Similarity Algorithms: Clustering File

David Teran

Created by David Teran on March 20, 2023

Clustering

This R Notebook will be using a dataset with more than 10K rows of data and will implement two clustering algorithms, grouping data into similar groups. The two clustering methods implemented will be k-means and hierarchical clustering, to which a model-based approach for the two methods will be taken for clustering the dataset.

The dataset used will be the results of an airline passenger customer satisfaction survey, where customers rate the airline based on several different factors. This dataset can be found in Kaggle over at:https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv.

The first clustering algorithm to be used will be the k-means clustering, which identifies the centers and groups data based on how close it is to a center. We will be using 'k' random observations to be the centroids and group the observations, repeating until convergence.

First, reading in the dataset to be used for clustering algorithms and prepare it by cleaning it from any NA or 0 values.

```
#Read the dataset in
current_path = rstudioapi::getActiveDocumentContext()$path
setwd(dirname(current_path ))
dataClust1 <- read.csv("train.csv")
dataClust2 <- read.csv("test.csv")
dataClust <- rbind(dataClust1, dataClust2)

#Drop unneeded Columns
dataClust <- dataClust[,-1:-2]

#Preparing Data
head(dataClust)
```

```
##   Gender     Customer.Type Age  Type.of.Travel    Class Flight.Distance
## 1   Male    Loyal Customer  13 Personal Travel Eco Plus             460
## 2   Male disloyal Customer  25 Business travel Business             235
## 3 Female    Loyal Customer  26 Business travel Business            1142
## 4 Female    Loyal Customer  25 Business travel Business             562
## 5   Male    Loyal Customer  61 Business travel Business             214
## 6 Female    Loyal Customer  26 Personal Travel      Eco            1180
##   Inflight.wifi.service Departure.Arrival.time.convenient
## 1                     3                                4
## 2                     3                                2
## 3                     2                                2
## 4                     2                                5
## 5                     3                                3
```

```
## 6                     3                                4
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1                      3             1              5               3
## 2                      3             3              1               3
## 3                      2             2              5               5
## 4                      5             5              2               2
## 5                      3             3              4               5
## 6                      2             1              1               2
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1            5                      5                4                3
## 2            1                      1                1                5
## 3            5                      5                4                3
## 4            2                      2                2                5
## 5            5                      3                3                4
## 6            1                      1                3                4
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1                4               4                5           5
## 2                3               1                4           1
## 3                4               4                4           5
## 4                3               1                4           2
## 5                4               3                3           3
## 6                4               4                4           1
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes            satisfaction
## 1                         25                       18 neutral or dissatisfied
## 2                          1                        6 neutral or dissatisfied
## 3                          0                        0               satisfied
## 4                         11                        9 neutral or dissatisfied
## 5                          0                        0               satisfied
## 6                          0                        0 neutral or dissatisfied
```

```
data(dataClust)
```

```
## Warning in data(dataClust): data set 'dataClust' not found
```

```
names(dataClust)
```

```
##  [1] "Gender"                          "Customer.Type"
##  [3] "Age"                             "Type.of.Travel"
##  [5] "Class"                           "Flight.Distance"
##  [7] "Inflight.wifi.service"           "Departure.Arrival.time.convenient"
##  [9] "Ease.of.Online.booking"          "Gate.location"
## [11] "Food.and.drink"                  "Online.boarding"
## [13] "Seat.comfort"                    "Inflight.entertainment"
## [15] "On.board.service"                "Leg.room.service"
## [17] "Baggage.handling"                "Checkin.service"
## [19] "Inflight.service"                "Cleanliness"
## [21] "Departure.Delay.in.Minutes"      "Arrival.Delay.in.Minutes"
## [23] "satisfaction"
```

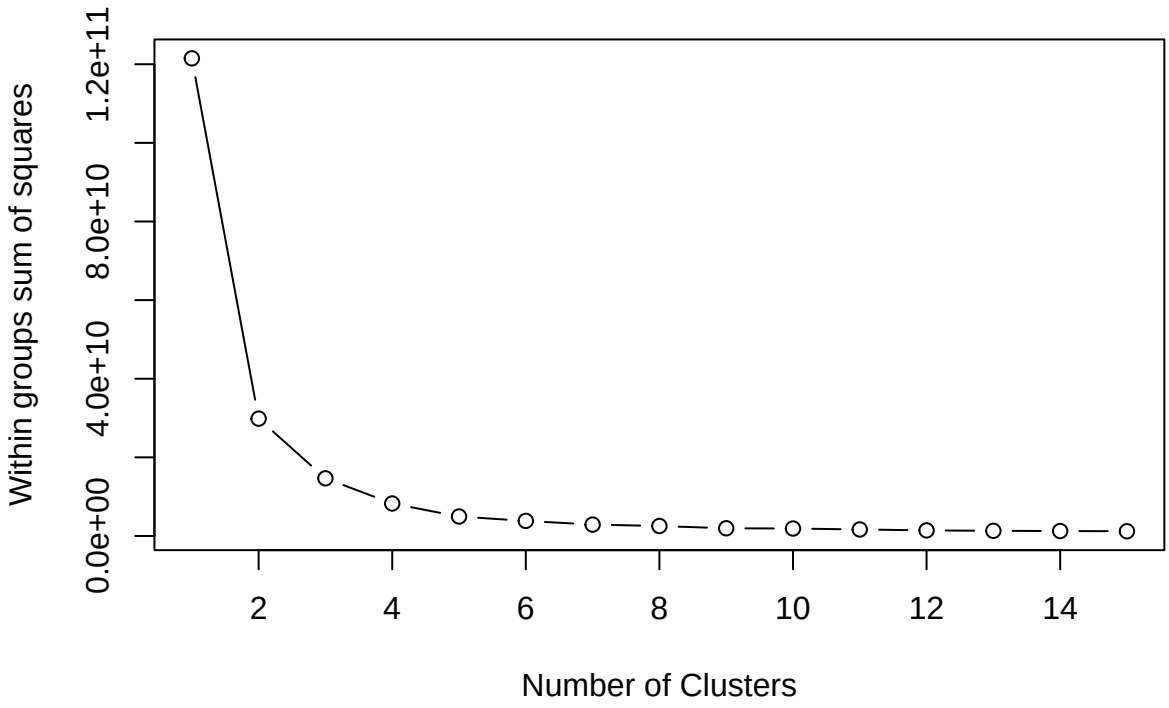
```
#Convert data columns to numeric data
dataClust$Gender <- ifelse(dataClust$Gender=="Female", 1, 0)
dataClust$Customer.Type <- ifelse(dataClust$Customer.Type=="Local Customer", 1, 0)
```

```
dataClust$Type.of.Travel <- ifelse(dataClust$Type.of.Travel=="Business travel", 1, 0)
dataClust$Class[dataClust$Class == "Eco"] <- 0
dataClust$Class[dataClust$Class == "Eco Plus"] <- 1
dataClust$Class[dataClust$Class == "Business"] <- 2
dataClust$satisfaction<-as.factor(dataClust$satisfaction)

#Remove 0's
dataClust <- na.omit(dataClust) #Clear missing data
dataClust <- dataClust[!(is.na(dataClust$Arrival.Delay.in.Minutes)),]
#Normalizing Data
dataClust$Class <- as.numeric(dataClust$Class)
dataClust$satisfaction <- as.numeric(dataClust$satisfaction)

dataClust <- dataClust[!(dataClust$Gate.location==0),]
dataClust <- dataClust[!(dataClust$Food.and.drink==0),]
dataClust <- dataClust[!(dataClust$Online.boarding==0),]
dataClust <- dataClust[!(dataClust$Seat.comfort==0),]
dataClust <- dataClust[!(dataClust$Inflight.entertainment==0),]
dataClust <- dataClust[!(dataClust$On.board.service==0),]
dataClust <- dataClust[!(dataClust$Leg.room.service==0),]
dataClust <- dataClust[!(dataClust$Checkin.service==0),]
dataClust <- dataClust[!(dataClust$Inflight.service==0),]
dataClust <- dataClust[!(dataClust$Cleanliness==0),]
dataClust <- dataClust[!(dataClust$Departure.Arrival.time.convenient==0),]
```

##K-means Clustering

The optimal number of k clusters must be determined in order to obtain a good result. To find 'k', the within sum of squares is used to determine the number of clusters to set for obtaining the optimal model.The withinss is then plotted onto a graph to see what value is k to produce the most optimal model for k-means.

```
#Determine optimal value for k using within sum of squares
wsplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data,centers=i)$withinss)
  }
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")
}
wsplot(dataClust)
```



With the within sum of squares, the number of clusters that will provide the best optimal model is where the "elbow" of the graph is. The drop from k=1 to k=2 is large compared to k=3 and onwards. This means that the number of clusters for k-means is 2. We can also use NbClust to find the best value for k

```
# library(NbClust)
# set.seed(1234)
# nc <- NbClust(test, min.nc=2, max.nc=15, method="kmeans")
# t <- table(nc$Best.n[1,])
# t
# barplot(t, xlab="Number of Clusters", ylab = "Criteria")
#
```

With k now known, the k-means clustering algorithm can be performed.The data will show the cluster sizes, cluster means, cluster vectors, within sum of squares, and available components to use for analysis.

```
set.seed(1234)
testClust <- kmeans(dataClust, 2, nstart=25)
```

Once the optimal model has been selected, then a model analysis can be done, using the size and the centers of the k-means to cross tabulate to see if we can find an agreement. Then it is possible to quantify the agreement between the value randomly assigned and the cluster by using the Rand index. The closer it is to 1, the greater the agreement. The closer it is to -1, the lower the agreement. However, if the Rand index is closer to 0, it means the clustering was done randomly.

```
#Show the size and centers of the clusters of data
testClust$size
```

```
## [1] 86588 33587
```

```
testClust$centers
```

```
##      Gender Customer.Type      Age Type.of.Travel     Class Flight.Distance
## 1 0.5080843             0 38.93367      0.6161593 0.7858479        683.5635
## 2 0.5038259             0 42.35925      0.8931730 1.7400185       2629.6770
##   Inflight.wifi.service Departure.Arrival.time.convenient
## 1              2.793667                          3.274091
## 2              2.798404                          3.052669
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1               2.819328      2.980482       3.169354        3.182658
## 2               2.956293      2.999821       3.332212        3.719326
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1     3.333568               3.269552         3.294995         3.269910
## 2     3.783309               3.662191         3.618751         3.658261
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1         3.586363        3.241292         3.601804    3.219753
## 2         3.755769        3.440885         3.752017    3.490964
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes satisfaction
## 1                   14.93303                 15.44097     1.337391
## 2                   14.47310                 14.75470     1.673773
```

```
testClust$withinss #within sum of squares
```

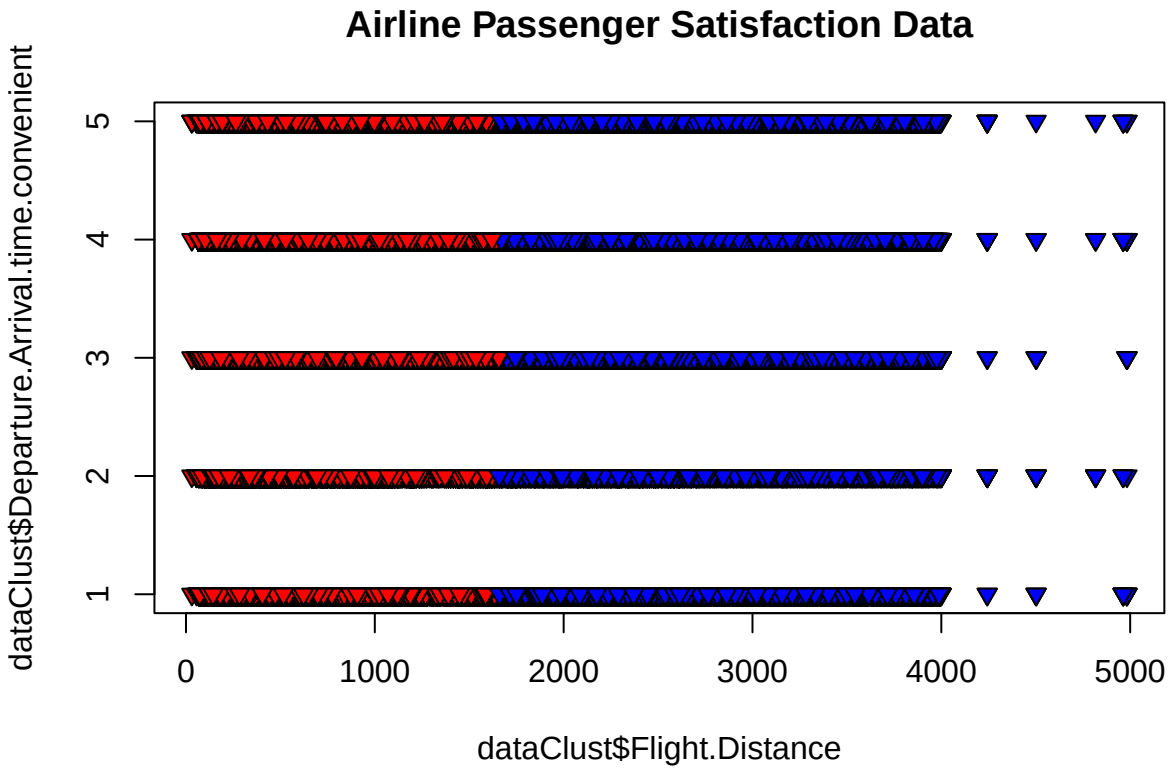
```
## [1] 14346470828 15499091777
```

```
testClust$tot.withinss #within sum of squares by cluster
```

```
## [1] 29845562605
```

We can also try a small plot graph comparing certain columns of data to see if the clusters form

```
plot(dataClust$Flight.Distance, dataClust$Departure.Arrival.time.convenient, pch=25, bg=c("red","blue")
[unclass(testClust$cluster)], main="Airline Passenger Satisfaction Data")
```



Now to obtain the Rand Index for the k-means clustering.

```
aggregate(dataClust, by=list(cluster=testClust$cluster), mean)
```

```
##   cluster    Gender Customer.Type      Age Type.of.Travel     Class
## 1       1 0.5080843             0 38.93367      0.6161593 0.7858479
## 2       2 0.5038259             0 42.35925      0.8931730 1.7400185
##   Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
## 1        683.5635              2.793667                          3.274091
## 2       2629.6770              2.798404                          3.052669
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1               2.819328      2.980482       3.169354        3.182658
## 2               2.956293      2.999821       3.332212        3.719326
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1     3.333568               3.269552         3.294995         3.269910
## 2     3.783309               3.662191         3.618751         3.658261
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1         3.586363        3.241292         3.601804    3.219753
## 2         3.755769        3.440885         3.752017    3.490964
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes satisfaction
## 1                   14.93303                 15.44097     1.337391
## 2                   14.47310                 14.75470     1.673773
```

```
ct.km <- table(dataClust$Departure.Arrival.time.convenient, testClust$cluster)
```

```
#Quantify Agreement
library(flexclust)
```

```
## Warning: package 'flexclust' was built under R version 4.2.3
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
randIndex(ct.km)
```

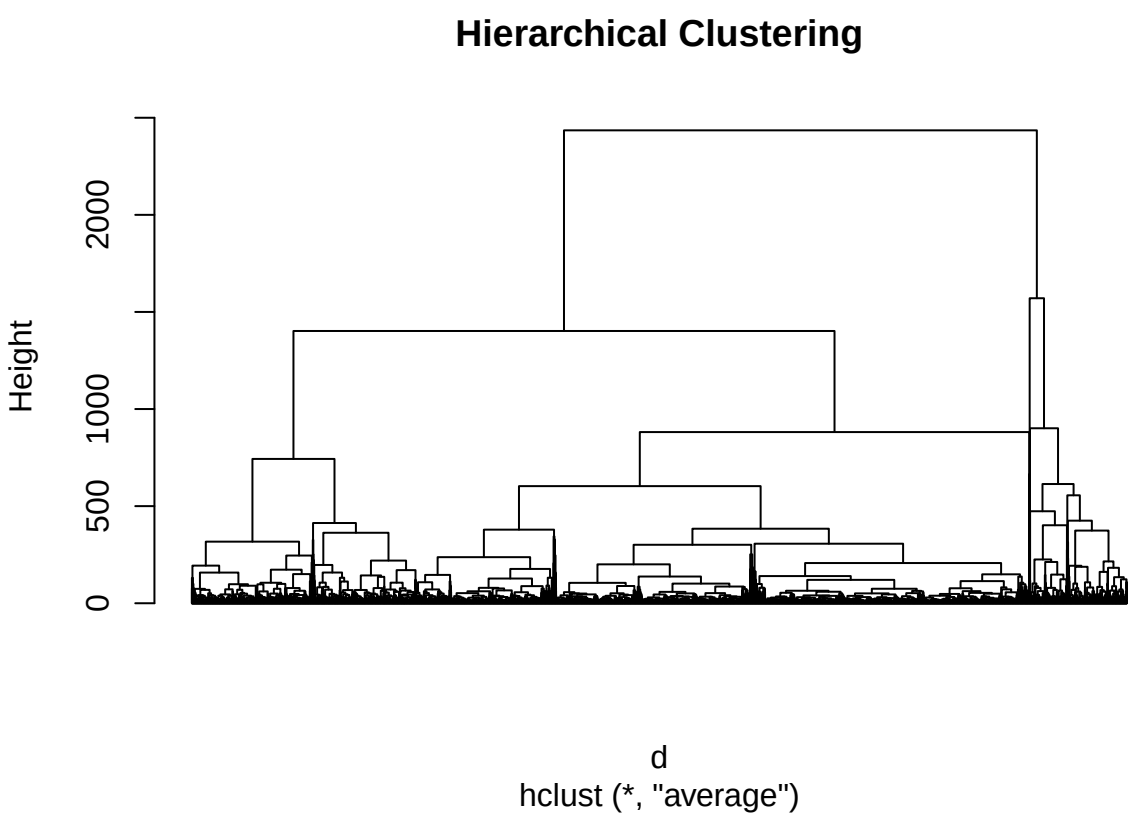
```
##        ARI
## 0.00604163
```

##Hierarchical Clustering

With clustering using the k-means algorithm, we can now move on to implementing a hierarchical clustering algorithm. Hierarchical clustering differs in that it does not require k to be specified beforehand. The other difference is that it creates a dendogram of the clustering of data.

For this dataset, the data must be split into subsets in order to perform the hierarchical clustering algorithm, since it is a greedy algorithm and will bog down with large datasets. There are 3 different ways to measure distance between clusters.With the same dataset, the euclidean distances will be calculated using average-linkage.

```
#Partition Data
set.seed(1234)
i <- sample(1:nrow(dataClust), nrow(dataClust)*0.1, replace=FALSE)
hier1 <- dataClust[i,]
hier2 <- dataClust[-i,]

d <- dist(hier1)
fit.average <- hclust(d, method="average")
plot(fit.average, hang=-1, cex=.001,
     main="Hierarchical Clustering")
```



The height will indicate the criterion value for where the clusters are joined. From here we can then begin cutting the dendogram and determine which cuts can provide the best Rand value.

```
for (c in 2:9){
 cluster_cut <- cutree(fit.average, c)
 table_cut <- table(cluster_cut, hier1$Departure.Arrival.time.convenient)
 print(table_cut)
 ri <- randIndex(table_cut)
 print(paste("cut=", c, "Rand index = ", ri))
}
```

```
##
## cluster_cut    1    2    3    4    5
##           1 1703 1814 1947 2773 2527
##           2  248  240  260  240  265
## [1] "cut= 2 Rand index =  0.00325562367429268"
##
## cluster_cut    1    2    3    4    5
##           1 1703 1814 1947 2773 2527
##           2  246  240  259  238  264
##           3    2    0    1    2    1
## [1] "cut= 3 Rand index =  0.00325797534626379"
##
## cluster_cut    1    2    3    4    5
##           1 1187 1307 1400 2111 1873
##           2  516  507  547  662  654
```

```
##           3  246  240  259  238  264
##           4    2    0    1    2    1
## [1] "cut= 4 Rand index =  0.00569456044489552"
##
## cluster_cut    1    2    3    4    5
##           1 1187 1307 1400 2111 1873
##           2  516  507  547  662  654
##           3  245  240  259  238  264
##           4    1    0    0    0    0
##           5    2    0    1    2    1
## [1] "cut= 5 Rand index =  0.00569533025116441"
##
## cluster_cut    1    2    3    4    5
##           1 1185 1307 1400 2109 1872
##           2  516  507  547  662  654
##           3  245  240  259  238  264
##           4    1    0    0    0    0
##           5    2    0    0    2    1
##           6    2    0    1    2    1
## [1] "cut= 6 Rand index =  0.00567968169839853"
##
## cluster_cut    1    2    3    4    5
##           1 1185 1307 1400 2109 1872
##           2  248  245  261  296  283
##           3  245  240  259  238  264
##           4  268  262  286  366  371
##           5    1    0    0    0    0
##           6    2    0    0    2    1
##           7    2    0    1    2    1
## [1] "cut= 7 Rand index =  0.00633802144734137"
##
## cluster_cut    1    2    3    4    5
##           1 1185 1307 1400 2109 1872
##           2  248  245  261  296  283
##           3  245  240  259  238  264
##           4  268  262  286  366  371
##           5    1    0    0    0    0
##           6    1    0    0    2    1
##           7    2    0    1    2    1
##           8    1    0    0    0    0
## [1] "cut= 8 Rand index =  0.00633801177436271"
##
## cluster_cut    1    2    3    4    5
##           1 1185 1307 1400 2109 1872
##           2  248  245  261  296  283
##           3  152  150  157  157  152
##           4  268  262  286  366  371
##           5   93   90  102   81  112
##           6    1    0    0    0    0
##           7    1    0    0    2    1
##           8    2    0    1    2    1
##           9    1    0    0    0    0
## [1] "cut= 9 Rand index =  0.00650851108089406"
```

Running the cuts reveal that the best cut is at 9, meaning cl

##Model-based Clustering Both k-means and hierarchical approaches to clustering are done, but we can also do a model-based approach to identify the most likely number of clusters and optimal model using maximum likelihood estimation and Bayes criteria. The model is selected based on the largest BIC (Bayes Information Criterion) for Expectation-Maximization.

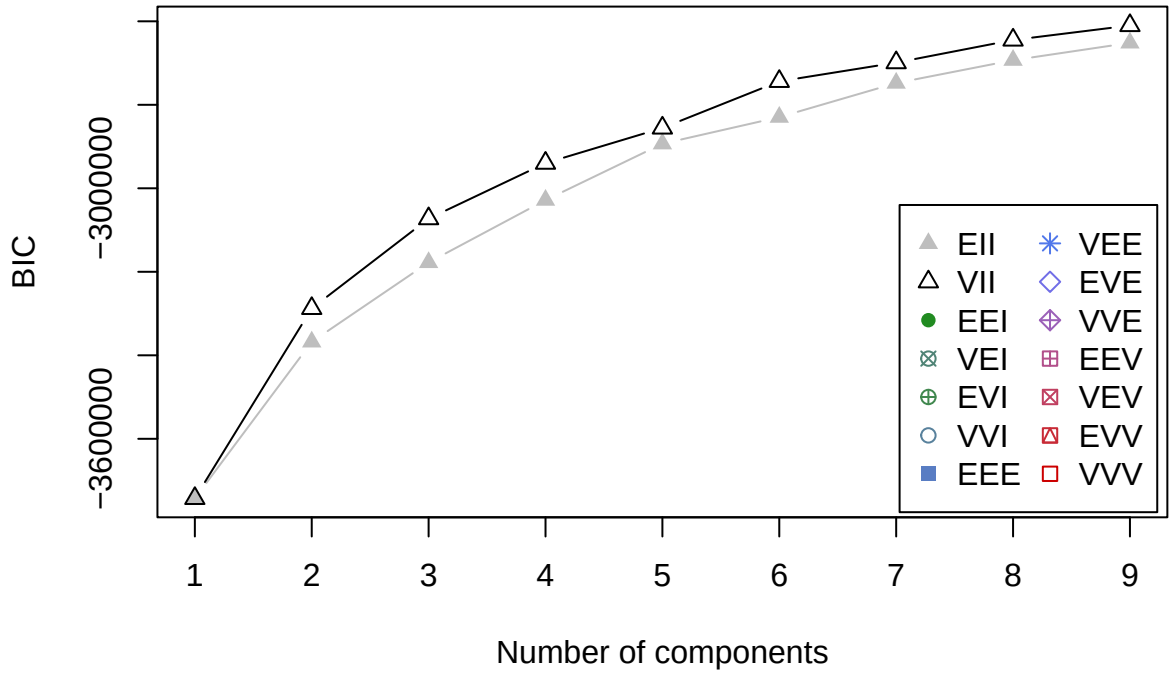
We can then use Mclust function to select the optimal model and plot the results for which model has the largest BIC and the number of clusters. However the data must be partitioned as well since the model-based algorithm can also take longer with large datasets.

```
# Model Based Clustering
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 4.2.3
```

```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

```
fitBIC <-mclustBIC(hier1)
plot(fitBIC, what = "BIC") # plot results
```

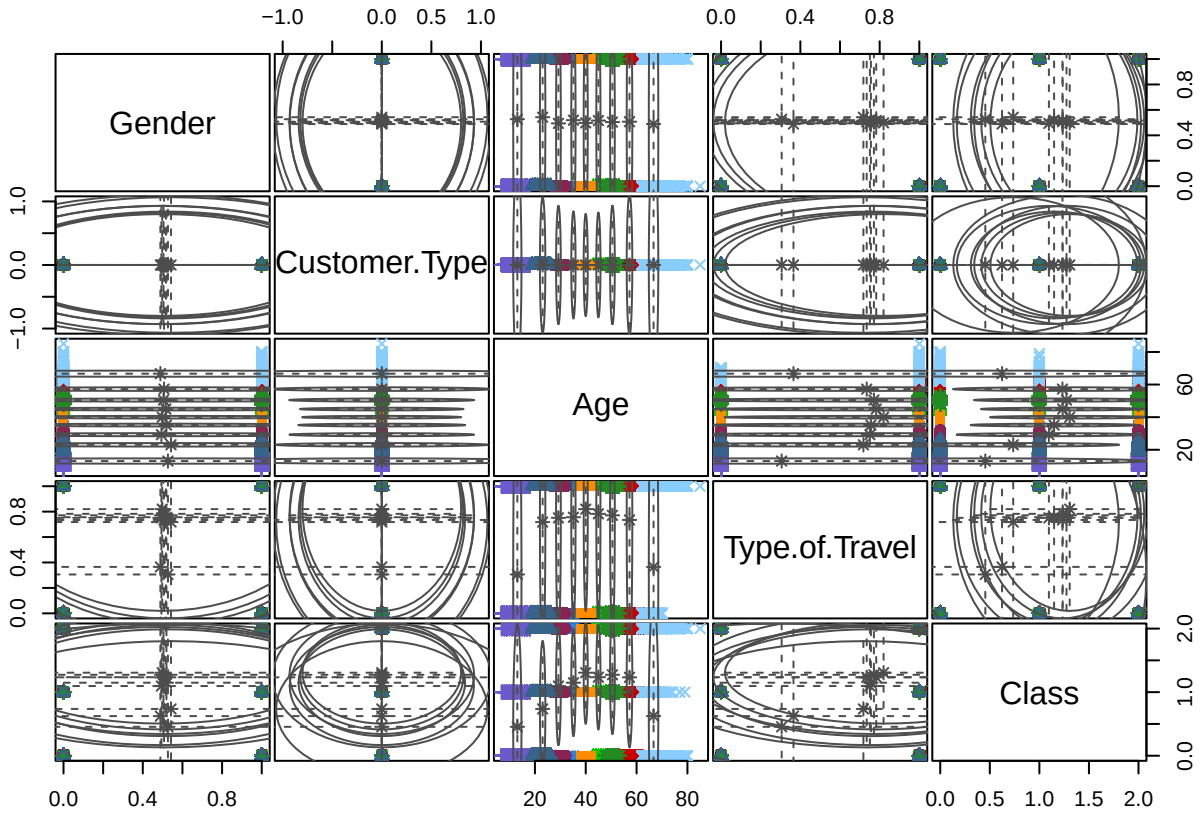



```
summary(fitBIC) # display the best model
```

```
## Best BIC values:
```

```
##              VII,9       VII,8       EII,9
## BIC      -2610051 -2644770.25 -2651669.22
## BIC diff        0   -34719.11   -41618.08
```

```
fitClass <-Mclust(hier1[,1:5])
plot(fitClass, what = "classification") # plot results
```



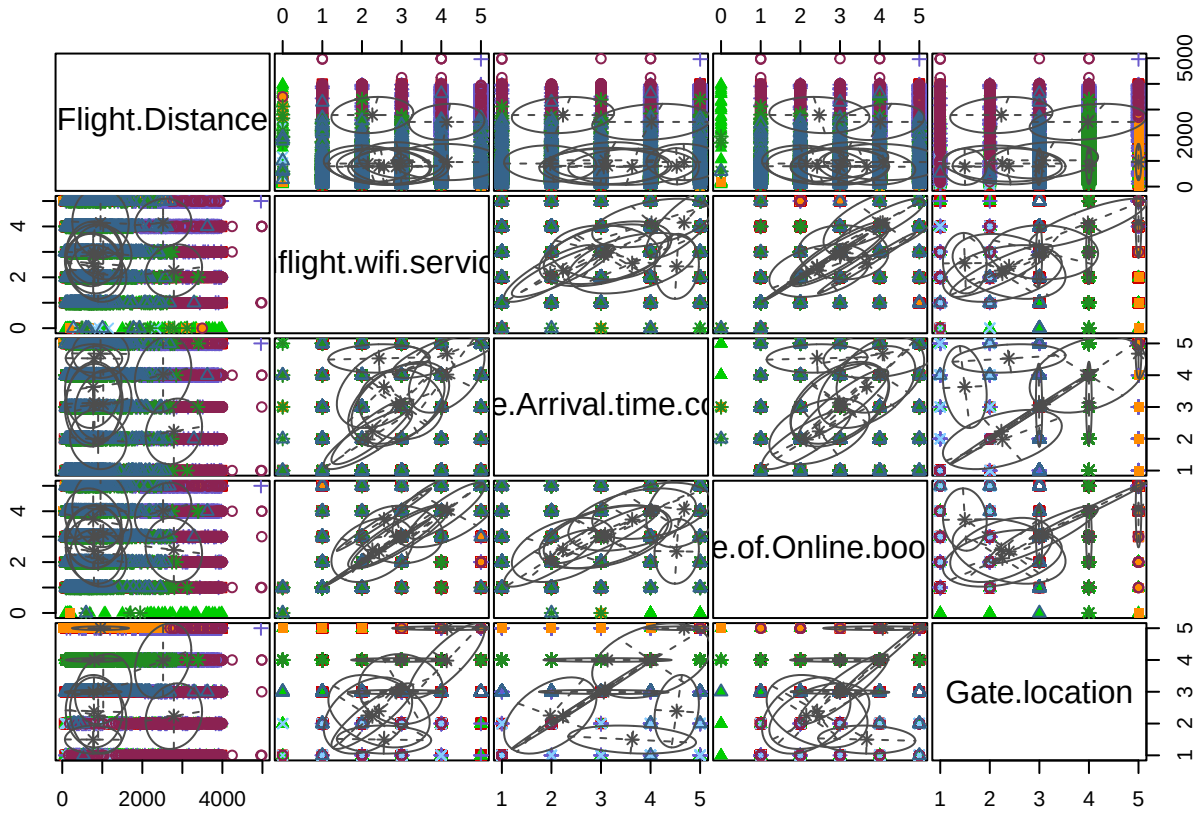

```
summary(fitClass, parameters=TRUE) # display the best model
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VII (spherical, varying volume) model with 9 components:
##
##  log-likelihood     n df       BIC       ICL
##       -111829.7 12017 62 -224241.8 -224611.9
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
## 1203 1642 1399  948 1544  808 1327 1526 1620
##
## Mixing probabilities:
##          1          2          3          4          5          6          7
## 0.10195521 0.13317065 0.11768914 0.08038683 0.12737660 0.06902757 0.11173190
```

```
##          8          9
## 0.12768004 0.13098206
##
## Means:
##                      [,1]       [,2]       [,3]       [,4]       [,5]
## Gender          0.5192844  0.5071939  0.5162231  0.5267483  0.4979521
## Customer.Type   0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## Age            35.1533348 57.2196206 44.9764196 13.0871053 39.9879224
## Type.of.Travel  0.7559300  0.7346589  0.7826770  0.3056823  0.8196832
## Class           1.1485800  1.2331310  1.2354838  0.4545934  1.3062265
##                      [,6]       [,7]       [,8]       [,9]
## Gender          0.4886919  0.4939018  0.5042524  0.5424111
## Customer.Type   0.0000000  0.0000000  0.0000000  0.0000000
## Age            66.6778556 29.2544545 50.5007870 23.0424981
## Type.of.Travel  0.3652801  0.7508458  0.7709818  0.7180250
## Class           0.6233659  1.0961260  1.2734863  0.7361557
##
## Variances:
## [,,1]
##                   Gender Customer.Type       Age Type.of.Travel     Class
## Gender         0.7060317     0.0000000 0.0000000      0.0000000 0.0000000
## Customer.Type  0.0000000     0.7060317 0.0000000      0.0000000 0.0000000
## Age            0.0000000     0.0000000 0.7060317      0.0000000 0.0000000
## Type.of.Travel 0.0000000     0.0000000 0.0000000      0.7060317 0.0000000
## Class          0.0000000     0.0000000 0.0000000      0.0000000 0.7060317
## [,,2]
##                   Gender Customer.Type      Age Type.of.Travel    Class
## Gender         1.214781      0.000000 0.000000       0.000000 0.000000
## Customer.Type  0.000000      1.214781 0.000000       0.000000 0.000000
## Age            0.000000      0.000000 1.214781       0.000000 0.000000
## Type.of.Travel 0.000000      0.000000 0.000000       1.214781 0.000000
## Class          0.000000      0.000000 0.000000       0.000000 1.214781
## [,,3]
##                   Gender Customer.Type       Age Type.of.Travel     Class
## Gender         0.6762341     0.0000000 0.0000000      0.0000000 0.0000000
## Customer.Type  0.0000000     0.6762341 0.0000000      0.0000000 0.0000000
## Age            0.0000000     0.0000000 0.6762341      0.0000000 0.0000000
## Type.of.Travel 0.0000000     0.0000000 0.0000000      0.6762341 0.0000000
## Class          0.0000000     0.0000000 0.0000000      0.0000000 0.6762341
## [,,4]
##                   Gender Customer.Type      Age Type.of.Travel    Class
## Gender         2.642639      0.000000 0.000000       0.000000 0.000000
## Customer.Type  0.000000      2.642639 0.000000       0.000000 0.000000
## Age            0.000000      0.000000 2.642639       0.000000 0.000000
## Type.of.Travel 0.000000      0.000000 0.000000       2.642639 0.000000
## Class          0.000000      0.000000 0.000000       0.000000 2.642639
## [,,5]
##                   Gender Customer.Type       Age Type.of.Travel     Class
## Gender         0.6407783     0.0000000 0.0000000      0.0000000 0.0000000
## Customer.Type  0.0000000     0.6407783 0.0000000      0.0000000 0.0000000
## Age            0.0000000     0.0000000 0.6407783      0.0000000 0.0000000
## Type.of.Travel 0.0000000     0.0000000 0.0000000      0.6407783 0.0000000
## Class          0.0000000     0.0000000 0.0000000      0.0000000 0.6407783
## [,,6]
```

```
##                  Gender Customer.Type     Age Type.of.Travel   Class
## Gender          3.48618       0.00000 0.00000        0.00000 0.00000
## Customer.Type   0.00000       3.48618 0.00000        0.00000 0.00000
## Age             0.00000       0.00000 3.48618        0.00000 0.00000
## Type.of.Travel 0.00000       0.00000 0.00000        3.48618 0.00000
## Class           0.00000       0.00000 0.00000        0.00000 3.48618
## [,,7]
##                    Gender Customer.Type       Age Type.of.Travel     Class
## Gender          0.8634853     0.0000000 0.0000000      0.0000000 0.0000000
## Customer.Type   0.0000000     0.8634853 0.0000000      0.0000000 0.0000000
## Age             0.0000000     0.0000000 0.8634853      0.0000000 0.0000000
## Type.of.Travel 0.0000000     0.0000000 0.0000000      0.8634853 0.0000000
## Class           0.0000000     0.0000000 0.0000000      0.0000000 0.8634853
## [,,8]
##                    Gender Customer.Type       Age Type.of.Travel     Class
## Gender          0.8676903     0.0000000 0.0000000      0.0000000 0.0000000
## Customer.Type   0.0000000     0.8676903 0.0000000      0.0000000 0.0000000
## Age             0.0000000     0.0000000 0.8676903      0.0000000 0.0000000
## Type.of.Travel 0.0000000     0.0000000 0.0000000      0.8676903 0.0000000
## Class           0.0000000     0.0000000 0.0000000      0.0000000 0.8676903
## [,,9]
##                    Gender Customer.Type      Age Type.of.Travel    Class
## Gender          1.136673      0.000000 0.000000       0.000000 0.000000
## Customer.Type   0.000000      1.136673 0.000000       0.000000 0.000000
## Age             0.000000      0.000000 1.136673       0.000000 0.000000
## Type.of.Travel 0.000000      0.000000 0.000000       1.136673 0.000000
## Class           0.000000      0.000000 0.000000       0.000000 1.136673
```

```
fitClass <-Mclust(hier1[,6:10])
plot(fitClass, what = "classification") # plot results
```




```
summary(fitClass, parameters=TRUE) # display the best model
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEV (ellipsoidal, equal volume and shape) model with 9 components:
##
##  log-likelihood     n  df       BIC       ICL
##       -146599.7 12017 148 -294589.7 -300524.5
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
## 1657 2666 1967  827  861  124 1518 1327 1070
##
## Mixing probabilities:
##         1         2         3         4         5         6         7
## 0.15092960 0.21145433 0.16321740 0.06758775 0.06511642 0.01057445 0.12612872
##         8         9
## 0.10327451 0.10171682
##
## Means:
##                                      [,1]        [,2]       [,3]        [,4]
## Flight.Distance                896.751848 1023.504617 792.201444 2518.130362
## Inflight.wifi.service            2.150962    2.982661   2.430281    4.136748
```

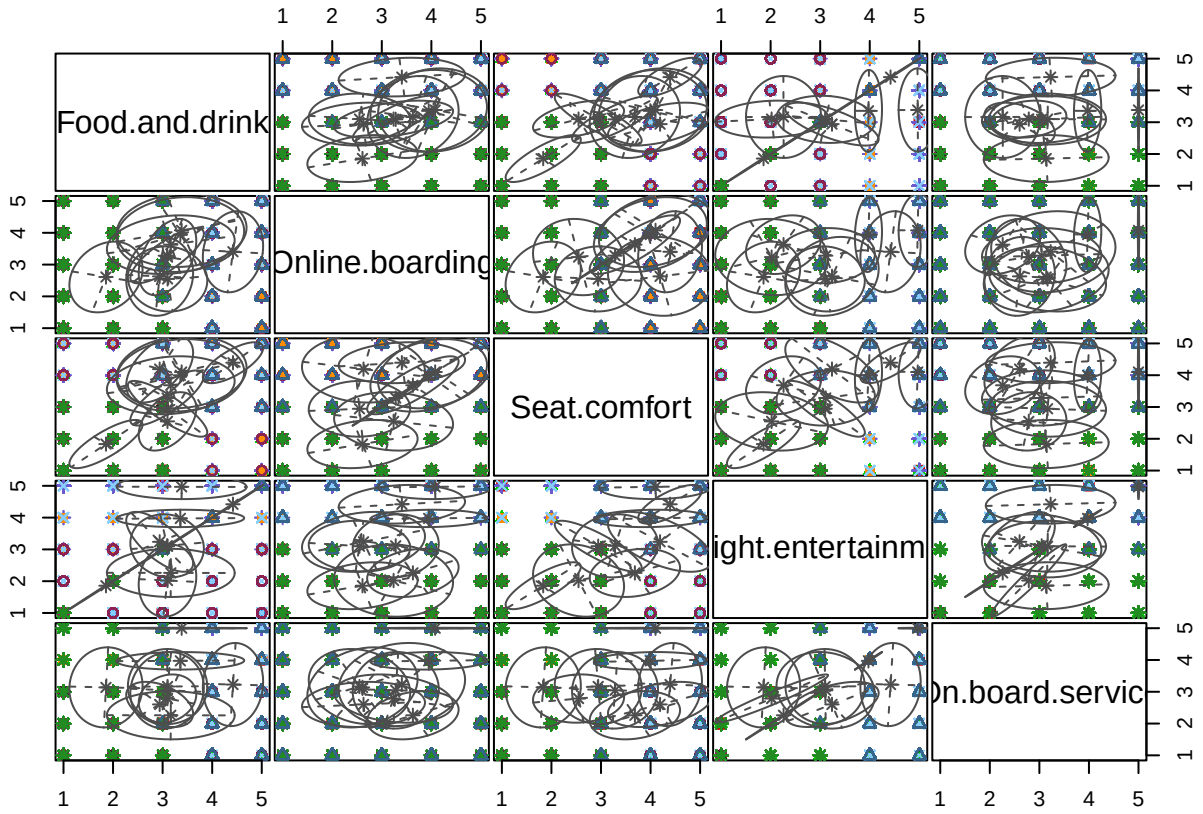
```
## Departure.Arrival.time.convenient   1.976532    3.082869   4.527429    4.018914
## Ease.of.Online.booking              2.149171    3.086140   2.422018    4.017461
## Gate.location                       2.253598    3.082956   2.388711    3.979293
##                                         [,5]       [,6]       [,7]       [,8]
## Flight.Distance                   956.255593 772.334475 2787.390064 805.361891
## Inflight.wifi.service               4.147281   2.555492   2.261937   2.966816
## Departure.Arrival.time.convenient   4.677692   3.628538   2.233094   3.132099
## Ease.of.Online.booking              4.077648   3.673534   2.478762   2.986085
## Gate.location                       5.000000   1.496068   2.233094   4.000000
##                                         [,9]
## Flight.Distance                   772.921477
## Inflight.wifi.service               2.866430
## Departure.Arrival.time.convenient   3.056492
## Ease.of.Online.booking              2.999178
## Gate.location                       2.993455
##
## Variances:
## [,,1]
##                                   Flight.Distance Inflight.wifi.service
## Flight.Distance                      498020.16788          -114.4850725
## Inflight.wifi.service                  -114.48507             1.2432858
## Departure.Arrival.time.convenient       -50.48661             0.8672409
## Ease.of.Online.booking                 -113.53411             1.2407279
## Gate.location                          -120.74692             0.4659214
##                                   Departure.Arrival.time.convenient
## Flight.Distance                                         -50.4866054
## Inflight.wifi.service                                     0.8672409
## Departure.Arrival.time.convenient                         0.8686644
## Ease.of.Online.booking                                    0.8682436
## Gate.location                                             0.5097093
##                                   Ease.of.Online.booking Gate.location
## Flight.Distance                             -113.5341106  -120.7469162
## Inflight.wifi.service                          1.2407279     0.4659214
## Departure.Arrival.time.convenient              0.8682436     0.5097093
## Ease.of.Online.booking                         1.2445028     0.4667805
## Gate.location                                  0.4667805     1.4400107
## [,,2]
##                                   Flight.Distance Inflight.wifi.service
## Flight.Distance                      498020.11207          -86.90230913
## Inflight.wifi.service                   -86.90231            1.13661370
## Departure.Arrival.time.convenient       146.34491           -0.04728127
## Ease.of.Online.booking                  143.31854           -0.03024517
## Gate.location                           146.33781           -0.04735839
##                                   Departure.Arrival.time.convenient
## Flight.Distance                                        146.34491446
## Inflight.wifi.service                                   -0.04728127
## Departure.Arrival.time.convenient                        1.21656147
## Ease.of.Online.booking                                   1.10945326
## Gate.location                                            1.21275748
##                                   Ease.of.Online.booking Gate.location
## Flight.Distance                           143.31853507  146.33780805
## Inflight.wifi.service                      -0.03024517   -0.04735839
## Departure.Arrival.time.convenient           1.10945326    1.21275748
## Ease.of.Online.booking                      1.28382092    1.10968734
```

```
## Gate.location                              1.10968734    1.21528556
## [,,3]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                    498020.207880         -105.87431911
## Inflight.wifi.service                -105.874319            1.63523931
## Departure.Arrival.time.convenient      -6.441762            0.04038129
## Ease.of.Online.booking               -107.569189            1.62476524
## Gate.location                         -19.613781           -0.44908756
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                       -6.44176195
## Inflight.wifi.service                                  0.04038129
## Departure.Arrival.time.convenient                      0.20037873
## Ease.of.Online.booking                                 0.04289972
## Gate.location                                          0.10326996
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                          -107.56918925   -19.6137807
## Inflight.wifi.service                      1.62476524    -0.4490876
## Departure.Arrival.time.convenient          0.04289972     0.1032700
## Ease.of.Online.booking                     1.62071178    -0.4416036
## Gate.location                             -0.44160355     1.3001376
## [,,4]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                      498020.0132            59.4394006
## Inflight.wifi.service                    59.4394             0.9111469
## Departure.Arrival.time.convenient       154.6608             0.5066857
## Ease.of.Online.booking                  194.9740             0.9056211
## Gate.location                           233.8572             0.9108213
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                      154.6608141
## Inflight.wifi.service                                  0.5066857
## Departure.Arrival.time.convenient                      1.4343085
## Ease.of.Online.booking                                 0.4852969
## Gate.location                                          0.4838802
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                          194.9740037   233.8572158
## Inflight.wifi.service                      0.9056211     0.9108213
## Departure.Arrival.time.convenient          0.4852969     0.4838802
## Ease.of.Online.booking                     1.2222740     1.2922491
## Gate.location                              1.2922491     1.3834108
## [,,5]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                    4.980202e+05         -3.097135e+01
## Inflight.wifi.service             -3.097135e+01          1.901415e+00
## Departure.Arrival.time.convenient -5.831240e+00          7.975077e-01
## Ease.of.Online.booking            -6.392393e+01          1.113895e+00
## Gate.location                      6.311402e-27         -6.545482e-28
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                     -5.831240e+00
## Inflight.wifi.service                                7.975077e-01
## Departure.Arrival.time.convenient                    5.827104e-01
## Ease.of.Online.booking                               3.313591e-01
## Gate.location                                        8.168790e-28
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                         -6.392393e+01  6.311402e-27
```

```
## Inflight.wifi.service                        1.113895e+00 -6.545482e-28
## Departure.Arrival.time.convenient            3.313591e-01  8.168790e-28
## Ease.of.Online.booking                       2.232779e+00  3.149508e-28
## Gate.location                                3.149508e-28  3.164604e-03
## [,,6]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                    498020.186941          -61.91298238
## Inflight.wifi.service                 -61.912982            1.41718813
## Departure.Arrival.time.convenient      87.665007            0.63916291
## Ease.of.Online.booking               -148.687198            1.44803762
## Gate.location                          -2.923253           -0.02224824
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                         87.6650069
## Inflight.wifi.service                                    0.6391629
## Departure.Arrival.time.convenient                        1.6613607
## Ease.of.Online.booking                                   0.6886641
## Gate.location                                           -0.1160991
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                             -148.6871983   -2.92325293
## Inflight.wifi.service                          1.4480376   -0.02224824
## Departure.Arrival.time.convenient              0.6886641   -0.11609908
## Ease.of.Online.booking                         1.5039626   -0.04032700
## Gate.location                                 -0.0403270    0.19489511
## [,,7]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                     498020.16990            98.4982180
## Inflight.wifi.service                  98.49822             1.0678760
## Departure.Arrival.time.convenient     105.92939             0.9245869
## Ease.of.Online.booking                -99.89324             0.4535436
## Gate.location                         105.93008             0.9245846
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                       105.9293948
## Inflight.wifi.service                                   0.9245869
## Departure.Arrival.time.convenient                       1.0662480
## Ease.of.Online.booking                                  0.5297583
## Gate.location                                           1.0630798
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                             -99.8932443   105.9300847
## Inflight.wifi.service                         0.4535436     0.9245846
## Departure.Arrival.time.convenient             0.5297583     1.0630798
## Ease.of.Online.booking                        1.5940853     0.5297589
## Gate.location                                 0.5297589     1.0662407
## [,,8]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                     4.980203e+05         -1.937219e+01
## Inflight.wifi.service              -1.937219e+01          1.519115e+00
## Departure.Arrival.time.convenient   2.281481e+01          6.846998e-01
## Ease.of.Online.booking              2.618143e+01          1.336053e+00
## Gate.location                      -1.562365e-61         -1.030639e-63
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                       2.281481e+01
## Inflight.wifi.service                                 6.846998e-01
## Departure.Arrival.time.convenient                     1.662105e+00
## Ease.of.Online.booking                                6.856227e-01
```

```
## Gate.location                                      -7.865802e-64
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                            2.618143e+01 -1.562365e-61
## Inflight.wifi.service                      1.336053e+00 -1.030639e-63
## Departure.Arrival.time.convenient          6.856227e-01 -7.865802e-64
## Ease.of.Online.booking                     1.528660e+00 -2.271340e-63
## Gate.location                             -2.271340e-63  3.164604e-03
## [,,9]
##                                  Flight.Distance Inflight.wifi.service
## Flight.Distance                     4.980202e+05          52.228460665
## Inflight.wifi.service               5.222846e+01           1.480516592
## Departure.Arrival.time.convenient   5.419746e+01           0.663848908
## Ease.of.Online.booking              1.713204e+01           1.322632830
## Gate.location                       3.738533e+00           0.006723742
##                                  Departure.Arrival.time.convenient
## Flight.Distance                                      54.197461553
## Inflight.wifi.service                                 0.663848908
## Departure.Arrival.time.convenient                     1.695933502
## Ease.of.Online.booking                                0.739197182
## Gate.location                                        -0.007889345
##                                  Ease.of.Online.booking Gate.location
## Flight.Distance                            17.13204320   3.738532892
## Inflight.wifi.service                       1.32263283   0.006723742
## Departure.Arrival.time.convenient           0.73919718  -0.007889345
## Ease.of.Online.booking                      1.54146560  -0.010019158
## Gate.location                              -0.01001916   0.003946407
```

```
fitClass <-Mclust(hier1[,11:15])
plot(fitClass, what = "classification") # plot results
```




```
summary(fitClass, parameters=TRUE) # display the best model
```

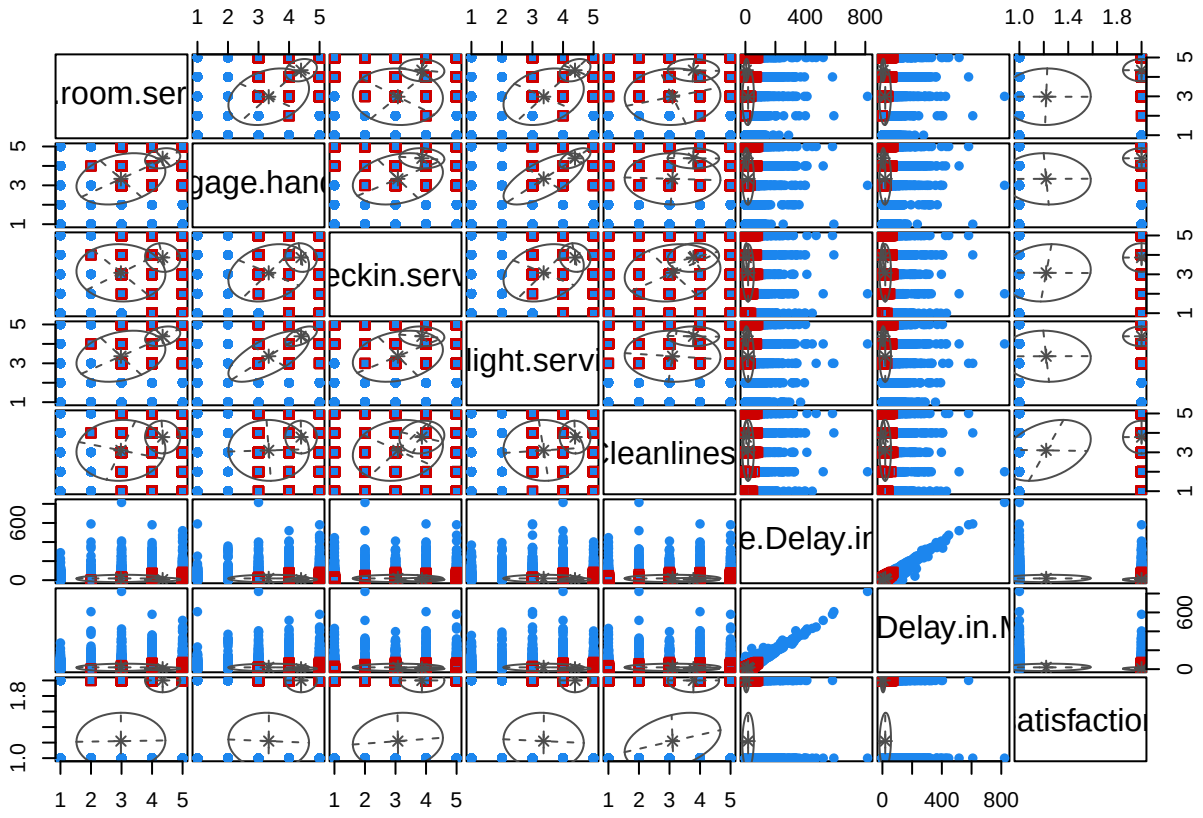
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEV (ellipsoidal, equal volume and shape) model with 9 components:
##
##  log-likelihood     n  df       BIC       ICL
##       -46012.15 12017 148 -93414.61 -97257.61
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
##   34    9 1051 1592 2030  248 1760 2745 2548
##
## Mixing probabilities:
##            1            2            3            4            5            6
## 0.0029054189 0.0007613175 0.0725201800 0.1311819285 0.1504606217 0.0340305558
##            7            8            9
## 0.1536947584 0.2236742116 0.2307710077
##
## Means:
##                              [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## Food.and.drink           2.940445 3.098145 3.100003 3.383900 3.357107 3.090254
## Online.boarding          2.630558 3.237359 2.549330 4.076144 3.945861 3.280736
```

```
## Seat.comfort           4.191644 2.547087 2.943670 4.100391 3.995500 3.283182
## Inflight.entertainment 3.237799 2.020137 3.100003 4.967605 3.974012 3.111558
## On.board.service       2.601291 2.786244 3.165665 5.000000 3.973459 3.046235
##                            [,7]     [,8]     [,9]
## Food.and.drink         3.164631 1.857065 4.419412
## Online.boarding        3.627149 2.610030 3.404558
## Seat.comfort           3.683571 1.834228 4.393404
## Inflight.entertainment 2.260318 1.857065 4.419412
## On.board.service       2.260318 3.137737 3.227395
##
## Variances:
## [,,1]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink             0.45596049      0.07576234   0.34569576
## Online.boarding            0.07576234      1.54149829  -0.08500379
## Seat.comfort               0.34569576     -0.08500379   1.10578585
## Inflight.entertainment    -0.45367635      0.21806018  -0.58325460
## On.board.service          -0.19951714      0.07238862   0.28546726
##                        Inflight.entertainment On.board.service
## Food.and.drink                      -0.4536764      -0.19951714
## Online.boarding                      0.2180602       0.07238862
## Seat.comfort                        -0.5832546       0.28546726
## Inflight.entertainment               0.8886815       0.34055410
## On.board.service                     0.3405541       0.50512478
## [,,2]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink             0.33789562      0.29462817   0.23995016
## Online.boarding            0.29462817      1.44223956  -0.10437788
## Seat.comfort               0.23995016     -0.10437788   0.84462011
## Inflight.entertainment     0.07201751     -0.06529393  -0.17211273
## On.board.service           0.07719046     -0.20759629   0.03755939
##                        Inflight.entertainment On.board.service
## Food.and.drink                      0.07201751       0.07719046
## Online.boarding                    -0.06529393      -0.20759629
## Seat.comfort                       -0.17211273       0.03755939
## Inflight.entertainment              1.29554457       0.83083029
## On.board.service                    0.83083029       0.57675107
## [,,3]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink            0.641976694     -0.04123225  -0.59348296
## Online.boarding          -0.041232248      1.04956678   0.01862030
## Seat.comfort             -0.593482961      0.01862030   0.77066926
## Inflight.entertainment    0.641816903     -0.04123225  -0.59348296
## On.board.service         -0.007654723     -0.20775599  -0.03334546
##                        Inflight.entertainment On.board.service
## Food.and.drink                     0.641816903     -0.007654723
## Online.boarding                   -0.041232248     -0.207755985
## Seat.comfort                      -0.593482961     -0.033345461
## Inflight.entertainment             0.641976694     -0.007654723
## On.board.service                  -0.007654723      1.392861493
## [,,4]
##                        Food.and.drink Online.boarding  Seat.comfort
## Food.and.drink           1.716931e+00    1.163471e-01  1.823444e-01
## Online.boarding          1.163471e-01    1.325987e+00  3.615252e-01
```

```
## Seat.comfort             1.823444e-01    3.615252e-01  1.305368e+00
## Inflight.entertainment  -5.232473e-03    1.442329e-02  2.518854e-02
## On.board.service         1.269497e-27    4.420042e-27 -4.251038e-28
##                        Inflight.entertainment On.board.service
## Food.and.drink                 -5.232473e-03     1.269497e-27
## Online.boarding                 1.442329e-02     4.420042e-27
## Seat.comfort                    2.518854e-02    -4.251038e-28
## Inflight.entertainment          1.486051e-01    -1.751951e-26
## On.board.service               -1.751951e-26     1.597908e-04
## [,,5]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink            1.66032514      0.14610658   0.14696593
## Online.boarding           0.14610658      1.38506528   0.39023520
## Seat.comfort              0.14696593      0.39023520   1.30094259
## Inflight.entertainment    0.01250060      0.02348740   0.03570232
## On.board.service          0.01335334      0.02403551   0.03697857
##                        Inflight.entertainment On.board.service
## Food.and.drink                      0.01250060       0.01335334
## Online.boarding                     0.02348740       0.02403551
## Seat.comfort                        0.03570232       0.03697857
## Inflight.entertainment              0.07468162       0.07519564
## On.board.service                    0.07519564       0.07603629
## [,,6]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink            0.55079160      0.51471946   0.51460858
## Online.boarding           0.51471946      0.75493022   0.75356474
## Seat.comfort              0.51460858      0.75356474   0.75254759
## Inflight.entertainment   -0.06253422      0.05638138   0.05135273
## On.board.service          0.06105777      0.02950089   0.03179289
##                        Inflight.entertainment On.board.service
## Food.and.drink                     -0.06253422       0.06105777
## Online.boarding                     0.05638138       0.02950089
## Seat.comfort                        0.05135273       0.03179289
## Inflight.entertainment              0.98573621      -0.14200948
## On.board.service                   -0.14200948       1.45304530
## [,,7]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink           1.663705981       0.1813029    0.1681028
## Online.boarding          0.181302865       0.8522860    0.6939567
## Seat.comfort             0.168102829       0.6939567    0.8323475
## Inflight.entertainment  -0.002806278       0.1659215    0.1475059
## On.board.service        -0.002806278       0.1659215    0.1475059
##                        Inflight.entertainment On.board.service
## Food.and.drink                    -0.002806278    -0.002806278
## Online.boarding                    0.165921496     0.165921496
## Seat.comfort                       0.147505915     0.147505915
## Inflight.entertainment             0.574355730     0.574195939
## On.board.service                   0.574195939     0.574355730
## [,,8]
##                        Food.and.drink Online.boarding Seat.comfort
## Food.and.drink            0.54561488       0.1891982   0.44949264
## Online.boarding           0.18919816       1.2388538   0.18427556
## Seat.comfort              0.44949264       0.1842756   0.57651277
## Inflight.entertainment    0.54545509       0.1891982   0.44949264
```

```
## On.board.service           0.03654873       0.2939930   0.04631668
##                         Inflight.entertainment On.board.service
## Food.and.drink                      0.54545509       0.03654873
## Online.boarding                     0.18919816       0.29399297
## Seat.comfort                        0.44949264       0.04631668
## Inflight.entertainment              0.54561488       0.03654873
## On.board.service                    0.03654873       1.59045461
## [,,9]
##                         Food.and.drink Online.boarding Seat.comfort
## Food.and.drink              0.37026980       0.1080642   0.28106699
## Online.boarding             0.10806425       1.6283647   0.11152536
## Seat.comfort                0.28106699       0.1115254   0.41498270
## Inflight.entertainment      0.37011001       0.1080642   0.28106699
## On.board.service            0.05467747       0.1893217   0.07339928
##                         Inflight.entertainment On.board.service
## Food.and.drink                      0.37011001       0.05467747
## Online.boarding                     0.10806425       0.18932172
## Seat.comfort                        0.28106699       0.07339928
## Inflight.entertainment              0.37026980       0.05467747
## On.board.service                    0.05467747       1.71316396
```

```
fitClass <-Mclust(hier1[,16:23])
plot(fitClass, what = "classification") # plot results
```



```r
summary(fitClass, parameters=TRUE) # display the best model
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 2 components:
##
##  log-likelihood     n df       BIC       ICL
##       -197367.8 12017 82 -395505.8 -395945.2
##
## Clustering table:
##    1    2
## 8582 3435
##
## Mixing probabilities:
##         1         2
## 0.7187118 0.2812882
##
## Means:
##                                [,1]     [,2]
## Leg.room.service           2.980394 4.346748
## Baggage.handling           3.334087 4.392968
## Checkin.service            3.077198 3.863921
## Inflight.service           3.365118 4.391487
## Cleanliness                3.100116 3.782879
## Departure.Delay.in.Minutes 18.050053 5.812908
## Arrival.Delay.in.Minutes   19.037109 4.696726
## satisfaction                1.219267 2.000000
##
## Variances:
## [,,1]
##                            Leg.room.service Baggage.handling Checkin.service
## Leg.room.service                 2.12275367       0.56645721     -0.11514091
## Baggage.handling                 0.56645721       1.74487739      0.51136292
## Checkin.service                 -0.11514091       0.51136292      2.17964297
## Inflight.service                 0.53556404       1.16773708      0.49492674
## Cleanliness                      0.11750938      -0.03289324      0.38274688
## Departure.Delay.in.Minutes       4.45949704       2.69559835      0.40675711
## Arrival.Delay.in.Minutes         5.08506947       3.02129091      0.75208525
## satisfaction                     0.01195883      -0.02548611      0.06119524
##                            Inflight.service Cleanliness
## Leg.room.service                 0.53556404  0.11750938
## Baggage.handling                 1.16773708 -0.03289324
## Checkin.service                  0.49492674  0.38274688
## Inflight.service                 1.72979192 -0.07623133
## Cleanliness                     -0.07623133  2.46158233
## Departure.Delay.in.Minutes       0.45360766  0.35656835
## Arrival.Delay.in.Minutes         0.61453105  0.56611427
## satisfaction                    -0.02868901  0.21233927
##                            Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
## Leg.room.service                           4.4594970               5.0850695
## Baggage.handling                           2.6955983               3.0212909
```

```
## Checkin.service                          0.4067571              0.7520852
## Inflight.service                         0.4536077              0.6145310
## Cleanliness                              0.3565684              0.5661143
## Departure.Delay.in.Minutes            1471.6570855           1422.0730128
## Arrival.Delay.in.Minutes              1422.0730128           1497.3566106
## satisfaction                             0.9554246              1.1694880
##                           satisfaction
## Leg.room.service           0.01195883
## Baggage.handling          -0.02548611
## Checkin.service            0.06119524
## Inflight.service          -0.02868901
## Cleanliness                0.21233927
## Departure.Delay.in.Minutes 0.95542464
## Arrival.Delay.in.Minutes   1.16948797
## satisfaction               0.13331292
## [,,2]
##                           Leg.room.service Baggage.handling Checkin.service
## Leg.room.service             3.323308e-01     6.536926e-02   -3.386816e-02
## Baggage.handling             6.536926e-02     2.697134e-01   -2.979732e-02
## Checkin.service             -3.386816e-02    -2.979732e-02    5.423890e-01
## Inflight.service             9.319671e-02     1.439696e-01   -3.513000e-02
## Cleanliness                 -1.621444e-02    -4.762887e-03    7.297670e-02
## Departure.Delay.in.Minutes   1.070755e+00     5.744468e-01    4.285744e-01
## Arrival.Delay.in.Minutes     9.791096e-01     5.365667e-01    3.504612e-01
## satisfaction                -2.476890e-10    -1.306497e-11   -1.951209e-10
##                           Inflight.service   Cleanliness
## Leg.room.service             9.319671e-02 -1.621444e-02
## Baggage.handling             1.439696e-01 -4.762887e-03
## Checkin.service             -3.513000e-02  7.297670e-02
## Inflight.service             2.574718e-01 -1.040431e-02
## Cleanliness                 -1.040431e-02  7.102715e-01
## Departure.Delay.in.Minutes   6.444027e-01  4.323672e-01
## Arrival.Delay.in.Minutes     5.909039e-01  3.448999e-01
## satisfaction                -1.951644e-11 -6.630192e-10
##                           Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
## Leg.room.service                        1.070755e+00             9.791096e-01
## Baggage.handling                        5.744468e-01             5.365667e-01
## Checkin.service                         4.285744e-01             3.504612e-01
## Inflight.service                        6.444027e-01             5.909039e-01
## Cleanliness                             4.323672e-01             3.448999e-01
## Departure.Delay.in.Minutes              3.521399e+02             2.887986e+02
## Arrival.Delay.in.Minutes                2.887986e+02             2.586411e+02
## satisfaction                            4.444782e-08             3.713270e-08
##                            satisfaction
## Leg.room.service          -2.476890e-10
## Baggage.handling          -1.306497e-11
## Checkin.service           -1.951209e-10
## Inflight.service          -1.951644e-11
## Cleanliness               -6.630192e-10
## Departure.Delay.in.Minutes  4.444782e-08
## Arrival.Delay.in.Minutes    3.713270e-08
## satisfaction                2.303335e-02
```

##Conclusion All 3 clustering algorithms gave varied results based on the data. For k-means, the results

showed that a model of 2 clusters was the most optimal model for this data set, but had a Rand index close to 0, meaning that the clustering is more randomized. The hierarchical clustering algorithm, though it is a greedy algorithm, had shown that the data with 9 clusters is the more optimal model to go with, though the rand index for that cut is still closer to 0. Lastly, the model-based algorithm, using Mclust showed that a model with 9 clusters is the most optimal, having a higher BIC out of the 9 components, the model indicating that it will be a spherical, uneven multivariate mixture.