OPENCLASSROOMS

Mission 7

Effectuez une prédiction de revenus

Mentor : Claire Della Vedova

Laurent Gourdon – Parcours Data Analyst

Agenda

Mission 1
- Consolidation des données
- Description des données

Mission 2
- Distribution logarithmique des revenus
- Courbes de Lorenz
- Evolution de l'indice Gini (2004-2013)
- Classement par indice Gini

Mission 3
- Génération de l'échantillon Gaussien
- Distributions conditionnelles
- Clonage de l'échantillon

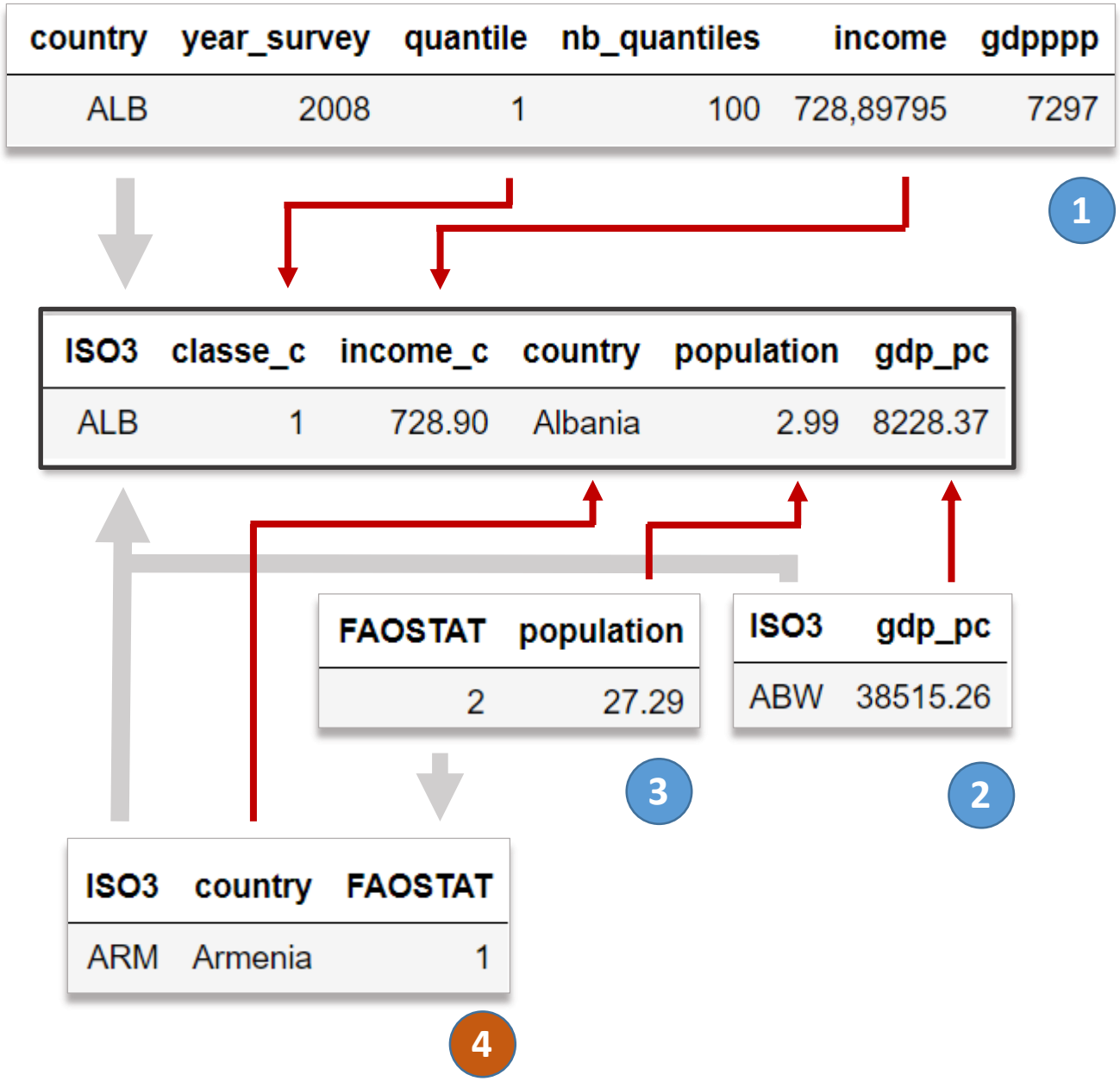
Mission 4 (Notebook Jupiter)
- ANOVA
- Régressions linéaires (income_c ~ gdp_pc + gini)
- Régression linéaire (income_c ~ gdp_pc +gini + classe_p)
- Choix du modèle
- Coéficient de régression de Gini

Misson 1

- Consolidation des données
- Description des données

Consolidation des données

1. 'data-projet7.csv' (OC)
2. GDP per capita 2008 (World Bank)
3. Population 2008 (World Bank)
4. Nomenclature pays (FAO)



Description des données

- Les quantiles sont au nombre de 100, ce sont donc des centiles ou percentiles
- Ils permettent de dépasser les limites de la moyenne qui masque les disparités de distribution par classe de revenus
- Le revenu est exprimé en Purchasing Power Parity ($PPP)
- L'année d'etude est répartie symétriquement en cloche autour de l'année 2008

- Population de l'étude : **6.15 milliards**
- Pourcentage de la population mondiale (2008) **: 91.8 %**

- Nombre de quantiles uniques : **100**
- Nombre de pays uniques : **111**

	classe_c	income_c	population	gdp_pc
count	11099.00	11099.00	11099.00	11099.00
mean	50.50	6148.37	55.42	15788.47
std	28.87	9466.70	172.37	15220.17
min	1.00	16.72	0.31	615.07
25%	25.50	922.41	4.77	3914.61
50%	51.00	2495.71	14.01	10236.54
75%	75.50	7622.36	43.27	20837.25
max	100.00	176928.55	1344.42	86693.90

year_survey	nb_ctry
2004	1
2006	5
2007	15
2008	76
2009	12
2010	6
2011	1

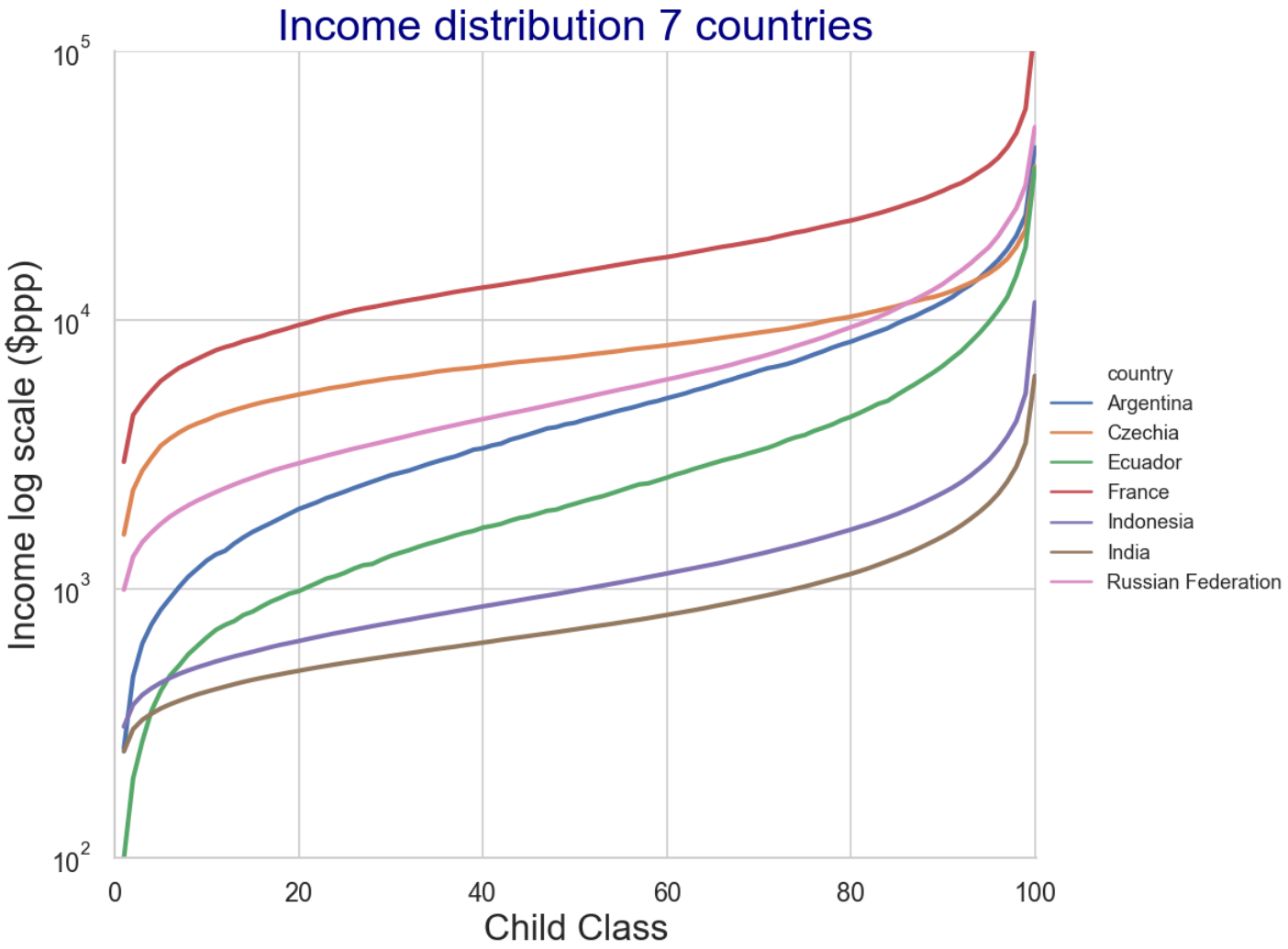
Misson 2

- Distribution logarithmique des revenus
- Courbes de Lorenz
- Evolution de l’indice Gini (2004-2013)
- Classement par indice Gini

Distribution logarithmique des revenus

Distribution des revenus des pays sélectionnés selon leur classe

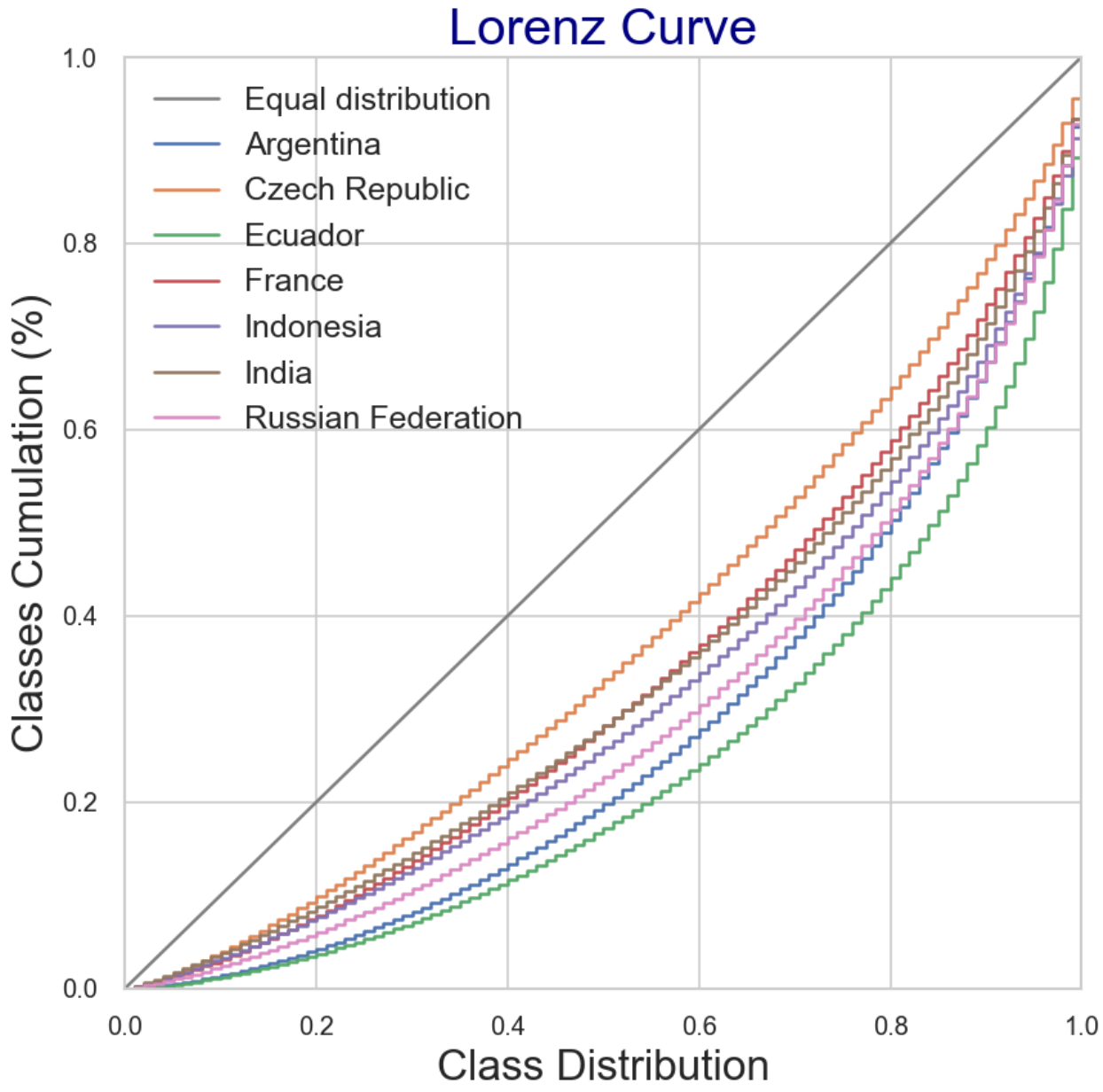
- Les courbes de Czechia, puis India et Indonesia sont les plus plates, les revenus sont mieux distribués entre les classes enfants
- Equador, puis Argentina et Russian Federation ont les pentes les plus marquées, la distribution des revenus est plus favorable aux classes élevées
- La France se situe au milieu de l'échantillon retenu



Courbes de Lorenz

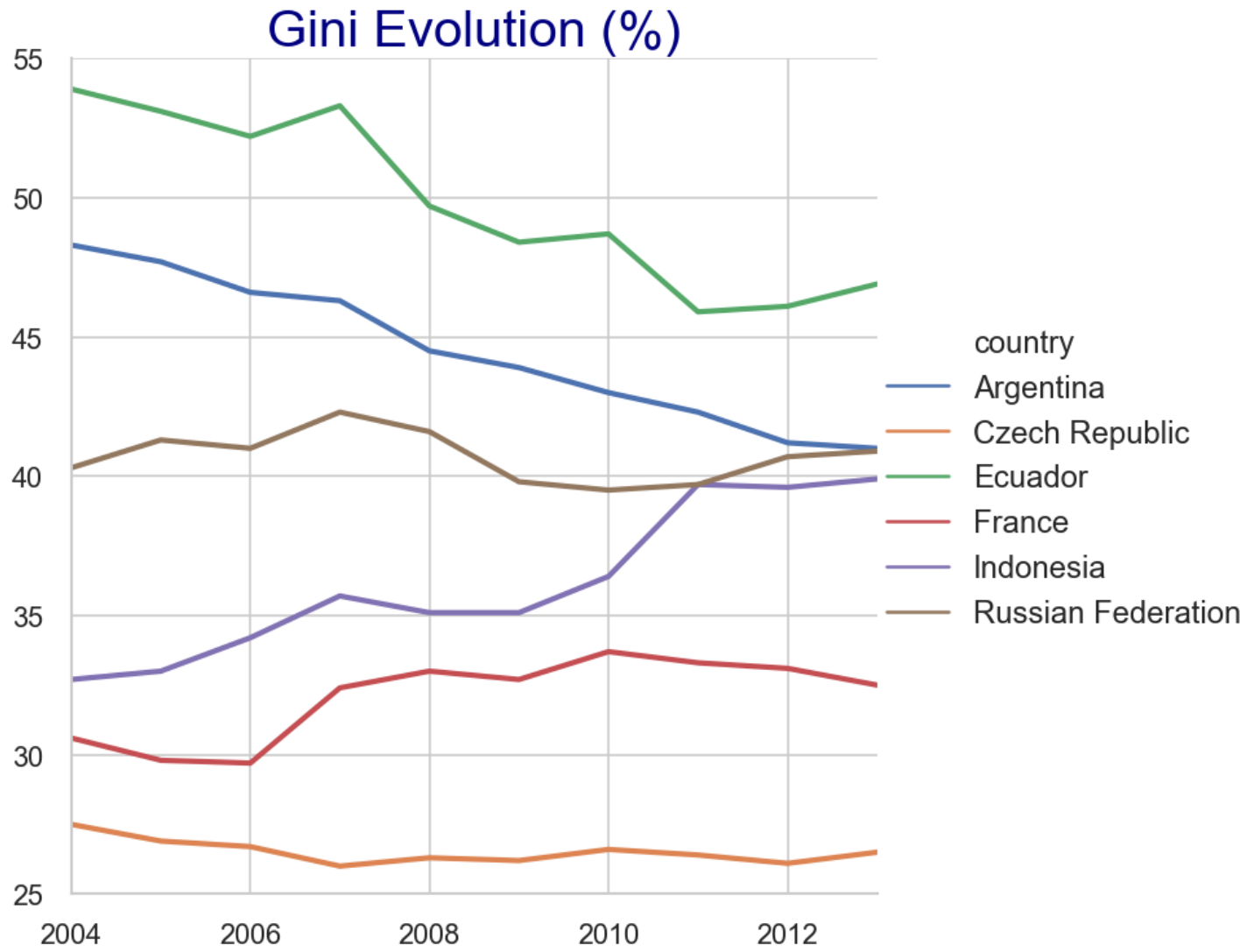
On représente la distribution ordonnée des revenus par classe et leur proportion pour visualiser les inégalités

- La courbe la plus proche d'une distribution équitable est celle de Czechia.
- Les courbes qui favorisent le plus les revenus élevés sont celle de Equador puis celle de Argentina
- La France est entre les deux, du côté des pays les moins inégalitaires, au même niveau que India



Evolution de l'indice Gini (2004-2013)

- 2 pays flat : Czech Republic , Russian Federation
- 2 pays tendent vers une meilleure répartition : Equador, Argentina
- L'inégalité croît pour Indonesia
- 3 courbes convergent à 40% en 2013 : Argentina, Russian Federation, Indonesia
- Czechia est le pays le plus stable et le plus égalitaire sur la période
- La France est le second pays le plus égalitaire de la liste, avec cependant une détérioration sur 2007-2013 (crise des subprimes)



Classement par indice Gini

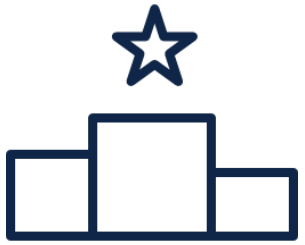
- Les 5 pays ayant le plus faible indice Gini en 2008 sont : Azerbaijan, Czech Republic, Slovak Republic, Denmark, Slovenia)
- Les 5 pays ayant le plus fort indice Gini en 2008 sont : Guatemala, Honduras, Colombia, Central African Republic, South Africa

Top 5 : pays les plus égalitaires

	ISO3	country	gini
0	SVN	Slovenia	23.70
1	DNK	Denmark	25.20
2	SVK	Slovak Republic	26.00
3	CZE	Czech Republic	26.30
4	AZE	Azerbaijan	26.60

Bottom 5 : pays les moins égalitaires

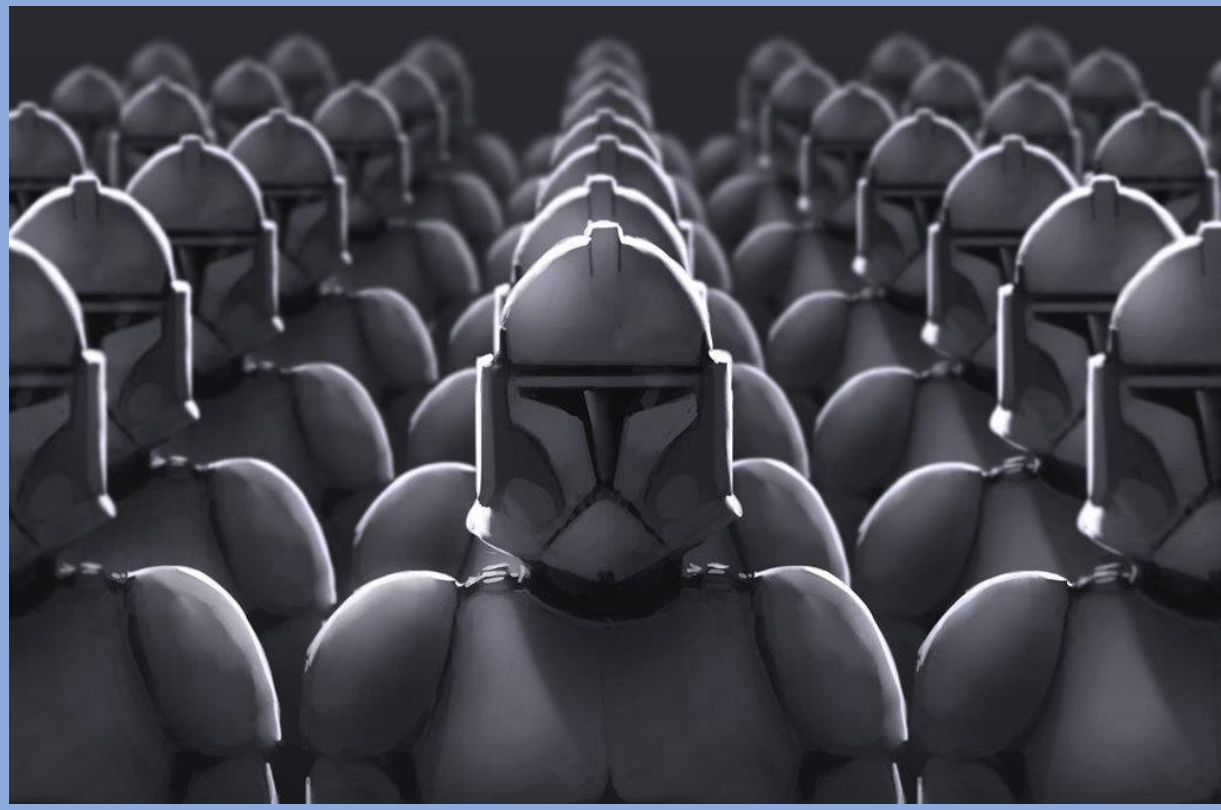
	ISO3	country	gini
106	GTM	Guatemala	54.60
107	HND	Honduras	55.50
108	COL	Colombia	55.50
109	CAF	Central African Republic	56.20
110	ZAF	South Africa	63.00



	ISO3	country	gini
37	FRA	France	33.00

La France est le 38e pays sur 111 en terme d'égalité

Mission 3



- Génération de l'échantillon Gaussien
- Distributions conditionnelles
- Clonage de l’échantillon

Génération de l'échantillon Gaussien

- Le coefficient d'élasticité est un ratio fourni pour chaque pays
- Il nous permet de générer un échantillon gaussien aléatoire de 1000 individus par centile
- Nous comptons les occurrences uniques pour chaque pays et les tranformons en poourcentage

ISO3	pj
ALB	0.82
ARG	0.40

Coefficient d'élasticité basé sur les données par la Banque mondiale complétées par elasticity.txt

```
# querry génération de l'échantillon
Q3 = pd.DataFrame()

for i in range(len(elasticity)):

    nb_quantiles = 100
    n = 1000*nb_quantiles
    pj = elasticity.loc[i]['pj']

    y_child, y_parents = sg.generate_incomes(n, pj)
    e = elasticity.loc[i]['ISO3']

    y_child = pd.Series(y_child)
    y_parents = pd.Series(y_parents)

    cq = sg.compute_quantiles(y_child, y_parents, nb_quantiles)
    cq['ISO3'] = e
    Q3 = Q3.append(cq, ignore_index = True)
```

	ISO3	c_i_child	c_i_parent	count
0	ALB	1	1	208
1	ALB	1	2	107

	ISO3	c_i_child	c_i_parent	percent
0	ALB	1	1	0.21
1	ALB	1	2	0.11

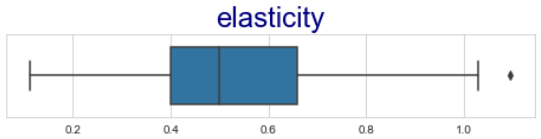
```
ISO3          1081214
c_i_child     1081214
c_i_parent    1081214
count         1081214
dtype: int64
```

Distributions conditionnelles

Distributions calculées sur des déciles de 1000 individus

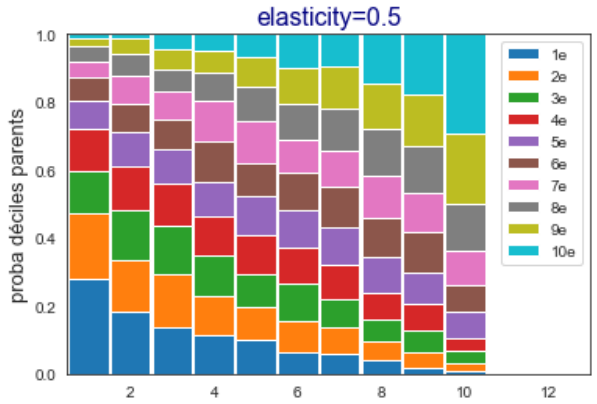
Valeurs caractéristiques du boxplot elasticity

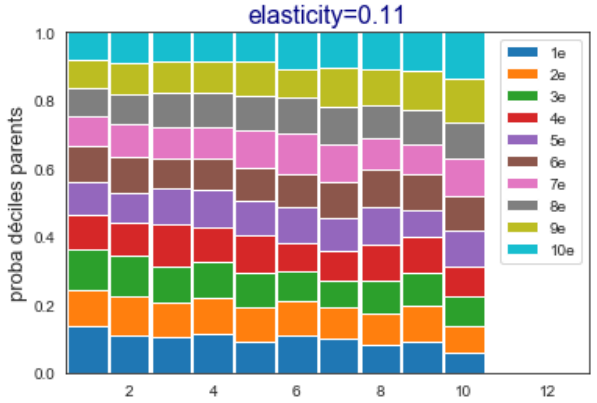
- Plus le coéficient d'élasticité augmente, plus l'inégalité entre les classes augmente
- Pour e=0.11, la probabilité d'avoir des parents dans les 4 premiers déciles est de 50%
- Pour e=1.1, la probabilité d'avoir des parents dans les 4 premiers déciles est de plus de 90%

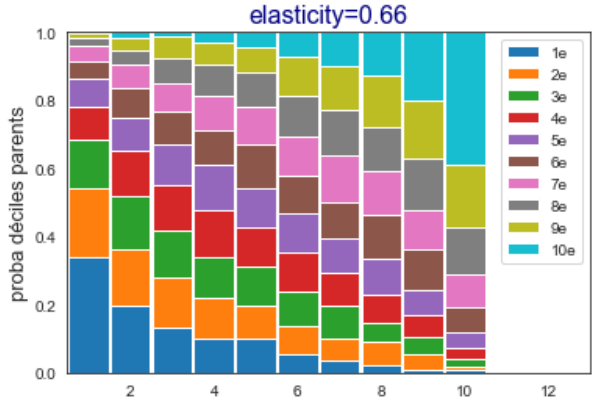


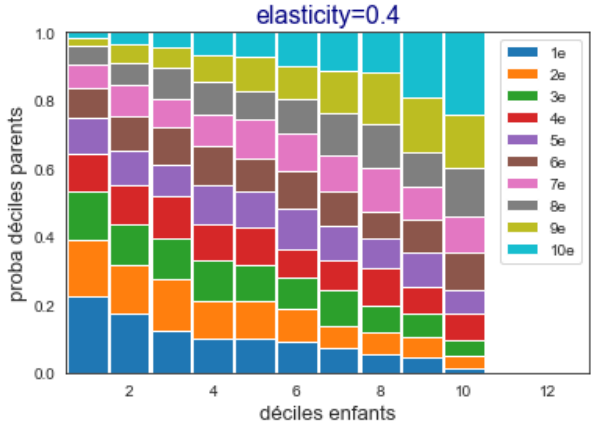
min	0.11
25%	0.40
50%	0.50
75%	0.66
max	1.10

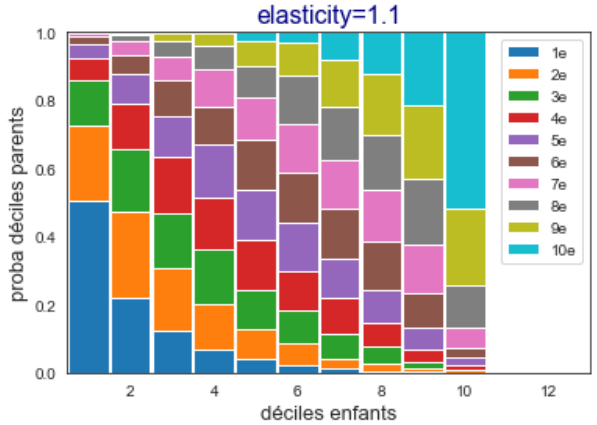
```
# choix des paramètres
nb_quantiles = 10
n = 1000*nb_quantiles
```











Clonage de l'échantillon

- On multiplie le pourcentage obtenu avec l'échantillon gaussien par 500 afin de réaliser l'échantillon requis
- Nous obtenons un dataframe final de 5.455.255 individus pour 109 pays

Nouvel échantillon, clonage

```
x = round(df_final['percent']*500,0)

#clonage du dataset
classes = pd.DataFrame(np.repeat(df_final.values,x,axis=0))
classes.columns = ['ISO3','c_i_child','c_i_parent','percent']
classes.drop(['percent'], axis=1, inplace=True)
classes[['c_i_child','c_i_parent']] = classes[['c_i_child','c_i_parent']].astype(int)
classes.columns=['ISO3','classe_c','classe_p']

print("Nombre d'individus",classes.ISO3.count())
print("Nombre de pays :",classes.ISO3.nunique())
classes.head(1)
```

```
Nombre d'individus 5555338
Nombre de pays : 111
```

	ISO3	classe_c	classe_p
0	ALB	1	1

	ISO3	country	classe_c	income_c	classe_p	income_p	gdp_pc	gini
0	ALB	Albania	1	728.90	1	728.90	8228.37	30.00

Des questions ?

lgourdon@gmail.com (+34 659 34 05 15)

Conclusion

Pour aller plus loin

Analyse de Income Child

- Fort skewness à droite (revenus élevés)
- La distribution des revenus par pays montre qu'il existe des disparités de revenu et de dispertion d'un pays à l'autre