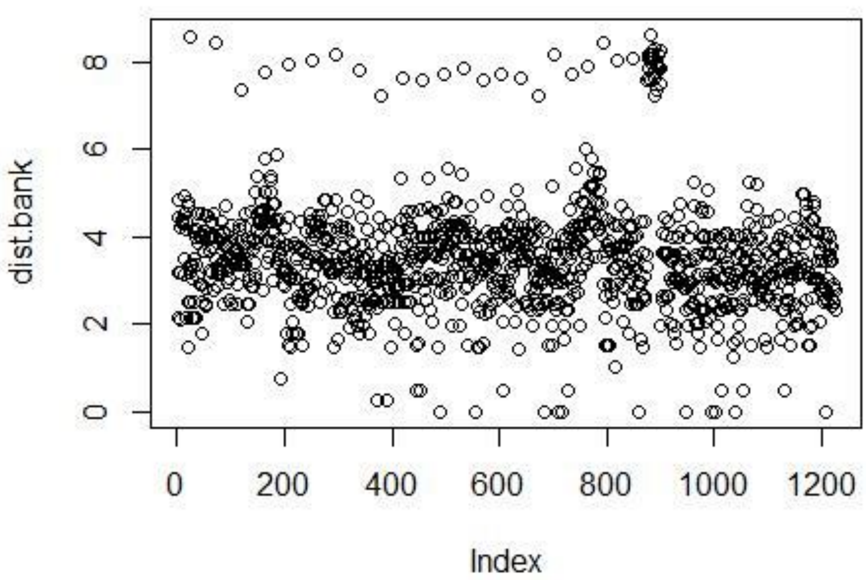
Assignment_5: Cluster Analysis

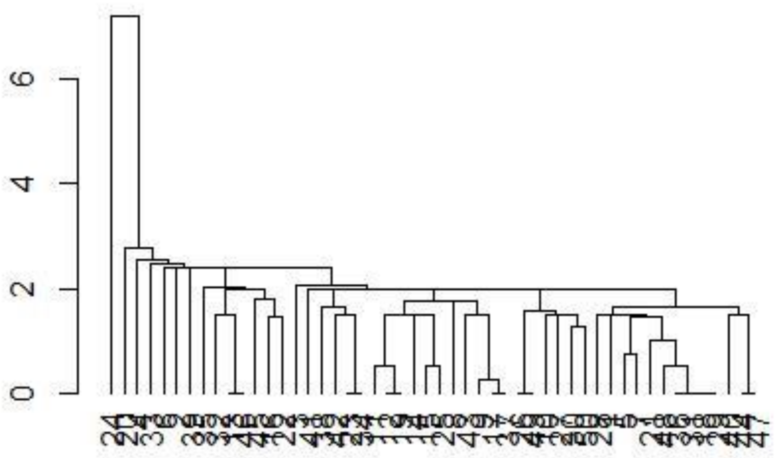
```
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(rmarkdown)
library(tinytex)
library(latexpdf)
library(latex2exp)
library(knitr)
#bankk <-
read_excel("C:/Users/Shamali/Desktop/RutgersSpring/multivariat/project/bank-
marketing-dataset/bankk.xlsx")


bank=read.csv("C:/Users/Shamali/Desktop/RutgersSpring/multivariat/project/New
folder/bank.csv",row.names=1,fill=TRUE)

#bank=bankk

#library(data.table)
#setDT(bank)

#Taking sample of data
#bank=bank[sample(.N,50)]
attach(bank)
dim(bank)

## [1] 50  7

#install.packages("cluster",lib="/Library/Frameworks/R.framework/Versions/3.5
/Resources/library")

library(cluster)

## Warning: package 'cluster' was built under R version 3.6.3

matstd.can <- scale(bank)
```

```
# Creating a (Euclidean) distance matrix of the standardized data

dist.bank <- dist(matstd.can, method="euclidean")

# Invoking hclust command (cluster analysis by single linkage method)
clusbank.nn <- hclust(dist.bank, method = "single") plot(dist.bank)
```




```
attach(bank)

## The following objects are masked from bank (pos = 4):
##
##     Default, Deposit, Education, Housing, Job, Loan, Marital

dim(bank)

## [1] 50  7

  # Invoking hclust command (cluster analysis by single linkage method)
clusbank.nn <- hclust(dist.bank, method = "single")

# Object "clusbank.nn" is converted into a object of class "dendrogram"
# in order to allow better flexibility in the (vertical) dendrogram plotting.

plot(as.dendrogram(clusbank.nn))
```

```
#Input dataset is a matrix where each row is a sample, and each column is
#a variable.Clustering is performed on a matrix (sample x sample)
#that provides the distance between samples. It can be computed using the
#dist() or the cor() function depending on the question.The hclust()
#function is used to perform the hierarchical clustering.
```



```
#Horizontal Dendrogram

dev.new()
plot(as.dendrogram(clusbank.nn))

# We will use agnes function as it allows us to select option for data
standardization, the distance measure and clustering algorithm in one single
function
#?agnes
(agn.bank <- agnes(bank, metric="euclidean", stand=TRUE, method = "single"))

## Call:     agnes(x = bank, metric = "euclidean", stand = TRUE, method =
"single")
## Agglomerative coefficient:  0.9339381
## Order of objects:
 ##  [1]  1  18  29  31  36  48  21  8  23  34  42  5   7   28  9   17  37  43  10  27  30  50  26
 49  40
## [26]44471112191415254138З 13463216223545332 396 4
20  24
## Height (summary):
##    Min. 1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
##   0.0000  0.5721  1.8604  2.0529  2.0325 25.5885
##
## Available components:
## [1] "order"     "height"    "ac"        "merge"     "diss"      "call"
## [7] "method"    "order.lab" "data"
```

```r
#View(agn.bank)

#  Description of cluster merging
agn.bank$merge
```

```
##       [,1] [,2]
##  [1,]  -44  -47
##  [2,]  -31  -36
##  [3,]  -292
##  [4,]  -26  -49
##  [5,]  -23  -34
##  [6,]  -183
##  [7,]  -17  -37
##  [8,]  -13  -46
##  [9,]  -12  -19
## [10,]   -9    7
## [11,]    6  -48
## [12,]  -15  -25
## [13,]  -119
## [14,]   -5   -7
## [15,]   11  -21
## [16,]  -30  -50
## [17,]  -27   16
## [18,]    5  -42
## [19,]  -14   12
## [20,]   13   19
## [21,]   10  -43
## [22,]   14  -28
## [23,]   -1   15
## [24,]  -40    1
## [25,]    8  -32
## [26,]  -10   17
## [27,]   26    4
## [28,]   25  -16
## [29,]   20  -41
## [30,]   27   24
## [31,]   30   29
## [32,]   21   31
## [33,]   -8   18
## [34,]   23   33
## [35,]   22   32
## [36,]   35  -38
## [37,]   28  -22
## [38,]   34   36
```

```
## [39,]   37  -35
## [40,]   39  -45
## [41,]   -2  -39
## [42,]   41   -6
## [43,]   38   -3
## [44,]   40  -33
## [45,]   43   44
## [46,]   45   42
## [47,]   46   -4
## [48,]   47  -20
## [49,]   48  -24

#Dendogram
plot(as.dendrogram(agn.bank))
```

```
#Interactive Plots

#plot(agn.bank,ask=TRUE)

#plot(agn.bank, which.plots=2)

#K-Means Clustering



#K-Means Clustering
#The purpose of clustering analysis is to identify patterns in your data and create groups
```

```
#according to those patterns. Therefore, if two points have similar
characteristics, that
#means they have the same patternand consequently, they belong to the same
group. By
#doing clustering analysis we should be able to check what #features usually
appear
#together and see what characterizes a group.
# Standardizing the data with scale()
matstd.employ <- scale(bank111sss)
# K-means
# Centers (k's) are numbers thus, 10 random sets are chosen
# Computing the percentage of variation accounted for. Two clusters
(kmeans2.employ <- kmeans(matstd.employ,2,nstart = 10))
```

```
bank111=na.omit(bank)
# Standardizing the data with scale()
matstd.bank <- scale(bank111)

# K-means, k=2, 3, 4
# Centers (k's) are numbers thus, 10 random sets are chosen

kmeans2.bank <- kmeans(matstd.bank,2) # this helps in omitting NA

kmeans2.bank

## K-means clustering with 2 clusters of sizes 27, 23
##
## Cluster means:
##          Job    Marital   Education    Housing       Loan    Default
Deposit
## 1 0.6921337 -0.4090295 -0.04611163 -0.2566536 -0.2602184 -0.1414214
0.2566536
## 2 -0.8125047 0.4801651 0.05413104 0.3012890 0.3054737 0.1660164 -
0.3012890
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26
##  1  1  2  1  1  1  1  2  2  1  1  1  2  1  1  2  2  1  1  1  1  2  2  2  2
1
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
##  2  1  1  1  1  2  2  2  2  1  2  2  1  1  1  1  2  2  2  2  2  1  1  2
##
## Within cluster sum of squares by cluster:
## [1] 138.3832 153.6727
##  (between_SS / total_SS =  14.9 %)
##
## Available components:
##
## [1] "cluster""centers""totss""withinss"
```

```
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
# Computing the percentage of variation accounted for. Two clusters
perc.var.2 <- round(100*(1 - kmeans2.bank$betweenss/kmeans2.bank$totss),1)
names(perc.var.2) <- "Perc. 2 clus"
perc.var.2
```

```
## Perc. 2 clus
##         85.1
```

```r
# Computing the percentage of variation accounted for. Three clusters
kmeans3.bank <- kmeans(matstd.bank,2) # this helps in omitting NA
kmeans3.bank
```

```
## K-means clustering with 2 clusters of sizes 26, 24
##
## Cluster means:
##          Job    Marital   Education    Housing       Loan     Default
Deposit
## 1 -0.5044163 0.5625219 0.07349814 0.4568998 0.3014775 0.1305428 -
0.2284499
## 2 0.5464510 -0.6093987 -0.07962299 -0.4949747 -0.3266006 -0.1414214
0.2474874
##
## Clustering vector:
##   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26
##   2  2  2  2  2  1  2  1  1  2  2  2  1  2  1  1  1  2  2  2  2  1  1  1  1
1
##  27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
##   1  2  2  2  2  1  1  1  1  2  1  1  2  2  2  2  1  1  1  1  1  2  1  1
##
## Within cluster sum of squares by cluster:
## [1] 173.1483 118.6563
##  (between_SS / total_SS =  14.9 %)
##
## Available components:
##
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"

## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
perc.var.3 <- round(100*(1 - kmeans3.bank$betweenss/kmeans3.bank$totss),1)

names(perc.var.3) <- "Perc. 3 clus"
perc.var.3
```

```
## Perc. 3 clus
##         85.1
```

```r
# Computing the percentage of variation accounted for. Four clusters
kmeans4.bank <- kmeans(matstd.bank,2) # this helps in omitting NA
kmeans4.bank

## K-means clustering with 2 clusters of sizes 11, 39
##
## Cluster means:
##          Job      Marital   Education     Housing       Loan     Default
## 1 -0.31098337 -0.027134637  0.26584866  0.26998623  1.8640133 -0.14142136
## 2  0.08771326  0.007653359 -0.07498295 -0.07614996 -0.5257473  0.03988807
##       Deposit
## 1 -0.08999541
## 2  0.02538332
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26
##  2  1  2  2  2  1  2  2  2  2  2  2  1  2  2  1  2  2  2  2  2  1  2  2  2
2
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
##  2  2  2  2  2  1  1  2  1  2  2  2  1  2  2  2  2  2  1  1  2  2  2  2
##
## Within cluster sum of squares by cluster:
## [1]  47.6737 242.5311
##  (between_SS / total_SS =  15.4 %)
##
## Available components:
##
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"

## [6] "betweenss"    "size"         "iter"         "ifault"

perc.var.4 <- round(100*(1 - kmeans4.bank$betweenss/kmeans4.bank$totss),1)
names(perc.var.4) <- "Perc. 4 clus"
perc.var.4

## Perc. 4 clus
##         84.6

# Saving four k-means clusters in a list
#
#kmeans4.bank$cluster == 1

is.na(kmeans4.bank)

##      cluster       centers         totss     withinss tot.withinss
betweenss
##        FALSE         FALSE         FALSE        FALSE        FALSE
FALSE
##         size          iter        ifault
##        FALSE         FALSE         FALSE
```

```r
clus1 <- matrix(names(kmeans4.bank$cluster[kmeans4.bank$cluster == 1]),
                ncol=1, nrow=length(kmeans4.bank$cluster[kmeans4.bank$cluster
== 1]))

colnames(clus1) <- "Cluster 1"
#clus1 <- matrix(names(kmeans4.bank$cluster[kmeans4.bank$cluster ==
1])),ncol=1, nrow=length(kmeans4.bank$cluster[kmeans4.bank$cluster == 1]))
clus2 <- matrix(names(kmeans4.bank$cluster[kmeans4.bank$cluster == 2]),
                 ncol=1, nrow=length(kmeans4.bank$cluster[kmeans4.bank$cluster
== 2]))
colnames(clus2) <- "Cluster 2"
clus3 <- matrix(names(kmeans4.bank$cluster[kmeans4.bank$cluster == 3]),
                ncol=1, nrow=length(kmeans4.bank$cluster[kmeans4.bank$cluster
== 3]))
colnames(clus3) <- "Cluster 3"
clus4 <- matrix(names(kmeans4.bank$cluster[kmeans4.bank$cluster == 4]),
                ncol=1, nrow=length(kmeans4.bank$cluster[kmeans4.bank$cluster
== 4]))
colnames(clus4) <- "Cluster 4"
list(clus1,clus2,clus3,clus4)
```

```
## [[1]]
##       Cluster 1
##  [1,] "2"
##  [2,] "6"
##  [3,] "13"
##  [4,] "16"
##  [5,] "22"
##  [6,] "32"
##  [7,] "33"
##  [8,] "35"
##  [9,] "39"
## [10,] "45"
## [11,] "46"
##
## [[2]]
##       Cluster 2
##  [1,] "1"
##  [2,] "3"
##  [3,] "4"
##  [4,] "5"
##  [5,] "7"
##  [6,] "8"
##  [7,] "9"
##  [8,] "10"
##  [9,] "11"
## [10,] "12"
## [11,] "14"
## [12,] "15"
## [13,] "17"
```

```
## [14,] "18"
## [15,] "19"
## [16,] "20"
## [17,] "21"
## [18,] "23"
## [19,] "24"
## [20,] "25"
## [21,] "26"
## [22,] "27"
## [23,] "28"
## [24,] "29"
## [25,] "30"
## [26,] "31"
## [27,] "34"
## [28,] "36"
## [29,] "37"
## [30,] "38"
## [31,] "40"
## [32,] "41"
## [33,] "42"
## [34,] "43"
## [35,] "44"
## [36,] "47"
## [37,] "48"
## [38,] "49"
## [39,] "50"
##
## [[3]]
##      Cluster 3
##
## [[4]]
##      Cluster 4
```