ACE Tutorial – Set Visualisation

The QuantiCode project is developing novel data mining and visualization tools and techniques, which will transform people's ability to analyse quantitative and coded longitudinal data. Such data are common in sectors such as health (e.g., electronic health records), local government (e.g., service provision) and retail (e.g., product sales). The project is funded by the Engineering and Physical Sciences Research Council (grant ref. EP/N013980/1; EP/K503836/1) and, through the Leeds Institute for Data Analytics (LIDA), supported by the Medical Research Council (MR/L01629X/1) and Economic and Social Research Council (ES/L011891/1).

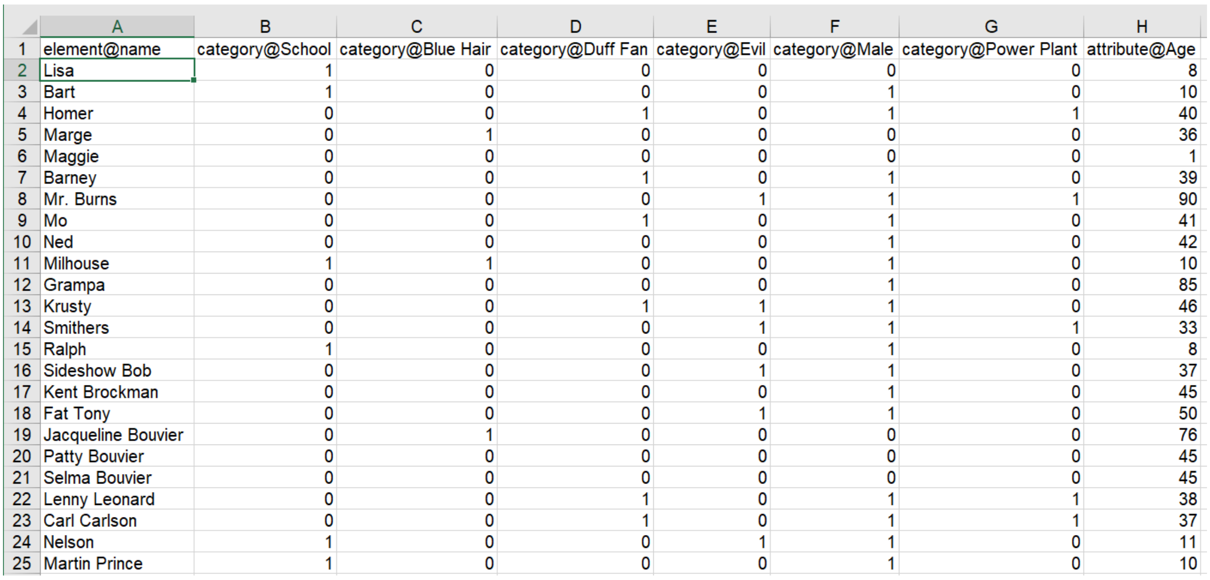
1. Introduction

This tutorial introduces the set visualisation functionality of ACE (Analysis of Combinations of Events) by describing the input data formats that ACE supports, and then providing two worked examples. The first example focuses on the high-level functionality of ACE and uses a publicly available dataset based upon The Simpsons. The second example covers some of the advanced functionality of ACE by analysing a publicly available dataset that provides information about movies.

2. Understanding the input file

In this section we learn more about the input file formats accepted by ACE. The files can have comma, tab, semi-colon, colon, space or pipe ('|') separators.

The first tutorial example (see Section 3.1, below) uses a dataset about characters in the TV show "The Simpsons". The original dataset can be found at https://github.com/VCG/upset/tree/master/data/simpsons and, even though the file is named "simpsons.csv", it actually uses tab separators. To make the file compatible with ACE you need to download the original file and edit is using a text editor, Excel or other software to prefix the name of the first column by "element@", the next six columns by "category@", and the last column by "attribute@" (see Figure 1).



	A	B	C	D	E	F	G	H
1	element@name	category@School	category@Blue Hair	category@Duff Fan	category@Evil	category@Male	category@Power Plant	attribute@Age
2	Lisa	1	0	0	0	0	0	8
3	Bart	1	0	0	0	1	0	10
4	Homer	0	0	1	0	1	1	40
5	Marge	0	1	0	0	0	0	36
6	Maggie	0	0	0	0	0	0	1
7	Barney	0	0	1	0	1	0	39
8	Mr. Burns	0	0	0	1	1	1	90
9	Mo	0	0	1	0	1	0	41
10	Ned	0	0	0	0	1	0	42
11	Milhouse	1	1	0	0	1	0	10
12	Grampa	0	0	0	0	1	0	85
13	Krusty	0	0	1	1	1	0	46
14	Smithers	0	0	0	1	1	1	33
15	Ralph	1	0	0	0	1	0	8
16	Sideshow Bob	0	0	0	1	1	0	37
17	Kent Brockman	0	0	0	0	1	0	45
18	Fat Tony	0	0	0	1	1	0	50
19	Jacqueline Bouvier	0	1	0	0	0	0	76
20	Patty Bouvier	0	0	0	0	0	0	45
21	Selma Bouvier	0	0	0	0	0	0	45
22	Lenny Leonard	0	0	1	0	1	1	38
23	Carl Carlson	0	0	1	0	1	1	37
24	Nelson	1	0	0	1	1	0	11
25	Martin Prince	1	0	0	0	1	0	10

Figure 1: Simpsons Dataset with Prefixes (simpsons - Format 1.csv).

The prefixes have the following purpose:

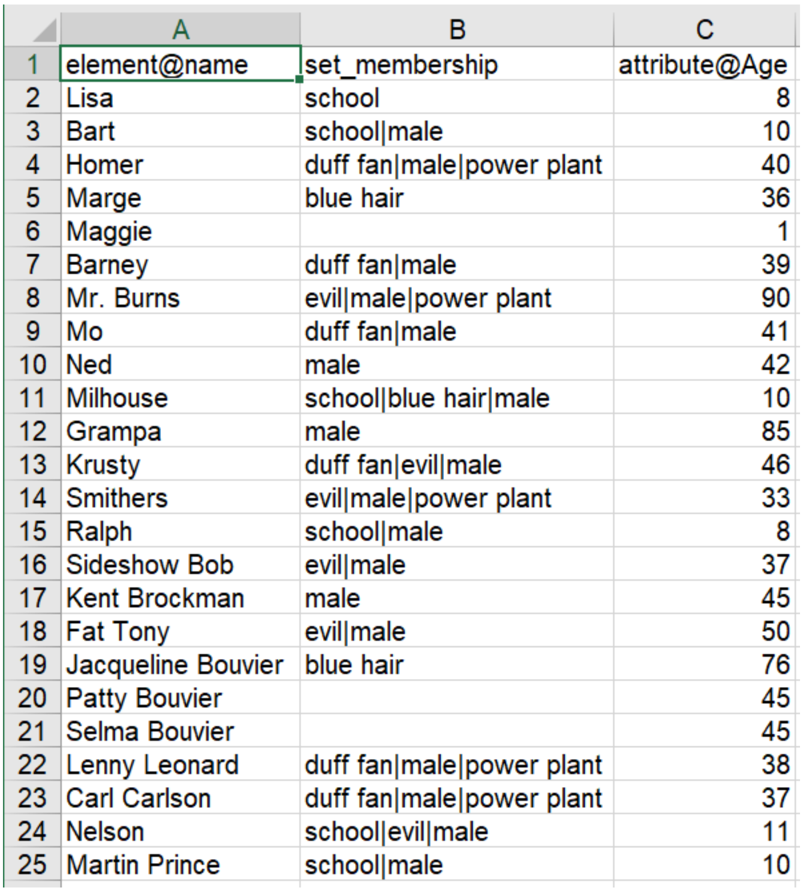
element@: The name of the element within the set/s. This prefix is optional.

category@: The set membership, expressed as either 1 for belonging or 0 for not belonging to the set. This prefix is required in order for ACE to automatically move to set visualization mode with this type of file.

attribute@: Additional properties of the element of the set. In the example above the characters age. This prefix is optional.

Caveat: In a dataset used by ACE for set analysis, all columns must both have headers and be annotated or it is assumed that missing data analysis is intended instead.

A shorter more concise form of the same file can be seen in Figure 2, whereby a **set_membership** column is used with a piped list ('|' separators) to denote the different sets to which the element belongs. This is in place of the **category@** prefix.



	A	B	C
1	element@name	set_membership	attribute@Age
2	Lisa	school	8
3	Bart	school|male	10
4	Homer	duff fan|male|power plant	40
5	Marge	blue hair	36
6	Maggie		1
7	Barney	duff fan|male	39
8	Mr. Burns	evil|male|power plant	90
9	Mo	duff fan|male	41
10	Ned	male	42
11	Milhouse	school|blue hair|male	10
12	Grampa	male	85
13	Krusty	duff fan|evil|male	46
14	Smithers	evil|male|power plant	33
15	Ralph	school|male	8
16	Sideshow Bob	evil|male	37
17	Kent Brockman	male	45
18	Fat Tony	evil|male	50
19	Jacqueline Bouvier	blue hair	76
20	Patty Bouvier		45
21	Selma Bouvier		45
22	Lenny Leonard	duff fan|male|power plant	38
23	Carl Carlson	duff fan|male|power plant	37
24	Nelson	school|evil|male	11
25	Martin Prince	school|male	10

Figure 2: A concise version of the same dataset with the "set_membership" column. That format is not used in this tutorial, but has been successfully used to analyse retail data that contained hundreds of thousands of customer transactions and more than 1 million different products[1]. The concise format indicates all of the products that were bought in a transaction in the set_membership column, instead of requiring the input file to have a separate column for each of the 1+ million products.

In the event prefixes are not used, the default assumption is that missing data analysis is intended. This however can be changed.

1. Go to "File" menu and select "Import data" (see Figure 1).
2. In newly opened window, navigate to the "ACE" folder and then the "datasets" folder. This time select the "simpsons.csv" and click "Open".
3. In the "Header presence" dialog box select "Yes" and click "OK".
4. A "Data Import Success" dialog box should open, providing a summary of the imported dataset. Click "OK".
5. Go to the "Mode" menu and select "Set Visualisation".
6. An option will now be displayed to select the columns relating to the set membership (See Figure 3).

[1] Adnan, M., & Ruddle, R.A. (2018). A set-based visual analytics approach to analyze retail data. Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA). https://eprints.whiterose.ac.uk/131939/.

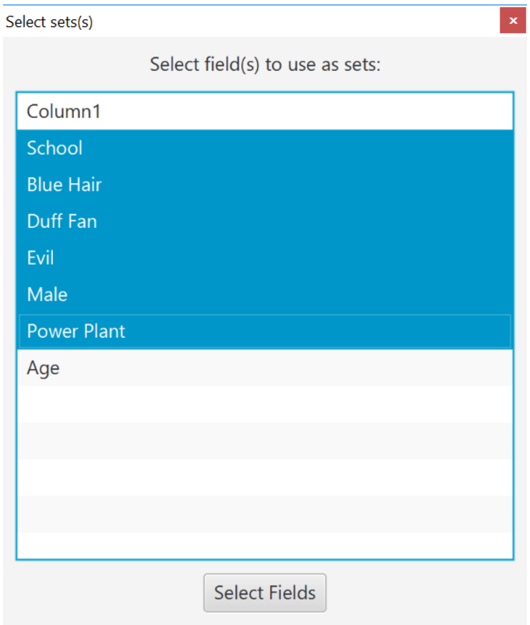


Figure 3: Selection of fields to use for set membership.

7. Select the sets “School” to “Power Plant” as shown in Figure 3. A range of items can be selected at once using the Shift key, and Ctrl key for selecting/deselecting individual items.
8. Once this is done Click “Select Fields”. This automatically performs the “Perform Computation” action which was previously shown in Section 3.1.

Note: Column1 is shown above in Figure 3, because the column header is blank in the original dataset. In Figure 1 and Figure 2, the first column’s header is annotated and given the value “name”.

3. Example 1 - Analysing set data with The Simpsons Dataset

The aim of this worked example is to familiarise users with:

- The prefixes used to identify element attribute and categorical data.
- Importing a dataset into ACE and computing the set intersections
- Analysing the field-level set membership using the “Set bar chart”.
- Analysing the combination-level intersections using the “Intersection heatmap”.
- Analysing unexpected intersections using the “data slicing (via intersection hiding)” functionality.
- Explaining unexpected set intersections using the “data mining” functionality.

The synthetic dataset used in this example has 24 records and 8 fields, and is identical to the publicly available Simpson’s dataset apart from the column headers, which have been modified as shown in see Figure 1. The first field represents the elements name followed by 6 fields showing set membership with values of either 0 or 1 and a final field that gives an attribute of the element called Age.

3.1. Import data and compute set intersections

1. Double click on “ACE.jar” to run the ACE.
2. Go to “File” menu and select “Import data” (see Figure 1).

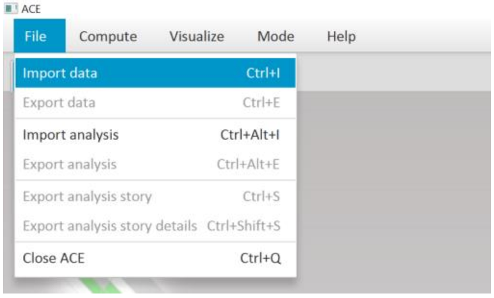
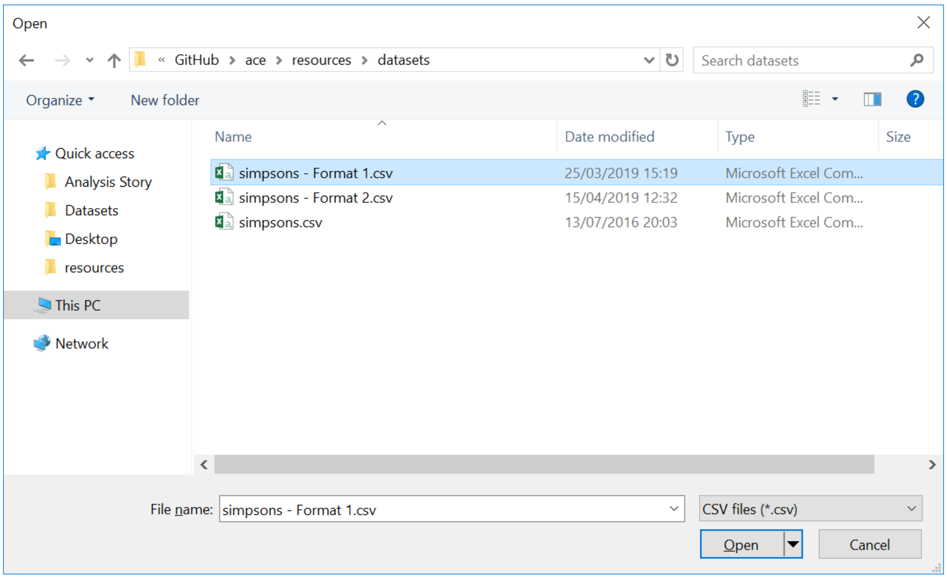

ACE
File
Compute
Visualize
Mode
Help
Import data Ctrl+I
Export data Ctrl+E
Import analysis Ctrl+Alt+I
Export analysis Ctrl+Alt+E
Export analysis story Ctrl+S
Export analysis story details Ctrl+Shift+S
Close ACE Ctrl+Q

Figure 4: Data import menu.

3. In newly opened window, navigate to the “ACE” folder and then the “datasets” folder. Select “Simpsons – Format 1.csv” and click “Open” (see Figure 5).



Name	Date modified	Type	Size
simpsons - Format 1.csv	25/03/2019 15:19	Microsoft Excel Com...	
simpsons - Format 2.csv	15/04/2019 12:32	Microsoft Excel Com...	
simpsons.csv	13/07/2016 20:03	Microsoft Excel Com...	

File name: simpsons - Format 1.csv

CSV files (*.csv)

Figure 5: Open file window.

4. In the “Header presence” dialog box select “Yes” and click “OK”.
5. A “Data Import Success” dialog box should open, providing a summary of the imported dataset. Click “OK”.
6. Go to the “Mode” menu and ensure that “Set Visualisation” is selected.
7. Go to the “Compute” menu and select “Perform Computation” (see Figure 6).

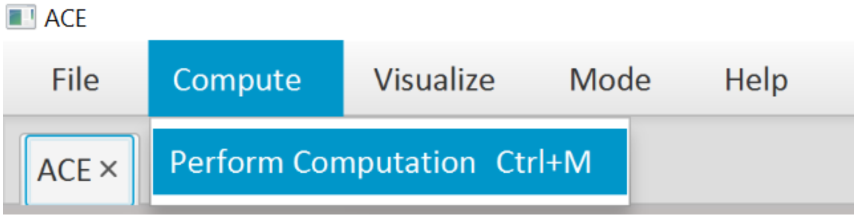


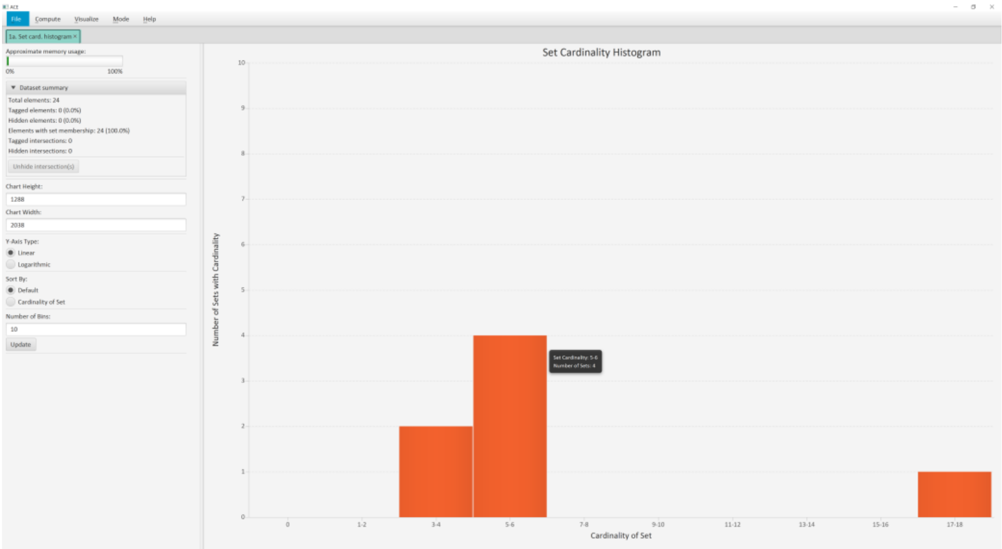
Figure 6: Perform computation menu.

8. A “Computed intersections successfully” dialog box should appear. Click “OK”. This completes the data import and set membership computation in ACE.

Note: In the event the original file has no prefixes then set visualization mode can be manually set in the top menu (at step 6). In this case the category fields must be specified upon changing mode. Step 7 which performs the computation is then run automatically. This is illustrated in Section 2.

3.1. Field-Level Set Membership – Set Cardinality Histogram and Set Bar Chart

Upon the successful computation of set membership, ACE automatically creates a “Set Cardinality Histogram” to act as a start point for analysing set membership within the dataset (see Figure 7), showing the number of sets that have a given cardinality. A new “Set Cardinality Histogram” may be created at any time by selecting “Set Cardinality Histogram” from the Visualize menu.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 24
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 24 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Unhide intersection(s)

Chart Height: 1288

Chart Width: 2038

Y-Axis Type:
● Linear
○ Logarithmic

Sort By:
● Default
○ Cardinality of Set

Number of Bins: 10

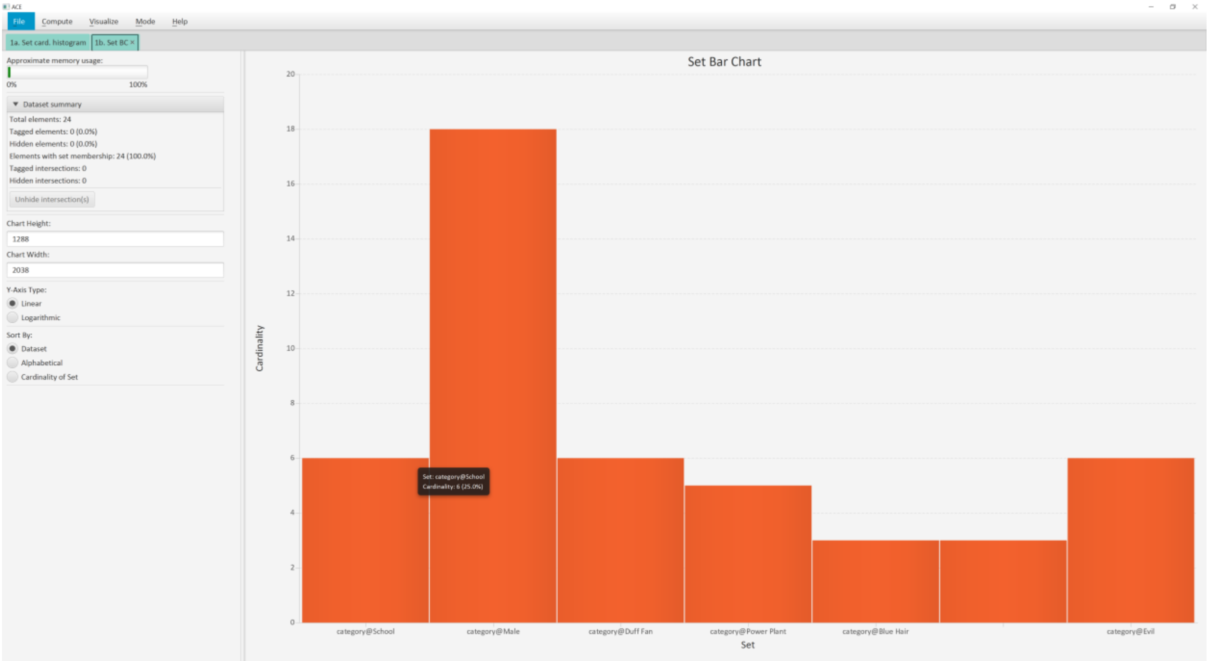
Update

Set Cardinality: 5-6
Number of Sets: 4

Figure 7: Set Cardinality summary at field level.

A slightly deeper view can be seen with the set bar chart as shown in Figure 8, where the cardinality of each set is shown, along with the percentage of all elements which belong to the set in the tooltip. This visualization is created by selecting Visualise on the menu bar and then selecting “Set Bar Chart”.

Note: that the empty set is shown as a column in the bar chart with a blank space.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 24
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 24 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Unhide intersection(s)

Chart Height: 1288

Chart Width: 2038

Y-Axis Type:
● Linear
○ Logarithmic

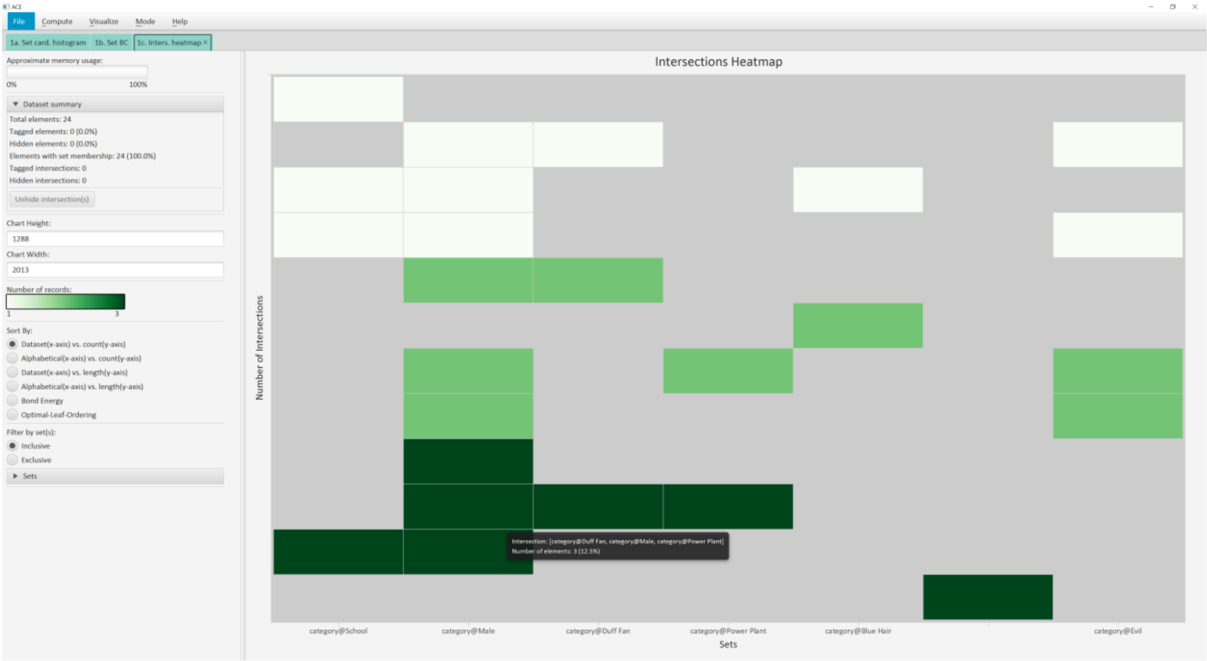
Sort By:
● Dataset
○ Alphabetical
○ Cardinality of Set

Set: category@School
Cardinality: 6 (25.0%)

Figure 8: A set bar chart with a tooltip.

3.1. Intersection-level set analysis - Intersection heatmap

A common form of analysis is to use the heatmap as shown in Figure 9, with the categories highlighted associated with Homer Simpson.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 24
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 24 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Chart Height: 1288

Chart Width: 2013

Number of records: 1 3

Sort By:
● Dataset(x-axis) vs. count(y-axis)
○ Alphabetical(x-axis) vs. count(y-axis)
○ Dataset(x-axis) vs. length(y-axis)
○ Alphabetical(x-axis) vs. length(y-axis)
○ Bond Energy
○ Optimal-Leaf-Ordering

Filter by set(s):
● Inclusive
○ Exclusive

▶ Sets

Intersection: [category@Duff Fan, category@Male, category@Power Plant]
Number of elements: 3 (12.5%)

Figure 9: A heat map showing how the sets overlap. Highlighted is Duff fans, that are male and at the power plant.

4. Example 2 - Analysing set data with the Movies Dataset

This worked example aims to familiarise users with some of the advanced functionality of ACE, which includes but is not limited to:

- Entropy analysis and grouping features
- Illustration of "importing and exporting analysis"

We used a publicly available dataset in this worked example, which can be downloaded from the following link: https://github.com/VCG/upset/tree/master/data/movies. This dataset has 3,883 records and 21 fields. The first field in the dataset gives the movies name (e.g., "Toy Story"), the second field gives the attribute the Release Date (e.g., "1995"). The remaining fields apart from the last two list the type of film (e.g., "Action", "Adventure" and "Horror"). The final two fields provides further attributes of the film namely the AvgRating and the number of Watches.

To make the file compatible with ACE you need to download the original file and edit is using a text editor, Excel or other software to:

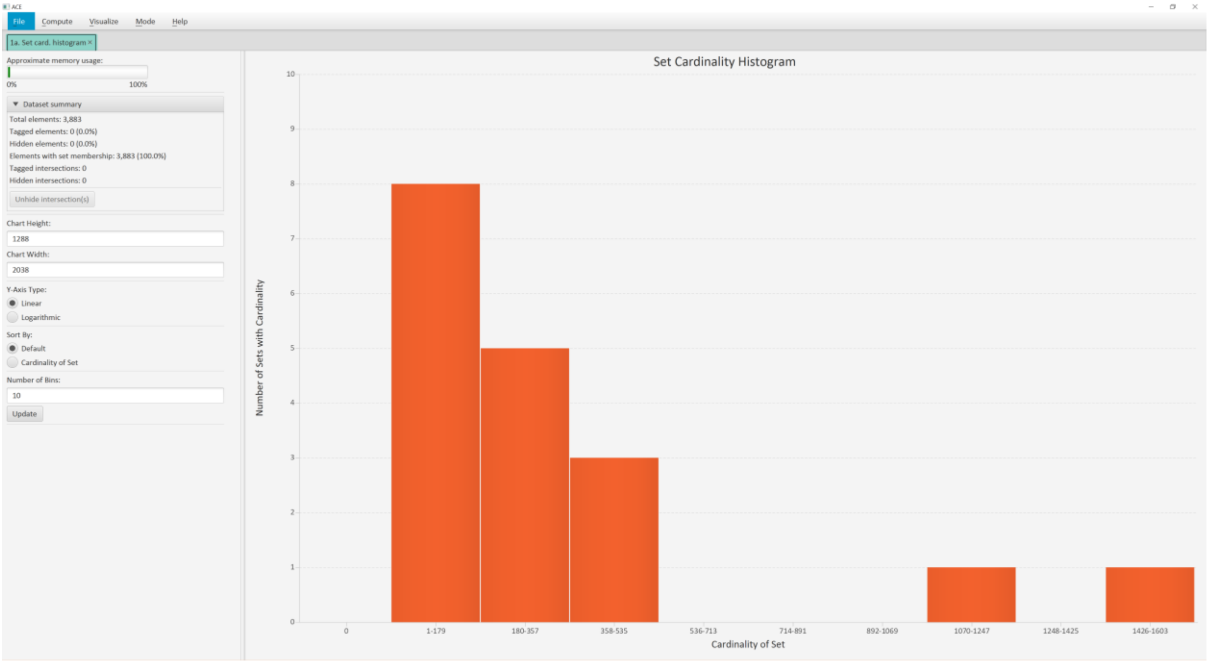
- Prefix the first column ("Name") by "element@"
- Prefix the "Release Date", "AvgRating" and "Watches" columns by "attribute@"
- Prefix all of the other columns by "category@"

To load the dataset and compute the set intersections:

1. Go to "File" menu and select "Import data" (see Figure 1).
2. In the newly opened window select the file type: "Semi-colon separated files (*.*)"
3. Now navigate to the "ACE" folder and then the "datasets" folder. Select "movies - Format 1 - semi-colon.txt" and click "Open".
4. In the "Header presence" dialog box select "Yes" and click "OK".
5. A "Data Import Success" dialog box should open, providing a summary of the imported dataset. Click "OK".
6. Go to the "Mode" menu and ensure that "Set Visualisation" is selected.
7. Go to the "Compute" menu and select "Perform Computation" (see Figure 6).
8. A "Computed intersections successfully" dialog box should appear. Click "OK". This completes the data import and set membership computation in ACE.

4.1. Field-level Set Membership – set cardinality histogram and set bar chart

Once the intersections have been computed successfully, the set cardinality histogram will be shown, as shown in Figure 10. Once we have done this click on the "Visualize" menu and select "Set Bar Chart". The set bar chart is shown in Figure 11.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 3,883 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Chart Height: 1288
Chart Width: 2038

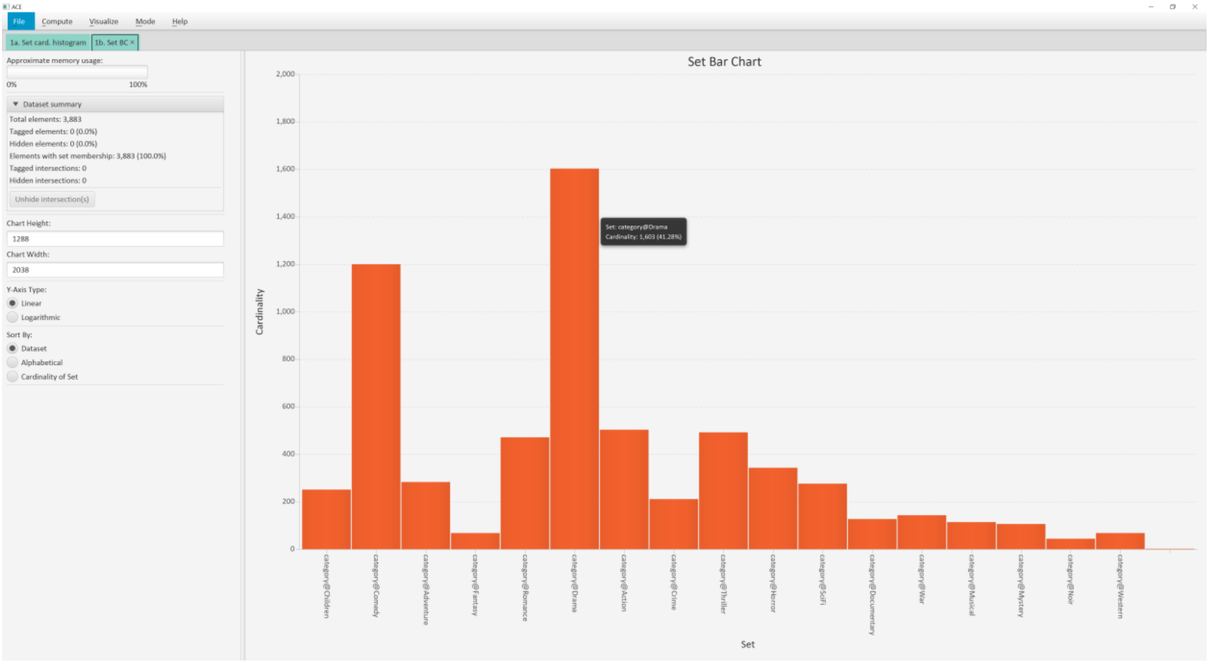
Y-Axis Type:
● Linear
○ Logarithmic

Sort By:
● Default
○ Cardinality of Set

Number of Bins: 10

Figure 10: Set Cardinalty historgram.

Looking at Figure 11 we can again find the most common set within the database i.e. the most common type of film which in this case is Drama.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 3,883 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Chart Height: 1288
Chart Width: 2038

Y-Axis Type:
● Linear
○ Logarithmic

Sort By:
● Dataset
○ Alphabetical
○ Cardinality of Set

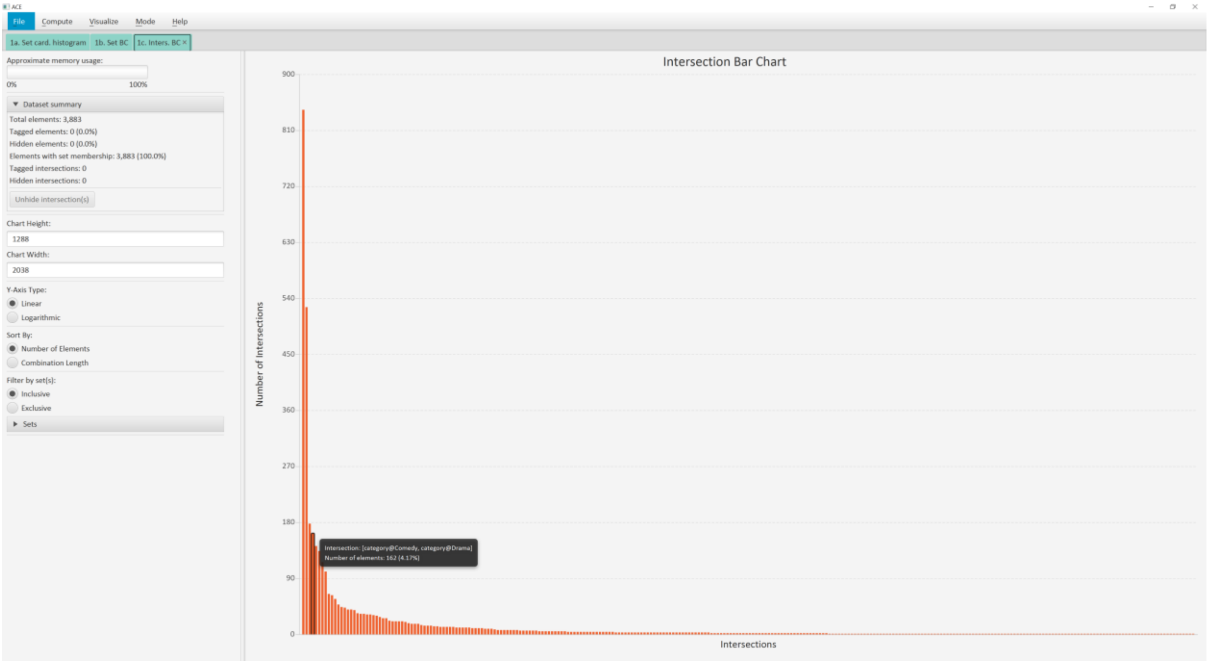
Set: category@Drama
Cardinality: 1,603 (41.28%)

Figure 11: The most common type of film.

In the event we want to save this data we can right click and click on the option “Save Graph Data”.

4.1. Set Intersection & Multiple Set Membership – Intersection Bar Chart

Elements of course can belong to multiple sets, essentially allowing sets to intersect. In the case of movies, a movie can for example be classed as both a comedy and a drama (Figure 12). The intersection bar chart can be created by selecting “Intersection Bar Chart” from the Visualize menu.



Approximate memory usage:
0% 100%

▼ Dataset summary
Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 3,883 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Unhide intersection(s)

Chart Height: 1288

Chart Width: 2038

Y-Axis Type:
● Linear
○ Logarithmic

Sort By:
● Number of Elements
○ Combination Length

Filter by set(s):
● Inclusive
○ Exclusive

▶ Sets

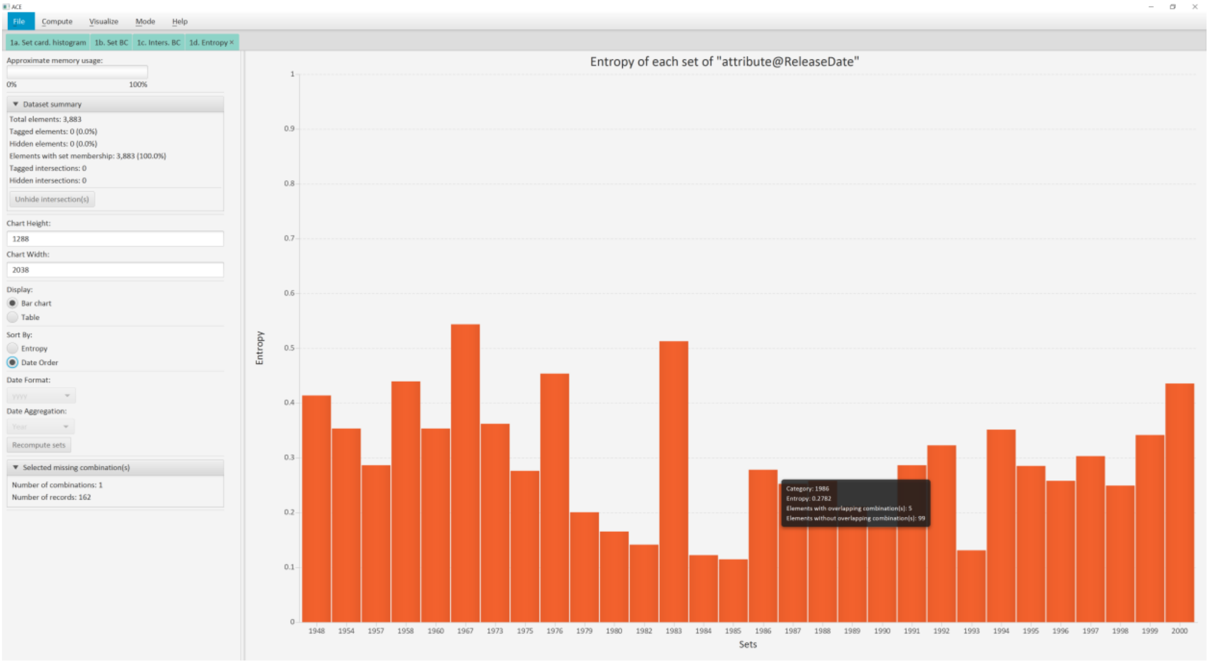
Intersection: [category@Comedy, category@Drama]
Number of elements: 162 (4.17%)

Figure 12: Illustration of set intersection as shown in the Intersection Bar Chart.

It is possible to perform analysis based upon entropy within this dataset.

1. Select the 4th most popular combination (i.e., “Comedy” and “Drama”) – with 162 elements).
2. Right-click and select “Explain combination(s) - Entropy” from the context menu.
3. Select “attribute@ReleaseDate” in the field selection dialog box and click OK. This creates an “Entropy bar chart” showing the correlation of categories in “attribute@ReleaseDate” field with the records of selected set intersection (see Figure 13). The entropy graph has been sorted in Date order. The tooltip shows that in “1986” there are 5 elements that have this intersection (i.e., 271 products that are both “comedy” and “drama”), whilst 99 elements do not have this same intersection.

Note: In the event the date field contains both month and year data the end user can specify the format of the date in the control panel and can perform aggregation based upon the following options "Month/Year", "Year", "Day of Week". Figure 13 shows these options but disables them as the date value only covers the movies’ release year.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 3,883 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Chart Height: 1288

Chart Width: 2038

Display:
● Bar chart
○ Table

Sort By:
○ Entropy
● Date Order

Date Format: yyyy

Date Aggregation: Year

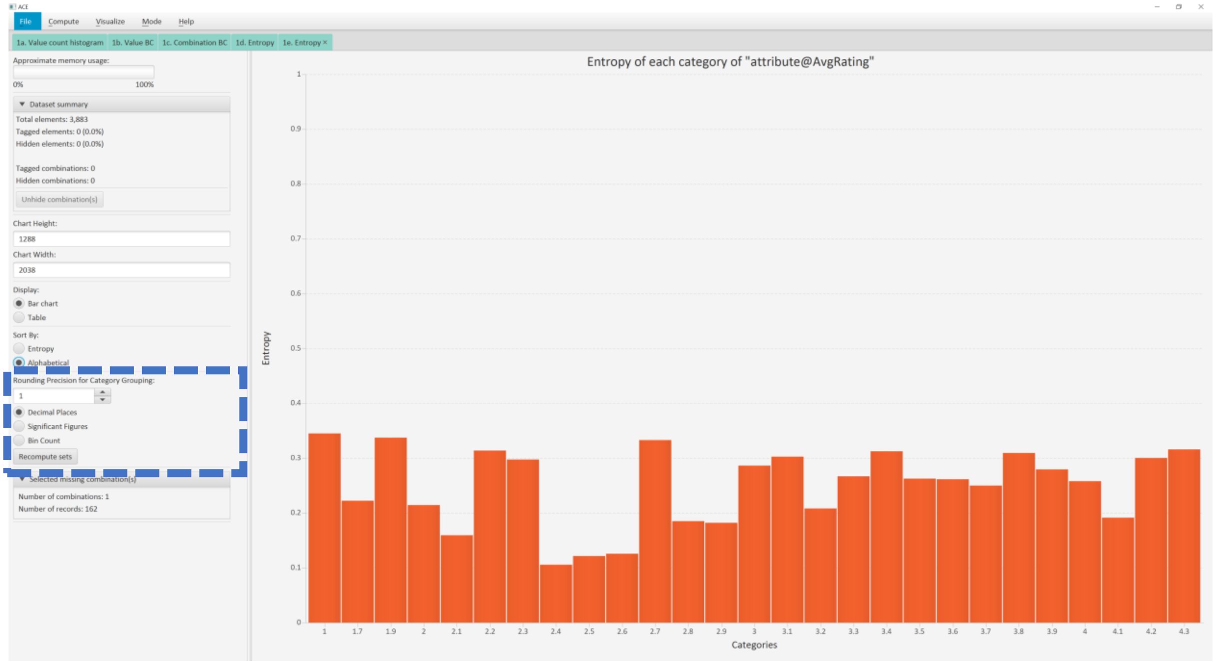
▼ Selected missing combination(s)

Number of combinations: 1
Number of records: 162

Category: 1986
Entropy: 0.2782
Elements with overlapping combination(s): 5
Elements without overlapping combination(s): 99

Figure 13: Comedy and Drama Entropy – sorted by release year.

Instead of selecting release date, it is also possible to select other attributes to calculate entropy. One such option is to select "AvgRating" as shown in Figure 14.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)

Tagged combinations: 0
Hidden combinations: 0

Chart Height: 1288

Chart Width: 2038

Display:
● Bar chart
○ Table

Sort By:
○ Entropy
● Alphabetical

Rounding Precision for Category Grouping: 1
● Decimal Places
○ Significant Figures
○ Bin Count

▼ Selected missing combination(s)

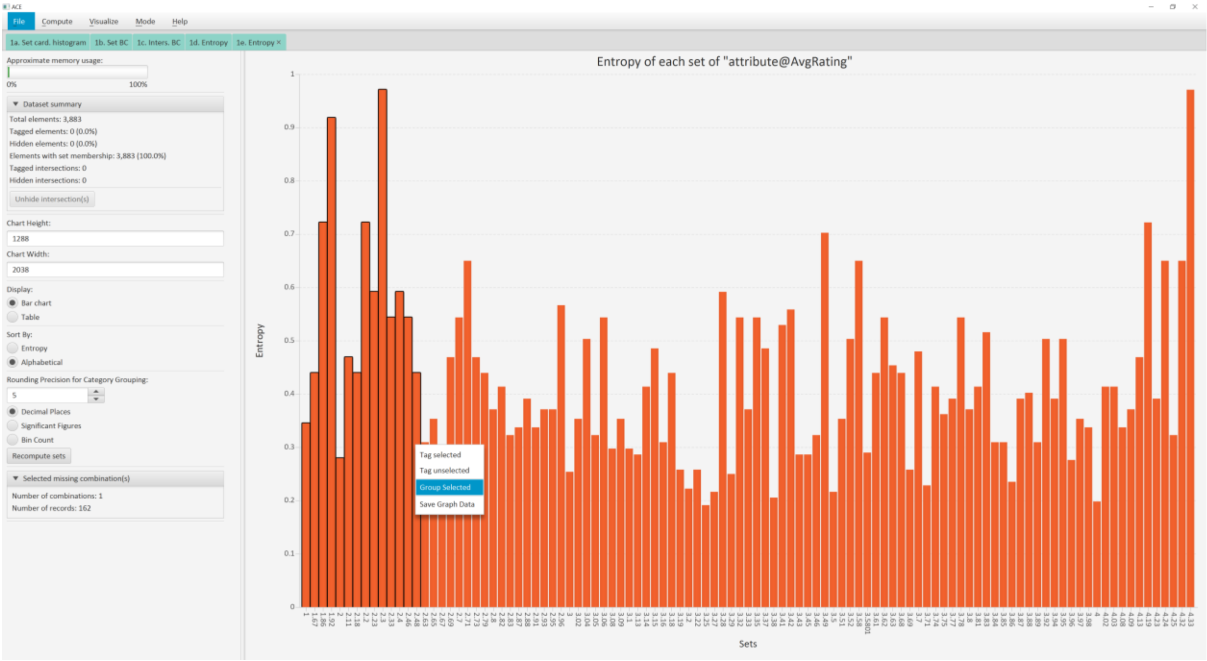
Number of combinations: 1
Number of records: 162

Figure 14: Numeric datatypes and rounding functionality.

In this case the field is numeric, which presents the possibility of performing rounding. Rounding is controlled by the control panel (see: Figure 14) and can include the following rounding options: by decimal place (including 0 i.e. integers), significant figures or using the same bin packing strategy as seen elsewhere within ACE.

For the purpose of rounding existing categories can be grouped together. In this case we group these values into low medium and high.

1. Change the rounding so that it is to 5 decimal places.
2. Change the sort order so that it is “Alphabetical”
3. Select a range of values (that might be considered low, < 2.5). (See Figure 15)
4. Right-click and select “Group Selected” from the context menu.
5. Now call this group “Low”.
6. Select a range of values (that might be considered low, < 3.5 but not the low category)
7. Right-click and select “Group Selected” from the context menu.
8. Now call this group “Medium”.
9. Finally select every column other than the “Low” and “Medium” columns
10. Right-click and select “Group Selected” from the context menu.
11. Now call this group “High”.



Approximate memory usage:

0% 100%

▼ Dataset summary

Total elements: 3,883

Tagged elements: 0 (0.0%)

Hidden elements: 0 (0.0%)

Elements with set membership: 3,883 (100.0%)

Tagged intersections: 0

Hidden intersections: 0

Chart Height: 1288

Chart Width: 2038

Display:
● Bar chart
○ Table

Sort By:
○ Entropy
● Alphabetical

Rounding Precision for Category Grouping: 5
● Decimal Places
○ Significant Figures
○ Bin Count

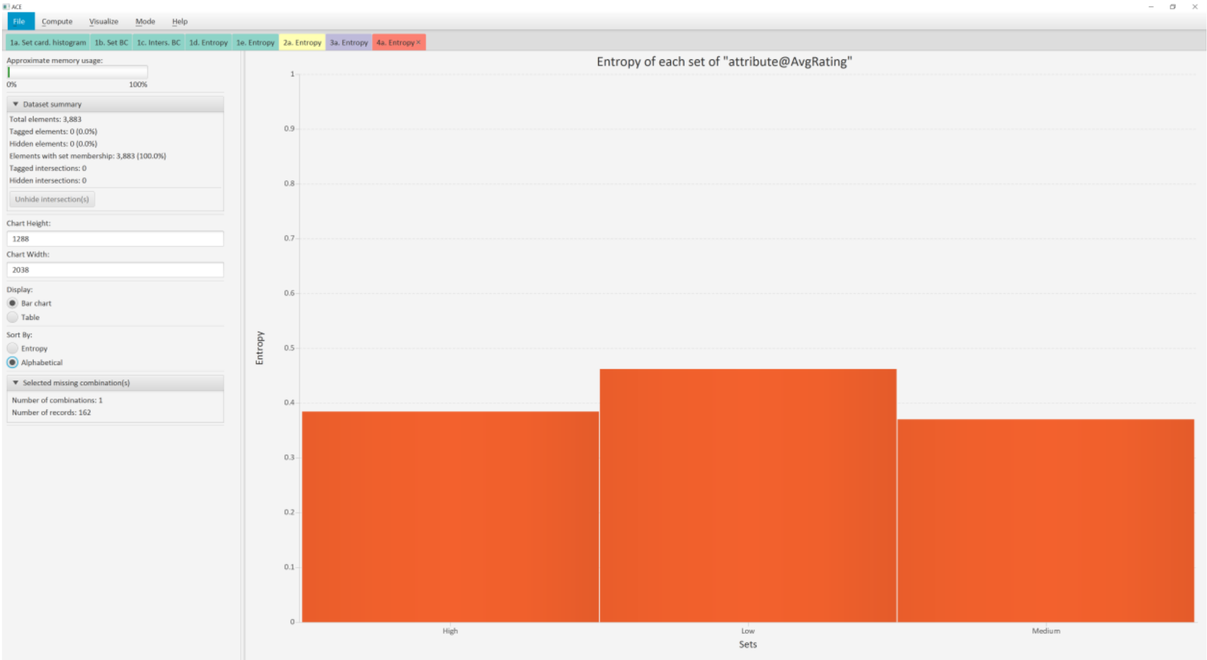
▼ Selected missing combination(s)

Number of combinations: 1

Number of records: 162

Figure 15: Grouping sets together 1st Stage, to create the “Low” group.

This results in the final categories: high medium and low and their respective entropy as shown in Figure 16.



Approximate memory usage:
0% 100%

▼ Dataset summary

Total elements: 3,883
Tagged elements: 0 (0.0%)
Hidden elements: 0 (0.0%)
Elements with set membership: 3,883 (100.0%)
Tagged intersections: 0
Hidden intersections: 0

Chart Height: 1288

Chart Width: 2038

Display:
● Bar chart
○ Table

Sort By:
○ Entropy
● Alphabetical

▼ Selected missing combination(s)

Number of combinations: 1
Number of records: 162

Figure 16: Grouping Sets Together Final Stage.

Note: It should be noted that once a numeric field is treated as categorical data the rounding functionality is disabled (See the difference between the control panel in Figure 15 and Figure 16). Thus it is best to ensure when you start rounding that the amount of decimal places shown is the maximum available.

4.2. Exporting Analysis

Now that this analysis has been completed, it is possible to export it to disk.

To perform the export the following actions should be performed:

1. Go to “File” menu and select “Export analysis”
2. Give the file a name
3. Click Save

It can then at a later time be reloaded again. To import the analysis back in the following actions should be performed:

1. Go to “File” menu and select “import analysis”
2. Click “OK” to discarding current data.
3. Select the file that you saved
4. Click “Open”