Athens University of Economics and Business Dept. of Informatics
M.Sc. Program in Business Analytics
Course: Data Mining Techniques – Assignment 1
Instructor: Y.Kotidis, I.Filippidou(Assistant)

Measuring Data Similarity and Dissimilarity

Introduction

In this assignment, our team wants to access the similarity or dissimilarity objects by comparing the given attributes of the customers of a Portuguese banking institution and suggest 10 most similar other customers. The dataset given includes 43192 bank customer profiles with 8 attributes each.

The workflow for our dissimilarity matrix is calculated on the fly starting with creating a function “matfunc”.

Step 1: Read the dataset

Most of our variables are categorical with various factor levels. Thus, we read the dataset and setting the necessary factors as shown below.

```
bank <- read_delim("bank.csv",

       ";",

       escape_double = FALSE,

       col_types = cols(Class = col_factor(levels = c("no","yes")),

       Default = col_factor(levels = c("no","yes")),

       Education = col_factor(levels = c("primary", "secondary", "tertiary")),

       Housing = col_factor(levels = c("no","yes")),

       Job = col_factor(levels = c("admin.","unknown", "unemployed",
       "management", "housemaid", "entrepreneur", "student", "blue-collar",
       "self-employed","retired", "technician", "services")),

       Loan = col_factor(levels = c("no","yes")),

       Marital = col_factor(levels = c("married", "divorced", "single"))),
       trim_ws = TRUE)
```

Step 2: Transformation of variables

Pre-processing data step is to remove variable “Class” as we do not need it for analysis. Next, we transform the ordinal and binary variables to numbers. And we add a column “ID” for the observation number to each row incrementally. This row ID will identify the customers by the id number assigned and it will assist us in the tabulation of the dissimilarity matrix.

```
# remove class
bank$Class <- NULL

# make the ordinal and binary variables into numbers (1,2,3 for
Education and 0,1
# for Housing, Loan, and Default)

bank$Education <- as.numeric(bank$Education)
bank$Housing<-as.numeric(bank$Housing)-1
bank$Loan<-as.numeric(bank$Loan)-1
bank$Default<-as.numeric(bank$Default)-1

# add a row of ID numbers

bank$ID <- seq.int(nrow(bank))
```

Step 3: Create function

After we have done the necessary transformation and removing redundant variable, now we are ready to start creating the function. We will start off our function by creating the matrices for our numeric variables. As the formula suggested, we are taking the absolute value of the differences between the variables and divide it by the max-min values.

Within the function, we also tackle the categorical variables with the similar concept. A categorical (nominal) variable is one that has two or more categories, but there is no intrinsic ordering to these categories. Therefore, the dissimilarity matrices will be 0 and 1 for 0 if they are the same and 1 if the values are different.

Lastly, we create the matrix for ordinal variable. The ordinal variable is the absolute value of the difference between the current ID and all other rows divided by the max-min values of the ordinal column.

```
matfunc <- function(rowID)
{


  age_numeric <- matrix(abs(bank$Age[rowID]-bank$Age))/(max(bank$Age)-
min(bank$Age))

  balance_numeric <- matrix(abs(bank$Balance[rowID]-bank$Balance))/
                    (max(bank$Balance)-min(bank$Balance))


 default_cat <- matrix(abs(bank$Default[rowID]-bank$Default))
 housing_cat <- matrix(abs(bank$Housing[rowID]- bank$Housing))
 loan_cat <- matrix(abs(bank$Loan[rowID]-bank$Loan))
 job_cat <- matrix(ifelse(bank$Job[rowID]!=bank$Job, 1, 0))
 marital_cat <- matrix(ifelse(bank$Marital[rowID]!=bank$Marital, 1, 0))

education_ord <- matrix(abs(bank$Education[rowID]-bank$Education)/
                  (max(bank$Education)-min(bank$Education)))
```

Step 4: Compute data (dis-)similarity

We are almost done with the function. In the last 2 steps, we put all the matrices above and calculate the total dissimilarity which is the average of them all. After we are done with it, we take the top 11 because we are finding the top 10 similar customers including the studied object as well.

```
#average all of the calculations to find the total dissimilarity

 bank$Dissimilarity <- matrix(age_numeric + balance_numeric +
default_cat + housing_cat + loan_cat +
                        job_cat + marital_cat + education_ord)

#return the 10 rows with the smallest dissimilarity. Specify to print 11 because
#this will include the ID we are looking at

  top_n(bank, -11, bank$Dissimilarity)
}
```

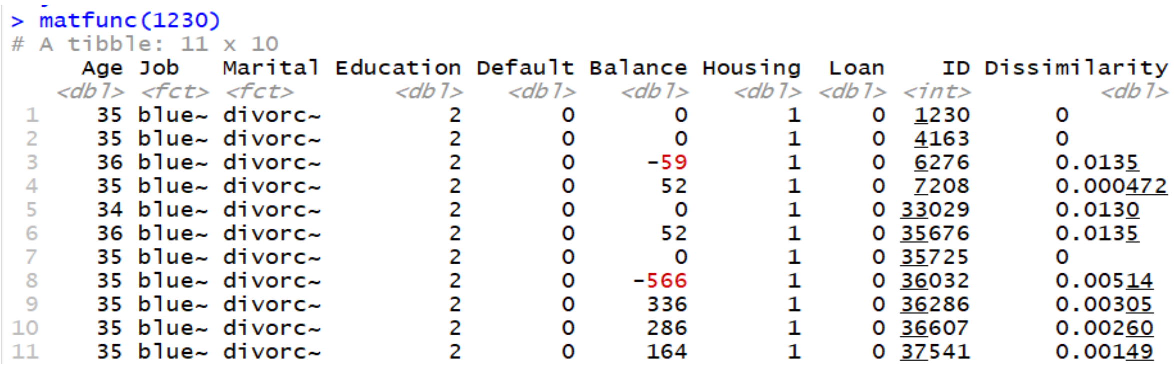
Results: Nearest Neighbour (NN) search

The function we've created allowed us to find the smallest dissimilar customers based on the characteristics based on customer id. Notice a trend on each group of customers below that people who are divorced generally have a low balance amount in the bank.

Together with customer id 1230, we will find some other customers that may be around the age of 35, also work as a blue collared worker and similar education background. For customer id 1230, it has the following characteristics:

Age: 35
Job: Blue Collar
Marital Status: Divorce
Education: Secondary
Balance: 0
Housing: Yes

This group has generally low balance amount in the bank, have housing which could means that they may require some form of finances to support either their daily necessities, or perhaps to pay off other bills. Or they do not have use this bank account as their dominant savings account. The bank may propose some loan scheme to them as they do not history to default payment or to encourage saving.

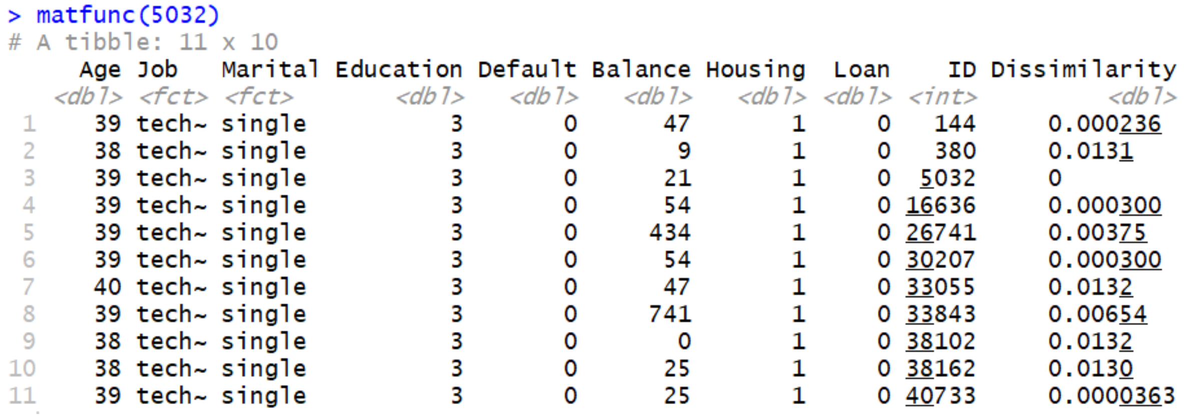


```
> matfunc(1230)
# A tibble: 11 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    35 blue~ divorc~          2       0       0       1     0  1230      0
 2    35 blue~ divorc~          2       0       0       1     0  4163      0
 3    36 blue~ divorc~          2       0     -59       1     0  6276      0.0135
 4    35 blue~ divorc~          2       0      52       1     0  7208      0.000472
 5    34 blue~ divorc~          2       0       0       1     0 33029      0.0130
 6    36 blue~ divorc~          2       0      52       1     0 35676      0.0135
 7    35 blue~ divorc~          2       0       0       1     0 35725      0
 8    35 blue~ divorc~          2       0    -566       1     0 36032      0.00514
 9    35 blue~ divorc~          2       0     336       1     0 36286      0.00305
10    35 blue~ divorc~          2       0     286       1     0 36607      0.00260
11    35 blue~ divorc~          2       0     164       1     0 37541      0.00149
```

For customer id 5032, it has the following characteristics:

Age: 39
Job: Technician
Marital Status: Single
Education: Tertiary
Balance: 47
Housing: Yes

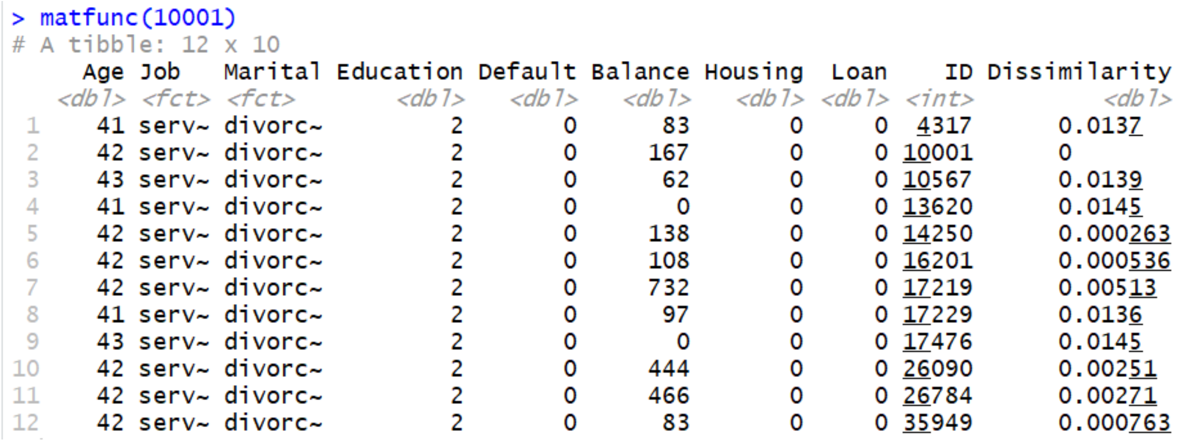
They are matured customers around the age of 39, single, with good tertiary education but still not much balance amount in the saving account. The bank may want to look into another matrix of credi card usage and expenditures. The bank can propose credit cards to these customers with targeted spending behavior. Also maybe loans such as cash line.



```
> matfunc(5032)
# A tibble: 11 x 10
     Age Job   Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct> <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    39 tech~ single          3       0      47       1     0   144     0.000236
 2    38 tech~ single          3       0       9       1     0   380     0.0131
 3    39 tech~ single          3       0      21       1     0  5032     0
 4    39 tech~ single          3       0      54       1     0 16636     0.000300
 5    39 tech~ single          3       0     434       1     0 26741     0.00375
 6    39 tech~ single          3       0      54       1     0 30207     0.000300
 7    40 tech~ single          3       0      47       1     0 33055     0.0132
 8    39 tech~ single          3       0     741       1     0 33843     0.00654
 9    38 tech~ single          3       0       0       1     0 38102     0.0132
10    38 tech~ single          3       0      25       1     0 38162     0.0130
11    39 tech~ single          3       0      25       1     0 40733     0.0000363
```

For customer id 10001, it has the following characteristics:

Age: 42
Job: Services
Marital Status: Divorce
Education: Secondary
Balance: 167
Housing: No

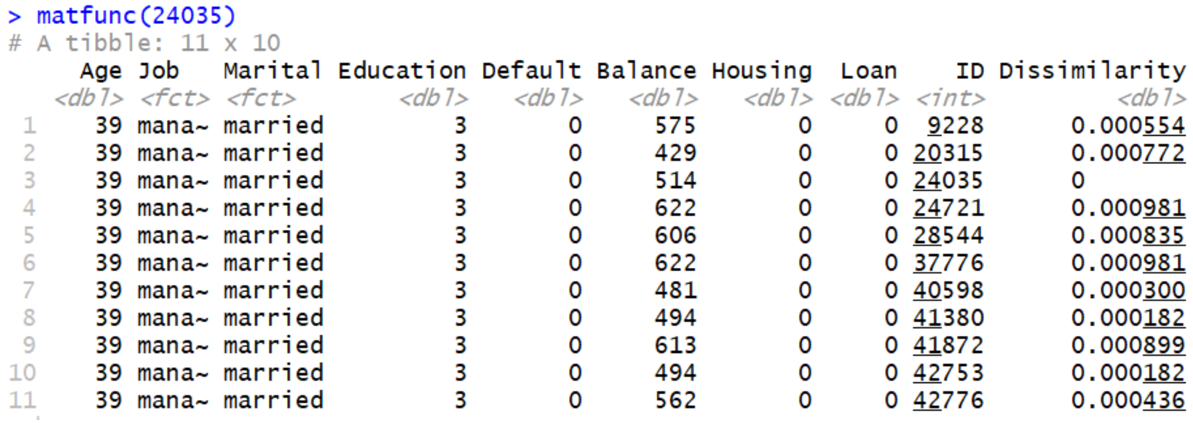


```
> matfunc(10001)
# A tibble: 12 x 10
     Age Job   Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct> <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    41 serv~ divorc~         2       0      83       0     0  4317      0.0137
 2    42 serv~ divorc~         2       0     167       0     0 10001      0
 3    43 serv~ divorc~         2       0      62       0     0 10567      0.0139
 4    41 serv~ divorc~         2       0       0       0     0 13620      0.0145
 5    42 serv~ divorc~         2       0     138       0     0 14250      0.000263
 6    42 serv~ divorc~         2       0     108       0     0 16201      0.000536
 7    42 serv~ divorc~         2       0     732       0     0 17219      0.00513
 8    41 serv~ divorc~         2       0      97       0     0 17229      0.0136
 9    43 serv~ divorc~         2       0       0       0     0 17476      0.0145
10    42 serv~ divorc~         2       0     444       0     0 26090      0.00251
11    42 serv~ divorc~         2       0     466       0     0 26784      0.00271
12    42 serv~ divorc~         2       0      83       0     0 35949      0.000763
```

Customer id 24035 are quite successful white collared management employee, married and have good tertiary education. However, we noticed that they do not have housing or exceptionally high savings. If this Portuguese bank has any investment products or home insurance products, we may try to introduce these customers to have a better security for their families and children.

Customer id 24035, it has the following characteristics:

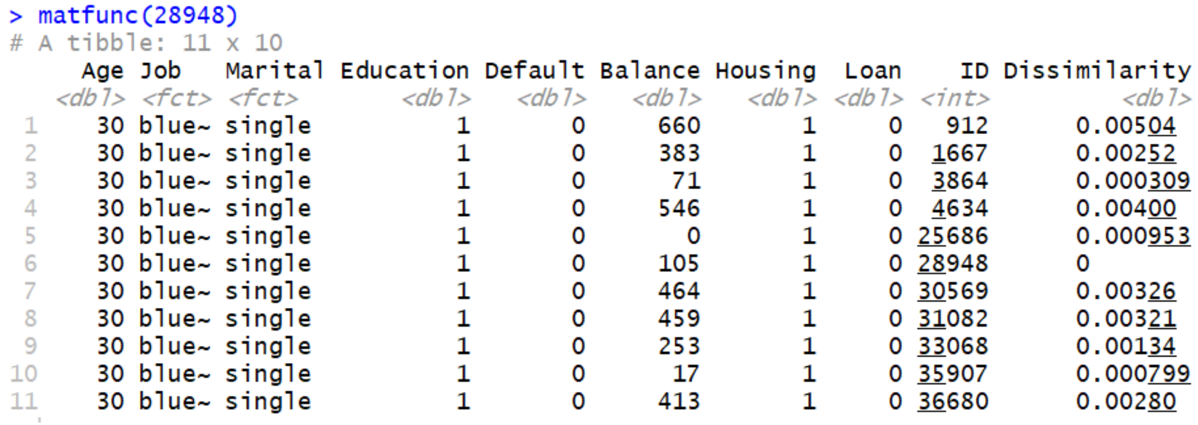
Age: 39
Job: Management
Marital Status: Married
Education: Tertiary
Balance: Around 514
Housing: No



```
> matfunc(24035)
# A tibble: 11 x 10
     Age Job   Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct> <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    39 mana~ married         3       0     575       0     0  9228      0.000554
 2    39 mana~ married         3       0     429       0     0 20315      0.000772
 3    39 mana~ married         3       0     514       0     0 24035      0       
 4    39 mana~ married         3       0     622       0     0 24721      0.000981
 5    39 mana~ married         3       0     606       0     0 28544      0.000835
 6    39 mana~ married         3       0     622       0     0 37776      0.000981
 7    39 mana~ married         3       0     481       0     0 40598      0.000300
 8    39 mana~ married         3       0     494       0     0 41380      0.000182
 9    39 mana~ married         3       0     613       0     0 41872      0.000899
10    39 mana~ married         3       0     494       0     0 42753      0.000182
11    39 mana~ married         3       0     562       0     0 42776      0.000436
```

For customer id 28948, it has the following characteristics:

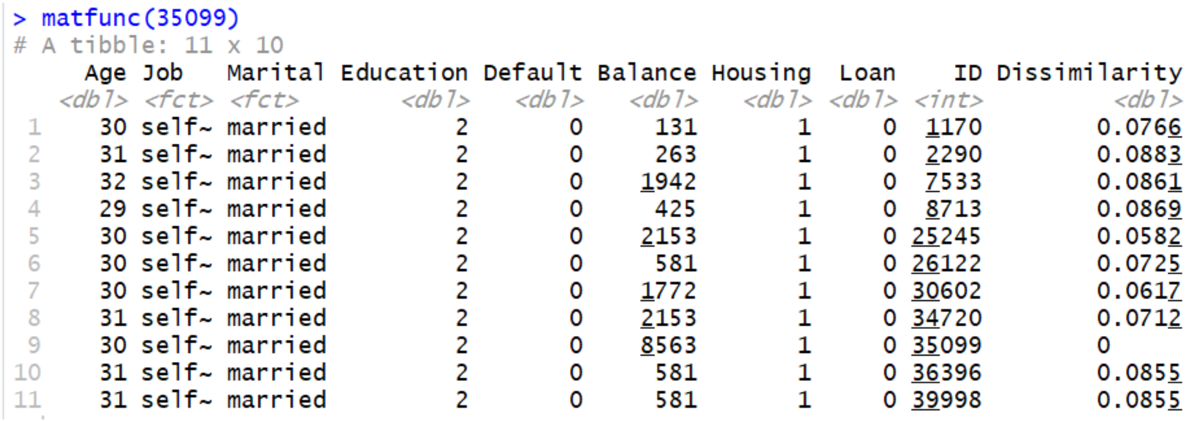
Age: 30
Job: Blue Collared
Marital Status: Single
Education: Primary
Balance: Around 105
Housing: Yes



```
> matfunc(28948)
# A tibble: 11 x 10
     Age Job   Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct> <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    30 blue~ single          1       0     660       1     0   912      0.00504 
 2    30 blue~ single          1       0     383       1     0  1667      0.00252 
 3    30 blue~ single          1       0      71       1     0  3864      0.000309
 4    30 blue~ single          1       0     546       1     0  4634      0.00400 
 5    30 blue~ single          1       0       0       1     0 25686      0.000953
 6    30 blue~ single          1       0     105       1     0 28948      0       
 7    30 blue~ single          1       0     464       1     0 30569      0.00326 
 8    30 blue~ single          1       0     459       1     0 31082      0.00321 
 9    30 blue~ single          1       0     253       1     0 33068      0.00134 
10    30 blue~ single          1       0      17       1     0 35907      0.000799
11    30 blue~ single          1       0     413       1     0 36680      0.00280 
```

For customer id 35099, it has the following characteristics:

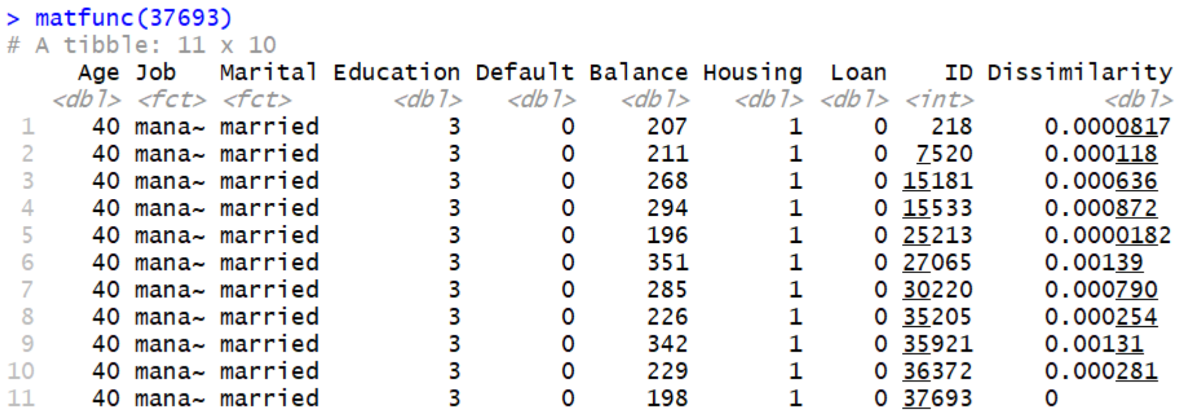
Age: 30
Job: Self-Employed
Marital Status: Married
Education: Secondary
Balance: Aaround a few thousand dollars
Housing: Yes



```
> matfunc(35099)
# A tibble: 11 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    30 self~ married          2       0     131       1     0  1170        0.0766
 2    31 self~ married          2       0     263       1     0  2290        0.0883
 3    32 self~ married          2       0    1942       1     0  7533        0.0861
 4    29 self~ married          2       0     425       1     0  8713        0.0869
 5    30 self~ married          2       0    2153       1     0 25245        0.0582
 6    30 self~ married          2       0     581       1     0 26122        0.0725
 7    30 self~ married          2       0    1772       1     0 30602        0.0617
 8    31 self~ married          2       0    2153       1     0 34720        0.0712
 9    30 self~ married          2       0    8563       1     0 35099        0     
10    31 self~ married          2       0     581       1     0 36396        0.0855
11    31 self~ married          2       0     581       1     0 39998        0.0855
```

For customer id 37693, it has the following characteristics:

Age: 40
Job: Management
Marital Status: Married
Education: Tertiary
Balance: around 198 dollars
Housing: Yes

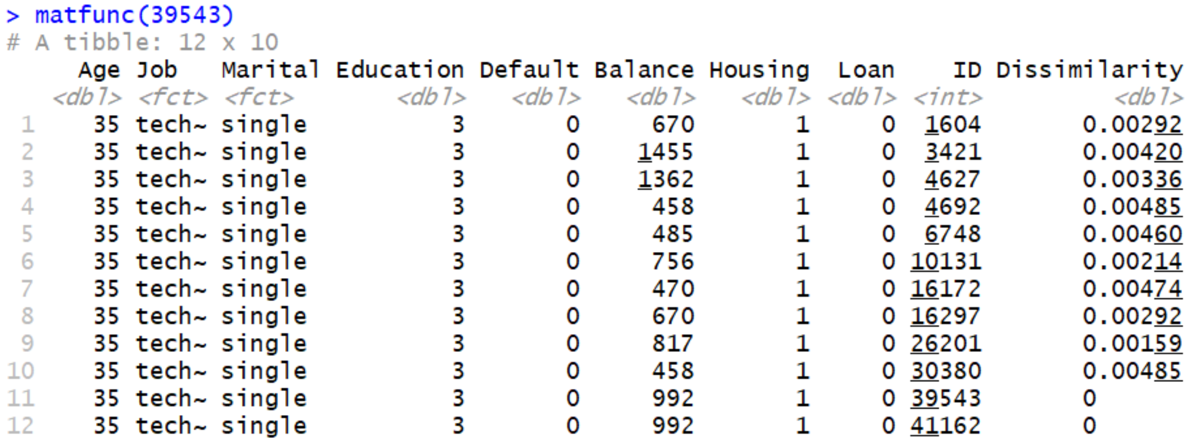


```
> matfunc(37693)
# A tibble: 11 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    40 mana~ married          3       0     207       1     0   218     0.0000817
 2    40 mana~ married          3       0     211       1     0  7520     0.000118 
 3    40 mana~ married          3       0     268       1     0 15181     0.000636 
 4    40 mana~ married          3       0     294       1     0 15533     0.000872 
 5    40 mana~ married          3       0     196       1     0 25213     0.0000182
 6    40 mana~ married          3       0     351       1     0 27065     0.00139  
 7    40 mana~ married          3       0     285       1     0 30220     0.000790 
 8    40 mana~ married          3       0     226       1     0 35205     0.000254 
 9    40 mana~ married          3       0     342       1     0 35921     0.00131  
10    40 mana~ married          3       0     229       1     0 36372     0.000281 
11    40 mana~ married          3       0     198       1     0 37693     0        
```

This group of customers has very similar profiling with customer id 5032 for job, marital status and education. But this group has better bank balance with the bank. We may want to try propose them similar product as the group with id 5032.

For customer id 39543, it has the following characteristics:

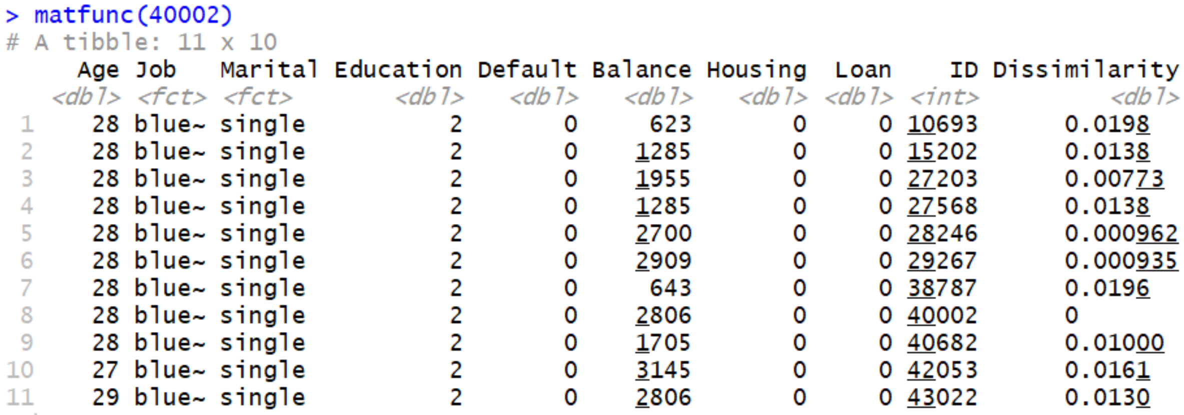
Age: 35
Job: Technician
Marital Status: Single
Education: Tertiary
Balance: 47
Housing: Yes



```
> matfunc(39543)
# A tibble: 12 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    35 tech~ single           3       0     670       1     0  1604       0.00292
 2    35 tech~ single           3       0    1455       1     0  3421       0.00420
 3    35 tech~ single           3       0    1362       1     0  4627       0.00336
 4    35 tech~ single           3       0     458       1     0  4692       0.00485
 5    35 tech~ single           3       0     485       1     0  6748       0.00460
 6    35 tech~ single           3       0     756       1     0 10131       0.00214
 7    35 tech~ single           3       0     470       1     0 16172       0.00474
 8    35 tech~ single           3       0     670       1     0 16297       0.00292
 9    35 tech~ single           3       0     817       1     0 26201       0.00159
10    35 tech~ single           3       0     458       1     0 30380       0.00485
11    35 tech~ single           3       0     992       1     0 39543       0
12    35 tech~ single           3       0     992       1     0 41162       0
```

This group has the highest saving/ balance amount with the bank. For customer id 40002, it has the following characteristics:

Age: 28
Job: Blue Collar
Marital Status: Single
Education: Secondary
Balance: Around 2806
Housing: Yes

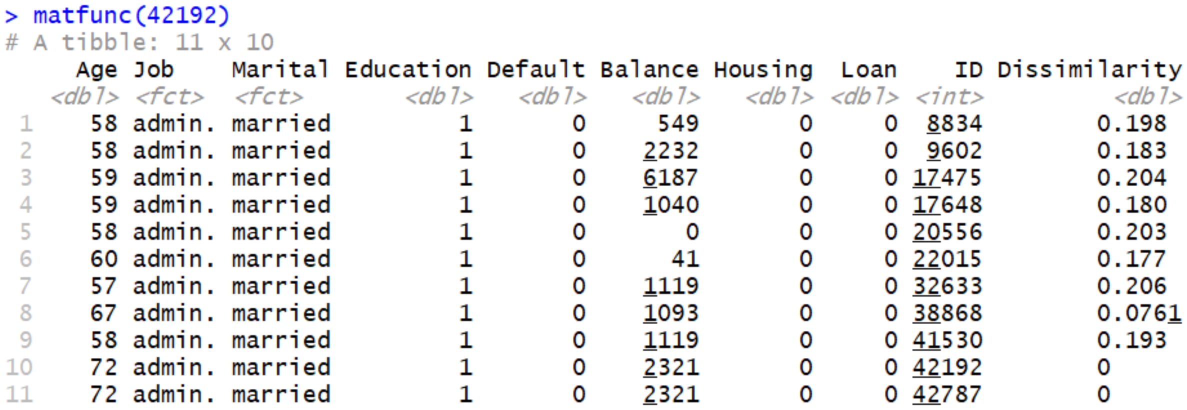


```
> matfunc(40002)
# A tibble: 11 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    28 blue~ single           2       0     623       0     0 10693      0.0198
 2    28 blue~ single           2       0    1285       0     0 15202      0.0138
 3    28 blue~ single           2       0    1955       0     0 27203      0.00773
 4    28 blue~ single           2       0    1285       0     0 27568      0.0138
 5    28 blue~ single           2       0    2700       0     0 28246      0.000962
 6    28 blue~ single           2       0    2909       0     0 29267      0.000935
 7    28 blue~ single           2       0     643       0     0 38787      0.0196
 8    28 blue~ single           2       0    2806       0     0 40002      0
 9    28 blue~ single           2       0    1705       0     0 40682      0.01000
10    27 blue~ single           2       0    3145       0     0 42053      0.0161
11    29 blue~ single           2       0    2806       0     0 43022      0.0130
```

This group of customers are approximately nearing the retirement age. They do not housing, does not default any loans. The bank could propose some form of retirement investment to these customers.

For customer id 42192, it has the following characteristics:

Age: 72
Job: Admin
Marital Status: Married
Education: Primary
Balance: 2321
Housing: No



```
> matfunc(42192)
# A tibble: 11 x 10
     Age Job    Marital Education Default Balance Housing  Loan    ID Dissimilarity
   <dbl> <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>         <dbl>
 1    58 admin. married         1       0     549       0     0  8834        0.198
 2    58 admin. married         1       0    2232       0     0  9602        0.183
 3    59 admin. married         1       0    6187       0     0 17475        0.204
 4    59 admin. married         1       0    1040       0     0 17648        0.180
 5    58 admin. married         1       0       0       0     0 20556        0.203
 6    60 admin. married         1       0      41       0     0 22015        0.177
 7    57 admin. married         1       0    1119       0     0 32633        0.206
 8    67 admin. married         1       0    1093       0     0 38868        0.0761
 9    58 admin. married         1       0    1119       0     0 41530        0.193
10    72 admin. married         1       0    2321       0     0 42192        0
11    72 admin. married         1       0    2321       0     0 42787        0
```