INSURANCE PREMIUM ANALYSIS USING KNIME

Project Guide:-
SAYAN CHAKBRABOTY

Presented by

1 RIAJ IKBAL

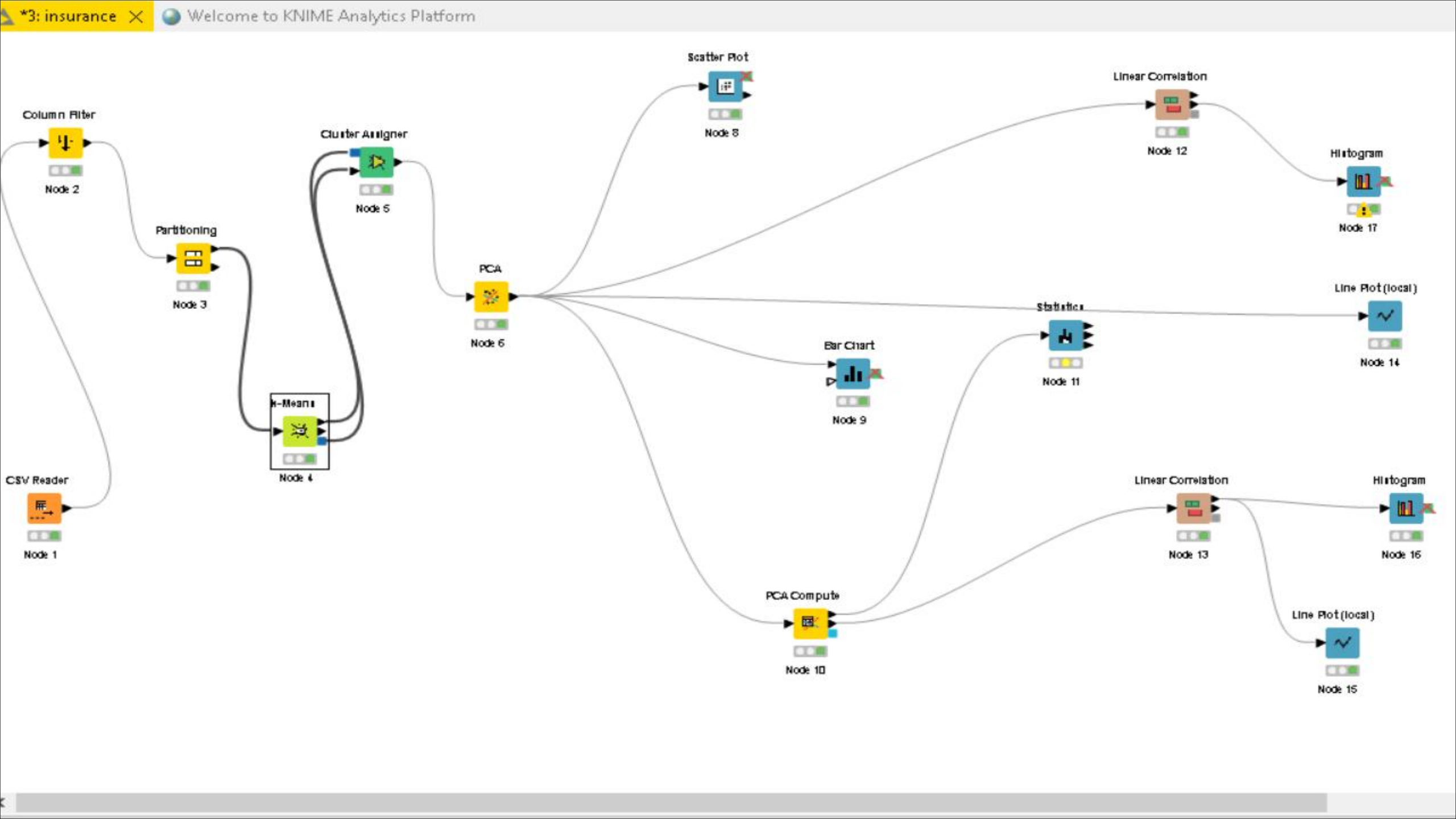
2 33200121135

3 BTECH | CSE | 3RD YEAR

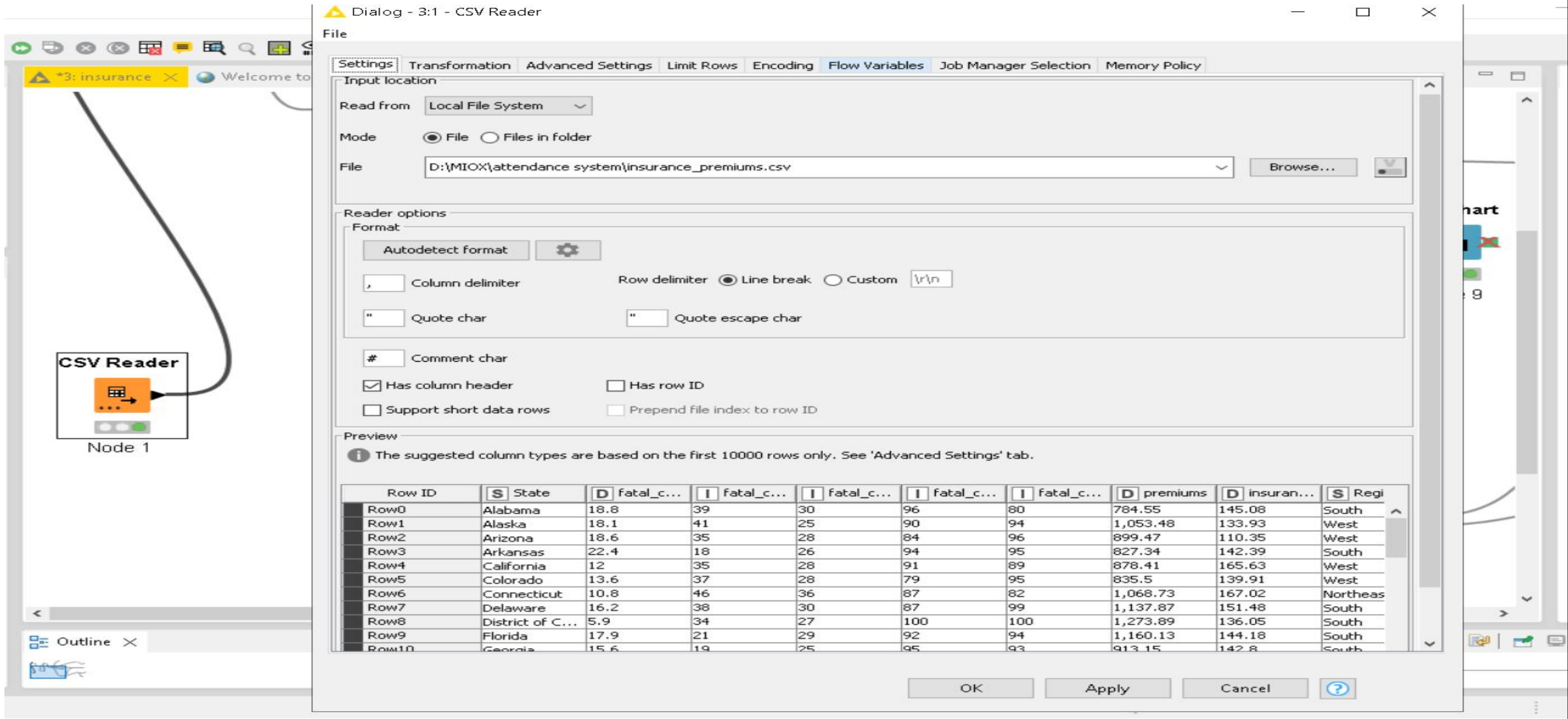
4 FROM TECHNO INTERNATIONAL BATANAGAR

motive

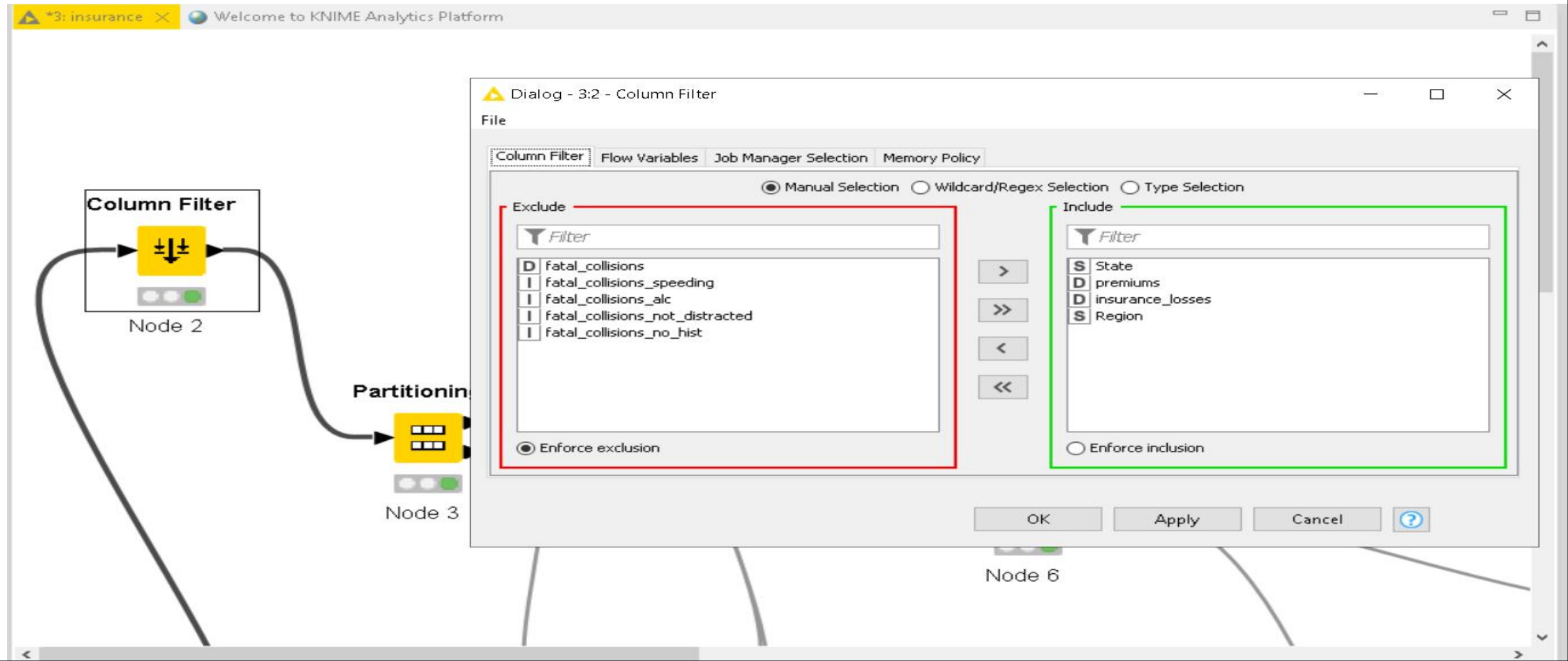
The loss ratio is constructed specifically to analyze the operation of an insurance company.


*3: insurance
Welcome to KNIME Analytics Platform
Column Filter
Node 2
Partitioning
Node 3
Cluster Assigner
Node 5
Scatter Plot
Node 8
Linear Correlation
Node 12
Histogram
Node 17
PCA
Node 6
Line Plot (local)
Node 14
Statistics
Node 11
Bar Chart
Node 9
k-Means
Node 4
CSV Reader
Node 1
Linear Correlation
Node 13
Histogram
Node 16
PCA Compute
Node 10
Line Plot (local)
Node 15

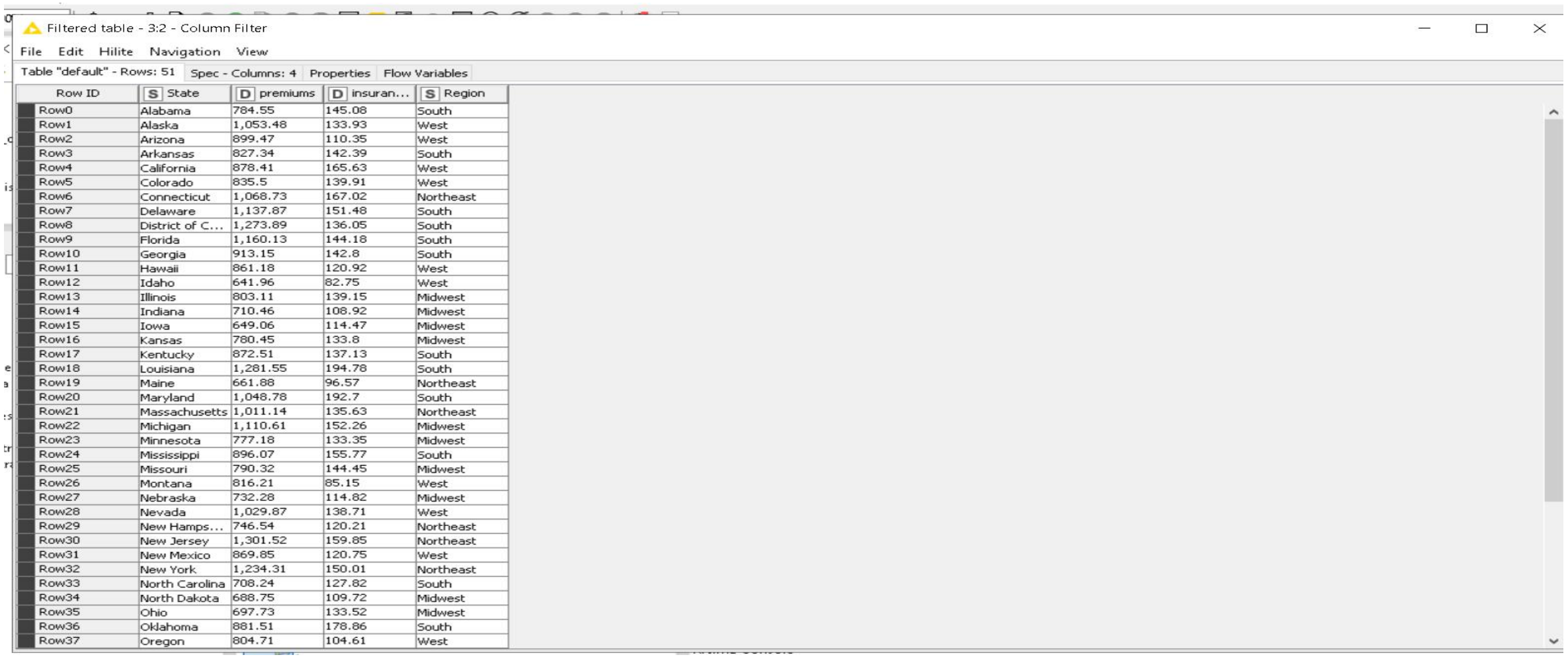
At first we need to read the data set so here i used csv reader



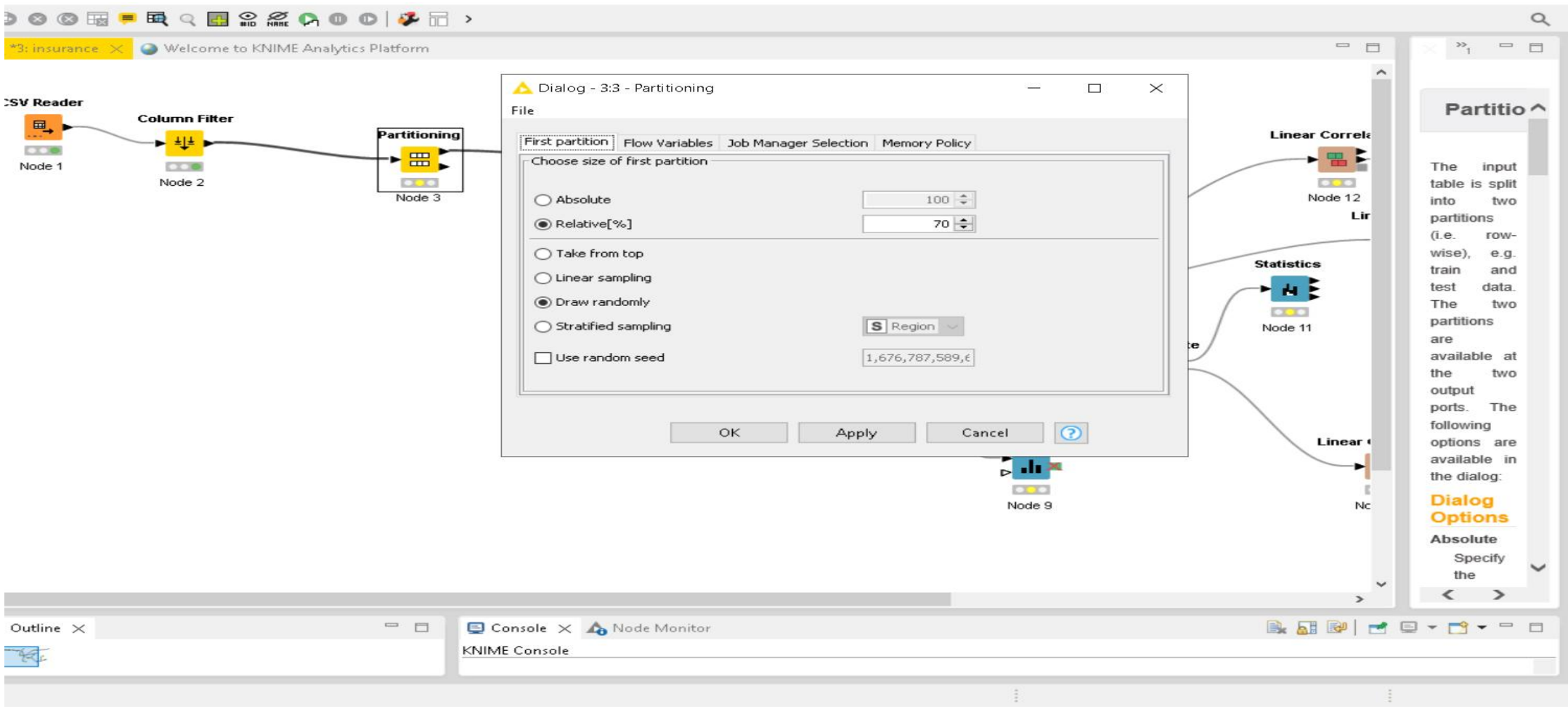
now here i have filtered the column accordingly and removed the unnecessary columns to the right side



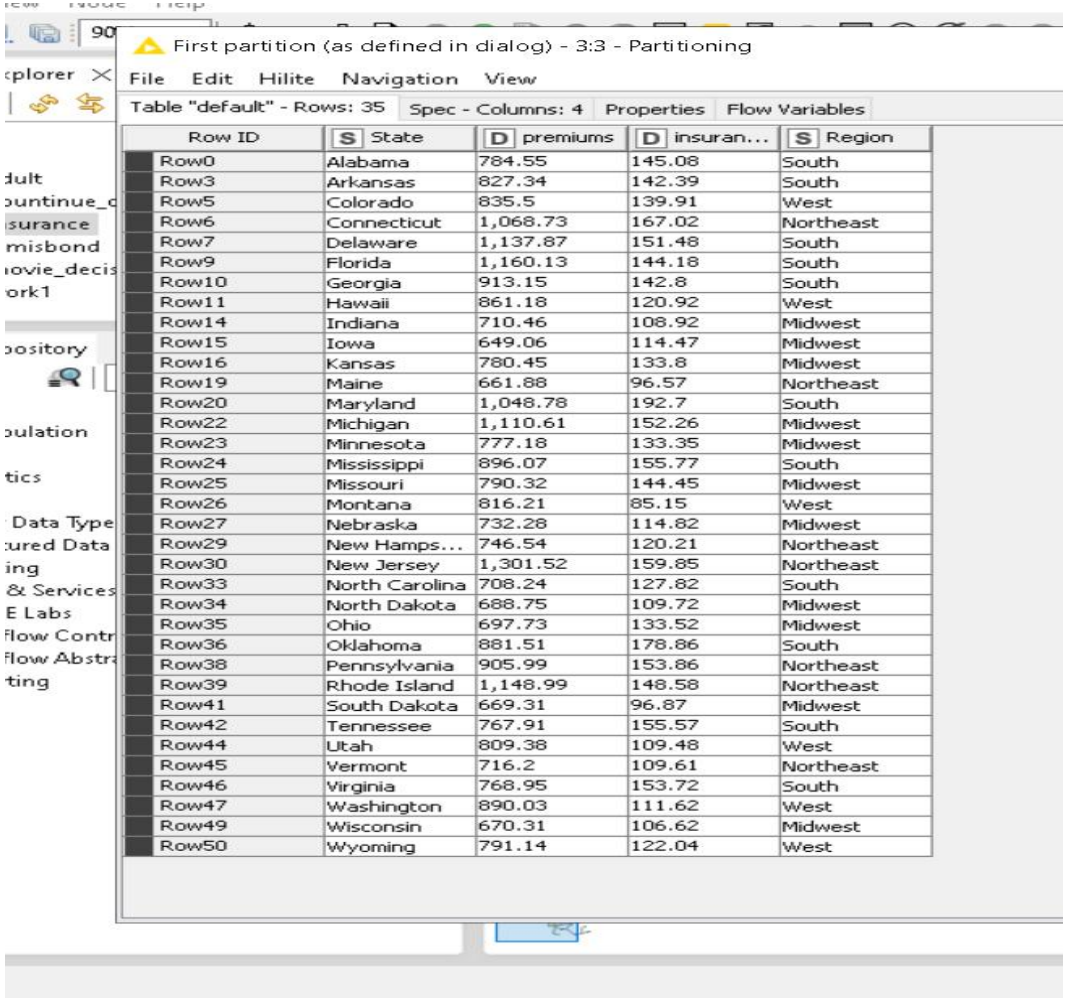
here’s the filtered table



Row ID	State	premiums	insuran...	Region
Row0	Alabama	784.55	145.08	South
Row1	Alaska	1,053.48	133.93	West
Row2	Arizona	899.47	110.35	West
Row3	Arkansas	827.34	142.39	South
Row4	California	878.41	165.63	West
Row5	Colorado	835.5	139.91	West
Row6	Connecticut	1,068.73	167.02	Northeast
Row7	Delaware	1,137.87	151.48	South
Row8	District of C...	1,273.89	136.05	South
Row9	Florida	1,160.13	144.18	South
Row10	Georgia	913.15	142.8	South
Row11	Hawaii	861.18	120.92	West
Row12	Idaho	641.96	82.75	West
Row13	Illinois	803.11	139.15	Midwest
Row14	Indiana	710.46	108.92	Midwest
Row15	Iowa	649.06	114.47	Midwest
Row16	Kansas	780.45	133.8	Midwest
Row17	Kentucky	872.51	137.13	South
Row18	Louisiana	1,281.55	194.78	South
Row19	Maine	661.88	96.57	Northeast
Row20	Maryland	1,048.78	192.7	South
Row21	Massachusetts	1,011.14	135.63	Northeast
Row22	Michigan	1,110.61	152.26	Midwest
Row23	Minnesota	777.18	133.35	Midwest
Row24	Mississippi	896.07	155.77	South
Row25	Missouri	790.32	144.45	Midwest
Row26	Montana	816.21	85.15	West
Row27	Nebraska	732.28	114.82	Midwest
Row28	Nevada	1,029.87	138.71	West
Row29	New Hamps...	746.54	120.21	Northeast
Row30	New Jersey	1,301.52	159.85	Northeast
Row31	New Mexico	869.85	120.75	West
Row32	New York	1,234.31	150.01	Northeast
Row33	North Carolina	708.24	127.82	South
Row34	North Dakota	688.75	109.72	Midwest
Row35	Ohio	697.73	133.52	Midwest
Row36	Oklahoma	881.51	178.86	South
Row37	Oregon	804.71	104.61	West

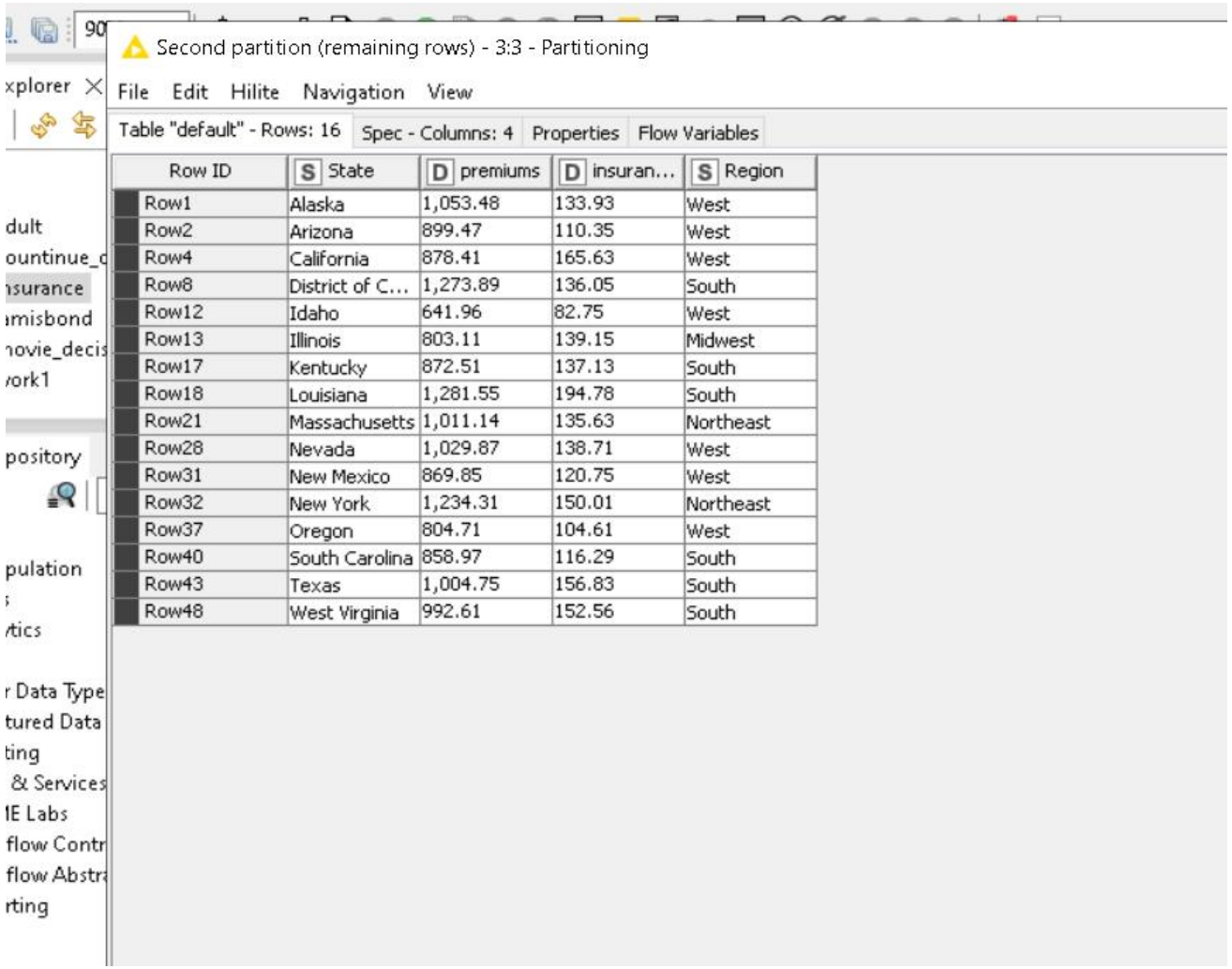

*3: insurance
Welcome to KNIME Analytics Platform
CSV Reader
Node 1
Column Filter
Node 2
Partitioning
Node 3
Dialog - 3:3 - Partitioning
File
First partition
Flow Variables
Job Manager Selection
Memory Policy
Choose size of first partition
Absolute
100
Relative[%]
70
Take from top
Linear sampling
Draw randomly
Stratified sampling
Region
Use random seed
1,676,787,589,€
OK
Apply
Cancel
Node 9
Linear Correla
Node 12
Statistics
Node 11
Partitio
The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports. The following options are available in the dialog:
Dialog Options
Absolute
Specify the
Outline
Console
Node Monitor
KNIME Console

first partition



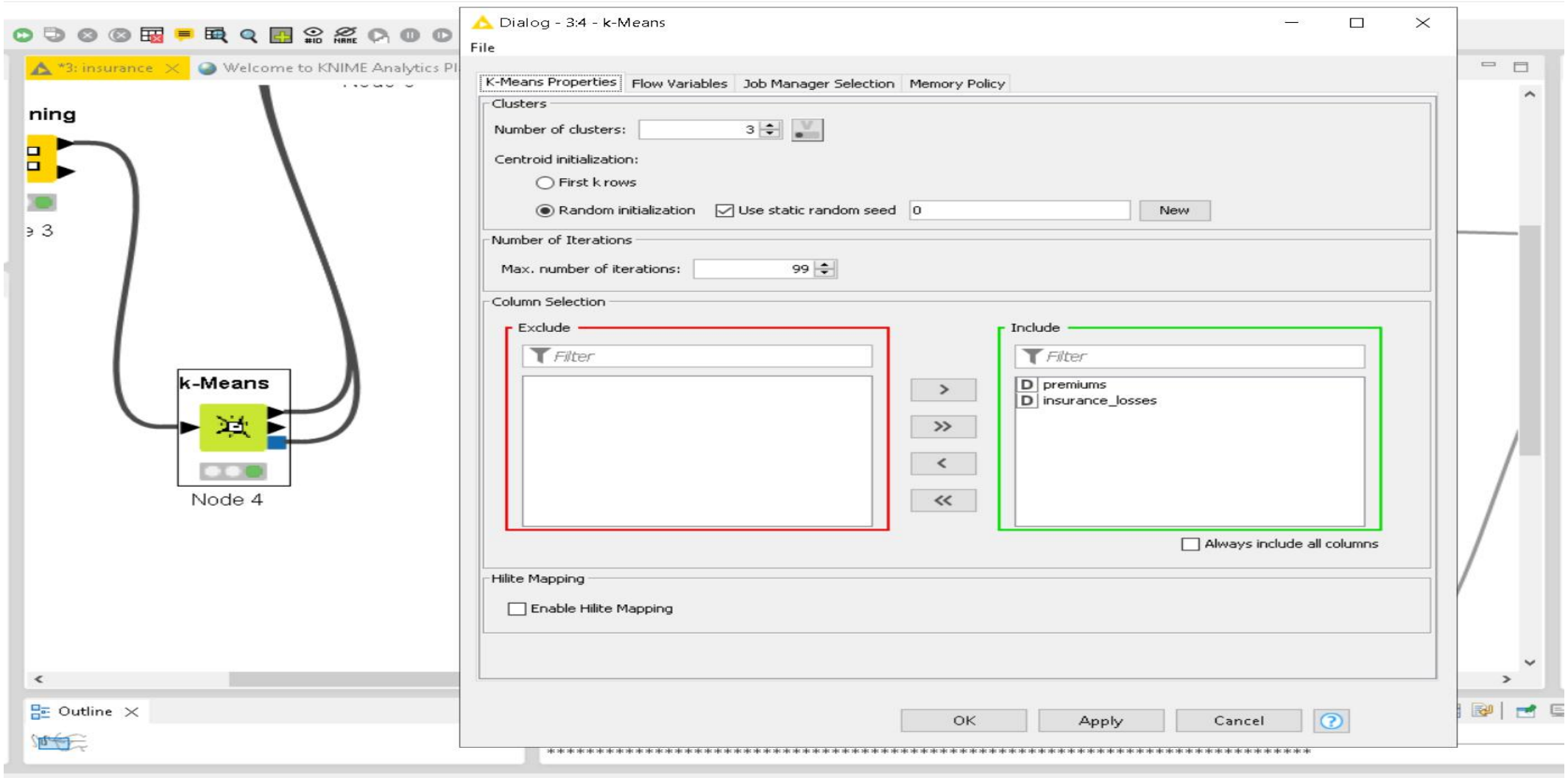
Row ID	State	premiums	insuran...	Region
Row0	Alabama	784.55	145.08	South
Row3	Arkansas	827.34	142.39	South
Row5	Colorado	835.5	139.91	West
Row6	Connecticut	1,068.73	167.02	Northeast
Row7	Delaware	1,137.87	151.48	South
Row9	Florida	1,160.13	144.18	South
Row10	Georgia	913.15	142.8	South
Row11	Hawaii	861.18	120.92	West
Row14	Indiana	710.46	108.92	Midwest
Row15	Iowa	649.06	114.47	Midwest
Row16	Kansas	780.45	133.8	Midwest
Row19	Maine	661.88	96.57	Northeast
Row20	Maryland	1,048.78	192.7	South
Row22	Michigan	1,110.61	152.26	Midwest
Row23	Minnesota	777.18	133.35	Midwest
Row24	Mississippi	896.07	155.77	South
Row25	Missouri	790.32	144.45	Midwest
Row26	Montana	816.21	85.15	West
Row27	Nebraska	732.28	114.82	Midwest
Row29	New Hamps...	746.54	120.21	Northeast
Row30	New Jersey	1,301.52	159.85	Northeast
Row33	North Carolina	708.24	127.82	South
Row34	North Dakota	688.75	109.72	Midwest
Row35	Ohio	697.73	133.52	Midwest
Row36	Oklahoma	881.51	178.86	South
Row38	Pennsylvania	905.99	153.86	Northeast
Row39	Rhode Island	1,148.99	148.58	Northeast
Row41	South Dakota	669.31	96.87	Midwest
Row42	Tennessee	767.91	155.57	South
Row44	Utah	809.38	109.48	West
Row45	Vermont	716.2	109.61	Northeast
Row46	Virginia	768.95	153.72	South
Row47	Washington	890.03	111.62	West
Row49	Wisconsin	670.31	106.62	Midwest
Row50	Wyoming	791.14	122.04	West

second partitionm

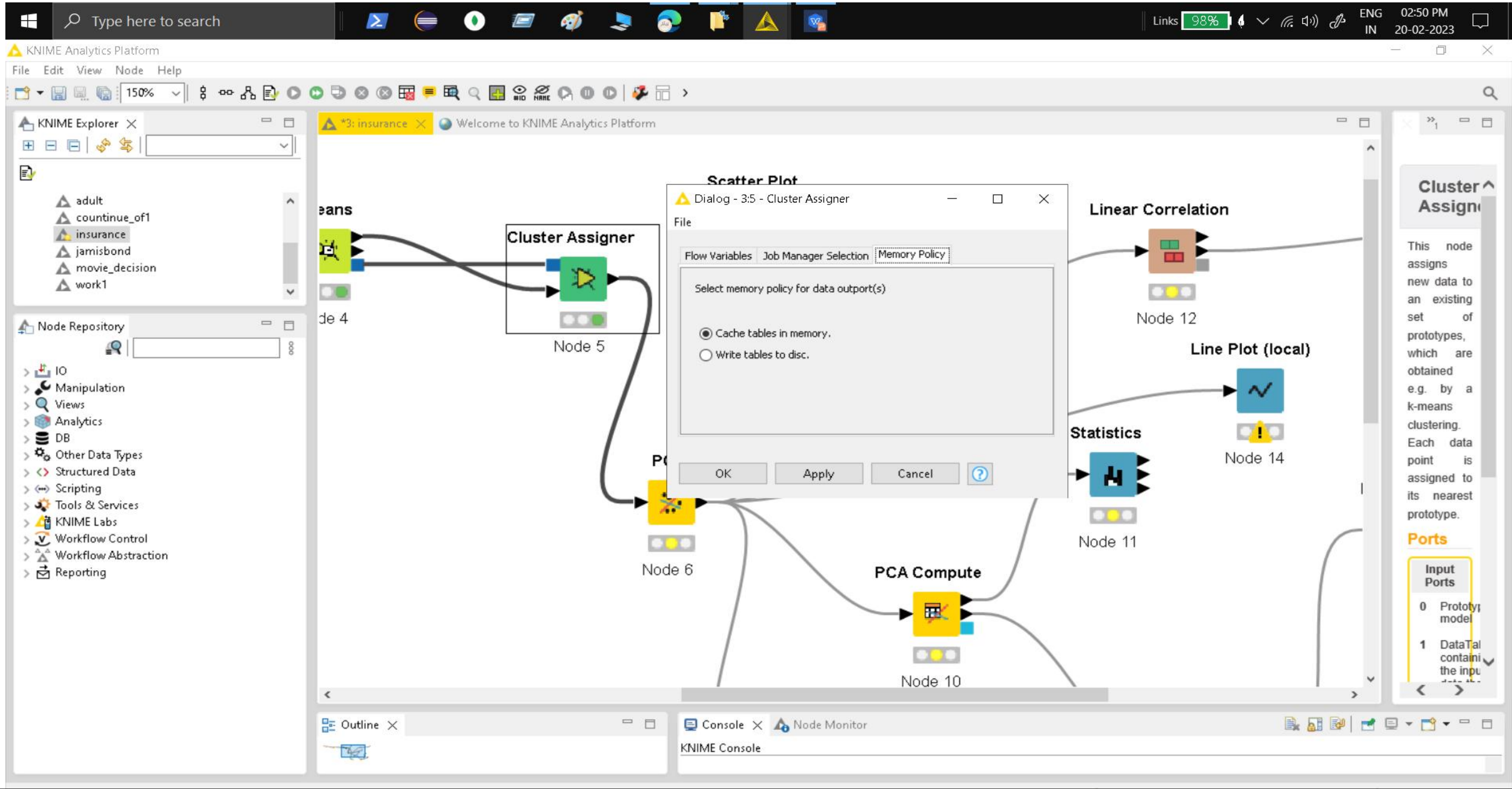


Row ID	State	premiums	insuran...	Region
Row1	Alaska	1,053.48	133.93	West
Row2	Arizona	899.47	110.35	West
Row4	California	878.41	165.63	West
Row8	District of C...	1,273.89	136.05	South
Row12	Idaho	641.96	82.75	West
Row13	Illinois	803.11	139.15	Midwest
Row17	Kentucky	872.51	137.13	South
Row18	Louisiana	1,281.55	194.78	South
Row21	Massachusetts	1,011.14	135.63	Northeast
Row28	Nevada	1,029.87	138.71	West
Row31	New Mexico	869.85	120.75	West
Row32	New York	1,234.31	150.01	Northeast
Row37	Oregon	804.71	104.61	West
Row40	South Carolina	858.97	116.29	South
Row43	Texas	1,004.75	156.83	South
Row48	West Virginia	992.61	152.56	South

K-means for grouping the unlabeled dataset into different clusters.



here clusters are created we can see the coverage, premiums of the insurance , and the loss on the insurance
after this i have used cluster assigner it assigns new data to an existing set of prototypes, which are obtained


Type here to search
Links 98%
ENG IN
02:50 PM 20-02-2023
KNIME Analytics Platform
File Edit View Node Help
150%
KNIME Explorer
adult
countinue_of1
insurance
jamisbond
movie_decision
work1
Node Repository
IO
Manipulation
Views
Analytics
DB
Other Data Types
Structured Data
Scripting
Tools & Services
KNIME Labs
Workflow Control
Workflow Abstraction
Reporting
*3: insurance
Welcome to KNIME Analytics Platform
Scatter Plot
Cluster Assigner
Node 5
Dialog - 3:5 - Cluster Assigner
File
Flow Variables
Job Manager Selection
Memory Policy
Select memory policy for data outport(s)
Cache tables in memory.
Write tables to disc.
OK
Apply
Cancel
Linear Correlation
Node 12
Line Plot (local)
Node 14
Statistics
Node 11
Node 6
PCA Compute
Node 10
Cluster Assign
This node assigns new data to an existing set of prototypes, which are obtained e.g. by a k-means clustering. Each data point is assigned to its nearest prototype.
Ports
Input Ports
0 Prototyp model
1 DataTal containi the inpu
Outline
Console
Node Monitor
KNIME Console

the new assigned data

Home Insert Design Transitions Animation Slide Show Review View

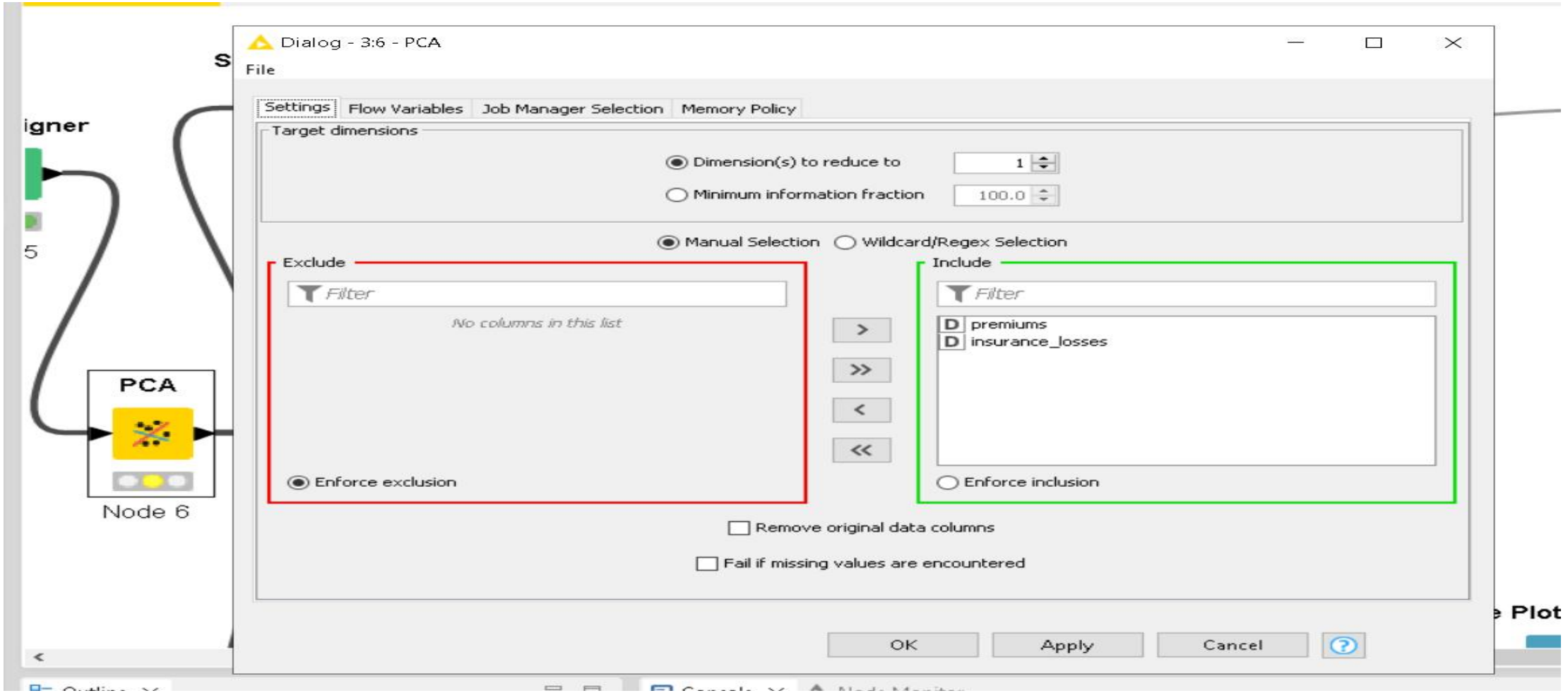
Assigned Data - 3:5 - Cluster Assigner

File Edit Hilite Navigation View

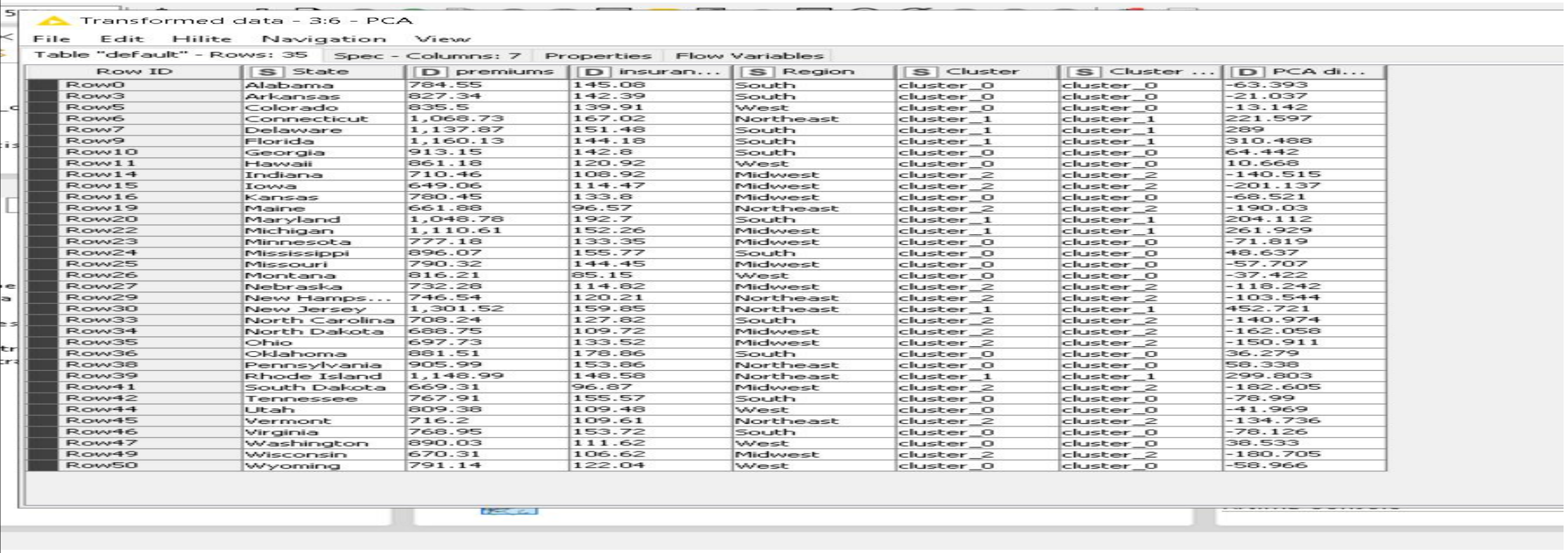
Table "default" - Rows: 35 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	State	premiums	insuran...	Region	Cluster	Cluster ...
Row0	Alabama	784.55	145.08	South	cluster_0	cluster_0
Row3	Arkansas	827.34	142.39	South	cluster_0	cluster_0
Row5	Colorado	835.5	139.91	West	cluster_0	cluster_0
Row6	Connecticut	1,068.73	167.02	Northeast	cluster_1	cluster_1
Row7	Delaware	1,137.87	151.48	South	cluster_1	cluster_1
Row9	Florida	1,160.13	144.18	South	cluster_1	cluster_1
Row10	Georgia	913.15	142.8	South	cluster_0	cluster_0
Row11	Hawaii	861.18	120.92	West	cluster_0	cluster_0
Row14	Indiana	710.46	108.92	Midwest	cluster_2	cluster_2
Row15	Iowa	649.06	114.47	Midwest	cluster_2	cluster_2
Row16	Kansas	780.45	133.8	Midwest	cluster_0	cluster_0
Row19	Maine	661.88	96.57	Northeast	cluster_2	cluster_2
Row20	Maryland	1,048.78	192.7	South	cluster_1	cluster_1
Row22	Michigan	1,110.61	152.26	Midwest	cluster_1	cluster_1
Row23	Minnesota	777.18	133.35	Midwest	cluster_0	cluster_0
Row24	Mississippi	896.07	155.77	South	cluster_0	cluster_0
Row25	Missouri	790.32	144.45	Midwest	cluster_0	cluster_0
Row26	Montana	816.21	85.15	West	cluster_0	cluster_0
Row27	Nebraska	732.28	114.82	Midwest	cluster_2	cluster_2
Row29	New Hamps...	746.54	120.21	Northeast	cluster_2	cluster_2
Row30	New Jersey	1,301.52	159.85	Northeast	cluster_1	cluster_1
Row33	North Carolina	708.24	127.82	South	cluster_2	cluster_2
Row34	North Dakota	688.75	109.72	Midwest	cluster_2	cluster_2
Row35	Ohio	697.73	133.52	Midwest	cluster_2	cluster_2
Row36	Oklahoma	881.51	178.86	South	cluster_0	cluster_0
Row38	Pennsylvania	905.99	153.86	Northeast	cluster_0	cluster_0
Row39	Rhode Island	1,148.99	148.58	Northeast	cluster_1	cluster_1
Row41	South Dakota	669.31	96.87	Midwest	cluster_2	cluster_2
Row42	Tennessee	767.91	155.57	South	cluster_0	cluster_0
Row44	Utah	809.38	109.48	West	cluster_0	cluster_0
Row45	Vermont	716.2	109.61	Northeast	cluster_2	cluster_2
Row46	Virginia	768.95	153.72	South	cluster_0	cluster_0
Row47	Washington	890.03	111.62	West	cluster_0	cluster_0
Row49	Wisconsin	670.31	106.62	Midwest	cluster_2	cluster_2
Row50	Wyoming	791.14	122.04	West	cluster_0	cluster_0

This node performs a principal component analysis (PCA) on the given data.
The input data is projected from its original feature space into a space of (possibly) lower dimension
with a minimum of information loss.



the transformed data



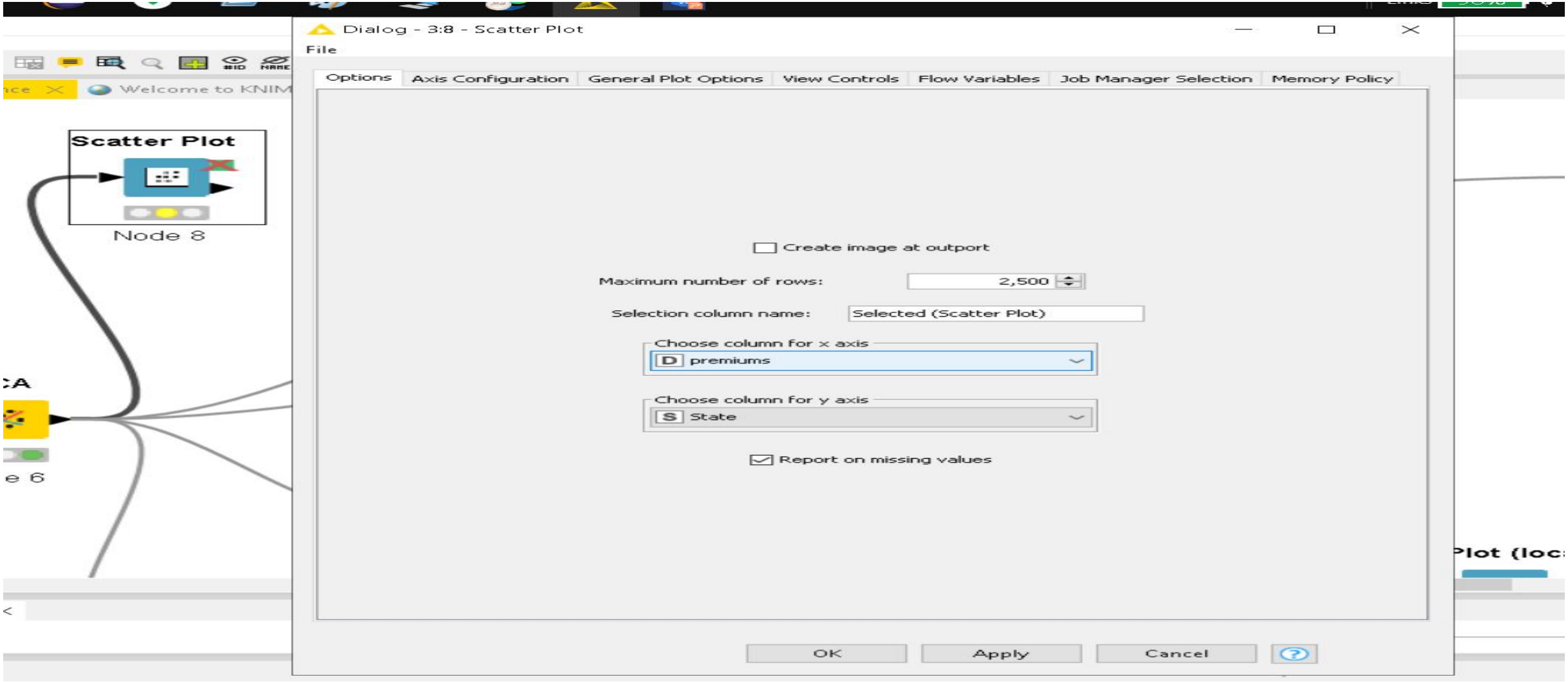
Transformed data - 3:6 - PCA

File Edit Hilite Navigation View

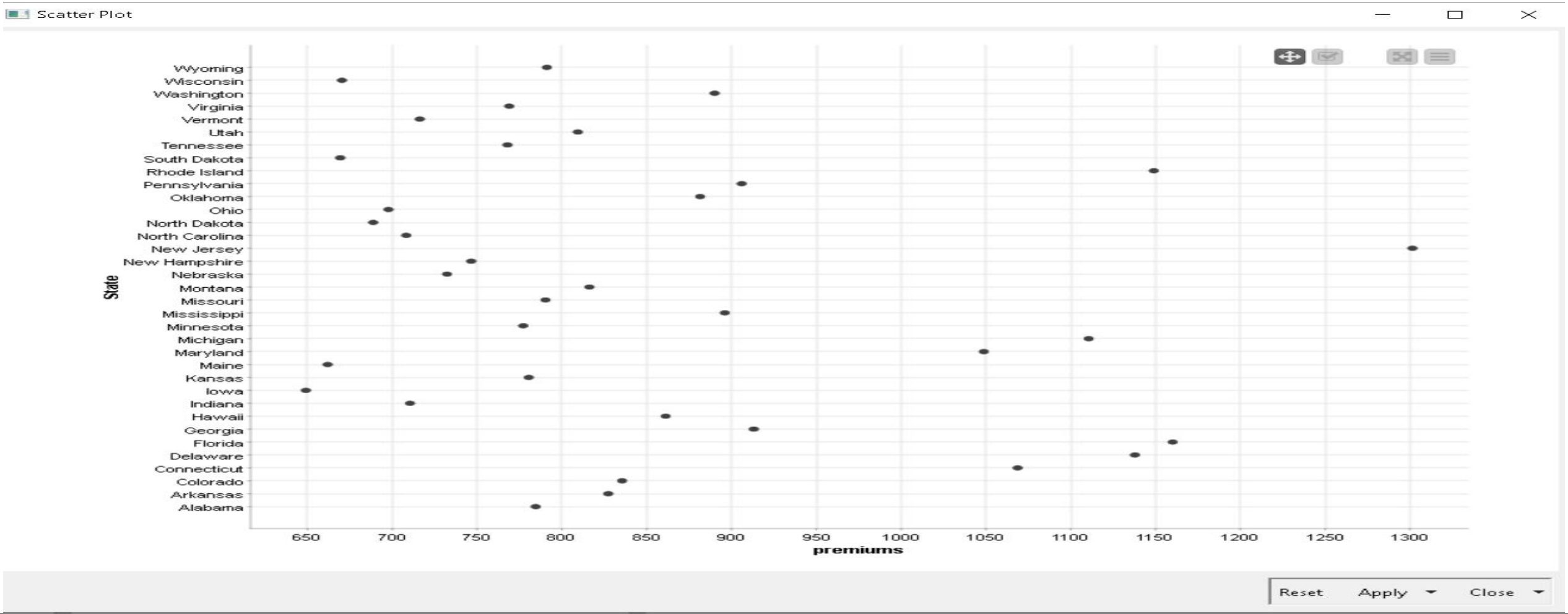
Table "default" - Rows: 35 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	State	premiums	insuran...	Region	Cluster	Cluster ...	PCA di...
Row0	Alabama	784.55	145.08	South	cluster_0	cluster_0	-63.393
Row3	Arkansas	827.34	142.39	South	cluster_0	cluster_0	-21.037
Row5	Colorado	835.5	139.91	West	cluster_0	cluster_0	-13.142
Row6	Connecticut	1,068.73	167.02	Northeast	cluster_1	cluster_1	221.597
Row7	Delaware	1,137.87	151.48	South	cluster_1	cluster_1	289
Row9	Florida	1,160.13	144.18	South	cluster_1	cluster_1	310.488
Row10	Georgia	913.15	142.8	South	cluster_0	cluster_0	64.442
Row11	Hawaii	861.18	120.92	West	cluster_0	cluster_0	10.668
Row14	Indiana	710.46	108.92	Midwest	cluster_2	cluster_2	-140.515
Row15	Iowa	649.06	114.47	Midwest	cluster_2	cluster_2	-201.137
Row16	Kansas	780.45	133.8	Midwest	cluster_0	cluster_0	-68.521
Row19	Maine	661.88	96.57	Northeast	cluster_2	cluster_2	-190.03
Row20	Maryland	1,048.78	192.7	South	cluster_1	cluster_1	204.112
Row22	Michigan	1,110.61	152.26	Midwest	cluster_1	cluster_1	261.929
Row23	Minnesota	777.18	133.35	Midwest	cluster_0	cluster_0	-71.819
Row24	Mississippi	896.07	155.77	South	cluster_0	cluster_0	48.637
Row25	Missouri	790.32	144.45	Midwest	cluster_0	cluster_0	-57.707
Row26	Montana	816.21	85.15	West	cluster_0	cluster_0	-37.422
Row27	Nebraska	732.28	114.82	Midwest	cluster_2	cluster_2	-118.242
Row29	New Hamps...	746.54	120.21	Northeast	cluster_2	cluster_2	-103.544
Row30	New Jersey	1,301.52	159.85	Northeast	cluster_1	cluster_1	452.721
Row33	North Carolina	708.24	127.82	South	cluster_2	cluster_2	-140.974
Row34	North Dakota	688.75	109.72	Midwest	cluster_2	cluster_2	-162.058
Row35	Ohio	697.73	133.52	Midwest	cluster_2	cluster_2	-150.911
Row36	Oklahoma	881.51	178.86	South	cluster_0	cluster_0	36.279
Row38	Pennsylvania	905.99	153.86	Northeast	cluster_0	cluster_0	58.338
Row39	Rhode Island	1,148.99	148.58	Northeast	cluster_1	cluster_1	299.803
Row41	South Dakota	669.31	96.87	Midwest	cluster_2	cluster_2	-182.605
Row42	Tennessee	767.91	155.57	South	cluster_0	cluster_0	-78.99
Row44	Utah	809.38	109.48	West	cluster_0	cluster_0	-41.969
Row45	Vermont	716.2	109.61	Northeast	cluster_2	cluster_2	-134.736
Row46	Virginia	768.95	153.72	South	cluster_0	cluster_0	-78.126
Row47	Washington	890.03	111.62	West	cluster_0	cluster_0	38.533
Row49	Wisconsin	670.31	106.62	Midwest	cluster_2	cluster_2	-180.705
Row50	Wyoming	791.14	122.04	West	cluster_0	cluster_0	-58.966

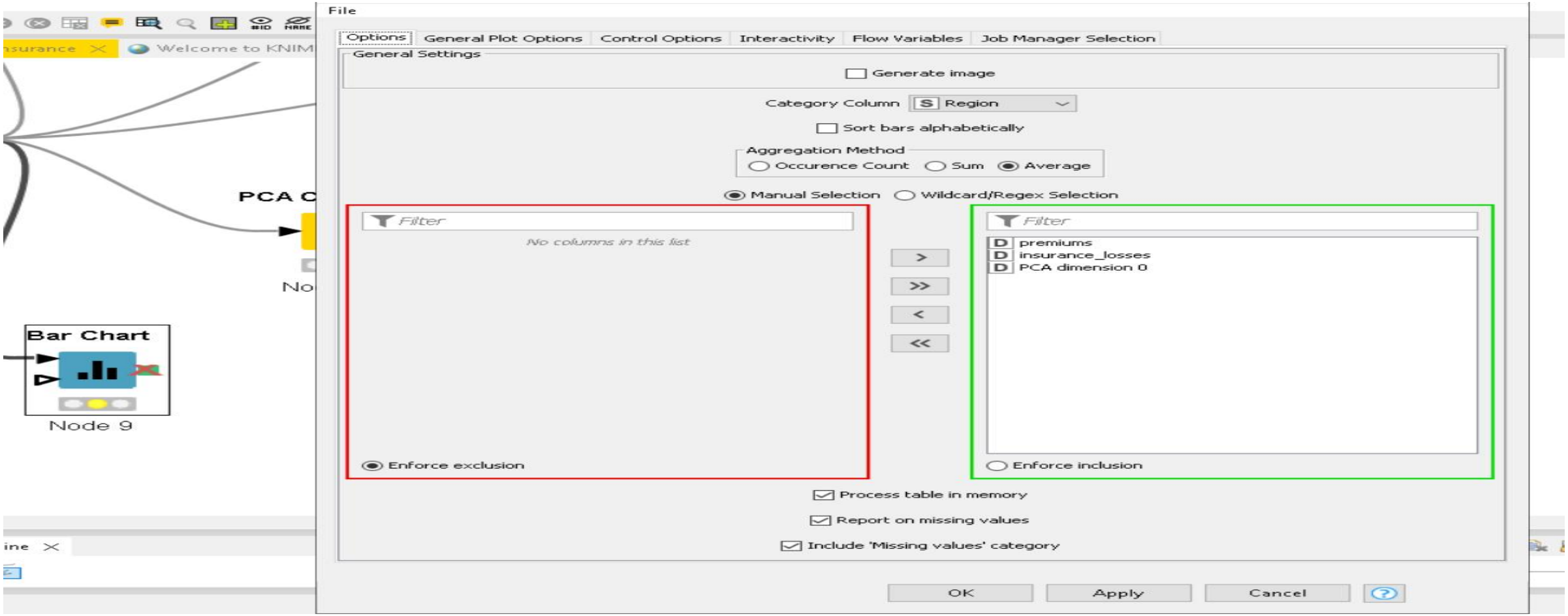
after that we plotting the data in x & y axis



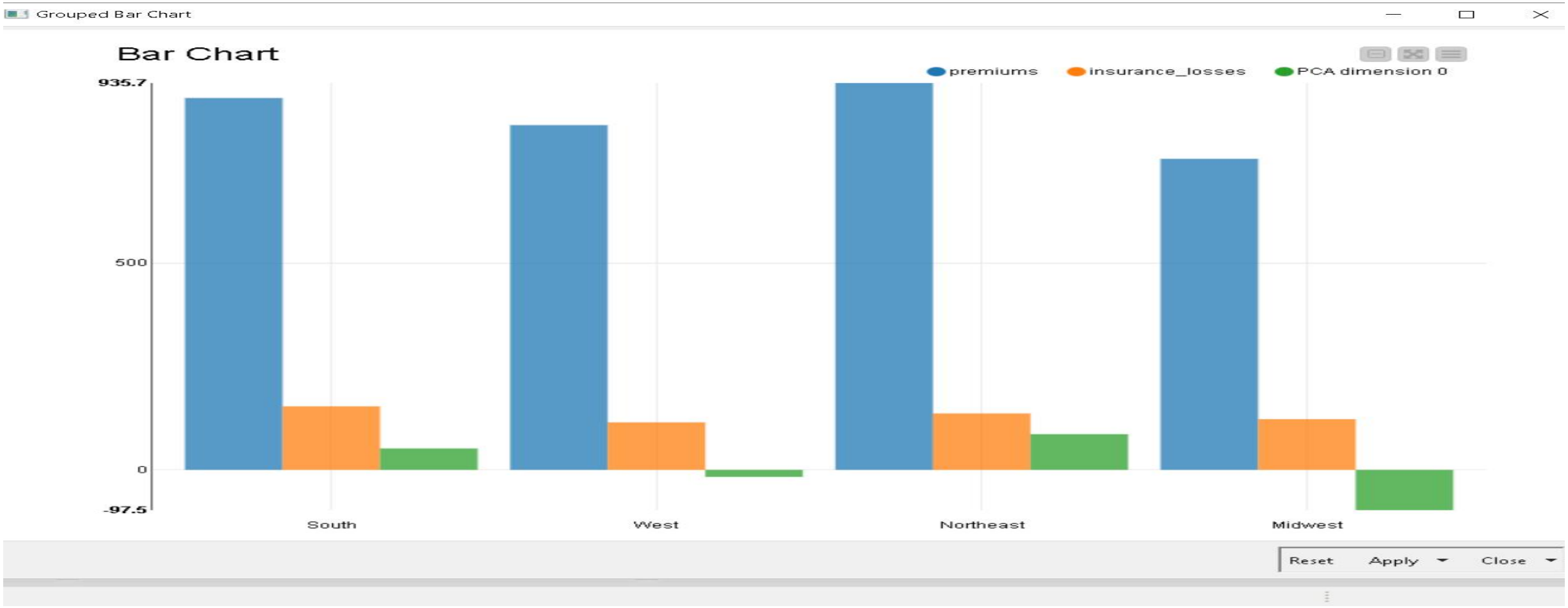
the states which have maximum premiums holder are from 750 -800



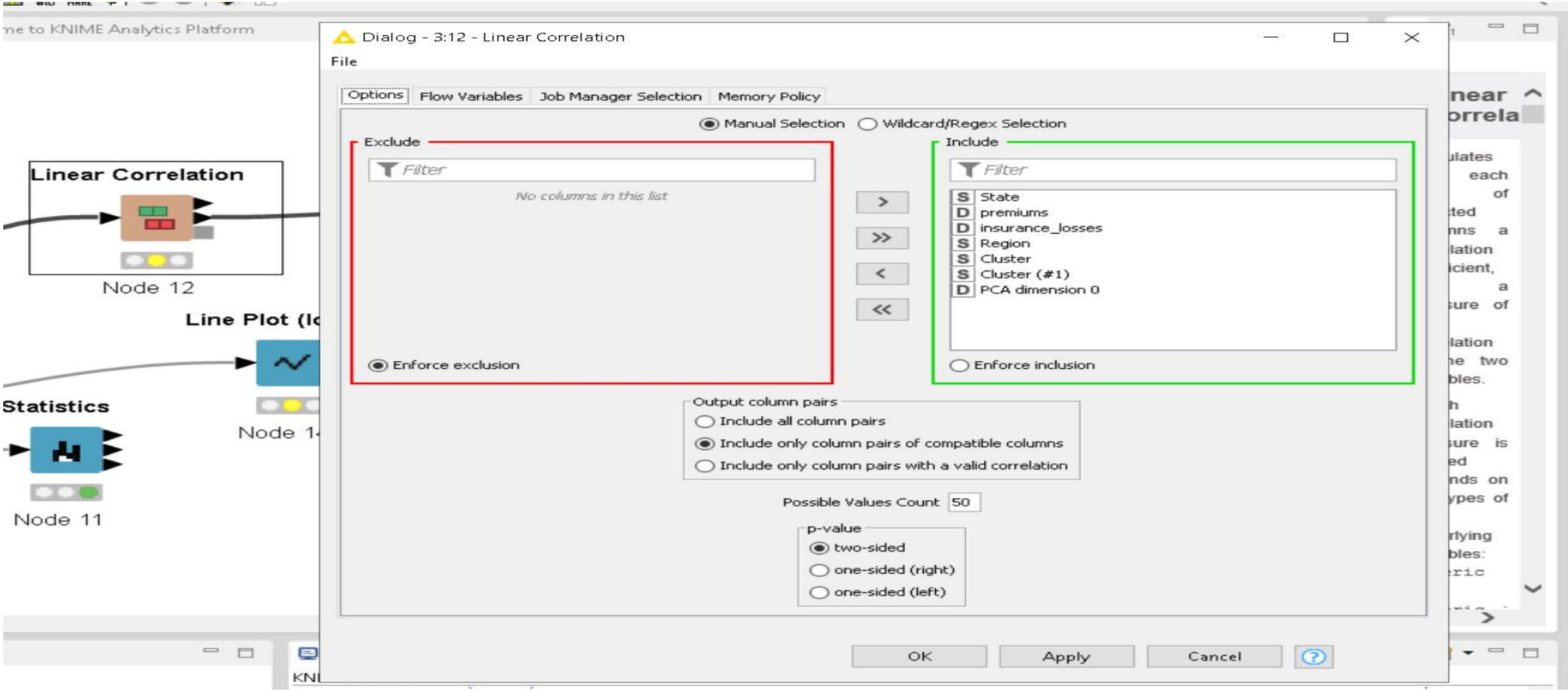
also in bar chart plotting is done

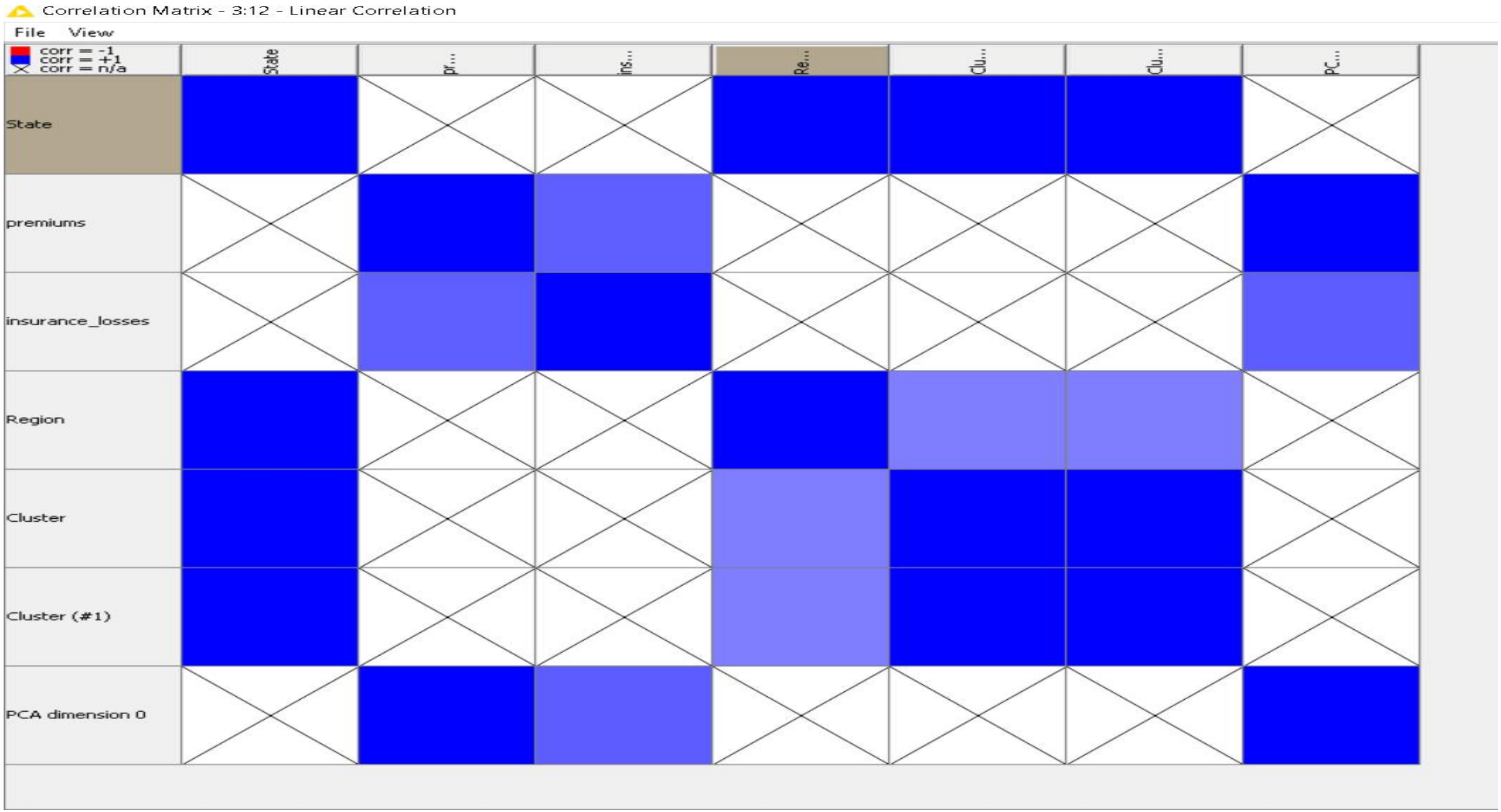


in midwest the coverage of insurance is going in negative
same in west
and northeast has the most premiums

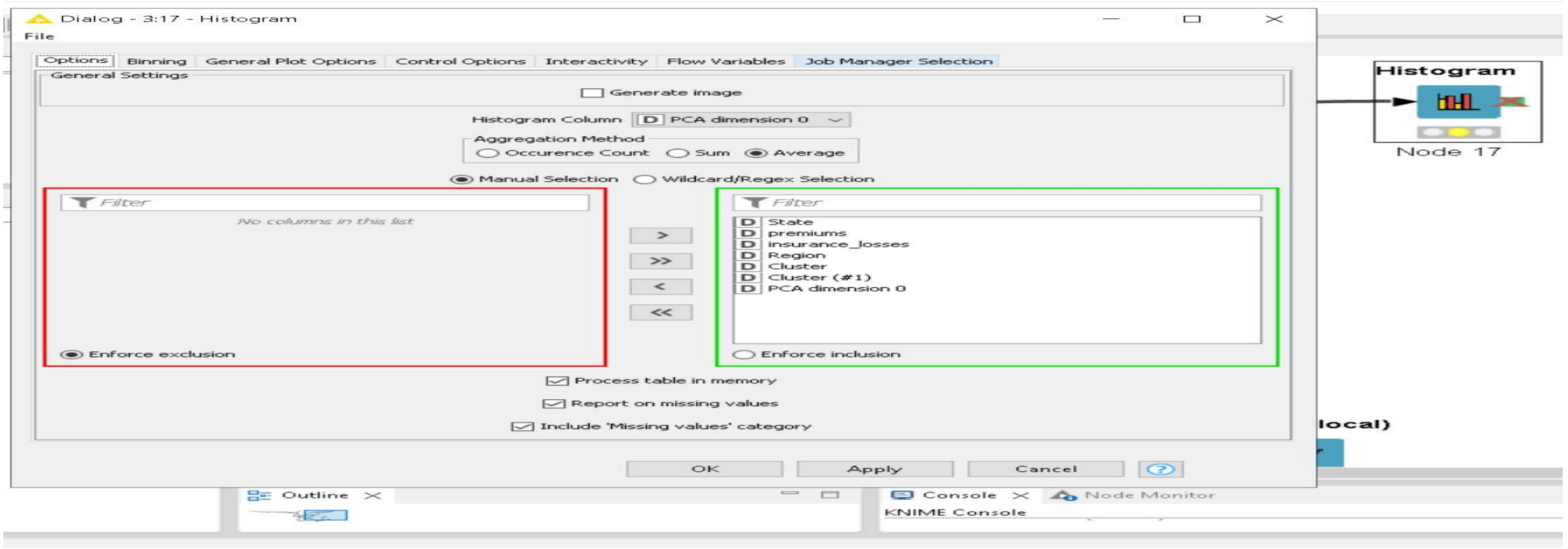


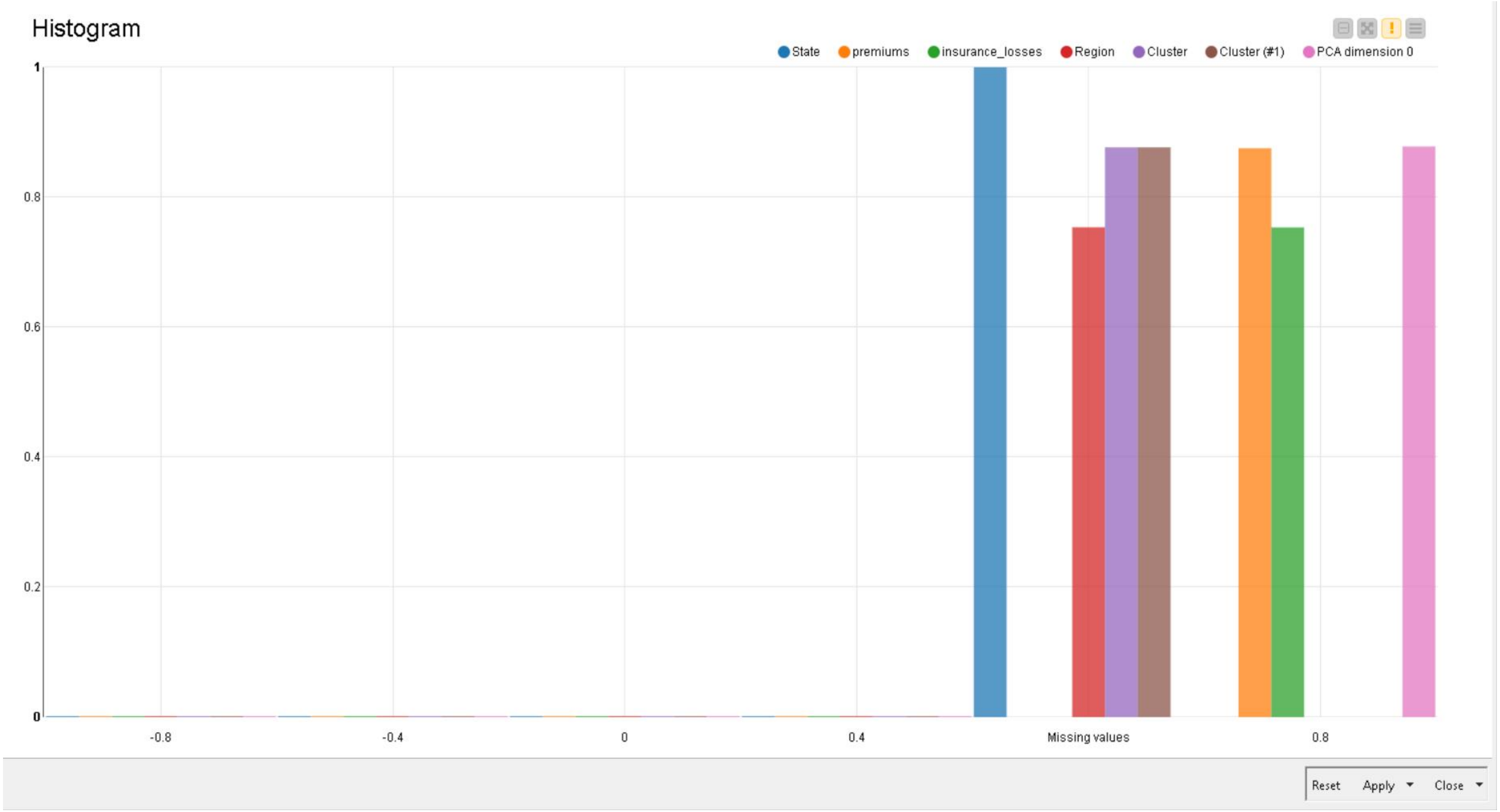
Calculates for each pair of selected columns a correlation coefficient, i.e. a measure of the correlation of the two variables.



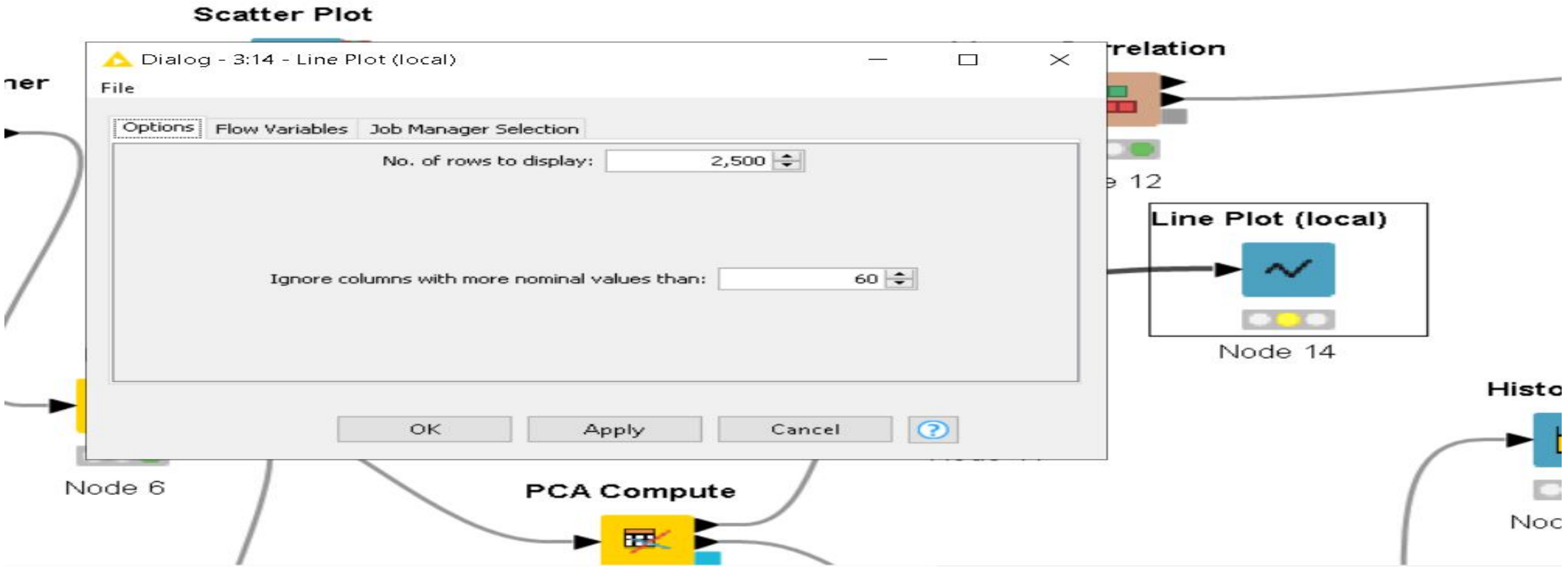

Correlation Matrix - 3:12 - Linear Correlation
File
View
corr = -1
corr = +1
corr = n/a
State
pr...
ins...
Re...
clu...
clu...
PC...
State
premiums
insurance_losses
Region
Cluster
Cluster (#1)
PCA dimension 0

after selection put CSS rules into a single string and set it as a flow variable 'customCSS so i have taken this columns

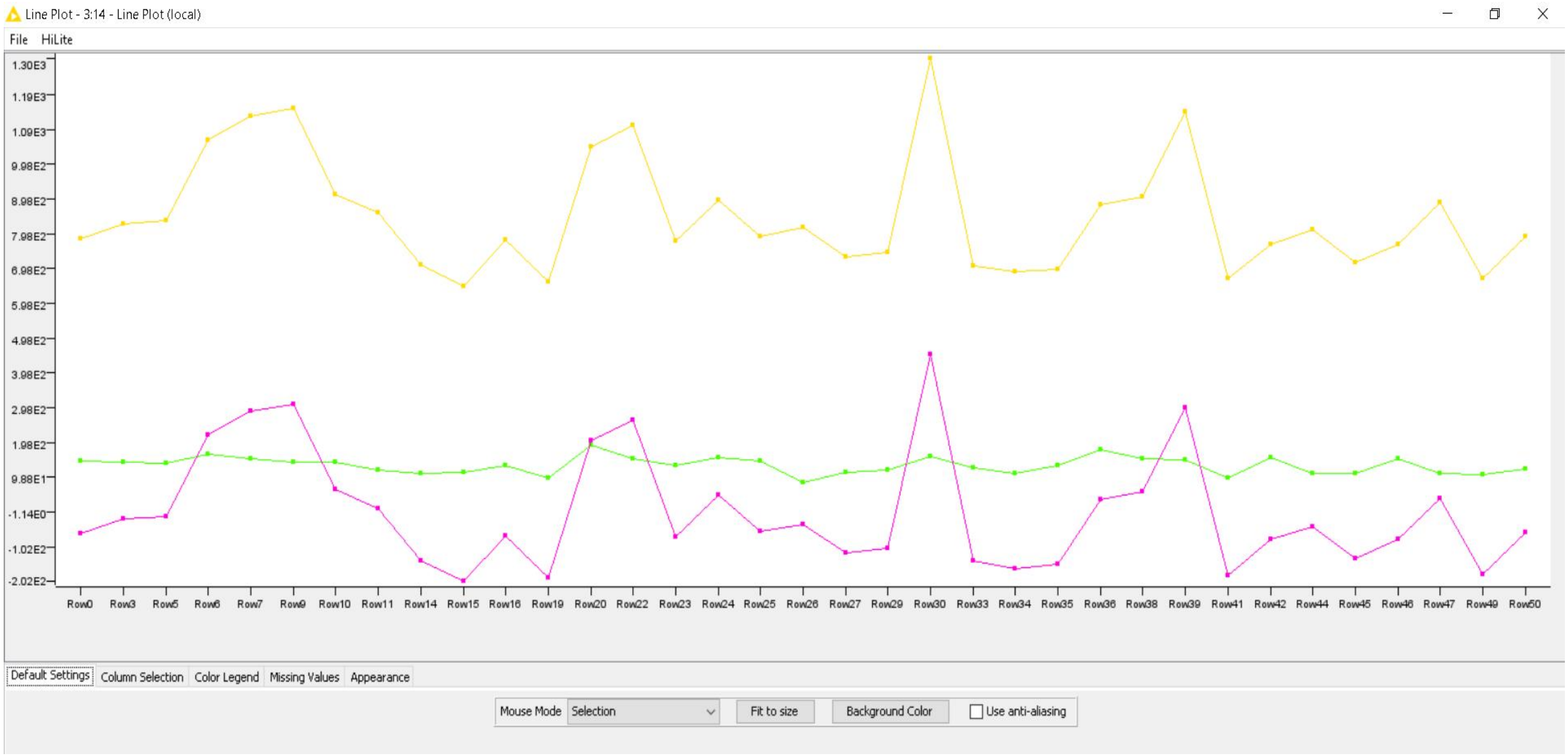



Histogram
State
premiums
insurance_losses
Region
Cluster
Cluster (#1)
PCA dimension 0
1
0.8
0.6
0.4
0.2
0
-0.8
-0.4
0
0.4
Missing values
0.8
Reset
Apply
Close

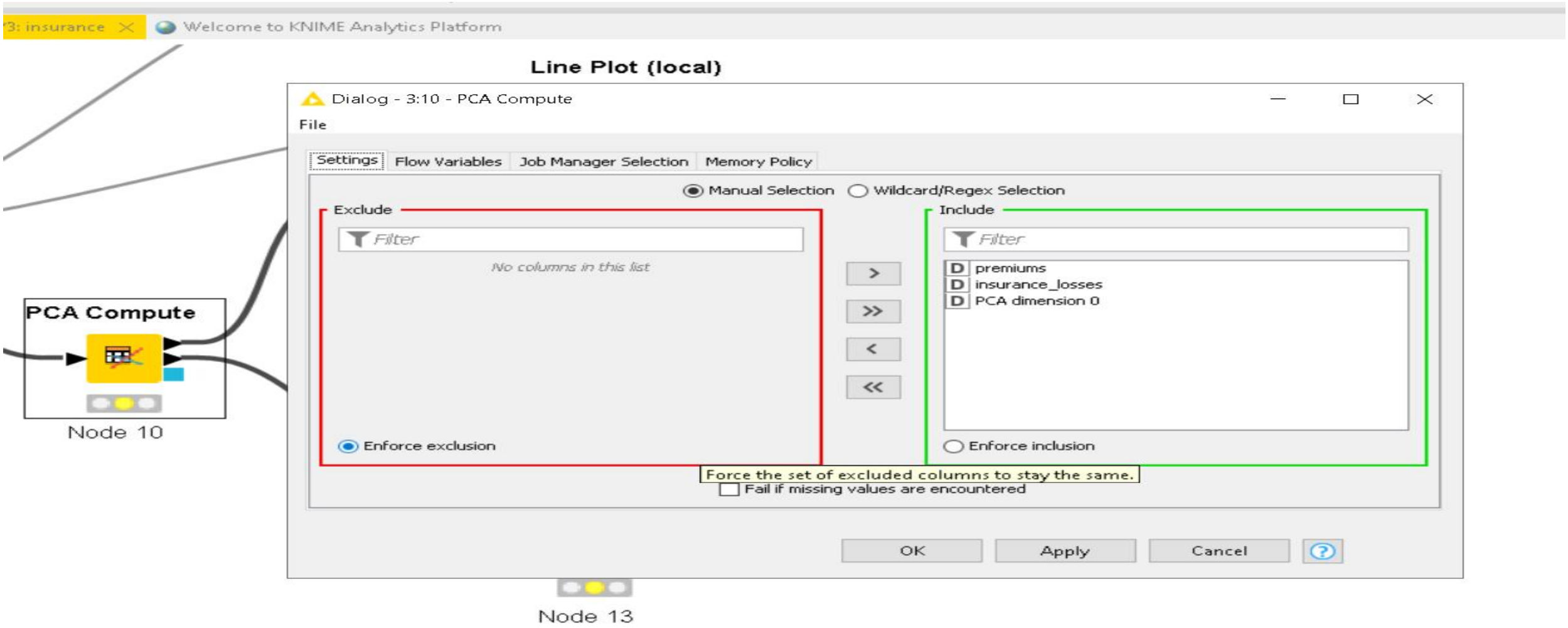
also line plot is necessary to understand the story

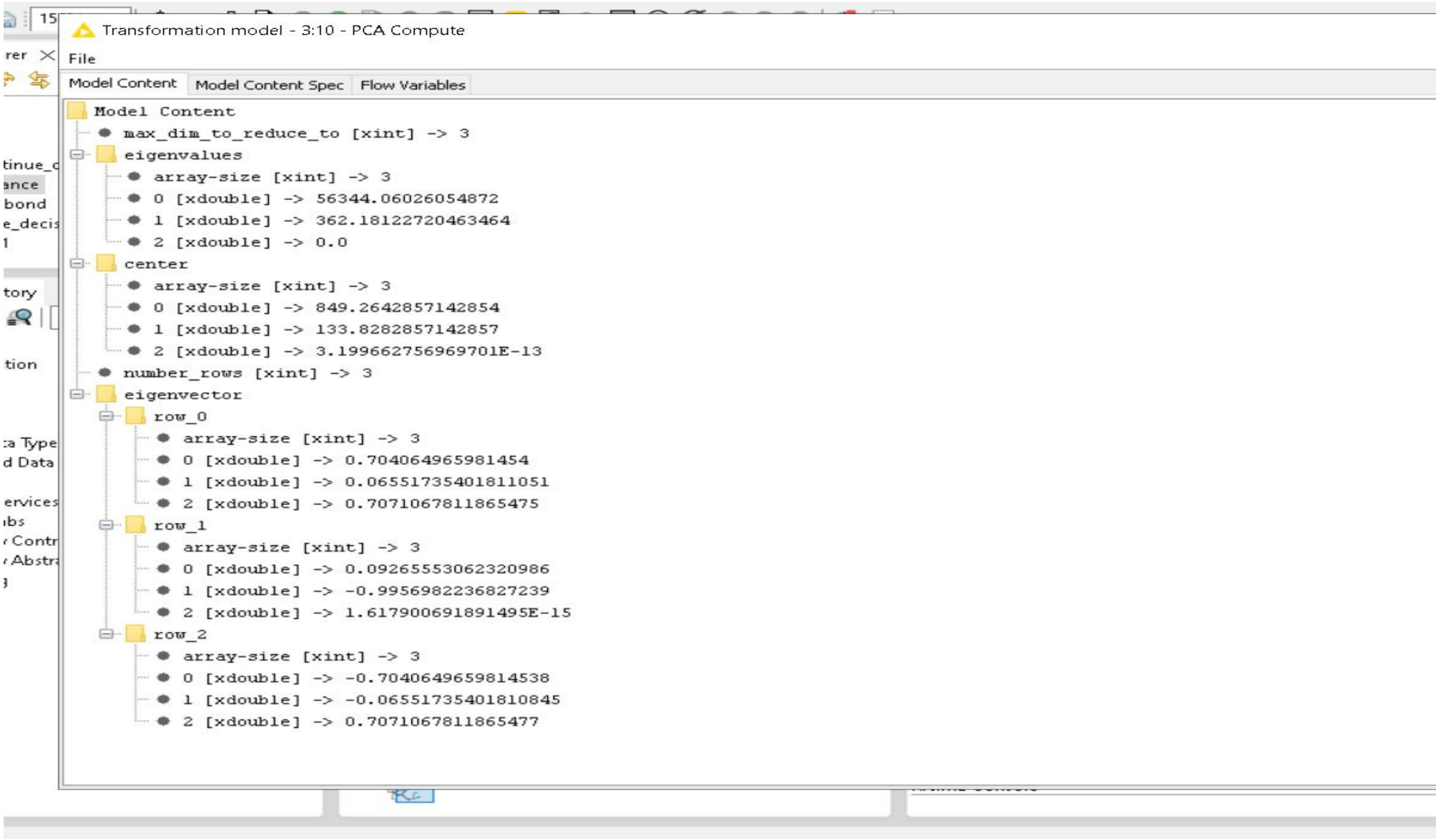


so company seems to be in heavy loss lets further investigate the data

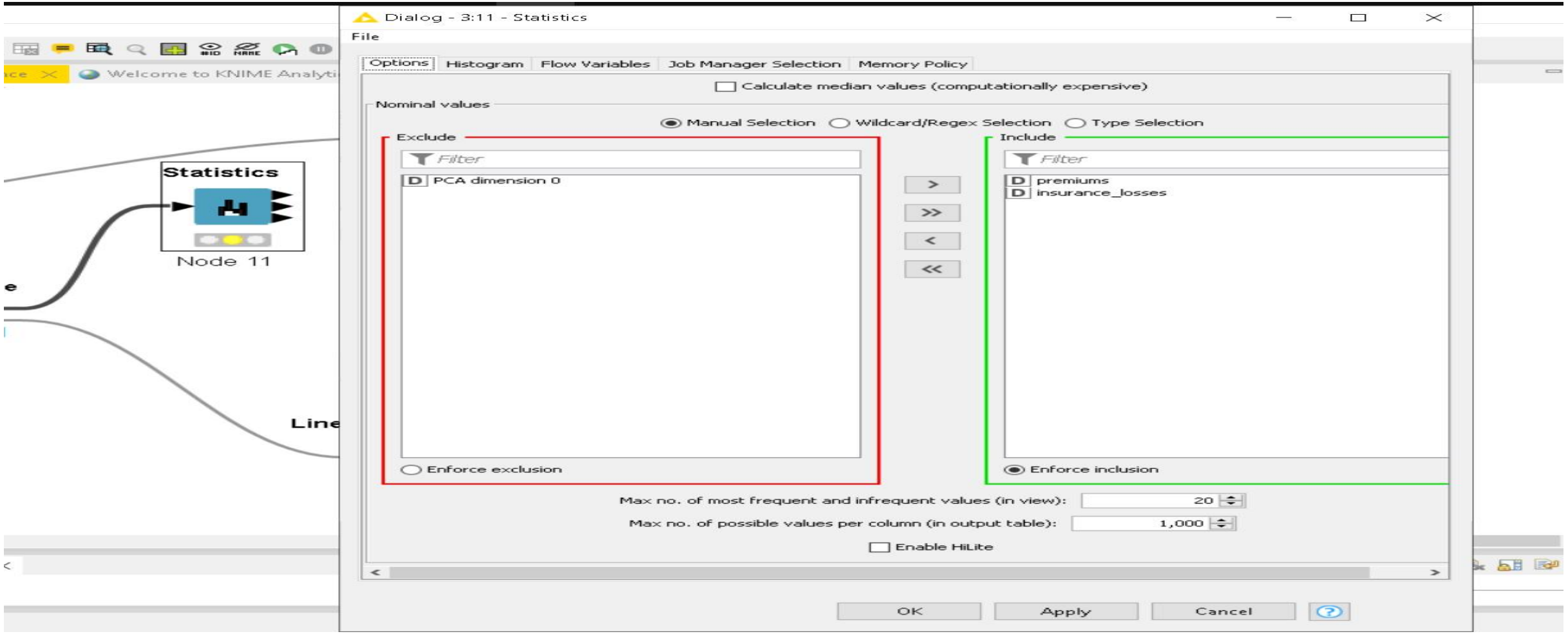


after watching all plots now connect pca with this node performs a principal component analysis (PCA) on the given input data. The directions of maximal variance (the principal components) are extracted and can be used in the PCA Apply node to project the input into a space of lower dimension while preserving a maximum of information.




Transformation model - 3:10 - PCA Compute
File
Model Content
Model Content Spec
Flow Variables
Model Content
max_dim_to_reduce_to [xint] -> 3
eigenvalues
array-size [xint] -> 3
0 [xdouble] -> 56344.06026054872
1 [xdouble] -> 362.18122720463464
2 [xdouble] -> 0.0
center
array-size [xint] -> 3
0 [xdouble] -> 849.2642857142854
1 [xdouble] -> 133.8282857142857
2 [xdouble] -> 3.199662756969701E-13
number_rows [xint] -> 3
eigenvector
row_0
array-size [xint] -> 3
0 [xdouble] -> 0.704064965981454
1 [xdouble] -> 0.06551735401811051
2 [xdouble] -> 0.7071067811865475
row_1
array-size [xint] -> 3
0 [xdouble] -> 0.09265553062320986
1 [xdouble] -> -0.9956982236827239
2 [xdouble] -> 1.617900691891495E-15
row_2
array-size [xint] -> 3
0 [xdouble] -> -0.7040649659814538
1 [xdouble] -> -0.06551735401810845
2 [xdouble] -> 0.7071067811865477

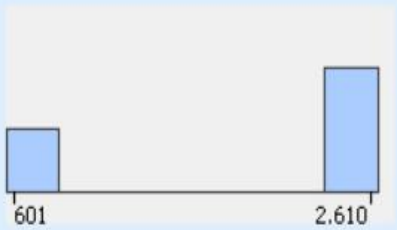
now i have used statistics to calculate statistical moments such as minimum, maximum, mean, standard deviation, variance, median, overall sum, number of missing values and row count across all numeric columns, and counts all nominal values together with their occurrences. The dialog offers two options for choosing the
median and/or nominal values calculations:



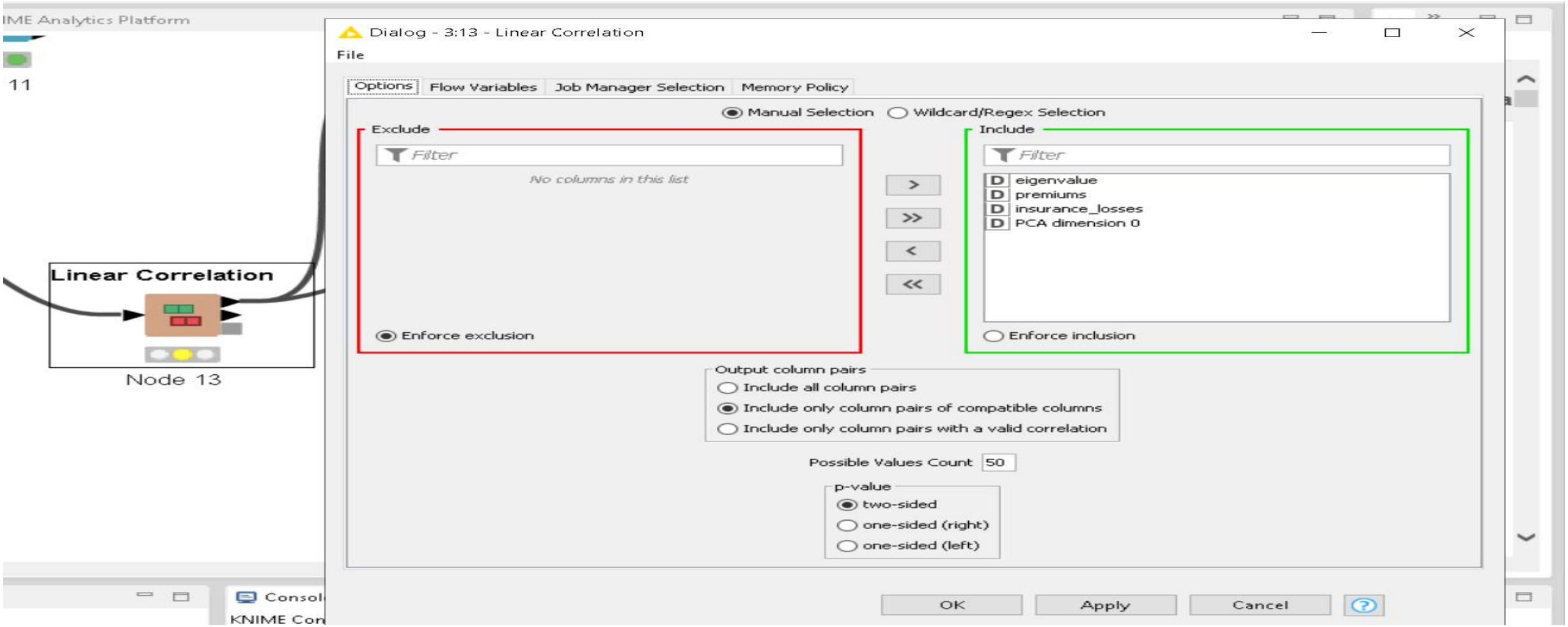
Statistics View - 3:11 - Statistics

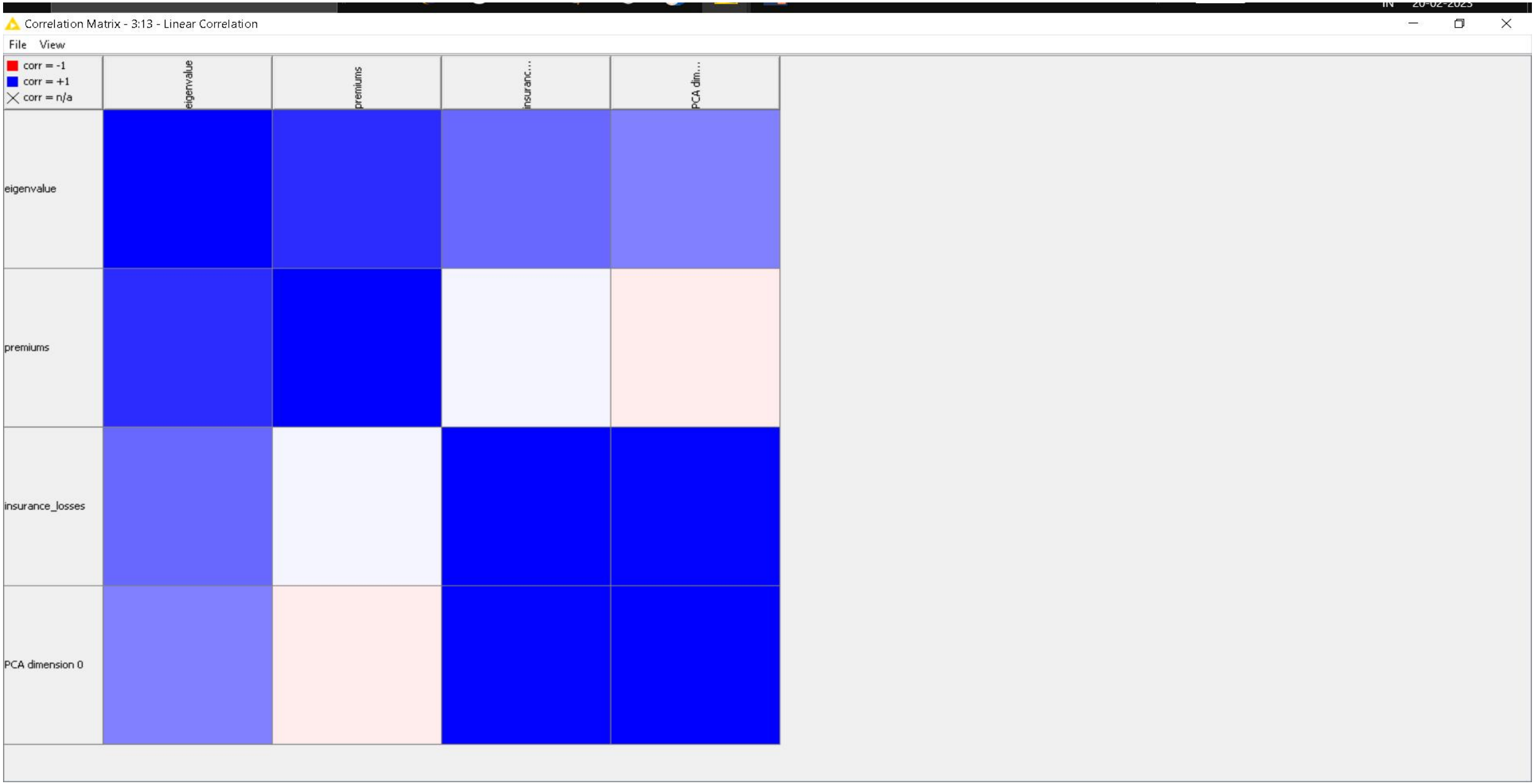
File

Numeric | Nominal | Top/bottom

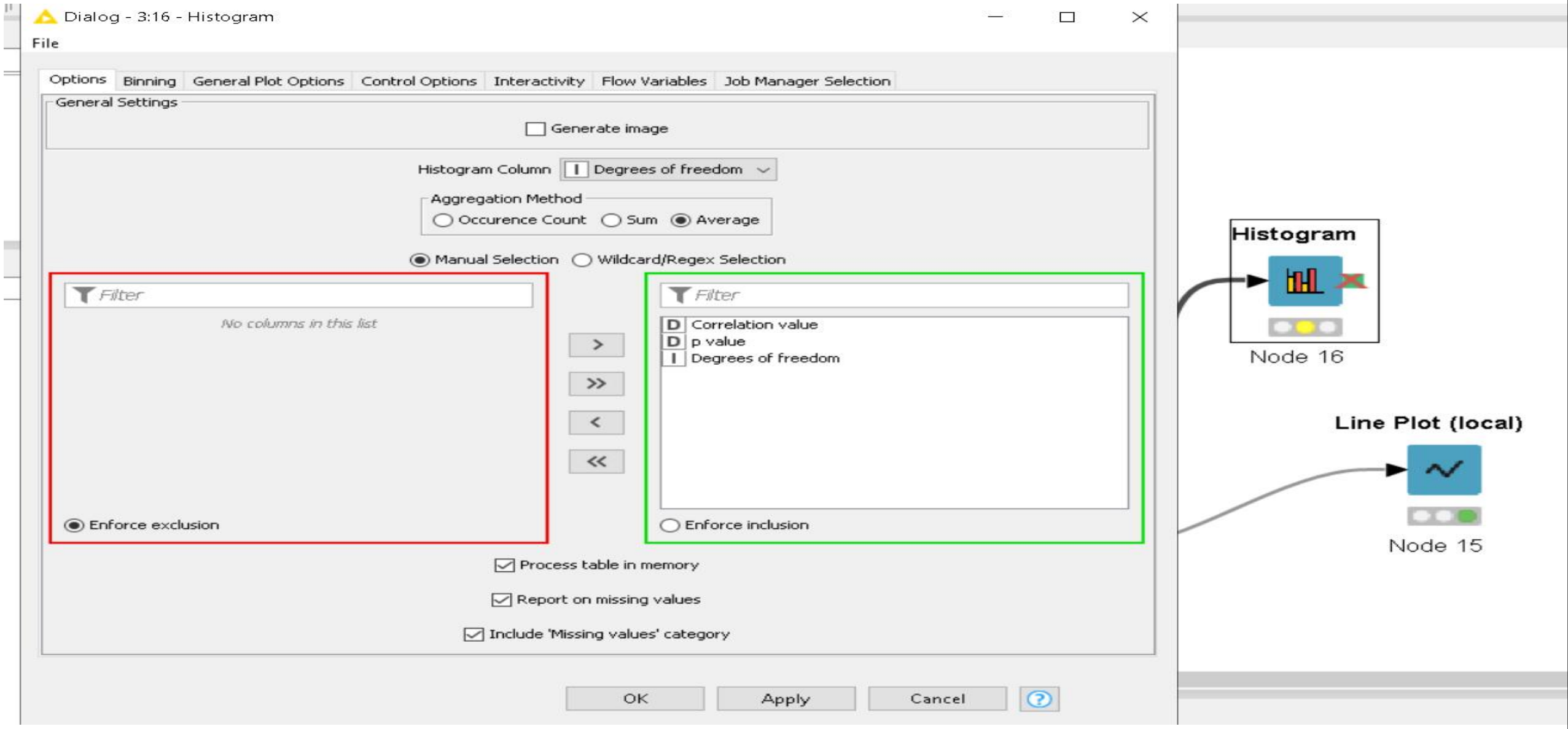
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
premiums	2,565.6518	19,516.5911	?	28,050.8404	14,680.0618	-1.7319	?	0	0	0	
insurance_losses	600.9301	1,925.6254	?	2,610.2944	1,147.4369	-1.7291	?	0	0	0	
PCA dimension 0	2,610.2944	19,611.055	?	28,172.0301	14,723.2152	-1.7319	?	0	0	0	

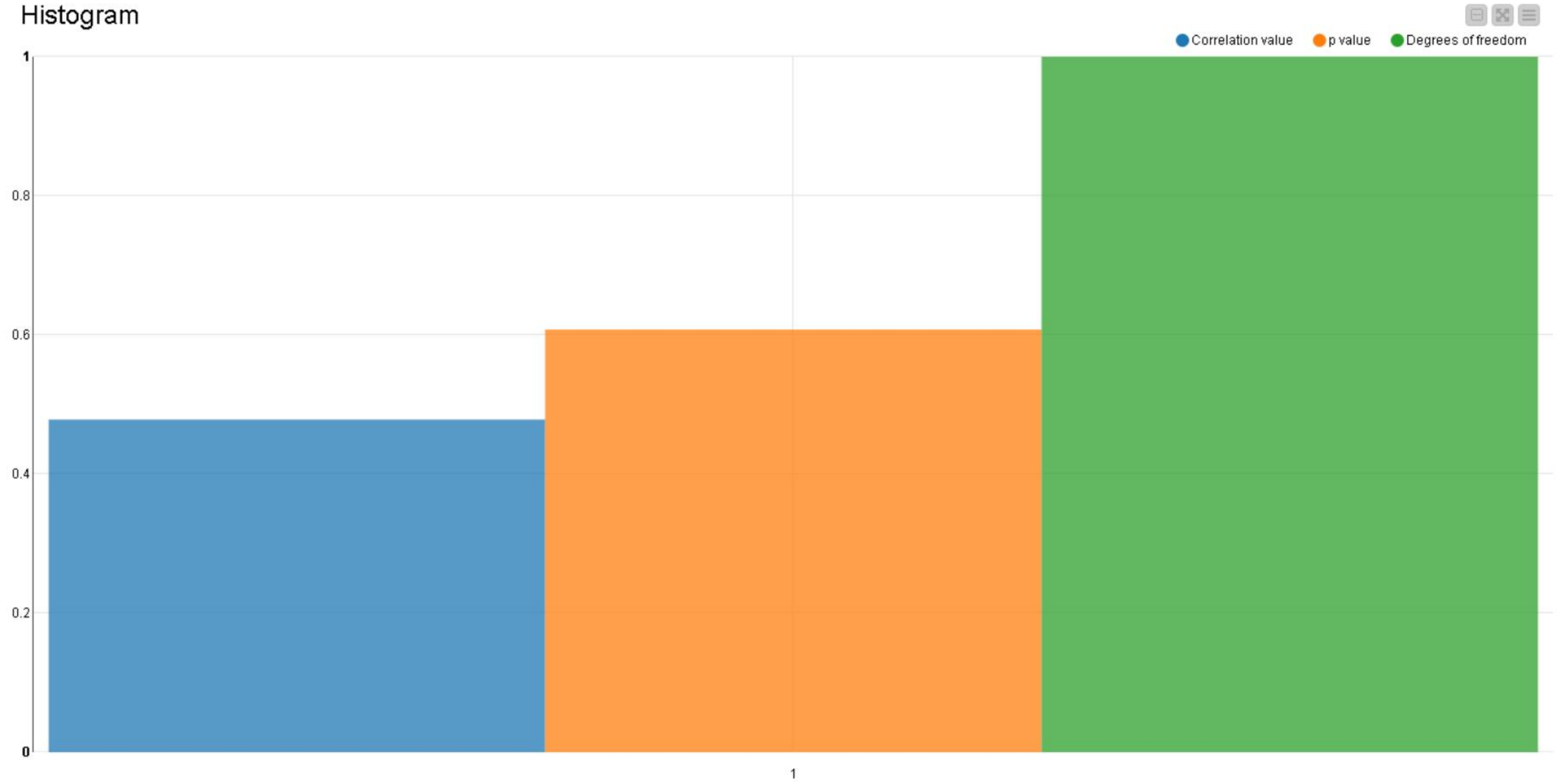
again finding the correlations



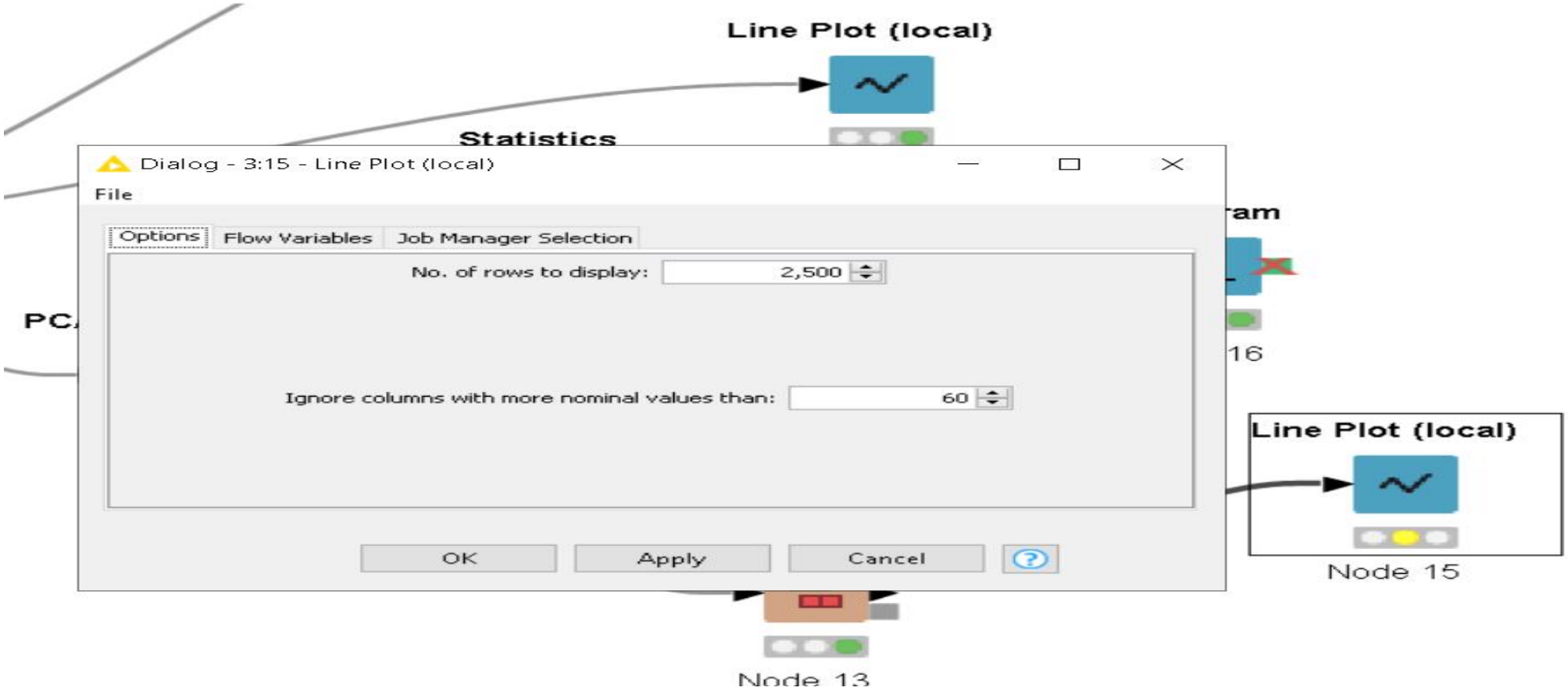

Correlation Matrix - 3:13 - Linear Correlation
File View
corr = -1
corr = +1
corr = n/a
eigenvalue
premiums
insuranc...
PCA dim...
eigenvalue
premiums
insurance_losses
PCA dimension 0

i have taken degree of freedom as column

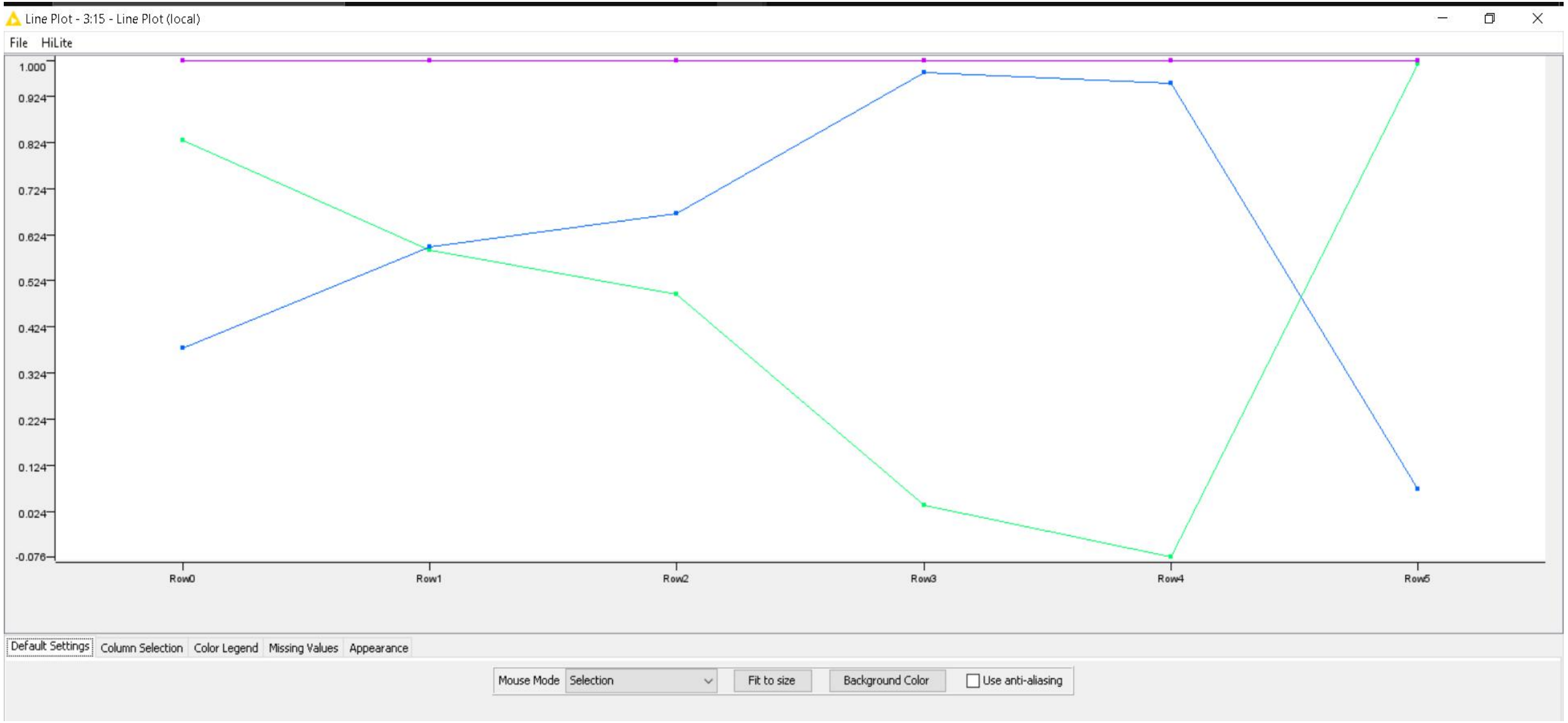



Histogram
Correlation value
p value
Degrees of freedom
1
0.8
0.6
0.4
0.2
0
1

now line plot



but as we can see their is a huge growth no loss in the line plot



Thanks for watching