CS52-Assignment-1 Report

Q1.

Usage Instruction:

1. Algorithms are in implemented in separate modules apriori.py, eclat.py, fpGrowth.py
2. To use them import apriori , eclat, fp methods from respective modules
3. The input format for all three functions is filename, support, categorical. Categorical variable needed to True if the categorical dataset is to be passed
4. The functions output frequent itemsets

Optimisation

Aproiri algorithm is implemented using hashing optimization

Dataset information

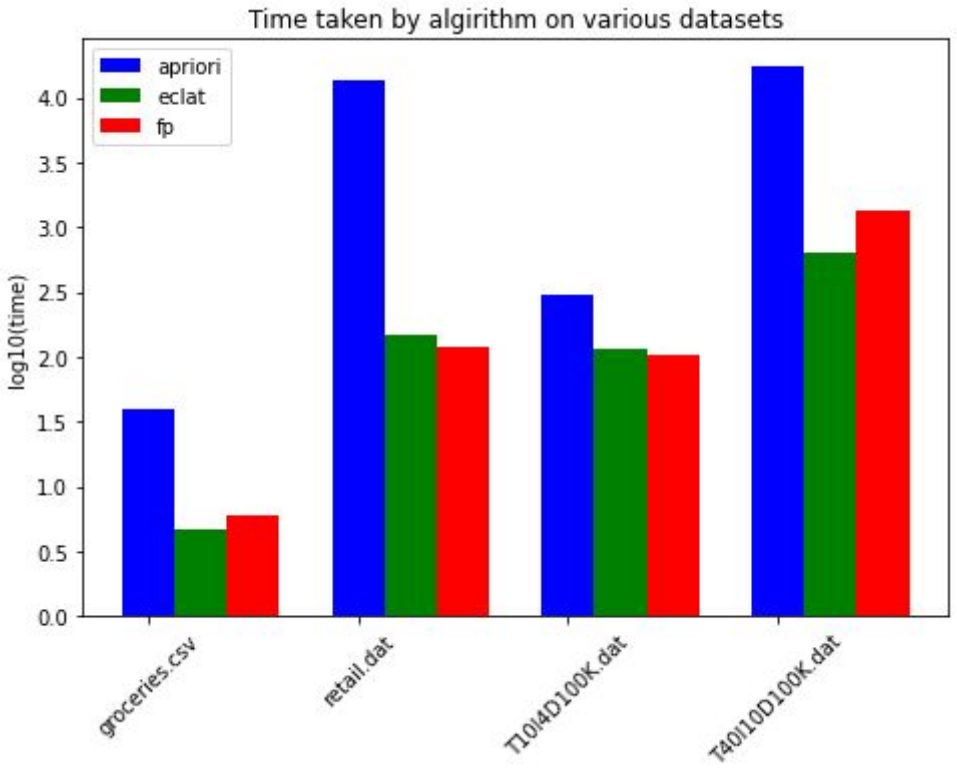
```
---------------------------
groceries.csv dataset info
no of transaction are  9835
no of items are  169
average width of transaction is  4.409456024402644
below values are obtained on support of 5% on eclat algo
maximum size of  frequent itemset is  2
size of maximum frequent itemsets is  3
```

```
---------------------------
retail.dat dataset info
no of transaction are  88162
no of items are  16470
average width of transaction is  10.305755314080896
below values are obtained on support of 5% on eclat algo
maximum size of  frequent itemset is  3
size of maximum frequent itemsets is  3
---------------------------
```

```
T10I4D100K.dat dataset info
no of transaction are  100000
no of items are  870
average width of transaction is  10.10228
below values are obtained on support of 5% on eclat algo
maximum size of  frequent itemset is  1
size of maximum frequent itemsets is  10
```

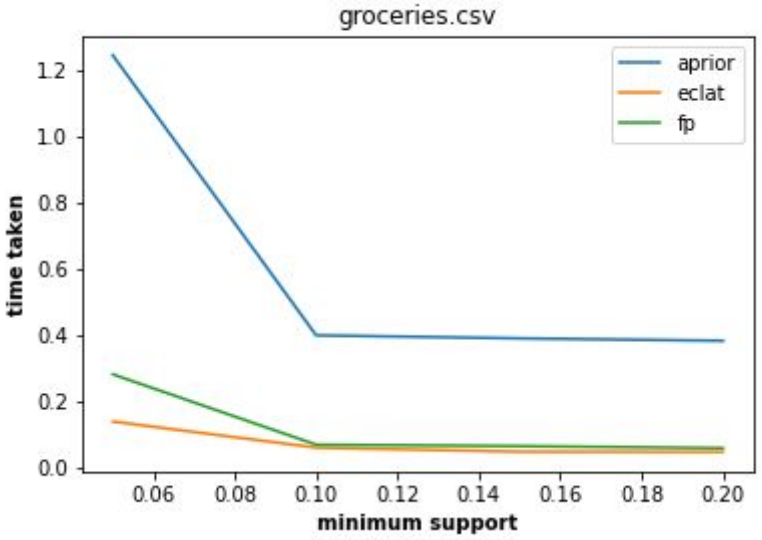
```
--------------------------
T40I10D100K.dat dataset info
no of transaction are  100000
no of items are  942
average width of transaction is  39.60507
below values are obtained on support of 5% on eclat algo
maximum size of  frequent itemset is  2
size of maximum frequent itemsets is  15
--------------------------
```

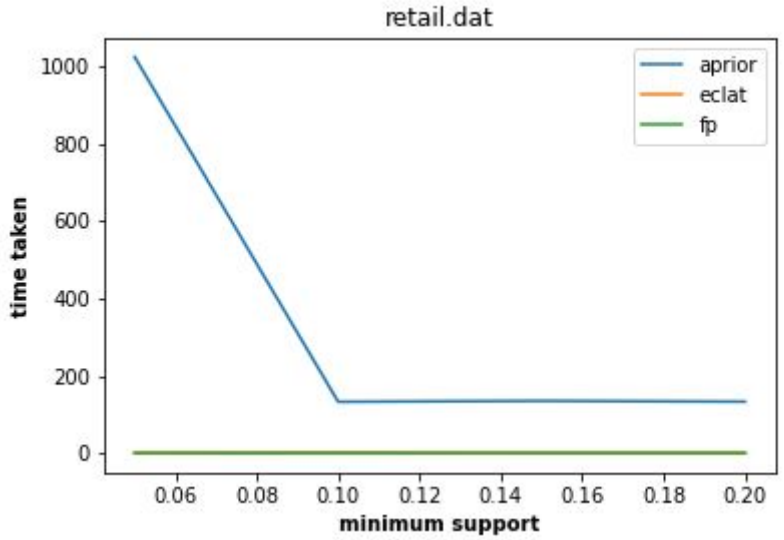
Algorithms Comparison:

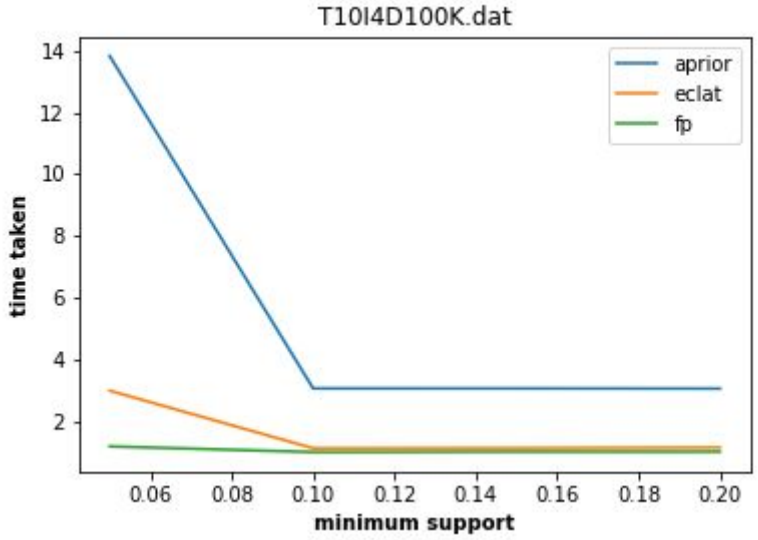


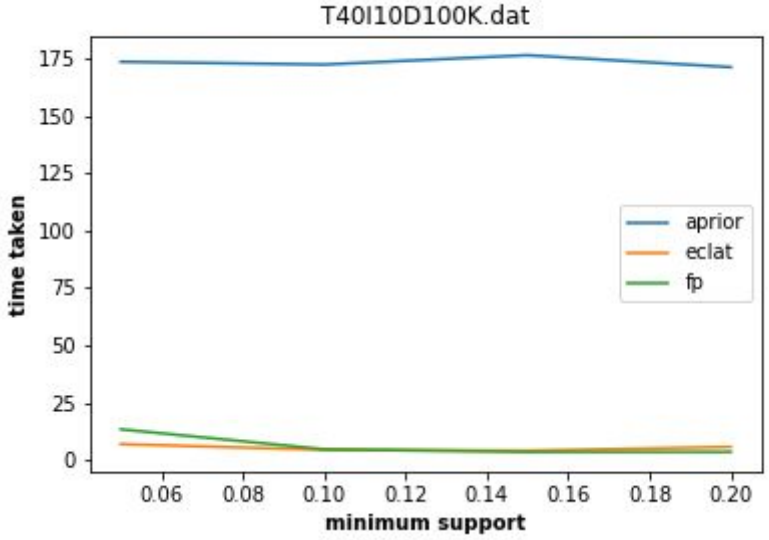
1) The graph takes time taken scales it 100 times and plots it log of times vs datasets as a bar graph

2) As show in the graph apriori is always slow
3) On the groceries dataset eclat is faster than fp because the dataset size very low and overhead for fp is large
4) On T40 dataset fp takes comparable time to aproiri as system I ran has very low and the dataset size is very large
5) Eclat is ideal algorithm because of fast running time and low memory
6) The following figures show time taken in seconds by three algorithm as function of min support on all four datasets. T401 dataset has support values of 0.15, 0.17, 0.18, 0.2 due to its very large run time on apriori and others have 0.05, 0.1, 0.15, 0.2 has min support





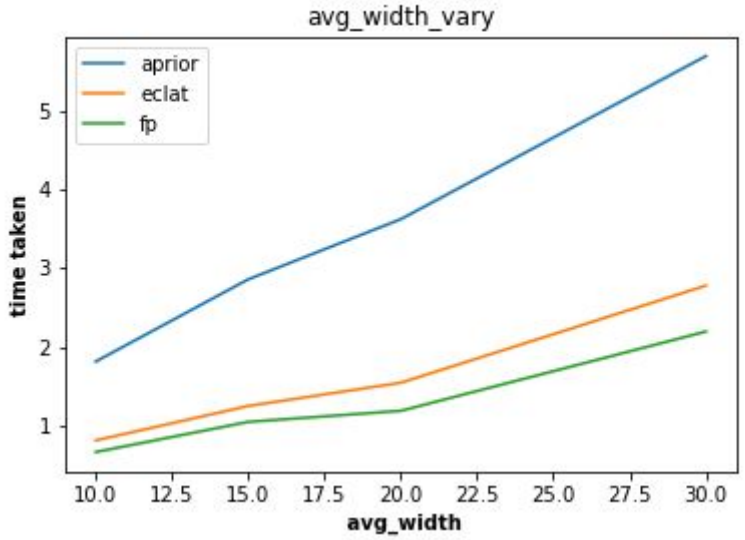


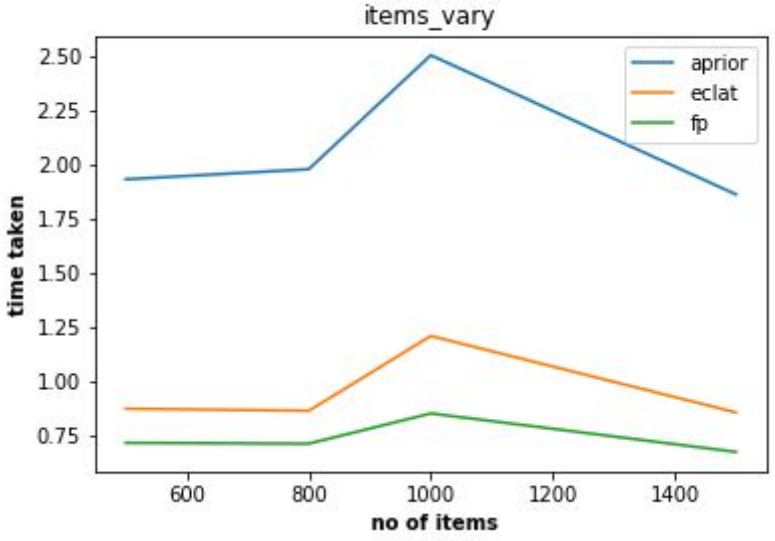


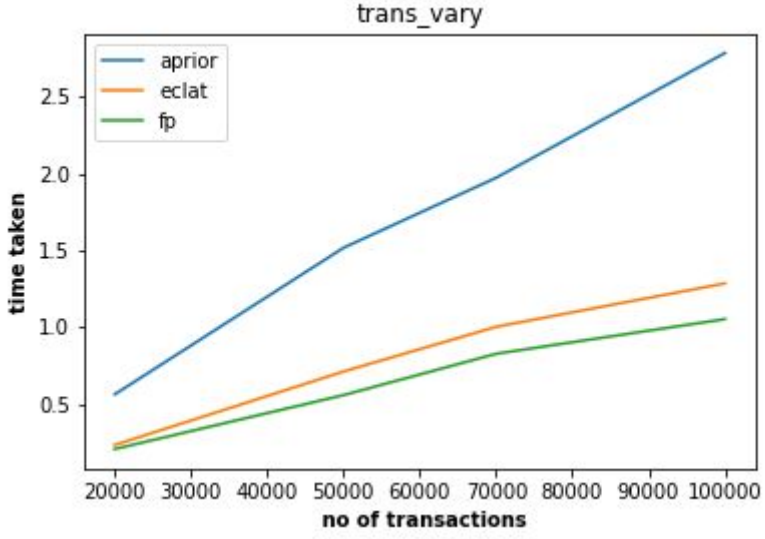
Q2:

1) Datasets are created using generate_dataset function datasets_generation module
2) Inputs are no of transaction, no of items, avg width of transaction

3) The following figures show time taken three algorithm as each parameter varies







4) As we can see no of items varying keeping transactions and average transactions constant almost have constant trend, this is due to total size of dataset remains same as no of items doesn’t affect it if avg width remains constant