PRÉDICTION DE REVENUS

CLADIERE Nathan, projet 7, Formation Data Analyst OC

Sommaire

- Introduction
  - *Contextes et objectifs*
  - *Données utilisées*
  - *Préparation des données*
- Présentation détaillée des données
- Différences entre les pays
- Détermination de la classe de revenus des parents
- Facteurs majeurs influençant le futur revenu

Contexte et objectifs

- Recherche de nouveaux clients potentiels
  - *Les jeunes qui vont ouvrir un compte*
  - *Les jeunes susceptibles d'avoir un haut niveau de revenu*
- Modèle permettant de déterminer quel revenu vont avoir les enfants.
  - *Revenu moyen du pays*
  - *Indice de Gini du pays*
  - *Classe de revenu des parents*
- Quels enfants cibler ? dans quel type de pays ?

Données utilisées

- Données issues d'études nationales:
  - *Pays*
  - *Années d'observation*
  - *Quantile*
  - *Nombre de quantiles*
  - *Revenus de la personne ($PPP)*
  - *Gdpppp (PIB par personne en $PPP)*

- Données calculées:
  - *Indice de Gini*
  - *Quantile des parents*

- Ajouts:
  - *Coefficient d'élasticité (Divers études et banque mondiale)*
  - *Population (banque mondiale)*

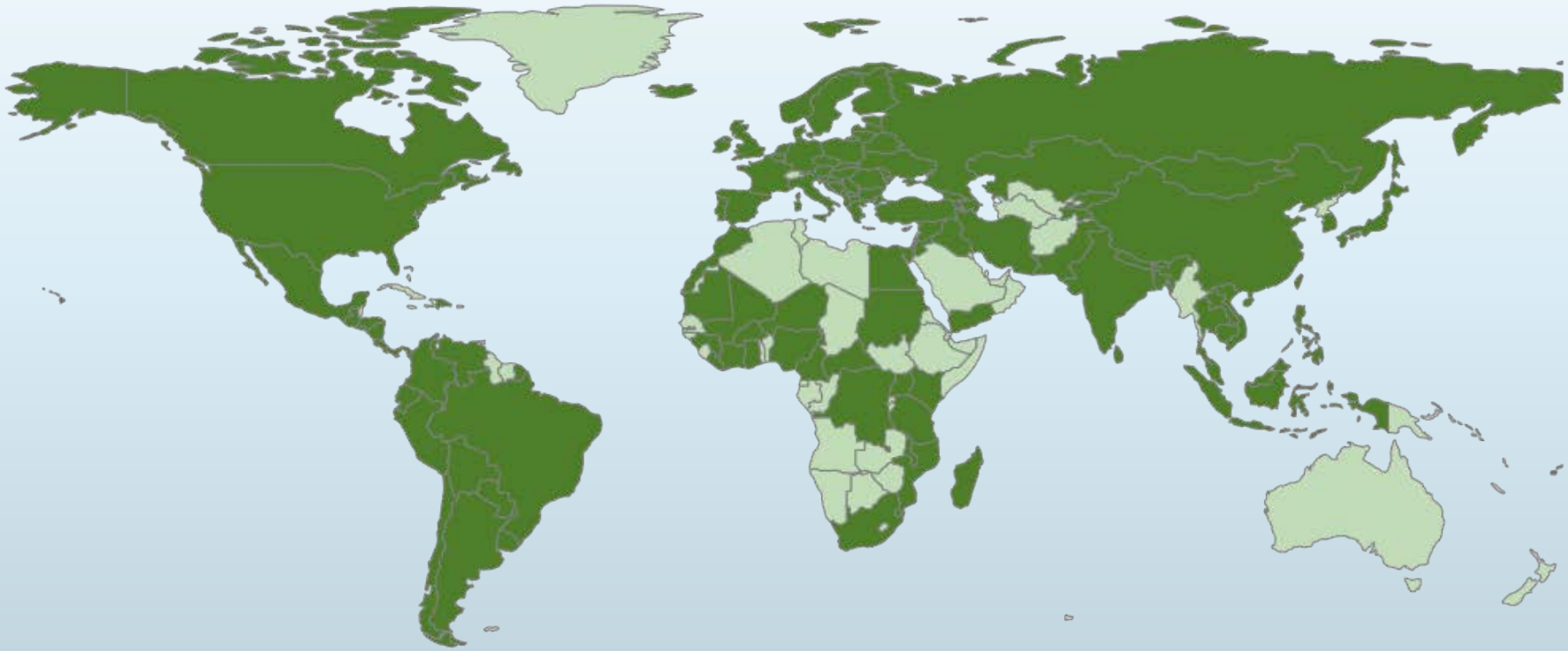
Préparation des données

- Ajouts:
  - *Population*
  - *Indice de Gini (calcul de l'aire sous la courbe de lorenz)*
- Données manquantes :
  - *GDPPP (Kosovo, Territoire palestinien)*
  - *Quantile (41, Lituanie)*
- Valeurs aberrantes :
  - *GDPPP (Fidji)*

PRÉSENTATION DÉTAILLÉES DES DONNÉES UTILISÉES

Mission 1

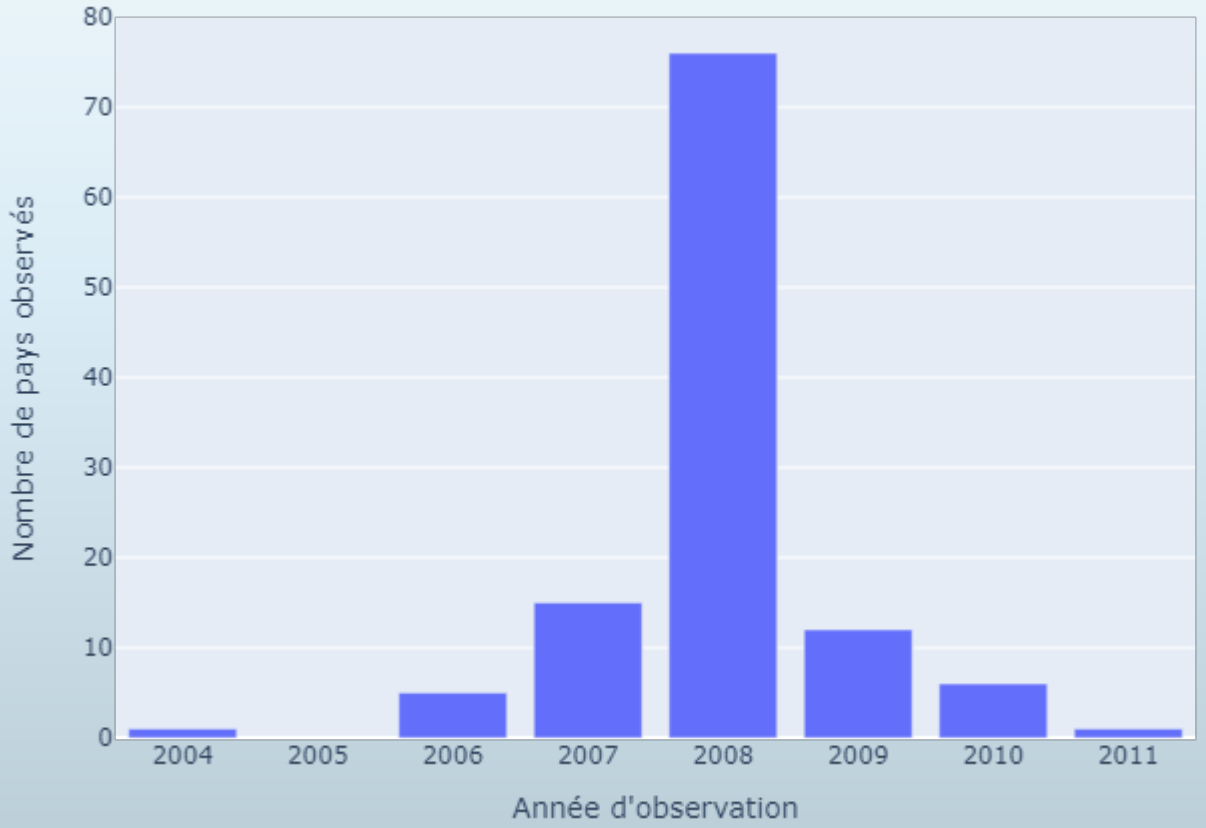
Pays observés



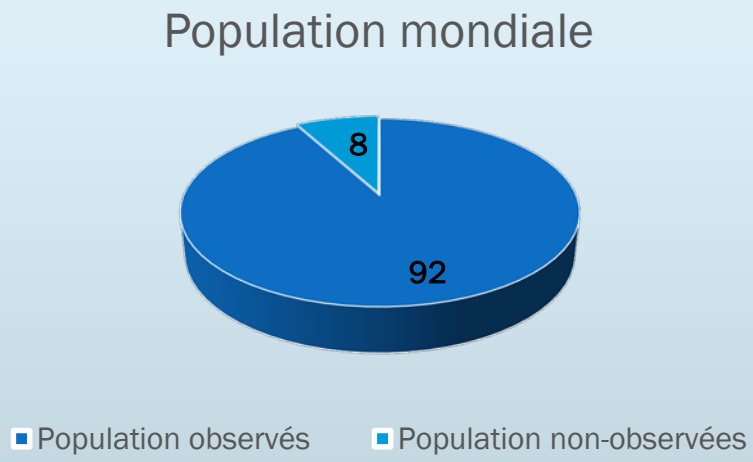
58% des pays observés

Années d’observation

- De 2004 à 2011
- Principalement en 2008



Pourcentage de population et échantillonnage



92% de la population mondiale étudiée

Le centile: un échantillonnage facilitant l'analyse

- Besoin de parler en pourcentage de population

Pourcentage de la population	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	…	100
Quartiles	Q1																									Q2 … Q3				…Q4
Déciles	D1										D2										D3…D9									…D10
Centiles	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	…	100

- 1% population ayant les même caractéristiques

Conclusion

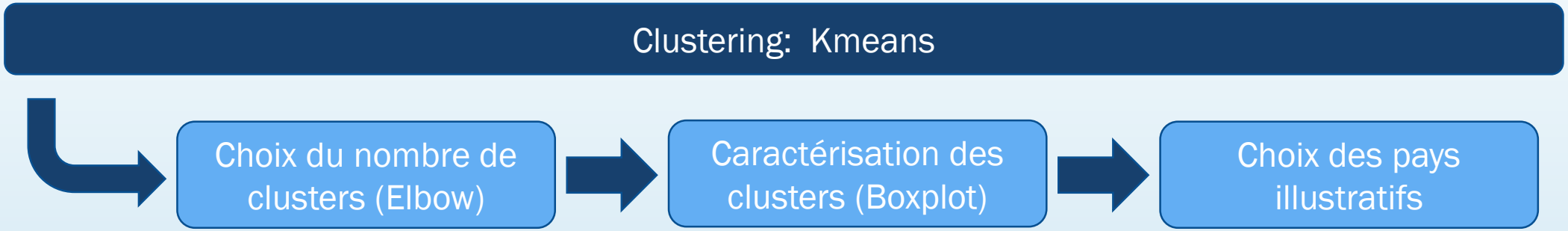
- Analyse sur une grande partie de la population
- Donnée assez récente (2008)
  - *Enfant sont en mesures d'ouvrir un compte (12 ans)*

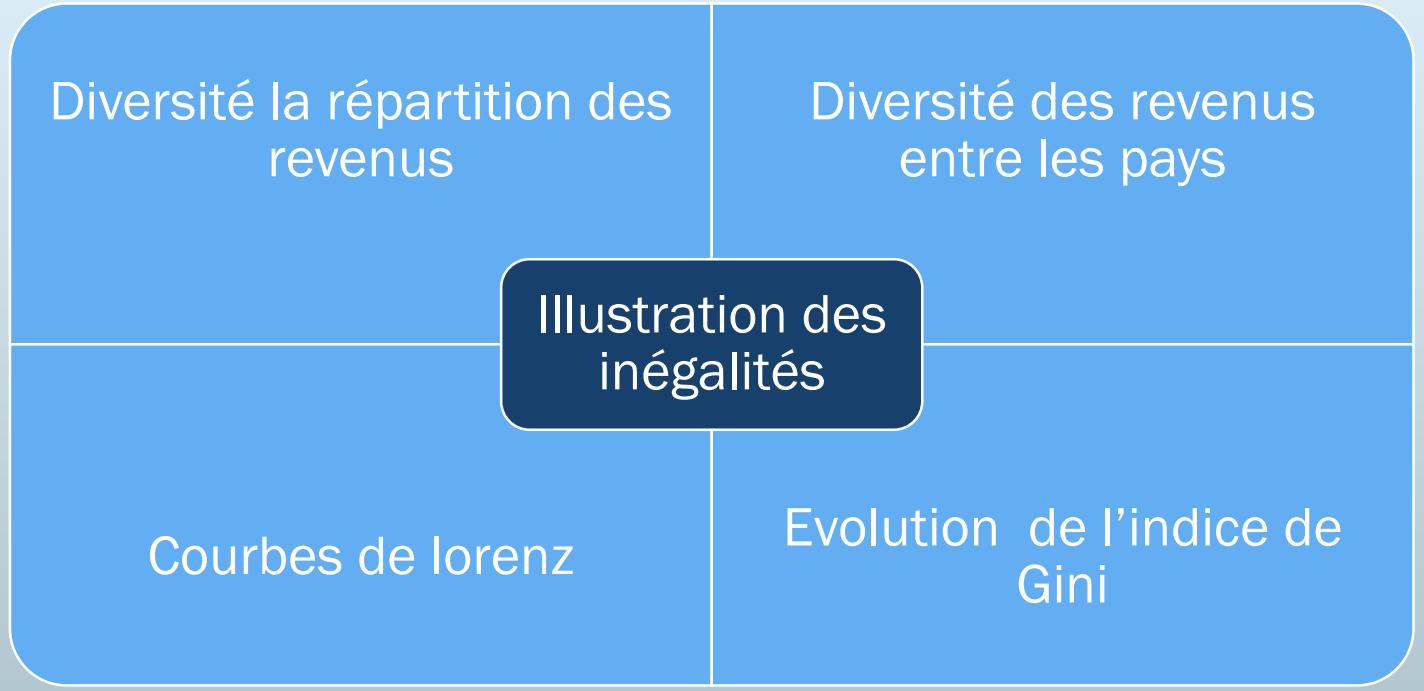
- Y-a-t-il des différences entre les pays ?
  - *Hauteurs de revenus*
  - *Répartition de la richesse*

DES PAYS QUI NE SE RESSEMBLENT PAS

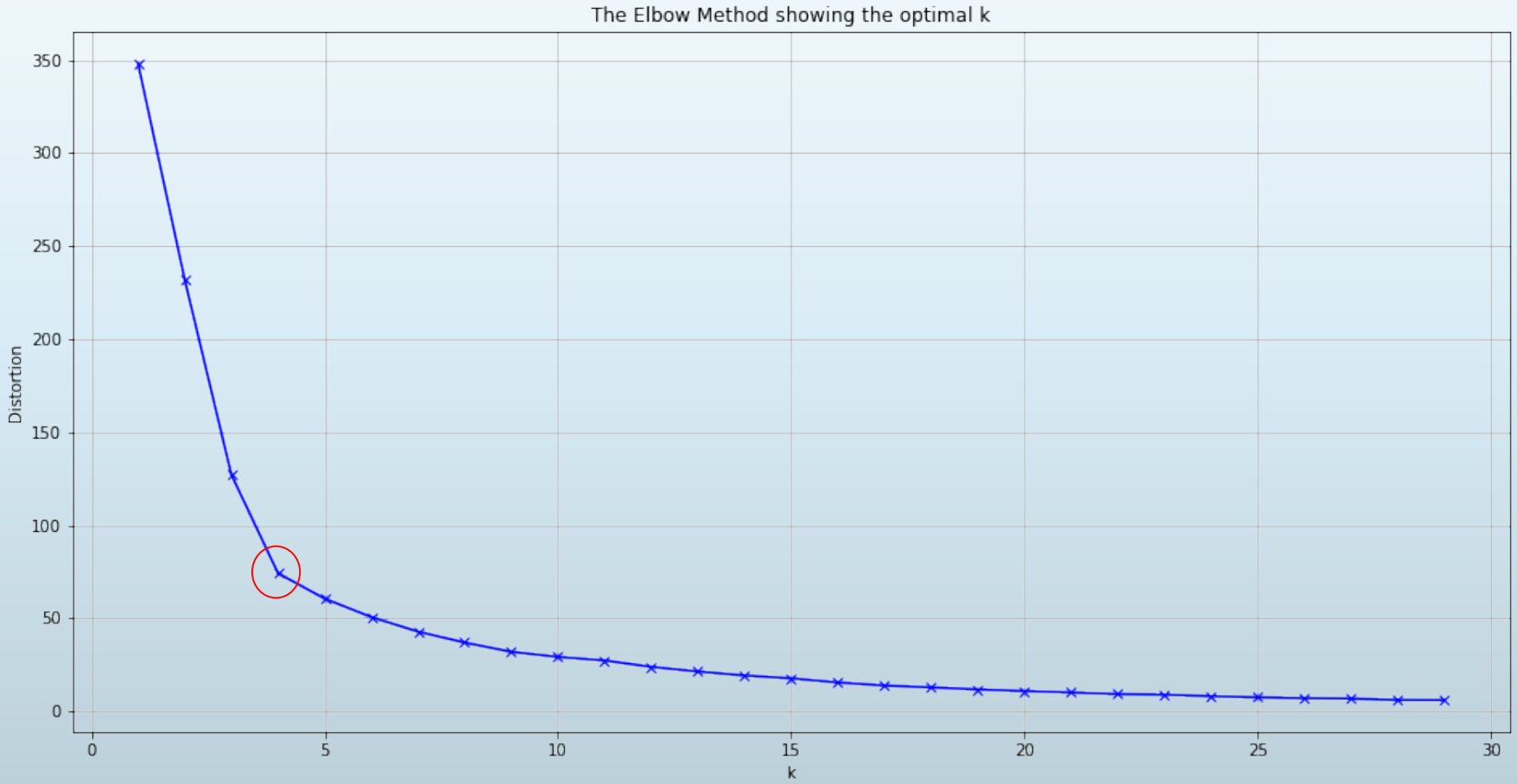
Mission 2

Démarche


Clustering: Kmeans
Choix du nombre de clusters (Elbow)
Caractérisation des clusters (Boxplot)
Choix des pays illustratifs

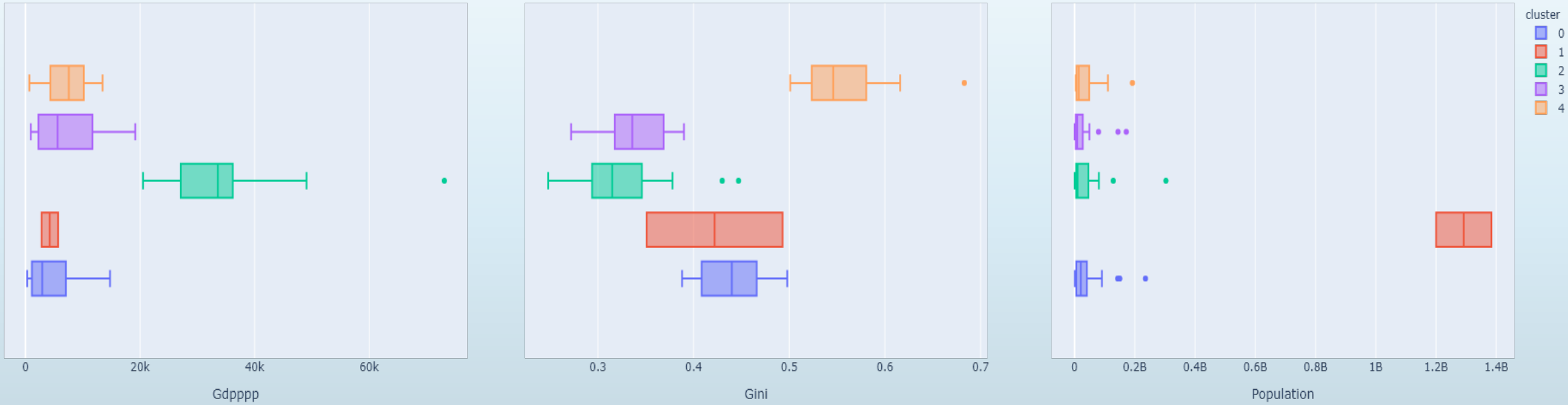

Diversité la répartition des revenus
Diversité des revenus entre les pays
Illustration des inégalités
Courbes de lorenz
Evolution de l'indice de Gini

Clustering: Choix du nombre de clusters



- 4 clusters suffisent
- Choix de 5 clusters pour plus de détails

Clustering: Caractéristiques des clusters



- Cluster 2: riches et égalitaires
- Cluster 1: très peuplés

- Cluster 4: très inégalitaires
- Cluster 3 et 0: Peu peuplés
  - *3 plus riches et égalitaires*

Clustering: quels pays

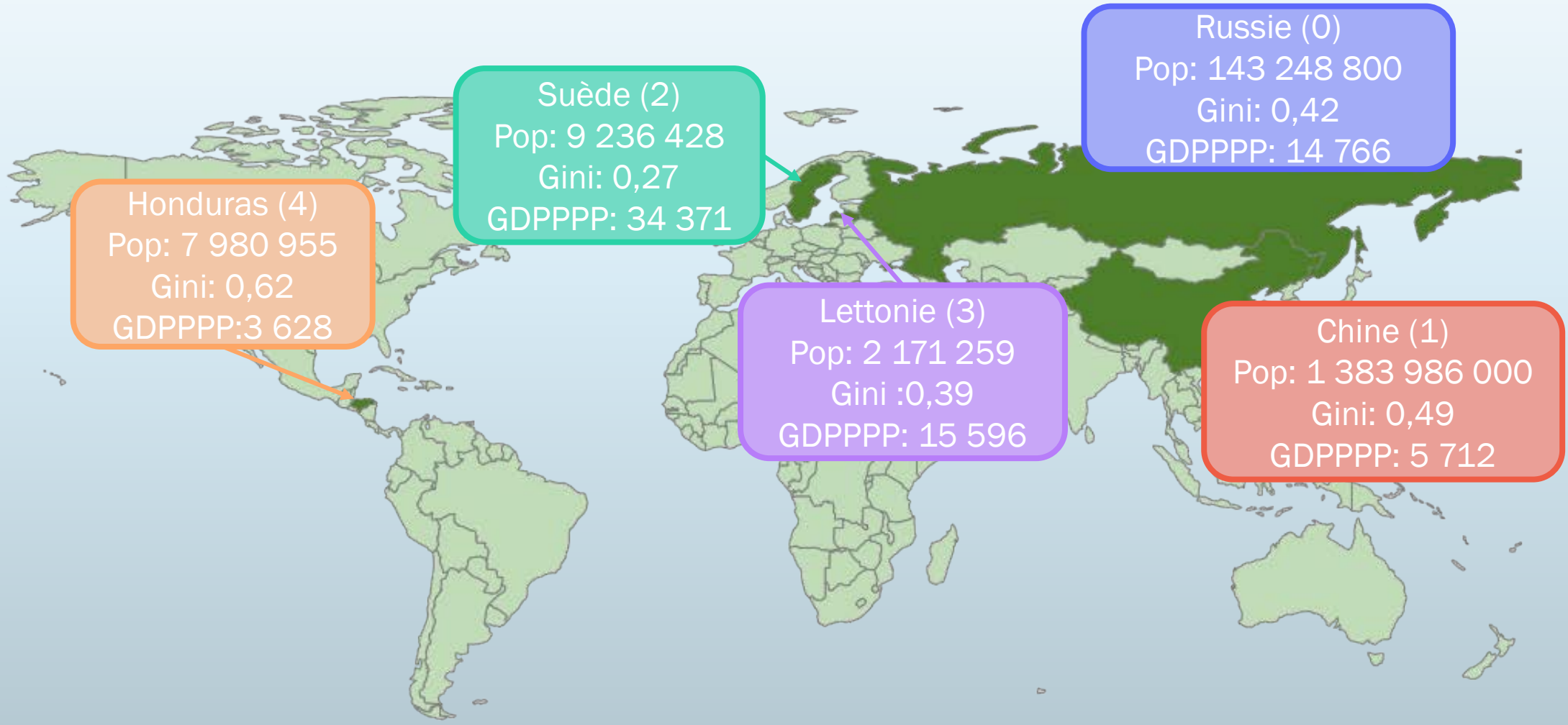


Illustration: Diversité de revenus

- Les plus pauvres de suède gagnent plus que 60% des Chinois et des Honduriens
- Les plus riches centiles chinois n'ont que le niveau de la classe moyenne suédoise
- 90% des Russes et lettons sont plus riches que les plus pauvres suédois

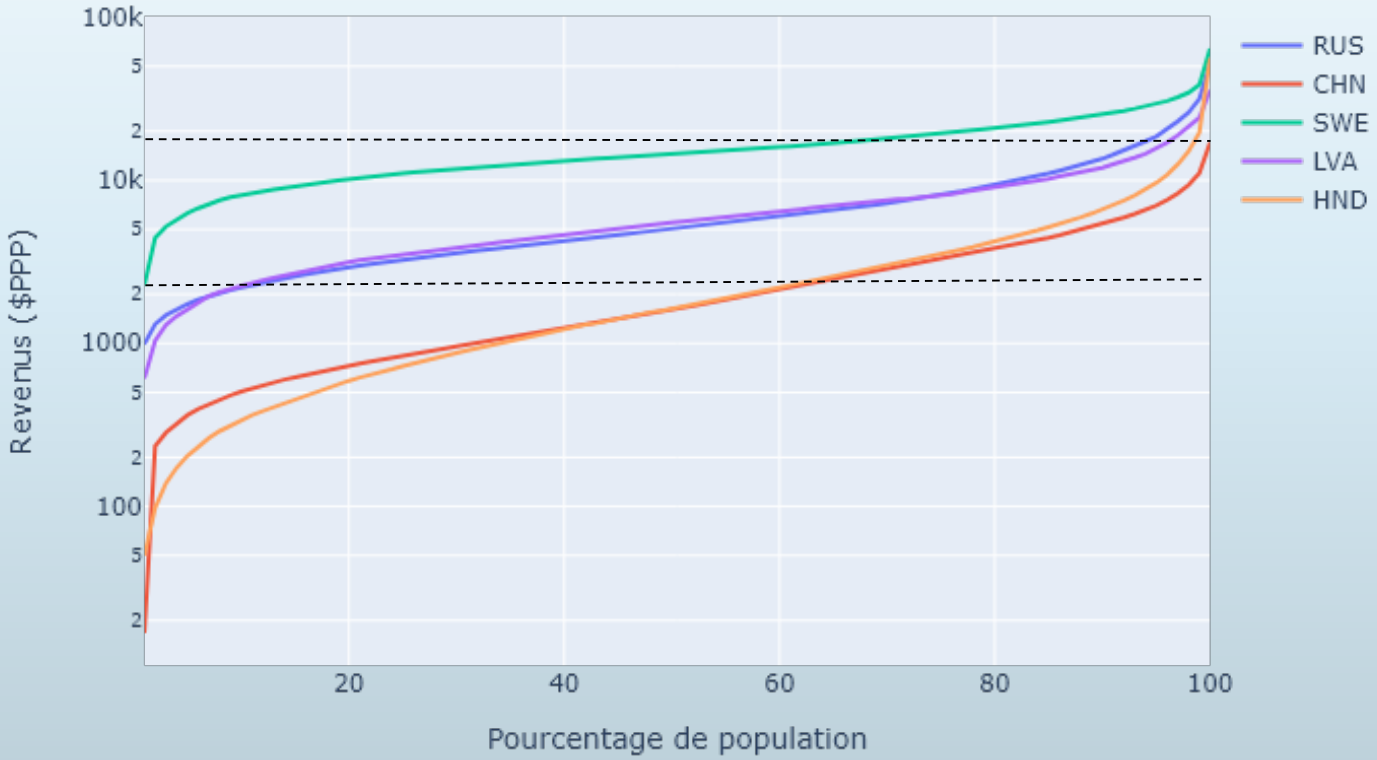


Illustration: Courbe de lorenz

- Pays très inégalitaires : Honduras
- Pays tendent à l'égalité : Suède

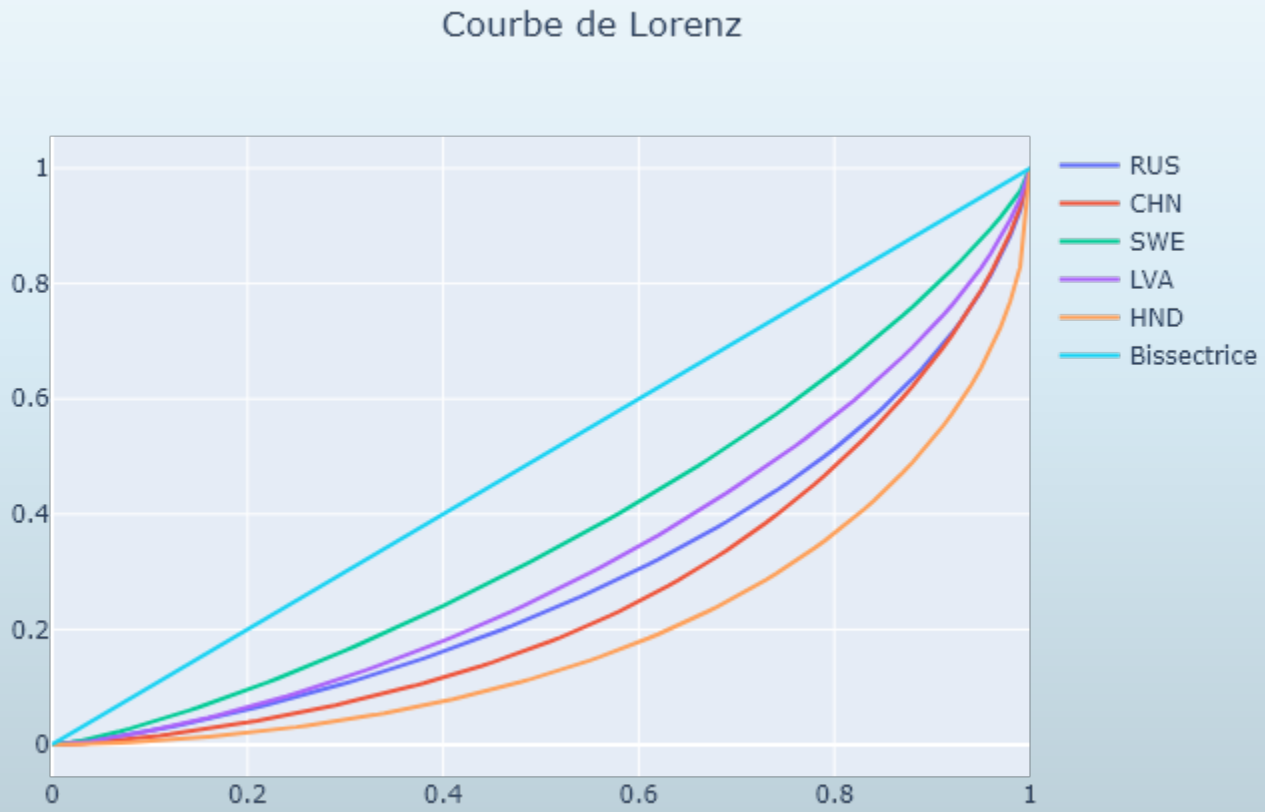
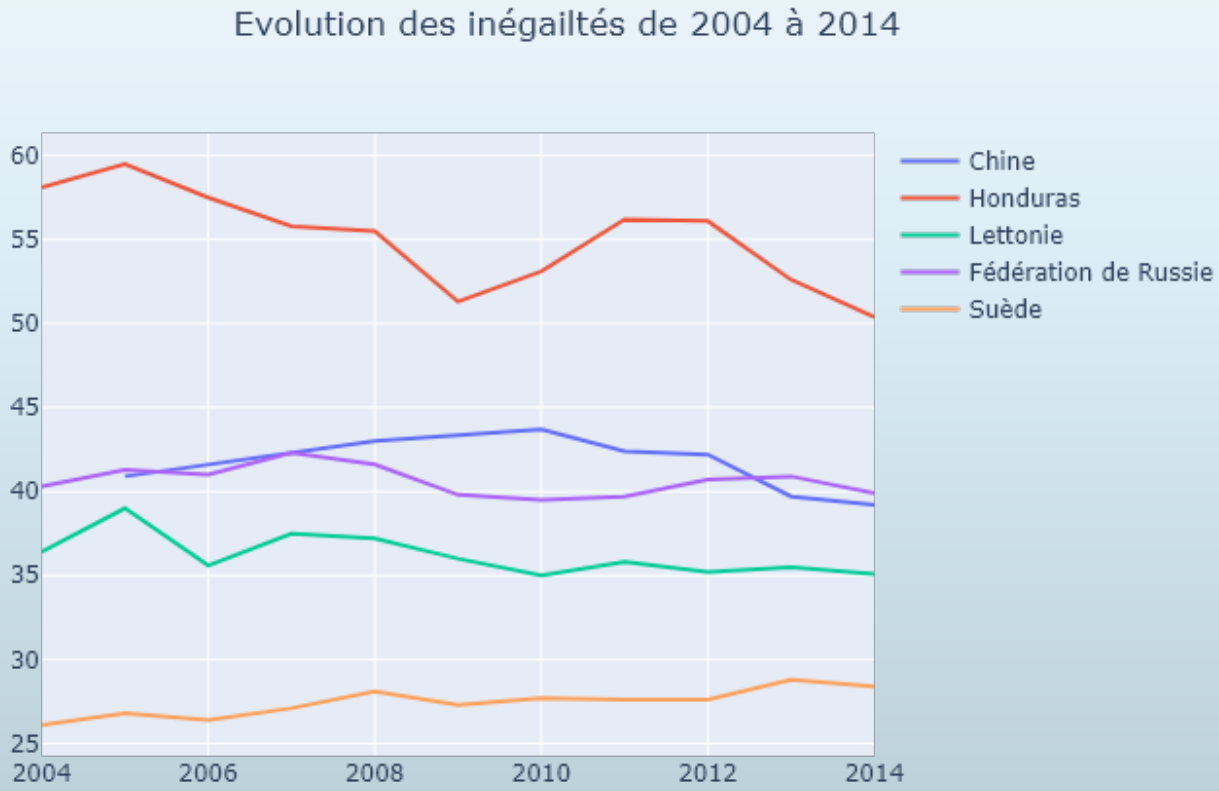


Illustration: évolution des inégalités de distribution

- Tendance générale à la baisse
- Exception avec la Suède

- Chine: Développement économique rapide , apparition de nombreuses classes aisées
- Honduras : Très pauvre, violence élevée
- Lettonie : pas de contrainte avec les hauts revenus, nouveau système d'imposition
- Russie: idée de régulation en 1992 mais pas appliquée
- Suède : mentalité d'égalitarisme, impossible de marginalité sociale avec le climat



Les extrêmes de l'inégalité

5 pays les moins égalitaires		Gini moyen Mondial	5 pays les plus égalitaires	
Afrique du Sud	0.683		Slovénie	0.248
Honduras	0.616		Slovaquie	0.265
Colombie	0.583	0.395	Tchéquie	0.270
Guatemala	0.582		Ukraine	0.272
Centre-Afrique	0.576		Suède	0.272
Position de la France			Gini de la France	
40ème			0.346	

Conclusion

- Diversités :
  - *Revenus*
  - *Distributions du revenus*
- 2 variables 3 :
  - *Revenus moyen du pays*
  - *Indice de Gini*

- Calcul de la dernière variable: le classe de revenu des parents

DÉTERMINATION DE LA CLASSE DE REVENU DES PARENTS

Mission 3

Démarche

1 Créations des individus

2 Affiliation de la classe de revenu du parent à chaque individu en fonction des PROBABILITÉS CONDITIONNELLES

Probabilités conditionnelles calculées en se basant sur l'équation :
$\ln(Y_{child}) = \alpha + \beta_1 \ln(Y_{parent}) + \varepsilon$

- Ychild/parent: classe de revenu
- α: négligeable pour le résultat
- B1: Coefficient d'élasticité
- ε: Résidus

a Génération des revenus parents, puis enfants (avec l'équation)

b Attribution des classes de revenus

c Calcul des probabilité conditionnelles

Générations des revenus

Génération aléatoire des revenus des parents

Calcul du revenus des enfants :

$$e^{(\beta 1 * \ln(Yparent) + \varepsilon)}$$

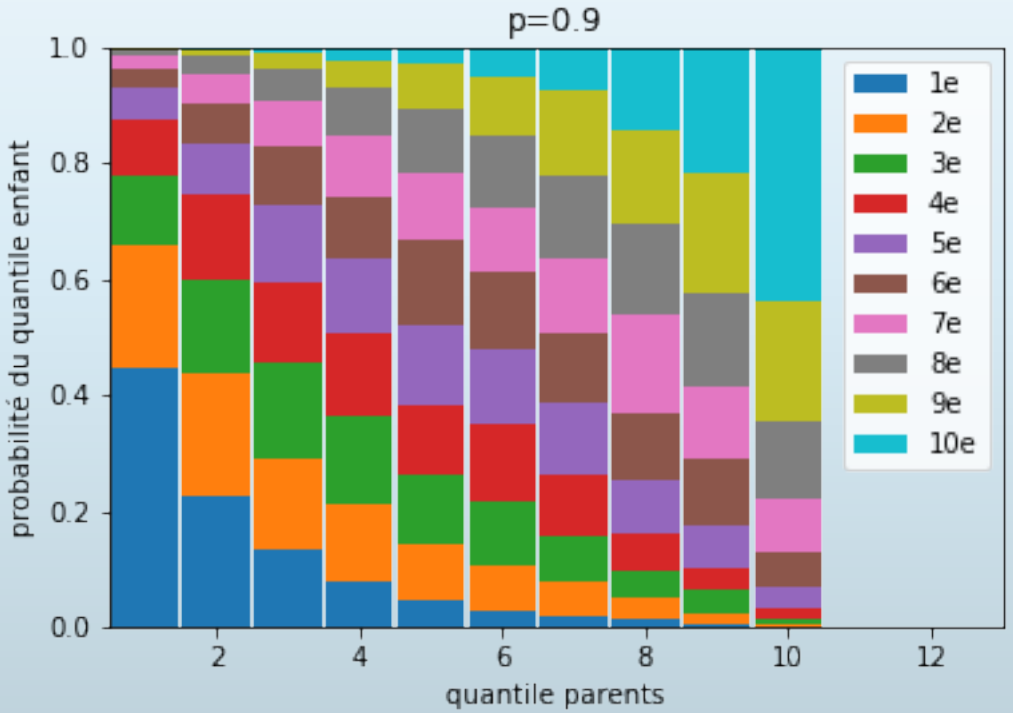
Affiliation de classes de revenus (quantiles)

Calcul des probabilités conditionnelles

Pour chaque quantile de revenus des enfants

Compte le nombre de parents issu des chaque classes de revenus

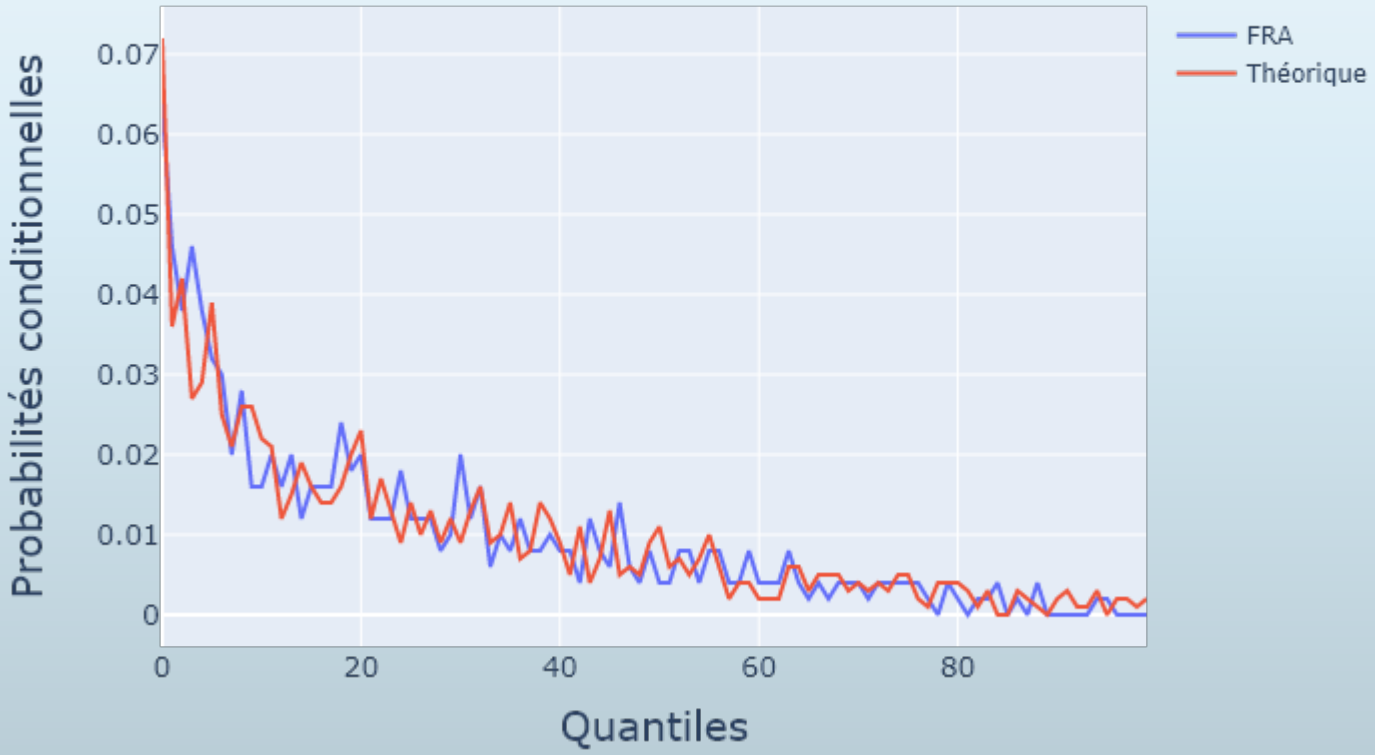
Calcul des probabilités conditionnelles



Vérification du DataFrame

- Pas identique : génération aléatoire des revenus pour calculer les probabilités conditionnelles

Probabilité conditionnelles du quantile 1
France dataframe final vs Génération théorique avec le même coefficient d'élasticité



Conclusion

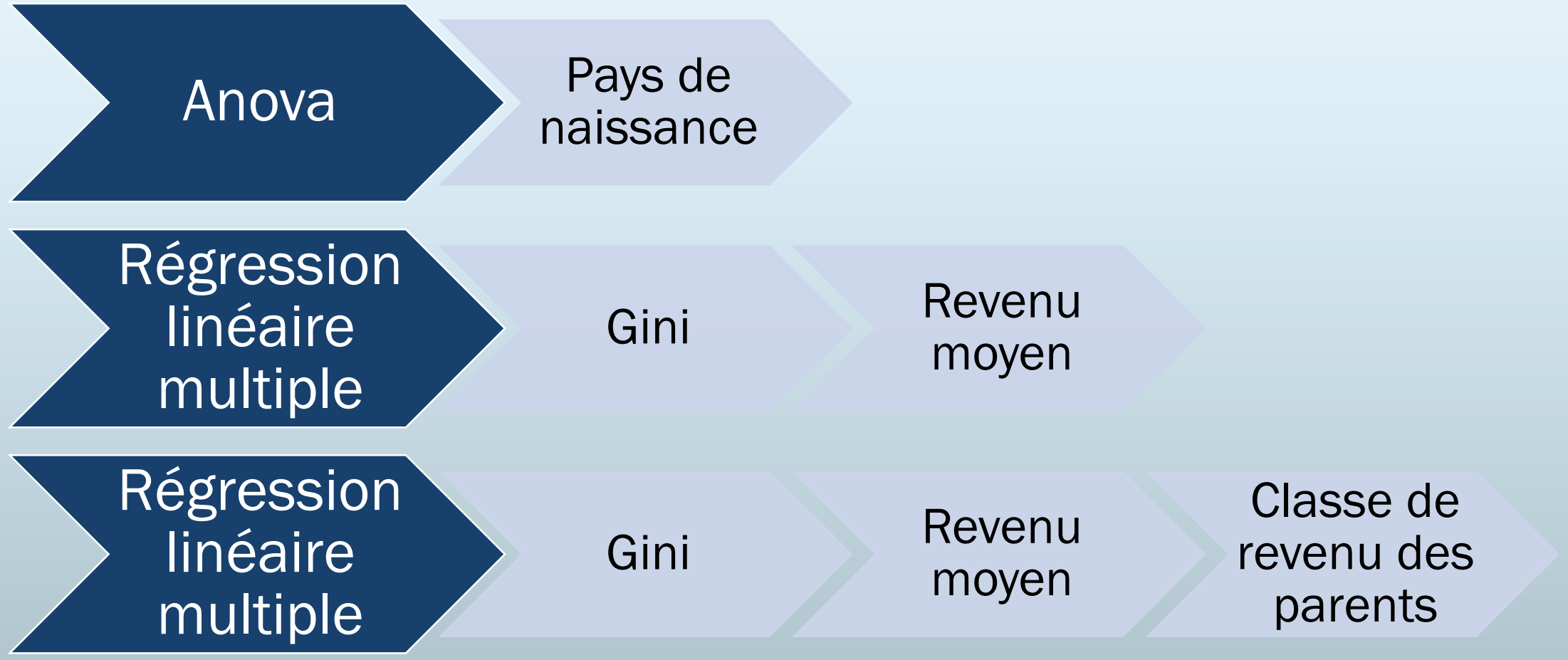
- 3 variables explicatives :
  - *Gini*
  - *Revenu des parents*
  - *Revenu moyen du pays*
- Quelles sont les variables influençant réellement le revenu de l'enfant ?

FACTEURS INFLUENÇANT LE REVENU D'UN INDIVIDU

Mission 4

Démarche

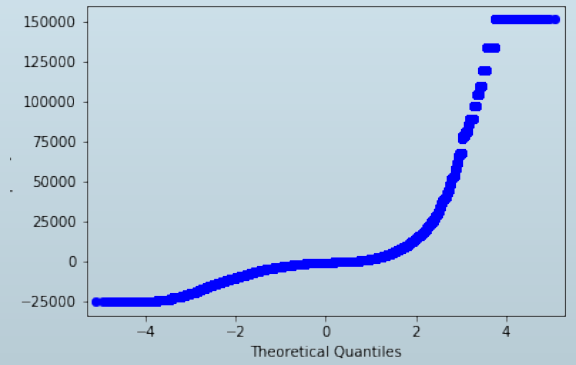
Corrélation du revenu avec :



Anova: Influence du pays

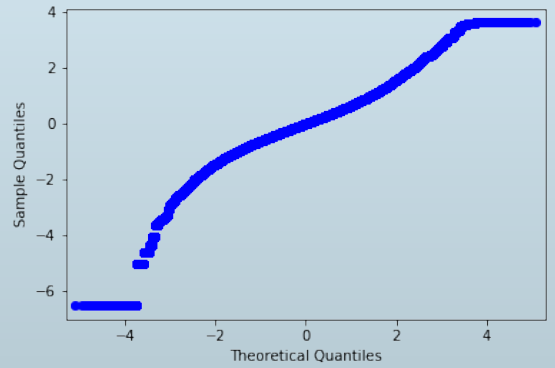
Sans log

- Décomposition de variance totale expliquée : 49 %
- Non-validité: pas de normalité des résidus



Avec log

- Décomposition de variance totale expliquée: 72 %
- Validité: normalité des résidus



Anova: Influence du pays

- Pvalue <0,05
- Le pays influence grandement le revenu que l'enfant va avoir

Régression linéaire multiple Revenu moyen et Gini

- Pvalue <0,05
- Décomposition de variance totale expliquée
  - *Caractéristiques du pays : 65%*
  - *Autres facteurs 35%*
- Individus défavorisés dans les pays avec un Gini élevé (coef Gini négatif)

OLS Regression Results

Dep. Variable:	np.log(income)	R-squared:	0.652
Model:	OLS	Adj. R-squared:	0.652
Method:	Least Squares	F-statistic:	5.436e+06
Date:	Tue, 10 Nov 2020	Prob (F-statistic):	0.00
Time:	09:51:46	Log-Likelihood:	-7.0403e+06
No. Observations:	5800000	AIC:	1.408e+07
Df Residuals:	5799997	BIC:	1.408e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7956	0.003	235.953	0.000	0.789	0.802
np.log(gdpppp)	0.8667	0.000	3019.161	0.000	0.866	0.867
gini	-1.4689	0.004	-366.988	0.000	-1.477	-1.461

Omnibus:	199687.168	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	542841.602
Skew:	-0.110	Prob(JB):	0.00
Kurtosis:	4.483	Cond. No.	128.

Régression linéaire multiple
Revenu moyen, Gini et revenu des parents

- Décomposition de variance totale expliquée
  - *Sans log: 47 %*
  - *Avec log: 70 %*

- Pvalue <0,05

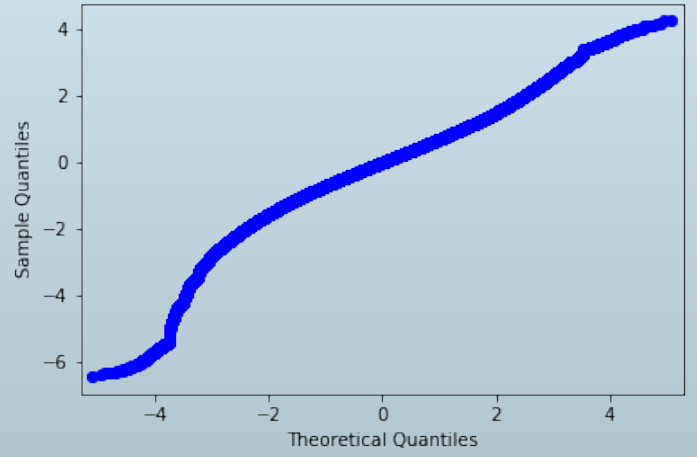
- Décomposition de variance totale expliquée
  - *Caractéristiques du pays et revenu des parents: 70%*
  - *Autres facteurs 30%*

Validité de la modélisation

- Théorème de Gauss-Markov
  - *Espérance des résidus* = 2.79e-12
  - *Erreurs non corrélées (VIF <5 ou 10)*

Gdpppp	Gini	Classe de revenu des parents
1.097	1.097	1.000

- Normalité des résidus :



Conclusion

- Revenus moyen du pays , Gini et la classe de revenu du parents expliquent à 70 % le revenu de l'enfant
- Pour avoir des nouveaux clients à haut revenu il faudra privilégier :
  - *Pays à haut revenu*
  - *Pays avec un Gini bas*
  - *Classe de revenu des parents haut*