

BLEUE BANK

Effectuer une prédiction de revenus

SOMMAIRE

- **OBJECTIF**
- **DONNÉES**
 - Nettoyage
- **ANALYSE DESCRIPTIVE**
- **ANALYSE EXPLORATOIRE**
 - Diversité des cas
- **MODÉLISATION**
 - Anova
 - Régression linéaire

OBJECTIFS

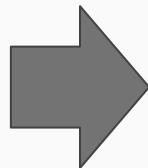
MISSION

Prospects

Nouveaux clients

Jeunes

Potentiel aux revenus



Différences pays

Modélisation

QUESTIONS :

- 1. Le revenus des enfants dépend t-il de la région, du type de pays ou du pays lui même ?**

- 2. Est-il possible de prédire le revenus des enfants de nos clients actuels ?**

DONNÉES

NETTOYAGE : Revenus

- **116 pays** représentés
- **2004 à 2011**
- Objectif de **conserver l'ensemble des pays.**

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.0
1	ALB	2008	2	100	916.66235	7297.0
2	ALB	2008	3	100	1010.91600	7297.0
3	ALB	2008	4	100	1086.90780	7297.0
4	ALB	2008	5	100	1132.69970	7297.0

- **Lituanie** : quantile 41 manquant
imputation par moyenne quantile inférieur et supérieur
- **Palestine** : gdpppp manquant
imputation par donnée externe (www.theglobaleconomy.com)
- **Kosovo** : gdpppp manquant
imputation par donnée externe (data.worldbank.org/)
- **Fidji** : valeur aberrante gdpppp
imputation moyenne région géographique



NETTOYAGE : Indice Gini

- Conservation données **2004 à 2011**
- **264 pays**
- **47 pays données complètes**

	Country Name	Country Code	2004	2006	2007	2008	2009	2010	2011
0	Aruba	ABW	NaN						
1	Afghanistan	AFG	NaN						
2	Angola	AGO	NaN	NaN	NaN	42.7	NaN	NaN	NaN
3	Albania	ALB	NaN	NaN	NaN	30.0	NaN	NaN	NaN
4	Andorra	AND	NaN						

NETTOYAGE : Indice Gini

Calcul indice Gini depuis les données de revenus par pays.

	country	year_survey	quantile	nb_quantiles	income	gdpppp	gini
0	ALB	2008	1	100	728.89795	7297.0	0.3
1	ALB	2008	2	100	916.66235	7297.0	0.3
2	ALB	2008	3	100	1010.91600	7297.0	0.3
3	ALB	2008	4	100	1086.90780	7297.0	0.3
4	ALB	2008	5	100	1132.69970	7297.0	0.3

Conservation de l'ensemble des données.

NETTOYAGE : Population

- Conservation données **2004 à 2011**
- **233 pays**

	Zone	Code zone	2004	2006	2007	2008	2009	2010	2011
0	Afghanistan	AFG	24.727	26.433	27.101	27.722	28.395	29.186	30.117
1	Afrique du Sud	ZAF	47.292	48.489	49.120	49.779	50.477	51.217	52.004
2	Albanie	ALB	3.105	3.063	3.034	3.003	2.973	2.948	2.929
3	Algérie	DZA	32.692	33.641	34.167	34.731	35.334	35.977	36.661
4	Allemagne	DEU	81.646	81.472	81.278	81.066	80.900	80.827	80.856

NETTOYAGE : Population

Calcul population **moyenne** par pays sur **période de l'étude**.

Left join sur données revenus.

	Zone	Code zone	pop
0	Afghanistan	AFG	27.669
1	Afrique du Sud	ZAF	49.768
2	Albanie	ALB	3.008
3	Algérie	DZA	34.743
4	Allemagne	DEU	81.149

	country	year_survey	quantile	nb_quantiles	income	gdpppp	gini	pop
0	ALB	2008	1	100	728.89795	7297.0	0.3	3.008
1	ALB	2008	2	100	916.66235	7297.0	0.3	3.008
2	ALB	2008	3	100	1010.91600	7297.0	0.3	3.008
3	ALB	2008	4	100	1086.90780	7297.0	0.3	3.008
4	ALB	2008	5	100	1132.69970	7297.0	0.3	3.008

4 pays toujours manquants :

- Chine
- Taiwan
- Soudan
- Kosovo



Imputation depuis données FAO ou World Bank.



NETTOYAGE : IGEincome

- Conservation données **parents (ensemble)** et **enfants (ensemble)**.
- **1940,1950,1960,1970,1980**
- **Nombreuses valeurs manquantes.**

	countryname	iso3	region	IGEincome
0	Afghanistan	AFG	South Asia	NaN
12	Angola	AGO	Sub-Saharan Africa	NaN
24	Albania	ALB	Europe & Central Asia	NaN
36	Albania	ALB	Europe & Central Asia	NaN
48	Albania	ALB	Europe & Central Asia	NaN

NETTOYAGE : IGEincome

- **Imputation données** région géographique (source externe).
- Imputation moyenne **région géographique**.
- Correction **4 pays avec IGEincome > 1**
 - Imputation moyenne région géographique.

ANALYSE DESCRIPTIVE

PREAMBULE

- 2004, 2006 à 2011
- 116 pays
- 92,3 % population mondiale



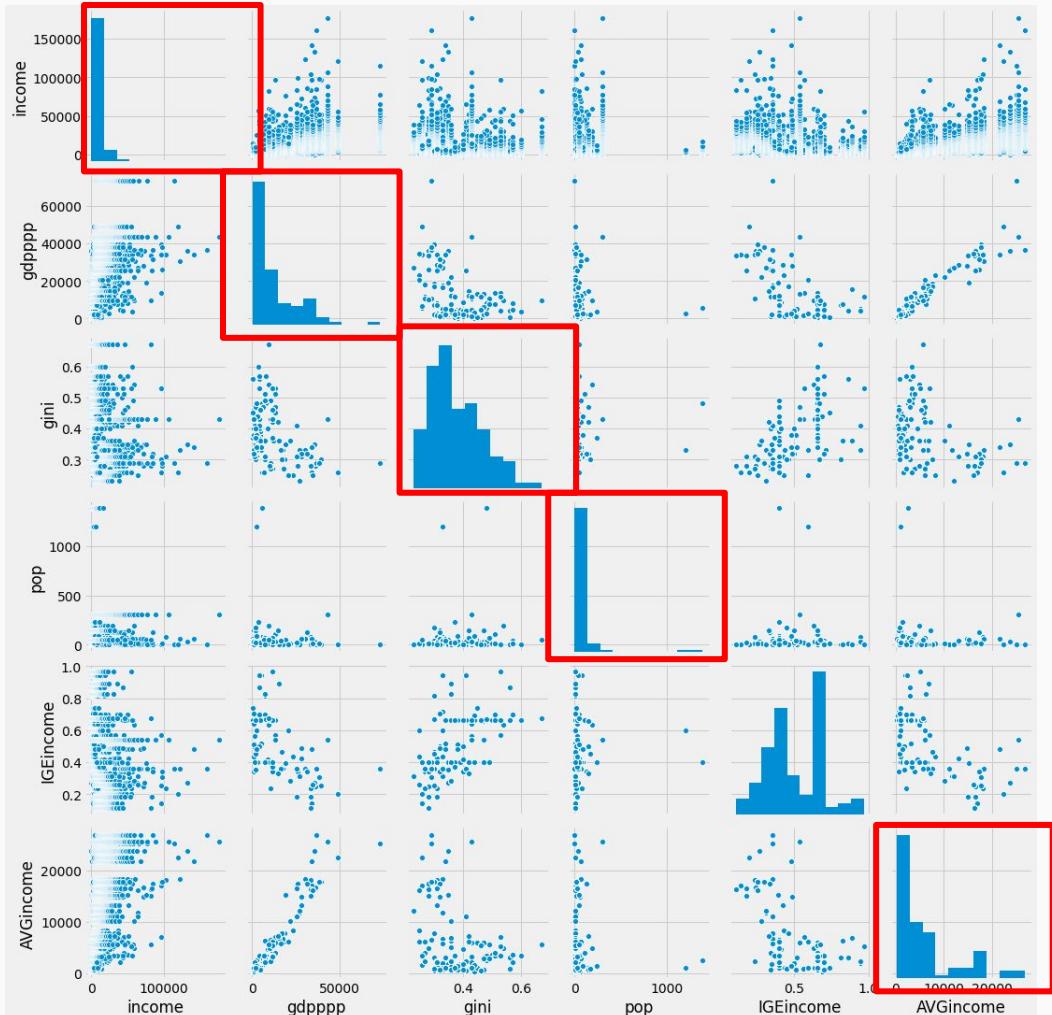
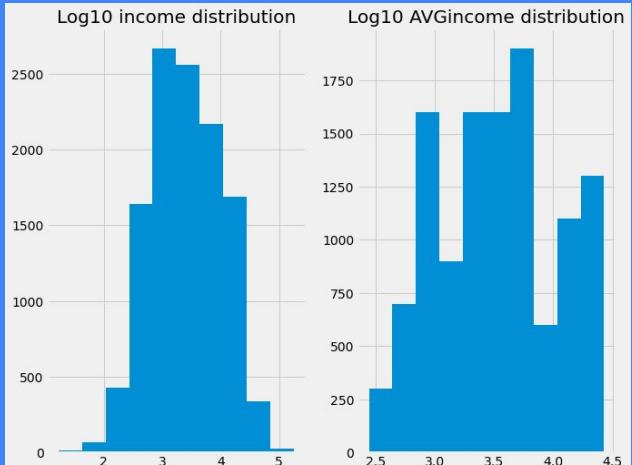
TENDANCES CENTRALES

	year_survey	quantile	nb_quantiles	income	gdpppp	gini	pop	IGEincome	AVGincome
count	11600.000000	11600.000000	11600.0	11600.000000	11600.000000	11600.000000	11600.000000	11600.000000	11600.000000
mean	2007.982759	50.500000	100.0	6069.121925	12412.282077	0.378707	54.014603	0.507329	6069.121925
std	0.909593	28.867314	0.0	9413.786596	13109.423699	0.089326	171.436697	0.183074	6632.479604
min	2004.000000	1.000000	100.0	16.719418	303.193050	0.230000	0.309000	0.112876	276.016044
25%	2008.000000	25.750000	100.0	900.768508	2577.500000	0.310000	4.694500	0.366551	1374.270126
50%	2008.000000	50.500000	100.0	2403.492950	7488.500000	0.360000	13.918000	0.480489	3287.174692
75%	2008.000000	75.250000	100.0	7515.313700	17679.250000	0.432500	40.474750	0.660000	7077.900152
max	2011.000000	100.000000	100.0	176928.550000	73127.000000	0.670000	1383.986000	0.966865	26888.511518

MATRICE PAR PAIRES

- Distribution fortement asymétrique.

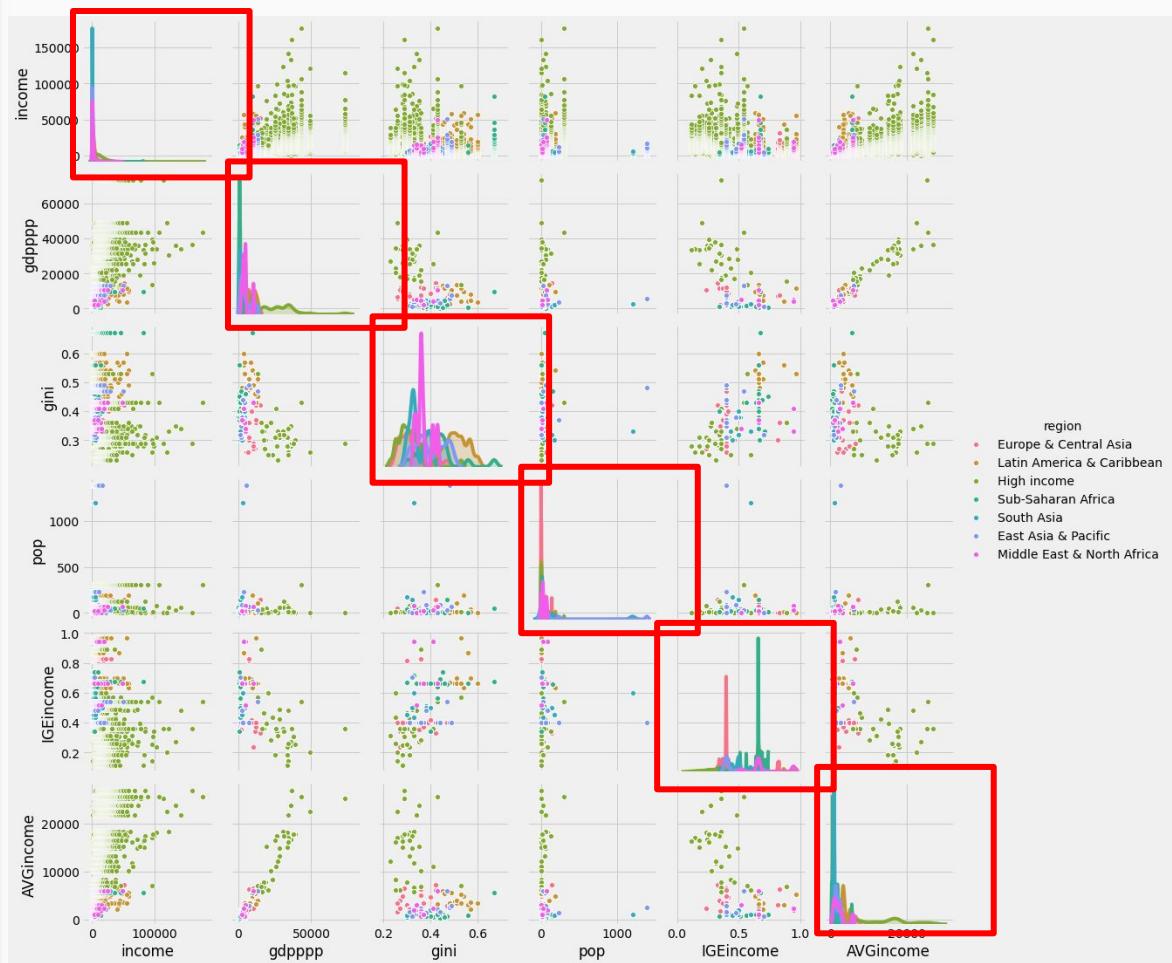
-> Transformation.



MATRICE PAR PAIRES

- **Disparités** inter-région géographiques
- **Disparités** intra-région géographiques

-> **Région géographique** probablement mauvais indicateur

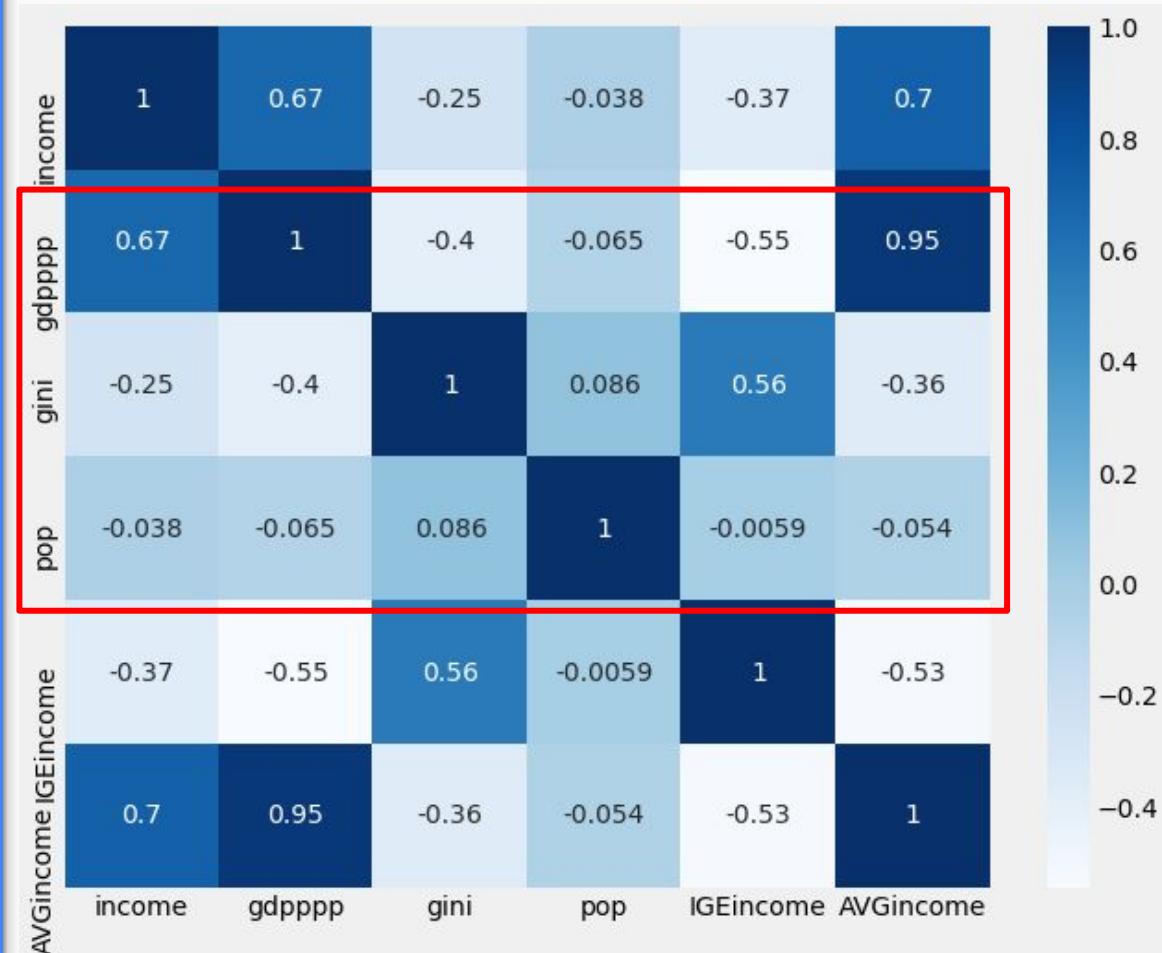


MATRICE CORRELATIONS

Variables descriptions
situation pays :

- gdppp
- gini

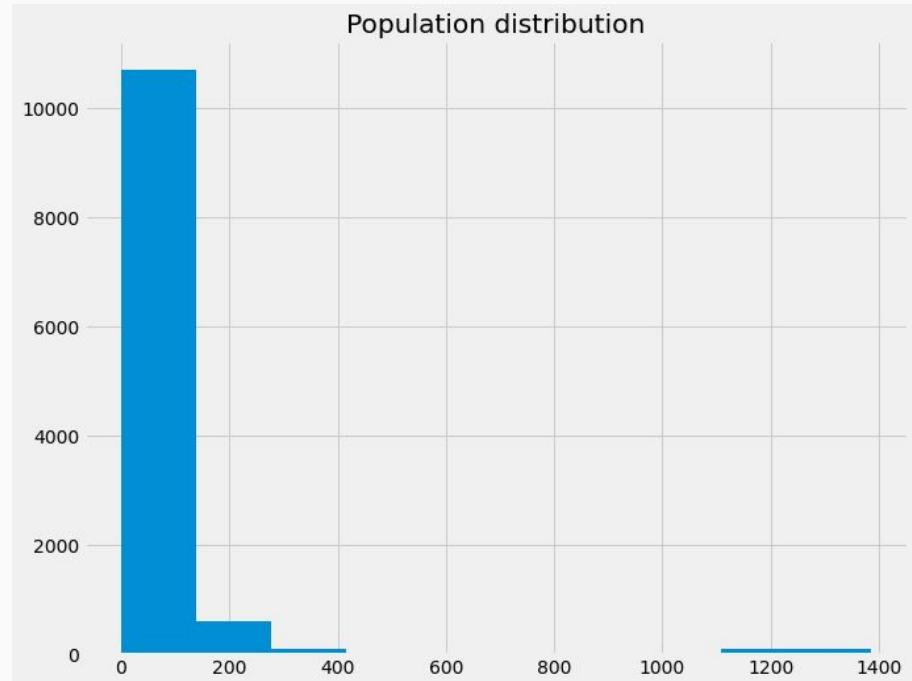
Population ?



Population

Trop sensible valeurs extrêmes

	Zone	Code zone	pop
41	Chine	CPR	1382.886
81	Inde	IND	1197.326
218	États-Unis d'Amérique	USA	303.016
82	Indonésie	IDN	235.146
30	Brésil	BRA	191.636



DIVERSITÉS DES CAS

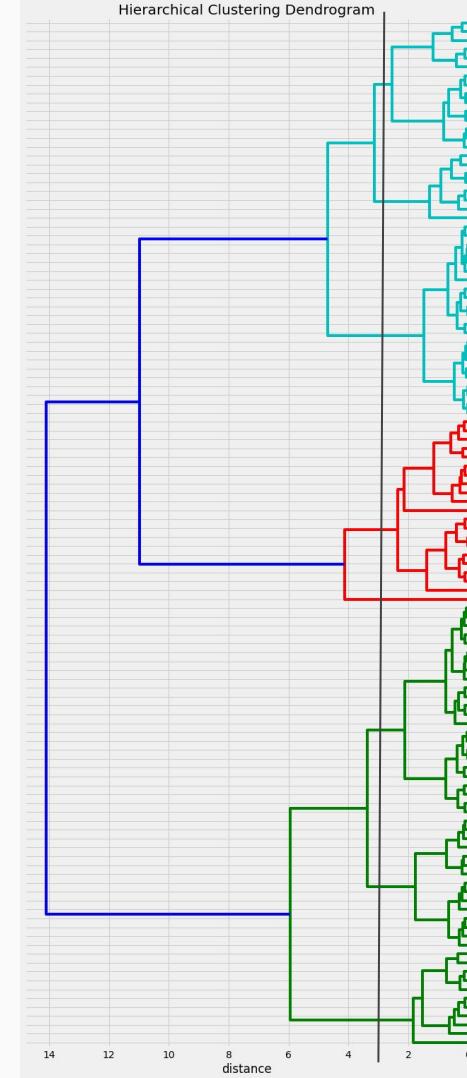
CLASSIFICATION

- Indice Gini
- GDPppp (corrélé avec AVGincome, mais moins asymétrique)



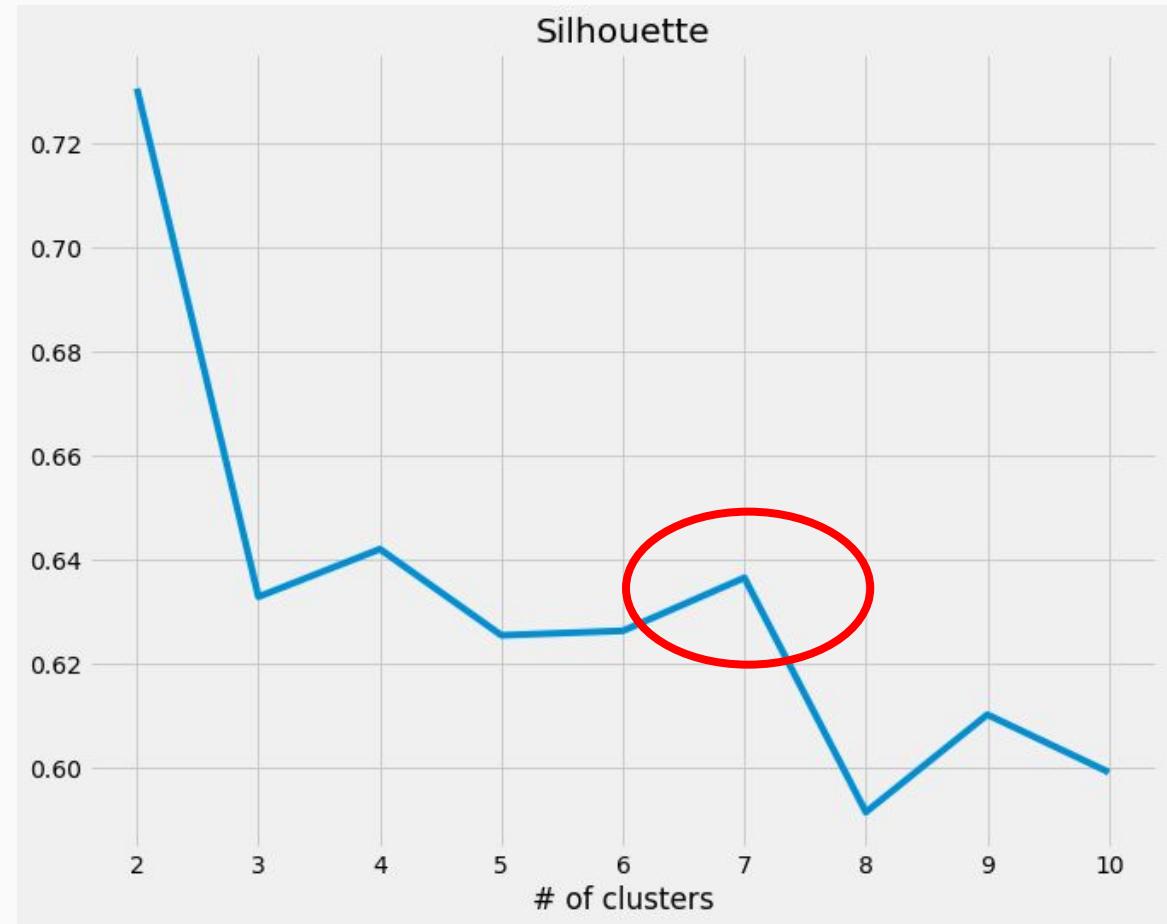
CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

- Comment définir le nombre de cas/cluster représentatifs ?
- Distance 3 ?



SILHOUETTE METRIC KMEANS

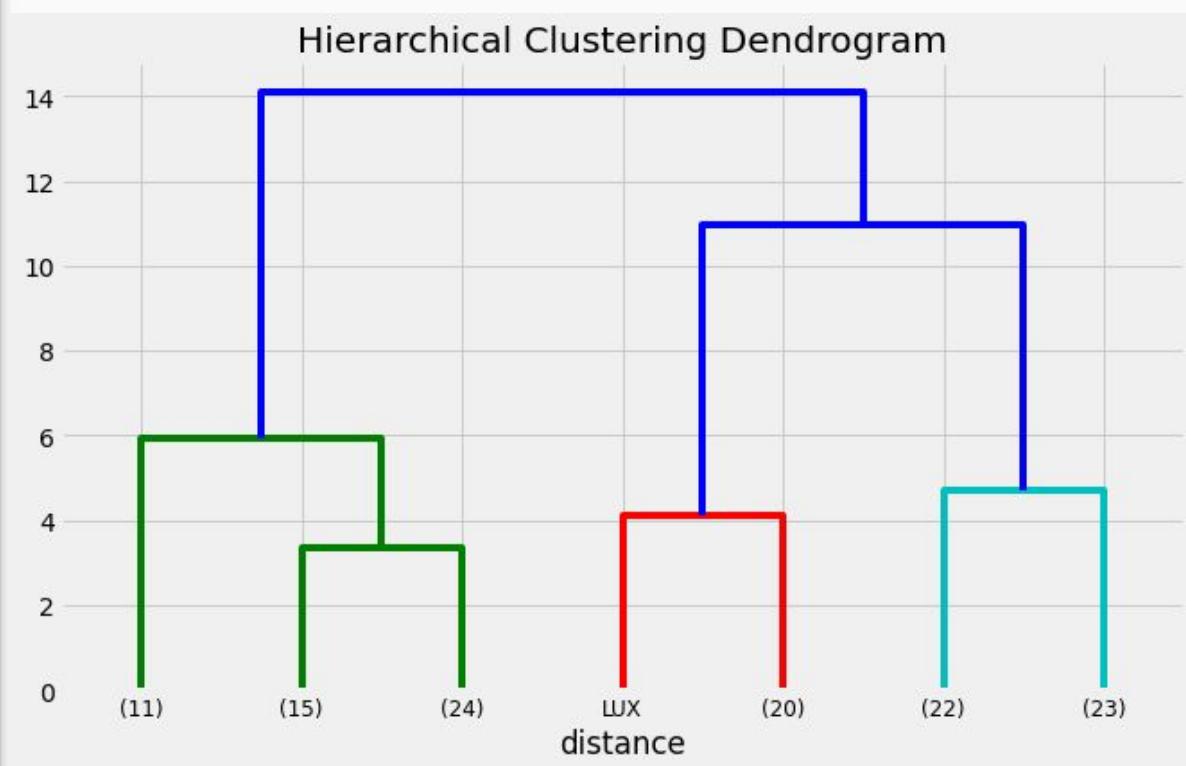
n = 7



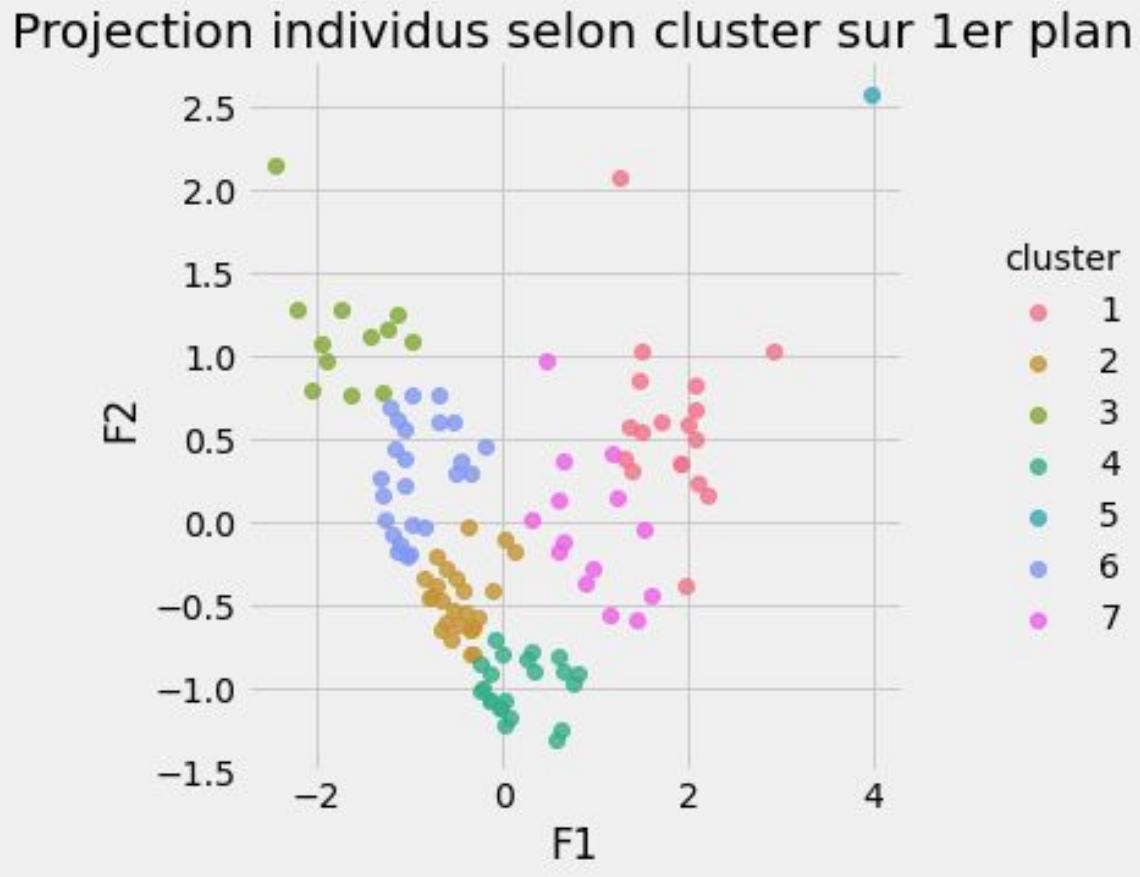
CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

7 Clusters

1 seul cluster individuel :
Luxembourg

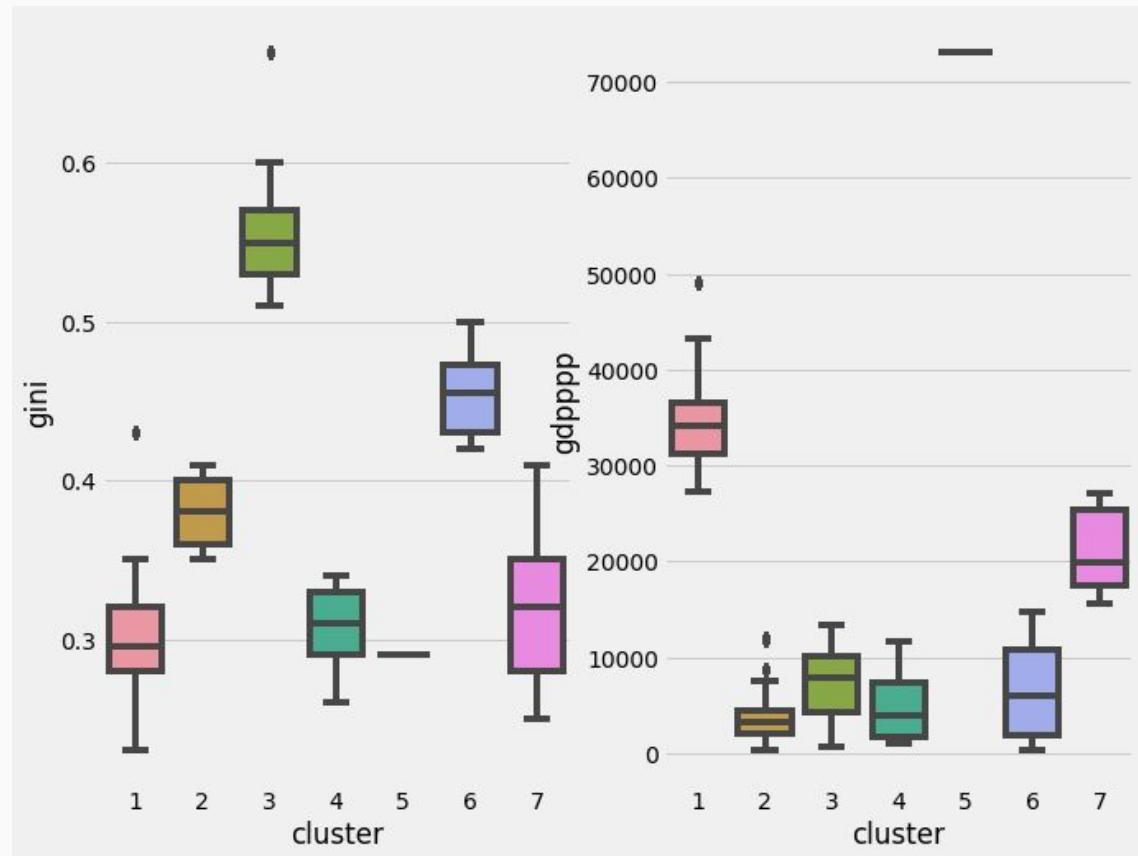


PROJECTION CLUSTER 1er PLAN



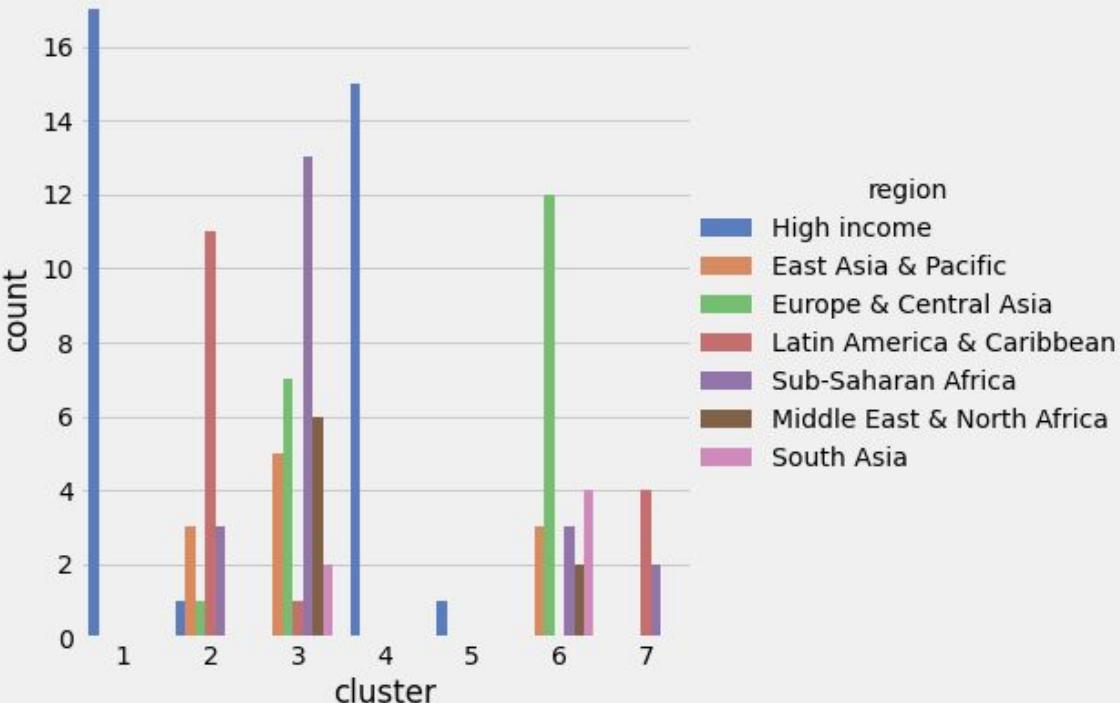
CARACTÉRISATION CLUSTERS

- Bonne différenciation des clusters



COMPOSITION DES CLUSTERS

Composition des clusters



- Région géographique mauvais indicateur des cas.

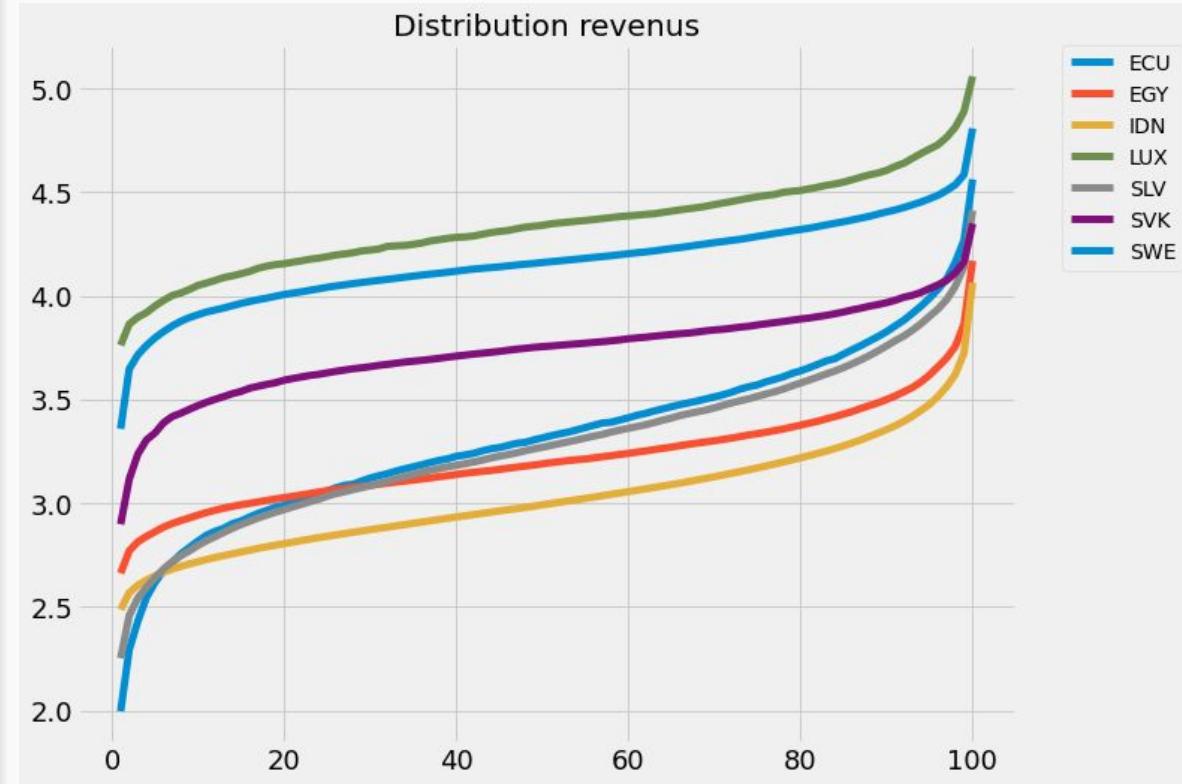
SÉLECTION PAYS REPRÉSENTATIFS

Distance Manhattan entre individus et centroïdes clusters

- 'ECU': Equateur
- 'EGY' : Egypte
- 'IDN' : Indonésie
- 'LUX' : Luxembourg
- 'SLV' : Slovénie
- 'SVK' : Slovaquie
- 'SWE' : Suède

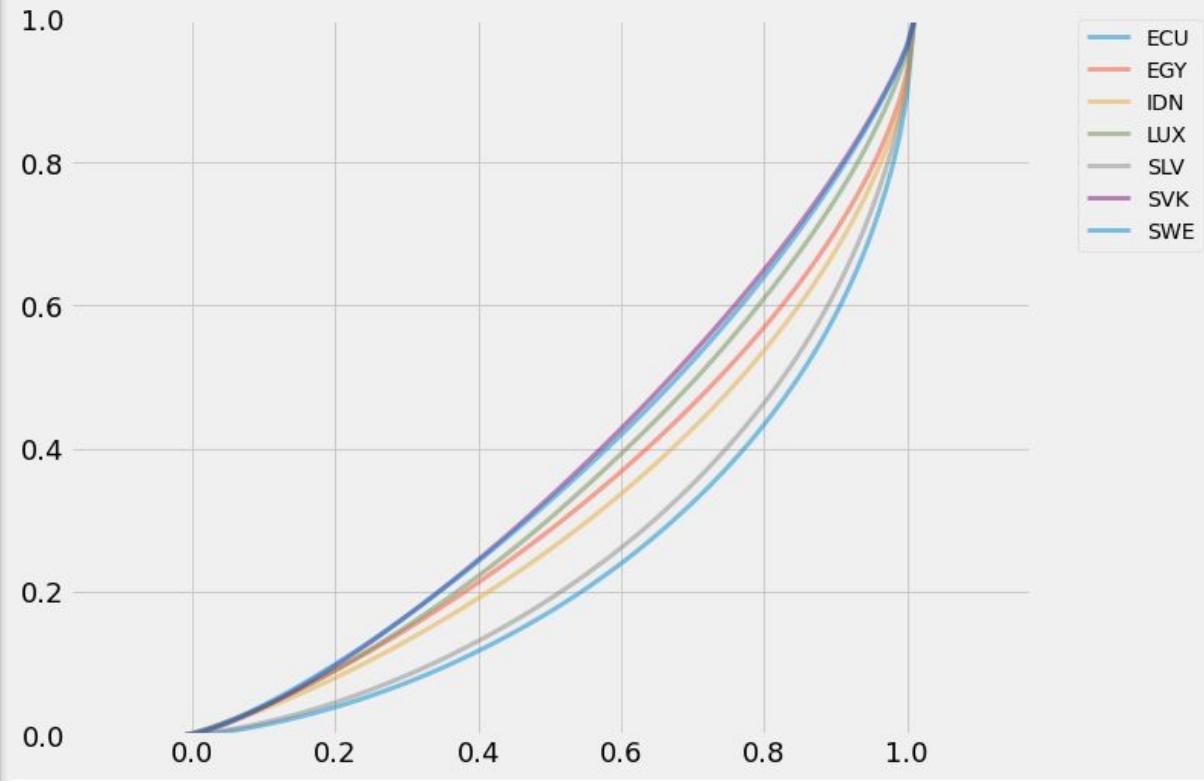
DISTRIBUTION REVENUS

- **Différences** de revenus par centile
- **Répartitions inégalitaires** des revenus pour certains pays (ECU)
- **Différences** intra-région

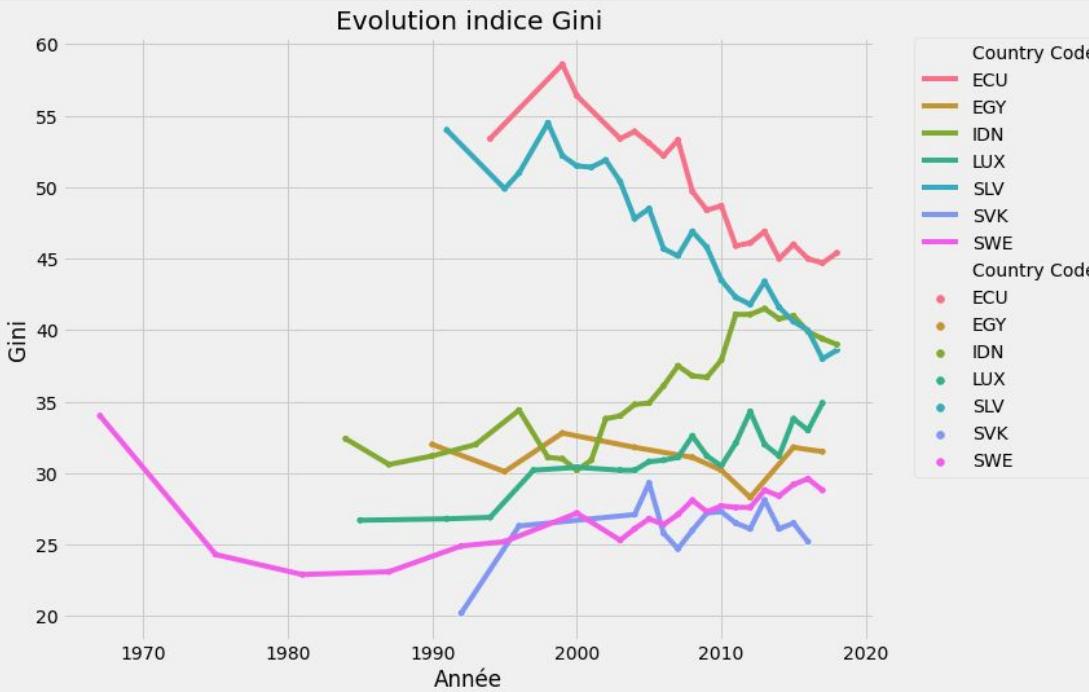


INDICE GINI

- Différences de cas intra-région
- Etude de cas **individuels**



INDICE GINI



- Tendance rapprochement vers moyenne
- Différences inter-région et intra-région

INDICE GINI

Gini moyen = 0.38



42 ème pays les plus égalitaires sur 116

Top 5 des pays les plus inégalitaires:

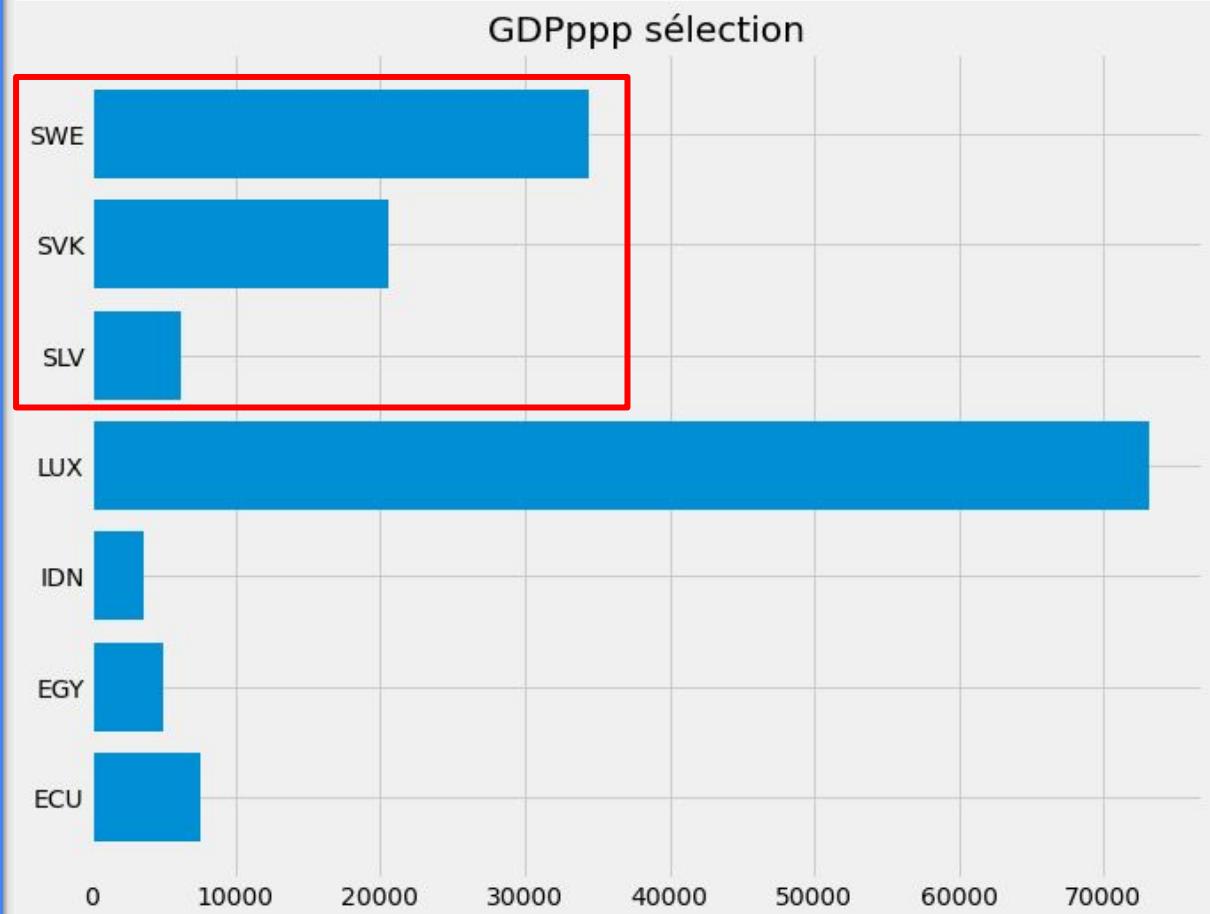
- Afrique du Sud
- Honduras
- Guatemala
- Colombie
- Bolivie

Top 5 des pays les plus égalitaires :

- Arménie
- Suède
- République-Tchèque
- Slovaquie
- Slovénie

GDPppp

- Différences de cas intra-région
- Etude de cas **individuels**



POINTS CLÉS

- **Grande variété** de cas de pays
- Pas spécifique à la **région géographique**
- Pas spécifique au **cluster**

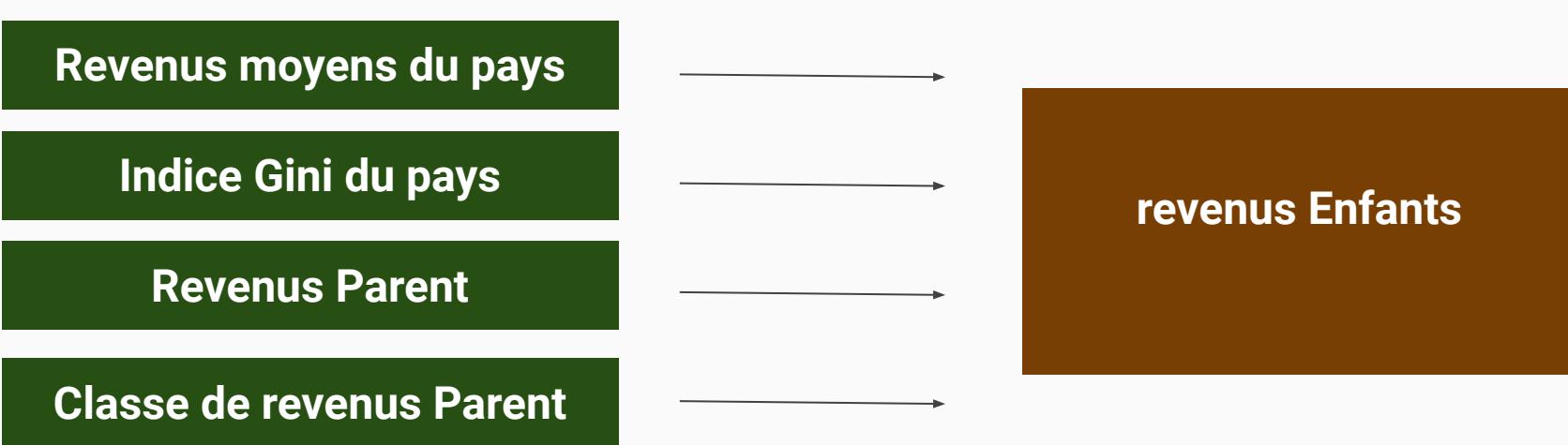
QUESTIONS :

1. **Le revenus des enfants dépend t-il de la région, du type de pays ou du pays lui même ?**
2. Est-il possible de prédire le revenus des enfants de nos clients actuels ?

-> **Le revenus des enfants dépend avant tout des caractéristiques du pays**

MODÉLISATION

PREDICTION REVENUS ENFANTS



-> Manque cible pour la modélisation

PREDICTION REVENUS ENFANTS

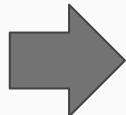
$$y_{child} = e^{\alpha + p_j \ln(y_{parent}) + \epsilon}$$

Revenus enfants/prospects Elasticité Revenus parent/clients

Génération aléatoire
de la classe de revenu
des parents

CLASSES ENFANT ET PARENT

- Revenus parent
- Revenus enfant

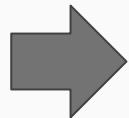


- cparent
- cchild

	country	ln_yparent	epsilone	pj	ychild	cparent	cchild
0	ALB	-0.720820	1.663780	0.815874	2.931989	24	80
1	ALB	0.995508	-0.366914	0.815874	1.560951	85	64
2	ALB	0.821614	1.162664	0.815874	6.252606	80	93
3	ALB	0.984201	0.483175	0.815874	3.618862	84	85
4	ALB	0.280341	-0.337515	0.815874	0.896917	62	47

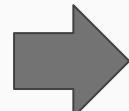
PROBABILITÉS

- c_child
- c_parent



$P(c_{parent} = 1 | c_{child} = 1) :$
nombre $c_{child} = 1 \& c_{parent} = 1$ / nombre $c_{child} = 1$

	country	ln_yparent	epsilone	pj	ychild	cparent	cchild
0	ALB	-0.720820	1.663780	0.815874	2.931989	24	80
1	ALB	0.995508	-0.366914	0.815874	1.560951	85	64
2	ALB	0.821614	1.162664	0.815874	6.252606	80	93
3	ALB	0.984201	0.483175	0.815874	3.618862	84	85
4	ALB	0.280341	-0.337515	0.815874	0.896917	62	47



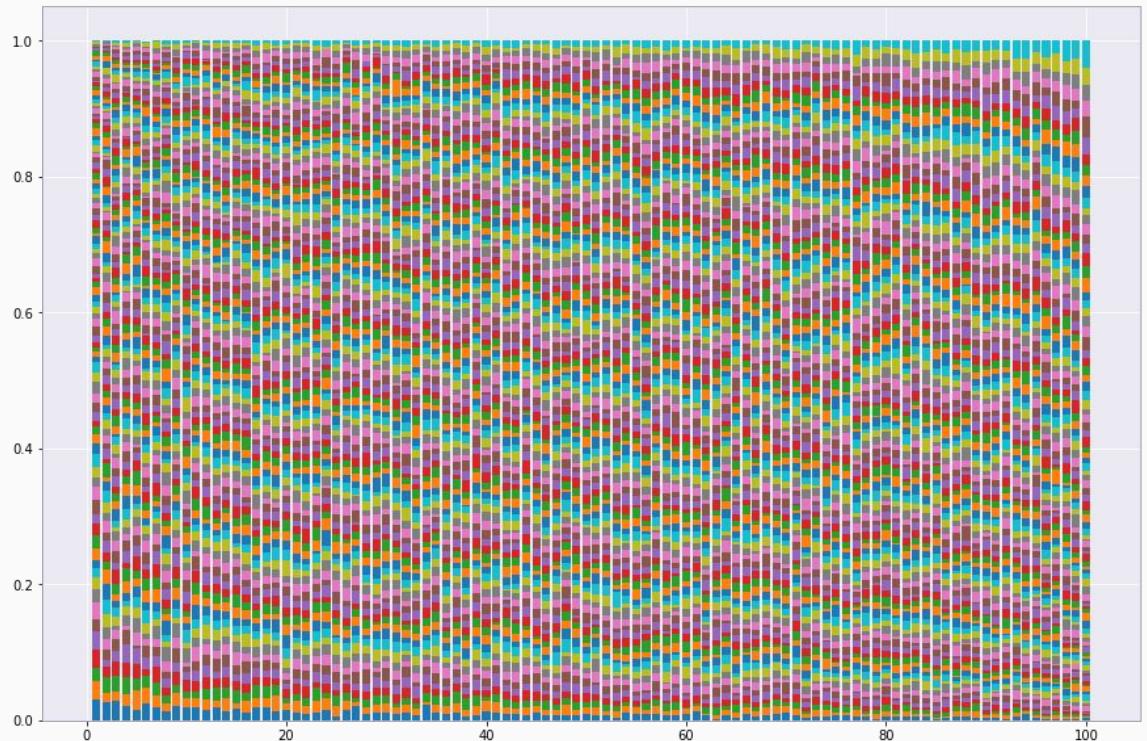
	country	c_child	c_parent	prob
0	ALB	1	1	0.216216
1	ALB	1	2	0.106106
2	ALB	1	3	0.083083
3	ALB	1	4	0.057057
4	ALB	1	5	0.055055

DISTRIBUTIONS CONDITIONNELLES



Gini = 0,33

42 ème pays les plus égalitaires sur 116

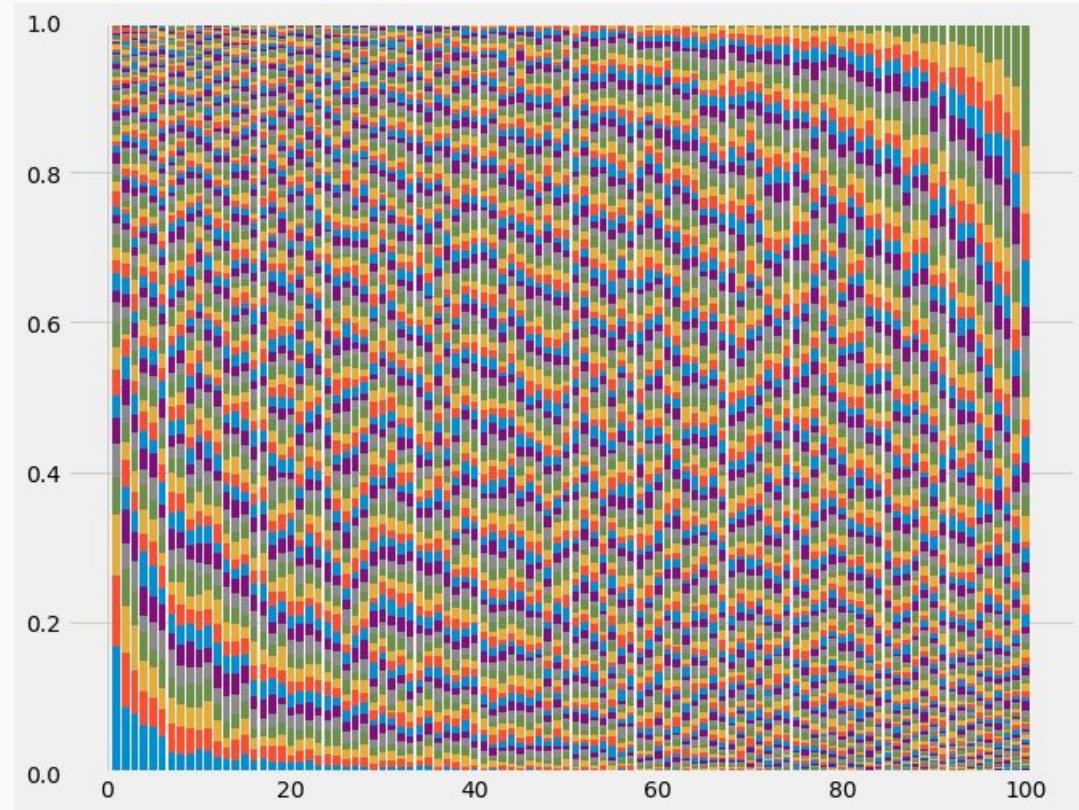




Gini = 0,67

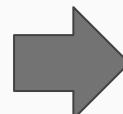
116 ème pays les plus égalitaires sur 116

ZAF



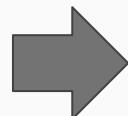
CLASSES / PROBABILITÉS

	country	ln_yparent	epsilone	pj	ychild	cparent	cchild
0	ALB	-0.720820	1.663780	0.815874	2.931989	24	80
1	ALB	0.995508	-0.366914	0.815874	1.560951	85	64
2	ALB	0.821614	1.162664	0.815874	6.252606	80	93
3	ALB	0.984201	0.483175	0.815874	3.618862	84	85
4	ALB	0.280341	-0.337515	0.815874	0.896917	62	47



	country	c_child	c_parent	prob
0	ALB	1	1	0.216216
1	ALB	1	2	0.106106
2	ALB	1	3	0.083083
3	ALB	1	4	0.057057
4	ALB	1	5	0.055055

- prob



n individus ($c_{parent} = 1 | c_{child} = 5$) :
 $500 * P(c_{parent} = 1 | c_{child} = 5)$

Données finales

left join data (country + quantile/c_parent)

	country	quantile	c_parent	gini	income_child	AVGincome	income_parent	region
0	ALB	1	1	0.3	728.89795	2994.829902	728.89795	Europe & Central Asia
1	ALB	1	1	0.3	728.89795	2994.829902	728.89795	Europe & Central Asia
2	ALB	1	1	0.3	728.89795	2994.829902	728.89795	Europe & Central Asia
3	ALB	1	1	0.3	728.89795	2994.829902	728.89795	Europe & Central Asia
4	ALB	1	1	0.3	728.89795	2994.829902	728.89795	Europe & Central Asia

PREDICTION REVENUS ENFANTS

Revenus moyens du pays



Indice Gini du pays



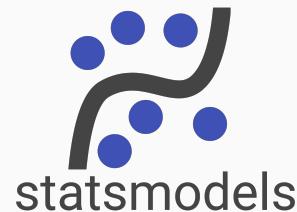
Revenus Parent



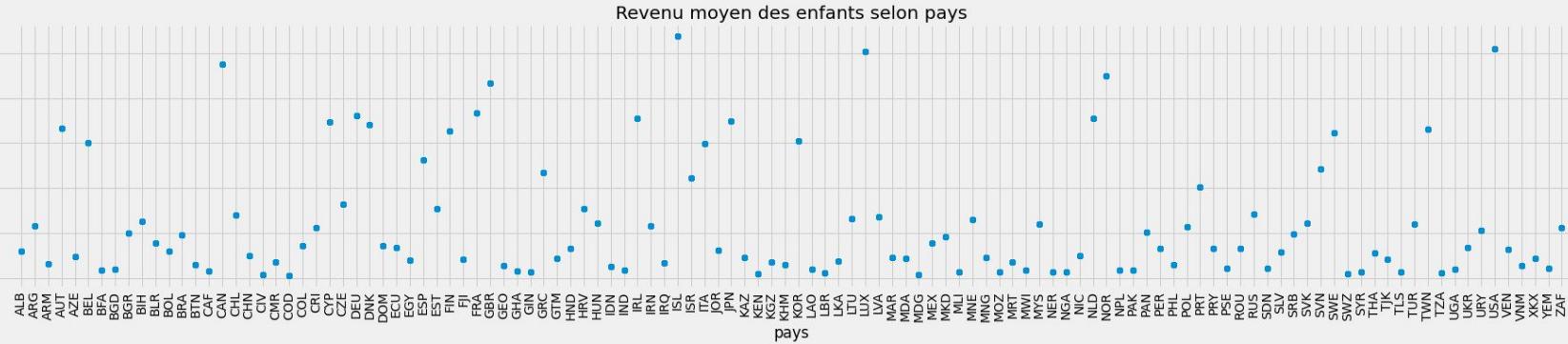
Classe de revenus Parent



Classe revenus Enfants



Le pays a t-il une influence sur le revenu moyen ?



RÉGRESSION LINÉAIRE

- Échantillonnage ($n = 5000$, stratifié sur pays)
- **income_child ~ pays**
- income_child ~ AVGincome + gini
- income_child ~ AVGincome + gini + income_parent
- log_income_child ~ log_AVGincome + gini
- **log _ income_child ~ log_AVGincome + gini + log_income_parent**
- **log _ income_child ~ log_AVGincome + gini + log_income_parent (cleaned)**

ANOVA

income_child :

- ~Pays

 H_0 : égalité de variance entre les pays

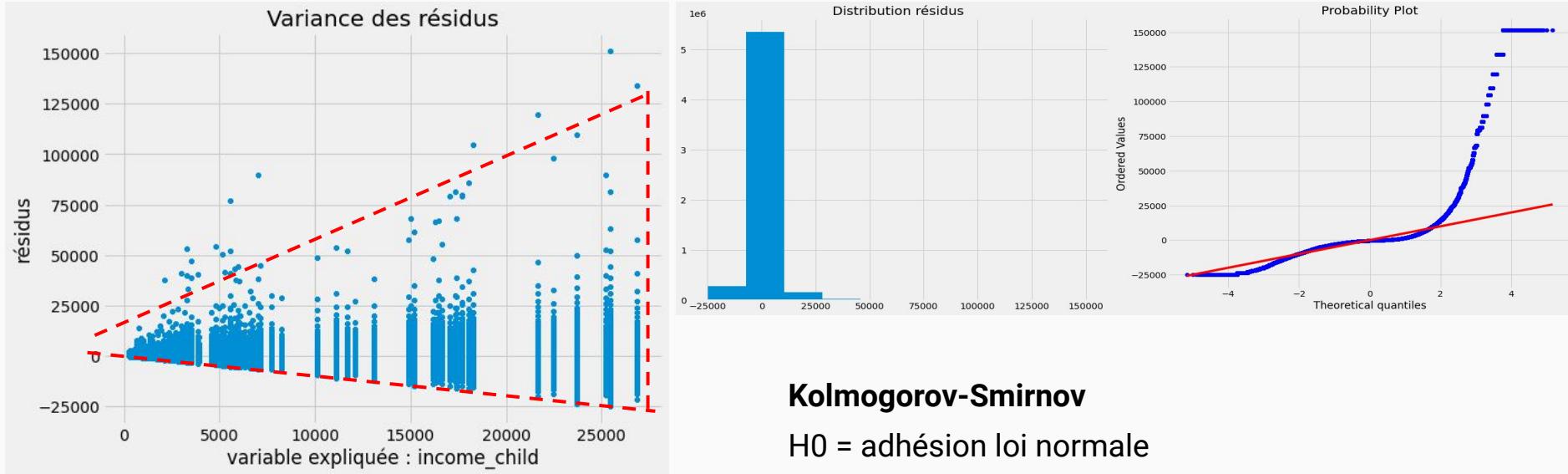
OLS Regression Results						
Dep. Variable:	income	R-squared:	0.497			
Model:	OLS	Adj. R-squared:	0.497			
Method:	Least Squares	F-statistic:	4.952e+04			
Date:	Thu, 21 May 2020	Prob (F-statistic):	0.00			
Time:	13:33:48	Log-Likelihood:	-5.8945e+07			
No. Observations:	5765267	AIC:	1.179e+08			
Df Residuals:	5765151	BIC:	1.179e+08			
Df Model:	115					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	2991.5544	29.960	99.853	0.000	2932.835	3050.274
country[T.ARG]	2839.5669	42.352	67.047	0.000	2756.558	2922.576

 $p < 0.05$ Rejet de H_0 .

Il y a une influence du pays sur le revenu moyen.

Le pays permet d'expliquer 50% de la variation du revenu.

income_child ~ pays : homoscédasticité & normalité des résidus



Kolmogorov-Smirnov

H_0 = adhésion loi normale

$p < 0.05$

Rejet de H_0 , pas d'adhésion loi normale.

RÉGRESSION LINÉAIRE

log_income_child :

~log_AVGincome

+gini

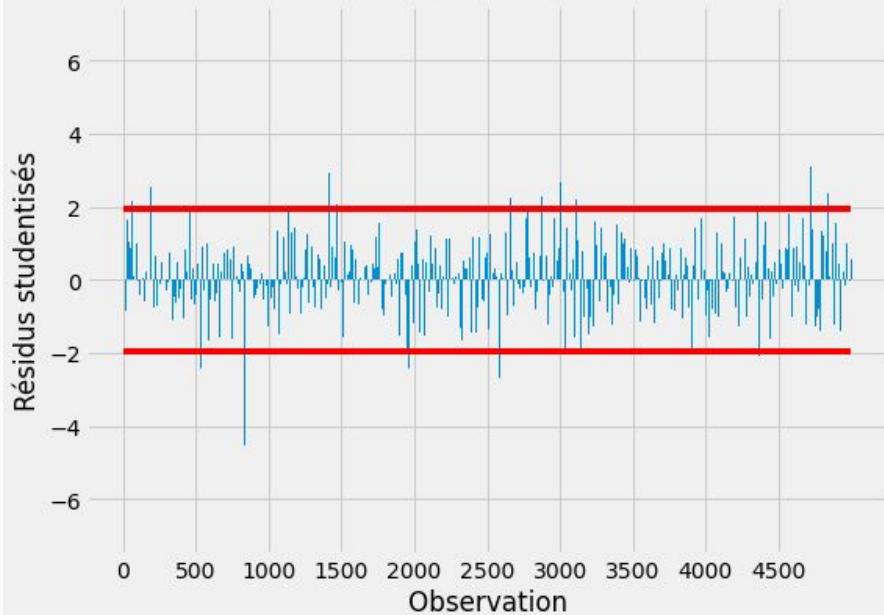
+log_income_parent

OLS Regression Results							
Dep. Variable:	log_income_child	R-squared:	0.782	Model:	OLS	Adj. R-squared:	0.782
Method:	Least Squares	F-statistic:	5982.	Date:	Fri, 29 May 2020	Prob (F-statistic):	0.00
Time:	10:07:27	Log-Likelihood:	-723.78	No. Observations:	5000	AIC:	1456.
Df Residuals:	4996	BIC:	1482.	Df Model:	3		
Covariance Type:	nonrobust						
	coef	std err	t	P> t	[0.025	0.975]	
const	0.1379	0.038	3.626	0.000	0.063	0.213	
log_AVGincome	0.4984	0.015	33.500	0.000	0.469	0.528	
gini	-0.3809	0.047	-8.089	0.000	-0.473	-0.289	
log_income_parent	0.4860	0.013	38.290	0.000	0.461	0.511	

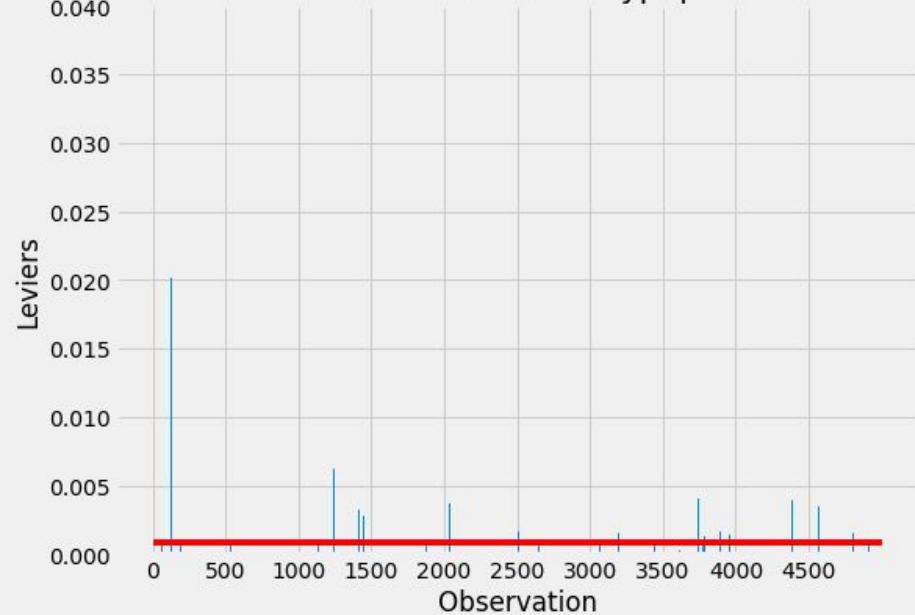
Le pays permet d'expliquer 78% de la variation du revenu.

$\log_{\text{income_child}} \sim \log_{\text{AVGincome+gini}} + \log_{\text{income_parent}}$: valeurs atypiques et influentes

Résidus studentisés



Influences valeurs atypiques



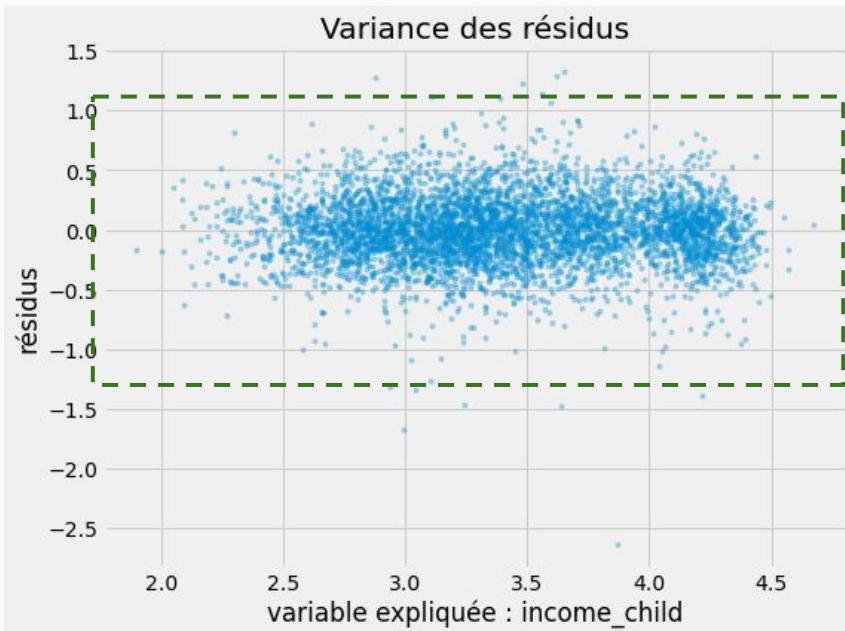
- **192 valeurs atypiques et influentes**
- soit **3.84%**

log _income_child ~ log_AVGincome+gini+log_income_parent : Colinéarité des variables

log_AVGincome	3.51
gini	1.13
log_income_parent	3.68

Variance inflation factor des variables (VIF) < 10

$\log_{\text{income}}_{\text{child}} \sim \log_{\text{AVG}}\text{income} + \text{gini} + \log_{\text{income}}_{\text{parent}}$: homoscédasticité



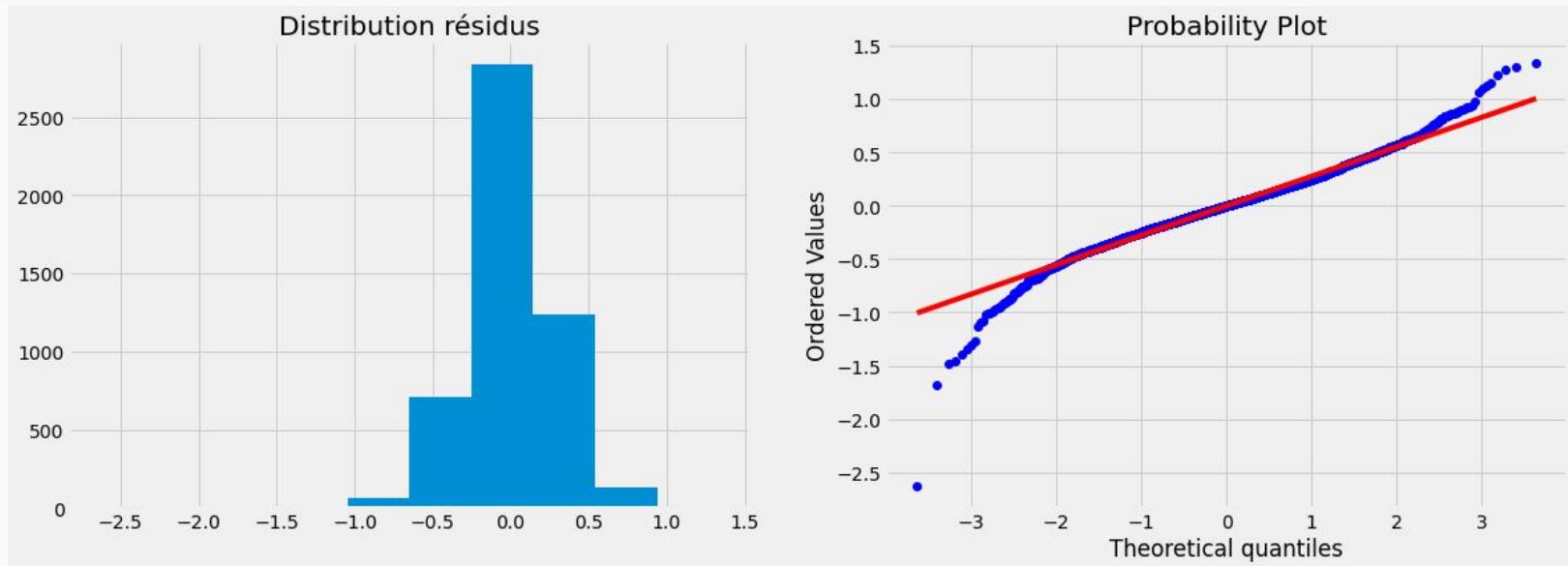
Breusch Pagan

H_0 = égalité de variance

$p < 0.05$

Rejet de H_0 , pas d'égalité de variance.

$\log_{\text{income}}_{\text{child}} \sim \log_{\text{AVGincome}} + \text{gini} + \log_{\text{income}}_{\text{parent}}$: Normalité des résidus



Kolmogorov-Smirnov

H_0 = adhésion loi normale

$p > 0.05$

Pas de rejet de H_0 , adhésion loi normale.

RÉGRESSION LINÉAIRE

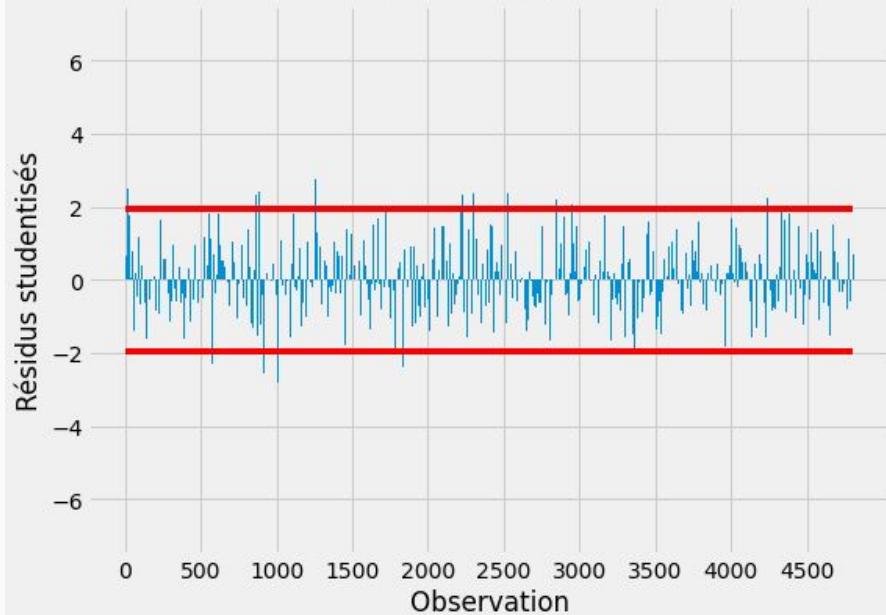
Sans valeurs atypiques et influentes

OLS Regression Results						
Dep. Variable:	log_income_child	R-squared:	0.839			
Model:	OLS	Adj. R-squared:	0.839			
Method:	Least Squares	F-statistic:	8348.			
Date:	Fri, 29 May 2020	Prob (F-statistic):	0.00			
Time:	10:15:48	Log-Likelihood:	165.24			
No. Observations:	4808	AIC:	-322.5			
Df Residuals:	4804	BIC:	-296.6			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.1321	0.033	4.040	0.000	0.068	0.196
log_AVGincome	0.5440	0.013	41.709	0.000	0.518	0.570
gini	-0.4582	0.041	-11.080	0.000	-0.539	-0.377
log_income_parent	0.4492	0.011	40.063	0.000	0.427	0.471

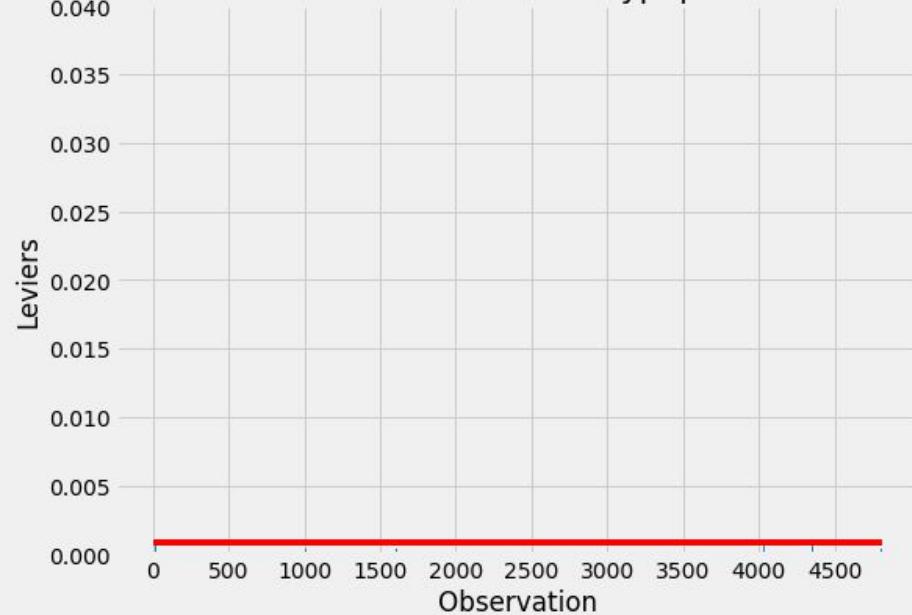
Le pays permet d'expliquer 84% de la variation du revenu.

$\log_{\text{e}} \text{income_child} \sim \log_{\text{e}} \text{AVGincome} + \text{gini} + \log_{\text{e}} \text{income_parent}$: valeurs atypiques et influentes

Résidus studentisés



Influences valeurs atypiques



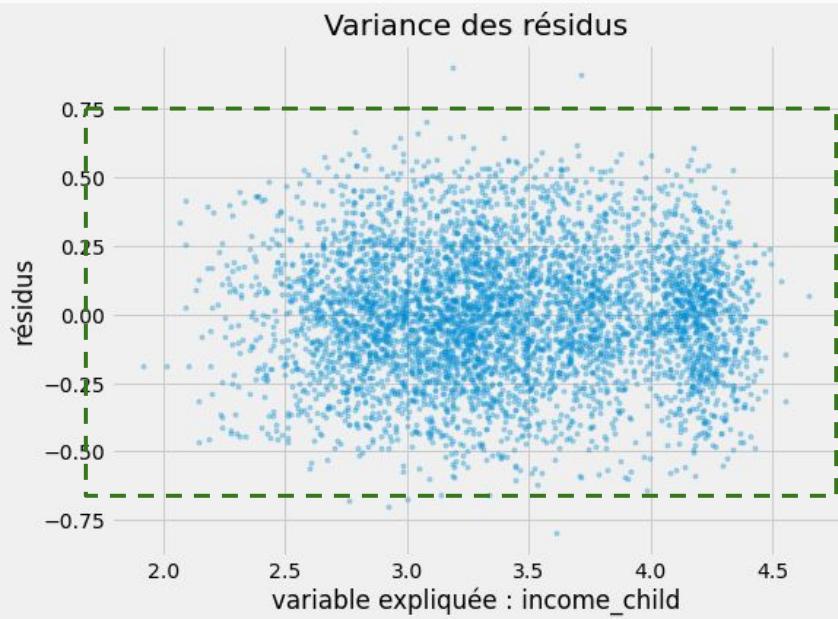
- **19 valeurs atypiques et influentes**
- soit **0.38%**

log _income_child ~ log_AVGincome+gini+log_income_parent : Colinéarité des variables

log_AVGincome	3.73
gini	1.13
log_income_parent	3.91

Variance inflation factor des variables (VIF) < 10

$\log_{\text{income}}_{\text{child}} \sim \log_{\text{AVG}}\text{income} + \text{gini} + \log_{\text{income}}_{\text{parent}}$: homoscédasticité



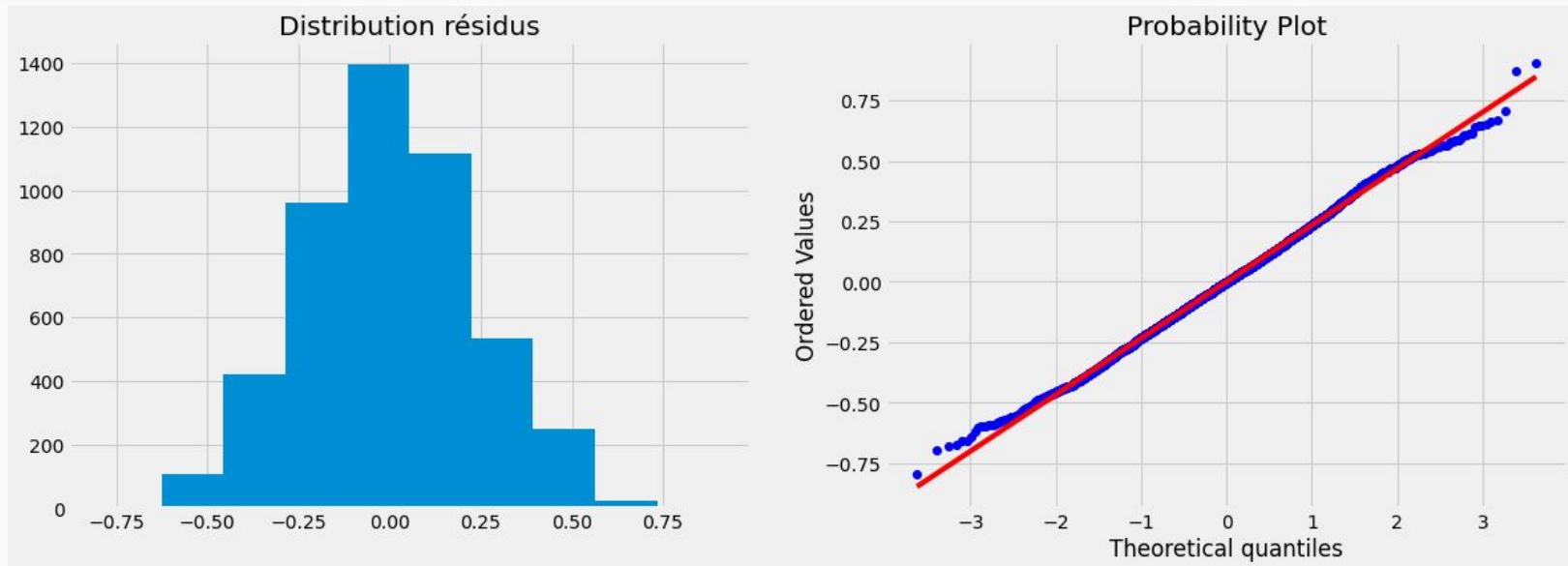
Breusch Pagan

H_0 = égalité de variance

$p < 0.05$

Rejet de H_0 , pas d'égalité de variance.

$\log_{\text{income}}_{\text{child}} \sim \log_{\text{AVGincome}} + \text{gini} + \log_{\text{income}}_{\text{parent}}$: Normalité des résidus



Kolmogorov-Smirnov

H_0 = adhésion loi normale

$p > 0.05$

Pas de rejet de H_0 , adhésion loi normale.

QUESTIONS :

1. Le revenus des enfants dépend t-il de la région, du type de pays ou du pays lui même ?
2. Est-il possible de prédire le revenus des enfants de nos clients actuels ?

log_AVGincome+gini+log_income_parent

CONCLUSION

Mission : créer un modèle permettant de déterminer le revenu potentiel d'une personne

- **84% de variance** expliquée.
- Impact **négatif** de l'indice de **gini** du pays sur le revenu enfant.
- Variances non expliquée peut provenir de **variables non prises en compte** de l'analyse :
 - niveau d'études
 - catégorie sociale
 - origine ethnique
 - sexe
 - Structure familiale

Mission : créer un modèle permettant de déterminer le revenu potentiel d'une personne

- Autres modèles linéaires (ridge, LASSO)
- Cross Validation
- Ajout d'autres variables

Conclusion : la suite

Mission : cibler les prospects les plus susceptibles d'avoir, plus tard dans leur vie, de hauts revenus, parmis nos clients.

log_AVGincome+gini+log_income_parent > = log(99eme quantile revenu mondial)

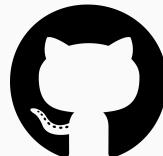
MERCI !



@xavbarbier



<https://www.linkedin.com/in/barbierxavier/>



<https://github.com/xavierbarbier/>



contact@xavierbarbier.com