Projet R

Introduction :

Le projet consiste à travailler avec l’ ACP

Il est constitué de 4 exercices.

Le but est d’appréhender l’ACP et apprendre à projeter les données de meilleur façon pour moins de perte.

Exo 1 : Répartition des entreprises automobile en fonction de leur chiffre d’affaire

1. (img : CA et ecart type)
2. (img : histo de fréquences)
3. (img : Polygone fréquence asc/desc )
4. (img : mediane, proportion entreprise > 3m)

Exo 2 : Import d’un fichier et calcul statistique

1. A
2. A
3. A
4. A
5. A
6. A

Exo 3 : Evolution du salaire en fonction de l’année

1. A
2. A
3. A

Exo 4 : l’ACP version 1

L’analyse de composante principale ou ACP est une méthode d’analyse des données qui consiste à transformer des variables corrélées en nouvelles variables décorrélé les unes des autres. On les appelles les composantes principales (ou axes principaux) .

C’est une méthode statistique d’exploration de données multivariées (données contenant plusieurs variables). Elle support l’analyse et la visualisation un jeu de données contenant des individus décrits pas plusieurs variables quantitatives.

But : identifier les variables corrélées réduire le nombre de variables et rendre l’information moins redondante plus facile à interpréter. En effet il est très compliqué de visualiser des données dans un espace multidimensionnelle.

En d’autres termes, l’ACP réduit les dimensions d’une donnée contenant plusieurs variables à deux ou trois composantes principales, qui peuvent être visualisées graphiquement, en perdant le moins possible d’information.

L’ACP est utilisé dans plusieurs domaines de nos jours tel que la biologie, le traitement d'images et le Big data dans la gestion et l'interprétation des données.

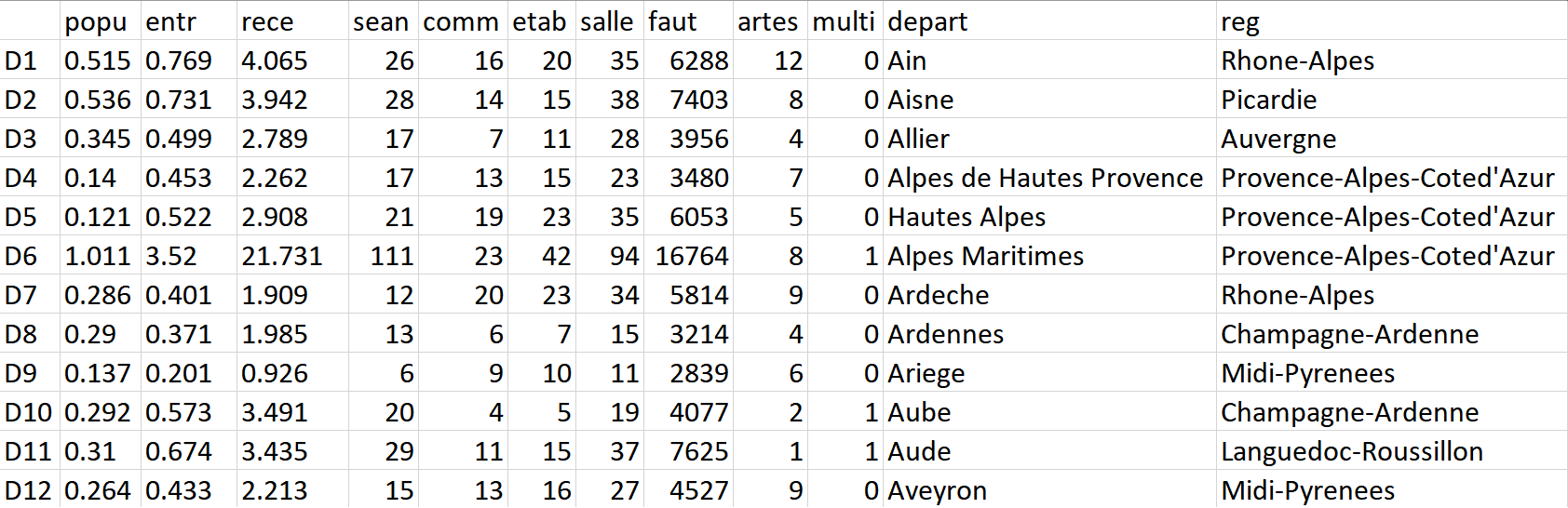
Le but de cet exercice sera de trouver les axes principaux d’analyse de cette donnée multivariée puis projeter et interpréter les données.

**Partie 1** : Première approche

On étude dix métriques différentes dans 94 départements français. Les variables sont les métriques et les individus étudiés sont les 94 départements de France.

Le programme Rstudio permet d’obtenir les tableau ci-dessous :

Data



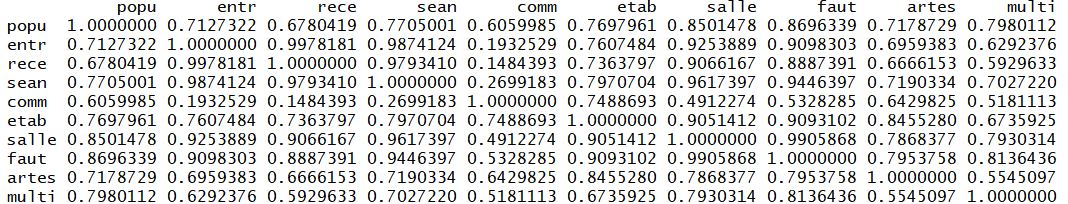
popu : population du département (en millions) entr : nombre d’entrées réalisées ( en millions)

rece : recettes (en millions d’euros) sean : nombre de séances ( en milliers)

comm : nombre de communes équipées de salles de cinéma etab : nombre de cinémas en activité

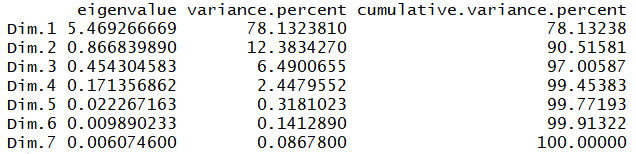
salle : nombre de salles en activité faut : nombre de fauteuils disponibles

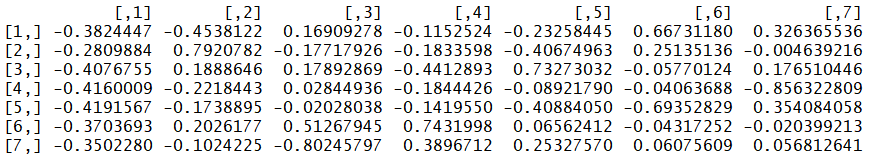
artes : nombre de salles d’art et d’essai multi : nombre de multiplexes ( au moins 8 salles)



**Tableau 1: Matrice de corrélation**

**Tableau 2: Valeurs propres**





**Tableau 3: Vecteurs propres**

1. Cf tableau 1.

Il faut regarder les valeurs fortement corrélées ( 90%): SEAN est fortement corrélé avec ENTR et SALLE

COMM est peu corrélé avec RECE

Tout d'abord, il faut noter qu'on considère une forte corrélation entre deux variables à partir de 0.90.

Les variables fortement corrélées sont :

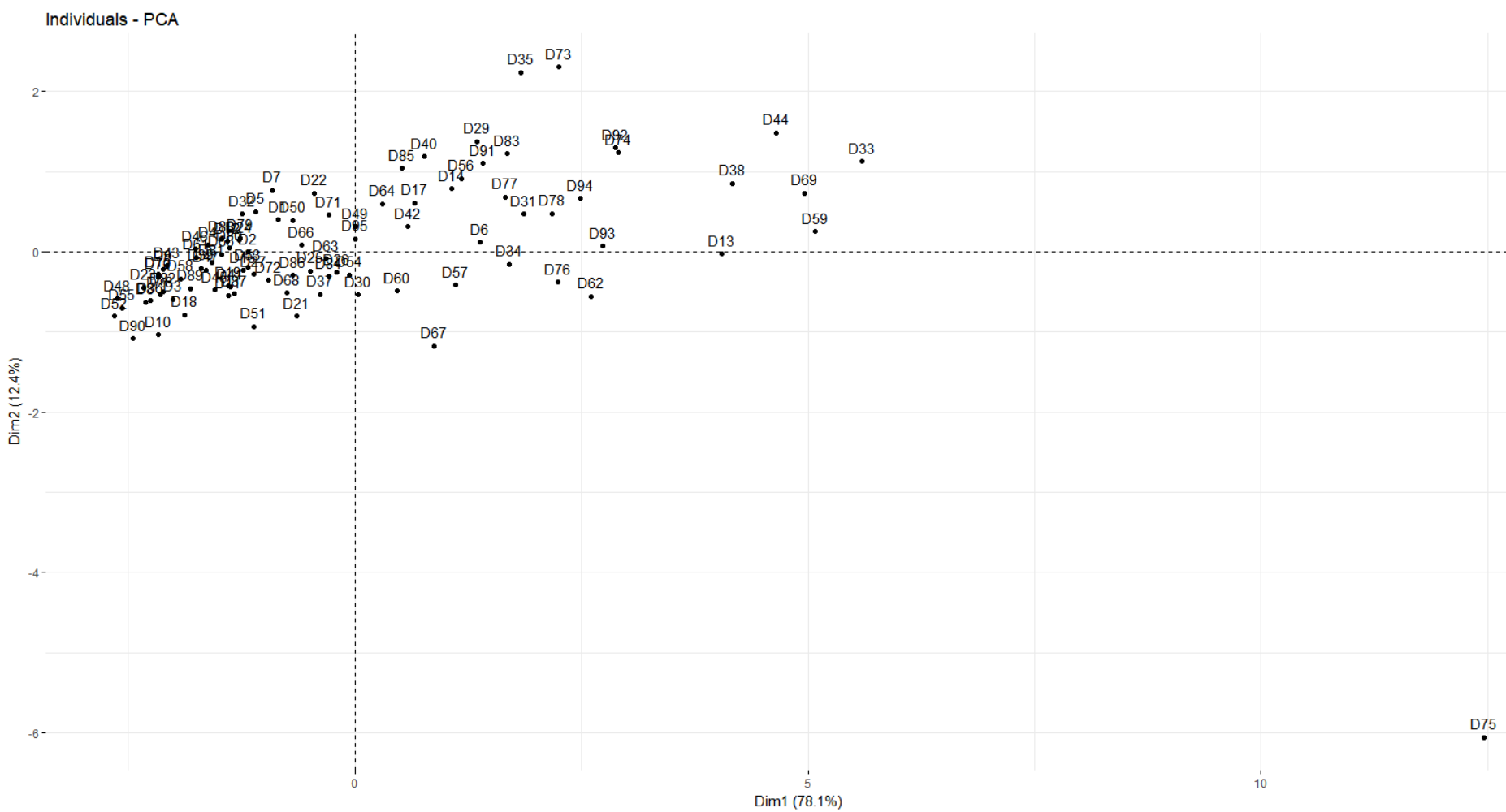
La plus grosse corrélation est entre ENTR et RECE (1.0)

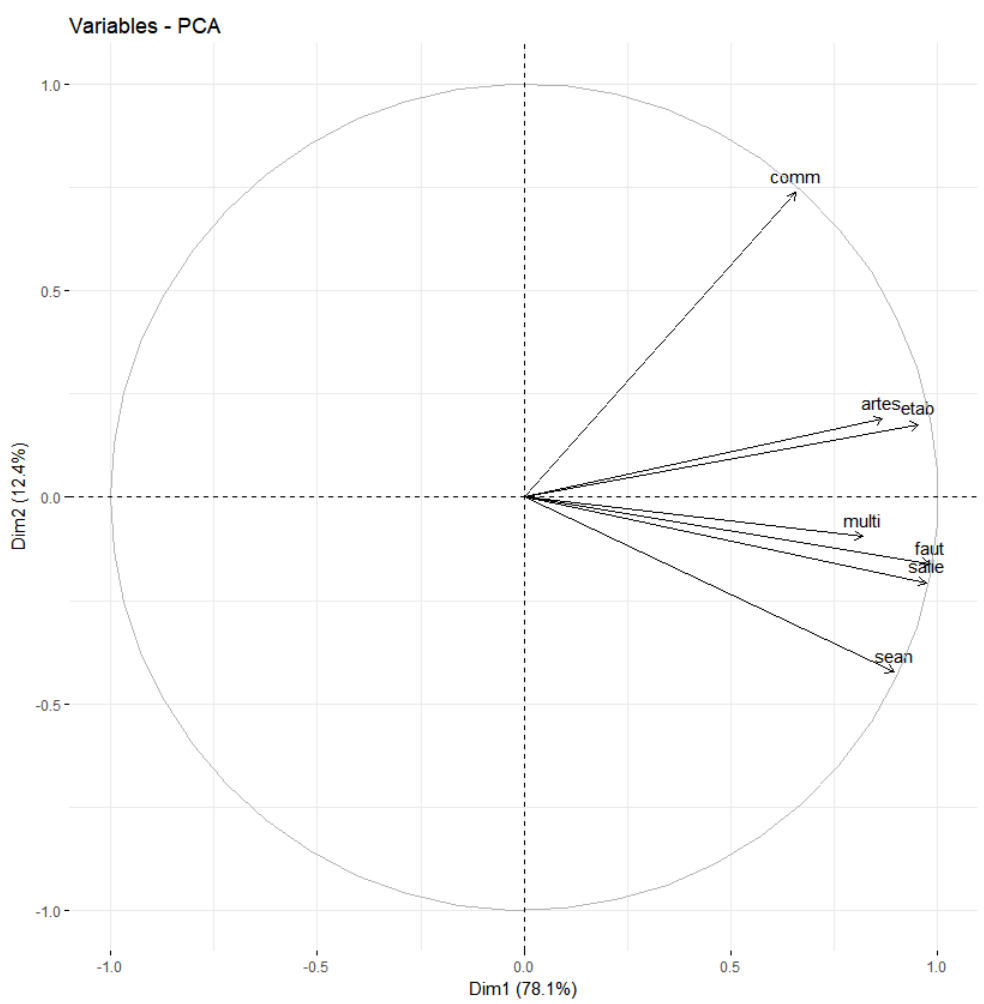
-> Ici, on peut aisément dire que plus le nombre d'entrées d'un film augmente et la recette de celui ci augmente.

Paris a une seule commune, un seul département, c'est pourquoi on peut observer que ses valeurs sont plus élevées que les autres.

De plus, c'est le deuxième département le plus peuplé (2 125 000 habitants)

**Partie 2** : ACP Version 1

Projection sur le plan d’inertie maximale

Cercle de corrélation

Par rapport à la première composante principale F1:

On remarque que les variables sont toute corrélées positivement à F1. On observe aussi que la variable sean est la moins corrélé et comm la plus corrélé

On admet qu'une hypothèse est vrai si sa probabilité qu'elle est une corrélation avec la première composante principale est supèrieur à 90%.

*Variables :* On sait que les variables contribuant le plus à la formation de l’axe 1 sont celles dont les coordonnées sur cet axe sont proches de 1 en valeur absolue. Pour repérer les contributions significatives, on utilise le tableau 3 : on compare les valeurs de la colonne Dim1 coordonnées du premier axe factoriel, à la racine de la contribution moyenne 1/38%, le signe donnant le sens de contribution. On obtient :

D'après le graphique, on a:

- les variables nombre de salles, nombre de fauteuils dispo et nombre de cinémas en activité qui sont fortement corrélé positivement. De plus, les variables représentant le nombre de fauteuils dispo et le nombre de salles en activité sont très proches, donc il y’a une corrélation entre ces deux variables. A l’inverse, on remarque que la variable représentant le nombre de communes équipé de salle de ciné est peu corrélée avec la première composante principale.

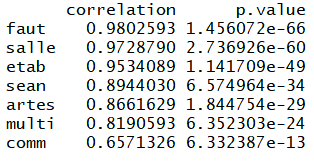
Par ailleurs, toutes les variables hormis multiplex semblent bien être représenté.

Hypothèses plausibles :

H0 = nombre de spectateur

H0 = nombre de vente de billet

(voir p-value pour départager)



Donc : [Resultat interpretation]

Par rapport à la première composante principale F2:

On observe que la variable comm est la plus corrélé et sean la moins corrélé, elle est même corrélé négativement.

- la variable nombre de salles de ciné est fortement corrélé à l’axe de la composante principale 2. De plus, les variables représentant le nombre de fauteuils dispo et le nombre de salles en activité sont très proches, donc il peut y’avoir une corrélation entre ces deux variables. Egalement, les variables nombre de salles d’arts et cinéma en activité sont aussi proches, ils doivent peut-être aussi être corrélé. A l’inverse, on remarque que la variable représentant le nombre de séances est corrélé négativement avec la deuxième composante principale.

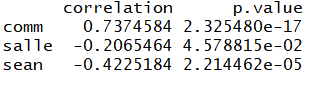
Hypothèse plausible :

H0.1 = nombre de salle pleine

H0.2 = nombre de salle vide

H0.3 = distance entre les visiteurs et le cine

(voir p-value pour départager)



Donc : [Resultat interpretation]

On observe que la tendance est presque la meme dans tout les regions sauf paris.

Nous avons vu à la question 2 que Paris est différentes des autres regions car elle ne possède qu'un département et une seule commune elle-même.

De plus, c'est le deuxième département le plus peuplé .

opposé à paris :

belfort

chaumont

Mende

meme tendance que paris sur l'axe 1:

marseille

lyon

bordeaux

lille

nante

grenoble

Strasbourg

meme tendance que paris sur l'axe 2:

D67, D90,D76,D57

opposé à paris sur l’axe 2:

chambery

rennes

les departements n'ont pas le meme nombre de population. Une sorte de normalisation des données. Permet de se concentrer plus sur la tendance de la population selon les regions.

**Partie 3** : ACP Version 2



Rendre les variables comparables

ecart type =1

moyenne = 0

ACP normée

On retient les 4 premiers axes principaux sur les sept.

Oui, la situation est meilleur car cette fois nous pouvons retenir deux axes, un de plus que pour les données non centrées réduites et sur des variables non normalisées.

Les valeurs propres peuvent être utilisées pour déterminer le nombre d’axes principaux à conserver après l’ACP. Il n’existe pas de méthode objective bien acceptée pour décider du nombre d’axes principaux qui suffisent.

Distance entre l’individu et le point d’origine du cercle qualité de représentation des variables. Plus le point est proche du cercle et plus la donnée sera vrai.

Pour savoir quels axes nous retenons, nous décidons de se référer au « critère de Kaiser ». Celui-ci nous dit que pour des variables centrées-réduites nos choix d’axes doivent se porter celles qui ont des valeurs propres supérieures à 1.

Nous allons donc retenir l’axe 1 et 2 qui totalise plus de 82% (3.71+2.04/7) des contributions, ce qui déjà pas mal.

D’après le cours, l’effet de taille est vérifié lorsque toutes les variables ont le même signe de corrélation avec la première composante principale. Dans notre exemple, on peut voir que toutes les valeurs de corrélations des premières composantes sont négatives, on peut donc dire qu’il y’a un effet de taille. les variables sont toutes du même côté de l’axe.

D’après le cours, l’effet de taille est vérifié lorsque toutes les variables ont le même signe de corrélation avec la première composante principale. Dans notre exemple, on peut voir que toutes les valeurs de corrélations des premières composantes sont négatives, on peut donc dire qu’il y’a un effet de taille.

Le poids des départements est moins disparate que pour l’ACP non normée

[ Interpretation ]

L’interprétation de l’ACP peut être achevée par une analyse des proximités entre points sur les plans factoriels. En effet, au vu des graphiques, on peut distinguer plusieurs groupes. Avant d’interpréter ces proximités, il s’agit de s’assurer de la qualité de représentation des points :

Je vois que le dernier sous espace est celui dont les données sont très clairement représentées c’est-à-dire que c’est le plan d’inertie maximum où les données peuvent être projeté. Nous voyons clairement que c’est le département 90 et 94 qui sont les moins bien représenté selon le tableau traduisant les critères de la qualité des représentations des individus sur le plan Axis1 :4.

Conclusion

L’ACP ou l’analyse de composante principale nous a aidé à résumer et interpréter les data set contenant plusieurs variables avec plus ou moins de précisions. Dans l’étude de composante principal, l’information contenue dans les données se résume en un ensemble de multiple variation de valeurs contenu dans celle-ci.

Cette méthode nous a permis de réduire les dimensions d’un data set contenant plusieurs variables puis visualisér graphiquement, en perdant le moins possible d’information.