INSEA

PROJET
ANALYSE DE DONNÉES

APPLICATION DE L'ACP SUR LES DONNÉES DES PRODUITS DE LA PECHE MARINE

Réaliser par :

ZENIBI Abdelhakim
(M2SI)

Encadré par :

Mme. EL HANNOUN Wafaa

2022/2023

Table des matières

Liste des figures

Introduction

La pêche côtière et artisanale joue un rôle vital dans l'économie du Maroc, en contribuant de manière significative à la sécurité alimentaire et à la subsistance de nombreuses communautés côtières. Cependant, la gestion efficace de cette ressource précieuse nécessite une compréhension approfondie des tendances, des schémas et des facteurs qui influencent la commercialisation des produits de la pêche.

Le présent rapport se penche sur une analyse approfondie des données relatives au volume des produits commercialisés issus de la pêche côtière et artisanale au Maroc, exprimés en tonnes. Cette analyse repose sur l'Application de l'Analyse en Composantes Principales (ACP), une technique statistique puissante qui permet de réduire la dimensionnalité des données tout en préservant les informations essentielles.

L'objectif principal de cette étude est de découvrir des schémas cachés dans les données, d'identifier des tendances significatives, et de mettre en lumière les relations complexes entre les différentes variables liées à la commercialisation des produits de la pêche. En fin de compte, nous visons à fournir des informations précieuses qui peuvent éclairer les décisions de gestion, soutenir le développement durable de la pêche côtière et artisanale, et contribuer à la prospérité économique du Maroc.

Ce rapport commence par une description du processus de prétraitement des données, suivi d'une explication détaillée de l'ACP appliquée aux données de volume des produits de la pêche. Nous explorerons également les résultats de l'ACP à travers diverses visualisations et interprétations. Enfin, nous conclurons en mettant en évidence les constatations clés de cette analyse et en fournissant des recommandations pour une gestion plus efficace de la ressource halieutique au Maroc.

Chapitre 1 : Analyse en Composantes Principales (ACP)

I . Introduction

L'analyse en composantes principales (ACP) est une technique statistique qui permet de réduire la complexité des données tout en préservant les informations essentielles. Elle consiste en plusieurs étapes, notamment la centralisation des données, le calcul de la covariance entre les variables, la détermination des vecteurs propres et des valeurs propres, et enfin la projection des données dans un espace de dimension réduite. L'objectif principal est de simplifier la visualisation et l'analyse des données tout en maintenant la variation la plus significative. L'ACP est largement utilisée dans divers domaines pour explorer, résumer et interpréter de grands ensembles de données.

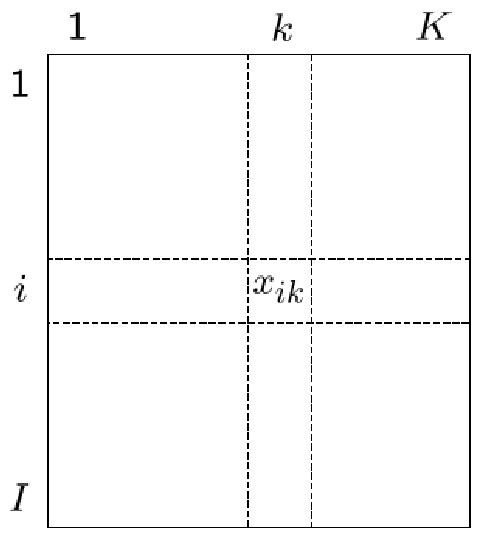
II . Comment fonctionne l'ACP

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

1. Standardiser de 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

$$x_{ik} \rightarrow \frac{x_{ik}-\overline{X}_k}{S_k}$$



Avec :

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

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

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

n : nombre de valeurs dans l'ensemble de données

2. Calculer la matrice de covariance

Pour séparer les variables fortement interdépendantes, on doit calculer la matrice de covariance. Une matrice de covariance est une matrice symétrique N x N qui contient les covariances de tous les ensembles de données possibles. Elle est donnée comme suit :

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

Avec :

$Y_1,\ldots,Y_k$: sont des variable

$Cov(Y_j, Y_p) = \overline{Y_j Y_p} - \overline{Y_j}\,\overline{Y_p}$: La covariance entre la variable Y_j et Y_p

3. Calculez le vecteur de caractéristiques

Pour déterminer le vecteur de caractéristiques, vous devez définir les valeurs propres et les vecteurs propres de la matrice de covariance. Les vecteurs propres seront les composantes principales et les valeurs propres décriront la quantité de variance expliquée par chaque composante principale. Tous simplement on choisit les composantes principales (vecteurs propres) qui sont associés à des variances (valeurs propres) grands où leurs sommes présente au moins 80%. On commence par les valeurs propres λ_i qui sont calculées par l'équation suivante :

$$|M - \lambda I| = 0$$

Où M est la matrice de covariance et I la matrice Identité. Ensuite, on calcule le vecteur propre ω_i associé à chaque valeur propre λ_i avec l'équation suivante :

$$M\lambda_i = M\omega_i$$

Enfin, nous trions en ordre décroissant les valeurs propres et leurs vecteurs propres correspondants, nos crions le vecteur de caractéristiques qui contient les premiers vecteurs propres qui la somme de leurs valeurs propres présente au moins 80% et nous ignorons le reste ?

4. Multipliez les données normalisées par les vecteurs propres

L'objectif de l'ACP est de réexprimer l'ensemble de données d'origine, et maintenant nous sommes enfin prêts à franchir cette étape et à générer les "composantes principales" réelles. Nous multiplions simplement notre jeu de données standardisé à l'étape 1 par le vecteur de caractéristiques que nous avons générée à l'étape 3.

Exemple

Pour comprendre PCA, utilisons un exemple avec des données sur trois personnes et leurs notes (de 0 à 20) dans trois matières : Math, Physique et Français.

Etudiant	Math	Physique	Français
Jean	19	16	8
Aline	18.5	15	9
Annie	18	14	7.5

Donc ces données peuvent en fait être présentées comme une matrice à 3 dimensions puisque nous ne nous intéressons qu'aux colonnes quantitatives

$$A = \begin{pmatrix} 19 & 16 & 8 \\ 18.5 & 15 & 9 \\ 18 & 14 & 7.5 \end{pmatrix}$$

• Centrer et réduire la matrice des données

Nous calculons la moyenne et l'écart type pour chaque variable

	Math	Science	Français
Moyenne	18.5	15	8.16
Ecart-type	0.40	0.81	0.62

Pour résoudre ce problème, on choisit de transformer les données en données centrées-réduites.
L'observation X_{ik} est alors remplacée par :

$$\frac{X_{ik} - \overline{X}_k}{S_k}$$

Ou :

$\overline{X}_k$: moyenne de la variable X_k

S_k : écart-type de la variable X_k

Ensuite, après la normalisation de chaque variable, voici les résultats sous forme matrice ci-dessous :

$$A = \begin{pmatrix} 1.25 & 1.23 & -0.25 \\ 0 & 0 & 1.35 \\ -1.25 & -1.23 & -1.06 \end{pmatrix}$$

• Calculer la matrice de covariance de données

Ainsi, nous pouvons calculer la covariance de deux variables X et Y en utilisant la formule suivante :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

À l'aide de la formule ci-dessus, nous pouvons trouver la matrice de covariance de A. De plus, le résultat serait une matrice carrée de 3 × 3 dimensions. Sa matrice de covariance serait :

$$\mathrm{Cov(X,\ Y)_A} = \begin{pmatrix} \mathrm{Var(math)} & \mathrm{Cov(math\,,physique)} & \mathrm{Cov(math\,,français)} \\ \mathrm{Cov(physique\,,math)} & \mathrm{Var(physique)} & \mathrm{Cov(physique\,,français)} \\ \mathrm{Cov(français\,,math)} & \mathrm{Cov(français\,,physique)} & \mathrm{Var(français)} \end{pmatrix}$$

$$\mathrm{Cov(X,\ Y)_A} = \begin{pmatrix} 1 & 1.02 & 0.33 \\ 1.02 & 1 & 0.33 \\ 0.33 & 0.33 & 1 \end{pmatrix}$$

• Calculer les vecteurs propres et les valeurs propres correspondantes

Les valeurs propres sont associées à des vecteurs propres en algèbre linéaire. Ces deux termes sont utilisés dans l'analyse des transformations linéaires. Les valeurs propres sont l'ensemble spécial des valeurs scalaires associées à l'ensemble d'équations linéaires très probablement dans les équations matricielles. Les vecteurs propres sont également appelés racines caractéristiques.

C'est un vecteur non nul qui peut être modifié au maximum de son facteur scalaire après l'application de transformations linéaires. Et le facteur correspondant qui met à l'échelle le vecteur propre est appelé une valeur propre.

En d'autres termes :

Soit A une matrice carrée, ν un vecteur et λ un scalaire qui satisfait

$$A\nu = \lambda\nu$$

Alors λ est appelée la valeur propre associée au vecteur propre ν de A. Les valeurs propres de A sont les racines de l'équation caractéristique.

$$det(A - \lambda I) = 0$$

En calculant d'abord det(A-λI), I est une matrice identité

$$det\left(\begin{bmatrix} 1 & 1.02 & 0.33 \\ 1.02 & 1 & 0.33 \\ 0.33 & 0.33 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)$$

En simplifiant d'abord la matrice, on pourra calculer le déterminant par la suite

$$\begin{bmatrix} 1 & 1.02 & 0.33 \\ 1.02 & 1 & 0.33 \\ 0.33 & 0.33 & 1 \end{bmatrix} - \lambda \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 1.02 & 0.33 \\ 1.02 & 1-\lambda & 0.33 \\ 0.33 & 0.33 & 1-\lambda \end{bmatrix}$$

Maintenant que nous avons notre matrice simplifiée, nous pouvons trouver le déterminant de la même :

$$det\left(\begin{bmatrix} 1-\lambda & 1.02 & 0.33 \\ 1.02 & 1-\lambda & 0.33 \\ 0.33 & 0.33 & 1-\lambda \end{bmatrix}\right)$$

Nous avons maintenant l'équation et nous avons besoin de résoudre pour λ, de manière à obtenir la valeur propre de la matrice. Ainsi, l'équation ci-dessus à zéro :

$$-\lambda^3 + 3\,\lambda^2 - 1.74\,\lambda - 0.03 = 0$$

Après avoir résolu cette équation pour la valeur de λ, nous obtenons la valeur suivante :

$$\lambda_1 = 2.20,\ \lambda_2 = 0.81\ \lambda_3 = -0.02.$$

Ainsi, après avoir résolu les vecteurs propres, nous obtiendrions la solution suivante pour les valeurs propres correspondantes

$$V_1 = (1{,}82,\ 1{,}82,\ 1),\ V_2 = (-0{,}27,\ -0{,}27{,}1),\ V_3 = (\ -1,\ 1,\ 0).$$

- ## Choisir k vecteurs de caractéristiques

Nous avons commencé par l'objectif de réduire la dimensionnalité de notre espace de caractéristiques, c'est-à-dire de projeter l'espace de caractéristiques via ACP sur un sous-espace plus petit, où les vecteurs propres formeront les axes de ce nouveau sous-espace de caractéristiques. Cependant, les vecteurs propres ne définissent que les directions du nouvel axe, puisqu'ils ont tous la même longueur unitaire 1.

Ainsi, pour décider quel(s) vecteur(s) propre(s) nous voulons laisser tomber pour notre sous-espace de dimension inférieure, nous devons examiner les valeurs propres correspondantes des vecteurs propres. Grosso modo, les vecteurs propres avec les valeurs propres les plus faibles portent le moins d'informations sur la distribution des données, et ce sont ceux que nous voulons laisser tomber.

L'approche courante consiste à classer les vecteurs propres de la valeur propre correspondante la plus élevée à la plus basse et à choisir les k vecteurs propres supérieurs. Ainsi, après avoir trié les valeurs propres par ordre décroissant, on a :

Valeur propre	Vecteur propre	Variance (%)
$\lambda_1 = 2.20$	$V_1 = (1{,}82, 1{,}82, 1)$	72.60 %
$\lambda_2 = 0.81$	$V_2 = (-0{,}27, -0{,}27, 1)$	26.73 %
$\lambda_3 = -0.02$	$V_3 = (-1, 1, 0)$	0.66 %

Maintenant, on choisit le pourcentage de variance des valeurs propres les plus élevées. Ce sont ses composants principaux :

$$V_1 = (1{,}82, 1{,}82, 1), \quad V_2 = (-0{,}27, -0{,}27, 1).$$

- **Reformuler les données sur les principaux axes des composantes**

Vous venez de sélectionner les composants principaux et de former un vecteur de caractéristiques. Pourtant, les données initiales restent les mêmes sur leurs axes d'origine. Cette étape vise à réorienter les données de leurs axes d'origine vers ceux que vous avez calculés

À partir des composantes principales. Cela peut être fait par la formule suivante :

Données centrées- réduites * Vecteur caractéristique = Données final

Données centrées réduites *	Vecteur caractéristique =	Données final
$\begin{pmatrix} 1.25 & 1.23 & -0.25 \\ 0 & 0 & 1.35 \\ -1.25 & -1.23 & -1.06 \end{pmatrix}$	$\begin{pmatrix} 1.82 & -0.27 \\ 1.82 & -0.27 \\ 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 4.26 & -0.91 \\ 1.35 & 1.35 \\ -5.57 & -0.39 \end{pmatrix}$

III .Conclusion

Donc, comme nous le voyons, ce sont les étapes moyennes de la méthode ACP, qui peuvent réduire l'ensemble de données de n dimensions à de petites dimensions telles que 2D, et c'est tellement intéressant car cela vous donne un graphique illustré de vos données, puis vous pouvez l'analyser et extraire ses caractéristiques.

Chapitre 2 : Prétraitement des Données

I. Introduction

Le prétraitement des données est une étape cruciale dans tout projet d'analyse de données, car il constitue la base pour obtenir des informations précises et significatives. Cette section décrit les étapes entreprises pour préparer le jeu de données en vue d'une analyse ultérieure, y compris le chargement, le nettoyage et la transformation des données.

II. Source et caractéristiques des données

Le jeu de données sur lesquelles on a appliqué l'ACP dans ce projet ont comme source la nouvelle Platform hcp.ma Elles ont comme titre : "indicateurs économiques : pêche maritime" et elles sont produites par Office National des Pêches. Ces données sont structurés comme un tableau Excel (extension.xlsx) et elles contient plusieurs indicateurs, statistiques et taux concernant volume des produits commercialisés issus de la pêche côtière et artisanale (en tonne) et elles sont associés aux délégués de la ville côtière qu'on appelle Les Individus tandis que les mois représentent des variables.

PECHE الصيد

Volume des produits commercialisés issus de la pêche côtière et artisanale حجم تسويق منتوجات الصيد الساحلي و التقليدي

(en tonne) (بالطن)

	2022			2021												
	مارس Mars	فبراير Fév.	يناير Janv.	دجنبر Dec	نوفمبر Nov	أكتوبر Oct	شتنبر Sept	غشت Août	يوليوز Juillet	يونيه Juin	ماي Mai	أبريل Avril	مارس Mars	فبراير Fév	يناير Janv	
Délégation Agadir	2 073	3 090	2 653	2 794	2 573	2 331	3 146	5 543	1 829	3 580	2 967	3 477	3 748	2 049	2 063	مندوبية اكادير
Délégation Sidi ifni	7 315	3 183	433	399	889	1 791	1 527	990	946	7 408	5 961	8 574	10 999	4 499	4 567	مندوبية سيدي إفني
Délégation Tantan	4 844	2 350	2 615	18 356	19 384	9 484	14 136	7 780	5 819	16 436	1 757	5 850	19 316	2 400	2 419	مندوبية طان طان
Délégation Casa	1 697	1 322	1 788	1 631	1 546	607	2 113	3 074	2 488	3 574	1 189	1 119	1 747	1 120	1 740	مندوبية الدار البيضاء
Délégation El Jadida	717	468	755	2 324	2 787	3 939	4 984	2 993	1 590	1 819	2 347	943	1 194	3 232	4 099	مندوبية الجديدة
Délégation Essaouira	921	1 153	1 956	2 478	2 647	1 949	1 148	905	512	1 442	793	515	1 200	325	408	مندوبية الصويرة
Délégation Mehdia	820	976	1 156	986	681	440	1 074	1 977	1 158	1 243	1 645	573	722	395	1 252	مندوبية مهدية
Délégation Mohamedia	163	284	671	397	484	384	169	101	264	471	24	60	132	141	353	مندوبية المحمدية
Délégation Rabat	156	24	32	59	1	1	1	44	48	405	59	21	20	17	39	مندوبية الرباط
Délégation Safi	4 283	3 400	3 976	6 615	7 545	7 958	8 983	6 459	2 061	5 767	2 358	4 598	4 170	4 383	4 913	مندوبية اسفي
Délégation Boujdour	3 573	5 246	2 396	7 784	10 504	10 731	9 045	1 519	2 701	6 473	5 264	7 554	2 306	671	2 394	مندوبية بوجدور
Délégation Dakhla	40 675	21 479	8 959	66 542	67 090	75 230	62 316	39 276	28 978	46 641	41 041	46 199	39 187	31 672	28 368	مندوبية الداخلة
Délégation Laayoune	8 165	21 349	17 856	34 769	44 385	40 819	45 421	35 152	19 038	18 791	2 624	14 015	12 569	19 408	16 666	مندوبية العيون
Délégation Hoceima	207	264	237	1 066	573	511	252	342	168	198	123	243	266	193	327	مندوبية الحسيمة
Délégation Jebha	78	81	97	352	212	239	120	130	111	176	71	43	67	65	183	مندوبية الجبهة
Délégation Larache	1 430	1 189	659	463	237	321	1 040	897	754	1 115	1 193	1 765	1 665	587	980	مندوبية العرائش
Délégation M'diq	155	134	147	242	251	217	350	230	300	422	225	233	219	170	432	مندوبية المضيق
Délégation Nador	431	655	710	323	63	73	443	628	431	487	363	377	430	413	676	مندوبية الناضور
Délégation Tanger	1 225	1 048	238	171	81	75	407	406	223	947	631	743	1 216	1 152	559	مندوبية طنجة
Total	**78 927**	**67 695**	**47 335**	**147 752**	**161 932**	**157 100**	**156 676**	**108 446**	**69 420**	**117 394**	**70 635**	**96 903**	**101 175**	**72 893**	**72 439**	**المجموع**

Figure 1: Le fichier du jeu de données (.csv)

III. Chargement des données

Dans ce segment initial du prétraitement des données, nous entamons la démarche d'acquisition du jeu de données brut depuis une source externe tout en configurant simultanément notre environnement de travail. L'objectif principal est de s'assurer que les données soient ingérées correctement et accessibles pour l'analyse ultérieure. Cette phase inclut la définition du répertoire de travail, où les données résident, et le chargement des données dans notre environnement analytique.

Pour atteindre cet objectif, les commandes suivantes ont été exécutées :

```
setwd("/Chemin du dossier contenant le fichier xlsx/")
file <- read.xlsx("Dataset.xlsx", sheetIndex = 1, header = TRUE)
```

- **setwd() :** Cette fonction établit le répertoire de travail, une étape cruciale car elle guide le script vers l'emplacement de la source de données.
- **read.xlsx("Dataset.xlsx", sheetIndex = 1, header = TRUE)** : Avec cette ligne, le script lit les données depuis un fichier Excel nommé "Dataset.xlsx". L'argument sheetIndex = 1 indique que l'extraction des données est effectuée depuis la première feuille du fichier Excel, tandis que header = TRUE spécifie que la première ligne du jeu de données contient les en-têtes de colonnes.

IV. Nettoyage des données

La phase de nettoyage des données implique l'élimination des données indésirables, telles que celles de l'année 2022, pour se concentrer uniquement sur les données de l'année 2021. Ainsi, chaque mois de 2021 devient une variable dans notre jeu de données. Voici les instructions exécutées pendant cette étape :

```
data <- file[c(3:22), c(1,5:16)]
```

Cette commande extrait des lignes spécifiques (de 3 à 22) et des colonnes spécifiques (colonne 1 et les colonnes de 5 à 16) des données chargées depuis Excel et stocke le résultat dans un jeu de données appelé **‘data’**

V. Transformation des données

La transformation des données est essentielle pour structurer le jeu de données de manière appropriée en vue de l'analyse. Cette étape peut inclure la réorganisation des colonnes, la définition des noms de lignes ou la conversion des types de données. Voici les actions de transformation des données effectuées :

```
names(data) <- data[1,]
data <- data[-c(1),]
```

Ces instructions attribuent les noms de colonnes en se basant sur les valeurs de la première ligne du jeu de données. Des noms de colonnes clairs et informatifs facilitent l'interprétation des données. Ensuite, la première ligne est supprimée du jeu de données car elle ne contient plus d'informations utiles pour l'analyse ultérieure.

```
rownames(data) <- data[, 1]
data <- data[, -1]
```

Ces commandes définit les noms de ligne du jeu de données 'data' en fonction des valeurs de la première colonne des données. Ensuite, la première colonne est supprimée du jeu de données **'data'**, car elle a été utilisée pour les noms de ligne.

```
dfata <- subset(data, select = c(12:1))
```

Réorganise les colonnes du jeu de données 'data' de la colonne 12 à la colonne 1. Elle réarrange les colonnes dans l'ordre inverse.

```
data[] <- lapply(data, type.convert, as.is = TRUE)
```

Cette commande convertit toutes les colonnes du jeu de données 'data' en types de données numériques en utilisant la fonction **'type.convert'**. L'argument **'as.is = TRUE'** garantit que les colonnes de caractères ne sont pas converties en facteurs.

Après avoir prétraité les données, voici le résultat final que l'on obtient :

	Janv	Fév	Mars	Avril	Mai	Juin	Juillet	Août	Sept	Oct	Nov	Dec
Délégation Agadir	2063.2318	2049.1711	3748.46830	3476.5231	2967.0594	3580.2380	1829.4766	5543.4701	3145.7519	2331.2041	2572.9792	2793.8360
Délégation Sidi ifni	4567.3875	4498.6455	10998.90950	8573.6835	5961.0586	7408.4446	945.7732	990.0075	1526.6024	1790.7085	889.3000	399.4590
Délégation Tantan	2419.4675	2399.8870	19316.37200	5850.1835	1756.7060	16436.3740	5819.4285	7779.7170	14135.7115	9484.1852	19384.4535	18355.6590
Délégation Casa	1739.6749	1119.5673	1746.97104	1119.3268	1189.0874	3573.7997	2488.2361	3073.9389	2113.0559	607.4283	1545.7475	1631.2223
Délégation El Jadida	4099.4425	3232.0610	1194.17850	942.8590	2347.4795	1818.5310	1589.7595	2993.2915	4984.2400	3938.5225	2787.4030	2324.2795
Délégation Essaouira	408.0961	325.2327	1200.24037	515.1946	793.1990	1441.7748	511.5761	905.2185	1148.0864	1949.3021	2647.1376	2478.2273
Délégation Mehdia	1251.9500	395.0980	721.79000	573.4190	1645.3410	1242.5850	1157.8040	1977.0910	1074.2390	440.0590	681.2050	986.1620
Délégation Mohamedia	353.1315	141.1625	132.35800	60.0035	23.7425	470.6225	264.0655	100.9975	169.2105	384.3350	483.6775	396.8995
Délégation Rabat	39.3475	17.3585	20.24200	21.4131	58.7580	404.9950	48.4365	43.5635	0.7760	0.7330	0.7500	59.2095
Délégation Safi	4913.4120	4383.0695	4169.58670	4597.9134	2357.5979	5766.7710	2060.5818	6459.1990	8982.6263	7958.3630	7545.0693	6615.0569
Délégation Boujdour	2394.4385	670.5630	2305.90800	7553.7315	5263.6395	6473.4915	2700.8840	1518.7295	9044.8465	10730.7985	10503.7795	7784.2810
Délégation Dakhla	28368.2245	31672.2410	39186.87250	46199.4505	41040.7200	46641.0855	28977.9656	39276.3275	62316.0202	75229.6142	67089.5955	66541.6610
Délégation Laayoune	16665.5927	19408.0932	12569.10170	14015.2066	2624.3621	18790.8019	19037.5769	35151.7945	45421.1045	40819.3844	44384.9606	34768.6437
Délégation Hoceima	326.5335	193.1915	266.34000	242.9260	123.3475	198.2240	168.2950	342.4590	252.1750	510.9570	572.7881	1065.6917
Délégation Jebha	182.9625	65.2475	67.44937	43.4055	71.1605	175.5036	111.0315	129.6619	120.4080	239.4340	211.9905	352.2655
Délégation Larache	979.5975	587.3240	1664.73440	1765.2257	1192.8379	1115.1012	754.1248	897.1069	1040.4860	320.6200	237.0085	463.4880
Délégation M'diq	431.5465	170.2865	219.41550	232.6500	225.2935	422.0560	300.3975	229.8880	350.4635	216.8355	250.8640	242.1180
Délégation Nador	676.2495	412.9295	430.07850	376.7115	363.1055	486.9400	431.3025	627.8050	442.8955	73.0125	62.7280	322.8565
Délégation Tanger	558.5370	1151.5645	1215.99450	742.7625	630.9315	946.9440	222.7980	406.0390	406.9830	74.8525	80.6515	171.4589

Figure 2: Chargement les données sur RStudio

Le processus complet de prétraitement des données dévoilé dans cette section constitue la pierre angulaire pour préparer un jeu de données propre et bien structuré, prêt pour les analyses ultérieures. Chaque phase du processus sert un objectif distinct visant à améliorer la qualité et l'utilisabilité des données, renforçant ainsi la fiabilité des insights générés dans les étapes suivantes du projet.

Chapitre 3 : Application d'ACP sur données

I . Introduction

Dans cette section, nous passons de la théorie à l'action. Notre objectif est de mettre en œuvre l'Analyse en Composantes Principales (ACP) sur nos données relatives à la pêche au Maroc. Comme nous l'avons déjà expliqué dans le chapitre précédent, nous avons détaillé l'origine de ces données et leur préparation. Maintenant, concentrons-nous sur la manière dont l'ACP est appliquée pour extraire des informations essentielles à partir de ces données. Cette étape nous rapproche davantage de la découverte des tendances significatives et des conclusions importantes pour le secteur de la pêche au Maroc.

II . Application d'ACP en R

Maintenant, on applique l'ACP avec une fonction prédéfinie dans la bibliothèque FactoMiner.

```
install.packge("FactoMineR")
library("FactoMineR")
resultACP <- PCA(data[, 1:12], scale.unit = TRUE, graph = TRUE)
```

Figure 3 : La commande PCA

Nous lançons l'ACP sur les données prétraitées à l'aide de la fonction PCA(). Voici un aperçu des paramètres spécifiés :

- **data[, 1:12]** : Nous utilisons les 12 premières colonnes des données pour l'ACP, ces colonnes étant les plus pertinentes pour notre analyse.
- **scale.unit = TRUE** : Cette option centre et réduit les données, garantissant une analyse adéquate des variables.
- **graph = TRUE** : Nous générons des graphiques pour visualiser les résultats de l'ACP, ce qui nous permettra de mieux interpréter les composantes principales.

On peut afficher un résumé détaillé des résultats de l'interprétation de l'ACP utilisant la fonction summary(resultACP). Ce résumé fournit des informations essentielles sur la variance expliquée, les charges des variables et les contributions des individus à ses composantes.

```
summary(resultACP)
```

```
Call:
PCA(X = df[, 1:12], scale.unit = TRUE, ncp = 5, graph = TRUE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
Variance              11.302   0.450   0.168   0.054   0.012   0.006   0.004   0.003   0.001   0.000   0.000   0.000
% of var.             94.184   3.747   1.403   0.452   0.097   0.052   0.030   0.025   0.006   0.002   0.001   0.000
Cumulative % of var.  94.184  97.931  99.334  99.786  99.883  99.935  99.965  99.991  99.997  99.999 100.000 100.000

Individuals (the 10 first)
                         Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2  
Délégation Agadir    |  0.845 | -0.783  0.285  0.857 |  0.084  0.083  0.010 |  0.050  0.078  0.003 |
Délégation Sidi ifni |  1.260 | -0.301  0.042  0.057 |  0.969 10.985  0.591 | -0.105  0.343  0.007 |
Délégation Tantan    |  2.073 |  1.178  0.647  0.323 |  0.247  0.715  0.014 | -1.684 88.644  0.660 |
Délégation Casa      |  1.161 | -1.130  0.595  0.948 | -0.049  0.028  0.002 |  0.023  0.017  0.000 |
Délégation El Jadida |  1.003 | -0.884  0.364  0.776 | -0.139  0.226  0.019 |  0.348  3.791  0.121 |
Délégation Essaouira |  1.428 | -1.416  0.933  0.983 |  0.010  0.001  0.000 |  0.003  0.000  0.000 |
Délégation Mehdia    |  1.419 | -1.402  0.916  0.977 |  0.011  0.001  0.000 |  0.145  0.660  0.010 |
Délégation Mohamedia |  1.672 | -1.666  1.293  0.993 | -0.021  0.005  0.000 |  0.083  0.213  0.002 |
Délégation Rabat     |  1.728 | -1.723  1.383  0.994 |  0.000  0.000  0.000 |  0.080  0.201  0.002 |
Délégation Safi      |  0.368 | -0.106  0.005  0.083 | -0.166  0.321  0.202 |  0.103  0.331  0.078 |

Variables (the 10 first)
                        Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2  
Janv                 |  0.985  8.584  0.970 | -0.057  0.733  0.003 |  0.124  9.106  0.015 |
Fév                  |  0.983  8.558  0.967 | -0.085  1.594  0.007 |  0.110  7.233  0.012 |
Mars                 |  0.929  7.639  0.863 |  0.236 12.390  0.056 | -0.268 42.715  0.072 |
Avril                |  0.971  8.345  0.943 |  0.215 10.328  0.046 |  0.074  3.270  0.006 |
Mai                  |  0.893  7.055  0.797 |  0.410 37.439  0.168 |  0.175 18.180  0.031 |
Juin                 |  0.982  8.528  0.964 |  0.129  3.691  0.017 | -0.134 10.736  0.018 |
Juillet              |  0.989  8.654  0.978 | -0.130  3.747  0.017 |  0.001  0.001  0.000 |
Août                 |  0.954  8.045  0.909 | -0.290 18.742  0.084 |  0.003  0.004  0.000 |
Sept                 |  0.982  8.540  0.965 | -0.179  7.088  0.032 | -0.015  0.128  0.000 |
Oct                  |  0.993  8.733  0.987 | -0.049  0.543  0.002 |  0.070  2.933  0.005 |
```

III . Choisir le nombre de composante principale

Comme nous pouvons observer dans la sortie de la fonction **"summary"** de l'Analyse en Composantes Principales (ACP), elle nous présente les valeurs propres ainsi que leurs vecteurs propres, ça va nous aider à choisir le nombre de composantes principales dont ‘on a besoin pour exprimer nos données

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
Variance              11.302   0.450   0.168   0.054   0.012   0.006   0.004   0.003   0.001   0.000   0.000   0.000
% of var.             94.184   3.747   1.403   0.452   0.097   0.052   0.030   0.025   0.006   0.002   0.001   0.000
Cumulative % of var.  94.184  97.931  99.334  99.786  99.883  99.935  99.965  99.991  99.997  99.999 100.000 100.000
```

Nous visons à conserver au moins 95 % de l'information après la réduction des dimensions. Pour déterminer le nombre optimal de composantes principales, Alors on commence à calculer le pourcentage « % of var » cumulé de variance expliquée, en commençant par la première dimension, jusqu'à atteindre une valeur supérieure ou égale à 95 %. Dans notre cas, nous avons atteint un pourcentage de 97,93 % dès la deuxième dimension. Par conséquent, nous avons relancé l'Analyse en Composantes Principales (ACP) en spécifiant le nombre de composantes (ncp) à 2. Le code correspondant est le suivant :

```
library("FactoMineR")
resultACP <- PCA(data[, 1:12], ncp=2, scale.unit = TRUE, graph = TRUE)
```

Cette approche nous permet de réduire efficacement la dimensionnalité tout en préservant une grande partie de l'information contenue dans nos données.

IV .Implémentation avec Python

Dans cette section, nous allons mettre en œuvre l'Analyse en Composantes Principales (ACP) en utilisant Python. En suivant les étapes précédemment mentionnées, nous explorerons de manière pratique le fonctionnement de l'ACP avec Python pour une meilleure compréhension.

Voici une explication étape par étape du code en python :

1. Standardisation des données

Avant de commencer l'ACP, il est essentiel de standardiser les données. La standardisation des données signifie que chaque variable est centrée autour de zéro (la moyenne de chaque variable devient zéro) et est mise à l'échelle pour avoir un écart-type égal à un. Cette étape est cruciale car elle garantit que toutes les variables ont une influence égale sur l'analyse, quel que soit leur ordre de grandeur initial. Pour ce faire, nous utilisons la classe **'StandardScaler'** de Scikit-Learn

```
# Bibliothèques nécessaires pour la standardisation des données
from sklearn.preprocessing import StandardScaler

# Standardize the data
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)
```

Le résultat de cette étape est un nouveau tableau de données **(df_scaled)** dans lequel toutes les variables ont été standardisées.

2. Calcul des valeurs propres et des vecteurs propres

Une fois que les données sont standardisées, nous calculons la matrice de covariance des données standardisées. La matrice de covariance est une mesure des relations linéaires entre les variables. Ensuite, nous utilisons la fonction **np.linalg.eig** de NumPy pour calculer les valeurs propres (eigenvalues) et les vecteurs propres (eigenvectors) de cette matrice de covariance.

```python
# Bibliothèques nécessaires pour le calcul des valeurs propres et des vecteurs propres
import numpy as np

# Calculate eigenvalues and eigenvectors
cov_matrix = np.cov(df_scaled.T)
eigenvalues, eigenvectors = np.linalg.eig(cov_matrix)
```

Les valeurs propres représentent l'importance de chaque composante principale, tandis que les vecteurs propres déterminent la direction de chaque composante principale dans l'espace des variables.

3. Calcul du ratio de variance expliquée

Une fois que nous avons obtenu les valeurs propres, nous calculons le ratio de variance expliquée pour chaque composante principale. Le ratio de variance expliquée nous indique la proportion de la variance totale des données qui est expliquée par chaque composante principale. Cela nous aide à comprendre quelle proportion de l'information est capturée par chaque composante principale.

```python
# Calculate explained variance ratio
explained_variance_ratio = eigenvalues / np.sum(eigenvalues)
```

Ce tableau de ratios nous permet de visualiser graphiquement la contribution de chaque composante principale à la variance totale des données et de prendre des décisions éclairées sur le nombre de composantes principales à retenir.

V . Comparaison des résultats

Voici une comparaison entre l’application PC A avec **FactomineR** et notre implémentation avec python :

```
> resultACP$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  1.130210e+01           94.184156815                          94.18416
comp 2  4.496437e-01            3.747030804                          97.93119
comp 3  1.683445e-01            1.402870446                          99.33406
comp 4  5.423276e-02            0.451939704                          99.78600
comp 5  1.168192e-02            0.097349350                          99.88335
comp 6  6.249849e-03            0.052082074                          99.93543
comp 7  3.590416e-03            0.029920137                          99.96535
comp 8  3.028982e-03            0.025241516                          99.99059
comp 9  7.571324e-04            0.006309437                          99.99690
comp 10 1.933982e-04            0.001611652                          99.99851
comp 11 1.651656e-04            0.001376380                          99.99989
comp 12 1.340221e-05            0.000111685                         100.00000
```

Figure 4: Résultat du PCA avec la commande PCA du package FactomineR

Entrée [7]:

```python
# Créez un DataFrame pour afficher le rapport de variance expliquée
explained_variance_df = pd.DataFrame({'eigenvalue': eigenvalues,
                                      'variance.percent': explained_variance_ratio * 100,
                                      'cumulative.variance.percent': explained_variance_ratio.cumsum() * 100})

# Renommez les index pour correspondre à la sortie souhaitée
explained_variance_df.index = ['Dim.' + str(i) for i in range(1, len(explained_variance_ratio) + 1)]

# Afficher le DataFrame
print(explained_variance_df)
```

```
        eigenvalue  variance.percent  cumulative.variance.percent
Dim.1    11.929993         94.184157                    94.184157
Dim.2     0.474624          3.747031                    97.931188
Dim.3     0.177697          1.402870                    99.334058
Dim.4     0.057246          0.451940                    99.785998
Dim.5     0.012331          0.097349                    99.883347
Dim.6     0.006597          0.052082                    99.935429
Dim.7     0.003790          0.029920                    99.965349
Dim.8     0.003197          0.025242                    99.990591
Dim.9     0.000799          0.006309                    99.996900
Dim.10    0.000014          0.000112                    99.997012
Dim.11    0.000204          0.001612                    99.998624
Dim.12    0.000174          0.001376                   100.000000
```

Figure 5: Résultat du PCA avec Python

Après avoir effectué la comparaison, nous avons constaté les résultats de l'analyse en composantes principales (ACP) obtenus avec FactoMineR (R) et votre implémentation en Python sont très similaires. Les valeurs propres et les pourcentages de variance expliquée par chaque composante principale concordent étroitement entre les deux approches. La première composante principale explique une part importante de la variance, soit environ 94%, dans les deux cas. Les résultats suggèrent que les deux implémentations sont fiables et produisent des résultats cohérents, indépendamment du langage de programmation utilisé. Cette concordance renforce la validité des analyses réalisées et confirme que les étapes de calcul de l'ACP ont été correctement mises en œuvre dans les deux

Chapitre 4 : Visualisation et Interprétation

I. Introduction

Dans ce chapitre, nous allons explorer la relation entre différents facteurs qui influencent les volumes des produits commercialisés issus de la pêche côtière et artisanale au Maroc à l'aide de techniques de visualisation et d'interprétation. Nous utiliserons une variété de techniques graphiques.

En visualisant et en interprétant les données, nous pourrons identifier les principaux facteurs qui ont le plus grand impact sur le volume des produits commercialisés issus de la pêche et obtenir des insights sur la manière dont ces facteurs interagissent entre eux.

II. Quantité d'informations expliquée par chaque composant

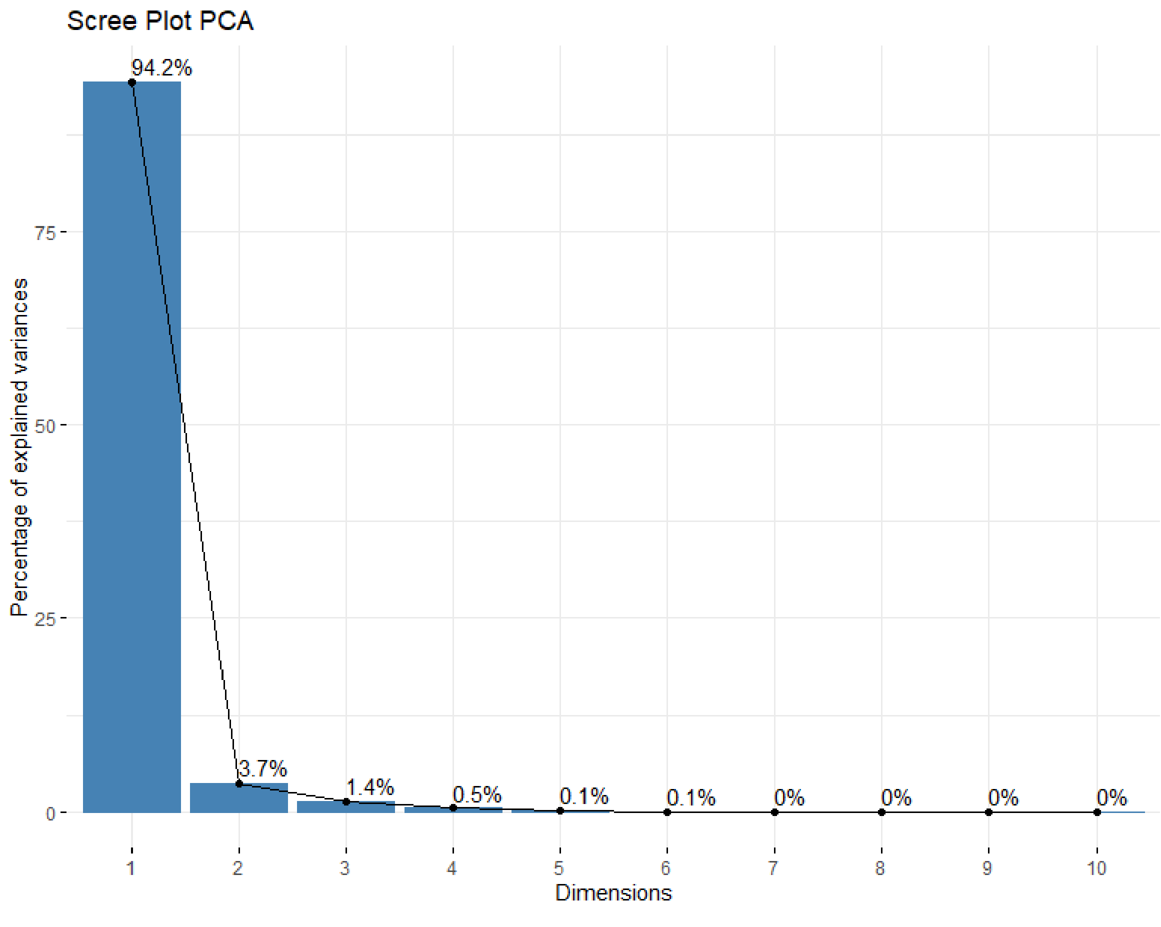


Figure 6: Screeplot

```r
# Scree plot of eigenvalues
fviz_screeplot(resultACP, type = "lines", addlabels = TRUE, main = "Scree Plot PCA")
```

fviz_screeplot est une fonction de la bibliothèque « factoextra » de R qui permet de créer un graphique en escalier, à partir des résultats de l'analyse des composantes principales (ACP).
Scree plot est un outil couramment utilisé pour visualiser et interpréter les résultats de l'ACP. Il représente graphiquement la variance expliquée par chaque composant principal, en ordonnant les composants principaux par ordre décroissant de variance expliquée. Il peut être utilisé pour déterminer le nombre de composants principaux à conserver pour l'analyse, en observant le point où la variance expliquée commence à diminuer de manière significative.
Après avoir examiné le graphique, nous pouvons constater que la première dimension explique 94,2 % de la variance, tandis que la deuxième dimension n'en explique que 3,7 %. Comme cela est clairement illustré dans le graphique en cascade, donc ces 2 dimensions sont plus que suffisantes pour représenter ce jeu de données.

III. Corrélation des variables

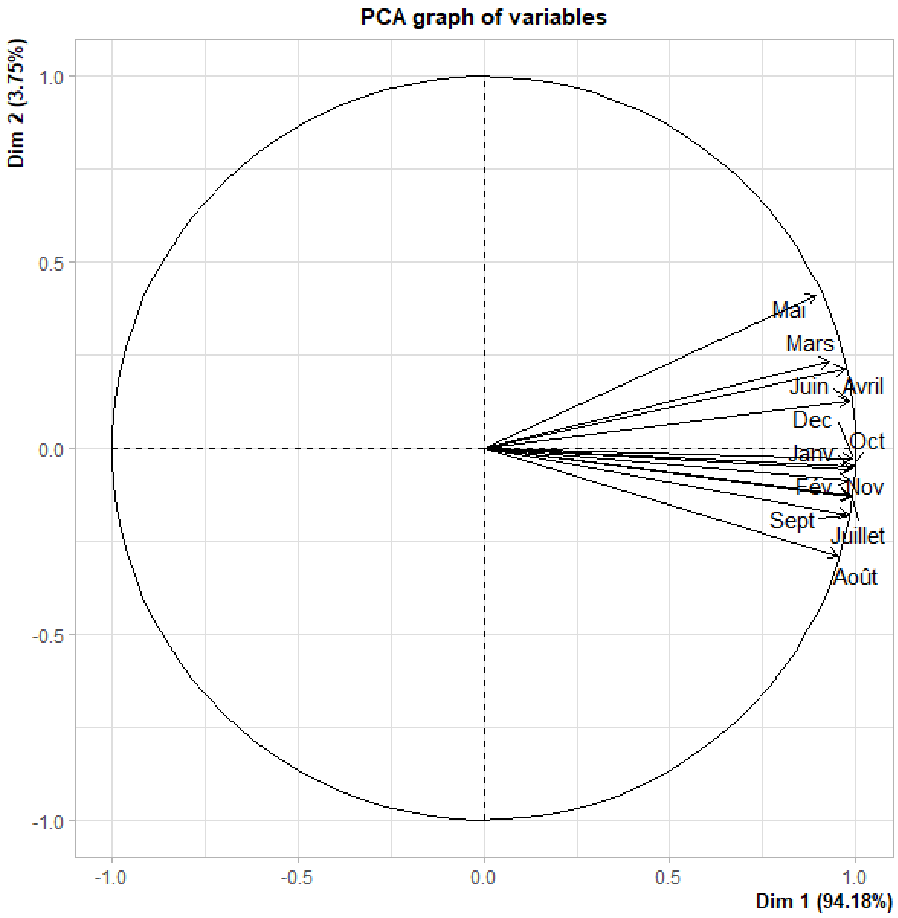


Figure 7: cercle de corrélation

La première observation à noter est que la plupart des variables sont bien réparties dans ce plan composé des deux premiers axes, car la distance entre la majorité des variables et le centre du cercle de corrélation est significative. De plus, il est notable que presque toutes les variables sont bien représentées sur la première dimension par rapport à la deuxième dimension, en se basant sur leurs mesures de corrélation.

IV. Contribution des variables/individues dans chaque CP

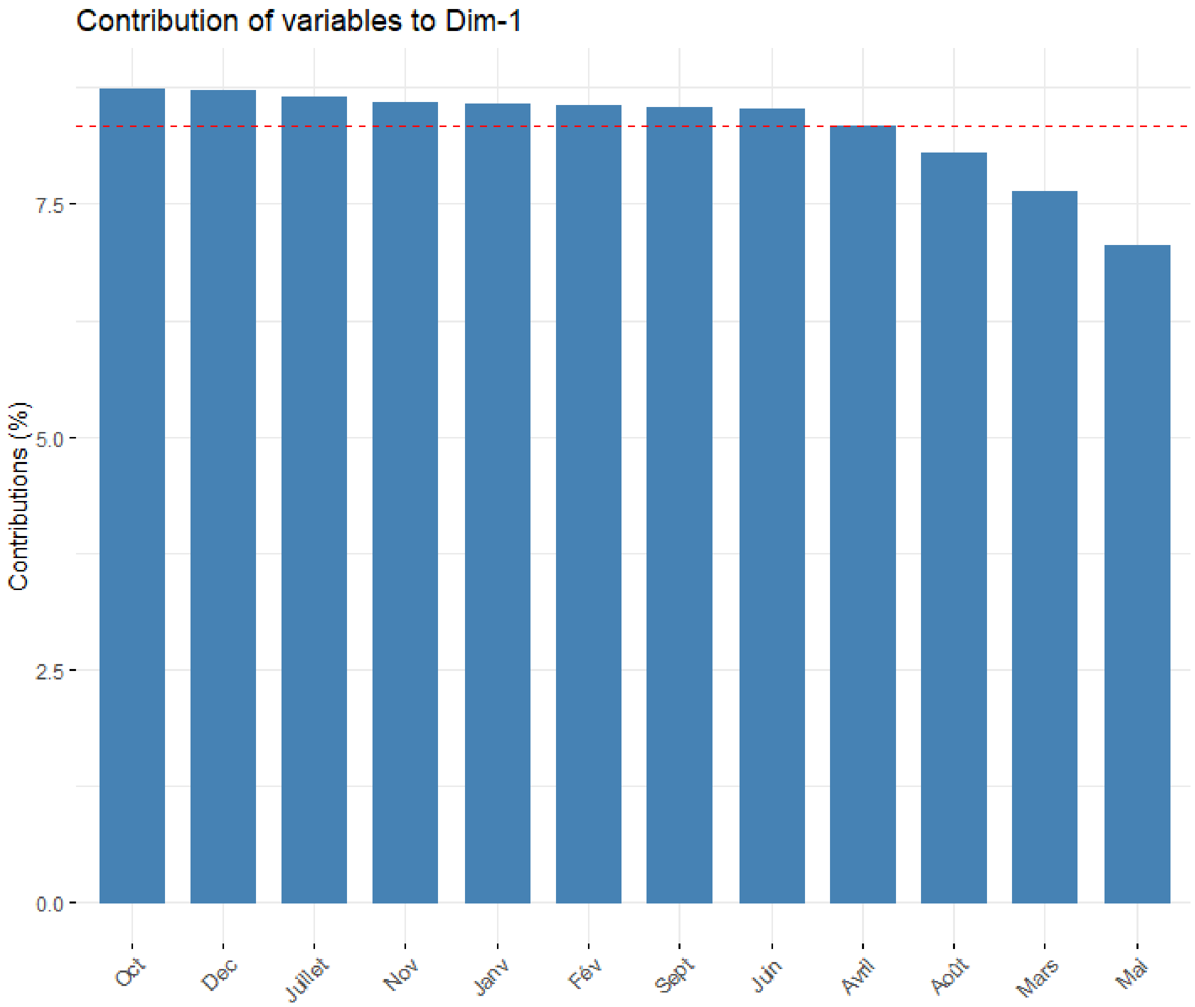


Figure 8: Histogramme de la Contribution des variables à la dimension 1

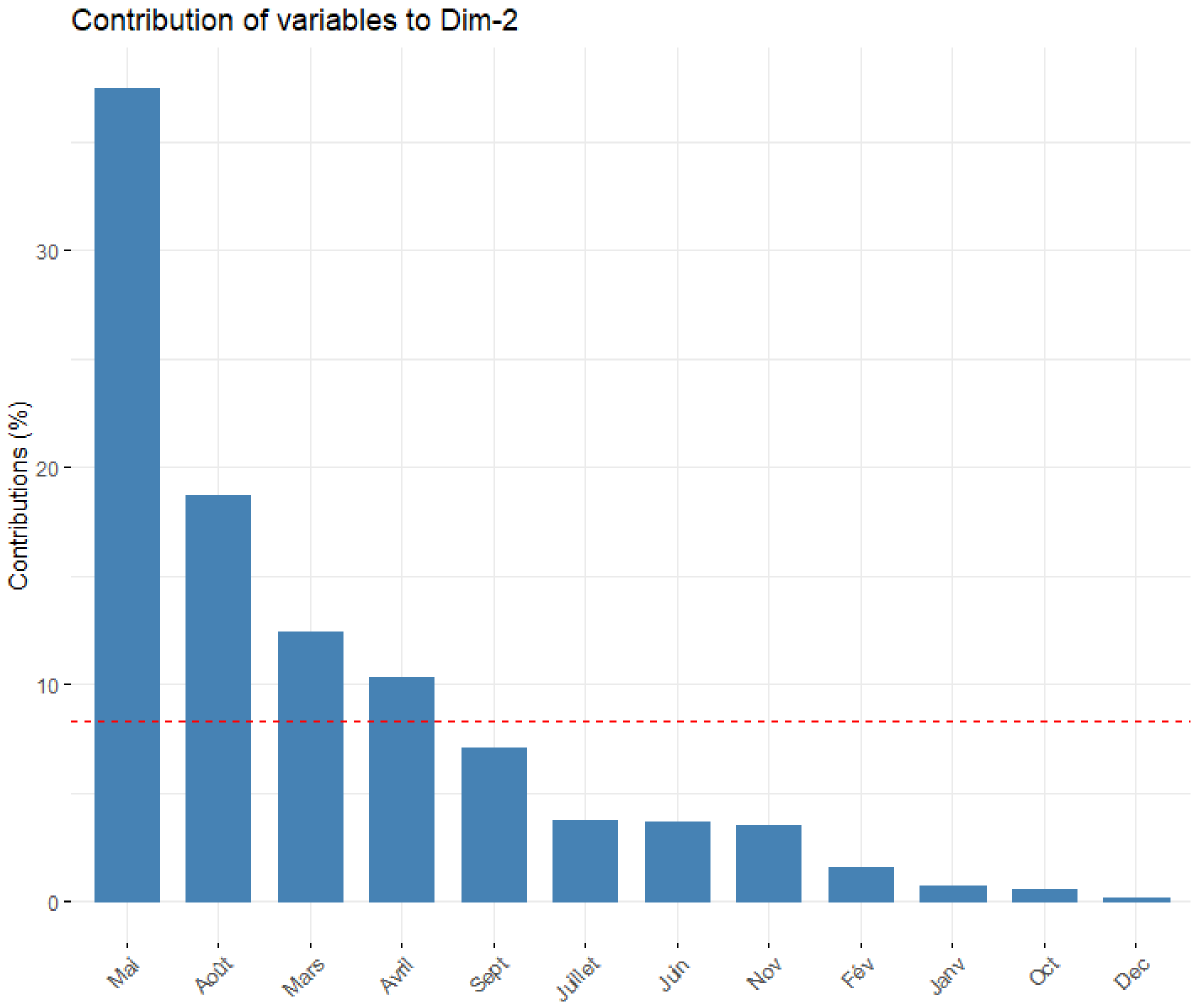


Figure 9: Histogramme de la Contribution des variables à la dimension 2

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, comme Août, Mars et Avril, et pour la deuxième dimension on voit que les variables : Mars, Aout, Mars, Avril sont les plus contributeurs.

Maintenant en passe à la contribution des individues dans chaque CP, voici les graphes :

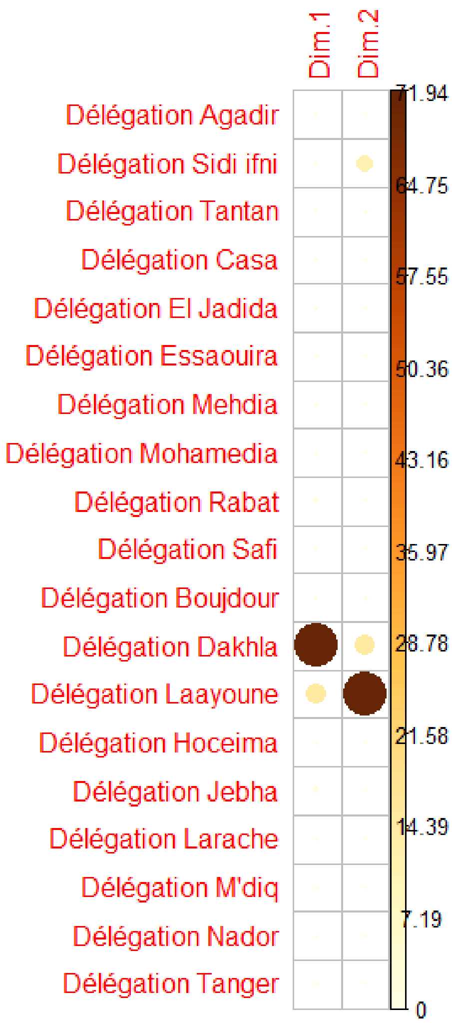


Figure 10: Corrélogramme de la Contribution des individus à la Dimension 1 et 2

On peut observer que deux principaux individus contribuent significativement aux deux dimensions, à savoir la délégation de Dakhla et celle de Laâyoune. De plus, il est à noter que la délégation de Sidi Ifni a également une contribution notable à la deuxième dimension.

V. Qualité de représentation des variables/individues dans chaque CP

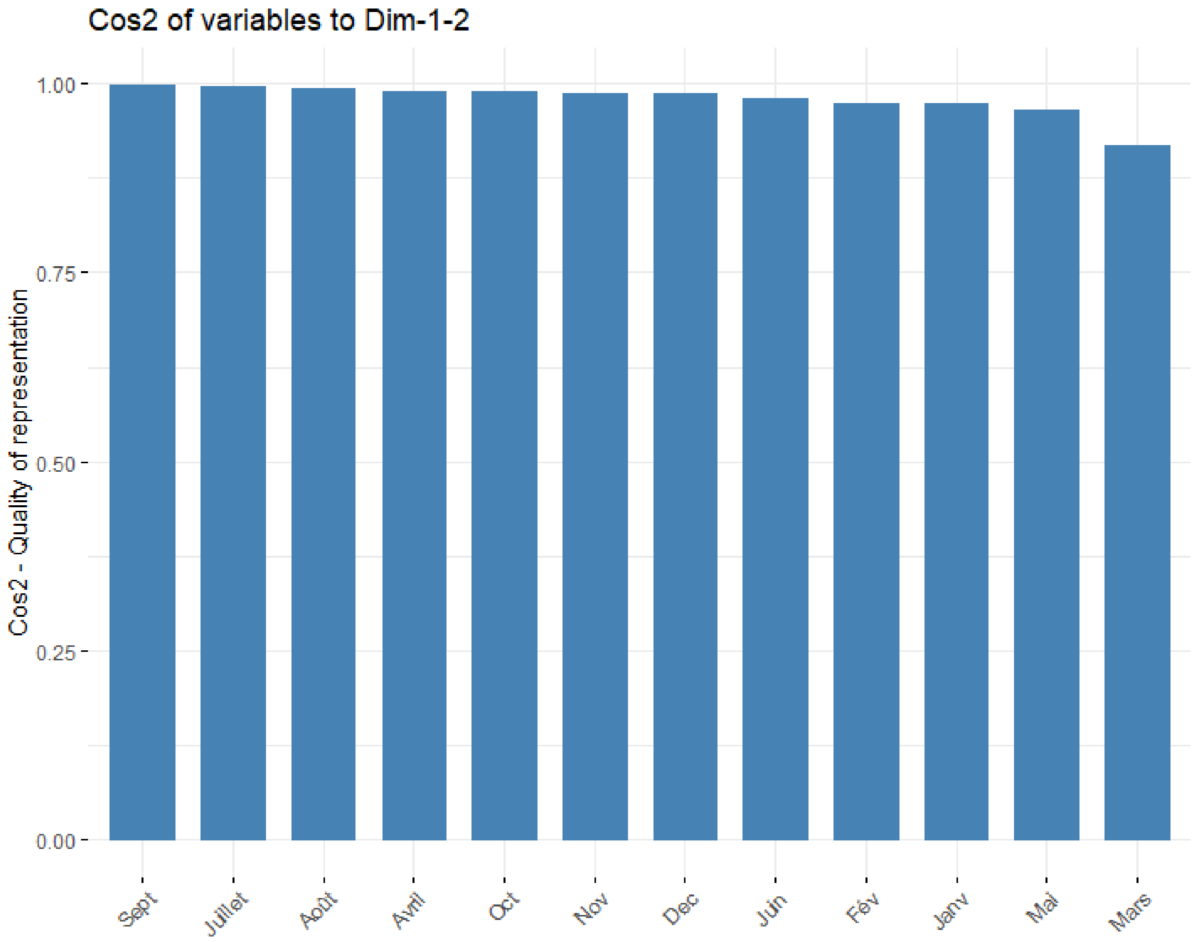


Figure 11: Histogramme de la Qualité de représentation des variables à la dimension 1 et 2

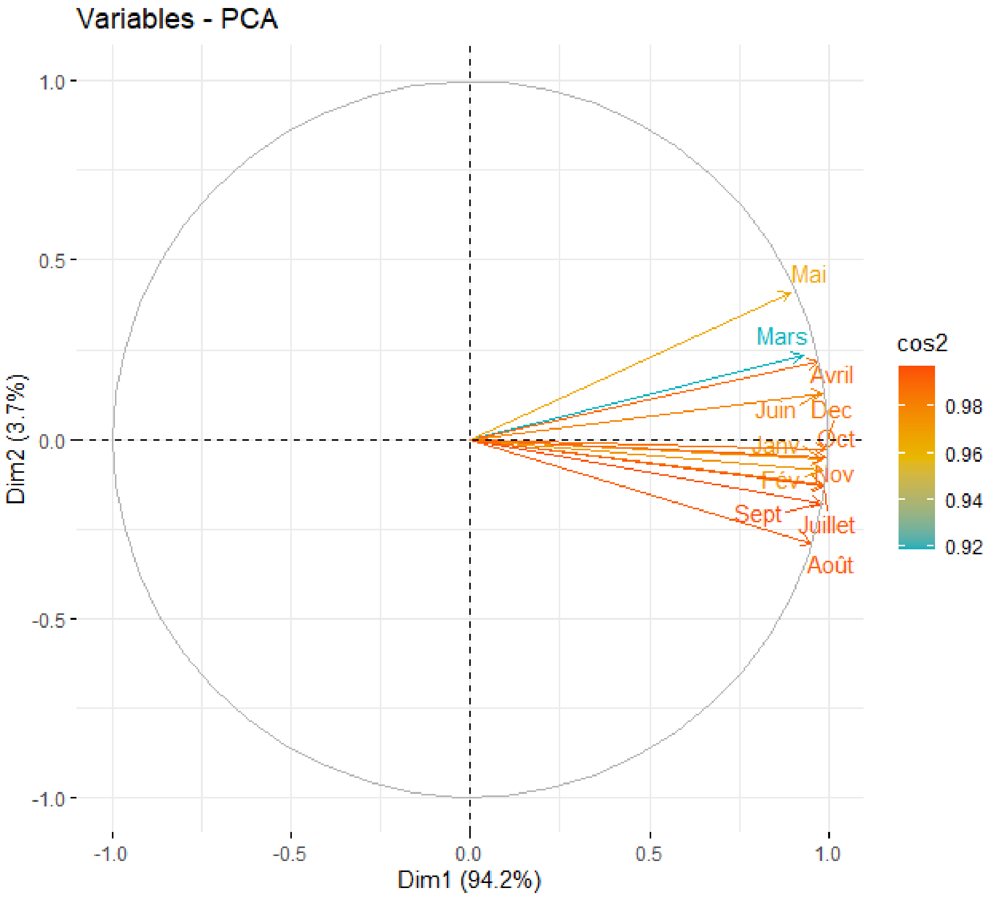


Figure 12: Cercle de la Qualité de représentation des variables à la dimension 1 et 2

Dans ces graphiques, nous pouvons clairement voir que les variables avec des valeurs de cos2 faibles, comme « mars », sont moins importantes et apparaissent en bleu, tandis que les variables avec des valeurs de cos2 modérées, comme « janvier », « février » et « mai », » apparaissent en orange. Ceci démontre son importance modérée. Les autres variables avec des valeurs de cos2 élevées sont colorées en rouge, soulignant leur forte influence dans l'analyse.

Maintenant, abordons la qualité de représentation dans chaque composante principale (CP). Voici le graphique correspondant :

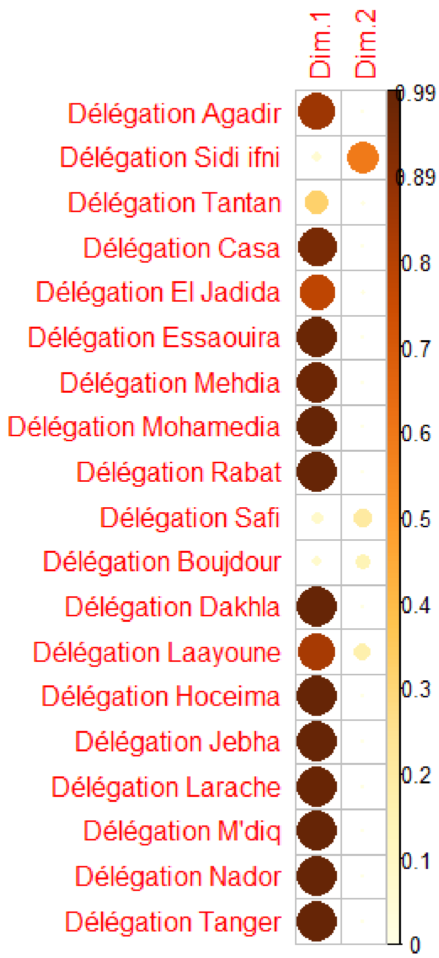


Figure 13: Corrélogramme de la Contribution des individus à la Dimension 1 et 2

On remarque que la grande majorité des individus sont fortement représentés dans la première dimension, à l'exception d'un petit groupe d'individus qui affichent une présence modérée à faible, tels que la "Délégation Sidi Ifni", la "Délégation Tan Tan", la "Délégation Safi" et la "Délégation Boujdour". En revanche, ces mêmes individus montrent une présence relativement plus faible dans la deuxième dimension.

VI. Interprétation par Biplot

Dans un **biplot**, les variables sont représentées par des vecteurs et les individus par des points. La longueur et la direction des vecteurs représentent la force et la direction des relations entre les variables, tandis que la position des points par rapport aux vecteurs représente les relations entre les individus et les variables. Pour afficher ce graphique on utilise la commande suivante :

```r
# Biplot of individuals and variables
fviz_pca_biplot(resultACP,
                #les individus
                geom.ind="point",
                axes=c(1,2),
                pointshape=21, pointsize=1,
                #les variables
                alpha.var="contrib",
                col.var= "cos2",
                gradient.cols=c("#00AFBB", "#E7B800", "#FC4E07"),
                repel = TRUE)
```

Cette commande nous permet également d'obtenir des informations sur la contribution des variables à la formation des composants principales et sur leur qualité de représentation.

Le résultat :

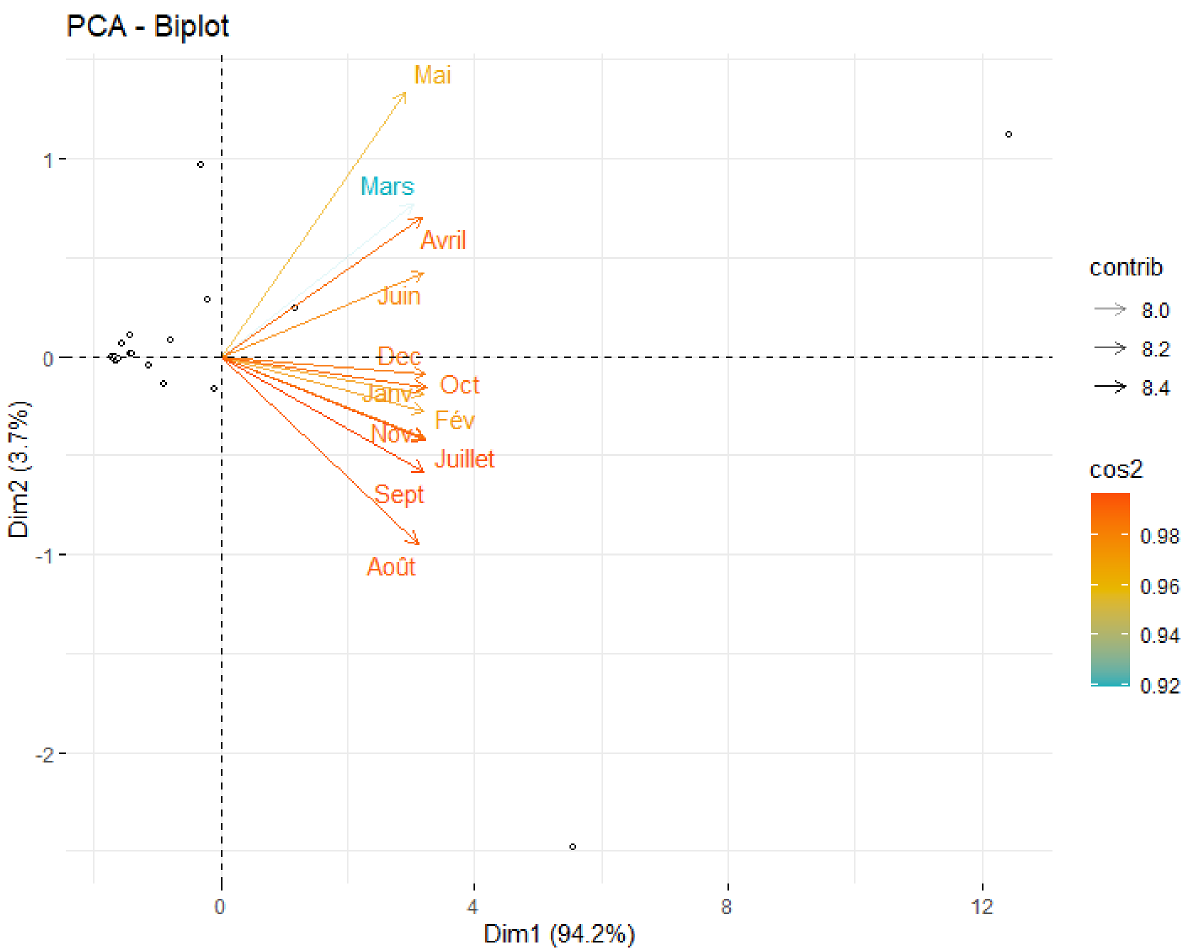


Figure 14: Biplot de la dimension 1 et 2

En constate d'abord que la variable « Mars » a une faible qualité de représentation donc il est difficile d'interpréter cette relation.

La dimension 1 oppose des individus caractérisés par une coordonnée fortement positive sur l'axe (à droite du graphe) à des individus caractérisés par une coordonnée fortement négative sur l'axe (à gauche du graphe).

- **Le groupe 1** (caractérisés par une coordonnée positive sur le premier l'axe) partage : de fortes valeurs pour toutes les variables comme « Dec », « oct » et etc.
- **Le groupe 2** partage : de faibles valeurs pour les variables, ce groupe est contient un seul variable « Mars ».

Notons aussi que les variables « Sept », « juillet » et « Aout », sont extrême ment corrélées à cette dimension.

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 « Mai », « Mars », « Avril » et « juin ».
- **Le groupe 2** (caractérisés par une coordonnées négative sur l'axe) partage : de faibles valeurs pour les variables « Dec », « Oct », « Janv », « Nov », « Fév », « juillet », « Sept » et « Aout » (de la plus extrême à la moins extrême).

En gros, un **biplot** peut être interprété comme suit : un individu qui se trouve du même côté d'une variable donnée a une valeur élevée pour cette variable ; un individu qui se trouve du côté opposé d'une variable donnée a une faible valeur pour cette variable.

VII. Quels sont les facteurs qui influencent le volume des produits commercialisés issus de la pêche maritime ?

Le volume des produits de la pêche maritime est influencé par divers facteurs. La saison de pêche, les conditions météorologiques, les réglementations gouvernementales, la disponibilité des espèces, les techniques de pêche, les facteurs économiques, l'environnement marin, les compétences des pêcheurs, l'infrastructure, et la démographie des délégués côtiers. En résumé, il s'agit d'une interaction complexe de facteurs biologiques, environnementaux, réglementaires, économiques, sociaux et culturels. Une gestion durable est essentielle pour préserver les ressources marines à long terme.

Conclusion

En général, l'analyse en composantes principales (ACP) est un outil efficace analyser des données. Elle permet d'identifier les variables et les observations les plus importantes dans le jeu de données. Dans ce travail, nous avons étudié les étapes nécessaires pour utiliser l'ACP, y compris la collecte de données, l'implémentation avec R et Python, la visualisation et l'interprétation des résultats. Grâce à la visualisation et l'interprétation, nous avons pu mieux comprendre les données et en tirer des conclusions significatives. L'ACP est une technique précieuse qui peut être appliquée à de nombreux ensembles de données et a de nombreuses applications dans les domaines de la statistique, de l'apprentissage automatique et de la science des données.

Références

- Abdi, Hervé, and Lynne J. Williams. 2010. “Principal Component Analysis.” *John Wiley and Sons, Inc. WIREs Comp Stat* 2: 433–59. http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf.

- Husson, Francois, Sebastien Le, and Jérôme Pagès. 2017. *Exploratory Multivariate Analysis by Example Using R*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://factominer.free.fr/bookV2/index.html.

- Jollife, I.T. 2002. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag. https://goo.gl/SB86SR.

- Kaiser, Henry F. 1961. “A Note on Guttman’s Lower Bound for the Number of Common Factors.” *British Journal of Statistical Psychology* 14: 1–2.

- Peres-Neto, Pedro R., Donald A. Jackson, and Keith M. Somers. 2005. “How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited.” *British Journal of Statistical Psychology* 49: 974–97.