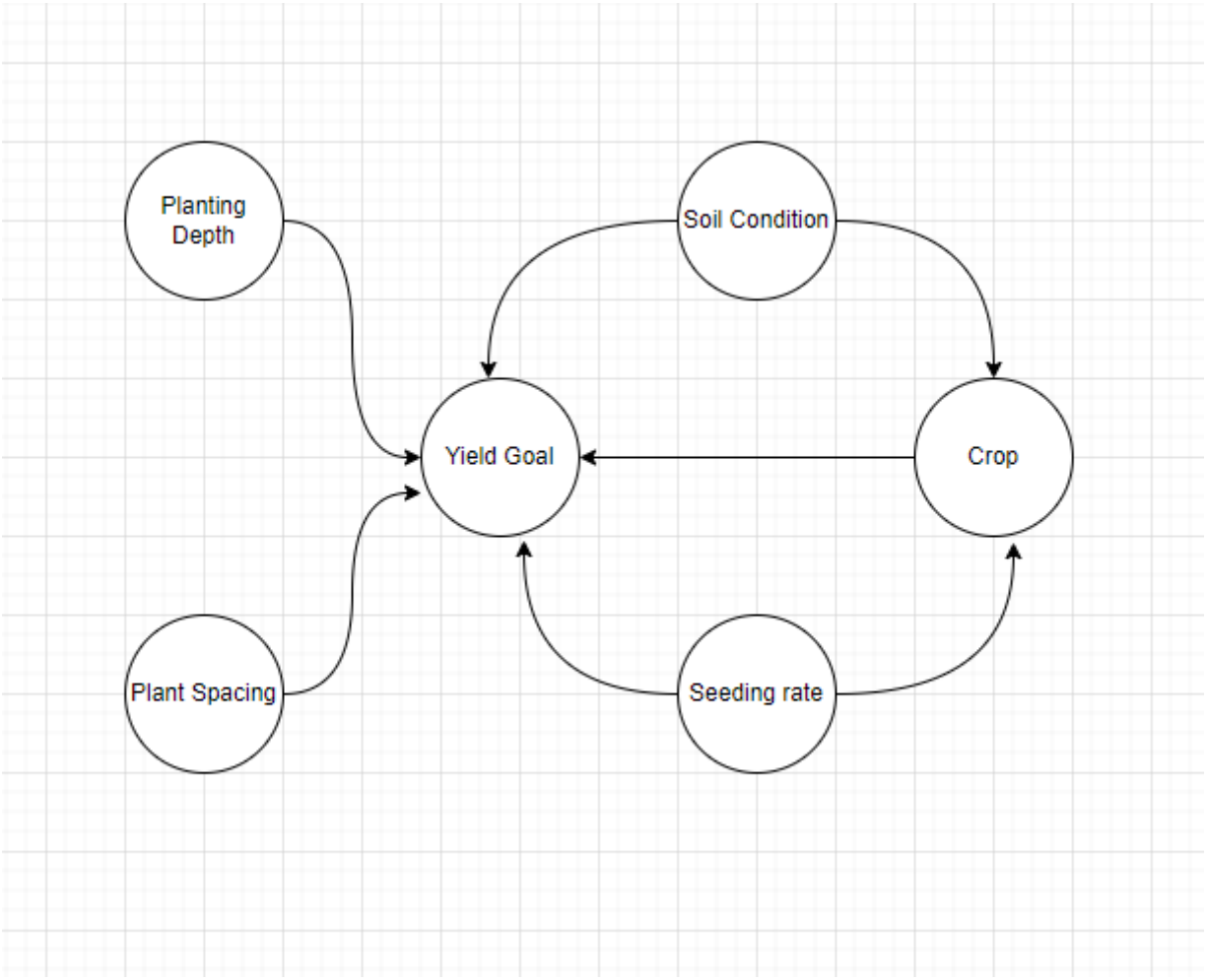
HOME | Timeline | Previous Week | Next Week

Monday 06/13

It's a fresh start to the week. I am going through the AgX dataset and built causal models for that.

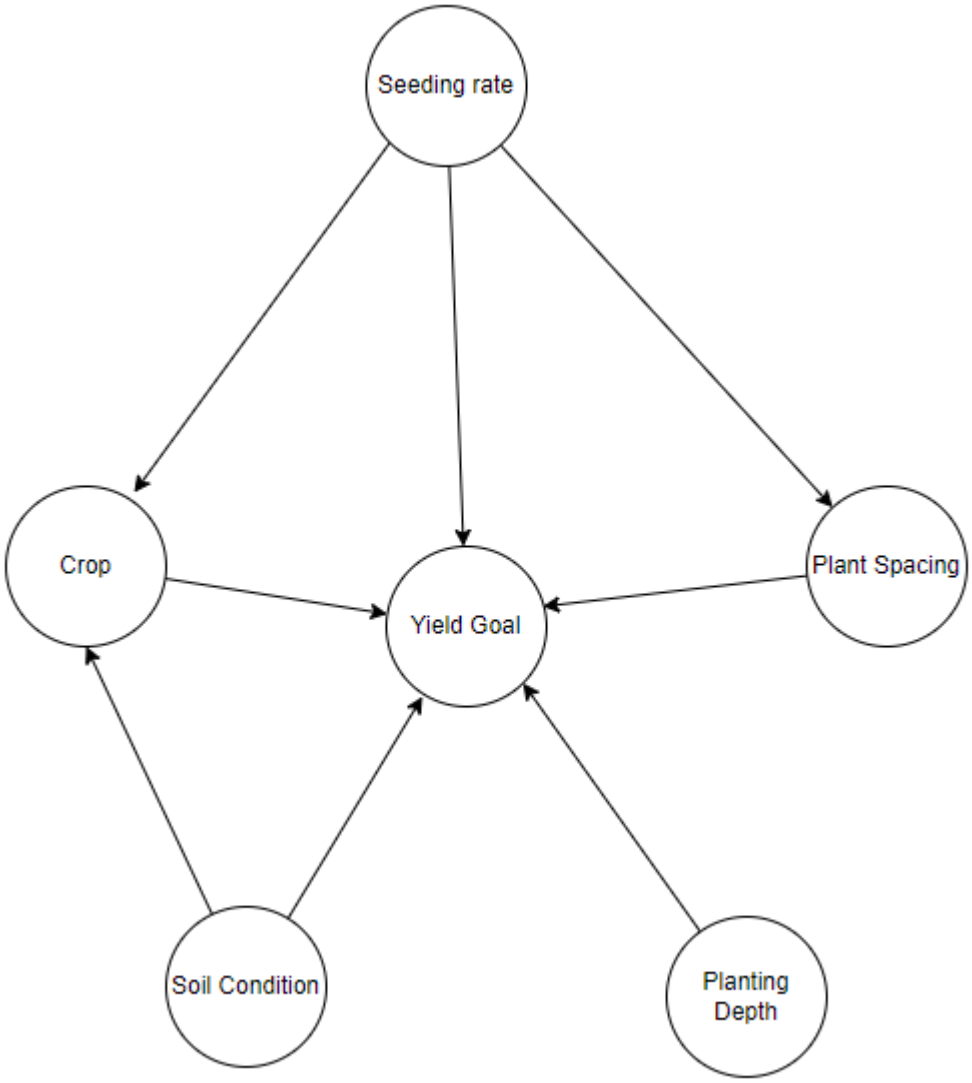


The thought process behind this model can be explained as follows:

In this model, does *Plant Spacing* and *Planting Depth* will have any effect on the *Yield Goal* of the crop? If the distance between two plants is less, it might reduce the sunlight reachability, and if the spacing is too much, there will be an irredundant waste of space. So I am going to include the distance between two plants in the model. And about, planting depth is obvious as, if the depth is extra, the seeds might not be able to germinate because of lack of water and sunlight, or excess heat.

Different types of crops are more suitable in different soil conditions. Based on this I assumed that *Soil condition* affects the yield goal, depending on the type of *crop*.

According to the dataset definition sheet, the *Seeding rate* is the number of seeds, plants, or trees being planted per acre. It depends on plant spacing, but what about crop? Does the number of trees/ seeds planted per acre depend on the crop type? Considering this second model is built as shown below.



After this, I have to discuss with Roger to choose the best model and for any suggestions.

Tuesday 06/14

In the AgX dataset, *Soil Condition* is an ordinal variable. So I wanted to discuss with Roger, whether is it better to encode and use it in the model or if will it be hindering the prediction.

For the covid19 model, I am thinking of finding a way to include week numbers ranging from 1 to 52 for the COVID19 dataset, so that I can keep track of which week refers to which season.

In the Team meeting, we discussed the following:

- A yield goal is an expected goal value, it is an estimation. So it is not the ideal choice for the causal model.
- Crop affects the seed rate, so I have to reverse the direction between crop and seeding rate.
- I should look for a definitive variable to include in the hypothesis. So these causal models have been discarded.
- The Latest version of the *because* module has been pushed, so I can continue testing the models and use them for causality analysis.

- I have been asked to proceed with any one of the datasets to apply causality. So I have decided to go with the Housing dataset to find a relationship between variables.

I was facing some issues with spraying the housing dataset file, Getting this error:

```
Failed: End of UTF record not found (wrong record terminator defined or need to increase maxRecordSize?) after processing 26476757 bytes!.
```

Once I have resolved this error, I started reading the housing dataset and verified the data types of each variable.

Wednesday 06/15

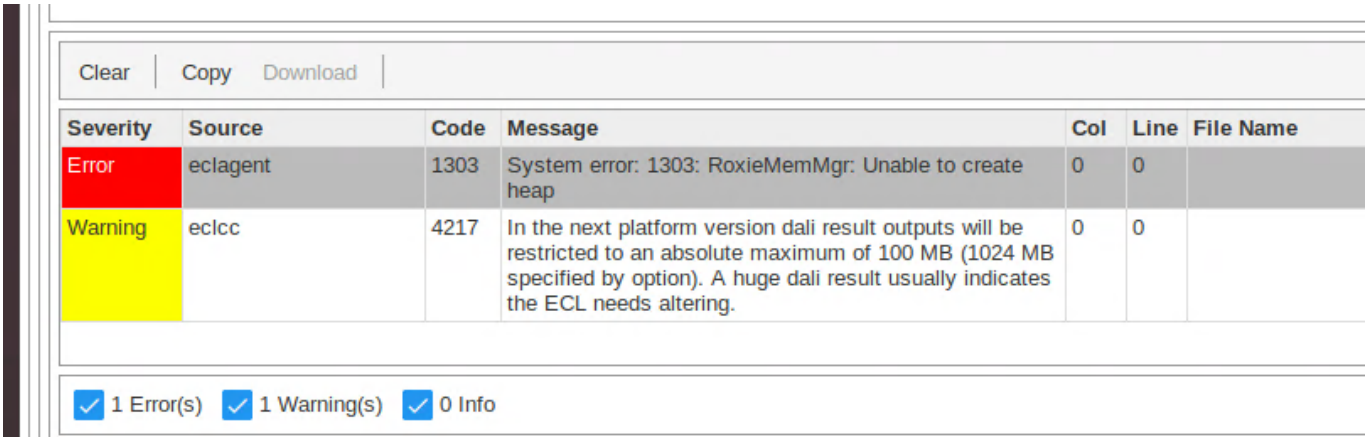
I have started analyzing the housing dataset and have observed the following.

1. There are 1307 values for which 'price' is 0.
2. There are 1036 rows for which 'sqfeet' is less than 120. (120 is the minimum square footage for a house. Ref: worldpopulationreview.com)
3. There are 3107 rows for which 'baths' is 0.
4. There are 10978 rows for which 'beds' is 0.
5. There is no entry for which both 'beds' and 'bath' are 0.

I had set up a meeting with Hugo to clarify some issues and doubts with ECL. In this meeting I learned about Data patterns and how can I get to know the best suitable variable type for each variable in the dataset. Along with that, I have also understood the importance of typing the exact variable type of each variable.

I have found outliers in the dataset. So I wanted to know what is the reasonable house size in sqfeet, below which I can delete the rows.

While using the dataset in coding, I have found that I am not able to display all the rows of the dataset. This is because of the restrictions in ECL memory management.



Thursday 06/16

To display all the data of the dataset, I can use

```
#OPTION('outputLimitMb','nn'); where nn is the number of MBs.
```

but there is an upper limit 100MBs for the ECL memory management. Even with the maxSize of 100MBs, I was not able to read all the data. But then I realized, there is no need to display all the data.

In the team meeting with Roger and Zheyu, Roger helped me to understand how to incorporate the causality toolkit from the *because* module which is in python, with the ECL Language to apply for a dataset. Transforming, Normalizing functions were discussed.

Further analyzing the data, I have found a few outliers in the dataset. For eg., The max value of the price was $2,768,307,249 for an *Apartment* with 2 beds & 1.5 baths located at Columbus, which seems an outlier. Similarly, the max size of the house was *8,388,607 sqfeet*, which is larger than Hawaii, having only 2 beds and 1 bath, and guess the cost of this apartment located in *Mankato, MN*. It is as cheap as it can be and at *$750* (If you guessed it correctly, give me a call. LOL). This is erroneous data, that might have been caused during the data entry.

So I have decided to apply this filter.

1. **500 <= Price <= 10000**, Assuming that the reasonable price for a house lies between that range.
2. **300 <= Square Feet <= 1200**, Assuming that the reasonable square footage for a house worth accommodating lies between that range.
3. **0 <= Bedrooms <= 4**, Assuming that the reasonable number of bedrooms for a house lies between that range.
4. **1 <= Bathrooms <= 4**, Assuming that the reasonable number of bathrooms for a house lies between that range.

Friday 06/17

I have worked on cleaning the dataset.

When applying *Normalize* function, I have found it cannot be applied directly to the *types* variables as it is a *STRING* type. So I am finding the ECL method to convert the *types* variables to *NUMERIC* types by encoding (1=apartment, 2=house, 3=condo etc). For this, I have found the *ENUM function*, but when applied with the sample example given in the documentation, I am facing the error.

Also, I have to apply the filtering option of selecting only required columns out of the whole dataset.

Along with that, I have looked at a few tutorial videos on ECL coding from LexisNexis. And the Probability module in HPCC_Causality toolkit helps in applying different functions. I have found an error in HPCC_Causality 'continuousTest.ecl' file when understanding the implementation of Probability testing.

HOME | Timeline | Previous Week | Next Week