Credit Prediction

```
library("FactoMineR")
library("factoextra")
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
library("ggplot2")
library("corrplot")
```

```
## corrplot 0.84 loaded
```

Our main analysis is in the python file of the project. However, we wanted to use R for PCA and MCA, to see if we could delete some redundant variables.

We first load data that has been pre-processed in python. Dates have been transformed to age in years or in months for variables "Prod_Decision_Date" and "Prod_Closed_Date". We then transform all columns later used for the MCA into factors.

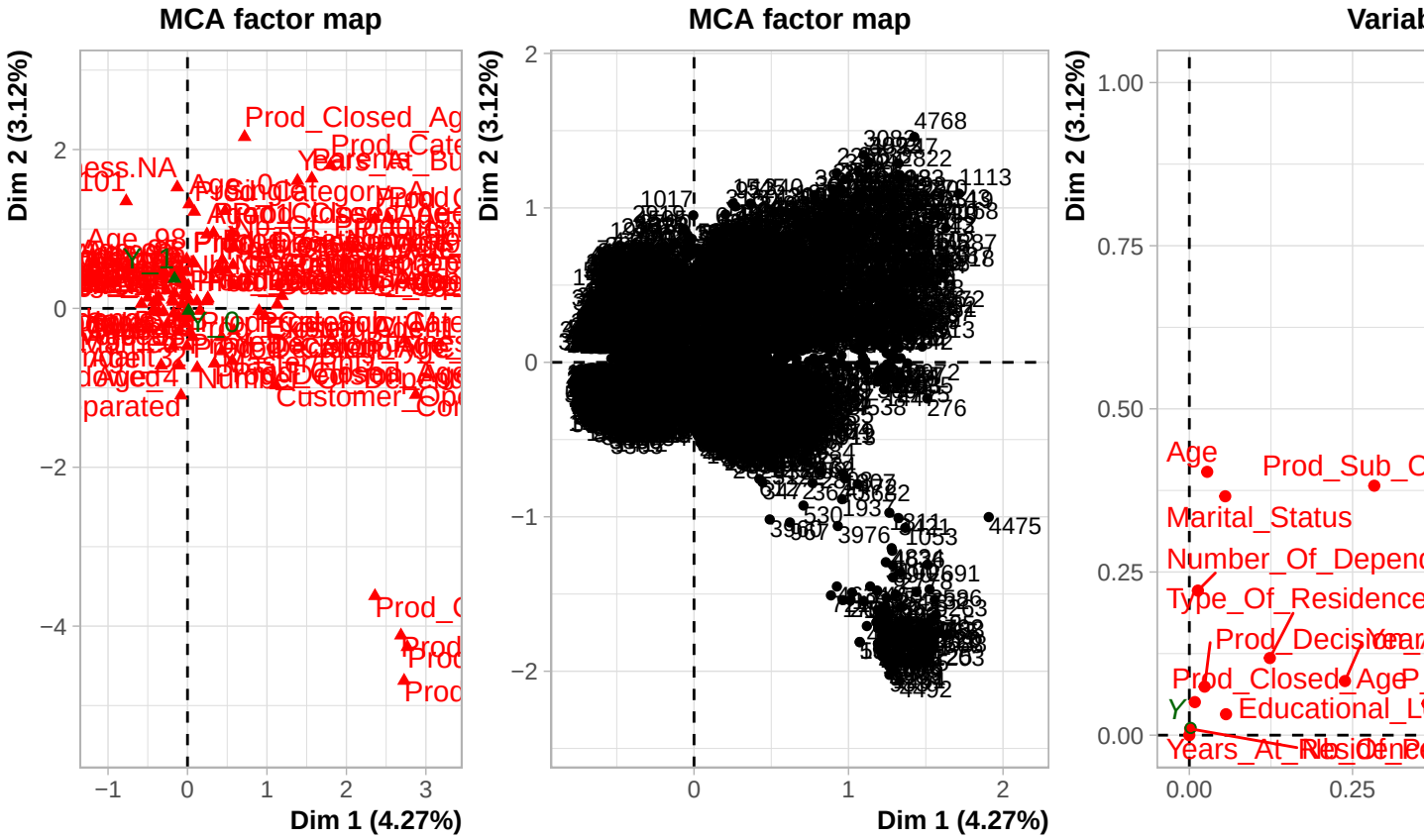
```
raw_data <- read.table("raw_data_age_transformed.csv",
                   sep = ",",
                   header = TRUE,
                   row.names = 1)
raw_data$Age <- as.factor(raw_data$Age)
raw_data$Customer_Open_Age <- as.factor(raw_data$Customer_Open_Age)
raw_data$Prod_Decision_Age <- as.factor(raw_data$Prod_Decision_Age)
raw_data$Prod_Closed_Age <- as.factor(raw_data$Prod_Closed_Age)
raw_data$Y <- as.factor(raw_data$Y)
raw_data$Number_Of_Dependant <- as.factor(raw_data$Number_Of_Dependant)
raw_data$Years_At_Residence <- as.factor(raw_data$Years_At_Residence)
raw_data$Years_At_Business <- as.factor(raw_data$Years_At_Business)
raw_data$Nb_Of_Products <- as.factor(raw_data$Nb_Of_Products)
```

Qualitative variables are all the factor variables used in the MCA

```
quali_data = raw_data[,sapply(raw_data, is.factor)]
```

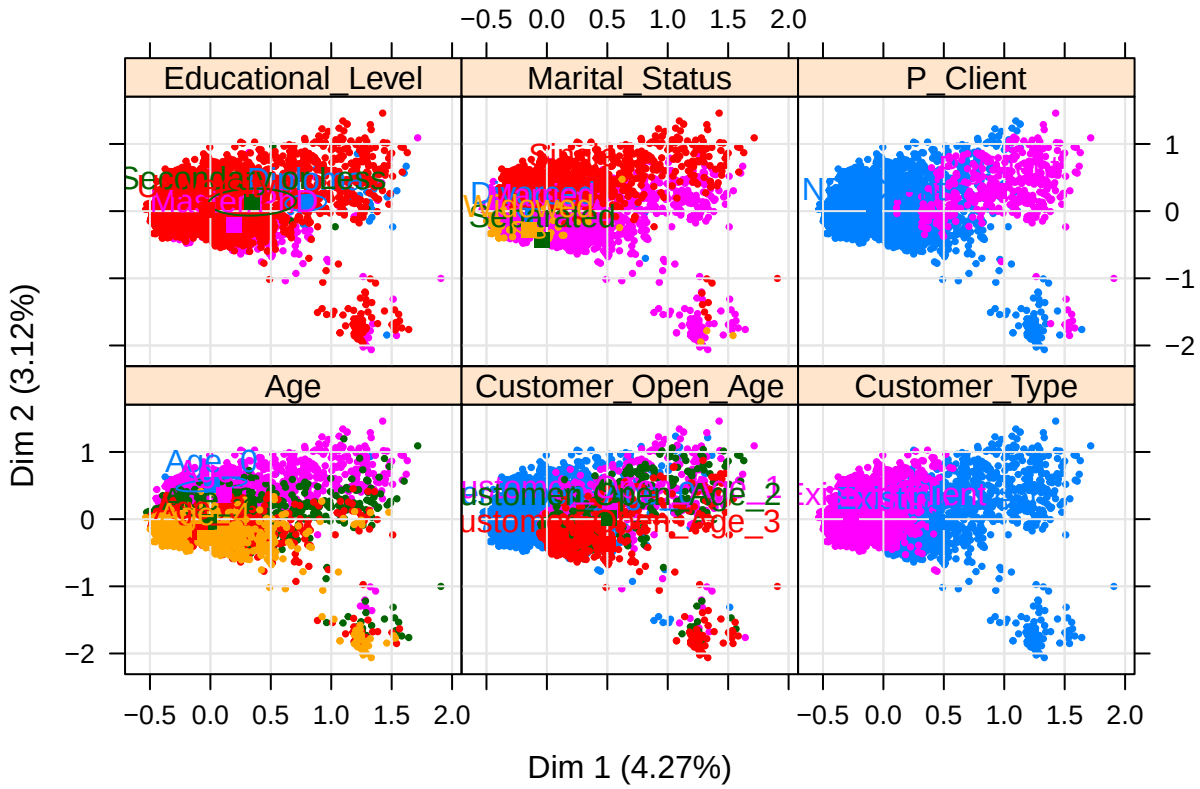
We plot here the representation of variables on the two main dimensions. The first one seems more focused on the type of client while the second one is focused on age that is naturally more correlated to a marital status or a number of dependencies. Y is used as supplementary variable.

```
res.mca <- MCA(quali_data, ncp = 5, quali.sup = 1)
```

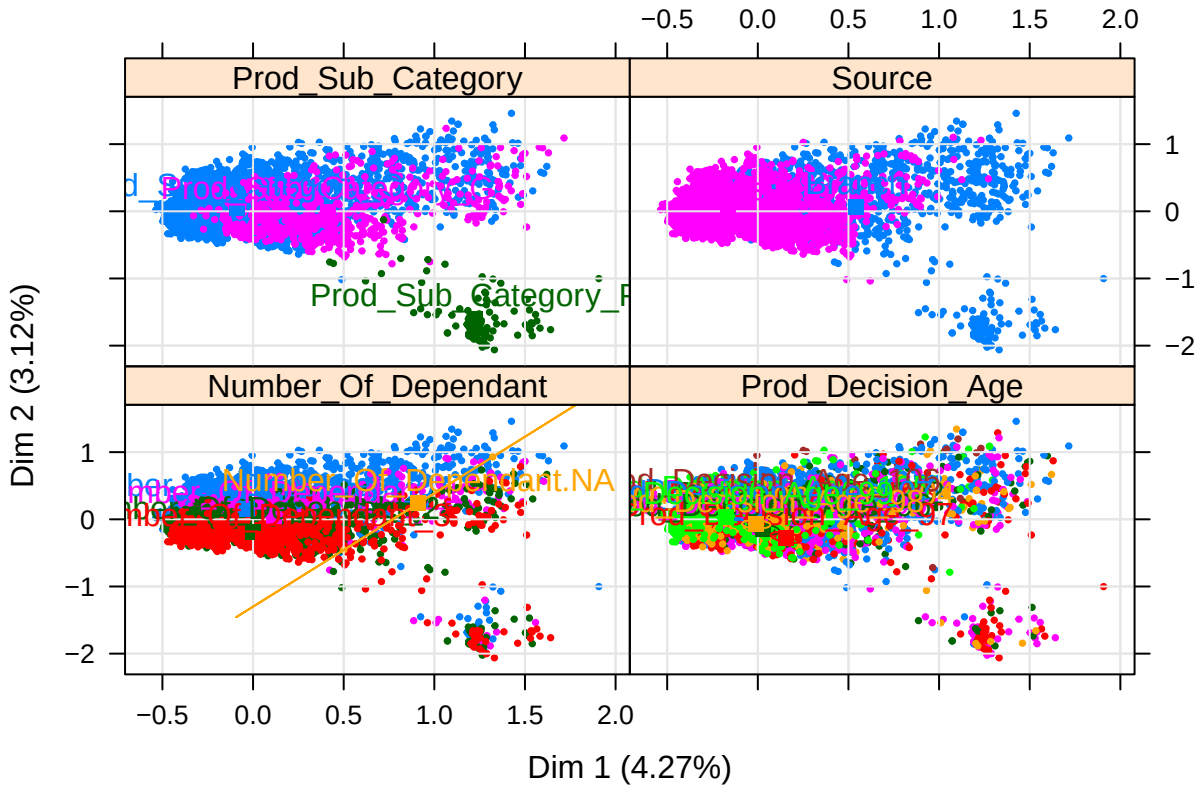


We plot here the repartition of variables given different categories to see if some variables are explained or not by one of the two first dimensions.

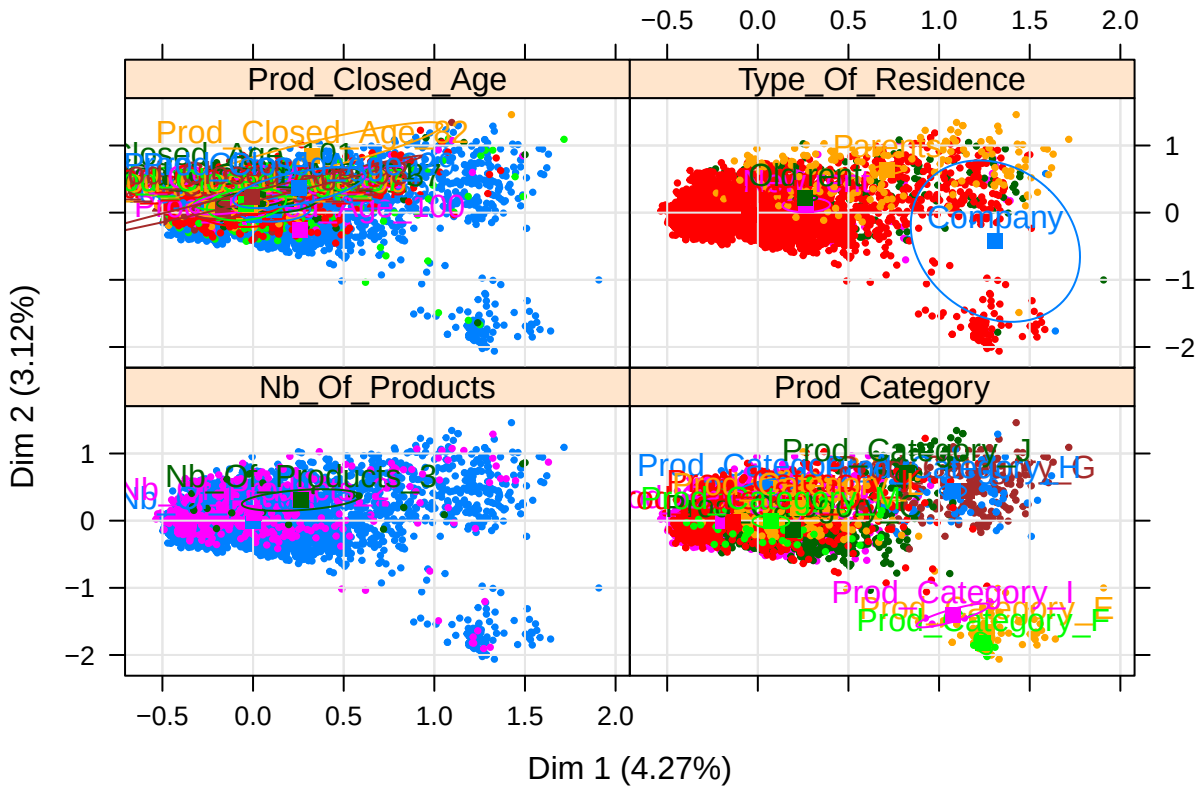
```
plotellipses(res.mca, keepvar = c("Customer_Type", "Customer_Open_Age", "Age", "P_Client", "Educational_
```




```
plotellipses(res.mca, keepvar = c("Number_Of_Dependant", "Prod_Sub_Category", "Source", "Prod_Decision_
```



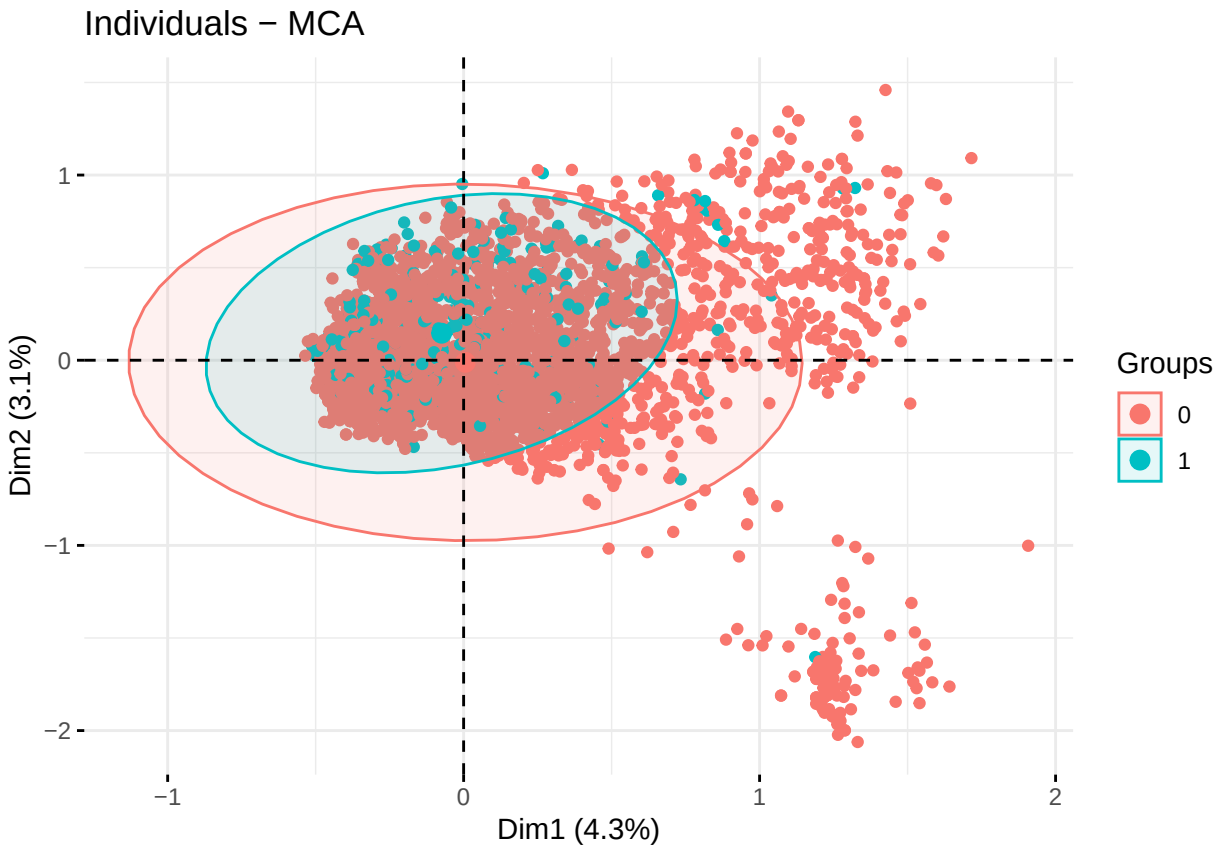
```
plotellipses(res.mca, keepvar = c("Type_Of_Residence", "Nb_Of_Products", "Prod_Closed_Age", "Prod_Categ
```



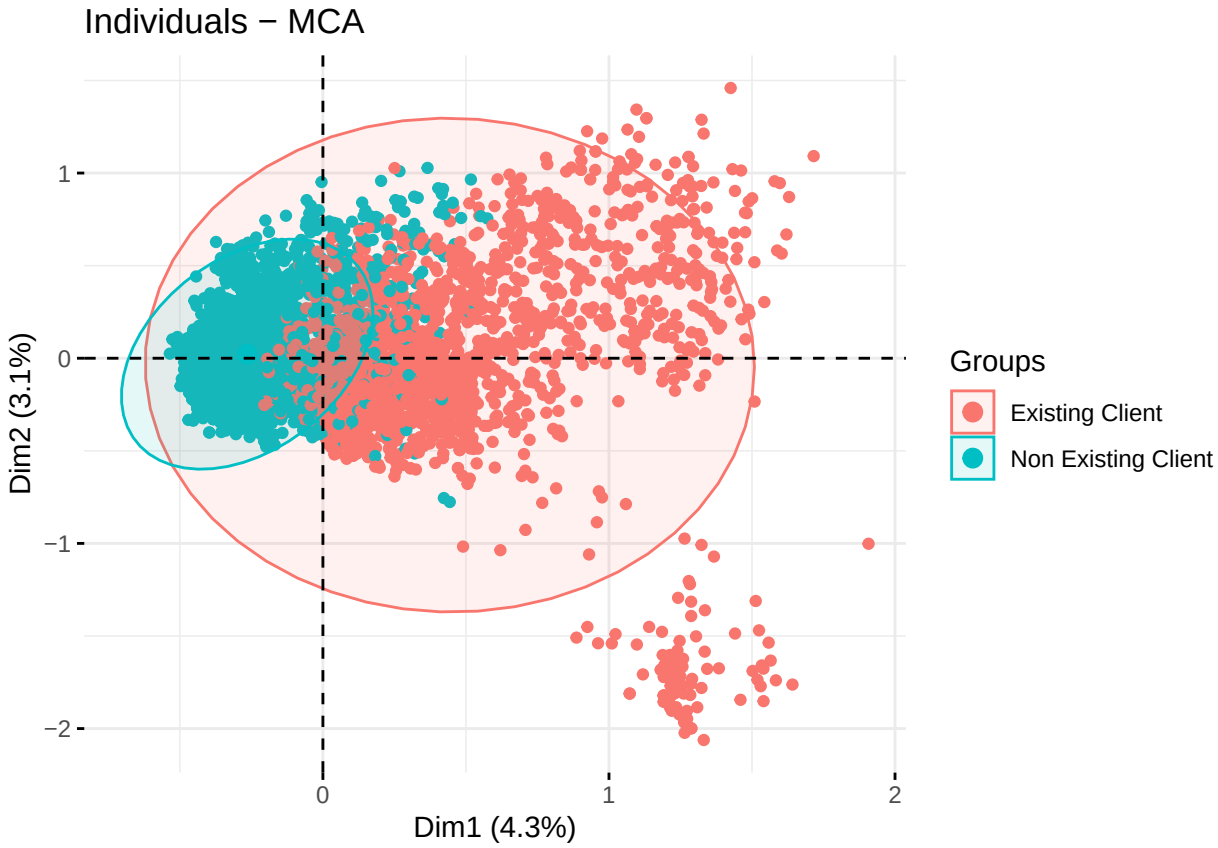
MCA didn't allow us to tell much things about Y individual repartition.

It is clear here that variables Customer_Type, P_Client and Source are explained by the first dimension. Non existing clients and "Sales" individuals seem to be overlapping, suggesting that sales tend to recrute non existing clients, which is understandable.

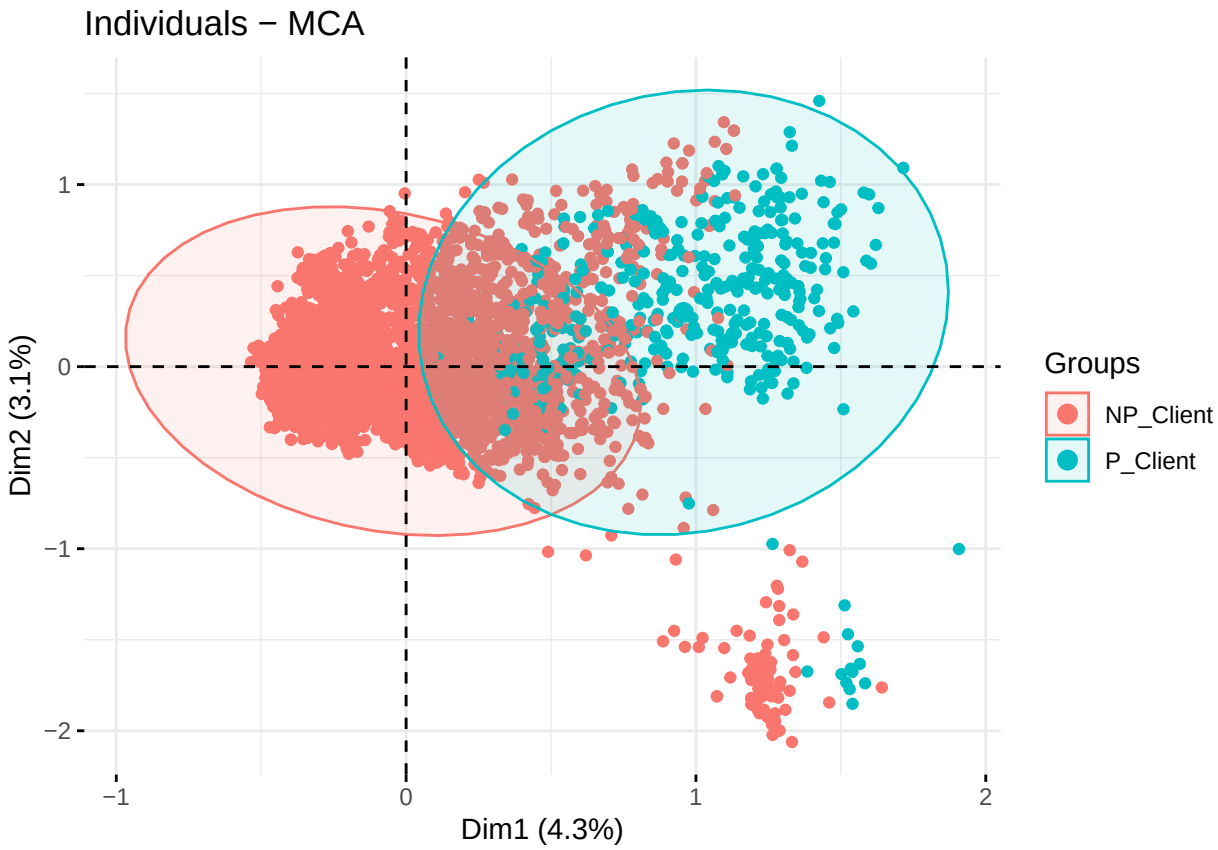
```
grp <- quali_data[, "Y"]
fviz_mca_ind(res.mca, label="none", repel = TRUE, habillage=grp,
      addEllipses=TRUE, ellipse.level=0.95)
```



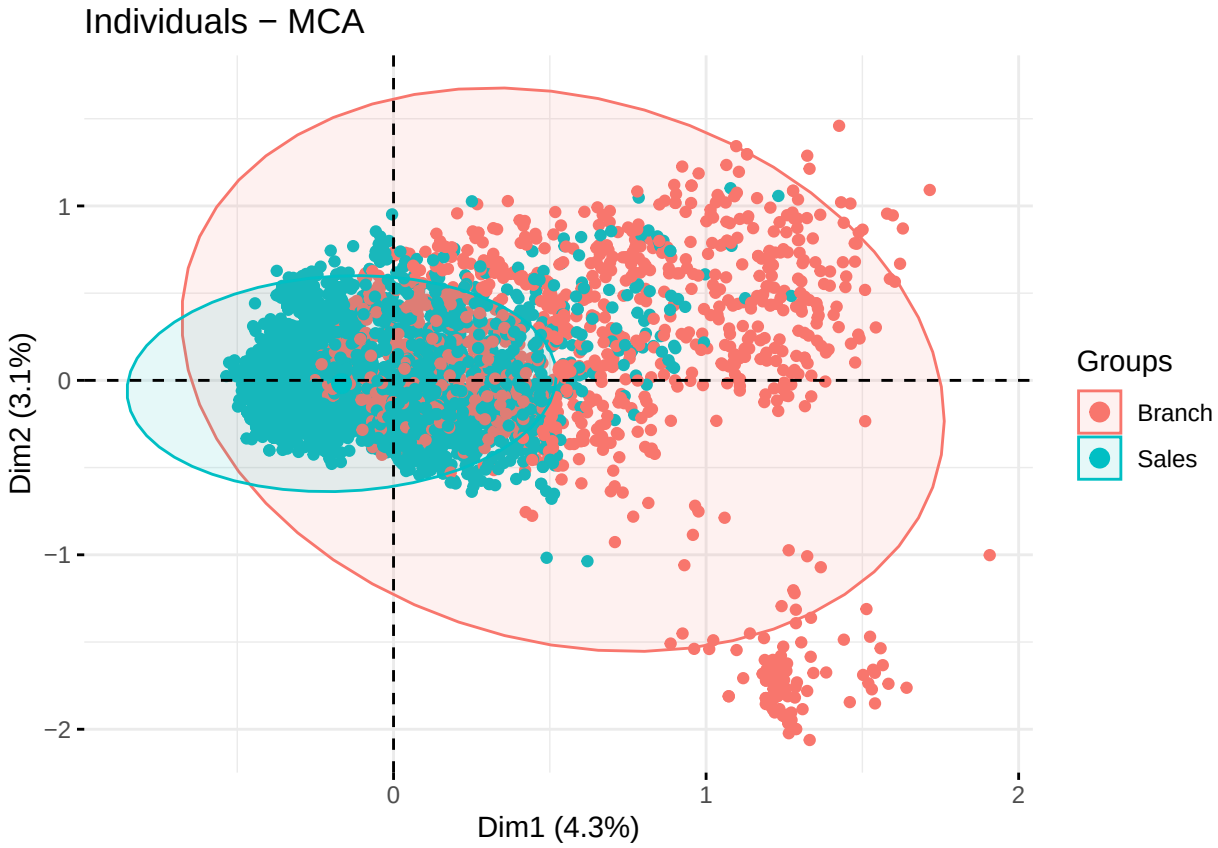

```
grp <- quali_data[, "Customer_Type"]
fviz_mca_ind(res.mca, label="none", repel = TRUE, habillage=grp,
      addEllipses=TRUE, ellipse.level=0.95)
```




```
grp <- quali_data[, "P_Client"]
fviz_mca_ind(res.mca, label="none", repel = TRUE, habillage=grp,
      addEllipses=TRUE, ellipse.level=0.95)
```



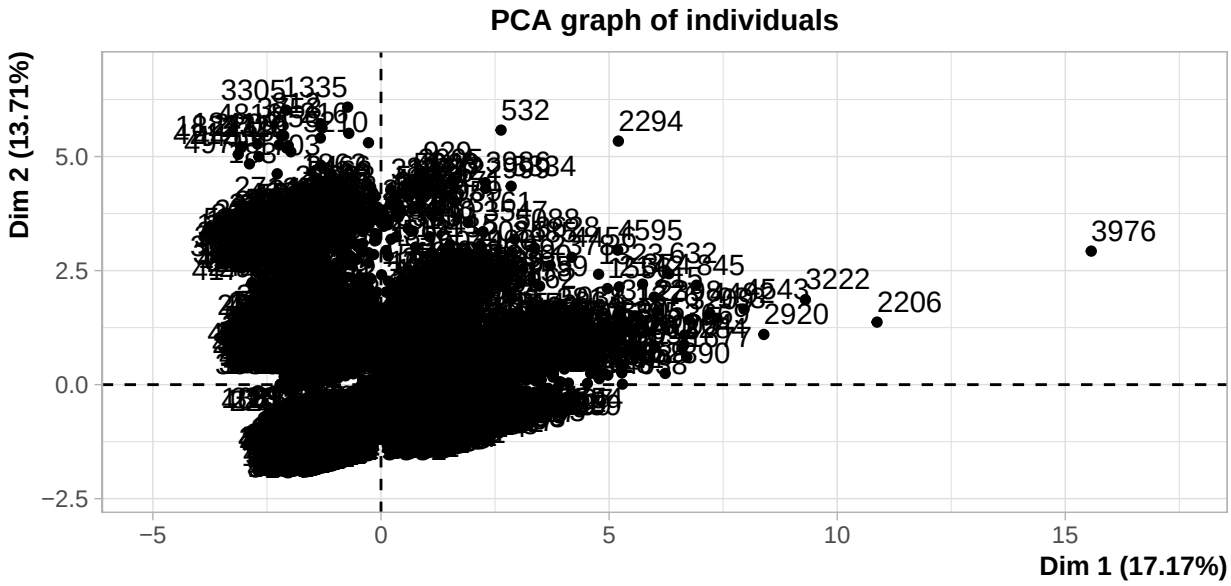

```
grp <- quali_data[, "Source"]
fviz_mca_ind(res.mca, label="none", repel = TRUE, habillage=grp,
      addEllipses=TRUE, ellipse.level=0.95)
```

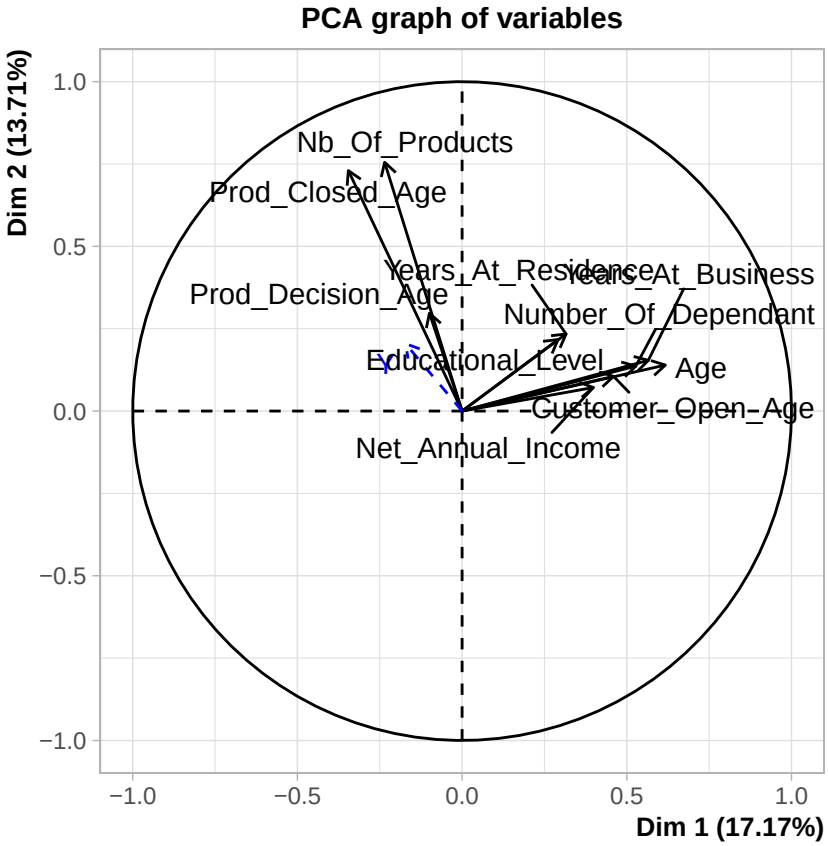


We now try to analyze the dataset with PCA. We kept all those variables : 'Y', 'Age', 'Educational_Level', 'Customer_Open_Age', 'Net_Annual_Income', 'Years_At_Business', 'Number_Of_Dependant', 'Years_At_Residence', 'Prod_Decision_Age', 'Nb_Of_Products', 'Prod_Closed_Age' Educational_Level has been transformed to quantitative variable in python as follows : Secondary or Less: 0 Diploma: 1 University: 2 Master/PhD: 3 It gives an order, Master/PhD being the most important. We thought it was good for educational level to have an order as qualitative variable.

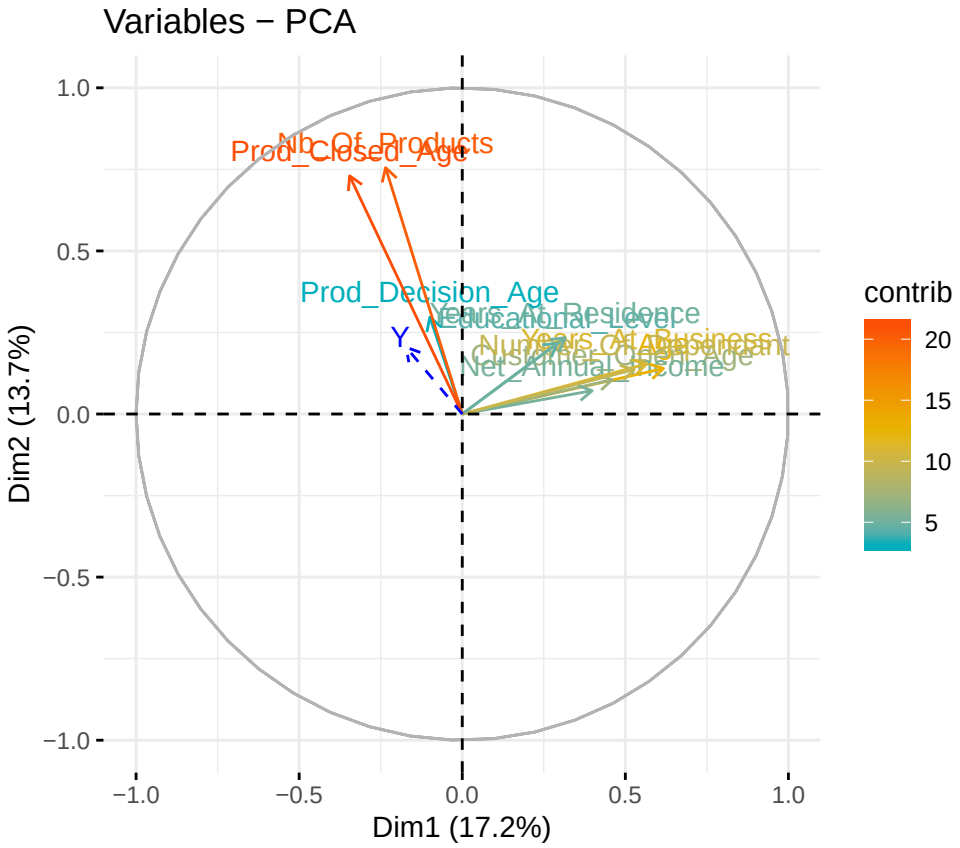
```
quanti_data <- read.table("quanti_databis.csv",
                  sep = ",",
                  header = TRUE,
                  row.names = 1)
#quanti_data, Y is used as supplementary variable
res.pca <- PCA(quanti_data, quanti.sup = 1)
```

```
## Warning in PCA(quanti_data, quanti.sup = 1): Missing values are imputed
## by the mean of the variable: you should use the imputePCA function of the
## missMDA package
```



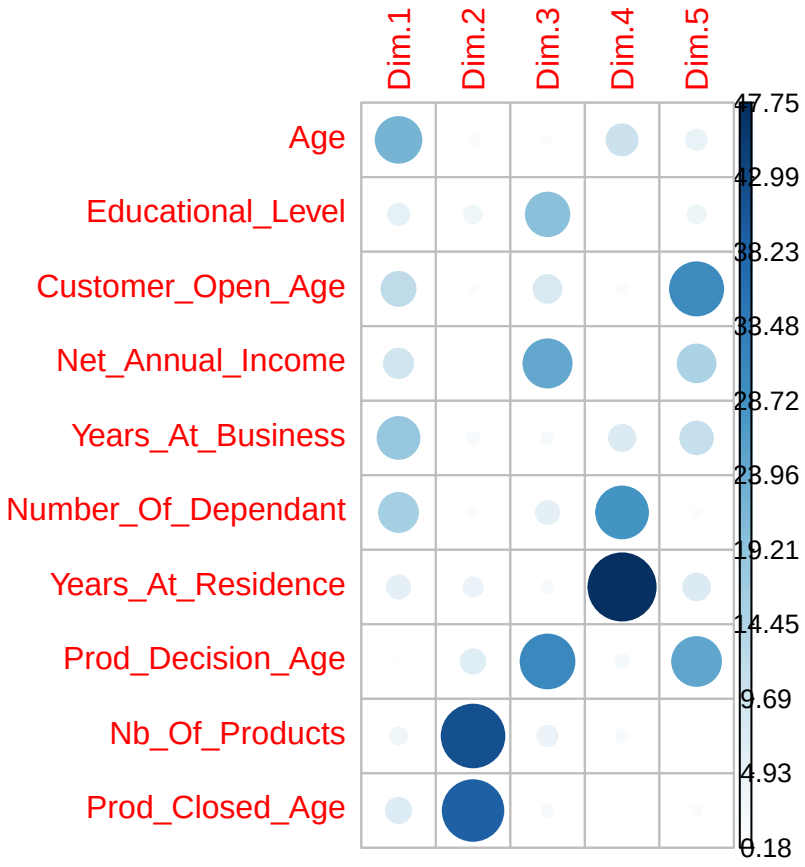



```
fviz_pca_var(res.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
             )
```



We can see here that Prod_Closed and Number_of_Products are highly correlated and contribute a lot to the second dimension.

```
var <- get_pca_var(res.pca)
corrplot(var$contrib, is.corr=FALSE)
```


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
Age
Educational_Level
Customer_Open_Age
Net_Annual_Income
Years_At_Business
Number_Of_Dependant
Years_At_Residence
Prod_Decision_Age
Nb_Of_Products
Prod_Closed_Age
47.75
42.99
38.23
33.48
28.72
23.96
19.21
14.45
9.69
4.93
0.18