HOME | Timeline | Previous Week | Next Week

Monday 07/11 & Tuesday 07/12

To apply the causality toolkit to the real-world dataset, I kept on searching for some more datasets. I have looked at US Accidents dataset, where some of the possible hypotheses were,

1. What will be the effect of weather conditions on the severity of the accident?
2. How does the side of the lane (left, right) affect the severity of the accident?

But the accident severity was already quantified in the dataset, so discovered inferences will not be rational enough to draw reasonable conclusions. So I have decided to look at some more datasets.

In the meeting with Roger, we resolved the *isIndependent* test error in Because module. I have continued testing the *makeGrid* bundle. Till now it seemed fine as there are no issues.

While updating the Because module with recent changes, I have faced the error of missing library installations. Reported to Roger to include sklearn in the default installation of packages.

Wednesday 07/13

For the dependence test for Housing dataset of 1.72 million data rows in the normalized form, it took around 6hrs, and later I aborted the test.

Scope	TimeElapsed
compile:compile c++:W20220705-101828.cpp	345.985ms
compile:compile c++:W20220705-101828.res.s	6.567ms
compile:compile c++:W20220705-101828_1.cpp	678.891ms
compile:compile c++:link	98.438ms
compile	1.020s
compile:parse	87.472ms
compile:generate	52.459ms
compile:generate:write c++	1.311ms
compile:compile c++	778.882ms
w1:graph1:sg1	16.000ms
w1:graph1:sg4	5.041s
w1:graph1:sg16	4.165s
w1:graph1:sg29	6h15m51s

So I ran *continuous test* and *mixed test* to check the time taken to execute a dependent and independent test for a synthetic dataset of the same size. And surprisingly, it was done in less than minutes. In python with the Because module, it executed in a fraction of a second. So I have decided to look into this issue. Since this issue was noticed after including the labelEncoded data of *types* variable for analysis, I thought of running multiple dependent and independent tests with and without including *types* variable.

After restarting the system and VM, I have observed that this issue is not related to labelEncoding of *types* variable. And to add more surprise, it took 2 minutes to execute the test.

			Total execution time	single query time	total execution time	Total query time	Time per query
	30 queries	including types	09s	2.20s	1m21s	1m13s	73/30=2.43
	24 queries	without types	10s	2.44s	1m02ss	55s	55/24=2.29

Then the time issue of the previous longer test might be due to the assigned Virtual Memory and heap. Assuming so, I tried to increase my VM memory to 16GB and it failed.

I have also kept on looking for some more datasets for the possible hypothesis.

Thursday 07/14 & Friday 07/15

I have received the CDC dataset schema from Roger. I have started looking at hypotheses from that.

I implemented the makeGrid module for including the three continuous variables. To integrate this with the discrete variables, I have to use the isDiscrete and histogram from the probability distribution of those variables. I am trying to extract the required details from the distribution but am not able to do it exactly.

I have also started analyzing the CDC dataset schema and looked at the real dataset for details. After that, I started cleaning the data and preparing it for analysis.

HOME | Timeline | Previous Week | Next Week