Projet ACP

Sami BEN BRAHIM - 2AMIndS-
Statistiques
ECOLE NATIONALE D'INGÉNIEURS DE TUNIS

May 30, 2020

On se propose d'effectuer une étude quantitative des caractéristiques de la population américaine aux années 2000-2001 via l'analyse ACP.

1/ Préparation du tableau de données:

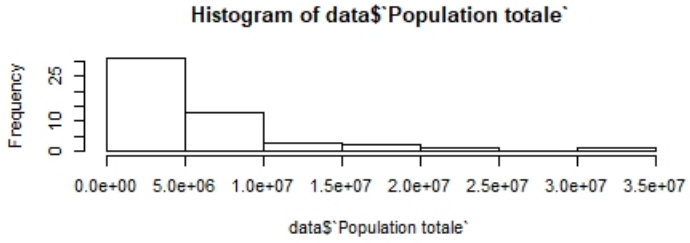
Tout d'abord, on importe les packages nécessaires et le dataset qui contient des caractéristiques démographiques des 51 états américains au cours de 2000-2001. Pour cela, on exécute les commandes suivantes :

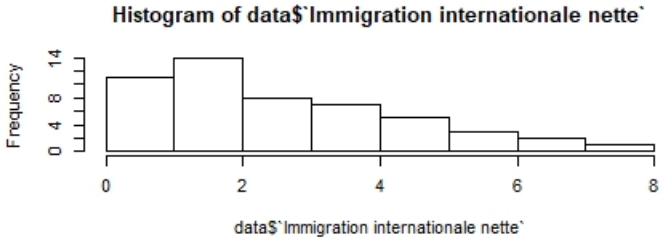
```
library(FactoMineR)
library(factoextra)
library(readxl)

demoPCA <- read_excel("demoPCA.xlsx")
View(demoPCA)
```

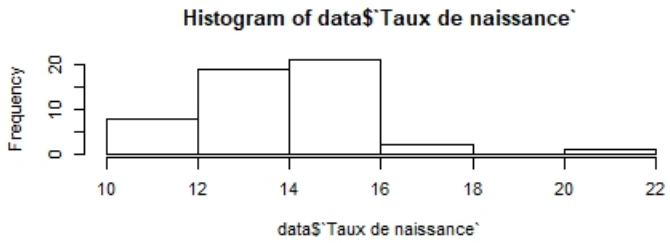
La figure ci-après montre les caractéristiques des 10 premiers états présents dans le dataset :

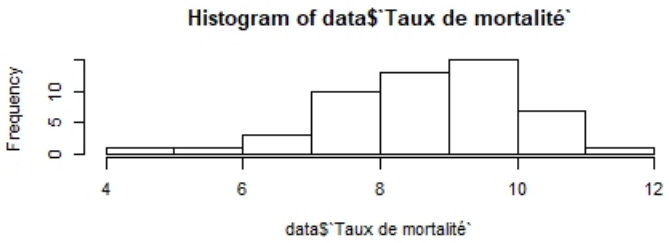
	Etat	Population totale	Immigration internationale nette	Taux de naissance	Taux de mortalité	Pop inf 65 ans	Pop sup 65 ans
1	Alabama	4464356	0.69	14.41	10.28	869.21	130.79
2	Alaska	634892	2.09	15.95	4.64	941.95	58.05
3	Arizona	5307331	4.29	15.88	7.77	869.54	130.46
4	Arkansas	2692090	1.07	14.35	10.51	861.06	138.94
5	California	34501130	7.88	15.37	6.72	894.03	105.97
6	Colorado	4417714	3.57	14.57	6.26	903.52	96.48
7	Connecticut	3425074	3.50	12.52	9.00	862.64	137.36
8	Delaware	796165	2.12	14.01	8.79	869.45	130.55
9	District of Columbia	571822	5.73	14.33	10.76	880.75	119.25
10	Florida	16396515	5.76	12.54	10.13	826.28	173.72

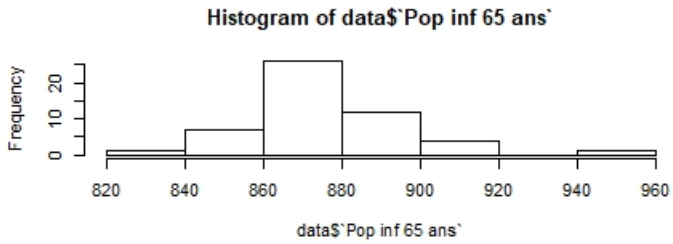
2/Description des différentes variables du dataset:

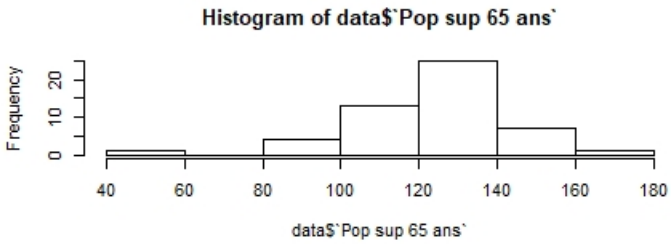












Analyse des histogrammes:

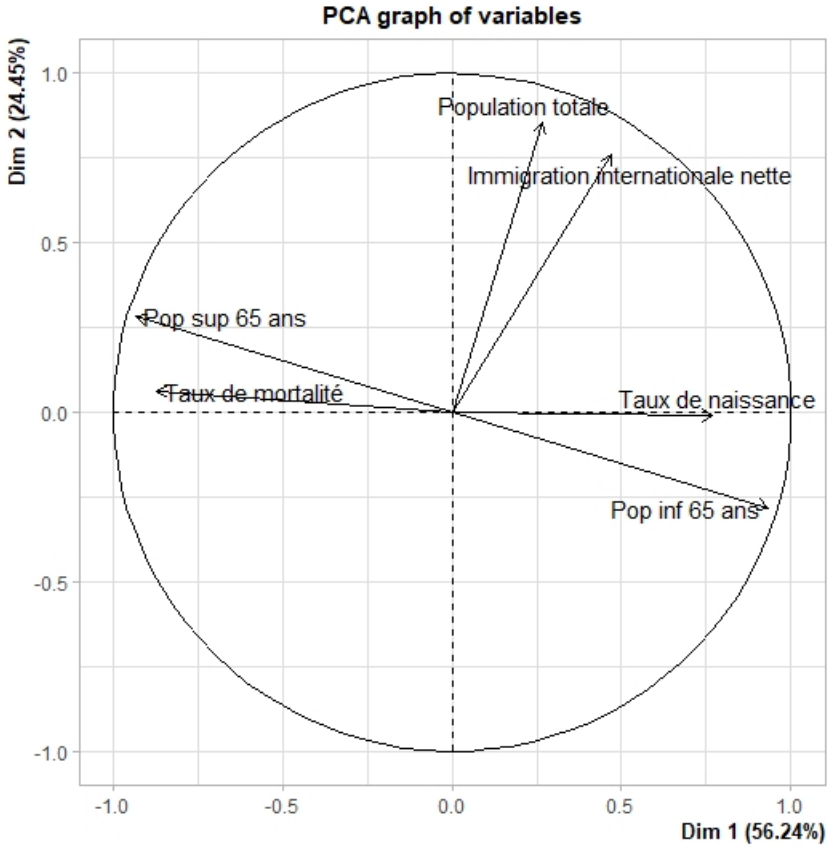
- Environ 70% des états comptent une population inférieure ou égale à 10 millions.
- 50% des états possèdent un taux d'immigration internationale nette dans l'intervalle $[0,2]$.
- Environ 75% des états ont un taux de naissance dans l'intervalle $[0.12,0.16]$ et un taux de mortalité dans l'intervalle $[0.07,0.1]$.
- Sur 1000 individus, environ 70% des états possèdent un total d'individus âgés moins de 65 ans dans l'intervalle $[860,900]$.

On applique la fonction PCA sur le jeu de données "demoPCA" après avoir supprimé la colonne des noms des états. (la fonction PCA est applicable sur un jeu de données numériques). On normalise aussi les données.

3/Contrôle de la linéarité entre les variables:

```
#suppression de la colonne de type "chr"
data=demoPCA[,2:ncol(demoPCA)]

res.pca=PCA(data,scale.unit = TRUE,graph = TRUE,ncp=5)
res.pca$var$cor
```

Ayant fixé le paramètre *graph* à TRUE, on obtient un graphe qui représente la corrélation des variables avec les deux premières dimensions.



De ce graphe, on peut déduire la corrélation linéaire des variables du dataset. D'ailleurs, on observe des variables :

- positivement corrélés : elles sont regroupées.
  Exp: *Population totale* et *Immigration internationale nette*, *Taux de naissance* et *Pop inf 65 ans*, *Pop sup 65 ans* et *Taux de mortalité*.

- négativement corrélées : elles sont situées dans les côtés opposés du graphe.
  Exp: *Pop inf 65 ans* et *Pop sup 65 ans*, *Taux de naissance* et *Taux de mortalité*.

On examine aussi la corrélation entre les variables et les axes factoriels. Pour cela, on applique *varcor* sur la variable *data*, comme le montre la figure ci-dessous.

```
> res.pca$var$cor
                                Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
Population totale             0.2672247  0.854269543  0.18360012 -0.39691831 -0.08695621
Immigration internationale nette 0.4714785  0.758154169 -0.24908801  0.32780892  0.18277527
Taux de naissance             0.7674603 -0.008975629  0.56335576  0.29350131 -0.08608973
Taux de mortalité            -0.8696145  0.063203914  0.36533032 -0.05295269  0.32172293
Pop inf 65 ans                0.9315571 -0.281187147 -0.02483984 -0.17164139  0.15184664
Pop sup 65 ans               -0.9315571  0.281187147  0.02483984  0.17164139 -0.15184664
```

On remarque d'après le tableau de mesure des corrélations qu'il existe une forte corrélation entre :

- la dimension 1 et les 4 quatre variables : taux de naissance, taux de mortalité, pop inf 65 ans et pop sup 65 ans.
- la dimension 2 et les 2 variables: population totale et immigration internationale nette.

4/Etude du tableau de valeurs propres:

```
> res.pca$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 3.374523e+00           5.624205e+01                          56.24205
comp 2 1.466782e+00           2.444637e+01                          80.68841
comp 3 5.478238e-01           9.130397e+00                          89.81881
comp 4 4.128714e-01           6.881189e+00                          96.70000
comp 5 1.980001e-01           3.300001e+00                         100.00000
comp 6 3.080017e-28           5.133362e-27                         100.00000
```

L'observation de la colonne de variance cumulée du tableau de valeurs propres montre que les deux premières composantes expliquent plus que 80% de la variance. C'est un pourcentage acceptable. Donc le plan de projection d'individus sera formé par les deux premières composantes principales CP.

5/Interprétation des composantes principales:
Pour bien interpréter les CP, on va chercher *contrib* et *cos2*.
En effet, le cos2 d'une variable représente :

- sa qualité de représentation par une CP.
- sa capacité à interpréter une CP.

contrib mesure la contribution d'une variable à une CP.
En premier lieu, on applique la fonction cos2 et on obtient du coup le tableau suivant :

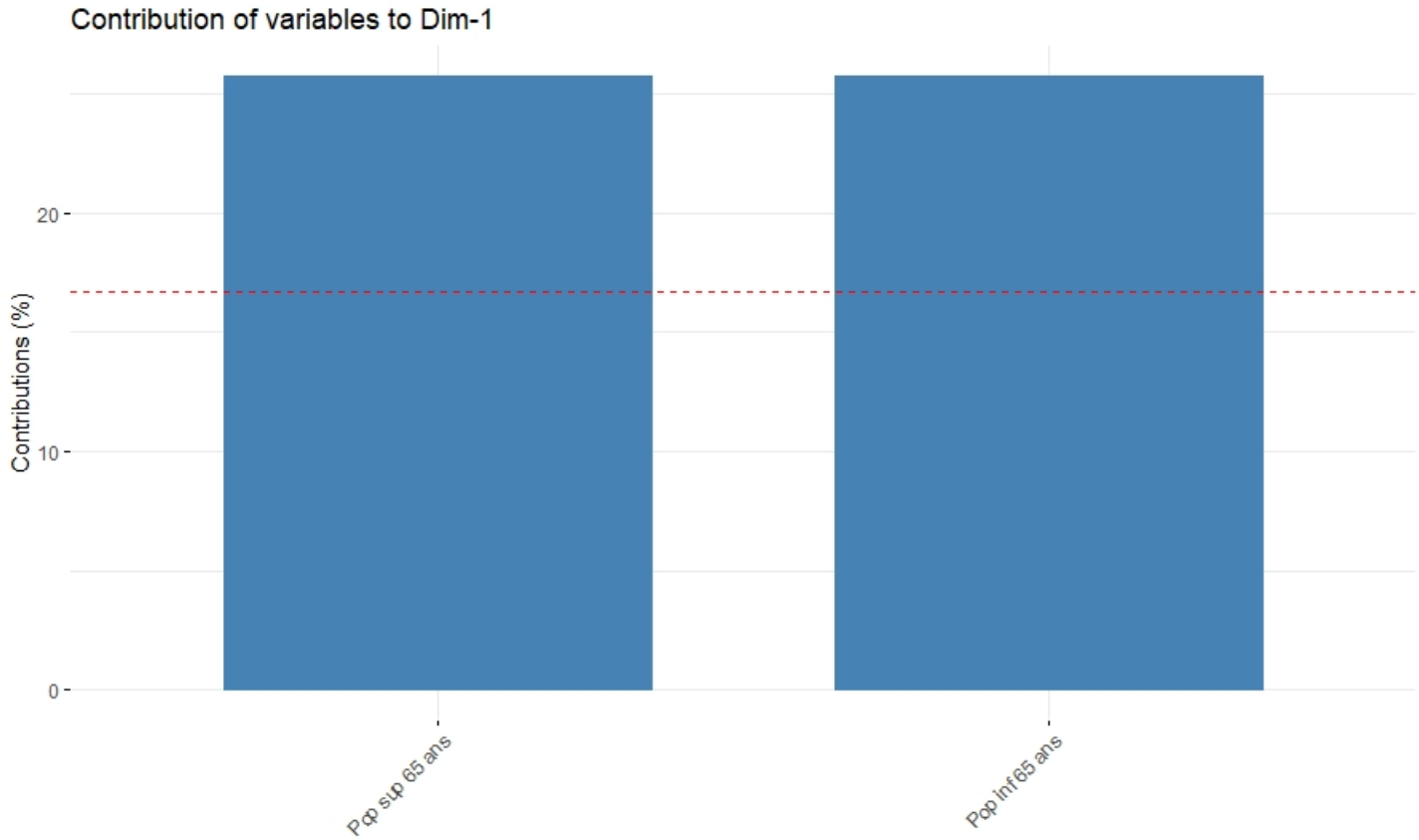
```
> res.pca$var$cos2
                                Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
Population totale             0.07140902 7.297765e-01 0.0337090042 0.157544146 0.007561383
Immigration internationale nette 0.22229193 5.747977e-01 0.0620448373 0.107458686 0.033406799
Taux de naissance             0.58899527 8.056192e-05 0.3173697120 0.086143017 0.007411442
Taux de mortalité             0.75622939 3.994735e-03 0.1334662425 0.002803987 0.103505643
Pop inf 65 ans                0.86779860 7.906621e-02 0.0006170178 0.029460766 0.023057401
Pop sup 65 ans                0.86779860 7.906621e-02 0.0006170178 0.029460766 0.023057401
```

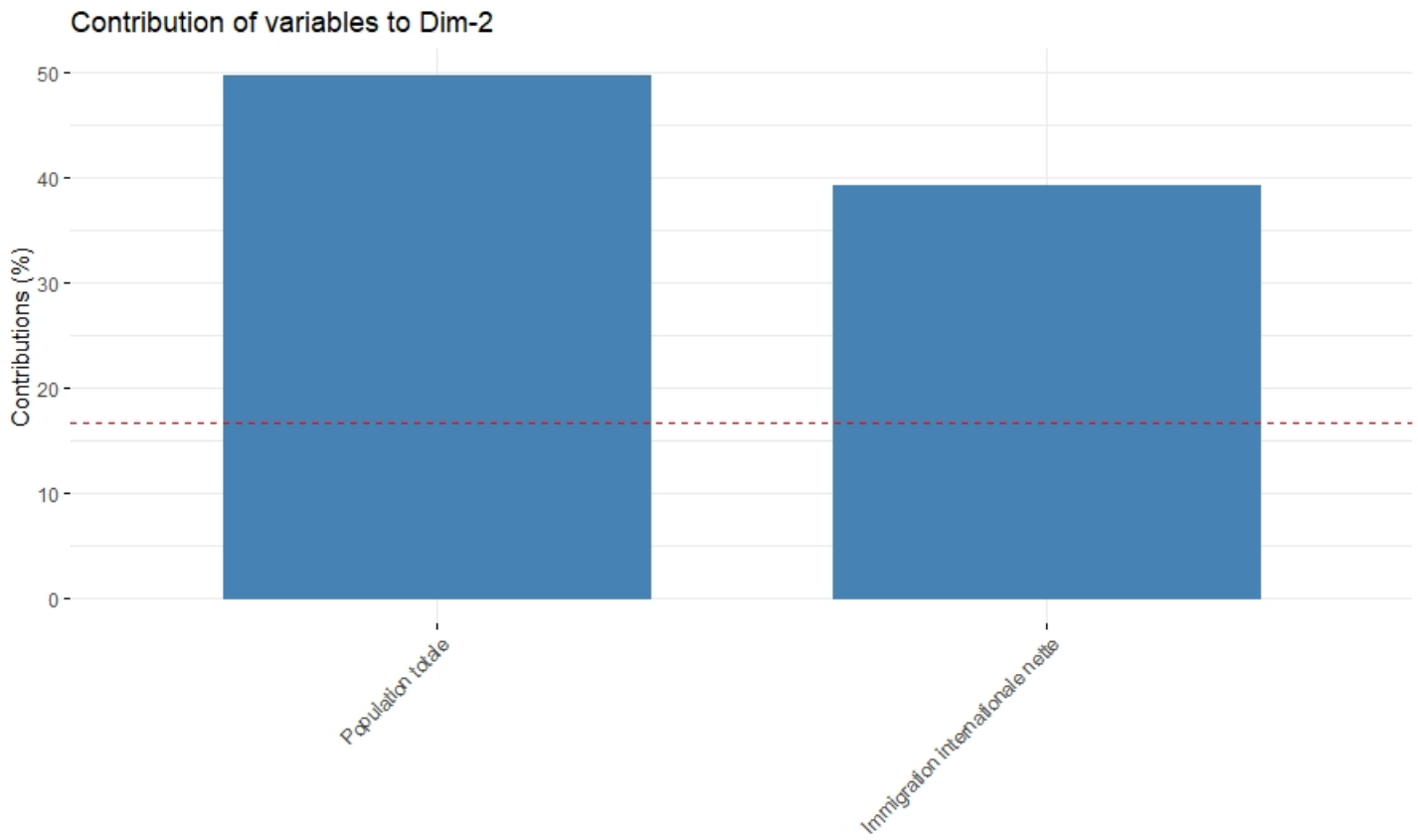
On remarque que :

- les deux variables *pop inf 65 ans* et *pop sup 65 ans* possèdent le cos2 maximal, donc elles interprètent le plus la première CP.
- la variable *pop totale* donne la meilleure interprétation de la deuxième CP.

En deuxième lieu, on applique la fonction *fviz_contrib* du package *factoextra* pour observer les 2 variables les plus contributives aux deux premières CP.

```
# Contributions des variables à PC1
fviz_contrib(res.pca, choice = "var", axes = 1, top = 2)
# Contributions des variables à PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 2)
```

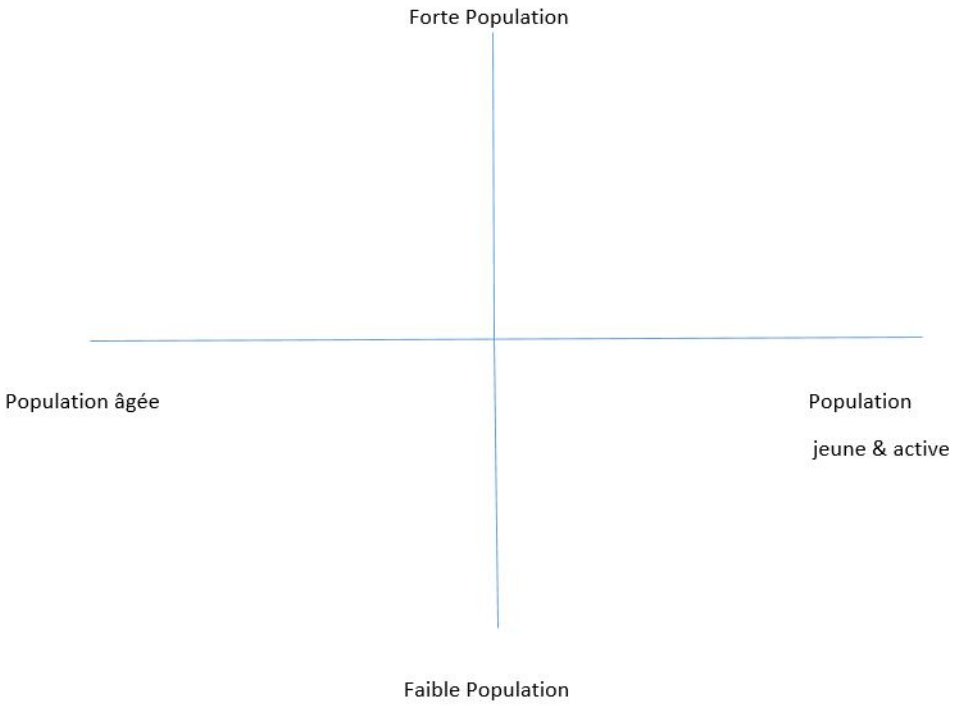




Donc, les deux variables *pop inf 65 ans* et *pop sup* sont les plus contributives à la première CP. Il faut signaler aussi que la première variable factorielle augmente quand le taux de naissance augmente(leur corrélation est proche de 1) et elle diminue quand le taux de mortalité augmente(leur corrélation est proche de -1).
Et la variable *population totale* contribue le plus à la deuxième CP.

Conclusion :

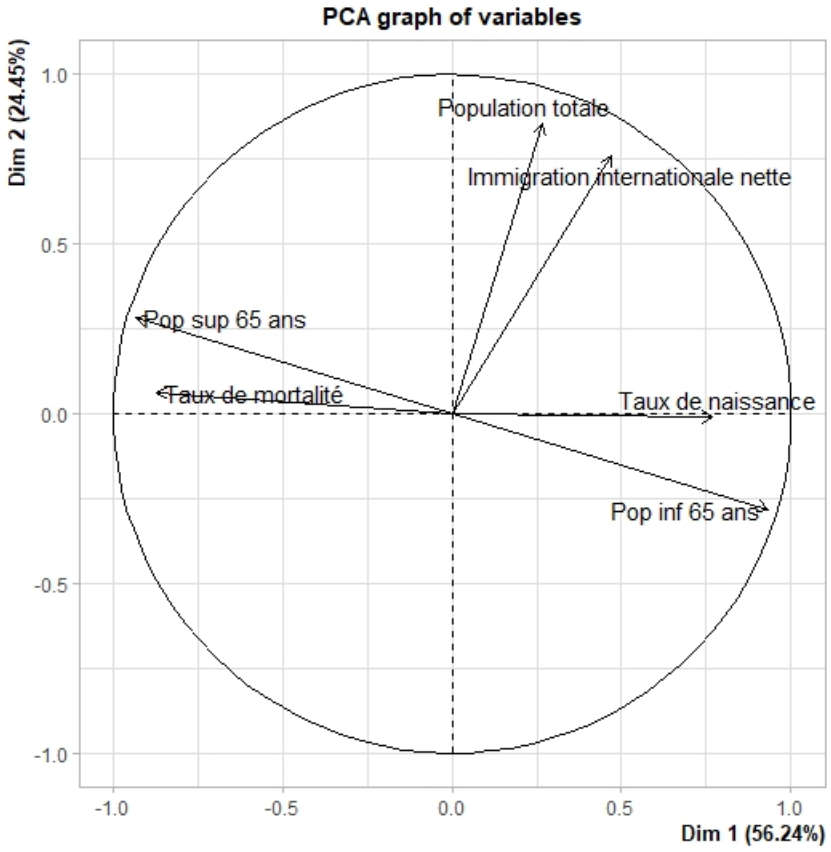
- La première variable factorielle correspond à l'âge de la population.
- La deuxième variable factorielle représente la taille de la population.



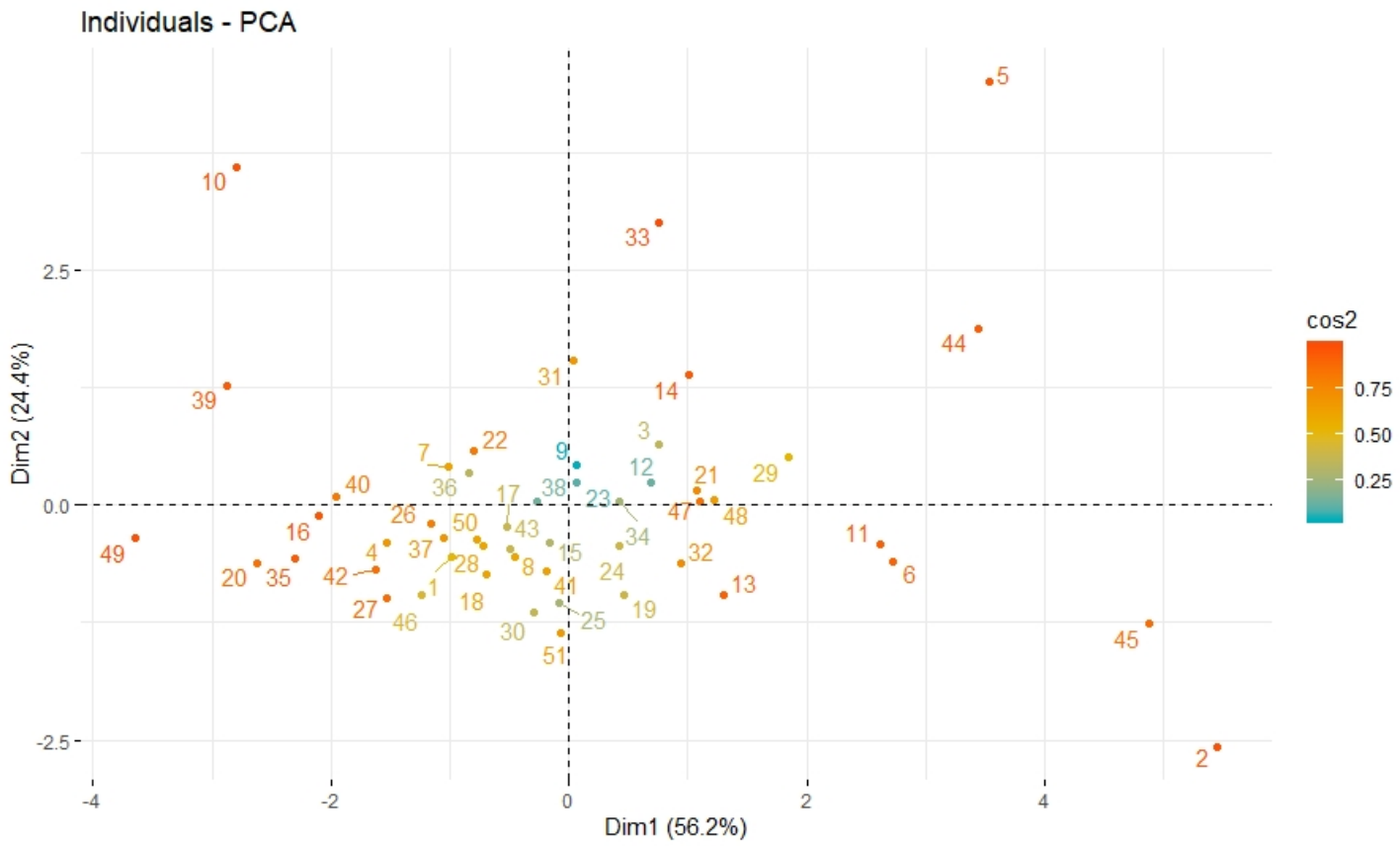
6/Plan principal/Plan des individus/plan des variables:

Le plan principal est formé par les deux composantes principales.

Les variables sont représentées sur le plan principal par leur corrélation comme on l'observe dans la figure ci-après.

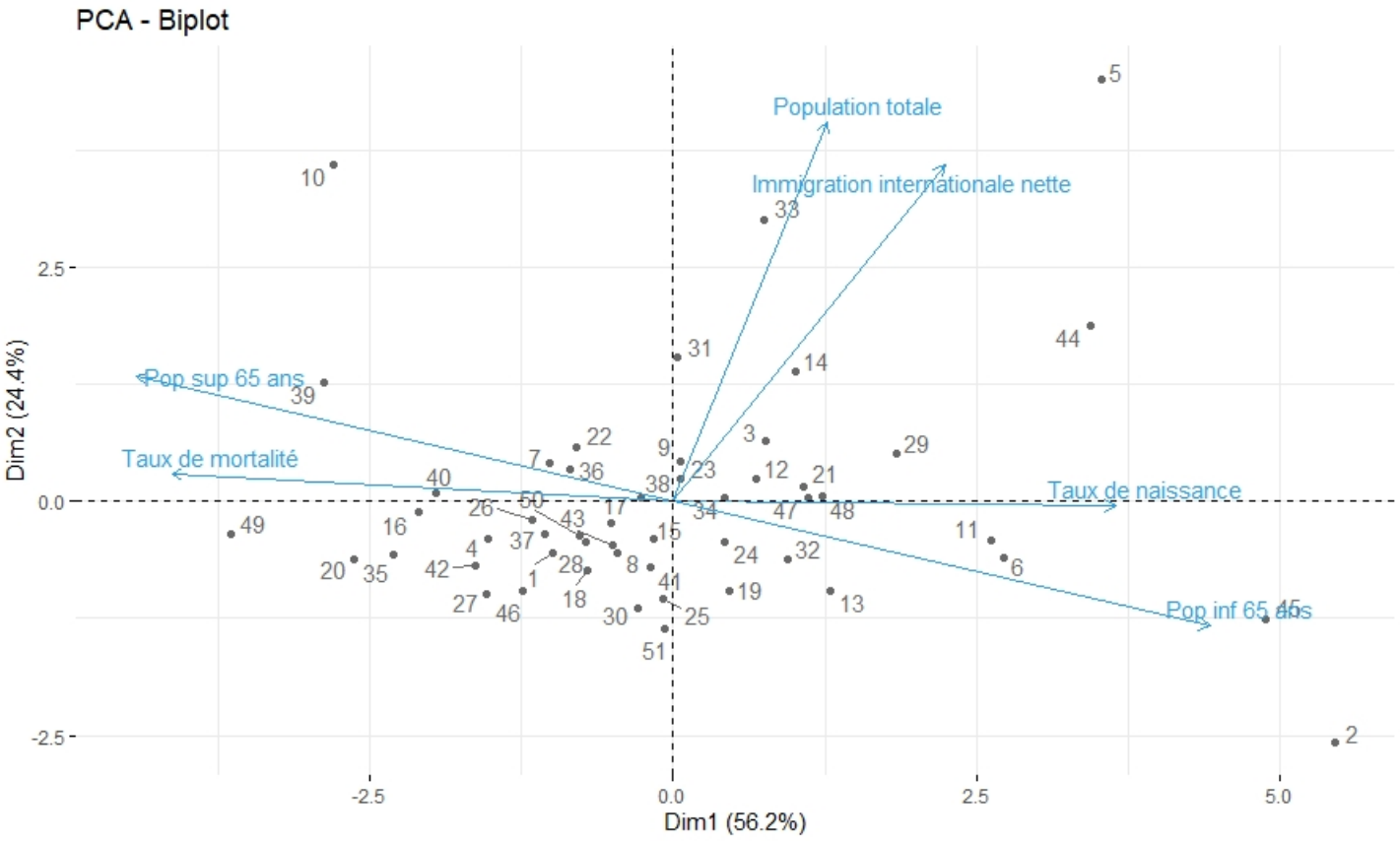


Par contre, les individus sont représentés par leur projection sur le plan principal. Le graph suivant illustre la projection des individus sur le plan principal.

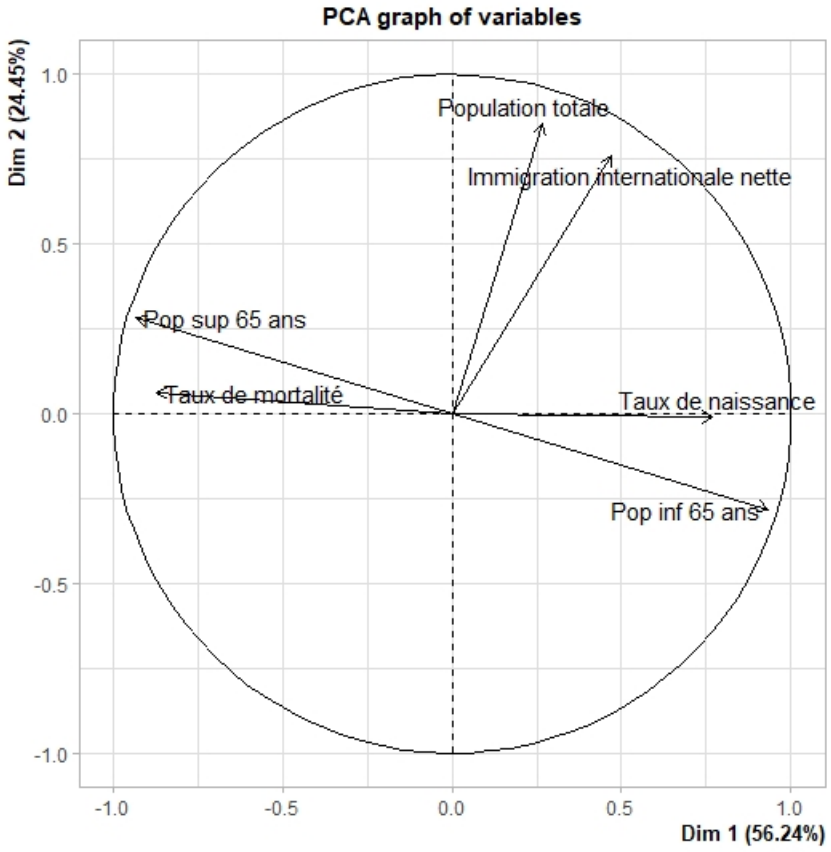


On peut aussi faire appel à la fonction *fviz_pca_biplot* pour représenter les individus et les variables sur le même graphe(voir la figure ci-dessous). Un tel graphe peut être interprété de la manière suivante :

- 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.



7-a/Graphique du plan de variables:



7-b et 7-c/ Corrélation des variables

*Corrélation entre les variables initiales:

L'observation du graphe des variables et de la matrice de corrélation(voir la matrice ci-dessous) montre que les 4 couples de variables suivants : (Pop inf 65 ans,taux de naissance), (Pop sup 65 ans,taux de mortalité), (pop inf 65 ans, pop sup 65 ans) et (Pop totale, immigration internationale nette) représentent les couples des variables les plus corrélées entre elles.

La matrice de corrélation est une matrice symétrique. Donc les signes (+) et (-) s'organisent symétriquement par rapport à la diagonale.

```
> mcor=cor(data)
> mcor
                            Population totale Immigration internationale nette Taux de naissance
Population totale                  1.00000000                       0.5819193         0.1918389
Immigration internationale nette   0.58191929                       1.0000000         0.2951882
Taux de naissance                  0.19183888                       0.2951882         1.0000000
Taux de mortalité                 -0.11827248                      -0.4116410        -0.5053897
Pop inf 65 ans                     0.05908841                       0.2037014         0.6400138
Pop sup 65 ans                    -0.05908841                      -0.2037014        -0.6400138
                            Taux de mortalité Pop inf 65 ans Pop sup 65 ans
Population totale                 -0.1182725     0.05908841    -0.05908841
Immigration internationale nette  -0.4116410     0.20370143    -0.20370143
Taux de naissance                 -0.5053897     0.64001381    -0.64001381
Taux de mortalité                  1.0000000    -0.77900102     0.77900102
Pop inf 65 ans                    -0.7790010     1.00000000    -1.00000000
Pop sup 65 ans                     0.7790010    -1.00000000     1.00000000
```

*Corrélation entre les variables et les CP:

D'après le plan des variables et le tableau de corrélation(voir le tableau ci-dessous), *Pop inf 65 ans* et *Pop sup 65 ans* sont les variables les plus corrélées avec la première CP. Alors que *Population totale* est la variable la plus corrélée avec la deuxième CP.

Les signes (+) et (-) s'organisent dans le tableau de coefficients selon le positionnement de la variable par rapport à l'axe principal :

- Les variables positivement corrélées sont regroupées.
- Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés).

```
> res.pca$var$cor
                               Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
Population totale          0.2672247  0.854269543  0.18360012 -0.39691831 -0.08695621
Immigration internationale nette  0.4714785  0.758154169 -0.24908801  0.32780892  0.18277527
Taux de naissance          0.7674603 -0.008975629  0.56335576  0.29350131 -0.08608973
Taux de mortalité         -0.8696145  0.063203914  0.36533032 -0.05295269  0.32172293
Pop inf 65 ans             0.9315571 -0.281187147 -0.02483984 -0.17164139  0.15184664
Pop sup 65 ans            -0.9315571  0.281187147  0.02483984  0.17164139 -0.15184664
```

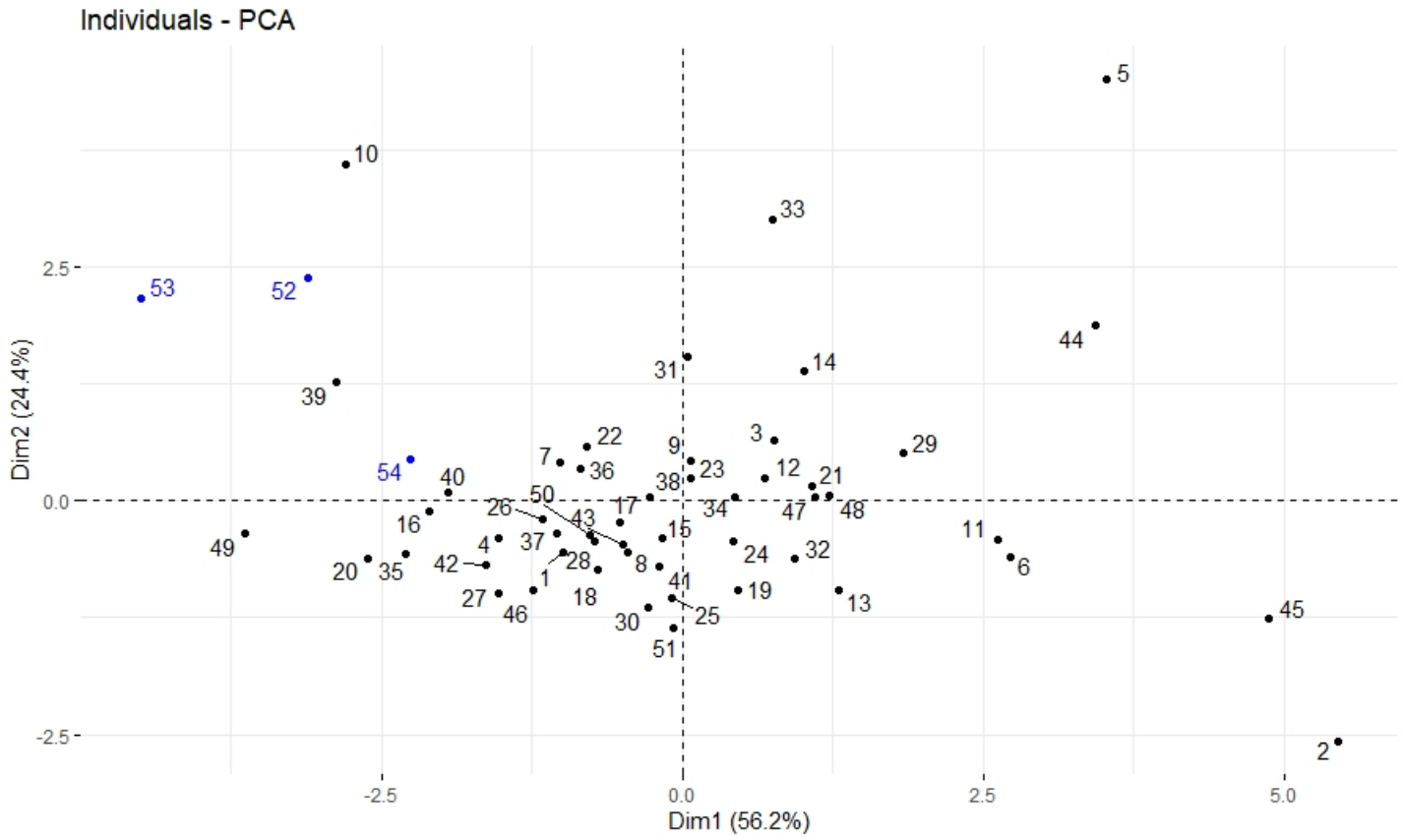
8/Ajout de variables supplémentaires:
On ajoute des données relatives à trois villes canadiennes.

52	Ontario	11 680 000	2.78	13.78	7.80	722.30	102.31
53	Toronto	2 481 000	3.69	15.30	9.70	692.10	108.60
54	Colombie britannique	4 039 000	2.10	11.20	8.30	788.32	98.75

9/Représentation des individus actifs et supplémentaires:
On recourt à la commande *PCA* pour appliquer une analyse en composantes principales au nouvel échantillon en indiquant par le paramètre *ind.sup* les indices des données supplémentaires(voir le code ci-dessous).

```
res.pca <- PCA(data, ind.sup = 52:54, graph=TRUE)
p <- fviz_pca_ind(res.pca, col.ind.sup = "blue", repel = TRUE)
p
```

La fonction *fviz_pca_ind* permet de représenter la projection des individus sur le plan principal. On marque dans le graphe ci-après les données supplémentaires en bleu pour mieux les distinguer.



Il faut intéressant de signaler que les individus supplémentaires ne sont pas utilisés pour la détermination des composantes principales. Leurs coordonnées sont prédites en utilisant uniquement les informations fournies par l'analyse en composantes principales effectuée sur les variables/individus actifs.