Ce notebook s'agit d'une ACP pour sst

Charger les données :

on change l'index pour qu'il soit les noms puis on supprime cette variable:

```
sst<-read.csv('data_acp_time-series-sst.csv',header = T)
rownames(sst)<-sst$index
sst$index<-NULL
#str(sst) # donne 656 obs et 288 variables
```

Charger les libraires :

```
library("FactoMineR")

## Warning: package 'FactoMineR' was built under R version 3.4.2

library("factoextra")

## Warning: package 'factoextra' was built under R version 3.4.2

## Loading required package: ggplot2

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

library("qtlcharts")

## Warning: package 'qtlcharts' was built under R version 3.4.2
```

Corrélation entre les variables : on choisit que les variables numériques

```
cor(sst)[1:5,1:5]

##            X1993_1  X1993_10  X1993_11  X1993_12   X1993_2
## X1993_1  1.0000000 0.8770626 0.9492017 0.9872703 0.9973695
## X1993_10 0.8770626 1.0000000 0.9769598 0.9141602 0.8559890
## X1993_11 0.9492017 0.9769598 1.0000000 0.9768533 0.9361258
## X1993_12 0.9872703 0.9141602 0.9768533 1.0000000 0.9833123
## X1993_2  0.9973695 0.8559890 0.9361258 0.9833123 1.0000000
```

La table de corrélation montre l'existence de plusieurs variables qui sont corrélées entre eux. Ceci est une bonne nouvelle dans le sens où l'ACP nécessite que les variables soient corrélées afin d'extraire de l'information contenue dans l'inertie totale.

On peut donc commencer l'analyse en composantes principales.

ACP

on va utiliser toutes les variables quantitatives (288 mois) pour effectuer notre ACP. Ainsi que tous les individus (656 individus).

```
res.pca=PCA(sst,graph=FALSE,scale.unit = FALSE)
print(res.pca)

## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 656 individuals, described by 288 variables
## *The results are available in the following objects:
##
##    name                description
## 1  "$eig"              "eigenvalues"
## 2  "$var"              "results for the variables"
## 3  "$var$coord"        "coord. for the variables"
## 4  "$var$cor"          "correlations variables - dimensions"
## 5  "$var$cos2"         "cos2 for the variables"
## 6  "$var$contrib"      "contributions of the variables"
## 7  "$ind"              "results for the individuals"
## 8  "$ind$coord"        "coord. for the individuals"
## 9  "$ind$cos2"         "cos2 for the individuals"
## 10 "$ind$contrib"      "contributions of the individuals"
## 11 "$call"             "summary statistics"
## 12 "$call$centre"      "mean of the variables"
## 13 "$call$ecart.type"  "standard error of the variables"
## 14 "$call$row.w"       "weights for the individuals"
## 15 "$call$col.w"       "weights for the variables"
```

res.pca est un objet qui contient plusieurs variables à analyser. Dans ce qui suit on va analyser chaque attribut de cette objet.

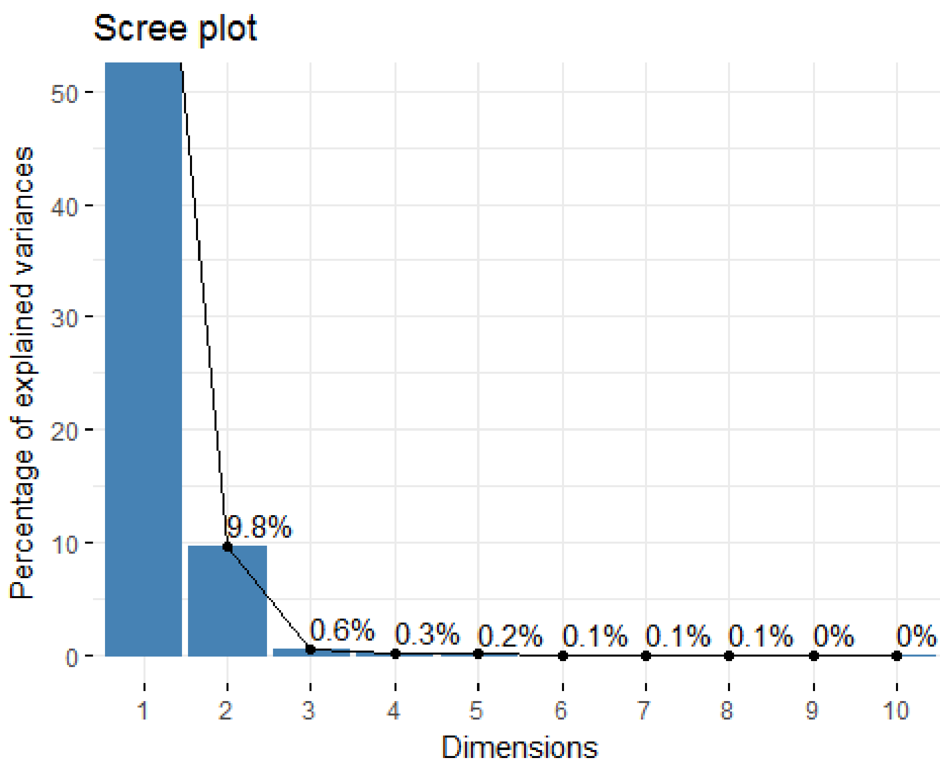
```
eig.val<-get_eigenvalue(res.pca)
head(eig.val)

##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1 3530.987713       88.4303476                    88.43035
## Dim.2  390.007182        9.7673720                    98.19772
## Dim.3   23.104044        0.5786196                    98.77634
## Dim.4   10.835955        0.2713766                    99.04772
## Dim.5    6.745755        0.1689412                    99.21666
## Dim.6    4.605448        0.1153392                    99.33200
```

D'après les résultats ci-dessus, on peut conserver 2 axes principales vu qu'ils expliquent 98.19% de l'inertie totale contenue dans notre jeu de données. A noter qu'ici on n'a pas le droit d'utiliser le critère de Kaîser pour choisir le nombre d'axe à conserver, car les données ne sont pas normalisées. On préfére la normalisation dans le cas où les variables sont hétérogénes.

==
===

Critére de Kaiser (1961) : est utilisé pour déterminer le nombre d'axes principaux à gardr après l'ACP. Une valeur propre> 1===> la CP en question represente plus de variance par rappot à une seule variable d'origine. Ceci est valable que si les données sont normalisées.
==
==
Afin de bien justifier notre choix on peut fournir les diagrammes suivants :

```
fviz_eig(res.pca,addlabels=T,ylim=c(0,50))
```



On peut, à partir du graph, se limiter à 2 composantes principales, soit 98.19% de la variance totale. Après ces deux axes, la variance cumulée ne change pas beaucoup.

Etude des variables :

Corrélation avec les CPs, Qualité de représentation, contributions aux CPs et cercles de corrélations.

Decription des dimensions :

dimdesc: est utilisée pour identifier les variables les plus significativement associées avec une CP donnée.

```
res.desc<-dimdesc(res.pca, axes=c(1,2), proba=0.05)
print('=====pour Dim.1======')

## [1] "=====pour Dim.1======"
```

```
res.desc$Dim.1$quanti[1:10,]
```

```
##           correlation p.value
## X2013_11    0.9901403       0
## X1997_10    0.9897789       0
## X2010_10    0.9891522       0
## X2003_10    0.9888726       0
## X2012_10    0.9886525       0
## X1999_5     0.9885681       0
## X2001_11    0.9884227       0
## X2009_10    0.9883685       0
## X2011_11    0.9882887       0
## X2001_10    0.9881368       0
```

```
print('=====pour Dim.2======')
```

```
## [1] "=====pour Dim.2======"
```

```
res.desc$Dim.2$quanti[1:5,]
```

```
##          correlation      p.value
## X2006_2    0.5559942 1.698791e-54
## X2010_2    0.5522167 1.233108e-53
## X2003_2    0.5449528 5.197615e-52
## X2011_2    0.5368994 2.960573e-50
## X2013_2    0.5229753 2.495742e-47
```

```
tail(res.desc$Dim.2$quanti)
```

```
##          correlation      p.value
## X2014_7   -0.4085847 8.716682e-28
## X2002_7   -0.4247202 4.101725e-30
## X1997_8   -0.4316750 3.717525e-31
## X1994_8   -0.4371488 5.399647e-32
## X1994_7   -0.4523604 2.099455e-34
## X2002_8   -0.4788597 6.635456e-39
```

Conclusion : - Les variables les plus corrélées avec la Dim.1 (CP1) sont les mois suivantes : 2013_11,1997_10,2010_10,2003_10,2012_10,1999_5,2001_11==> Dim1 peut être interprété comme étant la moyenne de sst sur ces mois. Remarque: Le premier axe prinicipale est très corrélés avec les mois 10,11 et 5. Intérprétation :

- Les variables les plus corrélées positivement avec Dim.2 (CP2) sont les mois suivantes : 2006_2,2010_2,2003_2, 2011_2 et 2013_2. On voit également ici que cet axe corréle fortement avec le mois 2.

Interprétation : - Les variables les plus corrélées négativement avec Dim.2 sont les mois suivantes : 2014_7,2002_7,1997_8, 1994_8 et 1994_7. L'axe principale 2 corréle négativement avec les mois 7 et 8.

Intérprétation :

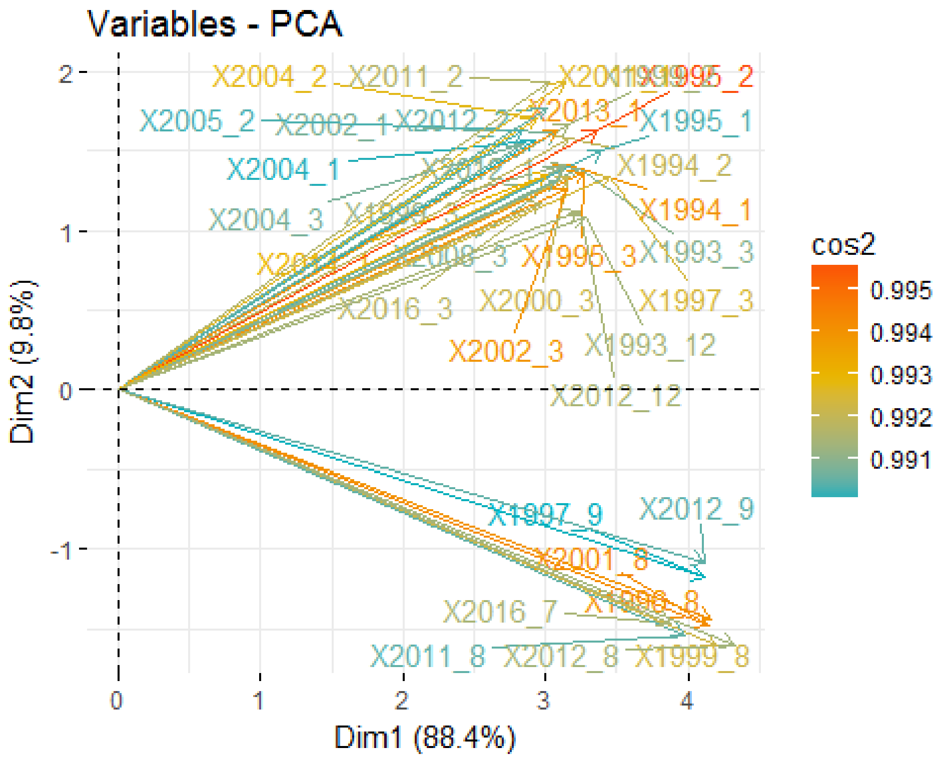
a.2 information générales sur les variables :

```
var<- get_pca_var(res.pca)
print('============ Coordonnées des dim :===========')

## [1] "============ Coordonnées des dim :==========="

head(var$coord)

##              Dim.1      Dim.2       Dim.3       Dim.4       Dim.5
## X1993_1  3.409211  1.3504778 -0.19625912 -0.05089616  0.14363062
## X1993_10 3.714602 -0.3024453 -0.19761346  0.12381590 -0.31532102
## X1993_11 3.442797  0.4294414 -0.13124092  0.02223662 -0.23502949
## X1993_12 3.278771  1.0895486 -0.08401327 -0.05595490 -0.09383399
## X1993_2  3.353832  1.4909989 -0.07421820 -0.03852766  0.17532967
## X1993_3  3.232180  1.3895130  0.07582810  0.02069417  0.09274438

print('=========== Qualité de représentation : ========')

## [1] "=========== Qualité de représentation : ========"

head(var$cos2)

##               Dim.1       Dim.2        Dim.3        Dim.4        Dim.5
## X1993_1  0.8531126 0.133866996 0.0028272115 1.901379e-04 0.0015142328
## X1993_10 0.9714321 0.006439938 0.0027492928 1.079297e-03 0.0069999314
## X1993_11 0.9648837 0.015012739 0.0014021378 4.025221e-05 0.0044967307
## X1993_12 0.8928044 0.098588751 0.0005861786 2.600222e-04 0.0007312308
## X1993_2  0.8254176 0.163134487 0.0004042146 1.089272e-04 0.0022558072
## X1993_3  0.8363835 0.154575108 0.0004603357 3.428551e-05 0.0006886357

print('========= Contributions des variables :==========')

## [1] "========= Contributions des variables :=========="

head(var$contrib)

##               Dim.1      Dim.2      Dim.3      Dim.4     Dim.5
## X1993_1  0.3291634 0.46762993 0.16671385 0.02390578 0.3058183
## X1993_10 0.3907764 0.02345423 0.16902271 0.14147694 1.4739247
## X1993_11 0.3356809 0.04728628 0.07455050 0.00456321 0.8188685
## X1993_12 0.3044570 0.30438315 0.03054976 0.02889409 0.1305238
## X1993_2  0.3185564 0.57000942 0.02384146 0.01369866 0.4557013
## X1993_3  0.2958659 0.49505405 0.02488699 0.00395211 0.1275101
```

a.3 : Les cercles de corrélations :

ci-dessous on fournit un cercle avec un gradient de couleurs.

```
fviz_pca_var(res.pca,
col.var="cos2",gradient.cols=c("#00AFBB","#E7B800","#FC4E07"),repel=T,select.
var = list(cos2=0.99))
```



->on représente les variables par leurs corrélation sur un cercle : -> Les variables positivement corréles sont regroupé par quart de cercle -> Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du cercle(quadrant opposés) -> La distance entre les variables et l'origine mesure la qualité de représentation des variables. -> Et donc les variables les plus loin de l'origine sont les bien représentés par l'ACP

Rq:

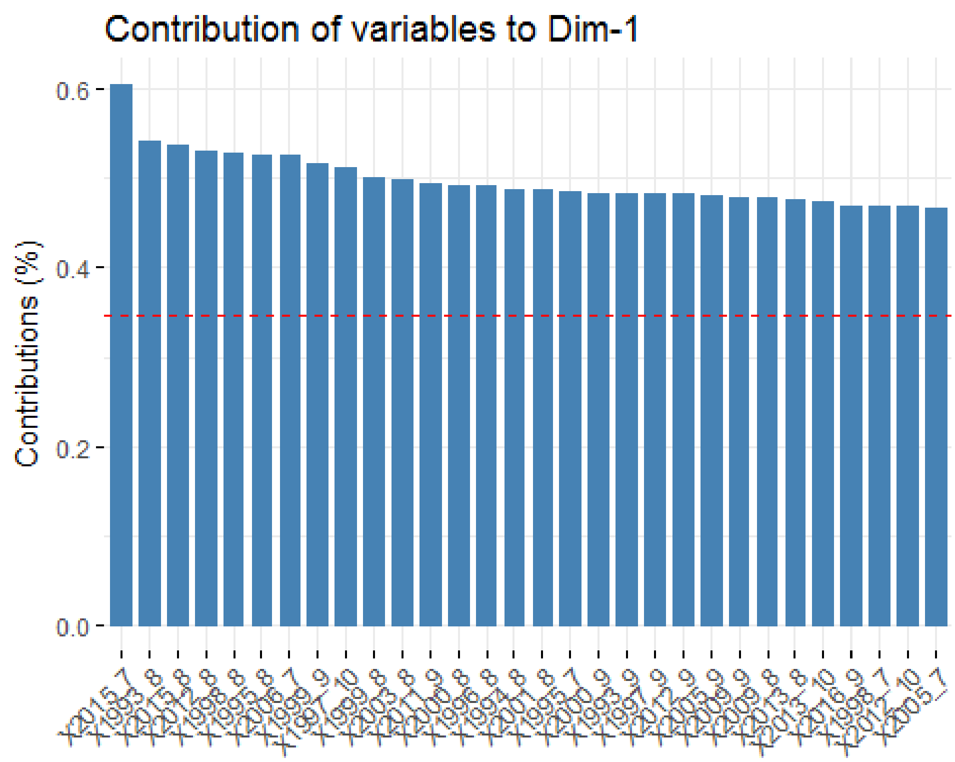
- Comme représenté sur la figure ci-dessus, les mois les plus bien représenté sont généralement les mois : 2,3 et 8.

- Les mois froids sont en bas à droite - Les mois chauds sont en hauts à droite.

a.4 : Contributions des variables aux axes principaux

Pour la dim.1 :

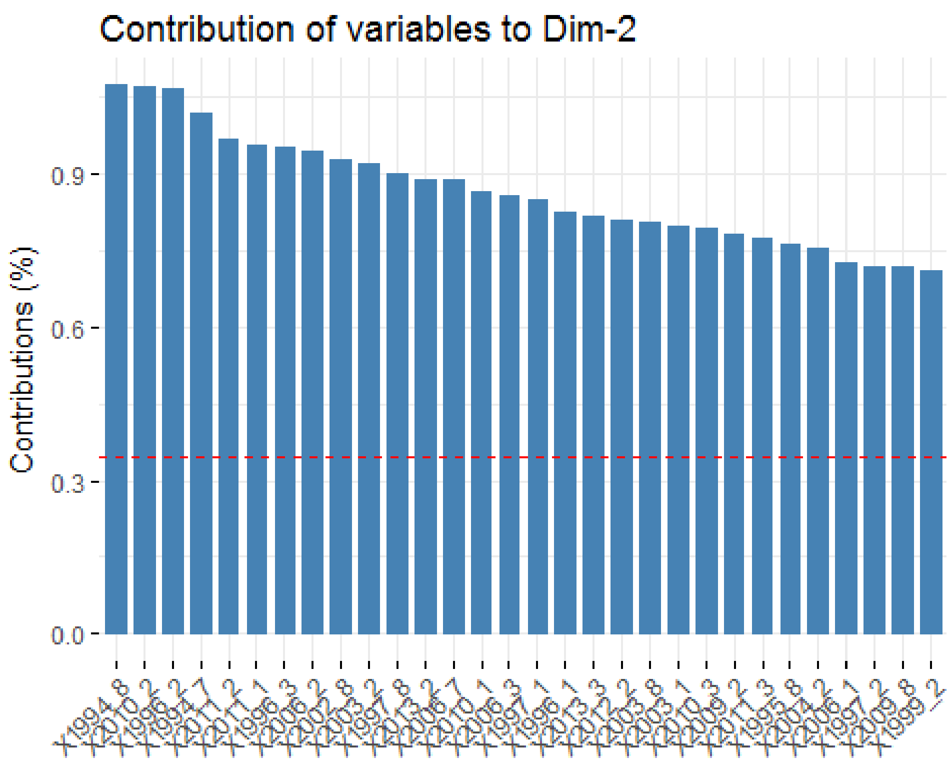
```
fviz_contrib(res.pca,choice="var", axes=1,top=30)
```



On voir que les mois 7 et 8 contribuent le plus à la dimension 1.

Pour la dim.2 :

```
fviz_contrib(res.pca,choice="var", axes=2,top=30)
```

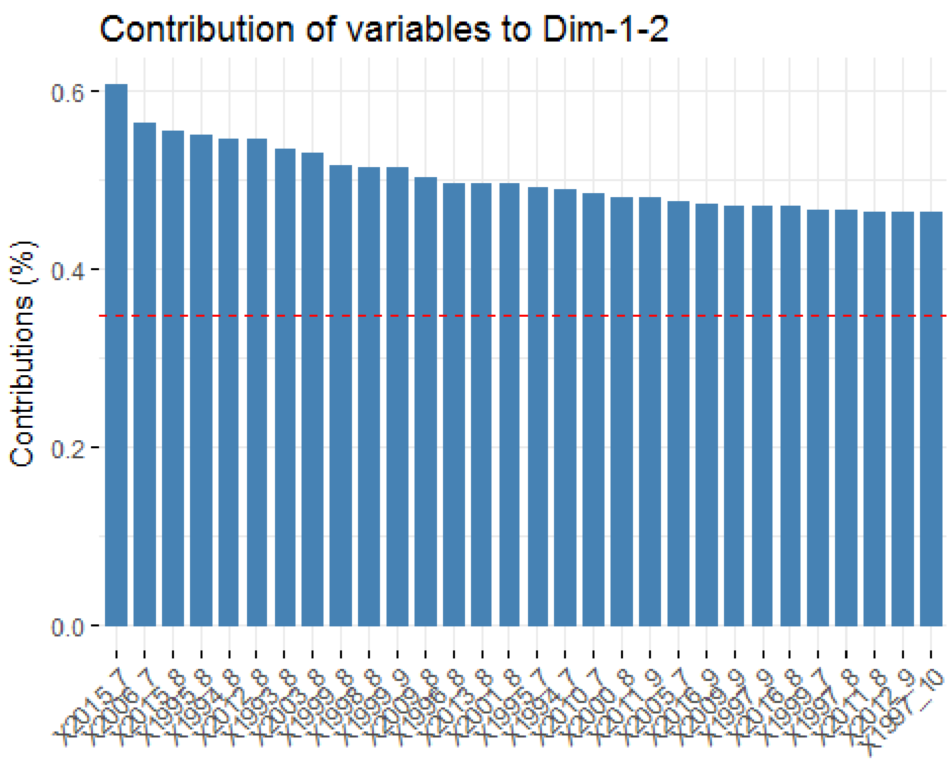


On voir que 8,2 et 7 sont les mois les plus contributifs à la dimension 2.

RQ : La ligne en pointillé rouge indique la contribution en moyenne attendue. si une variable dépasse ce seuil ==> elle est importante pour contribuer à la composante.

Notez que la contribution totale à PC1 et PC2 peut être obtenu avec le code R suivant:

```
fviz_contrib(res.pca,choice="var", axes=1:2,top=30)
```



Généralement, les mois les plus contributifs sont les mois de l'hiver. ###b. Etude des individus :

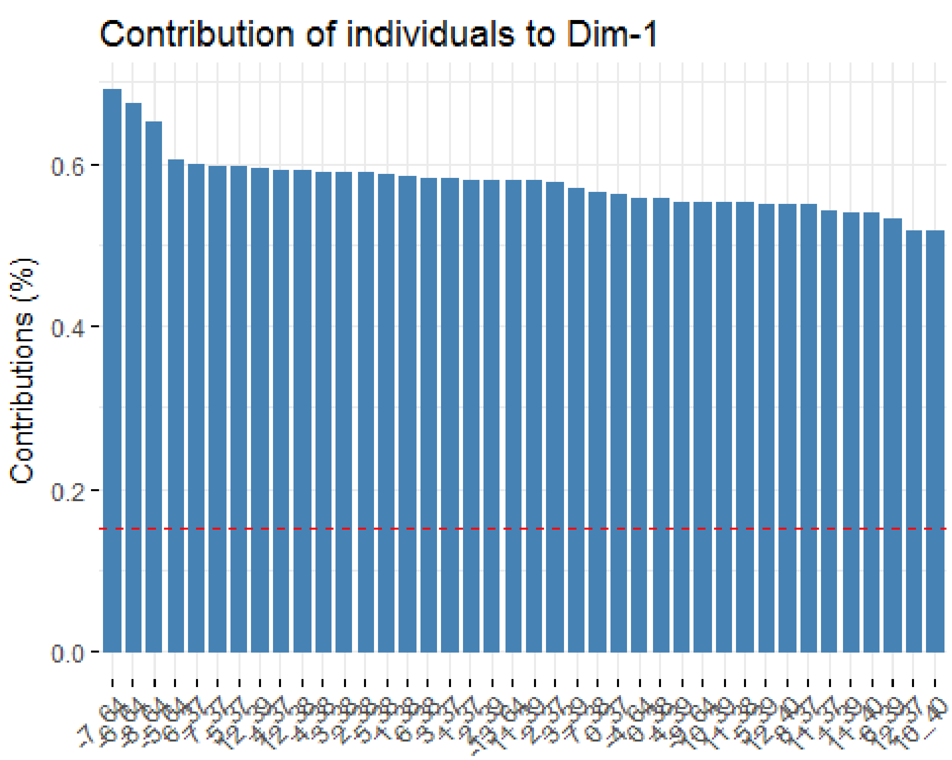
```
ind<-get_pca_ind(res.pca)
print('========= Coordonnées des individus :===========')

## [1] "========= Coordonnées des individus :==========="

head(ind$coord)

##            Dim.1    Dim.2       Dim.3     Dim.4     Dim.5
## -10_37 86.41839 22.24163 -0.55855108 2.2227634 -2.278238
## -10_38 81.81027 21.78118 -0.33371766 1.9373061 -1.804081
## -10_39 74.60298 19.19265 -0.28009448 1.6535489 -1.466811
## -10_40 68.36049 18.06186 -0.20644183 1.7878614 -1.660740
## -10_41 60.28080 17.88297 -0.06098609 1.0923526 -1.437483
## -10_42 50.70963 17.77450  0.16908443 0.7182815 -1.192142

print('======= Qualité de représentation des individus:=======')

## [1] "======= Qualité de représentation des individus:======="

head(ind$cos2)

##             Dim.1      Dim.2        Dim.3        Dim.4        Dim.5
## -10_37 0.9283807 0.06149602 3.878283e-05 0.0006141864 0.0006452263
## -10_38 0.9252896 0.06558805 1.539644e-05 0.0005188700 0.0004499600
## -10_39 0.9295906 0.06152468 1.310354e-05 0.0004566816 0.0003593583
## -10_40 0.9256993 0.06462263 8.442187e-06 0.0006331801 0.0005463396
```

```
## -10_41 0.9093730 0.08003193 9.307767e-07 0.0002986137 0.0005171177
## -10_42 0.8797211 0.10808337 9.780733e-06 0.0001765036 0.0004862060

print('======== Contributions des individus :=======')

## [1] "======== Contributions des individus :======="

head(ind$contrib)

##             Dim.1     Dim.2        Dim.3       Dim.4      Dim.5
## -10_37 0.3224128 0.1933556 2.058420e-03 0.069504905 0.11729080
## -10_38 0.2889453 0.1854327 7.347956e-04 0.052798971 0.07354912
## -10_39 0.2402771 0.1439771 5.176274e-04 0.038464762 0.04861989
## -10_40 0.2017485 0.1275113 2.811924e-04 0.044967285 0.06232593
## -10_41 0.1568765 0.1249980 2.453972e-05 0.016786289 0.04669507
## -10_42 0.1110148 0.1234862 1.886320e-04 0.007258022 0.03211602
```

On va utiliser des graphiques pour rapidement identifier les individus qui contribuent très bien à une telle Dimension principale.

->Contribution totale des individus sur PC1 :

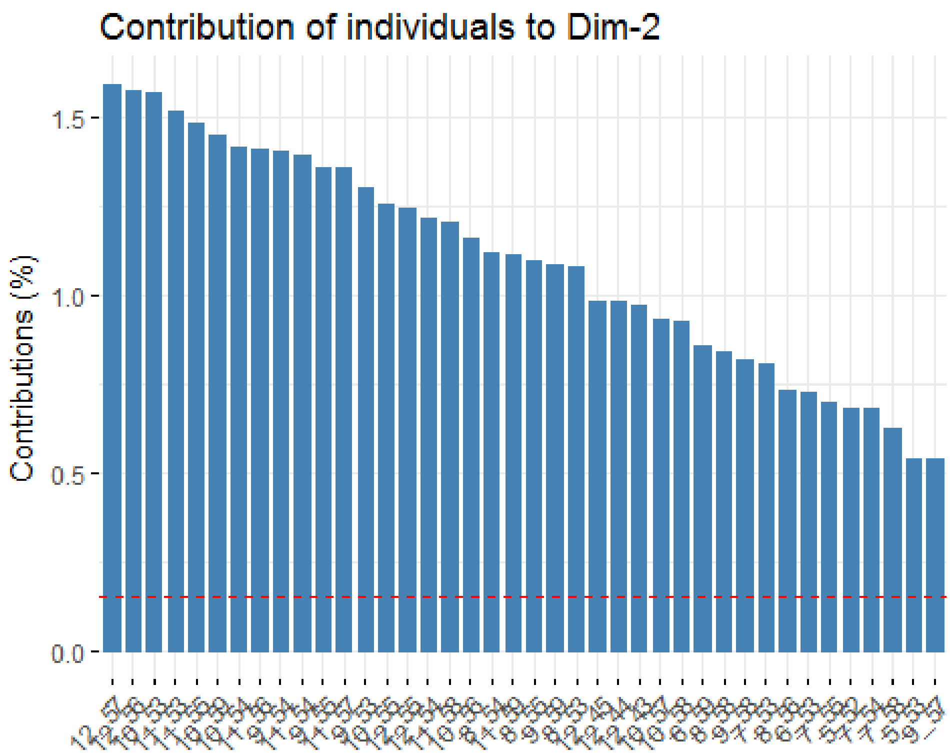
```
fviz_contrib(res.pca,choice="ind", axes=1,top = 40)
```



Donc la région qui contribue le plus à la définiton du premier axe principale est la région comprise entre : -13=<lon<=12 et 37=<lat<64.

Contribution totale des individus sur PC2 :

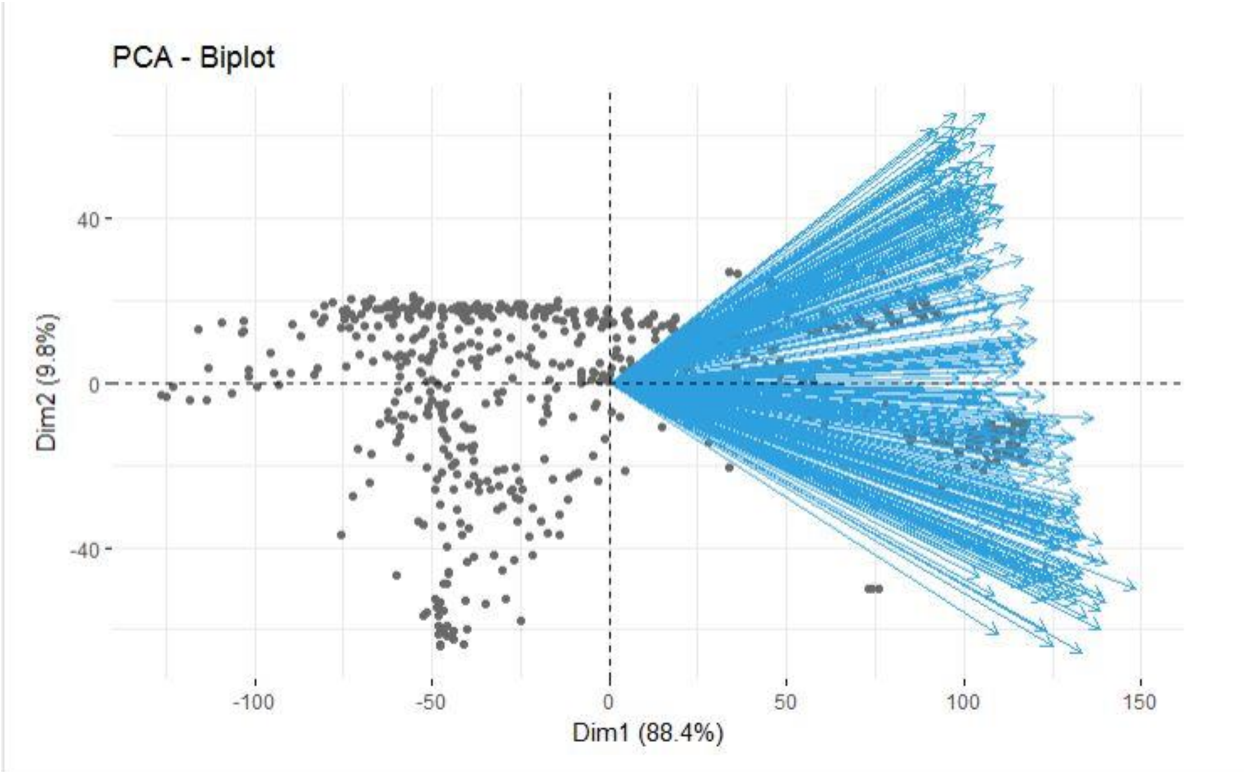
```
fviz_contrib(res.pca,choice="ind", axes=2,top=40)
```



Donc la région qui contribue le plus à la définition du premier axe principale est la région comprise entre : 5=<lon<=12 et 43=<lat<=59

Biplot :

Pour raison de lisibilité, j'ai supprimé tous les textes. Donc la figure ne contient que la représentation des variables(les flèches en bleu) ainsi que la nuage des individus(les points en noires).



Globalement un biplot peut être interprété comme suit : ->un individu qui se trouve du même côté d'une variable donnée a une valeur élevée pour cette variable ->un individu qui se trouve sur le côté opposé d'une variable donnée a une faible valeur pour cette variable.

NB: il faut se méfier des individus proches de l'origine : mal représentés, ou proches de la moyenne car ils sont mal représentés.
Commentaire pour notre jeu de données :

- Pour le quart du cercle en haut à droite : ces couple lon/lat (points noires) ont généralement de grandes valeurs de sst pendant ces mois qui sont sur le même quart.
- Pour le quart du cercle en bas à gauche : ces couple lon/lat ont généralement des petites valeurs de sst pendant les mêmes mois que le cas précédent.
- Pour le quart du cercle en bas à droite : ces couple lon/lat (points noires) ont généralement de grandes valeurs de sst pendant ces mois qui sont sur le même quart.
- Pour le quart du cercle en haut à gauche : ces couple lon/lat ont généralement des petites valeurs de sst pendant les mêmes mois que le cas précédent.

D'après le graphiques des variables toutes seules qu'on a présenté dans le section "Les cercles de corrélations", on peut fournir les conclusions suivantes: - Le nuage de points en haut à droite sont des zones chaudes pendant les mois froids. - Le nuage de points en bas à gauches sont des zones froides pendant les mois froids. - Le nuage de points en bas à droite sont des zones chaudes pendant les mois chauds. - Le nuage de points en haut à gauche sont des zones foides pendant les mois chauds.