Badge2, Lab-2 [Data Preprocessing]

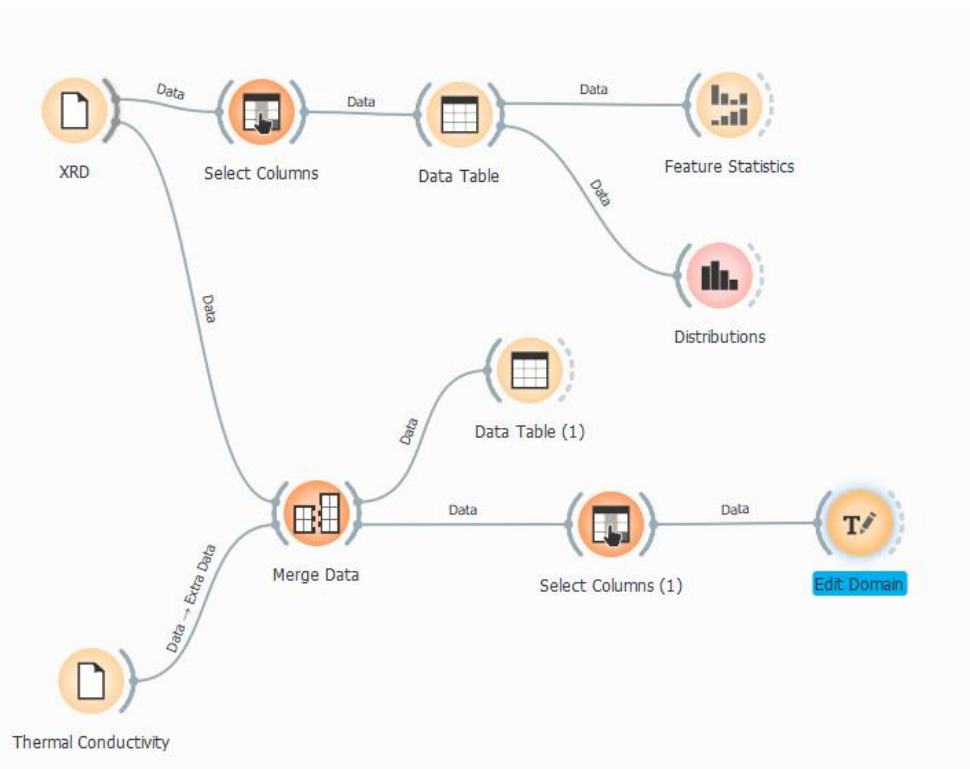
Out date: Jul 13, 2022
Due date: Jul 17, 2022 at 11:59pm

Objective: To review and understand data preprocessing techniques available in Orange and how to use it to improve model performance.

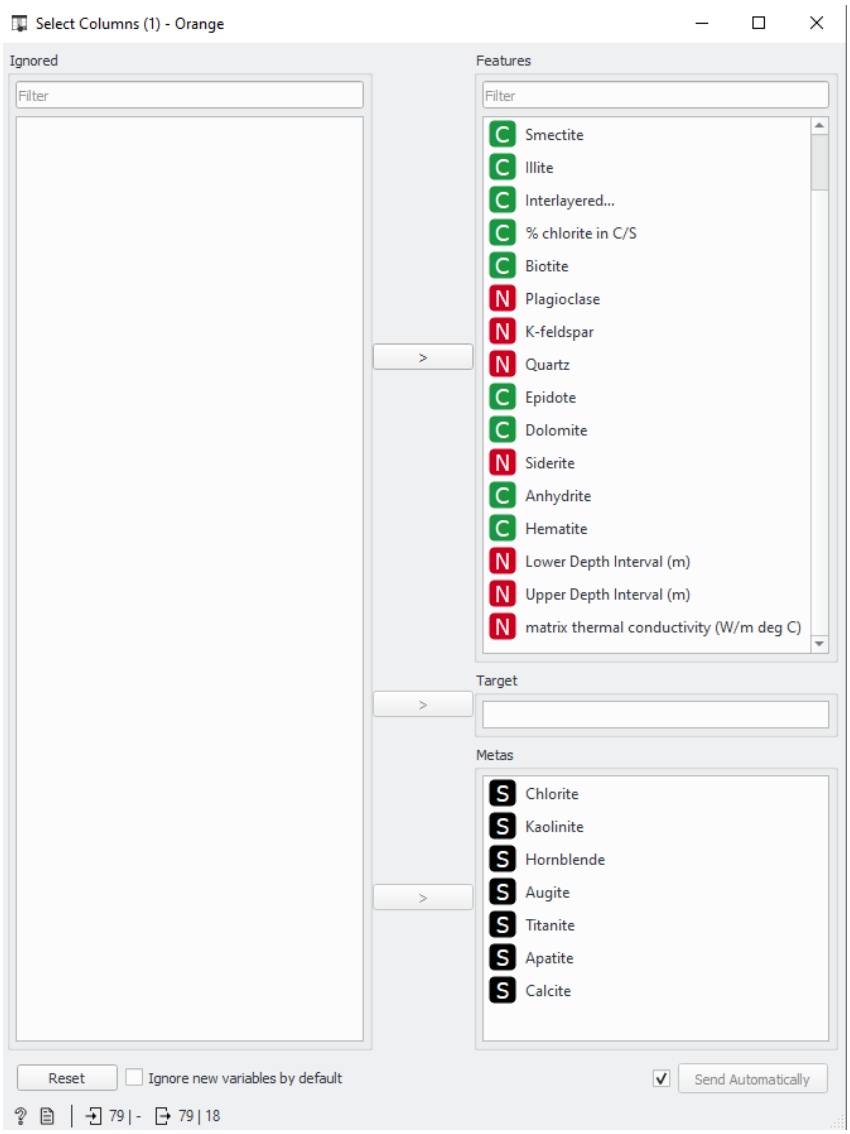
Lab Instructions

For this lab, we will continue with the Orange file that we created for Badge-1, Lab2.

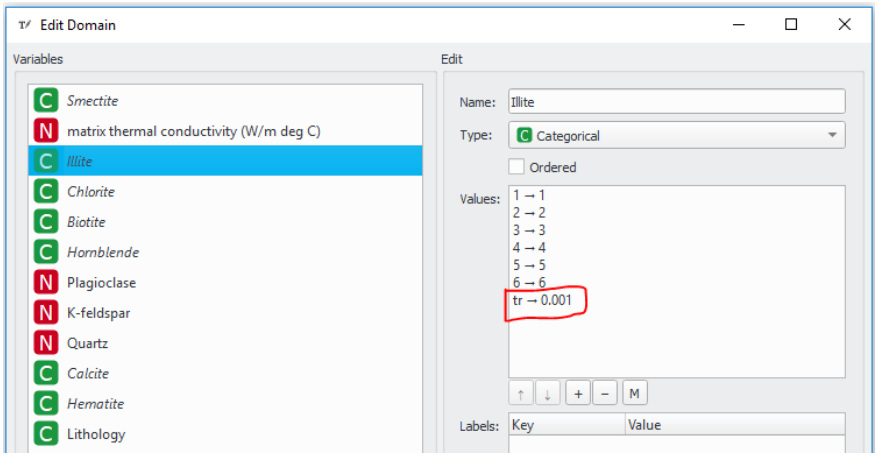
1. Add **Select Columns** and **Edit Domain** widget to your pipeline as shown below:



2. Open **Select Columns** widget and confirm feature selection is as shown below. Use **Feature Statistics** to confirm that only features with <60% missing values are selected.

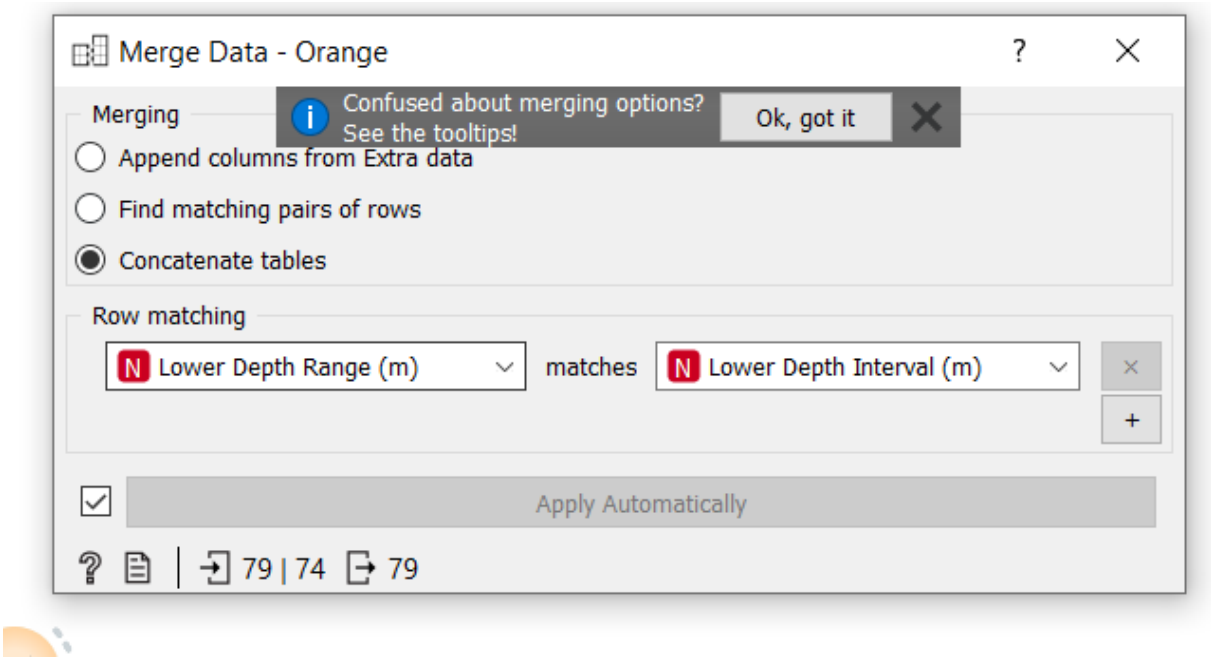


3. Open **Edit Domain** widget and edit as shown below for all categorical features with *t* or *tr* :

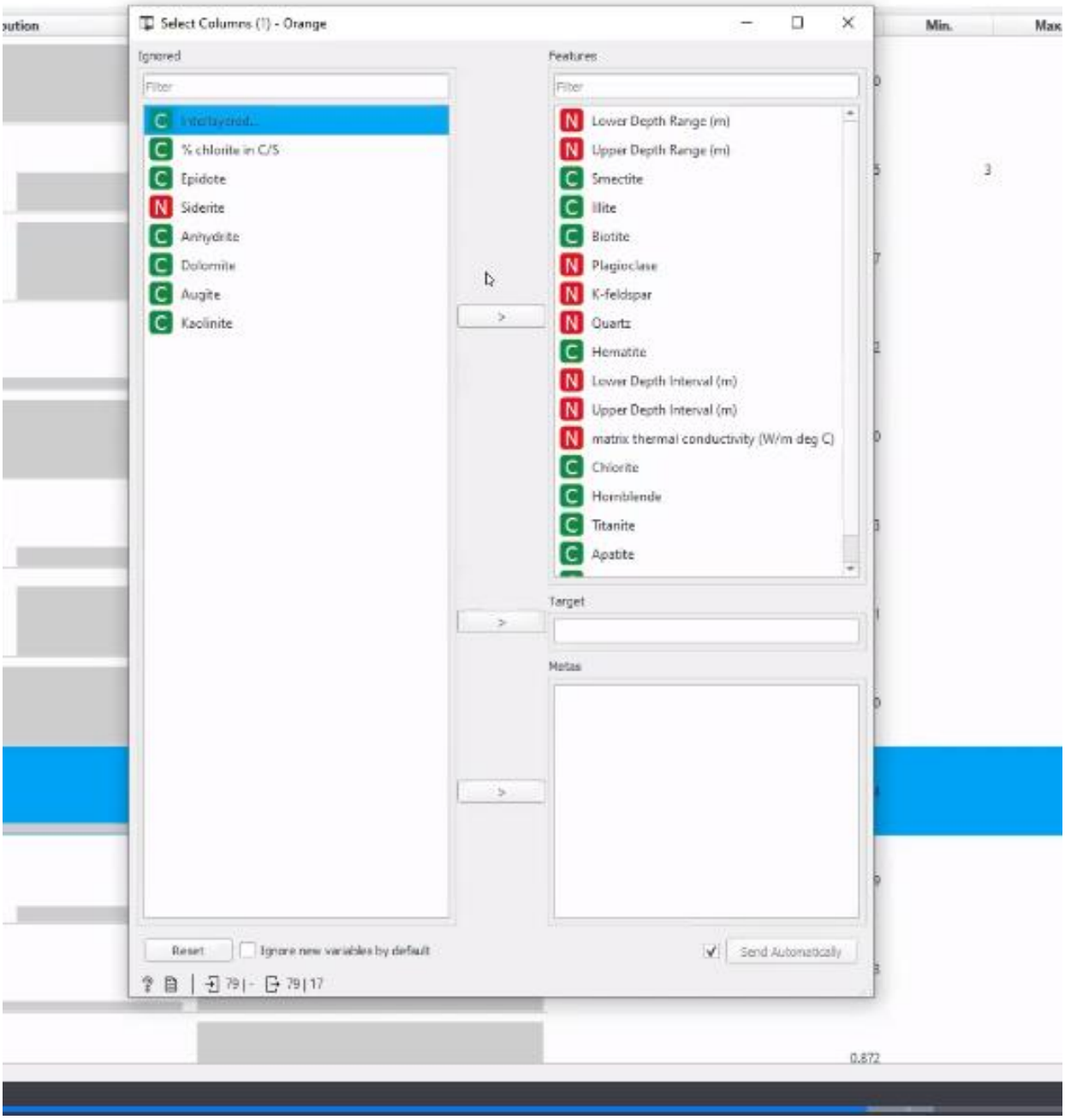


MAKE SURE XRD ALL TEXTS/META ARE CATEGORICAL/FEATURE

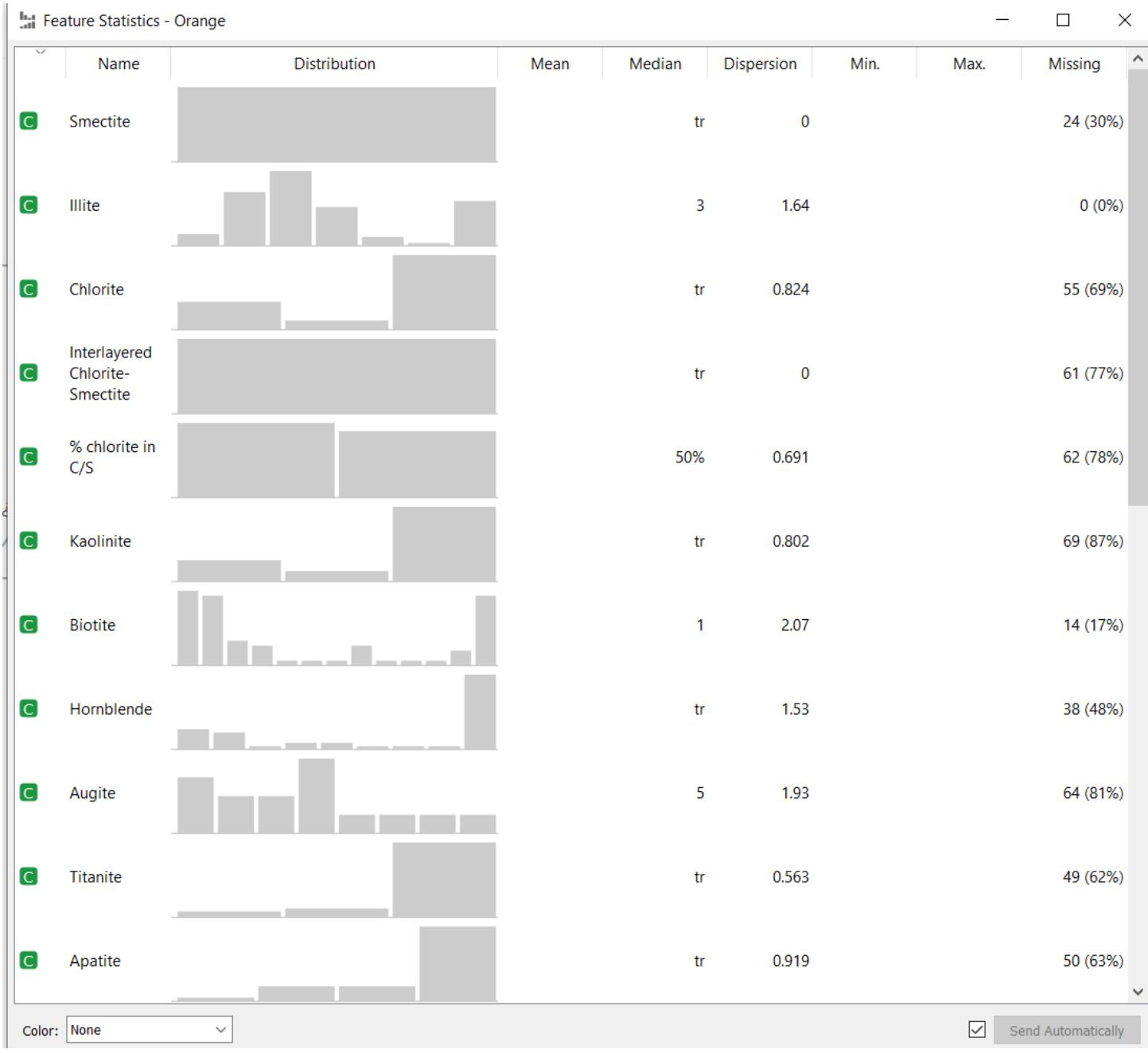
MAKE SURE:


Merge Data - Orange
Merging
Confused about merging options?
See the tooltips!
Ok, got it
Append columns from Extra data
Find matching pairs of rows
Concatenate tables
Row matching
Lower Depth Range (m)
matches
Lower Depth Interval (m)
Apply Automatically
79 | 74
79

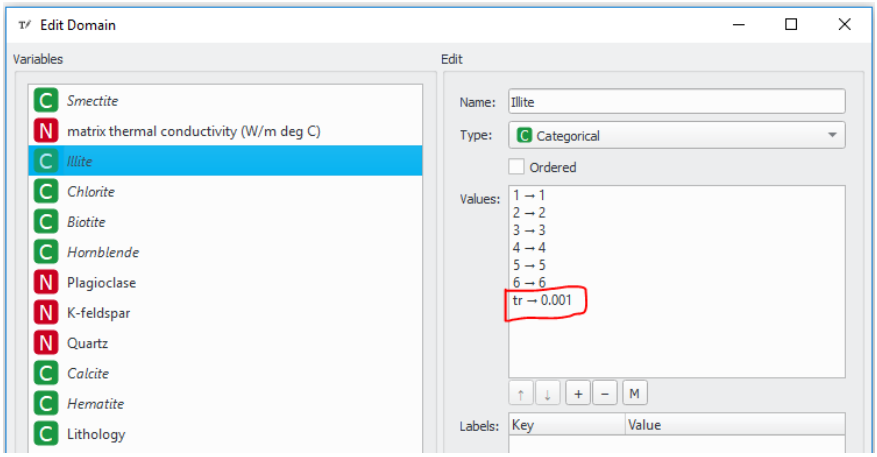
Use **Feature Statistics** to confirm that only features with <60% missing values are selected.



For Example:


Feature Statistics - Orange
Name
Distribution
Mean
Median
Dispersion
Min.
Max.
Missing
Smectite
tr
0
24 (30%)
Illite
3
1.64
0 (0%)
Chlorite
tr
0.824
55 (69%)
Interlayered Chlorite-Smectite
tr
0
61 (77%)
% chlorite in C/S
50%
0.691
62 (78%)
Kaolinite
tr
0.802
69 (87%)
Biotite
1
2.07
14 (17%)
Hornblende
tr
1.53
38 (48%)
Augite
5
1.93
64 (81%)
Titanite
tr
0.563
49 (62%)
Apatite
tr
0.919
50 (63%)
Color:
None
Send Automatically

Open **Edit Domain** widget and edit as shown below for all categorical features with *t* or *tr* :

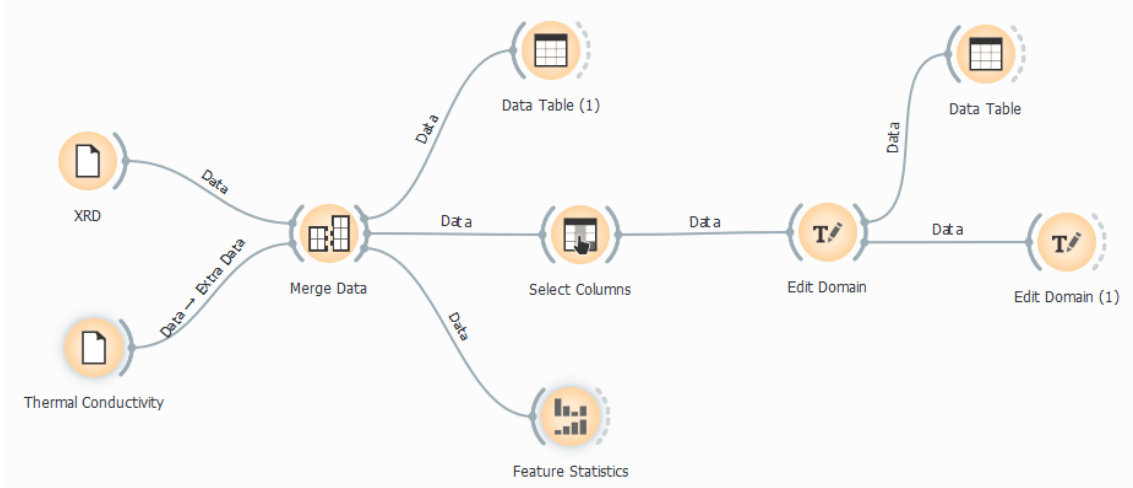


4. Add a **Data Table** widget to inspect the output of above step.

Do you see any 'tr' or 't' in the data table?	**yes**

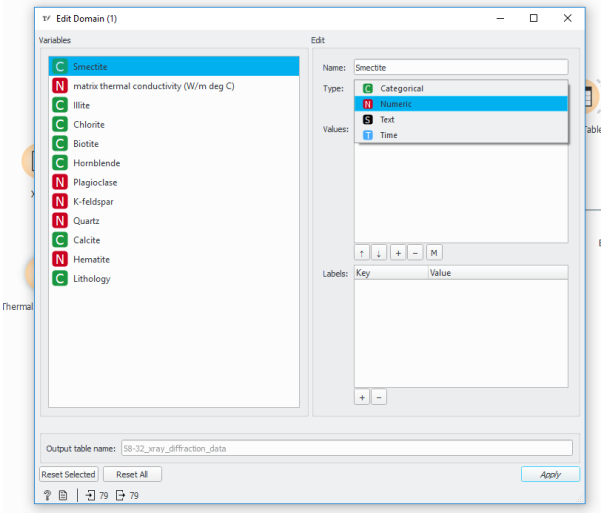
Replace tr and t to 0.001

5. Let us convert the Categorical features to Numeric now by adding another **Edit Domain** widget as shown below:

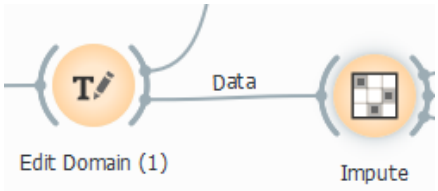


We do this (adding another Edit Domain-two step process) because it will not consider 0.001 value anymore or won't show it, the data table will show columns of ?'s/missing values)

6. Open **Edit Domain(1)** widget and change feature type to Numeric for all the features, as shown below. Verify and click apply.

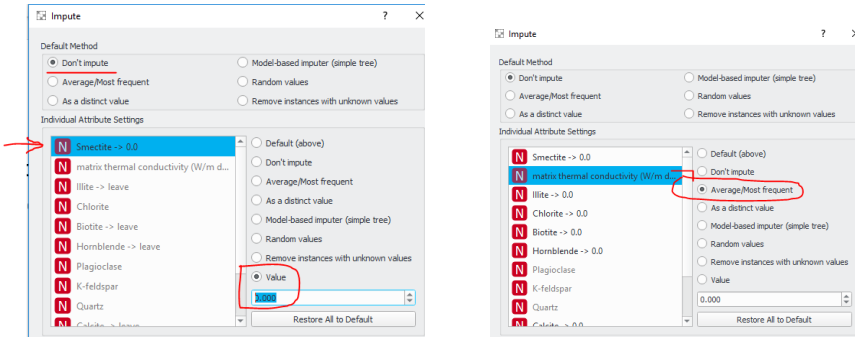


7. Add **Impute** widget to the above pipeline to address the missing values indicated by '? '.

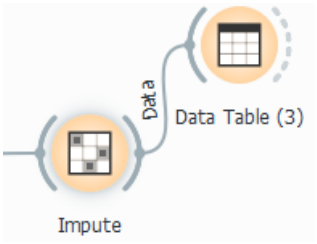


What do the missing values (?) for the categorical features converted to numeric in earlier step indicate in this data table? – Step 6	Need to impute missing values; didn't find those minerals, didn't register anything; some of them could have been a fixed value depending on what it is instead of a ?, where we need to impute these manually (thermal conductivity should be imputed as a mean, not zero tho)

8. Open and edit the **Impute** widget as shown below for features with missing values:

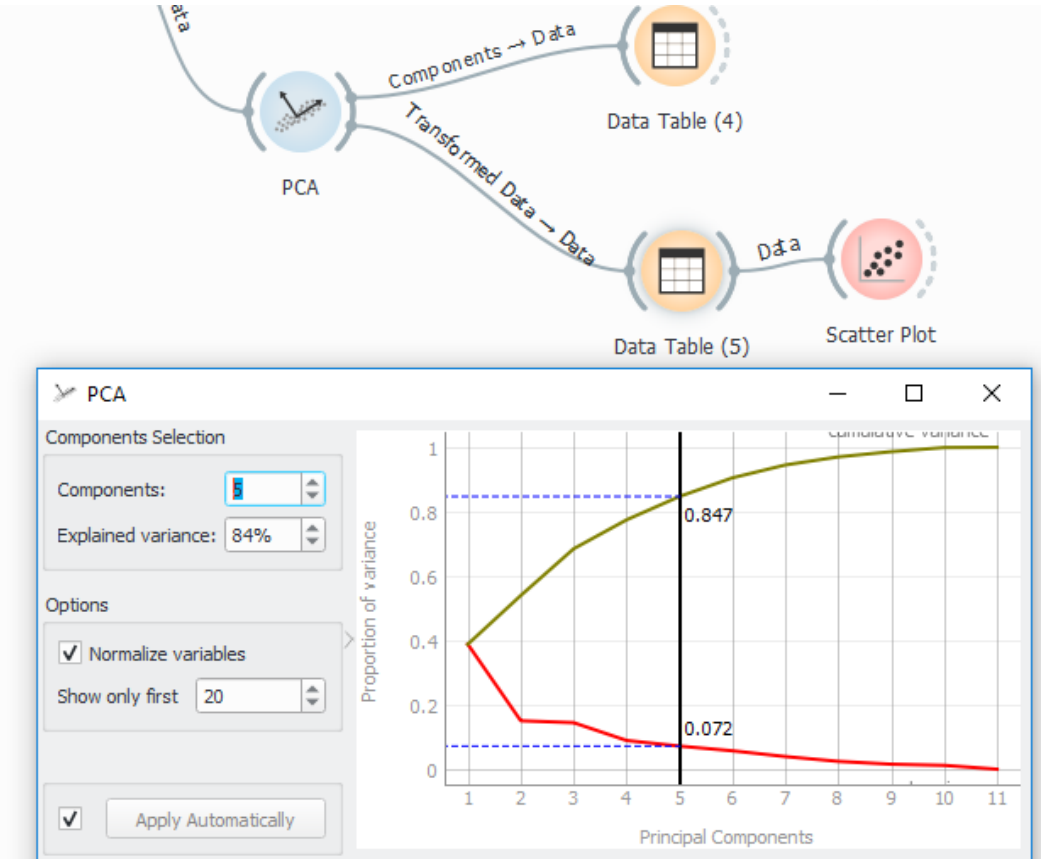


9. Inspect changes by adding a **Data Table** widget to Impute.



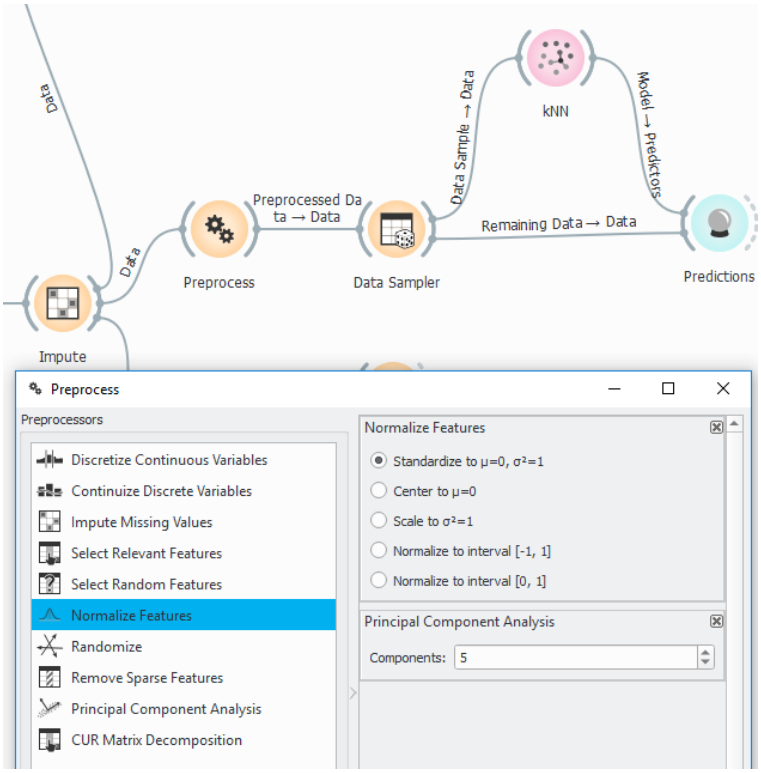
What do you observe?	There is no missing data (?), been replaced with what we imputed, like 0 and average

10. Let us do PCA. Add **PCA** widget as shown below and open the widget. Inspect the Components and Transformed Data Tables and the Scatter Plot widgets.

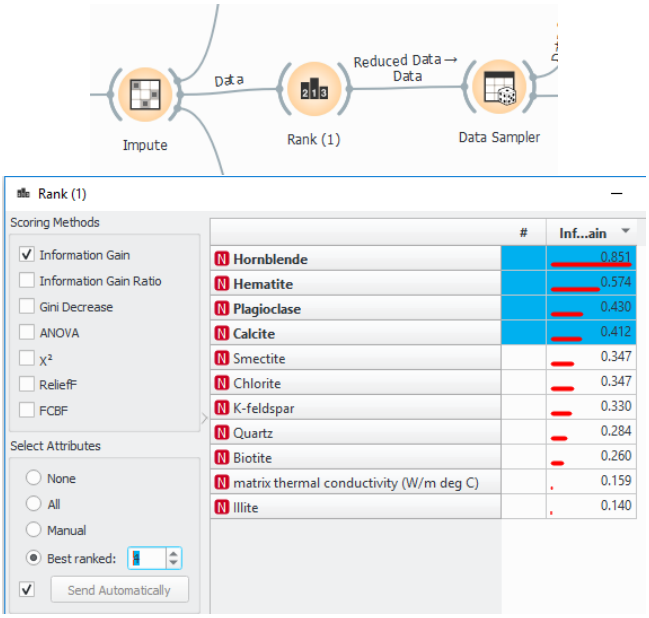


How many components are needed to get explained variance of at least 80%	5
Uncheck Normalize variables. How many components are needed now to get explained variance of at least 80%?	1

11. Let us add **Preprocess** widget to the pipeline as shown below:



12. Let us use the **Rank** widget to understand the importance of the features. Add this widget to the **Impute** widget and open. Select Information Gain as the Scoring Method and understand the rank of the features.



What are the top 2 ranked features?	Quartz and Smectite