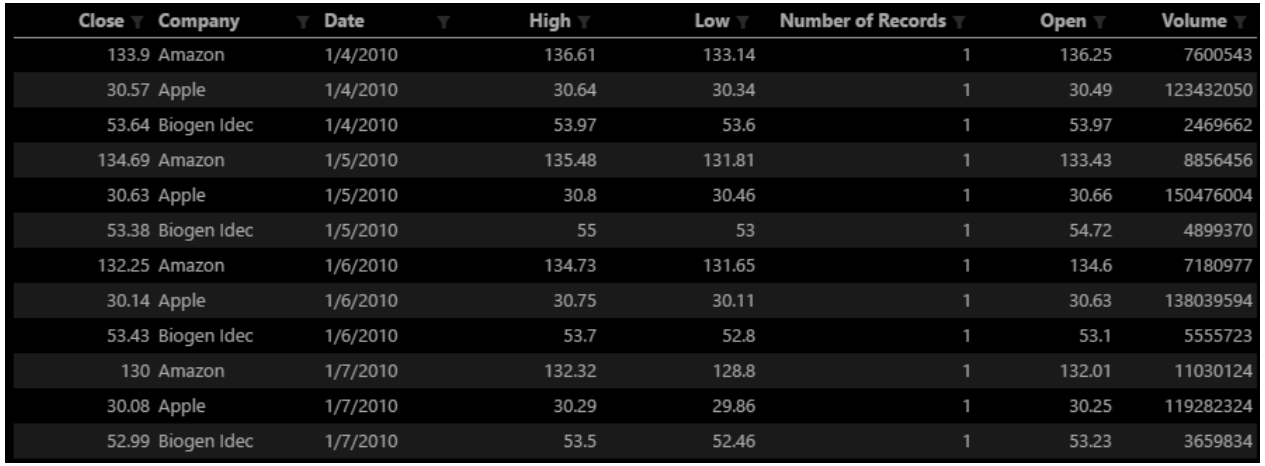
Knime Hands-on
World Indicators Dataset

Prof. Dr. Daniel A. Keim
University of Konstanz, Germany

2nd ACM Europe Summer School in Data Science 12 – 18 July 2018,
Athens, Greece

Stock Dataset

- `Close/Open`: stock price at open and close
- `High/Low`: max and min price during the day
- `Volume`: trading volume
- `Number of records`: the number of contained records.



Close	Company	Date	High	Low	Number of Records	Open	Volume
133.9	Amazon	1/4/2010	136.61	133.14	1	136.25	7600543
30.57	Apple	1/4/2010	30.64	30.34	1	30.49	123432050
53.64	Biogen Idec	1/4/2010	53.97	53.6	1	53.97	2469662
134.69	Amazon	1/5/2010	135.48	131.81	1	133.43	8856456
30.63	Apple	1/5/2010	30.8	30.46	1	30.66	150476004
53.38	Biogen Idec	1/5/2010	55	53	1	54.72	4899370
132.25	Amazon	1/6/2010	134.73	131.65	1	134.6	7180977
30.14	Apple	1/6/2010	30.75	30.11	1	30.63	138039594
53.43	Biogen Idec	1/6/2010	53.7	52.8	1	53.1	5555723
130	Amazon	1/7/2010	132.32	128.8	1	132.01	11030124
30.08	Apple	1/7/2010	30.29	29.86	1	30.25	119282324
52.99	Biogen Idec	1/7/2010	53.5	52.46	1	53.23	3659834

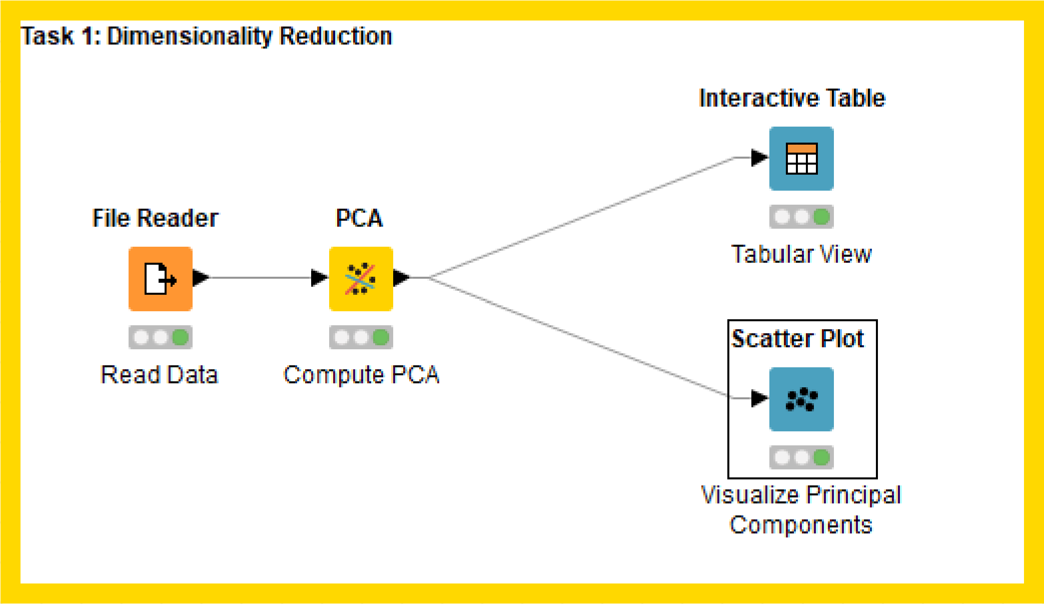
Stock Dataset – Task 1

- Find similar stocks in terms of the Dimensions *open*, *close*, *high*, and *low.*

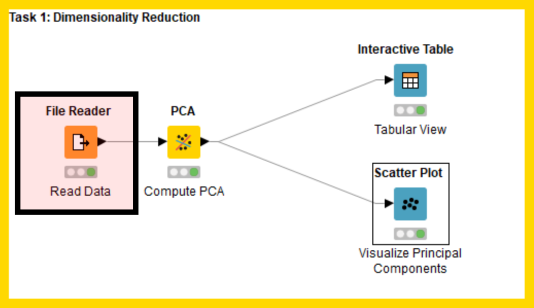
- Apply a dimension reduction technique.

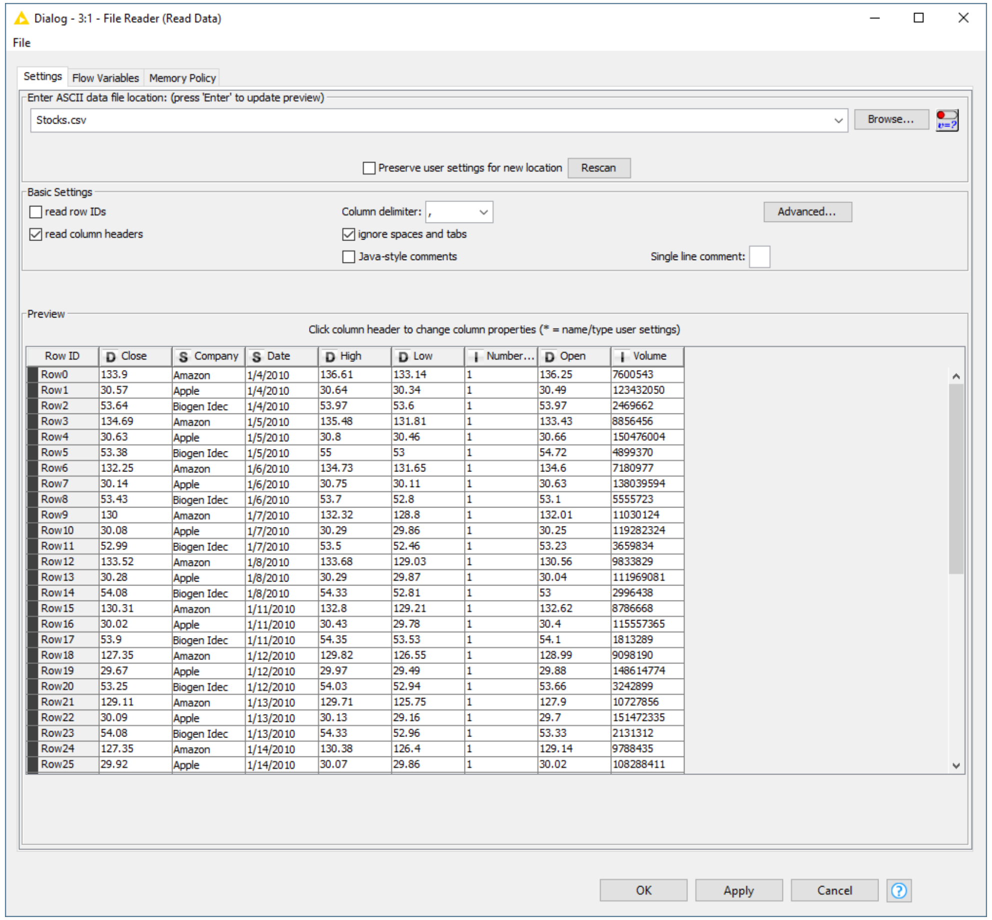
- Inspect the data in a table and visualize the projected dataset in a scatter plot. Identify similar stocks visually.

Task 1: Knime Workflow



Task 1: Read CSV





Enter ASCII data file location: (press 'Enter' to update preview)

Stocks.csv

□ Preserve user settings for new location

Basic Settings

□ read row IDs

☑ read column headers

Column delimiter: ,

☑ ignore spaces and tabs

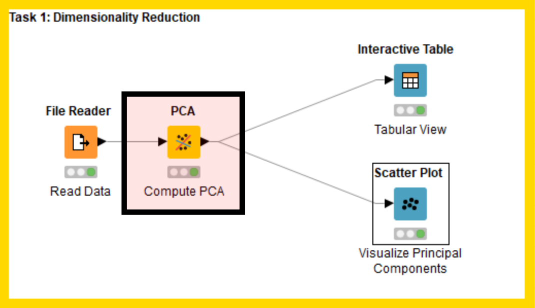
□ Java-style comments

Single line comment:

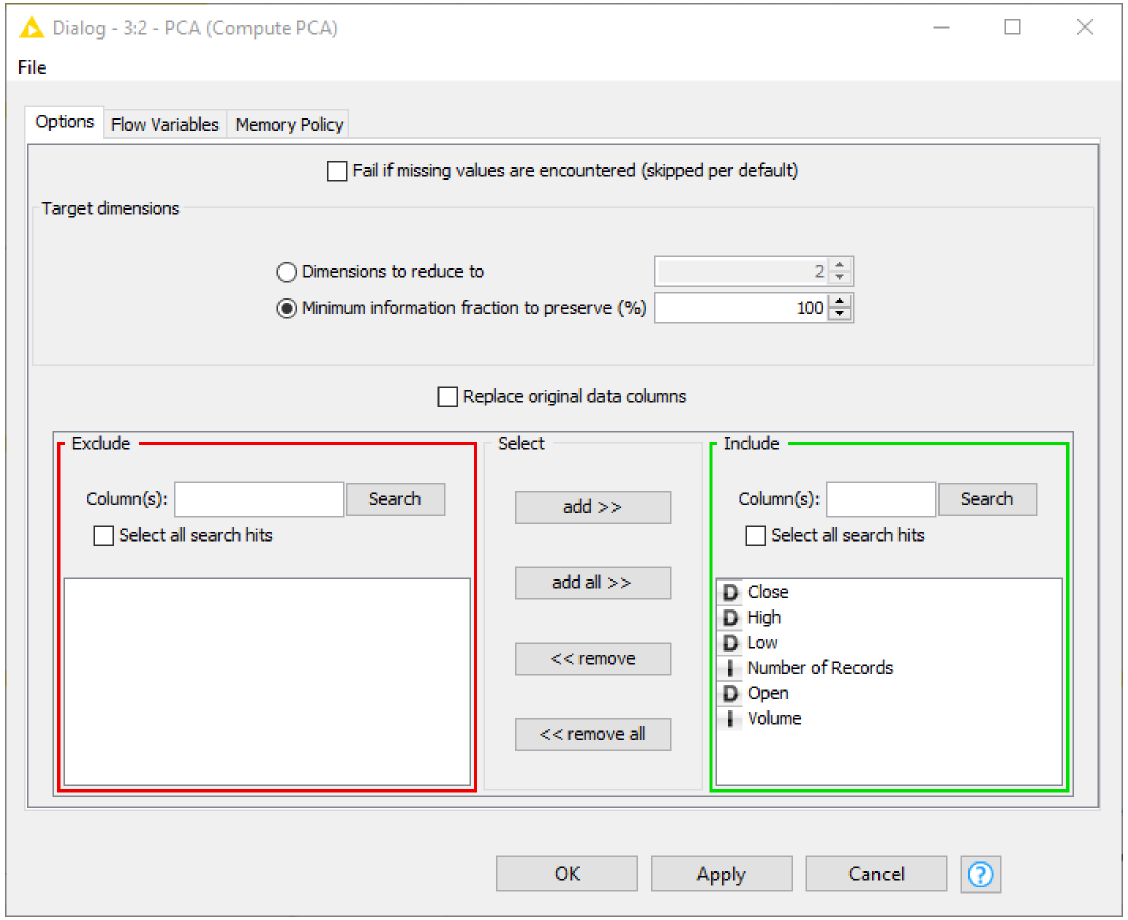
Preview

Click column header to change column properties (* = name/type user settings)

Row ID	D Close	S Company	S Date	D High	D Low	I Number...	D Open	I Volume
Row0	133.9	Amazon	1/4/2010	136.61	133.14	1	136.25	7600543
Row1	30.57	Apple	1/4/2010	30.64	30.34	1	30.49	123432050
Row2	53.64	Biogen Idec	1/4/2010	53.97	53.6	1	53.97	2469662
Row3	134.69	Amazon	1/5/2010	135.48	131.81	1	133.43	8856456
Row4	30.63	Apple	1/5/2010	30.8	30.46	1	30.66	150476004
Row5	53.38	Biogen Idec	1/5/2010	55	53	1	54.72	4899370
Row6	132.25	Amazon	1/6/2010	134.73	131.65	1	134.6	7180977
Row7	30.14	Apple	1/6/2010	30.75	30.11	1	30.63	138039594
Row8	53.43	Biogen Idec	1/6/2010	53.7	52.8	1	53.1	5555723
Row9	130	Amazon	1/7/2010	132.32	128.8	1	132.01	11030124
Row10	30.08	Apple	1/7/2010	30.29	29.86	1	30.25	119282324
Row11	52.99	Biogen Idec	1/7/2010	53.5	52.46	1	53.23	3659834
Row12	133.52	Amazon	1/8/2010	133.68	129.03	1	130.56	9833829
Row13	30.28	Apple	1/8/2010	30.29	29.87	1	30.04	111969081
Row14	54.08	Biogen Idec	1/8/2010	54.33	52.81	1	53	2996438
Row15	130.31	Amazon	1/11/2010	132.8	129.21	1	132.62	8786668
Row16	30.02	Apple	1/11/2010	30.43	29.78	1	30.4	115557365
Row17	53.9	Biogen Idec	1/11/2010	54.35	53.53	1	54.1	1813289
Row18	127.35	Amazon	1/12/2010	129.82	126.55	1	128.99	9098190
Row19	29.67	Apple	1/12/2010	29.97	29.49	1	29.88	148614774
Row20	53.25	Biogen Idec	1/12/2010	54.03	52.94	1	53.66	3242899
Row21	129.11	Amazon	1/13/2010	129.71	125.75	1	127.9	10727856
Row22	30.09	Apple	1/13/2010	30.13	29.16	1	29.7	151472335
Row23	54.08	Biogen Idec	1/13/2010	54.33	52.96	1	53.33	2131312
Row24	127.35	Amazon	1/14/2010	130.38	126.4	1	129.14	9788435
Row25	29.92	Apple	1/14/2010	30.07	29.86	1	30.02	108288411



Task 1: PCA Configuration



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

Target dimensions

○ Dimensions to reduce to 2

● Minimum information fraction to preserve (%) 100

□ Replace original data columns

Exclude

Column(s):

□ Select all search hits

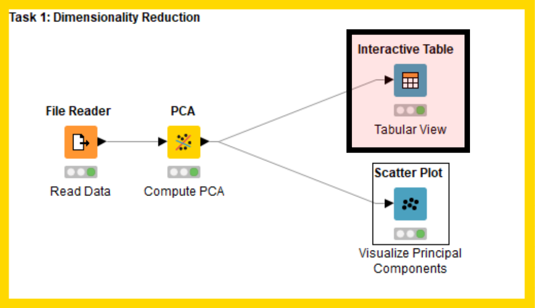
Select

Include

Column(s):

□ Select all search hits

- D Close
- D High
- D Low
- I Number of Records
- D Open
- I Volume



Task 1: PCA Tabular View

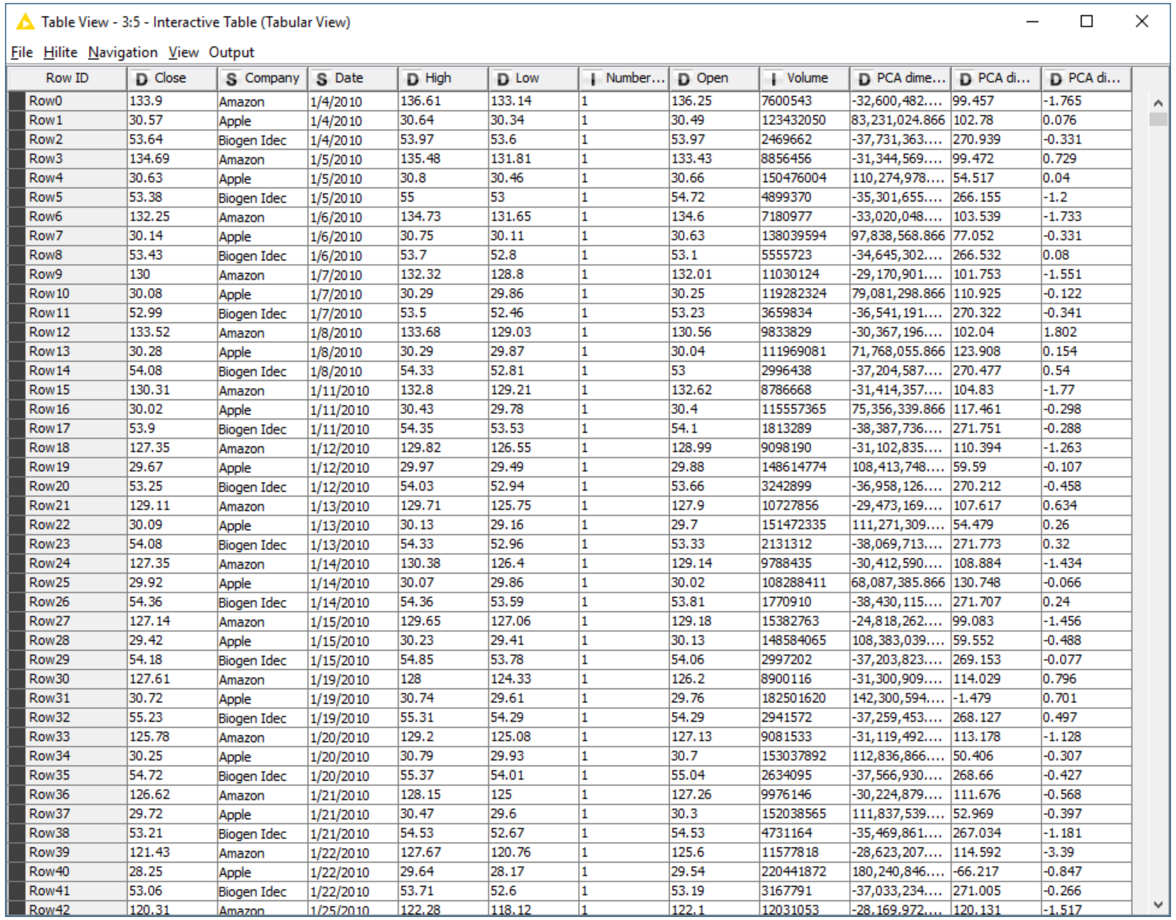
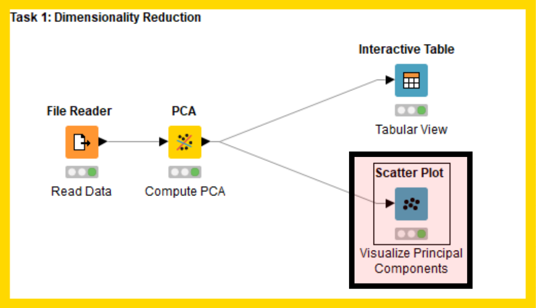
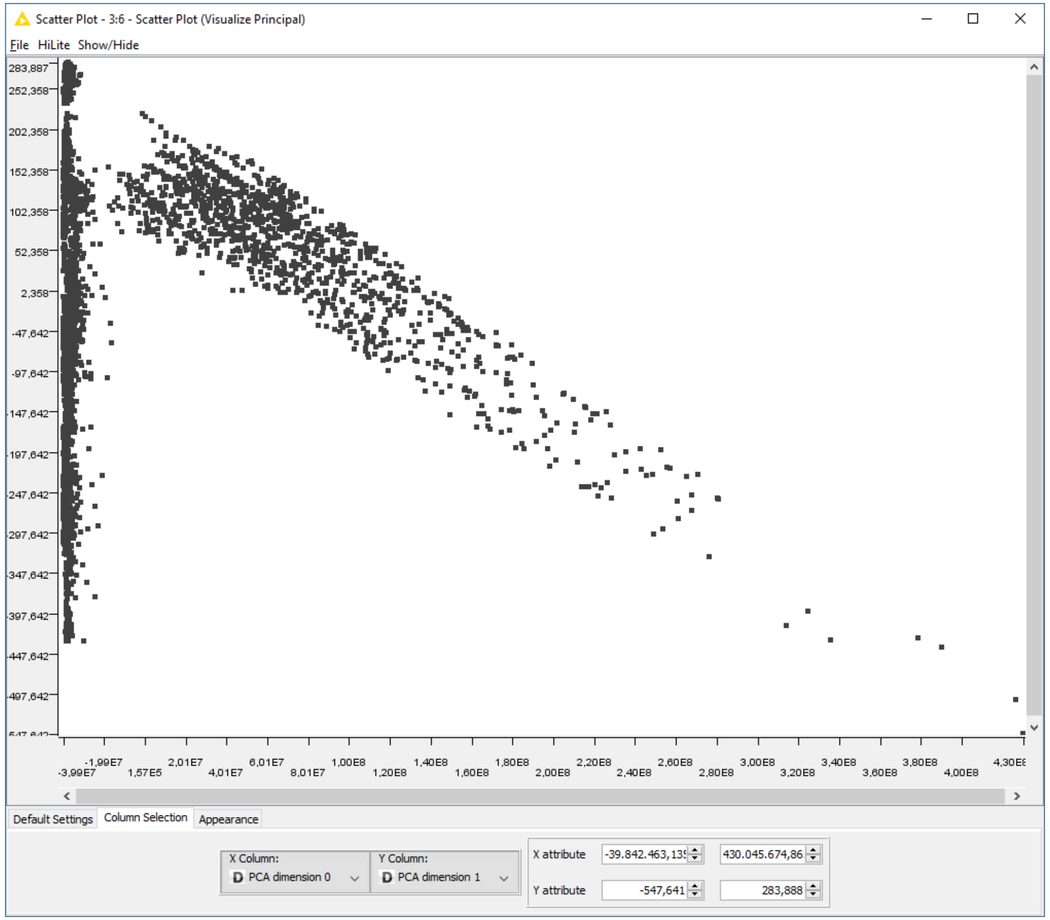


Table View - 3:5 - Interactive Table (Tabular View)
File Hilite Navigation View Output

Row ID	D Close	S Company	S Date	D High	D Low	I Number...	D Open	I Volume	D PCA dime...	D PCA di...	D PCA di...
Row0	133.9	Amazon	1/4/2010	136.61	133.14	1	136.25	7600543	-32,600,482....	99.457	-1.765
Row1	30.57	Apple	1/4/2010	30.64	30.34	1	30.49	123432050	83,231,024.866	102.78	0.076
Row2	53.64	Biogen Idec	1/4/2010	53.97	53.6	1	53.97	2469662	-37,731,363....	270.939	-0.331
Row3	134.69	Amazon	1/5/2010	135.48	131.81	1	133.43	8856456	-31,344,569....	99.472	0.729
Row4	30.63	Apple	1/5/2010	30.8	30.46	1	30.66	150476004	110,274,978....	54.517	0.04
Row5	53.38	Biogen Idec	1/5/2010	55	53	1	54.72	4899370	-35,301,655....	266.155	-1.2
Row6	132.25	Amazon	1/6/2010	134.73	131.65	1	134.6	7180977	-33,020,048....	103.539	-1.733
Row7	30.14	Apple	1/6/2010	30.75	30.11	1	30.63	138039594	97,838,568.866	77.052	-0.331
Row8	53.43	Biogen Idec	1/6/2010	53.7	52.8	1	53.1	5555723	-34,645,302....	266.532	0.08
Row9	130	Amazon	1/7/2010	132.32	128.8	1	132.01	11030124	-29,170,901....	101.753	-1.551
Row10	30.08	Apple	1/7/2010	30.29	29.86	1	30.25	119282324	79,081,298.866	110.925	-0.122
Row11	52.99	Biogen Idec	1/7/2010	53.5	52.46	1	53.23	3659834	-36,541,191....	270.322	-0.341
Row12	133.52	Amazon	1/8/2010	133.68	129.03	1	130.56	9833829	-30,367,196....	102.04	1.802
Row13	30.28	Apple	1/8/2010	30.29	29.87	1	30.04	111969081	71,768,055.866	123.908	0.154
Row14	54.08	Biogen Idec	1/8/2010	54.33	52.81	1	53	2996438	-37,204,587....	270.477	0.54
Row15	130.31	Amazon	1/11/2010	132.8	129.21	1	132.62	8786668	-31,414,357....	104.83	-1.77
Row16	30.02	Apple	1/11/2010	30.43	29.78	1	30.4	115557365	75,356,339.866	117.461	-0.298
Row17	53.9	Biogen Idec	1/11/2010	54.35	53.53	1	54.1	1813289	-38,387,736....	271.751	-0.288
Row18	127.35	Amazon	1/12/2010	129.82	126.55	1	128.99	9098190	-31,102,835....	110.394	-1.263
Row19	29.67	Apple	1/12/2010	29.97	29.49	1	29.88	148614774	108,413,748....	59.59	-0.107
Row20	53.25	Biogen Idec	1/12/2010	54.03	52.94	1	53.66	3242899	-36,958,126....	270.212	-0.458
Row21	129.11	Amazon	1/13/2010	129.71	125.75	1	127.9	10727856	-29,473,169....	107.617	0.634
Row22	30.09	Apple	1/13/2010	30.13	29.16	1	29.7	151472335	111,271,309....	54.479	0.26
Row23	54.08	Biogen Idec	1/13/2010	54.33	52.96	1	53.33	2131312	-38,069,713....	271.773	0.32
Row24	127.35	Amazon	1/14/2010	130.38	126.4	1	129.14	9788435	-30,412,590....	108.884	-1.434
Row25	29.92	Apple	1/14/2010	30.07	29.86	1	30.02	108288411	68,087,385.866	130.748	-0.066
Row26	54.36	Biogen Idec	1/14/2010	54.36	53.59	1	53.81	1770910	-38,430,115....	271.707	0.24
Row27	127.14	Amazon	1/15/2010	129.65	127.06	1	129.18	15382763	-24,818,262....	99.083	-1.456
Row28	29.42	Apple	1/15/2010	30.23	29.41	1	30.13	148584065	108,383,039....	59.552	-0.488
Row29	54.18	Biogen Idec	1/15/2010	54.85	53.78	1	54.06	2997202	-37,203,823....	269.153	-0.077
Row30	127.61	Amazon	1/19/2010	128	124.33	1	126.2	8900116	-31,300,909....	114.029	0.796
Row31	30.72	Apple	1/19/2010	30.74	29.61	1	29.76	182501620	142,300,594....	-1.479	0.701
Row32	55.23	Biogen Idec	1/19/2010	55.31	54.29	1	54.29	2941572	-37,259,453....	268.127	0.497
Row33	125.78	Amazon	1/20/2010	129.2	125.08	1	127.13	9081533	-31,119,492....	113.178	-1.128
Row34	30.25	Apple	1/20/2010	30.79	29.93	1	30.7	153037892	112,836,866....	50.406	-0.307
Row35	54.72	Biogen Idec	1/20/2010	55.37	54.01	1	55.04	2634095	-37,566,930....	268.66	-0.427
Row36	126.62	Amazon	1/21/2010	128.15	125	1	127.26	9976146	-30,224,879....	111.676	-0.568
Row37	29.72	Apple	1/21/2010	30.47	29.6	1	30.3	152038565	111,837,539....	52.969	-0.397
Row38	53.21	Biogen Idec	1/21/2010	54.53	52.67	1	54.53	4731164	-35,469,861....	267.034	-1.181
Row39	121.43	Amazon	1/22/2010	127.67	120.76	1	125.6	11577818	-28,623,207....	114.592	-3.39
Row40	28.25	Apple	1/22/2010	29.64	28.17	1	29.54	220441872	180,240,846....	-66.217	-0.847
Row41	53.06	Biogen Idec	1/22/2010	53.71	52.6	1	53.19	3167791	-37,033,234....	271.005	-0.266
Row42	120.31	Amazon	1/25/2010	122.28	118.12	1	122.1	12031053	-28,169,972....	120.131	-1.517

Task 1: PCA Scatter Plot (pc0, pc1)





X Column: PCA dimension 0

Y Column: PCA dimension 1

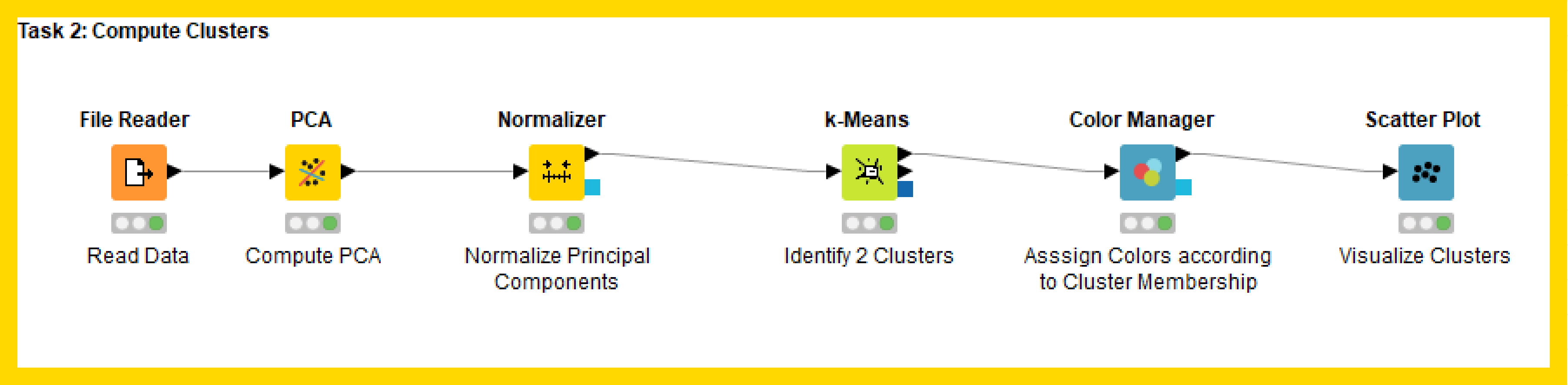
X attribute -39.842.463,13 430.045.674,86

Y attribute -547,641 283,888

Task 2 – Compute Clusters

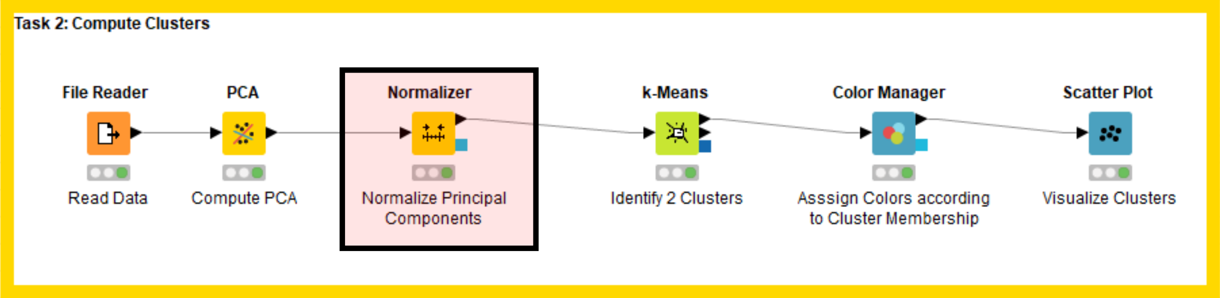
- Cluster the projected dataset using k-means.
- Assign colors corresponding to cluster membership.
- Visualize the cluster memberships.

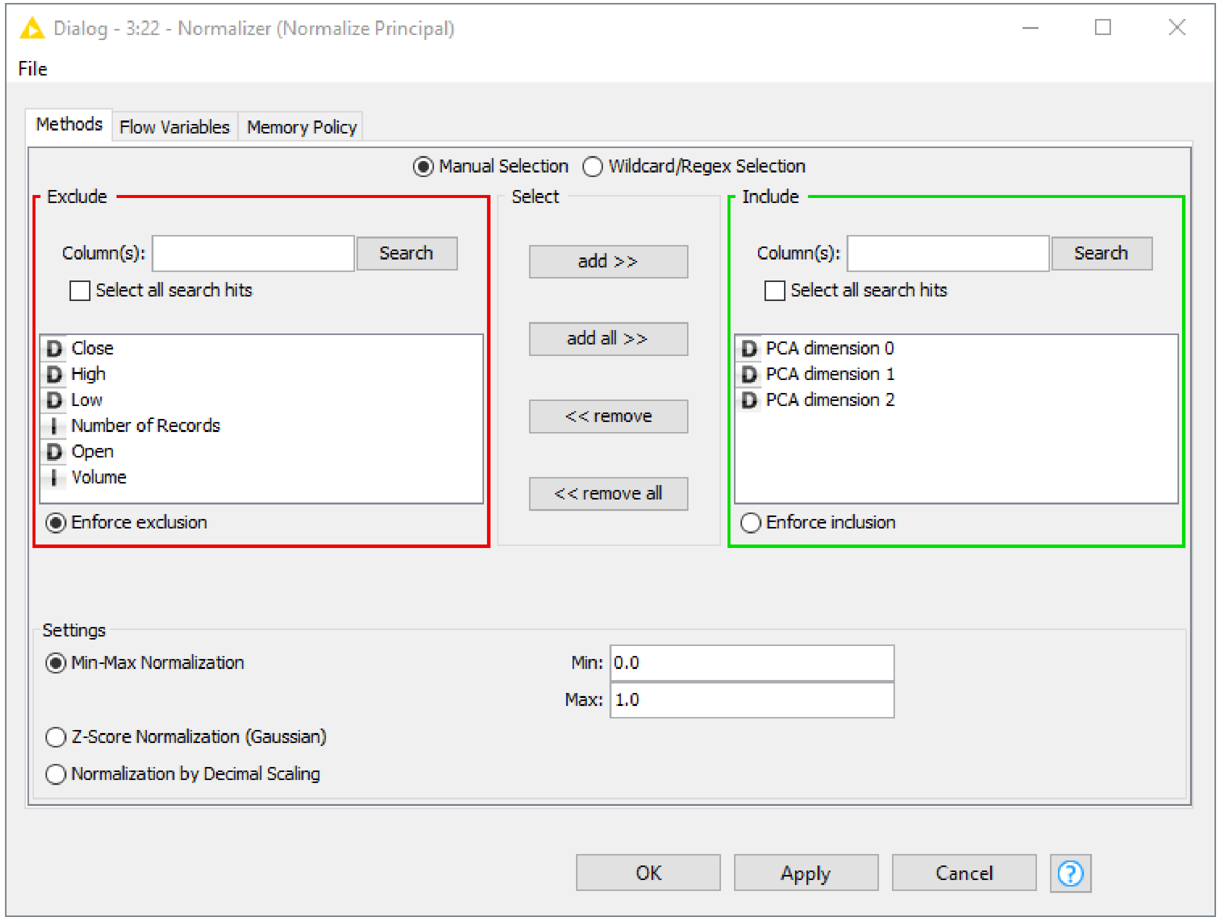
Task 2 – Knime Workflow



Task 2: Compute Clusters

Normalize PCs




Dialog - 3:22 - Normalizer (Normalize Principal)

File

Methods | Flow Variables | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Column(s): Search

□ Select all search hits

- D Close
- D High
- D Low
- I Number of Records
- D Open
- I Volume

● Enforce exclusion

Select

add >>

add all >>

<< remove

<< remove all

Include

Column(s): Search

□ Select all search hits

- D PCA dimension 0
- D PCA dimension 1
- D PCA dimension 2

○ Enforce inclusion

Settings

● Min-Max Normalization

Min: 0.0

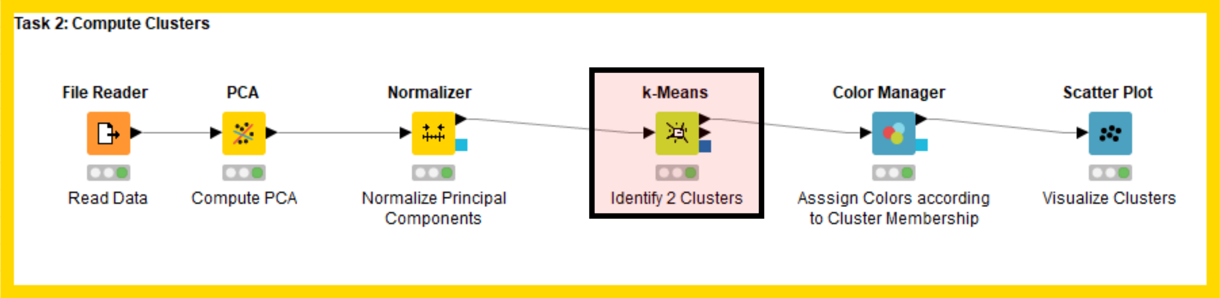
Max: 1.0

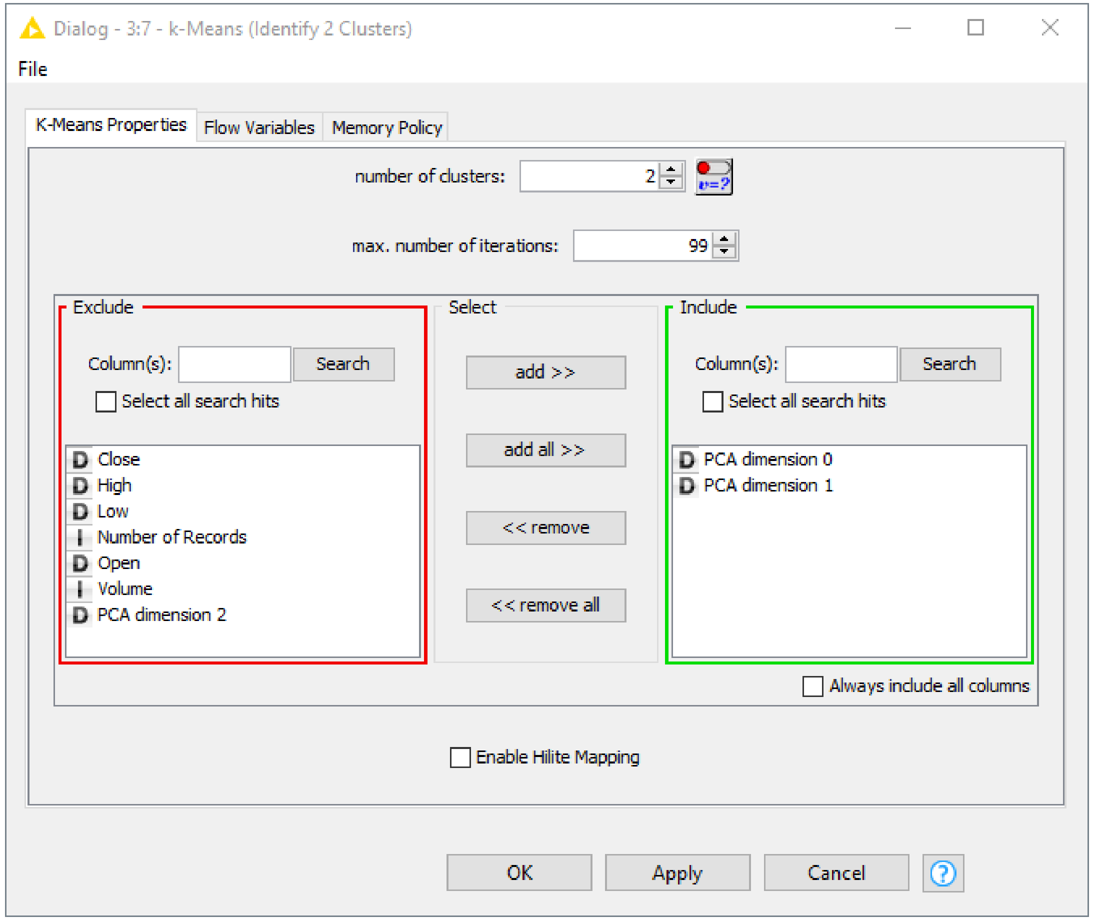
○ Z-Score Normalization (Gaussian)

○ Normalization by Decimal Scaling

OK | Apply | Cancel | ?

k-means Clustering





number of clusters: 2

max. number of iterations: 99

Exclude

Column(s):

□ Select all search hits

- D Close
- D High
- D Low
- I Number of Records
- D Open
- I Volume
- D PCA dimension 2

Include

Column(s):

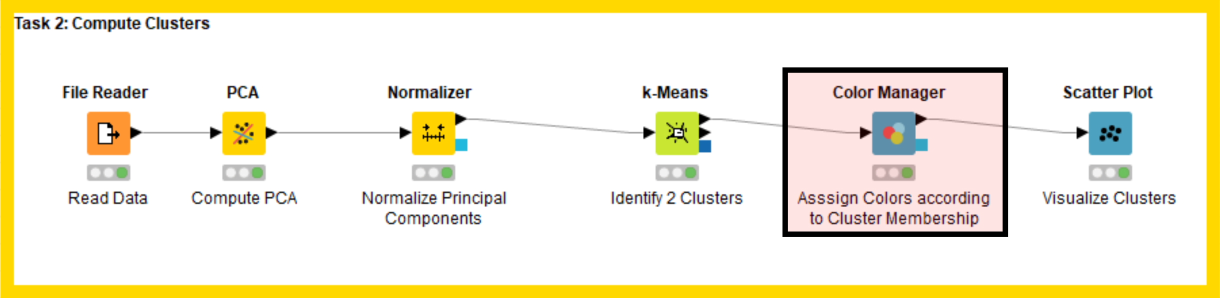
□ Select all search hits

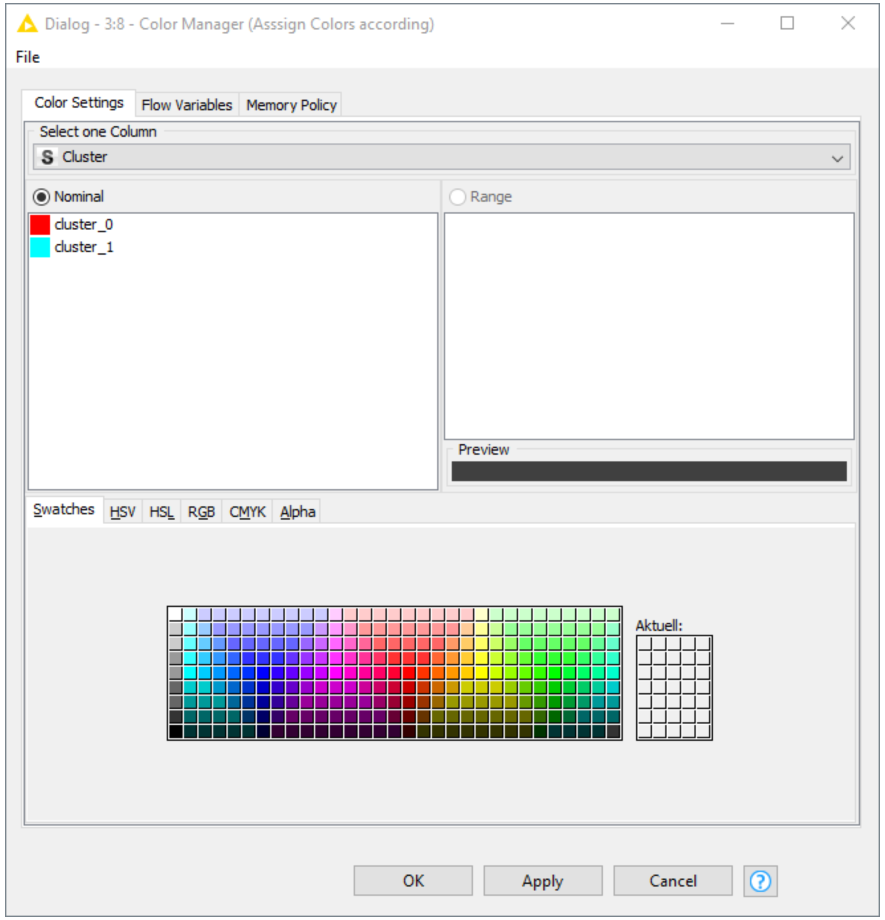
- D PCA dimension 0
- D PCA dimension 1

□ Always include all columns

□ Enable Hilite Mapping

Assign Colors to Clusters





Select one Column

S Cluster

● Nominal

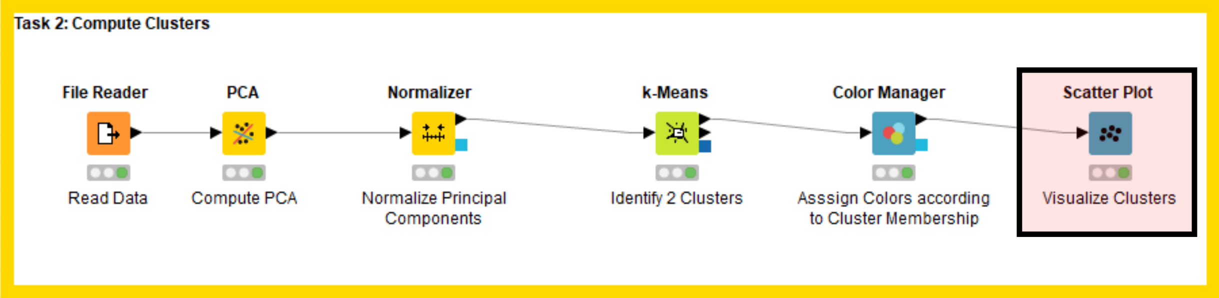
🟥 cluster_0

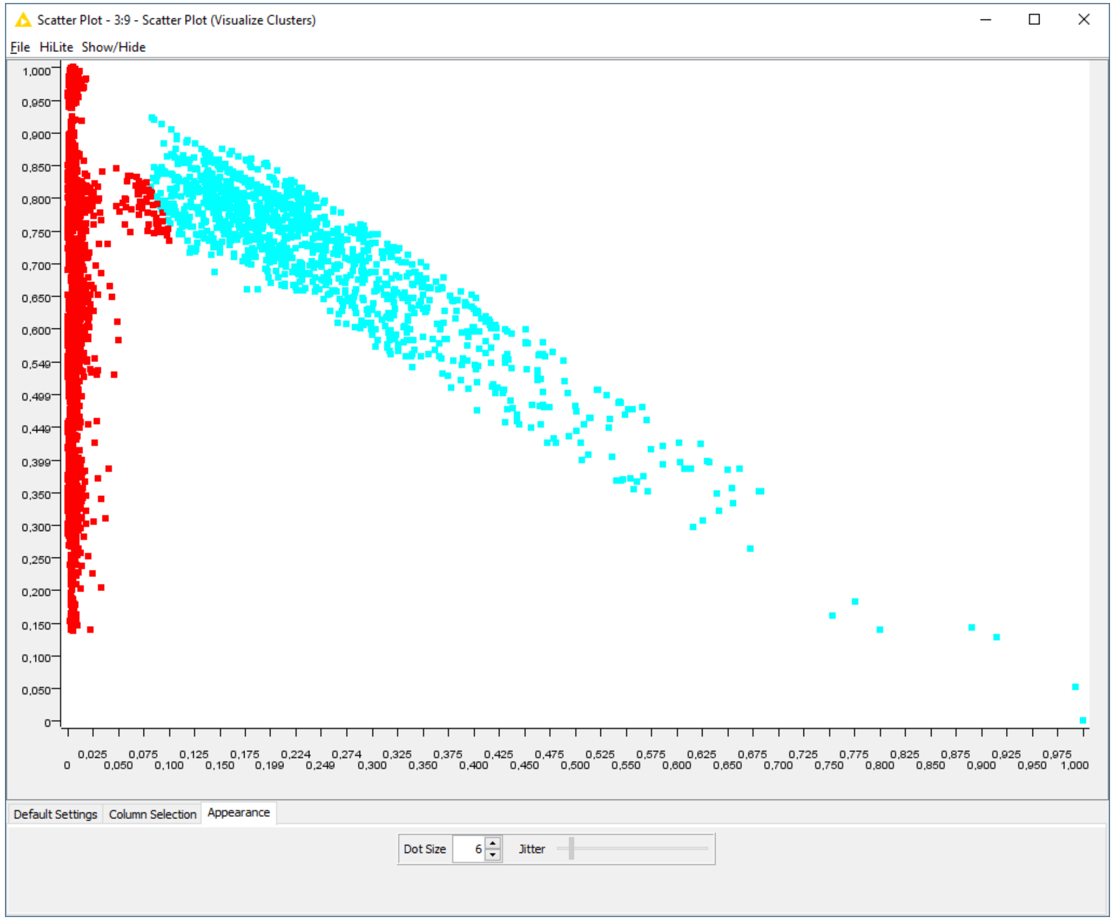
🟦 cluster_1

○ Range

Preview

Visualize Cluster Memberships





Dot Size 6

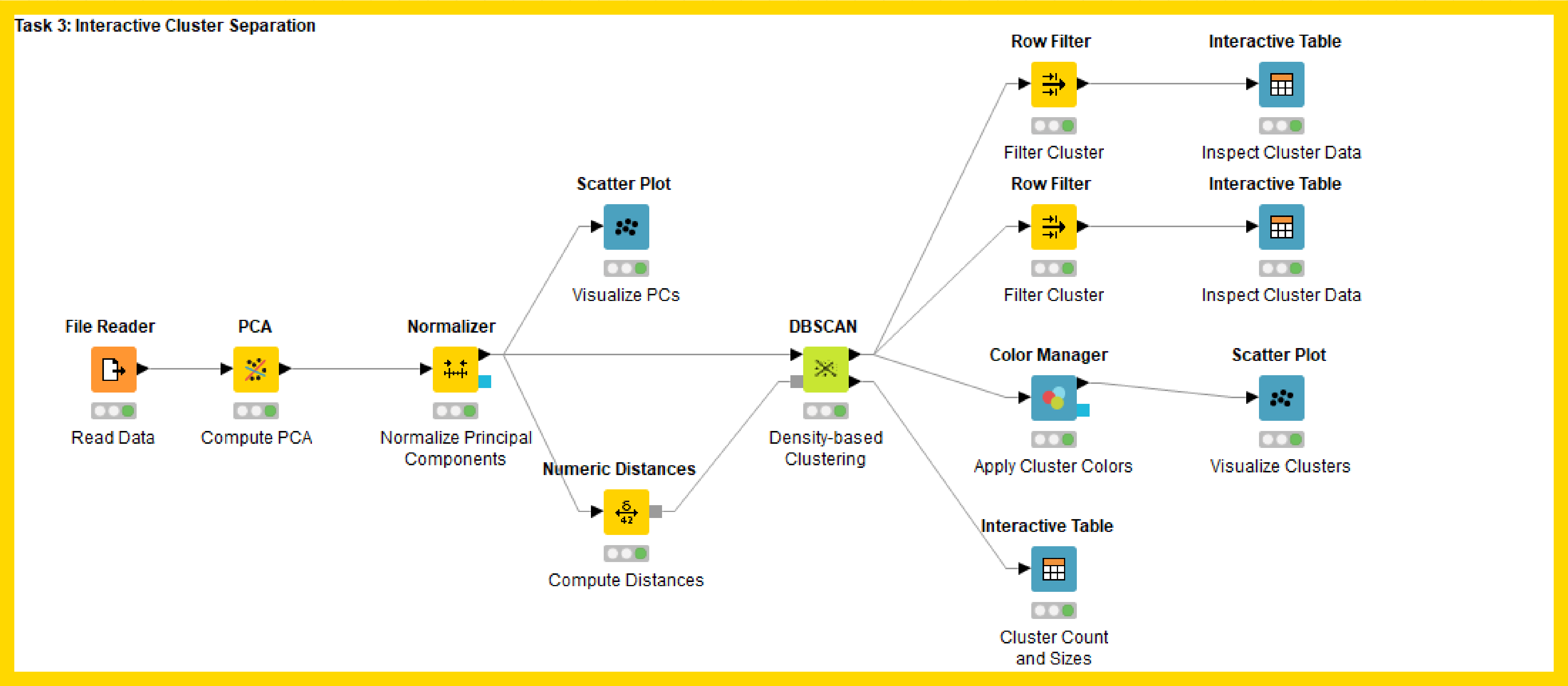
Task 3 – Interactive Cluster Separation

- Use a *better suited* clustering technique

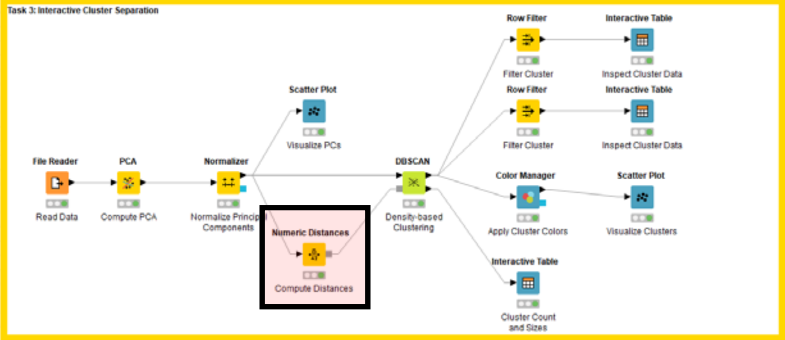
- Assign colors and visualize the cluster memberships

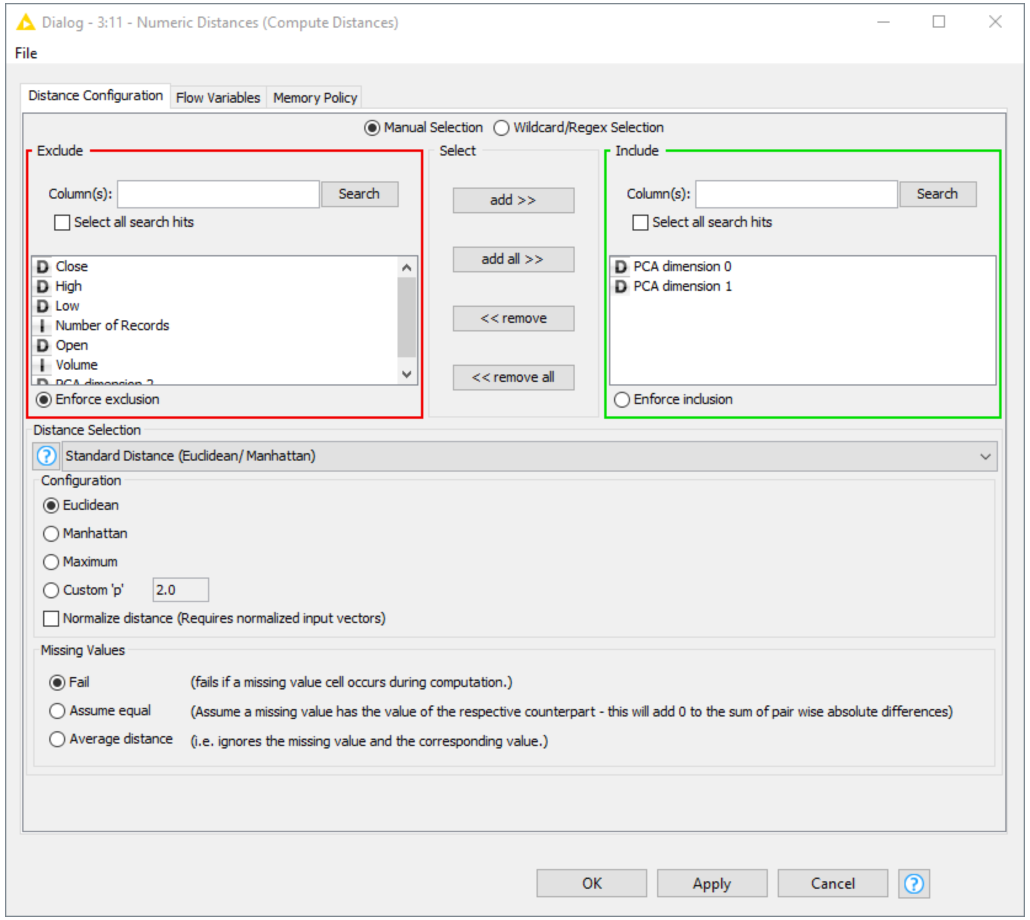
- Open two tables to show the data from at least two different clusters.

Task 3: Knime Workflow



Task 3: Distance Computations





● Manual Selection ○ Wildcard/Regex Selection

Exclude

Column(s): [ ] Search

□ Select all search hits

- D Close
- D High
- D Low
- I Number of Records
- D Open
- I Volume
- D PCA dimension 2

● Enforce exclusion

Select

Include

Column(s): [ ] Search

□ Select all search hits

- D PCA dimension 0
- D PCA dimension 1

○ Enforce inclusion

Distance Selection

Standard Distance (Euclidean/ Manhattan)

Configuration

● Euclidean

○ Manhattan

○ Maximum

○ Custom 'p' 2.0

□ Normalize distance (Requires normalized input vectors)

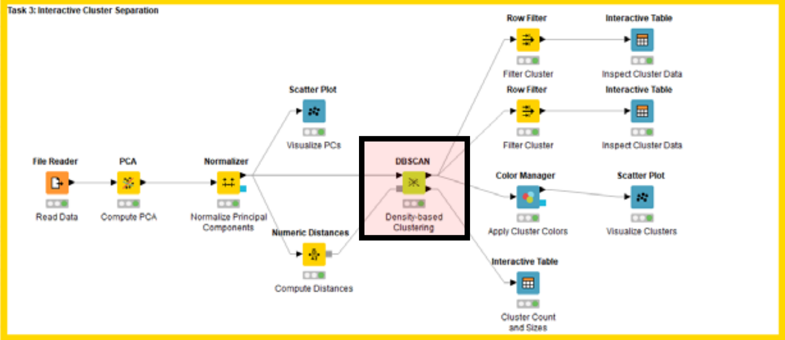
Missing Values

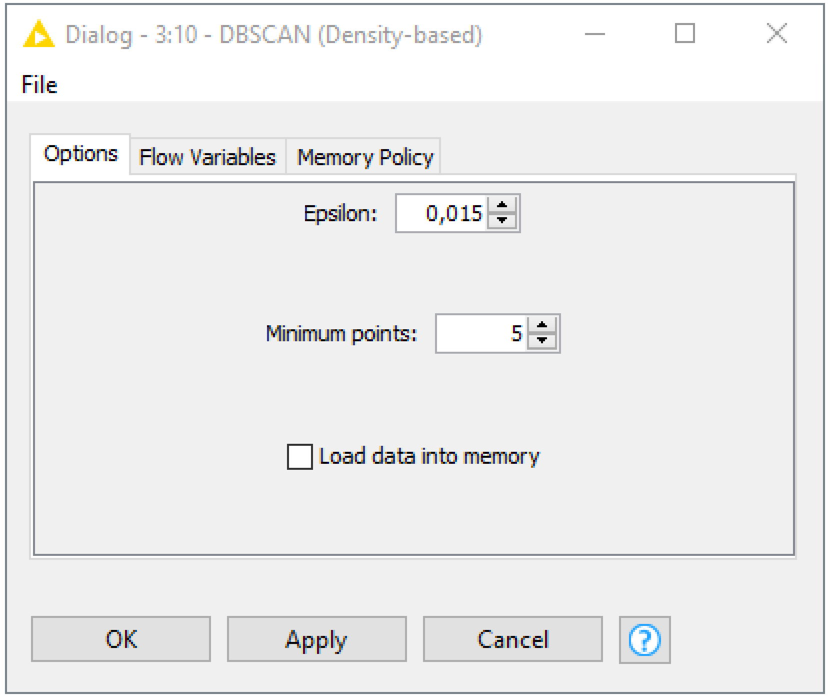
● Fail (fails if a missing value cell occurs during computation.)

○ Assume equal (Assume a missing value has the value of the respective counterpart - this will add 0 to the sum of pair wise absolute differences)

○ Average distance (i.e. ignores the missing value and the corresponding value.)

Task 3: DBScan Clustering



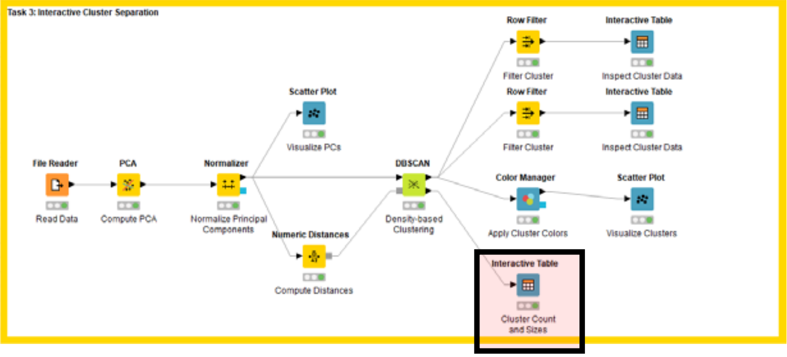


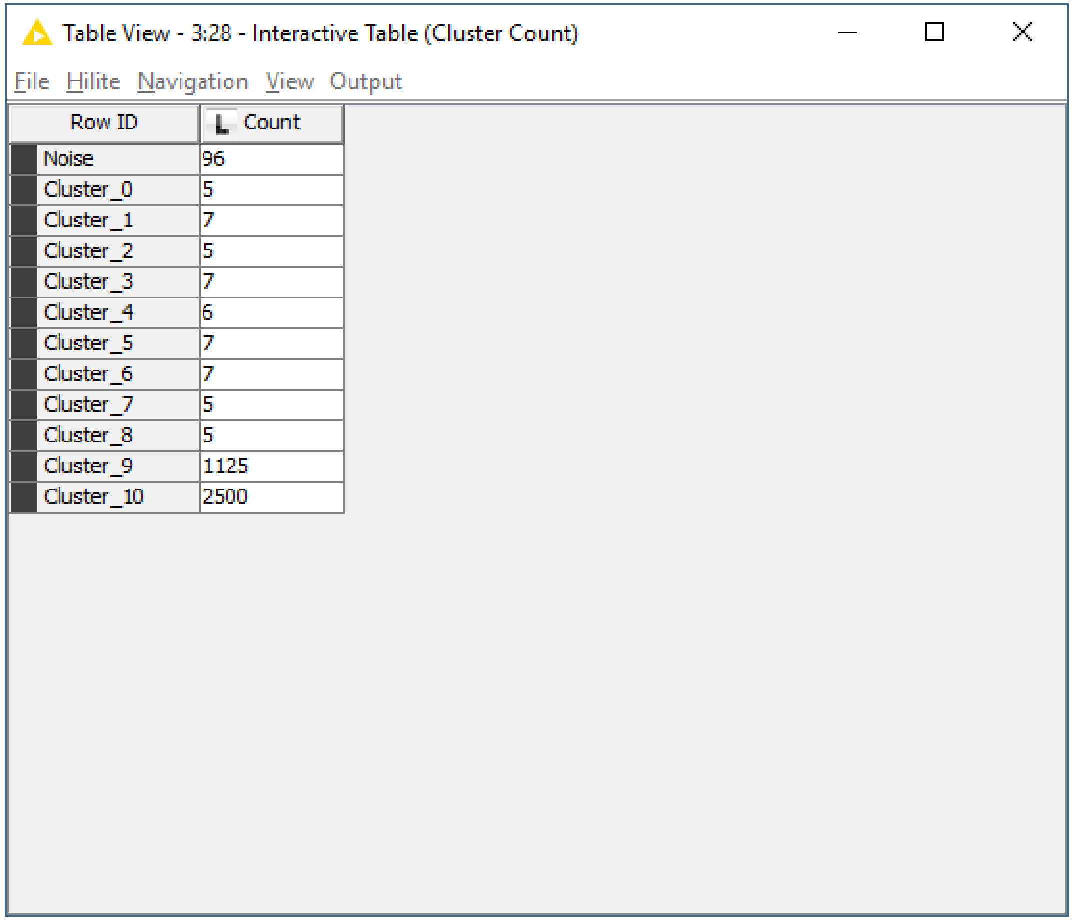
Epsilon: 0,015

Minimum points: 5

☐ Load data into memory

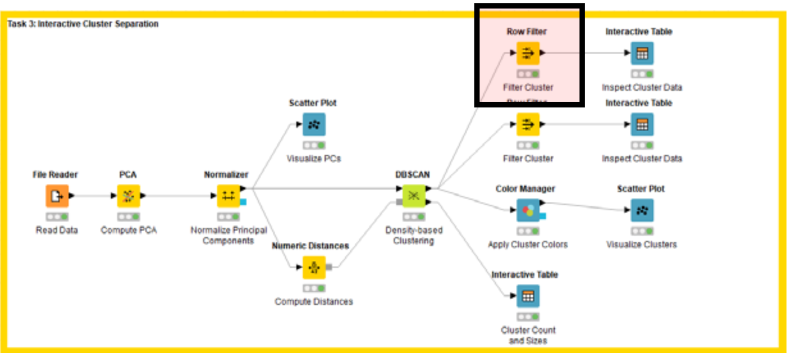
Task 3: DBScan Meta Data

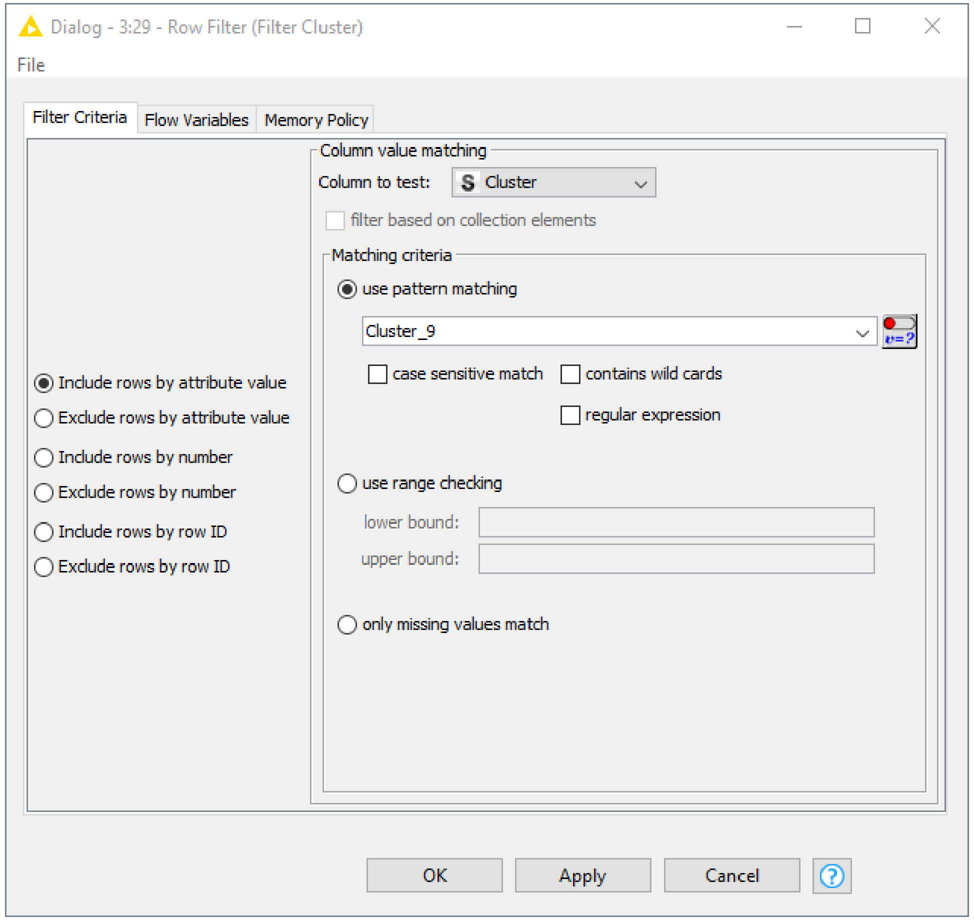




Row ID	Count
Noise	96
Cluster_0	5
Cluster_1	7
Cluster_2	5
Cluster_3	7
Cluster_4	6
Cluster_5	7
Cluster_6	7
Cluster_7	5
Cluster_8	5
Cluster_9	1125
Cluster_10	2500

Task 3: Select Instances of Cluster 9





● Include rows by attribute value
○ Exclude rows by attribute value
○ Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Column value matching

Column to test: S Cluster

□ filter based on collection elements

Matching criteria

● use pattern matching

Cluster_9

□ case sensitive match □ contains wild cards

□ regular expression

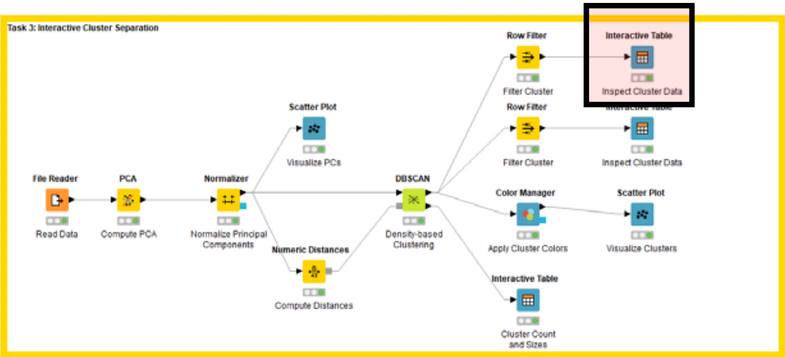
○ use range checking

lower bound:

upper bound:

○ only missing values match

Task 3: Inspect Instances of Cluster 9



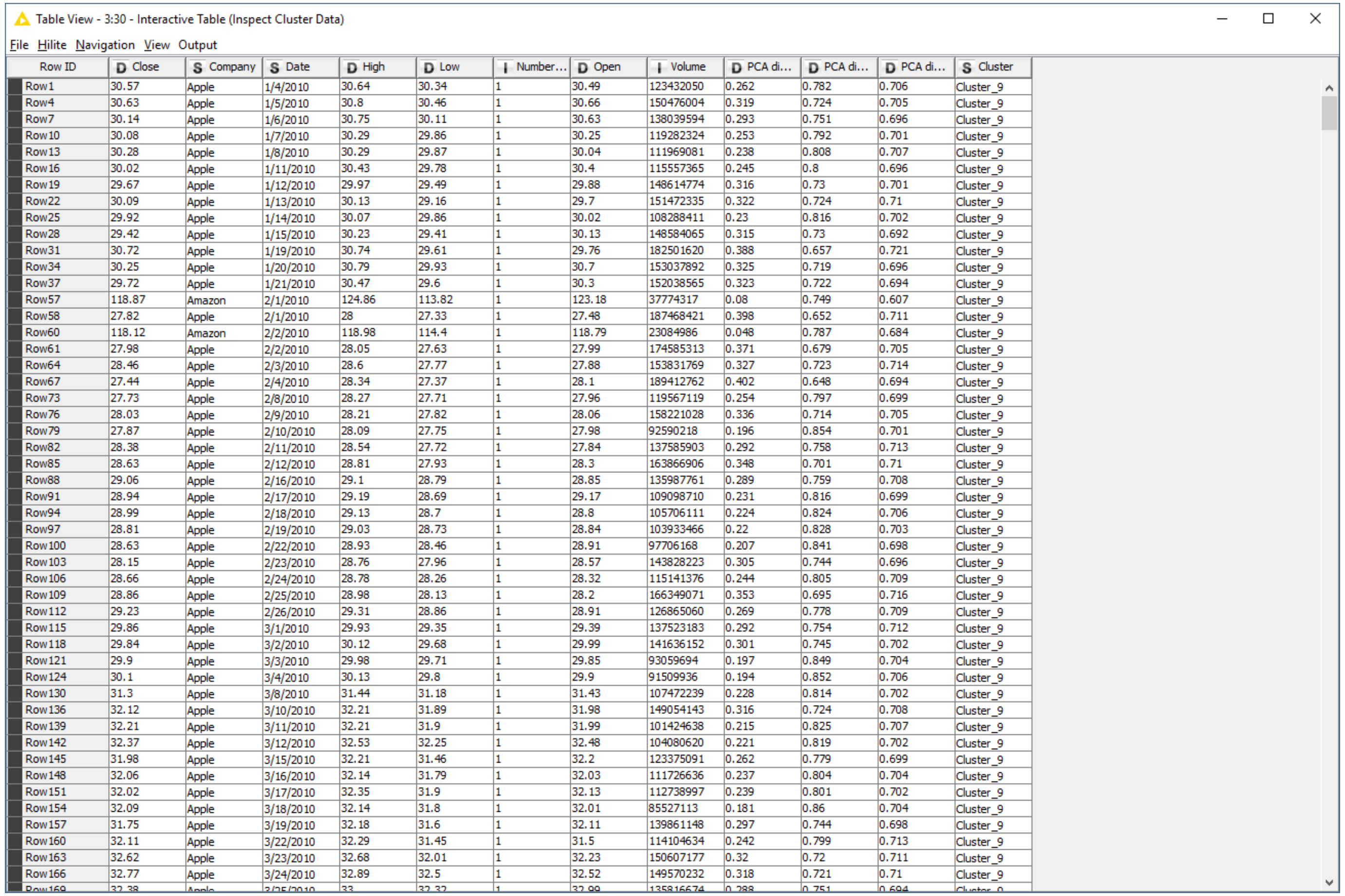


Table View - 3:30 - Interactive Table (Inspect Cluster Data)

Row ID	Close	Company	Date	High	Low	Number...	Open	Volume	PCA di...	PCA di...	PCA di...	Cluster
Row1	30.57	Apple	1/4/2010	30.64	30.34	1	30.49	123432050	0.262	0.782	0.706	Cluster_9
Row4	30.63	Apple	1/5/2010	30.8	30.46	1	30.66	150476004	0.319	0.724	0.705	Cluster_9
Row7	30.14	Apple	1/6/2010	30.75	30.11	1	30.63	138039594	0.293	0.751	0.696	Cluster_9
Row10	30.08	Apple	1/7/2010	30.29	29.86	1	30.25	119282324	0.253	0.792	0.701	Cluster_9
Row13	30.28	Apple	1/8/2010	30.29	29.87	1	30.04	111969081	0.238	0.808	0.707	Cluster_9
Row16	30.02	Apple	1/11/2010	30.43	29.78	1	30.4	115557365	0.245	0.8	0.696	Cluster_9
Row19	29.67	Apple	1/12/2010	29.97	29.49	1	29.88	148614774	0.316	0.73	0.701	Cluster_9
Row22	30.09	Apple	1/13/2010	30.13	29.16	1	29.7	151472335	0.322	0.724	0.71	Cluster_9
Row25	29.92	Apple	1/14/2010	30.07	29.86	1	30.02	108288411	0.23	0.816	0.702	Cluster_9
Row28	29.42	Apple	1/15/2010	30.23	29.41	1	30.13	148584065	0.315	0.73	0.692	Cluster_9
Row31	30.72	Apple	1/19/2010	30.74	29.61	1	29.76	182501620	0.388	0.657	0.721	Cluster_9
Row34	30.25	Apple	1/20/2010	30.79	29.93	1	30.7	153037892	0.325	0.719	0.696	Cluster_9
Row37	29.72	Apple	1/21/2010	30.47	29.6	1	30.3	152038565	0.323	0.722	0.694	Cluster_9
Row57	118.87	Amazon	2/1/2010	124.86	113.82	1	123.18	37774317	0.08	0.749	0.607	Cluster_9
Row58	27.82	Apple	2/1/2010	28	27.33	1	27.48	187468421	0.398	0.652	0.711	Cluster_9
Row60	118.12	Amazon	2/2/2010	118.98	114.4	1	118.79	23084986	0.048	0.787	0.684	Cluster_9
Row61	27.98	Apple	2/2/2010	28.05	27.63	1	27.99	174585313	0.371	0.679	0.705	Cluster_9
Row64	28.46	Apple	2/3/2010	28.6	27.77	1	27.88	153831769	0.327	0.723	0.714	Cluster_9
Row67	27.44	Apple	2/4/2010	28.34	27.37	1	28.1	189412762	0.402	0.648	0.694	Cluster_9
Row73	27.73	Apple	2/8/2010	28.27	27.71	1	27.96	119567119	0.254	0.797	0.699	Cluster_9
Row76	28.03	Apple	2/9/2010	28.21	27.82	1	28.06	158221028	0.336	0.714	0.705	Cluster_9
Row79	27.87	Apple	2/10/2010	28.09	27.75	1	27.98	92590218	0.196	0.854	0.701	Cluster_9
Row82	28.38	Apple	2/11/2010	28.54	27.72	1	27.84	137585903	0.292	0.758	0.713	Cluster_9
Row85	28.63	Apple	2/12/2010	28.81	27.93	1	28.3	163866906	0.348	0.701	0.71	Cluster_9
Row88	29.06	Apple	2/16/2010	29.1	28.79	1	28.85	135987761	0.289	0.759	0.708	Cluster_9
Row91	28.94	Apple	2/17/2010	29.19	28.69	1	29.17	109098710	0.231	0.816	0.699	Cluster_9
Row94	28.99	Apple	2/18/2010	29.13	28.7	1	28.8	105706111	0.224	0.824	0.706	Cluster_9
Row97	28.81	Apple	2/19/2010	29.03	28.73	1	28.84	103933466	0.22	0.828	0.703	Cluster_9
Row100	28.63	Apple	2/22/2010	28.93	28.46	1	28.91	97706168	0.207	0.841	0.698	Cluster_9
Row103	28.15	Apple	2/23/2010	28.76	27.96	1	28.57	143828223	0.305	0.744	0.696	Cluster_9
Row106	28.66	Apple	2/24/2010	28.78	28.26	1	28.32	115141376	0.244	0.805	0.709	Cluster_9
Row109	28.86	Apple	2/25/2010	28.98	28.13	1	28.2	166349071	0.353	0.695	0.716	Cluster_9
Row112	29.23	Apple	2/26/2010	29.31	28.86	1	28.91	126865060	0.269	0.778	0.709	Cluster_9
Row115	29.86	Apple	3/1/2010	29.93	29.35	1	29.39	137523183	0.292	0.754	0.712	Cluster_9
Row118	29.84	Apple	3/2/2010	30.12	29.68	1	29.99	141636152	0.301	0.745	0.702	Cluster_9
Row121	29.9	Apple	3/3/2010	29.98	29.71	1	29.85	93059694	0.197	0.849	0.704	Cluster_9
Row124	30.1	Apple	3/4/2010	30.13	29.8	1	29.9	91509936	0.194	0.852	0.706	Cluster_9
Row130	31.3	Apple	3/8/2010	31.44	31.18	1	31.43	107472239	0.228	0.814	0.702	Cluster_9
Row136	32.12	Apple	3/10/2010	32.21	31.89	1	31.98	149054143	0.316	0.724	0.708	Cluster_9
Row139	32.21	Apple	3/11/2010	32.21	31.9	1	31.99	101424638	0.215	0.825	0.707	Cluster_9
Row142	32.37	Apple	3/12/2010	32.53	32.25	1	32.48	104080620	0.221	0.819	0.702	Cluster_9
Row145	31.98	Apple	3/15/2010	32.21	31.46	1	32.2	123375091	0.262	0.779	0.699	Cluster_9
Row148	32.06	Apple	3/16/2010	32.14	31.79	1	32.03	111726636	0.237	0.804	0.704	Cluster_9
Row151	32.02	Apple	3/17/2010	32.35	31.9	1	32.13	112738997	0.239	0.801	0.702	Cluster_9
Row154	32.09	Apple	3/18/2010	32.14	31.8	1	32.01	85527113	0.181	0.86	0.704	Cluster_9
Row157	31.75	Apple	3/19/2010	32.18	31.6	1	32.11	139861148	0.297	0.744	0.698	Cluster_9
Row160	32.11	Apple	3/22/2010	32.29	31.45	1	31.5	114104634	0.242	0.799	0.713	Cluster_9
Row163	32.62	Apple	3/23/2010	32.68	32.01	1	32.23	150607177	0.32	0.72	0.711	Cluster_9
Row166	32.77	Apple	3/24/2010	32.89	32.5	1	32.52	149570232	0.318	0.721	0.71	Cluster_9
Row169	32.38	Apple	3/25/2010	33	32.32	1	32.99	135816674	0.288	0.751	0.694	Cluster_9