COL761: Data Mining
Assignment 1

Team Members:
Supreeti Kumari - 2020CS10396
Eshan Jain - 2020CS50424
Goonjan Saha - 2020CS10494

Contributions:

All team members have contributed equally to this assignment.

Main Algorithm:

1. We first construct the FP Tree for the given dataset. To do this, we order the elements of the dataset based on their frequency of occurrence in the various transactions.
2. Next, we perform multiple iterations to do the data compression. For this, we first look through all the edges in the constructed FP-Tree. Considering each edge as an itemset of size 2, we calculate the frequency of its occurrence using the FP-Tree. We choose the max frequency itemset among these and create a map that merges it. We treat the newly merged itemset as a single element and repeat the iterations.
3. These iterations terminate when the frequency of the most frequent itemset chosen in the iteration is below a threshold that we have suitably chosen.
4. The final compressed dataset and mapping is formed by traversing this final FP-Tree.
5. For decompression, we are just reading through each transaction and referring to the compression map to decode the transactions.

Design Decisions:

1. Since the threshold that we set for the value of max frequency of itemset at which the algorithm terminates causes a tradeoff in time taken versus compression achieved (i.e. a lower threshold leads to a higher compression but takes more time), we have set different values of threshold depending on the size of transaction dataset.
2. We also perform an initial pruning in the FP-Tree before beginning our iterative algorithm by removing the nodes that will never be traversed in our algorithm by checking if the frequency of the edges connected to it are below the threshold.

3. We also do an initial run in the algorithm that looks through paths on the tree and merges the itemsets where each element has the same frequency in the branch and they occur consecutively on the tree branch.