Prédiction de revenus

PARCOUS DATA ANALYST - OPENCLASSROOMS

Nayescha GOMES

SOMMAIRE

SOMMAIRE

LE JEU DES DONNEES ET SON TRAITEMENT

La source des données utilisées dans ce projet :

Database *: World Income Distribution de 2008*

Indices de Gini : *World Bank*

Population : *FAO 2008*

GDIM : *World Bank*

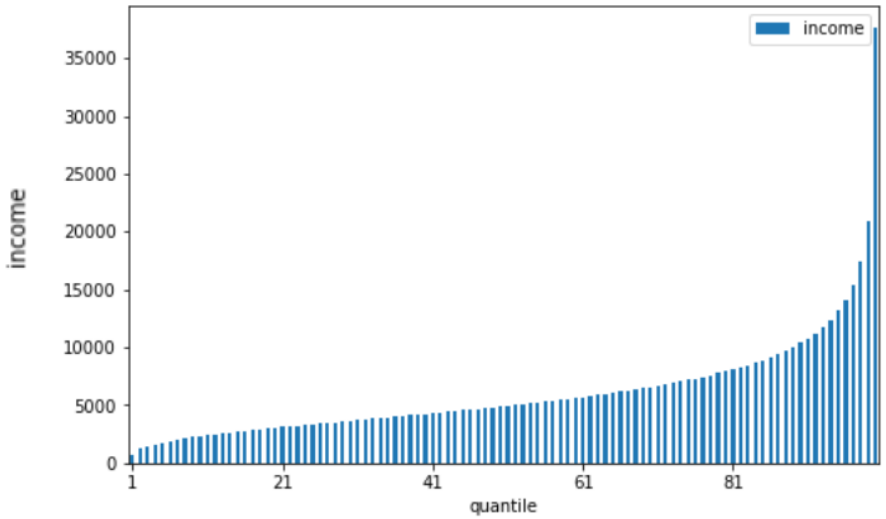
Le traitement des données :

- Vérification des doublons
- Recherche des valeurs nulles
- Vérification de la taille de l'échantillon des pays
- Remplissage des valeurs manquants pour le revenu moyen du quantile
- Vérification et remplissage des codes pays manquants
- Traitements dans le *dataframe* population

SOMMAIRE

MISSION I

- Notre base de données est composé par 116 pays, soit une représentativité de 77% de la population mondiale environ.
- Les données sont des années 2004, 2007, 2008, 2009, 2010, 2011.
- Les classes de revenus sont travaillés avec des centiles, ce qui réduit le risque d'avoir des outliers au sein d'un quantile mais crée aussi des grands écarts entre les quantiles dans les extrémités (les premiers et les derniers).
- L'utilisation des quantiles sert aussi à comparer différentes populations (avec une taille et structure différente)



SOMMAIRE

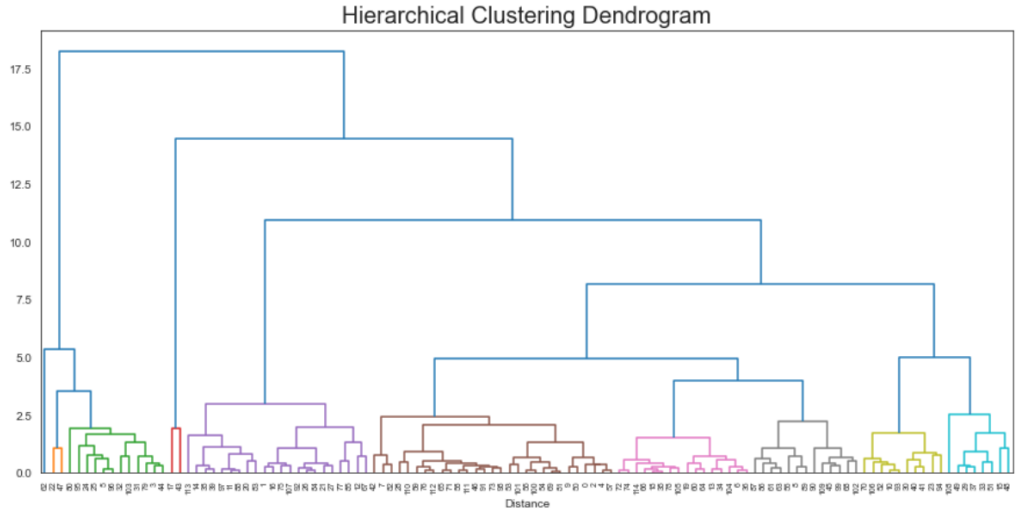
MISSION II

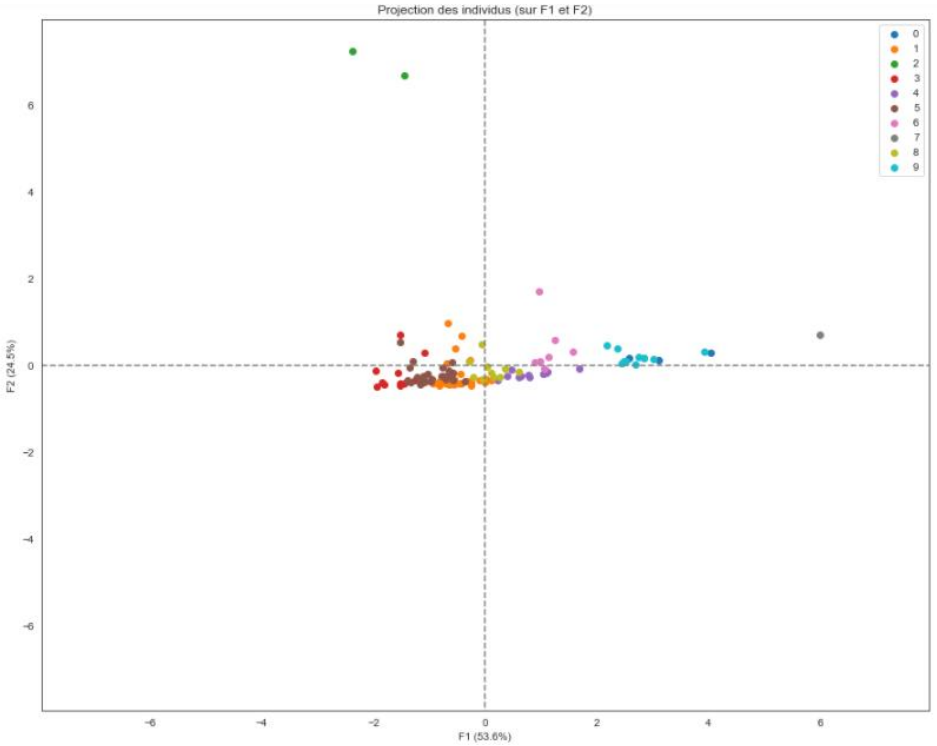
Les 5 premiers questions de la mission II ont été répondues dans le code :

- On génère les revenus des parents (exprimés en logs) selon une loi normale.
- La moyenne et variance n'ont aucune incidence sur le résultat final (ie. sur le calcul de la classe de revenu)
- Génération d'une réalisation du terme d'erreur epsilon
- Pour chacun des n individus générés, les classes de revenu sont calculées selon y_child et y_parent

MISSION II

Le *clustering* de la base en 10 et la projection des groupes sur le premier plan factoriel:





MISSION II

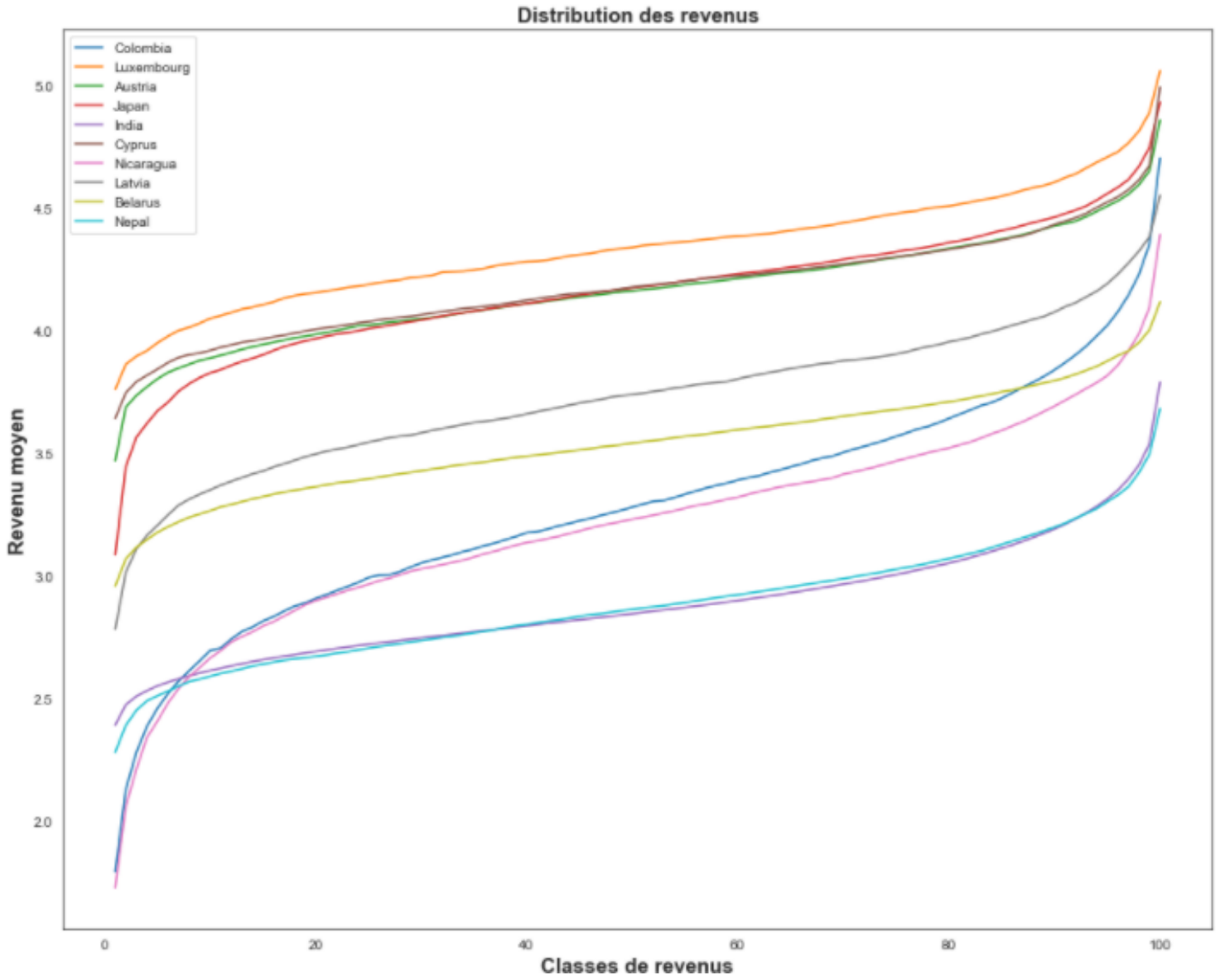
La représentation du cluster par le pays avec les valeurs les plus proches de la moyenne de son groupe :

	Pays	Code	year	cluster	income	gdpppp	Population	gini
27	Ecuador	ECU	2008	3	99.078545	7560	1.708436e+07	0.50
62	Luxembourg	LUX	2008	7	5780.837400	73127	6.042450e+05	0.33
3	Austria	AUT	2008	9	2958.076400	36193	8.891388e+06	0.30
51	Japan	JPN	2008	6	1224.340500	31307	1.272022e+08	0.35
43	India	IND	2007	2	247.838730	2796	1.352642e+09	0.33
22	Cyprus	CYP	2008	0	4406.335400	26273	1.189265e+06	0.32
63	Latvia	LVA	2008	8	609.161250	15596	1.928459e+06	0.37
10	Belarus	BLR	2008	4	913.442140	11651	9.452617e+06	0.28
105	Uganda	UGA	2009	5	121.179810	1067	4.272904e+07	0.44
81	Nepal	NPL	2010	1	192.238780	1048.1808	2.809571e+07	0.33

MISSION II

La diversité de distribution de revenus par pays :

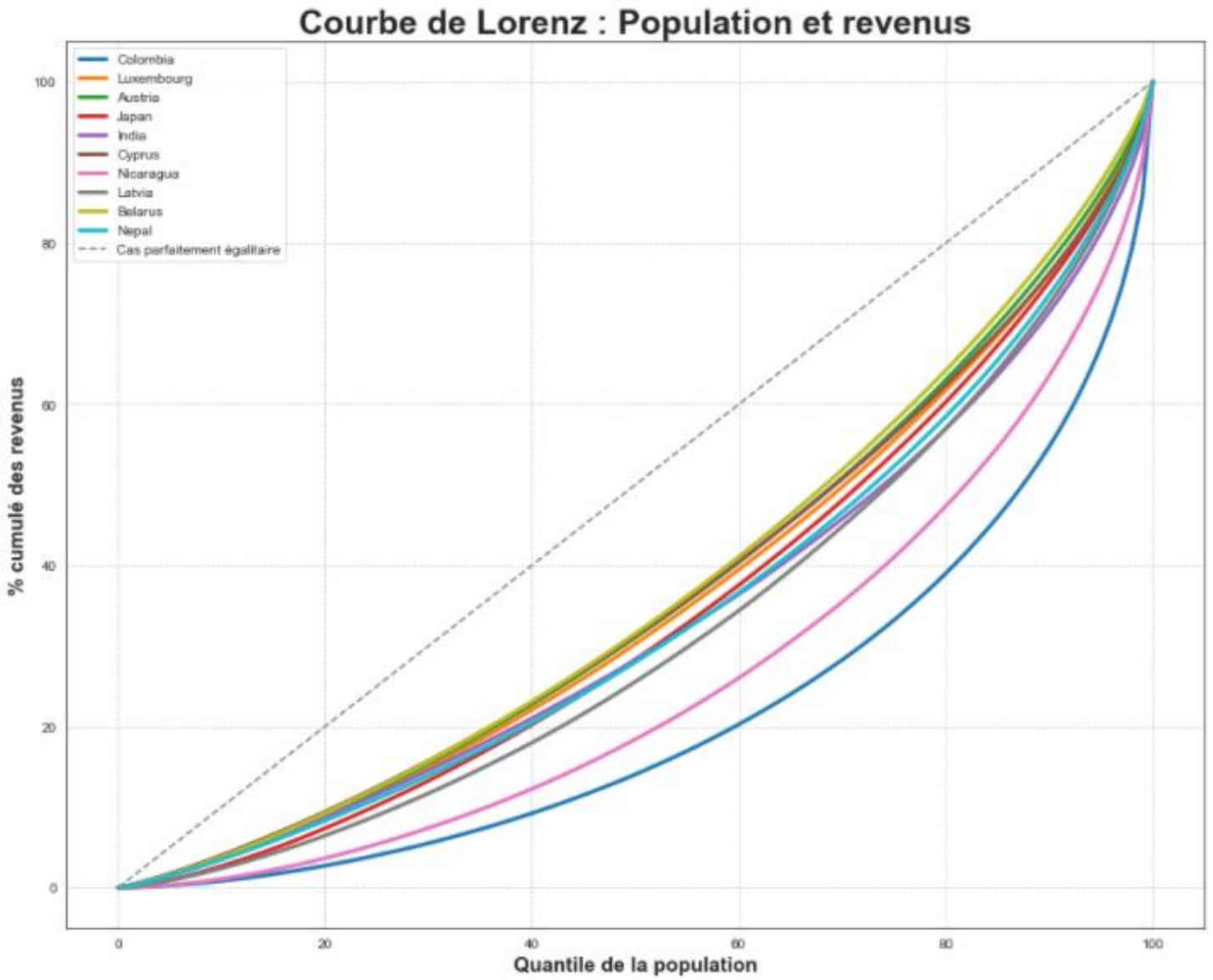
- Nous pouvons constater que la Nicaragua et la Colombie sont les pays les plus inégalitaires car leur courbe est la plus de accentué
- Le Luxembourg est le pays avec les plus aux revenus moyen toutes les classes de revenu



MISSION II

La courbe de Lorenz de chacun des pays:

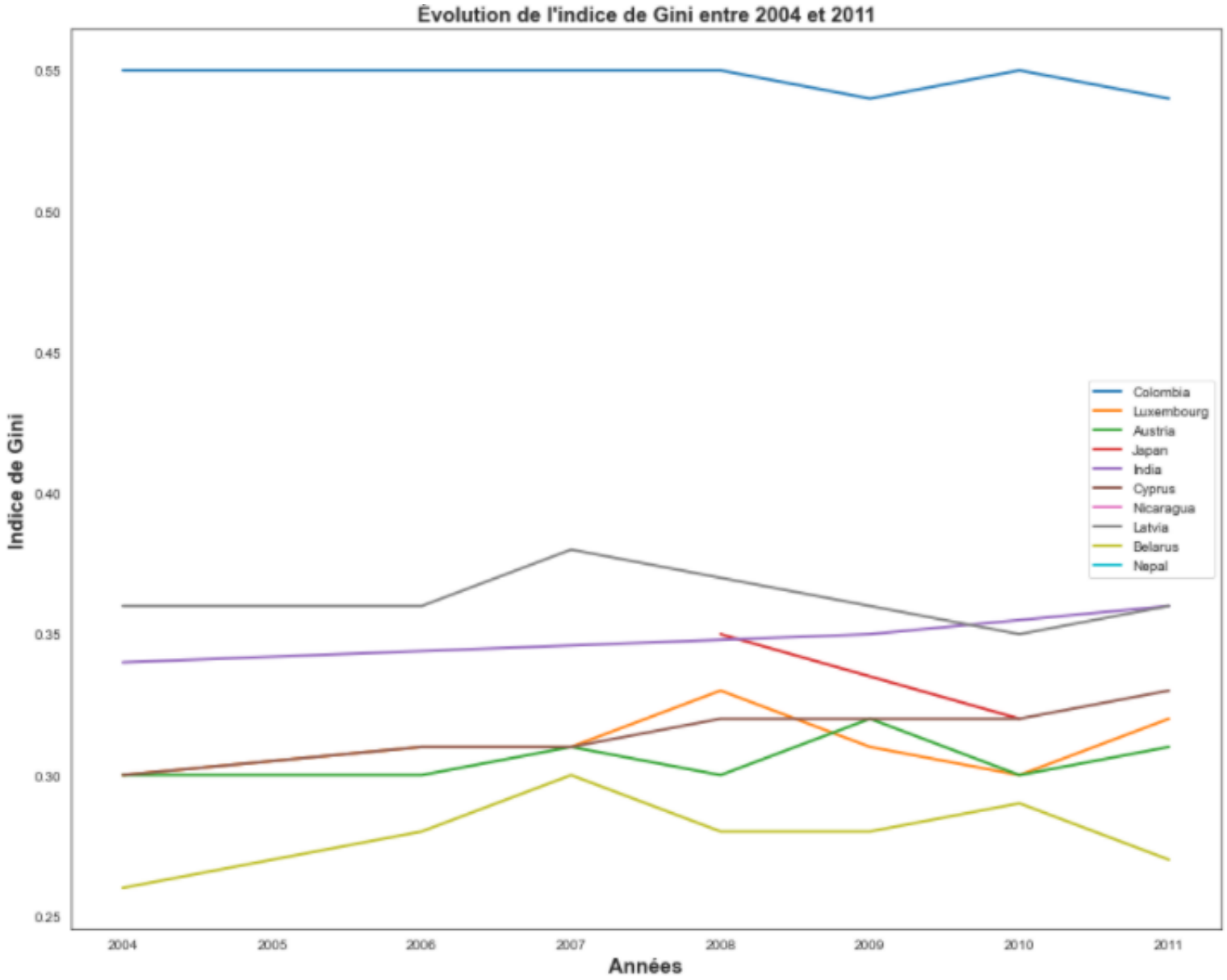
- Le Belarus est le pays avec le moins de inégalité entre les quantiles
- La Colombie est le pays le plus inégalitaire de la liste



MISSION II

L'évolution de l'indice de Gini au fil des ans pour chaque pays :

- La Colombie est le pays avec l’indice de Gini le plus haut, donc avec le pire résultat, suivi de la Latvia



MISSION II

Les 5 pays avec l'indice de Gini le plus élevé :

	Country Code	Country Name	gini	rang
141	ZAF	South Africa	0.63	144.0
94	NAM	Namibia	0.61	143.0
26	BWA	Botswana	0.60	142.0
27	CAF	Central African Republic	0.56	140.5
35	COM	Comoros	0.56	140.5

Les 5 pays ayant l'indice de Gini le plus faible :

	Country Code	Country Name	gini	rang
120	SVN	Slovenia	0.245714	1.0
41	DNK	Denmark	0.261429	2.0
119	SVK	Slovak Republic	0.262857	3.0
39	CZE	Czech Republic	0.265714	4.0
134	UKR	Ukraine	0.268571	5.0

La position de la France :

	Country Code	Country Name	gini	rang
50	FRA	France	0.322857	35.0

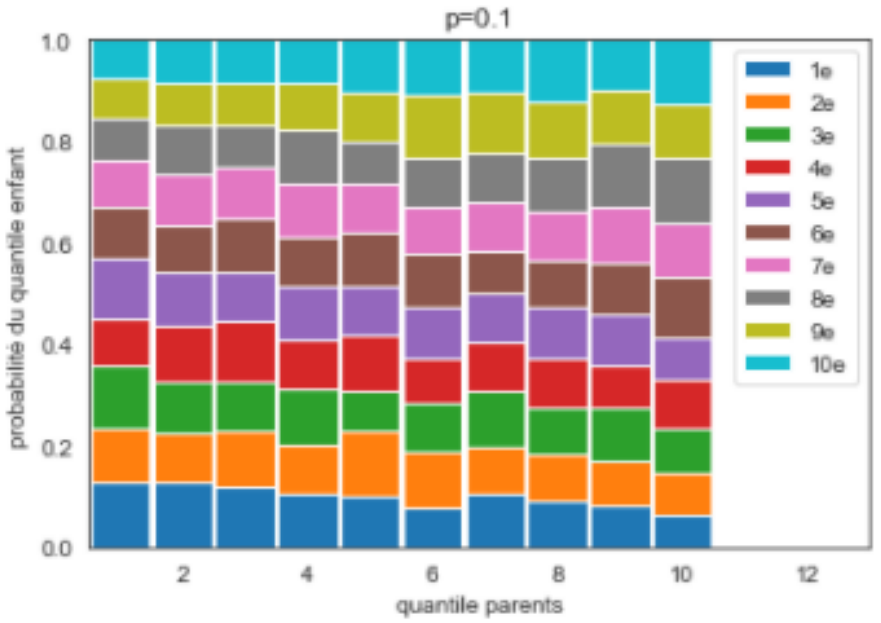
SOMMAIRE

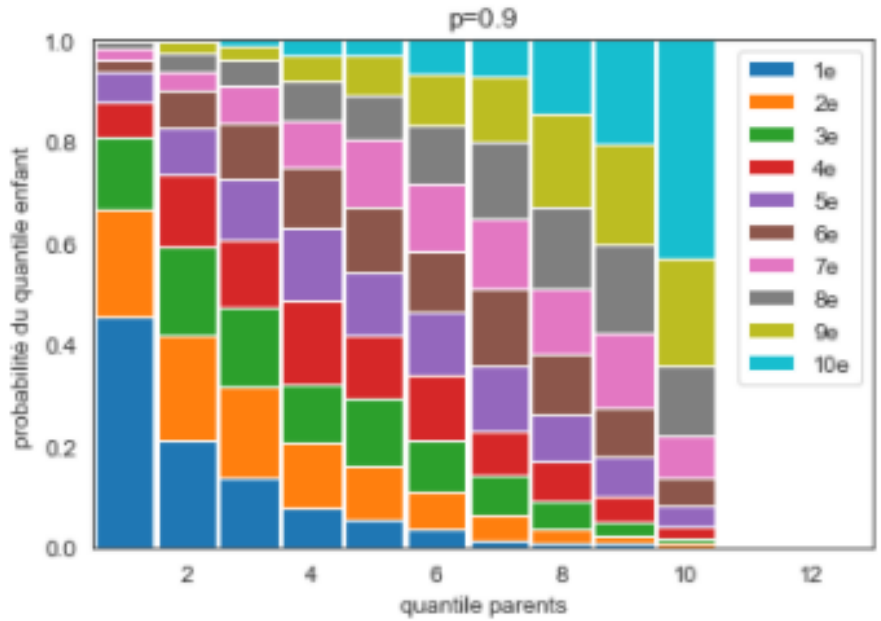
MISSION III

La représentation des distributions conditionnelles avec un coefficient d'élasticité faible et fort :

- Les deux exemples démontrent clairement des situations extrêmes révélatrices de la mobilité des classes de revenu enfants / parents.

- Dans une situation à faible mobilité intergénérationnelle (2éme graphique) les enfants resteront majoritairement dans la classe de revenu de leurs parents.





MISSION III

- Création d'une fonction qui fera le calcul des attributions de quantiles enfants x parent;
- Calcul du *df* avec 500 individus pour chaque pays et chaque quantile :

```
c_parent(liste_pays, liste_quantile, 500, 100)
```

	Country Code	Country Name	c_i_child	Population	income	gini	pj	c_i_parent
0	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2.0
11600	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2.0
23200	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2.0
34800	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2.0
46400	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2.0
...	...	...	...	...	...	...	...	...
5753598	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	1.0
5765198	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	1.0
5776798	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	1.0
5788398	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	1.0
5799998	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	1.0

5800000 rows × 8 columns

Plus de 26h de calcul

SOMMAIRE

MISSION IV

A cause d'un souci de mémoire interne, nous avons utilisé un *dataframe* avec un multiplicateur 50 au lieu de 500 pour la suite des analyses :

	country_code	country	c_i_child	Population	income	gini	pj	income_avg	c_i_parent
0	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2994.829902	1.0
1	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2994.829902	1.0
2	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2994.829902	1.0
3	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2994.829902	1.0
4	ALB	Albania	1	2882740.0	728.89795	0.30	0.815874	2994.829902	1.0
...	...	...	...	...	...	...	...	...	...
579995	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	276.016044	51.0
579996	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	276.016044	51.0
579997	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	276.016044	51.0
579998	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	276.016044	51.0
579999	COD	Congo	100	84068091.0	2243.12260	0.44	0.707703	276.016044	51.0

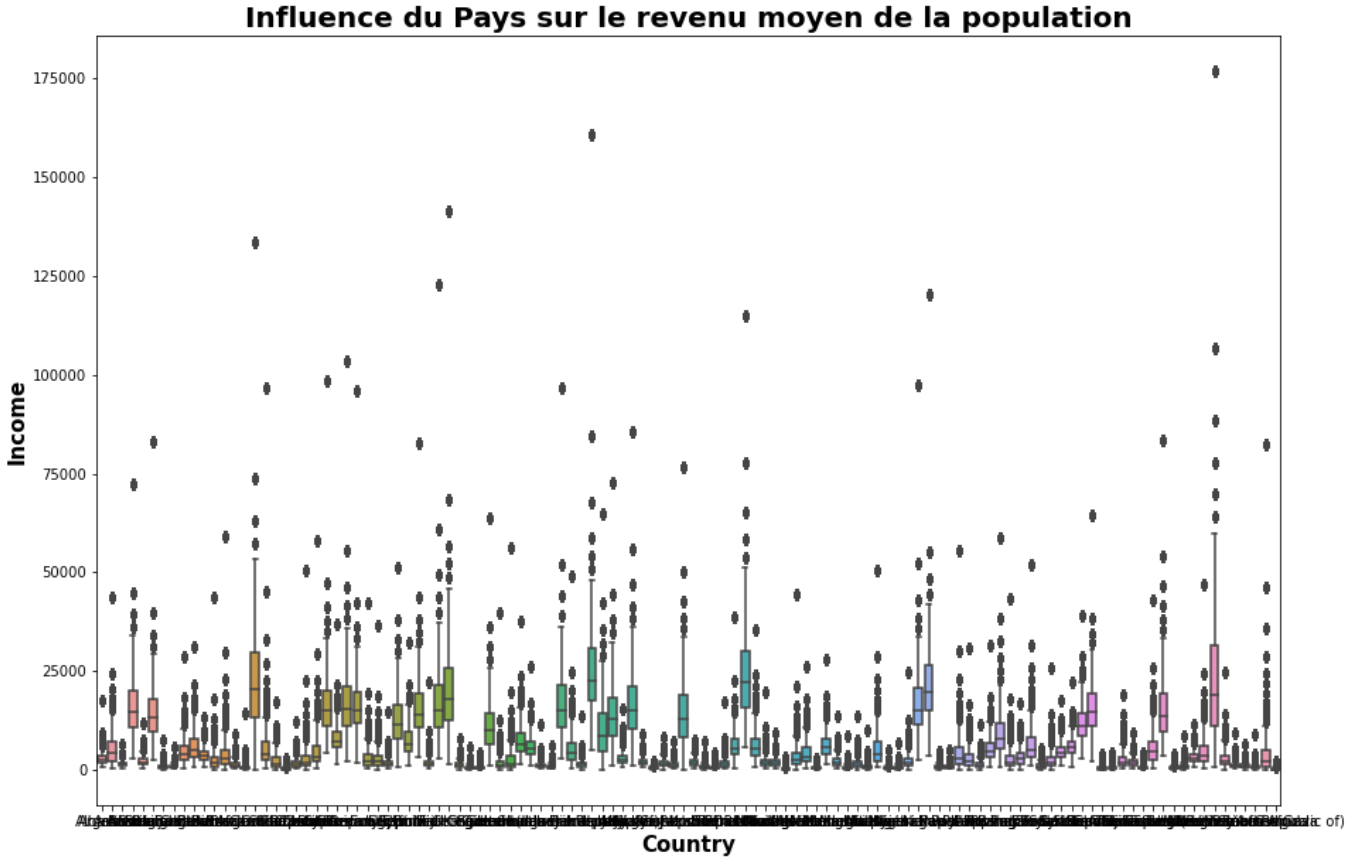
580000 rows × 9 columns

MISSION IV

Income ~ Country

Représentation des distributions

Les pays semblent assez différentes, même si l'ordre de grandeur de ces écarts n'est pas très grand. La question sera de savoir si ces écarts sont significatifs ou pas. **C'est l'ANOVA qui nous permettra de répondre à cette question.**



MISSION IV

Income ~ Country

ANOVA

Anova pour tester l'influence du pays de l'individu sur le revenu moyen des individus :

R^2 = 0.496, on peut en conclure que la variable explicative *Pays* explique près de *50% de la variance du revenu de l'individu*. Le reste, donc la moitié de la variance sur le revenu est expliquée par les autres facteurs non considérés dans ce modèle.

OLS Regression Results

Dep. Variable:	income	R-squared:	0.496
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	4970.
Date:	Wed, 24 Nov 2021	Prob (F-statistic):	0.00
Time:	11:23:21	Log-Likelihood:	-5.9310e+06
No. Observations:	580000	AIC:	1.186e+07
Df Residuals:	579884	BIC:	1.186e+07
Df Model:	115		
Covariance Type:	nonrobust		

	sum_sq	df	F	PR(>F)
country_code	2.551188e+13	115.0	4970.181077	0.0
Residual	2.588293e+13	579884.0	NaN	NaN

La p-valeur de ce test (~0.0) est très petite et largement inférieure à 5%. On rejette donc l'hypothèse H0 selon laquelle $α_{country1}=α_{country2}=α_{country...}=0$. *Le pays a donc bien une influence sur le revenu moyen des individus, comme nous en avions l'intuition en regardant les graphiques*

MISSION IV

ANOVA

log _Income ~ Country

Nouvelle Anova pour tester l'influence du pays de l'individu sur le logarithme du revenu moyen des individus :

En considérant le logarithme du revenu, la variance expliquée est plus concluante, 73% contre 50% précédemment.

OLS Regression Results

Dep. Variable:	ln_income	R-squared:	0.729
Model:	OLS	Adj. R-squared:	0.729
Method:	Least Squares	F-statistic:	1.358e+04
Date:	Wed, 24 Nov 2021	Prob (F-statistic):	0.00
Time:	11:24:44	Log-Likelihood:	-6.3135e+05
No. Observations:	580000	AIC:	1.263e+06
Df Residuals:	579884	BIC:	1.264e+06
Df Model:	115		
Covariance Type:	nonrobust		

MISSION IV

Première Régression Linéaire – modèle I

Revenu moyen du pays de l'individu et l'indice de Gini du pays de l'individu :

Ce modèle n'explique que 50% de la variance, nous ne sommes pas plus performant que l'ANOVA. Il peut s'agir d'un problème de linéarité, car les revenus ont tendance à évoluer de manière exponentielle.

OLS Regression Results

Dep. Variable:	income	R-squared:	0.496
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	2.858e+05
Date:	Wed, 24 Nov 2021	Prob (F-statistic):	0.00
Time:	11:24:45	Log-Likelihood:	-5.9310e+06
No. Observations:	580000	AIC:	1.186e+07
Df Residuals:	579997	BIC:	1.186e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.129e-09	47.203	1.93e-10	1.000	-92.516	92.516
gini	-1.508e-08	113.508	-1.33e-10	1.000	-222.472	222.472
income_avg	1.0000	0.001	714.121	0.000	0.997	1.003

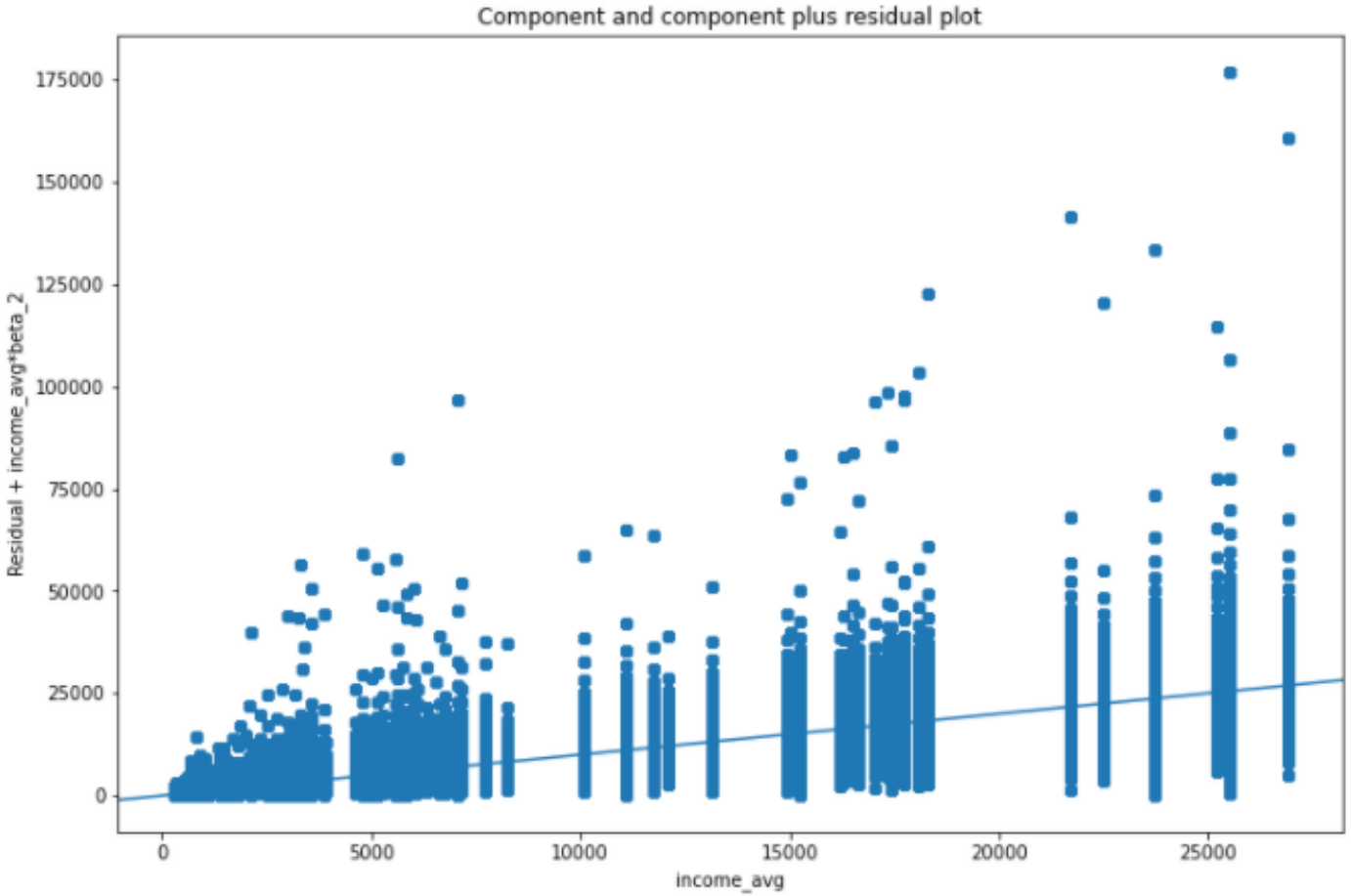
Omnibus:	729958.348	Durbin-Watson:	0.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	210371577.108
Skew:	6.739	Prob(JB):	0.00
Kurtosis:	95.322	Cond. No.	1.25e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+05. This might indicate that there are strong multicollinearity or other numerical problems.

MISSION IV

Première Régression Linéaire :

Modèle sans log



MISSION IV

OLS Regression Results

Dep. Variable:	ln_income	R-squared:	0.728
Model:	OLS	Adj. R-squared:	0.728
Method:	Least Squares	F-statistic:	7.769e+05
Date:	Wed, 24 Nov 2021	Prob (F-statistic):	0.00
Time:	11:24:48	Log-Likelihood:	-6.3247e+05
No. Observations:	580000	AIC:	1.265e+06
Df Residuals:	579997	BIC:	1.265e+06
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4802	0.009	51.957	0.000	0.462	0.498
gini	-1.7298	0.012	-144.577	0.000	-1.753	-1.706
ln_income_avg	0.9884	0.001	1158.290	0.000	0.987	0.990

Omnibus:	37774.824	Durbin-Watson:	0.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	177030.251
Skew:	-0.100	Prob(JB):	0.00
Kurtosis:	5.699	Cond. No.	122.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Seconde Régression Linéaire – modèle II :

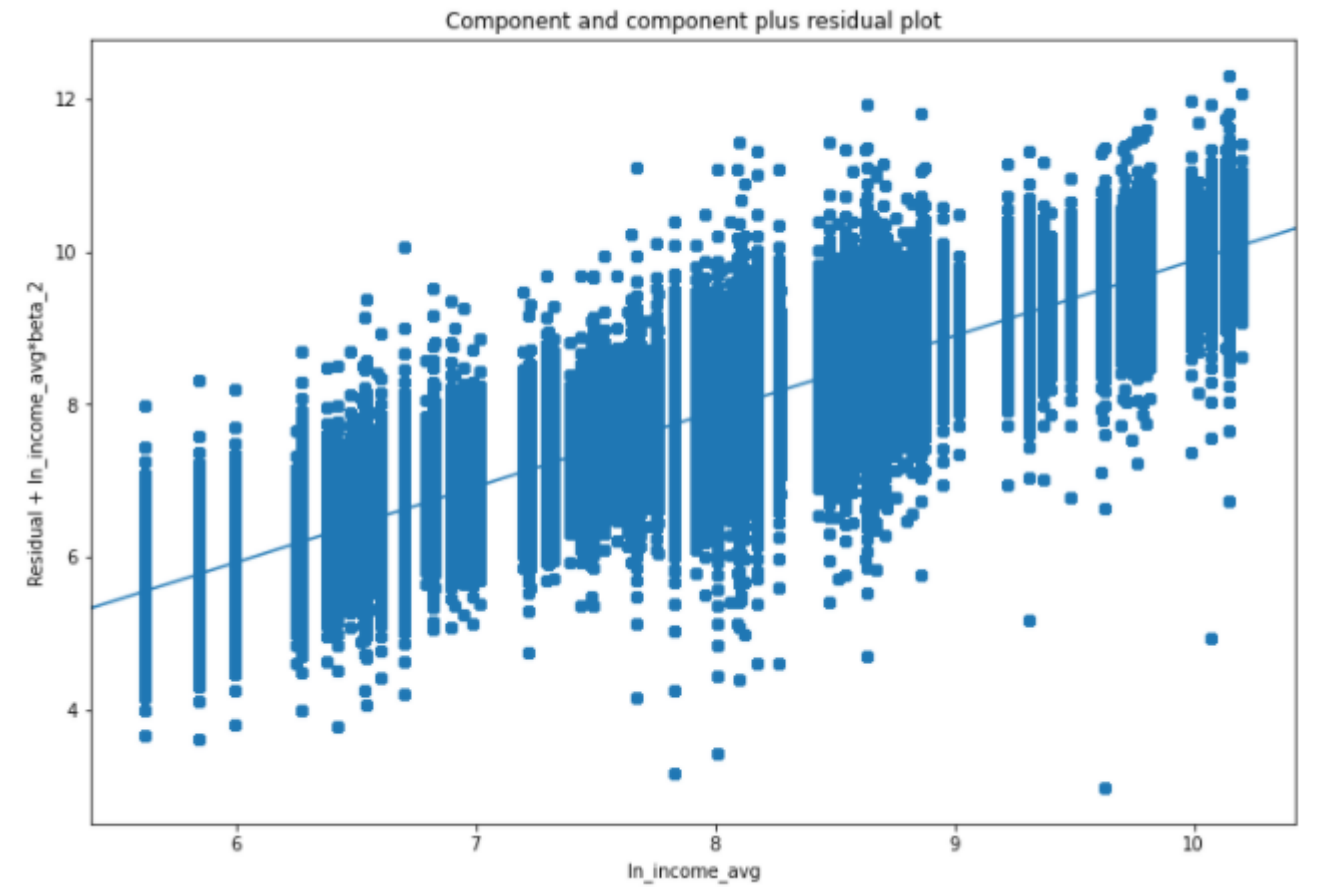
Logarithme du Revenu moyen du pays de l'individu et l'indice de Gini du pays de l'individu

- Comme lors de l'étude de l'ANOVA, la version logarithmique est plus performante.
- En prenant le logarithme du revenu et le logarithme du revenu moyen, la performance est plus optimale.
- Les p-valeurs sont d'ailleurs très faibles.
- Nous retrouvons le même niveau de performance que dans l'ANOVA du logarithme du revenu: l**e modèle 2 peut expliquer 73% de la variance;**
- le restant peut s'expliquer sur d'autres critères non pris en compte jusqu'ici, à savoir les classes de revenu des parents, ou encore des critères sociaux professionnels, etc…

MISSION IV

Seconde Régression Linéaire

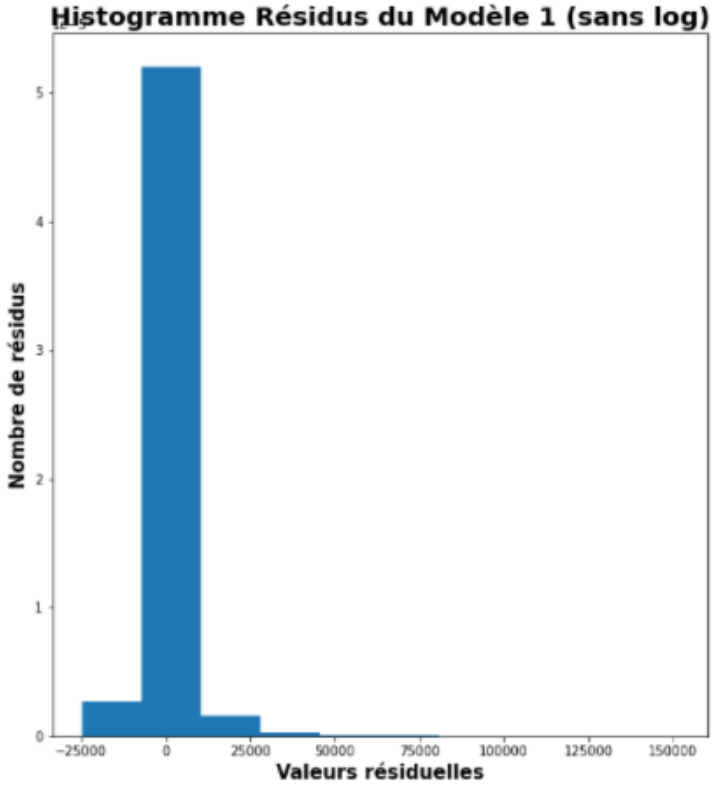
- Modèle avec log
- La seconde approche visuelle de la régression linéaire est également plus représentatrice.

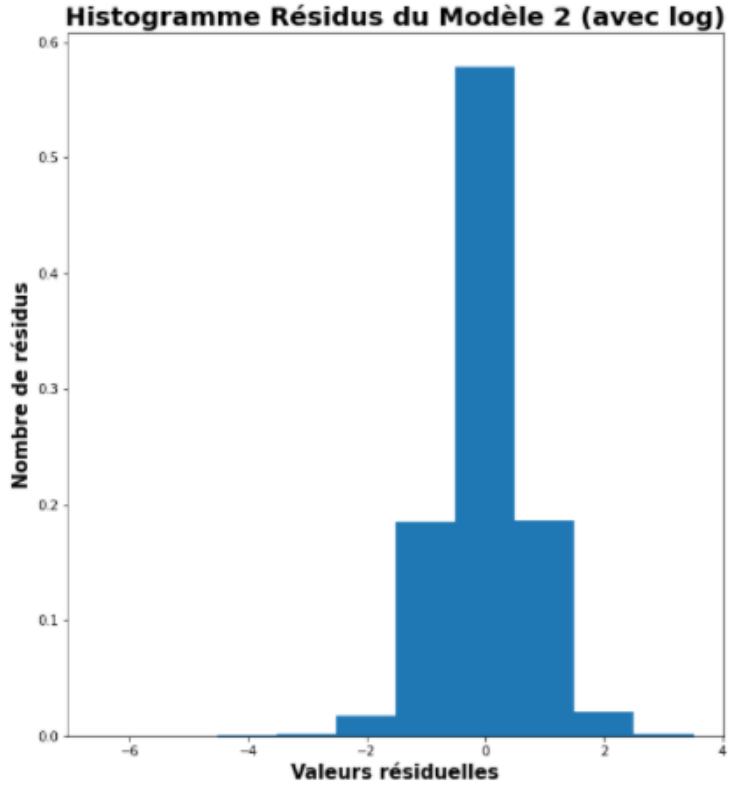


MISSION IV

Diagnostic de la régression linéaire des deux modèles (sans et avec logarithme):

L'inférence dans la régression linéaire multiple repose sur l'hypothèse de normalité des erreurs.





Le modèle 1 (sans logarithme) renvoie une distribution des résidus qui semble suivre plus difficilement une loi normale, tandis que le second modèle propose une distribution plus gaussienne, plus homogène

MISSION IV

Amélioration du modèle le plus performant (avec log) en incluant la classe de revenu des parents

- Le nouveau modèle est plus performant que son précédent (modele2), soit 77.8% au lieu de 72.8%
- Le coefficient de détermination R^2 est plus élevé, l'influence du revenu des parents sur le revenu de l'enfant est substantielle, mis en évidence par le coefficient d'élasticité.
- Pour rappel, la corrélation entre le revenu de l'individu et le revenu de ses parents est mesurée par ce coefficient, le coefficient d'élasticité (mesure de la mobilité intergénérationnelle du revenu).

OLS Regression Results

Dep. Variable:	ln_income	R-squared:	0.733
Model:	OLS	Adj. R-squared:	0.733
Method:	Least Squares	F-statistic:	5.305e+05
Date:	Mon, 06 Dec 2021	Prob (F-statistic):	0.00
Time:	10:39:47	Log-Likelihood:	-6.2739e+05
No. Observations:	580000	AIC:	1.255e+06
Df Residuals:	579996	BIC:	1.255e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3163	0.009	34.001	0.000	0.298	0.335
gini	-1.7009	0.012	-143.377	0.000	-1.724	-1.678
c_i_parent	0.0034	3.35e-05	101.272	0.000	0.003	0.003
ln_income_avg	0.9872	0.001	1166.856	0.000	0.986	0.989

Omnibus:	35626.000	Durbin-Watson:	0.022
Prob(Omnibus):	0.000	Jarque-Bera (JB):	164524.583
Skew:	-0.025	Prob(JB):	0.00
Kurtosis:	5.609	Cond. No.	833.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MISSION IV

Décomposition de la variance totale expliquée

- A partir de notre dernier modèle, nous pouvons conclure à un rôle clé du pays de naissance : **69% de la variance expliquée.**
- La classe de revenu des parents **explique 5.5% de la variance.**
- L'indice de Gini compte pour seulement **1% de la variance.**
- Pour le reste, les résidus sont **24%** et peuvent être expliqués par d'autres critères non traités ici, comme par exemple l'âge, le sexe, le niveau de qualification, etc…

SOMMAIRE

CONCLUSION

- Une question se pose concernant le coefficient de régression associé à l'indice de Gini, est-il possible d'affirmer que le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu'il n'en défavorise ?

Plus l'indice de Gini est élevé, plus les inégalités sont fortes, des écarts dans les salaires sont importants avec des revenus/individus plus bas dans ces pays. L'exemple du précédent modèle est révélateur, on peut comprendre ce lien par le coefficient négatif au sein du modèle.

- Pour finir sur le dernier modèle, la décomposition de variance totale est expliquée par différents éléments.
- **Concluons en disant que la prédiction du revenu potentiel d'une personne peut s'appréhender selon la logique suivante :**
  - ✓ Pour le dernier modèle inclut la classe de revenu des parents la loi normale est généralement suivie.
  - ✓ La normalisation des données apporte un gain sur le coefficient de détermination, $R^2 = 0.778$ (contre 0.728 sur le modèle précédent)
  - ✓ Il reste 24% de variance non expliquée par le modèle, à ce stade on peux supposer que ce restant correspond à d'autres critères non traités ici, comme par exemple l'âge, le sexe, le niveau de qualification, etc…