



La poule qui chante



Contexte

- L'objectif de l'entreprise ?

Exporter à l'international !



- Ma mission ?

Trouver les pays les plus intéressants pour exporter nos produits !

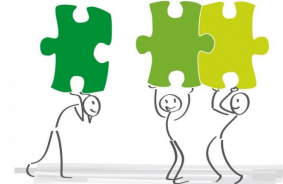


Les grandes lignes du projet

1) Choisir les données



2) Nettoyer et fusionner les données

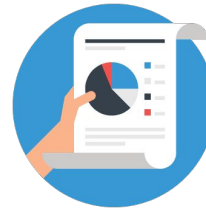


3) Analyser les données

- Création de groupes (Clustering hiérarchique, K-Means)
- Visualisation (ACP)



4) Choix du groupe





Outils

Langage de programmation :



Logiciel :



Choix des données





Choix des données

Quelles données choisir pour notre analyse ?

- **Indice consommation :**
Disponibilité de protéine (2020)
Disponibilité alimentaire (2020)
(Viande de poulet)
- **Indice économique :**
PIB par habitant (2020)
- **Indice politique :**
Stabilité politique (2020)
- **Indice population :**
Population total (2010 & 2020)
- **Indice commercial :**
Importation de viande de poulet (2020)
(en tonne)

Nettoyage et fusion des données





Nettoyage des données

Création d'une liste avec le nom des colonnes indésirable

```
# Liste des colonnes non nécessaires  
drop_columns = ['Code Domaine', 'Code Élément', 'Code année', 'Symbole', 'Description du Symbole', 'Code Produit']
```




Nettoyage des données

Fichier Disponibilité

Domaine	Code zone (M49)	Zone	Élément	Produit	Année	Unité	Valeur	
0	Compte Disponibilités et Utilisations (CDU) (2...	4	Afghanistan	Production	Viande, poulet, fraîche ou réfrigérée	2020	tonnes	27938.44
1	Compte Disponibilités et Utilisations (CDU) (2...	4	Afghanistan	Importations - Quantité	Viande, poulet, fraîche ou réfrigérée	2020	tonnes	35616.91
2	Compte Disponibilités et Utilisations (CDU) (2...	4	Afghanistan	Exportations - Quantité	Viande, poulet, fraîche ou réfrigérée	2020	tonnes	0.24
3	Compte Disponibilités et Utilisations (CDU) (2...	4	Afghanistan	Disponibilité alimentaire (Kcal/personne/jour)	Viande, poulet, fraîche ou réfrigérée	2020	Kcal/personne/jour	5.90
4	Compte Disponibilités et Utilisations (CDU) (2...	4	Afghanistan	Disponibilité de protéines en quantité (g/pers...	Viande, poulet, fraîche ou réfrigérée	2020	g/personne/jour	0.60
...
856	Compte Disponibilités et Utilisations (CDU) (2...	894	Zambie	Disponibilité de protéines en quantité (g/pers...	Viande, poulet, fraîche ou réfrigérée	2020	g/personne/jour	0.89
857	Compte Disponibilités et Utilisations (CDU) (2...	716	Zimbabwe	Production	Viande, poulet, fraîche ou réfrigérée	2020	tonnes	67211.11
858	Compte Disponibilités et Utilisations (CDU) (2...	716	Zimbabwe	Importations - Quantité	Viande, poulet, fraîche ou réfrigérée	2020	tonnes	1288.10
859	Compte Disponibilités et Utilisations (CDU) (2...	716	Zimbabwe	Disponibilité alimentaire (Kcal/personne/jour)	Viande, poulet, fraîche ou réfrigérée	2020	Kcal/personne/jour	15.40
860	Compte Disponibilités et Utilisations (CDU) (2...	716	Zimbabwe	Disponibilité de protéines en quantité (g/pers...	Viande, poulet, fraîche ou réfrigérée	2020	g/personne/jour	1.55

- **Données :**
Disponibilité alimentaire
Disponibilité de protéine
Importation
- **Exploration :**
186 pays différents
- **Nettoyage :**
Suppression des colonnes indésirable
Suppression des doublons



Nettoyage des données

Fichier Sécurité

	Domaine	Code zone (M49)	Zone	Élément	Produit	Année	Unité	Valeur
0	Données de la sécurité alimentaire	4	Afghanistan	Valeur	PIB par habitant, (\$ PPA internationaux consta...	2020	I\$	1970.60
1	Données de la sécurité alimentaire	4	Afghanistan	Valeur	Stabilité politique et absence de violence/ter...	2010	indice	-2.58
2	Données de la sécurité alimentaire	4	Afghanistan	Valeur	Stabilité politique et absence de violence/ter...	2020	indice	-2.73
3	Données de la sécurité alimentaire	710	Afrique du Sud	Valeur	PIB par habitant, (\$ PPA internationaux consta...	2020	I\$	12665.80
4	Données de la sécurité alimentaire	710	Afrique du Sud	Valeur	Stabilité politique et absence de violence/ter...	2010	indice	-0.03
...
578	Données de la sécurité alimentaire	894	Zambie	Valeur	Stabilité politique et absence de violence/ter...	2010	indice	0.52
579	Données de la sécurité alimentaire	894	Zambie	Valeur	Stabilité politique et absence de violence/ter...	2020	indice	-0.13
580	Données de la sécurité alimentaire	716	Zimbabwe	Valeur	PIB par habitant, (\$ PPA internationaux consta...	2020	I\$	3353.40
581	Données de la sécurité alimentaire	716	Zimbabwe	Valeur	Stabilité politique et absence de violence/ter...	2010	indice	-1.10
582	Données de la sécurité alimentaire	716	Zimbabwe	Valeur	Stabilité politique et absence de violence/ter...	2020	indice	-1.08

- **Données :**
PIB par habitant
Stabilité politique

- **Exploration :**
200 pays différents

- **Nettoyage :**
Suppression des colonnes
indésirable
Suppression des doublons



Nettoyage des données

Fichier Population

Domaine	Code zone (M49)	Zone	Élément	Produit	Année	Unité	Valeur
0 Séries temporelles annuelles	4	Afghanistan	Population totale	Population-Estimations	2010	1000 personnes	28189.672
1 Séries temporelles annuelles	4	Afghanistan	Population totale	Population-Estimations	2020	1000 personnes	38972.230
2 Séries temporelles annuelles	710	Afrique du Sud	Population totale	Population-Estimations	2010	1000 personnes	51784.921
3 Séries temporelles annuelles	710	Afrique du Sud	Population totale	Population-Estimations	2020	1000 personnes	58801.927
4 Séries temporelles annuelles	8	Albanie	Population totale	Population-Estimations	2010	1000 personnes	2913.399
...
460 Séries temporelles annuelles	887	Yémen	Population totale	Population-Estimations	2020	1000 personnes	32284.046
461 Séries temporelles annuelles	894	Zambie	Population totale	Population-Estimations	2010	1000 personnes	13792.086
462 Séries temporelles annuelles	894	Zambie	Population totale	Population-Estimations	2020	1000 personnes	18927.715
463 Séries temporelles annuelles	716	Zimbabwe	Population totale	Population-Estimations	2010	1000 personnes	12839.771
464 Séries temporelles annuelles	716	Zimbabwe	Population totale	Population-Estimations	2020	1000 personnes	15669.666

- **Données :**
Population total

- **Exploration :**
237 pays différents

- **Nettoyage :**
Suppression des colonnes
indésirable
Suppression des doublons



Fusion des données

Concaténation + pivot

Élément	Zone	Disponibilité alimentaire (Kcal/personne/jour)	Disponibilité de protéines en quantité (g/personne/jour)	Exportations - Quantité	Importations - Quantité	Population totale		Production	Valeur		
Produit		Viande, poulet, fraîche ou réfrigérée	Viande, poulet, fraîche ou réfrigérée	Viande, poulet, fraîche ou réfrigérée	Viande, poulet, fraîche ou réfrigérée	Population- Estimations		Viande, poulet, fraîche ou réfrigérée	PIB par habitant, (\$ PPA internationaux constants de 2017)	Stabilité politique et absence de violence/terrorisme (indice)	
Année		2020	2020	2020	2020	2010	2020	2020	2020	2010	2020
0	Afghanistan	5.90	0.60	0.24	35616.91	28189.672	38972.230	27938.44	1970.6	-2.58	-2.73
1	Afrique du Sud	138.50	13.57	45619.27	398923.26	51784.921	58801.927	1873238.00	12665.8	-0.03	-0.24
2	Albanie	64.98	4.59	NaN	23503.19	2913.399	2866.849	12233.91	13253.7	-0.19	0.08
3	Algérie	20.44	1.75	568.50	0.49	35856.344	43451.666	260268.79	10735.4	-1.26	-0.86
4	Allemagne	39.69	4.74	254316.47	470860.36	81325.090	83328.988	1066000.00	51423.2	0.80	0.67
...
232	Îles Salomon	14.52	1.46	NaN	3732.02	540.394	691.191	279.10	2482.4	0.34	0.64
233	Îles Turques-et-Caïques	NaN	NaN	NaN	NaN	29.726	44.276	NaN	NaN	NaN	NaN
234	Îles Vierges américaines	NaN	NaN	NaN	NaN	106.142	100.442	NaN	NaN	NaN	NaN
235	Îles Vierges britanniques	NaN	NaN	NaN	NaN	27.556	30.910	NaN	NaN	NaN	NaN
236	Îles Wallis-et-Futuna	NaN	NaN	NaN	NaN	13.142	11.655	NaN	NaN	NaN	NaN



Fusion des données

Nouveau DataFrame

	pays	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation
0	Afghanistan	5.90	0.60	1.382500	1970.6	-2.73	35616.91
1	Afrique du Sud	138.50	13.57	1.135503	12665.8	-0.24	398923.26
2	Albanie	64.98	4.59	0.984022	13253.7	0.08	23503.19
3	Algérie	20.44	1.75	1.211826	10735.4	-0.86	0.49
4	Allemagne	39.69	4.74	1.024641	51423.2	0.67	470860.36
...
232	Îles Salomon	14.52	1.46	1.279050	2482.4	0.64	3732.02
233	Îles Turques-et-Caïques	NaN	NaN	1.489470	NaN	NaN	NaN
234	Îles Vierges américaines	NaN	NaN	0.946298	NaN	NaN	NaN
235	Îles Vierges britanniques	NaN	NaN	1.121716	NaN	NaN	NaN
236	Îles Wallis-et-Futuna	NaN	NaN	0.886851	NaN	NaN	NaN



Nettoyage des données

	pays	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation
0	Afghanistan	5.90	0.60	1.382500	1970.6	-2.73	35616.91
1	Afrique du Sud	138.50	13.57	1.135503	12665.8	-0.24	398923.26
2	Albanie	64.98	4.59	0.984022	13253.7	0.08	23503.19
3	Algérie	20.44	1.75	1.211826	10735.4	-0.86	0.49
4	Allemagne	39.69	4.74	1.024641	51423.2	0.67	470860.36
...
165	Égypte	46.07	4.72	1.231658	11951.4	-1.21	19675.01
166	Émirats arabes unis	103.12	10.40	1.094970	63299.4	0.63	474939.04
167	États-Unis d'Amérique	205.19	17.71	1.079565	59755.2	-0.02	66140.69
168	Éthiopie	1.98	0.20	1.313243	2296.9	-1.74	451.20
169	Îles Salomon	14.52	1.46	1.279050	2482.4	0.64	3732.02

- Suppression des lignes aux valeurs manquantes
- Suppression des données de la France

```
df = df.drop(df.loc[df['pays'] == "France"].index)
```

Analyse des données

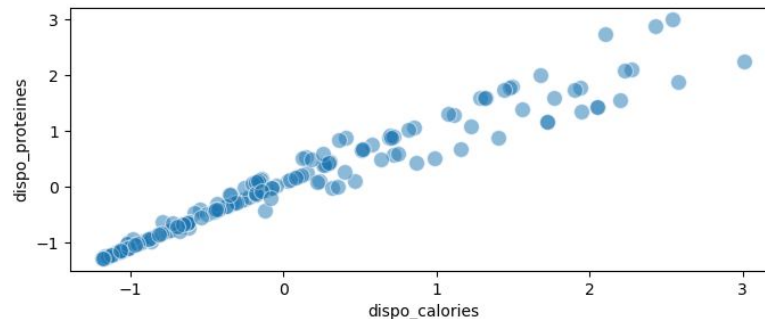
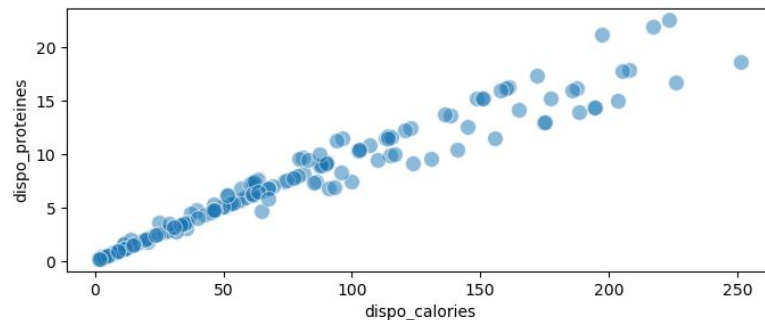




Analyse des données

Centrage et réduction

	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation
0	-1.111768	-1.196081	1.614546	-0.892507	-2.949617	-0.202899
1	1.112672	1.285322	-0.123926	-0.344630	-0.202078	2.733277
2	-0.120668	-0.432720	-1.190112	-0.314514	0.151020	-0.300800
3	-0.867852	-0.976065	0.413271	-0.443518	-0.886204	-0.490745
4	-0.544922	-0.404022	-0.904222	1.640771	0.802043	3.314660
...
164	-0.437894	-0.407848	0.552854	-0.381227	-1.272404	-0.331739
165	0.519152	0.678842	-0.409210	2.249146	0.757906	3.347623
166	2.231434	2.077381	-0.517643	2.067589	0.040677	0.043789
167	-1.177529	-1.272609	1.127084	-0.875792	-1.857222	-0.487102
168	-0.967163	-1.031547	0.886420	-0.866289	0.768940	-0.460587



Analyse des données

Classification ascendante hiérarchique

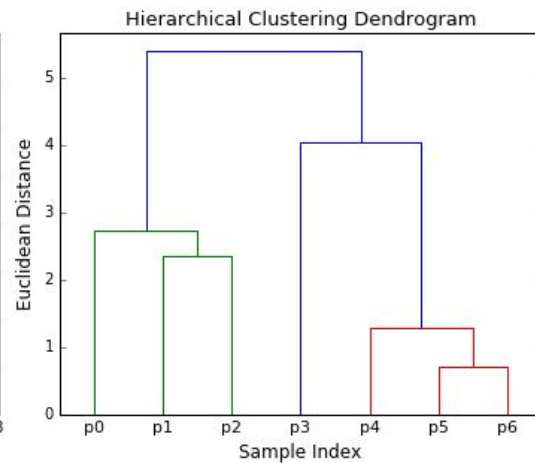
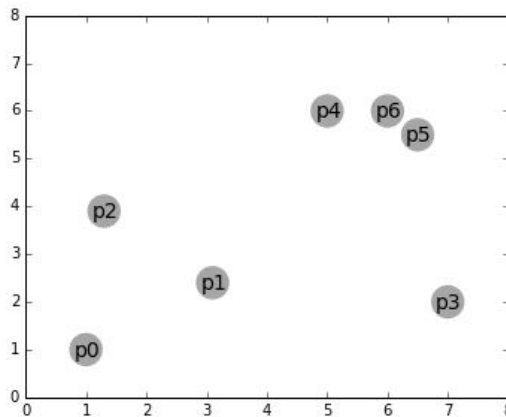




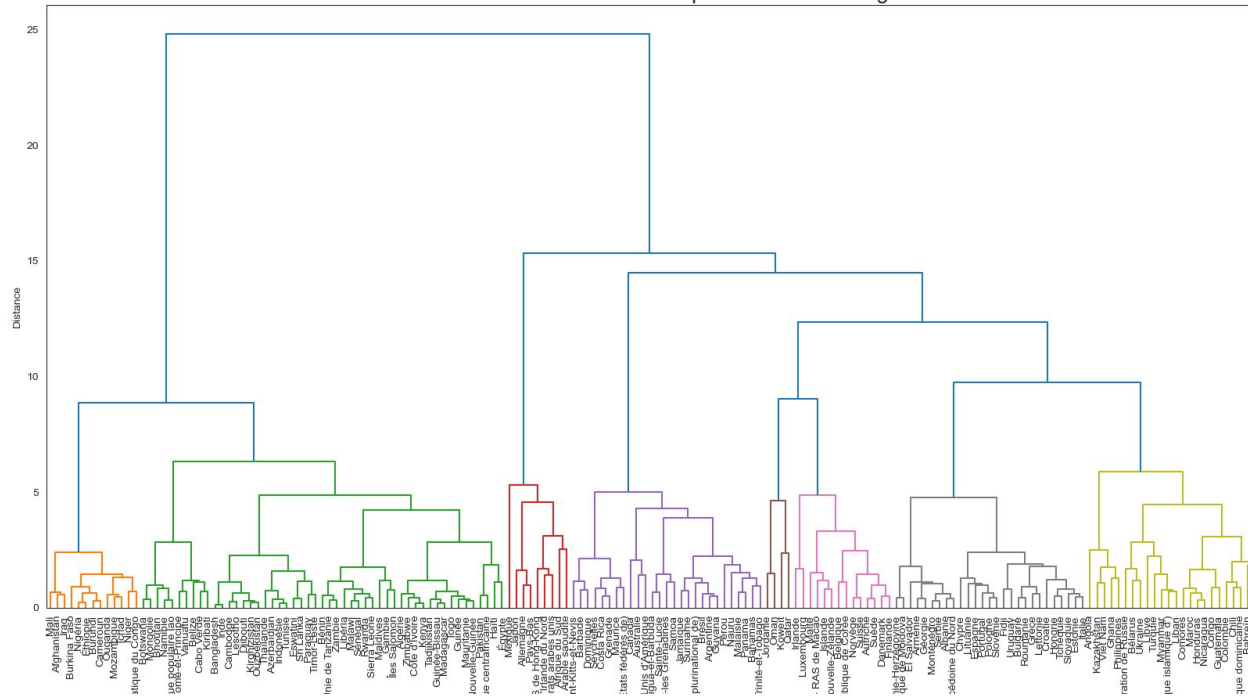
Classification ascendant hiérarchique

Définitions

*Méthode de clustering,
qui permet de regrouper des
données similaires en groupe
grâce à leur proximité !*



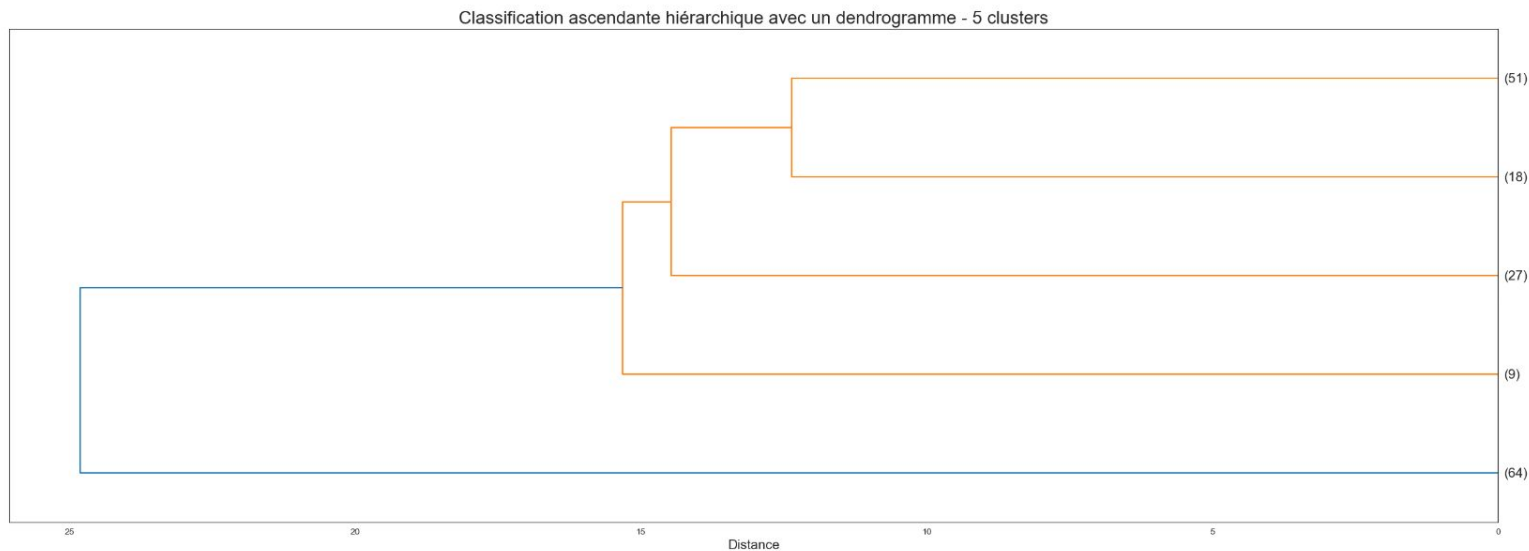
Classification ascendante hiérarchique avec un dendrogramme





Analyse des données

Dendrogramme (Clustering en 5 groupes)

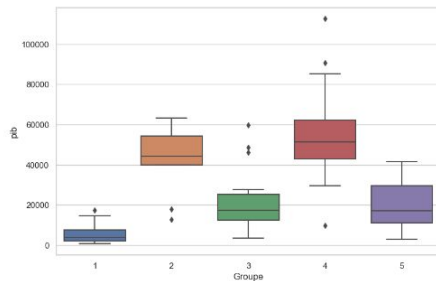
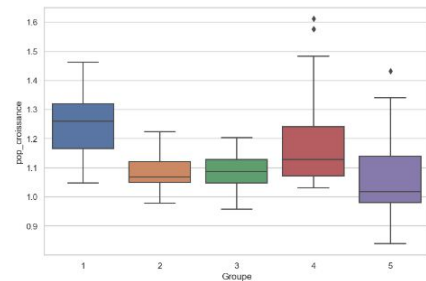
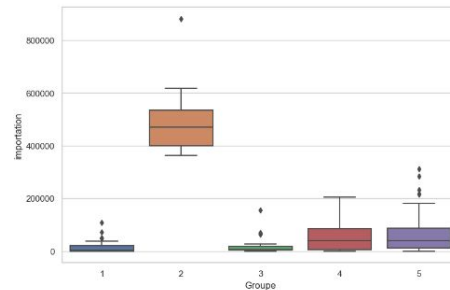
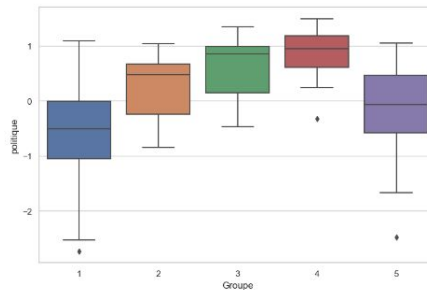
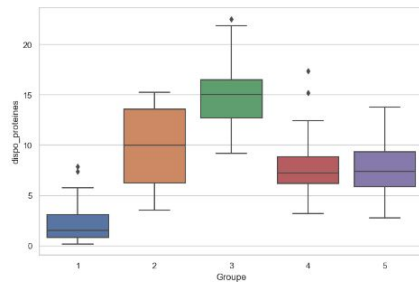
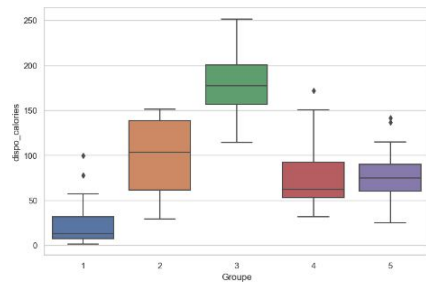


Groupe 1 : 64 pays / Groupe 2 : 9 pays / Groupe 3 : 27 pays / Groupe 4 : 18 pays /
Groupe 5 : 51 pays



Analyse des données

Comparaison des groupes par Boxplot



- **dispo calorie & dispo protéine** : Groupe 2 et 3
- **pop_croissance** : Groupe 1 et 4
- **pib** : Groupe 2 et 4
- **politique** : Groupe 3 et 4
- **importation** : Groupe 2

Analyse des données

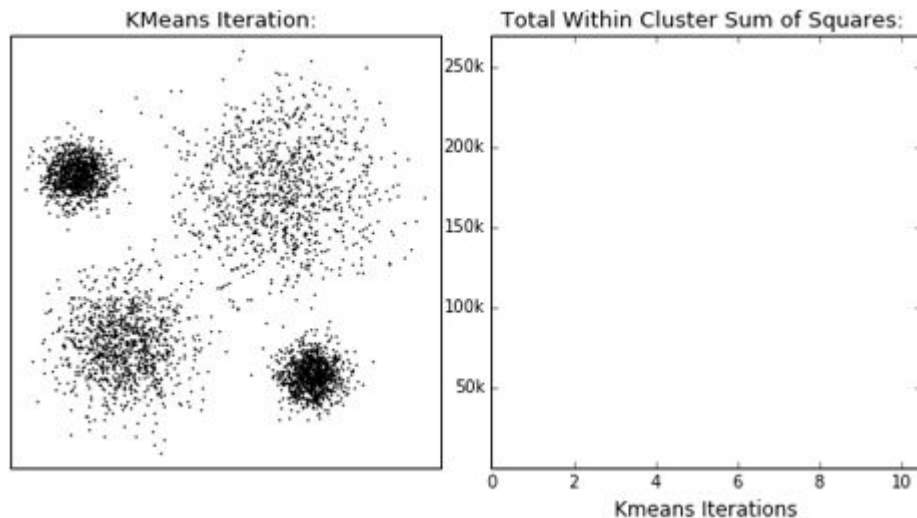
K-Means



K-Means

Définitions

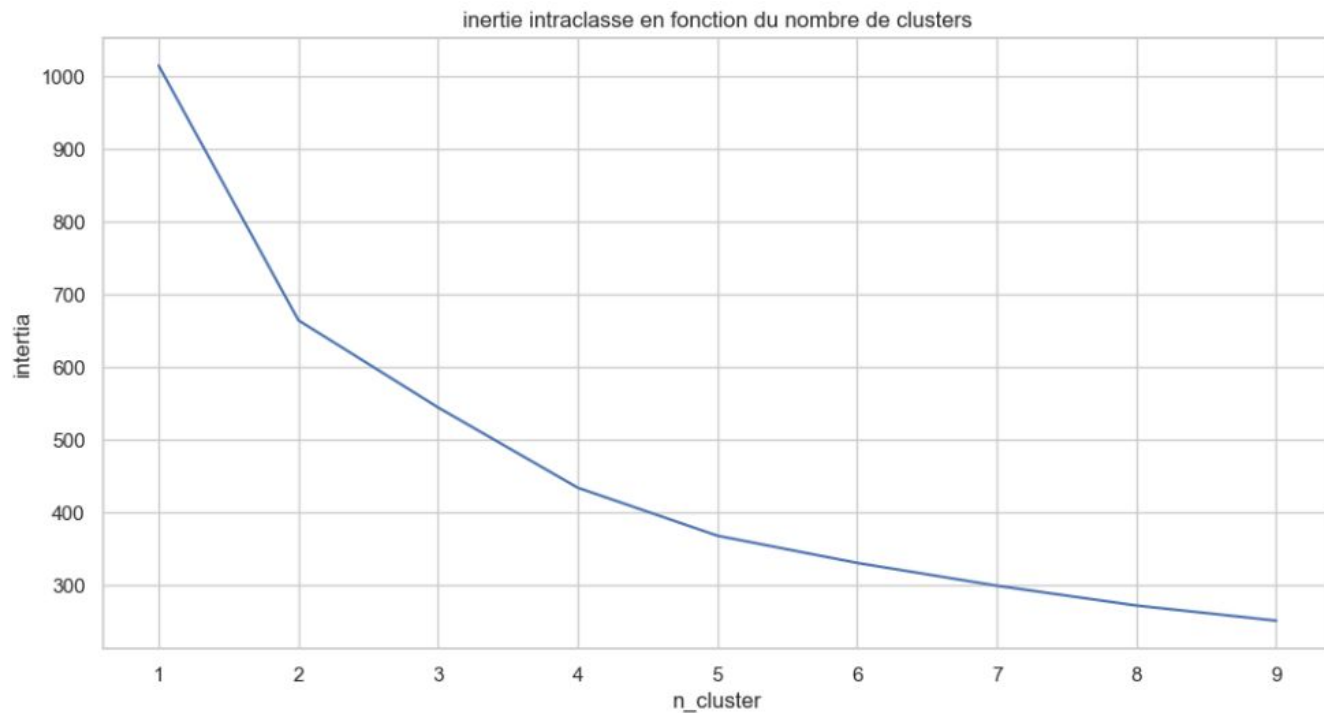
*Méthode de clustering,
qui permet de regrouper des
données similaires en groupe
grâce aux centroïdes !*





Analyse des données

Choix du nombre de clusters





Analyse des données

Définition du cluster pour chaque pays

```
kmeans = KMeans(n_clusters=5)  
kmeans.fit(df_std)
```

pays	
cluster	
a	55
b	32
c	44
d	29
e	9

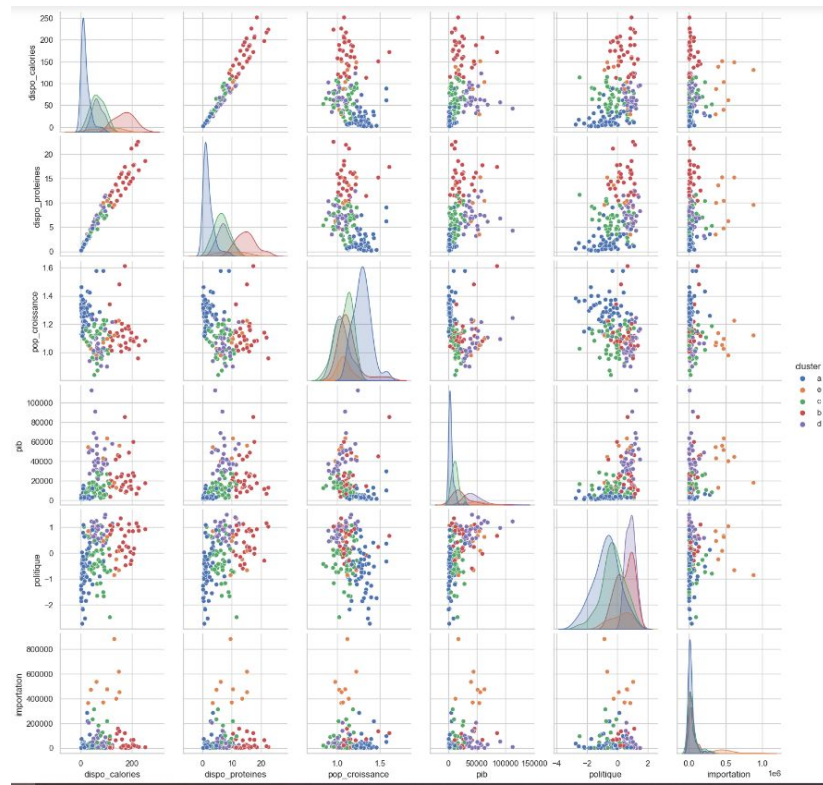
cluster		pays
0	a	Afghanistan
1	e	Afrique du Sud
2	c	Albanie
3	a	Algérie
4	e	Allemagne
...
164	a	Égypte
165	e	Émirats arabes unis
166	b	États-Unis d'Amérique
167	a	Éthiopie
168	a	Îles Salomon



Analyse des données

Comparaison des groupe par nuage de point

- dispo calorie & dispo protéine : Groupe **rouge**
- pop_croissance : Groupe **bleu**
- pib : Groupe **violet** et **orange**
- politique : Groupe **rouge** et **violet**
- importation : Groupe **orange**



Analyse des données

Analyse en Composantes Principales (ACP)

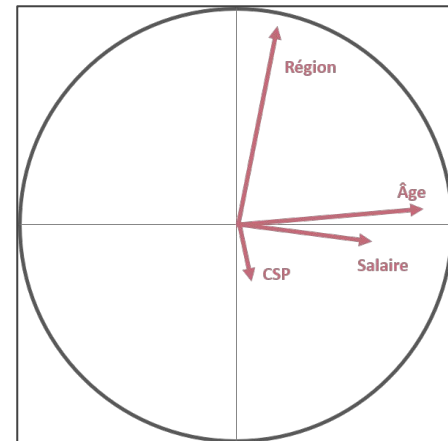
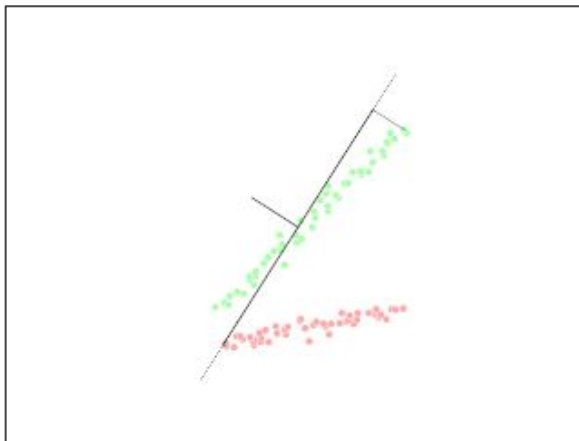




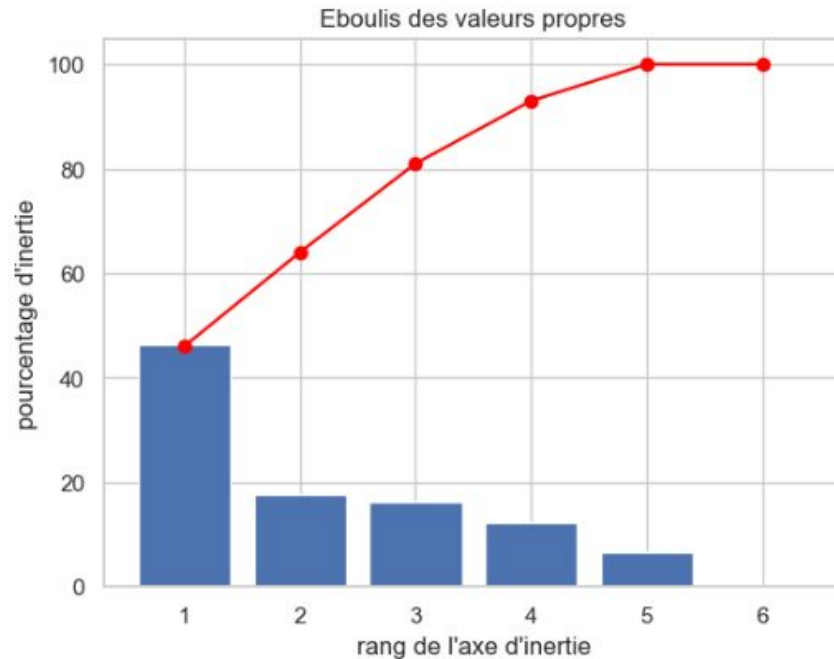
Analyse en Composantes Principales (ACP)

Définitions

*Méthode de réduction de dimension,
qui permet de résumer
l'information grâce aux
composantes principales !*



Analyse des données

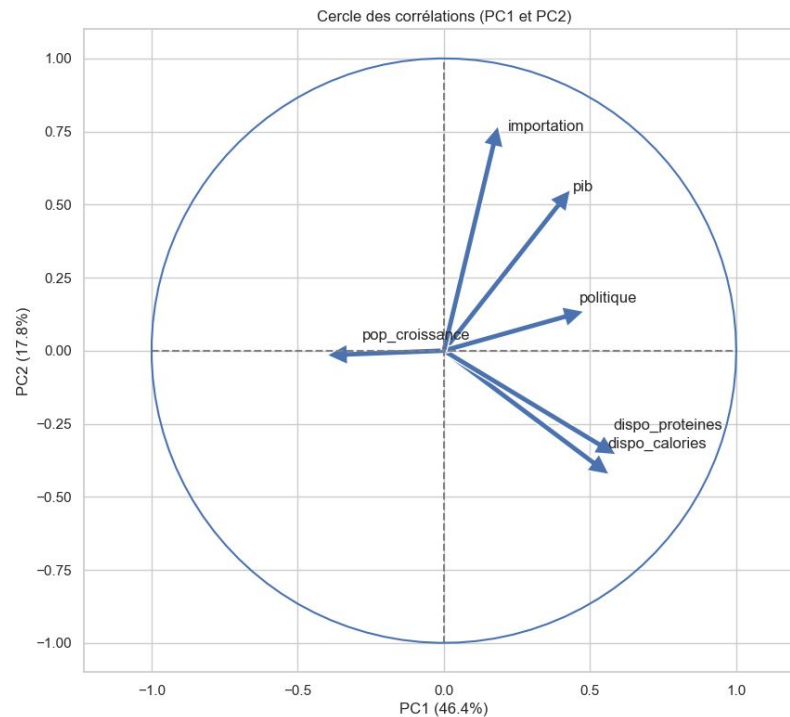


	PC1	PC2	PC3	PC4	PC5	PC6
dispo_calories	0.509658	-0.381894	0.334138	0.032651	0.000298	-0.694040
dispo_proteines	0.529585	-0.319843	0.310991	0.051999	0.060124	0.717081
pop_croissance	-0.330138	-0.012817	0.423698	0.828451	-0.157914	0.007511
piib	0.390079	0.495964	-0.197308	0.386581	0.639934	-0.062985
politique	0.412399	0.115763	-0.514210	0.299449	-0.680012	0.005375
importation	0.169005	0.701649	0.555674	-0.266054	-0.315462	-0.007103



Analyse des données

Cercle des corrélations

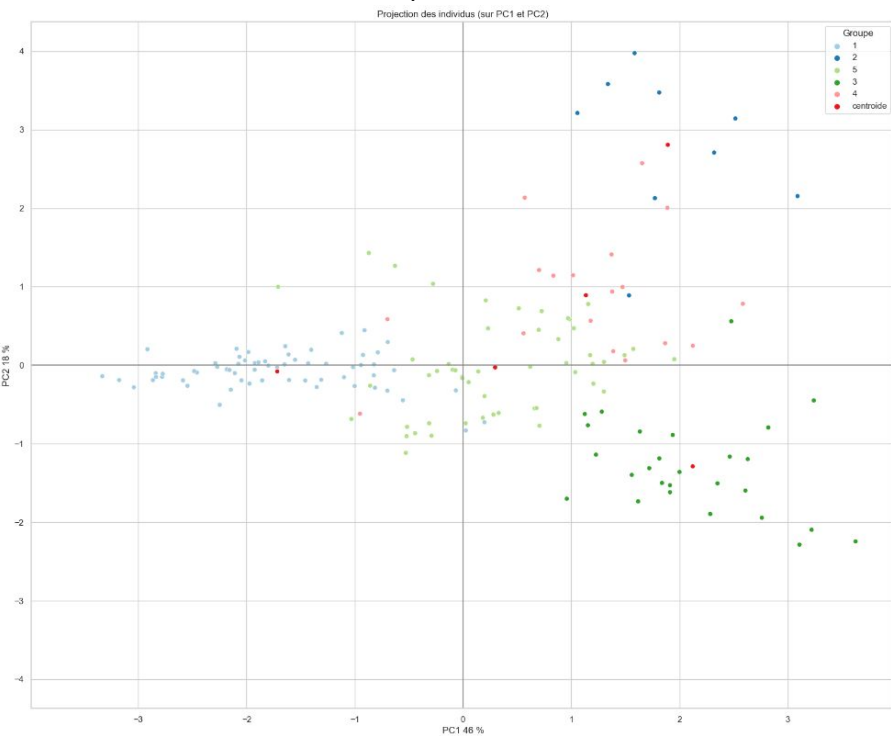




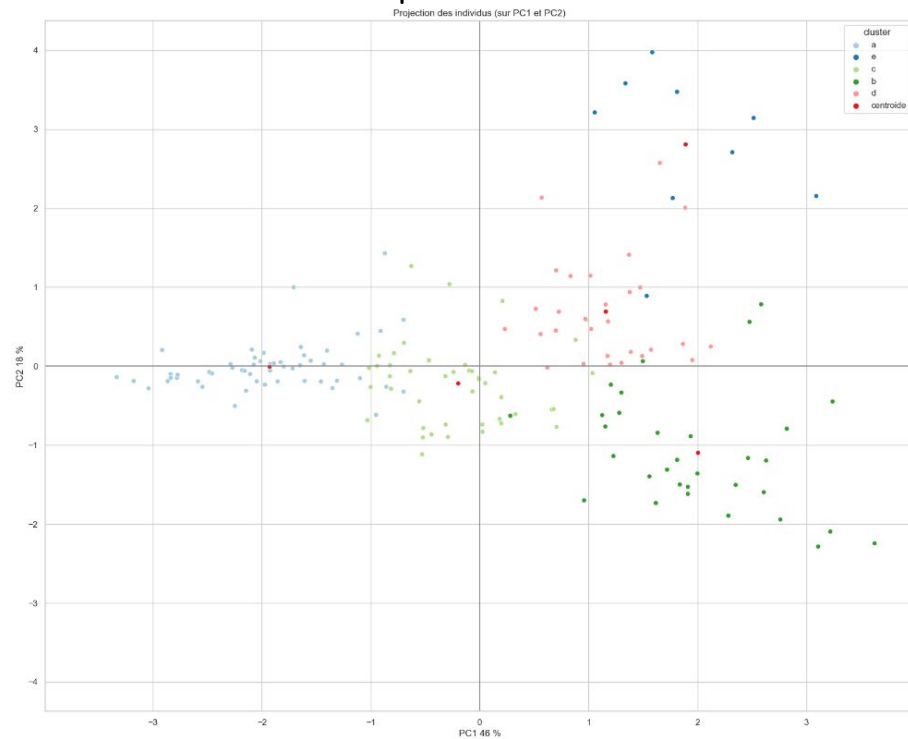
Analyse des données

Visualisation des données

Groupe : CAH



Groupe : K-Means



Choix du groupe





Choix du groupe

Groupe de pays intéressant

CAH :

	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation
Groupe						
1	20.601094	2.040781	1.244993	5264.589063	-0.553281	12518.246719
2	98.880000	9.820000	1.082856	42512.700000	0.222222	506905.343333
3	175.718519	15.112222	1.089664	20331.088889	0.575556	20073.403333
4	75.927778	8.101667	1.217942	55257.661111	0.860556	55270.762778
5	76.034510	7.550980	1.060911	19889.388235	-0.141765	65920.304510

K-Means :

	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation
cluster						
a	18.013273	1.852545	1.290747	4728.449091	-0.714909	26895.270000
b	170.558125	14.880000	1.131424	23693.218750	0.495312	32692.577500
c	65.825455	6.342045	1.078210	12626.877273	-0.291364	43696.905455
d	67.669655	7.326552	1.051448	45553.086207	0.851034	43168.716207
e	98.880000	9.820000	1.082856	42512.700000	0.222222	506905.343333



Première analyse

Les pays intéressants

	pays	dispo_calories	dispo_proteines	pop_croissance	pib	politique	importation	Groupe
1	Afrique du Sud	138.50	13.57	1.135503	12665.8	-0.24	398923.26	2
4	Allemagne	39.69	4.74	1.024641	51423.2	0.67	470860.36	2
7	Arabie saoudite	148.51	15.20	1.223895	44349.5	-0.66	617401.57	2
34	Chine - RAS de Hong-Kong	151.04	15.23	1.051668	55917.6	0.09	451454.48	2
74	Japon	61.67	6.22	0.977669	39935.5	1.04	535031.44	2
98	Mexique	130.79	9.57	1.119662	17878.3	-0.85	881419.27	2
118	Pays-Bas	29.14	3.48	1.049193	54324.4	0.85	363525.28	2
125	Royaume-Uni de Grande-Bretagne et d'Irlande du...	87.46	9.97	1.068506	42820.6	0.47	368593.39	2
165	Émirats arabes unis	103.12	10.40	1.094970	63299.4	0.63	474939.04	2



Synthèse

Choix des données :

- 5 indices différents
- 6 variables

Nettoyage et fusion :

- Suppression des données non-utilisable
- Suppression de la France

Analyse des données :

- 2 méthodes de clustering
- 1 méthodes de réduction des dimensions
- Comparaison des résultats

Choix du groupe :

- Démarcation au niveau de l'exportation
- Très bon résultat sur les autres points