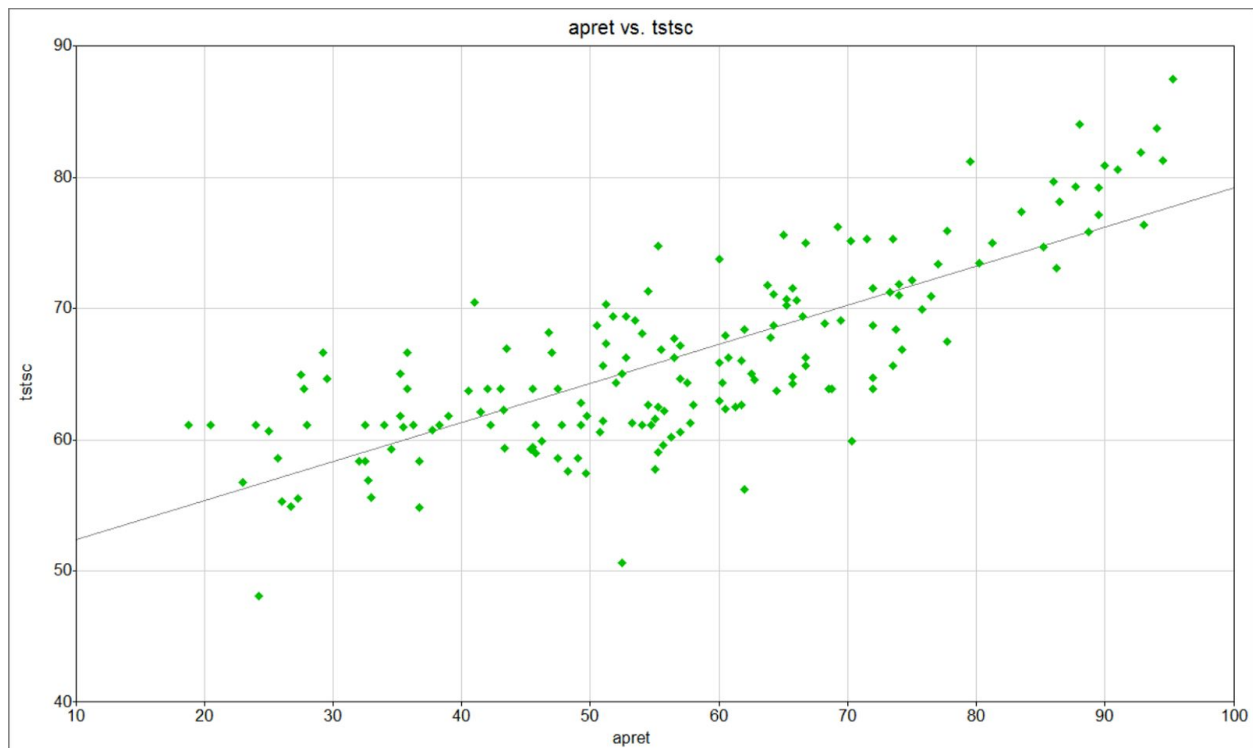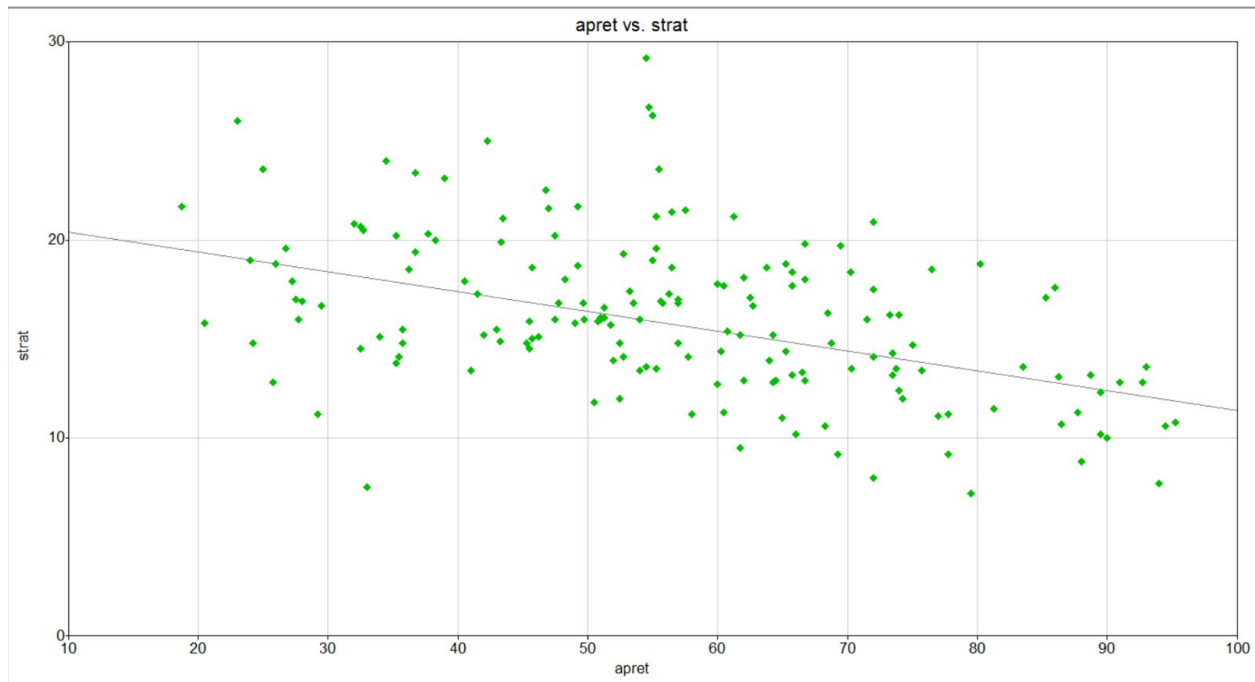# Assignment 5: Causal Relationship Discovery

Zhaoyan Ai (zha4), Ja-jan Hsu (jah247), Leilei Liu (lel74)

In the article, the two authors tried to find out the reasons for low freshman retention rate in US colleges. One apparently robust finding in this article is that student retention is directly related to the average test scores and high school class standing of the incoming freshmen.
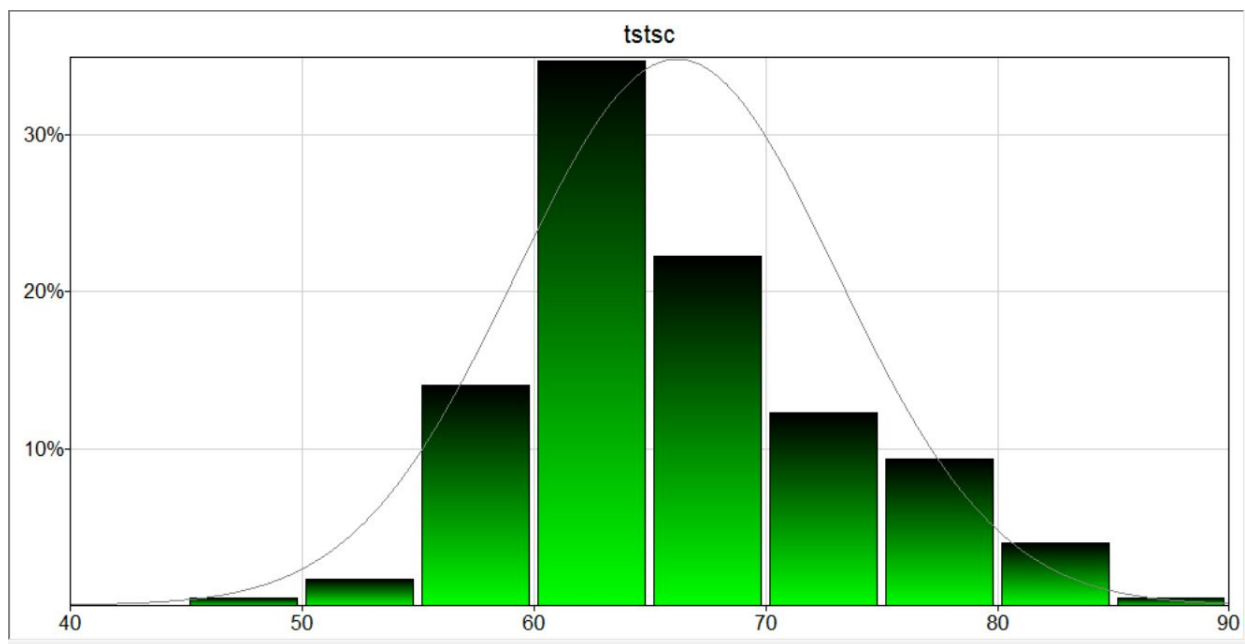
As we can see from the directed acyclic graph (DAG), both tstsc (test score) and strat (student teacher ratio) are parents of apret (average retention rate). Here we first plotted apret vs. tstsc and it is obvious that there is a linear relationship between apret and tstsc, though this relationship is not very strong.
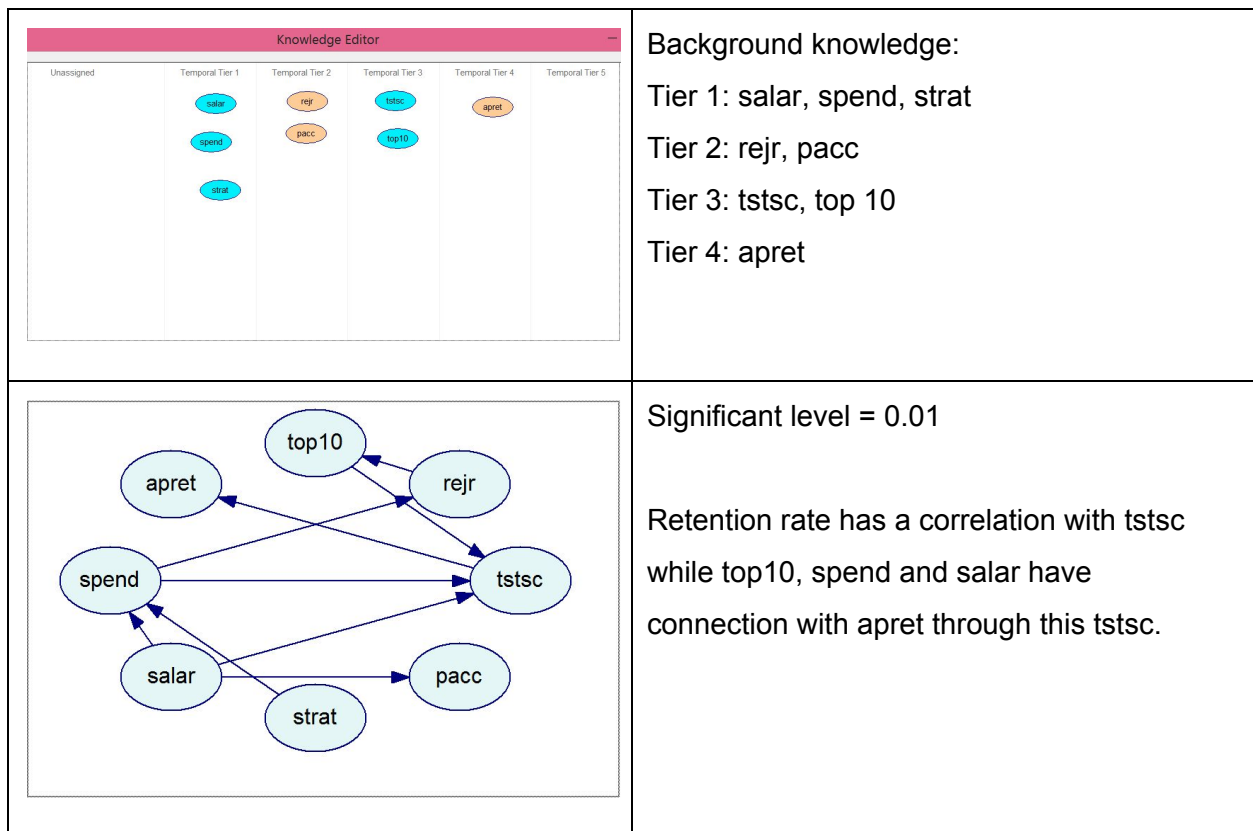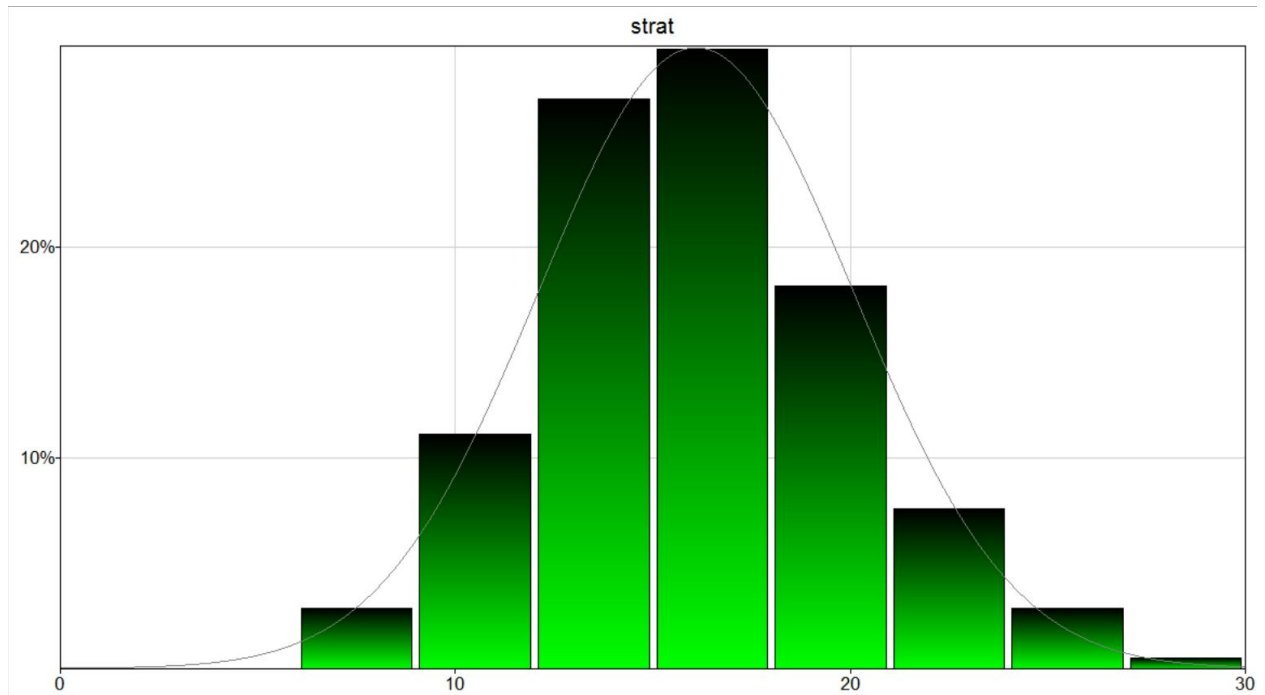

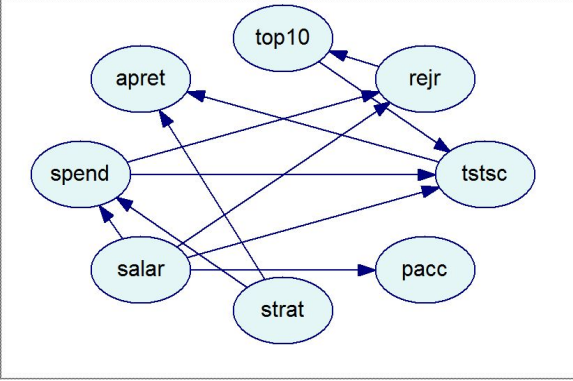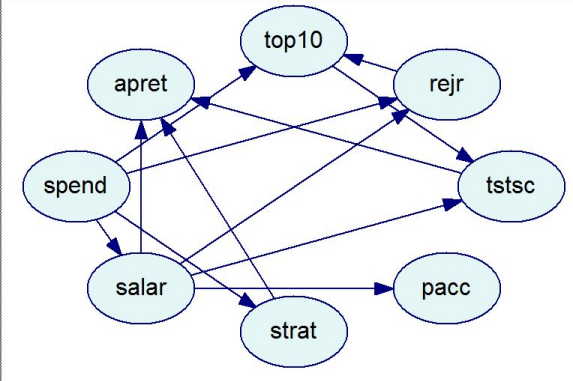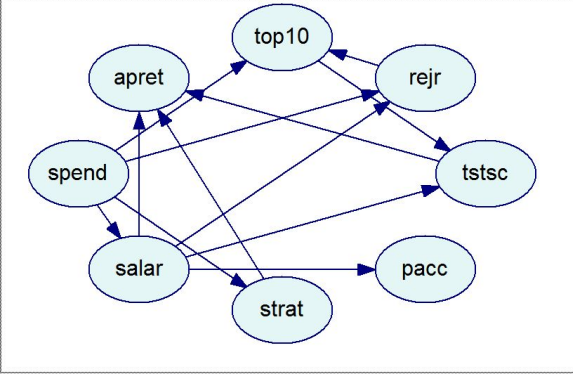There also seems to be a linear relationship between apret and strat as illustrated by the plot, this relationship is not as strong as that of apret vs. tstsc.

apret vs. strat

Further investigation into the distribution of tstsc and strat led to the observation that tstsc is slightly skewed-right while strat appears to more normal distributed. (Normal fit is displayed in both histograms)

tstsc

strat

Knowledge Editor	Background knowledge:
	Tier 1: salar, spend, strat Tier 2: rejr, pacc Tier 3: tstsc, top 10 Tier 4: apret
	Significant level = 0.01 Retention rate has a correlation with tstsc while top10, spend and salar have connection with apret through this tstsc.

	Significant level = 0.05 Retention rate has a correlation with tstsc and strat while top10, spend and salar have connection with apret through these two attributes.
	Significant level = 0.1 Retention rate has a correlation with tstsc, salar and strat while only spend and top 10 have connection with apret through salar/tstsc.
	Significant level = 0.15 Retention rate has a correlation with tstsc, salar and strat while only spend and top 10 have connection with apret through salar/tstsc.

From the four graph, we can conclude that as the significant level increases, apret has more and more parent nodes. When significant level is only 0.01, tstsc is the only possible cause of apret while as significant level increases to 0.05 and 0.1, strat and salar comes into play. We might infer that tstsc has a stronger relationship with apret than strat and then salar. No matter what the significant level is, top10 is always leading to tstsc, so we may infer that top10 is a non-negligible influencing factor.