# K Means

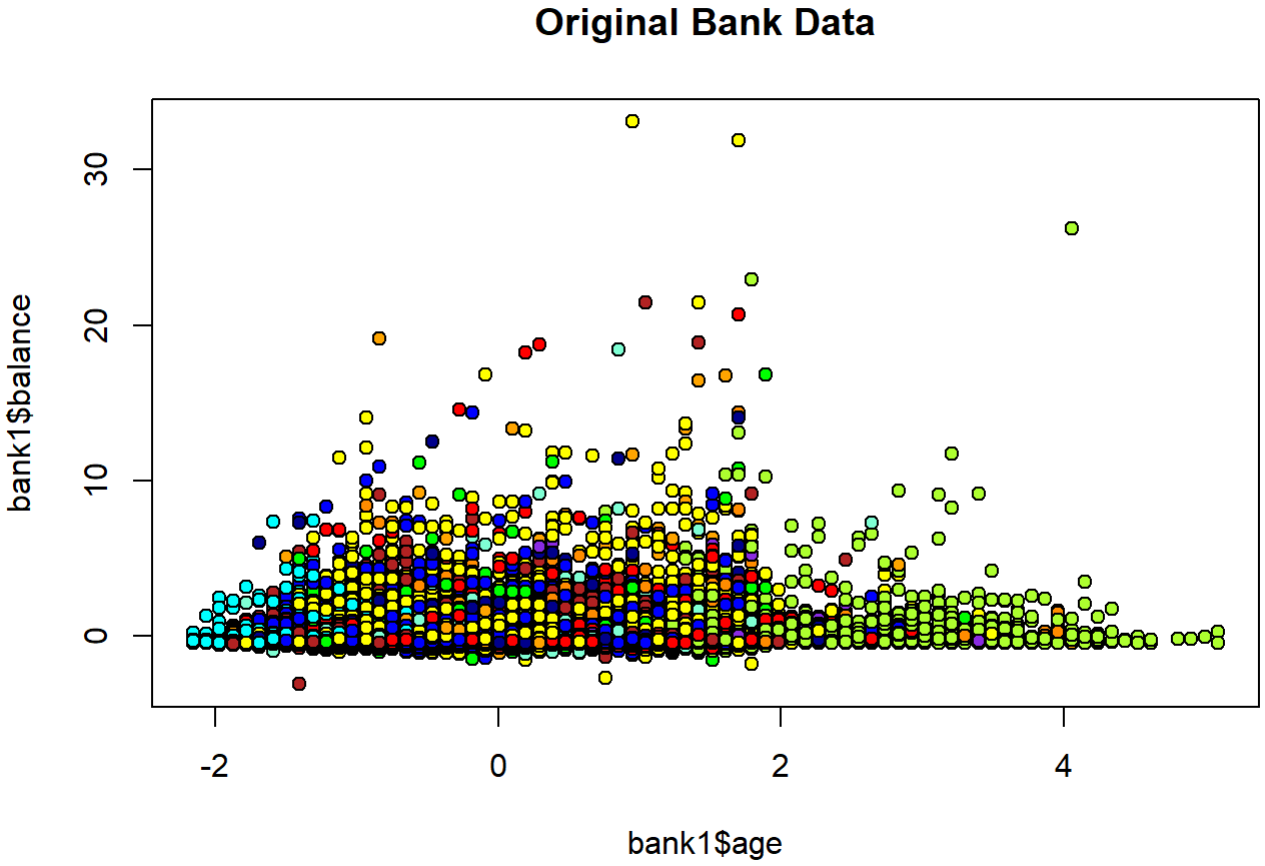
```
df <- read.csv("./bank-full.csv")
bank1 <- df[, c(1, 2, 6)]
bank1$job <- as.factor(bank1$job)
bank1[,-2] <- scale(bank1[,-2])
head(df)
```

```
##   age          job marital education default balance housing loan contact day
## 1  58   management married  tertiary      no    2143     yes   no unknown   5
## 2  44   technician  single secondary      no      29     yes   no unknown   5
## 3  33 entrepreneur married secondary      no       2     yes  yes unknown   5
## 4  47  blue-collar married   unknown      no    1506     yes   no unknown   5
## 5  33      unknown  single   unknown      no       1      no   no unknown   5
## 6  35   management married  tertiary      no     231     yes   no unknown   5
##   month duration campaign pdays previous poutcome Target
## 1   may      261        1    -1        0  unknown     no
## 2   may      151        1    -1        0  unknown     no
## 3   may       76        1    -1        0  unknown     no
## 4   may       92        1    -1        0  unknown     no
## 5   may      198        1    -1        0  unknown     no
## 6   may      139        1    -1        0  unknown     no
```

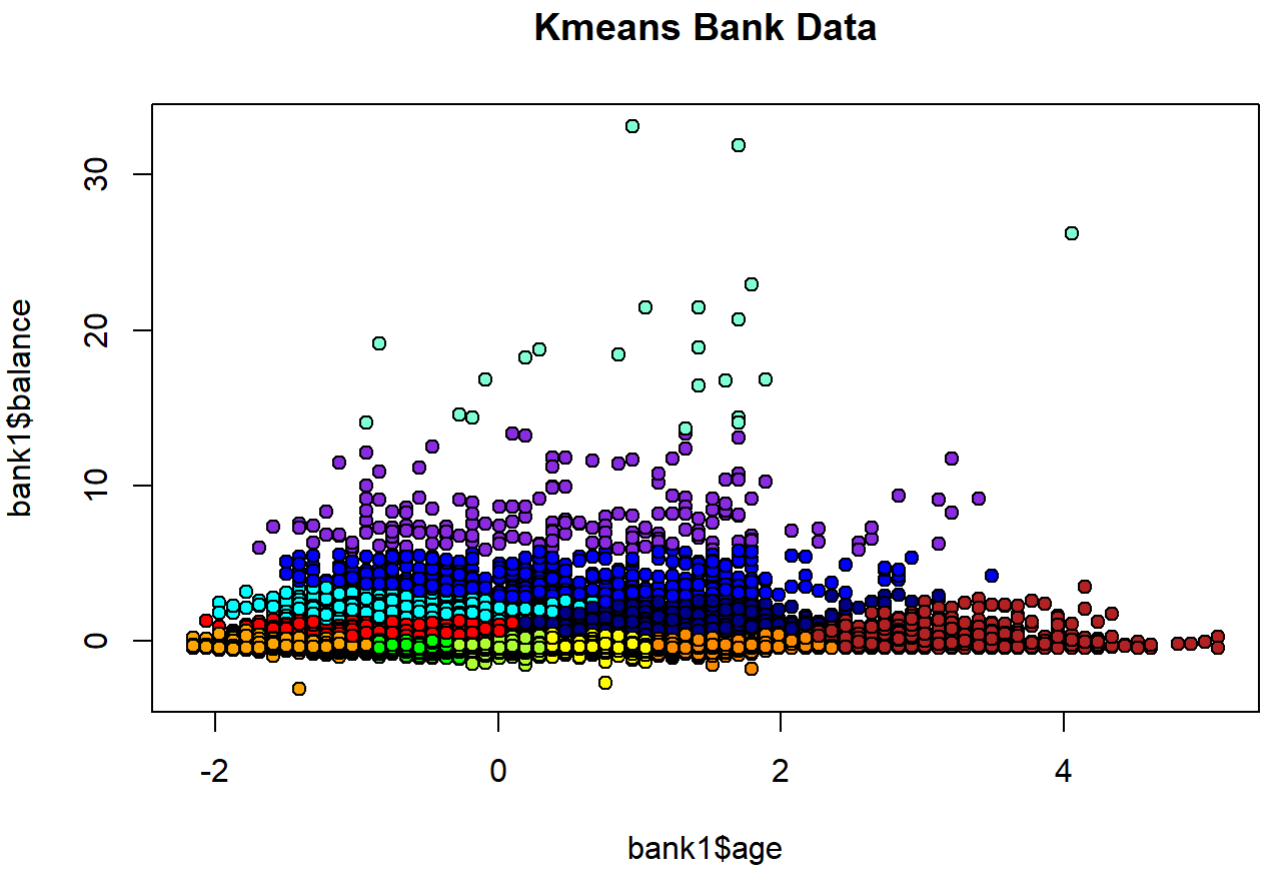
```
set.seed(1234)
bank1.km <- kmeans(bank1[, c(1, 3)], 12, nstart=25)

table(bank1.km$cluster, bank1$job)
```

```
##
##      admin. blue-collar entrepreneur housemaid management retired self-employed
##   1     347         611           89        52        825       3            124
##   2       6           5            3        28         33     612              7
##   3    1112        1679          174        89       1375       4            278
##   4     451         897          195       321        997    1184            179
##   5     857        1824          295       260       1384     134            250
##   6     893        1985          317       208       1574      45            262
##   7    1164        2099          276       156       2164      12            322
##   8       4           2            3         1          6       3              2
##   9      77         184           31        13        332       3             43
##   10     44          88           20        22        183      31             41
##   11    207         348           69        84        500     214             63
##   12      9          10           15         6         85      19              8
##
##      services student technician unemployed unknown
##   1       232      85        565         85      11
##   2         0       0         11          2      11
##   3       954     697       1336        232      18
##   4       335       0        672        172      72
##   5       673       4       1169        209      76
##   6       769      19       1404        230      40
##   7       944      78       1851        238      24
##   8         1       0          1          1       0
##   9        95      46        210         48       9
##   10       16       5         95         19       5
##   11      127       1        265         63      20
##   12        8       3         18          4       2
```

```
plot(bank1$age, bank1$balance, pch=21, cex=1, bg=c("red","firebrick","orange","darkorange","yell
ow","greenyellow","green","aquamarine","cyan","blue","darkblue","blueviolet")
     [unclass(bank1$job)], main="Original Bank Data")
```




```
plot(bank1$age, bank1$balance, pch=21, cex=1, bg=c("red","firebrick","orange","darkorange","yell
ow","greenyellow","green","aquamarine","cyan","blue","darkblue","blueviolet")
     [unclass(bank1.km$cluster)], main="Kmeans Bank Data")
```

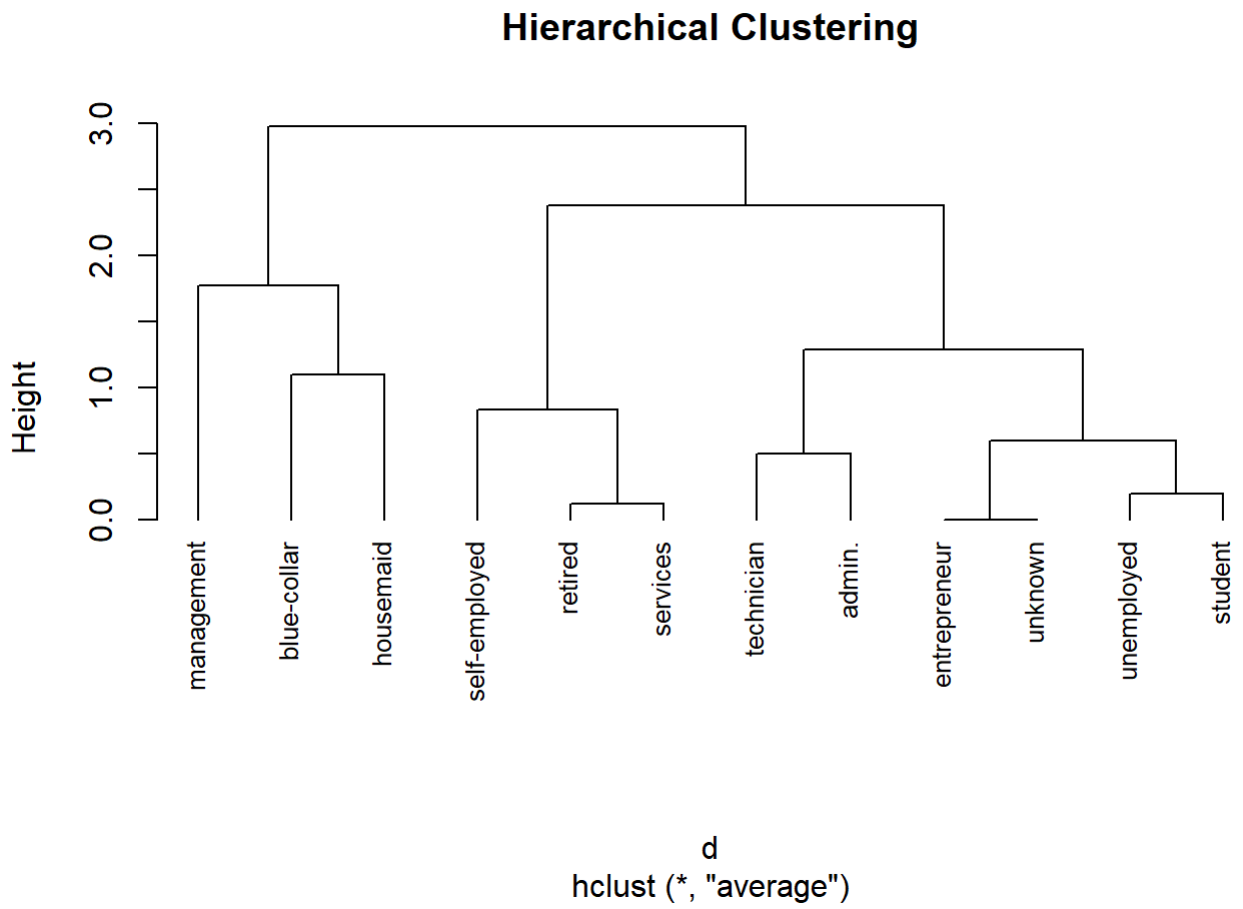



# Hierarchical

```
bank2 <- df[c(1,2,3,4,5,9,11,15,47,135,141,152), c(1, 2, 6)]
rownames(bank2) = unclass(bank2$job)
bank2[,-2] <- scale(bank2[,-2])

d <- dist(bank2)
```

```
## Warning in dist(bank2): NAs introduced by coercion
```

```
fit.average <- hclust(d, method="average")
plot(fit.average, hang=-1, cex=.8, main="Hierarchical Clustering")
```




```
cluster_cut <- cutree(fit.average, 12)
```

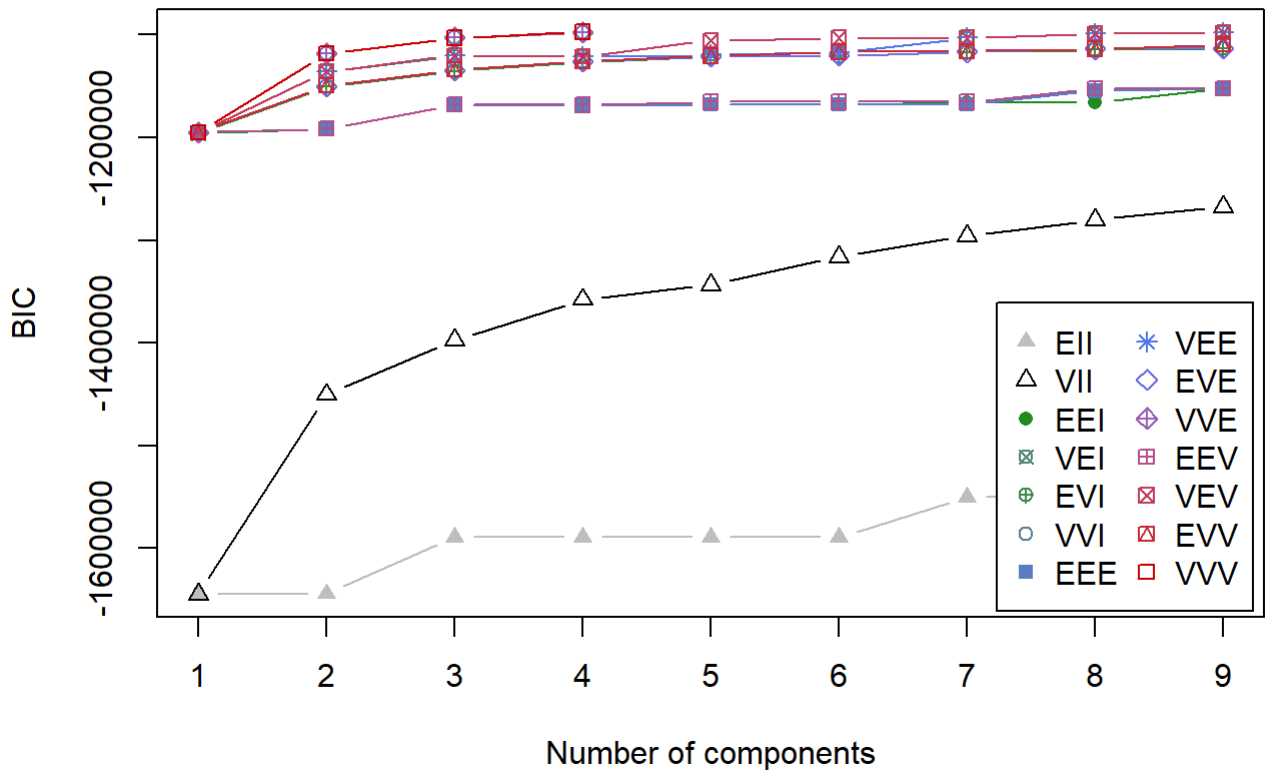
# Model Based Clustering

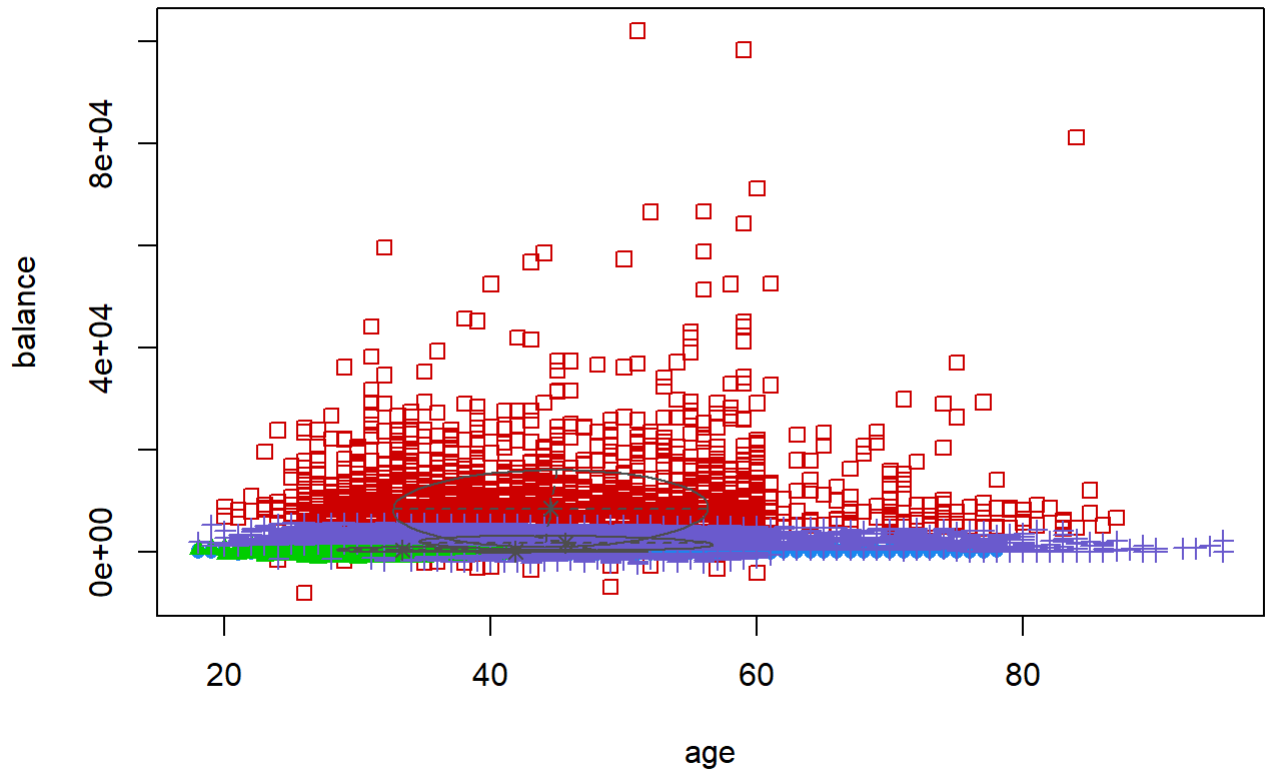
```
library(mclust)
```

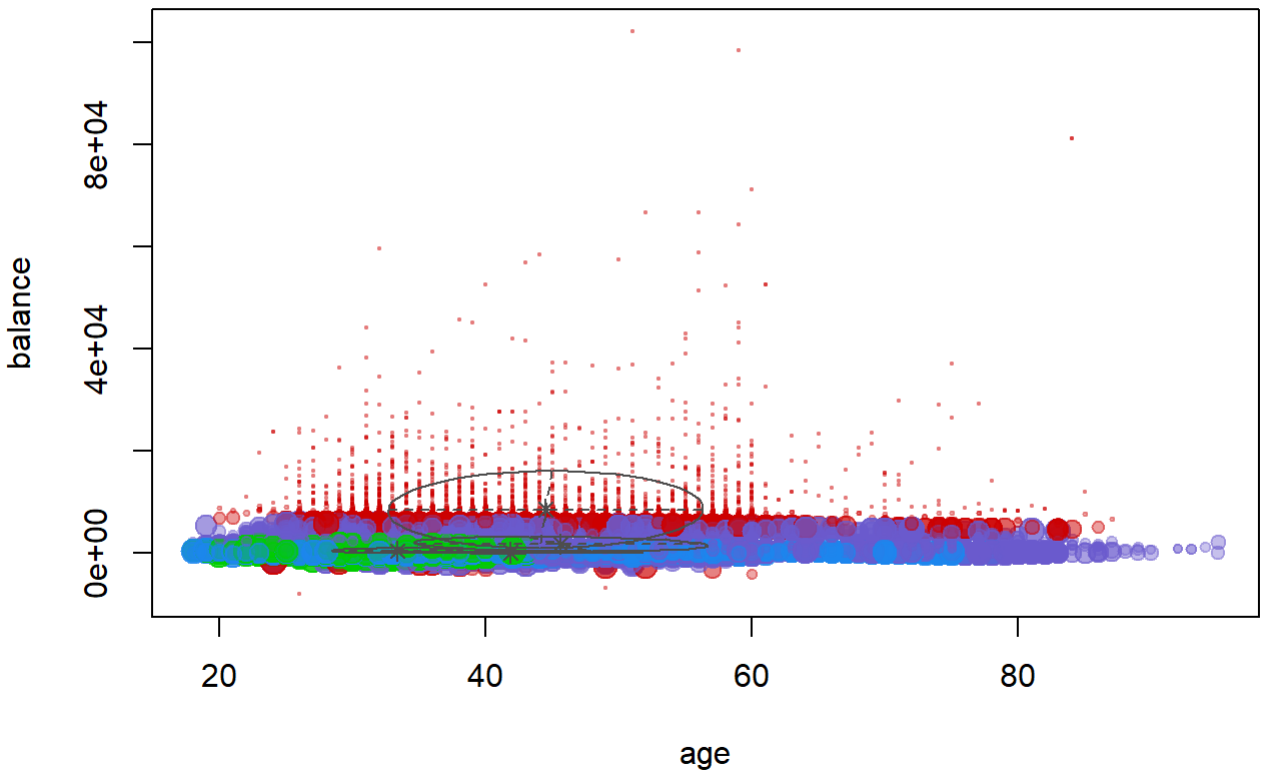
```
## Warning: package 'mclust' was built under R version 4.2.3
```

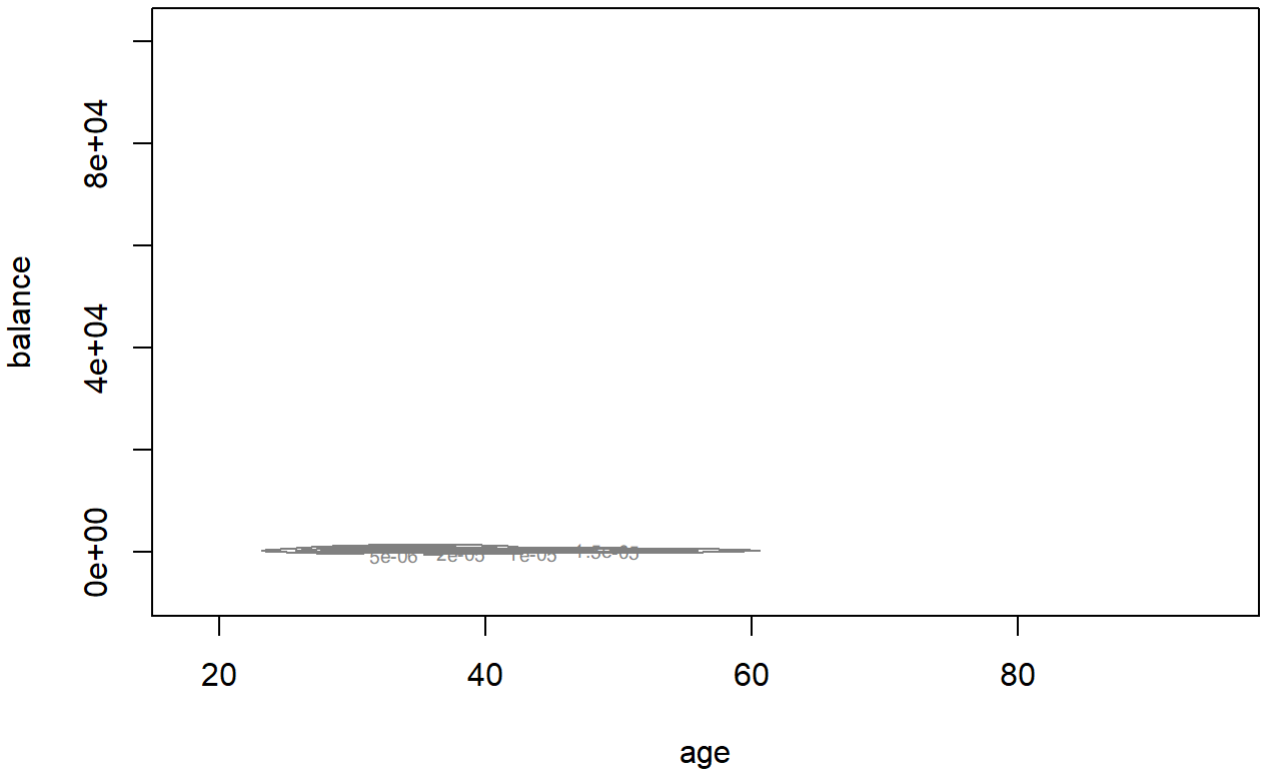
```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

```
bank3 <- df[, c(1, 6)]
fit <- Mclust(bank3)
plot(fit)
```


BIC
-1200000
-1400000
-1600000
1
2
3
4
5
6
7
8
9
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
VEE
EVE
VVE
EEV
VEV
EVV
VVV


balance
0e+00
4e+04
8e+04
20
40
60
80
age


balance
0e+00
4e+04
8e+04
20
40
60
80
age


balance
0e+00
4e+04
8e+04
20
40
60
80
age

```
summary(fit)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 4
## components:
##
##  log-likelihood     n df      BIC      ICL
##       -548469.2 45211 23 -1097185 -1120271
##
## Clustering table:
##     1     2     3     4
## 18489  2508 10701 13513
```

# Results

The results of K Means, Hierarchical, and Model Based clustering did not seem to accurately identify the 12 actual job clusters. This is most likely because job type cannot be predicted with only age and bank account balance, as well as the fact that the most of the actual clusters are highly mixed in the space of the two predictors. Still, the capabilities of the each algorithm were nicely shown. The K Means graph shows clearly distinct clusters, the Hierarchical dendrogram indicates some similarities between jobs that make some sense (e.g. technician and admin jobs are closely related), and the model based clustering results indicate that the best model for the data has four clusters.