Réaliser une prédiction de revenus

Analyse des données

Création d'un modèle

Modèle prédictif de revenus

SOUHAITE CIBLER DE NOUVEAUX JEUNES CLIENTS

EN CIBLANT DE FUTUR PERSONNES AVEC DES HAUTS REVENUS

CRÉER UN MODÈLE POUR DÉTERMINER LE REVENU POTENTIEL D'UNE PERSONNE

EMPLOYÉ DANS UNE BANQUE

Analyse des données

Données de la World Income distribution :
-Pays, quantile (centile), revenu (PPP), PIB

Données de la banque mondiale
-Indice de Gini le plus récent sur les quarante dernières années, et la médian des ginis
-IGE (relative IGM in income) qui correspond aux coefficients d'élasticités des revenus

Fao : population par pays en 2008, 2013 et 2018

Analyse des données

Donnée manquante

- J'ai réalisé la moyenne des incomes avec les quantile 40 et 42 pour remplacer cette valeur manquante

6238	LTU	39.0	4802.3680	17571.0	Lithuania	35.7	35.7	3198231.0
6239	LTU	40.0	4868.4507	17571.0	Lithuania	35.7	35.7	3198231.0
6240	LTU	42.0	4895.8306	17571.0	Lithuania	35.7	35.7	3198231.0
6241	LTU	43.0	4950.6380	17571.0	Lithuania	35.7	35.7	3198231.0

- L'ensemble des données rassemble plus de 90% de la population totale pour chaque années étudiées.

Analyse des données

- Le coefficient ou l'indice de gini est une mesure statistique permettant de se rendre compte de la répartition d'une variable (salaire/revenus/patrimoine dans notre cas) au sein d'un population

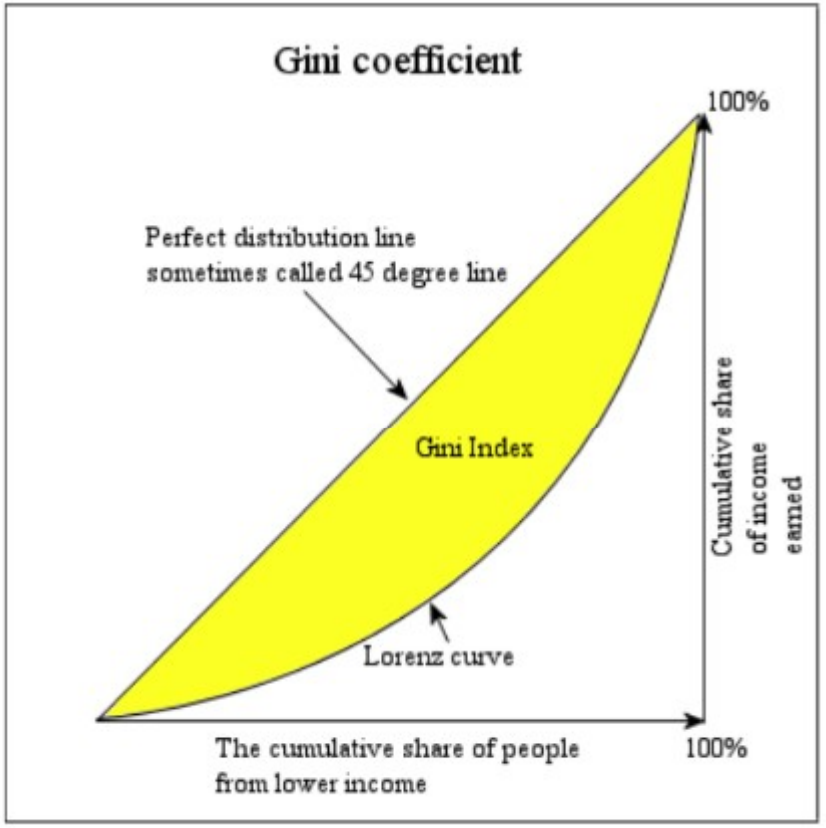


Image 1 : Graphique explicatif de l'indice de Gini et de la courbe de Lorenz

Création de 5 Clusters pour avoir des pays de profils différents

	Country Name	Gini_last	median_gini
0	South Africa	63.0	63.00
1	Central African Republic	56.2	58.75
2	Eswatini	54.6	53.85
3	Mozambique	54.0	50.30
4	Brazil	53.4	57.60

	Country Name	Gini_last	median_gini
107	Moldova	25.7	34.55
108	Belarus	25.3	27.65
109	Czech Republic	25.0	26.20
110	Slovak Republic	25.0	26.30
111	Slovenia	24.6	24.80

```
# mise sous forme de tableau pour créer des cluster
income_country_quantile = pd.pivot_table(data[['income_log', 'Country Name', 'quantile']],
                                         index='Country Name', columns='quantile', values='income_log',fill_value=0)

from scipy.cluster.hierarchy import linkage
linkage_matrix = linkage(income_country_quantile, method='median', metric='euclidean', optimal_ordering=True)

from scipy.cluster.hierarchy import fcluster
groups = fcluster(linkage_matrix, criterion='maxclust', t=5, depth=3)

income_country_quantile['group'] = groups.reshape(-1,1)
```

	Country Name	Gini_last	median_gini
82	France	32.4	32.4

Analyse des données

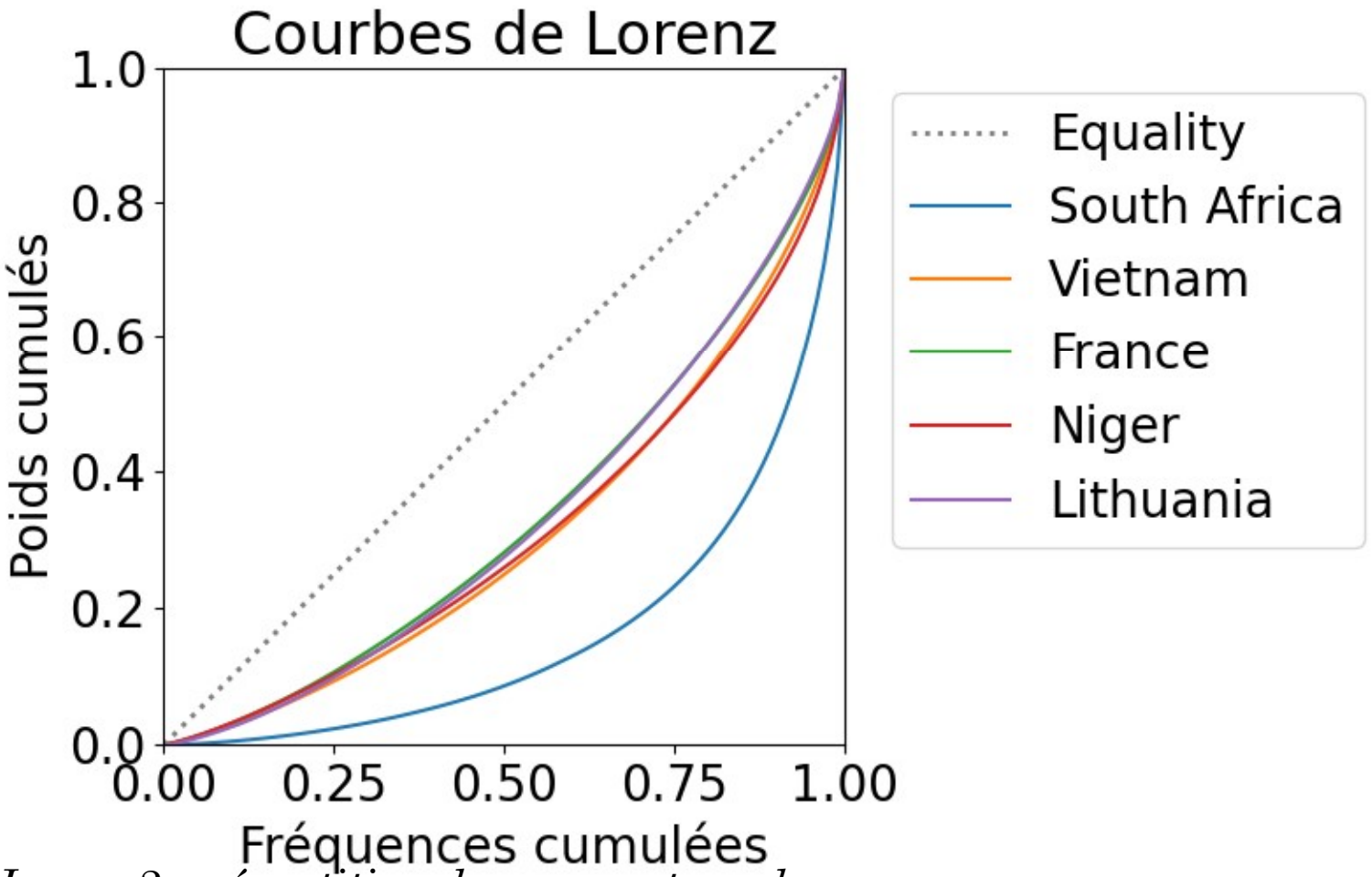


Image 2 : répartition du pourcentage de revenu par pourcentage de la population pour 5 pays avec des courbes de Lorenz

Analyse des données

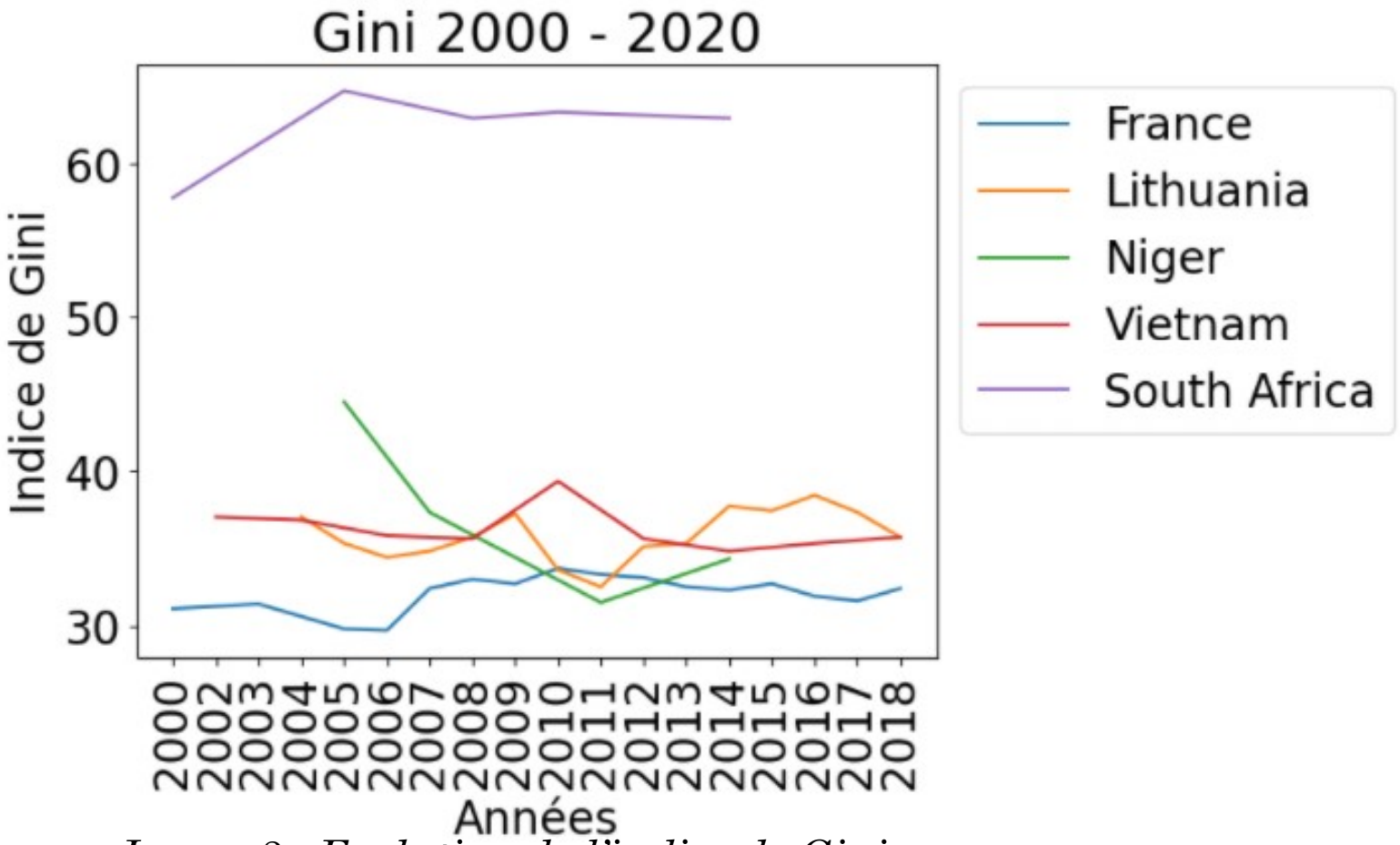


Image 3 : Evolution de l'indice de Gini de 5 pays entre 2000 et 2018

Analyse des données

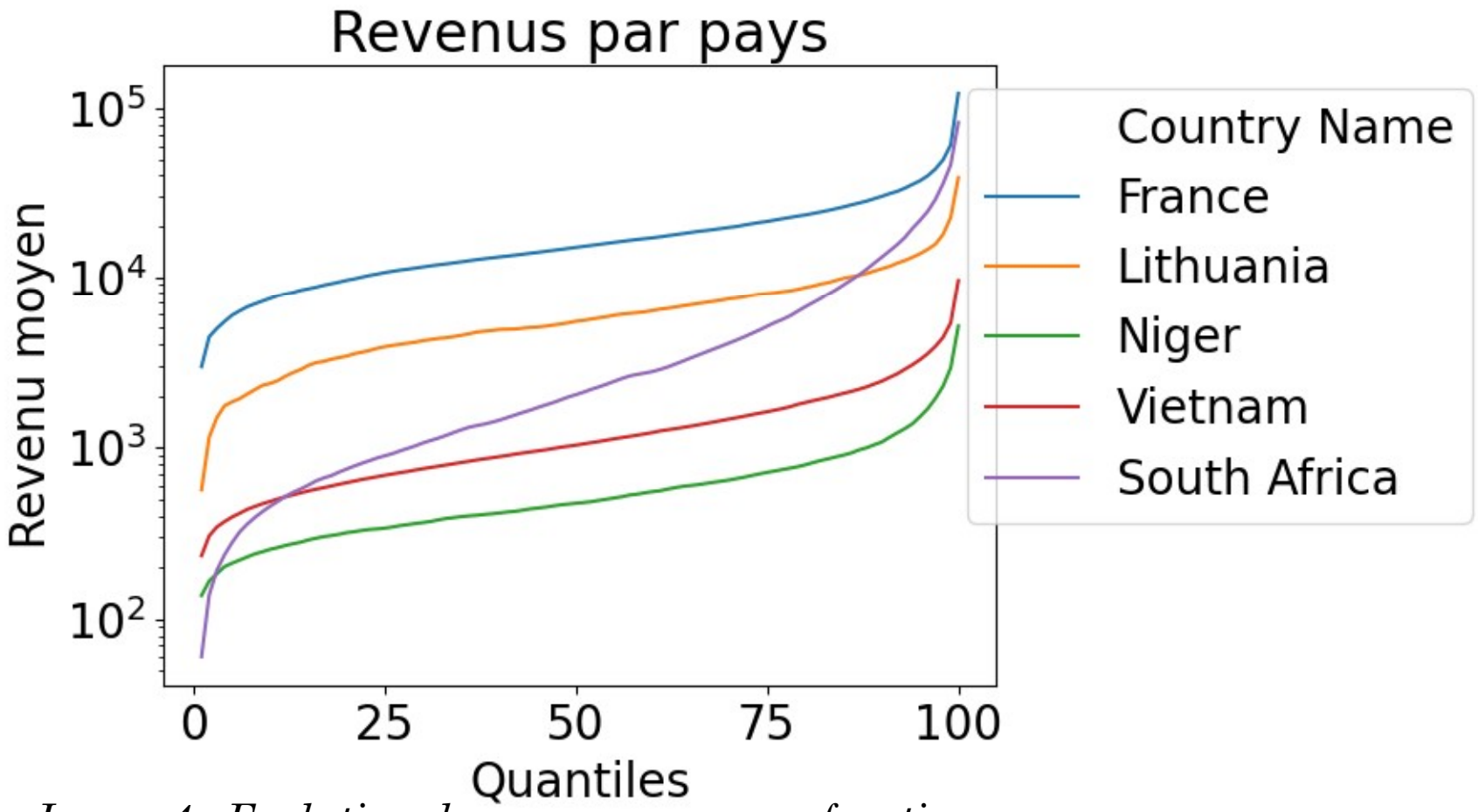


Image 4 : Evolution du revenu moyen en fonction de la classe de revenus pour 5 pays

Création du modèle

```python
def generate_incomes(n, pj): #n = nombre de parent, pj = coefficient d'élasticité des revenus du pays j
    # 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 caclul de la classe de revenu
    ln_y_parent = st.norm(0,1).rvs(size=n)
    # Génération d'une réalisation du terme d'erreur epsilon
    residues = st.norm(0,1).rvs(size=n)
    return np.exp(pj*ln_y_parent + residues), np.exp(ln_y_parent) # return y_child, y_parents (y = income)
```

```python
def quantiles(l, nb_quantiles):# l= nb d'échantillon (parent=enfant)
    #permet d'attribuer les classes de quantile
```

```python
def compute_quantiles(y_child, y_parents, nb_quantiles):  #revenu enfant et parent
    #return les classes de quantile des parents et des enfants
```

```python
def proba_cond(c_i_parent, c_i_child, mat): # determine probabilité que l'enfant soit
#     d'une classe de quantile en fonction de la classe de quantile du parent et du
#     coefficient d'elasticite du pays
```

```python
def plot_conditional_distributions(p, cd, nb_quantiles): # affiche graphiquement la proba
#     de quantile enfant en fonction du quantile parent
```

Création du modèle

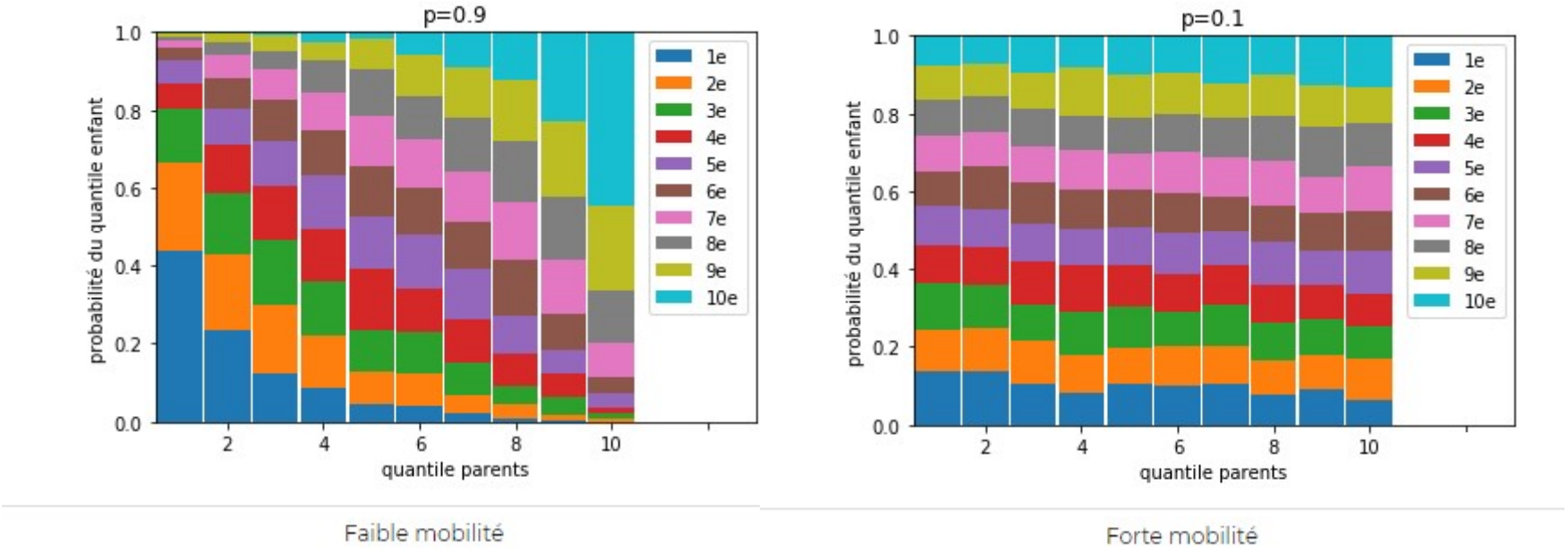
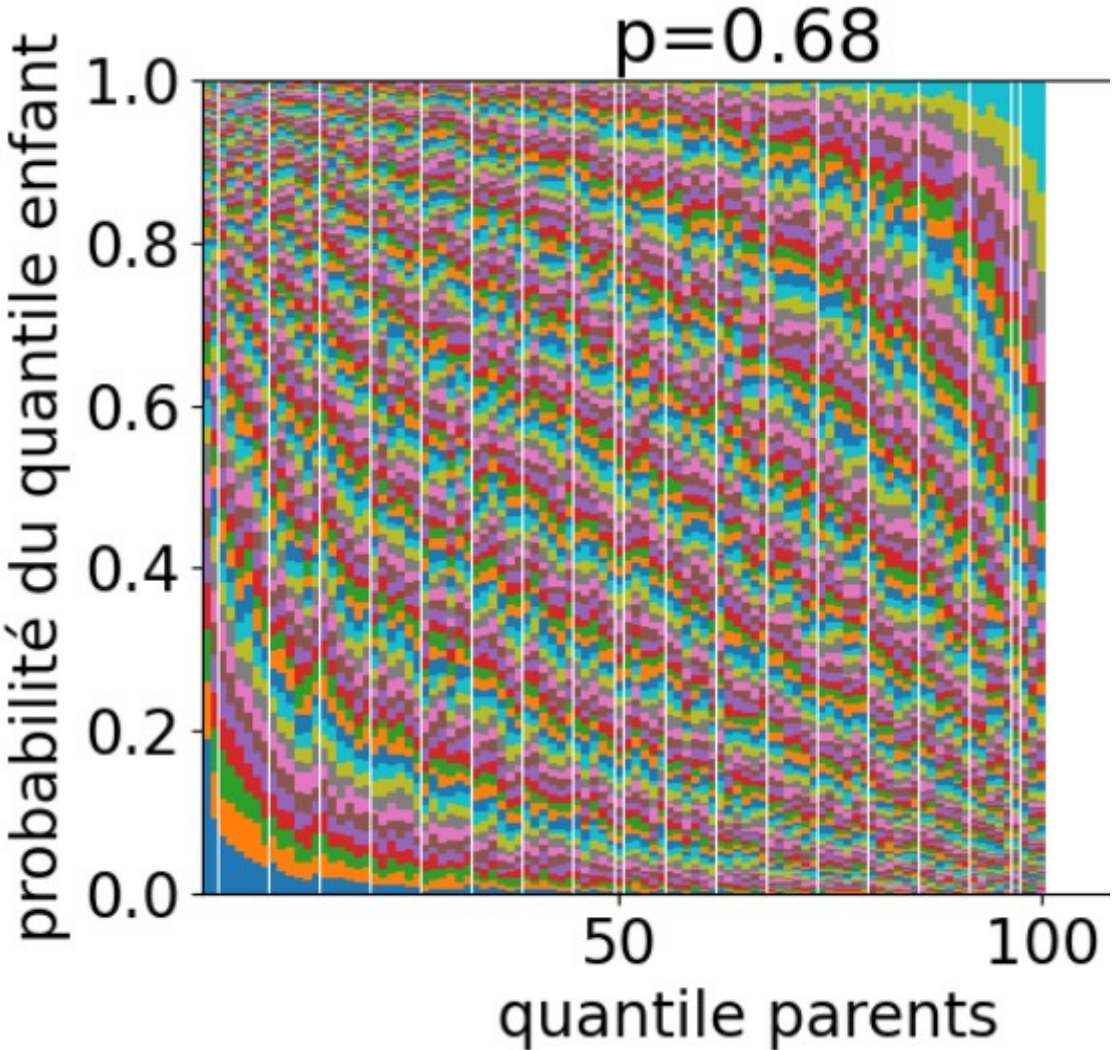


Image exemple : mobilités intergénérationnelles des revenus en fonction de la classe de revenus des parents et de l'indice de Gini du pays (en décille)

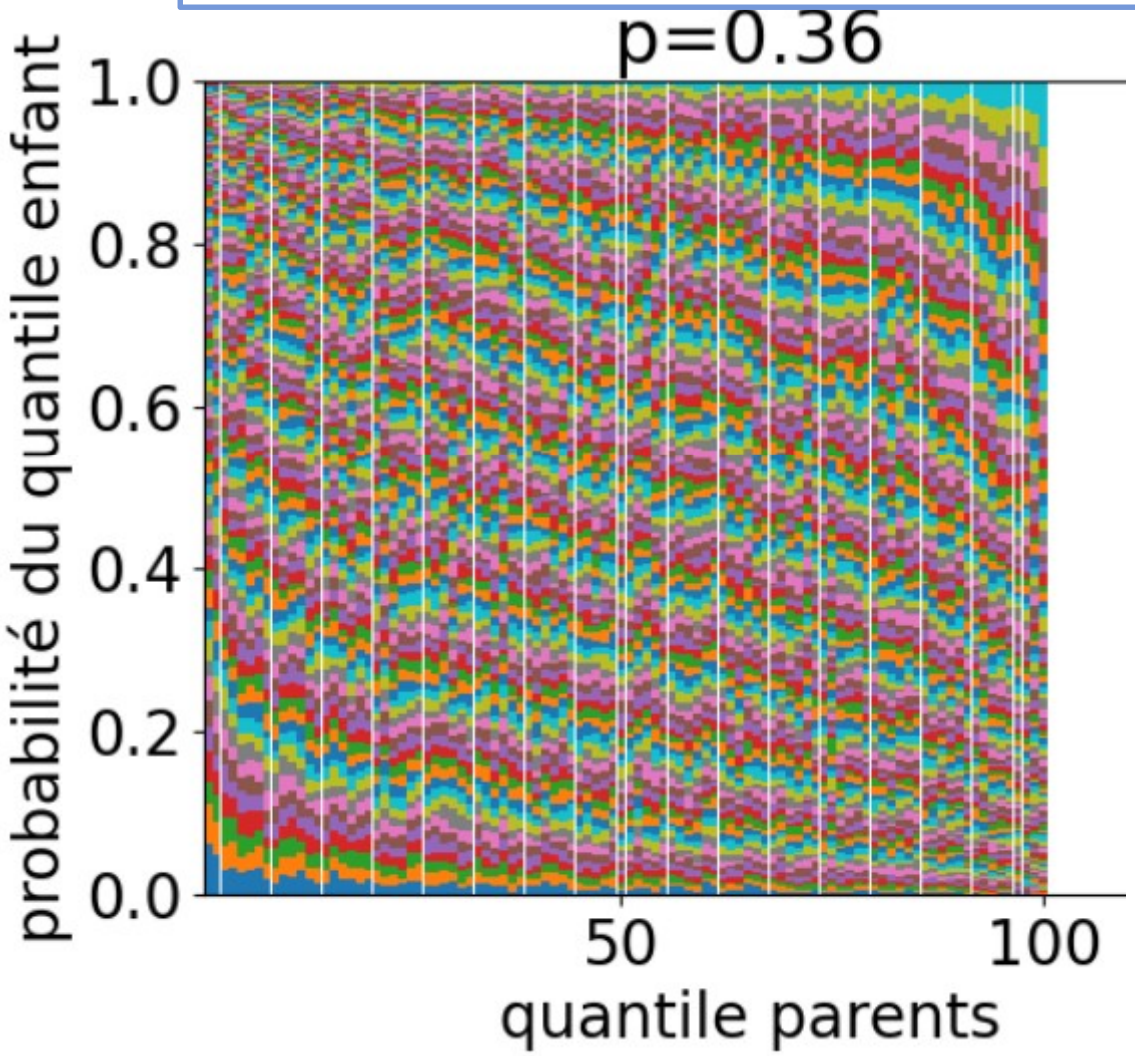
Création du modèle



Afrique du Sud

Image 5 : mobilités intergénérationnelles des revenus en fonction de la classe de revenus des parents en Afrique du sud

Création du modèle



France

Image 6 : mobilités intergénérationnelles des revenus en fonction de la classe de revenus des parents en France

Création du modèle

```
def generate_incomes(n, pj): #n = nombre de parent, pj = coefficient d'élasticité des revenus du pays j
def quantiles(l, nb quantiles):# l= nb d'échantillon (parent=enfant)
def compute_quantiles(y_child, y_parents, nb_quantiles):  #revenu enfant et parent
def proba_cond(c_i_parent, c_i_child, mat): # determine probabilité que l'enfant soit
```

Par exemple, des parents de la classe de revenus du 90ème quantile ont une probabilité de 1.2% que leurs enfants aient une classe de revenu du 90ème quantile

```
'\nP(c_i_parent = 80 | c_i_child = 90, pj = 0.36) = 0.012'
```

Pour la suite, étant donné que nous simulerons la création de 500 individus, il suffit de multiplier par 500 le résultat.
Ce qui donne donc 6 enfants sur les 500 qui appartiendrons à cette classe

```
def c_i_parent(elasticity_coefficient): # créer les revenus (enfant / parent)
#     en fonction du coef d'élasticité,
#     les répartis en fonction du nombre de quantile demandé
```

```
def elasticity_parent_class_dataframe(elasticity_coefficient):# Créer 500 individus avec les revenus
#     fonction du coefficient l'elasticite (c_i_parent)
#     les 500 individus sont répartis selon les distributions de prob_condi
```

```
for coefficient in elasticite['IGEincome'].unique():
    elasticity_dataframe = pd.concat([elasticity_dataframe,elasticity_parent_class_dataframe(coefficient)])
elasticity_dataframe.to_csv("elasticity_dataframe.csv")

elasticity_dataframe = pd.read_csv("elasticity_dataframe.csv")
```

Création du modèle

	Unnamed: 0	quantile	0	1	2	3	4	5	6	97	98	99	elasticity_coefficient	income	Gini_last
0	0	1	52.9	30.5	26.0	20.0	20.4	18.1	15.7	0.2	0.1	0.0	0.49	914.60840	46.6
1	1	2	30.6	24.0	19.5	19.1	13.9	15.4	14.2	0.3	0.1	0.0	0.49	1149.12230	46.6
2	2	3	23.8	22.9	17.4	18.8	15.6	13.2	13.9	0.3	0.1	0.1	0.49	2669.34100	46.6
3	3	4	23.2	18.9	17.6	15.2	13.9	12.2	12.0	0.1	0.2	0.3	0.49	5215.08840	46.6
4	4	5	21.0	15.0	16.9	14.8	13.2	12.8	11.4	0.1	0.3	0.0	0.49	946.51620	46.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5495	95	96	0.2	0.7	0.5	0.8	0.8	0.4	1.1	15.6	16.8	17.4	0.50	412.45360	40.5
5496	96	97	0.1	0.0	0.5	0.3	0.5	0.6	0.8	15.5	18.3	22.0	0.50	499.86500	40.5
5497	97	98	0.1	0.1	0.3	0.1	0.0	0.5	0.9	16.8	21.8	24.6	0.50	422.87726	40.5
5498	98	99	0.0	0.0	0.2	0.3	0.1	0.7	0.4	21.6	24.1	32.5	0.50	516.32370	40.5
5499	99	100	0.0	0.3	0.0	0.2	0.1	0.2	0.0	27.9	34.5	53.8	0.50	435.78415	40.5

Création du modèle

```
dataset = pd.melt(dataset, id_vars=['Country Code', 'Gini_last', 'income'],
                  value_name='parent_quantity', var_name='parent_quantile')
```

```
# # before repeating rows : deleting parent_quantity = 0
dataset = dataset[dataset['parent_quantity']!=0]
```

```
# # create parent quantiles
repeat_list = pd.Series(dataset['parent_quantity'])
dataset = dataset[['Country Code', 'Gini_last', 'income', 'parent_quantile']]


dataset = dataset.apply(np.repeat,repeats=(repeat_list))
```

	Country Code	Gini_last	income	parent_quantile
0	ALB	33.1875	2115	1
1100798	ZAF	63.0	3894	100
1100796	ZAF	63.0	6230.206	100

```
dataset['parent_quantile'] = dataset['parent_quantile'].astype(np.uint8)
dataset['income'] = dataset['income'].astype(np.uint32)
dataset['Gini_last'] = dataset['Gini_last'].astype(np.float16)
dataset['Country Code'] = dataset['Country Code'].astype('category')
```

Modèle prédictif

Intérêt : Prédire qu'un événement soit lié à certaines variables

Comment : Utiliser les données numériques, analyser la variance expliquée

Determiner : quelles sont les variables qui permettent de cibler des potentiels clients avec des revenus élevés.

Modèle prédictif

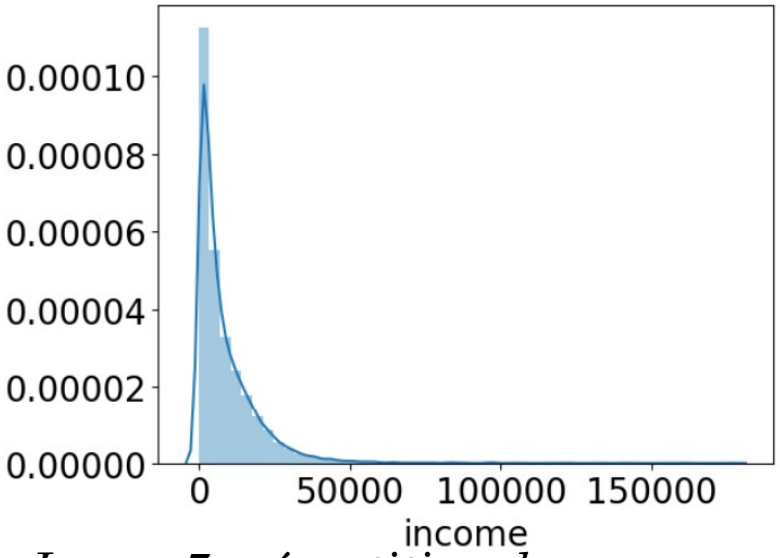


Image 7 : répartition des revenus en pourcentage de la population totale

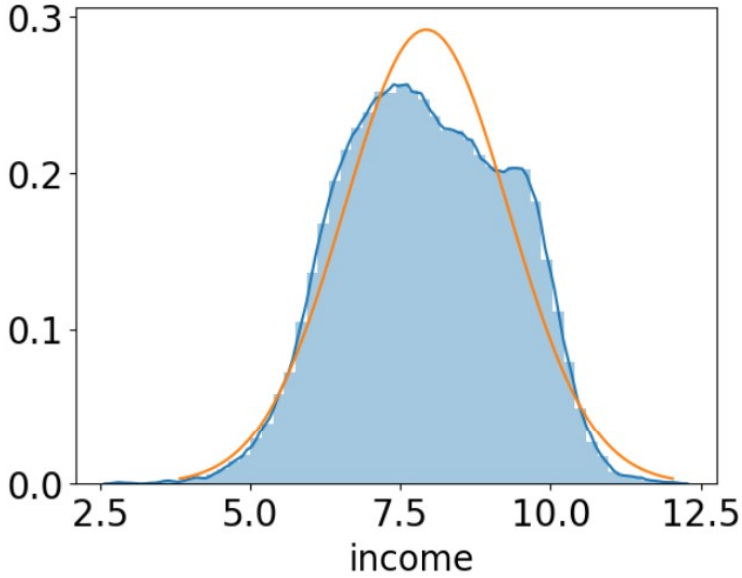


Image 8 : répartition des revenus logarithmiques en pourcentage de la population totale

Résultat avec les revenus normaux

OLS Regression Results

Dep. Variable:	income	R-squared:	0.490
Model:	OLS	Adj. R-squared:	0.485
Method:	Least Squares	F-statistic:	96.18
Date:	Fri, 10 Dec 2021	Prob (F-statistic):	0.00
Time:	13:22:29	Log-Likelihood:	-1.1165e+05
No. Observations:	10899	AIC:	2.235e+05
Df Residuals:	10790	BIC:	2.243e+05
Df Model:	108		
Covariance Type:	nonrobust		

Résultat avec les revenus en log

OLS Regression Results

Dep. Variable:	log_income	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.716
Method:	Least Squares	F-statistic:	255.7
Date:	Fri, 10 Dec 2021	Prob (F-statistic):	0.00
Time:	13:22:29	Log-Likelihood:	-11947.
No. Observations:	10899	AIC:	2.411e+04
Df Residuals:	10790	BIC:	2.491e+04
Df Model:	108		
Covariance Type:	nonrobust		

Modèle prédictif

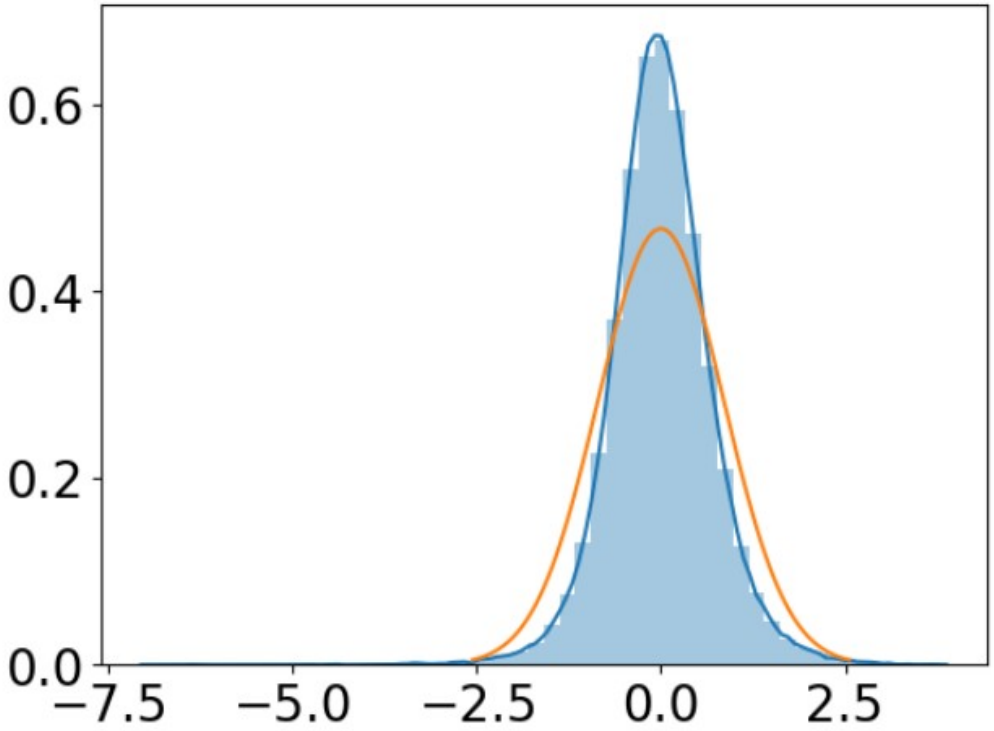


Image 9 : répartition des résidus du modèle

```
R² train : 0.71
R² test : 0.72
```

Résultat variance avec code pays

```
R² train : 0.75
R² test : 0.74
```

Résultat variance avec code pays et classe de revenus des parents

Modèle prédictif

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	-28686927.9064	99951242.8556	-0.2870	0.7741	-224587807.5696	167213951.7568
Gini_last	-6.958e+04	3.01e+06	-0.023	0.982	-5.97e+06	5.83e+06

Plus le Gini augmente et plus le revenu va diminuer à cause de son coefficient négatif. Un indice de Gini plus élevé implique un pays plus inégalitaire. Donc, plus un pays est égalitaire et plus cela va favoriser de personnes pour leur revenus

Problème la mobilité intergénérationnelle n'est pas forcement la même en fonction des classes de revenus

Modèle prédictif

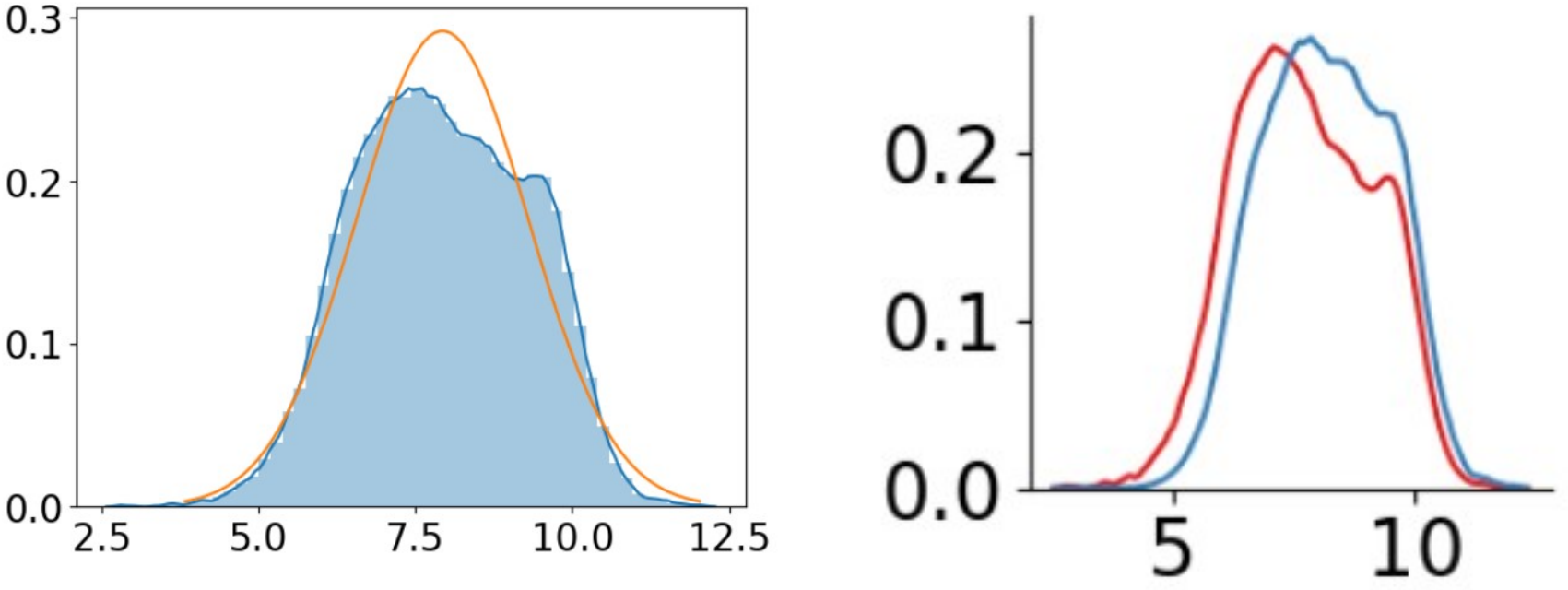


Image 10 : répartition des revenus logarithmique en pourcentage de la population totale et clustering de cette répartition

Conclusion

SOUHAITE CIBLER DE NOUVEAUX JEUNES CLIENTS

EN CIBLANT DE FUTUR PERSONNES AVEC DES HAUTS REVENUS

CRÉER UN MODÈLE POUR DÉTERMINER LE REVENU POTENTIEL D'UNE PERSONNE

EMPLOYÉ DANS UNE BANQUE