

Réalisé par :

Faiz Amine

Analyse de données ACP sous R

# **INTRODUCTION**

## Analyse en composantes principales sous R :

### Analyse des résultats :

Analyser les résultats d’un ACP, c’est répondre à trois questions :

* Les données sont-elles factorisables ?
* Combien de facteurs retenir ?
* Comment interpréter les résultats ?

1. Les données sont-elles factorisables ?

Pour répondre à cette question, dans un premier temps, il convient d’observer la matrice des corrélations. Si plusieurs variables sont corrélées (> 0.5), la factorisation est possible. Si non, la factorisation n’a pas de sens et n’est donc pas conseillée.

1. Combien de facteurs retenir ?

Deux règles sont applicables :

* 1ere règle : la règle de Kaiser qui veut qu’on ne retienne que les facteurs aux valeurs propres supérieures à 1.
* 2eme règle : on choisit le nombre d’axe en fonction de la restitution minimale d’information que l’on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l’information.

1. Interprétation des résultats ?

C’est la phase la plus délicate de l’analyse. On donne un sens à un axe à partir des coordonnées, COS² (qualité de représentation) et Contribution des variables et des individus. Ce sont les éléments extrêmes qui concourent à l’élaboration des axes.

PROBLEME : Salles de cinéma

On mesure pour 94 départements de France métropolitaine les dix variables décrites

**Popu** population du département (en millions d’habitants) ;

**Entr** nombres d’entrées réalisées (en millions) ;

**Rece** recettes (en millions d’euros) ;

**Sean** nombre de séances (en millions) ;

**Comm** nombre de communes équipées de salles de cinéma ;

**Etab** nombre de cinéma en activité ;

**Salle** nombre de salles en activité ;

**Faut** nombre de fauteuils disponibles ;

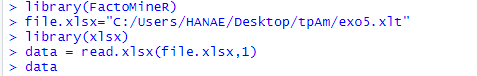
**Artes** nombre de salle d’art et essai (passant des films indépendants) ;

**Multi** nombre de multiplexes (au moins 8 salles).

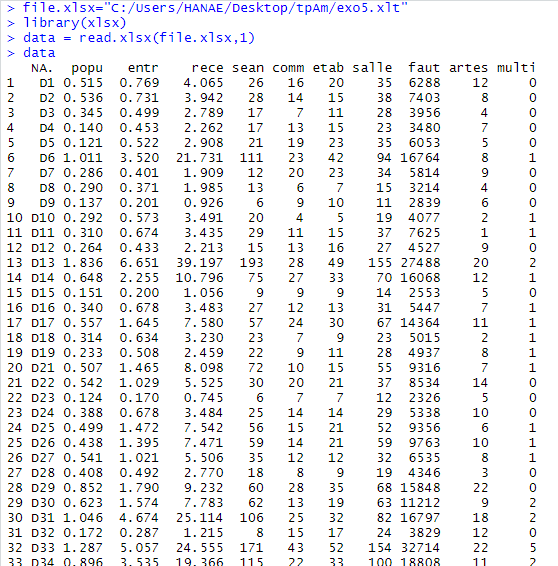
Partie 1 : première approche :

Premièrement j’ai importé la table de données d’un fichier Excel, pour établir nos analyses.

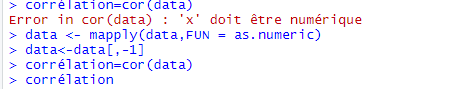
En utilisant le code suivant :



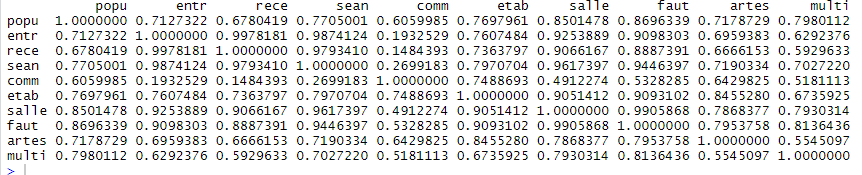
La table est bien chargée :



En utilisant le code suivant pour calculer la matrice de corrélation :



En obtient le résultat suivant :



D’après les résultats de la matrice de corrélation on constate que la méthode ACP est faisable pour notre problème ; on a une forte corrélation (> 0.5) entre les variables,

Les variables qui sont bien corrélées :

* Nombres d’entrées et Nombre de séances avec une corrélation de **0,99**
* Nombre de salle et Nombre de fauteuils avec une corrélation de **0,99**
* Recettes et Nombre de séances avec une corrélation de **0,98**
* Nombre de séances et Nombre de salle avec une corrélation de **0,96**
* Nombres d’entrées et Nombre de salle avec une corrélation de **0,93**
* Nombre de cinémas et Nombre de fauteuils avec une corrélation de **0,91**
* Nombre de fauteuils et Nombre de séances avec une corrélation de **0,94**

Question 3 :

Concernant Paris et d'après la base de données, on observe qu'elle a des valeurs plus grandes que les autres ; ce qui implique une dispersion par rapport à les autres.Partie 2 : ACP version1 :

Dans cette partie on s'intéresse par les variables d'offre de cinéma : **sean**, **comm**, **etab**, **salle**, **faut**, **artes** et multi, en appliquant **ACP**.

Question 4 :

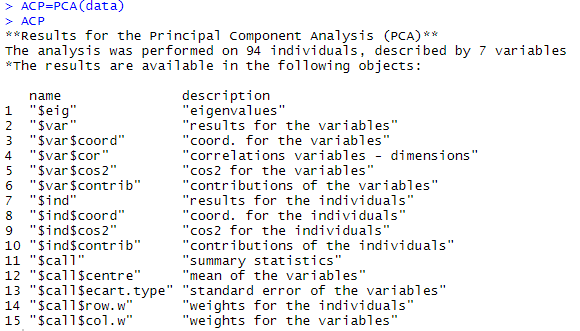
Afin d’obtient les deux traces (le cercle de corrélation et le plan de composantes principales), je vais utiliser le package :

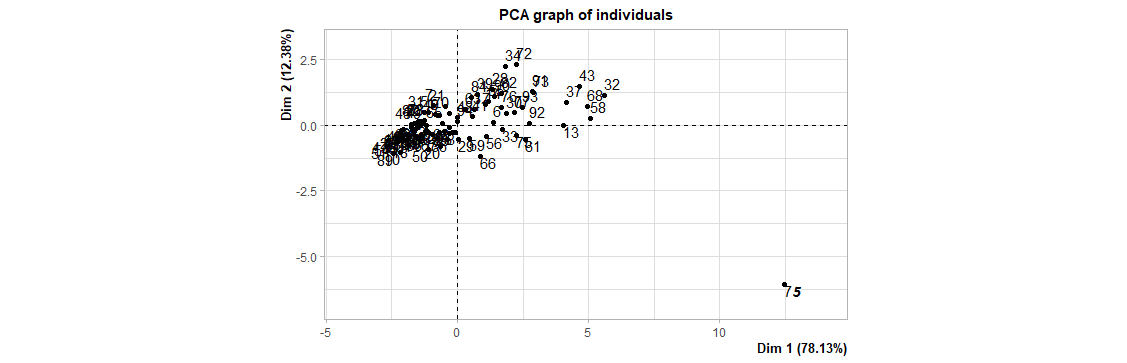
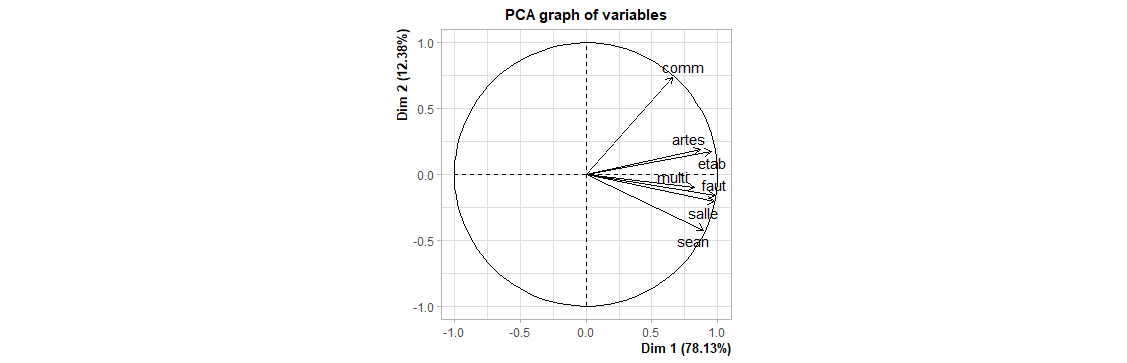
**install.packages("FactoMineR")**

Chargez-la dans R, en tapant ceci :

library("FactoMineR")

Et j’ai appliqué ACP à l’aide de code suivant :



J’ai obtenu comme résultat les deux traces :

Question 5 :

D’après le cercle on peut dire que les variables qui contribuent le plus de la construction de la première composante sont : **Etab**, **faut**, **salle** et **multi**.

Pour la seconde, elle est construite par les deux variables **Comm** et **Sean**.

Question 6 :

D’après la projection on observe que la plupart des individus sont bien représentés sur le premier plan, donc ils sont caractérisés par les variables **Etab**, **faut**, **salle** et **multi.**

Paris est particulier aux raisons de sa dispersée par rapport aux autres points.

Question 7 :

Cette approche est s’appelle : centrer et réduire les données, elle permet de donner aux variables la même importance, afin de construire les premières composantes principales.

Partie 2 : ACP version 2 :

Question 8 :

On constate que la quantité de variance expliquée par chaque axe est changée par rapport à la première ACP, et selon ces valeurs on retient deux axes principaux qui ont résumé un pourcentage important de l’inertie totale, et qui nous fournissent deux composantes principales.

Question 9 :

Selon les corrélations entre les variables et les quatre premiers axes, les variables qui déterminent les axes qu’on a retient sont : **etab**, **salle**, **faut** pour le premier axe, et pour deuxième axe est déterminer par **sean**, **multi** et **comm.**

Question 10 :

Selon les qualités de représentations des individus (les contributions), les départements qui contribuant le plus de la détermination des axes qu’on a retient sont :

Pour le premier axe : D5, D40, D73, D74, D95.

Pour le deuxième axe : D4, D5 D9, D15, D23, D28, D32, D38, D40, D46, D48, D67, D73, D74, D75, D95.

On a des départements sur-représentés : D4, D5, D40, D73, D74, D95.

Question 11 :

D’après les deux questions, on peut conclu que les départements qui ont une forte contribution avec l’axe 1, sont caractérisées par les variables : **etab**, **salle**, **faut**.

Et les départements qui ont une forte contribution avec l’axe 2, sont caractérisées par les variables **sean**, **multi** et **comm**.

Question 12 :

Les départements qui ont une qualité inférieure à 50% sont mal représentées et ayant une mauvaise qualité.

Les départements sont : D90 et D95.