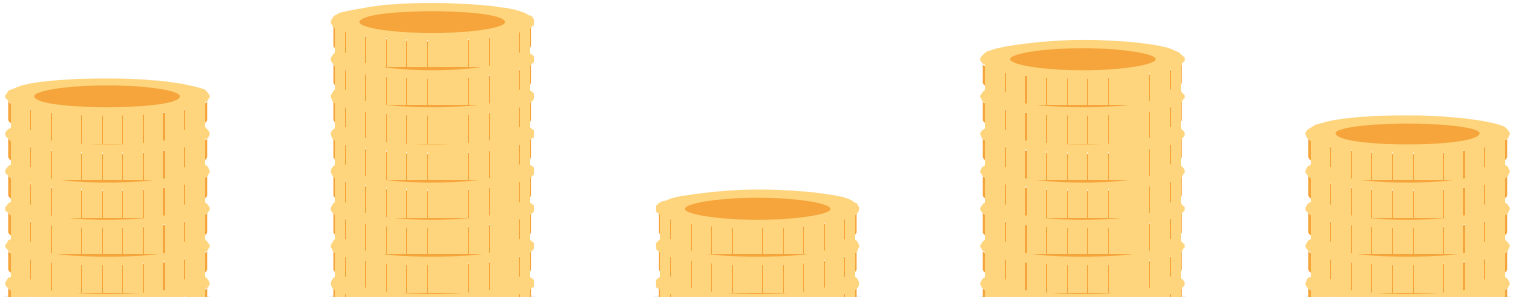
# - Banque -

Prédiction des revenus des enfants des clients



# Présentation du projet et enjeux spécifiques

01 Hauts revenus

02 International

# Avec quelles données ?

## World Income Distribution

- Revenu des classes de parents
- Revenu moyen du pays (gdp ppp)
- Indice de Gini (calcul)

## Banque Mondiale

- Indices de Gini

# SOMMAIRE

**Mission 1** — *Nettoyage*

**Mission 2** — *Analyse graphique*

**Mission 3** — *Algorithme*

**Mission 4** — *Analyse des contributions des variables*

# Mission 1

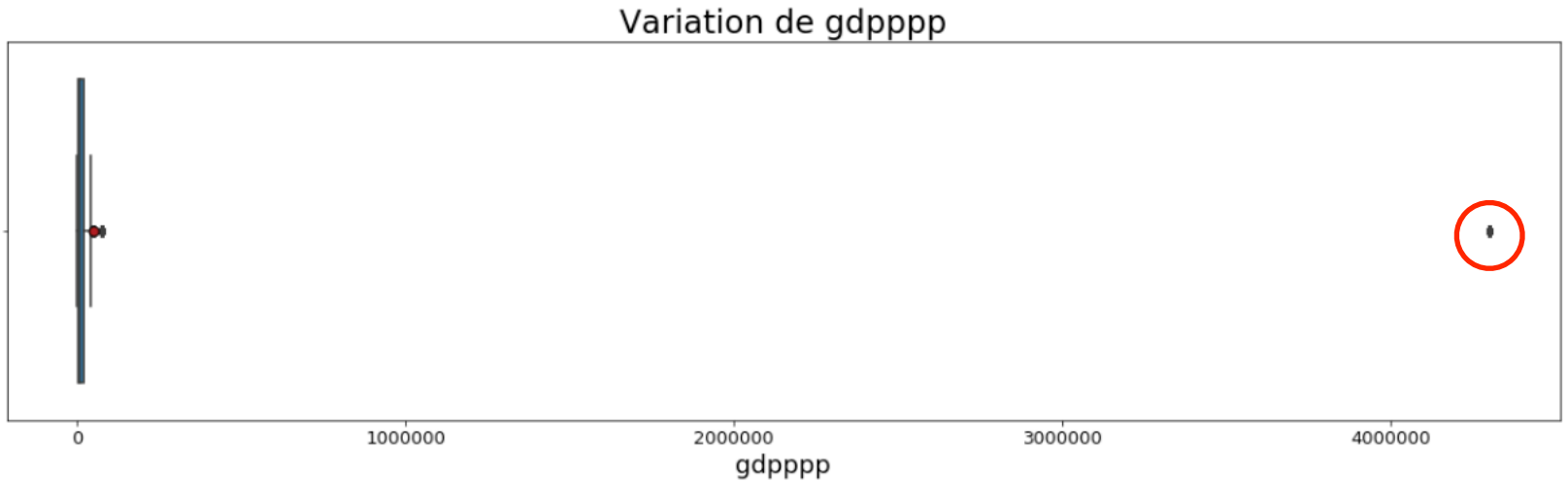
*Nettoyage*

# Problèmes sur les données

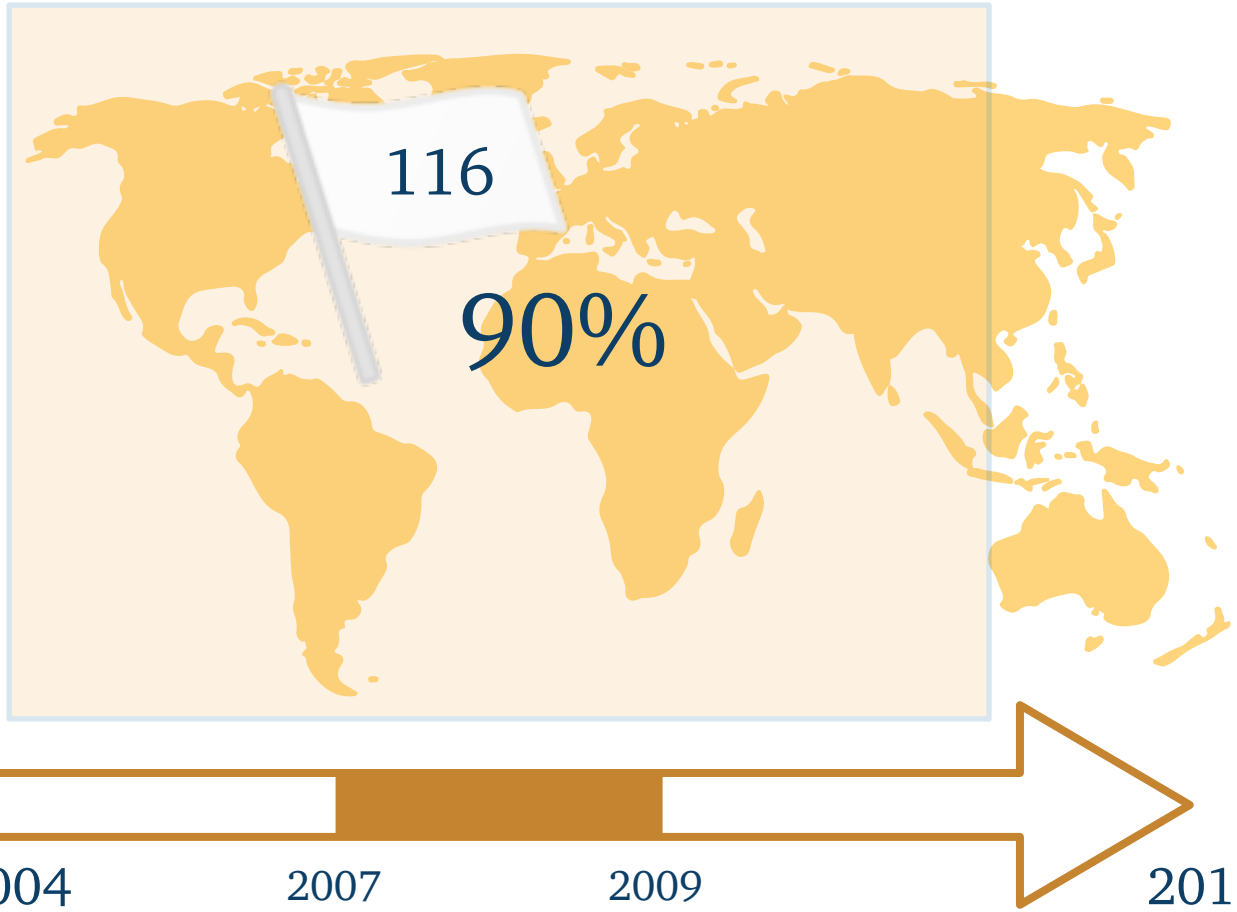
1 ligne manquante

200 NaN (gdpppp)

# Valeurs aberrantes



# Observations sur les données



# Mission 2

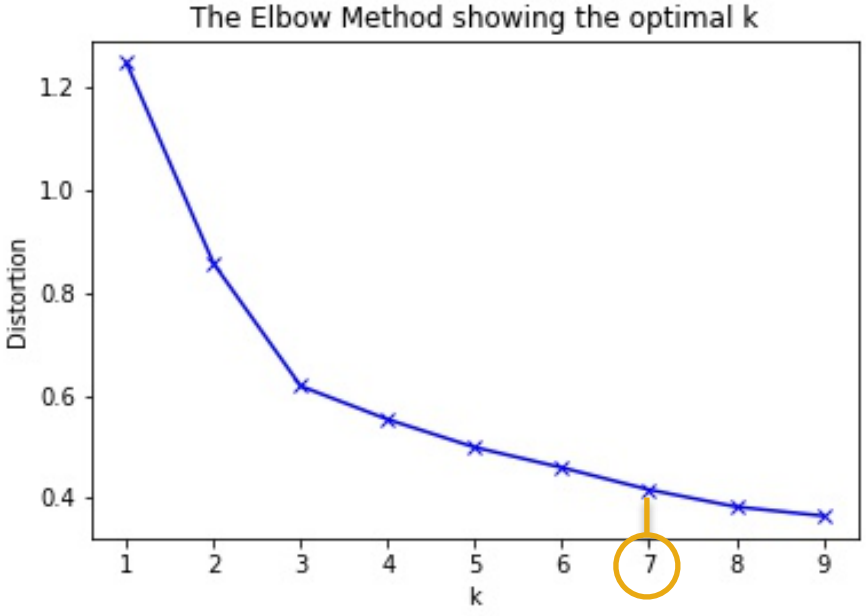
*Montrez la diversité des pays en termes de distribution de revenus*

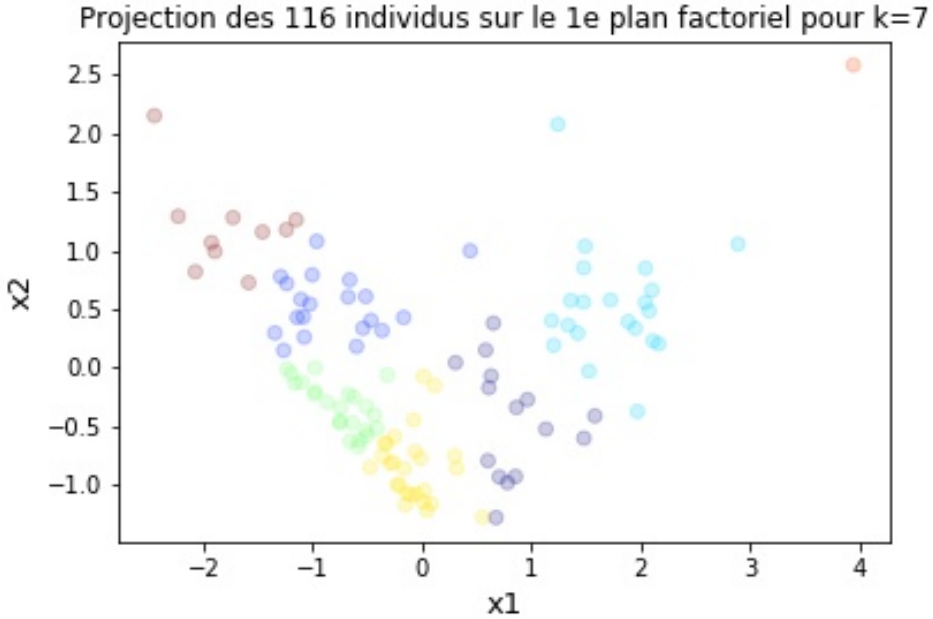

116

# Sélection des données pertinentes

| pays | population | gdpppp | gini | income |
|---|---|---|---|---|
| IDN | 238620563.0 | 3689.0 | 0.372115 | 425.47750 |
| LUX | 485405.0 | 73127.0 | 0.292935 | 58382.312 |
| … | … | … | … | … |
| … | … | … | … | … |

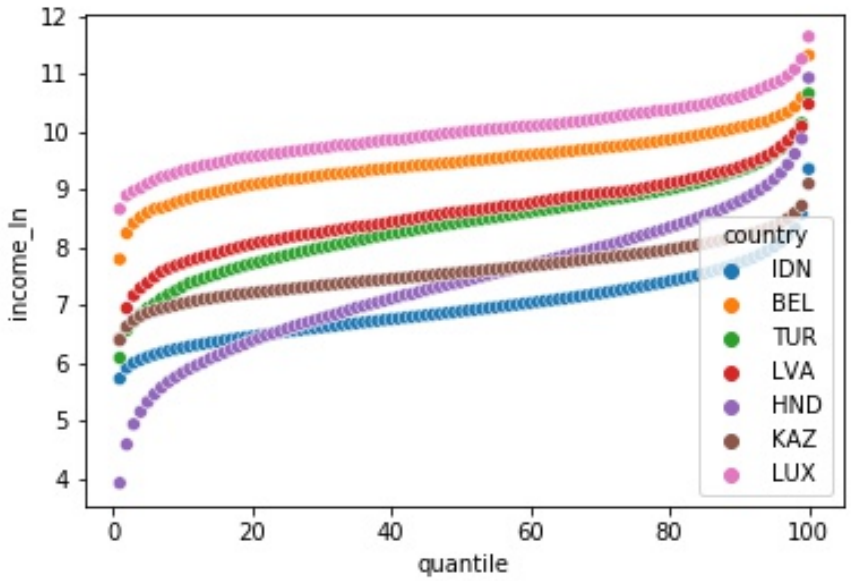
# Clustering


The Elbow Method showing the optimal k
Distortion
k
1.2
1.0
0.8
0.6
0.4
1
2
3
4
5
6
7
8
9

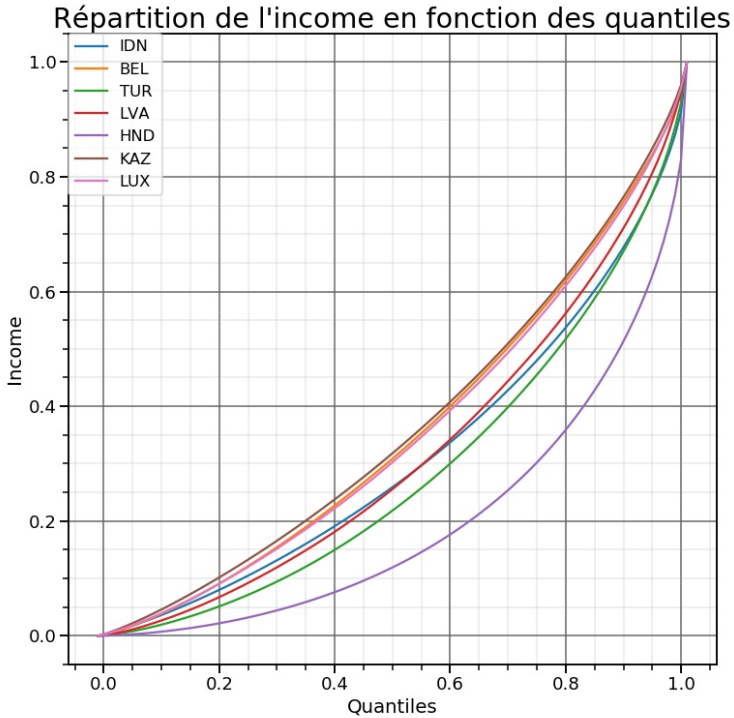

Projection des 116 individus sur le 1e plan factoriel pour k=7
x2
x1
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
3
4

# Observations graphiques

## Diversité des pays en termes de distribution des revenus

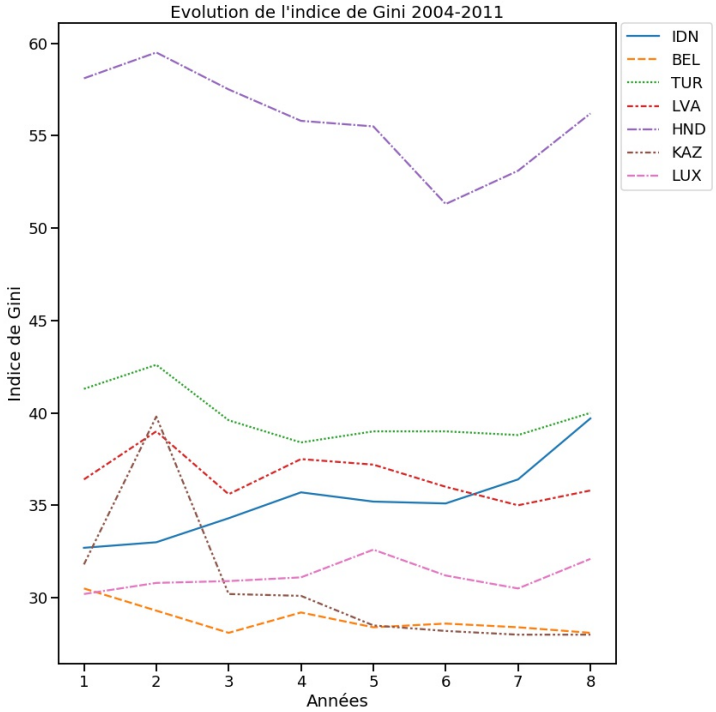



## Courbe de Lorenz des pays sélectionnés

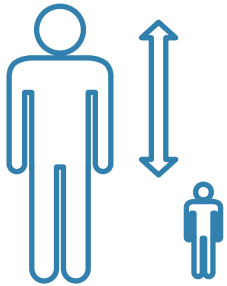


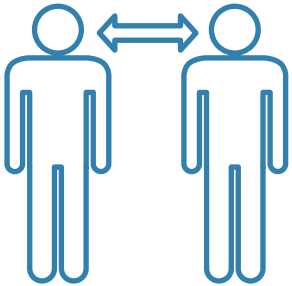
# Observations graphiques 2/2

## Evolution de l'indice de Gini 2004-2011



# Classement indice de Gini



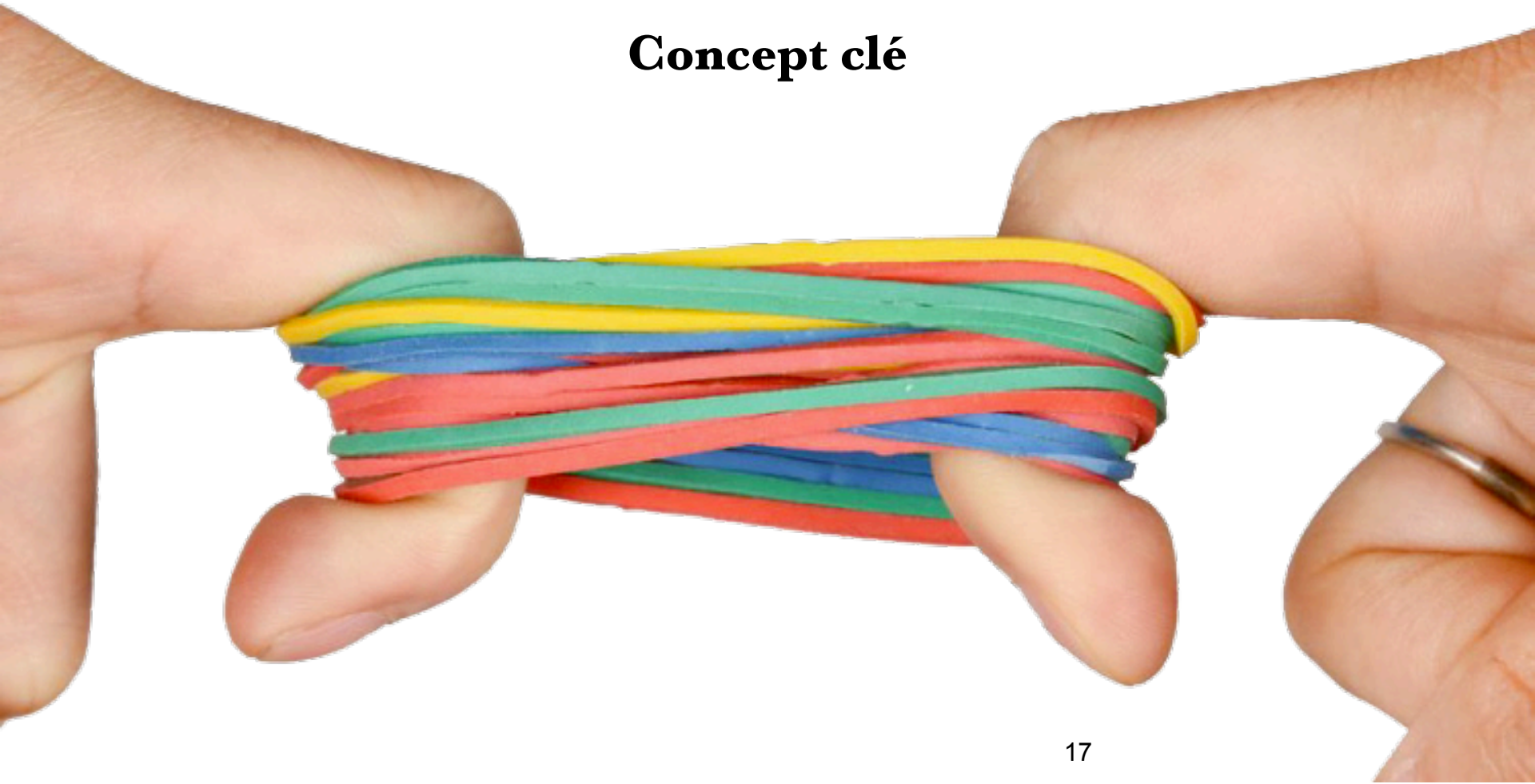


38

# Mission 3

*Algorithme*

# Concept clé



# Avec quelles données ?

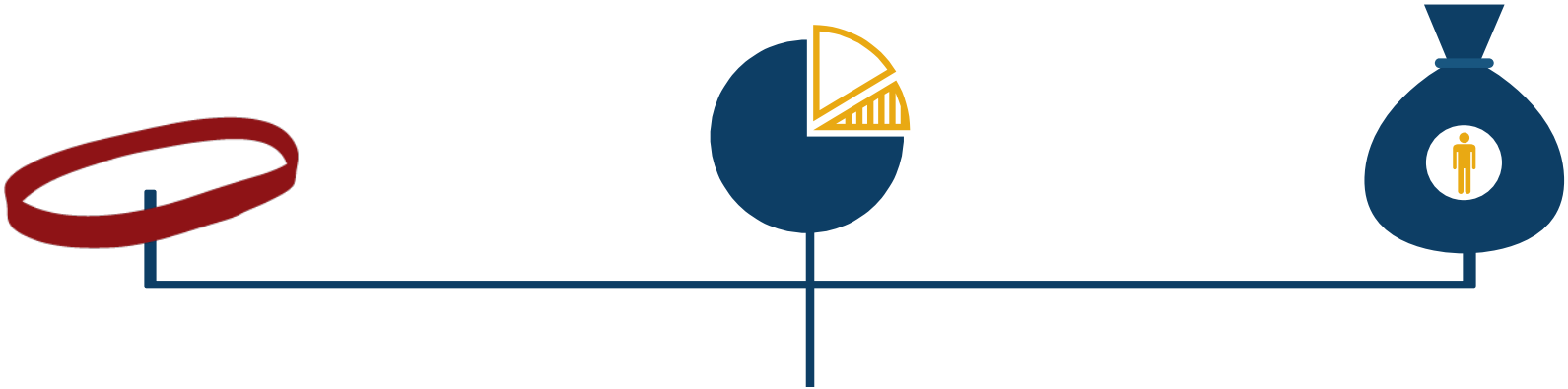
Banque Mondiale

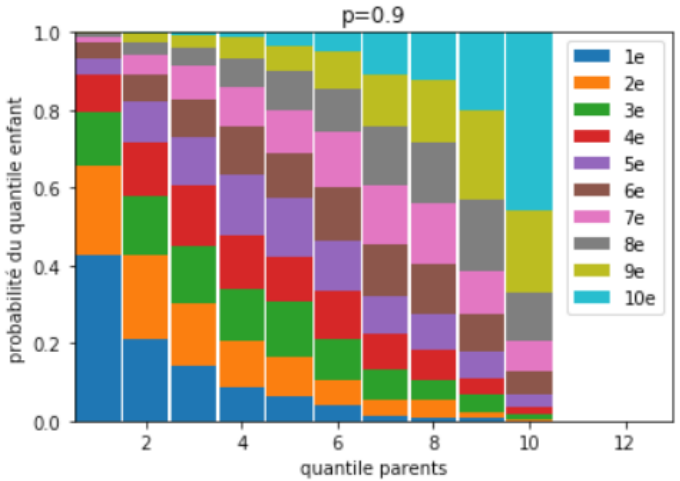
*Données manquantes*

Fichier elasticity

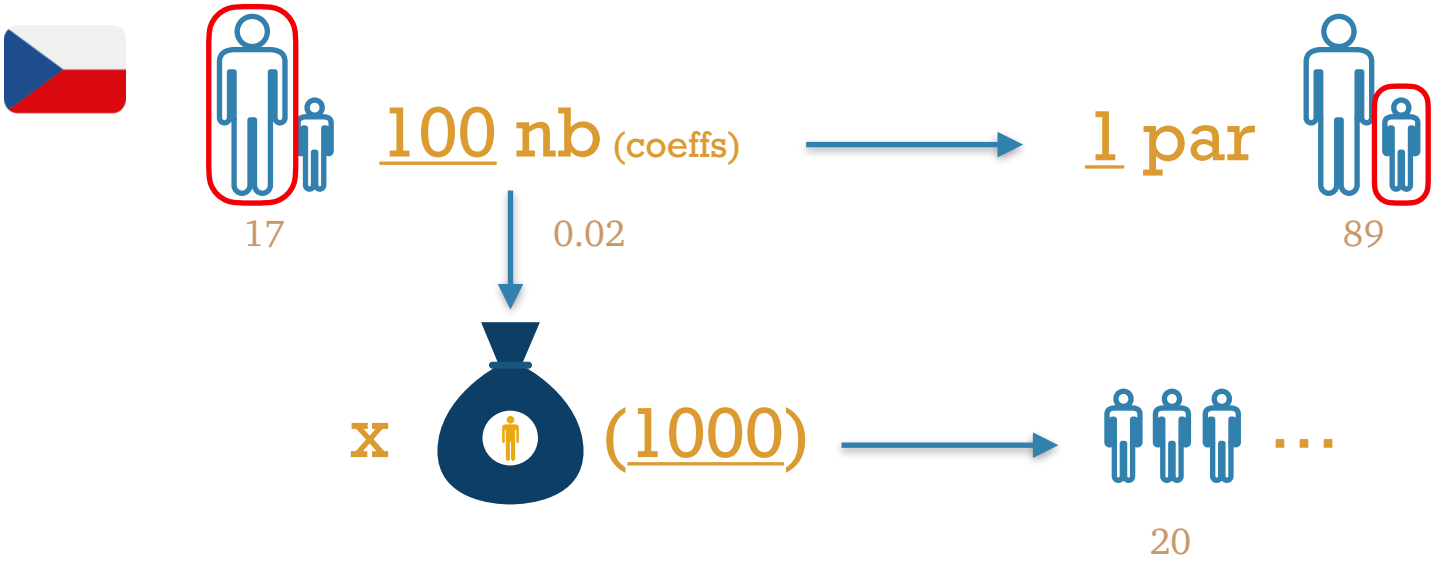
*Données très génériques*

# Fonctionnement de l'algorithme

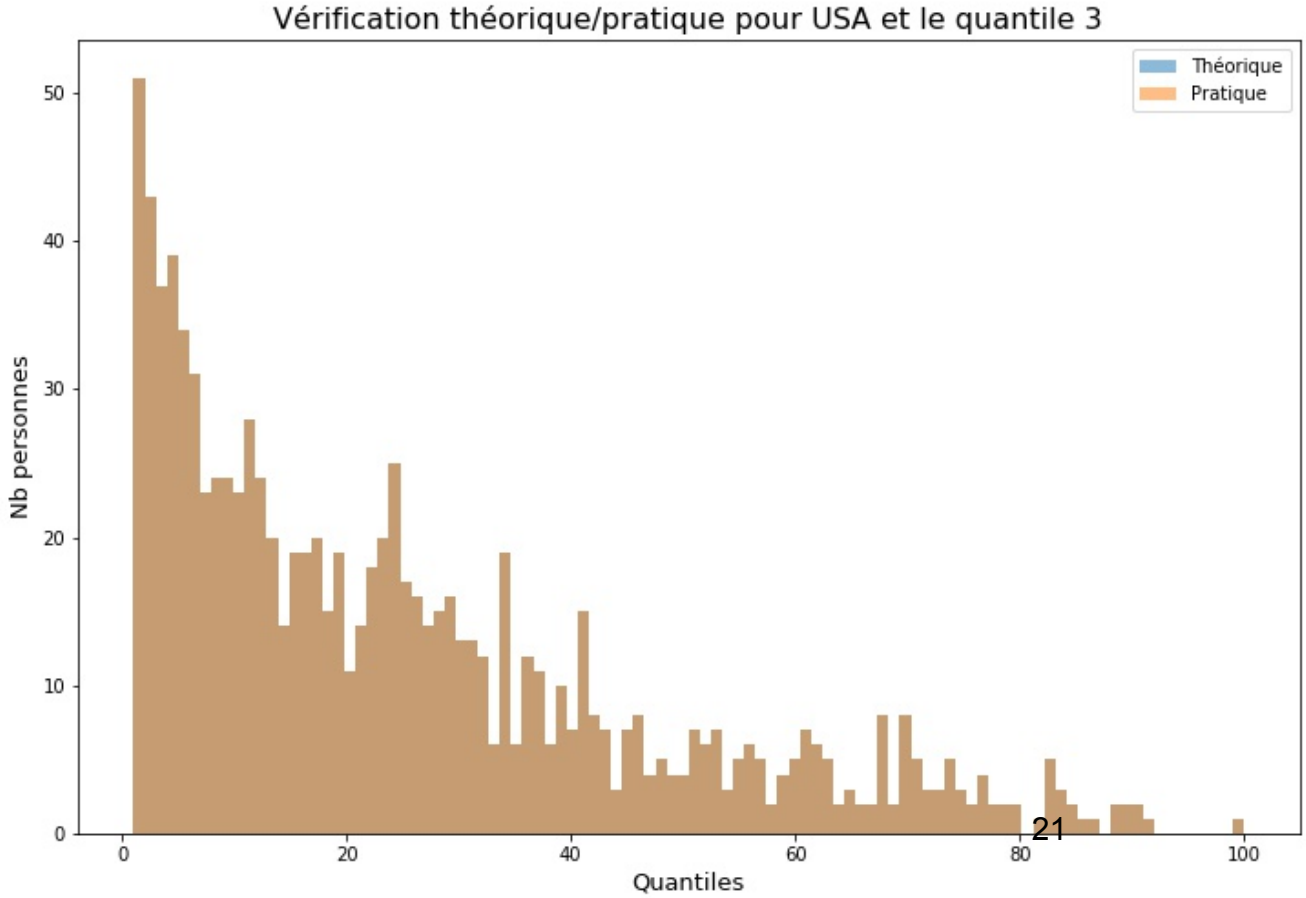




# Attribution à des individus



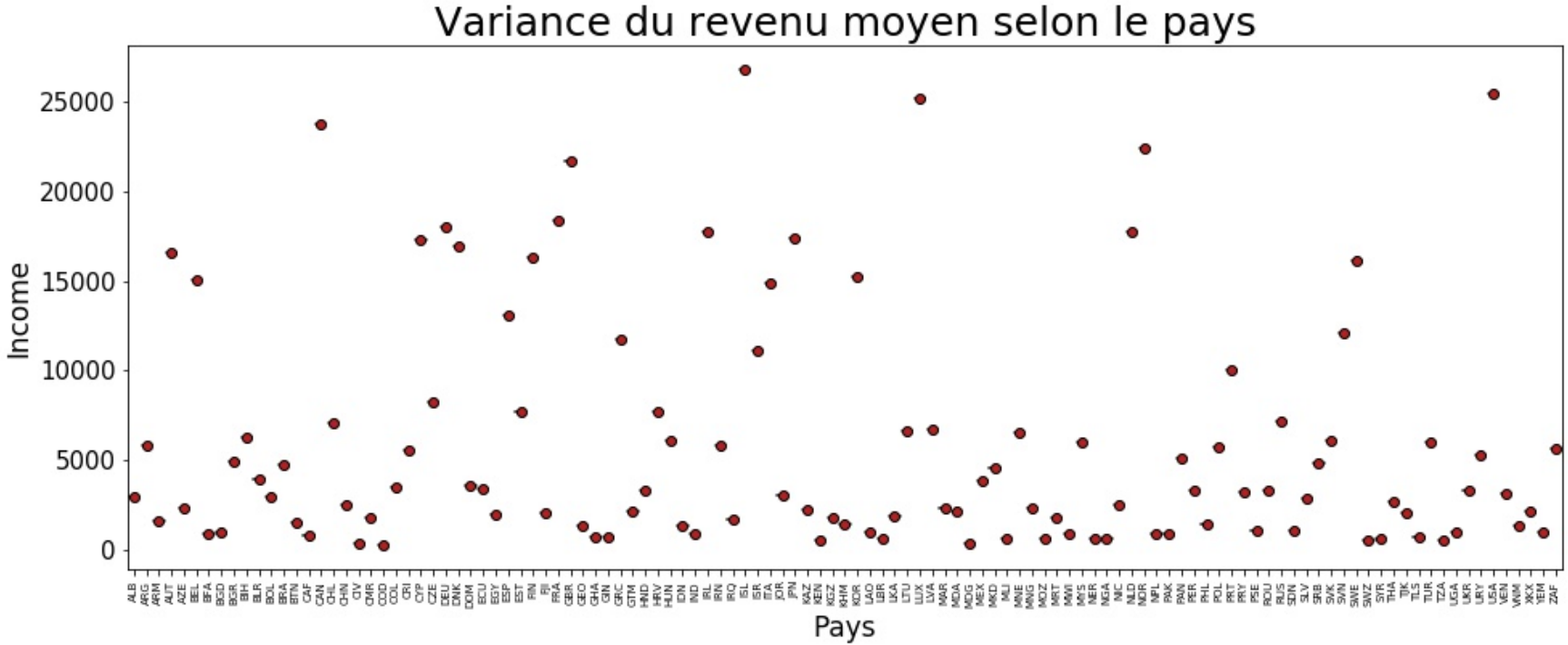
# Vérification de l'algorithme



# Mission 4

*Expliquer le revenu des individus en fonction de plusieurs variables explicatives.*

# ANOVA paramétrique




**Shapiro** - pvalue > 5%

# ANOVA non paramétrique

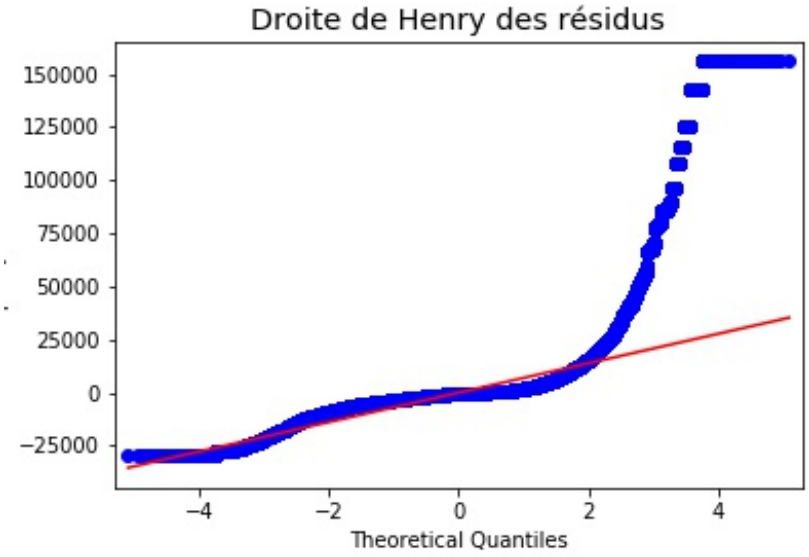
**Kruskal-Wallis** - pvalue < 5% 

# Régressions linéaires : sans la classe parent

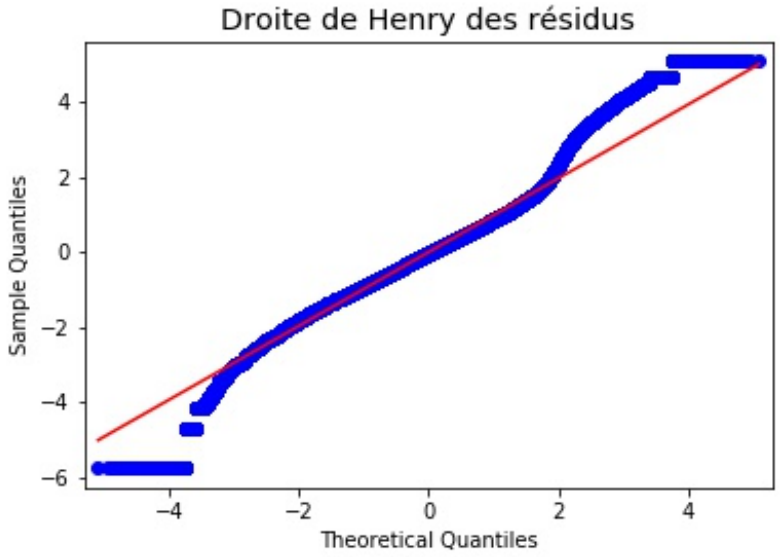
| | Sans log | Avec log |
|---|---|---|
| $R^2$ | 0.448  | 0.492  |
| Normalité (*Shapiro*) | ? | ? |
| Homoscédasticité (*Levene*) |  |  |
| Indépendance (*Durbin-Watson*) | 1.8  | 1  |

# Normalité : Droites de Henry

## Sans log




## Avec log

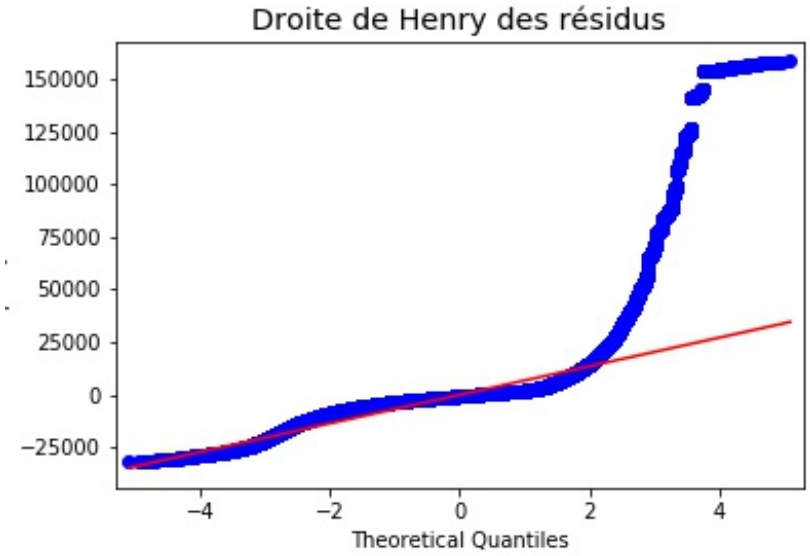


# Régressions linéaires avec la classe parent

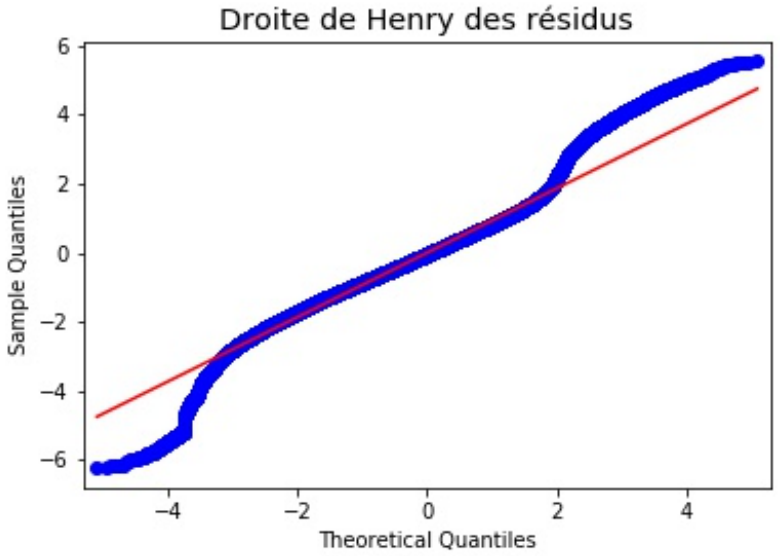
| | Sans log | Avec log |
|---|---|---|
| $R^2$ | 0.471 | 0.541 |
| Normalité (*Shapiro*) | ? | ? |
| Homoscédasticité (*Levene*) | | |
| Indépendance (*Durbin-Watson*) | 1.8 | 1 |

# Normalité : Droites de Henry

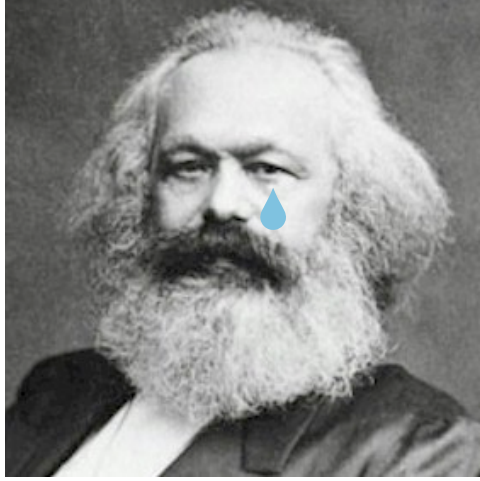
## Sans log




## Avec log



# CONCLUSION



🏳️ + 🧍 = 55%

🧍 = 5%

45% 💪 🍀 💋 🤝

# Recommandations

1

Gdpppp élevé
Gini faible

2

Gdpppp élevé
Gini moyen

3

Gdpppp très élevé
Gini moyen/fort

**Merci !**

**Avez-vous des questions ?**