Projet - Analyse de Données

Projet KikiCkisenVa - Analyse

```
fact.data <- function(data) {
  if (!is.null(data$Attrition))
    data$Attrition <- as.factor(data$Attrition)
  data$BusinessTravel <- as.factor(data$BusinessTravel)
  data$Department <- as.factor(data$Department)
  data$Education <- as.factor(data$Education)
  data$EducationField <- as.factor(data$EducationField)
  data$EnvironmentSatisfaction <- as.factor(data$EnvironmentSatisfaction)
  data$Gender <- as.factor(data$Gender)
  data$JobInvolvement <- as.factor(data$JobInvolvement)
  data$JobLevel <- as.factor(data$JobLevel)
  data$JobRole <- as.factor(data$JobRole)
  data$JobSatisfaction <- as.factor(data$JobSatisfaction)
  data$MaritalStatus <- as.factor(data$MaritalStatus)
  data$OverTime <- as.factor(data$OverTime)
  data$PerformanceRating <- as.factor(data$PerformanceRating)
  data$RelationshipSatisfaction <- as.factor(data$RelationshipSatisfaction)
  data$StockOptionLevel <- as.factor(data$StockOptionLevel)
  data$WorkLifeBalance <- as.factor(data$WorkLifeBalance)
  return(data)
}
```

Recuperation des donnees

```
data_train <- read.csv2("spreadsheets/data_train.csv", sep = ",")
data_train <- na.omit(data_train)
data_train <- fact.data(data_train)
dim(data_train)
```

```
## [1] 784  32
```

```
head(data_train)
```

```
##   Age Attrition BusinessTravel DailyRate              Department
## 1  50        No  Travel_Rarely      1126 Research & Development
## 2  36        No  Travel_Rarely       216 Research & Development
## 3  21       Yes  Travel_Rarely       337                   Sales
## 4  52        No  Travel_Rarely       994 Research & Development
```

```
## 5  33       Yes  Travel_Rarely      1277 Research & Development
## 6  47        No  Travel_Rarely      1001 Research & Development
##   DistanceFromHome Education EducationField EmployeeNumber
## 1                1         2       Medical            997
## 2                6         2       Medical            178
## 3                7         1     Marketing           1780
## 4                7         4 Life Sciences           1118
## 5               15         1       Medical            582
## 6                4         3 Life Sciences           1827
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       4   Male         66              3        4
## 2                       2   Male         84              3        2
## 3                       2   Male         31              3        1
## 4                       2   Male         87              3        3
## 5                       2   Male         56              3        3
## 6                       3 Female         92              2        3
##                    JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1         Research Director               4      Divorced         17399
## 2    Manufacturing Director               2      Divorced          4941
## 3      Sales Representative               2        Single          2679
## 4 Healthcare Representative               2        Single         10445
## 5                   Manager               3       Married         13610
## 6    Manufacturing Director               2      Divorced         10333
##   MonthlyRate NumCompaniesWorked OverTime PercentSalaryHike PerformanceRating
## 1        6615                 9       No                22                 4
## 2        2819                 6       No                20                 4
## 3        4567                 1       No                13                 3
## 4       15322                 7       No                19                 3
## 5       24619                 7      Yes                12                 3
## 6       19271                 8      Yes                12                 3
##   RelationshipSatisfaction StockOptionLevel TotalWorkingYears
## 1                        3                1                32
## 2                        4                2                 7
## 3                        2                0                 1
## 4                        4                0                18
## 5                        4                0                15
## 6                        3                1                28
##   TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole
## 1                     1               2              5                  4
## 2                     0               3              3                  2
## 3                     3               3              1                  0
## 4                     4               3              8                  6
## 5                     2               4              7                  6
## 6                     4               3             22                 11
##   YearsSinceLastPromotion YearsWithCurrManager
## 1                       1                    3
## 2                       0                    1
## 3                       1                    0
## 4                       4                    0
## 5                       7                    7
## 6                      14                   10
```

```
data_train_num <- data_train[, unlist(lapply(data_train, is.numeric))]
data_train_num[16] <- data_train["Attrition"]
```

```
dim(data_train_num)
```

```
## [1] 784  16
```

```
head(data_train_num)
```

```
##   Age DailyRate DistanceFromHome EmployeeNumber HourlyRate MonthlyIncome
## 1  50      1126                1             997         66         17399
## 2  36       216                6             178         84          4941
## 3  21       337                7            1780         31          2679
## 4  52       994                7            1118         87         10445
## 5  33      1277               15             582         56         13610
## 6  47      1001                4            1827         92         10333
##   MonthlyRate NumCompaniesWorked PercentSalaryHike TotalWorkingYears
## 1        6615                  9                22                32
## 2        2819                  6                20                 7
## 3        4567                  1                13                 1
## 4       15322                  7                19                18
## 5       24619                  7                12                15
## 6       19271                  8                12                28
##   TrainingTimesLastYear YearsAtCompany YearsInCurrentRole
## 1                     1              5                  4
## 2                     0              3                  2
## 3                     3              1                  0
## 4                     4              8                  6
## 5                     2              7                  6
## 6                     4             22                 11
##   YearsSinceLastPromotion YearsWithCurrManager Attrition
## 1                       1                    3        No
## 2                       0                    1        No
## 3                       1                    0       Yes
## 4                       4                    0        No
## 5                       7                    7       Yes
## 6                      14                   10        No
```

Stats descriptives

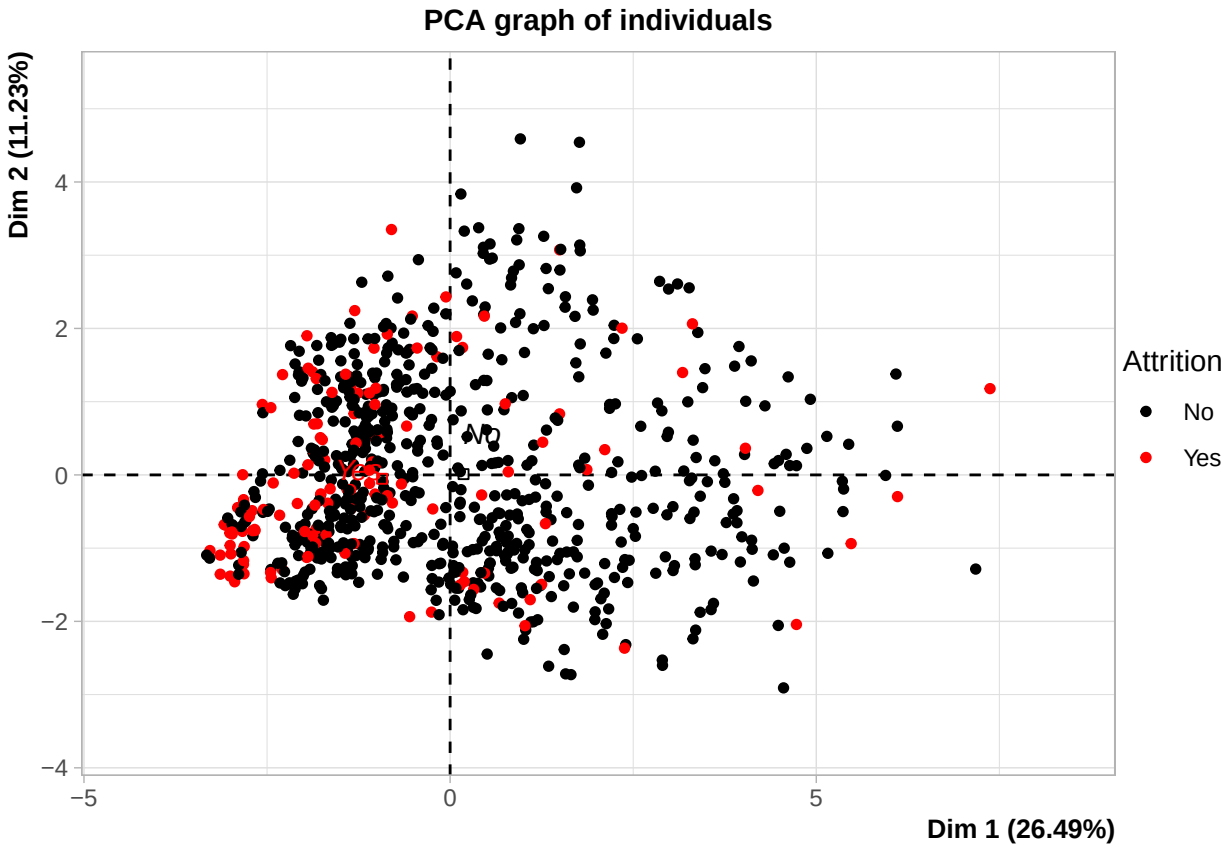
```
chisq.test(data_train_num[-16])
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_train_num[-16]
## X-squared = 3415673, df = 10962, p-value < 2.2e-16
```

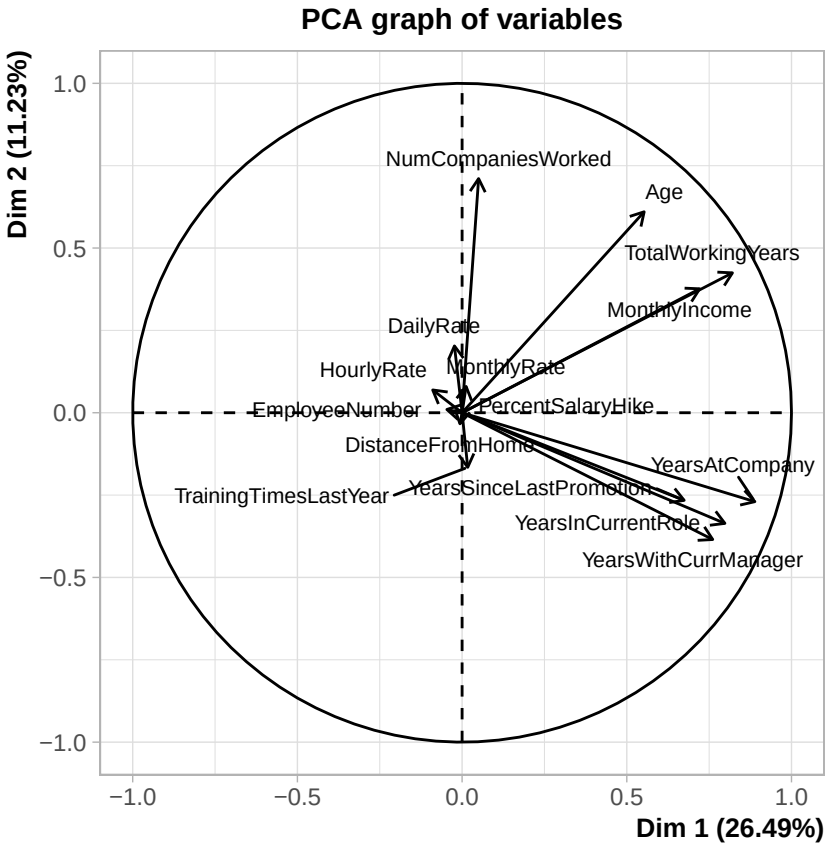
Toutes les variables ne semblent pas indépentantes entre elles.

ACP

```
library(FactoMineR)
res.pca <- PCA(data_train_num, scale.unit = TRUE, graph = FALSE, quali.sup = 16)
plot(res.pca, choix = "ind", habillage = 16, select = FALSE, unselect = 0)
```




```
plot(res.pca, choix = "var", cex = 0.7)
```



On voit apparaitre un effet taille. Pour contrer cela nous allons transformer les données en appliquant

Équilibrage des Données

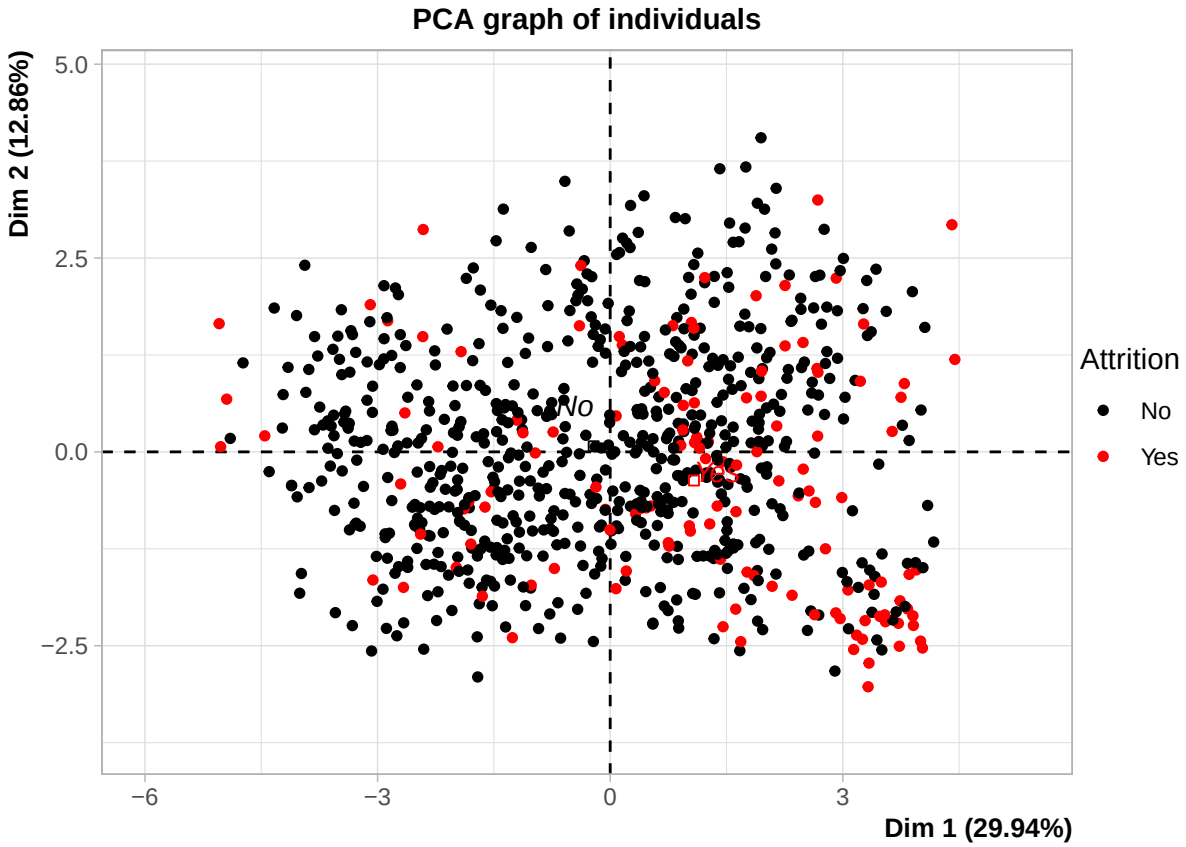
```
data_train_log <- log(data_train_num[,-16])
data_train_log[data_train_log == -Inf] <- 0
data_train_log <- t(scale(t(data_train_log)))
data_train_log <- as.data.frame(data_train_log)
data_train_log[16] <- data_train["Attrition"]

head(data_train_log)
```

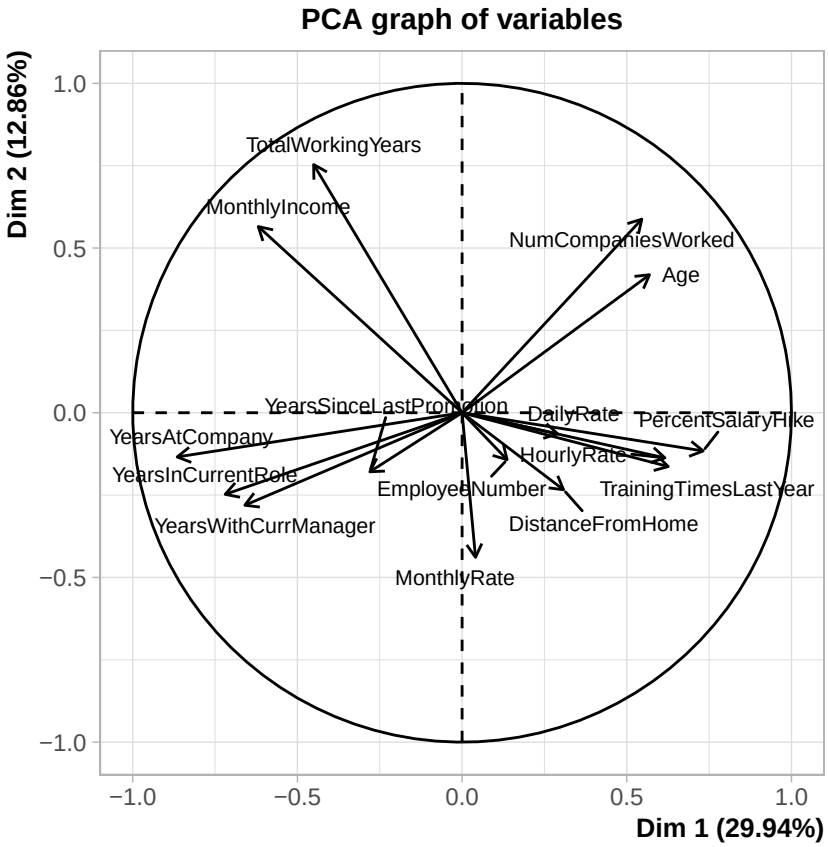
```
##           Age DailyRate DistanceFromHome EmployeeNumber HourlyRate
## 1  0.10875235 1.0787024       -1.1096083      1.0408076 0.19521799
## 2  0.20244958 0.8490462       -0.4441471      0.7792193 0.50821605
## 3  0.08343912 0.9616526       -0.2641722      1.4882483 0.20666944
## 4  0.03712663 1.0215512       -0.6319477      1.0607746 0.20884308
## 5 -0.14446568 1.0948206       -0.4117491      0.8284386 0.03481005
## 6 -0.12447781 0.9637158       -1.0010694      1.1777804 0.11447912
##   MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike
## 1      1.931345    1.630159         -0.4253045      -0.146933864
## 2      1.978593    1.776077         -0.4441471      -0.009666449
## 3      1.617606    1.786383         -0.8798765      -0.068302317
## 4      1.806337    1.934180         -0.6319477      -0.298791173
## 5      1.896983    2.097910         -0.6701111      -0.487393757
## 6      1.794230    2.015974         -0.7544610      -0.610204272
```

```
##   TotalWorkingYears TrainingTimesLastYear YearsAtCompany YearsInCurrentRole
## 1       -0.03023932            -1.1096083     -0.6083648         -0.6778607
## 2       -0.38851835            -1.0907437     -0.6942848         -0.8406060
## 3       -0.87987646            -0.5322652     -0.8798765         -0.8798765
## 4       -0.31683056            -0.8186621     -0.5873953         -0.6833797
## 5       -0.41174912            -1.0947920     -0.6701111         -0.7223675
## 6       -0.30875200            -1.0010694     -0.3945528         -0.6411612
##   YearsSinceLastPromotion YearsWithCurrManager Attrition
## 1              -1.1096083           -0.7674564        No
## 2              -1.0907437           -1.0907437        No
## 3              -0.8798765           -0.8798765       Yes
## 4              -0.8186621           -1.2811956        No
## 5              -0.6701111           -0.6701111       Yes
## 6              -0.5553604           -0.6750708        No
```

```
res.pca.log <- PCA(data_train_log, scale.unit = TRUE, graph = FALSE, quali.sup = 16)
plot(res.pca.log, choix = "ind", habillage = 16, select = FALSE, unselect = 0)
```



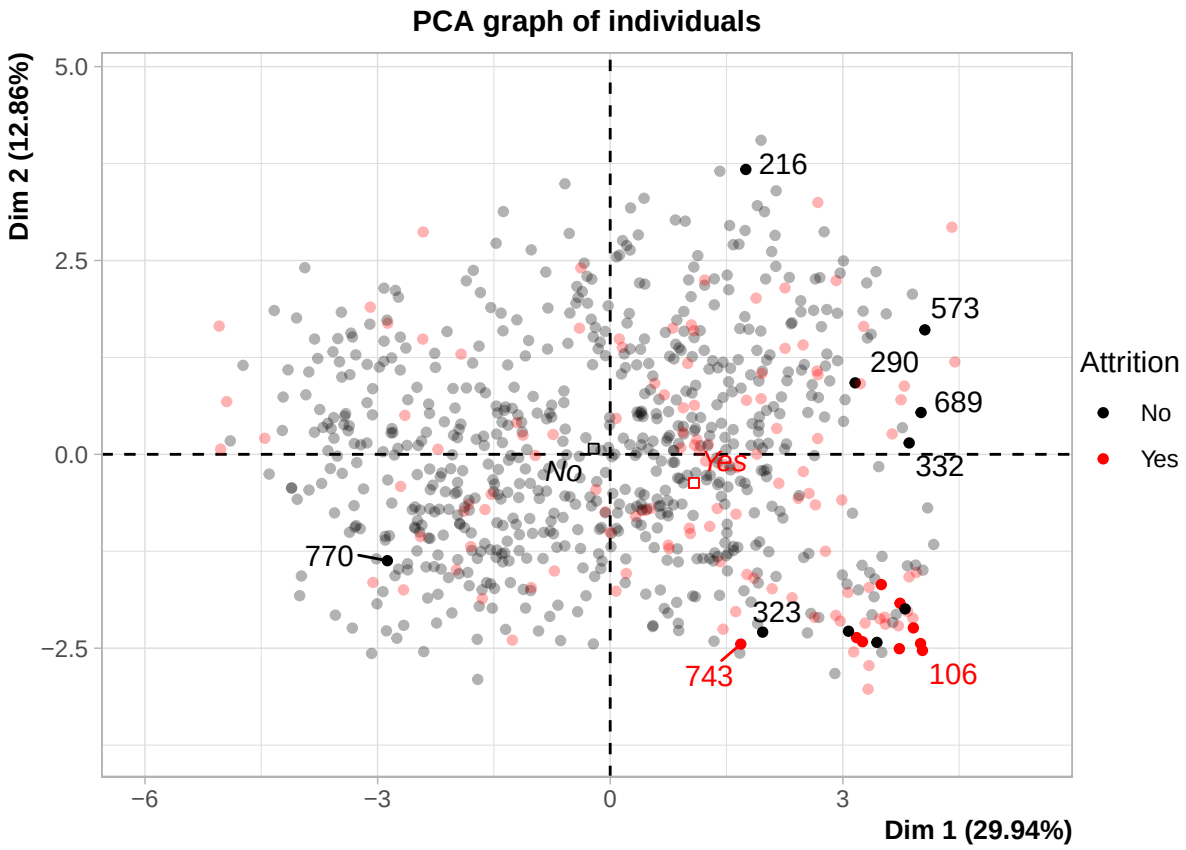

```
plot(res.pca.log, choix = "var", cex = 0.7)
```



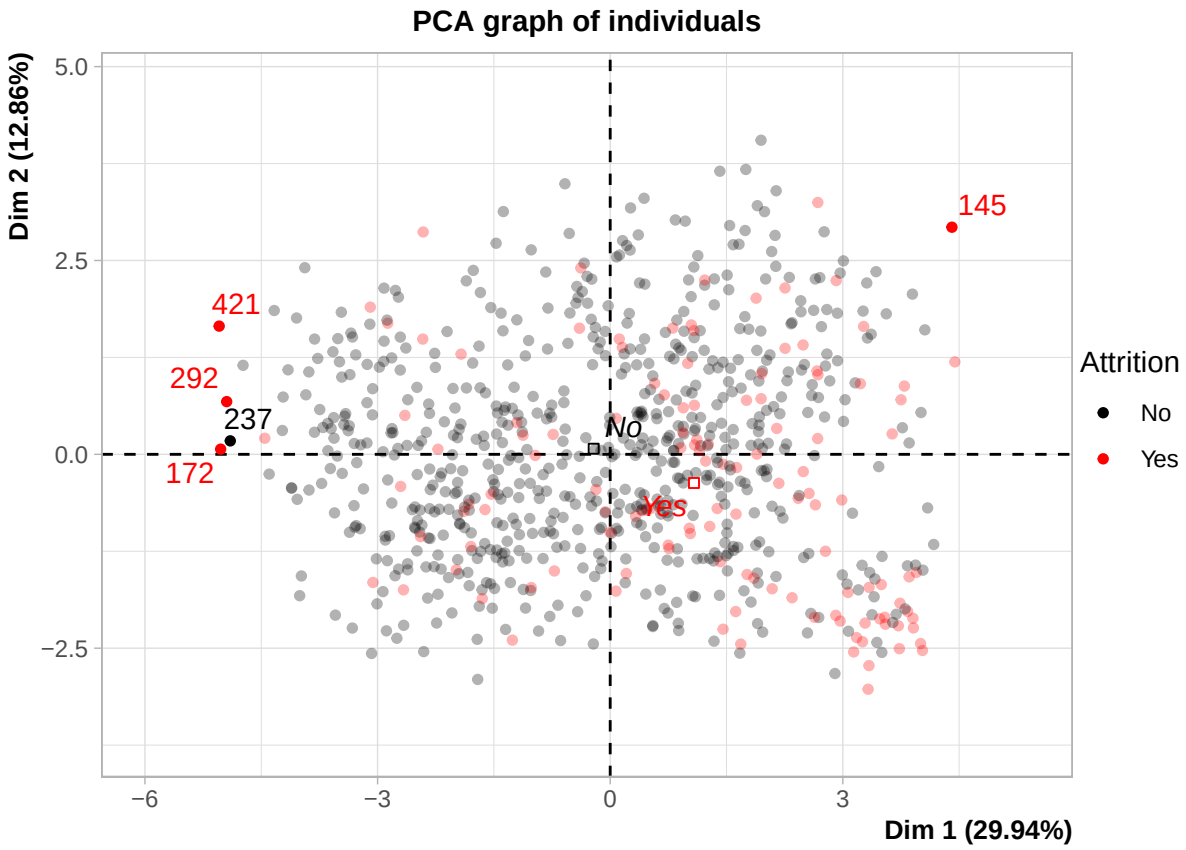
Contribution et représentation des données

```
plot(res.pca.log, select="cos2 0.82", choix="ind", habillage = 16)
```

```
## Warning: ggrepel: 10 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



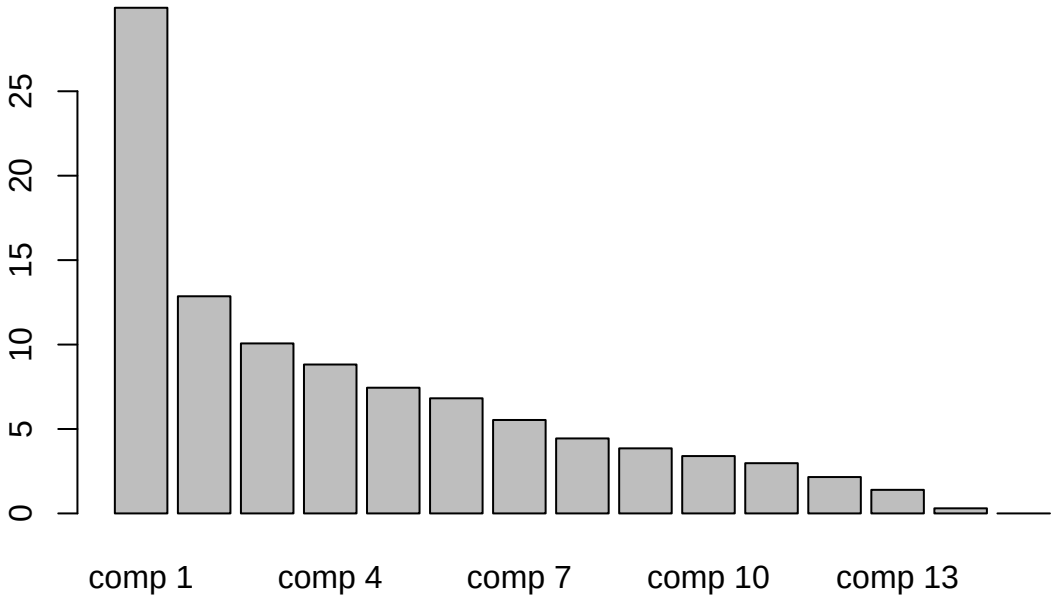

```
plot(res.pca.log, select="contrib 5", choix="ind", habillage = 16)
```




```
summary(res.pca.log$eig)
```

```
##    eigenvalue     percentage of variance cumulative percentage of variance
##  Min.   :0.0000   Min.   : 0.000         Min.   : 29.94
##  1st Qu.:0.3847   1st Qu.: 2.565         1st Qu.: 65.41
##  Median :0.6661   Median : 4.441         Median : 85.92
##  Mean   :1.0000   Mean   : 6.667         Mean   : 78.46
##  3rd Qu.:1.2197   3rd Qu.: 8.131         3rd Qu.: 97.23
##  Max.   :4.4915   Max.   :29.943         Max.   :100.00
```

```
barplot(res.pca.log$eig[,2])
```



L'inertie de chaque composante en pourcentage. On remarque que les 2 premiers axes suffisent car les autres apportent moins de 10%...

```
usefull_col <- (res.pca.log$var$contrib[,1] > median(res.pca.log$var$contrib[,1])) | (res.pca.log$var$co
usefull_col
```

```
##                    Age               DailyRate        DistanceFromHome
##                   TRUE                   FALSE                   FALSE
##         EmployeeNumber              HourlyRate          MonthlyIncome
##                  FALSE                    TRUE                    TRUE
##            MonthlyRate      NumCompaniesWorked       PercentSalaryHike
##                   TRUE                    TRUE                    TRUE
##      TotalWorkingYears   TrainingTimesLastYear          YearsAtCompany
##                   TRUE                    TRUE                    TRUE
##     YearsInCurrentRole YearsSinceLastPromotion    YearsWithCurrManager
##                   TRUE                   FALSE                    TRUE
```

AFC-MCA

```
data_train_fact <- data_train[, unlist(lapply(data_train, is.factor))]
dim(data_train_fact)
```

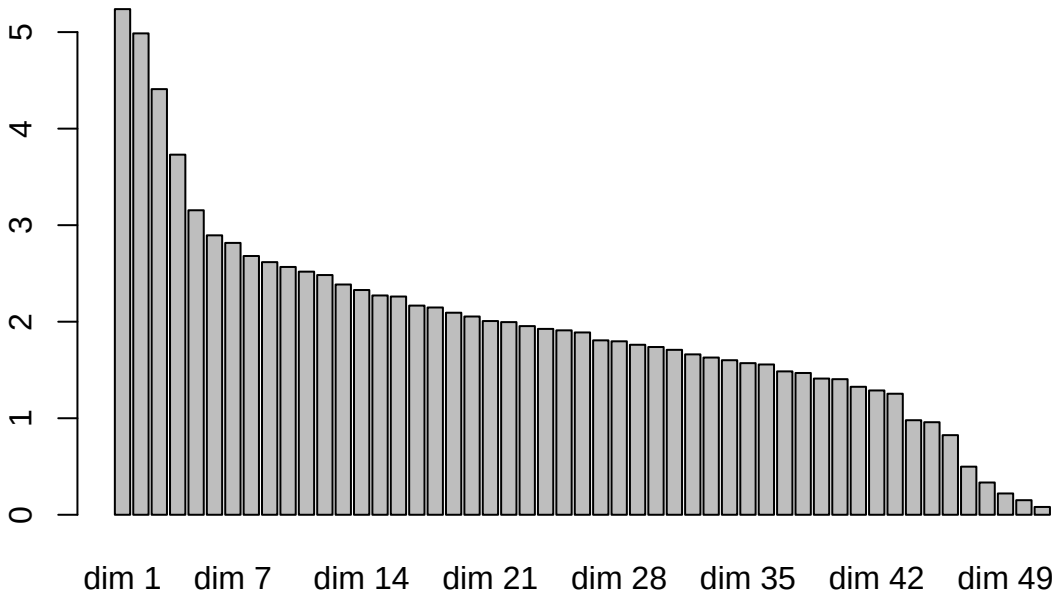
```
## [1] 784  17
```

```
head(data_train_fact)
```

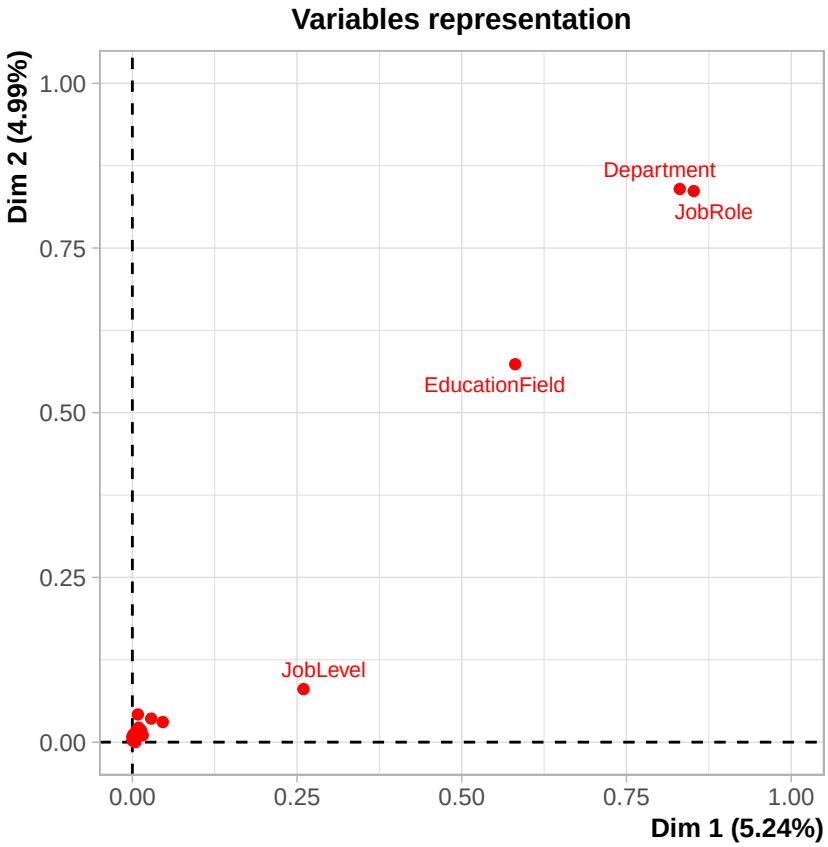
```
##   Attrition BusinessTravel              Department Education EducationField
## 1        No  Travel_Rarely Research & Development         2        Medical
## 2        No  Travel_Rarely Research & Development         2        Medical
## 3       Yes  Travel_Rarely                   Sales         1      Marketing
## 4        No  Travel_Rarely Research & Development         4  Life Sciences
```

```
## 5       Yes  Travel_Rarely Research & Development         1        Medical
## 6        No  Travel_Rarely Research & Development         3  Life Sciences
##   EnvironmentSatisfaction Gender JobInvolvement JobLevel
## 1                       4   Male              3        4
## 2                       2   Male              3        2
## 3                       2   Male              3        1
## 4                       2   Male              3        3
## 5                       2   Male              3        3
## 6                       3 Female              2        3
##                    JobRole JobSatisfaction MaritalStatus OverTime
## 1         Research Director               4      Divorced       No
## 2    Manufacturing Director               2      Divorced       No
## 3      Sales Representative               2        Single       No
## 4 Healthcare Representative               2        Single       No
## 5                   Manager               3       Married      Yes
## 6    Manufacturing Director               2      Divorced      Yes
##   PerformanceRating RelationshipSatisfaction StockOptionLevel WorkLifeBalance
## 1                 4                       3                1               2
## 2                 4                       4                2               3
## 3                 3                       2                0               3
## 4                 3                       4                0               3
## 5                 3                       4                0               4
## 6                 3                       3                1               3
```

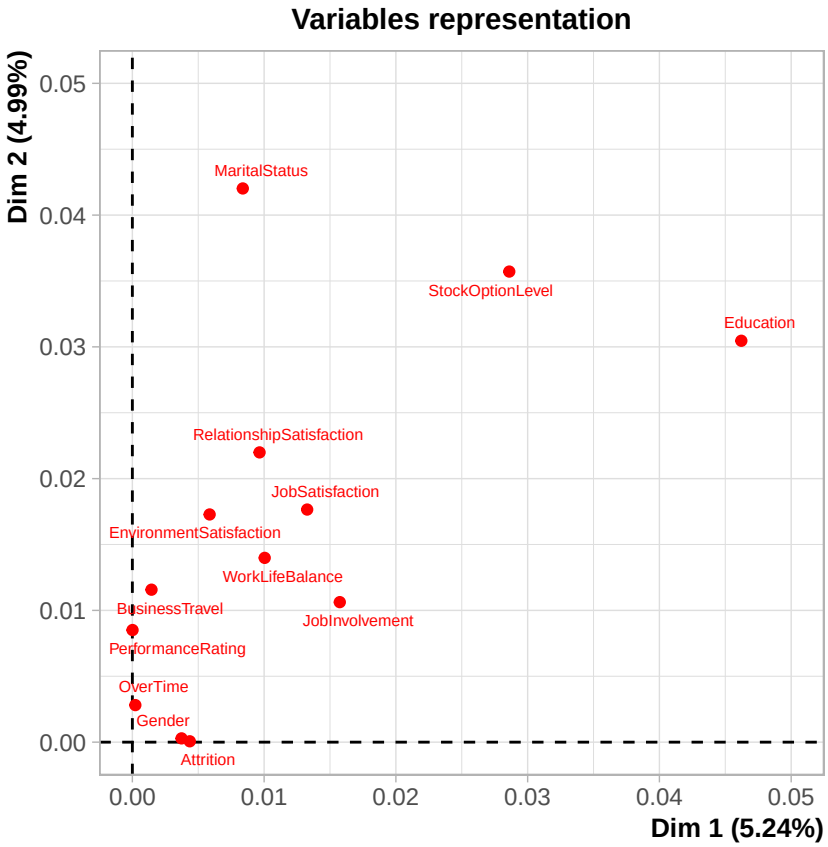
```
library(FactoMineR)
res.mca = MCA(data_train_fact, graph = FALSE)
barplot(res.mca$eig[,2])
```




```
plot(res.mca, choix = "var", cex = 0.7)
```




```
plot(res.mca, choix = "var", xlim = c(0, 0.05), ylim = c(0, 0.05), cex = 0.5)
```




```
attach(data_train)
chisq.test(table(EducationField, JobRole))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(EducationField, JobRole)
## X-squared = 506.77, df = 40, p-value < 2.2e-16
```

```
chisq.test(table(EducationField, Department))
```

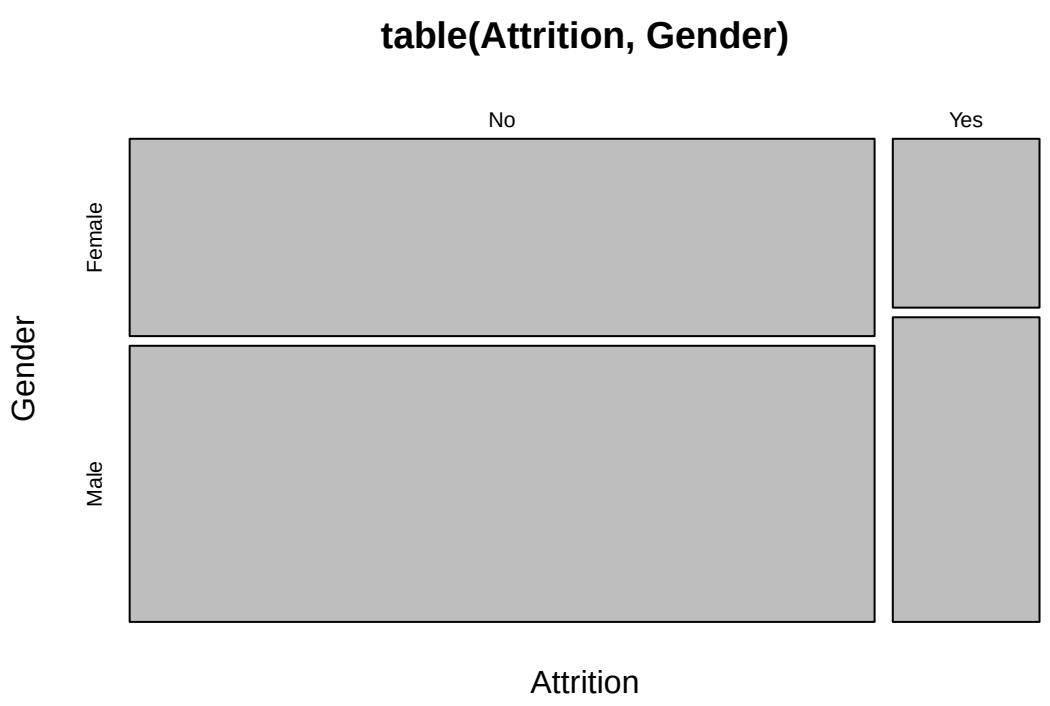
```
##
##  Pearson's Chi-squared test
##
## data:  table(EducationField, Department)
## X-squared = 620.76, df = 10, p-value < 2.2e-16
```

```
chisq.test(table(JobRole, Department))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(JobRole, Department)
## X-squared = 1351.6, df = 16, p-value < 2.2e-16
```

On a une p-value < 0.05, les variables sont donc effectivement liées.

```
attach(data_train)
plot(table(Attrition,Gender))
```




```
chisq.test(table(Attrition,Gender))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Attrition, Gender)
## X-squared = 1.3788, df = 1, p-value = 0.2403
```

Finalement le test chi 2 nous montre l'indépendance, démontrant que notre modèle n'est pas parfait.