CSCU9M5 Practical 1

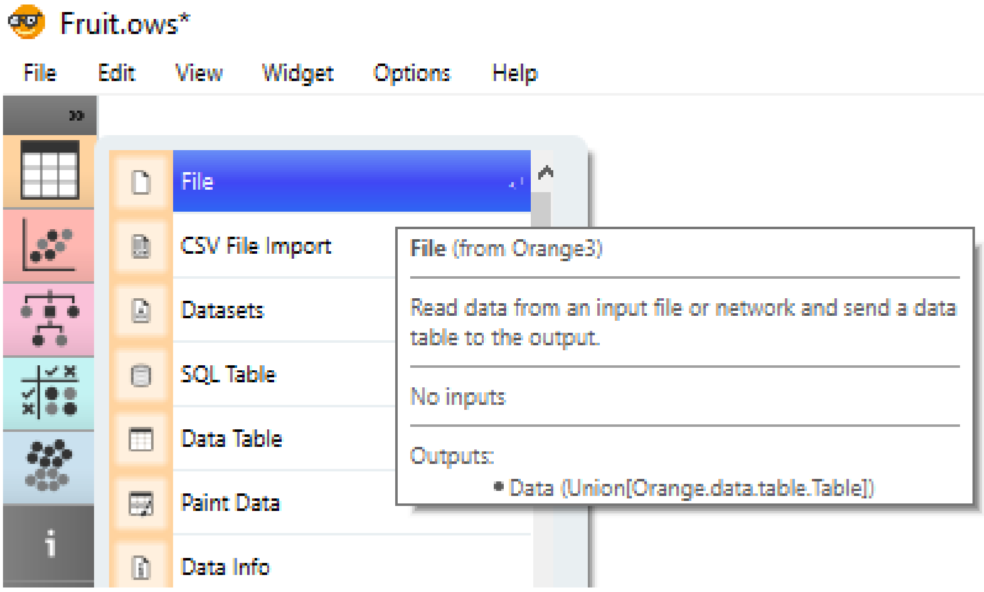
Let's have a go at using Orange with some data. First, download this file, which is in csv format: MotorPremiums.csv. Now run Orange and follow the steps below.

Orange Intro

Select New Project.

Importing Data

The first thing we need to do is load our data set into Orange. Use the File widget:



File (from Orange3)

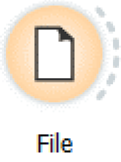
Read data from an input file or network and send a data table to the output.

No inputs

Outputs:

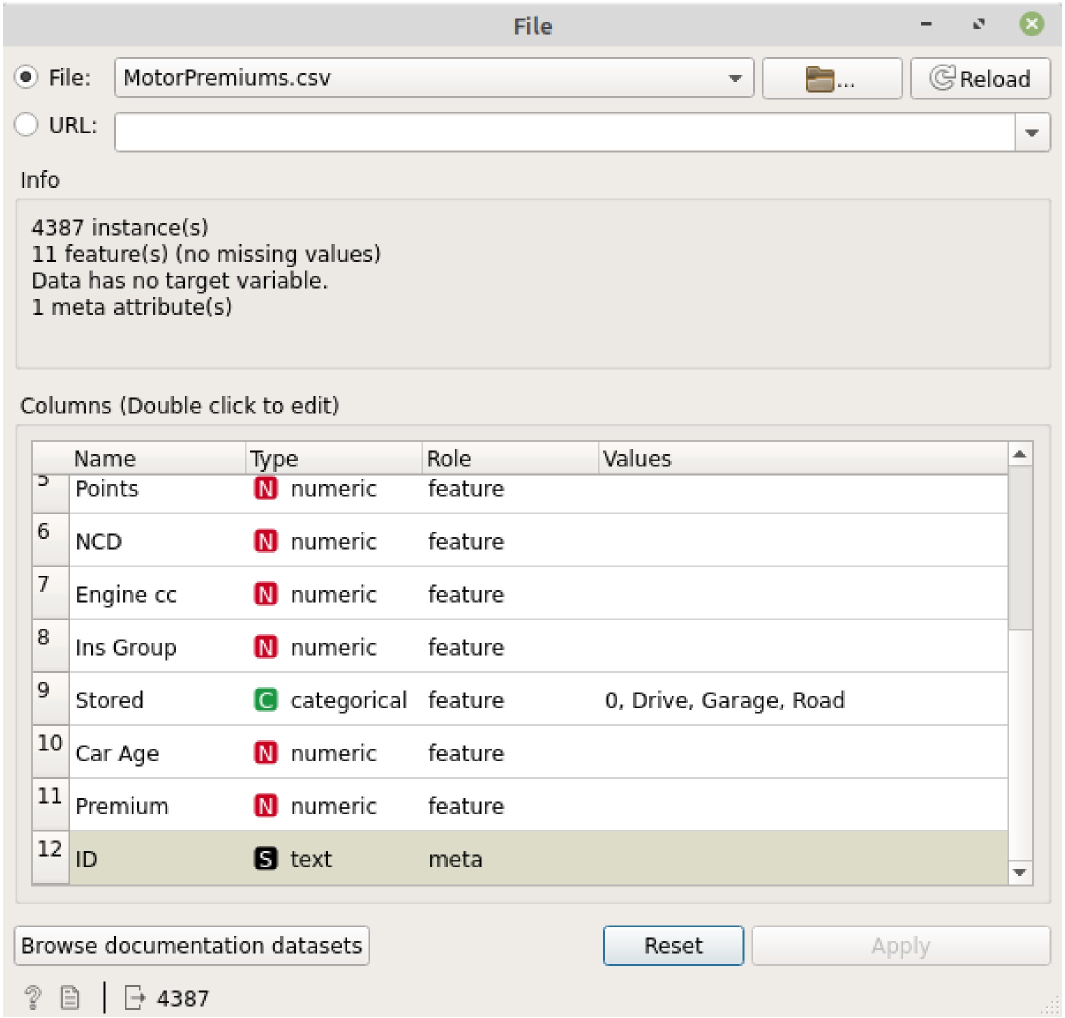
- Data (Union[Orange.data.table.Table])

The widget is then added to the workflow, looking like this:



Note the dotted lines. We can drag a line between widgets with these to make data flow between them. The File widget loads data; from Orange's perspective it is just a source for data, so there is only one dotted line, on the right, showing that data flows out of this widget. The left of the widget is its *input*, the right is its *output*.

Double click on the widget to open it. You will get a window like this:



File

● File: MotorPremiums.csv

○ URL:

Info

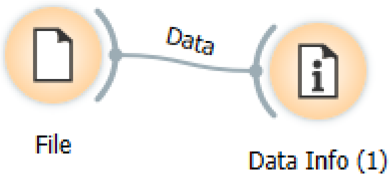
4387 instance(s)
11 feature(s) (no missing values)
Data has no target variable.
1 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
5	Points	N numeric	feature	
6	NCD	N numeric	feature	
7	Engine cc	N numeric	feature	
8	Ins Group	N numeric	feature	
9	Stored	C categorical	feature	0, Drive, Garage, Road
10	Car Age	N numeric	feature	
11	Premium	N numeric	feature	
12	ID	S text	meta	

Click on file picker (top right of window), choose the “MotorPremiums.csv” file, then OK and close the dialogue. Use the dropdowns on each row to ensure that (for example) Points and Car Age are numeric, and the other attributes are categorical as appropriate. Also change the “role” for Premium to “target” – because this is the attribute we’ll be trying to predict with our model.

Add a “Data Info” widget, from the same menu as the File widget. Drag a line to connect the two widgets together like this:



You can now double-click on the “Data Info” widget to see some summary statistics on the Data Set.

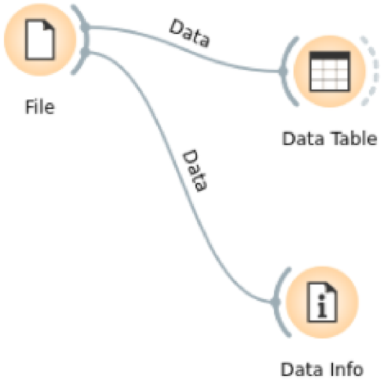
1. How many rows (instances) are there in the data?

2. How many columns (variables) are there?

Did these match what you saw in Excel?

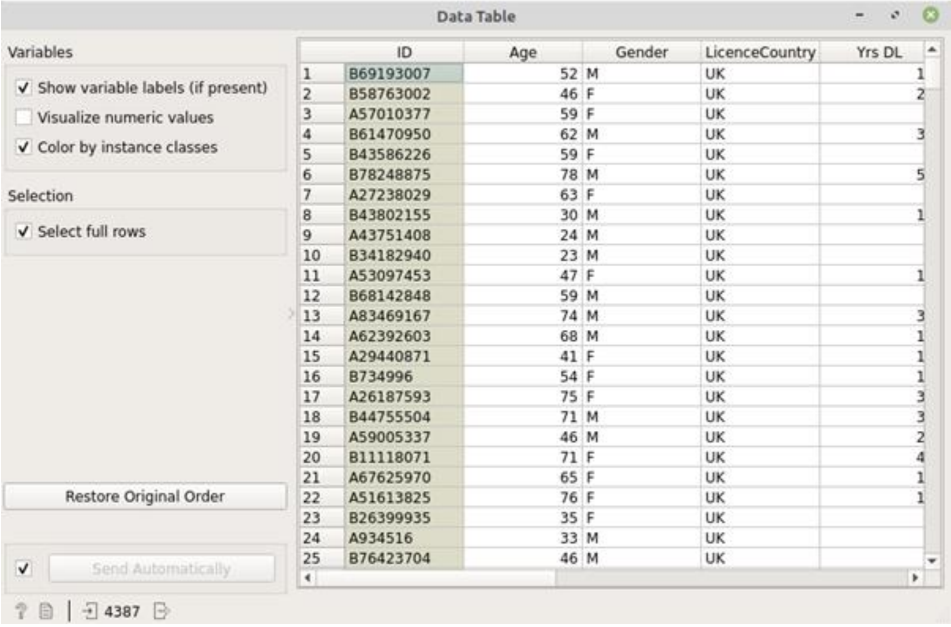
Well done: you've now successfully set up Orange, and have loaded a data set. Go to File > Save As to save your workflow. We'll come back to it in the next exercise.

Add a Data Table widget to your workflow, like this:



Data Table will let you see the raw values in your data set, much like you can in Excel or another spreadsheet. First, though, open the File widget again, by double clicking on it. You'll see that Orange has automatically assigned a type to each variable: numeric, categorical, or text (for the ID). "text" essentially means that Orange doesn't know what to do with a variable. You are able to change these assignments if you like, but don't do that just yet. Make a note of what type each variable has been categorised as.

You might have already noticed that "Numeric" and "Categorical" have not been further divided; that is just how Orange deals with things. You'll also see that Orange has assigned a "role" to each variable. "Feature" means a variable that can be used as an input for modelling. "Target" means that this is the variable we'll be trying to predict with our model. "Meta" and "Skip" both mean that the variable won't be used for modelling at all. Close the File widget, and open the Data Table widget. You should see something like this:



Data Table

Variables

☑ Show variable labels (if present)

□ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

☑ Send Automatically

4387

	ID	Age	Gender	LicenceCountry	Yrs DL
1	B69193007	52	M	UK	1
2	B58763002	46	F	UK	2
3	A57010377	59	F	UK	
4	B61470950	62	M	UK	3
5	B43586226	59	F	UK	
6	B78248875	78	M	UK	5
7	A27238029	63	F	UK	
8	B43802155	30	M	UK	1
9	A43751408	24	M	UK	
10	B34182940	23	M	UK	
11	A53097453	47	F	UK	1
12	B68142848	59	M	UK	
13	A83469167	74	M	UK	3
14	A62392603	68	M	UK	1
15	A29440871	41	F	UK	1
16	B734996	54	F	UK	1
17	A26187593	75	F	UK	3
18	B44755504	71	M	UK	3
19	A59005337	46	M	UK	2
20	B11118071	71	F	UK	4
21	A67625970	65	F	UK	1
22	A51613825	76	F	UK	1
23	B26399935	35	F	UK	
24	A934516	33	M	UK	
25	B76423704	46	M	UK	

Now, take a look at each variable in turn. Do you agree with the assignment of categorical or numeric that Orange made? Add to your notes:

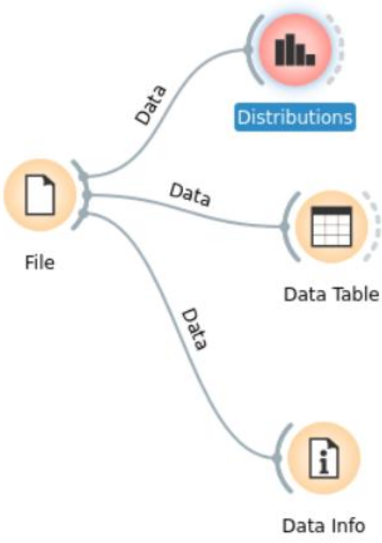
1. For the numeric variables, which are continuous and which are discrete?
2. For the categorical variables, which are nominal and which are ordinal?

Are any of the types or sub-types not present? If so, can you think of an example of a car-related variable that would be in that type?

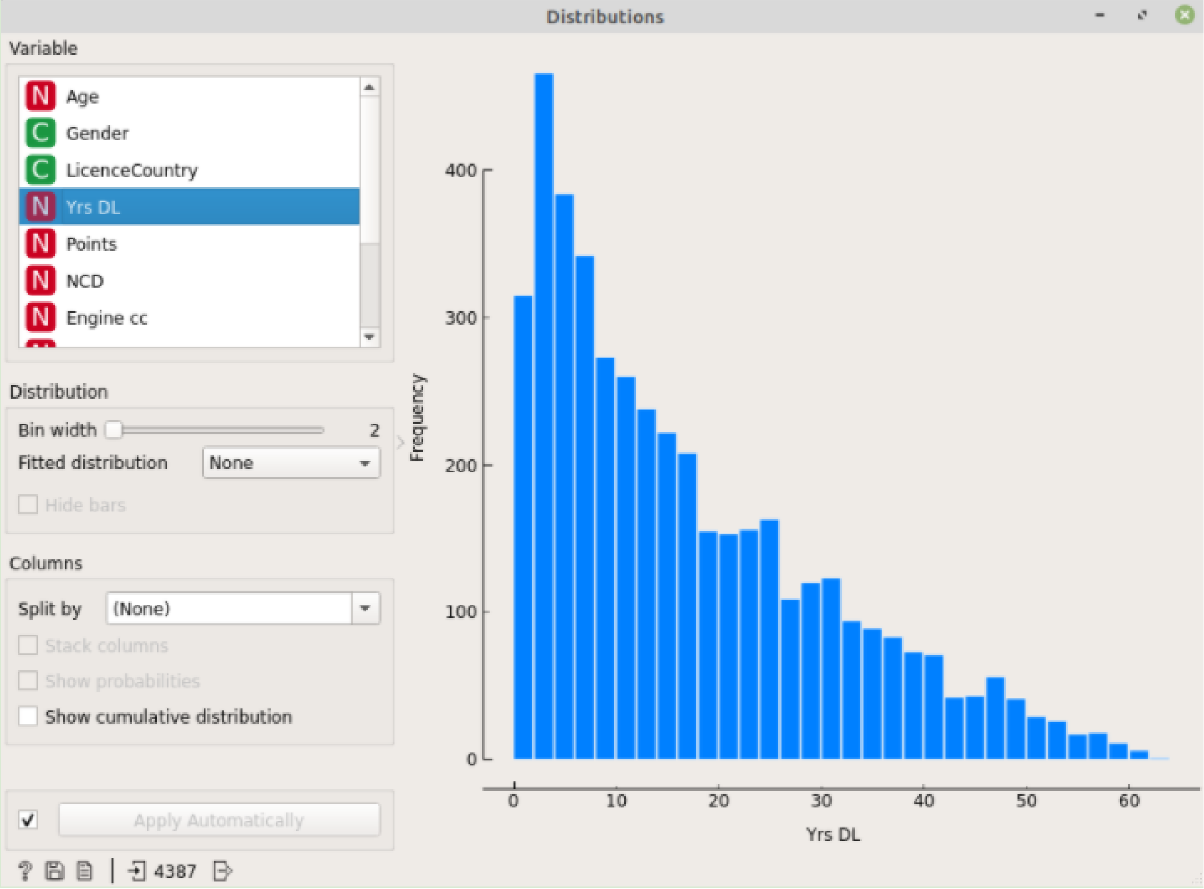
Could any of the variables arguably be considered as more than one type?

Distributions

Add a Distributions widget like this:



Open the Distributions widget, and you should get a window like this:



Distributions

Variable

N Age
C Gender
C LicenceCountry
N Yrs DL
N Points
N NCD
N Engine cc

Distribution

Bin width 2

Fitted distribution None

☐ Hide bars

Columns

Split by (None)

☐ Stack columns

☐ Show probabilities

☐ Show cumulative distribution

☑ Apply Automatically

4387

Click on the Variable names in the top-left list to see the distribution for that variable as a histogram or bar plot (depending on the variable's type). You can also choose to "Split by" one of the categorical variables so you have bars coloured according to the values in that variable, but we won't need that just yet.

Look at each variable in turn. For each one:

1. Is there an obvious trend or a shape to the distribution?
2. Are any of the variables following a normal or uniform distribution?
3. Given the trend, might this variable be suitable for modelling? (there are no right/wrong answers to this question at this stage)
4. For numeric variables, try changing the bin width: too many or too few bins and it might be hard to see the trend, so where is the ideal number? (Think about this especially for the Age variable)
5. Can you see anything else odd or noteworthy about the variables?

Data Cleaning

Open the Distributions widget, and look at each variable in turn. Make a note of the variables that show the characteristics listed below:

1. Outliers
2. Minority Values
3. Large Majority Values
4. Flat and Wide

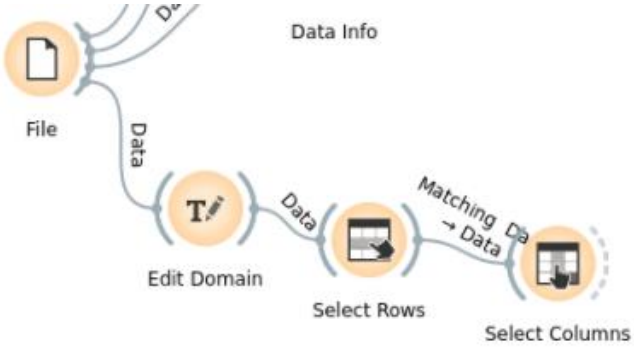
Are any of the above characteristics likely to be errors in the data?

Well, without giving too much away on the question above.... Gender appears to have "Female" sometimes when it should be "F", and likewise for "Male" and "M". "Stored" seems to have a few entries that are missing, and we might want to remove the "LicenceCountry" feature as it doesn't look all that useful.

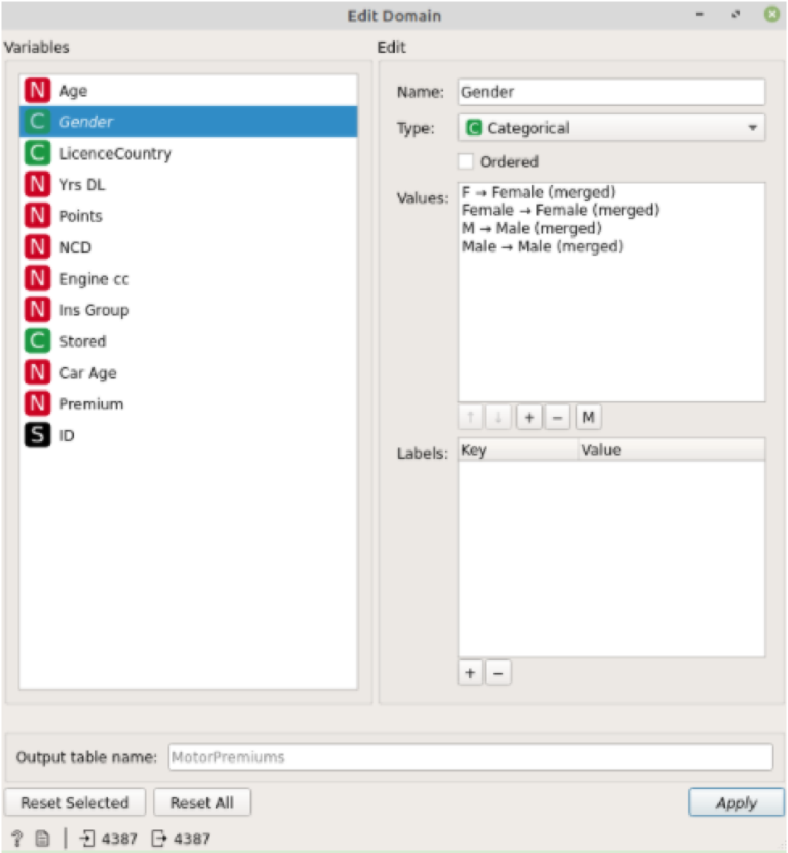
Basic cleaning

We can automatically correct the issues identified above:

Add the widgets "Edit Domain", "Select Rows" and "Select Columns" to your File widget, like this:



Edit Domain will allow you to recode the F and M to Female and Male. Left-click on Gender, then double click on "F" in the "Values" box, and type "Female", followed by enter. Do the same for "M" and "Male". The window should look like this:



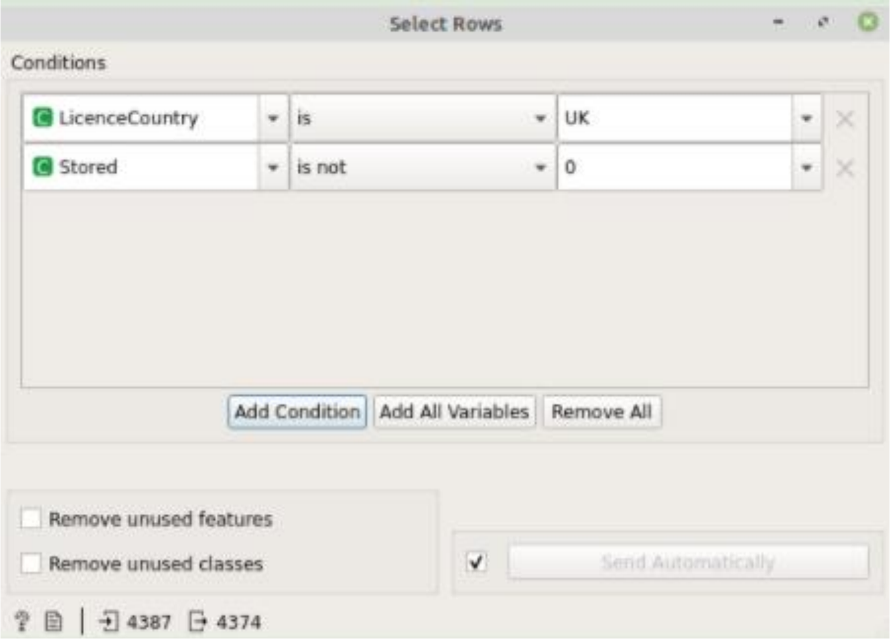
N Age
C Gender
C LicenceCountry
N Yrs DL
N Points
N NCD
N Engine cc
N Ins Group
C Stored
N Car Age
N Premium
S ID

Name: Gender
Type: Categorical
□ Ordered
Values:
F → Female (merged)
Female → Female (merged)
M → Male (merged)
Male → Male (merged)

Labels: Key Value

Output table name: MotorPremiums

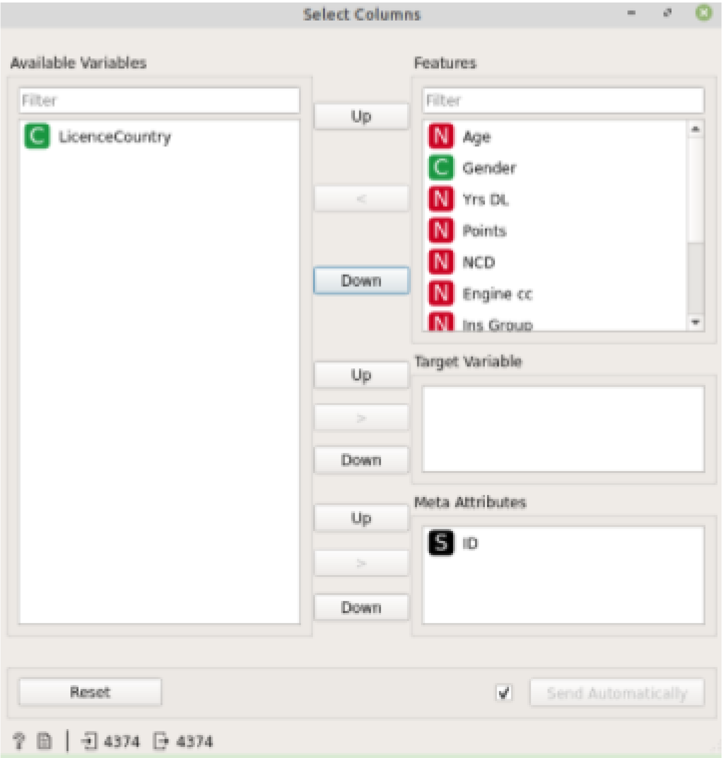
Click Apply and close the window. Then open the "Select Rows" widget. Here, we are choosing which elements of data to keep. We'll keep the records for customers in the UK, and for whom we are not missing the information about where their car is stored:



LicenceCountry	is	UK
Stored	is not	0

□ Remove unused features
□ Remove unused classes
☑ Send Automatically

Close this window, and now open the "Select Columns" widget. Click on LicenceCountry and the left arrow; this means that LicenceCountry will not be included in the features that can be used for modelling after this widget. Note: we could have click on "Remove unused features” in the “Select Rows” widget to do this automatically, but it’s useful to see how to explicitly drop features you don’t want too). The window should look like this:



Select Columns

Available Variables

- C LicenceCountry

Features

- N Age
- C Gender
- N Yrs DL
- N Points
- N NCD
- N Engine cc
- N Ins Group

Target Variable

Meta Attributes

- S ID

☑ Send Automatically

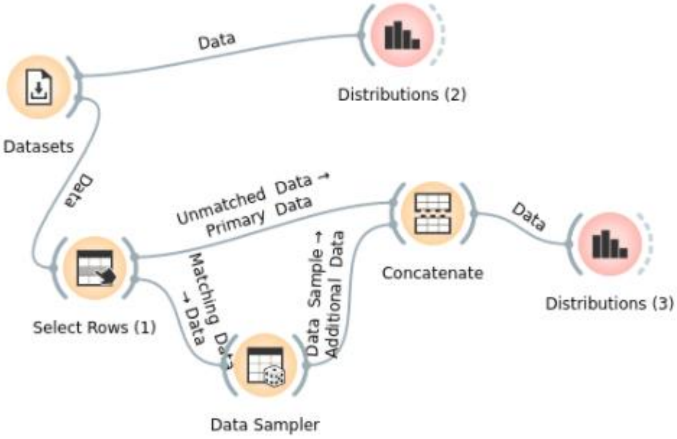
4374 4374

Close the window. You can now add another distributions widget after “Select Columns” to see how your data looks now. The errors should all now be gone.

There are several widgets that can help you with preprocessing (including a “preprocess” widget). Spend some time looking at the documentation for Orange to see what these can do (in the Help menu, under Documentation).

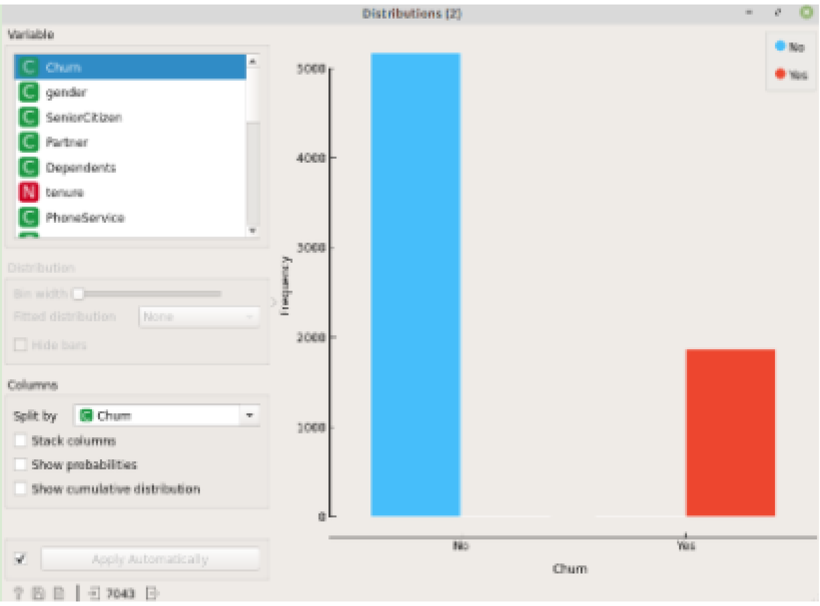
Unbalanced Data

We’ve one more thing to try. What if your data is unbalanced? How do we sample it down to rebalance it? For this we’ll use one of Orange’s built-in data sets. Add widgets and connections to your workflow like this:



(don't worry if the numbers after your widget names are different. They're just labels so they don't affect the outcomes. It's the connections that matter)

Double click on the Datasets widget, and on the list of datasets, double click on "Telecom customer churn". Close the window. If you look at the top Distributions widget (Distributions (2) in the picture above), you'll see the data is unbalanced. It could be worse, but this would definitely still cause problems for modelling:



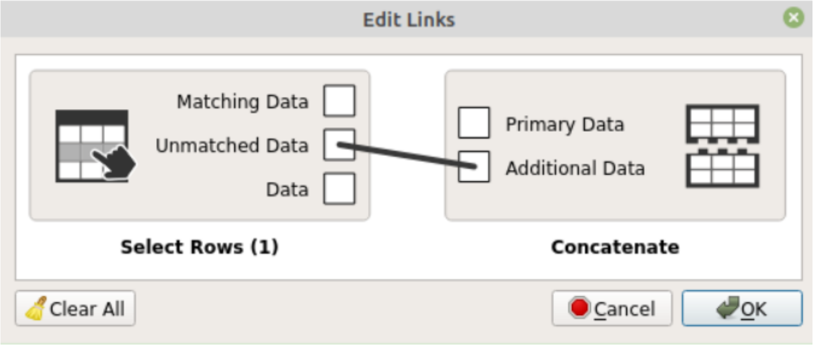
It's saying that customers were much more likely to stay with the company than move elsewhere. In fact, if you hover over the bars, Orange will tell you there are 1869 for "yes" and 5174 for "no". Ideally we'd like to delete some of the "no" customers until we have 1869 for both category.

To rebalance this data, we go through several steps:

1. We split the data into the two categories ("yes" and "no"), using the Select Rows (1) widget. Open the widget and make the condition "Churn" "is" "No".

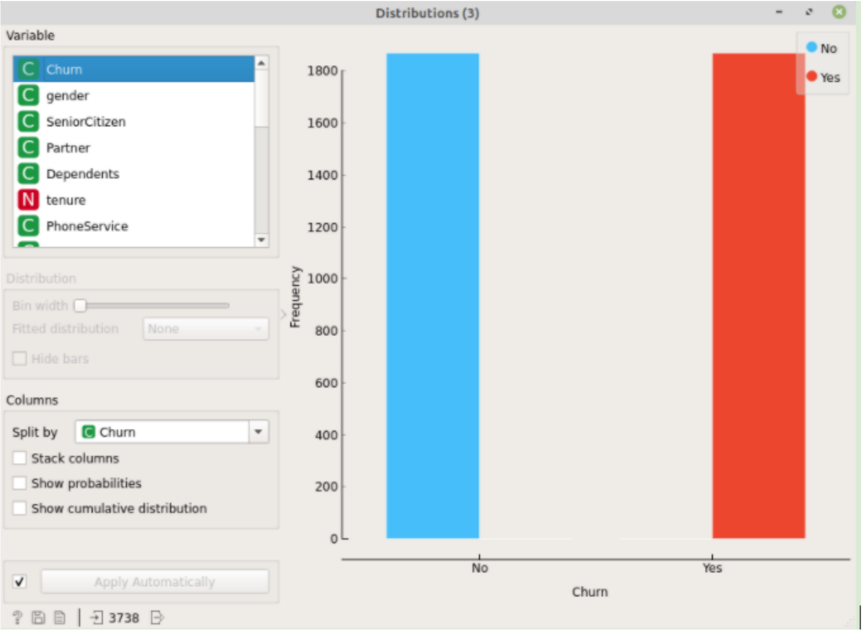
2. We choose a random sample of the larger category ("no"), using the Data Sampler widget. Open the widget and choose a "Fixed sample size" of 1869 instances.

3. We merge together the sampled "no" data with all of the "yes" data, using a Concatenate widget. For this to work, double click on the line connecting Select Rows (1) and Concatenate. Click on the line connecting "Matching Data" to "Additional Data" to delete it, then drag a line from "Unmatched Data" to "Additional Data"; the window should look like this:



This means the "unmatched data", that is, the "Yes" data, is passed from Select Rows (1) to Concatenate. Meanwhile, the "matched data" (the "no" data), still goes through to the Data Sampler.

4. Open the Distributions (3) widget to see what the data now looks like. You should see that it's evenly balanced:



Once you have completed this, you are ready to move on!