Data Mining Assignment Report

Ans 2:

Steps taken:

- Data from all the months was combined.
- Item attribute was set as nominal, set to flag
- Apriori was applied using IBM SPSS. The support and confidence threshold were taken to be 0 and 0.1% to extract the maximum number of rules possible. These rules were later refined manually.
- The consequent and antecedent were set so as to get both of them as items.
- The “rules_Q2.txt” file has all the frequent transactions stored.
- The “data_sets.py” file takes all the frequent itemsets input from “rules_Q2.txt”.
- Now, further from these frequent itemsets those itemsets have been taken whose at least one item has a rating below a minimum rate threshold. Let us call these new itemsets as collection C.
- In collection C, the itemsets are further sorted according to their total price and then no of items in an itemset.
- A group with a lower price is given higher priority as people will generally prefer to buy items that are cheap.
- Later, the priority has been decided by the no. of items in a group. A group that has more items is given more priority as it the discount for this combo will be less, and so will be the loss incurred.
- From this pool of rules, the rules that make actually become a combo are manually chosen. This is because many rules that contain similar items like plain maggi , paneer maggi have to removed, which can only be done manually.

- These final combo's are used in "profit_calcu.py" file as a list data.
- Finally, the total loss is calculated using "profit_calcu.py".
- The following results have been obtained:
    - Total number of combos offered: **176**
    - Total loss: **1.54%**
- The "final.py" file has been used to generate the csv file "combos.csv" in the format as mentioned in the question.
- The model can be evaluated considering the ratio of the number of combos offered v/s the total loss incurred. The higher this ratio is, the better is the performance of our solution.