PRÉDICTION DE REVENUS

Création d’un modèle permettant de déterminer le revenu potentiel d’une personne

SOMMAIRE

CONTEXTE ET ENJEUX

ciblage de nouveaux prospects

CONTEXTE
01

CONTEXTE ET ENJEUX

Le projet ici nous permet de nous mettre dans la peau d'un employé de banque.

Etablissement bancaire présent dans de nombreux pays à travers le monde souhaite cibler de nouveaux clients potentiels.

QUI ? Les jeunes en âge d'ouvrir leur premier compte bancaire

COMMENT ? A partir de plusieurs variables dont le revenus de leurs parents et le pays d'origine

MISSION Créer un modèle de prédiction

DÉMARCHE ET DONNÉES

présentation des données utilisées

DONNÉES
02

DONNÉES ET DÉMARCHES

Nous disposons de données téléchargées principalement via le site de la World Income Distribution et de la Banque mondiale.

Distribution en centiles / percentiles

WORLD INCOME DISTRIBUTION

année de référence
2008

\+

BANQUE MONDIALE

6,1 milliards d'individus
soit 91% personnes couvertes par l'étude

VARIABLES

- Country : pays
- Quantile : numéro du quantile
- Income : revenu
- Gdp PPP : PIB en parité de pouvoir d'achat
- Gini : niveau d'inégalité
- Coefficient d'élasticité : mesure la mobilité
- Population

ANALYSE

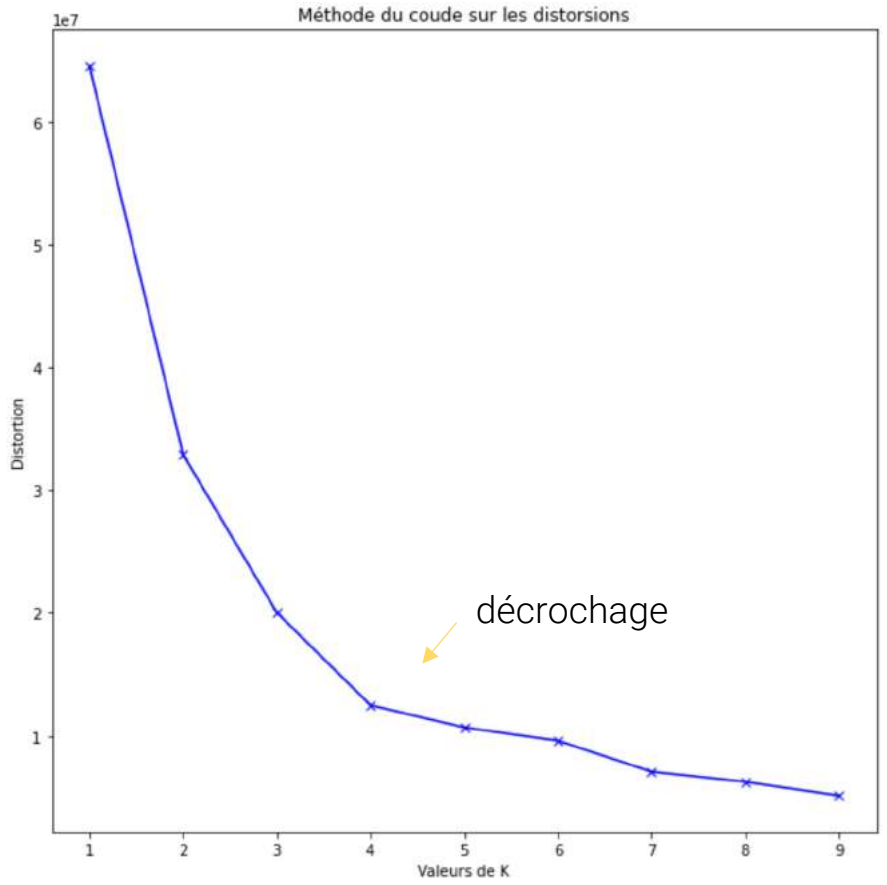
analyse des variables

ANALYSE
03

ANALYSE DES DONNÉES

Nous réalisons une analyse de la diversité des revenus sur une base de 5 pays représentatifs.

Méthode du coude déterminer le nombre de clusters idéal



clustering kmeans

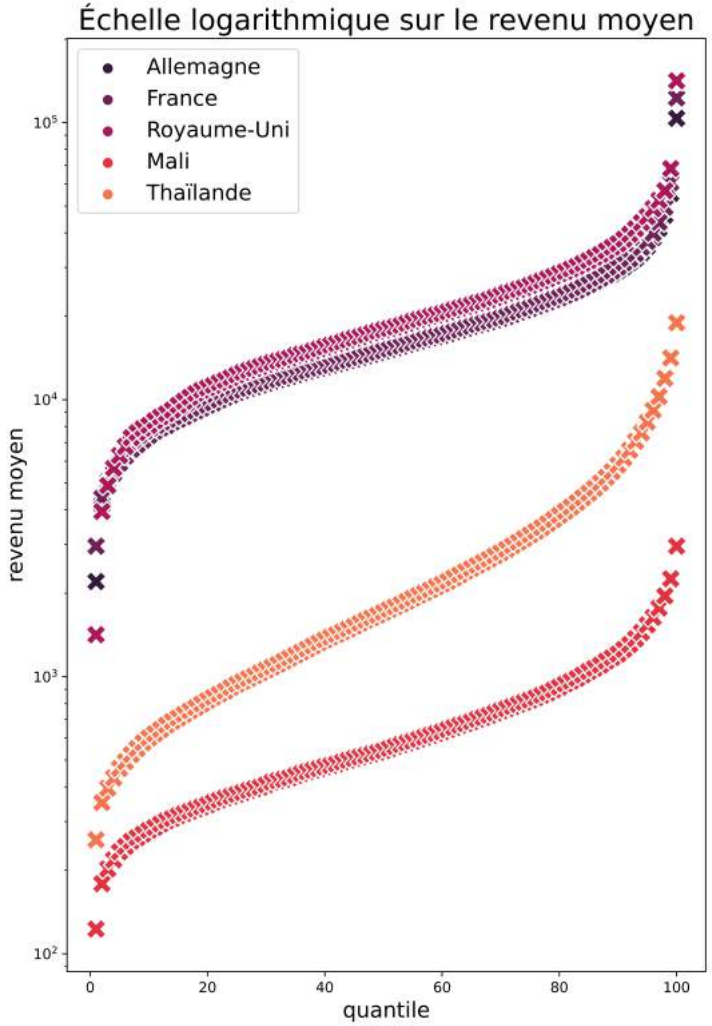
pays représentatifs : pays les plus proches des centroïdes

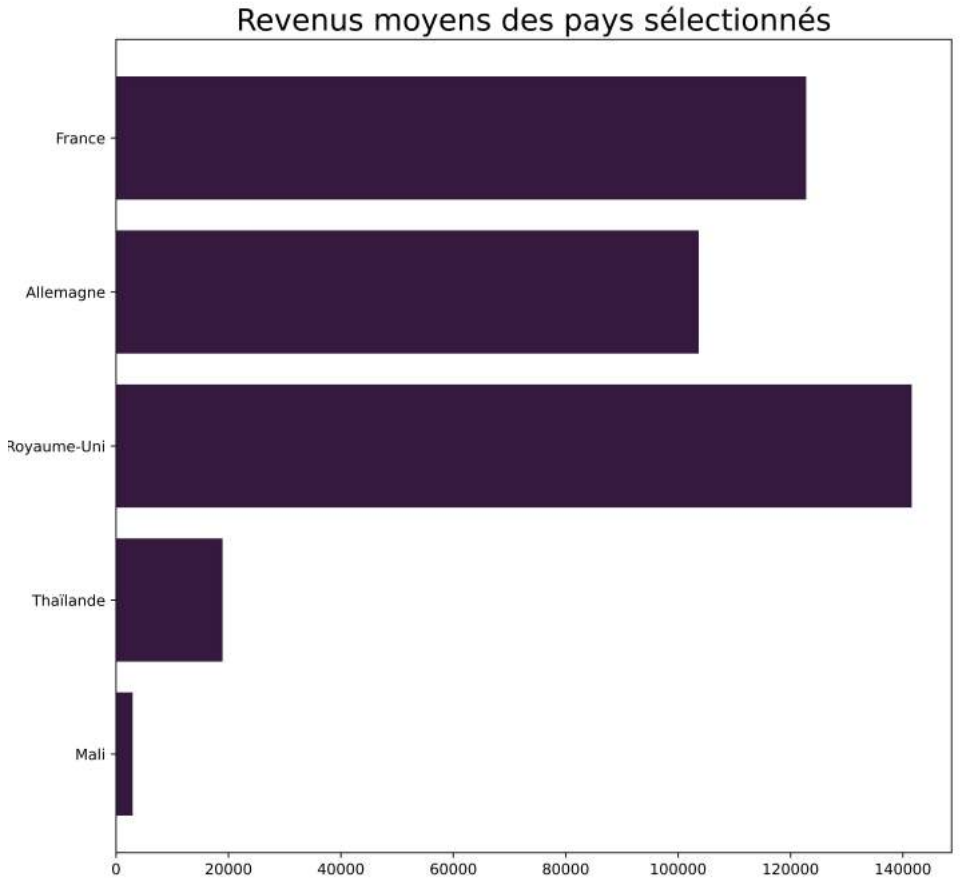
Royaume-Uni Allemagne Thaïlande Mali
+ France

ANALYSE
03

ANALYSE DES DONNÉES

Nous réalisons une analyse de la diversité des revenus sur une base de 5 pays représentatifs.



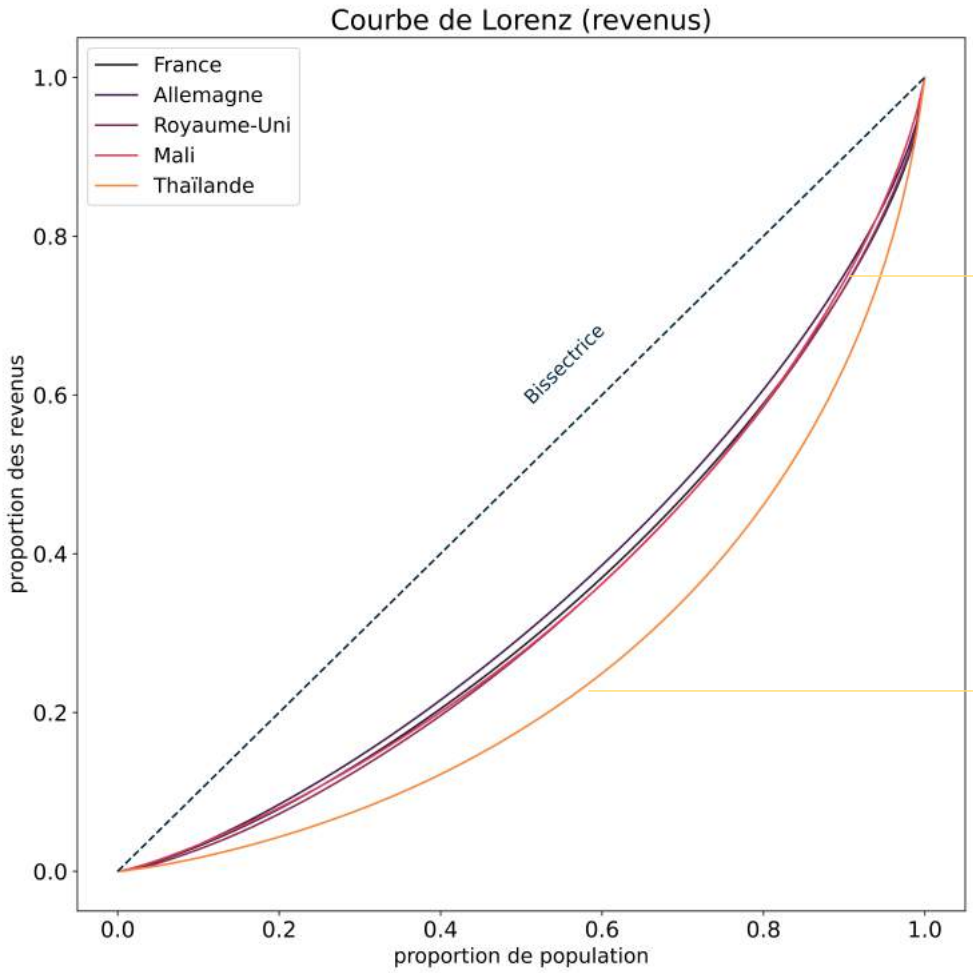


ANALYSE
03

ANALYSE DES DONNÉES

Nous réalisons une analyse de la diversité des revenus sur une base de 5 pays représentatifs.

Courbe de Lorenz distribution des revenus des pays



La *bissectrice* représente l'égalité parfaite

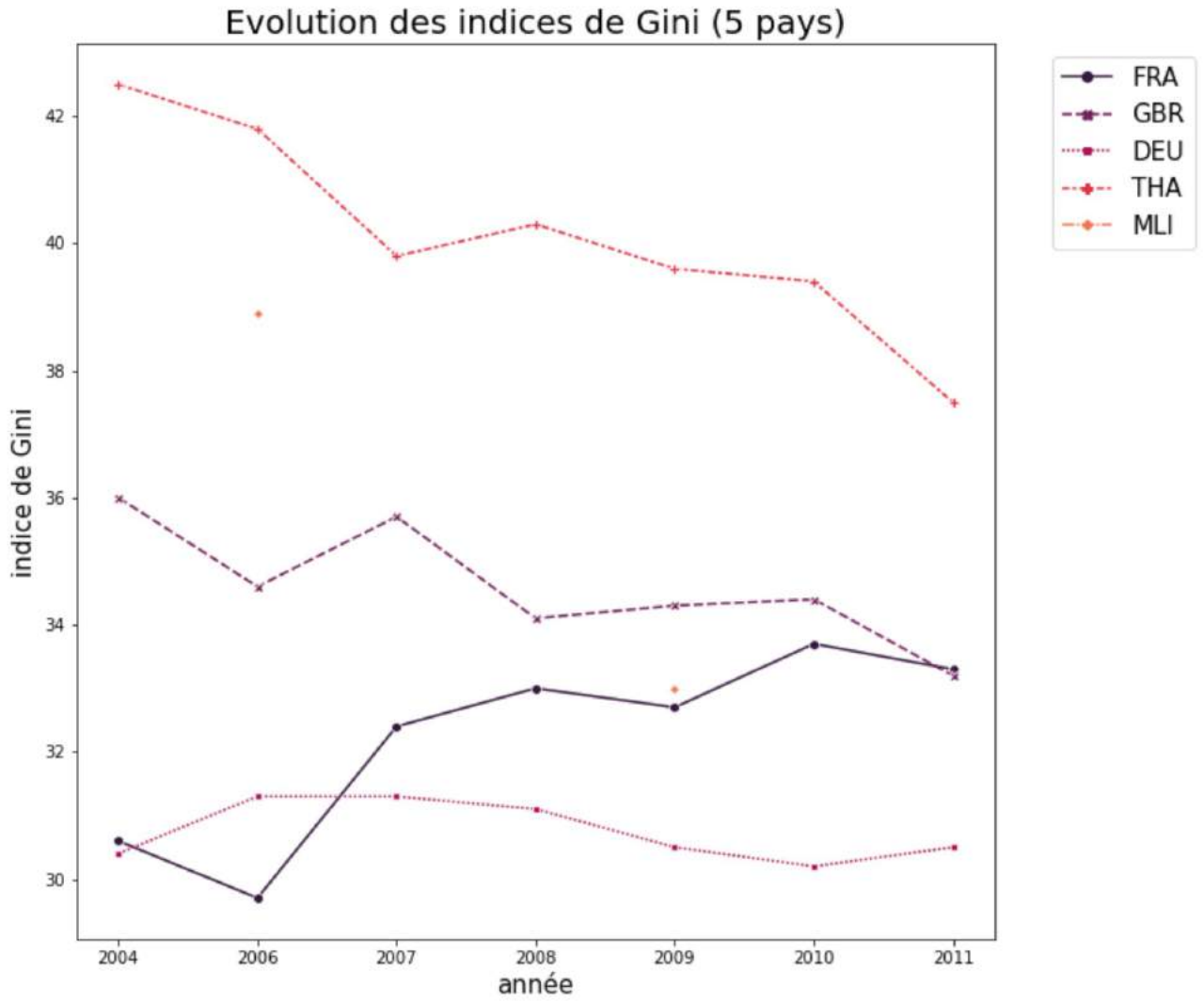
80% de la population **allemande** détient 70% des revenus du pays

60% de la population **thaïlandaise** détient 20% des revenus du pays

ANALYSE
03

ANALYSE DES DONNÉES

Nous réalisons ici une analyse des indices de Gini pour nos 5 pays. Cela nous permet de nous rendre compte des niveaux d'inégalité.



moyenne des indices de Gini : 0,38

TOP 5 des pays ayant un indice Gini **faible**

	Country Name	2008
220	Slovenia	23.7
56	Denmark	25.2
219	Slovak Republic	26.0
52	Czech Republic	26.3
246	Ukraine	26.6

TOP 5 des pays ayant un indice Gini **élevé**

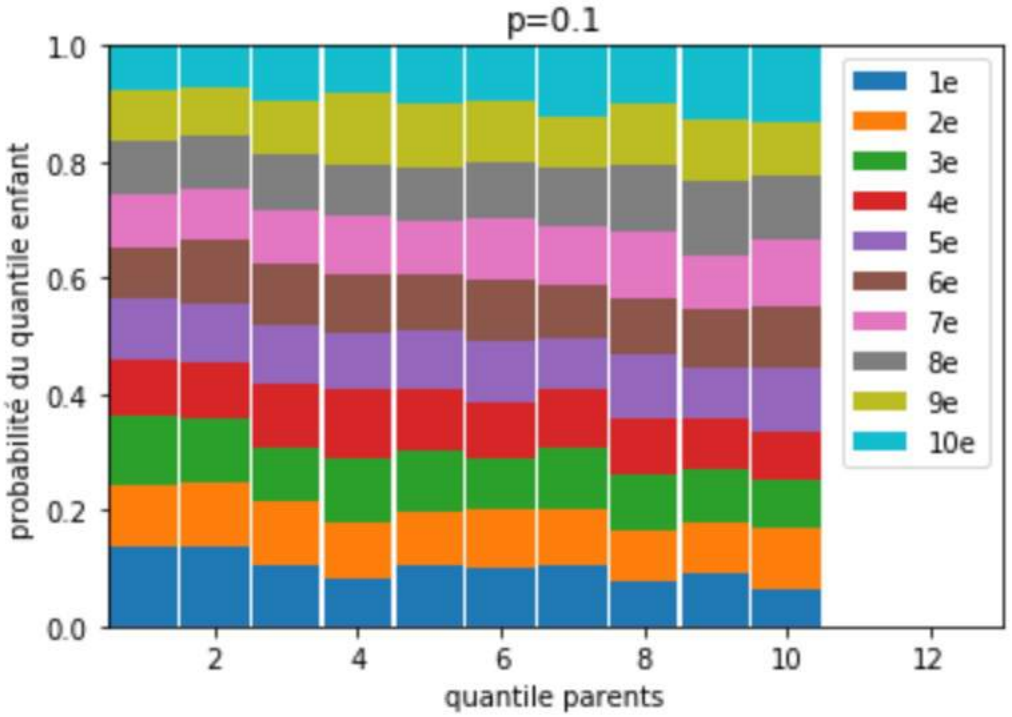
	Country Name	2008
261	South Africa	63.0
32	Central African Republic	56.2
95	Honduras	55.5
43	Colombia	55.3
27	Brazil	54.0

* Rang de la France : **75ème pays** le *plus égalitaire*

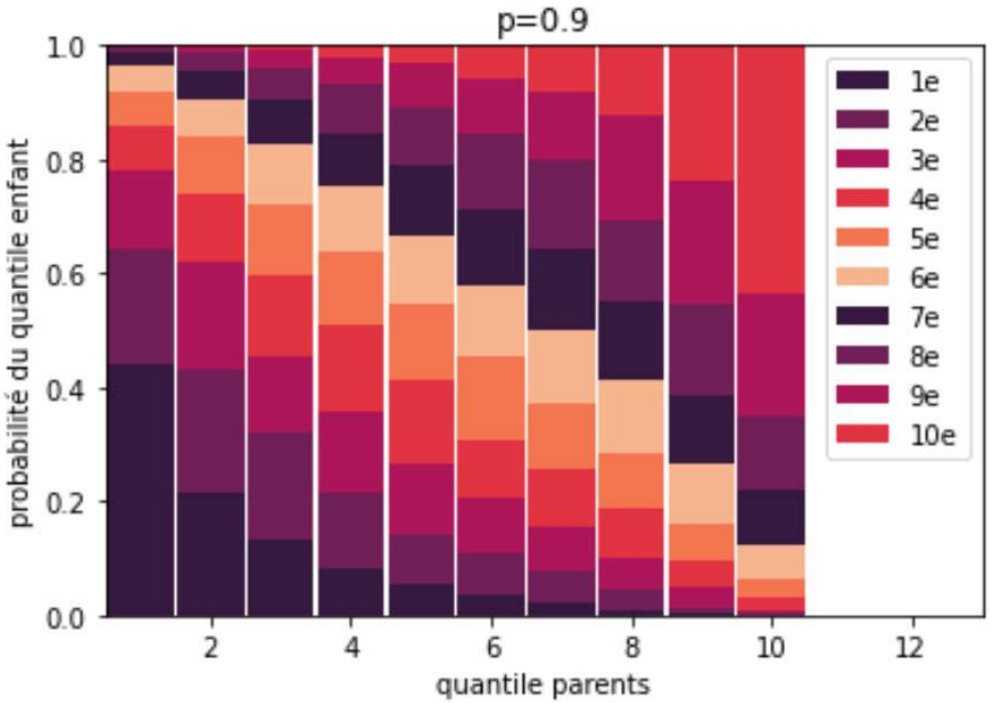
ANALYSE
03

ANALYSE DES DONNÉES

Nous avons généré la classe de revenu des parents de façon aléatoire. Cela a été permis grâce au coefficient d'élasticité et la classe de revenu des enfants.



Forte mobilité *sociale*



Faible mobilité *sociale*

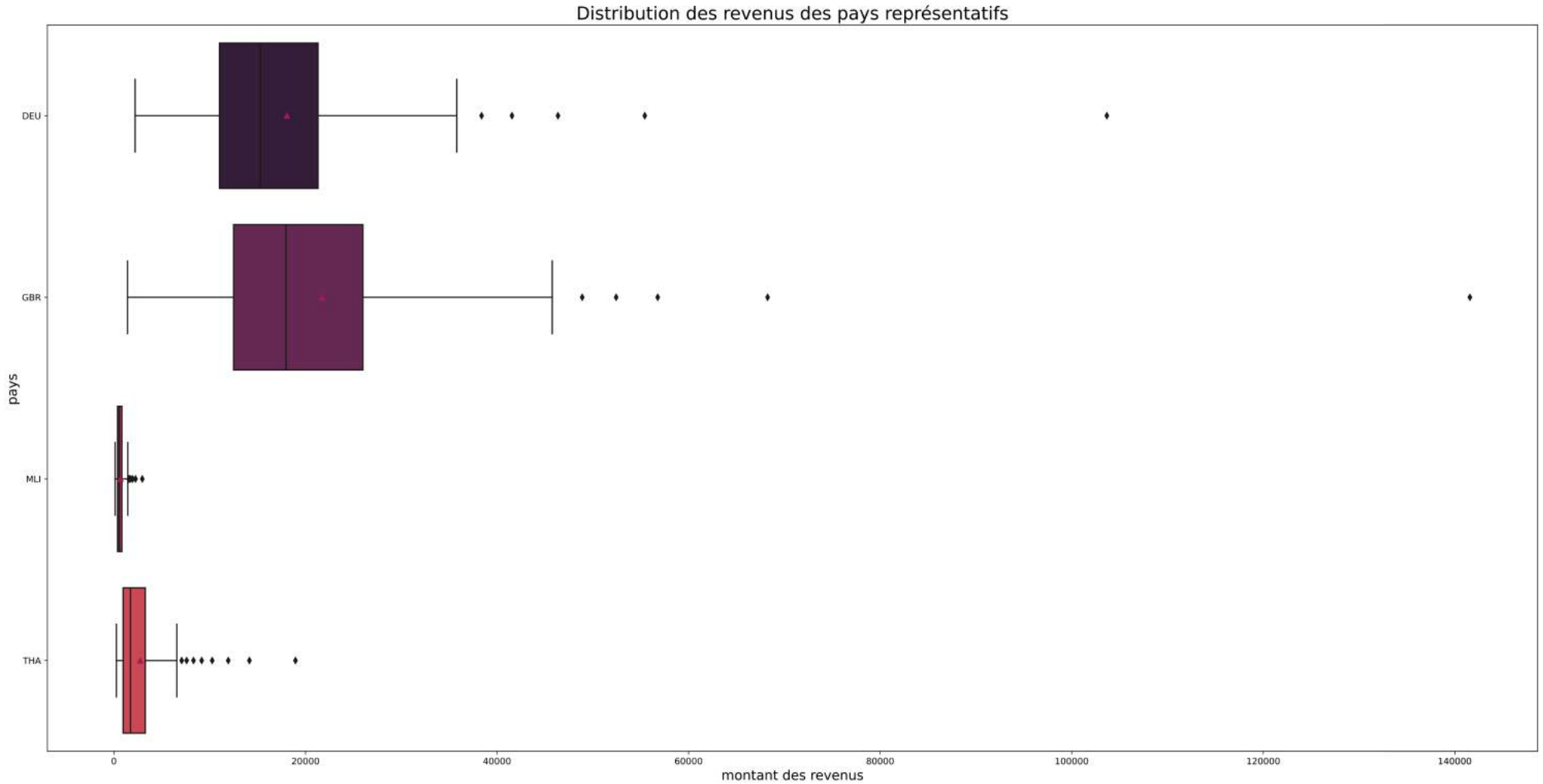
MODÈLE

création d’un modèle de prédiction

MODÈLE
04

MODÈLE STATISTIQUE

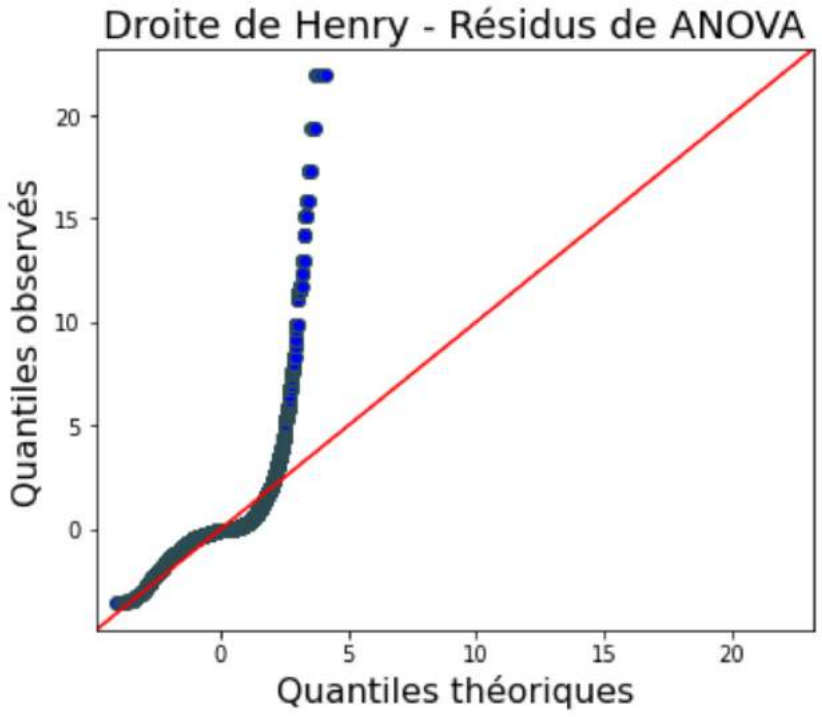
Nous avons appliqué une ANOVA sur nos données. Avant cela, nous devions justifier ce choix.



MODÈLE
04

MODÈLE STATISTIQUE

Pour appliquer une ANOVA, certaines conditions doivent être validées : la normalité des résidus et l'égalité des variances.



Les résidus *ne semblent pas suivre* une loi normale

Vérification des conditions

Test Kolmogorov-Smirnov

p-value = 0

Les résidus **ne suivent pas** une loi normale (à 5%)

Test de Breuch-Pagan

P-value : 2,3^{e-295}

les variances ne sont pas homogènes (à 5%)

Conditions non validées

MODÈLE
04

MODÈLE STATISTIQUE

Les conditions n'étant pas validées, nous réalisons une alternative à l'ANOVA pour savoir si le pays d'origine explique le revenu de l'individu.

Conditions de l'ANOVA non validées

Réalisation d'un test **non-paramétrique**

Test de Kruskal-Wallis

p-value = 0

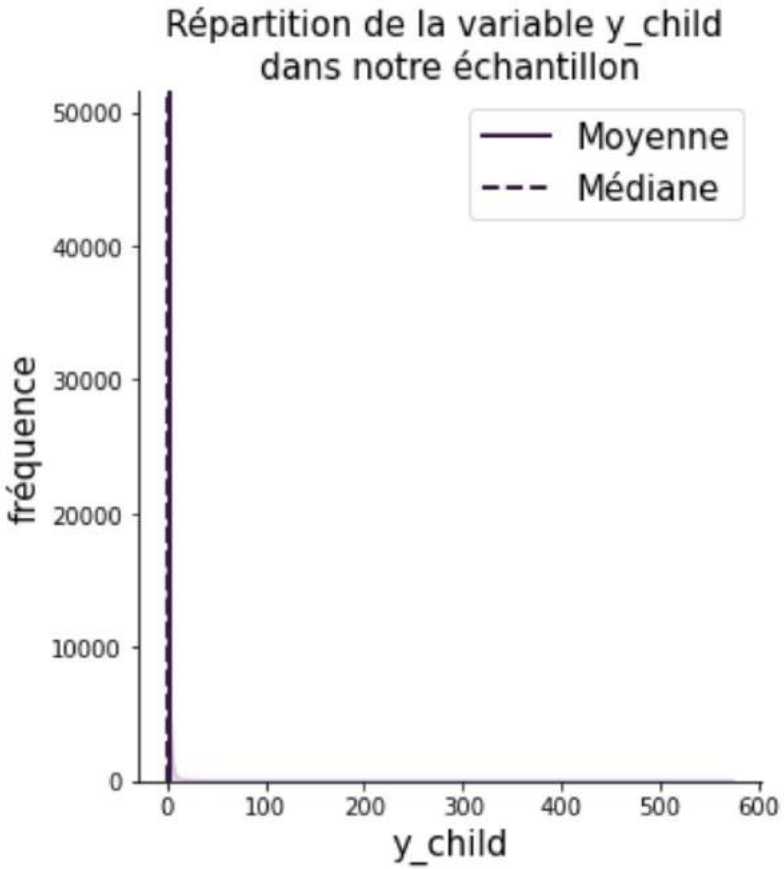
les revenus sont bien différents entre les pays

D'après la régression linéaire, le pays d'origine explique environ 47,8% le *revenu de l'individu*.

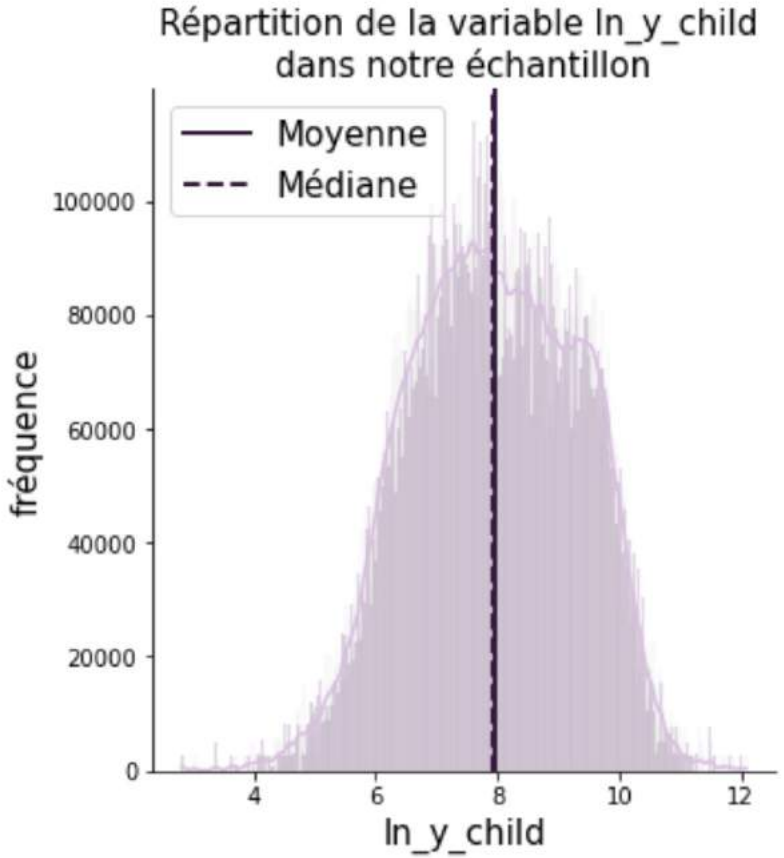
MODÈLE
04

MODÈLE STATISTIQUE

Nous chercherons à expliquer le revenu des individus en fonction de plusieurs variables explicatives.



Sans le passage aux logarithmes



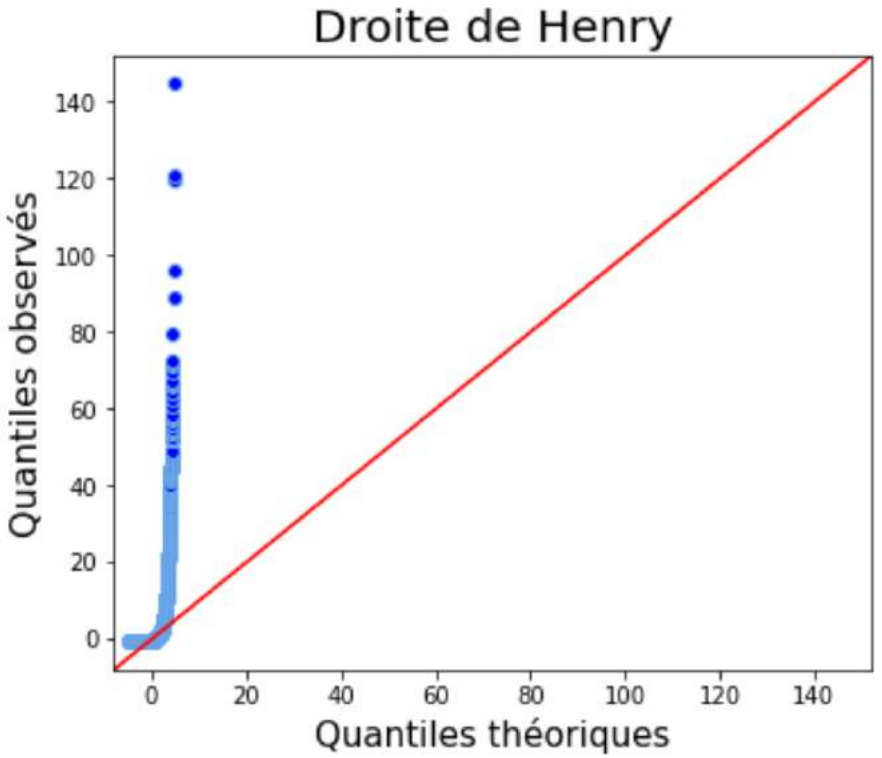
Passage aux logarithmes

Avantages du passage aux logarithmes :
Robustesse
Améliorer l'ajustement du modèle
Résidus symétriques

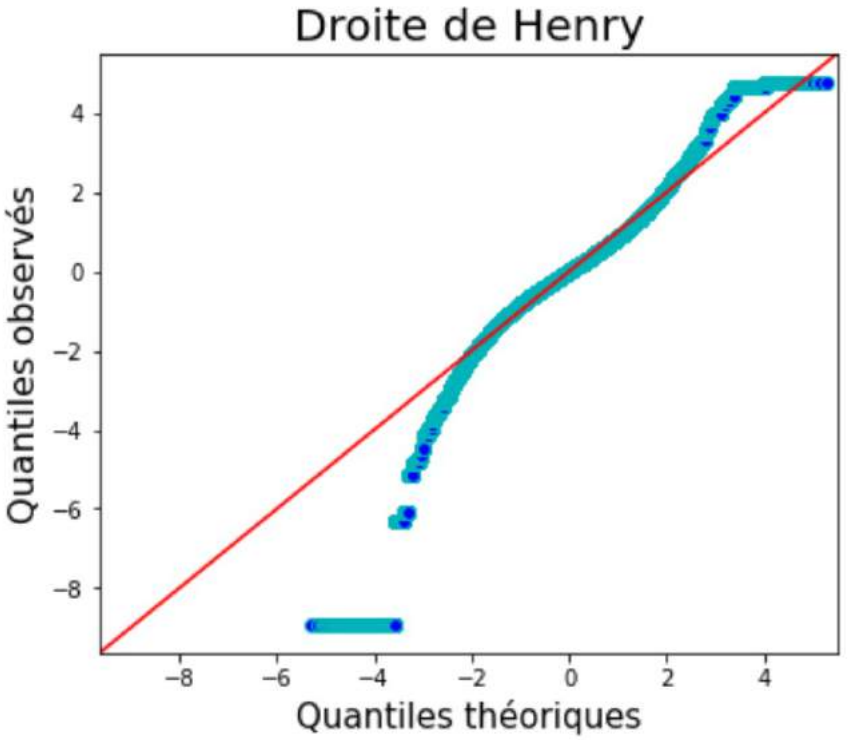
MODÈLE
04

MODÈLE STATISTIQUE

Pour accepter une régression linéaire, les résidus doivent suivre une loi normale. Le passage aux logarithmes permet de s'y rapprocher.



Sans le passage aux logarithmes



Passage aux logarithmes

La distribution est plus proche d'une **distribution gaussienne**

MODÈLE
04

MODÈLE STATISTIQUE

Nous chercherons à expliquer le revenu des individus en fonction de plusieurs variables explicatives.

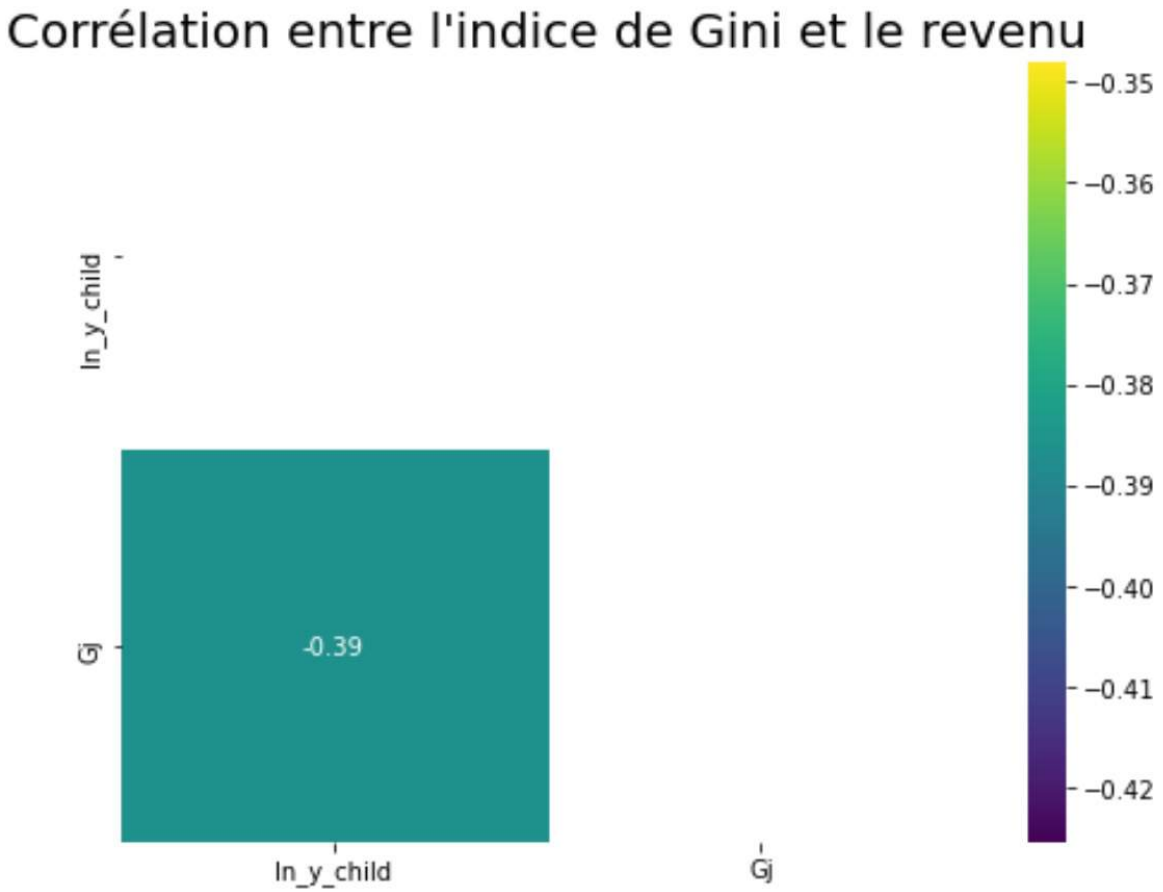
Récapitulatif des régressions linéaires appliquées

Le **revenu moyen du pays** et l'**indice de Gini** expliquent à 62% le *revenu de l'individu*.
L'ajout de la classe des parents ne change pas notre coefficient de détermination. L'ajout d'autres variables ne changent pas non plus notre coefficient de détermination.

MODÈLE
04

MODÈLE STATISTIQUE

Nous chercherons à expliquer le revenu des individus en fonction de plusieurs variables explicatives.



A noter - on observe une corrélation négative : plus l'indice de Gini est faible, plus le revenu à tendance à être élevé

BILAN
04

BILAN

Nous avons pu expliquer quelles étaient les variables permettant d'expliquer le revenu d'un individu.

Pour accepter une régression linéaire, et ainsi observer des relations entre nos variables, les résidus doivent suivre une loi normale. Le passage aux logarithmes nous a permis d'avoir des modèles fiables puisque ce passage permet de se rapprocher d'*une distribution normale*.

A la suite de nos modèles, on peut donc conclure que le revenu d'un individu dépend du revenu moyen de son pays d'origine et de l'indice de Gini. Cela a été expliqué à hauteur de 62% par l'analyse de la variance. Le reste, à savoir 38%, est expliqué par des *facteurs extérieurs non considérés* par le modèle comme la chance et les efforts, par exemple. D'autres facteurs peuvent expliquer le revenu d'un individu et peuvent améliorer notre modèle (*niveau d'études, sexe, classe sociale etc.*).

De plus, plus l'indice de Gini du pays d'origine *est faible*, plus le revenu *est élevé* et inversement.

MERCI

Anissa MANSOUR

parcours Data Analyst (2020/2021)