

데이터 마이닝 특강

Practice session

[Clustering]

Junseok Park

2018-08-23



Clustering

Introduction review

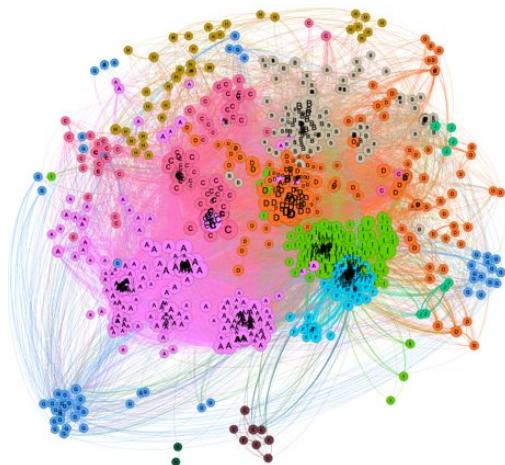
- **Datamining applications**
- **Datamining tool installation**
 - Orange3
 - Anaconda
 - Pycharm
- **Practice lecture resources**
 - <https://github.com/junseokpark/resources/datamining>
- **Cloud datamining tools**
 - Cloudera data science workbench
 - <https://www.cloudera.com/products/data-science-and-engineering/data-science-workbench.html>
 - Amazon machine learning (ML)
 - <https://aws.amazon.com/ko/machine-learning/>
 - Google cloud AI
 - <https://cloud.google.com/products/ai/>
 - Databricks unified analytic platform
 - <https://databricks.com/product/unified-analytics-platform>

Overview

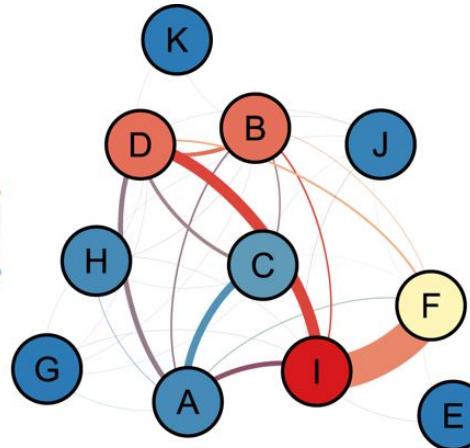
- **Concept of Clustering**
- **Practice of Clustering algorithms on toy data**
 - k-mean
 - Hierarchical-clustering
 - Self Organizing Map
- **Practice of Clustering algorithms on realworld data**
 - Data introduction
 - Data preprocessing
 - k-mean
 - Hierarchical-clustering

Introduction

- **Cluster: a collection of data objects**
 - Similar (or related) to one another within the same group
 - Dissimilar (or unrelated) to the objects in other groups
- **Cluster analysis (or clustering, data segmentation, ...)**
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- **Example^[1]**



(a) Effect-Size network



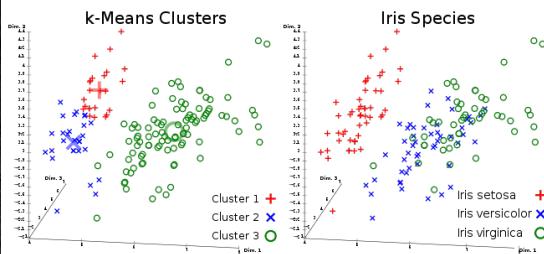
(b) Induced network

^[1] Valenzuela JF, Monterola C, Tong VJC, Ng TP, Larbi A, 2017. Health and disease phenotyping in old age using a cluster network analysis. *Scientific Reports* 7, 1 (2017/11/15), 15608. DOI= <http://dx.doi.org/10.1038/s41598-017-15753-3>.

Dataset

- **iris flower data set**

- multivariate data set
- introduced by Ronald Fisher in 1936
 - Statistician and Biologist



- the data to quantify the morphologic variation of Iris flowers of three related species



Iris setosa



Iris versicolor



Iris virginica

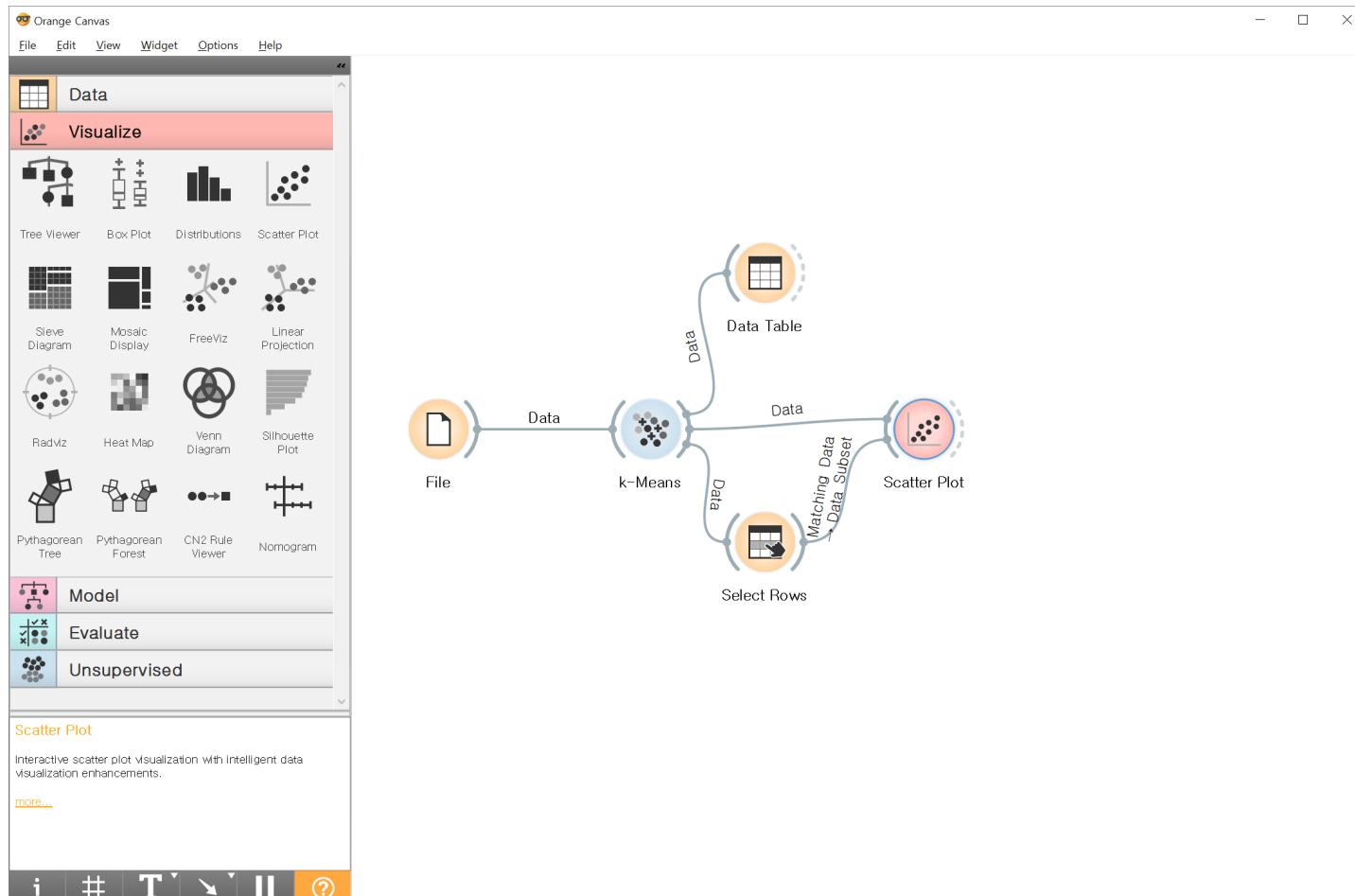
A screenshot of a software interface for managing datasets. At the top, there are 'File' and 'URL' input fields, and a 'Reload' button. Below is an 'Info' section with the title 'Iris flower dataset' and a description: 'Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor. 150 instance(s), 4 feature(s), 0 meta attribute(s). Classification: categorical class with 3 values,'. Underneath is a table titled 'Columns (Double click to edit)' with five rows. The columns are 'Name', 'Type', 'Role', and 'Values'. The rows are:

Name	Type	Role	Values
1 sepal length	N numeric	feature	
2 sepal width	N numeric	feature	
3 petal length	N numeric	feature	
4 petal width	N numeric	feature	
5 iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

A red box highlights the 'Apply' button at the bottom right of the interface.

k-means (Cont'd)

- Build schema



k-means (Cont'd)

- Load data

- Iris dataset into three clusters

The screenshot shows the Orange data mining software interface. The top bar has 'File' selected, with 'iris.tab' chosen as the file. Below the file menu are 'Info' and 'URL' fields. The 'Info' section displays details about the 'Iris flower dataset': 'Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.', '150 instance(s), 4 feature(s), 0 meta attribute(s)', and 'Classification: categorical class with 3 values.' The main area is titled 'Columns (Double click to edit)' and contains a table:

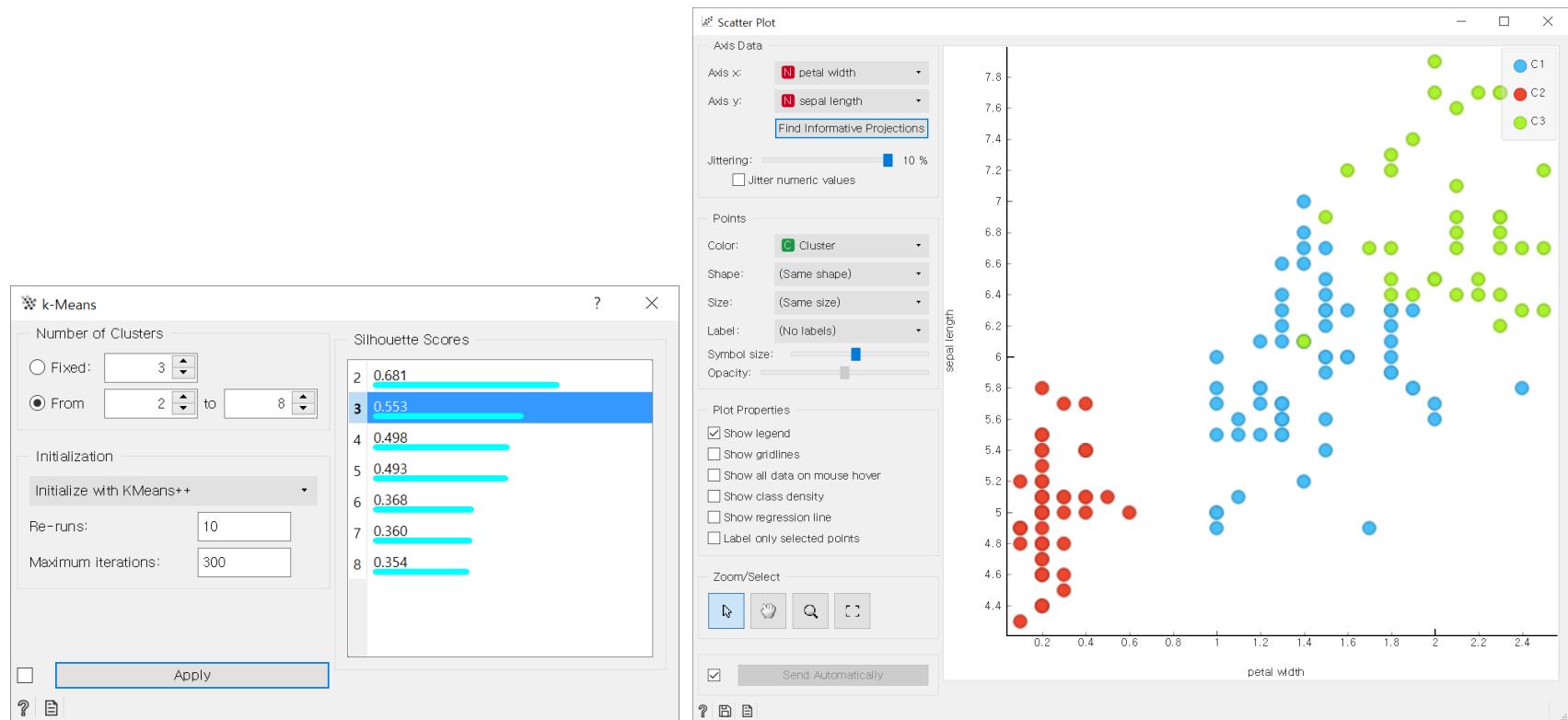
Name	Type	Role	Values
1 sepal length	N numeric	feature	
2 sepal width	N numeric	feature	
3 petal length	N numeric	feature	
4 petal width	N numeric	feature	
5 Iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

At the bottom are buttons for 'Browse documentation datasets' and 'Apply'.

k-means (Cont'd)

- Observation of k-means

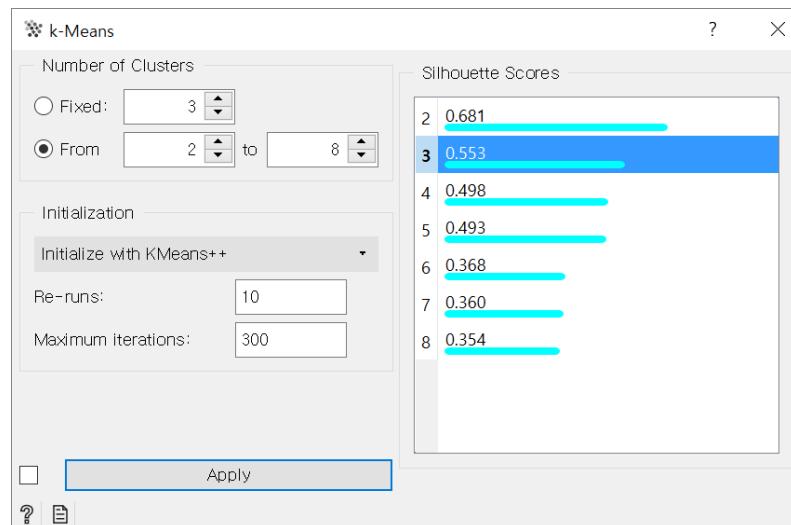
- k-means added the cluster index as a class attribute
- the scatter plot will color the points according to the clusters they are in



k-means (Cont'd)

- **Hyperparameters and configuration**

- Fixed : algorithm clusters data in a specified number of clusters.
- From : widget shows clustering scores for the selected cluster range
- Silhouette scores : contrasts average distance to elements in the same cluster with the average distance to elements in other clusters
- initialization
 - Kmeans++ : first center is selected randomly, subsequent are chosen from the remaining points with probability proportioned to squared distance from the closest center
 - Random : clusters are assigned randomly at first and then updated with further iterations
- Re-runs : how many times the algorithm is run
- Maximum iterations : the maximum number of iteration within each algorithm run



K-means (Cont'd)

- **Silhouette scores**
 - a method of interpretation and validation of consistency within [clusters of data^{\[1\]}](#)
 - a succinct graphical representation of how well each object lies within its cluster^[2]
- **Silhouette value**
 - a measure of how similar an object is to its own cluster (cohesion) compared to other cluster s (separation)
 - ranges from -1 to +1
- **Calculation methods**
 - Euclidian distance
 - Manhattan distance^[3]
- **Definition**
 - $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, which can be also written as $s(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i) \end{cases}$
 - $-1 \leq s(i) \leq 1$

[1] [https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))

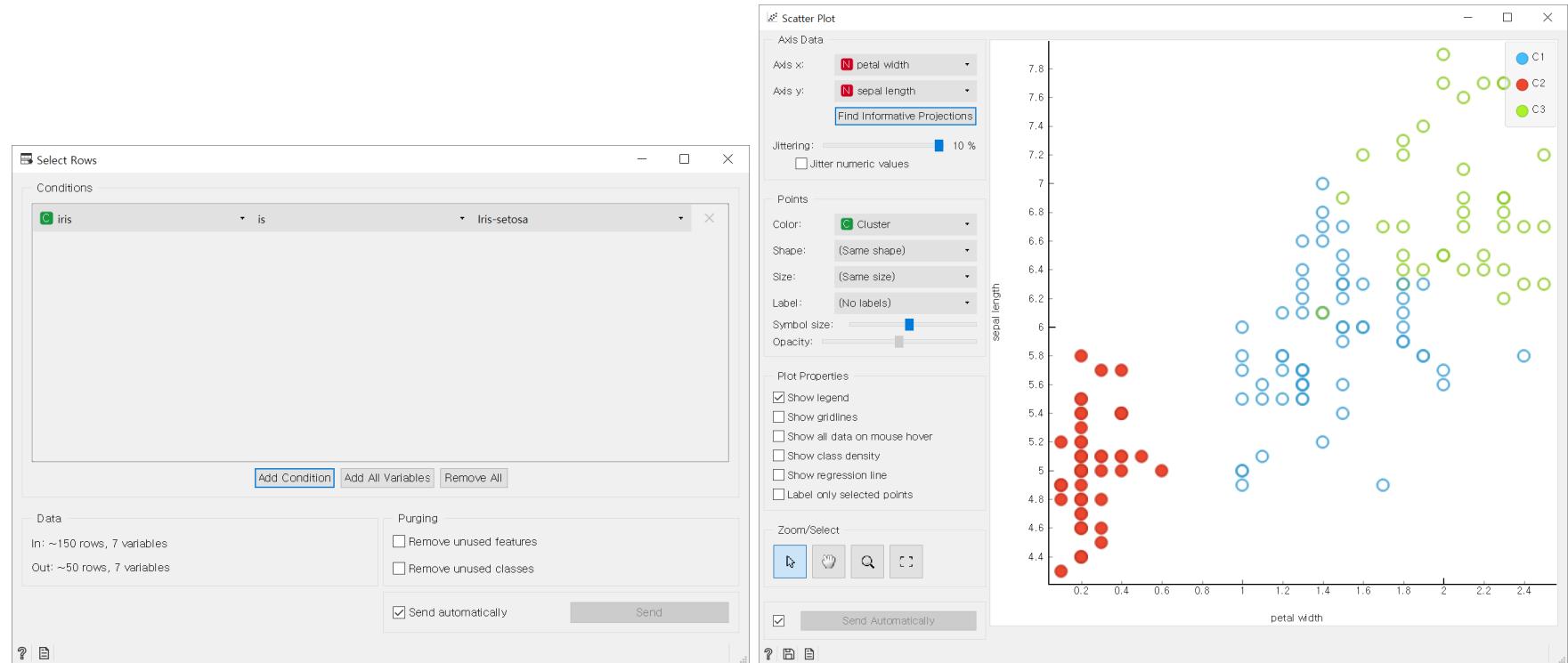
[2] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*. 20: 53–65.
[doi:10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).

[3] Paul E. Black, "Manhattan distance", in [Dictionary of Algorithms and Data Structures](#) [online], Vreda Pieterse and Paul E. Black, eds. 31 May 2006

k-means (Cont'd)

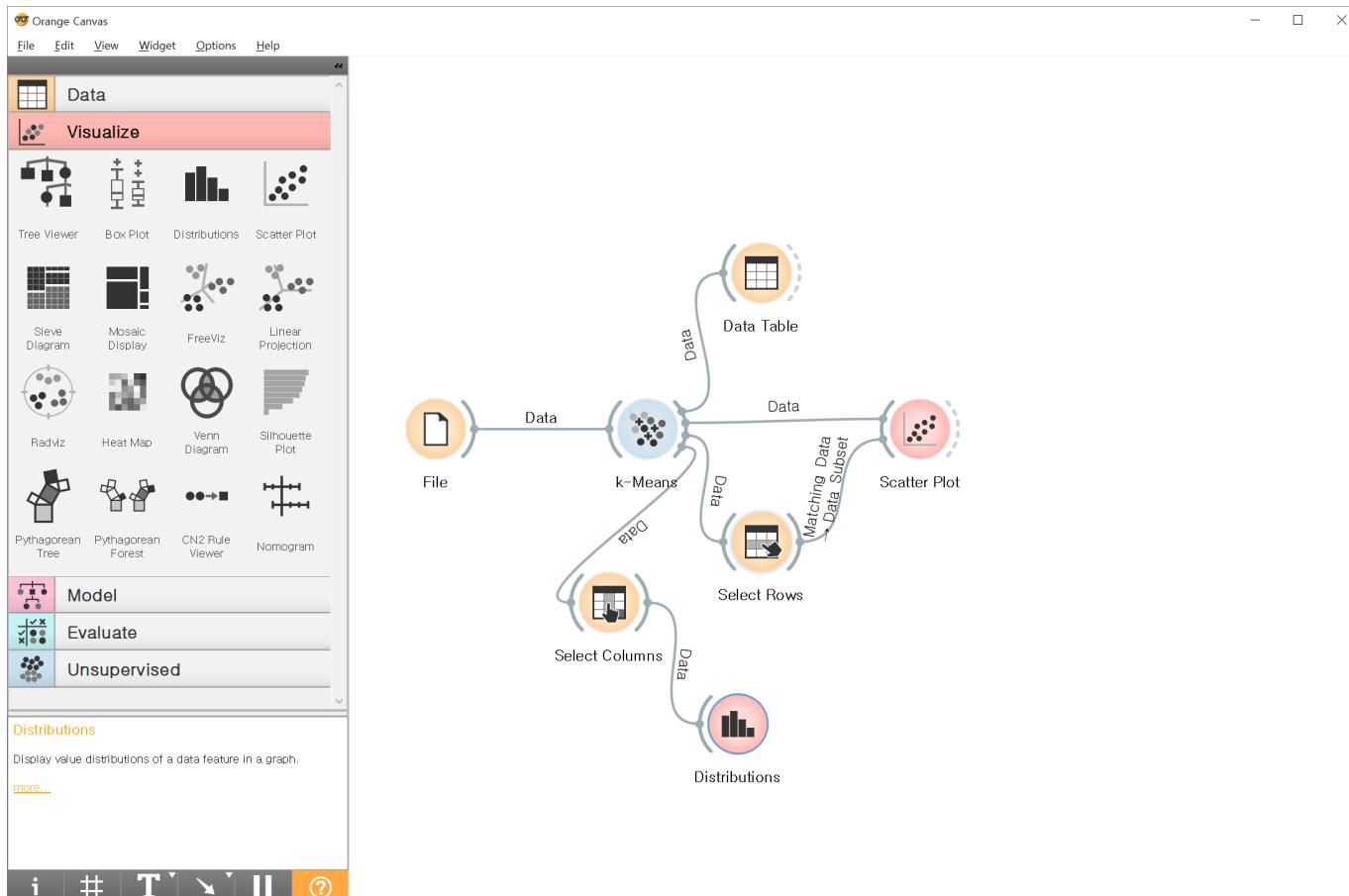
• Evaluation

- how well the clusters induced by the (unsupervised) clustering algorithm match the actual classes in the data
- select individual classes and have the corresponding points marked in the scatter plot



k-means (Cont'd)

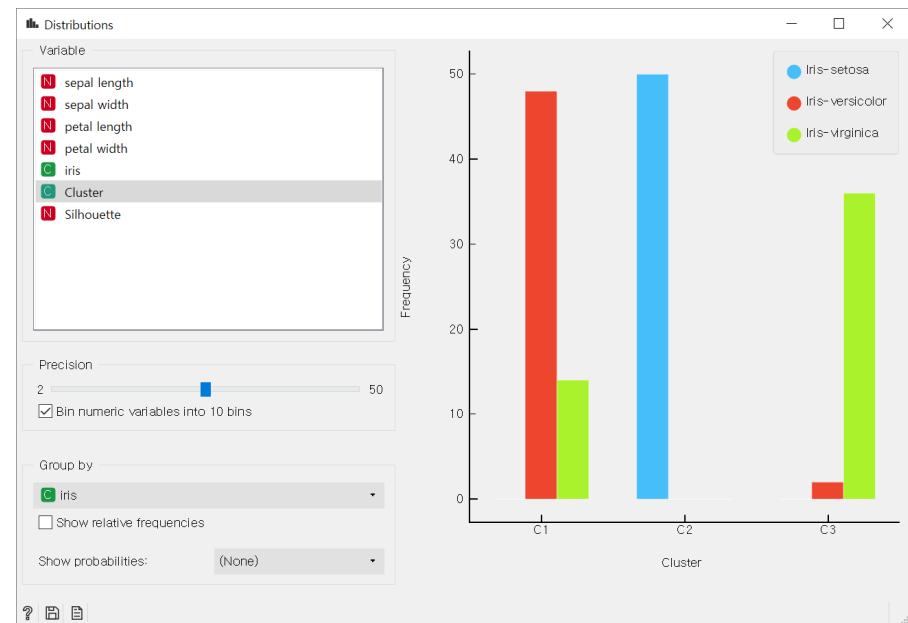
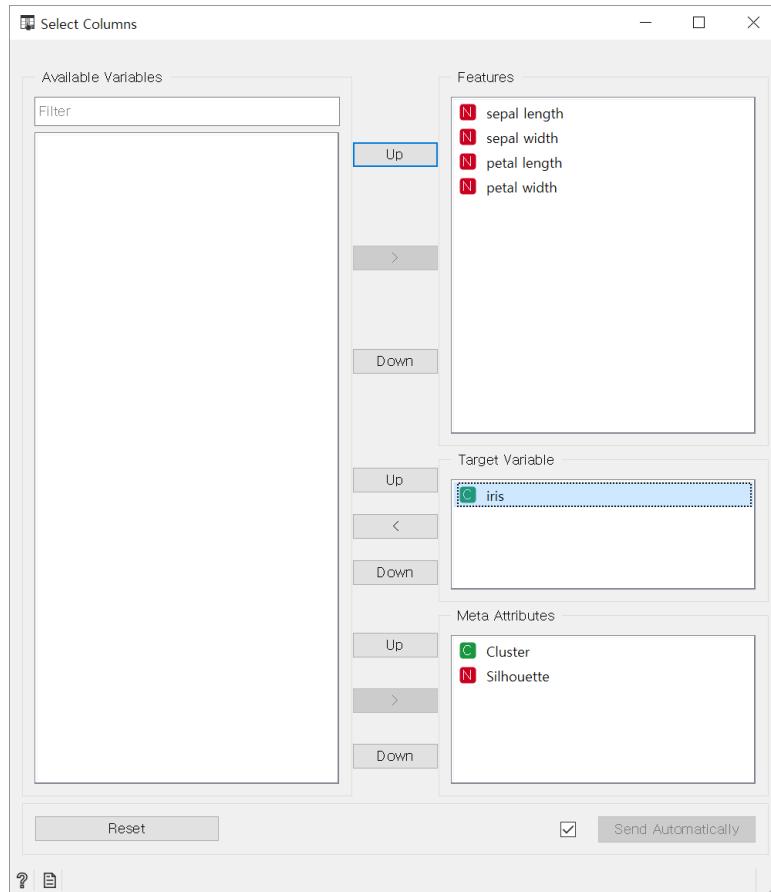
- **Distribution widget**
 - test the match between clusters and the original classes



k-means

- Select columns widget

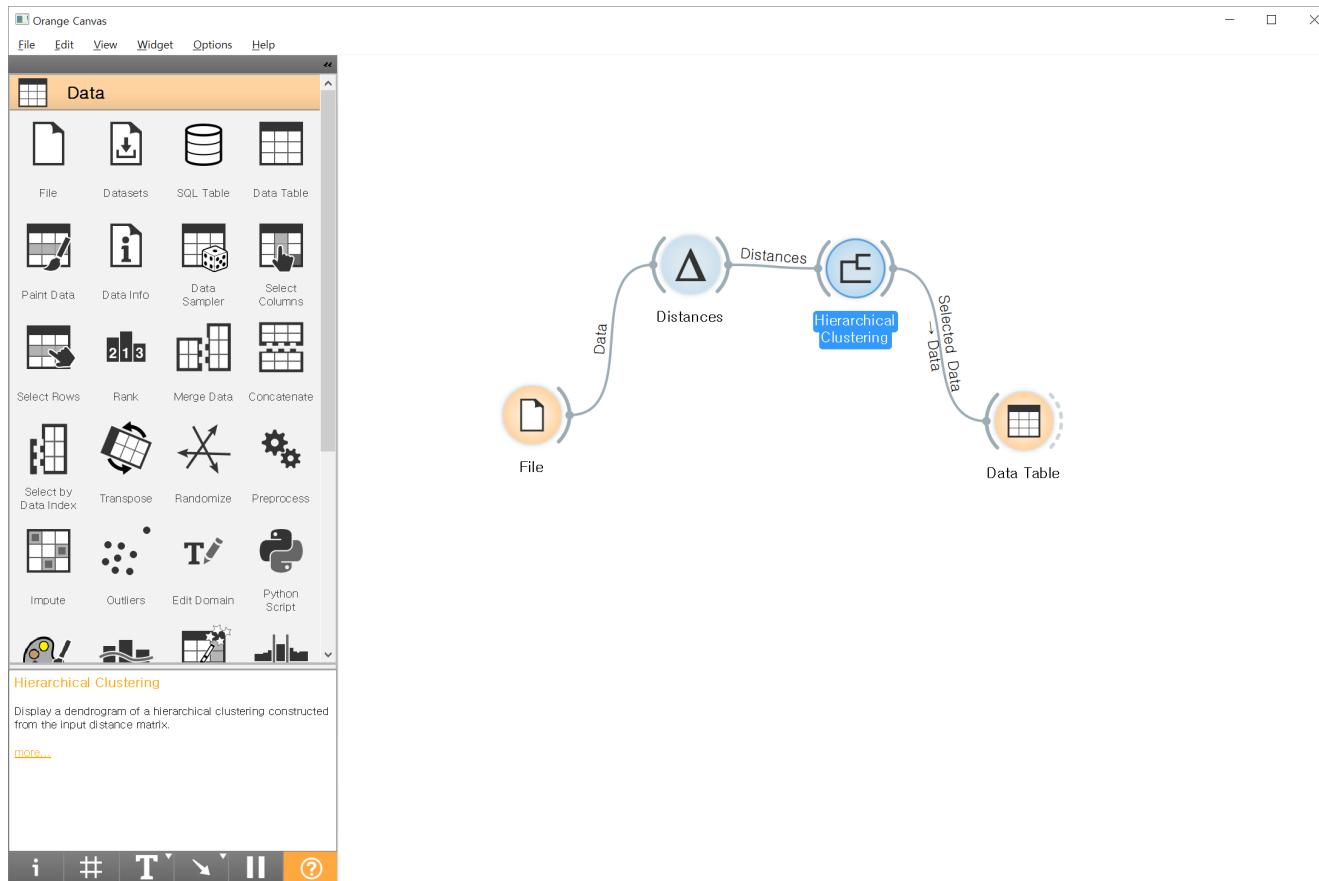
- reinstate the original class Iris as the class and put the cluster index among the attributes



Hierarchical Clustering (Cont'd)

- Build schema

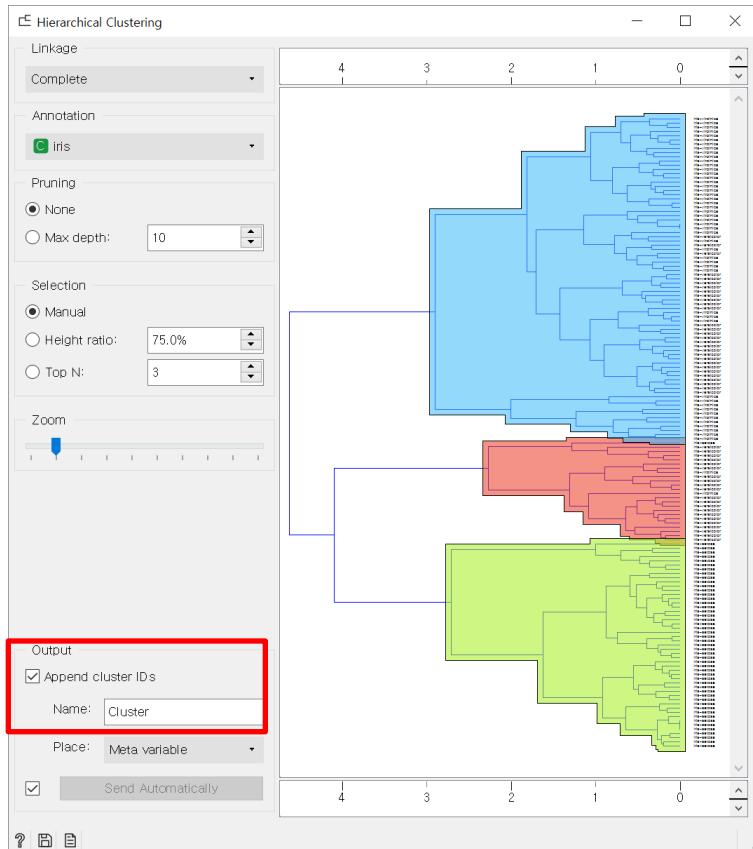
- skip load data (default is iris data)



Hierarchical Clustering (Cont'd)

- Observation of H-Clustering

- a way to check how hierarchical clustering clustered individual instance



The screenshot shows the Data Table node in KNIME displaying the Iris dataset. The table includes the following columns:

iris	Cluster	sepal length	sepal width	petal length	petal width
34	Iris-setosa	5.5	4.2	1.4	0.2
35	Iris-setosa	4.9	3.1	1.5	0.1
36	Iris-setosa	5.0	3.2	1.2	0.2
37	Iris-setosa	5.5	3.5	1.3	0.2
38	Iris-setosa	4.9	3.1	1.5	0.1
39	Iris-setosa	4.4	3.0	1.3	0.2
40	Iris-setosa	5.1	3.4	1.5	0.2
41	Iris-setosa	5.0	3.5	1.3	0.3
42	Iris-setosa	4.5	2.3	1.3	0.3
43	Iris-setosa	4.4	3.2	1.3	0.2
44	Iris-setosa	5.0	3.5	1.6	0.6
45	Iris-setosa	5.1	3.8	1.9	0.4
46	Iris-setosa	4.8	3.0	1.4	0.3
47	Iris-setosa	5.1	3.8	1.6	0.2
48	Iris-setosa	4.6	3.2	1.4	0.2
49	Iris-setosa	5.3	3.7	1.5	0.2
50	Iris-setosa	5.0	3.3	1.4	0.2
51	Iris-versicolor	7.0	3.2	4.7	1.4
52	Iris-versicolor	6.4	3.2	4.5	1.5
53	Iris-versicolor	6.9	3.1	4.9	1.5
54	Iris-versicolor	5.5	2.3	4.0	1.3
55	Iris-versicolor	6.5	2.8	4.6	1.5
56	Iris-versicolor	5.7	2.8	4.5	1.3
57	Iris-versicolor	6.3	3.3	4.7	1.6
58	Iris-versicolor	4.9	2.4	3.3	1.0
59	Iris-versicolor	6.6	2.9	4.6	1.3
60	Iris-versicolor	5.2	2.7	3.9	1.4
61	Iris-versicolor	5.0	2.0	3.5	1.0

Hierarchical Clustering (Cont'd)

- Hyperparameters and configuration

- Linkage

- Single : computes the distance between the closest elements of the two clusters
 - Average : computes the average distance between elements of the two clusters
 - Weighted : uses the WPGMA* method
 - Complete : computes the distance between the clusters' most distant elements

- Annotation

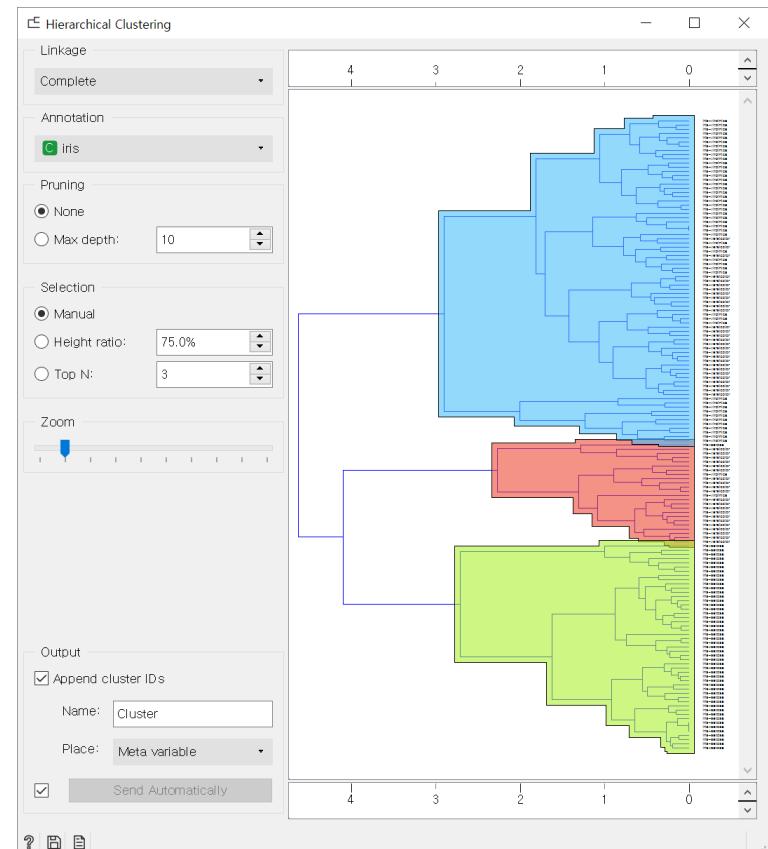
- Labels of nodes in the dendrogram

- Pruning

- Huge dendograms can be pruned in the Pruning box by selecting the maximum depth of the dendrogram
 - This only affects the display, not the actual clustering

- Selection methods

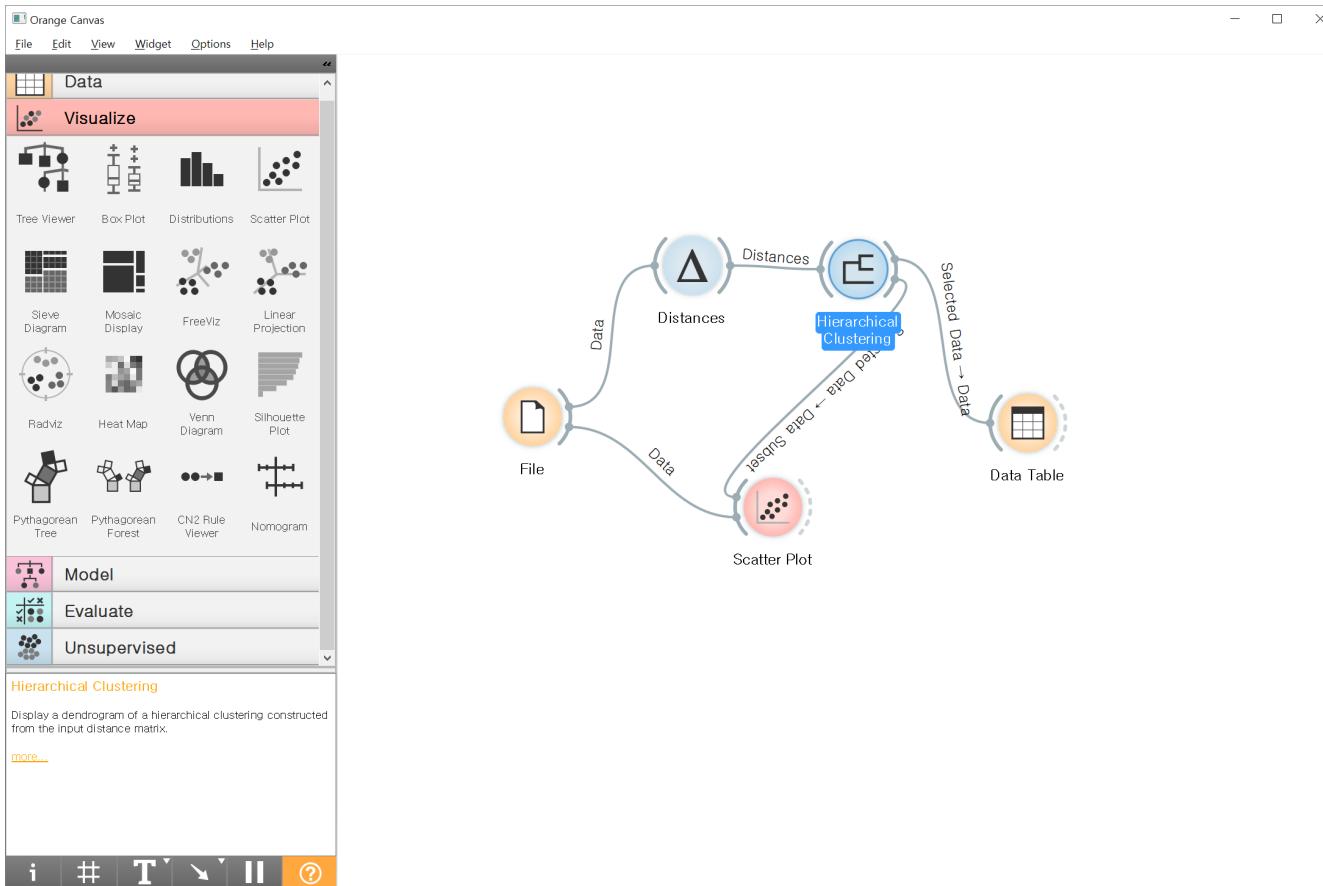
- Manual : Clicking inside the dendrogram will select a cluster
 - Height ratio : Clicking on the bottom or top ruler of the dendrogram places a cutoff line in the graph
 - Top N : Selects the number of top nodes



*WPGMA : Weighted Pair Group Method with Arithmetic Mean. It is a simple agglomerative (bottom-up) hierarchical clustering method, generally attributed to Sokal and Michene

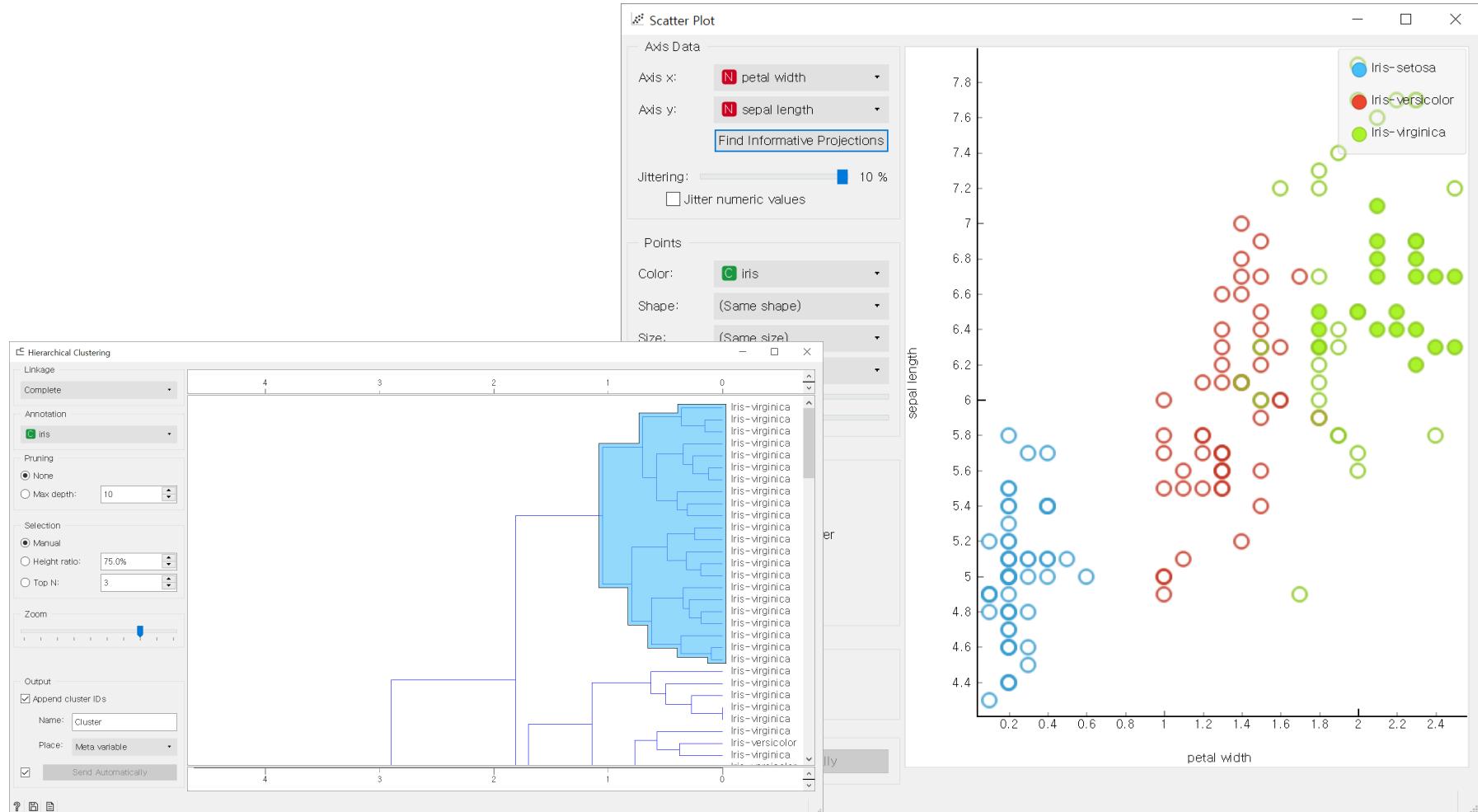
Hierarchical Clustering (Cont'd)

- Build schema for another observation



Hierarchical Clustering

- Observation of the position of the selected clusters in the projection



Self-organizing map Preparation

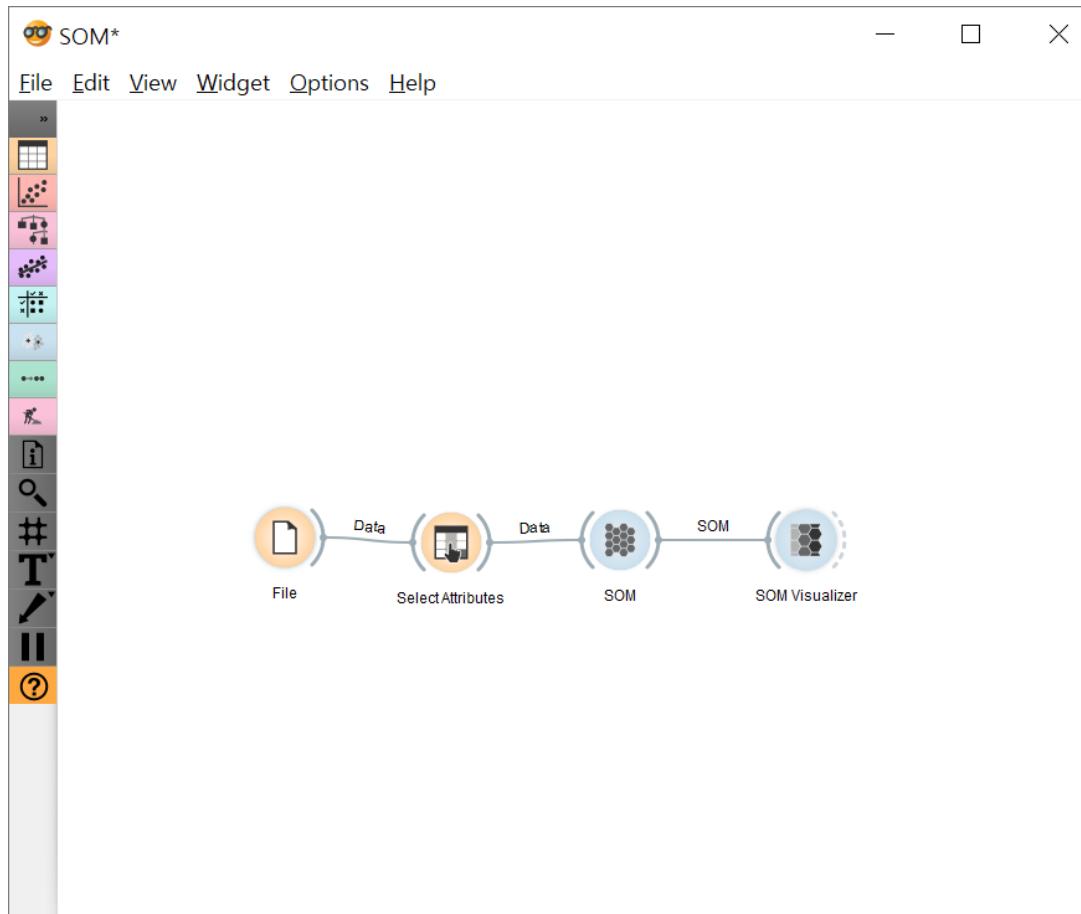
- Install Orange 2.7 for SOM
 - <https://orange.biolab.si/orange2/>

The screenshot shows a web browser window with multiple tabs open. The active tab is for the Orange 2.7 download page at <https://orange.biolab.si/orange2/>. The page features the Orange logo and navigation links for Home, Screenshots, Download, Docs, Blog, Training, and Donate. A message at the top states: "Orange 2.7 is a legacy version of Orange. We think you should upgrade to the [current version](#), but if you still need the version 2.7, please download it below." Below this message, there is a red box highlighting the link "Windows: Orange 2.7 installer for Windows". Other links for Mac OS and Other systems are also present. Further down the page, there is a "Documentation" section with a link to "Read documentation for Orange 2.7 [here](#)".

The screenshot shows the footer of the Orange website. It includes links for Orange License, Download Windows, Community Stack Exchange, FAQ Documentation, Developers GitHub, and a "Latest blog posts" section. The footer also contains links for Privacy Policy, Citation, Contact, and social media icons for Facebook, YouTube, and @OrangeDataMiner. At the bottom, it says "Copyright © University of Ljubljana".

Self-organizing map (Cont'd)

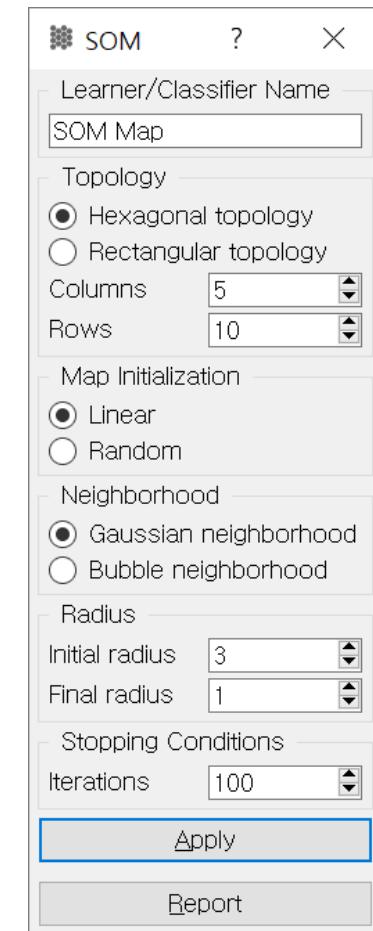
- Build schema (Orange 2.7)
 - Default file is *iris.tab*



Self-organizing map (Cont'd)

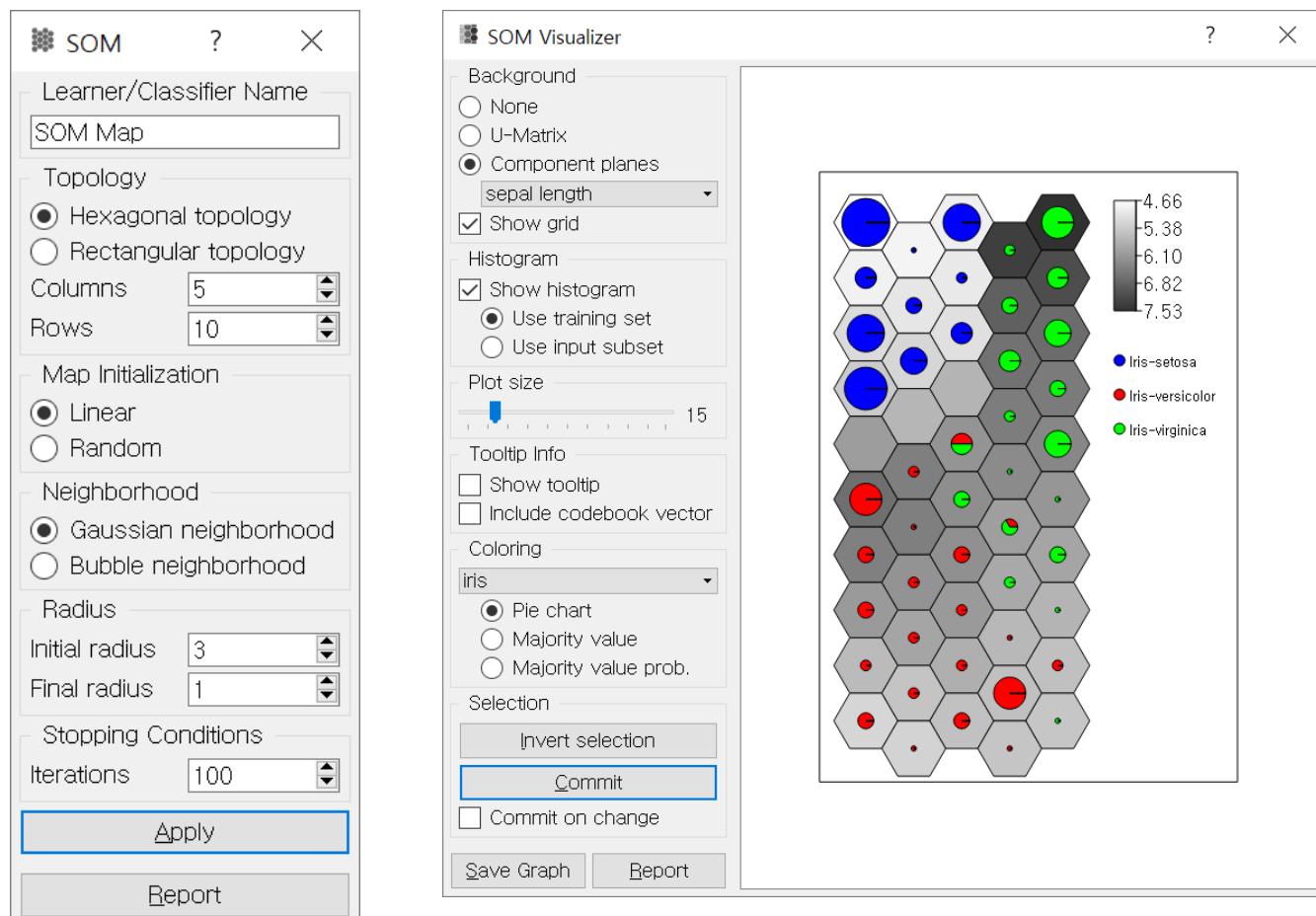
- **Hyperparameters and configuration**

- Topology : topology type id and map sizes
 - Hexagonal : cells are hexagon-shaped
 - Rectangular : cells are square-shaped
- Map initialization
 - Linear : Data instances are initially assigned to cells according to their two-dimensional PCA projection
 - Random : Data instances are initially randomly assigned to cells
- Neighborhood
 - Gaussian : smoothed neighborhood
 - Bubble : crisp neighborhood
- Radius : initial and final radius
- Iterations : iterations of a training steps



Self-organizing map

- Observation from SOM to SOM Visualizer



Practice with Realdta

Dataset

- **Realworld data**

- Online Retail Data Set^[1]

Data Set Characteristics:	Multivariate, Sequential, Time-Series	Number of Instances:	541909	Area:	Business
Attribute Characteristics:	Integer, Real	Number of Attributes:	8	Date Donated	2015-11-06
Associated Tasks:	Classification, Clustering	Missing Values?	N/A	Number of Web Hits:	231361

- a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.
- <https://archive.ics.uci.edu/ml/datasets/Online+Retail#>

- Attribute Information

Attributes	Description
InvoiceNo	Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product
Description	Product (item) name. Nominal.
Quantity	The quantities of each product (item) per transaction. Numeric.
InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice	Unit price. Numeric, Product price per unit in sterling.
CustomerID	Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country	Country name. Nominal, the name of the country where each customer resides.

^[1] Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK

Overview

- **Objective**

- to identify high- and low-value customers for marketing purposes

- **Feature selection**

- each customer's recency of last purchase
 - frequency of purchase
 - monetary value

- **Preprocessing**

- Cleansing and generate features from raw data

- **Develop a model**

- the k-means clustering technique

Preprocessing

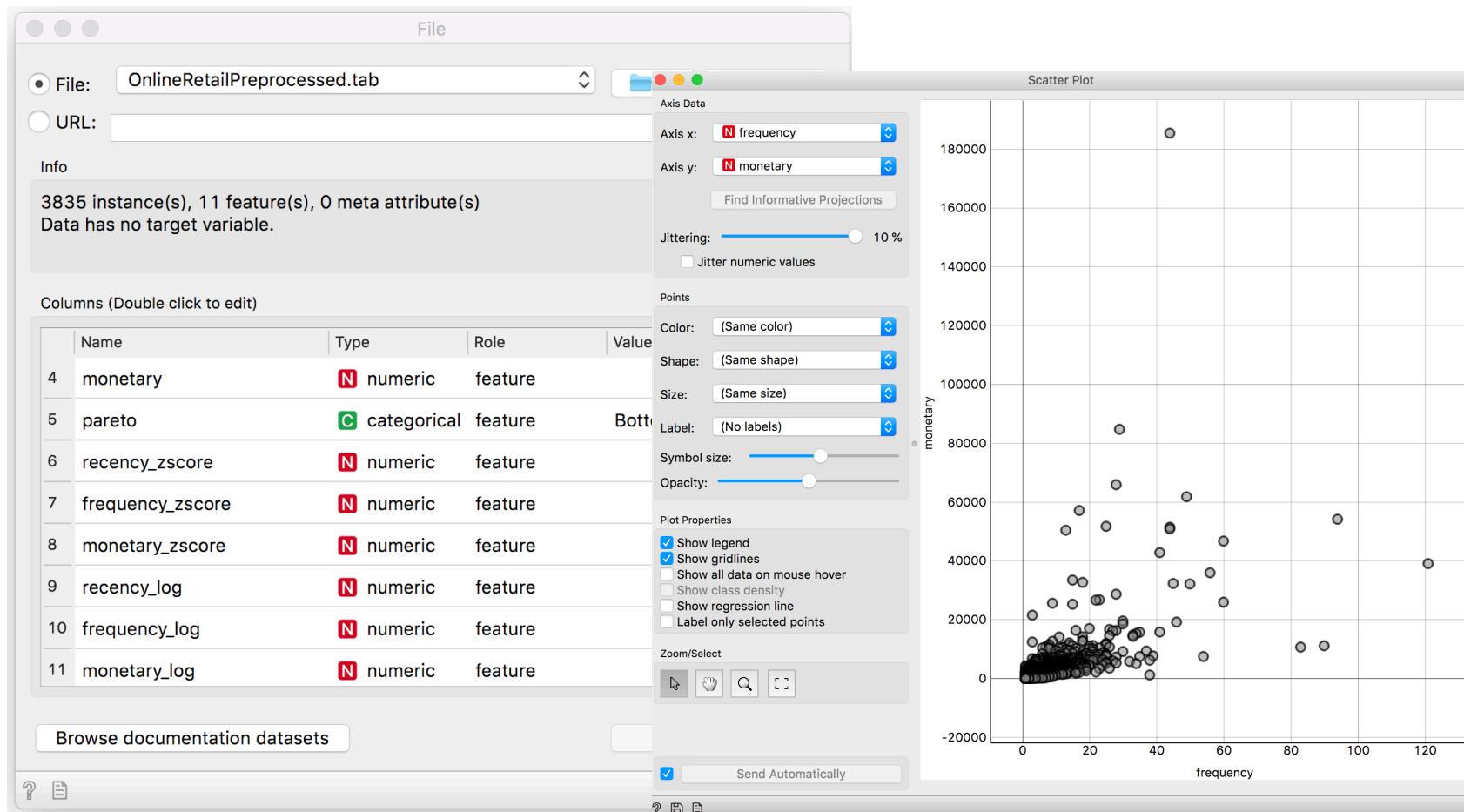
- Python code
 - <https://www.dropbox.com/s/kci0p07jcqkjzj4/PreprocessingForClustering.py?dl=0>
- Overall process of preprocessing



- delete null values
- Restrict country and date
- Frequency
- Recency
- Monetary
- log transform
- z-score

Load data

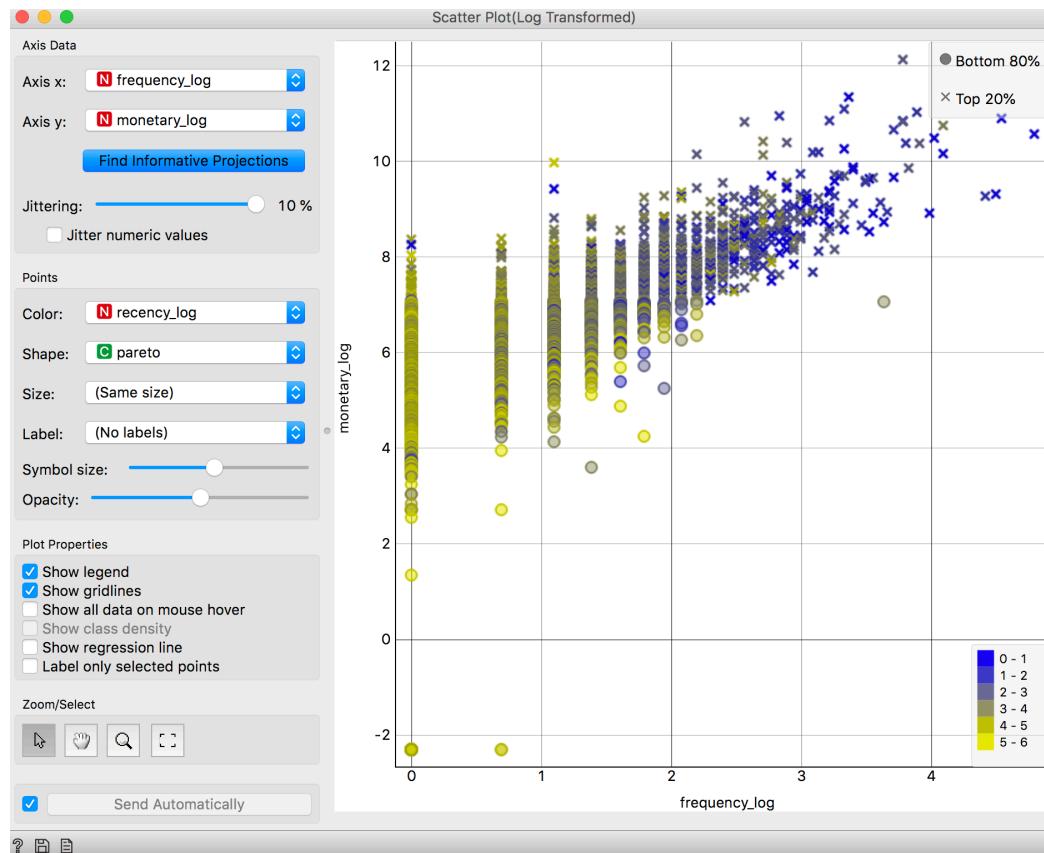
- Load preprocessed data
 - Analysis data for getting insights



Visualize data

- **Analysis data**

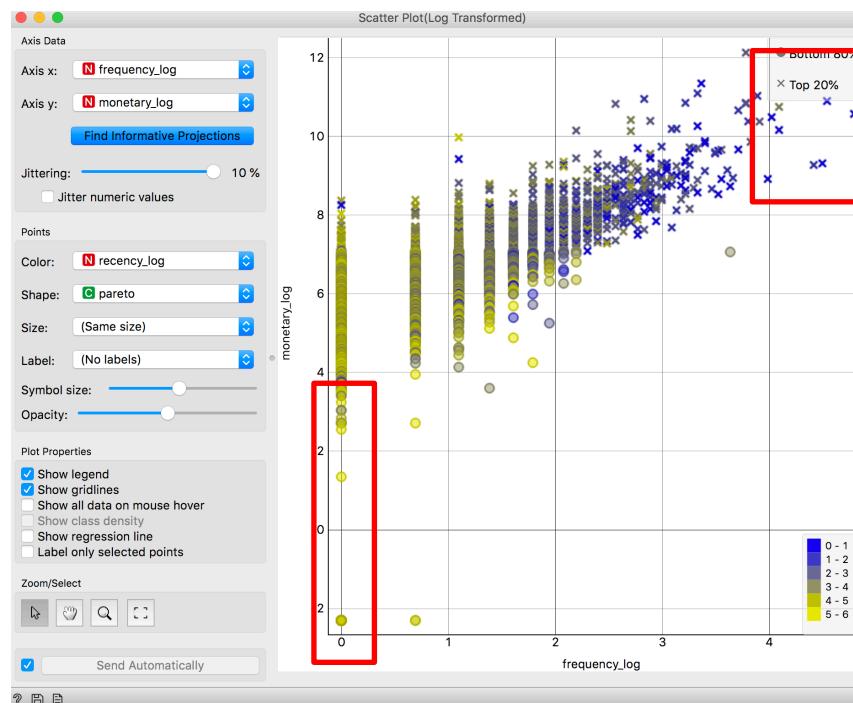
- a scattering of high-value, high-frequency customers in the top, right-hand corner of the graph
- data points are dark, indicating that they've purchased something recently



Handling outliers

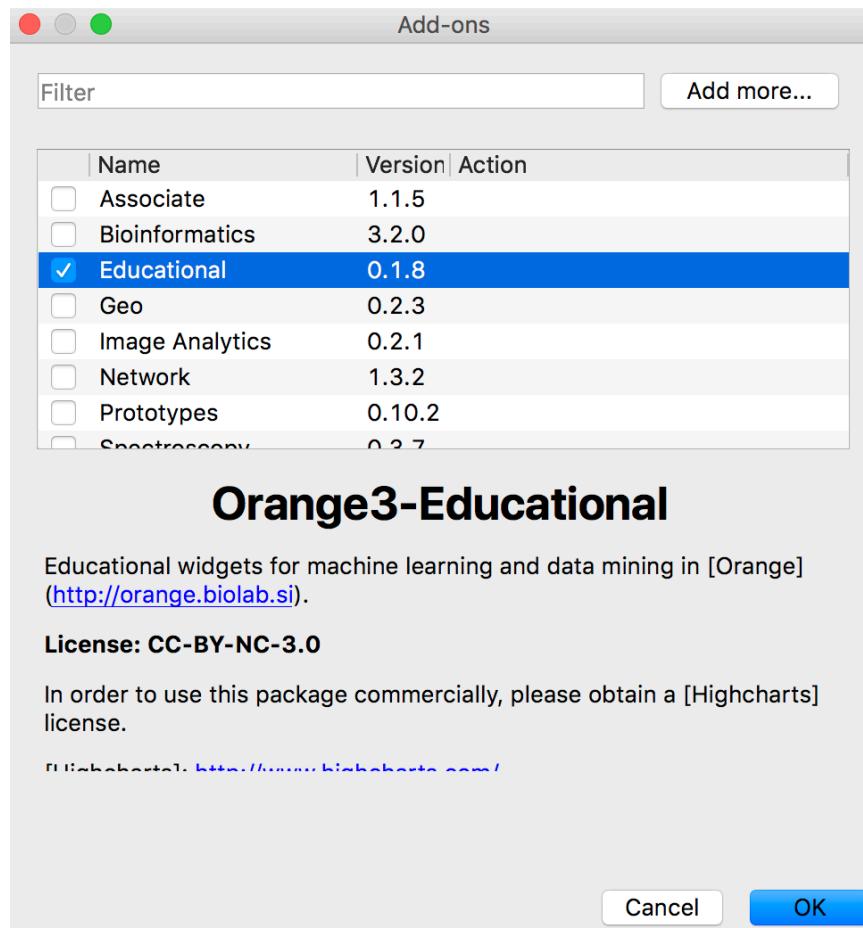
- **Analysis outliers**

- The eighteen, no-value customers we just investigated are all customers who returned every thing they bought
- the outliers may be the most important customers to understand
- right-hand corner, we have customers who are outliers in terms of being extraordinarily hig h-value, high-frequency shoppers
 - they're the customers we most want



Install add-on

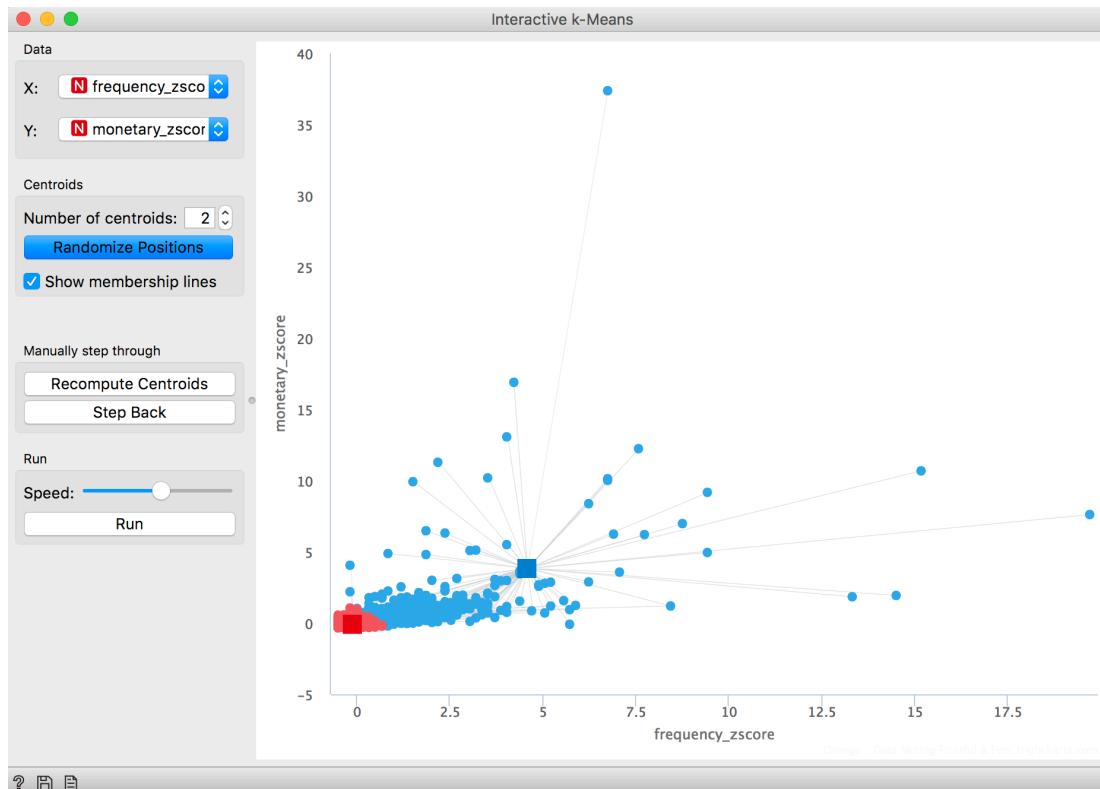
- Educational add-on
 - interactive k-mean



K-means (Cont'd)

- **Determine number of clusters / run k-means**

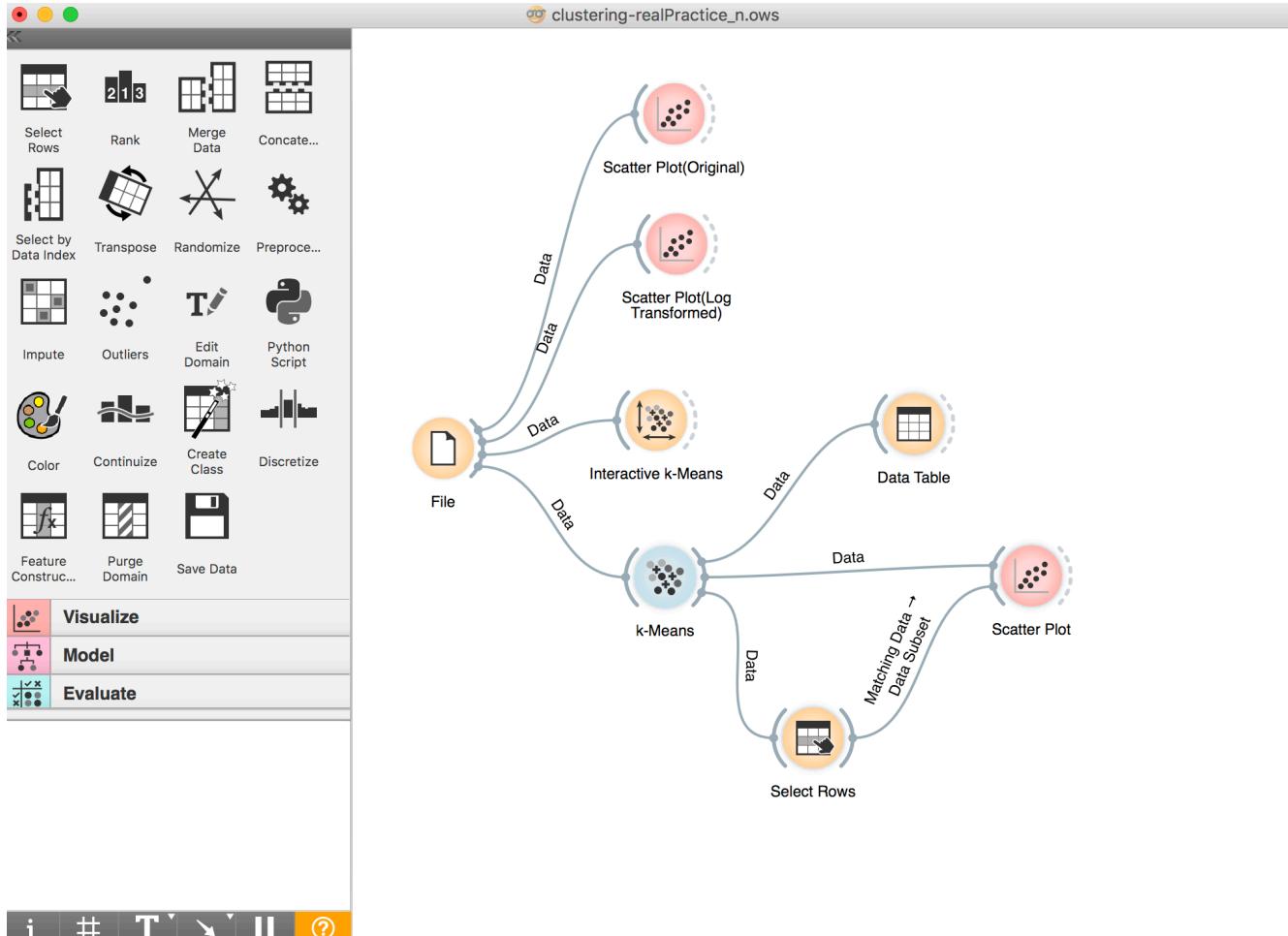
- the data didn't suggest an obvious choice for the number of clusters
- to determine how many clusters to extract (Judgement call)
- identifying high- and low-value customers for the business
- i.e. k=2 to 10



K-mean (Cont'd)

- K-mean Analysis

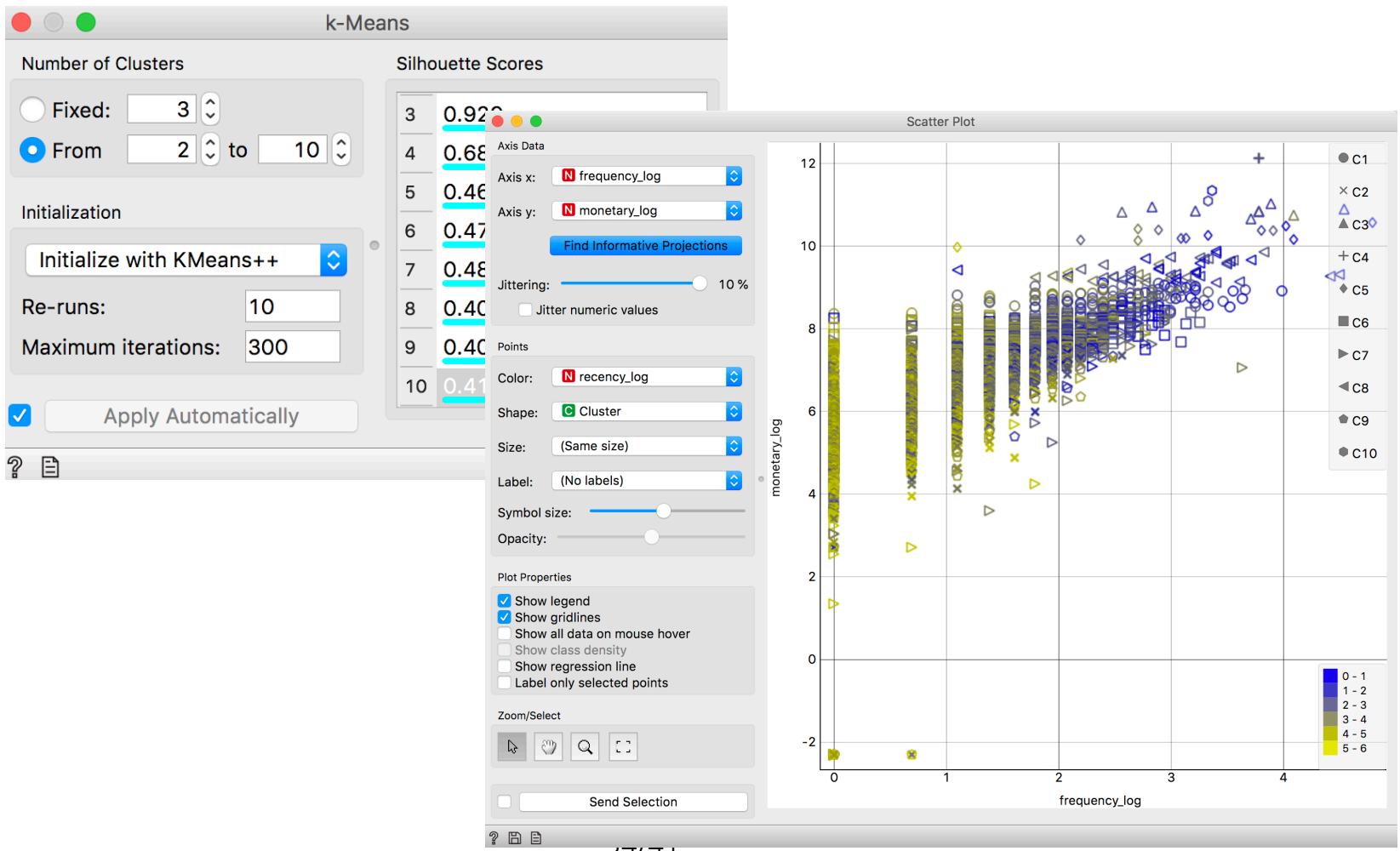
- Customer segmentation



K-mean (Cont'd)

- K-mean Analysis

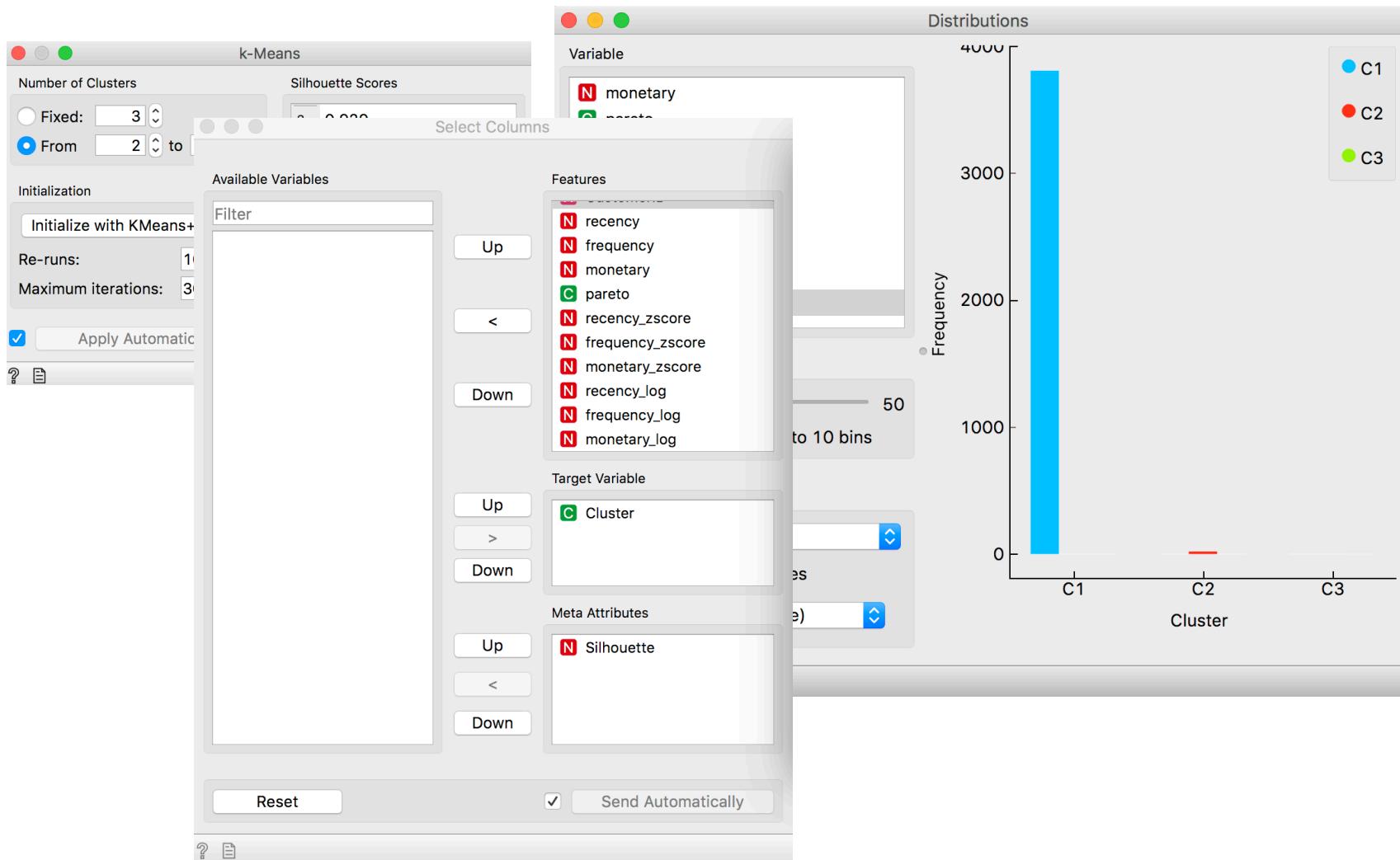
- Customer segmentation



K-mean (Cont'd)

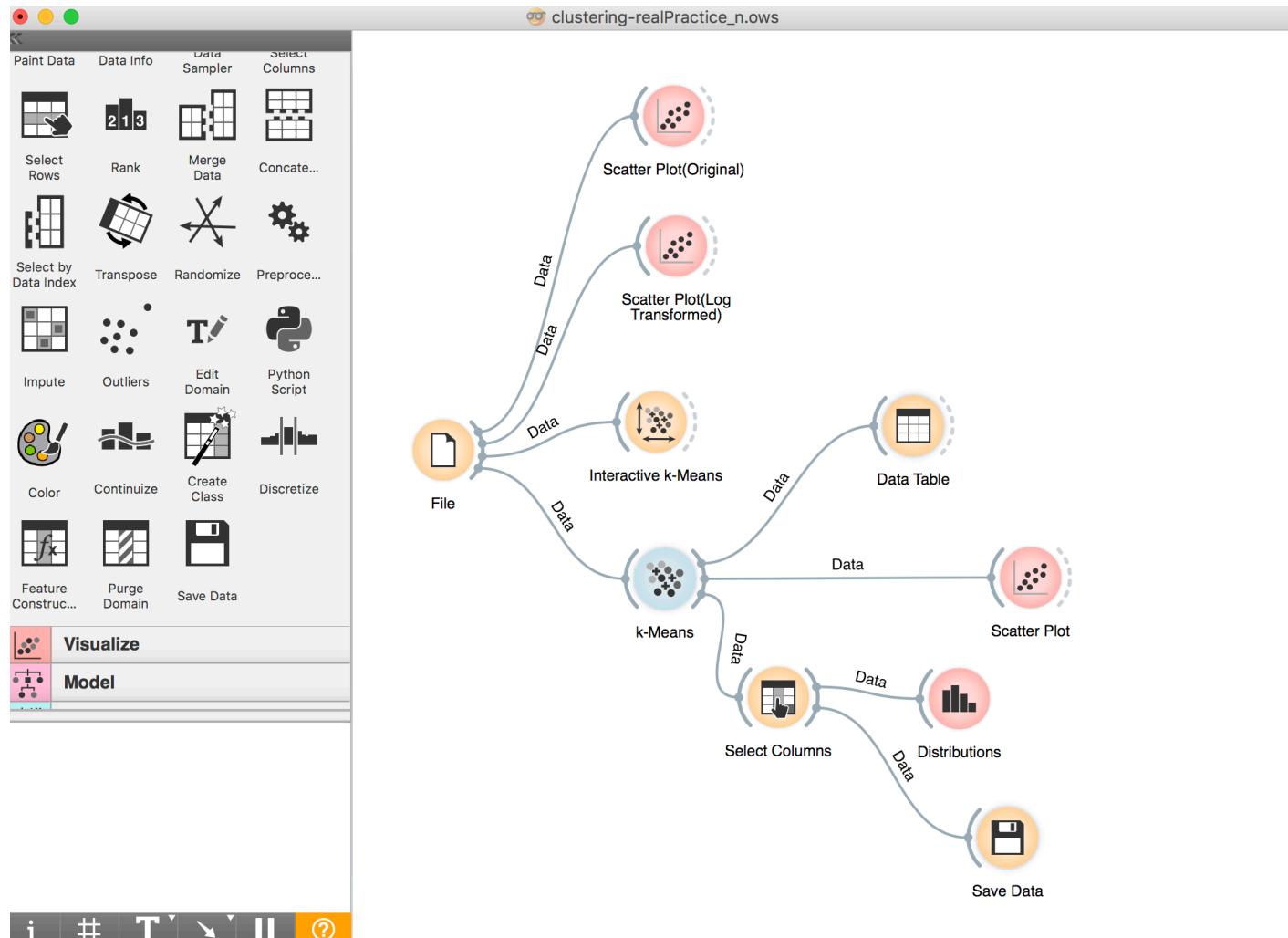
- K-mean Analysis

- Customer segmentation



K-mean (Cont'd)

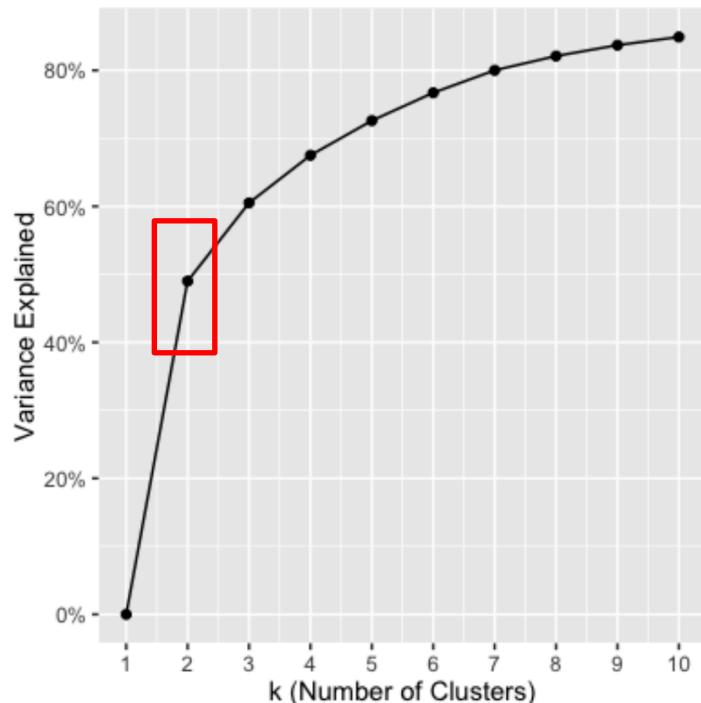
- Save result then further analysis



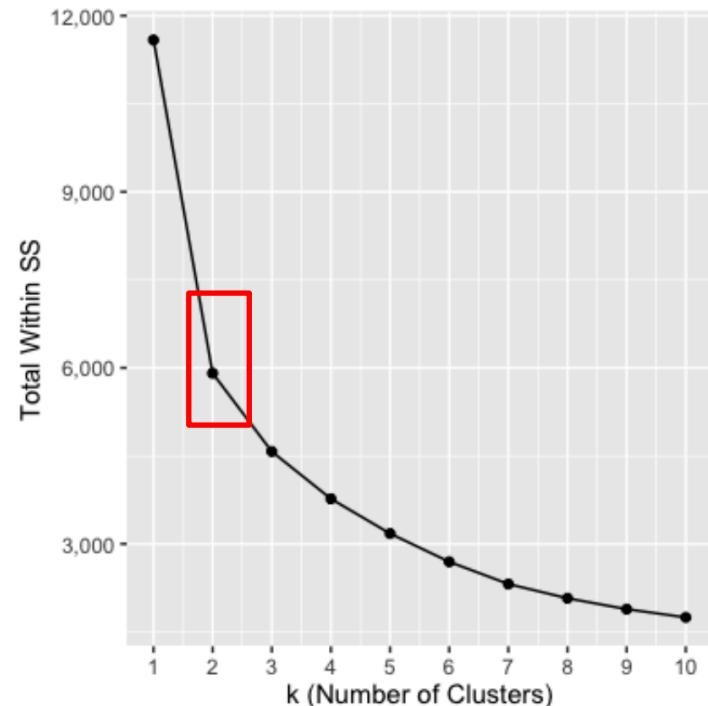
K-mean

- **Find best K**

- The decision should be based upon how the business plans to use the results, and the level of granularity they want to see in the clusters.



Graph variance explained by number of clusters



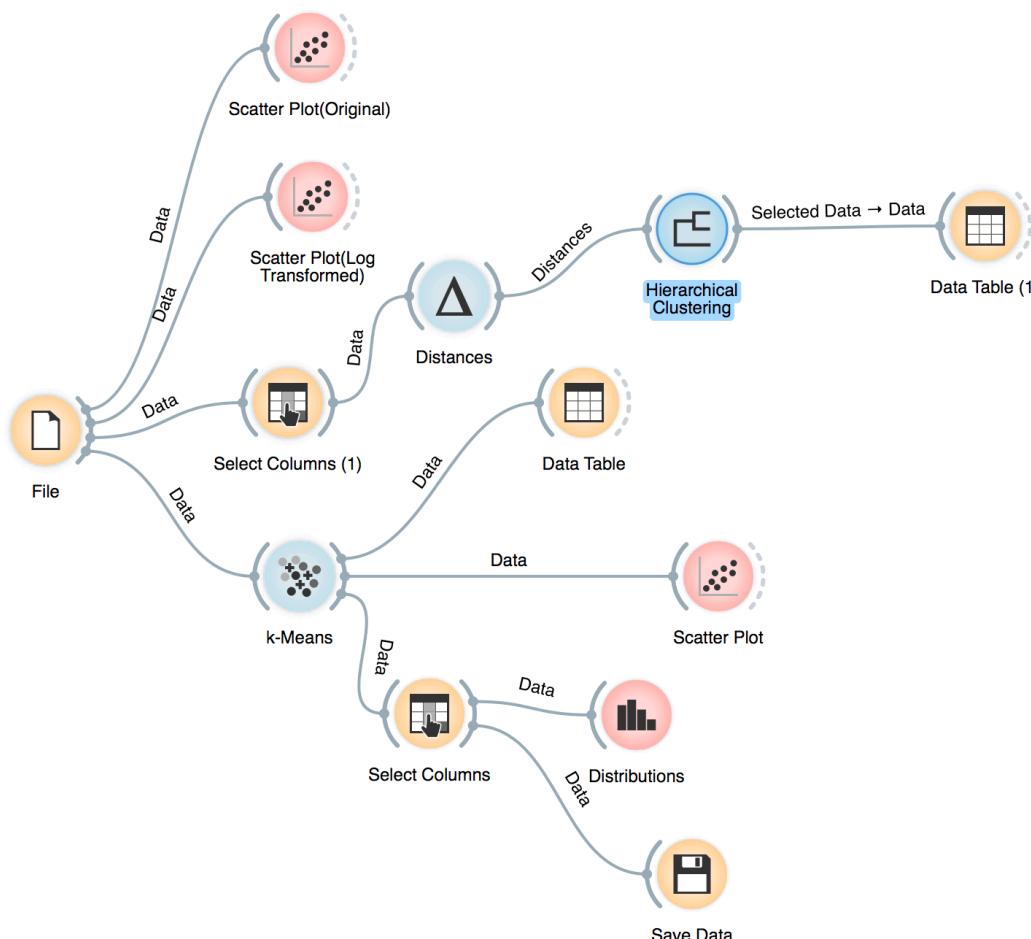
Graph within sums of squares by number of clusters

H-Clustering (Cont'd)

- Do same procedure By H-clustering algorithm
 - Time limit : 15min

H-Clustering

- Schema



References

- <https://docs.orange.biolab.si/3/visual-programming/widgets/unsupervised/kmeansclustering.html>
- <https://docs.orange.biolab.si/3/visual-programming/widgets/unsupervised/hierarchicalclustering.html>
- <https://docs.orange.biolab.si/2/reference/rst/Orange.projection.som.html>
- <http://www.algorithmsinnature.org/>
- <http://www.kimberlycoffey.com/blog/2016/8/k-means-clustering-for-customer-segmentation>

Thank you

