

IBM SPSS Direct Marketing 25

IBM

Note

Before using this information and the product it supports, read the information in "Notices" on page 21.

Product Information

This edition applies to version 25, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Direct marketing	1	Postal Code Response Rates	12
Direct Marketing	1	Settings.	13
RFM Analysis	1	Creating a Categorical Response Field	14
RFM Scores from Transaction Data	2	Propensity to purchase	15
RFM Scores from Customer Data	2	Settings.	17
RFM Binning	3	Creating a categorical response field	18
Saving RFM Scores from Transaction Data	5	Control Package Test	18
Saving RFM Scores from Customer Data	5		
RFM Output	6	Notices	21
Cluster analysis	7	Trademarks	23
Settings	8		
Prospect profiles	9	Index	25
Settings.	11		
Creating a categorical response field	11		

Direct marketing

The following direct marketing features are included in SPSS® Statistics Premium Edition or the Direct Marketing option.

Direct Marketing

Direct Marketing 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.

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.

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. See the topic “Prospect profiles” on page 9 for more information.

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. See the topic “Postal Code Response Rates” on page 12 for more information.

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. See the topic “Propensity to purchase” on page 15 for more information.

Control Package Test. This technique compares marketing campaigns to see if there is a significant difference in effectiveness for different packages or offers. See the topic “Control Package Test” on page 18 for more information.

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. See the topic "RFM Scores from Transaction Data" for more information.
- 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. See the topic "RFM Scores from Customer Data" for more information.

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.

Creating RFM Scores from Transaction Data

1. From the menus choose:
Analyze > Direct Marketing > Choose Technique
2. Select **Help identify my best contacts (RFM Analysis)** and click **Continue**.
3. Select **Transaction data** and click **Continue**.
4. Select the variable that contains transaction dates.
5. Select the variable that contains the monetary amount for each transaction.
6. Select the method for summarizing transaction amounts for each customer: Total (sum of all transactions), mean, median, or maximum (highest transaction amount).
7. 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.

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

1. From the menus choose:
Analyze > Direct Marketing > Choose Technique
2. Select **Help identify my best contacts (RFM Analysis)** and click **Continue**.
3. Select **Customer data** and click **Continue**.
4. Select the variable that contains the most recent transaction date or a number that represents a time interval since the most recent transaction.
5. Select the variable that contains the total number of transactions for each customer.
6. Select the variable that contains the summary monetary amount for each customer.
7. 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.

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 ideal value of 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 1. Assign Ties to Same Bin vs. Randomly Assign Ties.

ID	Most Recent Purchase (Recency)	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.

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: $(recency \times 100) + (frequency \times 10) + monetary$.

By default all available variables are included in the new dataset; so deselect 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.

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

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 be 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.

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) represents 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, postal 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.

Table 2. Measurement level icons

	Numeric	String	Date	Time
Scale (Continuous)		n/a		
Ordinal				
Nominal				

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.

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:

Analyze > Direct Marketing > Choose Technique

1. Select **Segment my contacts into clusters**.
2. Select the categorical (nominal, ordinal) and continuous (scale) fields that you want to use to create segments.
3. Click **Run** to run the procedure.

Settings

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.

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.

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. See the topic "Creating a categorical response field" on page 11 for more information.

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, postal 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.

Table 3. Measurement level icons

	Numeric	String	Date	Time
Scale (Continuous)		n/a		
Ordinal				
Nominal				

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.

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:

Analyze > Direct Marketing > Choose Technique

1. Select **Generate profiles of my contacts who responded to an offer.**
2. Select the field that identifies which contacts responded to the offer. This field must be nominal or ordinal.
3. 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.
4. Select the fields you want to use to create the profiles.
5. Click **Run** to run the procedure.

Related information:

Detailed examples

Settings

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:

$$\text{NewName}=\text{OldName}>0$$

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:

$$\text{NewName}=\text{NOT}(\text{MISSING}(\text{OldName}))$$

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

1. From the menus choose:
Transform > Compute Variable
2. For Target Variable, enter the new field (variable) name.
3. If negative responses are recorded as 0, for the Numeric Expression enter $\text{OldName}>0$, where *OldName* is the original field name.
4. If negative responses are recorded as missing (blank), for the Numeric Expression enter $\text{NOT}(\text{MISSING}(\text{OldName}))$, where *OldName* is the original field name.

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.

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 $N \times P \times (1-P)$, 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. See the topic "Creating a Categorical Response Field" on page 14 for more information.

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:

Analyze > Direct Marketing > Choose Technique

1. Select **Identify top responding postal codes**.
2. Select the field that identifies which contacts responded to the offer.

3. 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.
4. Select the field that contains the postal code.
5. 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

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 percentage (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: $(\text{Cost of mailing a package} / \text{Net revenue per response}) \times 100$. 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: $(0.75/56) \times 100 = 1.34\%$.

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

1. From the menus choose:
Transform > Compute Variable
2. For Target Variable, enter the new field (variable) name.
3. If negative responses are recorded as 0, for the Numeric Expression enter $OldName > 0$, where *OldName* is the original field name.
4. If negative responses are recorded as missing (blank), for the Numeric Expression enter $NOT(MISSING(OldName))$, where *OldName* is the original field name.

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:

1. 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.
2. 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.

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. See the topic “Creating a categorical response field” on page 18 for more information.

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, postal 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.

Table 4. Measurement level icons

	Numeric	String	Date	Time
Scale (Continuous)		n/a		
Ordinal				
Nominal				

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.

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:

Analyze > Direct Marketing > Choose Technique

1. Select **Select contacts most likely to purchase.**
2. Select the field that identifies which contacts responded to the offer.
3. 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.
4. Select the fields you want to use to predict propensity.
To save a model XML file to score other data files:
5. Select (check) **Export model information to XML file.**
6. Enter a directory path and file name or click **Browse** to navigate to the location where you want to save the model XML file.

7. Click **Run** to run the procedure.
To use the model file to score other datasets:
8. Open the dataset that you want to score.
9. Use the Scoring Wizard to apply the model to the dataset. From the menus choose:
Utilities > Scoring Wizard.

Settings

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 IBM 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:

$$\text{NewName} = \text{OldName} > 0$$

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:

$$\text{NewName} = \text{NOT}(\text{MISSING}(\text{OldName}))$$

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

1. From the menus choose:
Transform > Compute Variable
2. For Target Variable, enter the new field (variable) name.
3. If negative responses are recorded as 0, for the Numeric Expression enter $\text{OldName} > 0$, where *OldName* is the original field name.
4. If negative responses are recorded as missing (blank), for the Numeric Expression enter $\text{NOT}(\text{MISSING}(\text{OldName}))$, where *OldName* is the original field name.

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.

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:

Analyze > Direct Marketing > Choose Technique

1. Select **Compare effectiveness of campaigns**.
2. 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.
3. 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.

4. Click **Run** to run the procedure.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US*

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

*Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

*IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US*

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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 IBM, 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. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. _enter the year or years_. All rights reserved.

Trademarks

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

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 or registered trademarks of Oracle and/or its affiliates.

Index

C

cluster 7
cluster analysis 7
cluster analysis (Direct Marketing) 7
control package test 18

L

logistic regression (Direct Marketing) 15

P

postal code response rates 12
propensity to purchase 15
prospect profiles(Direct Marketing) 9

R

RFM 1, 5, 6
 binning 3
 customer data 2
 transaction data 2

Z

zip code response rates 12



Printed in USA