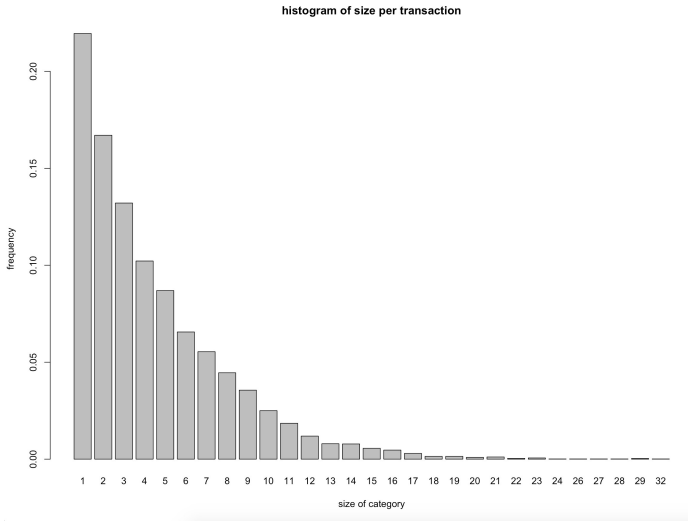
CMPT 459 Programming Assignment 4

Shawn An(301323174)

1.Plot a histogram of the number of items (categories) per transaction. What do you observe? How can you explain this observation?



According to the histogram, most transactions contain fewer categories. 21% of the transactions contains items in only 1 category and 16% contains items in 2 categories. When the size becomes larger, the frequency rate of such transaction becomes less. The observation matches the real-life scenario because people tend to purchase things in need at the moment most of the time. Thus, they would be more likely to visit the store and grab fewer items they need in a single transaction. A single transaction with more than 20 categories is rare and an example of the scenario would be buying necessary stuff after a house moving. Such scenario would not happen often in the real world either.

2.How many frequent/closed frequent/maximal frequent itemsets with min support = 0.001?

The number of frequent itemsets is 13492, the number of closed frequent itemsets is 13464 and the number of maximal frequent itemsets is 7794 with minimum support = 0.001.

3.How many frequent/closed frequent/maximal frequent itemsets with min support = 0.01?

The number of frequent itemsets is 333, the number of closed frequent itemsets is 333 and the number of maximal frequent itemsets is 243 with minimum support = 0.01.

4.What are the 10 itemsets with the highest support, and what is their support?

	itemset	support	count
1	{whole milk}	0.256	2513
2	{other vegetables}	0.193	1903
3	{rolls/buns}	0.184	1809
4	{soda}	0.174	1715
5	{yogurt}	0.140	1372
6	{bottled water}	0.111	1087
7	{root vegetables}	0.109	1072
8	{tropical fruit}	0.105	1032
9	{shopping bags}	0.099	969
10	{sausage}	0.094	924

5.How do you explain the relatively small number of frequent itemsets for the already low minimum support of 0.01? How do you explain the observation that the numbers of frequent itemsets, closed frequent itemsets, and maximal frequent itemsets are so similar?

The minimum support of 0.01 means the fraction of the itemsets occurs in the dataset has to be greater or equal to 0.01 to be considered. Since the dataset contains 9835 transactions, the total occurrence for the candidates with single category has to be greater or equal to 99(9835*0.01). The number of single-category candidates is only 88. Due to the anti-monotonicity property, the multi-category itemsets would be the candidate only if all its subset are frequent itemsets. By inspections, the number of 2-category itemsets is 213 and 3-category itemset is 32. Also, since there are a total of 169 items. The probability of repetitively showing the same item in the transaction would be low.

As we know the relationship of frequent itemsets(fset), closed frequent itemsets(cfset), and maximal frequent itemsets(mfset) is: mfset is a subset of fset, cfset is a subset of fset. In the dataset, the same number of cfset and fset indicates that all the itemsets with fewer categories have higher support than their supersets. Therefore, no itemset is represented by its superset in cfset. Since maximal frequent itemset is lossy compression of the frequent itemsets, it would not keep the support of the itemsets. A frequent itemset X is maximal if none of its supersets is frequent. The difference between the number of fset and mfset indicates that some itemsets have been combined with its superset.

6.At minimum support = 0.01, how many association rules do you obtain with minimum confidence = 0.9? How far do you need to lower the minimum confidence to obtain more than 10 rules?

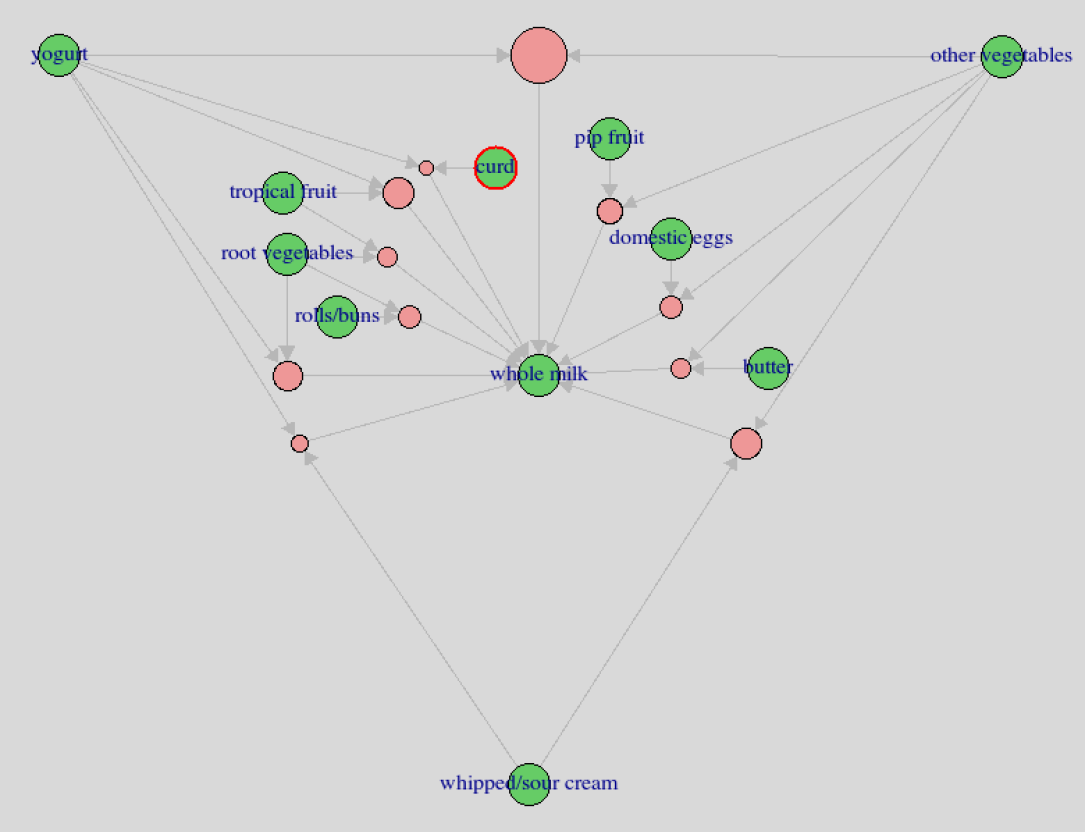
At minimum support 0.01, there is 0 association rules with minimum confidence of 0.9. I have to lower the confidence to 0.51 to get 12 rules. The rules are shown below.

```
> inspect(rules)
     lhs                                rhs                support confidence lift count
[1]  {curd,yogurt}                   => {whole milk}       0.010   0.58       2.3   99
[2]  {other vegetables,butter}       => {whole milk}       0.011   0.57       2.2  113
[3]  {other vegetables,domestic eggs} => {whole milk}      0.012   0.55       2.2  121
[4]  {yogurt,whipped/sour cream}     => {whole milk}       0.011   0.52       2.1  107
[5]  {pip fruit,other vegetables}    => {whole milk}       0.014   0.52       2.0  133
[6]  {citrus fruit,root vegetables}  => {other vegetables} 0.010   0.59       3.0  102
[7]  {tropical fruit,root vegetables} => {other vegetables} 0.012   0.58       3.0  121
[8]  {tropical fruit,root vegetables} => {whole milk}      0.012   0.57       2.2  118
[9]  {tropical fruit,yogurt}         => {whole milk}       0.015   0.52       2.0  149
[10] {root vegetables,yogurt}        => {whole milk}       0.015   0.56       2.2  143
[11] {root vegetables,rolls/buns}    => {whole milk}       0.013   0.52       2.0  125
[12] {other vegetables,yogurt}       => {whole milk}       0.022   0.51       2.0  219
```

7.Plot the rules that have "whole milk" in rhs for min support 0.01 and min confidence 0.5.

```
> inspect(rules2)
     lhs                                rhs          support confidence lift count
[1]  {curd,yogurt}                   => {whole milk} 0.010   0.58       2.3   99
[2]  {other vegetables,butter}       => {whole milk} 0.011   0.57       2.2  113
[3]  {tropical fruit,root vegetables} => {whole milk} 0.012   0.57       2.2  118
[4]  {root vegetables,yogurt}        => {whole milk} 0.015   0.56       2.2  143
[5]  {other vegetables,domestic eggs} => {whole milk} 0.012   0.55       2.2  121
[6]  {yogurt,whipped/sour cream}     => {whole milk} 0.011   0.52       2.1  107
[7]  {root vegetables,rolls/buns}    => {whole milk} 0.013   0.52       2.0  125
[8]  {pip fruit,other vegetables}    => {whole milk} 0.014   0.52       2.0  133
[9]  {tropical fruit,yogurt}         => {whole milk} 0.015   0.52       2.0  149
[10] {other vegetables,yogurt}       => {whole milk} 0.022   0.51       2.0  219
[11] {other vegetables,whipped/sour cream} => {whole milk} 0.015   0.51       2.0  144
```

The plot is shown below.



8.In task 7, which ones have the highest lift? Can you explain these rules? How interesting are they?

The one with highest lift is {curd,yogurt}=>{whole milk} with lift = 2.3

This rule is not particularly interesting even the lift is much greater than 1. It is because that curd yogurt and whole milk are all milk product and intuitively if one likes to eat curd and yogurt, he or she would also like to drink milk. Also, those items are closely located to each other normally.

```
[2]  {other vegetables,butter}                 => {whole milk}
[3]  {tropical fruit,root vegetables}          => {whole milk}
[4]  {root vegetables,yogurt}                  => {whole milk}
[5]  {other vegetables,domestic eggs}          => {whole milk}
[7]  {root vegetables,rolls/buns}              => {whole milk}
[8]  {pip fruit,other vegetables}              => {whole milk}
[9]  {tropical fruit,yogurt}                   => {whole milk}
[10] {other vegetables,yogurt}                 => {whole milk}
[11] {other vegetables,whipped/sour cream} => {whole milk}
```

These rules mean people who buy food from the left side tend to buy whole milk. It indicates that the customers intend to shop for food. People who buy fruits, vegetables and milk may tend to make a milkshake at home. People who buy buns, vegetables, and milk may tend to cook for breakfast. People who buy vegetable, yogurt and milk may be on their diet. Therefore, the store may want to put the fresh food and dairy section closer to promote sales.

```
[6]  {yogurt,whipped/sour cream}          => {whole milk}
```

The rule may be less interesting for the same reason as the first rule that both sides are from the super category of dairy product. If people like yogurt and cream, they would have a higher change to like milk. Since the items are often placed in the same area, the store may already know the knowledge.