

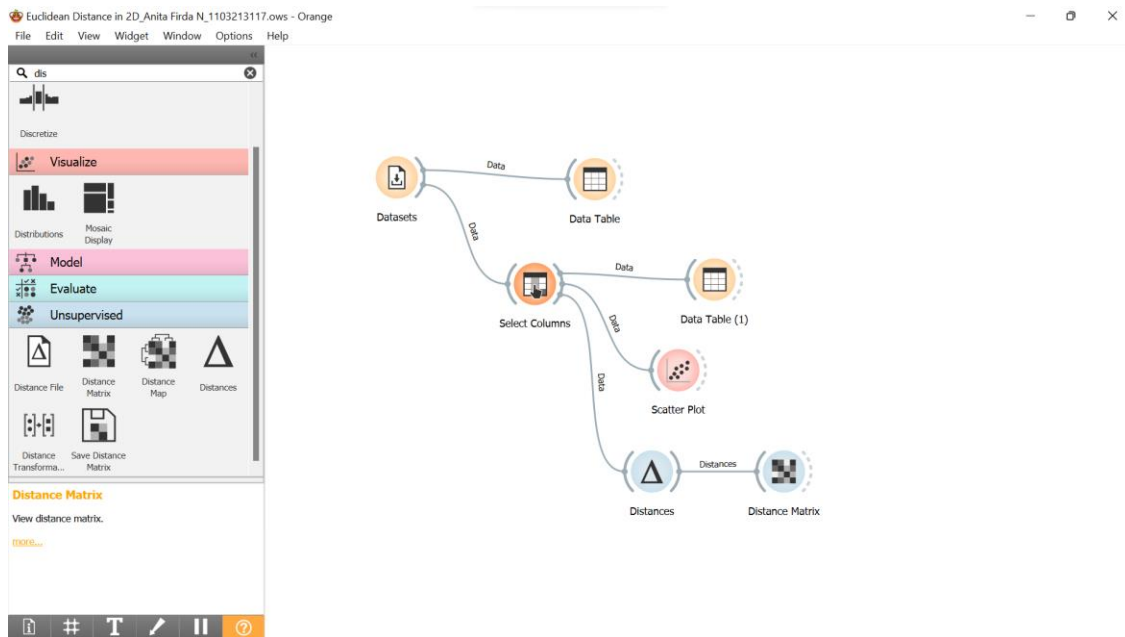
Machine Learning Week 6

Nama : Anita Firda Nuralifah

NIM : 1103213117

Euclidean Distance in 2D

Tampilan menu awal



Dataset

Memasukkan dataset bawaan yang sudah tersedia di orange, yaitu dataset course grades.

The screenshot shows the "Datasets - Orange" window. The search filter is "grad". The table below lists the datasets found:

Title	Size	Instances	Variables	Target	Tags
Course Grades	9.2 KB	16	8	none	synthetic, education
Grades for English and Math	265 bytes	12	3	none	synthetic, educational

The description for the "Course Grades" dataset is as follows:

Course Grades

A small dataset with grades on seven courses (English, French, History, Algebra, Biology, Physics, Physical) that was handcrafted to introduce hierarchical clustering.

The bottom status bar shows the number of instances: 16.

Data Table

Memperlihatkan isi dataset yang sedang dipakai.

Info

16 instances (no missing data)
7 features
No target variable.
1 meta attribute

Variables

☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

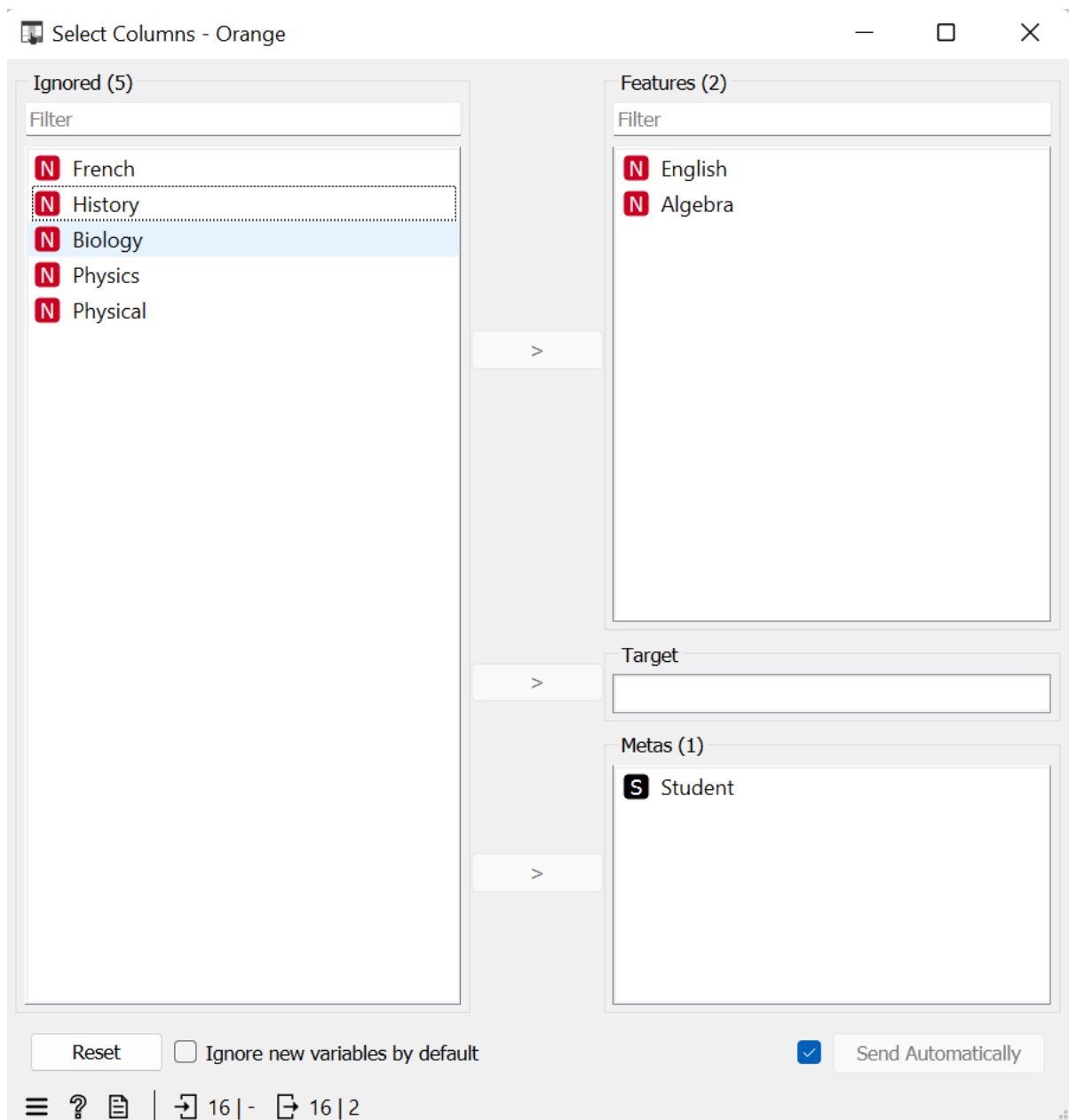
☒ Send Automatically

	Student	English	French	History	Algebra	Biology
1	Ana	22	30	32	21	
2	Bill	91	95	65	89	
3	Cynthia	51	89	21	100	
4	Demi	9	15	18	61	1
5	Eve	93	99	39	12	
6	Fred	49	17	17	92	
7	George	91	99	97	49	
8	Henry	12	30	32	34	
9	Ian	91	80	20	82	
10	Jena	39	18	19	99	
11	Katherine	20	50	10	71	
12	Lea	90	100	45	45	
13	Maya	100	98	97	32	
14	Nash	14	4	15	61	
15	Olga	9	22	8	100	
16	Phill	85	90	100	45	

16 | 16

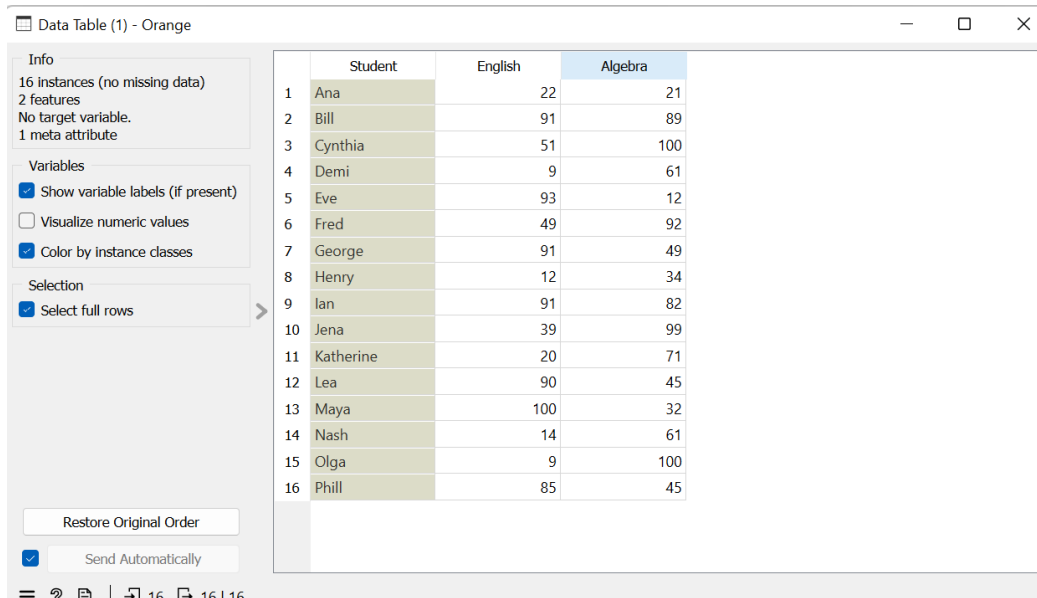
Select columns

Dengan menggunakan select columns kita bisa memilih kolom apa yang ingin kita lihat pada is a dan bisa kita pisahkan jika hanya ingin melihat English dan Algebra.



Data table

Seperti yang sudah kita lakukan sebelumnya, jika kita cek di data table, table tersebut hanya akan menampilkan kolom yang ingin kita lihat saja karena sebelumnya kolom lainnya sudah kita pisahkan.

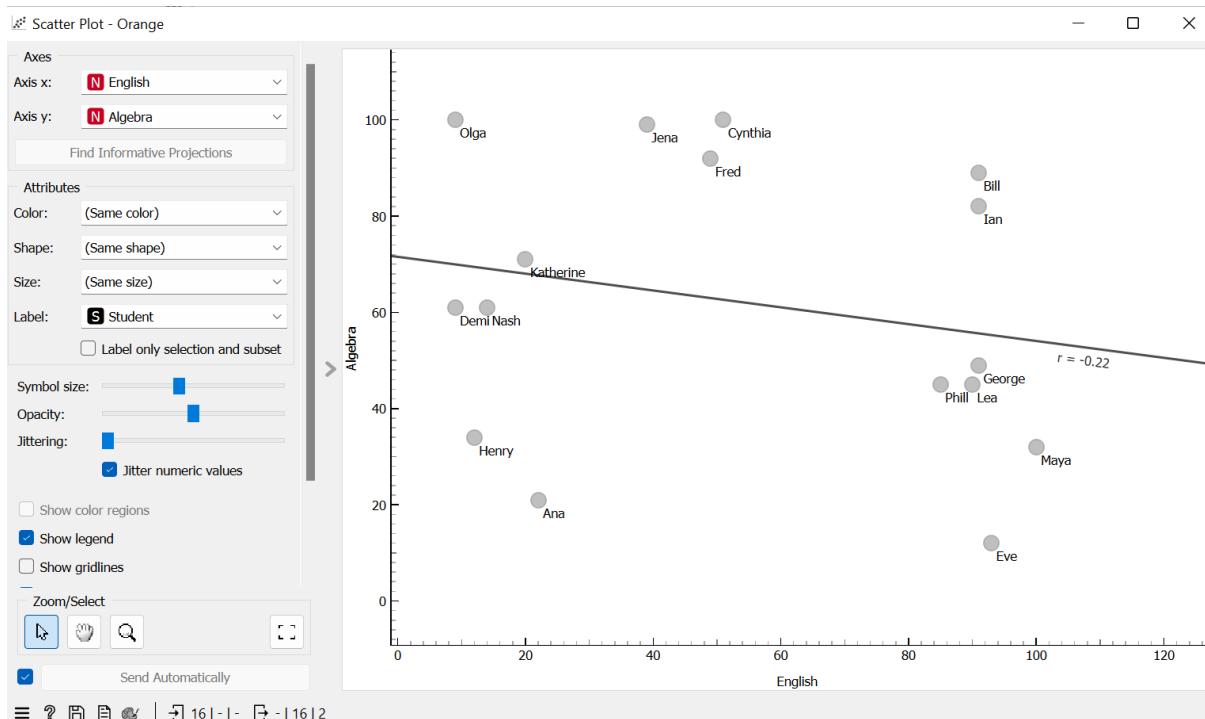


The screenshot shows the 'Data Table (1) - Orange' window. On the left, the 'Info' panel indicates 16 instances, 2 features, and 1 meta attribute. The 'Variables' panel has 'Show variable labels (if present)' checked. The 'Selection' panel has 'Select full rows' checked. The main table displays the following data:

	Student	English	Algebra
1	Ana	22	21
2	Bill	91	89
3	Cynthia	51	100
4	Demi	9	61
5	Eve	93	12
6	Fred	49	92
7	George	91	49
8	Henry	12	34
9	Ian	91	82
10	Jena	39	99
11	Katherine	20	71
12	Lea	90	45
13	Maya	100	32
14	Nash	14	61
15	Olga	9	100
16	Phill	85	45

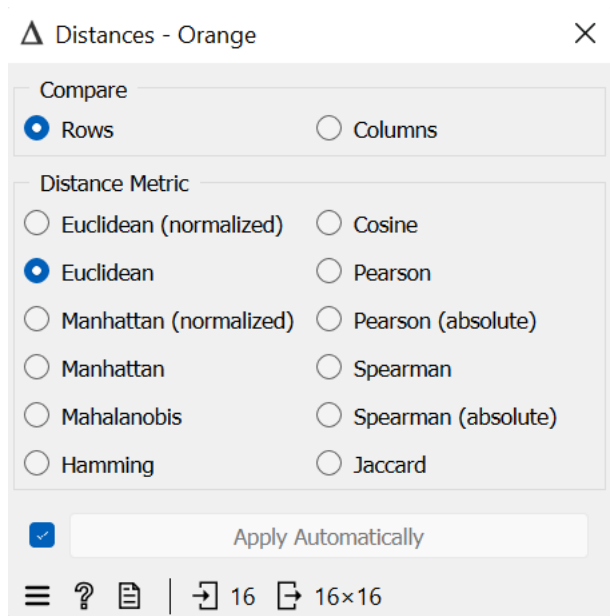
Scatter Plot

Scatter plot merupakan cara terbaik untuk melihat perkiraan nilai, karena saat ini kita hanya memiliki data dengan 2 kolom dimana, scatter plot ini diberi label nama murid pada titik titik tersebut.



Distances

Dengan menggunakan orange, untuk melihat jarak dari nilai para siswa tersebut, yaitu dengan menggunakan distance, dan distance metric Euclidean tanpa menggunakan normalized karena nilai English dan algebra dinyatakan dalam satuan yang sama maka normalized tidak diperlukan.



Distances matrix

Kita dapat melihat jarak nilai atau perbedaan nilai pada siswi tersebut dengan menggunakan distance matrix.

Distance Matrix - Orange

	Ana	Bill	Cynthia	Demi	Eve	Fred	George	Henry	Ian	Jena	Katherine	Lea	Maya	Nash
Ana		96,876	84,155	42,059	71,568	75,961	74,465	16,401	92,098	79,831	50,040	72,111	78,772	40,792
Bill	96,876		41,485	86,649	77,026	42,107	40,000	96,260	7,000	52,953	73,246	44,011	57,706	81,933
Cynthia	84,155	41,485		57,315	97,509	8,246	64,815	76,662	43,863	12,042	42,450	67,424	83,815	53,759
Demi	42,059	86,649	57,315		97,247	50,606	82,873	27,166	84,646	48,415	14,866	82,565	95,509	5,000
Eve	71,568	77,026	97,509	97,247		91,302	37,054	83,934	70,029	102,396	93,862	33,136	21,190	92,962
Fred	75,961	42,107	8,246	50,606	91,302		60,108	68,797	43,174	12,207	35,805	62,370	78,746	46,755
George	74,465	40,000	64,815	82,873	37,054	60,108		80,411	33,000	72,139	74,330	4,123	19,235	77,929
Henry	16,401	96,260	76,662	27,166	83,934	68,797	80,411		92,439	70,385	37,855	78,772	88,023	27,074
Ian	92,098	7,000	43,863	84,646	70,029	43,174	33,000	92,439		54,708	71,847	37,014	50,804	79,812
Jena	79,831	52,953	12,042	48,415	102,396	12,207	72,139	70,385	54,708		33,838	74,277	90,609	45,486
Katherine	50,040	73,246	42,450	14,866	93,862	35,805	74,330	37,855	71,847	33,838		74,673	89,000	11,662
Lea	72,111	44,011	67,424	82,565	33,136	62,370	4,123	78,772	37,014	74,277	74,673		16,401	77,666
Maya	78,772	57,706	83,815	95,509	21,190	78,746	19,235	88,023	50,804	90,609	89,000	16,401		90,758
Nash	40,792	81,933	53,759	5,000	92,962	46,755	77,929	27,074	79,812	45,486	11,662	77,666	90,758	
Olga	80,062	82,735	42,000	39,000	121,655	40,792	96,566	66,068	83,952	30,017	31,016	97,908	113,600	39,319

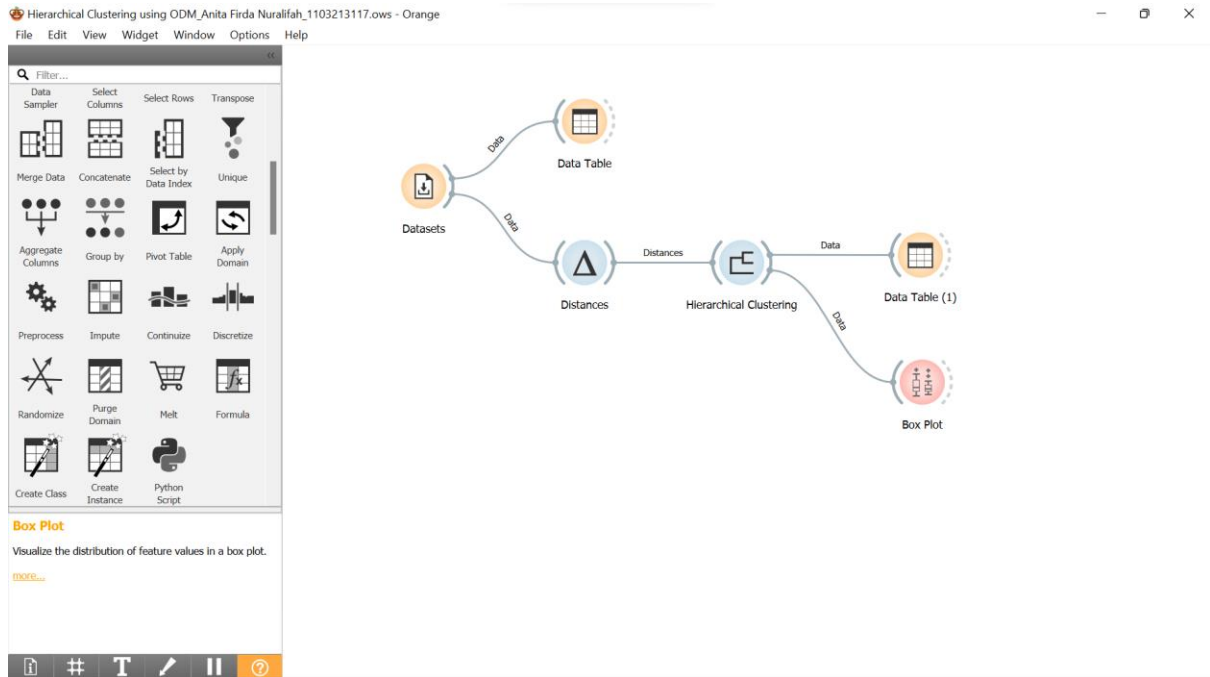
Labels: ☒ Student

☒ Send Automatically

16x16 1x1 | 1

Hierarchical Clustering using ODM

Tampilan menu awal



Dataset

Memasukkan dataset bawaan yang sudah tersedia di orange, yaitu data set course grades.

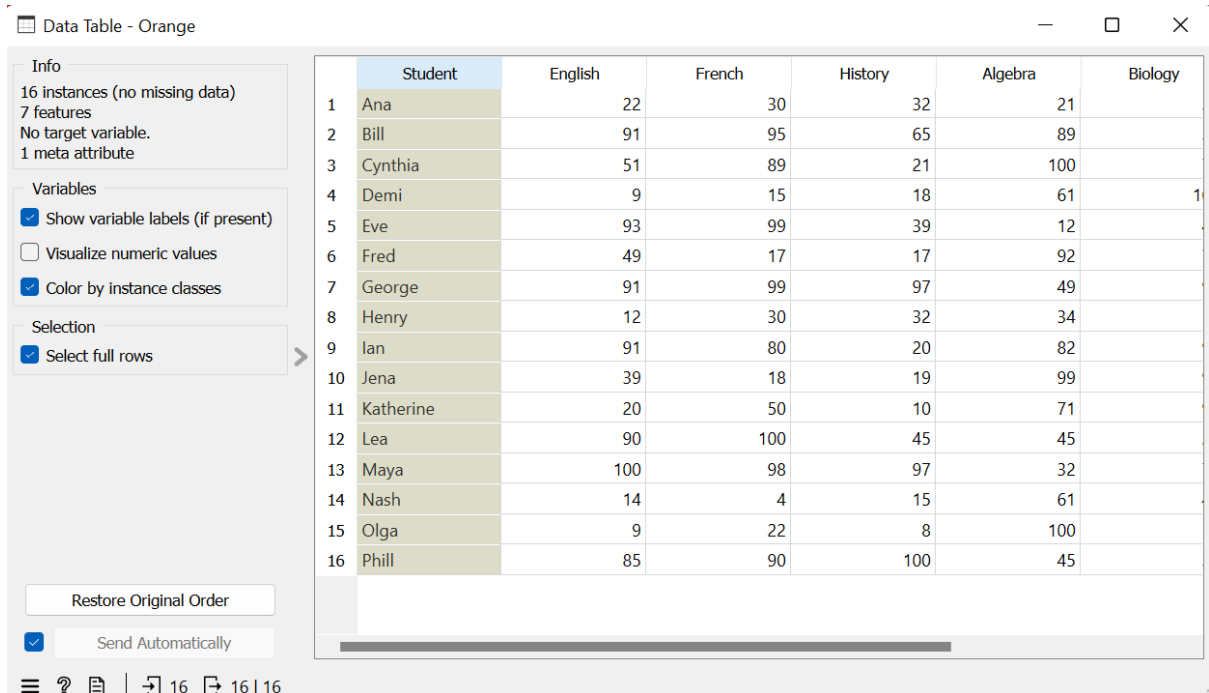
The screenshot shows the 'Datasets - Orange' window. It features a search bar at the top and a table of available datasets. The 'Course Grades' dataset is selected and highlighted in blue. Below the table, there is a 'Description' section for the 'Course Grades' dataset.

Title	Size	Instances	Variables	Target	Tags
Course Grades	9.2 KB	16	8	? none	synthetic, education
BBC3	2.6 MB	1407	3	C categorical	text, classification, news
Dendritic cells and monocytes in human blood	19.4 MB	1140	26595	C categorical	expression, human, homo-sapiens, ...
Dendritic cells and monocytes in human blood...	18.1 MB	1244	26595	C categorical	expression, human, homo-sapiens, ...
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3001	C categorical	genomics
HDI	45.2 KB	188	54	? none	economy, geo
ParlaMint	1.7 MB	1000	18	C categorical	text, classification, time, politics
SentiNews	5.0 MB	2000	8	C categorical	text, sentiment
TKI resistance	1.2 MB	280	468	C categorical	spectral
Abalone	187.5 KB	4177	9	N numeric	biology

Description
Course Grades
A small dataset with grades on seven courses (English, French, History, Algebra, Biology, Physics, Physical) that was handcrafted to introduce hierarchical clustering.

Data Table

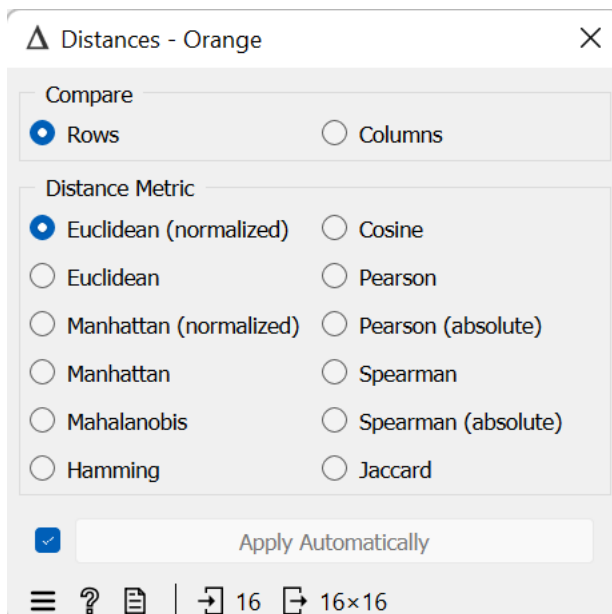
Memperlihatkan isi dataset yang sedang dipakai.



	Student	English	French	History	Algebra	Biology
1	Ana	22	30	32	21	
2	Bill	91	95	65	89	
3	Cynthia	51	89	21	100	
4	Demi	9	15	18	61	1
5	Eve	93	99	39	12	
6	Fred	49	17	17	92	
7	George	91	99	97	49	
8	Henry	12	30	32	34	
9	Ian	91	80	20	82	
10	Jena	39	18	19	99	
11	Katherine	20	50	10	71	
12	Lea	90	100	45	45	
13	Maya	100	98	97	32	
14	Nash	14	4	15	61	
15	Olga	9	22	8	100	
16	Phill	85	90	100	45	

Distances

Untuk mengukur jarak Euclidean antara setiap nilai siswa dan mengelompokkan datanya, kita bisa menggunakan distances.



Distances - Orange

Compare

☒ Rows ☐ Columns

Distance Metric

☒ Euclidean (normalized) ☐ Cosine

☐ Euclidean ☐ Pearson

☐ Manhattan (normalized) ☐ Pearson (absolute)

☐ Manhattan ☐ Spearman

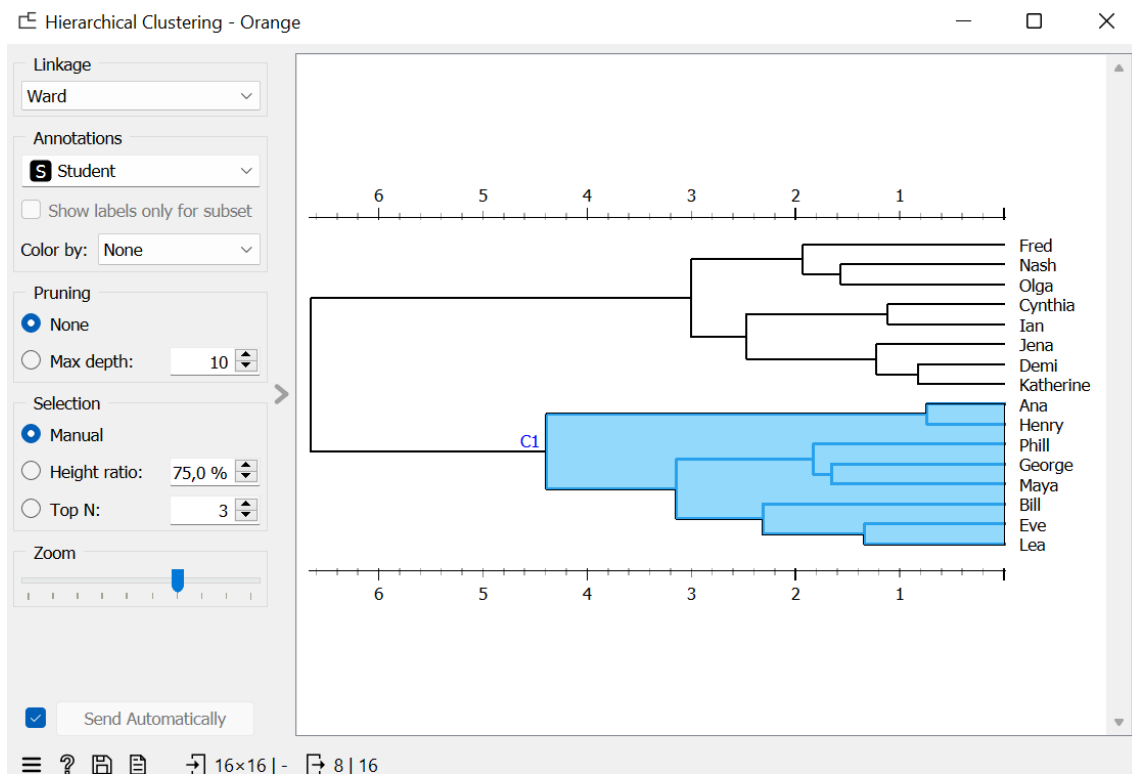
☐ Mahalanobis ☐ Spearman (absolute)

☐ Hamming ☐ Jaccard

☒ Apply Automatically

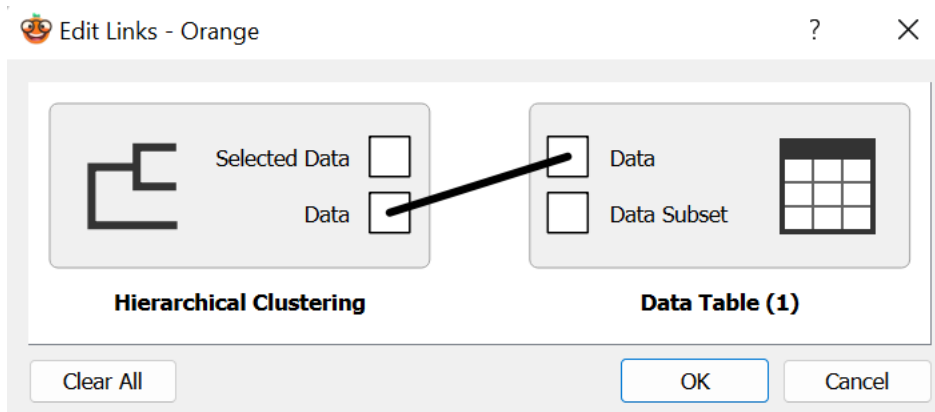
Hierarchical Clustering

Hierarchical clustering digunakan untuk menganalisis data dan menemukan pola dalam bentuk hierarki berdasarkan kedekatan atau kesamaan antar data.



Data table

Hubungkan terlebih dahulu koneksi antara hierarchial clustering dengan data table untuk mengkomunikasikan seluruh Kumpulan data.



Setelah melakukan hierarchial clustering, pengelompokkan tersebut akan memunculkan data yang dipilih serta seluruh dataset dengan kolom tambahan yang menunjukkan pilihan.

Data Table (1) - Orange

Info
16 instances (no missing data)
7 features
Target with 2 values
2 meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

☒ Send Automatically

	Selected	Student	Cluster	English	French	History
1	Yes	Ana	C1	22	30	
2	Yes	Bill	C1	91	95	
3	No	Cynthia	Other	51	89	
4	No	Demi	Other	9	15	
5	Yes	Eve	C1	93	99	
6	No	Fred	Other	49	17	
7	Yes	George	C1	91	99	
8	Yes	Henry	C1	12	30	
9	No	Ian	Other	91	80	
10	No	Jena	Other	39	18	
11	No	Katherine	Other	20	50	
12	Yes	Lea	C1	90	100	
13	Yes	Maya	C1	100	98	
14	No	Nash	Other	14	4	
15	No	Olga	Other	9	22	
16	Yes	Phill	C1	85	90	1

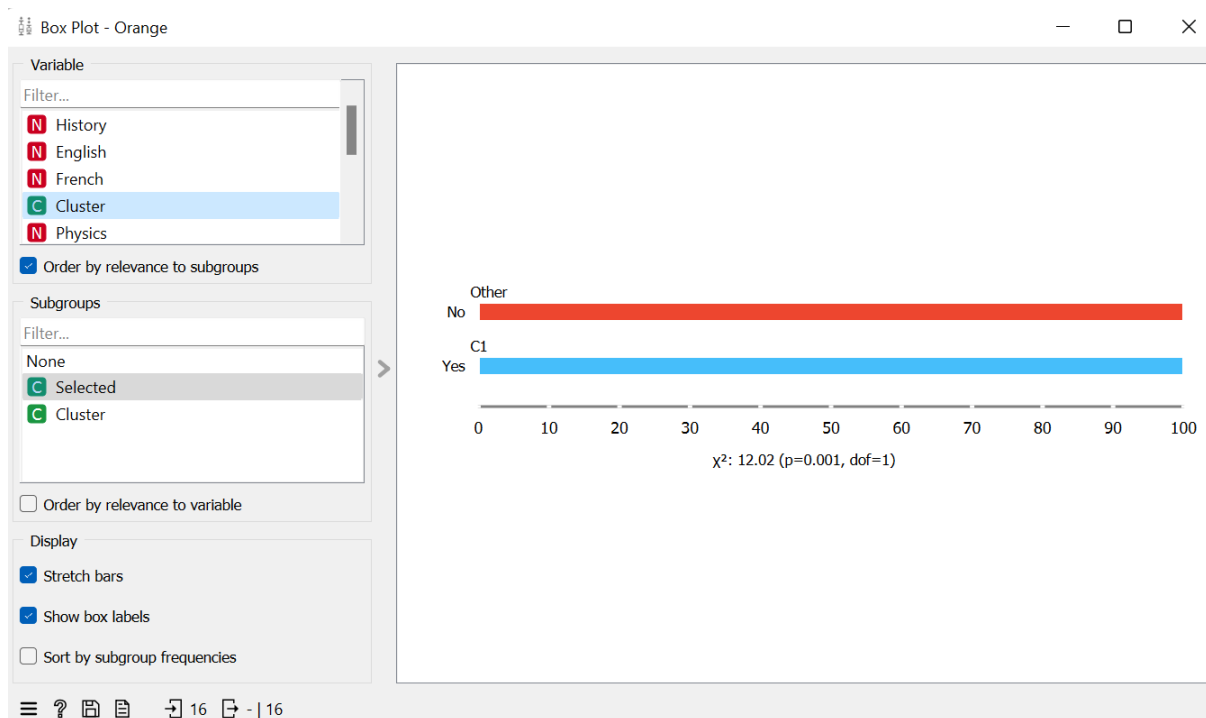
16 | 16 | 16

Box plot

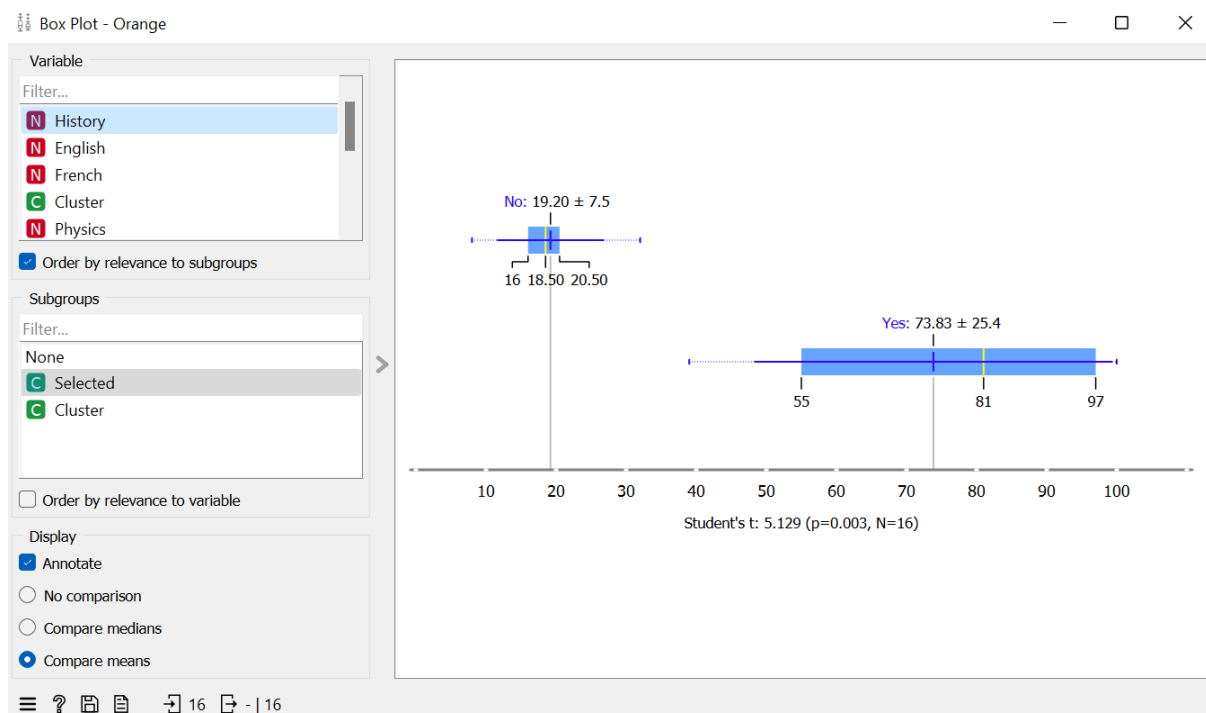
Hubungkan hierarchial clustering dengan box plot untuk mentransfer semua data, bukan hanya data pilihan saja.



Pada box plot kita pilih selected sebagai fitur subgroup dan centang order by relevance to subgroup.



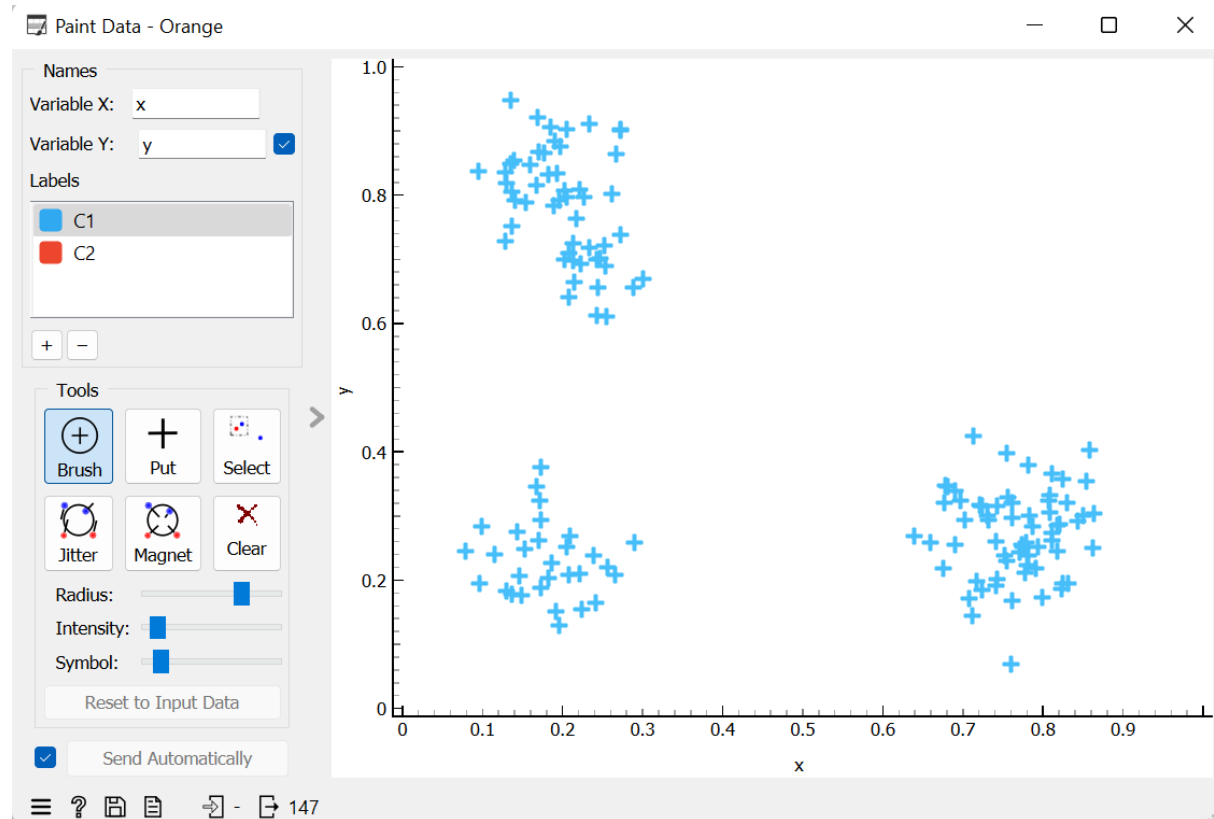
Lalu jika kita memilih mata pelajaran yang lain, akan terlihat nilai rata rata cluster masing masing siswa. Box plot menggunakan statistik T students untuk mengurutkan fitur berdasarkan perbedaan antara distribusinya di dalam cluster dan nilai fitur di luar cluster.



K-Means Clustering using ODM

Paint data

Jika ingin mengolah data tentu kita membutuhkan sebuah data, maka dari itu dengan paint data kita dapat menggambar beberapa cluster.



K-Means

Selanjutnya data akan dikirimkan ke K-Means.

The screenshot shows the 'k-Means' widget configuration window in Orange3. The window title is 'k-Means - Orange'. The 'Number of Clusters' section has two options: 'Fixed' (selected) with a value of 3, and 'From' (unselected) with a range from 2 to 8. The 'Preprocessing' section has a 'Normalize columns' checkbox which is checked. The 'Initialization' section has a dropdown menu set to 'Initialize with KMeans++'. Below this, the 'Re-runs' field is set to 10 and the 'Maximum iterations' field is set to 300. At the bottom, there is an 'Apply Automatically' checkbox which is checked. The bottom status bar shows icons for file operations and a count of 147 data points.

Data table

Pada data table kita melihat ada 2 kolom tambahan, dimana satu kolom menunjukkan cluster setiap poin dan satu lagi menunjukkan skor silhouette.

Data Table - Orange

Info
147 instances (no missing data)
2 features
No target variable.
2 meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

Restore Original Order

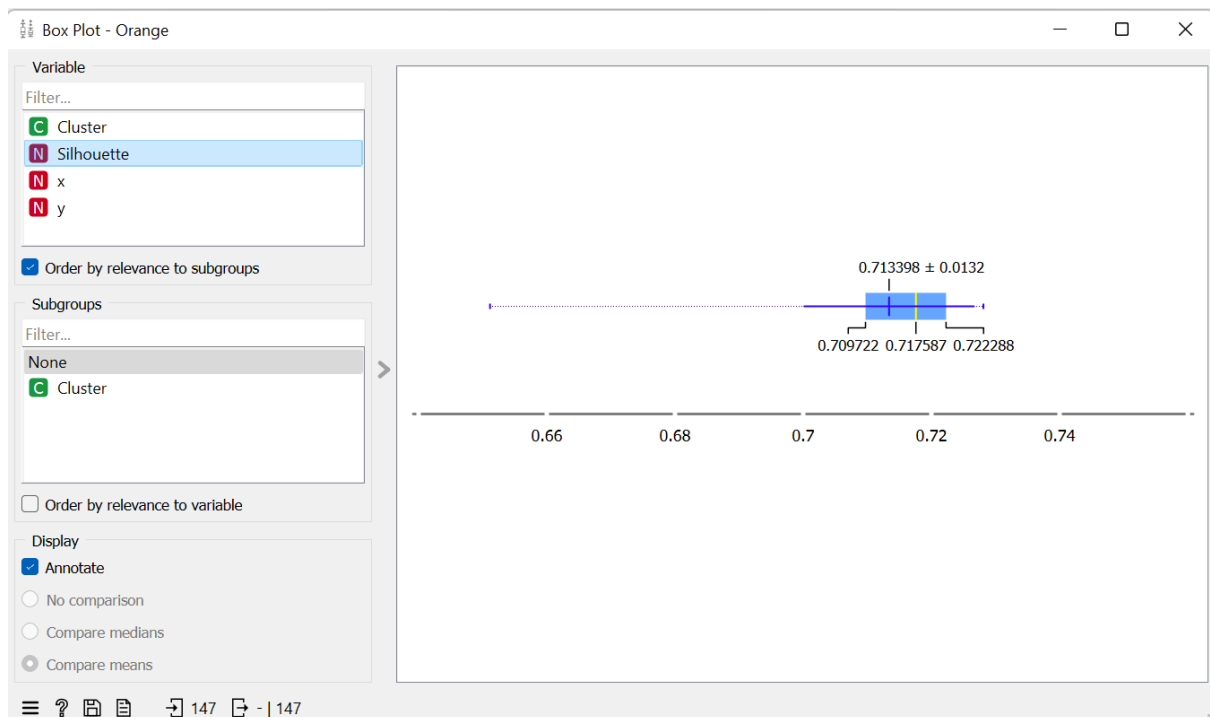
☒ Send Automatically

	Cluster	Silhouette	x	y
1	C2	0.709672	0.271958	0.90172
2	C2	0.714373	0.266719	0.863773
3	C2	0.71758	0.190178	0.884604
4	C2	0.709373	0.272879	0.903083
5	C2	0.71459	0.205686	0.902946
6	C2	0.722618	0.182041	0.831534
7	C2	0.717669	0.135664	0.848635
8	C2	0.722934	0.204704	0.796277
9	C2	0.717587	0.138683	0.854672
10	C2	0.717615	0.129258	0.817988
11	C2	0.719736	0.177135	0.865733
12	C2	0.722651	0.192805	0.833214
13	C2	0.722213	0.167108	0.815872
14	C2	0.720392	0.158936	0.847048
15	C2	0.717348	0.261966	0.801623
16	C2	0.72277	0.196109	0.791248
17	C2	0.706149	0.21262	0.698563
18	C2	0.690694	0.21415	0.663688
19	C2	0.723198	0.202738	0.807821

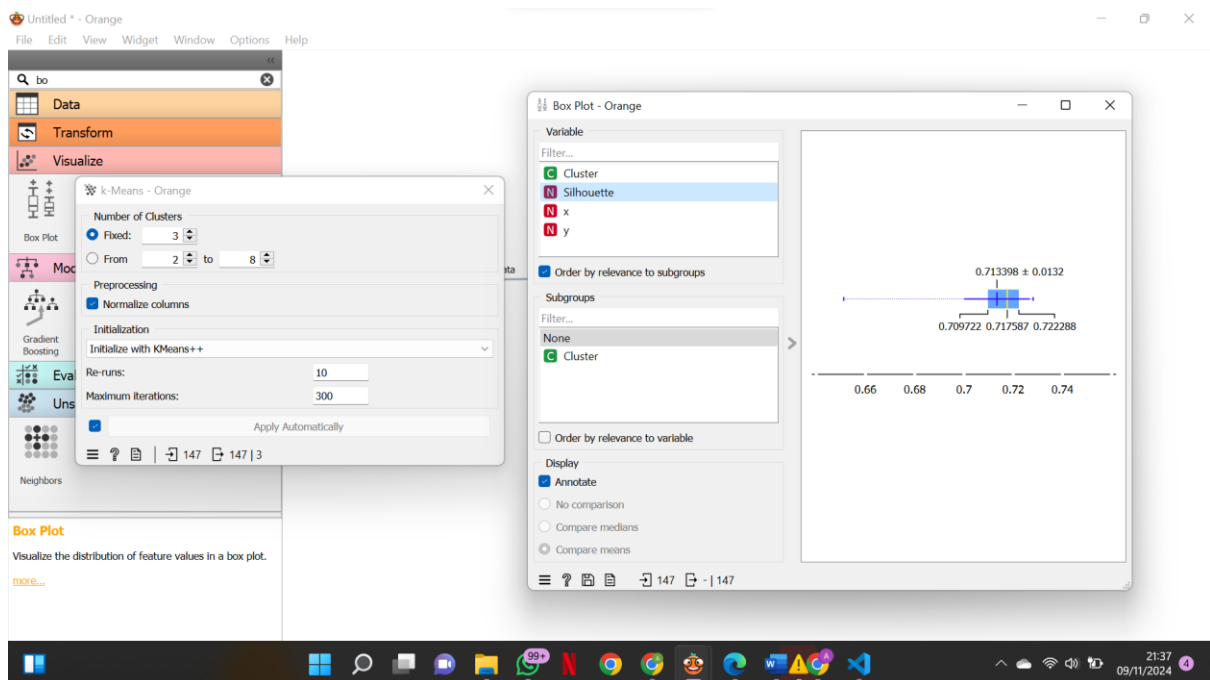
147 | 147

Box plot

Untuk menghitung rata rata silhouette, kita akan menggunakan box plot. Dalam contoh ini kita mendapatkan skor sekitar 0.72



Tempatkan K-Means dan boxplot secara berdampingan untuk melihat bagaimana pemilihan setelan yang berbeda memengaruhi skor silhouette rata rata. Dengan cara ini kita dapat menyempurnakan parameter k untuk menemukan pengelompokan data yang terbaik.



Scatter plot

Tampilan scatter plot dan paint data sama dan artinya penilaian dengan silhouette benar benar berfungsi.

