R Notebook | LUCAS ISCOVICI BINOME 9

Code

L'OBJECT DE L'ANALYSE: Analyser de données du jeu de données "villes", présentant par villes, différentes informations, tel que les salaires moyen de plusieurs métiers, les prix et quelques indicateurs essentiellement économiques.

Library utilisé: FactoMineR, factoextra, corrplot

Hide

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

Le jeu de données:

Hide

```
villes
```

ville <fctr>	annee <dbl>	region <dbl>	prxsl <dbl>	prxal <dbl>	salbrt <dbl>	salnet <dbl>	htrav <dbl>	vac <dbl>	achbrt <dbl>
AbuDhabi91	1	9	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
AbuDhabi94	2	9	71.3	78.1	29.9	38.0	2100	28.9	42.0
Amsterdam91	1	2	65.6	65.7	56.9	49.0	1714	31.9	86.7
Amsterdam94	2	2	68.8	70.3	61.4	53.7	1792	27.5	89.3
Athenes91	1	3	53.8	55.6	30.2	30.4	1792	23.5	56.1
Athenes94	2	3	54.4	56.7	27.7	28.5	1775	24.5	50.9
Bangkok91	1	6	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Bangkok94	2	6	64.6	70.1	9.5	11.2	2272	8.8	14.6
Bogota91	1	8	37.9	39.3	10.1	11.5	2152	17.4	26.6
Bogota94	2	8	52.8	54.4	13.5	16.0	2154	18.7	25.5

1-10 of 102 rows | 1-10 of 42 columns Previous **1** 2 3 4 5 6 … 11 Next

Dans ce jeu de données, nous allons garder pour l'analyse en composante principale, les 12 variables quantitatives indiquant le salaires moyen de plusieurs metiers (instit, chauffeur, meca, man, tourneur, cuisinier, chefserv, inge, banque, secr, vendeuse et ouvriere).

Ici dans cette etude nous allons nous intéréssée seulement aux données de l'édition 1994.

Hide

```
dim(villes)
```

```
[1] 102  42
```

Il ya donc 102 lignes et 41 colonnes

Nous effectuons ici tout le traitement de preparation des données

Hide

```
selectCol=c(3,8,14,29,(ncol(villes)-12):(ncol(villes)-1)) #les colonnes à selectionne
r.
sup=list(quanti=2:4,quali=c(1)) #indice var sup
lselectCol=length(selectCol) #taille des colonnes
ville94_12v=villes[c(F,T),selectCol] # Les données
colNormal=5:16 #indices des colonnes à etudier
row.names(ville94_12v)<-villes[c(F,T),]$ville #on indique au data.frame le nom des li
gnes
```

Nous affichons les données

	instit <dbl>	**chauffeur** <dbl>	**me…** <dbl>	**man** <dbl>	**tourneur** <dbl>	**cuisinier** <dbl>	**chefserv** <dbl>	**inge** <dbl>	**ban…** <dbl>
AbuDhabi94	19500	11400	9200	3500	6800	33900	95000	59700	47800
Amsterdam94	23800	24900	14300	13000	22000	15600	33600	32600	22500
Athenes94	10100	11300	6000	9700	9600	11000	12300	13000	11000
Bangkok94	4100	3400	2600	1700	6600	8500	27300	17900	12800
Bogota94	4100	4100	6500	1700	5500	11600	31500	19000	8400
Bombay94	1600	1700	1300	800	1400	2700	4300	2100	1800
Bruxelles94	16000	14900	12200	13200	18100	19000	30300	24600	20600
Budapest94	2100	3000	2200	1900	2600	4300	6400	4200	4200
BuenosAires94	4500	4500	8200	4500	6600	10900	30200	24000	33500
Caracas94	2500	900	1300	900	1600	6100	4400	8100	2000

1-10 of 51 rows | 1-10 of 12 columns Previous **1** 2 3 4 5 6 Next

Nous remarquons qu'il n'y a pas de valeurs manquantes. Il y a 51 lignes (villes)

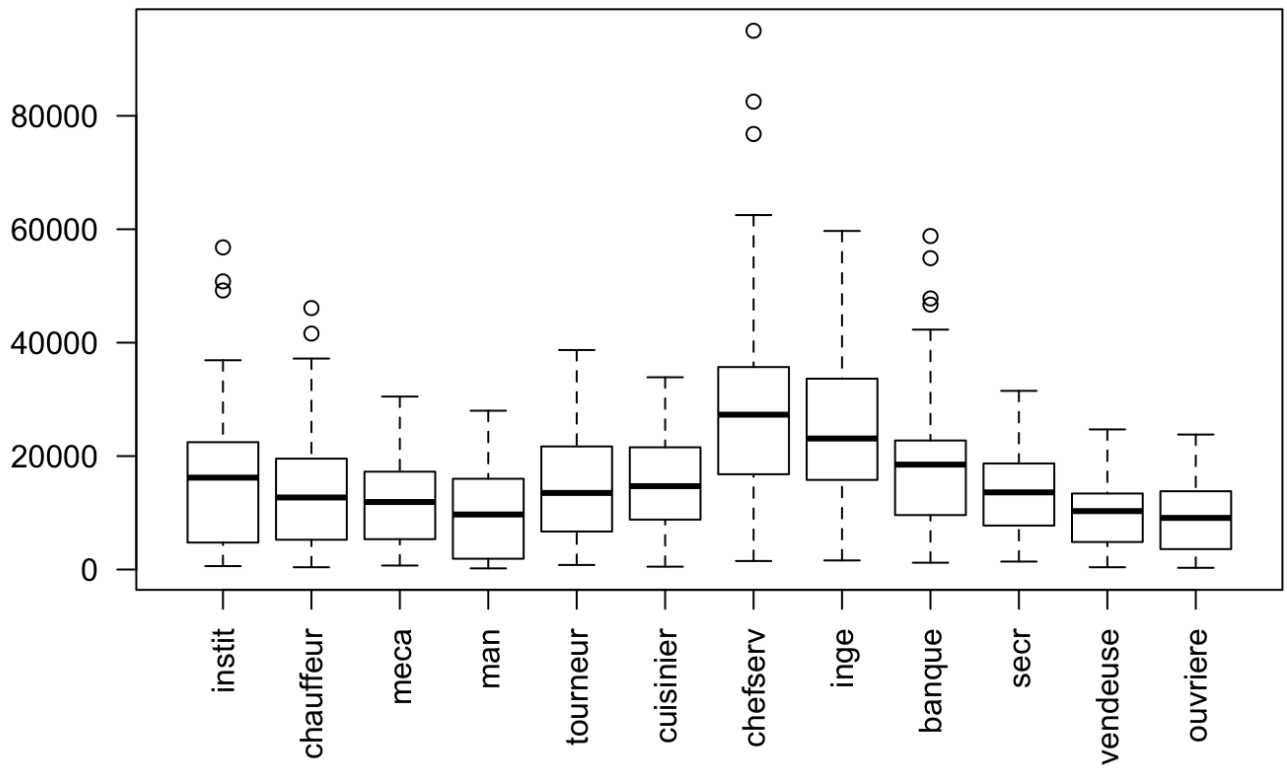
Nous allons maintenant faire une analyse univarié des variables.

1/Univarié et valeurs "anormales"

Hide

```
summary(ville94_12v_normal)
```

```
     instit        chauffeur          meca            man          tourneur    
 Min.   :  600   Min.   :  400   Min.   :  700   Min.   :  200   Min.   :  800  
 1st Qu.: 4750   1st Qu.: 5250   1st Qu.: 5350   1st Qu.: 1900   1st Qu.: 6700  
 Median :16200   Median :12700   Median :11900   Median : 9700   Median :13500  
 Mean   :16802   Mean   :14312   Mean   :12384   Mean   :10343   Mean   :15145  
 3rd Qu.:22450   3rd Qu.:19550   3rd Qu.:17250   3rd Qu.:16000   3rd Qu.:21700  
 Max.   :56800   Max.   :46100   Max.   :30500   Max.   :28000   Max.   :38700  
   cuisinier        chefserv          inge           banque          secr       
 Min.   :  500   Min.   : 1500   Min.   : 1600   Min.   : 1200   Min.   : 1400  
 1st Qu.: 8800   1st Qu.:16800   1st Qu.:15800   1st Qu.: 9600   1st Qu.: 7750  
 Median :14700   Median :27300   Median :23100   Median :18500   Median :13600  
 Mean   :15616   Mean   :30933   Mean   :24665   Mean   :18749   Mean   :13312  
 3rd Qu.:21550   3rd Qu.:35700   3rd Qu.:33650   3rd Qu.:22750   3rd Qu.:18700  
 Max.   :33900   Max.   :95000   Max.   :59700   Max.   :58800   Max.   :31500  
    vendeuse        ouvriere     
 Min.   :  400   Min.   :  300  
 1st Qu.: 4850   1st Qu.: 3600  
 Median :10300   Median : 9100  
 Mean   : 9659   Mean   : 9247  
 3rd Qu.:13400   3rd Qu.:13800  
 Max.   :24700   Max.   :23800  
```



On remarque qu'il y a quelques valeurs "anormales" pour les variables instit , chauffeur, chefserv et banque.

Hide

```
t(colValAb)
```

```
     [,1]     [,2]         [,3]        [,4]    
Var1 "instit" "chauffeur" "chefserv" "banque"
Freq "3"      "2"         "3"        "4"      
```

Pour les variables, je considere qu'à partir d'un quart du nombre des lignes, j'enleve la variables. (51/4=12.75), donc pas de pb.

Nous allons regarder plus en detail quelles villes sont concernées par des valeurs anormales.

Hide

```
tb
```

```
villesAberantes
  AbuDhabi94      Geneve94 Luxembourg94       Tokyo94      Zurich94
           2             2             2             2             4
```

Nous remarquons qu'il y a seulement 2 valeurs anormales pour AbuDhabi94, Geneve94, Luxembourg94 et Tokyo94, il y a 12 variables je considere qu'à partir d'un quart du nombre des colonnes j'enleve la ville (3).

Donc je vais enlever la villes Zurich car il y a 4 valeurs anormales. De plus l'acp est sensibles aux outliers.

Hide

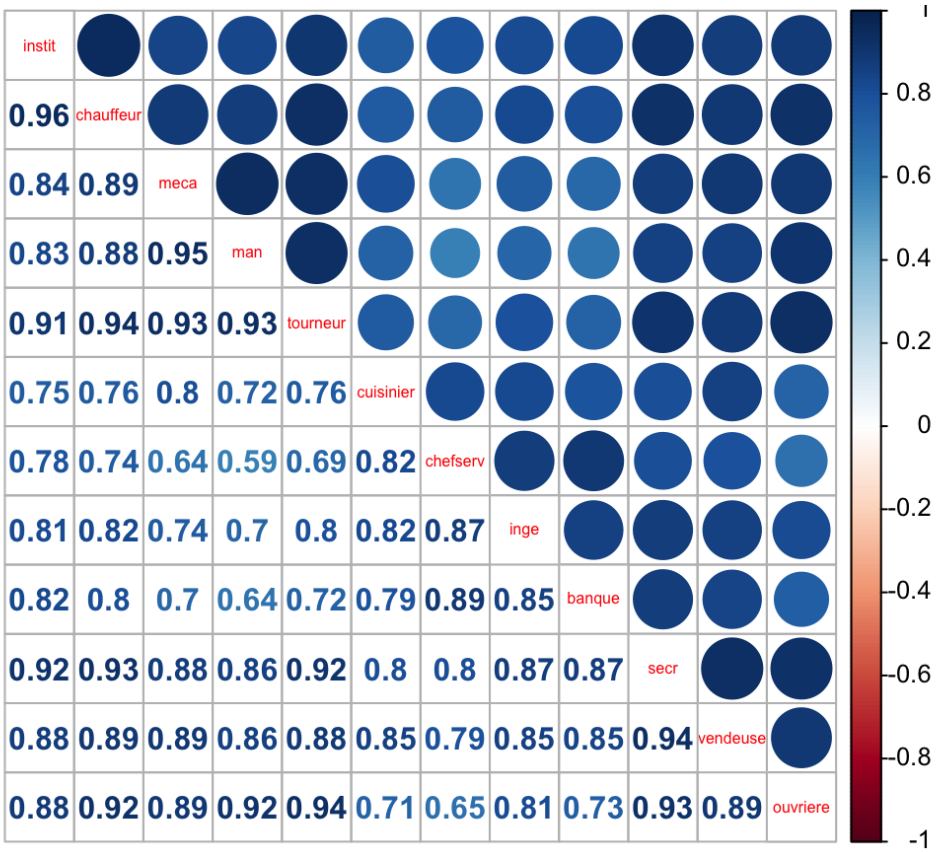
```
nrow(ville94_12v)
```

```
[1] 50
```

AINSI IL Y AURA 50 VILLES MAINTENANT

2/Multivarié - Correlations

Maintenant je vais regarder en premiere impression les correlations entre les variables.



On remarque que toutes les variables sont corrélées positivement et pour la plupart fortement.

correlation de plus de 0.9 et variables qui ont le plus de correlations avec les autres variables

		c
	<chr>	<int>
7	vendeuse	1
3	meca	2
1	instit	3
4	man	3
2	chauffeur	4
8	ouvriere	4
6	secr	5
5	tourneur	6

8 rows

On remarque donc que tourneur est corrélé avec la moitié des variables.

Hide

```
var_correlé
```

```
$instit
chauffeur  tourneur      secr
        1         2         3

$chauffeur
  instit tourneur     secr ouvriere
       1        2        3        4

$meca
     man tourneur
       1        2

$man
    meca tourneur ouvriere
       1        2        3

$tourneur
   instit chauffeur      meca       man      secr  ouvriere
        1         2         3         4         5         6

$cuisinier
named integer(0)

$chefserv
named integer(0)

$inge
named integer(0)

$banque
named integer(0)

$secr
   instit chauffeur  tourneur  vendeuse  ouvriere
        1         2         3         4         5

$vendeuse
secr
   1

$ouvriere
chauffeur       man  tourneur      secr
        1         2         3         4
```

3/ L’ACP

Nous allons commencer L’ACP:

Pour les variables supplémentaire, j’ai choisi la variable “region”" (qualitative) et les variables “alim”, “salhor” et “htrav”(quantitives).

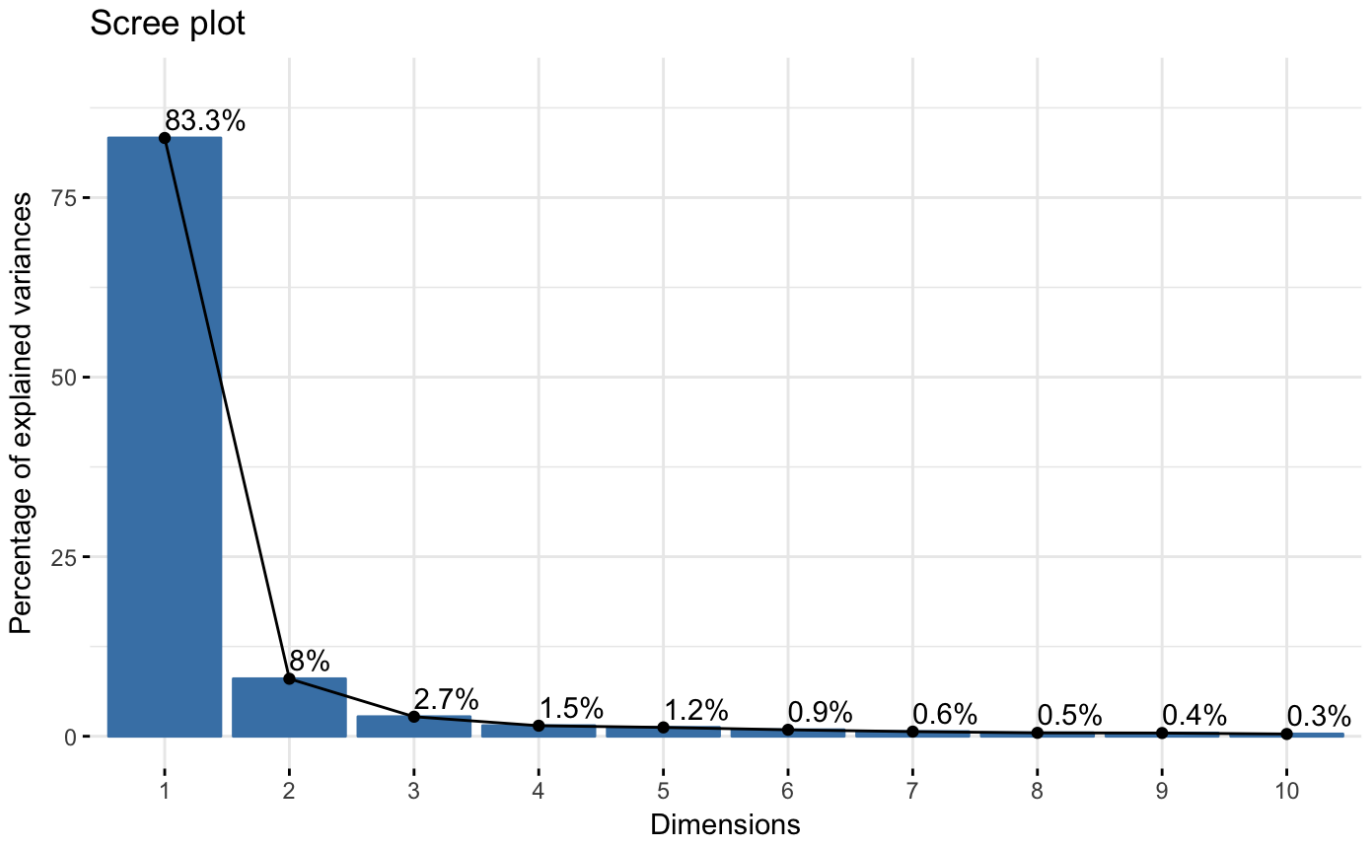
Hide

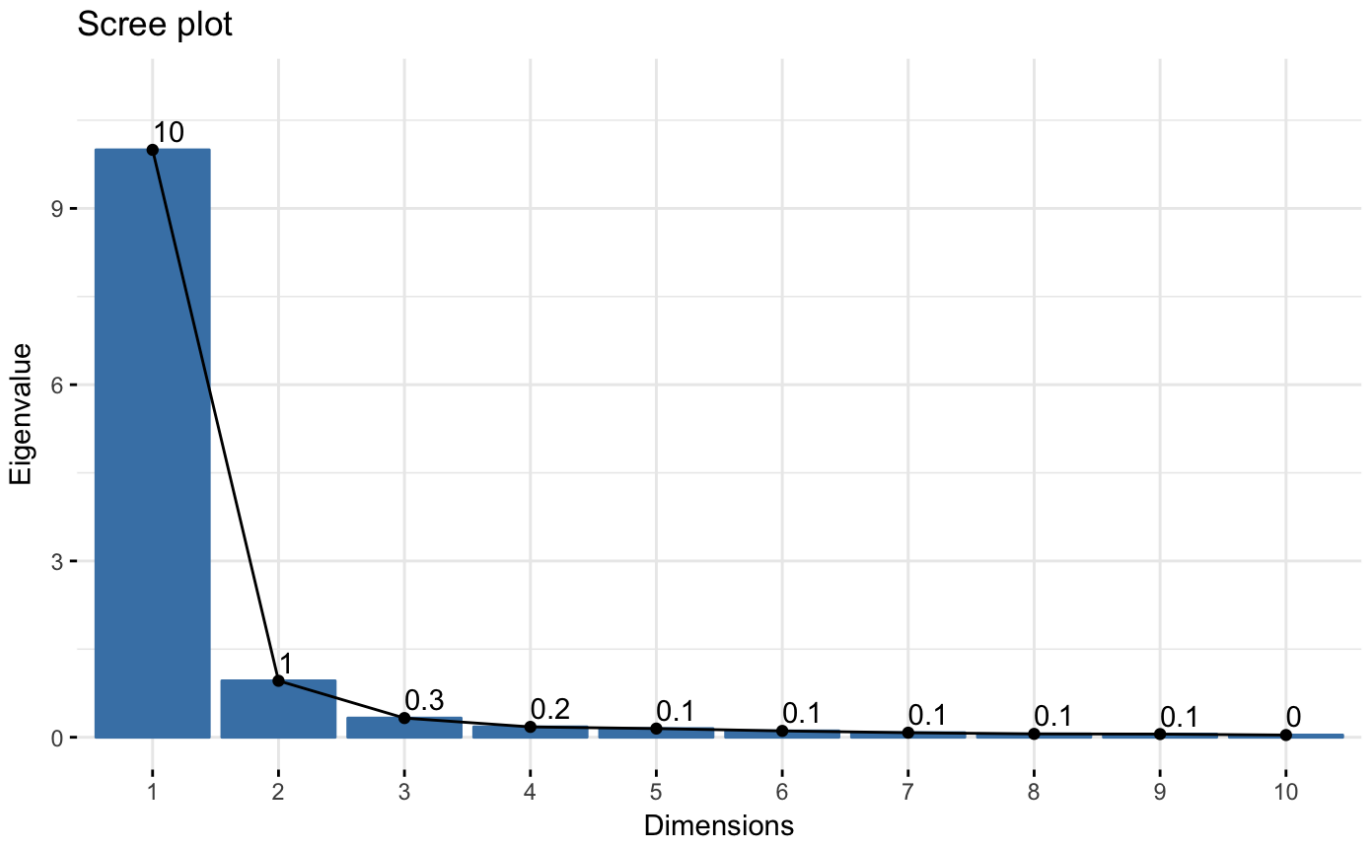
```
summary(k)
```

```
Call:
PCA(X = ville94_12v, quanti.sup = sup$quanti, quali.sup = sup$quali,
     graph = F)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8
  Dim.9  Dim.10  Dim.11  Dim.12
Variance              10.139   0.861   0.325   0.171   0.148   0.097   0.068   0.052
  0.051   0.033   0.031   0.023
% of var.             84.492   7.177   2.707   1.429   1.237   0.811   0.568   0.437
  0.421   0.277   0.257   0.188
Cumulative % of var.  84.492  91.668  94.375  95.804  97.040  97.851  98.419  98.857
99.278  99.555  99.812 100.000

Individuals (the 10 first)
                  Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
  cos2
AbuDhabi94     |  5.212 |  2.168  0.909  0.173 |  4.612 48.422  0.783 |  0.751  3.409
0.021 |
Amsterdam94    |  1.795 |  1.579  0.482  0.774 | -0.357  0.291  0.040 | -0.578  2.018
0.104 |
Athenes94      |  2.058 | -1.912  0.707  0.863 | -0.506  0.583  0.060 | -0.075  0.034
0.001 |
Bangkok94      |  3.142 | -2.971  1.707  0.894 |  0.819  1.527  0.068 | -0.220  0.291
0.005 |
Bogota94       |  2.672 | -2.474  1.183  0.857 |  0.658  0.987  0.061 |  0.138  0.114
0.003 |
Bombay94       |  4.595 | -4.565  4.030  0.987 | -0.310  0.219  0.005 | -0.255  0.394
0.003 |
Bruxelles94    |  0.862 |  0.612  0.072  0.504 | -0.175  0.070  0.041 |  0.250  0.376
0.084 |
Budapest94     |  4.212 | -4.186  3.388  0.987 | -0.242  0.134  0.003 | -0.165  0.164
0.002 |
BuenosAires94  |  2.321 | -0.887  0.152  0.146 |  1.227  3.430  0.280 | -0.309  0.578
0.018 |
Caracas94      |  4.251 | -4.242  3.480  0.996 | -0.065  0.010  0.000 | -0.009  0.000
0.000 |

Variables (the 10 first)
                 Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
instit         |  0.943  8.767  0.889 | -0.042  0.205  0.002 | -0.211 13.725  0.045 |
chauffeur      |  0.957  9.027  0.915 | -0.126  1.845  0.016 | -0.154  7.320  0.024 |
meca           |  0.923  8.401  0.852 | -0.271  8.554  0.074 |  0.193 11.429  0.037 |
man            |  0.898  7.962  0.807 | -0.370 15.921  0.137 |  0.111  3.763  0.012 |
tourneur       |  0.947  8.839  0.896 | -0.244  6.894  0.059 | -0.021  0.141  0.000 |
cuisinier      |  0.867  7.410  0.751 |  0.236  6.456  0.056 |  0.404 50.219  0.163 |
chefserv       |  0.839  6.942  0.704 |  0.487 27.594  0.238 | -0.011  0.037  0.000 |
inge           |  0.902  8.027  0.814 |  0.275  8.760  0.075 | -0.028  0.242  0.001 |
banque         |  0.876  7.563  0.767 |  0.379 16.640  0.143 | -0.129  5.154  0.017 |
secr           |  0.973  9.332  0.946 | -0.001  0.000  0.000 | -0.096  2.836  0.009 |

Supplementary continuous variables
                 Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
htrav          | -0.326  0.107 |  0.300  0.090 |  0.140  0.020 |
alim           |  0.580  0.336 |  0.062  0.004 | -0.012  0.000 |
```

```
salhor         |  0.985  0.970 | -0.135  0.018 | -0.020  0.000 |

Supplementary categories
                  Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2
v.test
region 1       |  1.083 |  0.741  0.468  0.601 | -0.647  0.357 -1.800 |  0.006  0.000
0.027 |
region 2       |  3.492 |  3.441  0.971  3.537 |  0.256  0.005  0.904 | -0.514  0.022
-2.954 |
region 3       |  1.272 | -1.180  0.860 -0.864 | -0.278  0.048 -0.698 |  0.011  0.000
0.044 |
region 4       |  3.927 | -3.903  0.988 -2.167 | -0.235  0.004 -0.448 | -0.220  0.003
-0.681 |
region 5       |  1.199 |  1.111  0.859  0.813 |  0.158  0.017  0.397 |  0.121  0.010
0.496 |
region 6       |  3.135 | -3.127  0.995 -2.290 |  0.014  0.000  0.036 | -0.071  0.001
-0.289 |
region 7       |  2.683 |  2.460  0.840  2.179 | -0.859  0.102 -2.611 |  0.503  0.035
2.489 |
region 8       |  2.800 | -2.736  0.955 -2.218 |  0.452  0.026  1.257 | -0.117  0.002
-0.528 |
region 9       |  2.402 | -0.019  0.000 -0.010 |  2.223  0.856  4.234 |  0.808  0.113
2.505 |
region 10      |  4.267 | -4.237  0.986 -1.901 | -0.278  0.004 -0.428 | -0.149  0.001
-0.373 |
```

Tout d'abord je vais regarder le pourcentage de variance pour les dimensions trouvées par l'acp:

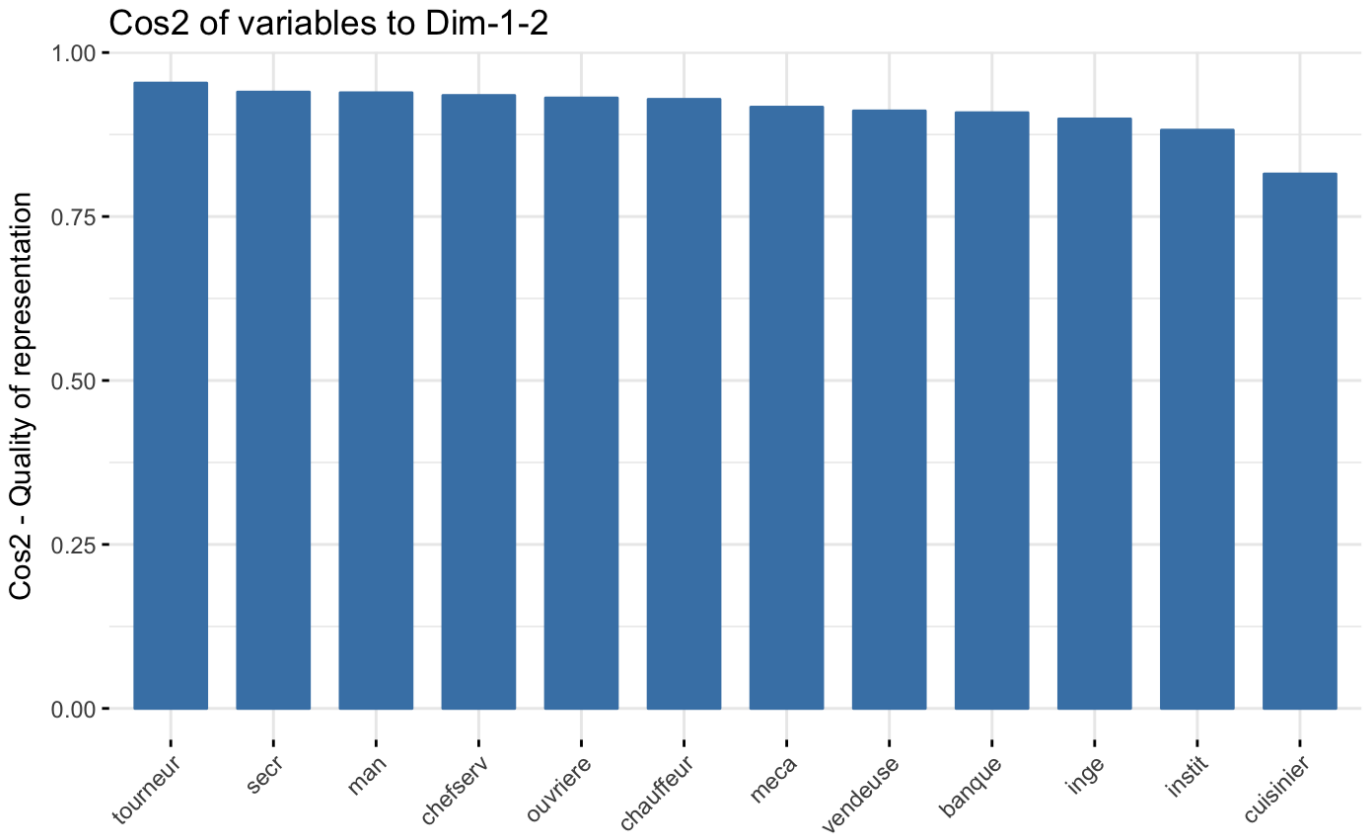


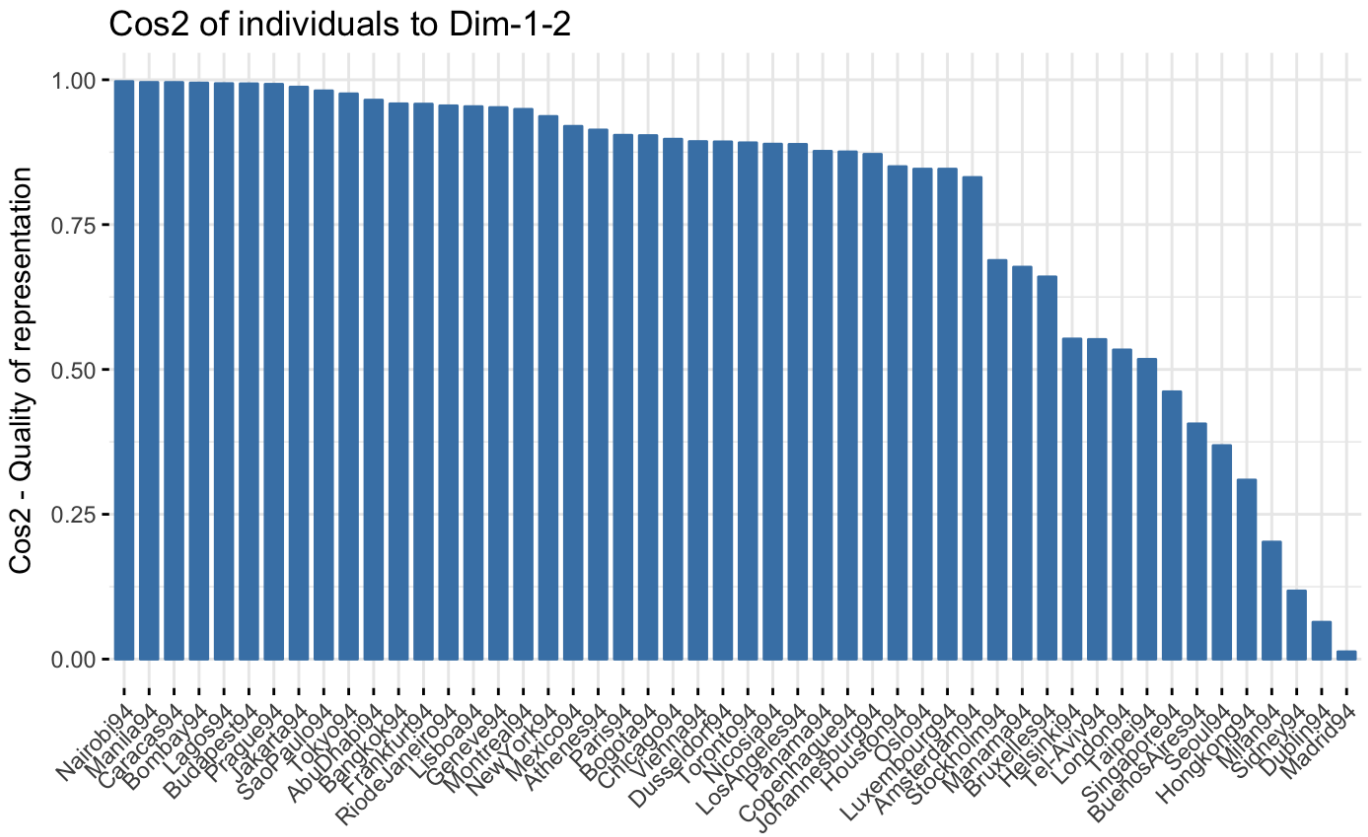


On remarque qu'il y a dans la dimension 1 83.3 % de variance ce qui est beaucoup, avec la dimension 2 on est a 91.3% ce qui est bien. De plus d'apres le critere de Kaiser, nous devons garder seulement les dimensions ayant une valeur propre >1. Donc les dimensions 1 et 2.

Nous allons garder les deux premieres dimensions dans cette etude.

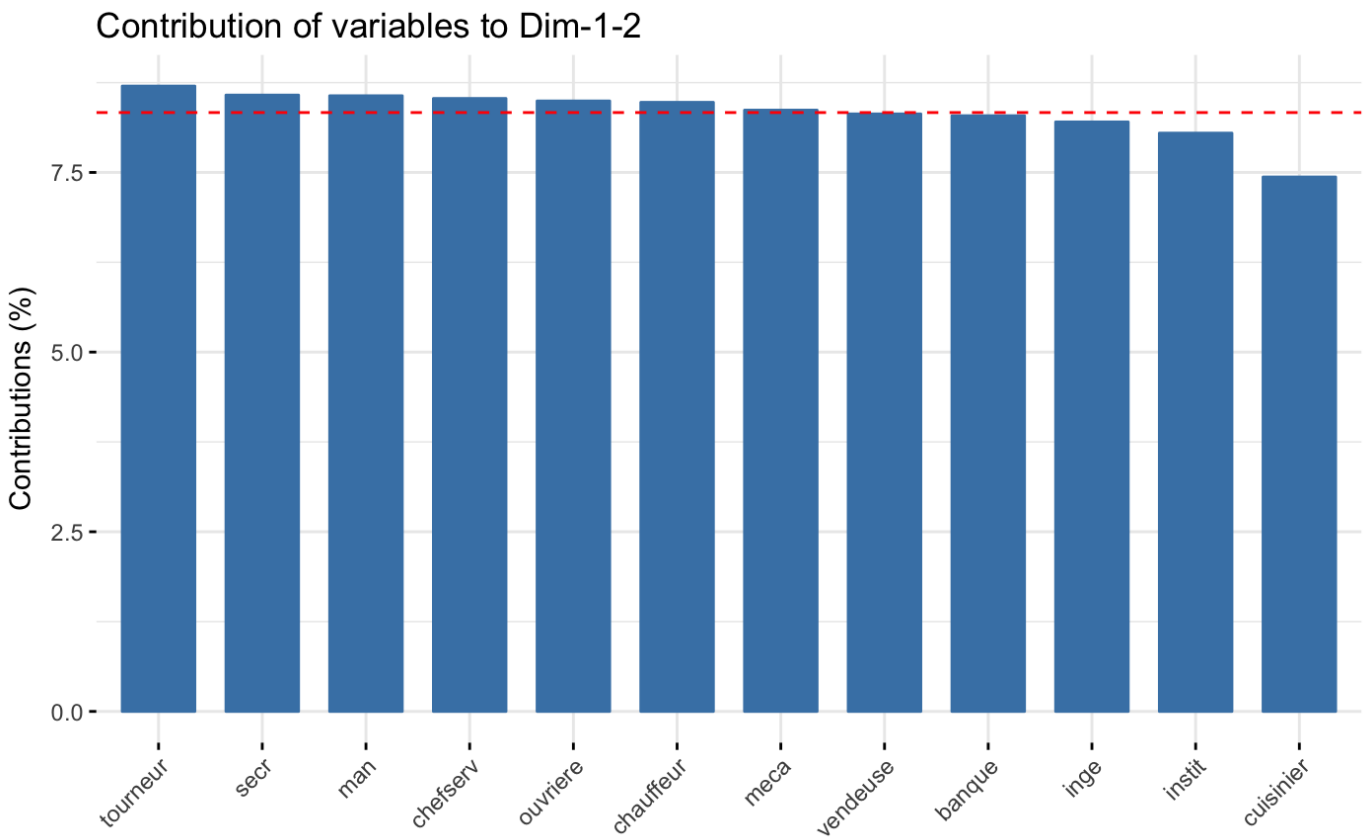
Interressons nous à present aux qualités de projections des variables et individus.

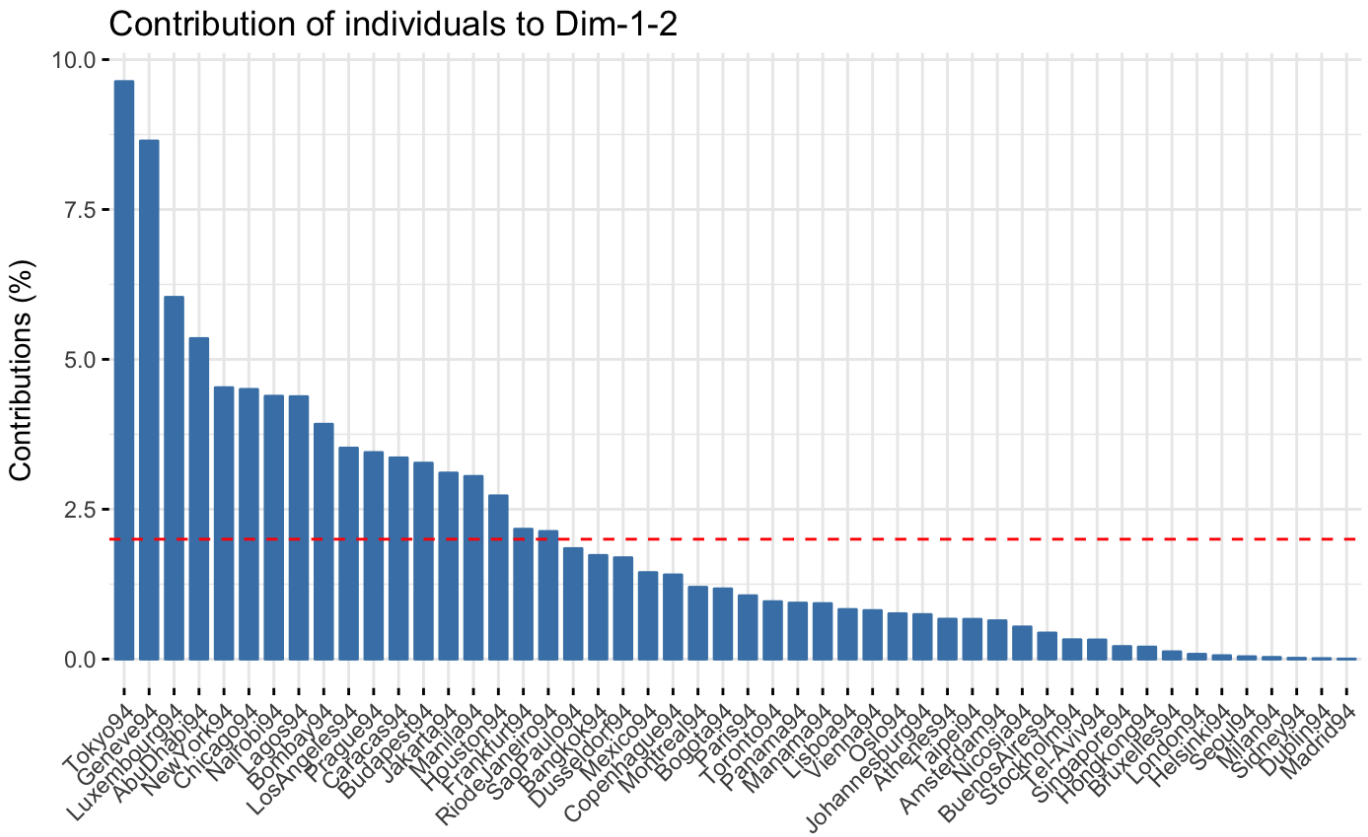




Les qualitées de projection sont bonnes pour toutes les variables (>0.75) Les qualitées de projection sont bonnes pour une bonnes parties des individus. Nous garderons seulement les individus ayant un cos2 > 0.75.

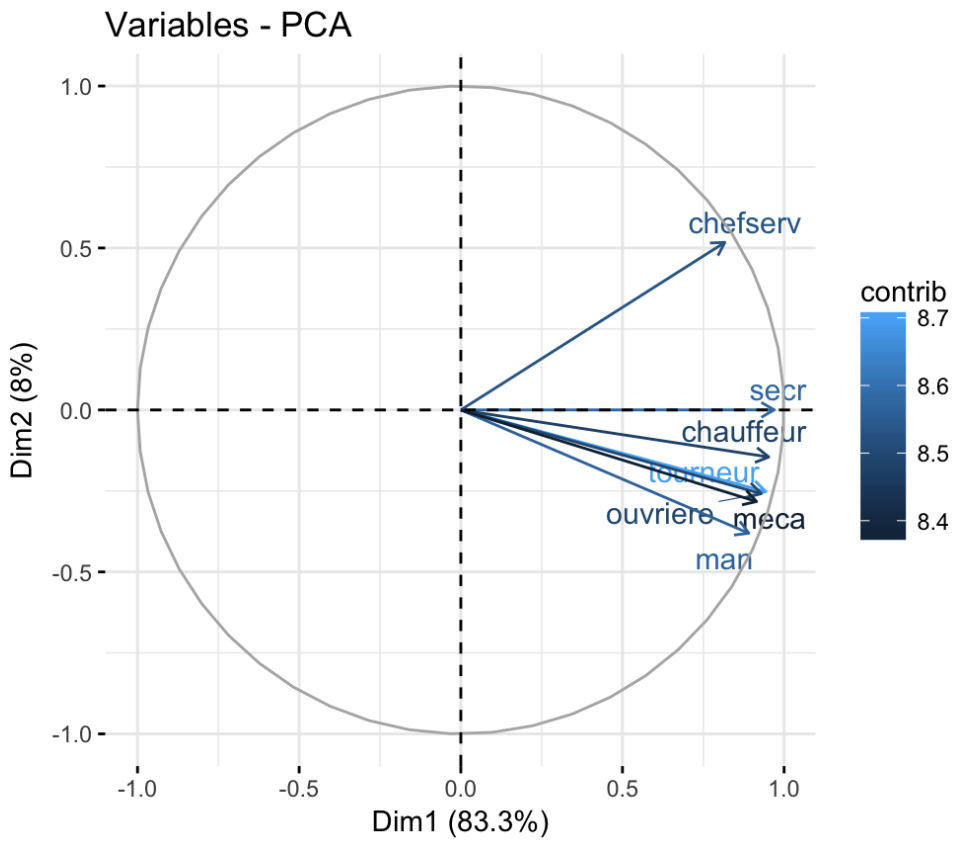
Au sujet des contributions, etudions celle des individus et des variables:

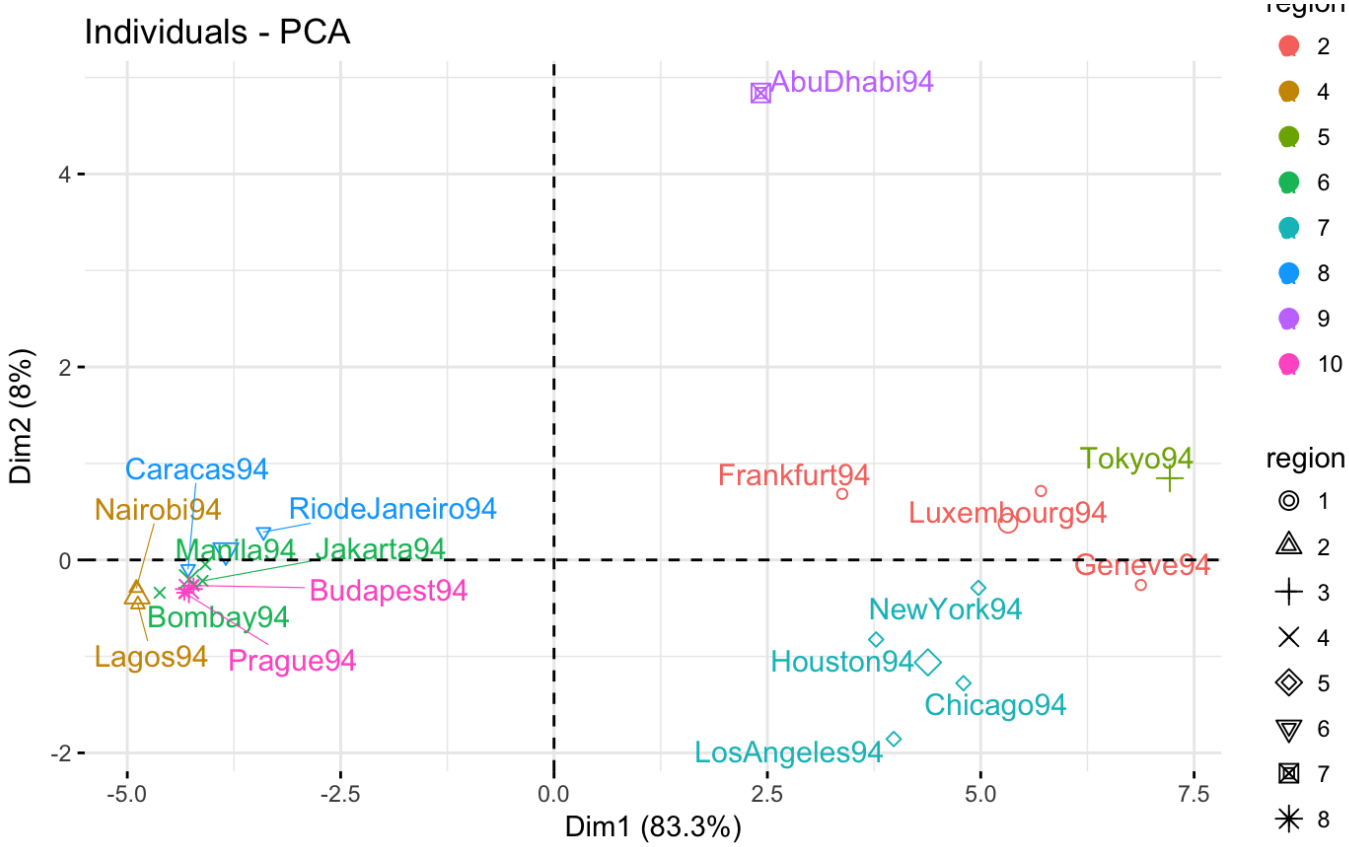




La ligne en pointillés rouge indique la contribution moyenne attendue pour les variables (>=1/12. (0.08333)) (si c'etait une loi uniforme). Nous garderons dans l'acp seulement les variables ayant une contributions >= 0.08333. Pour les individus nous garderons aussi ceux ayant une contributions >= 0.02 (1/50.)

Affichons le premier plan factoriel ainsi que le cercle de correlation de l'axe 1 et 2, avec un cos2 > 0.75 pour n'affichée que les données bien projetées. le premier plan factoriel sera affiché avec des individus ayant un minimun de contrib de 0.02 (top 18) le cercle de correlation sera affiché avec des variables ayant un minimun de contrib de 0.08 (top 7)

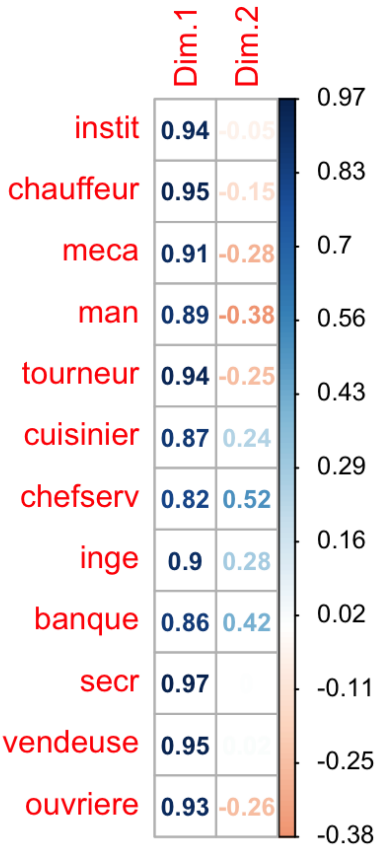




On confirme notre analyse avec la matrice de corrélation, en effet toutes les variables sont correllées positivement.

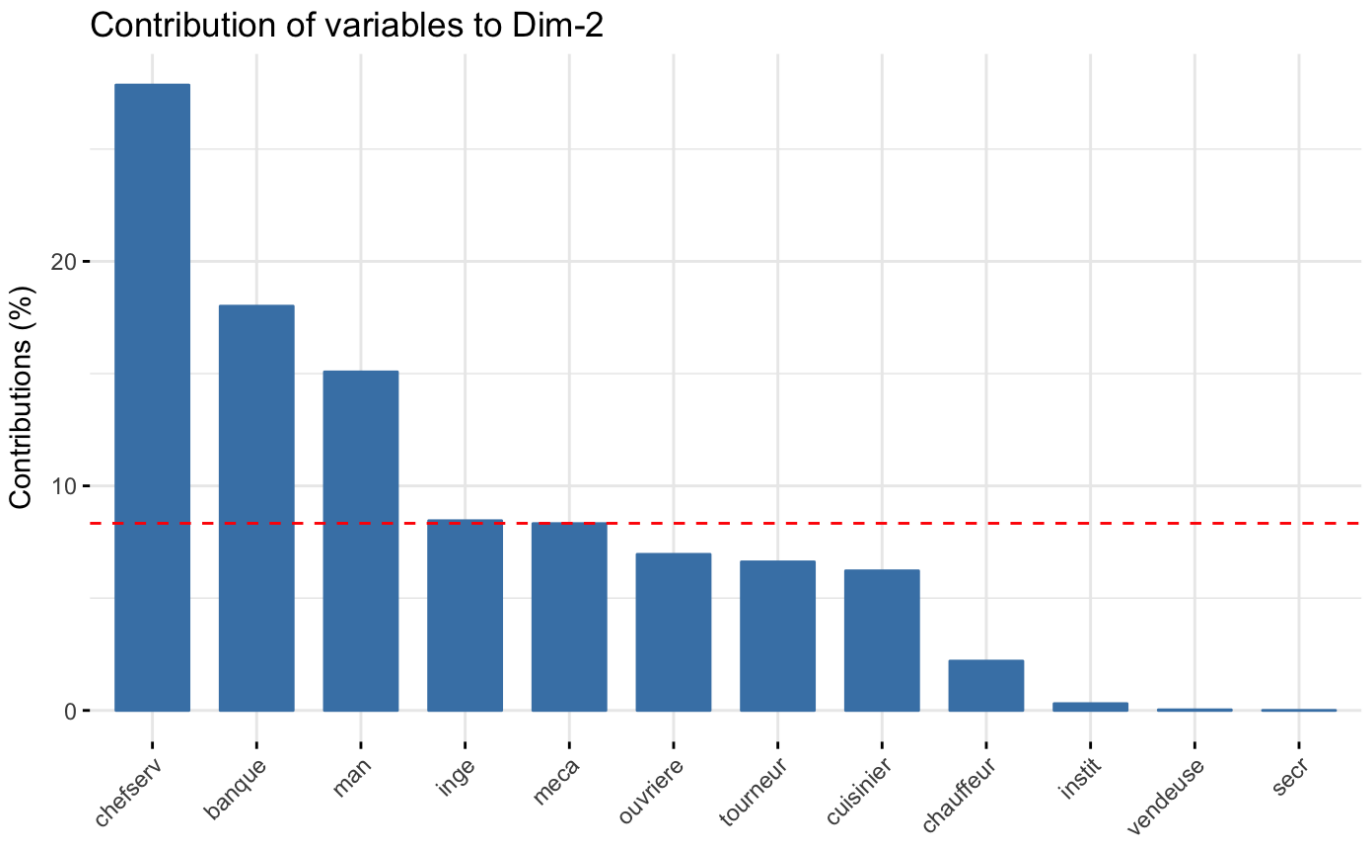
Sur le premiere axe, quasiment toutes les variables lui sont corrélés positivement à plus de 80%. cela veux dire que dans les villes, lorsqu'une de ces variables augmentent toutes augmentent et inversement. Sur le premiere axes nous aurons du coup a droite toutes les villes ayants globalement partout des hauts salaires(Tokyo,Geneve), et à gauche des villes avec globalement partout des bas salaires(Nairobi, Lagos) (partout=tous les metiers)

Interessons nous au corrélations des variables avec les dimensions.

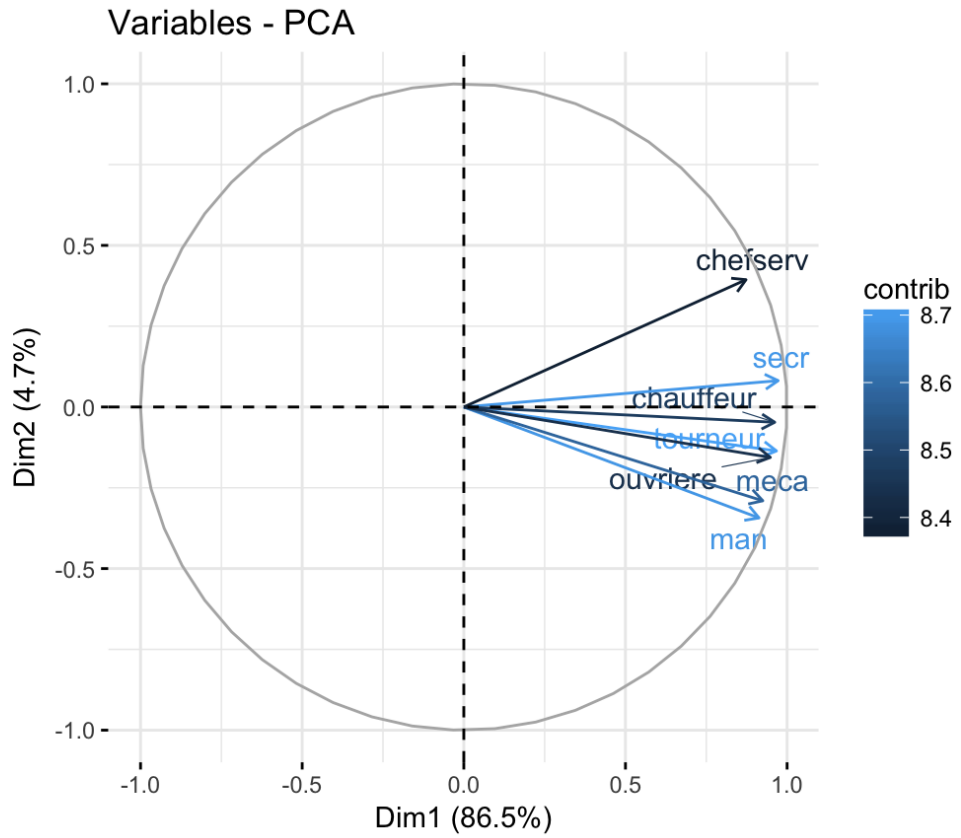


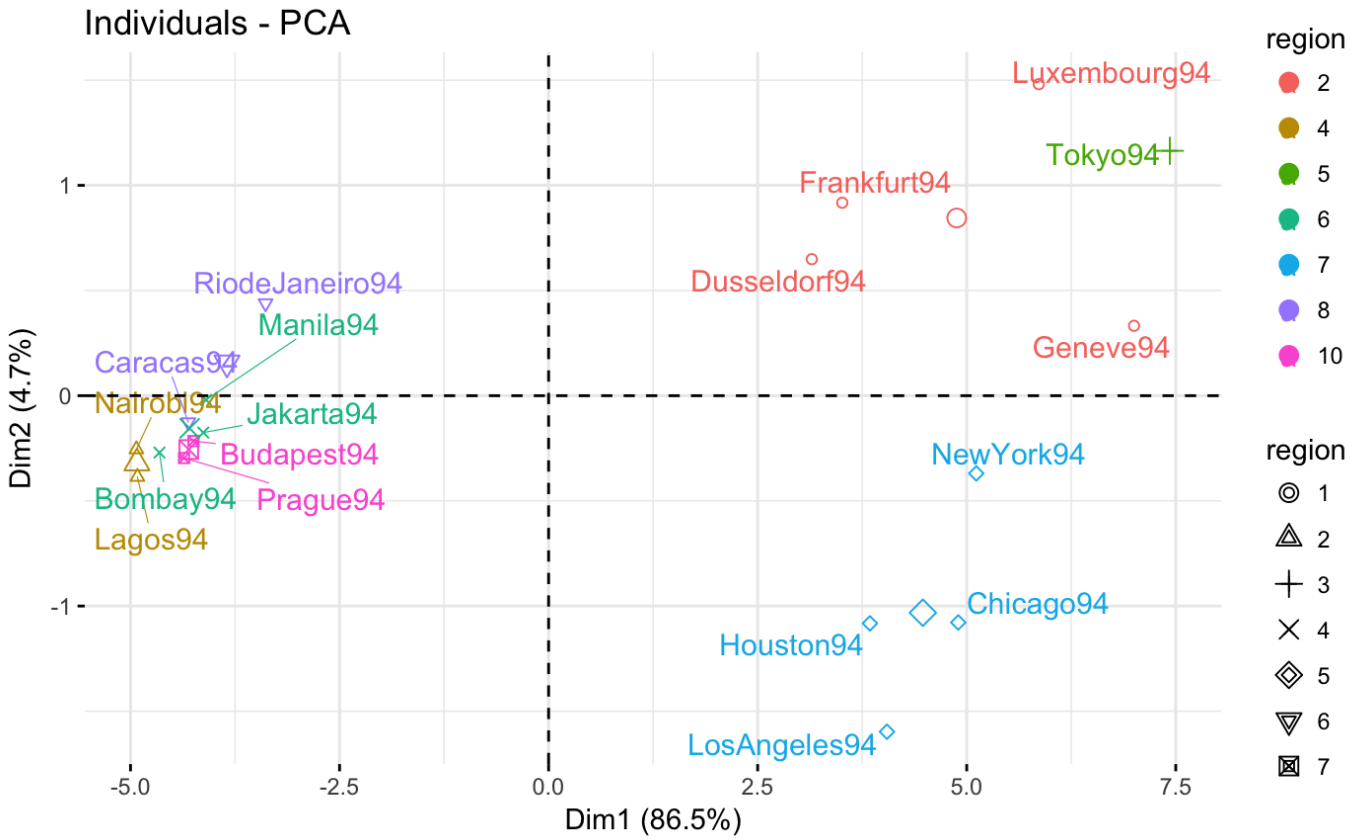
les vars chauffeur, vendeuse et secr sont tres corrélé à la premiere dimension. (nous gardons dans l'analyse seulement les vars chauffeur et secr) Nous rappellons de plus que ces deux variables sont correlé à 93%

Pour le deuxième axe, c'est moins evidents. Mais il y a une variables corrélés suffisament avec lui : chefserv. De plus chefserv contribue beaucoup a cette axe.(+25%) On en deduit que les valeurs ayant de forte valeurs en axe2 auront de bons salaires pour les chefservs. (n'oublions pas que cette dimension n'a que 8% , c'est peu est pas tres representatif).



Au sujet des individus, on remarque en tout premier lieu , l'excentricité de l'individus AbuDabi,nous recommencons l'acp, sans elle, pour voir d'eventuel changements.
De plus en regardant les individus par rapport au premiere axe,on distingue deux groupes d'individus: ceux avec des salaires moyen bas et ceux avec des salaires moyen haut. De plus nous remarquons que cela correspond aux regions auquelle appartienne ses individus. (regions en developpement, ou pauvre a gauche (afrique, amerique du sud, asie de sud est,Europe de l'Est), et regions developpées à droite(Amerique du Nord, Europe Centrale, Asie de l'est))





On voit que l'acp (cercle) est presque identique,sinon pour les individus, plutot 3 groupes se forment, distinguant l'amerique du Nord(USA) avec Europe Centrale et Asie de l'est.

A propos des variables supplementaire. On ne peux qu'interpreter "salhor" cat c'est la seul variable bien projeté. Elle confirme la corrélation positives entre toutes les variables de salaires.