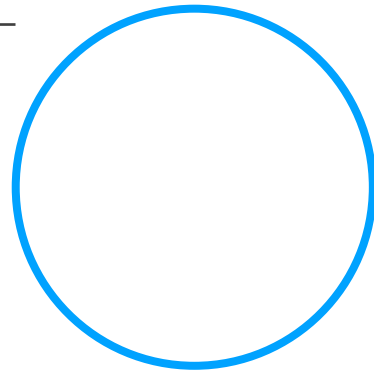


# Prédiction de revenus

---

MAXIMEBCH – DATA ANALYST



# INTRODUCTION

- Notre banque est présente dans de nombreux pays et doit continuer d'acquérir de nouveaux clients
- **Cible** : les jeunes en âge d'ouvrir leur premier compte bancaire et les plus susceptibles d'avoir de hauts revenus dans le futur
- **Méthode** : Création d'un modèle de prédiction basé sur plusieurs variables (revenu des parents, revenu moyen du pays, et indice de Gini)





# Sommaire

- 1 – DESCRIPTION DES DONNÉES
- 2 – ANALYSE DES DONNÉES
- 3 – MODÈLE DE PRÉDICTION



# 1 - DESCRIPTION DES DONNÉES



# 1 – DESCRIPTION DES DONNÉES

## FICHIERS SOURCE



1) Distribution des revenus dans le monde (*World Income Distribution*):

- *country* : le pays (uniquement le code ISO3)
- *year\_survey* : années des données
- *quantile* : la population est divisée en 100 parties égales, des centiles (ici des classes de revenus)
- *income* : revenu du quantile de population
- *gdpppp* : PIB en parité de pouvoir d'achat de la population totale calculé par la Banque Mondiale, permet de comparer en mettant les différents devises au même niveau avec un panier d'achat

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.00000
1	ALB	2008	2	100	916.66235	7297.00000
2	ALB	2008	3	100	1010.91600	7297.00000
3	ALB	2008	4	100	1086.90780	7297.00000
4	ALB	2008	5	100	1132.69970	7297.00000
...	...	...	...	...	...	...



# 1 – DESCRIPTION DES DONNÉES

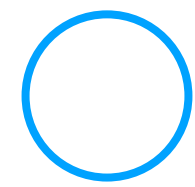
## FICHIERS SOURCE



2) Liste des codes ISO3 (ISO 3166-1)  
des pays :

- Table associant les noms des pays avec leur code ISO3
- Jointure avec le fichier *World Income Distribution* sur le code ISO3

	0	1	2	3	4	5
0	1	4	AF	AFG	Afghanistan	Afghanistan
1	2	8	AL	ALB	Albanie	Albania
2	3	10	AQ	ATA	Antarctique	Antarctica
3	4	12	DZ	DZA	Algérie	Algeria
4	5	16	AS	ASM	Samoa Américaines	American Samoa
...	...	...	...	...	...	...





# 1 – DESCRIPTION DES DONNÉES

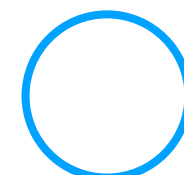
## FICHIERS SOURCE



3) Données de population des pays  
(source : FAO) :

- Contient la population totale des pays
- Jointure avec le dataframe principal

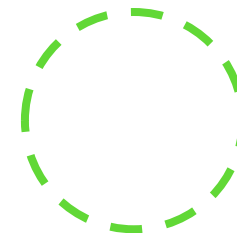
	Country	Année	population 100%
1	Afghanistan	2018	37171921.0
3	Afrique du Sud	2018	57792518.0
5	Albanie	2018	2882740.0
7	Algérie	2018	42228408.0
9	Allemagne	2018	83124418.0
...	...	...	...





# 1 – DESCRIPTION DES DONNÉES

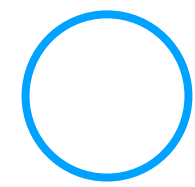
## FICHIERS SOURCE



4) Indice de Gini des pays (source : Banque mondiale) :

- Indicateur rendant compte du niveau d'inégalité pour une variable (ici le revenu)
- 0 = égalité parfaite
- 1 (ou 100%) = inégalité totale

	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965	...
0	Aruba	ABW	Gini index (World Bank estimate)	SI.POV.GINI	NaN	NaN	NaN	NaN	NaN	NaN	...
1	Africa Eastern and Southern	AFE	Gini index (World Bank estimate)	SI.POV.GINI	NaN	NaN	NaN	NaN	NaN	NaN	...
2	Afghanistan	AFG	Gini index (World Bank estimate)	SI.POV.GINI	NaN	NaN	NaN	NaN	NaN	NaN	...

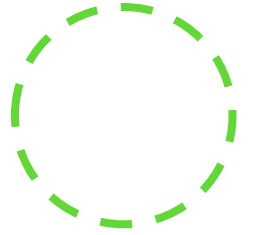






# 1 – DESCRIPTION DES DONNÉES

## DATAFRAME RÉUNISSANT LES FICHIERS SOURCES



1<sup>er</sup> dataframe, créé à partir des fichiers sources :

- 98 pays
- 70% de la population mondiale couverte
- *population\_quantile* : population totale divisée par 100

	iso_code	Country	year_survey	quantile	nb_quantiles	income	gdpppp	Année	population_quantile
0	ALB	Albanie	2008	1	100	728.89795	7297.0	2018	28827.40
1	ALB	Albanie	2008	2	100	916.66235	7297.0	2018	28827.40
2	ALB	Albanie	2008	3	100	1010.91600	7297.0	2018	28827.40
3	ALB	Albanie	2008	4	100	1086.90780	7297.0	2018	28827.40
4	ALB	Albanie	2008	5	100	1132.69970	7297.0	2018	28827.40
...	...	...	...	...	...	...	...	...	...




# 1 – DESCRIPTION DES DONNÉES

## DATAFRAME POUR ANALYSE



2<sup>ème</sup> dataframe, créé à partir du 1<sup>er</sup> et contenant uniquement :

- 5 pays (+ France) : Argentine, Equateur, Estonie, Islande, Russie



	iso_code	Country	year_survey	quantile	nb_quantiles	income	gdpppp	Année	population_quantile
349	ARG	Argentine	2008	50	100	4132.6655	13220.0	2018	443611.50
326	ARG	Argentine	2008	27	100	2429.5596	13220.0	2018	443611.50
327	ARG	Argentine	2008	28	100	2497.3577	13220.0	2018	443611.50
328	ARG	Argentine	2008	29	100	2566.5305	13220.0	2018	443611.50
329	ARG	Argentine	2008	30	100	2637.7153	13220.0	2018	443611.50



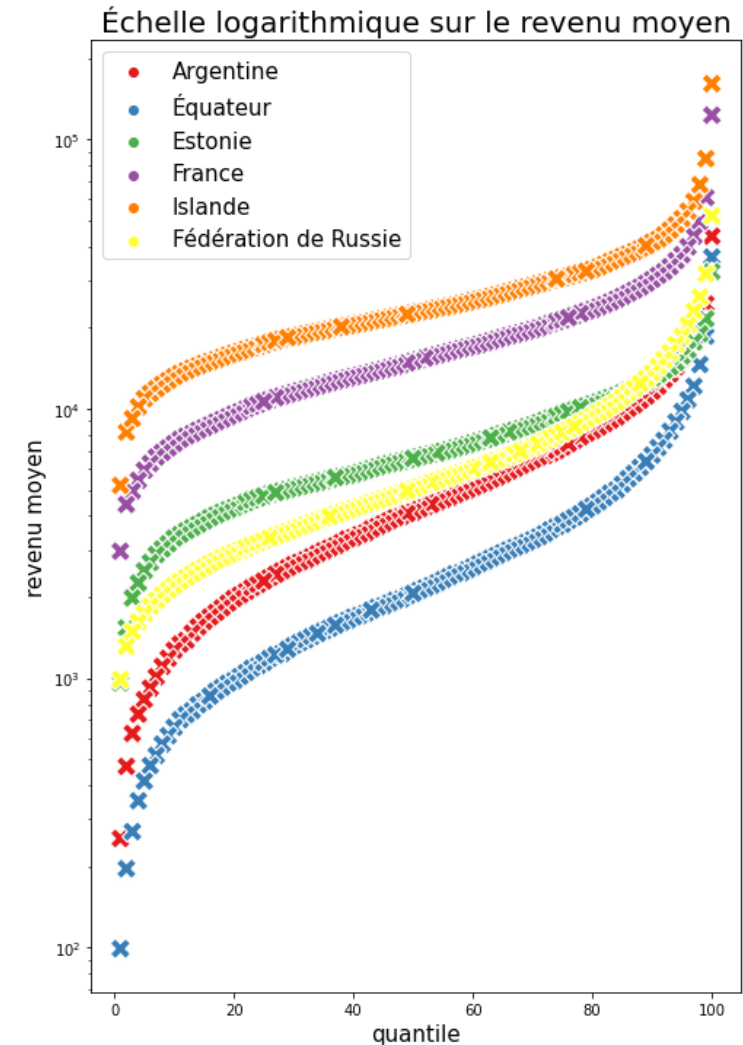
# 2 – ANALYSE DES DONNÉES

## 2 – ANALYSE DES DONNÉES

### DISTRIBUTION DES RICHESSES

Transposition du revenu moyen des quantiles sur une échelle logarithmique :

- Application de *math.log10()* (logarithme décimal / de base 10) sur les revenus pour créer des classes
- Permet de représenter une large gamme de valeurs sur un petit espace

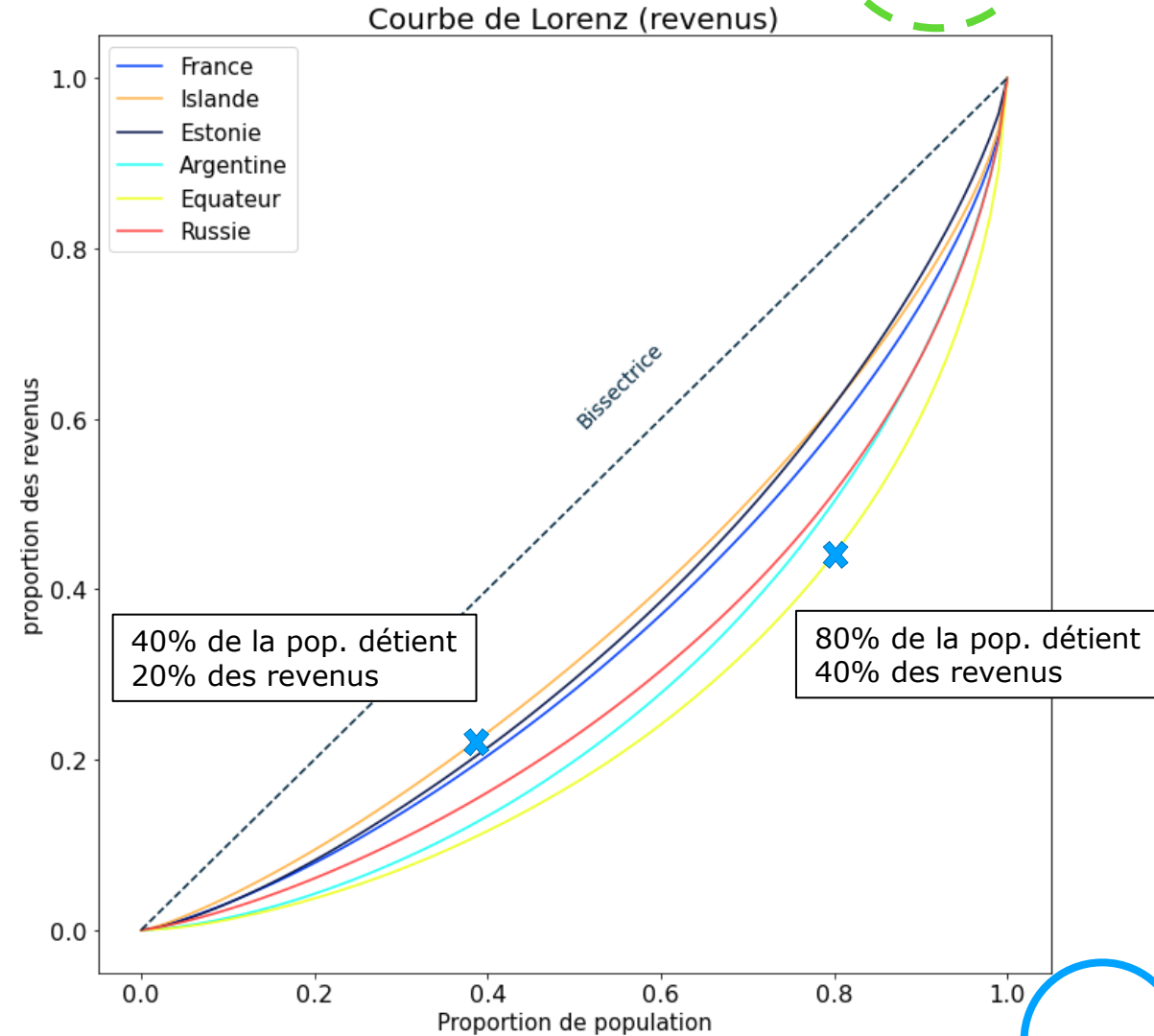


## 2 – ANALYSE DES DONNÉES

### DISTRIBUTION DES RICHESSES

La courbe de Lorenz :

- Créée pour représenter les inégalités de revenus (mais transposable à d'autres variables)
- Elle permet de calculer le coefficient de Gini

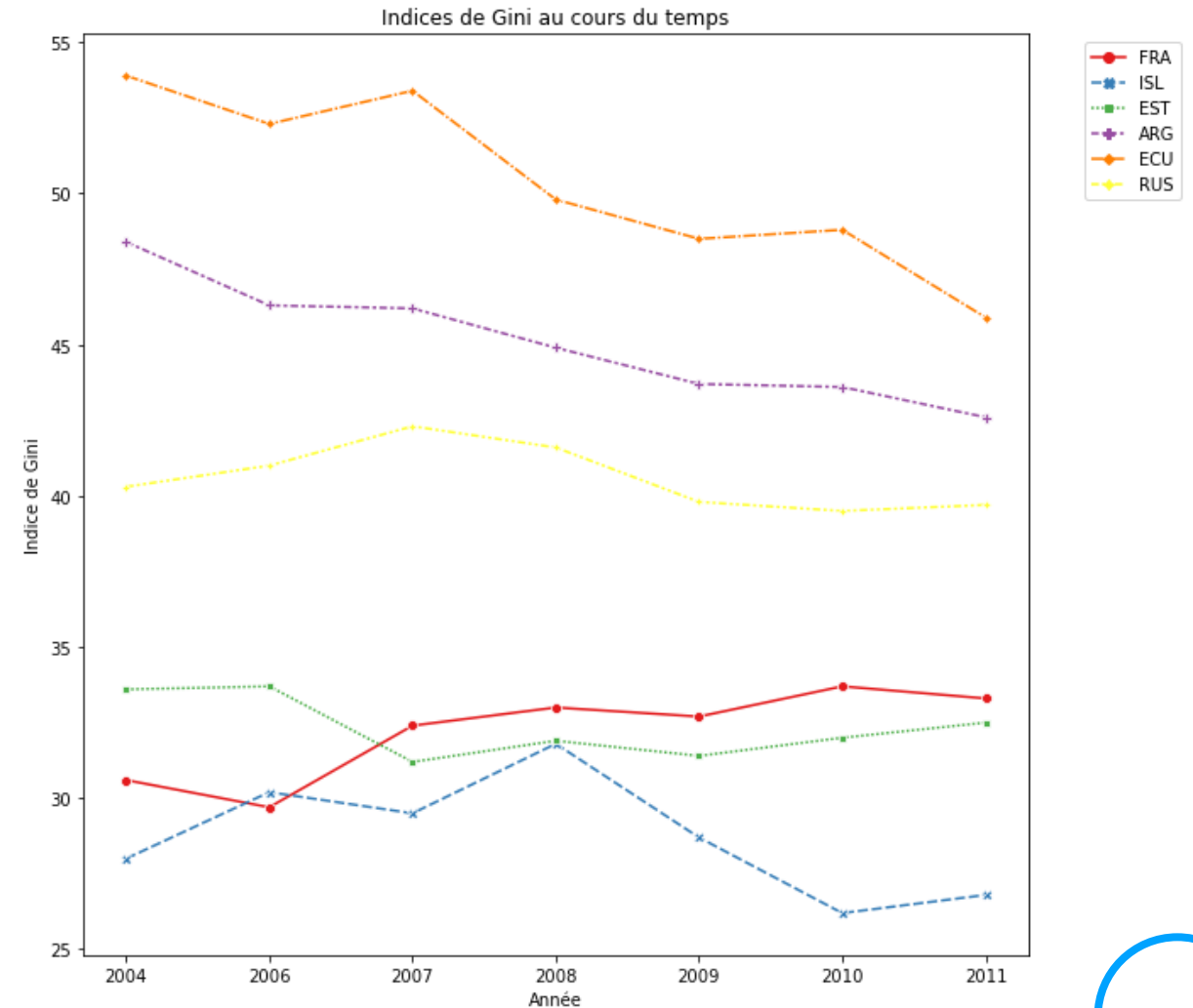


## 2 – ANALYSE DES DONNÉES

### DISTRIBUTION DES RICHESSES

Indice de Gini, en 2011, le classement du plus élevé au moins élevé était :

- Equateur
- Argentine
- Russie
- France
- Estonie
- Islande (après la crise de 2008 : a augmenté les salaires et les prestations sociales, a laissé les banques faire faillite, a donné le pouvoir au gouvernement d'intervenir sur les marchés financiers)



## 2 – ANALYSE DES DONNÉES

### SIMULATION DE LA CLASSE DE REVENU DES PARENTS

À partir du coefficient d'élasticité (source : Banque Mondiale), qui mesure la mobilité intergénérationnelle du revenu.

	iso3	Country	quantile	nb_quantiles	income	gdpppp	gini	population_quantile	IGEincome
0	ECU	Équateur	1	100	99.078545	7560.0	49.8	170843.58	0.525991
1	ECU	Équateur	2	100	196.350430	7560.0	49.8	170843.58	0.525991
2	ECU	Équateur	3	100	269.607300	7560.0	49.8	170843.58	0.525991
3	ECU	Équateur	4	100	350.863100	7560.0	49.8	170843.58	0.525991
4	ECU	Équateur	5	100	415.143740	7560.0	49.8	170843.58	0.525991



## 2 – ANALYSE DES DONNÉES

### SIMULATION DE LA CLASSE DE REVENU DES PARENTS



Génération aléatoire de la **classe de revenu des parents** à partir du coefficient d'élasticité (*IGEincome*) et de la classe de revenu de l'enfant (probabilité conditionnelle) :

- Echantillon d'individu  $n$  supérieur à 1000 fois le nombre de quantile
- Calcul du revenu de l'enfant ( $y_{child}$ ) à partir du revenu des parents ( $y_{parents}$ ) pour un coeff. d'élasticité donné
- Calcul de la classe de revenu des enfants ( $c_{i\_child}$ ) et parents ( $c_{i\_parent}$ )

	$y_{child}$	$y_{parents}$	$c_{i\_child}$	$c_{i\_parent}$
0	0.590742	2.896064	4	9
1	1.369858	0.706701	6	4
2	7.030097	1.372981	10	7
3	0.406454	0.088070	3	1
4	0.248834	0.627421	2	4




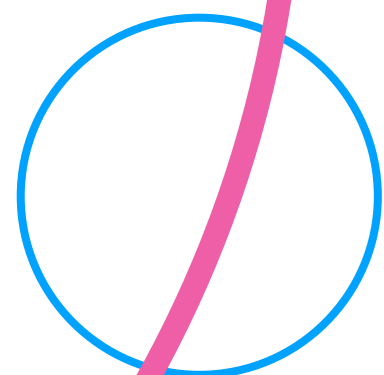

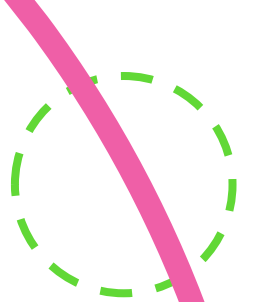
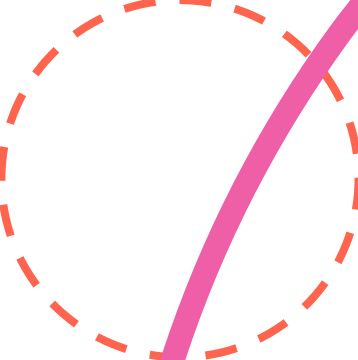
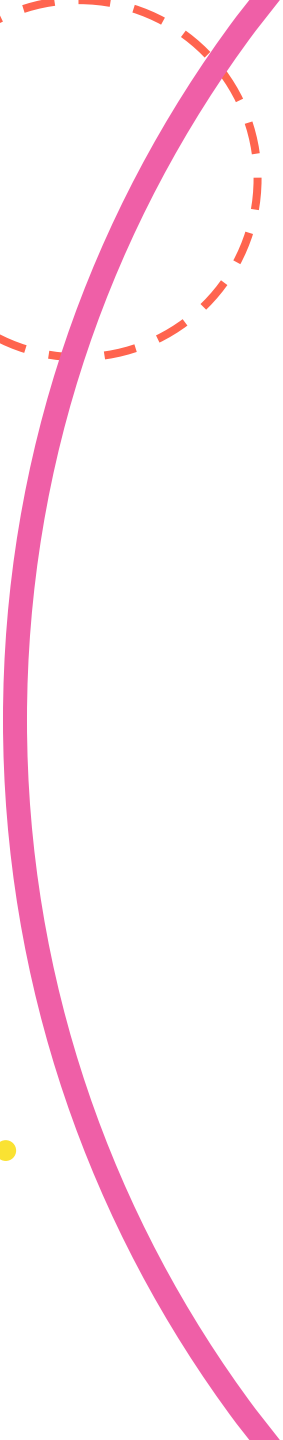
## 2 – ANALYSE DES DONNÉES

### GÉNÉRATION D'UN NOUVEL ÉCHANTILLON

Création d'un nouvel échantillon :

- En copiant 499 fois les individus du fichier *World Income Distribution*
- Attribution des classes parents aux 500 individus

	ln_y_parent	residus	country	income_child	coeff_elasticite	c_i_parent	Gj	y_child	y_parents
0	-0.417670	0.261167	ECU	99.078545	0.525991	34	49.8	1.042348	0.658580
1	0.778975	-0.528784	ECU	196.350430	0.525991	79	49.8	0.887763	2.179238
2	0.030515	-0.336599	ECU	269.607300	0.525991	52	49.8	0.725751	1.030985
3	-0.891634	0.139961	ECU	350.863100	0.525991	19	49.8	0.719621	0.409985
4	-0.452565	-0.125926	ECU	415.143740	0.525991	33	49.8	0.694911	0.635995



# 3 – MODÈLE DE PRÉDICTION

# 3 – MODÈLE DE PRÉDICTION

## ANOVA

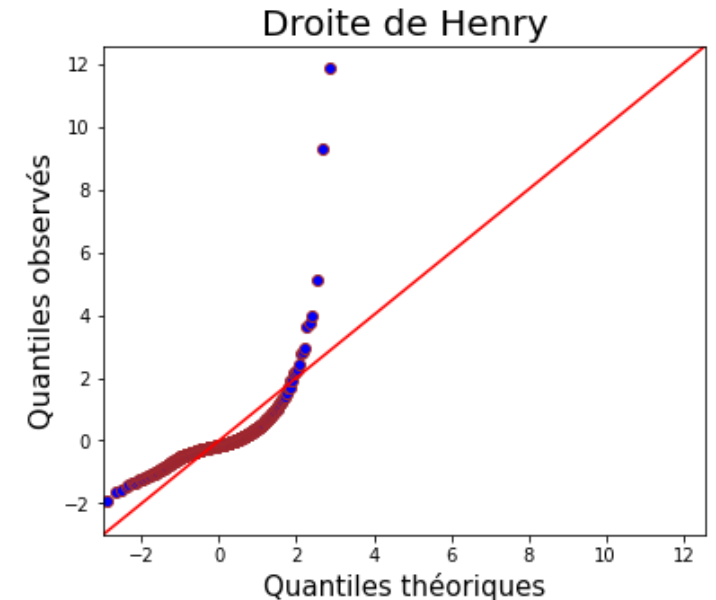
Après réalisation d'une ANOVA, la p-value indique que le pays d'origine influencerait donc les revenus.

Deux conditions à valider pour confirmer l'ANOVA :

- Normalité des résidus :
  - Droite de Henry (à droite) : ne semble pas normal
  - Test de Kolmogorov-Smirnov : - de 5% donc **ne suit pas une loi normale**
- Égalité des variances :
  - Test de Breusch-Pagan : + de 5% donc les variances **sont probablement hétérogènes**

L'ANOVA n'ayant pas été validée, on réalise un test non-paramétrique :

- Test de **Kruskal-Wallis** : la p-value est inférieure à 5% et confirme les résultats de l'ANOVA, le pays d'origine influe sur le revenus

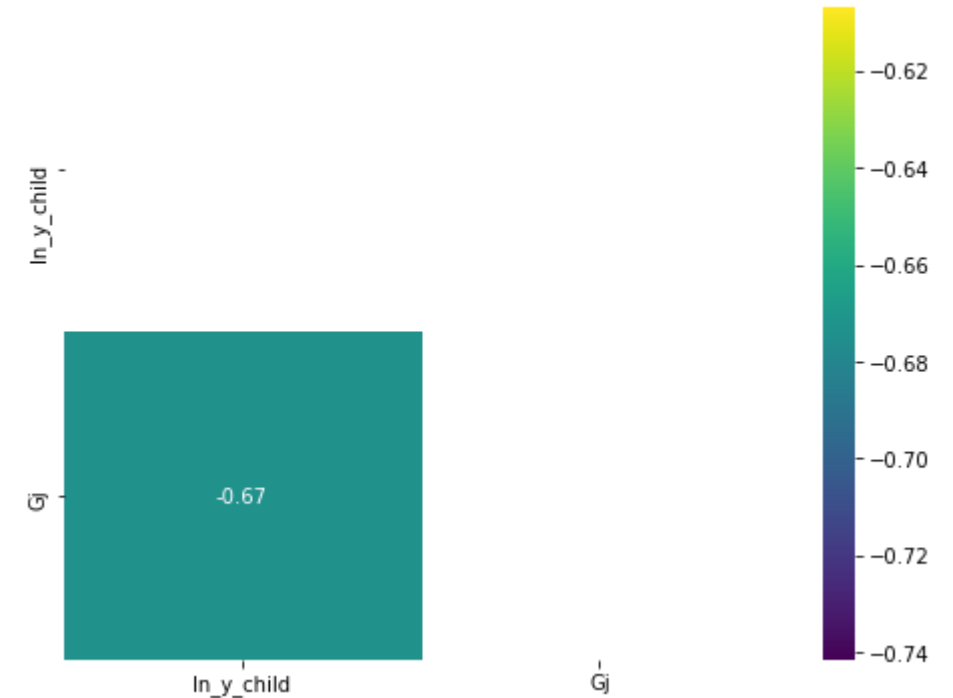


# 3 – MODÈLE DE PRÉDICTION

## CORRÉLATION ENTRE INDICE DE GINI ET REVENU DE L'ENFANT

On observe une corrélation négative entre l'indice de Gini et le revenu moyen de l'enfant : plus l'indice de Gini est bas, plus le revenu a tendance à être haut.

Corrélation entre l'indice de Gini et le revenu

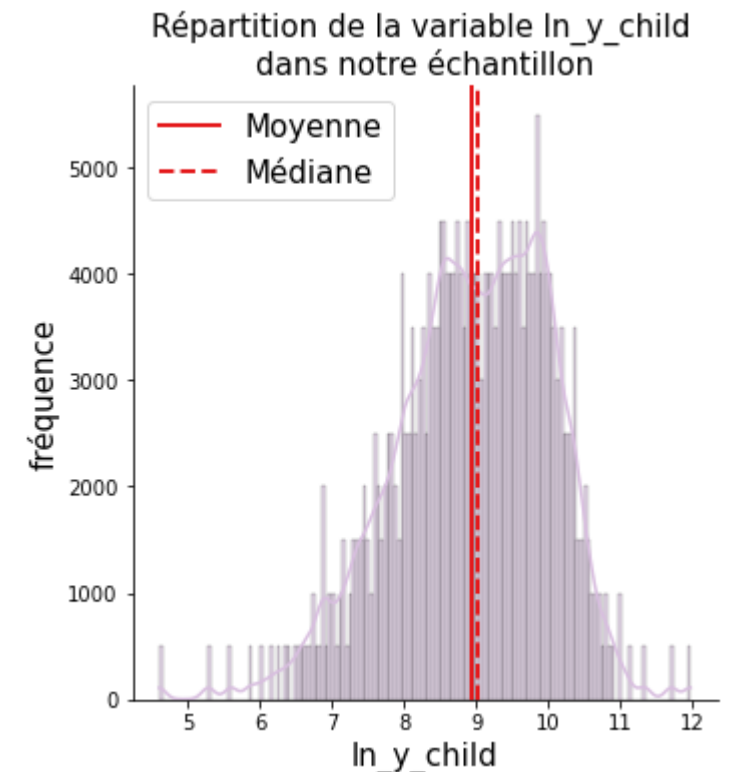
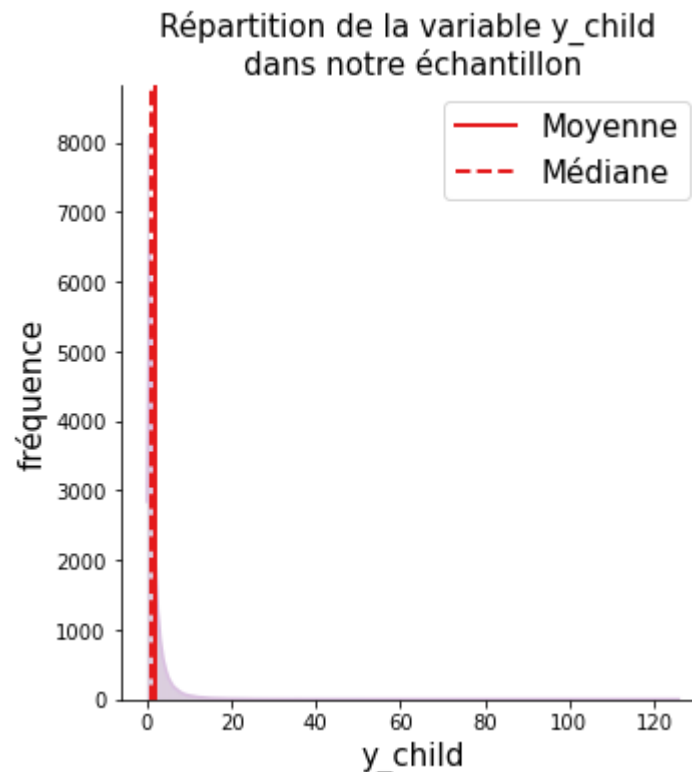


# 3 – MODÈLE DE PRÉDICTION

## PASSAGE AU LOGARITHME

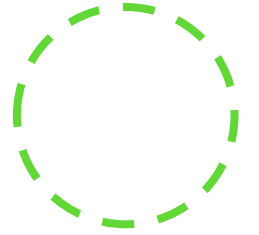
Passage à l'échelle logarithmique qui présente plusieurs intérêts :

- Normalise, lisse la distribution
- Représente des nombres aux ordres de grandeur différents sur un même graphique
- Données centrées et réduites
- Réduit les outliers (réduit la marginalité des quantiles)
- Réduit l'asymétrie positive



# 3 – MODÈLE DE PRÉDICTION

## RÉGRESSIONS LINÉAIRES



Plusieurs régressions linéaires (RL) sont réalisées pour expliquer le revenu de l'enfant.

RL1 = Revenu moyen du pays + Indice de Gini

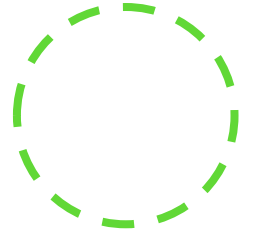
Indicateurs :

- Test de Kolmogorov-Smirnov ( $<0.05$ ) : ne suit pas une loi normale
- Test de Breush-Pagan ( $p < 0.05$ ) : homogénéité des variables

Dep. Variable:	ln_y_child	R-squared:	0.588			
Model:	OLS	Adj. R-squared:	0.588			
Method:	Least Squares	F-statistic:	1.784e+05			
Date:	Fri, 11 Feb 2022	Prob (F-statistic):	0.00			
Time:	16:37:46	Log-Likelihood:	-2.6450e+05			
No. Observations:	250000	AIC:	5.290e+05			
Df Residuals:	249997	BIC:	5.290e+05			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	9.9449	0.013	789.570	0.000	9.920	9.970
mj	6.421e-05	2.25e-07	284.813	0.000	6.38e-05	6.46e-05
Gj	-4.8229	0.028	-175.353	0.000	-4.877	-4.769
Omnibus:	10796.364	Durbin-Watson:	0.381			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36524.433			
Skew:	0.047	Prob(JB):	0.00			
Kurtosis:	4.870	Cond. No.	3.32e+05			

# 3 – MODÈLE DE PRÉDICTION

## RÉGRESSIONS LINÉAIRES



RL2 = Revenu moyen du pays (au logarithme) + Indice de Gini

Indicateurs :

- $R^2$  : renforcé par le passage au logarithme
- Test de Kolmogorov-Smirnov ( $<0.05$ ) : ne suit probablement pas une loi normale, mais plus probable après le passage au logarithme
- Test de Breush-Pagan ( $<0$ ) : homogénéité des variables

Dep. Variable:	ln_y_child	R-squared:	0.601			
Model:	OLS	Adj. R-squared:	0.601			
Method:	Least Squares	F-statistic:	1.880e+05			
Date:	Fri, 11 Feb 2022	Prob (F-statistic):	0.00			
Time:	16:37:48	Log-Likelihood:	-2.6061e+05			
No. Observations:	250000	AIC:	5.212e+05			
Df Residuals:	249997	BIC:	5.213e+05			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.3085	0.042	7.342	0.000	0.226	0.391
ln_mj	1.0081	0.003	302.604	0.000	1.002	1.015
Gj	-1.6778	0.034	-48.694	0.000	-1.745	-1.610
Omnibus:	10716.816	Durbin-Watson:	0.393			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35194.671			
Skew:	0.081	Prob(JB):	0.00			
Kurtosis:	4.831	Cond. No.	360.			

# 3 – MODÈLE DE PRÉDICTION

## RÉGRESSIONS LINÉAIRES

RL3 = Revenu moyen du pays (au logarithme) + Indice de Gini + classe de revenu des parents (*variable probablement pas significative*)

Indicateurs :

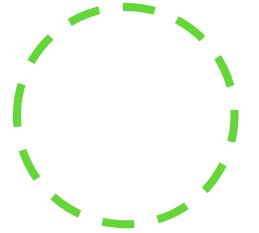
- $R^2$  : renforcé par le passage au logarithme
- Test de Kolmogorov-Smirnov ( $p < 0.05$ ) : ne suit probablement pas une loi normale, mais plus probable après le passage au logarithme
- Test de Breush-Pagan ( $p < 0.05$ ) : homogénéité des variables

Dep. Variable:	ln_y_child	R-squared:	0.601			
Model:	OLS	Adj. R-squared:	0.601			
Method:	Least Squares	F-statistic:	1.254e+05			
Date:	Fri, 11 Feb 2022	Prob (F-statistic):	0.00			
Time:	16:37:49	Log-Likelihood:	-2.6061e+05			
No. Observations:	250000	AIC:	5.212e+05			
Df Residuals:	249996	BIC:	5.213e+05			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.3090	0.042	7.344	0.000	0.227	0.391
ln_mj	1.0081	0.003	302.602	0.000	1.002	1.015
Gj	-1.6778	0.034	-48.693	0.000	-1.745	-1.610
c_i_parent	-1.144e-05	4.75e-05	-0.241	0.810	-0.000	8.18e-05
Omnibus:	10716.681	Durbin-Watson:	0.393			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35193.951			
Skew:	0.081	Prob(JB):	0.00			
Kurtosis:	4.831	Cond. No.	2.28e+03			



# 3 – MODÈLE DE PRÉDICTION

## RÉGRESSIONS LINÉAIRES



RL4 = Revenu moyen du pays (au logarithme) + Indice de Gini + classe de revenu des parents + revenu des parents (au logarithme) + coefficient d'élasticité

Indicateurs :

- **$R^2$  : Ce nouveau modèle explique environ 60,1% de la variance de la variable "ln\_y\_child" (39% restants sont expliqués par d'autres facteurs comme les études, genre, chance, efforts...)**
- Test de Kolmogorov-Smirnov ( $p < 0.05$ ) : ne suit probablement pas une loi normale
- Test de Breush-Pagan ( $p < 0.05$ ) : homogénéité des variables

Dep. Variable:	ln_y_child	R-squared:	0.601
Model:	OLS	Adj. R-squared:	0.601
Method:	Least Squares	F-statistic:	7.526e+04
Date:	Fri, 11 Feb 2022	Prob (F-statistic):	0.00
Time:	16:37:51	Log-Likelihood:	-2.6057e+05
No. Observations:	250000	AIC:	5.211e+05
Df Residuals:	249994	BIC:	5.212e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.4383	0.045	9.725	0.000	0.350	0.527
ln_mj	1.0054	0.003	300.839	0.000	0.999	1.012
Gj	-1.4376	0.042	-34.119	0.000	-1.520	-1.355
c_i_parent	-0.0002	0.000	-1.028	0.304	-0.001	0.000
ln_y_parent	0.0065	0.006	0.996	0.319	-0.006	0.019
coeff_elasticite	-0.4370	0.044	-9.894	0.000	-0.524	-0.350

Omnibus:	10639.801	Durbin-Watson:	0.393
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34576.550
Skew:	0.086	Prob(JB):	0.00
Kurtosis:	4.814	Cond. No.	2.39e+03

