Association (Revised)

Nitin Kalé, University of Southern California

Nancy Jones, San Diego State University

Part 1: titanic survival

# Objective

The objective of this exercise is to generate association rules (or affinity) for the survivability of the passengers on RMS Titanic[[1]](#footnote-1).

# Activities

* Import and prepare data
* Apply data mining algorithms
* Configure predictive models
* Create data visualizations
* Analyze and interpret output from models
* Publish results

# Software Prerequisites

* SAP Predictive Analytics (PA) 3.X

# UCC Products Required

* None

# Data Set

* Data file titled *Titanic\_E11\_1.xlsx*

# Scenario

Using an Excel data file containing information about the passengers of the Titanic, you will use association analysis to generate rules for their survivability.

# Association Analysis

Several predictive models are available in SAP Predictive Analytics. More can be integrated from the *R language*. We would like to discover the *associations* among items. These are presented as rules with values for *support*, *confidence,* and *lift* for each rule

1. We will now do an association analysis (using an *Apriori* algorithm) for the passenger data in the Titanic disaster [[2]](#footnote-2)
   1. Launch SAP Predictive Analytics
   2. Click on Expert Analytics, then on Expert Analytics.
   3. Create a new document. Choose MS Excel as Data Source. *Next.*
   4. Browse for the *titanic\_E11\_1.xlsx* file. *Create*.
2. We will now launch the prediction capabilities of SAP Expert Analytics
   1. Click on *Predict*. You are now in the *Designer* tab.
   2. You will see several *Algorithms* such as Regression, Outliers, Time Series, Decision Trees, Neural Network, Clustering and Association. See Figure 1.

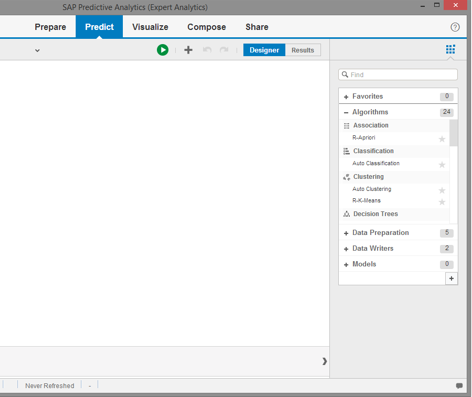
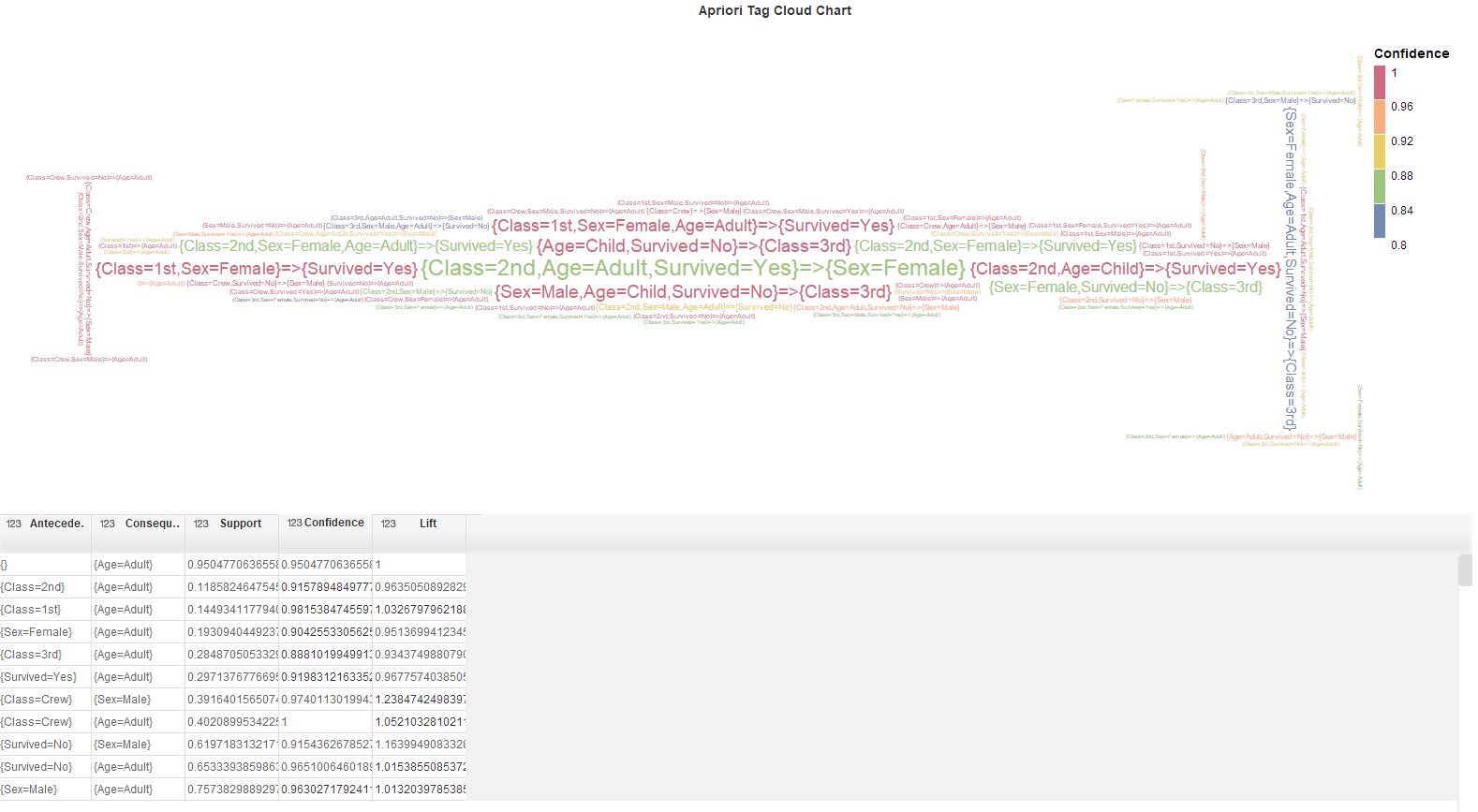


Figure 1

* 1. And you see the data source *titanic\_E11\_1.xlsx.*
  2. Double-click the *R-Apriori* algorithm. The algorithm is automatically connected to the data source
  3. Roll your mouse over the algorithm and click on *Configure Settings*
  4. Item Column(s) – Select *Class, Sex, Age* and *Survived*
  5. Support: 0.01, (leave confidence at .8)
  6. Done and Run

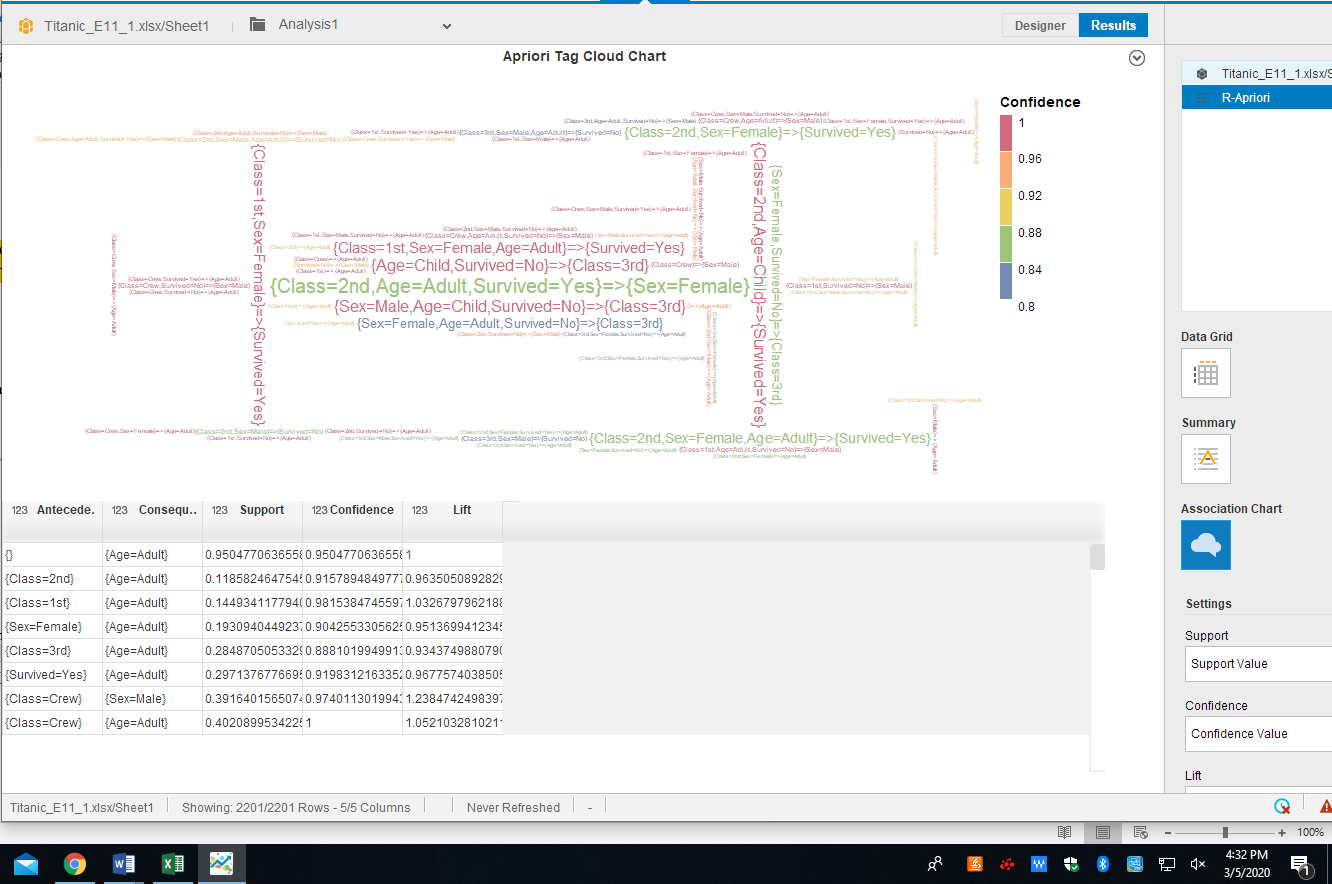
1. The algorithm is now generating the association rules. After the execution is complete, click OK to review the results. On the results screen, Click on the Association Chart.



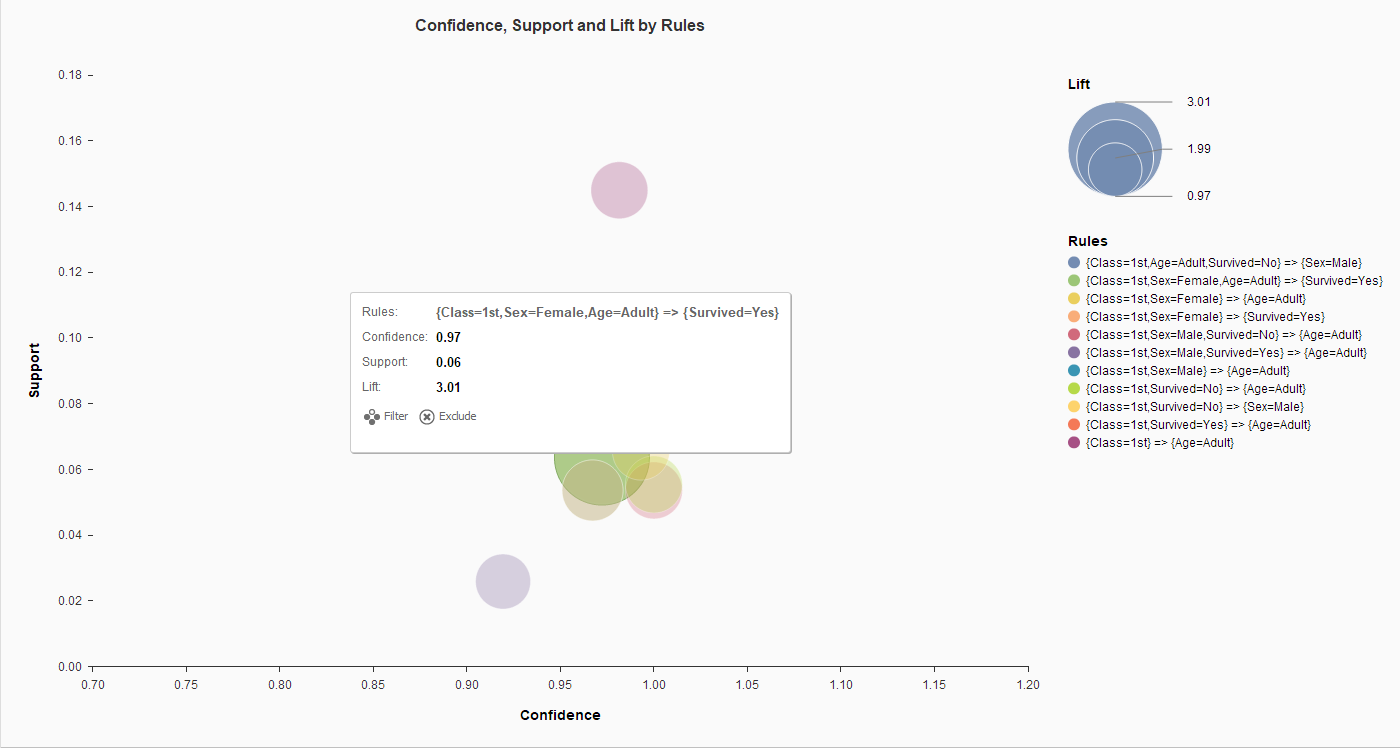
* 1. You see a table of *rules* that were generated.

Question 1: What do you conclude about the second rule? In other words, what does that result mean? Try to say something meaningful about support, confidence and lift.

Antecede is class 2 and consequent is adult. Support is 0.11 – support is very low so we cannot conclude based on this lets look further, confidence is 0.91 – for the 2nd class passengers are likely to be adults and lift is 0.96 – but the Lift has proved that not all passengers are adults in class 2 but children can also be there.

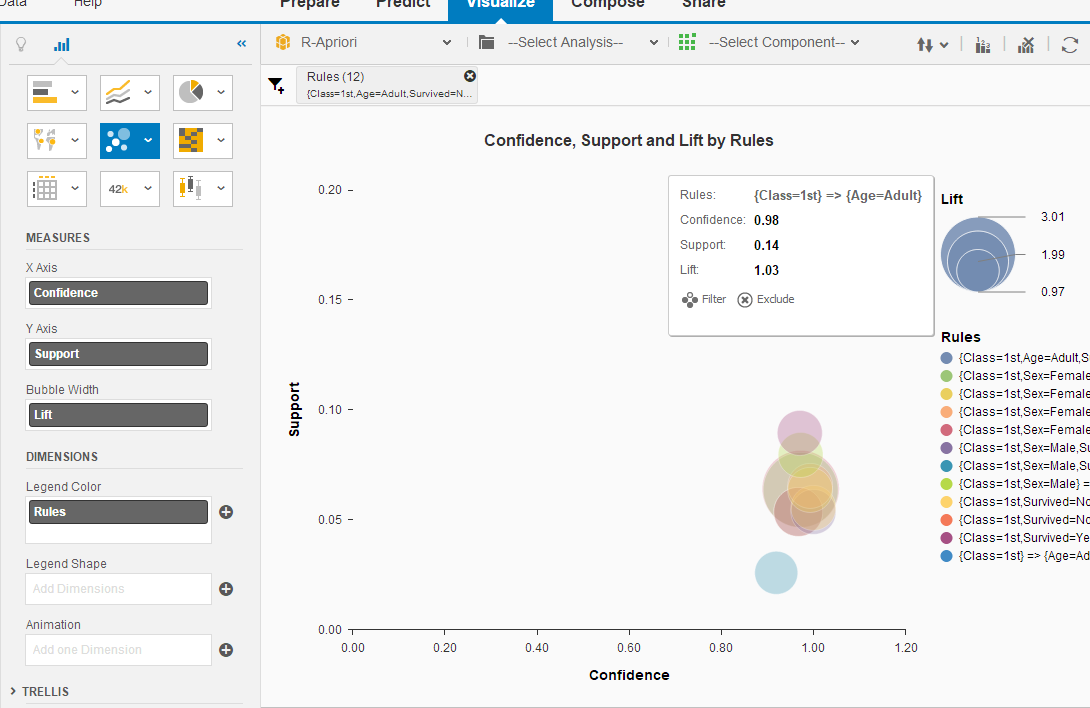


1. Click on *Visualize*
   1. Select component R-Apriori
   2. Convert the attributes *Confidence , Support,* and *Lift* to Measures (by right clicking on them and selecting ‘create a measure’)
   3. Change each measure’s aggregation to *None* (from the default *Sum*). You may also wish to rename each measure.
   4. Create a bubble chart (available under scatter plots). X-Axis – *Support*, Y Axis – *Confidence*, Bubble width – *Lift*
   5. Add the Rules from Attributes to *Dimensions: Legend Color*
   6. You can now see the large bubbles indicating the lift for that rule. Lift indicates the strength of a rule over the random co-occurrence of the independent and the dependent variables, given their individual support.
   7. Filter out all rules that do not involve 1st Class passengers. The result looks like Figure 3 below:

Figure 3

Question 2: Explain the meaning of this highest lift rule. Compare this rule to your earlier rule.

For the highest lift in this graph we can see the value of lift is 1.03 which is greater than 1. So all the 1st class members are likely to be adults.



1. Let’s examine which dimensions are related to survival.
   1. Edit the R-Apriori *properties* by selecting configure settings
   2. Click on *Advanced* tab.
   3. In Rhs Item(s) type: *Survived=No,Survived=Yes* (type without spaces in between)
   4. Choose *Default Appearance: Lhs Items*
   5. In the Performance tab, select Sort Type: *Descending*
   6. Done. Run the analysis again.
   7. View the results
   8. You now see the results for all the Rhs (right-hand side or *consequent*) for Survived (No, Yes). See Figure 2.

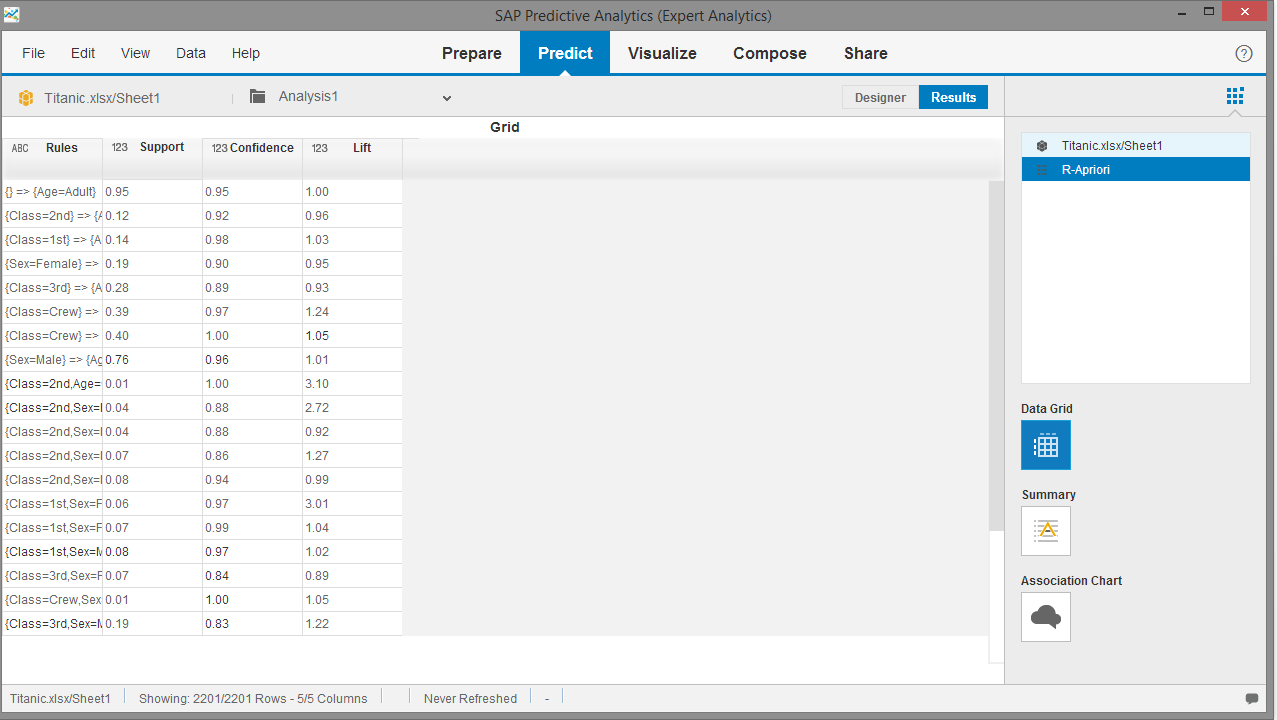


Figure 2: Association Rules

* 1. Click on *Association* *Chart*. Here you can see the results in a tag cloud format.
  2. Save and Close

part 2: sAM’s CLUB products

Nitin Kalé, University of Southern California

Nancy Jones, San Diego State University

# Objective

The objective of this exercise is to generate association rules (or affinity) for products sold at Sam’s Club. This exercise will provide hands-on experience with the big data sets from the Walton College of Business at the University of Arkansas.[[3]](#footnote-3)

# Activities

* Import and prepare data
* Apply data mining algorithms
* Configure predictive models
* Create data visualizations
* Analyze and interpret output from models
* Publish results

Software Prerequisites

* SAP Predictive Analytics 3.0 or newer
* Microsoft Excel

# Scenario: Sam’s Club Dataset

 ****

[Sam's Club](http://samsclub.com/), a division of Wal-Mart Stores, Inc., is a warehouse club that specializes in selling to small businesses. A membership-based store, Sam's Club offers goods and services for consumers and business owners as well as affordable luxury merchandise. Sam's Club keeps prices low by selling merchandise in bulk and at very low profit margins.

The Sam's Club Database contains retail sales information gathered from sales at Sam's Club stores. The process used to gather this information begins with a Sam's Club member gathering all of the items they intend to purchase during the current visit to Sam's Club. The member then proceeds to a register to check out. A Sam's Club associate scans the member's Sam's Club card, at which point a visit number (visit\_nbr) is generated and stored in the store\_visits table. The associate proceeds by scanning each item with a barcode reader. When all of the items have been scanned, summary information about each individual type of product (i.e. 6 packages of sop) purchased during that visit is recorded in the item\_scan table. When payment is tendered for items purchased on that visit, summary information for the total order (transaction time & date, amount spent, number of unique items purchased, etc) is recorded in the store\_visits table. Other tables are used to store information about stores, products, and members.

# Real Data and Data Integrity

The retail sales information in the ***UA\_SAMSCLUB*** database was provided to the Walton College of Business by Wal-Mart Stores, Inc. The database consists of 6 tables with more than 55 million rows populated and ready for use.



This is a gifted dataset that is based on real operational data. Like many real databases, integrity problems may be noted. This can provide a unique opportunity not only to expose students to real data but also to illustrate the effects of data integrity problems.

# SAM’S CLUB DATABASE SCHEMA

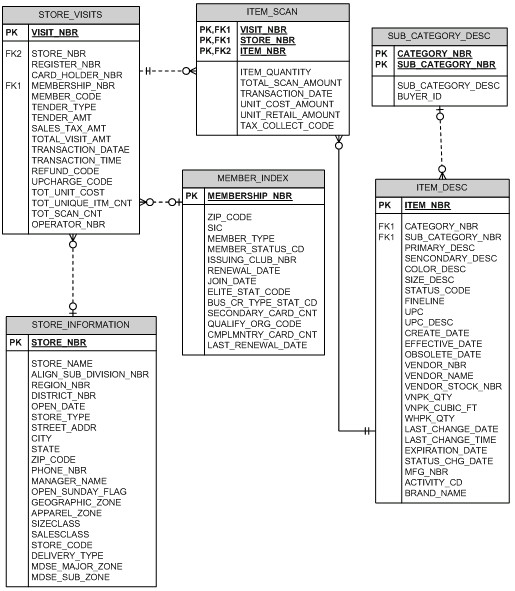


Figure 1: Sam's Club Schema

metadata for Sam's Club database

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Values |
| ACTIVITY\_CD | Activity Code | Y, N |
| BRAND\_NAME | Name of the brand associated with the item | Null, name of brand |
| BUS\_CR\_TYP\_STAT\_CD | Business Credit Type Status Code | 0-10 |
| CARD\_HOLDER\_NBR | Card holder within an account | 1-99 |
| CATEGORY\_NBR | Number assigned to a category of items | Null, 0-99 |
| CMPLMNTRY\_CARD\_CNT | Number of extra cards given to an account | 0-4 |
| COLOR\_DESC | Color description of an item | White, Almond, etc |
| CREATE\_DATE | Date the item was created | Date |
| EFFECTIVE\_DATE | Date the item began to be sold | Date |
| ELITE\_STAT\_CODE |  | 0-4 |
| EXPIRATION\_DATE | Expiration date of an item | Date |
| FINELINE | Combination of category\_nbr & sub\_category\_nbr | 4 digit number |
| ISSUING\_CLUB\_NBR | The club that the member originally joined | 1-150 |
| ITEM\_NBR | The number assigned to every different item for sale | Unique number (PK) |
| ITEM\_QUANTITY | The quantity of a unique item that is scanned |  |
| JOIN\_DATE | Date the member joined the club | Date |
| LAST\_RENEWAL\_DATE | Last date that the member renewed their membership | Date |
| MEMBER\_CODE |  | 1,A,D,E,G,V,W,X,Y |
| MEMBER\_STATUS\_CD |  | A,D,E,T |
| MEMBER\_TYPE |  | 1,A,E,G,V,W,X |
| MEMBERSHIP\_NBR | The number assigned to the member upon joining the club |  |
| MFG\_NBR | Number representing a manufacturer |  |
| OBSOLETE\_DATE | The date an item is no longer sold | Date |
| OPERATOR\_NBR |  |  |
| PRIMARY\_DESC | The description of an item | Teal X-Large etc |
| QUALIFY\_ORG\_CODE |  | Null, 015-3001 |
| REFUND\_CODE | Code to indicate a return transaction | 0 = Not Return, 1= Return |
| REGISTER\_NBR | The register identification number where the transaction took place | 1-85 |
| RENEWAL\_DATE | Date a membership should be renewed | Date |
| SALES\_TAX\_AMT | Tax charged for total visit |  |
| SECONDARY\_CARD\_CNT | Number of cards other than primary card assigned to the membership |  |
| SECONDARY\_DESC | Additional description of an item | Sweatshirt, gift set etc |
| SIC | Standard Industry Classification code | 783700, 443700 etc |
| SIZE\_DESC | Text description of the size of the item, including clothing and non-clothing items | 15CUFT, LARGE, etc |
| STATUS\_CHG\_DATE | The date an item last changed its status code | Date |
| STATUS\_CODE | Whether an item is active or deactive | A = Active, D = Deactive |
| STORE\_NAME | The name of the store |  |
| STORE\_NBR | Store identification number | 1-150 |
| SUB\_CATEGORY\_NBR | The number assigned to a sub\_category of items |  |
| TAX\_COLLECT\_CODE | Purchase taxable or not | 0,1 |
| TENDER\_AMT | The amount tendered for the purchase |  |
| TENDER\_TYPE | Type of payment used | 0 - Cash 1 - Check 2 - Gift Card 3 - Discover 4 - Direct Credit 5 - Business Credit 6 - Personal Credit |
| TOT\_SCAN\_CNT | Total number of scanned items per transaction |  |
| TOT\_UNIQUE\_ITM\_CNT | The number of unique items purchased per transaction | 0-84 |
| TOT\_UNIT\_COST | The cost of the item (scrubbed) |  |
| TOTAL\_SCAN\_AMOUNT | The total number of items scanned per visit number |  |
| TOTAL\_VISIT\_AMT | The total value of the entire transaction |  |
| TRANSACTION\_DATE | Date of the transaction |  |
| TRANSACTION\_TIME | The time of day that the transaction started |  |
| UNIT\_COST\_AMOUNT | Cost/Unit (scrubbed) |  |
| UNIT\_RETAIL\_AMOUNT | Purchase Price/Unit (scrubbed) |  |
| VENDOR\_NBR | The number of the vendor that supplies the item |  |
| VISIT\_NBR | Every time a member goes to the register and has their membership card scanned, this number is then created | 9 digit # |
| VNPK\_CUBIC\_FT | How many cubic feet does a vendor pack take up |  |
| VNPK\_QTY | The quantity of items in a vendor pack |  |
| ZIP\_CODE | The zip-code of the store |  |
| ZIP\_CODE | The zip-code of the member |  |

SAM’S CLUB STAR SCHEMA

For this exercise, the dataset you will use is the Sam’s Club star schema instead of the database schema listed above. The star schema is shown in Figure 2. It consists of one fact table called Item\_Scan\_Fact. There are six dimensions – *Date, Member, Item, Time, Scan Type* and *Store.*

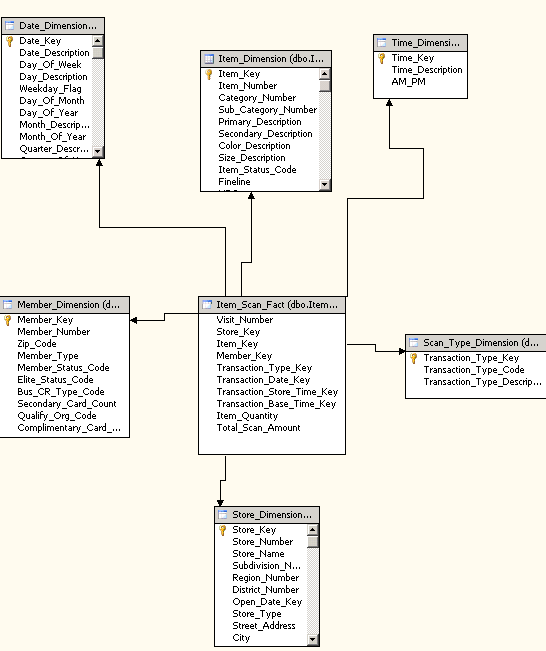


Figure 2: Star Schema for Sam’s Club

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Star Schema Table | | Table Detail | | Number of rows | Range of data |
| *Item Scan* | *Point of Sale item scan* | | *2,553,088* | |  |
| *Date* | *Date of transaction* | | *10,959* | | *01-11-1980 to 12-22-1999* |
| *Member* | *Customer* | | *5,668,375* | |  |
| *Item* | *Product* | | *432,223* | |  |
| *Time* | *Time of day* | | *86,400* | | *1:00:08 am to 9:59:40 pm* |
| *Scan Type* | *Type of scan* | | *2* | | *1 is purchase, 2 is return* |
| *Store* | *Retail Store* | | *150* | | *1-150* |

Table 1: Details of Star Schema Tables

# ASSOCIATION ANALYSIS

Several predictive models are available in SAP Predictive Analytics and more can be integrated from the *R language libraries*. We would like to discover the *association* among products bought within each sales order. These are presented as rules with values for *support*, *confidence* and *lift* for each rule. The algorithm we will use is called ***Apriori*** from the R programming language.

1. We will now use SAP Predictive Analytics (Expert Analytics) to *acquire* Sam’s Club data, and to *visualize* and *discover* association rules.
2. Launch SAP Predictive Analysis using *Start 🡪 All Programs 🡪 SAP Business Intelligence 🡪 SAP Predictive Analytics* or by clicking on the PA shortcut icon on your desktop.
3. Choose Expert Analytics.
4. Acquiring Sam’s Club data
   1. For the association analysis, we are interested in the products bought by customers in each visit. I have already downloaded data of only one day of transactions from the Teradata site and placed it into a .csv file called *SamClub.csv*. Select this file and acquire it.
   2. You should see 34,566 rows.
   3. *Save* your file.
5. Visualization
   1. To start visualizing the data, click on *Visualize*
   2. We would like to chart the number of products bought for each visit. Choose column chart.
   3. In the Dimensions panel, click on *Primary\_Description* settings. *Create a Measure.*
   4. Note that the aggregation for Primary\_Description is Count(Distinct). It cannot be sum or average because it is a non-numeric field.
   5. From *Measure*, drag *Primary\_Description* and drop in to Y axis
   6. In the setting for *Primary\_Description* (under Measures on the right side), Choose *Sort Descending*
   7. From *Dimensions* panel, drag *Visit\_Number* and drop in to X axis
   8. In Figure 3, you see the count of products per visit. You can roll over the bar chart to see individual visits.

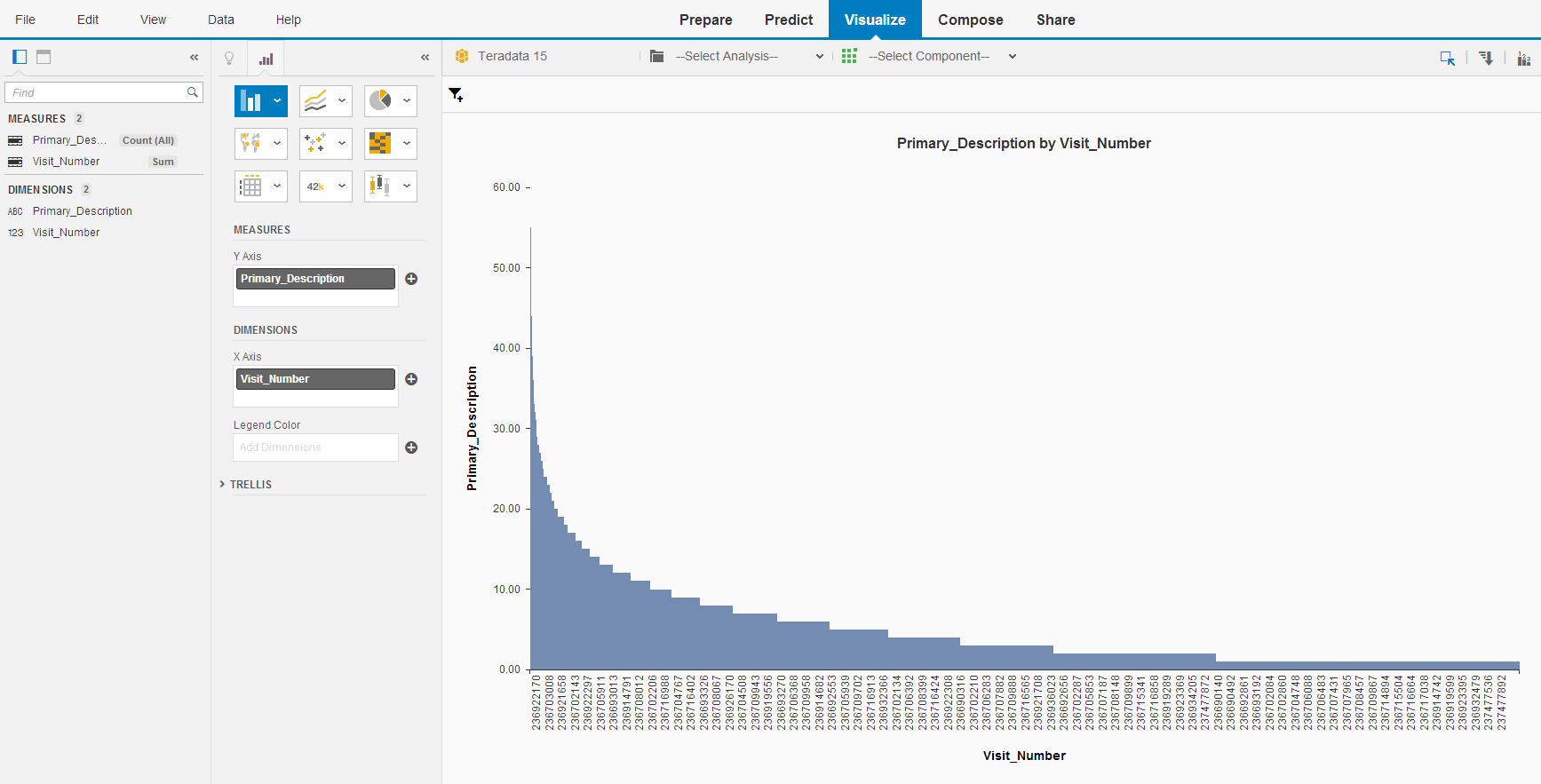


Figure 3: Data Visualization

1. We will now do the association analysis (using *Apriori* algorithm).
   1. Click on *Predict*. You are now in the *Designer* tab.
   2. You will see several *Algorithms* such as Regression, Outliers, Time Series, Decision Trees, Neural Network, Clustering and Association. See Figure 4.

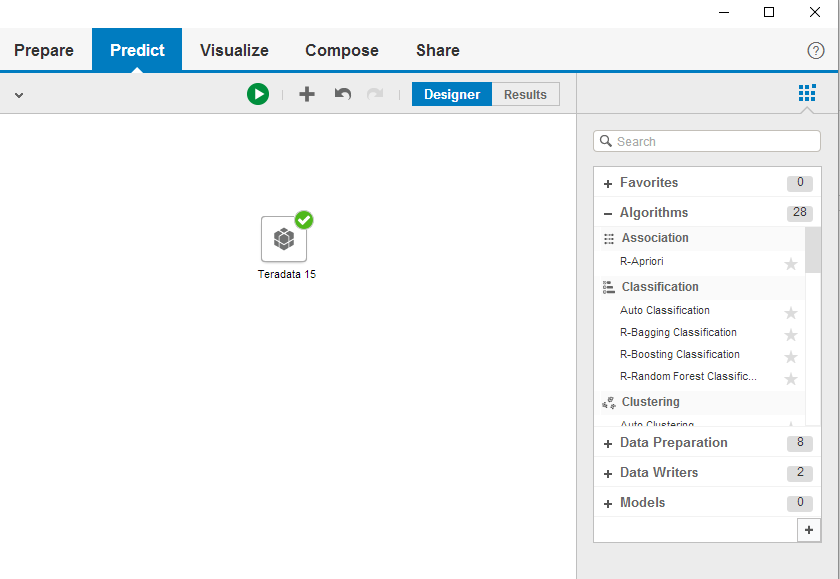


Figure 4: Data mining process

* 1. And you see the data source on the canvas.
  2. Double-click the *R-Apriori* algorithm or drag it to the canvas. The algorithm is automatically connected to the data source.
  3. Roll your mouse over the algorithm and click on *Configure Settings* or choose Configure Settings in the window on the right.
  4. Output Mode is still Rules
  5. Change the Input Data Format to *Transactions*
  6. Item Column – *Primary\_Description*
  7. TransactionID Column – *Visit\_Number*
  8. Support: 0.001, Confidence: 0.1
  9. *Done*
  10. Click *Run*

1. Results
   1. The algorithm is now generating the association rules. After the execution is complete, click OK to review the results.
   2. You see a table of *rules* that were generated as in Figure 5.

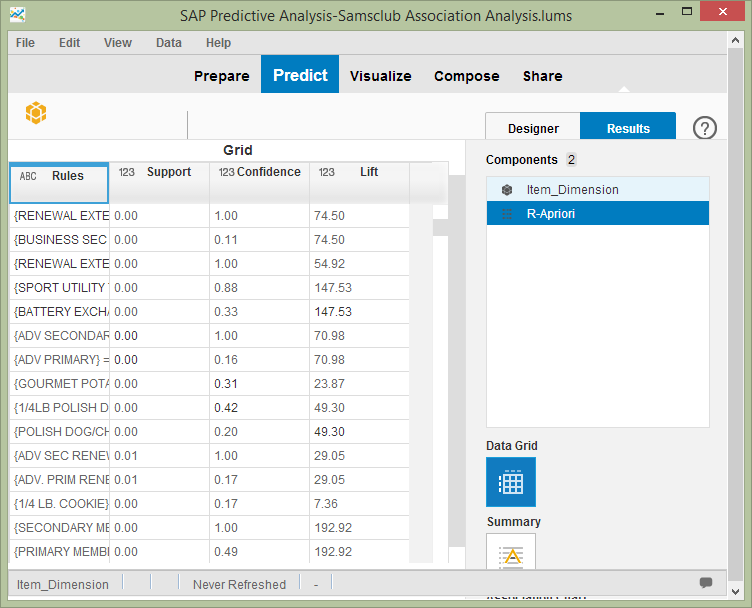


Figure 5: Association Rules

* 1. You see the rules that are reported as leading item(s) => dependent item(s). For each rule you see the support, confidence, and lift ratios.
  2. Look also at the Summary data for this analysis.

Question 1: What is the lift ratio for the rule ¼ lb. cookie -> soft drink, 20 oz.?

* 1. Now look at the Association Chart (tag cloud) generated by this analysis.



* 1. Save.

1. Click on *Visualize*
   1. Select *Component* and choose R-Apriori from the dropdown list*.*
   2. Convert the dimensions *Confidence , Support* and *Lift* to Measures (by clicking on their settings and selecting ‘create a measure’)
   3. Create a bubble chart (available under scatter plots). X-Axis – *Support*, Y Axis – *Confidence*, Bubble width – *Lift*
   4. Add the Rules from Dimensions to *Dimensions: Legend Color*
   5. You can now see the bubbles (Figure 7) indicating the lift for that rule. Lift indicates the strength of a rule over the random co-occurrence of the independent and the dependent variables, given their individual support.
   6. You can *Add Filters* to focus the bubble chart to the rules that you are interested in

Question 2: Select a rule and explain what is meant by support, confidence and lift?

Question 3: Using filters and other analysis, what other helpful results can you find.

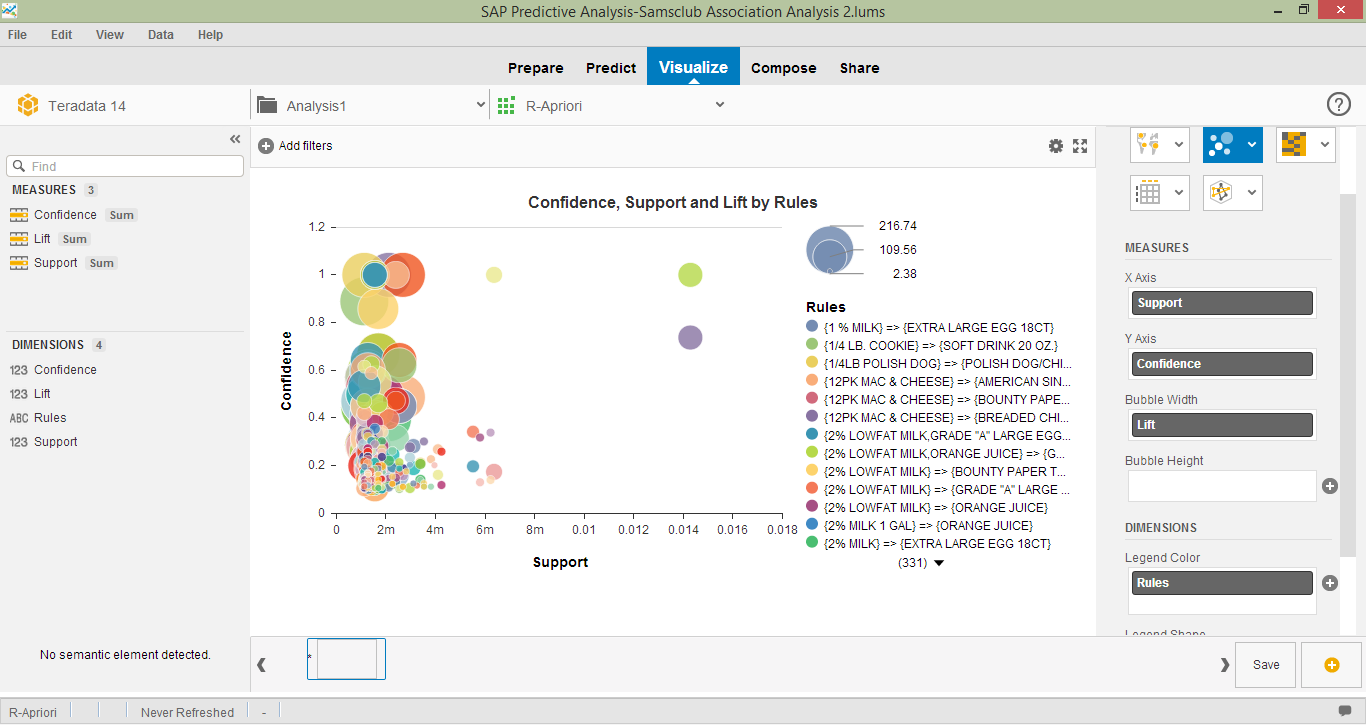


Figure 7: Bubble Chart of Support, Confidence and Lift

1. https://en.wikipedia.org/wiki/RMS\_Titanic [↑](#footnote-ref-1)
2. http://www.rdatamining.com/examples/association-rules [↑](#footnote-ref-2)
3. http://walton.uark.edu/enterprise/ [↑](#footnote-ref-3)