Executive Summary

Problem - 2

This problem has been solved using Association Rule Mining. During preprocessing, the August to November sales files have been used to extract bills, i.e, the slip obtained on each transaction at the ANC counter. This bill, encapsulated within the class Transaction, is the entity on which most of the analysis has been performed.

The transactions (as described above) obtained from the sales file have been used to create the file augToNovTrans.csv which lists all transactions over the four months in csv format.

Association Rule Mining has been used (with the Apriori algorithm) to extract itemsets occurring together frequently. Larger itemsets have been given a preference till the point where support for these sets does not take a plunge. A tradeoff has been made in this respect to balance the number of items in the potential combos against the support count in favour of these combos.

The top 30 association rules were extracted from the above process such that each rule contains 3 or more elements. The support threshold which resulted in the top 30 rules ensured the elimination of rules with 5 or more items. These 30 rules have been manually examined keeping the following considerations in mind:

1. At least one item should have a low frequency or a low average rating.
2. The items in a group should make logical sense.

The top 8 rules which satisfy these criteria (with higher support used for tie-breaking) have been used to populate a combo list.

Similarly, the top 15 association rules which contained an element of low frequency were obtained through a similar procedure and the top 2 rules based on the same criteria were selected. This resulted in a total of 10 combos.

The transactions from the test month, i.e. December have been obtained in a similar fashion as described above. Each transaction of December is then checked against the combo list in the following order:

1. The combos formed due to low frequency of purchase one of the items are given preference over the ones formed due to low rating of one of the items.
2. Within these two groups, the combos which cause a lower loss in total revenue generated are applied first.

When tested on the December sales data, the application of these combos gives a total revenue loss of 0.59%