Apriori Findings

I was given a dataset (*Market_Basket_Optimisation.csv*) that is comprised of a list of shoppers from a grocery store and the products that they purchased. My goal is to find trends and habits in purchase patterns so that the products placement can be optimized in order to increase sales.

First, I will turn the dataset into a sparse matrix. One column for each of the products (120 in total).

```
> ds = read.transactions('Market_Basket_Optimisation.csv', sep = ',', rm.duplicates = TRUE)
distribution of transactions with duplicates:
1
5
```

I found that there were 5 duplicates in the dataset (ie. the products being bought twice in a single transaction). I removed these as instructed.

I printed a quick summary of the dataset and received the following.

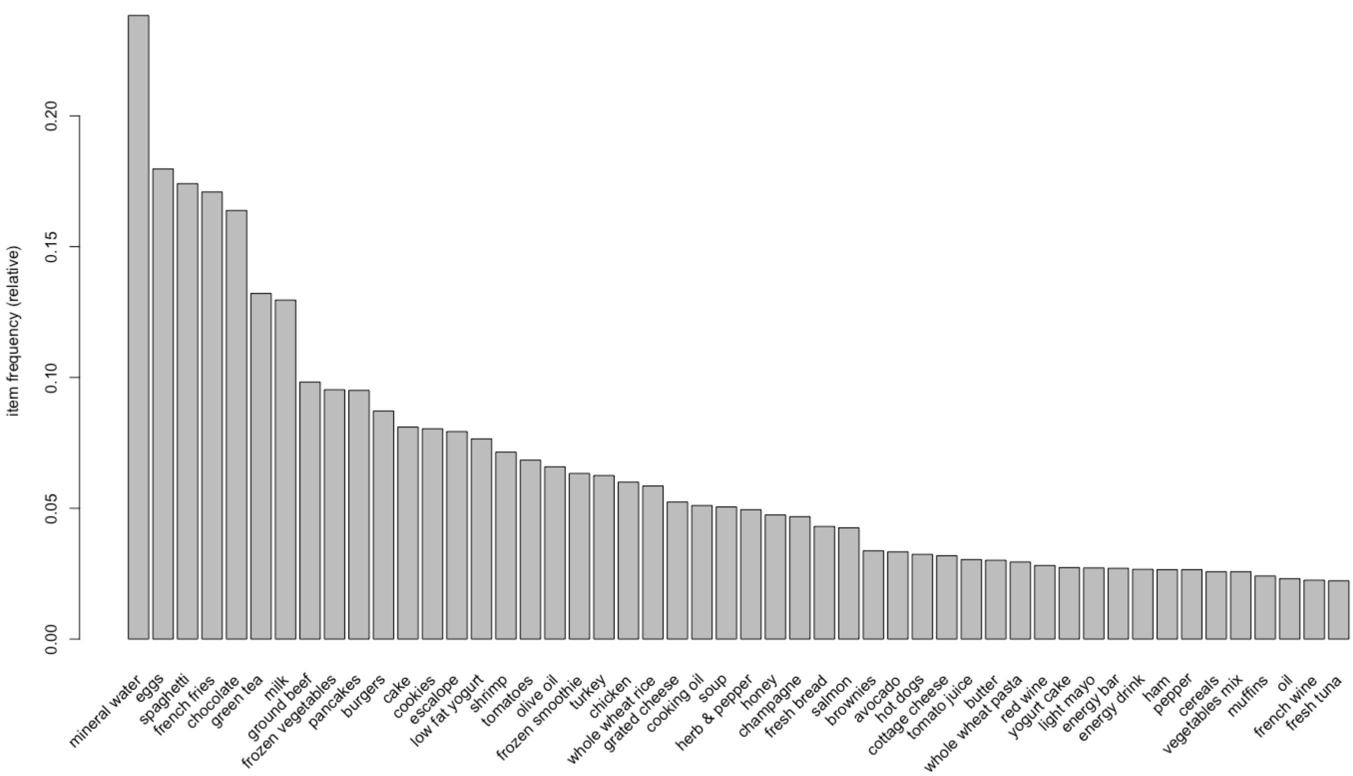
```
> summary(ds)
transactions as itemMatrix in sparse format with
 7501 rows (elements/itemsets/transactions) and
 119 columns (items) and a density of 0.03288973

most frequent items:
mineral water          eggs     spaghetti   french fries     chocolate       (Other)
         1788          1348          1306           1282          1229         22405

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   18   19   20
1754 1358 1044  816  667  493  391  324  259  139  102   67   40   22   17    4    1    2    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   3.914   5.000  20.000

includes extended item information - examples:
             labels
1           almonds
2 antioxydant juice
3         asparagus
```

From this we can observe

- 3% of the values are non-zero.
- Mineral water is the most bought item.

- 1754 people bought a single item, 1358 people bought two items, and so on.
- On average people purchased 3.9 products.

Below is a graph showing the top 50 items purchased



Next I ran the apriori algorithm in order to determine the rules (which products are often bought together). I started off with a *support* of 0.003 (products that are purchased at least 3 times a day across any given week)[1] and *confidence* of 0.8. This resulted in zero rules so I dropped the *confidence* by half to 0.4. This resulted in 281 rules, of which the top 10 are the following.

[1] This is determined by the following equation. *Daily purchase minimum*numbers of days in dataset / total number of rows in dataset.* Therefore, this equation is 3*7/7500 = 0.0028. For a minimum of a daily purchase of 4 the equations would be 4*7/7500 = 0.003733333333.

```
> inspect(sort(rules, by = 'lift')[1:10])
     lhs                                            rhs                 support     confidence lift     count
[1]  {mineral water,whole wheat pasta}           => {olive oil}         0.003866151 0.4027778  6.115863 29
[2]  {spaghetti,tomato sauce}                    => {ground beef}       0.003066258 0.4893617  4.980600 23
[3]  {french fries,herb & pepper}                => {ground beef}       0.003199573 0.4615385  4.697422 24
[4]  {cereals,spaghetti}                         => {ground beef}       0.003066258 0.4600000  4.681764 23
[5]  {frozen vegetables,mineral water,soup}      => {milk}              0.003066258 0.6052632  4.670863 23
[6]  {chocolate,herb & pepper}                   => {ground beef}       0.003999467 0.4411765  4.490183 30
[7]  {chocolate,mineral water,shrimp}            => {frozen vegetables} 0.003199573 0.4210526  4.417225 24
[8]  {frozen vegetables,mineral water,olive oil} => {milk}              0.003332889 0.5102041  3.937285 25
[9]  {cereals,ground beef}                       => {spaghetti}         0.003066258 0.6764706  3.885303 23
[10] {frozen vegetables,soup}                    => {milk}              0.003999467 0.5000000  3.858539 30
```

Here I have run into another problem. For example, rules 6 and 7. You will notice that both rules indicate that chocolate was a proponent in the purchase of ground beef and frozen vegetables. This could be the case, or it could be the case that because chocolate is a regularly purchased product it therefore happens to end up in a lot of baskets. Out of the two the more logical analysis is that chocolate is a frequently purchased item.

To refine the algorithm we can lower the confidence again. We set the confidence to 20% and receive 1348 rules of which the top 20 are.

```
> inspect(sort(rules, by = 'lift')[1:20])
     lhs                                             rhs                 support     confidence lift     count
[1]  {mineral water,whole wheat pasta}            => {olive oil}         0.003866151 0.4027778  6.115863 29
[2]  {frozen vegetables,milk,mineral water}       => {soup}              0.003066258 0.2771084  5.484407 23
[3]  {fromage blanc}                              => {honey}             0.003332889 0.2450980  5.164271 25
[4]  {spaghetti,tomato sauce}                     => {ground beef}       0.003066258 0.4893617  4.980600 23
[5]  {light cream}                                => {chicken}           0.004532729 0.2905983  4.843951 34
[6]  {pasta}                                      => {escalope}          0.005865885 0.3728814  4.700812 44
[7]  {french fries,herb & pepper}                 => {ground beef}       0.003199573 0.4615385  4.697422 24
[8]  {cereals,spaghetti}                          => {ground beef}       0.003066258 0.4600000  4.681764 23
[9]  {frozen vegetables,mineral water,soup}       => {milk}              0.003066258 0.6052632  4.670863 23
[10] {french fries,ground beef}                   => {herb & pepper}     0.003199573 0.2307692  4.665768 24
[11] {chocolate,frozen vegetables,mineral water}  => {shrimp}            0.003199573 0.3287671  4.600900 24
[12] {frozen vegetables,milk,mineral water}       => {olive oil}         0.003332889 0.3012048  4.573557 25
[13] {pasta}                                      => {shrimp}            0.005065991 0.3220339  4.506672 38
[14] {chocolate,herb & pepper}                    => {ground beef}       0.003999467 0.4411765  4.490183 30
[15] {chocolate,mineral water,shrimp}             => {frozen vegetables} 0.003199573 0.4210526  4.417225 24
[16] {cake,frozen vegetables}                     => {tomatoes}          0.003066258 0.2987013  4.367560 23
[17] {milk,tomatoes}                              => {soup}              0.003066258 0.2190476  4.335293 23
[18] {eggs,ground beef}                           => {herb & pepper}     0.004132782 0.2066667  4.178455 31
[19] {milk,olive oil}                             => {soup}              0.003599520 0.2109375  4.174781 27
[20] {whole wheat pasta}                          => {olive oil}         0.007998933 0.2714932  4.122410 60
```

If we raise the support to a minimum purchase of four times a day (support of 0.004) we receive 811 rules the top 20 of which are as follows.

```
> inspect(sort(rules, by = 'lift')[1:20])
     lhs                                        rhs                 support     confidence lift     count
[1]  {light cream}                           => {chicken}           0.004532729 0.2905983  4.843951  34
[2]  {pasta}                                 => {escalope}          0.005865885 0.3728814  4.700812  44
[3]  {pasta}                                 => {shrimp}            0.005065991 0.3220339  4.506672  38
[4]  {eggs,ground beef}                      => {herb & pepper}     0.004132782 0.2066667  4.178455  31
[5]  {whole wheat pasta}                     => {olive oil}         0.007998933 0.2714932  4.122410  60
[6]  {herb & pepper,spaghetti}               => {ground beef}       0.006399147 0.3934426  4.004360  48
[7]  {herb & pepper,mineral water}           => {ground beef}       0.006665778 0.3906250  3.975683  50
[8]  {tomato sauce}                          => {ground beef}       0.005332622 0.3773585  3.840659  40
[9]  {mushroom cream sauce}                  => {escalope}          0.005732569 0.3006993  3.790833  43
[10] {frozen vegetables,mineral water,spaghetti} => {ground beef}   0.004399413 0.3666667  3.731841  33
[11] {olive oil,tomatoes}                    => {spaghetti}         0.004399413 0.6111111  3.509912  33
[12] {frozen vegetables,spaghetti}           => {tomatoes}          0.006665778 0.2392344  3.498046  50
[13] {mineral water,soup}                    => {olive oil}         0.005199307 0.2254335  3.423030  39
[14] {ground beef,milk}                      => {olive oil}         0.004932676 0.2242424  3.404944  37
[15] {eggs,herb & pepper}                    => {ground beef}       0.004132782 0.3297872  3.356491  31
[16] {spaghetti,tomatoes}                    => {frozen vegetables} 0.006665778 0.3184713  3.341054  50
[17] {herb & pepper}                         => {ground beef}       0.015997867 0.3234501  3.291994 120
[18] {grated cheese,spaghetti}               => {ground beef}       0.005332622 0.3225806  3.283144  40
[19] {cooking oil,ground beef}               => {spaghetti}         0.004799360 0.5714286  3.281995  36
[20] {frozen vegetables,olive oil}           => {milk}              0.004799360 0.4235294  3.268410  36
```

From this data the store can then rearrange their products as they see fit to increase their overall sales.