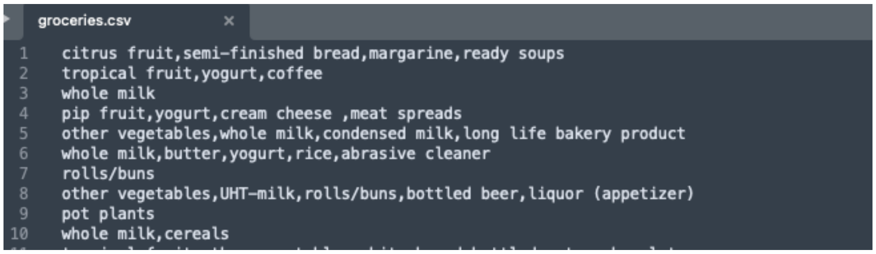
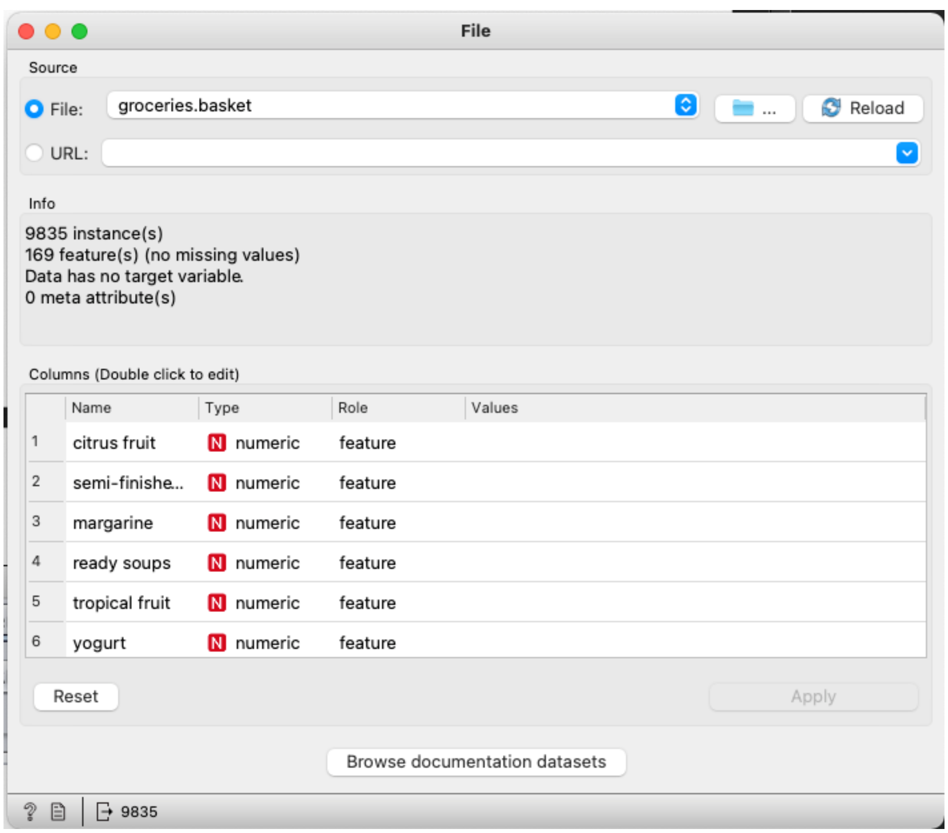
Orange Association Rules

Orange can either work with “basket” data, or normal csv data (with column headers). Basket data looks like this:


groceries.csv

```
citrus fruit,semi-finished bread,margarine,ready soups
tropical fruit,yogurt,coffee
whole milk
pip fruit,yogurt,cream cheese ,meat spreads
other vegetables,whole milk,condensed milk,long life bakery product
whole milk,butter,yogurt,rice,abrasive cleaner
rolls/buns
other vegetables,UHT-milk,rolls/buns,bottled beer,liquor (appetizer)
pot plants
whole milk,cereals
```

Note that even though this is a csv file, Orange won’t recognize it unless you save it as a “.basket” file. Once you’ve done that, you can open it with the File widget.



The limitation of this is that there is no easy way to manipulate the domain following a file without Orange converting features to “continuous,” which the association widget needs.

Alternatively, convert your file to a one-hot encoded feature table. To save yourself effort, use a character to indicate the presence (and absence, if you like) of features. E.g.,

	rice	canned vegetables	sauces	UHT-milk	bags	coffee	white bread	packaged frui
1	?	?	?	?	?	?	?	?
2	?	?	?	?	?	T	?	?
3	?	?	?	?	?	?	?	?
4	?	?	?	?	?	?	?	?
5	?	?	?	?	?	?	?	?
6	T	?	?	?	?	?	?	?
7	?	?	?	?	?	?	?	?
8	?	?	?	?	?	?	?	?
9	?	?	?	?	?	?	?	?
10	?	?	?	?	?	?	?	?
11	?	?	?	?	?	?	T	?

If you encode missing elements as well, you will get rules that include the absence of items on the left and right. Here's some code I used to transform my data.

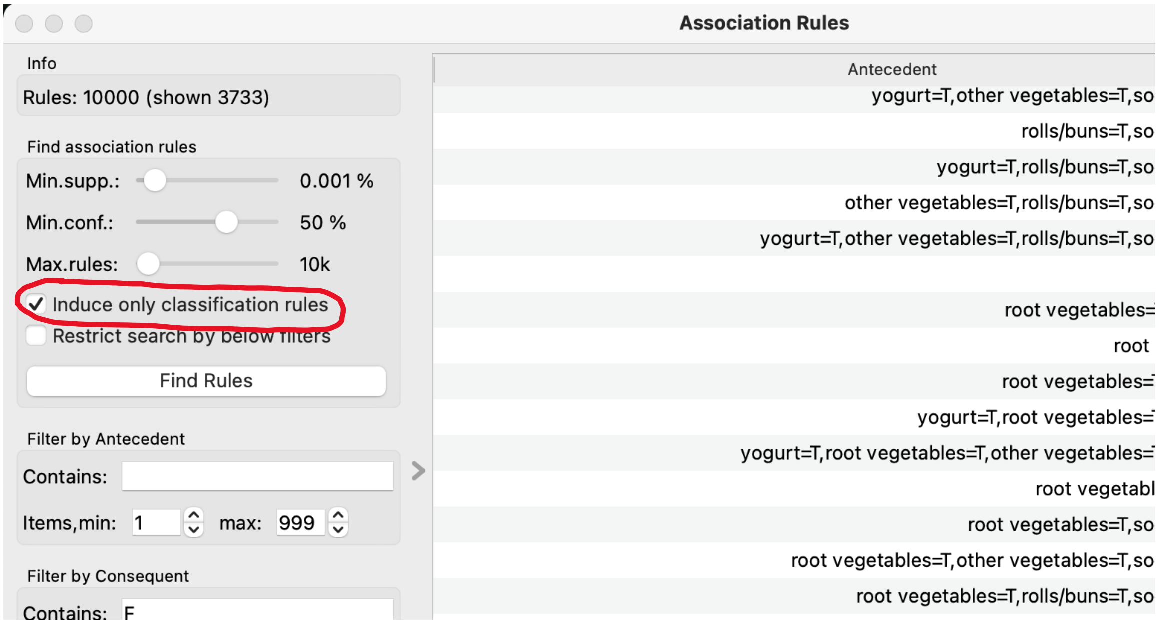
```
import csv

header = set()
rows = []
with open("groceries.csv") as f:
    for line in f:
        d_row = {}
        for word in line.split(","):
            word = word.strip()
            header.add(word)
            d_row[word] = "T"
        rows.append(d_row)


with open("groceries_one_hot.csv","w") as f:
    writer = csv.DictWriter(f,list(header))
    writer.writeheader()
    for r in rows:
        writer.writerow(r)
```

Classification Rules

In Orange, you can use the Association widget to develop classification rules by specifying a Target variable. Do this using the "Edit Domain" widget, and then select "Induce only classification rules" in the interface.



Exercise: Groceries

1) Try processing the "groceries" files. Try this with the basket file, and the "one-hot-encoded" version. Do you notice differences? What are they?
2) Note the other measures of rule quality are (Cover, Strength, Lever). What do these mean? Have a look at the Wikipedia page for association rule learning: https://en.wikipedia.org/wiki/Association_rule_learning
3) Use the Imputer (or a Python script) to replace missing values with a specific value and recalculate the rules. Play with the algorithm parameters to find the best (i.e., highest support / confidence, lift > 1) rule that predicts the *absence* of whole milk. What is it? Report confidence, support, lift.

Exercise: Titanic Data Set

Look at the Titanic (training) dataset. Load this dataset and in Preprocessing, use Discretize to convert Numeric attributes into Nominal.

First, play with the confidence measure. Try lowering the confidence to .8 in order to get more rules.

Now look at rules with Lift greater than 1 and as high Confidence as possible.

Post three interesting rules.