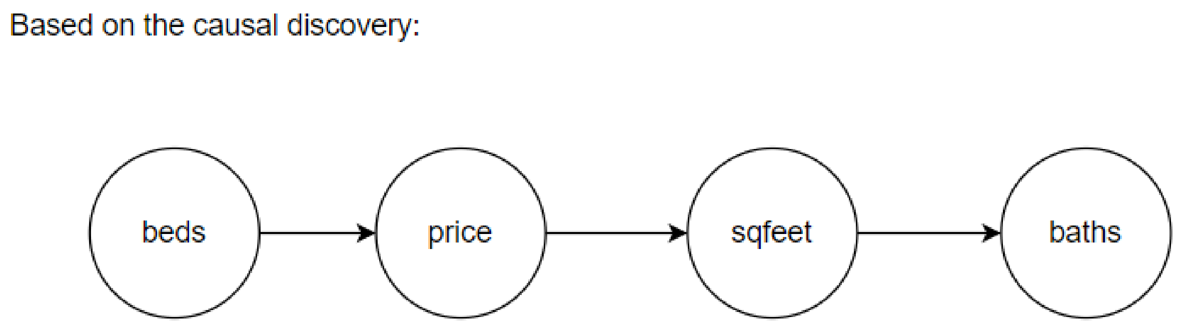
HOME | Timeline | Previous Week | Next Week

Monday 06/27

In 1:1 meeting with Roger, we started working on the makeGrid module. Since I was not able to draw any conclusions from the conditionality test results, I was asked to do CausalDiscovery on the data to see if there are any causal relationships between the variables.

I faced another error in the CausalDiscovery, that pointed to cScan.py of *because* module, but the error log was leading to the dead end, which may be because of, as later pointed out by Roger, memory overflow or something similar. But solving this issue, we found a few modifications that have to be done at the Causality bundle. After making those changes I have raised the pull request.

Based on the results of the CausalDiscovery, I was able to draw the following causal model.



But logically speaking, this model is not fitting the real-world case. so we have decided to look deeper into the model and dataset.

Some issues that I found today are:

1. When I combined the probTest with causal tests, why is it taking a long time to complete? but individually taking lesser than a minute?
2. Even with the same dataset and same code, executing multiple times, I am getting different models. What is the reason?

Tuesday 06/28

In a team meeting, I asked the reason for different results of the CausalDiscovery for the same code and data on different executions. Roger pointed out using the seed in the Causality toolkit to make the results consistent. But even after using the seed, I was getting different results.

I wanted to include 'types' variable in the housing analysis, but this was of string type, so I had to encode it. But to learn this I was re-implementing the LabelEncoder example from the ML_Core test folder. But there was an issue with importing the library. I had set up a meeting with Lili to figure out the issues with LabelEncoder from ML_Core bundle.

I have started working on the grid module.

Wednesday 06/29

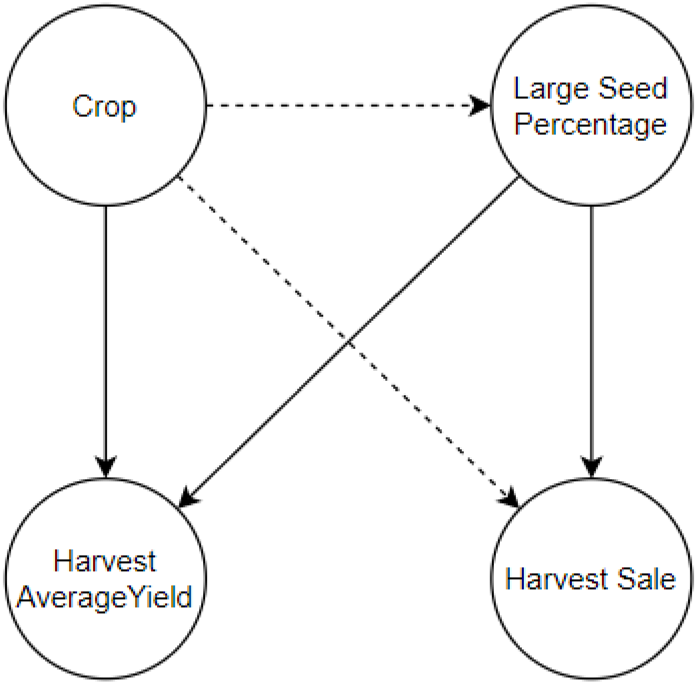
I have received the update from Lili on the LabelEncoder issue. the fix at ML_Core bundle has solved the issue in labelEncoder.

I have started learning more about the ECL language and its features from Introduction to ECL Part 2 course from LexisNexis. I have continued working on the grid module.

Thursday 06/30

I have started analyzing the AgX record schema for the possible hypothesis. I found a hypothesis which is as follows.

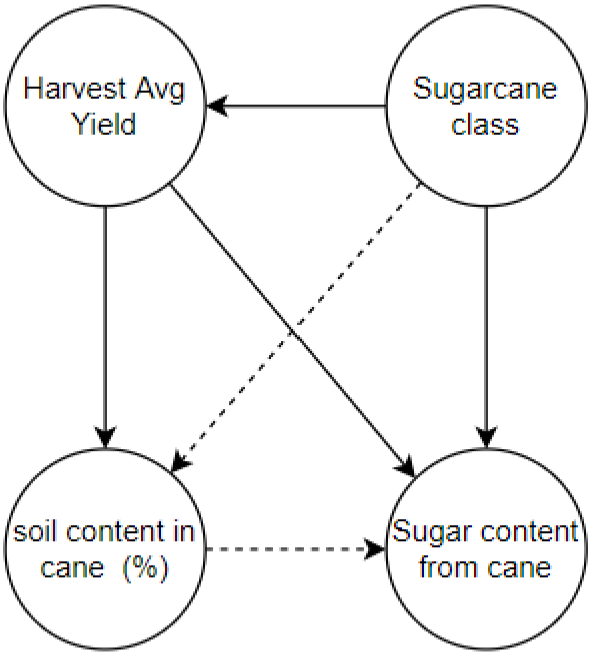
```
Depending on the type of the crop, will the large seeds affects the average yield?
How does large seeds affects the sale of the crop?
What is the effect of crop type on the Harvest Sale?
Does the yield affect the Harvest sale?
```



Grain and Seed harvest

Looking at the AgX record schema again, I wanted to find some intuitive hypotheses. So I started looking at sugarcane harvest data. I found another hypothesis which is as follows.

```
Considering we have different classes of sugarcane, does the yield of the crop
depend on the class of the crop?
What is the effect of the class on the sugar content that can be extracted from
the cane?
Does the soil percentage in the cane affected by class and the average yield?
How does soil percentage affects the sugar content of the cone?
```



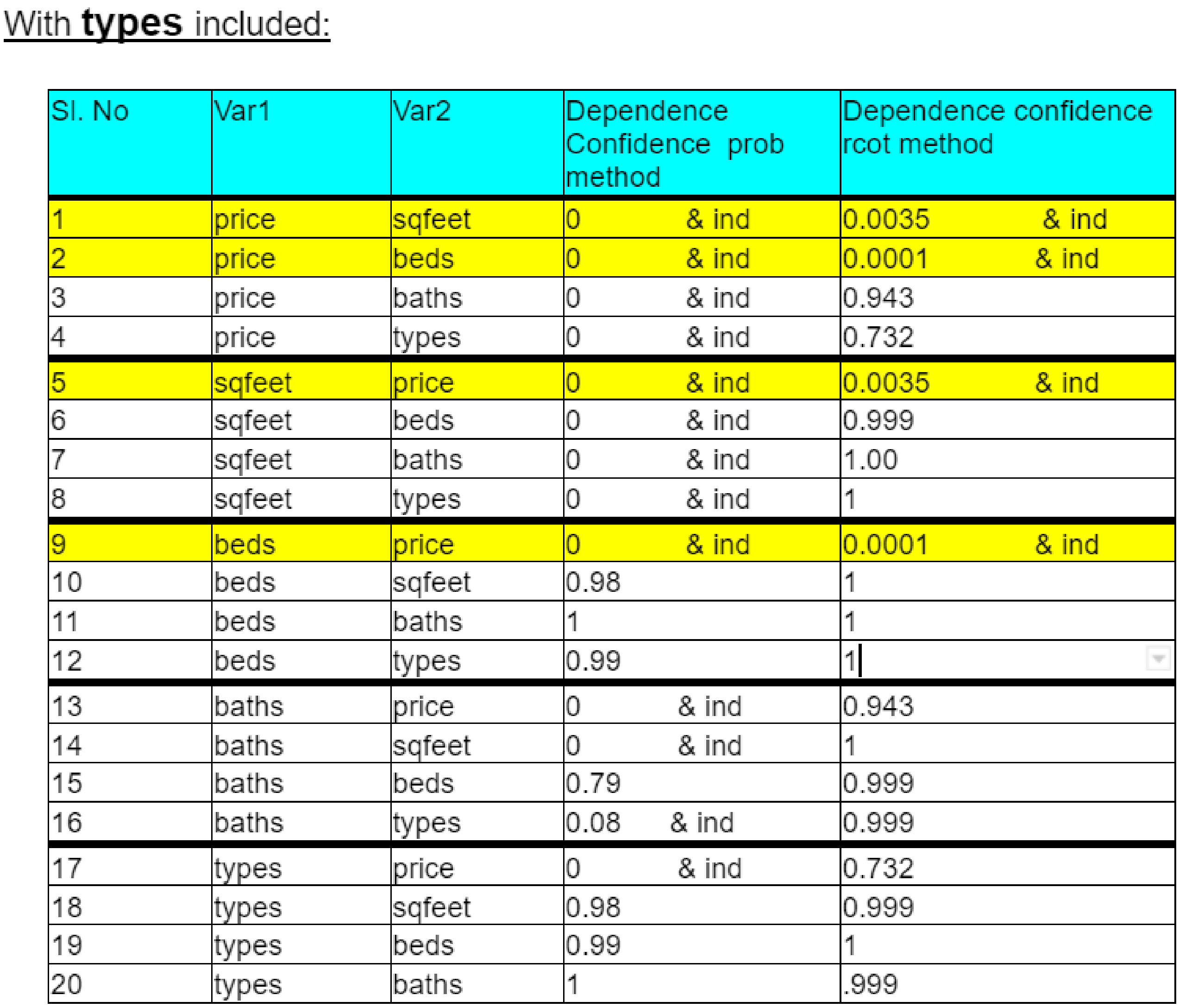
Sugarcane Harvest

Friday 07/01

I have implemented labelEncoder for housing *types* in the housing dataset. With the *types* being included in the dataset, now the total size will be 384977 rows. After normalizing the data, the total size is 1.92 million. And for each dependency test, it is taking about 18 minutes.

Results are interesting now when included the *types* parameter.

```
'price' is independent of 'sqfeet' and independent of 'beds' as well.
```



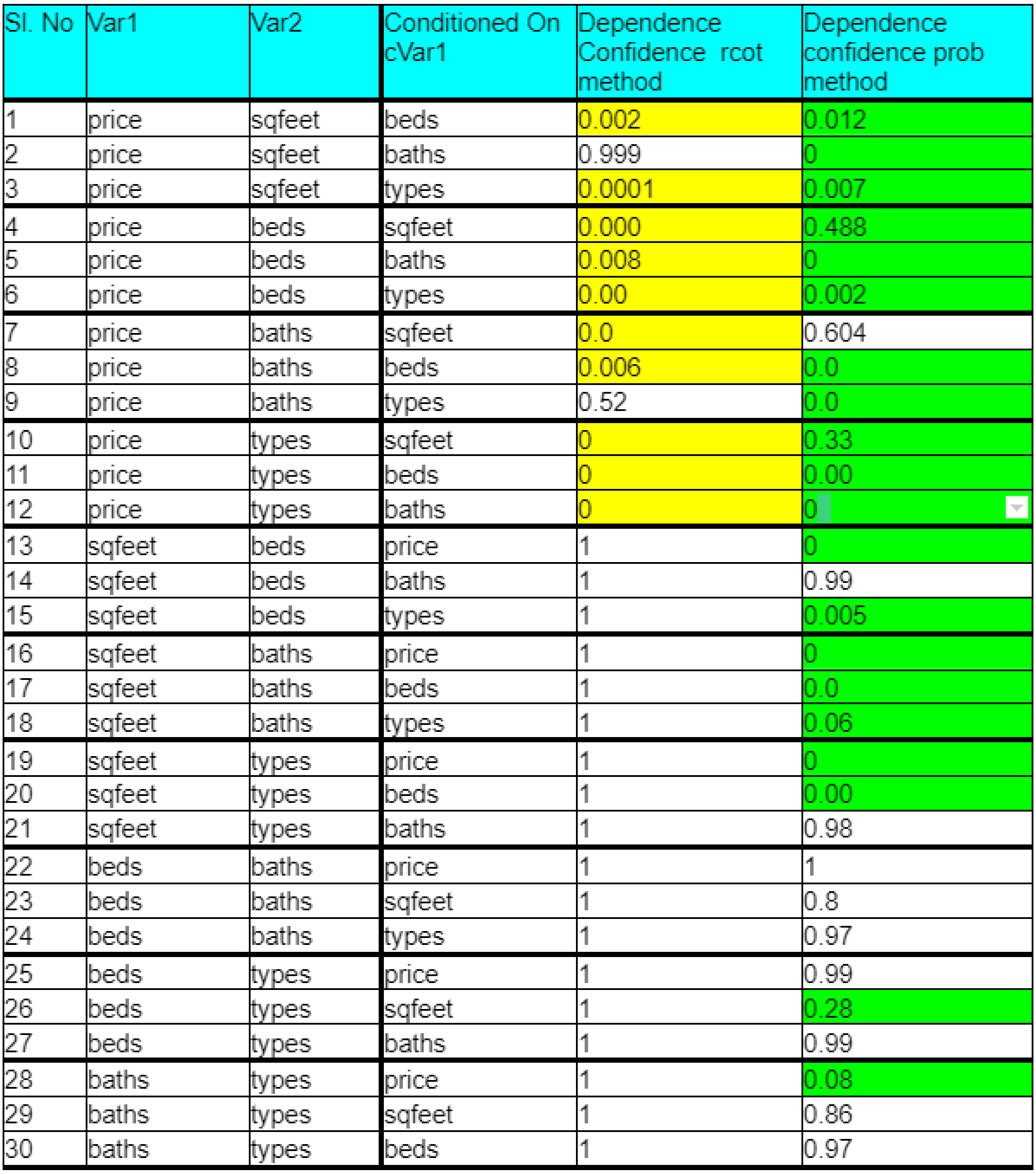
With **types** included:

Sl. No	Var1	Var2	Dependence Confidence prob method	Dependence confidence rcot method
1	price	sqfeet	0 & ind	0.0035 & ind
2	price	beds	0 & ind	0.0001 & ind
3	price	baths	0 & ind	0.943
4	price	types	0 & ind	0.732
5	sqfeet	price	0 & ind	0.0035 & ind
6	sqfeet	beds	0 & ind	0.999
7	sqfeet	baths	0 & ind	1.00
8	sqfeet	types	0 & ind	1
9	beds	price	0 & ind	0.0001 & ind
10	beds	sqfeet	0.98	1
11	beds	baths	1	1
12	beds	types	0.99	1
13	baths	price	0 & ind	0.943
14	baths	sqfeet	0 & ind	1
15	baths	beds	0.79	0.999
16	baths	types	0.08 & ind	0.999
17	types	price	0 & ind	0.732
18	types	sqfeet	0.98	0.999
19	types	beds	0.99	1
20	types	baths	1	.999

Another conclusion that can be drawn from this is that *prob* test result values are not symmetrical, which is expected probability behavior.

Implementing conditional dependency tests conditioned on 2 variables are as follows.

Conditioned on 1 variable with **types**



Sl. No	Var1	Var2	Conditioned On cVar1	Dependence Confidence rcot method	Dependence confidence prob method
1	price	sqfeet	beds	0.002	0.012
2	price	sqfeet	baths	0.999	0
3	price	sqfeet	types	0.0001	0.007
4	price	beds	sqfeet	0.000	0.488
5	price	beds	baths	0.008	0
6	price	beds	types	0.00	0.002
7	price	baths	sqfeet	0.0	0.604
8	price	baths	beds	0.006	0.0
9	price	baths	types	0.52	0.0
10	price	types	sqfeet	0	0.33
11	price	types	beds	0	0.00
12	price	types	baths	0	0
13	sqfeet	beds	price	1	0
14	sqfeet	beds	baths	1	0.99
15	sqfeet	beds	types	1	0.005
16	sqfeet	baths	price	1	0
17	sqfeet	baths	beds	1	0.0
18	sqfeet	baths	types	1	0.06
19	sqfeet	types	price	1	0
20	sqfeet	types	beds	1	0.00
21	sqfeet	types	baths	1	0.98
22	beds	baths	price	1	1
23	beds	baths	sqfeet	1	0.8
24	beds	baths	types	1	0.97
25	beds	types	price	1	0.99
26	beds	types	sqfeet	1	0.28
27	beds	types	baths	1	0.99
28	baths	types	price	1	0.08
29	baths	types	sqfeet	1	0.86
30	baths	types	beds	1	0.97

I have to draw the dependency relations and causal models using these results.

HOME | Timeline | Previous Week | Next Week