ACE Tutorial

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 functionality of ACE by providing two worked examples (ACE stands for Analysis of Combinations of Events). The first example focuses on the high-level functionality of ACE and uses a synthetic dataset that mirrors the missingness patterns that were found in an extract of health data (an Admitted Patient Care (APC) dataset from Hospital Episode Statistics (HES)). The second example covers some of the advanced functionality of ACE by analysing an open source dataset that provides nutritional information of various products.

2. Example 1 - Analysing missingness in synthetic APC data

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

- Importing a dataset into ACE and computing the missingness
- Analysing the field-level missingness using the "Value bar chart".
- Analysing the combination-level missingness using the "Combination heatmap".
- Analysing unexpected missing combinations using the "data slicing (via combination hiding)" functionality.
- Explaining unexpected missing combinations using the "data mining" functionality.

The synthetic dataset used in this example has 200 records and 15 fields. The first ten fields mimic the diagnosis codes for a patient's illness (DIAG_01 to DIAG_10) and the remaining five fields are a patient's sex (SEX), age at the date of admission (ADMIAGE), admission method (ADMIMETH), mortality status (MORTALITY), and the health care provider (PROCODE3).

2.1. Import data and compute missingness

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

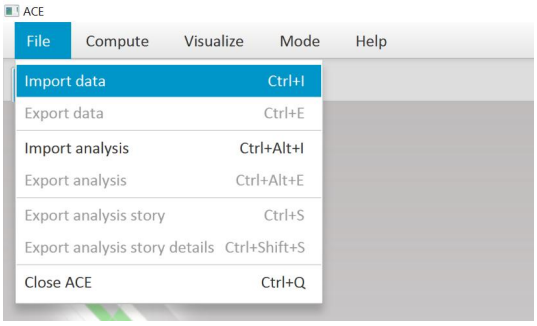

Figure 1: Data import menu

3. In newly opened window, navigate to the “ACE” folder and then the “datasets” folder. Select “Synthetic_APC_DIAG_Fields.csv” and click “Open” (see Figure 2).

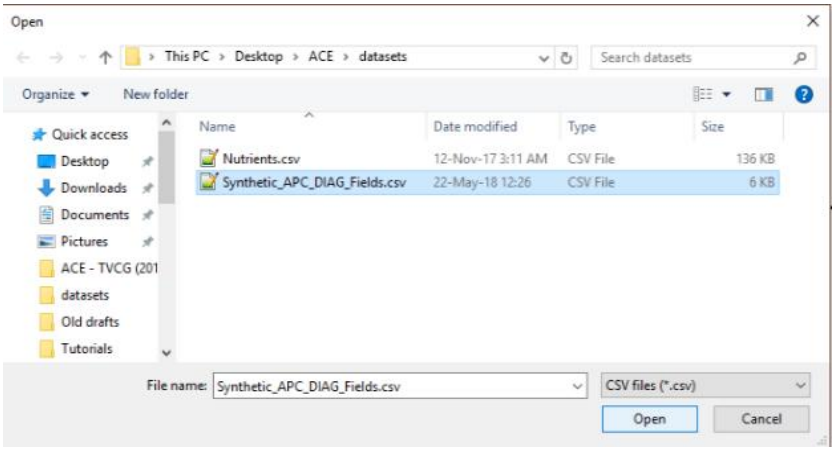

Figure 2: Open file window

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 “Compute” menu and select “Perform Computation” (see Figure 3).

Figure 3: Perform computation menu

7. A “Combinations computed successfully” dialog box should appear. Click “OK”. This completes the data import and missingness computation in ACE.

2.2. Field-level missingness - Value bar chart

Upon the successful computation of missingness, ACE automatically creates a “Value bar chart” to act as a start point for analysing missingness in a dataset (see Figure 4).

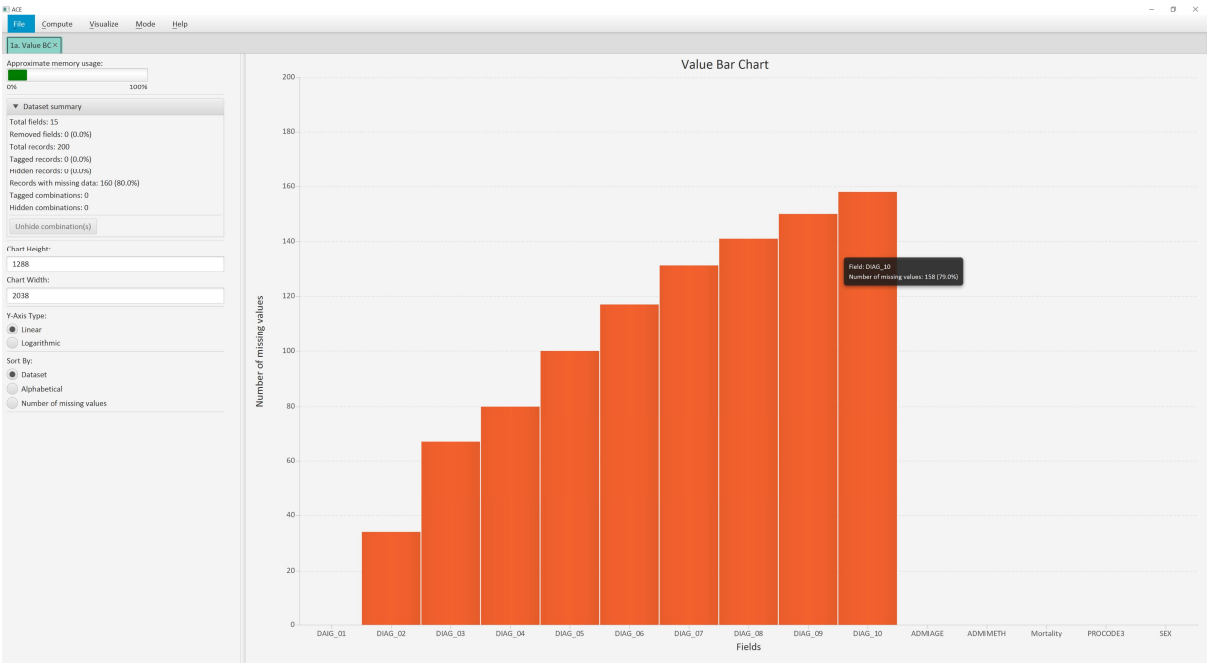

Figure 4: A value bar chart with a tooltip

ACE uses a tab pane layout, which places each newly created visualization in a new tab. Each tab is divided into two main panels: a visualization panel on the right and a control panel on the left. Users can move the divider between these two panels to change the amount of screen-space that is allocated to each panel. The visualization panel always has a title at the top and an interactive visualization (or a table) below the title. The control panel displays "approximate memory usage" by ACE, a "dataset summary", size of the displayed visualization, and other visualization-specific content (e.g., sorting methods).

The "Value bar chart" displays the number of missing values (y-axis) in each field (x-axis). Like any other visualization in ACE, it supports tooltip via mouseover and selection using left-click, <ctrl> left-click or rubber-band selection (i.e., hold left-click, drag and then release left-click).

This chart (see Figure 4) shows us that, as expected, the primary diagnosis field (DIAG_01) and the 5 categorical fields (ADMIAGE, ADMIMETH, MORTALITY, PROCODE3 and SEX) are never missing. Further, the secondary diagnosis fields are missing progressively more often from DIAG_02 to DIAG_10.

2.3. Combination-level missingness - Combination heatmap

To investigate missingness at the combination-level, go to "Visualize" menu and click "Combination heatmap". This creates a "Combination heatmap" and places it in a new tab. Users can switch between the tabs by left-clicking on their names (e.g., "1b. Combination heatmap"). The "combination heatmap" displays a matrix of fields on the x-axis and missing combinations on the y-axis. The number of records that are associated with each missing combination is encoded using a sequential colour map (see legend in the "control panel").

By default, "combination heatmaps" are sorted by "dataset(x-axis) vs. length(y-axis)", which means that the missing fields (x-axis) are sorted as they appear in the dataset and the missing combinations (y-axis) are sorted by their length (i.e., the number of fields in a missing combination) from bottom to top. Select the "Dataset(x-axis) vs. count(y-axis)" from the "control panel" to sort the missing combinations in the heatmap by their number of records (see Figure 5).

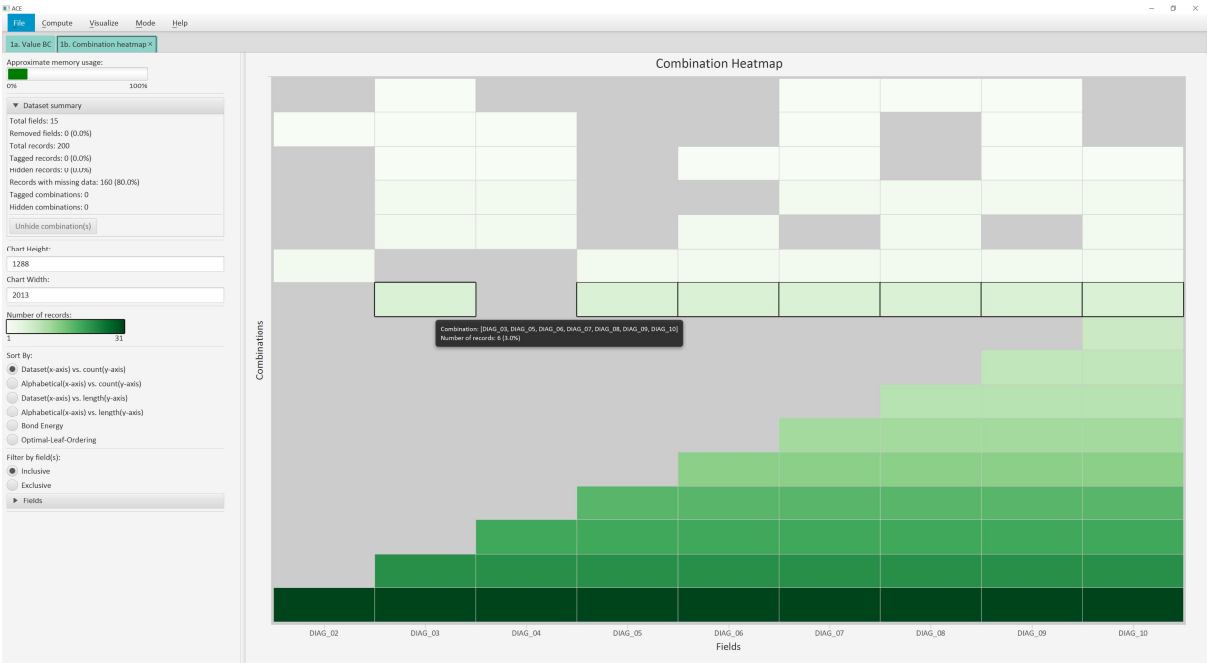

Figure 5: A combination heatmap sorted by dataset (x-axis) vs. count (y-axis) order. A combination with 7 missing fields (DIAG_03 and DIAG_05 to DIAG_010) is selected, with a tooltip showing its name and the associated number of records.

It is expected in this dataset that if any diagnosis field from DIAG_02 to DIAG_10 is missing then all of the subsequent diagnosis fields should also be missing. The “combination heatmap” highlights that this is indeed true for most of the missing records (see bottom 9 combinations in Figure 5). However, there are also 7 unexpected missing combinations (see top 7 combination in Figure 5), which have gaps in the diagnosis fields.

2.4. Analysing unexpected missing combinations - Data slicing

To further investigate the unexpected missing combinations:

1. Select the expected missing combinations (the bottom 9 combinations in the heatmap) via rubber-band selection (hold down the left mouse button, drag and then release the button). This will highlight all of the selected missing combinations (see Figure 6).
2. Right-click and select “Hide selected” (see Figure 6)

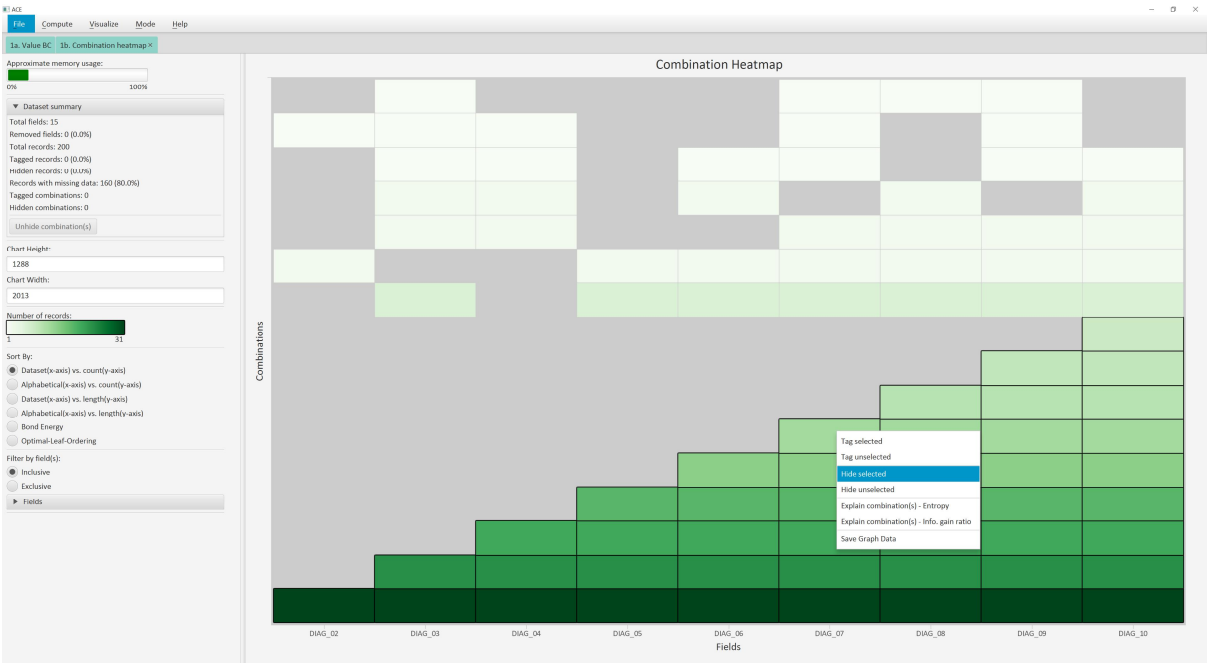

Figure 6: A combination heatmap showing a context menu for the selected missing combinations

The above operation will slice the data by excluding the records with expected missingness from the analysis. It will also create a new tab with a new “combination heatmap”, which will only have the unexpected missing combinations (see Figure 7). Also note the change in the number of “Hidden records” in the data summary panel on the top-left corner. The hidden records can be un-hidden at any stage during the analysis.

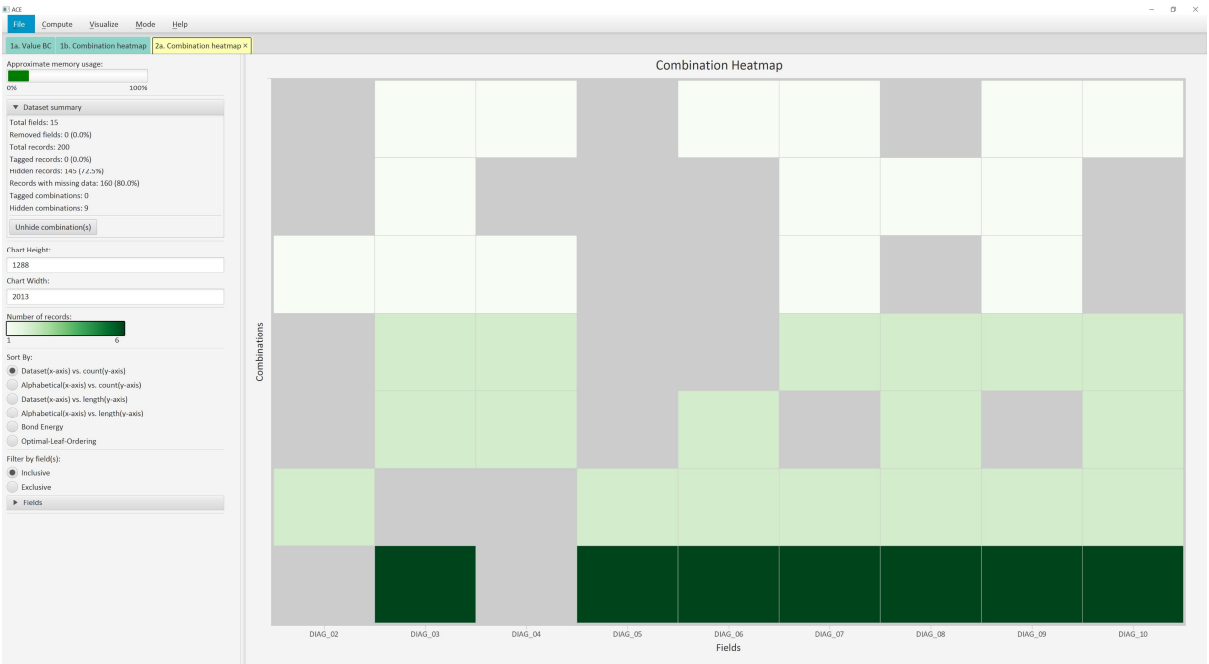

Figure 7: A combination heatmap showing the unexpected missing combinations

After excluding the records with expected missingness, go to the “Visualize” menu and select “Value bar chart”. This creates a new “Value bar chart” for the remaining records, which have the unexpected missingness (see Figure 8).

The new “value bar chart” (see Figure 8) highlights a very different pattern of missingness compared with before hiding the records (see Figure 4). From this view, DIAG_03 pop-out as a particular problem, because it is missing substantially more often than the subsequent diagnosis fields.

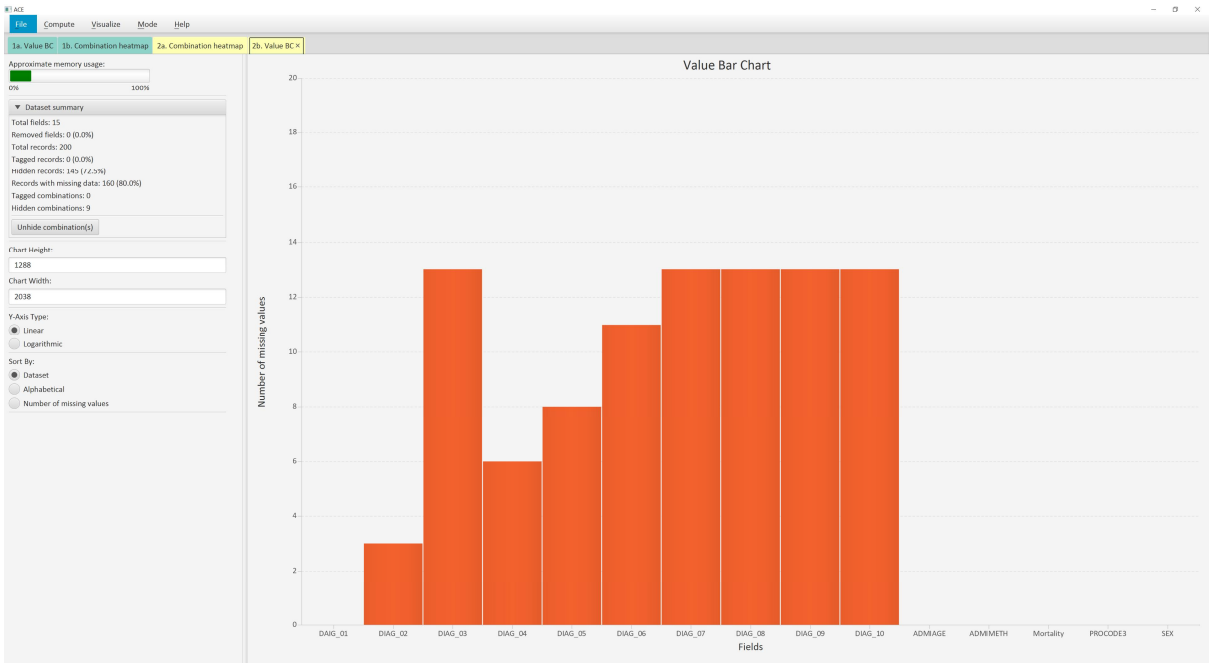

Figure 8: A value bar chart showing the field-level missingness after hiding the record with expected missingness

2.5. Explaining unexpected missing combinations – Data mining

To identify the source of unexpected DIAG_03 missingness:

1. Switch to the heatmap showing the unexpected missing combinations by left-clicking on tab “2a. Combination heatmap”.
2. Select the most frequent unexpected missing combination (DIAG_03 and DIAG_05 to DIAG_10 missing; 6 records).
3. Right-click and select “Explain combination(s) - Entropy” from the context menu.
4. Select “ADMIMETH” in the field selection dialog box and click OK (see Figure 9). This creates an “Entropy bar chart”.

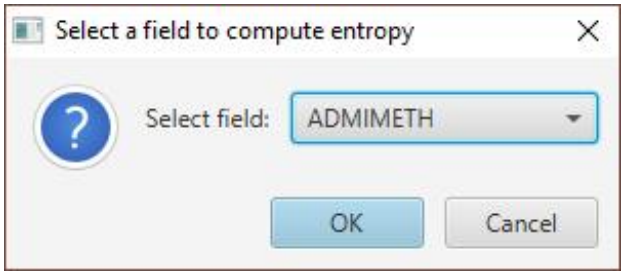

Figure 9: Field selection dialog box

5. From the visualization “control panel” on the left-hand-side, select “Table” radio button to display an entropy table (see Figure 10).

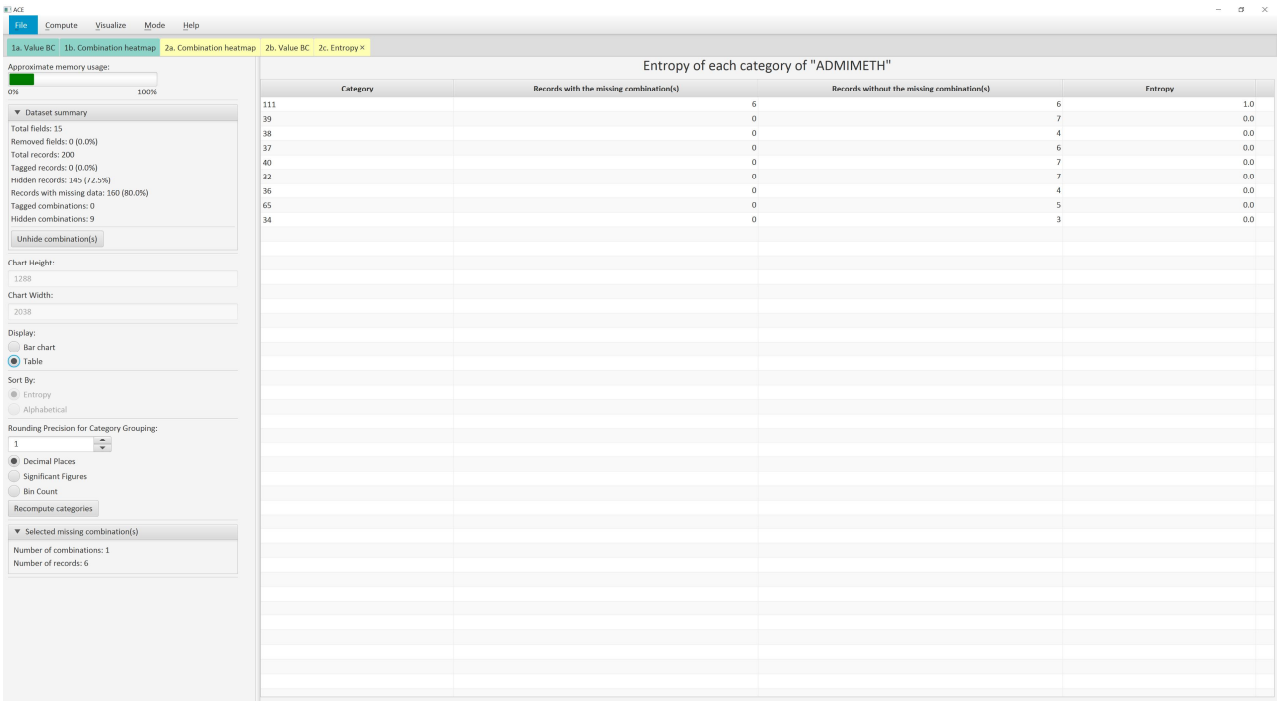

Figure 10: Entropy table showing the entropy of admission methods for the selected missing combination

The entropy table (see Figure 10) shows that all 6 records in the selected unexpected missing combination belong to admission method “111”.

Next, repeat steps 1 – 5 of the above process for the PROCODE3 field. This highlights that all the records in the selected unexpected missing combination are submitted by a single provider (i.e., “aaa”) (see Figure 11). In other words, the insight that you have gained is that this unexpected combination of missing fields is due to data for one admission method (“111”) from one provider (“aaa”). You can now clean the data (shift the values in DIAG_05 – DIAG_10, so that DIAG_10 is missing not DIAG_05) and send the provider feedback so that this problem does not occur in the future.

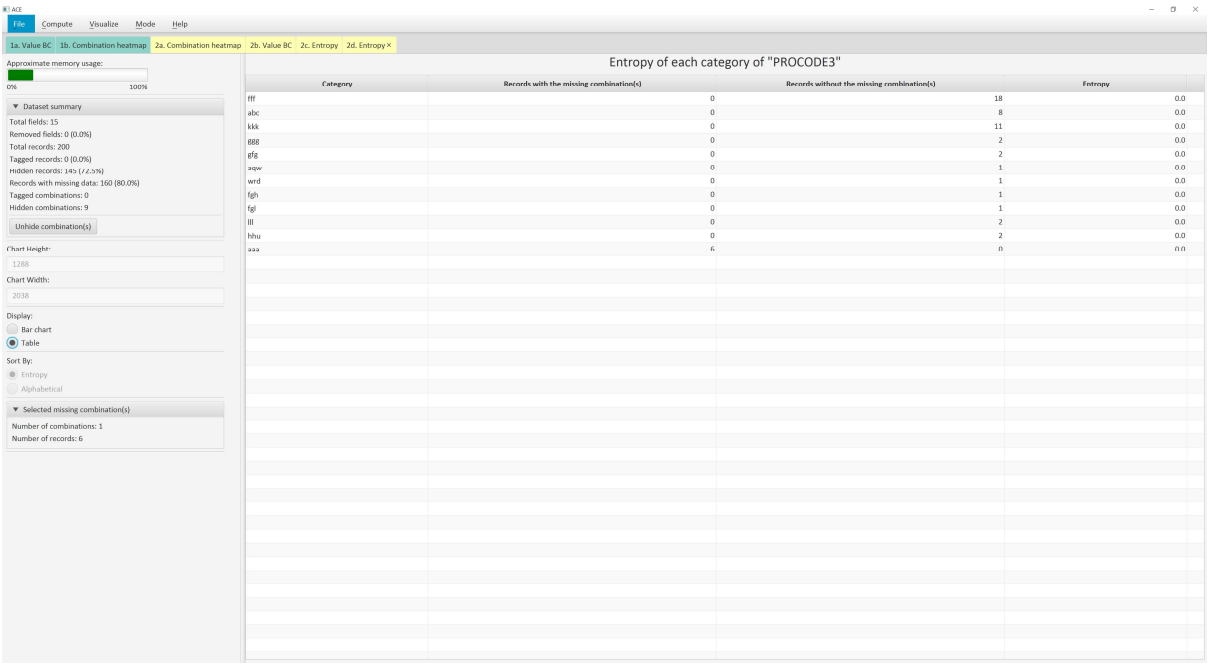

Figure 11: Entropy table showing the entropy of PROCODE3 for the selected missing combination

3. Example 2 - Analysing missingness in nutrient data

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

- Usage of the remaining four visualizations: “value count histogram”, “combination bar chart”, “combination count histogram” and “combination length histogram”.
- Explanation of “missing combination tagging” and “record tagging”
- Explanation of “information gain ratio” functionality
- Illustration of “analysis story export” and “data export”.

We used an open source dataset in this worked example, which can be downloaded from the following link: https://github.com/enjalot/Bay-Area-d3/blob/master/data/nutrients.csv. This dataset has 7,637 records and 16 fields. The first field in the dataset lists the products’ names (e.g., “Salt, table”), the second field describes their group (e.g., “Spices and Herbs”) and the remaining 14 fields list the nutritional information about the products (e.g., “protein (g)”, “calcium (g)” and “fibre (g)”).

3.1. Additional visualizations in ACE

1. To begin with, import the nutrient dataset (file name: Nutrient.csv) and compute the missingness in the dataset. See Section 2.1 for if you require any help.
2. The ACE will automatically create a “Value count bar chart”.
3. Go to “Visualize” menu and click “Value count histogram” (see Figure 12).

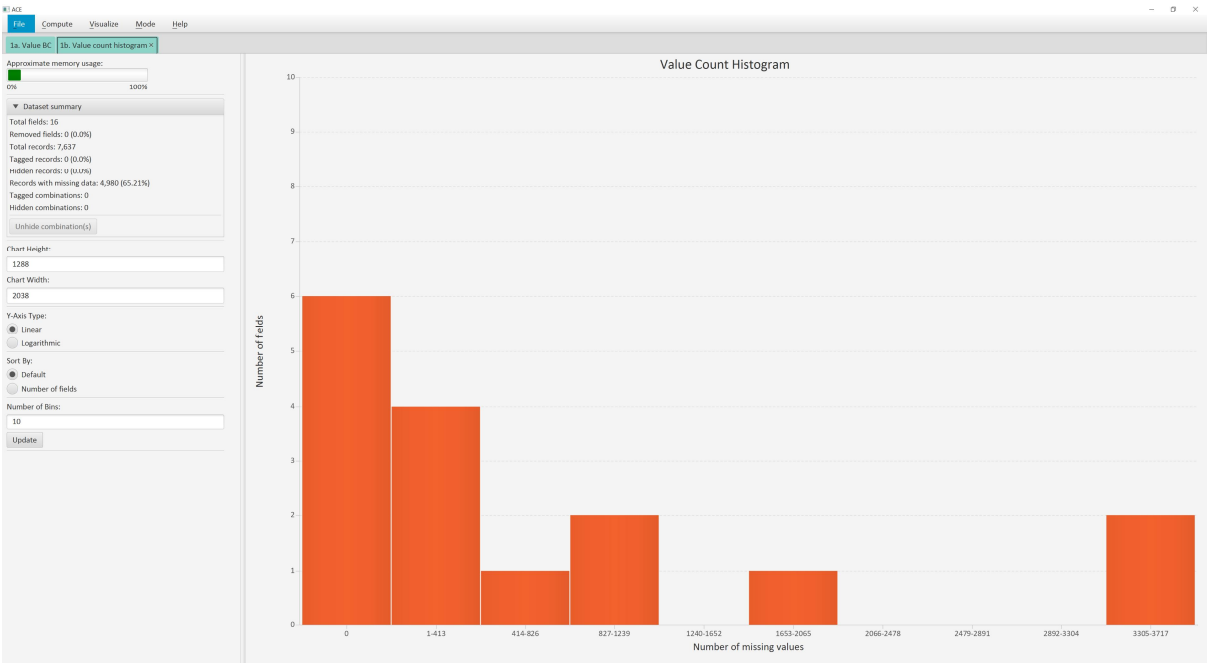

Figure 12: Value count histogram showing the distribution of the number of missing values for fields in nutrient data

The “Value count histogram” displays bins of the number of missing values on the x-axis and the number of different fields on the y-axis. By default, ACE creates ten bins for this visualization, the first of which collects all of the fields that are never missing. The remaining nine bins are of equal size. Users can interactively change the number of bins from the “control panel”. This visualization is especially useful when dealing with datasets with hundreds of fields.

The “Value count histogram” in Figure 12 shows that, in nutrient dataset, 6 fields are never missing, 4 fields are missing between 1-413 times, and so on.

4. Next, go to the “Visualize” menu and click “Combination bar chart” (see Figure 13).

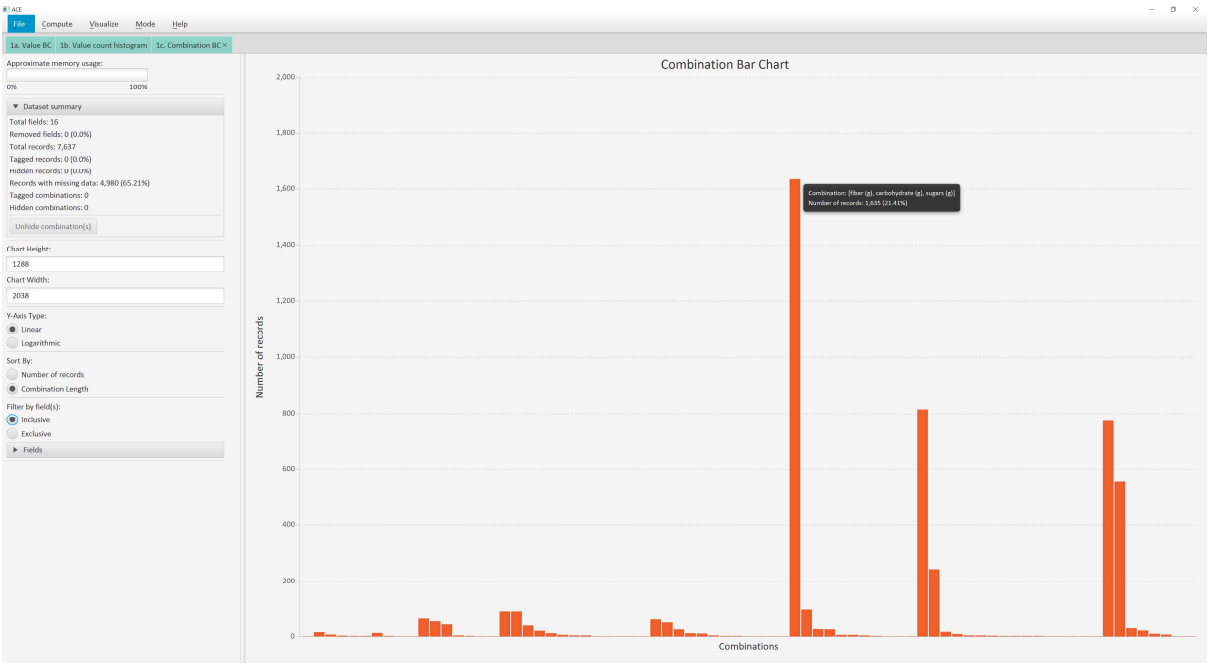

Figure 13: Combination bar chart showing all 77 missing combinations in nutrient data

The "Combination bar chart" displays missing combinations the x-axis and the associated number of records on the y-axis. Like any other bar chart in ACE, the use can sort this visualization and change the y-axis to "linear" or "logarithmic".

The visualization in Figure 13 displays all 77 missing combinations in the nutrient dataset. Using the tooltip, we can see that the most frequent missing combination contains three fields (i.e., "fiber (g)", "carbohydrate (g)" and "sugar (g)") and it is missing in 1,635 records.

5. Next, go to the "Visualize" menu and click "Combination count histogram" (see Figure 14).

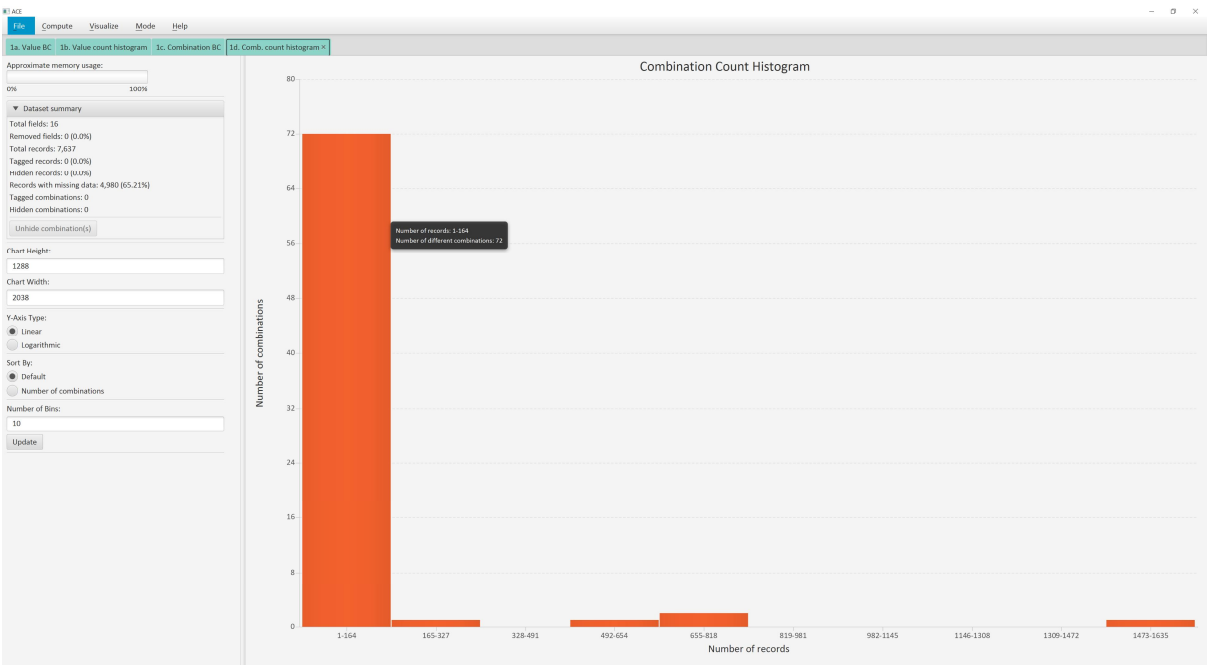

Figure 14: Combination count histogram showing all 77 missing combinations in 10 bins

The “Combination count histogram” displays bins of the number of records on the x-axis and the number of different missing combinations on the y-axis. This visualization is especially helpful when the number of missing combinations is so large (e.g., tens of thousands) that they cannot be visualized using the “Combination heatmap” or “Combination bar chart”.

The “Value count histogram” in Figure 14 shows that the frequency distribution of missing combinations in nutrient data is highly skewed, as 72 out of 77 combinations fall in the first bin of the histogram.

6. Lastly, go to the “Visualize” menu and click “Combination length histogram”.
7. In the “control panel”, change the “number of bins” from 10 to 5 and click “Update”. The resulting visualization is shown in Figure 15.

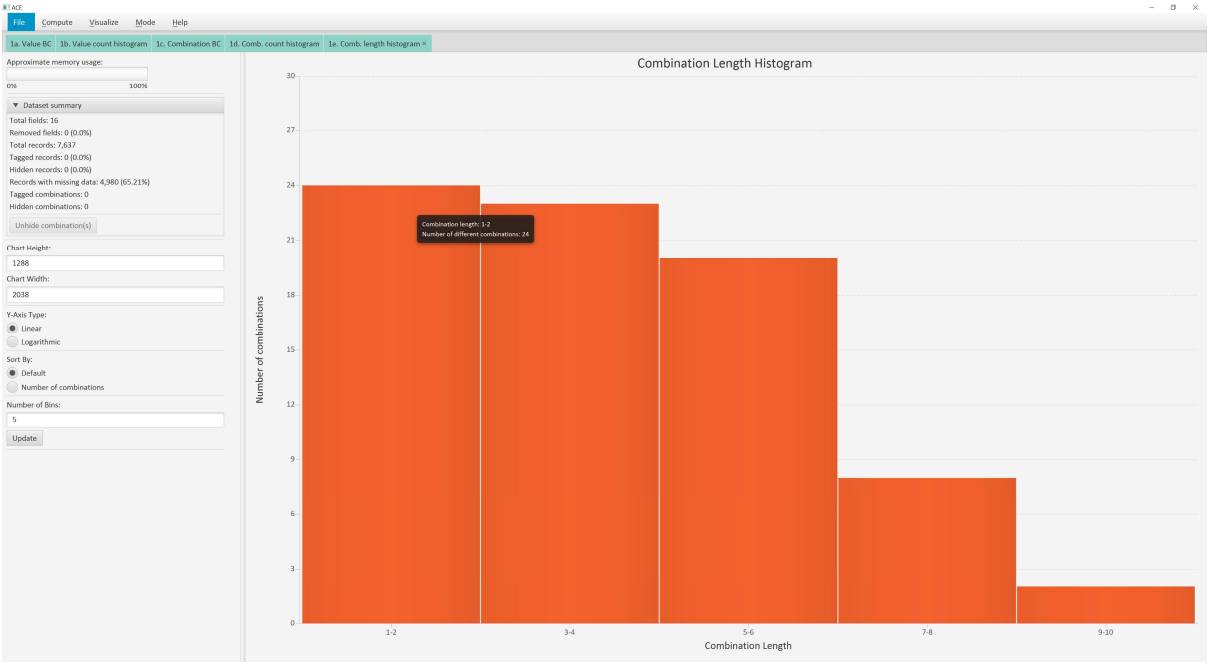

Figure 15: Combination length histogram showing all 77 missing combinations in 5 bins

The “Combination length histogram” displays bins of combination length (i.e., number of fields missing in combination) on the x-axis and the number of different missing combinations on the y-axis. Similar to “combination count histogram”, this visualization also highly scalable.

The histogram in Figure 15 displays all 77 missing combinations, in the nutrients dataset, using 5 bins. The tooltip shows that 24 combinations are missing between 1-2 fields.

3.2. Tag missing combinations and records

ACE allows an analyst to tag the selected missing combinations from any of the four missing combination visualizations: combination heatmap, combination bar chart, combination count histogram, and combination length histogram. Combination tagging prompts users to enter a comment for the analysis story (see Section 3.4). Similar to “combination hiding” (see Section 2.4), “combination tagging” also filters-out the selected missing combinations from further analysis. The difference between the two is that “hidden” combinations can be unhidden at any stage of the analysis; however, tagged combinations cannot be untagged unless the user closes a tab (i.e. undo an analysis step). To “tag” combinations:

1. Select length 1-2 combinations from the “Combination length histogram” (Figure 15) using a left-click.
2. Right-click and select “Tag selected” from the context menu.
3. Enter “Not interested in analyzing combinations where 2 or less fields are missing” and click OK. This will create a new tab (of different colour) with a “combination length histogram”, displaying combinations that are missing 3 or more fields (see Figure 16).

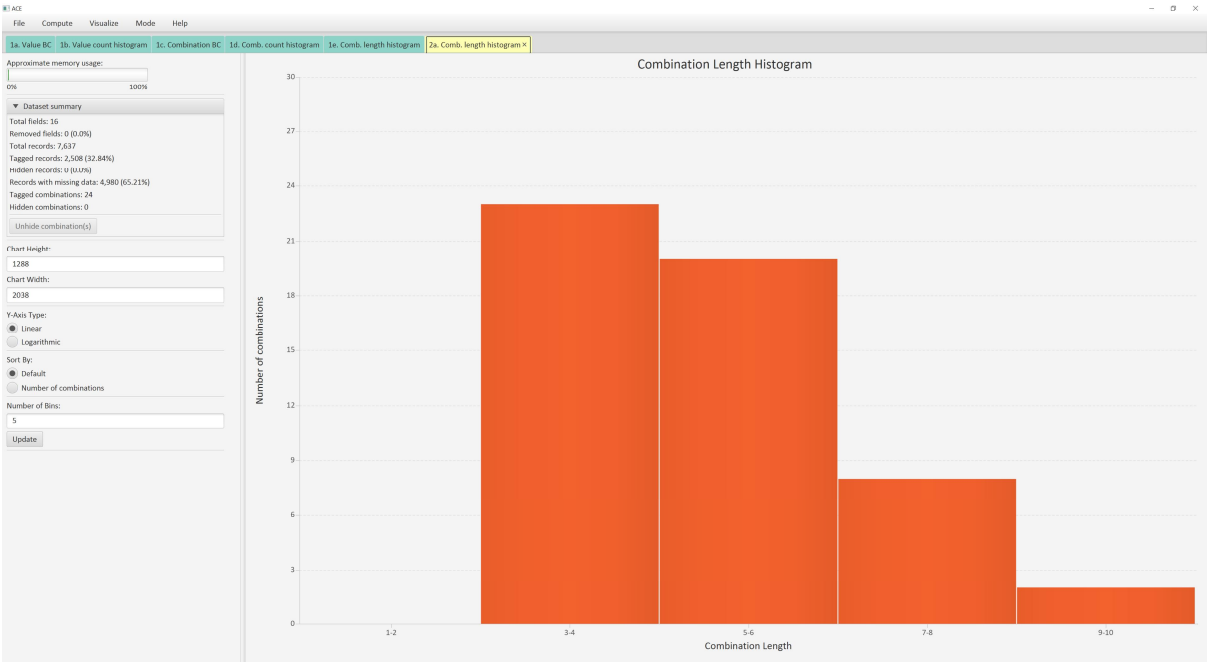

Figure 16: Combination length histogram showing combinations of length between 3-10

In addition to tagging all the records of selected missing combinations, ACE also allows user to tag specific records of selected combinations that are correlated with the values of other fields in a dataset. To do this:

1. Go to the “Visualize” menu and click “Combination bar chart”.
2. Select the most frequent missing combination (i.e., “fiber (g)”, “carbohydrate (g)” and “sugar (g)”) – missing 1,635 times).
3. Right-click and select “Explain combination(s) - Entropy” from the context menu.
4. Select “group” in the field selection dialog box and click OK. This creates an “Entropy bar chart” showing the correlation of categories in “group” field with the records of selected missing combination (see Figure 17). The tooltip shows that in “Poultry Products” category there are 271 records that have this missing combination (i.e., 271 products that are missing “fiber (g)”, “carbohydrate (g)” and “sugar (g)”).

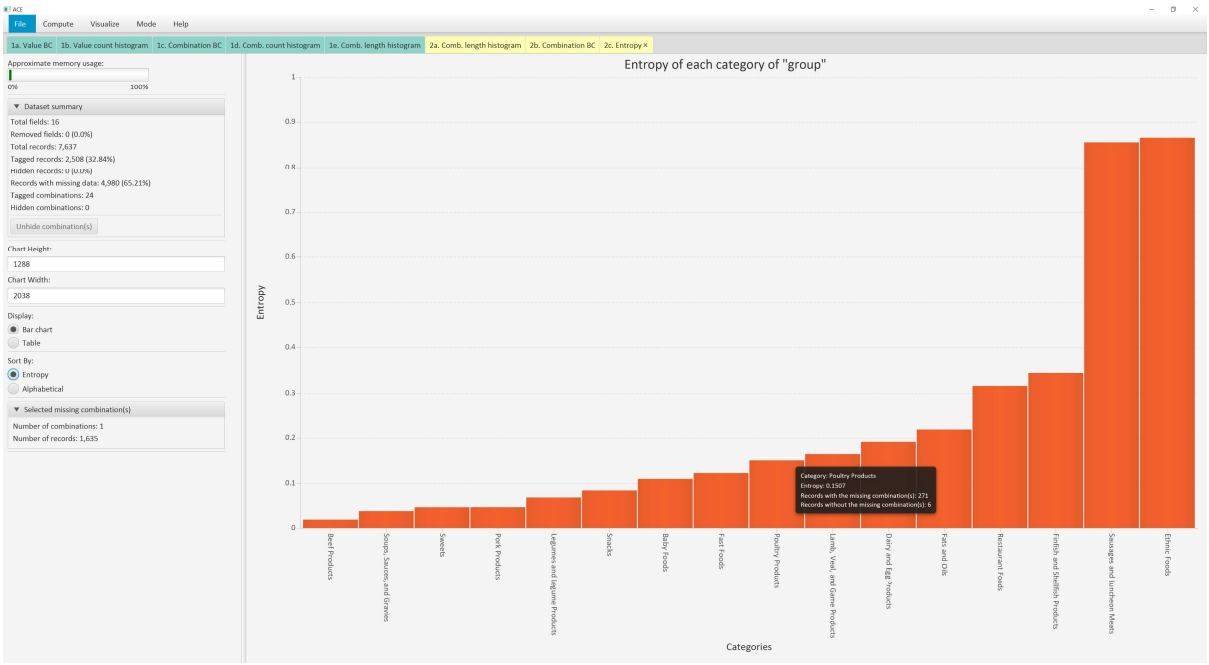

Figure 17: Entropy bar chart showing the correlation of categories in "group" field with the records of selected missing combination

5. Select "Poultry Products" category with left-click.
6. Right-click and select "Tag selected".
7. Enter "I understand why most of poultry products are missing these nutrients" in the comment dialog box and click OK. This will create a new tab (with a different colour) with an updated "Entropy bar chart".

By doing the above steps, we tagged only the 271 records of the selected missing combination that were associated with the "Poultry Products". Note the updated numbers in the "Dataset summary" and "Selected missing combination(s)" panels.

3.3. Information gain ratio

In addition to explaining missing combinations based on the values of a single field, ACE also allows users to rank multiple fields based on the correlation of their values with records that either are or are not members of selected missing combinations. For this:

1. Go to "Visualize" menu and click "Combination length histogram".
2. Select length 3 missing combinations.
3. Right-click and select "Explain combination(s) – Information gain ratio"
4. Select "calories", "water (g)" and "group" (via hold ctrl and left-click) from the field selection dialog and click "Calculate". This creates an "Information gain ratio" bar chart (see Figure 18).

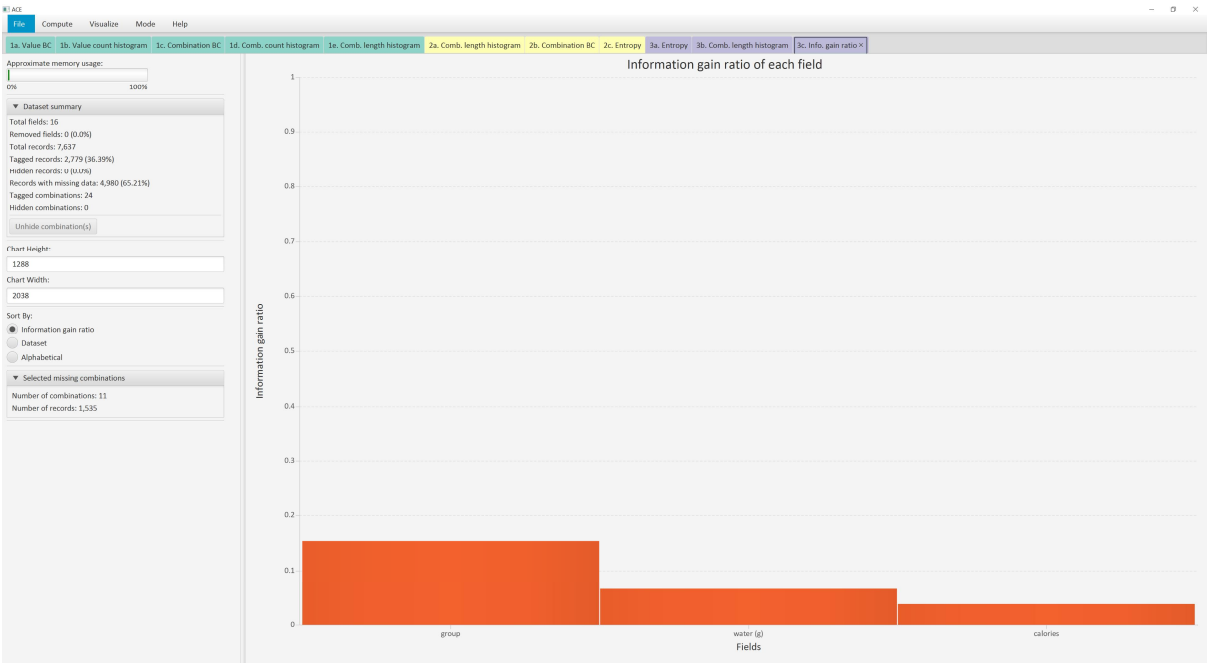

Figure 18: Information gain ratio of “calories”, “water (g)” and “group” for the selected combinations

The information gain ratio bar chart has the selected fields on the x-axis and information gain ratio on the y-axis. A higher information gain ratio indicates that a field’s categories are more homogenous than other selected fields, in terms of the presence (or absence) of the records of selected missing combinations.

This chart has tooltips, and the fields can be sorted by their order in the dataset, alphabetical order or by information gain ratio. Users can select any field from this chart (using left-click) to compute the correlation of its categories (using a context menu that appears with right-click) with the selected missing combinations.

3.4. Exporting analysis story and data

In section 3.2, we first tagged 24 length 1-2 missing combinations that were present in 2,508 records. Next, we tagged 271 records (out of 1,635) that belonged to a single length 3 combination. In total, we tagged 2,779 records so far in our analysis, out of 4,980 records that have missingness in this dataset.

There are six major actions that a user may perform during the analysis: remove fields, tag combinations, hide combinations, unhide combinations, tag records, and create a visualization. ACE captures these actions, together with the user's comments (if any), and can export them as an analysis story in a CSV file at two granularities.

At a coarse granularity, each analysis action is exported in a single row. Nine fields are always recorded (action number, action type, visualization type, total fields, remaining fields, total records, records with missing data, total tagged records, and total hidden records). Additionally, all actions except "create a visualization" have a few action-specific fields (e.g., number of removed fields, user comment, and the names of removed fields for the "remove fields action". To export a story:

1. Go to “File” menu and select “Export analysis story”
2. Enter the name of file in which you want to store the analysis story and click “save”.

A finer-grained analysis story file can contain multiple records for each action, with fields for the action number, action type, visualization type, and action details. Remove fields actions store a separate row for each field. Similarly, tag combinations, hide combinations, and unhide combinations actions store a separate row for each combination. The tag records and create a visualization actions are, as in the coarse-grained file, stored in a single row. For this:

1. Go to "File" menu and select "Export analysis story details"
2. Enter the name of file in which you want to store the analysis story and click "save".

In addition to "analysis story", ACE allows users to export a copy of the dataset with five extra fields, which contain automatically recorded data about the actions (action number, action type, visualization type, and action details) and any user-entered comment. This provides an audit trail that explains patterns of missingness at the level of individual records. To do this:

1. Go to "File" menu and select "Export data"
2. A dialog box appears with four different data export options (see

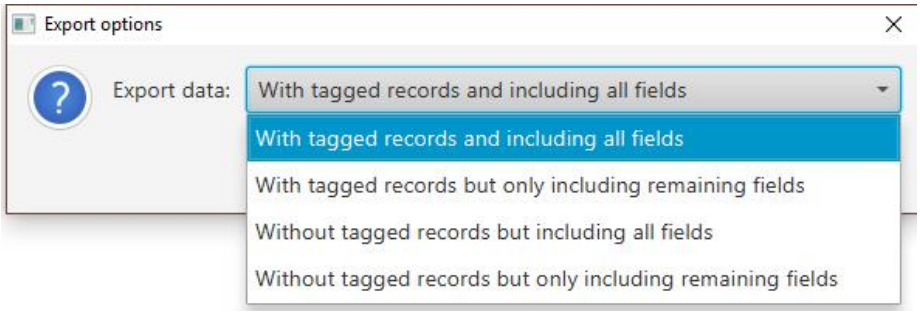

Figure 19).

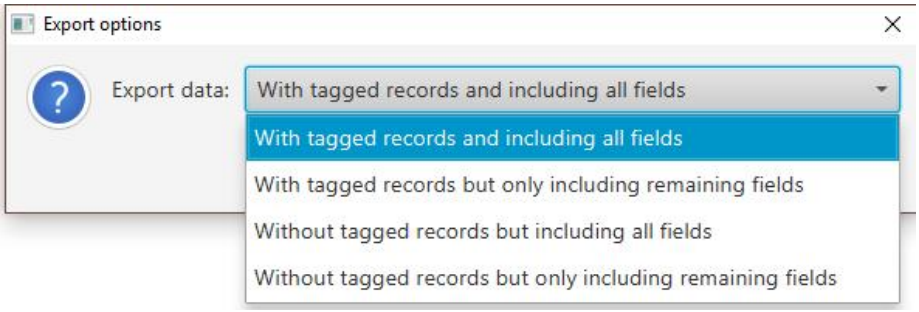

Figure 19: Data export dialog box

3. Select the first option "With tagged data and including all fields" and select OK.
4. Enter the name of file in which you want this data to be exported and click "save".