Data Analysis Module Assignment

Julian Hatwell

December 2, 2016

Student Number: S15142087
Tutor: Prof. Mohamed Medhat Gaber

1 Introduction

This is the individual individual write up of the Data Analysis module assignment. The Heart data set was chosen for reasons outlined in the group presentation; The mix of factor and numeric variables provided a variety of data cleansing challenges and the practical aspects of this data set for medical applications was appealing.

Exploratory analysis of the heart data set reveals very strong evidence of an association between chest pain and heart disease. In simple terms, relatively few patients with heart disease also suffer chest pain and so do not have a constant reminder of their condition. It is hypothesised that they may be less mindful of their risk. It is also hypothesised that there may be metabolic indicators in the data that are associated with the outward symptoms of atherosclerotic heart disease and chest pain. Finding these indicators may provide some additional information about these patients that might be useful in helping them better manage their condition.

Clustering will be used to look for patterns in the other variables of the dataset which might indicate an underlying association. The findings will be considered in the context supporting these patients.

2 Aims and Objectives

2.1 Aim

To determine whether clustering is useful in finding underlying associations or interactions between heart disease, chest pain and the other variables.

2.2 Objectives

1. Use two different clustering techniques to mine the Heart data set
2. Describe any alignment of the clusters with the pattern of association between heart disease and chest pain
3. Evaluate which of the techniques and tuning parameter settings gives closest alignment to the pattern (if any)
4. Critically assess the results in the context of a possible medical intervention for heart disease patients who do not feel chest pain

3 Exploratory Analysis

The main exploratory analysis for this investigation has been presented in the group presentation. This section contains some supplementary information relevent to the specific research question.

3.1 Categorical Variables

Using a frequency table, it is possible to see an association between the presence of heart disease and various categories of chest pain, including asymptomatic (no chest pain). See table 1.

For the purpose of this investigation, a new factor variable (sympt) is created from Chest Pain which is coded as either ”No” or ”Yes” for Chest Pain Symptoms. See table 2.

```
heart$sympt <- factor(ifelse(heart$ChestPain == "asymptomatic"
                             , "No", "Yes"))
```

The frequency table may be visualised as a mosaic plot with Pearson's residuals shading, as in Figure 1. This shows very strong evidence of an association between presence of heart disease and absence of symptoms of chest pain. The research question stems from this evidence and seeks to find any underlying associations in the other variables. These might be thought of as metabolic indicators of the reported symptoms.

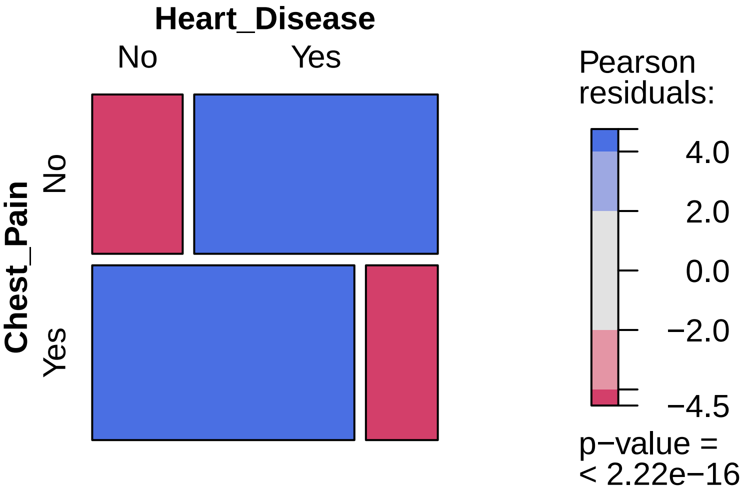

Figure 1: Mosaic plot of heart disease against presence of symptoms of chest pain. Pearson's residuals indicate that there is very strong evidence of an association between the two variables.

Table 1: Frequency Table of presence of heart disease against various types of reported symptoms of chest pain.

	No	Yes
asymptomatic	39	105
nonanginal	67	18
nontypical	41	9
typical	16	7

Table 2: Frequency Table of presence of heart disease against and reported symptoms of chest pain.

	No	Yes
No	39	105
Yes	124	34

4 Methodology

4.1 General Preprocessing

The preprocessing steps outlined in the group presentation have been implemented in the following code chunk and run on the data prior to any further steps described below.

```
# data is loaded as heart

# tidy data. Rename long var name and remove ID column X
names(heart) <- c(names(heart)[1:14], "HDisease")
heart <- heart[, -1]

# Apply the four transformations identified by team
# NA Ca
library(VIM)
hearttemp <-kNN(heart, "Ca")
heart <- hearttemp[,1:14]

# NA thal
heart$Thal[is.na(heart$Thal)] <- names(which.max(table(heart$Thal)))

# Outlier Chol
maxChol <- which.max(heart$Chol)
heart <- heart[-maxChol,]

# Skew Oldpeak
```

```
heart$Oldpeak <- sqrt(heart$Oldpeak)

# Scale the numeric variables
heart.unscaled <- heart # need this for unscaling later

# get numeric var names
vars.types <- sapply(heart, class)
num.vars <- names(vars.types[vars.types %in% c("integer", "numeric")])
other.vars <- names(vars.types[!(vars.types %in% c("integer", "numeric"))])

library(reshape)
scaled.vars <- sapply(heart[, num.vars]
                        , rescaler
                        , type = "range")

heart <- cbind(heart[, other.vars], scaled.vars)
```

4.2 Task Specific Pre-Processing

There is no need to split the data into train and test sets. Clustering is run on all instances of the data. The idea is to find structures and patterns without reference to a particular class label and there are no predictions against previously unseen samples, as is the case in predictive analytics.

Both the clustering algorithms used for this investigation use distance measures which require all the features to be scaled numeric variables but some of the variables are categorical. Various strategies exist to overcome this problem:

- Variations on the algorithms are available that can use categorical variables
- Excluding categorical variables from the analysis - this may lead to significant loss of information
- Finding ways to recode factors as binary or numeric while still retaining the information

A quick scan of this data set shows that most of the categorical variables have already been coded as binary (0 or 1) or a small ordinal (e.g. $Ca \in \{0, \ldots 3\}$). As this research question suggests removing the Heart Disease and Chest Pain variables before clustering, this only leaves Thal as a categorical variable. Further analysis shows that Thal could feasibly be reduced to a binary (normal or abnormal) as there are only 18 out of 303 $\approx$ 6% cases in one of the two abnormal categories. Given the minimal effort involved and the rather small loss of information, this is the approach taken. Further work might involve comparing results from this investigation with the extended algorithms.

```
heart$Thal <- ifelse(heart$Thal == "normal", 0, 1)
```

After that, it is simply a case of creating an experimental data set by removing the Heart Disease, Chest Pain and Chest Pain Symptoms variables.

```
rm.cols <- which(names(heart) %in% c("HDisease"
                                     , "ChestPain"
                                     , "sympt"))
heartx <- heart[, -rm.cols]
```

4.3 Clustering analysis with k-means

The k-means algorithm requires the researcher to set a value for k. It is useful to refer back to the research question when considering a reasonable base value or values. As the research question is characterised by a 2*2 frequency table, the parameter value $k \in \{2, 4\}$ is a good candidate for a first cut.

```
set.seed(2016)

K <- c(2, 4)

km <- list()
for (k in K) {
  km[[k]] <- kmeans(heartx , centers = k)
}
```

4.4 Hierarchical Clustering Analysis

Hierarchical Agglomerative Clustering also is attempted with the data. The algorithm is run for each of the following 4 methods and initially examined at the 2nd and 4th nodes:

1. Centroid
2. Average
3. Complete
4. Single

```
heartx.hclus1 <- hclust(dist(heartx)
                        , method="centroid")
heartx.hclus2 <- hclust(dist(heartx)
                        , method="average")
```

```
heartx.hclus3 <- hclust(dist(heartx)
                        , method="complete")
heartx.hclus4 <- hclust(dist(heartx)
                        , method="single")
```

4.5 Analysis of Clustering Outputs

The clustering outputs are tabulated and various types of plots are produced to assess the results visually and statistically.

Dendrograms are visually assessed by plotting, per method, per colouring scheme (Heart Disease and Chest Pain Symptoms variables from the original data set) and cutting the tree at various levels to identify a good separation between areas of generally uniform colour. See Figures 2, 3 & 4.

Parallel coordinates plots are useful for assessing the separation between clusters from their centroids. This is immediately useful for k-means clustering output. H-clust centroids must first be calculated (e.g. by taking the mean of each variable by the instance cluster Id) and then plotted. See Figures 5, 6 & 7.

Dot plots with jitter are used to identify which clusters are associated with each of the 4 symptom combinations. See Figures 8, 9 & 10.

Mosaic plots, with Pearson's residual shading provide similar information as the dot plots with the additional benefit of a measure of statistical significance of any associations. See Figures 11, 12 & 13.

5 Results

A large number of models were created but many did not provide useful results. This section contains only the results from the models that appear to best answer the research question.

5.1 H-clust Choice of Model

In all the resulting dendrograms, the label colours of the leaf level eventually show some alignment with the symptoms in the original data set. However, the most satisfactory result with the smallest number of splits is using the ”average” method with the split at the 5th node. One cluster contains most of the leaf nodes of one colour and the other 4 contain leaves mostly of the second colour. See Figures 2, 3 & 4.

5.2 Tabulated Cluster Centroids

The cluster centroid values for the k-means algorithm are listed in Tables 3 & 4.

In addition, the cluster centroids for the hierarchical clusters are calculated as the mean for each variable per cluster. These results are also presented below. See Table 5.

```r
hclust.final <- cutree(heartx.hclus2, 5)

clusterMeans <- function(x) {
  tapply(x, hclust.final, mean)
}
hclust.centroids <- sapply(heartx, clusterMeans)
```

Table 3: Cluster centroids for k-means, k=2

	Thal	Age	Sex	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca
1	0.90	0.56	0.91	0.37	0.40	0.18	0.55	0.53	0.55	0.42	0.41	0.33
2	0.01	0.50	0.47	0.34	0.42	0.12	0.44	0.66	0.11	0.21	0.19	0.13

Table 4: Cluster centroids for k-means, k=4

	Thal	Age	Sex	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca
1	0.01	0.50	0.34	0.33	0.42	0.09	0.26	0.66	0.08	0.21	0.17	0.11
2	0.51	0.55	0.92	0.41	0.41	0.15	1.00	0.65	0.00	0.32	0.32	0.28
3	0.96	0.53	0.94	0.33	0.38	0.19	0.01	0.55	0.51	0.39	0.38	0.27
4	0.64	0.56	0.78	0.38	0.43	0.22	0.98	0.48	1.00	0.44	0.44	0.34

Table 5: Cluster centroids for Hclust Average method, level 5

	Thal	Age	Sex	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca
1	1.00	0.57	0.71	0.55	0.46	1.00	0.93	0.50	0.79	0.48	0.57	0.43
2	0.00	0.60	0.57	0.35	0.43	0.11	0.70	0.53	1.00	0.33	0.39	0.24
3	1.00	0.55	0.94	0.36	0.40	0.09	0.44	0.55	0.49	0.40	0.37	0.28
4	0.00	0.49	0.51	0.33	0.41	0.13	0.44	0.67	0.03	0.21	0.18	0.13
5	1.00	0.68	0.20	0.48	0.51	0.00	0.90	0.54	0.00	0.74	0.70	0.93

The information in these tables is more easily visualised in a parallel coordinates plot, which can be found in the Appendix. See Figures 5, 6 & 7.

Based on the visual analysis, clusters have been found that cover the various symptom combinations defined in the research question:

k-means ($k = 2$) Model:

- Cluster 1, Heart Disease Yes, Symptoms No, Strong Association
- Cluster 2, Heart Disease No, Symptoms Yes, Strong Association

k-means ($k = 4$) Model:

- Cluster 1, Heart Disease No, Symptoms Yes, Strong Association
- Cluster 2, No Association
- Cluster 3, Heart Disease Yes, Symptoms No, Strong Association
- Cluster 4, Heart Disease Yes, Symptoms No, Strong Association

Hclust, Average Linkage, 5 Nodes Model:

- Cluster 1, Heart Disease Yes, Symptoms No, Strong Association
- Cluster 2, Heart Disease Yes, Symptoms No, Weak Association
- Cluster 3, Heart Disease Yes, Symptoms No, Strong Association
- Cluster 4, Heart Disease No, Symptoms Yes, Strong Association
- Cluster 5, Heart Disease Yes, Symptoms No, Strong Association, Very Small Group

6 Discussion

Returning to the research question, the ideal model will provide a distinct metabolic profile that provides new information on patients with Atherosclerotic Heart Disease but no symptoms of chest pain. The best possible outcome is for this profile to yield some medically useful insight. In this context, each of the models described in the previous section has their pros and cons.

k-means ($k = 2$) Model:
Cons: This appears to be an oversimplification. Lack of model flexibility and may not provide adequate separation of true clusters. It is not possible to correctly model the 2*2 table which characterises the research question, leading to a higher risk of false positives and false negatives. See Figure 9.
Pros: This is the simplest model and therefore easy to interpret. There is good separation between the clusters on 6 out of 12 variables.

k-means ($k = 4$) Model:
This is possibly the least satisfactory model. There are two clusters (3 & 4) of

interest which are broadly similar and somewhat separable from the other two in a few variables (Thal, MaxHR, ExAng, Oldpeak, Slope). However, Clusters 2 is noisy; It is not associated with the symptoms symptoms in any way and is spread over all four quadrants of the 2*2 table. This will make diagnosis and intervention less reliable. See Figures 10 & 13.

Hclust, Average Linkage, 5 Nodes Model:
Cons: The one really useful cluster is actually the opposite of the pattern of interest, so the model is a bit herder to interpret. Medical practitioners have to identify patients of interest by exclusion from this group.
Pros: This one cluster separates well from all the others, scoring significantly higher or lower than the other 4 clusters on 5 variables. Using thresholds on these 5 in combination with specific scores on some of the other variables provides the potential to clearly distinguish these patients in a way that is not possible with a 2 cluster model. It is especially useful that this one cluster covers all 3 quadrants of the 2*2 table that are not of interest. See Figure 8.

In order to interpret these results, a reverse of the scaling operation must be carried out on the cluster centroids, so their values are restored to the same scale as the original variables. A set of diagnostic rules could then be generated in flow-chart or decision tree format for use by a medical practitioner. See Table 6.

```
unscale <- function(x, name_x) {
    if (name_x %in% c("Sex", "Fbs", "ExAng", "Thal")) {
      x <- round(x)
    } else {
      x <- x * sum(range(heart.unscaled[[name_x]])) +
      min(heart[[name_x]])
    }
}

centroids.unscaled <-
    mapply(unscale
           , hclust.centroids[4, ]
           , names(hclust.centroids[4, ]))
```

Table 6: Possible diagnostic tool based on the profile of Hclust 4.

	remarks	Typical Values
Thal	Not equal to	0
Age	Greater than	52
Sex	May be male or female	1
RestBP	Greater than	98
Chol	Greater than	223
Fbs	1 is high risk	0
RestECG	Generally less than	1
MaxHR	Less than	182
ExAng	1 is high risk	0
Oldpeak	Square root is greater than	1
Slope	Greater than 1 is high risk	1
Ca	Greater than	0

7 Conclusions

From the exploratory analysis of the Heart data set, it was determined that there was a strong association between the presence of Atherosclerotic Heart Disease and a lack of any symptoms of chest pain, despite chest pain being commonly reported by other patients in the database.

Clustering was used on the dataset, excluding the symptomatic indicators, to determine whether there was some underlying association or pattern. Hierarchical clustering with various linkage methods was tried, as well as k-means with different values for k. The Hierarchical model with average linkage yielded good initial results as did the two k-means cluster models.

A critical evaluation of the three candidate models determined that the preferred model to use in medical practice would be to extract cluster 4 from the final Hierarchical model and identify patients of interest by lack of fit with this specific profile. It is hoped that some useful preventative medical intervention can be developed from this new information.

8 Appendix

8.1 Figures

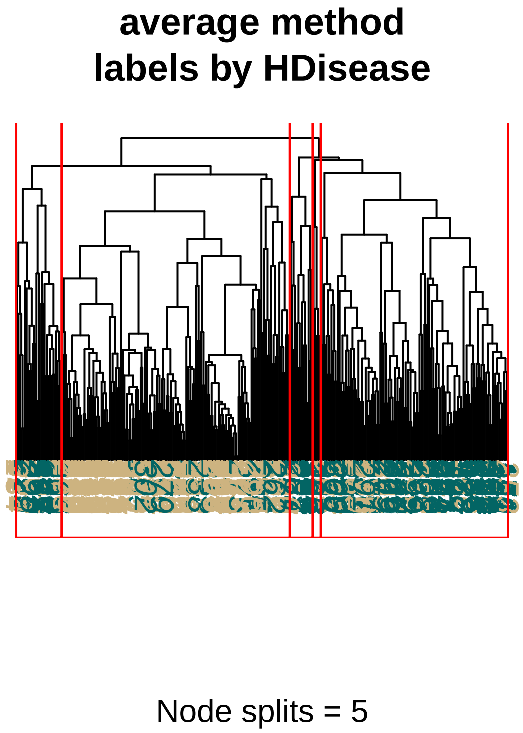

Figure 2: A selection of Cluster Dendrogram from the Heart data. Red lines indicate tree cut level. Label colours by Heart Disease. The best separation is achieved by the Average method with 5 clusters.

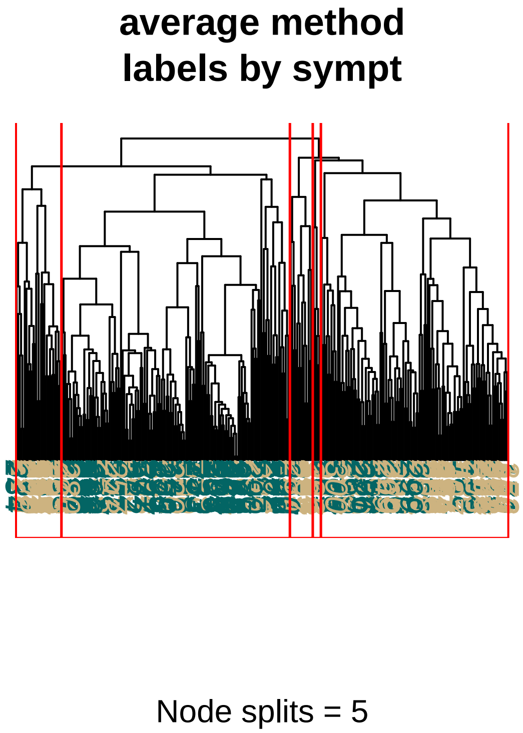

Figure 3: A selection of Cluster Dendrogram from the Heart data. Red lines indicate tree cut level. Label colours by Chest Pain Symptoms. The best separation is achieved by the Average method with 5 clusters.

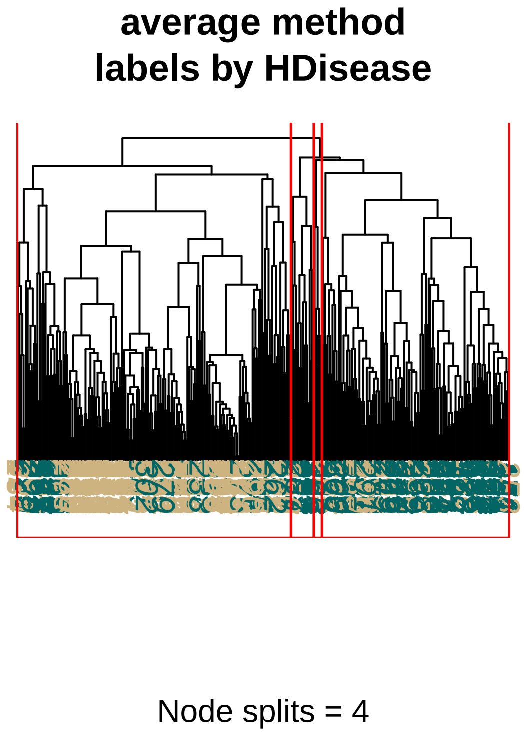

Figure 4: A selection of Cluster Dendrogram from the Heart data. Red lines indicate tree cut level. Label colours by Heart Disease. The best separation is achieved by the Average method with 5 clusters.

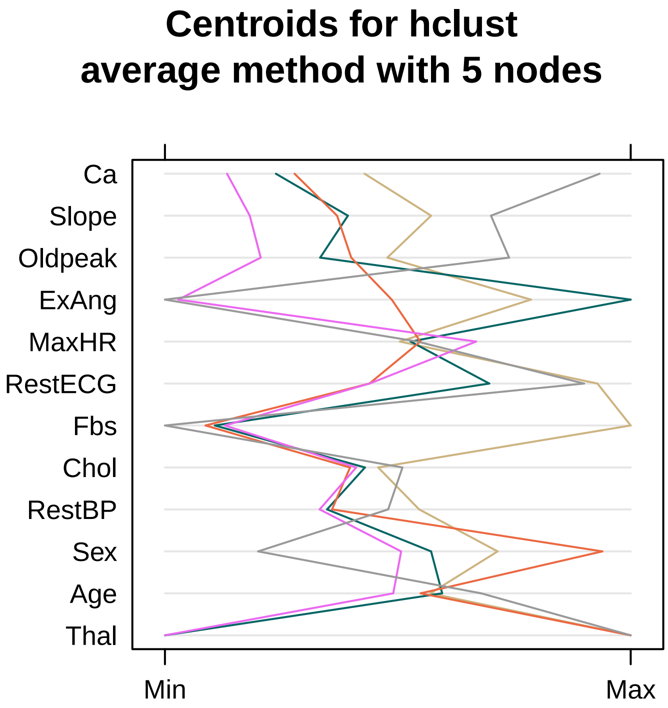

Figure 5: Parallel co-ordinates plot for centroids calculated at hclust node 5, showing values for the mean of each variable per cluster.

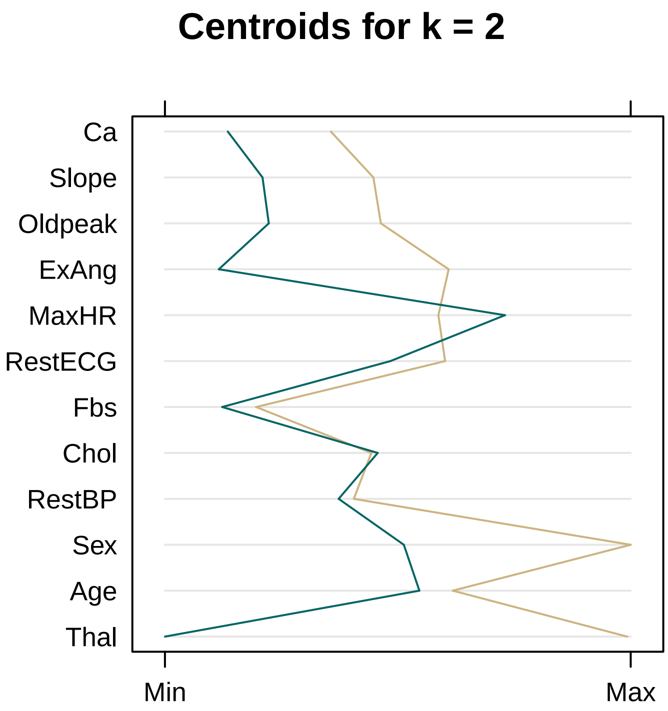

Figure 6: Parallel co-ordinates plot for k = 2 showing values of each variable for the cluster centroids.

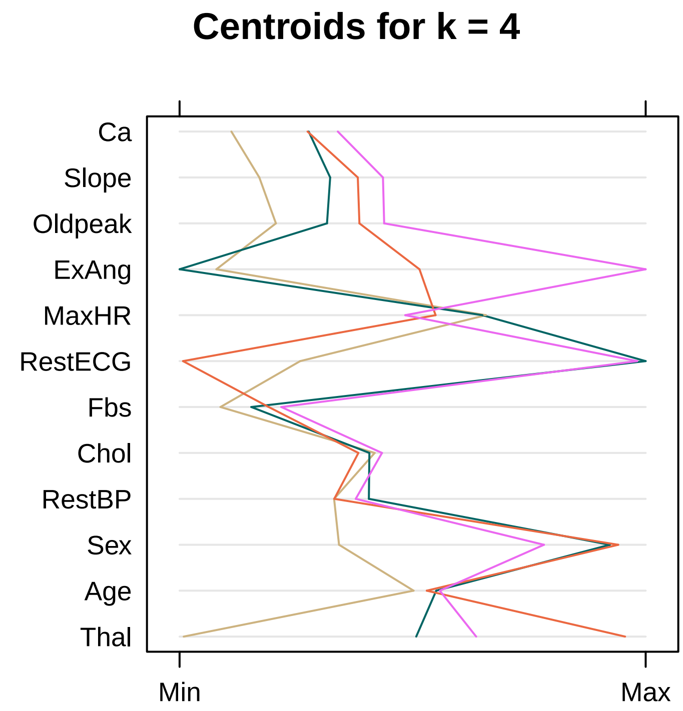

Figure 7: Parallel co-ordinates plot for k = 4 showing values of each variable for the cluster centroids.

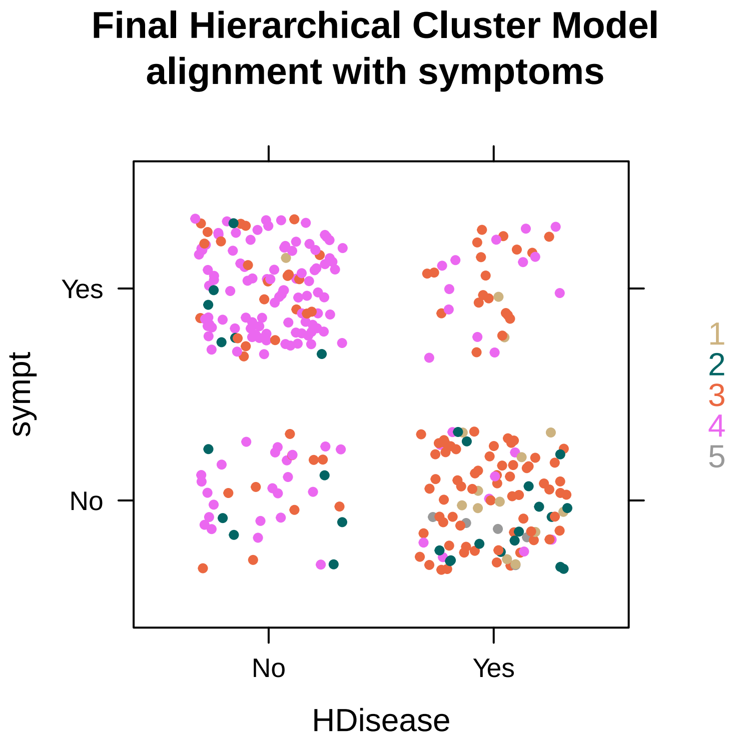

Figure 8: Final Hierarchical Cluster model alignment with symptoms.

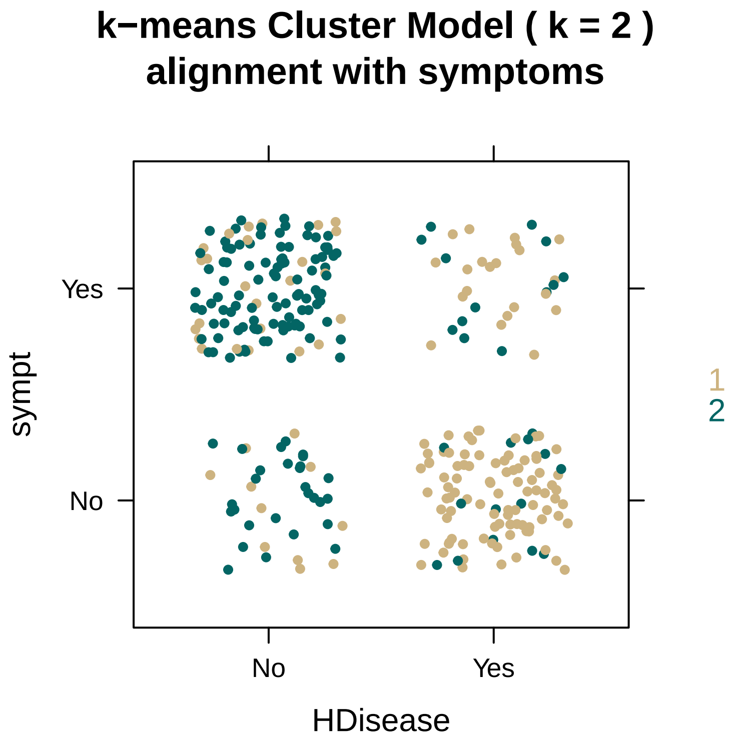

Figure 9: k-means cluster model alignment with symptoms with k = 2

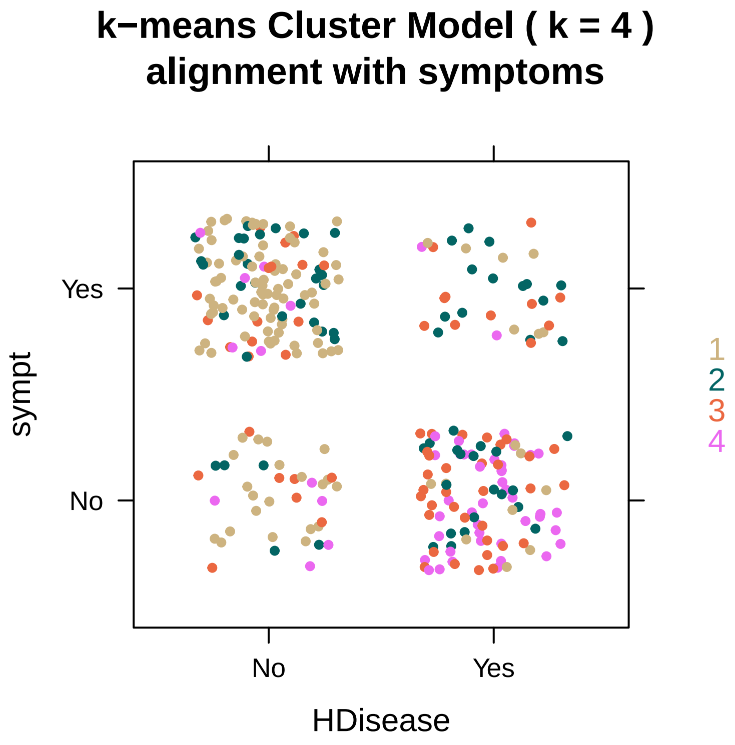

Figure 10: k-means cluster model alignment with symptoms with k = 4

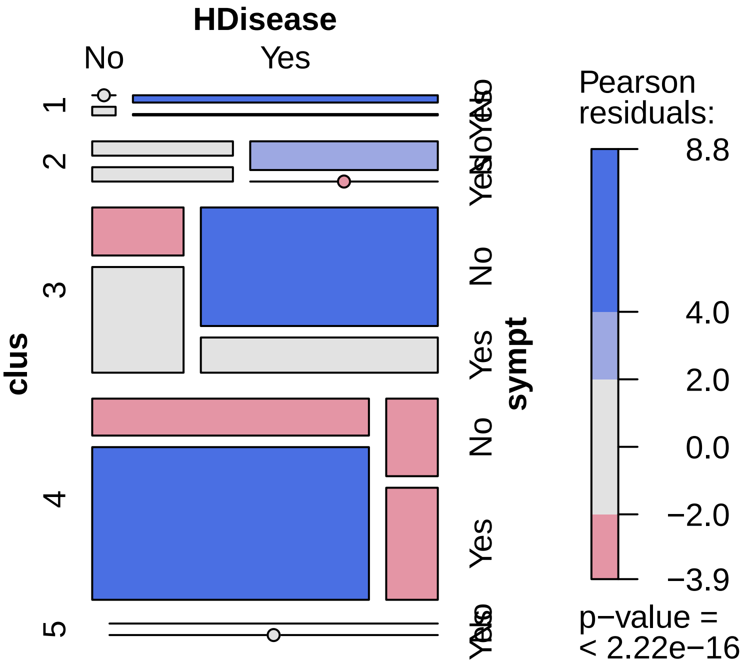

Figure 11: Mosaic Plot of the hierarchical cluster model with symptoms

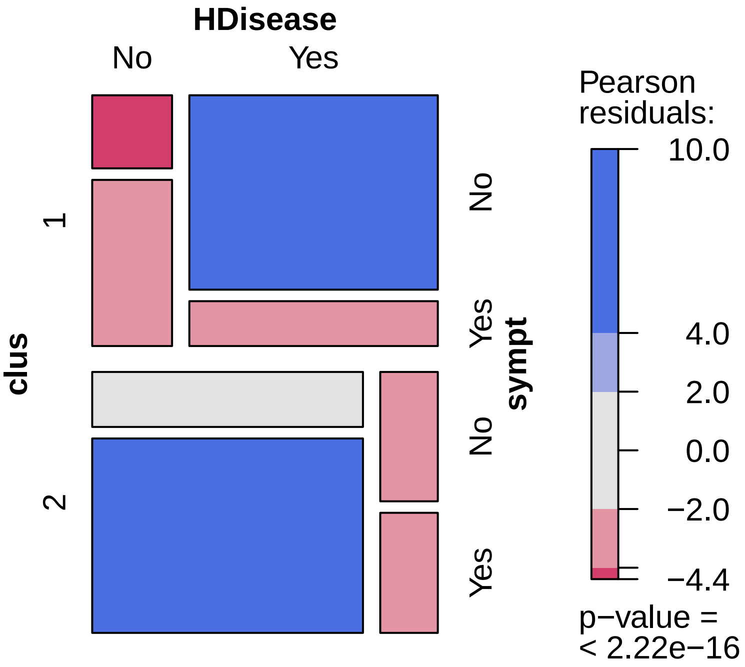

Figure 12: Mosaic Plot of the k = 2 k-means model with symptoms.

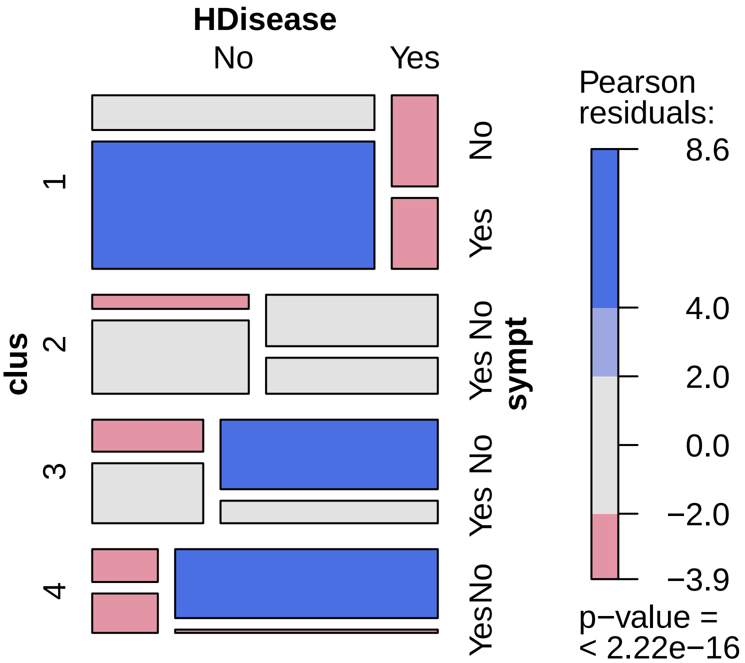

Figure 13: Mosaic Plot of the k = 4 k-means model with symptoms.