Effectuez une Prédiction de Revenus

Construction et interprétation d'un modèle de régression linéaire (pas de prédictions)

Glossaire

Indice de Gini (Gj) : indicateur synthétique d'inégalité de revenus

Quantile : classe de revenu

GDPPP : Gross Domestic Product based on Purchasing Power : PIB basé sur le pouvoir d'achat

Income (pour un quantile donné) (en $PPP) : revenu moyen des personnes appartenant à la classe de revenu correspondante à ce quantile

Classe de revenu : quantile

Coef d'élasticité (IGEIncome - pj) : mesure la mobilité intergénérationnelle du revenu

ACP : Analyse en Composantes Principales
objectif : rechercher la projection pour laquelle l'inertie des points est maximale.

Bon partitionnement : homogénéité intra-classe (les individus se ressemblent) et hétérogénéité inter-classe (les groupes diffèrent)

Composantes principales ou facteurs : nouvelles variables formées par combinaison linéaire des anciennes variables (variables synthétiques)

KMeans: algorithme de clustering

PPP : Purchasing Power Parities (= parités de pouvoir d'achat (PPA))

Recherche Jeunes à Hauts Revenus

Mission : créer un modèle permettant de déterminer le revenu potentiel d'un jeune (modèle valable pour la plupart des pays du monde)

Mode opératoire : régression linéaire avec revenu des parents, revenu moyen du pays d'origine, indice de Gini calculé sur les revenus des habitants du pays

Sommaire

01
Mission
Analyse et description des données

02
Mission
Description de la diversité des pays

03
Mission
Détermination de la classe de revenu de des parents

04
Mission
Interprétation du revenu des individus

Mission 1

Brève description des données

Importation et lecture du fichier csv

Fichier importé

Source: World Income Distribution datée de 2008
https://openclassrooms.com/fr/paths/65/projects/148/assignment

Contenu: distributions de revenus des populations

Nom Fichier: data-project7.csv

Dimensions: 11 599 lignes, 6 colonnes

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.0
1	ALB	2008	2	100	916.66235	7297.0

Caractéristiques

5 variables numériques: 2 type float, 3 type int

1 variable qualitative: type objet

Valeurs Nulls: 200 pour gdpppp

Valeurs Dupliquées: Aucune

Quantile: classe de revenu (1 à 100)

Fichier de la World Income Distribution

Fichier importé

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.0
1	ALB	2008	2	100	916.66235	7297.0

Caractéristiques

country: code ISO 3166-1 du pays

year_survey: Année de l'enquête

income: Revenu moyen de l'ind de la classe

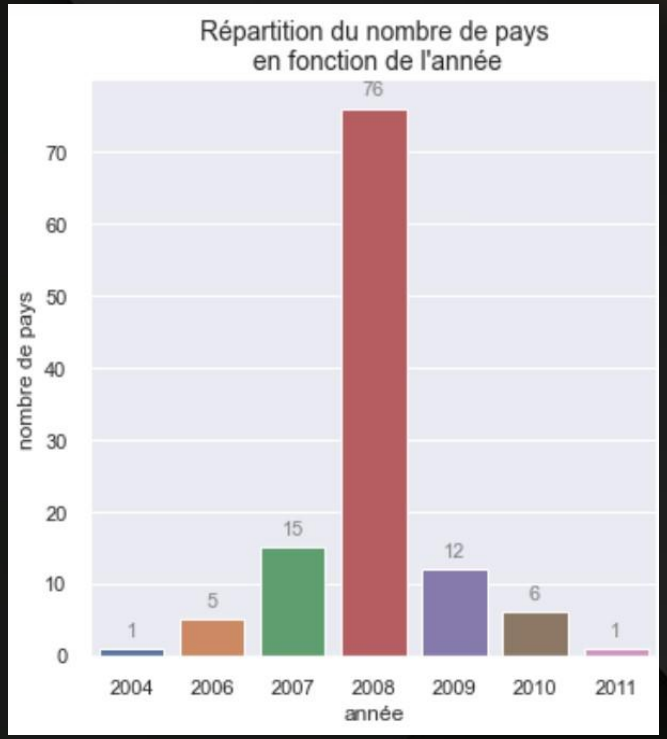
gdpppp: PIB basé sur le pouvoir d'achat

quantile: Classe de revenu (1 à 100)

Centile: Sépare la population en 100 parties égales (meilleure distribution)

Période Analyse

- 2004 - 2011 : période de l'analyse
- 7 années proposées
- 2005 : année manquante
- 2008 : majorité de pays renseignés
- 116 pays au total



Un Fichier avec des Anomalies

200 Valeurs GDP PPP Manquantes

Pays: Kosovo, Palestine (World Bank and Gaza)

Méthode: Implémentation des valeurs GDP PPP null avec les valeurs trouvées sur internet

https://www.indexmundi.com/g/g.aspx?c=kv&v=65

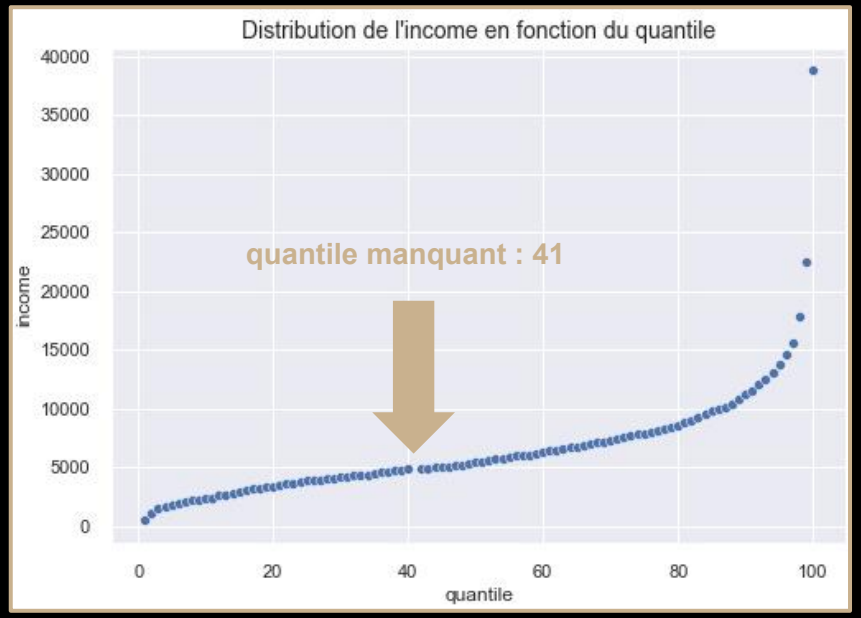
Un Pays avec un Centile Manquant

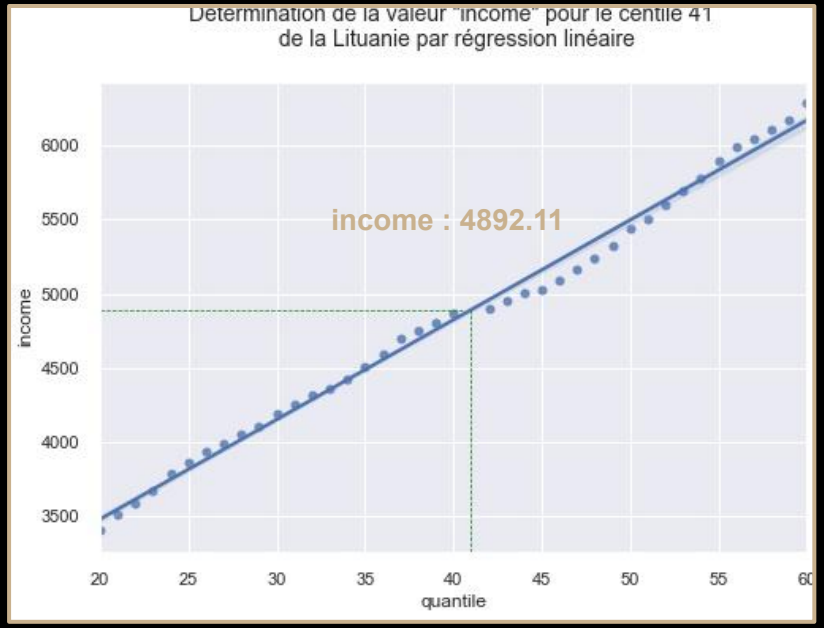
Pays: Lituanie

Méthode: Implémentation du quantile manquant par :
- régression linéaire
- interpolation
- calcul de l'écart moyen

Remarque: 116 pays et seulement 11 599 quantiles

Le centile implémenté par régression linéaire





Précision du modèle : 99 %

Méthodes d'Implémentation

rég. linéaire

income centile 41 :
4892.11

interpolate

income centile 41 :
4882.14

écart moyenne

entre q40 et q42
income centile 41 :
4882.14

Importation des Données Population

Distributions de Revenus de 2008

Population: Année 2008

Source: https://data.worldbank.org/indicator/SP.POP.TOTL

Caractéristiques: 266 lignes, 2 colonnes

	Country Code	2008
0	ABW	101362.0
1	AFE	491173160.0

Anomalies sur les Données

Code INX: Pas référencé dans la liste des codes

Méthode: Suppression de cette donnée

Taïwan: Pas d'habitants renseignés

Méthode: Ajout manuel

Code WLD: Nombre mondial d'habitants

Précaution: Ne pas confondre avec un pays

L' Analyse Porte sur ....

6,2 Mds

habitants

92 %

de la population mondiale

Finalisation Des Fichiers d' Analyse

Nom de Pays avec Code Iso 3

Fichier code: https://satvasolutions.com/

Méthode: Fusion avec le dataset d'analyse pour récupérer le nom du pays correspondant au code ISO 3

Nom de Pays avec Code Iso 3

Format: csv

Fichier: revenus_pays_code_pop.csv
(11600 lignes, 8 colonnes)

	country_code	nb_habitants_2008	year	quantile	nb_quantiles	income	gdpppp	country
0	ALB	2947314.0	2008	1	100	728.89795	7297.0	Albania
1	ALB	2947314.0	2008	2	100	916.66235	7297.0	Albania

Fichier: revenus_pays.csv
(11600 lignes, 7 colonnes)

	country_code	year	quantile	nb_quantiles	income	gdpppp	country
0	ALB	2008	1	100	728.89795	7297.0	Albania
1	ALB	2008	2	100	916.66235	7297.0	Albania

Mission 2

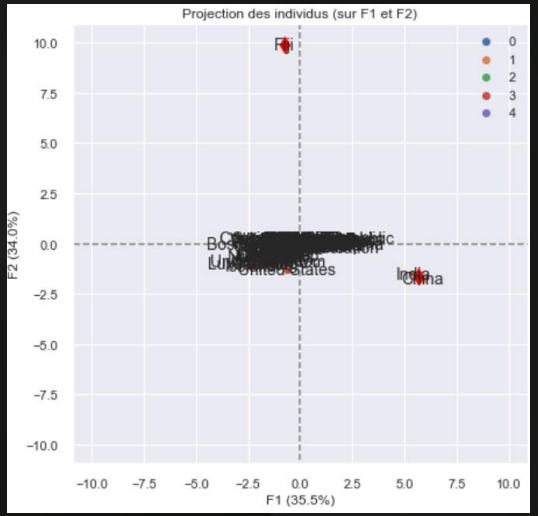
Etude des inégalités entre pays

Choix des Pays Étudiés

Clustering (5 groupes) des pays par revenu médian (KMeans)

ACP sur Les 5 Groupes KMeans

ACP



Présence d'Outliers

Pays 1 : Chine

Pays 2 : Fijis

Pays 3 : Inde

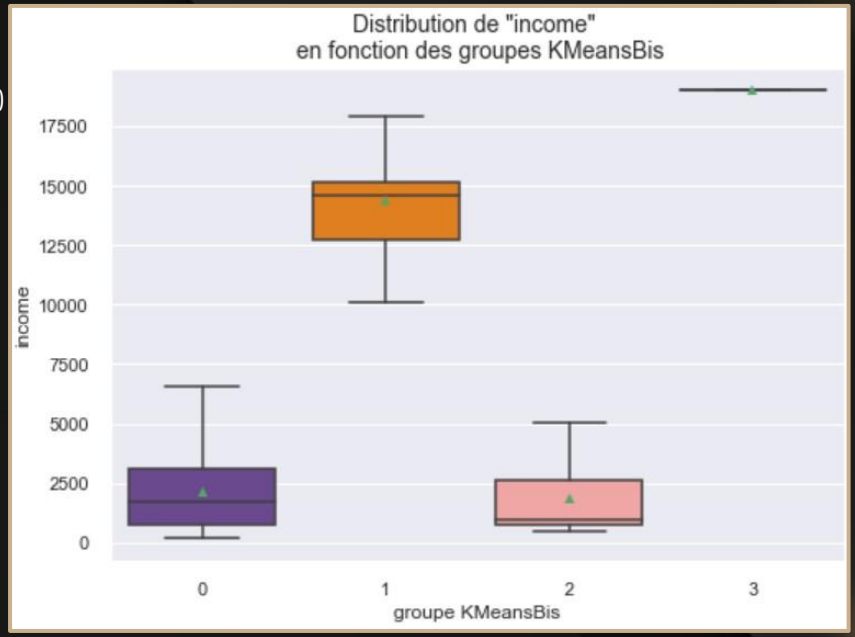
Méthode: Retrait des outliers pour une étude plus fine

Clustering (4 Groupes)

KMeans sans les outliers

Groupes Distincts ?

- Test de Levène d'égalité des variances (gpe 0 et 1)
- Hypothèses :
  - H0 : les variances sont égales
  - H1 : les variances ne sont pas égales
  - seuil alpha = 0.05
- p_value : 9.28e-05
- H0 rejetée au seuil alpha, les variances des 2 groupes sont inégales
- Les groupes sont distincts



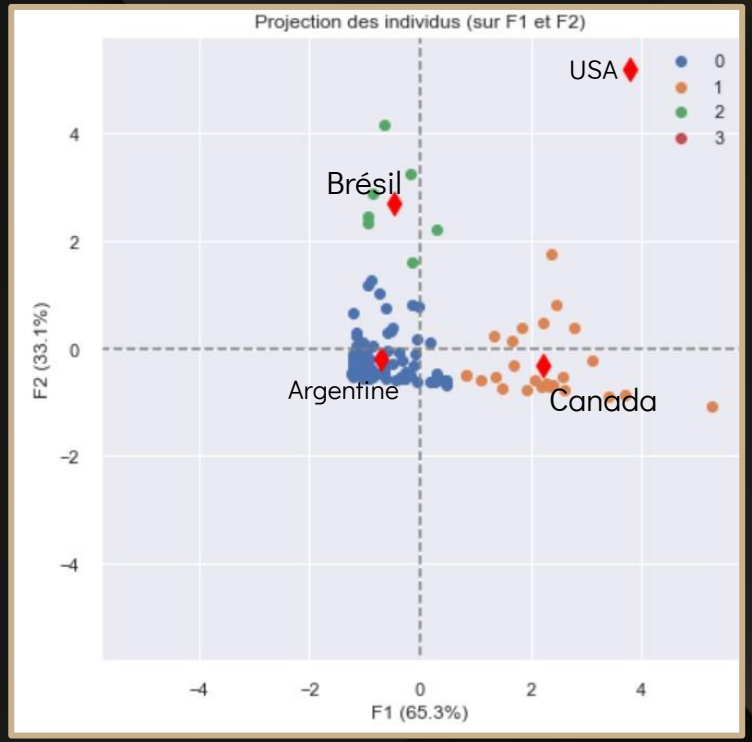
Détails des Groupes

- Groupe 0 : moyenne basse
- Groupe 1 : income fort, gdp ppp fort
- Groupe 2 : income le plus faible
- Groupe 3 : nb_habitants_2008 le plus fort, income le plus fort, gdp ppp le plus fort

Groupe 0	Argentine Burkina Faso...
Groupe 1	Europe, Canada...
Groupe 2	Brésil, Russie...
Groupe 3	USA

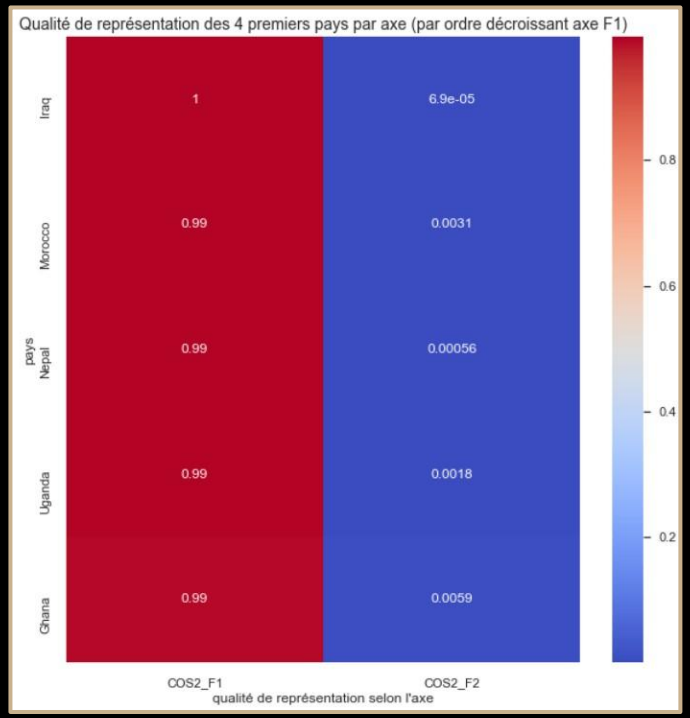
ACP sur Les Groupes

- Présence d'un outlier
  - USA

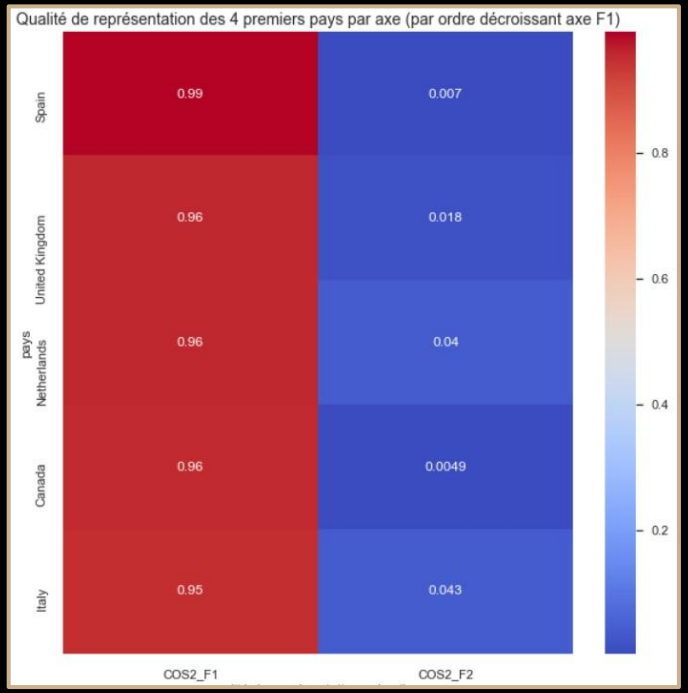


Qualité de Représentation des 4 Premiers Pays par Groupe

Groupe 0

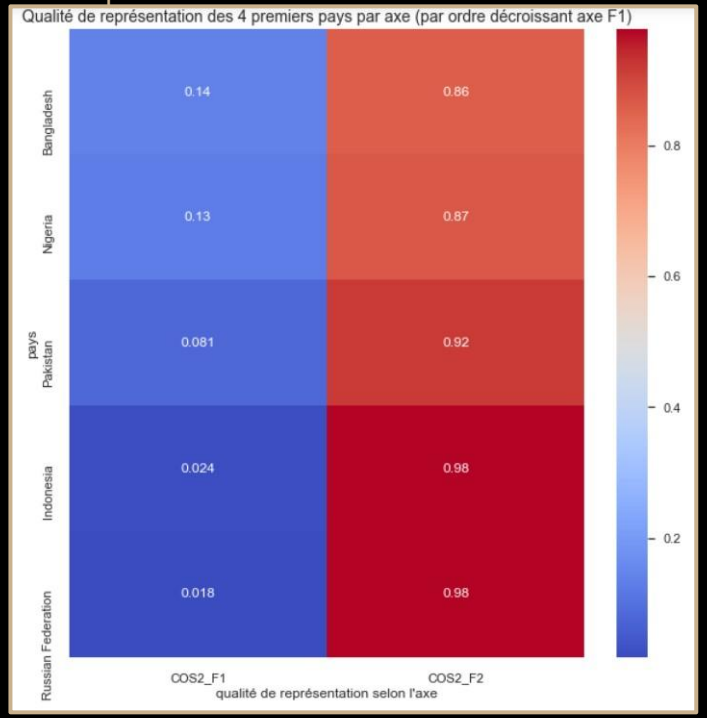


Groupe 1

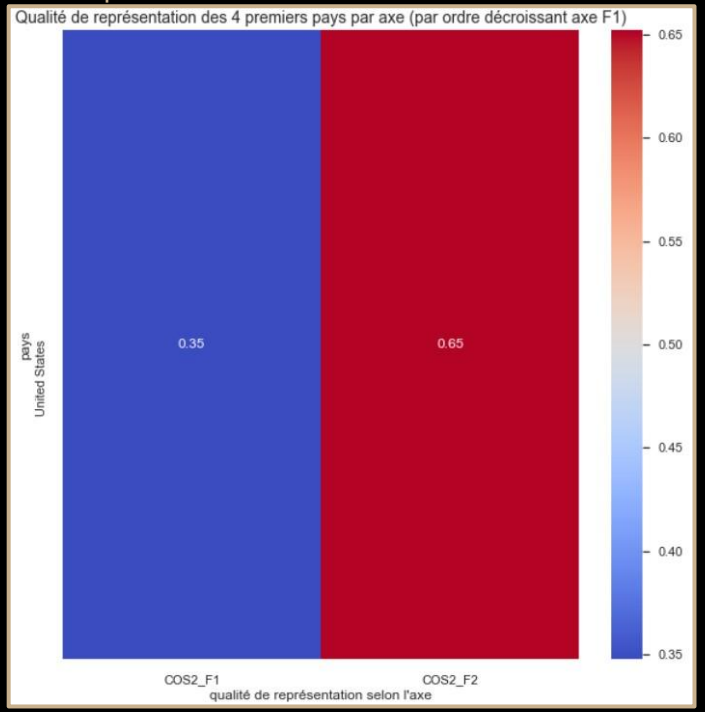


Qualité de Représentation des 4 Premiers Pays par Groupe

Groupe 2



Groupe 3



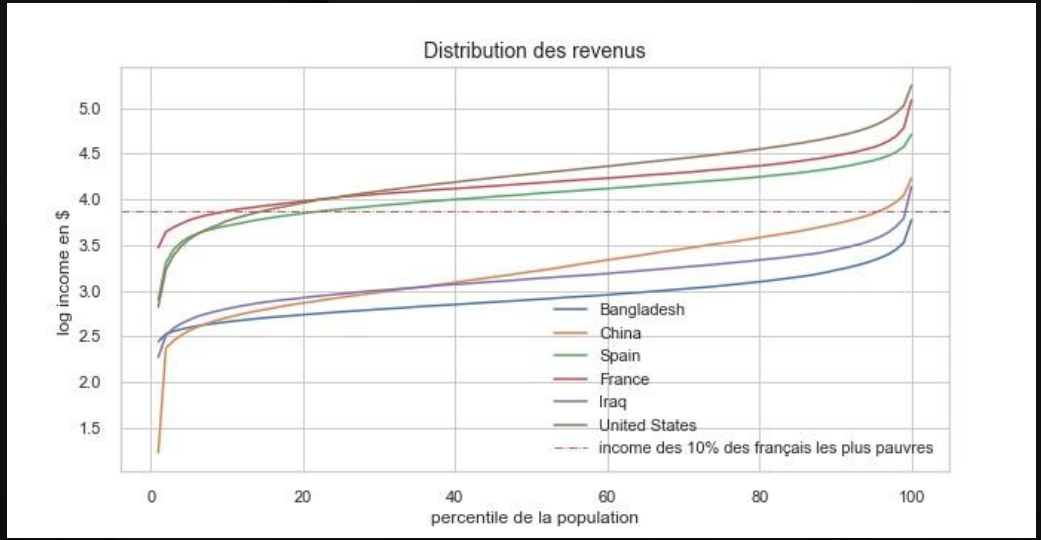
Pays Retenus Pour l’ Analyse

Groupe 0	Irak
Groupe 1	Espagne
Groupe 2	Bangladesh
Groupe 3	USA
Groupe 4	Chine
Groupe 5	France

Diversité des Pays en Terme de Revenus

Diversité des Revenus

- Seuls les ~ 5% des plus riches chinois sont plus riches que les 10% des plus pauvres français ...
- Les 20% des plus pauvres français sont plus riches que les 20 % des plus pauvres américains puis inversion des courbes
- Les 80% des plus riches américains sont plus riches que les 80% français les plus riches



Mesure des Inégalités

Courbes de Lorentz

Inégalitaires

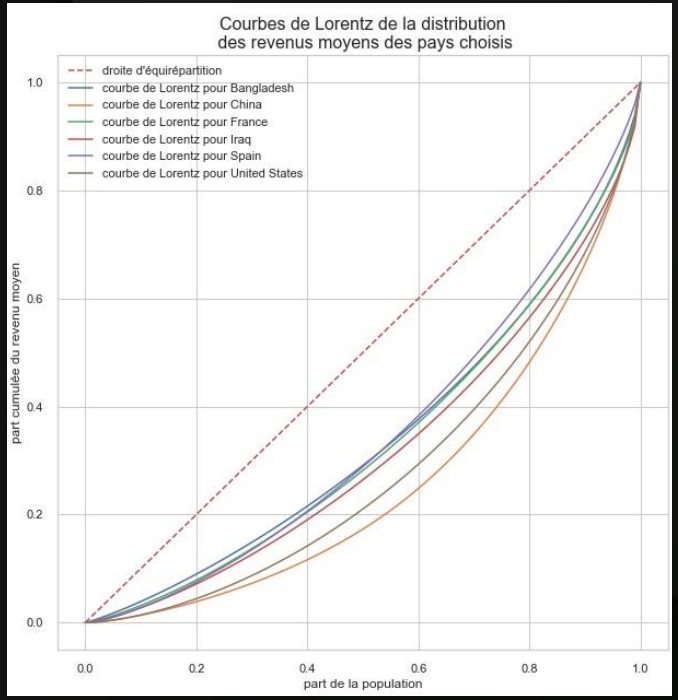
Chine - USA

Égalitaire

Bangladesh

Intermédiaire

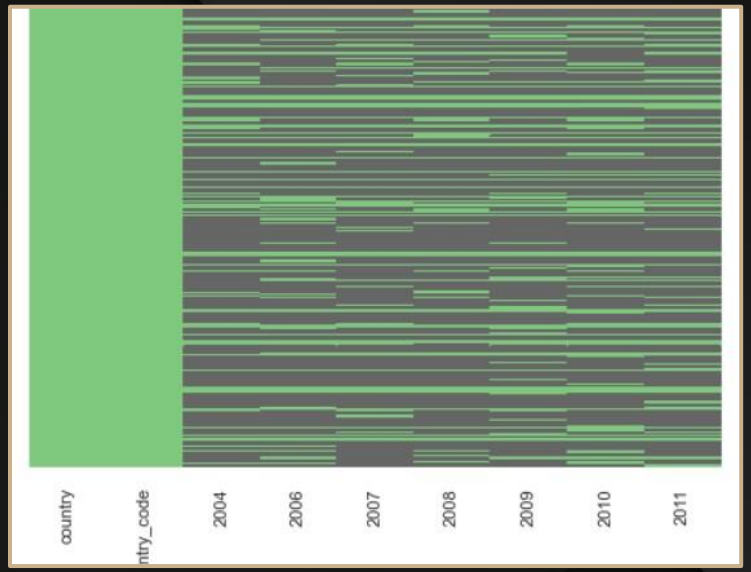
France



Evolution de l’Indice de Gini

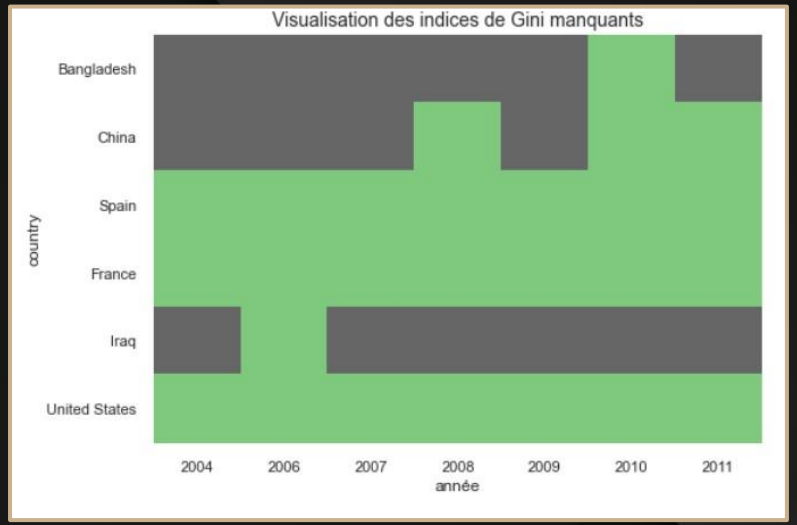
Trop d'Indices de Gini Manquants

- Importation du fichier csv comportant les indices de Gini
- Nom : " WOLD_BANK_gini.csv "
- Source : The World Bank
  Gini index (World Bank estimate) | Data



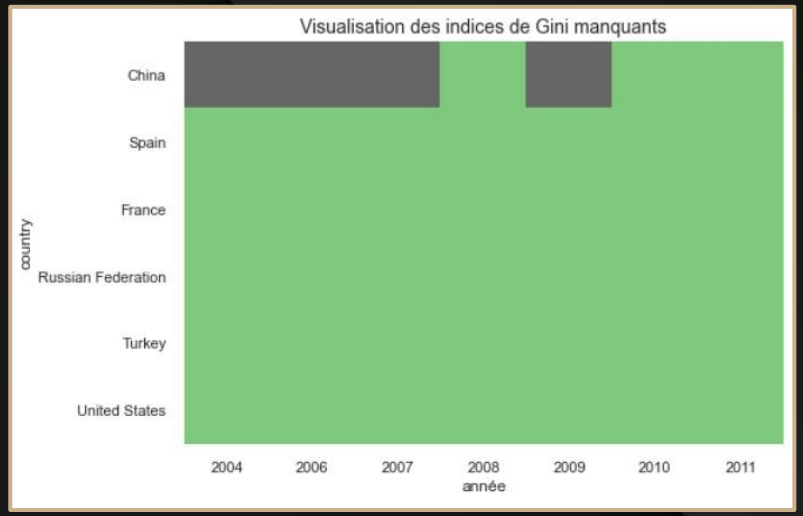
Indices de Gini Manquants pour les pays choisis

- Choisir d'autres pays pour lesquels l'indice de gini est renseigné
- Appartenir au même cluster que le pays initialement choisi
- Avoir bonne qualité de représentation (ACP)



Nouveaux Pays

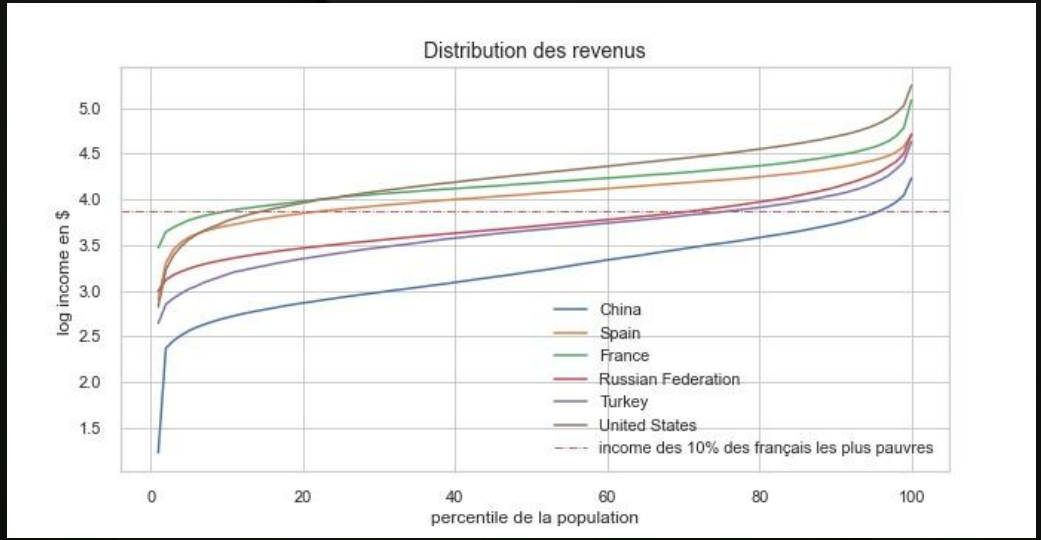
- Turquie remplace l'Irak
- Espagne
- Russie remplace le Bangladesh
- USA
- Chine : indices de Gini récupérés sur le net et calcul pour celui de 2007
- source https://www.ceicdata.com/en/china/resident-income-distribution/gini-coefficient



Diversité des Pays en Terme de Revenus

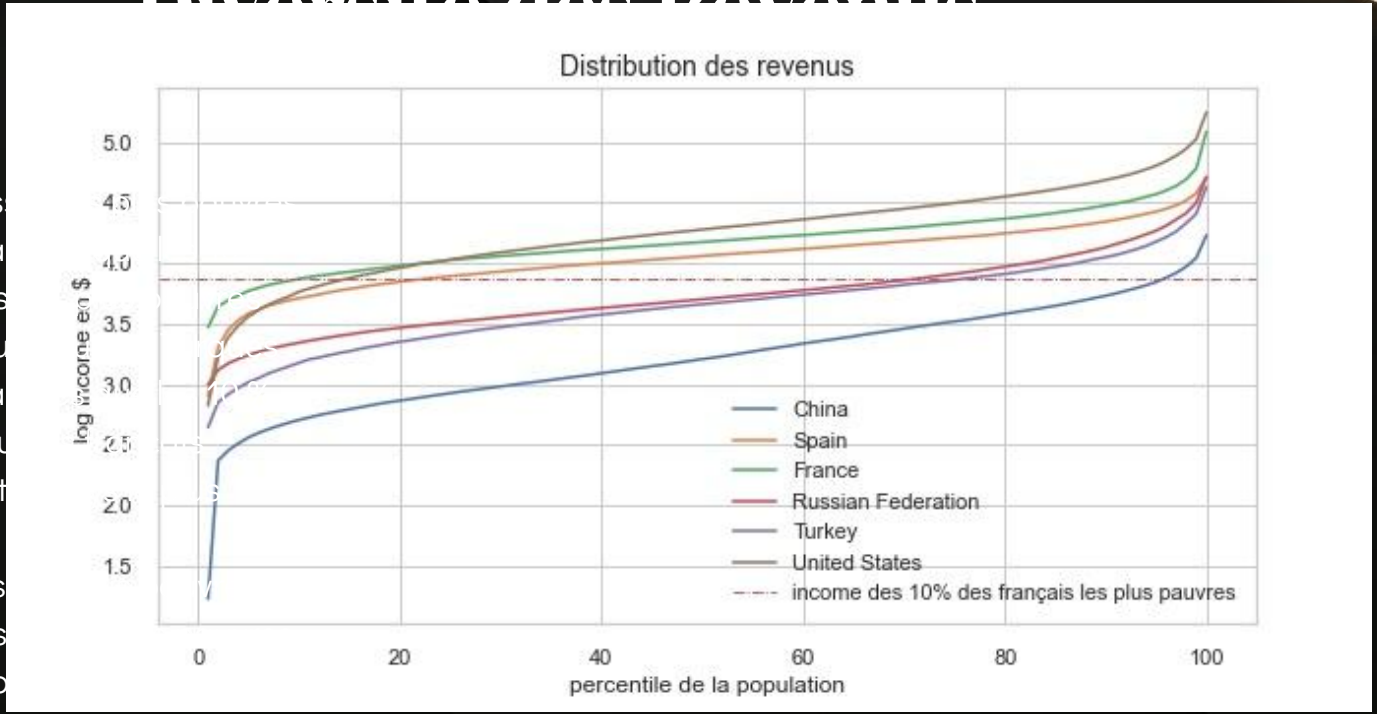
Diversité des Revenus

- 70% des russes les plus pauvres sont plus pauvres que les 10% des français les plus pauvres
- 78 % des plus pauvres Turques sont plus pauvres que les 10 % des plus pauvres français
- La Chine est le pays le plus pauvre
- Les français les plus pauvres sont les plus riches des pauvres des autres pays



Diversité des Revenus

- 70% des rus
  sont plus pa
  des français
- 78 % des plu
  sont plus pa
  des plus pau
- La Chine est
  pauvre
- Les français
  sont les plus
  des autres p
-



Mesure des Inégalités

Courbes de Lorentz

Inégalitaires

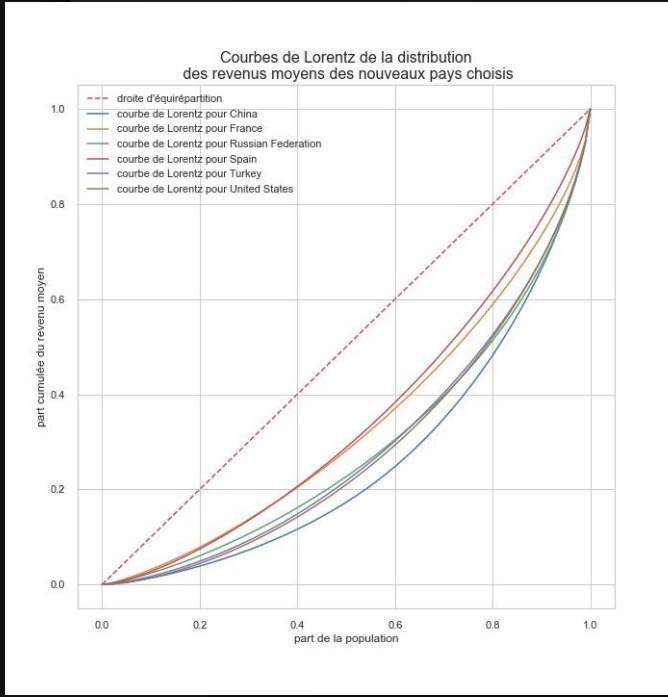
Chine - USA

Égalitaire

Espagne

Intermédiaire

France



Courbes

Inégalitaires

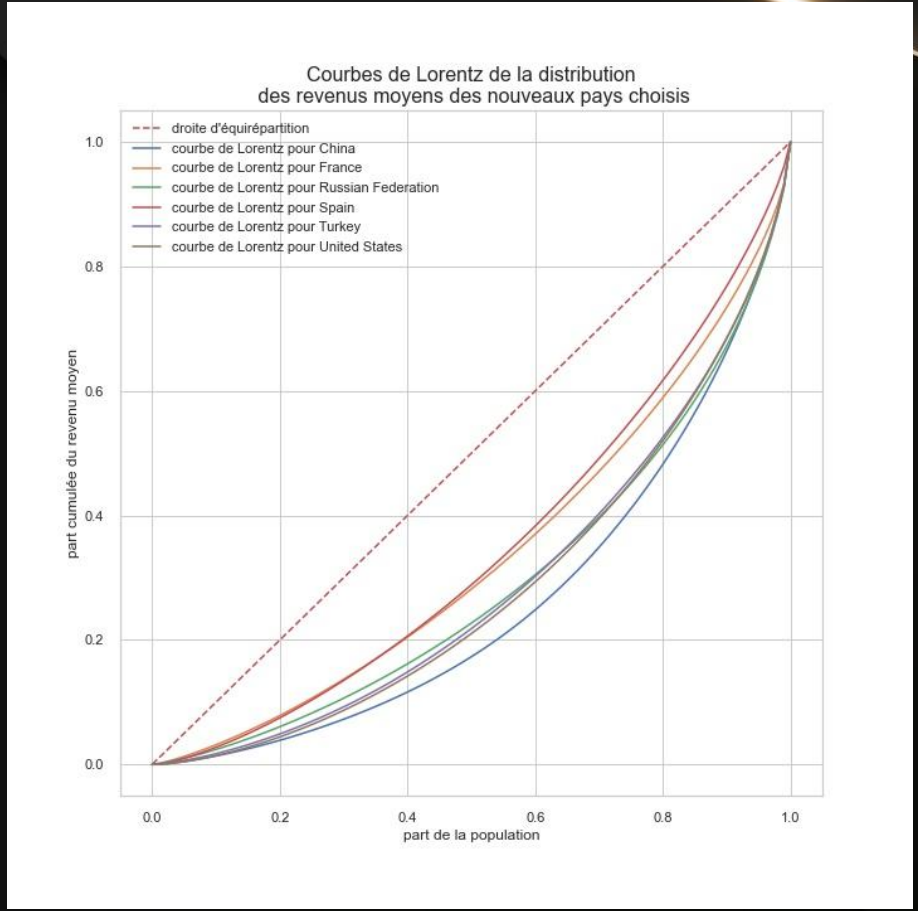
Chine - USA

Égalitaire

Espagne

Intermédiaire

France



Evolution de l’Indice de Gini

Evolution des Indices de Gini à la...

△ Hausse

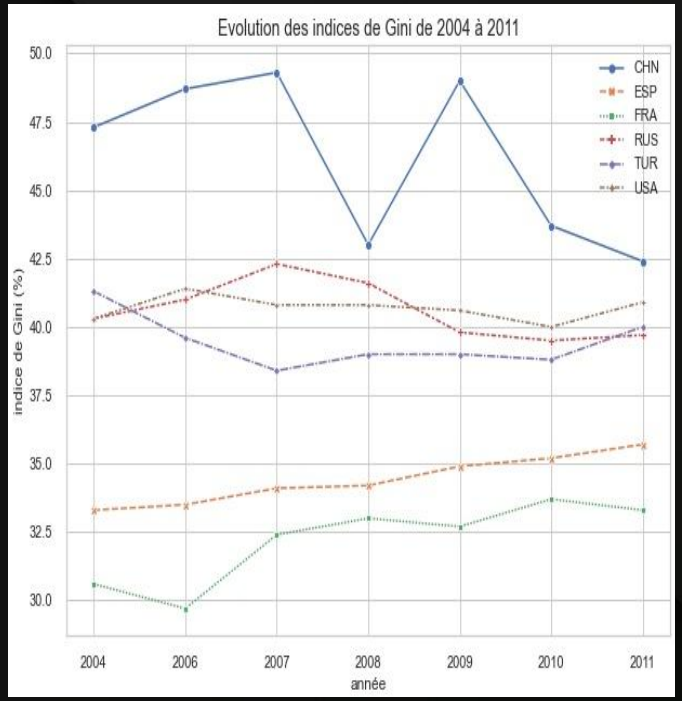
Espagne - France - USA

△ Baisse

Chine - Russie - Turquie

△ Tendance

Évolution globale des inégalités peu marquée



Evolution des Indices de Gini à la...

△ Hausse

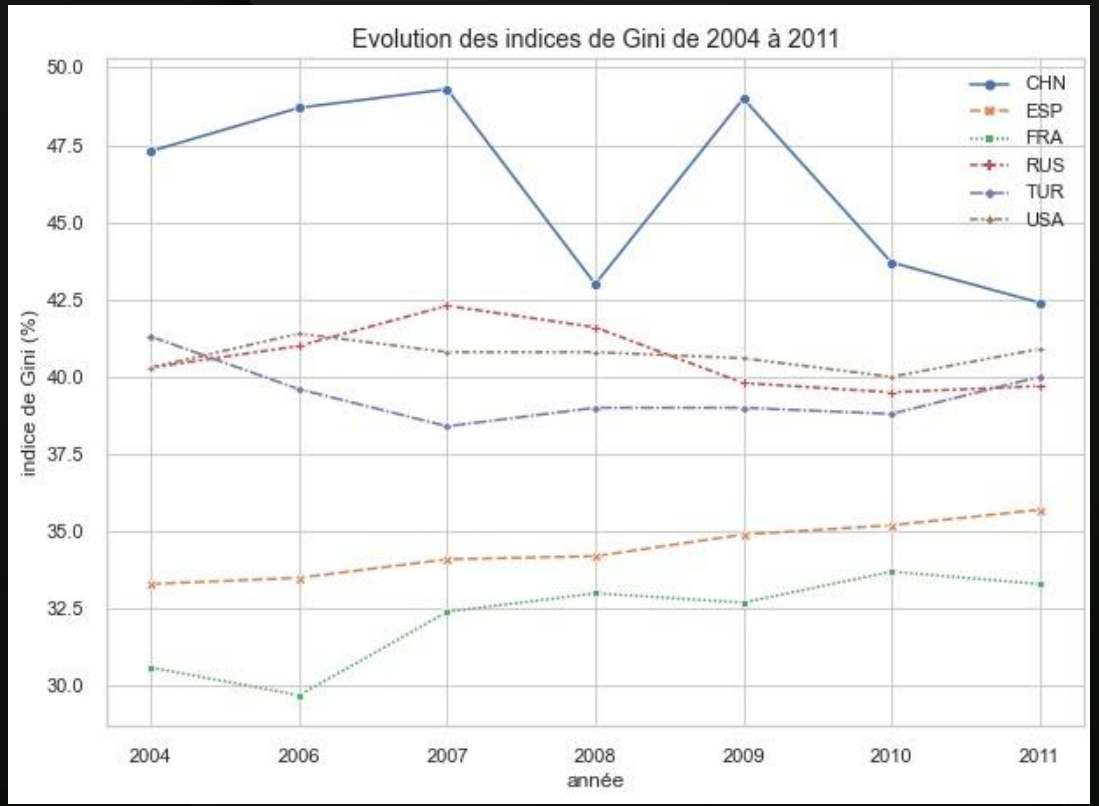
Espagne - France - USA

△ Baisse

Chine - Russie - Turquie

△ Tendance

Évolution peu marquée



Classement des Pays par l’Indice de Gini

8 Pays Sans Aucun Indice de Gini

-
  - Implémentation par le calcul

Ghana	44.32
Kenya	31.58
Cambodge	33.04
Montenegro	30.75
Serbie	29.23
Syrie	37.39
Taïwan	33.15

Palmarès des Pays par Indice de Gini

Classement par Indice Gini Moyen

Indices de Gini les plus forts

Afrique du Sud	63.2
Centrafrique	56.2
Honduras	54.8
Guatemala	54.6
Brésil	54.6

Indices de Gini les plus faibles

Slovénie	24.6
Danemark	26.2
Slovaquie	26.4
Tchéquie	26.5
Azerbaijan	26.6

La France, un Pays “Égalitaire”

31

position

32,2

indice de Gini

Mission 3

Un échantillon 500 fois plus grand !

Classes des 500 Nouveaux Individus

Attribution conforme aux distributions

Source : https://openclassrooms.com/fr/paths/65/projects/148/assignment

Importation des Coef d'Élasticité

Coefficients d'Élasticité

Définition: Mobilité intergénérationnelle des revenus

Proche de 0: Forte mobilité - revenus parent/enfant non liés

Proche de 1: Faible mobilité - enfant hérite du revenu des parents

Source: World Bank
What is the GlobalDatabaseonIntergenerational Mobility (GDIM)? (worldbank.org)

Nom du Fichier: GDIMMay2018.csv

Caractéristiques

Dimension: 6504 lignes , 66 colonnes

Coef d'Élasticité pj: IGEincome

Valeurs Manquantes: 5651

Méthode: Implémentation par la moyenne de l' IGEincome de la région du pays

Calcul des classes de revenu

Revenu Parent

Définition: $\ln(Ychild) = \alpha + pj \ln(Yparent) + \epsilon$

Protocole: Générer les revenus des parents (en log) selon une loi normale

Méthode: Générer n réalisations du terme d'erreur ϵ selon une loi normale ($\mu = 0$, std = 1)

Méthode: Générer les revenus des enfants (Ychild) en fonction de celui des parents (Yparent) et pj

Classe Revenu Parent/Enfant

Méthode: Calcul des classes de revenu à partir des Ychild et Yparent

Méthode: Création d'un dataframe comportant les revenus et classes de revenu enfant / parent

	y_child	y_parents	c_i_child	c_i_parent
13738	0.041265	0.178576	1	5
85798	0.033826	0.170195	1	4

Pour Chaque Classe Enfant , la Distribution Conditionnelle de la Classe Parent

- Etablissement d'une liste des fréquences des différentes combinaisons de classes enfants-parents
- Comptage des différentes combinaisons de classes enfants-parents
- Calcul de la probabilité conditionnelle à partir des classes enfants et des classes parents
- Notation

  $P(c\_i\_parent = 8 \mid c\_i\_child = 5, pj = 0.9) = 0.03$

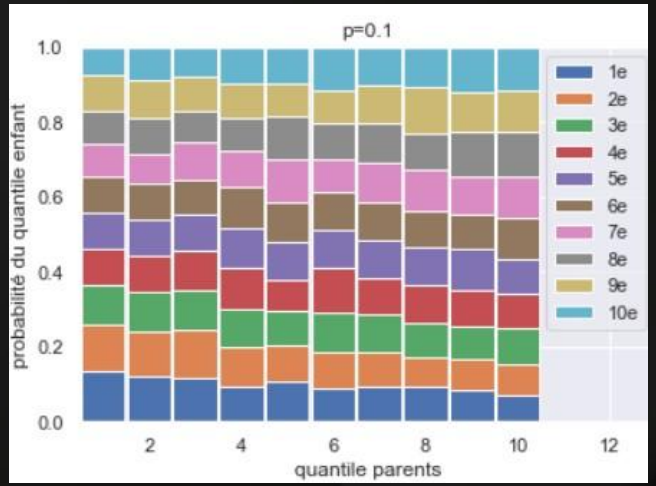
- Exemple

  - pj = 0.9
  - 6 individus avec à la fois c_i_child = 5 et c_i_parent = 8
  - 200 individus sur 20000 avec c_i_child = 5

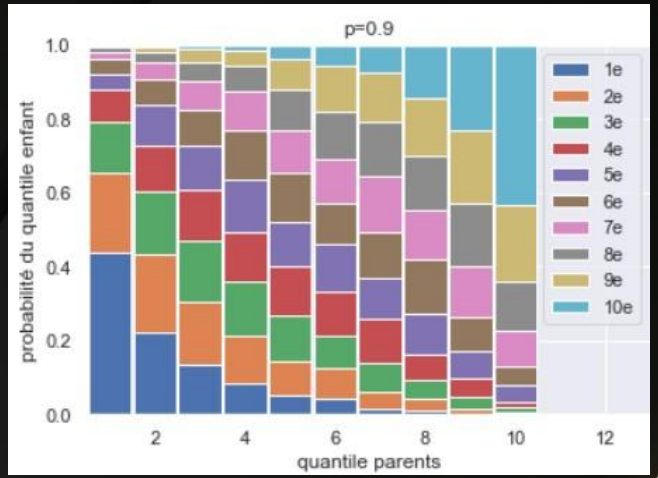
  alors la probabilité d'avoir c_i_parent = 8 sachant c_i_child = 5 et sachant pj = 0.9 estimée à 6/200

Distributions Conditionnelles

Forte Mobilité



Faible Mobilité



Un Échantillon 500 Fois Plus Grand !

Mode Opératoire

- Pour chaque individu de la Wold Income Distribution, créer 499 "clones"
- Attribuer aux 500 individus leurs classes parents conformément aux distributions trouvées précédemment
- Utilisation des pj correspondants aux pays
- Typage des variables pour réduire la place en mémoire (category, int16 et float32)

Remarques

Vitesse Programme: Très lent (~16 minutes)

Classe Enfant: Suppression de cette variable (inutile dans la mission 4)

Dimensions: 5 800 000 lignes, 6 colonnes

	country	country_code	income	pj	c_i_parent	mj	Gj
0	Albania	ALB	728.89795	0.535604	1.0	2994.829902	30.0
1	Albania	ALB	728.89795	0.535604	1.0	2994.829902	30.0

Mission 4

Expliquer le revenu des individus en fonction de plusieurs variables explicatives

ANOVA

Variable explicative : le pays de l'individu
Etude de l'effet du pays sur les revenus d'un habitant

Dataset d'Analyse

Caractéristiques

Nouvelles Variables: Log de income
Log revenu moyen du pays (mj)

Accélérer le Traitement des Données: Nouvel échantillonnage en prenant TOUS les pays mais seulement 5000 individus par pays choisis aléatoirement (au lieu de 500 * 100 individus)

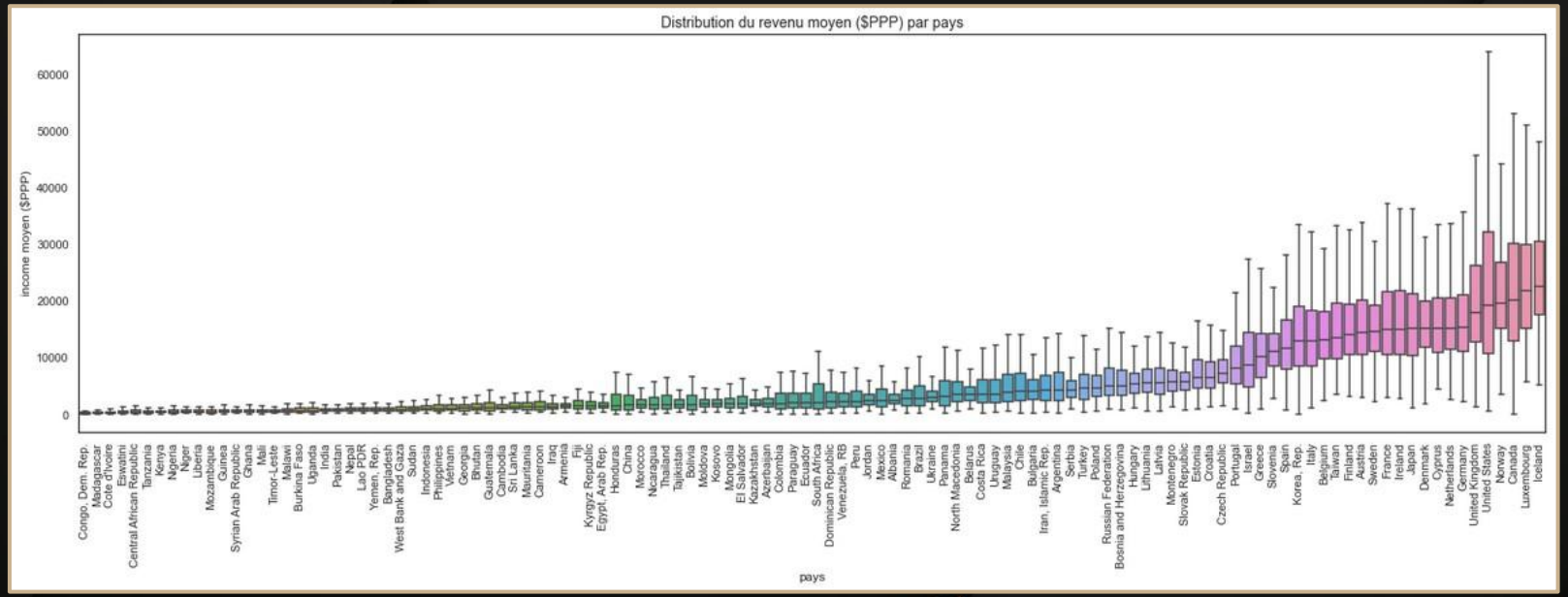
Fichier

Dimensions Initiales: 5 800 000 colonnes, 9 colonnes

Dimensions Finales: 580 000 colonnes, 9 colonnes

	country	country_code	income	pj	c_i_parent	mj	Gj	ln_income	ln_mj
0	Albania	ALB	3061.0693	0.535604	64.0	2994.829902	30.0	8.026520	8.004643
1	Albania	ALB	17754.3240	0.535604	93.0	2994.829902	30.0	9.784384	8.004643

Des Revenus Inégalement Répartis Selon les Pays



ANOVA sans Log

Le Pays Influe Sur Les Revenus

ANOVA

Question: Existe-t-il une différence statistiquement significative dans les moyennes de revenus des individus des différents pays

Tests d'Hypothèse: H0 : les moyennes sont égales
H1 : les moyennes sont différentes
Seuil alpha : 0,05

Résultat

p_value: 0

H0 rejetée: p_value inférieure au seuil fixé

Moyennes des Incomes: Significativement différentes selon le pays

Pays d'Origine: Influe sur les revenus de l'individu

Modèle: Expliqué à 49,74 % par la variable Country

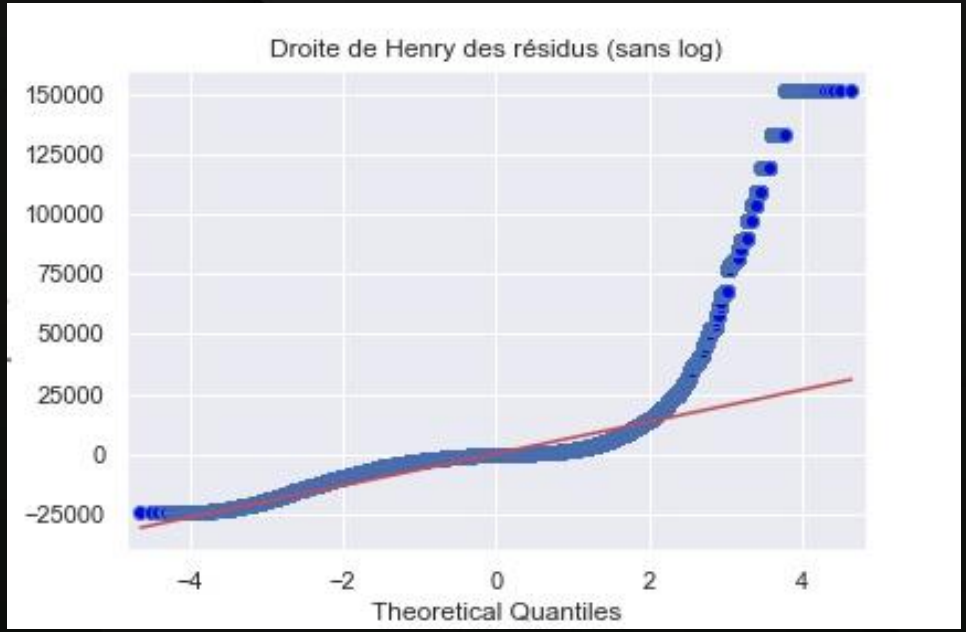
Droite de Henry des Résidus

△ Normalité

Pas de distribution normale

△ Mais ...

Échantillon de grande taille donc les écarts par rapport à la normalité n'ont pas d'impact considérable sur les résultats (TCL)



ANOVA avec Log

Le Pays Influe Sur Les Revenus

ANOVA

Tests d'Hypothèse: H0 : les moyennes sont égales
H1 : les moyennes sont différentes
Seuil alpha : 0,05

Conclusion: Le modèle est amélioré par les données logarithmiques (baisse de l'AIC et du BIC)

Résultat

p_value: 0

H0 rejetée: p_value inférieure au seuil fixé

Moyennes des Incomes: Significativement différentes selon le pays

Pays d'Origine: Influe sur les revenus de l'individu

Modèle: Expliqué à 73 % par la variable Country

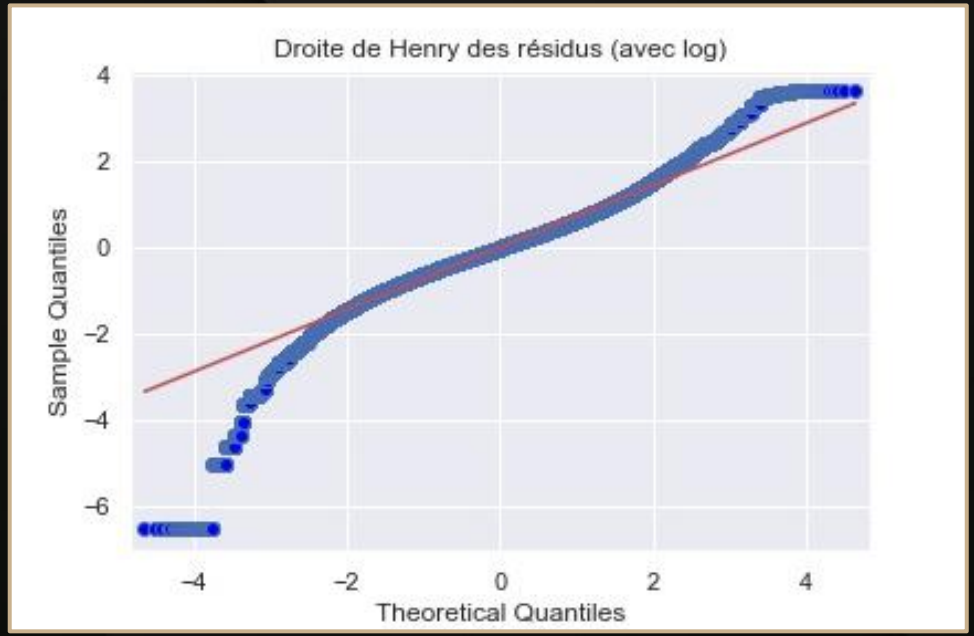
Droite de Henry des Résidus

Normalité

Distribution normale

Log

Transformer les income en log a permis de rapprocher les valeurs extrêmes



Régression linéaire

Variables explicatives :
revenu moyen et indice de Gini du pays de l'individu

Régression sans log

Modèle Income

- Variable cible : income
- Variables explicatives : mj, Gj
- Résidus : au seuil de 5%, l'hypothèse null de normalité des résidus peut être rejetée (Prob(JB):0)
- Prob (F-statistic): 0 : la p_value est inférieure au seuil alpha de 0.05. Le modèle est globalement significatif
- Pourcentage de variance expliquée par le modèle 49.7 % (R^2 : 0.497)
- Variable significative au seuil de 5% : mj

OLS Regression Results

Dep. Variable:	income	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	2.869e+05
Date:	Sat, 18 Sep 2021	**Prob (F-statistic):**	0.00
Time:	17:06:44	**Log-Likelihood:**	-5.9304e+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	53.4386	48.388	1.104	0.269	-41.401	148.278
mj	1.0004	0.001	710.000	0.000	0.998	1.003
Gj	-1.3818	1.166	-1.185	0.236	-3.668	0.904

Omnibus:	730698.302	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	212466196.805
Skew:	6.749	**Prob(JB):**	0.00
Kurtosis:	95.787	**Cond. No.**	4.97e+04

Régression avec log

Modèle Income

- Variable cible : log income (ln_income)
- Variables explicatives : log mj (ln_mj) et Gj
- Résidus : au seuil de 5%, l'hypothèse null de normalité des résidus peut être rejetée (Prob(JB):0)
- Prob (F-statistic): 0 : la p_value est inférieure au seuil alpha de 0.05. Le modèle est globalement significatif
- Pourcentage de variance expliquée par le modèle 72,8 % (R^2 : 0.728)
- Variable significative au seuil de 5% : ln_mj et Gj

OLS Regression Results

Dep. Variable:	ln_income	**R-squared:**	0.728
Model:	OLS	**Adj. R-squared:**	0.728
Method:	Least Squares	**F-statistic:**	7.773e+05
Date:	Sat, 18 Sep 2021	**Prob (F-statistic):**	0.00
Time:	17:07:20	**Log-Likelihood:**	-6.3268e+05
No. Observations:	580000	**AIC:**	1.265e+06
Df Residuals:	579997	**BIC:**	1.265e+06
Df Model:	2		
Covariance Type:	nonrobust		

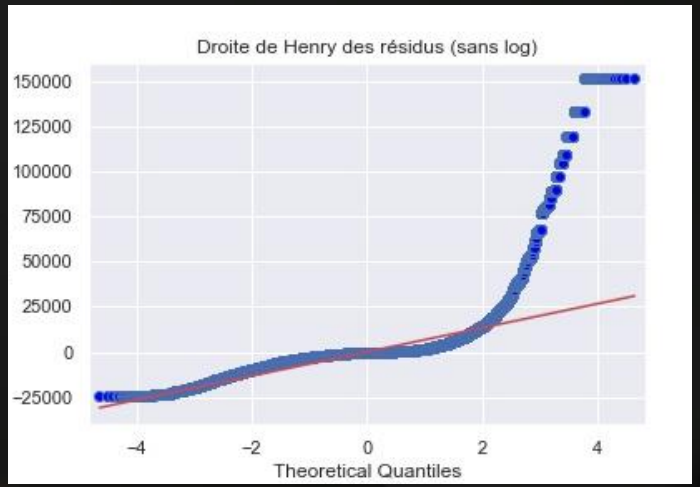
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4953	0.009	52.963	0.000	0.477	0.514
ln_mj	0.9881	0.001	1154.725	0.000	0.986	0.990
Gj	-0.0177	0.000	-144.194	0.000	-0.018	-0.017

Omnibus:	37762.545	**Durbin-Watson:**	1.992
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	176361.518
Skew:	-0.104	**Prob(JB):**	0.00
Kurtosis:	5.693	**Cond. No.**	390.

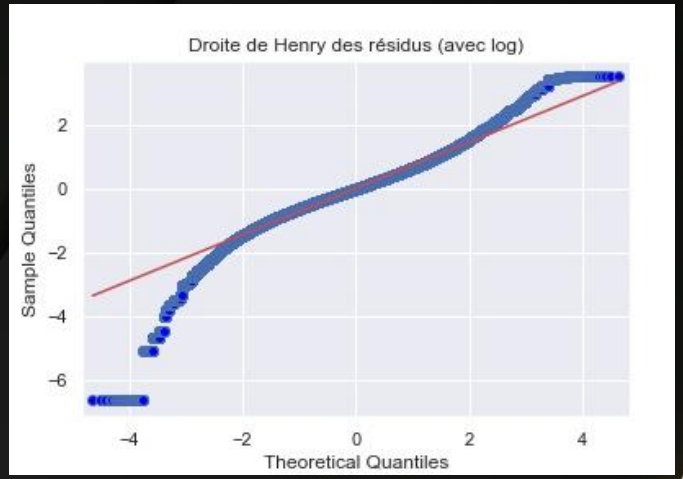
Variable ln_mj : explique 49,73 % de la variance ln_income
Variable Gj : explique 0,97 % de la variance ln_income

Droites de Henry des Résidus

Sans Log



Avec Log



Test de Kolmogorov-Smirnov sur les Résidus

Sans Log

Tests d'Hypothèse: H0 : la variable suit une loi normale
H1 : la variable ne suit pas une loi normale
Seuil alpha : 0,05

p_value: 5.69e-52

Conclusion: HO rejetée au seuil alpha, la variable résidu ne suit pas une loi normale MAIS échantillon suit le théorème central limite.

Avec Log

Tests d'Hypothèse: H0 : la variable suit une loi normale
H1 : la variable ne suit pas une loi normale
Seuil alpha : 0,05

p_value: 0.075

Conclusion: HO ne peut pas être rejetée au seuil alpha, la variable résidu suit une loi NORMALE

Les données log sont plus proches de la tendance gaussienne et donc plus intéressantes à utiliser dans un modèle de régression linéaire

Régression linéaire

Variable explicative supplémentaire :
classe de revenu des parents

Régression sans log

Modèle Income

- Variable cible : income
- Variables explicatives : mj, Gj et classe revenu parent (c_i_parent)
- Résidus : au seuil de 5%, l'hypothèse null de normalité des résidus peut être rejetée (Prob(JB):0)
- Prob (F-statistic): 0 : la p_value est inférieure au seuil alpha de 0.05. Le modèle est globalement significatif
- Pourcentage de variance expliquée par le modèle 52,.4 % (R^2 : 0.524)
- Variable significative au seuil de 5% : mj, c_i_parent

OLS Regression Results

Dep. Variable:	income	R-squared:	0.524
Model:	OLS	Adj. R-squared:	0.524
Method:	Least Squares	F-statistic:	2.129e+05
Date:	Sat, 18 Sep 2021	Prob (F-statistic):	0.00
Time:	17:07:35	Log-Likelihood:	-5.9146e+06
No. Observations:	580000	AIC:	1.183e+07
Df Residuals:	579996	BIC:	1.183e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2640.9424	49.392	-53.469	0.000	-2737.750	-2544.135
mj	0.9999	0.001	729.309	0.000	0.997	1.003
Gj	-1.2851	1.135	-1.132	0.258	-3.510	0.939
c_i_parent	53.3585	0.296	180.526	0.000	52.779	53.938

Omnibus:	742668.822	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	236841197.570
Skew:	6.927	Prob(JB):	0.00
Kurtosis:	101.022	Cond. No.	5.21e+04

Conclusion : l'ajout de la classe de revenu des parents a amélioré le modèle de régression.

Régression avec log

Modèle Income avec Log : Meilleur Modèle !

- Variable cible : ln_income
- Variables explicatives : ln_mj, Gj et classe revenu parent (c_i_parent)
- Résidus : au seuil de 5%, l'hypothèse null de normalité des résidus peut être rejetée (Prob(JB):0)
- Prob (F-statistic): 0 : la p_value est inférieure au seuil alpha de 0.05. Le modèle est globalement significatif
- Pourcentage de variance expliquée par le modèle 78,4 % (R^2 : 0.784)
- Toutes les variables significatives au seuil de 5 %

OLS Regression Results

Dep. Variable:	ln_income	R-squared:	0.784
Model:	OLS	Adj. R-squared:	0.784
Method:	Least Squares	F-statistic:	7.035e+05
Date:	Sat, 18 Sep 2021	Prob (F-statistic):	0.00
Time:	17:07:42	Log-Likelihood:	-5.6554e+05
No. Observations:	580000	AIC:	1.131e+06
Df Residuals:	579996	BIC:	1.131e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0739	0.008	-8.736	0.000	-0.090	-0.057
ln_mj	0.9875	0.001	1295.653	0.000	0.986	0.989
Gj	-0.0176	0.000	-161.637	0.000	-0.018	-0.017
c_i_parent	0.0114	2.92e-05	388.727	0.000	0.011	0.011

Omnibus:	39781.433	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	186071.136
Skew:	-0.151	Prob(JB):	0.00
Kurtosis:	5.758	Cond. No.	684.

Variable ln_mj : explique 71,85 % de la variance ln_income
Variable Gj : explique 0,97 % de la variance ln_income
Variable c_i_parent : explique 5,62 % de la variance ln_income

Analyse du Modèle

Atypisme & Influence

Levier: 2,04 % de points atypiques

Résidus Studentisés: 2,76 % de points atypiques

Distance de Cook: 5,62 % de points influents

Conclusion: Le modèle n'est pas totalement satisfaisant. D'autres variables explicatives devraient être ajoutées au modèle (éducation...)

Etudes Variables

Colinéarité: Variables non colinéaires
VIF : tous les coefficients sont inférieurs à 10 (1,077 - 1,077 - 1,000)

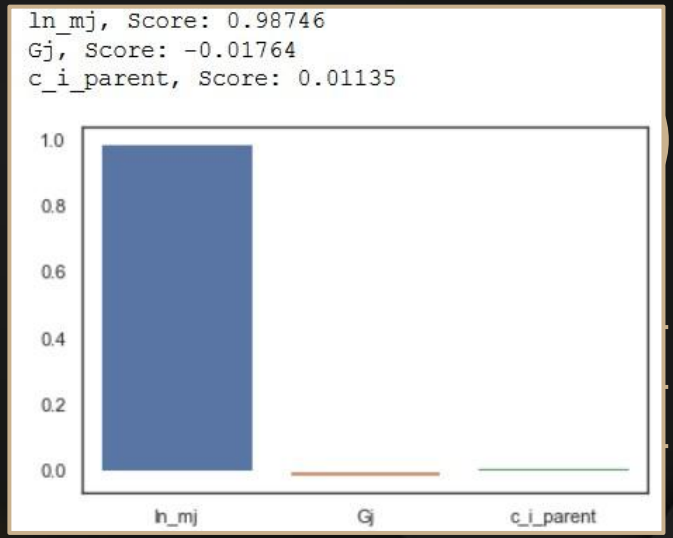
Homoscédasticité: Résidus à variances différentes
Test de Breusch-Pagan
H0 : homoscédasticité - p_value : 0
HO : rejetée

Normalité des Résidus: Test de Kolmogorov-Smirnov
- résidus non gaussiens
MAIS distribution symétrique
n > 50 donc résultats de la régression acceptables

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 ?

Coefficient de Gini

Influence des Variables



Influence du Coef de Gini

Coefficient de Gini: - 0,0176

Négatif: Croît inversement à ln_income

Très Faible: Peu d'impact, pas pertinent

Conclusion: Le coefficient de régression associé à l'indice de Gini ne permet pas d'affirmer que le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu'il n'en défavorise

Merci

Des questions?

linkedin.com/in/isabelle-barbier