dun & bradstreet

Market Insight

Modelling Module

Training Manual v3.2

D&B Market Insight

Modelling Module

Manual Version:	3.2
Software Version:	2016 Q4
System:	Training (US)

D&B Market Insight is powered by
FastStats Technology from Apteco Ltd

Contents

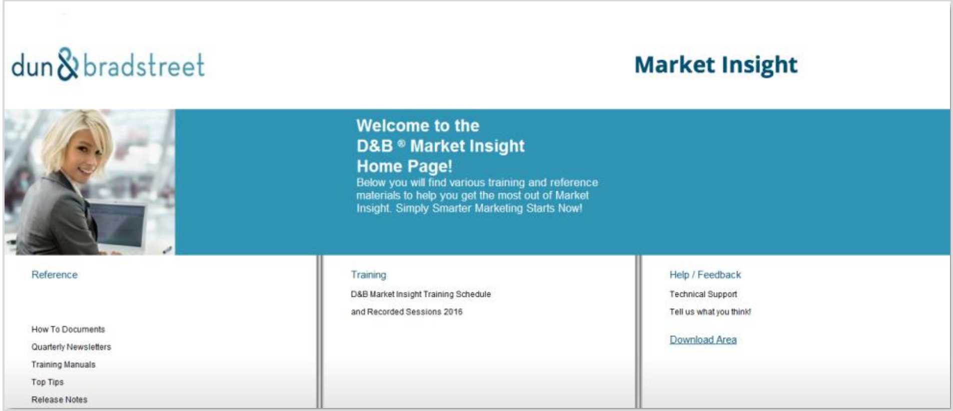
Introduction

Market Insight provides powerful and interactive marketing analysis of customer data overlaid on a D&B data universe. The system is web based with a truly easy to use Windows interface. Using a consistent and intuitive “drag and drop” approach throughout, every action automatically results in a query that can be saved and reused with ease. With a wide range of descriptive and predictive analytical tools, Market Insight’s analysis options are virtually unlimited as any technique can be applied to any results in any order. Market Insight provides a unique combination of speed, power and accessibility for data exploration and understanding.

Market Insight holds your data overlaid on a D&B universe. This enables you to accurately measure your customer data in proportion to the opportunities in the market place. Hence the product’s name: it enables insight of your activities in comparison to the market place rather than just within your business.

The D&B data universe in your Market Insight system will be adjusted to suit your licensing and measurement requirements. Your customer data is loaded from extract file(s) you provide and although this process allows for some cleaning and manipulation of the data, what you see within Market Insight is a reflection of the data you provide.

The Market Insight view of the data is a snapshot at the time that the data was loaded. Market Insight is an analytical system able to provide insight and understanding but it can also provide data feeds to your operational marketing systems to implement your targeting decisions.



dun&bradstreet

Market Insight

Welcome to the
D&B ® Market Insight
Home Page!
Below you will find various training and reference materials to help you get the most out of Market Insight. Simply Smarter Marketing Starts Now!

Reference

How To Documents
Quarterly Newsletters
Training Manuals
Top Tips
Release Notes

Training

D&B Market Insight Training Schedule
and Recorded Sessions 2016

Help / Feedback

Technical Support
Tell us what you think!

Download Area

Market Insight Homepage

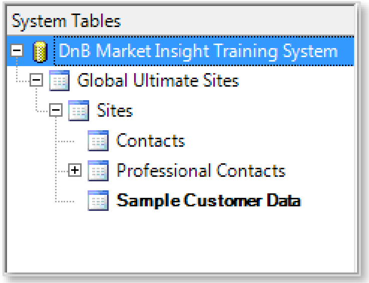
N.B. The counts and figures in this manual may differ to those seen when you use this system as the data changes over time. Not all the functionality shown in this manual may be available in the system you are using.

Data Structure

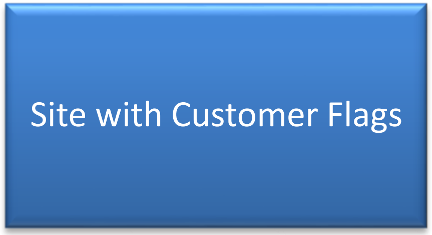
The structure of your Market Insight system can vary. The elements shown here are typical – each Site may be simply flagged with Customer data or can have many related Contacts. A Site may also have many matched Customer Accounts, and / or many other types of related data (such as one or more URLs, UCC Filings and Corporate Family Linkage Plus entries, etc.) The data loaded for each matched Customer Account is configurable – for example you may have multiple Transactions or Divisional Summaries or Product Summaries etc.

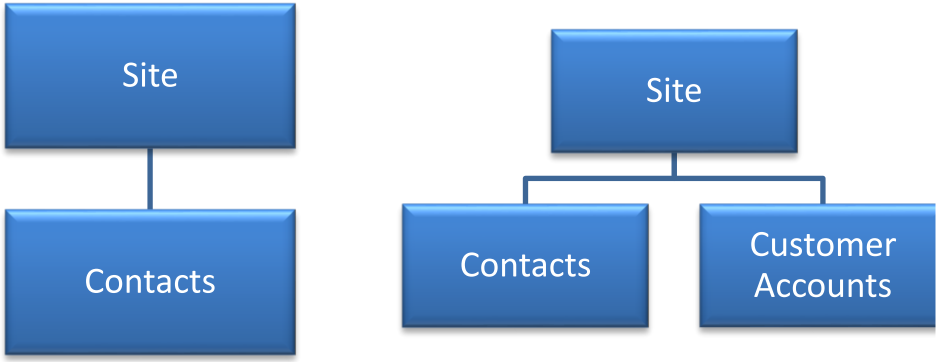
The detail present on each table of data depends on the Market Insight administrator. The data is arranged into folders to assist the user to navigate and find data items.

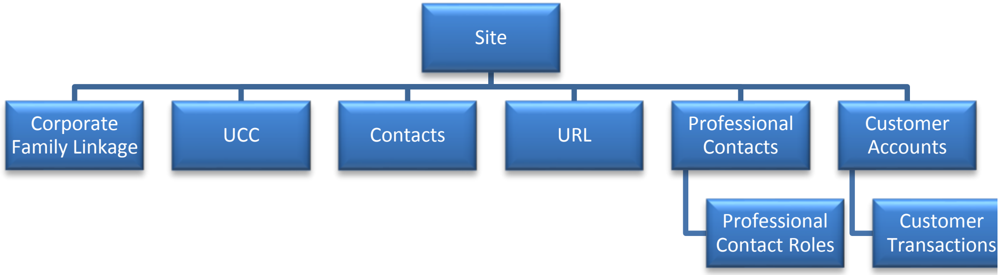
The structure used in the Training System, illustrated in this manual, uses a simple structure that has sample customer data and contact tables that are linked to a Sites table and these sites are in turn linked to the Global Ultimate Sites table (e.g. worldwide headquarters).



System Tables
- DnB Market Insight Training System
  - Global Ultimate Sites
    - Sites
      - Contacts
      - Professional Contacts
      - **Sample Customer Data**







Accessing Market Insight

The Market Insight software is downloaded automatically to your PC when you click a link to launch the system. Once the software has been downloaded, it will automatically update from the server whenever necessary.

To access Market Insight you need:

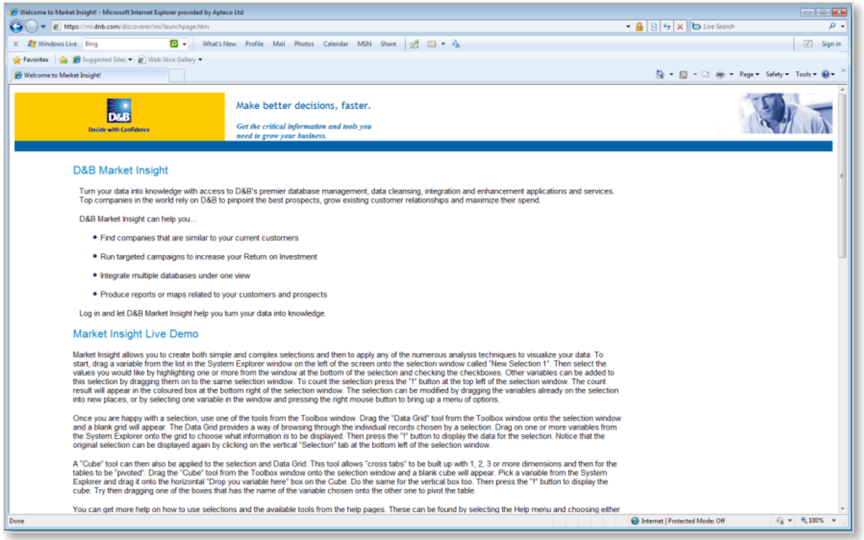
- Windows PC – Market Insight is a Windows.NET application that combines the best of the Windows interface with web based systems. Market Insight is not available on Mac or UNIX computers
- The latest Windows.NET framework version installed. This can be obtained by visiting www.windowsupdate.com or from your IT team

To launch your Market Insight system, use a browser to view:

https://mi.dnb.com/discoverer/mi/launchpage.htm

N.B. The “https” prefix, which establishes a secure connection between your browser and the D&B Server.

➢ Click on the **“Click here to run Market Insight”** link to launch the software installer


Welcome to Market Insight! - Microsoft Internet Explorer provided by Apteco Ltd
https://mi.dnb.com/discoverer/mi/launchpage.htm
Live Search
Windows Live
Bing
What's New
Profile
Mail
Photos
Calendar
MSN
Share
Sign in
Favorites
Suggested Sites
Web Slice Gallery
Welcome to Market Insight!
Page
Safety
Tools
D&B
Decide with Confidence
Make better decisions, faster.
Get the critical information and tools you need to grow your business.
Done
Internet | Protected Mode: Off
100%

D&B Market Insight

Turn your data into knowledge with access to D&B's premier database management, data cleansing, integration and enhancement applications and services. Top companies in the world rely on D&B to pinpoint the best prospects, grow existing customer relationships and maximize their spend.

D&B Market Insight can help you...

- Find companies that are similar to your current customers
- Run targeted campaigns to increase your Return on Investment
- Integrate multiple databases under one view
- Produce reports or maps related to your customers and prospects

Log in and let D&B Market Insight help you turn your data into knowledge.

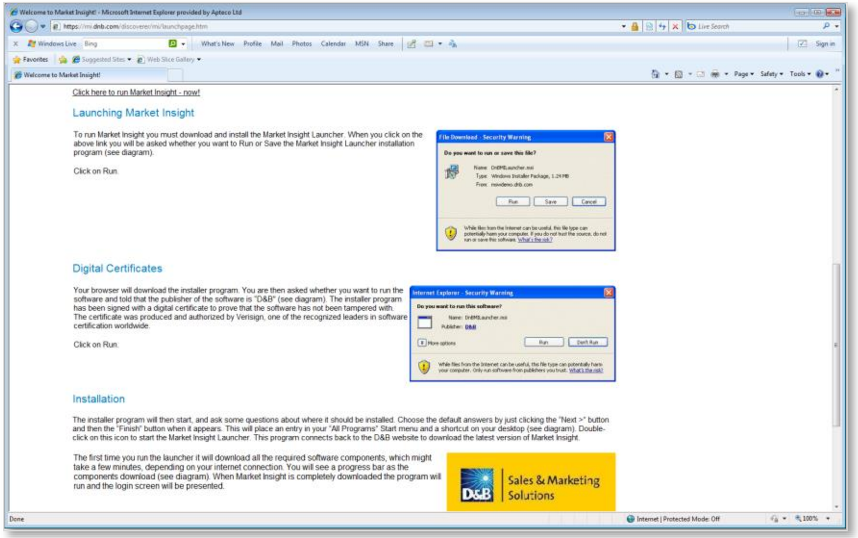
Market Insight Live Demo

Market Insight allows you to create both simple and complex selections and then to apply any of the numerous analysis techniques to visualize your data. To start, drag a variable from the list in the System Explorer window on the left of the screen onto the selection window called "New Selection 1". Then select the values you would like by highlighting one or more from the window at the bottom of the selection and checking the checkboxes. Other variables can be added to this selection by dragging them on to the same selection window. To count the selection press the "!" button at the top left of the selection window. The count result will appear in the coloured box at the bottom right of the selection window. The selection can be modified by dragging the variables already on the selection into new places, or by selecting one variable in the window and pressing the right mouse button to bring up a menu of options.

Once you are happy with a selection, use one of the tools from the Toolbox window. Drag the "Data Grid" tool from the Toolbox window onto the selection window and a blank grid will appear. The Data Grid provides a way of browsing through the individual records chosen by a selection. Drag on one or more variables from the System Explorer onto the grid to choose what information is to be displayed. Then press the "!" button to display the data for the selection. Notice that the original selection can be displayed again by clicking on the vertical "Selection" tab at the bottom left of the selection window.

A "Cube" tool can then also be applied to the selection and Data Grid. This tool allows "cross tabs" to be built up with 1, 2, 3 or more dimensions and then for the tables to be "pivoted". Drag the "Cube" tool from the Toolbox window onto the selection window and a blank cube will appear. Pick a variable from the System Explorer and drag it onto the horizontal "Drop you variable here" box on the Cube. Do the same for the vertical box too. Then press the "!" button to display the cube. Try then dragging one of the boxes that has the name of the variable chosen onto the other one to pivot the table.

You can get more help on how to use selections and the available tools from the help pages. These can be found by selecting the Help menu and choosing either


Welcome to Market Insight! - Microsoft Internet Explorer provided by Apteco Ltd
https://mi.dnb.com/discoverer/mi/launchpage.htm
Live Search
Windows Live
Bing
What's New
Profile
Mail
Photos
Calendar
MSN
Share
Sign in
Favorites
Suggested Sites
Web Slice Gallery
Welcome to Market Insight!
Page
Safety
Tools
File Download - Security Warning
Run
Save
Cancel
Internet Explorer - Security Warning
More options
Run
Don't Run
D&B
Sales & Marketing Solutions
Done
Internet | Protected Mode: Off
100%

Click here to run Market Insight - now!

Launching Market Insight

To run Market Insight you must download and install the Market Insight Launcher. When you click on the above link you will be asked whether you want to Run or Save the Market Insight Launcher installation program (see diagram).

Click on Run.

Do you want to run or save this file?

Name: DnBMILauncher.msi

Type: Windows Installer Package, 1.24 MB

From: midemo.dnb.com

While files from the Internet can be useful, this file type can potentially harm your computer. If you do not trust the source, do not run or save this software. What's the risk?

Digital Certificates

Your browser will download the installer program. You are then asked whether you want to run the software and told that the publisher of the software is "D&B" (see diagram). The installer program has been signed with a digital certificate to prove that the software has not been tampered with. The certificate was produced and authorized by Verisign, one of the recognized leaders in software certification worldwide.

Click on Run.

Do you want to run this software?

Name: DnBMILauncher.msi

Publisher: D&B

While files from the Internet can be useful, this file type can potentially harm your computer. Only run software from publishers you trust. What's the risk?

Installation

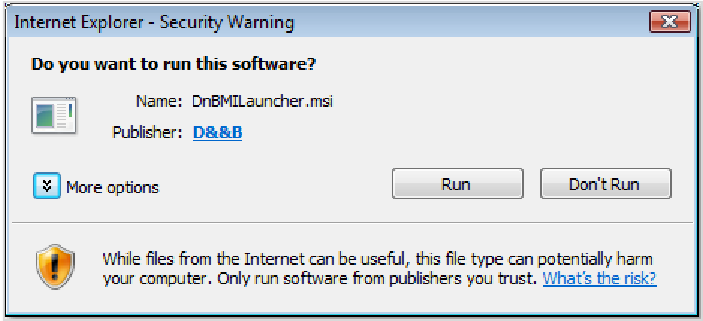
The installer program will then start, and ask some questions about where it should be installed. Choose the default answers by just clicking the "Next >" button and then the "Finish" button when it appears. This will place an entry in your "All Programs" Start menu and a shortcut on your desktop (see diagram). Double-click on this icon to start the Market Insight Launcher. This program connects back to the D&B website to download the latest version of Market Insight.

The first time you run the launcher it will download all the required software components, which might take a few minutes, depending on your internet connection. You will see a progress bar as the components download (see diagram). When Market Insight is completely downloaded the program will run and the login screen will be presented.

Instructions to Download Launcher

You can elect to save the file or run it directly from the web site.

- If you elected to save the file, navigate to where you saved and double click it. Agree to run when prompted, and then follow the on screen instructions
- The installation process will result in an icon on your desktop and in a D&B Start Menu folder
- On subsequent uses of Market Insight, you can simply double click this icon. The software will automatically update from the D&B server whenever new releases are made available
- You can install Market Insight on as many computers as you wish – it is your user id that controls your access. This means, for example, you can use Market Insight when working from home



Internet Explorer - Security Warning

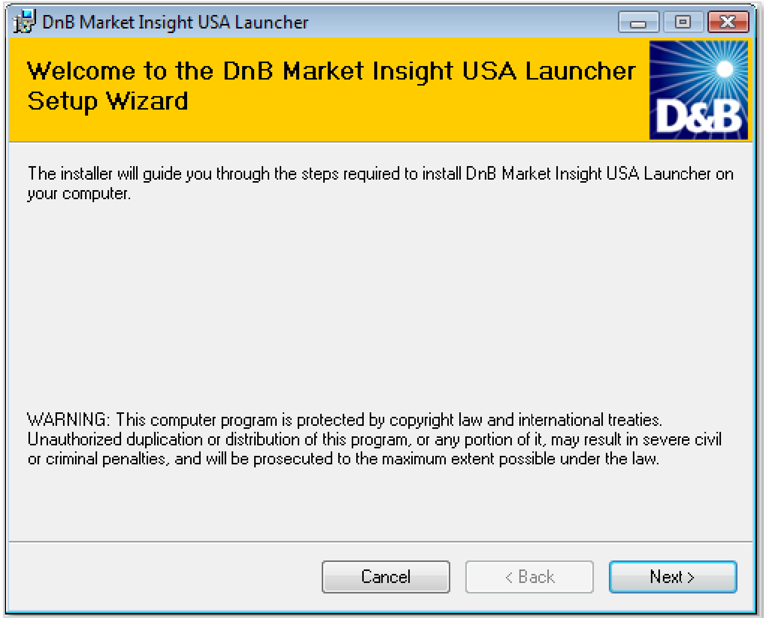
Do you want to run this software?

Name: DnBMILauncher.msi

Publisher: D&B

While files from the Internet can be useful, this file type can potentially harm your computer. Only run software from publishers you trust. What's the risk?

Security Warning Prompt



DnB Market Insight USA Launcher

Welcome to the DnB Market Insight USA Launcher Setup Wizard

The installer will guide you through the steps required to install DnB Market Insight USA Launcher on your computer.

WARNING: This computer program is protected by copyright law and international treaties. Unauthorized duplication or distribution of this program, or any portion of it, may result in severe civil or criminal penalties, and will be prosecuted to the maximum extent possible under the law.

Launcher Setup Wizard

How to Login

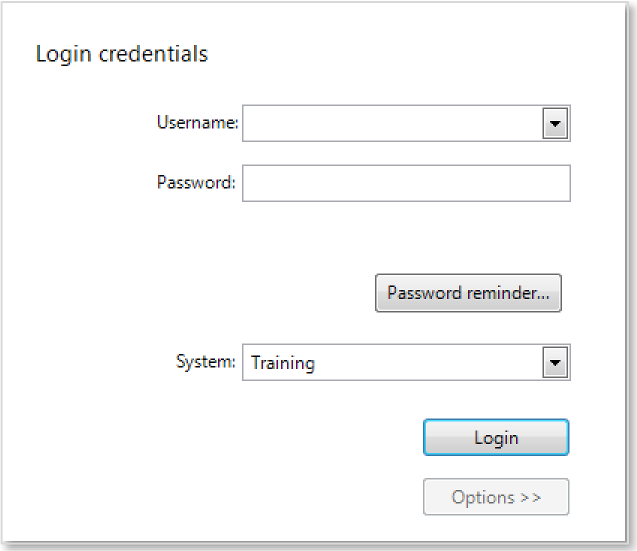
To use Market Insight, you need to have an Internet connection.

Start Market Insight by:

- Clicking on the **Market Insight** icon on your desktop, or
- Navigating via:

  Start → All Programs → DnB → Market Insight

In the upper left hand corner of the screen you will see a Login window that gives you the opportunity to connect to a Market Insight system containing data available to you for analysis.



Login credentials

Username:

Password:

System: Training

Login Window

Enterprise Tab

Your Market Insight system operates on a secure and resilient web connected server enabling you to access the system from any location with an Internet connection. A number of users may access the system at the same time, each of whom is authorized by a user account and password. Your Market Insight Administrator will provide you with a Username and Password.

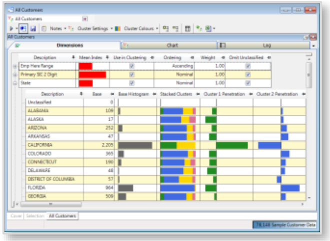
Introduction to Modelling

The Market Insight Modelling Module gives you the tools to *describe* your current data and *predict* possible activity in the future. The Marketer will often start by wanting an overview of *existing* customers and their *existing* behavior. Profile and Decision Tree can be very useful in *describing* a particular set of customers although their main purpose is *predictive*. The purpose of Cluster is purely *descriptive* and it identifies sub-groups who share common characteristics. There are all sorts of different applications of clustering including both tactical and strategic (e.g. classifying all customers as High Value, Average Value, Dormant, etc.).

At some point, you will want to acquire *new* customers or to *change* behavior of existing customers. In this case you will need a *predictive* tool, such as Profile or Decision Tree. You may want to predict which customers within an Average Value cluster are most likely to move up to the High Value and so are worth marketing to. You may simply want to identify likely responders when planning a campaign. Having previously created a Cluster model may be helpful in building a predictive model such as this, but it is not a pre-requisite.

In order for a model to be useful, the patterns and characteristics encapsulated in the model must apply more generally to data other than that used to build the model, for example, to companies who are not yet customers. A Cluster model encapsulates natural patterns *within one* set of data, whereas a predictive model is based on the distinguishing characteristics *between two* sets of data. The Model Report can be used to test how well predictive models apply to new data, for example, predicting behavior of non-customers.

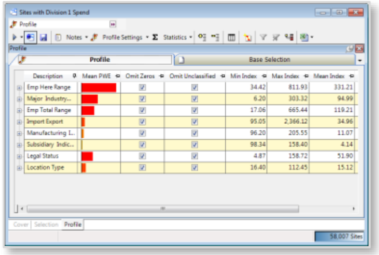
All models encapsulate the essential features of the real world at the time they are built. As time passes models should be reviewed to ensure that they are still representative of current customer behavior.



Description	Mean Index	Use in Clustering	Ordering	Weight	Omit Unclassified
Emp Here Range	🟥	☑	Ascending	1.00	☑
Primary SIC 2 Digit	🟥	☑	Nominal	1.00	☑
State	🟥	☑	Nominal	1.00	☑

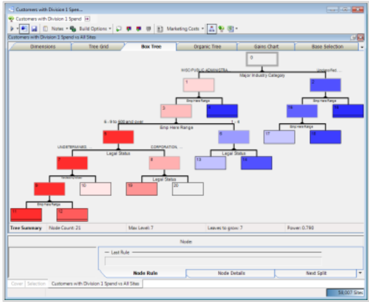
Description	Base	Base Histogram	Stacked Clusters	Cluster 1 Penetration	Cluster 2 Penetration
Unclassified	0				
ALABAMA	109				
ALASKA	17				
ARIZONA	252				
ARKANSAS	47				
CALIFORNIA	2,205				
COLORADO	365				
CONNECTICUT	190				
DELAWARE	48				
DISTRICT OF COLUMBIA	57				
FLORIDA	964				
GEORGIA	509				

Cluster



Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index
Emp Here Range	🟥	☑	☑	34.42	811.93	331.21
Major Industry...	🟥	☑	☑	6.20	303.32	94.99
Emp Total Range	🟥	☑	☑	17.06	665.44	119.21
Import Export	🟥	☑	☑	95.05	2,366.12	34.96
Manufacturing L...	🟥	☑	☑	96.20	205.55	11.07
Subsidiary Indic...	🟥	☑	☑	98.34	158.40	4.14
Legal Status	🟥	☑	☑	4.87	158.72	51.90
Location Type	🟥	☑	☑	16.40	112.45	15.12

Profile



Decision Tree

Introduction to the Modelling Environment

The Modelling Environment provides an easy way to create and evaluate alternative models. It allows modelling users to define multiple candidate selections and dimensions and then apply a subset to the Cluster, Profile or Decision Tree, creating alternative models which can then be compared. This environment can then be saved and reloaded to keep in one place all the modelling decisions and experiments made by the user.

At the bottom of the Modelling Environment window there are three tabs:

Selections

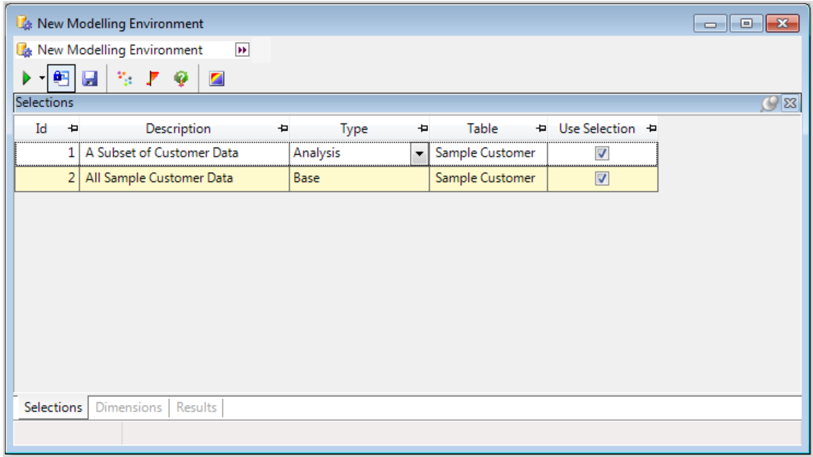
Selections of records can be dragged onto this area for possible use in each of the Modelling tools. By default, when you drag on the first selection it will be set as the Analysis group and another selection will be auto generated and set as the Base group. Further groups can be added and then defined in the Type column. By ticking or unticking a group in the Use Selection column you can define what groups to use. You will need to define an Analysis and Base group for Profile and Decision Tree and just an Analysis group for the Cluster.

Dimensions

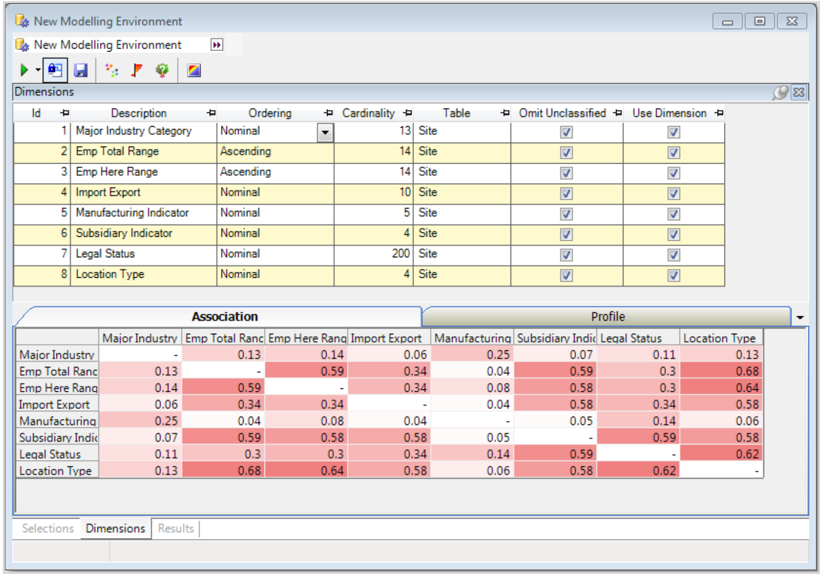
Drag onto this area the variables you would like the tools to use when calculating and generating their models. The Association tab shows the relationship between the variables; the closer the figure is to 1 the more closely related the variables are. This is discussed further in a later section.

Results

This area lists each model generated through the tools, allowing you to evaluate and compare models and creating a history of your activity.



Id	Description	Type	Table	Use Selection
1	A Subset of Customer Data	Analysis	Sample Customer	☑
2	All Sample Customer Data	Base	Sample Customer	☑

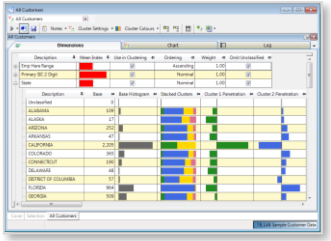


Id	Description	Ordering	Cardinality	Table	Omit Unclassified	Use Dimension
1	Major Industry Category	Nominal	13	Site	☑	☑
2	Emp Total Range	Ascending	14	Site	☑	☑
3	Emp Here Range	Ascending	14	Site	☑	☑
4	Import Export	Nominal	10	Site	☑	☑
5	Manufacturing Indicator	Nominal	5	Site	☑	☑
6	Subsidiary Indicator	Nominal	4	Site	☑	☑
7	Legal Status	Nominal	200	Site	☑	☑
8	Location Type	Nominal	4	Site	☑	☑

	Major Industry	Emp Total Ranc	Emp Here Rang	Import Export	Manufacturing	Subsidiary Indic	Legal Status	Location Type
Major Industry	-	0.13	0.14	0.06	0.25	0.07	0.11	0.13
Emp Total Ranc	0.13	-	0.59	0.34	0.04	0.59	0.3	0.68
Emp Here Rang	0.14	0.59	-	0.34	0.08	0.58	0.3	0.64
Import Export	0.06	0.34	0.34	-	0.04	0.58	0.34	0.58
Manufacturing	0.25	0.04	0.08	0.04	-	0.05	0.14	0.06
Subsidiary Indic	0.07	0.59	0.58	0.58	0.05	-	0.59	0.58
Legal Status	0.11	0.3	0.3	0.34	0.14	0.59	-	0.62
Location Type	0.13	0.68	0.64	0.58	0.06	0.58	0.62	-

Introduction to Cluster

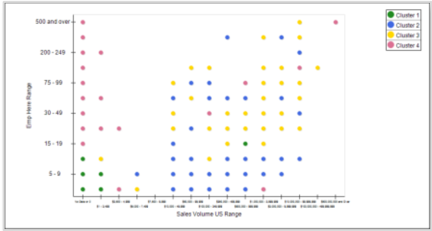
Cluster analysis is about exploring and identifying natural groupings in a set of data points. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". In Market Insight we are interested in taking a selection of records (e.g. Sites) and automatically detecting groups with similar characteristics. We can use these groups or clusters to better visualize our customer population and segment them for marketing purposes.



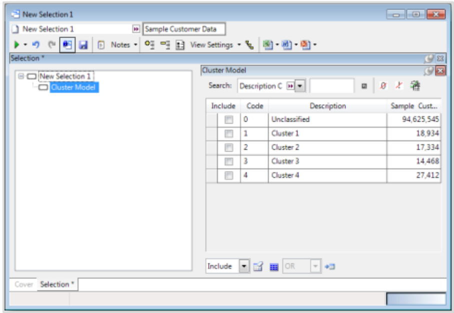
Description	Mean Index	Use in Clustering	Ordering	Weight	Omit Unclassified
Emp Here Range	🟥	☑	Ascending	1.00	☑
Primary SIC 2 Digit	🟥	☑	Nominal	1.00	☑
State	🟥	☑	Nominal	1.00	☑

Description	Base	Base Histogram	Stacked Clusters	Cluster 1 Penetration	Cluster 2 Penetration
Unclassified	0				
ALABAMA	109				
ALASKA	17				
ARIZONA	252				
ARKANSAS	47				
CALIFORNIA	2,205				
COLORADO	365				
CONNECTICUT	190				
DELAWARE	48				
DISTRICT OF COLUMBIA	57				
FLORIDA	964				
GEORGIA	509				

Market Insight allows us to create a Cluster Model from the initial analysis so that the cluster records can be represented as categories in a selection variable. The data represented in this way allows the other Market Insight tools to be used in the analysis; Cubes to further breakdown the results by other dimensions (variables), Data Grids to identify individual records or create mailing files etc.



This module can be used on its own or as a compliment to the other Modelling tools, Profiler and Decision Tree.



Include	Code	Description	Sample Cust...
□	0	Unclassified	94,625,545
□	1	Cluster 1	18,934
□	2	Cluster 2	17,334
□	3	Cluster 3	14,468
□	4	Cluster 4	27,412

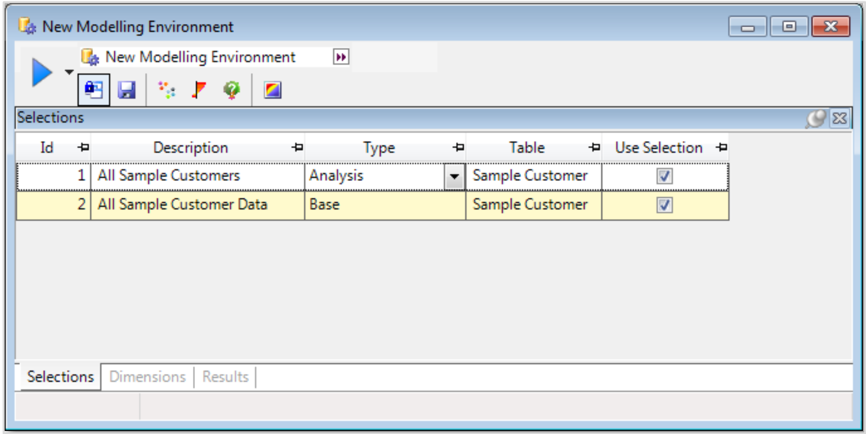
A Cluster Worked Example

Clustering looks for the similarities between data points. In this example those data points are Customers plotted in terms of the categories they fall into in multiple dimensions. Using the default settings these data points will be grouped into 4 clusters. The records to be used in the Cluster analysis are identified by creating a query through the Market Insight Selection page. In this example all the Customers on the database will be used.

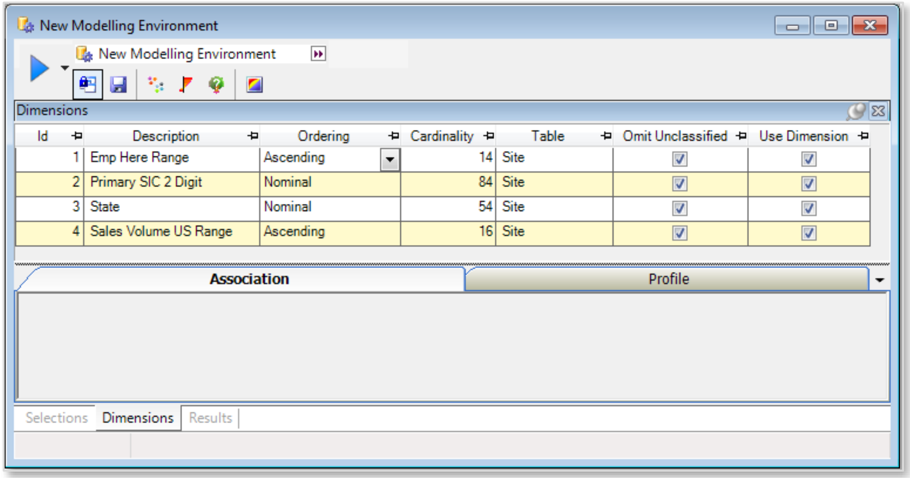
Setting up within the Modelling Environment

- Drag the **Modelling Environment** tool onto the workspace
- Drag the **Customer Flags** variable onto the workspace and select **Customer**
- Click the ▶ **Build** button to see the total number of **Customers** to be used in the **Cluster** analysis. Rename as **All Sample Customers**
- Drag this selection onto the **Selections** area of the **New Modelling Environment** window
- Click onto the **Dimensions** tab and then from the **System Explorer** drag and drop the variables **Emp Here Range, Primary SIC 2 Digit, State and Sales Volume US Range**
- Save the **New Modelling Environment** in your **Private** folder
- Click the **Cluster** button to take you into the Cluster window

N.B. The **Use in Clustering** box can be unchecked for variables you do not want to use in the calculation but you do want to help you describe your end results.



Id	Description	Type	Table	Use Selection
1	All Sample Customers	Analysis	Sample Customer	☑
2	All Sample Customer Data	Base	Sample Customer	☑



Id	Description	Ordering	Cardinality	Table	Omit Unclassified	Use Dimension
1	Emp Here Range	Ascending	14	Site	☑	☑
2	Primary SIC 2 Digit	Nominal	84	Site	☑	☑
3	State	Nominal	54	Site	☑	☑
4	Sales Volume US Range	Ascending	16	Site	☑	☑

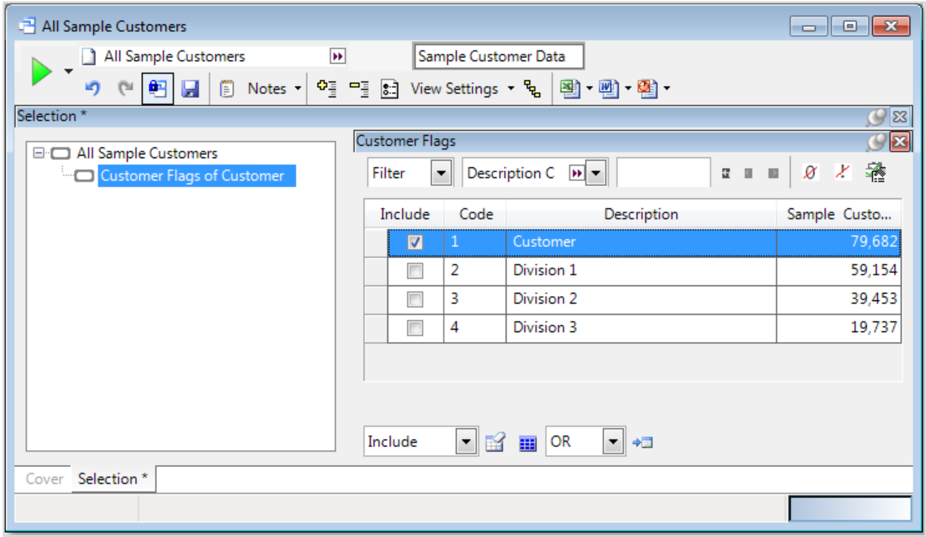
- Click the ▶ **Build** button to create the clusters

Setting up outside the Modelling Environment

- Drag the **Customer Flags** variable onto the workspace and select **Customer**

- Click the ▶ **Build** button to see the total number of **Customers** to be used in the **Cluster** analysis. Rename as **All Sample Customers**

- From the **Toolbox** drag and drop the **Cluster** tool onto the open selection page



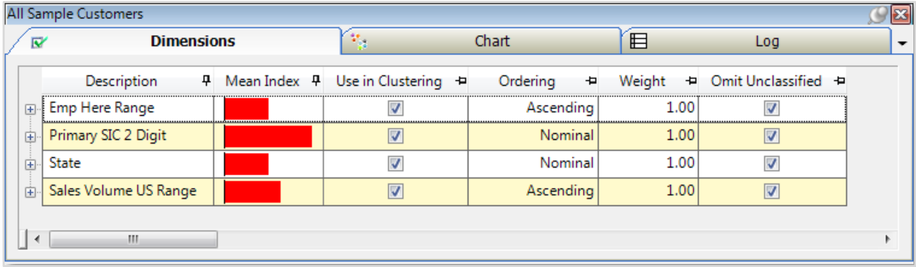
Include	Code	Description	Sample Custo...
☑	1	Customer	79,682
□	2	Division 1	59,154
□	3	Division 2	39,453
□	4	Division 3	19,737

Having determined the analysis group you now need to specify which variables will be used for the analysis.

- Ensure the **Dimensions** tab is upper most by clicking on it

- Click onto the **Dimensions** tab and then from the **System Explorer** drag and drop the variables **Emp Here Range, Primary SIC 2 Digit, State and Sales Volume US Range**

- Click the ▶ **Build** button to create the clusters

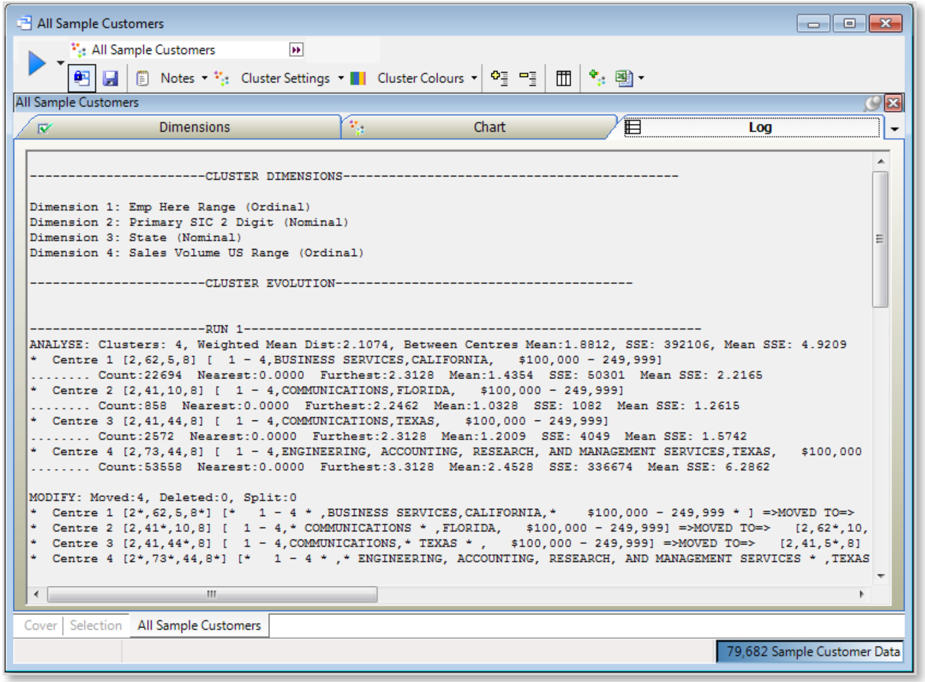


Description	Mean Index	Use in Clustering	Ordering	Weight	Omit Unclassified
Emp Here Range	🟥	☑	Ascending	1.00	☑
Primary SIC 2 Digit	🟥	☑	Nominal	1.00	☑
State	🟥	☑	Nominal	1.00	☑
Sales Volume US Range	🟥	☑	Ascending	1.00	☑

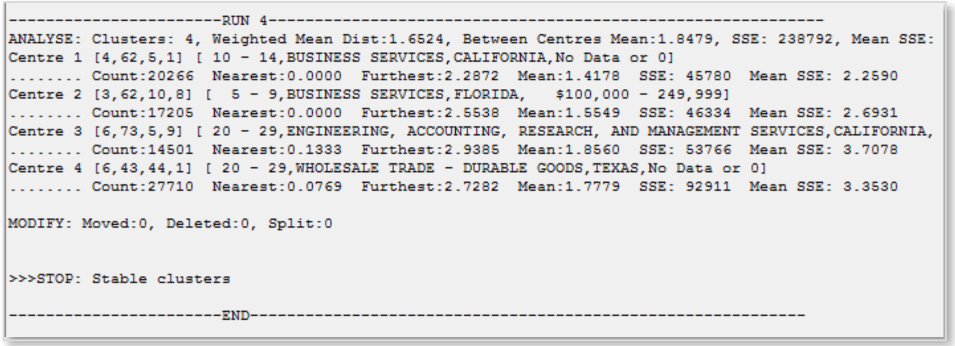
Once the Build button has been pressed Market Insight starts to use the K-Means technique to run what is called the Cluster Evolution (see **Appendix 1** for more details).

The following points are an overview of the process and can be seen on the Log window:

- Starting points are calculated by finding the highest populated points using the dimensions assigned
- Each point is then assigned to its nearest cluster starting point
- For all the points assigned to a particular cluster the midpoint of those points are calculated
- The calculated midpoint of each cluster becomes the new center of the cluster
- Market Insight will now perform another run through the data points and reallocate them to their nearest cluster. You can see these changes under the Modify section of each run in the Log window
- This process will repeat for the required number of runs. When the cluster centers no longer move the message STOP: Stable clusters will appear



```
---------------------CLUSTER DIMENSIONS------------------------------------------

Dimension 1: Emp Here Range (Ordinal)
Dimension 2: Primary SIC 2 Digit (Nominal)
Dimension 3: State (Nominal)
Dimension 4: Sales Volume US Range (Ordinal)

---------------------CLUSTER EVOLUTION------------------------------------


---------------------RUN 1-------------------------------------------------------------
ANALYSE: Clusters: 4, Weighted Mean Dist:2.1074, Between Centres Mean:1.8812, SSE: 392106, Mean SSE: 4.9209
*  Centre 1 [2,62,5,8] [  1 - 4,BUSINESS SERVICES,CALIFORNIA,   $100,000 - 249,999]
........ Count:22694  Nearest:0.0000  Furthest:2.3128  Mean:1.4354  SSE: 50301  Mean SSE: 2.2165
*  Centre 2 [2,41,10,8] [  1 - 4,COMMUNICATIONS,FLORIDA,   $100,000 - 249,999]
........ Count:858  Nearest:0.0000  Furthest:2.2462  Mean:1.0328  SSE: 1082  Mean SSE: 1.2615
*  Centre 3 [2,41,44,8] [  1 - 4,COMMUNICATIONS,TEXAS,   $100,000 - 249,999]
........ Count:2572  Nearest:0.0000  Furthest:2.3128  Mean:1.2009  SSE: 4049  Mean SSE: 1.5742
*  Centre 4 [2,73,44,8] [  1 - 4,ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES,TEXAS,   $100,000
........ Count:53558  Nearest:0.0000  Furthest:3.3128  Mean:2.4528  SSE: 336674  Mean SSE: 6.2862

MODIFY: Moved:4, Deleted:0, Split:0
*  Centre 1 [2*,62,5,8*] [*   1 - 4 * ,BUSINESS SERVICES,CALIFORNIA,*    $100,000 - 249,999 * ] =>MOVED TO=>
*  Centre 2 [2,41*,10,8] [  1 - 4,* COMMUNICATIONS * ,FLORIDA,   $100,000 - 249,999] =>MOVED TO=>   [2,62*,10,
*  Centre 3 [2,41,44*,8] [  1 - 4,COMMUNICATIONS,* TEXAS * ,   $100,000 - 249,999] =>MOVED TO=>   [2,41,5*,8]
*  Centre 4 [2*,73*,44,8*] [*   1 - 4 * ,* ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES * ,TEXAS
```

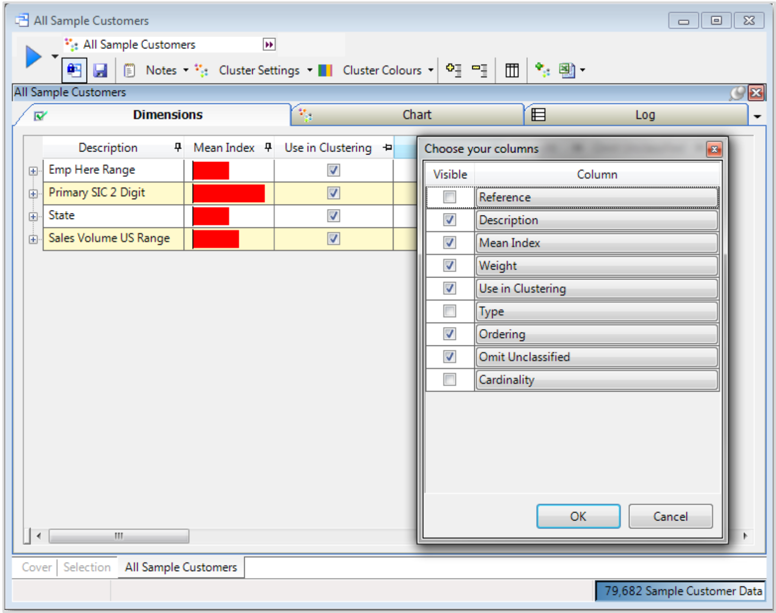


```
-----------------------RUN 4---------------------------------------------------------------
ANALYSE: Clusters: 4, Weighted Mean Dist:1.6524, Between Centres Mean:1.8479, SSE: 238792, Mean SSE:
Centre 1 [4,62,5,1] [ 10 - 14,BUSINESS SERVICES,CALIFORNIA,No Data or 0]
........ Count:20266  Nearest:0.0000  Furthest:2.2872  Mean:1.4178  SSE: 45780  Mean SSE: 2.2590
Centre 2 [3,62,10,8] [  5 - 9,BUSINESS SERVICES,FLORIDA,   $100,000 - 249,999]
........ Count:17205  Nearest:0.0000  Furthest:2.5538  Mean:1.5549  SSE: 46334  Mean SSE: 2.6931
Centre 3 [6,73,5,9] [ 20 - 29,ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES,CALIFORNIA,
........ Count:14501  Nearest:0.1333  Furthest:2.9385  Mean:1.8560  SSE: 53766  Mean SSE: 3.7078
Centre 4 [6,43,44,1] [ 20 - 29,WHOLESALE TRADE - DURABLE GOODS,TEXAS,No Data or 0]
........ Count:27710  Nearest:0.0769  Furthest:2.7282  Mean:1.7779  SSE: 92911  Mean SSE: 3.3530

MODIFY: Moved:0, Deleted:0, Split:0


>>>STOP: Stable clusters

-----------------------END--------------------------------------------------------------
```

Reviewing the Cluster Report

N.B. ***The figures you see on your screen may well be different to those in the following screen shots as the data is dynamic e.g. as time progresses the sales of a Customer may well change. This is a true reflection of reality as the shape of your clusters evolves over time.***

The visualization of the cluster results is found under the three tabs; Dimensions, Chart and Log. The Dimensions graphic shows the distribution of the clusters across the categories of the dimensions and other variables if required. The Scatter Plot chart shows how points for any two dimensions are assigned to the clusters. The Log shows descriptions of the clusters at each stage of evolution.



Description	Mean Index	Use in Clustering
Emp Here Range		☑
Primary SIC 2 Digit		☑
State		☑
Sales Volume US Range		☑

Choose your columns

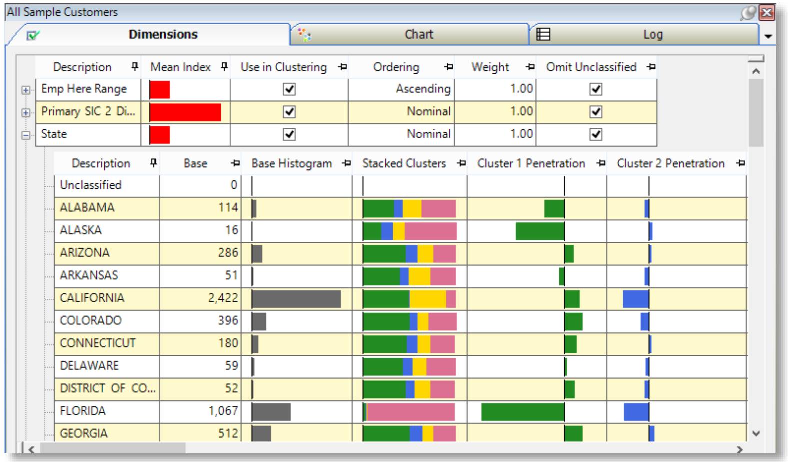
Visible	Column
□	Reference
☑	Description
☑	Mean Index
☑	Weight
☑	Use in Clustering
□	Type
☑	Ordering
☑	Omit Unclassified
□	Cardinality

Dimensions

The Dimensions tab displays a summary line for each variable in the analysis. It is then possible to drill down into the results for each category of that variable.

Not all of the information columns may be on display. To show further information:

- ➢ Right click on one of the existing column headings (variable or category) and select **Column Chooser...**
- ➢ Mark the options to display and click **OK**



Description	Mean Index	Use in Clustering	Ordering	Weight	Omit Unclassified
Emp Here Range		☑	Ascending	1.00	☑
Primary SIC 2 Di...		☑	Nominal	1.00	☑
State		☑	Nominal	1.00	☑

Description	Base	Base Histogram	Stacked Clusters	Cluster 1 Penetration	Cluster 2 Penetration
Unclassified	0				
ALABAMA	114				
ALASKA	16				
ARIZONA	286				
ARKANSAS	51				
CALIFORNIA	2,422				
COLORADO	396				
CONNECTICUT	180				
DELAWARE	59				
DISTRICT OF CO...	52				
FLORIDA	1,067				
GEORGIA	512				

The following page describes each of the column displays at both the variable and category level. This information can help in the understanding of the makeup of each cluster created and the predominant characteristics of the people within them.

Variable Display Options

Reference is the actual name of the variable e.g. Emp Here Range is siEmpHR

Description displays the name assigned to that particular variable

Mean Index gives the average index across all the clusters showing the influence of this variable on the clusters

Weight can be applied to a variable to reflect its importance in creating the clusters e.g. a variable with a weight of 2 is given twice as much importance as a variable with a weight of 1

Use in Clustering if ticked the variable will be used in creating the clusters, if not, the variable will not influence the creation of the clusters but will be used to describe them afterwards (see pg. 14)

Type of variable being used e.g. Numeric or Selector

Omit Unclassified if ticked Market Insight will not use this category in the calculations. It should only be unticked if the unclassified category has a special or real meaning

Cardinality displays the number of categories in the variable

N.B. There are a number of columns displayed. Right clicking on a column header allows a user to filter the columns displayed to those they are particularly interested in. Results for a cluster or particular types of columns can be shown or hidden.

Category Display Options

Description displays the name assigned to that particular variable category

Base figure displays the total number of records in this category

Base Histogram shows graphically the relative number of records in a category overall

Stacked Clusters shows graphically the relative size of each cluster within each category and overall. The Total row shows the overall size of each cluster

Cluster *N* Penetration displays a histogram with an Index value shown centered around 100. Histogram bars to the left of the 100 center line show under representation. Histogram bars to the right of the center line show over representation.

Cluster *N* is the number of records in the cluster in this category

Cluster *N* Index is the ratio of the Cluster *N* and Base figure percentages multiplied by 100

Cluster *N* Z-Score displays a figure which is the standardized measure of how confident we can be that the result presented is a true characteristic of the data and not a quirk of the data sample used. For each category, the Z-Score measures the number of standard deviations the result is away from the expected result of the category.

Cluster *N* Histogram graphically displays the figures in Cluster *N*

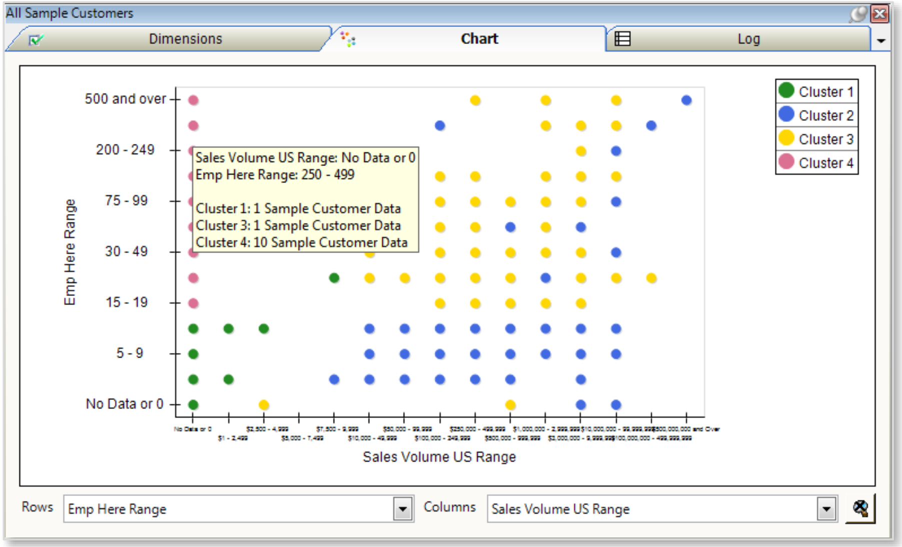
Cluster *N* Base shows a histogram comparison of the Base and Cluster volume figures

Chart

The Chart tab on the Cluster tool displays a Scatter Plot graph. The graph will show how points for any two dimensions are assigned to the clusters.

If more than two variables have been used in the analysis, the drop down display for Rows and Columns allows you to select which of the variables to be used in the chart.

The chart uses a sample 1000 records for the display and the color of each point is assigned to the cluster with highest number of records. A breakdown of each point can be seen in the tool tip box when the mouse pointer moves over the point.



Sales Volume US Range: No Data or 0
Emp Here Range: 250 - 499

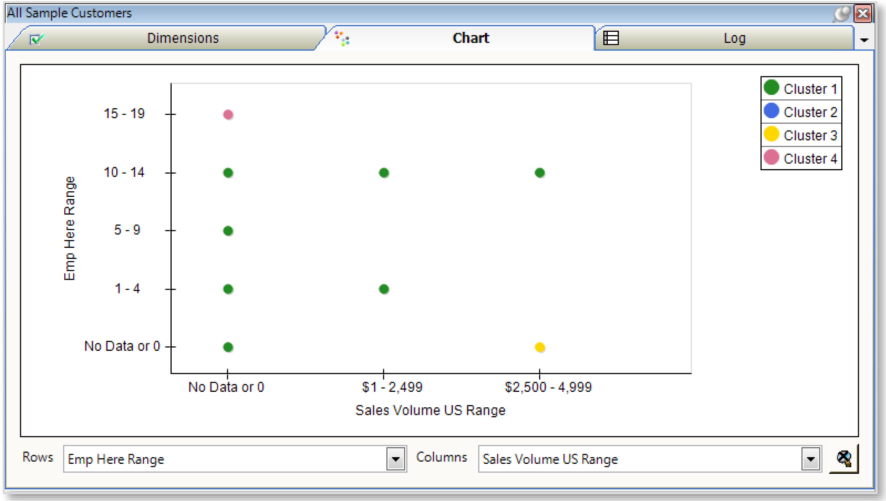
Cluster 1: 1 Sample Customer Data
Cluster 3: 1 Sample Customer Data
Cluster 4: 10 Sample Customer Data

Rows: Emp Here Range
Columns: Sales Volume US Range

The Scatter Plot graph allows for an overall impression of which points belong to each cluster. The screen shot opposite shows the clusters described by Emp Here Range and Sales Volume US Range.

Cluster 1 (Green, bottom left section) show customers with both low numbers of employees and sales.

The chart screen shot opposite has been zoomed into for easier review.



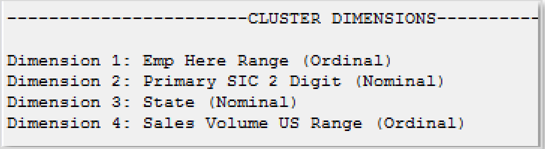
Rows: Emp Here Range
Columns: Sales Volume US Range

Log

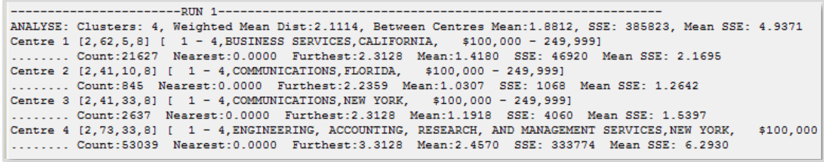
The Log tab describes the actions performed in generating the clusters. After listing the dimensions (variables) at the top of the window the Cluster Evolution section shows each run in calculating the final clusters based upon the Cluster Settings (see the next section).

Having used the default settings in this example i.e. number of clusters 4, City Block, Medioid and Frequency, the cluster tool will use a technique called K-Means to detect a fixed number of clusters from a set of data points (see **Appendix 1**).

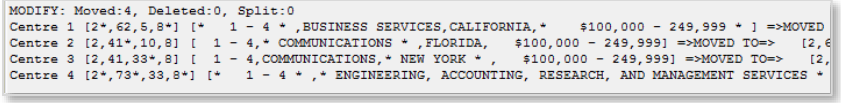
In this example, 4 centers will be located based upon the points with highest number of records. The K-Means technique will allocate each point to its nearest cluster center. Once this has been done the midpoint of each cluster (in this case Medioid) is calculated and used as the new center of the cluster. The process repeats itself to now reallocate the points to the nearest new cluster center. This will continue until no points are reallocated to a different cluster center or the maximum number of iterations has been met.



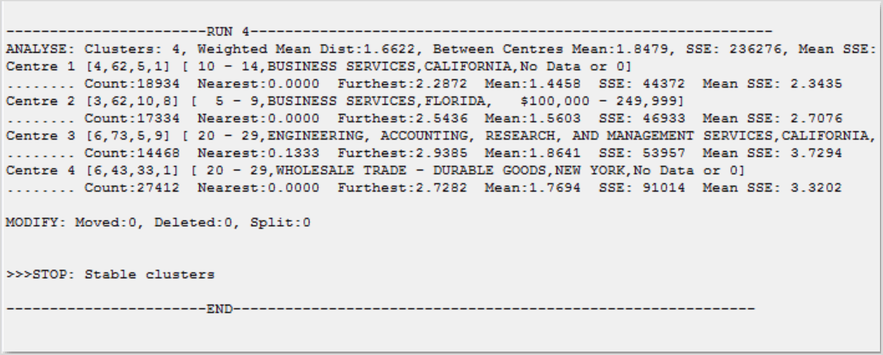
```
-----------------------CLUSTER DIMENSIONS----------

Dimension 1: Emp Here Range (Ordinal)
Dimension 2: Primary SIC 2 Digit (Nominal)
Dimension 3: State (Nominal)
Dimension 4: Sales Volume US Range (Ordinal)
```



```
----------------------RUN 1-----------------------------------------------------------
ANALYSE: Clusters: 4, Weighted Mean Dist:2.1114, Between Centres Mean:1.8812, SSE: 385823, Mean SSE: 4.9371
Centre 1 [2,62,5,8] [  1 - 4,BUSINESS SERVICES,CALIFORNIA,   $100,000 - 249,999]
........ Count:21627  Nearest:0.0000  Furthest:2.3128  Mean:1.4180  SSE: 46920  Mean SSE: 2.1695
Centre 2 [2,41,10,8] [  1 - 4,COMMUNICATIONS,FLORIDA,   $100,000 - 249,999]
........ Count:845  Nearest:0.0000  Furthest:2.2359  Mean:1.0307  SSE: 1068  Mean SSE: 1.2642
Centre 3 [2,41,33,8] [  1 - 4,COMMUNICATIONS,NEW YORK,   $100,000 - 249,999]
........ Count:2637  Nearest:0.0000  Furthest:2.3128  Mean:1.1918  SSE: 4060  Mean SSE: 1.5397
Centre 4 [2,73,33,8] [  1 - 4,ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES,NEW YORK,   $100,000
........ Count:53039  Nearest:0.0000  Furthest:3.3128  Mean:2.4570  SSE: 333774  Mean SSE: 6.2930
```



```
MODIFY: Moved:4, Deleted:0, Split:0
Centre 1 [2*,62,5,8*] [*   1 - 4 * ,BUSINESS SERVICES,CALIFORNIA,*    $100,000 - 249,999 * ] =>MOVED
Centre 2 [2,41*,10,8] [  1 - 4,* COMMUNICATIONS * ,FLORIDA,   $100,000 - 249,999] =>MOVED TO=>   [2,6
Centre 3 [2,41,33*,8] [  1 - 4,COMMUNICATIONS,* NEW YORK * ,   $100,000 - 249,999] =>MOVED TO=>   [2,
Centre 4 [2*,73*,33,8*] [*   1 - 4 * ,* ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES *
```



```
-----------------------RUN 4-----------------------------------------------------------
ANALYSE: Clusters: 4, Weighted Mean Dist:1.6622, Between Centres Mean:1.8479, SSE: 236276, Mean SSE:
Centre 1 [4,62,5,1] [ 10 - 14,BUSINESS SERVICES,CALIFORNIA,No Data or 0]
........ Count:18934  Nearest:0.0000  Furthest:2.2872  Mean:1.4458  SSE: 44372  Mean SSE: 2.3435
Centre 2 [3,62,10,8] [  5 - 9,BUSINESS SERVICES,FLORIDA,   $100,000 - 249,999]
........ Count:17334  Nearest:0.0000  Furthest:2.5436  Mean:1.5603  SSE: 46933  Mean SSE: 2.7076
Centre 3 [6,73,5,9] [ 20 - 29,ENGINEERING, ACCOUNTING, RESEARCH, AND MANAGEMENT SERVICES,CALIFORNIA,
........ Count:14468  Nearest:0.1333  Furthest:2.9385  Mean:1.8641  SSE: 53957  Mean SSE: 3.7294
Centre 4 [6,43,33,1] [ 20 - 29,WHOLESALE TRADE - DURABLE GOODS,NEW YORK,No Data or 0]
........ Count:27412  Nearest:0.0000  Furthest:2.7282  Mean:1.7694  SSE: 91014  Mean SSE: 3.3202

MODIFY: Moved:0, Deleted:0, Split:0


>>>STOP: Stable clusters

-----------------------END-----------------------------------------------------------
```

Reviewing the Cluster Settings

There are a number of options that can be changed in the Cluster Settings to determine the number, size and calculations used in the creation of clusters.

- Click on the **Cluster Settings** button to reveal the window opposite

Basic Options

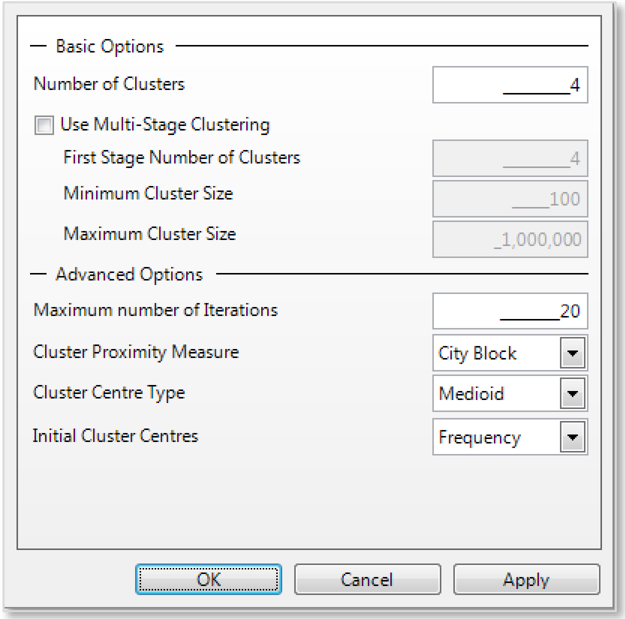
Number of Clusters

A figure that defines the required number of clusters to be found.

Use Multi-Stage Clustering

When the Multi-Stage Clustering option is used, the standard K-Means algorithm described in **Appendix 1** is used to create a stable set of First-Stage clusters. The multi-stage algorithm then progressively adds or removes clusters one at a time until the **Number of Clusters** is achieved - at each stage the standard K-Means algorithm is used until a stable set of clusters has formed. This process is similar to basic Divisive Clustering and Agglomerative Clustering, with the advantage that people can change cluster at each of the multiple stages so that the final solution is less sensitive to the initial set of cluster centers.

Divisive Clustering – traditionally this starts by treating all points as if they are part of a single large cluster, and then partitions the cluster into smaller and smaller clusters. By making the First Stage Number of Clusters figure *smaller* than the Number of Clusters figure, the multi-stage algorithm progressively increases the number of cluster centers, forming a stable set of smaller clusters each time



Basic Options

Number of Clusters 4

□ Use Multi-Stage Clustering

First Stage Number of Clusters 4

Minimum Cluster Size 100

Maximum Cluster Size 1,000,000

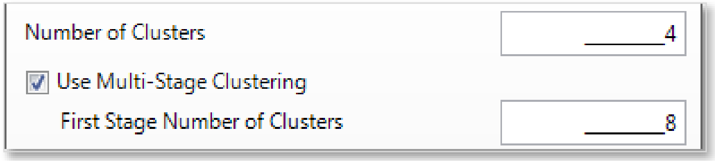
Advanced Options

Maximum number of Iterations 20

Cluster Proximity Measure City Block

Cluster Centre Type Medioid

Initial Cluster Centres Frequency



Number of Clusters 4

☑ Use Multi-Stage Clustering

First Stage Number of Clusters 8

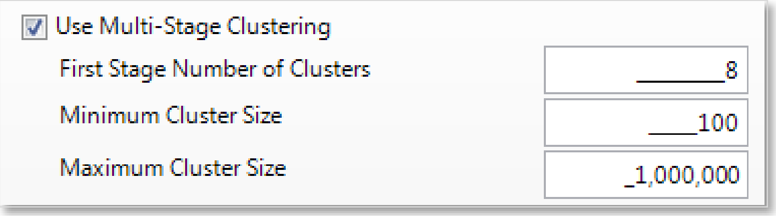
Agglomerative Clustering – traditionally starts by treating each point as a separate cluster, and then combines them into bigger and bigger clusters. By making the First Stage Number of Clusters figure *larger* than the Number of Clusters figure, the multi-stage algorithm progressively removes one of the cluster centers, forming a stable set of larger clusters each time

Minimum Cluster Size

This determines the minimum number of points in a cluster as the tool works through the Multi-Stage clustering process.

Maximum Cluster Size

This determines the maximum number of points in a cluster as the tool works through the Multi-Stage clustering process.



☑ Use Multi-Stage Clustering

First Stage Number of Clusters 8

Minimum Cluster Size 100

Maximum Cluster Size 1,000,000

Advanced Options

Maximum number of Iterations

This figure is the maximum number of runs through the cluster analysis before the final results are shown. This figure may not be reached if it is determined the clusters are stable after fewer runs.

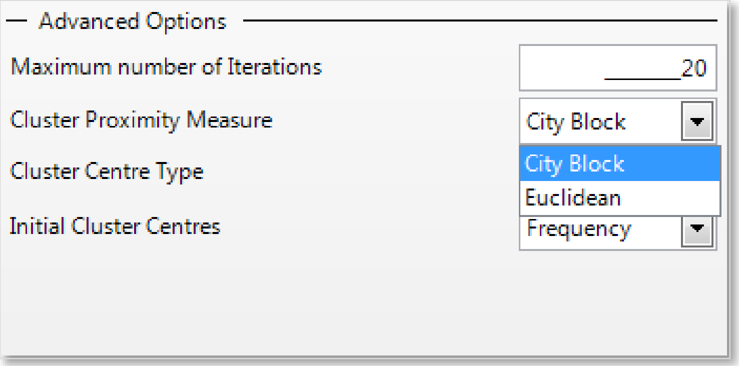
Cluster Proximity Measure

Euclidean

In mathematics, the Euclidean distance or Euclidean metric is the "straight line" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula.

City Block

This is a form of geometry in which the usual metric of Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their coordinates. Also known as Manhattan Distance whose name alludes to the grid layout of most streets on the island of Manhattan, which causes the shortest path a car, could take between two points in the borough.



Advanced Options

Maximum number of Iterations 20

Cluster Proximity Measure City Block

Cluster Centre Type

Initial Cluster Centres Frequency

Cluster Centre Type

Medioid

This is the middle point of a cluster and selects an actual point in the data. (see **Appendix 1** – Midpoint Cluster Centre)

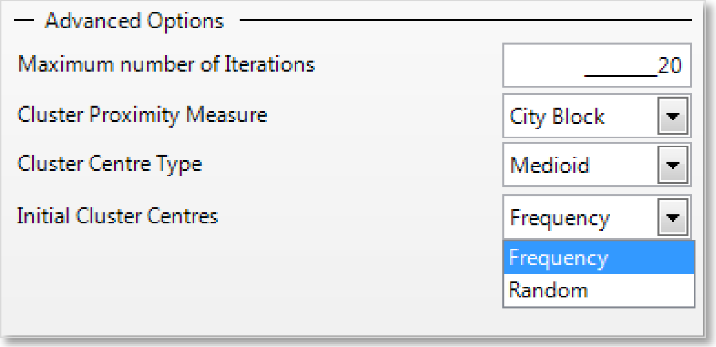
Initial Cluster Centers

Frequency

This uses the point with the highest number of records to set the initial cluster center. Other cluster centers are selected by the next highest but distant from the other initial centers, with consideration taken to ensure that the centers are not too close together.

Random

This uses random starting points to set the initial cluster centers.



Advanced Options

Maximum number of Iterations: 20

Cluster Proximity Measure: City Block

Cluster Centre Type: Medioid

Initial Cluster Centres: Frequency

Output Clusters as a Virtual Variable

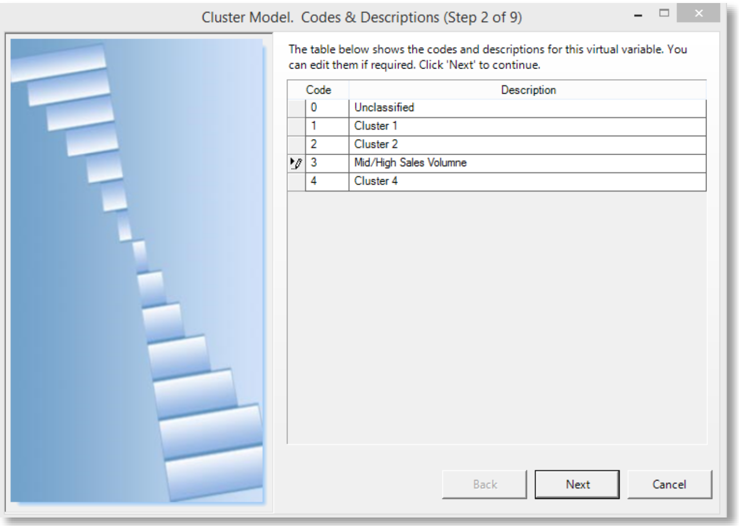
A Virtual Variable needs to be created to access the records identified from the clustering analysis. With the records identified and accessible through a selector, other Market Insight functionality (Cubes, Data Grids etc.) and other modelling tools (Profiler & Decision Tree) can be used to explore these groups further.

To create a Virtual Variable from any Cluster Analysis use the Cluster Model wizard from the Wizard window. Within an open Cluster model:

- Click on the **Create Cluster Model** button on the icon bar

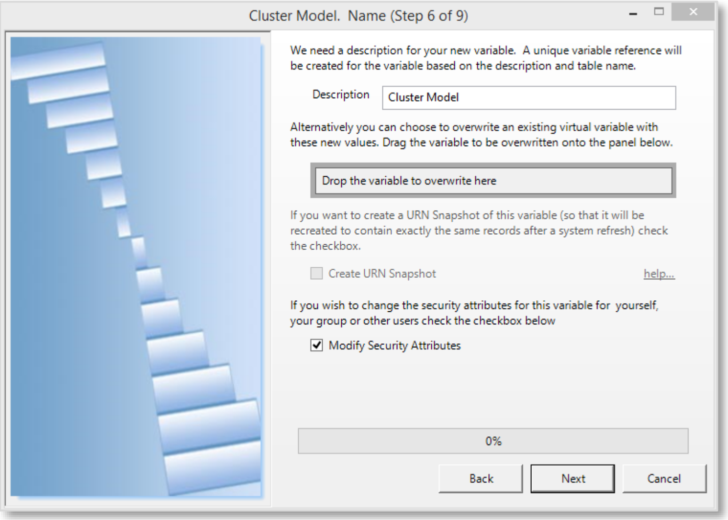
A series of step by step instructions act as a guide through the creation of the Virtual Variable.

- Step 1 – Will only be displayed if the Create Cluster Model option is selected from the Wizard window and the Cluster Analysis to be used needs identifying
- Step 2 – This step displays the current Codes and Descriptions for this virtual variable; overtype the descriptions with relevant names for each cluster then click **Next**
- Step 3 – Leave blank as we want all records, otherwise a selection can be used here to narrow down what records are used in the virtual variable, click **Next**
- Step 4 – Click, or create, the folder where the variable is to be saved. Leave on others in this case and click **Next**
- Step 5 – Enter any notes, click **Next**



The table below shows the codes and descriptions for this virtual variable. You can edit them if required. Click 'Next' to continue.

Code	Description
0	Unclassified
1	Cluster 1
2	Cluster 2
3	Mid/High Sales Volumne
4	Cluster 4



We need a description for your new variable. A unique variable reference will be created for the variable based on the description and table name.

Description: Cluster Model

Alternatively you can choose to overwrite an existing virtual variable with these new values. Drag the variable to be overwritten onto the panel below.

Drop the variable to overwrite here

If you want to create a URN Snapshot of this variable (so that it will be recreated to contain exactly the same records after a system refresh) check the checkbox.

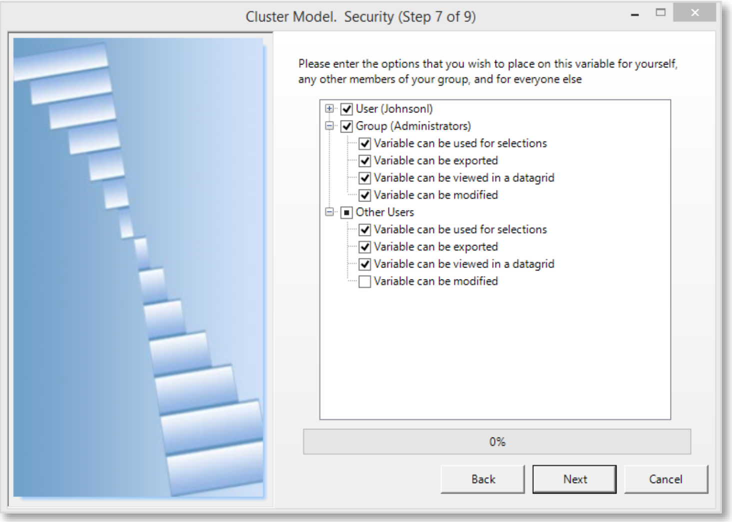
□ Create URN Snapshot

If you wish to change the security attributes for this variable for yourself, your group or other users check the checkbox below

☑ Modify Security Attributes

- Step 6 – Enter the description **Cluster Model,** click **Next**

- Step 7 – Security feature to control access and use of the variable (not applicable on local system)



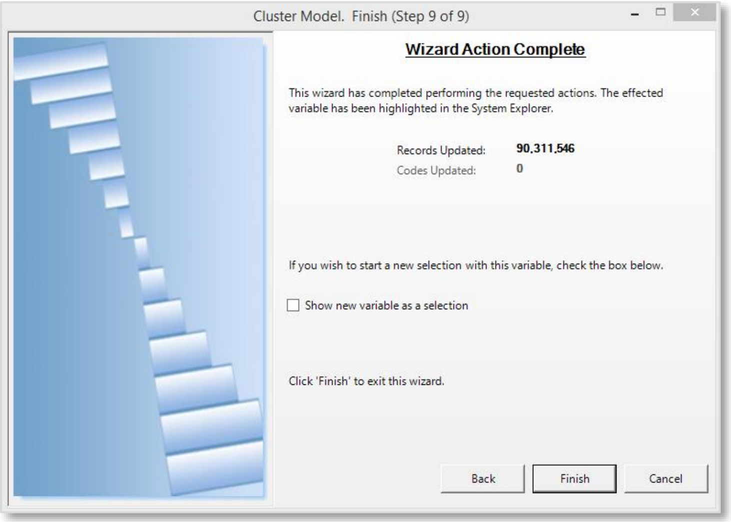
Cluster Model. Security (Step 7 of 9)

Please enter the options that you wish to place on this variable for yourself, any other members of your group, and for everyone else

- ☑ User (Johnsonl)
- ☑ Group (Administrators)
  - ☑ Variable can be used for selections
  - ☑ Variable can be exported
  - ☑ Variable can be viewed in a datagrid
  - ☑ Variable can be modified
- ■ Other Users
  - ☑ Variable can be used for selections
  - ☑ Variable can be exported
  - ☑ Variable can be viewed in a datagrid
  - □ Variable can be modified

0%

- Step 9 – Click **Finish**



Cluster Model. Finish (Step 9 of 9)

Wizard Action Complete

This wizard has completed performing the requested actions. The effected variable has been highlighted in the System Explorer.

Records Updated: 90,311,546

Codes Updated: 0

If you wish to start a new selection with this variable, check the box below.

□ Show new variable as a selection

Click 'Finish' to exit this wizard.

Introduction to Profile

One of the key mechanisms of Market Insight is the ability to profile a selection. A profile identifies the significant characteristics of the selected records when compared with another set of records. A profile can be built of any group of records selected using any of the methods supported by Market Insight, including those generated from the Cluster model. You can control whether the profile is built against the whole database (the universe) or just a part of it.

Understanding the characteristics of your current customers and your strengths and weaknesses across market sectors is a prerequisite for any marketing activity. Whether the marketing strategy is to expand to new sectors or up-sell in existing sectors it is necessary to consider your current sector performance and the remaining market potential before embarking on list acquisition and expensive direct marketing promotions.

Profiling compares your current customer base against a data “universe“. A profile report highlights market sectors in which your business is over or under performing relative to the universe population. This information can be used directly to guide marketing strategy and in particular prospect list selection.

Until now profile reports have required bespoke consultancy and hours of processing. Market Insight produces profiles on millions of records in just a few seconds and most importantly is accessible to marketers without background statistical knowledge.

Profiling is not limited to detecting simple customer characteristics: any segment of the data may be analyzed against any background. For example the customers for a particular product may be analyzed against just the applicable industry sector or even the set of prospects that have been offered the product.

The speed and flexibility of Market Insight make it practical to include multiple profiling stages in a marketing strategy for research, selection and response analysis. Leading on from Profiling, the Market Insight Modelling & Scoring technology enables you to score prospects on the Marketing Universe based on your market profile and responses to close a virtuous marketing loop.

Why Profile?

- Because any analysis of a group of records in isolation will only provide a limited understanding of the characteristics of those records.

- By comparing the records with other records a fuller understanding can be reached.

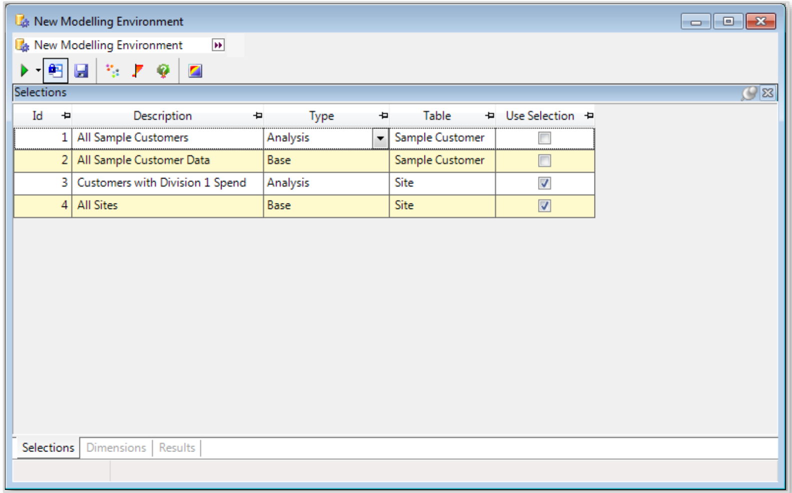
A Profile Worked Example

Profiling compares one group of records with another to identify the characteristics that differentiate them. In this example we will use Sites where they have spent with our company's Division 1. The profile will then let us easily identify those sites on the rest of the database that "look like" the Division 1 spenders and therefore make a good marketing target.

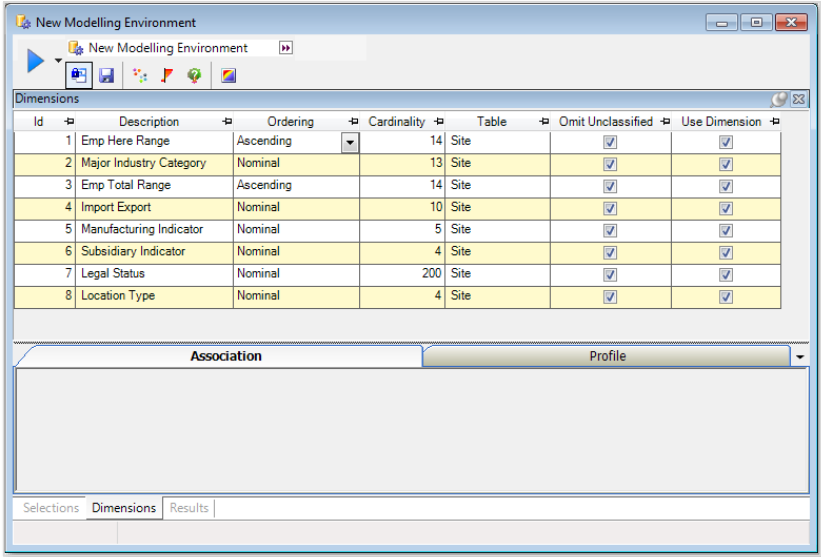
Profiling detects characteristics by identifying results that occur in your selected data and which are highly unlikely to occur in a truly random sample of data.

Setting up within the Modelling Environment

- Redisplay from your Private folder the **New Modelling Environment** window on the workspace if it has been closed
- Create a **Site** selection of **Division 1 Spend** of **>0**. Rename the selection window **Sites with Division 1 Spend** and drag onto the **Selections** window of the **New Modelling Environment**
- Create a **Site** selection of **All Sites** and drag on as above
- Set **Sites with Division 1 Spend** to be the **Analysis** selection and **All Sites** to be the **Base** selection.
- Click onto the **Dimensions** tab and select all the variables except **Emp Here Range** and with a right click select **Delete**
- Drag on the following variables – **Major Industry Category, Emp Total Range, Import Export, Manufacturing Indicator, Subsidiary Indicator, Legal Status** and **Location Type**



Id	Description	Type	Table	Use Selection
1	All Sample Customers	Analysis	Sample Customer	□
2	All Sample Customer Data	Base	Sample Customer	□
3	Customers with Division 1 Spend	Analysis	Site	☑
4	All Sites	Base	Site	☑



Id	Description	Ordering	Cardinality	Table	Omit Unclassified	Use Dimension
1	Emp Here Range	Ascending	14	Site	☑	☑
2	Major Industry Category	Nominal	13	Site	☑	☑
3	Emp Total Range	Ascending	14	Site	☑	☑
4	Import Export	Nominal	10	Site	☑	☑
5	Manufacturing Indicator	Nominal	5	Site	☑	☑
6	Subsidiary Indicator	Nominal	4	Site	☑	☑
7	Legal Status	Nominal	200	Site	☑	☑
8	Location Type	Nominal	4	Site	☑	☑

- Save the **New Modelling Environment** in your **Private** folder

Setting up <u>outside</u> the Modelling Environment

- Create a **Site** selection of **Division 1 Spend** of **>0**. Rename the selection window **Sites with Division 1 Spend**

- From the **Toolbox** window drag and drop the **Profile** tool onto the open selection page

The target records (Sites with Division 1 Spend) are referred to as the **Analysis** selection.

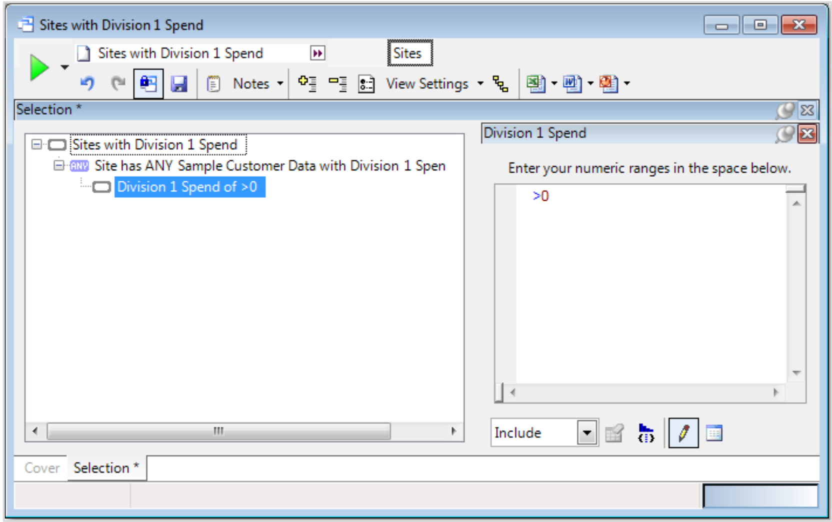
The data which we will be profiling against is referred to as the **Base** selection.

- Click on the **Base Selection** tab (we will leave this blank so it will use the entire database by default as the base selection)

Having determined the Analysis and Base group we now need to specify the Variables for inclusion in the profile.

- Ensure the **Profile** tab is upper most by clicking on it

- Drag on the following variables – **Emp Here Range, Major Industry Category, Emp Total Range, Import Export, Manufacturing Indicator, Subsidiary Indicator, Legal Status** and **Location Type**

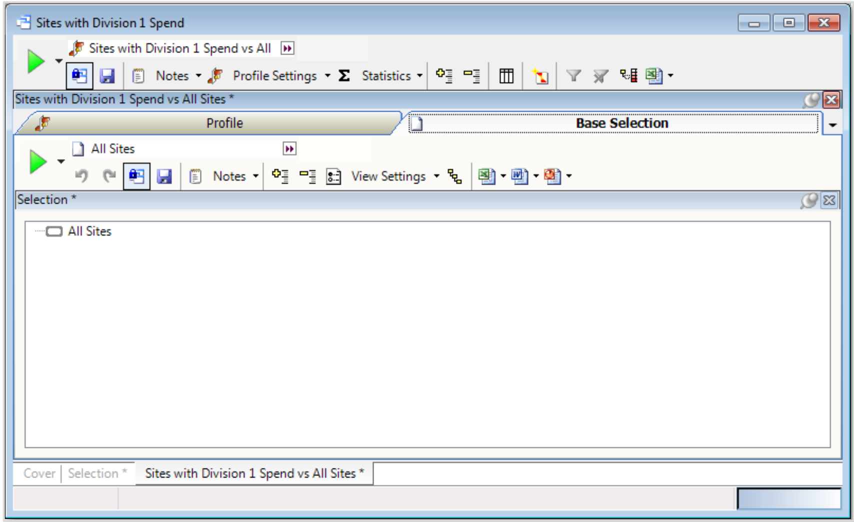


Sites with Division 1 Spend
Site has ANY Sample Customer Data with Division 1 Spen
Division 1 Spend of >0

Enter your numeric ranges in the space below.

>0

Include


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Profile Settings
Statistics
Sites with Division 1 Spend vs All Sites *
Profile
Base Selection
All Sites
Notes
View Settings
Selection *
Cover
Selection *
Sites with Division 1 Spend vs All Sites *

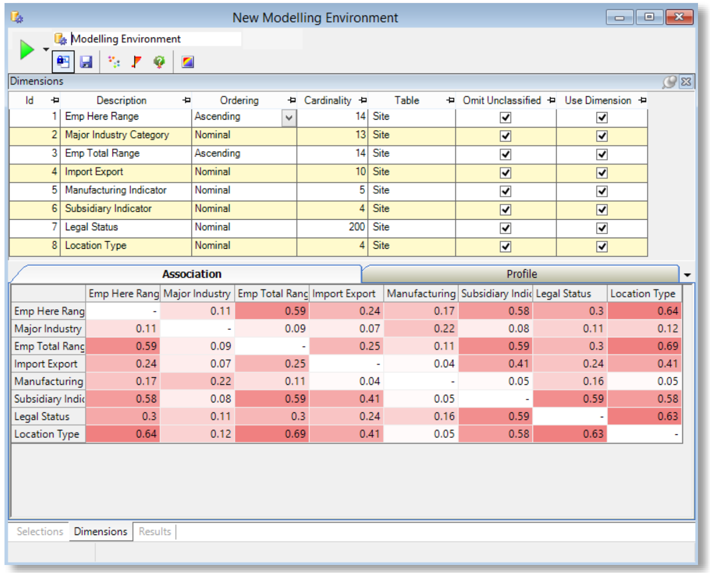
All Sites

Within the dimensions tab of the modelling environment we can view the associations between the dimensions that have been dragged on.

- Ensure the dimensions tab is uppermost and click the ▶ **Build** button

In the resulting cube the closer to 1.00 the more direct the association between the two variables. If there are variables that are closely associated we may wish to consider if we should include both when building the profile.

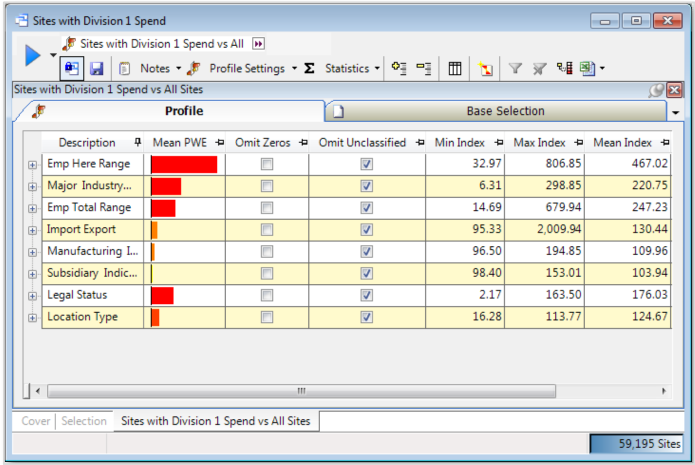
N.B. As well as Selector variables you can use Numeric, Currency, Date variables and selection queries in your profile.



Id	Description	Ordering	Cardinality	Table	Omit Unclassified	Use Dimension
1	Emp Here Range	Ascending	14	Site	☑	☑
2	Major Industry Category	Nominal	13	Site	☑	☑
3	Emp Total Range	Ascending	14	Site	☑	☑
4	Import Export	Nominal	10	Site	☑	☑
5	Manufacturing Indicator	Nominal	5	Site	☑	☑
6	Subsidiary Indicator	Nominal	4	Site	☑	☑
7	Legal Status	Nominal	200	Site	☑	☑
8	Location Type	Nominal	4	Site	☑	☑

	Emp Here Rang	Major Industry	Emp Total Ranc	Import Export	Manufacturing	Subsidiary Indic	Legal Status	Location Type
Emp Here Rang	-	0.11	0.59	0.24	0.17	0.58	0.3	0.64
Major Industry	0.11	-	0.09	0.07	0.22	0.08	0.11	0.12
Emp Total Ranc	0.59	0.09	-	0.25	0.11	0.59	0.3	0.69
Import Export	0.24	0.07	0.25	-	0.04	0.41	0.24	0.41
Manufacturing	0.17	0.22	0.11	0.04	-	0.05	0.16	0.05
Subsidiary Indic	0.58	0.08	0.59	0.41	0.05	-	0.59	0.58
Legal Status	0.3	0.11	0.3	0.24	0.16	0.59	-	0.63
Location Type	0.64	0.12	0.69	0.41	0.05	0.58	0.63	-

- When you are satisfied with the dimensions click the ▶ **Build** button to run the profile


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Profile Settings
Σ Statistics
Sites with Division 1 Spend vs All Sites
Profile
Base Selection
Cover
Selection
Sites with Division 1 Spend vs All Sites
59,195 Sites

Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index
Emp Here Range		☐	☑	32.97	806.85	467.02
Major Industry...		☐	☑	6.31	298.85	220.75
Emp Total Range		☐	☑	14.69	679.94	247.23
Import Export		☐	☑	95.33	2,009.94	130.44
Manufacturing I...		☐	☑	96.50	194.85	109.96
Subsidiary Indic...		☐	☑	98.40	153.01	103.94
Legal Status		☐	☑	2.17	163.50	176.03
Location Type		☐	☑	16.28	113.77	124.67

Reviewing the Profile Report

We will start by drilling down to the Category level and view the results.

- Click on the **Profile** tab

- Click the + sign next to the **Major Industry Category** variable

The expanded report lists all the variable categories together with the counts of the Analysis and Base selections and measures calculated from them.

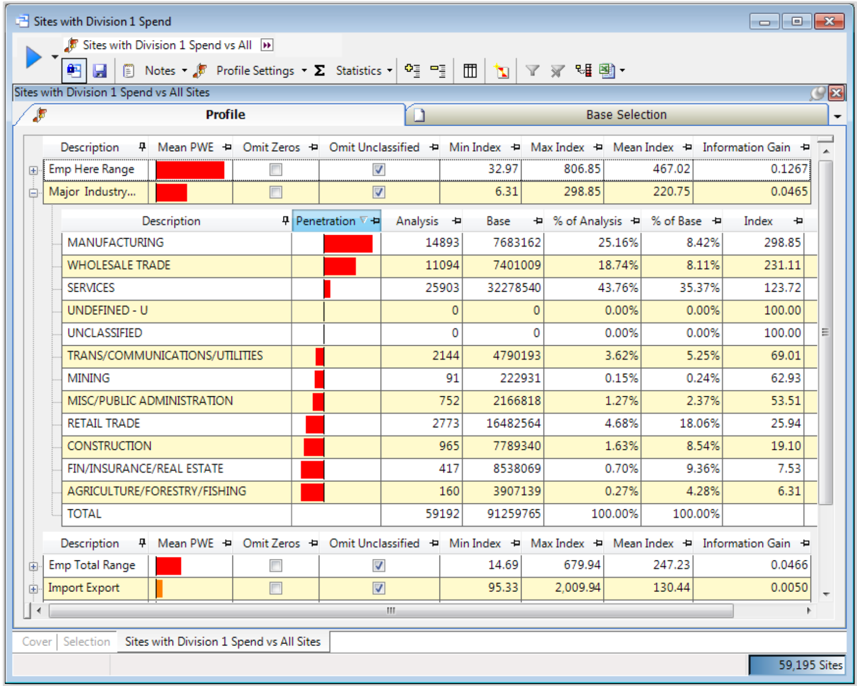
The Penetration histogram gives a quick visual interpretation of the over or under representation of the Analysis selection in each category. The size and direction of the bar is determined by the Index and the color by the statistical significance (as measured by the Z-score yellow low → red high).

- Right click on the **Penetration** column and select **Sort Descending** to order the display

In the report opposite the categories Manufacturing, Wholesale Trade and Services show positive penetration whereas Finance, Agriculture etc. show strong negative penetration.

The simple marketing interpretation is:

"Our Best Customers when compared to the average seem to have an MIC of Manufacturing, Wholesale Trade, and Services and are less likely to be within Finance, Agriculture etc."


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Profile Settings
Statistics
Sites with Division 1 Spend vs All Sites
Profile
Base Selection

Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index	Information Gain
Emp Here Range	🟥	□	☑	32.97	806.85	467.02	0.1267
Major Industry...	🟥	□	☑	6.31	298.85	220.75	0.0465

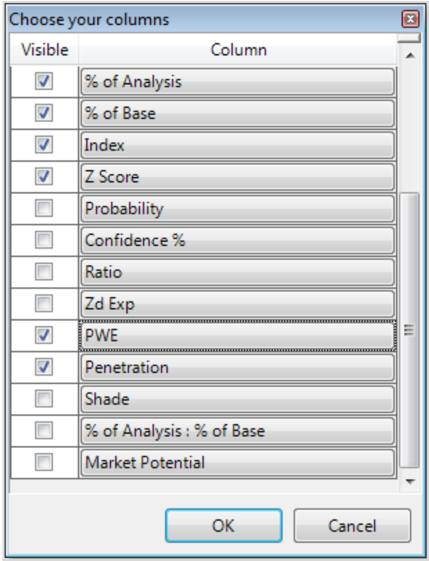
Description	Penetration	Analysis	Base	% of Analysis	% of Base	Index
MANUFACTURING	🟥	14893	7683162	25.16%	8.42%	298.85
WHOLESALE TRADE	🟥	11094	7401009	18.74%	8.11%	231.11
SERVICES	🟥	25903	32278540	43.76%	35.37%	123.72
UNDEFINED - U		0	0	0.00%	0.00%	100.00
UNCLASSIFIED		0	0	0.00%	0.00%	100.00
TRANS/COMMUNICATIONS/UTILITIES	🟥	2144	4790193	3.62%	5.25%	69.01
MINING	🟥	91	222931	0.15%	0.24%	62.93
MISC/PUBLIC ADMINISTRATION	🟥	752	2166818	1.27%	2.37%	53.51
RETAIL TRADE	🟥	2773	16482564	4.68%	18.06%	25.94
CONSTRUCTION	🟥	965	7789340	1.63%	8.54%	19.10
FIN/INSURANCE/REAL ESTATE	🟥	417	8538069	0.70%	9.36%	7.53
AGRICULTURE/FORESTRY/FISHING	🟥	160	3907139	0.27%	4.28%	6.31
TOTAL		59192	91259765	100.00%	100.00%	

Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index	Information Gain
Emp Total Range	🟥	□	☑	14.69	679.94	247.23	0.0466
Import Export	🟧	□	☑	95.33	2,009.94	130.44	0.0050

Category Columns

Not all of the information columns may be on display. To show further information:

- ➢ Right click on one of the existing column headings (variable or category) and select **Column Chooser...**
- ➢ Tick the options to display (e.g. **PWE**) and click **OK**


Choose your columns

Visible	Column
☑	% of Analysis
☑	% of Base
☑	Index
☑	Z Score
☐	Probability
☐	Confidence %
☐	Ratio
☐	Zd Exp
☑	PWE
☑	Penetration
☐	Shade
☐	% of Analysis : % of Base
☐	Market Potential

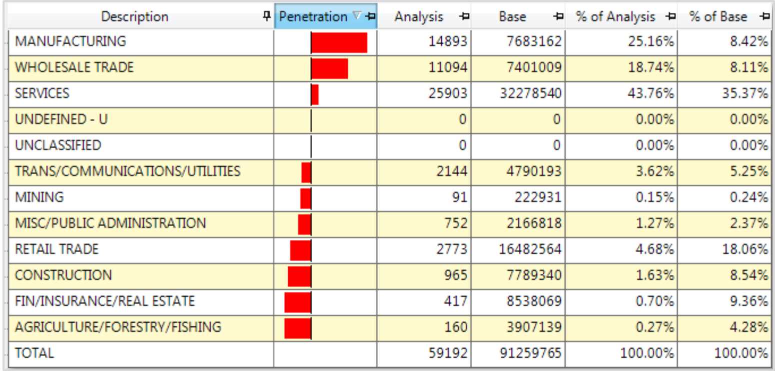
The displayed columns at the category level are described below:

Analysis is the actual number of the analysis records.

Base is the actual number of the base records.

% of Analysis is the percentage of analysis records in each category.

% of Base is the percentage of base records in each category.



Description	Penetration	Analysis	Base	% of Analysis	% of Base
MANUFACTURING		14893	7683162	25.16%	8.42%
WHOLESALE TRADE		11094	7401009	18.74%	8.11%
SERVICES		25903	32278540	43.76%	35.37%
UNDEFINED - U		0	0	0.00%	0.00%
UNCLASSIFIED		0	0	0.00%	0.00%
TRANS/COMMUNICATIONS/UTILITIES		2144	4790193	3.62%	5.25%
MINING		91	222931	0.15%	0.24%
MISC/PUBLIC ADMINISTRATION		752	2166818	1.27%	2.37%
RETAIL TRADE		2773	16482564	4.68%	18.06%
CONSTRUCTION		965	7789340	1.63%	8.54%
FIN/INSURANCE/REAL ESTATE		417	8538069	0.70%	9.36%
AGRICULTURE/FORESTRY/FISHING		160	3907139	0.27%	4.28%
TOTAL		59192	91259765	100.00%	100.00%

✎ **N.B.** There is no point in comparing the numbers in the analysis and base columns in a penetration profile, the analysis is a sub set of the base and will therefore always be a smaller quantity. To make these results comparable we use the relative population of the categories by calculating the Analysis and Base percentages.

Index is the ratio of the percentages multiplied by 100. (see **Appendix 2** for further details)

Z-Score is the standardized measure of how confident we can be that the result presented is a true characteristic of the data and not a quirk of the data sample used.

For each category, the Z-Score measures the number of standard deviations our result is away from the expected result of the category.

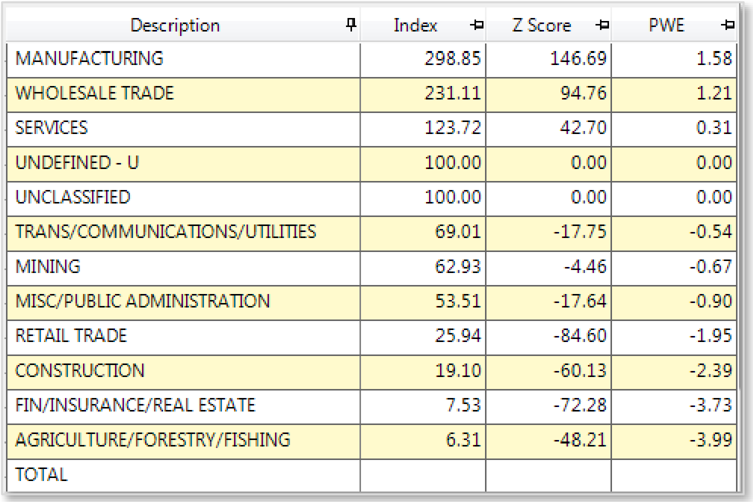
The further a result is from the average (or "expected") result, the less chance there is that this result is just a quirk of data sampling and conversely the more chance there is that the result is revealing some bias in the records analyzed. (see **Appendix 2** for further details)

PWE (Predictive Weight of Evidence)* this value is the amount of evidence that membership of that category gives to support a record being in the analysis set. This is done by examining the proportion of people in the analysis selection, comparing the proportion observed in each category with the observed overall. This provides a PWE value for each category, where a positive weight indicates that people from the category are more likely to be in the analysis selection.

Category	Index	PWE	Interpretation
Manufacturing	298.85	1.58	Strong evidence for
Agriculture	6.31	-3.99	Strong evidence against

Penetration is a graphical view of the index value. Histogram bars to the left of the 100 center line show under representation. Histogram bars to the right of the center line show over representation. The shading is as shown opposite.

A description of all the category options can be found in the Help files.



Description	Index	Z Score	PWE
MANUFACTURING	298.85	146.69	1.58
WHOLESALE TRADE	231.11	94.76	1.21
SERVICES	123.72	42.70	0.31
UNDEFINED - U	100.00	0.00	0.00
UNCLASSIFIED	100.00	0.00	0.00
TRANS/COMMUNICATIONS/UTILITIES	69.01	-17.75	-0.54
MINING	62.93	-4.46	-0.67
MISC/PUBLIC ADMINISTRATION	53.51	-17.64	-0.90
RETAIL TRADE	25.94	-84.60	-1.95
CONSTRUCTION	19.10	-60.13	-2.39
FIN/INSURANCE/REAL ESTATE	7.53	-72.28	-3.73
AGRICULTURE/FORESTRY/FISHING	6.31	-48.21	-3.99
TOTAL			

* **PWE** is a hybrid model building technique based on Information Theory and Bayesian Probability. For more details on how PWE Models are created, see the Apteco White Paper – "FastStats PWE Models: Further Details".

Color	**Z-Score**	**Confidence**
Red	> + / - 3.29	>99.9%
Red Orange	> +/ - 2.576	>99%
Orange	> +/ - 1.96	>95%
Yellow Orange	> + / - 1.65	>90%
Yellow	<= + / - 1.65	<= 90%

Variable Column Summary

The default displayed columns at the variable level are described below:

Description is the name of the variable within Market Insight.

Omit Zeros will hide categories with zero values in the display.

Omit Unclassified will not use the unclassified categories in the analysis.

Min. Index - Minimum Index – the minimum index from a statistically significant (Z-Score > 1.96) category of the variable.

Max. Index - Maximum Index – the maximum index from a statistically significant (Z-Score > 1.96) category of the variable.

Mean Index – the average index from a statistically significant (Z-Score > 1.96) category of the variable.

Information Gain – is a standard statistic that measures the relationship between each variable and the Analysis selection. The higher the value (range 0 to 1) the stronger the relationship. This figure is useful in deciding which variables you want to include in the Profile, especially if there are many variables to choose from.

Mean PWE is calculated as a weighted average of the evidence provided by the categories of the variable in predicting membership of the Analysis selection. So Mean PWE is a summary of generally how useful a variable is to us in predicting whether a prospect will be in the Analysis selection.

Mean PWE is calculated separately for each variable by using the base percentage to weigh the PWE measure for those categories of the variable that have statistically significant results (Z-Score above 1.96 i.e. 95% confidence).

The weighting uses the base percentage to ensure that a large difference in a category with few records is less significant than a moderate difference in a category with more records. You can also interpret the Mean PWE as a summary of how much of a relationship exists between the variable and the analysis selection. The higher the Mean PWE, the greater the relationship.

The Mean PWE measures are comparable between variables. The shading emphasizes the magnitude of the Mean PWE. The higher the value the more information is provided by the variable. In the example above Income has the highest Mean PWE, so knowledge of a prospects Income level is the most valuable in predicting whether they will be in the Analysis selection.

N.B. A very high Mean PWE measure might indicate a suspiciously direct correlation between the variable and your analysis criteria. This would perhaps be a variable that is determined by the analysis selection.

Working with the Profile

Once the profile has been created you may want to use some of the following features to view the information.

Column Formatting

You can use the grid controls to move and resize columns. If you “pin” a column it will remain in view while you scroll across.

➢ Click on the pin symbol next to the **Penetration** column heading

The column will now be positioned next to the Description column at the beginning of the display. As you scroll to view the other columns they will slide underneath those that have been pinned. To undo this action:

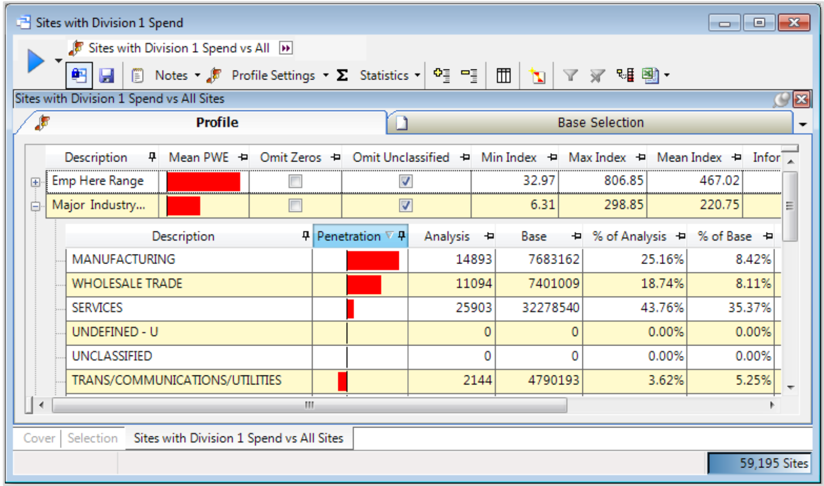
➢ Click on the pin symbol once more and then drag the column back to its original position or another if you so prefer

Column Chooser

This enables you to display or hide columns. These settings are maintained when you save the profile.

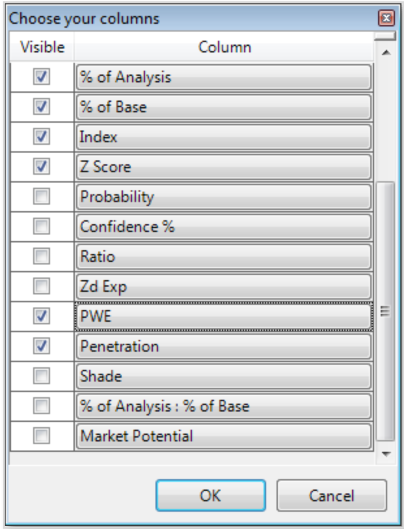
➢ Right click on one of the column headings and select **Column Chooser**

➢ Tick the columns you would like to display and click **OK**


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Profile Settings
Statistics
Sites with Division 1 Spend vs All Sites
Profile
Base Selection

Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index	Infor
Emp Here Range	🟥	☐	☑	32.97	806.85	467.02	
Major Industry...	🟥	☐	☑	6.31	298.85	220.75	

Description	Penetration	Analysis	Base	% of Analysis	% of Base
MANUFACTURING	🟥	14893	7683162	25.16%	8.42%
WHOLESALE TRADE	🟥	11094	7401009	18.74%	8.11%
SERVICES	🟥	25903	32278540	43.76%	35.37%
UNDEFINED - U		0	0	0.00%	0.00%
UNCLASSIFIED		0	0	0.00%	0.00%
TRANS/COMMUNICATIONS/UTILITIES	🟥	2144	4790193	3.62%	5.25%

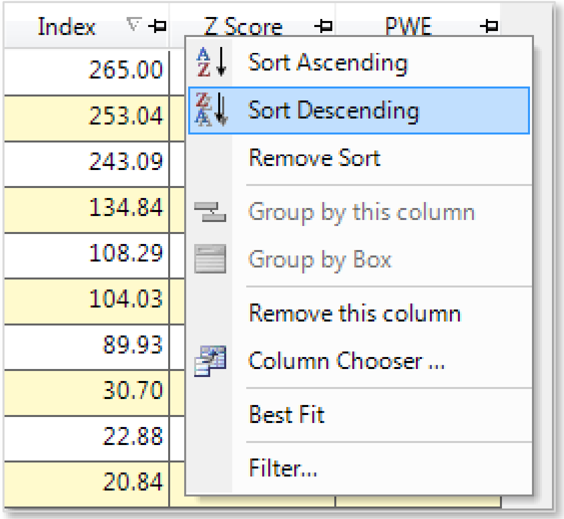


Visible	Column
☑	% of Analysis
☑	% of Base
☑	Index
☑	Z Score
☐	Probability
☐	Confidence %
☐	Ratio
☐	Zd Exp
☑	PWE
☑	Penetration
☐	Shade
☐	% of Analysis : % of Base
☐	Market Potential

Sort

This enables you to sort the rows of the profile report by this column. It is often useful to sort by Index – Descending to show from best to worst.

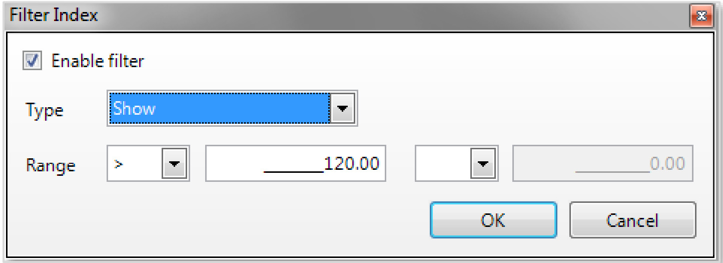
- Right click on the **Index** column heading and select **Sort Descending**



Filter

This enables you to filter the rows by the value of this column. It is often useful to filter to show only high Index + high z-score rows. Note that you can flick the filter on and off using the Filter button.

- Right click on the Index column and select **Filter**
- Enable the Filter and complete as per the screen shot opposite
- Click **OK** to see the results
- Use the ▽ button to toggle between the filter view and the full view
- Use the ✗ button to cancel the filter


Filter Index

☑ Enable filter

Type: Show

Range: > 120.00 0.00

Save Profile

You can save the profile by naming and dragging onto the File Explorer or My Favorites. It is also often useful to save as a Template (by dragging onto Templates) in which case just the variables and format of the profile are saved ready for you to reuse with a different selection.

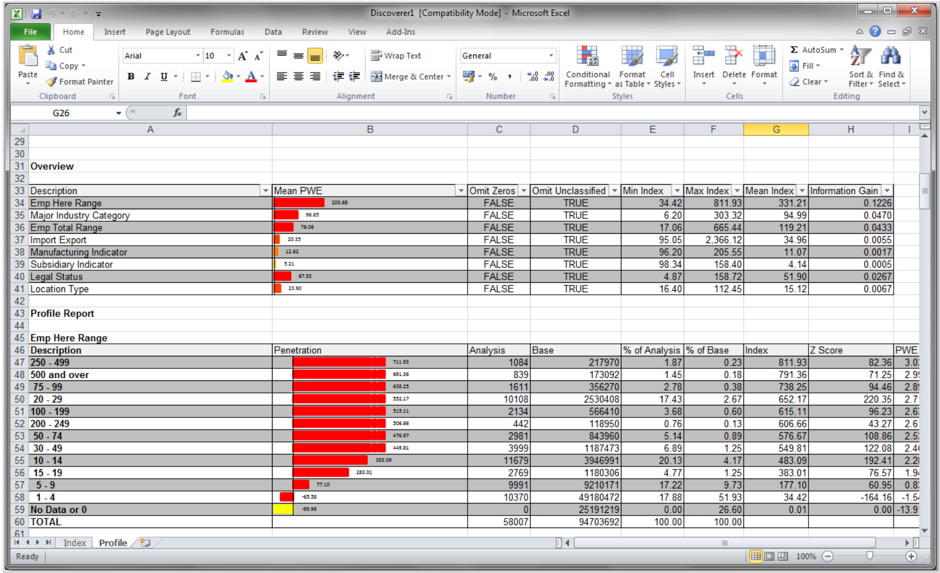
Report Profile

A Profile can be dragged into a Market Insight Report.

Output to MS Excel

If you have MS Excel 2003 or later installed you can transfer your profile into a workbook. This is a particularly useful for adding your own custom calculations or formatting inside an office document.

- ➢ Click on the **Transfer to Excel** button



Overview

Description	Mean PWE	Omit Zeros	Omit Unclassified	Min Index	Max Index	Mean Index	Information Gain
Emp Here Range	203.68	FALSE	TRUE	34.42	811.93	331.21	0.1226
Major Industry Category	96.65	FALSE	TRUE	6.20	303.32	94.99	0.0470
Emp Total Range	76.06	FALSE	TRUE	17.06	665.44	119.21	0.0433
Import Export	20.35	FALSE	TRUE	95.05	2,366.12	34.96	0.0055
Manufacturing Indicator	12.62	FALSE	TRUE	96.20	205.55	11.07	0.0017
Subsidiary Indicator	5.21	FALSE	TRUE	98.34	158.40	4.14	0.0005
Legal Status	67.53	FALSE	TRUE	4.87	158.72	51.90	0.0267
Location Type	23.90	FALSE	TRUE	16.40	112.45	15.12	0.0067

Profile Report

Emp Here Range

Description	Penetration	Analysis	Base	% of Analysis	% of Base	Index	Z Score	PWE
250 - 499	711.93	1084	217970	1.87	0.23	811.93	82.36	3.0
500 and over	691.36	839	173092	1.45	0.18	791.36	71.25	2.9
75 - 99	638.25	1611	356270	2.78	0.38	738.25	94.46	2.8
20 - 29	552.17	10108	2530408	17.43	2.67	652.17	220.35	2.7
100 - 199	515.11	2134	566410	3.68	0.60	615.11	96.23	2.6
200 - 249	506.66	442	118950	0.76	0.13	606.66	43.27	2.6
50 - 74	476.67	2981	843960	5.14	0.89	576.67	108.86	2.5
30 - 49	449.81	3999	1187473	6.89	1.25	549.81	122.08	2.4
10 - 14	383.09	11679	3946991	20.13	4.17	483.09	192.41	2.2
15 - 19	283.01	2769	1180306	4.77	1.25	383.01	76.57	1.9
5 - 9	77.10	9991	9210171	17.22	9.73	177.10	60.95	0.8
1 - 4	-65.58	10370	49180472	17.88	51.93	34.42	-164.16	-1.5
No Data or 0	-99.99	0	25191219	0.00	26.60	0.01	0.00	-13.9
TOTAL		58007	94703692	100.00	100.00			

Selecting from the Profile

It is possible to make a selection directly from the Profile report by highlighting the categories you want and dragging them onto the workspace.

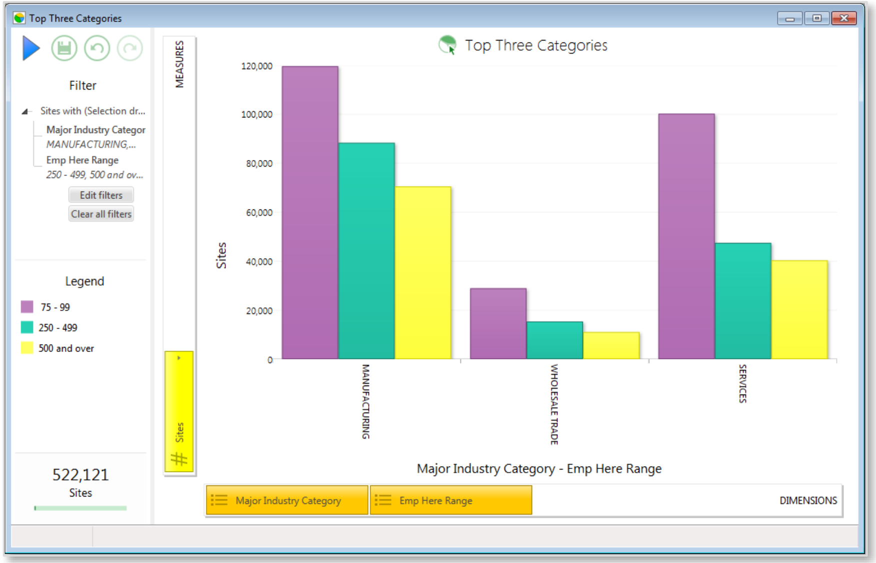
- Highlight the top three categories by **Penetration** for the **Major Industry Category** variable and drag onto the workspace
- Highlight the top three categories by **Penetration** for the **Emp Here Range** variable and drag onto the selection window created above

You now have a selection of the top sites who meet the Sites with Division 1 Spend profile in terms of the variables selected.

This method has its limitations in that you are only selecting sites that happen to fall into one of the categories from both variables selected.

A more accurate approach would be to create a total score for each site, based upon the PWE value calculated for each variable category used in the profile. These scores can then be ranked to find those sites that most closely match the analysis group.

To gain access to these scores you will need to create a virtual variable as described in the next section.



Selection shown as a Chart

Output PWE Scores as a Virtual Variable

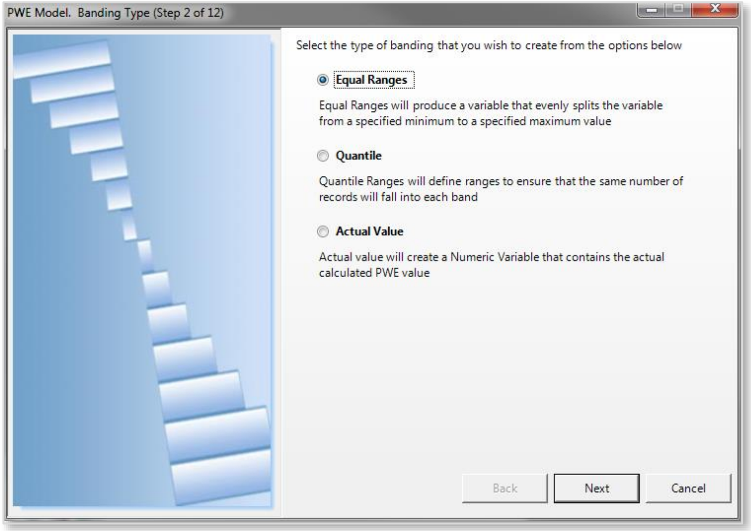
We will now take our Profile report and use the PWE Model wizard to score the database. The variable created will let us select people based on how well they fit the profile. The variable can also be used in a Model Report to compare models.

- Redisplay the **Sites with Division 1 Spend Report** if you have closed it
- Click on the **PWE Model** button
- Step 1 – This step requires you to drag on the source profile but because we have accessed the wizard from the profile window this step is missed
- Step 2 – Select the radio button next to **Equal Ranges** to define the banding output for the virtual variable, click **Next**

Quantiles are also a good choice for PWE bands as they segment the prospects into equal sized groups ranked by PWE score. Actual Value captures the actual score as a numeric rather than creating a banding.

- Step 3 – This step allows you to select the number of bands or band size and minimum and maximum banding values. Select the **Size of each Band** radio button and type a size of 1.00, click **Next**

This will allow you to view each individual score as a band and the number of people that have it. In this way you can select the very best prospects in a ranked order.



PWE Model. Banding Type (Step 2 of 12)

Select the type of banding that you wish to create from the options below

● Equal Ranges

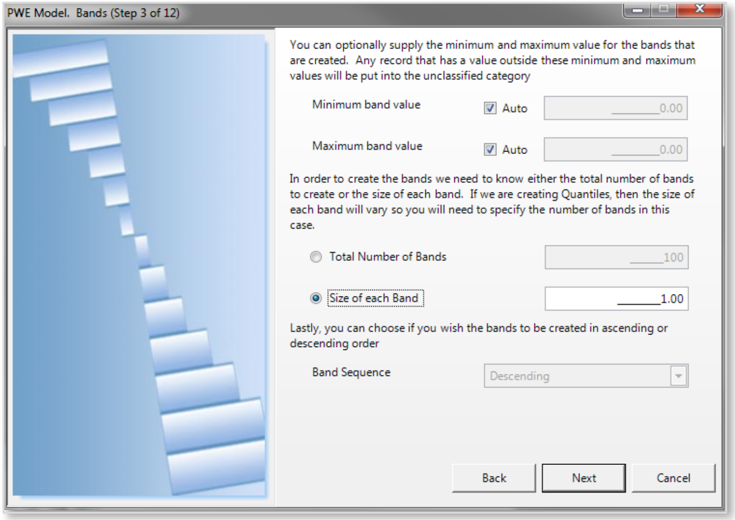
Equal Ranges will produce a variable that evenly splits the variable from a specified minimum to a specified maximum value

○ Quantile

Quantile Ranges will define ranges to ensure that the same number of records will fall into each band

○ Actual Value

Actual value will create a Numeric Variable that contains the actual calculated PWE value



PWE Model. Bands (Step 3 of 12)

You can optionally supply the minimum and maximum value for the bands that are created. Any record that has a value outside these minimum and maximum values will be put into the unclassified category

Minimum band value ☑ Auto 0.00

Maximum band value ☑ Auto 0.00

In order to create the bands we need to know either the total number of bands to create or the size of each band. If we are creating Quantiles, then the size of each band will vary so you will need to specify the number of bands in this case.

○ Total Number of Bands 100

● Size of each Band 1.00

Lastly, you can choose if you wish the bands to be created in ascending or descending order

Band Sequence Descending

- Step 4 – This step gives you the opportunity to create a **Holdout Sample** (see below). Leave the box unticked and click **Next**

You can specify a holdout sample that will be automatically removed from the records that are used to build the model. The holdout sample records will then automatically be used in any subsequent model reports.

In fact because the PWE method uses a linear combination of scores and the PWE category scores can be accurately estimated from a sample, the use of a holdout sample is not crucial.

- Step 5 – This step allows you to specify a selection of records to score. For example you may only wish to score prospects who have not already bought a product or those from a particular test region. Leave this section blank to score all the records, click **Next**

- Step 6 - Click or create the folder where the variable is to be created. Leave on **Others**, click **Next**

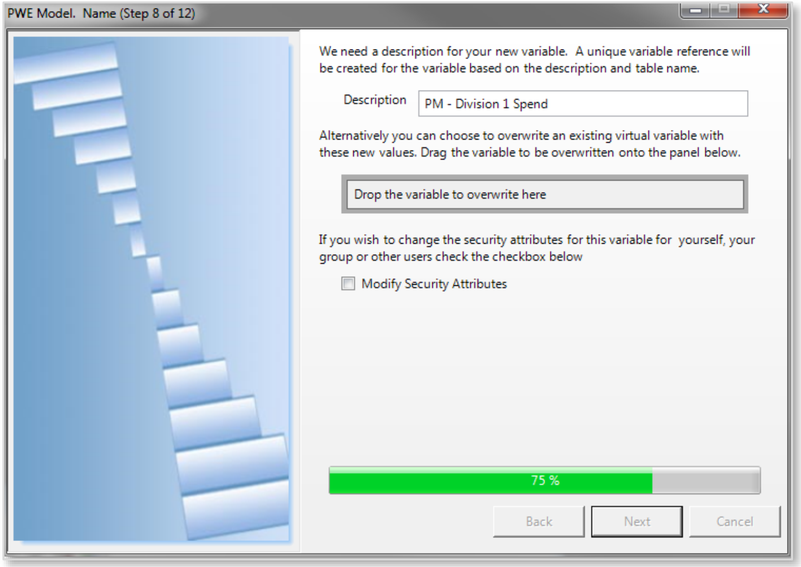
- Step 7 - Enter optional notes, click **Next**

- Step 8 - Enter the description **PM – Division 1 Spend**, click **Next**

- Step 9 - Security feature to control access and use of the variable (not applicable on local system) click **Next**

- Step 11 – Click **Finish**

This virtual variable can also be used in the Model Report tool to examine how well the profile performs at selecting Customers with Division 1 Spend.



PWE Model. Name (Step 8 of 12)

We need a description for your new variable. A unique variable reference will be created for the variable based on the description and table name.

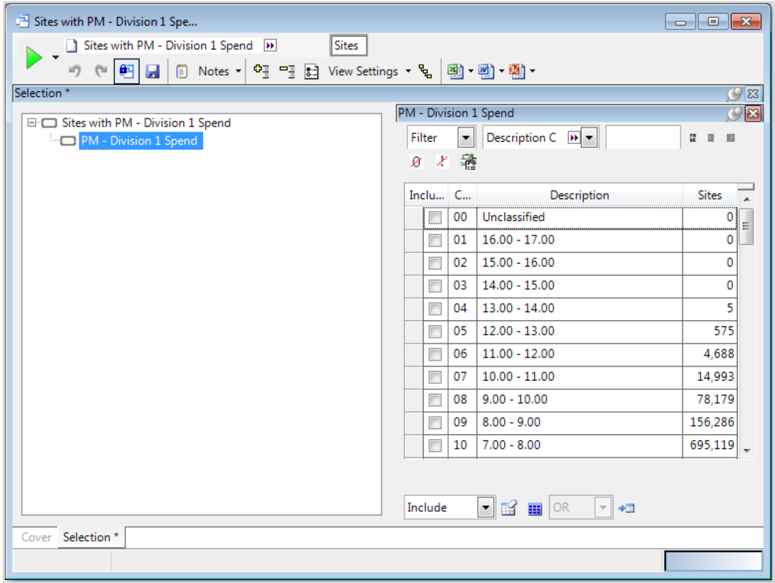
Description: PM - Division 1 Spend

Alternatively you can choose to overwrite an existing virtual variable with these new values. Drag the variable to be overwritten onto the panel below.

Drop the variable to overwrite here

If you wish to change the security attributes for this variable for yourself, your group or other users check the checkbox below

□ Modify Security Attributes



Sites with PM - Division 1 Spe...

Sites with PM - Division 1 Spend
PM - Division 1 Spend

PM - Division 1 Spend

Filter: Description C

Inclu...	C...	Description	Sites
□	00	Unclassified	0
□	01	16.00 - 17.00	0
□	02	15.00 - 16.00	0
□	03	14.00 - 15.00	0
□	04	13.00 - 14.00	5
□	05	12.00 - 13.00	575
□	06	11.00 - 12.00	4,688
□	07	10.00 - 11.00	14,993
□	08	9.00 - 10.00	78,179
□	09	8.00 - 9.00	156,286
□	10	7.00 - 8.00	695,119

The Limitations of Profiling

The most likely problem with profiling is that you do not distinguish between true characteristics of the set of records you profile and correlations that are present because of the way the data has been collected or processed.

Profiling always needs a set of records to start from – it is not a self-starting process. In marketing terms, you must start the process by making an offer to some potential customers before you can profile the responses and build a model to use in the next round of marketing.

Profiling is a regressive technique when used with modeling and selection – it will always look for a subset of characteristics that apply to the data you analyze. Beware that you do not end up spiraling into a smaller and smaller set of prospects and lose sight of the bigger picture. Some marketers pull other groups (for example a random selection of low scoring records) into their promotions in order to introduce "new blood" into the profiling and marketing process.

Profiling is limited by the base universe data and by the available variables. For example, the most useful variable might be "how likely are you to change bank", but this variable is unlikely to be available.

Profiling is univariate – it considers each predictor variable in isolation. This has benefits in terms of simplicity and transparency but ignores interactions between variables. Interactions between variables can be best explored using Decision Trees.

Choosing Variables for the Model

The profile report guides your choice of variables for the model. In general you should include the variables that indicate the greatest predictive power through a high Mean PWE.

However PWE is relatively robust to missing values in variables. If a value is missing for a particular record then the evidence accumulated is zero (i.e. no evidence for or against). This means that it is practical to include variables that are only partially populated although be aware that the population distribution of the included variables will influence the result.

Beware of relationships in the data. For example the following fields are closely correlated: City, Zip Code and State are all representations of the location of the business.

You should in general only include one of each of these sets of fields.

N.B. If you use multiple correlated fields, you multiply the strength of that information in the PWE calculation and bias the result. In some examples, this improves the model (for example if only one of the fields is usually populated per record), but should be treated with caution.

Introduction to Decision Tree

One of the key mechanisms of Market Insight is the ability to identify the significant characteristics of a selected group of records.

Understanding the characteristics of your current customers and your strengths and weaknesses across market sectors is a prerequisite for any marketing activity. Whether the marketing strategy is to expand to new sectors or up-sell in existing sectors it is necessary to consider your current sector performance and the remaining market potential before embarking on list acquisition and expensive direct marketing promotions.

The Decision Tree module complements the PWE (Predictive Weight of Evidence) model within Profiler, providing a very visual and interactive way of modelling a selection of records.

A Decision Tree enables you to define a selection of customers that you are interested in and then to identify characteristics that are typical of these people. The characteristics are **built into rules as opposed to a set of scores within Profiler**, which can then be used to make selections, perhaps for a marketing campaign.

There are two main reasons why you might want to do this:

To ***describe*** your current customers so that you have a better understanding of the different types who spend with a Division of your organization.

To ***predict*** which companies should be targeted by a marketing campaign related to spending with a Division of your organization.

This could either be identifying the best prospects from a prospect database, or identifying companies on your existing database who don’t currently with a particular division, but have similar characteristics to those companies that do.

There are two main statistical algorithms available to create your Decision Tree, Predictive Weight of Evidence (PWE) and CHAID:

The Binary PWE algorithm creates two-way splits by separating categories with a high proportion of the target group from those with a low proportion. This process is very fast even with lots of categories and always results in a split.

The CHAID algorithm, based on Chi-square statistical tests, creates multi-way splits by merging together categories that have a similar proportion of the target group. This merging process is slow (several seconds) when there are lots of categories (e.g. thousands) and is not always able to produce a split.

A Decision Tree Worked Example

Decision Tree compares one group of records with another to identify the characteristics that differentiate them. In this example we will use the Customers with Division 1 Spend compared with the wider universe.

Decision Tree detects characteristics by identifying results that occur in your selected data and which are highly unlikely to occur in a truly random sample of data.

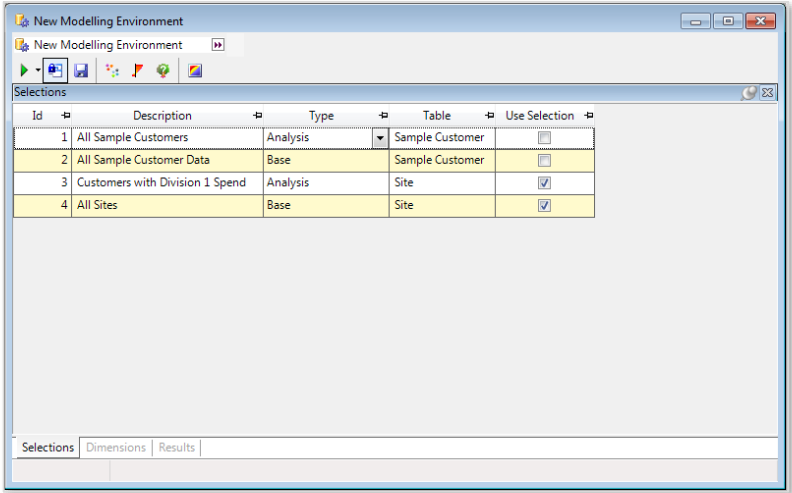
Setting up within the Modelling Environment

- Redisplay from your Private folder the **New Modelling Environment** window on the workspace if it has been closed

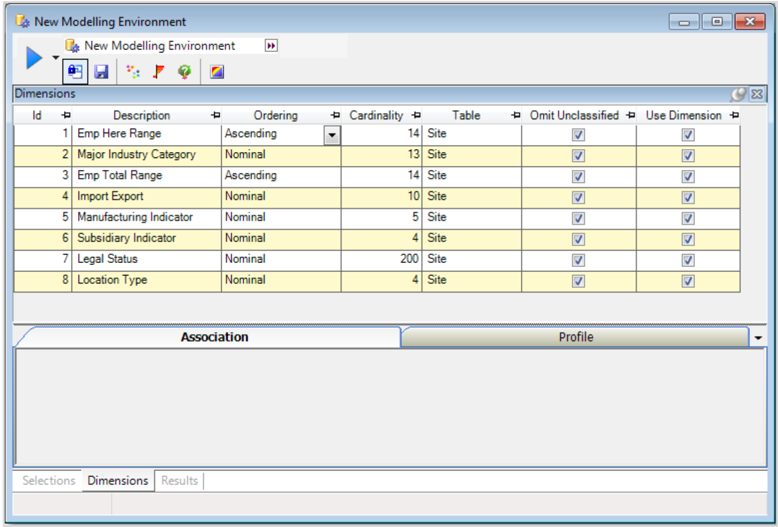
We are going to use the same settings as we did for the Profile so nothing needs to be changed on this occasion.

- Click the **Decision Tree** button to take you into the Decision Tree window

- Click the **Build** button to create the Decision Tree



Id	Description	Type	Table	Use Selection
1	All Sample Customers	Analysis	Sample Customer	□
2	All Sample Customer Data	Base	Sample Customer	□
3	Customers with Division 1 Spend	Analysis	Site	☑
4	All Sites	Base	Site	☑



Id	Description	Ordering	Cardinality	Table	Omit Unclassified	Use Dimension
1	Emp Here Range	Ascending	14	Site	☑	☑
2	Major Industry Category	Nominal	13	Site	☑	☑
3	Emp Total Range	Ascending	14	Site	☑	☑
4	Import Export	Nominal	10	Site	☑	☑
5	Manufacturing Indicator	Nominal	5	Site	☑	☑
6	Subsidiary Indicator	Nominal	4	Site	☑	☑
7	Legal Status	Nominal	200	Site	☑	☑
8	Location Type	Nominal	4	Site	☑	☑

Setting up outside the Modelling Environment

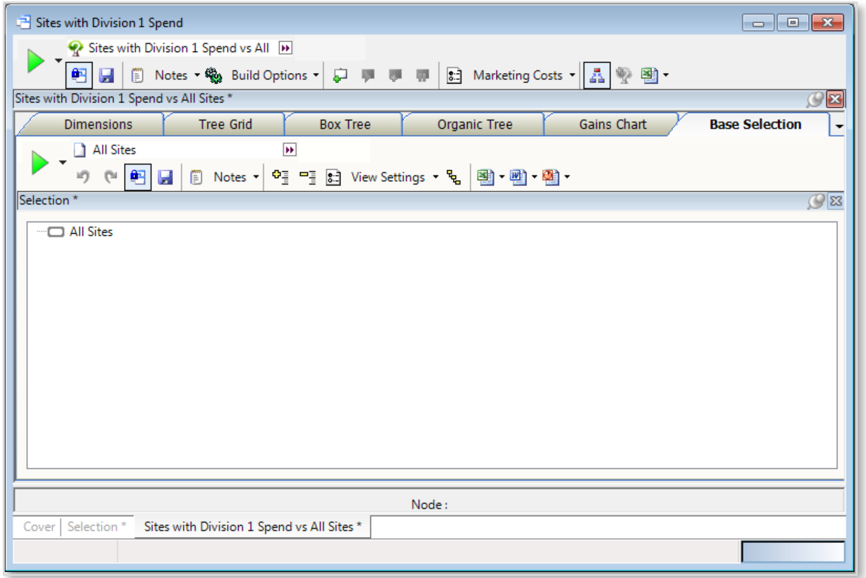
- ➢ Create a **Site** selection of **Division 1 Spend** of **>0**. Rename the selection window **Sites with Division 1 Spend**

- ➢ From the **Toolbox** window drag and drop the **Decision Tree** tool onto the open selection page

The target records (Sites with Division 1 Spend) are referred to as the **Analysis** selection.

The data which we will be comparing with is referred to as the **Base** selection.

- ➢ Click on the **Base Selection** tab (we will leave this blank so it will use the entire database by default as the base selection)

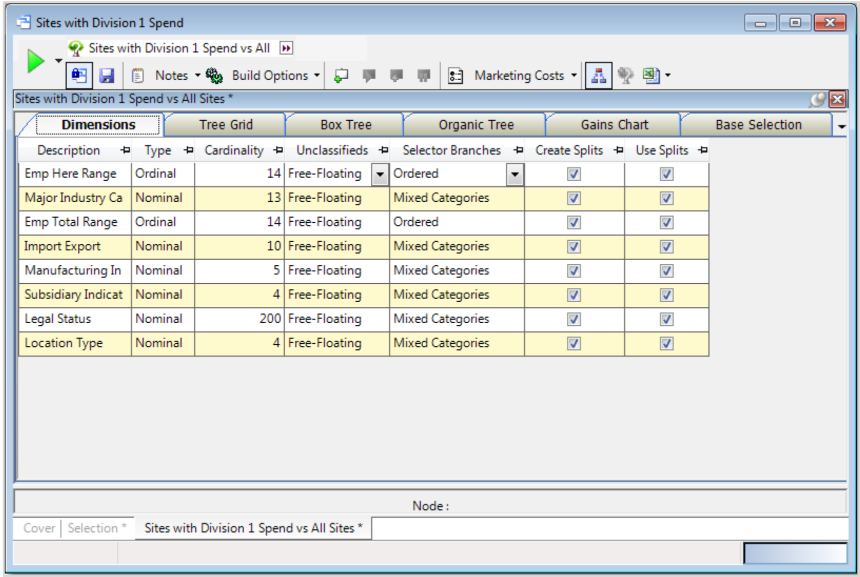


Having determined the Analysis and Base group we now need to specify the Variables for inclusion in the Decision Tree.

- ➢ Ensure the **Dimensions** tab is upper most by clicking on it

- ➢ Drag on the following variables – **Emp Here Range, Major Industry Category, Emp Total Range, Import Export, Manufacturing Indicator, Subsidiary Indicator, Legal Status** and **Location Type**

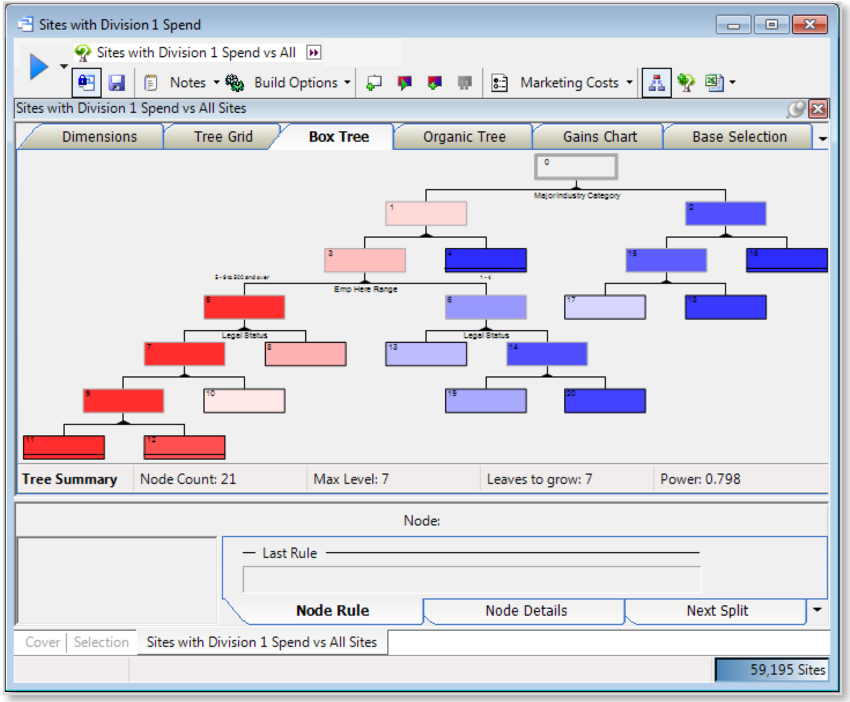
- ➢ Click the ▶ **Build** button to run the Decision Tree


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Build Options
Marketing Costs
Sites with Division 1 Spend vs All Sites *
Dimensions
Tree Grid
Box Tree
Organic Tree
Gains Chart
Base Selection
Node :
Cover
Selection *
Sites with Division 1 Spend vs All Sites *

Description	Type	Cardinality	Unclassifieds	Selector Branches	Create Splits	Use Splits
Emp Here Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Major Industry Ca	Nominal	13	Free-Floating	Mixed Categories	☑	☑
Emp Total Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Import Export	Nominal	10	Free-Floating	Mixed Categories	☑	☑
Manufacturing In	Nominal	5	Free-Floating	Mixed Categories	☑	☑
Subsidiary Indicat	Nominal	4	Free-Floating	Mixed Categories	☑	☑
Legal Status	Nominal	200	Free-Floating	Mixed Categories	☑	☑
Location Type	Nominal	4	Free-Floating	Mixed Categories	☑	☑

Reviewing the Decision Tree Results

We have already discussed two of the tabs, Dimensions and Base Selection. The remaining tabs are split into two with the top half (Tree Panel) displaying alternate views of the same tree. The bottom half (Focus Panel) displays details relating to a single node. We will discuss this later. To make more sense of the information generated for us we can start by looking at the Box Tree view more closely. We will look at each of these tabs in more depth later in the course.



Tree Summary | Node Count: 21 | Max Level: 7 | Leaves to grow: 7 | Power: 0.798

59,195 Sites

Box Tree

In this view we start with a Root node which is then split by the Dimension (variable) which has been calculated to be the best indicator of being in our target group i.e. Customers with Division 1 Spend.

In this example Income has been identified as the best predictor and the node has been split grouping the categories within the variable accordingly.

The color of a node indicates the increase or decrease in the Analysis % of the node compared to the root node.

Red relates to those people in the **analysis** section and is determined by a Gain of **more** than 1 – the more intense the red, the **higher** the proportion of people in the analysis selection.

Blue relates to those people **not** in the analysis section and is determined by a Gain of **less** than 1 – the more intense the blue, the **lower** the proportion of people in the analysis selection.

Organic Tree

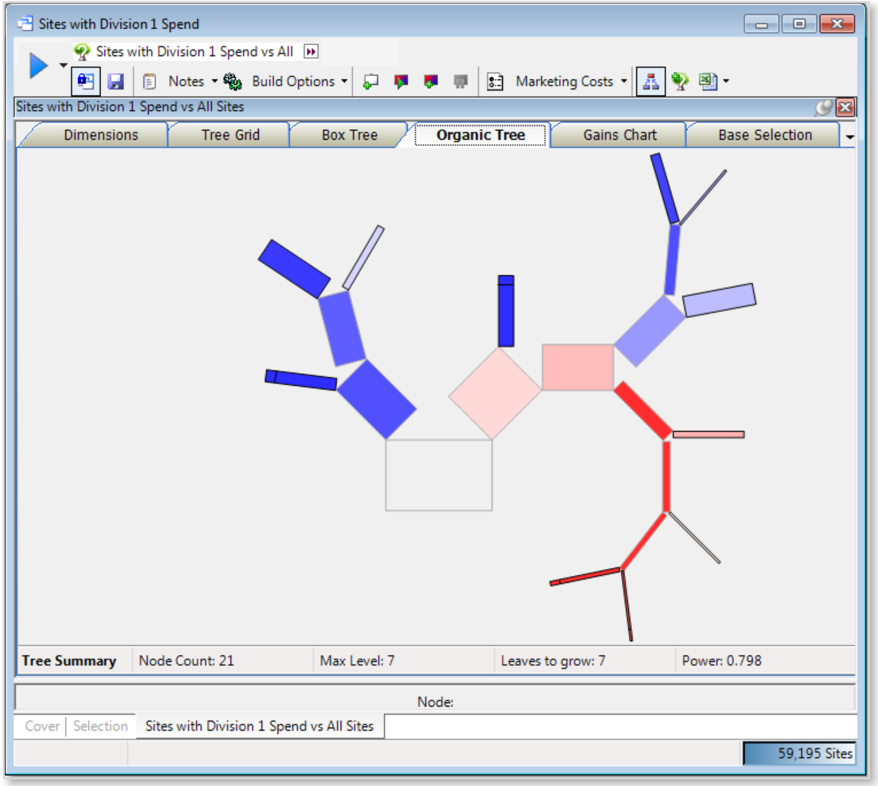
The Organic Tree provides an overview of the statistical relationship between nodes. From the example shown opposite you can see how the branches represent the exact same nodes as shown in the Box Tree view.

The major differences we can see here are the **direction** in which the braches grow and the variation in their **thickness**.

- The width of each branch is proportional to the number of people in that node.
- The angle of the branch indicates how significant the difference in Analysis% within the branch is compared to its parent.
  - A branch almost in line with its parent does not differ that significantly in Analysis%
  - A branch sticking out at a large angle shows a significant difference in Analysis%

The color reference is the same as described under the Box Tree section. As such you will notice Red branches tend to bend to the right and Blue branches tend to bend to the left.

In this example, the node selected to split is based upon the default under Build Options – Best Split. (See pg. 49 and the Help function for more details)


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Build Options
Marketing Costs
Sites with Division 1 Spend vs All Sites
Dimensions
Tree Grid
Box Tree
Organic Tree
Gains Chart
Base Selection
Cover
Selection
Sites with Division 1 Spend vs All Sites

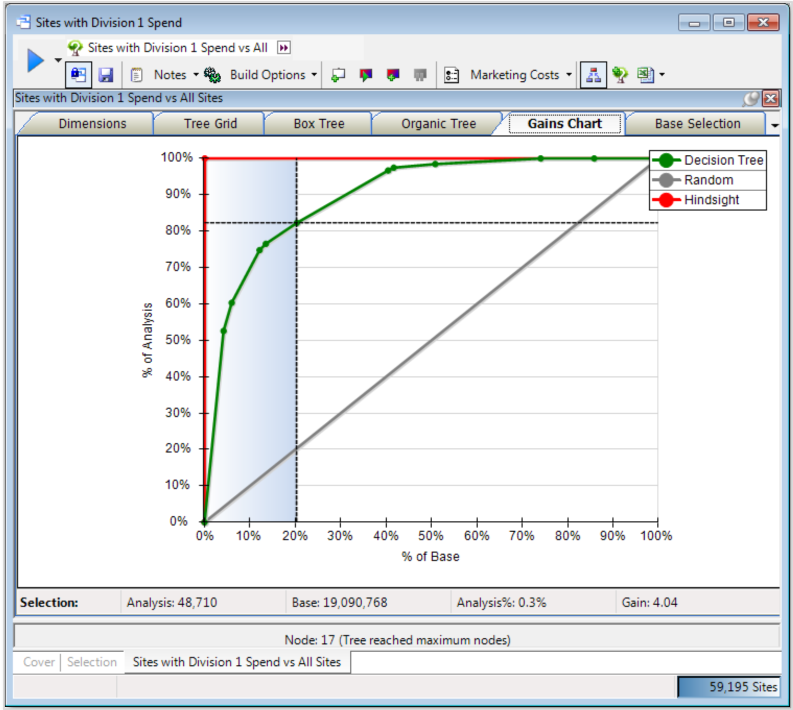
Tree Summary | Node Count: 21 | Max Level: 7 | Leaves to grow: 7 | Power: 0.798

Node:

59,195 Sites

Gains Chart

This chart plots the cumulative %Analysis and %Base to give the classic representation of the power of the model.



Selection: Analysis: 48,710 Base: 19,090,768 Analysis%: 0.3% Gain: 4.04

Node: 17 (Tree reached maximum nodes)

The Gains Chart shows the proportion of data used across the bottom and the proportion of Analysis Records found up the side.

- Each point represents a leaf node. They are sorted from best to worst (by descending gain). .
- For comparison the grey line (Random) shows the result of a random choice – there is a direct and linear relationship between the number of records selected and the number of analysis records found.
- For comparison the red line (Hindsight) shows the perfect result using our current knowledge of the Analysis set to plot the best outcome the model could possibly achieve.
- The green line (Decision Tree) shows the outcome achieved by the model. The steeper the green line and the closer it remains to the red line the better. However if it is too close to the red line we might suspect that the model is "too good to be true" i.e. over-fitted.
- The area between the green and grey lines is an indication of the power of the model – the larger the area, the higher the power and the stronger the model.

Tree Grid

The Tree Grid displays the numeric details behind each node. This view can also be used to see the implications of the Marketing Costs.

Hierarchical

Enables you to drill down, expanding and collapsing nodes to see the structure of the tree. This is the default view.

Flat ☰

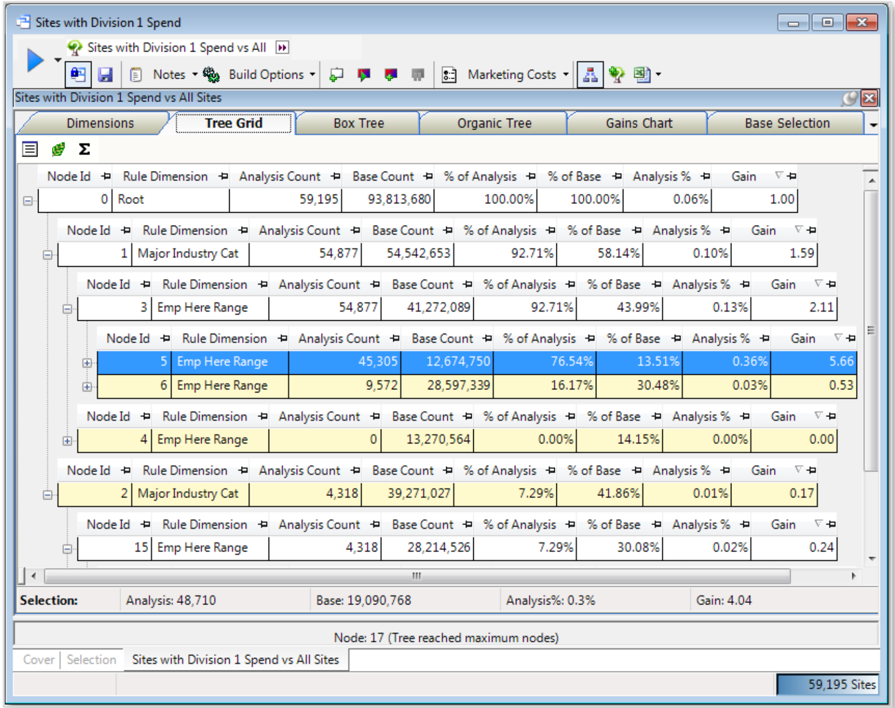
Displays all nodes from all levels in a single table. This is useful for sorting the nodes by a particular column and for making selections at any level.

Non-Cumulative Gains Table

This just displays the leaf nodes from whatever level of the tree in a single table. This is useful for sorting the nodes by a particular column and for making selections of leaves.

Cumulative Gains Table Σ

Displays just leaf nodes from whatever level of the tree in a single table. By default the table is sorted by descending gain. This is also the order in which the counts and percentages are cumulated. It is useful in conjunction with the Gains Chart to see the total number of records selected in the top N nodes.


Sites with Division 1 Spend
Sites with Division 1 Spend vs All
Notes
Build Options
Marketing Costs
Sites with Division 1 Spend vs All Sites
Dimensions
Tree Grid
Box Tree
Organic Tree
Gains Chart
Base Selection

Node Id	Rule Dimension	Analysis Count	Base Count	% of Analysis	% of Base	Analysis %	Gain
0	Root	59,195	93,813,680	100.00%	100.00%	0.06%	1.00
1	Major Industry Cat	54,877	54,542,653	92.71%	58.14%	0.10%	1.59
3	Emp Here Range	54,877	41,272,089	92.71%	43.99%	0.13%	2.11
5	Emp Here Range	45,305	12,674,750	76.54%	13.51%	0.36%	5.66
6	Emp Here Range	9,572	28,597,339	16.17%	30.48%	0.03%	0.53
4	Emp Here Range	0	13,270,564	0.00%	14.15%	0.00%	0.00
2	Major Industry Cat	4,318	39,271,027	7.29%	41.86%	0.01%	0.17
15	Emp Here Range	4,318	28,214,526	7.29%	30.08%	0.02%	0.24

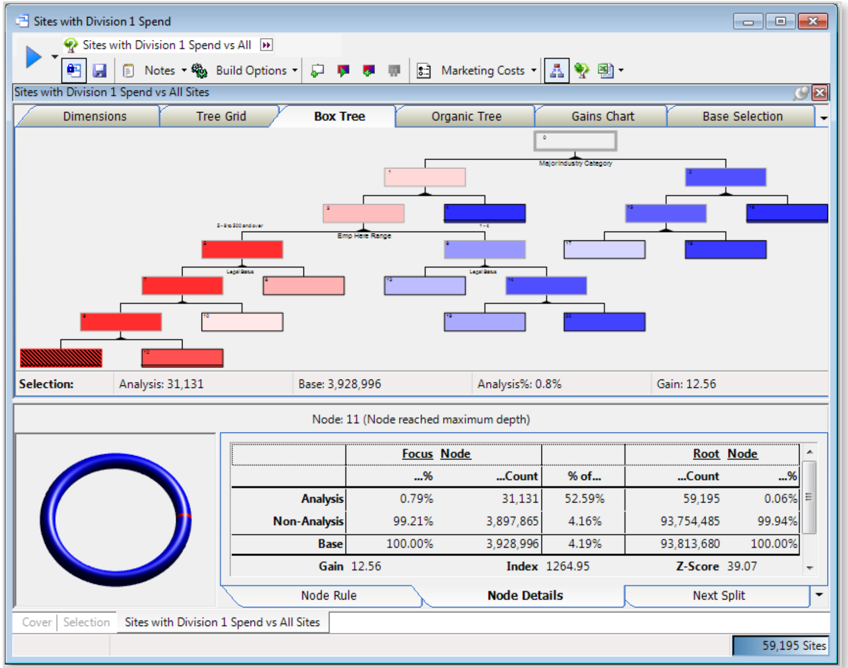
Selection: Analysis: 48,710 Base: 19,090,768 Analysis%: 0.3% Gain: 4.04

Node: 17 (Tree reached maximum nodes)

N.B. Remember you can right click on any of the headings and select the Column Chooser option to display further details.

Focus Panel

The Focus Panel is made up of 4 areas; 3 of which are available through the tabs, the 4th being the ever present graphical display. All the details held here relate to the focus node i.e. the one highlighted with a red border in the Tree Panel.



Selection: Analysis: 31,131 Base: 3,928,996 Analysis%: 0.8% Gain: 12.56

Node: 11 (Node reached maximum depth)

	Focus Node			Root Node	
	...%	...Count	% of...	...Count	...%
Analysis	0.79%	31,131	52.59%	59,195	0.06%
Non-Analysis	99.21%	3,897,865	4.16%	93,754,485	99.94%
Base	100.00%	3,928,996	4.19%	93,813,680	100.00%
Gain	12.56	Index	1264.95	Z-Score	39.07

The Torus (Graphical display)

Sometimes referred to as the donut this image indicates information through color and size. The red segment represents the Analysis% of the node. The girth, or fatness of the torus represents how many of the people in the root node are contained within this node i.e. % of the base.

Node Rule

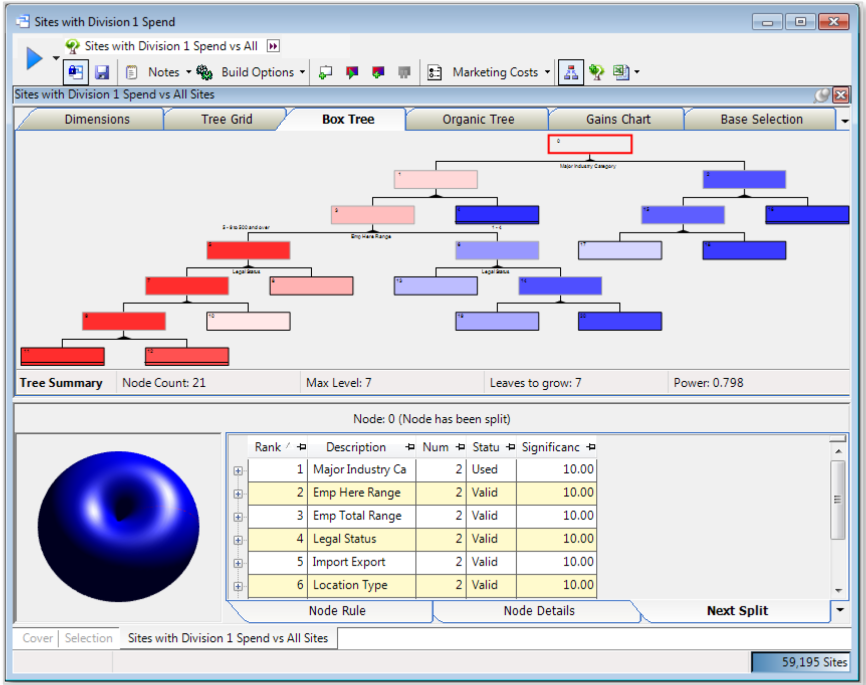
This tab displays the selection rule behind a single node in terms of the dimensions used to build the tree. You can view the last rule used, the full set of rules used or a combined set of rules where it is possible to simplify the display.

Node Details

This tab displays a selection of statistics that are available in the Tree Grid, but just relating to the focus node.

Next Split

This tab allows you to investigate the decisions made at any particular split. As each node is split, each of the dimensions is examined and essentially a profile like report is generated showing the analysis and base counts within this node for each value of the dimension e.g. major Industry Category.



Sites with Division 1 Spend

Sites with Division 1 Spend vs All

Sites with Division 1 Spend vs All Sites

Tree Summary	Node Count: 21	Max Level: 7	Leaves to grow: 7	Power: 0.798

Node: 0 (Node has been split)

Rank	Description	Num	Statu	Significanc
1	Major Industry Ca	2	Used	10.00
2	Emp Here Range	2	Valid	10.00
3	Emp Total Range	2	Valid	10.00
4	Legal Status	2	Valid	10.00
5	Import Export	2	Valid	10.00
6	Location Type	2	Valid	10.00

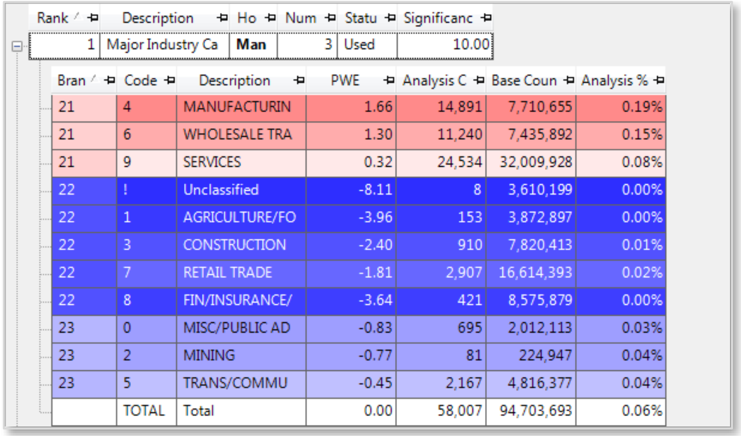
Manual Split

You may wish to override the split generated for you and define your own branches:

- Highlight the **Root Node** and then click on the **Next Split** tab
- Click on the + sign next to **Major Industry Category** to reveal the color coded generated split
- Click on the column headers to sort the table (e.g. by Branch)

You may feel that the categories may be of more interest split differently e.g. have 3 Branches:

- Highlight the rows **with codes 0, 2 & 5** right click and select **Define Branch**. Click **Yes**
- Right click on the **Major Industry Category** row and select **Use Valid Split: Major Industry Category**



Rank	Description	Ho	Num	Statu	Significanc
1	Major Industry Ca	**Man**	3	Used	10.00

Bran	Code	Description	PWE	Analysis C	Base Coun	Analysis %
21	4	MANUFACTURIN	1.66	14,891	7,710,655	0.19%
21	6	WHOLESALE TRA	1.30	11,240	7,435,892	0.15%
21	9	SERVICES	0.32	24,534	32,009,928	0.08%
22	!	Unclassified	-8.11	8	3,610,199	0.00%
22	1	AGRICULTURE/FO	-3.96	153	3,872,897	0.00%
22	3	CONSTRUCTION	-2.40	910	7,820,413	0.01%
22	7	RETAIL TRADE	-1.81	2,907	16,614,393	0.02%
22	8	FIN/INSURANCE/	-3.64	421	8,575,879	0.00%
23	0	MISC/PUBLIC AD	-0.83	695	2,012,113	0.03%
23	2	MINING	-0.77	81	224,947	0.04%
23	5	TRANS/COMMU	-0.45	2,167	4,816,377	0.04%
	TOTAL	Total	0.00	58,007	94,703,693	0.06%

To reset to the default split:

- Right click on the **Major Industry Category** row and select **Reset to Default Split**
- Click the **Continue Build** button to continue the build

Making a Selection from the Decision Tree

Having built the Decision Tree the best customers/prospects are in the nodes which are most red.

By clicking on the Box Tree and Organic Tree you can see there are 3 end nodes with a red color.

- Click on the **Tree Grid** tab

- Click on the **Leaves Only** 🍃 button

If you scroll across to the Gain column you will see the first three nodes all have a Gain greater than 1 i.e. they have a higher % of the Analysis group than the Root node.

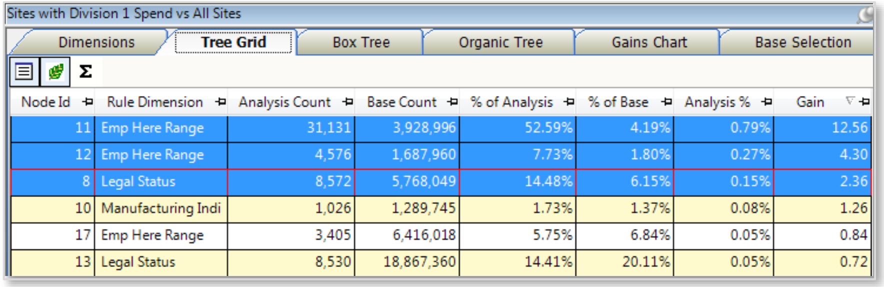
- Highlight the first 3 rows of the **Tree Grid**

If you look at the Box Tree and Organic Tree you will see a number of the red nodes are now highlighted.

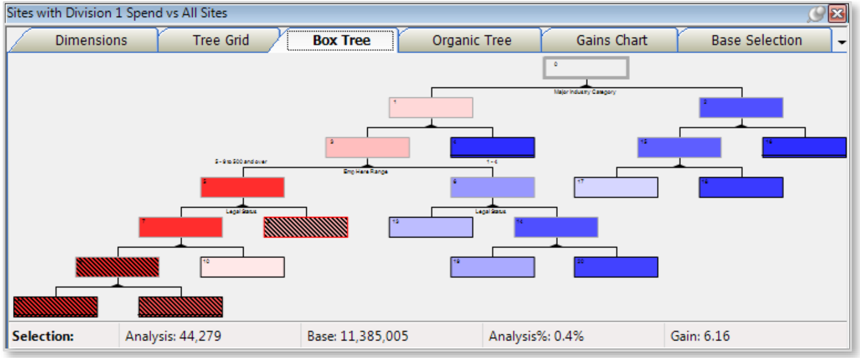
- Click on the **Gains Chart** tab

Each leaf node is plotted along the green Decision Tree line. The 3rd node has been selected to shade the area of the graph to show that, using this analysis only 13% of the Base would need to be contacted to find 75% of the people in the Analysis group.

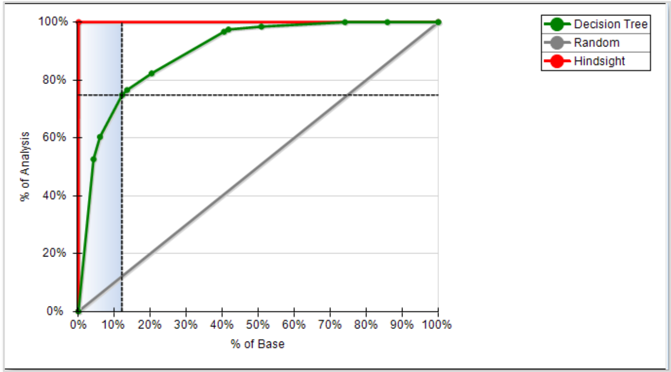
You can drag off from the diagram to make a Selection from the highlighted nodes. The default is to simplify this Selection as much as possible. In the above example, this would result in a Selection that gives the rules for nodes 8 and 9 (note that node 9 is the parent of the highlighted nodes 11 and 12).



Node Id	Rule Dimension	Analysis Count	Base Count	% of Analysis	% of Base	Analysis %	Gain
11	Emp Here Range	31,131	3,928,996	52.59%	4.19%	0.79%	12.56
12	Emp Here Range	4,576	1,687,960	7.73%	1.80%	0.27%	4.30
8	Legal Status	8,572	5,768,049	14.48%	6.15%	0.15%	2.36
10	Manufacturing Indi	1,026	1,289,745	1.73%	1.37%	0.08%	1.26
17	Emp Here Range	3,405	6,416,018	5.75%	6.84%	0.05%	0.84
13	Legal Status	8,530	18,867,360	14.41%	20.11%	0.05%	0.72



Selection: Analysis: 44,279 Base: 11,385,005 Analysis%: 0.4% Gain: 6.16



To identify the individuals behind these nodes we need to display them in a selection window.

- ➢ Drag the first 3 rows of the **Tree Grid** onto the workspace (You may want to turn off the **Simplify Query** option)
- ➢ Save as **Best Fit**

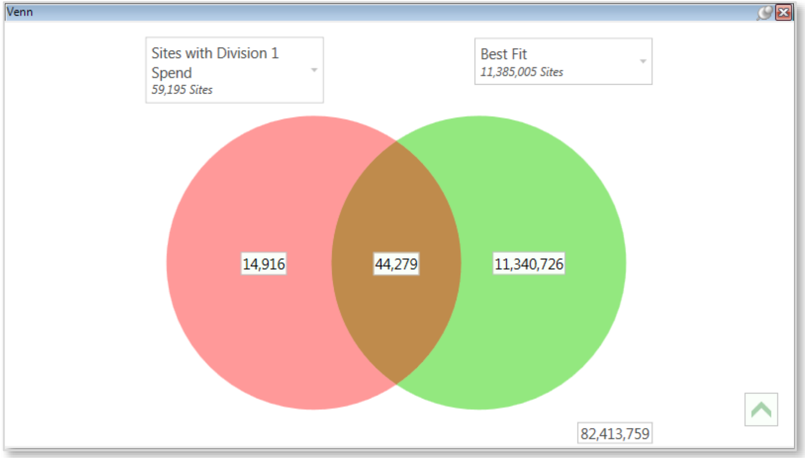
To identify the individuals who meet the rules, but yet to become a Division 1 Spender we can use the Venn tool.

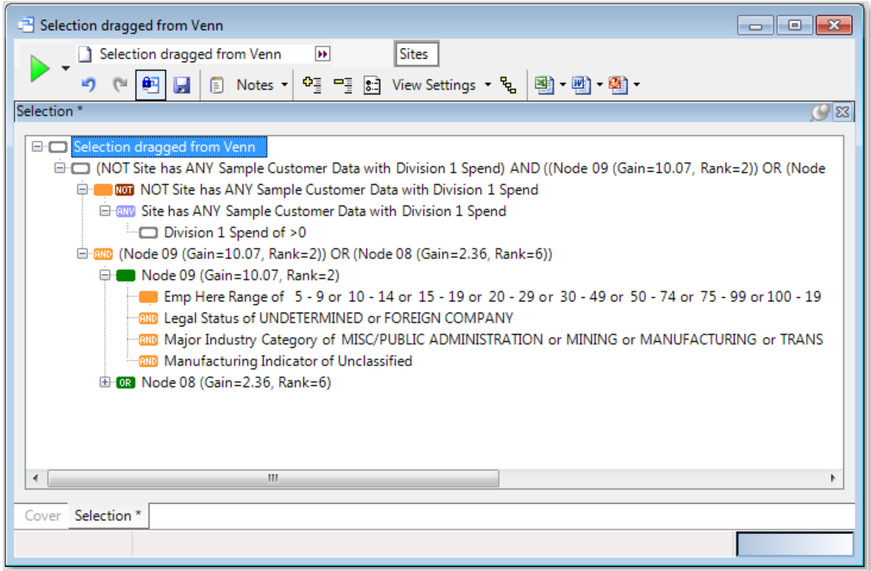
- ➢ Drag a **Venn** from the **Toolbox** onto the workspace
- ➢ Add to the **Venn** the original **Sites with Division 1 Spend** selection followed by the **Best Fit** selection
- ➢ Click on the **Build** button

The green segment represents all those individuals who meet the rules we have identified in the Decision Tree but which have yet to become a Division 1 Spender.

- ➢ Drag out the green segment onto the workspace

We could now save this selection of sites and further restrict the criteria with other tools until we have a list ready to purchase.





Selection dragged from Venn
(NOT Site has ANY Sample Customer Data with Division 1 Spend) AND ((Node 09 (Gain=10.07, Rank=2)) OR (Node
NOT NOT Site has ANY Sample Customer Data with Division 1 Spend
ANY Site has ANY Sample Customer Data with Division 1 Spend
Division 1 Spend of >0
AND (Node 09 (Gain=10.07, Rank=2)) OR (Node 08 (Gain=2.36, Rank=6))
Node 09 (Gain=10.07, Rank=2)
Emp Here Range of 5 - 9 or 10 - 14 or 15 - 19 or 20 - 29 or 30 - 49 or 50 - 74 or 75 - 99 or 100 - 19
AND Legal Status of UNDETERMINED or FOREIGN COMPANY
AND Major Industry Category of MISC/PUBLIC ADMINISTRATION or MINING or MANUFACTURING or TRANS
AND Manufacturing Indicator of Unclassified
OR Node 08 (Gain=2.36, Rank=6)

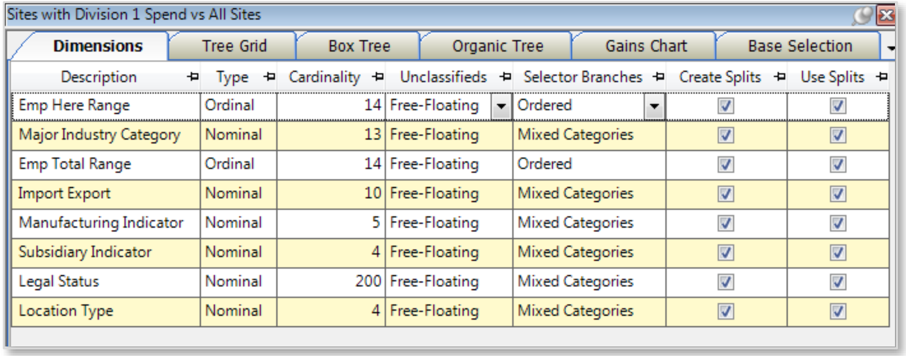
Reviewing the Options

Dimensions

When including variables to be used in the Decision Tree you have a number of options when dealing with Unclassified categories and the way in which branches are created.

Unclassified can be treated as:

Free Floating	where they can be grouped with any other category or on their own.
Low	where they are grouped with the lowest category or on their own.
High	where they are grouped with the highest category or on their own.
Either End	where they can be grouped with the lowest or highest categories or on their own.
Keep Separate	where they are forced into a separate node if chosen.
Omit	where they are removed from the tree at the point that the variable is chosen.



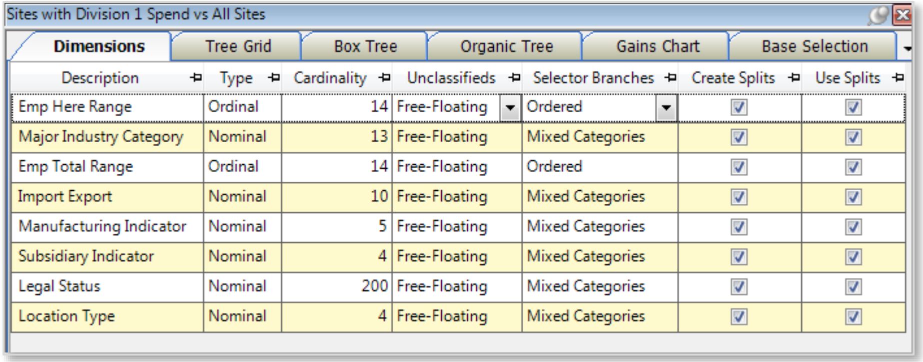
Description	Type	Cardinality	Unclassifieds	Selector Branches	Create Splits	Use Splits
Emp Here Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Major Industry Category	Nominal	13	Free-Floating	Mixed Categories	☑	☑
Emp Total Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Import Export	Nominal	10	Free-Floating	Mixed Categories	☑	☑
Manufacturing Indicator	Nominal	5	Free-Floating	Mixed Categories	☑	☑
Subsidiary Indicator	Nominal	4	Free-Floating	Mixed Categories	☑	☑
Legal Status	Nominal	200	Free-Floating	Mixed Categories	☑	☑
Location Type	Nominal	4	Free-Floating	Mixed Categories	☑	☑

Selector Branches can be handled as follows:

Mixed Categories	allows any combination of categories in the branch.
Ordered	allows each branch to only contain consecutive categories.
Cyclic	allows each branch to only contain consecutive categories, but also allows lowest and highest categories to be joined.

Create/Use Splits can be used as follows:

Neither Create Split nor Use Split	Counts will be created for all dimensions that are dragged on to the dimensions tab. Categories will not be combined to make splits. You can manually group categories into branches and then select to use the split. Using this option saves processing time. Creating CHAID splits for high cardinality variables can take considerable time.
Create Split	Categories are combined to make splits, but the split will not be selected automatically for use in creating child nodes. You can manually select a split that has been created by right clicking on the split.
Create Split and Use Split	Categories are combined to make splits, and the split can be selected automatically for use in creating child nodes.



Description	Type	Cardinality	Unclassifieds	Selector Branches	Create Splits	Use Splits
Emp Here Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Major Industry Category	Nominal	13	Free-Floating	Mixed Categories	☑	☑
Emp Total Range	Ordinal	14	Free-Floating	Ordered	☑	☑
Import Export	Nominal	10	Free-Floating	Mixed Categories	☑	☑
Manufacturing Indicator	Nominal	5	Free-Floating	Mixed Categories	☑	☑
Subsidiary Indicator	Nominal	4	Free-Floating	Mixed Categories	☑	☑
Legal Status	Nominal	200	Free-Floating	Mixed Categories	☑	☑
Location Type	Nominal	4	Free-Floating	Mixed Categories	☑	☑

Build Options

As part of the Build Options we have a series of Stopping Conditions that dictate how large the tree will grow.

- Click on the **Build Options** button

Maximum Nodes: The tree will be allowed to grow further while it contains less than the number of nodes specified by the *Maximum Nodes* stopping condition. Since the tree grows by 2 nodes each time it could end up with more nodes than this condition specifies as a result of the final split made.

Maximum Depth: A node will be allowed to grow further while it is at a depth less than that specified by the *Maximum Depth* stopping condition. A tree with just the root node has a depth of 1.

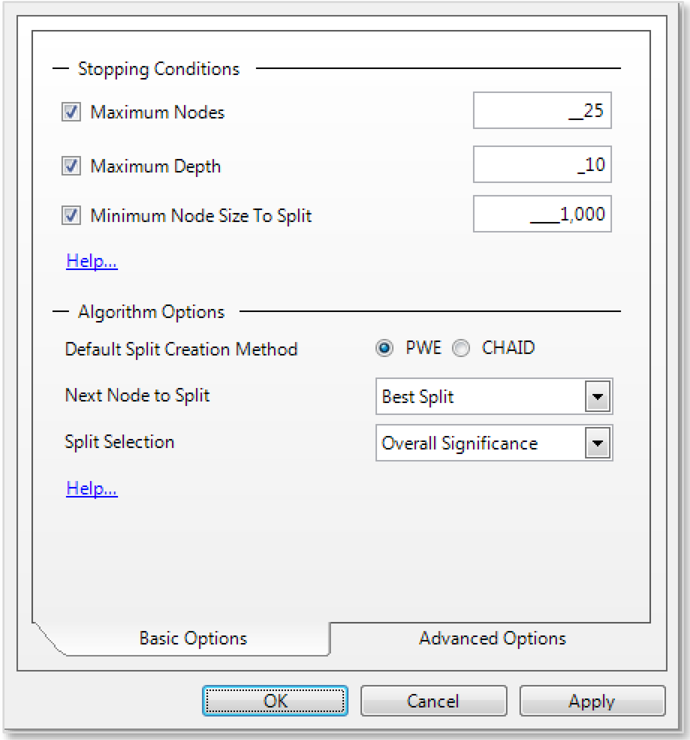
Minimum Node Size to Split: A node will be allowed to grow further while its base count is higher than that specified by the *Minimum Node Size to Split* stopping condition. It could then split into two nodes which may have fewer records than this condition.

- Set the **Stopping Conditions** as shown in the screen shot opposite, click **OK** and then **Yes** on the warning box

The various Algorithm Options are explained in more detail under the online Help accessed on this screen.

The various Advanced Options are explained in more detail under the online Help accessed on that screen.

For a further explanation of the PWE and CHAID algorithms see **Appendix 3**.



Stopping Conditions

☑ Maximum Nodes 25

☑ Maximum Depth 10

☑ Minimum Node Size To Split 1,000

Algorithm Options

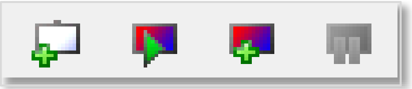
Default Split Creation Method ● PWE ○ CHAID

Next Node to Split: Best Split

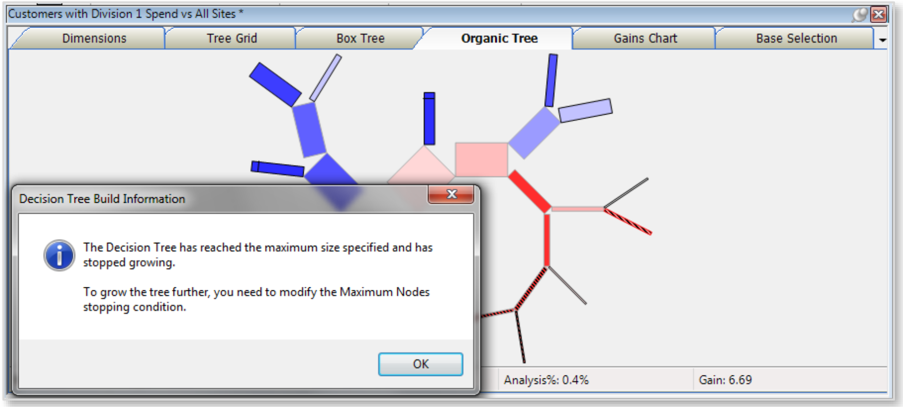
Split Selection: Overall Significance

Manual Build Controls

Before we Build the Decision Tree again with these new conditions it is useful to note that we have the ability to Pause and step through the build at any time using the appropriate buttons at the top of the window.



- Click on the **Organic Tree** tab
- Click on the **Build** button
- Click the **Pause Build** button
- Click on the **Single Step** button to stop after each split of the branches
- Click on the **Continue Build** button to continue the building of the Decision Tree until one of the Stopping Conditions has been met



Decision Tree Build Information

The Decision Tree has reached the maximum size specified and has stopped growing.

To grow the tree further, you need to modify the Maximum Nodes stopping condition.

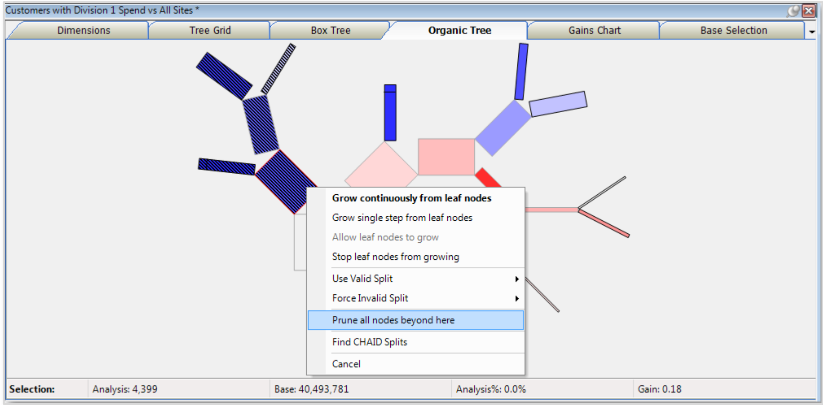
The Create Root button will start a new Decision Tree and stop after creating the Root Node.

The Market Insight Build button will build a new Decision Tree, but only if any of the settings have changed.

Growing and Pruning the Decision Tree

So far we have allowed our Decision Tree to grow according to the Build Options we have set in place. There may be occasions where we want to Prune back the branches where we believe they will bear no fruit (i.e. the blue nodes), or to experiment with different alternatives.

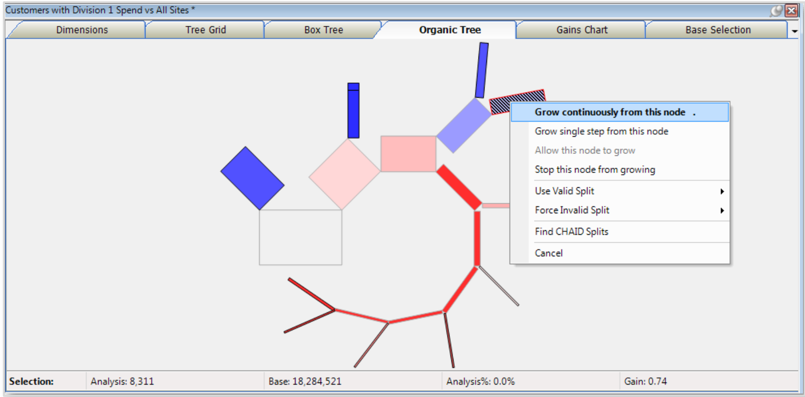
- Click on an appropriate node (Notice how all the branches beyond that node have been selected)
- Right click on the node and select the **Prune all nodes beyond here** option. Click **Yes** on the warning box


Customers with Division 1 Spend vs All Sites *
Dimensions
Tree Grid
Box Tree
Organic Tree
Gains Chart
Base Selection
Grow continuously from leaf nodes
Grow single step from leaf nodes
Allow leaf nodes to grow
Stop leaf nodes from growing
Use Valid Split
Force Invalid Split
Prune all nodes beyond here
Find CHAID Splits
Cancel

Selection: Analysis: 4,399 Base: 40,493,781 Analysis%: 0.0% Gain: 0.18

Note that all the branches have now been removed from the display.

- Right click on a leaf node and select the **Grow continuously from leaf nodes** option



Selection: Analysis: 8,311 Base: 18,284,521 Analysis%: 0.0% Gain: 0.74

The branches will grow back in accordance with the Build Options.

- Spend a few minutes Pruning branches and growing from different nodes to see what happens

You may want to apply new build options to allow the tree to grow further.

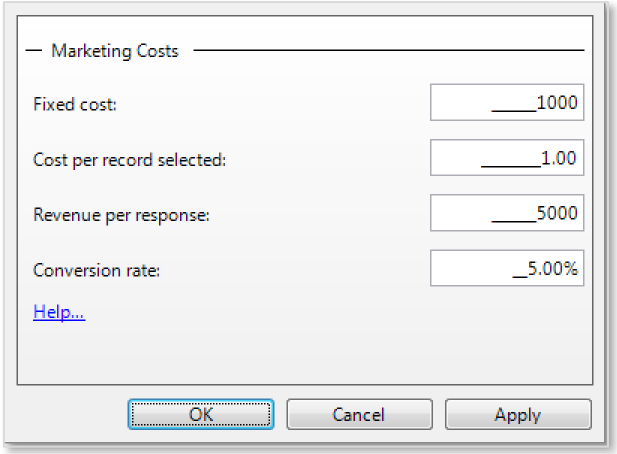
- Reset to the original **Build Options** and rebuild the Decision Tree (Max Nodes 20 & Max Depth 7)

You can use the **Allow this node to grow** and **Stop this node from growing** options to control which nodes grow when you use the **Single Step** and **Continue Build** buttons.

Marketing Costs

To get some idea of the possible returns in terms of revenue and profit, figures can be calculated by completing the Marketing Costs boxes.

- Click on the **Marketing Costs** button and complete as shown opposite. Click **OK**



Marketing Costs

Fixed cost:	____1000
Cost per record selected:	______1.00
Revenue per response:	____5000
Conversion rate:	__5.00%

The Marketing Cost inputs are as follows:

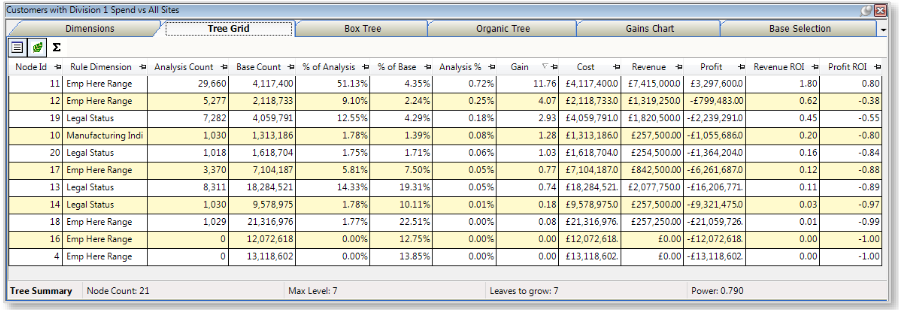
Fixed cost
This is the cost forecast to occur in initiating the marketing activity *independently* of the volume of data mailed. For example, this cost might include the costs of copywriting and leaflet design.

Cost per record selected
This is the cost forecast to occur for each record dispatched. This cost would typically include the data licensing cost, the letter print costs, the mailing cost, etc.

Revenue per Response [Customer]
This indicates the forecast revenue that can be expected for each new customer (successfully converted respondent).

Conversion Rate
This is the number of people who actually become customers as a proportion of those who are forecast to respond to the campaign.

- Click on the **Tree Grid** tab to see the calculated results



Customers with Division 1 Spend vs All Sites

Node Id	Rule Dimension	Analysis Count	Base Count	% of Analysis	% of Base	Analysis %	Gain	Cost	Revenue	Profit	Revenue ROI	Profit ROI
11	Emp Here Range	29,660	4,117,400	51.13%	4.35%	0.72%	11.76	£4,117,400.0	£7,415,000.0	£3,297,600.0	1.80	0.80
12	Emp Here Range	5,277	2,118,733	9.10%	2.24%	0.25%	4.07	£2,118,733.0	£1,319,250.0	-£799,483.00	0.62	-0.38
19	Legal Status	7,282	4,059,791	12.55%	4.29%	0.18%	2.93	£4,059,791.0	£1,820,500.0	-£2,239,291.0	0.45	-0.55
10	Manufacturing Indi	1,030	1,313,186	1.78%	1.39%	0.08%	1.28	£1,313,186.0	£257,500.00	-£1,055,686.0	0.20	-0.80
20	Legal Status	1,018	1,618,704	1.75%	1.71%	0.06%	1.03	£1,618,704.0	£254,500.00	-£1,364,204.0	0.16	-0.84
17	Emp Here Range	3,370	7,104,187	5.81%	7.50%	0.05%	0.77	£7,104,187.0	£842,500.00	-£6,261,687.0	0.12	-0.88
13	Legal Status	8,311	18,284,521	14.33%	19.31%	0.05%	0.74	£18,284,521.	£2,077,750.0	-£16,206,771.	0.11	-0.89
14	Legal Status	1,030	9,578,975	1.78%	10.11%	0.01%	0.18	£9,578,975.0	£257,500.00	-£9,321,475.0	0.03	-0.97
18	Emp Here Range	1,029	21,316,976	1.77%	22.51%	0.00%	0.08	£21,316,976.	£257,250.00	-£21,059,726.	0.01	-0.99
16	Emp Here Range	0	12,072,618	0.00%	12.75%	0.00%	0.00	£12,072,618.	£0.00	-£12,072,618.	0.00	-1.00
4	Emp Here Range	0	13,118,602	0.00%	13.85%	0.00%	0.00	£13,118,602.	£0.00	-£13,118,602.	0.00	-1.00

Tree Summary Node Count: 21 Max Level: 7 Leaves to grow: 7 Power: 0.790

You may need to turn on these columns by using the Column Chooser. Remember these results will depend upon an assumption that your prospects will react in a similar way to your current responders.

N.B. These figures are purely projections based upon the figures you input.

Output Decision Tree Model as Virtual Variable

To gain access to the results from the selection rules generated in a Decision Tree model you create a Virtual Variable. This will let you select people as well as evaluate the model in a Model Report.

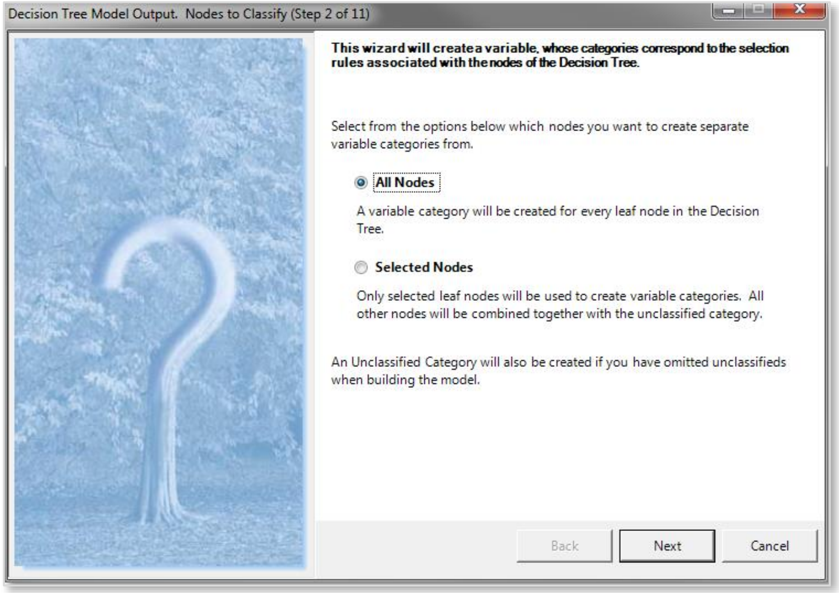
N.B. You can evaluate models in the Modelling Environment without having to create a virtual variable – a temporary one is created for you behind the scenes so that models can be compared.

- Click on the **Output Model As Variable** button

A series of step by step instructions will now guide you through the creation of your Virtual Variable. Use the following selections to complete this example:

- Step 2 – Select **All Nodes**, click **Next**
- Step 3 – Select **All Records**, click **Next** (jump to step 5)
- Step 5 - Leave on **Others**, click **Next**
- Step 6 – Enter optional notes, click **Next**
- Step 7 – Enter the description **DTM – Division 1 Spend**, click **Next**
- Step 8 – Security feature (not applicable on local system)
- Step 10 – Click **Finish**

Drag out the **DTM – Division 1 Spend** variable and use as any other variable. See the Help files for more information on this wizard.



Decision Tree Model Output. Nodes to Classify (Step 2 of 11)

This wizard will create a variable, whose categories correspond to the selection rules associated with the nodes of the Decision Tree.

Select from the options below which nodes you want to create separate variable categories from.

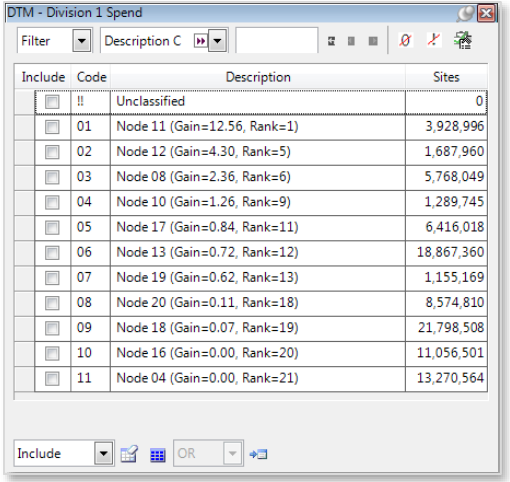
● **All Nodes**

A variable category will be created for every leaf node in the Decision Tree.

○ **Selected Nodes**

Only selected leaf nodes will be used to create variable categories. All other nodes will be combined together with the unclassified category.

An Unclassified Category will also be created if you have omitted unclassifieds when building the model.



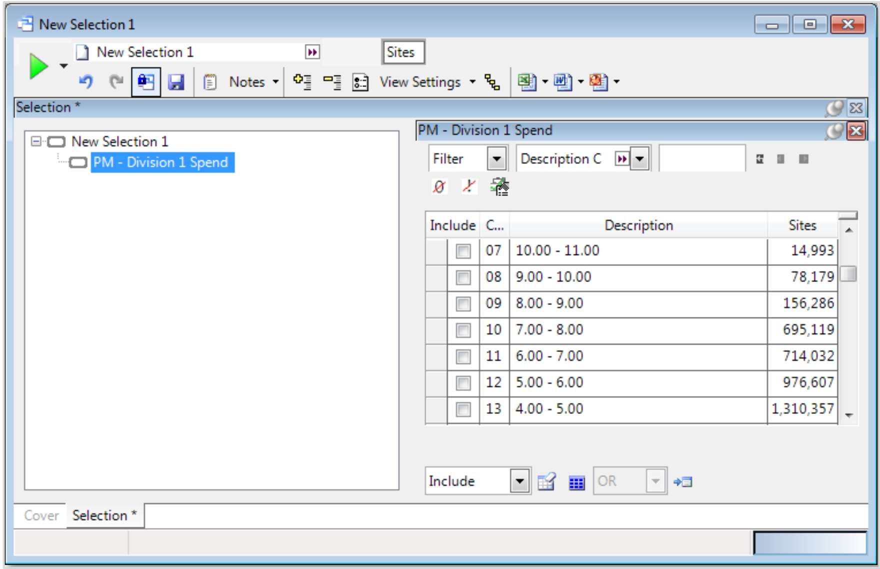
DTM - Division 1 Spend

Include	Code	Description	Sites
□	!!	Unclassified	0
□	01	Node 11 (Gain=12.56, Rank=1)	3,928,996
□	02	Node 12 (Gain=4.30, Rank=5)	1,687,960
□	03	Node 08 (Gain=2.36, Rank=6)	5,768,049
□	04	Node 10 (Gain=1.26, Rank=9)	1,289,745
□	05	Node 17 (Gain=0.84, Rank=11)	6,416,018
□	06	Node 13 (Gain=0.72, Rank=12)	18,867,360
□	07	Node 19 (Gain=0.62, Rank=13)	1,155,169
□	08	Node 20 (Gain=0.11, Rank=18)	8,574,810
□	09	Node 18 (Gain=0.07, Rank=19)	21,798,508
□	10	Node 16 (Gain=0.00, Rank=20)	11,056,501
□	11	Node 04 (Gain=0.00, Rank=21)	13,270,564

Introduction to Model Reports

Market Insight includes a powerful model reporting suite that can help you evaluate a Profile PWE model or a Decision Tree model as well as selector variables in identifying a target group of records.

The model reporting suite runs the model to see how well it can discriminate between the records that are in the Analysis selection and those that are not.

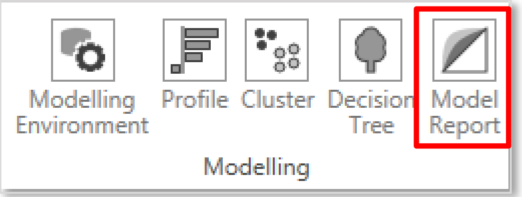


Include	C...	Description	Sites
□	07	10.00 - 11.00	14,993
□	08	9.00 - 10.00	78,179
□	09	8.00 - 9.00	156,286
□	10	7.00 - 8.00	695,119
□	11	6.00 - 7.00	714,032
□	12	5.00 - 6.00	976,607
□	13	4.00 - 5.00	1,310,357

Why use Model Reports?

…to ensure that the modelling process has succeeded in creating rules which identify the records that we started with. If the model is poor, the high scored records will include many that are not part of the group that we started with – the proportion of target records will be low. If the model is good, the high scored records will include few that are not part of the group that we started with but the proportion of target records will be high.

If the proportion of target records is very high, the model may be over-fitted i.e. so specific to the data that it was trained on that it becomes useful only to identify those records and useless in the general case. We want models that are powerful in the general case.

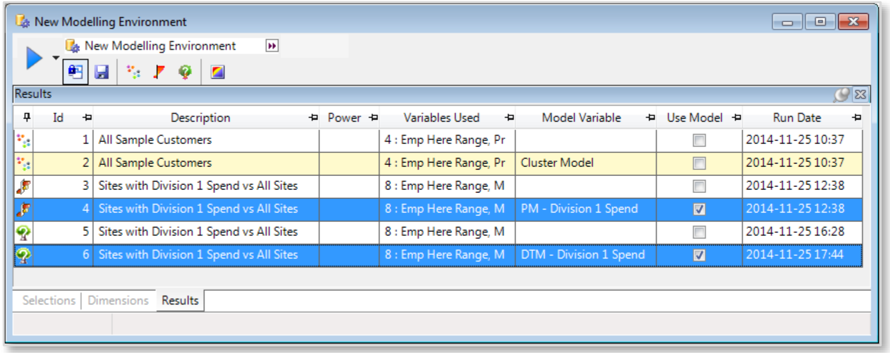


A Model Report Worked Example

Setting up within the Modelling Environment

Model Reporting can be used at any time within the Modelling Environment by selecting the models you want to look at from the Results window.

- Click on the **Results** tab of the **Modelling Environment**
- From the **Use Model** column put a tick against **PM – Division 1 Spend** and **DTM – Division 1 Spend**



	Id	Description	Power	Variables Used	Model Variable	Use Model	Run Date
	1	All Sample Customers		4 : Emp Here Range, Pr		□	2014-11-25 10:37
	2	All Sample Customers		4 : Emp Here Range, Pr	Cluster Model	□	2014-11-25 10:37
	3	Sites with Division 1 Spend vs All Sites		8 : Emp Here Range, M		□	2014-11-25 12:38
	4	Sites with Division 1 Spend vs All Sites		8 : Emp Here Range, M	PM - Division 1 Spend	☑	2014-11-25 12:38
	5	Sites with Division 1 Spend vs All Sites		8 : Emp Here Range, M		□	2014-11-25 16:28
	6	Sites with Division 1 Spend vs All Sites		8 : Emp Here Range, M	DTM - Division 1 Spend	☑	2014-11-25 17:44

N.B. You can add selector variables on the Results window and treat them as a model to be compared with the others

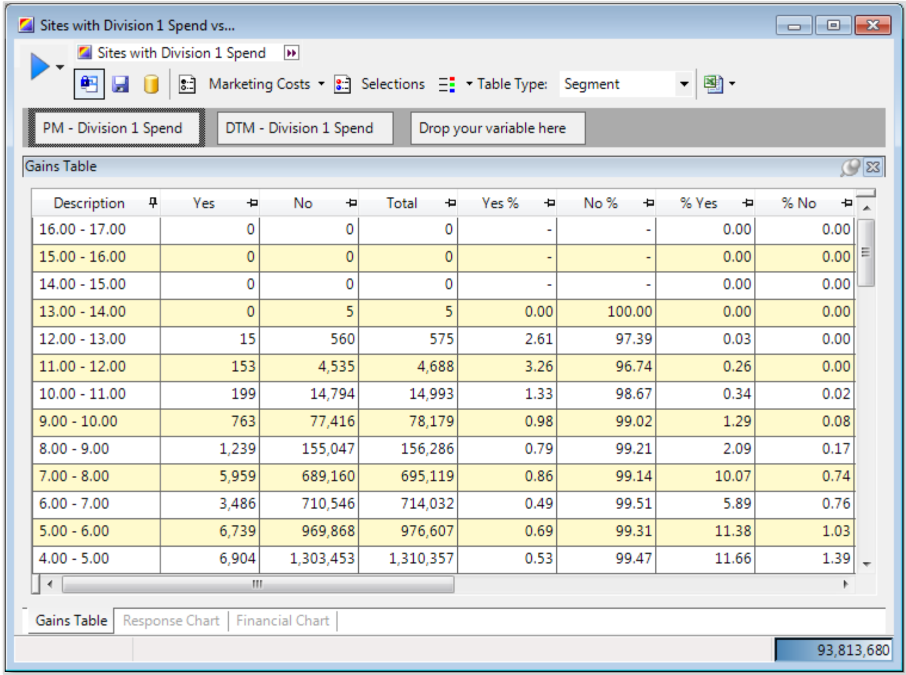
To obtain a quick indication of which model is the better predictor we can calculate the Power figure for each within the Modelling Environment.

- Click on the drop down arrow next to the ▶ **Build** button
- From the menu select **Evaluate Models**

A figure will appear in the Power column against each model. The closer that figure is to 1 the better the predictive power of that model.

For more details we can launch the Model Report tool:

- Click the **Create Model Report** button to take you into the Model Report window
- Click the ▶ **Build** button to create the report



Description	Yes	No	Total	Yes %	No %	% Yes	% No
16.00 - 17.00	0	0	0	-	-	0.00	0.00
15.00 - 16.00	0	0	0	-	-	0.00	0.00
14.00 - 15.00	0	0	0	-	-	0.00	0.00
13.00 - 14.00	0	5	5	0.00	100.00	0.00	0.00
12.00 - 13.00	15	560	575	2.61	97.39	0.03	0.00
11.00 - 12.00	153	4,535	4,688	3.26	96.74	0.26	0.00
10.00 - 11.00	199	14,794	14,993	1.33	98.67	0.34	0.02
9.00 - 10.00	763	77,416	78,179	0.98	99.02	1.29	0.08
8.00 - 9.00	1,239	155,047	156,286	0.79	99.21	2.09	0.17
7.00 - 8.00	5,959	689,160	695,119	0.86	99.14	10.07	0.74
6.00 - 7.00	3,486	710,546	714,032	0.49	99.51	5.89	0.76
5.00 - 6.00	6,739	969,868	976,607	0.69	99.31	11.38	1.03
4.00 - 5.00	6,904	1,303,453	1,310,357	0.53	99.47	11.66	1.39

Setting up outside the Modelling Environment

Model Reporting can be used at any time on variables created by Profile PWE models (or Decision Tree models and Selector variables).

- ➢ Drag and drop the **Model Report** from the **Toolbox** onto the workspace
- ➢ From the **System Explorer** → **Others** folder drag the **PM – Division 1 Spend** virtual variable onto the drop box indicated on the **Model Report** window. Repeat for **DTM – Division 1 Spend**

From the Marketing Costs button you can optionally complete the summary marketing parameters. (see pg. 59)

- ➢ Click the ▶ **Build** button to create the report

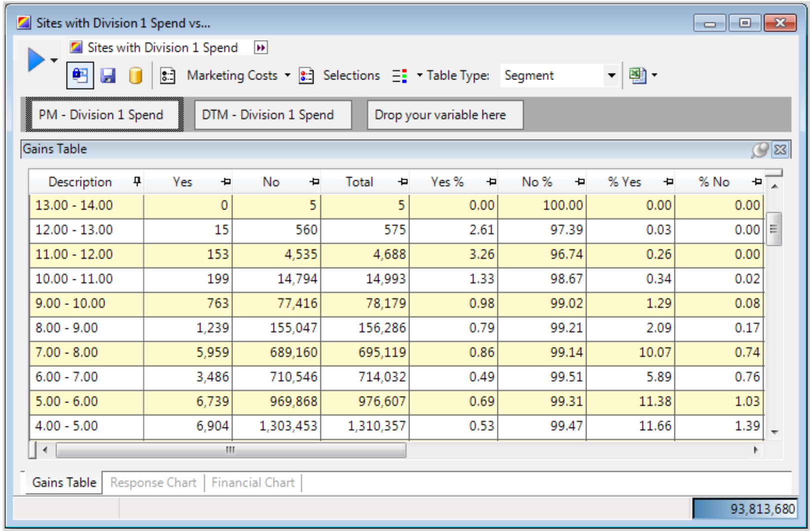


Table Type: Segment

Gains Table

Description	Yes	No	Total	Yes %	No %	% Yes	% No
13.00 - 14.00	0	5	5	0.00	100.00	0.00	0.00
12.00 - 13.00	15	560	575	2.61	97.39	0.03	0.00
11.00 - 12.00	153	4,535	4,688	3.26	96.74	0.26	0.00
10.00 - 11.00	199	14,794	14,993	1.33	98.67	0.34	0.02
9.00 - 10.00	763	77,416	78,179	0.98	99.02	1.29	0.08
8.00 - 9.00	1,239	155,047	156,286	0.79	99.21	2.09	0.17
7.00 - 8.00	5,959	689,160	695,119	0.86	99.14	10.07	0.74
6.00 - 7.00	3,486	710,546	714,032	0.49	99.51	5.89	0.76
5.00 - 6.00	6,739	969,868	976,607	0.69	99.31	11.38	1.03
4.00 - 5.00	6,904	1,303,453	1,310,357	0.53	99.47	11.66	1.39

93,813,680

Selections Control

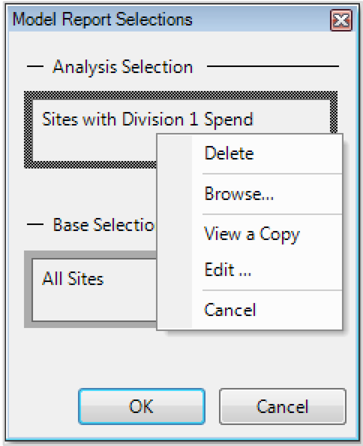
It is possible to see the selections used in the model:

- ➢ Click on the **Selections** button

The description of both the Analysis and Base selections are shown. To see the selection criteria:

- ➢ Right click on the selection name and then click on the option **View a Copy**

For more information go to the Help file – **Model Reports** → **Evaluating Models**

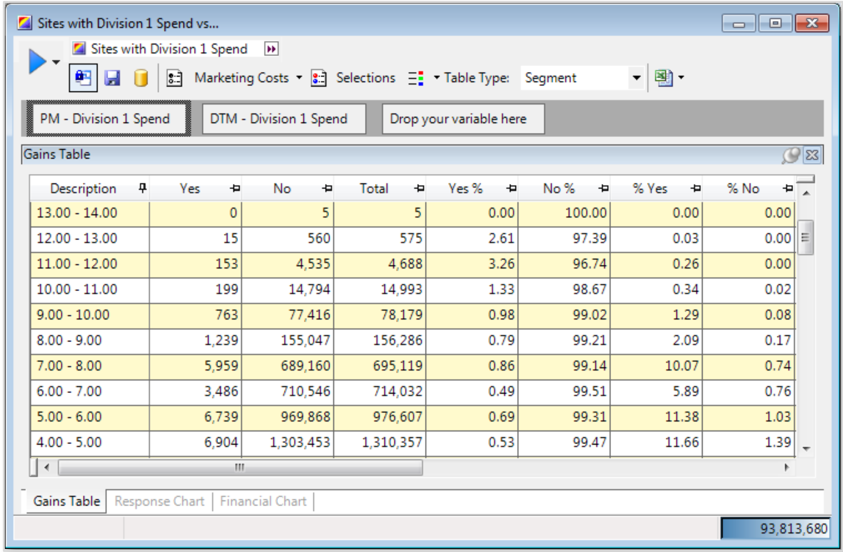


Reviewing the Model Report

Along the bottom of the report window there are three tabs that display a Gains Table, a Response Chart and a Financial Chart. These are explained as follows using the PM – Division 1 Spend.

Gains Table

First review the **Segment Gains** Table that shows the results for each segment – in this case each centile. Segment Gains show the value of the model reported in each individual segment.



Sites with Division 1 Spend vs...

Sites with Division 1 Spend

Table Type: Segment

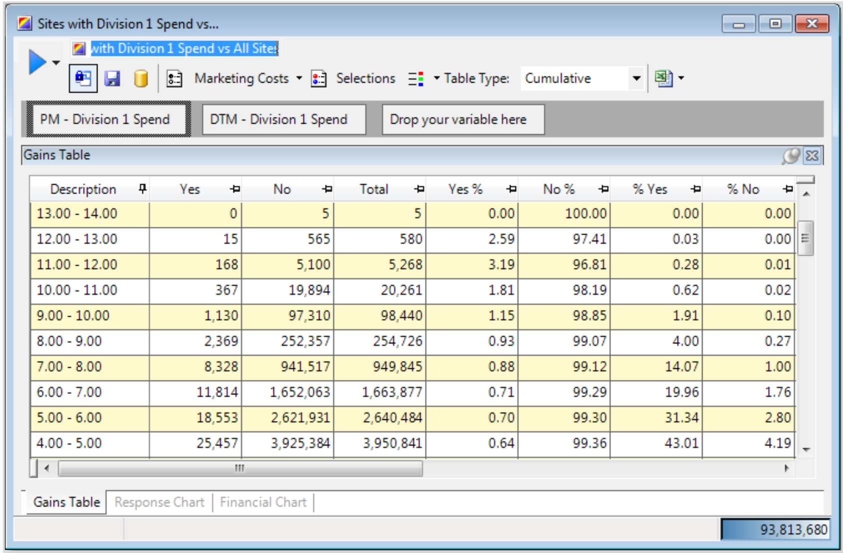
Gains Table

Description	Yes	No	Total	Yes %	No %	% Yes	% No
13.00 - 14.00	0	5	5	0.00	100.00	0.00	0.00
12.00 - 13.00	15	560	575	2.61	97.39	0.03	0.00
11.00 - 12.00	153	4,535	4,688	3.26	96.74	0.26	0.00
10.00 - 11.00	199	14,794	14,993	1.33	98.67	0.34	0.02
9.00 - 10.00	763	77,416	78,179	0.98	99.02	1.29	0.08
8.00 - 9.00	1,239	155,047	156,286	0.79	99.21	2.09	0.17
7.00 - 8.00	5,959	689,160	695,119	0.86	99.14	10.07	0.74
6.00 - 7.00	3,486	710,546	714,032	0.49	99.51	5.89	0.76
5.00 - 6.00	6,739	969,868	976,607	0.69	99.31	11.38	1.03
4.00 - 5.00	6,904	1,303,453	1,310,357	0.53	99.47	11.66	1.39

93,813,680

However it is more useful to look at the **Cumulative Gains** Table as this shows the cumulative result of marketing to each segment from best to worst (i.e. first to last centile).

- Click on the drop down arrow next to **Table Type** and select **Cumulative**



Sites with Division 1 Spend vs...

with Division 1 Spend vs All Sites

Table Type: Cumulative

Gains Table

Description	Yes	No	Total	Yes %	No %	% Yes	% No
13.00 - 14.00	0	5	5	0.00	100.00	0.00	0.00
12.00 - 13.00	15	565	580	2.59	97.41	0.03	0.00
11.00 - 12.00	168	5,100	5,268	3.19	96.81	0.28	0.01
10.00 - 11.00	367	19,894	20,261	1.81	98.19	0.62	0.02
9.00 - 10.00	1,130	97,310	98,440	1.15	98.85	1.91	0.10
8.00 - 9.00	2,369	252,357	254,726	0.93	99.07	4.00	0.27
7.00 - 8.00	8,328	941,517	949,845	0.88	99.12	14.07	1.00
6.00 - 7.00	11,814	1,652,063	1,663,877	0.71	99.29	19.96	1.76
5.00 - 6.00	18,553	2,621,931	2,640,484	0.70	99.30	31.34	2.80
4.00 - 5.00	25,457	3,925,384	3,950,841	0.64	99.36	43.01	4.19

93,813,680

The purpose of the Gains Table is to estimate the Response Rate from each model grade, if it were to be used in marketing. Clearly the marketing has not yet taken place, and so this estimate is made by looking at existing data. The Analysis records represent people who have already behaved in the way we want responders to behave (e.g. in this example, to be a Division 1 Spender). By looking at the count of and proportion of Analysis records in each model segment (the Yes and Yes% columns), we can estimate the number of responders and the Response Rate.

Each of the columns is described on the next page.

Gains Table Columns

- The **Description** at the left edge shows the score bands used in the PWE model, with highest scores at the top.
- The **Yes** column is an estimate of the number of Responders and is based on the number of successfully located Analysis records within that model grade.
- **No** shows the number of records within that model grade that were not Analysis records.
- **Total** shows the total number of records in that grade.
- **Yes%** is an estimate of the Response rate and is based on the percentage of successfully located analysis records found in this grade.
- **No%** shows the percentage of non-Analysis records included within the grade.
- **% Yes** shows the percentage of all Analysis records included in this grade.
- **% No** shows the percentage of non-Analysis records included in this grade.
- **% Total** shows the percentage of all Base records included in this grade.

N.B. Right click on a heading and select Column Chooser. Tick the options for Alt Cost, Alt Revenue and Alt Profit. These figures will relate to all the records not in the analysis group.

- **Gain** shows how much higher a proportion of analysis records this grade contains compared with an average selection of the same number of records from the database. The Gain is a measure of the success of the grade of the model. Higher scoring grades will typically have a high proportion of Analysis records i.e. “Gain” times better than the overall average.
- **Revenue** shows the total forecast Revenue i.e. Number of successfully located Analysis records * conversion rate * revenue per response NOTE: The revenue calculation is programmed to calculate the number of successfully targeted recipients based on membership of the analysis set. This will not be appropriate where the mailing is being sent to non-analysis set records only.
- **Costs** shows the total forecast costs based on the mailing volume multiplied by the Cost per Record figure given on the Costs tab. BEWARE: The cost calculation is programmed based on a mailing of all the base records including the analysis records. This may not be the case where the user wishes to mail only to net new records.
- **Profit** shows the difference between Revenue and Costs.
- **Revenue ROI** shows the return on investment for the revenue figure. A ROI above 1 shows a successful marketing activity – for each pound invested we receive more than one pound in revenue. By contrast a Revenue ROI figure of below one shows a grade that is losing money.
- **Profit ROI** shows the return on investment for the profit figure – this is the normal ROI quoted.

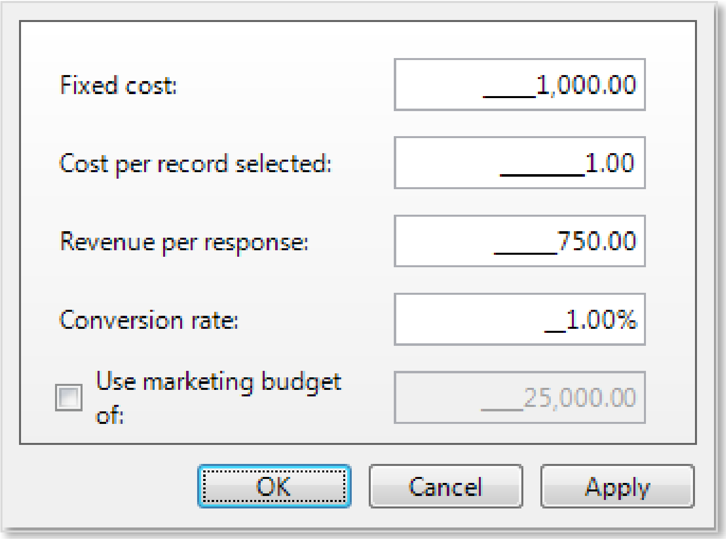
Marketing Costs

The **Fixed cost** is the costs forecast to occur in initiating the marketing activity independently of the volume of data mailed. For example, this cost might include the costs of copywriting and leaflet design.

The **Cost per record** selected is the cost forecast to occur for each record dispatched. This cost would typically include the data licensing cost, the letter print costs, the mailing cost, etc.

The **Revenue per response** indicates the forecast revenue that can be expected for each successfully targeted recipient.

The **Conversion rate** enables you to adjust the Response Rate calculations in the **Gains Table (Yes%)** made using the model data to suit your particular business. In the case of a two stage marketing process, the Conversion Rate lets you specify the proportion of positive responses which go on to generate revenue.



Fixed cost: 1,000.00

Cost per record selected: 1.00

Revenue per response: 750.00

Conversion rate: 1.00%

□ Use marketing budget of: 25,000.00

N.B. Clearly Market Insight is intended to increase revenue by increasing the response rate. However, this increase will be relative to existing performance, and the Conversion rate can also be used to adjust the forecasting calculations to reflect this.

The **Use marketing budget of:** enables you to specify a threshold value to be used for determining how much of the model data can be used.

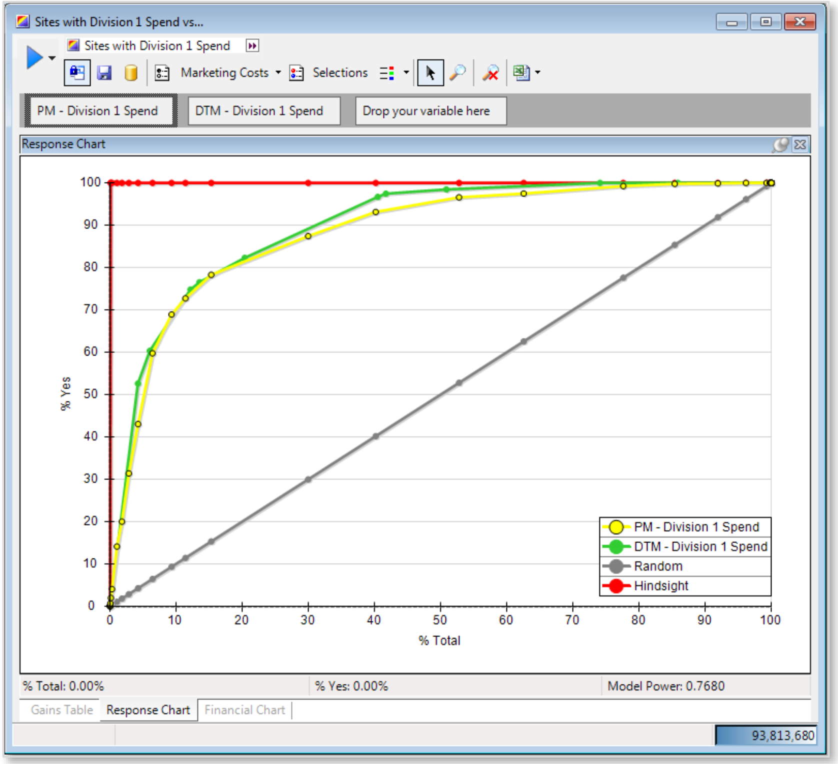
Response Chart

This chart plots the cumulative %Yes and %Total to give the classic representation of the power of the model.

➢ Click on the **Response Chart** tab at the bottom of the **Model Report**

The Response Chart shows the proportion of data used across the bottom and the proportion of Analysis Records found up the side.

- For comparison the **Grey line (Random)** shows the result of a random choice – there is a direct and linear relationship between the number of records selected and the number of analysis records found
- For comparison the **Red line (Hindsight)** shows the perfect result using our current knowledge of the Analysis set to plot the best outcome the model could possibly achieve
- The **Yellow line (Modelled)** shows the outcome achieved by the model. The steeper the yellow line and the closer it remains to the red line the better. However if it is too close to the red line we might suspect that the model is "too good to be true" i.e. over-fitted



% Total: 0.00% % Yes: 0.00% Model Power: 0.7680

93,813,680

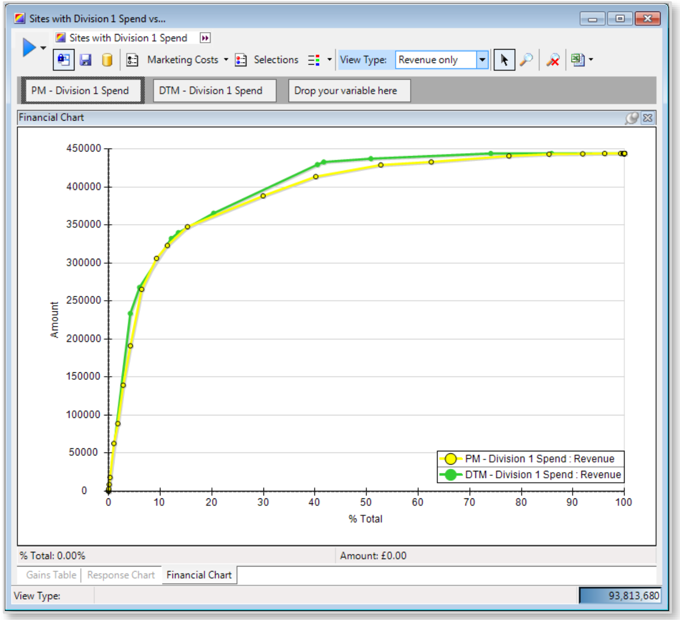
The area between the yellow and grey lines is an indication of the power of the model – the larger the area the stronger the model.

Use the Model Power figure to compare the effectiveness of more than one model showing in the report.

Financial Chart

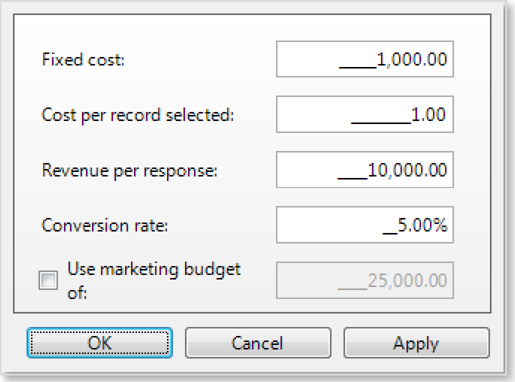
The financial chart allows you to view the plotting of Cost, Revenue, Profit or all three.

- Click on the **Financial Chart** tab at the bottom of the **Model Report**
- Click on the drop down arrow next to **View Type** to select the appropriate financial measure to view



This chart is helpful in understanding one of the financial measures in more detail.

The user can position the cross-hairs on the Single Financial Chart to adjust the blue highlight on the Cumulative Gains Table and override the budget setting.



Fixed cost: 1,000.00

Cost per record selected: 1.00

Revenue per response: 10,000.00

Conversion rate: 5.00%

☐ Use marketing budget of: 25,000.00

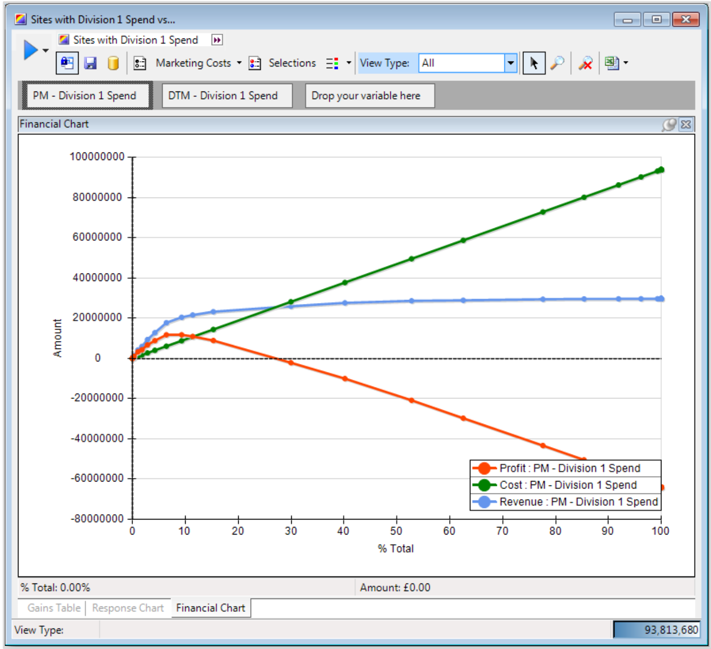
To adjust the figures used in the financial chart:

- Click on the **Marketing Costs** button and complete as opposite
- Click **Apply**

On all the charts you can use the Zoom tool (magnifying glass) to rescale the axes and magnify the interesting part of the chart.

The pointer tool is used to cumulatively select segments based on the model reports. In the example opposite just the more profitable segments have been selected.

The selection is linked between all the charts and the cumulative gains table. You can drag off the gains table to use the selection to drive a data extraction or other Market Insight visualization.



Comparing Different Model Types

It is possible to have a number of models evaluated at any time on the Model Report. In the example opposite three variables are being evaluated:

PM – Division 1 Spend

This variable is created from the PWE scores used in Profile as calculated and shown previously. It is the currently active display as indicated by the boarder around the circles which represent this model on the Response Chart.

DTM – Division 1 Spend

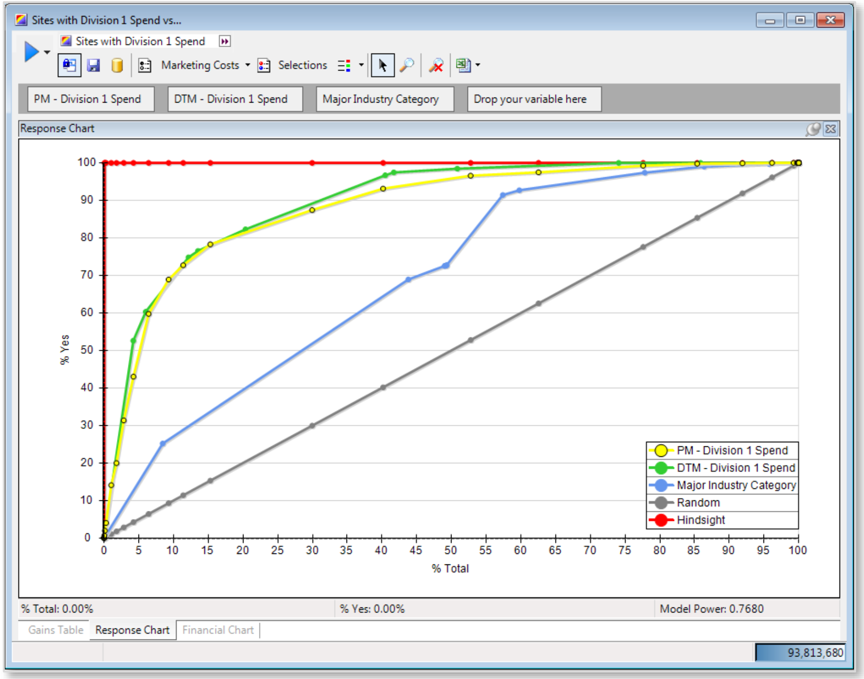
This variable has been created using Decision Tree with the same Analysis and Base selection used in Profile. The variable consists of categories based on the leaf nodes in the Decision Tree evaluation. (See the Decision Tree section of the manual for more details)

Notice that both the Profile and Decision Tree models are very close in their ability to identify members of the target group.

Major Industry Category

This variable is an item that was in the original Market Insight system and may form part of the more complex models above. As you might imagine a single variable on its own is unlikely to be as good a predictor of identifying members of the target group.

However, you may want to compare a number of single variables to see which the better predictor is.



% Total: 0.00%

% Yes: 0.00%

Model Power: 0.7680

93,813,680

The Limitations of Model Reporting

It is important that all users appreciate the model reporting suite is simply a calculator. It is not able to predict what will happen if you market to a certain set of records. The model reporting is simply provided to help evaluate the models by automating and structuring the calculation of cost and response estimates. There is absolutely no guarantee that the real world results will be anything like the estimates produced by the model reporting. The results presented by the model reporting are directly derived from the inputs provided by the user. If the user inputs a higher response rate, the model reporting suite will show a more successful outcome. However, this does not mean that the same will be achieved in the real world.

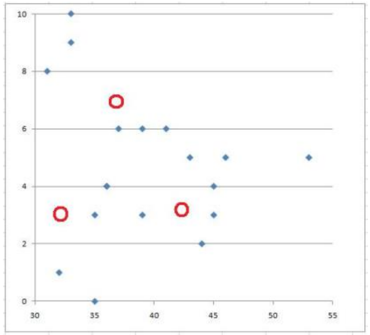
N.B.   Model Reporting is only a calculator that provides certain useful results for marketers. It is the user's sole responsibility to decide what marketing to undertake, and which records to use. Apteco/D&B accept no responsibility for marketing activity undertaken by the user.

Appendix 1 – Cluster Evolution

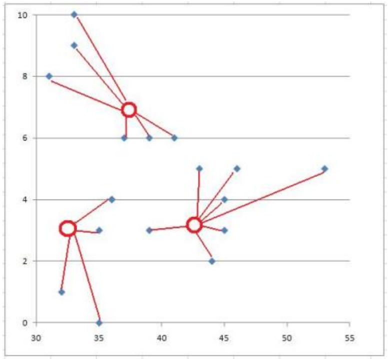
The standard and effective technique called K-Means used to identify a fixed number of clusters requires three defined elements. How to choose the initial starting points for cluster centers, what is meant by 'nearest' and what is the 'midpoint'?

The K-Means technique essentially works as follows:

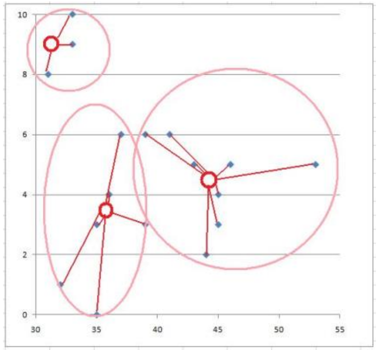
- A number of **starting points** are selected for the clusters to center upon
- Each data point is assigned to its **nearest** cluster
- The **midpoint** is calculated for all the points within each cluster
- The calculated midpoint of each cluster becomes the new center of the cluster
- Each data point is reassessed and assigned to its nearest cluster
- This process continues until the cluster centers no longer move (see the diagrams opposite)



Cluster Centre Starting Points



Initial Cluster Centres

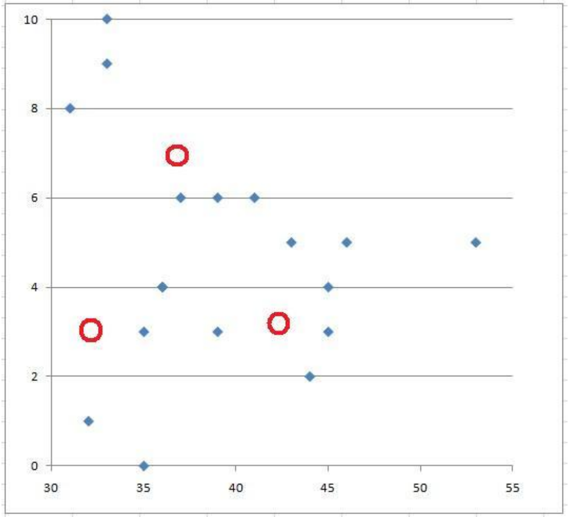


Final Cluster Centres

Starting Point – Initial Cluster Centre

As described above the K-Means algorithm adjusts the cluster centers iteratively until the center points stabilize. Unfortunately the final center points can often be influenced by the choice of starting points. The standard technique to overcome this is to repeatedly try random starting points, checking that similar final centers are found. This method can be used in the Market Insight Cluster Module by selecting the Random option under Initial Cluster Centers from the Cluster Settings.

The Cluster Module goes further using the Frequency option to try and find the best starting points. In Market Insight there is usually a large amount of categorical data so many points will be occupied by many records. A reasonable guess for the likely cluster centers is the points that are most heavily populated. This can be found by using a sparse cube limited to find the N most populated cells. The initial points do not want to be close together so this method carefully picks the points that are most populated but distant from the other initial centers.



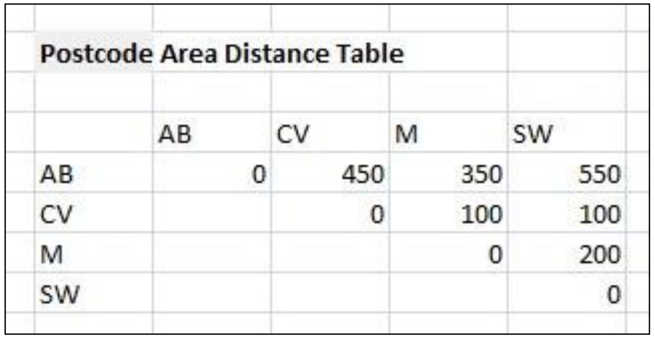
Cluster Centre Starting Points

Nearest – Distance and Similarity Measures

In our example one of the dimensions is numeric and the other is a numbered scale (ordinal) so we can just measure the distance. But this only works with ordinal data where the scales are similar.

In general we need some way of quantifying the distance between two records along each dimension and then some way of combining these together to give the overall distance between the records.

For some dimensions there is an obvious distance table. For example with UK postal areas we could use the drive time distance as shown in the screen shot opposite:



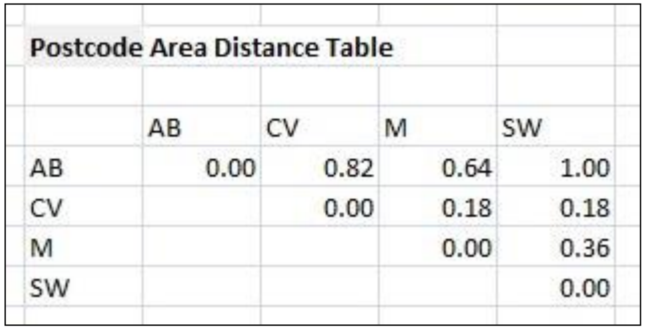
Postcode Area Distance Table

	AB	CV	M	SW
AB	0	450	350	550
CV		0	100	100
M			0	200
SW				0

This only shows a few postal areas but gives the idea.

The table is symmetric so we only need complete one triangle. The diagonal is zero.

We really need to standardize the range of the distances so that we can work with multiple dimensions where distances are measured in different size units e.g. miles, years, $. If we divide by the biggest distance we have a distance measure standardized between 0 and 1.



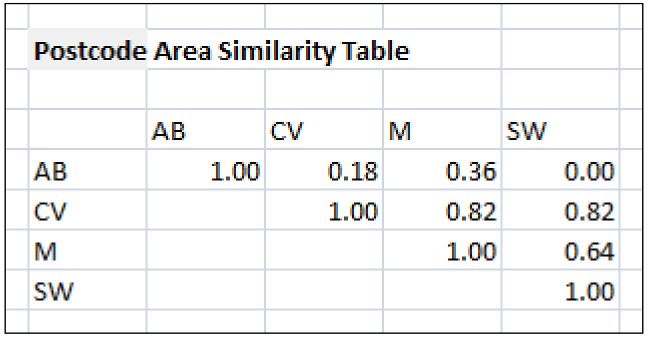
Postcode Area Distance Table

	AB	CV	M	SW
AB	0.00	0.82	0.64	1.00
CV		0.00	0.18	0.18
M			0.00	0.36
SW				0.00

As we are talking about nearest it is often easier to use a similarity measure rather than distance. Similarity is just 1.0 - distance but is a bit more intuitive as then an entry of 0.0 indicates complete lack of similarity whereas 1.0 means identically similar.

We can also refer to a proximity table or matrix as a generalization of similarity or distance. Market Insight will let you enter measures in terms of distance or similarity but internally uses similarity.

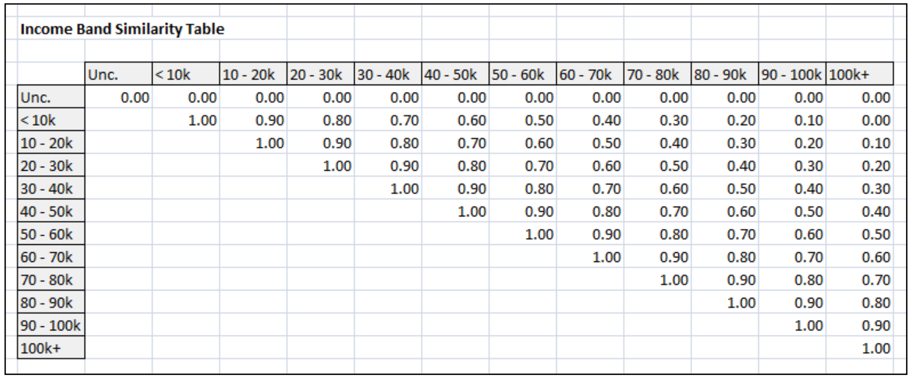
So finally we have a similarity table.



Postcode Area Similarity Table

	AB	CV	M	SW
AB	1.00	0.18	0.36	0.00
CV		1.00	0.82	0.82
M			1.00	0.64
SW				1.00

For ordinal data we can use some sort of scaling to calculate the similarity measures. For example with Income Band a linear scaling gives...(see screen shot opposite)



Income Band Similarity Table

	Unc.	< 10k	10 - 20k	20 - 30k	30 - 40k	40 - 50k	50 - 60k	60 - 70k	70 - 80k	80 - 90k	90 - 100k	100k+
Unc.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
< 10k		1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00
10 - 20k			1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
20 - 30k				1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20
30 - 40k					1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30
40 - 50k						1.00	0.90	0.80	0.70	0.60	0.50	0.40
50 - 60k							1.00	0.90	0.80	0.70	0.60	0.50
60 - 70k								1.00	0.90	0.80	0.70	0.60
70 - 80k									1.00	0.90	0.80	0.70
80 - 90k										1.00	0.90	0.80
90 - 100k											1.00	0.90
100k+												1.00

For nominal data we need to use some intrinsic or background knowledge. This was easy enough with the postal areas but would be more subjective with something like Occupation category. Market Insight enables you to manually specify a proximity matrix or derive one by using the PWE measure from a profile report (*this feature is not currently in the User Interface*). If you use the PWE method then the similarity is with respect to the analysis set.

If no proximity matrix is specified for a nominal dimension the default is 1.0 for matches on the diagonal and 0.0 otherwise.

In Market Insight we generally say that Unclassified / Missing are 0.0 similarity to any other category including Unclassified.

Midpoint – Cluster Centre

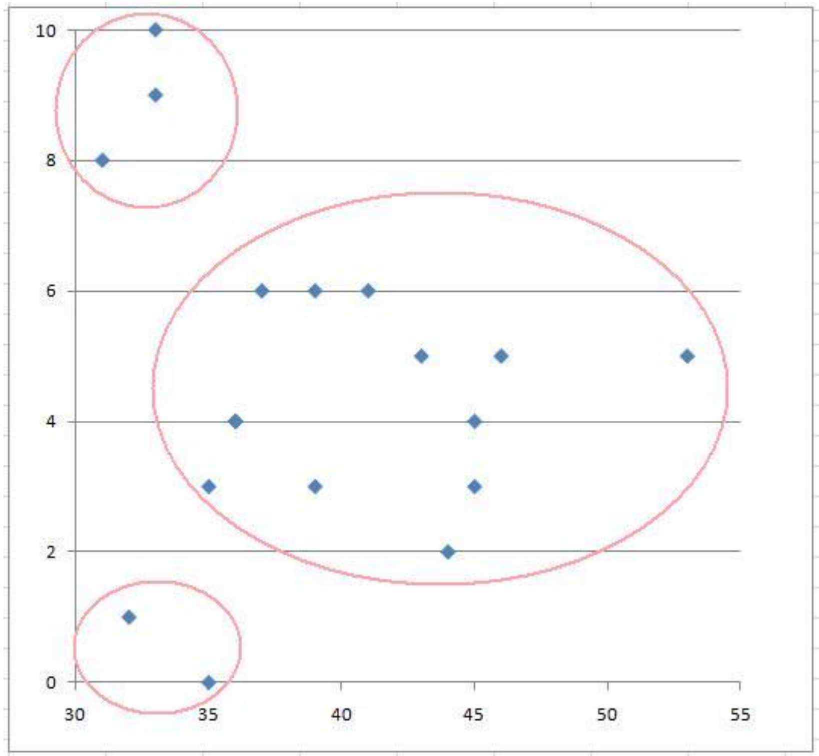
Looking at the largest clusters in the chart opposite it is easy to point to the center. But in our technique we can't easily see the line enclosing the cluster. A point is in the cluster if it is nearest to its center.

Also in real data there can be many records at the same point (e.g. people with same postal area / income band). So we choose the center of gravity of the points in the cluster as the midpoint i.e. weighted by the number of records at each point.

Depending on the distribution of points the center of gravity might be in between points (like the mean). It is more intuitive to choose an actual point that is nearest the middle (rather like a median). This is especially true when some or all dimensions are nominal - otherwise you could end up with a center defined part way between say male and female.

The exact center of gravity is termed centroid and the middle point is called medioid.

Medioid really implies that the data record actually exists (like a median). In theory any point is allowed i.e. combination of dimension values. But in practice it will always be an actual data record because of the way the “middle” is chosen. For each cluster there are frequency counts for each category for each dimension. For an ordinal dimension go along the categories until the cumulative total is half the total records in the cluster. For nominal the software picks the category with the most records. In either case the category identified as a coordinate of the center point must have a frequency count. Therefore the point is always populated.



Cluster Chart with 3 identified clusters

Include or Omit Unclassified

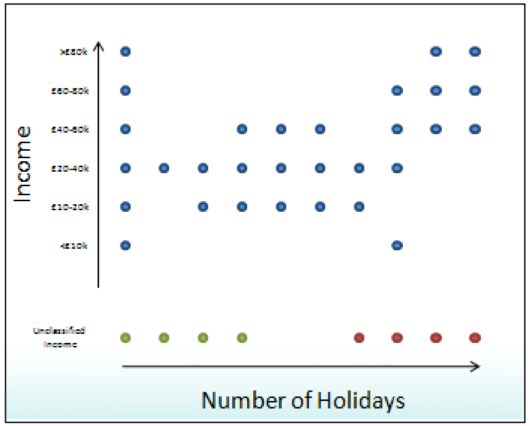
When Unclassified is included:

- Unclassified Income is treated as a long way from all other Income bands

- This will mean clusters tend to form within the unclassified, rather than between unclassified and other Income bands

- This leaves fewer clusters to subdivide the rest of the records where the income is known; in this example forming one big cluster

- ➢ This option allows the Unclassified Category to influence the clusters and should only be used if the category has a valid meaning. You can then use the distance matrix to control how close the category is to other values



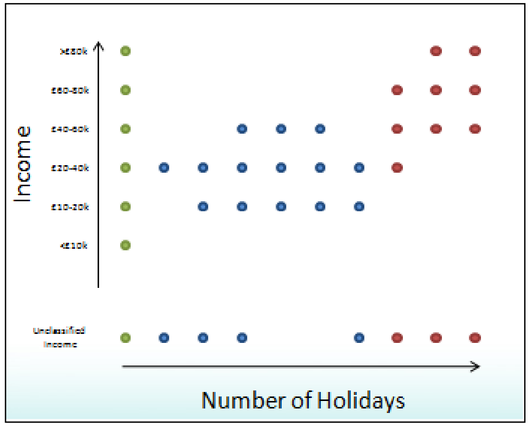
When Unclassified is Omitted:

- The records with an Unclassified Income are still included in the analysis, but their income value is ignored

- Clusters will form based on patterns in records with known Income bands and Number of Holidays

- Records with Unclassified Income join clusters based on their known data, i.e. Number of Holidays

- ➢ This option prevents the Unclassified Category from influencing the clusters, while allowing known data for these records still to be used. Other categories can also be omitted in this way with the Omit Category option



Appendix 2 – Profile Statistics

Index

The index for a category of a variable will be above 100 if a higher proportion of analysis records than base records have that value.

The index for a value of a variable will be below 100 if a lower proportion of analysis records than base records have that value.

The index for a value of a variable will be exactly 100 if the same proportion of analysis records as base records has that value.

Where the index is close to 100, we can say the analysis records follow the same proportions as the base records for these variable values. These variable values will not be useful to try to discriminate between records that do and do not fit the characteristics of the records we have analyzed.

Where the index is notably above 100, we can say the analysis set contains a higher proportion of these values than we would expect from a truly random selection of records from the base. We can use this information to help select records that are more likely to fit the characteristics of the records we have analyzed.

When the index is notably below 100, we can say the analysis set contains a lower proportion of these values than we would expect from a truly random selection of records. We can use this information to help avoid records that are less likely to fit the characteristics of the records we have analyzed.

Analysis% : Base%

Where the green analysis percentage bar is longer than the blue base percentage bar, you will see over-representation and a Penetration Index histogram to the right. Similarly where the green bar is shorter than the blue bar, you will see under representation. This is because the index is the ratio of these two percentages.

The Analysis %: Base % graphic is useful to see where you have a significant portion of your customer base even if there is no significant under or over representation.

Z-Score

The Z-Score identifies which results in our profile are sufficiently unlikely to occur in a truly random data sample that they indicate the data analyzed is most likely not a random sample but actually has some other characteristic.

The Z-Score is the "switch" that determines which category values of which variables in our profile give us useful decision making data and which are just the arbitrary results from any data sample. Only those variable category values that reach a certain Z-Score threshold (e.g. +/- 3 and above) will be used in our decision making (or "modelling" process).

If the Z-Score validates the use of a category of a variable for decision making, it is the index (and more precisely the PWE result derived from the index) that shows the true characteristic of the data.

The Z-Score protects us from making business decisions based on an insufficient sample of data. A data sample can be insufficient in terms of either the sample size or the detected difference between observed and expected results.

Z-Score calculation

The Z-Score calculation is used to measure the characteristics of samples taken from a population. The Z-Score calculation in general is:

(observed value - expected value)
----
standard deviation of observed value

In profiling, the analysis set is a sample from the base set and our results are proportions expressed as percentages, so

- the observed value is the analysis percentage,
- the expected value is the base percentage and
- the standard deviation is applied to the analysis percentage.

The standard deviation for the difference between proportions can be estimated for large populations, so in our case the calculation is

(Analysis% - Base%)

SQRT(Base% * (100 - Base%) / Analysis Total)

Note that the Z-Score used is actually Z-Score for Difference from Expected Value(ZdExp), which is appropriate for PWE modelling. ZdExp is the "adjusted residual" calculation, and produces similar values to the standard Z-Score for large samples

Z-Score Explained

If we took a truly random sample of our records, and analyzed the distribution of data, we would expect the analysis percentage for each category to be pretty close to the base percentage for that category. We would accept that the results may be not exactly equal as the randomness of the sample might give a slightly skewed result. However, if we repeated a truly random sampling many times we would get far more analysis percentage results close to the base percentage than far away. If we plotted these many analysis percentage results on a bar chart we would get a normal distribution “bell” curve of results around the base percentage result. The standard deviation result used in the z-score calculation measures how fat this bell curve is - the spread of the results.

If the number of records sampled (the analysis total) was small then we would be more likely to get some unrepresentative samples and the curve would be fatter. If the number of records sampled (the analysis total) was large then you would be less likely to get some unrepresentative samples and the curve would be thinner.

If the analysis percentage was very small then you would be more likely to get some unrepresentative samples (e.g. if there are only 3 out of 1000 you might get 2 in your first sample of 10) and the curve would be fatter. The same is true if the analysis percentage is very large.

A More General Z-Score Example

If we know that the average age of all of our customers is 35, whenever we take a large but truly random sample of our customers we would expect the average age to be close to 35. However, if we take a sample of customers who have bought a particular product, we might find the average age of those customers is just 21. This difference from the expected average for all customers is unlikely to have occurred by chance in the data sample we have tested, so (ignoring sample size etc. for this example) we can conclude that the customers who bought the particular product are not a truly random sample from the whole population and that therefore there is some correlation between the particular product and customer age. We would probably conclude we should market that product to younger customers.

Interpreting Z-Scores

A small z-score indicates that the analysis % result is just a slightly different but otherwise truly random sample from the base population.

A large z-score indicates that the analysis % result may just occur by chance in a truly random data sample, but the odds are very small – it is much more likely that the analyzed data is not actually a truly random but is biased in some way i.e. membership of that category might be related to membership of the analysis set.

The z-score gives us a standardized ruler with which to measure the real difference between the analysis percentage and base percentage taking into account the sampling process.

The Z-Score of a result will be positive if the analysis % was greater than the base % (i.e. observed result greater than expected result) and negative for the converse. The strength of the Z-Score is unaffected by whether it is positive or negative.

An easy interpretation of Z-Scores is the confidence we can associate with the result. A high Z-Score indicates a result that is unlikely (but still possible) to occur by chance in a truly random data sample. We can express our confidence in terms of how often such an unusual truly random sample would occur. For example, a Z-Score above +/- 2.576 indicates a result that would only occur one time in one hundred truly random samples. We can say in this case we have a confidence of 99% that the result shows the data is not a truly random sample (and therefore is biased in some way we can rely on).

Common thresholds for using Z-scores in marketing are:

+/- 3.290 or above, which equates to 99.9% confidence
+/- 2.576 or above, which equates to 99% confidence
+/- 1.960 or above, which equates to 95% confidence

You can display the equivalent confidence percentage in the profile report by right clicking on the column headings and using the Column Chooser.

Standard Measures

The Cramer's V and Phi measures are based upon the Chi-square statistic.

A chi square (X^2) statistic is used to investigate whether distributions of categorical variables differ from one another.

Phi

The phi coefficient attempts to reduce the influence of the sample size N and is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Cramer's V

Everitt describes several normalized versions of the χ^2-statistic. The only measure that overcomes limitations for the general case where r≠c is Cramer's *V* calculated as:

$$V = \sqrt{\frac{\chi^2 / N}{\min(r-1, c-1)}}$$

This scaling suggested by Cramer results in *V* ranging from 0 (complete independence) to 1 (complete association) for all values of r,c and N.

Information Gain

In general terms, the expected information gain is the change in entropy from a prior state to a state that takes some information as given:

IG(*Ex*,*a*) = *H*(*Ex*) - *H*(*Ex* | *a*)

Formal definition

Let *Attr* be the set of all attributes and *Ex* the set of all training examples, *value*(*x*,*a*) with $x \in Ex$ defines the value of a specific example *x* for attribute $a \in Attr$, *H* specifies the entropy. The information gain for an attribute $a \in Attr$ is defined as follows:

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \frac{|\{x \in Ex \wedge value(x,a) = v\}|}{|Ex|} \bullet H(\{x \in Ex | value(x,a) = v\})$$

The information gain is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. In this case the relative entropies subtracted from the total entropy are 0.

Appendix 3 – Decision Tree PWE v CHAID

PWE extracts predictive information which can be used to identify records in the analysis selection. This is done by examining the proportion of people in the target group, comparing the proportion observed in each category with that observed overall. This provides a PWE weight for each category, where a positive weight indicates that people from the category are more likely to be in the target group

The Binary PWE Decision Tree algorithm produces a tree which gradually homes in on the target group, without over fragmenting the data at any stage. This binary approach is often better when the performance of the overall tree is assessed. One disadvantage is that successive splits on the same variable, although efficient, make it harder for the user to recognize that the net effect may be a simple multi-way split.

Often a multi-way split is desired (e.g. low, medium and high income). CHAID can produce this in a single step, although there is no guarantee that CHAID will produce only a 3-way split, and there is a tendency for CHAID to over split. The Binary PWE split can always be modified manually to create a third group, or successive splits using the same variable will achieve an equivalent result.

The splits produced by CHAID are based on statistical tests so that all categories within a group are statistically similar, and each group is statistically different from the others. This can be an advantage, although with the high volumes of marketing data, differences can often be statistically significant when the user would prefer to combine branches together.

N.B. For further information regarding PWE and other statistical algorithms refer to the relevant White Papers published on the Forum.

The Binary PWE Decision Tree algorithm separates categories with a high PWE from those with a low PWE. The point chosen for making this split is the mean PWE score across all the categories. Thus categories which are more likely to contain the target group go in to one branch, and those less likely go into the other branch. Restrictions concerning unclassifieds and the selector branches are enforced after this initial split. The Binary PWE algorithm is fast and is always able to create a split.

There is a multi-way PWE Decision Tree algorithm which is again based on separating categories based on their PWE value. When the “mixed categories” option is used this is a simple matter of ranking the categories by their PWE value, and then creating a split where the change in PWE from one category to the next is greater than the threshold PWE value specified. This is a fast process. However, when categories are required to be ordered, then a slower more complex agglomerative clustering approach is used which ensures that categories within each node never differ in PWE by more than the threshold value.

The CHAID algorithm starts by considering all categories to be in separate branches and then iteratively combines branches which are statistically similar (based on a Chi-square test). This process results in a set of branches which are both significantly different from each other (in terms of the proportion of records in the target group) and also each comprised of categories that are statistically similar. The CHAID algorithm in Market Insight implements the recognized Chi-square Automatic Interaction Detection (CHAID) process for a binary target variable.

Market Insight imposes an upper limit on the number of branches that can be created in a multi-way split. Therefore the multi-way algorithms are not guaranteed to produce a split containing a valid number of branches.

Appendix 4 – Decision Tree Statistics

The various statistics used in the Decision Tree are explained below.

Analysis Count
The number of records in the analysis selection

Non-Analysis Count
The number of records in the base selection but not in the analysis selection

Base Count
The number of records in the base selection

Analysis %
The proportion of analysis selection records within a particular node or dimension value. i.e. the *Analysis Count* for the node or dimension value, as a percentage of the *Base Count* for that node or dimension value.

Non-Analysis %
The proportion of non-analysis selection records within a particular node or dimension value. i.e. the *Non-Analysis Count* for the node or dimension value, as a percentage of the *Base Count* for that node or dimension value.

Dimension Value
In the case of a selector variable, this is one of the variable categories (e.g. YOB = 1970)

% of [all] Analysis
The proportion of the whole analysis selection within a particular node or dimension value.

In the case of a node this is the *Analysis Count* for the particular node, as a percentage of the *Analysis Count* for the root node.

In the case of a dimension value this is the *Analysis Count* for the particular dimension value, as a percentage of the *Analysis Count* for the variable (all dimension values).

% of [all] Non-Analysis
The proportion of the whole non-analysis selection within a particular node or dimension value.

In the case of a node this is the *Non-Analysis Count* for the particular node, as a percentage of the *Non-Analysis Count* for the root node.

In the case of a dimension value this is the *Non-Analysis Count* for the particular dimension value, as a percentage of the *Non-Analysis Count* for all dimension values.

% of [all] Base
The proportion of the whole base selection within a particular node or dimension value.

In the case of a node this is the *Base Count* for the particular node, as a percentage of the *Base Count* for the root node.

In the case of a dimension value this is the *Base Count* for the particular dimension value, as a percentage of the *Base Count* for all dimension values.

Index

The ratio *% of Analysis* / *% of Base* for a node or dimension value, expressed around 100.

An index of 100: *% of Analysis* = *% of Base*
An index of greater than 100: *% of Analysis* greater than *% of Base*
An index of less than 100: *% of Analysis* less than *% of Base*

Gain

The *Analysis %* for the node divided by the *Analysis %* in the root node.

A gain of 1.0 = this node has the same *Analysis %* as the root node.
A gain of more than 1.0 = this node has a higher *Analysis %* as the root node.
A gain of less than 1.0 = this node has a lower *Analysis %* as the root node.

Z-Score

The *Z-Score* of a node measures the significance of any difference in terms of *Analysis %,* between this node and its parent, or between a dimension value and the dimension overall.

Z-Score = (Difference in *Analysis %*) / Standard Error.

Standard Error = Sqrt (Parent *Analysis %* * (1-Parent *Analysis %*) / Node Base Count)

A large *Z-Score* indicates that the child node is significantly different to its parent.
A small *Z-Score* indicates that any difference in *Analysis %* is not very significant. This is more likely when some or all of the following apply: a small difference in *Analysis %*, a small node base count, or when dealing with very high (e.g. >95%) or very low (e.g. <5%)
Analysis %.

Cramer's V

This provides a measure of association between the dimension being examined and the probability that someone is in the analysis selection.

It is a commonly used measure for assessing the association in a 2-way contingency table, based on the Chi-square value, but adjusting for the number of records and the shape of the table.

Cramer's V values range from 0.0 (no association) to 1.0 (maximum association)

Significance (P value)

The significance of the split is assessed using a Chi Square test to measure the association between the child nodes and the analysis selection.
The higher the Significance figure the more significant the child nodes. These figures are capped at 10 (equivalent to 0.0000000001, i.e. 10 decimal places).

A Significance of 0 indicates that the child nodes are not significant.

The Significance figure is calculated as - Log10 (P value), where the P value is taken directly from the Chi Square test. The smaller the P value the more significant the child nodes. Typically the values are so small they appear as 0.0000. The Significance value may therefore be more useful.

Minimum Z Score (after splits)

This is a measure of how different the least distinct child node is compared to the parent node. When a split is made, a Z-Score is calculated for each of the child nodes and this measure is the minimum (absolute) value of these scores.

Bonferroni Adjustment

The Bonferroni Adjustment reduces the P value to allow for the fact that any split is chosen from a large number of possible splits. It divides the P value by the number of possible splits that could be created for a particular node – this number depends on the number of categories, the number of nodes being created and the type of splits allowed (e.g. mixed, ordered, cyclic). In a simple example, where there 2 nodes are being made from a variable with 4 ordered categories, the possible splits are A | BCD, AB | CD and ABC | D. There will be considerably more possible splits if Mixed Categories are allowed, including AC | BD, AD | BC, ABD | C, ACD | B etc.

The P value is the probability of making a mistake when deciding to use a split (i.e. the probability that you make a split when actually there are no real differences in the data that warrant dividing a node further). Setting a P value of 0.05 means you want the chance of making a mistake to be 5% overall.

If you allowed a 5% chance of making a mistake when considering a single split, then in a situation such as above, where there are say 20 theoretically possible splits that could be considered, your overall chance of making a mistake when considering all possible splits is 20 x 5% = a guaranteed 100%. Arguably, not all the theoretically possible splits were actually considered by the Decision Tree algorithm: the creation of splits is not a random process.

However, a conservative approach commonly adopted to avoid over-fitting is to divide out the error allowed between the possible splits. So if 20 splits were theoretically possible, you would insist that a single split would only be accepted if the chance of error was 0.25% (5% / 20). This is a very conservative approach, but does mean that your probability of making a mistake is no more than 5% (the P value) overall.

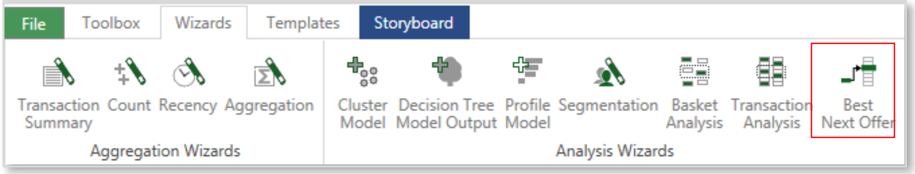
Gini

The Gini Impurity measure for the node (range = 0 to 1)
Min = 0, when all of node belong to same target group (Yes or No)

Calculated from “1 - Sum of Squared Probabilities of membership to each group”

Appendix 5 – Best Next Offer Wizard

The Best Next Offer (BNO) Wizard is available, via the Wizards ribbon bar, to any user with a modelling license. The wizard will suggest the best next offer for a customer based upon their previous transactions and/or those of all customers.



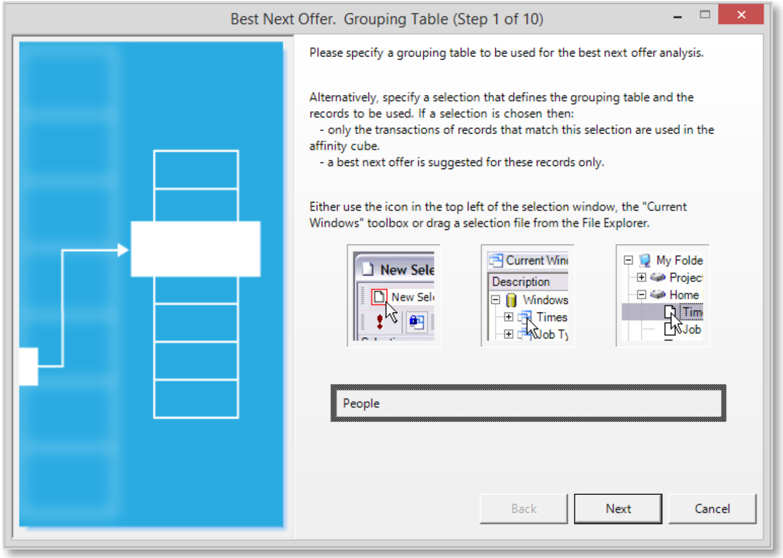
Firstly, the wizard generates an affinity cube identifying “what is bought with what?” The user then has the ability to apply popularity and propensity weightings before generating the virtual variable.

In this example we have used a system based upon a Holiday company to demonstrate the use of the wizard with a transactional variable.

Example

Create a Virtual Variable that will identify the best next holiday destination to offer a customer.

- ➢ Click on the **Best Next Offer** wizard link within the Analysis section in the wizards ribbon bar.
- ➢ **Step 1** - Drag a selection, or relevant table, on to the drop zone to identify the records you wish to be examined. We will select the People table. Click **Next**
- ➢ **Step 2** – Drag the transactional records to be used on to the drop zone. In this case we will use the bookings table. Click **Next**


Best Next Offer. Grouping Table (Step 1 of 10)

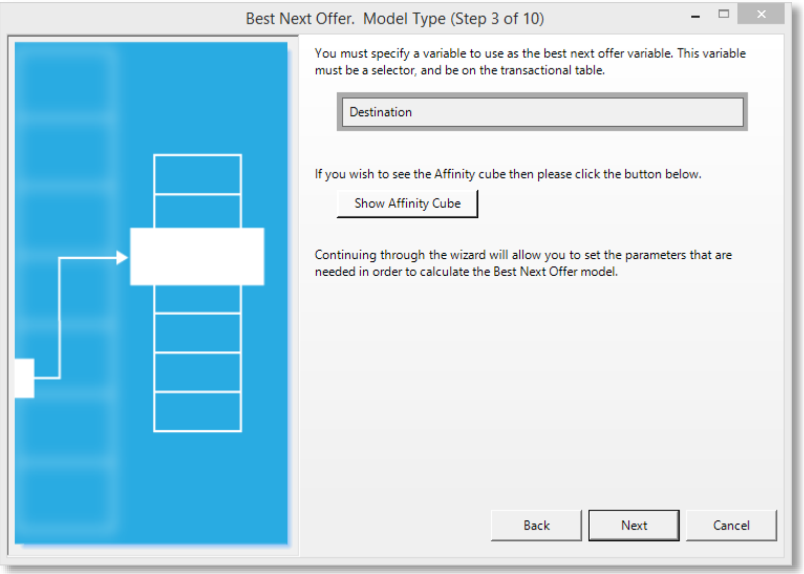
Please specify a grouping table to be used for the best next offer analysis.

Alternatively, specify a selection that defines the grouping table and the records to be used. If a selection is chosen then:
 - only the transactions of records that match this selection are used in the affinity cube.
 - a best next offer is suggested for these records only.

Either use the icon in the top left of the selection window, the "Current Windows" toolbox or drag a selection file from the File Explorer.

People

- **Step 3 –** Drag on the selector variable that relates to the offer. It must come from the transactional table used at step 2. We will use Destination
    - If you wish to view the Affinity Cube click **Show Affinity Cube**
    - To move to the next step click **Next**



You must specify a variable to use as the best next offer variable. This variable must be a selector, and be on the transactional table.

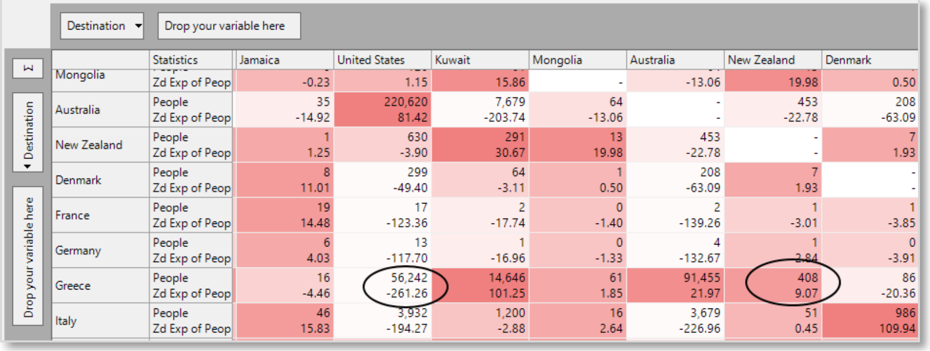
Destination

If you wish to see the Affinity cube then please click the button below.

Continuing through the wizard will allow you to set the parameters that are needed in order to calculate the Best Next Offer model.

Clicking on Show Affinity Cube presents you with a cube identifying which products are purchased with which other products. Destination is on both axis and each cell shows the number of customers who have purchased both products. The Z score is also included for each cell so we can see if the number of customers in that cell is higher or lower than would be expected.

We can see a large number of records who have been to Greece and the US however there is a minus Zd Exp hence those records are fewer than would be expected. Conversely far fewer records have been to Greece and New Zealand but the Zd Exp is much higher. Therefore we can conclude someone who has been on holiday to Greece is more likely to buy a holiday to New Zealand and less likely to buy one to the United States.



	Statistics	Jamaica	United States	Kuwait	Mongolia	Australia	New Zealand	Denmark
Mongolia	Zd Exp of Peop	-0.23	1.15	15.86	-	-13.06	19.98	0.50
Australia	People Zd Exp of Peop	35 -14.92	220,620 81.42	7,679 -203.74	64 -13.06	- -	453 -22.78	208 -63.09
New Zealand	People Zd Exp of Peop	1 1.25	630 -3.90	291 30.67	13 19.98	453 -22.78	- -	7 1.93
Denmark	People Zd Exp of Peop	8 11.01	299 -49.40	64 -3.11	1 0.50	208 -63.09	7 1.93	- -
France	People Zd Exp of Peop	19 14.48	17 -123.36	2 -17.74	0 -1.40	2 -139.26	1 -3.01	1 -3.85
Germany	People Zd Exp of Peop	6 4.03	13 -117.70	1 -16.96	0 -1.33	4 -132.67	1 2.84	0 -3.91
Greece	People Zd Exp of Peop	16 -4.46	56,242 -261.26	14,646 101.25	61 1.85	91,455 21.97	408 9.07	86 -20.36
Italy	People Zd Exp of Peop	46 15.83	3,932 -194.27	1,200 -2.88	16 2.64	3,679 -226.96	51 0.45	986 109.94

- **Step 4** - Here you can set the Parameters for the Wizard

Propensity – Increasing the Propensity will weight the results of the variable based on the previous transactions of the individual.

A high waiting may produce suggestions that are less common products. They will still be statistically significant and might show interesting niches to explore.

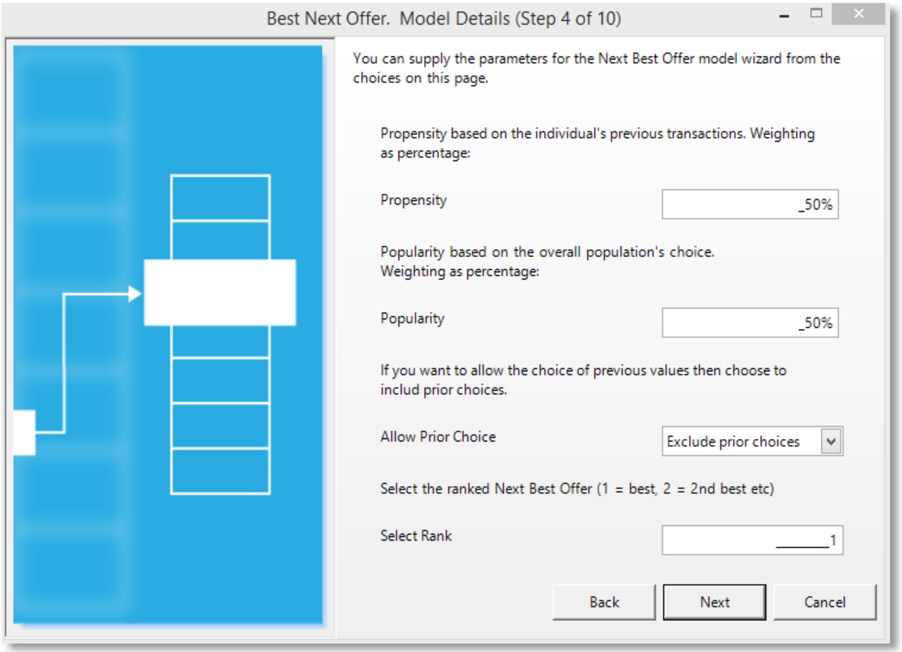
Popularity – Increasing the Popularity will weight the results of the variable based on the popularity of the overall population's transactions.

A high weighting will identify the most popular choice.

Allow Prior Choices – Here you can chose to allow or exclude prior choices, in this case holiday destinations person has been to before. In the case of a holiday a user may choose to allow prior choices. In the case of electronic goods, for example, a user may choose to exclude prior choices as a person would be unlikely to buy a washing machine immediately after having just bought one.

Select Rank – Here you can select a number to determine if you want 1st best or 2nd best etc.

Leave the settings as they are and click **Next**



Best Next Offer. Model Details (Step 4 of 10)

You can supply the parameters for the Next Best Offer model wizard from the choices on this page.

Propensity based on the individual's previous transactions. Weighting as percentage:

Propensity: _50%

Popularity based on the overall population's choice. Weighting as percentage:

Popularity: _50%

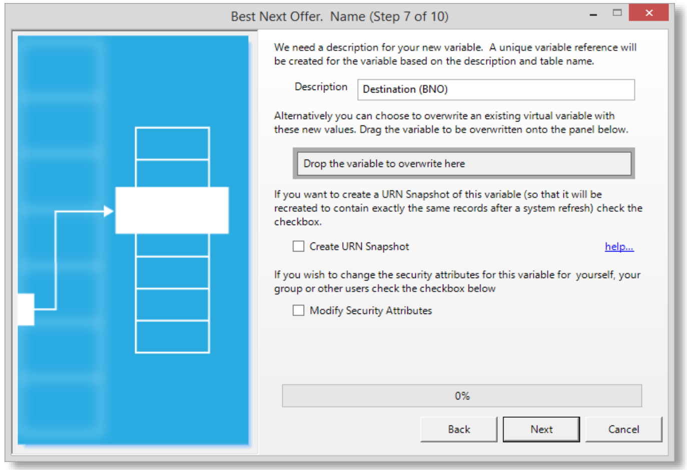
If you want to allow the choice of previous values then choose to includ prior choices.

Allow Prior Choice: Exclude prior choices

Select the ranked Next Best Offer (1 = best, 2 = 2nd best etc)

Select Rank: 1

- **Step 5** – Choose the folder in which you wish to save the variable. Click **Next**
- **Step 6** – Add any notes. Click **Next**
- **Step 7** – Enter the Description – Destination (BNO). Click **Next**
- **Step 10** – Click **Finish**



We need a description for your new variable. A unique variable reference will be created for the variable based on the description and table name.

Description: Destination (BNO)

Alternatively you can choose to overwrite an existing virtual variable with these new values. Drag the variable to be overwritten onto the panel below.

Drop the variable to overwrite here

If you want to create a URN Snapshot of this variable (so that it will be recreated to contain exactly the same records after a system refresh) check the checkbox.

□ Create URN Snapshot

help...

If you wish to change the security attributes for this variable for yourself, your group or other users check the checkbox below

□ Modify Security Attributes

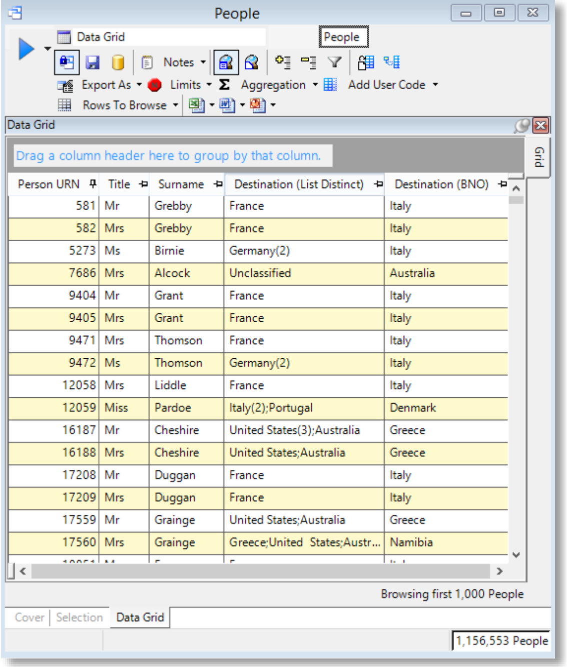
0%

The Best Next Offer has now been written in to a virtual variable which can now be used to personalise the next marketing communication a customer receives.

We can now use a selection of individuals we would like to market to, possibly high earners or those at risk of becoming lapsed customers, and identify the best marketing offer for them.

- Create a **People** level selection
- Drop a **Data Grid** on to the selection and add some relevant variables to view the data including **Destination (BNO)**. In the case opposite the data grid has aggregated bookings up to the people level.

We can see that as a result of her visiting Italy twice and Portugal once that we should offer Miss Pardoe a holiday to Denmark.



Drag a column header here to group by that column.

Person URN	Title	Surname	Destination (List Distinct)	Destination (BNO)
581	Mr	Grebby	France	Italy
582	Mrs	Grebby	France	Italy
5273	Ms	Birnie	Germany(2)	Italy
7686	Mrs	Alcock	Unclassified	Australia
9404	Mr	Grant	France	Italy
9405	Mrs	Grant	France	Italy
9471	Mrs	Thomson	France	Italy
9472	Ms	Thomson	Germany(2)	Italy
12058	Mrs	Liddle	France	Italy
12059	Miss	Pardoe	Italy(2);Portugal	Denmark
16187	Mr	Cheshire	United States(3);Australia	Greece
16188	Mrs	Cheshire	United States;Australia	Greece
17208	Mr	Duggan	France	Italy
17209	Mrs	Duggan	France	Italy
17559	Mr	Grainge	United States;Australia	Greece
17560	Mrs	Grainge	Greece;United States;Austr...	Namibia

Browsing first 1,000 People

1,156,553 People