Assignment 4

```
#Necessary Packages
library(dplyr,quietly = TRUE)
library(ROracle,quietly = TRUE)
library(BBmisc,quietly = TRUE)
library(factoextra,quietly = TRUE)
library(fastcluster,quietly = TRUE)
library(cluster,quietly = TRUE)
library(kableExtra,quietly = TRUE)
library(DMwR,quietly = TRUE)
```

Introduction

This report focuses onproducing a model, using clustering methods to determine links iwthin the data. It analyses the links to determine the most signifcant and mot valuable one to the business.

```
#Data Preparation
set.seed(124)
loaddata <- function(user,password){
  drv <- dbDriver("Oracle")
  user <- "A13599863"
  password <- "A13599863"
  host <- "oracle.vittl.it.bond.edu.au"
   sid <- 'inft320'
  port <- '1521'
  dbname <- paste(
  "(DESCRIPTION=",
  "(ADDRESS=(PROTOCOL=tcp)(HOST=", host, ")(PORT=", port, "))",
  "(CONNECT_DATA=(SID=", sid, ")))", sep = "")
  statement <- "select * from brucedba.BankMarketing"

  db <- {dbConnect(Oracle(), user = user , password = password , host = host , dbname = dbname , port =
  loandata <- dbGetQuery(db , statement = statement )
  return(loandata)
}

loandata <- loaddata("A13599863","A13599863")
```

```
#Clean DataSet And Remove Unecessary Variable(s)
clean_loan <- select(loandata,-c("EMP_VAR_RATE","CONS_PRICE_IDX","CONS_CONF_IDX","EURIBOR3M"))
```

```
#Create Sample Data Set To Work With
loan.set <- sample_n(tbl = clean_loan, size = (41188*0.45))
```

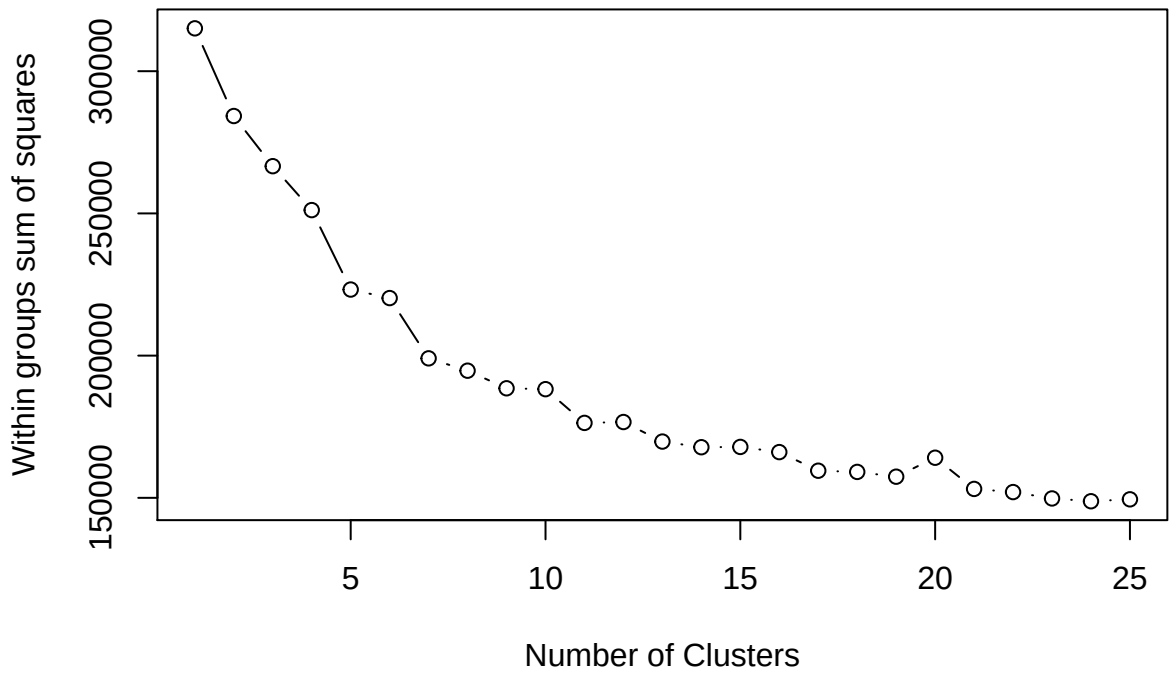
```
#Convert Factors to Numeric
factornames <- c("AGE","PDAYS","PREVIOUS","NR_EMPLOYED","CAMPAIGN","DURATION")

loan.set[,factornames] <- lapply(factornames, function (x) as.numeric(as.character(loan.set[,x])))
```

```
#Convert Factors to Categorical
factornames1 <- c("JOB","MARITAL","EDUCATION","DEFAULTCREDIT","HOUSING","LOAN","CONTACT","MONTH","DAY_O

loan.set[,factornames1] <- lapply(factornames1, function (x) as.factor(as.character(loan.set[,x])))
```

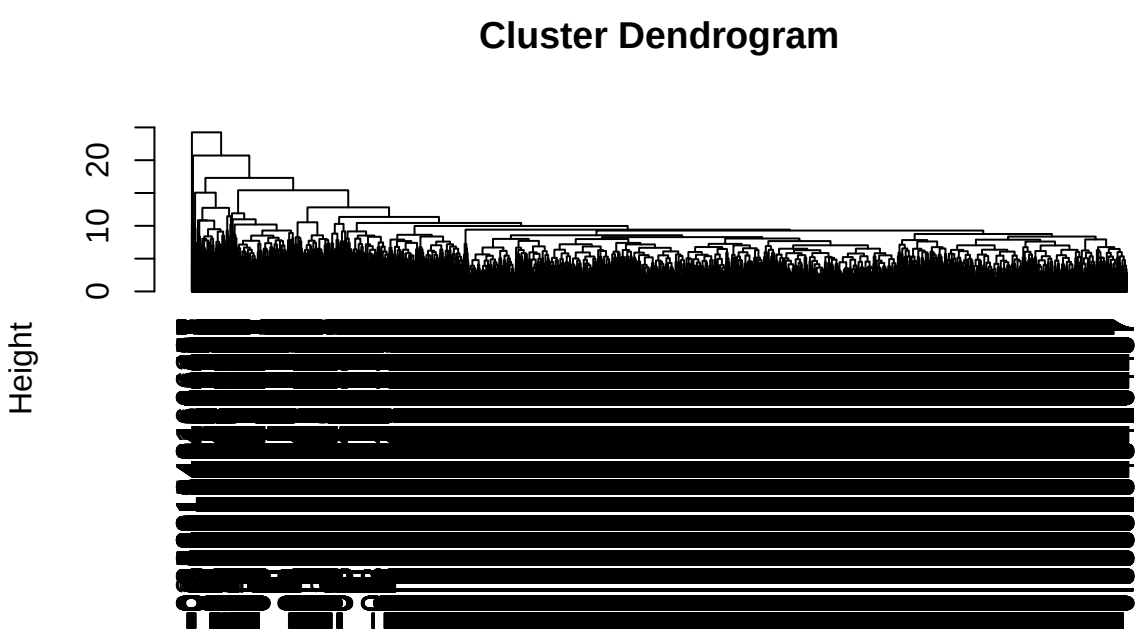
```r
#Convert Categorical Data To Numeric Data
factornames2 <- c("JOB","MARITAL","EDUCATION","DEFAULTCREDIT","HOUSING","LOAN","CONTACT","MONTH","DAY_O

loan.set[,factornames2] <- lapply(factornames2, function (x) as.numeric(as.factor(loan.set[,x])))
```

```r
#Scaling Data
numericdata <- loan.set

#Determine Mean And Standard Deviation
mean <- apply(numericdata, 2, mean)
sd <- apply(numericdata, 2, sd)

# Scale Data
scalefunc <- function(x){(x-mean(x, na.rm = T))/sd(x, na.rm = T)}
loan.norm<- apply(numericdata, 2, scalefunc)
```

```r
set.seed(453)
#Elbow Method To Select Optimal k Value

cplot <- function(data, nc=15){
  clus <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    clus[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, clus, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")}

cplot(loan.norm, nc=25)
```

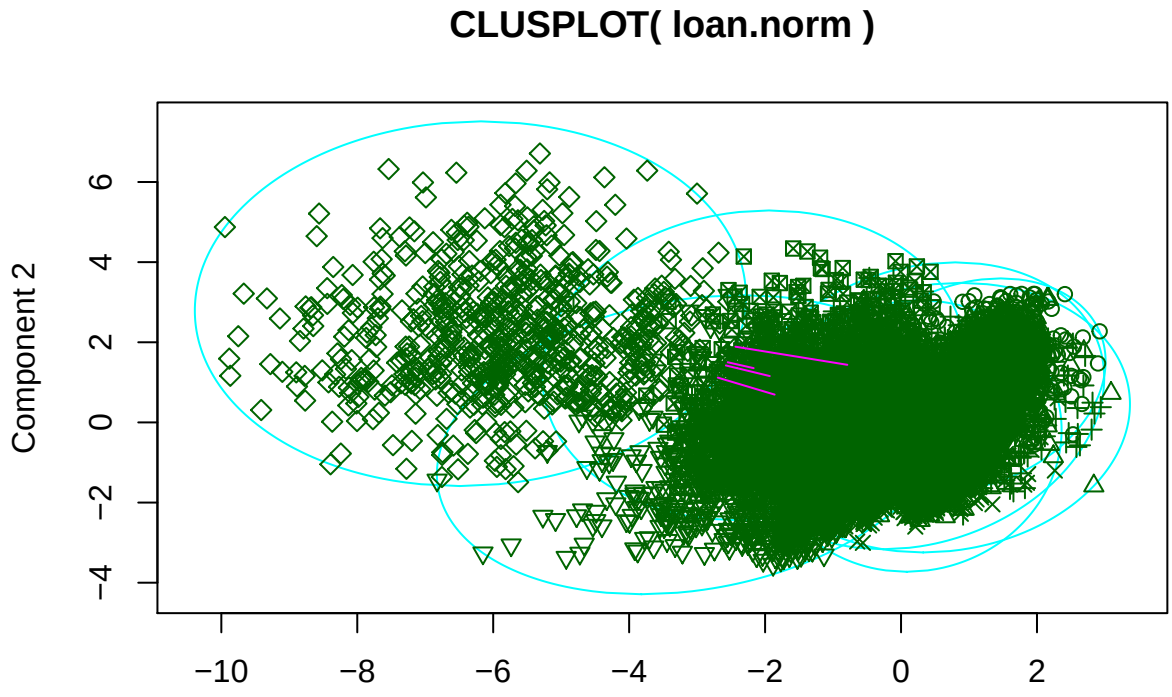



```
set.seed(453)
#Dendogram Plot
hcplotdist <- function(data, range = 1:17, target){
  hcplot <- hclust(d = dist(x=data[,range]),method="complete")
  return(plot(x=hcplot, hang = -1, labels = data [,target]))
}

hcplotdist(data = loan.norm,target="Y")
```

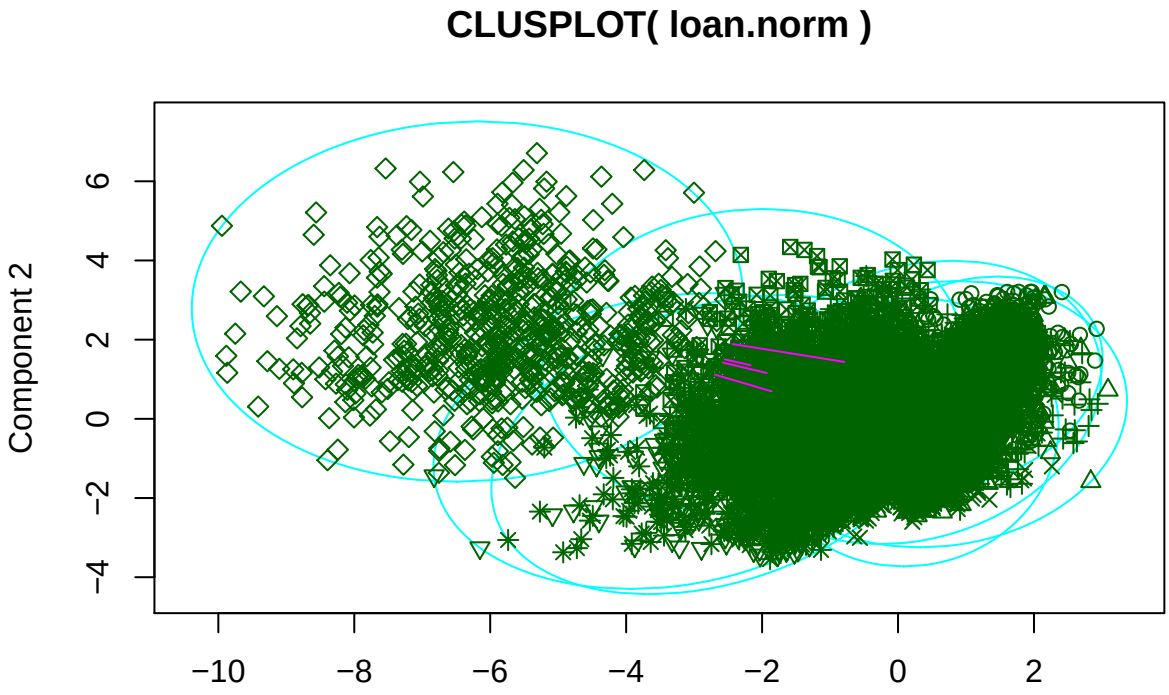


Looking at the Dendogram and Elbow Method Plot it seems that k = 7 is the optimal value for the plot. We will use k = 7 for the analysis of the clusters.

```
set.seed(453)
#Plot Cluster With K = 7
kml <- kmeans(x = loan.norm,center = 7)
csplot <- function(data,object){
  clusplot(data,object$cluster,color = TRUE, shade = TRUE,labels=2,lines=0)
}

clusplot(loan.norm, kml$cluster)
```




```
set.seed(453)
#Plot Cluster With K = 8
kml2 <- kmeans(x = loan.norm,center = 8)
csplot <- function(data,object){
  clusplot(data,object$cluster,color = TRUE, shade = TRUE,labels=2,lines=0)
}

clusplot(loan.norm, kml2$cluster)
```




```
set.seed(9874)
#Use K = 8 Cluster
attributes(kml2)
```

```
## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
##
## $class
## [1] "kmeans"
```

```
#Proportion Of Yes / NO Table (1 = No and 2 = Yes)
yntable<- table(loan.set$Y, kml2$cluster)
yntable
```

```
##
##        1    2    3    4    5    6    7    8
##   1 2952 2277 5756 3568  238  389   54 1243
##   2    1    1    0    0  428  114 1429   84
```

```
#Select One Group With More Y > N and Vice Versa
set7<- kml2$centers[7,]
set8<- kml2$centers[8,]

#Unscale Data With Below
un7 <- t((t(set7) * sd) + mean)
un8 <- t((t(set8) * sd) + mean)
```

```r
#Group Data To Tabulate

#Refer back to numeric dataframe and manually crosscheck values to determine values

Variables<- c("AGE", "JOB","MARITAL","DEFAULTCREDIT", "HOUSING", "LOAN","CONTACT","PREVIOUS","%YES")
Cluster1<- c(39, "blue-collar", "married","unknown","yes", "yes","cellular",1, "11%" )
Cluster2<- c(42, "technician", "married","no","yes", "no","telephone",2, "64%" )

clusgroup<- data.frame(cbind(Variables,Cluster1,Cluster2))
```

Results

```r
#Tabulate Data
kable(clusgroup) %>%
  kable_styling(bootstrap_options= "striped", font_size= 10, full_width = F)
```

Variables	Cluster1	Cluster2
AGE	39	42
JOB	blue-collar	technician
MARITAL	married	married
DEFAULTCREDIT	unknown	no
HOUSING	yes	yes
LOAN	yes	no
CONTACT	cellular	telephone
PREVIOUS	1	2
%YES	11%	64%

Discussion

In the table above there are two different groups that present unique business opportunities. Only 11% of customers in the first group are likely to take up the offer whereas customers in the second group are more likely to take up the offer at 64%.

People in the first group have an uknown default credit rating and hence consideration should be taken in asthere isn't a definite answer to if they will be able to pay back or subtantiate the payments. The telemarketing club should not focus on these customers until they have to as there customers are a potential risk due to their unknown default-credit rating and there low take up percentage rate.

Customers in the second group are a target group for the telemarketing company as they have a high take up rate and they also have no default credit rating. In addition these customers are not likely to default in nature unlike customers in group one. These customers have also been previosuly contacted and a potential change in the method of contact might help as they were being contacted via telephone and not cellular mobile which is now more common. These customers do not also have a loan and hence are a target customer type as they might be seeking a loan or be in need of a loan aswell. Resource would be better allocated to customers in group 2 rather than group 1 and hence resource allocation should be diverted into group 2 until they need to move to another group.

Reccomendation

Cluster 1:

1. The Telemarketing Club should be offering some sort of security to these customers as they are known to default and a type of insurance should be added onto their products.

2. Customers need to be checked and updated regularly for internal purposes to make sure they are keeping up to date with all payments and requirements. If customer falls into red zone, customer might need to be removed.

Cluster 2:

1. Keep in contact with these customers as they have a loan and due to the high percentage of take up there is room for an increased uptake. These are valubale customers for the company.
2. Incentivise these customers by offering lower interest rates or other tangible benefits.

Conclusion

The clusters present unique business opportunities for the Portuguese Telemarketing Club to become more efficient and run more effectively and hence increase profits.