Exercise 1.3

Compare the two algorithms on the datasets T10I4D100K, T40I10D100K, chess,connect, mushroom, pumsb, pumsb star

Accuracy :

Compared by SON algorithm ,randomised algorithm has less accuracy in lower simple size .because randomised algorithm may generate degree with false positive and false negative

Efficiency:

In different transaction ,for the simple randomised algorithm ,we can resize the file efficiently which can reduce the scope of search.The SON algorithm have to divide file into chunks ,also once we have lower threshold , we need to handle much more candidates in every files. And SON algorithm need to do map reduce which cause much time for sorting. Hence ,SON algorithm has less efficiency.
In same transaction .there is no efficiency with two algorithms

T10I4D100K

Size is smaller than other files and different transactions .
Threshold : 1000
Simple randomised : 3931 ms
Son algorithm : 59784ms

Chess,connect, mushroom

Transactions are quite same and both algorithms can effectively calculate with a sufficiently large support threshold.

T10I4D100K,mushroom, pumsb, pumsb star

Item-sets are diverse and size are larger than others
Simple randomised : less efficient with files in lower threshold . But we have seen a significant improvement in larger threshold .
SON : get a accuracy result with enough support threshold

Exercise 1.4

Experiment with different sample sizes in the simple randomised algorithm such as 1, 2, 5, 10% and compare your results (including the result produced by the SON algorithm).

In our experiment , we found different sample sizes causes different degree with false positive and false negative .

Simple randomised algorithm in 1 % sample size , which generates a bunch of false positive inputs .Then we turn to simple randomised algorithm in 2 % sample size , which generates less false positive inputs . We found outputs instability in 1&2% sample sizes. However ,in 5% and 10% sample size ,we can get much more stable and accuracy outputs and the frequent itemsets when we run the code .

According to our experiment, we have more accurate and stable output with the size increasing. In comparison, the impact of efficiency is not obvious with different sample sizes.