



University of Minho  
School of Engineering



# 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

2

The Elbow Method

Quality Measures

HTTP Requests

Hands On

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

# Clustering and Recommender Systems

3

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# Clustering and Recommender Systems

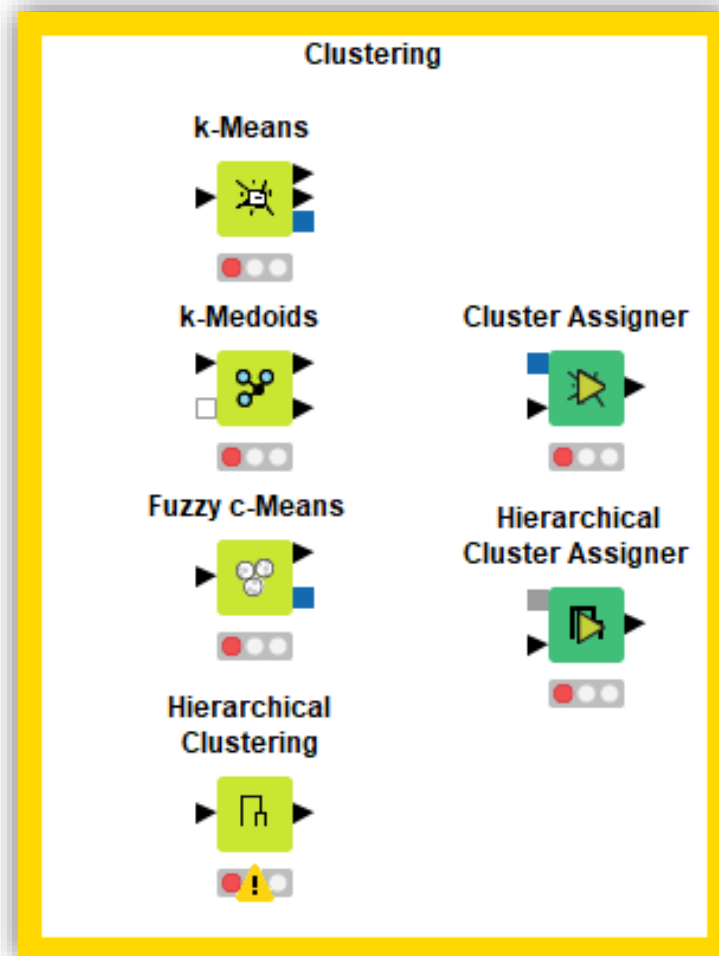
4

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# Is it clustered?

## Principal Components Analysis

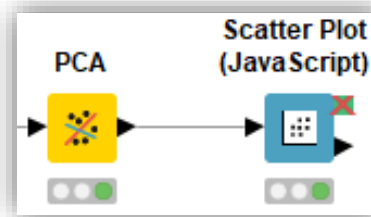
5

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



Dialog - 0:142:0:107 - PCA

File

Options Flow Variables Memory Policy

☐ Fail if missing values are encountered (skipped per default)

Target dimensions

☒ Dimensions to reduce to

☐ Minimum information fraction to preserve (%)

☐ Replace original data columns

Exclude

Filter

movieId

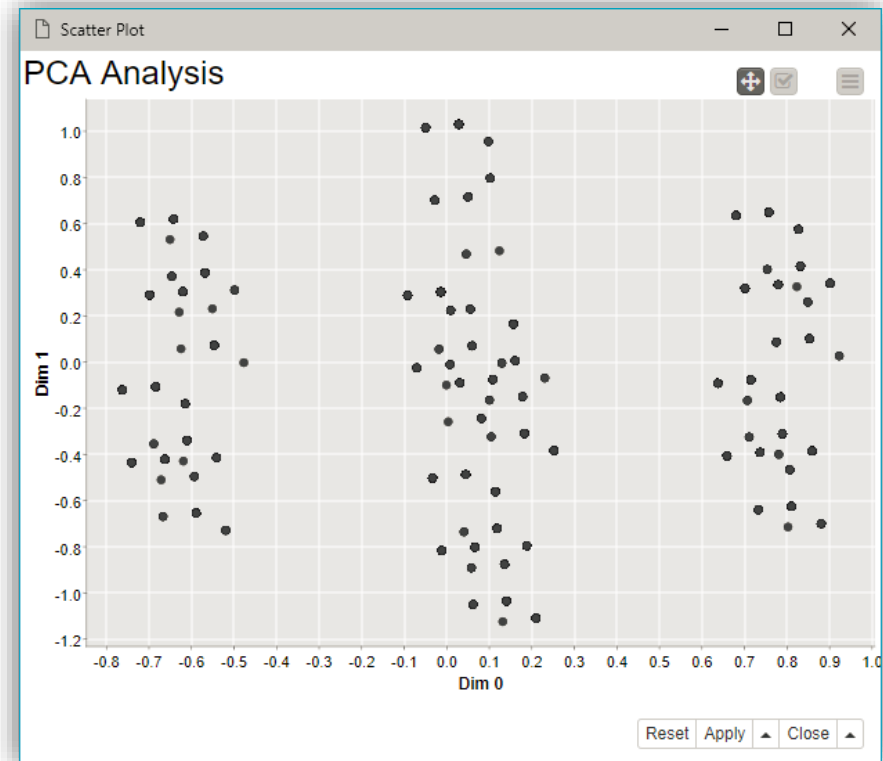
> >> < <<

Include

Filter

Adventure  
Action  
Animation  
Comedy  
Drama  
Crime  
Other

OK Apply Cancel ?



# Is it clustered?

## Scatter Matrix

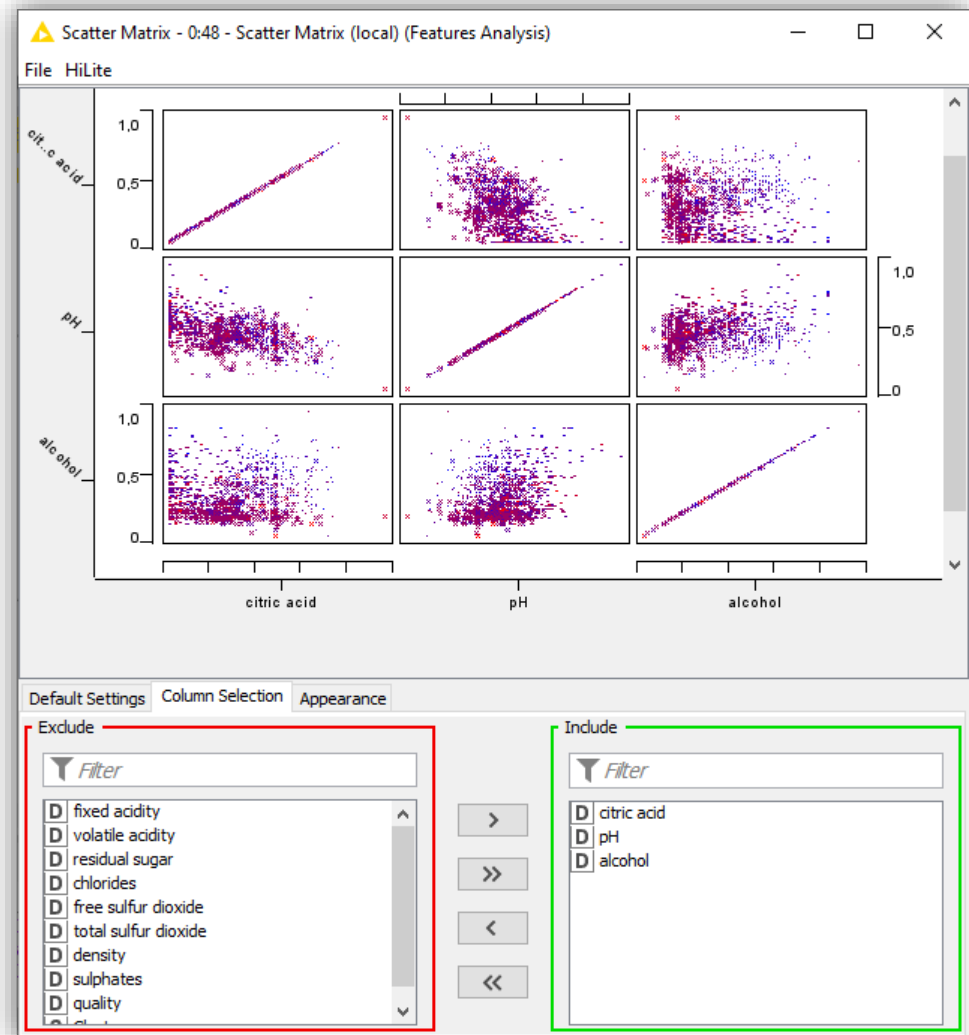
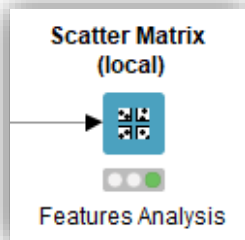
6

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# K-Means Settings

7

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On

Dialog - 0:21 - k-Means

File

K-Means Properties Flow Variables Memory Policy

number of clusters: 3

max. number of iterations: 99

Exclude

Filter

Include

Filter

> >> < <<

☐ Always include all columns

☐ Enable Hilite Mapping

OK Apply Cancel ?

Node Description

### k-Means

This node outputs the cluster centers for a predefined number of clusters (no dynamic number of clusters). K-means performs a crisp clustering that assigns a data vector to exactly one cluster. The algorithm terminates when the cluster assignments do not change anymore.

The clustering algorithm uses the Euclidean distance on the selected attributes. The data is not normalized by the node (if required, you should consider to use the "Normalizer" as a preprocessing step).

#### Dialog Options

**number of clusters**

The number of clusters (cluster centers) to be created.

**max number of iterations**

The number of iterations after which the algorithm terminates, independent of the accuracy improvement of the cluster centers.

**Enable Hilite Mapping**

If enabled, the hiliting of a cluster row (2nd output) will hilite all rows of this cluster in the input table and the 1st output table. Depending on the number of rows, enabling this feature might consume a lot of memory.

#### Ports

**Input Ports**

0 Input to clustering. All numerical values and only these are considered for clustering.

# But ... How many clusters?

8

**THE ELBOW METHOD**

Quality Measures

HTTP Requests

Hands On



**The Elbow Method**



# The Elbow Method

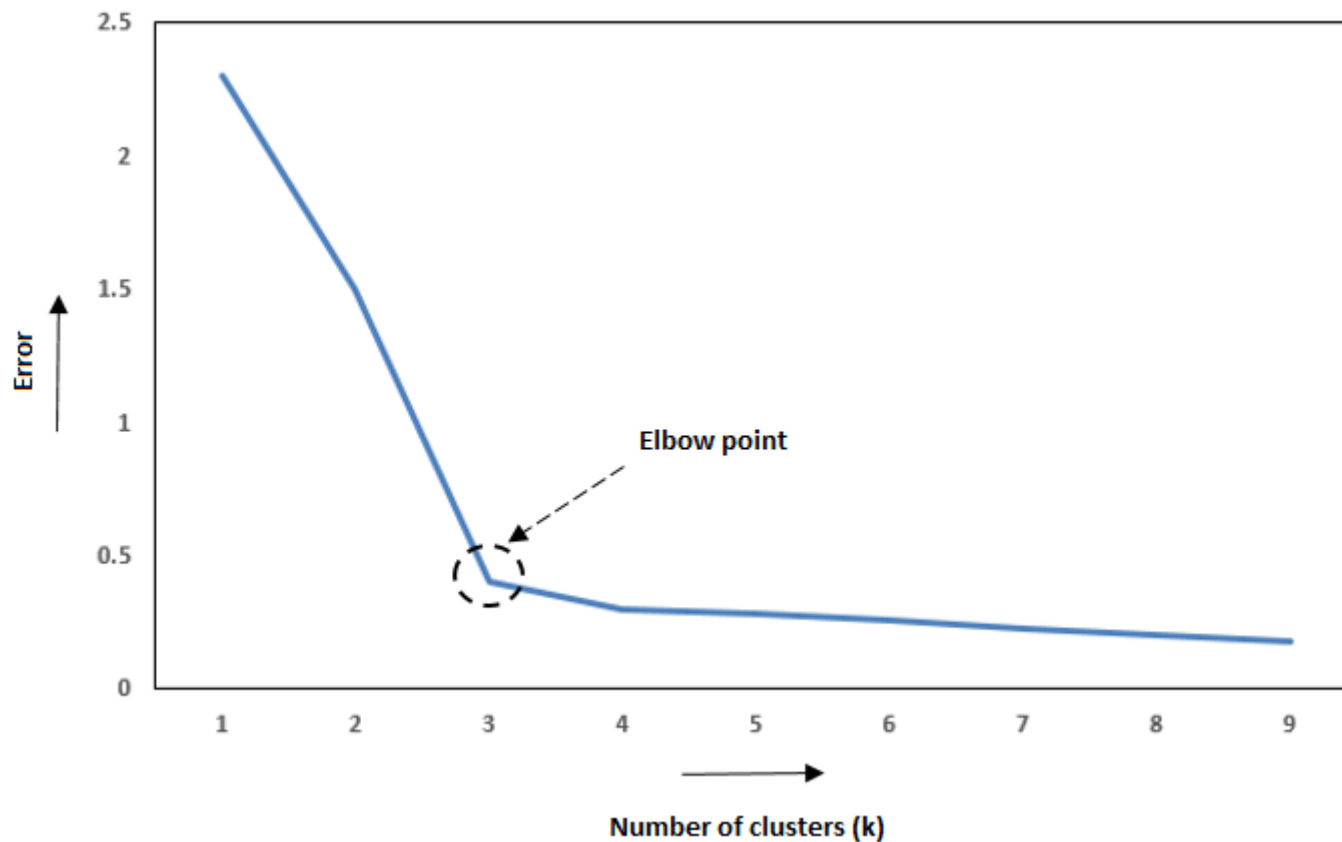
9

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# The Elbow Method

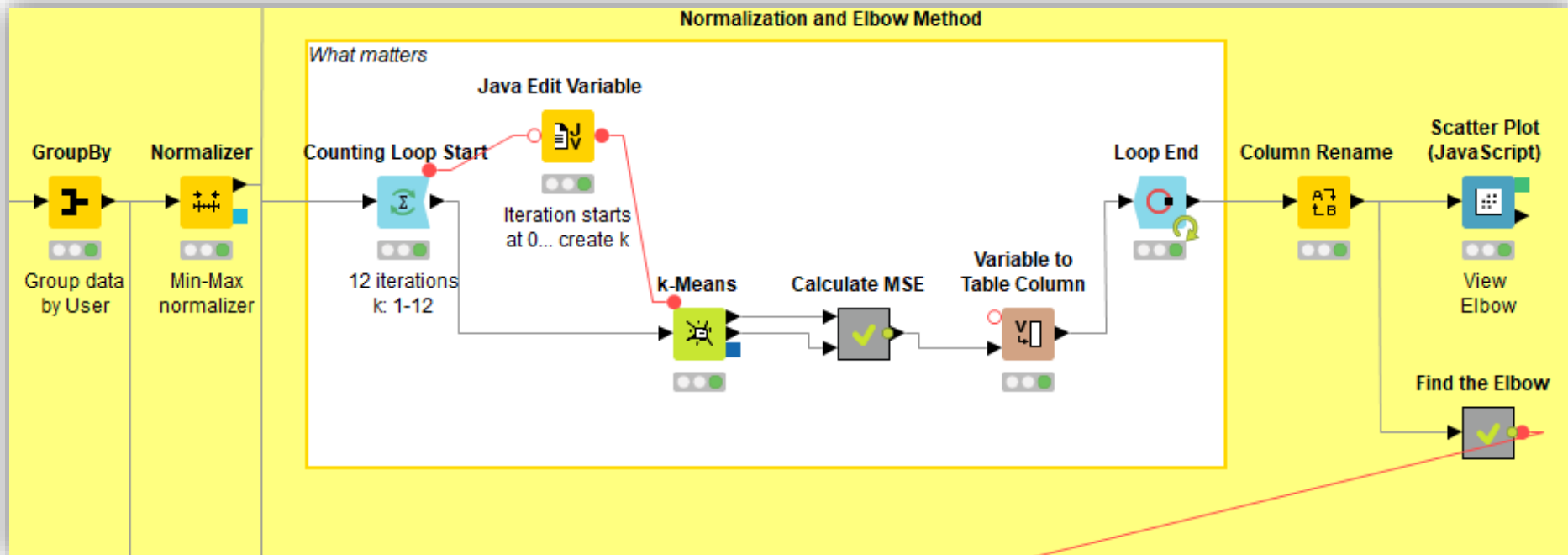
10

THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# The Elbow Method

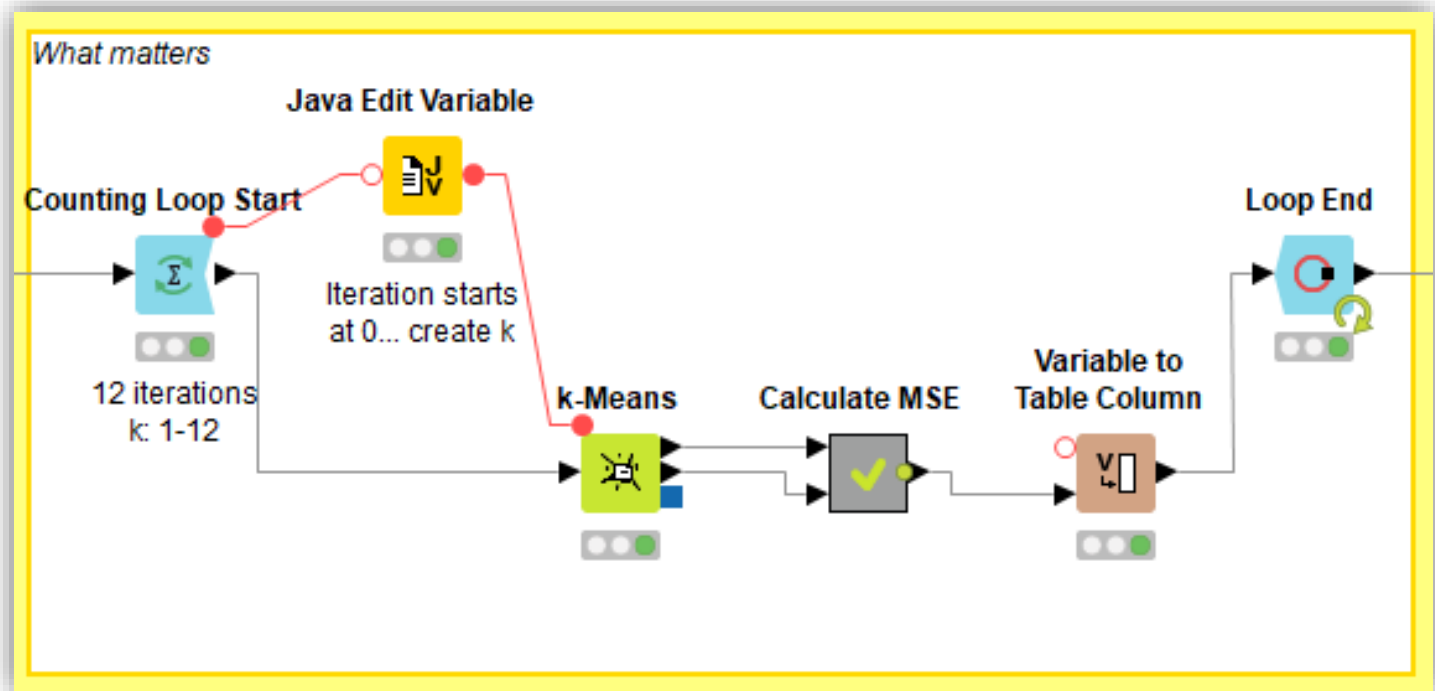
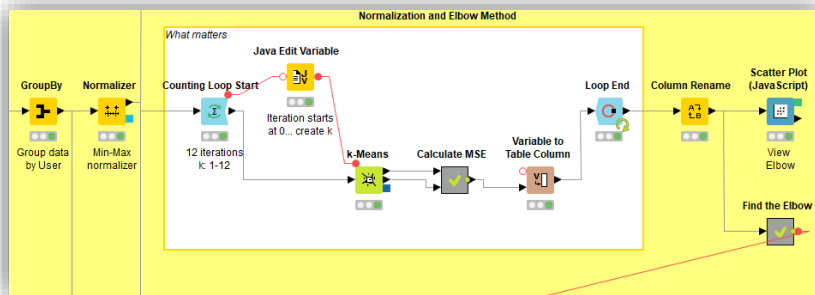
11

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# The Elbow Method

12

## THE ELBOW METHOD

Quality Measures

HTTP Requests

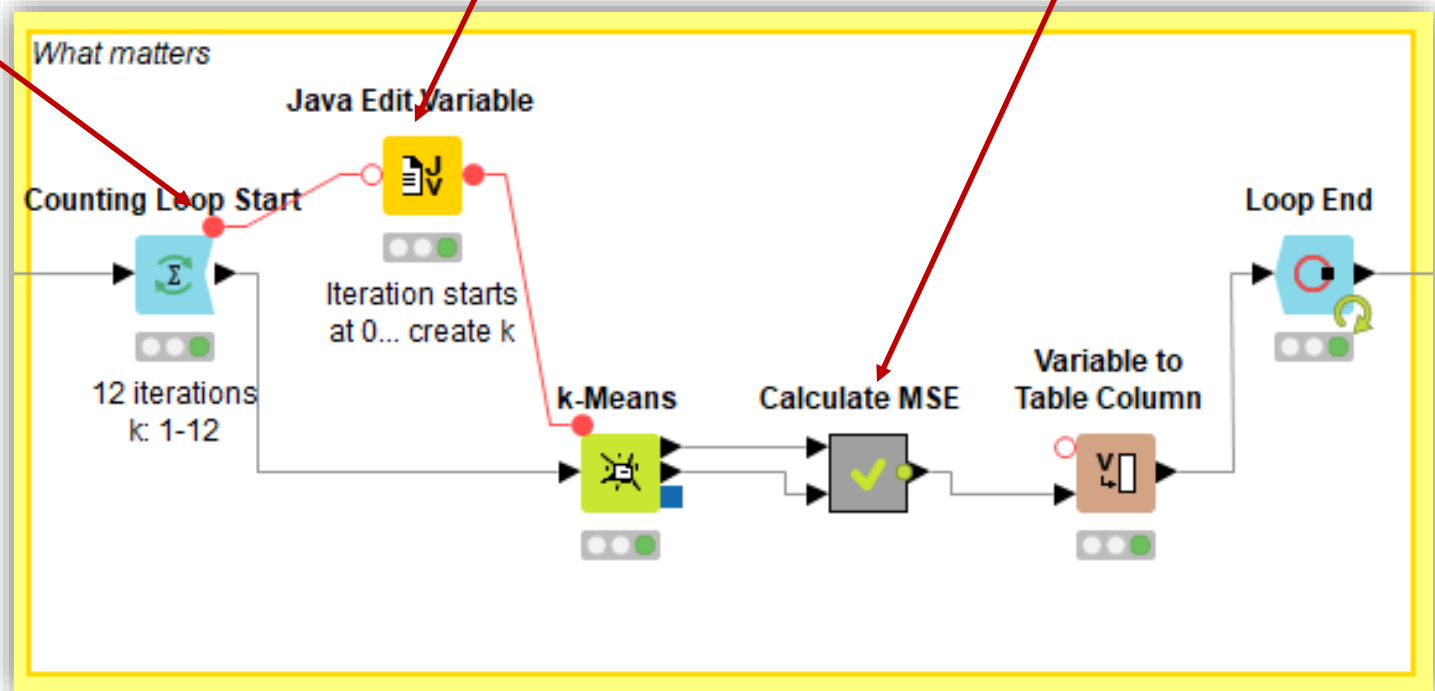
Hands On

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

13

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On

The image shows a KNIME workflow on the left and a 'Dialog - 2:142:0:123 - Java Edit Variable (iteration starts)' window on the right.

**Workflow:** The workflow consists of several nodes: 'GroupBy' (Group data by User), 'Normalizer' (Min-Max normalizer), 'Counting Loop Start' (12 iterations, k: 1-12), 'Java Edit Variable' (highlighted with a red box), and 'k-Means'. A red arrow points from the 'Java Edit Variable' node to the dialog window.

**Dialog - 2:142:0:123 - Java Edit Variable (iteration starts):**

- File:** Java Snippet, Additional Libraries, Additional Bundles, Templates, Flow Variables, Memory Policy.
- Flow Variable List:** currentIteration, knime.workspace, maxIterations.
- Code Editor:** Contains Java code for the loop iteration. The code includes comments for system imports, custom imports, system variables, custom variables, expression start, and expression end. The main code line is `out_k = v_currentIteration+1;`.
- Input:** A table with columns: Flow Variable, Java Type, Java Field. It contains one entry: `currentIteration` (Integer) mapped to `v_currentIteration`.
- Output:** A table with columns: Replace, Flow Variable, KNIME Type, Java Type, Java Field. It contains one entry: `k` (int) mapped to `out_k`.

Buttons at the bottom: OK, Apply, Cancel, and a help icon.

# The Elbow Method

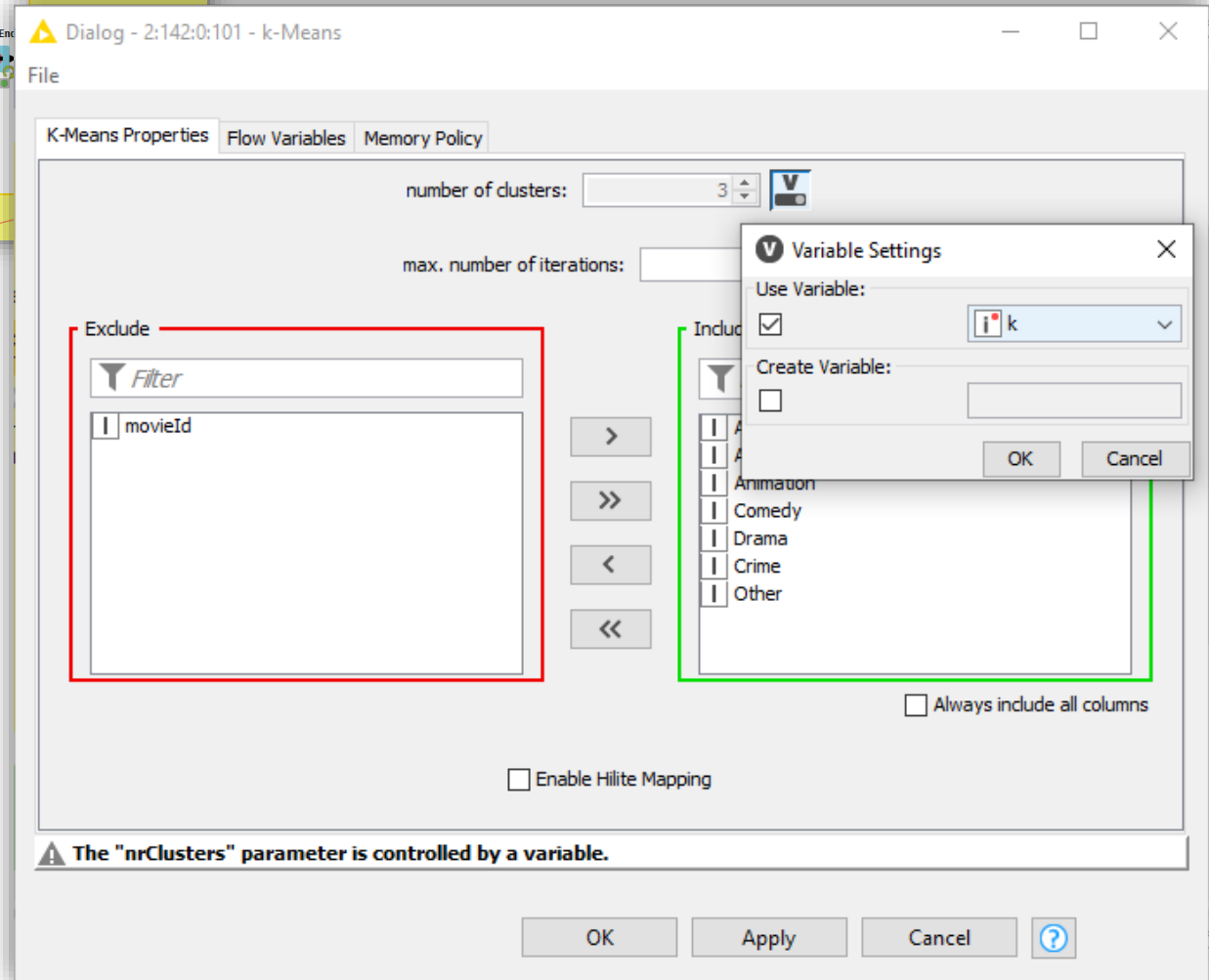
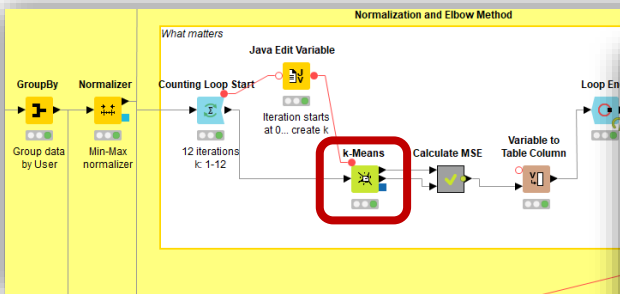
14

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# The Elbow Method

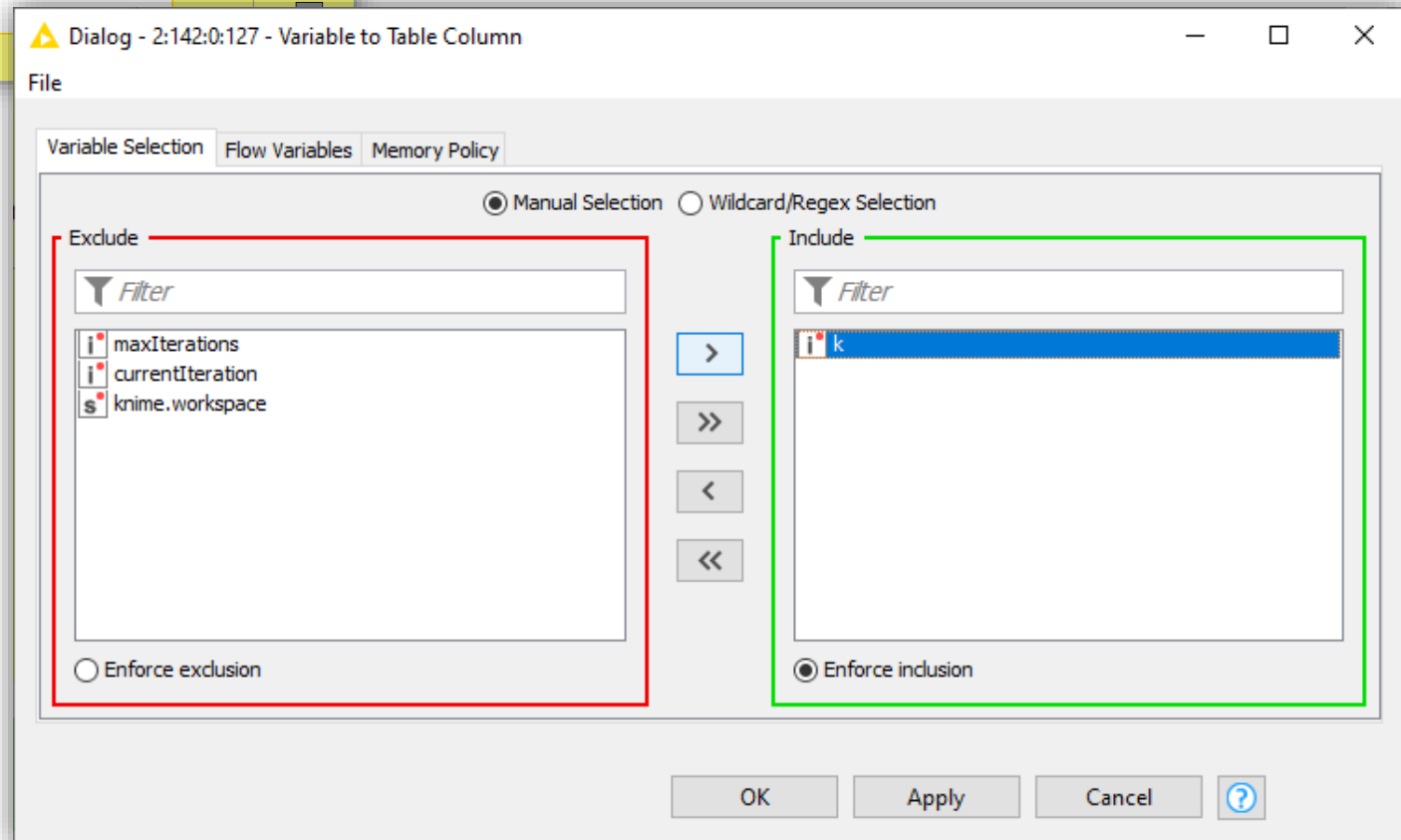
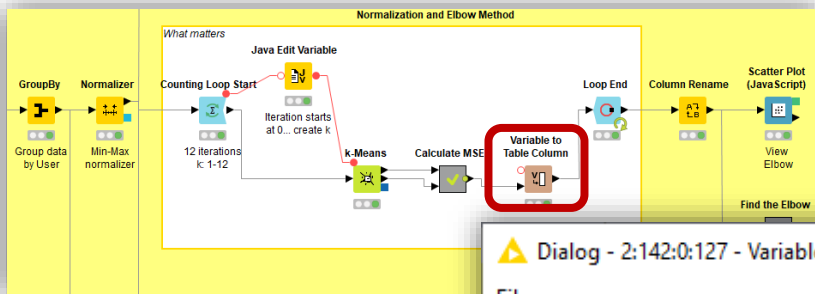
15

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



# The Elbow Method

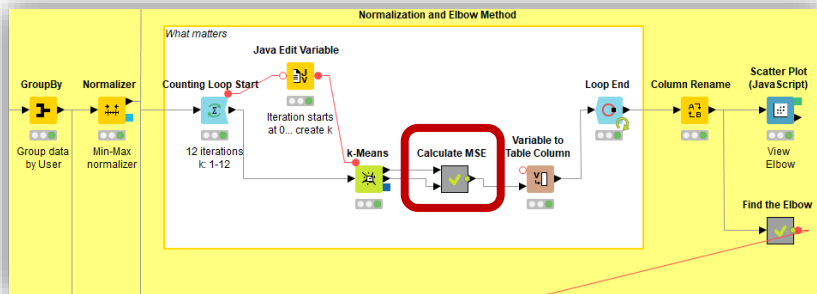
16

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



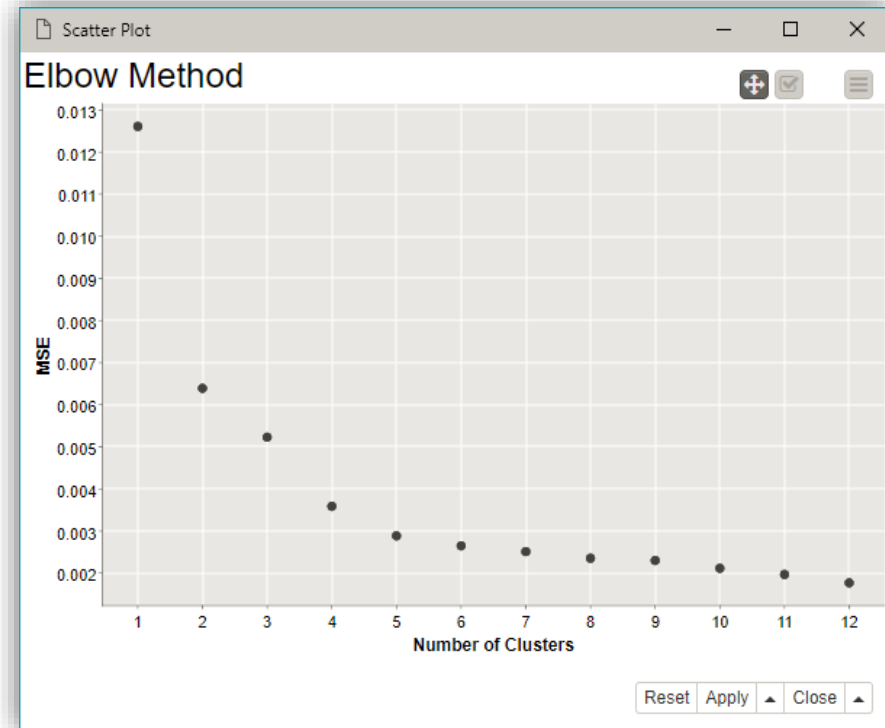
Collected results - ...

File Hilite Navigation View

Properties Table "default" - Rows: 12

Flow Variables 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





# The Elbow Method

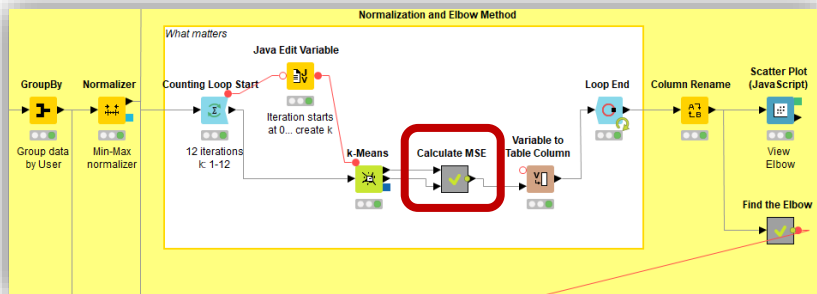
17

## THE ELBOW METHOD

Quality Measures

HTTP Requests

Hands On



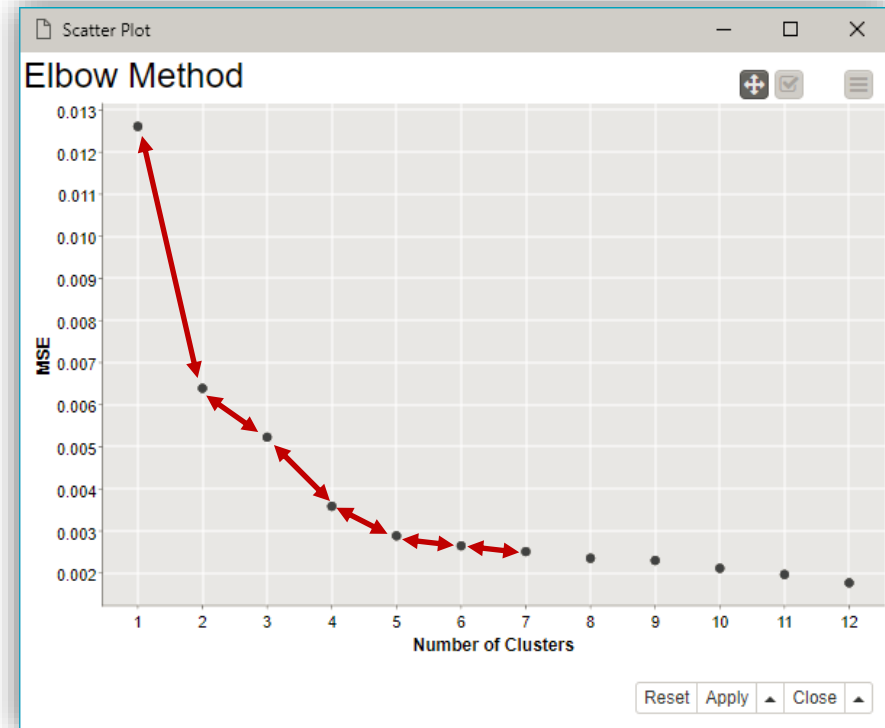
Collected results - ...

File Hilite Navigation View

Properties Table "default" - Rows: 12

Flow Variables 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

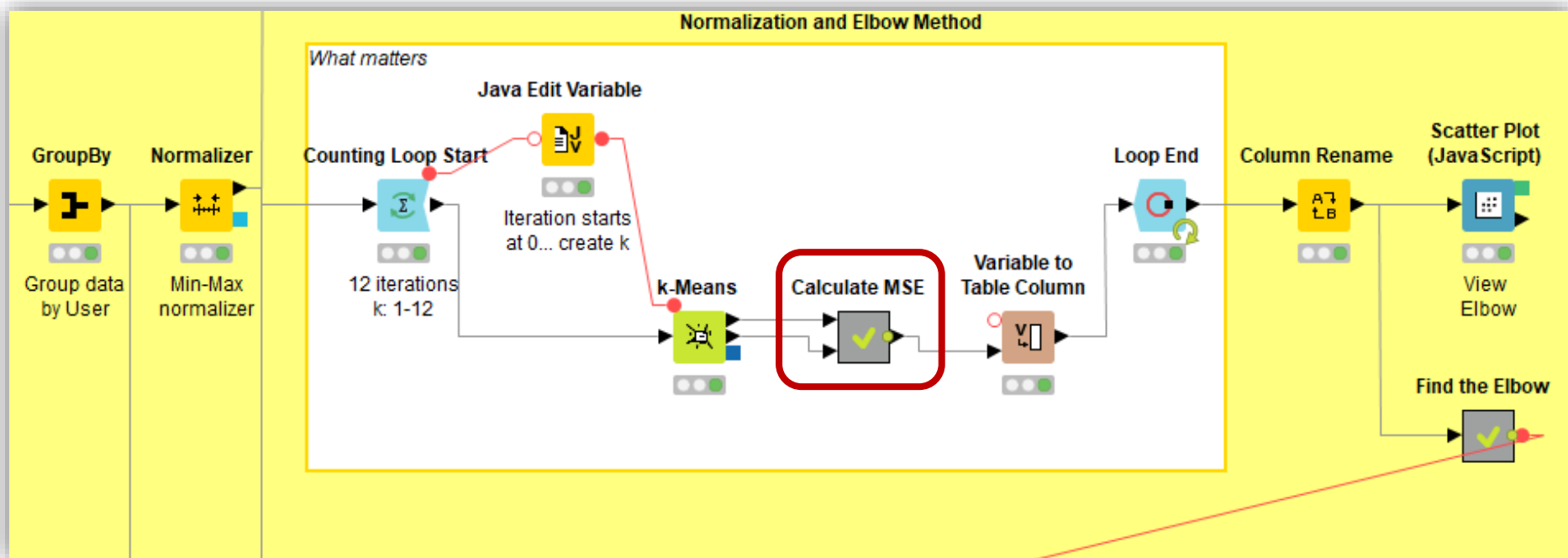
18

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



# Quality Measures

## MAE, MSE and RMSE

19

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On

### 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

20

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On

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

### 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  $\infty$  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

21

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On

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

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

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

#	Error	Error	Error <sup>2</sup>
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	Error <sup>2</sup>
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

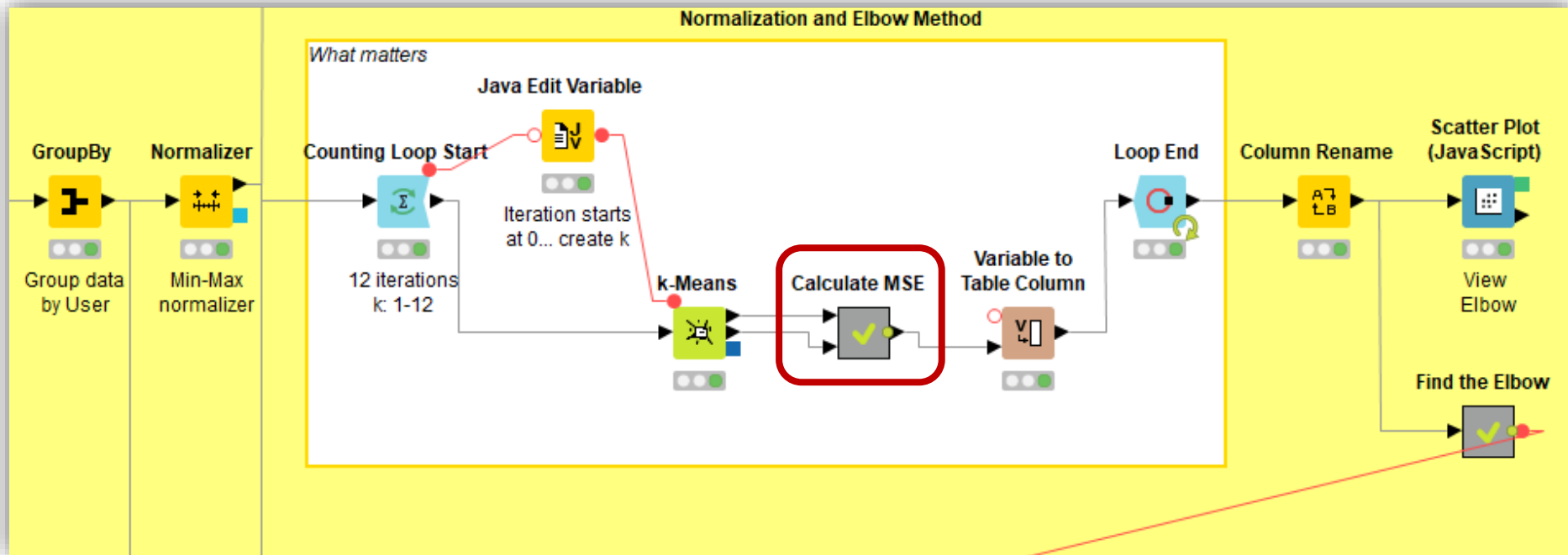
22

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



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

# Quality Measures for Clustering

## The Elbow Method

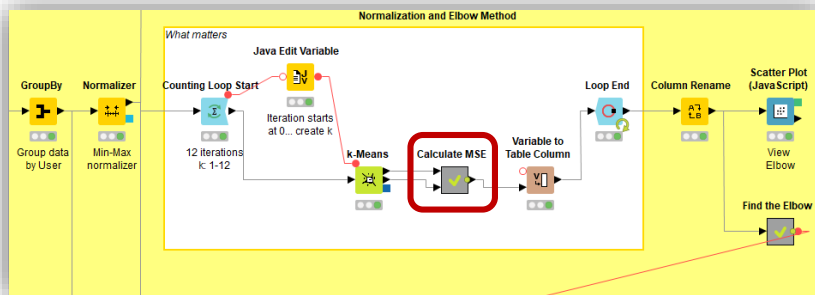
23

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



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	D Animation	D Comedy	D Crime	D Drama	D 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	D Comedy	D Crime	D Drama	D Other	S 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

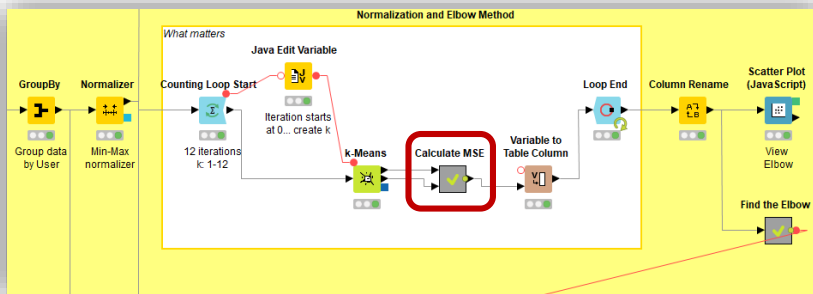
24

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



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	D Animation	D Comedy	D Crime	D Drama	D 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	D Comedy	D Crime	D Drama	D Other	S 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

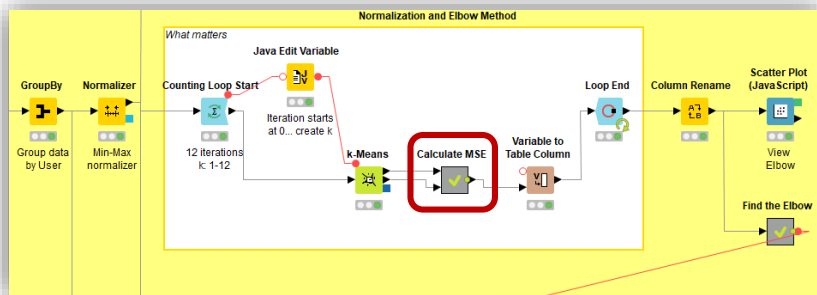
25

The Elbow Method

QUALITY MEASURES

HTTP Requests

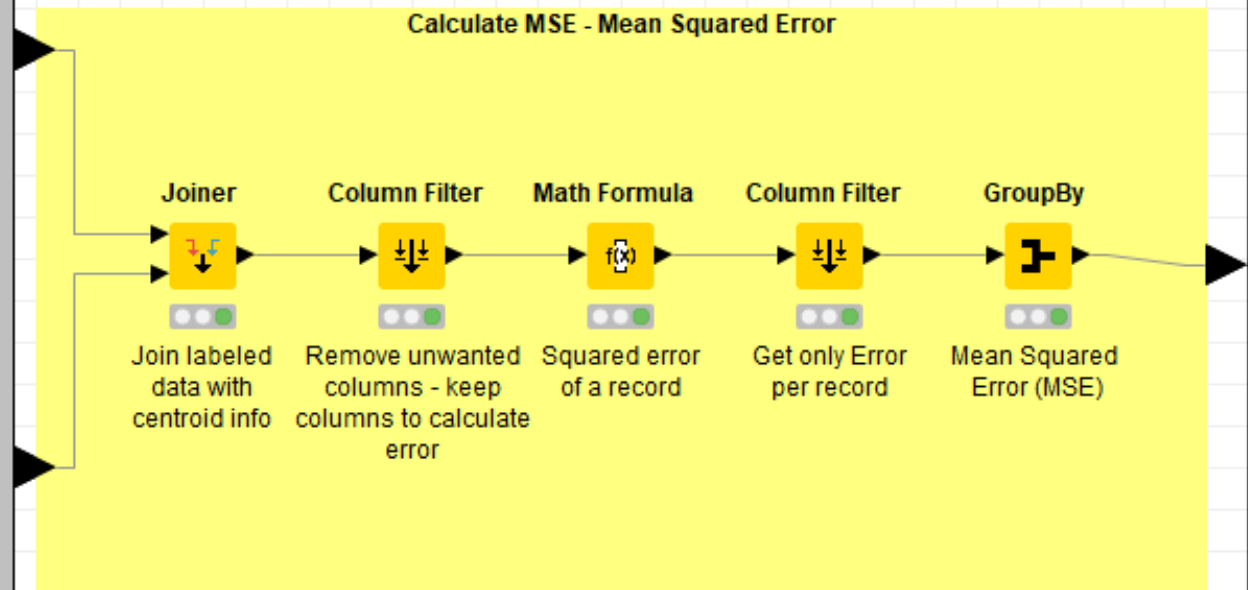
Hands On



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

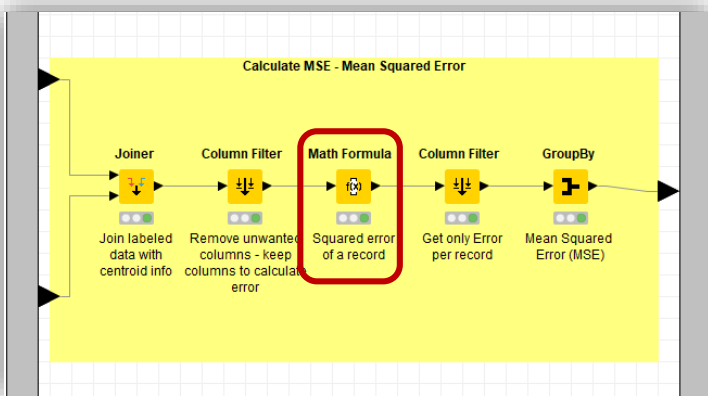
26

The Elbow Method

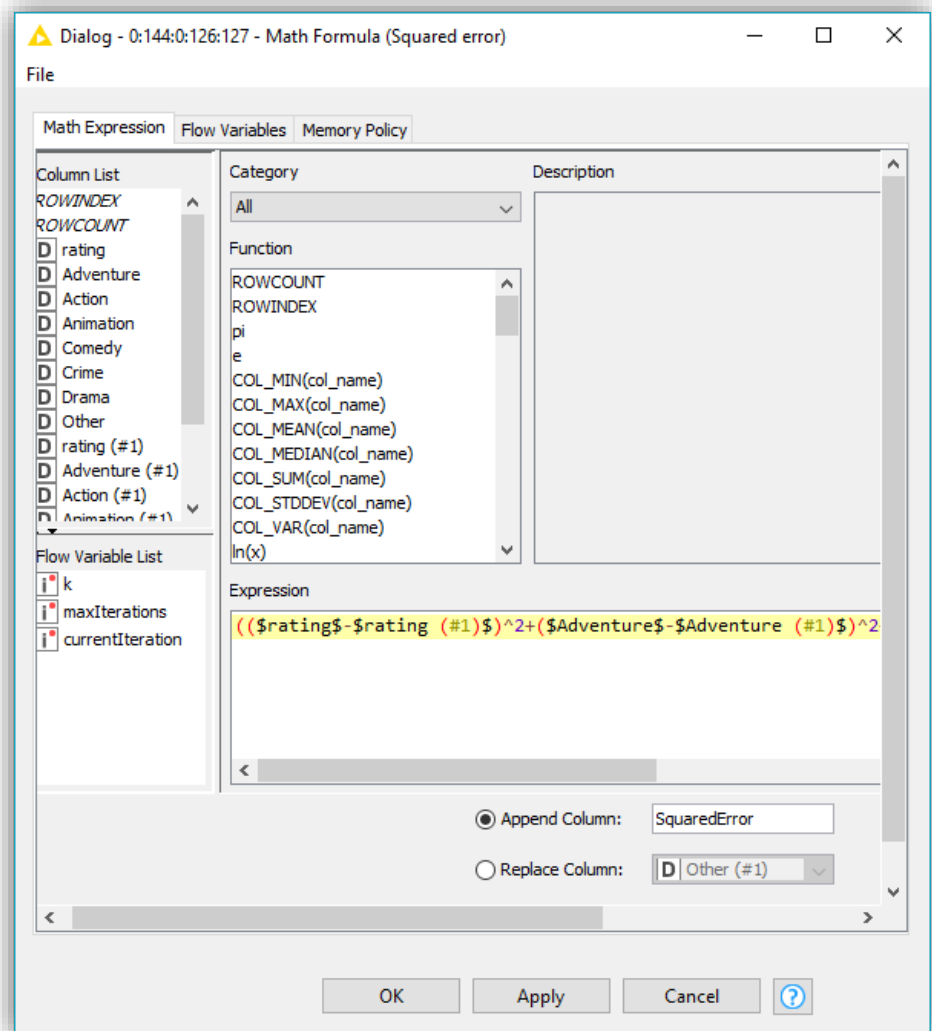
QUALITY MEASURES

HTTP Requests

Hands On



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



# Quality Measures for Clustering

## The Elbow Method Metanode

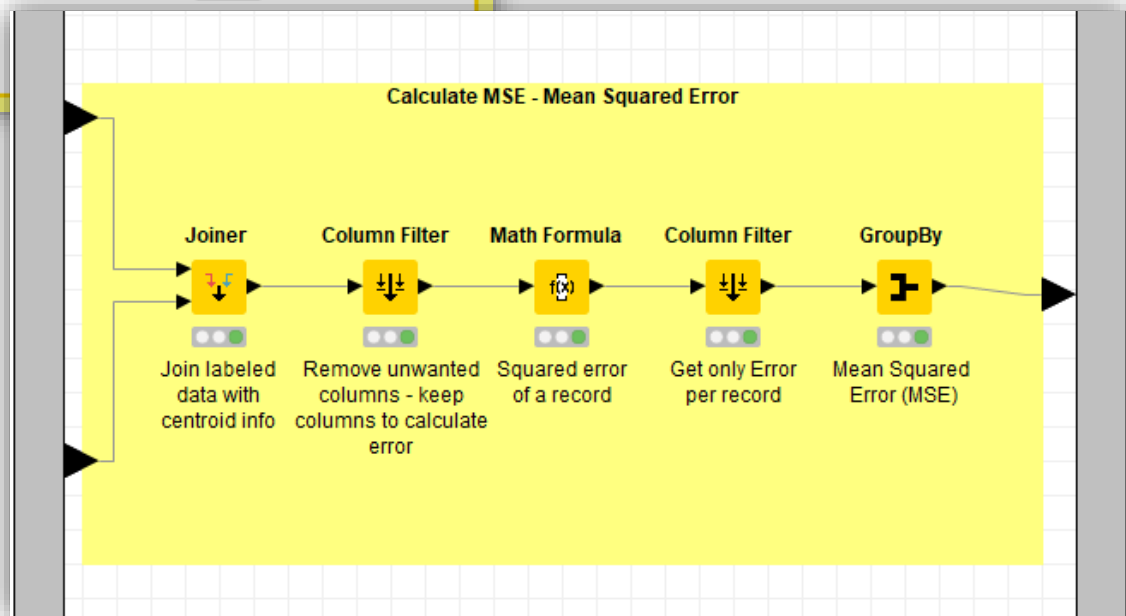
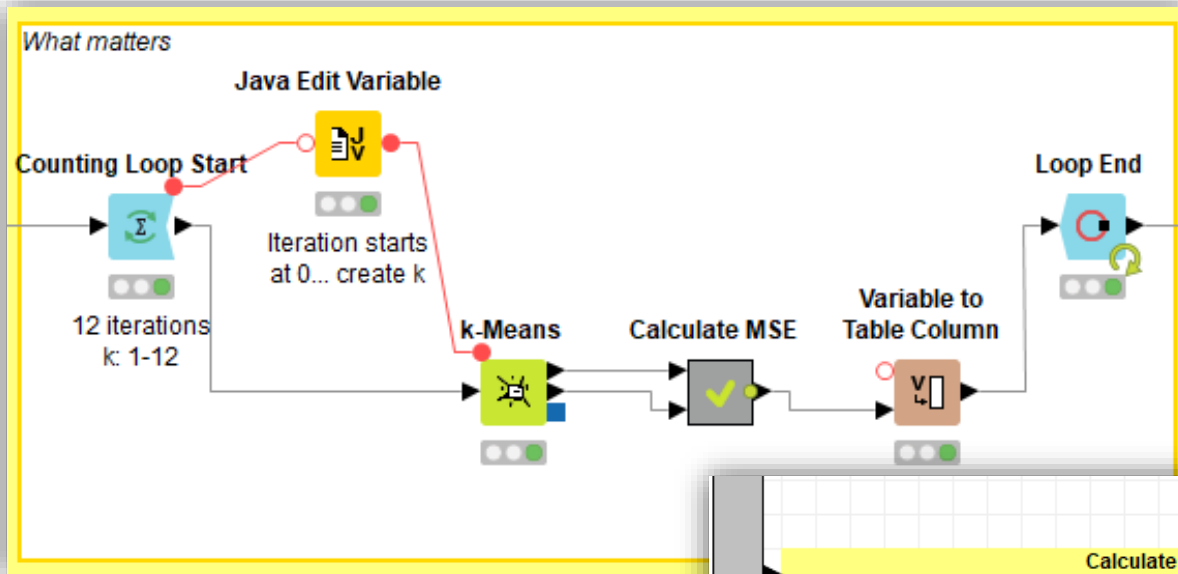
27

## The Elbow Method

## QUALITY MEASURES

## HTTP Requests

## Hands On



# Quality Measures for Clustering

## The Elbow Method Metanode

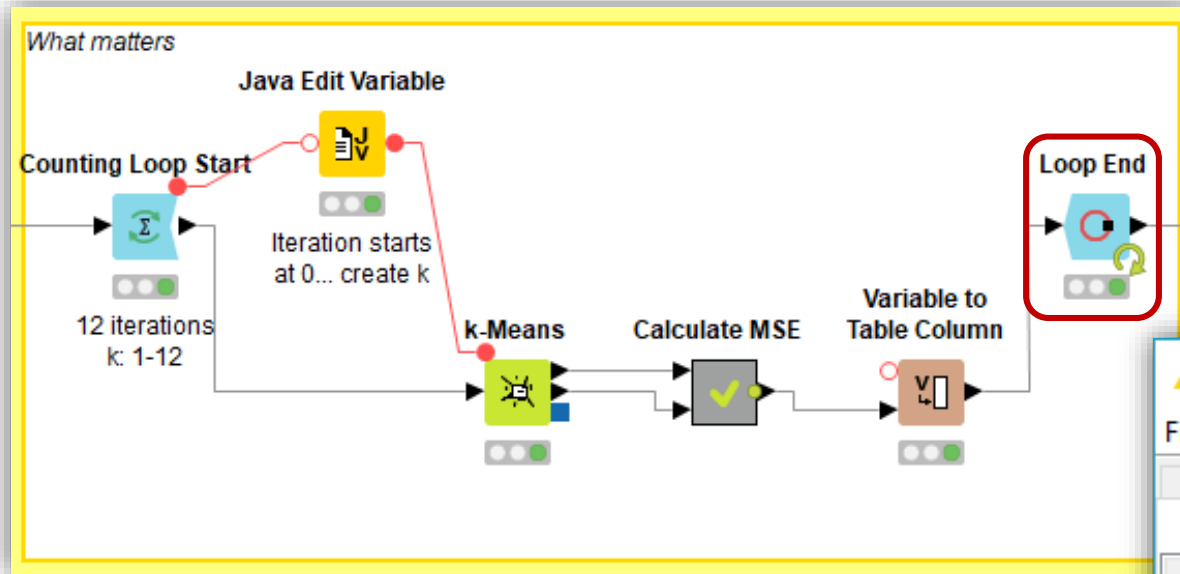
28

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



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

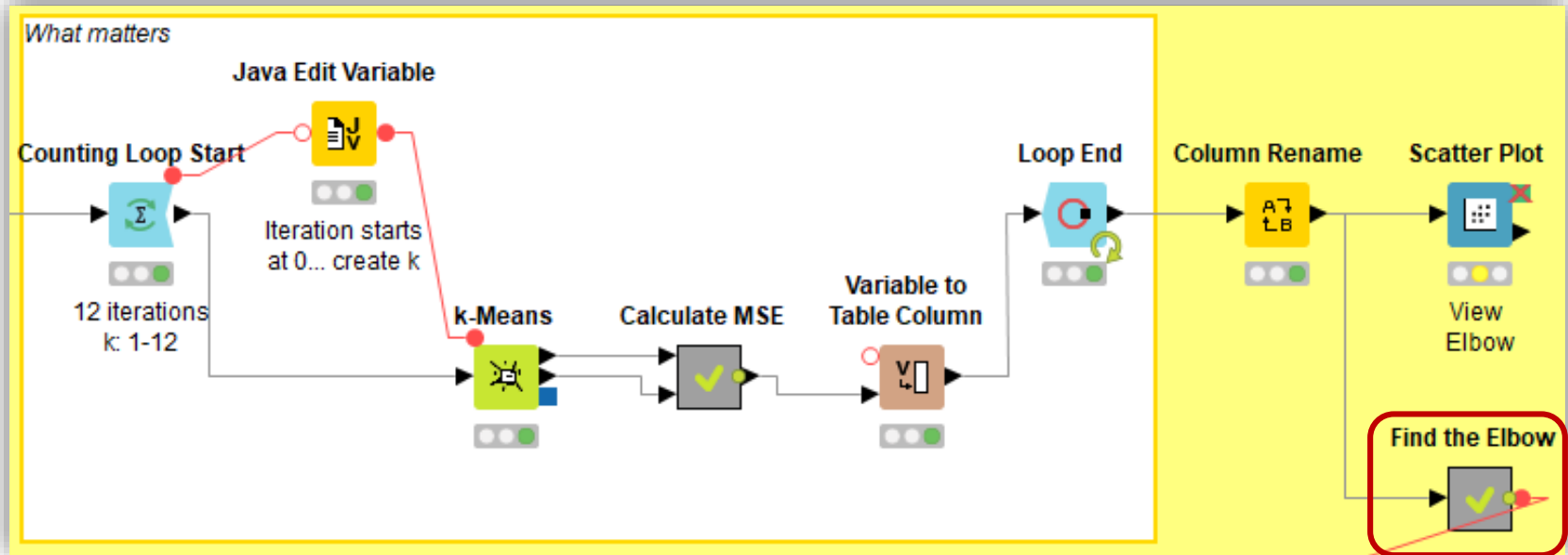
29

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



# Quality Measures for Clustering

## Finding the Elbow ... Automatically

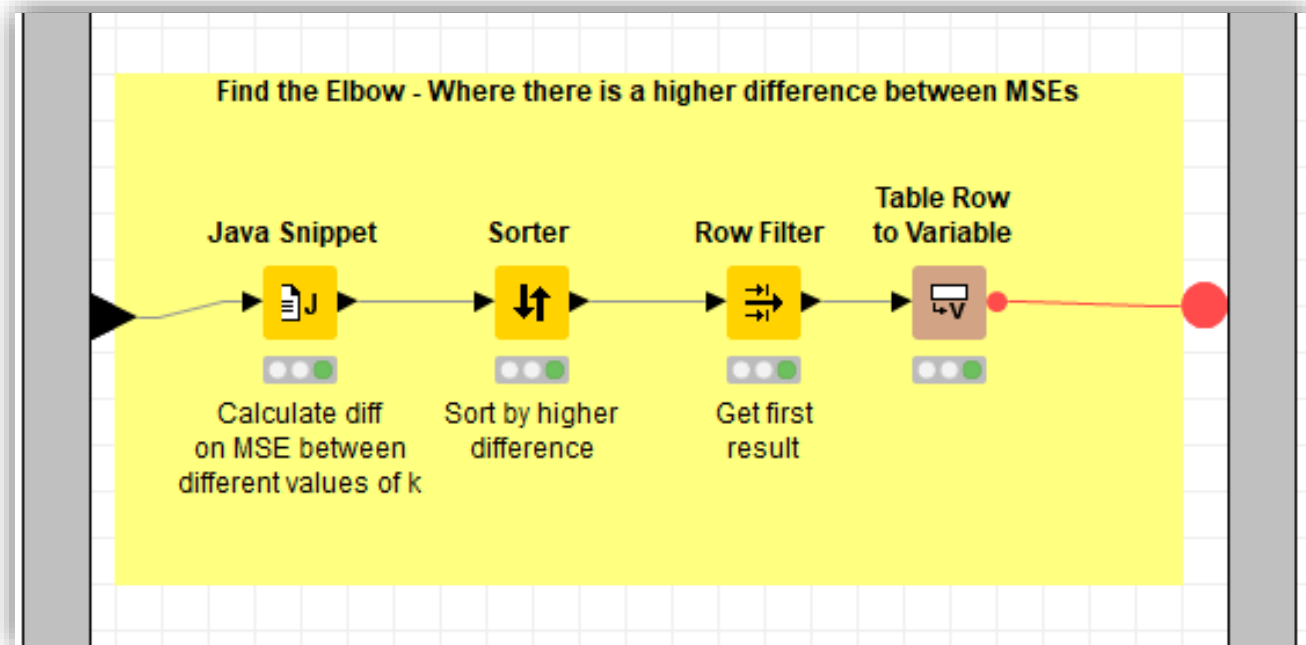
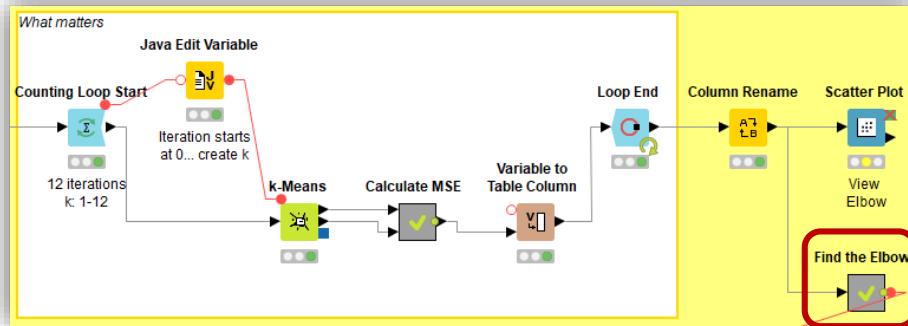
30

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On



# Quality Measures for Clustering

## Finding the Elbow ... Automatically

31

The Elbow Method

QUALITY MEASURES

HTTP Requests

Hands On

Find the Elbow - Where there is a higher difference between MSEs

Java Snippet

Sorter

Row Filter

Table Row  
to Variable

Calculate diff  
on MSE between  
differ

Sort by higher  
difference

Get first  
result

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

File Hilite Navigation View

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

Row ID	D MSE	I k	D 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	D MSE	I k	D 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

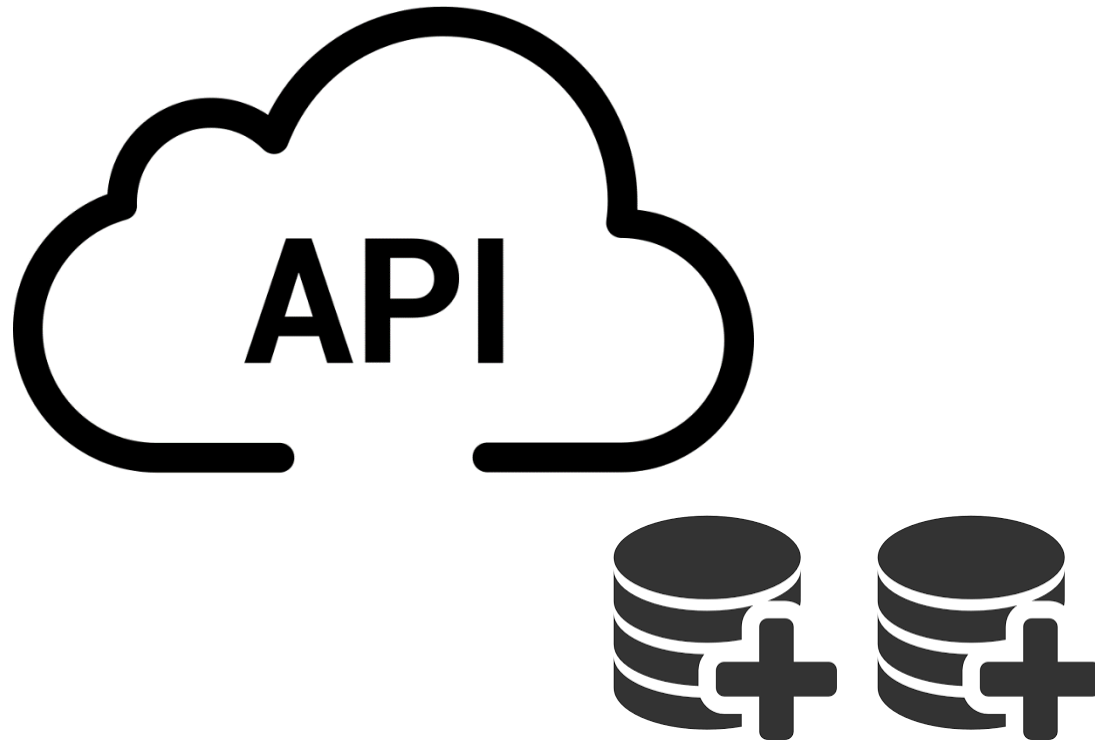
32

The Elbow Method

Quality Measures

**HTTP REQUESTS**

Hands On





# HTTP Requests

## API Calls

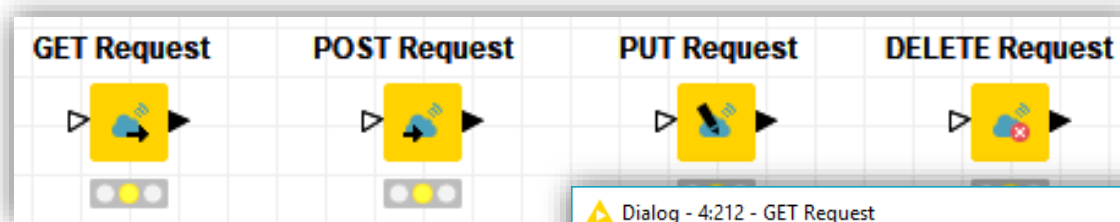
33

The Elbow Method

Quality Measures

HTTP REQUESTS

Hands On



Dialog - 4:212 - GET Request

File

Connection Settings Authentication Request Headers Response Headers Flow Variables Memory Policy

☒ URL:

☐ URL column:

☐ Delay (ms):

Concurrency:

SSL

☐ Ignore hostname mismatches

☐ Trust all certificates

☒ Fail on connection problems (e.g. timeout, certificate errors, ...)

☒ Fail on http errors (e.g. page not found)

☒ Follow redirects

Timeout (s)

Body column:

OK Apply Cancel ?

# HTTP Requests

## JSON to Table

34

The Elbow Method

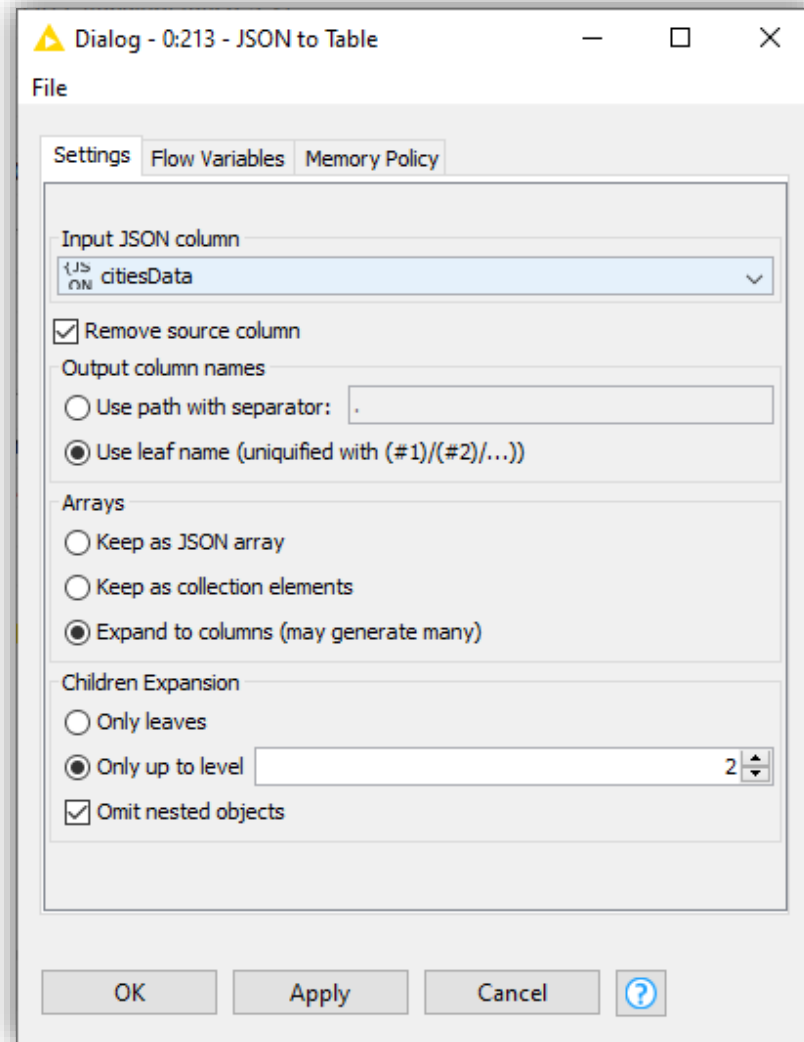
Quality Measures

HTTP REQUESTS

Hands On



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



# HTTP Requests

## Handle the JSON

35

The Elbow Method

Quality Measures

HTTP REQUESTS

Hands On

GET results - 0:212 - GET Request

File Hilite Navigation View

Table "default" - Rows: 1 Spec - Columns: 3 Properties Flow Variables

Row ID	Status	Content type	citiesData
Row0	200	application/json; charset=utf-8	<pre>{   "meta": {     "name": "openaq-api",     "license": "CC BY 4.0",     "website": "https://docs.op   },   "results": [     {       "country": "PT",       "name": "Aveiro",       "city": "Aveiro",       "count": 171326,       "locations": 4     },     {       "country": "PT",       "name": "Braga",       "city": "Braga",       "count": 81279,       "locations": 3     }   ] }</pre>

Transposed Table - 0:214 - Transpose

File Hilite Navigation View

Table "default" - Rows: 16 Spec - Column: 1 Properties Flow Variables

Row ID	Row0
results (#1)	<pre>{   "country": "PT",   "name": "Aveiro",   "city": "Aveiro",   "count": 171326,   "locations": 4 }</pre>
results (#2)	<pre>{   "country": "PT",   "name": "Braga",   "city": "Braga",   "count": 81279,   "locations": 3 }</pre>
results (#3)	<pre>{   "country": "PT",   "name": "Castelo Branco",   "city": "Castelo Branco",   "count": 57894,   "locations": 1 }</pre>
results (#4)	<pre>{   "country": "PT",   "name": "Coimbra",   "city": "Coimbra",   "count": 86655,   "locations": 3 }</pre>
results (#5)	<pre>{   "country": "PT",   "name": "Évora",   "city": "Évora",   "count": 55441,   "locations": 1 }</pre>

Extracted values - 0:215 - JSON to Table

File Hilite Navigation View

Table "default" - Rows: 16 Spec - Columns: 5 Properties Flow Variables

Row ID	country	name	city	count	locations
results (#1)	PT	Aveiro	Aveiro	171326	4
results (#2)	PT	Braga	Braga	81279	3
results (#3)	PT	Castelo Branco	Castelo Branco	57894	1
results (#4)	PT	Coimbra	Coimbra	86655	3
results (#5)	PT	Évora	Évora	55441	1
results (#6)	PT	Faro	Faro	158717	4
results (#7)	PT	Ilha da Mad...	Ilha da Mad...	110060	3
results (#8)	PT	Ilha do Faial	Ilha do Faial	60336	1
results (#9)	PT	Leiria	Leiria	56116	1
results (#10)	PT	Lisboa	Lisboa	705655	15
results (#11)	PT	Porto	Porto	478174	15
results (#12)	PT	Santarém	Santarém	66310	1
results (#13)	PT	Setúbal	Setúbal	653196	12
results (#14)	PT	Viana do Ca...	Viana do Ca...	40303	1
results (#15)	PT	Vila Real	Vila Real	59815	1
results (#16)	PT	Viseu	Viseu	32882	1

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

# HTTP Requests

## A Workflow

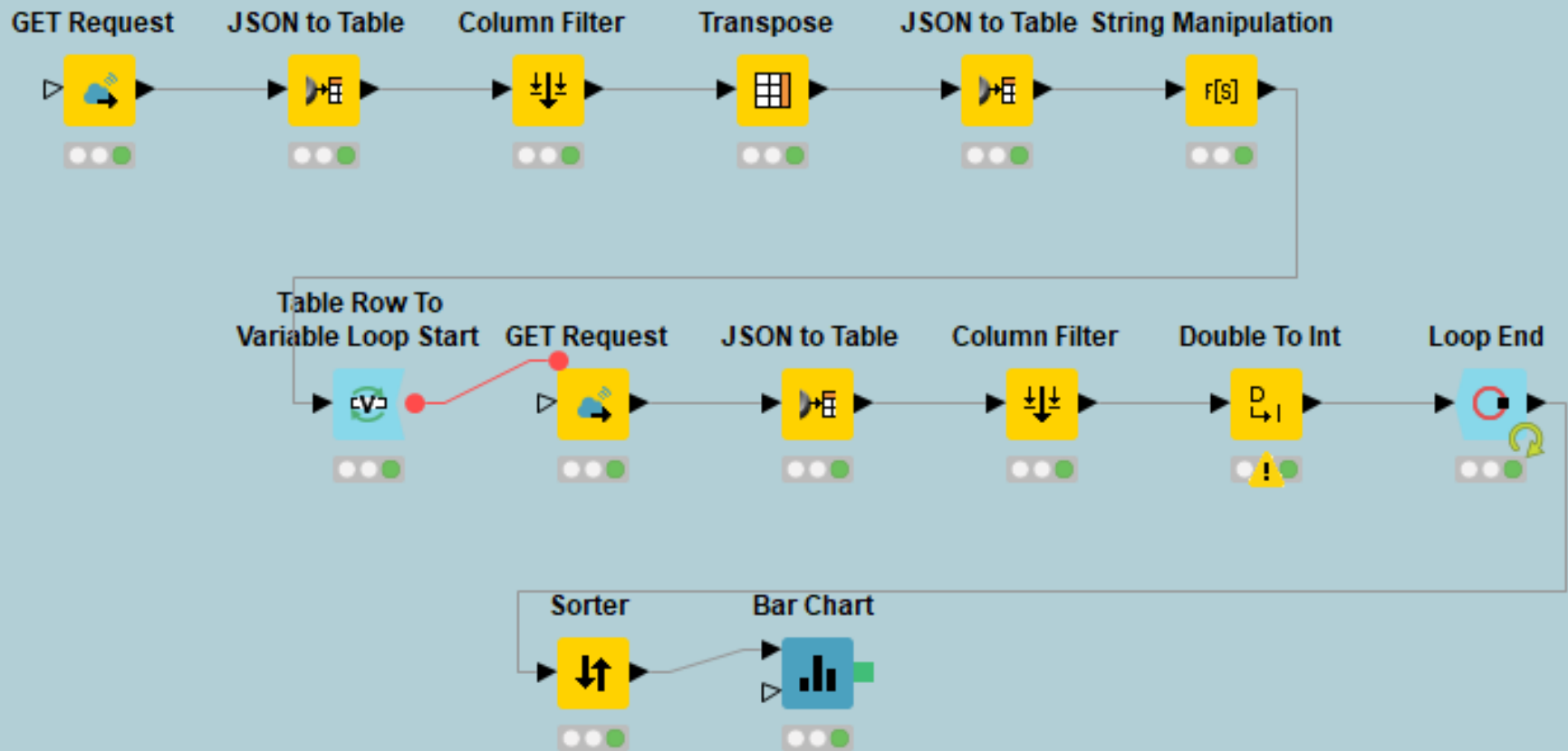
36

The Elbow Method

Quality Measures

HTTP REQUESTS

Hands On



# HTTP Requests

## A Workflow

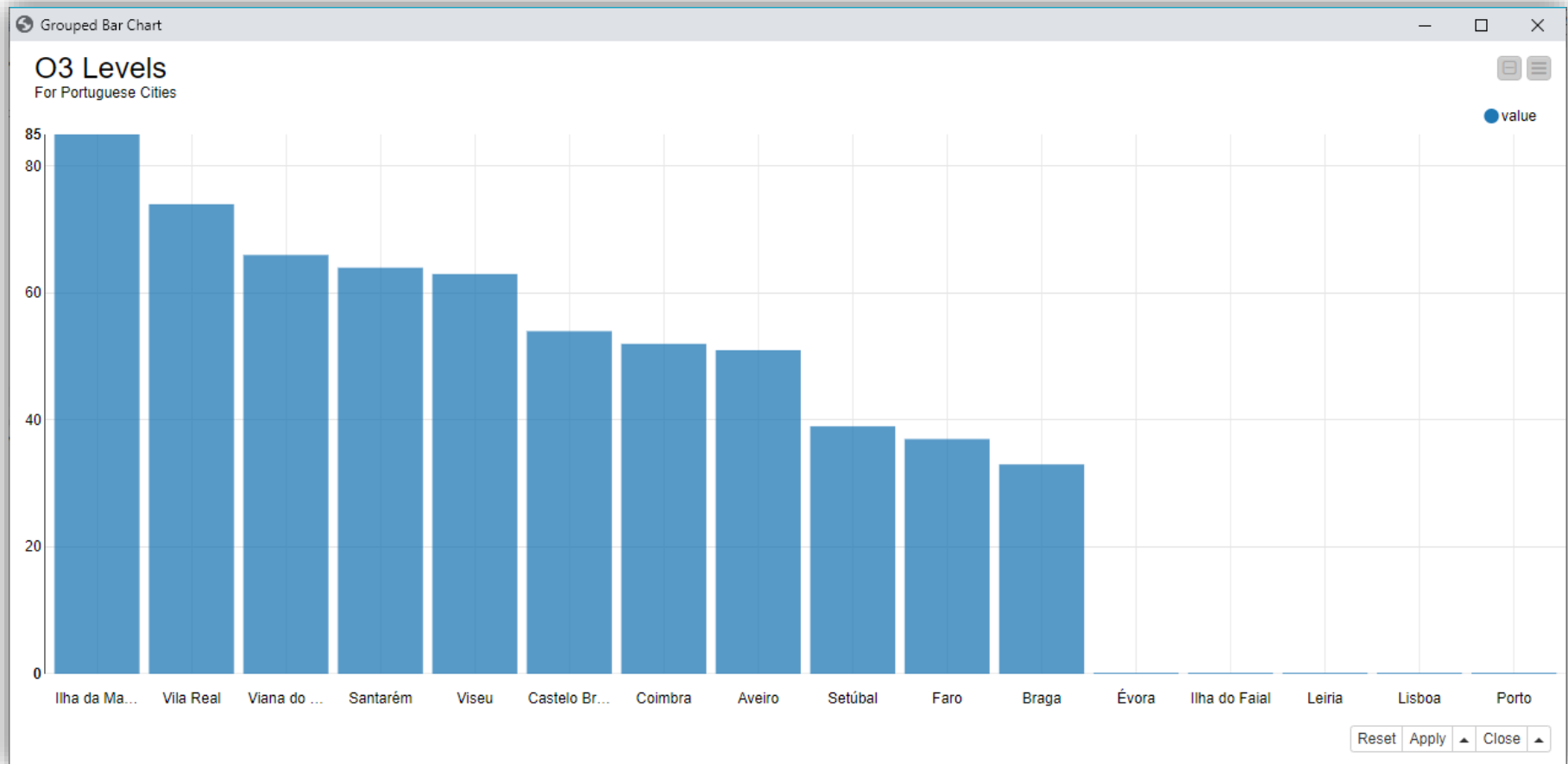
37

The Elbow Method

Quality Measures

**HTTP REQUESTS**

Hands On



# HTTP Requests

## API Calls

38

The Elbow Method

Quality Measures

**HTTP REQUESTS**

Hands On



<https://openweathermap.org/api>



<https://docs.openaq.org/>



FOR DEVELOPERS

<https://developer.tomtom.com/>



<https://www.openuv.io/uvindex>



<https://developers.google.com/>



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



<https://developers.coinbase.com/>

# Hands On

39

The Elbow Method

Quality Measures

HTTP Requests

**HANDS ON**

