PRODUCT DOCUMENTATION

bigml®

Association Discovery

BigML Inc.
2851 NW 9th, Corvallis,
OR 97330, U.S.
https://bigml.com
info@bigml.com

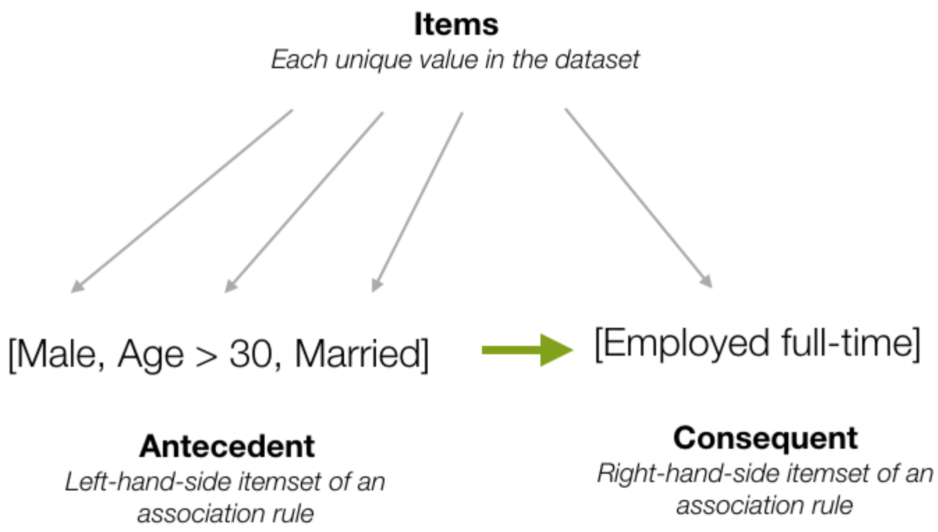
1. INTRODUCTION

Useful concepts

Association Discovery is a well-known method to find out associations among values in high-dimensional datasets. It can discover meaningful relations among values across thousands of variables.

BigML acquired Magnum Opus from Professor Webb in July 2015 to combine the best-in-class association discovery technology with BigML's easy-to-use platform.

Association rule example:


Items
Each unique value in the dataset
[Male, Age > 30, Married]
[Employed full-time]
Antecedent
Left-hand-side itemset of an association rule
Consequent
Right-hand-side itemset of an association rule

The above rule indicates that if the person is male, he is more than 30 years old and he is married, it is more likely that he is also a full-time employee. Note that association rules look for co-occurrence between items and don't imply causality. In this example, being a full-time employee is not a consequence of being a 30-year-old married male.

Measures

Given the Association rule [A → C] where A is the Antecedent itemset of the rule and C is the Consequent, find below some measures definitions:

- **Support**: the proportion of instances in the dataset which contain an itemset. The support of an association is the portion of instances in the dataset which contain the rule's antecedent and rule's consequent together over the total number of instances (N) in the dataset. It gives a measure of the importance of the rule:

$$\textbf{Support}\ \text{(itemset)} = \frac{|\ \text{instances} \in \mathbf{D}\ \textbf{(itemset)} \subseteq \text{instance}\ |}{N}$$

$$\textbf{Support}\ (A \rightarrow C) = \textbf{Support}\ (A \cup C)$$

- **Coverage**: the support of the Antecedent of an association rule, i.e. the portion of instances in the dataset which contain the Antecedent itemset. It measures how often a rule could be applied.

$$\textbf{Coverage}\ (A \rightarrow C) = \textbf{Support}\ (A)$$

- **Confidence** (or Strength): the percentage of instances which contain the consequent and antecedent together over the number of instances which only contain the antecedent. It can be thought of as an estimate of the probability that the consequent will occur in case the antecedent occurs. Confidence is computed using the support of the association rule over the coverage of the antecedent.

$$\textbf{Confidence}\ (A \rightarrow C) = \frac{\textbf{Support}\ (A \rightarrow C)}{\textbf{Support}\ (A)}$$

- **Leverage**: it measures the difference between the probability of the rule and the expected probability if the items were statistically independent. Leverage ranges between [-1, 1]. A leverage of 0 suggest there is no association between the items. Higher leverage for positive values suggests stronger positive association between the antecedent and consequent. Negative values for leverage suggest a negative relationship.

$$\textbf{Leverage}\ (A \rightarrow C) = \textbf{Support}\ (A \rightarrow C) - (\textbf{Support}\ (A) \times \textbf{Support}(C))$$

- **Lift**: how many times more often antecedent and consequent occur together than expected if they were statistically independent. For example, a lift of 5 for the following rule (onions → potatoes) means that buying onions makes it 5 times more likely to buy potatoes. Lift is always a real positive number. A lift of 1 suggests there is no association between the items. A lift between 0 and 1 indicates a negative correlation. Higher values suggest stronger relationships between the items.

$$\textbf{Lift}\ (A \rightarrow C) = \frac{\textbf{Support}\ (A \rightarrow C)}{(\textbf{Support}\ (A) \text{ x } \textbf{Support}(C))}$$

There are no measures more important than others and there are no general thresholds to take into account as rule-of-thumbs. You will have to analyze your results according to your main goals which may be different depending on the problem you are trying to solve. For example, you may be interested in very frequent associations so you will have to pay more attention to the rule's support. Or maybe you want to find some more infrequent associations but with a stronger relationship between the items (a higher lift for example). Usually it is not one single measure, but the combination and coherence of all measures what makes a rule stronger than others.

2. PREPARING YOUR DATA

How to structure your data

It is quite usual in Association Discovery to have a great number of different values per instance. For example a commercial dataset containing the transactions with all the products bought by customers; or medical datasets containing all the medicines prescribed per patient. This kind of data can be organized in several ways, see below an example:

transaction_id: 12345; product_A; product_B; product_C; product_D; product_F
transaction_id: 67890; product_A; product_G
transaction_id: 67890; product_B; product_C; product_H

This transactional data can be structured in several ways:

Binary Data Representation

Transaction ID	product_A	product_B	product_C	product_D	product_F	product_G	product_H
12345	1	1	1	1	1	0	0
67890	1	0	0	0	0	1	0
67890	0	1	1	0	0	0	1

Vertical Data Layout

Transaction ID	1st Product	2nd Product	3rd Product	4th Product	5th Product
12345	product_A	product_B	product_C	product_D	product_F
67890	product_A	product_G			
67890	product_B	product_C	product_H		

Horizontal Data Layout

Transaction ID	Products
12345	product_A, product_B, product_C, product_D, product_F
67890	product_A, product_G
67890	product_B, product_C, product_H

The correct way to structure your data for Association Discovery is the one shown in the **Horizontal Data Layout** example. If you use any of the other options you may get lots of missing values in your association rules. By using this data structure the field "Products" will be considered an **items field** and each product will be a unique item. Remember that you need to separate your items by a unique separator (see items fields description below).

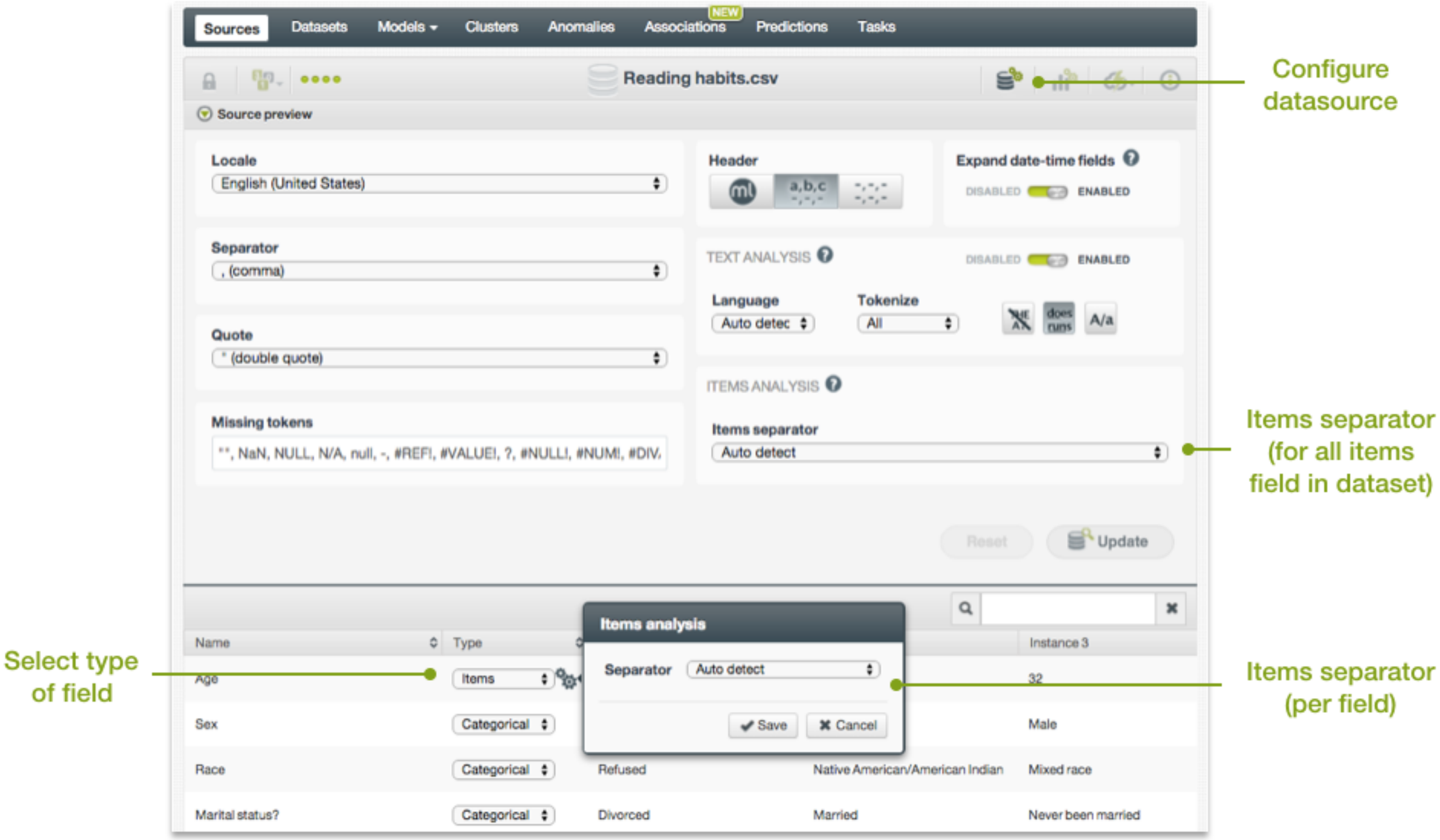
Create datasource

Upload your datasource with one of the several options available in BigML. The datasource can contain any number of fields and instances.

Associations support any type of fields:

- Categorical: each unique category will be considered a different item.
- Numerical: numeric fields will be discretized in different segments at the moment of the Associations creation. For example, a numeric field with values ranging from 0 to 600 may be split into 3 segments: segment 1 → [0, 200), segment 2 → [200, 400), segment 3 → [400, 600]. Each different segment will be considered an item. You can configure discretization parameters before creating an Association (see "Discretization" in the parameters table below).
- Text: for text fields each unique term will be considered a separate item by default.
- Items: items fields usually have many different categorical values per instance separated by non-alphanumeric characters. BigML supports up to 10,000 different items per field (for greater number of different items it will be considered a text field). You need to separate your items by a unique separator and BigML will try to

automatically detect it. For example, for the following itemset {hot dog; milk, skimmed; chocolate}, the best separator is the semicolon which yields three different items: “hot dog”, “milk, skimmed” and “chocolate”. By configuring your source you can also specify which separator you want to set for your items using the item separator selector (see screenshot below).



Locale: English (United States)

Separator: , (comma)

Quote: " (double quote)

Missing tokens: "", NaN, NULL, N/A, null, -, #REF!, #VALUE!, ?, #NULL!, #NUM!, #DIV.

Header

Expand date-time fields: DISABLED ENABLED

TEXT ANALYSIS: DISABLED ENABLED

Language: Auto detec

Tokenize: All

ITEMS ANALYSIS

Items separator: Auto detect

Separator: Auto detect

Name	Type			Instance 3
Age	Items			32
Sex	Categorical			Male
Race	Categorical	Refused	Native American/American Indian	Mixed race
Marital status?	Categorical	Divorced	Married	Never been married

Associations can find relations between values (items) across fields and within the same field.

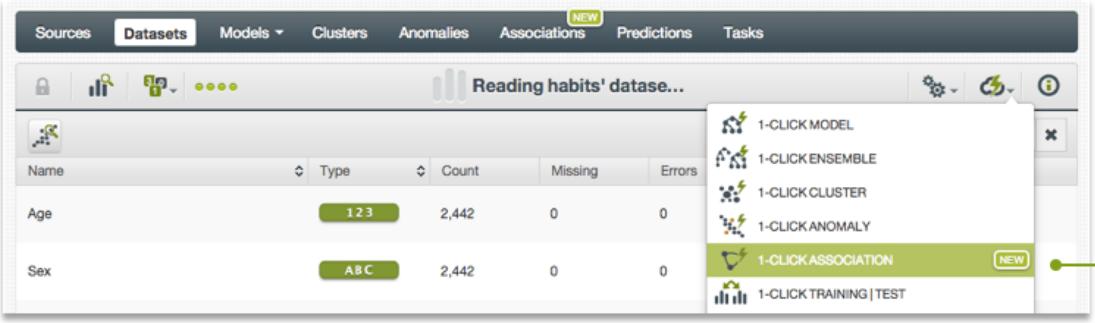
Create a dataset

You need to create a dataset from your datasource to create Associations. You can use the 1-click dataset or the configuration option from the datasource menu.

3. CREATING ASSOCIATIONS

1-click Associations

You can use the 1-click option from your dataset view to create Associations. Default values will be set for all parameters (see default values in the parameters table below).



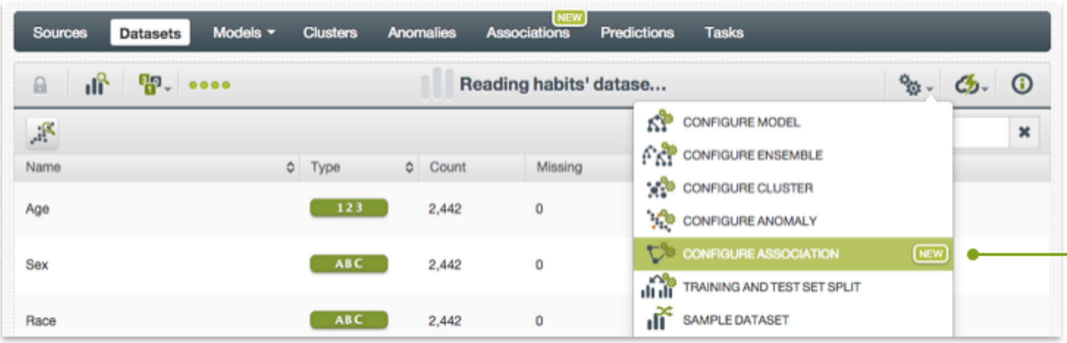
Name	Type	Count	Missing	Errors
Age	123	2,442	0	0
Sex	ABC	2,442	0	0

1-click Association

Depending on your dataset size, it may take a bit of time to get your Associations results. You can always check the progress from the Associations section in your Dashboard.

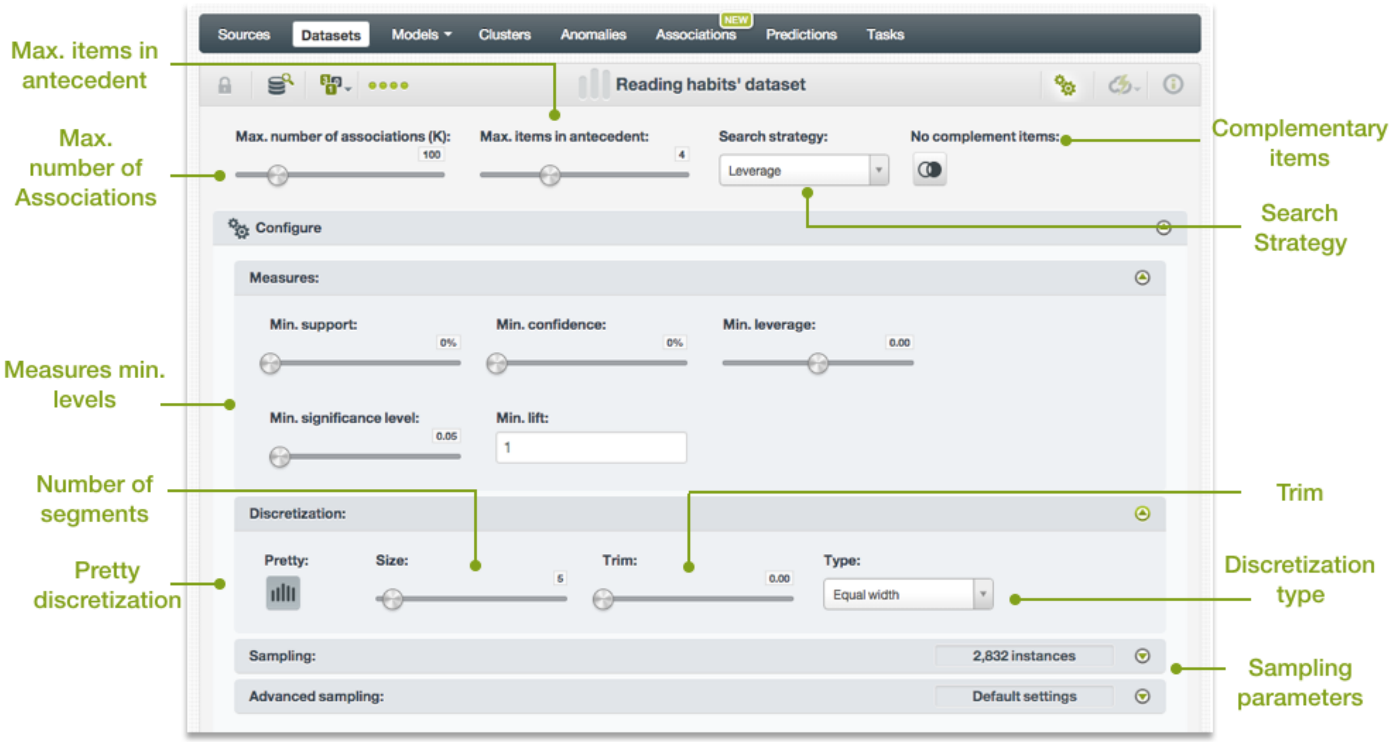
Configuring Associations

You can configure the following parameters to create Associations:



Name	Type	Count	Missing
Age	123	2,442	0
Sex	ABC	2,442	0
Race	ABC	2,442	0

Configure Association



Reading habits' dataset

Max. number of associations (K): 100
Max. items in antecedent: 4
Search strategy: Leverage
No complement items:

Min. support: 0%
Min. confidence: 0%
Min. leverage: 0.00
Min. significance level: 0.05
Min. lift: 1

Pretty:
Size: 5
Trim: 0.00
Type: Equal width

Sampling: 2,832 instances
Advanced sampling: Default settings

	Parameter	Description	Available Values	Default Value
Association Rules	Max. number of associations (k)	Maximum number of associations to be discovered. Higher numbers may take longer to calculate.	1 - 500	100
	Max. number of items in antecedent	Maximum number of items to be considered within the antecedent itemset. The consequent itemset will always contain one item	1 - 10	4
	Search strategy	Select the measure to prioritize the associations discovered. You can use the leverage, lift, coverage, support and confidence so rules with higher values for the measure chosen will be prioritized. Leverage is one of the measures that usually gives relevant results in most of the cases. Two other measures frequently used are confidence and lift. The strategy chosen should be set according to your main purpose.	Leverage, Lift, Support, Confidence and Coverage	Leverage

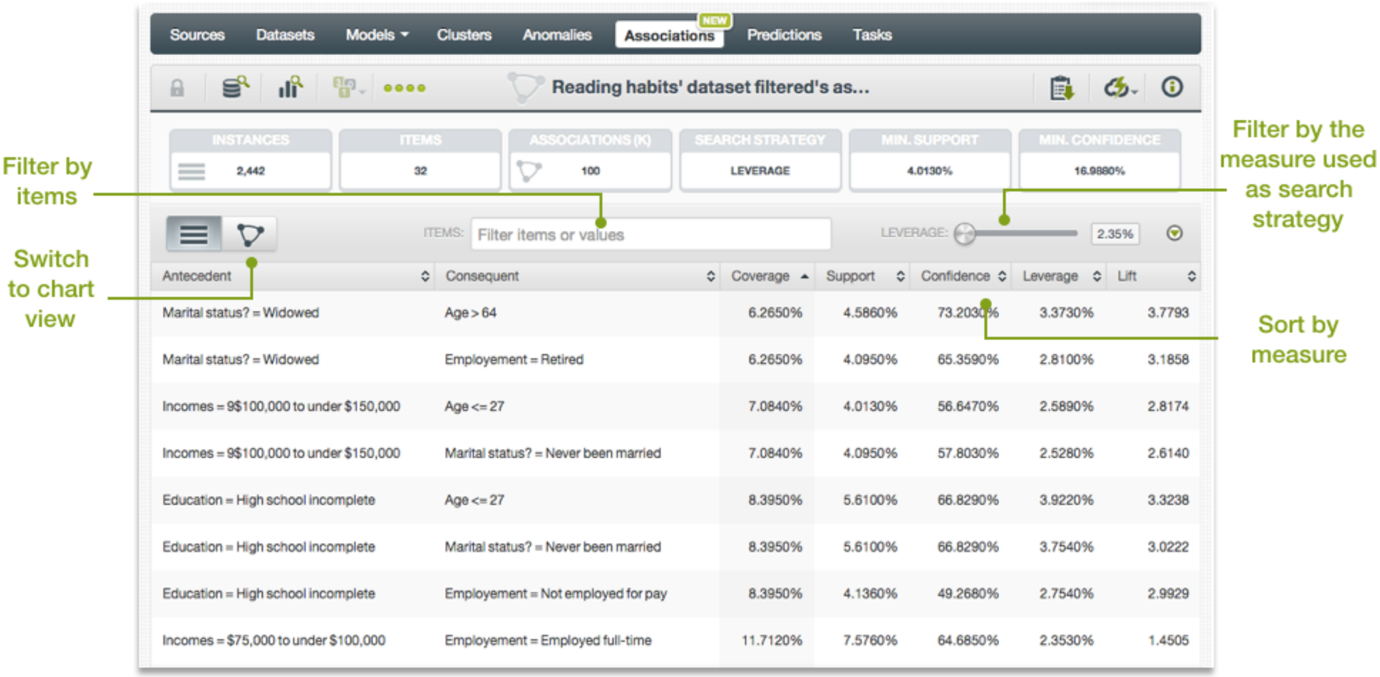
	Complementary items	By enabling this parameter, complementary items are also taken into account. For example, for the item (coffee) the complement (NOT coffee) will be also considered to find rules. Complementary items will be represented with an exclamation mark (coffee —> coffee !)	True / False	False
Measures	Minimum support	Associations below this support will be discarded	(0%, 100%)	0%
	Minimum confidence	Associations below this confidence will be discarded	(0%, 100%)	0%
	Minimum leverage	Associations below this leverage will be discarded	(-100%, 100%)	0
	Significance level	Statistical tests are applied to control the risk of finding spurious associations. The significance level is the maximum level of risk you are willing to take to discover a spurious association	(0, 1)	0.05
	Minimum lift	Associations below this lift will be discarded	Positive real number	1
Discretization	Pretty	By enabling "pretty", segment boundaries for numerical fields will be set in a way that are easy to read. For example, instead of "segment > 20.678" you will get "segment > 20". If pretty is enabled the parameter type (equal width or population) may be ignored for some segments and the specified size may act as a maximum	True / False	True
	Size	The number of equal segments. If pretty is enable this value acts as a maximum size	Postive real number	5
	Trim	The portion of the overall population that may be removed from either tail of the distribution. Setting this parameter to 1% usually gives good results.	(0, 10%)	0
	Type	Whether the field is discretized using an equal width or equal population strategy for each segment	Equal width, Equal population	Width
Sampling	Rate	In case you want to use just a sample of your dataset, this parameter allows you to set the proportion of the dataset to create the association	(0, 100%)	1

Advanced sampling	Range	It allows you to specify a linear subset of the dataset instances to create the association (example: from instance 5 to instance 1,000). The rate you set before will be computed over the range configured	(1, max. rows in dataset)	(1, max. rows in dataset)
	Sampling	It allows you to set a random sampling or a deterministic sampling so the random-number generator will always use the same seed, producing repeatable results	True/ False	False
	Replacement	Sampling with replacement allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance is selected exactly once	True/ False	False
	Out_of_bag	If an instance is not selected as part of a deterministic sampling, it is considered out-of-bag. Enabling this will select only the out-of-bag instances for the currently defined sample. This can be useful for splitting a dataset into training and testing subsets. It is only selectable when a sample is deterministic and the sample rate is less than 100%	True/ False	False

4. VISUALIZING ASSOCIATIONS

Once Associations are created you will get a **table** that summarizes all the rules discovered. Each row of the table contains a rule which is composed of two parts: the antecedent itemset (with 1 or more items) and the consequent itemset (it will always contain one item). For each rule you will find five different measures that describe the relationship between both parts of the rule (see "Measures" for more details).

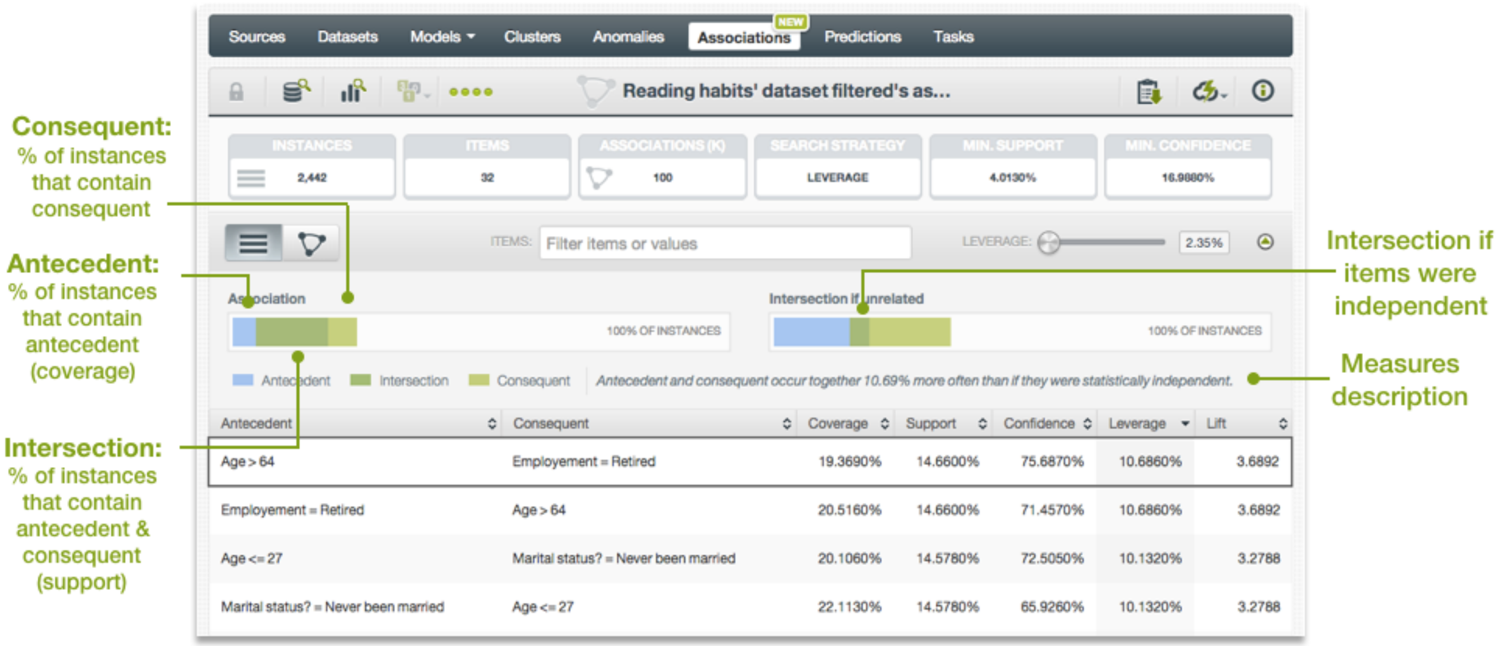
You will see two **filters** at the top of the table: a text box to filter rules by items names and a slider to filter rules by the measure used as the search strategy to find the rules.



INSTANCES	ITEMS	ASSOCIATIONS (K)	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE
2,442	32	100	LEVERAGE	4.0130%	16.9880%

Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
Marital status? = Widowed	Age > 64	6.2650%	4.5860%	73.2030%	3.3730%	3.7793
Marital status? = Widowed	Employement = Retired	6.2650%	4.0950%	65.3590%	2.8100%	3.1858
Incomes = 9$100,000 to under $150,000	Age <= 27	7.0840%	4.0130%	56.6470%	2.5890%	2.8174
Incomes = 9$100,000 to under $150,000	Marital status? = Never been married	7.0840%	4.0950%	57.8030%	2.5280%	2.6140
Education = High school incomplete	Age <= 27	8.3950%	5.6100%	66.8290%	3.9220%	3.3238
Education = High school incomplete	Marital status? = Never been married	8.3950%	5.6100%	66.8290%	3.7540%	3.0222
Education = High school incomplete	Employement = Not employed for pay	8.3950%	4.1360%	49.2680%	2.7540%	2.9929
Incomes = $75,000 to under $100,000	Employement = Employed full-time	11.7120%	7.5760%	64.6850%	2.3530%	1.4505

If you select an association in the table you will get a **graphic representation** of the rule with two diagrams, one indicating the actual intersection between the antecedent and consequent itemsets (Association diagram) and the other one indicating the intersection if both itemsets were independent. This graphic allows you to get a visual overview of the rules importance. The **blue bar** represents the portion of instances in the dataset that contain the antecedent items (coverage) and the **green bar** is the portion of instances that contain the consequent itemset. The **intersection** between them is the portion of instances that contain both itemsets (in the Association diagram it is the support of the rule). You can also get a visual insight of the rule's lift and leverage (represented by the differences of the actual Association intersection and the intersection if both itemsets were independent). If you mouse over the different measures in the table, the **legend** in the diagrams view will change to show the corresponding description for each measure.



Reading habits' dataset filtered's as...

INSTANCES	ITEMS	ASSOCIATIONS (K)	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE
2,442	32	100	LEVERAGE	4.0130%	16.9880%

Antecedent and consequent occur together 10.69% more often than if they were statistically independent.

Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
Age > 64	Employment = Retired	19.3690%	14.6600%	75.6870%	10.6860%	3.6892
Employment = Retired	Age > 64	20.5160%	14.6600%	71.4570%	10.6860%	3.6892
Age <= 27	Marital status? = Never been married	20.1060%	14.5780%	72.5050%	10.1320%	3.2788
Marital status? = Never been married	Age <= 27	22.1130%	14.5780%	65.9260%	10.1320%	3.2788

Consequent: % of instances that contain consequent

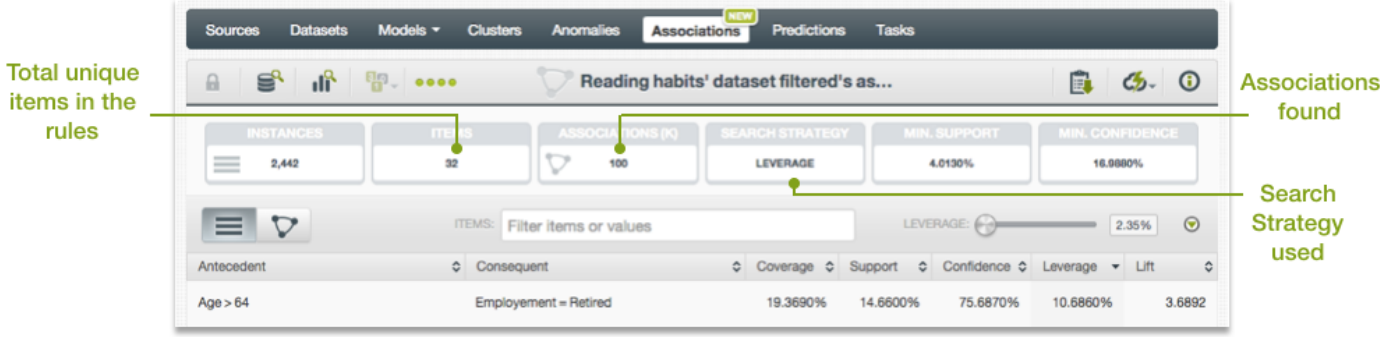
Antecedent: % of instances that contain antecedent (coverage)

Intersection: % of instances that contain antecedent & consequent (support)

Intersection if items were independent

Measures description

At the top of the Associations results you will also find a **summary** which contains the number of instances in the dataset, the number of unique items, the number of Associations discovered, the search strategy used and the minimum support and minimum confidence found in the rules.



Reading habits' dataset filtered's as...

INSTANCES	ITEMS	ASSOCIATIONS (K)	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE
2,442	32	100	LEVERAGE	4.0130%	16.9880%

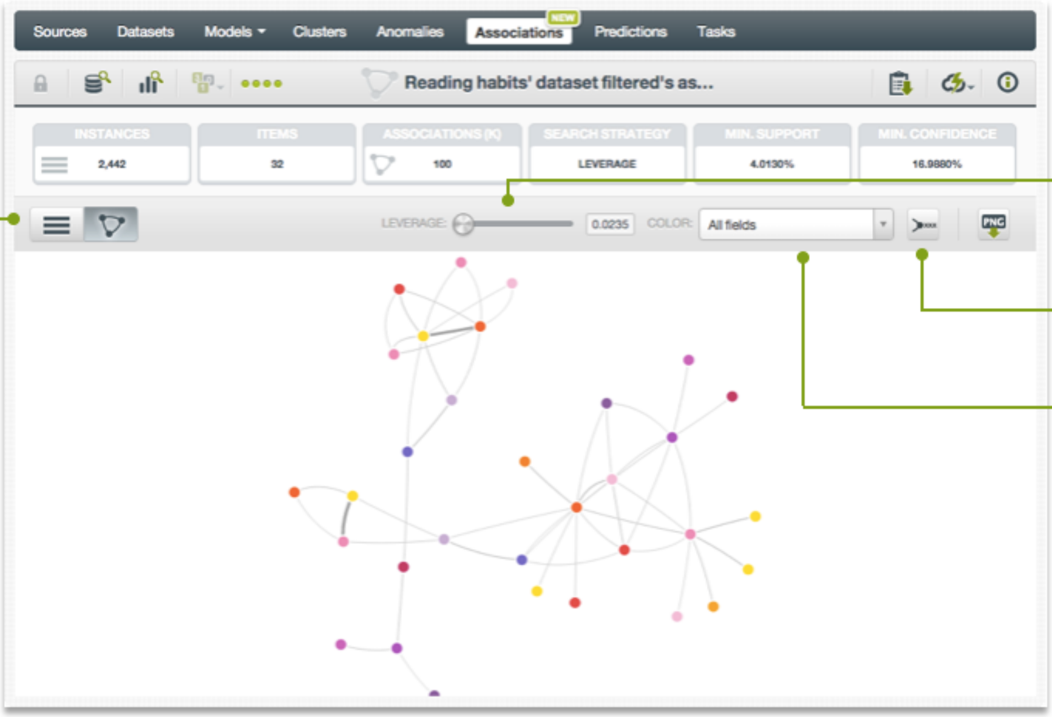
Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
Age > 64	Employment = Retired	19.3690%	14.6600%	75.6870%	10.6860%	3.6892

Total unique items in the rules

Associations found

Search Strategy used

You can also switch the view to visualize the rules in a **network chart**. The chart will give you an overview of which items are connected to others. You can show and hide items labels, filter according to the measure used as the search strategy and color the chart points by field.



Reading habits' dataset filtered's as...

INSTANCES	ITEMS	ASSOCIATIONS (K)	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE
2,442	32	100	LEVERAGE	4.0130%	16.9880%

LEVERAGE: 0.0235 COLOR: All fields

Return to table view

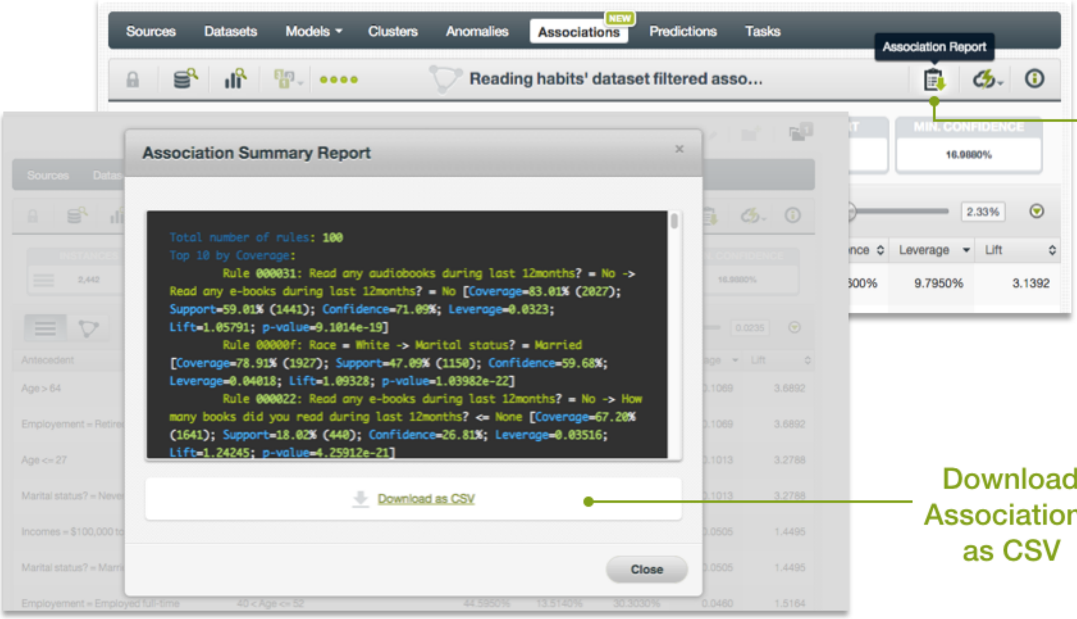
Search strategy measure filter

Show/hide labels

Color by field

5. EXPORTING ASSOCIATIONS

You can export your Associations as a CSV file and the chart in PNG format.



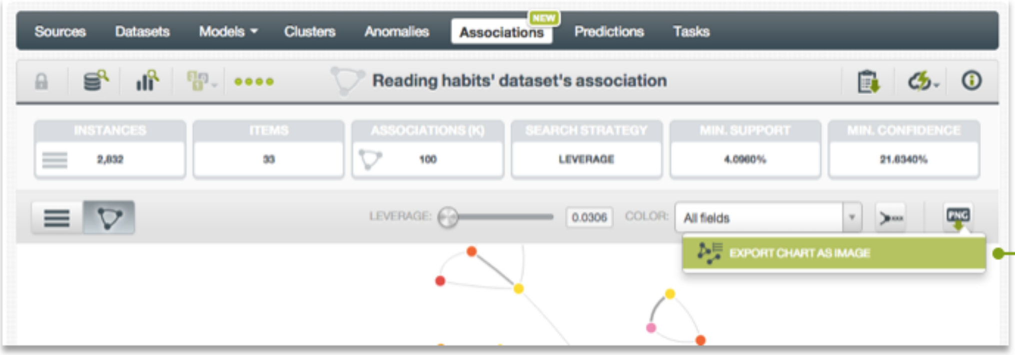
Association Summary Report

```
Total number of rules: 100
Top 10 by Coverage:
        Rule 000031: Read any audiobooks during last 12months? = No ->
Read any e-books during last 12months? = No [Coverage=83.01% (2027);
Support=59.01% (1441); Confidence=71.09%; Leverage=0.0323;
Lift=1.05791; p-value=9.1014e-19]
        Rule 00000f: Race = White -> Marital status? = Married
[Coverage=78.91% (1927); Support=47.09% (1150); Confidence=59.68%;
Leverage=0.04018; Lift=1.09328; p-value=1.03982e-22]
        Rule 000022: Read any e-books during last 12months? = No -> How
many books did you read during last 12months? <= None [Coverage=67.20%
(1641); Support=18.02% (440); Confidence=26.81%; Leverage=0.03516;
Lift=1.24245; p-value=4.25912e-21]
```

Download as CSV

Association Report

Download Associations as CSV



Reading habits' dataset's association

INSTANCES: 2,832
ITEMS: 33
ASSOCIATIONS (K): 100
SEARCH STRATEGY: LEVERAGE
MIN. SUPPORT: 4.0960%
MIN. CONFIDENCE: 21.6340%

LEVERAGE: 0.0306
COLOR: All fields

Export Chart as PNG

bigml®