M2SI-INSEA

Analyse de Données

Étude des facteurs influençant l'espérance de vie via PCA

Encadré par :
Prof. Marya Sadki

Réalisé par:
M. Yassine DARIF

Plan

01 »

Introduction

Introduction

Le projet consiste à appliquer l'Analyse en Composantes Principales (ACP) sur un jeu de données lié à l'espérance de vie moyenne dans différents pays. L'objectif est de comprendre les facteurs qui influencent l'espérance de vie et d'identifier les relations entre ces facteurs de manière synthétique.

- Réduire le nombre de variables tout en conservant l'essentiel de l'information.
- Identifier les liens entre les facteurs influençant l'espérance de vie.
- Regrouper les pays aux profils similaires selon leurs données
- Créer des graphiques clairs pour interpréter les résultats.

02 »

Méthodologie

0. Technologies

✓ Environnement

✓ Python Packages

```python
# Pour les calculs numériques et les opérations sur les tableaux
import numpy as np

# Pour la manipulation et l'analyse de données
import pandas as pd

# Pour la création de graphiques de base
import matplotlib.pyplot as plt

# Pour les visualisations statistiques avancées
import seaborn as sns

# Pour l'analyse en composantes principales (ACP)
from sklearn.decomposition import PCA

# Pour la normalisation et le prétraitement des données
from sklearn.preprocessing import StandardScaler
```

1. Chargement des Données

```
file_path = "Life-Expectancy-Data-Averaged.csv"
data = pd.read_csv(file_path, encoding='ISO-8859-1', sep = ",")
```

- AnnéeDécès_infantiles
- Décès_sous_cinq_ans
- Mortalité_adulte
- PIB_par_habitant
- Population_en_millions
- Maigreur_10_19_ans
- Maigreur_5_9_ans
- Scolarité
- Statut_économique
- Espérance_de_vie
- …..

	Country	Region	Year	Infant_deaths	Under_five_deaths	Adult_mortality	Alcohol_consumption	Hepatitis_B	Measles	BMI	Polio	Diphtheria
0	Afghanistan	Asia	2007.5	71.08125	98.61250	265.804969	0.016125	64.5625	24.3750	22.46250	55.3750	55.1250
1	Albania	Rest of Europe	2007.5	15.25625	17.14375	83.132969	4.696875	98.0000	95.9375	25.85625	98.1250	98.0625
2	Algeria	Africa	2007.5	26.75625	31.19375	113.439281	0.400625	88.3125	93.2500	24.86875	91.7500	91.8750
3	Angola	Africa	2007.5	88.76875	144.16250	297.844063	4.935625	68.8125	64.0000	22.51875	35.7500	55.5625
4	Antigua and Barbuda	Central America and Caribbean	2007.5	9.47500	11.51875	142.478813	7.755000	98.2500	75.4375	25.85000	96.9375	98.3125

1. Chargement des Données

Analyse du type de variables

```
data.dtypes
```

```
Country                          object
Region                           object
Year                            float64
Infant_deaths                   float64
Under_five_deaths               float64
Adult_mortality                 float64
Alcohol_consumption             float64
Hepatitis_B                     float64
Measles                         float64
BMI                             float64
Polio                           float64
Diphtheria                      float64
Incidents_HIV                   float64
GDP_per_capita                  float64
Population_mln                  float64
Thinness_ten_nineteen_years     float64
Thinness_five_nine_years        float64
Schooling                       float64
Economy_status                  float64
Life_expectancy                 float64
dtype: object
```

le nombre de lignes et de colonnes

```
data.shape
```

```
(179, 20)
```

Analyse de données manquantes

```
data.isnull().sum()
```

```
Country                          0
Region                           0
Year                             0
Infant_deaths                    0
Under_five_deaths                0
Adult_mortality                  0
Alcohol_consumption              0
Hepatitis_B                      0
Measles                          0
BMI                              0
Polio                            0
Diphtheria                       0
Incidents_HIV                    0
GDP_per_capita                   0
Population_mln                   0
Thinness_ten_nineteen_years      0
Thinness_five_nine_years         0
Schooling                        0
Economy_status                   0
Life_expectancy                  0
dtype: int64
```

2. Prétraitement de Données

01

- Renommer les colonnes en français

```
data = data.rename(columns={
    'Country': 'Pays',
    'Region': 'Région',
    'Year': 'Année',
    'Infant_deaths': 'Décès_infantiles',
    'Under_five_deaths': 'Décès_sous_cinq_ans',
    'Adult_mortality': 'Mortalité_adulte',
```

02

- Mettre la colonne pays comme index

```
data.set_index('Pays', inplace=True)
#Enlever le nom de l'index
data.index.name = None
```

04

- Normaliser les données

```
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_clean)
data_scaled
```

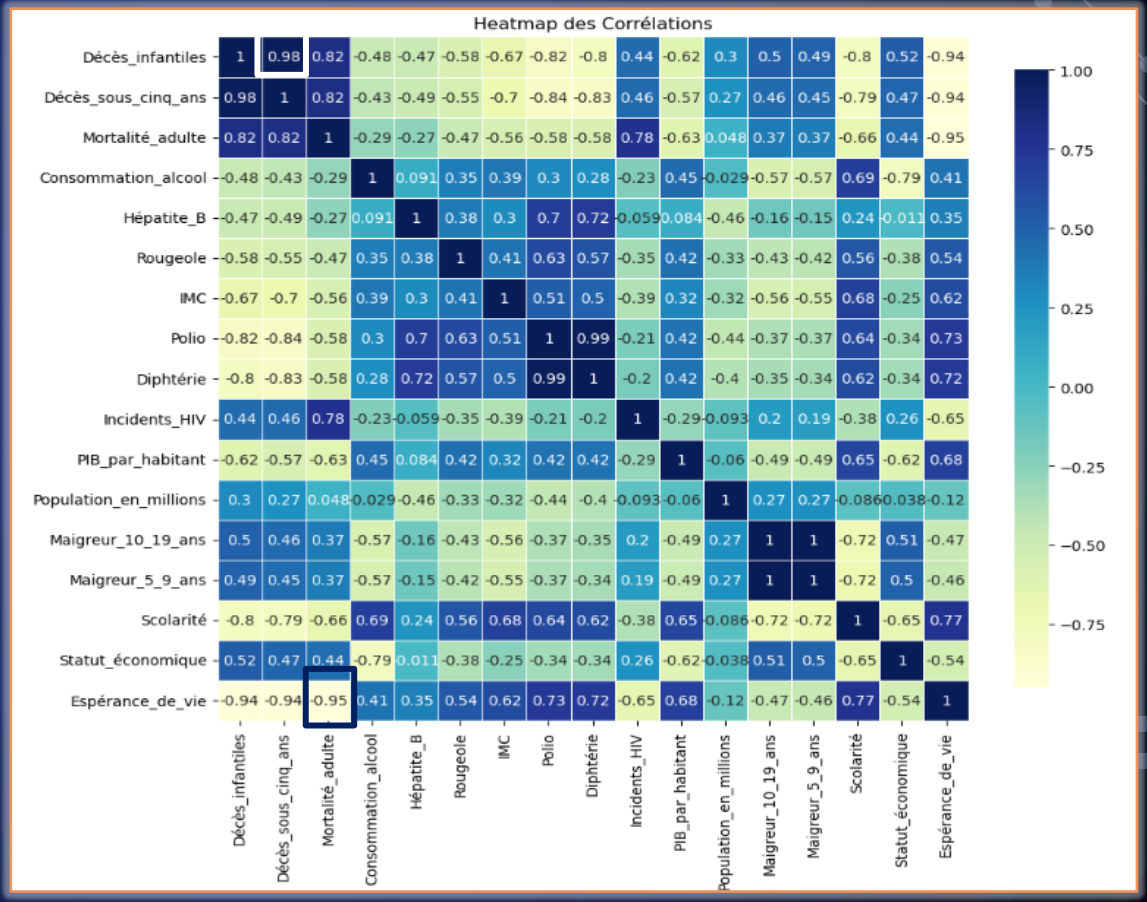
3. Analyse bivariées

- **Calcul et Affichage de la Matrice de Corrélation**

```
corr_matrix = data_clean.corr()
corr_matrix
```

La matrice de corrélation montre comment les variables d'un jeu de données sont liées entre elles, avec des valeurs entre -1 et 1

- 1 : Corrélation positive parfaite
- 0 : Pas de corrélation
- -1 : Corrélation négative parfaite



4. Analyse multivariée avec l'ACP

Construction de l'ACP

```
pca = PCA()
X_pca = pca.fit_transform(data_scaled)

#nombre de composantes calculées
print(pca.n_components_) #
```

```
17
```

- **X_pca** contient les nouvelles coordonnées des données dans l'espace des composantes principales.
- il s'agit des données originales **projetées** sur les nouvelles dimensions définies par les composantes principales.

4. Analyse multivariée avec l'ACP

Calcul des valeurs propres et des vecteurs propres

```
# Valeurs propres
eigenvalues = pca.explained_variance_
print("Valeurs propres :", eigenvalues)
eigenvalues

# Vecteurs propres
eigenvectors = pca.components_
print("Vecteurs propres :\n", eigenvectors)
```

- Les composantes principales sont les vecteurs propres de la matrice de covariance (ou de corrélation) des données.
- Ces composantes sont ordonnées par leur capacité à expliquer la variance totale des données.

4. Analyse multivariée avec l'ACP

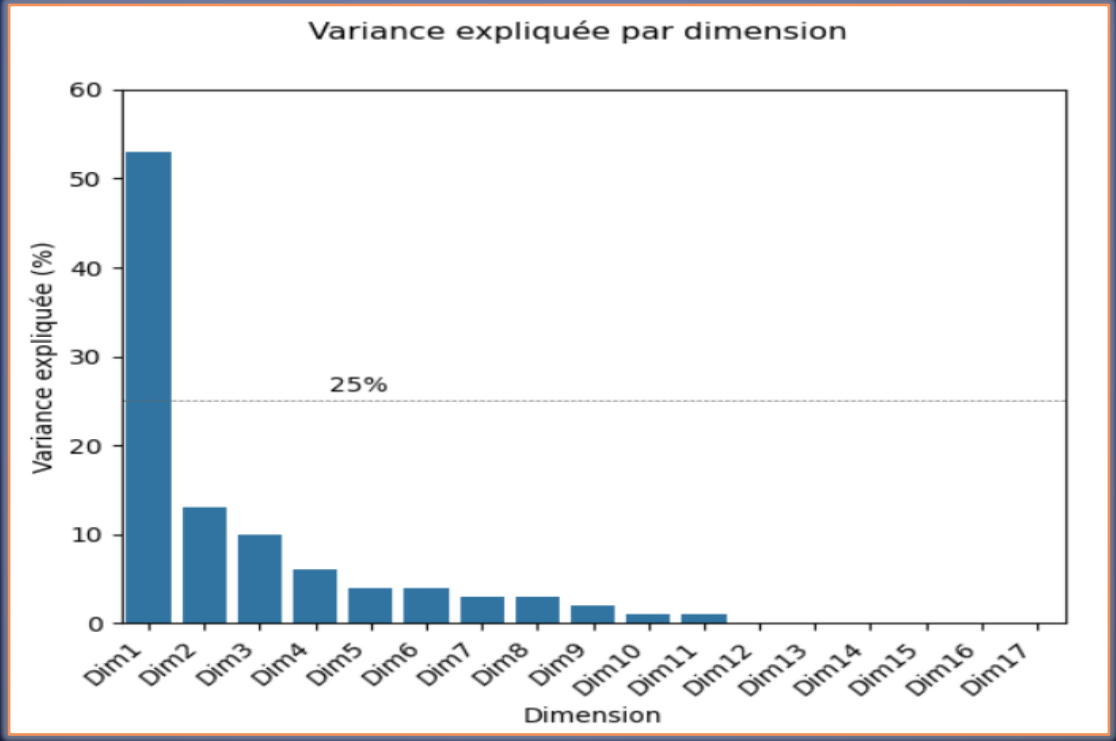
Tableau des Résultats de l'Analyse en Composantes Principales (ACP)

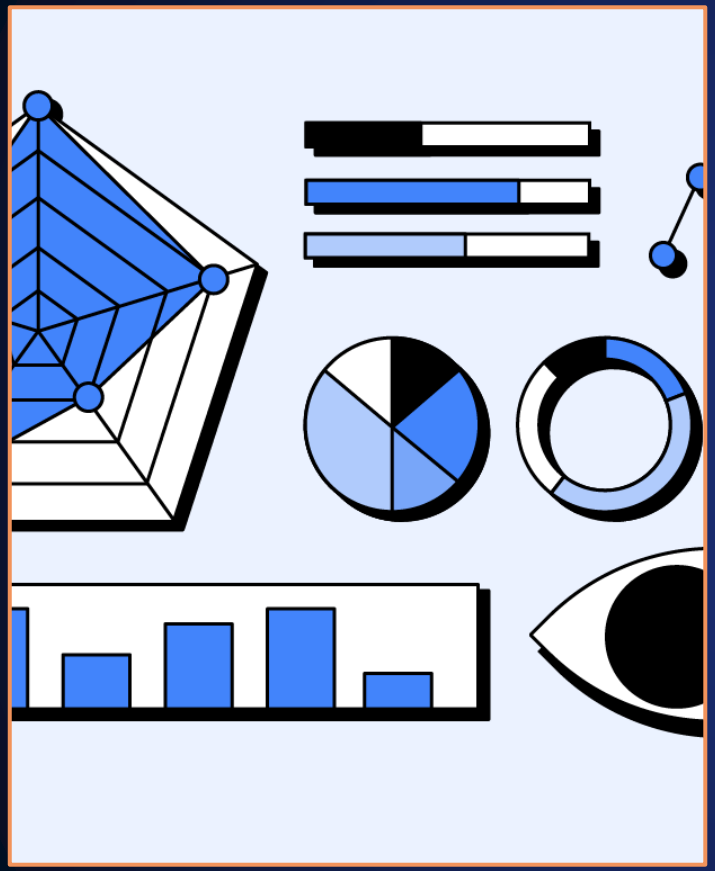
```
n_dim = len(pca.explained_variance_)
comp = pd.DataFrame(
    {
        "Dimension" : ["Dim" + str(x + 1) for x in range(n_dim)],
        "Valeur propre" : pca.explained_variance_,
        "% variance expliquée" : np.round(pca.explained_variance_ratio_ * 100),
        "% cum. var. expliquée" : np.round(np.cumsum(pca.explained_variance_ratio_) * 100)
    },
    columns = ["Dimension", "Valeur propre", "% variance expliquée", "% cum. var. expliquée"]
)
comp
```

	Dimension	Valeur propre	% variance expliquée	% cum. var. expliquée
0	Dim1	9.320976	53.0	53.0
1	Dim2	2.327425	13.0	67.0
2	Dim3	1.771847	10.0	77.0
3	Dim4	1.046859	6.0	83.0
4	Dim5	0.684410	4.0	87.0
5	Dim6	0.643238	4.0	90.0
6	Dim7	0.517276	3.0	93.0
7	Dim8	0.448682	3.0	96.0
8	Dim9	0.281001	2.0	97.0
9	Dim10	0.157787	1.0	98.0
10	Dim11	0.142912	1.0	99.0
11	Dim12	0.084059	0.0	100.0
12	Dim13	0.035523	0.0	100.0
13	Dim14	0.013552	0.0	100.0
14	Dim15	0.005195	0.0	100.0
15	Dim16	0.004700	0.0	100.0
16	Dim17	0.000271	0.0	100.0

4. Analyse multivariée avec l'ACP

Pourcentage d'information (d'inertie) expliqué par chaque axe

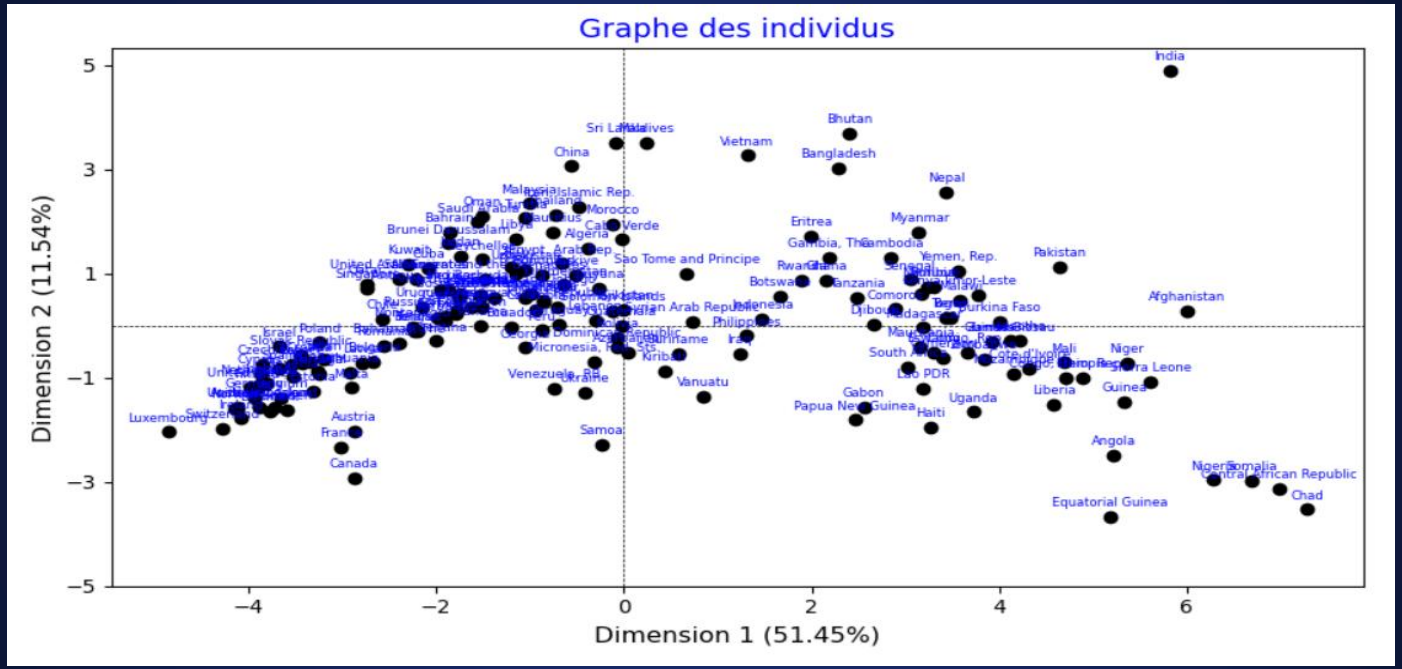




03 »

Visualisation et interprétation

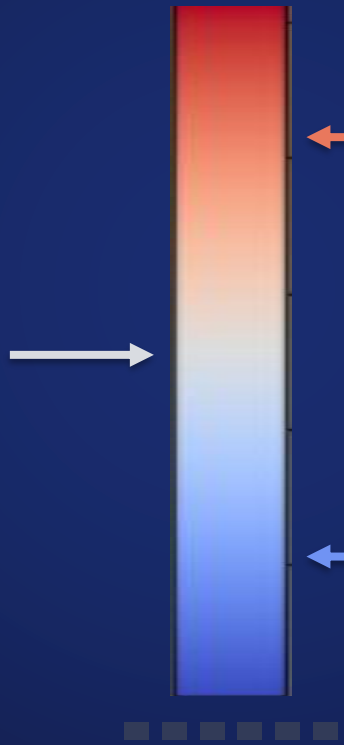
0. Graphe des individus de l'ACP



- Comment interpréter les axes principaux ?
- Quels éléments différencient les pays éloignés sur le graphique ?

➔ Besoin d'analyser les variables pour comprendre ces différences.

1. Palette de couleurs (ou échelle de couleurs graduée)

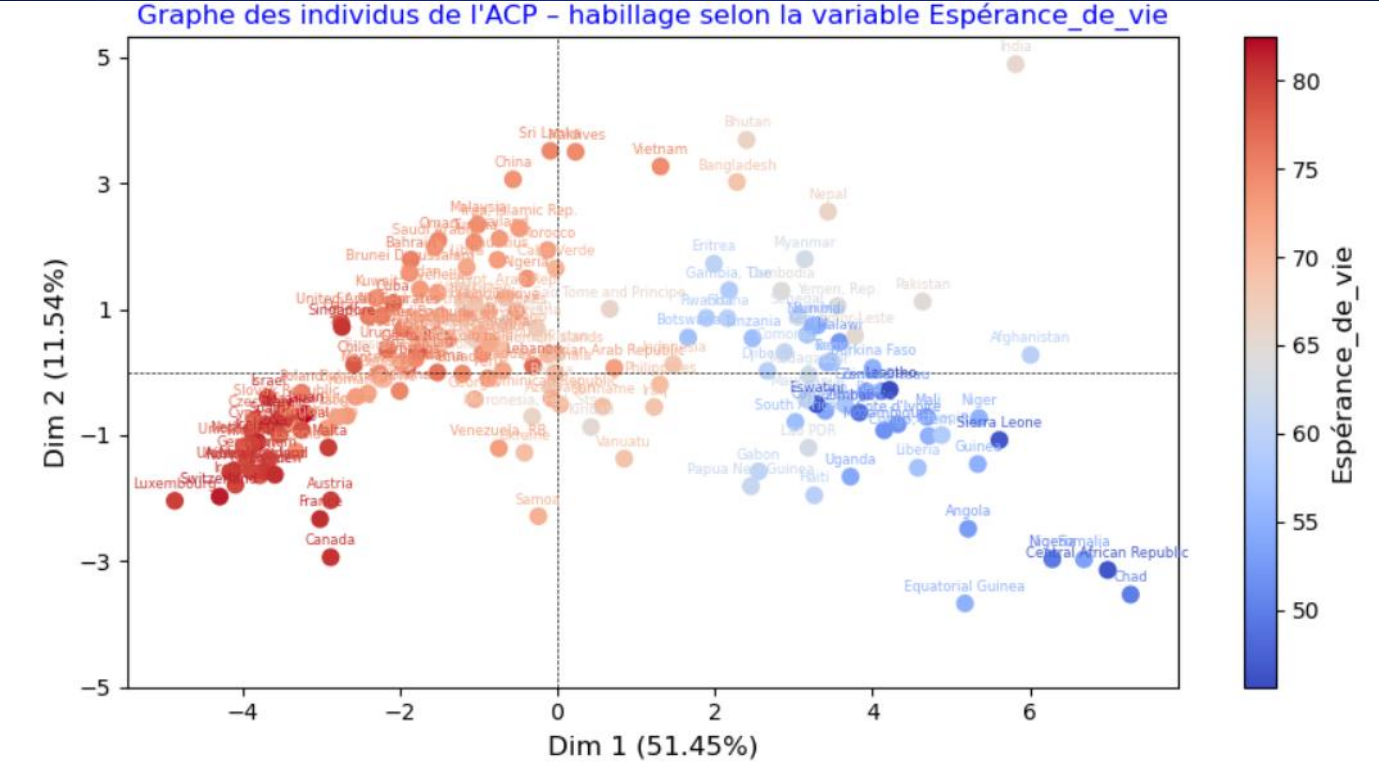


Rouge (en haut) : représente des valeurs **élevées**, positives ou maximales.

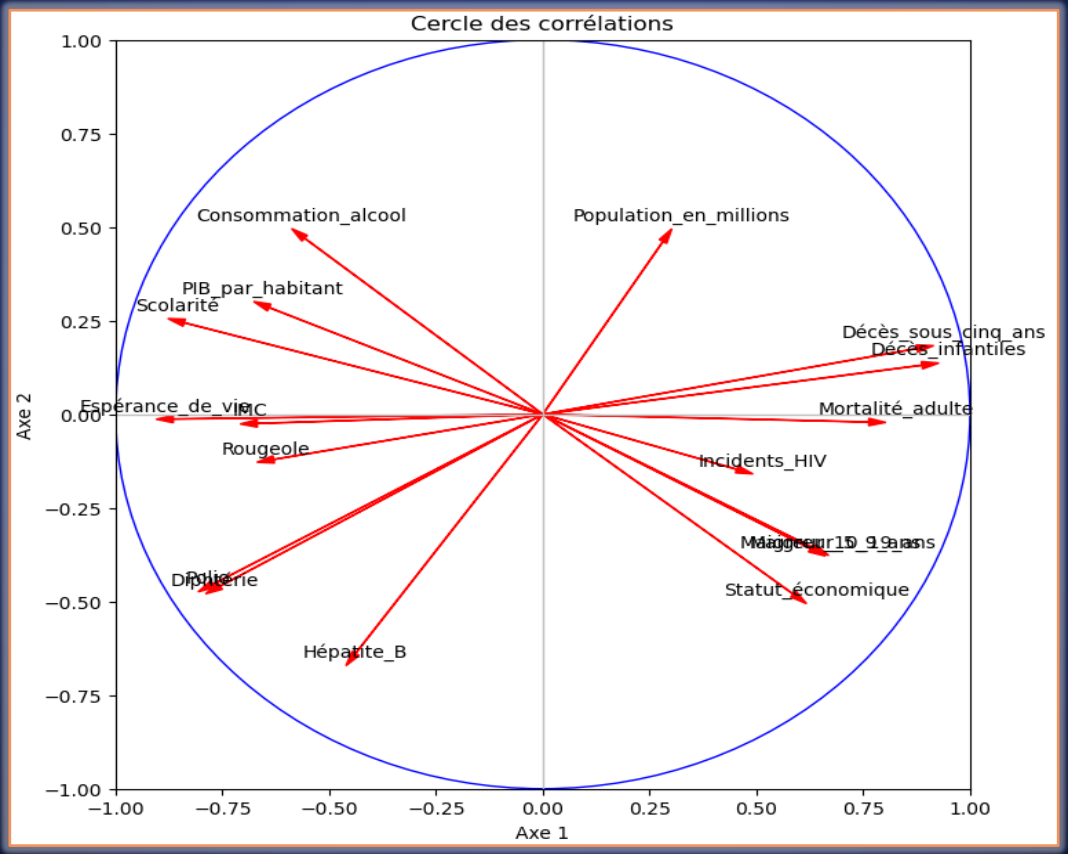
Blanc (au centre) : correspond à des valeurs intermédiaires ou neutres (souvent zéro).

Bleu (en bas) : indique des valeurs **basses**, négatives ou minimales.

2. Graphe des individus de l'ACP – habillage selon la variable Espérance_de_vie

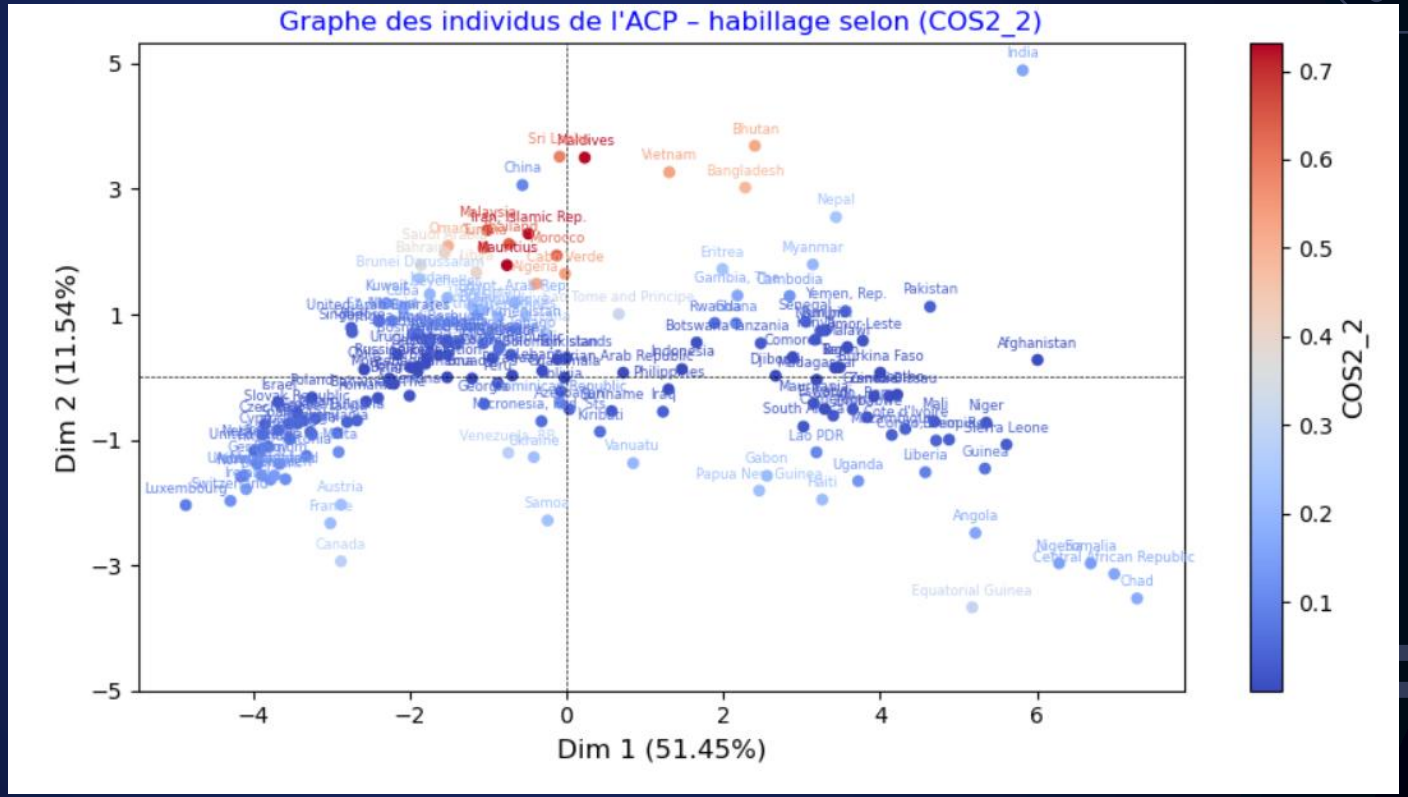


3. Cercle des corrélations



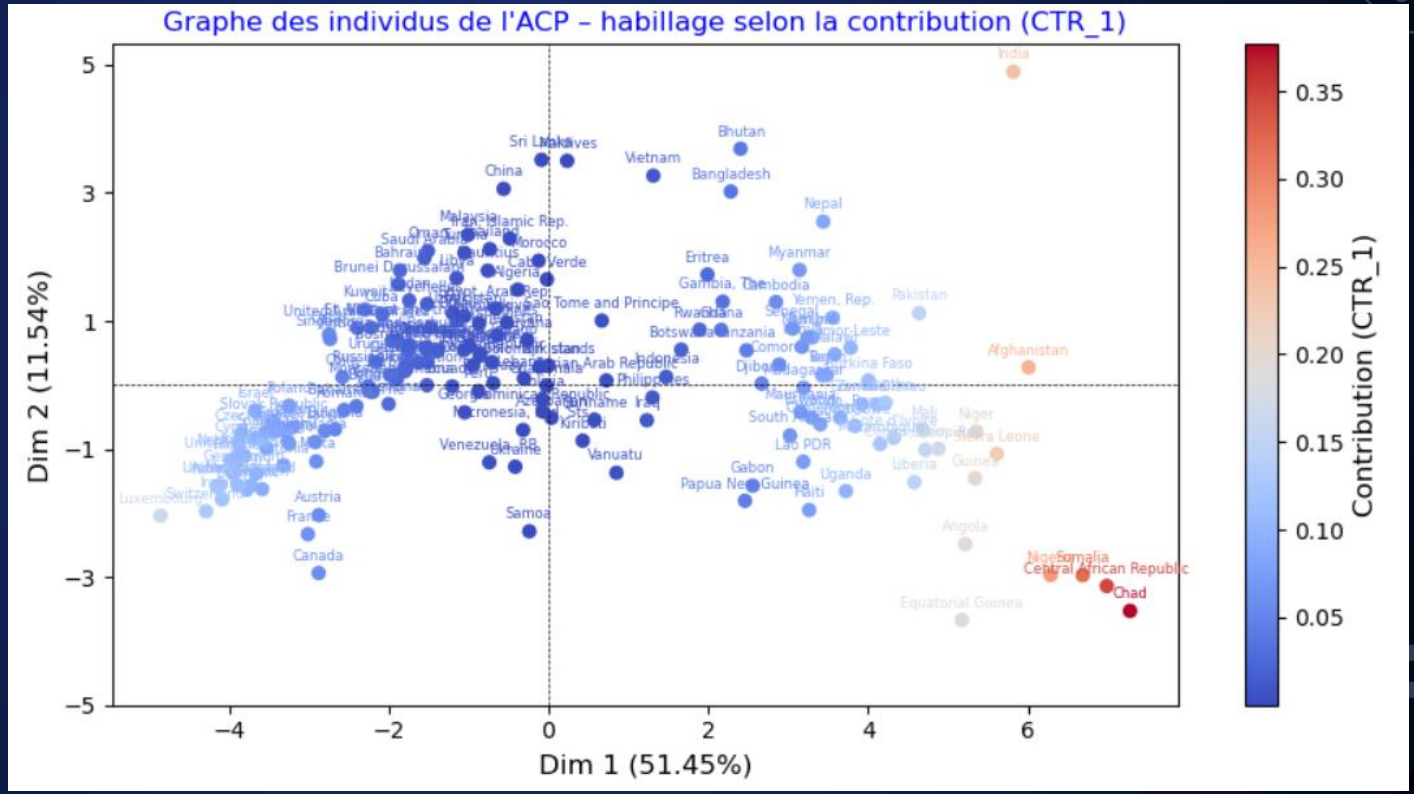
4. La qualité de représentation des individus selon l'axe 2

	id	COS2_1	COS2_2
0	Ireland	0.771496	0.105555
1	Belize	0.411764	0.116359
2	Grenada	0.487971	0.132397
3	Nepal	0.443391	0.093100
4	Iran, Islamic Rep.	0.025447	0.260719
5	Bolivia	0.006291	0.009752
6	Oman	0.314864	0.473475
7	Bulgaria	0.655102	0.024118
8	Guatemala	0.002210	0.023159
9	Suriname	0.017279	0.004710
10	Denmark	0.741963	0.118248
11	Cuba	0.595338	0.156745
12	Mozambique	0.722259	0.003007
13	Japan	0.365723	0.167533
14	Guinea	0.821979	0.060152
15	Belgium	0.812838	0.097735
16	Namibia	0.381655	0.076082
17	Turkmenistan	0.093508	0.220379
18	Guinea-Bissau	0.828810	0.000943
19	Niger	0.814504	0.025884
20	Singapore	0.512442	0.036116
21	Canada	0.348224	0.360274
22	Azerbaijan	0.004172	0.024969
23	Nigeria	0.612207	0.304869
24	Syrian Arab Republic	0.033401	0.000715
25	Bhutan	0.213865	0.447017
26	Romania	0.622561	0.015859
27	Guyana	0.055924	0.404736
28	Lithuania	0.610954	0.030051
29	Pakistan	0.525309	0.034511
30	Malawi	0.517326	0.060726
31	United Arab Emirates	0.503098	0.069022
32	Maldives	0.000009	0.602704
33	Serbia	0.612921	0.012564
34	Cameroon	0.761505	0.003014
35	Netherlands	0.844443	0.071039

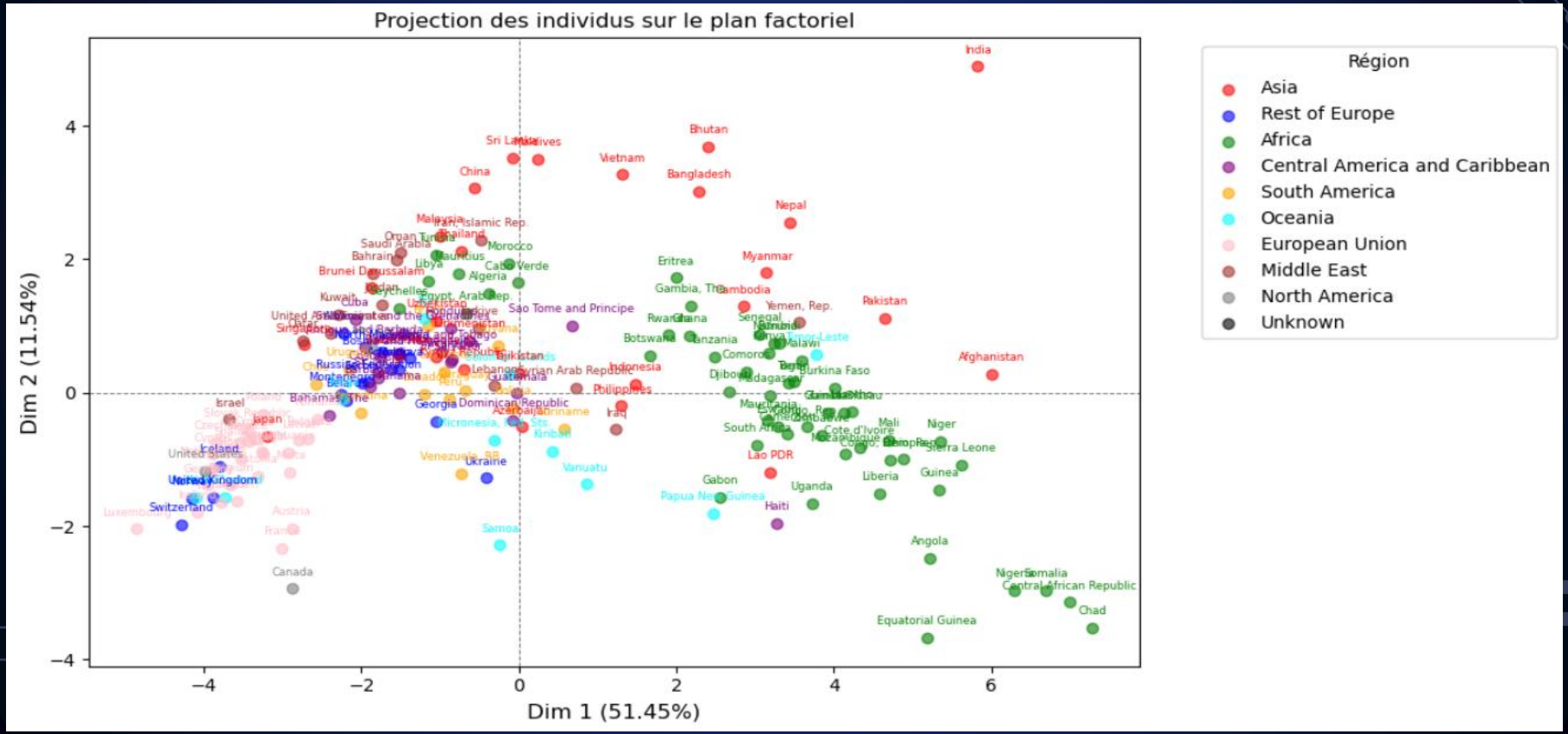


5. Graphe les individus qui contribuent fortement à la définition de l'axe 1

	id	CTR_1	CTR_2
0	Ireland	1.115745e-01	0.061136
1	Belize	2.238333e-02	0.025331
2	Grenada	1.676690e-02	0.018219
3	Nepal	6.557858e-02	0.055146
4	Iran, Islamic Rep.	1.384778e-03	0.056819
5	Bolivia	1.437190e-04	0.000892
6	Oman	1.806828e-02	0.108812
7	Bulgaria	4.707380e-02	0.006941
8	Guatemala	6.391971e-05	0.002682
9	Suriname	1.750985e-03	0.001911
10	Denmark	9.585581e-02	0.061181
11	Cuba	2.868826e-02	0.030250
12	Mozambique	1.158907e-01	0.001932
13	Japan	5.579279e-02	0.102356
14	Guinea	1.712643e-01	0.050193
15	Belgium	9.037263e-02	0.043518
16	Namibia	7.604646e-02	0.060712
17	Turkmenistan	3.511999e-03	0.033148
18	Guinea-Bissau	1.070189e-01	0.000487
19	Niger	1.740492e-01	0.022151
20	Singapore	5.003402e-02	0.014122
21	Canada	6.356987e-02	0.263398
22	Azerbaijan	1.493634e-04	0.003580
23	Nigeria	2.576403e-01	0.513824
24	Syrian Arab Republic	1.829948e-03	0.000157
25	Bhutan	2.945724e-02	0.246582
26	Romania	4.244924e-02	0.004331
27	Guyana	8.523390e-04	0.024704
28	Lithuania	5.695778e-02	0.011220
29	Pakistan	1.443178e-01	0.037971
30	Malawi	9.133982e-02	0.042939
31	United Arab Emirates	4.374435e-02	0.024035
32	Maldives	7.827588e-07	0.219606
33	Serbia	2.730347e-02	0.002241
34	Cameroon	7.426476e-02	0.001177
35	Netherlands	1.003092e-01	0.033795



6. Distribution des pays sur les axes principaux en fonction de leur région



04 »

Conclusion

Conclusion

L'ACP est une méthode puissante pour analyser des jeux de données complexes et multidimensionnels. Ce projet permet d'explorer les relations entre les variables et de simplifier les données tout en offrant des outils pour une meilleure prise de décision.