ANALYSE DE DONNEES

DESCHAUX Pierre-Honoré

NARDI Léo

Dans le cadre de ce projet nous allons travailler sur la base de données “longley”

1-Principes de l’ACP

STATISTIQUES DESCRIPTIVES

Cette base contient 6 variables économiques observées de 1947 à 1962 :

le RNB (GNP)
son déflateur (GNP.deflator)
le nombre de chômeurs (Unemployed)
le nombre de personnes engagées dans l’armée (Armerd.Forces)
la population civile hors établissement institutionnel (Population)
le nombre de personnes en emploi (Employed)
et l’année (Year)

```
> summary(longley)
  GNP.deflator          GNP          Unemployed     Armed.Forces
 Min.   : 83.00   Min.   :234.3   Min.   :187.0   Min.   :145.6
 1st Qu.: 94.53   1st Qu.:317.9   1st Qu.:234.8   1st Qu.:229.8
 Median :100.60   Median :381.4   Median :314.4   Median :271.8
 Mean   :101.68   Mean   :387.7   Mean   :319.3   Mean   :260.7
 3rd Qu.:111.25   3rd Qu.:454.1   3rd Qu.:384.2   3rd Qu.:306.1
 Max.   :116.90   Max.   :554.9   Max.   :480.6   Max.   :359.4
   Population         Year         Employed
 Min.   :107.6   Min.   :1947   Min.   :60.17
 1st Qu.:111.8   1st Qu.:1951   1st Qu.:62.71
 Median :116.8   Median :1954   Median :65.50
 Mean   :117.4   Mean   :1954   Mean   :65.32
 3rd Qu.:122.3   3rd Qu.:1958   3rd Qu.:68.29
 Max.   :130.1   Max.   :1962   Max.   :70.55
```

Il n’est pas pertinent ici d'étudier les statistiques descriptives de Year.

Population et Employed n’ont pas une forte volatilité, car la différence entre MAX et MIN n’est pas si élevée que ça comparé aux autres variables.
Il est intéressant de prendre en compte le déflateur du GNP puisque son écart type est plus faible que le GNP.

Armed Force a une moyenne nettement inférieure à la médiane suggérant que les données ne sont pas symétriques et donc qu’il y a une queue à gauche.

Avec la règle des 1.5 fois l’écart interquartile (IQR), on peut détecter les valeurs aberrantes. Selon cette règle, une valeur est considérée comme une valeur aberrante si elle est inférieure à Q1 - 1.5 * IQR ou supérieure à Q3 + 1.5 * IQR.

Par exemple, si nous prenons la variable GNP.deflator, le premier quartile (Q1) est de 94.53 et le troisième quartile (Q3) est de 111.25. L'écart interquartile (IQR) est donc de 111.25 - 94.53 = 16.72. Selon la règle des 1.5 fois l'IQR, toute valeur inférieure à 94.53 - 1.5 * 16.72 = 69.43 ou supérieure à 111.25 + 1.5 * 16.72 = 136.35 serait considérée comme une valeur aberrante pour cette variable. Il n'y a donc pas de valeurs aberrantes.
De la même manière, il n'y a pas de valeur aberrante dans les autres variables.

Ces données sont dans des unités différentes, il faut donc les centrer et réduire.

MATRICE VARIANCE COVARIANCE / CORRELATION

```
> round(var(longley_cr),3)
             GNP.deflator   GNP Unemployed Armed.Forces Population  Year Employed
GNP.deflator        1.000 0.992      0.621        0.465      0.979 0.991    0.971
GNP                 0.992 1.000      0.604        0.446      0.991 0.995    0.984
Unemployed          0.621 0.604      1.000       -0.177      0.687 0.668    0.502
Armed.Forces        0.465 0.446     -0.177        1.000      0.364 0.417    0.457
Population          0.979 0.991      0.687        0.364      1.000 0.994    0.960
Year                0.991 0.995      0.668        0.417      0.994 1.000    0.971
Employed            0.971 0.984      0.502        0.457      0.960 0.971    1.000
```

Les deux matrices sont identiques car les données sont centrées réduites.

La forte corrélation positive entre GNP et GNP.deflator peut s'expliquer par le fait que le déflateur du RNB (GNP.deflator) mesure les changements de prix des biens et services produits dans une économie. Lorsque les prix augmentent (inflation), le RNB nominal augmente également, même si la production réelle de biens et services reste constante.

La corrélation positive entre GNP et Population peut s'expliquer par le fait que la population est un facteur clé de la croissance économique. Une population plus importante signifie plus de travailleurs pour produire des biens et services et plus de consommateurs pour les acheter.

La corrélation positive entre GNP et Employed peut s'expliquer par le fait que l'emploi est un facteur clé de la production économique. Lorsque plus de personnes sont employées, la production de biens et services augmente, ce qui entraîne une augmentation du RNB.

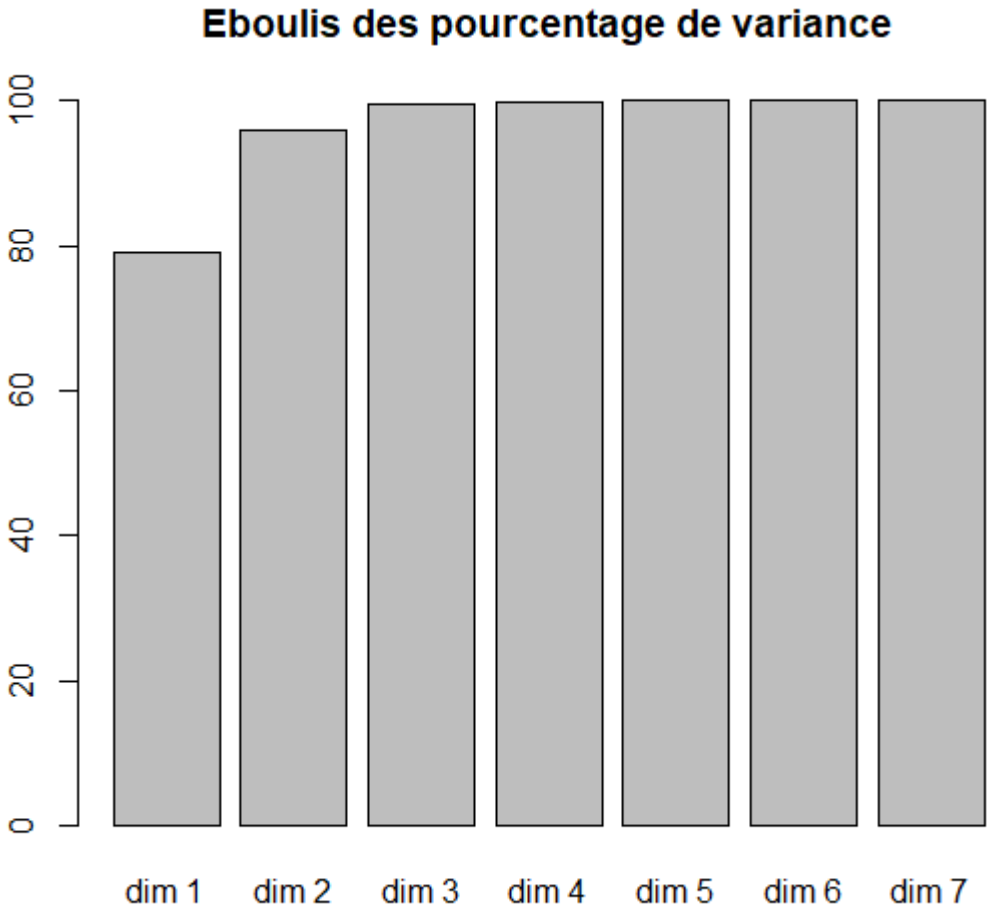
La corrélation négative entre Unemployed et Armed.Forces peut s'expliquer par le fait que lorsque l'économie est en récession et que le chômage augmente, le gouvernement peut augmenter les dépenses militaires pour stimuler l'économie. Cela peut entraîner une augmentation du nombre de personnes engagées dans l'armée et une diminution du nombre de chômeurs.

La corrélation positive entre Unemployed et Population peut s'expliquer par le fait que lorsque la population augmente, le nombre de personnes cherchant un emploi augmente également. Si l'offre d'emplois ne suit pas, cela peut entraîner une augmentation du chômage.

La corrélation positive entre Employed et Population peut s'expliquer par le fait que lorsque la population augmente, le nombre de personnes en âge de travailler augmente également. Si l'économie est en croissance et crée des emplois, cela peut entraîner une augmentation du nombre de personnes employées.

ACP

```
> acp$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 5.533067679           79.043823979                           79.04382
comp 2 1.187554644           16.965066347                           96.00889
comp 3 0.252216311            3.603090161                           99.61198
comp 4 0.015238522            0.217693171                           99.82967
comp 5 0.010636265            0.151946637                           99.98162
comp 6 0.001027941            0.014684876                           99.99631
comp 7 0.000258638            0.003694829                          100.00000
```



Les résultats ont calculé 7 dimensions. On peut observer que la première composante principale explique 79.04% de la variance dans les données. La deuxième composante principale explique 16.97% de la variance supplémentaire. Ensemble, les deux premières composantes principales expliquent 96.01% de la variance totale dans les données. Les autres composantes principales expliquent une faible proportion de la variance restante et c'est pour cela que l'on retient que les 2 premières dimensions.

VECTEUR PROPRE

```
> vecpropres
                   Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
GNP.deflator 0.4225559  0.03331479 -0.04154777  0.68544289 -0.5384123
GNP          0.4232763  0.03003906 -0.15667625 -0.18863757 -0.2148123
Unemployed   0.2791521 -0.61726052  0.67738968  0.07297657  0.2248161
Armed.Forces 0.1887305  0.77656258  0.58598762 -0.04866667  0.1046508
Population   0.4218501 -0.06924115 -0.08434184 -0.66741256 -0.2739382
Year         0.4247835 -0.02365912 -0.03157581 -0.04933020  0.1339253
Employed     0.4127227  0.09259423 -0.40419980  0.19752696  0.7137896
```

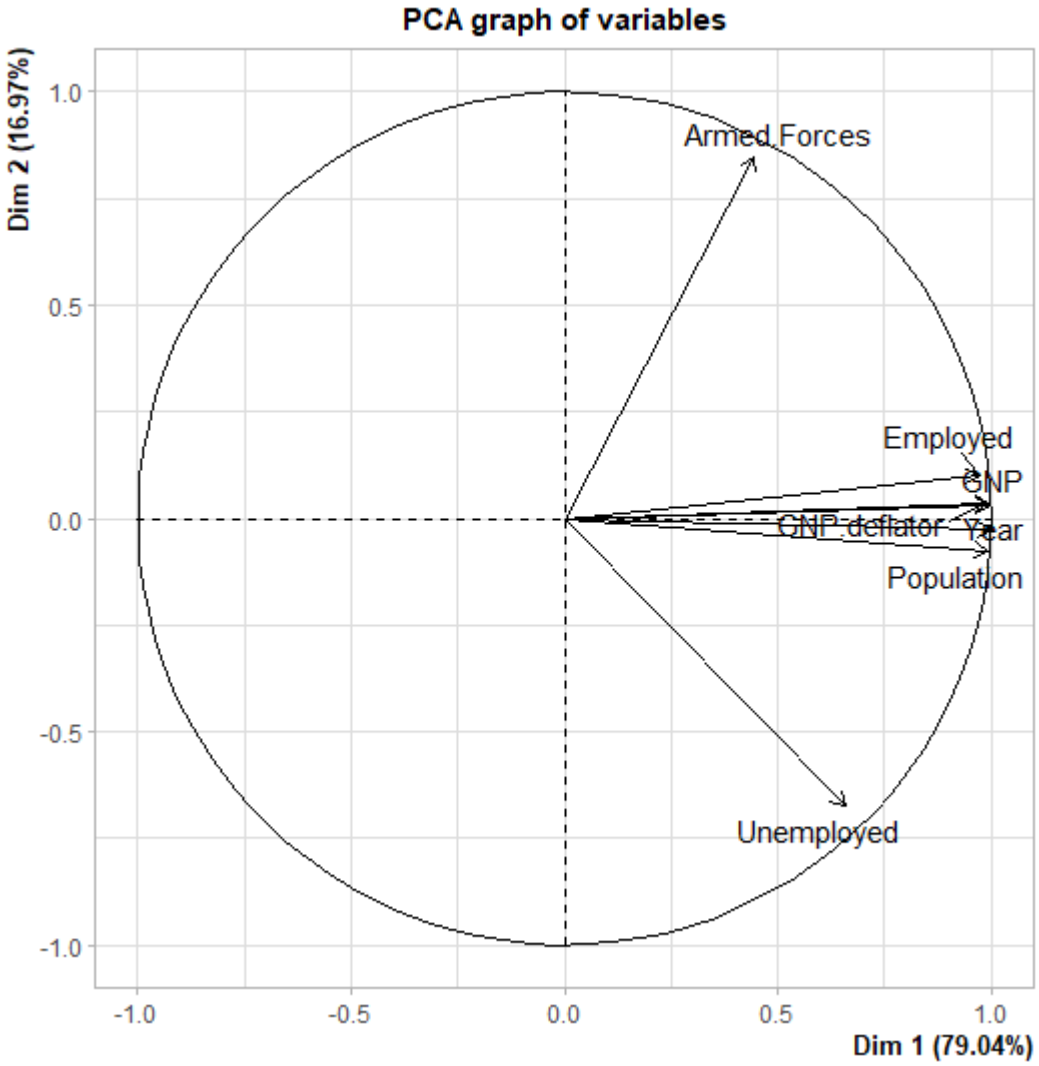
D'après ces résultats, on peut observer que pour la première dimension, toutes les variables ont des coefficients positifs, ce qui indique qu'elles contribuent toutes positivement à cette dimension. Les variables GNP.deflator, GNP et Population ont les coefficients les plus élevés.
Pour la deuxième dimension, les variables Unemployed et Armed.Forces ont des coefficients opposés et élevés en magnitude, ce qui indique qu'elles ont des contributions importantes mais opposées à cette dimension.

```
> round(acp$var$cos2[,1:3],4)
             Dim.1  Dim.2  Dim.3
GNP.deflator 0.9879 0.0013 0.0004
GNP          0.9913 0.0011 0.0062
Unemployed   0.4312 0.4525 0.1157
Armed.Forces 0.1971 0.7162 0.0866
Population   0.9847 0.0057 0.0018
Year         0.9984 0.0007 0.0003
Employed     0.9425 0.0102 0.0412
```

Les résultats du vecteur propre sont soutenus par ceux de la qualité de représentation des variables. On peut observer que pour la première dimension, les variables GNP.deflator, GNP, Population, Year et Employed ont des valeurs proches de 1, ce qui indique qu'elles sont bien représentées par cette dimension. Les variables Unemployed et Armed.Forces ont des valeurs plus faibles pour cette dimension, ce qui indique qu'elles ne sont pas aussi bien représentées.
Pour la deuxième dimension, la variable Armed.Forces a une valeur proche de 1, ce qui indique qu'elle est bien représentée par cette dimension. Les autres variables ont des valeurs plus faibles pour cette dimension, ce qui indique qu'elles ne sont pas aussi bien représentées puisqu'elles le sont dans la dimension 1.

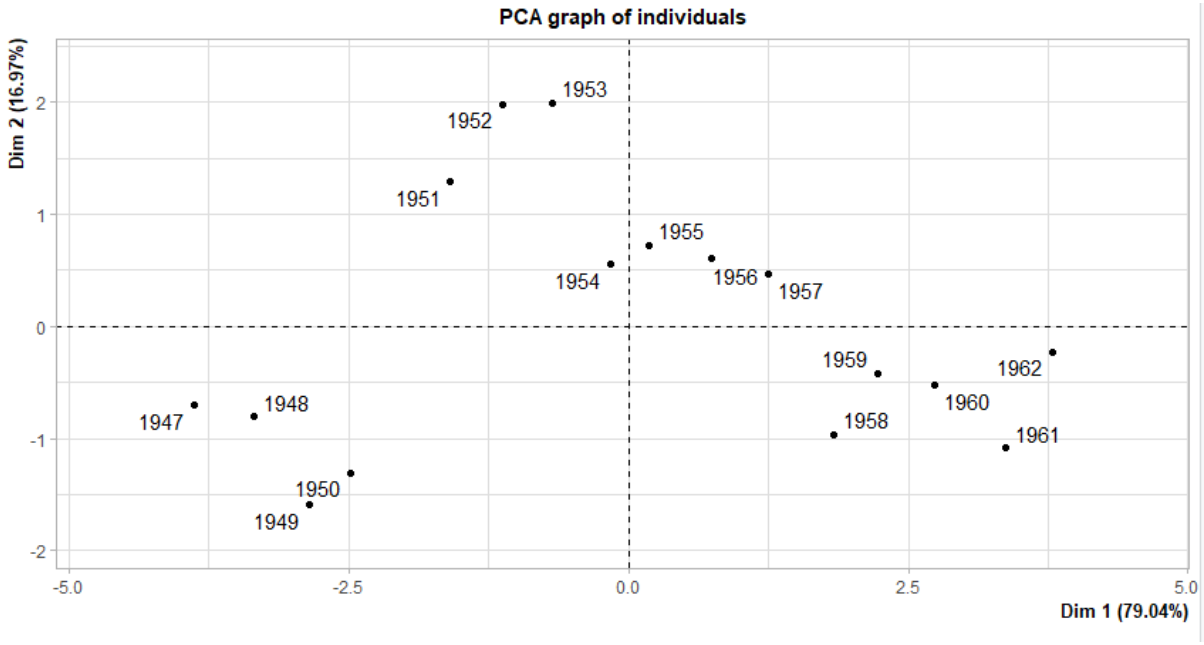
COORDONNEE DES VARIABLES

```
> round (acp$var$coord [,1:2],4)
                Dim.1   Dim.2
GNP.deflator 0.9940  0.0363
GNP          0.9957  0.0327
Unemployed   0.6566 -0.6727
Armed.Forces 0.4439  0.8463
Population   0.9923 -0.0755
Year         0.9992 -0.0258
Employed     0.9708  0.1009
> round (acp$var$cor[,1:2],4)
                Dim.1   Dim.2
GNP.deflator 0.9940  0.0363
GNP          0.9957  0.0327
Unemployed   0.6566 -0.6727
Armed.Forces 0.4439  0.8463
Population   0.9923 -0.0755
Year         0.9992 -0.0258
Employed     0.9708  0.1009
```



Ce graphique valide nos résultats car Unemployed et Armed.Forces varient énormément selon la dimension 2

D’après ces coordonnées, on peut observer que les variables GNP.deflator, GNP, Population, Year et Employed sont proches les unes des autres dans l’espace des composantes principales et forment des angles aigus entre elles, ce qui indique des liens positifs entre ces variables. La variable Unemployed et Armed.Forces forment des angles plus élevés avec ces variables, ce qui indique des liens positifs moins prononcés. Unemployed et Armed.Forces ont un angles obtus ce qui indique leur non lien.



Ici on voit que les années 1947 1948 1949 et 1950 sont proches entre elles dans les 2 dimensions, cela veut dire que ces années ont des caractéristiques similaires au regard des variables qui sont représentées sur ces axes.
De plus on peut apercevoir 3 autres groupes d’années:
un 1er groupe composée de 1958 1959 1960 1961 1962
Un second groupe composée de 1951 1952 1953
Un troisième un peu composé de 1954 1955 1956 1957

Cette étude avec des années nous permet d’identifier clairement la période des dynamiques de nos variables