

Rapport de projet analyse de données

Analyse En Composantes Principales Sur Espérance De Vie En France



Réalisé par

BASSOUL Ahmed
El MSSAFI Ibrahim

Enseignant responsable

EL HANNOUN Wafaa

Date

12/02/2024

Table des matières

i. Introduction	4
ii. Chapitre 1 : Préparation des données	5
ii.1. Introduction	5
ii.2. Collection de données	5
iii. Chapitre 2 : Analyse en Composantes Principales (ACP)	8
iii.1. Introduction	8
iii.2. Comment fonctionne l'ACP	8
iii.2.1. Définition	8
iii.2.2. Standardisation des données	8
iii.2.3. Calcul de la matrice de covariance	9
iii.2.4. Calcul du vecteur de caractéristiques	9
iii.2.5. Multiplication des données normalisées par les vecteurs propres	10
iii.3. Exemple	10
iii.4. Conclusion	12
iv. Chapitre 3 : Application d'ACP sur données	13
iv.1. Introduction	13
iv.2. Chargement des bibliothèques	13
iv.3. Chargement des données	13
iv.4. Transformation des données	14
iv.5. Application d'ACP	15
iv.5.1. Matrice de corrélation	15
iv.5.2. Choisir le nombre de composante principale	16
iv.5.3. Application de cpa	17
iv.6. Implementation avec Python	18
iv.7. Comparaison des résultats R/Python	19
v. Chapitre 4 : Visualisation et Interprétation	20
v.1. Introduction	20
v.2. Quantité d'informations expliquée par chaque composant	20
v.3. Corrélation des variables	22
v.4. Contribution des variables/individus dans chaque CP	23
v.5. Qualité de représentation des variables/individus dans chaque CP	24
v.6. Interprétation par Biplot	25
v.7. Résumé	27
vi. Conclusion	28
vii. Référence	29

Table des figures

1.	Le site web <code>www.insee.fr</code>	5
2.	examination du site web	6
3.	library	6
4.	accéder au site web	6
5.	stockage de jeu de donnée csv	7
6.	Le fichier <code>donnees_tableau.csv</code>	7
7.	bibliothèques	13
8.	chargement des données	14
9.	Notre Dataset	14
10.	transformation des données	14
11.	Calcul de la Matrice de corrélation	15
12.	Valeur de la Matrice de corrélation	15
13.	<code>cortest.bartlett(matrCorr, n = 78)</code>	16
14.	<code>valeur_propre=get_eigenvalue(resultACP)</code>	16
15.	Commande PCA	17
16.	Interpretation de resultat ACP	17
17.	Implementation ACP en Python	18
18.	Resultat d'analyse en python	19
19.	Commande <code>screen plot</code>	20
20.	resultat de <code>screen plot</code>	21
21.	Corrélation des variables	22
22.	Histogramme de la Contribution des variables à la dimension 1	23
23.	Histogramme de la Contribution des variables à la dimension 2	24
24.	Commande de Contribution	24
25.	Corrélogramme	25
26.	Commande pour afficher le biplot	25
27.	Biplot de la dimension 1 et 2	26

i. Introduction

L'espérance de vie est un indicateur crucial de la santé et du bien-être d'une population. En France, comme dans de nombreux autres pays, elle est étroitement surveillée et constitue un aspect central des politiques de santé publique et de planification sociale.

Ce rapport se penche sur une analyse approfondie de l'espérance de vie en France, en examinant les tendances, les déterminants et les variations régionales de cet indicateur clé. En effet, comprendre les facteurs qui influent sur l'espérance de vie est essentiel pour orienter les politiques de santé, allouer les ressources de manière appropriée et améliorer la qualité de vie de la population.

L'analyse sera basée sur l'Application de l'Analyse en Composantes Principales (ACP), une méthode statistique puissante qui permettra de mettre en évidence les principaux déterminants de l'espérance de vie et d'explorer les relations complexes entre les différents facteurs.

L'objectif principal de cette étude est de fournir des insights précieux sur les tendances de l'espérance de vie en France, d'identifier les disparités socio-économiques et géographiques, et de proposer des recommandations pour promouvoir la santé et le bien-être de la population française. En fin de compte, cette analyse vise à soutenir les efforts visant à améliorer la qualité de vie et à favoriser le vieillissement en bonne santé.

ii. Chapitre 1 : Préparation des données

ii.1. Introduction

La collecte et la préparation des données constituent des étapes fondamentales dans notre étude sur l'Analyse en Composantes Principales (ACP) de l'espérance de vie en France. Nous avons recueilli nos données à partir du site de l'Institut National de la Statistique et des Études Économiques (INSEE), une source réputée et fiable en matière de statistiques démographiques et sociales en France.

ii.2. Collection de données

Pour obtenir les données nécessaires à notre étude sur l'Analyse en Composantes Principales (ACP) de l'espérance de vie en France, nous avons décidé de recourir à la technique du web scraping sur le site de l'Institut National de la Statistique et des Études Économiques (INSEE) à l'adresse <https://www.insee.fr>. Le choix du web scraping s'est avéré être une solution efficace pour extraire un jeu de données répondant à nos besoins spécifiques. En explorant le site de l'INSEE, nous avons identifié les sections pertinentes contenant les données sur l'espérance de vie et d'autres indicateurs démographiques clés. Pour automatiser ce processus, nous avons utilisé la bibliothèque Python BeautifulSoup, un outil puissant pour extraire des données à partir de pages web. En combinant BeautifulSoup avec d'autres bibliothèques et outils de scraping, nous avons pu récupérer rapidement et efficacement les données nécessaires pour notre analyse. Cette approche nous a offert une source de données actualisée et fiable, nous permettant ainsi de mener notre étude avec un ensemble de données complet et pertinent.

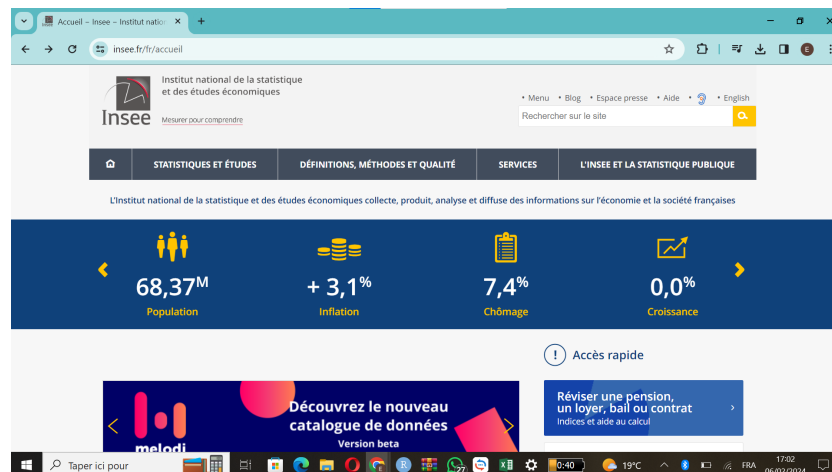


FIGURE 1 – Le site web www.insee.fr

La prochaine étape consiste à examiner le site Web pour comprendre sa structure et les données que nous souhaitons extraire.

Espérance de vie à divers âges et taux de mortalité infantile, France métropolitaine												
Année	Espérance de vie des hommes (en années)					Espérance de vie des femmes (en années)						
	0 à 1 an	1 à 20 ans	20 à 40 ans	40 à 60 ans	60 à 65 ans	0 à 1 an	1 à 20 ans	20 à 40 ans	40 à 60 ans	60 à 65 ans		
1946	59,9	64,4	48,0	30,8	15,4	12,2	68,9	52,2	34,7	18,0	14,3	
1947	61,2	65,3	48,4	30,9	15,5	12,3	66,7	70,0	52,9	35,1	18,2	14,5
1948	62,7	65,9	48,5	30,9	15,6	12,5	68,8	71,2	53,6	35,6	18,7	15,0

FIGURE 2 – examination du site web

Le code utilisé pour extraire les données est découpé en plusieurs étapes :
 Première étape : Nous importons les outils nécessaires à notre programme, comme requests pour accéder à des pages web, BeautifulSoup pour analyser le code HTML et csv pour manipuler des fichiers CSV.

```

1 import requests
2 from bs4 import BeautifulSoup
3 import csv

```

FIGURE 3 – library

Deuxième étape : Nous définissons l'adresse web de la page contenant le tableau que nous voulons extraire, et nous envoyons une demande au serveur web pour obtenir le contenu de la page web spécifiée. et Nous vérifions si la requête a réussi. Aussi nous utilisons BeautifulSoup pour analyser le contenu HTML de la page et extraire les données qui nous intéressent

```

1 url = 'https://www.insee.fr/fr/statistiques/7746166?sommaire=7746197'
2
3 response = requests.get(url)
4
5 if response.status_code == 200:
6     soup = BeautifulSoup(response.text, 'html.parser')
7

```

FIGURE 4 – accéder au site web

Troisième étape : Dans cette partie du code, nous nous concentrons sur l'extraction et le stockage des données du tableau présent sur la page web. Tout d'abord, nous utilisons l'ID spécifique du tableau pour le localiser dans le code HTML. Ensuite, nous parcourons le tableau pour extraire les données, en naviguant ligne par ligne et cellule par cellule. Une fois toutes les données extraites, nous les stockons dans un fichier CSV pour une utilisation ultérieure. Cette approche nous permet d'automatiser le processus d'extraction et de stockage des données, rendant ainsi leur analyse plus facile et plus efficace.

```

1  tableau = soup.find('table', {'id': 'produit-tableau-figure2'})
2
3  if tableau:
4      with open('donnees_tableau.csv', 'w', newline='', encoding='utf-8') as csvfile:
5          csvwriter = csv.writer(csvfile)
6          headers = [header.text.strip() for header in tableau.select('thead tr th')]
7          csvwriter.writerow(headers)
8          for row in tableau.select('tbody tr'):
9              cells = row.find_all(['th', 'td'])
10             row_data = [cell.text.strip() for cell in cells]
11             csvwriter.writerow(row_data)
12

```

FIGURE 5 – stockage de jeu de donnée csv

Et finalement, un fichier nommé "donnees_tableau.csv" est créé.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	50,06	A 0 ans	A 1 an	A 20 ans	A 40 ans	A 60 ans	A 80 ans	A 90 ans	A 95 ans										
2	1946	59,9	64,4	48,0	30,8	15,4	12,2	65,2	68,9	52,2	34,7	18,0	14,3						
3	1947	61,2	65,3	48,4	30,9	15,5	12,3	66,7	70,0	52,9	35,1	18,2	14,5						
4	1948	62,7	65,9	48,5	30,9	15,6	12,5	68,8	71,2	53,6	35,6	18,7	15,0						
5	1949	62,2	65,5	48,2	30,3	14,9	11,7	67,6	70,2	52,7	34,5	17,7	14,0						
6	1950	63,4	66,2	48,7	30,7	15,4	12,2	69,2	71,4	53,6	35,2	18,4	14,6						
7	1951	63,1	65,9	48,2	30,2	14,9	11,8	68,9	71,0	53,2	34,8	17,9	14,2						
8	1952	64,4	66,8	49,1	30,9	15,5	12,3	70,2	72,1	54,2	35,6	18,6	14,8						
9	1953	64,3	66,4	48,6	30,3	15,0	11,8	70,3	71,8	53,8	35,1	18,1	14,4						
10	1954	65,0	67,1	49,2	31,0	15,5	12,4	71,2	72,8	54,7	35,9	18,9	15,1						
11	1955	65,2	67,1	49,2	30,9	15,4	12,3	71,5	73,0	54,8	36,0	18,9	15,1						
12	1956	65,2	66,9	48,9	30,6	15,2	12,0	71,7	72,9	54,7	35,9	18,7	14,9						
13	1957	65,5	67,0	49,1	30,8	15,3	12,2	72,2	73,4	55,2	36,3	19,0	15,2						
14	1958	66,8	68,3	50,2	31,8	16,0	12,8	73,2	74,3	56,0	37,0	19,5	15,6						
15	1959	66,8	68,1	50,1	31,7	15,9	12,8	73,4	74,3	56,0	37,0	19,8	15,7						
16	1960	67,0	68,1	50,0	31,5	15,7	12,6	73,6	74,4	56,0	37,0	19,5	15,6						
17	1961	67,5	68,5	50,3	32,0	16,1	13,0	74,4	75,1	56,7	37,6	20,1	16,1						
18	1962	67,0	68,0	49,9	31,5	15,7	12,5	73,9	74,6	56,2	37,1	19,6	15,7						
19	1963	66,8	67,8	49,7	31,2	15,5	12,4	73,8	74,3	56,1	37,1	19,5	15,6						
20	1964	67,7	68,5	50,3	31,9	16,0	12,9	74,8	75,2	57,0	37,9	20,3	16,3						
21	1965	67,5	68,1	50,0	31,6	15,7	12,6	74,7	75,0	56,7	37,7	20,1	16,1						
22	1966	67,8	68,5	50,3	31,9	16,1	12,9	75,2	75,5	57,2	38,1	20,5	16,5						
23	1967	67,8	68,4	50,2	31,8	15,9	12,8	75,2	75,4	57,1	38,0	20,4	16,5						
24	1968	67,8	68,3	50,2	31,7	15,8	12,7	75,2	75,5	57,2	38,0	20,4	16,4						
25	1969	67,4	67,9	49,8	31,4	15,6	12,5	75,1	75,4	57,0	37,8	20,2	16,3						
26	1970	68,4	68,8	50,7	32,3	16,2	13,0	75,9	76,1	57,6	38,5	20,8	16,8						
27	1971	68,3	68,7	50,5	32,1	16,2	13,0	75,9	76,1	57,7	38,5	20,8	16,8						
28	1972	68,5	68,7	50,6	32,3	16,4	13,1	76,2	76,3	57,9	38,7	21,1	17,0						
29	1973	68,7	68,9	50,8	32,3	16,4	13,1	76,3	76,4	57,9	38,8	21,0	17,0						
30	1974	68,5	68,1	50,5	32,5	16,5	13,5	76,7	76,7	58,1	39,1	21,3	17,2						

FIGURE 6 – Le fichier donnees_tableau.csv

iii. Chapitre 2 : Analyse en Composantes Principales (ACP)

iii.1. Introduction

L'analyse en composante principale ACP (Jolliffe, 1986) est une méthode basée sur des statistiques descriptives multidimensionnelles permettant de traiter simultanément un nombre quelconque de variables quantitatives. Le cas de plusieurs individus (n individus) mesurés par rapport à un grand nombre de variables numériques. Ces variables sont la plupart du temps corrélées entre elles.

Elle consiste à rechercher des facteurs en nombre restreint en résumant le mieux possible les données considérées. Elle aboutit à des représentations graphiques des données (des individus comme des variables) par rapport à ces facteurs représentés comme des axes. Ces représentations graphiques sont du type nuage de points.

Proposée par Hotelling en 1933 mais elle n'est devenue une technique opérationnelle qu'à partir des années 60 avec le développement des outils informatiques.

Cette méthode a été réinterprétée sous un formalisme probabiliste par Tipping et Bishop en 1999, elle a de nombreuses applications comprennent la compression de données, le traitement de l'image, la visualisation, l'analyse exploratoire des données, la reconnaissance des formes et la prévision des séries chronologiques.

iii.2. Comment fonctionne l'ACP

iii.2.1. Définition

Selon le manuel, l'ACP est définie comme "une procédure statistique qui utilise une transformation orthogonale pour convertir un ensemble d'observations de variables potentiellement corrélées en un ensemble de valeurs de variables linéairement non corrélées appelées composantes principales". En d'autres termes, l'ACP permet de réduire la dimensionnalité des données en transformant les variables d'origine en un nouveau jeu de variables, les composantes principales, qui capturent l'essentiel de la variation dans les données sans corrélation entre elles.

La réalisation d'une Analyse en composantes principales nécessite les 4 étapes suivantes :

iii.2.2. Standardisation des données

Tout d'abord, nous devons standardiser les données car cela garantit que toutes les fonctionnalités sont à la même échelle, ce qui est nécessaire pour que ACP fonctionne correctement.

$$x_{ik} \rightarrow \frac{x_{ik} - \bar{X}_k}{S_k} \quad \text{avec}$$

x_{ik} : la valeur associée à l'individu i et la variable k

$$S_k = \sqrt{\frac{\sum_{i=1}^n (x_{ik} - \bar{X}_k)^2}{n-1}} : \text{L'écart-type des valeurs dans la colonne (variable) } k \quad (1)$$

$$\bar{X}_k = \frac{1}{n} \sum_{i=1}^n x_{ik} \quad \text{La moyenne des valeurs dans la colonne (variable) } k$$

n : nombre de valeurs dans l'ensemble de données

iii.2.3. Calcul de la matrice de covariance

Une fois les données standardisées, Nous devons calculer la matrice de covariance. Cette matrice mesure la variance conjointe entre chaque paire de variables. Elle est utilisée pour calculer les vecteurs propres et les valeurs propres qui définissent les composantes principales.

$$M = \begin{pmatrix} \text{Cov}(Y_1, Y_1) & \dots & \text{Cov}(Y_1, Y_k) \\ \vdots & \ddots & \vdots \\ \text{Cov}(Y_k, Y_1) & \dots & \text{Cov}(Y_k, Y_k) \end{pmatrix}$$

— Y_1, Y_2, \dots, Y_k : ce sont les variables.

— $\text{Cov}(Y_i, Y_j) = \bar{Y}_j \bar{Y}_i - \bar{Y}_j \cdot \bar{Y}_i$: la covariance entre la variable Y_j et Y_i .

iii.2.4. Calcul du vecteur de caractéristiques

Ensuite, Nous calculons les vecteurs propres de la matrice de covariance. Ces vecteurs propres, également appelés vecteurs caractéristiques, représentent les directions dans lesquelles les données varient le plus. Ils sont utilisés pour déterminer les composantes principales de vos données.

Pour le calculer, nous commençons par les valeurs propres λ_i qui sont calculées par l'équation suivante :

$$|\mathbf{M} - \lambda \mathbf{I}| = 0 \quad (2)$$

où M est la matrice de covariance et I la matrice Identité.

Ensuite, nous calculons les vecteurs propres ω_i associés à chaque valeur propre λ_i en utilisant l'équation suivante :

$$M \cdot \omega_i = \lambda_i \cdot \omega_i$$

Enfin, nous classons les valeurs propres et leurs vecteurs propres correspondants par ordre décroissant. Nous sélectionnons ensuite le vecteur de caractéristiques qui comprend

les premiers vecteurs propres. Cette sélection est basée sur le critère que la somme des valeurs propres de ces vecteurs représente au moins 80% de la variation totale. Les autres vecteurs propres sont ignorés.

iii.2.5. Multiplication des données normalisées par les vecteurs propres

Enfin, pour obtenir les scores des composantes principales pour chaque observation dans vos données, vous multipliez les données standardisées par les vecteurs propres. Cela projette les données dans l'espace des composantes principales, où chaque dimension représente une combinaison linéaire des variables d'origine.

iii.3. Exemple

Une présentation très élémentaire de cette démarche est proposée sur un exemple jouet de données. Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

Étudiant	Math	Phys	Fran	Angl
Jean	6.00	6.00	5.00	5.50
Alan	8.00	8.00	8.00	8.00
Anni	6.00	7.00	11.0	9.50
Moni	14.5	14.5	15.5	15.0

1. Standardisation de données :

Étudiant	Math	Phys	Fran	Angl
Jean	-0.23	-0.25	-0.32	-0.33
Alan	-0.04	-0.07	-0.12	-0.12
Anni	-0.23	-0.16	0.07	0
Moni	0.51	0.50	0.37	0.45

2. Calculer la matrice de covariance :

$$M = \begin{pmatrix} \text{Var}(\text{Math}) & \text{Cov}(\text{Phys}, \text{Math}) & \text{Cov}(\text{Fran}, \text{Math}) & \text{Cov}(\text{Angl}, \text{Math}) \\ \text{Cov}(\text{Math}, \text{Phys}) & \text{Var}(\text{Phys}) & \text{Cov}(\text{Fran}, \text{Phys}) & \text{Cov}(\text{Angl}, \text{Phys}) \\ \text{Cov}(\text{Math}, \text{Fran}) & \text{Cov}(\text{Phys}, \text{Fran}) & \text{Var}(\text{Fran}) & \text{Cov}(\text{Angl}, \text{Fran}) \\ \text{Cov}(\text{Math}, \text{Angl}) & \text{Cov}(\text{Phys}, \text{Angl}) & \text{Cov}(\text{Fran}, \text{Angl}) & \text{Var}(\text{Angl}) \end{pmatrix}$$

$$M = \begin{pmatrix} 1 & 0.08 & 0.06 & 0.07 \\ 0.08 & 1 & 0.06 & 0.07 \\ 0.06 & 0.06 & 1 & 0.07 \\ 0.07 & 0.07 & 0.07 & 1 \end{pmatrix}$$

3. Calculez le vecteur de caractéristiques :

$$\begin{aligned}
 v_1 &= \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} & \lambda_1 &= 0.92 \\
 v_2 &= \begin{pmatrix} -0.22 \\ -0.22 \\ -0.58 \\ 1 \end{pmatrix} & \lambda_2 &= 0.92 \\
 v_3 &= \begin{pmatrix} -2.02 \\ -2.02 \\ 3.3 \\ 1 \end{pmatrix} & \lambda_3 &= 0.95 \\
 v_4 &= \begin{pmatrix} 1 \\ 1 \\ 0.93 \\ 1 \end{pmatrix} & \lambda_4 &= 1.2
 \end{aligned}$$

Dans ce cas, la quatrième composante principale (c'est-à-dire le quatrième vecteur propre) explique le plus de variance dans les données (30.00%) et la troisième composante principale explique la variance suivante (23.75%). La deuxième composante principale explique aussi de variance (23.00%) et alors on peut ignorer la première.

Le vecteur de caractéristiques sera :

$$\begin{pmatrix} 1 & 2.02 & 0.22 \\ 1 & 2.02 & 0.22 \\ 0.93 & 3.3 & 3.3 \\ 1 & 1 & 1 \end{pmatrix}$$

4. Multipliez les données normalisées par les vecteurs propres :

Données Standardisées x Vecteur Caractéristiques =

$$\begin{pmatrix} -0.23 & -0.25 & -0.32 & -0.33 \\ -0.04 & -0.07 & -0.12 & -0.12 \\ -0.23 & -0.16 & 0.07 & 0 \\ 0.51 & 0.50 & 0.37 & 0.45 \end{pmatrix} \times \begin{pmatrix} 1 & 2.02 & 0.22 \\ 1 & 2.02 & 0.22 \\ 0.93 & 3.3 & 3.3 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -1.1076 & -2.3556 & -1.4916 \\ -0.3416 & -0.7382 & -0.5402 \\ -0.3249 & -0.5568 & 0.1452 \\ 1.8041 & 3.7112 & 1.8932 \end{pmatrix}$$

Les données finales :

Étudiant	CP1	CP2	CP3
Jean	-1.1076	-2.3556	-1.4916
Alan	-0.3416	-0.7382	-0.5402
Anni	-0.3249	-0.5568	0.1452
Moni	1.8041	3.7112	1.8932

Nous pouvons ensuite tracer les données transformées sur un plan 3D , avec la CP1 sur l'axe des x et la CP2 sur l'axe des y et CP3 sur l'axe des z.

iii.4. Conclusion

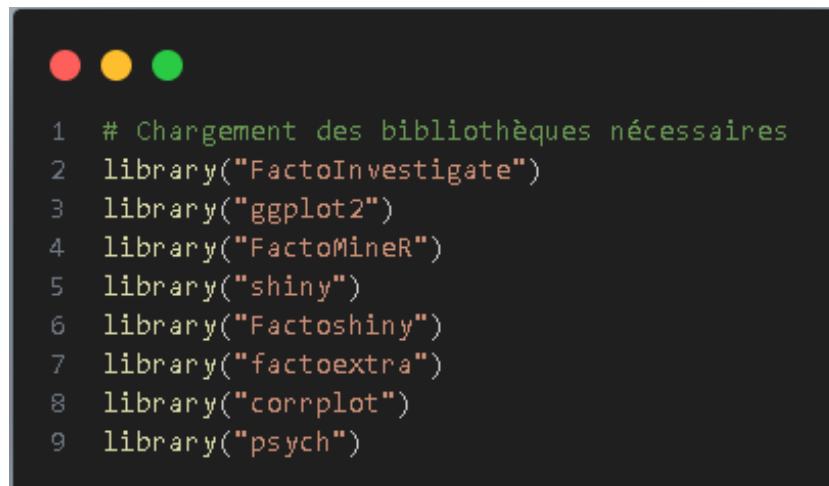
Ainsi, comme nous pouvons le constater, ce sont les étapes fondamentales de la méthode de l'Analyse en Composantes Principales (ACP). Ces étapes permettent de réduire le nombre de dimension , telles que 2D. Cette réduction est particulièrement intéressante car elle vous offre une représentation graphique de vos données, ce qui facilite leur analyse et l'extraction de leurs caractéristiques.

iv. Chapitre 3 : Application d'ACP sur données

iv.1. Introduction

Dans ce chapitre, nous mettrons en pratique les concepts abordés précédemment en utilisant les langages de programmation R et Python. Nous appliquerons ces outils de manière automatique à l'ensemble des données collectées lors du premier chapitre.

iv.2. Chargement des bibliothèques



```
1 # Chargement des bibliothèques nécessaires
2 library("FactoInvestigate")
3 library("ggplot2")
4 library("FactoMineR")
5 library("shiny")
6 library("Factoshiny")
7 library("factoextra")
8 library("corrplot")
9 library("psych")
```

FIGURE 7 – bibliothèques

ggplot2 : Une bibliothèque très populaire pour la création de graphiques et de visualisations de données.

FactoMineR : Une bibliothèque pour l'analyse exploratoire de données multidimensionnelles, y compris l'analyse en composantes principales (ACP).

shiny : Une bibliothèque qui permet de créer des applications web interactives en R.

Factoshiny : Il s'agit d'une extension de la bibliothèque FactoMineR qui permet d'intégrer des fonctionnalités interactives dans des applications Shiny.

factoextra : Une bibliothèque complémentaire à FactoMineR qui fournit des outils supplémentaires pour l'analyse exploratoire de données.

corrplot : Une bibliothèque pour la création de graphiques de matrices de corrélation.

psych : Une bibliothèque pour l'analyse statistique des données.

iv.3. Chargement des données

Le code ci-dessous permettra de charger simplement les données stockées sur notre ordinateur au format CSV :

header = TRUE : Indique que la première ligne du fichier CSV contient les noms des variables.

row.names = "Annees" : Utilise la colonne "Annees" comme noms de lignes dans le dataframe.

sep = ';' : Utilise le point-virgule ';' comme séparateur de champ dans le fichier CSV.

```

1 # Définir le répertoire de travail
2 setwd("C:/Users/Ahmed/Desktop/AnalyseDeDonnes/")
3
4 # Charger votre ensemble de données depuis un fichier CSV
5 data <- read.csv('donnees_tableau.csv', header = TRUE, row.names = "Annees", sep = ';')
6

```

FIGURE 8 – chargement des données

Consulter la dataset :

```

> data
      à 0. an..H.  à 1. an..H.  à 20. ans..H.  à 40. ans..H.  à 60. ans..H.  à 65. ans..H.  à 0. an..F.  à 1. an..F.  à 20. ans..F.  à 40. ans..F.  à 60. ans..F.  à 65. ans..F.
1946      59,9      64,4      48      30,8      15,4      12,2      65,2      68,9      52,2      34,7      18      14,3
1947      61,2      65,3      48,4      30,9      15,5      12,3      66,7      70      52,9      35,1      18,2      14,5
1948      62,7      65,9      48,5      30,9      15,6      12,3      68,8      71,2      53,6      35,6      18,7      15
1949      62,2      65,5      48,2      30,3      14,9      11,7      67,6      70,2      52,7      34,5      17,7      14
1950      63,4      66,2      48,7      30,7      15,4      12,2      69,2      71,4      53,6      35,2      18,4      14,6
1951      63,1      65,9      48,2      30,2      14,9      11,8      68,9      71      53,2      34,8      17,9      14,2
1952      64,4      66,8      49,1      30,9      15,5      12,3      70,2      72,1      54,2      35,6      18,6      14,8
1953      64,3      66,4      48,6      30,3      15      11,8      70,3      71,8      53,8      35,1      18,1      14,4
1954      65      67,1      49,2      31      15,5      12,4      71,2      72,8      54,7      35,9      18,9      15,1
1955      65,1      67,1      49,2      30,9      15,4      12,3      71,5      73      54,8      36      18,9      15,1
1956      65,2      66,9      48,9      30,6      15,2      12      71,7      72,9      54,7      35,9      18,7      14,9
1957      65,5      67      49,1      30,8      15,3      12,2      72,2      73,4      55,2      36,3      19      15,2
1958      66,8      68,3      50,2      31,8      16      12,8      73,2      74,3      56      37      19,5      15,6
1959      66,9      68,1      50,1      31,7      15,9      12,8      73,4      74,3      56      37      19,6      15,7
1960      67      68,1      50      31,5      15,7      12,6      73,6      74,4      56      37      19,5      15,6
1961      67,55      68,5      50,3      32      16,1      13      74,4      75,1      56,7      37,6      20,1      16,1
1962      67,1      68      49,9      31,5      15,7      12,5      73,9      74,6      56,2      37,1      19,6      15,7
1963      66,7      67,8      49,7      31,2      15,5      12,4      73,8      74,3      56,1      37,1      19,5      15,6
1964      67,7      68,5      50,3      31,9      16      12,9      74,8      75,2      57      37,9      20,3      16,3
1965      67,5      68,1      50      31,6      15,7      12,6      74,7      75      56,7      37,7      20,1      16,1
1966      67,9      68,5      50,3      31,9      16,1      12,9      75,2      75,5      57,2      38,1      20,5      16,5
1967      67,85      68,4      50,2      31,8      15,9      12,8      75,2      75,4      57,1      38      20,4      16,5
1968      67,8      68,3      50,2      31,7      15,8      12,7      75,2      75,5      57,2      38      20,4      16,4
1969      67,4      67,9      49,8      31,4      15,6      12,5      75,1      75,4      57      37,8      20,2      16,3
1970      68,4      68,8      50,7      32,3      16,2      13      75,9      76,1      57,6      38,5      20,8      16,8
1971      68,3      68,7      50,5      32,1      16,2      13      75,9      76,1      57,7      38,5      20,8      16,8
1972      68,5      68,7      50,6      32,3      16,4      13,1      76,2      76,3      57,9      38,7      21,1      17
1973      68,7      68,9      50,8      32,3      16,4      13,1      76,3      76,4      57,9      38,8      21      17
1974      68,9      69,1      50,9      32,5      16,5      13,3      76,7      76,7      58,3      39,1      21,3      17,2
1975      69      69,1      50,9      32,4      16,5      13,2      76,9      76,8      58,3      39,1      21,3      17,2

```

FIGURE 9 – Notre Dataset

iv.4. Transformation des données

Suite à l'exécution de la fonction `str(df)`, il a été observé que les données sont stockées sous forme de chaînes de caractères. Pour optimiser l'analyse et utiliser les fonctions statistiques correctement, il est recommandé de convertir ces données en type numérique à l'aide de la fonction `as.numeric()`. Cela permettra de s'assurer que les opérations numériques peuvent être effectuées correctement sur les données, garantissant ainsi l'exactitude des résultats obtenus lors de l'analyse des données.

```

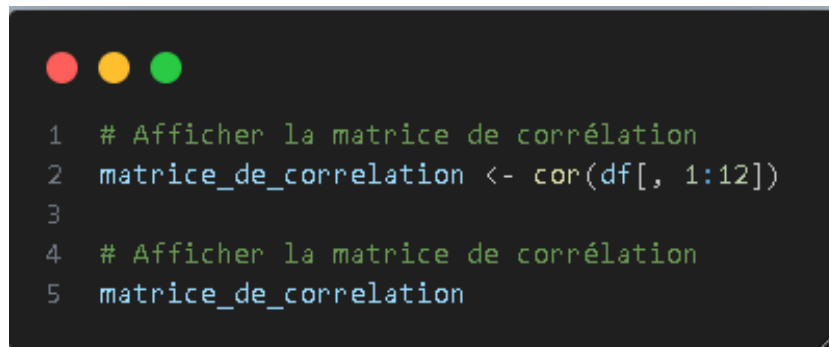
1 # Afficher la structure des données
2 str(df)
3
4 # Résoudre le problème de la structure des données (conversion des nombres au format string en numérique)
5 df[] <- lapply(df, function(x) as.numeric(gsub(",", ".", x)))
6

```

FIGURE 10 – transformation des données

iv.5. Application d'ACP

iv.5.1. Matrice de corrélation



```

1 # Afficher la matrice de corrélation
2 matrice_de_correlation <- cor(df[, 1:12])
3
4 # Afficher la matrice de corrélation
5 matrice_de_correlation

```

FIGURE 11 – Calcule du Matrice de corrélation

```

> print(matrice_de_correlation, width = 1000)
à.0.an..H. à.1.an..H. à.20.an..H. à.40.an..H. à.60.an..H. à.65.an..H. à.0.an..F. à.1.an..F. à.20.an..F. à.40.an..F. à.60.an..F. à.65.an..F.
à.0.an..H. 1.0000000 0.9904284 0.9819668 0.9759691 0.9718905 0.9722806 0.9863482 0.9943910 0.9944847 0.9914462 0.9902365 0.9900004
à.1.an..H. 0.9904284 1.0000000 0.9983777 0.9952789 0.9930507 0.9934150 0.9561107 0.973167 0.9708647 0.9823451 0.9860394 0.9880344
à.20.an..H. 0.9819668 0.9983777 1.0000000 0.9986018 0.9971325 0.9973089 0.9410731 0.9642035 0.9721140 0.9762829 0.9817632 0.9844862
à.40.an..H. 0.9759691 0.9952789 0.9986018 1.0000000 0.9992616 0.9993703 0.9347696 0.9600975 0.9686884 0.9760975 0.9823373 0.9852294
à.60.an..H. 0.9722806 0.9930507 0.9971325 0.9992616 1.0000000 0.9996026 0.9299512 0.9565185 0.9657226 0.9741411 0.9812672 0.9842861
à.65.an..H. 0.9722806 0.9934150 0.9973089 0.9993703 0.9996026 1.0000000 0.9300246 0.9564791 0.9655134 0.9737203 0.9807337 0.9838083
à.0.an..F. 0.9863482 0.9761107 0.9410731 0.9347696 0.9299512 0.9300246 1.0000000 0.9966094 0.9927322 0.9856750 0.9792989 0.9763672
à.1.an..F. 0.9944847 0.973167 0.9642035 0.9600975 0.9565185 0.9564791 0.9966094 1.0000000 0.9990440 0.9953444 0.9902602 0.9903157
à.20.an..F. 0.9914462 0.9708647 0.9721140 0.9686884 0.9657226 0.9655134 0.9927322 0.9990440 1.0000000 0.9985671 0.9962893 0.9950172
à.40.an..F. 0.9823451 0.9860394 0.9817632 0.9823373 0.9760975 0.9741411 0.9737203 0.9856750 0.9953444 0.9985671 1.0000000 0.9991445
à.60.an..F. 0.9852294 0.9842861 0.9838083 0.9823373 0.9812672 0.9807337 0.9792989 0.9763672 0.9902602 0.9962893 0.9950172 1.0000000
à.65.an..F. 0.9900004 0.9880344 0.9844862 0.9852294 0.9842861 0.9838083 0.9763672 0.9903157 0.9950172 0.9984186 0.9997801 1.0000000

```

FIGURE 12 – Valeur du Matrice de corrélation

Le test de Bartlett est une méthode statistique utilisée pour évaluer si les variables d'un ensemble de données sont corrélées entre elles. Il s'appuie sur deux hypothèses :

H0 (hypothèse nulle) : Toutes les corrélations entre les variables sont nulles, ce qui signifie que la matrice de corrélation est une matrice identité, où toutes les variables sont non corrélées.

H1 (hypothèse alternative) : Au moins une paire de variables a une corrélation différente de zéro.

Après avoir appliqué le test de Bartlett à nos données, nous avons obtenu une valeur de la statistique de test chi-carré de 5946.272 avec 66 degrés de liberté, et une valeur p extrêmement faible ($p < 0.05$). Ceci indique un rejet significatif de l'hypothèse nulle, ce qui suggère que les variables de notre ensemble de données présentent des corrélations significatives entre elles. Ainsi, nous concluons que les variables ne sont pas toutes indépendantes et qu'il existe des associations significatives entre elles.

```

> cortest.bartlett(matrice_de_correlation, n = 78)
$chisq
[1] 5946.272

$p.value
[1] 0

$df
[1] 66

```

FIGURE 13 – `cortest.bartlett(matrCorr, n = 78)`

iv.5.2. Choisir le nombre de composante principale

Pour déterminer le nombre optimal de dimensions à retenir dans notre analyse en composantes principales (PCA), nous avons utilisé la méthode des valeurs propres. Cette méthode consiste à examiner les valeurs propres extraites lors de l'analyse, qui représentent la quantité de variance expliquée par chaque composante principale. Nous avons affiché les valeurs propres pour évaluer la décroissance de leur importance.

En examinant les valeurs propres, nous avons identifié le point où il y avait un coude ou une cassure dans la décroissance des valeurs propres. Ce point a été considéré comme le nombre optimal de dimensions à retenir. En d'autres termes, nous avons sélectionné le nombre de dimensions qui capturerait le plus efficacement la variance des données tout en évitant le sur-ajustement.

```

> print(valeur_propre)

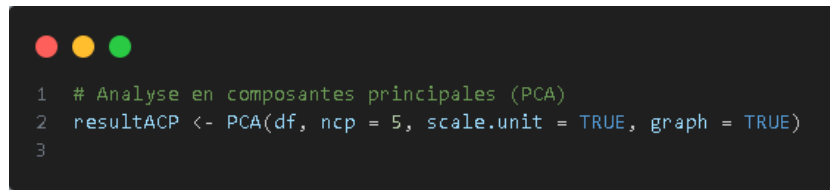
```

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.179738e+01	9.831146e+01	98.31146
Dim.2	1.771205e-01	1.476004e+00	99.78747
Dim.3	2.332802e-02	1.944001e-01	99.98187
Dim.4	8.602749e-04	7.168957e-03	99.98903
Dim.5	5.231360e-04	4.359466e-03	99.99339
Dim.6	2.259364e-04	1.882804e-03	99.99528
Dim.7	1.553623e-04	1.294686e-03	99.99657
Dim.8	1.358354e-04	1.131962e-03	99.99770
Dim.9	1.124215e-04	9.368458e-04	99.99864
Dim.10	7.855184e-05	6.545986e-04	99.99930
Dim.11	4.550441e-05	3.792034e-04	99.99967
Dim.12	3.909458e-05	3.257882e-04	100.00000

FIGURE 14 – `valeur_propre=get_eigenvalue(resultACP)`

Pour obtenir une représentation concise des données tout en préservant une proportion significative de la variance totale, nous avons choisi de retenir deux dimensions principales dans notre analyse en composantes principales (PCA). En examinant les valeurs propres, nous avons observé que les deux premières dimensions expliquent ensemble 99.79% (>80%) de la variance totale des données, ce qui indique une représentation robuste des données en deux dimensions. Cette approche nous permet de conserver une grande partie de la structure et de la variabilité des données tout en simplifiant leur représentation pour une interprétation plus aisée.

iv.5.3. Application de cpa



```

1 # Analyse en composantes principales (PCA)
2 resultACP <- PCA(df, ncp = 5, scale.unit = TRUE, graph = TRUE)
3

```

FIGURE 15 – Commande PCA

df : Le dataframe contenant les données à analyser.

ncp = 5 : Le nombre de composantes principales à extraire.

scale.unit = TRUE : Indique si les variables doivent être mises à l'échelle.

graph = TRUE : Indique si des graphiques doivent être produits pour visualiser les résultats.

```

Variance          Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6 Dim.7 Dim.8 Dim.9
% of var.         11.797 0.177 0.023 0.001 0.001 0.000 0.000 0.000 0.000
Cumulative % of var. 98.311 99.787 99.982 99.989 99.993 99.995 99.997 99.998 99.999
Variance          Dim.10 Dim.11 Dim.12
% of var.         0.000 0.000 0.000
Cumulative % of var. 99.999 100.000 100.000

Individuals (the 10 first)
      Dist Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2
1946 | 6.006 | -5.818 3.678 0.938 | -1.352 13.234 0.051 | -0.619 21.038 0.011 |
1947 | 5.497 | -5.363 3.126 0.952 | -1.129 9.223 0.042 | -0.420 9.699 0.006 |
1948 | 4.910 | -4.836 2.541 0.970 | -0.788 4.500 0.026 | -0.315 5.468 0.004 |
1949 | 5.641 | -5.579 3.382 0.978 | -0.837 5.069 0.022 | -0.048 0.129 0.000 |
1950 | 4.944 | -4.895 2.604 0.980 | -0.689 3.441 0.019 | -0.045 0.110 0.000 |
1951 | 5.320 | -5.286 3.036 0.987 | -0.599 2.600 0.013 | 0.043 0.102 0.000 |
1952 | 4.562 | -4.527 2.227 0.985 | -0.554 2.223 0.015 | 0.075 0.307 0.000 |
1953 | 4.933 | -4.914 2.624 0.992 | -0.390 1.101 0.006 | 0.201 2.215 0.002 |
1954 | 4.256 | -4.237 1.951 0.991 | -0.379 1.038 0.008 | 0.098 0.528 0.001 |
1955 | 4.230 | -4.218 1.934 0.994 | -0.286 0.592 0.005 | 0.119 0.781 0.001 |

Variables (the 10 first)
      Dim.1 ctr cos2 Dim.2 ctr cos2 Dim.3 ctr cos2
à.0.an..H. | 0.995 8.393 0.990 | 0.061 2.092 0.004 | 0.078 26.050 0.006 |
à.1.an..H. | 0.995 8.391 0.990 | -0.076 3.263 0.006 | 0.065 17.851 0.004 |
à.20.ans..H. | 0.991 8.327 0.982 | -0.125 8.808 0.016 | 0.043 7.857 0.002 |
à.40.ans..H. | 0.990 8.303 0.980 | -0.142 11.348 0.020 | -0.003 0.033 0.000 |
à.60.ans..H. | 0.988 8.271 0.976 | -0.153 13.239 0.023 | -0.022 2.150 0.001 |
à.65.ans..H. | 0.988 8.270 0.976 | -0.154 13.349 0.024 | -0.016 1.104 0.000 |
à.0.an..F. | 0.976 8.069 0.952 | 0.217 26.607 0.047 | 0.029 3.638 0.001 |
à.1.an..F. | 0.990 8.309 0.980 | 0.140 10.998 0.019 | 0.009 0.385 0.000 |
à.20.ans..F. | 0.994 8.376 0.988 | 0.108 6.535 0.012 | -0.015 0.939 0.000 |
à.40.ans..F. | 0.996 8.410 0.992 | 0.071 2.831 0.005 | -0.051 11.308 0.003 |
>

```

FIGURE 16 – Interpretation de resultat ACP

iv.6. Implementation avec Python

Voici le code que nous avons écrit :

```
1 import pandas as pd
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.decomposition import PCA
4 import matplotlib.pyplot as plt
5
6 data = pd.read_csv('AnalyseDeDonnes/donnees_tableau.csv', header=0, index_col="Annees", sep=';')
7
8 df = data.iloc[0:78, 0:12]
9
10 df = df.apply(lambda x: pd.to_numeric(x.astype(str).str.replace(',', '.'), errors='coerce'))
11
12 scaler = StandardScaler()
13 df_standardized = scaler.fit_transform(df)
14
15 pca = PCA(n_components=2)
16 pca_results = pca.fit_transform(df_standardized)
17
18 print("Composantes principales:")
19 print(pca_results)
20
21 print("Variance expliquée par chaque composante principale:")
22 print(pca.explained_variance_ratio_)
23
24 print("Pourcentage de variance totale expliquée:")
25 print(sum(pca.explained_variance_ratio_) * 100)
26
```

FIGURE 17 – Implementation ACP en Python

Importation des bibliothèques nécessaires : Les bibliothèques pandas, scikit-learn (pour le prétraitement des données, la PCA et la standardisation) et matplotlib (pour la visualisation) sont importées. Chargement des données :

Les données sont chargées à partir du fichier CSV spécifié (donnees_tableau.csv) en utilisant la fonction `pd.read_csv()`. Les données sont stockées dans un DataFrame pandas. Sélection des colonnes et lignes pertinentes :

Seules les colonnes et les lignes pertinentes de l'ensemble de données sont sélectionnées pour l'analyse. Dans ce cas, les 78 premières lignes et les 12 premières colonnes sont sélectionnées. Prétraitement des données :

Les données sont prétraitées pour résoudre les problèmes de format. La fonction `apply()` est utilisée avec une lambda fonction pour convertir les valeurs au format string en nombres numériques. Cela est accompli en remplaçant les virgules (,) par des points (.) dans les valeurs et en convertissant les valeurs en nombre. Standardisation des données :

Les données sont standardisées en utilisant la classe `StandardScaler` de scikit-learn. Cela permet de centrer et de mettre à l'échelle les données, en les transformant de sorte que leur moyenne soit égale à zéro et leur variance soit égale à un. Analyse en Composantes Principales (ACP) :

La classe `PCA` de scikit-learn est utilisée pour effectuer l'Analyse en Composantes Principales. Dans ce cas, nous spécifions `n_components=2` pour indiquer que nous voulons réduire les dimensions à deux composantes principales. La méthode `fit_transform()` est utilisée pour ajuster le modèle PCA aux données et transformer les données dans l'espace des composantes principales. Affichage des résultats :

Les résultats de l'ACP sont affichés, y compris les composantes principales (`pca_results`), la variance expliquée par chaque composante principale (`explained_variance_ratio_`) et le pourcentage de variance totale expliquée.

```
C:\Users\Ahmed\Desktop\ProjetAD.py:1: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

import pandas as pd
Composantes principales:
[[5.8179613  1.3521663 ]
 [5.36318383  1.12880348]
 [4.83587763  0.78846341]
 [5.57852943  0.83680167]
 [4.89481959  0.68949753]
 [5.28589217  0.59932134]
 [4.52785653  0.95444533]
 [4.91368912  0.38999223]
 [4.23746795  0.3787113 ]
 [4.2181889  0.28591884]]
Variance expliquée par chaque composante principale:
[0.98311462 0.01476804]
Pourcentage de variance totale expliquée:
99.7874655179629
```

FIGURE 18 – Resultat d'analyse en python

iv.7. Comparaison des résultats R/Python

R :

Les composantes principales ont été générées à l'aide de la bibliothèque FactoMineR. Les coordonnées des individus sur les deux premières dimensions ont été extraites et affichées sous forme de tableau. La variance expliquée par chaque dimension ainsi que le pourcentage de variance totale expliquée ont été calculés.

Python :

L'ACP a été réalisée en utilisant les bibliothèques pandas, scikit-learn et matplotlib. Les résultats comprennent les coordonnées des individus sur les deux premières dimensions ainsi que la variance expliquée. Des graphiques de dispersion ont été générés pour visualiser la répartition des individus sur les deux dimensions.

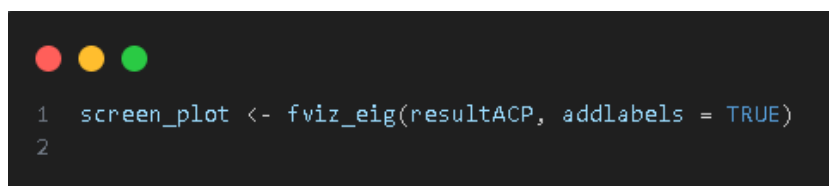
En conclusion, cette comparaison des résultats de l'ACP entre R et Python met en évidence la cohérence des analyses réalisées dans les deux environnements. Ces résultats renforcent la confiance dans les conclusions tirées de notre étude et soulignent l'importance de vérifier la robustesse des analyses à travers différents outils et environnements de programmation.

v. Chapitre 4 : Visualisation et Interprétation

v.1. Introduction

Dans ce chapitre, nous explorerons l'importance de la visualisation et de l'interprétation des données dans le processus d'analyse. La visualisation des données est un outil essentiel pour comprendre la structure sous-jacente des données, identifier les tendances, les schémas et les anomalies. De plus, une interprétation précise des résultats est nécessaire pour tirer des conclusions significatives et prendre des décisions éclairées. Nous examinerons diverses techniques de visualisation de données, telles que les graphiques, les tableaux de bord interactifs et les cartes, ainsi que des méthodes d'interprétation pour analyser et comprendre les résultats obtenus à partir des analyses statistiques. En combinant la visualisation et l'interprétation des données, nous serons en mesure d'extraire des insights significatifs et de communiquer efficacement nos résultats aux parties prenantes. Ce chapitre constituera une étape cruciale dans notre parcours d'analyse des données, en nous permettant de transformer les données brutes en informations exploitables et en connaissances exploitables.

v.2. Quantité d'informations expliquée par chaque composant

A screenshot of a terminal window with a dark background and three colored window control buttons (red, yellow, green) in the top left corner. The terminal displays two lines of R code: line 1: `screen_plot <- fviz_eig(resultACP, addlabels = TRUE)` and line 2: .

```
1 screen_plot <- fviz_eig(resultACP, addlabels = TRUE)
2
```

FIGURE 19 – Commande screen plot

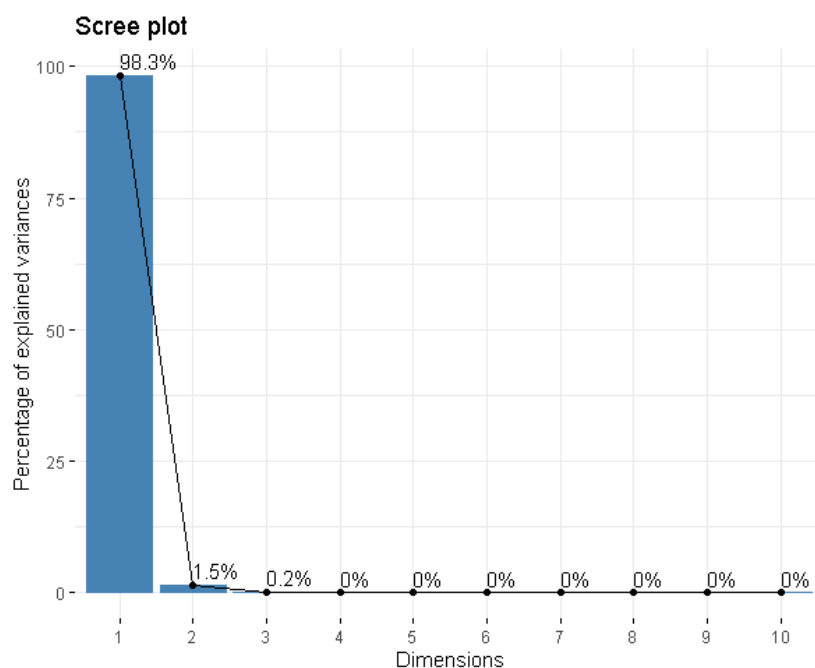


FIGURE 20 – resultat de screen plot

Le premier axe (CP1) est toujours très important et explique la majorité de la variance. Le deuxième axe (CP2) pourrait expliquer une partie significative de la variance restante. Il est possible que CP2 capture des patterns importants dans les données qui ne sont pas représentés par CP1. Les axes suivants (CP3 et plus) ne sont probablement pas significatifs et peuvent être ignorés.

v.3. Corrélation des variables

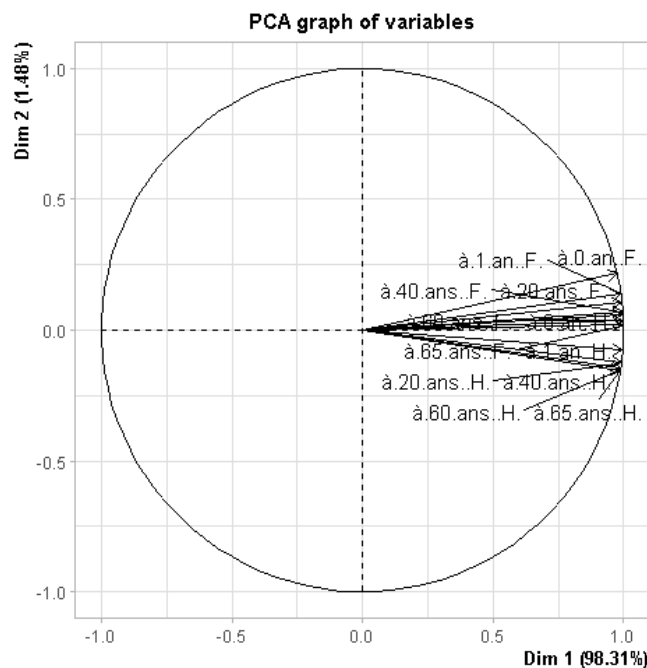


FIGURE 21 – Corrélation des variables

La dimension 1 oppose des individus tels que 2013, 2014, 2016, 2018, 2015, 2023, 2019, 2017, 2021 et 2022 (à droite du graphe, caractérisés par une coordonnée fortement positive sur l'axe) à des individus comme 1951, 1949, 1950, 1952, 1953, 1948 et 1947 (à gauche du graphe, caractérisés par une coordonnée fortement négative sur l'axe).

Le groupe auquel les individus 2013, 2014, 2016, 2018, 2015, 2023, 2019, 2017, 2021 et 2022 appartiennent (caractérisés par une coordonnée positive sur l'axe) partage :

de fortes valeurs pour des variables telles que à.60.an..H., à.65.an..H., à.40.an..H., à.20.an..H., à.1.an..H., à.65.an..F., à.60.an..F., à.40.an..F., à.0.an..H. et à.20.an..F. (de la plus extrême à la moins extrême). Le groupe auquel les individus 1951, 1949, 1950, 1952, 1953, 1948 et 1947 appartiennent (caractérisés par une coordonnées négative sur l'axe) partage :

de faibles valeurs pour des variables telles que à.0.an..F., à.1.an..F., à.20.an..F., à.0.an..H., à.40.an..F., à.60.an..F., à.65.an..F., à.1.an..H., à.20.an..H. et à.40.an..H. (de la plus extrême à la moins extrême). Le groupe 3 (caractérisés par une coordonnées négative sur l'axe) partage :

de faibles valeurs pour des variables telles que à.60.an..H., à.65.an..H., à.40.an..H., à.20.an..H., à.1.an..H., à.65.an..F., à.60.an..F., à.40.an..F., à.0.an..H. et à.20.an..F. (de la plus extrême à la moins extrême). Notons que les variables à.0.an..H., à.1.an..H., à.20.an..H., à.40.an..H., à.60.an..H., à.65.an..H., à.0.an..F., à.1.an..F., à.20.an..F. et à.40.an..F. sont extrêmement corrélées à cette dimension (corrélations respectives de

0.99, 0.99, 0.98, 0.98, 0.98, 0.98, 0.95, 0.98, 0.99, 0.99, 1, 1). Ces variables pourraient donc résumer à elles seules la dimension 1.

v.4. Contribution des variables/individus dans chaque CP

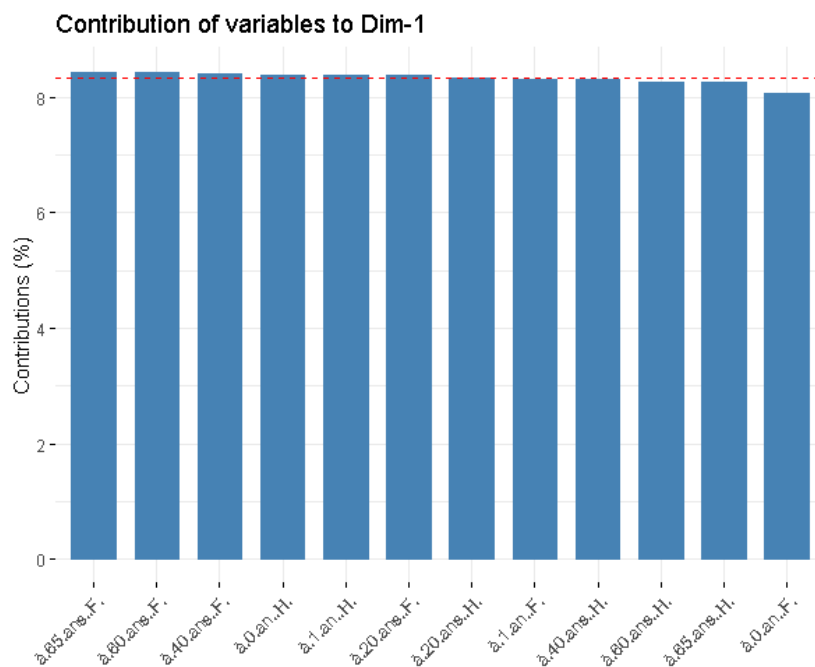


FIGURE 22 – Histogramme de la Contribution des variables à la dimension 1

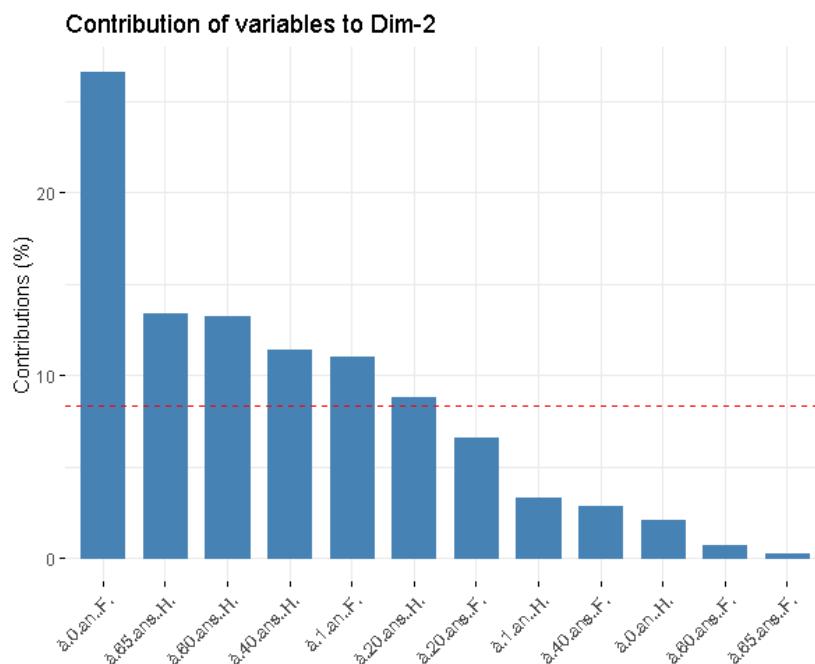


FIGURE 23 – Histogramme de la Contribution des variables à la dimension 2

```

1  fviz_contrib(resultACP, choice = "var", axes = 1, top = 13)
2  fviz_contrib(resultACP, choice = "var", axes = 2, top = 13)
3

```

FIGURE 24 – Commande de Contribution

En observant ces graphiques, nous pouvons évaluer l'impact de chaque variable sur contribution à la création de chaque dimension. Dans notre cas, nous nous concentrons sur seulement deux dimensions. Il est clair que toutes les variables contribuent à créer la première dimension, à l'exception de certaines variables, qui ont besoin aussi de la deuxième dimension.

v.5. Qualité de représentation des variables/individus dans chaque CP

Première composante (Dmin1) : Toutes les variables semblent contribuer significativement à la première composante (Dmin1). Cela signifie que l'âge et le genre ont un impact sur la variation observée dans les données le long de cette dimension. Deuxième composante (Dmin2) : La deuxième composante (Dmin2) semble être principalement influencée par les groupes d'âge masculins (à 0 an.H, à 1 an.H, etc.). Les autres groupes

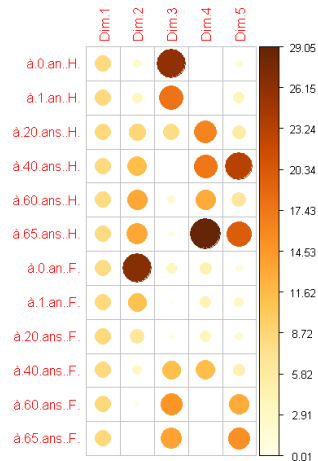


FIGURE 25 – Corréléogramme

d'âge féminins (à 0 an.F, à 1 an.F, etc.) ont des valeurs plus faibles dans cette dimension. Troisième et quatrième composantes (Dmin3 et Dmin4) : Les valeurs continuent de diminuer pour Dmin3 et Dmin4. Ces dimensions supplémentaires peuvent expliquer davantage de variation dans les données.

v.6. Interprétation par Biplot

un biplot est une représentation graphique qui affiche à la fois les individus et les variables sur un même graphique. Les individus sont représentés par des points dans l'espace, tandis que les variables sont représentées par des vecteurs partant de l'origine. La direction et la longueur de chaque vecteur indiquent la contribution de la variable à la composante principale correspondante. Ainsi, la position des points par rapport aux vecteurs reflète les relations entre les individus et les variables, offrant ainsi une visualisation intuitive des structures et des corrélations dans les données.

Pour afficher, on utilise l'instruction :

```

1  fviz_pca_biplot(resultACP,
2                      geom.ind = "point",
3                      axes = c(1, 2),
4                      pointshape = 21,
5                      pointsize = 1,
6                      alpha.var = "contrib",
7                      col.var = "cos2",
8                      gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
9                      repel = TRUE
10 )

```

FIGURE 26 – Commande pour afficher le biplot

resultat :

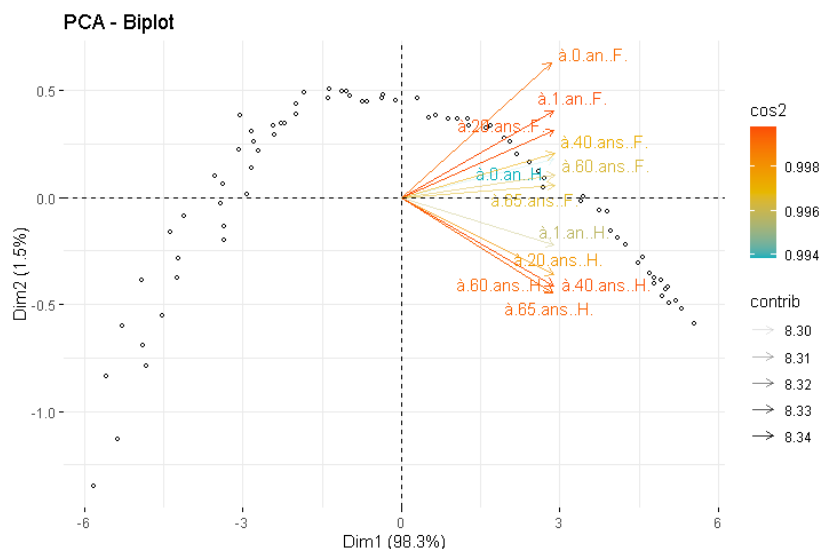


FIGURE 27 – Biplot de la dimension 1 et 2

On peut constater que la majorité des variables ont une bonne qualité de représentation, à l'exception de (A 0 an (H)), qui présente une mauvaise qualité de représentation.

Sur la dimension 1, les individus se répartissent entre ceux dont la coordonnée est fortement positive sur l'axe (à droite du graphe) et ceux dont la coordonnée est fortement négative sur l'axe (à gauche du graphe).

Le groupe 1, caractérisé par une coordonnée positive sur l'axe, présente des valeurs élevées pour toutes les variables.

Le groupe 2, caractérisé par une coordonnée négative sur l'axe, présente des valeurs faibles pour toutes les variables.

La dimension 2 oppose des individus caractérisés par une coordonnée fortement positive sur l'axe (en haut du graphe) à des individus caractérisés par une coordonnée fortement négative sur l'axe (en bas du graphe).

Le groupe 1 (caractérisés par une coordonnée positive sur l'axe) partage : de fortes valeurs pour les variables (A 0an F), (A 1an F), (A 20ans F), (A 40ans F), (A 60ans), (A 0an H), (A 65an F).

Le groupe 2 (caractérisés par une coordonnées négative sur l'axe) partage : de fortes valeurs pour les variables (A 1an H), (A 20ans H), (A 40ans H), (A 60ans H), (A 65ans H).

v.7. Résumé

La tendance générale indique une augmentation de l'espérance de vie au fil des années pour les deux sexes et pour toutes les tranches d'âge. Les hommes présentent généralement une espérance de vie légèrement inférieure à celle des femmes dans toutes les tranches d'âge, mais cette disparité tend à diminuer au fil du temps. Les femmes, quant à elles, ont une espérance de vie plus élevée que les hommes dans toutes les tranches d'âge, avec une stabilité et une croissance plus régulières au fil des années. Globalement, l'espérance de vie a progressé de manière constante pour les deux sexes et pour toutes les tranches d'âge au cours des années. Bien que l'espérance de vie tende à augmenter avec l'âge de manière générale, cette tendance peut varier légèrement d'une année à l'autre. Enfin, les femmes présentent une longévité supérieure à celle des hommes dans toutes les tranches d'âge, bien que cette différence puisse fluctuer légèrement au fil du temps.

vi. Conclusion

La méthode ACP (Analyse en Composantes Principales) est un outil puissant et polyvalent utilisé en statistiques multivariées pour explorer et analyser des données complexes. Grâce à cette méthode, nous pouvons réduire la dimensionnalité de nos données tout en préservant au mieux l'information contenue dans celles-ci. En résumé, l'ACP permet de transformer un ensemble de variables corrélées en un ensemble de variables non corrélées, appelées composantes principales, qui capturent l'essentiel de la variabilité des données.

En conclusion, la méthode ACP est un outil essentiel pour l'analyse exploratoire des données, offrant une approche robuste et efficace pour réduire la complexité des données, identifier des structures sous-jacentes et extraire des informations pertinentes pour la prise de décision. Son utilisation judicieuse peut conduire à une meilleure compréhension des phénomènes étudiés et à des insights précieux pour améliorer la performance et la compréhension dans divers domaines d'application.

vii. Référence

- [1] L’Institut national de la statistique et des études économiques (INSEE) :
<https://www.insee.fr/fr/statistiques/7746166?sommaire=7746197>
- [2] INC, T. E. (s. d.). Step-By-Step Guide to Principal Component Analysis With Example. <https://www.turing.com/kb/guide-to-principal-component-analysis>
- [3] RICHARDSON, L. (2007). Beautiful Soup documentation. April.