

Rule Mining

Allie Touchstone

8/16/2021

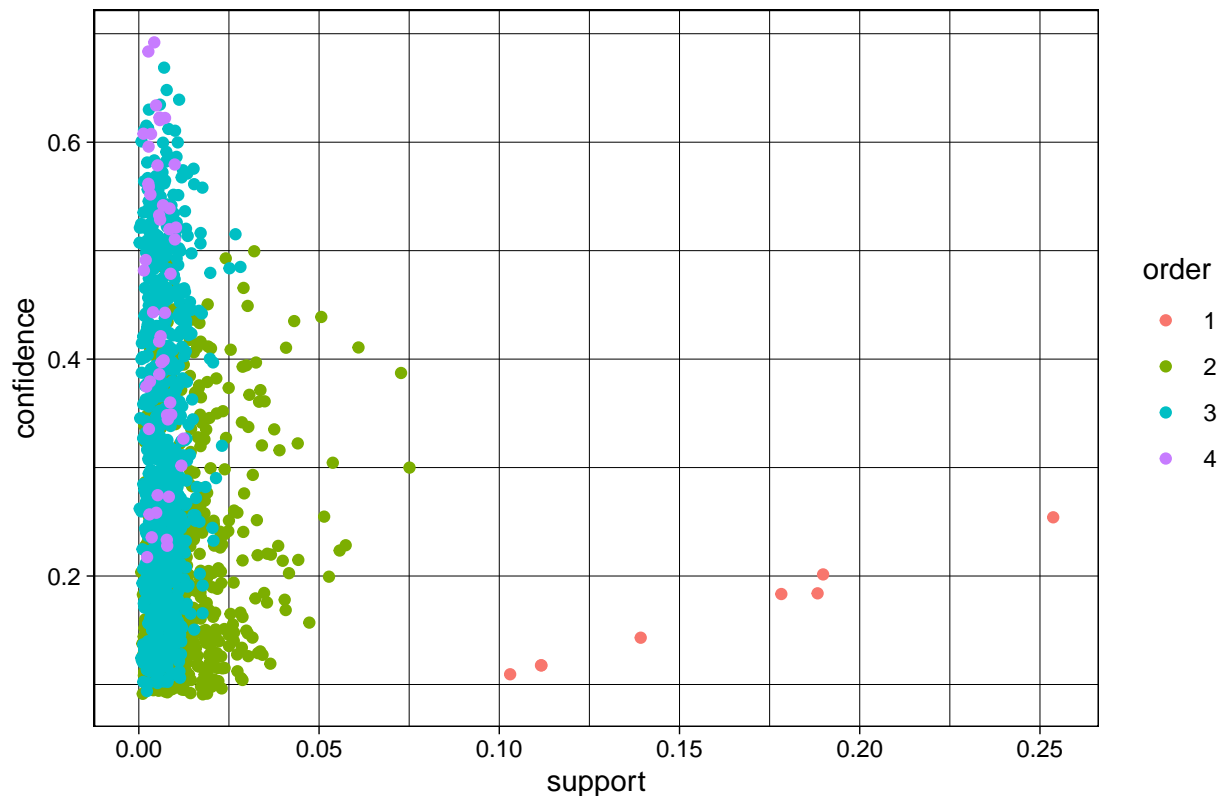
First we separated the data out into a format that the code could easily read. Then we put parameters on the data and are only looking at the data that has at more than a confidence of 0.005 and a support of 0.1.

The pertinent information from the summary of this subset from the groceries.txt file is that in all there are 1582 rules.

This is a plots of the rules where the groceries data is grouped off into 4 orders.

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```

Scatter plot for 1582 rules



These are the rules where the support is larger than 0.05.

```
##      lhs      rhs      support  confidence coverage
## [1] {}      => {bottled water}  0.11052364 0.1105236  1.0000000
```

```

## [2] {} => {tropical fruit} 0.10493137 0.1049314 1.0000000
## [3] {} => {root vegetables} 0.10899847 0.1089985 1.0000000
## [4] {} => {soda} 0.17437722 0.1743772 1.0000000
## [5] {} => {yogurt} 0.13950178 0.1395018 1.0000000
## [6] {} => {rolls/buns} 0.18393493 0.1839349 1.0000000
## [7] {} => {other vegetables} 0.19349263 0.1934926 1.0000000
## [8] {} => {whole milk} 0.25551601 0.2555160 1.0000000
## [9] {yogurt} => {whole milk} 0.05602440 0.4016035 0.1395018
## [10] {whole milk} => {yogurt} 0.05602440 0.2192598 0.2555160
## [11] {rolls/buns} => {whole milk} 0.05663447 0.3079049 0.1839349
## [12] {whole milk} => {rolls/buns} 0.05663447 0.2216474 0.2555160
## [13] {other vegetables} => {whole milk} 0.07483477 0.3867578 0.1934926
## [14] {whole milk} => {other vegetables} 0.07483477 0.2928770 0.2555160
## lift count
## [1] 1.000000 1087
## [2] 1.000000 1032
## [3] 1.000000 1072
## [4] 1.000000 1715
## [5] 1.000000 1372
## [6] 1.000000 1809
## [7] 1.000000 1903
## [8] 1.000000 2513
## [9] 1.571735 551
## [10] 1.571735 551
## [11] 1.205032 557
## [12] 1.205032 557
## [13] 1.513634 736
## [14] 1.513634 736

```

You can see from these rules which items are the top 8 grocery items bought on there own. Each item bought at least 1000 times a piece. (There are 9835 entrys in the data).

These are the rules where the confidence is larger than 0.6. While it might look overwhelming due to formatting issues, the reason this is being included is to point out the rhs column.

##	lhs	rhs	support	confidence	coverage	lift	count
## [1]	{onions,	=> {other vegetables}	0.005693950	0.6021505	0.009456024	3.112008	56
## [2]	root vegetables}						
## [2]	{curd,	=> {whole milk}	0.006507372	0.6336634	0.010269446	2.479936	64
## [3]	tropical fruit}						
## [3]	{domestic eggs,	=> {whole milk}	0.005185562	0.6219512	0.008337570	2.434099	51
## [4]	margarine}						
## [4]	{butter,	=> {whole milk}	0.005998983	0.6210526	0.009659380	2.430582	59
## [5]	domestic eggs}						
## [5]	{butter,	=> {whole milk}	0.006710727	0.6600000	0.010167768	2.583008	66
## [6]	whipped/sour cream}						
## [6]	{bottled water,	=> {whole milk}	0.005388917	0.6022727	0.008947636	2.357084	53
## [7]	butter}						
## [7]	{butter,	=> {whole milk}	0.006202339	0.6224490	0.009964413	2.436047	61
## [8]	tropical fruit}						
## [8]	{butter,	=> {whole milk}	0.008235892	0.6377953	0.012913066	2.496107	81
## [9]	root vegetables}						
## [9]	{butter,	=> {whole milk}	0.009354347	0.6388889	0.014641586	2.500387	92
##	yogurt}						

```

## [10] {domestic eggs,
##       pip fruit}      => {whole milk}      0.005388917  0.6235294 0.008642603 2.440275 53
## [11] {domestic eggs,
##       tropical fruit} => {whole milk}      0.006914082  0.6071429 0.011387900 2.376144 68
## [12] {pip fruit,
##       whipped/sour cream} => {other vegetables} 0.005592272  0.6043956 0.009252669 3.123610 55
## [13] {pip fruit,
##       whipped/sour cream} => {whole milk}      0.005998983  0.6483516 0.009252669 2.537421 59
## [14] {fruit/vegetable juice,
##       other vegetables,
##       yogurt}          => {whole milk}      0.005083884  0.6172840 0.008235892 2.415833 50
## [15] {other vegetables,
##       root vegetables,
##       whipped/sour cream} => {whole milk}      0.005185562  0.6071429 0.008540925 2.376144 51
## [16] {other vegetables,
##       pip fruit,
##       root vegetables}  => {whole milk}      0.005490595  0.6750000 0.008134215 2.641713 54
## [17] {pip fruit,
##       root vegetables,
##       whole milk}       => {other vegetables} 0.005490595  0.6136364 0.008947636 3.171368 54
## [18] {other vegetables,
##       pip fruit,
##       yogurt}          => {whole milk}      0.005083884  0.6250000 0.008134215 2.446031 50
## [19] {citrus fruit,
##       root vegetables,
##       whole milk}       => {other vegetables} 0.005795628  0.6333333 0.009150991 3.273165 57
## [20] {root vegetables,
##       tropical fruit,
##       yogurt}          => {whole milk}      0.005693950  0.7000000 0.008134215 2.739554 56
## [21] {other vegetables,
##       tropical fruit,
##       yogurt}          => {whole milk}      0.007625826  0.6198347 0.012302999 2.425816 75
## [22] {other vegetables,
##       root vegetables,
##       yogurt}          => {whole milk}      0.007829181  0.6062992 0.012913066 2.372842 77

```

These rules show how whole milk is bought with just about everything, as well other vegetables are commonly bought with a wide variety of other items.

This first plot is considering all rules where the confidence and the support are greater than 0.03.

```

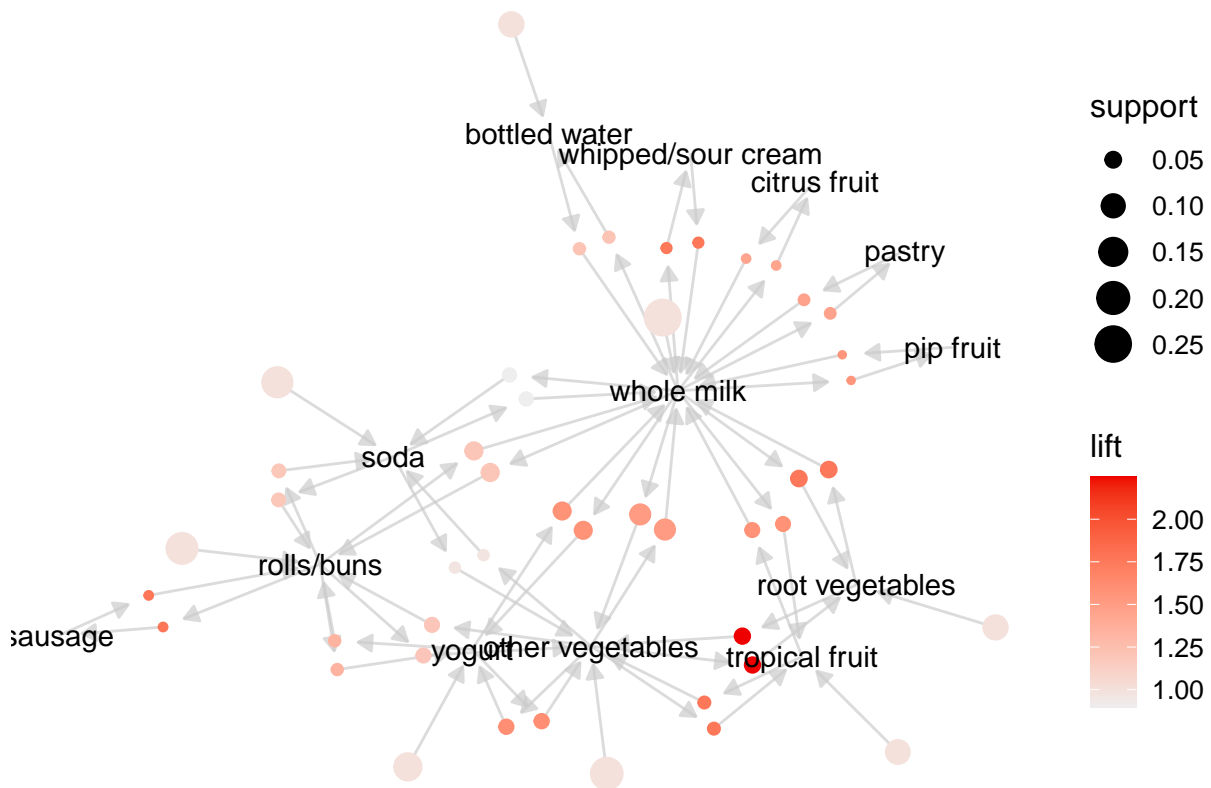
## set of 46 rules
##
## rule length distribution (lhs + rhs):sizes
## 1 2
## 8 38
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000  2.000   2.000   1.826   2.000   2.000
##
## summary of quality measures:
##      support      confidence      coverage      lift
##    Min.    :0.03010  Min.    :0.1049  Min.    :0.07168  Min.    :0.8991
##    1st Qu.:0.03353  1st Qu.:0.1671  1st Qu.:0.13950  1st Qu.:1.0488

```

```

## Median :0.04230   Median :0.2200   Median :0.19349   Median :1.4424
## Mean   :0.06175   Mean   :0.2420   Mean   :0.32140   Mean   :1.3938
## 3rd Qu.:0.05648   3rd Qu.:0.3112   3rd Qu.:0.25552   3rd Qu.:1.6007
## Max.   :0.25552   Max.   :0.4496   Max.   :1.00000   Max.   :2.2466
##      count
## Min.    : 296.0
## 1st Qu.: 329.8
## Median  : 416.0
## Mean    : 607.3
## 3rd Qu.: 555.5
## Max.    :2513.0
##
## mining info:
##      data ntransactions support confidence
## groceries      9835    0.005      0.1

```



Following it up we have this graph used in gephi to break it up into 7 orders (each shown as a different color) and how the grocery items are connected.

This second group of plots is when there is a more strict set of rules on the data. Here the confidence has to be larger than 0.3 and the support has to be larger than 0.03.

```

## set of 14 rules
##
## rule length distribution (lhs + rhs):sizes
## 2

```

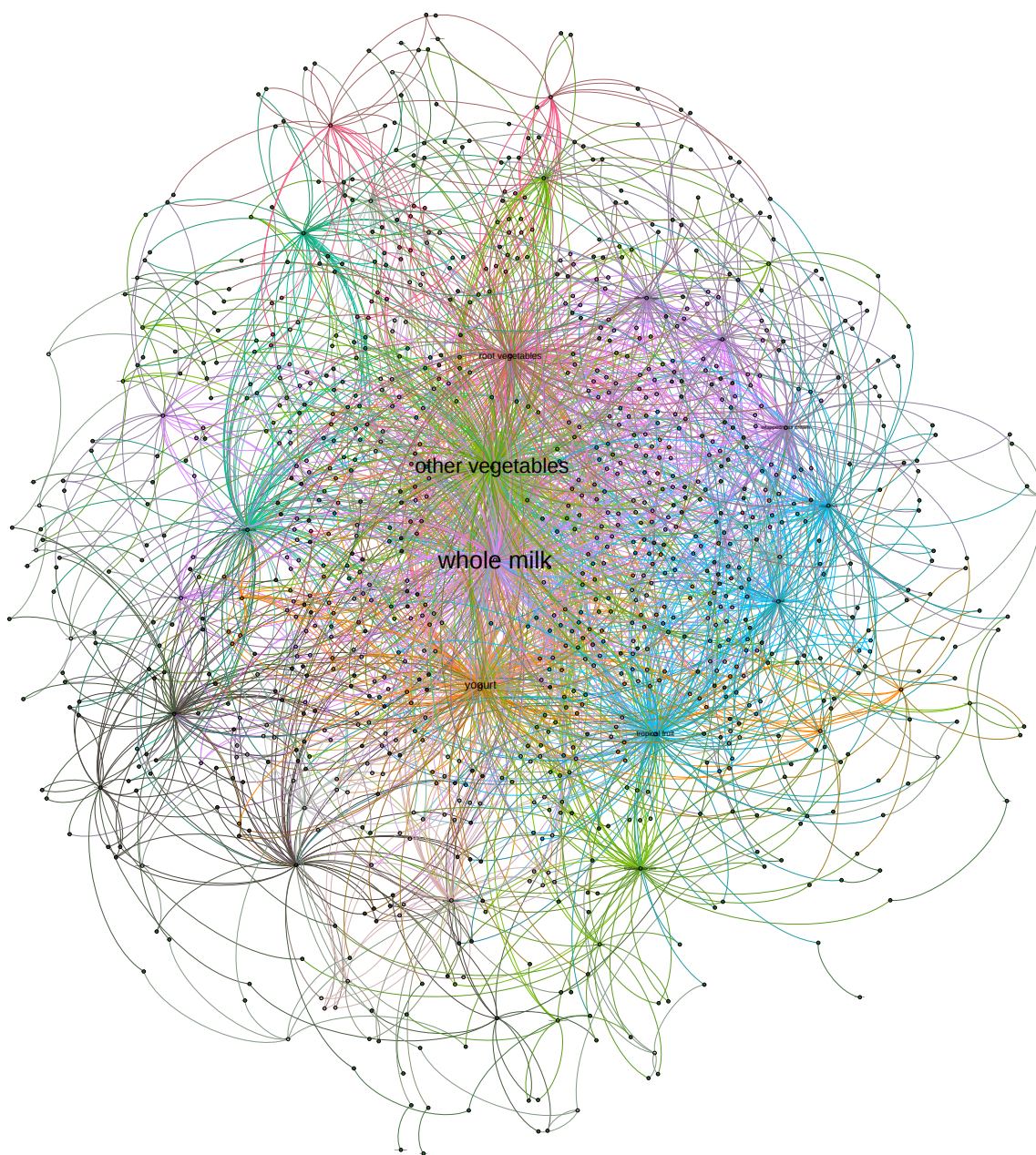
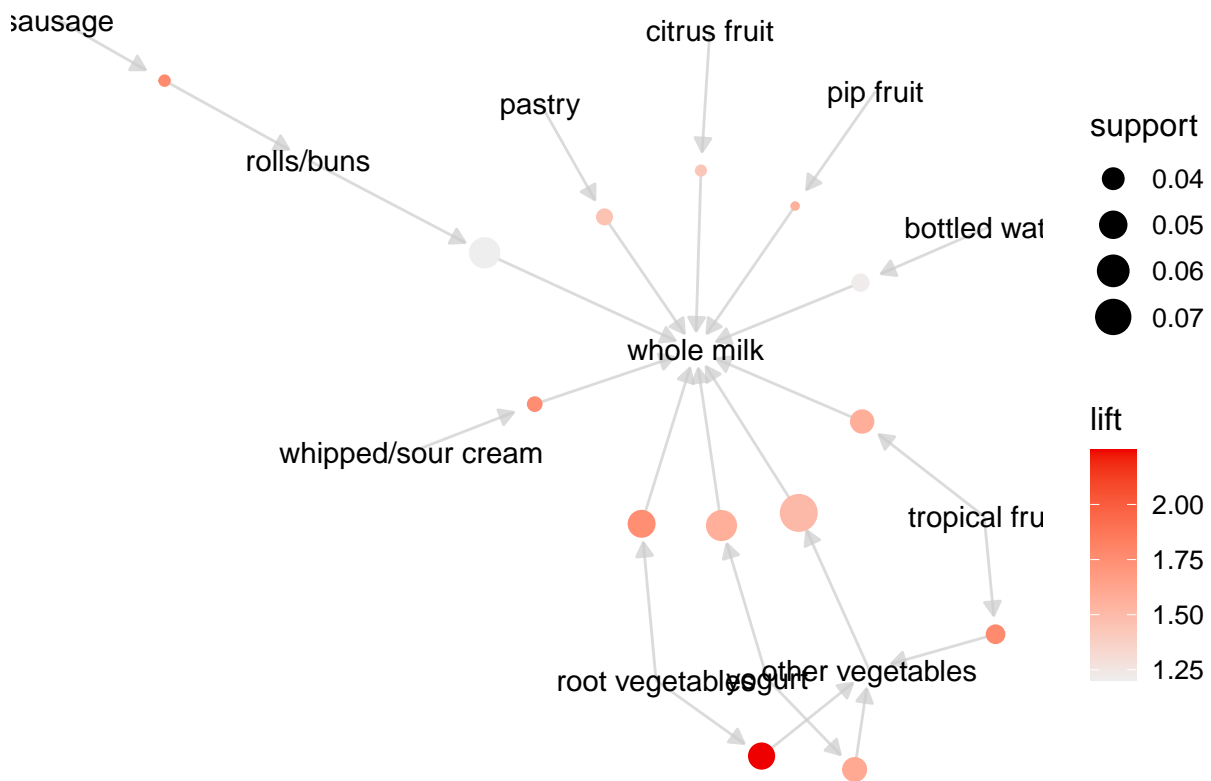


Figure 1: Support > 0.03 , Confidence > 0.03

```

## 14
##
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##        2        2        2        2        2        2
##
## summary of quality measures:
##      support      confidence      coverage      lift
##  Min.   :0.03010  Min.   :0.3079  Min.   :0.07168  Min.   :1.205
## 1st Qu.:0.03249  1st Qu.:0.3298  1st Qu.:0.09021  1st Qu.:1.475
## Median :0.03910  Median :0.3802  Median :0.10696  Median :1.575
## Mean   :0.04260  Mean   :0.3759  Mean   :0.11484  Mean   :1.604
## 3rd Qu.:0.04853  3rd Qu.:0.4027  3rd Qu.:0.13226  3rd Qu.:1.759
## Max.   :0.07483  Max.   :0.4496  Max.   :0.19349  Max.   :2.247
##      count
##  Min.   :296.0
## 1st Qu.:319.5
## Median :384.5
## Mean   :419.0
## 3rd Qu.:477.2
## Max.   :736.0
##
## mining info:
##      data ntransactions support confidence
## groceries      9835    0.005      0.1

```



Once again, the following graph created in gephi breaks up the grocery items into 7 orders (each shown as

a different color) and how they items are connected to each other.

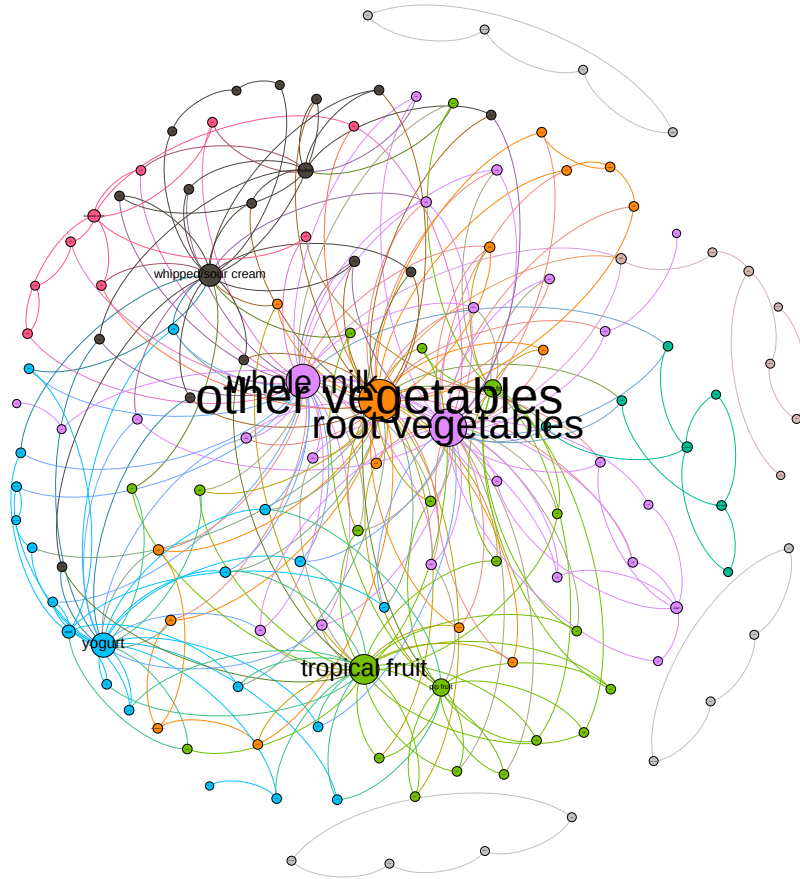


Figure 2: Support > 0.03, Confidence > 0.3

As we can see this is a much simpler visual of the data and both versions of the plot can help us easily determine certain things. Such as the first shows some of the other items commonly bought with groceries such as pasties, fruit, and bottle water. While in the second plot it is easier to what items are more connected to items in other groups, and how connected things like vegetables and whole milk are to the rest of the information.