# Data Mining Learning Resource

## An Introduction to Data Mining Processes

# Market Basket in SAS

# A walk-through of Market Basket Analysis using SAS Enterprise Miner.

We will be performing this Market Basket Analysis using the "Transactions" example data source in SAS Enterprise Miner Workstation 7.3.

The Transactions Data set will be accessible in the Further Reading and Multimedia page.

This will be undertaken in the 6-step CRISM-DM process.

## Step 1: Business Understanding

Without even running any analysis, we first need to decide on what we need to obtain from this data analysis; what do we want to know? In the case of this data set; we are provided with the Product purchased (Candy Bar, Bow, etc.), Store Number (from 1 to 10), the quantity of product purchased, and also the Transaction ID (unique for each product at each store – so Candy Bar at shop 1 will have a different number to Candy Bar at shop 2). In this analysis, we will attempt to discover meaningful patterns and information from this data; in particular discovering which items are purchased together with high confidence, and which items are not purchased with others – in order to help the business decide on a marketing strategy to become more profitable.

# Step 2: Data Understanding

This step in the data mining process, in our example, is about importing the data into SAS, as it has already been collected by the supermarket chain – and collate it into a meaningful format in order for the following steps to be able to perform a quantitative analysis. This module assumes you have basic knowledge of SAS Miner and have created a new project and have access to the Transaction data file.

# Step 3: Data Preparation

In this stage, we will be creating our Transactions Data Source inside of SAS Miner and doing a brief exploration of the data.

-Create a new data source, using the Transactions File.

-Clicking through to Step 4, use the Basic option

-Highlight all 4 variables, and click the Explore button

On this screen, we can see as follows:



(https://dataminingandvisualisation.files.wordpress.com/2013/11/dataprep.jpg)
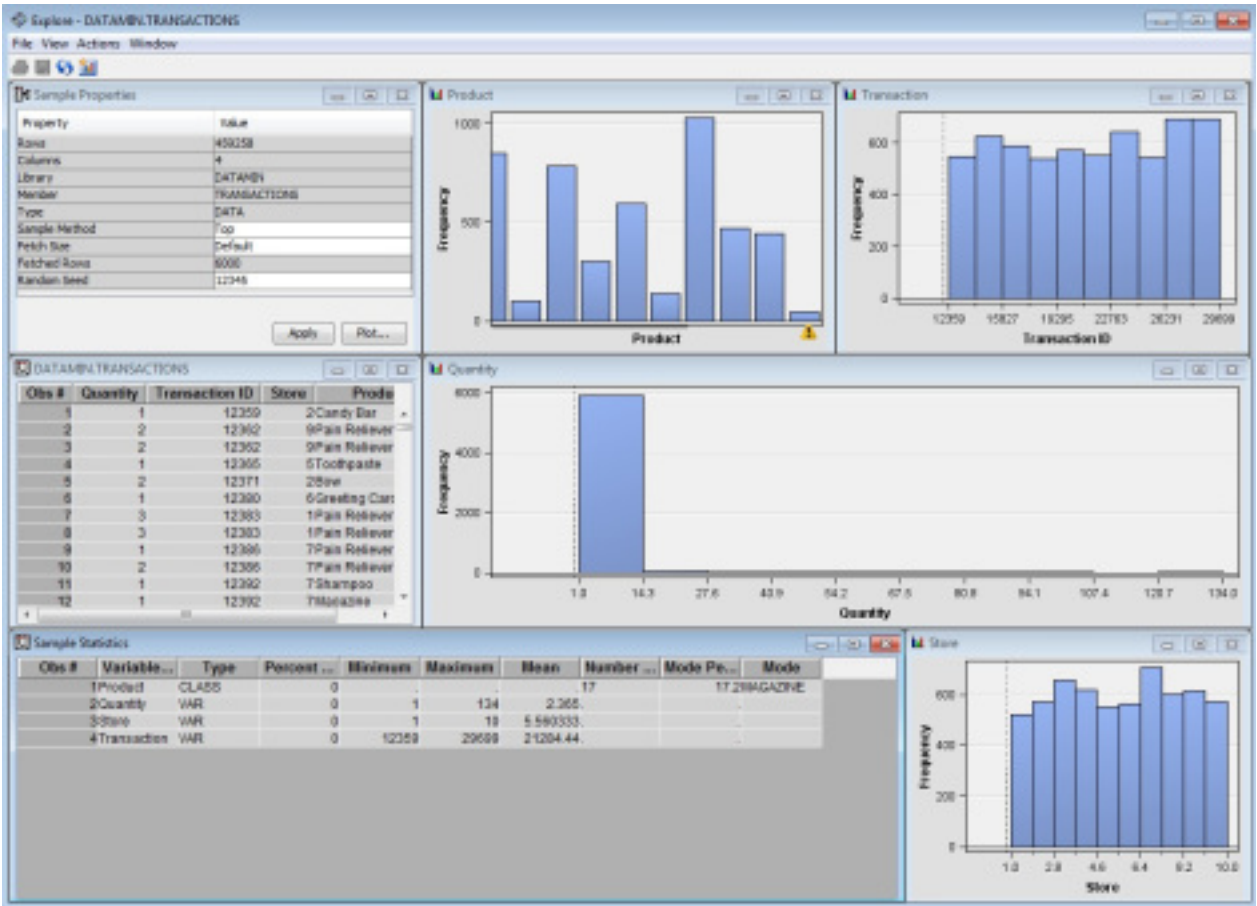
This gives us a very good summary of the data and we can already make a few inferences about the data – In the Sample Statistics, we can see that the mean value for products purchased in a transaction is 2.3 – so on average, we can say that most customers buy 2 or 3 items in any given transaction.

Expanding the Product graph to full size shows us the more popular products, which in this example are Magazines and Candy Bars.

Closing out of this screen, we can change the Product "Role" to Target, as it is the variable we want to be able to see the connections in. Quantity and Store are "input", and Transaction is the "ID" (and also change it to Nominal "Level" for this analysis), or identification between the data (as each transaction ID is unique for a certain combination of products).

# Step 4: Modelling.

In this step, we will use our imported Transactions data source to perform a Market Basket analysis of the data. We will first need to create a new Diagram, and drag the Transactions data source into it.

For our first stage of analysis, we will be dragging an **Association** node from the **Explore** tab, and then connecting the two as follows on the next page:



(https://dataminingandvisualisation.files.wordpress.com/2013/11/datamarketb.jpg)
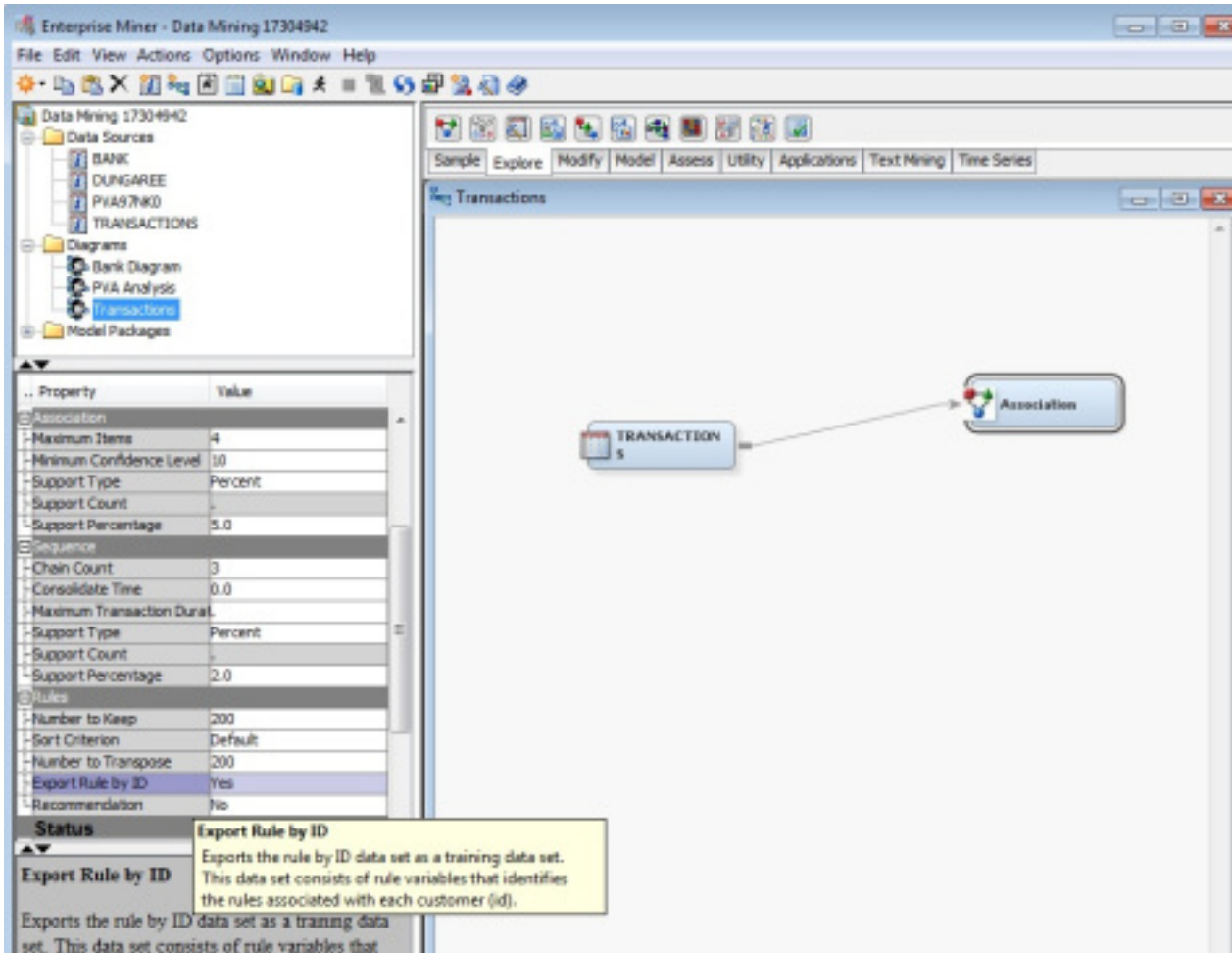
We will be setting the Export Rule by ID property to Yes, and this will allow us to view the Rule Description table later on when the diagram is Run. Right click the Association node now, and click Run. The process may take a minute, but when it is complete – press the Results button.



(https://dataminingandvisualisation.files.wordpress.com/2013/11/datamarketb2.jpg)

From this Results pane, we can see that the confidence and support of buying certain products per transaction are not relatively high (45% at the highest confidence and 4.5% at the highest Support), but many of the variables have very high lift (up to 3.60, or very positively correlated). This lets us say that people who buy Toothbrushes or Perfumes (which have a correlation, or lift with each other of 3.6) are much more likely to buy the other, than any other item in the shop – although we cannot say that they will with any high level of confidence.

Maximise the Statistics Line Plot – and remember in the introduction to Market Basket, that the Support of A -> B is the probability that the customer has both, while the Confidence of A -> B is the probability of the customer purchasing B, considering that they already have A. The Expected Confidence of A -> B is the probability that the customer will have B. Also note that Rules are Symmetric, in that A -> B is the same as B -> A.

Maximising the Rule Matrix, this graph plots each rule, having the Left side on the Y Axis and the Right side on the X Axis, with the points being colored based on their confidence, with the highest being red, with Magazine & Greeting Card -> Candy Bar at that point.

To continue on, let's explore the associations in a visual sense, through a link graph. To access this, click View -> Rules -> Link Graph:
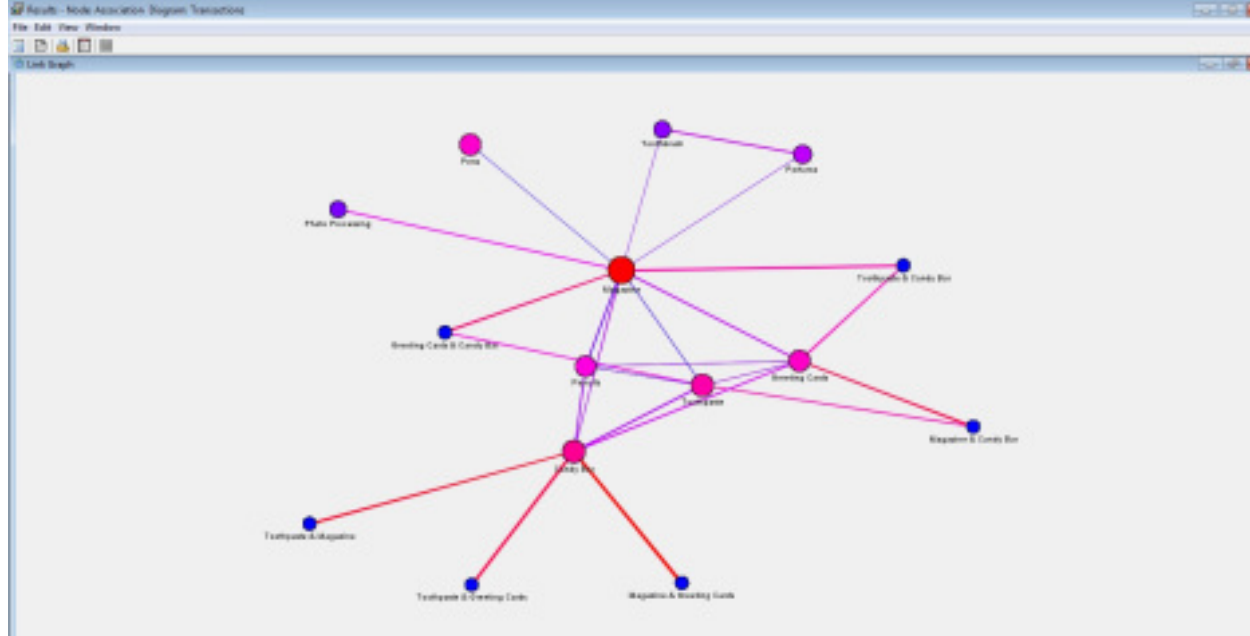
In this graph, the size and color of a link indicates the number of transactions inside of the data set, with larger nodes having greater counts. The color and thickness of a link determines the confidence level – with thicker links having higher confidence levels. In this, we can see that the only people who bought pens, had also purchased magazines in some of their purchases – with pens having no link to any other product node. This could be key information to a shop, which would tell them that they should sell Magazines and Stationary (including pens) near each other to maximise sales.

Another way to plot this rule information is through the plotting of the Rules table, which is found through View -> Rules -> Rules Table. This table allows you to sort and view the rules ascending or descending based on their lift, confidence or any other factor listed – which can show some interesting statistics – such as the Lift between Pens and Magazine being 0.61, which shows that they are in fact not positively correlated, yet negatively – which can then infer that people who buy pens are in fact less likely to buy a Magazine than a random selected customer.

Clicking the Plot Wizard Icon and selecting Three-Dimensional Scatter for the type will allow us to visualise the relationship between Lift, Confidence and Support in three dimensions. Setting Support = X, Lift = Y and Conf = Z as the variables to plot gives us this information:



(https://dataminingandvisualisation.files.wordpress.com/2013/11/marketb5.jpg)
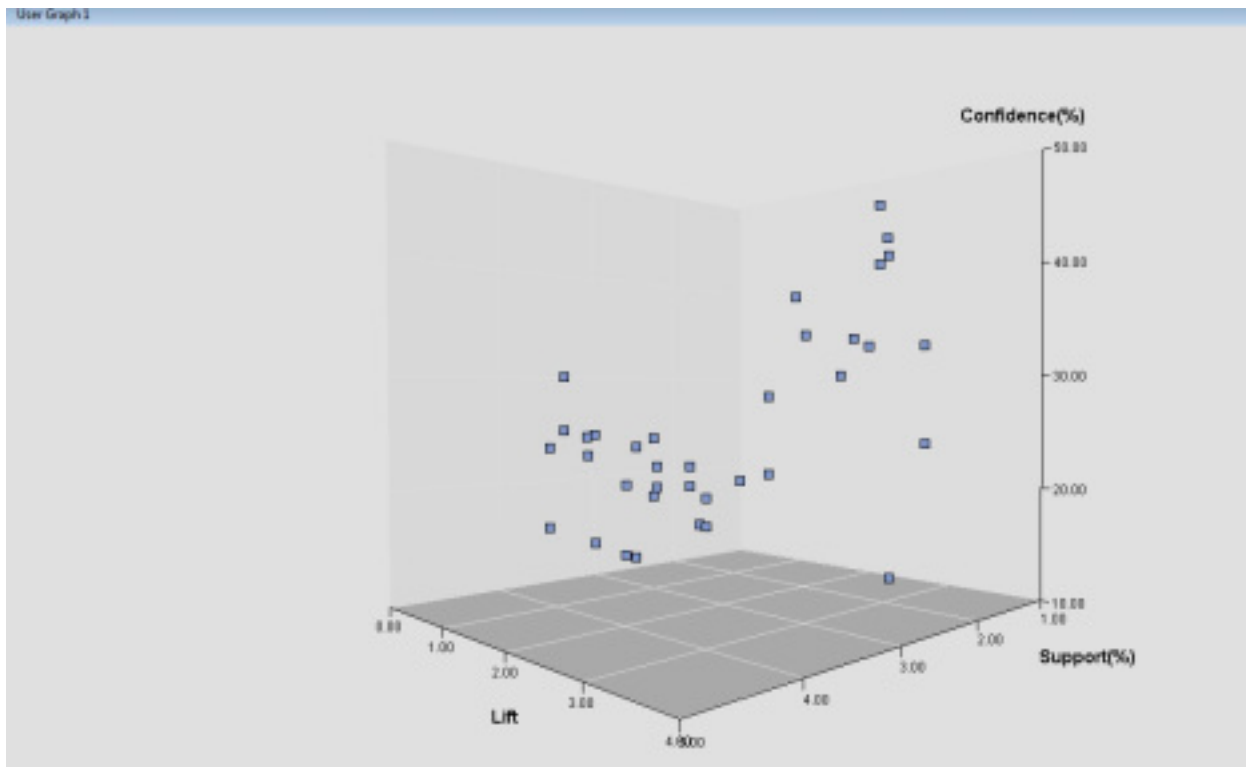
# Step 5: Evaluation

As stated in the introduction – in this phase we must ask ourselves if our outcomes have fulfilled our objectives. Our objective was "in particular discovering which items are purchased together with high confidence, and which items are not purchased with others" – and through this analysis, those conditions have been met; through the discovery and analysis of the products with high and low lifts with each other – for example, the lift of 3.60 for Toothbrush -> Perfume, which suggests that people who buy one are much, much more likely to purchase the other, compared with picking a customer at random.

Conversely, Pens -> Magazine has a lift of 0.61, which suggests that people who buy pens are actually much less likely to purchase a magazine in a given transaction, compared with picking a customer at random. The most commonly occurring transaction (found by sorting the "Transaction Count" column in the Rules Table descendingly) is in fact Candy Bars -> Greeting Cards and vice versa. This, along with the transaction's lift value of 1.74,

(https://dataminingandvisualisation.files.wordpress.com/2013/11/mb6.jpg)

# Step 6: Deployment

The deployment of this model would ideally be done on a weekly or monthly basis, where this data can be analysed and used to help devise plans for the layout of the shops – for example, it would be logical to have greeting cards and candy bars available together during major holidays throughout the year (such as Haloween or Christmas), as these are the times that Greeting Cards are usually purchased, which would therefore increase sales of Candy Bars. It could also be tested that the business tries placing their less successful products near the more successful ones, in order to increase the probability of the customer making a purchase of these.

# One response to "Market Basket in SAS"

1. *Monika M. Wahi, MPH, CPH* 07/12/2015 at 5:20 pm · · Reply →
   That's for this great walkthrough! Your straightforward explanations are very helpful.