Compsci 361 Assignment 3

Hasnain Cheena  
190411106  
hche737

## Part A & B

The dataset contains information related to grocery shopping. It contains 124 unique items and 4627 unequal length transactions. The longest length transaction in the dataset is 49 items.

Supermarket transactions

The tool used in this report is the *mlxtend* library in Python (Python 3.6.9). The algorithm used within the library is *fpgrowth.*Preprocessing of the dataset (avaliable in a .csv format) was done prior to analysis. This preprocessing chain involved:  
1. Using pandas to read in the csv data (specifying the number of columns to be equal to the longest length transaction).   
2. Removing all NaN values and converting the structure to a list of lists in which each transaction is a list stored within a bigger list which contains all the transactions.  
3. Transforming the dataset using a Transactional Encoder to get it into the correct format for association rule mining.

## Part C

The final parameters selected were *minsup* of 0.2, *minconf* of 0.7 and *minlevel* of 1.3.

Lift higher than 1 was the min lift. Tried every value of lift between 1.5 to 2.5   
Table of number

Threshold of 2

The minimum support parameter was selected first. It was selected systematically by iterating through a range of support levels of 0.05 to 0.7 at intervals of 0.05. 0.05 was choosen as a lower limit as the memory required and compute resources required at support levels lower were infeasible. 0.7 was choosen as the upper end as the highest support in the dataset is around 0.72 (support of item ‘bread & cake’). 0.2 was selected as the final minsup parameter as to generate useful and interesting rules on this dataset 0.2 seemed reasonable.

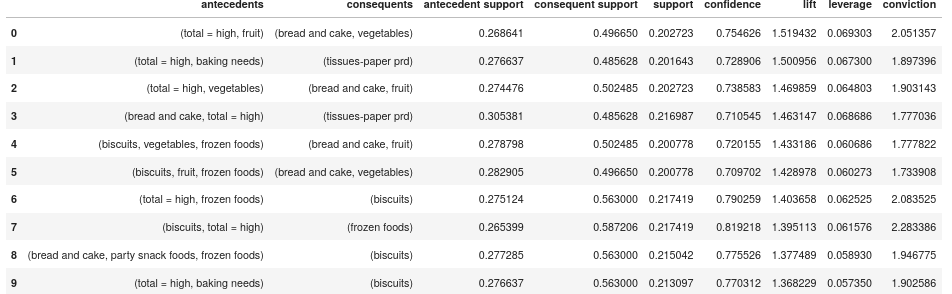
Then the minimum confidence parameter was set by deciding what level of confidence was accepted in the final rules. Higher confidence rules are desired and thus confidence levels were evaluated from 0.7 to 1 at intervals of 0.1. Confidence levels 0.9 and 1 produced no rules. It was difficult to pick between 0.7 and 0.8. 0.7 was selected as it produced rules with higher lift values than 0.8.

Further pruning involved checking whether confidence of the rule is greater than consequent support to avoid misleading rules. Additionally, to only rules with lift greater than 1.3 were selected.

To calculate timing characteristics of the *fpgrowth* algorithm multiple runs of the algorithm were performed using the *timeit* functionality in Python. The results show that the algorithm on average at 600 ms with a standard deviation of 182 ms.

## Part D

Interestingness measures use to select rules were confidence and lift. Rules with a sufficiently high confidence of 0.7 and lift above 1.3 were selected. The number of rules after filtering is 44.



The rules selected are above.

Need to discuss and describe the rules.

Why didn’t use confidence – confidence has problem

Negatively correlated items set’s cause problems

Prob of