Projet ACP Sujet 4

Introduction :

Nous avions comme objectif d’effectuer une ACP sur une base de données qui nous a été fournie. Notre base de données a pour individus des villes de France au nombre de 29, et des informations sur celles-ci, qui sont nos 11 variables. Comme nos différentes variables avaient des données de différentes unités (certaines étaient en pourcentage par exemple) et échelle, et une variable qualitative nous avons décidé d’effectuer une ACP normée.

Modification :

Tout d’abord après avoir téléchargé le fichier Excel qui nous a été attribué nous avons dû en créer une copie en .txt afin de pouvoir l’importer sur R.

Nous avons ensuite effectué quelques changements sur ce fichier .txt. D’abord nous avons remplacé toutes les virgules par des points pour que R puisse lire correctement les séparateurs. Ensuite nous avons enlevé le titre et le mot ville et qui était dans la première ligne sinon R aurait pris la colonne des villes comme une variable et non comme les individus. De plus nous avons modifié le nom des variables pour les raccourcir pour éviter les décallages et que cela soit plus clair.

Explication du code :

* villes<-read.table("Sujet4.txt",header=T)

Cette ligne permet de donner la valeur de ville au sujet4.txt qui se trouve dans nos fichiers. Cela permet de l’appeler plus simplement.

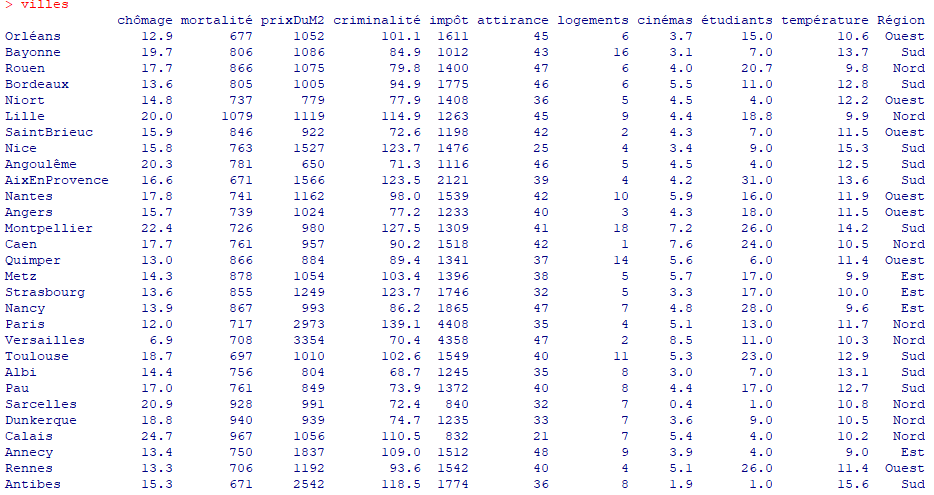
Il ressort :



* villes

Ici on appelle ville et cela permet d’afficher les données que nous allons étudier.

Il ressort :



* dim(villes)

Cette ligne permet de ressortir le nombre d’individus de notre base de données (les lignes) et le nombre de variable de notre base de données (les colonnes).

Ici il nous ressort :

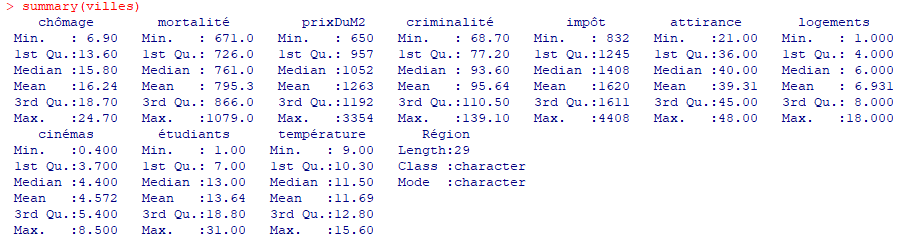


Notre base de données se compose donc de 29 individus et de 11 variables.

* summary(villes)

Cela nous permet de ressortir pour chacune des variables suivantes : le min et max, la médiane, le 1er et 3ème quartile et la médiane.

De plus il nous permet de connaître nos variables qualitatives en affichant, class : character. Grâce à cela nous avons pu voir que notre variable “Région” était une variable qualitative.



On effectue une méthode pour afficher la matrice de corrélation.

* villes1<-villes[,1:10]

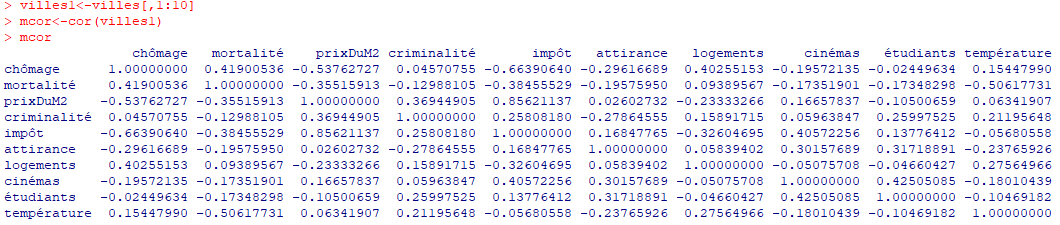
Ici villes1 prend la valeur de notre base de données sans notre variable qualitative.

* mcor<-cor(villes1)

Ici mcor prend la valeur de la matrice de corrélation.

* mcor

On affiche la matrice de corrélation.



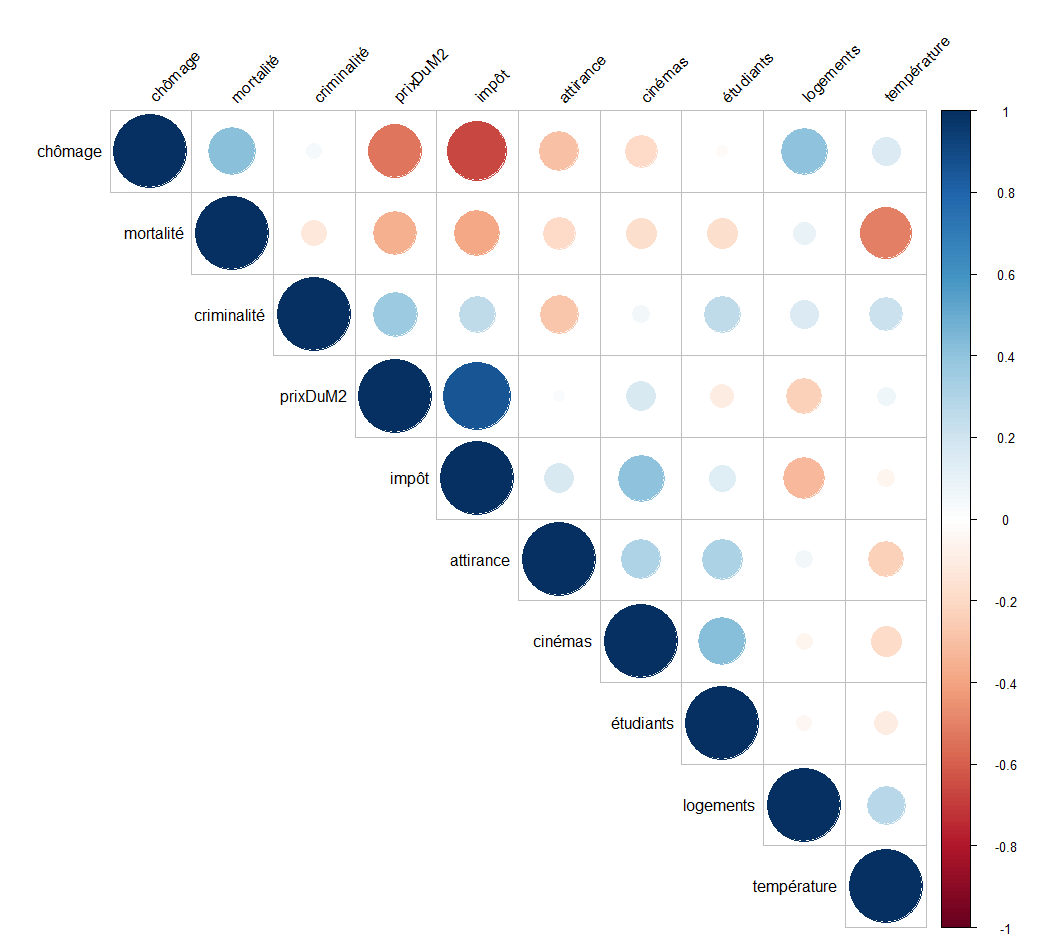
Afin de pouvoir plus facilement l’analyser nous allons faire une représentation graphique de notre matrice de corrélation en effectuant la méthode corrplot suivante.

* install.packages("corrplot")

Nous n’avions pas installé le package corrplot nous avons donc dû l’installer.

* library(corrplot)
* corrplot(mcor,type="upper",order="hclust",tl.col="black",tl.srt=45)

Grâce à cette méthode le graphique suivant ressort :



Avec ceci nous pouvons remarquer qu’il y a une assez forte corrélation entre les variables prix du mètre carré et impôts, nous pouvons en déduire que si le prix du mètre carré augmente, les impôts augmenteront aussi.

De plus nous pouvons remarquer que les variables chômage et impôts sont corrélées négativement. Ceci sous-entendrait que si le taux de chômage augmente les impôts diminueront.

Aussi nous pouvons constater que les variables prix du mètre carré et attirance, et chômage et étudiants, sont décorrélés entre eux.

En observant notre matrice de corrélation et notre graphique nous pouvons constater que dans la globalité nos variables sont faiblement corrélées entre elles. Ce qui est plutôt logique car nos variables ne sont pas directement liées entre elles, en effet aucune de nos variables ne sont de même unité.

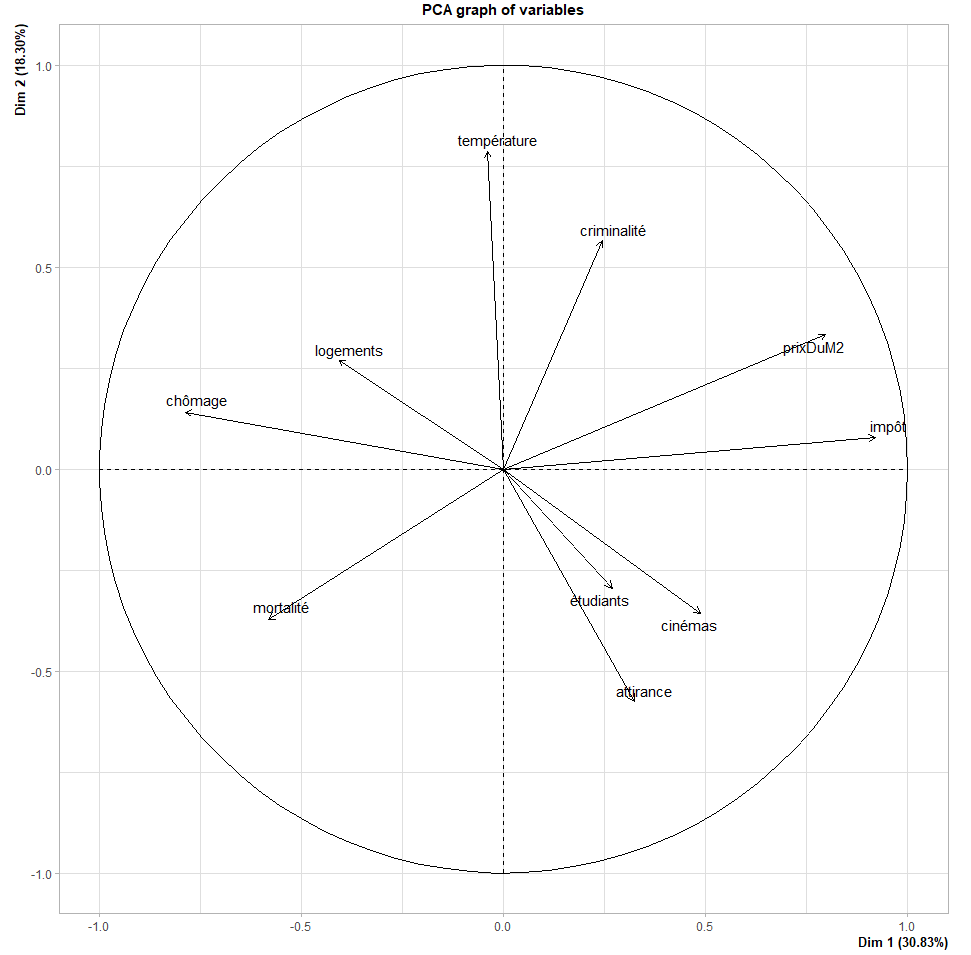
Nous allons ensuite faire une représentation graphique de nos variables selon nos individus :

* library(FactoMineR)

Pour avoir cette représentation nous devons utiliser le package FactoMineR. Nous l’avions déjà installé donc nous avons juste à le lancer.

* res.pca=PCA(villes[,1:10],scale.unit=TRUE,ncp=5,graph=T)

Grâce à ce code nous pouvons avoir une représentation de nos variables :



Nous avons observé que la variable “Logement” était trop éloignée du cercle unité donc il n’est pas intéressant.

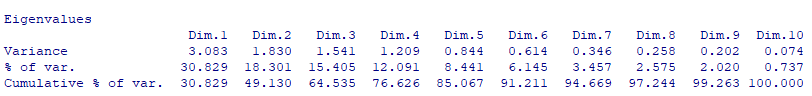
Aussi comme dis précédemment les variables “Chômage” et “Impôts” sont diamétralement opposés et sont donc corrélées négativement.

De plus les variables “Cinémas” et “Criminalités” sont assez orthogonales, nous pouvons donc en déduire qu’elles sont décorrélées.

Afin de savoir combien d’axe garder nous avons effectué une méthode pour récupérer les valeurs propres :

* summary(res.pca)

Dans ce qui s’affiche nous pouvons récupérer les valeurs propres. Comme nous sommes en ACP normée nous utilisons la règle de Kaiser et nous pouvons garder que les valeurs propres supérieures à 1, soit :



Nous ne gardons donc que les 4 premiers axes. De plus ils représentent 76,626% de l’inertie totale.

Nous connaissons les axes à garder, nous pouvons donc changer dans le code du “res.pca” le ncp, qui était égal à 5 de base, et mettre ncp=4.

De plus comme nous disposons de 11 variables et seulement 10 sont représentées dans notre graphe, nous ajoutons notre 11ème variable qui est qualitative, la variable “Région”, grâce au code suivant :

* res.pca <-PCA(villes, quali.sup=11, scale.unit=TRUE, ncp=4, graph=T)

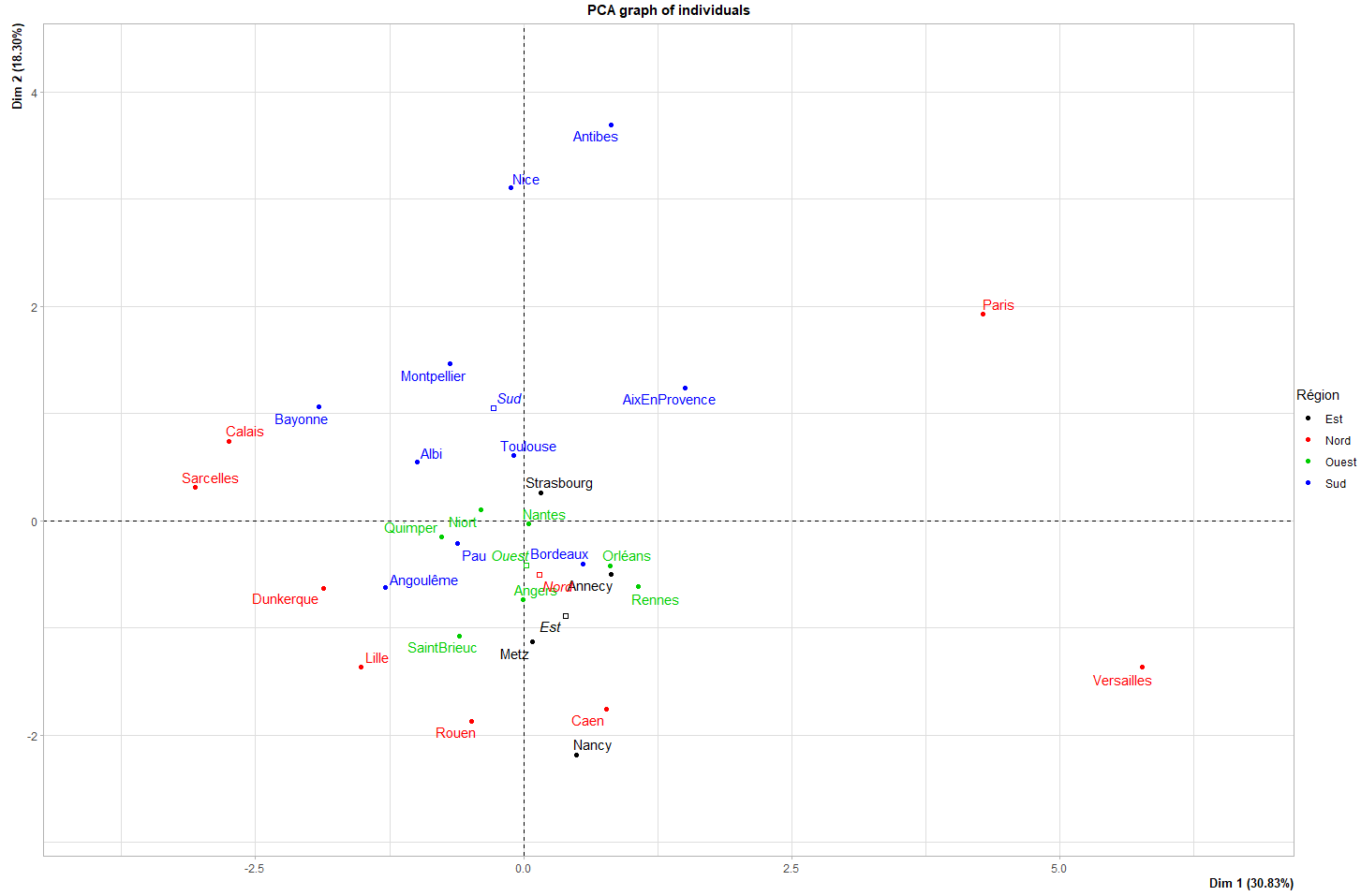
Grâce à “quali.sup” nous rajoutons la variable “Région” en tant que variable qualitative, et nous obtenons le même graphe car c’est une représentation de nos variables.

Nous voulons voir nos individus. Nous écrivons donc les lignes suivantes et les exécutons :

* plot.PCA(res.pca, axes=c(1, 2), choix="ind", habillage=11)
* plot.PCA(res.pca, axes=c(1, 3), choix="ind", habillage=11)
* plot.PCA(res.pca, axes=c(2, 3), choix="ind", habillage=11)

Pour voir les individus nous mettons choix=”ind” et comme nous avons 4 axes nous devons choisir les axes représentés, axes=c(axe, axe).

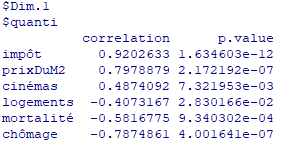
Par exemple pour la représentation des individus sur l’axe 1 et l’axe 2 (le 1er code) nous obtenons :



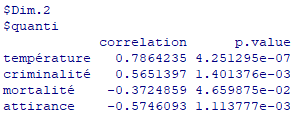
Grâce à la méthode suivante nous avons pu voir quelles variables étaient les plus corrélées avec les axes 1 et 2.

* dimdesc(res.pca, axes=c(1,2))

Nous avons donc obtenu que les variables les plus corrélées avec l’axe 1 sont les suivants, avec leur corrélation :



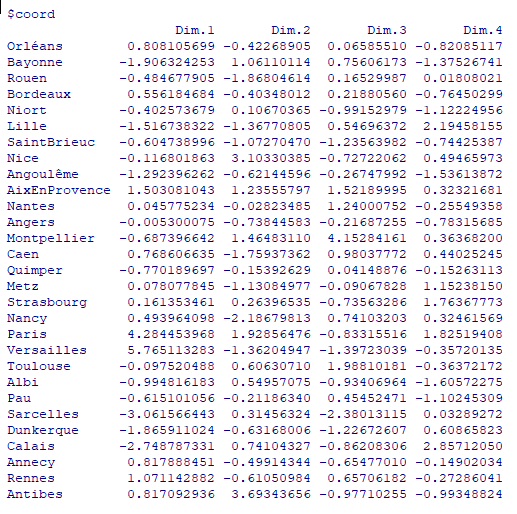
Et les variables les plus corrélées avec l’axe 2 sont les suivantes, avec leur corrélation :



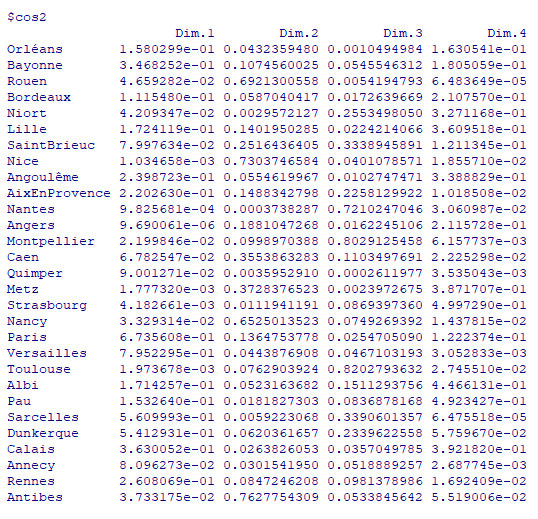
Ensuite grâce à :

* res.pca$ind

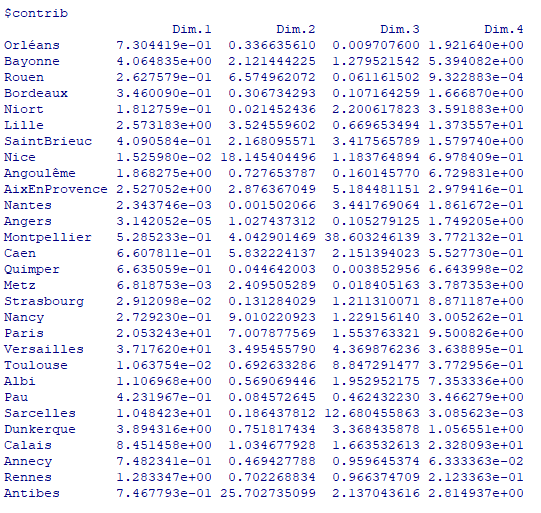
Nous avons pu obtenir les coordonnées des individus sur les 4 axes retenus :



Ainsi que la qualité de représentation des individus sur les 4 axes retenus :



Et pour finir la contribution des individus sur les 4 axes retenus :

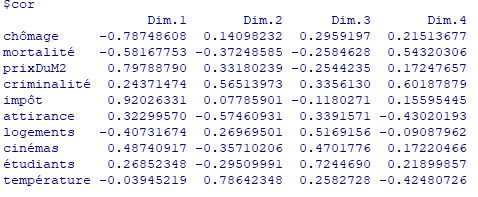


Avec la commande :

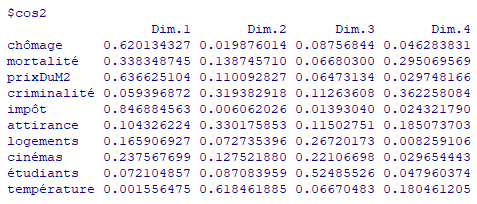
* res.pca$var

Nous obtenons des informations entre les variables et les 4 axes retenus, soit :

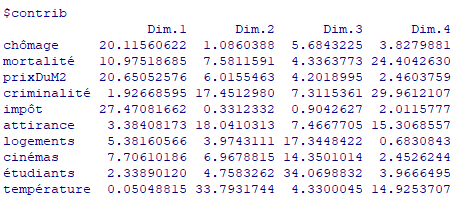
La corrélation/coordonnées



La qualité de représentation



La contribution



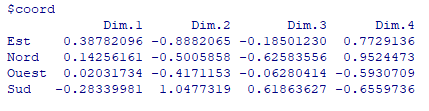
Ici on remarque que le chômage, les impôts, la mortalité et le prix du m² sont les variables qui contribuent le plus à l’axe 1. Pour l’axe 2 se sont la température, l’attirance des actifs et la criminalité qui contribuent le plus. Les logements, les cinémas et les étudiants contribuent le plus à l’axe 3 et la mortalité, la criminalité, l’attirance des actifs et la température contribuent le plus à l’axe 4.

Avec la méthode :

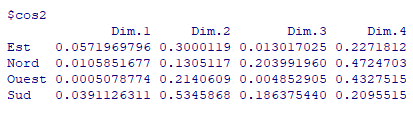
* res.pca$quali.sup

Nous avons pu obtenir des informations entre la variable qualitative ajoutée, “Région”, et les 4 axes retenus, soit :

Les coordonnées



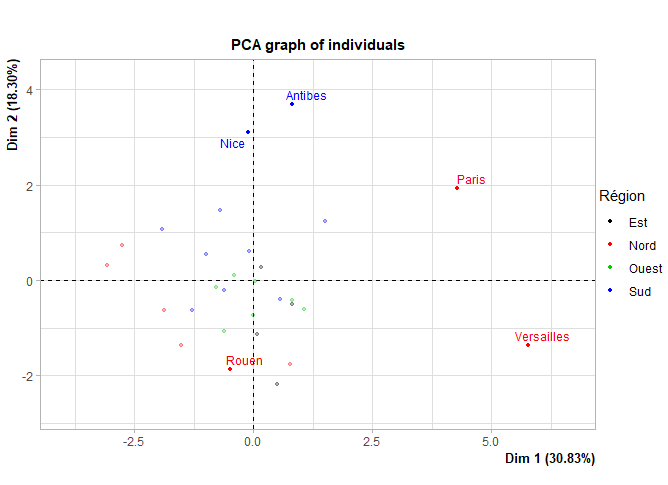
La qualité de représentation



Nous allons maintenant nous intéresser aux individus qui sont le mieux représentés à l’aide de la méthode :

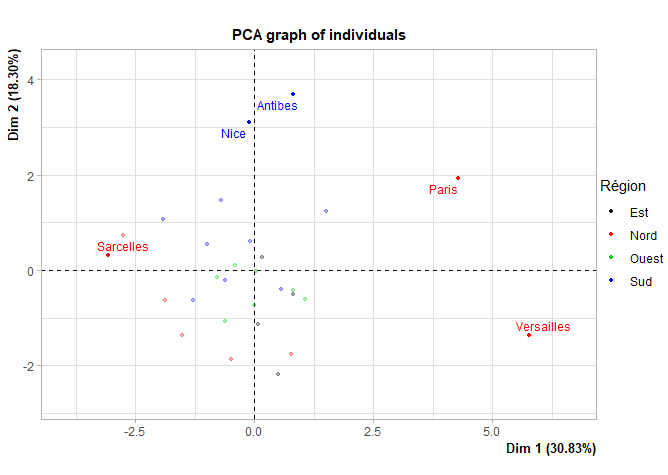
* x11( )
* plot(res, cex=0.8, shadow=TRUE, habillage=11, select="cos2 0.7", invisible=c("quali"))

On choisit la police grâce à “cex”, on met en couleur les individus selon leur région grâce à “habillage=11”, on sélectionne les individus qui ont une qualité de représentation supérieure à 0,7 grâce à “select="cos2 0.7"”, et on met en invisible la variable région grâce à “invisible=c("quali"))”.



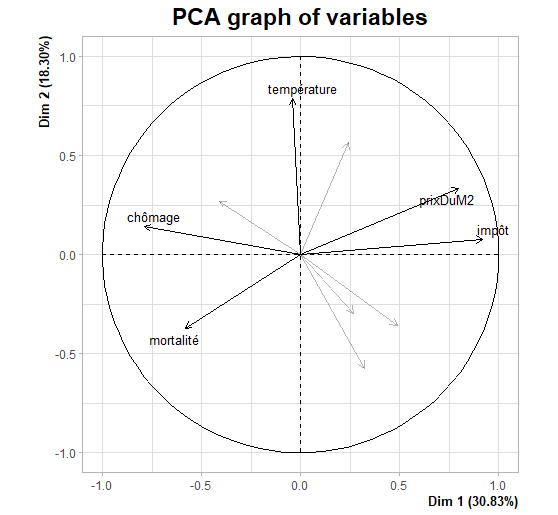
Nous choisissons ensuite de sélectionner les 5 individus qui ont le plus de contribution sur les axes 1 et 2, grâce à :

* plot(res, cex=0.8, shadow=TRUE, habillage=11, select="contrib 5")



Nous avons ensuite décidé d’afficher les 5 variables qui contribuaient le plus à la formation des axes grâce à :

* plot(res, choix”var”, shadow=”TRUE”, select=”contrib 5”, cex=0.9, cex.main=1.5)



On peut interpréter ces résultats. Tout d’abord on retient les individus ayant le plus contribués aux axes en utilisant la racine carrée de alpha multiplié à la valeur propre. On prendra alpha = 1.

On se fixe un seuil de 70% pour lequel une variable est considérée comme ayant contribuée significativement à la création d’un axe.

|  |  |
| --- | --- |
| AXE | 1 |
| - | + |
| Bayonne  Sarcelles  Dunkerque  Calais | Paris  Versailles |

Ici on peut en déduire que Paris et Versailles ont un prix du m² plus élevés et les personnes y habitants payent plus d’impôts qu’à Bayonne, Sarcelles, Dunkerque et Calais qui elles ont un taux de chômage plus élevés.

|  |  |
| --- | --- |
| AXE | 2 |
| - | + |
| Rouen  Lille  Caen  Nancy  Versailles | Paris  Nice  Aix en Provence  Montpellier  Antibes |

Ici on voit que la température est généralement plus élevés à Paris, Nice, Aix en Provence, Montpellier et Antibes qu’à Rouen, Lille, Caen, Nancy, Versailles.

Pour finir nous avons effectué un diagramme circulaire sur les régions afin d’avoir une représentation visuelle de la part des individus dans chaque région.

* region<-c(31,21,50,10)
* names(region)<-c(“Nord”, “Sud”, “Ouest”, “Est”)
* pie(region, main = “Diagramme circulaire pour région”, col = rainbow(lenght(region)))

