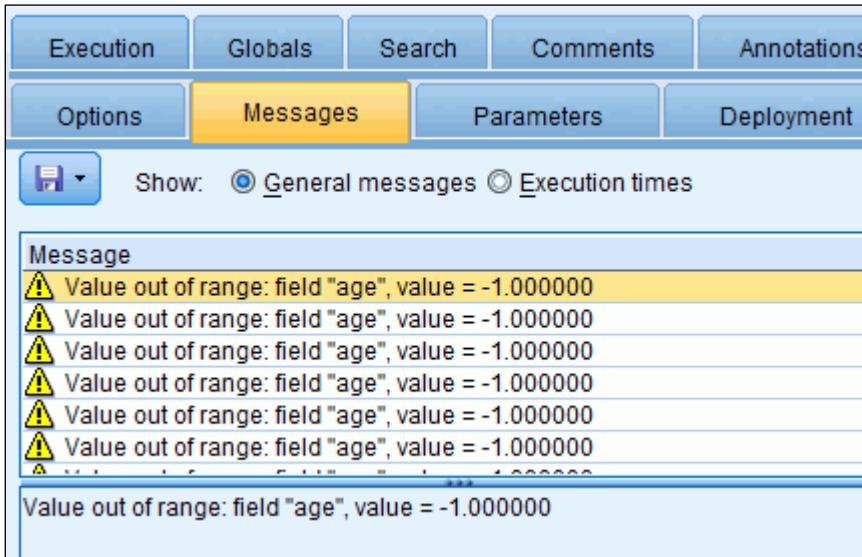
A section of the results appear as follows:



The Show stream messages button in the status bar shows an exclamation mark, indicating you have new messages.

7. Click the **Show stream messages** button.

A section of the result appear as follows:



8. Click **OK** to close the **Stream messages** window.

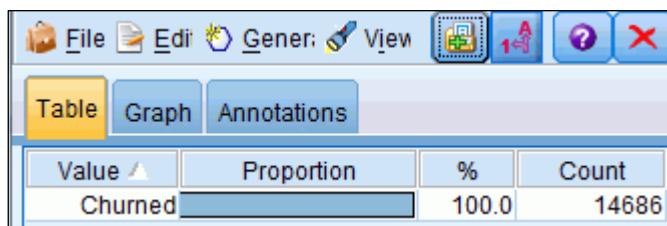
As a second example of dealing with missing values, suppose that records with an undefined (\$null\$) value for end\_date must be discarded.

9. Edit the **Type** node, and then set the action to **Discard** in the **Check** column for **end\_date**.

You will examine the effect of removing these records.

10. Add a **Distribution** graph (Output palette) downstream from the **Type** node.  
 11. Edit the **Distribution** node, select **churn**, and then click **Run**.

A section of the results appear as follows:



To understand what has happened, run a Table on the original data (instructions not shown here).

A section of the results appear as follows:

| Table Annotations |            |               |            |        |          |         |  |
|-------------------|------------|---------------|------------|--------|----------|---------|--|
| ite               | end_date   | dropped_calls | pay method | tariff | handset  | churn   |  |
| 5121              | \$null\$   | 0.000         | Pre Pay    | CA...  | S80      | Active  |  |
| 5122              | \$null\$   | 2.000         | Pre Pay    | CA...  | ASAD1... | Active  |  |
| 5123              | 2007-03-13 | 0.000         | Pre Pay    | CA...  | ASAD90   | Churned |  |
| 5124              | \$null\$   | 1.000         | Pre Pay    | CA...  | S80      | Active  |  |
| 5125              | \$null\$   | 9.000         | Pre Pay    | CA...  | ASAD1... | Active  |  |
| 5126              | \$null\$   | 2.000         | Pre Pay    | CA...  | ASAD1... | Active  |  |
| 5127              | 2009-03-29 | 2.000         | Pre Pay    | CA...  | ASAD90   | Churned |  |
| 5128              | 2005-12-17 | 1.000         | Pre Pay    | CA...  | ASAD90   | Churned |  |
| 5129              | 2007-03-08 | 0.000         | Pre Pay    | CA...  | S80      | Churned |  |

The **end\_date** field is always undefined (\$null\$) for customers who did not churn. When you remove all records with **end\_date** missing you will remove all active customers, making modeling later impossible (to build a model to predict churn, both churners and active customers are needed).

The crucial concern is whether there is a pattern to the missing data such that there is a difference between the records with missing data and those without missing data. If there is, then your model can be misestimated and cannot be applied to the full population of interest. Removing records that have an undefined value for **end\_date** is an extreme example of this.

All in all, you will need to know the business and your data to take the appropriate action.

Records having an undefined end\_date cannot be discarded, so you will undo the action.

12. Edit the **Type** node, and then set action to **None** for **end\_date**.
13. Click **OK** to close the **Type** dialog box.

Leave the stream open for the next task.

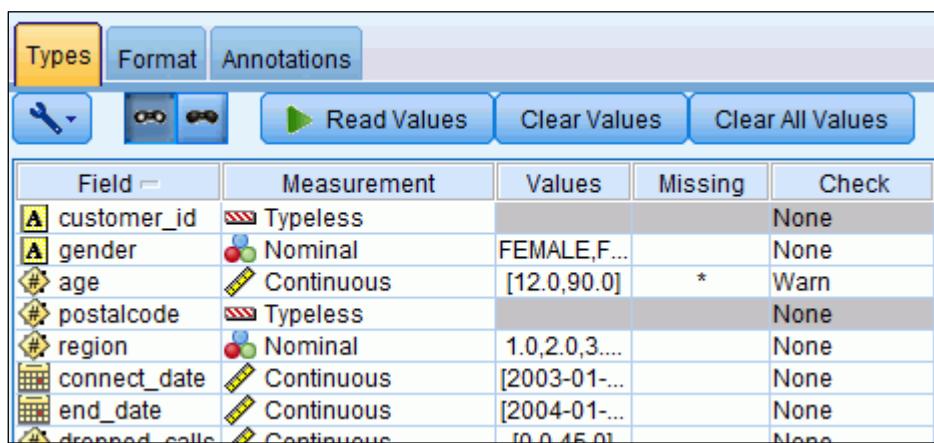
### Task 3. Declaring blank values.

In the previous task, a warning message was issued when a value was out of range for age. However, it is known beforehand that the value -1 can be present in the data, because this value is used when age is unknown. This value needs to be declared as blank value, so that out-of-range messages will only report values that are unexpectedly out-of-range.

In this task you will build from the previous stream.

1. Edit the **Type** node and then:
  - click in the cell for field **age** in the **Missing** column, and click **Specify**
  - enable the **Define blanks** option
  - under **Missing values**, type **-1**
2. Click **OK** to close the **Values** dialog box.

A section of the specifications in the Types tab dialog box appear as follows:



The screenshot shows the 'Types' tab of the SPSS Modeler dialog box. The table lists fields, their measurements, and data type specifications. The 'customer\_id' field is Typeless. The 'gender' field is Nominal. The 'age' field is Continuous, with a measurement range of [12.0, 90.0] and a missing value of \* (Warn). The 'postalcode' field is Typeless. The 'region' field is Nominal. The 'connect\_date' and 'end\_date' fields are Continuous, with 'connect\_date' having a range of [2003-01-...]. The 'dropped\_calls' field is also Continuous. The 'Missing' column shows 'None' for most fields except 'age' which has a warning symbol (\*).

| Field         | Measurement | Values        | Missing | Check |
|---------------|-------------|---------------|---------|-------|
| customer_id   | Typeless    |               | None    |       |
| gender        | Nominal     | FEMALE,F...   | None    |       |
| age           | Continuous  | [12.0,90.0]   | *       | Warn  |
| postalcode    | Typeless    |               | None    |       |
| region        | Nominal     | 1.0,2.0,3.... | None    |       |
| connect_date  | Continuous  | [2003-01-...] | None    |       |
| end_date      | Continuous  | [2004-01-...] | None    |       |
| dropped_calls | Continuous  | [0.0,1E+01]   | None    |       |

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

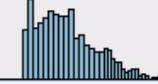
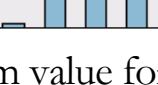
An asterisk in the Missing column for age shows that blanks are defined for this field.

3. Click **OK** to close the **Type** dialog box.

You will verify that warning messages will no longer be issued.

4. Run the **Data Audit** node.

The results appear as follows:

| Field    | Sample Graph  | Measurement | Min    | Max    | Mean   | Std. Dev | Skewness | Unique | Valid |
|----------|---|-------------|--------|--------|--------|----------|----------|--------|-------|
| A gender |  | Nominal     | --     | --     | --     | --       | --       | 6      | 31781 |
| # age    |  | Continuous  | 12.000 | 82.000 | 30.327 | 12.858   | 0.829    | --     | 31766 |
| ¤ region |  | Nominal     | 1.000  | 4.000  | --     | --       | --       | 4      | 31781 |

Minimum value for age is 12, instead of -1 that you had in the previous task.

5. Click **OK** to close the **Data Audit** output window.

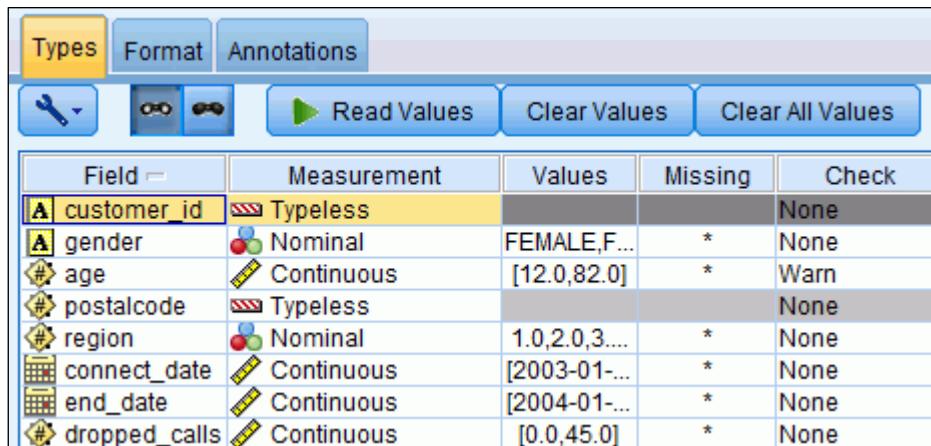
The Show stream messages button does not show an exclamation mark, which indicates that no out-of-range values are reported.

In the data checking process you probably do not want to discard records, or have a warning message issued, when undefined (\$null\$) values are encountered. You will declare the undefined value as blank for all fields.

6. Edit the **Type** node, and then:

- right-click any field, and choose **Select All** from the context menu
- right-click any field, and choose **Set Missing\On (\*)** from the context menu

A section of the results appear as follows:



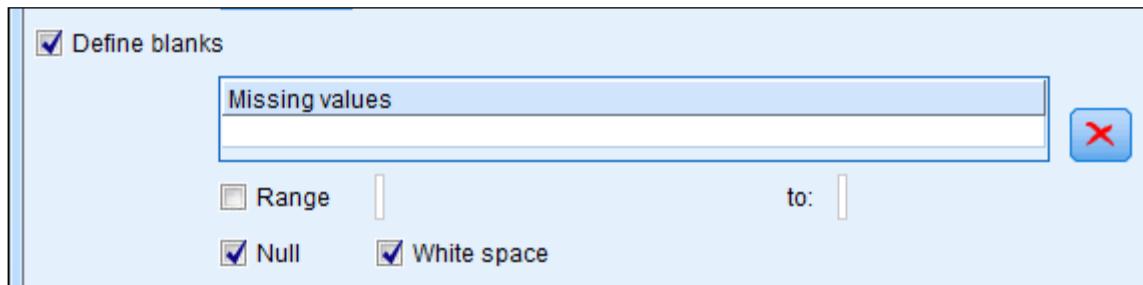
The screenshot shows the 'Types' tab of the 'Values' dialog box. It lists various fields with their measurement types and other properties. Fields include customer\_id (Typeless), gender (Nominal), age (Continuous), postalcode (Typeless), region (Nominal), connect\_date (Continuous), end\_date (Continuous), and dropped\_calls (Continuous). The 'Missing' column contains an asterisk (\*) for most fields, except for customer\_id and postalcode which have 'None'. The 'Check' column shows 'None' for all fields.

| Field         | Measurement | Values        | Missing | Check |
|---------------|-------------|---------------|---------|-------|
| customer_id   | Typeless    |               | None    |       |
| gender        | Nominal     | FEMALE,F...   | *       | None  |
| age           | Continuous  | [12.0,82.0]   | *       | Warn  |
| postalcode    | Typeless    |               | None    |       |
| region        | Nominal     | 1.0,2.0,3.... | *       | None  |
| connect_date  | Continuous  | [2003-01-...  | *       | None  |
| end_date      | Continuous  | [2004-01-...  | *       | None  |
| dropped_calls | Continuous  | [0.0,45.0]    | *       | None  |

All fields, except the typeless fields, have an asterisk in the Missing column, meaning blanks are defined for each of them. To examine the difference with the situation before, examine the blank definitions for one of the fields.

- Click in the cell for **region**, **Missing** column, and then select **Specify**.

A section of the dialog box appear as follows:



The Define blanks option is enabled, and also the Null option. This means that null values are declared as blank values for this field, and the same goes for the other fields.

- Click **OK** to close the **Values** dialog box, and then click **OK** to close the **Type** node.

This completes the demo for this module. You will find the solution results in **demo\_understanding\_your\_data\_completed.str**, located in the **05-Understanding\_Your\_Data\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Is the following statement true or false? The Data Audit node will select the appropriate graph and statistics, depending on the measurement level of the field.

- A. True
- B. False

Question 2: Is the following statement true or false? A Data Audit node for a field of measurement level Typeless will display the field's mean and standard deviation.

- A. True
- B. False

Question 3: Which of the following statements are correct? For a continuous field, the Data Audit node produces:

- A. a histogram graph
- B. a distribution graph
- C. the minimum and maximum value
- D. the mean

Question 4: Is the following statement true or false? An undefined (\$null\$) value is considered an invalid value by default.

- A. True
- B. False

Question 5: Is the following statement true or false? Nominal fields cannot be checked for invalid values.

- A. True
- B. False

Question 6: A field named NUMBER OF CHILDREN is read from a certain data source, and running a Table node shows the value 999 for this field. It appears that this value is used when the number of children is unknown. True or false: The best choice for the Check column is then to coerce values (with a defined range of [0, 12]).

- A. True
- B. False

Question 7: Is the following statement true or false? The number of records that has a white space value on field X is always greater than or equal to the number of records that has an empty string value.

- A. True
- B. False

Question 8: Which of the following statements are correct?

- A. By default an undefined (\$null\$) value is not considered to be a blank value.
- B. By default a white space value for a string field is not considered to be a blank value.
- C. Suppose that the data show a value "unknown" for a field region. This value can be defined as blank value.
- D. For a continuous field you can specify discrete values (for example, -9, -99) as blank values, or a range of values (for example, 99 thru 999) as blank values, but not both.

Question 9: Which of the following statements are correct? Suppose that your dataset includes a string field and that you enable the White space as blank option. Then:

- A. a series of spaces will be treated as blank value
- B. an empty string will be treated as blank value
- C. the value "unknown" will be treated as blank value
- D. the value "white space" will be treated as blank value

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Answers to questions:

Answer 1: A. True. Output is determined by the measurement level of a field.

Answer 2: B. False. Typeless fields will not be included in the output of a Data Audit.

Answer 3: A, C, D.

Answer 4: A. True. Undefined (\$null\$) values are considered to be invalid, unless there are declared as blank.

Answer 5: B. False. For nominal fields, type the valid values, and any value different from those will be marked as invalid.

Answer 6: B. False. Coercing will replace 999 with the maximum value, 12. However, this choice is arbitrary. A better choice is it to nullify the 999 value.

Answer 7: True. White space = empty string ("") + white space (" " ).

Answer 8: A, B, C.

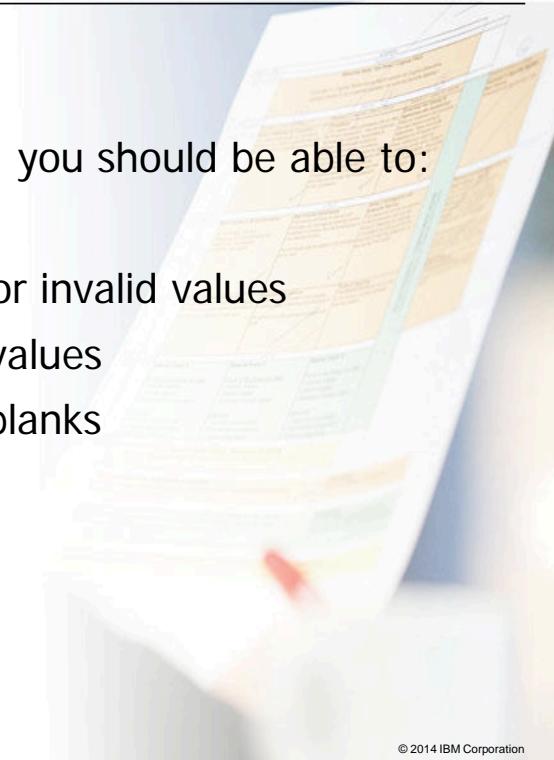
Answer 9: A, B.

Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - audit the data
  - explain how to check for invalid values
  - take action for invalid values
  - explain how to define blanks



© 2014 IBM Corporation

# Workshop 1

## Understanding Your Data



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

The following (synthetic) file is used in this workshop:

- **ACME customer data.txt** – a text file, storing data on customers for a (fictitious) company named ACME

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

5-31

## Workshop 1: Understanding Your Data

You are working at ACME, a company selling sport products. ACME wants to promote a new product through direct mail. ACME has sent out a test mailing and has collected data on the response for this test mailing. It is your job, in preparation of building models later, to examine the quality of one of ACME's datasets, and to take corrective action where needed.

- Import the data from **ACME customer data.txt**, run a Table node to view the data, and audit the data.

What is the number of records in the dataset?

Do you trust the estimate for the mean AGE?

What is the number of valid records for AGE? Can you explain the difference between the number of records in the dataset and the number of valid records for AGE?

- Set a [15, 75] valid range for AGE. Also, ensure that a warning is issued when an out-of-range value is encountered for AGE.

Which values are encountered in the data that are out- of-range?

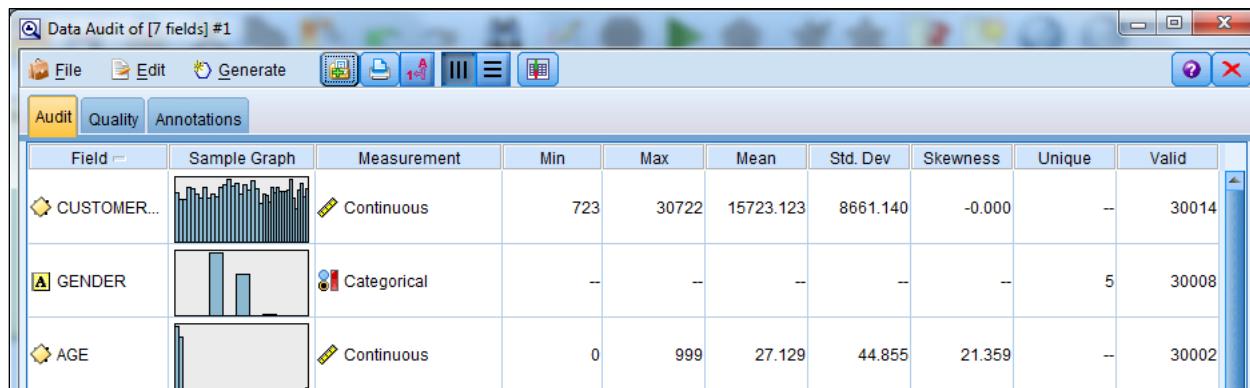
- Declare values 0 and 999 as blank values for AGE. Also, ensure that the \$null\$ value is declared as blank. (Keep the same valid range for AGE as before ([15, 75]).)

What is the mean AGE? Is this statistic correct?

# Workshop 1: Tasks and Results

## Task 1. Explore the data.

- Use the **Var. File** node to import data from **ACME customer data.txt** (use default settings for the import).
- Running a **Table** node shows that the dataset has 30,014 records.
- Running a **Data Audit** node shows that AGE ranges from 0 to 999, with mean AGE 27.1. This statistic is incorrect, because the values 0 and 999 are included in the computation.

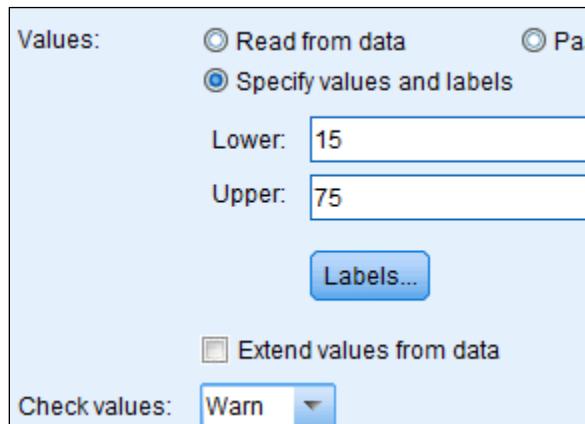


The number of valid records for AGE is 30,002. Apparently, 12 records have an undefined (\$null\$) value for AGE.

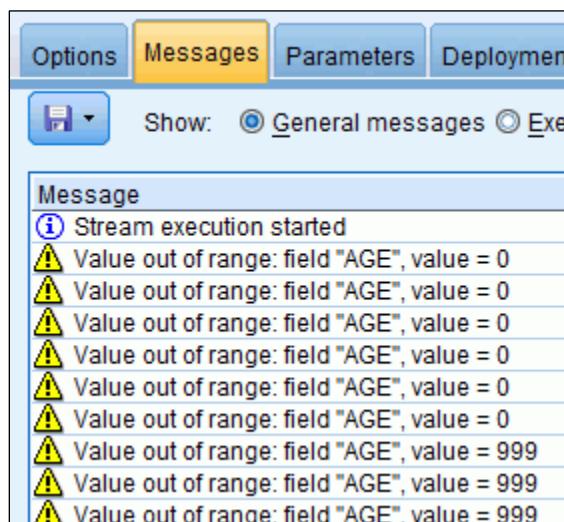
## Task 2. Setting ranges and taking action.

- Add a **Type** node downstream from the **Var. File** node, click in the cell for AGE in the values column and set a [15, 75] range for AGE (alternatively, set the ranges in the Types tab of the Var. File node). Also, set the action to **Warn**.

A section of the specifications in the Values dialog box appear as follows:



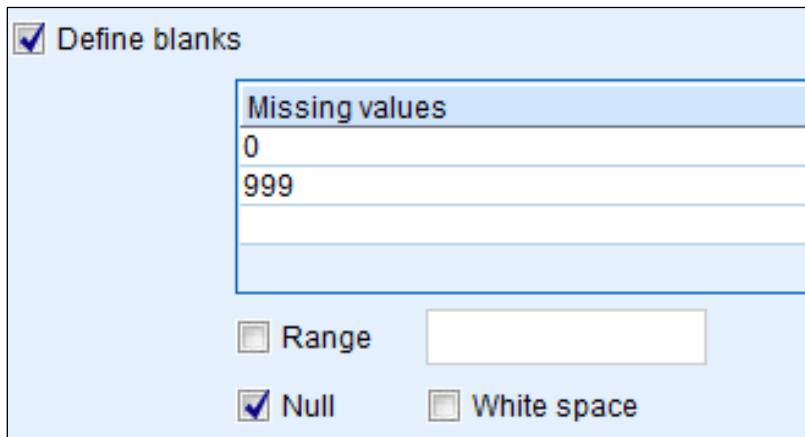
- After having executed a **Table** node downstream the **Type** node, the messages window will report the out-of-range values.
- Click the Show stream messages button. The values 0 and 999 are reported as out-of-range.



It seems as if undefined (\$null\$) values are not reported, but MODELER reports these values as 0 also (which is somewhat confusing).

### Task 3. Declaring blanks.

- Edit the **Type** node, and then select **Specify** in the **Missing** column for **AGE**.
- Enable the **Define blanks** options, and then declare **0** and **999** as blanks. Also, ensure that the **Null** option is enabled.



- Running a **Data Audit** node shows that the mean AGE is 25.1, which is the correct number (values 0 and 999 are not included in the computation because they are declared as blank values).

The stream **workshop\_understanding\_your\_data\_completed.str**, located in the **05-Understanding\_Your\_Data\Solution Files** sub folder provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



## Setting the Unit of Analysis

IBM SPSS Modeler (v16)



**Business Analytics software**

© 2014 IBM Corporation

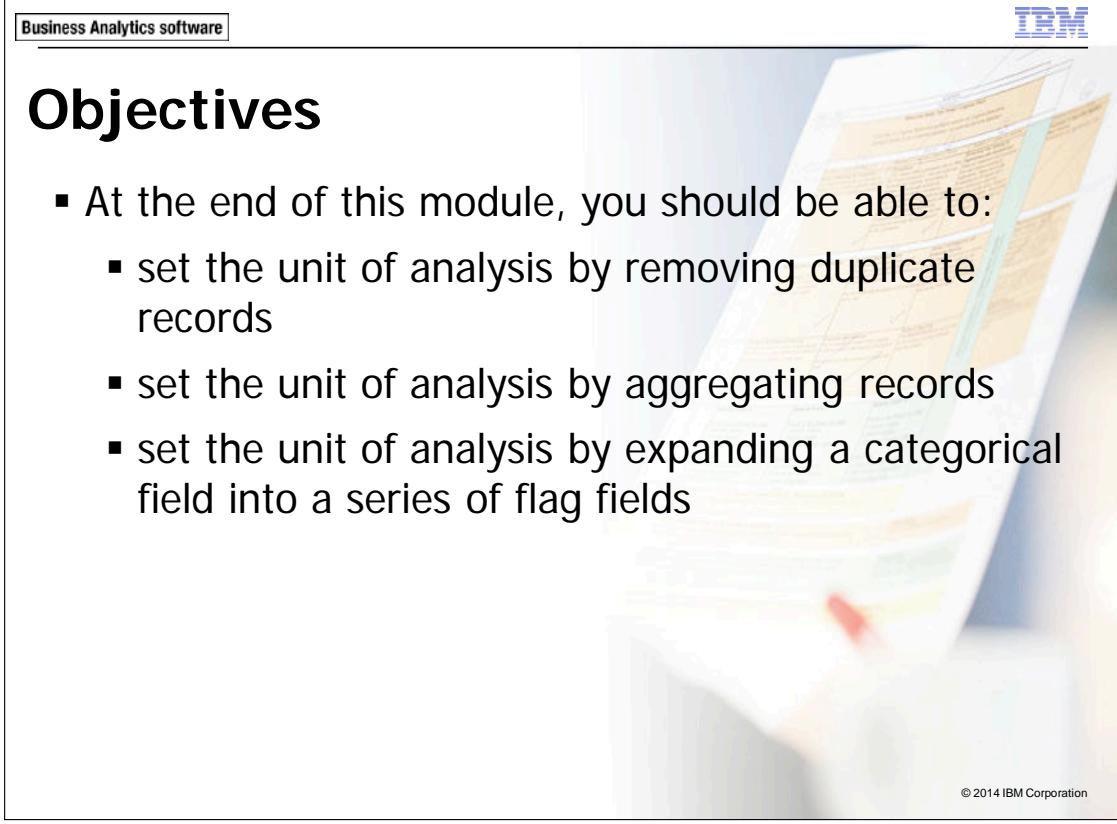
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Objectives

- At the end of this module, you should be able to:
  - set the unit of analysis by removing duplicate records
  - set the unit of analysis by aggregating records
  - set the unit of analysis by expanding a categorical field into a series of flag fields



© 2014 IBM Corporation

After importing and exploring the data, the next task is to set the unit of analysis, one of the tasks in the Data Preparation stage in the CRISP-DM process model. This module presents three methods how you can set the unit of analysis.

Before reviewing this module you should be familiar with:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data
- methods to explore the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

**Business Analytics software**

**IBM**

## The Required Unit of Analysis

| ID | AGE | CHURNED |
|----|-----|---------|
| 1  | 21  | F       |
| 2  | 43  | F       |
| 3  | 56  | F       |
| 4  | 38  | F       |
| 5  | 32  | T       |
| 5  | 32  | T       |

**Data errors**

| ID | PRODUCT | YEAR | REVENUES |
|----|---------|------|----------|
| 1  | A       | 2012 | 100      |
| 2  | B       | 2014 | 50       |
| 2  | C       | 2011 | 200      |
| 3  | B       | 2011 | 50       |
| 3  | C       | 2007 | 200      |
| 3  | D       | 2005 | 10       |

**Transactional data**



© 2014 IBM Corporation

A dataset may not have the required unit of analysis. For example, when your dataset has a field age and you want to compute the mean age, your dataset should have only one record per person.

When the same record appears more than once it may be that your dataset is erroneous and in this case you should remove duplicate records. Another situation where you can expect multiple records is when information is collected at different moments in time. For example, a customer may have purchased products at different times, and each purchase is a record. This is a so-called transactional dataset, which is especially occurring in databases. Another example of information that is collected at different points in time is when a patient visits a hospital several times and each visit is a record.

When you have a transactional dataset as shown on this slide, a business question such as the frequency of co-occurrence of two products cannot be answered in MODELER. To answer that question MODELER requires a data structure where the products make up the fields, not the records.

## Methods to Create Datasets with the Required Unit of Analysis (1 of 3)

**Distinct**

| ID | PR | YEAR | REV |
|----|----|------|-----|
| 1  | A  | 2012 | 100 |
| 2  | B  | 2014 | 50  |
| 2  | C  | 2011 | 200 |
| 3  | B  | 2011 | 50  |
| 3  | C  | 2007 | 200 |
| 3  | D  | 2005 | 10  |

| ID | PR | YEAR | REV |
|----|----|------|-----|
| 1  | A  | 2012 | 100 |
| 2  | B  | 2014 | 50  |
| 3  | B  | 2011 | 50  |

© 2014 IBM Corporation



You can create a dataset with one record per customer in three ways. This slide shows the option to keep one of the records in the group. This option is called Distinct.

In this example the group of records is defined by ID and the first record of each ID is retained. Because the data are sorted descending by year, you will retain the most recent record.

There are different choices in how you want to create the record, which will be presented later in this module.

Business Analytics software

IBM

## Methods to Create Datasets with the Required Unit of Analysis (2 of 3)

**Aggregate**

| ID | PR | YEAR | REV |
|----|----|------|-----|
| 1  | A  | 2012 | 100 |
| 2  | B  | 2014 | 50  |
| 2  | C  | 2011 | 200 |
| 3  | B  | 2011 | 50  |
| 3  | C  | 2007 | 200 |
| 3  | D  | 2005 | 10  |

| ID | REV_SUM | RECORD_COUNT |
|----|---------|--------------|
| 1  | 100     | 1            |
| 2  | 250     | 2            |
| 3  | 260     | 3            |

© 2014 IBM Corporation

Another method is to summarize the information over the records in the group. This option is called Aggregate.

In this example, the group of records is defined by ID, and the sum of the REV field is computed per ID. Also, a field is created that keeps track of the number of records in the source file.

## Methods to Create Datasets with the Required Unit of Analysis (3 of 3)

### SetToFlag

| ID | PR | YEAR | REV |
|----|----|------|-----|
| 1  | A  | 2012 | 100 |
| 2  | B  | 2014 | 50  |
| 2  | C  | 2011 | 200 |
| 3  | B  | 2011 | 50  |
| 3  | C  | 2007 | 200 |
| 3  | D  | 2005 | 10  |

| ID | A | B | C | D |
|----|---|---|---|---|
| 1  | T | F | F | F |
| 2  | F | T | T | F |
| 3  | F | T | T | T |

© 2014 IBM Corporation



The last method is useful to transform a nominal field into a series of flag fields, so that the categories make up the columns of the dataset instead of the rows. This operation is called SetToFlag.

In this example ID defines a group of records, and the nominal field PR with categories A, B, C and D is transformed into a new dataset with one record per ID, with the fields A, B, C and D flagging if one has the particular product.

The method that you will use (Distinct, Aggregate, or SetToFlag) will be determined by your analysis requirements. For example, if you want to keep track of the products purchased, then SetToFlag is the preferred method. You can also choose more than one method and combine the datasets later.

Note: MODELER also provides an option to restructure the dataset from records into fields and retaining the values of the continuous field. In the example above you would create the fields revenues\_A, revenues\_B, and revenues\_C. restructuring data in this way is presented in the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course.

Business Analytics software

IBM

## Distincting Records

- A group of records is defined by key fields
- By default, only one record of the group will be retained
- Use the Distinct node (Record Ops)



© 2014 IBM Corporation



The Distinct node, located in the Record Ops palette, removes duplicate records. Duplicate records are defined by key fields and records with the same values on the key fields are treated as duplicate records. When you define all fields as key you can identify identical records and the data can be cleansed. To deal with transactional data, you should only specify a field such as customer\_id as key.

The Distinct node also enables you to keep all but the first record of the group so that you can zoom in on the duplicate records. Furthermore, the Distinct node has the option to create a composite record, which is presented later.

The screenshot shows the 'Business Analytics software' interface with the 'IBM' logo in the top right. A dialog box titled 'Explore the Distinct Dialog Box (1 of 2)' is open. The 'Settings' tab is selected. Under 'Mode', the option 'Include only the first record in each group' is chosen. In the 'Key fields for grouping' section, 'ID' is listed. Under 'Within groups, sort records by:', there is a table with 'Field' (YEAR) and 'Order' (Descending). The bottom right of the dialog box features a decorative graphic of interconnected hexagons.

The Distinct dialog box has three main tabs: Settings, and Composite and Optimization. The first two tabs are presented in this module. The Optimization tab sets options to improve performance; refer to the online Help for more information.

On the Settings tab, Mode controls how records are formed in the output dataset. MODELER provides three modes. The options Include only the first record in each group and Discard only the first record in each group either include the first record or exclude the first record. The first option is useful for removing duplicate records, the second option is useful for examining the duplicate records. The third option is Create a composite record for each group, which will make the Composite tab available. Refer to the next slide for a presentation.

Groups are defined by the fields specified under Key fields for grouping. You can sort the records within each group, ascending or descending, to ensure that a particular record is the first record in the group. Sorting records in the Distinct node itself makes a separate Sort node upstream from the Distinct node redundant, and is recommended.

Business Analytics software

**IBM**

## Explore the Distinct Dialog Box (2 of 2)

| Field | Fill with values base |
|-------|-----------------------|
| PR    | mostFreq (Ties:first) |
| YEAR  | First record in group |
| REV   | Max                   |

Include record count in field Record\_Count

© 2014 IBM Corporation

On the Composite tab you specify how the new record is created from the source records. For example, you can have the first record's value for field X, and the last record's value for field Y.

This slide shows an example of creating a composite record for a transactional dataset. For a categorical field named PR, the most frequent product will be output for each group of records. Suppose that a customer has 5 records with products A A B C D, then the value that is output will be A. When there are more products with the same frequency of occurrence ,for example AA B B D, then the first product will be output, A in this example.

The Composite tab also provides an option to create an extra field, named Record\_Count by default, which returns the number of input records that were grouped to form an output record.

## Aggregating Records

- A group of records is defined by key fields
- Aggregation will summarize information over all records in each group
- Use the Aggregate node (Record Ops)

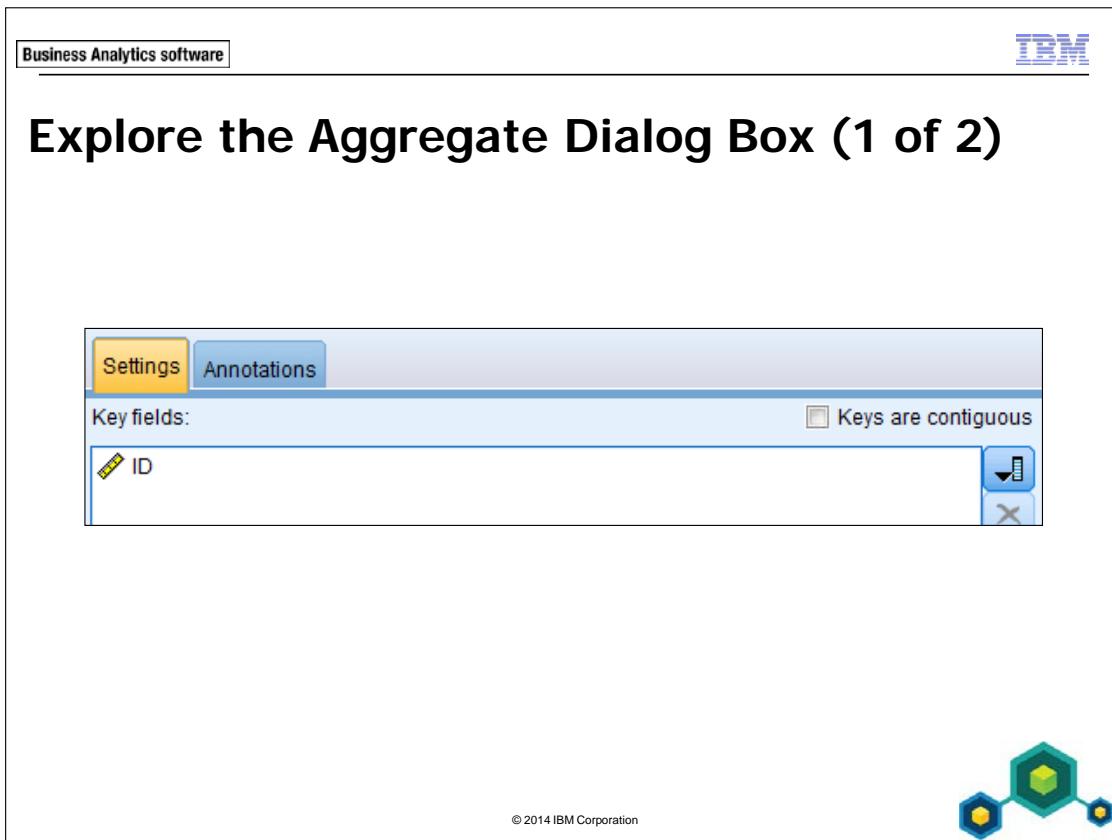


© 2014 IBM Corporation



You can deal with multiple records per customer by using the Distinct node. Another option is that you aggregate information over all the records that a customer has. For example, you can have the mean value of a certain field, computed over all the customer's records. The Distinct node mimics this functionality by creating composite records and offering aggregate statistics such as minimum, mean and maximum, but the Aggregate node has more statistics.

The Aggregate node, located in the Record Ops palette, replaces a group of input records with one aggregated output record. After passing data through the Aggregate node, the overall file structure has changed because the record definition is altered.



As in the Distinct node a group of records is defined by key fields. A key field such as ID will group the records of a customer into one record.

When no key field is specified, the aggregation will be over all the records in the dataset, and thus will result in one record.

If you want to retain a field value that is constant for all records in an aggregate group, such as gender, add the field to the list of key fields.

To improve performance you can enable the Keys are contiguous option if the data are already sorted on the key fields.

Business Analytics software

IBM

## Explore the Aggregate Dialog Box (2 of 2)

**Basic Aggregates**

Aggregate fields:

| Field | Sum                                 | Mean                                | Min                      | Max                                 | SDev                     | Media                    |
|-------|-------------------------------------|-------------------------------------|--------------------------|-------------------------------------|--------------------------|--------------------------|
| REV   | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/>            | <input type="checkbox"/> | <input type="checkbox"/> |
| YEAR  | <input type="checkbox"/>            | <input type="checkbox"/>            | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Default mode:  .  ...   .  ...  M...

New field name extension:

Include record count in field  Record\_Count

**Aggregate Expressions**

| Field     | Expression            |
|-----------|-----------------------|
| RANGE_REV | MAX (REV) - MIN (REV) |

© 2014 IBM Corporation



Under Basic Aggregates, select the fields that you want to aggregate values for and select the statistic(s). You can choose Sum, Mean, Min, Max, SDev (standard deviation), Median, Count (the number of records having a non-\$null\$ value), Variance, 1<sup>st</sup> and 3<sup>rd</sup> Quartile (25th and 75th percentile). Enable the Include record count in field option to create a field that stores the number of records aggregated to form each output record.

Aggregate Expressions operate on the group of records as defined by the key field(s). For example, although the range (the difference between maximum and minimum value) is not available as one of the statistics, it can be created by using the built-in aggregate functions MAX en MIN. When you imported from a database, you can also use the functions supported by your database.

Note: User-defined blanks are included in the computation of the statistics. Suppose, for example, that the value 999 is declared as blank value for a field AGE. Requesting the Max statistic for AGE will return 999 although this value is declared as blank. Also the Mean statistic will be affected by this blank value.

## Setting To Flag Fields

- Transform a categorical field into a series of flag fields
- Especially relevant in the case of transactional data
- Use the SetToFlag node (Field Ops)

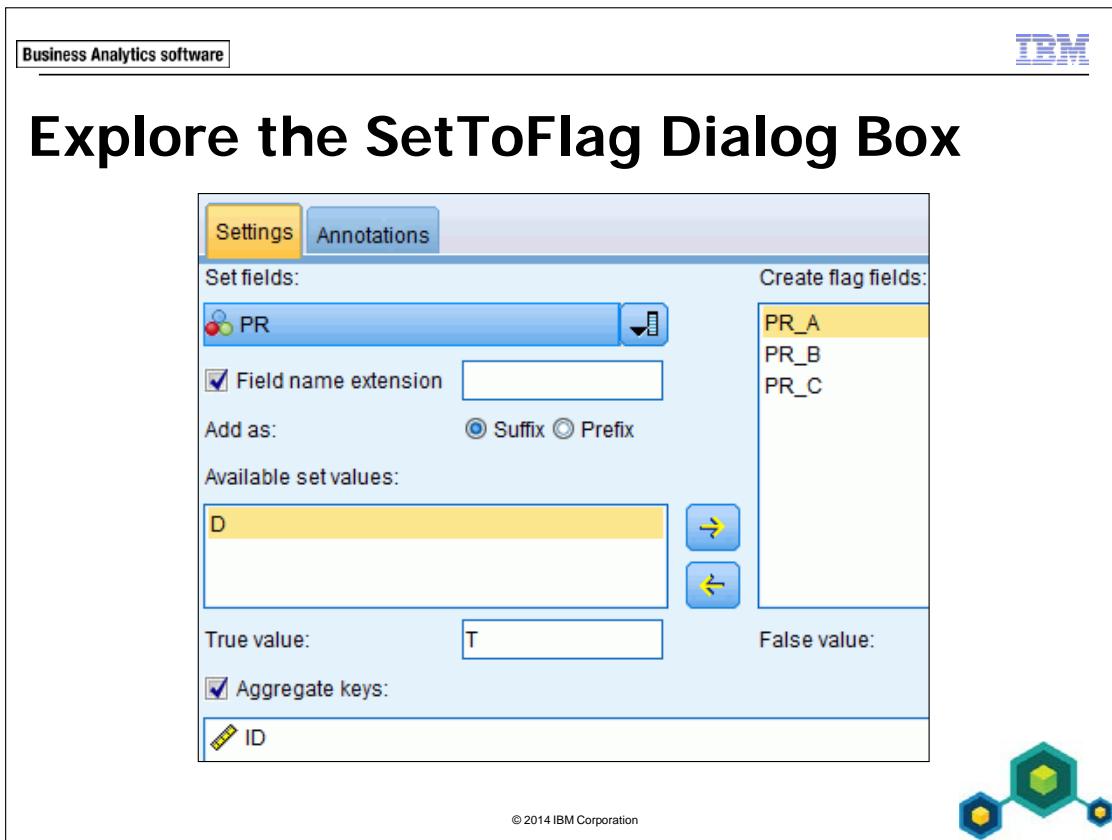


© 2014 IBM Corporation



It may be necessary to convert information held in a categorical field into a collection of flag fields. This is especially true when the data are of a transactional nature.

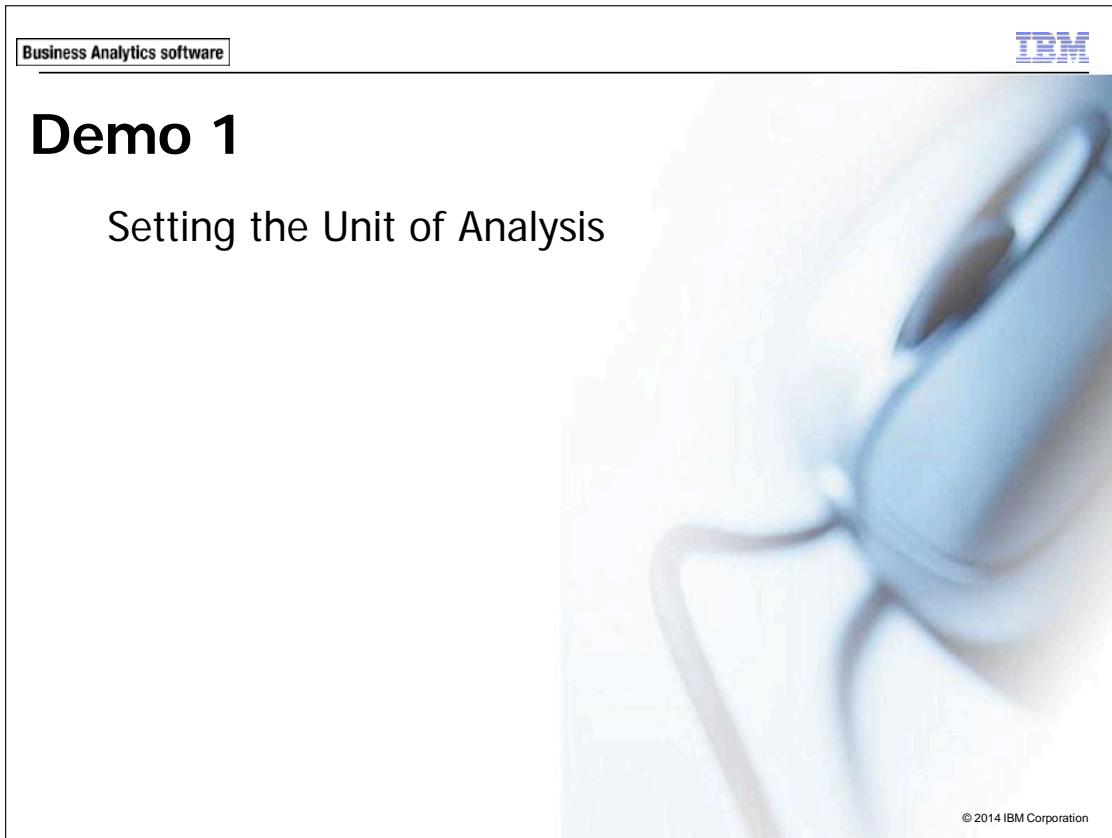
The SetToFlag node, located in the Field Ops palette, expands a nominal field into a series of flag fields, with the option to aggregate the data.



On the Settings tab, under Set fields, select the categorical field that you want to expand in flags. The area under Available set values will be populated with the categories of the selected categorical field, provided that the field is instantiated. If no values are available, add a Type node upstream from the SetToFlag node and instantiate the field. Move the categories that you want to create flag fields for to the Create flag field area. Optionally, extend the field name for the new flag fields, either as suffix or prefix.

By default, the true value and the false value will be T and F, respectively. You can change these values if you want.

Enable the option Aggregate keys and select the appropriate key field(s) to change the unit of analysis. If you do not specify an aggregate key field, the unit of analysis will not change and you will have as many records downstream from the SetToFlag node as you had upstream from the SetToFlag node.



The slide features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The main title "Demo 1" is centered at the top, followed by the subtitle "Setting the Unit of Analysis". The background is a blurred image of a person wearing a hard hat and safety glasses, looking at a blueprint or map. A copyright notice "© 2014 IBM Corporation" is visible in the bottom right corner of the slide area.

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)

Two (synthetic) datasets from a (fictitious) telecommunications firm are used to demonstrate how you can change the unit of analysis:

- **telco x customer info.xlsx**: a Microsoft Excel file storing demographic and churn data
- **telco x products.dat**: a text file with information on which gadgets were purchased (a transactional dataset; a customer has as many records as he has gadgets)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Demo 1: Setting the Unit of Analysis

### Purpose:

You are working as a data miner for a telecommunications firm. It is your job, in order to merge the datasets later, to remove duplicate records in the customer dataset and to transform a transactional dataset into a dataset that has one record per customer.

### Task 1. Cleanse data by removing duplicate records.

1. Import **telco x customer info.xlsx** (a Microsoft Excel 2007 file – ensure that the File type is Excel 2007, 2010 \*.xlsx).

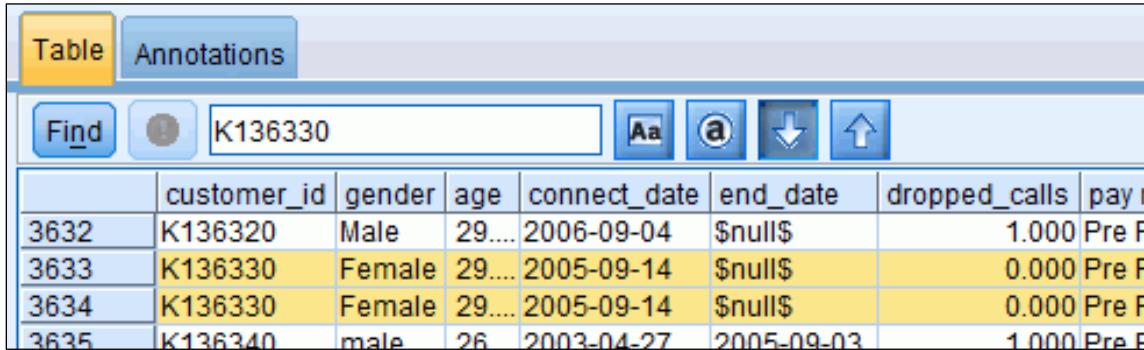
You will explore the data to find out if there are any duplicate records.

2. Add a **Table** node downstream from the **Excel** node, and then run the **Table** node.

The Table output window opens.

3. In the **Table** output window, click the **Search**  button; type **K136330** and then click the **Find**  button.

A section of the results appear as follows:



The screenshot shows a software interface for a Table output window. At the top, there are two tabs: "Table" (which is selected) and "Annotations". Below the tabs is a toolbar with several icons: "Find" (highlighted in blue), a magnifying glass, a search input field containing "K136330", a font size icon, a bold icon, and three navigation icons (down, up, left). The main area displays a table with the following data:

|      | customer_id | gender | age    | connect_date | end_date   | dropped_calls | pay r       |
|------|-------------|--------|--------|--------------|------------|---------------|-------------|
| 3632 | K136320     | Male   | 29.... | 2006-09-04   | \$null\$   |               | 1.000 Pre P |
| 3633 | K136330     | Female | 29.... | 2005-09-14   | \$null\$   |               | 0.000 Pre P |
| 3634 | K136330     | Female | 29.... | 2005-09-14   | \$null\$   |               | 0.000 Pre P |
| 3635 | K136340     | male   | 26     | 2003-04-27   | 2005-09-03 |               | 1.000 Pre P |

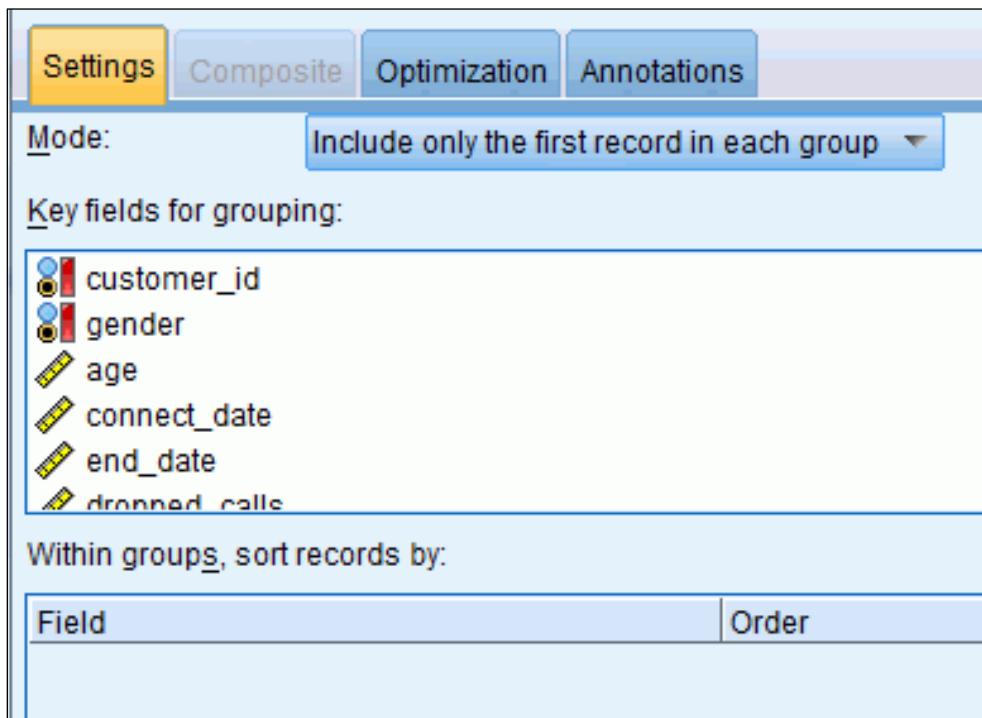
Customer K136330 has two identical records.

4. Click **OK** to close the **Table** output window.
5. Add a **Distinct** node (Record Ops palette) downstream from the **Excel** source node.

6. Edit the **Distinct** node, and then:

- for **Mode**, select **Include only the first record in each group**
- for **Key fields for grouping**, select all fields (records are duplicates if values for all fields are the same)

A section of the specifications in the Distinct dialog box appear as follows:



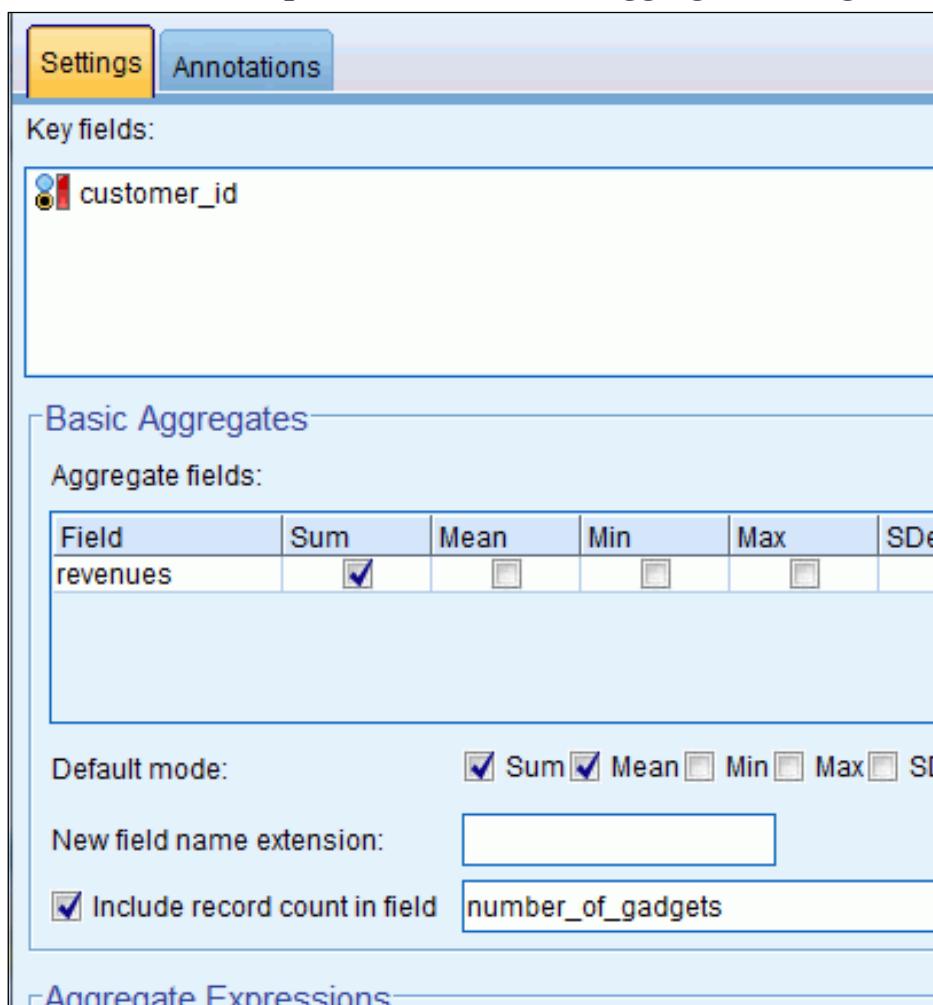
7. Click **OK** to close the **Distinct** dialog box.
8. Add a **Table** node downstream from the **Distinct** node, and then run the **Table** node.  
The Table output window opens.
9. In the **Table** output window, click **Search**; type **K136330** and then click **Find**. You will now find only one record for this customer.
10. Click **OK** to close the **Table** output window.  
Leave the stream open for the next task.

## Task 2. Obtain the number of gadgets bought and the total revenues per customer.

In this task you will build from the previous stream.

1. Import **telco x products.dat** (a tab-delimited text file; use the Var. File node, deselect Comma as delimiter, and select Tab as delimiter).
2. Add an **Aggregate** node (Record Ops palette) downstream from the **Var. File** node.
3. Edit the **Aggregate** node, and then:
  - for **Key field for merge**, select **customer\_id**
  - for **Aggregate fields**, select **revenues**
  - enable the **Include record count in field** option, and rename the field to **number\_of\_gadgets**

A section of the specifications in the Aggregate dialog box appear as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4. Click **Preview**.

A section of the results appear as follows:

| customer_id | revenues_Sum | number_of_gadgets |
|-------------|--------------|-------------------|
| K100010     | 250          | 4                 |
| K100020     | 141          | 3                 |
| K100030     | 455          | 7                 |
| K100040     | 72           | 3                 |
| K100070     | 0            | 0                 |

There is one record per customer, with the sum of revenues stored in the field revenues\_Sum. Also the field number\_of\_gadgets is created, the number of gadgets the customer had in the source dataset. The first customer purchased 4 gadgets, representing total revenue of 250.

5. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Aggregate** dialog box.

Leave the stream open for the next task.

### Task 3. Expand a categorical field into a series of flag fields with one record per customer.

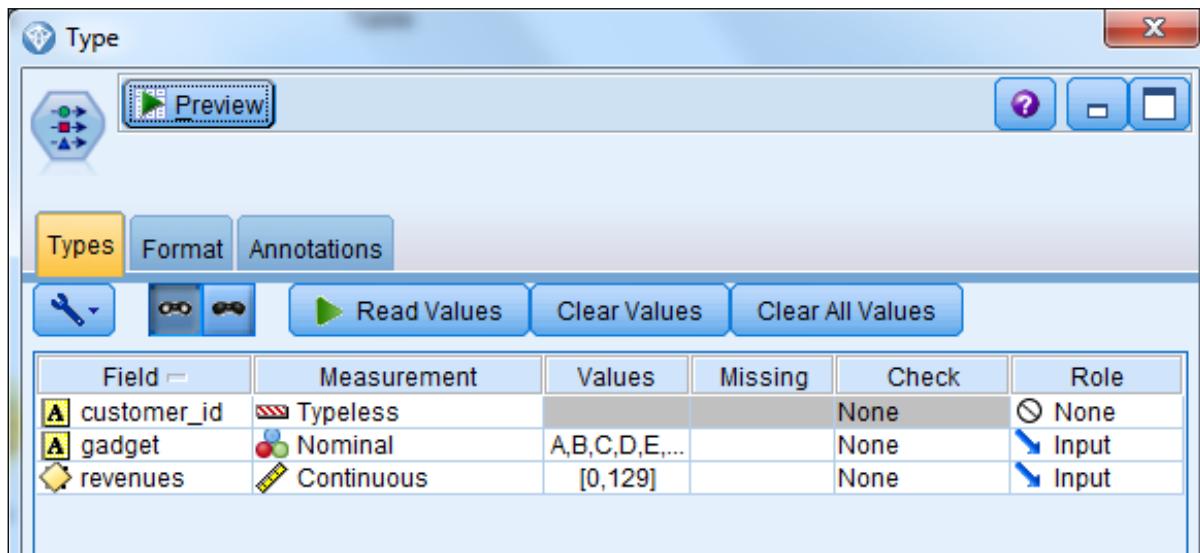
For the dataset comprising product information (**telco x products.dat**), create a dataset where the gadget field is expanded into a series of flags, and at the same time ensure that there is only one record per customer. To accomplish this, you will use the SetToFlag node.

When you use the SetToFlag node, it is important to instantiate the data first, so that the values of the categorical field are available in the SetToFlag node. You can instantiate the data in the Types tab of the Var. File node, or in a separate Type node upstream from the SetToFlag node. In this task the latter is preferred to emphasize that instantiation is a separate step.

In this task you will build from the previous stream.

1. Add a **Type** node (Field Ops palette) downstream from the **Var. File** node.
2. Edit the **Type** node, and then click the **Read Values** button.

A section of the specifications in the Type dialog box appear as follows:



The gadget field is instantiated.

3. Click **OK** to close the **Type** dialog box.
4. Add a **SetToFlag** node (Field Ops palette) downstream from the **Type** node.
5. Edit the **SetToFlag** node, and then:
  - for **Set fields**, select **gadget**
  - click all available set values to the **Create flag fields** area

At this point, the flag fields have been created, but there is not yet one record per customer. To verify that a customer still has multiple records, you will preview the data.

6. Click **Preview**.

A section of the results appear as follows:

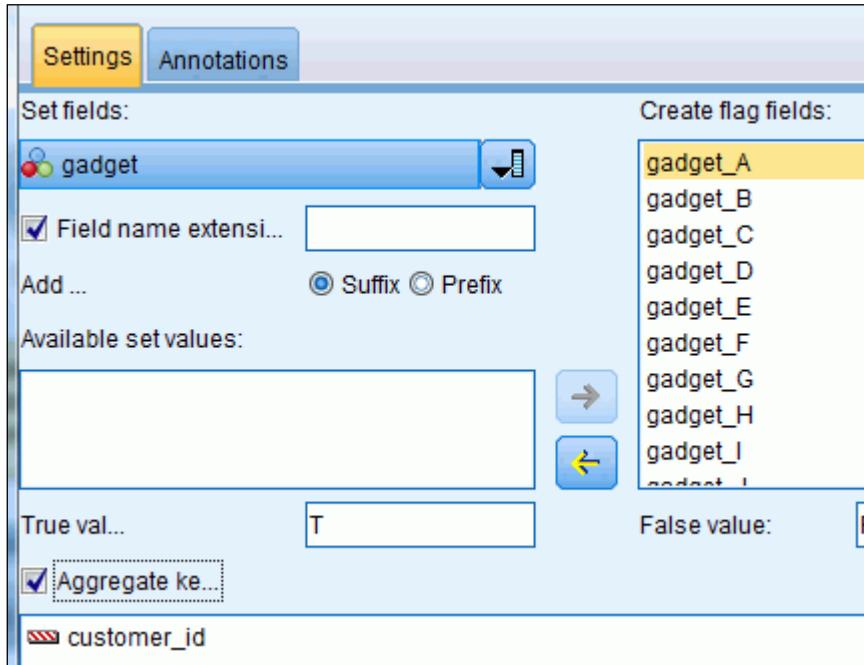
| customer_id | gadget | revenues | gadget_A | gadget_B | gadget_C | gadget_D | gadget_E | gadget_F |
|-------------|--------|----------|----------|----------|----------|----------|----------|----------|
| K100010     | C      | 28 F     | F        | T        | F        | F        | F        | F        |
| K100010     | E      | 52 F     | F        | F        | F        | T        | F        | F        |
| K100010     | F      | 61 F     | F        | F        | F        | F        | F        | T        |
| K100010     | K      | 109 F    | F        | F        | F        | F        | F        | F        |
| K100020     | A      | 11 T     | F        | F        | F        | F        | F        | F        |

Notice that only one value is true for a record.

7. Click **OK** to close the **Preview** output window.

8. Enable the **Aggregate keys** option, and then select **customer\_id** as field to aggregate on.

A section of the specifications in the SetToFlag dialog box appear as follows:



9. Click **Preview**.

A section of the results appear as follows:

| customer_id | gadget_A | gadget_B | gadget_C | gadget_D | gadget_E | gadget_F | gadget_G | gadget_H | gadget_I | gadget_J |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| K100010     | F        | F        | T        | F        | T        | T        | F        | T        | F        | F        |
| K100020     | T        | F        | F        | F        | F        | T        | T        | F        | T        | T        |
| K100030     | T        | T        | F        | T        | F        | F        | F        | F        | F        | F        |
| K100040     | T        | T        | F        | T        | F        | F        | F        | F        | F        | F        |

The flag fields have been created, and there is one record per customer.

10. Click **OK** to close the **Preview** output window.

This completes the demo for this module. You will find the solution results in the file **demo\_setting\_the\_unit\_of\_analysis\_completed.str**, in the **06-Setting the Unit of Analysis\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Which of the following is the correct statement?

- A. The Distinct node can be used to remove duplicate fields.
- B. The Distinct node can be used to remove duplicate records.
- C. A maximum of 8 key fields can be used to identify a duplicate record.
- D. The Distinct node is a terminal node.

Question 2: With respect to the Distinct node, which of the following statements are correct?

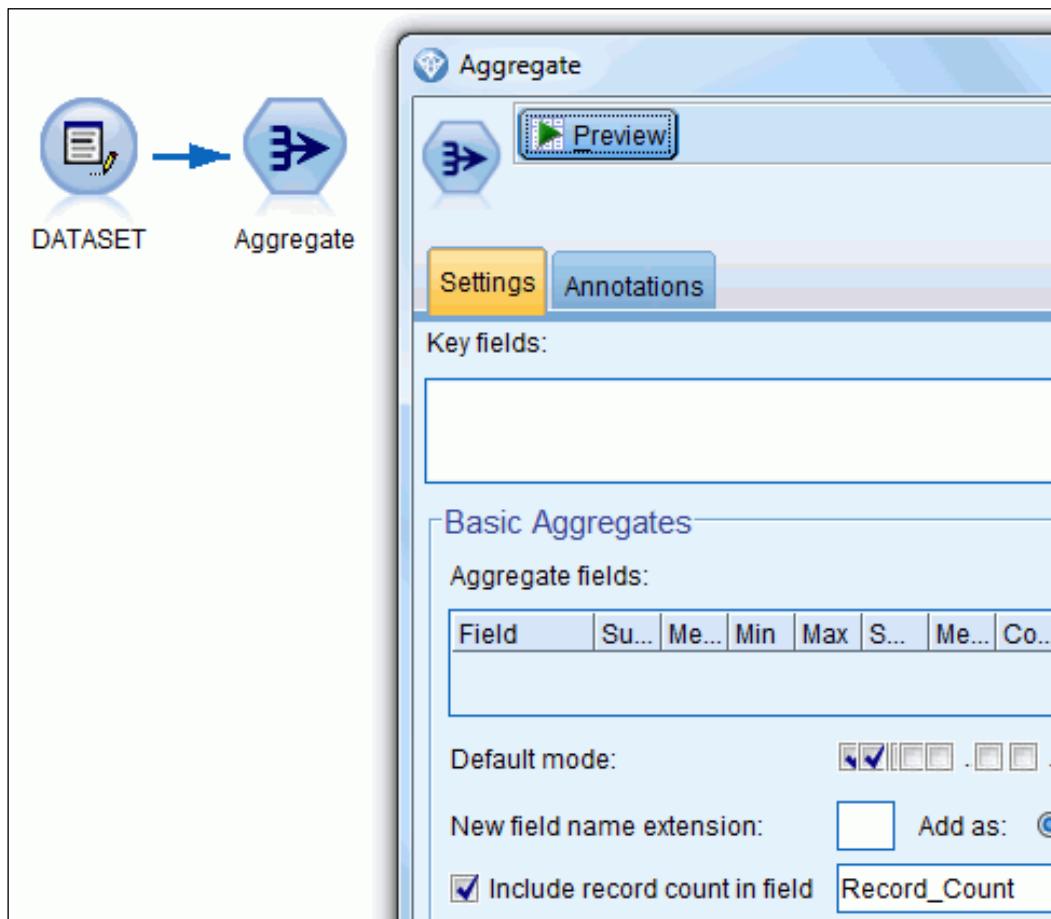
- A. The option Discard only the first record in each group retains the first distinct record as defined by the key field(s), and discards all duplicate records.
- B. A field of any measurement level can be used to define duplicate records.
- C. A field of any storage can be used to define duplicate records.
- D. The Distinct node is located in the Field Ops palette.

Question 3: In which of the following nodes can you use the median as aggregate statistic?

- A. SetToFlag
- B. Aggregate
- C. Distinct

Question 4: Refer to the figure that follows. Which of the following is the correct statement? Clicking the Preview button:

- A. returns the number of fields in the dataset
- B. returns the number of records in the dataset
- C. returns both the number of fields and the number of records in the dataset
- D. none of the above statements are valid



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Question 5: Given the dataset and the SetToFlag dialog box that are depicted below, which of the following is the correct statement? The reason that the area Available set values is not populated with the products is that:

- A. the product field is a continuous field
- B. the product field has only undefined (\$null\$) values
- C. the product field is not instantiated
- D. none of the above statements are correct

| CUSTOMER_ID | PRODUCT |
|-------------|---------|
|             | 1A      |
|             | 1A      |
|             | 1B      |
|             | 1A      |
|             | 1C      |
|             | 1D      |
|             | 1B      |
|             | 1C      |
|             | 1D      |
|             | 1E      |

**SetToFlag**

**Preview**

**Settings** **Annotations**

**Set fields:** PRODUCT

**Create flag fields:**

Field name extension

Add as:  Suffix  Prefix

**Available set values:**

→ ←

Question 6: Starting from the dataset depicted in the figure that follows, which node is used to arrive at the new dataset?

Note: The field names have been renamed in the new dataset.

- A. Distinct
- B. Aggregate
- C. SetToFlag
- D. none of the above statements are correct

### SOURCE DATASET

| <b>id</b> | <b>name</b>    | <b>gender</b> | <b>age</b> | <b>visit</b> | <b>code_medicin</b> | <b>costs_medicin</b> |
|-----------|----------------|---------------|------------|--------------|---------------------|----------------------|
| 1         | Rob Johnson    | m             | 45         | 07-Jul-2012  | A102                | 8                    |
| 1         | Robert Johnson | m             | 47         | 02-Jan-2014  | A107                | 22                   |
| 2         | Deb Peterson   | f             | 29         | 24-Jan-2012  | B481                | 10                   |
| 2         | Debby Peterson | f             | 29         | 18-Nov-2012  | B481                | 10                   |
| 2         | Debby Peterson | f             | 30         | 31-Mar-2013  | D141                | 5                    |

### NEW DATASET

| <b>id</b> | <b>name</b>  | <b>gender</b> | <b>age</b> | <b>most_recent_visit</b> | <b>most_frequent_medicin</b> | <b>total_costs_medicin</b> |
|-----------|--------------|---------------|------------|--------------------------|------------------------------|----------------------------|
| 1         | Rob Johnson  | m             | 47         | 02-Jan-2014              | A102                         | 30                         |
| 2         | Deb Peterson | f             | 30         | 31-Mar-2013              | B481                         | 25                         |

**Answers to questions:**

Answer 1: B. The Distinct node can be used to remove duplicate records.

Answer 2: B, C. A field of any measurement level and any storage can be used to define duplicate records.

Answer 3: B. You can choose the median in the Aggregate node.

Answer 4: B. An Aggregate node without a key field will aggregate over all records in the dataset, and when the option Include record count in field is enabled, this will return the number of records in the dataset.

Answer 5: C. The product is a categorical field, but apparently is not instantiated.

Answer 6: A. The Distinct node, with the option to create composite records.

Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - set the unit of analysis by removing duplicate records
  - set the unit of analysis by aggregating records
  - set the unit of analysis by expanding a categorical field into a series of flag fields

© 2014 IBM Corporation

Three methods to set the unit of analysis were presented in this module. Which method you will choose is determined by the requirements of your analysis. You can also use more than one method and combine the datasets later.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

6-28

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Workshop 1

## Setting the Unit of Analysis



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened the stream file **workshop\_setting\_the\_unit\_of\_analysis.str** in MODELER, located in the **06-Setting the Unit of Analysis\Start Files** sub folder

The following (synthetic) files are used in this workshop:

- **ACME customer data.xls**: a Microsoft Excel 2003 file with formation on customers
- **ACME purchases 1999 - 2004.dat**: a text file storing purchases made by customers, from 1999 to 2004

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

6-29

- **ACME orderlines 1999 - 2004.sav**: an IBM SPSS Statistics file with orders per purchase, from 1999 to 2004
- **ACME mailing history.xlsx**: a Microsoft Excel 2010 file, storing data on three mailings
- **workshop\_setting\_the\_unit\_analysis.str**: a MODELER stream file that opens the data files, and which the starting point for the workshop

## Workshop 1: Setting the Unit of Analysis

You are working in the database marketing department of ACME, a company that sells sport products. It is your job to import data from several sources, and create datasets with the required unit of analysis (one record per customer), so that these datasets can be merged later.

- For the data source storing customer data, the Microsoft Excel file **ACME customer data.xls**, remove duplicate records (records are duplicates if they are identical).

Note: After you have opened **workshop\_setting\_the\_unit\_of\_analysis.str** in the **06-Setting the Unit of Analysis\Start Files sub folder**, you will have the nodes that are needed for the data import in this and the next tasks on the stream canvas.

To check your results: The cleansed dataset is comprised of 30,000 records.

- From the data source storing purchases, **ACME purchases 1999 - 2004.dat** (each record represents a purchase made by a customer), create a dataset that has only one record per customer, with the most recent order date and the number of purchases made by the customer.
- From the data source storing orders, **ACME orderlines 1999 - 2004.sav** (each record represents one of the items bought at a certain moment of purchase), create a dataset that has only one record per purchase, with the total price per purchase (the sum of the PRICE field), and the number of items bought for that purchase.
- From the data source storing the mailing history, **ACME mailing history.xlsx** (each record represents a mailing sent to a customer), create a dataset with one record per customer, with fields indicating whether the customer was included in the first, second and/or third mailing.

For more information about where to work and the workshop results, refer to the Task and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

## Workshop 1: Tasks and Results

Task 1. Remove duplicate records in ACME's customer data.

- Add a **Distinct** node downstream from the **Excel** source node named **ACME customer data.xls**
- Edit the **Distinct** node, and then:
  - for **Mode**, select **Include only the first record in each group**
  - for **Key fields for grouping**, select all fields (if values in all fields are the same, records are identical (and duplicate))
- Add a **Table** node downstream from the **Distinct** node, and then run the **Table** node.

This will show that you have 30,000 records in the dataset.

Task 2. Create a dataset where customers are unique in ACME's purchases data.

- Add an **Aggregate** node downstream from the **Var. File** source node.
- Edit the **Aggregate** node, and then:
  - for **Key fields**, select **CUSTOMER\_ID**
  - for **Aggregate fields**, select **ORDERDATE**, and select **Max**
  - enable the **Include record count in field** option; then type a name such as **number\_of\_purchases**
- **Preview** the data.

A section of the results appear as follows:

| CUSTOMER_ID | ORDERDATE_Max | number_of_purchases |
|-------------|---------------|---------------------|
| 724         | 1999-07-11    | 1                   |
| 727         | 1999-02-21    | 1                   |
| 728         | 2004-05-22    | 1                   |
| 730         | 2002-11-29    | 1                   |
| 731         | 2003-07-17    | 3                   |

Note: Instead of using Aggregate, you can use the Distinct node, with the option to create a composite record.

### Task 3. Create a dataset where customers are unique in ACME's order lines data.

- Add an **Aggregate** node downstream from the **Statistics File** source node.
- Edit the **Aggregate** node, and then:
  - for **Key fields**, select **PURCHASE\_ID**
  - for **Aggregate fields**, select **PRICE**, and select **Sum**
  - enable the **Include record count in field** option; and type a name such as **number\_of\_items\_purchased**
- **Preview** the data.

A section of the results appear as follows:

| PURCHASE_ID | PRICE_Sum | items_bought |
|-------------|-----------|--------------|
| 5723        | 948.480   | 4            |
| 5724        | 114.840   | 1            |
| 5726        | 217.770   | 1            |
| 5727        | 466.130   | 3            |
| 5729        | 985.520   | 3            |

Note: Instead of using Aggregate, you can use the Distinct node, with the option to create a composite record.

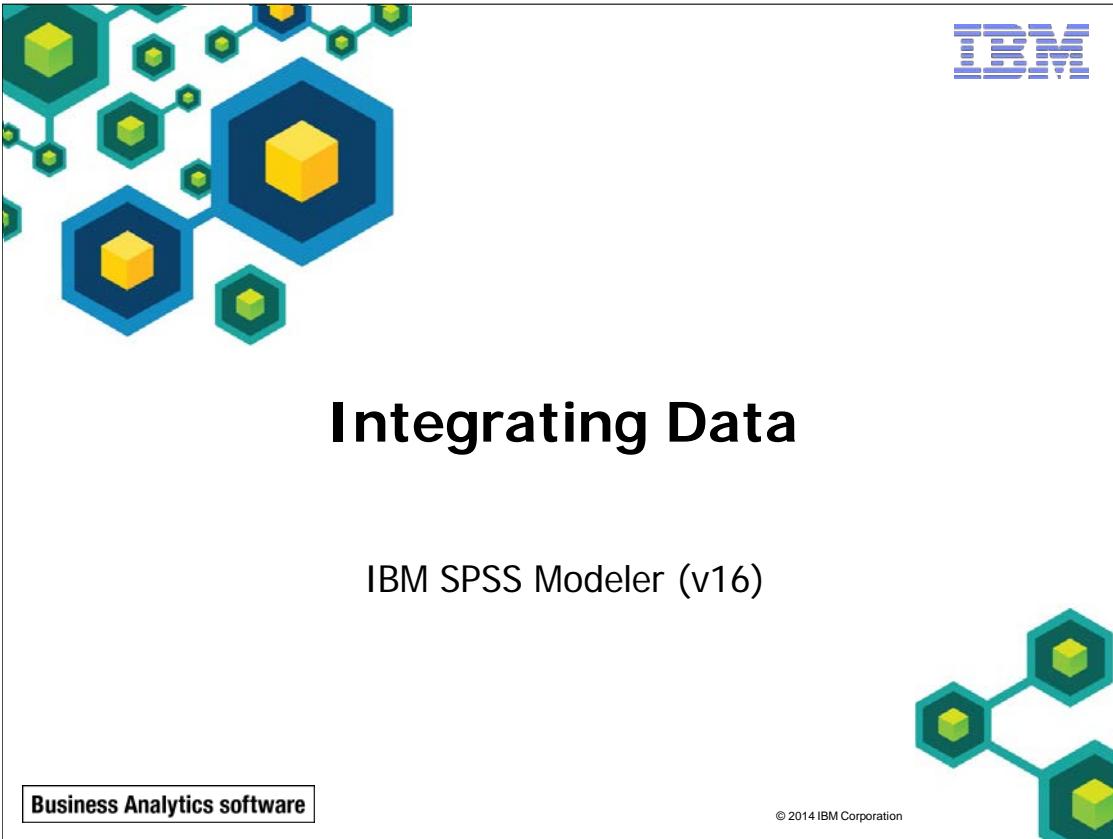
## Task 4. Create a dataset where customers are unique in ACME's mailing history data.

- Add a **Type** node downstream from the **Excel** source node named **ACME mailing history.xlsx**.
- Edit the **Type** node, and then click **Read Values** to instantiate the data (the values for the mailing field will then be available in the SetToFlag node that will be added downstream).
- Add a **SetToFlag** node downstream from the **Type** node.
- Edit the **SetToFlag** node and then:
  - for Set fields, select mailing
  - select all values under **Available set values**, and move them to the **Create flag fields** area
  - enable the option **Aggregate keys**, and select **customer\_id**
- **Preview** the data.

A section of the results appear as follows:

| customer_id | mailing_Standard tennis... | mailing_XL original orange ... | mailing_XS original red soccer ... |
|-------------|----------------------------|--------------------------------|------------------------------------|
| 723.000 F   | T                          | T                              |                                    |
| 724.000 F   | T                          | F                              |                                    |
| 725.000 F   | T                          | F                              |                                    |

Note: The stream **workshop\_setting\_the\_unit\_of\_analysis\_completed.str** in the **06-Setting\_the\_Unit\_of\_Analysis\Solution Files** sub folder provides a solution to the workshop tasks.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Objectives

- At the end of this module, you should be able to:
  - integrate data by appending records from multiple datasets
  - integrate data by merging fields from multiple datasets
  - sample records

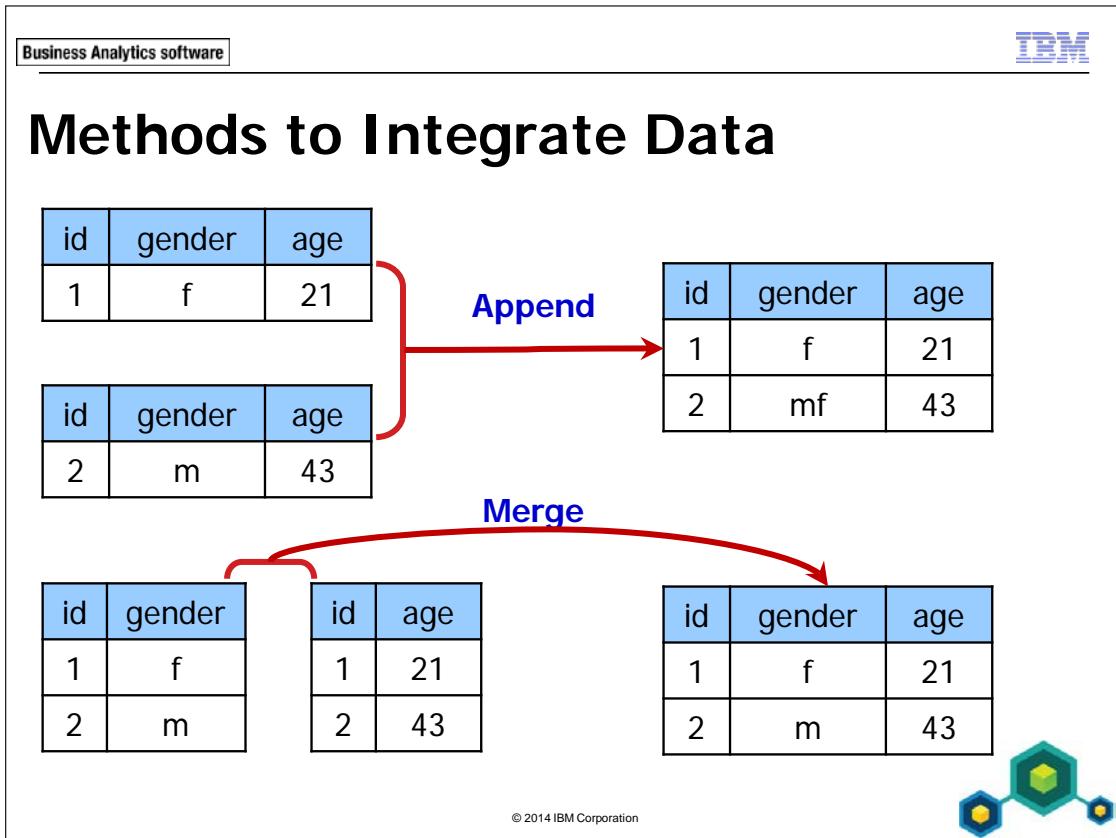
© 2014 IBM Corporation

Similar pieces of information may be stored in different datasets. These datasets must be combined into a single dataset for analyses. Typically, datasets are combined after the unit of analysis is set correctly for each of the datasets involved.

Combining datasets is referred to in the CRISP-DM process model as integrating data.

Before reviewing this module you should be familiar with:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data
- methods to explore the data
- methods to set the unit of analysis



There are two methods of combining datasets:

- Append: Add records from one dataset to another
- Merge: Add fields from one dataset to another

This module presents these two methods of combining datasets.

## Appending Records

- Similar pieces of information for different groups of records stored in different datasets
- For analysis and modeling, create a single dataset
- Use the Append node (Record Ops)



© 2014 IBM Corporation

Similar pieces of information for different groups of records may be stored in different datasets. Examples of this include:

- fraud information for different local offices
- bank account information for different financial years
- examination results for different academic years
- transaction data for different weeks

There is often a need to collectively analyze such data, possibly to compare performance over subsequent years or to discover group differences. To analyze such information the datasets must be combined into a single dataset.

Use the Append node (located in the Record Ops palette) to append datasets.

Business Analytics software

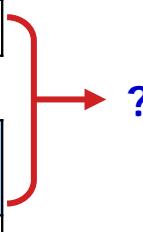
IBM

## Options to Append Records

| id | gender | age |
|----|--------|-----|
| 1  | f      | 21  |

| id | gender | married |
|----|--------|---------|
| 2  | m      | no      |



- Main dataset: one dataset is leading
- All datasets: equal role of the datasets

© 2014 IBM Corporation

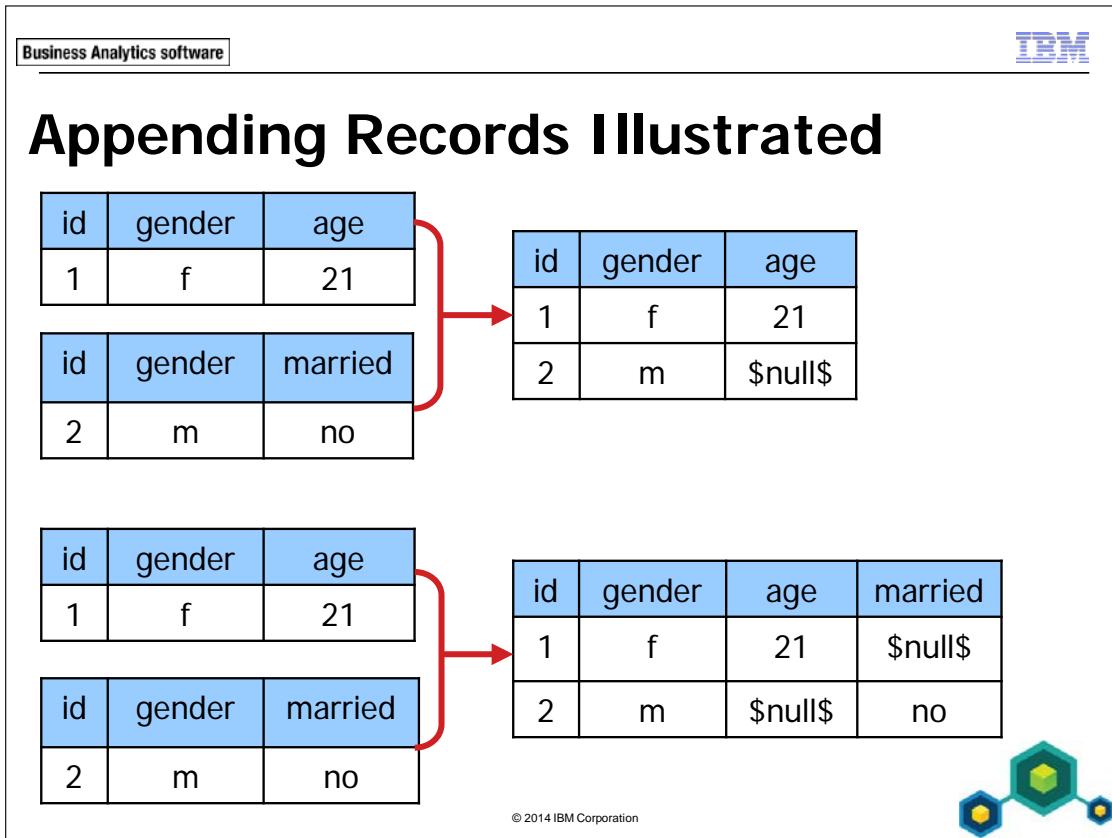


In appending records from one dataset to another, various options are available for the situation where not all fields are the same.

Two choices can be made when appending records from datasets: to treat one of the datasets as the main dataset or to let the datasets play an equal role.

Using one of the datasets as the main dataset means that only the fields present in the main dataset will be output to the new dataset. Fields only present in the secondary dataset are lost.

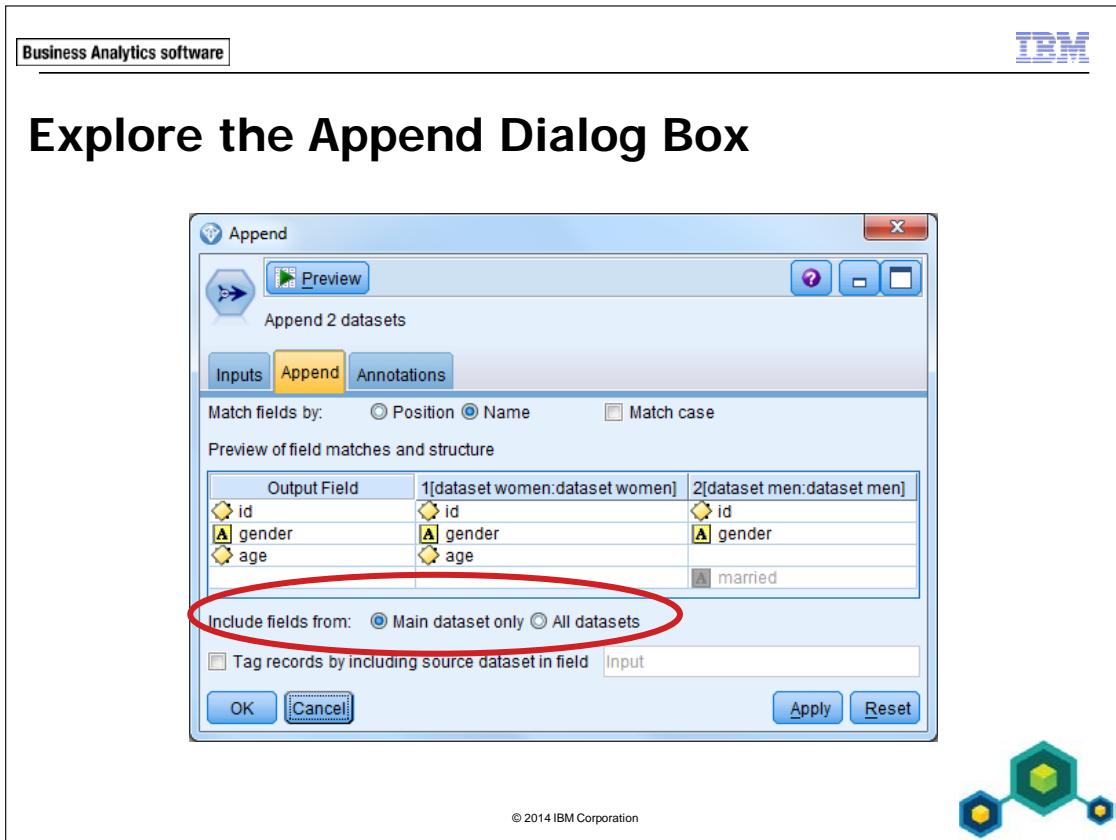
When the datasets play an equal role, all fields from all datasets will be output to the new dataset.



The first figure on this slide gives an example, with the upper dataset as the main dataset. Only the fields id, gender, and age are output and the married field is lost.

The second figure on this slide gives an example of giving the datasets an equal role in the append operation.

When you append records from different datasets, some records may have undefined values, regardless of the method used. For example, in the first example id 2 will have an undefined value for age, because age was not observed for this person. Likewise id 1 has an undefined value for married and id 2 has an undefined value for age in the second example.



In the Append dialog box, the Append tab controls how the datasets are appended. For Include fields from, select whether you want to use a leading dataset (the Main dataset only option) or that you want all datasets to play an equal role (the All datasets option).

It is recommended to tag records by adding a field to the combined dataset whose values indicate the source dataset for each record.

Fields can be matched by position or name (the default). In general, matching fields by position is not recommended. When matching on name, there is an extra option to enable case sensitivity when matching names. For example a field named id will match with field name ID, but when this option is enabled they will not match.

On the Inputs tab in the Append dialog you can specify the main dataset (if this option is selected), or the order wherein the datasets are appended (when all datasets play an equal role).

## Merging Fields

- Different pieces of information for the same records are stored in different datasets
- Create a single dataset for analyses
- Use the Merge node (Record Ops)



© 2014 IBM Corporation



In many organizations, different pieces of information for individuals are held in separate locations. Examples of this include:

- customer information that is held separately from purchase information
- account details that is held in a database separate from transactions
- a housing organization may hold information at an individual and property level

To analyze such data, for example to find patterns in purchase behavior and link them back to demographics, the individual datasets have to be combined into one single dataset. The way to join these datasets is not by adding records from one dataset to another, but by adding fields.

You can use the Merge node, located in the Record Ops palette) to combine fields from different datasets into a single dataset.

Business Analytics software

IBM

## Options to Merge Fields

|    |        |
|----|--------|
| id | gender |
| 1  | f      |
| 2  | m      |

+

|    |     |
|----|-----|
| id | age |
| 2  | 25  |
| 3  | 43  |

= ?

- Inner join
- Full outer join
- Partial outer join
- Anti-join

© 2014 IBM Corporation



When datasets do not have the same records, three choices can be made for the merge:

1. Only matching records are passed through the Merge node, known as an inner join.
2. Records from all datasets are passed, known as an outer join.
3. One dataset is leading in the merge and enriched with fields from other datasets, but no records are added. This is known as a partial outer join.

There is also an option to retain only the records that are unique to a specific dataset. This is known as an anti-join.

## Merging Fields Illustrated (1 of 2)

| id | gender |
|----|--------|
| 1  | f      |
| 2  | m      |

+

| id | age |
|----|-----|
| 2  | 25  |
| 3  | 43  |

=

| id | gender | age |
|----|--------|-----|
| 2  | m      | 25  |

| id | gender |
|----|--------|
| 1  | f      |
| 2  | m      |

+

| id | age |
|----|-----|
| 2  | 25  |
| 3  | 43  |

=

| id | gender   | age      |
|----|----------|----------|
| 1  | f        | \$null\$ |
| 2  | m        | 25       |
| 3  | \$null\$ | 43       |

© 2014 IBM Corporation



This slide illustrates an inner join (the upper figure in each of the cases) and an outer join (the lower figure in each of the cases).

In an inner join only data from records present in all source datasets will be merged. In this example only id 2 will be output.

In an outer join every record from each dataset is passed through the Merge node. The undefined (\$null\$) value is added to the missing fields. In this example id 1 will have the undefined value for the age field, and id 3 will have the undefined value for gender.

Business Analytics software

IBM

## Merging Fields Illustrated (2 of 2)

The diagram illustrates two cases of merging fields:

**Case 1 (Partial Outer Join):**

- Left Dataset: id | gender  
1 | f  
2 | m
- Right Dataset: id | age  
2 | 25  
3 | 43
- Output: id | gender | age  
1 | f | \$null\$  
2 | m | 25

**Case 2 (Anti-Join):**

- Left Dataset: id | gender  
1 | f  
2 | m
- Right Dataset: id | age  
2 | 25  
3 | 43
- Output: id | gender  
1 | f

© 2014 IBM Corporation

This slide illustrates a partial outer join (the upper figure in each of the cases) and an anti-join (the lower figure in each of the cases).

In a partial outer join one of the datasets is leading in the merge, in the sense that all records from the leading dataset will be in the output dataset and only information from matching records in the secondary datasets will be added. In this example, taking the left dataset as the leading dataset, id's 1 and 2 will be output, with id 1 having the undefined (\$null\$) values for age.

An anti-join takes one dataset as leading and retains only records from that dataset that are not present in the secondary datasets. No fields will be added. In this example, with the left dataset as the leading dataset, this will only keep id 1.

Business Analytics software

IBM

## A One-to-Many Merge

| id | age | zip code |
|----|-----|----------|
| 1  | 43  | 105      |
| 2  | 22  | 105      |
| 3  | 51  | 481      |
| 4  | 63  | 481      |

+

| zip code | % with car |
|----------|------------|
| 105      | 25         |
| 481      | 43         |
| 585      | 12         |

=

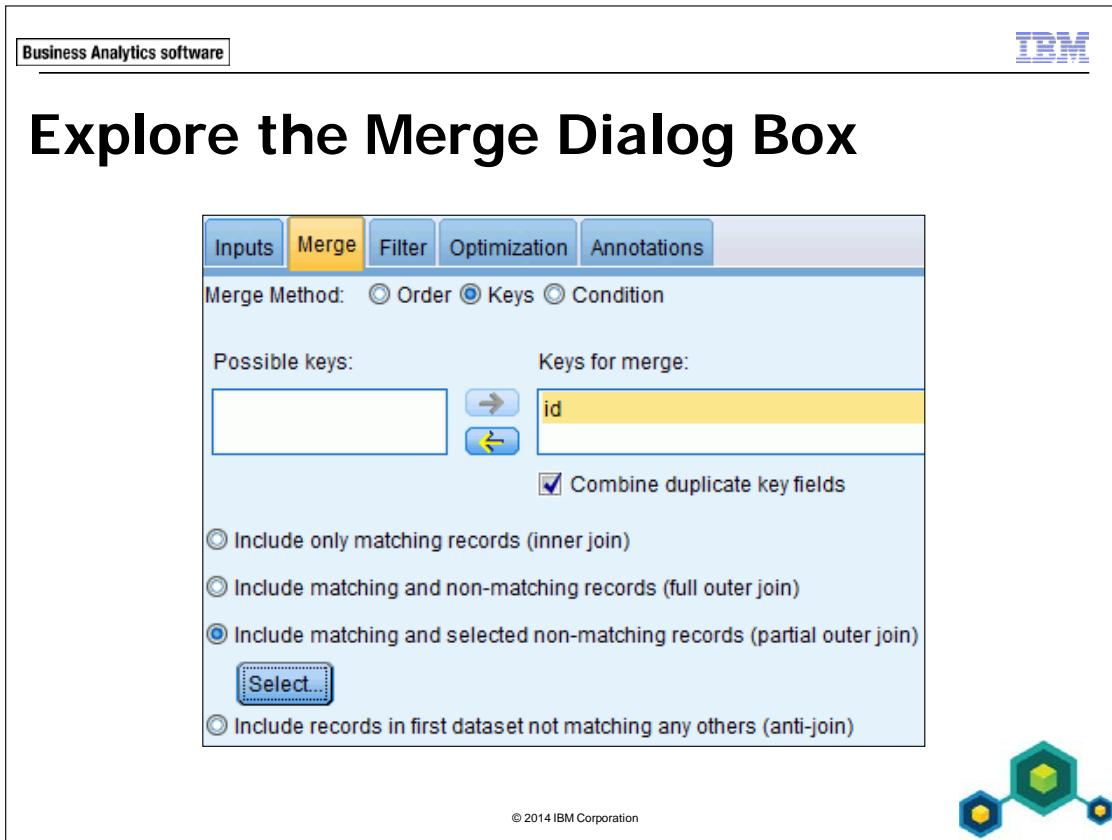
| id | age | zip code | % with car |
|----|-----|----------|------------|
| 1  | 43  | 105      | 25         |
| 2  | 22  | 105      | 25         |
| 3  | 51  | 481      | 43         |
| 4  | 63  | 481      | 43         |

© 2014 IBM Corporation



In the previous examples the situation was one of a 1-to-1 match: one record in one dataset matches one (or no) record in the other dataset.

The Merge node accommodates as well a 1-to-many match. This slide shows an example, which is frequently encountered in database marketing. A customer dataset can be enriched with aggregated data by using the zipcode as key field for the merge. Typically, a partial outer join will be used in this situation, with the customer dataset leading.



In the Merge dialog box, the Merge tab controls how datasets are merged. You can merge records by order (not recommended), by using one or more key fields (the most common situation), or by specifying a condition.

Under Possible keys, fields contained in all input datasets are listed. Field names in MODELER are case sensitive, and field names have to match in case otherwise they will not appear. You can use a Filter node upstream from the Merge node to rename fields, or you can match by condition to solve this issue.

Enabling the option Combine duplicate key fields ensures that you will have only output key field with a given name. When this option is disabled, duplicate key fields must be renamed or excluded using the Filter tab in the Merge dialog box.

Select the type of the merge that you want. For a partial outer join, click Select to choose the leading dataset.

The Input tab determines the order in which the records are read. Also, it sets the main dataset for an anti-join.

## Sampling Records

- Processing big amounts of data can be inefficient in terms of time and memory
- Sample the records and if the stream does the job, use all records
- Use the Sample node (Record Ops)

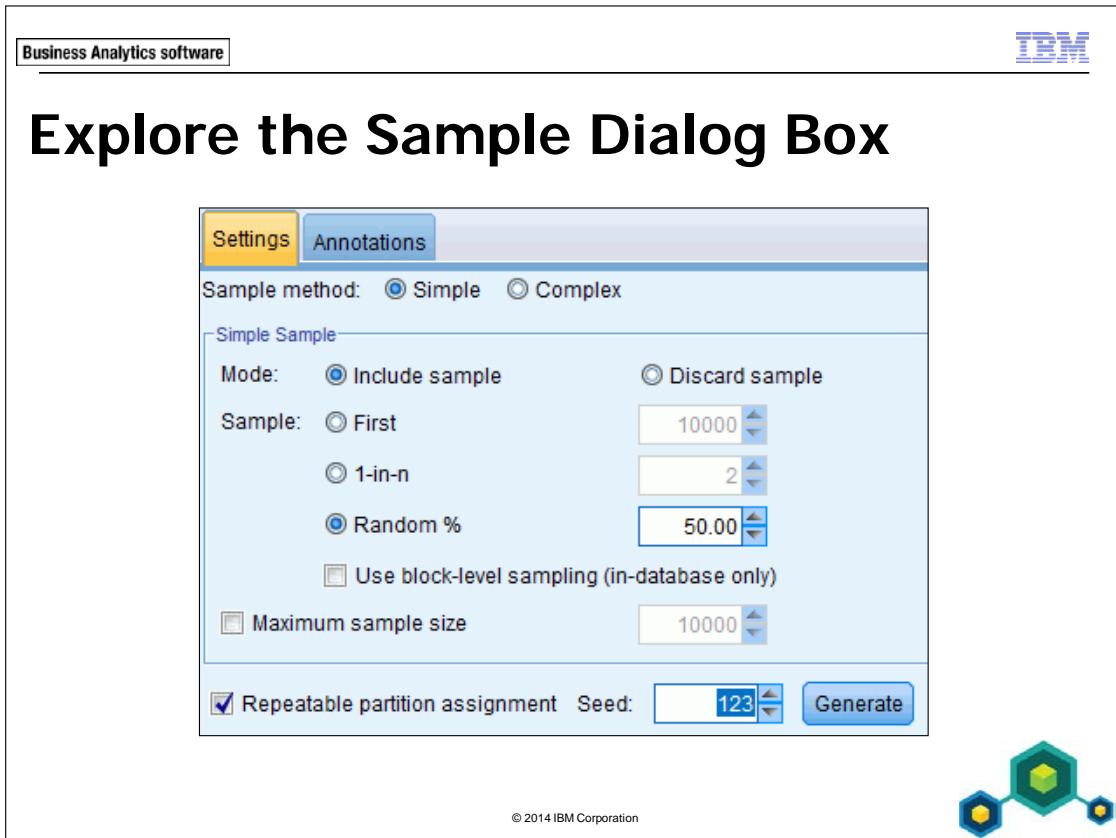


© 2014 IBM Corporation

It is not uncommon to have hundreds of thousands, if not millions, of records available in one or more datasets. When appending or merging this size of datasets, data processing can take a huge amount of time.

In general, using all the records can be quite inefficient in terms of processing time and memory. To build and test the stream it is recommended to first sample the records and if the stream does the job as required, have a final test with all records.

To sample records, use the Sample node (located in the Record Ops palette).



MODELER offers two sampling methods: simple and complex. The simple sampling method is presented in this module. Refer to the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for a presentation of complex sampling.

You can either select or deselect records (the Include sample option and Discard sample option, respectively).

You have three options for sampling: select the first  $n$  records in the dataset (where  $n$  needs to be specified), select every  $n^{\text{th}}$  record or draw a random sample of a certain percentage. The latter option will draw a different sample each time that records pass through the Sample node. If you want to replicate a sample, type or generate a seed value for the algorithm, so that the same records will be sampled.

## Caching Data

- MODELER processes data starting at the source nodes
- With huge amounts of data, cache the data
  - Right-click the node, then select Cache\Enable
- Cached nodes:

-  empty cache
-  full cache

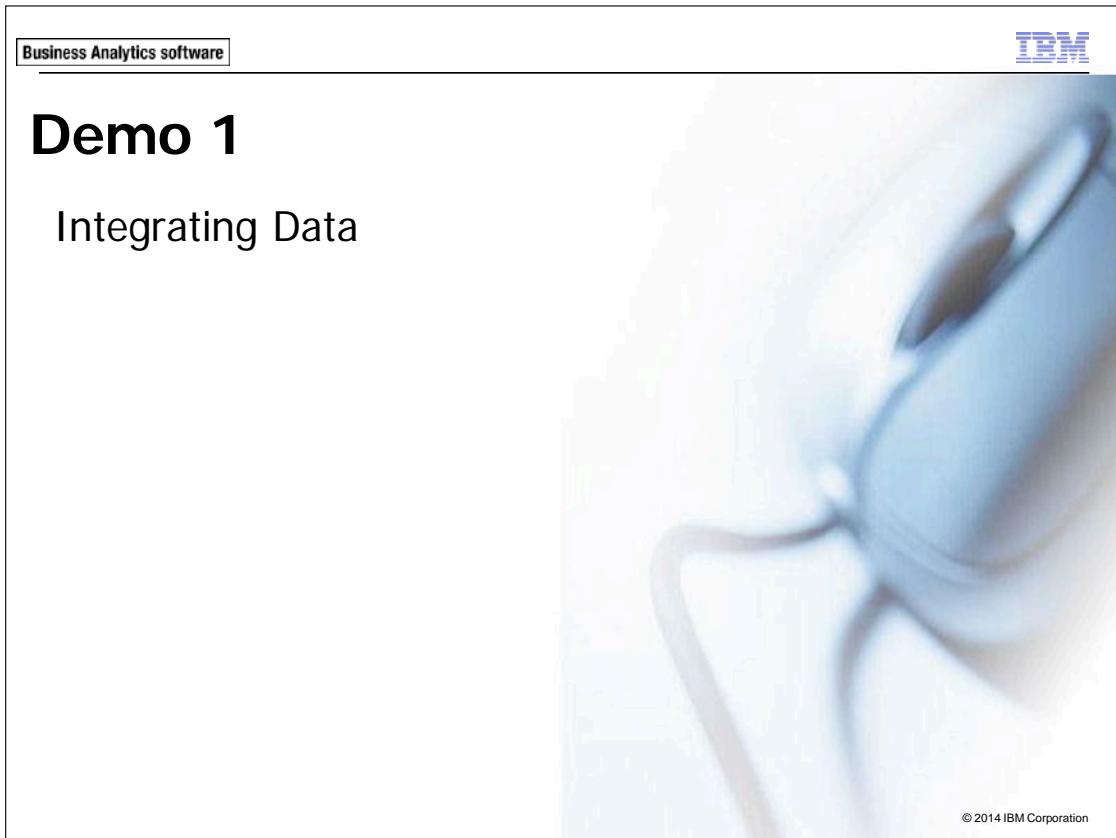
© 2014 IBM Corporation



When you sample records, the Sample node pulls data from the data source. This process repeats every time that the stream is run. This means that each time data must be read and records must be sampled (and in case of a fixed seed value, the same records will be sampled). This is highly inefficient in the case of huge amounts of data. Therefore, MODELER provides the option to cache the data. Caching creates a temporary file behind the scenes, storing the data.

When a node carries a cache (enabled by selecting Cache\Enable from the context menu on a node), a document icon is displayed at the top right corner. At first, the icon will be empty, but when data flow through the cached node, the cache fills and the icon turns green. If changes are made upstream from a cached node, the cache will become emptied, as its contents will no longer be valid. It will refill when data are passed through the node.

Data can be cached at any non-terminal node. A data cache is also useful at time-consuming nodes such as a Sort, Append, Merge, or Aggregate. Caching on or after these nodes can substantially improve performance.



The slide features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The main title "Demo 1" is centered at the top, followed by the subtitle "Integrating Data". The background is a blurred image of a person's face.

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **demo\_integrating\_data.str**, located in the **07-Integrating\_Data\Start Files** sub folder.

The following (synthetic) files from a (fictitious) telecommunications firm are used to demonstrate how you can integrate data:

- **telco x call data q1.sav**: three months (first quarter) of call detail records for each customer)
- **telco x call data q2.sav**: three months (second quarter) of call detail records for each customer

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- **telco x customer info.xlsx** : a Microsoft Excel data file, storing demographic and churn data on the company's customers
- **telco x products.dat**: a tab-delimited text data file, storing data on gadgets that customers purchased (a transactional dataset - a customer has as many records as he has gadgets)
- **telco x tariff.dat**: a text file - information on tariffs for this telecommunications company
- **demo\_integrating\_data.str** – a stream file that imports the datasets, sets the required unit of analysis for some of the datasets, and serves as the starting point for the demo

## Demo 1: Integrating Data

### Purpose:

You work for a telecommunications company and you have to combine a number of datasets into a single dataset as a preparation for analyses and modeling later.

### Task 1. Append records from two datasets.

You will combine two IBM SPSS Statistics .sav files that store call detail records into a single dataset.

1. Place an **Append** node on the stream canvas to the **right** of the two **Statistics File** sources nodes named **telco x call data q1.sav** and **telco x call data q2.sav**.  
Note: These nodes are on the stream canvas after you have opened **demo\_integrating\_data.str**, located in the **07-Integrating\_Data\Start Files** sub folder.
2. Connect the **Var. File** node named **telco x call data q1.sav** to the **Append** node.
3. Connect the **Var. File** named **telco x call data q2.sav** to the **Append** node.

4. Edit the **Append** node.

A section of the specifications in the Append dialog box appear as follows:

The screenshot shows the 'Append' tab selected in the dialog box. It includes settings for matching fields by position or name, and an option to match case. A preview table shows the mapping of fields from two datasets. Fields matching by name are shown in the first column, while unique fields from the second dataset are listed in the second and third columns.

| Output Field          | 1[telco x call data q1.sav...] | 2[telco x call data q2.sav...] |
|-----------------------|--------------------------------|--------------------------------|
| customer_id           | customer_id                    | customer_id                    |
| # peak_calls          | # peak_calls                   | # peak_calls                   |
| # peak_mins           | # peak_mins                    | # peak_mins                    |
| # offpeak_calls       | # offpeak_calls                | # offpeak_calls                |
| # offpeak_mins        | # offpeak_mins                 | # offpeak_mins                 |
| # weekend_calls       | # weekend_calls                | # weekend_calls                |
| # weekend_mins        | # weekend_mins                 | # weekend_mins                 |
| # international_mins  | # international_mins           |                                |
| # International_calls | # International_calls          |                                |
| month                 | month                          | month                          |
|                       |                                | # internat_mins                |
|                       |                                | # Internat_calls               |

Below the table, there are options to include fields from the main dataset only or all datasets, and a checkbox for tagging records with source dataset information.

Fields are matched by name by default. Fields that will be included in the new dataset are listed in the Output Field column. In this example, the field names match except those that are related to international phone calling. Fields unique to the second dataset (telco x call data q2.sav) will be lost and records coming from the second dataset will be assigned the undefined (\$null\$) value for international\_mins and international\_calls.

You can force the second dataset to act as the main dataset by reordering the datasets in the Inputs tab. In this case, that is no solution to the problem, because data will still be lost (just for the other dataset). Neither would it help if you let both datasets play an equal role: the information about international calling would be in different fields. This can be handled downstream by further data preparation, but in this demo the simpler solution is preferred that makes use of the fact that international\_mins and international\_calls in the first dataset are in the same position as internat\_mins and internat\_calls are in the second dataset. You will match the datasets by position and to check the results, you will tag the records.

5. On the **Append** tab:

- for **Match fields by**, select **Position**
- enable the **Tag records by including source dataset in field** option
- rename the Input field to **quarter**

6. Close the **Append** dialog box.

When the Append node is processed, the first block of data that is output comes from the first dataset, followed by a block of data from the second dataset. To have a better view on the data, you will sort the data on **customer\_id** and **month**.

7. Add a **Sort** node downstream from the **Append** node.

8. Edit the **Sort** node, and then:

- for **Sort keys**, select **customer\_id** and **month**
- click **Preview**

A section of the results appear as follows:

| customer_id | peak_calls | peak_mins | ... | ... | ... | ... | ... | ... | month  | quarter |
|-------------|------------|-----------|-----|-----|-----|-----|-----|-----|--------|---------|
| K100010     | 2.000      | 6.086     | ... | ... | ... | ... | ... | ... | month1 | 1       |
| K100010     | 2.000      | 6.060     | ... | ... | ... | ... | ... | ... | month2 | 1       |
| K100010     | 2.000      | 5.494     | ... | ... | ... | ... | ... | ... | month3 | 1       |
| K100010     | 2.000      | 4.605     | ... | ... | ... | ... | ... | ... | month4 | 2       |
| K100010     | 2.000      | 5.561     | ... | ... | ... | ... | ... | ... | month5 | 2       |
| K100010     | 4.000      | 8.325     | ... | ... | ... | ... | ... | ... | month6 | 2       |
| K100020     | 7.000      | 5.538     | ... | ... | ... | ... | ... | ... | month1 | 1       |
| K100020     | 9.000      | 6.875     | ... | ... | ... | ... | ... | ... | month2 | 1       |
| K100020     | 8.000      | 6.172     | ... | ... | ... | ... | ... | ... | month3 | 1       |
| K100020     | 9.000      | 7.034     | ... | ... | ... | ... | ... | ... | month4 | 2       |

The Preview output window shows that the first customer has 6 records of data, one record for each month. The quarter field indicates the data source.

9. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Sort** dialog box.

Appending the two datasets has been completed successfully. For the analyses later in the project, however, a data structure is required with one record per customer. You can use the Distinct node or the Aggregate node to arrive at that data structure. Refer to the *Setting the Unit of Analysis* module in this course for details on each of these operations. Which method is preferred depends on the objective for the analysis. Here, you will aggregate the data to customer-level by summing the values.

10. Add an **Aggregate** node downstream from the **Sort** node.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

11. Edit the **Aggregate** node, and then:

- for **Key fields**, select **customer\_id**
- for **Aggregate fields**, select **peak\_calls** to **international\_calls**
- disable the **Include record count in field** option
- click **Preview**

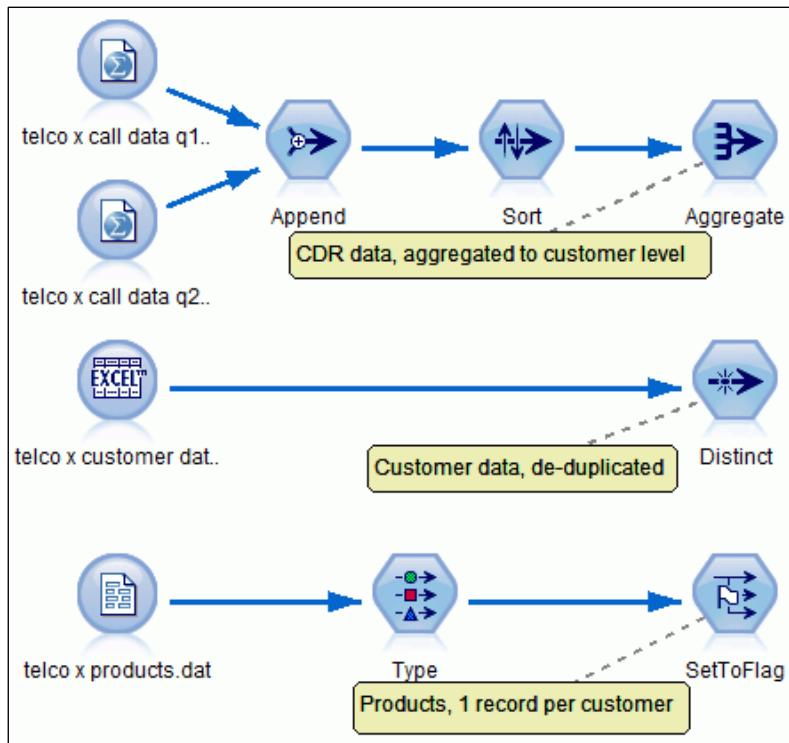
A section of the results appear as follows:

| customer_id | peak_calls_Sum | peak_mins_Sum | offpeak_calls_Sum | offpeak_mins_Sum |
|-------------|----------------|---------------|-------------------|------------------|
| K100010     | 14.000         | 36.131        | 10.000            | 7.973            |
| K100020     | 54.000         | 39.437        | 34.000            | 21.153           |
| K100030     | 44.000         | 72.600        | 1.000             | 27.600           |
| K100040     | 44.000         | 72.600        | 1.000             | 27.600           |
| K100050     | 32.000         | 40.608        | 14.000            | 18.824           |
| K100060     | 56.000         | 46.260        | 6.000             | 11.085           |

This dataset has the required unit of analysis.

12. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Aggregate** dialog box.
13. Add the following text as a comment to the **Aggregate** node: **CDR data, aggregated to customer level**.

A section of the results appear as follows:



Leave the stream open for the next task.

## Task 2. Merge fields from three datasets.

Different datasets have to be merged: customer data (with duplicate records removed by using a Distinct node), call detail records (aggregated to a 6 month period in the previous task) and product data (transformed into a dataset with one record per customer using a SetToFlag node). These datasets all have one record per customer, and thus can be merged on the key field `customer_id`.

In this demo, the customer information dataset is taken as the leading dataset in the merge, so you will use a partial outer join.

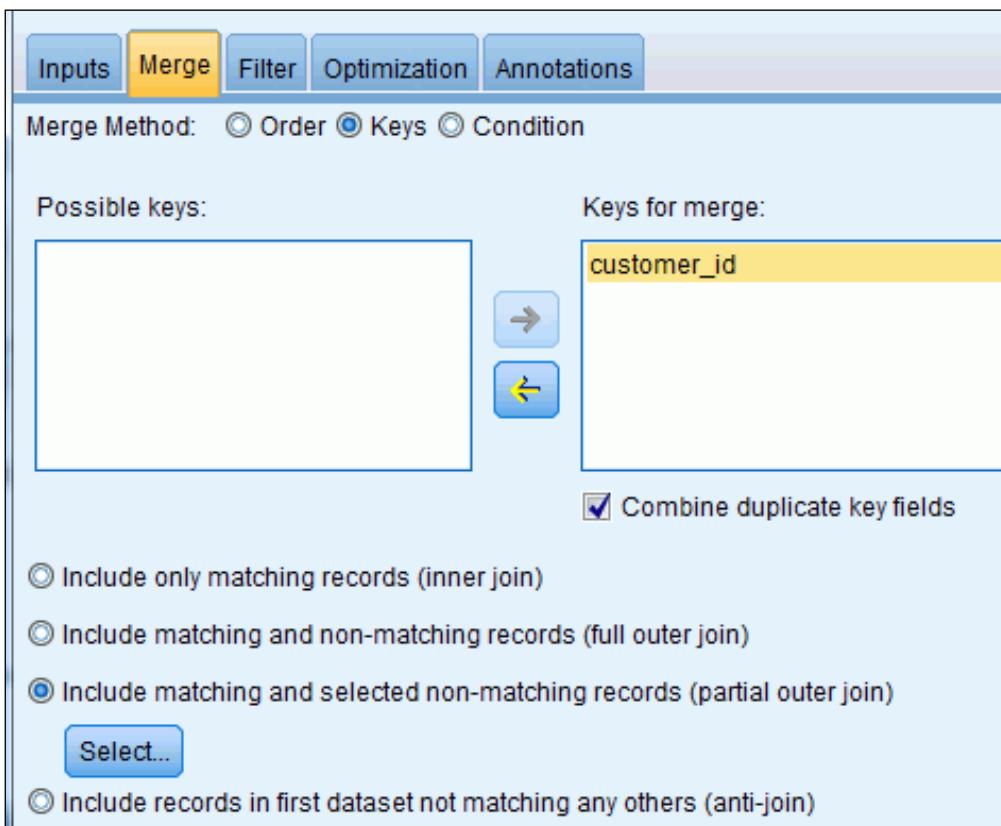
In this task you will build from the previous stream.

1. Place a **Merge** node (Record Ops palette) on the stream canvas, to the right of the **Distinct** node (do not connect the nodes – we want to explicitly connect the nodes to be in control of the order of connections).
2. Connect the **Distinct** node to the **Merge** node.
3. Connect the **Aggregate** node to the **Merge** node.
4. Connect the **SetToFlag** node to the **Merge** node.

5. Edit the **Merge** node, and then:

- for **Merge Method**, select **Keys**
- for **Keys for merge**, select **customer\_id**
- enable the **Include matching and selected non-matching records (partial outer join)** option

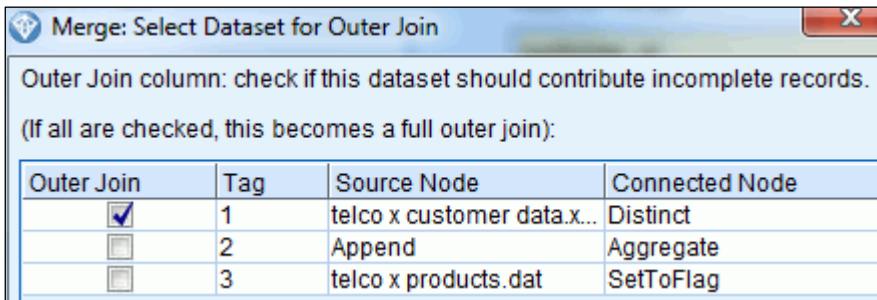
A section of the specifications in the Merge dialog box appear as follows:



The customer dataset must be the leading dataset in the merge, so you will mark this dataset as main dataset.

6. Click **Select**, and then ensure that the customer dataset (**telco x customer data.xlsx**) is the leading dataset.

A section of the specifications in the Merge sub dialog box appears as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

7. Click **OK** to close the **Merge: Select Dataset for Outer Join** sub dialog box.
8. Click **Preview**.

A section of the results appear as follows:

| id           | gadget_A | gadget_B | gadget_C | gadget_D | gadget_E | gadget_F | gadget_G | gadget_H |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|
| 000 F        | F        | T        | F        | T        | T        | F        | F        |          |
| 000 T        | F        | F        | F        | F        | T        | T        | F        |          |
| 000 T        | T        | F        | T        | F        | F        | F        | T        |          |
| 000 T        | T        | F        | T        | F        | F        | F        | F        |          |
| 000 \$null\$ |          |
| 000 \$null\$ |          |
| 000 F        | F        | T        | F        | F        | T        | F        | F        |          |
| 000 F        | F        | F        | F        | F        | F        | F        | F        |          |
| 000 F        | F        | F        | F        | F        | F        | F        | T        |          |
| 000 T        | T        | F        | T        | F        | F        | F        | F        |          |

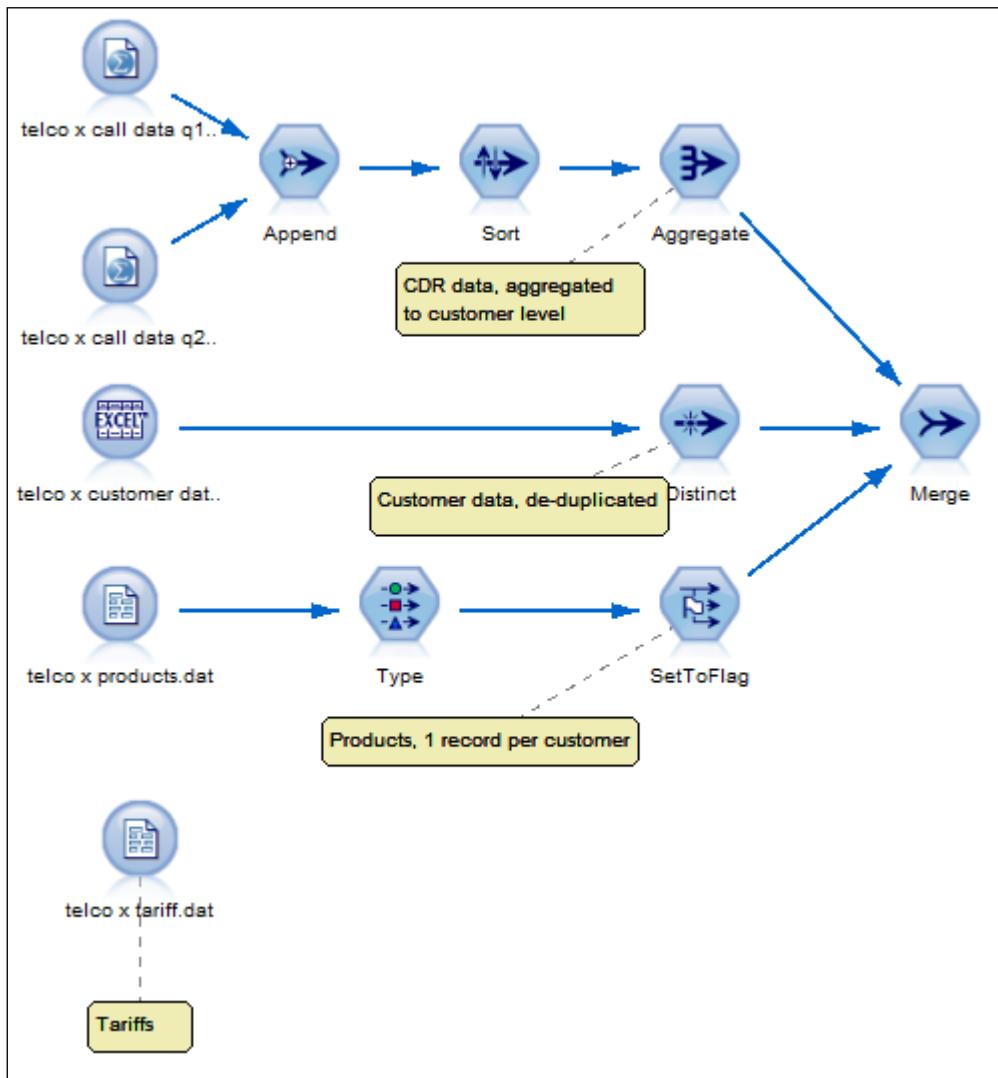
Three datasets are combined into one. The order of the fields is determined by the order of the datasets in the Inputs tab in the Merge node.

The key field is the first field in the combined dataset. This will always be the case.

Notice that some customers have undefined (\$null\$) values for all gadget fields, for example, all gadgets are \$null\$ for record with ID K100050. This customer was not in the product dataset, so the undefined (\$null\$) value was assigned for the fields related to gadgets.

9. Click **OK** to close the **Preview** output window, and click **OK** to close the **Merge** dialog box.

A section of the results appear as follows:



Leave the stream open for the next task.

### Task 3. Enrich a dataset with aggregated data.

Apart from the three datasets merged in the previous task, there is additional information on tariffs. This dataset will be added to the single dataset that was created in the previous task, but it should be noted that the tariff dataset does not store customer data, but stores data on the aggregated level of tariffs. This will require a separate merge.

In this task you will build from the previous stream.

1. Place a second **Merge** node on the stream canvas, to the right of the data source **telco x tariff.dat** (but do not connect it)
2. Connect the **first Merge** node to the **second Merge** node
3. Connect the **Var. File** node named **telco x tariff.dat** to the **second Merge** node.
4. Edit the **second Merge** node, and then:
  - for **Merge Method**, select **Keys**
  - for **Keys for merge**, select **tariff**
  - enable the option **Include matching and selected non-matching records (partial outer join)**
  - click **Select**
  - ensure that the formerly merged dataset is the leading dataset
  - click **OK** to close the **Merge: Select Dataset for Outer Join** sub dialog
5. Click **Preview**

A section of the results appear as follows:

| dget_L | fixed_cost | free_minutes | peak_rate | offpeak_rate | weekend_rate |
|--------|------------|--------------|-----------|--------------|--------------|
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |
|        | 17.500     | 100          | 15        | 5            | 5            |

The fields coming from the tariff dataset are added to the (merged) customer data. Customers in the same tariff group will have the same values on the fields originating from the tariff dataset. From a database perspective, this is not ideal; however, MODELER needs a rectangular data structure, so there is no way around this.

6. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Merge** dialog box.

Leave the stream open for the next task.

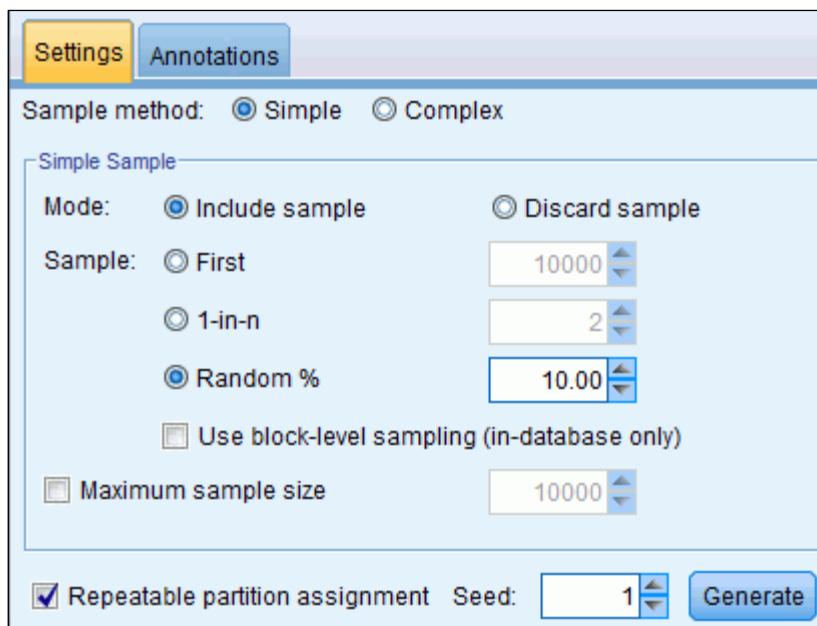
## Task 4. Sample records.

From the combined dataset (storing data from four different sources), a random sample is drawn of approximately 10%, and the data are cached. Furthermore, to ensure that the student samples the same records as are sampled in this demo, the value for the random seed will be fixed.

In this task you will build from the previous stream.

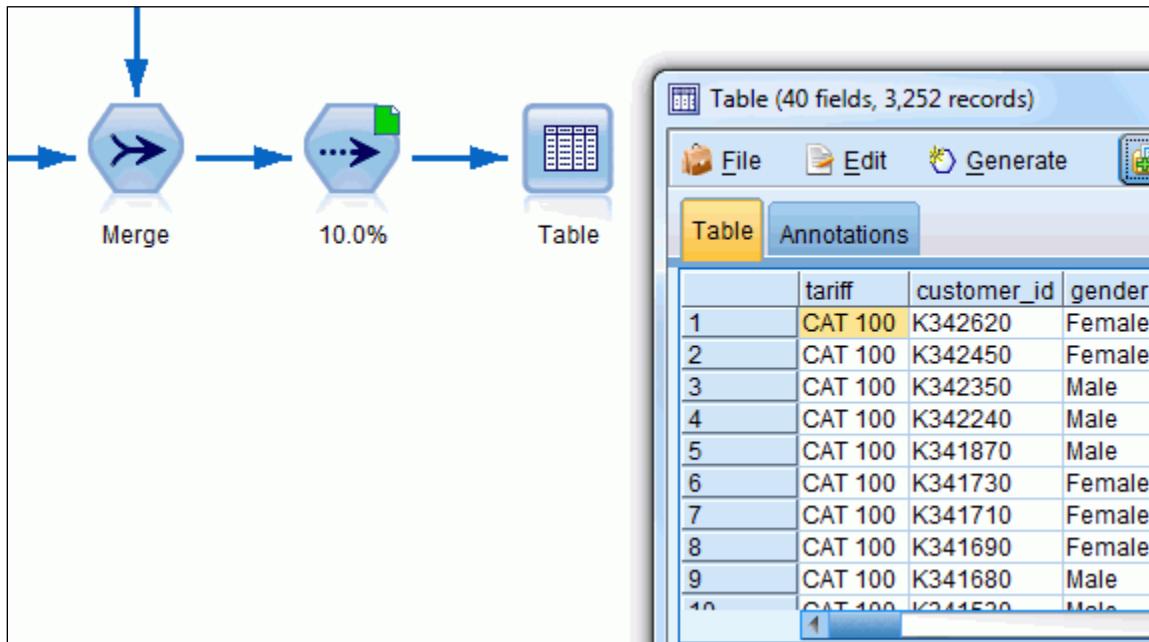
1. Add a **Sample** node (Record Ops palette) downstream from the **second Merge node**.
2. Right-click the **Sample** node, and then select **Cache...Enable** from the context menu.
3. Edit the **Sample** node, and then (on the Settings tab):
  - for **Sample**, select **Random %**
  - type **10** for the percentage to sample
  - enable the **Repeatable partition assignment** option
  - for **Seed**, type **1**

A section of the specifications in the Sample dialog box appear as follows:



4. Click **OK** to close the **Sample** dialog box.
5. Add a **Table** node downstream from the **Sample** node.
6. Run the **Table** node.

A section of the results appear as follows:



The cache is filled (the document icon has turned green) and 3,252 records are sampled. The next time when the Table node is executed, it will run from the cache, returning the same 3,252 records (to be clear: this is the effect of the cache, not because of the fixed seed value).

This completes the demo for this module. You will find the solution results in **demo\_integrating\_data\_completed.str**, located in the **07-Integrating\_Data\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Is the following statement true or false? The Append node can only add records from two datasets together.

- A. True
- B. False

Question 2: Is the following statement true or false? The Merge node can only merge two datasets at the same time.

- A. True
- B. False

Question 3: Refer to the figure that follows. Is the following statement true or false?  
The pupil field must be used as key field to merge the two datasets.

- A. True
- B. False

| A       |           |               | B       |           |            |
|---------|-----------|---------------|---------|-----------|------------|
| pupil   | classroom | grade_english | pupil   | classroom | grade_math |
| John    |           | 1 A           | John    |           | 1 B        |
| Michael |           | 1 D           | Michael |           | 1 C        |
| Richard |           | 1 B           | Nancy   |           | 1 B        |
| John    |           | 2 D           | John    |           | 2 C        |
| Nancy   |           | 2 D           | Richard |           | 2 D        |

Question 4: Refer to the figure that follows. Is the following statement true or false?  
These datasets cannot be merged, because the unit of analysis is id in dataset A and zipcode in dataset B.

- A. True
- B. False

| A  |         |        | B       |                              |
|----|---------|--------|---------|------------------------------|
| id | zipcode | gender | zipcode | pct with car in zipcode area |
| 1  | 10      | male   | 10      | 40                           |
| 2  | 10      | female | 11      | 65                           |
| 3  | 11      | male   | 12      | 81                           |
| 4  | 11      | female |         |                              |
| 5  | 12      | male   |         |                              |

Question 5: Which of the following is the correct statement? Refer to the figure that follows. What was the merge type used to merge datasets A and B into dataset C?

- A. An inner join.
- B. An outer join.
- C. A partial outer join, with dataset A as the leading dataset.
- D. An anti-join, with dataset A as the main dataset.

| A  |         |        | B  |     |  | C  |         |        |
|----|---------|--------|----|-----|--|----|---------|--------|
| id | zipcode | gender | id | age |  | id | zipcode | gender |
| 1  | 10      | male   |    |     |  | 1  | 10      | male   |
| 2  | 10      | female |    |     |  | 2  | 10      | female |
| 3  | 11      | male   |    |     |  | 3  | 11      | male   |
| 4  | 11      | female | 4  | 22  |  |    |         |        |
| 5  | 12      | male   | 5  | 33  |  |    |         |        |
|    |         |        | 6  | 44  |  |    |         |        |
|    |         |        | 7  | 55  |  |    |         |        |

Question 6: Is the following statement true or false? The Sample node enables you to discard the first 1000 records.

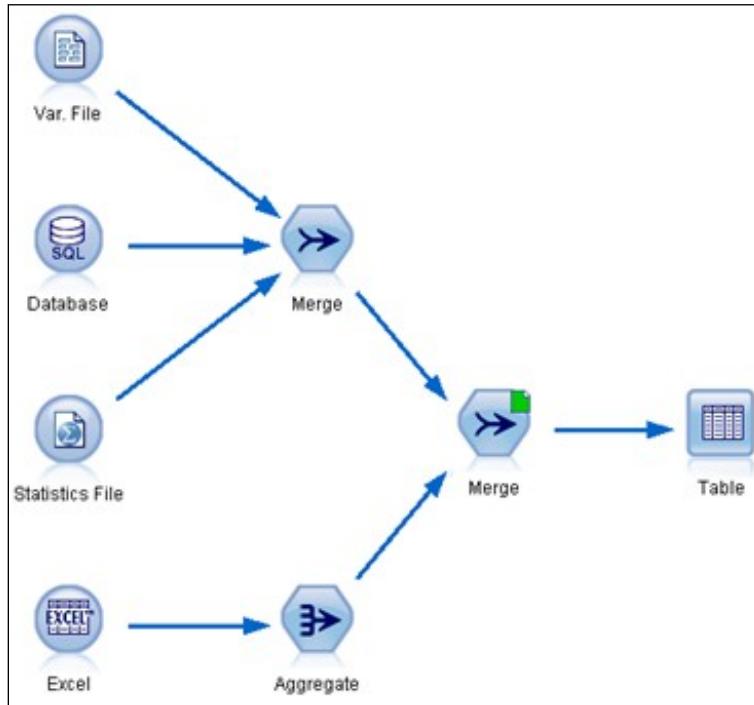
- A. True
- B. False

Question 7: Is the following statement true or false? When a sample must be replicated in a MODELER session next week, you can set the value for the seed.

- A. True
- B. False

Question 8: Is the following statement true or false? Refer to the figure below. When you execute the Table node the data will be read from the source nodes.

- A. True
- B. False



## Answers to questions:

Answer 1: B. False. More than 2 datasets can be input to the Append node.

Answer 2: B. False. More than 2 datasets can be input to the Merge node.

Answer 3: B. False. The datasets must be merged on pupil and classroom.

Answer 4: B. False. These datasets can be merged, because MODELER supports a 1-to-many merge.

Answer 5: D. Anti-join, because only records unique to dataset 1 are in the output dataset.

Answer 6: A. True. Select the option Discard records, in combination with First n.

Answer 7: A. True. To replicate a sample in another MODELER session, fix the seed value.

Answer 8: B. False. There is a cache on the Merge data, so executing the Table node will read the data from the cache established on the Merge node.

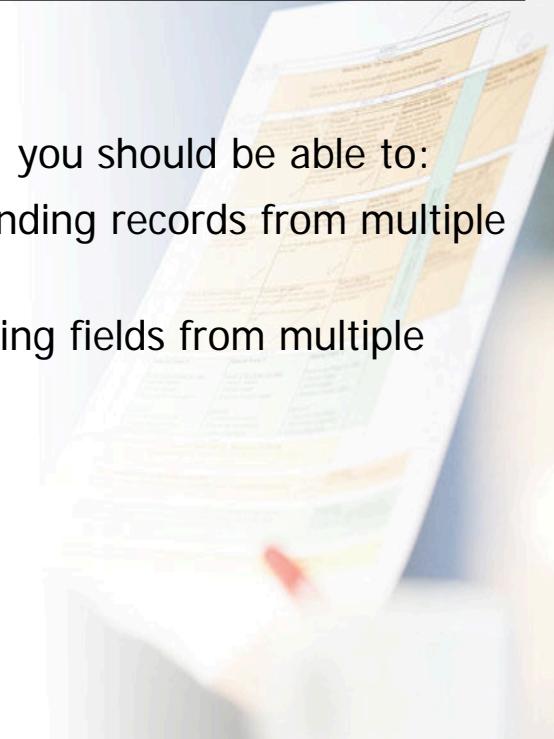
Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - integrate data by appending records from multiple datasets
  - integrate data by merging fields from multiple datasets
  - sample records

© 2014 IBM Corporation



This module presented two methods to integrate datasets: appending records and merging fields.

This module also introduced you to sampling records, and the possibility to cache the data. Sampling records and putting a cache on nodes may save you a significant amount of time.

# Workshop 1

## Integrating data



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **workshop\_integrating\_data.str**, located in the **07-Integrating\_Data\Start Files** sub folder.

The following (synthetic) files are used in this workshop:

- **ACME customer data.xls**: a Microsoft Excel file storing background information on customers; such as gender ( duplicate records are removed so that there is one record per customer)
- **ACME purchases: 1999 - 2004.dat**: purchases made by customers from period 1999 to 2004 (a customer can have multiple records, because he can have more than one purchase)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- **ACME purchases 2005 - 2010.dat**: purchases, for the period 2005 – 2010
- **ACME orderlines 1999 - 2004.sav**: an IBM SPSS Statistics file, storing orders per purchase, from 1999 to 2004 (a purchase can have multiple records, every item purchased makes up one record)
- **ACME orderlines 2005 - 2010.sav**: an IBM SPSS Statistics file with order lines, for the period 2005 – 2010
- **ACME mailing history.xlsx**: a Microsoft Excel file with data on mailings, with flag fields created to have a dataset with one record per customer
- **ACME zip data.csv**: information on postal codes
- **workshop\_integrating\_data.str**: a stream file that imports the datasets, which is the starting point for the workshop

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Workshop 1: Integrating Data

You are working for ACME, a company that sells sport products. It is your job to combine a number of datasets into a single dataset, so that models can be built using the information from all these datasets later.

- Consider the two datasets storing information about ACME purchases, one dataset for the 1999- 2004 period (**ACME purchases 1999 - 2004.dat** - 4,018 records), the other dataset for the 2005 – 2010 period (**ACME purchases 2005 - 2010.dat** - 67,109 records). Create a single dataset storing purchases for the whole 1999 – 2010 period.

Note: The nodes to import these datasets are already on the stream canvas after you have opened **workshop\_integrating\_data.str**, located in the **07-Integrating\_Data\Start Files** sub folder.

To check your results: The new dataset should be comprised of  $4,018 + 67,109 = 71,127$  records.

- Consider the two datasets storing information about the items that were ordered (**ACME orderlines 1999 - 2004.sav** and **ACME orderlines 2005 - 2010.sav**). Create a single dataset containing all order lines for the whole 1999 - 2010 period.

To check your results: The new dataset should be comprised of 169,734 records.

- From the dataset storing the order lines for the whole 1999 - 2010 period (the dataset 169,734 records created in the previous task), create a dataset so that there is only one record per PURCHASE ID, and a field that gives the total price for each PURCHASE\_ID (do not include a record count field).

To check your results: There should be 71,127 records in the new dataset. Also, check that the PURCHASE ID 5723 has a total price of 948.48.

- Create a single dataset from:

- the dataset storing the purchase information (created in the first task - 71,127 records)
- the dataset storing aggregated information for the purchases that you created in the previous task (also 71,127 records)

You may assume that both datasets include the same PURCHASE\_ID's.

- For the dataset created in the previous task, ensure that there is 1 record per customer, with fields CUSTOMER\_ID, most recent order date, the total price the customer has paid for all purchases, and the number of purchases the customer made.

To check your results: There should be 30,000 records in this dataset.

- Create a single dataset from:

- the dataset having customer background information originating from **ACME customer data.xls** (already on the stream canvas)
- the dataset that is comprised of the customer purchase information (created in the previous task)
- the dataset with the customer mail history originating from **ACME mailing history.xlsx** (already on the stream canvas)

The dataset having the customer information must be the leading dataset in combining these datasets.

Note: When combining datasets on a key field, the key field name must be identical (in name and case) in order to have the field appear in the **Possible keys** area. Use a Filter node where needed to ensure that the name of the key field is the same in all datasets.

- Enrich the dataset that you created in the previous task with zipcode information stored in the file **ACME zip data.csv** (already on the stream canvas).
- Export a 25% random sample of the data to a comma-separated text file (use **123** for the seed value); name the file **ACME sample.dat**.

To check your results: 7,568 records will be sampled.

For more information about where to work and the workshop results, refer to the Task and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

## Workshop 1: Tasks and Results

### Task 1. Create a single dataset for purchases.

- Use the **Append** node to join the two datasets. No edits are needed in the Append node.
- Running a **Table** node downstream from the **Append** node will show that you have 71,127 records in the new dataset.

### Task 2. Create a single dataset for order lines.

- Use the **Append** node to join the two datasets. No edits are needed in the Append node.
- Running a **Table** node downstream from the **Append** node will show that the new dataset is comprised of 169,734 records.

### Task 3. Create a dataset with one record per purchase for order lines.

- Add an **Aggregate** node downstream from the **Append** node.
- Edit the **Aggregate** node, and then:
  - for **Key fields**, select **PURCHASE\_ID**
  - for **Aggregate fields**, select **PRICE**, and statistic **Sum**
  - **disable** the option **Include record count in field**
- Run a **Table** node downstream from the **Aggregate** node to verify that the aggregated dataset has 71,127 records, and PURCHASE ID 5723 has a PRICE\_Sum of 948.48.

### Task 4. Create a single dataset from the purchases dataset and the (aggregated) order lines dataset.

- Add a **Merge** node to the stream canvas, and connect the **Append** node (originating from the appended datasets storing the purchases) and the **Aggregate** node (originating from the appended datasets storing order lines) to it.

- Edit the **Merge** node, and then:
  - for **Keys for merge**, select **PURCHASE\_ID**
  - use the (default) **Inner join** method for merging (both datasets include the same PURCHASE IDs, so an inner, outer, or partial join will give the same results)

A section of the results appear as follows.

| PURCHASE_ID | CUSTOMER_ID | ORDERDATE  | PRICE_Sum |
|-------------|-------------|------------|-----------|
| 5723        | 5030        | 2003-06-24 | 948.480   |
| 5724        | 3119        | 2002-12-03 | 114.840   |
| 5726        | 2764        | 2002-03-13 | 217.770   |
| 5727        | 2190        | 2003-01-21 | 466.130   |
| 5729        | 1887        | 1999-07-03 | 985.520   |
| 5730        | 783         | 2002-05-26 | 935.350   |
| 5731        | 5584        | 1999-02-12 | 746.840   |
| 5732        | 2009        | 2000-10-16 | 940.540   |
| 5734        | 3178        | 2001-11-07 | 460.000   |
| 5736        | 5634        | 2003-11-18 | 605.410   |

## Task 5. Create a dataset with one record per customer from the purchases and order lines dataset.

- Add an **Aggregate** node downstream from the **Merge** node.
- Edit the **Aggregate** node, and then:
  - for **Keys fields**, select **CUSTOMER\_ID**
  - for **Aggregate fields**, select **ORDERDATE** and statistic **Max**; also select **PRICE\_Sum** and statistic **Sum**
  - ensure that the option **Include record count in field** is enabled

- Run a **Table** node downstream from the **Aggregate** node to verify that you have 30,000 records.

A section of the results appear as follows (note: the data depicted below are sorted on CUSTOMER\_ID):

| CUSTOMER_ID | ORDERDATE_Max | PRICE_Sum_Sum | number_of_purchases |
|-------------|---------------|---------------|---------------------|
| 723         | 2009-07-02    | 546.731       | 2                   |
| 724         | 1999-07-11    | 306.480       | 1                   |
| 725         | 2007-02-27    | 1796.016      | 4                   |
| 726         | 2010-06-14    | 2489.259      | 6                   |
| 727         | 1999-02-21    | 377.210       | 1                   |
| 728         | 2004-05-22    | 624.015       | 1                   |
| 729         | 2006-07-13    | 34.724        | 1                   |
| 730         | 2002-11-29    | 788.150       | 1                   |
| 731         | 2003-07-17    | 2413.870      | 3                   |
| 732         | 2003-05-10    | 1158.180      | 2                   |

## Task 6. Create a single dataset for customer information, their purchases and their mailing history.

- Ensure that the key field has the same name, CUSTOMER\_ID, in all three datasets. In the dataset storing the mailing history, the field name is not CUSTOMER\_ID, but customer\_id. To rename this field, add a **Filter** node downstream from the **SetToFlag** node, and replace **customer\_id** with **CUSTOMER\_ID**.
- Add a **Merge** node to the stream canvas, and connect the **Aggregate** node (from the previous task), the **Distinct** node and the **Filter** node to it.
- Edit the **Merge** node, and then:
  - for **Keys for merge**, select **CUSTOMER\_ID** (after having renamed customer\_id to CUSTOMER\_ID; this field is available as key field)
  - select the **Partial outer join** merge method, click **Select**, and ensure that **ACME customer data.xls** is selected as the leading dataset in the merge
- Run a **Table** node downstream from the **Merge** node to verify your results.

A section of the results appear as follows:

| CUSTOMER_ID | GENDER | CREDITLIMIT | ZODIAC | E-MAIL ADDRESS       | ZIP    | ORDERDATE_Max | PRICE_Sum_Sum | number_of_purchases | mailing_Standard tennisracket |
|-------------|--------|-------------|--------|----------------------|--------|---------------|---------------|---------------------|-------------------------------|
| 723.000 F   |        | 9026.000    | 1.000  | name7502@tnet.fr     | 1818BO | 2009-07-02    | 546.731       | 2 F                 |                               |
| 724.000 M   |        | 5223.000    | 3.000  | name25485@wwmail.org | 1132DG | 1999-07-11    | 306.480       | 1 F                 |                               |
| 725.000 F   |        | 668.000     | 8.000  | name15543@wwmail.de  | 1803YT | 2007-02-27    | 1796.016      | 4 F                 |                               |
| 726.000 M   |        | 8302.000    | 6.000  | name28335@zigzag.be  | 1205WR | 2010-06-14    | 2489.259      | 6 F                 |                               |
| 727.000 F   |        | 997.000     | 8.000  | name5354@tnet.jp     | 1711ON | 1999-02-21    | 377.210       | 1 T                 |                               |
| 728.000 M   |        | 1955.000    | 3.000  | name20637@wwmail.es  | 1055FG | 2004-05-22    | 624.015       | 1 F                 |                               |
| 729.000 F   |        | 6898.000    | 5.000  | name20636@wwmail.es  | 1254MR | 2008-07-13    | 34.724        | 1 F                 |                               |
| 730.000 F   |        | 6488.000    | 11.000 | name10414@tnet.inc   | 1723DG | 2002-11-29    | 788.150       | 1 T                 |                               |
| 731.000 F   |        | 5006.000    | 2.000  | name23372@wwmail.inc | 1713AQ | 2003-07-17    | 2413.870      | 3 F                 |                               |
| 732.000 F   |        | 929.000     | 5.000  | name20635@wwmail.es  | 1264EC | 2003-05-10    | 1158.180      | 2 F                 |                               |

## Task 7. Enrich the data with zipcode information.

- Add a **Merge** node to the stream, and connect the **Merge** node (from the previous task), and the **Var. File** node importing **ACME zip data.csv** to it.
- Edit the **Merge** node, and then:
  - for **Keys for merge**, select **ZIP**
  - select the **Partial outer join** merge method, click **Select**, and ensure that the dataset originating from the Merge node in the previous task is the leading dataset in the merge
- Run a **Table** node downstream from the **Merge** node to verify your results.

A section of the results appear as follows:

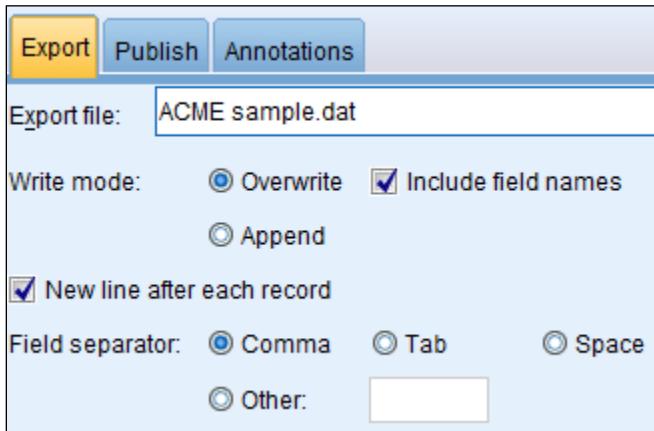
| ZIP    | CUSTOMER_ID | GENDER | CREDITLIMIT | ZODIAC | E-MAIL ADDRESS      |
|--------|-------------|--------|-------------|--------|---------------------|
| 1001EZ | 4015.000 F  |        | 3692.000    | 2.000  | name25156@wwmail.o  |
| 1001EZ | 9016.000 F  |        | 3692.000    | 1.000  | name2048@lomejor.es |
| 1001EZ | 24016.000 F |        | 3692.000    | 9.000  | name6707@tnet.fr    |
| 1001EZ | 19016.000 F |        | 3692.000    | 3.000  | name8954@tnet.inc   |
| 1001EZ | 29016.000 F |        | 3692.000    | 8.000  | name8113@tnet.fr    |
| 1001EZ | 14016.000 F |        | 3692.000    | 4.000  | name18111wwmail.es  |
| 1001FA | 26952.000 F |        | 6747.000    | 5.000  | name21012@wwmail.in |
| 1001FA | 11952.000 F |        | 6747.000    | 1.000  | name8125@tnet.fr    |
| 1001FA | 16952.000 F |        | 6747.000    | 10.000 | name25364@wwmail.o  |
| 1001FA | 21952.000 F |        | 6747.000    | 11.000 | name14306@tnet.uk   |

Note: The customer dataset was the first dataset in the merge. The fields originating from the zipcode dataset are the last fields in the new dataset and are not shown in this figure.

## Task 8. Export a random sample of 25%.

- Add a **Sample** node downstream from the **Merge** node (from the previous task)
- Edit the **Sample** node, and then:
  - for **Sample**, select **Random %**, and set the value to **25**
  - enable the option **Repeatable partition assignment**, and set **Seed** to **123**
- Run a **Table** node downstream from the **Sample** node to verify the results (7,568 records should be sampled).
- Add a **Flat File** node (Export palette) downstream from the **Sample** node.
  - Edit the **Flat File** node, and then, for **Export file**, type **ACME sample.dat**

A section of the specifications in the Flat File dialog box appear as follows:

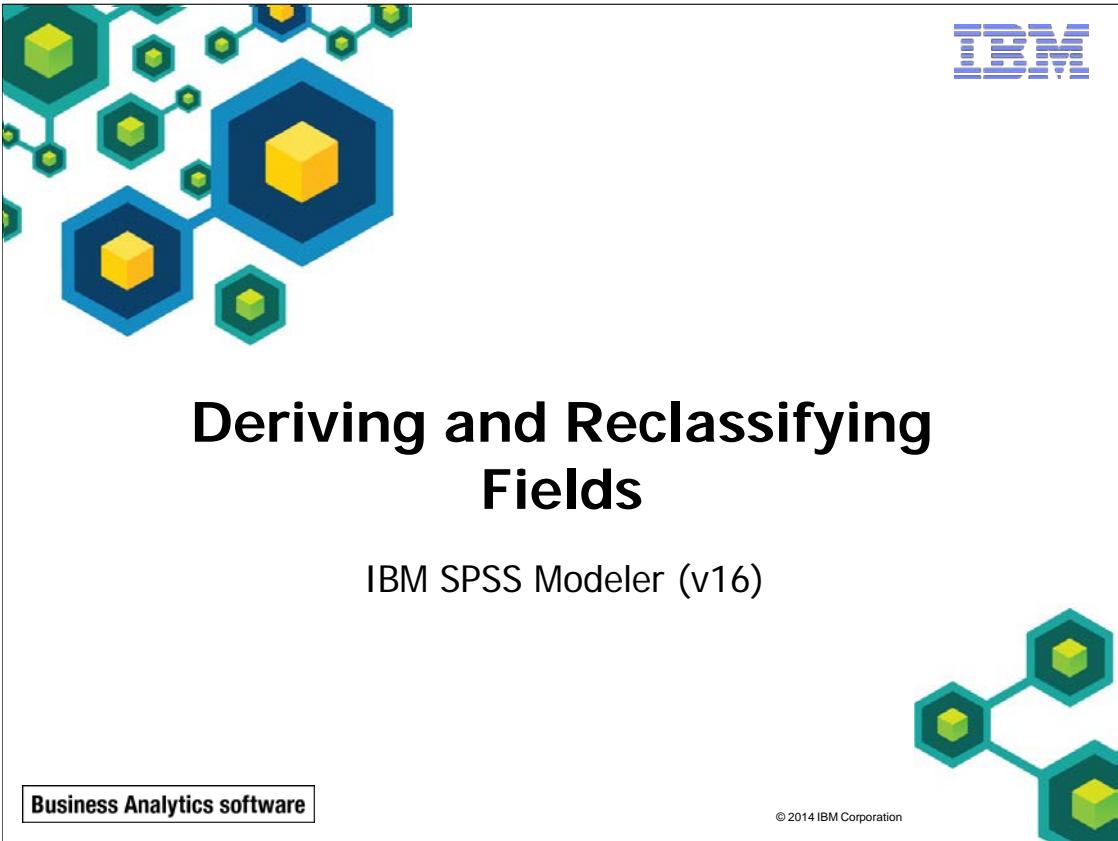


**Note.** The stream **workshop\_integrating\_data\_completed.str**, located in the **07-Integrating\_Data\Solution Files** sub folder, provides a solution to the workshop exercises.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The advertisement features a white background with a decorative pattern of blue and green hexagonal shapes containing yellow cubes. In the top right corner is the IBM logo. Below the pattern, the title "Deriving and Reclassifying Fields" is displayed in a large, bold, black sans-serif font. Underneath the title, the text "IBM SPSS Modeler (v16)" is shown in a smaller, regular black font. At the bottom left, a small rectangular box contains the text "Business Analytics software". At the bottom right, there is a copyright notice: "© 2014 IBM Corporation".

# Deriving and Reclassifying Fields

IBM SPSS Modeler (v16)

Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

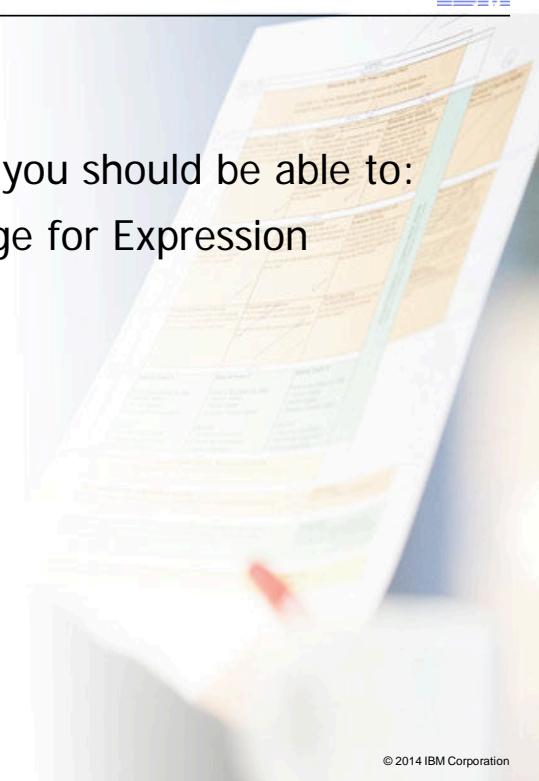
This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Objectives

- At the end of this module, you should be able to:
  - use the Control Language for Expression Manipulation (CLEM)
  - derive new fields
  - reclassify field values



© 2014 IBM Corporation

The focus in this module is on another task in the data preparation stage: construct the final dataset for modeling by cleansing and enriching your data.

Before reviewing this module you should be familiar with:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data
- methods to explore the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

## Methods to Create Fields

| ID | BDATE      | age | agecat | adult | GENDER | gender_ok |
|----|------------|-----|--------|-------|--------|-----------|
| 1  | 01/24/1940 | 71  | 3      | T     | Fem    | Female    |
| 2  | 05/11/1968 | 43  | 2      | T     | F      | Female    |
| 3  | 09/11/1989 | 22  | 1      | T     | Female | Female    |
| 4  | 10/14/1992 | 19  | 1      | F     | MALE   | Male      |

**Derive**

**Reclassify**



© 2014 IBM Corporation

This module presents two field operations nodes that can be used for cleansing and your data. The Derive node computes new fields; the Reclassify node recodes the values of a categorical field.

This slide shows some examples:

- Based on BIRTHDATE, three new fields are derived: age, age category (1 junior, 2 middle-age, 3 senior) and a field flagging if the person is 21 years or older.
- The GENDER field shows inconsistencies in spelling and is reclassified into a new field with values Female and Male.

MODELER offers many more field operations nodes to prepare your data for analyses and modeling. Refer to the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for more information.

## Introducing the Control Language for Expression Manipulation (CLEM)

- MODELER's native language to build expressions
- Derive node, Select node use CLEM
- Refer to the online Help for a full discussion of CLEM

© 2014 IBM Corporation



MODELER implements a powerful language for specifying expressions called Control Language for Expression Manipulation (CLEM). CLEM is used in a number of nodes, among which the Select and the Derive node.

CLEM enables you to:

- specify conditions, for example, income < 10000
- specify expressions to assign values to fields, for example, tax = income \* 0.1

Refer to the online Help for a detailed presentation of CLEM. In this module you are introduced to the basic concepts.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

# Introducing CLEM Expressions

The diagram illustrates the structure of a CLEM expression with the following components and their corresponding labels:

- Function:** `datetime_year`
- Field:** `(connect_date)`
- Operator:** `=`
- Value:** `2000`
- Operator:** `and`
- Field:** `handset`
- Operator:** `=`
- Value:** `"ASAD"`

© 2014 IBM Corporation

This slide shows an example of a CLEM expression, specifying a condition. The condition returns true for a customer if he connected in the year 2000 and has handset ASAD.

CLEM expressions are constructed from values, fields, operators and functions. When writing CLEM expressions, take note of the following:

- Values smaller than 1 (in absolute value), must be specified with a leading zero, or else MODELER will issue an error message such as *CLEM error: Illegal token'. in expression: probability > .5* (the value should have been specified as 0.5).
- String values should be within single or double quotes, for example 'male' or "male", 'married' or "married". Occasionally, in specific functions such as the locchar function, a string value need to be enclosed in single backquotes ` and `.

- Field names are case sensitive. For example, the field names age and AGE are not the same. When a field name contains a blank or if it is a special field name it needs to be enclosed in single quotes for example: 'INCOME 2012', '\$RC-churn'.
- MODELER offers many functions. Refer to the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for more details or the online Help for a complete overview. Function names are case-sensitive.
- Operators can be:
  - arithmetic: +, -, \*, /, \*\* (raise to the power)
  - relational: >, <, <=, >=, =, /= (unequal)
  - logical: and, or, not (all in lower case).

To be not dependent on how MODELER evaluates an expression, it is advised to use parentheses in compound conditions. For example, if you want to refer to men, or those working full time with an income greater than 10000:

(gender= "male") or (job\_status="full time" and income > 10000)

The screenshot shows the Expression Builder dialog box. On the left, there is a list of 'General Functions' with columns for 'Function' and 'Return'. One function, 'is\_integer(ITEM)', is highlighted. In the center, there are buttons for arithmetic operators (+, -, \*, /, mod, %) and comparison operators (>, >=, <, <=, =, /=, and, or, not). To the right, there is a list of 'Fields' with columns for 'Type', 'Field', and 'Storage'. A field named 'customer\_id' is highlighted. At the bottom of the dialog, the description for 'is\_integer(ITEM)' is displayed: 'Returns a value of true if ITEM type is an integer. Otherwise, returns a value of false.' The dialog has a header 'Expression Builder' and a footer '© 2014 IBM Corporation'.

In nodes such as Select and Derive you can type your CLEM expression, but that is probably not an efficient option, especially because field names and function names are case-sensitive. Instead of, or in conjunction with typing CLEM expressions, you can use the Expression Builder to create expressions.

You can invoke the Expression Builder by clicking the Launch expression builder button in the Select or Derive dialog box.

This slide shows the Expression Builder dialog box. Build your expression by selecting and pasting the various elements (fields, functions, values and operators) to the area where the expression must be specified.

Functions are grouped by categories, such as string functions, date and time functions, numeric functions, and logical functions. When you select a function you will have a description of its use at the bottom of the dialog box.

When you need to specify a value of a categorical field MODELER offers the very user-friendly feature to pick the value from a list of values, provided that the field is instantiated. If the field is not instantiated, its values will not be available.

## Deriving Fields

- Create new fields
- Use the Derive node (Field Ops)
- Derive types:
  - Formula
  - Flag
  - Nominal
  - Conditional



© 2014 IBM Corporation

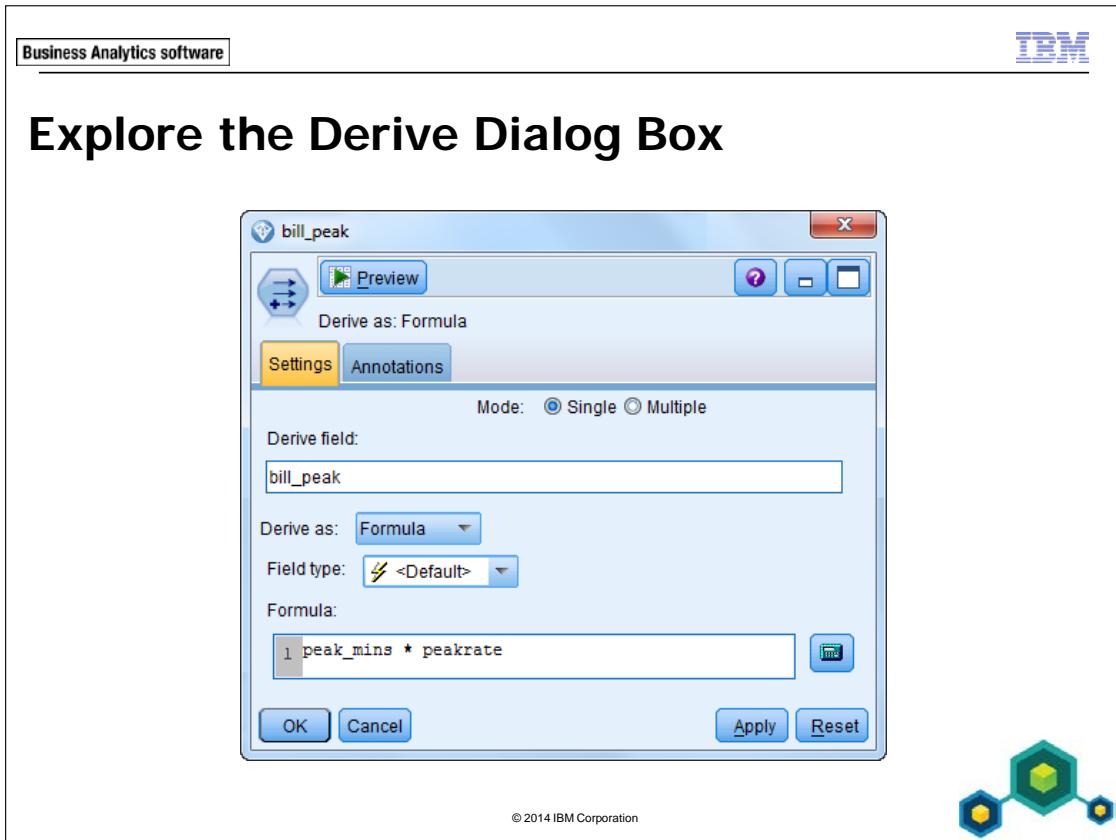


The Derive node, located in the Field Ops palette, will add a new field to the dataset. The Derive node does not let you overwrite an existing field. If you want to overwrite a field, use a Filler node.

Using the Derive node you can derive a field of one of the following types:

- Formula: An outcome of a formula. For example a new field TAX derived as:  $TAX = INCOME * 0.20$ .
- Flag: A T/F field. For example: a new field ADULT, T when  $AGE \geq 21$ , else F.
- Nominal: A categorical field. For example: a new field AGECAT, 1 when  $AGE \leq 35$ , 2 when  $AGE > 35$  and  $AGE \leq 70$ , and 3 when  $AGE > 70$ .
- Conditional: An outcome of a formula, but computed conditionally. For example, a new field  $TAX = 0.1 * INCOME$  if  $INCOME \leq 100000$ , and  $TAX = 10000 + 0.2 * (INCOME - 100000)$  if  $INCOME > 100000$ .

For other derive types or for a presentation of the Filler node, refer to the online Help or the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course.



This slide shows the Derive dialog box. A handy feature is to derive multiple fields with one Derive node. This applies when multiple fields must be derived using the same calculation. For example, when you want to derive 5 new fields by deriving 5 source fields by 100, use the multiple mode instead of using 5 separate Derive nodes.

Specify the name of the new field in the Derive field text box, and select the derive type (Formula/Flag/Nominal/State/Count/Conditional) in the Derive as drop-down list.

For Field Type, select a measurement level for the newly derived field. When the field type is not specified, MODELER will automatically assign a measurement level. For example, MODELER will assign measurement level Flag when you derive a flag field. As one of the few examples where you will set the measurement level manually, think of deriving a field such as AGE CATEGORY with values 1, 2 and 3. This field will be derived as nominal and autotyped as nominal, while its measurement level should be ordinal.

The dialog box will reflect the derive type that is selected. This slide shows the dialog box when the derive type is Formula.

# Deriving Fields and Blanks

**Derived fields**

| ID | AGE      | INCOME   | adult    | income_in_1000s |
|----|----------|----------|----------|-----------------|
| 1  | \$null\$ | 21000    | \$null\$ | 21              |
| 2  | 23       | \$null\$ | T        | \$null\$        |
| 3  | 19       | -1       | F        | -0.001          |
| 4  | 999      | 10000    | T        | 10              |

© 2014 IBM Corporation



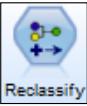
MODELER will treat user-defined blank values as valid values when a field is derived. This slide shows a few examples of how MODELER handles blanks. The value 999 is declared as blank value for AGE, and -1 is declared as blank value for INCOME.

Based on AGE and INCOME two fields are derived. The first field, adult, equals T when AGE is greater than 20, else it returns F. The second field, income\_in\_1000s, is derived as INCOME/1000. Although 999 is declared as a blank value AGE, MODELER will treat the blank as any other value and so adult equals T when AGE equals 999. When INCOME equals -1, the blank value for this field, income\_in\_1000s field is computed as -1/1000, with -0.001 as the result.

When the original values are undefined (\$null\$), the result is also undefined (\$null\$), which is as it should be.

Note: The undefined value (\$null\$) is referred to in MODELER's user-interface as undef. When you specify undef as value, MODELER will return \$null\$.

## Reclassifying Fields

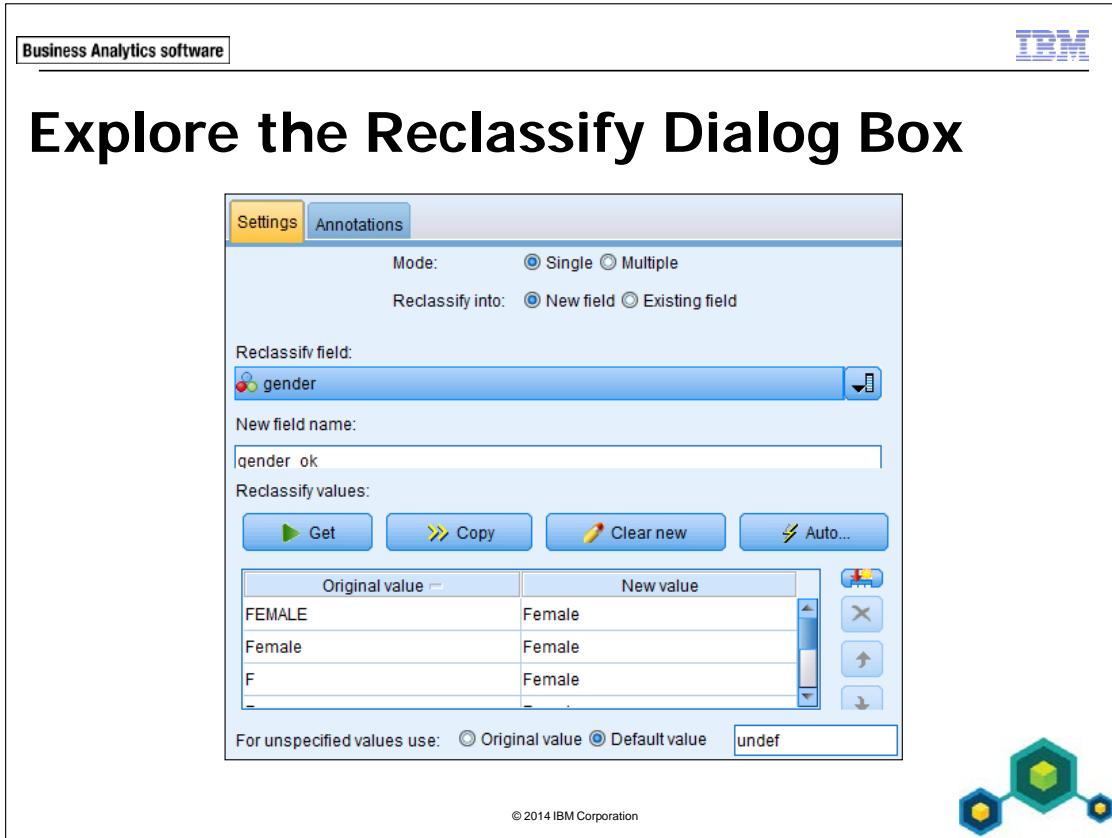
- Recode the values of a categorical field into broader categories
- Use the Reclassify node (Field Ops) 

© 2014 IBM Corporation

Sometimes you need to recode the values of a categorical field into broader categories. For example, a field that stores a customer's specific job position may be more useful for prediction if it is reclassified into broader job categories.

The Reclassify node, located in the Field Ops palette, enables you to reclassify values of a categorical field.

It should be emphasized that the Reclassify node is to recode categorical fields only. For example, if you want to recode the continuous field AGE into age categories, you will use the Derive node, and not the Reclassify node.

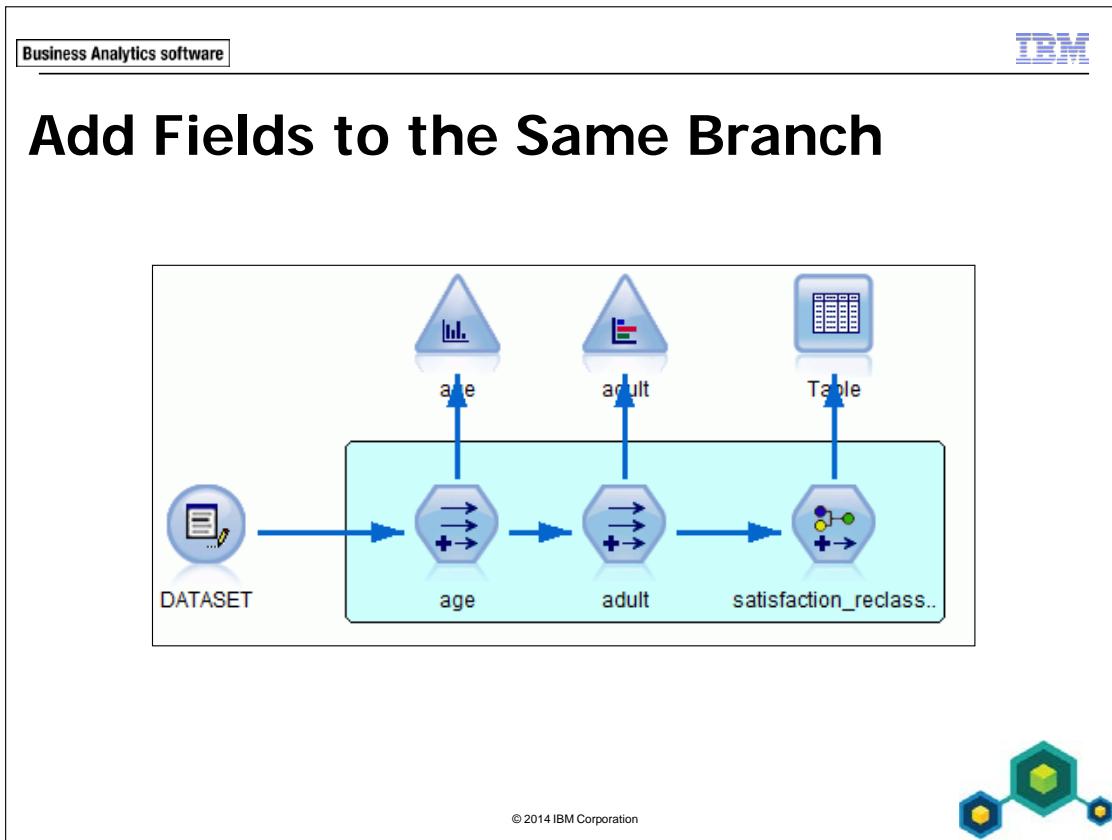


In the Reclassify dialog box, select single mode when you want to reclassify only one field. Multiple mode is useful when the same reclassification rules must be applied to a number of fields. For example, when you want to recode 15 satisfaction fields that are measured on a 7-point scale into 3-point scales, you would select the Multiple Mode option instead of using 15 separate Reclassify nodes.

The new values will be placed in a new field by default. Alternatively, you can choose to Reclassify into the same field and replace the field's values.

Under Reclassify click the Get button to populate the Original value column with values (assuming the field is instantiated). You can click the Copy button to copy the values in the Original value column into the New Value column. This is useful if you want to retain most of the original values, reclassifying only a few.

When a value is encountered in the source field that is not listed in the Original value column, either the value itself can be output or the default value can be assigned, which is the undef (\$null\$) value.



When you derive and reclassify fields, it is recommended to add the nodes to the same branch of the stream. This makes it possible to create a field from an earlier created field. Also, adding new fields downstream in the same branch is necessary to use all of these new fields together in modeling.

This slide shows an example. Two Derive nodes and one Reclassify node are added to the stream to create a dataset that includes all fields.

## Checking Your Results

- Preview the data, or run a Table node
- Use a Matrix node
- Use an Aggregate node
- Run a Data Audit

© 2014 IBM Corporation

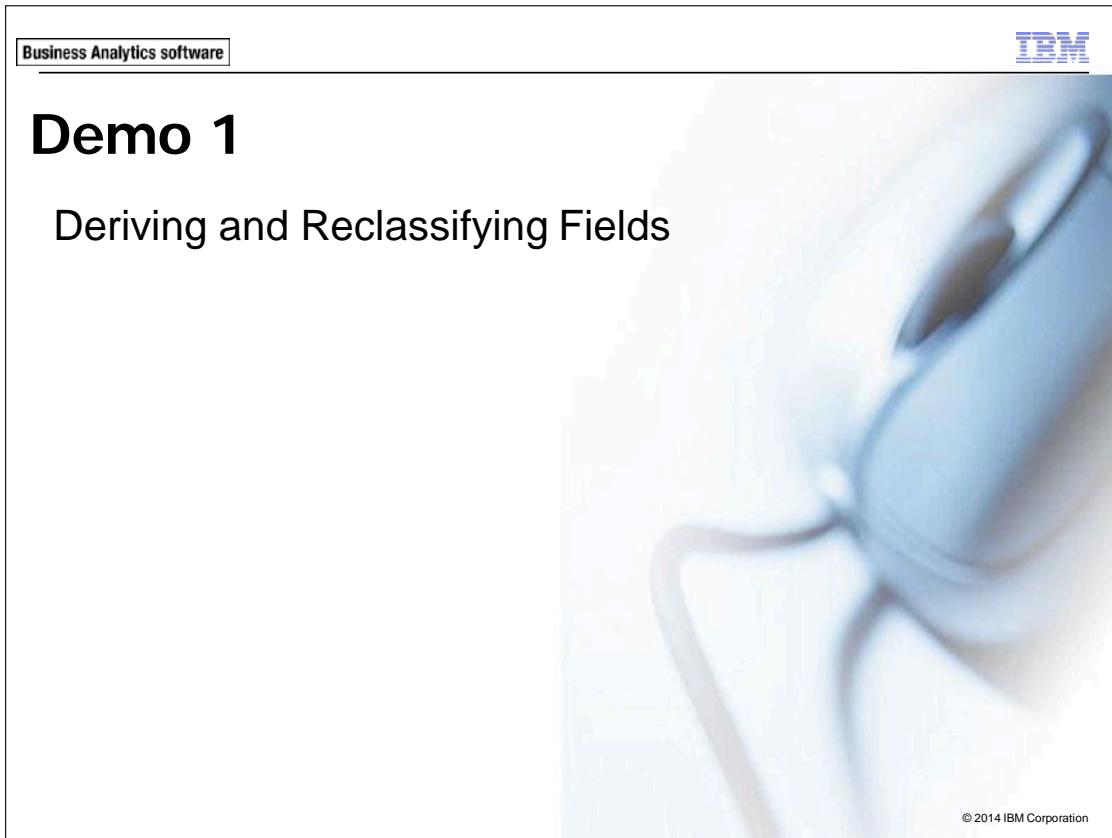


Whenever new fields are created it is advised to check the results. A formula can be specified incorrectly, values for a field can be specified incorrectly, conditions can be specified following human instead of computer logic, blank values can affect the results, and so forth. In many of these situations MODELER will not issue an error message, so simply not receiving an error does not mean that the new field is correct.

How to check a new field depends on how that field was created:

- For fields created from formulas, simply review the output from a Table node and calculate a few values to check the equation.
- For a categorical fields created from another categorical field, use a Matrix node to cross tabulate the fields.
- For categorical fields created from a continuous field, use the Aggregate node with the new field as key field and request the minimum/maximum of the original field.

Also, running a Data Audit node will show the minimum and maximum and provides a quick check.



The slide features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The main title "Demo 1" is centered at the top, followed by the subtitle "Deriving and Reclassifying Fields". The background of the slide is a blurred image of a person wearing a hard hat and safety glasses, looking at a blueprint or map. A small copyright notice "© 2014 IBM Corporation" is visible in the bottom right corner of the slide area.

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **demo\_deriving\_and\_reclassifying\_fields.str**, located in the **08-Deriving\_and\_Reclassifying\_Fields\Start Files** sub folder.

The files that are used in this demo are:

- **telco x data.txt**: a (synthetic) text file storing data on customers of a (fictitious) telecommunications firm
- **demo\_deriving\_and\_reclassifying\_fields.str**: a stream file that imports the data, sets measurement levels, and serves as the starting point for the demo.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Demo 1: Deriving and Reclassifying Fields

### Purpose:

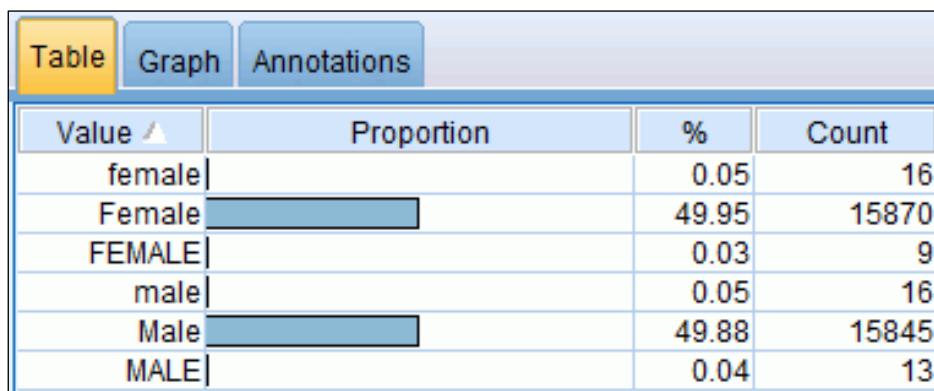
You work for a telecommunications firm and you need to cleanse and enrich a dataset, to build models later.

Task 1. Cleanse data and derive fields for modeling.

- Run the **Distribution** node for gender, downstream from the **Type** node.

Note: You will find the Distribution node when you have opened `demo_deriving_and_reclassifying_fields.str`, located in the **08-Deriving\_and\_Reclassifying\_Fields\Start Files** sub folder..

A section of the results appear as follows:



The screenshot shows a table titled "Distribution" with three tabs: Table, Graph, and Annotations. The Table tab is selected, displaying the following data:

| Value  | Proportion | %     | Count |
|--------|------------|-------|-------|
| female |            | 0.05  | 16    |
| Female |            | 49.95 | 15870 |
| FEMALE |            | 0.03  | 9     |
| male   |            | 0.05  | 16    |
| Male   |            | 49.88 | 15845 |
| MALE   |            | 0.04  | 13    |

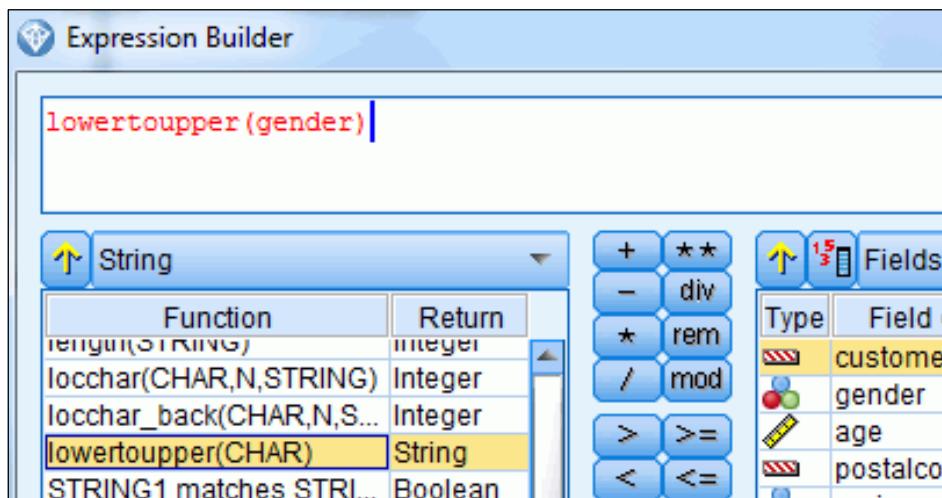
The values of gender are not spelled consistently. This must be repaired, and a new field named `gender_OK` will be derived, with values MALE and FEMALE. You could use the Reclassify node for this purpose, but it is quicker to use the Derive node with an appropriate string function, so you will use the Derive node.

- Close the **Distribution** output window.
- Add a **Derive** node downstream from the **Type** node.

4. Edit the **Derive** node, and then:

- in the **Derive field** text box, type **gender\_OK**
- for **Derive as**, use **Formula** (the default; also leave Field type to its default value. MODELER will autotype the measurement level)
- click the **Launch expression builder**  button to launch the Expression Builder
- select the **String** group of functions
- select the **lowertoupper** function in the list of string functions
- click the **lowertoupper** function to the area where the expression must be built
- select **gender** as the argument for the **lowertoupper** function

A section of the specifications in the Expression Builder dialog box appear as follows:



5. Click **OK** to close the **Expression Builder** window.

6. Click **Preview**.

A section of the results appear as follows:

|   | international_rate | voicemail | SMS | gender_OK |
|---|--------------------|-----------|-----|-----------|
| 1 | 30                 | 10        | 15  | FEMALE    |
| 2 | 30                 | 10        | 15  | MALE      |
| 3 | 30                 | 10        | 15  | MALE      |
| 4 | 30                 | 10        | 15  | MALE      |
| 5 | 30                 | 10        | 15  | MALE      |
| 6 | 30                 | 10        | 15  | FEMALE    |
| 7 | 30                 | 10        | 15  | MALE      |

7. Click **OK** to close **Preview** output window, and then click **OK** to close the **Derive** dialog box.
8. Add a **Distribution** node downstream from the **Derive** node named **gender\_OK**.
9. Edit the **Distribution** node, and then:
  - select **gender\_OK**
  - click **Run**

The results appear as follows:

| Value  | Proportion | %     | Count |
|--------|------------|-------|-------|
| FEMALE |            | 50.03 | 15895 |
| MALE   |            | 49.97 | 15874 |

The values of the new field are okay.

As a note, the Derive node will always create a new field. If you want to replace the values in the existing field gender you can use the Reclassify node, as demonstrated in the next task. You can also use the Filler node; refer to the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for a presentation of the Filler node.

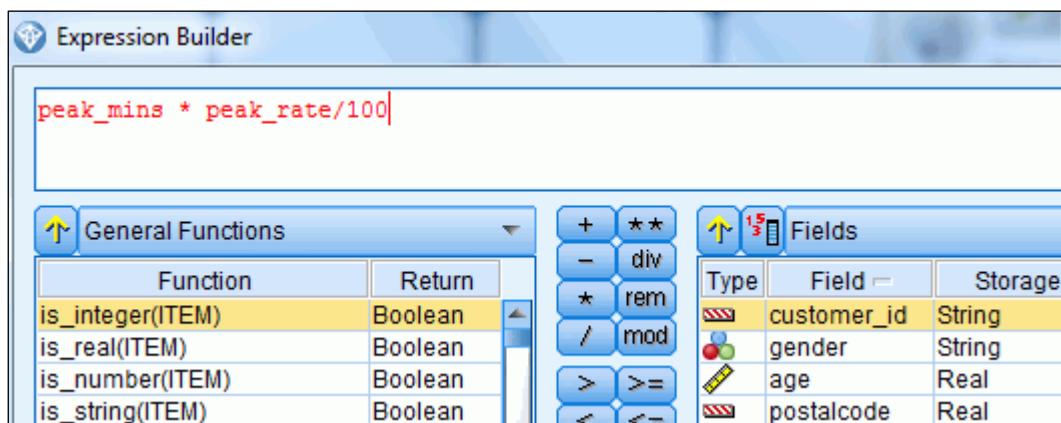
10. Click **OK** to the **Distribution** output window.

You will derive fields giving the bill for calling in the peak hours, and the bill for calling in the off-peak hours (these fields can be valuable predictors for churn).

Each field is computed by multiplying the minutes by the corresponding rate. Rates are given in cents, so you will divide by 100.

11. Add a **Derive** node downstream from the **Derive** node named **gender\_OK**.
12. Edit the **Derive** node, and then:
  - in the **Derive field** text box, type **bill\_peak**
  - for **Derive as**, use **Formula** (the default)
  - launch the **Expression Builder**, and construct the expression **peak\_mins \* peak\_rate/100**

A section of the specifications in the Expression Builder dialog box appear as follows:



13. Click **OK** to close the **Expression Builder** window.
14. Click **Preview**.

A section of the results appear as follows:

|   | bicemail | SMS | gender_OK | bill_peak |
|---|----------|-----|-----------|-----------|
| 1 | 10       | 15  | FEMALE    | 79.380    |
| 2 | 10       | 15  | MALE      | 81.180    |
| 3 | 10       | 15  | MALE      | 61.290    |
| 4 | 10       | 15  | MALE      | 77.940    |
| 5 | 10       | 15  | MALE      | 48.510    |
| 6 | 10       | 15  | FEMALE    | 74.700    |

15. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Derive** dialog box.
16. Add a **Derive** node downstream from the **Derive** node named **bill\_peak**, and derive a field **bill\_offpeak** (being **offpeak\_mins \* offpeak\_rate / 100**) in the same way as it was demonstrated in the previous step.

Now that you have derived these two fields you will sum their values to get the total bill.

17. Add a **Derive** node downstream from the **Derive** node named **bill\_offpeak**.
18. Edit the **Derive** node, and then:
  - in the **Derive field** text box, type **bill\_total**
  - for **Derive as**, use **Formula** (the default)
  - build the expression **bill\_peak + bill\_offpeak**
19. Click **Preview**.

A section of the results appear as follows:

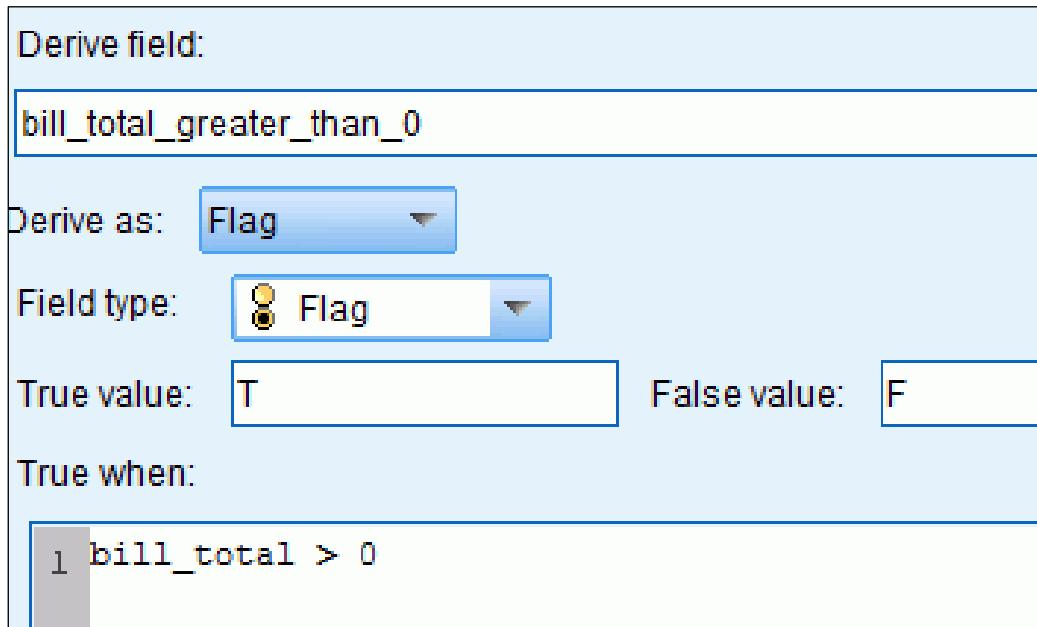
| S | gender_OK | bill_peak | bill_offpeak | bill_total |
|---|-----------|-----------|--------------|------------|
| 5 | FEMALE    | 79.380    | 15.540       | 94.920     |
| 5 | MALE      | 81.180    | 4.110        | 85.290     |
| 5 | MALE      | 61.290    | 4.695        | 65.985     |
| 5 | MALE      | 77.940    | 8.760        | 86.700     |
| 5 | MALE      | 48.510    | 6.105        | 54.615     |
| 5 | FEMALE    | 74.700    | 18.765       | 93.465     |

20. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Derive** dialog box.

Next, you will derive a field that flags whether **bill\_total** is greater than 0.

21. Add a **Derive** node downstream from the **Derive** node named **bill\_total**.
22. Edit the **Derive** node, and then:
  - for **Derive field**, type **bill\_total\_greater\_than\_0**
  - for **Derive as**, select **Flag**
  - for **Field type**, accept **Flag** (autotyped by MODELER, because the Derive type is Flag)
  - for **True when**, type **bill\_total > 0**

A section of the specifications in the Derive dialog box appear as follows:

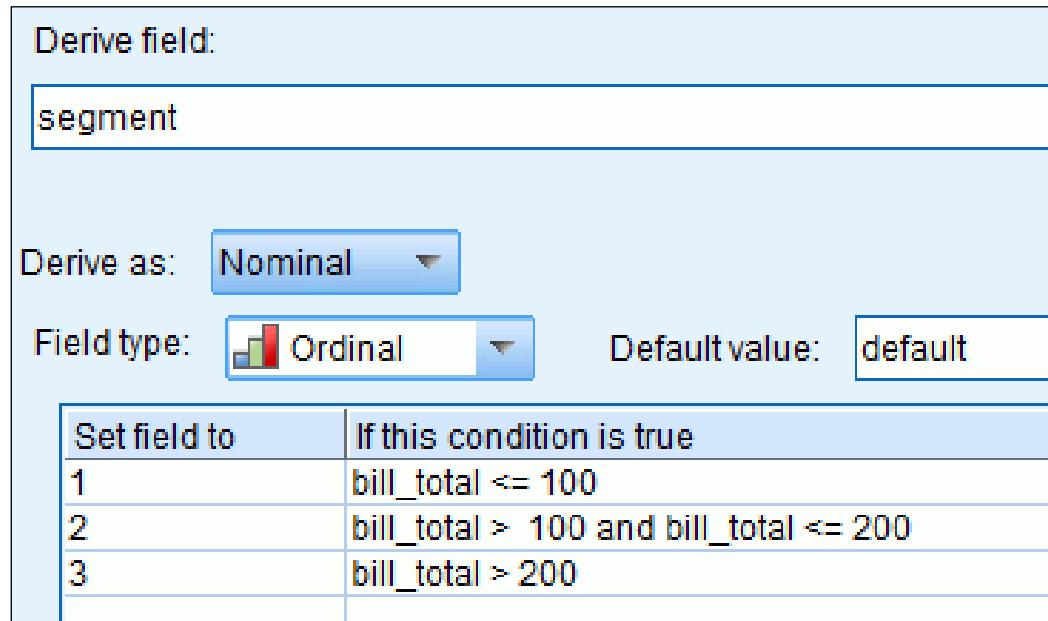


23. Click **OK** to close the **Derive** dialog box.

Based on the field **bill\_total** you will derive a new field named **segment**. This field classifies a customer in one of three categories: 1 ( $\text{bill\_total} \leq 100$ ), 2 ( $\text{bill\_total} > 100$  and  $\text{bill\_total} \leq 200$ ), 3 ( $\text{bill\_total} > 200$ ). Because this is an ordinal field you will set the field's measurement to **Ordinal**.

24. Add a **Derive** node downstream from the **Derive** node named **bill\_total\_greater\_than\_0**.
25. Edit the **Derive** node, and then:
  - for **Derive field**, type **segment**
  - for **Derive as**, select **Nominal**
  - for **Field type**, select **Ordinal**
  - under **Set field to**, type value **1**, under **If this condition is true**, specify **bill\_total <= 100**
  - repeat for the values and expressions: **2 - bill\_total > 100 and bill\_total <= 200; 3 - bill\_total > 200**

A section of the specifications in the **Derive** dialog box appear as follows:



- Click **OK** to close the **Derive** dialog box.

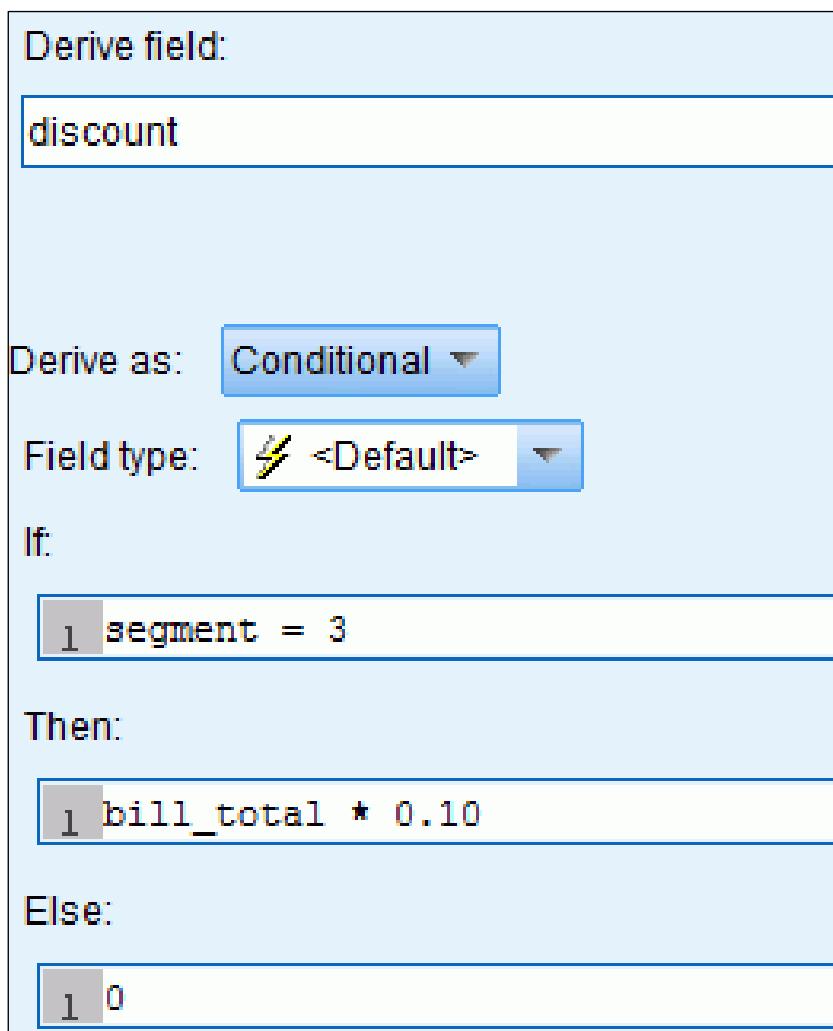
Finally, suppose that customers in segment 3 receive a 10% discount on their total bill, while customers in segments 1 and 2 do not receive a discount. You will derive a field named **discount** conditionally.

- Add a **Derive** node downstream from the **Derive** node named **segment**.

28. Edit the **Derive** node, and then:

- for **Derive field**, type **discount**
- for **Derive as**, select **Conditional**
- for **Field type**, keep the default **Continuous**
- for **If**, type **segment = 3**
- for **Then:**, type **bill\_total \* 0.1**
- for **Else**, type **0**

A section of the specifications in the Derive dialog box appear as follows:



29. Click **Preview**.

A section of the results appear as follows:

| bill_offpeak | bill_total | bill_total_greater_than_0 | segment | discount |
|--------------|------------|---------------------------|---------|----------|
| 15.540       | 94.920 T   |                           | 1       | 0.000    |
| 4.110        | 85.290 T   |                           | 1       | 0.000    |
| 4.695        | 65.985 T   |                           | 1       | 0.000    |
| 8.760        | 86.700 T   |                           | 1       | 0.000    |
| 6.105        | 54.615 T   |                           | 1       | 0.000    |
| 18.765       | 93.465 T   |                           | 1       | 0.000    |

30. Click **OK** to close the **Preview** output window, and then click **OK** to close the **Derive** dialog box.

Leave the stream open for the next task.

## Task 2. Cleanse data and reclassify fields for modeling.

In this task you will repair the different spellings used for gender. In the previous task, this was done by using the Derive node and adding a new field. In this task the Reclassify node will be used which has the advantage that the field's values can be overwritten.

Also, you will derive a new field recoding the handsets into a (smaller) number of brands.

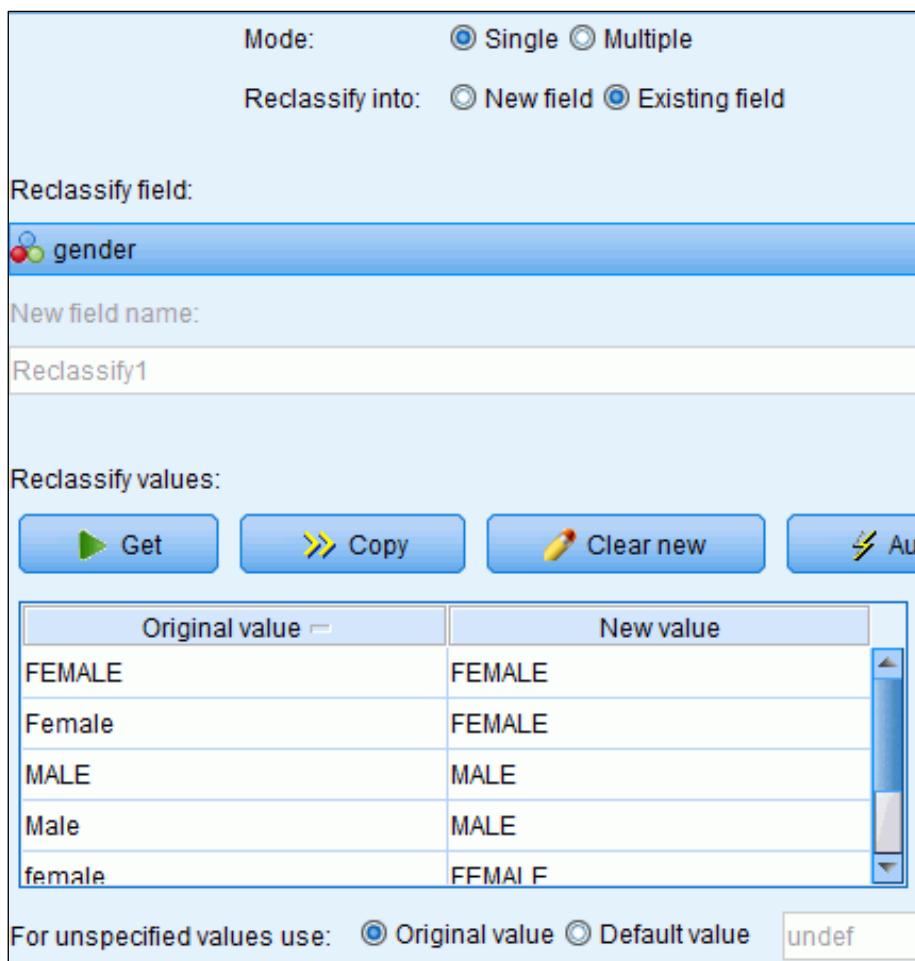
In this task you will build from the stream created in the previous task.

1. Add a **Reclassify** node (Field Ops palette) downstream from the **Derive** node named **discount**.

2. Edit the **Reclassify** node, and then:

- for **Reclassify into**, select **Existing field**
- for **Reclassify field**, select **gender**
- click the **Get** button to populate the **Original value** column
- click the **Copy** button to populate the **New value** column (the values that were originally in upper case are correct now)
- go through the values in the **New value** column, and ensure that the new value is in upper case

A section of the specifications in the Reclassify dialog box appear as follows:



- Click the **Preview** button.

A section of the results appear as follows:

| customer_id | gender | age    | postal_code |
|-------------|--------|--------|-------------|
| K338270     | FEMALE | 20.... | 6599        |
| K342660     | MALE   | 21.... | 1639        |
| K342650     | MALE   | 17.... | 2149        |
| K342640     | MALE   | 20.... | 7788        |
| K342630     | MALE   | 46.... | 5221        |
| K342620     | FEMALE | 24.... | 9141        |
| K342610     | MALE   | 22     | 4221        |

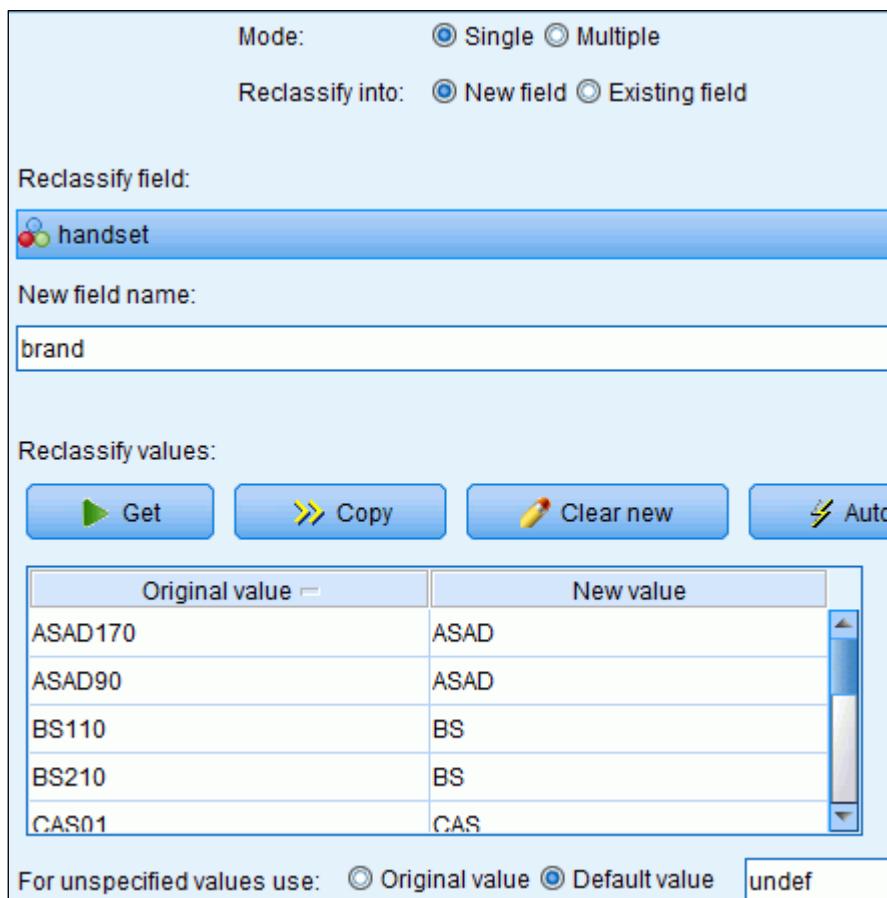
- Click **OK** to close the **Preview** output window, and then click **OK** to close the **Reclassify** dialog box.

Next, you will recode the values of handset. The handsets have a brand name and a type number, for example CAS30. For modeling, only the brand name is relevant.

- Add a **Reclassify** node downstream from the **Reclassify** node named **Reclassify**.

6. Edit the **Reclassify** node, and then:
- for **Reclassify field**, select **handset**
  - for **New field name**, type **brand**
  - click the **Get** button to populate the **Original values** column
  - for the first new brand name, type **ASAD** in the New value column
  - for the second new brand name, type **ASAD**, or pick **ASAD** from the drop-down list
  - repeat for the other brands
  - for values not in the list of Original values, set default value to **undef** (if this Reclassify node would be applied to another dataset, with possibly different handsets, these handsets will be recoded into the \$null\$ value)

A section of the specifications in the Reclassify dialog box appear as follows:

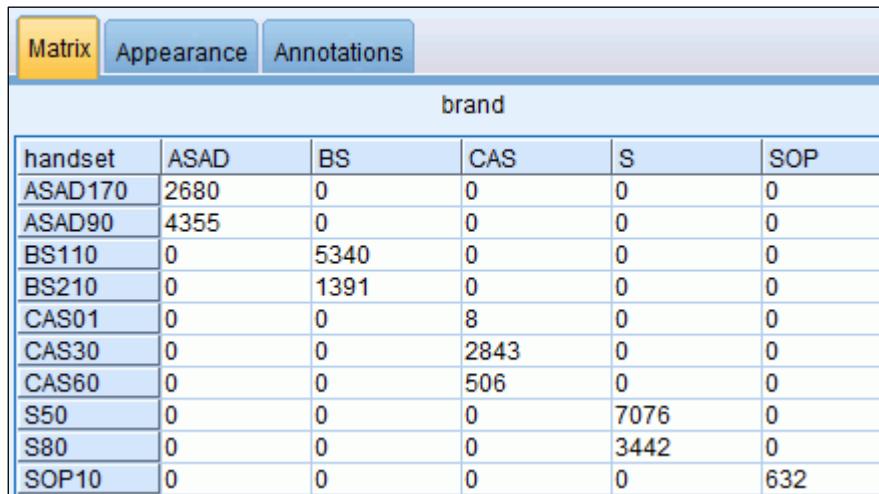


7. Click **OK** to close the **Reclassify** dialog box.

To check the recoding for handsets into brands, you will cross tabulate the two fields. A Matrix node can be used for this purpose.

8. Add a **Matrix** node (Output palette) downstream from the Reclassify node named brand.
9. Edit the **Matrix** node, and then:
  - for **Rows**, select **handset**
  - for **Columns**, select **brand**
  - click **Run**

The results appear as follows:



The screenshot shows a software interface with a tab bar at the top labeled 'Matrix', 'Appearance', and 'Annotations'. The 'Matrix' tab is selected and highlighted in yellow. Below the tabs is a table titled 'brand'. The table has a header row with columns labeled 'handset' and 'ASAD', 'BS', 'CAS', 'S', and 'SOP'. The data rows show counts for various handset models across different brands. For example, ASAD170 has 2680 units in the ASAD brand, while BS110 has 5340 units in the BS brand.

| handset | ASAD | BS   | CAS  | S    | SOP |
|---------|------|------|------|------|-----|
| ASAD170 | 2680 | 0    | 0    | 0    | 0   |
| ASAD90  | 4355 | 0    | 0    | 0    | 0   |
| BS110   | 0    | 5340 | 0    | 0    | 0   |
| BS210   | 0    | 1391 | 0    | 0    | 0   |
| CAS01   | 0    | 0    | 8    | 0    | 0   |
| CAS30   | 0    | 0    | 2843 | 0    | 0   |
| CAS60   | 0    | 0    | 506  | 0    | 0   |
| S50     | 0    | 0    | 0    | 7076 | 0   |
| S80     | 0    | 0    | 0    | 3442 | 0   |
| SOP10   | 0    | 0    | 0    | 0    | 632 |

10. Click **OK** to close the **Matrix** output window.

This completes the demo for this module. You will find the solution results in **demo\_deriving\_and\_reclassifying\_fields\_completed.str**, located in the **08-Deriving\_and\_Reclassifying\_Fields\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Is the following statement true or false? MODELER provides an Expression Builder to help you build CLEM statements

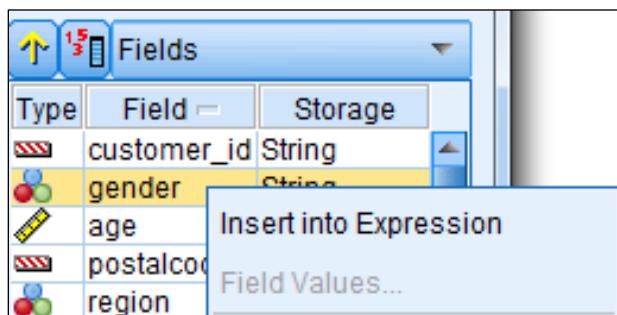
- A. True
- B. False

Question 2: Is the following statement true or false? Logical operators (and, or, not) can be specified in lower case, upper case, or a mix thereof.

- A. True
- B. False

Question 3: Which of the following is the correct statement? Refer to the figure that follows. What is the reason that the option Field Values is greyed out, when the field gender is right-clicked?

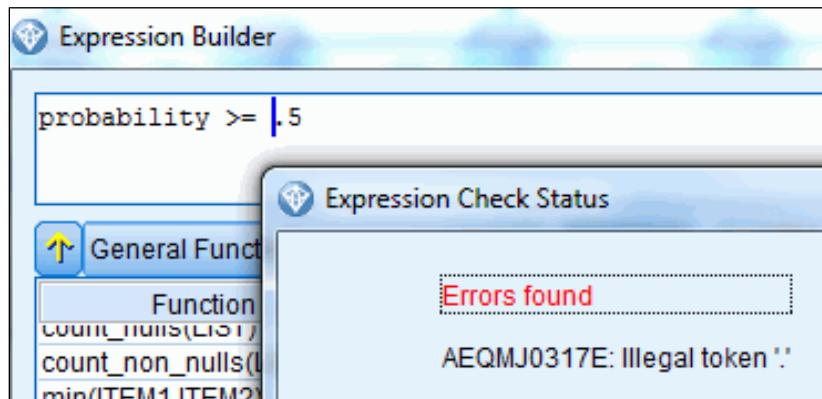
- A. The field is not instantiated.
- B. The option Field Values is only relevant for continuous fields.
- C. Both Insert into Expression and Field Values do the same, so if Insert to Expression can be selected, the option Field Values is greyed out.
- D. None of the answers A through C are correct.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Question 4: When the Check button is clicked in the Expression Builder dialog box, the message depicted in the figure that follows appears. What is the reason for this message?

- A. The probability field does not exist in the data.
- B. The value should be specified as 0.5 instead of .5.
- C. The value .5 should be quoted, so ".5".
- D. The operator  $\geq$  is invalid and should be replaced with  $\geq$ .



Question 5: True or false. The expression AGE\_CATEGORY = "junior" and AGE\_CATEGORY = "senior" in a Select node will select all junior and senior persons.

- A. True
- B. False

Question 6: Which of the following is the correct statement? To select all records with ages between 18 and 65 (including boundary values), the correct expression is:

- A. age  $\geq$  18 and  $\leq$  65
- B. age  $\geq$  18 and age  $\leq$  65
- C. age  $\geq$  18 AND age  $\leq$  65
- D. age  $\geq$  18 or age  $\leq$  65

Question 7: Which of the following is the correct statement?

- A. You can derive multiple fields using a single Derive node.
- B. Within the Derive node, you can choose between creating a new field and overwriting an existing field.
- C. New field names cannot contain blanks.

Question 8: . Refer to the figure that follows. A new field is derived, named ratio, being (income/partner\_income). What is the value for ratio for record 4 and for record 5?

- A. The value for record 4 for ratio equals \$null\$; for record 5 it also equals \$null\$.
- B. The value for record 4 for ratio equals \$null\$; for record 5 it equals 0.
- C. The value for record 4 for ratio equals 0; for record 5 it equals \$null\$.
- D. The value for record 4 for ratio equals 0; for record 5 it equals 0.

| <b>id</b> | <b>income</b> | <b>has_partner</b> | <b>partner_income</b> |
|-----------|---------------|--------------------|-----------------------|
| 1         | 11111         | Y                  | 12345                 |
| 2         | 22222         | Y                  | 23456                 |
| 3         | 33333         | Y                  | 34567                 |
| 4         | 44444         | Y                  | 0                     |
| 5         | 55555         | N                  | \$null\$              |

Question 9: Suppose that a certain dataset has 10 satisfaction fields, all rated on a 5-point scale, and that these 10 fields must be reclassified into "-" (original values: 1 very unsatisfied, 2 unsatisfied), "+/-" (original value: 3 neutral) and "+" (original values: 4 satisfied, 5 very satisfied), where the input fields must be overwritten with the new values.

Which settings will you choose in the Reclassify dialog box?

- A. Mode: Single; Reclassify into: New field.
- B. Mode: Single; Reclassify into: Existing field.
- C. Mode: Multiple; Reclassify into: New field.
- D. Mode: Multiple; Reclassify into: Existing field.

**Answers to questions:**

Answer 1: A. True. Modeler provides an Expression Builder to build CLEM.

Answer 2: B. False. These operators should be specified in lower case.

Answer 3: A. The field is not instantiated and that is why the field's values are not available.

Answer 4: B. Real numbers smaller than 1 in absolute value must have a leading 0.

Answer 5: B. False. It will leave no records. The and operator should be replaced with or to select the juniors and seniors.

Answer 6: B. The operator must be lower case, and the field name age must be repeated.

Answer 7: A. You can derive multiple fields using a single Derive. The Derive node will always create new fields. The Filler node (presented in the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course) can replace a field's values.

Answer 8: A. Both records have an undefined (\$null\$) value.

Answer 9: D. Reclassify multiple fields, and overwrite the fields' values.

Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - use the Control language for Expression Manipulation (CLEM)
  - derive new fields
  - reclassify field values

© 2014 IBM Corporation



In this module you were introduced to two methods to cleanse and enrich your data. You can derive new fields and you reclassify fields.

There are many more options to prepare your data for analyses and modeling. Refer to the *Advanced Data Preparation Using IBM SPSS Modeler (v16)* course for more information.

# Workshop 1

## Deriving and Reclassifying Fields



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **workshop\_deriving\_and\_reclassifying\_fields.str**, located in the **08-Deriving\_and\_Reclassifying\_Fields\Start Files** sub folder.

The following files are used in this workshop:

- **ACME data part 1.dat**: a text data file storing information on ACME's customers and their purchases
- **workshop\_deriving\_and\_reclassifying\_fields.str**: a MODELER stream file, importing the data, setting measurement levels, and instantiating the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Workshop 1: Deriving and Reclassifying Fields

You are working at ACME where you are responsible for data-preparation. It is your job to cleanse the data and to enrich the data with a number of new fields, so that better models can be built later.

- Create a field that gives the difference between AMOUNT\_SPENT and CREDITLIMIT. The difference should return a negative number when AMOUNT\_SPENT exceeds CREDITLIMIT.

Note: Use the **workshop\_deriving\_and\_reclassifying\_fields.str**, located in the **08-Deriving\_and\_Reclassifying\_Fields\Start** sub folder, to import the data.

To check your results: The customer with ID 723.000 (the first record in the dataset) has a credit limit of 9026.000, her amount spent is 546.731, so the difference is 8479.269.

- The fields CREDITLIMIT and AMOUNT\_SPENT are both in euro (a European currency). Also, the field that shows the difference between the two fields (derived in the previous task) is in euro. Create three new fields, being the credit limit, amount spent, and difference in US dollar (1 euro = 1.4 US dollar).

To check your results: The customer with ID 723.000 (the first record in the dataset) has a credit limit of 12636.400 dollar, spent 765.424 dollar, and the difference equals 11870.976 dollar.

- Create a field that flags if the amount spent exceeds the credit limit.

To check your results: There are 2,987 customers who exceeded their credit limit.

- Based on amount spent (in dollars), create a field with three categories:

- bronze: amount\_spent \_dollar up to 2000 dollar (including 2000)
- silver: amount\_spent \_dollar between 2000 dollar and 5000 dollar (excluding 2000, including 5000 dollar).
- gold: amount\_spent \_dollar greater than 5000 dollar

To check your results: There are 26,147 bronze customers, 3,734 silver customers, and 119 gold customers.

- Suppose that gold customers receive a bonus of 5% of their amount spent (in dollars), and bronze and silver customers get no bonus at all. Create this field bonus.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

To check your results: The customer with ID 723.000 (the first record in the dataset) has a bonus of 0.

- The GENDER field shows different spellings. Cleanse the data so that the field's values are Female and Male.

To check your results: Run a Distribution graph for GENDER; you should only have the values Female and Male.

- You have a field ZODIAC, ranging from 1 to 12. Recode this field into a new field named SEASON\_BORN, with 4 categories, defined as follows:

- Winter: zodiac equals 12, 1 or 2
- Spring: zodiac equals 3, 4, or 5
- Summer: zodiac equals 6, 7 or 8
- Autumn: zodiac equals 9, 10 or 11

To check your results: Run a Matrix node (Output palette), with ZODIAC cross tabulated by SEASON\_BORN.

- For the brave: create field ORDERDATE\_YEAR, which returns the year out of the order date.

Hint: use an appropriate Date & Time function.

To check your results: The customer with ID 723.000 (the first record in the dataset) ordered in the year 2009.

For more information about where to work and the workshop results, refer to the Task and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

## Workshop 1: Tasks and Results

Task 1. Compute the difference between amount spent and credit limit.

- Add a **Derive** node downstream from the data source node.
- Edit the **Derive** node, and then:
  - for **Derive field**, type **DIFFERENCE\_CL\_AS**
  - for **Derive type**, keep the default value **Formula**
  - for **Formula**, specify **CREDITLIMIT - AMOUNT\_SPENT**
- Previewing the data shows that the difference for the first record equals 8479.269.

Task 2. Compute fields in dollars from fields in euro.

Notice that the same formula applies to three fields: all fields must be multiplied by 1.4.

- Add a **Derive** node downstream from the **Derive** node named **DIFFERENCE\_CL\_AS**.
- Edit the Derive node, and then:
  - for **Mode**, select **Multiple**
  - for **Derive from**, select **CREDITLIMIT**, **AMOUNT\_SPENT**, and **DIFFERENCE\_CL\_AS**
  - for **Field name extension**, type **\_Dollar**
  - for **Formula**, specify **@FIELD \* 1.4**
- Previewing the data shows that customer with ID 723.000 (the first record in the dataset) has a credit limit of 12636.400 dollar, spent 765.424 dollar, and the difference between the two equals 11870.976 dollar.

### Task 3. Flag if amount spent exceeds the credit limit.

- Add a **Derive** node downstream from the **Derive** node named **\_Dollar**.
- Edit the **Derive** node, and then:
  - for **Derive field**, type **SPENT\_TOO MUCH**
  - for **Derive type**, select **Flag**
  - for **True when**, specify **AMOUNT\_SPENT\_Dollar > CREDITLIMIT\_Dollar** (alternatively, use the DIFFERENCE field created in a previous task)
- Run a **Distribution** graph for **SPENT\_TOO MUCH**. This will show that 2,987 customers exceeded their credit limit.

### Task 4. Create a segment field.

- Add a **Derive** node downstream from the **Derive** node named **SPENT\_TOO MUCH**.
- Edit the **Derive** node, and then:
  - for **Derive field**, type **SEGMENT**
  - for **Derive type**, select **Nominal**
  - specify the values and conditions as follows:

| Set field to | If this condition is true                                  |
|--------------|--|
| Bronze       | AMOUNT_SPENT_Dollar <= 2000                                |
| Silver       | AMOUNT_SPENT_Dollar < 2000 and AMOUNT_SPENT_Dollar <= 5000 |
| Gold         | AMOUNT_SPENT_Dollar > 5000                                 |

- Running a **Distribution** graph for **SEGMENT** shows that there are 26,147 bronze customers, 3,734 silver customers, and 119 gold customers.

## Task 5. Create a field returning the bonus.

- Add a **Derive** node downstream from the **Derive** node named SEGMENT.
- Edit the **Derive** node, and then:
  - for **Derive field**, type **BONUS**
  - for **Derive type**, select **Conditional**
  - for **If**, specify **SEGMENT = "Gold"**
  - for **Then**, specify **AMOUNT\_SPENT\_Dollar \* 0.05**
  - for **Else**, type **0**
- Previewing the data shows that customer with ID 723.000 (the first record in the dataset) does not have a bonus (BONUS equals 0).

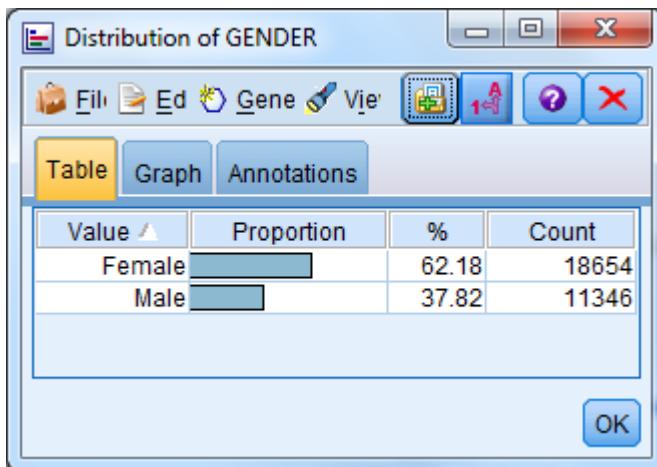
## Task 6. Cleanse the data for gender.

Although the Derive node could be used to cleanse the data for GENDER, here a Reclassify node is preferred, because Reclassify enables you to overwrite the existing field's values, so no new field will be added to the data.

- Add a **Reclassify** node downstream from the **Derive** node named **BONUS**.
- Edit the **Reclassify** node, and then:
  - for **Reclassify into**, select **Existing field**
  - for **Reclassify field**, select **GENDER**
  - click the **GET** button to populate the **Original value** column (if you receive a message that values are not available, instantiate the data upstream from the Reclassify node)
  - for **New values**, specify **Female and Male**, as depicted in the figure below:

| Original value | New value |
|----------------|-----------|
| F              | Female    |
| M              | Male      |
| f              | Female    |
| m              | Male      |

- Run a **Distribution** graph for **GENDER**. This will display the values Female and Male.



## Task 7. Recode zodiac into broader categories.

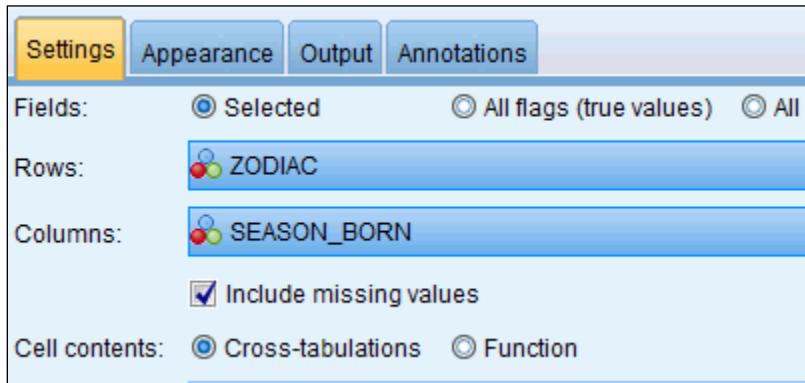
- Add a **Reclassify** node downstream from the **Reclassify** node named **GENDER**.
- Edit the **Reclassify** node, and then:
  - for **Reclassify into**, select **New field**
  - for **New field name**, type **SEASON\_BORN**
  - for **Reclassify field**, select **ZODIAC**
  - click the **GET** button to populate the **Original value** column (if you receive a message that values are not available, instantiate the data upstream from the Reclassify node)
  - for **New values**, specify the values as depicted in the figure that follows:

| Original value | New value |
|----------------|-----------|
| 1              | Winter    |
| 2              | Winter    |
| 3              | Spring    |
| 4              | Spring    |
| 5              | Spring    |
| 6              | Summer    |
| 7              | Summer    |
| 8              | Summer    |
| 9              | Autumn    |
| 10             | Autumn    |
| 11             | Autumn    |
| 12             | Winter    |

- Add a **Matrix** node (Output palette) downstream from the **Reclassify** node named **SEASON\_BORN**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Edit the **Matrix** node, select **ZODIAC** as **Row** field, and then **SEASON\_BORN** as **Column** field.



- Running the **Matrix** node shows that the original values for **ZODIAC** and the new values for **SEASON\_BORN** match.

| SEASON_BORN |        |        |        |        |
|-------------|--------|--------|--------|--------|
| ZODIAC      | Autumn | Spring | Summer | Winter |
| 1.0         | 0      | 0      | 0      | 1335   |
| 10.0        | 2790   | 0      | 0      | 0      |
| 11.0        | 2716   | 0      | 0      | 0      |
| 12.0        | 0      | 0      | 0      | 1306   |
| 2.0         | 0      | 0      | 0      | 2759   |
| 3.0         | 0      | 2751   | 0      | 0      |
| 4.0         | 0      | 2654   | 0      | 0      |
| 5.0         | 0      | 2730   | 0      | 0      |
| 6.0         | 0      | 0      | 2737   | 0      |
| 7.0         | 0      | 0      | 2711   | 0      |
| 8.0         | 0      | 0      | 2764   | 0      |

Note: The stream **workshop\_deriving\_and\_reclassifying\_fields\_completed.str**, located in the **08-Deriving\_and\_Reclassifying\_Fields\Solution Files** sub folder, provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

The advertisement features a white background with a decorative pattern of blue and green hexagons containing yellow cubes, resembling a molecular or data structure. In the top right corner is the classic blue IBM logo. Below the pattern, the text "Looking for Relationships" is displayed in a large, bold, black sans-serif font. Underneath this, "IBM SPSS Modeler (v16)" is written in a smaller, regular black font. At the bottom left, a small rectangular box contains the text "Business Analytics software". On the right side, there is a graphic of three interconnected hexagons, each with a yellow cube inside, connected by teal lines. A small copyright notice, "© 2014 IBM Corporation", is located near the bottom of this graphic.

**Business Analytics software**

IBM SPSS Modeler (v16)

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

# Objectives

- At the end of this module, you should be able to:
  - examine the relationship between two categorical fields
  - examine the relationship between a categorical field and a continuous field
  - examine the relationship between two continuous fields

© 2014 IBM Corporation

Although building powerful models is key in data mining projects, investigating the relationships between the target field (churn, fraud, credit risk, response, and so on) and the predictors can still be helpful in answering the questions that motivated the project. You may find that revenue is directly related to length of time as a customer, or that customers with a certain mobile phone plan are more likely to switch providers. Although these patterns are not substitutes for a full model, they can often be used along with a model.

This module presents methods to examine the relationship between two fields. Before reviewing this module you should be familiar with the following topics:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data
- methods to explore the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

## Methods to Examine the Relationship Between Two Fields

| Relationship between                  | Tabular output | Graphical output |
|---------------------------------------|----------------|------------------|
| Two categorical fields                | Matrix         | Distribution     |
| One categorical, one continuous field | Means          | Histogram        |
| Two continuous fields                 | Statistics     | Plot             |

© 2014 IBM Corporation



The methods used to examine relationships between two fields depend on the measurement level of the fields in question. This goes for exploring the relationship in tabular and graphical format. This slide outlines which node to use.

When you use the dialog boxes, you will notice that MODELER restricts the fields in the field lists to fields of a particular measurement level. For example, you can only select categorical fields in a Distribution node. If a field such as HAS\_CHURNED is typed as continuous because its storage is numeric, you will not be able to run a distribution of the field. Hence, it is of great importance to set the fields' measurement levels correctly.

# Explore Matrix Output

Matrix   Appearance   Annotations

GENDER

| RISK  |          | female | male   | Total  |
|-------|----------|--------|--------|--------|
| bad   | Count    | 457    | 449    | 906    |
|       | Column % | 22.003 | 22.010 | 22.006 |
| good  | Count    | 1620   | 1591   | 3211   |
|       | Column % | 77.997 | 77.990 | 77.994 |
| Total | Count    | 2077   | 2040   | 4117   |
|       | Column % | 100    | 100    | 100    |

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 0, df = 1, probability = 0.996



© 2014 IBM Corporation

To examine the relationship between two categorical fields, cross tabulate the two fields.

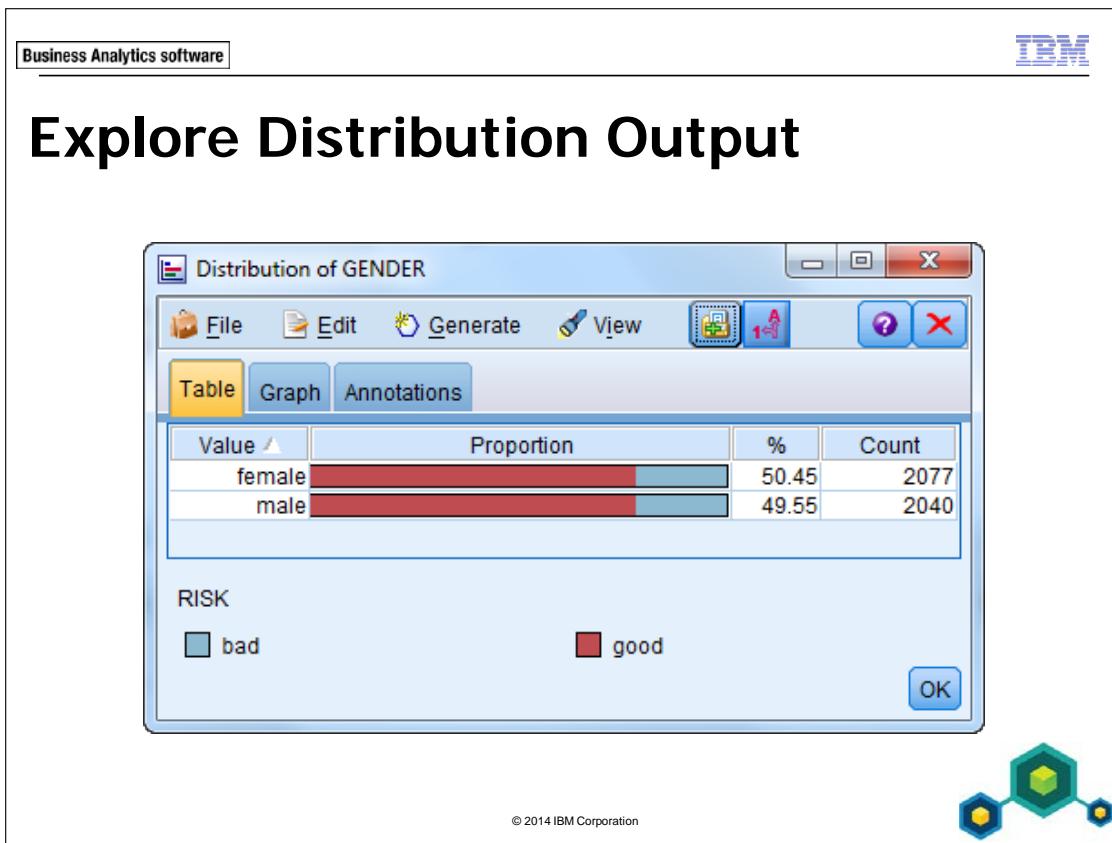
The output on this slide shows the relationship between GENDER and RISK, both categorical fields. Percentages in the table are based on the total column count. For example, 457 of the 2077 women are classified as bad risk, a percentage of  $(457/2077) * 100 = 22.003\%$ . This percentage is 22.010% for men.

Examining the percentages, there appears to be no difference between men and women in percentage bad risk, this is confirmed by a statistical test, the Chi-square test. This test computes the probability that the difference between the percentages of bad risk is caused by the sampling process. This probability ranges between 0 and 1. The probability that the difference between the percentages of bad risk can be attributed to the sampling process in this dataset equals .996, almost 1. Based on this probability you may conclude that the difference that you have observed is only a sample difference, and that men and women in the population, from which this sample was drawn, do not differ with respect to bad risk.

The Matrix node, located in the Output palette, is used to cross tabulate two categorical fields. On the Settings tab, select the row and the column field. Only categorical fields are eligible in the field lists. By default records with missing values on any of the fields will be included in the table (and in the computation of the Chi-square statistic).

On the Appearance tab, request the different statistics such as counts and percentages. Cells with the highest or lowest values in the table can be highlighted by entering the number of cells in the Highlight top/bottom options. This feature is useful when percentages are displayed.

It is recommended to include row and column totals, so that percentages can be interpreted easier.



Instead of showing a table with counts and percentages, you can visualize the relationship in a distribution graph.

In the graph on this slide the bars have the same length, because normalization was applied. Also notice that the categories of GENDER make up the bars, and these categories are overlaid with the values of RISK. In this way you can best compare men and women for their bad risk rates.

The figure confirms the findings for the cross tabulation: men and women appear do behave the same with respect to credit risk.

For a visualization of the relationship between two categorical fields, use the Distribution node, located in the Graphs palette.

On the Plot tab, select the field whose categories will make up the bars. Also select the field to overlay the bars with. Optionally check the Normalize by color option to scale bars so that all bars take up the full width of the graph. This makes comparisons across categories easier.

Notice the option to request a distribution graph for a series of flag fields. This will create one bar for each flag field, with each bar representing the percentage T for the respective field.

Business Analytics software

IBM

## Explore Means Output

**Means** Annotations

Sort by: Field View: Simple

Grouping field: RISK

\*Cells contain: Mean

| Field | bad*   | good*  | Importance           |
|-------|--------|--------|----------------------|
| AGE   | 38.228 | 30.011 | 1.000<br>★ Important |

© 2014 IBM Corporation

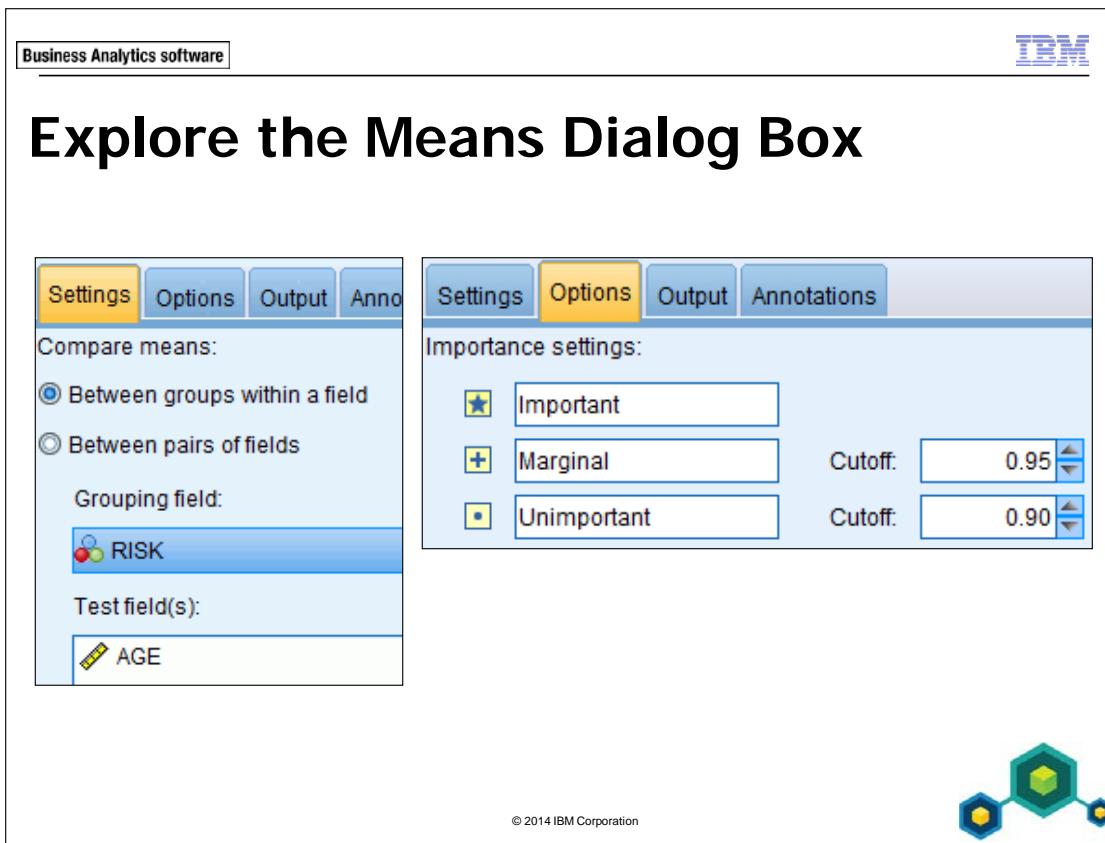


To investigate the relationship between a categorical field and a continuous field, compare the group means.

On this slide, the mean age for the bad credit risk group is 38.228, against 30.011 for the good credit risk group. So, on average, bad risk customers are older than good risk customers.

The difference (38.228 versus 30.011) is labeled important, as shown in the Importance column. Importance equals  $1 - \text{probability}$ , and as in the Chi-square test this is the probability that the sample difference between the means is caused by the sampling process. An importance of 1 means that the probability must have been 0, and you may conclude that the sample difference cannot be attributed to the sampling process. Thus, there must be another reason why you have observed the difference: it is not only in the sample that the groups differ in their mean age, but also in the population.

All in all, you may conclude that there are differences in mean age between the two groups.



You can use the Means node, located in the Output palette, to compare group means. The categorical field that defines the groups is specified under Grouping field, the continuous field for the means must be computed is specified under Test field(s).

The Options tab in the Means dialog box enables you to set threshold probability values used to label results as important, marginal and unimportant (or other labels if you prefer those). By default, importance values below 0.90 are considered Unimportant. Recall that Importance is equal to 1–probability, so an importance less than 0.9 means that the probability (the probability that the difference in means is caused by the sampling process) is greater than 0.1.

Importance values between 0.90 and 0.95 are labeled Marginal, and values greater than 0.95 are labeled Important (an importance greater than .95 means that the probability value must have been smaller than 0.05, which is the most-often used level of significance).

Business Analytics software

IBM

# Explore Histogram Output

© 2014 IBM Corporation



Instead of, or in conjunction with, a table of means, you can present the relationship graphically. A histogram is used when one field is categorical, and the other continuous.

The figure on this slide shows a normalized histogram. Apparently, older people have higher rates of bad risk than youngsters.

The Histogram dialog box (not shown here) is similar to the Distribution dialog box. When you want to normalize the bars, click the Options tab and enable the Normalize by color option (in the Distribution dialog box, this option is not on the Options tab but on the Settings tab).

The screenshot shows the 'Explore Statistics Output' window in IBM SPSS Modeler. The 'Statistics' tab is selected. At the top, there are 'Collapse All' and 'Expand All' buttons. Below is a tree view of correlations. The path selected is 'NUMKIDS > Pearson Correlations > LOANS'. The correlation value is 0.697, which is labeled 'Strong'. The bottom right corner features the IBM logo.

To examine the relationship between two continuous fields, request the Pearson correlation.

The Pearson correlation measures the extent to which two continuous fields are linearly associated, that is the degree to which the relationship between two fields can be described by a straight line.

The correlation coefficient ranges from  $-1$  to  $+1$ , where:  $+1$  represents a perfect positive linear relationship (as one field increases the other field increases at a constant rate),  $-1$  represents a perfect negative relationship (as one field increases the other decreases at a constant rate), and  $0$  represents no linear relationship between the two fields.

In this example, the correlation between NUMKIDS and LOANS equals of  $.697$  and is labeled Strong.

**Business Analytics software**

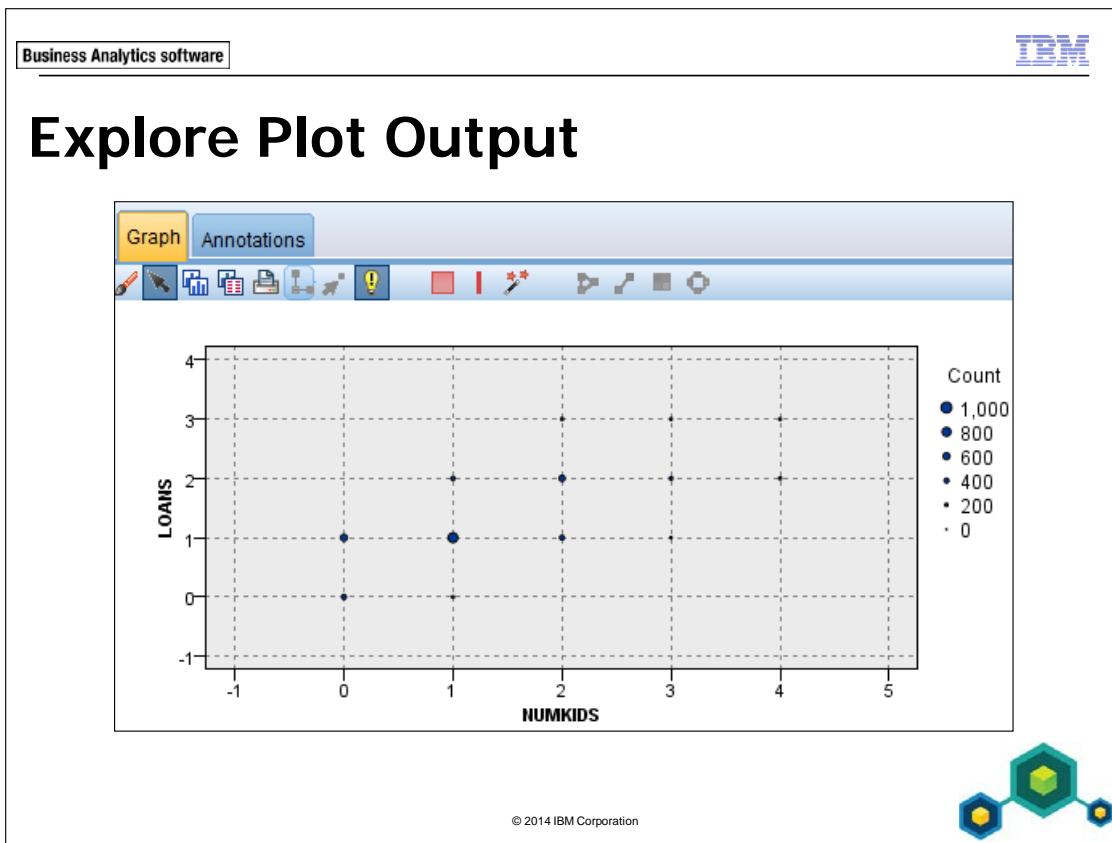
**IBM**

## Explore the Statistics Dialog Box

© 2014 IBM Corporation

The Statistics node, located in the Output palette, computes the correlation between fields specified in the Examine list, and fields in the Correlate list. When you click the Correlations Settings button can change the thresholds for the importance labels, and the labels themselves.

By default, the strength of the correlation is defined by the importance ( $1 - \text{probability}$ ). Even very small correlations can be statistically significant, provided the dataset has enough records. For example, with 100,000 records, a correlation of 0.01 will be statistically significant. An alternative form of assessing the strength is based on the absolute value of the correlation. In general, it is recommended that correlations are labeled based on importance rather than by absolute value since the first decision about a correlation is whether it is significant (important). On the other hand, if your dataset uses thousands, maybe millions of records almost all correlations will be significant and show an importance of 1. So the larger the sample size, the more one should rely on the actual value of the correlation. The smaller the sample size, look first at the importance, then at the correlation.

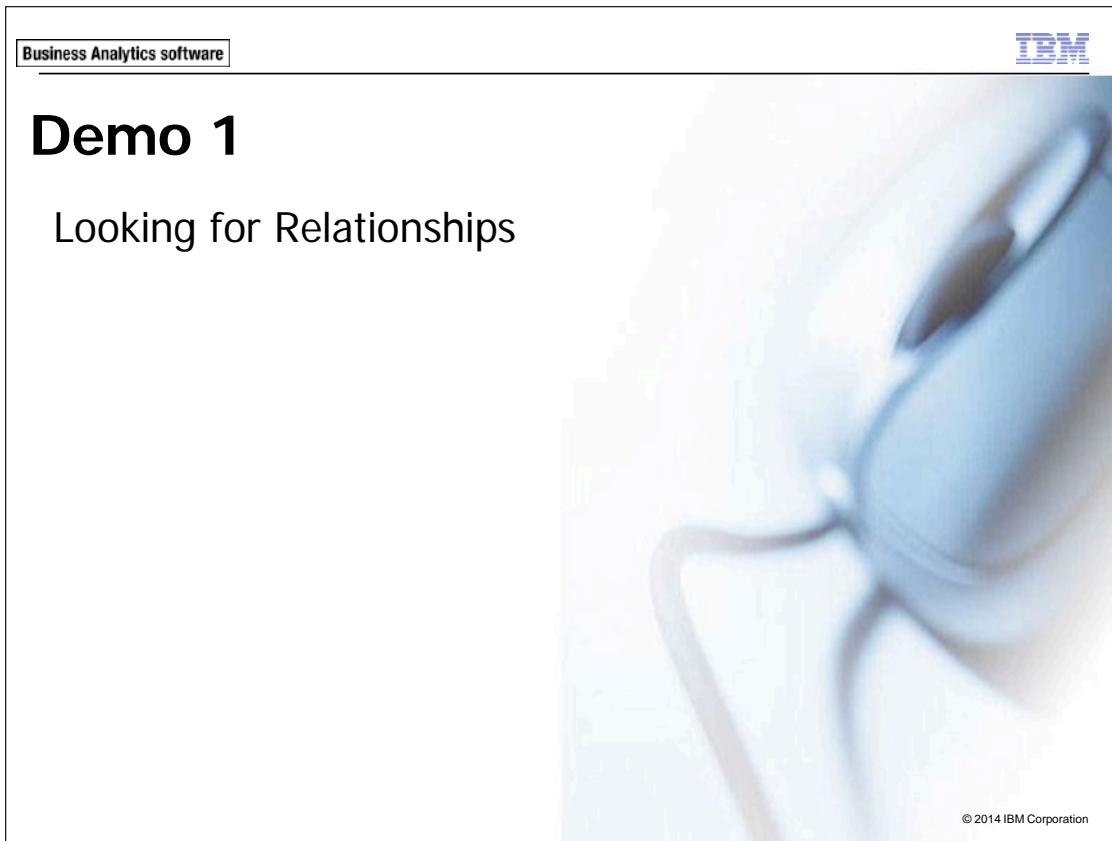


To visualize the relationship between two continuous fields, plot one field against the other.

In general, it is hard to deduce the value of the correlation from this plot. On a previous slide, we noticed that the correlation between NUMKIDS and LOANS was 0.697, and labeled Strong, which you probably cannot tell from this plot.

Having said this, the plot does give an impression of the form of the relationship. Non-linear relationships can be detected in this way (and may explain why the correlation between the two fields is low).

To visualize the relationship between two continuous fields, use the Plot node, located in the Graphs palette. The specifications are straightforward: just specify the X-axis field and Y-axis field.



The slide features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The main title "Demo 1" is at the top left, followed by the subtitle "Looking for Relationships". A large, blurry background image of a person's face is visible. In the bottom right corner of the slide area, there is a small copyright notice: "© 2014 IBM Corporation".

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **demo\_looking\_for\_relationships.str**, located in the **09-Looking\_for\_Relationships\Start Files** sub folder.

The following files are used in this demo:

- **telco x data.txt** a (synthetic) dataset combining various data sources from a (fictitious) telecommunications firm
- **demo\_looking\_for\_relationships.str**: a MODELER stream that imports the data, sets measurement levels, instantiates the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

9-15

## Demo 1: Looking for Relationships

### Purpose:

**Is churn related to handset? Is churn related to the number of dropped calls? And how strong is the relationship between number of gadgets purchased and revenues?**

**You will answer these questions in this demo.**

### Task 1. Assessing the relationship between churn and handset.

In this task we will investigate whether the type of handset is related to churn. Both fields are categorical, so you will use a Matrix node and Distribution.

1. Add a **Matrix** node (Output palette) downstream from the **Type** node.

Note: The Matrix is on the stream canvas after that you have opened **demo\_looking\_for\_relationships.str**, located in the **09-Looking\_for\_Relationships\Start Files** sub folder.

2. Click the **Settings** tab, and then:

- for **Rows**, select **handset**
- for **Columns**, select **churn**

Churn rates must be computed by handset, so that the handsets can be compared. Because the handset field makes up the row of the table, you will request row percentages.

3. Click the **Appearance** tab, and then:

- select the **Percentage of row** option
- select the **Include row and column totals** option

4. Click **Run**.

A section of the result appear as follows:

| Matrix   | Appearance | Annotations |         |       |
|--|------------|-------------|---------|-------|
| churn  |            |             |         |       |
| handset  |            | Active      | Churned | Total |
| ASAD170  | Count      | 2556        | 124     | 2680  |
|  | Row %      | 95.373      | 4.627   | 100   |
| ASAD90   | Count      | 224         | 4131    | 4355  |
|  | Row %      | 5.144       | 94.856  | 100   |
| BS110  | Count      | 2991        | 2349    | 5340  |
|  | Row %      | 56.011      | 43.989  | 100   |
| BS210  | Count      | 989         | 402     | 1391  |
|  | Row %      | 71.100      | 28.900  | 100   |
| CAS01  | Count      | 3           | 5       | 8     |
|  | Row %      | 37.500      | 62.500  | 100   |
| CAS30  | Count      | 153         | 2690    | 2843  |
|  | Row %      | 5.382       | 94.618  | 100   |
| mean   | count      | 170         | 20      | 190   |
| Cells contain: cross-tabulation of fields (including missing values) |            |             |         |       |
| Chi-square = 13,941.022, df = 11, probability = 0                    |            |             |         |       |

For handset ASAD170, the churn rate is 4.627%, while it is 94.856 for those with handset ASAD90 churn. So, there are huge differences between the handsets in churn rates. The Chi-square statistic also points to a significant relationship between handset and churn (the probability is 0).

This is not to say, that all handsets differ in churn rate. For example, ASAD90 and CAS30 both have a churn rate of about 95%. So, although the Chi-square test tells you that there are differences in churn rates, it does not tell you which handsets differ in churn rate. Further analyses, or a model (in this case a CHAID model) would be needed to investigate this.

5. Click **OK** to close the **Matrix** output window.

You will request a Distribution graph to support your findings.

6. Add a **Distribution** node (Graphs palette) downstream from the **Type** node.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

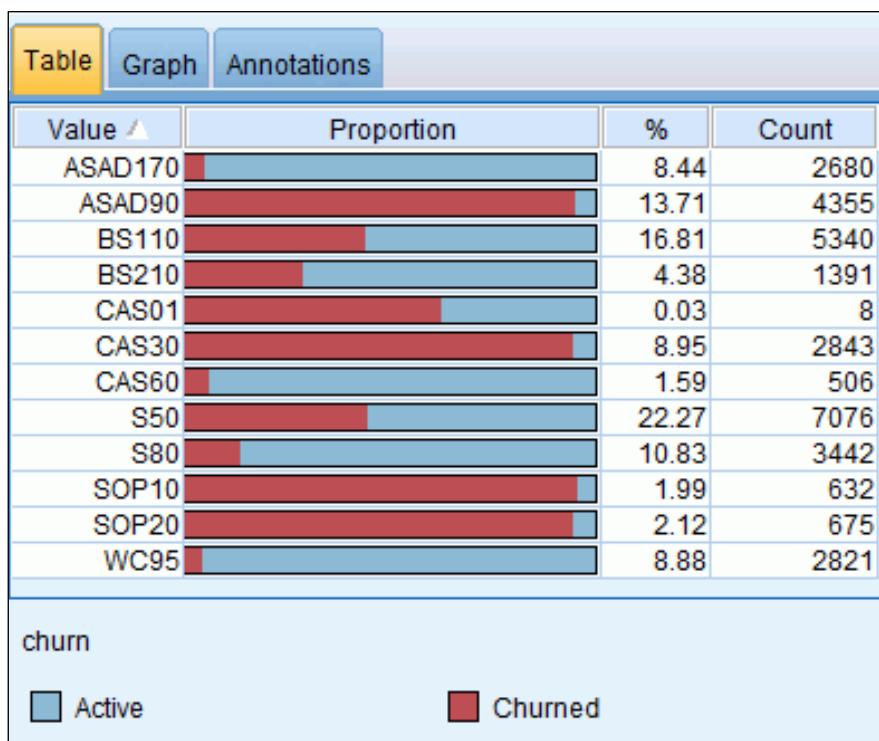
© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

7. Edit the **Distribution** node, and then:

- click the **Plot** tab
- for **Field**, select **handset**
- for **Color**, select **churn**
- select the **Normalize by colors** option (so that churn rates can be compared better)
- click **Run**

A section of the results appear as follows:



The Distribution graph shows differences in churn rates. It also shows that some handsets (ASAD90, CAS30, SOP10, SOP20) have similar high churn rates. So customers with these handsets are at risk for churning.

8. Close the **Distribution** output window.

Leave the stream open for the next task.

## Task 2. Assessing the relationship between churn and number of dropped calls.

In this task you will investigate if the number of dropped calls is related to churn. The first field is continuous, the second field is categorical, so you will use a Means node and a Histogram node.

In this task, you will build from the stream in the previous task.

1. Add a **Means** node (Output palette) downstream from the **Type** node.
2. Edit the **Means** node, and then:
  - for **Grouping** field, select **churn**
  - for **Test field(s)**, select **dropped\_calls**
  - click **Run**

A section of the results appear as follows:

| Field         | Active* | Churned* | Importance         |
|---------------|---------|----------|--------------------|
| dropped_calls | 2.573   | 3.906    | 1.000<br>Important |

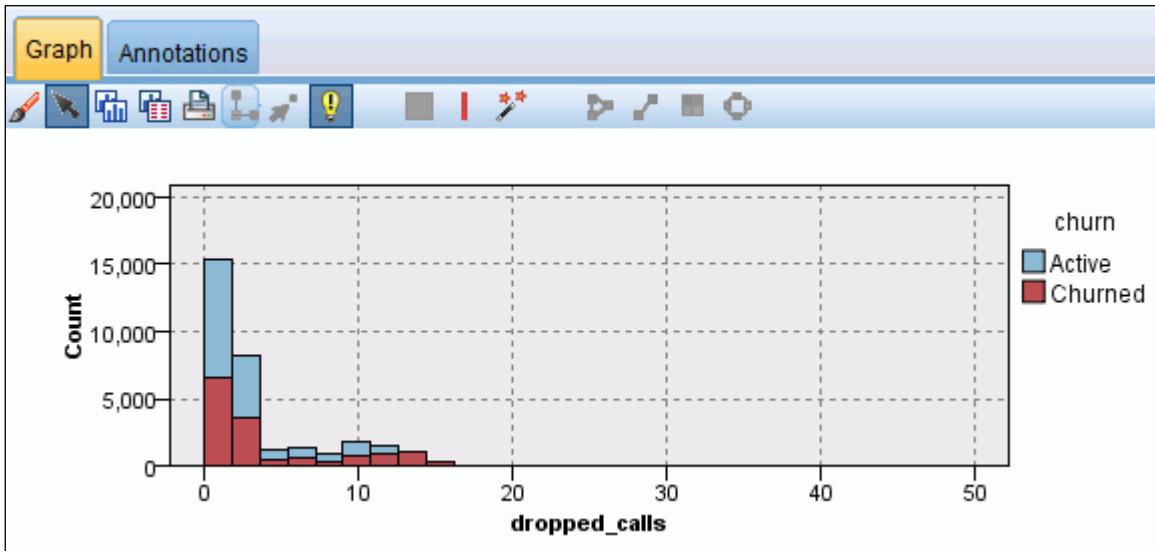
Customers that churned have 3.9 dropped calls on average, while active customers have 2.5 dropped calls on average. So, the number of dropped calls is higher for churners than for active customers. This difference is labeled Important (statistically significant).

To visualize the relationship, you will request a Histogram.

3. Click **OK** to close the **Means** output window.
4. Add a **Histogram** node downstream from the **Type** node.

5. Click the **Plot** tab, and then:
  - for **Field**, select **dropped\_calls**
  - for **Color**, select **churn**
6. Click **Run**.

A section of the results appear as follows:



The histogram confirms that higher values for dropped calls go along with higher churn rates.

7. Click **OK** to close the Histogram output window.

Leave the stream open for the next task.

### Task 3. Assessing the relationship between number of products and revenues.

In this task you will explore the relationship between two continuous fields, number of gadgets and revenues.

In this task, you will build from the stream in the previous task.

1. Add a **Statistics** node (Output palette) downstream from the **Type** node.
2. Edit the **Statistics** node, and then:
  - for **Examine**, select **revenues**
  - for **Correlate**, select **number\_of\_gadgets**

3. Click **Run**.

A section of the results appear as follows:

The screenshot shows the 'Statistics' tab of the IBM SPSS Modeler interface. At the top, there are tabs for 'Statistics' (selected) and 'Annotations'. Below them are buttons for 'Collapse All' and 'Expand All'. The main area displays a hierarchical tree structure under the 'revenues' node. Under 'Statistics', there is a table of descriptive statistics:

|                        |           |
|------------------------|-----------|
| Count                  | 30590     |
| Mean                   | 311.628   |
| Min                    | 5         |
| Max                    | 796       |
| Range                  | 791       |
| Variance               | 23927.646 |
| Standard Deviation     | 154.686   |
| Standard Error of Mean | 0.884     |

Under 'Pearson Correlations', there is a row for 'number\_of\_gadgets' with a correlation value of 0.809 and the label 'Strong'.

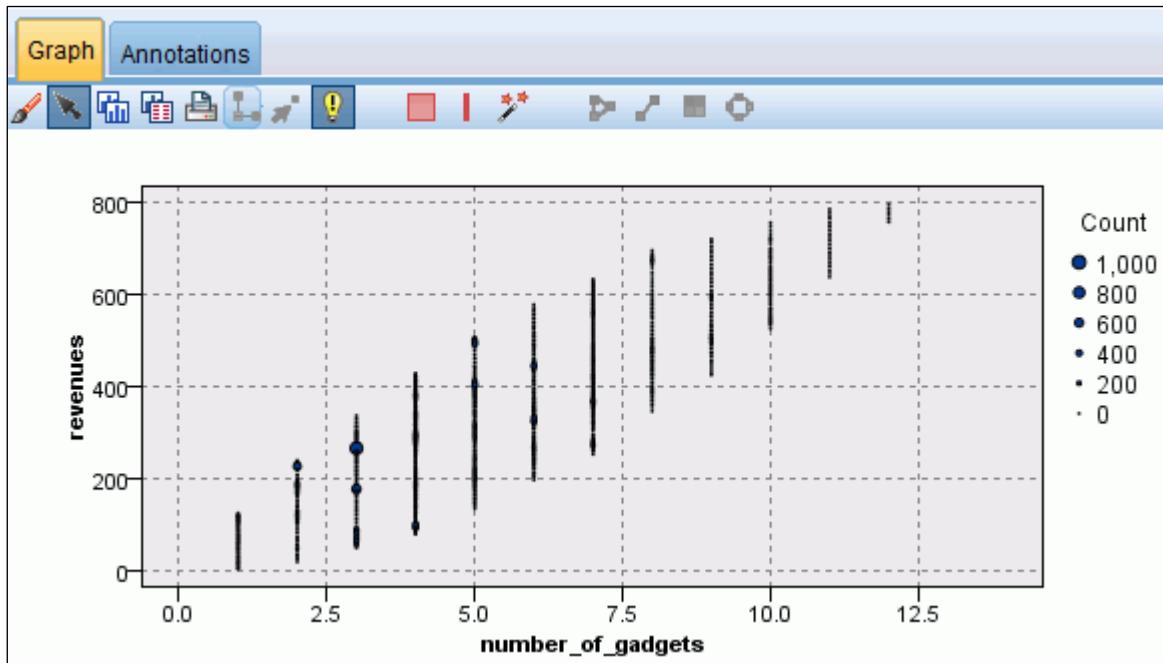
The correlation indicates a strong relationship (meaning significant, because the default correlation settings were used) between the two fields. This may come as no surprise, because number of gadgets is the number of products the customer purchased, and revenues is the total price he or she paid for it.

4. Add a **Plot** node (Graphs palette) downstream from the **Type** node.

5. Edit the **Plot** node, and then:

- for **X field**, select **number\_of\_gadgets**
- for **Y field**, select **revenues**

A section of the results appear as follows:



The plot shows a nice linear trend.

6. Click **OK** to close the **Plot** output window.

This completes the demo for this module. You will find the solution results in **demo\_looking\_for\_relationships\_completed.str**, located in the **09-Looking\_for\_Relationships\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Which of the following statements are correct? Refer to the figure that follows.

- A. To get better insight into the relationship between MARITAL and CREDIT RISK, row percentages must be included, not column percentages.
- B. To get better insight into the relationship between MARITAL and CREDIT RISK, column percentages must be included, not row percentages.
- C. The Chi-square statistic indicates that there is a statistically significant relationship between MARITAL and CREDIT RISK.
- D. The Chi-square statistic will not change if MARITAL is placed in the row, and CREDIT RISK in the column.

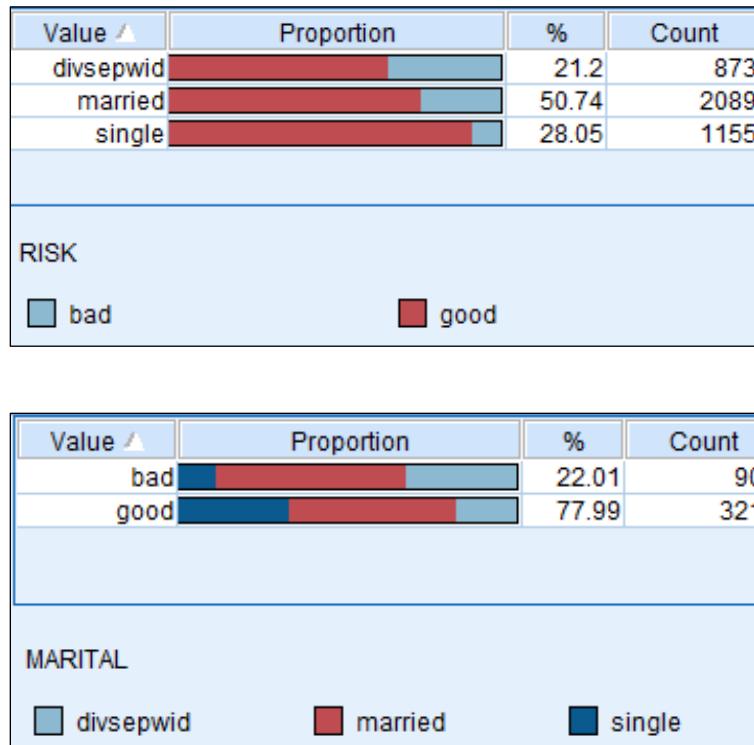
|       |           | MARITAL |        |       |  |
|-------|-----------|---------|--------|-------|--|
| RISK  | divsepwid | married | single | Total |  |
| bad   | 297       | 507     | 102    | 906   |  |
| good  | 576       | 1582    | 1053   | 3211  |  |
| Total | 873       | 2089    | 1155   | 4117  |  |

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 196.467, df = 2, probability = 0

Question 2: Is the following statement true or false? With huge datasets (millions of records), the Chi-square statistic can be statistically significant, although there are hardly differences in percentages.

- A. True
- B. False

Question 3: Refer to the figure that follows. Distribution \_\_\_ (A/B) is the best way to examine if credit risk is different for the various marital groups



Question 4: Is the following statement true or false? The importance measure which uses the probability is derived from the probability by: Importance = 100 \* probability.

- A. True
- B. False

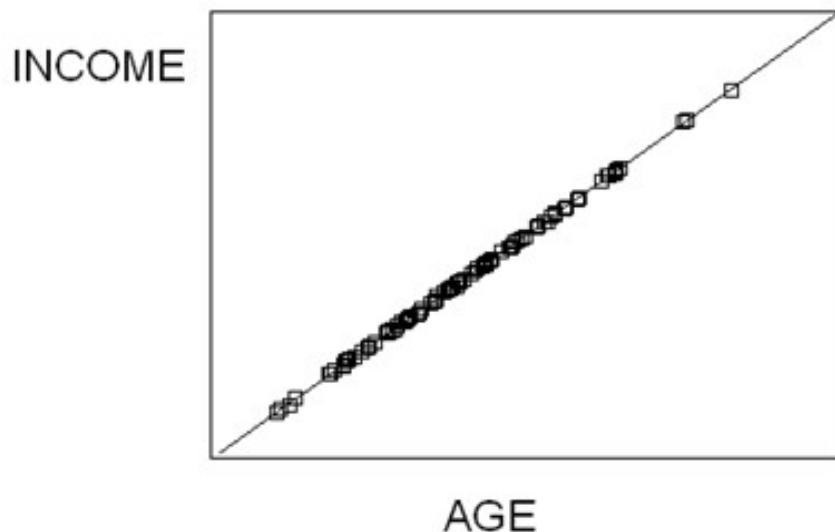
Question 5: Which of the following is the correct statement? Refer to the figure that follows.

- A. The probability value equals 0.879.
- B. There is no statistically significant difference between men and women in mean INCOME.
- C. There is a statistically significant difference between men and women in mean INCOME.
- D. If there would 1000 times as many men, and 1000 times as many women, the Importance value would not change.

| Means                  | Annotations |           |  |
|------------------------|-------------|-----------|--|
| Sort by:               | Field       | View:     | Simple   |
| Grouping field: GENDER |             |           |  |
| *Cells contain: Mean   |             |           |  |
| Field                  | female*     | male*     | Importance   |
| INCOME                 | 25370.144   | 25794.090 | 0.879<br><input checked="" type="checkbox"/> Unimportant |

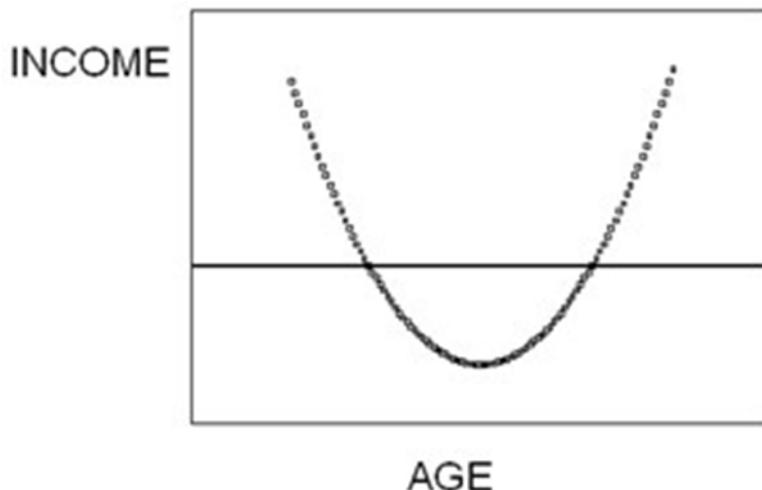
Question 6: Refer to the figure that follows, displaying a plot of AGE against INCOME, with the best fitting straight line superimposed. True or false: the correlation between AGE and INCOME is +1.

- A. True
- B. False



Question 7: Which of the following statements are correct? Refer to the figure below, which displays the plot of AGE against INCOME, with the best fitting straight line (the horizontal line) superimposed.

- A. The correlation between AGE and INCOME is +1.
- B. The correlation between AGE and INCOME is 0.
- C. The correlation between AGE and INCOME is -1.
- D. There is a perfect relationship between AGE and INCOME.



## Answers to questions:

Answer 1: B, C, D.

Answer 2: A. True. The result of a statistical test is dependent on the sample size.

Answer 3: A. Use the dependent field as overlay field (and credit risk is dependent on marital status, not the other way around).

Answer 4: B. False. Importance equals  $100 - \text{probability}$ .

Answer 5: B.

Answer 6: A. True. The figure depicts a perfect positive (the higher age, the higher income) linear relationship, so the correlation is +1.

Answer 7: B, D. The relationship is perfect, but the relationship is not-linear. Because the best fitting straight line is horizontal, the correlation equals 0.

## Summary

- At the end of this module, you should be able to:
  - examine the relationship between two categorical fields
  - examine the relationship between a categorical field and a continuous field
  - examine the relationship between two continuous fields

© 2014 IBM Corporation

This module presented methods to investigate the relationship between two fields. The measurement level of the fields involved determines which nodes you will use.

Business Analytics software

IBM

# Workshop 1

## Looking for relationships



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **workshop\_looking\_for\_relationships.str**, located in the **09-Looking\_for\_Relationships\Start Files** sub folder.

The following (synthetic) files are used in this workshop:

- **ACME analysis data.sav**: an IBM SPSS Statistics data file storing data for a (fictitious) company named ACME
- **workshop\_looking\_for\_relationships.str**: a stream that imports the, sets the fields' measurement levels, instantiates the data

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

## Workshop 1: Looking for Relationships

You are working for ACME, a company that sells sports products. Before you start building models you want to examine the relationships in the data, especially which fields are related to response.

- Examine the relationship between `response_to_test_mailing` and `gender`, both in tabular and graphical output.

Note: the file `workshop_looking_for_relationships.str` imports the data for you.

What is the percentage of women that has responded to the test mailing, and what is this percentage for men?

Is the association between the two fields statistically significant?

Does the graph support this conclusion?

- Which relationship is strongest in terms of statistical significance?
  - between `monetary_value` and `response_to_test_mailing`
  - between `frequency` and `response_to_test_mailing`
  - between `recency` and `response_to_test_mailing`.
- Examine the relationship between `response_to_test_mailing` and `creditlimit`, both in tabular and graphical output .

What is the mean credit limit for those that responded positively to the test mailing, and what is the mean credit limit for those that did not respond positive to the test mailing?

For more information about where to work and the workshop results, refer to the Task and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

## Workshop 1: Tasks and Results

Task 1. Examine the relationship between response and gender.

Both fields are categorical, so use a Matrix node for tabular output, and a Distribution node for graphical output.

- Add a **Matrix** node to the **Type** node.
- Edit the **Matrix** node, and then:
  - in the **Settings** tab, for **Rows**, select **gender**, for **Columns** select **response\_to\_test\_mailing**
  - in the **Appearance** tab, select **Percentage of row** (we want to compare men with women, and these categories make up the rows of the table, per specification in the Settings tab); also select **Include row and column totals** (so the percentages are easier to explain)
- Run the **Matrix** node.

A section of the results appear as follows:

The screenshot shows a Matrix node output window. At the top, there are three tabs: Matrix (selected), Appearance, and Annotations. Below the tabs is a title "response\_to\_test\_mailing". The main area is a table with the following data:

| gender |       | F      | T     | Total |
|--------|-------|--------|-------|-------|
| female | Count | 5958   | 181   | 6139  |
| male   | Count | 3681   | 180   | 3861  |
|        | Row % | 97.052 | 2.948 | 100   |
|        | Row % | 95.338 | 4.662 | 100   |
| Total  | Count | 9639   | 361   | 10000 |
|        | Row % | 96.390 | 3.610 | 100   |

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 20.003, df = 1, probability = 0

About 2.9% of the 6,139 women responded positive to the mailing, versus 4.6% of the men. Although the difference in percentages is small, it is statistically significant: the probability of the difference is caused by the sampling process

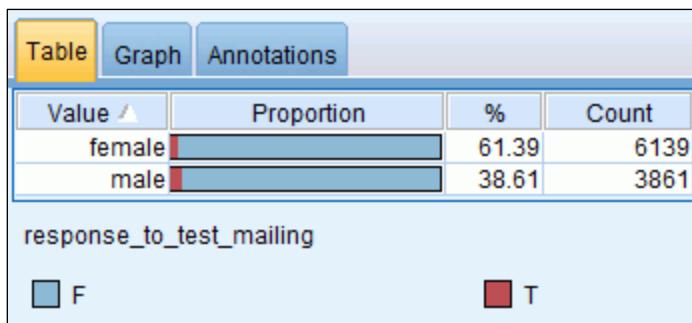
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

equals 0, so the conclusion is that there must be a difference in response rate between the groups in the population.

The fact that such a small difference is statistically significant is on the account of the sample size (10,000).

- Add a **Distribution** graph downstream from the **Type** node.
- Edit the **Distribution** node, and then:
  - for **Field**, select **gender**
  - for **Color**, select **response\_to\_test\_mailing**
  - enable the **Normalize by color option**
- Run the **Distribution** node.

A section of the results appear as follows:



The graph shows that the difference in response rate is small.

## Task 2. Examine the relationship between response and RFM fields.

Experience learns, that important fields for predicting response (or churn) are recency (how long ago did the customer purchase a product from the company), monetary value (the total amount of all purchases made by the customer), and frequency (how many times did the customer purchase a product). These fields are known as RFM fields.

In this dataset, the RFM fields are categorical, so again a Matrix node is needed to assess their relationship with response to the test mailing.

The Matrix node only enables you to specify one row field, and one column field, so to examine the relationship of the RFM fields with response, three Matrix nodes are needed.

- Add a **Matrix** node to the **Type** node.
- Edit the **Matrix** node, and then:
  - in the **Settings** tab, for **Rows**, select **monetary\_value**, for **Columns** select **response\_to\_test\_mailing**
  - in the **Appearance** tab, select **Percentage of row** (we want to compare the categories of monetary value, for example if those who are in the low category have a higher response rate than those in the category high).
- Click **Run**.

A section of the results appear as follows:

| Matrix                   | Appearance | Annotations |       |       |  |
|--------------------------|------------|-------------|-------|-------|--|
| response_to_test_mailing |            |             |       |       |  |
| monetary_value           |            | F           | T     | Total |  |
| 1 low                    | Count      | 3340        | 14    | 3354  |  |
|                          | Row %      | 99.583      | 0.417 | 100   |  |
| 2 medium                 | Count      | 3194        | 42    | 3236  |  |
|                          | Row %      | 98.702      | 1.298 | 100   |  |
| 3 high                   | Count      | 3105        | 305   | 3410  |  |
|                          | Row %      | 91.056      | 8.944 | 100   |  |
| Total                    | Count      | 9639        | 361   | 10000 |  |
|                          | Row %      | 96.390      | 3.610 | 100   |  |

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 426.807, df = 2, probability = 0

About 0.4% of the 3,354 low category customers responded to the mailing, while this percentage is 8.9 for the high category. The relationship is statistically significant (probability equals 0). Although the test does not tell you where the differences lie (for example, do the response percentages for the low and medium category differ), it tells you that monetary value is an important field, and in modeling should be included (a model such as CHAID will tell you where the differences between the categories are).

- Repeat the steps for the **frequency** field.

A section of the results appear as follows:

| Matrix                   | Appearance | Annotations |       |       |
|--------------------------|------------|-------------|-------|-------|
| response_to_test_mailing |            |             |       |       |
| frequency                |            | F           | T     | Total |
| 1 low                    | Count      | 2864        | 66    | 2930  |
|                          | Row %      | 97.747      | 2.253 | 100   |
| 2 medi                   | Count      | 2978        | 170   | 3148  |
|                          | Row %      | 94.600      | 5.400 | 100   |
| 3 high                   | Count      | 3797        | 125   | 3922  |
|                          | Row %      | 96.813      | 3.187 | 100   |
| Total                    | Count      | 9639        | 361   | 10000 |
|                          | Row %      | 96.390      | 3.610 | 100   |

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 46.526, df = 2, probability = 0

Again, the test tells you that there is a relationship between frequency and response, although the differences between the categories are smaller than for monetary value (the probability is rounded to 0, so you cannot tell that from the probability value; however, the Chi-square value is less than the Chi-square in the previous table).

- Repeat the steps for the **recency** field.

A section of the results appear as follows:

| Matrix                   | Appearance | Annotations |       |       |
|--------------------------|------------|-------------|-------|-------|
| response_to_test_mailing |            |             |       |       |
| frequency                |            | F           | T     | Total |
| 1 low                    | Count      | 2864        | 66    | 2930  |
|                          | Row %      | 97.747      | 2.253 | 100   |
| 2 medi                   | Count      | 2978        | 170   | 3148  |
|                          | Row %      | 94.600      | 5.400 | 100   |
| 3 high                   | Count      | 3797        | 125   | 3922  |
|                          | Row %      | 96.813      | 3.187 | 100   |
| Total                    | Count      | 9639        | 361   | 10000 |
|                          | Row %      | 96.390      | 3.610 | 100   |

Cells contain: cross-tabulation of fields (including missing values)  
Chi-square = 46.526, df = 2, probability = 0

Again, the relationship is statistically significant, with the high category having the highest response rate.

All in all, these fields are candidates for including them in model to predict response. A model, such as CHAID, will, based on statistical criteria, group the categories that have a similar response, and also a model will show the interaction between the predictors.

As a note, the RFM fields in this dataset were recorded before the test mailing went out. These RFM fields are not contaminated by the results of the test mailing itself. If they would include the results of the test mailing, they would no longer be predictors.

### Task 3. Examine the relationship between response and credit limit.

In this case, use the Means and the Histogram node, because it concerns the relationship between a categorical and a continuous field.

- Add a **Means** node to the **Type** node.
- Edit the **Means** node, and then:
  - for **Grouping field**, select **response\_to\_test\_mailing**
  - for **Test field(s)**, select **creditlimit**
- Click **Run**.

A section of the results appear as follows:

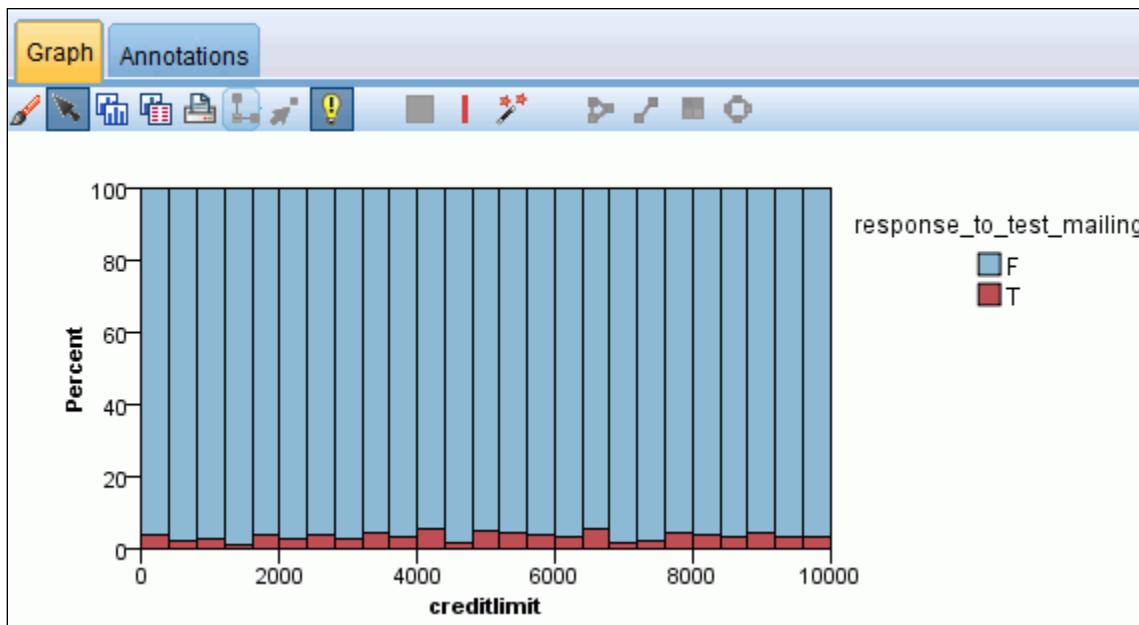
| Means                                    | Annotations |          |  |
|--|-------------|----------|--|
| Sort by: Field ▾ View: Simple ▾          |             |          |  |
| Grouping field: response_to_test_mailing |             |          |  |
| *Cells contain: Mean                     |             |          |  |
| Field                                    | F*          | T*       | Importance   |
| creditlimit                              | 4670.315    | 4898.460 | 0.853<br><input checked="" type="checkbox"/> Unimportant |

The credit limit for those that did not respond to the test mailing equals 4670.3, while the credit limit is 4898.4 for those that did respond to the test mailing. So, it seems like those with higher credit limits are more inclined to respond to the test mailing. This difference, however, is unimportant from a statistical point of view, because it is labeled as unimportant.

A graph enables you to examine the relationship closer.

- Add a **Histogram** node downstream from the **Type** node.
- Edit the **Histogram** node, and then:
  - for **Field**, select **creditlimit**
  - for **Color**, select **response\_to\_test\_mailing**
  - click the **Options** tab, and enable the **Normalize by color** option
- Run the **Histogram** node.

A section of the results appear as follows:



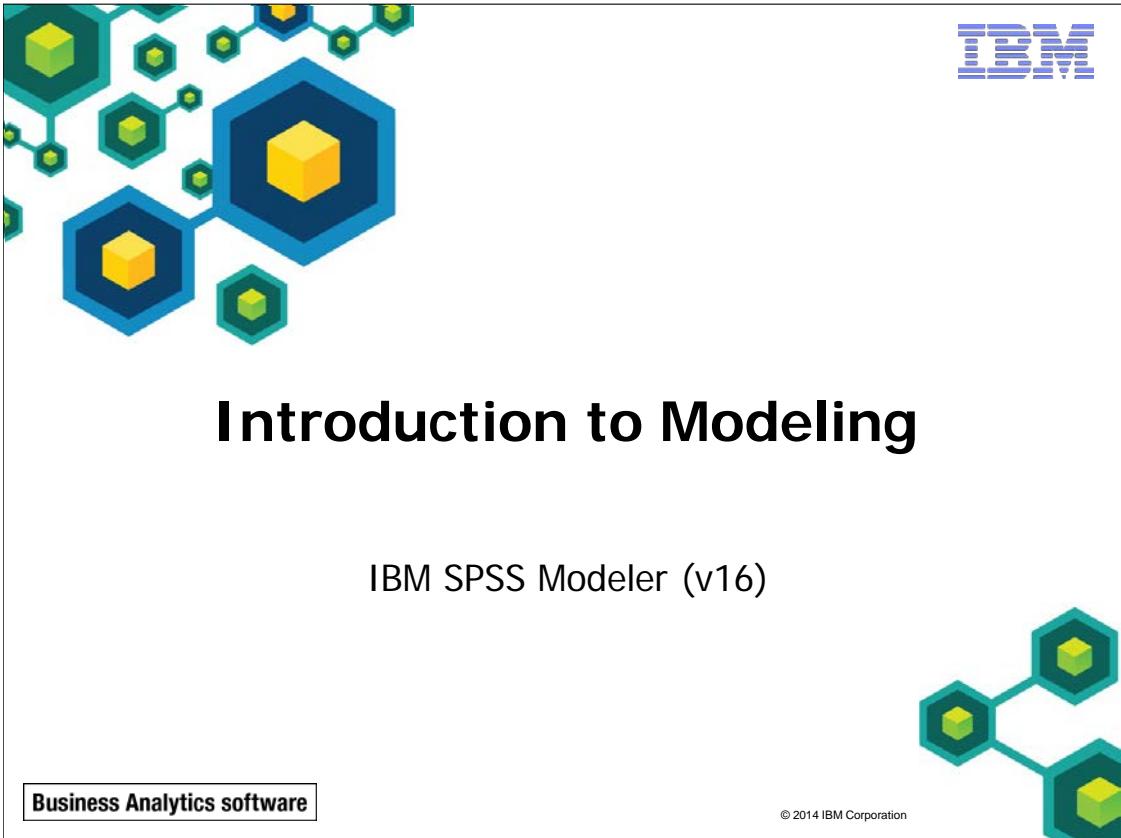
The graph supports the finding that there is no relationship between credit limit and response to the test mailing.

Note: The stream **workshop\_looking\_for\_relationships\_completed.str**, located in the **09-Looking\_for\_Relationships\Solution Files** sub folder provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The slide features a white background with a decorative pattern of blue hexagons containing yellow cubes and smaller green hexagons at the top left. The IBM logo is in the top right corner. The main title "Introduction to Modeling" is centered in bold black font. Below it, the subtitle "IBM SPSS Modeler (v16)" is also centered. A small box at the bottom left contains the text "Business Analytics software". At the bottom right, there is a graphic of three interconnected hexagons and the copyright notice "© 2014 IBM Corporation".

# Introduction to Modeling

## IBM SPSS Modeler (v16)

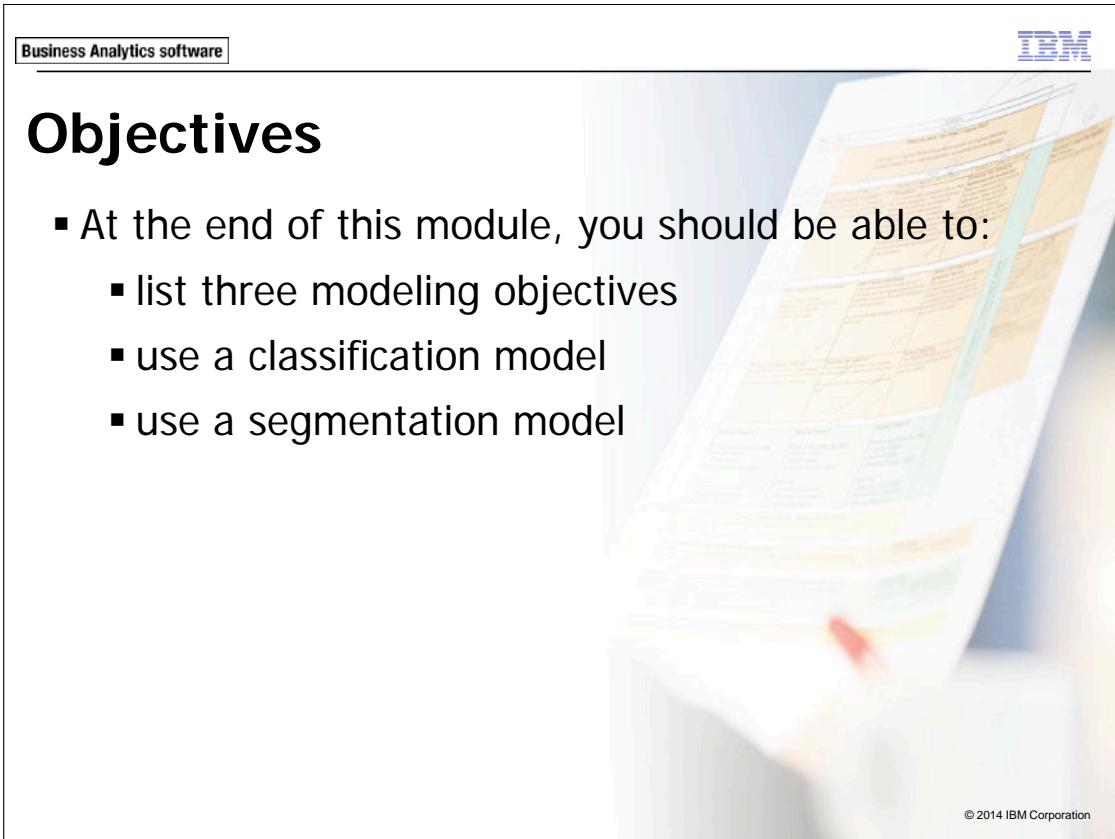
Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The image shows a screenshot of IBM Business Analytics software. In the top left corner, there is a small box labeled "Business Analytics software". In the top right corner, the "IBM" logo is visible. The main area of the interface displays a "Modeling" palette with various nodes and connections, typical of a data mining or modeling application like MODELER.

This module focuses on the Modeling stage in the CRISP-DM process model.

MODELER offers many modeling nodes, which are located in the Modeling palette. These modeling nodes can be classified into three categories, depending on the modeling objective.

This module starts with a review of the three modeling objectives and then presents two classes of models, corresponding to two modeling objectives.

Before reviewing this module you should be familiar with:

- CRISP-DM
- MODELER streams, nodes and palettes
- methods to collect initial data
- methods to explore the data
- methods to examine the relationship between two fields

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-3

# Modeling Objectives

- Classification
- Segmentation
- Association

© 2014 IBM Corporation



Although there are different taxonomies, modeling objectives can be classified as follows:

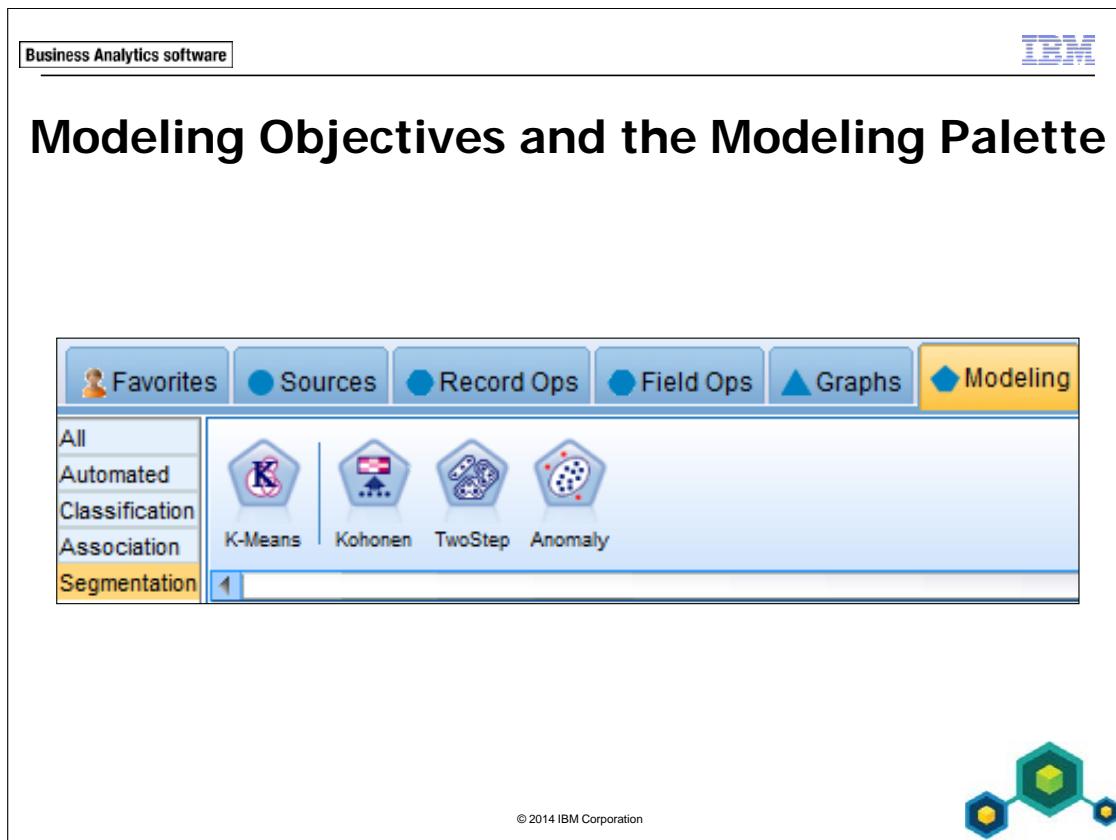
- Classification: Predict a target field, using one or more predictors. Examples of classification include:
  - Telco. Predict if a customer churns, service usage data can be used to predict which customers are liable to transfer to another provider.
  - Banking. Predict if a customer fails on paying back a loan; possible predictors are age, sex, and income.
- Segmentation: Discover groups of records with similar values or patterns. Segmentation is also known as clustering. Examples of segmentation include:
  - Marketing: Cluster customers on RFM (Recency, Frequency, Monetary value) fields into gold, silver, and bronze segments, and approach each segment differently.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Insurance: Cluster claims and look for unusual cases within the groups. This is also known as anomaly detection and a useful method to detect fraud.
- As a data preparation step prior to predictive modeling. For example, you can use the cluster group field as an additional predictor.
- Association: Looking for relationships between fields. A typical example of association is in so-called market basket analysis:

Rule #1: 40 % of the customers have products A and B and 90 % of these customers also have C

Rule #2: 30 % of the customers have products C, D, and G and 75% of these customers also have F



The Modeling palette is organized by the modeling objectives: each of the objectives is an item. Selecting one of the items will show all modeling nodes suitable for that task.

This slide shows an example. When you select the Segmentation item on the Modeling palette, the segmentation models will appear.

# Objectives and Roles in the Type Node

| Modeling Objectives | Field Roles       |
|---------------------|-------------------|
| Classification      | Input(s), Targets |
| Segmentation        | Input(s)          |
| Association         | Both              |

© 2014 IBM Corporation



Each objective needs different specifications in the Role column of the Type node.

Classification models need one or more input fields (predictors), and a target field.

In Segmentation models you will only have fields with role Input. The records will be clustered on the basis of these fields. No target field is specified.

Field roles in association models are typically set to Both. For example, consider the rules:

- Rule #1: 40 % of the customers have products A and B and 90 % of these customers also have C
- Rule #2: 30 % of the customers have products C, D, and G and 75% of these customers also have F

Here, product C is a target field in rule #1, and an input field in rule #2.

For other roles, refer to the online Help.

## Types of Classification Models

- Rule induction models
- Traditional statistical models
- Machine learning models

© 2014 IBM Corporation

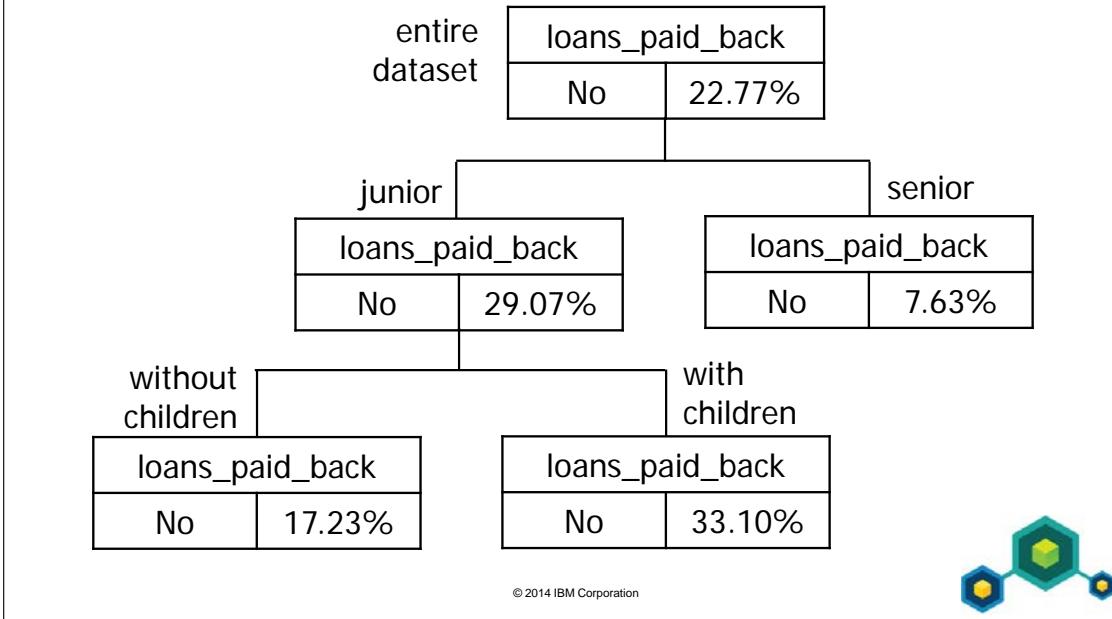


MODELER offers a great number of classification models. Three classes of classification models can be distinguished:

- rule induction models
- traditional statistical models
- machine learning models

In this section an example of each type will be presented. Refer to the IBM SPSS predictive modeling courses for a presentation of these models.

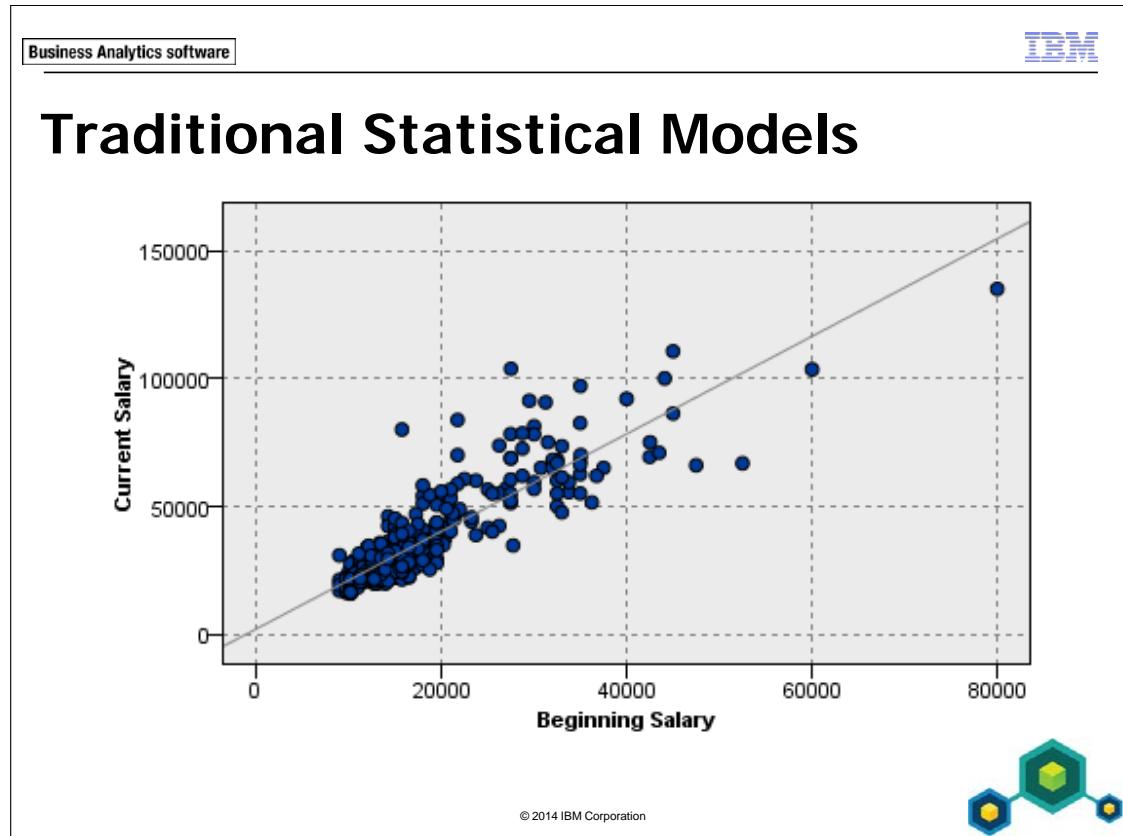
# Rule Induction Models



Rule induction models derive a set of rules that describe distinct segments within the data in relation to the target. Models that produce trees belong to this class of models. Some rule induction models can be used for a target of any measurement level. Other rule induction models are only relevant to a target of a particular measurement level.

One of the most popular rule induction models is CHAID. CHAID supports both categorical and continuous targets. This slide gives an example for the flag target field loans paid back. In the entire dataset, 22.77% did pay back their loans. The most important predictor, in terms of significance as assessed by the Chi-square test, is age category. Within the senior group, there are no further subgroups; within the junior group it matters whether one has children or not. The terminal nodes make up the model:

1. age category = "senior" 7.63% did not pay their loan back)
2. age category = "junior" and has children = "no" (17.23% did not pay back)
3. age category = "junior" and has children = "yes" (33.10% did not pay back)



Traditional statistical classification models make stronger assumptions than Rule Induction models or machine learning models do. For example, statistical models assume certain distributions.

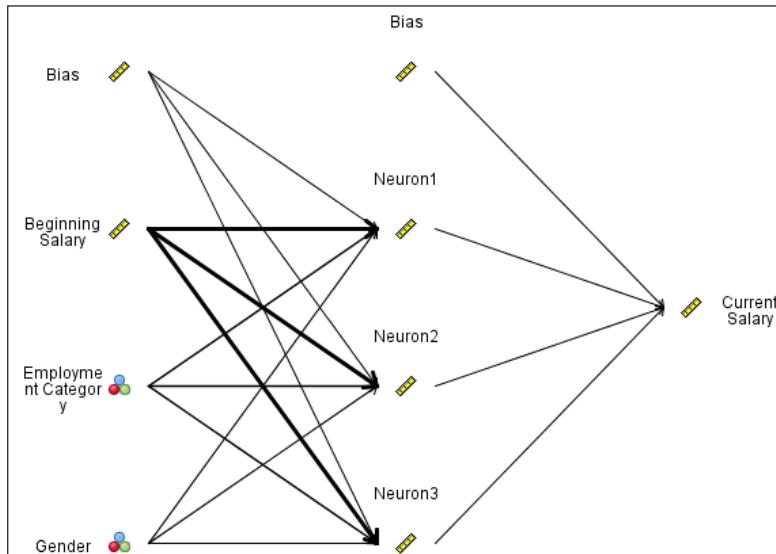
The outcome produced by these models is expressed by an equation, and statistical tests can guide field selection in the model.

Again, the measurement level of the target is a decider in which model to choose. For example, you will use linear regression for continuous targets, and you will use logistic regression for categorical targets.

This slide sketches the idea of one model in this category, linear regression. The plot depicts the relationship between Beginning Salary and Current Salary. Superimposed on the plot is the linear line, or regression line that best describes the linear relationship between the two fields. A point represents the actual current salary, the line represents the predicted current salary, given a certain beginning salary. The line is expressed mathematically by an equation:  $\text{Current Salary} = a + b * \text{Beginning Salary}$ , and MODELER will find the coefficients  $a$  and  $b$ .

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

# Machine Learning Models



© 2014 IBM Corporation



Machine Learning models are optimized to learn complex patterns. No assumptions are made as traditional statistical models do, they do not produce an equation or a set of rules.

The most well-known model in this area is the Neural Net, depicted on this slide. Historically, neural networks attempted to solve problems in a way modeled on how the human brain operates. Neural networks are generally viewed as powerful models, but the interpretation of the results of the mathematical model that is used behind the scenes is difficult. That is why Neural Networks are considered to be black box models.

# Which Classification Model to Use?

- Determined by:
  - business context
  - measurement level of the target
  - your choice in how models handle details
- Alternatively, combine several models (AutoClassifier node, AutoNumeric node).

© 2014 IBM Corporation



From the previous presentation it follows that more than one classification model can be used to predict a target. Which model should you then choose?

The business context is a first decider in the choice of the model. For example, if the model has to be presented in a number of rules, then a Rule Induction model will be preferred above a black box model such as a neural network. Or, if the business context is such that the model itself is of no interest, but only that the model has to be as accurate as possible in predicting the target, then each model is a candidate for the modeling task.

A second decider in the choice of a classification model is the measurement level of the target. Some models are only appropriate for categorical targets, others for continuous targets. For example, the C5.0 node will not work for a continuous target. It cannot be stressed enough that setting correct measurement levels is a necessary condition for an analysis to make sense.

But even when one type of models is preferred over another, for a target of a certain measurement level, more models within that type of models are available. In the predictive modeling courses it is shown that models differ in:

- how missing values are handled
- how categorical predictors are handled
- how continuous predictors are handled
- how a model scores data

In conclusion, there are many subtle differences between models. In the end it is always the business user, balancing all pros and cons, who decides which model should be used. Perhaps you decide to run multiple models and combine them into a single "overall" model. The AutoClassifier and AutoNumeric node automate finding the best model, including combined models. Again, there is a wide range of possibilities and it is only the business user who can decide what to do.

## Running Classification Models

- In a Type node upstream the modeling node:
  - set measurement levels
  - set roles (Inputs, Target)
- Add the modeling node.
- Run the modeling node.

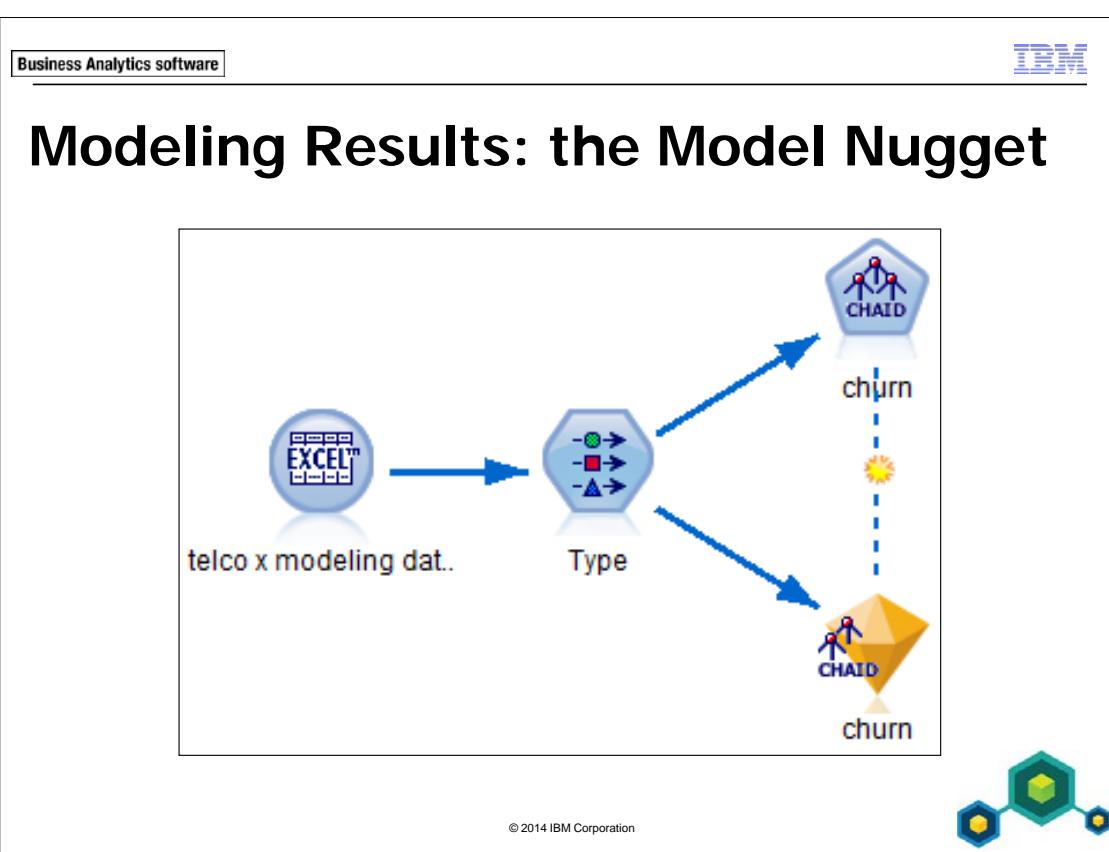
© 2014 IBM Corporation



Critical for any model is that measurement levels and roles are set correctly. To do this, use a Type node and set roles and measurements levels in the Types tab of the node. Using a Type node guarantees that modeling nodes downstream from this Type node will use the same settings.

Add the modeling node(s) downstream from the Type node. The modeling node will be labeled with the target field name if the model that was selected is a valid modeling technique given the measurement level of the target. If the model does not support the target's measurement level, the modeling node will be labeled with "no target".

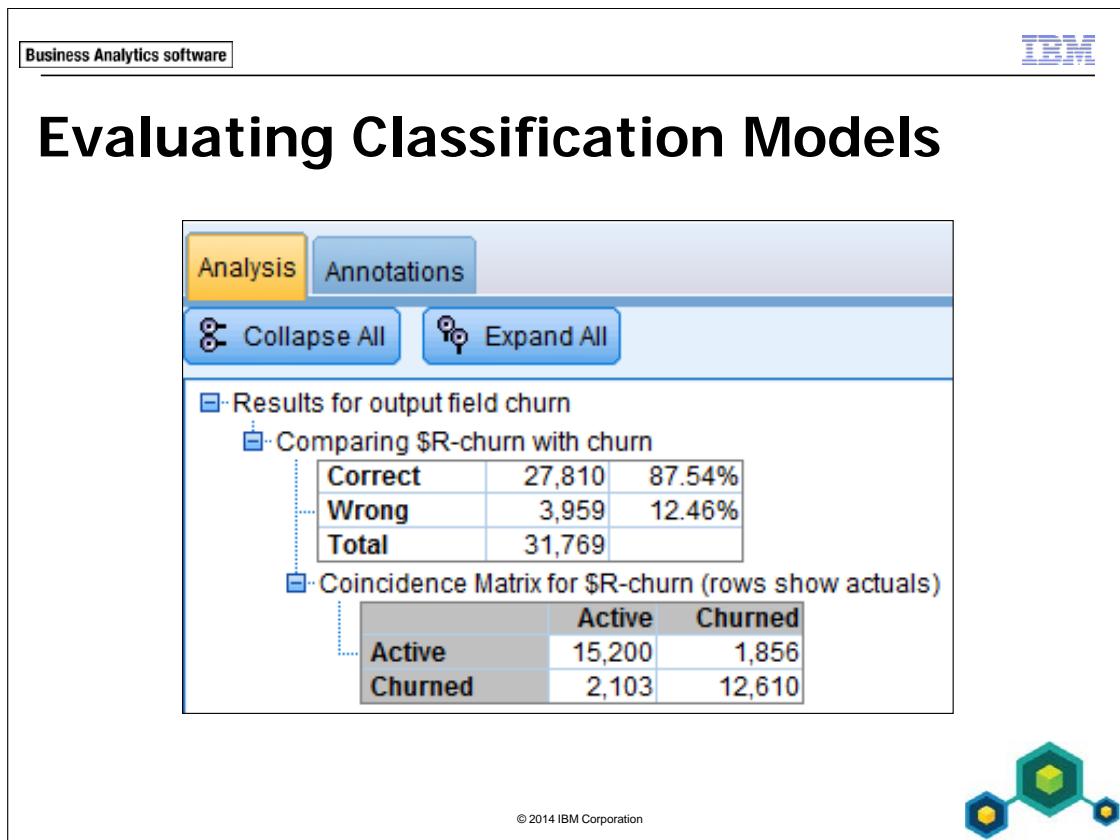
If required, the user can customize the settings for the modeling technique. This assumes that the user is familiar with the modeling technique. In general, default values are a good starting point.



Running a modeling node will add a model nugget to the stream. The model nugget stores the results of the analysis and is linked to the modeling node. The link ensures that when you rerun the model, for example with other predictors, that the model nugget is updated with the new result.

If you do not want to update the model nugget when the model is rerun, break the link between the modeling node en the model nugget.

To view the model's output, edit the model nugget. The output depends on which modeling node was run. For example, you will have a tree when you have run a CHAID model.



The screenshot shows the IBM SPSS Modeler interface. At the top, there's a header bar with the IBM logo and the text "Business Analytics software". Below the header is a toolbar with buttons for "Analysis" (which is highlighted in yellow) and "Annotations", and buttons for "Collapse All" and "Expand All". The main content area displays analysis results for an output field named "churn". It includes a section titled "Comparing \$R-churn with churn" which contains three tables:

- Correct:** 27,810 (87.54%)
- Wrong:** 3,959 (12.46%)
- Total:** 31,769

Below this is a "Coincidence Matrix for \$R-churn (rows show actuals)" table:

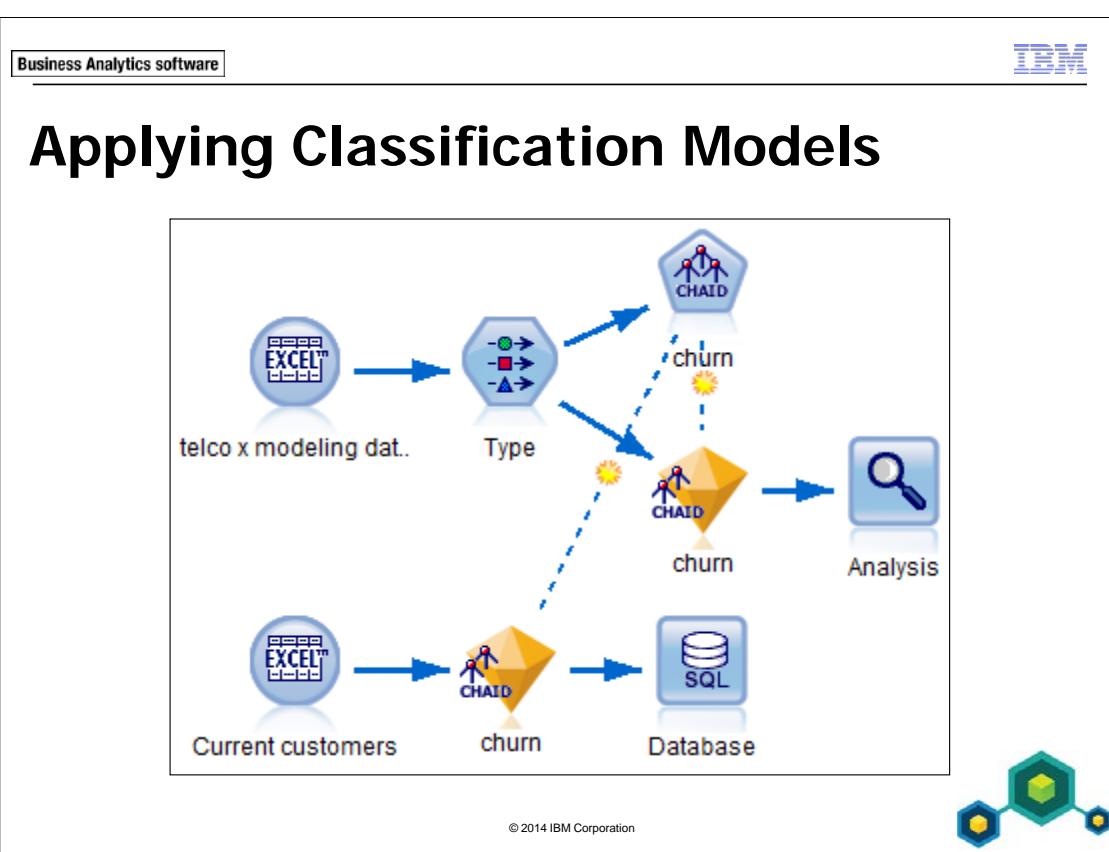
|         | Active | Churned |
|---------|--------|---------|
| Active  | 15,200 | 1,856   |
| Churned | 2,103  | 12,610  |

At the bottom of the interface, there's a small decorative graphic of interconnected hexagons and the text "© 2014 IBM Corporation".

You can use an Analysis node, located in the Output palette, to evaluate your predictive model. The Analysis node compares the target's actual values to the target's predicted values. In a perfect model, the actual values and predicted values coincide. A perfect model will only occur in trivial datasets, or indicates that the wrong predictors have been used.

The slide here shows the Analysis output when the flag target churn has been predicted. The coincidence matrix shows the target's actual values in the rows, and the target's predicted values in the column. Here 15,200 customers were predicted to be active, while they were actually active, and 12,610 customers were predicted to churn, while they actually churned. This make a total of  $15,200 + 12,610 = 27,810$  correct predictions, a 87.54% percentage of correct predictions.

When the target is continuous, other statistics measure the model's accuracy.



Because the model nugget contains the model, it can be used to score records. In this sense, it is not different from a Derive node, although the expression that is in the nugget will be more complex than an expression in a Derive node.

Classification models generate model nuggets that add the predicted value to the data, and the confidence for the predicted value. Previewing the data in the model nugget or running a Table node downstream the model nugget will show the new field(s).

This slide presents an example, when the CHAID modeling node was executed. A model nugget was generated and was used to score the current customer. The results predictions are exported to a database for further processing.

# Segmentation Models

- Three segmentations methods:
  - K-Means
  - Kohonen
  - Two-Step
- AutoCluster node to automate the analysis.

© 2014 IBM Corporation



MODELER offers three segmentation methods:

- K-Means: A quick method for exploring clusters in data. The user sets the number of clusters (K) to be created, and the algorithm iterates until each record is assigned to the nearest of the K clusters. Since the user must set the number of clusters, this procedure is typically run several times, assessing the results for different numbers of clusters (values of K).
- Kohonen networks: A neural network approach to segmentation. A Kohonen network is an  $m \times n$  grid of neurons, where a neuron represents a profile for a cluster. When a record is presented to the grid, its input fields are compared with the neurons within the grid. The neuron with the pattern most like that of the input wins the record. This resembles K-Means as the number of clusters is set upfront (there will be  $m \times n$  clusters in the solution). However, the algorithms differ in how they assign records to the clusters and how clusters are updated.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

10-18

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- TwoStep: In the first step, all records are classified into a great number of pre-clusters. These pre-clusters are designed to be well separated. In the second step, another cluster method is used to successively combine the pre-clusters. Compared to the other two clustering techniques, the two-step clustering method thus has the advantage of automatically selecting the number of clusters. The disadvantage is that TwoStep assumes certain distributions for the input fields.

Refer to the *Clustering and Association Models using IBM SPSS Modeler (v16)* course for more details on segmentation models.

## Running Segmentation Models

- Set Roles in a Type node
- Add the modeling node downstream from the Type node
- Run
- Edit the model nugget to view the results

© 2014 IBM Corporation



To run a segmentation model, set the measurement levels and roles in a Type node upstream from the segmentation modeling node. Segmentation models do not have targets, so that only the Input role is relevant.

Similar to running classification models, a model nugget will be generated, added downstream from the Type node, and linked to the modeling node. The model nugget will add a cluster-membership field to the data.

The model nugget will also store the modeling results, and editing the model nugget will show these results.

In segmentation models, the so-called Silhouette statistic is used to evaluate the quality of the cluster solution. The Analysis node is not relevant here, because the Analysis node compares the actual target field values to the predicted target field values, but in segmentation models there is no target field.

## Examining the Results: Cluster Profiles

|                    | Cluster 1 | Cluster 2 |
|--------------------|-----------|-----------|
| bill_peak          | 20.1      | 57.8      |
| bill_offpeak       | 17.3      | 102.11    |
| bill_international | 9.9       | 94.48     |

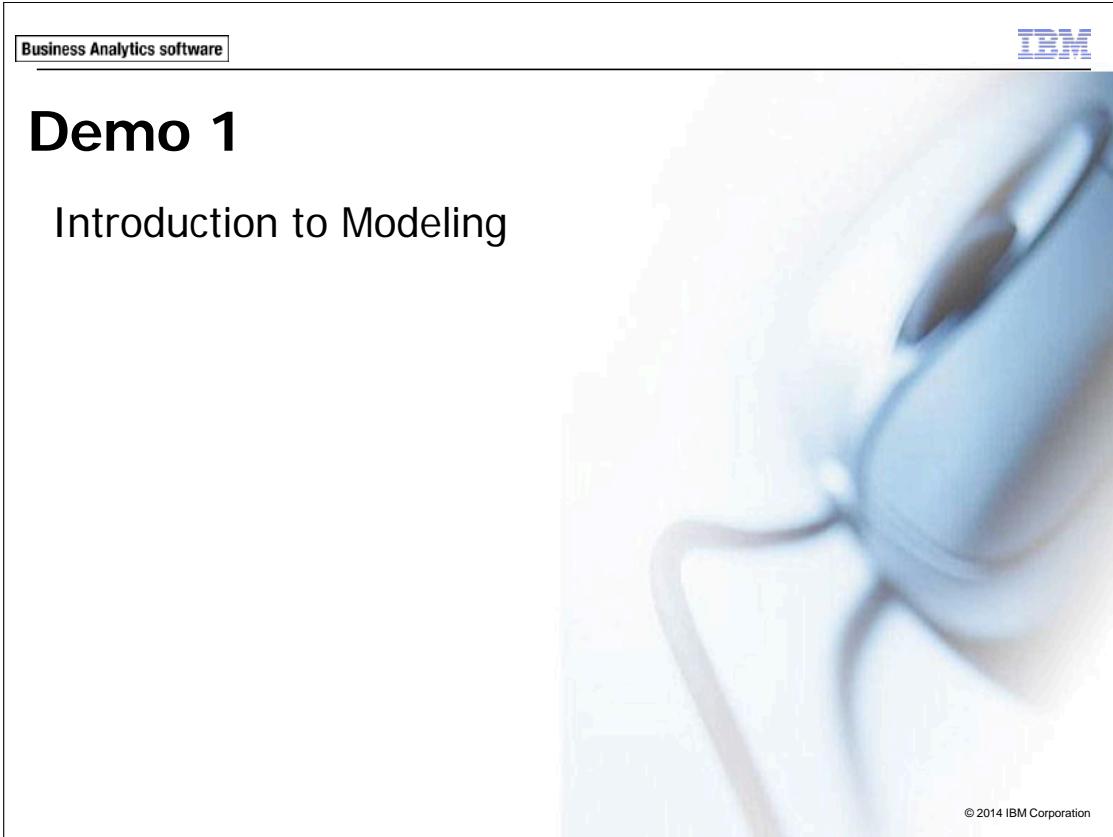
© 2014 IBM Corporation



When you edit the model nugget, you will have a table showing the mean of the input fields (when the inputs are continuous) for each cluster. This is a helpful table to profile the clusters, and to label them.

The example on this slide shows the results for an analysis with three continuous inputs, where two clusters were found. The first cluster has low mean values on the fields, so you may label them as inactive customers.. The second cluster may be labeled as active customers.

For more information, refer to the *Clustering and Association Models Using IBM SPSS Modeler (v16)* course.



The slide is titled "Demo 1" and "Introduction to Modeling". It features the IBM logo in the top right corner and a small "Business Analytics software" badge in the top left. A large, abstract blue and white graphic of a stylized figure or flower occupies the right side. The bottom right corner contains the text "© 2014 IBM Corporation".

This demo uses a (synthetic) dataset coming from a (fictitious) telecommunications firm to introduce you classification and segmentation.

Before you begin with the demo, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **demo\_introduction\_to\_modeling.str**, located in the **10-Introduction\_to\_Modeling\Start Files** sub folder.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

The files used in this demo are:

- **telco x modeling data.xls**: a Microsoft Excel file, combining data from various sources in a single dataset, and including a number of derived fields
- **demo\_introduction\_to\_modeling.str**: MODELER stream that imports, sets the fields' measurement levels, and instantiates the data

## Demo 1: Introduction to Modeling

### Purpose:

**In the first task, you will use data from a telecommunications firm to predict churn by running a CHAID model and a Neural Net model. You will also compare the accuracy of these models.**

**In the second task you will find groups (clusters) of similar customers, based on usage.**

### Task 1. Predicting churn.

1. When you have opened **demo\_introduction\_to\_modeling.str**, located in the **Introduction\_to\_Modeling\Start Files** sub folder, scroll to the **upper branch** of the stream, and then add a **CHAID** node (Modeling palette – Classification item) downstream from the **Type** node.

2. Add a **Neural Net** node (Modeling palette – Classification item) downstream from the **Type** node.

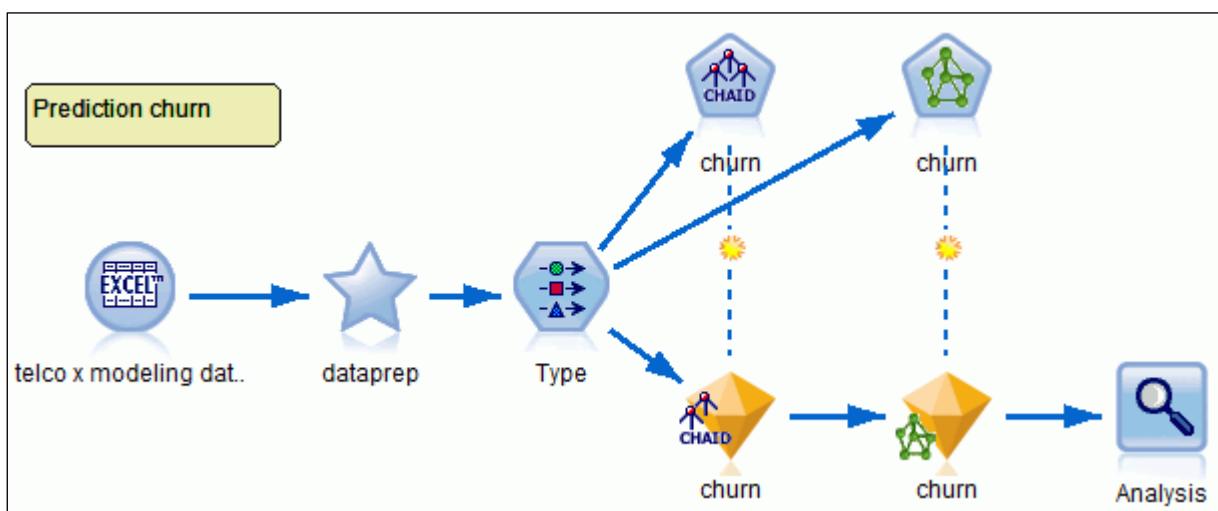
3. Run the **CHAID** node, and then run the **Neural Net** node.

When the nodes have been executed, two model nuggets are added downstream from the **Type** node (and linked to the respective model node). To make comparison of the models easier, you will rearrange the stream so that both model nuggets are in the same branch of the stream.

4. Disconnect the **Neural Net** model nugget from the **Type** node, place it downstream from the **CHAID** model nugget, and then connect the **CHAID** model nugget to the **Neural Net** model nugget.

5. Add an **Analysis** node (Output palette) downstream from the **Neural Net** model nugget.
6. Edit the **Analysis** node, and then:
  - enable the option Coincidence matrices (for symbolic targets)
  - click **OK** to close the **Analysis** dialog box

A section of the results appear as follows:



7. Run the **Analysis** node.

A section of the results appear as follows:

The screenshot shows the 'Analysis' tab selected in the top navigation bar. Below it are buttons for 'Collapse All' and 'Expand All'. The main content area displays a hierarchical tree of results for the 'churn' output field. Under 'Individual Models', there are two entries: 'Comparing \$R-churn with churn' and 'Comparing \$N-churn with churn'. Each entry has a table showing 'Correct', 'Wrong', and 'Total' counts, along with their percentages. Below each model entry is a 'Coincidence Matrix' table. The first matrix (for \$R-churn) has columns 'Active' and 'Churned', and rows 'Active' and 'Churned'. The second matrix (for \$N-churn) has columns 'Active', 'Churned', and '\$null\$', and rows 'Active' and 'Churned'.

|         | Active | Churned |
|---------|--------|---------|
| Active  | 15,200 | 1,856   |
| Churned | 2,103  | 12,610  |

|         | Active | Churned | \$null\$ |
|---------|--------|---------|----------|
| Active  | 15,560 | 1,493   | 3        |
| Churned | 2,227  | 12,474  | 12       |

The accuracy (the percentage of correct predictions) of the Neural Net model (\$N- churn) is slightly better than the accuracy of the CHAID model (\$R- churn).

The coincidence matrix shows the actual category in the rows, the predicted category in the column. You can verify that the CHAID model predicted 15,200 active customers and 12,610 churners correctly. The Neural Net predicted more active customers correctly (15,560), but fewer churners (12,474).

The CHAID model gives more insight, Also, the CHAID model is actionable, because the tree shows where the customers at risk are located, plus CHAID identifies more churners. All in all, CHAID is the preferred model.

As a note, comparing models is usually done on a test dataset. A Partition node, located in the Field Ops palette, upstream from the Type node accomplishes this. Refer to the online Help for more information on the Partition node; or refer the *Advanced Data Preparation with IBM SPSS Modeler* (v16) course.

8. Click **OK** to close the **Analysis** output window.

Leave the stream open for the next task.

## Task 2. Finding groups of similar customers.

In this task you will build from the previous stream.

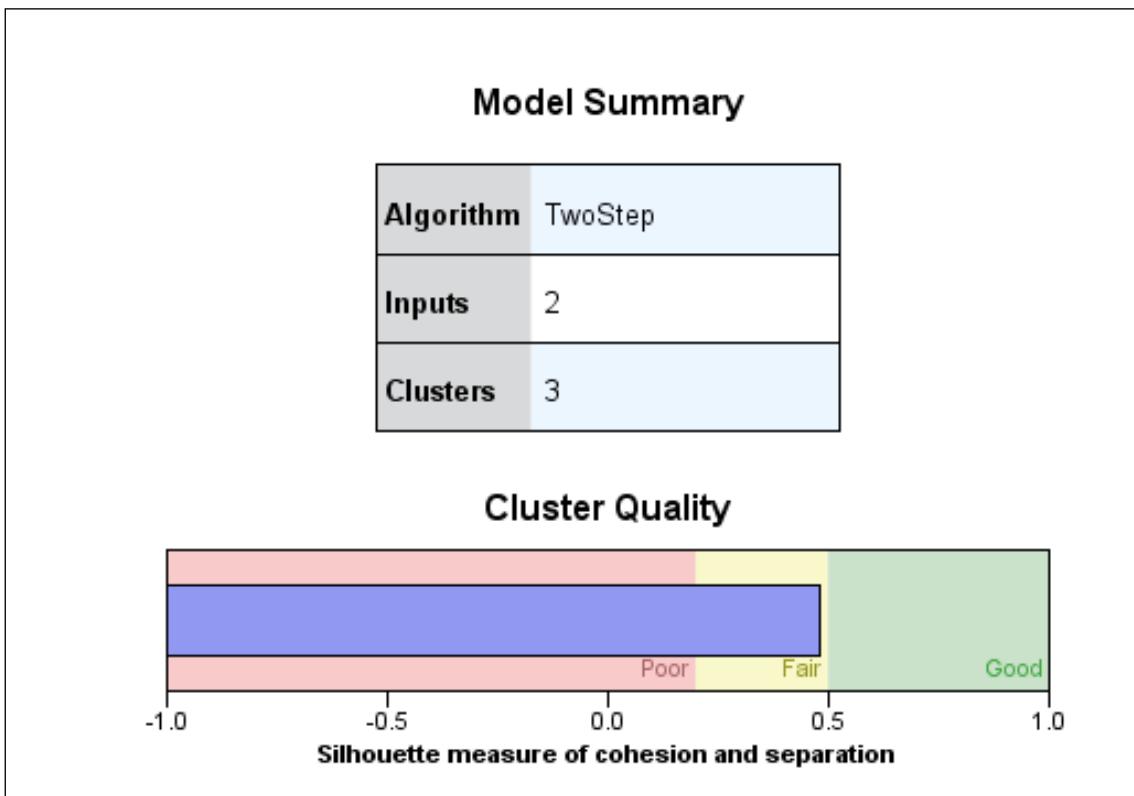
1. Scroll to the **lower branch** of the stream.

You will use TwoStep clustering because this model will find the number of clusters automatically.

2. Add a **TwoStep** node (Modeling palette, Segmentation item) downstream from the **Type** node.
3. Run the **TwoStep** node.

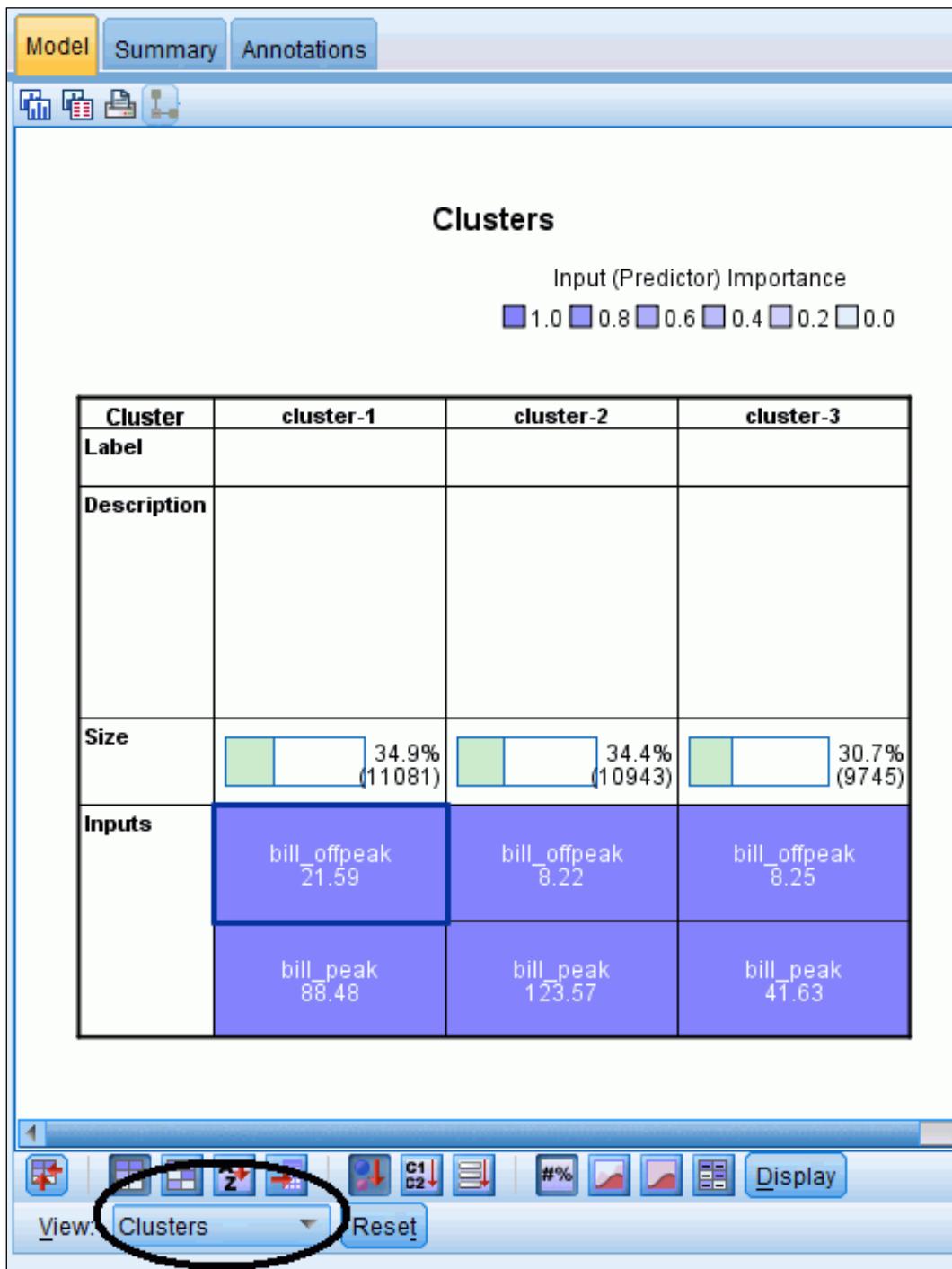
4. Edit the **TwoStep** model nugget that was generated.

A section of the results appear as follows:



Three clusters are found, in a solution that is considered to be fair.

5. For View, select **Clusters**.
6. A section of the results appear as follows:



Cluster 1 is comprised of customers phoning off-peak, cluster 2 is characterized by a high usage in the peak hours, and cluster 3 is a segment of customers not being very active in both peak and off-peak hours.

It is up to you to decide if this solution is useful. You could also explore if the cluster field is a valuable predictor for churn.

This completes the demo for this module. You will find the solution results in **demo\_introduction\_to\_modeling\_completed.str**, located in the located in the **10-Introduction\_to\_Modeling\Solution Files** sub folder.

## Apply Your Knowledge

Use the questions in this section to test your knowledge of the course material.

Question 1: Which of the following is the correct statement? Based on transactional data such as minutes of outgoing calls, minutes of incoming calls, and text messaging, a telecommunications firm clusters their customers and finds groups such as leaders and followers. This is an example of:

- A. classification
- B. segmentation
- C. association

Question 2: Which of the following is the correct statement? A retailer runs an analysis on what customers have in their shopping carts to find out what popular product combinations are. This is an example of:

- A. classification
- B. segmentation
- C. association

Question 3: Which of the following is the correct statement? An Insurance company has historical data on claims, such as claim amount, gender of the policy holder, age of policy holder, claim type, number of claims in a one year period. The company has found out that claim amount is related to number of claims within a one year period, claim type and gender of the policy holder. This is an example of:

- A. classification
- B. segmentation
- C. association

Question 4: Is the following statement true or false? In the Type node, the field's role is set in the Measurement Level column.

- A. True
- B. False

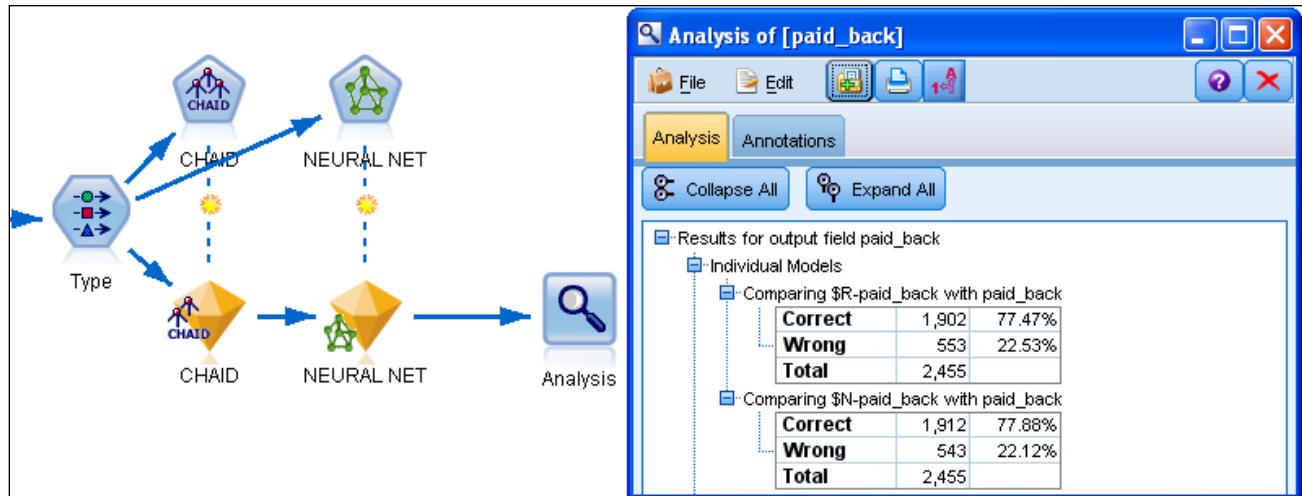
Question 5: Refer to the figure that follows. Which of the following statements holds for the record with customer\_id K103540:

- A. CHAID and Neural Net predict another category for the target.
- B. The confidence for the CHAID prediction is higher than the confidence for the Neural Net prediction.
- C. The CHAID prediction is correct.

| customer_id    | gender        | age           | tariff         | dropped_calls | handset    | churn         | \$R-churn     | \$RC-churn   | \$N-churn      | \$NC-churn   |
|----------------|---------------|---------------|----------------|---------------|------------|---------------|---------------|--------------|----------------|--------------|
| K102840        | FEMALE        | 26....        | CAT 100        | 6.000         | S50        | Active        | Active        | 0.600        | Churned        | 0.534        |
| K103330        | FEMALE        | 27....        | CAT 100        | 10.000        | S50        | Active        | Active        | 0.600        | Churned        | 0.650        |
| <b>K103540</b> | <b>FEMALE</b> | <b>40....</b> | <b>CAT 100</b> | <b>11.000</b> | <b>S50</b> | <b>Active</b> | <b>Active</b> | <b>0.718</b> | <b>Churned</b> | <b>0.506</b> |
| K103550        | FEMALE        | 26....        | CAT 100        | 11.000        | S50        | Active        | Active        | 0.600        | Churned        | 0.707        |
| K106100        | FEMALE        | 27....        | CAT 100        | 10.000        | S50        | Churned       | Active        | 0.600        | Churned        | 0.650        |
| K108820        | FEMALE        | 26....        | CAT 100        | 9.000         | S50        | Active        | Active        | 0.600        | Churned        | 0.627        |

Question 6: Referring to the figure that follows, is the following statement true or false? The percentage correctly classified is higher for CHAID than for Neural Net.

- A. True
- B. False



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Question 7: Which of the following statements are correct?

- A. In K-means, the user specifies the number of clusters upfront.
- B. In Kohonen, the number of clusters is determined by an  $m * n$  grid.
- C. In TwoStep, the number of clusters is always 2.
- D. A model nugget generated by K-Means, Kohonen or Two-Step Cluster, adds the cluster membership.

Question 8: Is the following statement true or false? Adding an Analysis node downstream from a TwoStep model nugget will compare the predicted cluster membership with the actual cluster membership.

- A. True
- B. False

## Answers to questions:

Answer 1: B. Segmentation, because the objective is to find clusters of records.

Answer 2: C. Association, because the objective is to find popular product combinations.

Answer 3: A. Classification, because fraud is predicted.

Answer 4: B. False, roles are set in the Role column.

Answer 5: A, B, C.

Answer 6: B. False.

Answer 7: A, B, D.

Answer 8: B. False. There is no target field in segmentation models (and the Analysis node compares target fields (the actual target field with the predicted target field)).

Business Analytics software

IBM

## Summary

- At the end of this module, you should be able to:
  - list three modeling objectives
  - use a classification model
  - use a segmentation model

© 2014 IBM Corporation

In this module you were introduced to MODELER's modeling capabilities. Three types of models are supported: classification, segmentation, and association models. Each type needs different field roles.

Two types of models were given a closer look: classification and segmentation models.

# Workshop 1

## Introduction to Modeling



© 2014 IBM Corporation

Before you begin with the workshop, ensure that:

- You have started MODELER.
- You have set MODELER's working folder. (In MODELER, select **File\Set Directory**. When you are working in an IBM operated environment, browse to the folder **C:\Train\0A005** and then click **Set** to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- You have opened **workshop\_introduction\_to\_modeling.str**, located in the **10-Introduction\_to\_Modeling\Start Files** sub folder.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

10-36

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

The following files are used in this workshop:

- **ACME analysis data.sav**: an IBM SPSS Statistics data file, ready for modeling
- **workshop\_introduction\_to\_modeling.str**: a MODELER stream file that imports the data, sets the fields' measurement levels, instantiates the data, and is the starting point for the workshop

## Workshop 1:Introduction to Modeling

You are working at ACME, where you prepared a dataset for modeling. Now the time has come to build a model to predict response to the test mailing, and, if the model is satisfactory, to apply it to the customers that were not included in the test mailing. Also, you will use a segmentation model to find clusters of customers in the ACME database.

- Use CHAID to predict response\_to\_test\_mailing, with the RFM (recency, frequency and monetary value) fields as predictors.

The stream **workshop\_introduction\_to\_modeling.str**, located in the **10-Introduction\_to\_Modeling\Start Files** sub folder.

When the model nugget has been added to your stream, examine the tree. What is the predicted response for customers having in the high frequency category, high recency category and high monetary value category?

- Assess the model's accuracy.

What is the percentage predicted correctly? How many of the responders have been identified as such by the model?

- Assuming the model found is satisfactory, apply the model to the customers that were not in test mailing, select all those customers that are predicted to respond, and then export their data to a text file.

How many customers (not included in the test mailing), are predicted to respond?

- Use the TwoStep segmentation model to cluster records (all 30,000 customers), based on the RFM fields.

How many clusters are found? Is the solution acceptable?

Profile the clusters in terms of the input fields (the RFM fields).

For more information about where to work and the workshop results, refer to the Task and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

## Workshop 1: Tasks and Results

### Task 1. Build a CHAID Model to predict response.

A model will be built on only those customers that were included in the test mailing.

- Add a **Select** node downstream from the source node named **ACME analysis data.sav**.
- Edit the **Select** node, and then type the condition **has\_received\_test\_mailing = "yes"** (alternatively, and recommended, generate the **Select** node from a Table output window).
- Add a **Type** node downstream from the **Select** node.
- Edit the **Type** node, set **Role** for **recency**, **frequency** and **monetary value** to **Input**, **Role** for **response\_to\_test\_mailing** to **Target**, and then click **Read Values** to instantiate the data.
- Add a **CHAID** node (Modeling palette – Classification item) downstream from the **Type** node, and then run the **CHAID** node.

A model nugget is generated, connected to the Type node, and linked to the CHAID node.

## Task 2. Assess the model's accuracy.

- Add an **Analysis** downstream from the **CHAID** model nugget.
- Edit the **Analysis** node, and then check the **Coincidence matrices (for symbolic targets)** option.
- Click **Run**.

A section of the results appear as follows:

The screenshot shows the 'Analysis' tab selected in the top navigation bar. Below it are 'Collapse All' and 'Expand All' buttons. The results pane displays the following information:

- Results for output field response\_to\_test\_mailing**
- Comparing \$R-response\_to\_test\_mailing with response\_to\_test\_mailing**

|         |        |        |
|---------|--------|--------|
| Correct | 9,704  | 97.04% |
| Wrong   | 296    | 2.96%  |
| Total   | 10,000 |        |
- Coincidence Matrix for \$R-response\_to\_test\_mailing (rows show actuals)**

|   | F     | T   |
|---|-------|-----|
| F | 9,542 | 97  |
| T | 199   | 162 |

The accuracy is 97.04%; 162 customers were identified by the model as responders.

### Task 3. Apply the model to other customers.

Assuming that the accuracy is satisfactory, apply the model to the customers that were not in the test mailing: select these customers, and then use the model nugget.

- Add a **Select** node downstream from the source node.
- Edit the **Select** node, and then type the condition **has\_received\_test\_mailing = "no"** (alternatively, generate the node from a Table output window).
- Copy and paste the **CHAID** model nugget downstream from the **Select** node.
- Add a **Select** node downstream from the model nugget, and then type the expression '**\$R-response\_to\_test\_mailing**' = "T" (alternatively, generate the **Select** node from a Table output window).
- Add a **Table** node downstream from the **Select** node, and then run the **Table** node.

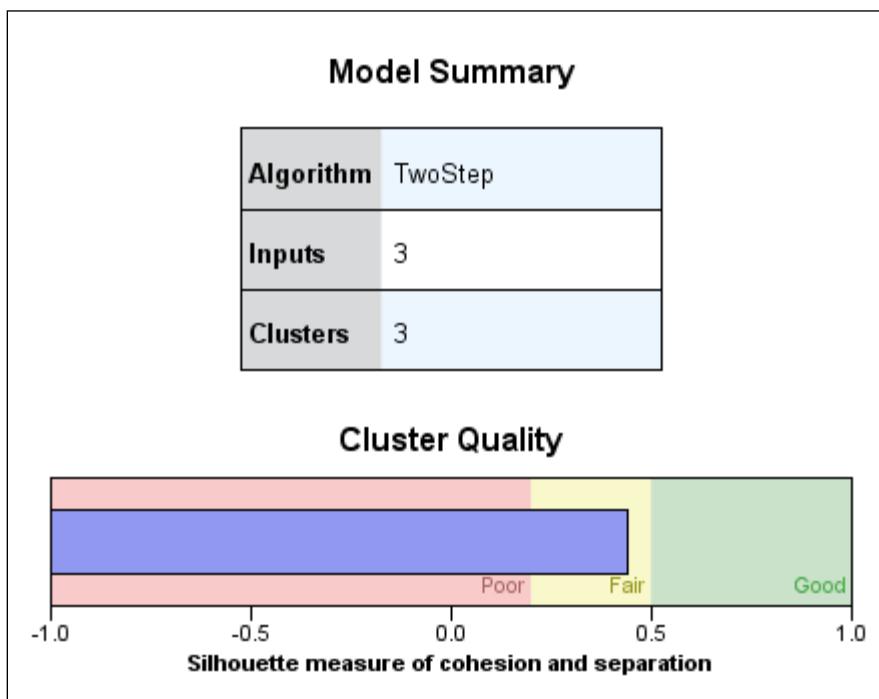
This will show that 254 customers are predicted to respond.

- To export the data for the 254 customers predicted to respond: add a **Flat File** node (Export palette) downstream from the **Select** node, and then run the **Flat File** node.

## Task 4. Find clusters of customers.

- Add a **Type** node downstream from the source node named **ACME analysis data.sav**.
- Edit the **Type** node, set **Role** for **recency**, **frequency** and **monetary value** to **Input**. Ensure that all other fields have **Role None**, and then click **Read Values** to instantiate the data.
- Add a **Two-Step** node (Modeling palette – Segmentation item) downstream from the **Type** node, and then run the **TwoStep** node.
- Edit the **model nugget**.

A section of the results appear as follows:



The model summary shows that three clusters are found, and that the solution is considered to be fair.

To profile the clusters in terms of the input fields, you will examine the Inputs by Clusters grid.

- For View, select **Clusters**.

A section of the results appear as follows:

| <b>Cluster</b>     | <b>cluster-1</b>                   | <b>cluster-3</b>                  | <b>cluster-2</b>                 |
|--------------------|------------------------------------|-----------------------------------|----------------------------------|
| <b>Label</b>       |                                    |                                   |                                  |
| <b>Description</b> |                                    |                                   |                                  |
| <b>Size</b>        | 33.7%<br>(10112)                   | 33.3%<br>(10000)                  | 33.0%<br>(9888)                  |
| <b>Inputs</b>      | frequency<br>2 medi (55.7%)        | frequency<br>1 low (80.2%)        | frequency<br>3 high (73.8%)      |
|                    | monetary_value<br>2 medium (98.9%) | monetary_value<br>3 high (100.0%) | monetary_value<br>1 low (100.0%) |
|                    | recency<br>2 medium (36.3%)        | recency<br>3 high (47.3%)         | recency<br>1 low (49.5%)         |

The screenshot shows the SPSS Modeler interface. At the top, there's a menu bar with 'File', 'Edit', 'View', 'Data', 'Statistics', 'Model', 'Output', 'Help'. Below the menu is a toolbar with various icons: a red asterisk, a blue folder, a green folder, a yellow folder, a blue square with 'A1', a blue square with 'Z', a blue square with 'C1', a blue square with 'C2', a blue square with a downward arrow, a blue square with a right arrow, a blue square with '#%', a blue square with a line graph, a blue square with a bar graph, a blue square with a scatter plot, and a blue square with 'Display'. A black oval highlights the 'Clusters' icon in the toolbar. Below the toolbar is a menu bar with 'View: Clusters' and a 'Reset' button.

Cluster 1 is a group of medium customers. Cluster 3 is a group of high value customers, although they do not have many transactions. Customers in Cluster 2 represent low monetary value, although they have many transactions.

Note: The stream **workshop\_introduction\_to\_modeling\_completed.str**, located in the sub folder **Introduction\_to\_Modeling\Solution Files** provides a solution to the workshop tasks.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing*  
IBM Corporation  
North Castle Drive  
Armonk, NY 10504-1785  
U.S.A.

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE