IBM SPSS Direct Marketing 19

Note: Before using this information and the product it supports, read the general information under Notices on p. 105.

Preface

IBM® SPSS® Statistics is a comprehensive system for analyzing data. The Direct Marketing optional add-on module provides the additional analytic techniques described in this manual. The Direct Marketing add-on module must be used with the SPSS Statistics Core system and is completely integrated into that system.

About SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, is a leading global provider of predictive analytic software and solutions. The company's complete portfolio of products — data collection, statistics, modeling and deployment — captures people's attitudes and opinions, predicts outcomes of future customer interactions, and then acts on these insights by embedding analytics into business processes. SPSS Inc. solutions address interconnected business objectives across an entire organization by focusing on the convergence of analytics, IT architecture, and business processes. Commercial, government, and academic customers worldwide rely on SPSS Inc. technology as a competitive advantage in attracting, retaining, and growing customers, while reducing fraud and mitigating risk. SPSS Inc. was acquired by IBM in October 2009. For more information, visit *http://www.spss.com*.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using SPSS Inc. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the SPSS Inc. web site at *http://support.spss.com* or find your local office via the web site at *http://support.spss.com/default.asp?refpage=contactus.asp*. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Customer Service

If you have any questions concerning your shipment or account, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*. Please have your serial number ready for identification.

Training Seminars

SPSS Inc. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*.

Additional Publications

The *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion*, and *SPSS Statistics: Advanced Statistical Procedures Companion*, written by Marija Norušis and published by Prentice Hall, are available as suggested supplemental material. These publications cover statistical procedures in the SPSS Statistics Base module, Advanced Statistics module and Regression module. Whether you are just getting starting in data analysis or are ready for advanced applications, these books will help you make best use of the capabilities found within the IBM® SPSS® Statistics offering. For additional information including publication contents and sample chapters, please see the author's website: *http://www.norusis.com*

Contents

Part I: User's Guide

Part II: Examples

Appendices

Part I:
User’s Guide

Chapter

1

Direct Marketing

The Direct Marketing option provides a set of tools designed to improve the results of direct marketing campaigns by identifying demographic, purchasing, and other characteristics that define various groups of consumers and targeting specific groups to maximize positive response rates.

RFM Analysis. This technique identifies existing customers who are most likely to respond to a new offer. For more information, see the topic RFM Analysis in Chapter 2 on p. 2.

Cluster Analysis. This is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics. For more information, see the topic Cluster analysis in Chapter 3 on p. 14.

Prospect Profiles. This technique uses results from a previous or test campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns. For more information, see the topic Prospect profiles in Chapter 4 on p. 19.

Postal Code Response Rates. This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns. For more information, see the topic Postal Code Response Rates in Chapter 5 on p. 25.

Propensity to Purchase. This technique uses results from a test mailing or previous campaign to generate propensity scores. The scores indicate which contacts are most likely to respond. For more information, see the topic Propensity to purchase in Chapter 6 on p. 32.

Control Package Test. This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. For more information, see the topic Control Package Test in Chapter 7 on p. 39.

Chapter

2

RFM Analysis

RFM analysis is a technique used to identify existing customers who are most likely to respond to a new offer. This technique is commonly used in direct marketing. RFM analysis is based on the following simple theory:

- The most important factor in identifying customers who are likely to respond to a new offer is **recency**. Customers who purchased more recently are more likely to purchase again than are customers who purchased further in the past.
- The second most important factor is **frequency**. Customers who have made more purchases in the past are more likely to respond than are those who have made fewer purchases.
- The third most important factor is total amount spent, which is referred to as **monetary**. Customers who have spent more (in total for all purchases) in the past are more likely to respond than those who have spent less.

How RFM Analysis Works

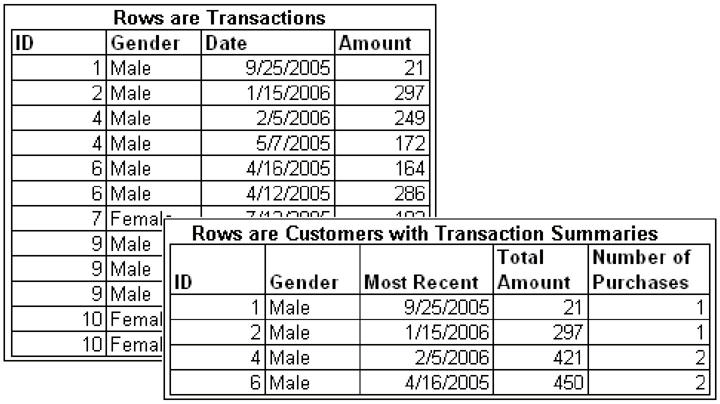
- Customers are assigned a recency score based on date of most recent purchase or time interval since most recent purchase. This score is based on a simple ranking of recency values into a small number of categories. For example, if you use five categories, the customers with the most recent purchase dates receive a recency ranking of 5, and those with purchase dates furthest in the past receive a recency ranking of 1.
- In a similar fashion, customers are then assigned a frequency ranking, with higher values representing a higher frequency of purchases. For example, in a five category ranking scheme, customers who purchase most often receive a frequency ranking of 5.
- Finally, customers are ranked by monetary value, with the highest monetary values receiving the highest ranking. Continuing the five-category example, customers who have spent the most would receive a monetary ranking of 5.

The result is four scores for each customer: recency, frequency, monetary, and combined RFM score, which is simply the three individual scores concatenated into a single value. The “best” customers (those most likely to respond to an offer) are those with the highest combined RFM scores. For example, in a five-category ranking, there is a total of 125 possible combined RFM scores, and the highest combined RFM score is 555.

Data Considerations

- If data rows represent transactions (each row represents a single transaction, and there may be multiple transactions for each customer), use RFM from Transactions. For more information, see the topic RFM Scores from Transaction Data on p. 3.
- If data rows represent customers with summary information for all transactions (with columns that contain values for total amount spent, total number of transactions, and most recent transaction date), use RFM from Customer Data. For more information, see the topic RFM Scores from Customer Data on p. 4.

Figure 2-1
Transaction vs. customer data



Rows are Transactions

ID	Gender	Date	Amount
1	Male	9/25/2005	21
2	Male	1/15/2006	297
4	Male	2/5/2006	249
4	Male	5/7/2005	172
6	Male	4/16/2005	164
6	Male	4/12/2005	286
7	Femal		
9	Male		
9	Male		
9	Male		
10	Femal		
10	Femal		

Rows are Customers with Transaction Summaries

			Total	Number of
ID	Gender	Most Recent	Amount	Purchases
1	Male	9/25/2005	21	1
2	Male	1/15/2006	297	1
4	Male	2/5/2006	421	2
6	Male	4/16/2005	450	2

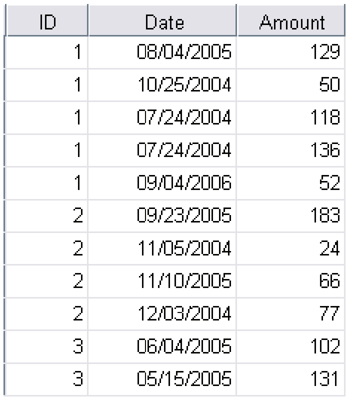
RFM Scores from Transaction Data

Data Considerations

The dataset must contain variables that contain the following information:

- A variable or combination of variables that identify each case (customer).
- A variable with the date of each transaction.
- A variable with the monetary value of each transaction.

Figure 2-2
RFM transaction data



ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

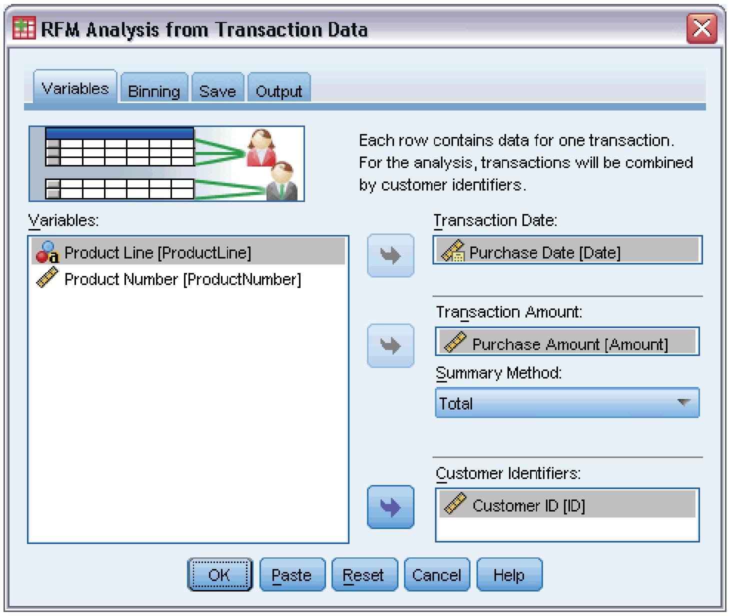
Creating RFM Scores from Transaction Data

▶ From the menus choose:
Direct Marketing > Choose Technique

▶ Select Help identify my best contacts (RFM Analysis) and click Continue.

▶ Select Transaction data and click Continue.

Figure 2-3
Transactions data, Variables tab



Each row contains data for one transaction. For the analysis, transactions will be combined by customer identifiers.

Variables:
Product Line [ProductLine]
Product Number [ProductNumber]

Transaction Date: Purchase Date [Date]

Transaction Amount: Purchase Amount [Amount]

Summary Method: Total

Customer Identifiers: Customer ID [ID]

▶ Select the variable that contains transaction dates.

▶ Select the variable that contains the monetary amount for each transaction.

▶ Select the method for summarizing transaction amounts for each customer: Total (sum of all transactions), mean, median, or maximum (highest transaction amount).

▶ Select the variable or combination of variables that uniquely identifies each customer. For example, cases could be identified by a unique ID code or a combination of last name and first name.

RFM Scores from Customer Data

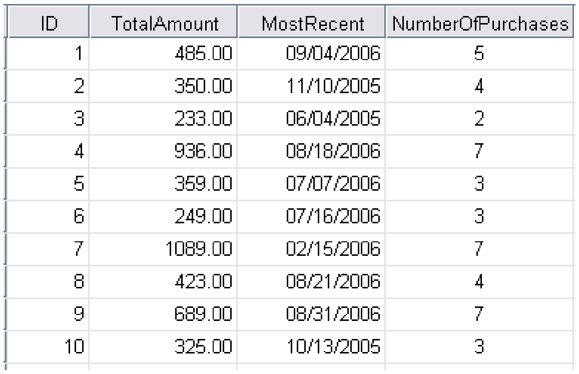
Data Considerations

The dataset must contain variables that contain the following information:

- Most recent purchase date or a time interval since the most recent purchase date. This will be used to compute recency scores.

- Total number of purchases. This will be used to compute frequency scores.
- Summary monetary value for all purchases. This will be used to compute monetary scores. Typically, this is the sum (total) of all purchases, but it could be the mean (average), maximum (largest amount), or other summary measure.

Figure 2-4
RFM customer data



ID	TotalAmount	MostRecent	NumberOfPurchases
1	485.00	09/04/2006	5
2	350.00	11/10/2005	4
3	233.00	06/04/2005	2
4	936.00	08/18/2006	7
5	359.00	07/07/2006	3
6	249.00	07/16/2006	3
7	1089.00	02/15/2006	7
8	423.00	08/21/2006	4
9	689.00	08/31/2006	7
10	325.00	10/13/2005	3

If you want to write RFM scores to a new dataset, the active dataset must also contain a variable or combination of variables that identify each case (customer).

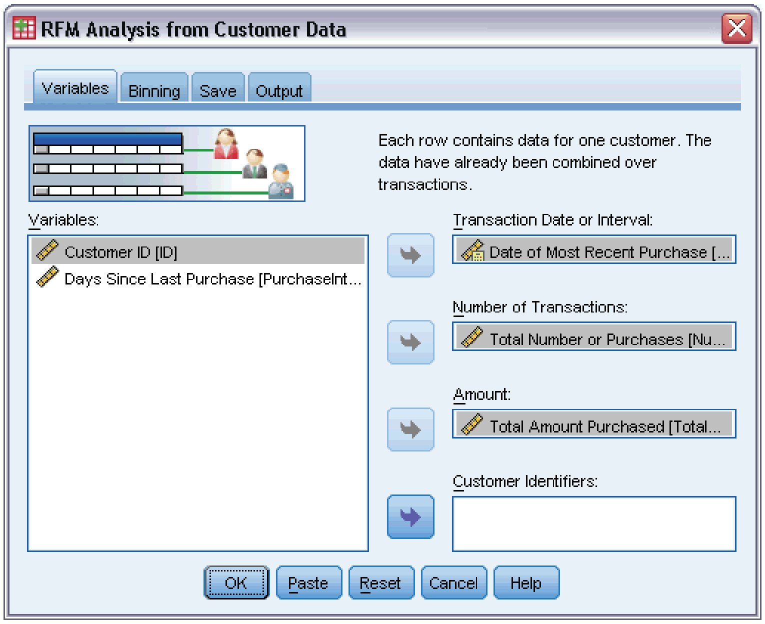
Creating RFM Scores from Customer Data

▶ From the menus choose:
Direct Marketing > Choose Technique

▶ Select Help identify my best contacts (RFM Analysis) and click Continue.

▶ Select Customer data and click Continue.

Figure 2-5
Customer data, Variables tab



RFM Analysis from Customer Data

Each row contains data for one customer. The data have already been combined over transactions.

Variables:
- Customer ID [ID]
- Days Since Last Purchase [PurchaseInt...

Transaction Date or Interval: Date of Most Recent Purchase [...

Number of Transactions: Total Number or Purchases [Nu...

Amount: Total Amount Purchased [Total...

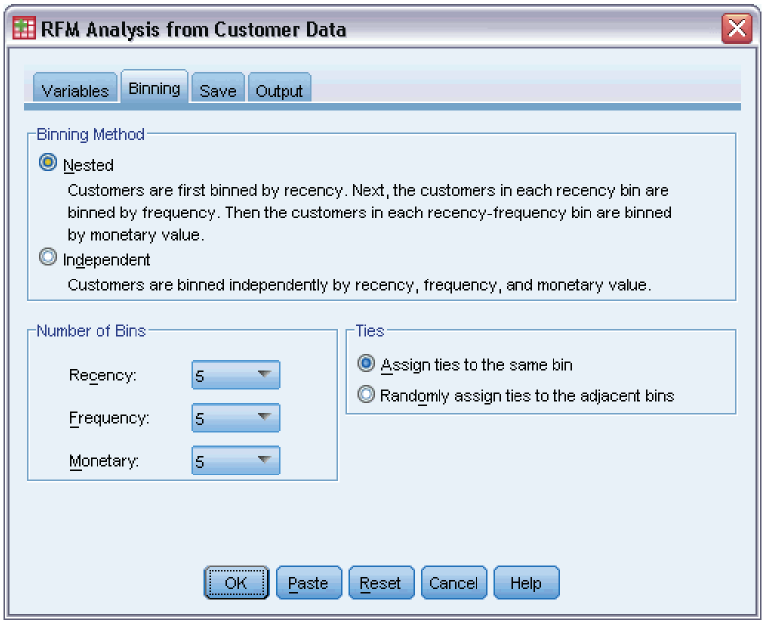
Customer Identifiers:

- ▶ Select the variable that contains the most recent transaction date or a number that represents a time interval since the most recent transaction.
- ▶ Select the variable that contains the total number of transactions for each customer.
- ▶ Select the variable that contains the summary monetary amount for each customer.
- ▶ If you want to write RFM scores to a new dataset, select the variable or combination of variables that uniquely identifies each customer. For example, cases could be identified by a unique ID code or a combination of last name and first name.

RFM Binning

The process of grouping a large number of numeric values into a small number of categories is sometimes referred to as **binning**. In RFM analysis, the bins are the ranked categories. You can use the Binning tab to modify the method used to assign recency, frequency, and monetary values to those bins.

Figure 2-6
RFM Binning tab



Binning Method

● Nested

Customers are first binned by recency. Next, the customers in each recency bin are binned by frequency. Then the customers in each recency-frequency bin are binned by monetary value.

○ Independent

Customers are binned independently by recency, frequency, and monetary value.

Number of Bins

Recency: 5

Frequency: 5

Monetary: 5

Ties

● Assign ties to the same bin

○ Randomly assign ties to the adjacent bins

Binning Method

Nested. In nested binning, a simple rank is assigned to recency values. Within each recency rank, customers are then assigned a frequency rank, and within each frequency rank, customer are assigned a monetary rank. This tends to provide a more even distribution of combined RFM scores, but it has the disadvantage of making frequency and monetary rank scores more difficult to interpret. For example, a frequency rank of 5 for a customer with a recency rank of 5 may not mean the same thing as a frequency rank of 5 for a customer with a recency rank of 4, since the frequency rank is dependent on the recency rank.

Independent. Simple ranks are assigned to recency, frequency, and monetary values. The three ranks are assigned independently. The interpretation of each of the three RFM components is therefore unambiguous; a frequency score of 5 for one customer means the same as a frequency score of 5 for another customer, regardless of their recency scores. For smaller samples, this has the disadvantage of resulting in a less even distribution of combined RFM scores.

Number of Bins

The number of categories (bins) to use for each component to create RFM scores. The total number of possible combined RFM scores is the product of the three values. For example, 5 recency bins, 4 frequency bins, and 3 monetary bins would create a total of 60 possible combined RFM scores, ranging from 111 to 543.

- The default is 5 for each component, which will create 125 possible combined RFM scores, ranging from 111 to 555.
- The maximum number of bins allowed for each score component is nine.

Ties

A "tie" is simply two or more equal recency, frequency, or monetary values. Ideally, you want to have approximately the same number of customers in each bin, but a large number of tied values can affect the bin distribution. There are two alternatives for handling ties:

- **Assign ties to the same bin**. This method always assigns tied values to the same bin, regardless of how this affects the bin distribution. This provides a consistent binning method: If two customers have the same recency value, then they will always be assigned the same recency score. In an extreme example, however, you might have 1,000 customers, with 500 of them making their most recent purchase on the same date. In a 5-bin ranking, 50% of the customers would therefore receive a recency score of 5, instead of the desired 20%.

  Note that with the nested binning method "consistency" is somewhat more complicated for frequency and monetary scores, since frequency scores are assigned within recency score bins, and monetary scores are assigned within frequency score bins. So two customers with the same frequency value may not have the same frequency score if they don't also have the same recency score, regardless of how tied values are handled.

- **Randomly assign ties.** This ensures an even bin distribution by assigning a very small random variance factor to ties prior to ranking; so for the purpose of assigning values to the ranked bins, there are no tied values. This process has no effect on the original values. It is only used to disambiguate ties. While this produces an even bin distribution (approximately the same number of customers in each bin), it can result in completely different score results for customers who appear to have similar or identical recency, frequency, and/or monetary values — particularly if the total number of customers is relatively small and/or the number of ties is relatively high.

Table 2-1
Assign Ties to Same Bin vs. Randomly Assign Ties

ID	Most Recent Purchase (Recency)	Recency Ranking	
		Assign Ties to Same Bin	Randomly Assign Ties
1	10/29/2006	5	5
2	10/28/2006	4	4
3	10/28/2006	4	4
4	10/28/2006	4	5
5	10/28/2006	4	3
6	9/21/2006	3	3
7	9/21/2006	3	2
8	8/13/2006	2	2
9	8/13/2006	2	1
10	6/20/2006	1	1

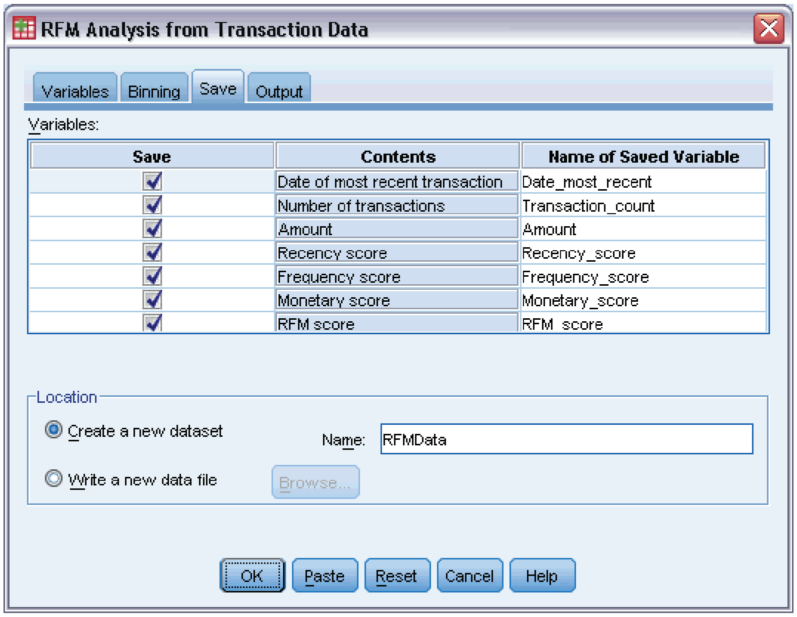
- In this example, assigning ties to the same bin results in an uneven bin distribution: 5 (10%), 4 (40%), 3 (20%), 2 (20%), 1 (10%).
- Randomly assigning ties results in 20% in each bin, but to achieve this result the four cases with a date value of 10/28/2006 are assigned to 3 different bins, and the 2 cases with a date value of 8/13/2006 are also assigned to different bins.

Note that the manner in which ties are assigned to different bins is entirely random (within the constraints of the end result being an equal number of cases in each bin). If you computed a second set of scores using the same method, the ranking for any particular case with a tied value could change. For example, the recency rankings of 5 and 3 for cases 4 and 5 respectively might be switched the second time.

Saving RFM Scores from Transaction Data

RFM from Transaction Data always creates a new aggregated dataset with one row for each customer. Use the Save tab to specify what scores and other variables you want to save and where you want to save them.

Figure 2-7
Transaction data, Save tab



Variables:

Save	Contents	Name of Saved Variable
☑	Date of most recent transaction	Date_most_recent
☑	Number of transactions	Transaction_count
☑	Amount	Amount
☑	Recency score	Recency_score
☑	Frequency score	Frequency_score
☑	Monetary score	Monetary_score
☑	RFM score	RFM_score

Location

● Create a new dataset Name: RFMData

○ Write a new data file Browse...

Variables

The ID variables that uniquely identify each customer are automatically saved in the new dataset. The following additional variables can be saved in the new dataset:

- **Date of most recent transaction for each customer.**
- **Number of transactions.** The total number of transaction rows for each customer.
- **Amount.** The summary amount for each customer based on the summary method you select on the Variables tab.
- **Recency score.** The score assigned to each customer based on most recent transaction date. Higher scores indicate more recent transaction dates.
- **Frequency score.** The score assigned to each customer based on total number of transactions. Higher scores indicate more transactions.

- **Monetary score.** The score assigned to each customer based on the selected monetary summary measure. Higher scores indicate a higher value for the monetary summary measure.
- **RFM score.** The three individual scores combined into a single value: *(recency x 100) + (frequency x 10) + monetary.*

By default all available variables are included in the new dataset; so deselect (uncheck) the ones you don't want to include. Optionally, you can specify your own variable names. Variable names must conform to standard variable naming rules.

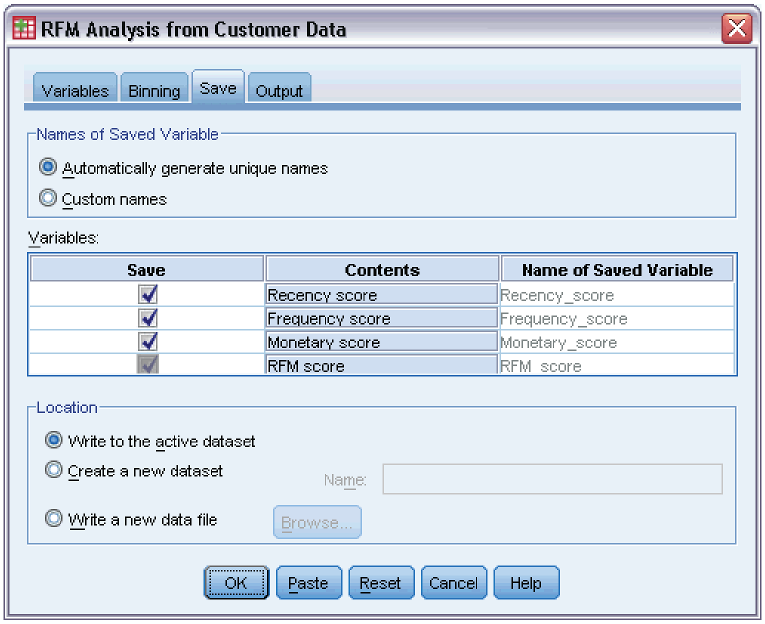
Location

RFM from Transaction Data always creates a new aggregated dataset with one row for each customer. You can create a new dataset in the current session or save the RFM score data in an external data file. Dataset names must conform to standard variable naming rules. (This restriction does not apply to external data file names.)

Saving RFM Scores from Customer Data

For customer data, you can add the RFM score variables to the active dataset or create a new dataset that contains the selected scores variables. Use the Save Tab to specify what score variables you want to save and where you want to save them.

Figure 2-8
Customer data, Save tab



Names of Saved Variable

● Automatically generate unique names

○ Custom names

Variables:

Save	Contents	Name of Saved Variable
☑	Recency score	Recency_score
☑	Frequency score	Frequency_score
☑	Monetary score	Monetary_score
☑	RFM score	RFM_score

Location

● Write to the active dataset

○ Create a new dataset   Name:

○ Write a new data file   Browse...

Names of Saved Variables

- **Automatically generate unique names.** When adding score variables to the active dataset, this ensures that new variable names are unique. This is particularly useful if you want to add multiple different sets of RFM scores (based on different criteria) to the active dataset.
- **Custom names.** This allows you to assign your own variable names to the score variables. Variable names must conform to standard variable naming rules.

Variables

Select (check) the score variables that you want to save:

- **Recency score.** The score assigned to each customer based on the value of the Transaction Date or Interval variable selected on the Variables tab. Higher scores are assigned to more recent dates or lower interval values.
- **Frequency score.** The score assigned to each customer based on the Number of Transactions variable selected on the Variables tab. Higher scores are assigned to higher values.
- **Monetary score.** The score assigned to each customer based on the Amount variable selected on the Variables tab. Higher scores are assigned to higher values.
- **RFM score.** The three individual scores combined into a single value: *(recency*100)+(frequency*10)+monetary.*

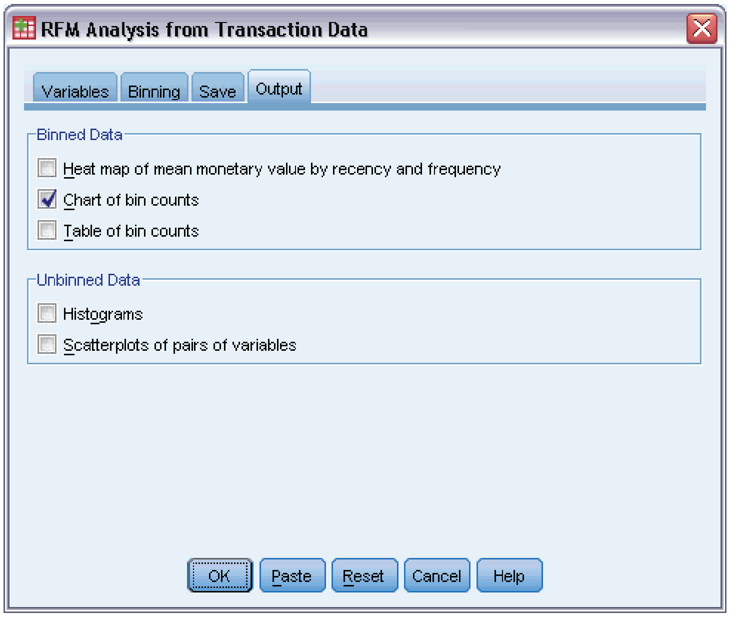
Location

For customer data, there are three alternatives for where you can save new RFM scores:

- **Active dataset.** Selected RFM score variables are added to active dataset.
- **New Dataset.** Selected RFM score variables and the ID variables that uniquely identify each customer (case) will be written to a new dataset in the current session. Dataset names must conform to standard variable naming rules. This option is only available if you select one or more Customer Identifier variables on the Variables tab.
- **File.** Selected RFM scores and the ID variables that uniquely identify each customer (case) will be saved in an external data file. This option is only available if you select one or more Customer Identifier variables on the Variables tab.

RFM Output

Figure 2-9
RFM Output tab



Binned Data

☐ Heat map of mean monetary value by recency and frequency
☑ Chart of bin counts
☐ Table of bin counts

Unbinned Data

☐ Histograms
☐ Scatterplots of pairs of variables

Binned Data

Charts and tables for binned data are based on the calculated recency, frequency, and monetary scores.

Heat map of mean monetary value by recency and frequency. The heat map of mean monetary distribution shows the average monetary value for categories defined by recency and frequency scores. Darker areas indicate a higher average monetary value.

Chart of bin counts. The chart of bin counts displays the bin distribution for the selected binning method. Each bar represents the number of cases that will be assigned each combined RFM score.

- Although you typically want a fairly even distribution, with all (or most) bars of roughly the same height, a certain amount of variance should be expected when using the default binning method that assigns tied values to the same bin.
- Extreme fluctuations in bin distribution and/or many empty bins may indicate that you should try another binning method (fewer bins and/or random assignment of ties) or reconsider the suitability of RFM analysis.

Table of bin counts. The same information that is in the chart of bin counts, except expressed in the form of a table, with bin counts in each cell.

Unbinned Data

Chart and tables for unbinned data are based on the original variables used to create recency, frequency, and monetary scores.

Histograms. The histograms show the relative distribution of values for the three variables used to calculate recency, frequency, and monetary scores. It is not unusual for these histograms to indicate somewhat skewed distributions rather than a normal or symmetrical distribution.

The horizontal axis of each histogram is always ordered from low values on the left to high values on the right. With recency, however, the interpretation of the chart depends on the type of recency measure: date or time interval. For dates, the bars on the left represent values further in the past (a less recent date has a lower value than a more recent date). For time intervals, the bars on the left represent more recent values (the smaller the time interval, the more recent the transaction).

Scatterplots of pairs of variables. These scatterplots show the relationships between the three variables used to calculate recency, frequency, and monetary scores.

It's common to see noticeable linear groupings of points on the frequency scale, since frequency often represents a relatively small range of discrete values. For example, if the total number of transactions doesn't exceed 15, then there are only 15 possible frequency values (unless you count fractional transactions), whereas there could by hundreds of possible recency values and thousands of monetary values.

The interpretation of the recency axis depends on the type of recency measure: date or time interval. For dates, points closer to the origin represent dates further in the past. For time intervals, points closer to the origin represent more recent values.

Chapter

3

Cluster analysis

Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.

Example. Retail and consumer product companies regularly apply clustering techniques to data that describe their customers' buying habits, gender, age, income level, etc. These companies tailor their marketing and product development strategies to each consumer group to increase sales and build brand loyalty.

Cluster Analysis data considerations

Data. This procedure works with both continuous and categorical fields. Each record (row) represent a customer to be clustered, and the fields (variables) represent attributes upon which the clustering is based.

Record order. Note that the results may depend on the order of records. To minimize order effects, you may want to consider randomly ordering the records. You may want to run the analysis several times, with records sorted in different random orders to verify the stability of a given solution.

Measurement level. Correct measurement level assignment is important because it affects the computation of the results.

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.
- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- **Continuous.** A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

An icon next to each field indicates the current measurement level.

Measurement Level	Data Type			
	Numeric	String	Date	Time

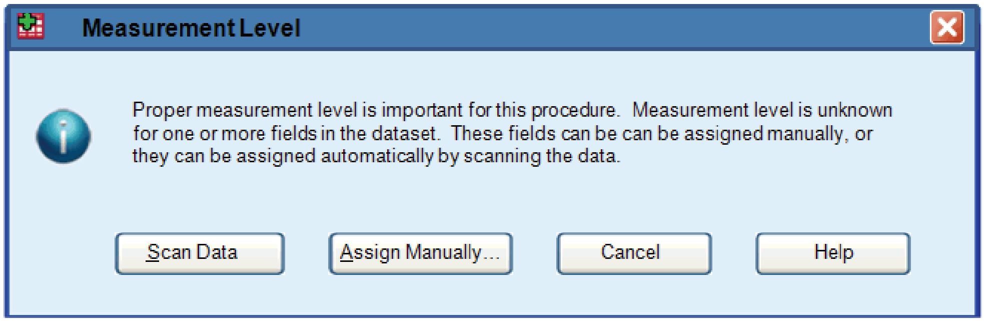
Scale (Continuous)	📏	n/a	📏📅	📏🕒
Ordinal	📊	📊a	📊📅	📊🕒
Nominal	●	●a	●📅	●🕒

You can change the measurement level in Variable View of the Data Editor or you can use the Define Variable Properties dialog to suggest an appropriate measurement level for each field.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 3-1
Measurement level alert



Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be can be assigned manually, or they can be assigned automatically by scanning the data.

- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.
- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

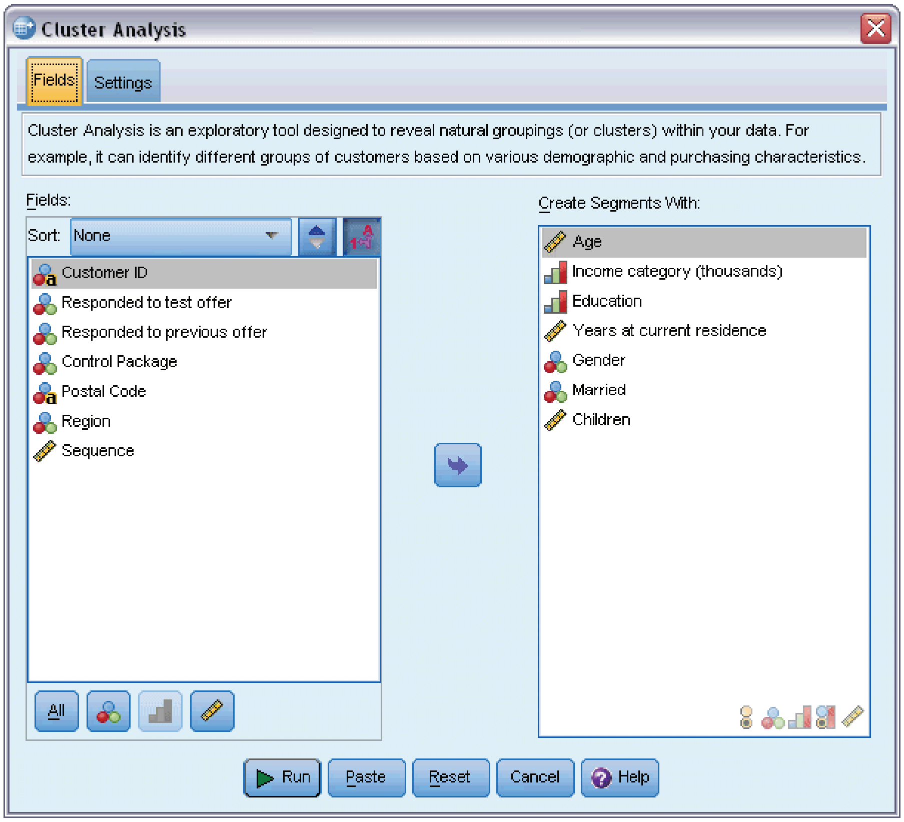
To obtain Cluster Analysis

From the menus choose:

Direct Marketing > Choose Technique

▶ Select Segment my contacts into clusters.

Figure 3-2
Cluster Analysis Fields tab



Cluster Analysis

Fields | Settings

Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.

Fields:

Sort: None

- Customer ID
- Responded to test offer
- Responded to previous offer
- Control Package
- Postal Code
- Region
- Sequence

Create Segments With:

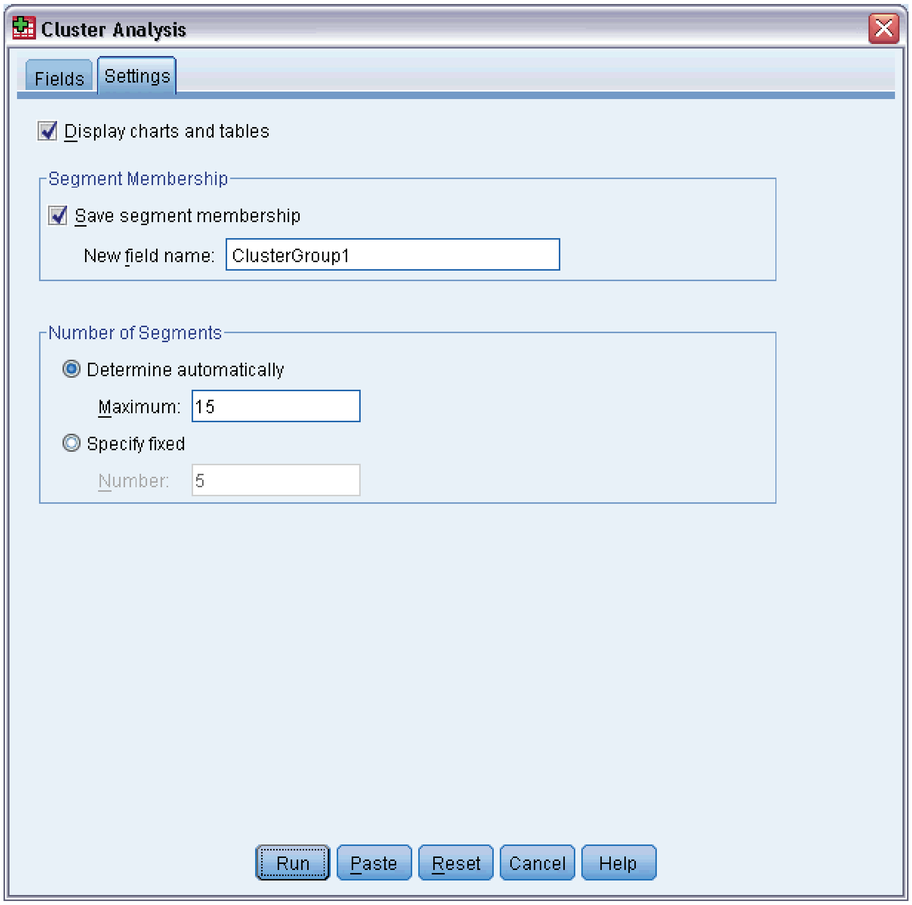
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Children

▶ Select the categorical (nominal, ordinal) and continuous (scale) fields that you want to use to create segments.

▶ Click Run to run the procedure.

Settings

Figure 3-3
Cluster Analysis Settings tab



☑ Display charts and tables

Segment Membership

☑ Save segment membership

New field name: ClusterGroup1

Number of Segments

● Determine automatically

Maximum: 15

○ Specify fixed

Number: 5

The Settings tab allows you to show or suppress display of charts and tables that describe the segments, save a new field in the dataset that identifies the segment (cluster) for each record in the dataset, and specify how many segments to include in the cluster solution.

Display charts and tables. Displays tables and charts that describe the segments.

Segment Membership. Saves a new field (variable) that identifies the segment to which each record belongs.

- Field names must conform to IBM® SPSS® Statistics naming rules.
- The segment membership field name cannot duplicate a field name that already exists in the dataset. If you run this procedure more than once on the same dataset, you will need to specify a different name each time.
- **Number of Segments.** Controls how the number of segments is determined.
- **Determine automatically.** The procedure will automatically determine the “best” number of segments, up to the specified maximum.

Specify fixed. The procedure will produce the specified number of segments.

Chapter
4
Prospect profiles

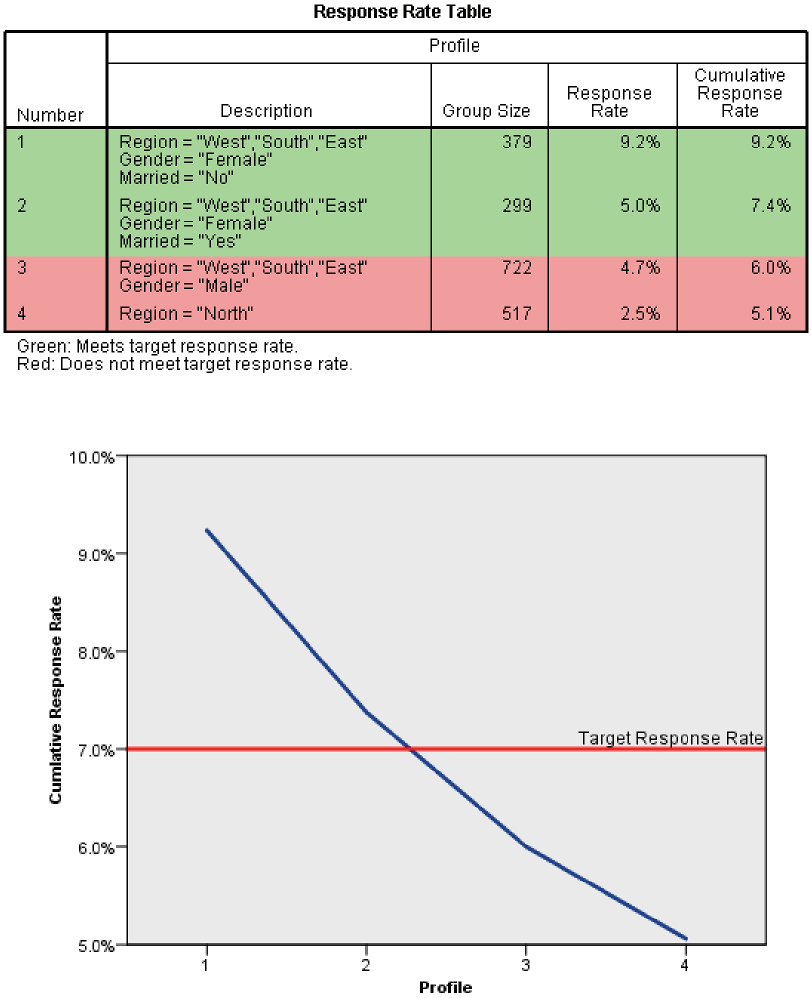
This technique uses results from a previous or test campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns. The Response field indicates who responded to the previous or test campaign. The Profiles list contains the characteristics that you want to use to create the profile.

Example. Based on the results of a test mailing, the direct marketing division of a company wants to generate profiles of the types of customers most likely to respond to an offer, based on demographic information.

Output

Output includes a table that provides a description of each profile group and displays response rates (percentage of positive responses) and cumulative response rates and a chart of cumulative response rates. If you include a target minimum response rate, the table will be color-coded to show which profiles meet the minimum cumulative response rate, and the chart will include a reference line at the specified minimum response rate value.

Figure 4-1
Response rate table and chart



Response Rate Table

Number	Profile			
	Description	Group Size	Response Rate	Cumulative Response Rate

1	Region = "West","South","East"
	Gender = "Female"
	Married = "No"
		379	9.2%	9.2%

2	Region = "West","South","East"
	Gender = "Female"
	Married = "Yes"
		299	5.0%	7.4%

3	Region = "West","South","East"
	Gender = "Male"
		722	4.7%	6.0%

4	Region = "North"	517	2.5%	5.1%

Green: Meets target response rate.
Red: Does not meet target response rate.

Prospect Profiles data considerations

Response Field. The response field must be nominal or ordinal. It can be string or numeric. If this field contains a value that indicates number or amount of purchases, you will need to create a new field in which a single value represents all positive responses. For more information, see the topic Creating a categorical response field on p. 24.

Positive response value. The positive response value identifies customers who responded positively (for example, made a purchase). All other non-missing response values are assumed to indicate a negative response. If there are any defined value labels for the response field, those labels are displayed in the drop-down list.

Create Profiles with. These fields can be nominal, ordinal, or continuous (scale). They can be string or numeric.

Measurement level. Correct measurement level assignment is important because it affects the computation of the results.

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.
- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- **Continuous.** A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

An icon next to each field indicates the current measurement level.

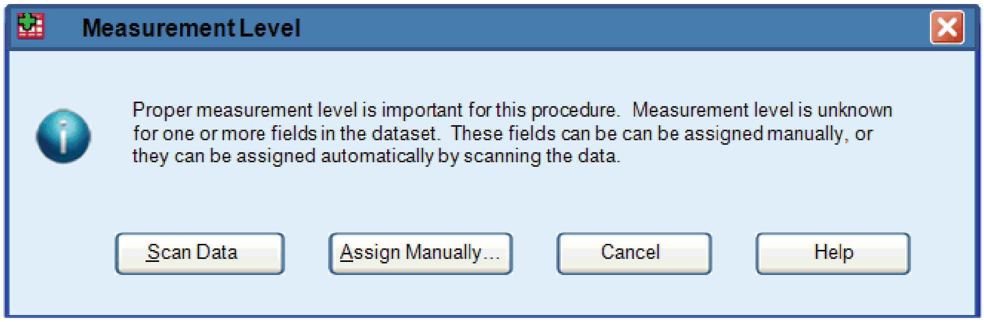
Measurement Level	**Data Type**			
	Numeric	String	Date	Time
Scale (Continuous)	📏	n/a	📏📅	📏🕒
Ordinal	📊	📊a	📊📅	📊🕒
Nominal	●	●a	●📅	●🕒

You can change the measurement level in Variable View of the Data Editor or you can use the Define Variable Properties dialog to suggest an appropriate measurement level for each field.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 4-2
Measurement level alert



Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be can be assigned manually, or they can be assigned automatically by scanning the data.

- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.
- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

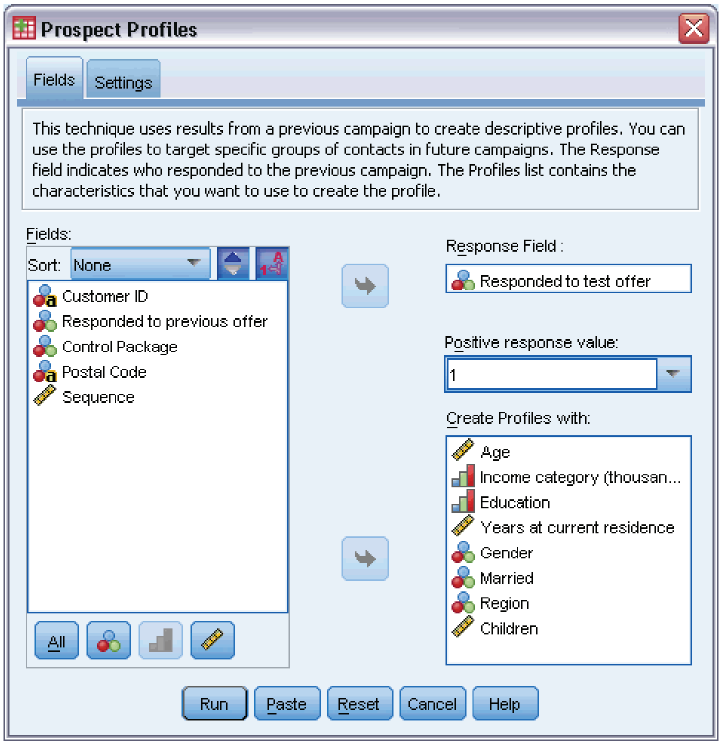
To obtain prospect profiles

From the menus choose:

Direct Marketing > Choose Technique

▶ Select Generate profiles of my contacts who responded to an offer.

Figure 4-3
Prospect Profiles Fields tab



This technique uses results from a previous campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns. The Response field indicates who responded to the previous campaign. The Profiles list contains the characteristics that you want to use to create the profile.

Fields:
Sort: None
Customer ID
Responded to previous offer
Control Package
Postal Code
Sequence

Response Field: Responded to test offer

Positive response value: 1

Create Profiles with:
Age
Income category (thousan...
Education
Years at current residence
Gender
Married
Region
Children

▶ Select the field that identifies which contacts responded to the offer. This field must be nominal or ordinal.

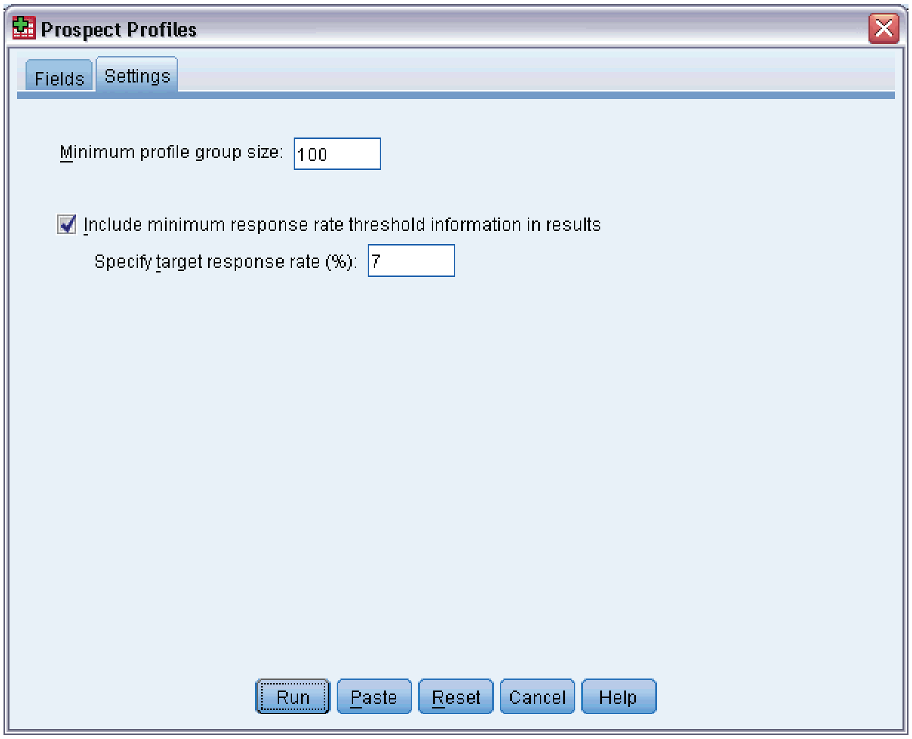
▶ Enter the value that indicates a positive response. If any values have defined value labels, you can select the value label from the drop-down list, and the corresponding value will be displayed.

▶ Select the fields you want to use to create the profiles.

▶ Click Run to run the procedure.

Settings

Figure 4-4
Prospect Profiles Settings tab



Minimum profile group size: 100

☑ Include minimum response rate threshold information in results

Specify target response rate (%): 7

The Settings tab allows you to control the minimum profile group size and include a minimum response rate threshold in the output.

Minimum profile group size. Each profile represents the shared characteristics of a group of contacts in the dataset (for example, females under 40 who live in the west). By default, the smallest profile group size is 100. Smaller group sizes may reveal more groups, but larger group sizes may provide more reliable results. The value must be a positive integer.

Include minimum response rate threshold information in results. Results include a table that displays response rates (percentage of positive responses) and cumulative response rates and a chart of cumulative response rates. If you enter a target minimum response rate, the table will be color-coded to show which profiles meet the minimum cumulative response rate, and the chart will include a reference line at the specified minimum response rate value. The value must be greater than 0 and less than 100.

Creating a categorical response field

The response field should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values.

- If negative responses are recorded as 0 (not blank, which is treated as missing), this can be computed with the following formula:

  NewName=OldName>0

  where *NewName* is the name of the new field and *OldName* is the name of the original field. This is a logical expression that assigns a value of 1 to all non-missing values greater than 0, and 0 to all non-missing values less than or equal to 0.

- If no value is recorded for negative responses, then these values are treated as missing, and the formula is a little more complicated:

  NewName=NOT(MISSING(OldName))

  In this logical expression, all non-missing response values are assigned a value of 1 and all missing response values are assigned a value of 0.

- If you cannot distinguish between negative (0) response values and missing values, then an accurate response value cannot be computed. If there are relatively few truly missing values, this may not have a significant effect on the computed response rates. If, however, there are many missing values — such as when response information is recorded for only a small test sample of the total dataset — then the computed response rates will be meaningless, since they will be significantly lower than the true response rates.

To Create a Categorical Response Field

▶ From the menus choose:

Transform > Compute Variable

▶ For Target Variable, enter the new field (variable) name.

▶ If negative responses are recorded as 0, for the Numeric Expression enter OldName>0, where *OldName* is the original field name.

▶ If negative responses are recorded as missing (blank), for the Numeric Expression enter NOT(MISSING(OldName)), where *OldName* is the original field name.

Chapter
5
Postal Code Response Rates

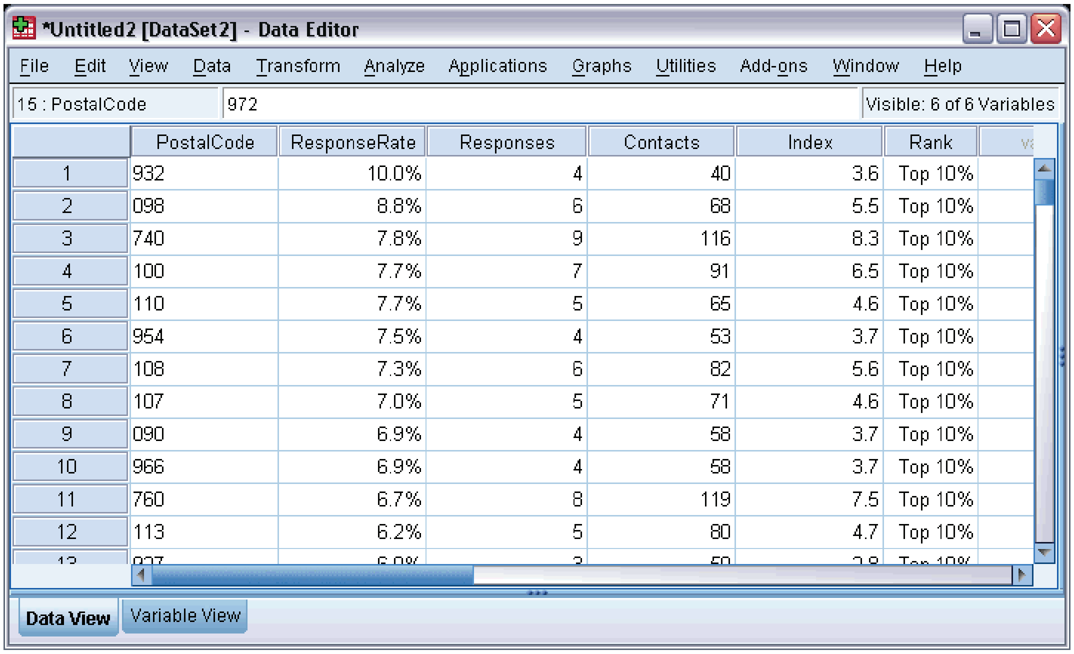
This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns. The Response field indicates who responded to the previous campaign. The Postal Code field identifies the field that contains the postal codes.

Example. Based on the results of a previous mailing, the direct marketing division of a company generates response rates by postal codes. Based on various criteria, such as a minimum acceptable response rate and/or maximum number of contacts to include in the mailing, they can then target specific postal codes.

Output

Output from this procedure includes a new dataset that contains response rates by postal code, and a table and chart that summarize the results by decile rank (top 10%, top 20%, etc.). The table can be color-coded based on a user-specified minimum cumulative response rate or maximum number of contacts.

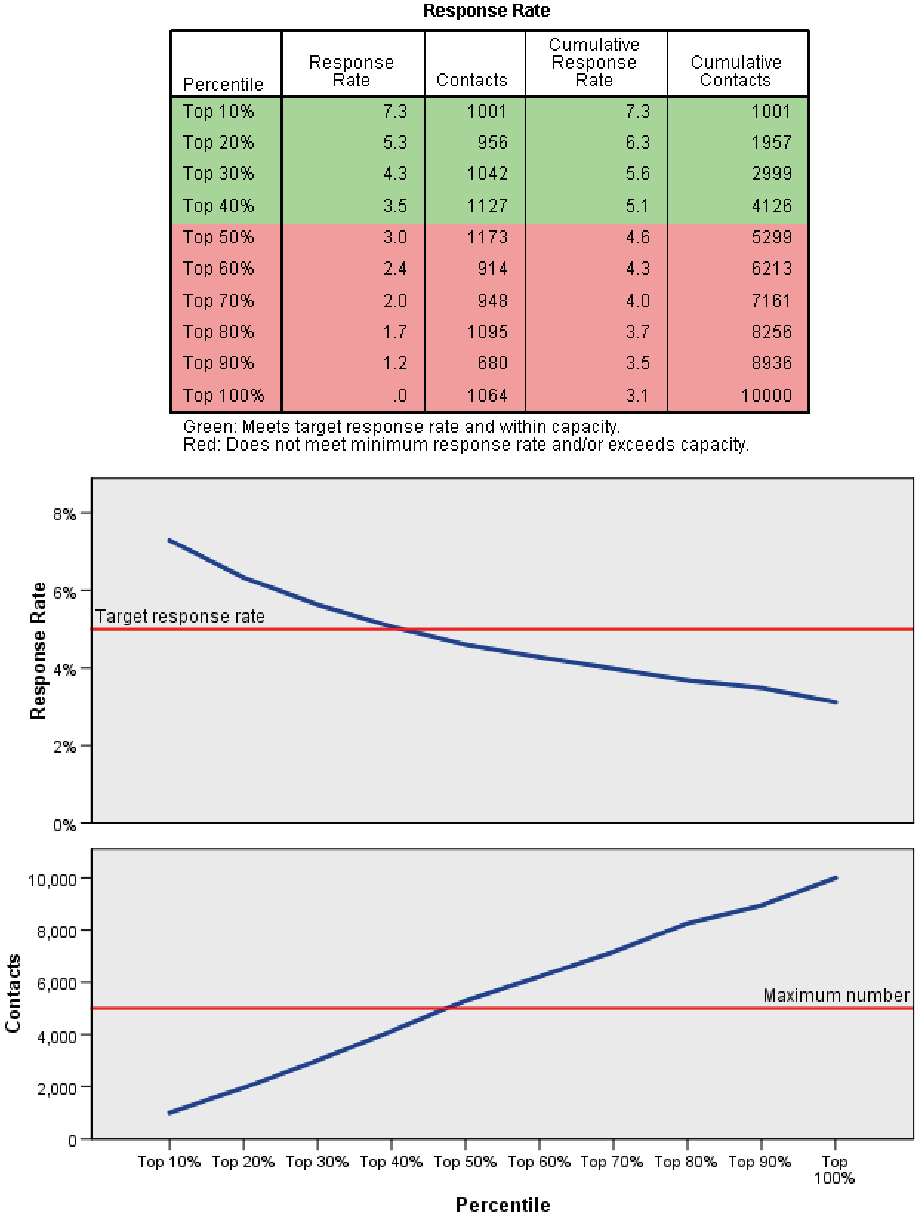
Figure 5-1
Dataset with response rates by postal code



15 : PostalCode 972 Visible: 6 of 6 Variables

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
1	932	10.0%	4	40	3.6	Top 10%
2	098	8.8%	6	68	5.5	Top 10%
3	740	7.8%	9	116	8.3	Top 10%
4	100	7.7%	7	91	6.5	Top 10%
5	110	7.7%	5	65	4.6	Top 10%
6	954	7.5%	4	53	3.7	Top 10%
7	108	7.3%	6	82	5.6	Top 10%
8	107	7.0%	5	71	4.6	Top 10%
9	090	6.9%	4	58	3.7	Top 10%
10	966	6.9%	4	58	3.7	Top 10%
11	760	6.7%	8	119	7.5	Top 10%
12	113	6.2%	5	80	4.7	Top 10%
13	927	6.0%	3	50	2.8	Top 10%

Figure 5-2
Summary table and chart



Response Rate

Percentile	Response Rate	Contacts	Cumulative Response Rate	Cumulative Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Green: Meets target response rate and within capacity.
Red: Does not meet minimum response rate and/or exceeds capacity.

The new dataset contains the following fields:

- **Postal code.** If postal code groups are based on only a portion of the complete value, then this is the value of that portion of the postal code. The header row label for this column in the Excel file is the name of the postal code field in the original dataset.
- **ResponseRate.** The percentage of positive responses in each postal code.
- **Responses.** The number of positive responses in each postal code.

- **Contacts.** The total number of contacts in each postal code that contain a non-missing value for the response field.
- **Index.** The "weighted" response based on the formula *N x P x (1-P)*, where *N* is the number of contacts, and *P* is the response rate expressed as a proportion.
- **Rank.** Decile rank (top 10%, top 20% , etc.) of the cumulative postal code response rates in descending order.

Postal Code Response Rates Data Considerations

Response Field. The response field can be string or numeric. If this field contains a value that indicates number or monetary value of purchases, you will need to create a new field in which a single value represents all positive responses. For more information, see the topic Creating a Categorical Response Field on p. 31.

Positive response value. The positive response value identifies customers who responded positively (for example, made a purchase). All other non-missing response values are assumed to indicate a negative response. If there are any defined value labels for the response field, those labels are displayed in the drop-down list.

Postal Code Field. The postal code field can be string or numeric.

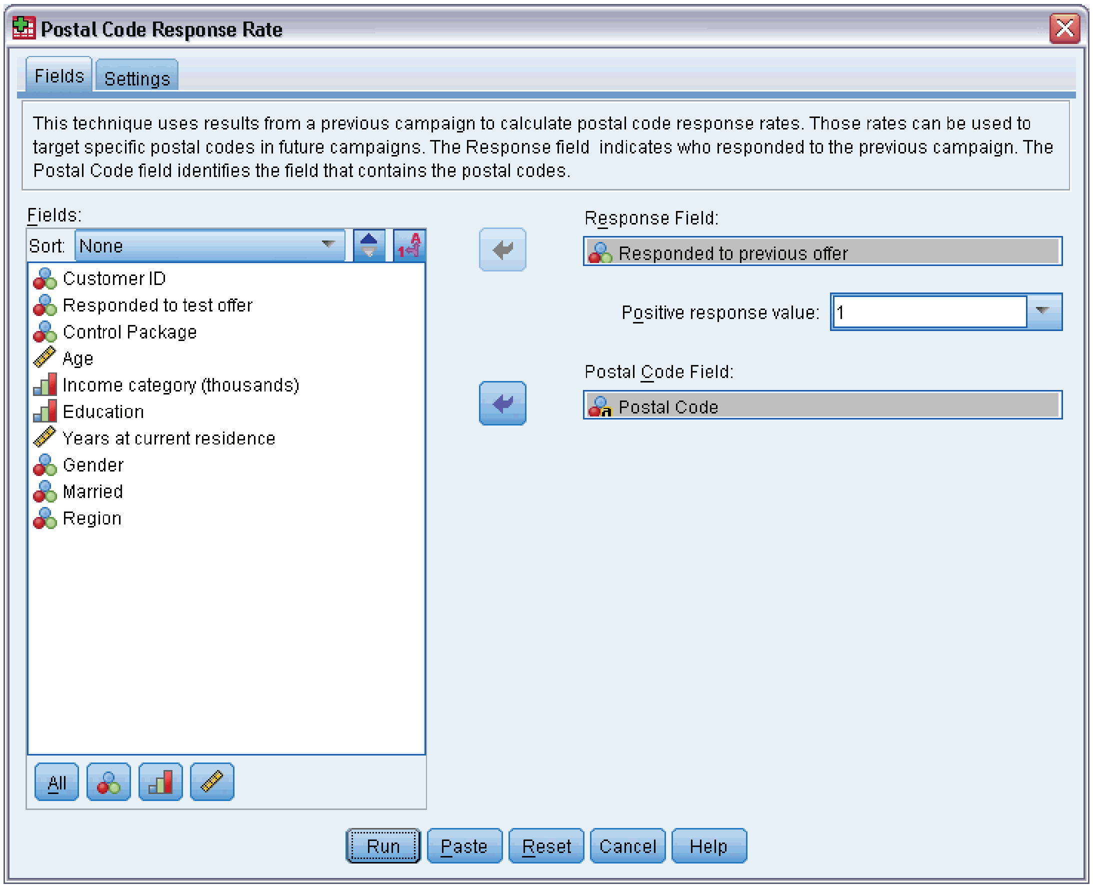
To Obtain Postal Code Response Rates

From the menus choose:

Direct Marketing > Choose Technique

▶ Select Identify top responding postal codes.

Figure 5-3
Postal Code Response Rates Fields tab



Postal Code Response Rate

This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns. The Response field indicates who responded to the previous campaign. The Postal Code field identifies the field that contains the postal codes.

Fields:
Sort: None
- Customer ID
- Responded to test offer
- Control Package
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Region

Response Field: Responded to previous offer

Positive response value: 1

Postal Code Field: Postal Code

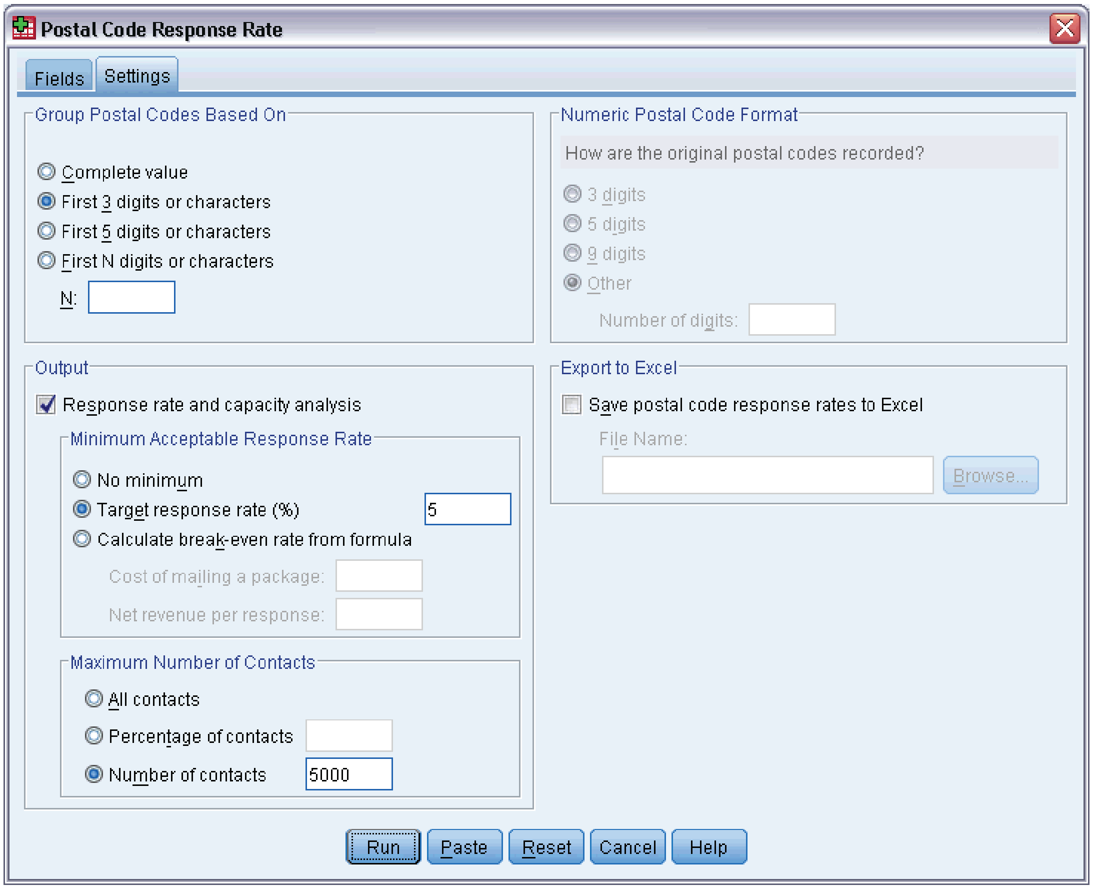
- ▶ Select the field that identifies which contacts responded to the offer.
- ▶ Enter the value that indicates a positive response. If any values have defined value labels, you can select the value label from the drop-down list, and the corresponding value will be displayed.
- ▶ Select the field that contains the postal code.
- ▶ Click Run to run the procedure.

Optionally, you can also:

- Generate response rates based on the first *n* characters or digits of the postal code instead of the complete value
- Automatically save the results to an Excel file
- Control output display options

Settings

Figure 5-4
Postal Code Response Rates Settings tab



Postal Code Response Rate

Fields | Settings

Group Postal Codes Based On

○ Complete value
● First 3 digits or characters
○ First 5 digits or characters
○ First N digits or characters

N:

Numeric Postal Code Format

How are the original postal codes recorded?

○ 3 digits
○ 5 digits
○ 9 digits
● Other

Number of digits:

Output

☑ Response rate and capacity analysis

Minimum Acceptable Response Rate

○ No minimum
● Target response rate (%) 5
○ Calculate break-even rate from formula

Cost of mailing a package:

Net revenue per response:

Maximum Number of Contacts

○ All contacts
○ Percentage of contacts
● Number of contacts 5000

Export to Excel

□ Save postal code response rates to Excel

File Name:

Group Postal Codes Based On

This determines how records are grouped to calculate response rates. By default, the entire postal code is used, and all records with the same postal code are grouped together to calculate the group response rate. Alternatively, you can group records based on only a portion of the complete postal code, consisting of the first *n* digits or characters. For example, you might want to group records based on only the first 5 characters of a 10-character postal code or the first three digits of a 5-digit postal code. The output dataset will contain one record for each postal code group. If you enter a value, it must be a positive integer.

Numeric Postal Code Format

If the postal code field is numeric and you want to group postal codes based on the first *n* digits instead of the entire value, you need to specify the number of digits in the original value. The number of digits is the *maximum* possible number of digits in the postal code. For example, if the postal code field contains a mix of 5-digit and 9-digit zip codes, you should specify 9 as the number of digits.

Note: Depending on the display format, some 5-digit zip codes may appear to contain only 4 digits, but there is an implied leading zero.

Output

In addition to the new dataset that contains response rates by postal code, you can display a table and chart that summarize the results by decile rank (top 10%, top 20%, etc.). The table displays response rates, cumulative response rates, number of records, and cumulative number of records in each decile. The chart displays cumulative response rates and cumulative number of records in each decile.

Minimum Acceptable Response Rate. If you enter a target minimum response rate or break-even formula, the table will be color-coded to show which deciles meet the minimum cumulative response rate, and the chart will include a reference line at the specified minimum response rate value.

- **Target response rate.** Response rate expressed as a percerntage (percentage of positive responses in each postal code group). The value must be greater than 0 and less than 100.
- **Calculate break-even rate from formula.** Calculate minimum cumulative response rate based on the formula: *(Cost of mailing a package/Net revenue per response) x 100*. Both values must be positive numbers. The result should be a value greater than 0 and less than 100. For example, if the cost of mailing a package is $0.75 and the net revenue per response is $56, then the minimum response rate is: (0.75/56) x 100 = 1.34%.

Maximum Number of Contacts. If you specify a maximum number of contacts, the table will be color-coded to show which deciles do not exceed the cumulative maximum number of contacts (records) and the chart will include a reference line at that value.

- **Percentage of contacts.** Maximum expressed as percentage. For example, you might want to know the deciles with the highest response rates that contain no more than 50% of all the contacts. The value must be greater than 0 and less than 100.
- **Number of contacts.** Maximum expressed as a number of contacts. For example, if you don't intend to mail out more than 10,000 packages, you could set the value at 10000. The value must be a positive integer (with no grouping symbols).

If you specify both a minimum acceptable response rate and a maximum number of contacts, the color-coding of the table will be based on whichever condition is met first.

Export to Excel

This procedure automatically creates a new dataset that contains response rates by postal code. Each record (row) in the dataset represents a postal code. You can automatically save the same information in an Excel file. This file is saved in Excel 97-2003 format.

Creating a Categorical Response Field

The response field should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values.

- If negative responses are recorded as 0 (not blank, which is treated as missing), this can be computed with the following formula:

  NewName=OldName>0

  where *NewName* is the name of the new field and *OldName* is the name of the original field. This is a logical expression that assigns a value of 1 to all non-missing values greater than 0, and 0 to all non-missing values less than or equal to 0.

- If no value is recorded for negative responses, then these values are treated as missing, and the formula is a little more complicated:

  NewName=NOT(MISSING(OldName))

  In this logical expression, all non-missing response values are assigned a value of 1 and all missing response values are assigned a value of 0.

- If you cannot distinguish between negative (0) response values and missing values, then an accurate response value cannot be computed. If there are relatively few truly missing values, this may not have a significant effect on the computed response rates. If, however, there are many missing values — such as when response information is recorded for only a small test sample of the total dataset — then the computed response rates will be meaningless, since they will be significantly lower than the true response rates.

To Create a Categorical Response Field

▶ From the menus choose:
Transform > Compute Variable

▶ For Target Variable, enter the new field (variable) name.

▶ If negative responses are recorded as 0, for the Numeric Expression enter OldName>0, where *OldName* is the original field name.

▶ If negative responses are recorded as missing (blank), for the Numeric Expression enter NOT(MISSING(OldName)), where *OldName* is the original field name.

Chapter

6

Propensity to purchase

Propensity to Purchase uses results from a test mailing or previous campaign to generate scores. The scores indicate which contacts are most likely to respond. The Response field indicates who replied to the test mailing or previous campaign. The Propensity fields are the characteristics that you want to use to predict the probability that contacts with similar characteristics will respond.

This technique uses binary logistic regression to build a predictive model. The process of building and applying a predictive model has two basic steps:

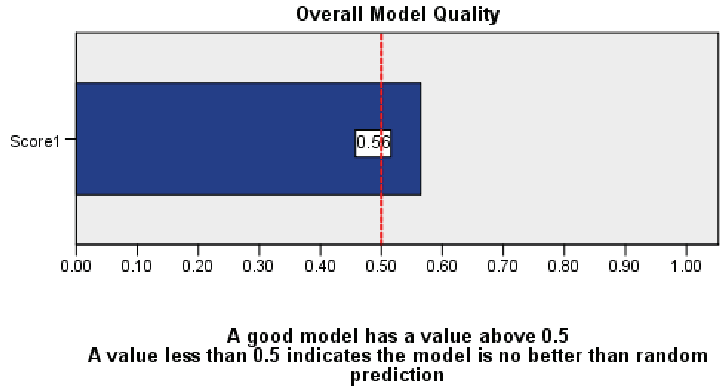
- ▶ Build the model and save the model file. You build the model using a dataset for which the outcome of interest (often referred to as the **target**) is known. For example, if you want to build a model that will predict who is likely to respond to a direct mail campaign, you need to start with a dataset that already contains information on who responded and who did not respond. For example, this might be the results of a test mailing to a small group of customers or information on responses to a similar campaign in the past.
- ▶ Apply that model to a different dataset (for which the outcome of interest is not known) to obtain predicted outcomes.

Example. The direct marketing division of a company uses results from a test mailing to assign propensity scores to the rest of their contact database, using various demographic characteristics to identify contacts most likely to respond and make a purchase.

Output

This procedure automatically creates a new field in the dataset that contain propensity scores for the test data and an XML model file that can be used to score other datasets. Optional diagnostic output includes an overall model quality chart and a classification table that compares predicted responses to actual responses.

Figure 6-1
Overall model quality chart



Propensity to Purchase data considerations

Response Field. The response field can be string or numeric. If this field contains a value that indicates number or monetary value of purchases, you will need to create a new field in which a single value represents all positive responses. For more information, see the topic Creating a categorical response field on p. 38.

Positive response value. The positive response value identifies customers who responded positively (for example, made a purchase). All other non-missing response values are assumed to indicate a negative response. If there are any defined value labels for the response field, those labels are displayed in the drop-down list.

Predict Propensity with. The fields used to predict propensity can be string or numeric, and they can be nominal, ordinal, or continuous (scale) — but it is important to assign the proper measurement level to all predictor fields.

Measurement level. Correct measurement level assignment is important because it affects the computation of the results.

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.
- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.
- **Continuous.** A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

An icon next to each field indicates the current measurement level.

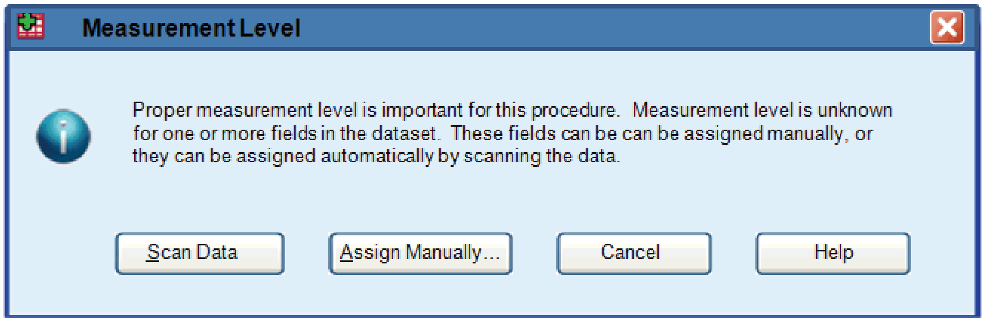
Measurement Level	**Data Type**			
	Numeric	String	Date	Time
Scale (Continuous)	📏	n/a	📏📅	📏🕒
Ordinal	📊	📊a	📊📅	📊🕒
Nominal	●	●a	●📅	●🕒

You can change the measurement level in Variable View of the Data Editor or you can use the Define Variable Properties dialog to suggest an appropriate measurement level for each field.

Fields with unknown measurement level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 6-2
Measurement level alert



Measurement Level

Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be can be assigned manually, or they can be assigned automatically by scanning the data.

- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.
- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

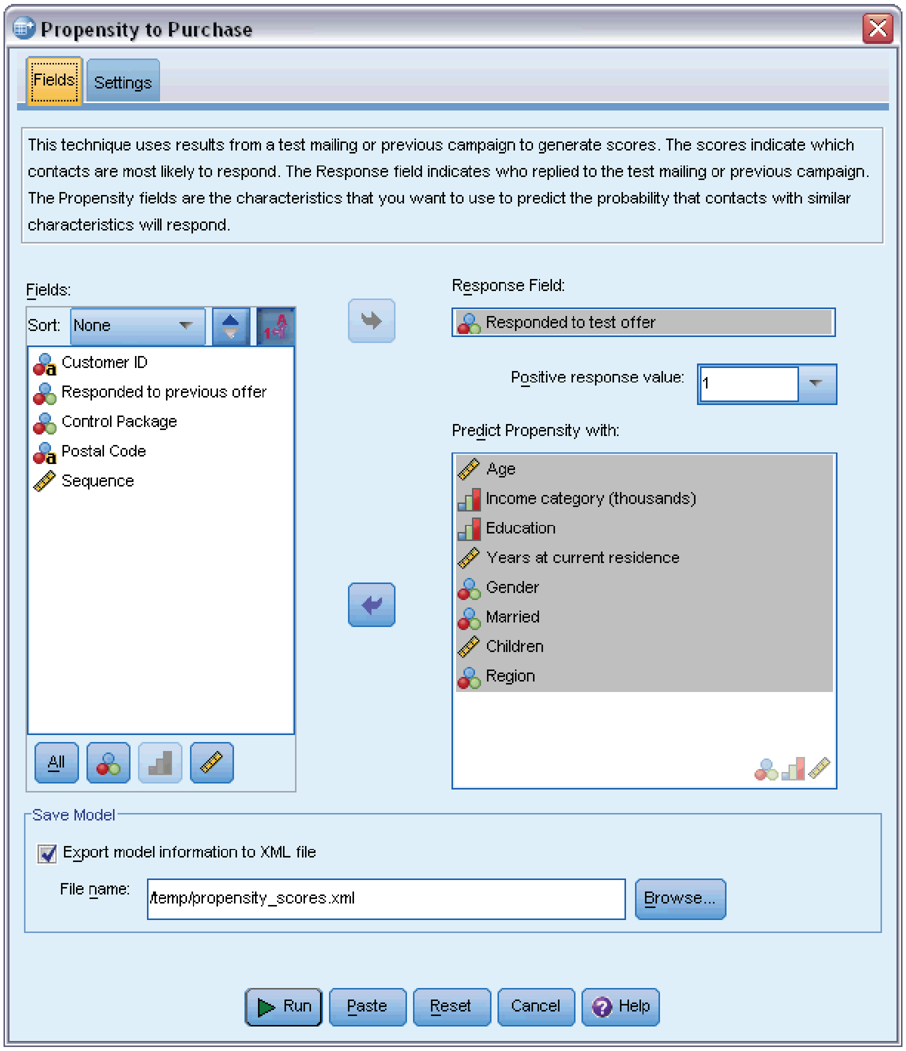
To obtain propensity to purchase scores

From the menus choose:

Direct Marketing > Choose Technique

- ▶ Select Select contacts most likely to purchase.

Figure 6-3
Propensity to Purchase Fields tab



Propensity to Purchase

This technique uses results from a test mailing or previous campaign to generate scores. The scores indicate which contacts are most likely to respond. The Response field indicates who replied to the test mailing or previous campaign. The Propensity fields are the characteristics that you want to use to predict the probability that contacts with similar characteristics will respond.

Fields:

Sort: None

- Customer ID
- Responded to previous offer
- Control Package
- Postal Code
- Sequence

Response Field: Responded to test offer

Positive response value: 1

Predict Propensity with:

- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Children
- Region

Save Model

☑ Export model information to XML file

File name: /temp/propensity_scores.xml

- ▶ Select the field that identifies which contacts responded to the offer.
- ▶ Enter the value that indicates a positive response. If any values have defined value labels, you can select the value label from the drop-down list, and the corresponding value will be displayed.
- ▶ Select the fields you want to use to predict propensity.

To save a model XML file to score other data files:

- ▶ Select (check) Export model information to XML file.
- ▶ Enter a directory path and file name or click Browse to navigate to the location where you want to save the model XML file.

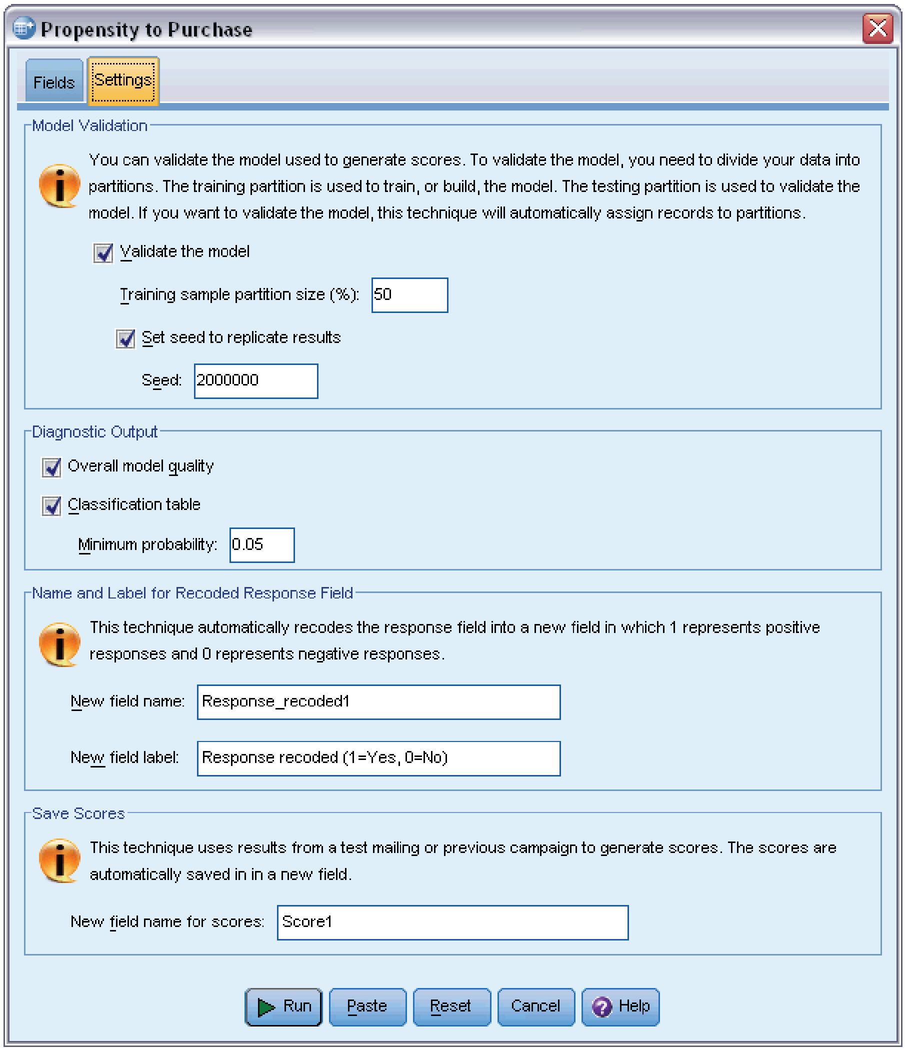
- ▶ Click Run to run the procedure.

  To use the model file to score other datasets:

- ▶ Open the dataset that you want to score.
- ▶ Use the Scoring Wizard to apply the model to the dataset. From the menus choose:
  Utilities > Scoring Wizard.

Settings

Figure 6-4
Propensity to Purchase, Settings tab



Model Validation

You can validate the model used to generate scores. To validate the model, you need to divide your data into partitions. The training partition is used to train, or build, the model. The testing partition is used to validate the model. If you want to validate the model, this technique will automatically assign records to partitions.

☑ Validate the model

Training sample partition size (%): 50

☑ Set seed to replicate results

Seed: 2000000

Diagnostic Output

☑ Overall model quality

☑ Classification table

Minimum probability: 0.05

Name and Label for Recoded Response Field

This technique automatically recodes the response field into a new field in which 1 represents positive responses and 0 represents negative responses.

New field name: Response_recoded1

New field label: Response recoded (1=Yes, 0=No)

Save Scores

This technique uses results from a test mailing or previous campaign to generate scores. The scores are automatically saved in in a new field.

New field name for scores: Score1

Model Validation

Model validation creates training and testing groups for diagnostic purposes. If you select the classification table in the Diagnostic Output section, the table will be divided into training (selected) and testing (unselected) sections for comparison purposes. Do not select model validation unless you also select the classification table. The scores are based on the model generated from the training sample, which will always contain fewer records than the total number of available records. For example, the default training sample size is 50%, and a model built on only half the available records may not be as reliable as a model built on all available records.

- **Training sample partition size (%).** Specify the percentage of records to assign to the training sample. The rest of the records with non-missing values for the response field are assigned to the testing sample. The value must be greater than 0 and less than 100.
- **Set seed to replicate results.** Since records are randomly assigned to the training and testing samples, each time you run the procedure you may get different results, unless you always specify the same starting random number seed value.

Diagnostic Output

Overall model quality. Displays a bar chart of overall model quality, expressed as a value between 0 and 1. A good model should have a value greater than 0.5.

Classification table. Displays a table that compares predicted positive and negative responses to actual positive and negative responses. The overall accuracy rate can provide some indication of how well the model works, but you may be more interested in the percentage of correct predicted positive responses.

- **Minimum probability.** Assigns records with a score value greater than the specified value to the predicted positive response category in the classification table. The scores generated by the procedure represent the probability that the contact will respond positively (for example, make a purchase). As a general rule, you should specify a value close to your minimum target response rate, expressed as a proportion. For example, if you are interested in a response rate of at least 5%, specify 0.05. The value must be greater than 0 and less than 1.

Name and Label for Recoded Response Field

This procedure automatically recodes the response field into a new field in which 1 represents positive responses and 0 represents negative responses, and the analysis is performed on the recoded field. You can override the default name and label and provide your own. Names must conform to IBM® SPSS® Statistics naming rules.

Save Scores

A new field containing propensity scores is automatically saved to the original dataset. Scores represent the probability of a positive response, expressed as a proportion.

- Field names must conform to SPSS Statistics naming rules.
- The field name cannot duplicate a field name that already exists in the dataset. If you run this procedure more than once on the same dataset, you will need to specify a different name each time.

Creating a categorical response field

The response field should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values.

- If negative responses are recorded as 0 (not blank, which is treated as missing), this can be computed with the following formula:

  NewName=OldName>0

  where *NewName* is the name of the new field and *OldName* is the name of the original field. This is a logical expression that assigns a value of 1 to all non-missing values greater than 0, and 0 to all non-missing values less than or equal to 0.

- If no value is recorded for negative responses, then these values are treated as missing, and the formula is a little more complicated:

  NewName=NOT(MISSING(OldName))

  In this logical expression, all non-missing response values are assigned a value of 1 and all missing response values are assigned a value of 0.

- If you cannot distinguish between negative (0) response values and missing values, then an accurate response value cannot be computed. If there are relatively few truly missing values, this may not have a significant effect on the computed response rates. If, however, there are many missing values — such as when response information is recorded for only a small test sample of the total dataset — then the computed response rates will be meaningless, since they will be significantly lower than the true response rates.

To Create a Categorical Response Field

▶ From the menus choose:

Transform > Compute Variable

▶ For Target Variable, enter the new field (variable) name.

▶ If negative responses are recorded as 0, for the Numeric Expression enter OldName>0, where *OldName* is the original field name.

▶ If negative responses are recorded as missing (blank), for the Numeric Expression enter NOT(MISSING(OldName)), where *OldName* is the original field name.

Chapter

7

Control Package Test

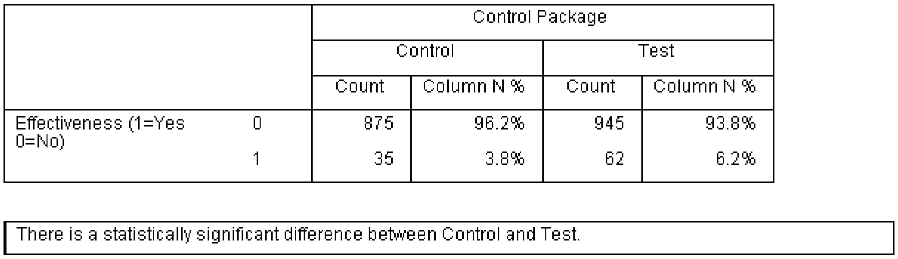
This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. Campaign effectiveness is measured by responses. The Campaign Field identifies different campaigns, for example Offer A and Offer B. The Response Field indicates if a contact responded to the campaign. Select Purchase Amount when the response is recorded as a purchase amount, for example “99.99”. Select Reply when the response simply indicates if the contact responded positively or not, for example “Yes” or “No”.

Example. The direct marketing division of a company wants to see if a new package design will generate more positive responses than the existing package. So they send out a test mailing to determine if the new package generates a significantly higher positive response rate. The test mailing consists of a control group that receives the existing package and a test group that receives the new package design. The results for the two groups are then compared to see if there is a significant difference.

Output

Output includes a table that displays counts and percentages of positive and negative responses for each group defined by the Campaign Field and a table that identifies which groups differ significantly from each other.

Figure 7-1
Control Package Test output



		Control Package			
		Control		Test	
		Count	Column N %	Count	Column N %
Effectiveness (1=Yes 0=No)	0	875	96.2%	945	93.8%
	1	35	3.8%	62	6.2%

There is a statistically significant difference between Control and Test.

Control Package Test Data Considerations and Assumptions

Campaign Field. The Campaign Field should be categorical (nominal or ordinal).

Effectiveness Response Field. If you select Purchase amount for the Effectiveness Field, the field must be numeric, and the level of measurement should be continuous (scale).

If you cannot distinguish between negative (for purchase amount, a value of 0) response values and missing values, then an accurate response rate cannot be computed. If there are relatively few truly missing values, this may not have a significant effect on the computed response rates. If, however, there are many missing values — such as when response information is recorded for only a small test sample of the total dataset — then the computed response rates will be meaningless, since they will be significantly lower than the true response rates.

Assumptions. This procedure assumes that contacts have been randomly assigned to each campaign group. In other words, no particular demographic, purchase history, or other characteristics affect group assignment, and all contacts have an equal probability of being assigned to any group.

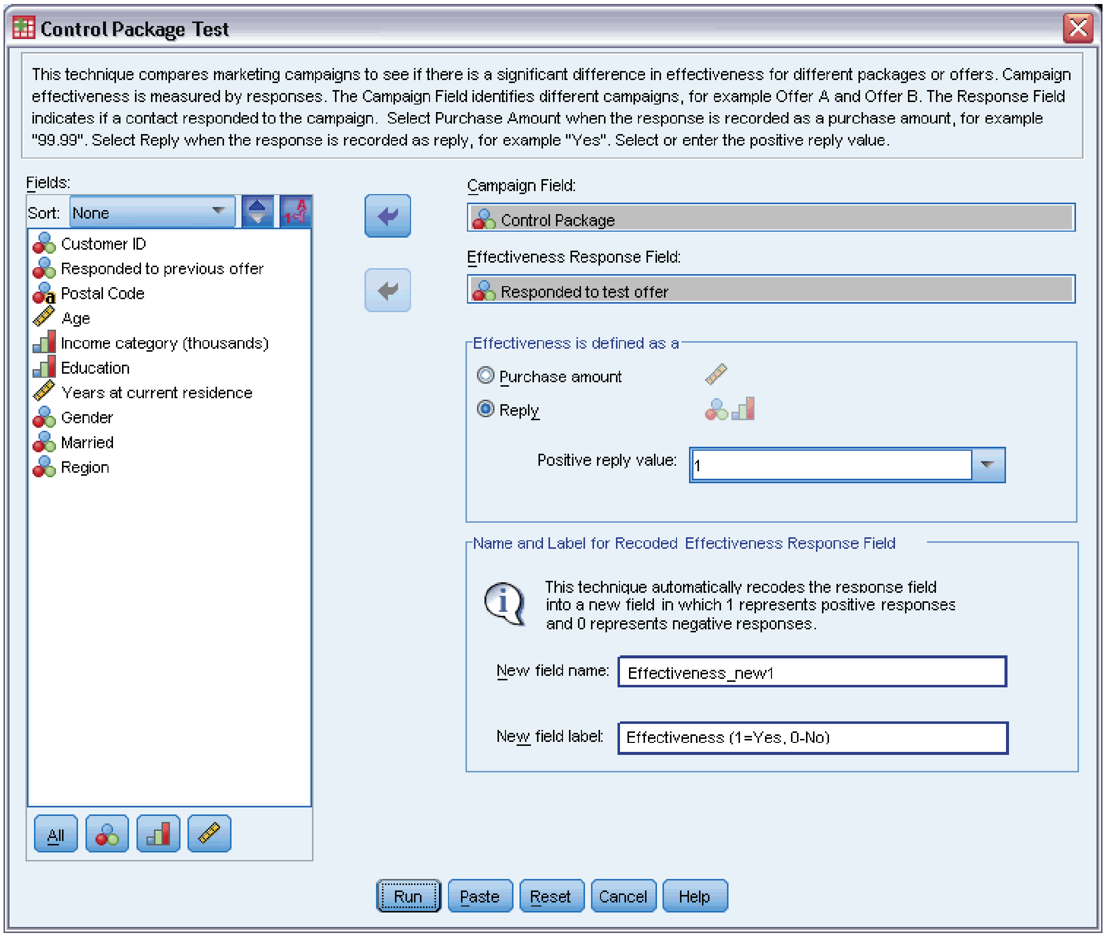
To Obtain a Control Package Test

From the menus choose:

Direct Marketing > Choose Technique

▶ Select Compare effectiveness of campaigns.

Figure 7-2
Control Package Test dialog



Control Package Test

This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. Campaign effectiveness is measured by responses. The Campaign Field identifies different campaigns, for example Offer A and Offer B. The Response Field indicates if a contact responded to the campaign. Select Purchase Amount when the response is recorded as a purchase amount, for example "99.99". Select Reply when the response is recorded as reply, for example "Yes". Select or enter the positive reply value.

Fields:
Sort: None

- Customer ID
- Responded to previous offer
- Postal Code
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Region

Campaign Field: Control Package

Effectiveness Response Field: Responded to test offer

Effectiveness is defined as a
○ Purchase amount
● Reply

Positive reply value: 1

Name and Label for Recoded Effectiveness Response Field

This technique automatically recodes the response field into a new field in which 1 represents positive responses and 0 represents negative responses.

New field name: Effectiveness_new1

New field label: Effectiveness (1=Yes, 0-No)

- Select the field that identifies which campaign group each contact belongs to (for example, offer A, offer B, etc.) This field must be nominal or ordinal.
- Select the field that indicates response effectiveness.

  If the response field is a purchase amount, the field must be numeric.

  If the response field simply indicates if the contact responded positively or not (for example "Yes" or "No"), select Reply and enter the value that represents a positive response. If any values have defined value labels, you can select the value label from the drop-down list, and the corresponding value will be displayed.

  A new field is automatically created, in which 1 represents positive responses and 0 represents negative responses, and the analysis is performed on the new field. You can override the default name and label and provide your own. Names must conform to IBM® SPSS® Statistics naming rules.
- Click Run to run the procedure.

Part II: Examples

Chapter

8

RFM Analysis from Transaction Data

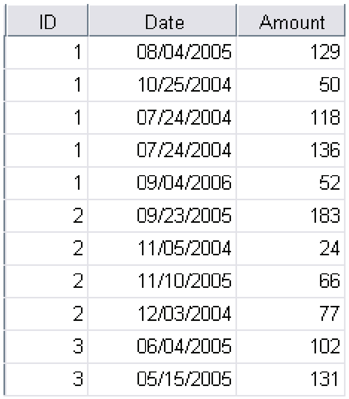
In a transaction data file, each row represents a separate transaction, rather than a separate customer, and there can be multiple transaction rows for each customer. This example uses the data file *rfm_transactions.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

Transaction Data

The dataset must contain variables that contain the following information:

- A variable or combination of variables that identify each case (customer).
- A variable with the date of each transaction.
- A variable with the monetary value of each transaction.

Figure 8-1
RFM transaction data



ID	Date	Amount
1	08/04/2005	129
1	10/25/2004	50
1	07/24/2004	118
1	07/24/2004	136
1	09/04/2006	52
2	09/23/2005	183
2	11/05/2004	24
2	11/10/2005	66
2	12/03/2004	77
3	06/04/2005	102
3	05/15/2005	131

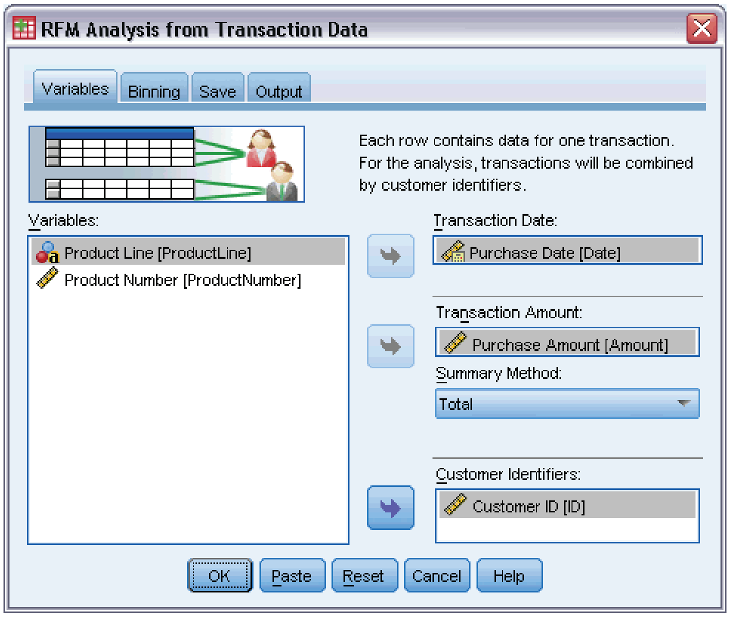
Running the Analysis

▶ To calculate RFM scores, from the menus choose:
Direct Marketing > Choose Technique

▶ Select Help identify my best contacts (RFM Analysis) and click Continue.

▶ Click Transaction data and then click Continue.

Figure 8-2
RFM from Transactions, Variables tab



Each row contains data for one transaction. For the analysis, transactions will be combined by customer identifiers.

Variables:
- Product Line [ProductLine]
- Product Number [ProductNumber]

Transaction Date: Purchase Date [Date]

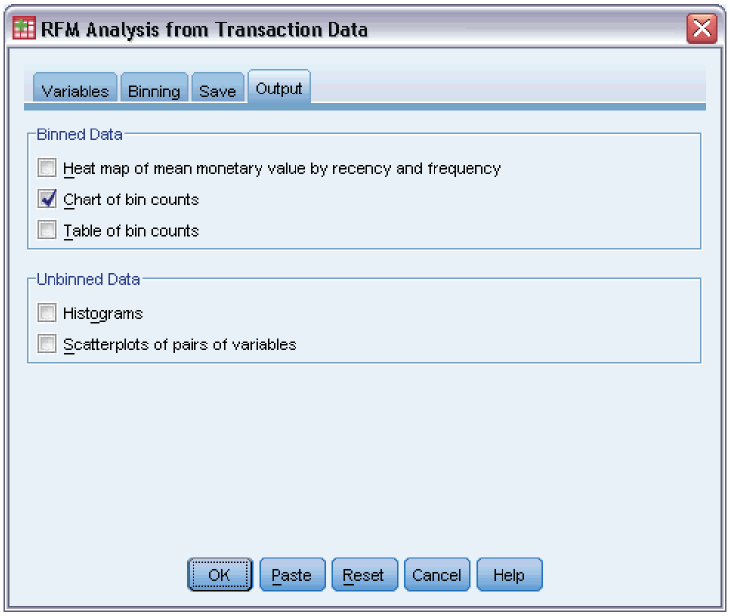
Transaction Amount: Purchase Amount [Amount]

Summary Method: Total

Customer Identifiers: Customer ID [ID]

- ▶ Click Reset to clear any previous settings.
- ▶ For Transaction Date, select *Purchase Date [Date]*.
- ▶ For Transaction Amount, select *Purchase Amount [Amount]*.
- ▶ For Summary Method, select Total.
- ▶ For Customer Identifiers, select *Customer ID [ID]*.
- ▶ Then click the Output tab.

Figure 8-3
RFM for Transactions, Output tab



Binned Data

- □ Heat map of mean monetary value by recency and frequency
- ☑ Chart of bin counts
- □ Table of bin counts

Unbinned Data

- □ Histograms
- □ Scatterplots of pairs of variables

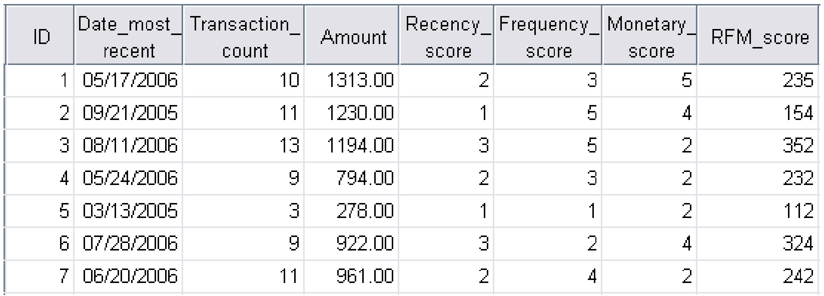
▶ Select (check) Chart of bin counts.

▶ Then click OK to run the procedure.

Evaluating the Results

When you compute RFM scores from transaction data, a new dataset is created that includes the new RFM scores.

Figure 8-4
RFM from Transactions dataset



ID	Date_most_recent	Transaction_count	Amount	Recency_score	Frequency_score	Monetary_score	RFM_score
1	05/17/2006	10	1313.00	2	3	5	235
2	09/21/2005	11	1230.00	1	5	4	154
3	08/11/2006	13	1194.00	3	5	2	352
4	05/24/2006	9	794.00	2	3	2	232
5	03/13/2005	3	278.00	1	1	2	112
6	07/28/2006	9	922.00	3	2	4	324
7	06/20/2006	11	961.00	2	4	2	242

By default, the dataset includes the following information for each customer:

- Customer ID variable(s)
- Date of most recent transaction
- Total number of transactions

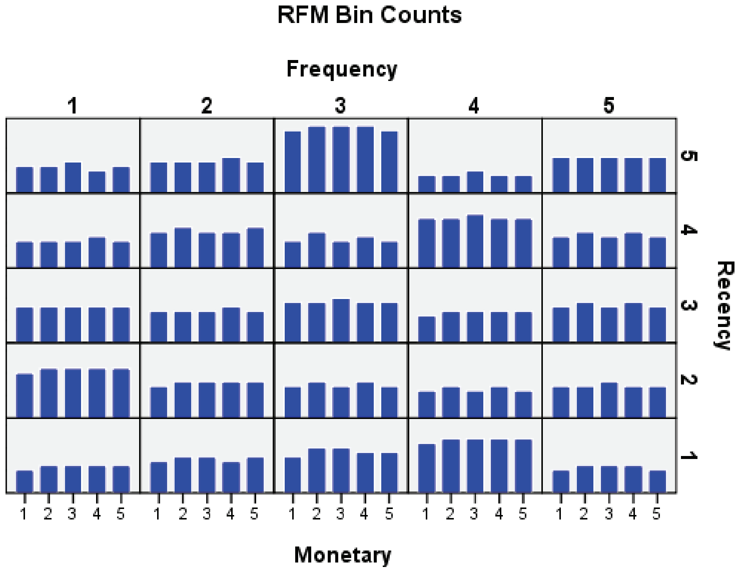
- Summary transaction amount (the default is total)
- Recency, Frequency, Monetary, and combined RFM scores

The new dataset contains only one row (record) for each customer. The original transaction data has been aggregated by values of the customer identifier variables. The identifier variables are always included in the new dataset; otherwise you would have no way of matching the RFM scores to the customers.

The combined RFM score for each customer is simply the concatenation of the three individual scores, computed as: (recency x 100) + (frequency x 10) + monetary.

The chart of bin counts displayed in the Viewer window shows the number of customers in each RFM category.

Figure 8-5
Chart of bin counts



Using the default method of five score categories for each of the three RFM components results in 125 possible RFM score categories. Each bar in the chart represents the number of customers in each RFM category.

Ideally, you want a relatively even distribution of customers across all RFM score categories. In reality, there will usually be some amount of variation, such as what you see in this example. If there are many empty categories, you might want to consider changing the binning method.

There are a number of strategies for dealing with uneven distributions of RFM scores, including:

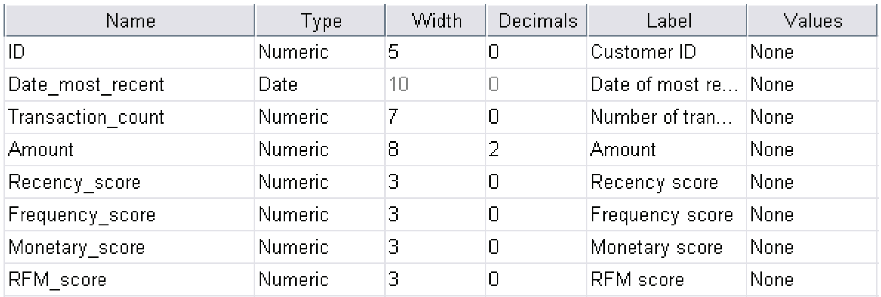
- Use nested instead of independent binning.
- Reduce the number of possible score categories (bins).
- When there are large numbers of tied values, randomly assign cases with the same scores to different categories.

For more information, see the topic RFM Binning in Chapter 2 on p. 6.

Merging Score Data with Customer Data

Now that you have a dataset that contains RFM scores, you need to match those scores to the customers. You could merge the scores back to the transaction data file, but more typically you want to merge the score data with a data file that, like the RFM score dataset, contains one row (record) for each customer — and also contains information such as the customer's name and address.

Figure 8-6
RFM score dataset in Variable View

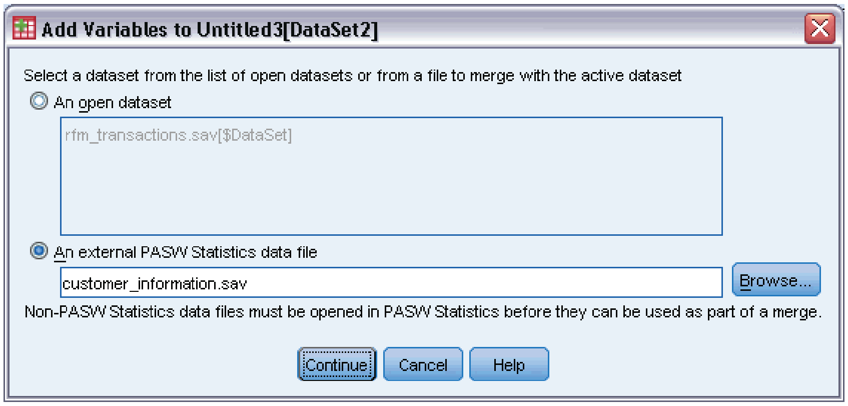


Name	Type	Width	Decimals	Label	Values
ID	Numeric	5	0	Customer ID	None
Date_most_recent	Date	10	0	Date of most re...	None
Transaction_count	Numeric	7	0	Number of tran...	None
Amount	Numeric	8	2	Amount	None
Recency_score	Numeric	3	0	Recency score	None
Frequency_score	Numeric	3	0	Frequency score	None
Monetary_score	Numeric	3	0	Monetary score	None
RFM_score	Numeric	3	0	RFM score	None

▶ Make the dataset that contains the RFM scores the active dataset. (Click anywhere in the Data Editor window that contains the dataset.)

▶ From the menus choose:
Data > Merge Files > Add Variables

Figure 8-7
Add Variables, select files dialog



Add Variables to Untitled3[DataSet2]

Select a dataset from the list of open datasets or from a file to merge with the active dataset

○ An open dataset

rfm_transactions.sav[$DataSet]

● An external PASW Statistics data file

customer_information.sav

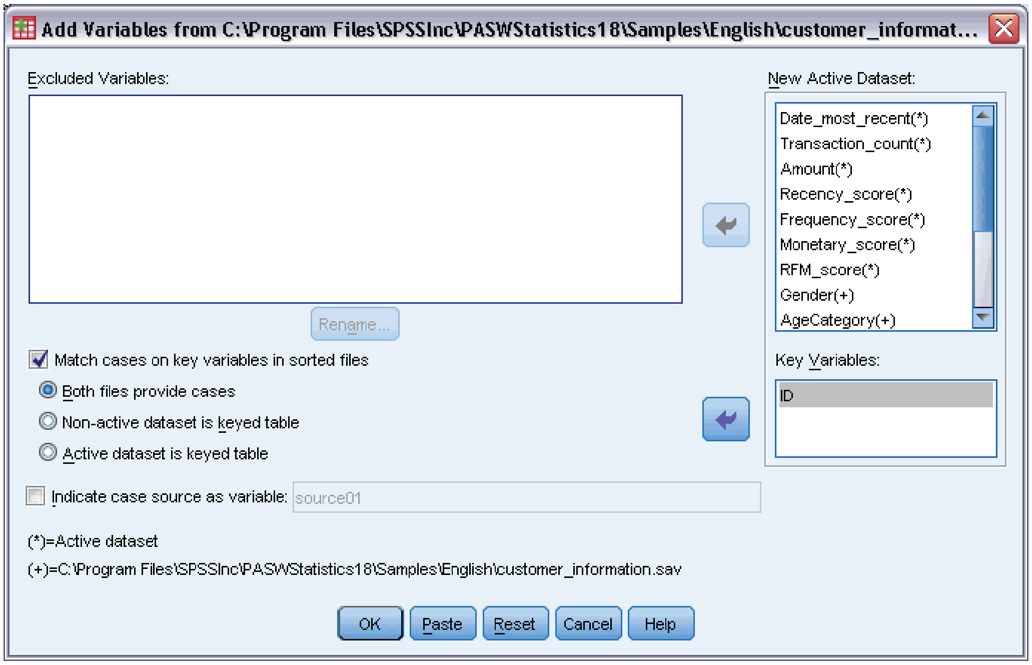
Non-PASW Statistics data files must be opened in PASW Statistics before they can be used as part of a merge.

▶ Select An external data file.

▶ Use the Browse button to navigate to the *Samples* folder and select *customer_information.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

▶ Then click Continue.

Figure 8-8
Add Variables, select variables dialog



Add Variables from C:\Program Files\SPSSInc\PASWStatistics18\Samples\English\customer_informat...

Excluded Variables:

New Active Dataset:

Date_most_recent(*)
Transaction_count(*)
Amount(*)
Recency_score(*)
Frequency_score(*)
Monetary_score(*)
RFM_score(*)
Gender(+)
AgeCategory(+)

Key Variables:

ID

☑ Match cases on key variables in sorted files

● Both files provide cases

○ Non-active dataset is keyed table

○ Active dataset is keyed table

□ Indicate case source as variable: source01

(*)=Active dataset

(+)=C:\Program Files\SPSSInc\PASWStatistics18\Samples\English\customer_information.sav

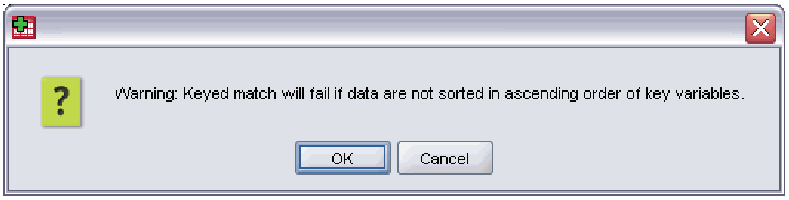
- ▶ Select (check) Match cases on key variables in sorted files.

- ▶ Select Both files provide cases.

- ▶ Select *ID* for the Key Variables list.

- ▶ Click OK.

Figure 8-9
Add Variables warning message



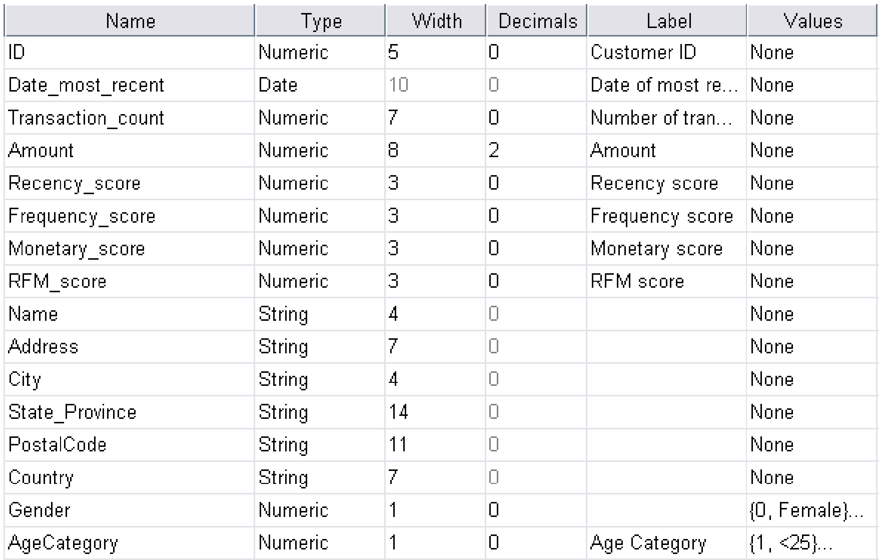
Warning: Keyed match will fail if data are not sorted in ascending order of key variables.

Note the message that warns you that both files must be sorted in ascending order of the key variables. In this example, both files are already sorted in ascending order of the key variable, which is the customer identifier variable we selected when we computed the RFM scores. When you compute RFM scores from transaction data, the new dataset is automatically sorted in ascending order of the customer identifier variable(s). If you change the sort order of the score dataset or the data file with which you want to merge the score dataset is not sorted in that order, you must first sort both files in ascending order of the customer identifier variable(s).

- ▶ Click OK to merge the two datasets.

The dataset that contains the RFM scores now also contains name, address and other information for each customer.

Figure 8-10
Merged datasets



Name	Type	Width	Decimals	Label	Values
ID	Numeric	5	0	Customer ID	None
Date_most_recent	Date	10	0	Date of most re...	None
Transaction_count	Numeric	7	0	Number of tran...	None
Amount	Numeric	8	2	Amount	None
Recency_score	Numeric	3	0	Recency score	None
Frequency_score	Numeric	3	0	Frequency score	None
Monetary_score	Numeric	3	0	Monetary score	None
RFM_score	Numeric	3	0	RFM score	None
Name	String	4	0		None
Address	String	7	0		None
City	String	4	0		None
State_Province	String	14	0		None
PostalCode	String	11	0		None
Country	String	7	0		None
Gender	Numeric	1	0		{0, Female}...
AgeCategory	Numeric	1	0	Age Category	{1, <25}...

Chapter

9

Cluster analysis

Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.

For example, the direct marketing division of a company wants to identify demographic groupings in their customer database to help determine marketing campaign strategies and develop new product offerings.

This information is collected in *dmdata.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

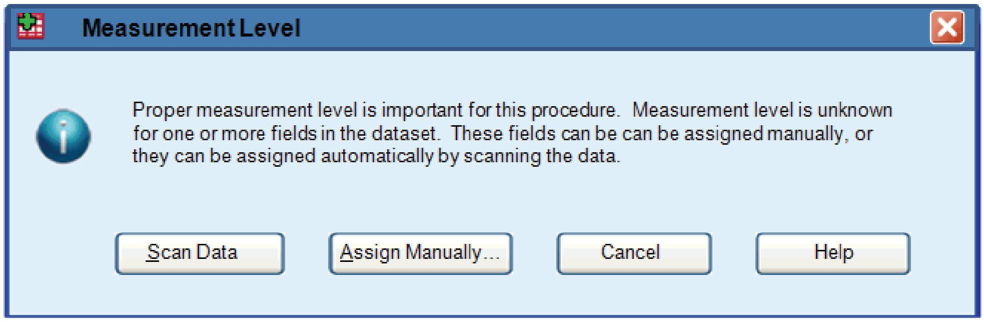
Running the analysis

▶ To run a Cluster Analysis, from the menus choose:
Direct Marketing > Choose Technique

▶ Select Segment my contacts into clusters and click Continue.

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 9-1
Measurement level alert



Measurement Level

Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be can be assigned manually, or they can be assigned automatically by scanning the data.

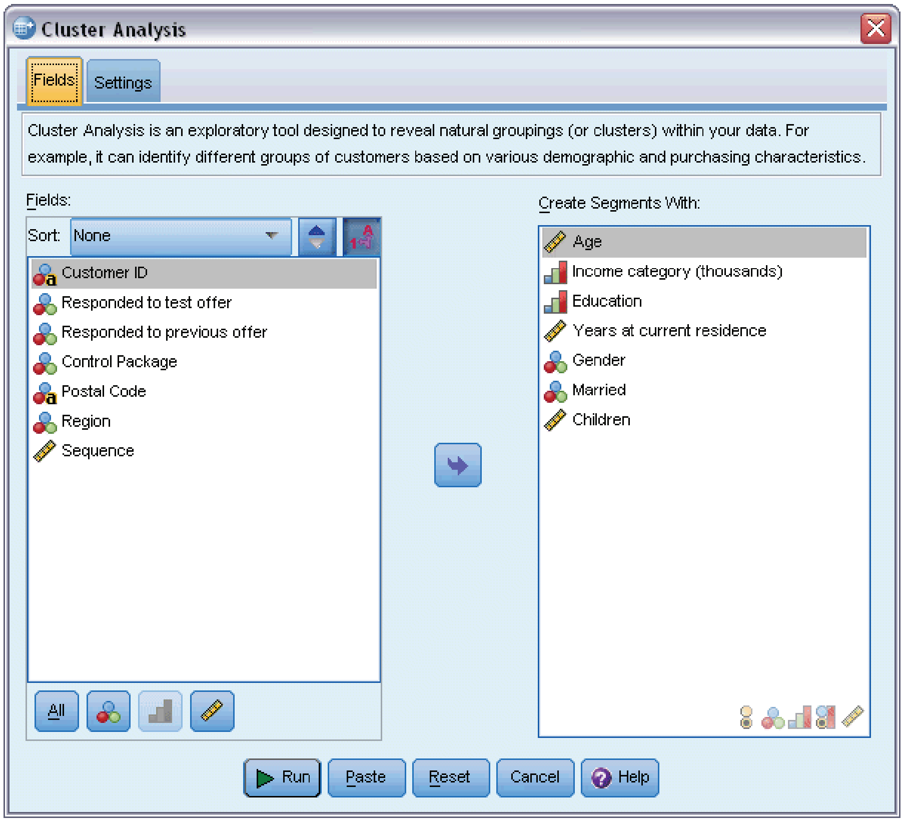
- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

In this example file, there are no fields with an unknown measurement level, and all fields have the correct measurement level; so the measurement level alert should not appear.

Figure 9-2
Cluster Analysis, Fields tab



Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within your data. For example, it can identify different groups of customers based on various demographic and purchasing characteristics.

Fields:

Sort: None

- Customer ID
- Responded to test offer
- Responded to previous offer
- Control Package
- Postal Code
- Region
- Sequence

Create Segments With:

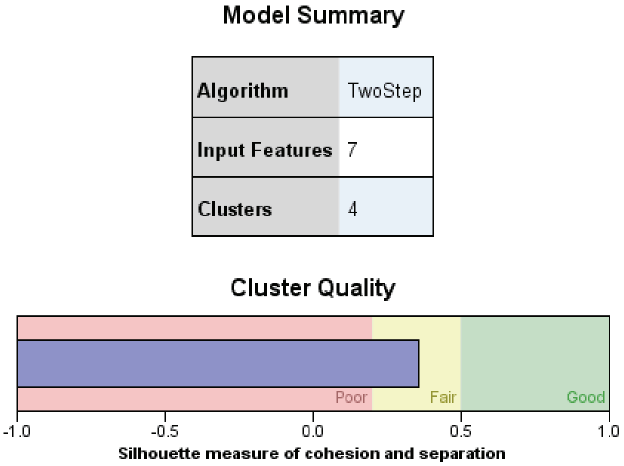
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Children

▶ Select the following fields to create segments: *Age*, *Income category*, *Education*, *Years at current residence*, *Gender*, *Married*, and *Children*.

▶ Click Run to run the procedure.

Output

Figure 9-3
Cluster model summary



Model Summary

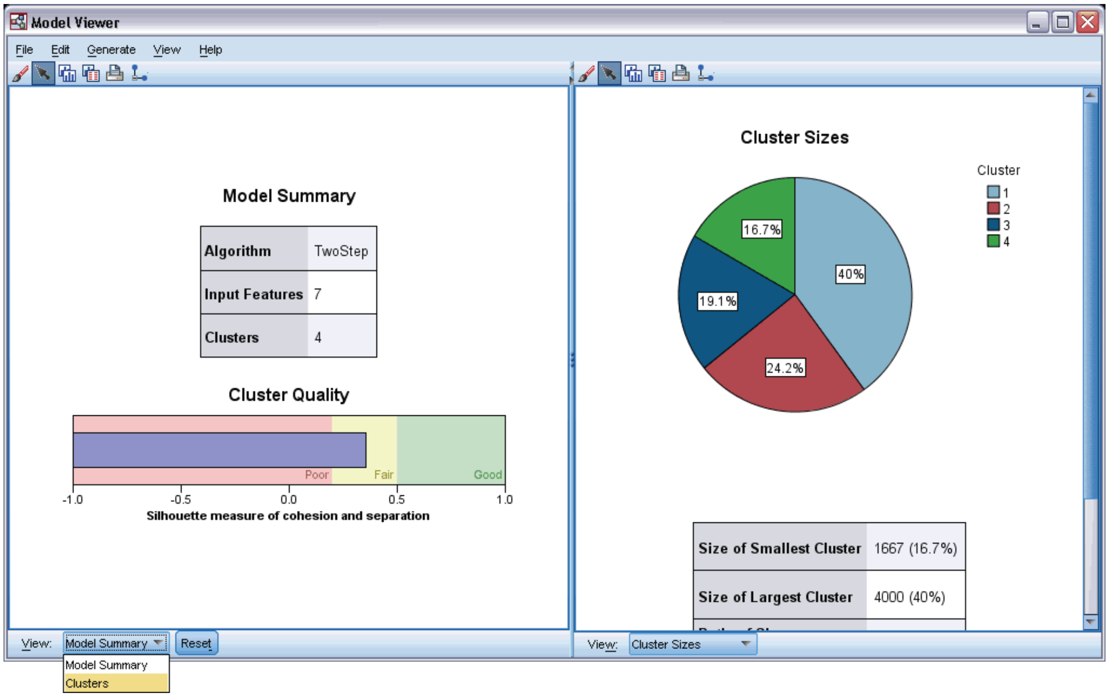
Algorithm	TwoStep
Input Features	7
Clusters	4

The results are displayed in the Cluster Model Viewer.

- The model summary indicates that four clusters were found based on the seven input features (fields) you selected.
- The cluster quality chart indicates that the overall model quality is in the middle of the “Fair” range.

▶ Double-click the Cluster Model Viewer output to activate the Model Viewer.

Figure 9-4
Activated Cluster Model Viewer



Algorithm	TwoStep
Input Features	7
Clusters	4

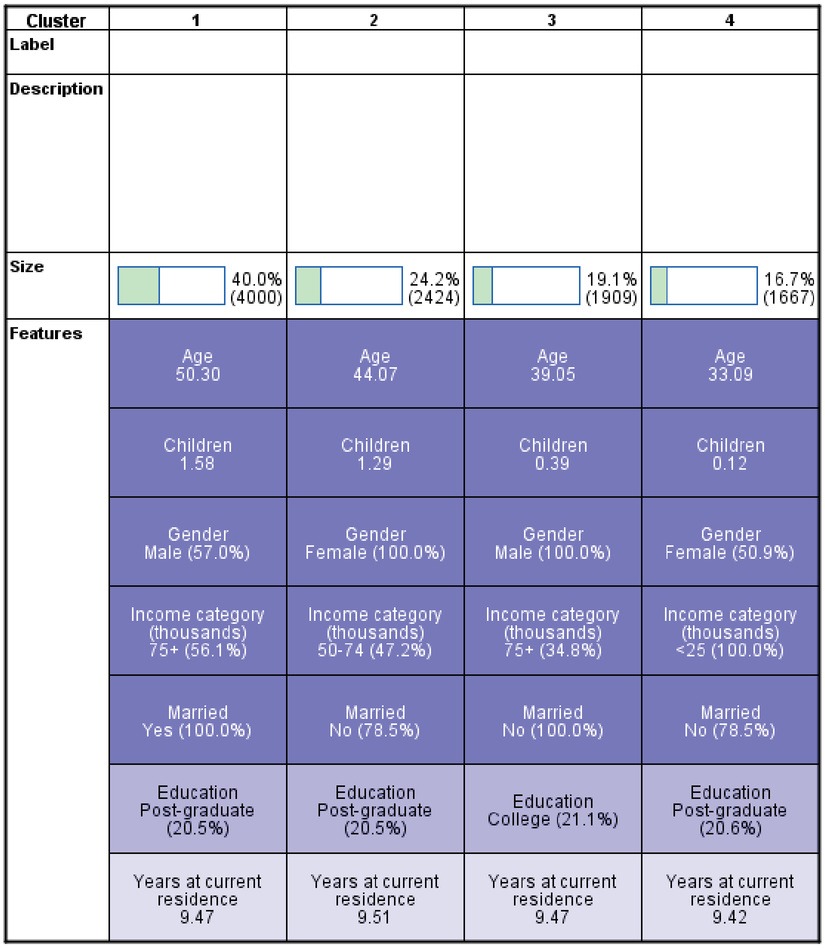
Size of Smallest Cluster	1667 (16.7%)
Size of Largest Cluster	4000 (40%)

View: Model Summary

View: Cluster Sizes

▶ From the View drop-down list at the bottom of the Cluster Model Viewer window, select Clusters.

Figure 9-5
Cluster view



Cluster	1	2	3	4
Label				
Description				
Size	40.0% (4000)	24.2% (2424)	19.1% (1909)	16.7% (1667)
Features	Age 50.30	Age 44.07	Age 39.05	Age 33.09
	Children 1.58	Children 1.29	Children 0.39	Children 0.12
	Gender Male (57.0%)	Gender Female (100.0%)	Gender Male (100.0%)	Gender Female (50.9%)
	Income category (thousands) 75+ (56.1%)	Income category (thousands) 50-74 (47.2%)	Income category (thousands) 75+ (34.8%)	Income category (thousands) <25 (100.0%)
	Married Yes (100.0%)	Married No (78.5%)	Married No (100.0%)	Married No (78.5%)
	Education Post-graduate (20.5%)	Education Post-graduate (20.5%)	Education College (21.1%)	Education Post-graduate (20.6%)
	Years at current residence 9.47	Years at current residence 9.51	Years at current residence 9.47	Years at current residence 9.42

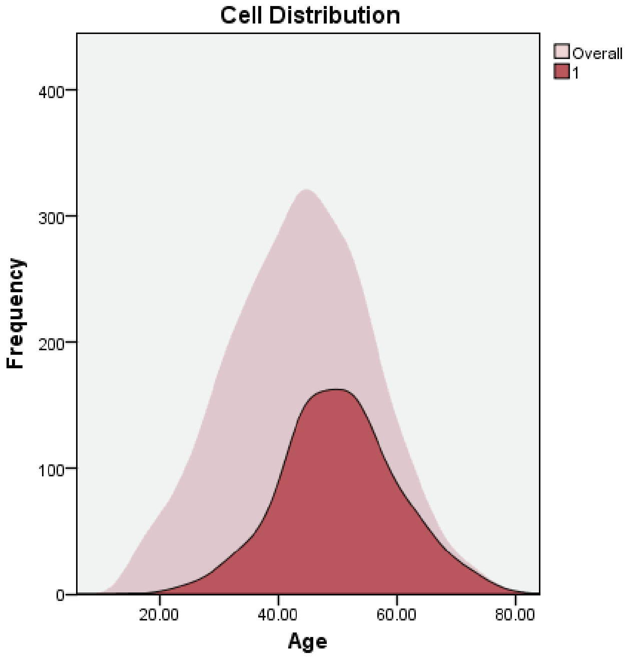
The Cluster view displays information on the attributes of each cluster.

- For continuous (scale) fields, the mean (average) value is displayed.
- For categorical (nominal, ordinal) fields, the mode is displayed. The mode is the category with the largest number of records. In this example, each record is a customer.
- By default, fields are displayed in the order of their overall importance to the model. In this example, *Age* has the highest overall importance. You can also sort fields by within-cluster importance or alphabetical order.

If you select (click) any cell in Cluster view, you can see a chart that summarizes the values of that field for that cluster.

▶ For example, select the *Age* cell for cluster 1.

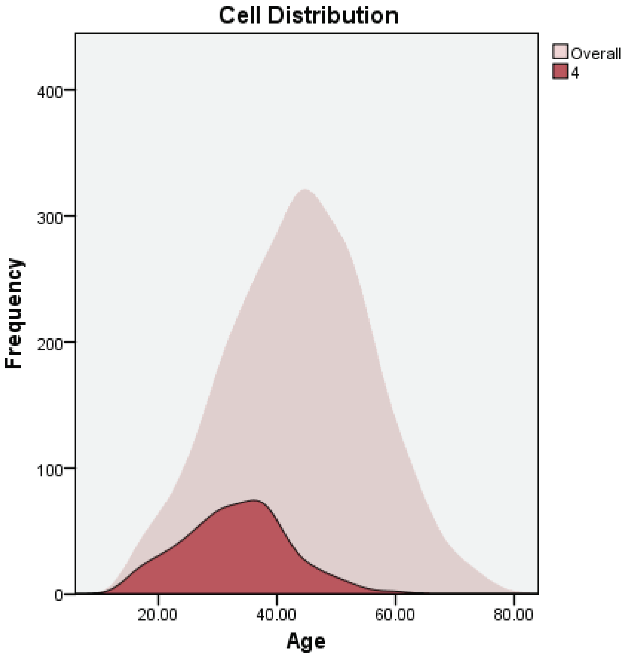
Figure 9-6
Age histogram for cluster 1



For continuous fields, a histogram is displayed. The histogram displays both the distribution of values within that cluster and the overall distribution of values for the field. The histogram indicates that the customers in cluster 1 tend to be somewhat older.

- ▶ Select the *Age* cell for cluster 4 in the Cluster view.

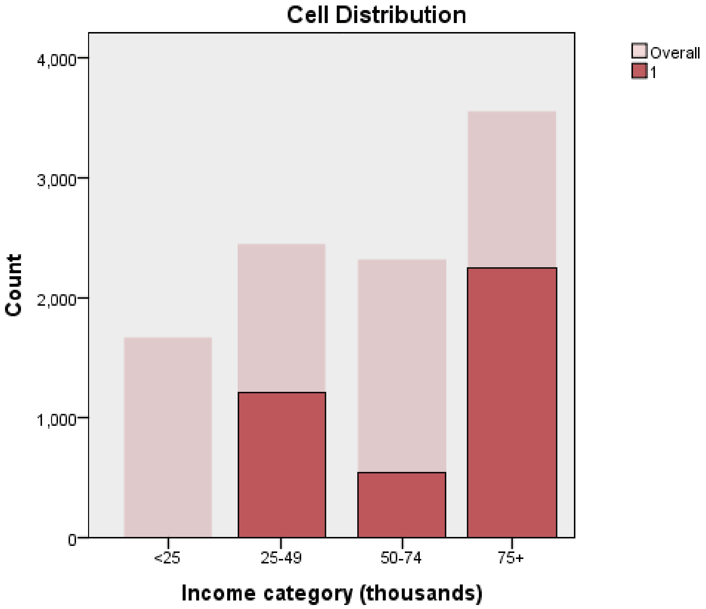
Figure 9-7
Age histogram for cluster 4



In contrast to cluster 1, the customers in cluster 4 tend to be younger than the overall average.

▶ Select the *Income category* cell for cluster 1 in the Cluster view.

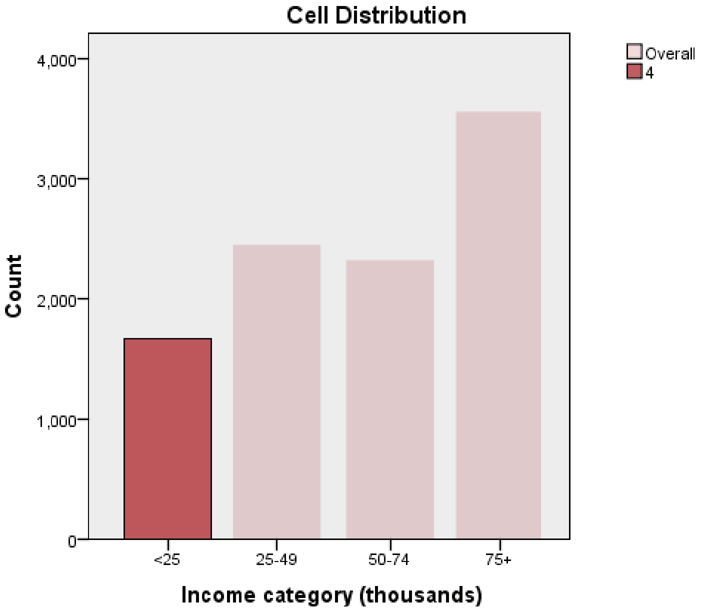
Figure 9-8
Income category bar chart for cluster 1



For categorical fields, a bar chart is displayed. The most notable feature of the income category bar chart for this cluster is the complete absence of any customers in the lowest income category.

▶ Select the *Income category* cell for cluster 4 in the Cluster view.

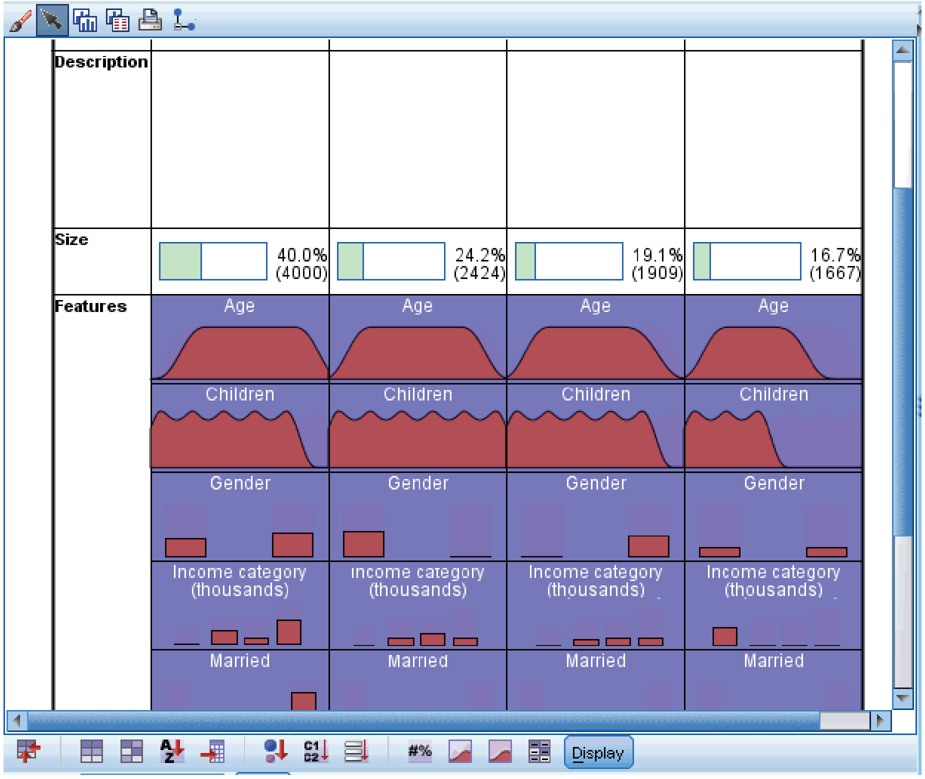
Figure 9-9
Income category bar chart for cluster 4



In contrast to cluster 1, all of the customers in cluster 4 are in the lowest income category.

You can also change the Cluster view to display charts in the cells, which makes it easy to quickly compare the distributions of values between clusters by using the toolbar at the bottom of Model Viewer window to change the view.

Figure 9-10
Charts displayed in the Cluster

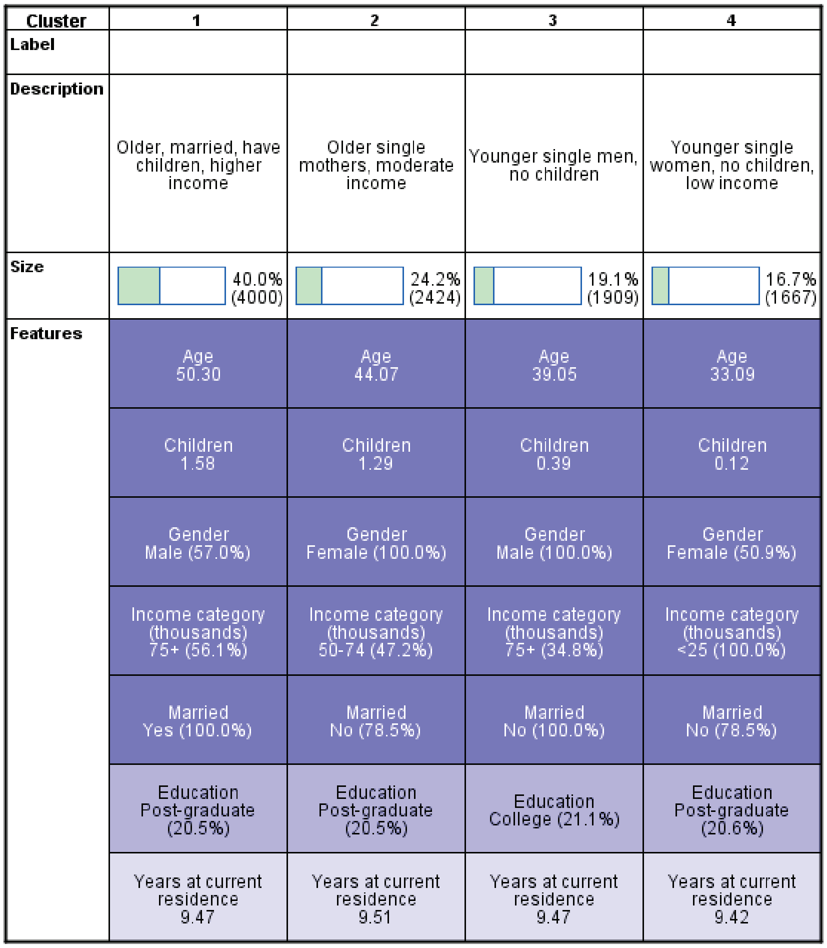


Looking at the Cluster view and the additional information provided in the charts for each cell, you can see some distinct differences between the clusters:

- Customers in cluster 1 tend to be older, married people with children and higher incomes.
- Customers in cluster 2 tend to be somewhat older single mothers with moderate incomes.
- Customers in cluster 3 tend to be younger, single men without children.
- Customers in cluster 4 tend to be younger, single women without children and with lower incomes.

The Description cells in the Cluster view are text fields that you can edit to add descriptions of each cluster.

Figure 9-11
Cluster view with cluster descriptions



Cluster	1	2	3	4
Label				
Description	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no children	Younger single women, no children, low income
Size	40.0% (4000)	24.2% (2424)	19.1% (1909)	16.7% (1667)
Features	Age 50.30	Age 44.07	Age 39.05	Age 33.09
	Children 1.58	Children 1.29	Children 0.39	Children 0.12
	Gender Male (57.0%)	Gender Female (100.0%)	Gender Male (100.0%)	Gender Female (50.9%)
	Income category (thousands) 75+ (56.1%)	Income category (thousands) 50-74 (47.2%)	Income category (thousands) 75+ (34.8%)	Income category (thousands) <25 (100.0%)
	Married Yes (100.0%)	Married No (78.5%)	Married No (100.0%)	Married No (78.5%)
	Education Post-graduate (20.5%)	Education Post-graduate (20.5%)	Education College (21.1%)	Education Post-graduate (20.6%)
	Years at current residence 9.47	Years at current residence 9.51	Years at current residence 9.47	Years at current residence 9.42

Selecting records based on clusters

You can select records based on cluster membership in two ways:

- Create a filter condition interactively in the Cluster Model Viewer.
- Use the values of the cluster field generated by the procedure to specify filter or selection conditions.

Creating a filter in the Cluster Model Viewer

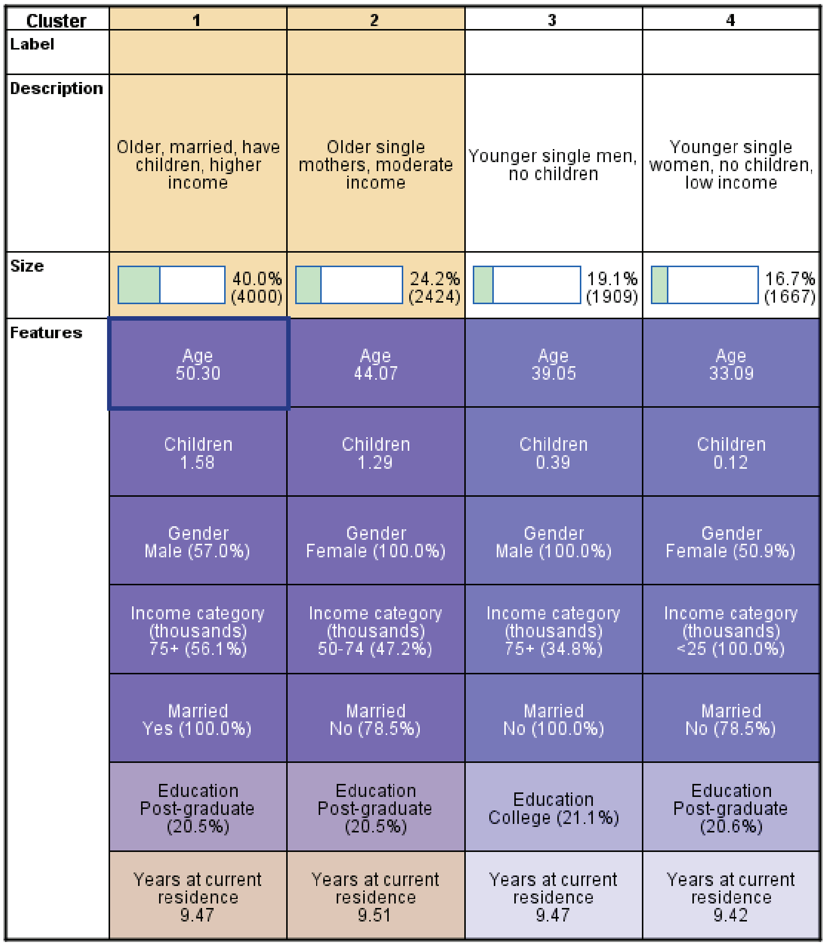
To create a filter condition that selects records from specific clusters in the Cluster Model Viewer:

▶ Activate (double-click) the Cluster Model Viewer.

▶ From the View drop-down list at the bottom of the Cluster Model Viewer window, select Clusters.

▶ Click the cluster number for the cluster you want at the top of the Cluster View. If you want to select multiple clusters, Ctrl-click on each additional cluster number that you want.

Figure 9-12
Clusters selected in Cluster view

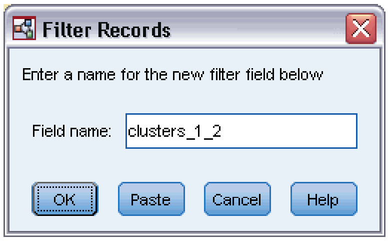


Cluster	1	2	3	4
Label				
Description	Older, married, have children, higher income	Older single mothers, moderate income	Younger single men, no children	Younger single women, no children, low income
Size	40.0% (4000)	24.2% (2424)	19.1% (1909)	16.7% (1667)
Features	Age 50.30	Age 44.07	Age 39.05	Age 33.09
	Children 1.58	Children 1.29	Children 0.39	Children 0.12
	Gender Male (57.0%)	Gender Female (100.0%)	Gender Male (100.0%)	Gender Female (50.9%)
	Income category (thousands) 75+ (56.1%)	Income category (thousands) 50-74 (47.2%)	Income category (thousands) 75+ (34.8%)	Income category (thousands) <25 (100.0%)
	Married Yes (100.0%)	Married No (78.5%)	Married No (100.0%)	Married No (78.5%)
	Education Post-graduate (20.5%)	Education Post-graduate (20.5%)	Education College (21.1%)	Education Post-graduate (20.6%)
	Years at current residence 9.47	Years at current residence 9.51	Years at current residence 9.47	Years at current residence 9.42

▶ From the Cluster Model Viewer menus, choose:

Generate > Filter records

Figure 9-13
Filter Records dialog



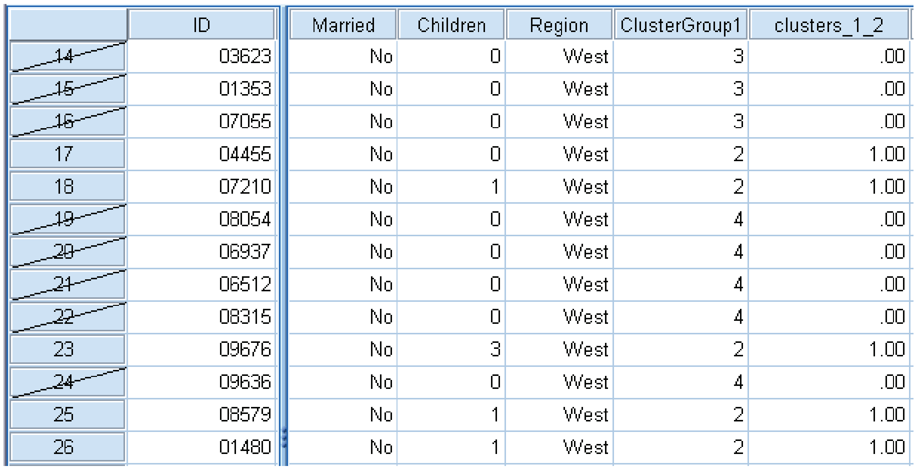
Filter Records

Enter a name for the new filter field below

Field name: clusters_1_2

▶ Enter a name for the filter field and click OK. Names must conform to IBM® SPSS® Statistics naming rules.

Figure 9-14
Filtered records in Data Editor



	ID	Married	Children	Region	ClusterGroup1	clusters_1_2
14	03623	No	0	West	3	.00
15	01353	No	0	West	3	.00
16	07055	No	0	West	3	.00
17	04455	No	0	West	2	1.00
18	07210	No	1	West	2	1.00
19	08054	No	0	West	4	.00
20	06937	No	0	West	4	.00
21	06512	No	0	West	4	.00
22	08315	No	0	West	4	.00
23	09676	No	3	West	2	1.00
24	09636	No	0	West	4	.00
25	08579	No	1	West	2	1.00
26	01480	No	1	West	2	1.00

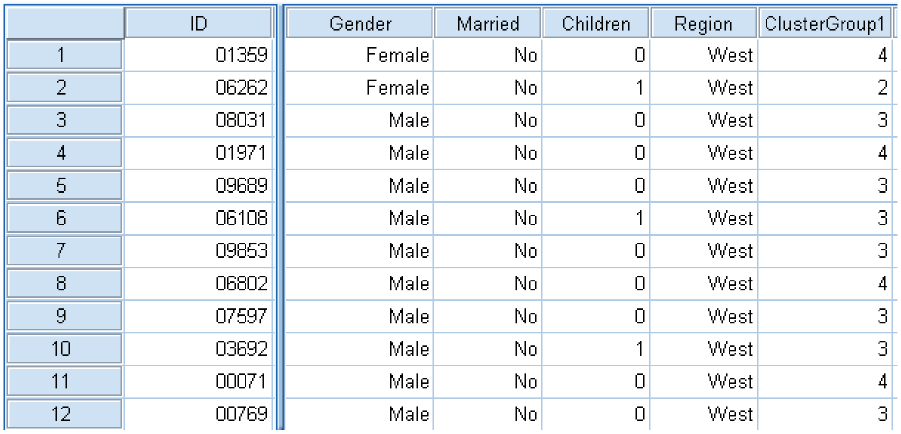
This creates a new field in the dataset and filters records in the dataset based on the values of that field.

- Records with a value of 1 for the filter field will be included in subsequent analyses, charts, and reports.
- Records with a value of 0 for the filter field will be excluded.
- Excluded records are not deleted from the dataset. They are retained with a filter status indicator, which is displayed as a diagonal slash through the record number in the Data Editor.

Selecting records based on cluster field values

By default, Cluster Analysis creates a new field that identifies the cluster group for each record. The default name of this field is *ClusterGroupn*, where *n* is an integer that forms a unique field name.

Figure 9-15
Cluster field added to dataset

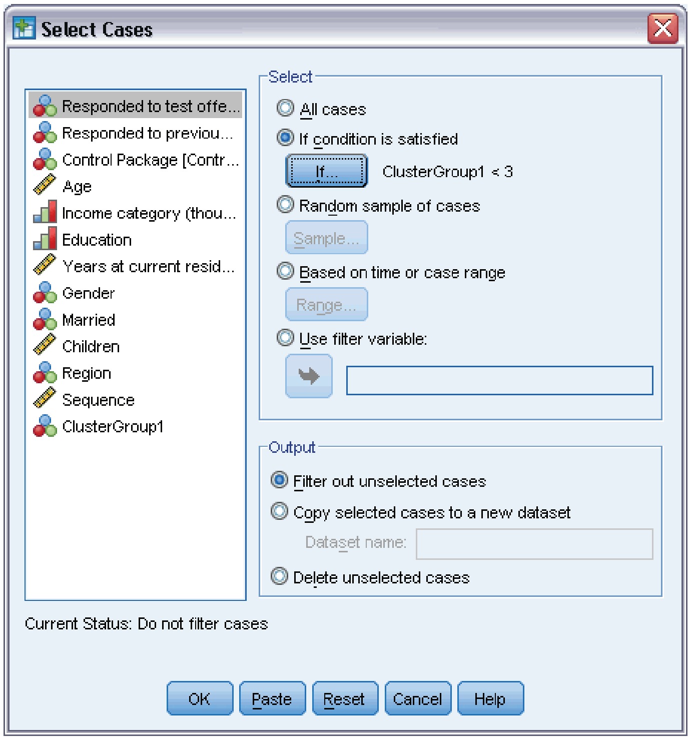


	ID	Gender	Married	Children	Region	ClusterGroup1
1	01359	Female	No	0	West	4
2	06262	Female	No	1	West	2
3	08031	Male	No	0	West	3
4	01971	Male	No	0	West	4
5	09689	Male	No	0	West	3
6	06108	Male	No	1	West	3
7	09853	Male	No	0	West	3
8	06802	Male	No	0	West	4
9	07597	Male	No	0	West	3
10	03692	Male	No	1	West	3
11	00071	Male	No	0	West	4
12	00769	Male	No	0	West	3

To use the values of the cluster field to select records in specific clusters:

▶ From the menus choose:
Data > Select Cases

Figure 9-16
Select Cases dialog



Select Cases

Responded to test offe...
Responded to previou...
Control Package [Contr...
Age
Income category (thou...
Education
Years at current resid...
Gender
Married
Children
Region
Sequence
ClusterGroup1

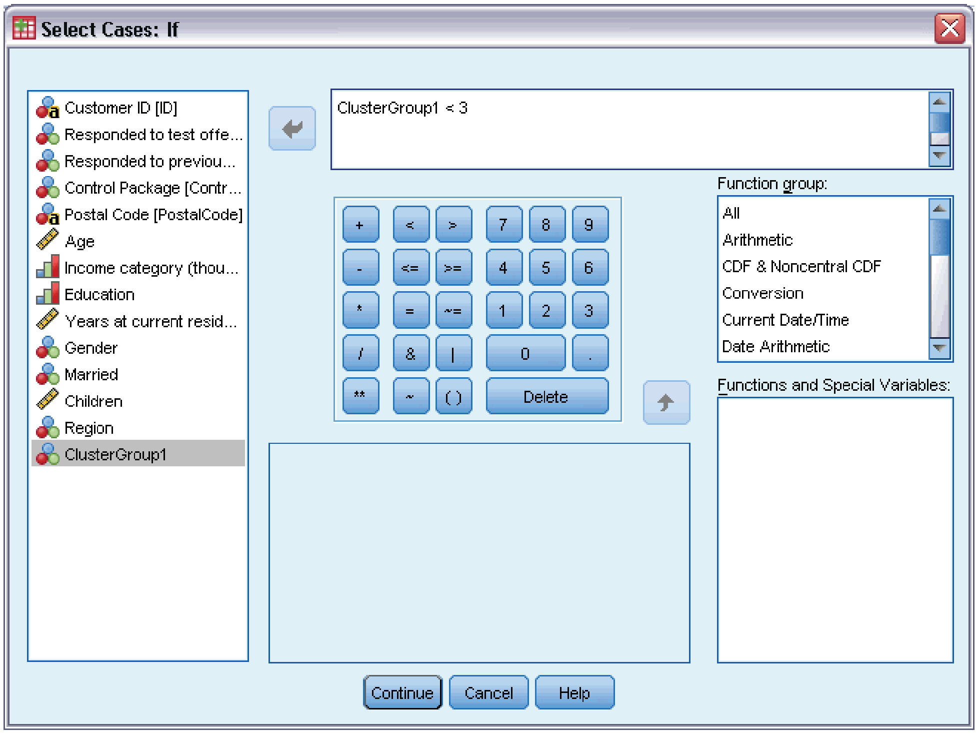
Select
○ All cases
● If condition is satisfied
If... ClusterGroup1 < 3
○ Random sample of cases
Sample...
○ Based on time or case range
Range...
○ Use filter variable:

Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

▶ In the Select Cases dialog, select If condition is satisfied and then click If.

Figure 9-17
Select Cases: If dialog



ClusterGroup1 < 3

- ▶ Enter the selection condition.

For example, ClusterGroup1 < 3 will select all records in clusters 1 and 2, and will exclude records in clusters 3 and higher.

- ▶ Click Continue.

In the Select Cases dialog, there are several options for what to do with selected and unselected records:

Filter out unselected cases. This creates a new field that specifies a filter condition. Excluded records are not deleted from the dataset. They are retained with a filter status indicator, which is displayed as a diagonal slash through the record number in the Data Editor. This is equivalent to interactively selecting clusters in the Cluster Model Viewer.

Copy selected cases to a new dataset. This creates a new dataset in the current session that contains only the records that meet the filter condition. The original dataset is unaffected.

Delete unselected cases. Unselected records are deleted from the dataset. Deleted records can be recovered only by exiting from the file without saving any changes and then reopening the file. The deletion of cases is permanent if you save the changes to the data file.

The Select Cases dialog also has an option to use an existing variable as a filter variable (field). If you create a filter condition interactively in the Cluster Model Viewer and save the generated filter field with the dataset, you can use that field to filter records in subsequent sessions.

Summary

Cluster Analysis is a useful exploratory tool that can reveal natural groupings (or clusters) within your data. You can use the information from these clusters to determine marketing campaign strategies and develop new product offerings. You can select records based on cluster membership for further analysis or targeted marketing campaigns.

Chapter

10

Prospect profiles

Prospect Profiles uses results from a previous or test campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns. For example, based on the results of a test mailing, the direct marketing division of a company wants to generate profiles of the types of people most likely to respond to a certain type of offer, based on demographic information. Based on those results, they can then determine the types of mailing lists they should use for similar offers.

For example, the direct marketing division of a company sends out a test mailing to approximately 20% of their total customer database. The results of this test mailing are recorded in a data file that also contains demographic characteristics for each customer, including age, gender, marital status, and geographic region. The results are recorded in a simple yes/no fashion, indicating which customers in the test mailing responded (made a purchase) and which ones did not.

This information is collected in *dmdata.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

Data considerations

The response field should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values.For more information, see the topic Creating a categorical response field in Chapter 4 on p. 24.

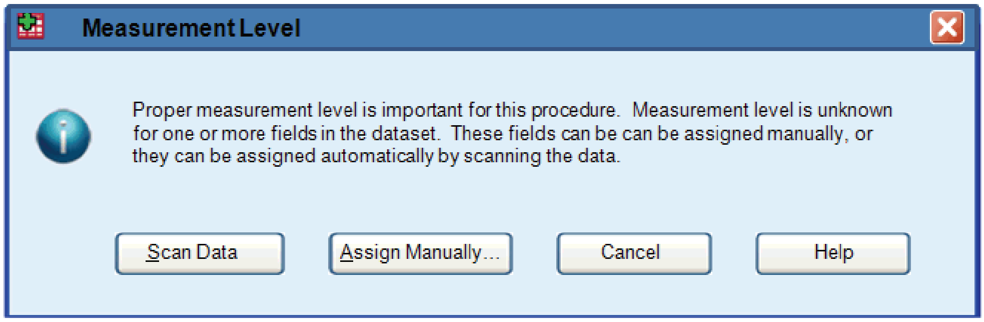
Running the analysis

▶ To run a Prospect Profiles analysis, from the menus choose:
Direct Marketing > Choose Technique

▶ Select Generate profiles of my contacts who responded to an offer and click Continue.

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 10-1
Measurement level alert



Measurement Level

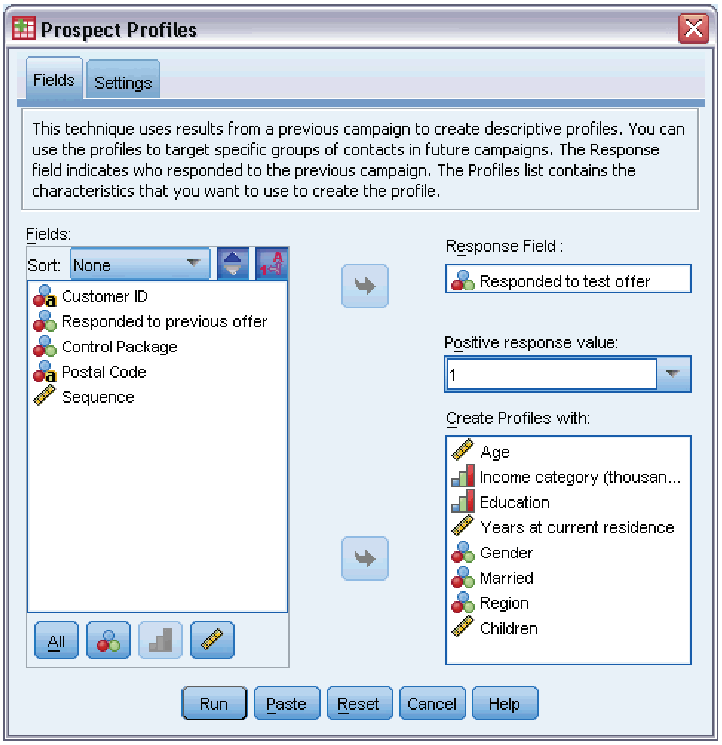
Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be can be assigned manually, or they can be assigned automatically by scanning the data.

- **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.
- **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

In this example file, there are no fields with an unknown measurement level, and all fields have the correct measurement level; so the measurement level alert should not appear.

Figure 10-2
Prospect Profiles, Fields tab



This technique uses results from a previous campaign to create descriptive profiles. You can use the profiles to target specific groups of contacts in future campaigns. The Response field indicates who responded to the previous campaign. The Profiles list contains the characteristics that you want to use to create the profile.

Fields:
- Customer ID
- Responded to previous offer
- Control Package
- Postal Code
- Sequence

Response Field : Responded to test offer

Positive response value: 1

Create Profiles with:
- Age
- Income category (thousan...
- Education
- Years at current residence
- Gender
- Married
- Region
- Children

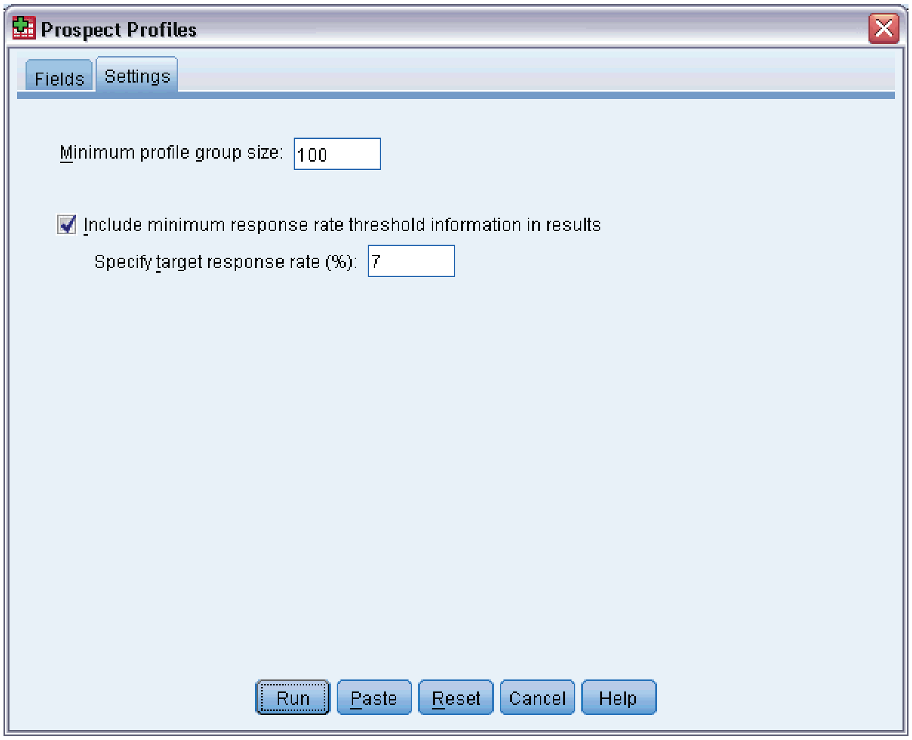
▶ For Response Field, select *Responded to test offer*.

▶ For Positive response value, select *Yes* from the drop-down list. A value of 1 is displayed in the text field because "Yes" is actually a value label associated with a recorded value of 1. (If the positive response value doesn't have a defined value label, you can just enter the value in the text field.)

▶ For Create Profiles with, select *Age*, *Income category*, *Education*, *Years at current residence*, *Gender*, *Married*, *Region*, and *Children*.

▶ Click the Settings tab.

Figure 10-3
Prospect Profiles, Settings tab



Minimum profile group size: 100

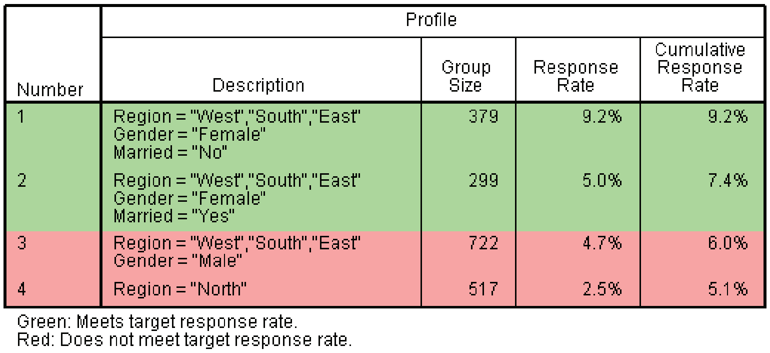
☑ Include minimum response rate threshold information in results

Specify target response rate (%): 7

- ▶ Select (check) Include minimum response rate threshold information in results.
- ▶ For the target response rate, enter 7.
- ▶ Then click Run to run the procedure.

Output

Figure 10-4
Response rate table



	Profile			
Number	Description	Group Size	Response Rate	Cumulative Response Rate

1	Region = "West","South","East"
	Gender = "Female"
	Married = "No"
		379	9.2%	9.2%

2	Region = "West","South","East"
	Gender = "Female"
	Married = "Yes"
		299	5.0%	7.4%

3	Region = "West","South","East"
	Gender = "Male"
		722	4.7%	6.0%

4	Region = "North"	517	2.5%	5.1%

Green: Meets target response rate.
Red: Does not meet target response rate.

The response rate table displays information for each profile group identified by the procedure.

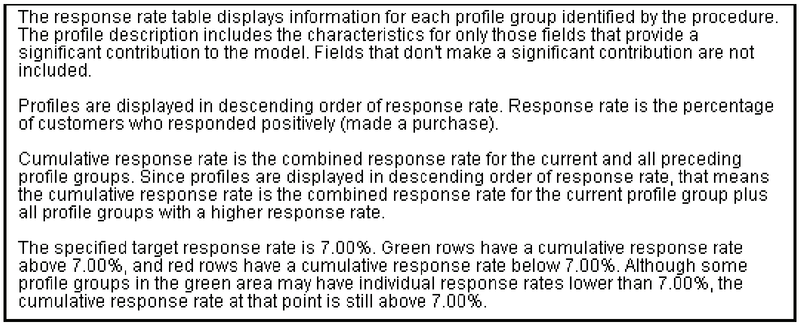
- Profiles are displayed in descending order or response rate.
- Response rate is the percentage of customer who responded positively (made a purchase).
- Cumulative response rate is the combined response rate for the current and all preceding profile groups. Since profiles are displayed in descending order of response rate, that means the cumulative response rate is the combined response rate for the current profile group plus all profile groups with a higher response rate.
- The profile description includes the characteristics for only those fields that provide a significant contribution to the model. In this example, region, gender, and marital status are included in the model. The remaining fields — age, income, education, and years at current address — are not included because they did not make a significant contribution to the model.
- The green area of the table represents the set of profiles with a cumulative response rate equal to or greater than the specified target response rate, which in this example is 7%.
- The red area of the table represents the set of profiles with a cumulative response rate lower than the specified target response rate.
- The cumulative response rate in the last row of the table is the overall or average response rate for all customers included in the test mailing, since it is the response rate for all profile groups.

The results displayed in the table suggest that if you target females in the west, south, and east, you should get a response rate slightly higher than the target response rate.

Note, however, that there is a substantial difference between the response rates for unmarried females (9.2%) and married females (5.0%) in those regions. Although the cumulative response rate for both groups is above the target response rate, the response rate for the latter group alone is, in fact, lower than the target response rate, which suggests that you may want to look for other characteristics that might improve the model.

Smart output

Figure 10-5
Smart output



The response rate table displays information for each profile group identified by the procedure. The profile description includes the characteristics for only those fields that provide a significant contribution to the model. Fields that don't make a significant contribution are not included.

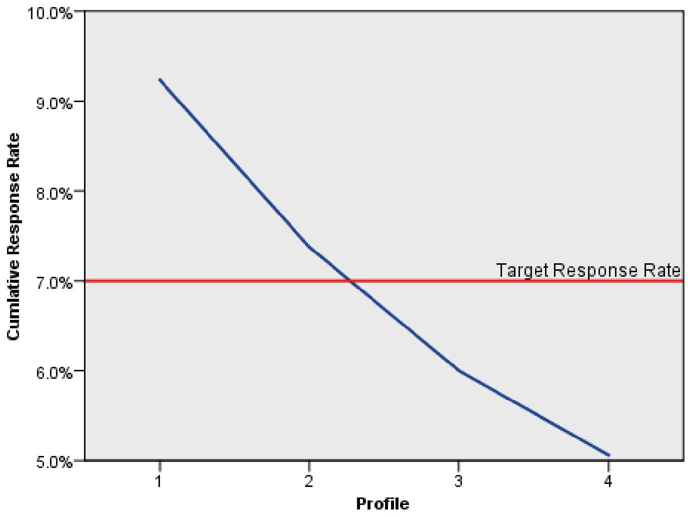
Profiles are displayed in descending order of response rate. Response rate is the percentage of customers who responded positively (made a purchase).

Cumulative response rate is the combined response rate for the current and all preceding profile groups. Since profiles are displayed in descending order of response rate, that means the cumulative response rate is the combined response rate for the current profile group plus all profile groups with a higher response rate.

The specified target response rate is 7.00%. Green rows have a cumulative response rate above 7.00%, and red rows have a cumulative response rate below 7.00%. Although some profile groups in the green area may have individual response rates lower than 7.00%, the cumulative response rate at that point is still above 7.00%.

The table is accompanied by "smart output" that provide general information on how to interpret the table and specific information on the results contained in the table.

Figure 10-6
Cumulative response rate chart



The cumulative response rate chart is basically a visual representation of the cumulative response rates displayed in the table. Since profiles are reported in descending order of response rate, the cumulative response rate line always goes down for each subsequent profile. Just like the table, the chart shows that the cumulative response rate drops below the target response rate between profile group 2 and profile group 3.

Summary

For this particular test mailing, four profile groups were identified, and the results indicate that the only significant demographic characteristics that seem to be related to whether or not a person responded to the offer are gender, region, and marital status. The group with the highest response rate consists of unmarried females, living in the south, east, and west. After that, response rates drop off rapidly, although including married females in the same regions still yields a cumulative response rate higher than the target response rate.

Chapter

11

Postal code response rates

This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns.

For example, based on the results of a previous mailing, the direct marketing division of a company generates response rates by postal codes. Based on various criteria, such as a minimum acceptable response rate and/or maximum number of contacts to include in the mailing, they can then target specific postal codes.

This information is collected in *dmdata.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

Data considerations

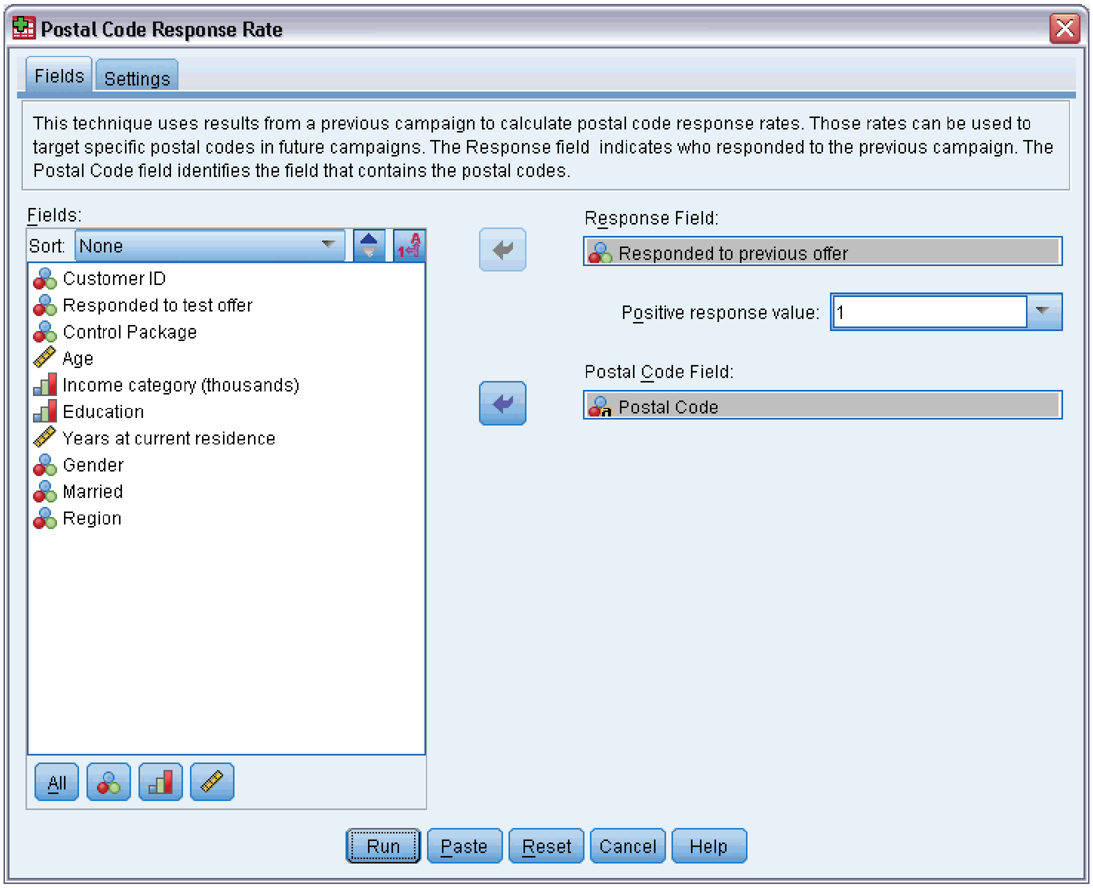
The response field should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values. For more information, see the topic Creating a Categorical Response Field in Chapter 5 on p. 31.

Running the analysis

▶ To calculate postal code response rates, from the menus choose:
Direct Marketing > Choose Technique

▶ Select Identify the top respondng postal codes and click Continue.

Figure 11-1
Postal Code Response Rates, Fields tab



Postal Code Response Rate

This technique uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns. The Response field indicates who responded to the previous campaign. The Postal Code field identifies the field that contains the postal codes.

Fields:
Sort: None

- Customer ID
- Responded to test offer
- Control Package
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Region

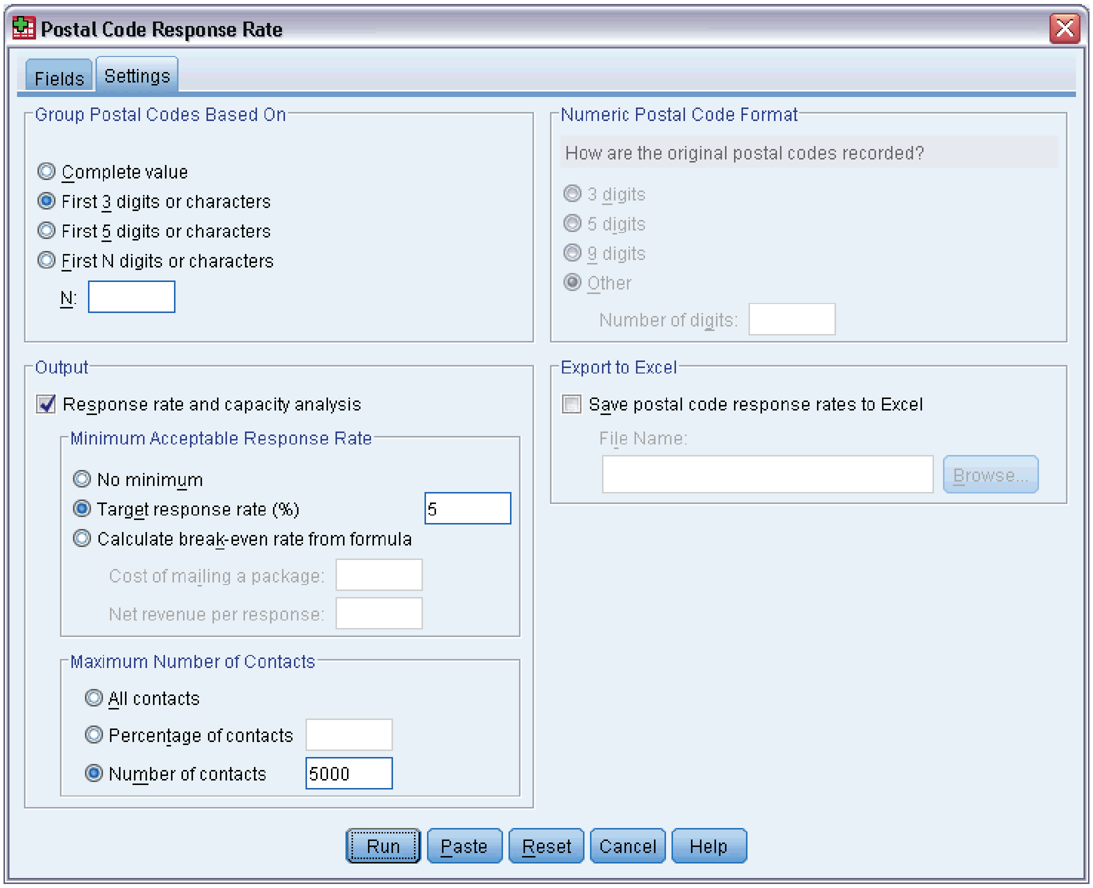
Response Field: Responded to previous offer

Positive response value: 1

Postal Code Field: Postal Code

- ▶ For Response Field, select *Responded to previous offer*.
- ▶ For Positive response value, select *Yes* from the drop-down list. A value of 1 is displayed in the text field because “Yes” is actually a value label associated with a recorded value of 1. (If the positive response value doesn’t have a defined value label, you can just enter the value in the text field.)
- ▶ For Postal Code Field, select *Postal Code*.
- ▶ Click the Settings tab.

Figure 11-2
Postal Code Response Rates, Settings tab



Postal Code Response Rate

Fields | Settings

Group Postal Codes Based On

○ Complete value
● First 3 digits or characters
○ First 5 digits or characters
○ First N digits or characters

N:

Numeric Postal Code Format

How are the original postal codes recorded?

○ 3 digits
○ 5 digits
○ 9 digits
● Other

Number of digits:

Output

☑ Response rate and capacity analysis

Minimum Acceptable Response Rate

○ No minimum
● Target response rate (%): 5
○ Calculate break-even rate from formula

Cost of mailing a package:

Net revenue per response:

Maximum Number of Contacts

○ All contacts
○ Percentage of contacts
● Number of contacts: 5000

Export to Excel

□ Save postal code response rates to Excel

File Name:

- ▶ In the Group Postal Codes Based On group, select First 3 digits or characters. This will calculate combined response rates for all contacts that have postal codes that start with the same three digits or characters. For example, the first three digits of a U.S. zip code represent a common geographic area that is larger than the geographic area defined by the full 5-digit zip code.

- ▶ In the Output group, select (check) Response rate and capacity analysis.

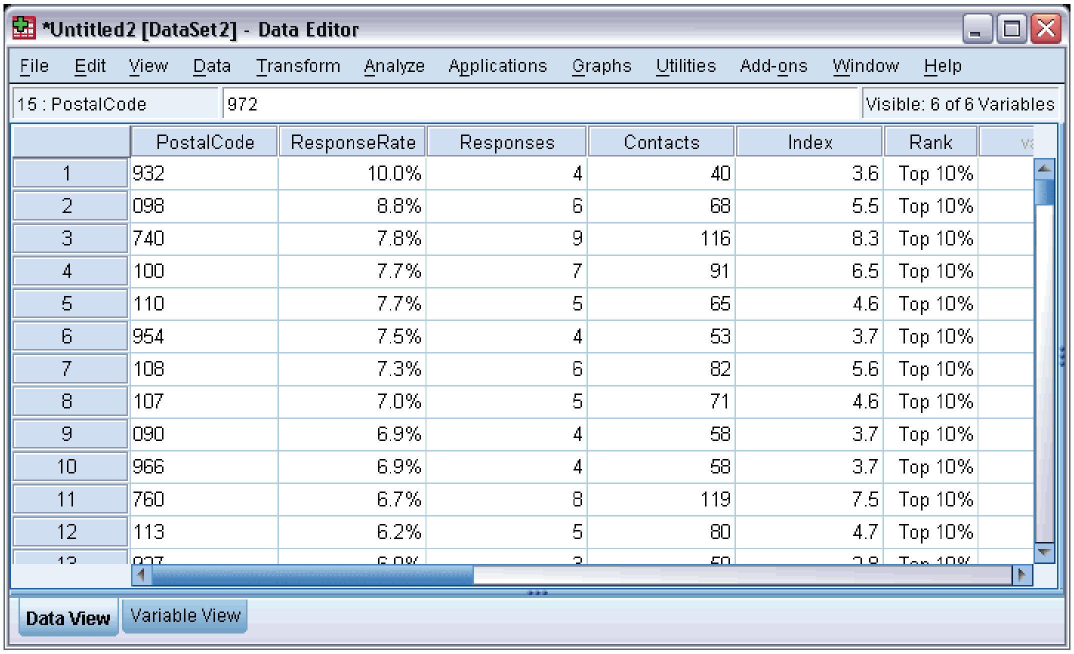
- ▶ Select Target response rate and enter a value of 5.

- ▶ Select Number of contacts and enter a value of 5000.

- ▶ Then click Run to run the procedure.

Output

Figure 11-3
New dataset with response rates by postal code



*Untitled2 [DataSet2] - Data Editor
File Edit View Data Transform Analyze Applications Graphs Utilities Add-ons Window Help
15 : PostalCode
972
Visible: 6 of 6 Variables
Data View
Variable View

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank	va
1	932	10.0%	4	40	3.6	Top 10%	
2	098	8.8%	6	68	5.5	Top 10%	
3	740	7.8%	9	116	8.3	Top 10%	
4	100	7.7%	7	91	6.5	Top 10%	
5	110	7.7%	5	65	4.6	Top 10%	
6	954	7.5%	4	53	3.7	Top 10%	
7	108	7.3%	6	82	5.6	Top 10%	
8	107	7.0%	5	71	4.6	Top 10%	
9	090	6.9%	4	58	3.7	Top 10%	
10	966	6.9%	4	58	3.7	Top 10%	
11	760	6.7%	8	119	7.5	Top 10%	
12	113	6.2%	5	80	4.7	Top 10%	
13	927	6.0%	3	50	3.8	Top 10%	

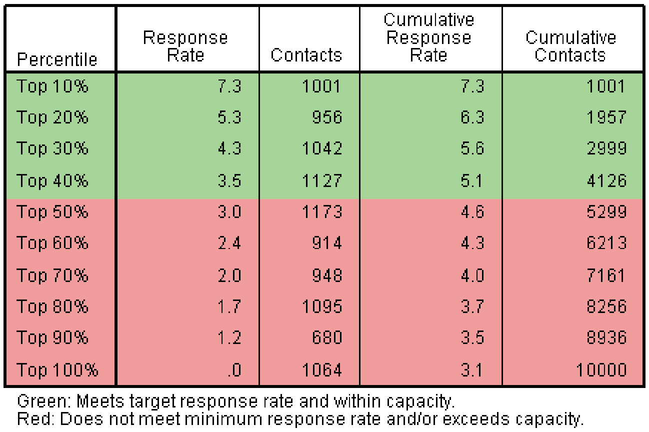
A new dataset is automatically created. This dataset contains a single record (row) for each postal code. In this example, each row contains summary information for all postal codes that start with the same first three digits or characters.

In addition to the field that contains the postal code, the new dataset contains the following fields:

- **ResponseRate.** The percentage of positive responses in each postal code. Records are automatically sorted in descending order of response rates; so postal codes that have the highest response rates appear at the top of the dataset.
- **Responses.** The number of positive responses in each postal code.
- **Contacts.** The total number of contacts in each postal code that contain a non-missing value for the response field.
- **Index.** The "weighted" response based on the formula *N x P x (1-P)*, where *N* is the number of contacts, and *P* is the response rate expressed as a proportion. For two postal codes with the same response rate, this formula will assign a higher index value to the postal code with the larger number of contacts.
- **Rank.** Decile rank (top 10%, top 20%, etc.) of the cumulative postal code response rates in descending order.

Since we selected Response rate and capacity analysis on the Settings tab of the Postal Code Response Rates dialog, a summary response rate table and chart are displayed in the Viewer.

Figure 11-4
Response rate table



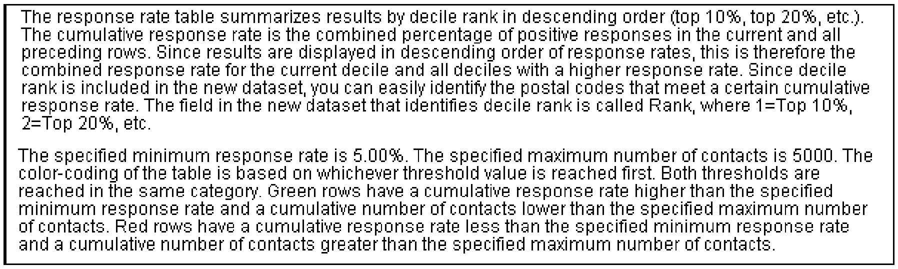
Percentile	Response Rate	Contacts	Cumulative Response Rate	Cumulative Contacts
Top 10%	7.3	1001	7.3	1001
Top 20%	5.3	956	6.3	1957
Top 30%	4.3	1042	5.6	2999
Top 40%	3.5	1127	5.1	4126
Top 50%	3.0	1173	4.6	5299
Top 60%	2.4	914	4.3	6213
Top 70%	2.0	948	4.0	7161
Top 80%	1.7	1095	3.7	8256
Top 90%	1.2	680	3.5	8936
Top 100%	.0	1064	3.1	10000

Green: Meets target response rate and within capacity.
Red: Does not meet minimum response rate and/or exceeds capacity.

The table summarizes results by decile rank in descending order (top 10%, top 20%, etc.).

- The cumulative response rate is the combined percentage of positive responses in the current and all preceding rows. Since results are displayed in descending order of response rates, this is therefore the combined response rate for the current decile and all deciles with a higher response rate.
- The table is color-coded based on the values you entered for target response rate and maximum number of contacts. Rows with a cumulative response rate equal to or greater than 5% and 5,000 or fewer cumulative contacts are colored green. The color-coding is based on whichever threshold value is reached first. In this example, both threshold values are reached in the same decile.

Figure 11-5
Smart output for response rate table

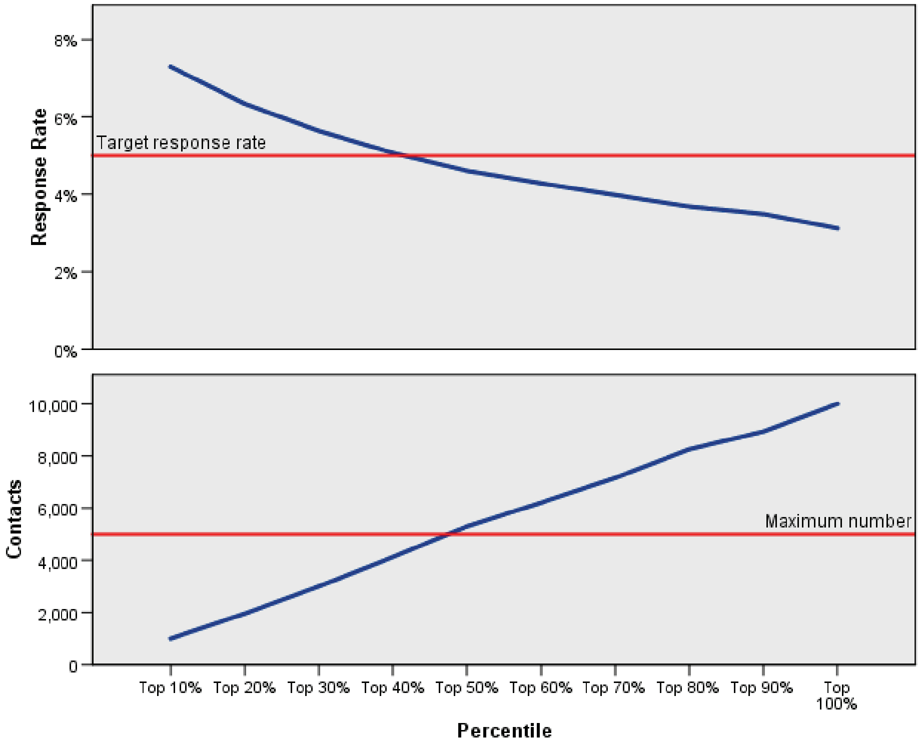


The response rate table summarizes results by decile rank in descending order (top 10%, top 20%, etc.). The cumulative response rate is the combined percentage of positive responses in the current and all preceding rows. Since results are displayed in descending order of response rates, this is therefore the combined response rate for the current decile and all deciles with a higher response rate. Since decile rank is included in the new dataset, you can easily identify the postal codes that meet a certain cumulative response rate. The field in the new dataset that identifies decile rank is called Rank, where 1=Top 10%, 2=Top 20%, etc.

The specified minimum response rate is 5.00%. The specified maximum number of contacts is 5000. The color-coding of the table is based on whichever threshold value is reached first. Both thresholds are reached in the same category. Green rows have a cumulative response rate higher than the specified minimum response rate and a cumulative number of contacts lower than the specified maximum number of contacts. Red rows have a cumulative response rate less than the specified minimum response rate and a cumulative number of contacts greater than the specified maximum number of contacts.

The table is accompanied by text that provides a general description of how to read the table. If you have specified either a minimum response rate or a maximum number of contacts, it also includes a section describing how the results relate to the threshold values you specified.

Figure 11-6
Cumulative response rate chart

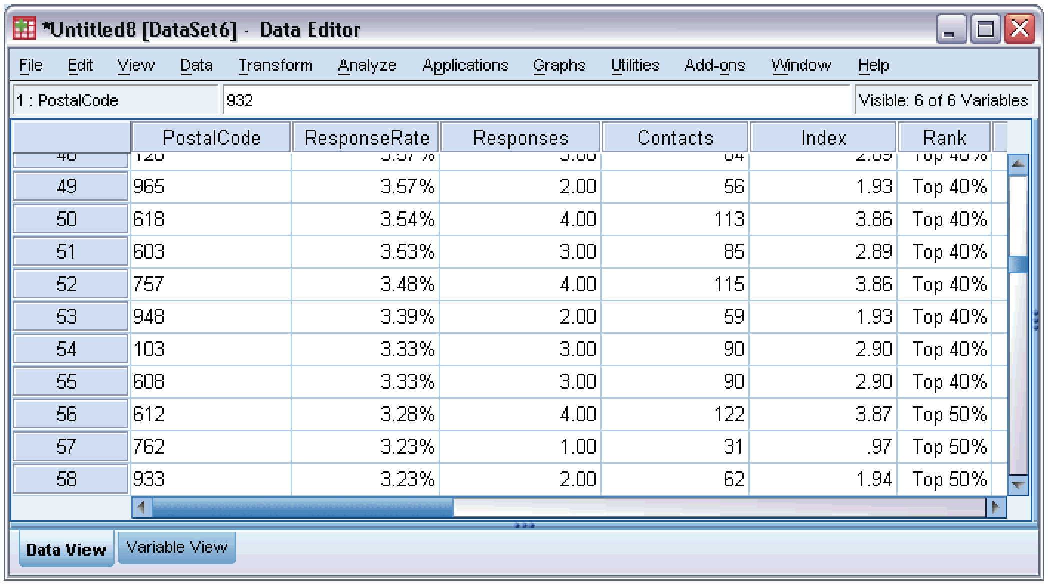


The chart of cumulative response rate and cumulative number of contacts in each decile is a visual representation of the same information displayed in the response rate table. The threshold for both minimum cumumlative response rate and maximum cumulative number of contacts is reached somewhere between the 40th and 50th percentile.

- Since the chart displays cumulative response rates in descending order of decile rank of response rate, the cumulative response rate line always goes down for each subsequent decile.
- Since the line for number of contacts represents cumulative number of contacts, it always goes up.

The information in the table and chart tell you that if you are want to achieve a response rate of at least 5% but don't want to include more than 5,000 contacts in the campaign, you should focus on the postal codes in the top four deciles. Since decile rank is included in the new dataset, you can easily identify the postal codes that meet the top 40% requirement.

Figure 11-7
New dataset



1 : PostalCode 932

Visible: 6 of 6 Variables

	PostalCode	ResponseRate	Responses	Contacts	Index	Rank
48	128	3.57%	3.00	84	2.89	Top 40%
49	965	3.57%	2.00	56	1.93	Top 40%
50	618	3.54%	4.00	113	3.86	Top 40%
51	603	3.53%	3.00	85	2.89	Top 40%
52	757	3.48%	4.00	115	3.86	Top 40%
53	948	3.39%	2.00	59	1.93	Top 40%
54	103	3.33%	3.00	90	2.90	Top 40%
55	608	3.33%	3.00	90	2.90	Top 40%
56	612	3.28%	4.00	122	3.87	Top 50%
57	762	3.23%	1.00	31	.97	Top 50%
58	933	3.23%	2.00	62	1.94	Top 50%

Note: *Rank* is recorded as an integer value from 1 to 10. The field has defined value labels, where 1= Top 10%, 2=Top 20%, etc. You will see either the actual rank values or the value labels in Data View of the Data Editor, depending on your View settings.

Summary

The Postal Code Response Rates procedure uses results from a previous campaign to calculate postal code response rates. Those rates can be used to target specific postal codes in future campaigns. The procedure creates a new dataset that contains response rates for each postal code. Based on information in the response rate table and chart and decile rank information in the new dataset, you can identify the set of postal codes that meet a specified minimum cumulative response rate and/or cumulative maximum number of contacts.

Chapter

12

Propensity to purchase

Propensity to Purchase uses results from a test mailing or previous campaign to generate propensity scores. The scores indicate which contacts are most likely to respond, based on various selected characteristics.

This technique uses binary logistic regression to build a predictive model. The process of building and applying a predictive model has two basic steps:

- ▶ Build the model and save the model file. You build the model using a dataset for which the outcome of interest (often referred to as the **target**) is known. For example, if you want to build a model that will predict who is likely to respond to a direct mail campaign, you need to start with a dataset that already contains information on who responded and who did not respond. For example, this might be the results of a test mailing to a small group of customers or information on responses to a similar campaign in the past.

- ▶ Apply that model to a different dataset (for which the outcome of interest is not known) to obtain predicted outcomes.

This example uses two data files: *dmdata2.sav* is used to build the model, and then that model is applied to *dmdata3.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

Data considerations

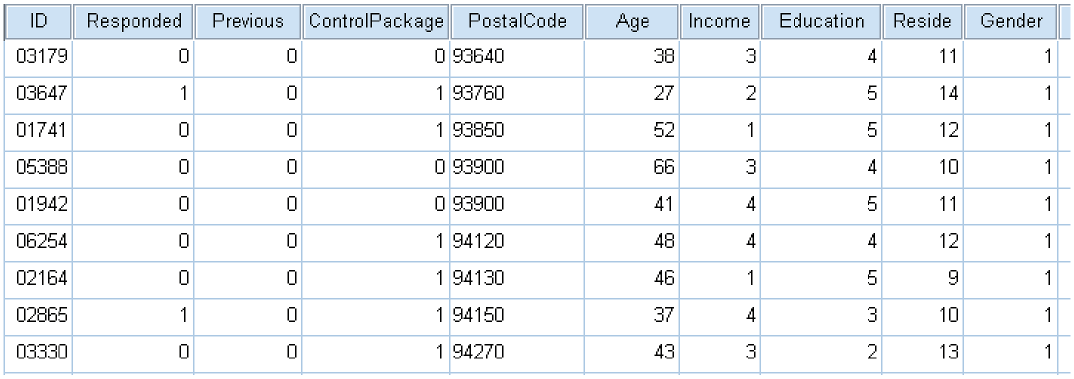
The response field (the target outcome of interest) should be categorical, with one value representing all positive responses. Any other non-missing value is assumed to be a negative response. If the response field represents a continuous (scale) value, such as number of purchases or monetary amount of purchases, you need to create a new field that assigns a single positive response value to all non-zero response values.For more information, see the topic Creating a categorical response field in Chapter 6 on p. 38.

Building a predictive model

- ▶ Open the data file *dmdata2.sav*.

This file contains various demographic characteristics of the people who received the test mailing, and it also contains information on whether or not they responded to the mailing. This information is recorded in the field (variable) *Responded*. A value of 1 indicates that the contact responded to the mailing, and a value of 0 indicates that the contact did not respond.

Figure 12-1
Contents of data file in the Data Editor

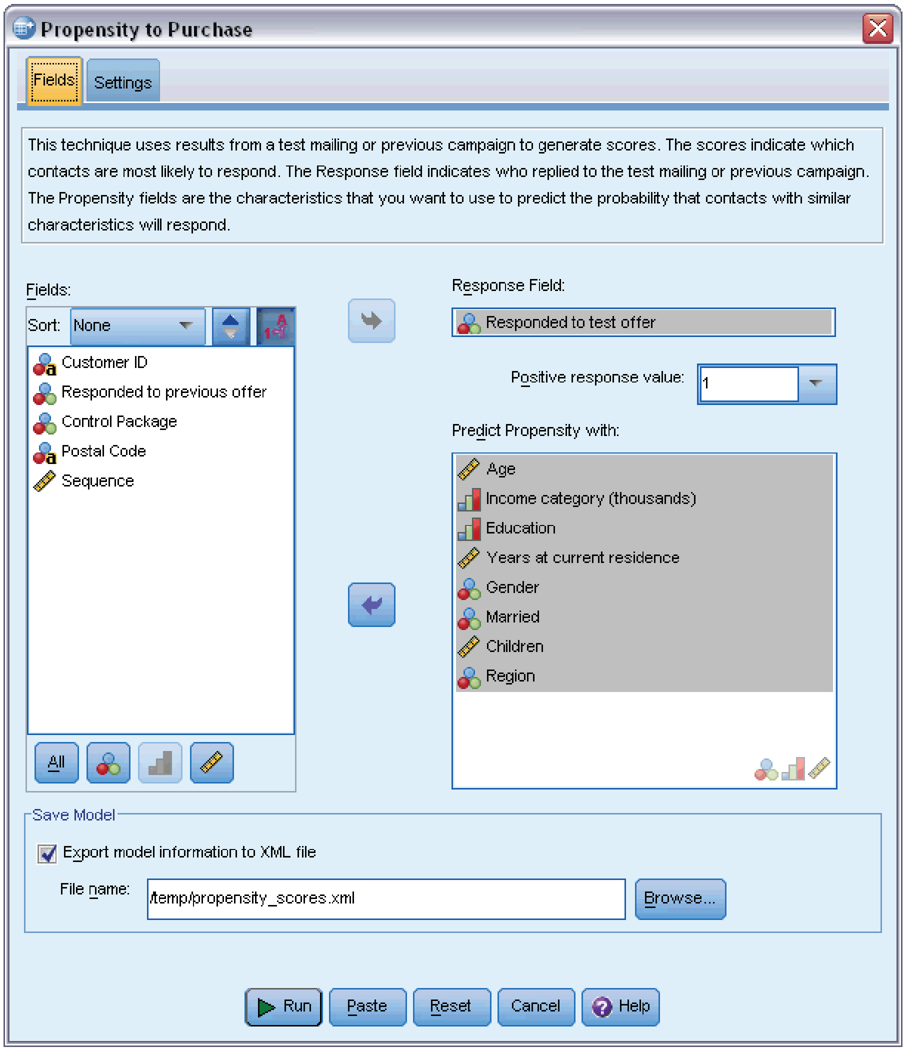


ID	Responded	Previous	ControlPackage	PostalCode	Age	Income	Education	Reside	Gender
03179	0	0	0	93640	38	3	4	11	1
03647	1	0	1	93760	27	2	5	14	1
01741	0	0	1	93850	52	1	5	12	1
05388	0	0	0	93900	66	3	4	10	1
01942	0	0	0	93900	41	4	5	11	1
06254	0	0	1	94120	48	4	4	12	1
02164	0	0	1	94130	46	1	5	9	1
02865	1	0	1	94150	37	4	3	10	1
03330	0	0	1	94270	43	3	2	13	1

▶ From the menus choose:
Direct Marketing > Choose Technique

▶ Select Select contacts most likely to purchase and click Continue.

Figure 12-2
Propensity to Purchase, Fields tab



Propensity to Purchase

Fields | Settings

This technique uses results from a test mailing or previous campaign to generate scores. The scores indicate which contacts are most likely to respond. The Response field indicates who replied to the test mailing or previous campaign. The Propensity fields are the characteristics that you want to use to predict the probability that contacts with similar characteristics will respond.

Fields:

Sort: None

- Customer ID
- Responded to previous offer
- Control Package
- Postal Code
- Sequence

Response Field: Responded to test offer

Positive response value: 1

Predict Propensity with:

- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Children
- Region

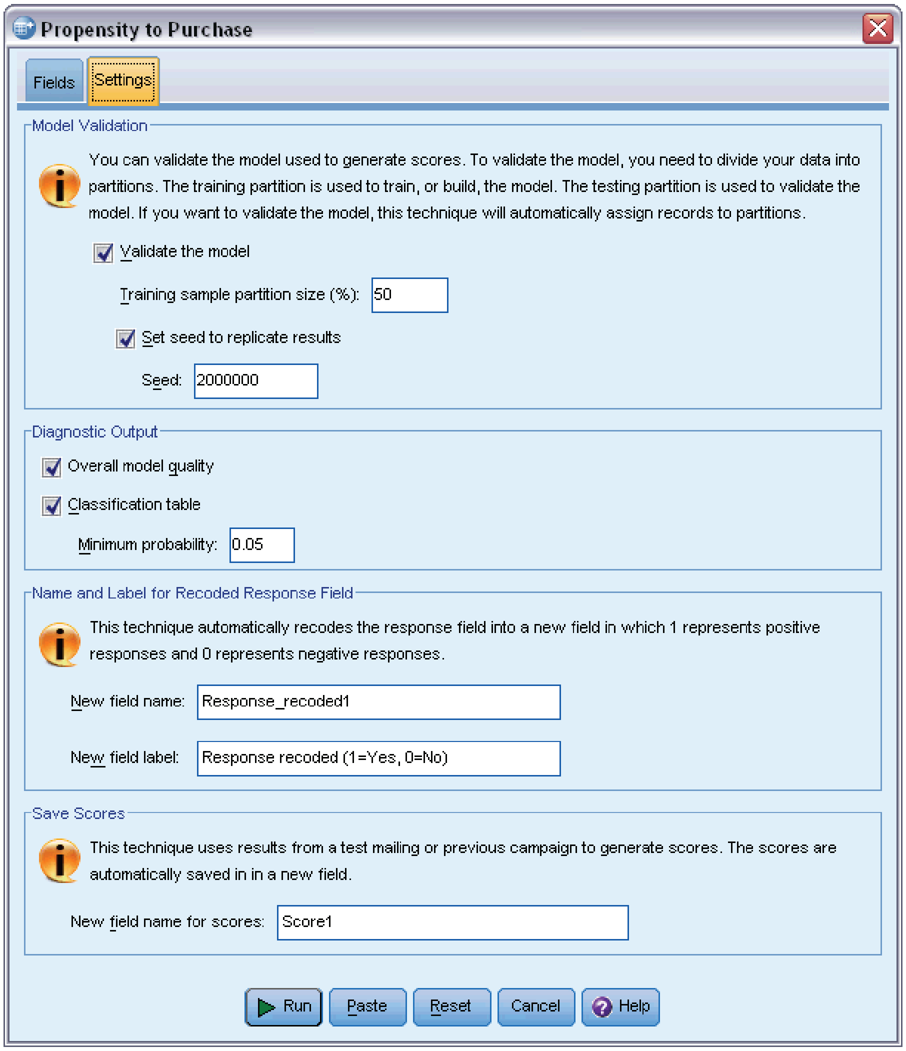
Save Model

☑ Export model information to XML file

File name: /temp/propensity_scores.xml

- ▶ For Response Field, select *Responded to test offer*.
- ▶ For Positive response value, select *Yes* from the drop-down list. A value of 1 is displayed in the text field because “Yes” is actually a value label associated with a recorded value of 1. (If the positive response value doesn’t have a defined value label, you can just enter the value in the text field.)
- ▶ For Predict Propensity with, select *Age*, *Income category*, *Education*, *Years at current residence*, *Gender*, *Married*, *Region*, and *Children*.
- ▶ Select (check) Export model information to XML file.
- ▶ Click Browse to navigate to where you want to save the file and enter a name for the file.
- ▶ In the Propensity to Purchase dialog, click the Settings tab.

Figure 12-3
Propensity to Purchase, Settings tab



Propensity to Purchase

Model Validation

You can validate the model used to generate scores. To validate the model, you need to divide your data into partitions. The training partition is used to train, or build, the model. The testing partition is used to validate the model. If you want to validate the model, this technique will automatically assign records to partitions.

☑ Validate the model

Training sample partition size (%): 50

☑ Set seed to replicate results

Seed: 2000000

Diagnostic Output

☑ Overall model quality

☑ Classification table

Minimum probability: 0.05

Name and Label for Recoded Response Field

This technique automatically recodes the response field into a new field in which 1 represents positive responses and 0 represents negative responses.

New field name: Response_recoded1

New field label: Response recoded (1=Yes, 0=No)

Save Scores

This technique uses results from a test mailing or previous campaign to generate scores. The scores are automatically saved in in a new field.

New field name for scores: Score1

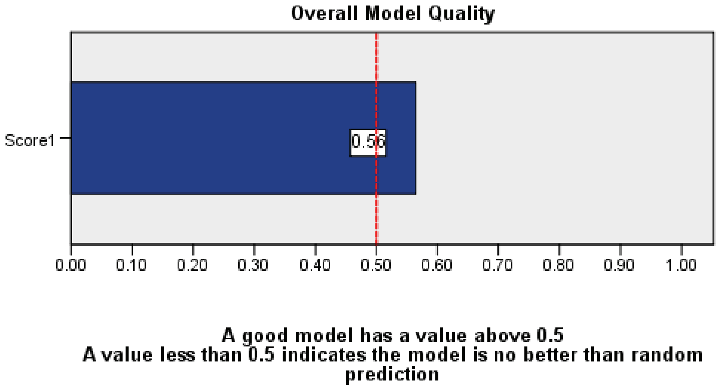
- In the Model Validation Group, select (check) Validate model and Set seed to replicate results.
- Use the default training sample partition size of 50% and the default seed value of 2000000.
- In the Diagnostic Output group, select (check) Overall model quality and Classification table.
- For Minimum probability, enter 0.05. As a general rule, you should specify a value close to your minimum target response rate, expressed as a proportion. A value of 0.05 represents a response rate of 5%.
- Click **Run** to run the procedure and generate the model.

Evaluating the model

Propensity to Purchase produces an overall model quality chart and a classification table that can be used to evaluate the model.

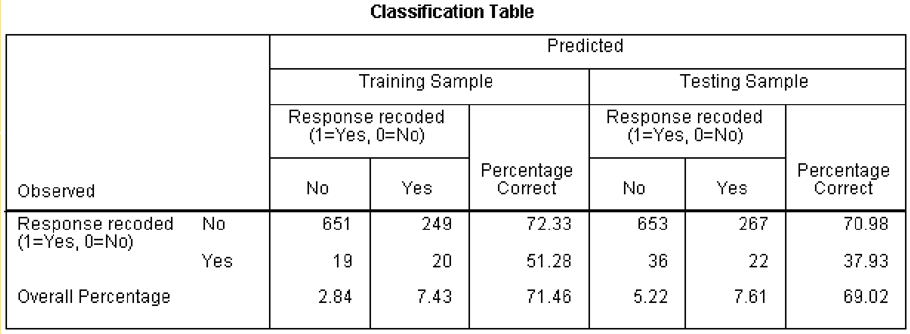
The overall model quality chart provides a quick visual indication of the model quality. As a general rule, the overall model quality should be above 0.5.

Figure 12-4
Overall model quality chart



To confirm that the model is adequate for scoring, you should also examine the classification table.

Figure 12-5
Classification table



Classification Table

		Predicted					
		Training Sample			Testing Sample		
		Response recoded (1=Yes, 0=No)		Percentage Correct	Response recoded (1=Yes, 0=No)		Percentage Correct
Observed		No	Yes		No	Yes	
Response recoded (1=Yes, 0=No)	No	651	249	72.33	653	267	70.98
	Yes	19	20	51.28	36	22	37.93
Overall Percentage		2.84	7.43	71.46	5.22	7.61	69.02

The classification table compares predicted values of the target field to the actual values of the target field. The overall accuracy rate can provide some indication of how well the model works, but you may be more interested in the percentage of correct predicted positive responses, if the goal is to build a model that will identify the group of contacts likely to yield a positive response rate equal to or greater than the specified minimum positive response rate.

In this example, the classification table is split into a **training sample** and a **testing sample**. The training sample is used to build the model. The model is then applied to the testing sample to see how well the model works.

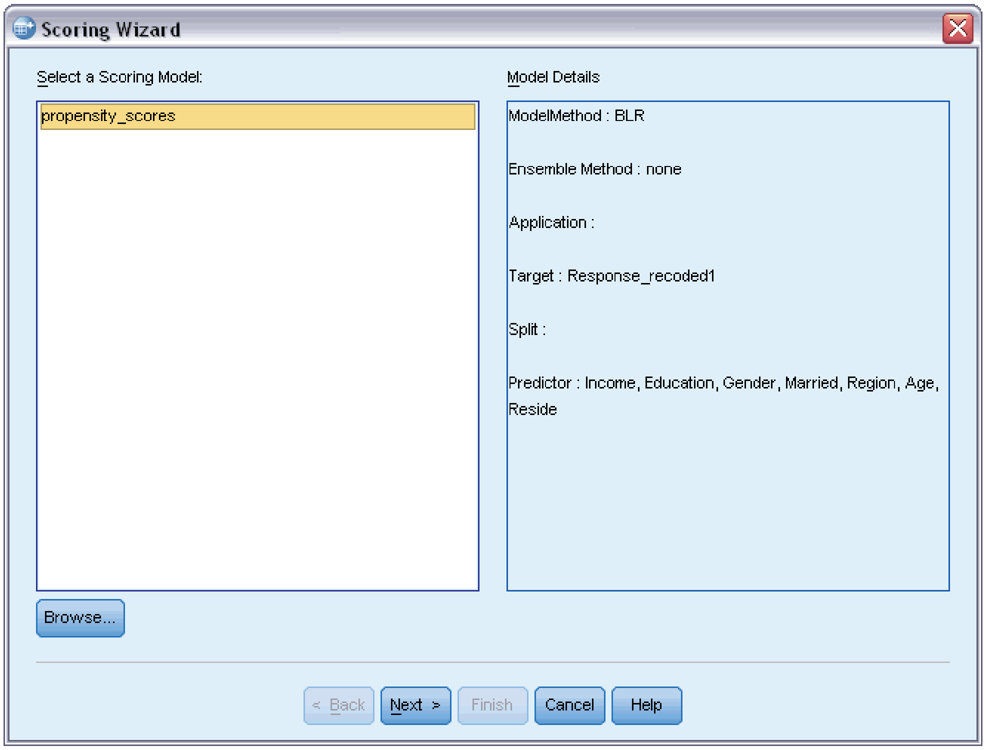
The specified minimum response rate was 0.05 or 5%. The classification table shows that the correct classification rate for positive responses is 7.43% in the training sample and 7.61% in the testing sample. Since the testing sample response rate is greater than 5%, this model should be able to identify a group of contacts likely to yield a response rate greater than 5%.

Applying the model

▶ Open the data file *dmdata3.sav*. This data file contains demographic and other information for all the contacts that were not included in the test mailing. For more information, see the topic Sample Files in Appendix A on p. 96.

▶ Open the Scoring Wizard. To open the Scoring Wizard, from the menus choose:
Utilities > Scoring Wizard

Figure 12-6
Scoring Wizard, Select a Scoring Model



Scoring Wizard

Select a Scoring Model:

propensity_scores

Model Details

ModelMethod : BLR

Ensemble Method : none

Application :

Target : Response_recoded1

Split :

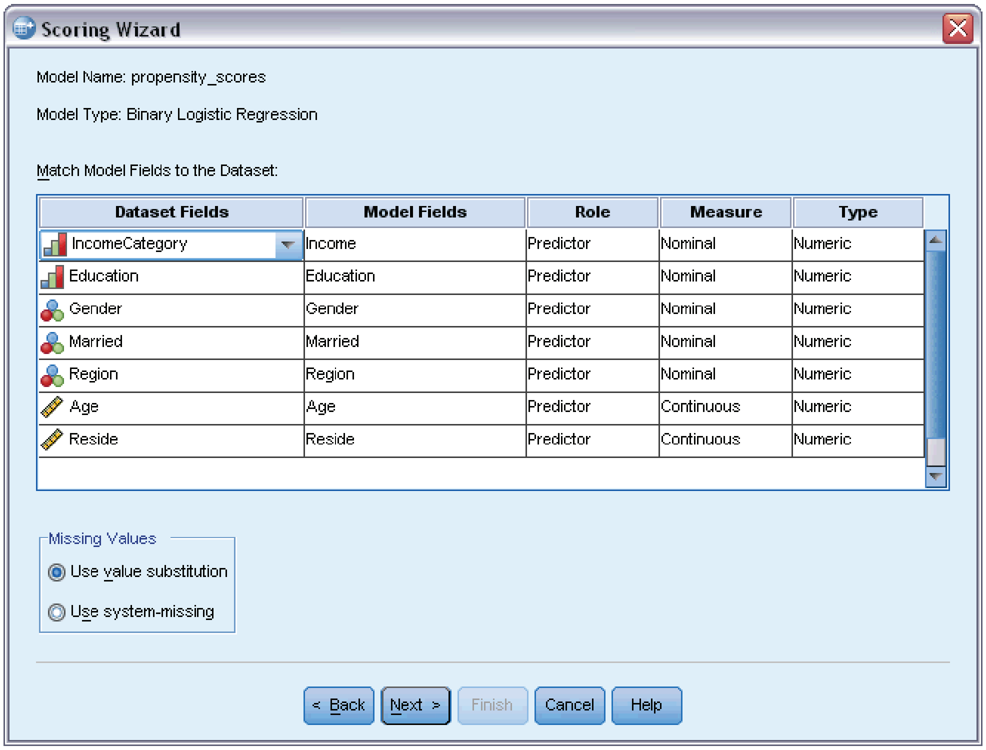
Predictor : Income, Education, Gender, Married, Region, Age, Reside

▶ Click Browse to navigate to the location where you saved the model XML file and click Select in the Browse dialog.

All files with an .xml or .zip extension are displayed in the Scoring Wizard. The extensions are not displayed. If the selected file is recognized as a valid model file, a description of the model is displayed.

▶ Select the model XML file you created and then click Next.

Figure 12-7
Scoring Wizard, Match Model Fields



Model Name: propensity_scores

Model Type: Binary Logistic Regression

Match Model Fields to the Dataset:

Dataset Fields	Model Fields	Role	Measure	Type
IncomeCategory	Income	Predictor	Nominal	Numeric
Education	Education	Predictor	Nominal	Numeric
Gender	Gender	Predictor	Nominal	Numeric
Married	Married	Predictor	Nominal	Numeric
Region	Region	Predictor	Nominal	Numeric
Age	Age	Predictor	Continuous	Numeric
Reside	Reside	Predictor	Continuous	Numeric

Missing Values

● Use value substitution

○ Use system-missing

In order to score the active dataset, the dataset must contain fields (variables) that correspond to all the predictors in the model. If the model also contains split fields, then the dataset must also contain fields that correspond to all the split fields in the model.

- By default, any fields in the active dataset that have the same name and type as fields in the model are automatically matched.
- Use the drop-down list to match dataset fields to model fields. The data type for each field must be the same in both the model and the dataset in order to match fields.
- You cannot continue with the wizard or score the active dataset unless all predictors (and split fields if present) in the model are matched with fields in the active dataset.

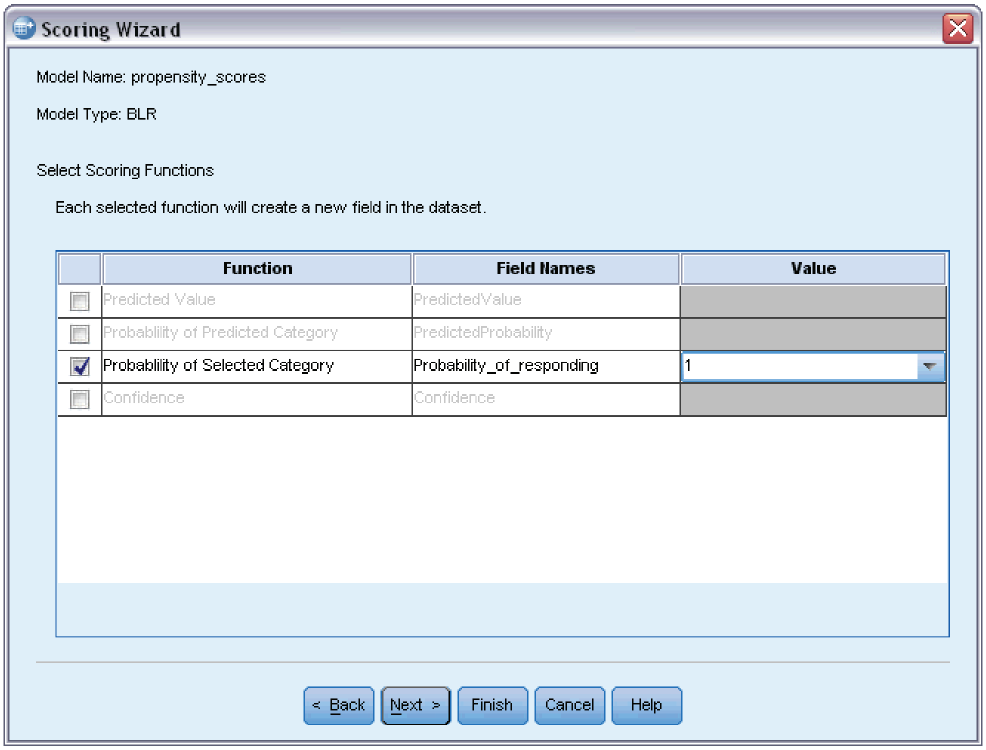
The active dataset does not contain a field named *Income*. So the cell in the Dataset Fields column that corresponds to the model field *Income* is initially blank. You need to select a field in the active dataset that is equivalent to that model field.

▶ From the drop-down list in the Dataset Fields column in the blank cell in the row for the *Income* model field, select *IncomeCategory*.

Note: In addition to field name and type, you should make sure that the actual data values in the dataset being scored are recorded in the same fashion as the data values in the dataset used to build the model. For example, if the model was built with an *Income* field that has income divided into four categories, and *IncomeCategory* in the active dataset has income divided into six categories or four different categories, those fields don't really match each other and the resulting scores will not be reliable.

Click Next to continue to the next step of the Scoring Wizard.

Figure 12-8
Scoring Wizard: Select Scoring Functions



Model Name: propensity_scores

Model Type: BLR

Select Scoring Functions

Each selected function will create a new field in the dataset.

	Function	Field Names	Value
☐	Predicted Value	PredictedValue	
☐	Probability of Predicted Category	PredictedProbability	
☑	Probability of Selected Category	Probability_of_responding	1
☐	Confidence	Confidence	

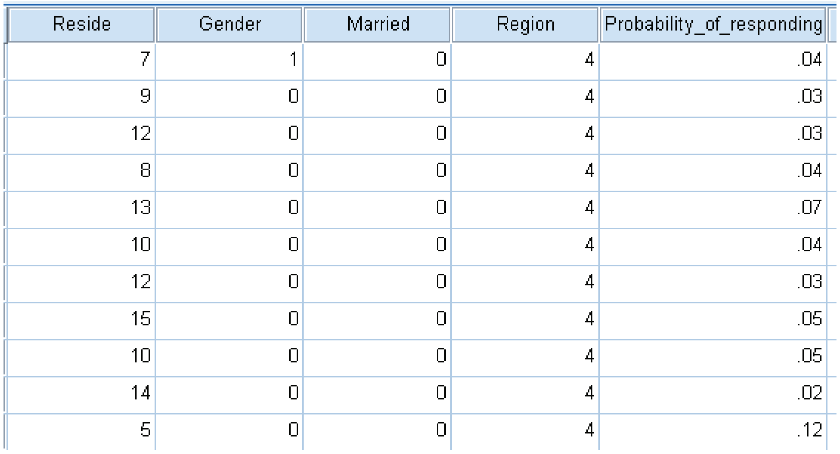
The scoring functions are the types of "scores" available for the selected model. The scoring functions available are dependent on the model. For the binary logistic model used in this example, the available functions are predicted value, probability of the predicted value, probability of a selected value, and confidence.

In this example, we are interested in the predicted probability of a positive response to the mailing; so we want the probability of a selected value.

- ▶ Select (check) Probability of Selected Category.
- ▶ In the Value column, select 1 from the drop-down list. The list of possible values for the target is defined in the model, based on the target values in the data file used to build the model.
- ▶ Deselect (uncheck) all the other scoring functions.
- ▶ Optionally, you can assign a more descriptive name to the new field that will contain the score values in the active dataset. For example, *Probability_of_responding*.
- ▶ Click Finish to apply the model to the active dataset.

The new field that contains the probability of a positive response is appended to the end of the dataset.

Figure 12-9
Dataset with new probability field



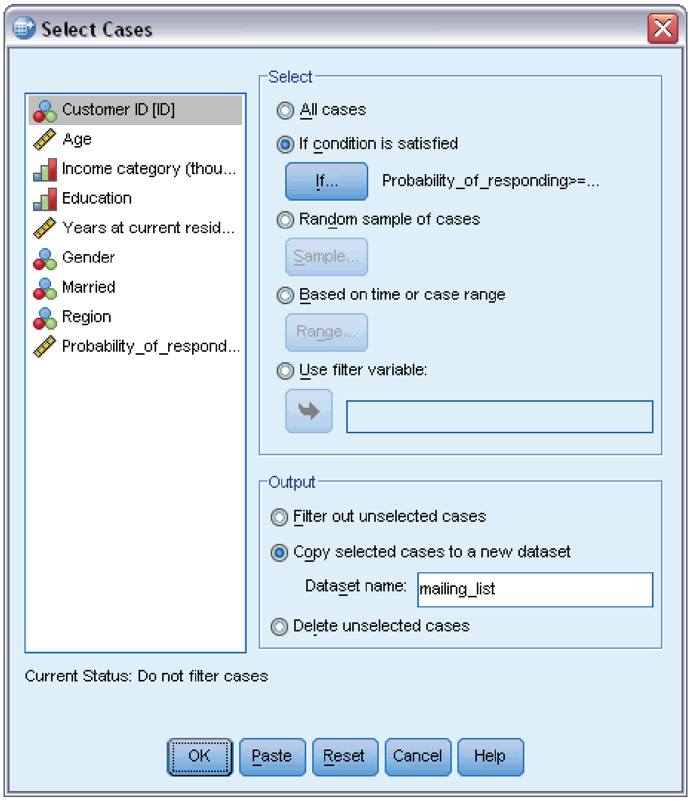
Reside	Gender	Married	Region	Probability_of_responding
7	1	0	4	.04
9	0	0	4	.03
12	0	0	4	.03
8	0	0	4	.04
13	0	0	4	.07
10	0	0	4	.04
12	0	0	4	.03
15	0	0	4	.05
10	0	0	4	.05
14	0	0	4	.02
5	0	0	4	.12

You can then use that field to select the subset of contacts that are likely to yield a positive response rate at or above a certain level. For example, you could create a new dataset that contains the subset of cases likely to yield a positive response rate of at least 5%.

- ▶ From the menus choose:

  Data > Select Cases

Figure 12-10
Select Cases dialog



● If condition is satisfied

If... Probability_of_responding>=...

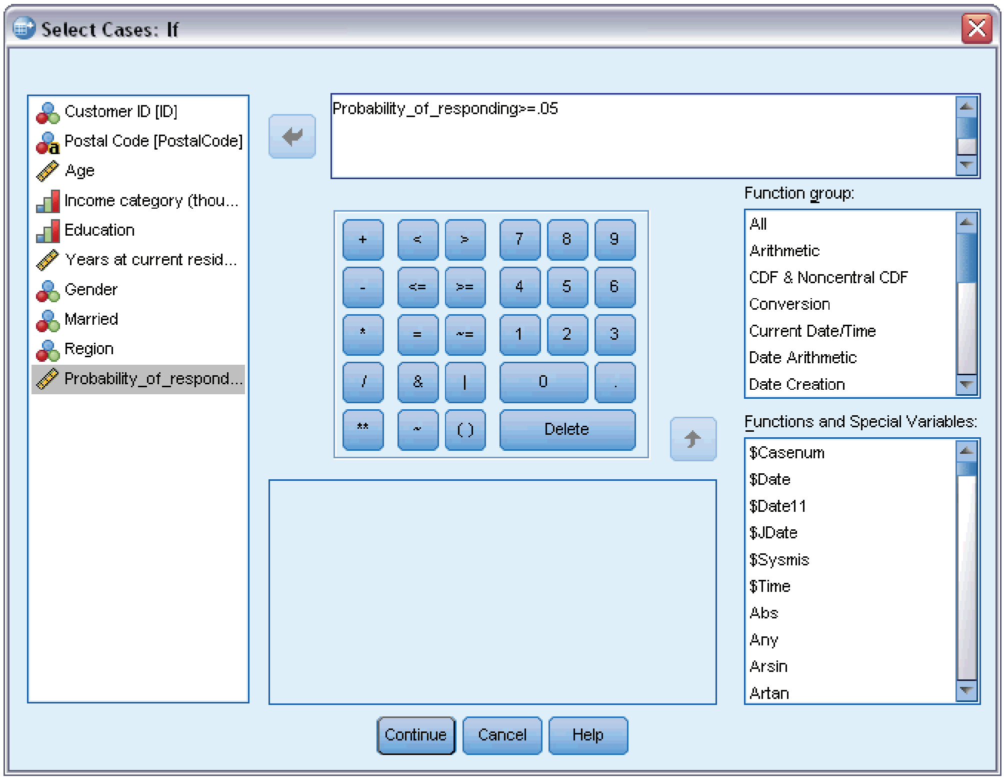
● Copy selected cases to a new dataset

Dataset name: mailing_list

Current Status: Do not filter cases

- ▶ In the Select Cases dialog, select If condition is satisfied and click If.

Figure 12-11
Select Cases: If dialog



Probability_of_responding>=.05

- ▶ In the Select Cases: If dialog enter the following expression:

  Probability_of_responding >=.05

  Note: If you used a different name for the field that contains the probability values, enter that name instead of Probability_of_responding. The default name is *SelectedProbability*.

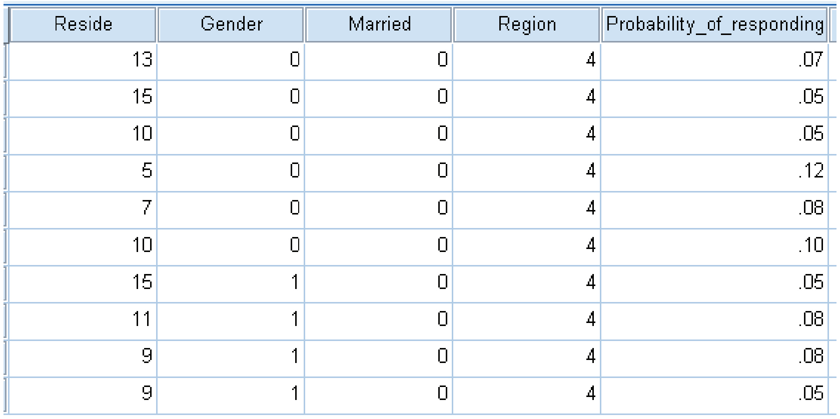
- ▶ Click Continue.

- ▶ In the Select Cases dialog, select Copy selected cases to a new dataset and enter a name for the new dataset. Dataset names must conform to field (variable) naming rules.

- ▶ Click OK to create the dataset with the selected contacts.

The new dataset contains only those contacts with a predicted probability of a positive response of at least 5%.

Figure 12-12
New dataset with selected contacts



Reside	Gender	Married	Region	Probability_of_responding
13	0	0	4	.07
15	0	0	4	.05
10	0	0	4	.05
5	0	0	4	.12
7	0	0	4	.08
10	0	0	4	.10
15	1	0	4	.05
11	1	0	4	.08
9	1	0	4	.08
9	1	0	4	.05

Summary

Propensity to Purchase uses results from a test mailing or previous campaign to generate propensity scores. The scores indicate which contacts are most likely to respond, based on various selected characteristics. This techniques builds a predictive model that can then be applied to dataset to obtain propensity scores.

Chapter

13

Control package test

This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. Campaign effectiveness is measured by responses.

For example, The direct marketing division of a company wants to see if a new package design will generate more positive responses than the existing package. So they send out a test mailing to determine if the new package generates a significantly higher positive response rate. The test mailing consists of a control group that receives the existing package and a test group that receives the new package design. The results for the two groups are then compared to see if there is a significant difference.

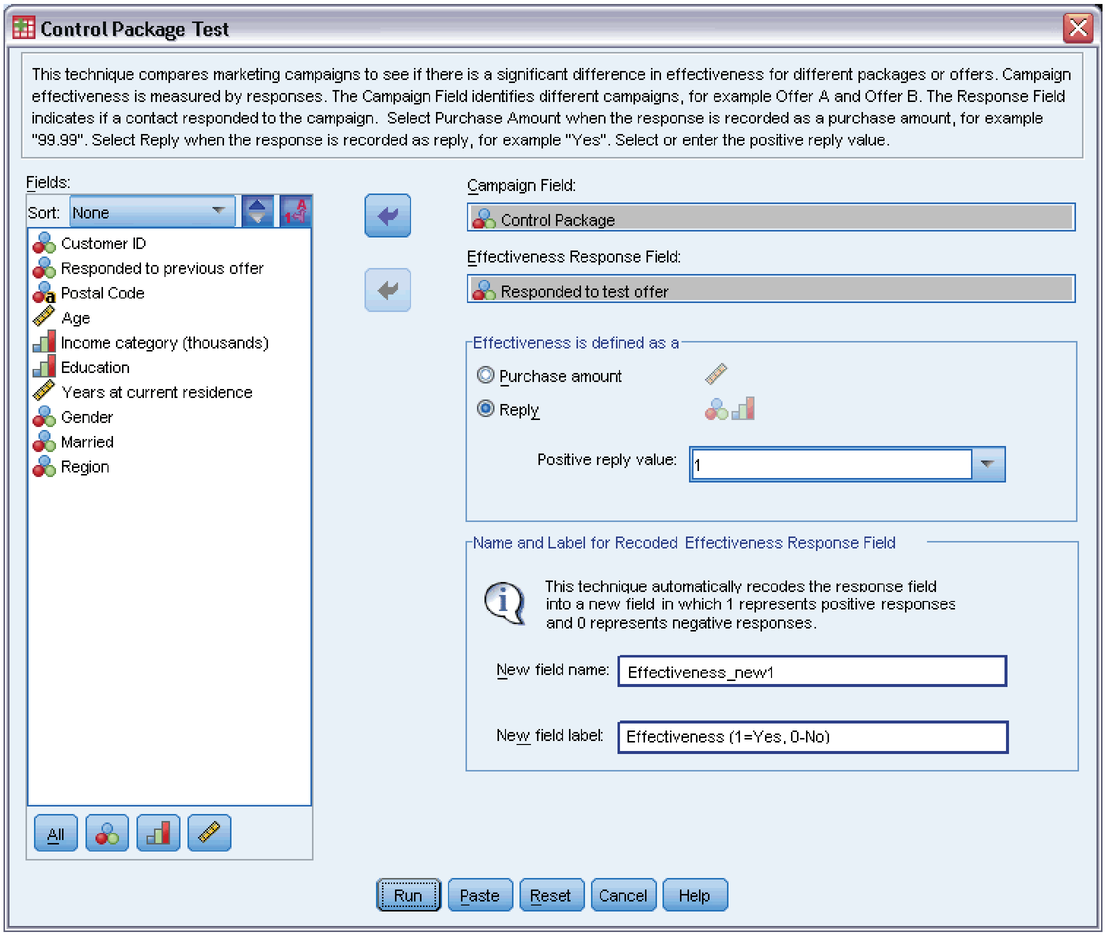
This information is collected in *dmdata.sav*. For more information, see the topic Sample Files in Appendix A on p. 96.

Running the analysis

- ▶ To obtain a control package test, from the menus choose:
  Direct Marketing > Choose Technique

- ▶ Select Compare effectiveness of campaigns (Control Package Test) and click Continue.

Figure 13-1
Control Package Test, Fields tab



Control Package Test

This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. Campaign effectiveness is measured by responses. The Campaign Field identifies different campaigns, for example Offer A and Offer B. The Response Field indicates if a contact responded to the campaign. Select Purchase Amount when the response is recorded as a purchase amount, for example "99.99". Select Reply when the response is recorded as reply, for example "Yes". Select or enter the positive reply value.

Fields:
Sort: None

- Customer ID
- Responded to previous offer
- Postal Code
- Age
- Income category (thousands)
- Education
- Years at current residence
- Gender
- Married
- Region

Campaign Field: Control Package

Effectiveness Response Field: Responded to test offer

Effectiveness is defined as a

○ Purchase amount
● Reply

Positive reply value: 1

Name and Label for Recoded Effectiveness Response Field

This technique automatically recodes the response field into a new field in which 1 represents positive responses and 0 represents negative responses.

New field name: Effectiveness_new1

New field label: Effectiveness (1=Yes, 0-No)

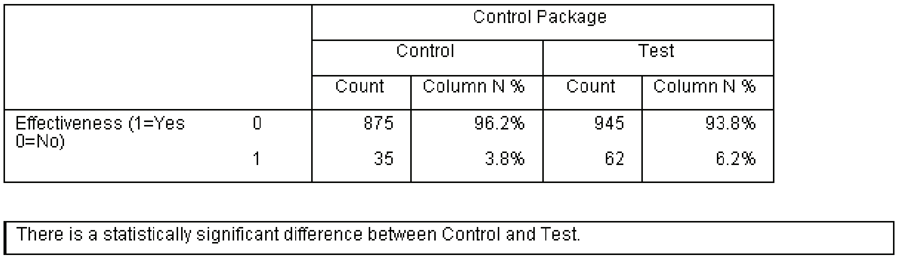
- ▶ For Campaign Field, select *Control Package*.
- ▶ For Effectiveness Response Field, select *Responded to test offer*.
- ▶ Select Reply.
- ▶ For Positive response value, select *Yes* from the drop-down list. A value of 1 is displayed in the text field because “Yes” is actually a value label associated with a recorded value of 1. (If the positive response value doesn’t have a defined value label, you can just enter the value in the text field.)

A new field is automatically created, in which 1 represents positive responses and 0 represents negative responses, and the analysis is performed on the new field. You can override the default name and label and provide your own. For this example, we’ll use the field name already provided.

- ▶ Click Run to run the procedure.

Output

Figure 13-2
Control Package Test output



		Control Package			
		Control		Test	
		Count	Column N %	Count	Column N %
Effectiveness (1=Yes 0=No)	0	875	96.2%	945	93.8%
	1	35	3.8%	62	6.2%

There is a statistically significant difference between Control and Test.

The output from the procedure includes a table that displays counts and percentages of positive and negative responses for each group defined by the Campaign Field and a table that indicates if the group response rates differ significantly from each other.

- *Effectiveness* is the recoded version of the response field, where 1 represents positive responses and 0 represents negative responses.
- The positive response rate for the control package is 3.8%, while the positive response rate for the test package is 6.2%.

The simple text description below the table indicates that the difference between the groups is significantly different, which means that the higher response rate for the test package probably isn't the result of random chance. This text table will contain a comparison for each possible pair of groups included in the analysis. Since there are only two groups in this examples, there is only one comparison. If there are more than five groups, the text description table is replaced with the Comparison of Column Proportions table.

Summary

The Control Package Test compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. In this example, the positive response of 6.2% for the test package was significantly higher than the positive response rate of 3.8% for the control package. This suggests that you should use the new package design instead of the old one, but there may be other factors that you need to consider, such as any additional costs associated with the new package design.

Appendix

A

Sample Files

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Descriptions

Following are brief descriptions of the sample files used in various examples throughout the documentation.

- **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.
- **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.
- **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs..
- **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the alfatoxin levels in parts per billion (PPB).
- **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.
- **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given

loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.

- **bankloan_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.
- **behavior.sav.** In a classic example , 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.
- **behavior_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.
- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.
- **breakfast.sav.** In a classic study , 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1="most preferred" to 15="least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only."
- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, "Overall preference," only.
- **broadband_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.
- **broadband_2.sav.** This data file is identical to *broadband_1.sav* but contains data for three additional months.
- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.
- **car_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.
- **car_sales_uprepared.sav.** This is a modified version of *car_sales.sav* that does not include any transformed versions of the fields.
- **carpet.sav.** In a popular example , a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.

- **carpet_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet_plan.sav*.
- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.
- **catalog_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.
- **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.
- **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.
- **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.
- **clothing_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.
- **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.
- **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.
- **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.
- **customer_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.
- **customer_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.
- **customer_subset.sav.** A subset of 80 cases from *customer_dbase.sav*.

- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.
- **debate_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.
- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.
- **demo_cs_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.
- **demo_cs_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.
- **demo_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.
- **dmdata.sav.** This is a hypothetical data file that contains demographic and purchasing information for a direct marketing company. *dmdata2.sav* contains information for a subset of contacts that received a test mailing, and *dmdata3.sav* contains information on the remaining contacts who did not receive the test mailing.
- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" . Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.
- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.
- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases at the University of California, Irvine.
- **grocery_1month.sav.** This hypothetical data file is the *grocery_coupons.sav* data file with the weekly purchases "rolled-up" so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.
- **grocery_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.

- **guttman.sav.** Bell presented a table to illustrate possible social groups. Guttman used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).
- **health_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.
- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.
- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.
- **insurance_claims.sav.** This is a hypothetical data file that concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Each case represents a separate claim.
- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.
- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.
- **kinship_dat.sav.** Rosenberg and Kim set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained. Each source corresponds to a 15 × 15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.
- **kinship_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship_dat.sav*.
- **kinship_var.sav.** This data file contains independent variables *gender*, *gener*(ation), and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.
- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.

- **nhis2000_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accessed 2003.
- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers , , among others found nonlinearities among these variables, which hinder standard regression approaches.
- **pain_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.
- **patient_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **patlos_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **poll_cs.sav.** This is a hypothetical data file that concerns pollsters' efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.
- **poll_cs_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added the data file after the sample as taken.
- **property_assess.sav.** This is a hypothetical data file that concerns a county assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.
- **property_assess_cs.sav.** This is a hypothetical data file that concerns a state assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.
- **property_assess_cs_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property_assess_cs.sav*. The sample was taken according to the design specified in the *property_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.

- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.
- **recidivism_cs_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the sampling plan specified in *recidivism_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.
- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.
- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.
- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks .
- **shampoo_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.
- **ships.sav.** A dataset presented and analyzed elsewhere that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.
- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.
- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. (*http://dx.doi.org/10.3886/ICPSR02934*) Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.
- **stroke_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in the Data Preparation option.

- **stroke_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.
- **stroke_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.
- **stroke_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.
- **survey_sample.sav.** This data file contains survey data, including demographic data and various attitude measures. It is based on a subset of variables from the 1998 NORC General Social Survey, although some data values have been modified and additional fictitious variables have been added for demonstration purposes.
- **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.
- **telco_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.
- **telco_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.
- **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.
- **testmarket_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.
- **tree_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.
- **tree_missing_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.
- **tree_score_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.

- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.
- **ulcer_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere .
- **ulcer_recurrence_recoded.sav.** This file reorganizes the information in *ulcer_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere .
- **verd1985.sav.** This data file concerns a survey . The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.
- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.
- **wheeze_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children . The data contain repeated binary measures of the wheezing status for children from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.
- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.

Appendix

B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, 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. SPSS Inc. 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-SPSS and 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 SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS 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.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

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

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

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

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Index