Association Rule Analysis for the NLTCS data

The NLTCS data lists the proficiency of 21574 older Americans in sixteen daily activities, including dressing, cooking, and managing money. Six of these are categorised as activities of daily living, and a further ten are categorised as instrumental activities of daily living. In the data, 0 indicates that the person has no problem in performing the task, whilst 1 indicates that they have problems in performing it. The frequency of the sixteen categories ranges from a minimum of 10.5% for Y1 (eating) – meaning that 10% of those surveyed have problems in eating; through to a maximum of 67.5% for Y7 (heavy house work). In other words, 67.5% of those surveyed have problems in performing heavy house work. It is possible for an adult to have no problems in performing any of the activities. It is also possible to have problems, possibly inter-related, in several activities. The median number of disabilities per person is 4, and the mean is number of disabilities per person is 5.3. There is a standard deviation of 4.7.

The *a priori* algorithm was used to complete an association rule analysis of the data. The minimum support threshold (showing the co-occurrence of two disabilities) was set to 10% (or 0.1) as the least frequent disability (eating) occurs in 10.5% of those surveyed. A threshold greater than this would result in this rule not being taken into account. A number of confidence thresholds were tried out. It was noticed that the choice of confidence threshold had a significant effect on the number of rules returned (as shown in table 1 below). Eventually a confidence threshold of 90% (or 0.9) was selected, meaning that the consequent (rhs) of any rule has probability of at least 0.9 given the antecedent (lhs). The number of rules found, using these settings came to a (manageable) 19.

Table 1 – Relationship between choice of confidence level and number of rules found (*support = 0.1)*

Confidence Threshold	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
# of rules (len=2)	164	115	85	49	19
# of rules (len=3)	985	887	710	533	257

The following nineteen rules were found (table 2) and are sorted using the support measure of quality on the left and the lift measure of quality on the right.

Table 2 – The nineteen association rules in the NLCTS data

Sorted by support (high -> low)

```
    lhs       rhs    support   confidence lift
19 {Y11} => {Y7}   0.4453045 0.9169610  1.357105
18 {Y3}  => {Y12}  0.3748957 0.9299759  1.676832
17 {Y9}  => {Y7}   0.3457402 0.9755428  1.443806
16 {Y2}  => {Y7}   0.2572078 0.9330755  1.380954
14 {Y2}  => {Y3}   0.2539631 0.9213049  2.285412
13 {Y10} => {Y7}   0.2528507 0.9758497  1.444260
15 {Y2}  => {Y12}  0.2492352 0.9041534  1.630272
12 {Y10} => {Y11}  0.2421433 0.9345259  1.924355
11 {Y6}  => {Y7}   0.2294428 0.9257528  1.370117
10 {Y8}  => {Y7}   0.2161398 0.9982873  1.477468
5  {Y14} => {Y7}   0.2087235 0.9098808  1.346626
8  {Y8}  => {Y9}   0.2022805 0.9342753  2.636157
9  {Y8}  => {Y11}  0.2017243 0.9317063  1.918548
7  {Y4}  => {Y7}   0.1975526 0.9507027  1.407043
4  {Y15} => {Y7}   0.1950959 0.9246485  1.368482
6  {Y4}  => {Y5}   0.1904144 0.9163507  2.088459
3  {Y16} => {Y7}   0.1318717 0.9048982  1.339252
2  {Y1}  => {Y7}   0.1029480 0.9719912  1.438550
1  {Y1}  => {Y5}   0.1021137 0.9641138  2.197316
```

Sorted by lift (high -> low)

```
    lhs       rhs    support   confidence lift
8  {Y8}  => {Y9}   0.2022805 0.9342753  2.636157
14 {Y2}  => {Y3}   0.2539631 0.9213049  2.285412
1  {Y1}  => {Y5}   0.1021137 0.9641138  2.197316
6  {Y4}  => {Y5}   0.1904144 0.9163507  2.088459
12 {Y10} => {Y11}  0.2421433 0.9345259  1.924355
9  {Y8}  => {Y11}  0.2017243 0.9317063  1.918548
18 {Y3}  => {Y12}  0.3748957 0.9299759  1.676832
15 {Y2}  => {Y12}  0.2492352 0.9041534  1.630272
10 {Y8}  => {Y7}   0.2161398 0.9982873  1.477468
13 {Y10} => {Y7}   0.2528507 0.9758497  1.444260
17 {Y9}  => {Y7}   0.3457402 0.9755428  1.443806
2  {Y1}  => {Y7}   0.1029480 0.9719912  1.438550
7  {Y4}  => {Y7}   0.1975526 0.9507027  1.407043
16 {Y2}  => {Y7}   0.2572078 0.9330755  1.380954
11 {Y6}  => {Y7}   0.2294428 0.9257528  1.370117
4  {Y15} => {Y7}   0.1950959 0.9246485  1.368482
19 {Y11} => {Y7}   0.4453045 0.9169610  1.357105
5  {Y14} => {Y7}   0.2087235 0.9098808  1.346626
3  {Y16} => {Y7}   0.1318717 0.9048982  1.339252
```

The rule with the highest support is {grocery shopping} -> {heavy house work} (Y11->Y7). The support of this rule tells us that 44.5% of the respondents have this combination of disabilities. The confidence value tells us that amongst all respondents that have problems in grocery shopping, 91.6% also have problems with heavy house work. The lift value tells us that the two problems co-occur 1.35 times more than if these two disabilities were independent.

The rule with the highest lift is {doing light house work} -> {doing laundry} (Y8->Y9), telling us that problems with doing these tasks co-occur 2.63 more times than if the two were independent. Some of this may come from the design of the survey, as many respondents might reasonably consider the two activities to be very similar. This can also explain the strong relationship between 'getting in / out of bed' (Y2) and 'getting around inside' (Y3), with the relatively high lift value of 2.28.

It is noticable that 'heavy house work' (Y7), co-occurs with 11 of the remaining 15 conditions. This also happens, though on a much smaller basis, with 'bathing' (Y5), 'grocery shopping' (Y11) and 'getting about outside' (Y12).

All of this can be represented graphically as in the two figures below.

Figure 1 – Scatter Plot of the 19 rules

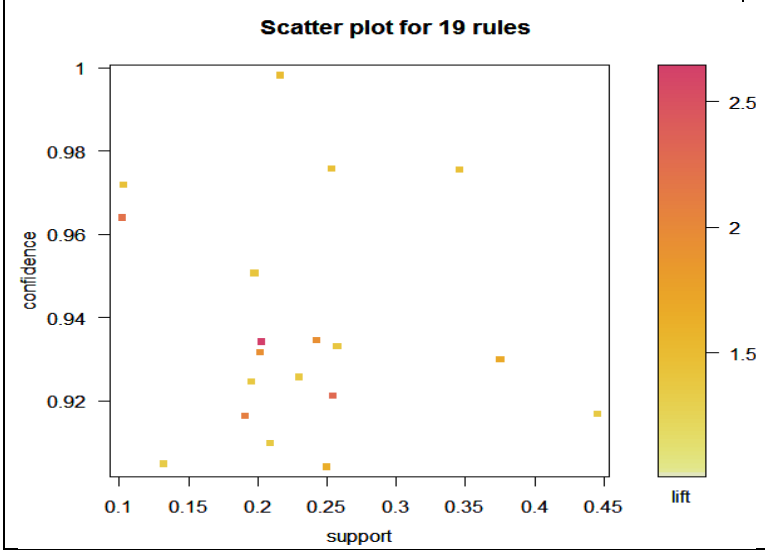
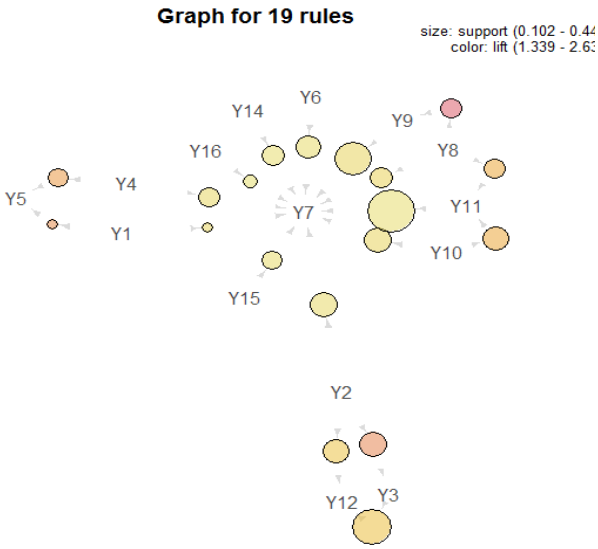


Figure 2 – Graph for the 19 rules



The two figures tell complementary stories about the data. The one on the left is useful for an overview. It examines the relative relationship between confidence, support and lift. It can be noted that two of the rules have low support values (0.1) combined with high levels of lift and confidence. A further outlier has a high level of support (0.45) combined with low levels of lift and confidence. The majority of the rules are clustered with levels of support between 0.17 and 0.25, confidence levels between 0.9 and 0.93, and lift levels just below 1.

The graph on the right shows the central part that 'doing heavy house work' (Y7) plays in the data, and it can be noted how a majority of each of the other rules point to it, with the relationship between 'grocery shopping' (Y11) and 'heavy house work' (Y7) being particularly prominent, in terms of high support. The strong relationships between 'getting in/out of bed' -> 'getting around inside' (Y2->Y3) and 'getting around inside' -> 'getting about outside' (Y3->Y12) can also be examined at the bottom right of the chart – something that might not have been noticed by simply glancing at the rule list in table 2.