2016

Assignment 7

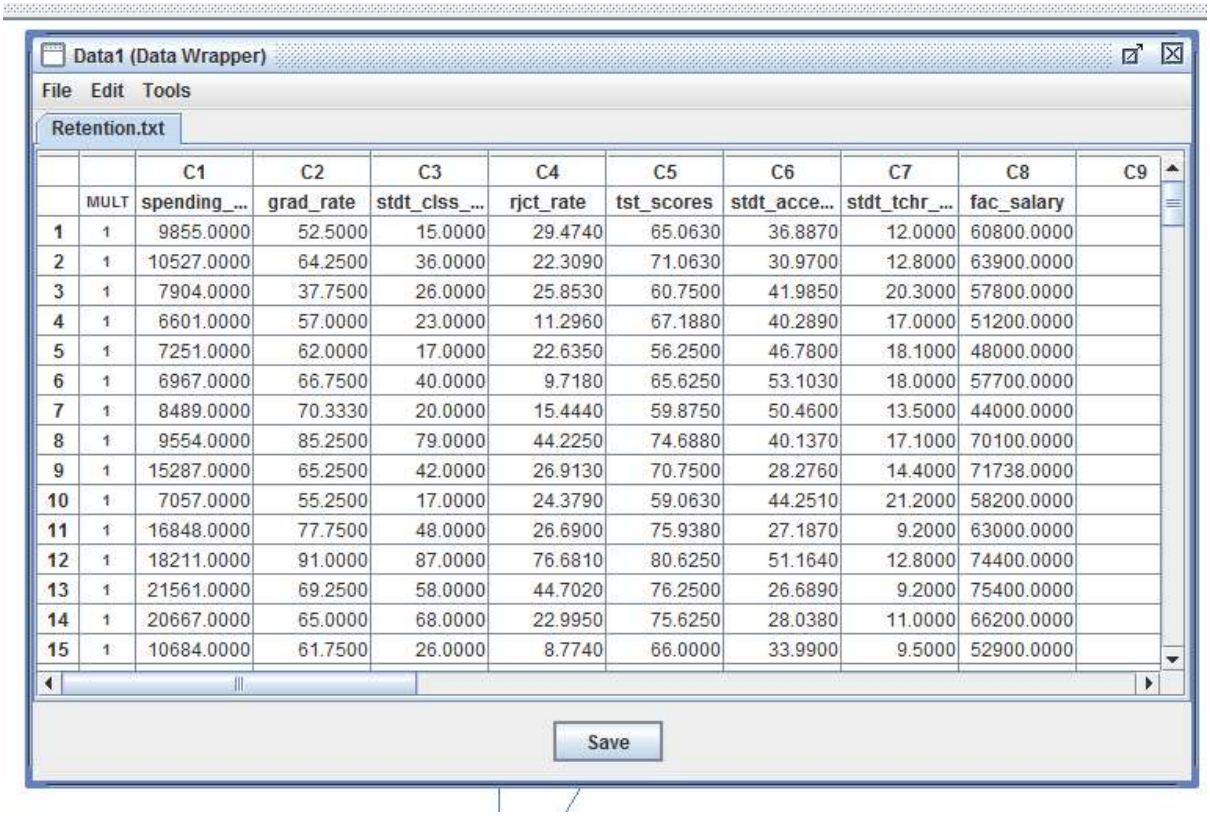
DATA ANALYTICS – CAUSAL DISCOVERY

LIJU ROBIN GEORGE & JAYASHANKAR MALEPATI

Report:

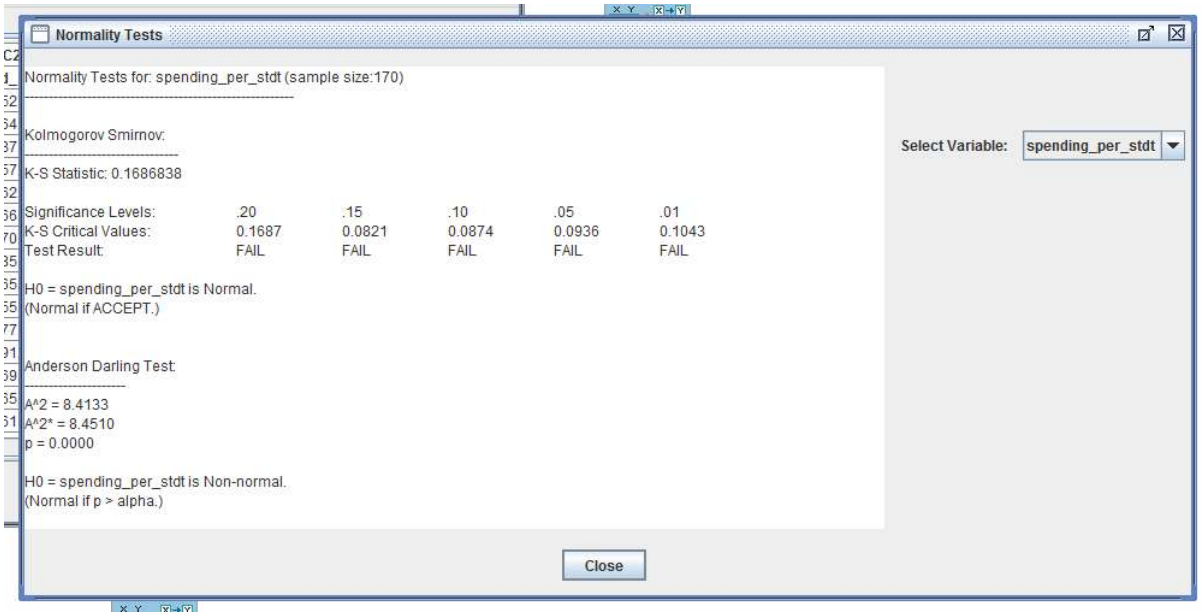
We performed the analysis using tetrad 5.3 as it seemed quite stable and bug free. We did the analysis step by step, as was done in the paper for the 1992 data. At each stage, we compare what we found with what was mentioned in the previous paper.

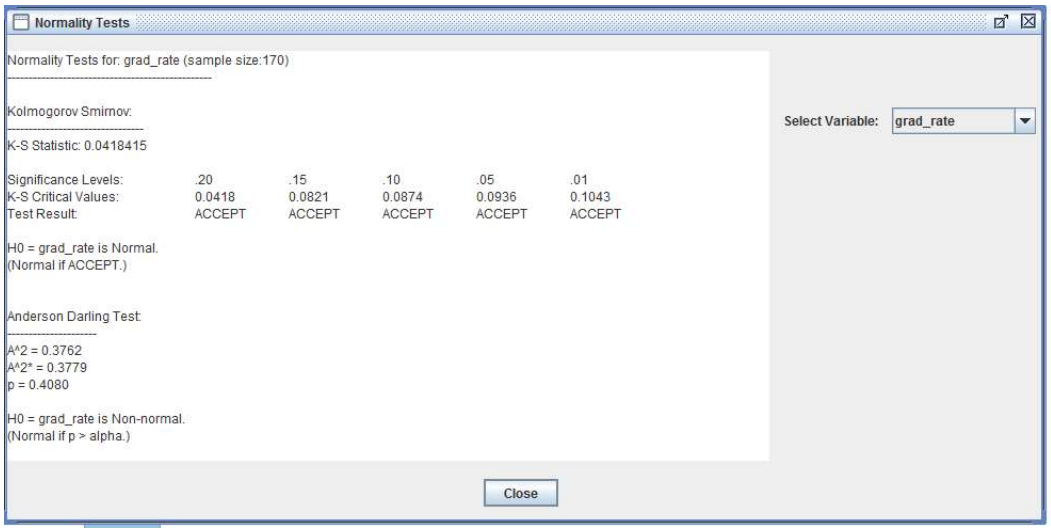
We started off with reading the data in:

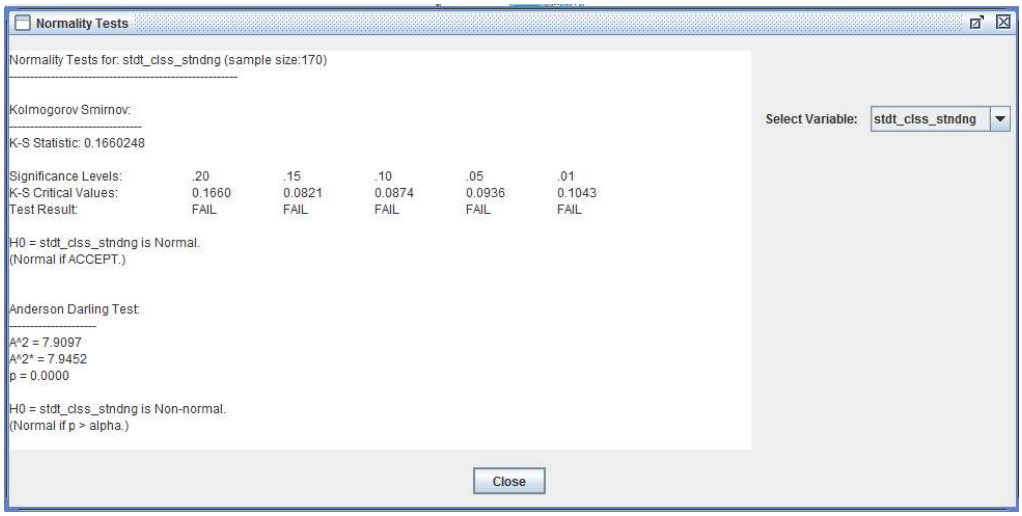


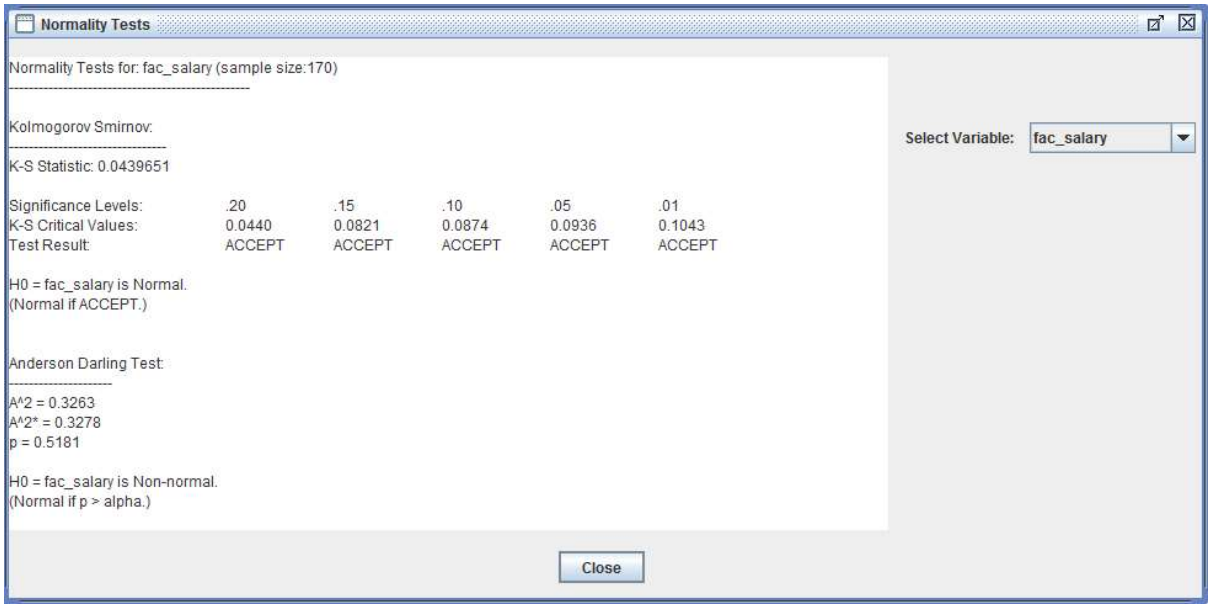
		C1	C2	C3	C4	C5	C6	C7	C8	C9
	MULT	spending_...	grad_rate	stdt_clss_...	rjct_rate	tst_scores	stdt_acce...	stdt_tchr_...	fac_salary	
1	1	9855.0000	52.5000	15.0000	29.4740	65.0630	36.8870	12.0000	60800.0000	
2	1	10527.0000	64.2500	36.0000	22.3090	71.0630	30.9700	12.8000	63900.0000	
3	1	7904.0000	37.7500	26.0000	25.8530	60.7500	41.9850	20.3000	57800.0000	
4	1	6601.0000	57.0000	23.0000	11.2960	67.1880	40.2890	17.0000	51200.0000	
5	1	7251.0000	62.0000	17.0000	22.6350	56.2500	46.7800	18.1000	48000.0000	
6	1	6967.0000	66.7500	40.0000	9.7180	65.6250	53.1030	18.0000	57700.0000	
7	1	8489.0000	70.3330	20.0000	15.4440	59.8750	50.4600	13.5000	44000.0000	
8	1	9554.0000	85.2500	79.0000	44.2250	74.6880	40.1370	17.1000	70100.0000	
9	1	15287.0000	65.2500	42.0000	26.9130	70.7500	28.2760	14.4000	71738.0000	
10	1	7057.0000	55.2500	17.0000	24.3790	59.0630	44.2510	21.2000	58200.0000	
11	1	16848.0000	77.7500	48.0000	26.6900	75.9380	27.1870	9.2000	63000.0000	
12	1	18211.0000	91.0000	87.0000	76.6810	80.6250	51.1640	12.8000	74400.0000	
13	1	21561.0000	69.2500	58.0000	44.7020	76.2500	26.6890	9.2000	75400.0000	
14	1	20667.0000	65.0000	68.0000	22.9950	75.6250	28.0380	11.0000	66200.0000	
15	1	10684.0000	61.7500	26.0000	8.7740	66.0000	33.9900	9.5000	52900.0000	

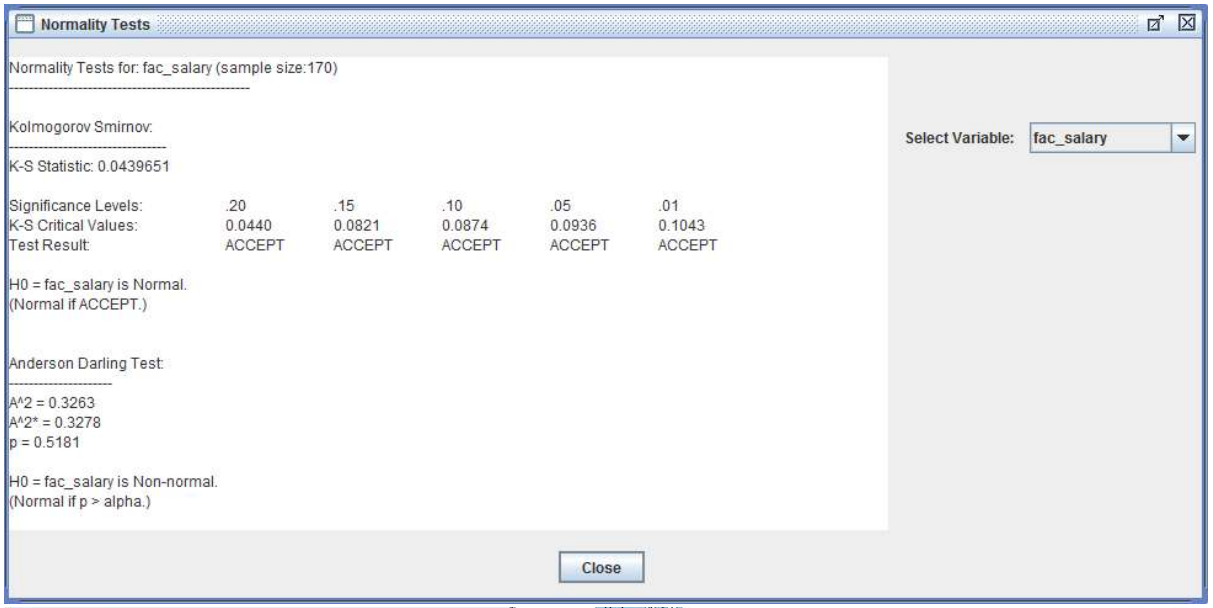
The next step we checked whether the variables of concern were normalized as was mentioned in the paper. We got the following set of results. The normality tests showed that a few variables were not normalize at all, 3 of them specifically spending_pre_stdt, stdt_class_standing, rjct_rate.

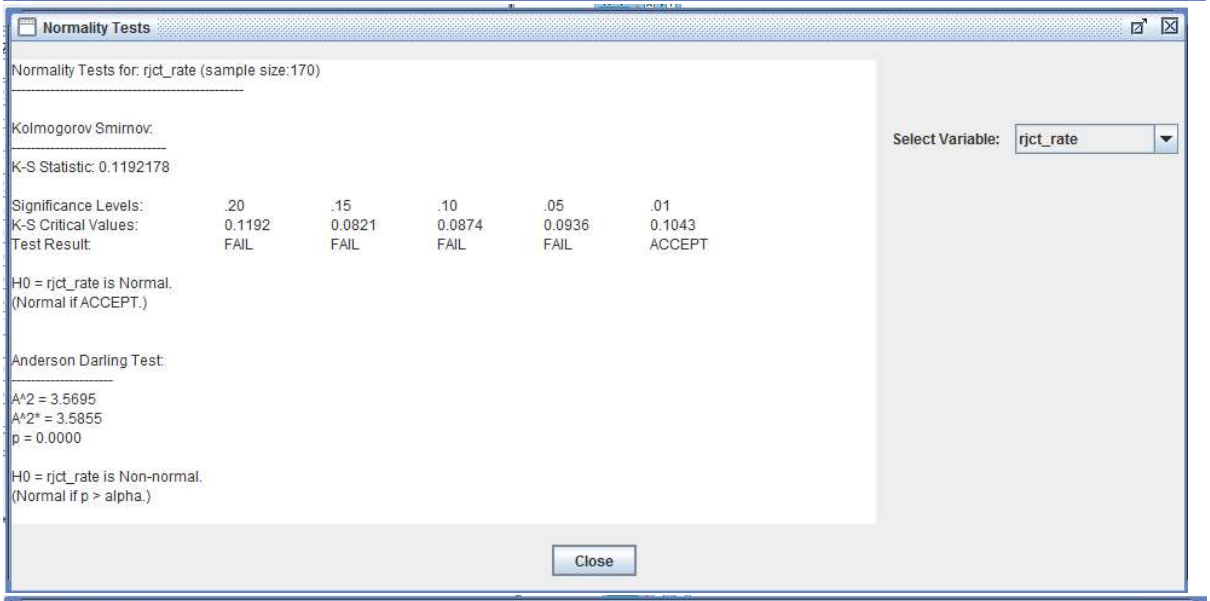

Normality Tests
Normality Tests for: spending_per_stdt (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.1686838
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.1687 0.0821 0.0874 0.0936 0.1043
Test Result: FAIL FAIL FAIL FAIL FAIL
H0 = spending_per_stdt is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 8.4133
A^2* = 8.4510
p = 0.0000
H0 = spending_per_stdt is Non-normal.
(Normal if p > alpha.)
Select Variable: spending_per_stdt
Close

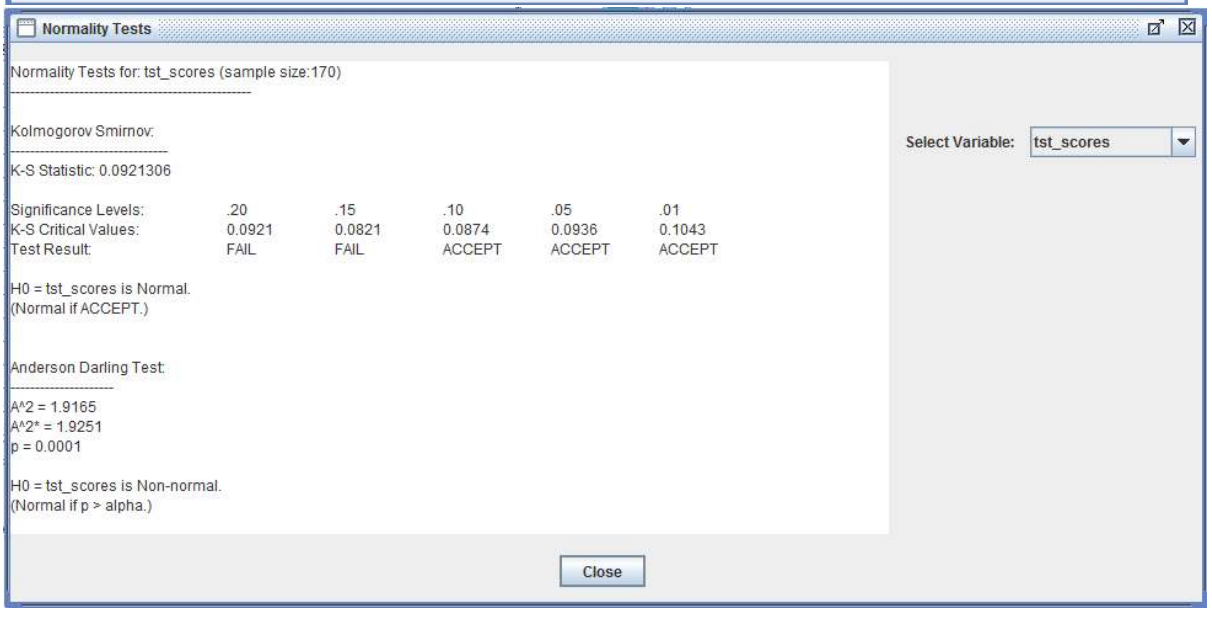

Normality Tests
Normality Tests for: grad_rate (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.0418415
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.0418 0.0821 0.0874 0.0936 0.1043
Test Result: ACCEPT ACCEPT ACCEPT ACCEPT ACCEPT
H0 = grad_rate is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 0.3762
A^2* = 0.3779
p = 0.4080
H0 = grad_rate is Non-normal.
(Normal if p > alpha.)
Select Variable: grad_rate
Close

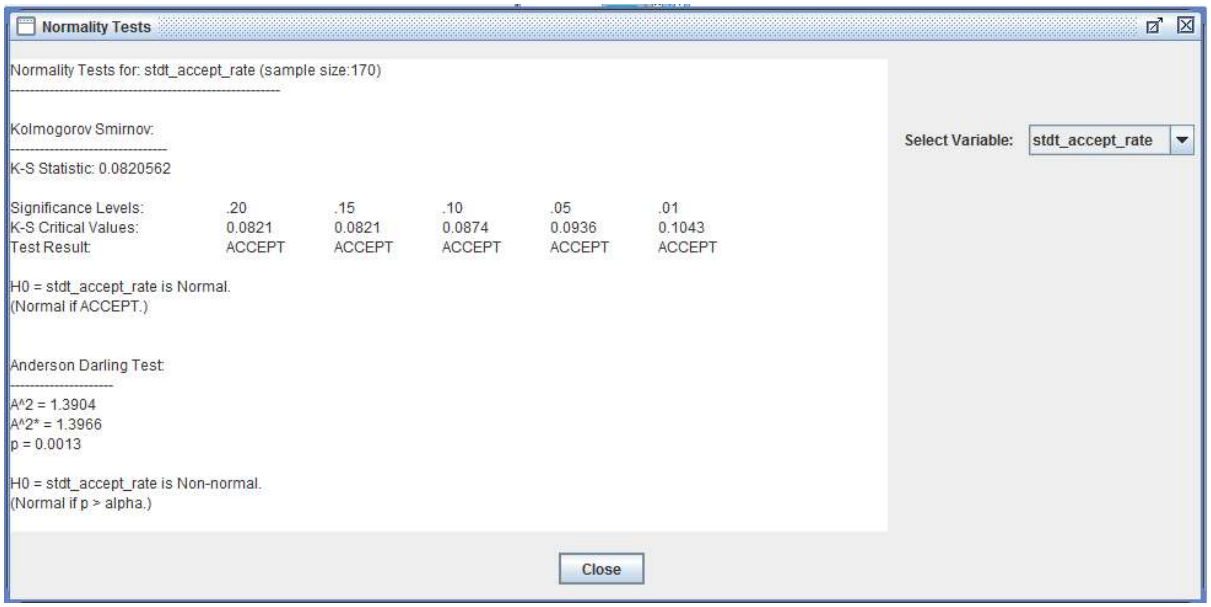

Normality Tests
Normality Tests for: stdt_clss_stndng (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.1660248
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.1660 0.0821 0.0874 0.0936 0.1043
Test Result: FAIL FAIL FAIL FAIL FAIL
H0 = stdt_clss_stndng is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 7.9097
A^2* = 7.9452
p = 0.0000
H0 = stdt_clss_stndng is Non-normal.
(Normal if p > alpha.)
Select Variable: stdt_clss_stndng
Close

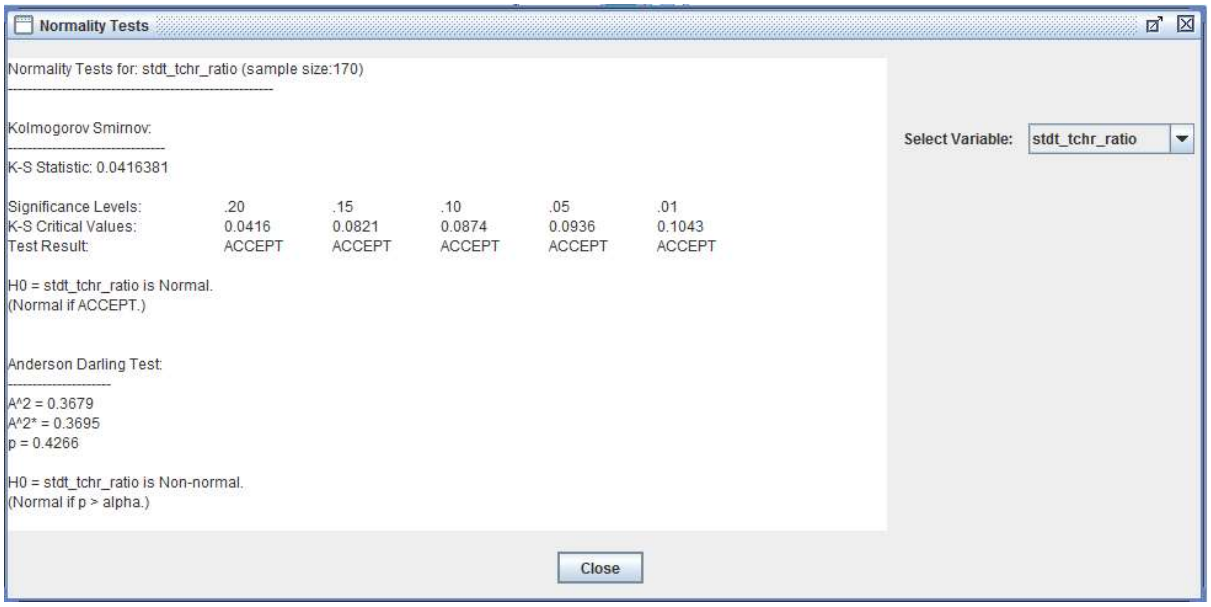

Normality Tests
Normality Tests for: fac_salary (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.0439651
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.0440 0.0821 0.0874 0.0936 0.1043
Test Result: ACCEPT ACCEPT ACCEPT ACCEPT ACCEPT
H0 = fac_salary is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 0.3263
A^2* = 0.3278
p = 0.5181
H0 = fac_salary is Non-normal.
(Normal if p > alpha.)
Select Variable:
fac_salary
Close


Normality Tests
Normality Tests for: fac_salary (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.0439651
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.0440 0.0821 0.0874 0.0936 0.1043
Test Result: ACCEPT ACCEPT ACCEPT ACCEPT ACCEPT
H0 = fac_salary is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 0.3263
A^2* = 0.3278
p = 0.5181
H0 = fac_salary is Non-normal.
(Normal if p > alpha.)
Select Variable: fac_salary
Close

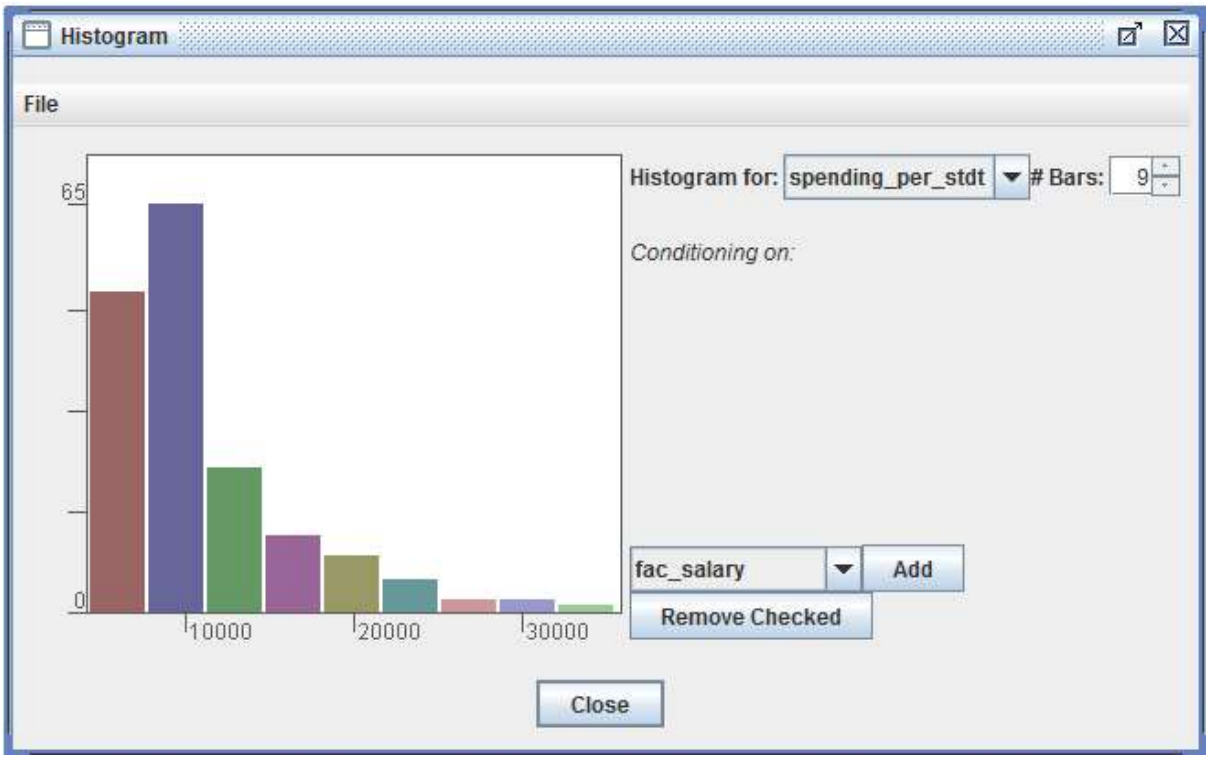

Normality Tests
Normality Tests for: rjct_rate (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.1192178
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.1192 0.0821 0.0874 0.0936 0.1043
Test Result: FAIL FAIL FAIL FAIL ACCEPT
H0 = rjct_rate is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 3.5695
A^2* = 3.5855
p = 0.0000
H0 = rjct_rate is Non-normal.
(Normal if p > alpha.)
Select Variable: rjct_rate
Close

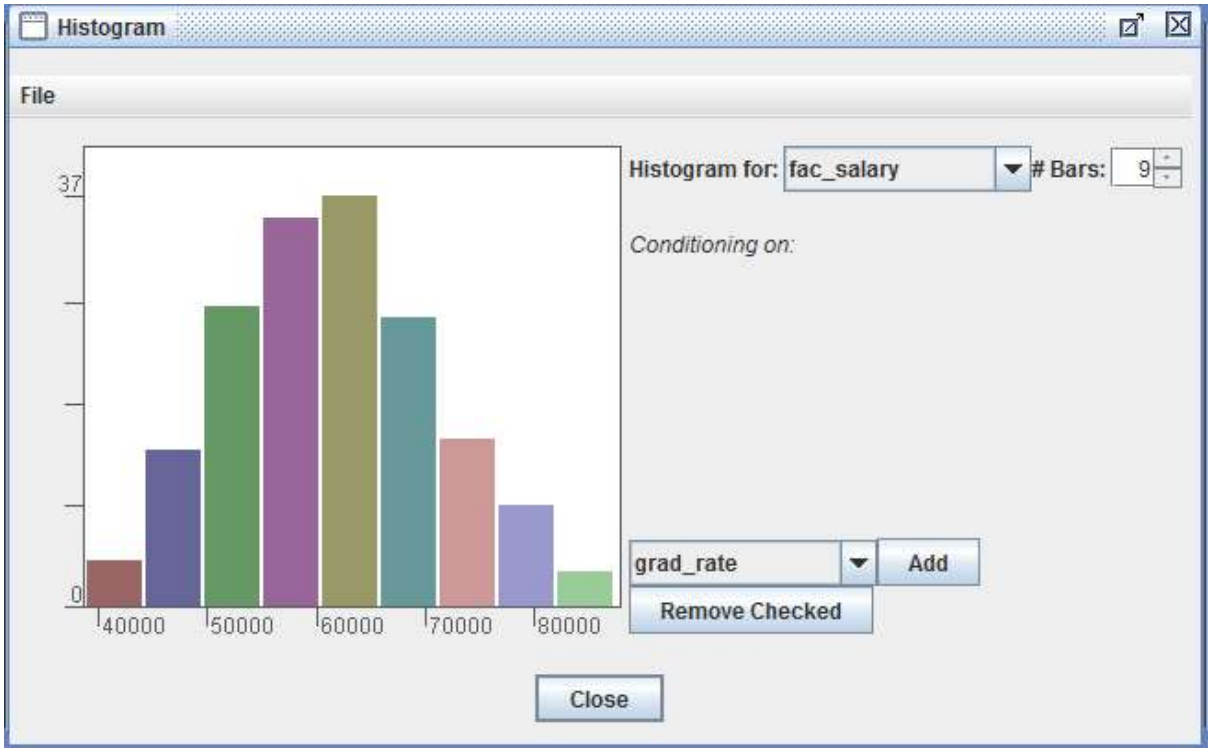

Normality Tests
Normality Tests for: tst_scores (sample size:170)
Kolmogorov Smirnov:
K-S Statistic: 0.0921306
Significance Levels: .20 .15 .10 .05 .01
K-S Critical Values: 0.0921 0.0821 0.0874 0.0936 0.1043
Test Result: FAIL FAIL ACCEPT ACCEPT ACCEPT
H0 = tst_scores is Normal.
(Normal if ACCEPT.)
Anderson Darling Test:
A^2 = 1.9165
A^2* = 1.9251
p = 0.0001
H0 = tst_scores is Non-normal.
(Normal if p > alpha.)
Select Variable: tst_scores
Close

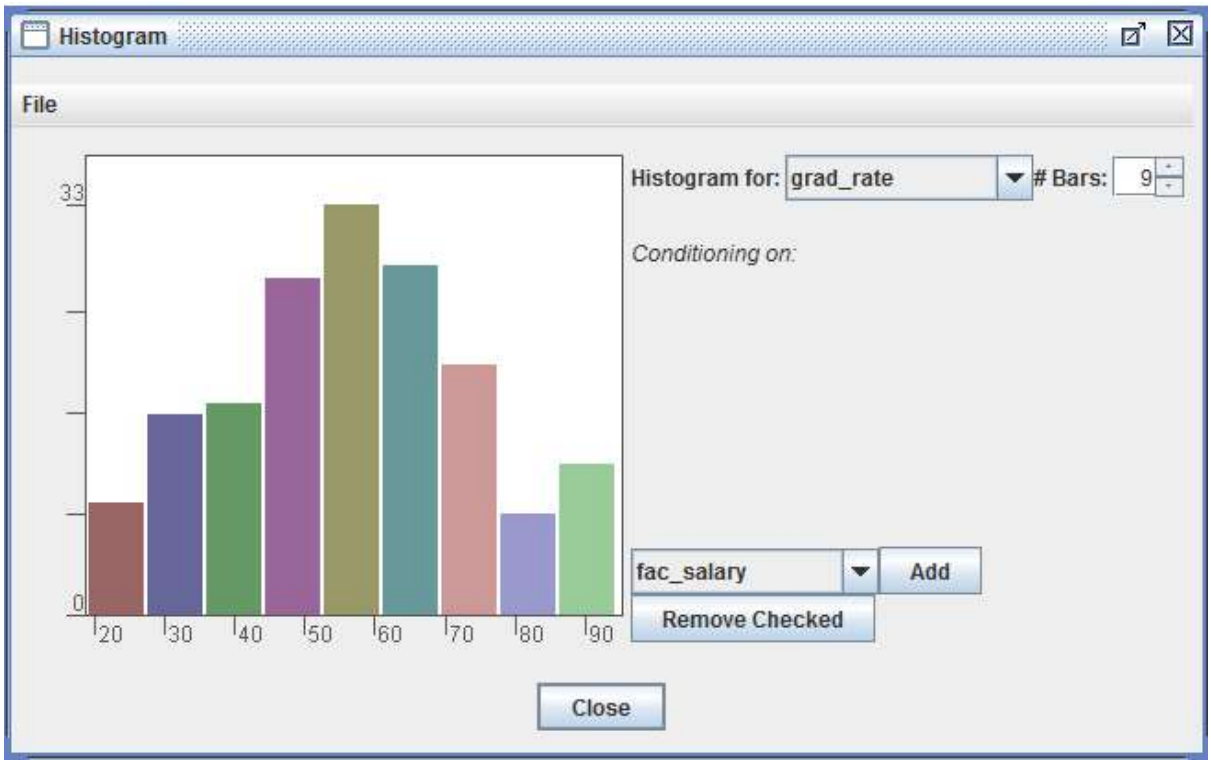


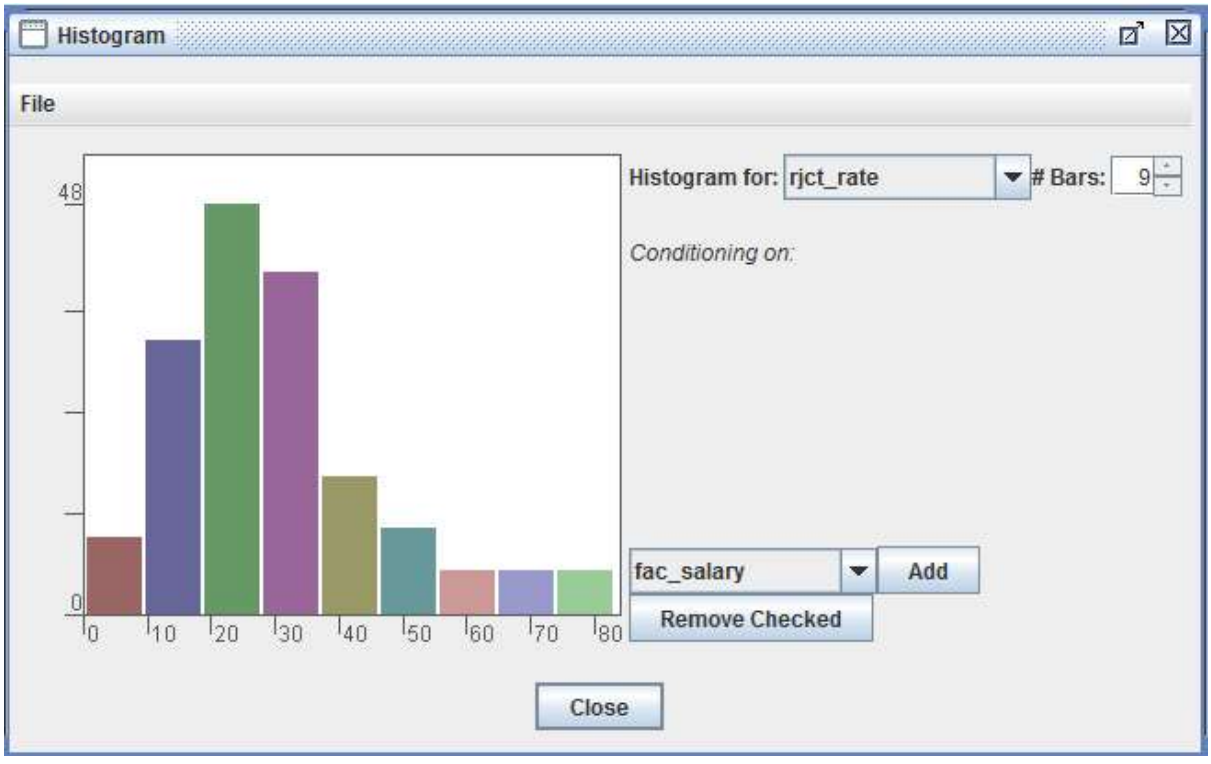


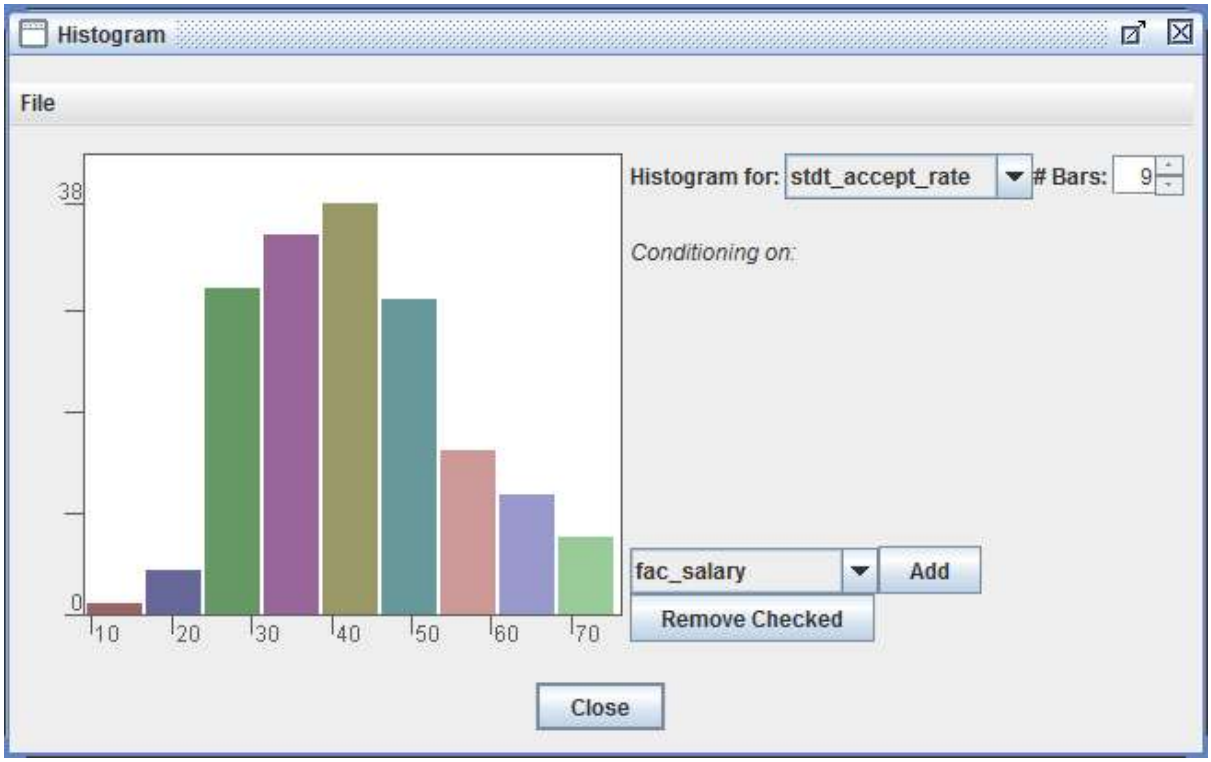
The following histograms also prove this fact. We increased the bin sizes also to see if there were any variations. But did not find much.


Histogram
File
65
0
10000
20000
30000
Histogram for: spending_per_stdt
Bars: 9
Conditioning on:
fac_salary
Add
Remove Checked
Close

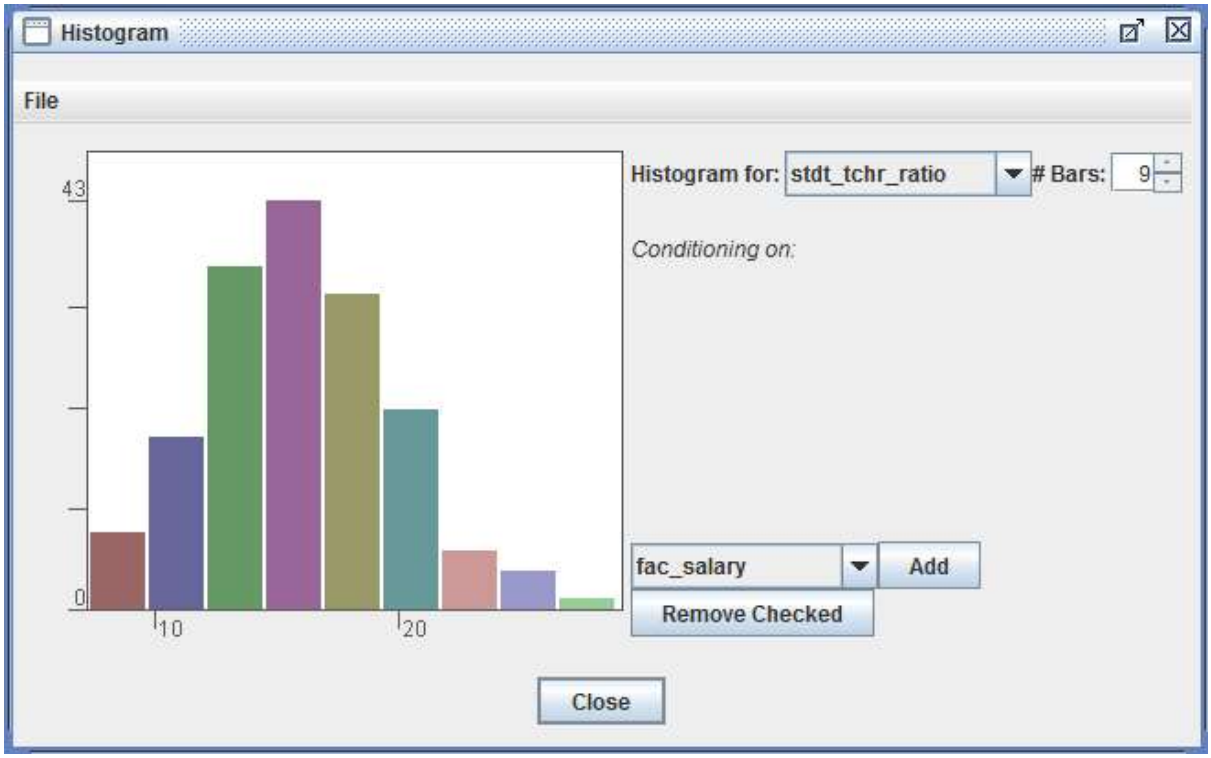

Histogram
File
37
0
40000
50000
60000
70000
80000
Histogram for: fac_salary
Bars: 9
Conditioning on:
grad_rate
Add
Remove Checked
Close


Histogram
File
33
0
20
30
40
50
60
70
80
90
Histogram for: grad_rate
Bars: 9
Conditioning on:
fac_salary
Add
Remove Checked
Close


Histogram
File
48
0
0
10
20
30
40
50
60
70
80
Histogram for: rjct_rate
Bars: 9
Conditioning on:
fac_salary
Add
Remove Checked
Close


Histogram
File
38
0
10
20
30
40
50
60
70
Histogram for:
stdt_accept_rate
Bars:
9
Conditioning on:
fac_salary
Add
Remove Checked
Close


Histogram
File
52
0
10
20
30
40
50
60
70
80
90
Histogram for:
stdt_clss_stndng
Bars:
9
Conditioning on:
fac_salary
Add
Remove Checked
Close


Histogram
File
43
0
10
20
Histogram for:
stdt_tchr_ratio
Bars:
9
Conditioning on:
fac_salary
Add
Remove Checked
Close


Histogram
File
42
0
50
60
70
80
Histogram for:
tst_scores
Bars:
9
Conditioning on:
fac_salary
Add
Remove Checked
Close

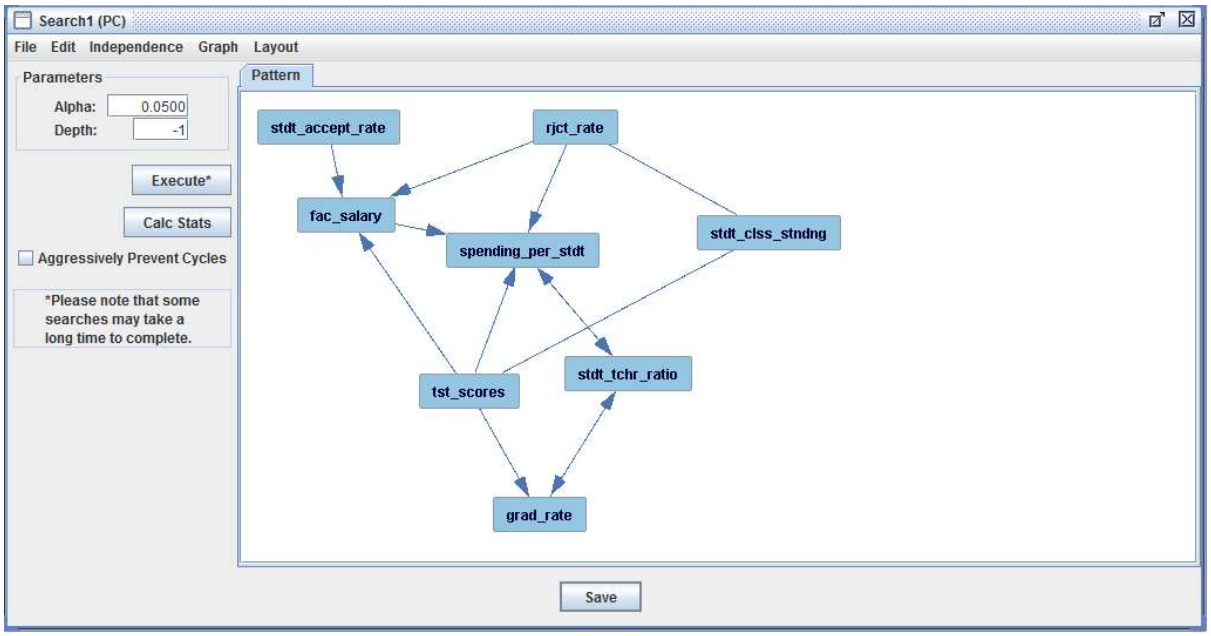
PC Algorithm:

The next step was to run the PC algorithm search for the data that we have. The analysis was done in two stage: Without any pre-knowledge and with pre-knowledge. We tested for various significance levels as shown below.

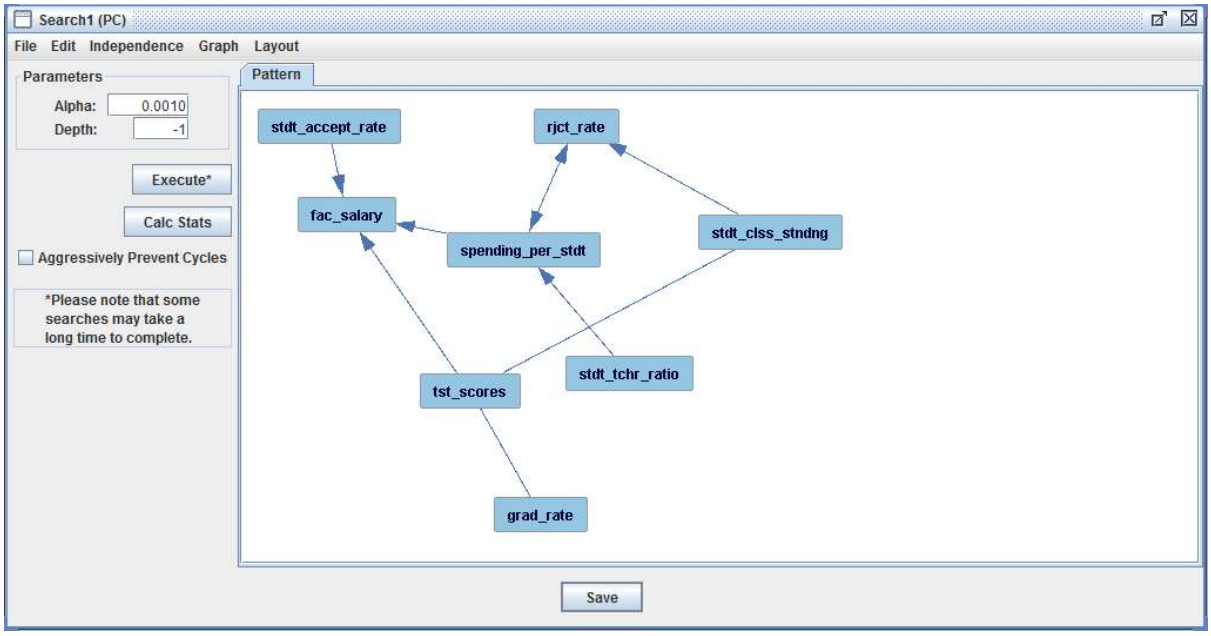
Without Pre-Knowledge:

Without any pre-knowledge, we just put the data into out PC search algorithm and it gave the following causal graph. This was totally random. The results did not show any specific valuable information. Majority of nodes were hitting on spending_stdt_student. We checked this for various significance levels as shown below.

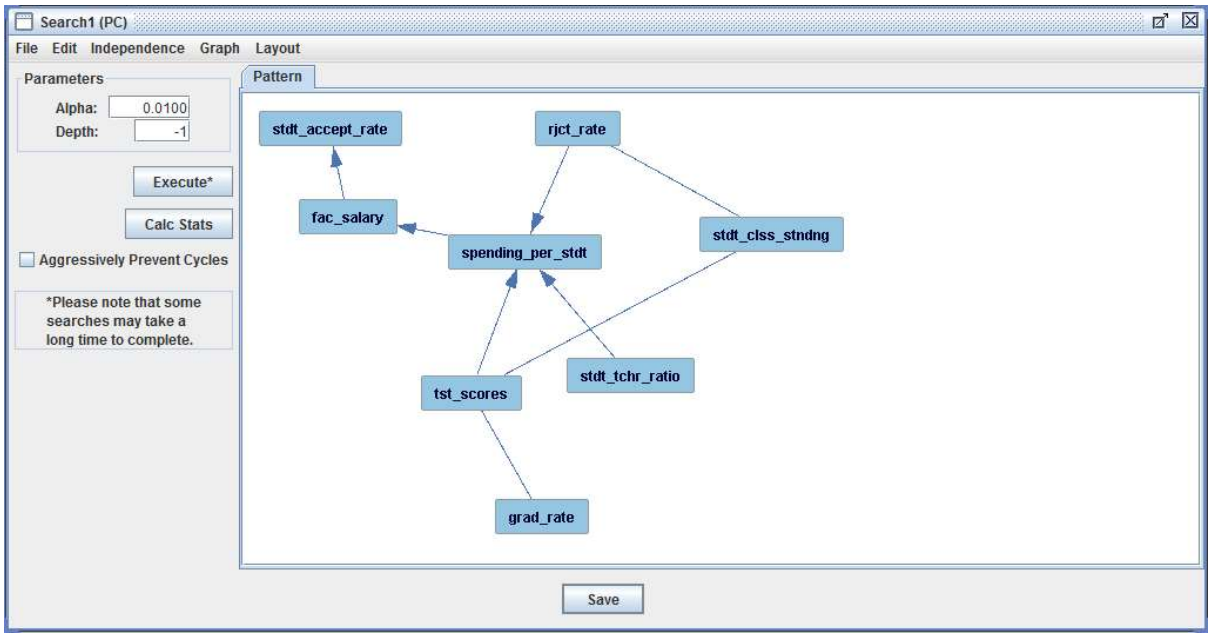
Significance: p = 0.05: Gives a more connected graph.



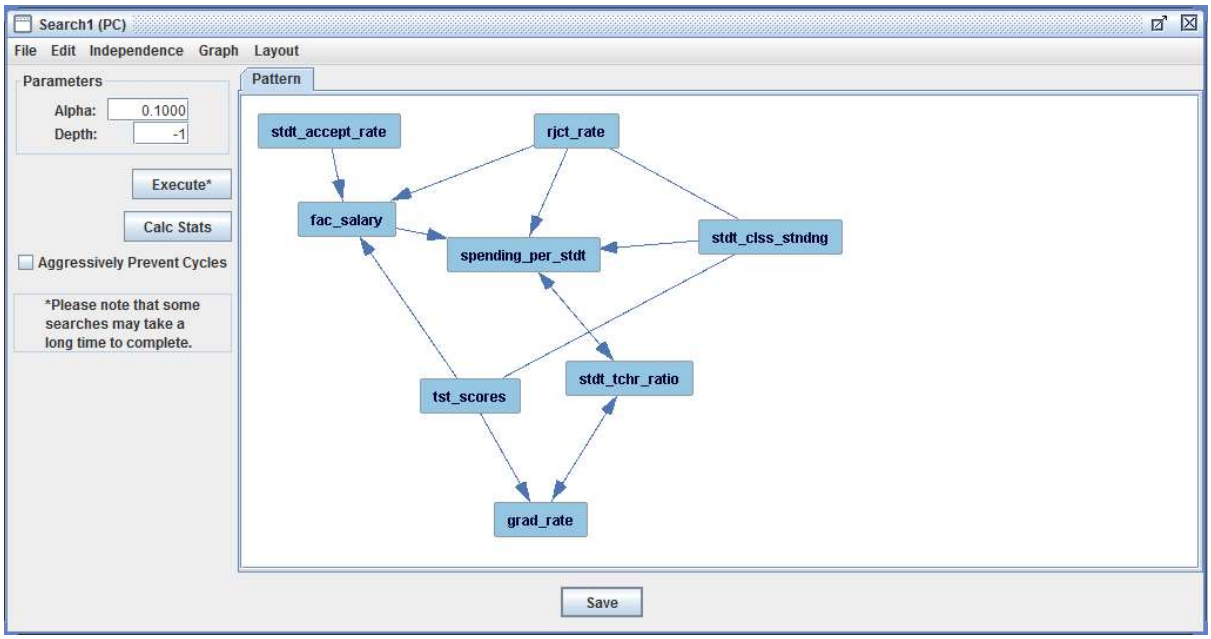
Significance: p= 0.001: Connections are less and some have changed orientations



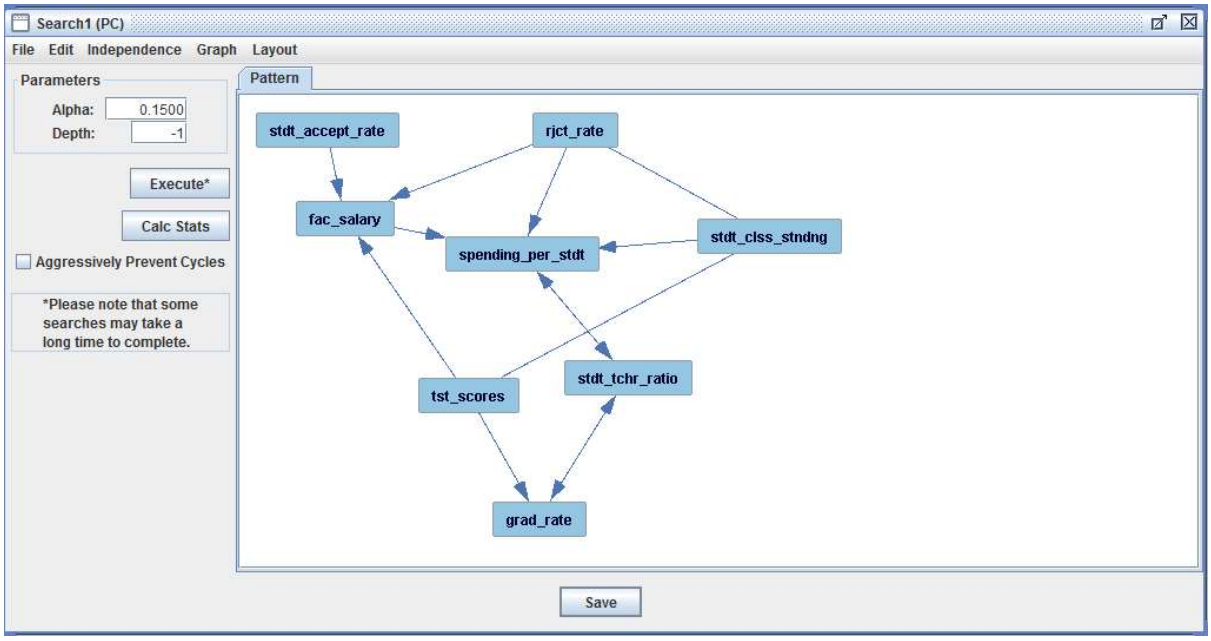
Significance: p= 0.01: Lesser connections and directions



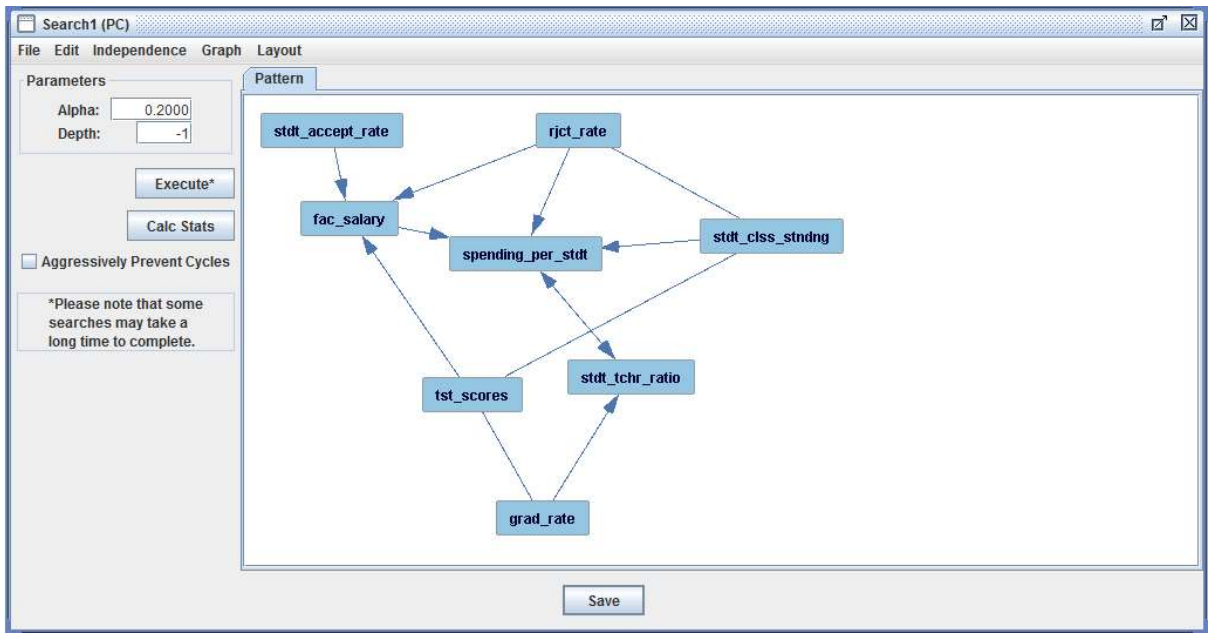
Significance: p= 0.1: Gave more connections but close to 0.05



Significance: p= 0.15

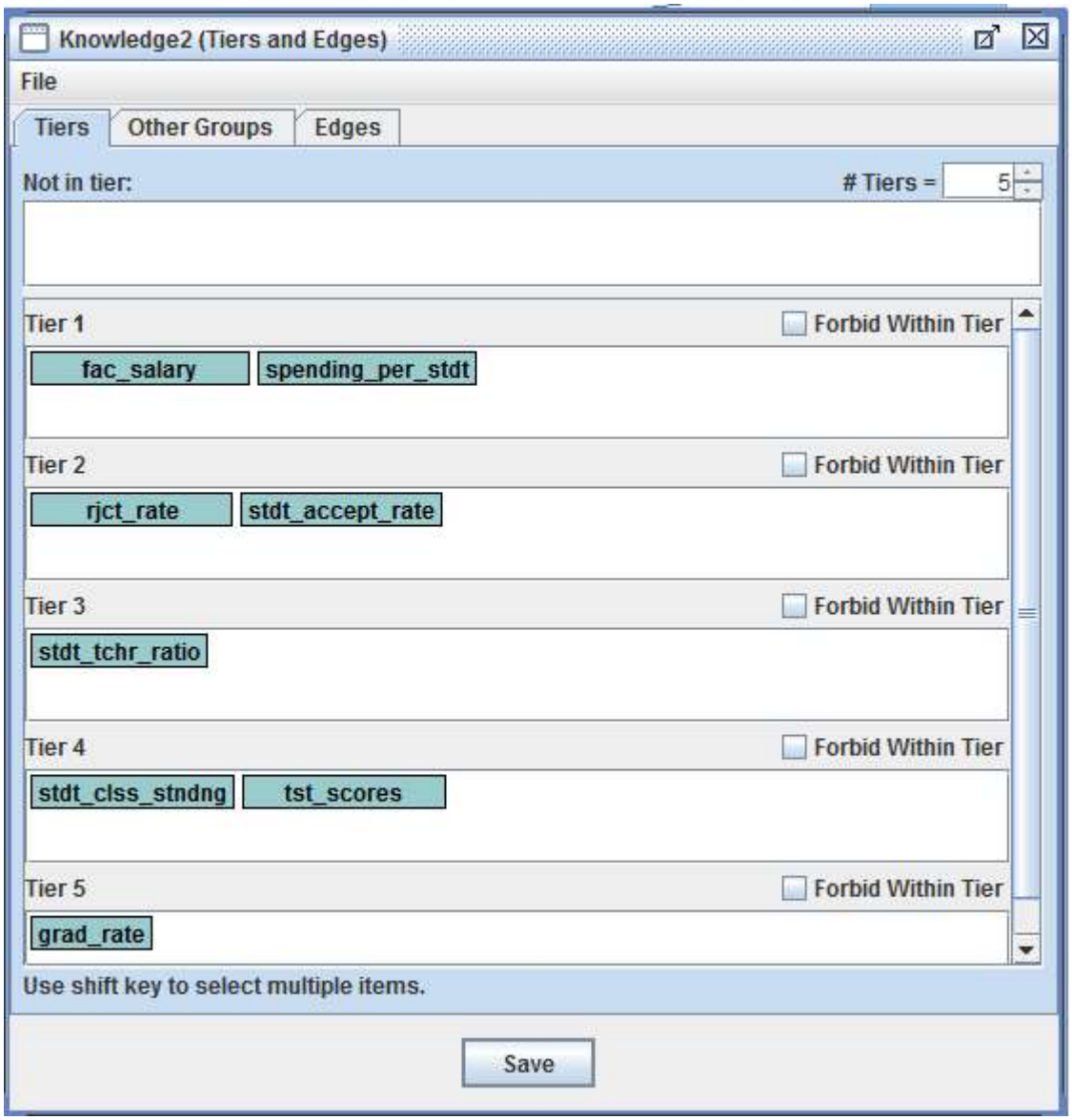


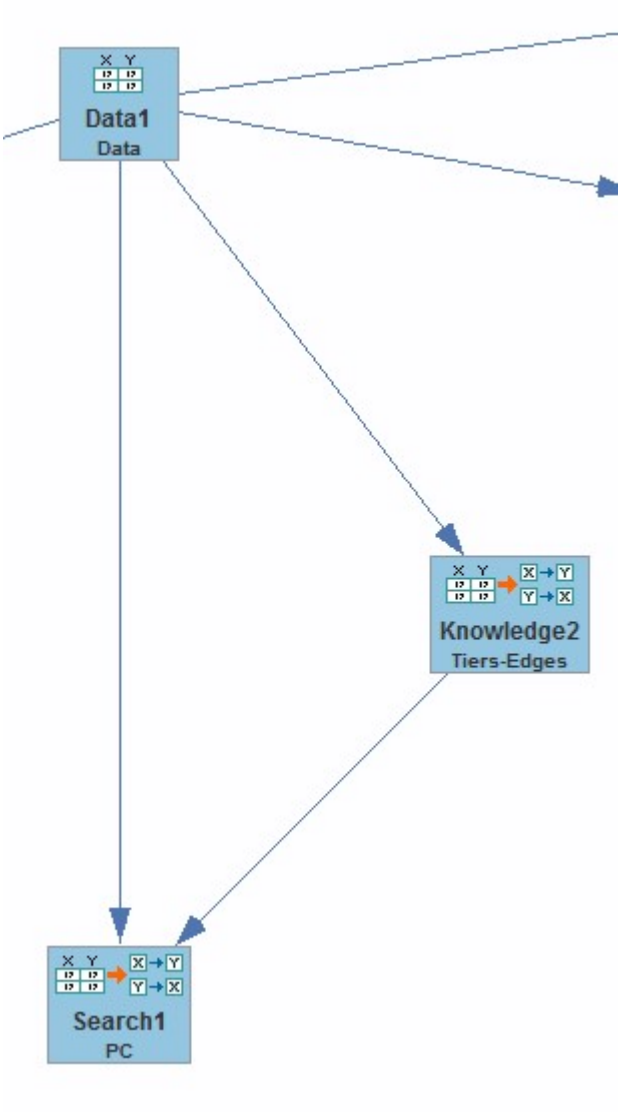
Significance: p= 0.2



With Pre-knowledge: Temporal Order:

Please note we have seen a relation between student_teacher_ratio and graduate rate and decided to put that into a tier just after the student acceptance and reject rates and before the testscores and class standings, because we feel that the student_teacher_ratio would be determined once the number of students joining are decided.



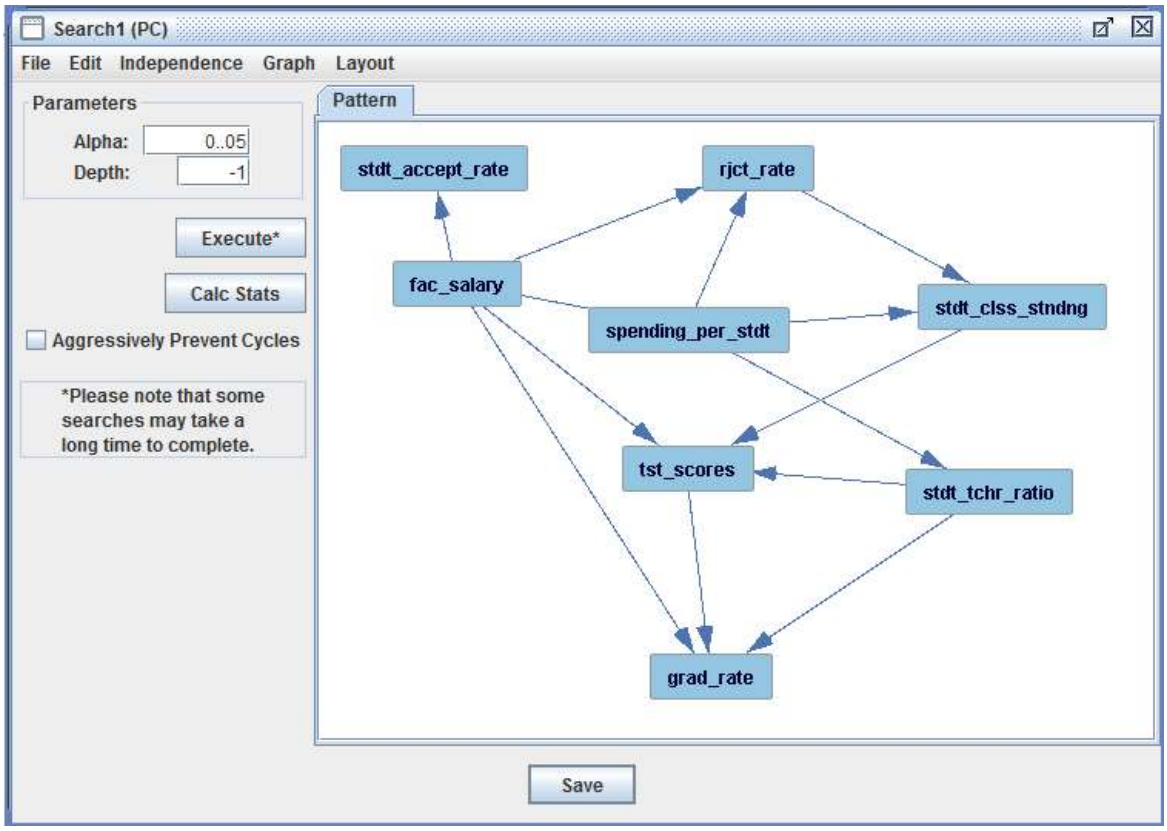


Interestingly we saw that now we have a graph which looked more like the one predicted in the paper. Most variables pass through tst_scores and determine the graduation rate.

However, there seems to be a direct relation between std_teacher_ratio and graduation rate as discussed above. This was not seen in the paper for the 1992 data. This relation makes sense too, the better the student_teacher_ratio the more attention a student gets, hence can influence better understanding in class and a higher graduation rate.

We also noted that there is a strange relation shown between fac_salary and rjct_rate. In the analysis below that, we tried removing the edge and seeing the graph. This brought about an interesting change, the connection between std_teacher_ratio and tst_scores were removed, but there is an addition of an edge between spending_per_stdt and tst_scores, which might signal that, fac_salary plays a role in the std_teacher_ratio, spending_per_stdt and hence the tst_scores which again affects the graduation rate.

With Significance: p = 0.05



With Significance: p = 0.05 and edge between fac_salary and rjct_rate removed.