Prédiction de revenus

A la recherche de nos futurs clients

Projet 7 - DA – Marc Sellam 10.2020

Introduction :

La banque souhaite cibler les clients les plus susceptibles d'avoir, plus tard dans leur vie, de hauts revenus en se basant sur le revenus de leurs parents, avec comme informations fournies les données regroupant 116 pays.

Sommaire :

Mission 1 : préparation des données

Il nous a été fourni un tableau de données contenant les informations de revenus de 116 pays.

Nous complétons les données ‘gdpppp’ pour les pays ‘XKX’ , ’PSE’ et nous rajoutons à ‘LTU’ son quantile manquant.

Nous travaillerons donc a partir d’un tableau concernant 116 pays, des années 2004 à 2011, l’année 2008 concerne 74 pays.

Nous corrigeons la valeur aberrante du ‘gdpppp’ du pays ‘Fiji’ par le remplacement de sa véritable valeur.

Mission 1 : préparation des données

Constantly updated lists of world countries, territories and areas of geographical interest, with associated alpha-2, alpha-3 and numeric codes as defined by the ISO 3166-1 standard, published and maintained by the International Organization for Standardization, available in SQL, CSV, JSON and PHP formats, in multiple languages and with national flags included.
https://github.com/stefangabos/world_countries/blob/7d19c162fc6df8773e5660c5dabe1b6e9cdf9c8a/data/sk/countries.json

```
: #"region":1 = Nordic European countries and Canada |    0.2
  #"region":2 = Europe (except nordic countries)    |    0.4
  #"region":3 = Australia/New Zealand/USA |    0.4
  #"region":4 = Asia                      |    0.5
  #"region":5 = Latin America/Africa      |    0.66

  ccode=[{"id":4,"name":"Afghanistan","alpha2":"af","alpha3":"afg","region":4},
  {"id":8,"name":"Albania","alpha2":"al","alpha3":"alb","region":2},
  {"id":12,"name":"Algeria","alpha2":"dz","alpha3":"dza","region":5},
  {"id":20,"name":"Andorra","alpha2":"ad","alpha3":"and","region":2},
  {"id":24,"name":"Angola","alpha2":"ao","alpha3":"ago","region":5},
  {"id":28,"name":"Antigua and Barbuda","alpha2":"ag","alpha3":"atg","region":3},
  {"id":32,"name":"Argentina","alpha2":"ar","alpha3":"arg","region":5},
  {"id":51,"name":"Armenia","alpha2":"am","alpha3":"arm","region":2},
  {"id":36,"name":"Australia","alpha2":"au","alpha3":"aus","region":3},
  {"id":40,"name":"Austria","alpha2":"at","alpha3":"aut","region":2},
  {"id":31,"name":"Azerbaijan","alpha2":"az","alpha3":"aze","region":4},
  {"id":44,"name":"Bahamas","alpha2":"bs","alpha3":"bhs","region":3},
  {"id":48,"name":"Bahrain","alpha2":"bh","alpha3":"bhr","region":4},
  {"id":50,"name":"Bangladesh","alpha2":"bd","alpha3":"bgd","region":4},
```

Mission 1 : préparation des données

	country	year_survey	quantile	nb_quantiles	income	gdpppp	Area	population
0	ALB	2008	1	100	728.89795	7297.00000	Albania	3002678
1	ALB	2008	2	100	916.66235	7297.00000	Albania	3002678
2	ALB	2008	3	100	1010.91600	7297.00000	Albania	3002678
3	ALB	2008	4	100	1086.90780	7297.00000	Albania	3002678
4	ALB	2008	5	100	1132.69970	7297.00000	Albania	3002678
...	...	...	...	...	...	...	...	...
11595	COD	2008	96	100	810.62330	303.19305	Congo	4011486
11596	COD	2008	97	100	911.78340	303.19305	Congo	4011486
11597	COD	2008	98	100	1057.80740	303.19305	Congo	4011486
11598	COD	2008	99	100	1286.60290	303.19305	Congo	4011486
11599	COD	2008	100	100	2243.12260	303.19305	Congo	4011486

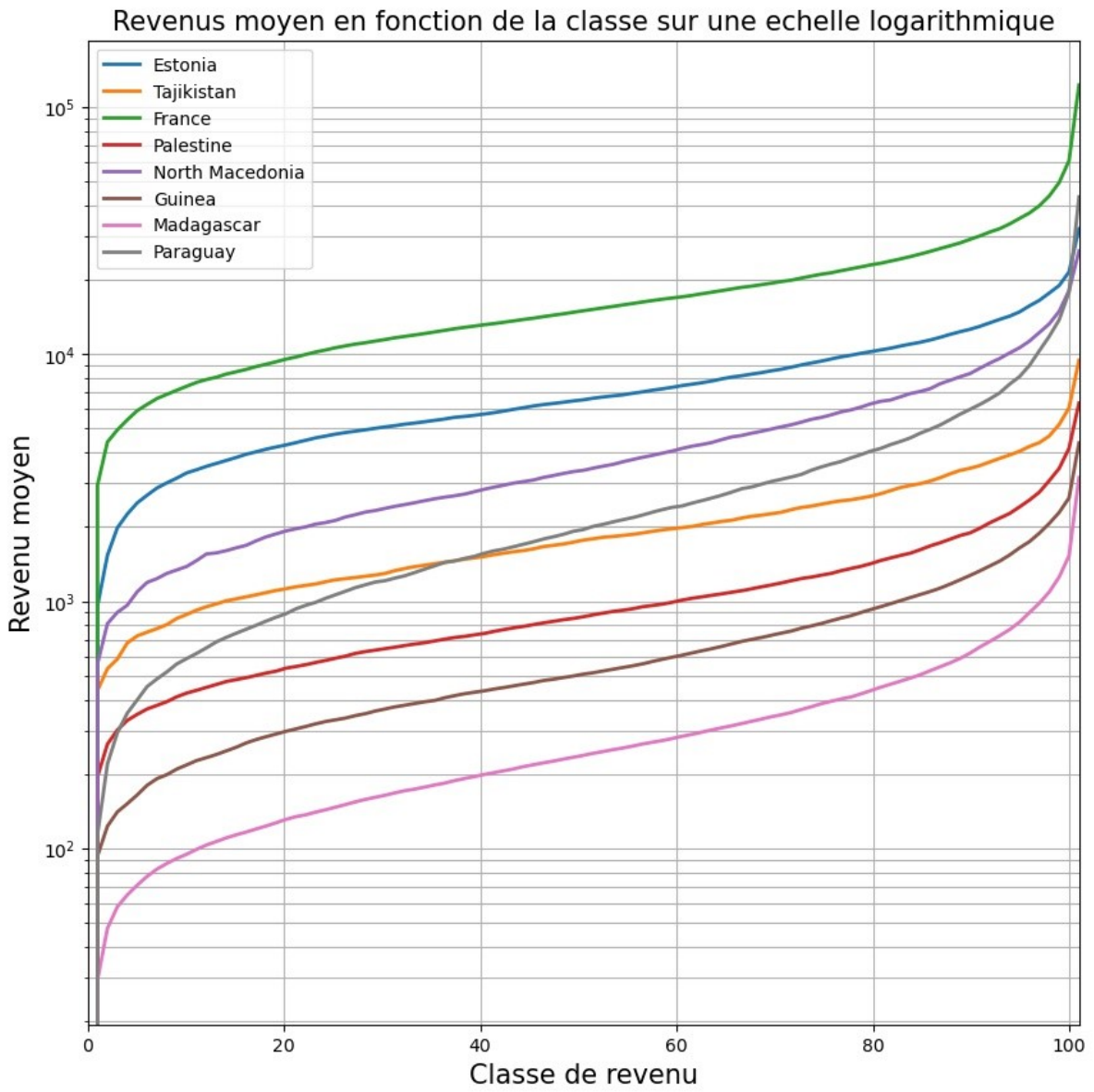
11600 rows × 8 columns

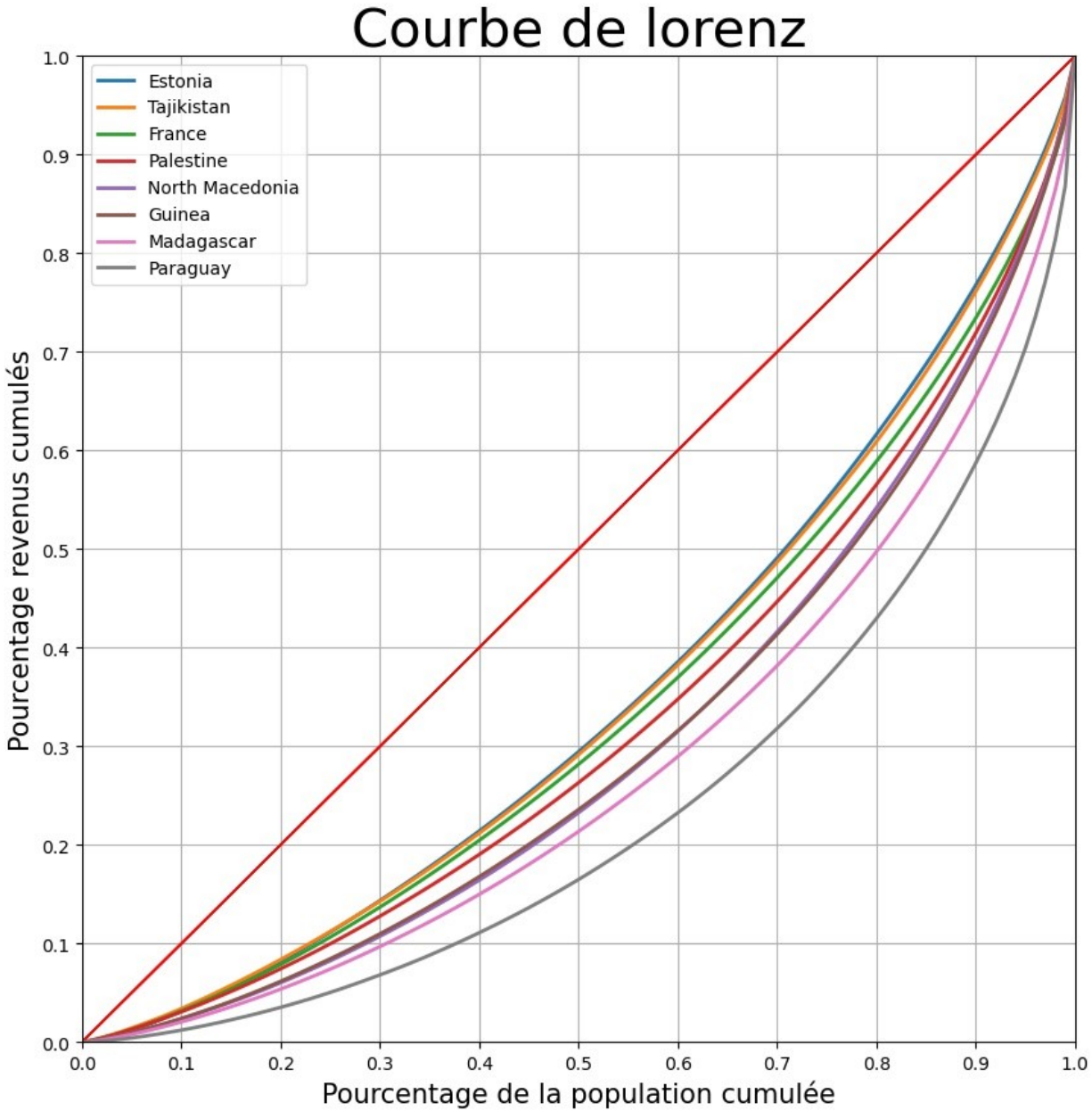
Les données utilisées représentent 91,56 % de la population mondiale.

Mission 2 : les indices de Gini

Segmentation des 116 pays en 8 groupes avec l'algorithme kmeans .

Puis, choix d'un pays représentatif de son groupe, défini par sa proximité minimum avec le centre de celui-ci.



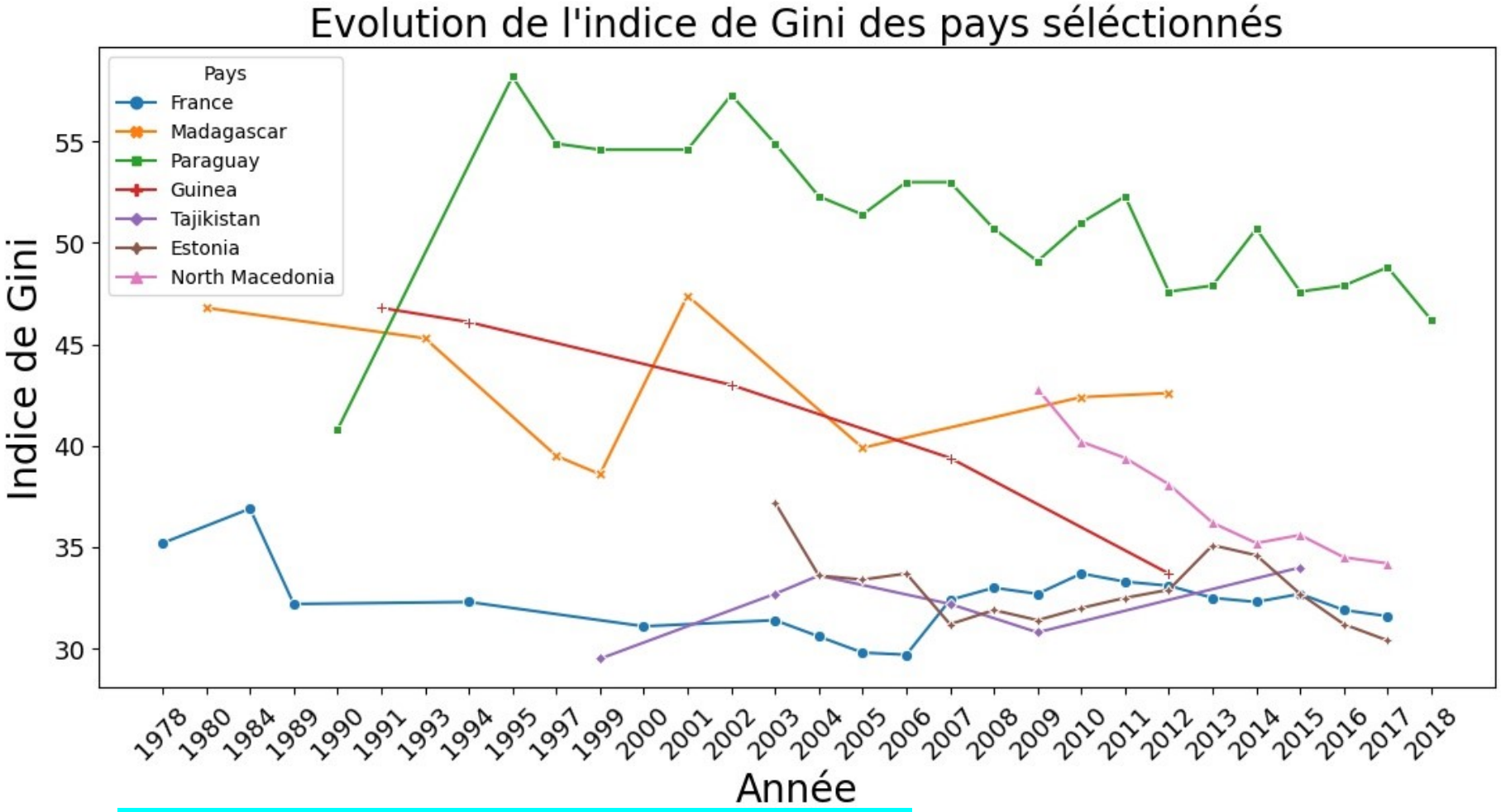


Mission 2 :

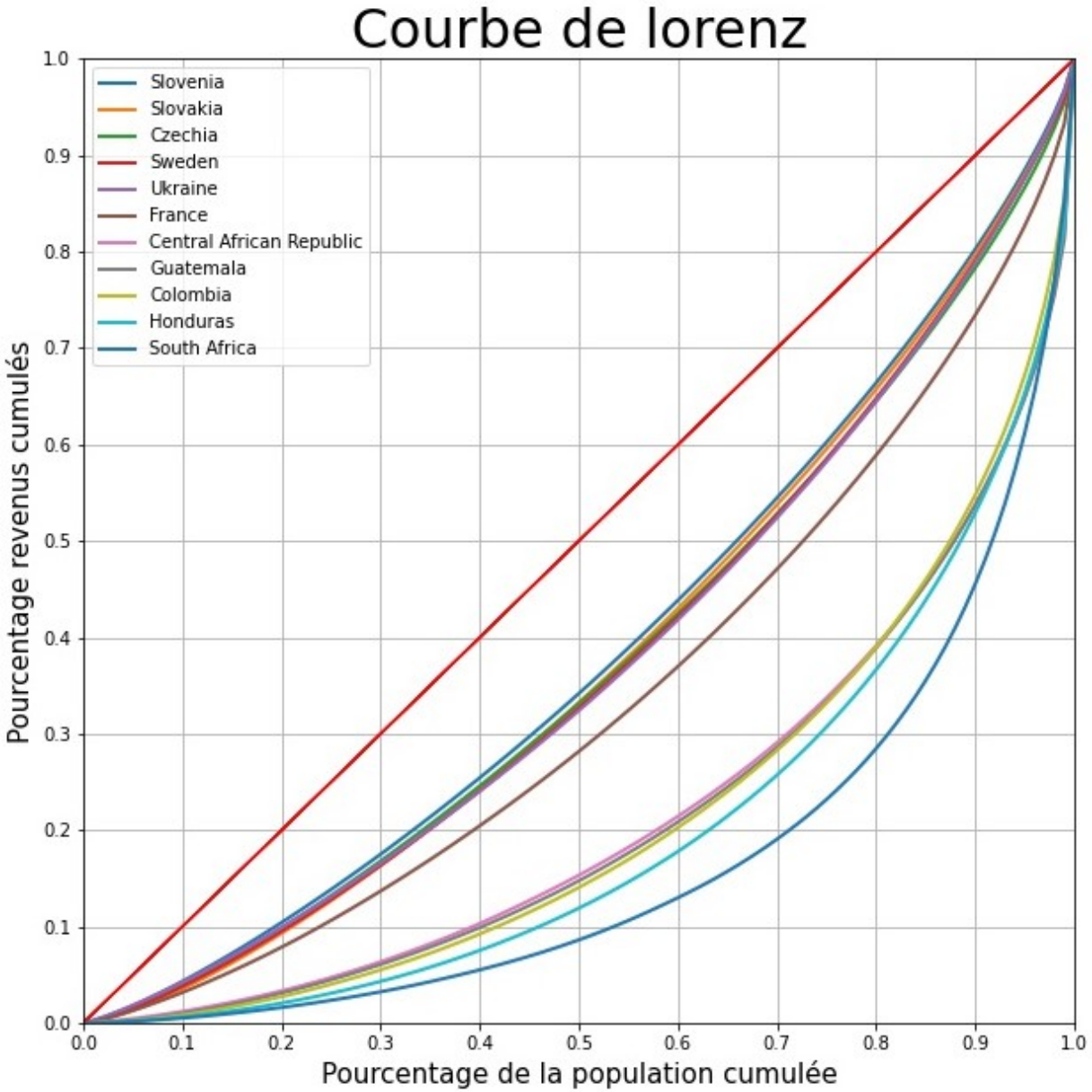
Indices Gini des 8 pays représentatifs de chaque groupe:

Estonia	**: 0.30**
Tajikistan	**: 0.31**
France	**: 0.33**
Palestine	**: 0.36**
North Macedonia	**: 0.40**
Guinea	**: 0.40**
Madagascar	**: 0.44**
Paraguay	**: 0.53**

Mission 2 :



L'Estonie et la France ont les indice de Gini les plus bas et tendent à s'améliorer. .
La macédoine du nord fais de gros progrès en terme d'évolution, tout comme le Paraguay dans une moindre mesure.
Les autres pays ont une tendance a la hausse et/ou des données trop anciennes, voir inexistantes.



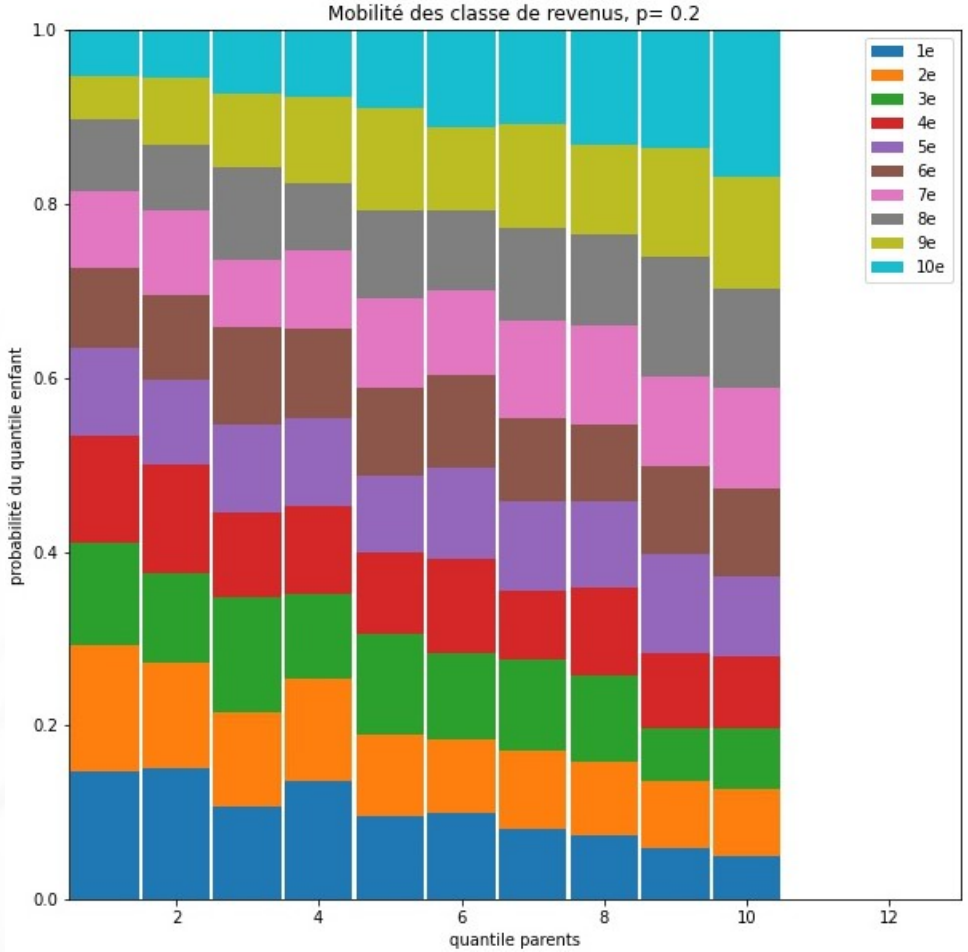
Mission 2 :

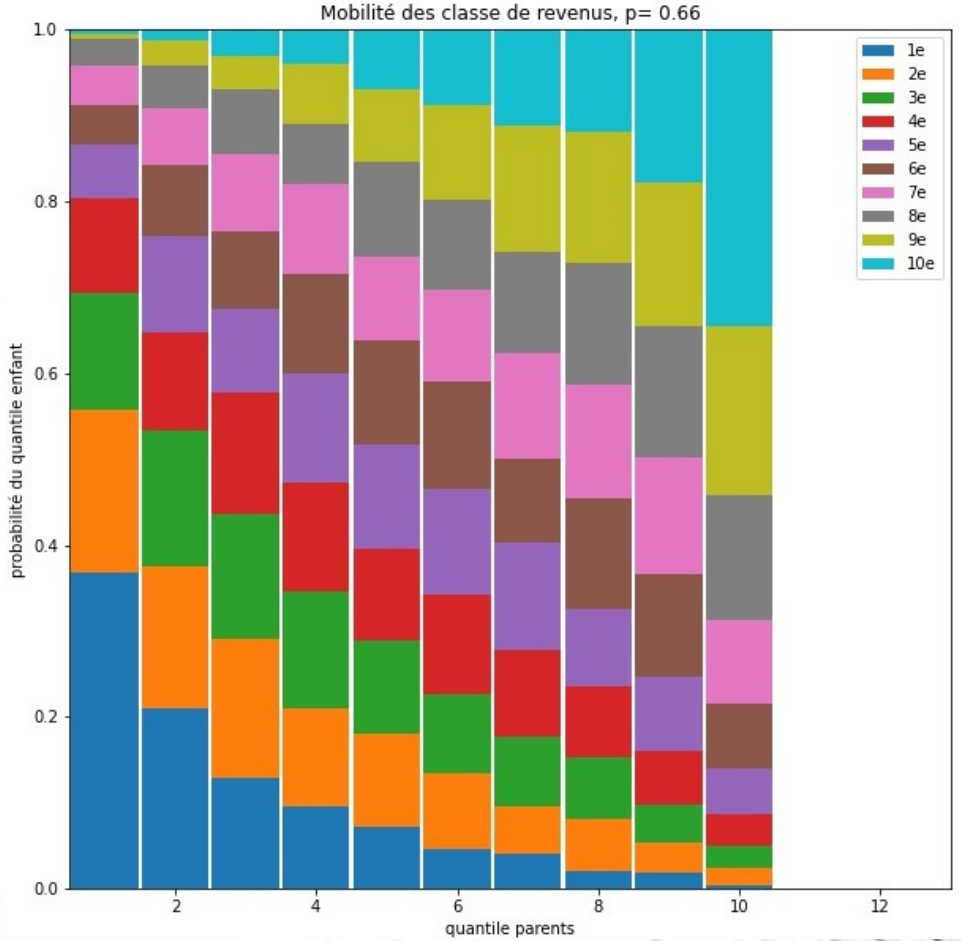
Indices de Gini:

Slovenia	**0.23**
Slovakia	**0.25**
Czechia	**0.25**
Sweden	**0.25**
Ukraine	**0.26**
France (40e)	**0.33**
Central African	**0.56**
Guatemala	**0.57**
Colombia	**0.57**
Honduras	**0.60**
South Africa	**0.67**
Moyenne des 116 pays	**0,38**

Mission 3 : Comparatif de mobilité intergénérationnelle

la suède (coefficient d'élasticité : 0,2) et l'Afrique du sud (coefficient d'élasticité : 0,66)





Mission 4 : Création d'un nouvel échantillon

Intégration du coefficient d'élasticité pj et génération de la colonne proba_cond

	Area	y_child	y_parents	c_i_child	c_i_parent	pj	Gini	sal_moy	population	proba_cond
0	Albania	0.470637	0.662840	25	35	0.40	0.304624	2994.83	3002678.0	0.008
1	Albania	0.384282	1.206952	19	58	0.40	0.304624	2994.83	3002678.0	0.007
2	Albania	0.948058	0.959296	49	49	0.40	0.304624	2994.83	3002678.0	0.012
3	Albania	1.560411	0.330967	67	14	0.40	0.304624	2994.83	3002678.0	0.010
4	Albania	0.214171	0.221732	8	7	0.40	0.304624	2994.83	3002678.0	0.028
...	...	...	...	...	...	...	...	...	...	...
11599995	Congo	0.471758	0.333330	27	14	0.66	0.443997	276.02	4011486.0	0.014
11599996	Congo	0.087938	0.227316	3	7	0.66	0.443997	276.02	4011486.0	0.035
11599997	Congo	7.514530	2.067218	96	77	0.66	0.443997	276.02	4011486.0	0.013
11599998	Congo	2.403203	1.191634	77	57	0.66	0.443997	276.02	4011486.0	0.013
11599999	Congo	4.386279	4.186727	90	93	0.66	0.443997	276.02	4011486.0	0.018

11600000 rows × 10 columns

Mission 4 : Création d'un nouvel échantillon

Création d'un tableau de données 500 fois plus grand que le tableau de données d'origine .

	Area	c_i_child	c_i_parent	proba_cond	income	gdpppp	population	Gini	sal_moy	pj
0	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40
1	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40
2	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40
3	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40
4	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40
...	...	...	...	...	...	...	...	...	...	...
5799995	Congo	99	100	0.080	2243.12260	303.19305	4011486	0.443997	276.02	0.66
5799996	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66
5799997	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66
5799998	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66
5799999	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66

5800000 rows × 10 columns

Mission 4 : Création d'un nouvel échantillon

Ajout de la colonne revenus enfants 'income_c' .

	Area	c_i_child	c_i_parent	proba_cond	income	gdpppp	population	Gini	sal_moy	pj	income_c
0	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40	728.89795
1	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40	728.89795
2	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40	728.89795
3	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40	728.89795
4	Albania	1	1	0.077	728.89795	7297.00000	3002678	0.304624	2994.83	0.40	728.89795
...	...	...	...	...	...	...	...	...	...	...	...
5799995	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66	2243.12260
5799996	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66	2243.12260
5799997	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66	2243.12260
5799998	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66	2243.12260
5799999	Congo	100	100	0.133	2243.12260	303.19305	4011486	0.443997	276.02	0.66	2243.12260

5800000 rows × 11 columns

Mission 5 : Création d'un modèle

ANOVA du revenus des enfants avec comme variable explicative le pays de l'individu :

```
anova_inca = smf.ols('income_c~Area', data=f_data).fit()
print(anova_inca.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 income_c   R-squared:                       0.496
Model:                             OLS   Adj. R-squared:                  0.496
Method:                  Least Squares   F-statistic:                 4.971e+04
Date:                 Wed, 21 Oct 2020   Prob (F-statistic):               0.00
Time:                         02:26:01   Log-Likelihood:            -5.9310e+07
No. Observations:              5800000   AIC:                         1.186e+08
Df Residuals:                  5799884   BIC:                         1.186e+08
Df Model:                          115
Covariance Type:            nonrobust
==============================================================================
```

Mission 5 : Régression linéaire

Régression linéaire avec comme variables explicatives le revenu moyen du pays de l'individu en logarithme ainsi que l'indice Gini :

```
reg_multi = smf.ols('np.log(income_c)~np.log(gdpppp)+Gini', data=f_data).fit()
print(reg_multi.summary())
#la fonction log a bien amélioré le model
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       np.log(income_c)   R-squared:                       0.653
Model:                            OLS   Adj. R-squared:                  0.653
Method:                 Least Squares   F-statistic:                5.447e+06
Date:                Wed, 21 Oct 2020   Prob (F-statistic):               0.00
Time:                        02:28:45   Log-Likelihood:            -7.0366e+06
No. Observations:             5800000   AIC:                         1.407e+07
Df Residuals:                 5799997   BIC:                         1.407e+07
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.7917      0.003    239.365      0.000       0.785       0.798
np.log(gdpppp) 0.8654      0.000   3021.971      0.000       0.865       0.866
Gini          -1.4969      0.004   -378.325      0.000      -1.505      -1.489
==============================================================================
```

Tableau de l'analyse de la variance :

	sum_sq	df	F	PR(>F)
np.log(gdpppp)	6.051754e+06	1.0	9.132310e+06	0.0
Gini	9.484873e+04	1.0	1.431301e+05	0.0
Residual	3.843514e+06	5799997.0	NaN	NaN

Mission 5 : Régression linéaire

Régression linéaire avec comme variables explicatives le revenu moyen du pays (en logarithme), l'indice Gini de l'individu ainsi que la classe des revenus des parents:

Modèle retenu

np.log(income_c) =
0.2231
+ 0.8655*np.log(gdpppp)
+ 0.0113*c_i_parent
-1,4973*Gini

```
reg_multi = smf.ols('np.log(income_c)~np.log(gdpppp)+c_i_parent+Gini', data=f_data).fit()
print(reg_multi.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       np.log(income_c)   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.708
Method:                 Least Squares   F-statistic:                 4.686e+06
Date:                Wed, 21 Oct 2020   Prob (F-statistic):               0.00
Time:                        02:28:53   Log-Likelihood:            -6.5330e+06
No. Observations:             5800000   AIC:                         1.307e+07
Df Residuals:                 5799996   BIC:                         1.307e+07
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.2231      0.003     72.433      0.000       0.217       0.229
np.log(gdpppp) 0.8655      0.000   3296.142      0.000       0.865       0.866
c_i_parent     0.0113   1.07e-05   1048.699      0.000       0.011       0.011
Gini          -1.4973      0.004   -412.745      0.000      -1.504      -1.490
==============================================================================
```

Tableau de l'analyse de la variance

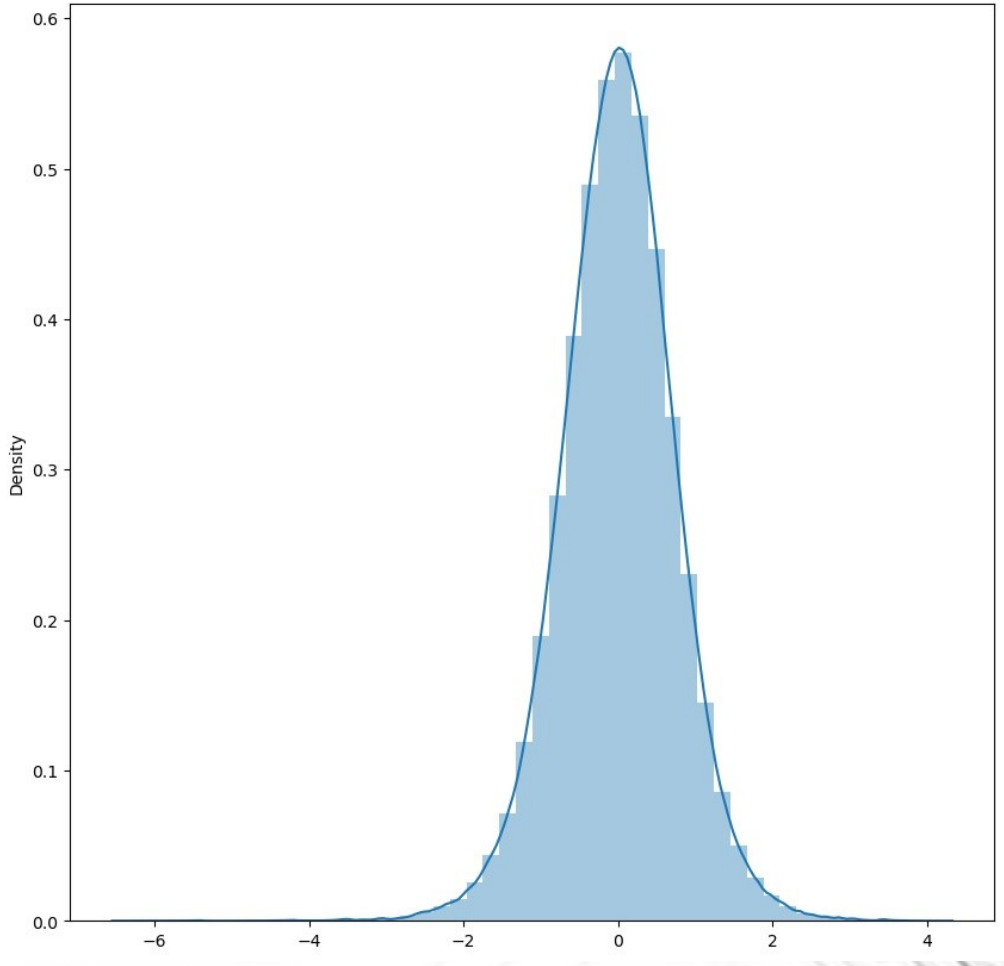
	sum_sq	df	F	PR(>F)
np.log(gdpppp)	6.052096e+06	1.0	1.086455e+07	0.0
c_i_parent	6.126266e+05	1.0	1.099770e+06	0.0
Gini	9.489833e+04	1.0	1.703588e+05	0.0
Residual	3.230887e+06	5799996.0	NaN	NaN

Mission 6 : tests statistiques du modèle retenu

Test de normalité des résidus :

```
test Kolmogorov-Smirnov
P-value = 0.0
```

l'hypothèse de normalité est remise en cause (p-value < 0.05).
Cependant l'observation des résidus permettent de la contredire.



Mission 6 : tests statistiques du modèle retenu

Vérification de la colinéarité des variables :

```
variance_inflation_factor, ou vif
```

```
[1.0949730068035852, 1.000000026088889, 1.0949730097750463]
```

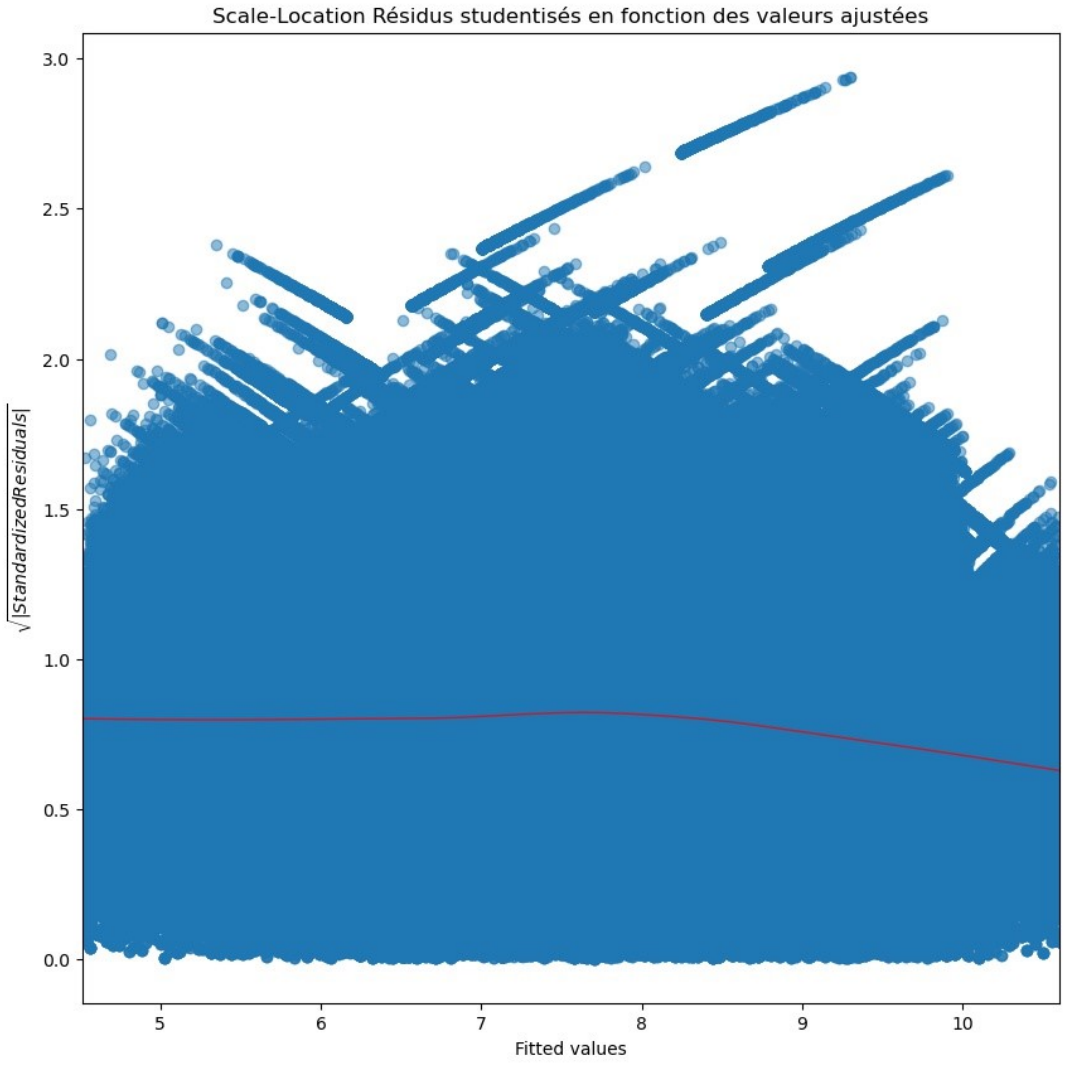
Ici, tous les coefficients sont inférieurs à 10, il n'y a donc pas de problème de colinéarité.

Mission 6 : tests statistiques du modèle retenu

Test d'homoscédasticité des résidus :

```
test Breusch Pagan
P-value: 0.0
```

La p-value ici est inférieure à 5 %, on rejette l'hypothèse $H0$ selon laquelle les variances sont homogènes (homoscédasticité).



Conclusion:

Vivre dans un pays inégalitaire favorise les bas salaires.

Notre modèle basé sur le salaire moyen du pays, l'indice Gini et la classe des parents explique pour plus de 70% les revenus de l'enfant, cependant d'autres paramètres pourraient améliorer notre modèle :

-Age/sex
-Niveau d'études
-Les écoles fréquentées
-Prix au m2 de la ville du domicile
-La motivation à réussir de l'enfant
-Que les parents soient dépensiers ou pas
….