University of Minho
School of Engineering

ISLab

Machine Learning with Knime

Similarity Based Systems

Perfil ML:FA@MiEI/4º ano - 1º Semestre
@MES/2º ano - 1º Semestre
Bruno Fernandes, Paulo Novais

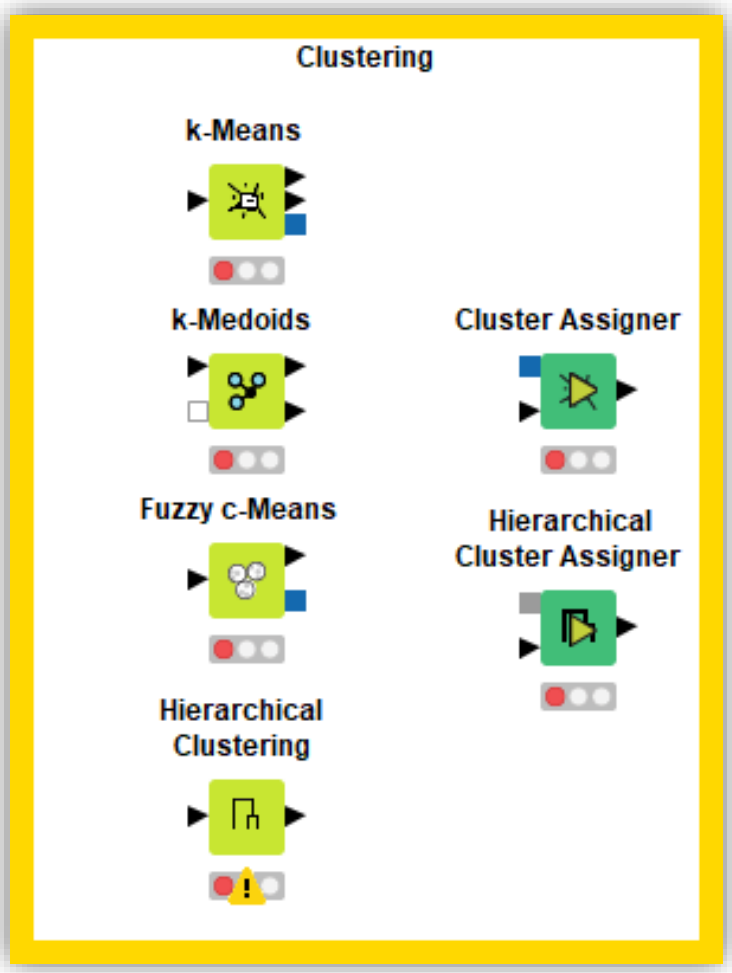
05/12/2019

Contents

- Clustering and Recommender Systems
  - The Elbow Method
- Quality measures
  - MAE
  - MSE
  - RMSE
- HTTP Requests (API calls)
- Hands On

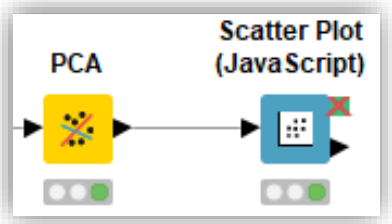
Clustering and Recommender Systems

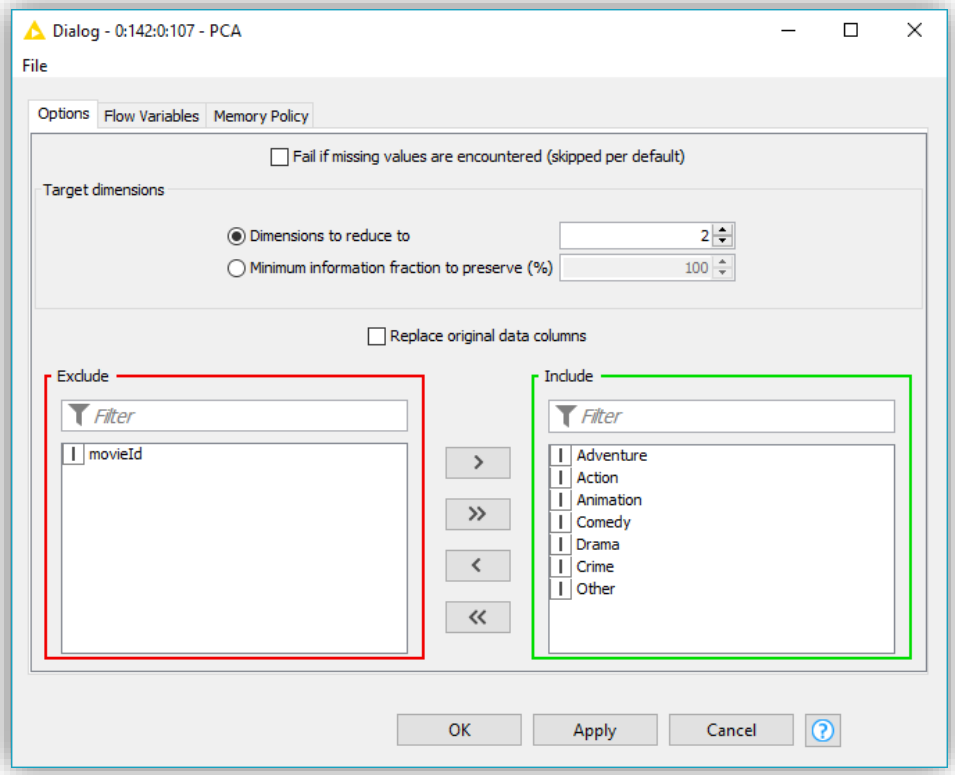
Clustering and Recommender Systems

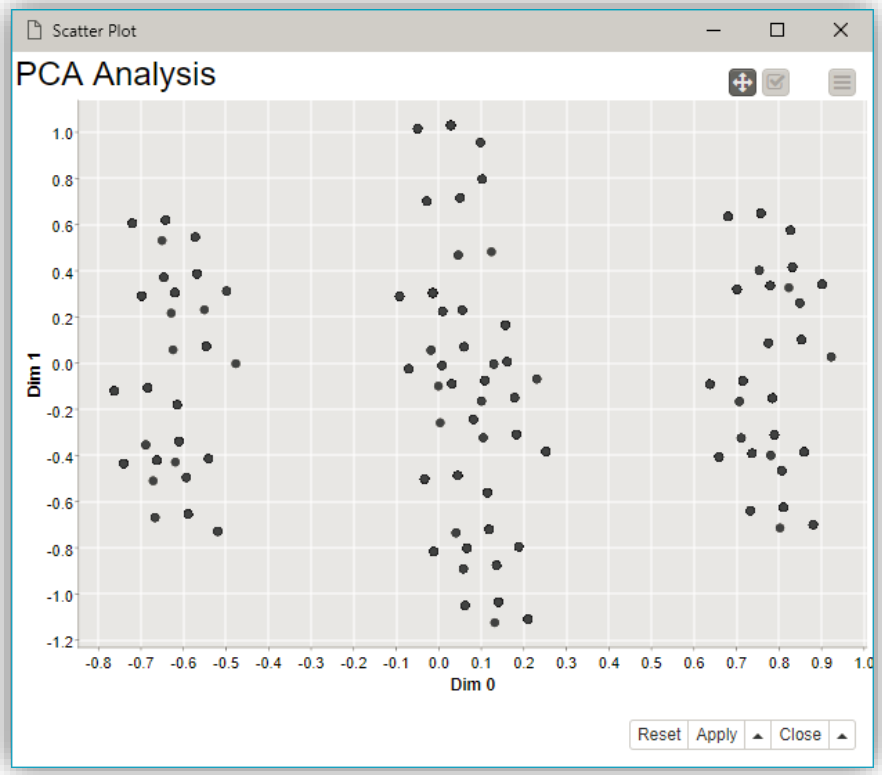


Is it clustered?

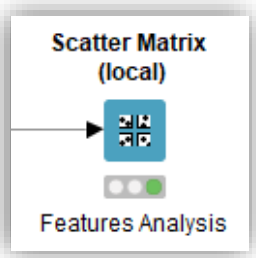
Principal Components Analysis

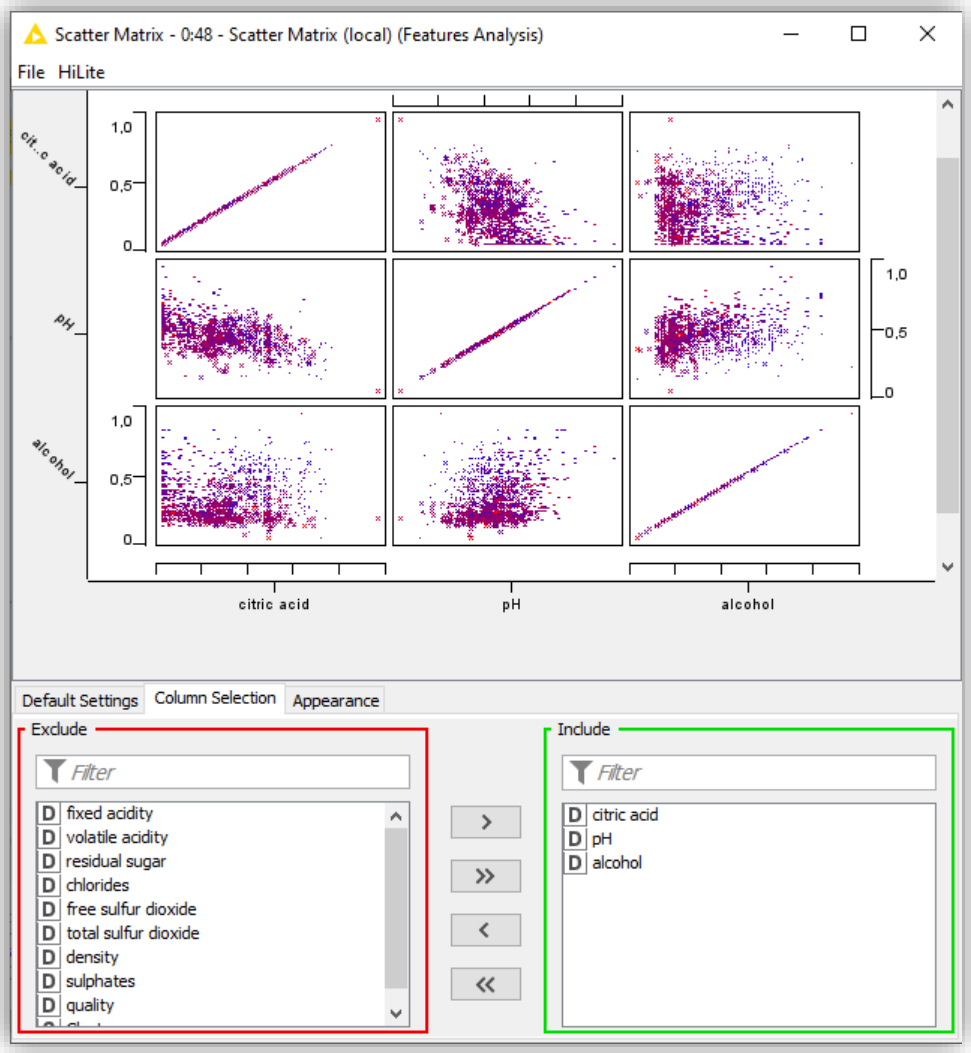




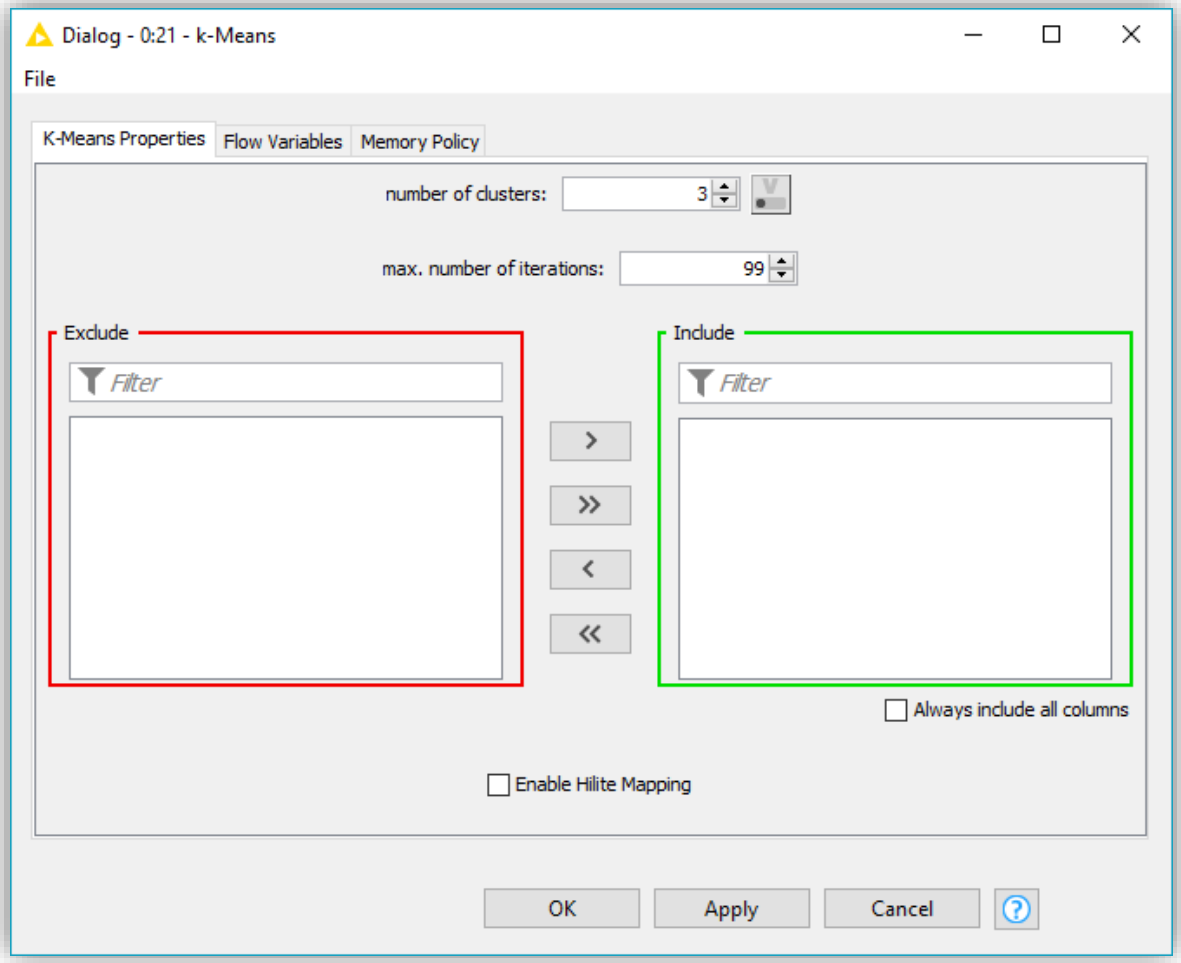


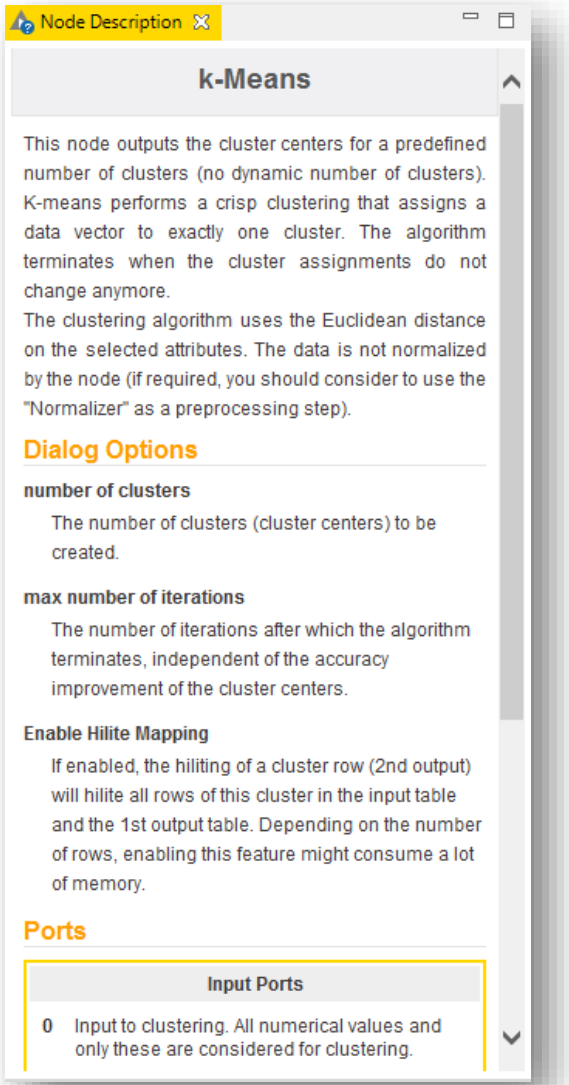
Is it clustered?
Scatter Matrix





K-Means Settings

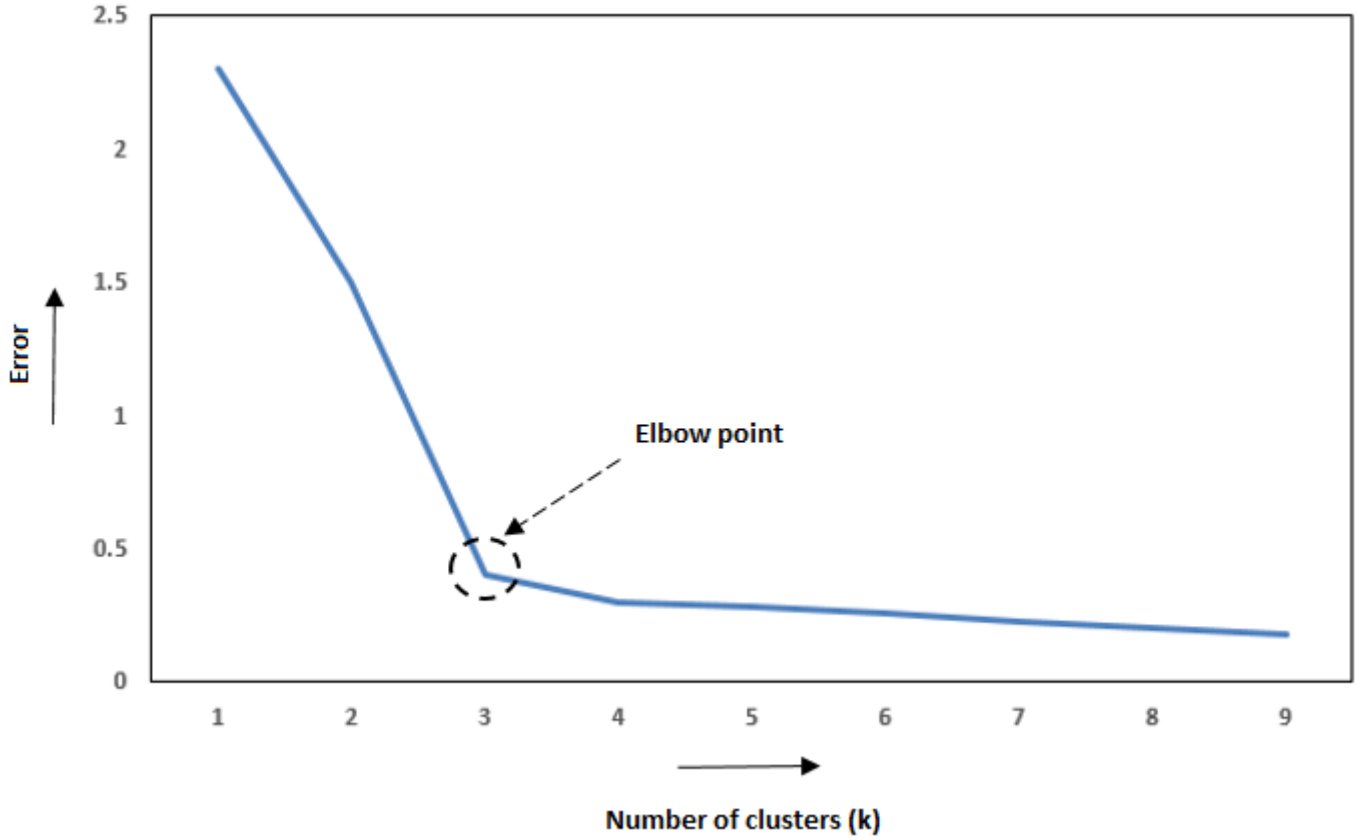




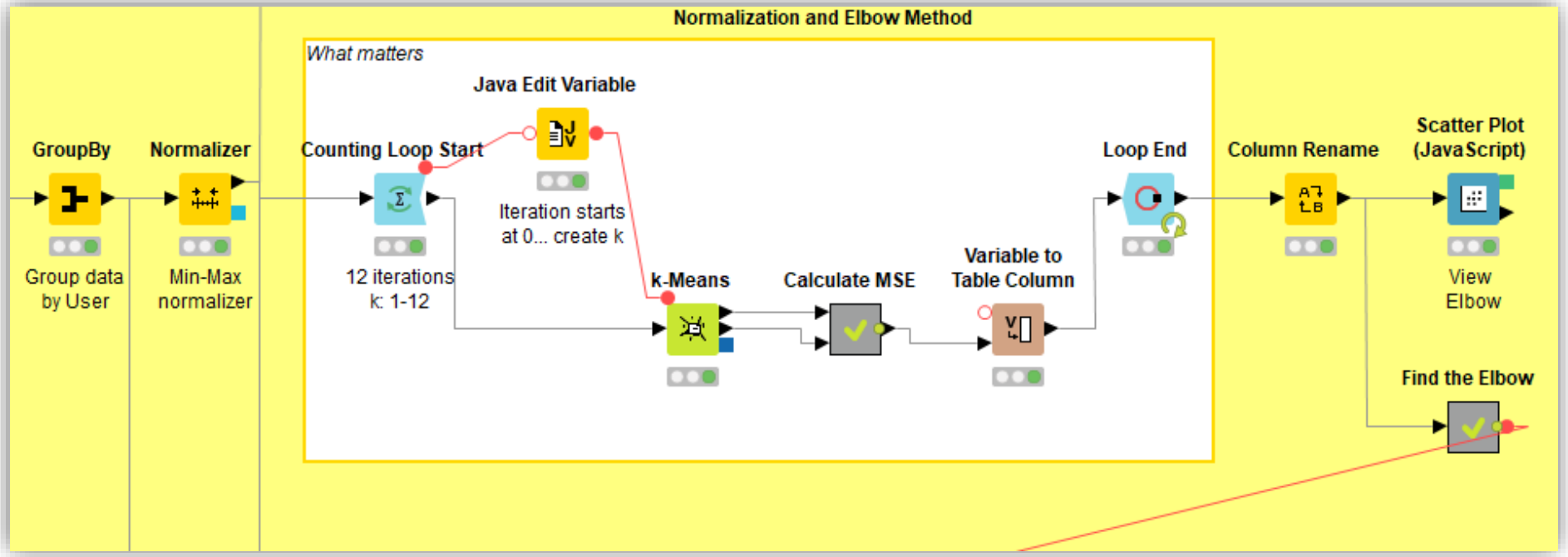
But … How many clusters?

The Elbow Method

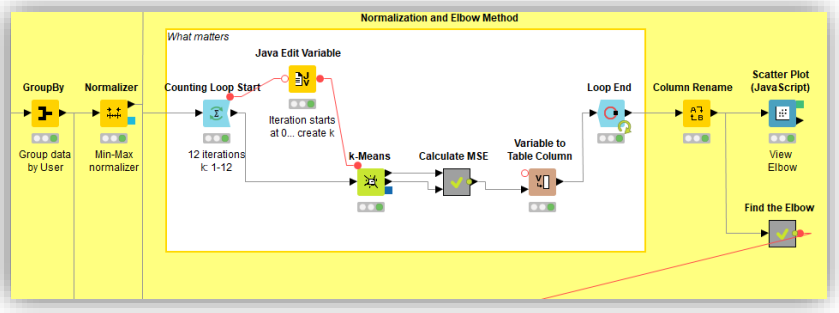
The Elbow Method

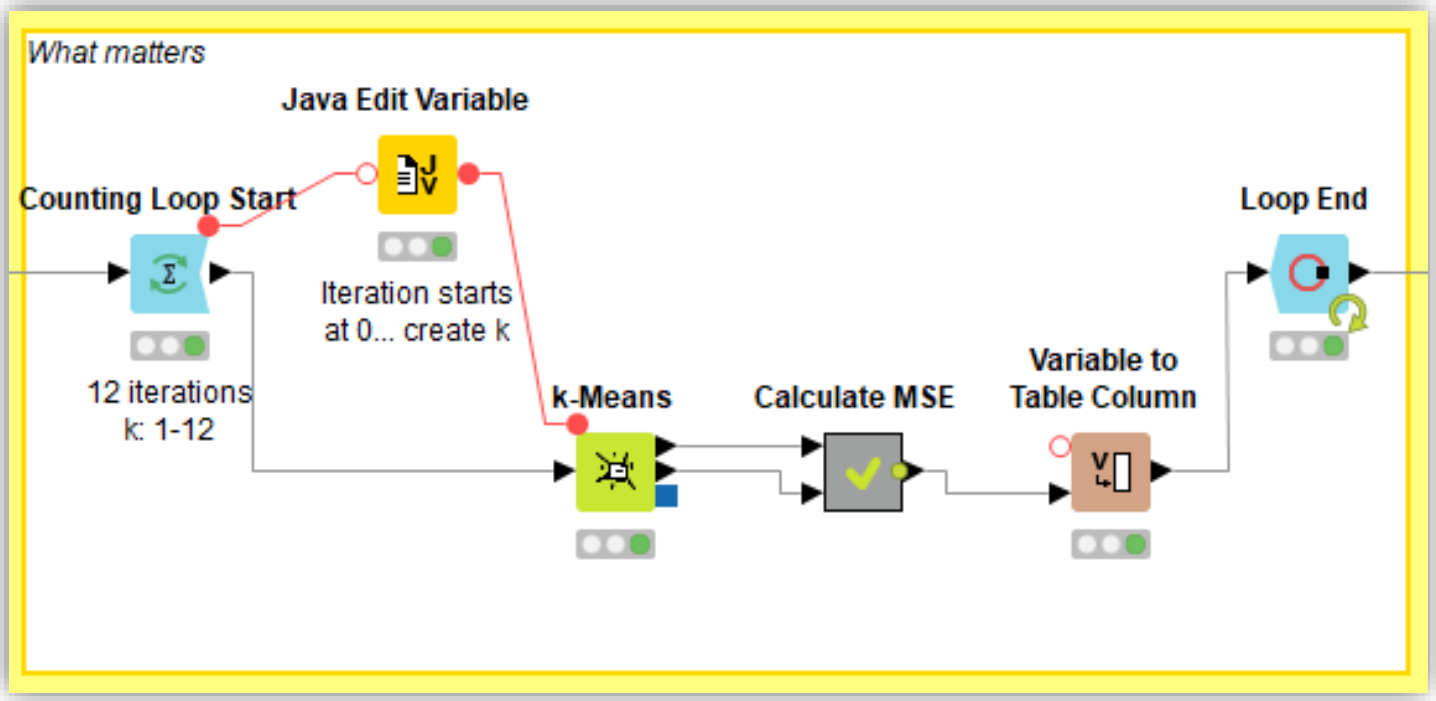


The Elbow Method



The Elbow Method



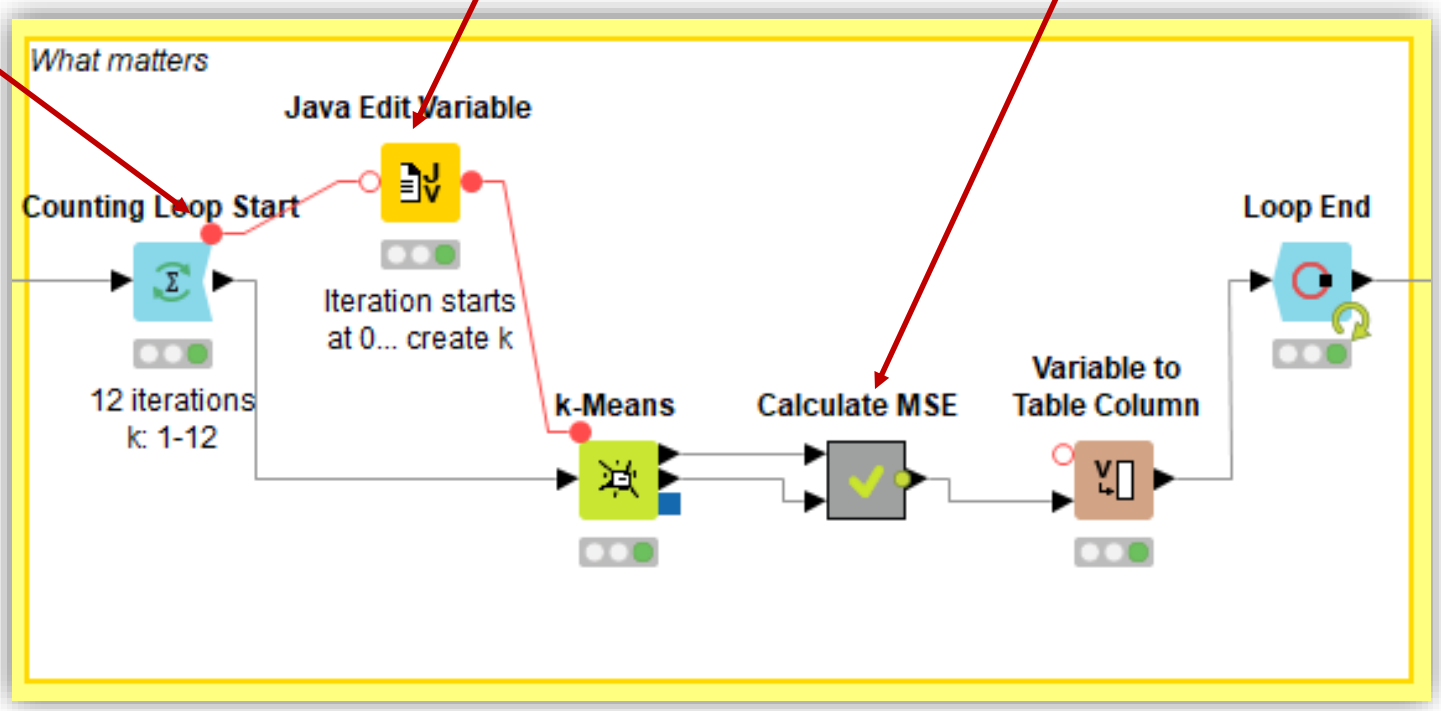


The Elbow Method

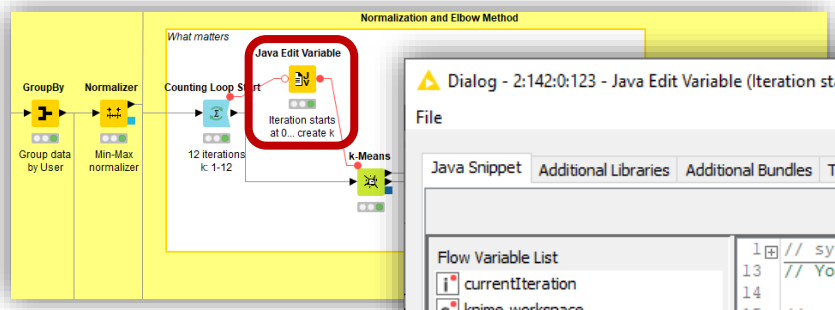
Flow Variable made available by the Counting Loop Start: *currentIteration*

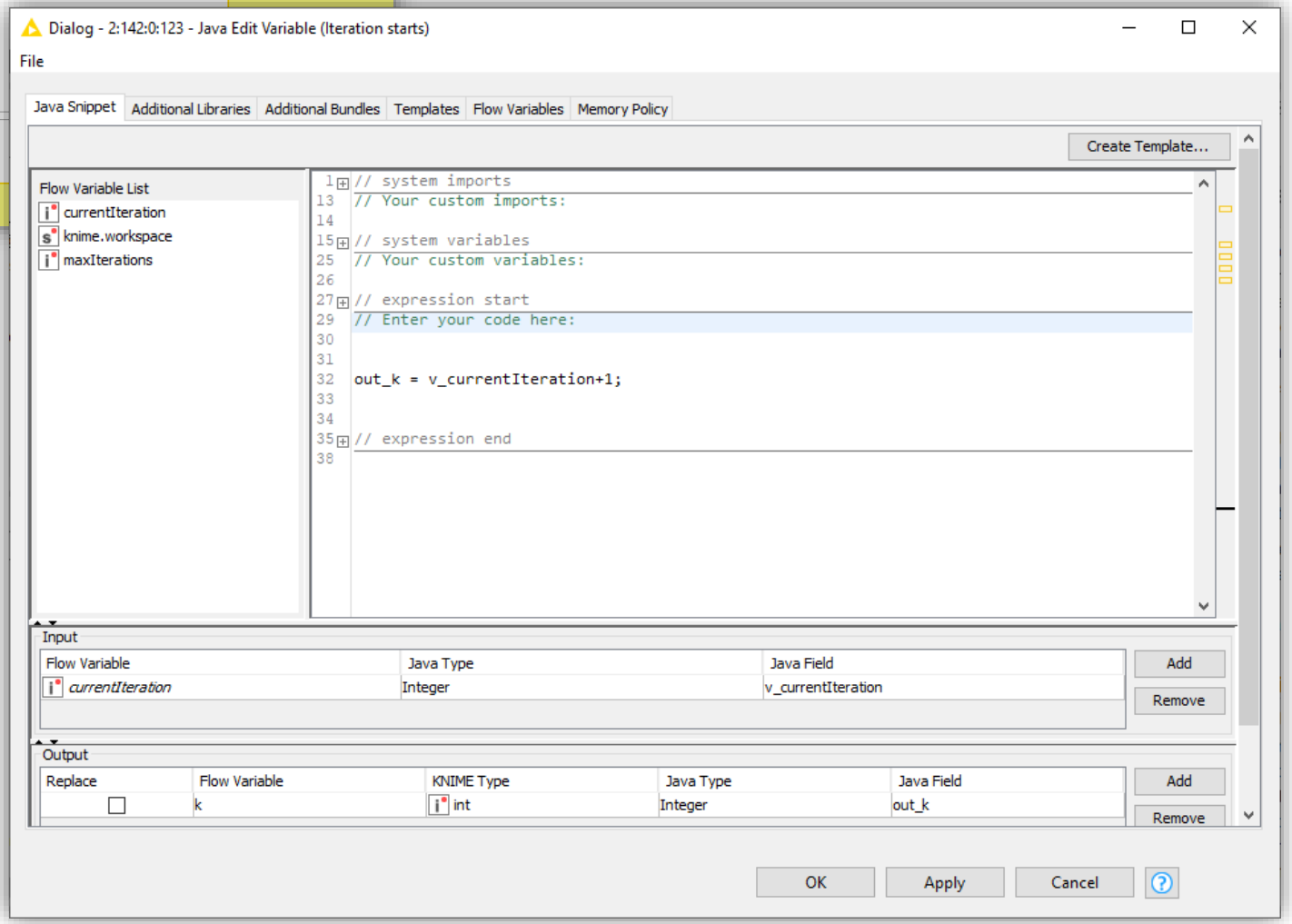
Makes use of *currentIteration* to define the k of k-Means. Adds one because it starts at 0!

Calculates a error metric (MSE) to quantify k!

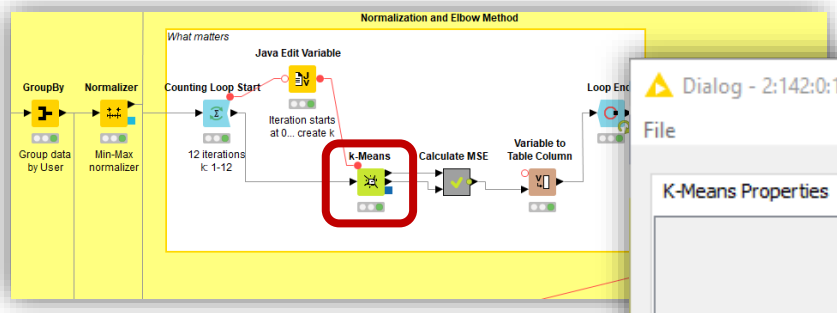


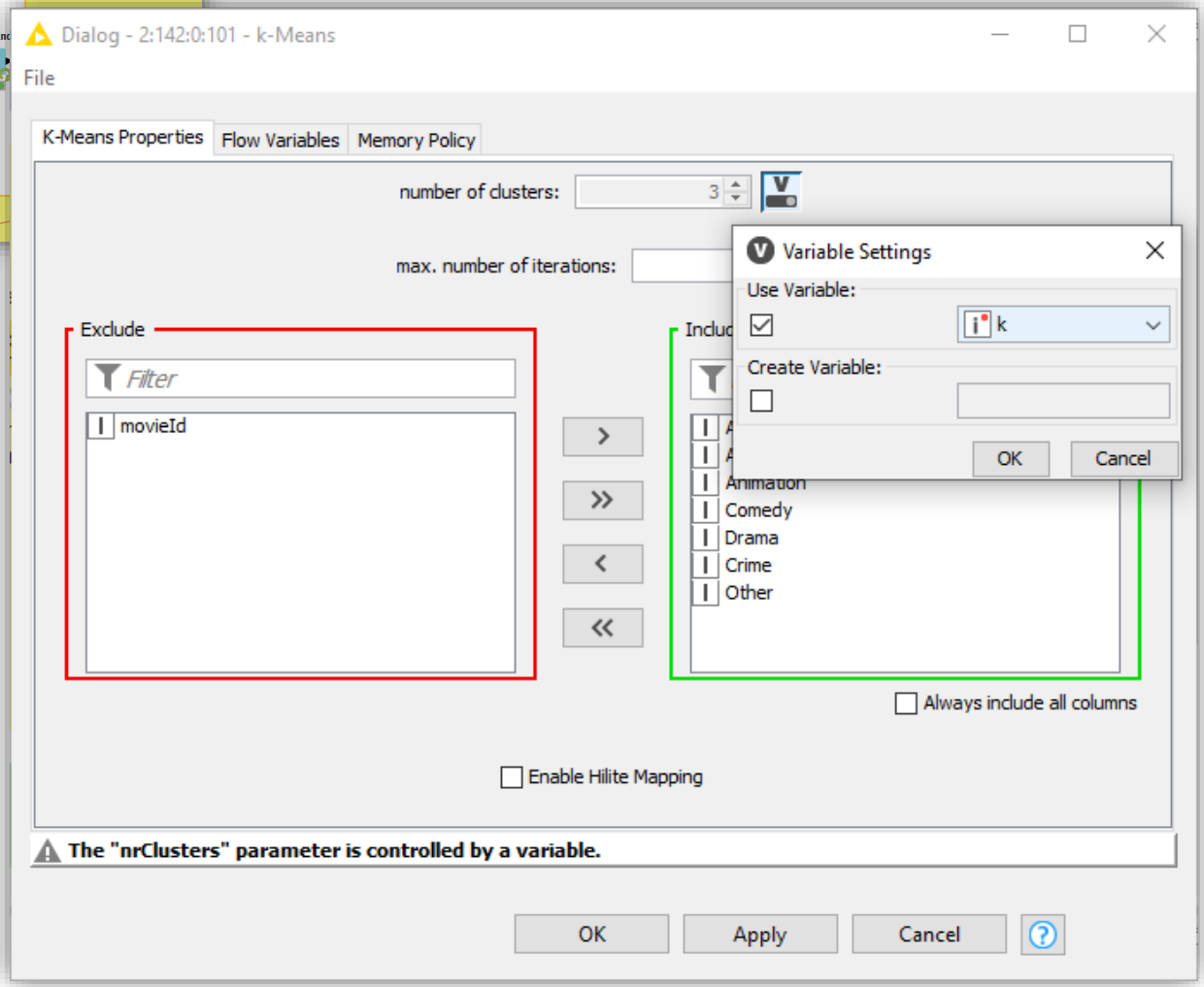
The Elbow Method



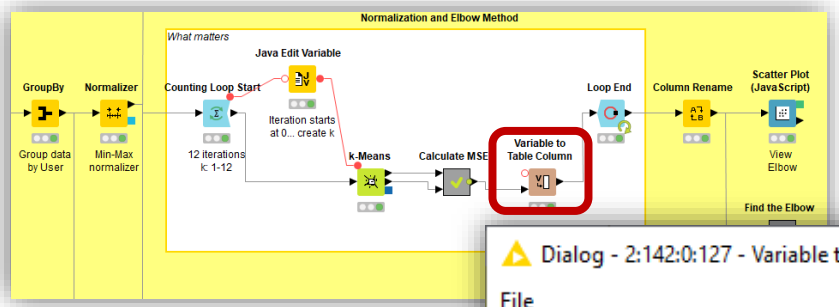


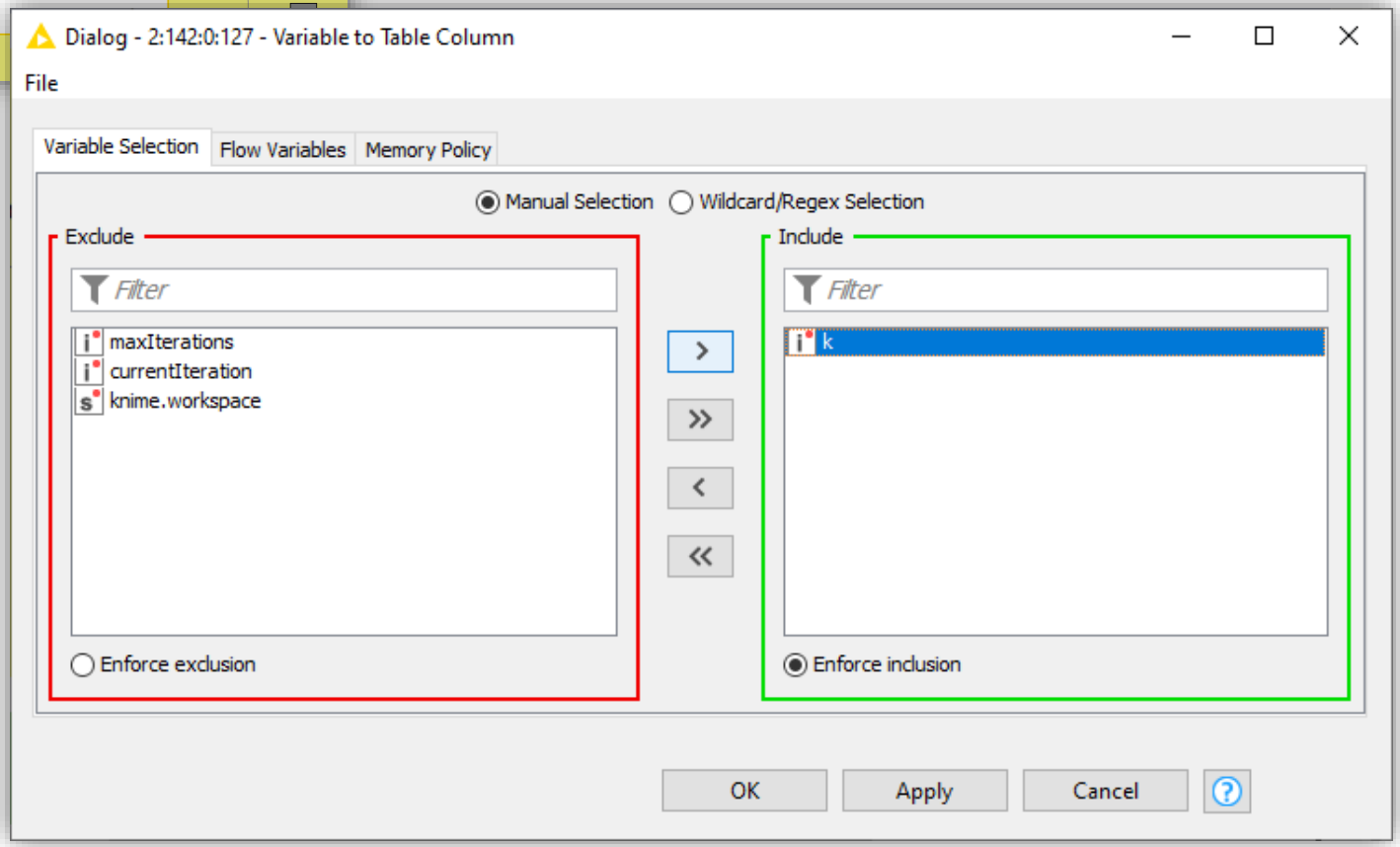
The Elbow Method



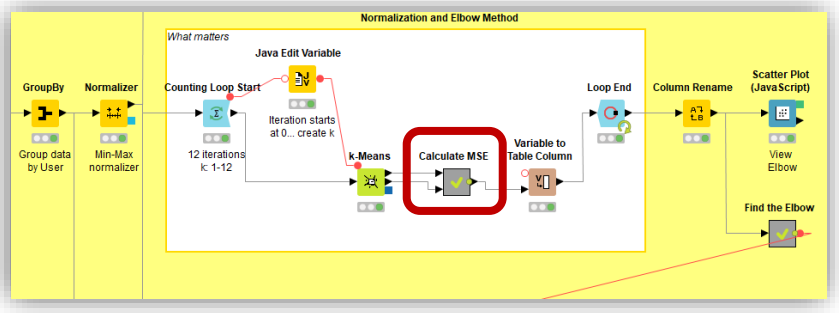


The Elbow Method





The Elbow Method



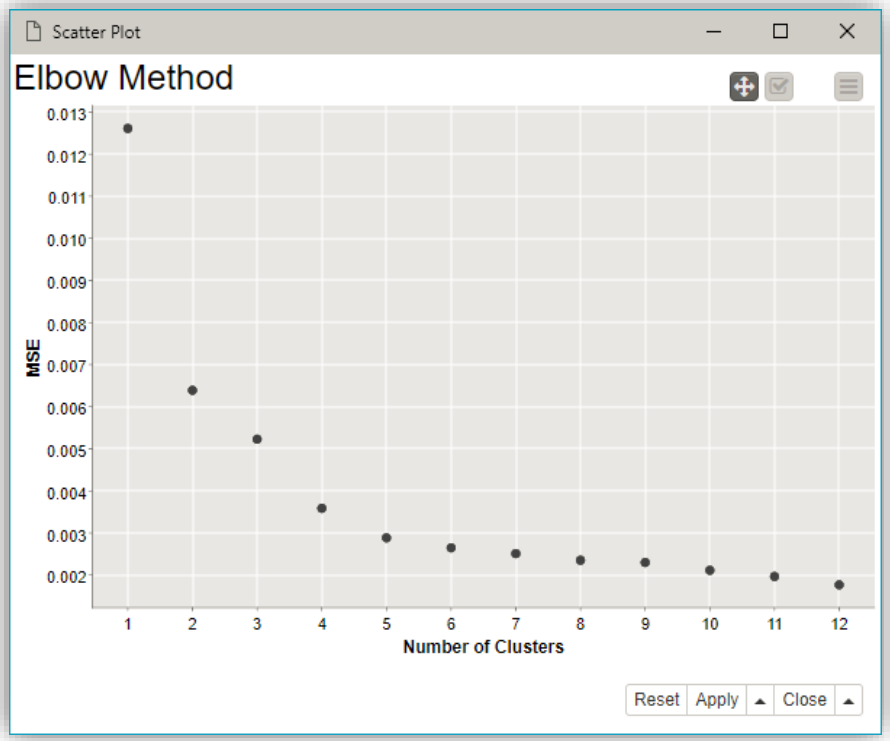
Collected results - ...

File Hilite Navigation View

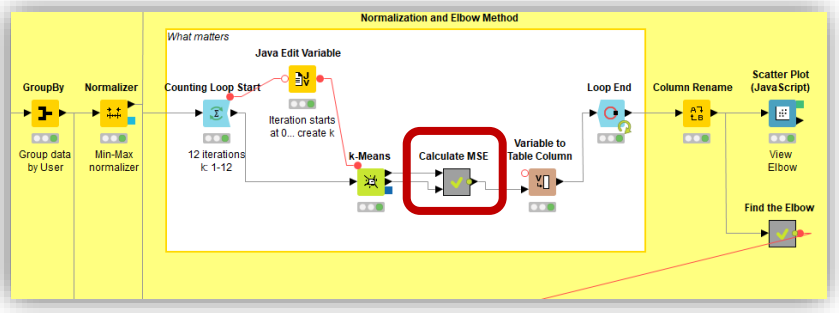
Properties | Flow Variables

Table "default" - Rows: 12 | Spec - Columns: 2

Row ID	D Mean(S...	I k
Row0#0	0.013	1
Row0#1	0.006	2
Row0#2	0.005	3
Row0#3	0.004	4
Row0#4	0.003	5
Row0#5	0.003	6
Row0#6	0.003	7
Row0#7	0.002	8
Row0#8	0.002	9
Row0#9	0.002	10
Row0#10	0.002	11
Row0#11	0.002	12



The Elbow Method



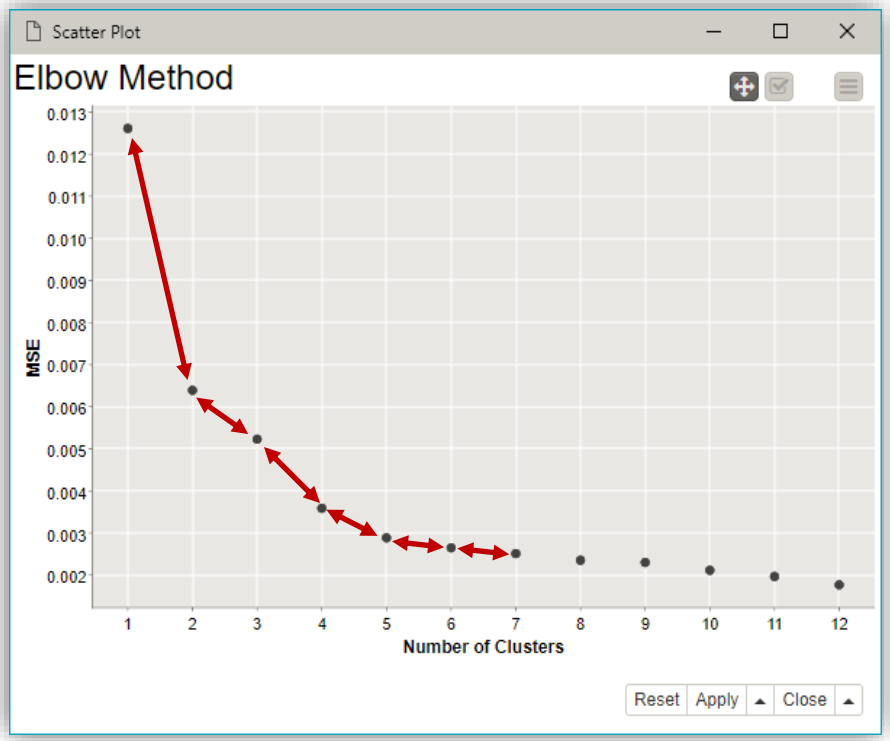
Collected results - ...

File Hilite Navigation View

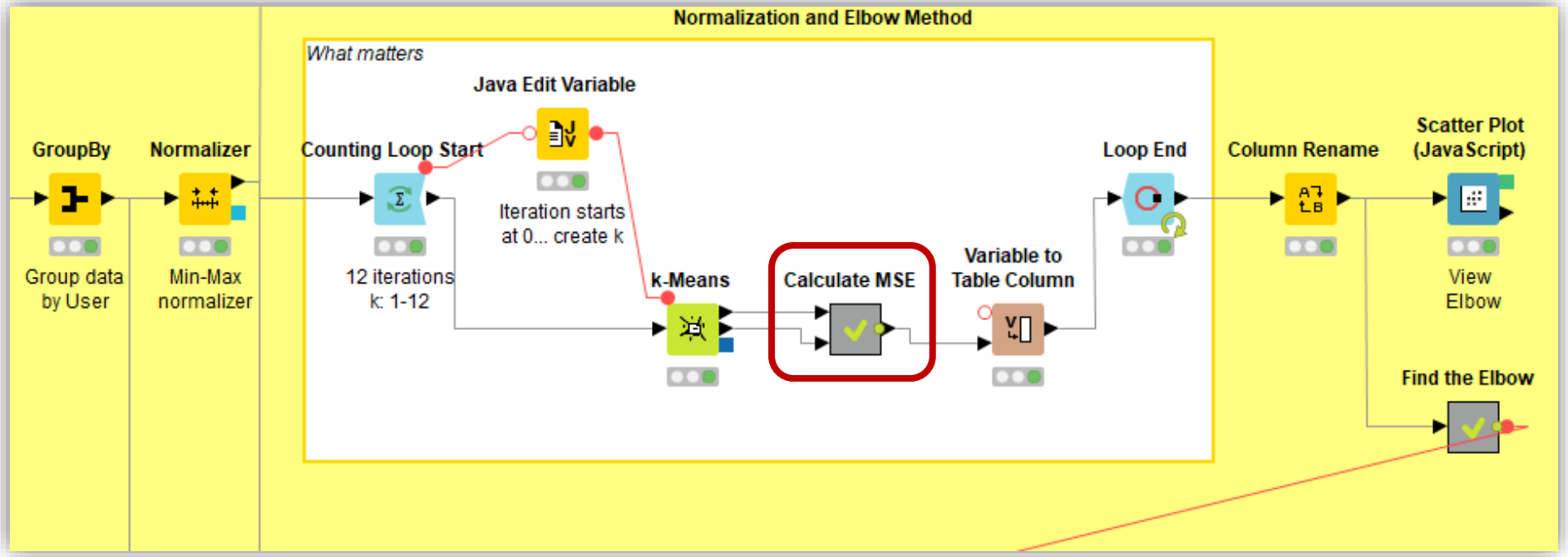
Properties | Flow Variables

Table "default" - Rows: 12 | Spec - Columns: 2

Row ID	D Mean(S...	I k
Row0#0	0.013	1
Row0#1	0.006	2
Row0#2	0.005	3
Row0#3	0.004	4
Row0#4	0.003	5
Row0#5	0.003	6
Row0#6	0.003	7
Row0#7	0.002	8
Row0#8	0.002	9
Row0#9	0.002	10
Row0#10	0.002	11
Row0#11	0.002	12



Quality Measures



Quality Measures
MAE, MSE and RMSE

MAE

Mean Absolute Error measures the average magnitude of the errors in a set of predictions, without considering their direction.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

MSE

Mean Squared Error consists of the average of squared differences between the prediction and the actual observation, without considering their direction

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

RMSE

Root Mean Squared Error consists of the square root of the average of squared differences between the prediction and the actual observation, without considering their direction

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Where n is the number of observations, and y_j and $\hat{y}_j$ are the actual observation and the predicted value, respectively.

Quality Measures
MAE, MSE and RMSE

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \qquad RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Important notes:

- Three of the most common metrics used to measure accuracy for continuous variables
- All express average model prediction error (lower values are better)
- All range from 0 to ∞ and are indifferent to the direction of errors
- MAE and RMSE express the prediction error in units of the variable of interest
- MSE and RMSE, by squaring the error, gives a relatively high weight to large errors
- Hence, MSE and RMSE are more useful when large errors are particularly undesirable

Quality Measures
MAE, MSE and RMSE

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \qquad RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

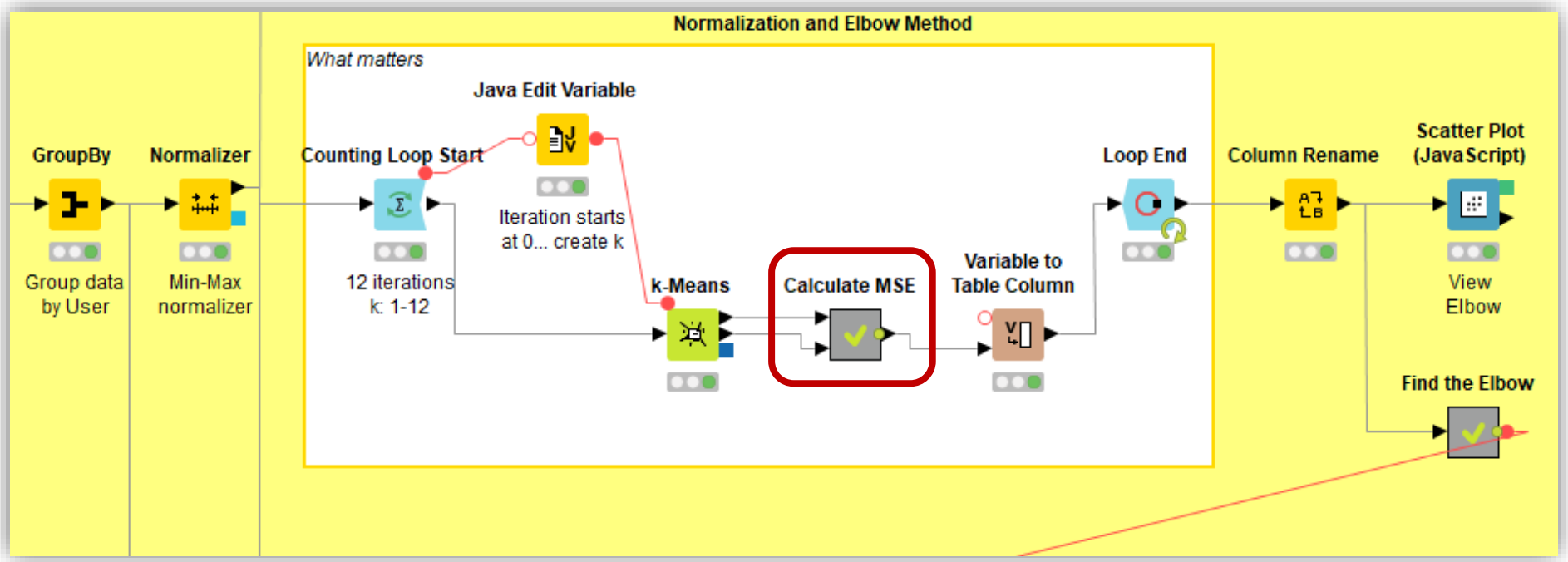
#	Error	|Error|	Error²
1	1	1	1
2	1	1	1
3	3	3	9
4	3	3	9

MAE	MSE	RMSE
2	5	2.24

#	Error	|Error|	Error²
1	0	0	0
2	0	0	0
3	0	0	0
4	10	10	100

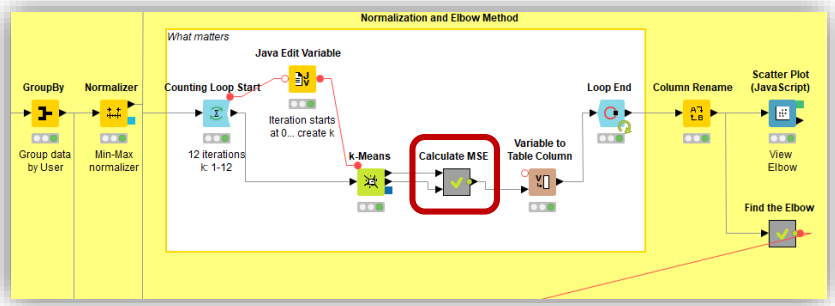
MAE	MSE	RMSE
2.5	25	5

Quality Measures for Clustering
The Elbow Method



Distance from each point to the centroid of the cluster it belongs to

Quality Measures for Clustering
The Elbow Method



The two outputs of the k-Means node:

(1) Input data labeled with the cluster;

(2) The created clusters and centroids.

Clusters - 0:144:0:141 - k-Means (2)

File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Animation	Comedy	Crime	Drama	Other
cluster_0	0.747	0.767	0.76	0.672	0.786
cluster_1	0.017	0.015	0.021	0.016	0.014
cluster_2	0.026	0.02	0.024	0.018	0.022
cluster_3	0.113	0.107	0.127	0.098	0.119
cluster_4	0.153	0.085	0.073	0.058	0.085
cluster_5	0.397	0.386	0.392	0.315	0.408
cluster_6	0.477	0.257	0.154	0.14	0.229
cluster_7	0.02	0.019	0.021	0.016	0.017
cluster_8	0.025	0.023	0.025	0.021	0.021
cluster_9	0.167	0.186	0.237	0.17	0.223
cluster_10	0.049	0.051	0.066	0.055	0.059
cluster_11	0.026	0.014	0.03	0.021	0.019

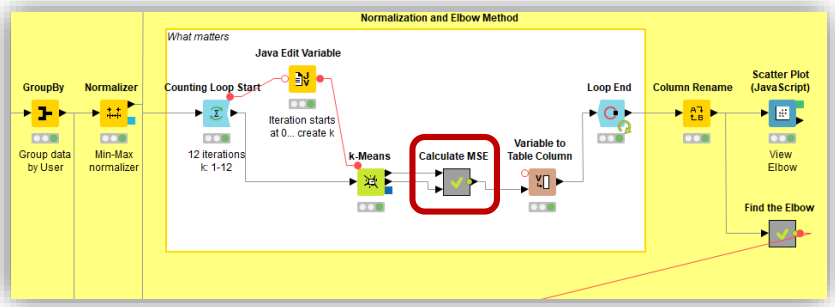
Labeled input - 0:144:0:141 - k-Means (1)

File Hilite Navigation View

Table "default" - Rows: 610 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Comedy	Crime	Drama	Other	Cluster
Row0	0.077	0.107	0.051	0.101	cluster_3
Row1	0.006	0.024	0.012	0.005	cluster_1
Row2	0.008	0.005	0.011	0.012	cluster_2
Row3	0.096	0.064	0.091	0.069	cluster_10
Row4	0.014	0.029	0.018	0.013	cluster_7
Row5	0.118	0.084	0.106	0.116	cluster_4
Row6	0.045	0.062	0.043	0.067	cluster_4
Row7	0.022	0.021	0.014	0.016	cluster_7
Row8	0.014	0.017	0.015	0.01	cluster_8
Row9	0.073	0.031	0.054	0.058	cluster_8
Row10	0.011	0.031	0.018	0.023	cluster_1
Row11	0.017	0.002	0.011	0.007	cluster_11
Row12	0.01	0.007	0.011	0.007	cluster_7
Row13	0.016	0.017	0.018	0.014	cluster_7
Row14	0.026	0.045	0.041	0.06	cluster_4
Row15	0.023	0.057	0.036	0.039	cluster_10
Row16	0.018	0.076	0.038	0.037	cluster_11
Row17	0.137	0.329	0.166	0.202	cluster_9
Row18	0.35	0.172	0.115	0.308	cluster_6
Row19	0.097	0.06	0.054	0.107	cluster_4
Row20	0.186	0.174	0.071	0.189	cluster_6

Quality Measures for Clustering
The Elbow Method



Observation with id Row0 was assigned to cluster 3. Its value for comedy is 0.077. How far is it from the centroid's center of cluster 3 (0.107)? And for the other genres?

How far is this observation from the centroid of the cluster?

Clusters - 0:144:0:141 - k-Means

File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	Animation	Comedy	Crime	Drama	Other
cluster_0	0.747	0.767	0.76	0.672	0.786
cluster_1	0.017	0.015	0.021	0.016	0.014
cluster_2	0.026	0.02	0.024	0.018	0.022
cluster_3	0.113	0.107	0.127	0.098	0.119
cluster_4	0.153	0.085	0.073	0.058	0.085
cluster_5	0.397	0.386	0.392	0.315	0.408
cluster_6	0.477	0.257	0.154	0.14	0.229
cluster_7	0.02	0.019	0.021	0.016	0.017
cluster_8	0.025	0.023	0.025	0.021	0.021
cluster_9	0.167	0.186	0.237	0.17	0.223
cluster_10	0.049	0.051	0.066	0.055	0.059
cluster_11	0.026	0.014	0.03	0.021	0.019

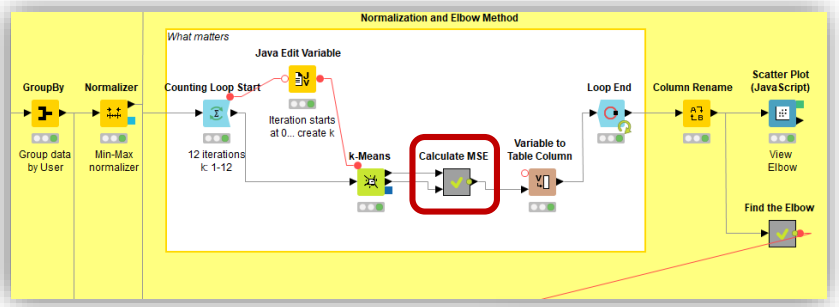
Labeled input - 0:144:0:141 - k-Means

File Hilite Navigation View

Table "default" - Rows: 610 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Comedy	Crime	Drama	Other	Cluster
Row0	0.077	0.107	0.051	0.101	cluster_3
Row1	0.006	0.024	0.012	0.005	cluster_1
Row2	0.008	0.005	0.011	0.012	cluster_2
Row3	0.096	0.064	0.091	0.069	cluster_10
Row4	0.014	0.029	0.018	0.013	cluster_7
Row5	0.118	0.084	0.106	0.116	cluster_4
Row6	0.045	0.062	0.043	0.067	cluster_4
Row7	0.022	0.021	0.014	0.016	cluster_7
Row8	0.014	0.017	0.015	0.01	cluster_8
Row9	0.073	0.031	0.054	0.058	cluster_8
Row10	0.011	0.031	0.018	0.023	cluster_1
Row11	0.017	0.002	0.011	0.007	cluster_11
Row12	0.01	0.007	0.011	0.007	cluster_7
Row13	0.016	0.017	0.018	0.014	cluster_7
Row14	0.026	0.045	0.041	0.06	cluster_4
Row15	0.023	0.057	0.036	0.039	cluster_10
Row16	0.018	0.076	0.038	0.037	cluster_11
Row17	0.137	0.329	0.166	0.202	cluster_9
Row18	0.35	0.172	0.115	0.308	cluster_6
Row19	0.097	0.06	0.054	0.107	cluster_4
Row20	0.186	0.174	0.071	0.189	cluster_6

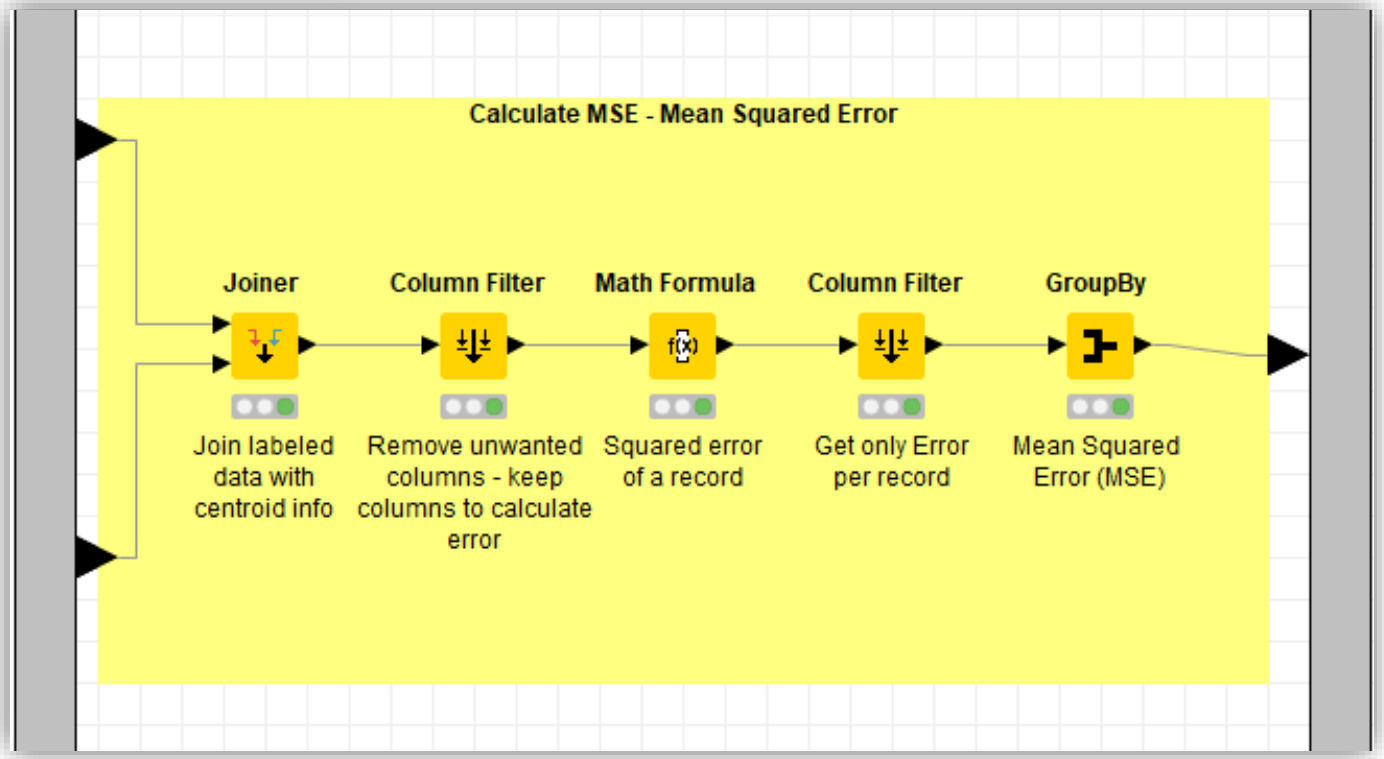
Quality Measures for Clustering
The Elbow Method



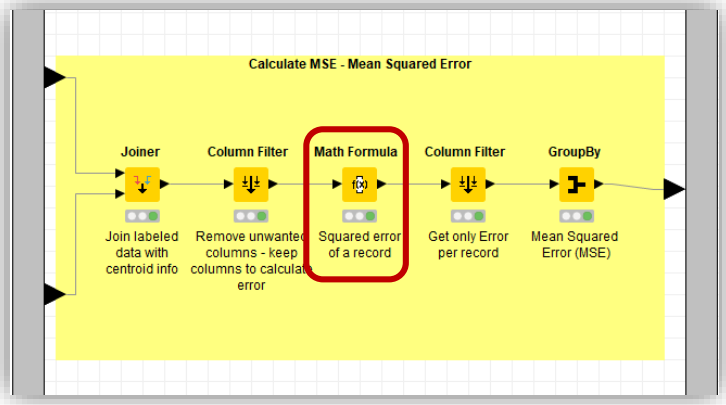
We may use MSE, MAE or RMSE to compute this error metric, i.e., how far are records from the centroid's of their cluster.

Input: The input data labeled with the cluster they belong.

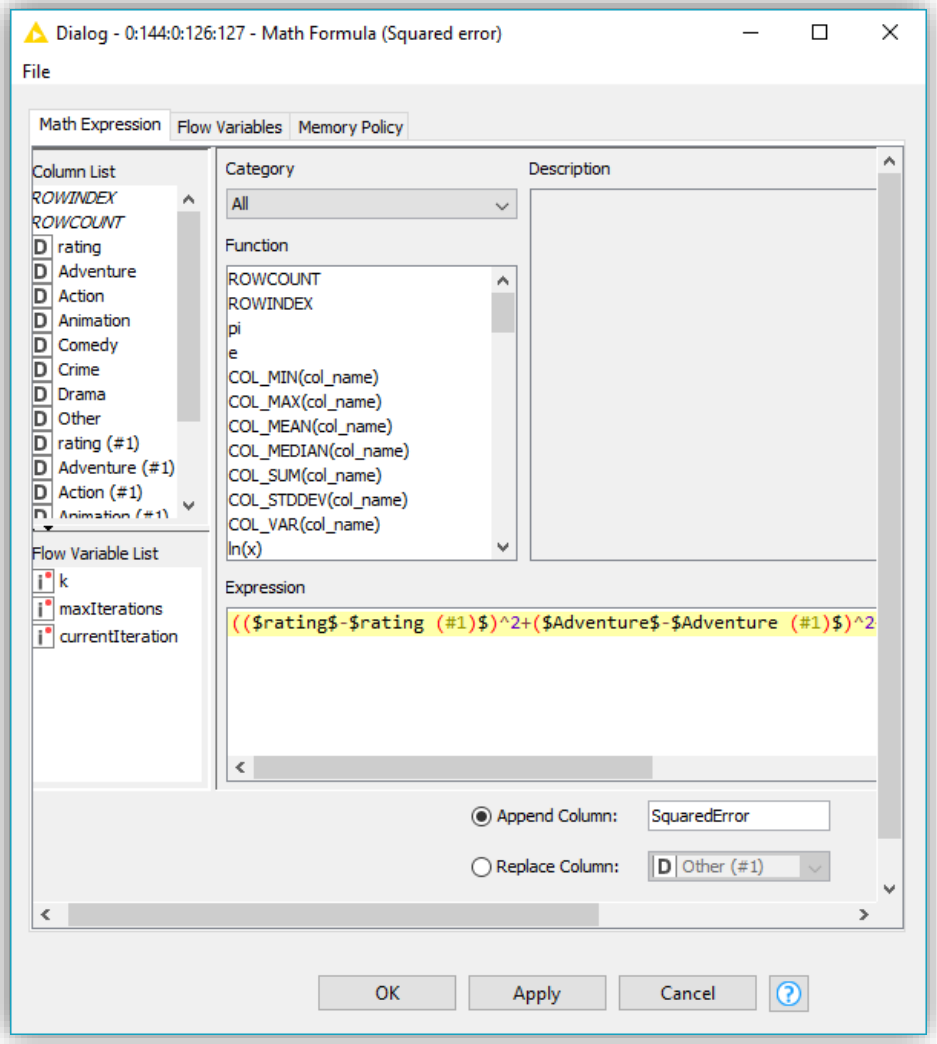
Input: The created clusters and centroids.



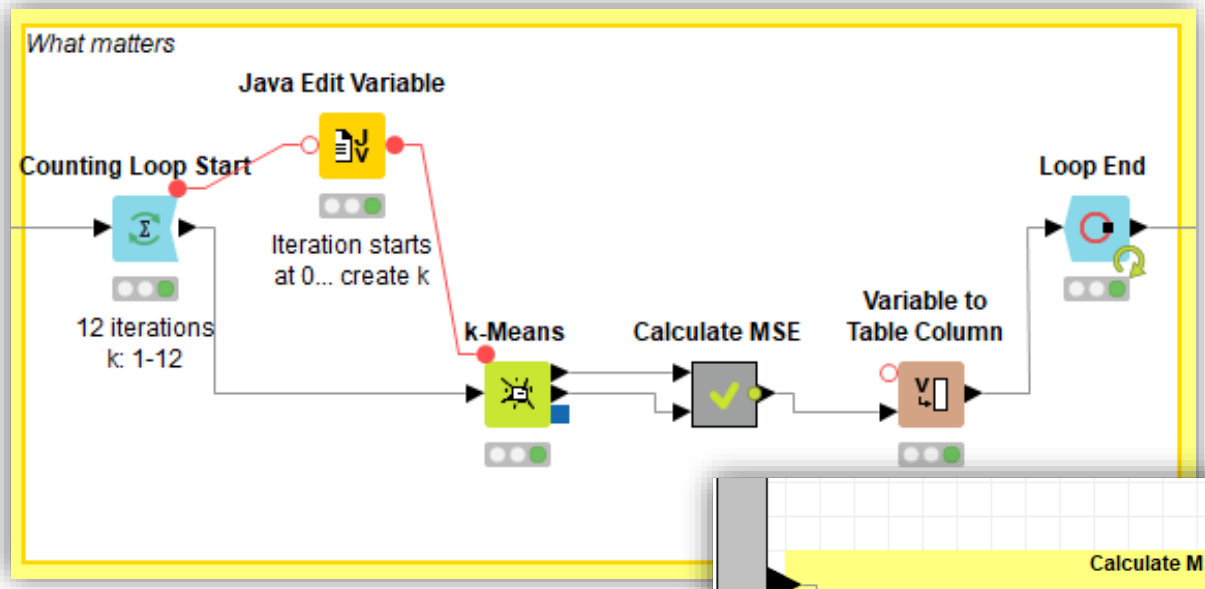
Quality Measures for Clustering
The Elbow Method

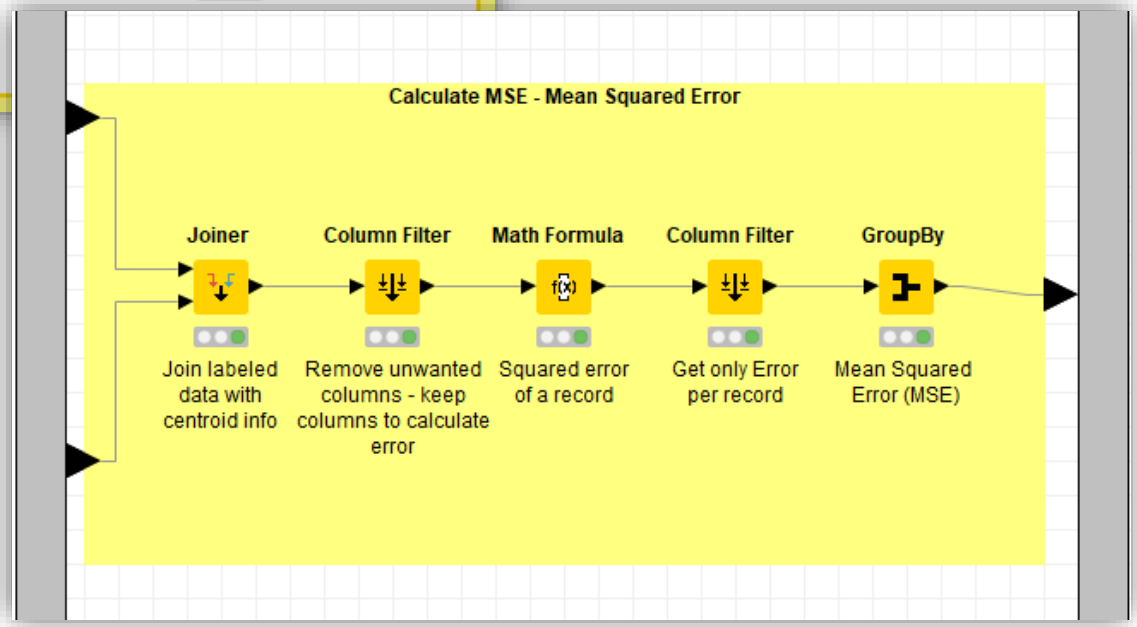


$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2$$

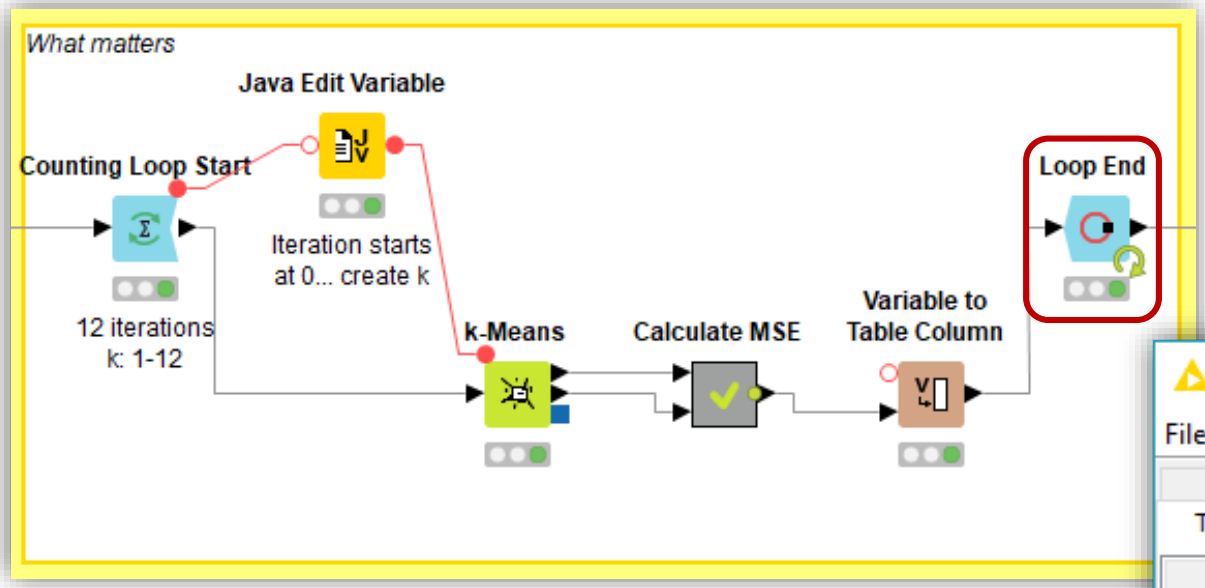


Quality Measures for Clustering
The Elbow Method Metanode





Quality Measures for Clustering
The Elbow Method Metanode



Collected results - ...

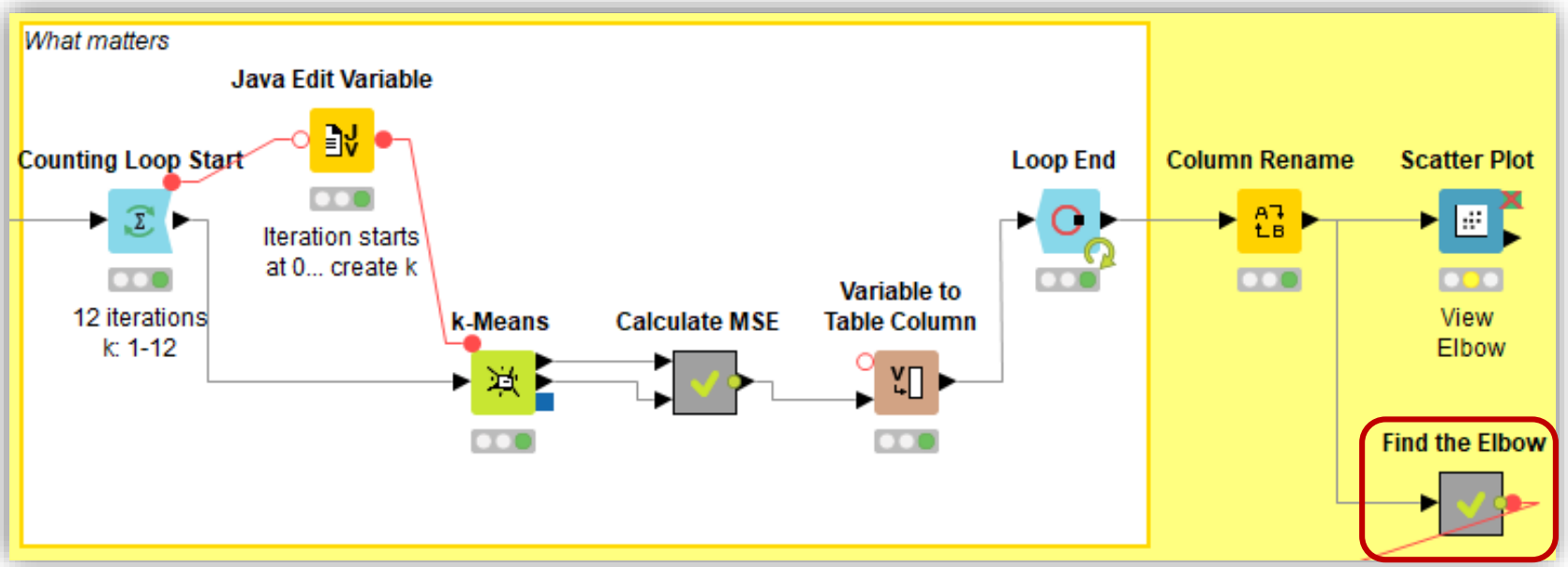
File Hilite Navigation View

Properties | Flow Variables

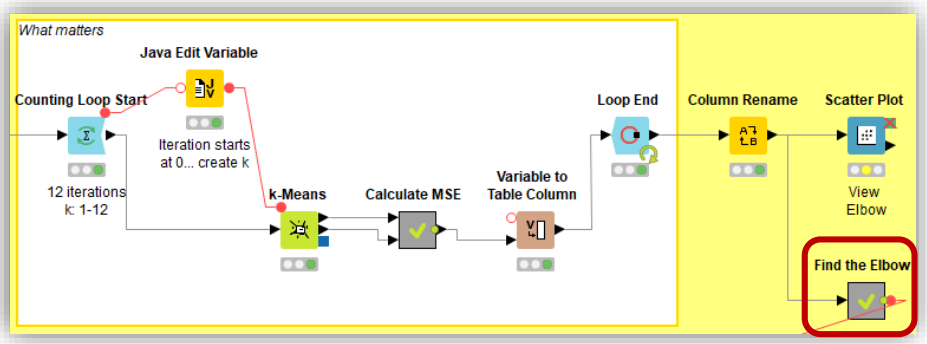
Table "default" - Rows: 12 | Spec - Columns: 2

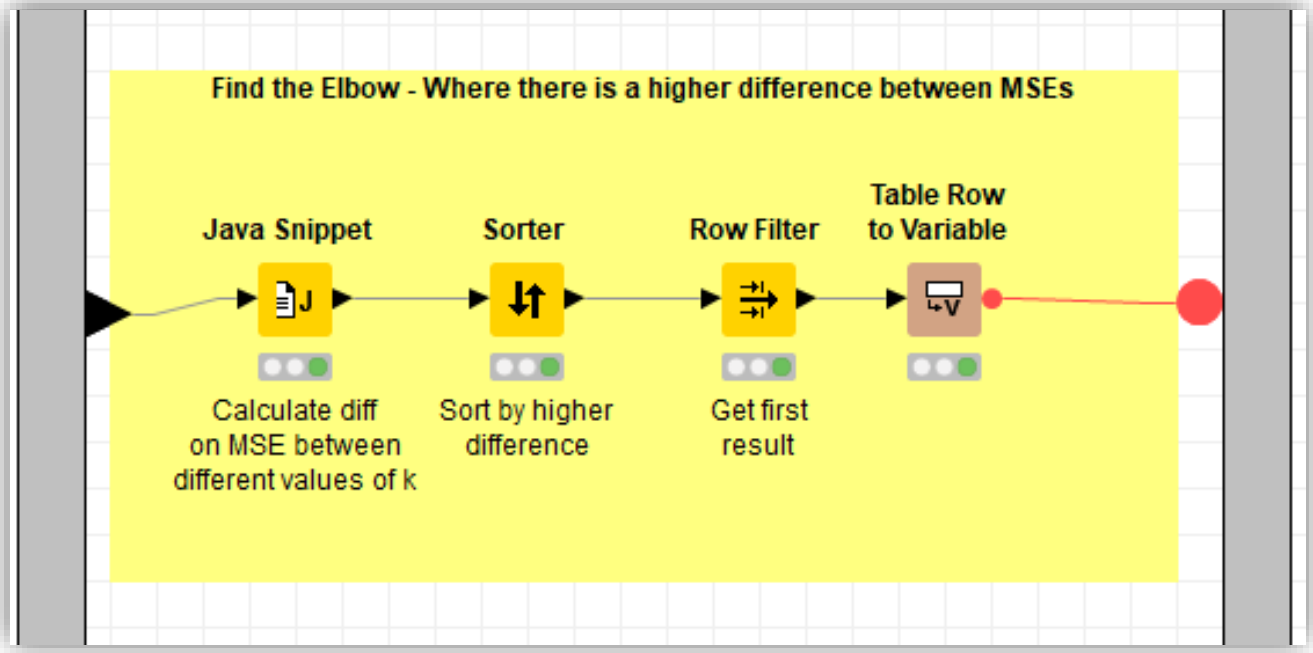
Row ID	Mean(S...	k
Row0#0	0.013	1
Row0#1	0.006	2
Row0#2	0.005	3
Row0#3	0.004	4
Row0#4	0.003	5
Row0#5	0.003	6
Row0#6	0.003	7
Row0#7	0.002	8
Row0#8	0.002	9
Row0#9	0.002	10
Row0#10	0.002	11
Row0#11	0.002	12

Quality Measures for Clustering
Finding the Elbow … Automatically

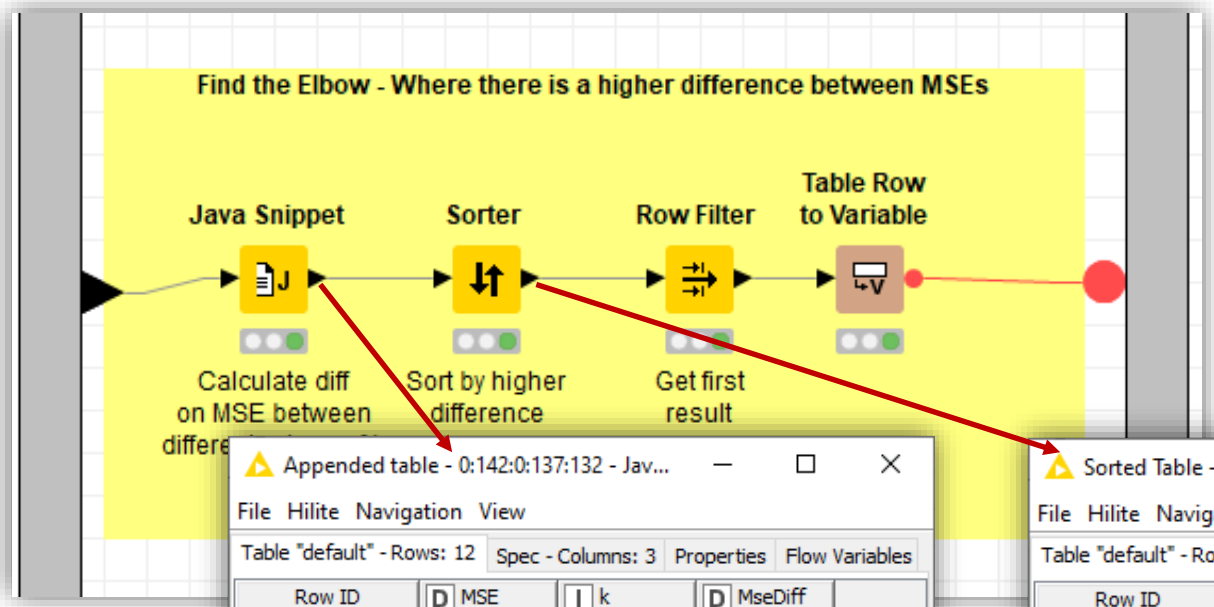


Quality Measures for Clustering
Finding the Elbow … Automatically





Quality Measures for Clustering
Finding the Elbow … Automatically



Appended table - 0:142:0:137:132 - Jav...

File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	MSE	k	MseDiff
Row0#0	0.163	1	0
Row0#1	0.124	2	0.039
Row0#2	0.106	3	0.019
Row0#3	0.093	4	0.012
Row0#4	0.085	5	0.009
Row0#5	0.073	6	0.011
Row0#6	0.069	7	0.004
Row0#7	0.064	8	0.005
Row0#8	0.061	9	0.003
Row0#9	0.059	10	0.002
Row0#10	0.053	11	0.006
Row0#11	0.053	12	0

Sorted Table - 0:142:0:137:133 - Sorter (...

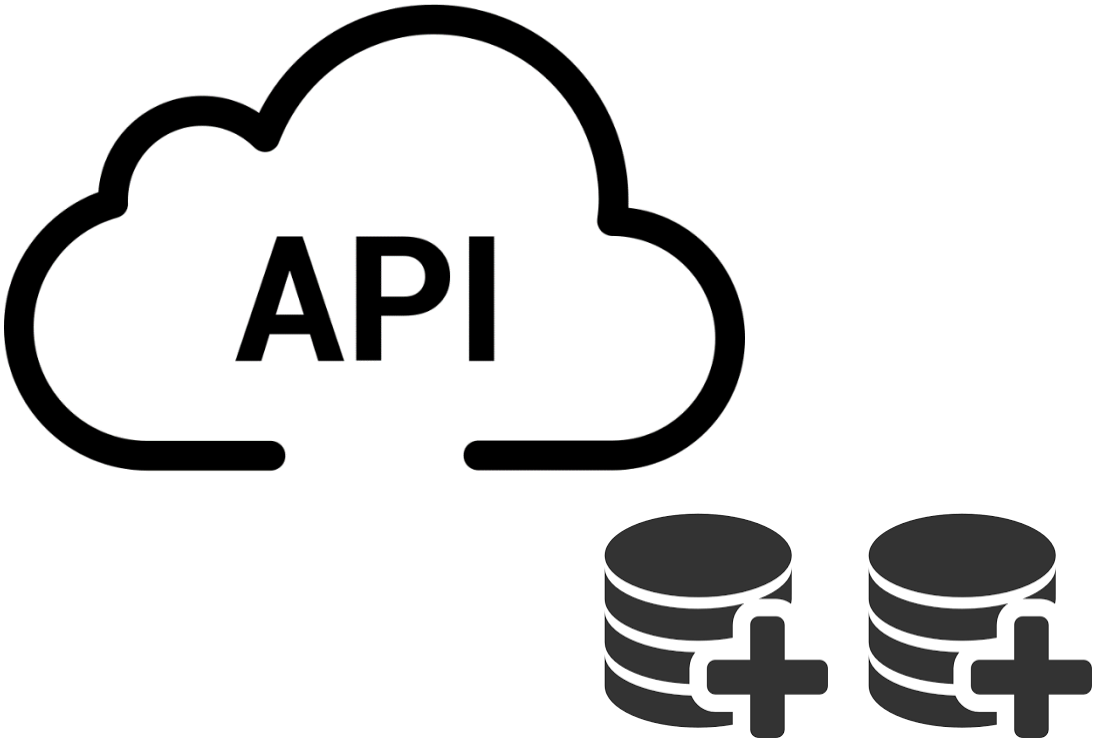
File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 3 | Properties | Flow Variables

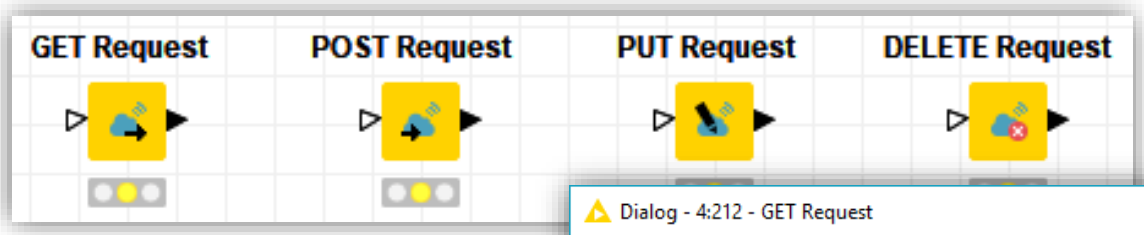
Row ID	MSE	k	MseDiff
Row0#1	0.124	2	0.039
Row0#2	0.106	3	0.019
Row0#3	0.093	4	0.012
Row0#5	0.073	6	0.011
Row0#4	0.085	5	0.009
Row0#10	0.053	11	0.006
Row0#7	0.064	8	0.005
Row0#6	0.069	7	0.004
Row0#8	0.061	9	0.003
Row0#9	0.059	10	0.002
Row0#0	0.163	1	0
Row0#11	0.053	12	0

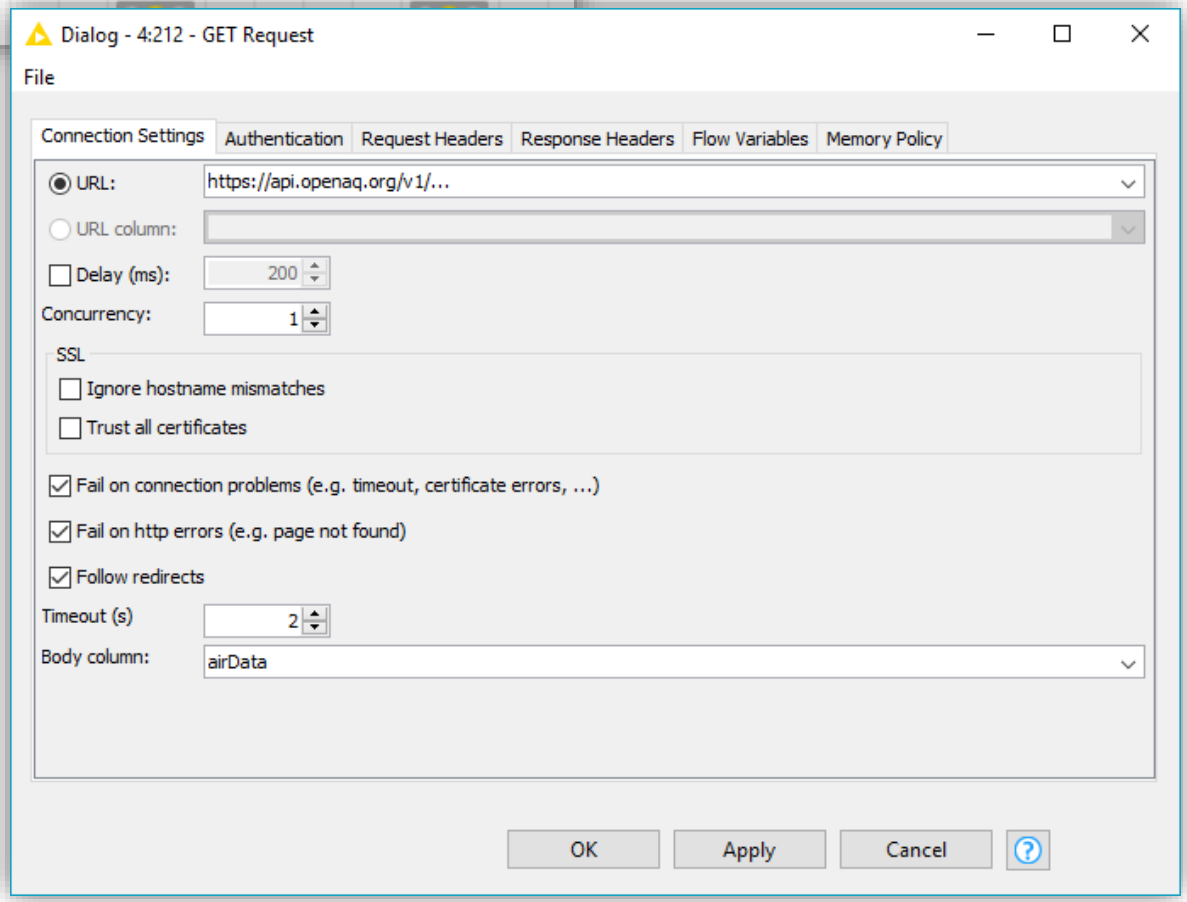
HTTP Requests

API Calls



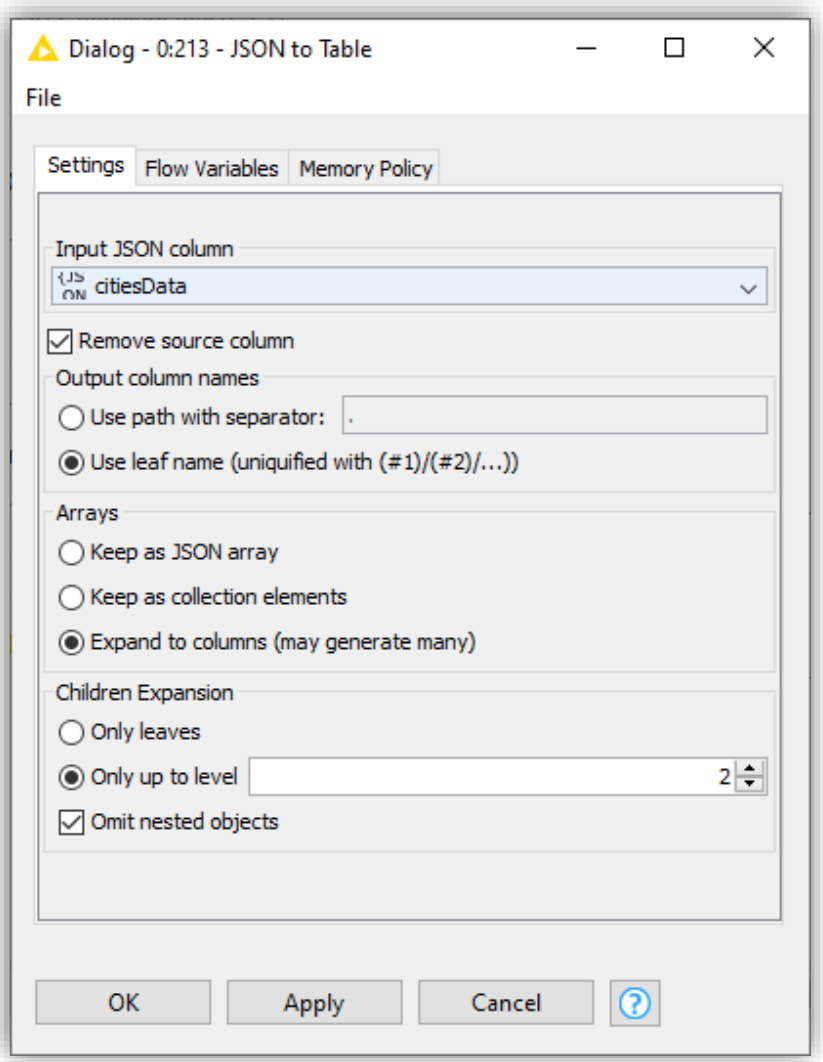
HTTP Requests
API Calls



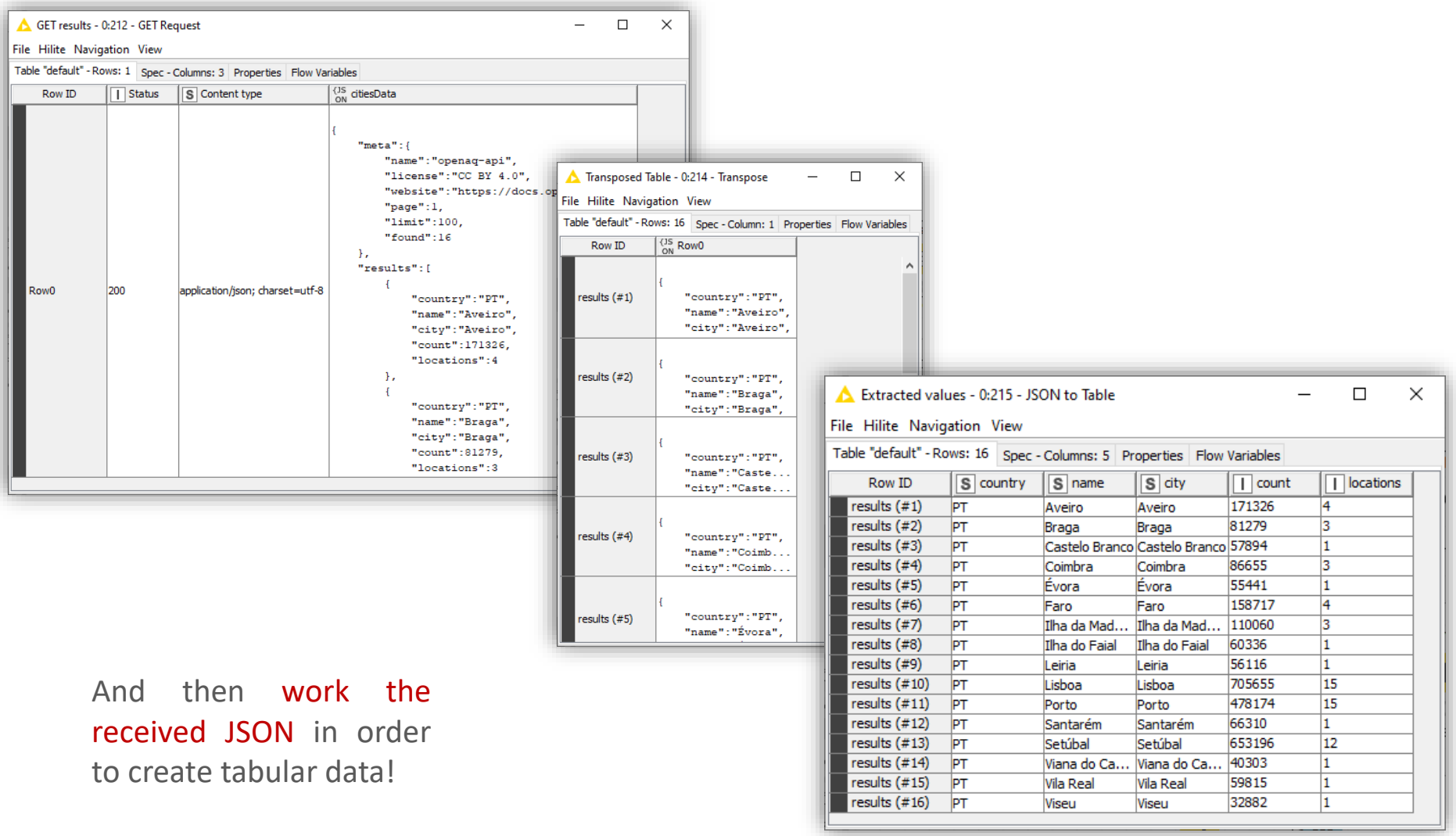


HTTP Requests
JSON to Table

And then work the received JSON in order to create tabular data!

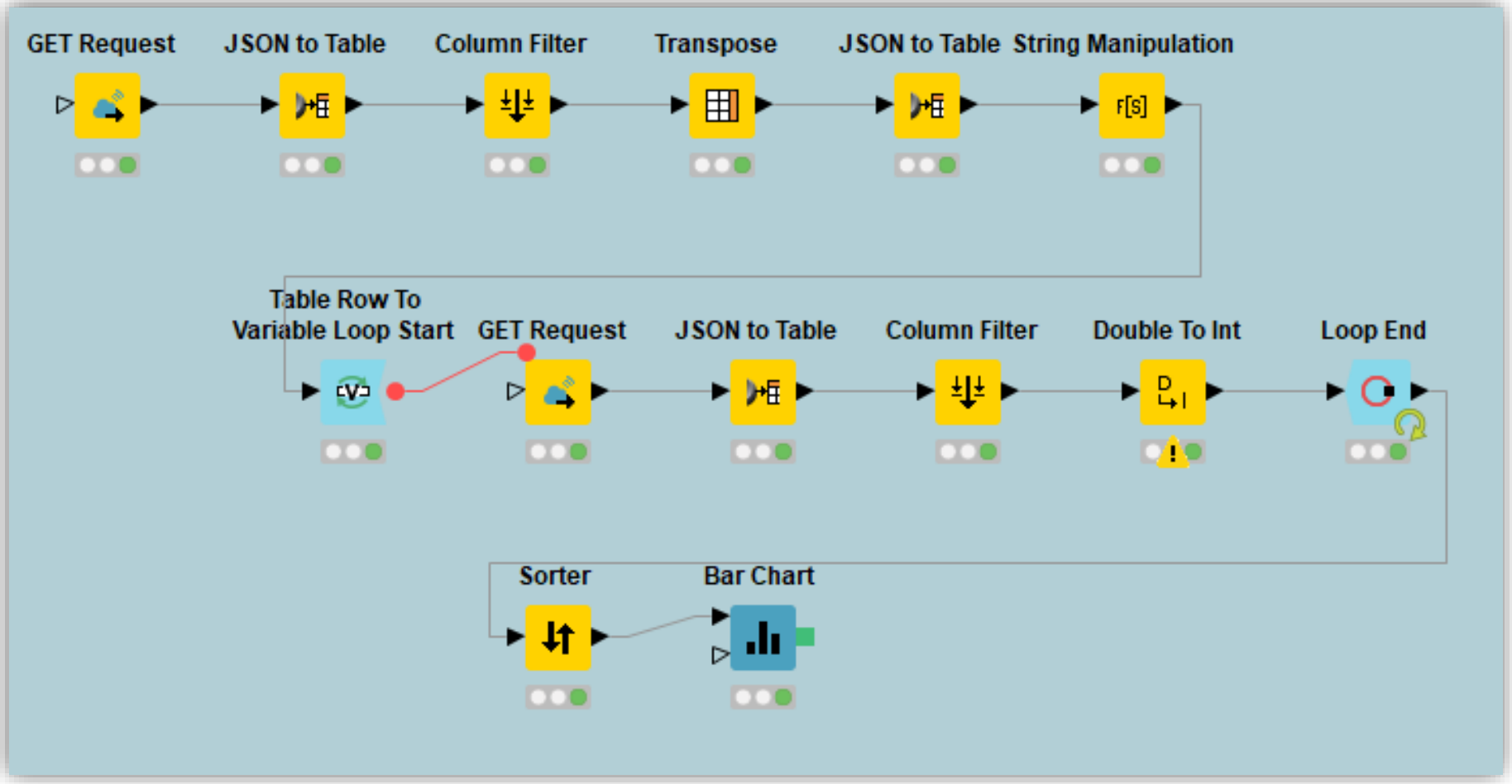


HTTP Requests
Handle the JSON

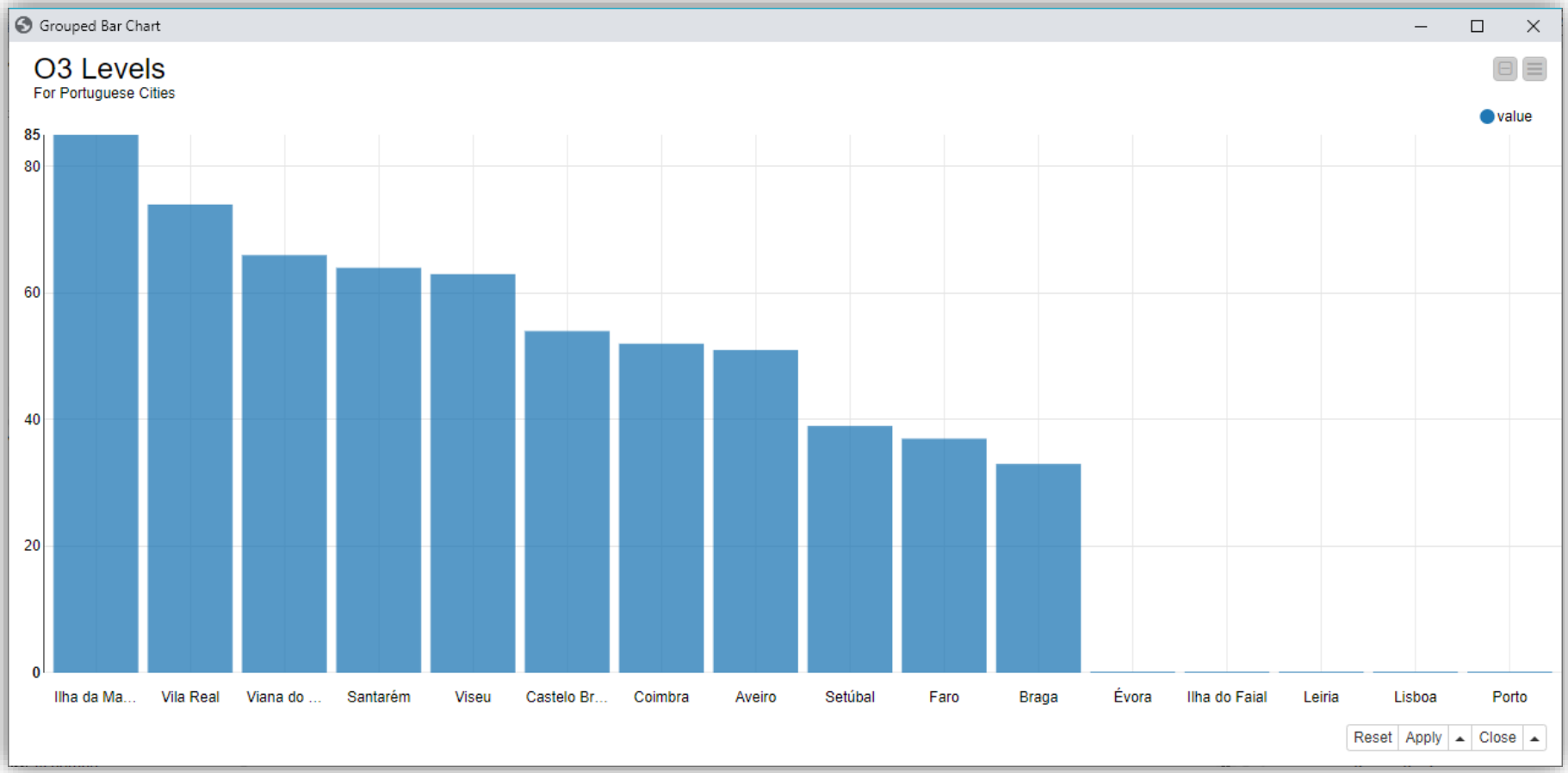


And then work the received JSON in order to create tabular data!

HTTP Requests
A Workflow



HTTP Requests
A Workflow



HTTP Requests
API Calls

OpenWeatherMap

https://openweathermap.org/api

openaq

https://docs.openaq.org/

TOMTOM FOR DEVELOPERS

https://developer.tomtom.com/

https://www.openuv.io/uvindex

whitepages PRO

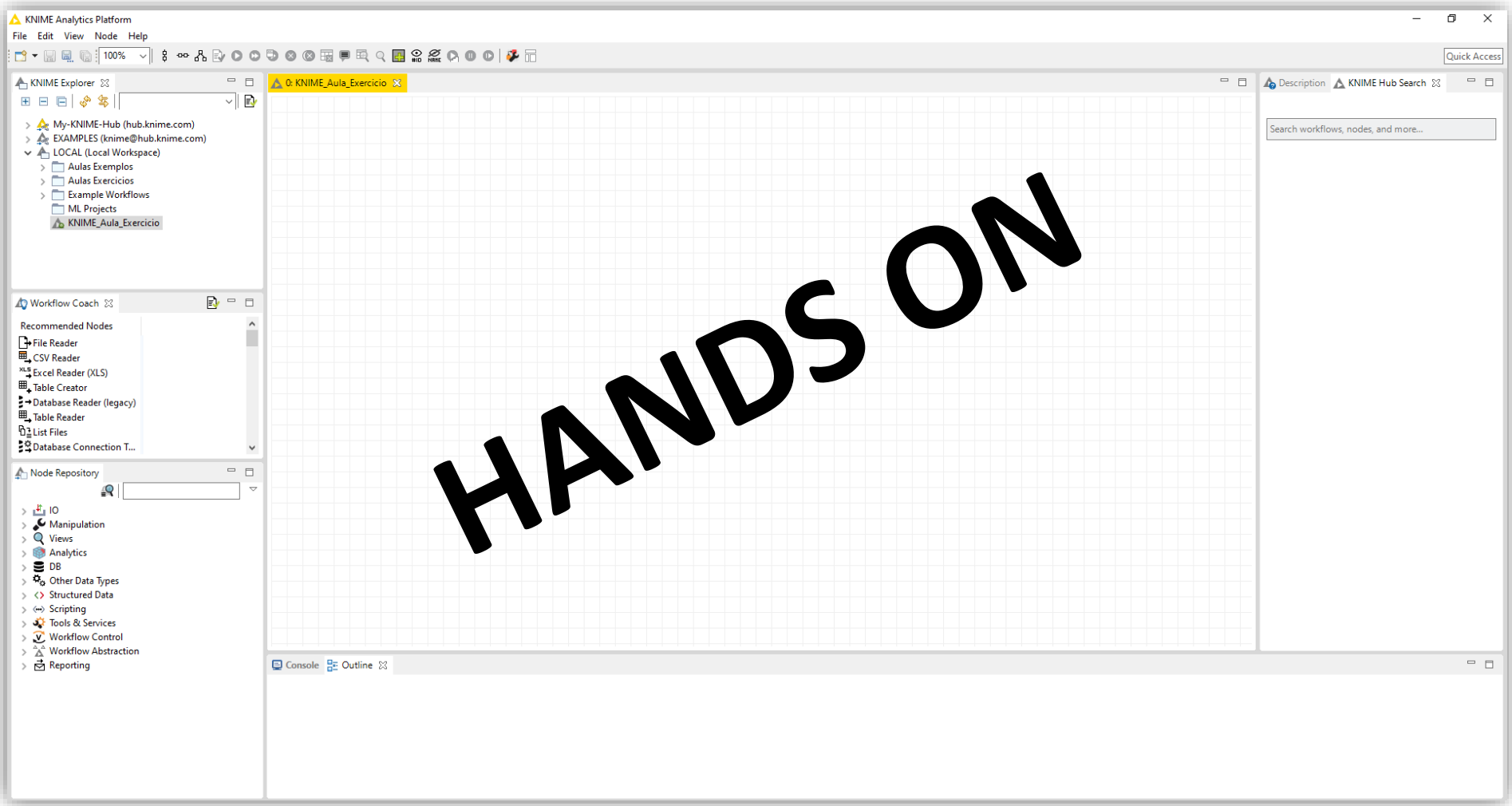
https://pro.whitepages.com/apis/

https://developers.google.com/

coinbase

https://developers.coinbase.com/

Hands On


KNIME Analytics Platform
File Edit View Node Help
KNIME Explorer
My-KNIME-Hub (hub.knime.com)
EXAMPLES (knime@hub.knime.com)
LOCAL (Local Workspace)
Aulas Exemplos
Aulas Exercicios
Example Workflows
ML Projects
KNIME_Aula_Exercicio
Workflow Coach
Recommended Nodes
File Reader
CSV Reader
Excel Reader (XLS)
Table Creator
Database Reader (legacy)
Table Reader
List Files
Database Connection T...
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Structured Data
Scripting
Tools & Services
Workflow Control
Workflow Abstraction
Reporting
KNIME_Aula_Exercicio
HANDS ON
Description
KNIME Hub Search
Search workflows, nodes, and more...
Console
Outline