Variable Clustering in R

#### ***Ariful Mondal***

#### ***25 September 2016***

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

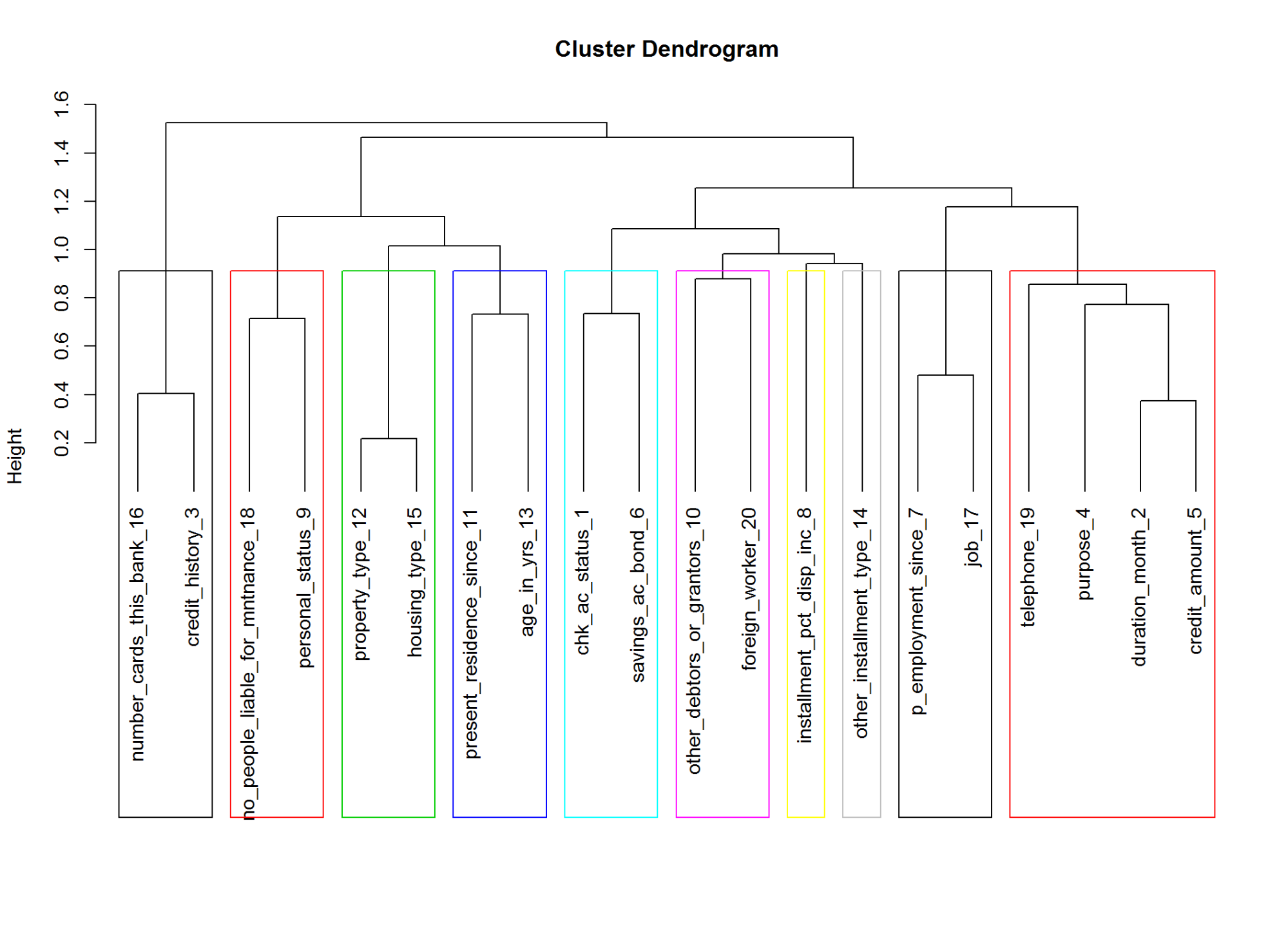
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# Set working directory  
setwd("C:/creditscoring")  
  
# List of libraries   
library(lattice) # for Histogram  
library(knitr) # for kable  
  
# Read data into R (tab delimitted)  
cdata<-read.table("data.txt", h=T, sep="")

# Varclus  
library(ClustOfVar) # for variable clustering  
library(ggplot2)  
library(GPArotation)   
library(ape) # for as.phylo  
  
# Step 1:  
vars\_quali<- cdata[,c(1,3,4,6,7,9,10,12,14,15,17,19,20)]  
vars\_quanti <- cdata[, c(2,5,8,11,13,16,18)]  
str(vars\_quanti)

## 'data.frame': 1000 obs. of 7 variables:  
## $ duration\_month\_2 : int 6 48 12 42 24 36 24 36 12 30 ...  
## $ credit\_amount\_5 : int 1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...  
## $ installment\_pct\_disp\_inc\_8 : int 4 2 2 2 3 2 3 2 2 4 ...  
## $ present\_residence\_since\_11 : int 4 2 3 4 4 4 4 2 4 2 ...  
## $ age\_in\_yrs\_13 : int 67 22 49 45 53 35 53 35 61 28 ...  
## $ number\_cards\_this\_bank\_16 : int 2 1 1 1 2 1 1 1 1 2 ...  
## $ no\_people\_liable\_for\_mntnance\_18: int 1 1 2 2 2 2 1 1 1 1 ...

tree <- hclustvar(X.quanti=vars\_quanti,X.quali=vars\_quali)  
plot(tree)  
rect.hclust(tree, k=10, border = 1:10)



# add colors randomly  
plot(as.phylo(tree), type = "fan",  
 tip.color = hsv(runif(15, 0.65, 0.95), 1, 1, 0.7),  
 edge.color = hsv(runif(10, 0.65, 0.75), 1, 1, 0.7),   
 edge.width = runif(20, 0.5, 3), use.edge.length = TRUE, col = "gray80")



summary.phylo(as.phylo(tree))

##   
## Phylogenetic tree: as.phylo(tree)   
##   
## Number of tips: 20   
## Number of nodes: 19   
## Branch lengths:  
## mean: 0.2405273   
## variance: 0.02336192   
## distribution summary:  
## Min. 1st Qu. Median 3rd Qu. Max.   
## 0.02006 0.10900 0.20250 0.36730 0.56050   
## No root edge.  
## First ten tip labels: duration\_month\_2   
## credit\_amount\_5  
## installment\_pct\_disp\_inc\_8  
## present\_residence\_since\_11  
## age\_in\_yrs\_13  
## number\_cards\_this\_bank\_16  
## no\_people\_liable\_for\_mntnance\_18  
## chk\_ac\_status\_1  
## credit\_history\_3  
## purpose\_4  
## No node labels.

#stab<-stability(tree,B=10) # Bootstrap 10 times  
#plot(stab,main="Stability of the partitions")  
#boxplot(stab$matCR[,1:12])  
#boxplot(stab$matCR)  
  
part<-cutreevar(tree,12)  
print(part)

##   
## Call:  
## cutreevar(obj = tree, k = 12)  
##   
##   
##   
## name description   
## "$var" "list of variables in each cluster"  
## "$sim" "similarity matrix in each cluster"  
## "$cluster" "cluster memberships"   
## "$wss" "within-cluster sum of squares"   
## "$E" "gain in cohesion (in %)"   
## "$size" "size of each cluster"   
## "$scores" "score of each cluster"

summary(part)

##   
## Call:  
## cutreevar(obj = tree, k = 12)  
##   
##   
##   
## Data:   
## number of observations: 1000  
## number of variables: 20  
## number of numerical variables: 7  
## number of categorical variables: 13  
## number of clusters: 12  
##   
## Cluster 1 :   
## squared loading  
## duration\_month\_2 0.69  
## credit\_amount\_5 0.77  
## purpose\_4 0.39  
##   
##   
## Cluster 2 :   
## squared loading  
## installment\_pct\_disp\_inc\_8 1  
##   
##   
## Cluster 3 :   
## squared loading  
## present\_residence\_since\_11 0.63  
## age\_in\_yrs\_13 0.63  
##   
##   
## Cluster 4 :   
## squared loading  
## number\_cards\_this\_bank\_16 0.8  
## credit\_history\_3 0.8  
##   
##   
## Cluster 5 :   
## squared loading  
## no\_people\_liable\_for\_mntnance\_18 0.64  
## personal\_status\_9 0.64  
##   
##   
## Cluster 6 :   
## squared loading  
## chk\_ac\_status\_1 0.63  
## savings\_ac\_bond\_6 0.63  
##   
##   
## Cluster 7 :   
## squared loading  
## p\_employment\_since\_7 0.76  
## job\_17 0.76  
##   
##   
## Cluster 8 :   
## squared loading  
## other\_debtors\_or\_grantors\_10 1  
##   
##   
## Cluster 9 :   
## squared loading  
## property\_type\_12 0.89  
## housing\_type\_15 0.89  
##   
##   
## Cluster 10 :   
## squared loading  
## other\_installment\_type\_14 1  
##   
##   
## Cluster 11 :   
## squared loading  
## telephone\_19 1  
##   
##   
## Cluster 12 :   
## squared loading  
## foreign\_worker\_20 1  
##   
##   
## Gain in cohesion (in %): 73.53

#part$scores

*duration\_month\_2* age\_in\_yrs\_13 credit\_amount\_5 installment\_pct\_disp\_inc\_8 chk\_ac\_status\_1 credit\_history\_3 savings\_ac\_bond\_6 purpose\_4 property\_type\_12 p\_employment\_since\_7