INTRODUCTION

This dataset contains information on the quality of life measures of 380 hospital patients. There are 22 quality of life variables (WORK2, HOBBY2, BREATH2, PAIN2, REST2, SLEEP2, APPET2, NAUSEA2, VOMIT2, CONST2, DIARR2, TIRED2, PAINADL2, CONCEN2, TENSE2, WORRY2, IRRIT2, DEPRES2, REMEM2, FAMILY2, SOCIAL2, FINANC2) and three other variables (sex, age, live with).

The quality of life variables have the following ranking: 1 (do not have this quality at all), 2 (have this quality a little), 3 (have this quality quite a bit), and 4 (have this quality very much).

In this dataset, the rows of any of the 22 quality of life measures which had missing values were removed. 78 rows were removed.

Different kinds of clustering methods were performed with this dataset to identify different groups of patients and to interpret the output. The following methods were used: K-means, Partitioning around medoids (pam), Gaussian mixture (Mclust) and Latent class (poLCA).

K-MEANS

K-means was carried out by using 4 clusters and treating the 22 quality of life variables as continuous. The procedure was run 200 times with different starting values. After that, the minimum total within sums of squares was obtained (3445.71). The clusters obtained had the following sizes: 50, 95, 62, 95. The composition of each cluster can be seen in the below table:

Quality of life variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
WORK2	3.780000	1.431579	3.354839	2.705263
HOBBY2	3.600000	1.284211	3.290323	2.578947
BREATH2	2.760000	1.273684	1.951613	1.589474
PAIN2	3.040000	1.284211	2.612903	1.789474
REST2	3.560000	1.642105	3.306452	2.421053
SLEEP2	3.320000	1.589474	2.451613	1.852632
APPET2	2.720000	1.168421	1.903226	1.473684
NAUSEA2	2.760000	1.073684	1.661290	1.305263
VOMIT2	2.000000	1.000000	1.193548	1.084211
CONST2	2.540000	1.231579	1.677419	1.378947
DIARR2	1.440000	1.147368	1.419355	1.305263
TIRED2	3.620000	1.810526	3.338710	2.463158
PAINADL2	3.000000	1.052632	2.483871	1.473684
CONCEN2	2.900000	1.168421	2.064516	1.568421
TENSE2	3.180000	1.400000	2.209677	1.852632
WORRY2	3.280000	1.684211	2.403226	2.294737
IRRIT2	3.040000	1.284211	1.870968	1.642105
DEPRES2	3.040000	1.273684	1.935484	1.694737
REMEM2	2.240000	1.315789	1.854839	1.610526
FAMILY2	3.400000	1.168421	2.693548	1.747368
SOCIAL2	3.540000	1.252632	3.274194	2.305263
FINANC2	2.480000	1.210526	1.516129	1.452632

Table 1. K-means clustering with 4 clusters

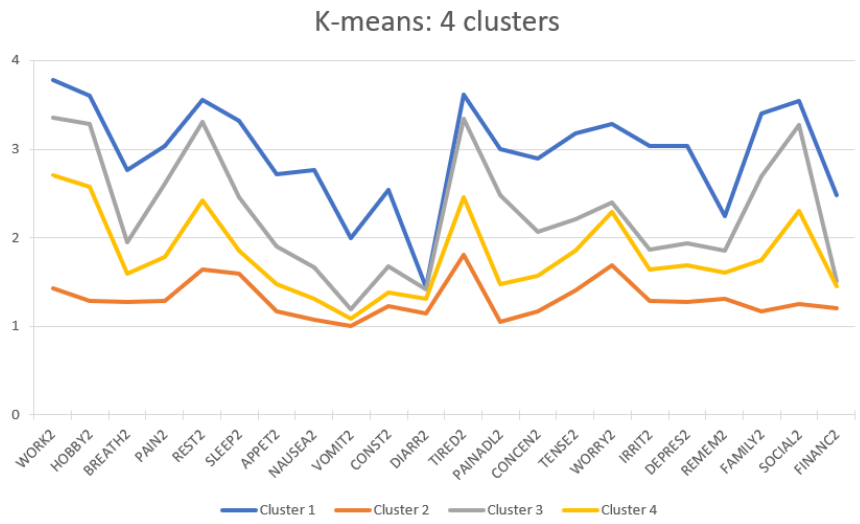

Figure 1. Plot K-means clustering with 4 clusters

The characteristics of each cluster group are as follows:

Cluster 1: In general, the quality of life for these patients is very bad when it is compared against the quality of life of the other groups of patients. In other words, all the variables of quality of life for this group have a greater ranking than the other groups. They have a high ranking when it comes to working limitations, hobby limitations, tiredness and social limitations. However, the variable "diarr" has the lowest ranking (1.4) for these patients.

Cluster 2: Overall, this group of patients has the best quality of life when it is compared against the other groups. Within this group, the quality of life variables which suffer the highest ranking are: "tired", "rest", "sleep", and "worry". However, these are between 1.5 and 1.8 which is relatively low. These patients do not suffer from vomiting.

Cluster 3 and Cluster 4: The trajectories of these groups in the plot are in the middle of cluster 1 and cluster 2. The shape of lines in the plot of cluster 3 and cluster 4 is very similar, but the patients who belong to cluster 4 have a better quality of life than cluster 3. Within these groups, the quality of life variables which suffer the highest ranking are: "work", "hobby", "rest", "tired" and "social". Alternately, the variables which have the lowest ranking are: "vomit" and "nausea". For cluster 3, the variables of quality of life are between 1.19 and 3.35 and for cluster 4 they are between 1.08 and 2.70.

Then the elbow method was performed to determine the optimal number of clusters for k-means clustering. This method allows us to decide the number of principal components. (James, 2007). In Figure 2, the scree plot is shown.

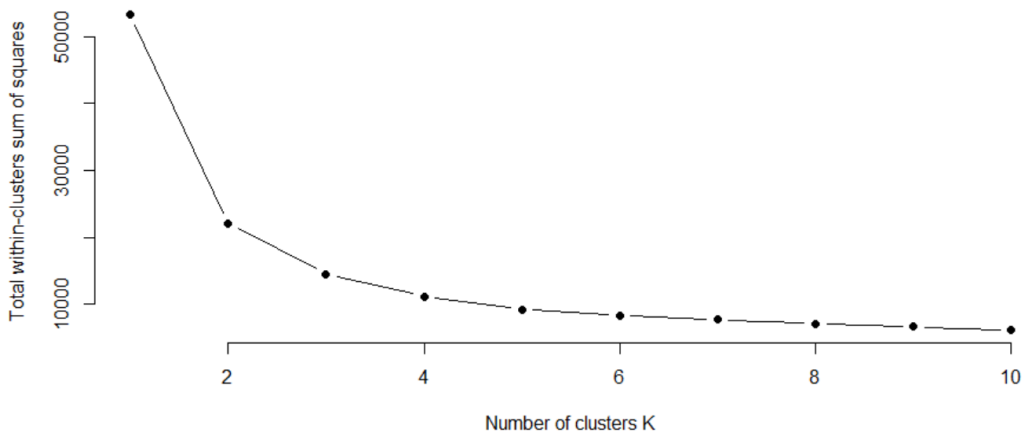

Figure 2. Scree plot

The scree plot looks very smooth, so it is very difficult to get a conclusion about the best number of clusters. Therefore, the NbClust package was used because "it provides 30 indices which determine the number of clusters in a data set and it offers also the best clustering scheme from different results to the user" (Charrad, 2014). This method suggests that the best number of clusters is 2.

```
*******************************************************************
* Among all indices:
* 12 proposed 2 as the best number of clusters
* 9 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 1 proposed 14 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

Figure 3. Result of using NbClust

After using NbClust, K-means was carried out using 2 clusters. The procedure was run 200 times with different starting values. It was very stable and the value of the minimum total within sums of squares was 4140.47. The clusters obtained had the following sizes: 195, 107. The composition of each cluster can be seen in the following table:

Quality of life variables	**Cluster 1**	**Cluster 2**
WORK2	3.570093	2.092308
HOBBY2	3.429907	1.969231
BREATH2	2.345794	1.435897
PAIN2	2.822430	1.558974
REST2	3.429907	2.061538
SLEEP2	2.897196	1.717949
APPET2	2.308411	1.323077
NAUSEA2	2.17757	1.20000
VOMIT2	1.579439	1.041026
CONST2	2.065421	1.323077
DIARR2	1.448598	1.220513
TIRED2	3.485981	2.158974

PAINADL2	2.719626	1.297436
CONCEN2	2.457944	1.384615
TENSE2	2.700935	1.620513
WORRY2	2.878505	1.964103
IRRIT2	2.429907	1.466667
DEPRES2	2.485981	1.476923
REMEM2	2.056075	1.461538
FAMILY2	3.037383	1.482051
SOCIAL2	3.411215	1.810256
FINANC2	2.009346	1.312821

Table 2. K-means clustering with 2 clusters

In Figure 3, the cluster means and their trajectories can be observed.

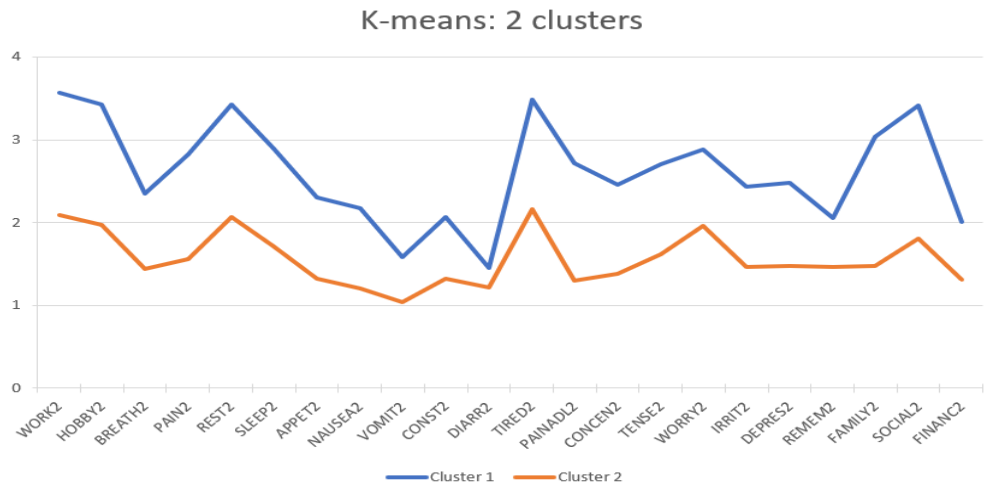

Figure 3. Plot K-means clustering with 2 clusters

The characteristics of each cluster group are as follows:

Cluster 1: Overall, this group of patients has a less favourable quality of life than cluster 2. They have a high ranking when it comes to "work" (3.57), "hobby" (3.42), "rest" (3.42), "tired" (3.48), and "social" (3.41). The variables "vomit" (1.57) and "diarr" (1.44) have the lowest ranking.

Cluster 2: In general, the quality of life of these patients is between 1 and 2, which means they do not have this quality (1) or have a little (2). These patients do not suffer of vomiting and they suffer more of "tired" (2.15), "worry" (1.96), "rest" (2.06), "worry" (1.96) and "hobby" (1.96).

PARTITIONING AROUND MEDOIDS

"Partitioning around medoids" (pam) technique with 4 clusters was used. Using this technique, "each cluster is represented by one of the data point in the cluster. These points are named cluster medoids" (STHDA, 2017). In other words, the medoids are the objects that represent clusters. Pam was carried out using 4 clusters. The medoids of each cluster can be seen in the following table:

Quality of life variables	Medoid 1	Medoid 2	Medoid 3	Medoid 4
ID	**135**	**215**	**353**	**360**
WORK2	3	4	2	1
HOBBY2	3	4	2	1
BREATH2	2	2	1	1
PAIN2	1	4	2	1
REST2	3	4	2	1
SLEEP2	2	3	1	1
APPET2	1	2	1	1
NAUSEA2	1	3	1	1
VOMIT2	1	3	1	1
CONST2	1	3	1	1
DIARR2	1	1	1	1
TIRED2	3	4	2	1
PAINADL2	1	4	1	1
CONCEN2	2	2	1	1
TENSE2	2	3	2	1
WORRY2	2	4	2	1
IRRIT2	2	3	2	1
DEPRES2	2	4	2	1
REMEM2	2	2	1	1
FAMILY2	2	4	2	1
SOCIAL2	3	4	2	1
FINANC2	2	2	1	1

Table 3. Pam technique with 4 clusters

Evidence that that the clusters are not isolated was gotten by using “Isolation”:

```
out2$isolation
```

```
1 no
2 no
3 no
4 no
```

Figure 4. Isolation of 4 clusters

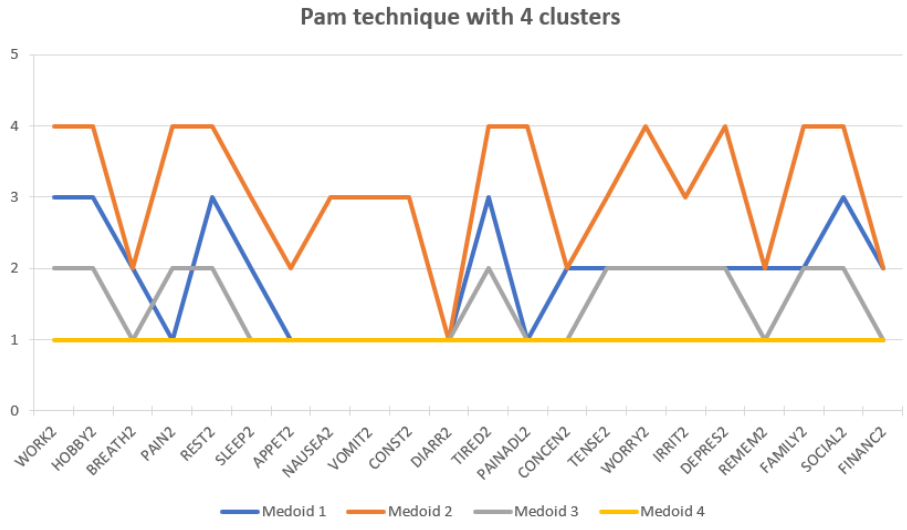

Figure 5. Plot Pam technique with 4 clusters

The description of each medoid is as follows:

Medoid 1 and Medoid 3: these medoids are between medoid 2 and medoid 4. Medoid 1 has more work, social and hobby limitations, short of breath, need of rest and tiredness, than medoid 3. In terms of pain, medoid 1 does not suffer from pain and medoid 3 suffers from a little amount of pain. Both of them have a little amount of "tense", "worry", "irrit" and "depress".

Medoid 2: This group tends to be very unhealthy because the value of their quality of life is very high if it is compared to the other medoids. For example, patient "215" has the highest value for each variable of quality of life, when the trajectories are compared against the other medoids.

Medoid 4: These types of patients tend not to have any qualities of life, which means they tend to be very healthy.

The pam result is not similar to the k-means analysis, because k-means clusters do no overlap in the different quality of life variables. Conversely, Pam clusters have some overlaps between each group such as: "tense", "worry", "irrit" and "depress".

Then, the average silhouette width was performed to determine the optimal number of clusters.

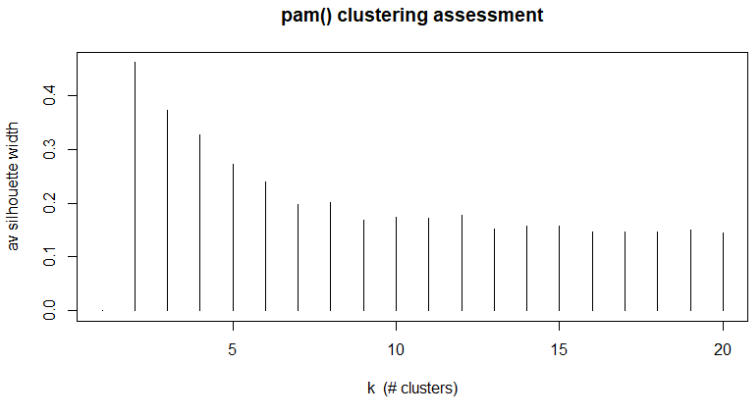

Figure 6. Result of using average silhouette width

Figure 6 shows that the silhouette-optimal number of clusters is 2. After this, 4 clusters are not a reasonable choice. Therefore, Pam technique with 2 clusters was used and the medoids of each cluster can be seen in the following table:

Quality of life variables	**Cluster 1**	**Cluster 2**
ID	**353**	**212**
WORK2	2	3
HOBBY2	2	4
BREATH2	1	3
PAIN2	2	3
REST2	2	3
SLEEP2	1	3
APPET2	1	2
NAUSEA2	1	1
VOMIT2	1	1
CONST2	1	1
DIARR2	1	1
TIRED2	2	3

PAINADL2	1	3
CONCEN2	1	3
TENSE2	2	3
WORRY2	2	3
IRRIT2	2	2
DEPRES2	2	2
REMEM2	1	3
FAMILY2	2	4
SOCIAL2	2	4
FINANC2	1	2

Table 4. Pam technique with 2 clusters

Evidence that that the clusters are not isolated was gotten by using “Isolation”:

```
out3$isolation
```

```
1  no
2  no
```

Figure 7. Isolation of 2 clusters

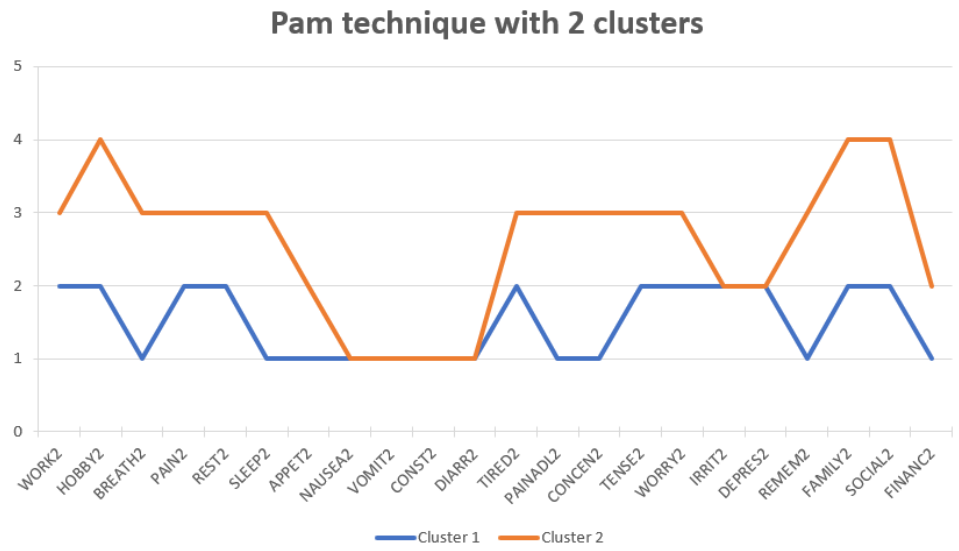

Figure 8. Plot Pam technique with 2 clusters

The description of each medoid is as follows:

Medoid 1: In general, these patients have a better quality of life than medoid 2. The quality of life variables are between 1 and 2. They tend not to suffer from “breath”, “appet”, “nausea”, “vomit”, “const”, “diarr”, “painadl”, “concen”, “remem” and “financ”.

Medoid 2: Overall, this group tends to be less healthy than medoid 2. They tend to have a high ranking of “hobby”, “family” and “social”. They tend not to have “nausea”, “vomit”, “const” and “diarr”.

Between medoid 1 and 2, the following variables overlap: “vomit”, “const”, “diarr”, “irrit” and “depress”.

GAUSSIAN MIXTURE MODEL USING Mclust

"Mclust is a popular R package for model-based clustering, classification, and density estimation based on finite Gaussian mixture modelling" (Scrucca, 2017). Mclust was used to identify the optimal number of groups. In Figure 9, the proposed best model using Mclust can be observed, where one cluster is composed of 200 patients and the other of 102 patients.

```
mc=Mclust(patient_clus)
mc
```

```
'Mclust' model object:
 best model: ellipsoidal, equal shape and orientation (VEE) with 2 components

Clustering table:
  1   2
200 102
```

Figure 9. Best model using Mclust and composition of each cluster

Moreover, the BIC trajectories were plotted using Mclust data. In Figure 10, it can be observed that "VEE" with 2 clusters has the maximum value of BIC and its trajectory does not show more than 2 components. However, "VEI" and "VII" with 3 clusters have a similar value of BIC as "VEE" with 2 components. "VEI" and "VII" do not show more than 3 components.

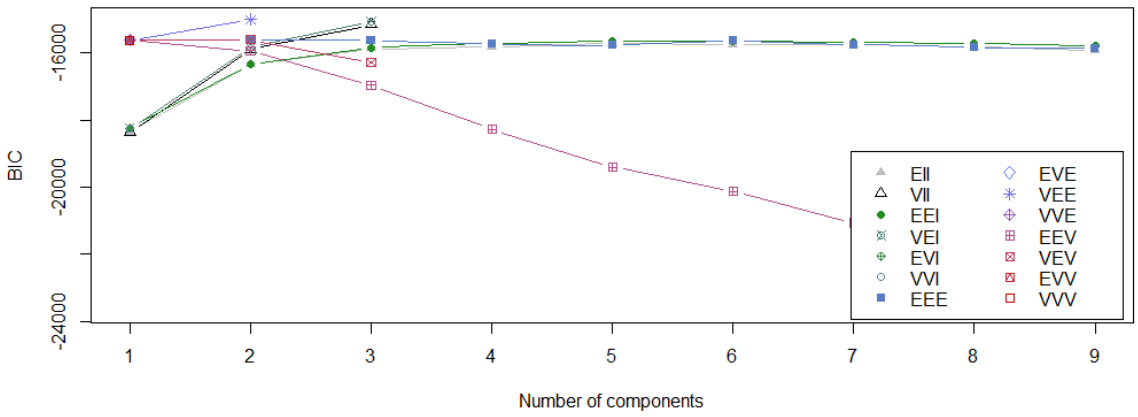

Figure 10. BIC trajectories

The classification and the density plot were created, but they were very difficult to interpret because there are many variables. The following table shows the composition of each cluster:

Quality of life variables	Cluster 1	Cluster 2
WORK2	3.055685	1.708474
HOBBY2	2.947742	1.535601
BREATH2	2.005655	1.247866
PAIN2	2.330239	1.338904
REST2	2.920724	1.773929
SLEEP2	2.421382	1.546441
APPET2	1.928112	1.144133
NAUSEA2	1.765557	1.094083
VOMIT2	1.335355	1.018098
CONST2	1.790240	1.164876

DIARR2	1.409273	1.078594
TIRED2	3.032207	1.797489
PAINADL2	2.125213	1.133046
CONCEN2	2.054439	1.167496
TENSE2	2.295838	1.399740
WORRY2	2.597805	1.649022
IRRIT2	2.047481	1.313717
DEPRES2	2.070993	1.346351
REMEM2	1.847759	1.309924
FAMILY2	2.357138	1.364550
SOCIAL2	2.819477	1.465518
FINANC2	1.714168	1.240688

Table 5. Mclust with 2 clusters

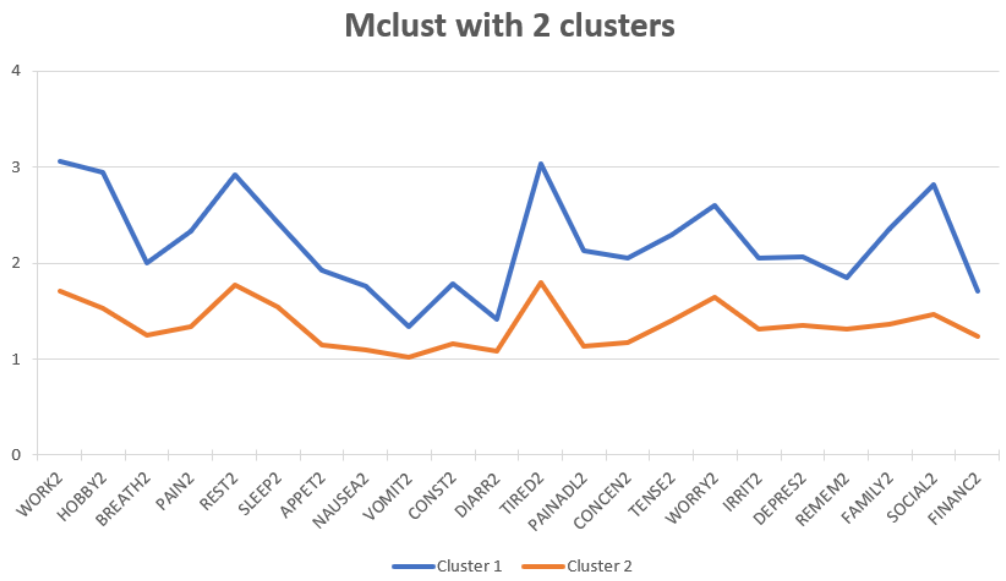

Figure 11. Plot Mclust with 2 clusters

In Figure 11, it can be seen that the line shape of each cluster is very similar to K-means with 2 clusters.

The characteristics of each cluster group are as follows:

Cluster 1: Overall, this group of patients have a less favourable quality of life than cluster 2. They have a high ranking of “work”, “hobby” and “social” and “tired”. They suffer less from “vomit” and “diarr”.

Cluster 2: In general, the quality of life of these patients is between 1 and 2, which means they do not have this quality (1) or have a little (2). These patients do not suffer from vomiting, and they have more of “tired”, “worry”, “rest”, “hobby” and “work”.

LATENT CLASS MODEL USING poLCA

poLCA package was used. The 22 quality of life variables were converted to binary. They were recoded to yes-> 2 and no-> 1. polCA was performed from 2 classes to 6 classes. 50 start sets were used, with 200 of iterations.

Number of classes	BIC
2	5921.20
3	5738.66
4	5726.57
5	5749.20
6	5817.56

Table 6. polCA – Number of classes and BIC

And the lowest value of BIC was using 4 classes. The value of BIC with 4 classes was 5726.57. The following plot shows the composition of each cluster:

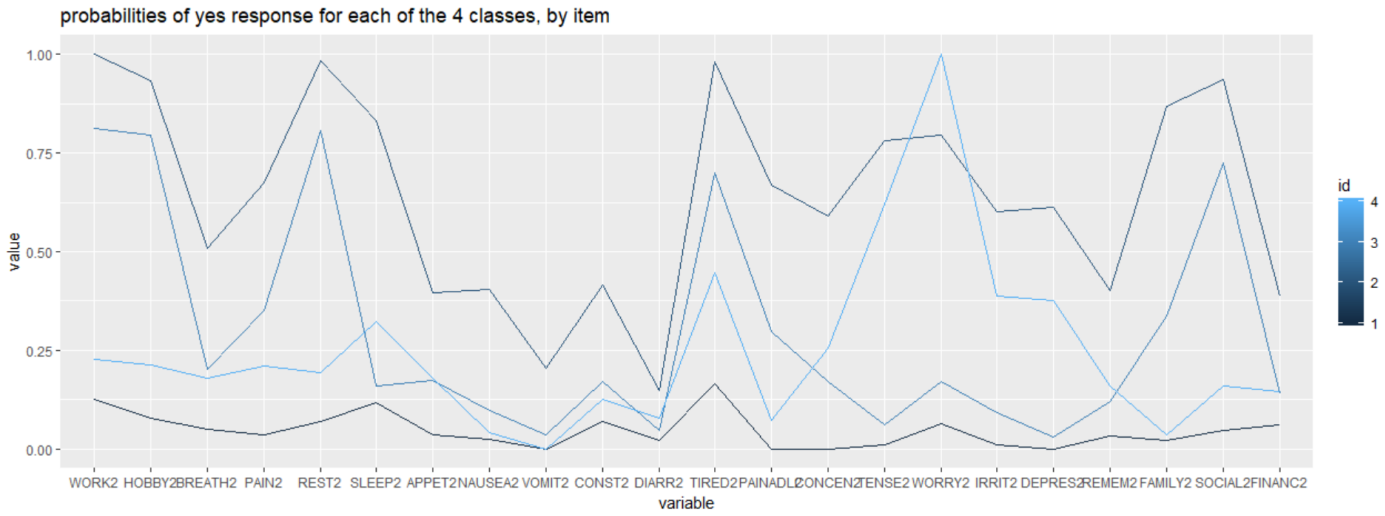

Figure 12. BIC trajectories

Class 1: Overall, this class has a better quality of life when it is compared against the other classes. The highest probability is to suffer from tiredness with 0.16, and this value is very low.

Class 2: In general, they have a high ranking of these qualities of life when it is compared against the other classes. These types of patients have a high probability of “work”, “hobby”, “painadl”, “rest”, “sleep”, “tired”, “family” and “social”. They have lowest probabilities of diarrhoea and vomiting.

Class 3: These kinds of patients have a high probability of working, social and hobby limitations, tired and rest. They have less probability of suffering from diarrhoea, tense, depression and vomiting. They have a better quality of life than class 2.

Class 4: They have a very high probability of being worried. This variable is the highest for this class. The variables with the lowest values are vomiting and family limitations. The value of the other variables are less than 0.5.

After this, the latent class model was extended to include the other variables (sex, age, live with), as covariates. For this, the missing values were removed from the original dataset. 85 rows were removed.

COVARIATES					
Gender		**livewith**		**Age**	
Number of classes	**BIC**	**Number of classes**	**BIC**	**Number of classes**	**BIC**
2	5798.43	2	5798.05	2	5798.41
3	5631.38	3	5627.86	3	5632.38
4	5621.28	4	5622.74	4	5624.75
5	5644.84	5	5652.73	5	5655.33

Table 7. polCA – Number of classes and BIC including the covariates

The lowest value of BIC was using 4 classes for all the covariates. However, the p-values are not significant for those covariates (Figure 13). So, the covariates gender, livewith and age are not relevant.

```
=========================================================
Fit for 4 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -0.12728      0.8037   -0.158     0.874
GENDER         -0.03701      0.4709   -0.079     0.937
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -3.70941     2.38037   -1.558     0.121
GENDER          1.40410     1.26945    1.106     0.270
=========================================================
4 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     0.57096     0.63370    0.901     0.369
GENDER         -0.13298     0.37503   -0.355     0.723
=========================================================
```

```
=========================================================
Fit for 4 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -2.09626     1.66343   -1.260     0.209
LIVEWITH        0.26710     0.67280    0.397     0.692
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -0.72022     0.59878   -1.203     0.230
LIVEWITH        0.08642     0.25224    0.343     0.732
=========================================================
4 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -0.26111     0.54501   -0.479     0.632
LIVEWITH       -0.03159     0.23215   -0.136     0.892
=========================================================
```

```
=========================================================
Fit for 4 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     1.07584     1.81295    0.593     0.554
AGE             0.00635     0.02964    0.214     0.831
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     1.40458     1.94576    0.722     0.471
AGE            -0.00786     0.03200   -0.246     0.806
=========================================================
4 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     0.13026     1.85890    0.070     0.944
AGE             0.01633     0.03019    0.541     0.589
=========================================================
number of observations: 295
```

Figure 13. Latent classes - pvalues

CONCLUSION

Different kinds of clustering methods were performed with the patient dataset to identify diverse groups of patients and to interpret the output. It can be observed that the different groups of patients using different methods sometimes are very similar. For example, regarding the 2 clusters obtained using k-means and Mclust, the plots were very similar. In other words, the characteristics of the clusters were very similar. However, the clusters created using the other methods were different wherein sometimes they have quality of life variables which overlap.

REFERENCES

Charrad, M. *et al.* (2014) '**NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set', *Journal of Statistical Software*, 61(6). doi: 10.18637/jss.v061.i06.

James, G. *et al.* (2007) *An Introduction to Statistical Learning with Applications in R*, *Performance Evaluation*. doi: 10.1016/j.peva.2007.06.006.

Scrucca, L. (2017) A quick tour of mclust.

https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

STHDA http://www.sthda.com/english/articles/27-partitioning-clustering-essentials/88-k-medoids-essentials/

APPENDIX

```
#K-means
library(dplyr)
library(scatterplot3d)
library(cluster)

setwd("D:/Lancaster University/Modules/Statistical learning/coursework task-20180209")
patient <- read.table("patient.dat", header = TRUE)
#View(patient)
dim(patient)#381  25
names(patient)
patient_clus<- select(patient, WORK2, HOBBY2, BREATH2, PAIN2, REST2, SLEEP2, APPET2, NAUSEA2,
VOMIT2, CONST2, DIARR2, TIRED2, PAINADL2, CONCEN2, TENSE2, WORRY2, IRRIT2, DEPRES2, REMEM2,
FAMILY2, SOCIAL2, FINANC2)
names(patient_clus)
patient_clus<- na.omit(patient_clus)
dim(patient_clus)#302 22
head(patient_clus)

#this code was used to get 4 and 2 clusters
outwss = numeric(200)
outkmlist=array(list(NULL), 200)
for (i in 1:200)
{                                    # kmeans 200 times with
outkm = kmeans(patient_clus, 4, nstart=200, iter.max=200) # different starting values
outkmlist[[i]]=outkm                 # store the kmeans object
outwss[i]=outkm$tot.withinss        # and the within ss
}
outwss

j=which.min(outwss)       # what is the index of the minimum wss
            # and the result of the fit
j#3

outwss[j]                 # print out the wss value - minimum total within sums of squares


outkmlist[[j]]

hist(outwss)               # see how variable the results are
table(outwss)

k.max=10

wss <- sapply(1:k.max,
              function(k){
kmeans(patient_clus, k, nstart=200,iter.max = 20 )$tot.withinss

})
wss

plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")

#The plot looks very smooth, with no sudden changes of direction (elbow). It is hard to judge
how many clusters
#from this plot. We try the NbClust function

library(NbClust)
nb <- NbClust(patient_clus, diss=NULL, distance = "euclidean",
              min.nc=2, max.nc=15, method = "kmeans",
              index = "all", alphaBeale = 0.1)
```

```
#pam
out2<-pam(patient_clus,4)
out2
out2$id.med
out2$isolation
plot(out2)# as before produces a silhouette plot.
asw<-numeric(20)
for (k in 2:20)
       asw[k] <- pam(patient_clus, k) $ silinfo $ avg.width
k.best <- which.max(asw)
k.best#it is 2

plot(1:20, asw, type= "h",
main = "pam() clustering assessment",
xlab= "k  (# clusters)",
ylab = "av silhouette width") #silhouette-optimal number of clusters:2

#I tried with 2
out3<-pam(patient_clus,2)
out3
out3$id.med
out3$isolation

plot(out3)# as before produces a silhouette plot.

#mclust
library(mclust)
mc=Mclust(patient_clus)
mc
summary(mc)
mc$parameters$mean#?
mc$parameters$variance$sigma#?
plot(mc, what="BIC")
patient3VEE= Mclust(patient_clus,G=2,c("VEE"))
patient3VEE
summary(patient3VEE)
plot(mc, what="classification")
plot(mc, what="density")
mc
mc$classification
mc$G
mc$parameters

#poLCA
patient_clusbi<- patient_clus
#WORK2
patient_clusbi$WORK2[patient_clusbi$WORK2==1 | patient_clusbi$WORK2==2] <- "1"
patient_clusbi$WORK2[patient_clusbi$WORK2==3 | patient_clusbi$WORK2==4] <- "2"
#HOBBY2
patient_clusbi$HOBBY2[patient_clusbi$HOBBY2==1 | patient_clusbi$HOBBY2==2] <- "1"
patient_clusbi$HOBBY2[patient_clusbi$HOBBY2==3 | patient_clusbi$HOBBY2==4] <- "2"
#BREATH2
patient_clusbi$BREATH2[patient_clusbi$BREATH2==1 | patient_clusbi$BREATH2==2] <- "1"
patient_clusbi$BREATH2[patient_clusbi$BREATH2==3 | patient_clusbi$BREATH2==4] <- "2"
#PAIN2
patient_clusbi$PAIN2[patient_clusbi$PAIN2==1 | patient_clusbi$PAIN2==2] <- "1"
patient_clusbi$PAIN2[patient_clusbi$PAIN2==3 | patient_clusbi$PAIN2==4] <- "2"
#REST2
patient_clusbi$REST2[patient_clusbi$REST2==1 | patient_clusbi$REST2==2] <- "1"
patient_clusbi$REST2[patient_clusbi$REST2==3 | patient_clusbi$REST2==4] <- "2"
#SLEEP2
patient_clusbi$SLEEP2[patient_clusbi$SLEEP2==1 | patient_clusbi$SLEEP2==2] <- "1"
patient_clusbi$SLEEP2[patient_clusbi$SLEEP2==3 | patient_clusbi$SLEEP2==4] <- "2"
#APPET2
patient_clusbi$APPET2[patient_clusbi$APPET2==1 | patient_clusbi$APPET2==2] <- "1"
patient_clusbi$APPET2[patient_clusbi$APPET2==3 | patient_clusbi$APPET2==4] <- "2"
#NAUSEA2
patient_clusbi$NAUSEA2[patient_clusbi$NAUSEA2==1 | patient_clusbi$NAUSEA2==2] <- "1"
patient_clusbi$NAUSEA2[patient_clusbi$NAUSEA2==3 | patient_clusbi$NAUSEA2==4] <- "2"
#VOMIT2
patient_clusbi$VOMIT2[patient_clusbi$VOMIT2==1 | patient_clusbi$VOMIT2==2] <- "1"
```

```
patient_clusbi$VOMIT2[patient_clusbi$VOMIT2==3 | patient_clusbi$VOMIT2==4] <- "2"
#CONST2
patient_clusbi$CONST2[patient_clusbi$CONST2==1 | patient_clusbi$CONST2==2] <- "1"
patient_clusbi$CONST2[patient_clusbi$CONST2==3 | patient_clusbi$CONST2==4] <- "2"
#DIARR2
patient_clusbi$DIARR2[patient_clusbi$DIARR2==1 | patient_clusbi$DIARR2==2] <- "1"
patient_clusbi$DIARR2[patient_clusbi$DIARR2==3 | patient_clusbi$DIARR2==4] <- "2"
#TIRED2
patient_clusbi$TIRED2[patient_clusbi$TIRED2==1 | patient_clusbi$TIRED2==2] <- "1"
patient_clusbi$TIRED2[patient_clusbi$TIRED2==3 | patient_clusbi$TIRED2==4] <- "2"
#PAINADL2
patient_clusbi$PAINADL2[patient_clusbi$PAINADL2==1 | patient_clusbi$PAINADL2==2] <- "1"
patient_clusbi$PAINADL2[patient_clusbi$PAINADL2==3 | patient_clusbi$PAINADL2==4] <- "2"
#CONCEN2
patient_clusbi$CONCEN2[patient_clusbi$CONCEN2==1 | patient_clusbi$CONCEN2==2] <- "1"
patient_clusbi$CONCEN2[patient_clusbi$CONCEN2==3 | patient_clusbi$CONCEN2==4] <- "2"
#TENSE2
patient_clusbi$TENSE2[patient_clusbi$TENSE2==1 | patient_clusbi$TENSE2==2] <- "1"
patient_clusbi$TENSE2[patient_clusbi$TENSE2==3 | patient_clusbi$TENSE2==4] <- "2"
#WORRY2
patient_clusbi$WORRY2[patient_clusbi$WORRY2==1 | patient_clusbi$WORRY2==2] <- "1"
patient_clusbi$WORRY2[patient_clusbi$WORRY2==3 | patient_clusbi$WORRY2==4] <- "2"
#IRRIT2
patient_clusbi$IRRIT2[patient_clusbi$IRRIT2==1 | patient_clusbi$IRRIT2==2] <- "1"
patient_clusbi$IRRIT2[patient_clusbi$IRRIT2==3 | patient_clusbi$IRRIT2==4] <- "2"
#DEPRES2
patient_clusbi$DEPRES2[patient_clusbi$DEPRES2==1 | patient_clusbi$DEPRES2==2] <- "1"
patient_clusbi$DEPRES2[patient_clusbi$DEPRES2==3 | patient_clusbi$DEPRES2==4] <- "2"
#REMEM2
patient_clusbi$REMEM2[patient_clusbi$REMEM2==1 | patient_clusbi$REMEM2==2] <- "1"
patient_clusbi$REMEM2[patient_clusbi$REMEM2==3 | patient_clusbi$REMEM2==4] <- "2"
#FAMILY2
patient_clusbi$FAMILY2[patient_clusbi$FAMILY2==1 | patient_clusbi$FAMILY2==2] <- "1"
patient_clusbi$FAMILY2[patient_clusbi$FAMILY2==3 | patient_clusbi$FAMILY2==4] <- "2"
#SOCIAL2
patient_clusbi$SOCIAL2[patient_clusbi$SOCIAL2==1 | patient_clusbi$SOCIAL2==2] <- "1"
patient_clusbi$SOCIAL2[patient_clusbi$SOCIAL2==3 | patient_clusbi$SOCIAL2==4] <- "2"
#FINANC2
patient_clusbi$FINANC2[patient_clusbi$FINANC2==1 | patient_clusbi$FINANC2==2] <- "1"
patient_clusbi$FINANC2[patient_clusbi$FINANC2==3 | patient_clusbi$FINANC2==4] <- "2"
head(patient_clusbi)

#WORK2
patient_clusbi$WORK2<- as.numeric(patient_clusbi$WORK2)
#HOBBY2
patient_clusbi$HOBBY2<- as.numeric(patient_clusbi$HOBBY2)
#BREATH2
patient_clusbi$BREATH2<- as.numeric(patient_clusbi$BREATH2)
#PAIN2
patient_clusbi$PAIN2<- as.numeric(patient_clusbi$PAIN2)
#REST2
patient_clusbi$REST2<- as.numeric(patient_clusbi$REST2)
#SLEEP2
patient_clusbi$SLEEP2<- as.numeric(patient_clusbi$SLEEP2)
#APPET2
patient_clusbi$APPET2<- as.numeric(patient_clusbi$APPET2)
#NAUSEA2
patient_clusbi$NAUSEA2<- as.numeric(patient_clusbi$NAUSEA2)
#VOMIT2
patient_clusbi$VOMIT2<- as.numeric(patient_clusbi$VOMIT2)
#CONST2
patient_clusbi$CONST2<- as.numeric(patient_clusbi$CONST2)
#DIARR2
patient_clusbi$DIARR2<- as.numeric(patient_clusbi$DIARR2)
#TIRED2
patient_clusbi$TIRED2<- as.numeric(patient_clusbi$TIRED2)
#PAINADL2
patient_clusbi$PAINADL2<- as.numeric(patient_clusbi$PAINADL2)
#CONCEN2
patient_clusbi$CONCEN2<- as.numeric(patient_clusbi$CONCEN2)
#TENSE2
```

```
patient_clusbi$TENSE2<- as.numeric(patient_clusbi$TENSE2)
#WORRY2
patient_clusbi$WORRY2<- as.numeric(patient_clusbi$WORRY2)
#IRRIT2
patient_clusbi$IRRIT2<- as.numeric(patient_clusbi$IRRIT2)
#DEPRES2
patient_clusbi$DEPRES2<- as.numeric(patient_clusbi$DEPRES2)
#REMEM2
patient_clusbi$REMEM2<- as.numeric(patient_clusbi$REMEM2)
#FAMILY2
patient_clusbi$FAMILY2<- as.numeric(patient_clusbi$FAMILY2)
#SOCIAL2
patient_clusbi$SOCIAL2<- as.numeric(patient_clusbi$SOCIAL2)
#FINANC2
patient_clusbi$FINANC2<- as.numeric(patient_clusbi$FINANC2)

set.seed(42)
pbind=cbind(WORK2,HOBBY2,BREATH2,PAIN2,REST2,SLEEP2,APPET2,NAUSEA2,VOMIT2,CONST2,DIARR2,TIRE
D2,PAINADL2,CONCEN2,TENSE2,WORRY2,IRRIT2,DEPRES2,REMEM2,FAMILY2,SOCIAL2,FINANC2)~1
pbind
k2=poLCA(pbind, data=patient_clusbi, nclass=2,  maxiter=200, nrep =50,verbose=F)
k2$bic#5921.20438800557
k3=poLCA(pbind, data=patient_clusbi, nclass=3, maxiter=200, nrep =50,verbose=F)
k3$bic#5738.66006190546
k4=poLCA(pbind, data=patient_clusbi, nclass=4, maxiter=200, nrep =50,verbose=F)
k4$bic #5726.57282290333
k5=poLCA(pbind, data=patient_clusbi, nclass=5, maxiter=200, nrep =50,verbose=F)
k5$bic#5749.20947996148
k6=poLCA(pbind, data=patient_clusbi, nclass=6, maxiter=200, nrep =50,verbose=F)
k6$bic#5817.56
k7=poLCA(pbind, data=patient_clusbi, nclass=7, maxiter=200, nrep =50,verbose=F)
k7$bic#5891.52232940461
k8=poLCA(pbind, data=patient_clusbi, nclass=8, maxiter=200, nrep =50,verbose=F)
k8$bic#5865.80201704499
k9=poLCA(pbind, data=patient_clusbi, nclass=9, maxiter=200, nrep =50,verbose=F)
k9$bic
k4
library(GGally)
cb=cbind(k4$probs[[1]][,2], k4$probs[[2]][,2],k4$probs[[3]][,2],
         k4$probs[[4]][,2], k4$probs[[5]][,2], k4$probs[[6]][,2],
         k4$probs[[7]][,2], k4$probs[[8]][,2], k4$probs[[9]][,2],
         k4$probs[[10]][,2],k4$probs[[11]][,2], k4$probs[[12]][,2],
         k4$probs[[13]][,2],k4$probs[[14]][,2], k4$probs[[15]][,2],
         k4$probs[[16]][,2],k4$probs[[17]][,2], k4$probs[[18]][,2],
         k4$probs[[19]][,2],k4$probs[[20]][,2], k4$probs[[21]][,2],
         k4$probs[[22]][,2], factor(c(1,2,3,4)))
colnames(cb)=c("WORK2",
"HOBBY2","BREATH2","PAIN2","REST2","SLEEP2","APPET2","NAUSEA2","VOMIT2","CONST2","DIARR2","T
IRED2","PAINADL2","CONCEN2","TENSE2","WORRY2","IRRIT2","DEPRES2","REMEM2","FAMILY2","SOCIAL2
","FINANC2","id")
ggparcoord(cb,columns=1:22, groupColumn="id", scale = 'globalminmax', title="probabilities of
yes response for each of the 4 classes, by item")
cb
k4$P
patient_clus2<- na.omit(patient)
dim(patient_clus2)
patient_clusbi2<- patient_clus2
#WORK2
patient_clusbi2$WORK2[patient_clusbi2$WORK2==1 | patient_clusbi2$WORK2==2] <- "1"
patient_clusbi2$WORK2[patient_clusbi2$WORK2==3 | patient_clusbi2$WORK2==4] <- "2"
#HOBBY2
patient_clusbi2$HOBBY2[patient_clusbi2$HOBBY2==1 | patient_clusbi2$HOBBY2==2] <- "1"
patient_clusbi2$HOBBY2[patient_clusbi2$HOBBY2==3 | patient_clusbi2$HOBBY2==4] <- "2"
#BREATH2
patient_clusbi2$BREATH2[patient_clusbi2$BREATH2==1 | patient_clusbi2$BREATH2==2] <- "1"
patient_clusbi2$BREATH2[patient_clusbi2$BREATH2==3 | patient_clusbi2$BREATH2==4] <- "2"
#PAIN2
patient_clusbi2$PAIN2[patient_clusbi2$PAIN2==1 | patient_clusbi2$PAIN2==2] <- "1"
patient_clusbi2$PAIN2[patient_clusbi2$PAIN2==3 | patient_clusbi2$PAIN2==4] <- "2"
#REST2
patient_clusbi2$REST2[patient_clusbi2$REST2==1 | patient_clusbi2$REST2==2] <- "1"
```

```
patient_clusbi2$REST2[patient_clusbi2$REST2==3 | patient_clusbi2$REST2==4] <- "2"
#SLEEP2
patient_clusbi2$SLEEP2[patient_clusbi2$SLEEP2==1 | patient_clusbi2$SLEEP2==2] <- "1"
patient_clusbi2$SLEEP2[patient_clusbi2$SLEEP2==3 | patient_clusbi2$SLEEP2==4] <- "2"
#APPET2
patient_clusbi2$APPET2[patient_clusbi2$APPET2==1 | patient_clusbi2$APPET2==2] <- "1"
patient_clusbi2$APPET2[patient_clusbi2$APPET2==3 | patient_clusbi2$APPET2==4] <- "2"
#NAUSEA2
patient_clusbi2$NAUSEA2[patient_clusbi2$NAUSEA2==1 | patient_clusbi2$NAUSEA2==2] <- "1"
patient_clusbi2$NAUSEA2[patient_clusbi2$NAUSEA2==3 | patient_clusbi2$NAUSEA2==4] <- "2"
#VOMIT2
patient_clusbi2$VOMIT2[patient_clusbi2$VOMIT2==1 | patient_clusbi2$VOMIT2==2] <- "1"
patient_clusbi2$VOMIT2[patient_clusbi2$VOMIT2==3 | patient_clusbi2$VOMIT2==4] <- "2"
#CONST2
patient_clusbi2$CONST2[patient_clusbi2$CONST2==1 | patient_clusbi2$CONST2==2] <- "1"
patient_clusbi2$CONST2[patient_clusbi2$CONST2==3 | patient_clusbi2$CONST2==4] <- "2"
#DIARR2
patient_clusbi2$DIARR2[patient_clusbi2$DIARR2==1 | patient_clusbi2$DIARR2==2] <- "1"
patient_clusbi2$DIARR2[patient_clusbi2$DIARR2==3 | patient_clusbi2$DIARR2==4] <- "2"
#TIRED2
patient_clusbi2$TIRED2[patient_clusbi2$TIRED2==1 | patient_clusbi2$TIRED2==2] <- "1"
patient_clusbi2$TIRED2[patient_clusbi2$TIRED2==3 | patient_clusbi2$TIRED2==4] <- "2"
#PAINADL2
patient_clusbi2$PAINADL2[patient_clusbi2$PAINADL2==1 | patient_clusbi2$PAINADL2==2] <- "1"
patient_clusbi2$PAINADL2[patient_clusbi2$PAINADL2==3 | patient_clusbi2$PAINADL2==4] <- "2"
#CONCEN2
patient_clusbi2$CONCEN2[patient_clusbi2$CONCEN2==1 | patient_clusbi2$CONCEN2==2] <- "1"
patient_clusbi2$CONCEN2[patient_clusbi2$CONCEN2==3 | patient_clusbi2$CONCEN2==4] <- "2"
#TENSE2
patient_clusbi2$TENSE2[patient_clusbi2$TENSE2==1 | patient_clusbi2$TENSE2==2] <- "1"
patient_clusbi2$TENSE2[patient_clusbi2$TENSE2==3 | patient_clusbi2$TENSE2==4] <- "2"
#WORRY2
patient_clusbi2$WORRY2[patient_clusbi2$WORRY2==1 | patient_clusbi2$WORRY2==2] <- "1"
patient_clusbi2$WORRY2[patient_clusbi2$WORRY2==3 | patient_clusbi2$WORRY2==4] <- "2"
#IRRIT2
patient_clusbi2$IRRIT2[patient_clusbi2$IRRIT2==1 | patient_clusbi2$IRRIT2==2] <- "1"
patient_clusbi2$IRRIT2[patient_clusbi2$IRRIT2==3 | patient_clusbi2$IRRIT2==4] <- "2"
#DEPRES2
patient_clusbi2$DEPRES2[patient_clusbi2$DEPRES2==1 | patient_clusbi2$DEPRES2==2] <- "1"
patient_clusbi2$DEPRES2[patient_clusbi2$DEPRES2==3 | patient_clusbi2$DEPRES2==4] <- "2"
#REMEM2
patient_clusbi2$REMEM2[patient_clusbi2$REMEM2==1 | patient_clusbi2$REMEM2==2] <- "1"
patient_clusbi2$REMEM2[patient_clusbi2$REMEM2==3 | patient_clusbi2$REMEM2==4] <- "2"
#FAMILY2
patient_clusbi2$FAMILY2[patient_clusbi2$FAMILY2==1 | patient_clusbi2$FAMILY2==2] <- "1"
patient_clusbi2$FAMILY2[patient_clusbi2$FAMILY2==3 | patient_clusbi2$FAMILY2==4] <- "2"
#SOCIAL2
patient_clusbi2$SOCIAL2[patient_clusbi2$SOCIAL2==1 | patient_clusbi2$SOCIAL2==2] <- "1"
patient_clusbi2$SOCIAL2[patient_clusbi2$SOCIAL2==3 | patient_clusbi2$SOCIAL2==4] <- "2"
#FINANC2
patient_clusbi2$FINANC2[patient_clusbi2$FINANC2==1 | patient_clusbi2$FINANC2==2] <- "1"
patient_clusbi2$FINANC2[patient_clusbi2$FINANC2==3 | patient_clusbi2$FINANC2==4] <- "2"
head(patient_clusbi)
#WORK2
patient_clusbi2$WORK2<- as.numeric(patient_clusbi2$WORK2)
#HOBBY2
patient_clusbi2$HOBBY2<- as.numeric(patient_clusbi2$HOBBY2)
#BREATH2
patient_clusbi2$BREATH2<- as.numeric(patient_clusbi2$BREATH2)
#PAIN2
patient_clusbi2$PAIN2<- as.numeric(patient_clusbi2$PAIN2)
#REST2
patient_clusbi2$REST2<- as.numeric(patient_clusbi2$REST2)
#SLEEP2
patient_clusbi2$SLEEP2<- as.numeric(patient_clusbi2$SLEEP2)
#APPET2
patient_clusbi2$APPET2<- as.numeric(patient_clusbi2$APPET2)
#NAUSEA2
patient_clusbi2$NAUSEA2<- as.numeric(patient_clusbi2$NAUSEA2)
#VOMIT2
patient_clusbi2$VOMIT2<- as.numeric(patient_clusbi2$VOMIT2)
```

```
#CONST2
patient_clusbi2$CONST2<- as.numeric(patient_clusbi2$CONST2)
#DIARR2
patient_clusbi2$DIARR2<- as.numeric(patient_clusbi2$DIARR2)
#TIRED2
patient_clusbi2$TIRED2<- as.numeric(patient_clusbi2$TIRED2)
#PAINADL2
patient_clusbi2$PAINADL2<- as.numeric(patient_clusbi2$PAINADL2)
#CONCEN2
patient_clusbi2$CONCEN2<- as.numeric(patient_clusbi2$CONCEN2)
#TENSE2
patient_clusbi2$TENSE2<- as.numeric(patient_clusbi2$TENSE2)
#WORRY2
patient_clusbi2$WORRY2<- as.numeric(patient_clusbi2$WORRY2)
#IRRIT2
patient_clusbi2$IRRIT2<- as.numeric(patient_clusbi2$IRRIT2)
#DEPRES2
patient_clusbi2$DEPRES2<- as.numeric(patient_clusbi2$DEPRES2)
#REMEM2
patient_clusbi2$REMEM2<- as.numeric(patient_clusbi2$REMEM2)
#FAMILY2
patient_clusbi2$FAMILY2<- as.numeric(patient_clusbi2$FAMILY2)
#SOCIAL2
patient_clusbi2$SOCIAL2<- as.numeric(patient_clusbi2$SOCIAL2)
#FINANC2
patient_clusbi2$FINANC2<- as.numeric(patient_clusbi2$FINANC2)
dim(patient_clusbi2)
names(patient_clusbi2)
pbind_gender=cbind(WORK2,HOBBY2,BREATH2,PAIN2,REST2,SLEEP2,APPET2,NAUSEA2,VOMIT2,CONST2,DIAR
R2,TIRED2,PAINADL2,CONCEN2,TENSE2,WORRY2,IRRIT2,DEPRES2,REMEM2,FAMILY2,SOCIAL2,FINANC2)~GEND
ER
pbind_gender
kg2=poLCA(pbind_gender, data=patient_clusbi2, nclass=2, maxiter=200, nrep =50,verbose=F)
kg2$bic#5798.43022801817
kg3=poLCA(pbind_gender, data=patient_clusbi2, nclass=3, maxiter=200, nrep =50,verbose=F)
kg3$bic #5631.38070151742
kg4=poLCA(pbind_gender, data=patient_clusbi2, nclass=4, maxiter=200, nrep =50,verbose=F)
kg4$bic#5621.28956775403---------------->best
kg5=poLCA(pbind_gender, data=patient_clusbi2, nclass=5, maxiter=200, nrep =50,verbose=F)
kg5$bic

kg6=poLCA(pbind_gender, data=patient_clusbi2, nclass=6, maxiter=200, nrep =50,verbose=F)
kg6$bic
kg4
pbind_age=cbind(WORK2,HOBBY2,BREATH2,PAIN2,REST2,SLEEP2,APPET2,NAUSEA2,VOMIT2,CONST2,DIARR2,
TIRED2,PAINADL2,CONCEN2,TENSE2,WORRY2,IRRIT2,DEPRES2,REMEM2,FAMILY2,SOCIAL2,FINANC2)~AGE
pbind_age
ka2=poLCA(pbind_age, data=patient_clusbi2, nclass=2, maxiter=200, nrep =50,verbose=F)
ka2$bic
ka3=poLCA(pbind_age, data=patient_clusbi2, nclass=3, maxiter=200, nrep =50,verbose=F)
ka3$bic
ka4=poLCA(pbind_age, data=patient_clusbi2, nclass=4, maxiter=200, nrep =50,verbose=F)
ka4$bic #5622.74341493913----->best
ka4
ka5=poLCA(pbind_age, data=patient_clusbi2, nclass=5, maxiter=200, nrep =50,verbose=F)
ka5$bic
ka6=poLCA(pbind_age, data=patient_clusbi2, nclass=6, maxiter=200, nrep =50,verbose=F)
ka6$bic
pbind_live=cbind(WORK2,HOBBY2,BREATH2,PAIN2,REST2,SLEEP2,APPET2,NAUSEA2,VOMIT2,CONST2,DIARR2
,TIRED2,PAINADL2,CONCEN2,TENSE2,WORRY2,IRRIT2,DEPRES2,REMEM2,FAMILY2,SOCIAL2,FINANC2)~LIVEWI
TH
pbind_live
kl2=poLCA(pbind_live, data=patient_clusbi2, nclass=2, maxiter=200, nrep =50,verbose=F)
kl2$bic
kl3=poLCA(pbind_live, data=patient_clusbi2, nclass=3, maxiter=200, nrep =50,verbose=F)
kl3$bic
kl4=poLCA(pbind_live, data=patient_clusbi2, nclass=4, maxiter=200, nrep =50,verbose=F)
kl4$bic#5624.7586092205----best
kl4
kl5=poLCA(pbind_live, data=patient_clusbi2, nclass=5, maxiter=200, nrep =50,verbose=F)
kl5$bic
```