Prédiction de revenus

Contexte

Une banque cherche à créer un modèle permettant de prédire le revenu potentiel des enfants de ses clients.

Le modèle

Il devra expliquer le revenu des enfants en fonction de 3 variables :

- Le revenu moyen du pays de résidence
- L'indice de Gini de ce pays
- La classe de revenu des parents

Les différentes étapes

- Récupération les données de la World Income Distribution (revenus par pays)
- Réplication des données de manière à simuler 500 individus par pays
- Génération de la classe de revenu des parents (car nous n'avons pas cette donnée)
- Interprétation du modèle
- Conclusion

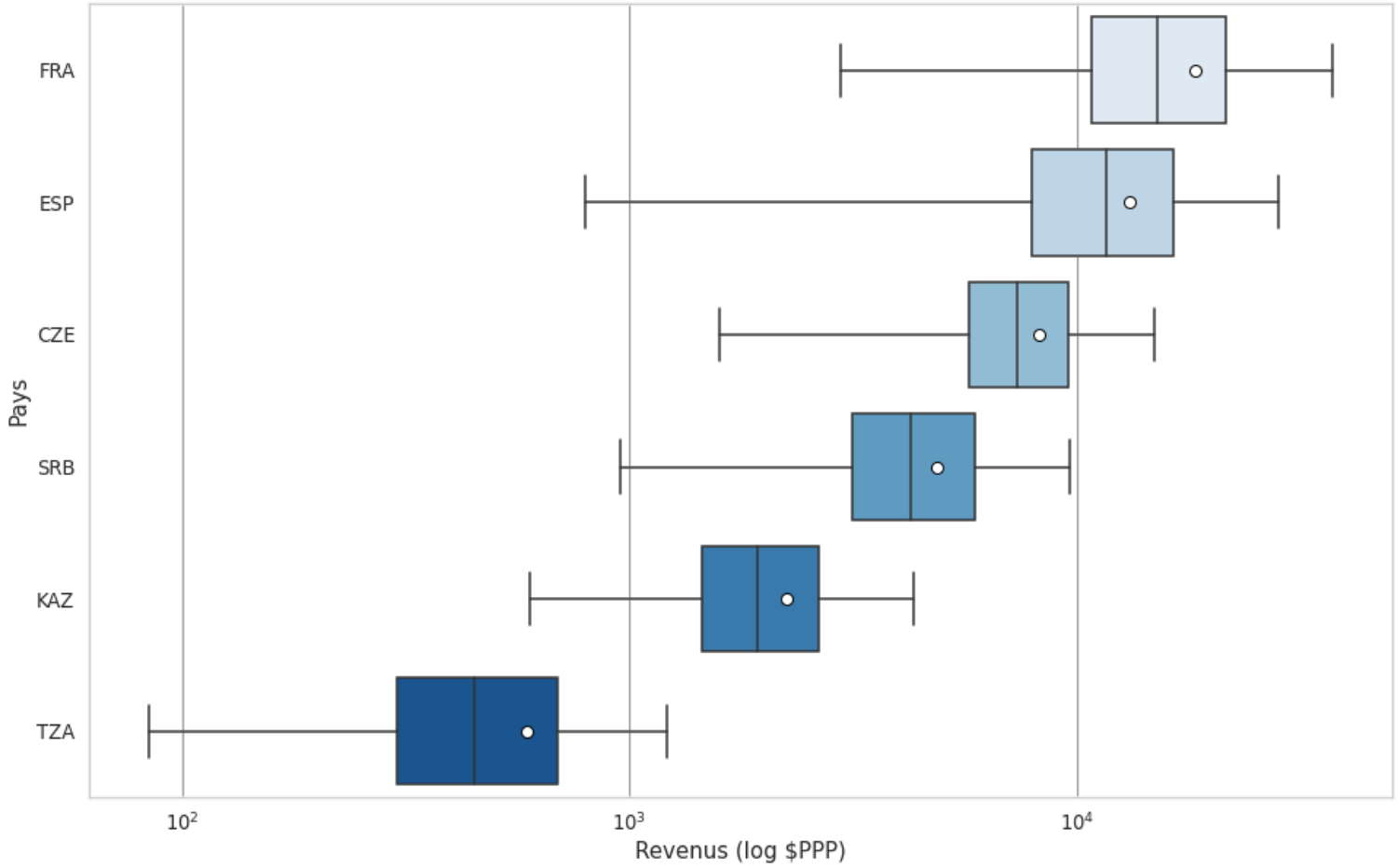
Collecte des données

- Le revenu moyen du pays --> source World Income Distribution
- L'indice de Gini du pays --> source Banque mondiale
- La classe de revenu des parents : à générer

World Income Distribution : études nationales sur la distribution des revenus, la population est échantillonnée par centiles, 110 pays sont représentés (soit 88 % de population mondiale), les études vont de 2004 à 2011 (l'année 2005 est manquante).

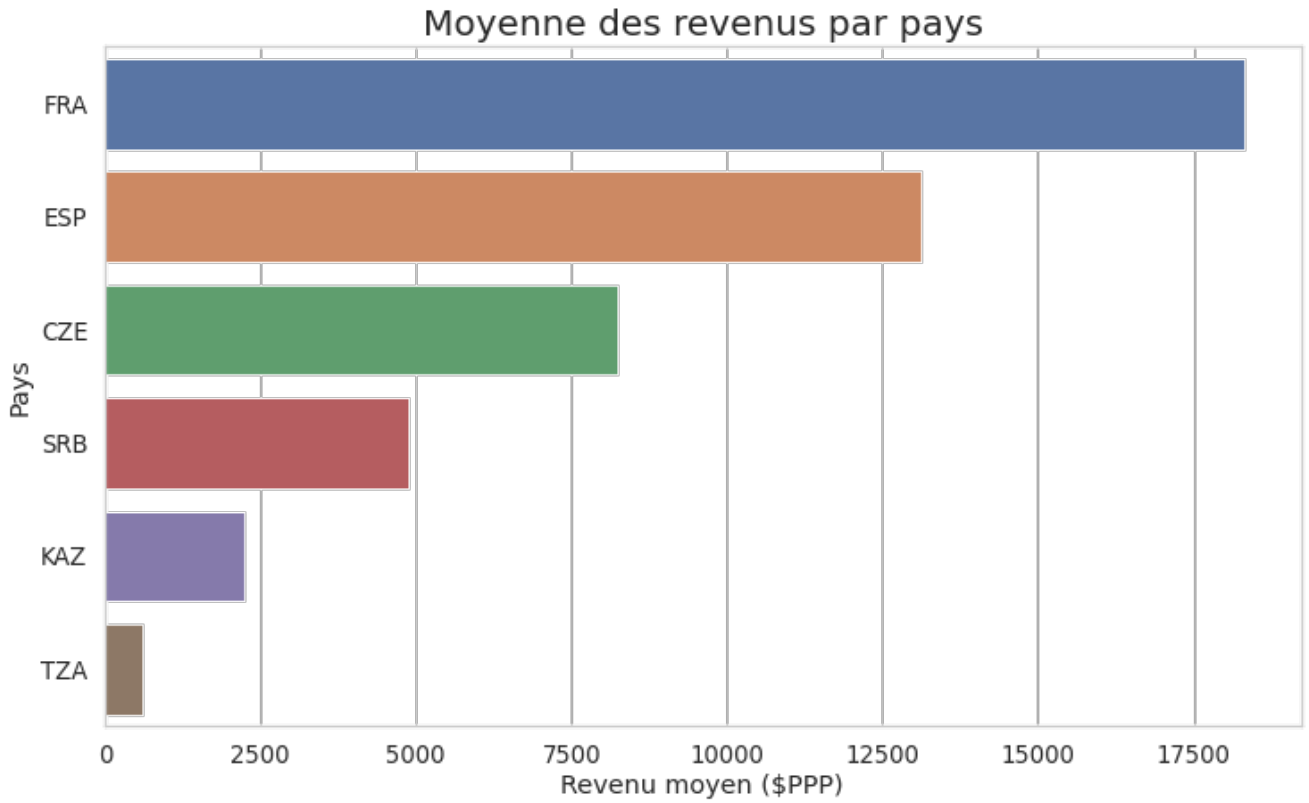
Indice de Gini : indice moyen de la période 2000 – 2020 pour chaque pays.

Distribution des revenus, exemple de 6 pays

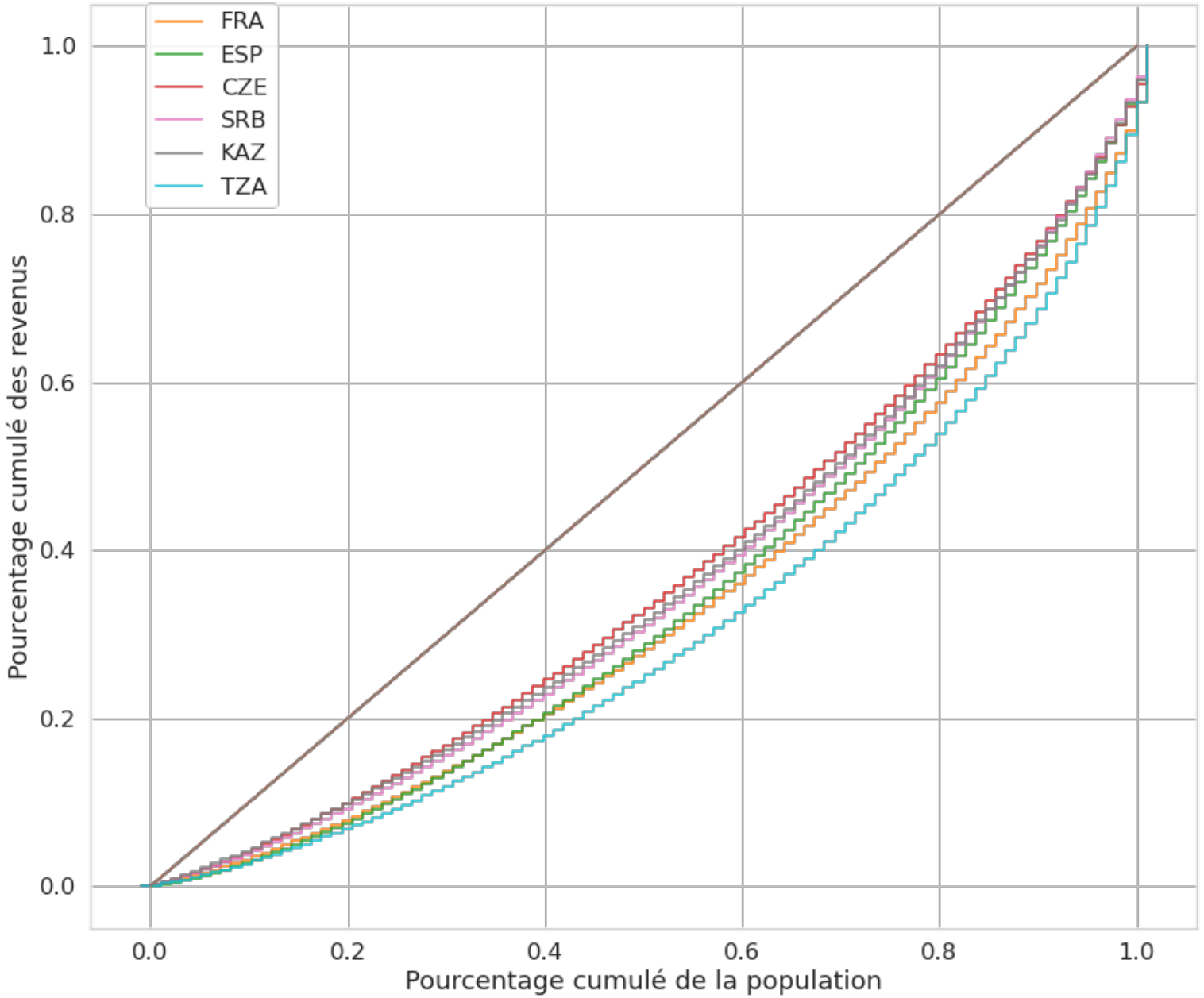


- France (FRA)
- Espagne (ESP)
- République Tchèque (CZE)
- Serbie (SRB)
- Kazakhstan (KAZ)
- Tanzanie (TZA)

Moyenne des revenus des 6 pays



- France (FRA)
- Espagne (ESP)
- République Tchèque (CZE)
- Serbie (SRB)
- Kazakhstan (KAZ)
- Tanzanie (TZA)



Courbes de Lorentz (6 pays)

- République Tchèque (CZE)
- Kazakhstan (KAZ)
- Serbie (SRB)

- Espagne (ESP)
- France (FRA)
- Tanzanie (TZA)

Classement des pays par indice de Gini

Classement	Pays	Indice de Gini
1	Slovenia	25
2	Czech Republic	26
3	Ukraine	26
4	Slovak Republic	26
5	Norway	27
34	France	32
159	Brazil	54
160	Central African Republic	56
161	Botswana	60
162	Namibia	61
163	South Africa	62

Génération de la classe de revenu des parents

Pour chaque pays, on calcule une matrice de probabilités conditionnelles qui permettra d'attribuer les classes de revenu des parents.

Exemple pour une valeur de probabilité :

P(classe_parent = 8 | classe_enfant = 5 et coeff d'élasticité = 0.9) = 0.03, avec P probabilité conditionnelle qui permet d'assigner la classe_parent = 8 à 15 (500*0.03) des 500 individus ayant classe_enfant = 5, et un pays avec coefficient d'élasticité (mesure de mobilité intergénérationnelle du revenu) de 0.9.

Dataset final (extrait)

- 5 500 000 lignes (500 individus * 100 quantiles * 110 pays)
- 9 colonnes

index	code	quantile	income	population	pays	Gini index	elasticity	id	quantile_parent
0	ALB	1	728.89795	3002678	Albania	32	0.668573	1	1
1	ALB	1	728.89795	3002678	Albania	32	0.668573	2	1
2	ALB	1	728.89795	3002678	Albania	32	0.668573	3	1
3	ALB	1	728.89795	3002678	Albania	32	0.668573	4	1
4	ALB	1	728.89795	3002678	Albania	32	0.668573	5	1
…	…	…	…	…	…	…	…	…	…
5499995	ZAF	100	82408.55	49779471	South Africa	62	0.638598	496	100
5499996	ZAF	100	82408.55	49779471	South Africa	62	0.638598	497	100
5499997	ZAF	100	82408.55	49779471	South Africa	62	0.638598	498	100
5499998	ZAF	100	82408.55	49779471	South Africa	62	0.638598	499	100
5499999	ZAF	100	82408.55	49779471	South Africa	62	0.638598	500	100

Modèle linéaire : voir notebook

Le modèle permet d’expliquer 77 % du revenu des individus avec 3 variables :

- le revenu moyen du pays de naissance
- l'indice de Gini du pays de naissance
- le revenu des parents

Les 33 % non expliqués peuvent être liés à des facteurs non considérés dans le modèle comme la chance, le niveau d'étude, la volonté...