

## Projet 7 - Prédictions de Revenus



# Introduction

## Ce projet a été réalisé :

- Dans le cadre de la formation DATA ANALYST d'OpenClassRooms
- Avec l'aide/support de **M. César Clavé** (Mentor Openclassrooms)
- **Contexte** → Banque Internationale (*NEURONAL PRIVATE BANKING*)
- **Mission** → Cibler de nouveaux clients potentiels, susceptibles d'avoir de hauts revenus.
- **Proposition** → Conception d'un Modèle/Programme de prédiction de revenu d'un individu en utilisant les paramètres suivants:
  - *revenu moyen du pays d'origine* (Source OpenClassrooms)
  - *indice de gini* (Source World Bank Open Data)
  - *Coefficient de mobilité intergénérationnelle* (Fourni OpenClassrooms)
  - *Classe(tranche) de revenus de ses parents* (Estimée par étude)

## Sources



# Sommaire

## Introduction (remerciements)

## Data Import - Analyse descriptive

- *World Income Distribution (dataset fourni)*
- *Indices de Gini (World bank open Data)*
- *Coefficient d'élasticité (World bank open Data)*
- *Synthèse chiffrée, justification périmètre*

## Etude des revenus par pays (égalités/inégalités)

- *Etudes des Indices de Gini*
- *Courbes de Lorenz*
- *Visualisations sur 6 pays*

## Algorithme de Détermination Classe de revenu parent

- *Etude du coefficient d'élasticité (mobilité intergénérationnelle)*
- *Création programme d'estimation du revenu en fonction du coeff.*
- *Constitution Dataset final consolidé avec les 3 variables indicateurs  
Indice Gini - Revenu moyen classe - Classe de revenu parent*

## Anova & Régression Logistique - Modèle Prédictif – Etudes Résidus

- Application d'une ANOVA sur la Dataset consolidé
- Application de modèles de régressions linéaires multiple sur Dataset \* 500
- Meilleur modèle (3variables avec logarithme)
- Etude performance de ce modèle.

## Conclusion / Questions

## Annexes



# Data Import - Analyse descriptive

1<sup>er</sup> Contrôle visuel (rapide)  
(re)Mise en forme de données



Nettoyage données (en Python)  
«Cleaning» détaillé dans un notebook jupyter



## DATASET Principal → WID (world income distribution)

Constat Année 2008 : Nombre de quantiles n'est pas un multiple de 100 sur l'année 2008.  
Recherche des pays ayant un nombre de quantiles annuel <> 100  
De plus certains données manquantes en terme de gdppp sur 2008 et 2009

year	Nb_Quantiles	Nb_Pays	count(income)/100	count(gdppp)/100
2004	100	1	1	1
2006	500	5	5	5
2007	1500	15	15	15
2008	7599	75	75	74
2009	1200	12	12	11
2010	600	6	6	6
2011	100	1	1	1

Il manque la classe de revenu du **quantile 41** pour l'année 2008 et pour la Lituanie.

Création d'une ligne pour ce quantile manquant en prenant la valeur intermédiaire entre les quantiles 40 & 42.

Avant					Après						
	country	year	quantile	income	gdppp		country	year	quantile	income	gdppp
6237	LTU	2008	38	4,756.434	17,571.000	6238	LTU	2008	39	4,802.368	17,571.000
6238	LTU	2008	39	4,802.368	17,571.000	6239	LTU	2008	40	4,868.451	17,571.000
6239	LTU	2008	40	4,868.451	17,571.000	6240	LTU	2008	41	4,882.141	17,571.000
6240	LTU	2008	42	4,895.831	17,571.000	6241	LTU	2008	42	4,895.831	17,571.000

# Data Import - Analyse descriptive

Manquent aussi les données gdppp pour la Palestine (PSE) et le Kosovo (XK).

Imputation des valeurs d'après les sources :

- (PSE) gdppp Palestine = 3612.14
- (XK) gdppp Kosovo = 7236.41

Valeur énorme en gdppp pour les FIDJI

Modification de la valeur d'après la source :

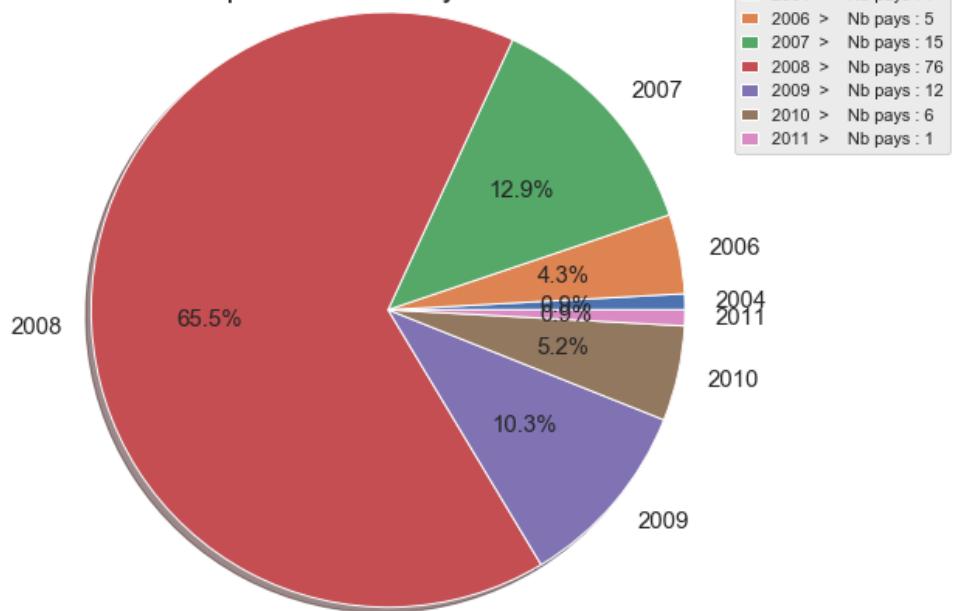
- (XK) gdppp Fidji = 7078.62

country	year	quantile	income	gdppp
3200	FJI	2008	1	308.173 4,300,332.000
3201	FJI	2008	2	384.332 4,300,332.000
3202	FJI	2008	3	436.593 4,300,332.000
3203	FJI	2008	4	486.814 4,300,332.000
3204	FJI	2008	5	520.197 4,300,332.000

DATASET Principal → WID (world income distribution)

116 Pays

Repartition des Pays / Année



2004 > Nb pays : 1
2006 > Nb pays : 5
2007 > Nb pays : 15
2008 > Nb pays : 76
2009 > Nb pays : 12
2010 > Nb pays : 6
2011 > Nb pays : 1

country_code	year	quantile	income	gdppp
		ALB	2008	1 728.898 7,297.000
	2008	2	916.662	7,297.000
	2008	3	1,010.916	7,297.000
	2008	4	1,086.908	7,297.000
	2008	5	1,132.700	7,297.000

	quantile	income	gdppp
count	11,600.000	11,600.000	11,600.000
mean	50.500	6,069.122	12,435.296
std	28.867	9,413.787	13,097.663
min	1.000	16.719	303.193
25%	25.750	900.769	2,577.500
50%	50.500	2,403.493	7,488.500
75%	75.250	7,515.314	17,679.250
max	100.000	176,928.550	73,127.000

# Data Import - Analyse descriptive

## DATASET → *Gini* (Coefficients de Gini issus de la banque mondiale)

### Origine

		country_name	incomegroup	gini
country_code	year			
<b>AGO</b>	2000	Angola	Lower middle	52.000
	2008	Angola	Lower middle	42.700
<b>ALB</b>	1996	Albania	Upper middle	27.000
	2002	Albania	Upper middle	31.700
<b>2005</b>	Albania	Upper middle	30.600	
	2008	Albania	Upper middle	30.000
<b>2012</b>	Albania	Upper middle	29.000	
country_code	country_name	incomegroup	year	gini
count	1486	1486	1486	1,486.000 1,486.000
unique	164	164	4	nan nan
top	BRA	Brazil	Upper middle	nan nan
freq	33	33	551	nan nan
mean	NaN	NaN	NaN 2,005.193	39.157
std	NaN	NaN	NaN 8.089	9.423
min	NaN	NaN	NaN 1,979.000	21.000
25%	NaN	NaN	NaN 2,001.000	31.800
50%	NaN	NaN	NaN 2,007.000	37.100
75%	NaN	NaN	NaN 2,011.000	46.075
max	NaN	NaN	NaN 2,017.000	65.800

Constitution d'un dataframe *gini\_evo* de l'évolution des indices de Gini.

Sur la période de notre périmètre (2004 à 2011)

- Filtre sur données existantes dans "wid" : Clef --> **country\_code**
- Pivoter les données annuelles en colonnes

### Constitution d'un dataframe *gini\_moy*

Moyenne des indices sur la période de notre périmètre (2004 à 2011)

<b><i>gini_moy</i></b>		
country_code	country_name	gini
AGO	Angola	42.700
ALB	Albania	30.300
ARG	Argentina	45.325
ARM	Armenia	31.375
AUS	Australia	34.400
AUT	Austria	30.213
AZE	Azerbaijan	26.600
BDI	Burundi	33.400
BEL	Belgium	28.812
BEN	Benin	43.400

<b><i>gini_evo</i></b>								
country_code	gini04	gini05	gini06	gini07	gini08	gini09	gini10	gini11
ALB	nan	30.600	nan	nan	30.000	nan	nan	nan
ARG	48.300	47.700	46.600	46.300	44.500	43.900	43.000	42.300
ARM	37.500	36.000	29.700	31.200	29.200	28.000	30.000	29.400
AUT	29.800	28.700	29.600	30.600	30.400	31.500	30.300	30.800
AZE	26.600	26.600	nan	nan	nan	nan	nan	nan
BEL	30.500	29.300	28.100	29.200	28.400	28.500	28.400	28.100
BFA	nan	nan	nan	nan	39.800	nan	nan	nan

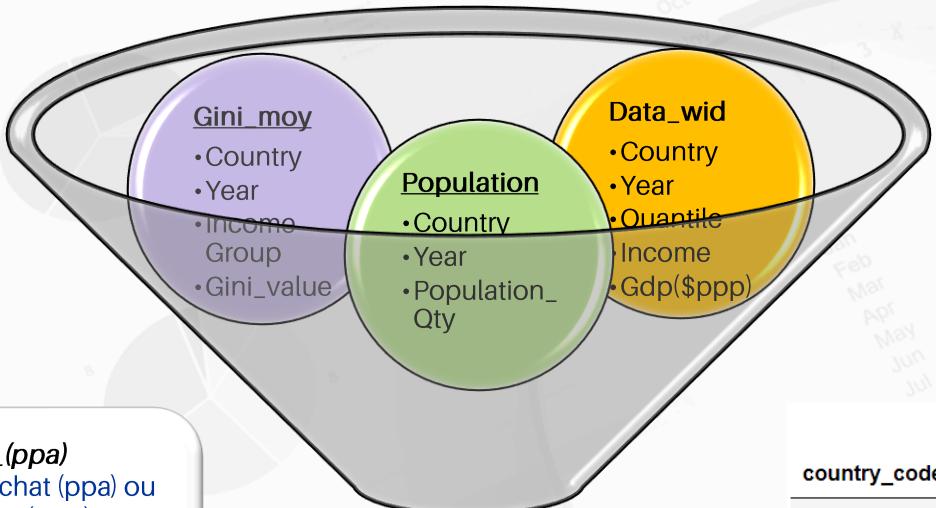
114 Pays de wid trouvés

Taiwan (TWN) et le Cambodge (KHM) n'ont pas de données (gini) entre 2004 & 2011.



# Data Import - Analyse descriptive

**DATASET → gen** (*fusion de wid, gini\_moy & pop*)



**GDP\_(ppp) ou PIB\_(ppa)**

Parité de pouvoir d'achat (ppa) ou Power Purchase Parity (ppp)

Méthode utilisée en économie.  
Permet de corriger l'effet de la devise sur la valeur d'un panier de biens "moyen" et donc l'effet des taxes et des réglementations locales.

```
<class 'pandas.core.frame.DataFrame'>
Index: 11400 entries, ALB to COD
Data columns (total 6 columns):
country_name    11400 non-null object
pop             11400 non-null float64
quantile        11400 non-null int64
income          11400 non-null float64
gini            11400 non-null float64
gdpppp         11400 non-null float64
dtypes: float64(4), int64(1), object(1)
memory usage: 623.4+ KB
```

	country_name	pop	quantile	income	gini	gdpppp
country_code						
<b>ALB</b>	Albania	2,947,314.000	1	728.898	30.300	7,297.000
<b>ALB</b>	Albania	2,947,314.000	2	916.662	30.300	7,297.000
<b>ALB</b>	Albania	2,947,314.000	3	1,010.916	30.300	7,297.000
<b>ALB</b>	Albania	2,947,314.000	4	1,086.908	30.300	7,297.000
<b>ALB</b>	Albania	2,947,314.000	5	1,132.700	30.300	7,297.000

**Années des données utilisées**

2004 - 2006 - 2007 - 2008 - 2009 - 2010 - 2011

Approximation → Année considérée 2008

Suppression de la colonne année



# Data Import - Analyse descriptive

## DATASET → gdim

Il permet de déduire la probabilité (%) d'un individu appartenant à une classe donnée (parents) d'évoluer, de changer de classe.

### Données "mobilité intergénérationnelle" de la Banque Mondiale :

- Période de couverture des Naissances : 1940-1989
- Années d'étude : 1991-2016
- Pourcentage de la population couverte : 96(%)

### Approximation

- 1ère Etude (income, ind\_gini) :

Population couverte par l'analyse (114 pays) = 91.26(%)

- 2ème Etude (coeff.elasticité) :

Population couverte par l'analyse (75 pays) = 76.93(%)

Utilisation de la moyenne du coefficient de mobilité intergénérationnelle de la région du pays pour combler les données manquantes de ce pays.

region	IGEincome
East Asia & Pacific	0.504
Europe & Central Asia	0.466
High income	0.346
Latin America & Caribbean	0.897
Middle East & North Africa	0.817
South Asia	0.505
Sub-Saharan Africa	0.665



150 Pays, mais seulement 75 avec une valeur de gdim

country_code	country_name	region	IGEincome
AFG	Afghanistan	South Asia	nan
AGO	Angola	Sub-Saharan Africa	nan
ALB	Albania	Europe & Central Asia	0.816
ARG	Argentina	Latin America & Caribbean	nan
ARM	Armenia	Europe & Central Asia	nan

2 Pays exclus de l'étude  
le Kosovo(KSV) et la Syrie(SYR) n'ont pas de données « gdim »

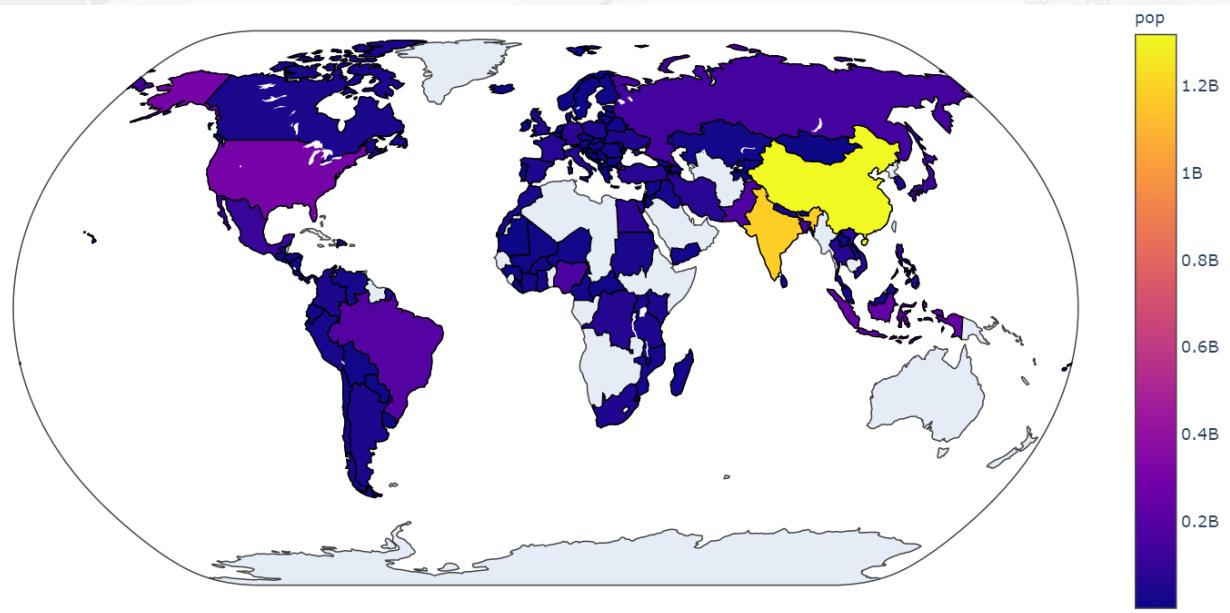
country_code	country_name	region	IGEincome
AFG	Afghanistan	South Asia	0.505
AGO	Angola	Sub-Saharan Africa	0.665
ALB	Albania	Europe & Central Asia	0.816
ARG	Argentina	Latin America & Caribbean	0.897
ARM	Armenia	Europe & Central Asia	0.466
AUS	Australia	High income	0.275
AUT	Austria	High income	0.245
AZE	Azerbaijan	Europe & Central Asia	0.466
BEL	Belgium	High income	0.183
BEN	Benin	Sub-Saharan Africa	0.855

## Synthèse / Périmètre Etude

112 pays

Taux Couverture Etude → 90,93% de la population mondiale

Année 2008 : Ratio entre la  $\Sigma$  population des 112 pays / Population mondiale 2008



country_code	country_name	pop	c_i_child	income	gini	gdpppp	pj
ALB	Albania	2,947,314.000	1	728.898	30.300	7,297.000	0.816
ALB	Albania	2,947,314.000	2	916.662	30.300	7,297.000	0.816
ALB	Albania	2,947,314.000	3	1,010.916	30.300	7,297.000	0.816
ALB	Albania	2,947,314.000	4	1,086.908	30.300	7,297.000	0.816
ALB	Albania	2,947,314.000	5	1,132.700	30.300	7,297.000	0.816

Ici la colonne **c\_i\_child** est découpée en *centiles*. Oui, car permet de bien voir la répartition par tranches, de segmenter et de classer rapidement une valeur. De plus le fait d'avoir des centiles induit rapidement une notion de (%) pourcentage très pratique (étude gini par ex.)

# Etude des revenus / pays (Indices de gini)

**Gini** (prend ses valeurs entre [0 et 1])

Le coefficient de Gini, ou indice de Gini, est une mesure statistique permettant de rendre compte de la répartition d'une variable (salaire, revenus, patrimoine) au sein d'une population.

L'inégalité est d'autant plus forte que l'indice de Gini est élevé.

Calcul → voir « courbe de Lorenz » ci-après

## Graphique Diversité Pays (revenus)

6 pays dont la FRANCE (imposée).

### Filtres :

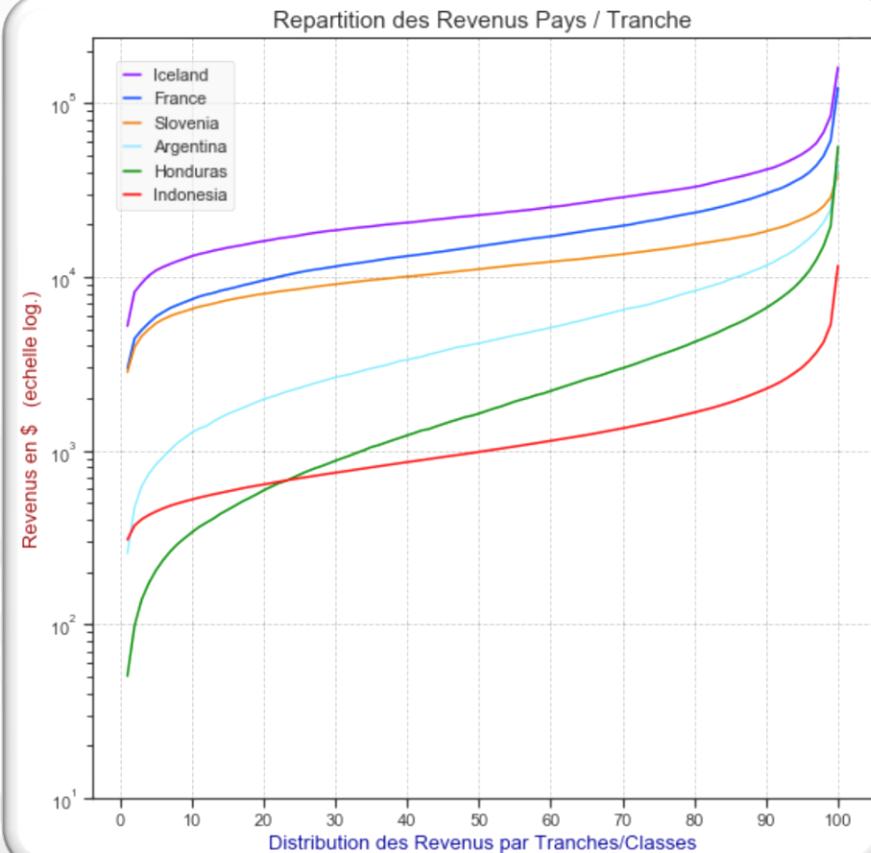
- pays ayant une évolution complète des ginis de 2004 à 2011 (45 pays)
- pays *mini* et *maxi* en "income" et "gini\_moyen".
- Argentine, pays intermédiaire.

En abscisse, la classe de revenu (centiles).

En ordonnée, le revenu moyen cumulé (en log.)

		income	gini
country_code	country_name		
ARG	Argentina	5,847.885	45.325
FRA	France	18,309.408	31.900
HND	Honduras	3,296.268	55.875
IDN	Indonesia	1,334.618	35.238
ISL	Iceland	26,888.512	28.775
SVN	Slovenia	12,106.007	24.563

(Annexe)



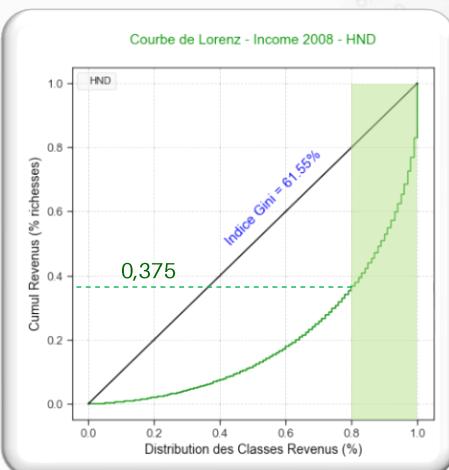
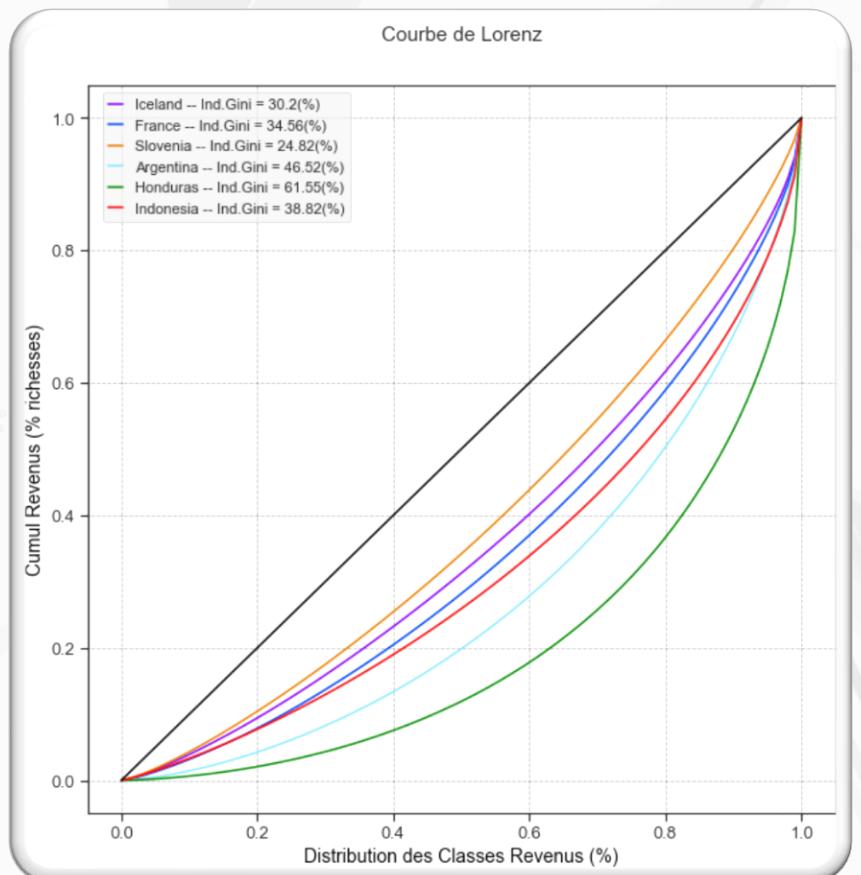
# Etude des revenus / pays (Indices de gini)

## Courbe de Lorenz

Une courbe de Lorenz illustre la répartition d'une donnée dans un population de cette donnée. Plus cette courbe est éloignée de la bissectrice, plus les inégalités sont fortes.

Permet d'en déduire le calcul du coefficient de Gini à partir de l'aire sous la courbe.

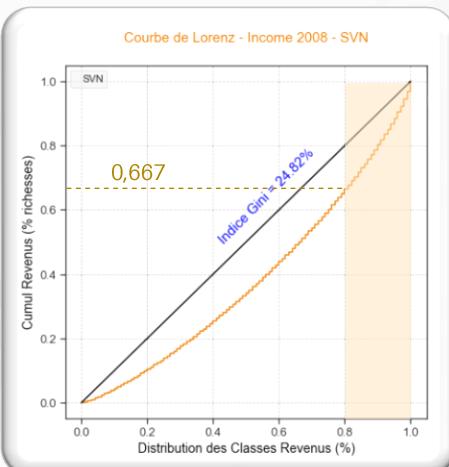
$$G = 2 \times (0,5 - \text{Aire Surface courbe/bissectrice})$$



## Cas Honduras

On constate que **20%** de la population, représentent **62,5%** des revenus cumulés du pays ( $1-0,375$ )...

*Très inégalitaire*



## Cas Slovénie

A contario On constate que **20%** de la population, représentent **33%** des revenus cumulés du pays ( $1-0,667$ )...

*Plutôt égalitaire*

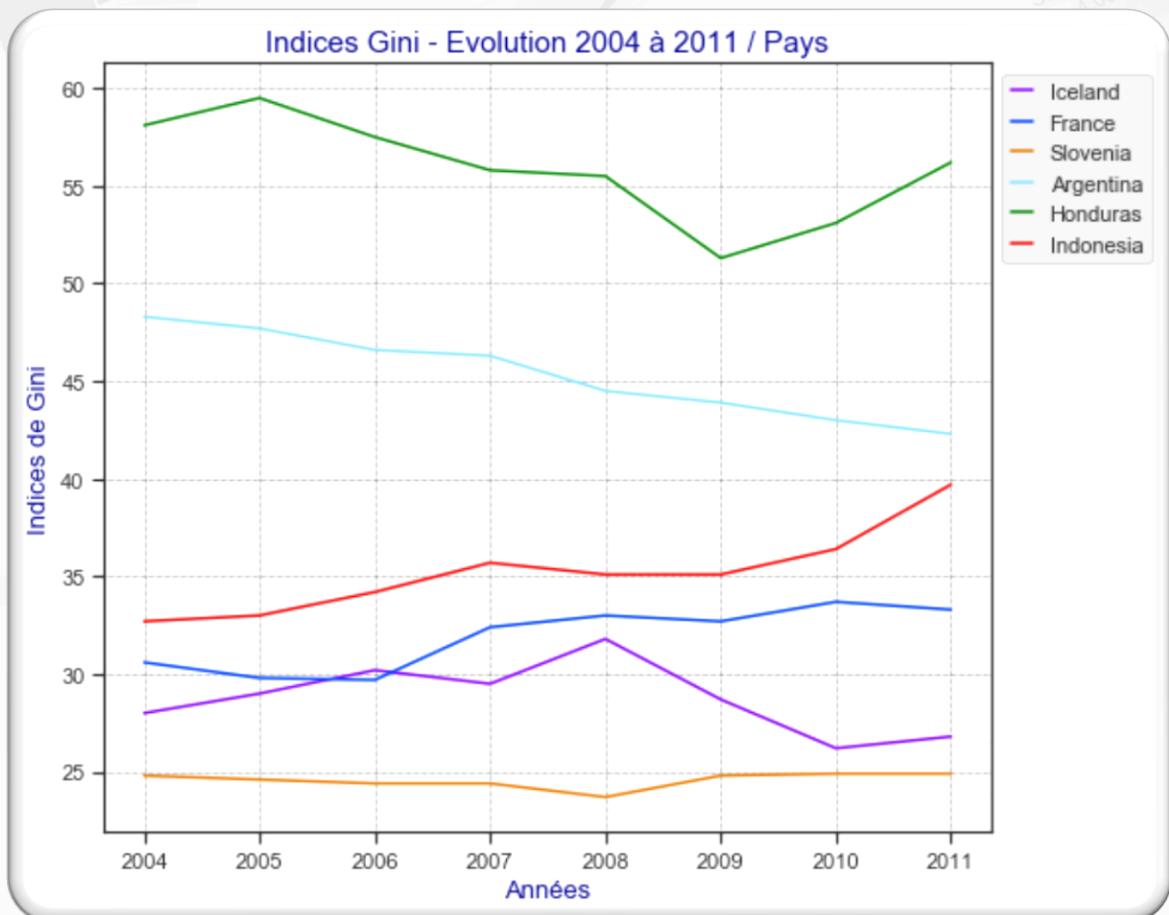
# Etude des revenus / pays (Indices de gini)

## Evolution Indices Gini 2004 - 2011

L'Indonésie et la France n'évoluent pas dans le bons sens et deviennent moins égalitaires sur la période

Slovénie et Honduras sont constants sur la période. La Slovénie déjà un bon élève, peut difficilement faire mieux.

L'Islande a baissé sur la fin de la période, tandis que l'Argentine est en baisse constante signe de meilleur équilibre entre les gens.



# Etude des revenus / pays (Indices de gini)

## Quelques chiffres

### 5 Pays les plus inégalitaires (gini max)

country_code	country_name	gini
ZAF	South Africa	63.733
NAM	Namibia	61.000
BWA	Botswana	60.500
CAF	Central African Republic	56.200
COM	Comoros	55.900

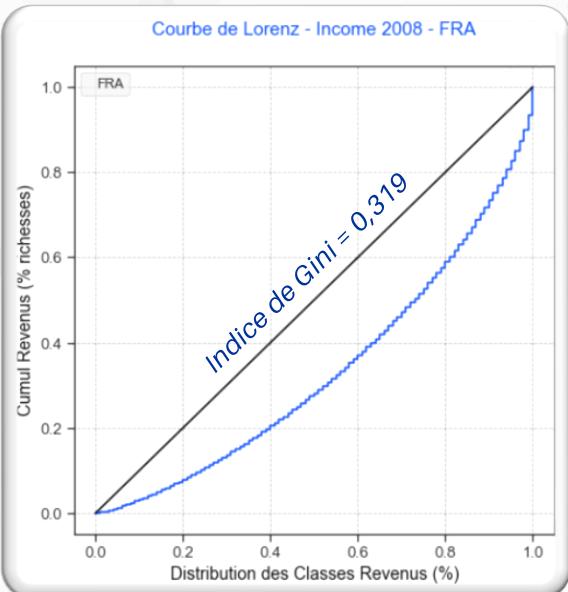
### 5 Pays les plus égalitaires (gini min)

country_code	country_name	gini
SVN	Slovenia	24.563
DNK	Denmark	26.075
CZE	Czech Republic	26.575
AZE	Azerbaijan	26.600
SVK	Slovak Republic	26.738

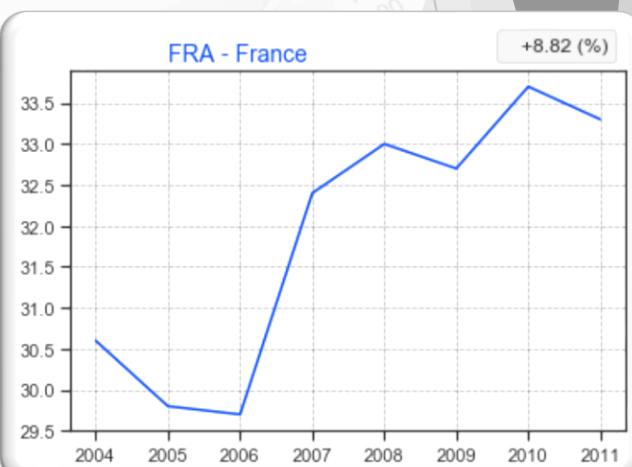
## La France dans ce classement

Rang → 34<sup>ème</sup>

rank	country_code	country_name	gini
27	LUX	Luxembourg	31.175
28	CYP	Cyprus	31.312
29	ARM	Armenia	31.375
30	EGY	Egypt, Arab Rep.	31.467
31	ETH	Ethiopia	31.500
32	PAK	Pakistan	31.500
33	LBN	Lebanon	31.800
34	FRA	France	31.900
35	KOR	Korea, Rep.	32.000
36	JPN	Japan	32.100
37	TJK	Tajikistan	32.200
38	KGZ	Kyrgyz Republic	32.250
39	HRV	Croatia	32.433
40	EST	Estonia	32.463
41	IRL	Ireland	32.587



	2013	2014	2015
Indice de Gini	0,288	0,289	0,292



# Détermination Classe de revenu parent

**Rappel:** Coefficient d'élasticité ou de mobilité intergénérationnelle (gdim)

Il nous permet de déduire la probabilité (%) d'un individu appartenant à une classe donnée ( $c_{i\_parent}$ ) parent d'évoluer, de changer de classe ( $c_{i\_child}$ ).

Soient       $i$       → individu  
                  $j$       → pays  
                 Y      → revenu

Ce que l'on connaît

$\rho_j$       → coefficient d'élasticité (mobilité intergénérationnelle gdim)

Ce que l'on cherche à déterminer

$c_{i,parent}$       → classe de revenu des parent d'un individu  $i$

On simulera cette information avec le coefficient d'élasticité d'un pays.

Il mesure la mobilité intergénérationnelle du revenu d'un pays.

Soit la corrélation entre le revenu d'un individu  $i$  et le revenu de ses parents  
Selon la formule logarithmique suivante

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

Paramètres entrée

n      → taille de l'échantillon

nb\_qt      → nombre de quantiles (classes)

pj      → coefficient d'élasticité (du pays)

Programme de  
calcul des  
probabilités  
conditionnelles  
- Restitue une  
Table

Exemple : Table résultante

- n = 1000

- nb\_q = 100 quantiles

- pj = 0,9

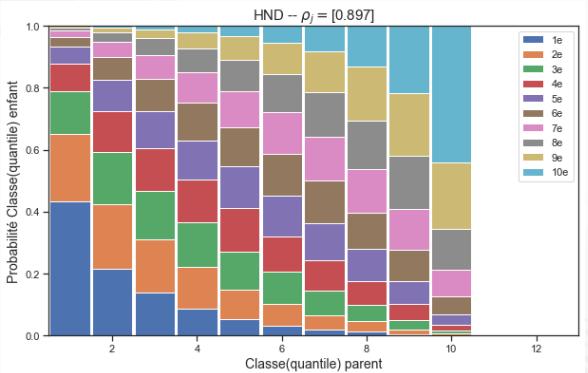
y_child	y_parent	c_i_child	c_i_parent
1361	0.025	0.083	1
1483	0.029	0.048	1
1604	0.012	0.066	1
1864	0.014	0.042	1
2215	0.024	0.086	1
...	...	...	...
98145	84.088	14.179	100
98275	30.803	10.375	100
99295	61.104	41.325	100
99434	37.754	10.607	100
99719	27.311	11.979	100

100000 rows × 4 columns

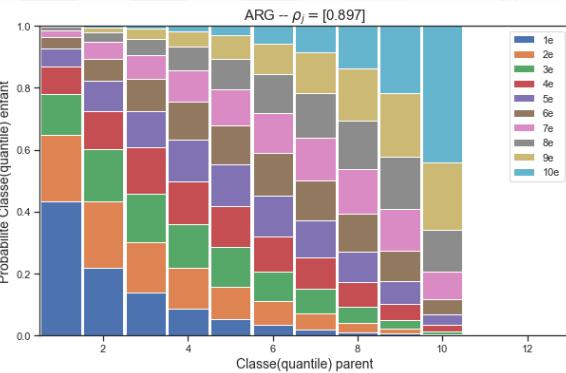


# Détermination Classe de revenu parent

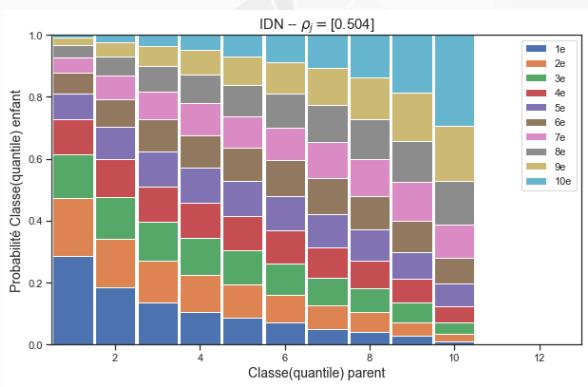
Tests → Sur coefficients élasticité de la sélection de pays de l'étude précédente



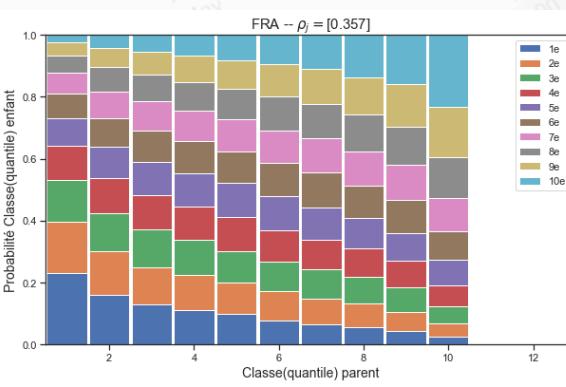
*Honduras*  
*Argentine*  
 $\rho_j = 0,897$



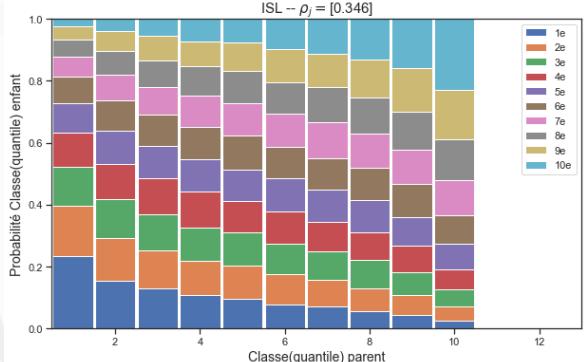
*Découpage Revenus*  
→ Déciles



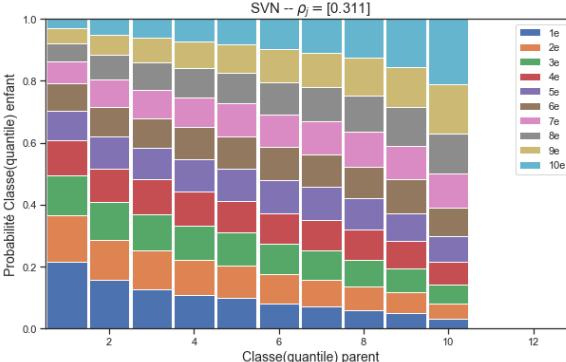
*Indonésie*  
 $\rho_j = 0,504$   
*France*  
 $\rho_j = 0,357$



Plus le coefficient baisse,  
plus il y a de mobilité  
sociale en terme de  
revenus dans le pays :  
- Ex. Slovénie



*Islande*  
 $\rho_j = 0,346$   
*Slovénie*  
 $\rho_j = 0,311$



En revanche, une valeur  
élevée indique que l'on a  
moins de chance de  
changer de classe donc  
d'évoluer quand on  
appartient aux classes  
inférieures :  
- Ex. Honduras/Argentine  
42% de chances de rester  
en classe 1



# Détermination Classe de revenu parent

**Application sur notre Dataset → Crédation de 500 « clones »**

Pour avoir donc plus d'individus par tranche/classes de revenus dans un pays.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5600000 entries, 0 to 11099
Data columns (total 6 columns):
country_code      category
c_i_child         int8
pj                float64
income             float64
gini               float64
gdpppp             float64
dtypes: category(1), float64(4), int8(1)
memory usage: 224.3 MB
```

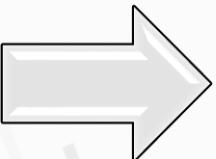
wid500						
country_code	c_i_child	pj	income	gini	gdpppp	c_i_parent
ALB	1	0.816	728.898	30.300	7,297.000	1
ALB	2	0.816	916.662	30.300	7,297.000	1
ALB	3	0.816	1,010.916	30.300	7,297.000	1
ALB	4	0.816	1,086.908	30.300	7,297.000	1
ALB	5	0.816	1,132.700	30.300	7,297.000	1

**Constitution des classes « parent » → Ventilation du (%) sur 500**

Exemple, si la probabilité  $P(c_i.parent=8|c_i.child=5,pj=0.9)=3\%$   
 ➔ 15 lignes sur 500 auront la classe parent 8 ( $500 \times 0.03 = 15$ )

country_code	pj
ALB	0.816
ARG	0.897
ARM	0.466
AUT	0.245
AZE	0.466

Programme de Ventilation  
En fonction de probabilités



Class					
y_child	y_parent	c_i_child	c_i_parent	country_code	c_i_parent
0.044	0.067	1	1	ALB	1
0.021	0.024	1	1	ALB	1
0.049	0.048	1	1	ALB	1
0.037	0.096	1	1	ALB	1
0.041	0.097	1	1	ALB	1

**DataSet – Anova / Régressions Linéaires**

country_code	c_i_child	pj	income	gini	gdpppp	c_i_parent
0	ALB	1	0.816	728.898	30.300	7,297.000
0	ALB	1	0.816	728.898	30.300	7,297.000
0	ALB	1	0.816	728.898	30.300	7,297.000
0	ALB	1	0.816	728.898	30.300	7,297.000
0	ALB	1	0.816	728.898	30.300	7,297.000
...	...	...	...	...	...	...
11099	ZAF	100	0.677	82,408.550	63.733	9,602.000
11099	ZAF	100	0.677	82,408.550	63.733	9,602.000
11099	ZAF	100	0.677	82,408.550	63.733	9,602.000
11099	ZAF	100	0.677	82,408.550	63.733	9,602.000
11099	ZAF	100	0.677	82,408.550	63.733	9,602.000

5600000 rows × 7 columns

# Anova - Régression Linéaire - Etude Résidus

Le test d'ANOVA (analyse de la variance) ou  $R^2$  ( $\eta^2$ ) permet de savoir si une donnée qualitative et une donnée quantitative sont corrélées.

Le résultat obtenu est compris entre 0 et 1. Plus il est proche de 1 plus les données sont corrélées..

## Formule de Décomposition de la Variance

ANOVA		
(ANalysis Of VAriance)		
SCT	=SCE	+SCR
$\sum_{i=1}^n (y_i - \bar{y})^2$	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$

## Coefficient de Détermination

$R^2$

$$R^2 = \frac{SCE}{SCT}$$

Ce coefficient  $R^2$  est compris entre [0,1]

En effet,  $0 \leq SCE \leq SCT$

- **SCT** (Somme des Carrés Totale) traduit la **variation totale** de Y.
- **SCE** (Somme des Carrés Expliquée) traduit la **variation interclasse** (expliquée par le modèle).
- **SCR** (Somme des Carrés Résiduelle) traduit la **variation intraclasses** (inexpliquée par le modèle, erreurs).

Si  $R^2 = 1$ , on a alors  $SCE=SCT$  : toute la variation est expliquée par le modèle.

Si  $R^2 = 0$ , on a alors  $SCR=SCT$  : aucune variation n'est expliquée par le modèle.

Décomposition de la variance :

$$SCT = SCM + SCE$$

observations = modèle + bruit

Definition (le coefficient de détermination  $R^2$ )

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} = \frac{SCM}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT}$$

Données = Modèle sous-jacent + Bruit indépendant  
(Résidus - Variance inexpliquée)

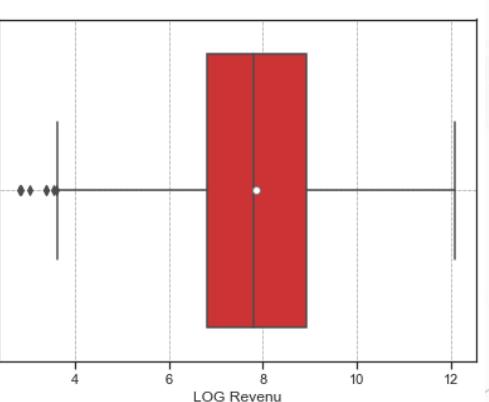
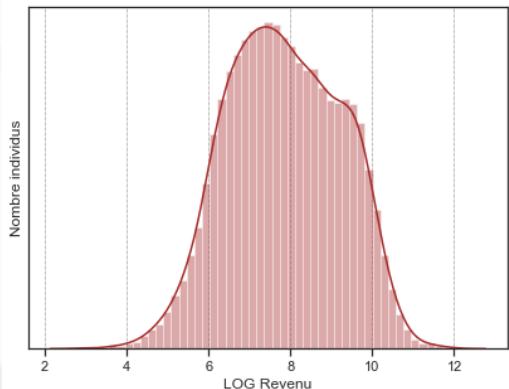
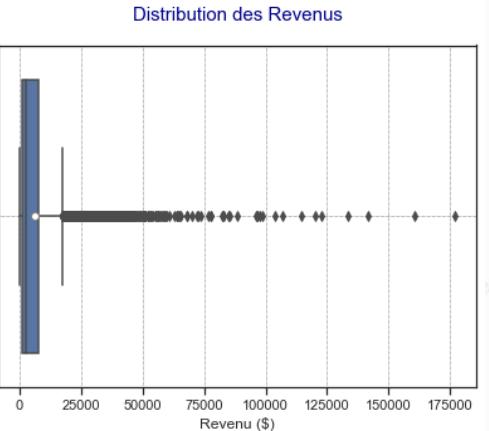
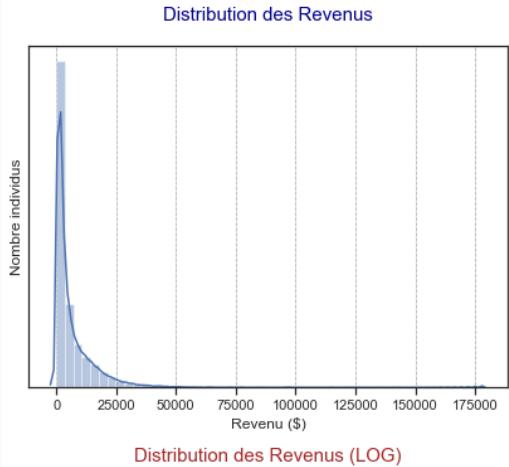


# Anova - Régression Linéaire - Etude Résidus

On cherche à expliquer le revenu d'un individu en utilisant le "PAYS" comme variable explicative.  
Existe-t-il une dépendance entre le revenus et le pays.

En d'autres termes, on cherche à savoir s'il y a une différence statistiquement significative entre les moyennes des revenus des pays par exemple.

Graphes des Revenus moyens / centiles



## 1<sup>er</sup> Constat

Le passage au Logarithme  
→ Rend la distribution "normale (gaussienne)"

Nous verrons par la suite ce que cela implique lors d'une régression linéaire.



# Anova - Régression Linéaire - Etude Résidus

## Résultat ANOVA revenu <- fct(pays)

### OLS Regression Results

```
=====
Dep. Variable: income    R-squared:      0.494
Model:          OLS         Adj. R-squared:  0.489
Method:         Least Squares F-statistic:   97.55
Date:           Wed, 07 Aug 2019 Prob (F-statistic):  0.00
Time:           20:40:36   Log-Likelihood: -1.1458e+05
No. Observations: 11200    AIC:             2.294e+05
Df Residuals:   11088    BIC:             2.302e+05
Df Model:        111
Covariance Type: nonrobust
```



Le pays a donc bien un effet sur le revenu moyen

comme nous en avions l'intuition en regardant la distribution ci-contre

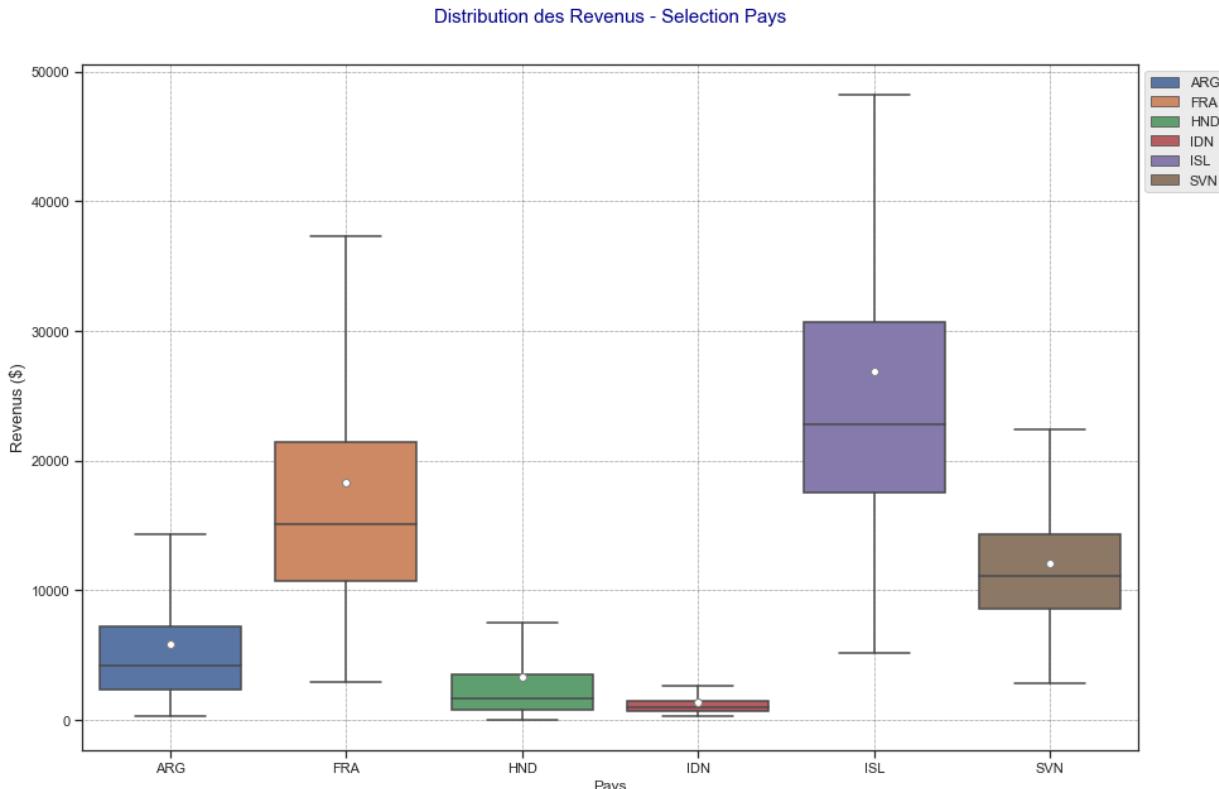
	sum_sq	df	F	PR(>F)
country_code	492,832,628,998.266	111.000	97.546	0.000
Residual	504,687,076,205.208	11,088.000	nan	nan

Ce qui nous intéresse réellement, c'est le test de Fisher.

La p-valeur de ce test (0) est :

- largement < au risque de 1<sup>ère</sup> espèce 5%.

On rejette donc l'hypothèse nulle (H0) selon laquelle le revenu moyen par centile est identique pour chaque pays.



# Anova - Régression Linéaire - Etude Résidus

OPENCLASSROOMS

NEURONAL  
PRIVATE  
BANKING

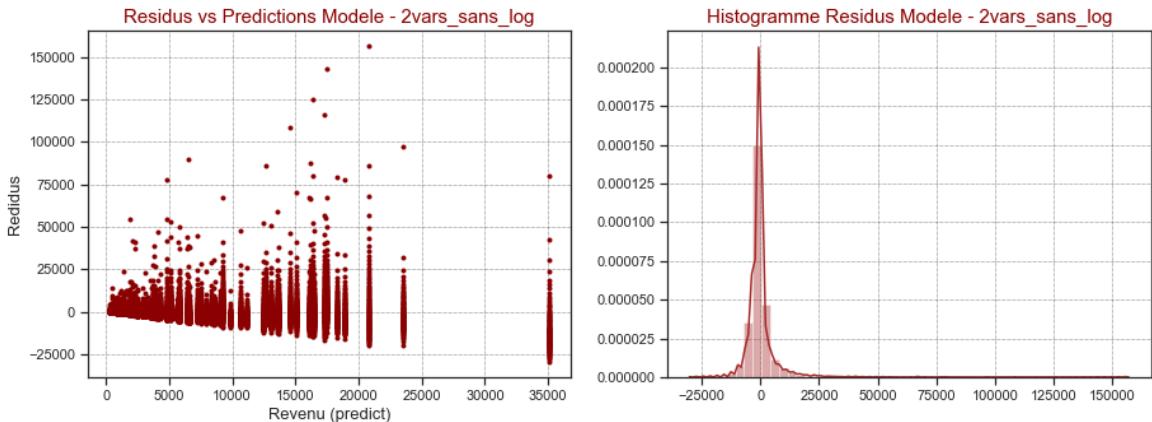
## Régressions Linéaires 2 variables :

- *Revenu moyen du pays (gdppp)*
- *Indice de gini*

2 versions avec ou sans passage au logarithme de ces variables

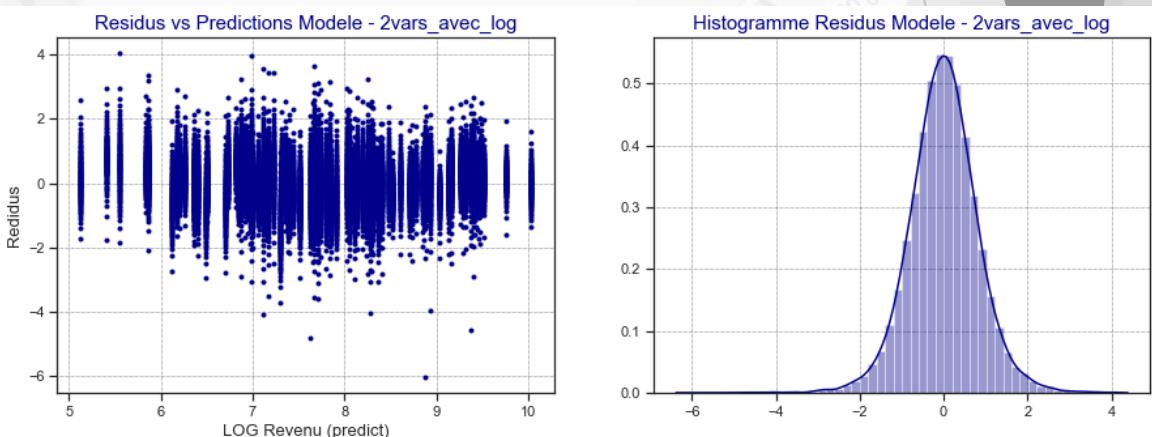
### Sans Logarithme

44,6(%) de Variance expliquée (Coeff.R<sup>2</sup>)  
55,4(%) de variance non expliquée (résidus)



### Avec Logarithme

65,9(%) de Variance expliquée (Coeff.R<sup>2</sup>)  
34,1(%) de variance non expliquée (résidus)



## Régressions Linéaires 3 variables :

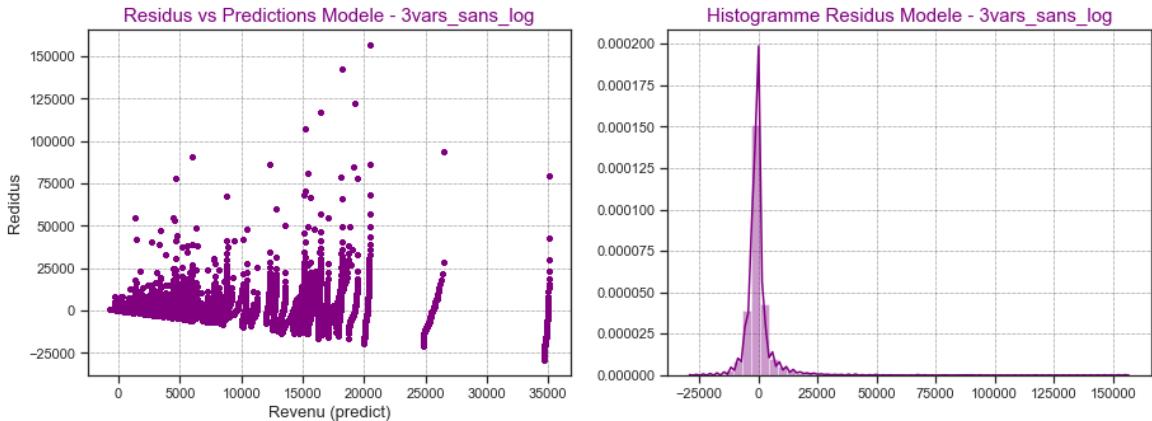
- *Revenu moyen du pays (gdppp)*
- *Indice de gini*
- *Classe de revenus des parents*

2 versions avec ou sans passage au logarithme de ces variables

### Sans Logarithme

45,5(%) de Variance expliquée (Coeff.R<sup>2</sup>)

54,5(%) de variance non expliquée (résidus)

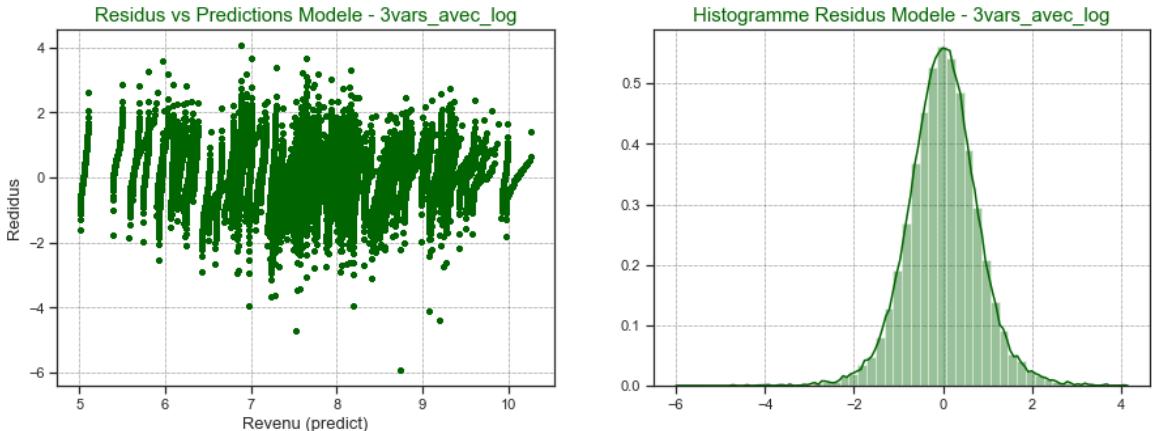


### Avec Logarithme

67,2(%) de Variance expliquée (Coeff.R<sup>2</sup>)

32,8(%) de variance non expliquée (résidus)

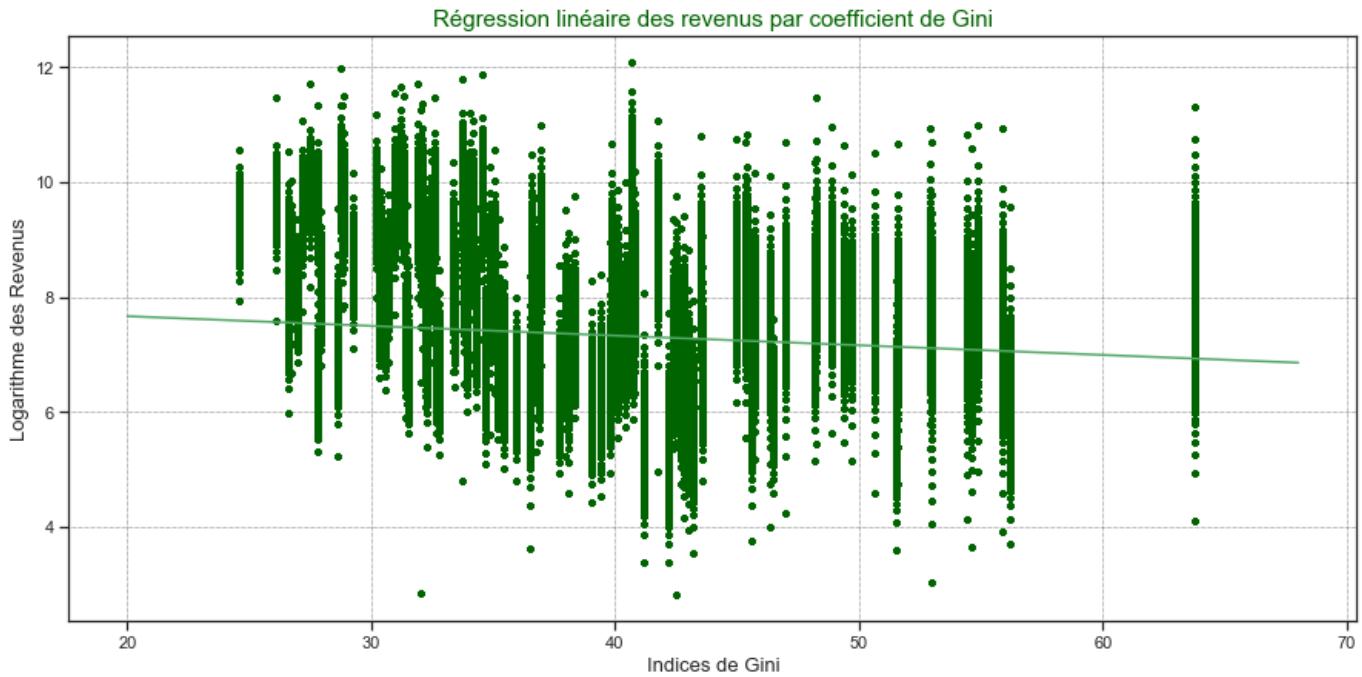
**Meilleur modèle des 4**



# Anova - Régression Linéaire - Etude Résidus

## Question ?

En observant le coefficient de régression associé à l'indice de Gini, peut-on affirmer que le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu'il n'en défavorise ?



L'indice de Gini va de 0 (égalité parfaite) à 1 (inégalité parfaite).

Le coefficient directeur de la droite de régression linéaire de Gini étant négatif, plus l'indice de Gini va augmenter, plus les revenus vont diminuer.

Donc NON, on ne peut pas affirmer que le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu'il n'en défavorise.

## Modèle

Régression Linéaire

3 variables (Log)

- Revenu moyen du pays
- Indice de gini
- Classe de revenus des parents



## Un bon résidu

est un résidu sans structure ou sans structure apparente :

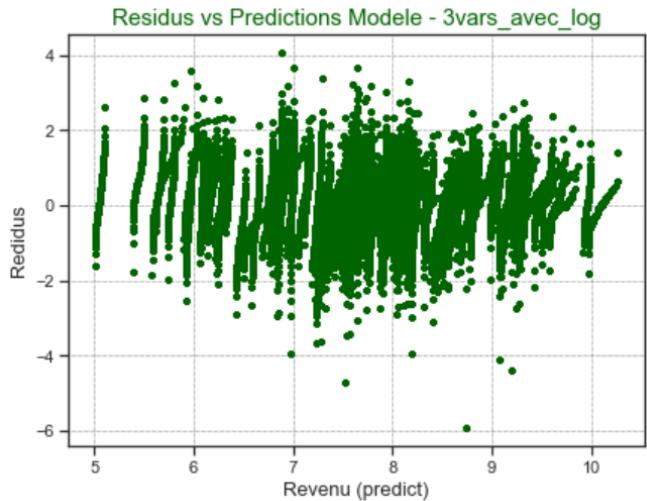
- o Les résidus non structurés ont une variance constante (*Homoscédasticité*)  
Tests → *Breusch Pagan / Bartlett* (si distr.gaussienne) / *Levene*
- o Leur distribution est «normale»  
Tests → *Shapiro-wilk, Kolmogorov-Smirnov*
- o Il n'y a pas de points aberrants.  
Calcul de *Leviers, Distance de cook, Résidus «Studentisés»*

## Test d'homoscédasticité (variances constantes)

- p value test Breusch Pagan: 0.0

La p-valeur ici est très nettement inférieure au seuil de risque de 1<sup>ère</sup> espèce 5 % :

On rejette donc l'hypothèse nulle ( $H_0$ ) selon laquelle les variances sont constantes (l'hypothèse d'homoscédasticité).



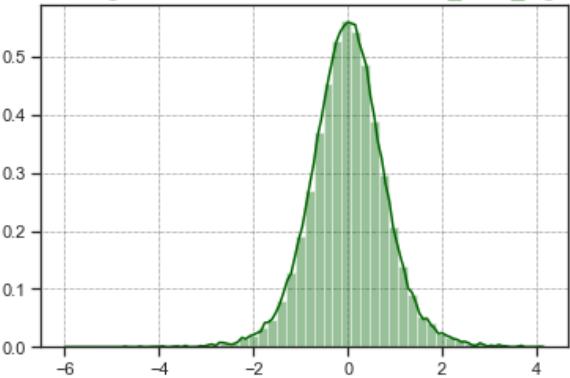
## Test normalité des résidus

- $p$  value test Shapiro-Wilk : 0.0

Ici, l'hypothèse de normalité est remise en cause ( $p\text{-value} = 0.0 < 0.05$ ).

Néanmoins, l'observation des résidus, le fait qu'ils ne soient pas très différents d'une distribution symétrique, et le fait que l'échantillon soit de taille gigantesque (5'600'000), permettent de dire que les résultats obtenus par le modèle linéaire gaussien ne sont pas absurdes, même si le résidu n'est pas considéré comme étant gaussien par shapiro-Wilk.

Histogramme Residus Modele - 3vars\_avec\_log



De plus un avertissement nous est fourni sur la robustesse du Test de Shapiro sur les grands échantillons

```
shapiro(reglin3log.resid)
```

```
C:\Users\boiss\Anaconda3\Lib\site-packages\scipy\stats\morestats.py:1653: UserWarning: p-value may not be accurate for N > 5000.  
warnings.warn("p-value may not be accurate for N > 5000.")
```

```
(0.9956063628196716, 0.0)
```

# Anova - Régression Linéaire - Etude Résidus

OPENCLASSROOMS

NEURONAL  
PRIVATE  
BANKING

## Points Aberrants ?

Soient :

$n$ , le nombre d'individus de l'échantillon,  
 $p$ , le nombre de variables.

### Calcul des Leviers

Un levier est considéré comme un point atypique. On peut calculer les leviers comme ceci, en sachant que le seuil des leviers est de

$$\text{Seuil} = 2 * \frac{p+1}{n}$$

Nombre d'observations dont levier > seuil :  
**288500**  
(5,15% du Dataset)

country_code	nb_obs
ZAF	50000
TLS	41000
NOR	22500
SVN	21000
CAF	21000
GTM	19500
URY	18000
RUS	15000
MYS	14500
YEM	10500

### Calcul des Résidus studentisés

Le seuil des résidus studentisés est une loi de Student à  $n-p-1$  degrés de liberté

country_code	nb_obs
SWZ	21500
ZAF	11000
HND	11000
BOL	10000
NIC	7000
COL	7000
CIV	6500
MEX	6000
PAN	6000
BRA	6000

Nombre d'observations dont résidus hors seuil  
**297000**  
(5,30% du Dataset)

### Calcul Distance de Cook

Ici une observation sera très influente, si la distance supérieure au seuil de

$$\text{Seuil} = \frac{4}{(n - p)}$$

Nombre d'observations dont dist\_cook > seuil  
**642**  
(5,73% du Dataset)

country_code	nb_obs
ZAF	54
SWZ	50
HND	36
CAF	34
LBR	33
COL	26
BOL	25
BRA	25
PAN	23
GTM	19

country_code	pj	gini	gdpppp
BOL	BOL	0.866	52.971
CAF	CAF	0.665	56.200
COL	COL	1.095	54.433
HND	HND	0.897	55.875
SWZ	SWZ	0.665	51.500
ZAF	ZAF	0.677	63.733

Bolivia  
Central African Republic  
Colombia  
Honduras  
Eswatini  
South Africa

Calcul fait sur dataset non clonés car sinon beaucoup trop long à exécuter  
(11'200 lignes vs 5'600'000)

# Projet 7 - Prédictions de Revenus

Ce Modèle de régression linéaire multiple sur 3 variables & conversion log

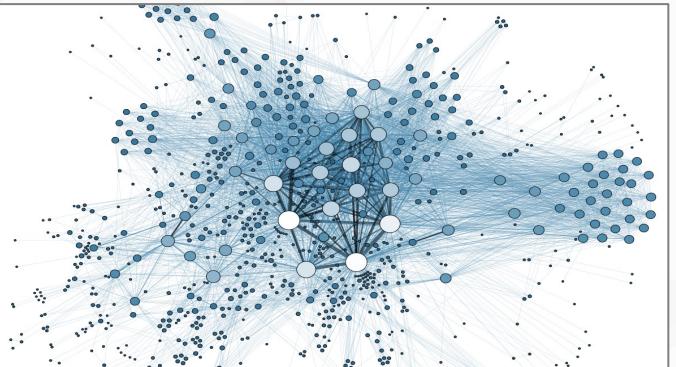
**N'est pas optimal.**

L'étude des résidus nous indique qu'il pourrait être amélioré soit en rajoutant des variables explicatives manquantes (efforts, niveau d'étude, etc...)

Ou bien en écartant certaines données de l'étude correspondant par exemple aux pays associés à l'observations des points aberrants (étude résidus).

Cependant on peut tout de même conclure que le fait de naître dans un pays riche et le fait d'appartenir à une classe de revenus élevée restent des critères assez prépondérants pour déterminer la classe de revenus d'un individu (client cible).

Merci de votre attention !



**QUESTIONS ?**

Pour plus d'informations, vous pouvez me contacter en suivant le lien directement ci-dessous

[frederic.boissy@gmail.com](mailto:frederic.boissy@gmail.com)



# Annexe 1

Explications de la sélection de pays utilisés pour le graphique de distributions des indices de Gini.

Choix des 4 pays ayant des valeurs extrêmes tant en terme de revenu\_moyen qu'en terme d'incide de gini moyen sur la période 2004-2011

Visualisation de France & Argentine, sélectionnées arbitrairement

	country_code	country_name	rev_moy	gini_moy		country_code	country_name	rev_moy	gini_moy
0	SVN	Slovenia	12,106.01	24.56	0	IDN	Indonesia	1,334.62	35.24
1	DNK	Denmark	17,043.15	26.07	1	GEO	Georgia	1,363.76	38.05
2	CZE	Czech Republic	8,235.29	26.57	2	ARM	Armenia	1,628.38	31.38
3	SVK	Slovak Republic	6,096.58	26.74	3	KGZ	Kyrgyz Republic	1,773.22	32.25
4	UKR	Ukraine	3,349.39	27.00	4	MDA	Moldova	2,149.17	33.93
5	SWE	Sweden	16,184.22	27.14	5	KAZ	Kazakhstan	2,239.15	30.57
6	NOR	Norway	22,483.38	27.49	6	SLV	El Salvador	2,855.22	45.71
7	FIN	Finland	16,306.33	27.80	7	PRY	Paraguay	3,278.08	51.60
8	BLR	Belarus	3,921.16	27.91	8	HND	Honduras	3,296.27	55.88
9	ISL	Iceland	26,888.51	28.78	9	PER	Peru	3,330.53	48.16
10	BEL	Belgium	15,024.61	28.81	10	UKR	Ukraine	3,349.39	27.00
11	NLD	Netherlands	17,728.64	28.89	11	ECU	Ecuador	3,383.74	50.65
12	HUN	Hungary	6,101.34	29.24	12	DOM	Dominican Republic	3,558.40	49.38
13	AUT	Austria	16,637.60	30.21	13	BLR	Belarus	3,921.16	27.91
14	KAZ	Kazakhstan	2,239.15	30.57	14	PAN	Panama	5,135.14	52.91
15	DEU	Germany	18,061.72	30.96	15	CRI	Costa Rica	5,580.39	48.84
16	LUX	Luxembourg	25,217.56	31.18	16	POL	Poland	5,741.72	33.80
17	CYP	Cyprus	17,345.39	31.31	17	ARG	Argentina	5,847.88	45.32
18	ARM	Armenia	1,628.38	31.38	18	TUR	Turkey	6,050.47	39.84
19	FRA	France	18,309.41	31.90	19	SVK	Slovak Republic	6,096.58	26.74
20	KGZ	Kyrgyz Republic	1,773.22	32.25	20	HUN	Hungary	6,101.34	29.24
21	EST	Estonia	7,702.06	32.46	21	LTU	Lithuania	6,623.66	35.06
22	IRL	Ireland	17,710.74	32.59	22	LVA	Latvia	6,764.47	36.55
23	POL	Poland	5,741.72	33.80	23	RUS	Russian Federation	7,156.77	40.69
24	MDA	Moldova	2,149.17	33.93	24	EST	Estonia	7,702.06	32.46
25	ITA	Italy	14,925.21	34.01	25	CZE	Czech Republic	8,235.29	26.57
26	ESP	Spain	13,116.99	34.16	26	PRT	Portugal	10,098.68	36.97
27	GRC	Greece	11,727.27	34.18	27	GRC	Greece	11,727.27	34.18
28	GBR	United Kingdom	21,709.60	34.57	28	SVN	Slovenia	12,106.01	24.56
29	LTU	Lithuania	6,623.66	35.06	29	ESP	Spain	13,116.99	34.16
30	IDN	Indonesia	1,334.62	35.24	30	ITA	Italy	14,925.21	34.01
31	LVA	Latvia	6,764.47	36.55	31	BEL	Belgium	15,024.61	28.81
32	PRT	Portugal	10,098.68	36.97	32	SWE	Sweden	16,184.22	27.14
33	GEO	Georgia	1,363.76	38.05	33	FIN	Finland	16,306.33	27.80
34	TUR	Turkey	6,050.47	39.84	34	AUT	Austria	16,637.60	30.21
35	RUS	Russian Federation	7,156.77	40.69	35	DNK	Denmark	17,043.15	26.07
36	ARG	Argentina	5,847.88	45.32	36	CYP	Cyprus	17,345.39	31.31
37	SLV	El Salvador	2,855.22	45.71	37	IRL	Ireland	17,710.74	32.59
38	PER	Peru	3,330.53	48.16	38	NLD	Netherlands	17,728.64	28.89
39	CRI	Costa Rica	5,580.39	48.84	39	DEU	Germany	18,061.72	30.96
40	DOM	Dominican Republic	3,558.40	49.38	40	FRA	France	18,309.41	31.90
41	ECU	Ecuador	3,383.74	50.65	41	GBR	United Kingdom	21,709.60	34.57
42	PRY	Paraguay	3,278.08	51.60	42	NOR	Norway	22,483.38	27.49
43	PAN	Panama	5,135.14	52.91	43	LUX	Luxembourg	25,217.56	31.18
44	HND	Honduras	3,296.27	55.88	44	ISL	Iceland	26,888.51	28.78

# Annexes

## Guide de Choix d'un Test Statistique (Excel)

Guide Choix Tests Statistiques.xlsx - Excel

	B	C	D	E	F	G
1	Données	Hypothèse nulle (H0)	Exemples	Tests paramétriques "Echantillons Appariés" (mêmes individus)	Conditions de validité (tests paramétriques)	Tests non-paramétriques "Echant. Non appariés"
2	mesures sur 1 échantillon ; moyenne théorique (1 chiffre)	moyenne observée = moyenne théorique	Comparaison à une norme d'un taux de pollution mesuré	Test t ( <a href="#">Student</a> ) pour un échantillon	2	Test de <a href="#">Wilcoxon</a> pour un échantillon
3	mesures sur 2 échantillons	Les positions* sont identiques	Comparaison de notes d'étudiants entre deux classes	Test t ( <a href="#">Student</a> ) pour échantillons indépendants	1 ; 3 ; 5	<a href="#">Mann-Whitney</a>
4	mesures sur plusieurs échantillons	Les positions* sont identiques	Comparaison du rendement de maïs selon 4 engrains différents	<a href="#">ANOVA</a>	1 ; 3 ; 4 ; 6	Kruskal-Wallis
5	deux séries de mesures quanti sur les mêmes individus (avant-après)	Les positions* sont identiques	Comparaison du taux d'hémoglobine moyen avant / après l'application d'un traitement sur un groupe de patients	Test t ( <a href="#">Student</a> ) pour échantillons appariés	10	<a href="#">Wilcoxon</a>
6	Plusieurs séries de mesures quanti sur les mêmes individus (avant-après)	Les positions* sont identiques	Suivi de la concentration d'un élément trace au cours du temps au sein d'un groupe de plantes	<a href="#">ANOVA</a> à mesures répétées; <a href="#">modèles mixtes</a>	10 ; Sphéricité	Friedman
7	Plusieurs séries de mesures binaires sur les mêmes individus (avant-après)	Les positions* sont identiques	Différents juges évaluent la présence/l'absence d'un attribut sur différents produits			Test Q de <a href="#">Cochran</a>
8	Mesures sur deux échantillons	variance(1) = variance(2)	Comparaison de la dispersion naturelle de la taille de 2 variétés d'un fruit	Test de <a href="#">Fisher</a>		
9	Mesures sur plusieurs échantillons	variance(1) = variance(2) = variance(n)	Comparaison de la dispersion naturelle de la taille de plusieurs variétés d'un fruit	Test de <a href="#">Levene</a> Test de <a href="#">Bartlett</a>		
10	une proportion observée ; son effectif associé ; une proportion théorique	proportion observée = proportion théorique	Comparaison de la proportion de femelles à une proportion de 0.5 dans un échantillon	Test pour une proportion ( <a href="#">chi²</a> )		
11	Effectif de chaque catégorie	proportion(1) = proportion(2) = proportion(n)	Comparaison des proportions de 3 couleurs d'yeux dans un échantillon	<a href="#">chi²</a>		
12	Proportion théorique et effectif associés à chaque catégorie	proportions observées = proportions théoriques	Comparer les proportions de génotypes obtenus par croisement F1xF1 à des proportions mendéliennes (1/2, 1/4, 1/2)	Test d'ajustement multinomial		
13	Tableau de contingence	variable 1 et variable 2 sont indépendantes	La présence d'un attribut est-elle liée à la présence d'un autre attribut?	<a href="#">chi²</a> sur un tableau de contingence	1 ; 9	Test exact de <a href="#">Fisher</a> ; méthode de Monte Carlo
14	mesures de deux variables sur un échantillon	variable 1 et variable 2 sont indépendantes	La biomasse de plante change-t-elle avec la concentration de Pb?	Corrélation de <a href="#">Pearson</a> (& <a href="#">Neyman</a> )	7 ; 8	Corrélation de <a href="#">Spearman</a>
	Mesures d'une variable quantitative sur un échantillon; paramètres de la distribution	Les distributions observée et théorique sont les mêmes	Les salaires d'une société suivent-ils une distribution normale de moyenne 2500 et			<a href="#">Kolmogorov-Smirnov</a>

