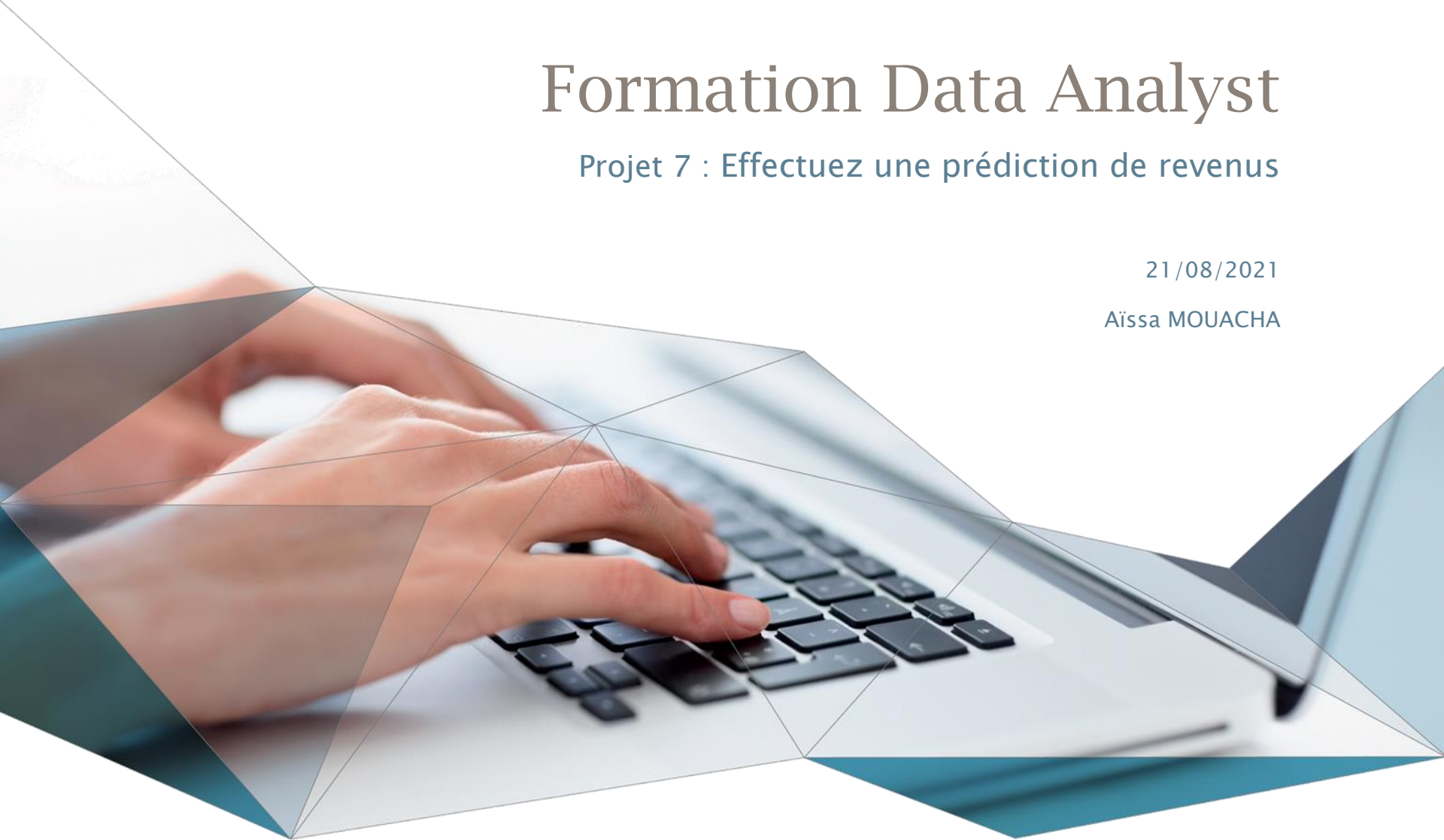


Formation Data Analyst


Projet 7 : Effectuez une prédiction de revenus

21/08/2021

Aïssa MOUACHA



SOMMAIRE

- ❑ Introduction
 - ❑ Data cleaning et statistique descriptive
 - ❑ Classification et illustrations
 - ❑ Algorithme de génération classe_parent
 - ❑ ANOVA et régressions multilinéaires
 - ❑ Conclusion
- 

Votre banque souhaite cibler de nouveaux clients potentiels, plus particulièrement les jeunes en âge d'ouvrir leur tout premier compte bancaire.



Stratégie : Modélisez les données à l'aide d'une régression linéaire.

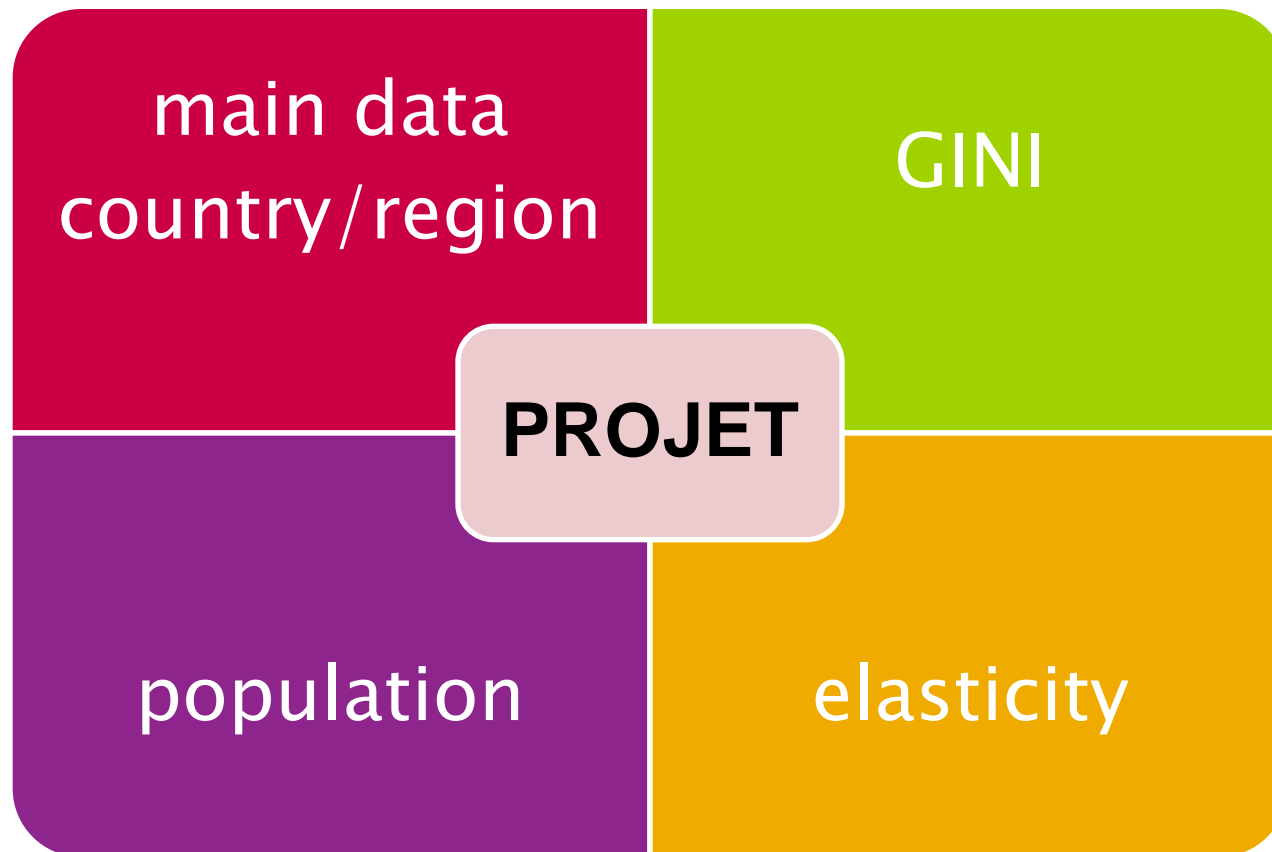
Objectif : Créer un programme capable de prédire les revenus d'un enfant.

Source : Site OC + World Income

Mission 1 : Statistique descriptive

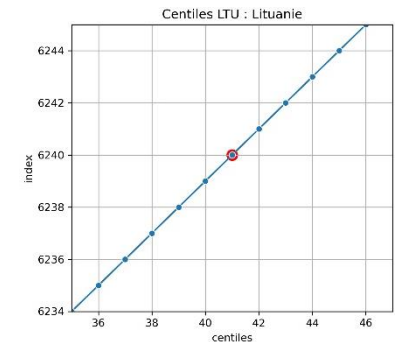
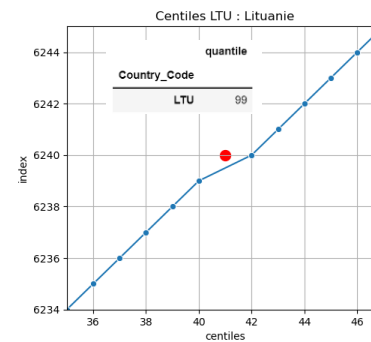
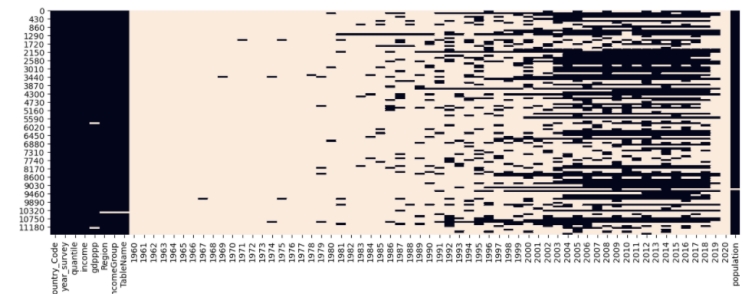
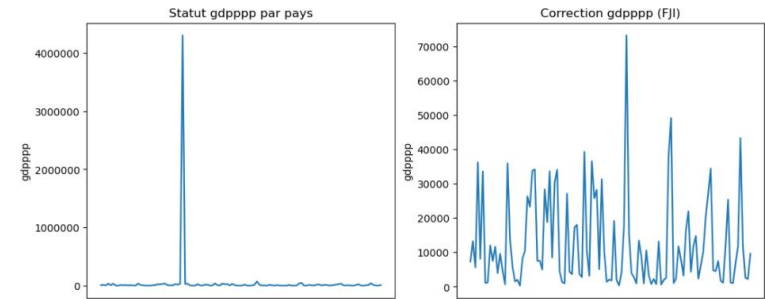
Décrire, résumer, représenter la donnée

Tables de l'étude



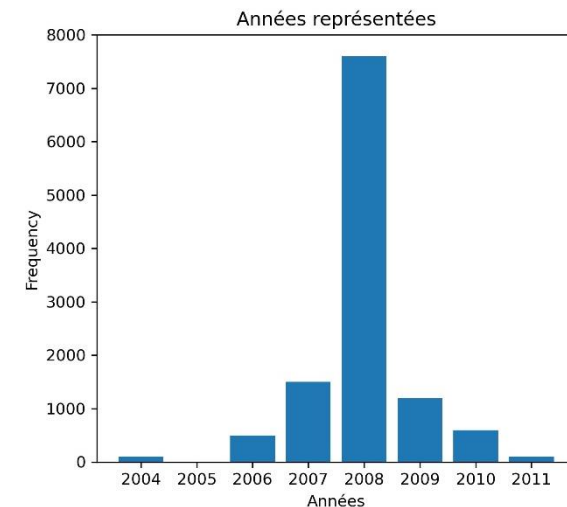
Data cleaning

- Nb total pays : 116
 - ✓ **Fidji** gdp PPP correction
 - ✓ Rename, drop
 - ✓ **Kosovo** population
 - ✓ **Soudan** population
 - ✓ **Taïwan** informations
 - ✓ **Palestine & Kosovo** gdp PPP
 - ✓ **Lituanie** centile manquant



Statistique descriptive

- Représentativité: 89% de la population Monde
- Les quantiles :
 - ✓ population séparée en **n classes d'égal effectif**
 - ✓ **situent** très rapidement **un sujet** au sein d'une population parente
- L'OCDE définit les Parités de Pouvoir d'Achat (PPP) comme :
 - ✓ Les **TAUX de CONVERSION** monétaire qui **EGALISENT** les **POUVOIRS D'ACHAT** des différentes monnaies

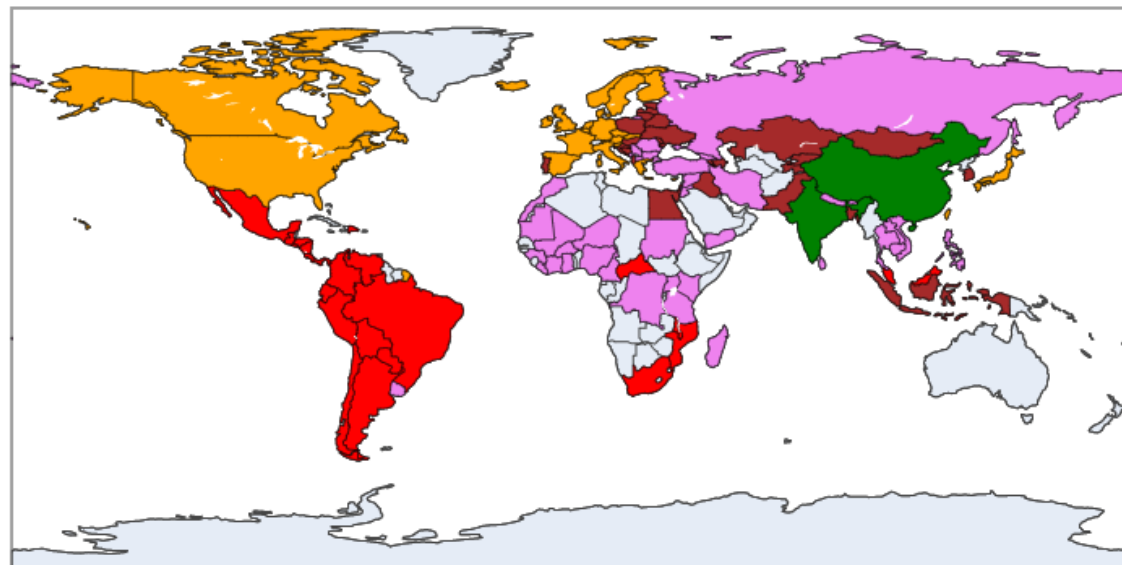
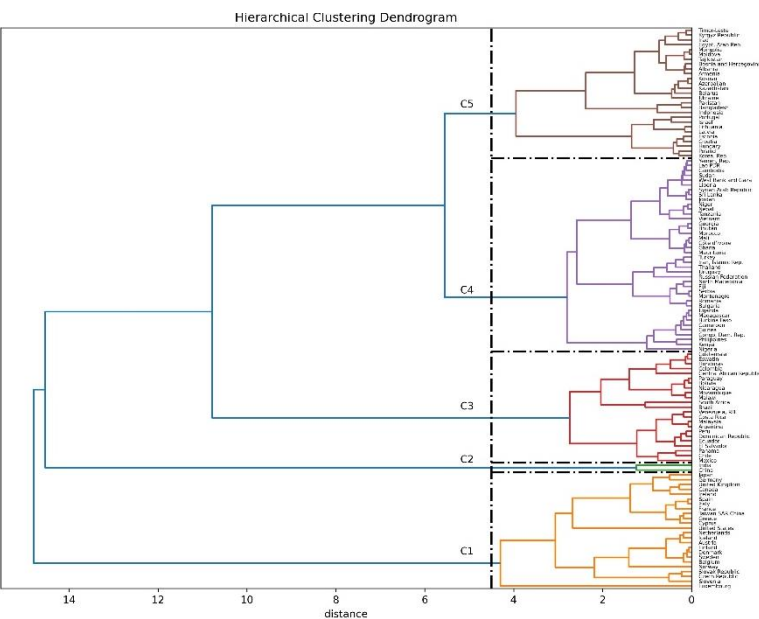


Mission 2: CAH

Illustrer la diversité des pays en termes de distribution de revenus et ainsi aboutir à un clustering

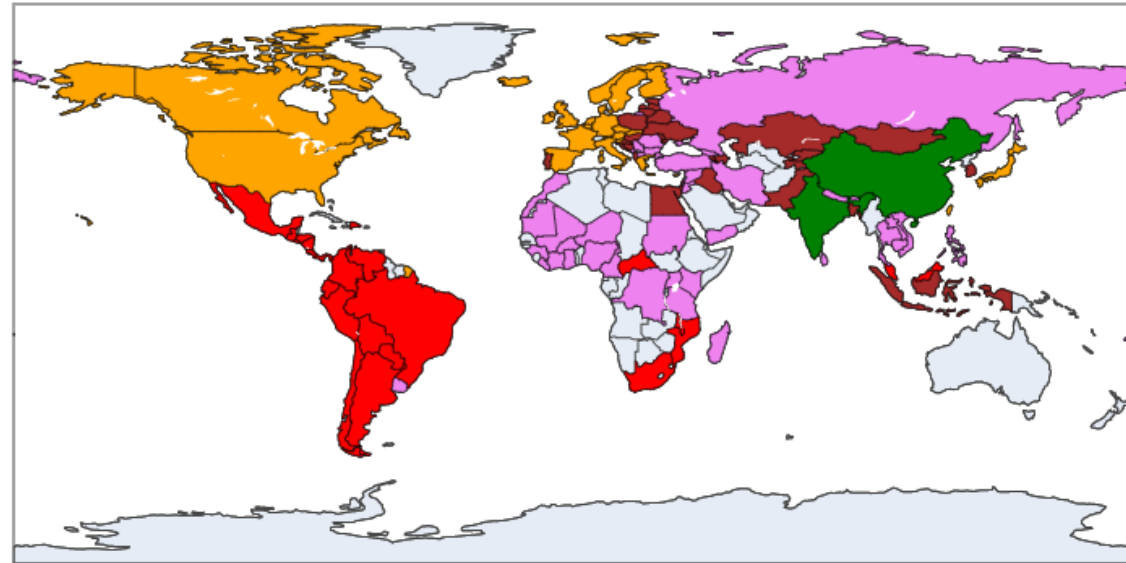
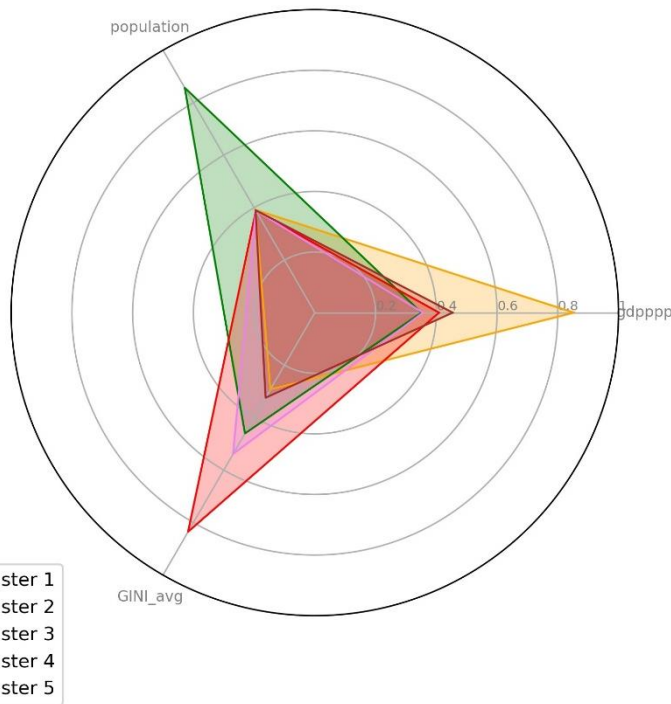
Classification & illustrations

- Classification : 5 clusters suivant les variables [[gdp PPP , population , GINI]]



Classification & illustrations

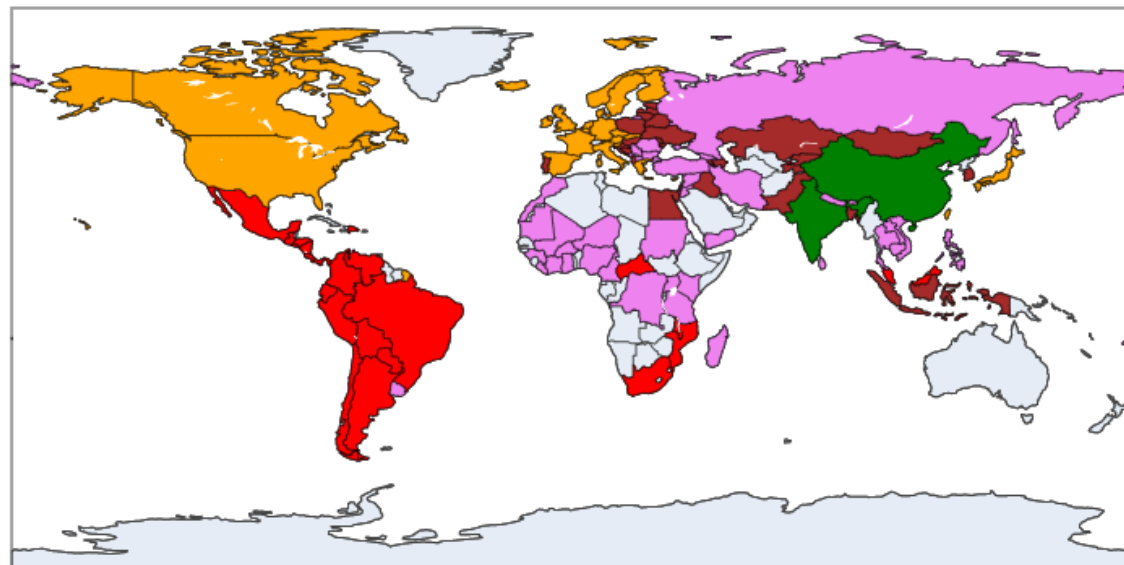
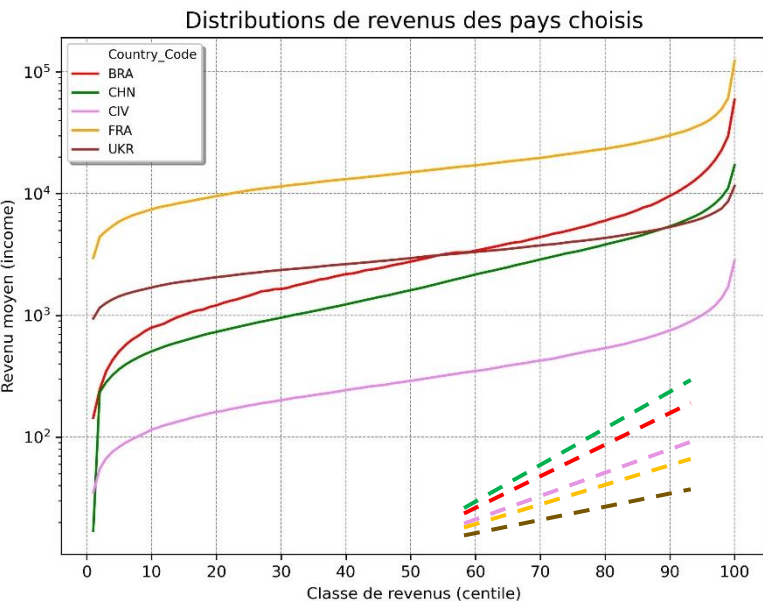
- Caractérisation des clusters suivant les variables



- On retrouve CHN et IND caractérisées par leur population et des inégalités
- Cluster 3 regroupe les pays les plus inégalitaires
- Cluster 1 regroupe les pays les plus développés (au fort pouvoir d'achat)

Classification & illustrations

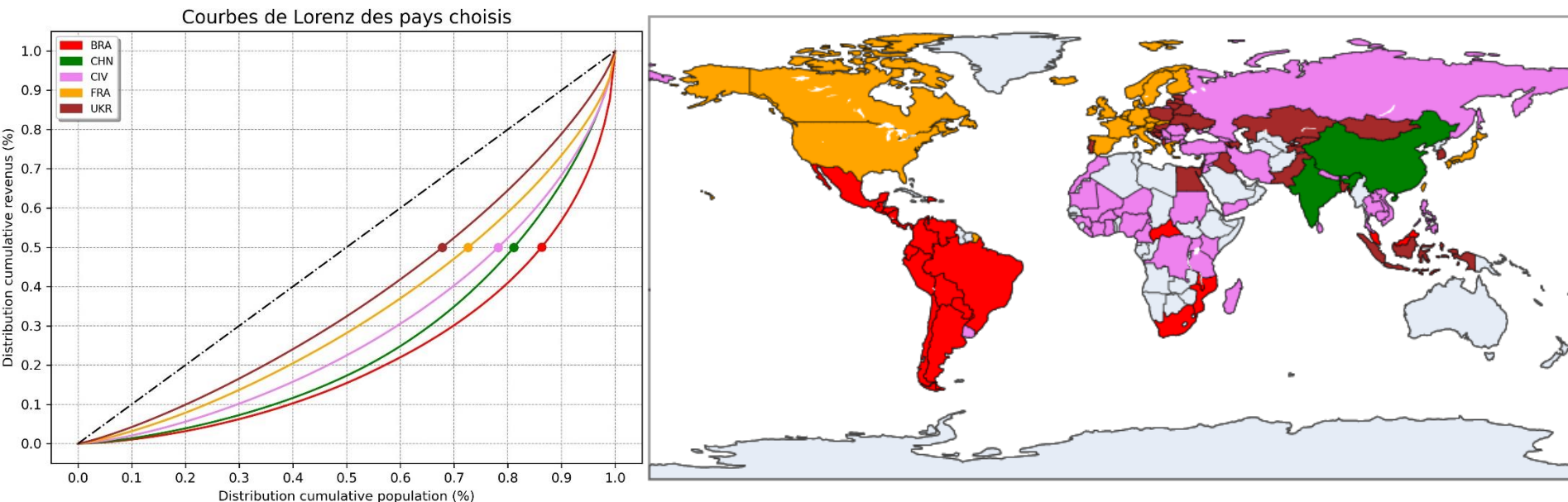
- Différenciation des revenus en absolu et les écarts observés entre classe (pente)



- Ecarts importants entre classes les plus pauvres (p1 vs p2 «CHN») et les plus riches (p99 vs p100 «BRA»)

Classification & illustrations

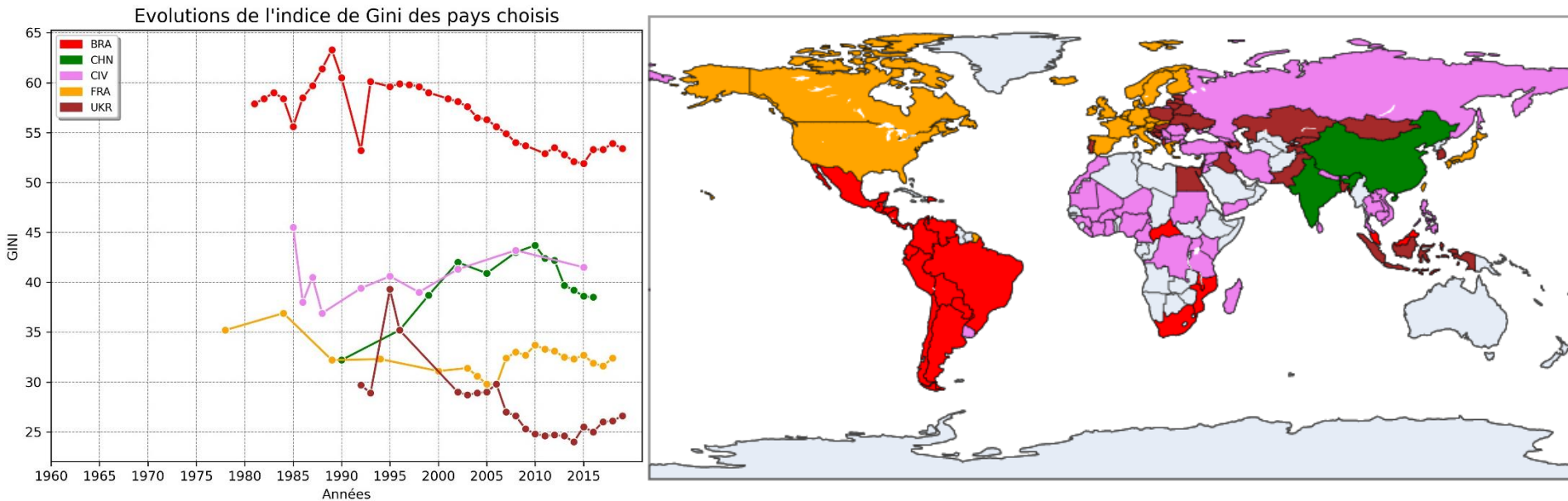
- Courbes de Lorenz



- 13% des brésiliens les plus riches possèdent 50% des richesses du pays
- La répartition de la richesse en Ukraine est « plus égalitaire »

Classification & illustrations

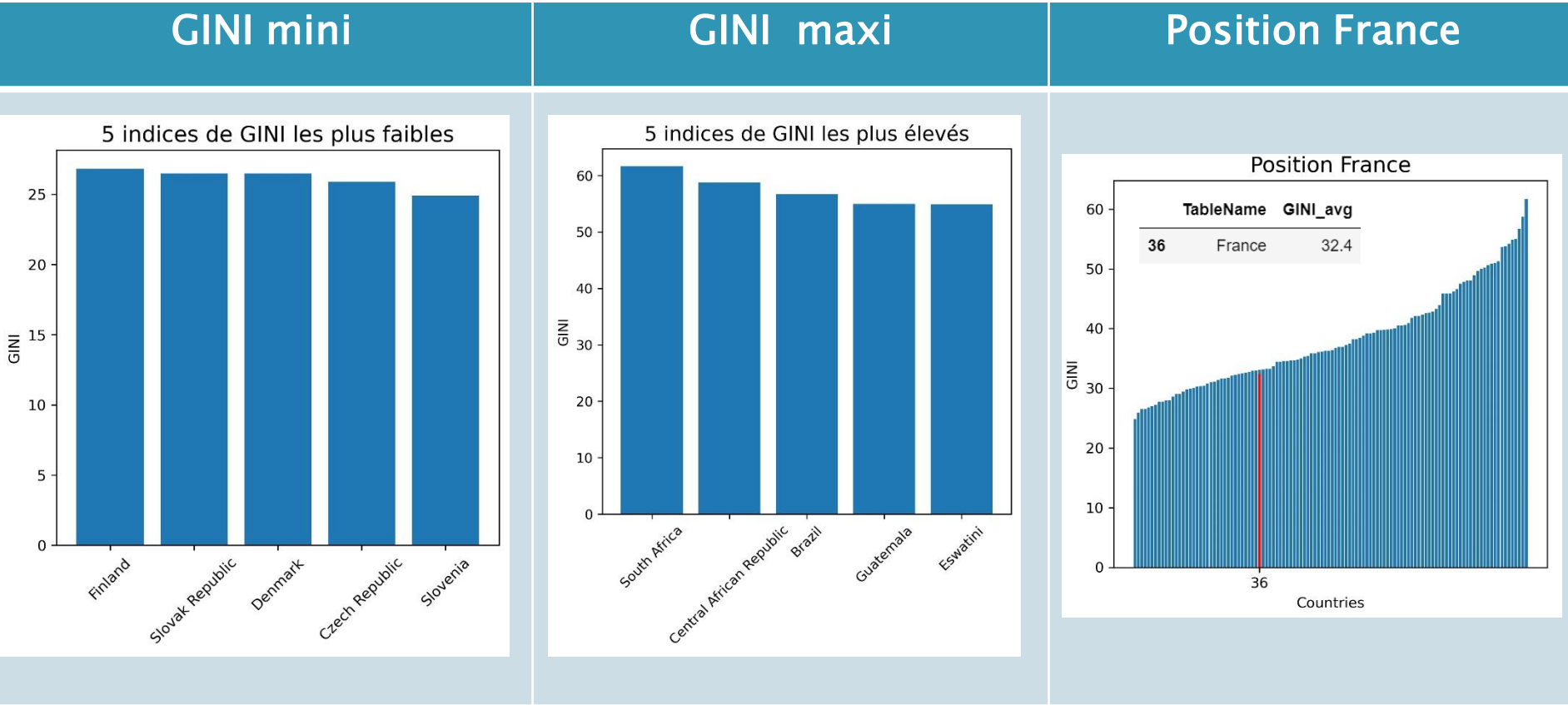
- Evolution de l'indice de GINI



- GINI conforme aux Lorenz

Classification & illustrations

- Classement par indice de GINI



- Rappel : prédiction par régression basée sur 3 variables (2 sont disponibles)
 - Objectif : générer la classe de revenu du parent (3ème variable)

Mission 3: Génération de données

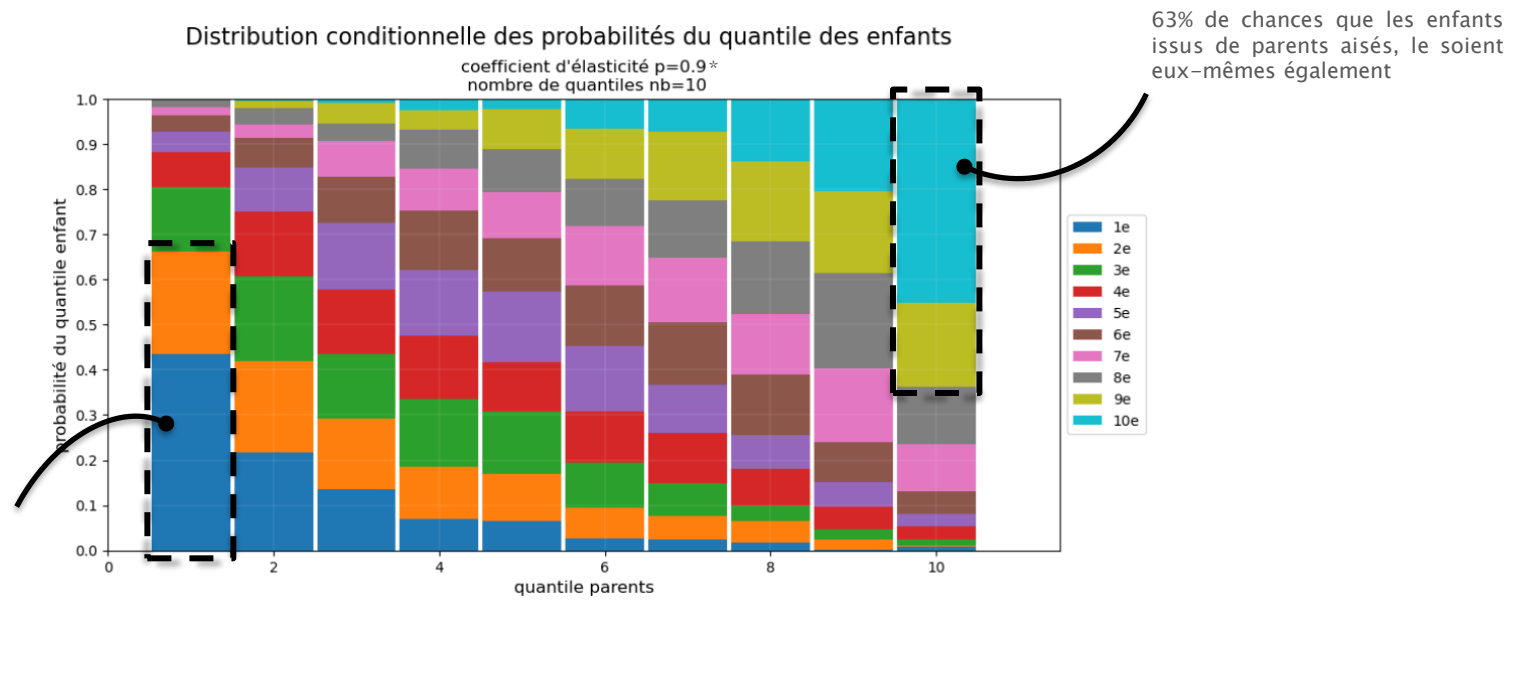
*Introduction de la notion d'élasticité
intergénérationnelle*

*Simuler des données manquantes utiles pour la
régression finale*

$$\ln(Y_{child}) = \alpha + \beta_1 \ln(Y_{parent}) + \varepsilon$$

Algorithme de génération

- Interprétation de la distribution conditionnelle « faible mobilité » (ex : *pays africains*)



- Une distribution conditionnelle « forte mobilité » verra une répartition plus équilibrée des chances d'être soit favorisé soit défavorisé (ex : *pays scandinaves*)

* % revenus transmis en moyenne entre générations

Méthodologie

- Clonage données $df \times 500 \rightarrow 5,8M$ rows
- Récupération coefficients d'élasticité (116)
- Initialisation dataframe vide avec var [y_child , y_parents , c_i_child , c_i_parent]
- For Loop avec {nb_q = 100, size = $500 \times nb_q$ }

```
for i in range(len(coeff)):
    pj = coeff[i]
    y_child, y_parents = generate_incomes(n, pj)
    sample = compute_quantiles(y_child, y_parents, nb_quantiles)
    sample_init = sample_init.append(sample)
```

- Obtention Sample $\rightarrow 5,8M$ rows
- Join variable [Country_Code]
- Correction type (object \rightarrow int)
- Merge du df principal à Sample sur [quantile] + [Country_Code]
- Conservation des variables utiles à la mission 4

	Country_Code	c_i_parent	income	gdpppp	GINI_avg	IGEincome
0	ALB	66	4670.0	7297.0	31.411111	0.815874
1	ALB	40	1087.0	7297.0	31.411111	0.815874
2	ALB	10	2938.0	7297.0	31.411111	0.815874
3	ALB	31	2892.0	7297.0	31.411111	0.815874
4	ALB	11	2058.0	7297.0	31.411111	0.815874

- A ce stade, obtention d'un dataset regroupant les variables souhaitées pour la mission 4
 - Objectif : expliquer le revenu des individus en fonction de plusieurs variables explicatives (pays, l'indice de Gini, la classe de revenus des parent)

Mission 4: Régression linéaire

Montrer une relation de dépendance entre une variable à expliquer et une série de variables explicatives

Modéliser cette association par un modèle mathématique

- Avant toute chose, il est intéressant de se demander si la variable cible peut et/ou doit-être utilisée telle quelle...

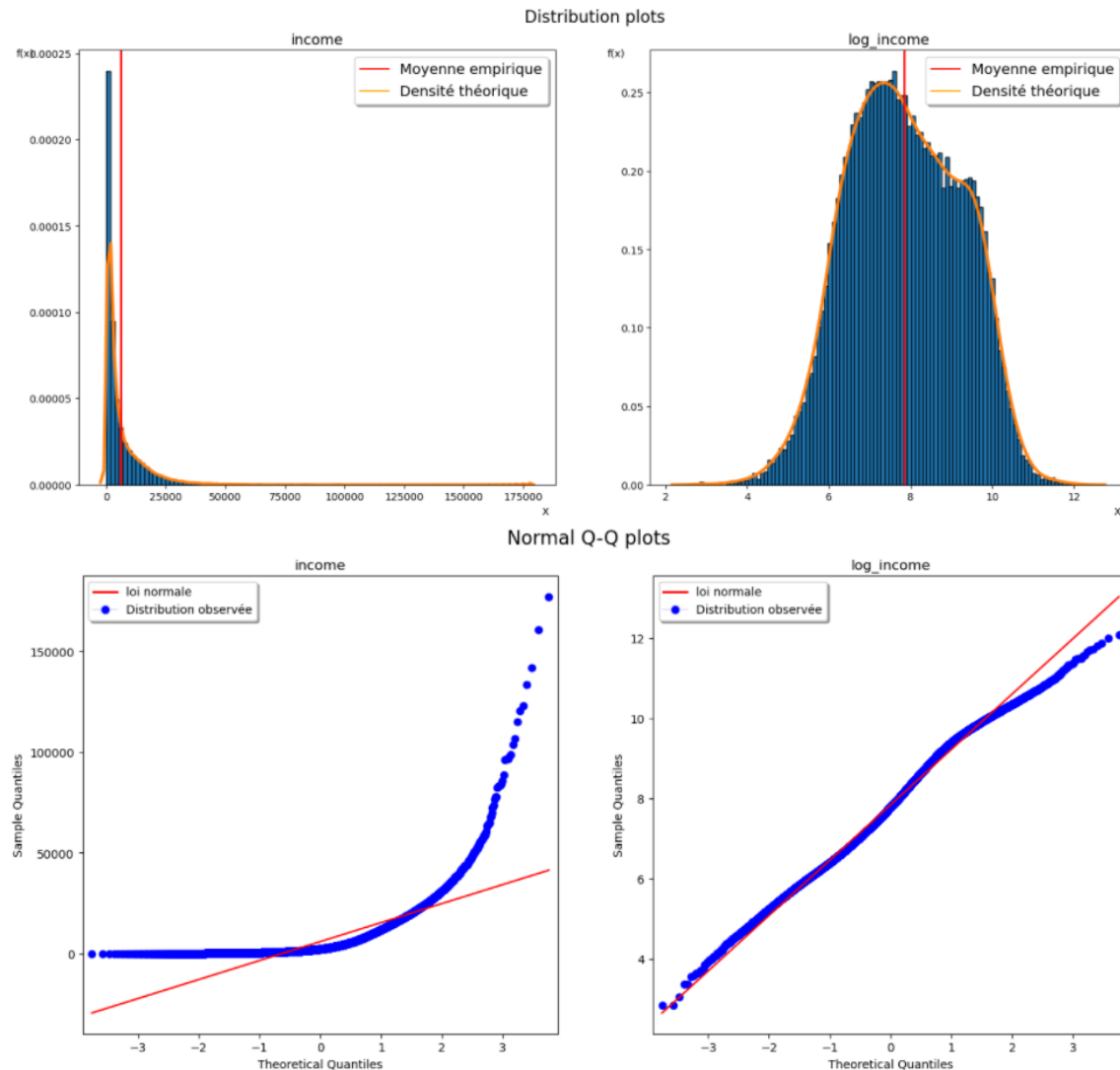
Présentation des revenus sous 2 formes (brute vs log)

- Bimodalité d'une distribution : indication forte de distribution non-normale.*

	income	log_income
W_statistic	0.598772	0.991711
p_value	0.000000	0.000000
Normality	#	#



- On retiendra la version en log plus proche d'une loi normale



ANOVA : Ordinary Least Squares regression (ols)

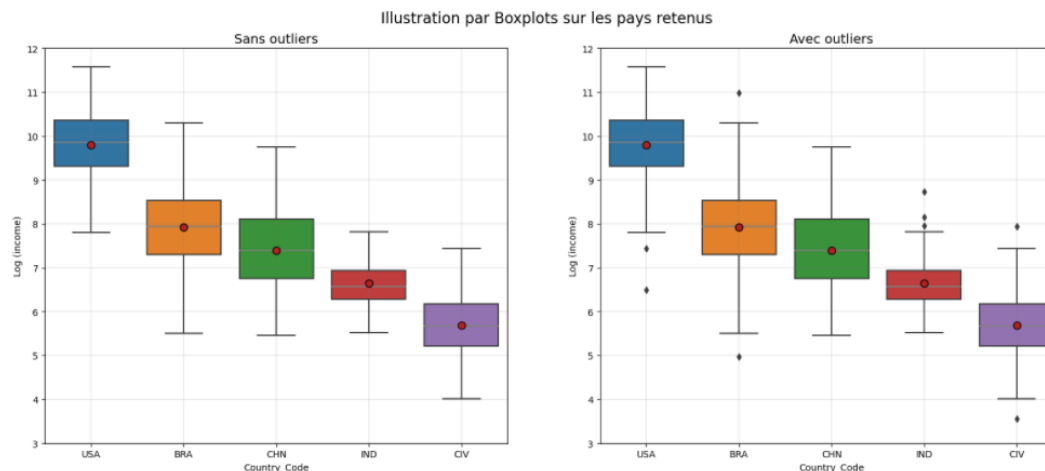
- Application d'une ANOVA sous Statsmodels

```
anova_vln = sm.tools.add_constant(anova_vln)
X = "log_income"
Y = "Country_Code"
# Ordinary Least Squares regression
# 1st fit linéaire (ite 1)
model_0 = ols('log_income ~ Country_Code', data=anova_vln).fit()
```

- Lecture des résultats, statistique de Fisher ($F \gg 1$) et $\eta^2 \gg$

LOG_INCOME vs COUNTRY_CODE						
	sum_sq	df	F	PR(>F)	EtaSq	
Country_Code	16134.136079	115.0	268.953709	0.0	0.729238	
Residual	5990.506186	11484.0	NaN	NaN	NaN	
F statistic = 269.0						
p-value for F statistics = 0.0						
η^2 = 0.729						

- Influence réelle du pays sur les revenus et forte intensité de la corrélation



Performance du modèle suite à régression (facultatif)

- Régression sur données catégorielles ici : [Country_Code]
- Variables transformées Country_Code[T.xxx]

OLS Regression Results						
=====						
Dep. Variable:	log_income	R-squared:	0.729			
Model:	OLS	Adj. R-squared:	0.727			
Method:	Least Squares	F-statistic:	269.0			
Date:	Mon, 23 Aug 2021	Prob (F-statistic):	0.00			
Time:	15:37:15	Log-Likelihood:	-12627.			
No. Observations:	11600	AIC:	2.549e+04			
Df Residuals:	11484	BIC:	2.634e+04			
Df Model:	115					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	7.8517	0.072	108.713	0.000	7.710	7.993
Country Code[T.ARG]	0.4503	0.102	4.409	0.000	0.250	0.651

- Itération n°2 : élimination variables transformées non-significatives (pvalue > 5%)

OLS Regression Results			
=====			
Dep. Variable:	log_income	R-squared:	0.757
Model:	OLS	Adj. R-squared:	0.755

	coef	odds ratio	pvalue	name
Country_Code[T.ROU]	-0.002038	0.997964	0.984078	ROU
Country_Code[T.JOR]	-0.013742	0.986352	0.892980	JOR
Country_Code[T.MEX]	-0.070608	0.931827	0.489401	MEX
Country_Code[T.BRA]	0.078907	1.082103	0.439818	BRA
Country_Code[T.DOM]	-0.118307	0.888423	0.246777	DOM
Country_Code[T.VEN]	-0.126321	0.881332	0.216211	VEN
Country_Code[T.ZAF]	-0.134118	0.874487	0.189165	ZAF
Country_Code[T.PER]	-0.156034	0.855530	0.126630	PER
Country_Code[T.UKR]	0.159648	1.173098	0.118075	UKR
Country_Code[T.PAN]	0.170641	1.186065	0.094820	PAN

- Amélioration du modèle de 3% avec variance expliquée ~76%

Régression multilinéaire 1

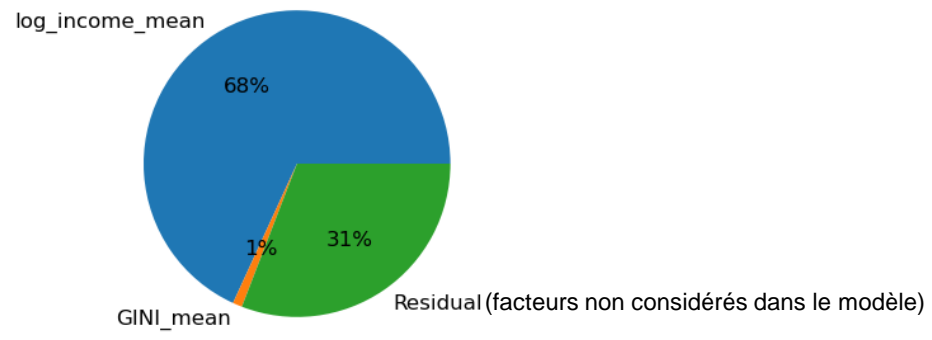
- Modèle 1 : $\log(\text{income}) \sim \log(\text{income_mean})_{|\text{pays}} + \text{GINI}_{|\text{pays}}$

OLS Regression Results						
Dep. Variable:	log_income	R-squared:	0.727			
Model:	OLS	Adj. R-squared:	0.727			
Method:	Least Squares	F-statistic:	1.545e+04			
Date:	Fri, 27 Aug 2021	Prob (F-statistic):	0.00			
Time:	16:51:45	Log-Likelihood:	-12672.			
No. Observations:	11600	AIC:	2.535e+04			
Df Residuals:	11597	BIC:	2.537e+04			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.5053	0.068	7.424	0.000	0.372	0.639
log_income_mean	0.9833	0.006	160.043	0.000	0.971	0.995
GINI_mean	-0.0167	0.001	-19.338	0.000	-0.018	-0.015



	sum_sq	df	F	PR(>F)	EtaSq
log_income_mean	13334.685346	1.0	25613.845878	0.000000e+00	0.681494
GINI_mean	194.692167	1.0	373.973216	4.964476e-82	0.009950
Residual	6037.451256	11597.0	NaN	NaN	0.308555

Décomposition de la variance du modèle multilinéaire 1



Régression multilinéaire 2

- Modèle 2 : $\log(\text{income}) \sim \log(\text{income_mean})_{\text{pays}} + \text{GINI}_{\text{pays}} + \text{classe_parent}$

OLS Regression Results

Dep. Variable: log_income
Model: OLS
Method: Least Squares
Date: Fri, 27 Aug 2021
Time: 16:53:06
No. Observations: 5800000
Df Residuals: 5799996
Df Model: 3
Covariance Type: nonrobust

R-squared: 0.774
Adj. R-squared: 0.774
F-statistic: 6.626e+06
Prob (F-statistic): 0.00
Log-Likelihood: -5.7841e+06
AIC: 1.157e+07
BIC: 1.157e+07

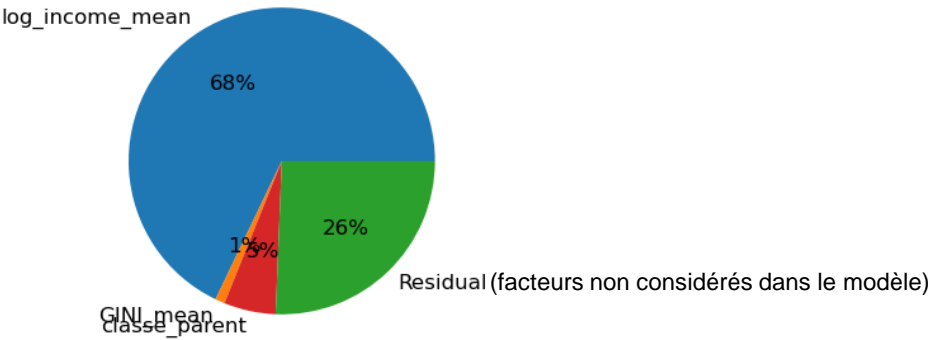
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-0.0253	0.003	-8.971	0.000	-0.031	-0.020
log_income_mean	0.9834	0.000	3918.485	0.000	0.983	0.984
GINI_mean	-0.0167	3.49e-05	-479.239	0.000	-0.017	-0.017
classe_parent	0.0105	9.44e-06	1112.047	0.000	0.010	0.011

Coef régression GINI <<0
→ impact négligeable sur la situation des personnes

↓

	sum_sq	df	F	PR(>F)	EtaSq
log_income_mean	6.606401e+06	1.0	1.535452e+07	0.0	0.678778
GINI_mean	9.881741e+04	1.0	2.296703e+05	0.0	0.010153
classe_parent	5.320772e+05	1.0	1.236648e+06	0.0	0.054669
Residual	2.495493e+06	5799996.0	NaN	NaN	0.256401

Décomposition de la variance du modèle multilinéaire 2



Cross validation (Scikit-learn)

```
from sklearn.model_selection import cross_validate
from sklearn.linear_model import LinearRegression

X = XTrain
y = yTrain
model = LinearRegression()
model.fit(X,y)

scores = cross_validate(model, X, y, cv = 5, scoring=('r2', 'neg_mean_squared_error'), return_train_score=True)
```

- Performance très proche training vs. test set (sur 5 splits)

```
Train_R² crossV values: [0.77362787 0.77198907 0.77478373 0.7759848 0.77416014]
Test_R² crossV values: [0.77589084 0.78216905 0.77120973 0.76659614 0.77352937]
```

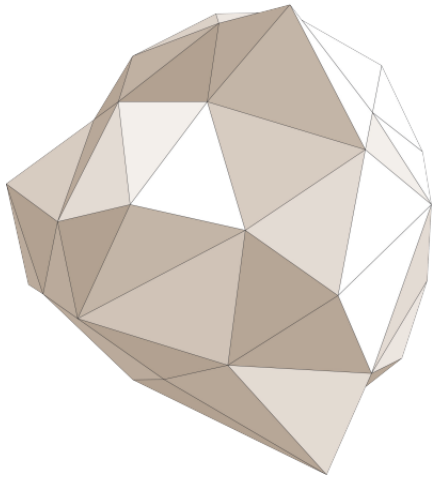
- Pas de surapprentissage observé

		test accuracy(%)	train accuracy(%)
0	Scores_min	77.2	76.7
1	Scores_mean	77.4	77.4
2	Scores_max	77.6	78.2

- Modèle conforme aux résultats obtenus sous Statsmodels

- Modèle fortement porté par le **revenu moyen du pays**
- **68% Variance expliquée avec 1 seule variable**
- Limite étude : données [classe parent] **simulées**
- **Faible poids** du GINI et classe parent
- Amélioration du modèle possible par :
 - intégration d'autres critères (éducation, typologie territoriale...)
 - étude poussée des leviers et observations influentes

Merci

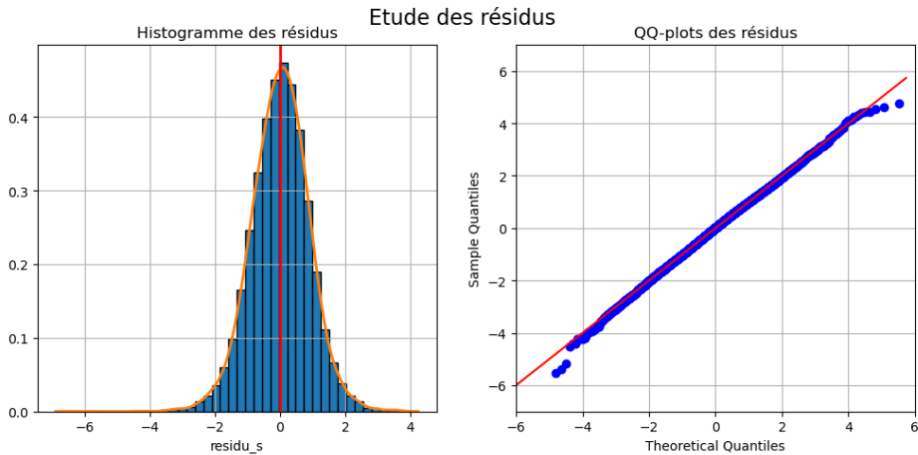


Annexes

*Normalité, Homoscédasticité et Colinéarité des
Résidus*

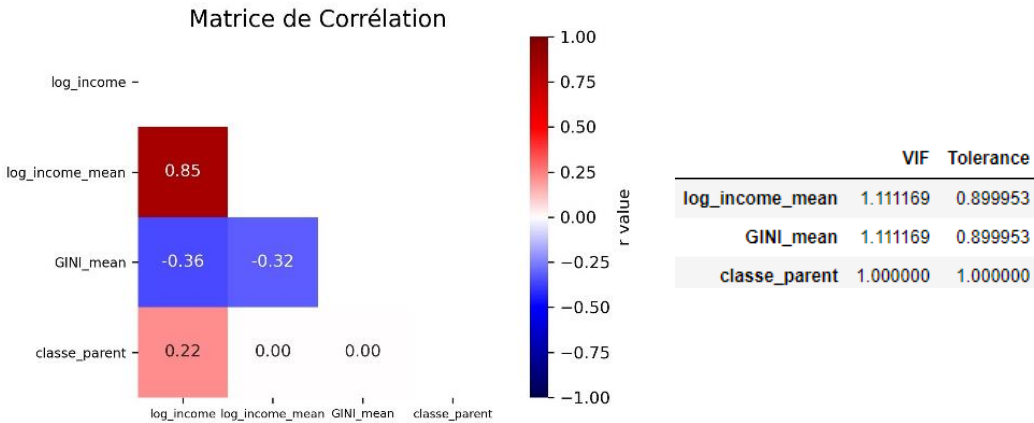
Leviers et Distance de Cook

- Etude de la normalité des résidus



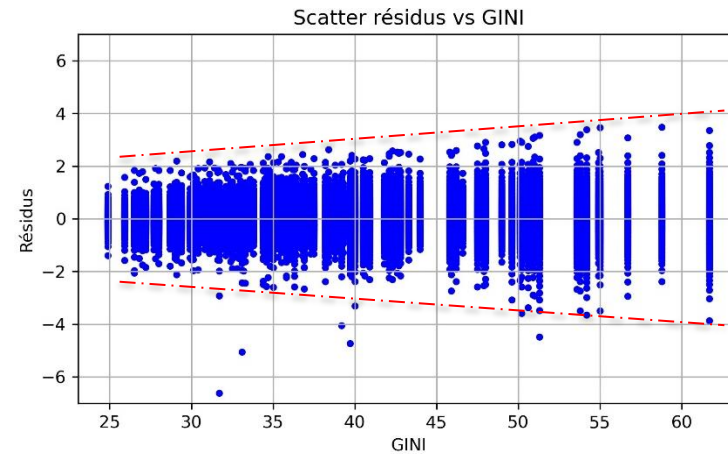
Distribution à 98% similaire à une loi normale, mais non-homogène selon la pvalue
 . Kurtosis ~ 2,97 (proche de 3 -> normal)
 . Skewness ~ -0,17 (symétrique)

- Corrélation et colinéarité



Pas de multi-colinéarité (VIF <10)

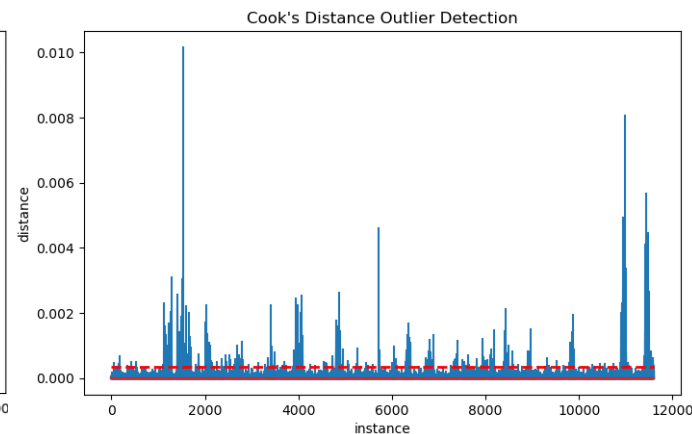
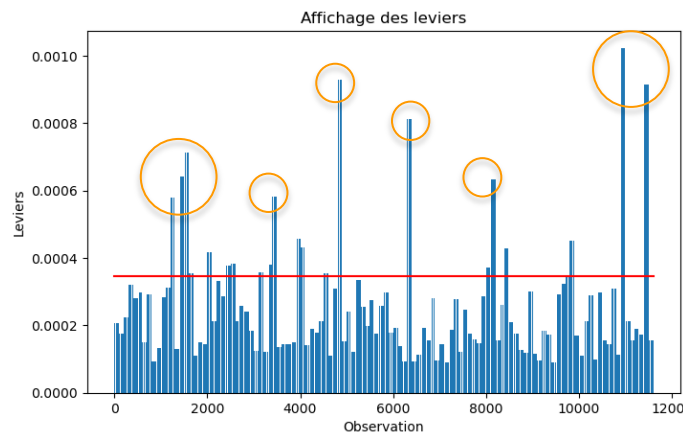
- Homoscédasticité des résidus



Variabilité des résidus change en fonction de la variable GINI → signe hétéroscédasticité

Confirmé par Test statistique de Breusch Pagan : pvalue << 5%

- Leviers et Distance de Cook



Identification des observations résiduelles à fort effet de levier

Identification des observations pouvant impacter la régression (valeurs de prédiction aberrantes)