EFFECTUEZ UNE PRÉDICTION DE REVENUS

PROBLÉMATIQUE

CIBLER DE NOUVEAUX JEUNES CLIENTS AVEC DES HAUTS REVENUS.

SOLUTION

CRÉER UN MODÈLE PERMETTANT DE DÉTERMINER LE REVENU POTENTIEL D'UNE PERSONNE.

LES SOURCE DES DONNÉES

WID.WORLD

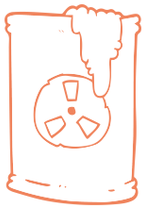
BASE DE DONNÉES SUR LES INÉGALITÉS MONDIALES PROPOSE UN ACCÈS À L'ÉVOLUTION HISTORIQUE DE LA RÉPARTITION MONDIALE DES RICHESSES.

THE WORLD BANK
IBRD • IDA

LE GROUPE BANQUE MONDIALE EST UNE SOURCE ESSENTIELLE D'APPUI FINANCIER ET TECHNIQUE POUR LES PAYS EN DÉVELOPPEMENT DU MONDE ENTIER.

Organisation des Nations Unies pour l'alimentation et l'agriculture
FAO homepage

L'ORGANISATION POUR L'ALIMENTATION ET L'AGRICULTURE (FAO) EST L'AGENCE SPÉCIALISÉE DES NATIONS UNIES QUI MÈNE LES EFFORTS INTERNATIONAUX VERS L'ÉLIMINATION DE LA FAIM.

PROBLÈMES RENCONTRÉS	SOLUTIONS
 BEAUCOUP VALEURS MANQUANTES	UTILISER « INTERPOLATE » CALCULER DES DONNÉES DE LA WID
DIFFERENCES DE NOMENCLATURES	RENOMMER LES PAYS PROBLÉMATIQUES UTILISER LE ISO 3
 DONNÉES TRÈS SALES	ELIMINER VARIABLES UTILISER « PIVOT TABLE »
TEMPS DE CALCUL BESOIN DE MÉMOIRE RAM	UTILISER GOOGLE COLAB

DF FINAL

	Country_Code	Year	Quantile	Nb_quantiles	Income	Gdpppp	Country_Name	Population	Gini
0	ALB	2008	1	100	728.89795	7297.0	Albania	3002678.0	0.32141
1	ALB	2008	2	100	916.66235	7297.0	Albania	3002678.0	0.32141
2	ALB	2008	3	100	1010.91600	7297.0	Albania	3002678.0	0.32141
3	ALB	2008	4	100	1086.90780	7297.0	Albania	3002678.0	0.32141
4	ALB	2008	5	100	1132.69970	7297.0	Albania	3002678.0	0.32141

LES VARIABLES

«INCOME » : REVENUES PAR CENTILES.

«GDPPPP » : UNITÉ QUI PERMET DE COMPARER LE POUVOIR D'ACHAT ENTRE DEUX PAYS SANS DISTORSION DUE AUX TAUX DE CHANGE.

«GINI » : INDICE DE GINI.

MISSION 1

RÉSUMER LES DONNÉES UTILISÉES

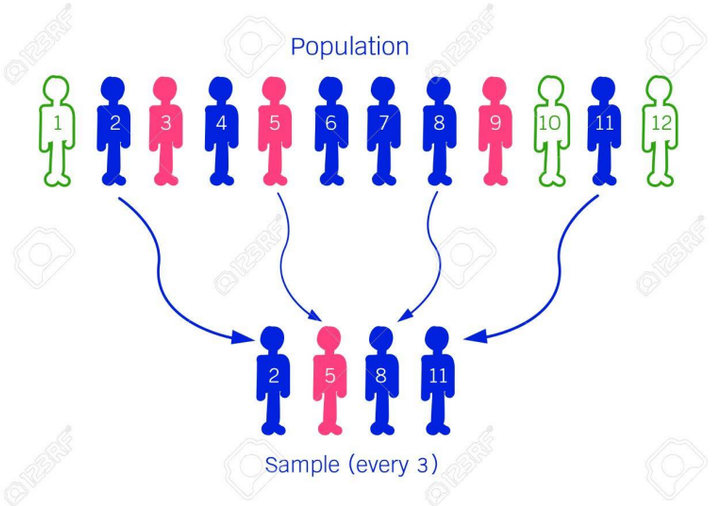
ANNÉE(S)	NOMBRE DE PAYS	POPULATION COUVERTE	%
2004	1	17827825	0
2006	5	287548000	0,2
2007	14	1374001111	0,6
2008	75	2245356494	33
2009	12	475559459	0,5
2010	6	383832444	0,3
2011	1	14948801	0

DE QUEL TYPE DE QUANTILE S'AGIT IL ?

CENTILES

PAYS → 100 QUANTILES

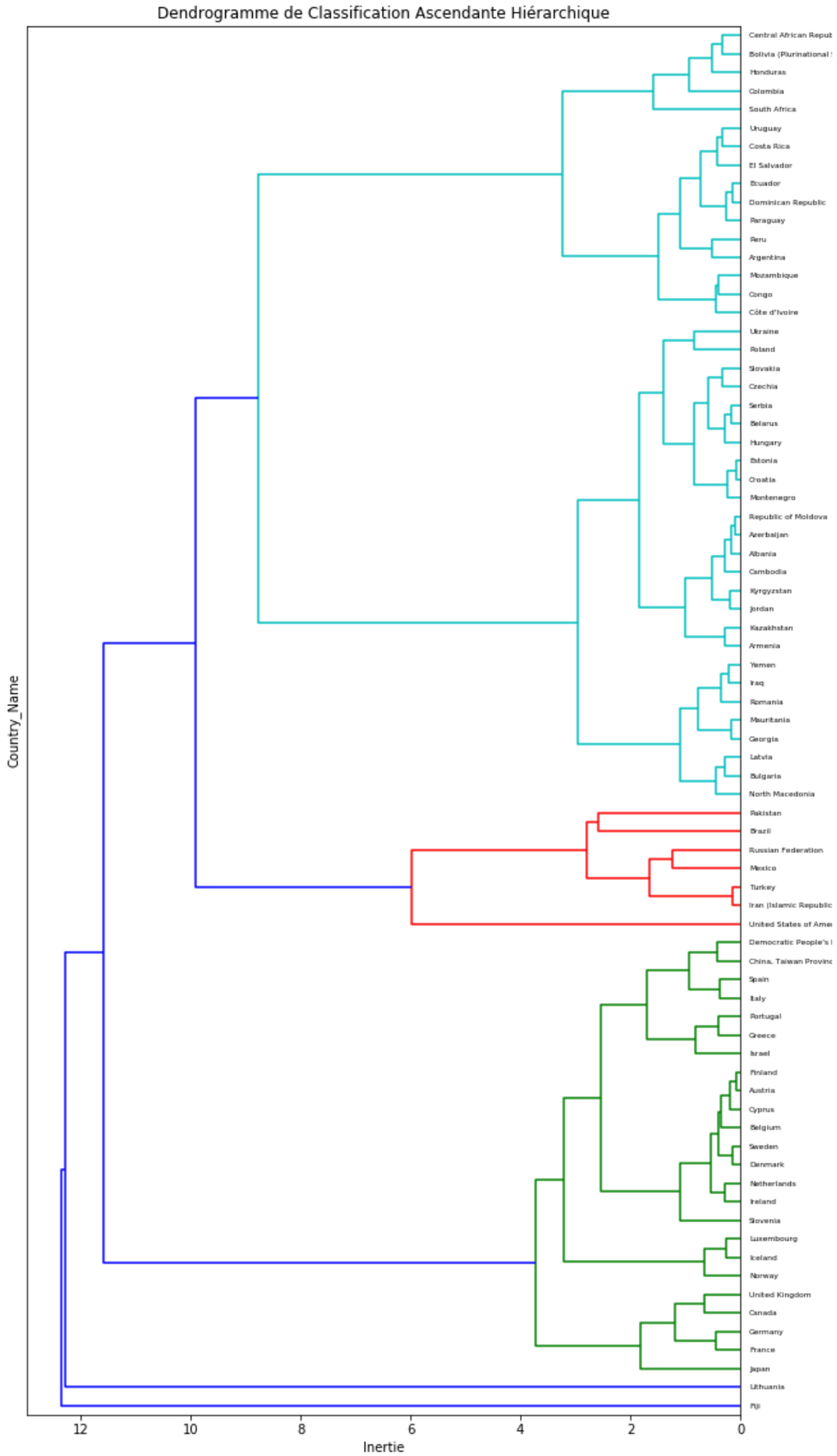
ECHANTILLONNER UNE POPULATION EST UNE BONNE MÉTHODE ?



PLUS FACILE À REPRÉSENTER

MISSION 2

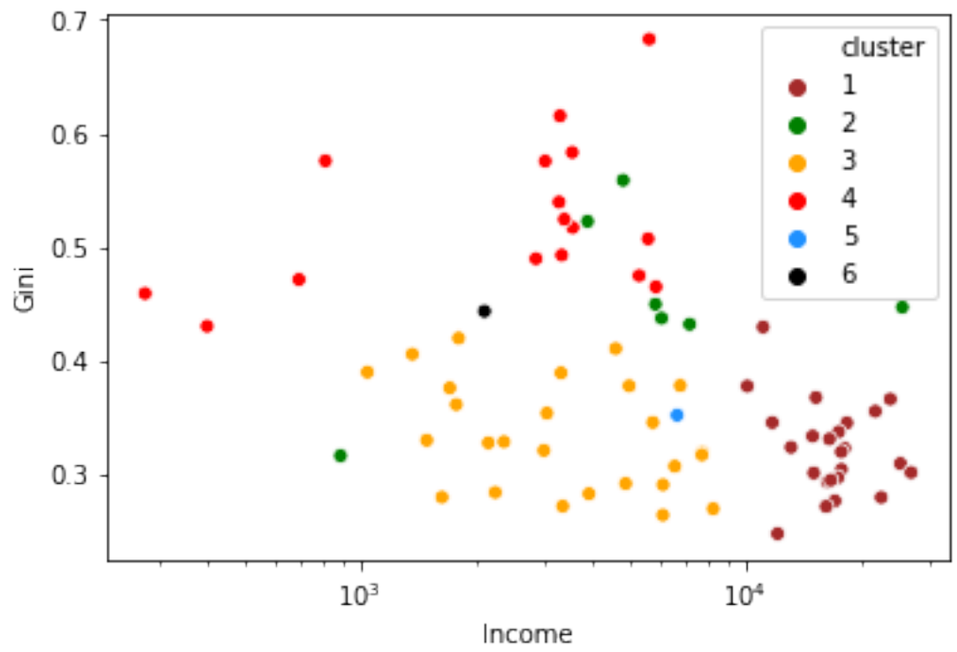
DIVERSITÉ DES PAYS- CAH



6 CLUSTER

PAYS

CENTROIDES



DIVERSITÉ DES PAYS- CHOIX DES PAYS

PAYS DONNÉS PAR LA CLASSIFICATION

KOREA

LITHUANIA

FIJI

RUSSIE

LA RÉPUBLIQUE TCHÈQUE

CÔTE D’IVOIRE

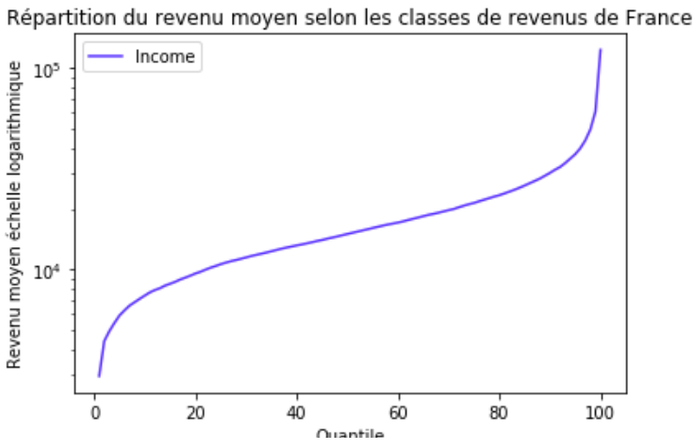
+

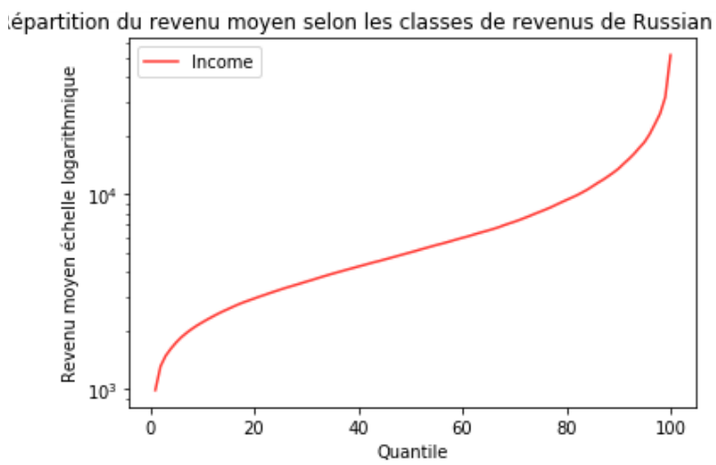
FRANCE

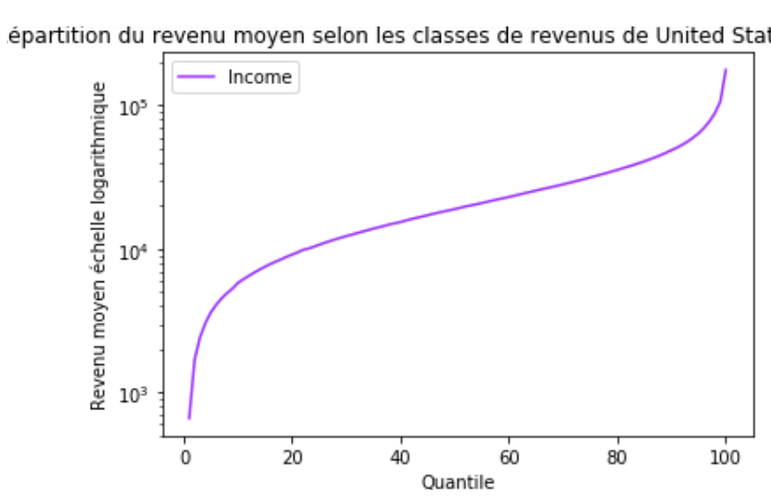
UNITED STATES

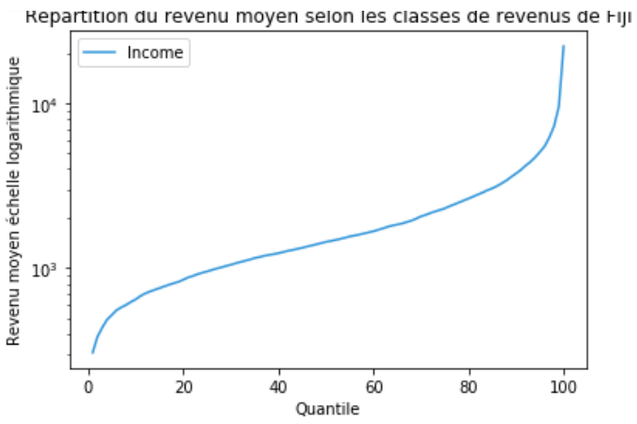
ARGENTINA

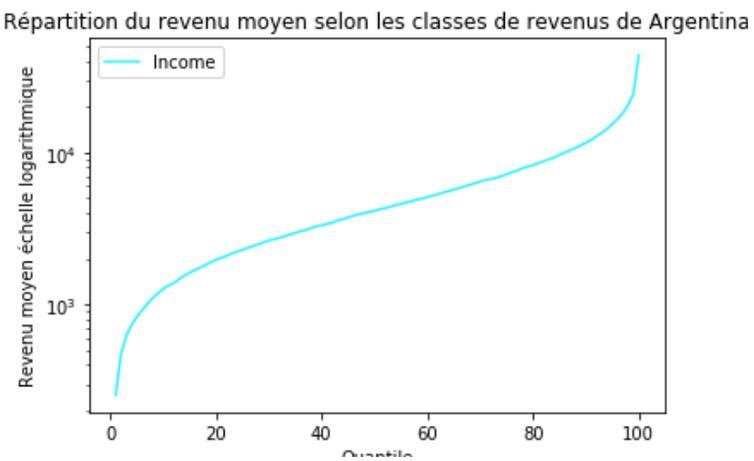
DIVERSITÉ DES PAYS- RÉPARTITION DE REVENUS

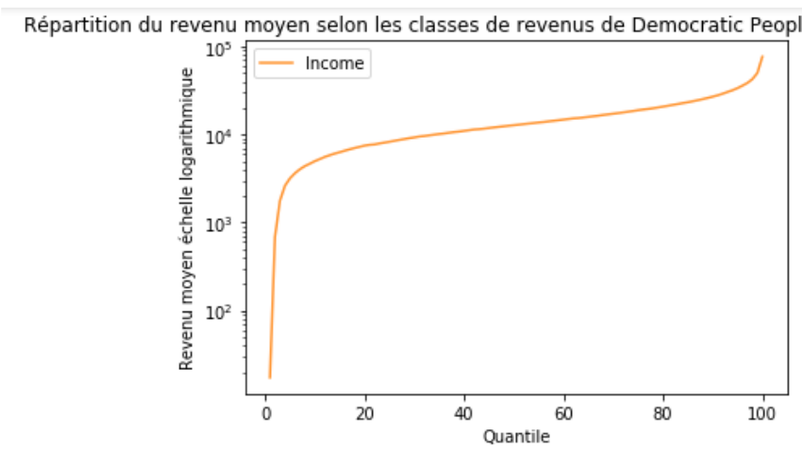


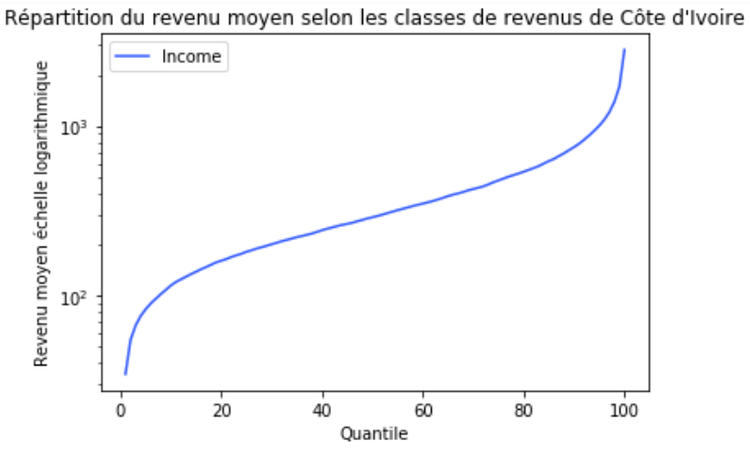


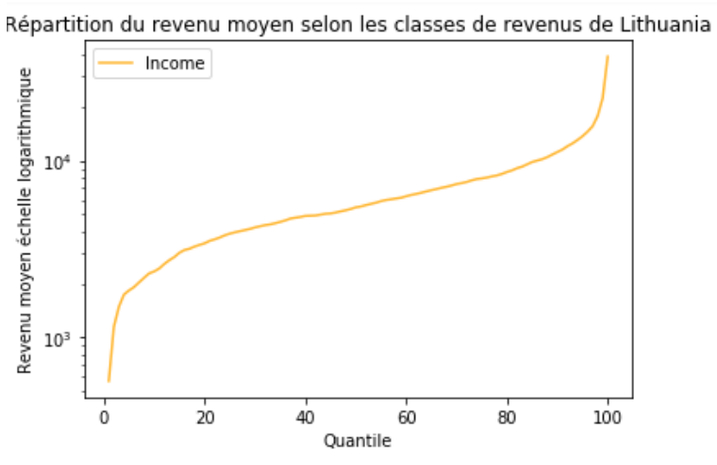


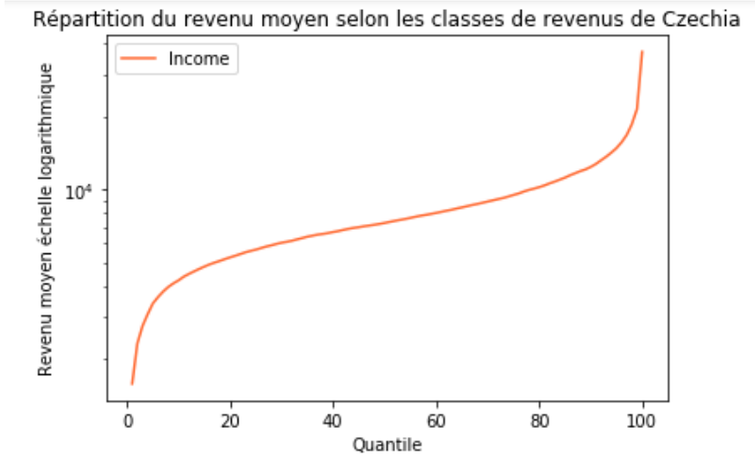




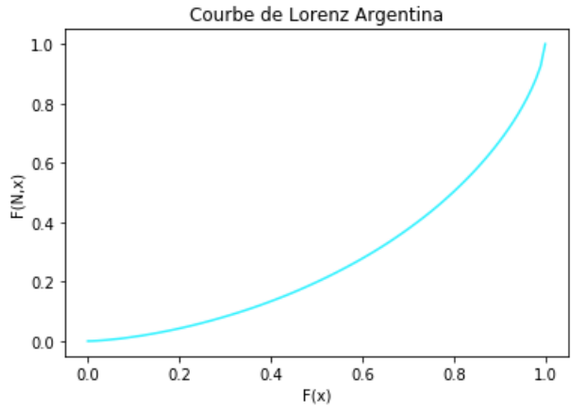




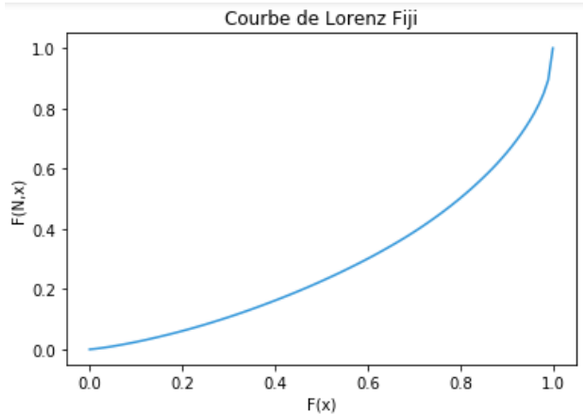




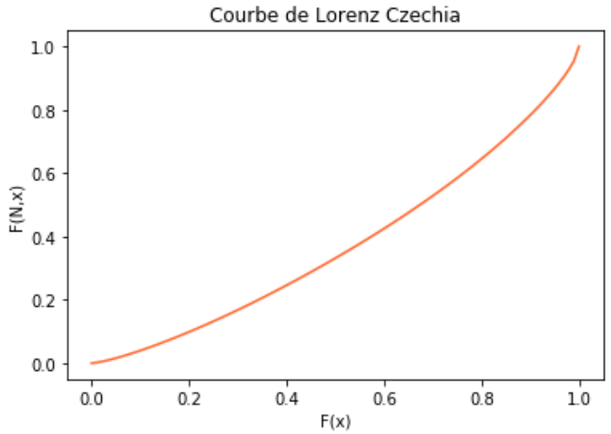
DIVERSITÉ DES PAYS- COURBE DE LORENZ



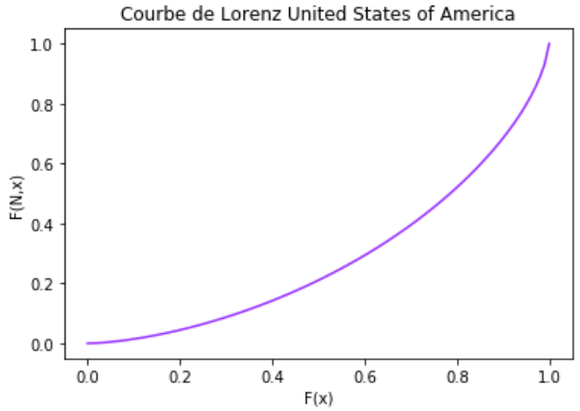
L'indice de gini est 0.47



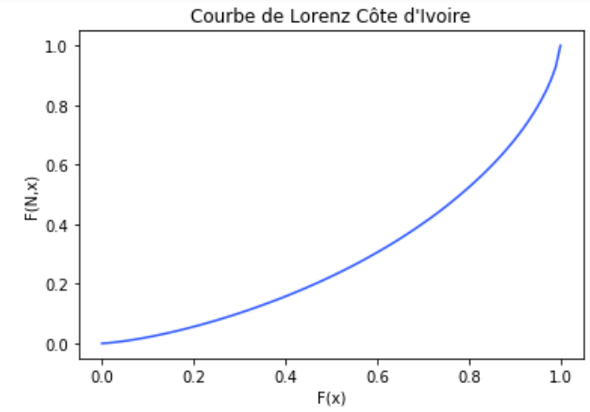
L'indice de gini est 0.44



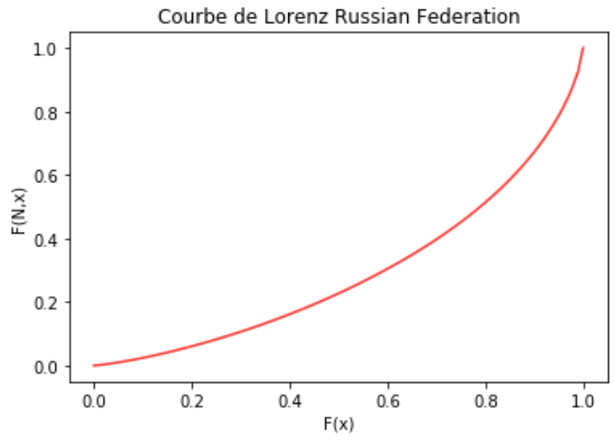
L'indice de gini est 0.27



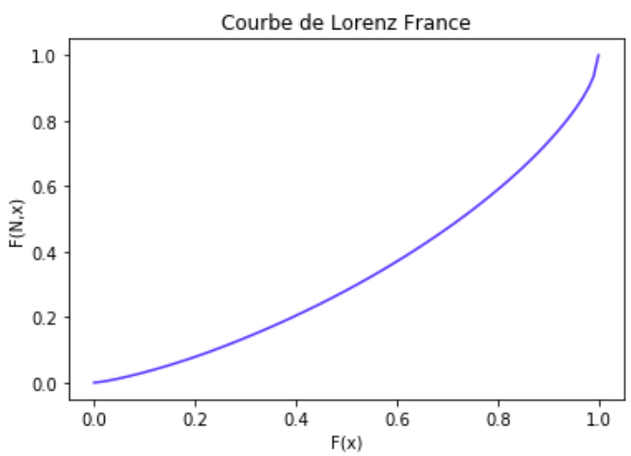
L'indice de gini est 0.45



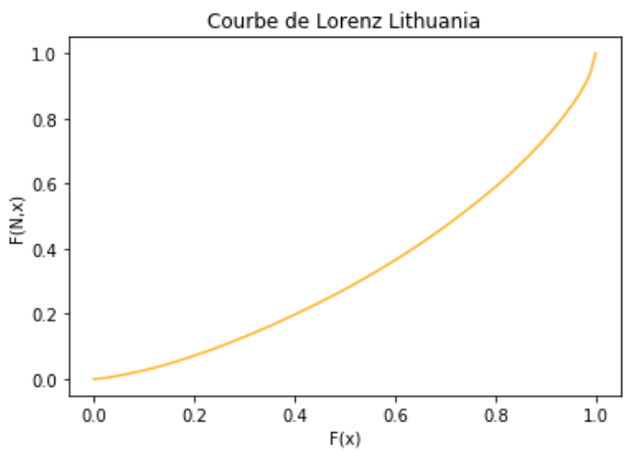
L'indice de gini est 0.43



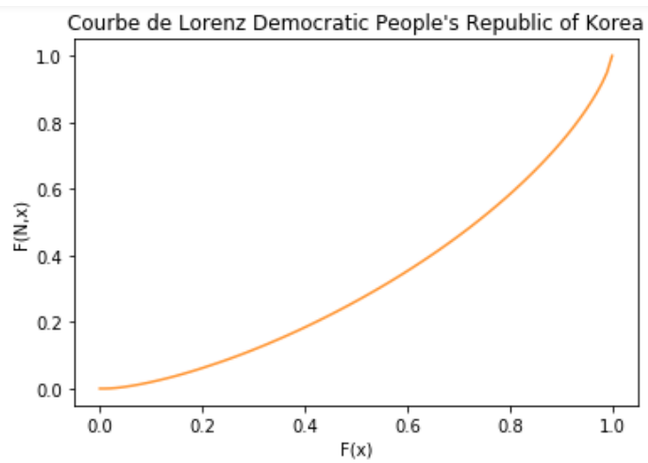
L'indice de gini est 0.43



L'indice de gini est 0.35

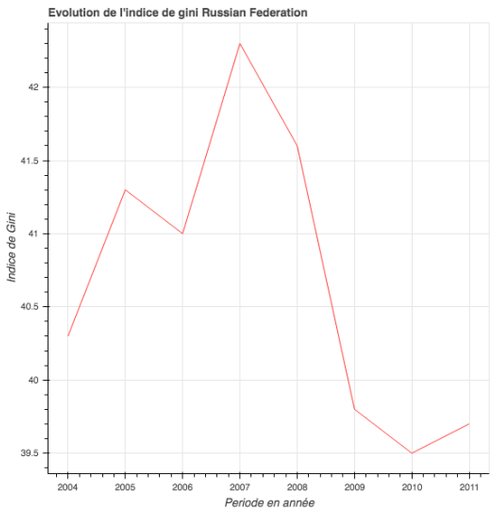


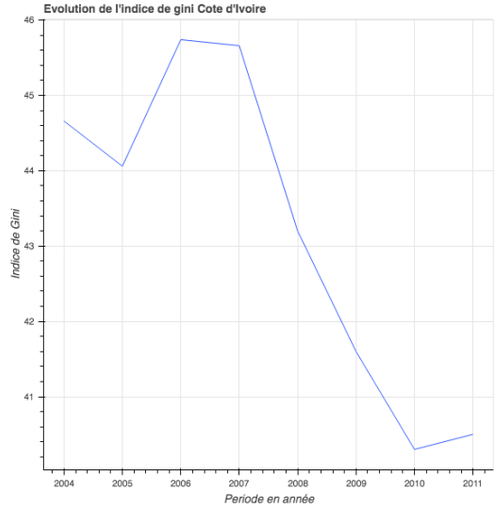
L'indice de gini est 0.35

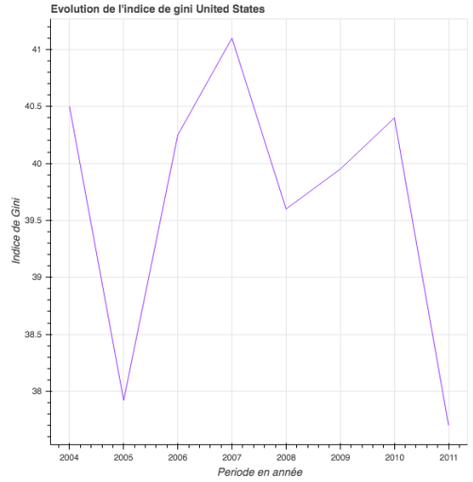


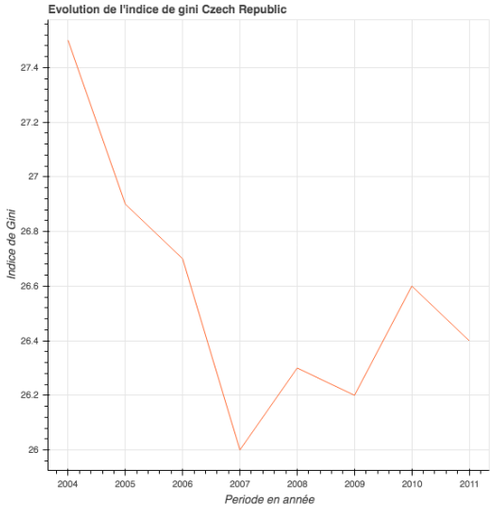
L'indice de gini est 0.37

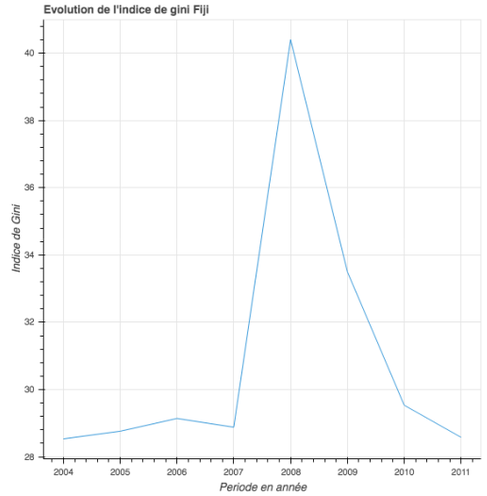
DIVERSITÉ DES PAYS- INDICE DE GINI

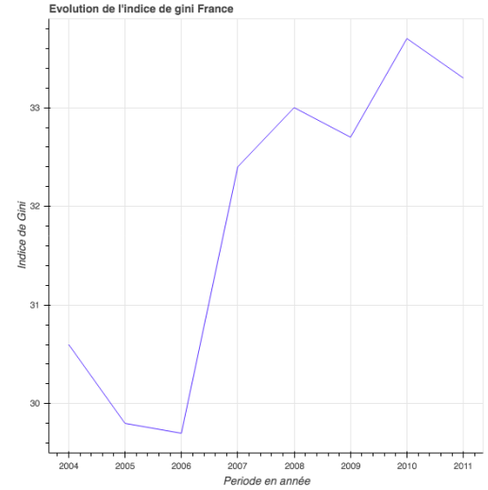


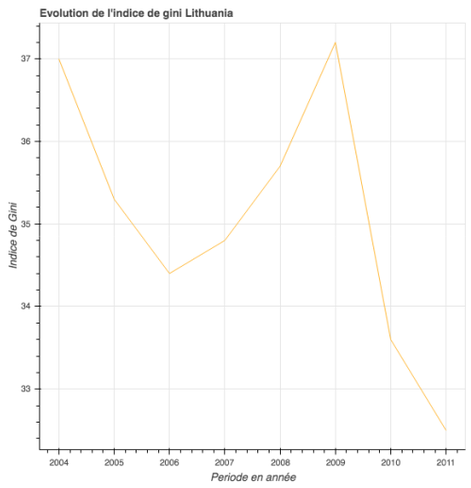


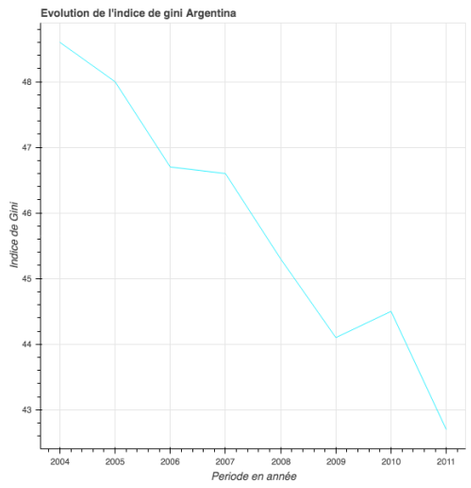












DIVERSITÉ DES PAYS- INDICE DE GINI

5 PAYS AYANT L'INDICE DE GINI LE PLUS ÉLEVÉ

HONDURAS -HAITI - PANAMA- COLOMBIA- BRAZIL

5 PAYS AYANT L'INDICE DE GINI LE PLUS FAIBLE

SLOVENIA - DENMARK -CZECH REPUBLIC -SLOVAK REPUBLIC - UKRAINE

FRANCE

	Country_Name	Year	Gini
40	France	2007.5	31.9

MISSION 3

COEFFICIENT D'ÉLASTICITÉ

IMPORT DE DATASET

↓

REPÉRER LA VARIABLE : « GEINCOME »

↓

TRAITEMENT DE VALEURS MANQUANTES

↓

DF FINAL : 75 PAYS - ANNÉE 2008

GÉNÉRATION DU DATASET PARENT/ENFANT

$$ln(Y_{child}) = \alpha + p_j \, ln(Y_{parent}) + \epsilon$$

	Country_Code	ln_yparent	epsilone	pj	ychild	cparent	cchild
0	ALB	-0.204323	-0.474299	0.800179	0.528457	42	32
1	ALB	-0.940798	-0.404884	0.800179	0.314211	18	19
2	ALB	-0.309873	-0.528270	0.800179	0.460140	38	28
3	ALB	1.522322	0.095780	0.800179	3.720698	94	85
4	ALB	-0.675399	0.138446	0.800179	0.668984	25	38

```
(7500000, 7)
```

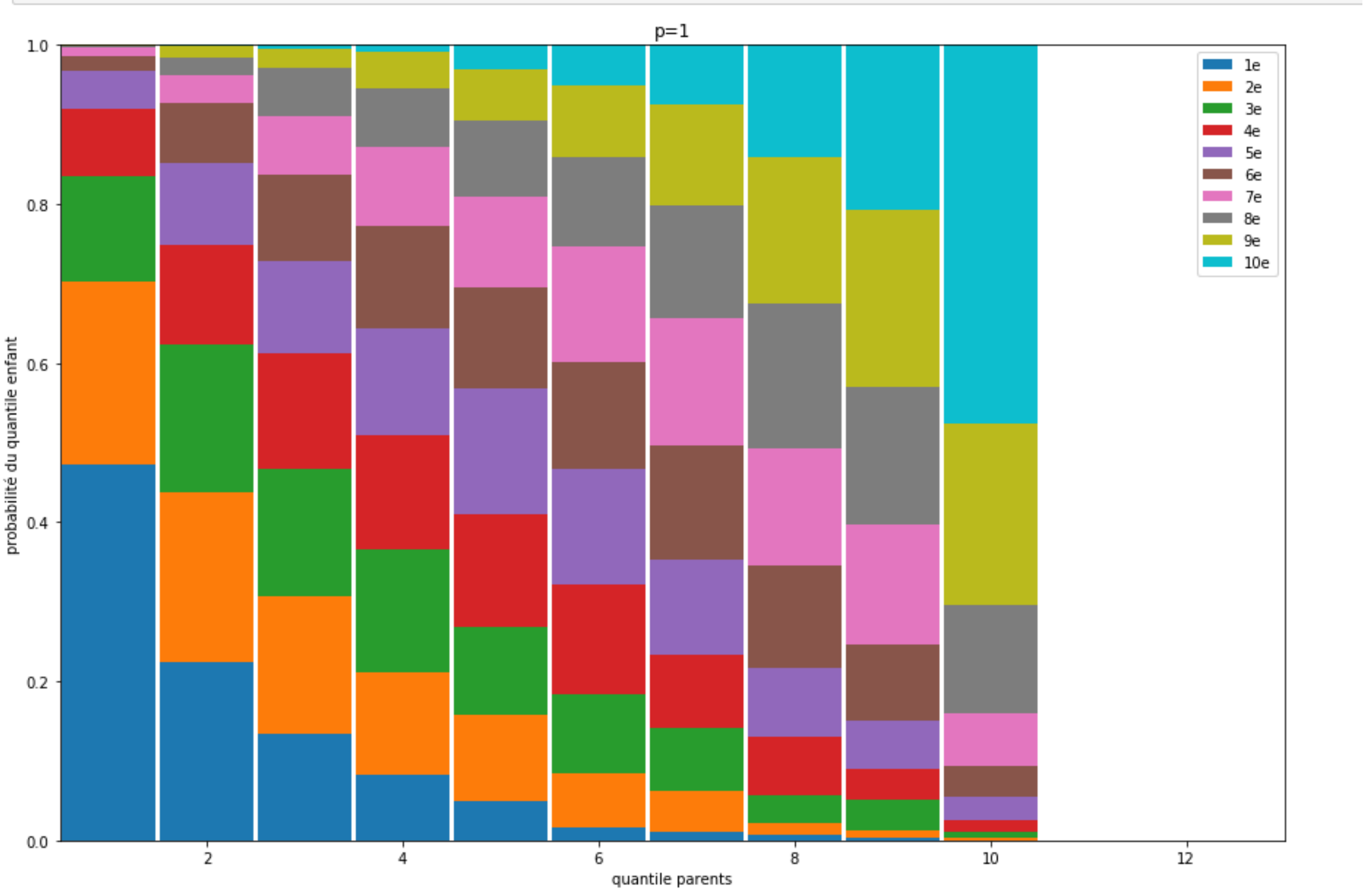
GÉNÉRATION DE PROBABILITÉS CONDITIONNELLES

$$P(c_{i,parent} | c_{i,child}, j) .$$

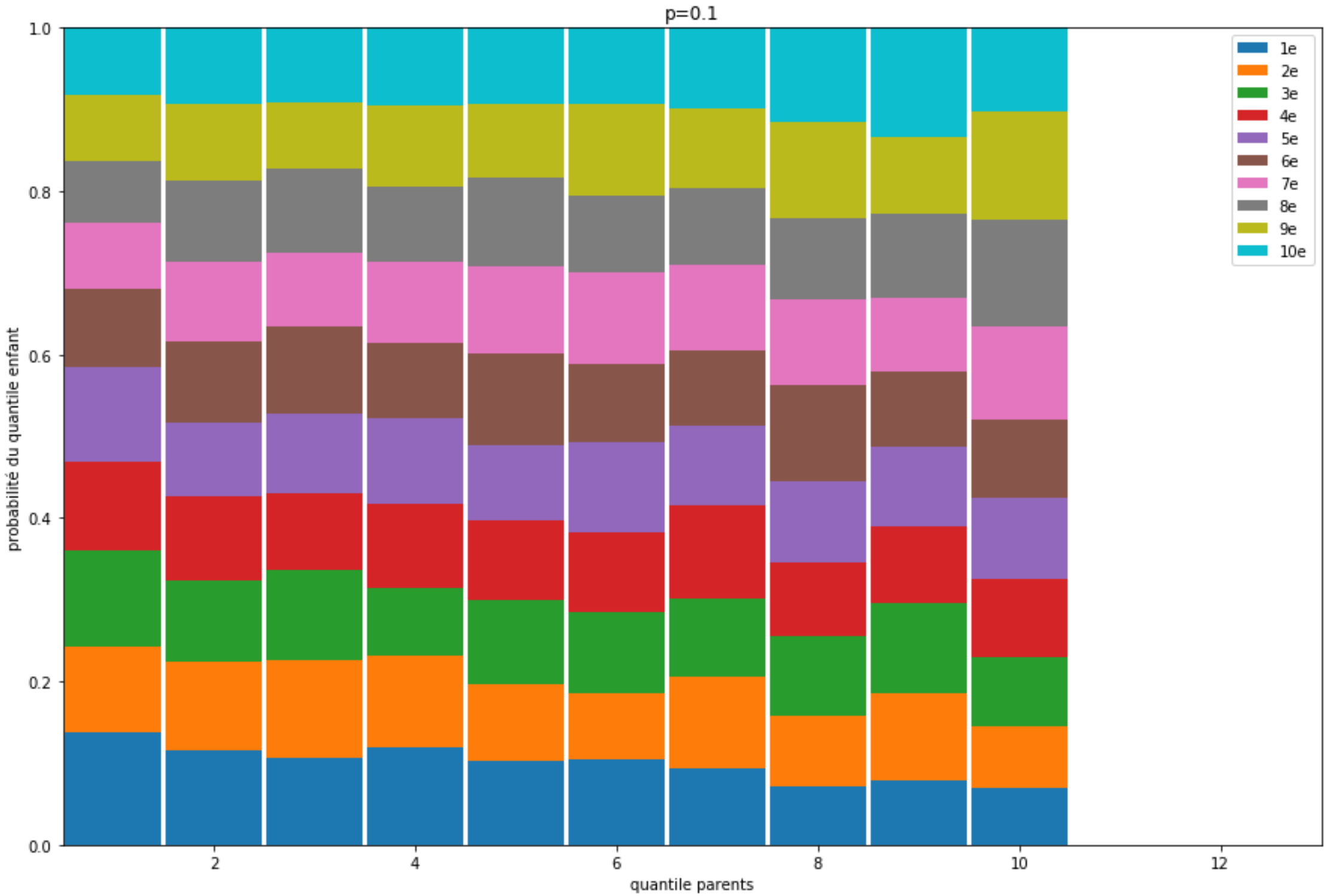
(750000, 4)

	Country_Code	c_child	c_parent	prob
0	ALB	1	1	0.204204
1	ALB	1	2	0.123123
2	ALB	1	3	0.069069
3	ALB	1	4	0.072072
4	ALB	1	5	0.054054

PROBABILITÉS CONDITIONNELLES : COLOMBIE (PJ = 1)



PROBABILITÉS CONDITIONNELLES : FINLANDE (PJ = 0, 1)



GÉNÉRATION DE DATASET FINAL

DF_WID

500 FOIS PLUS

DF_FINAL

	Country_Code	Country_Name	Quantile	c_parent	Population	Gini	Gdpppp	Income
0	ALB	Albania	1	1	3002678.0	0.32141	7297.0	728.89795
1	ALB	Albania	1	1	3002678.0	0.32141	7297.0	728.89795
2	ALB	Albania	1	1	3002678.0	0.32141	7297.0	728.89795
3	ALB	Albania	1	1	3002678.0	0.32141	7297.0	728.89795
4	ALB	Albania	1	1	3002678.0	0.32141	7297.0	728.89795

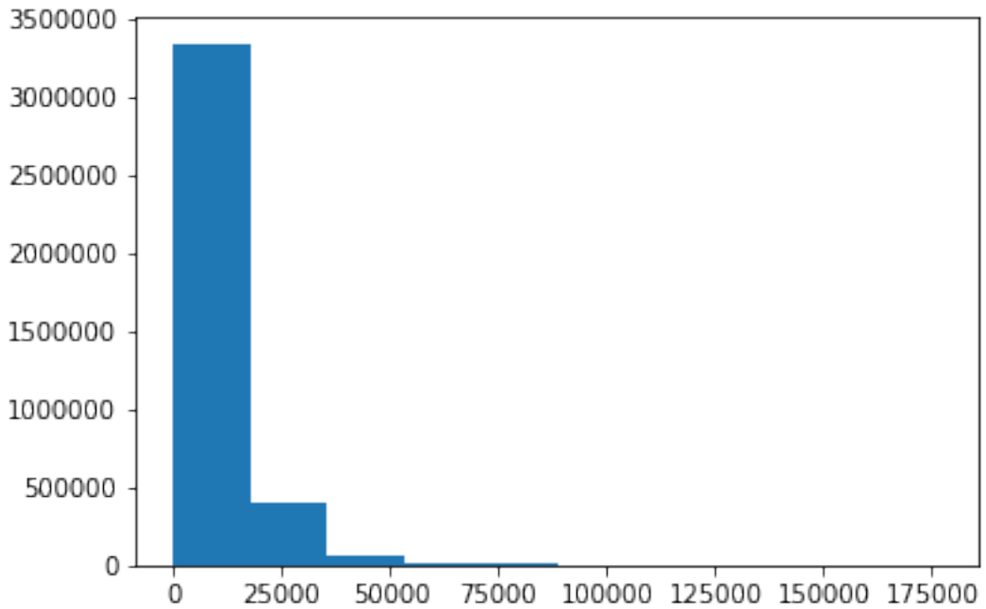
MISSION 4

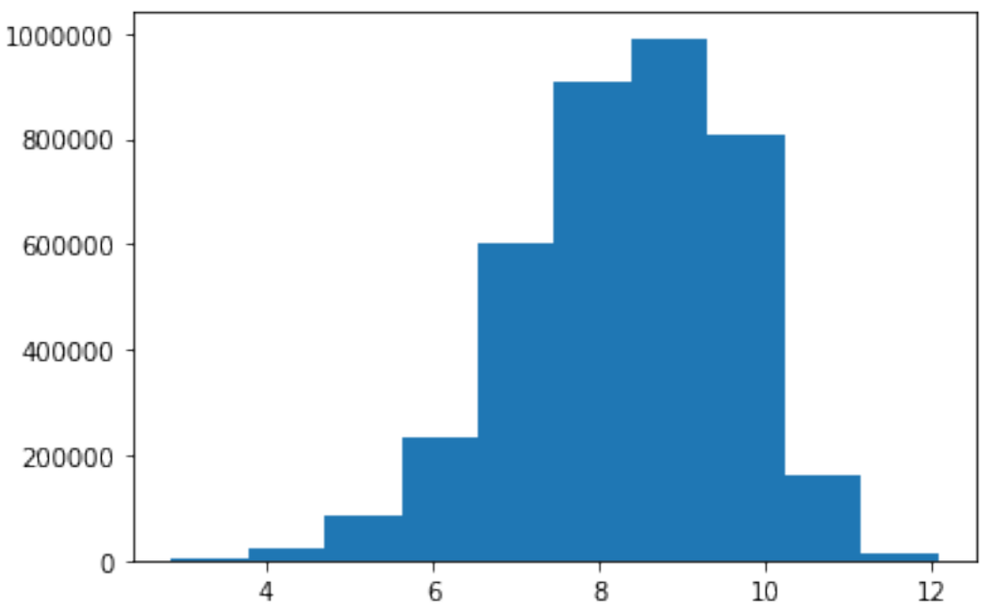
ANOVA - TEST DE NORMALITÉ DE LA VARIABLE « INCOME »

AVANT

LOG

APRÈS



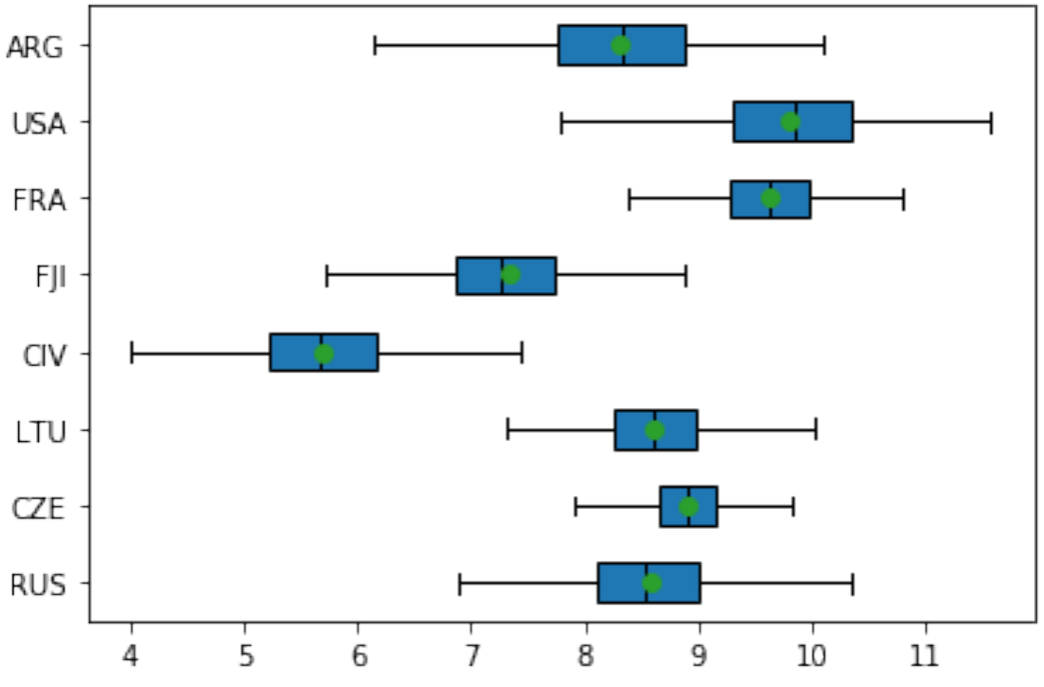


PAS GAUSSIENNE

ANALYSE DE LA VARIANCE

H0 : LES VARIABLES SONT INDÉPENDANTES (LE FACTEUR N'A AUCUNE INFLUENCE SUR LA VARIABLE DÉPENDANTE.

H1 : LES VARIABLES SONT DÉPENDANTES (LE FACTEUR A UNE INFLUENCE SUR LA VARIABLE DÉPENDANTE.



	df	sum_sq	mean_sq	F	PR(>F)
Country_Code	7.0	2.475397e+13	3.536281e+12	30568.034086	0.0
Residual	397336.0	4.596605e+13	1.156856e+08	NaN	NaN

ANALYSE DE LA VARIANCE

INTERPRÉTATIONS

LA P-VALUE EST INFINIMENT FAIBLE, LE PAYS EST DONC UN FACTEUR D'INFLUENCE STATISTIQUEMENT FIABLE.

LE MODÈLE PERMET SEULEMENT D'EXPLIQUER 35% DE LA SOMME DES CARRÉS

OLS Regression Results

Dep. Variable:	Income	**R-squared:**	0.350
Model:	OLS	**Adj. R-squared:**	0.350
Method:	Least Squares	**F-statistic:**	3.057e+04
Date:	Thu, 16 Jan 2020	**Prob (F-statistic):**	0.00
Time:	15:08:46	**Log-Likelihood:**	-4.2524e+06
No. Observations:	397344	**AIC:**	8.505e+06
Df Residuals:	397336	**BIC:**	8.505e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5841.5480	48.281	120.992	0.000	5746.920	5936.176
Country_Code[T.CIV]	-5442.4767	68.281	-79.707	0.000	-5576.306	-5308.647
Country_Code[T.CZE]	2392.4926	68.246	35.057	0.000	2258.733	2526.252
Country_Code[T.FJI]	-3747.1179	68.196	-54.946	0.000	-3880.781	-3613.455
Country_Code[T.FRA]	1.247e+04	68.240	182.671	0.000	1.23e+04	1.26e+04
Country_Code[T.LTU]	794.6305	68.390	11.619	0.000	660.588	928.673
Country_Code[T.RUS]	1300.7921	68.220	19.068	0.000	1167.084	1434.501
Country_Code[T.USA]	1.962e+04	68.274	287.370	0.000	1.95e+04	1.98e+04

Omnibus:	461739.538	**Durbin-Watson:**	0.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74312252.428
Skew:	6.020	**Prob(JB):**	0.00
Kurtosis:	68.906	**Cond. No.**	8.89

RÉGRESSION LINÉAIRE

MODELE 1

X_TRAIN= GINI + MEAN_INCOME

Y_TRAIN = INCOME

COEFFICIENTS: -22.36358521 0.99985338

RESIDUAL SUM OF SQUARES: 62803839.49

VARIANCE SCORE: 0.44

INTERPRÉTATIONS

R2 ÉTANT ENVIRON ÉGAL À 45%.

L'INDICE DE GINI NE SEMBLE PAS SIGNIFICATIF.

L'ANALYSE MONTRE QUE LES DONNÉES NE SONT PAS LINÉAIRES.

RÉGRESSION LINÉAIRE

MODELE 2

X_TRAIN= GINI + MEAN_INCOME_LOG

Y_TRAIN = INCOME_LOG

COEFFICIENTS: -1.6712121 0.98831956

RESIDUAL SUM OF SQUARES: 0.51

VARIANCE SCORE: 0.69

INTERPRÉTATIONS

A CE STADE, ON EST CAPABLE D'EXPLIQUER 69.47%% DE LA VARIANCE.

RÉGRESSION LINÉAIRE

MODELE 3

X_TRAIN= GINI + MEAN_INCOME_LOG + C_PARENT

Y_TRAIN = INCOME_LOG

COEFFICIENTS: -1.6712121 0.98831956 0.00991195

RESIDUAL SUM OF SQUARES: 0.43

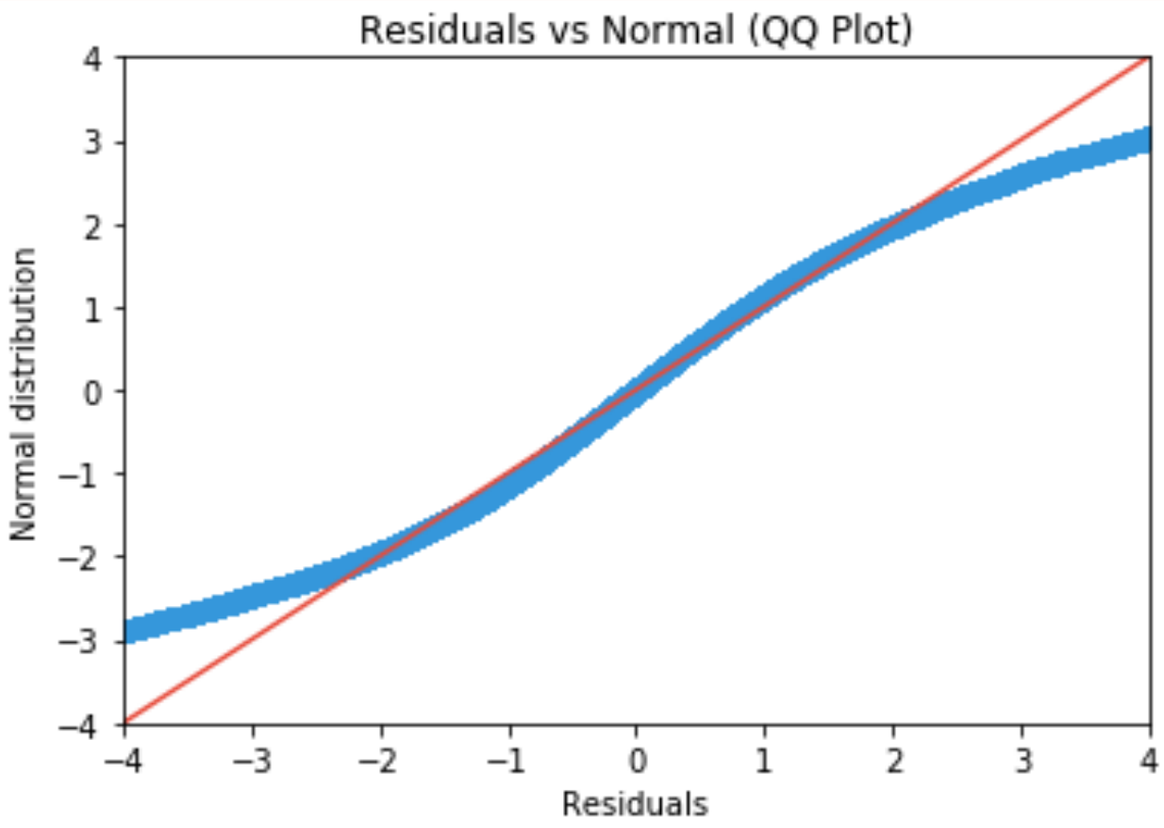
VARIANCE SCORE: 0.74

INTERPRÉTATIONS

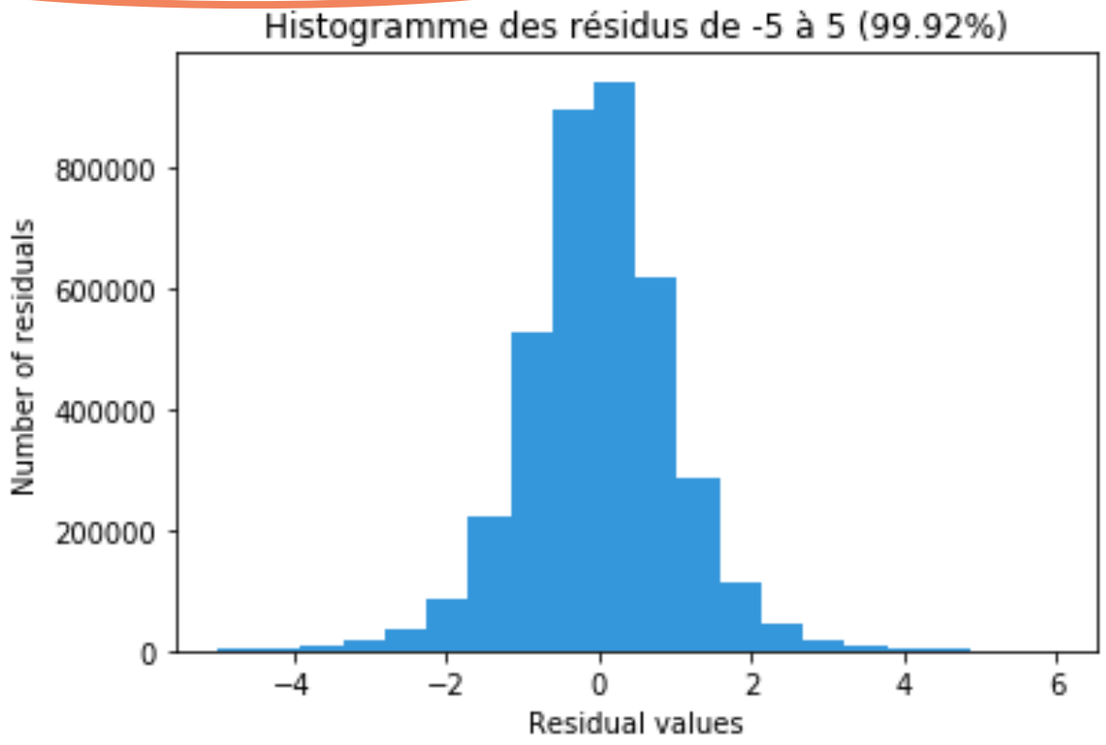
EN INCLUANT LA CLASSE DE REVENU DES PARENTS, ON GAGNE 5 POINTS SUR LE COEFFICIENT DE DÉTERMINATION.

L'INFLUENCE DU REVENU DES PARENTS SUR LE REVENU DE L'ENFANT EST SUBSTANTIEL ET EST MIS EN ÉVIDENCE PAR LE COEFFICIENT D'ÉLASTICITÉ.

RÉGRESSION LINÉAIRE- MODELE 3 RÉSIDUS



Shapiro pvalue : 0.0



RÉGRESSION LINÉAIRE- CONCLUSIONS

DÉSORMAIS, IL RESTE 25% NON EXPLIQUÉ.

LES 25% POURRAIENT INCLURE :

LE NIVEAU D'ÉTUDES

L'ÂGE

LE SEXE

SI ON POUVAIT INCLURE CES FACTEURS LÀ, LE MODÈLE GAGNERAIT CERTAINEMENT GRANDEMENT EN PERFORMANCE.

ENFIN, PLUS L'INDICE DE GINI EST ÉLEVÉ, PLUS LE SALAIRE SERA BAS. CECI EST MIS EN ÉVIDENCE PAR LE COEFFICIENT NÉGATIF AU SEIN DU MODÈLE.

...FIN