Projet 2: Analyse des données de systèmes éducatifs

Mark Creasey,

OpenClassrooms, Parcours « Data Scientist » 10/2021

Mission

- L’entreprise **academy**, qui propose des contenus de formation en ligne pour un public de niveau lycée et université.

- Objectifs commerciaux
  - Quel sont les **pays avec un fort potentiel** de clients pour nos services ?
  - Pour chacun de ces pays, quelle sera **l’évolution de ce potentiel** de clients ?
  - Dans **quels pays** l'entreprise doit-elle opérer en **priorité** ?

- Objectifs d’analyse
  - proposer des réponses à ces questions,

    en utilisant les données sur l'éducation du World Bank

Étapes du projet

Formulation du problème

Description des données

Validation

Nettoyage

Sélection des Indicateurs

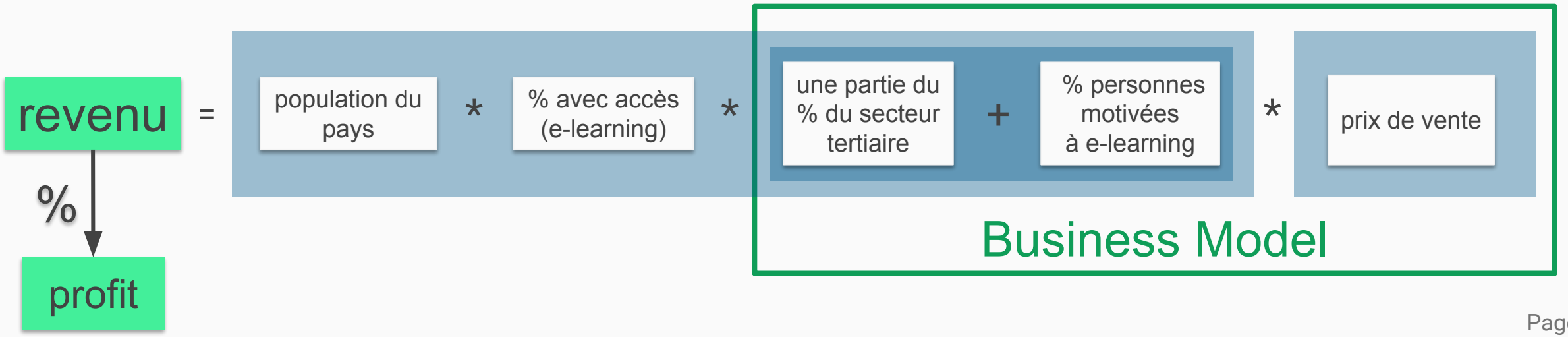
Analyse exploratoire

Conclusion

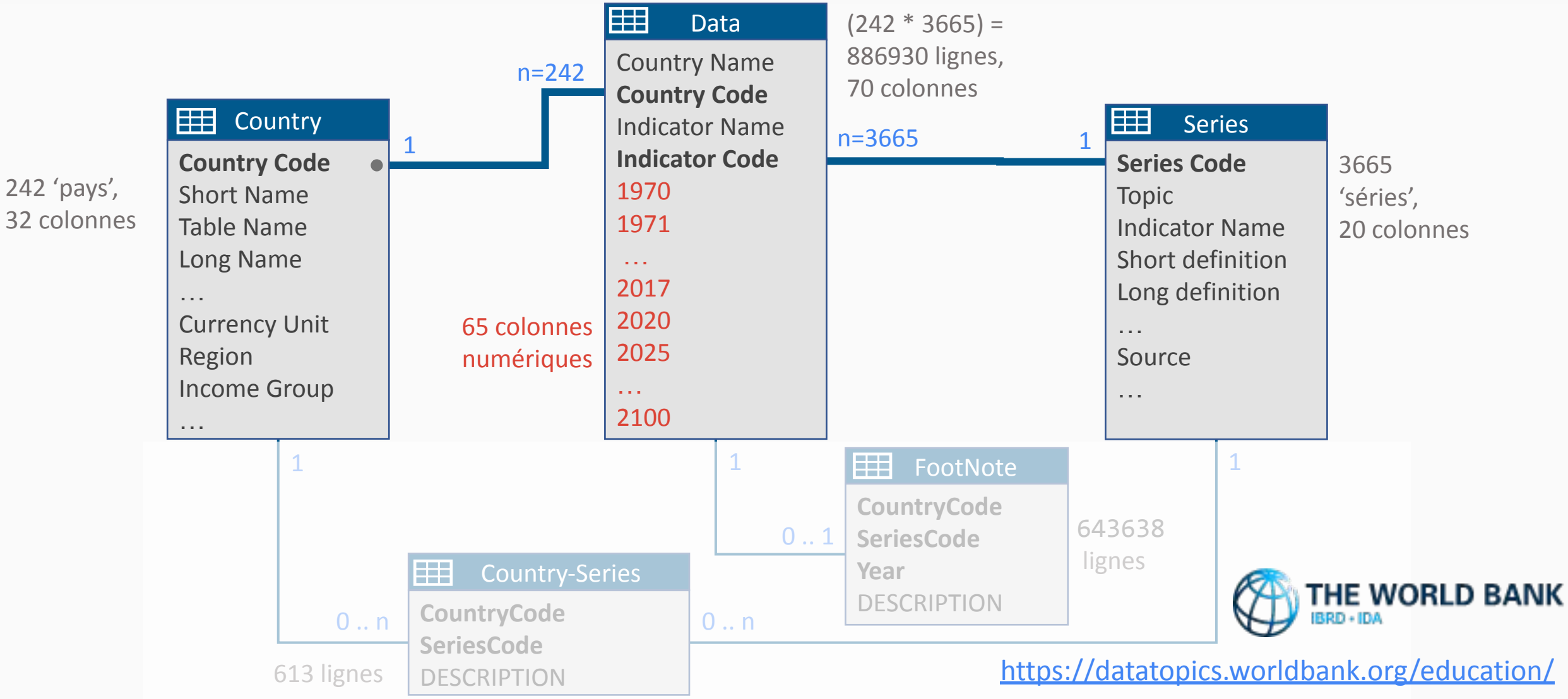
1. Formulation du problème

Chercher des indicateurs dans les données pour discriminer
dans quels pays on peut vendre :

- la **quantité maximum de cours** *(e-learning)*
- au **prix maximum**
- au **plus grand secteur de la population** *(secteur tertiaire)*
- pour **le maximum de profit** *(nombre de clients * prix de vente - dépenses)*



2.1 Description des données

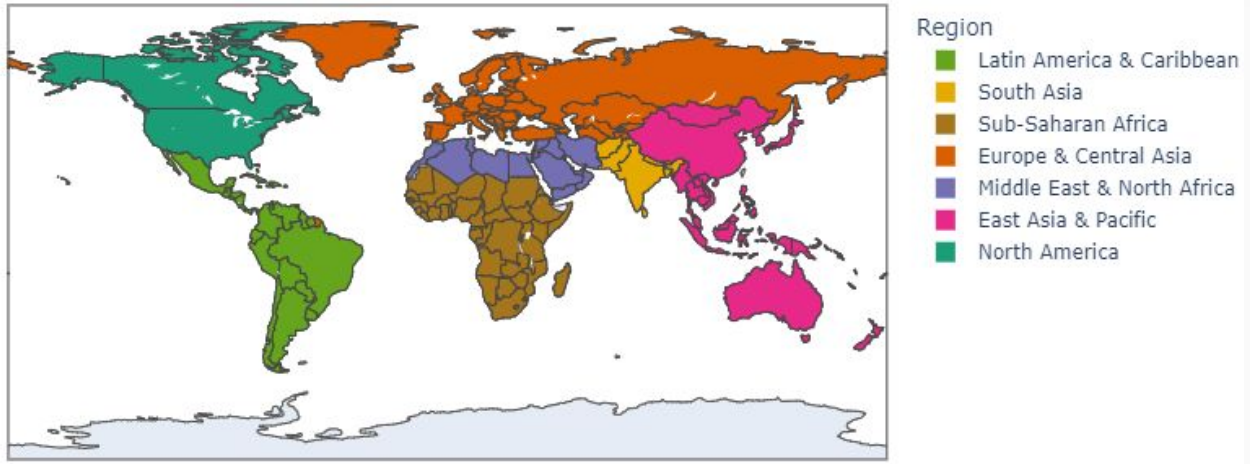


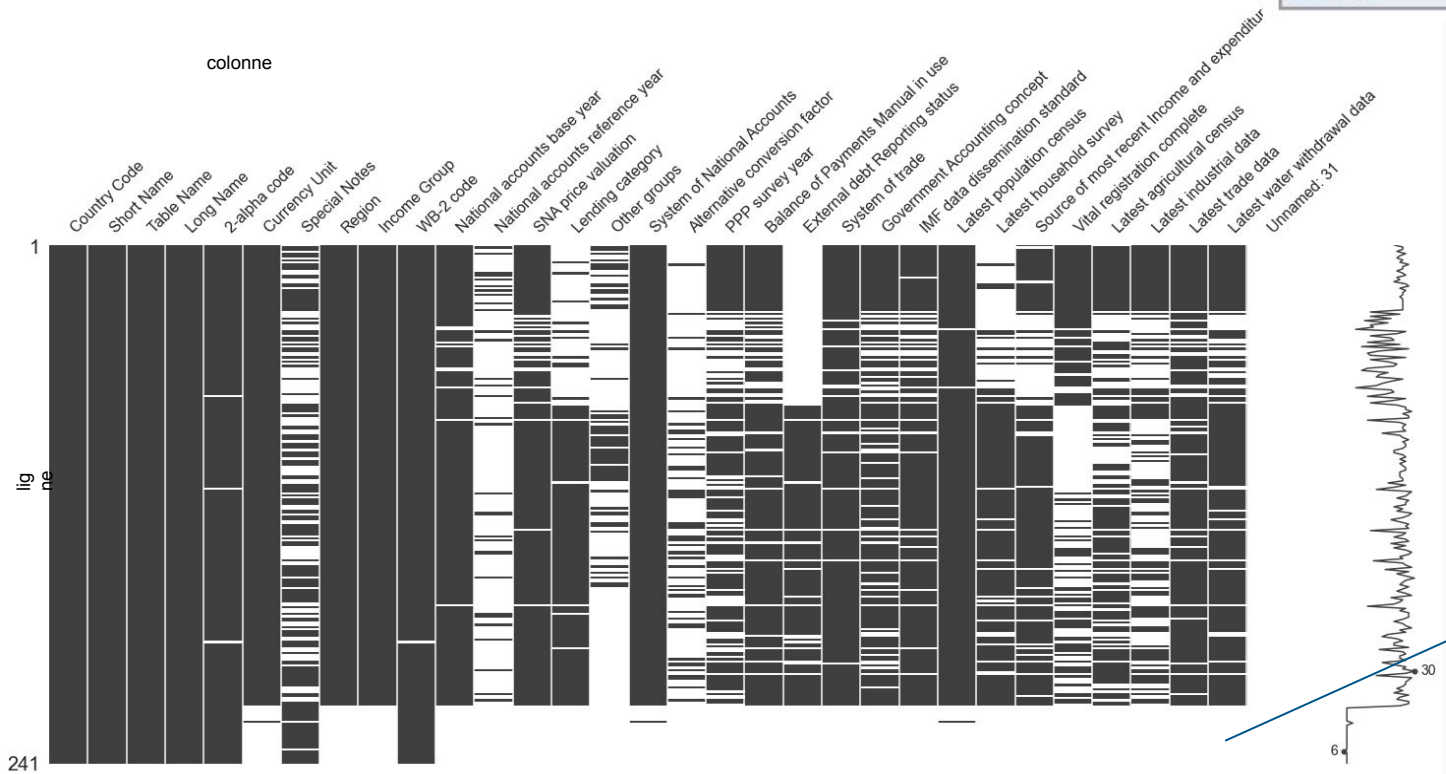
https://datatopics.worldbank.org/education/

2.2 Validation des données (1)

Country

- 214 pays en 7 régions du monde
- Aussi 28 non-pays qui sont des groupements de pays (comme EU, Arab world, OECD, …)
- Pays identifiés par ‘Country Code’ ISO-3 code (clé primaire)



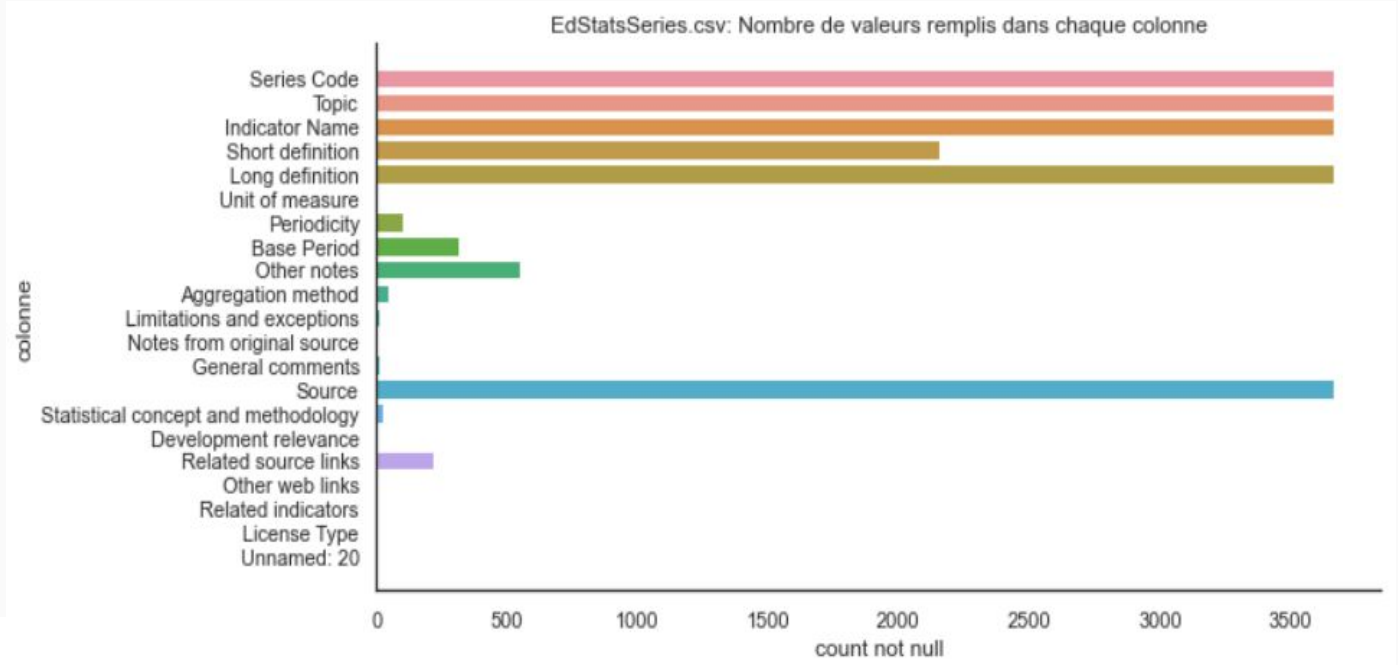


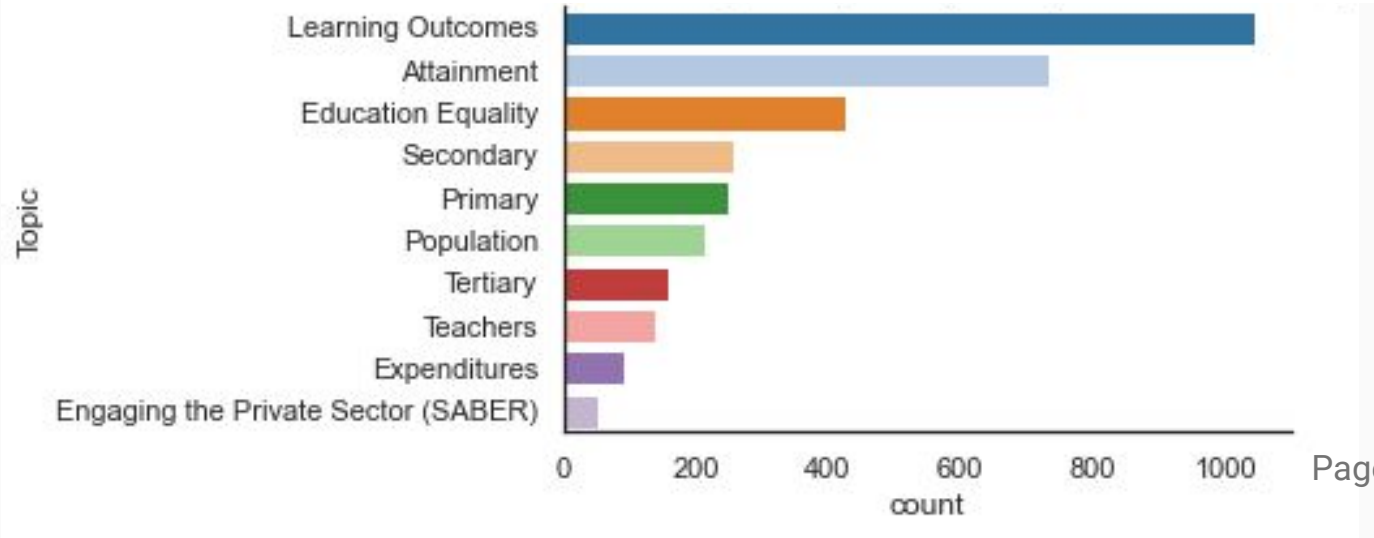
« Non-pays » (beaucoup de valeurs manquantes)

2.2 Validation des données (2)

Series

- 37 topics
- Identifié par « Series Code » (clé primaire)
- Clés ne coïncident pas toujours avec « Indicator Code » dans « Data »
- Pas de lignes dupliquées
- Seulement 6 colonnes bien remplies
- 6 colonnes vides

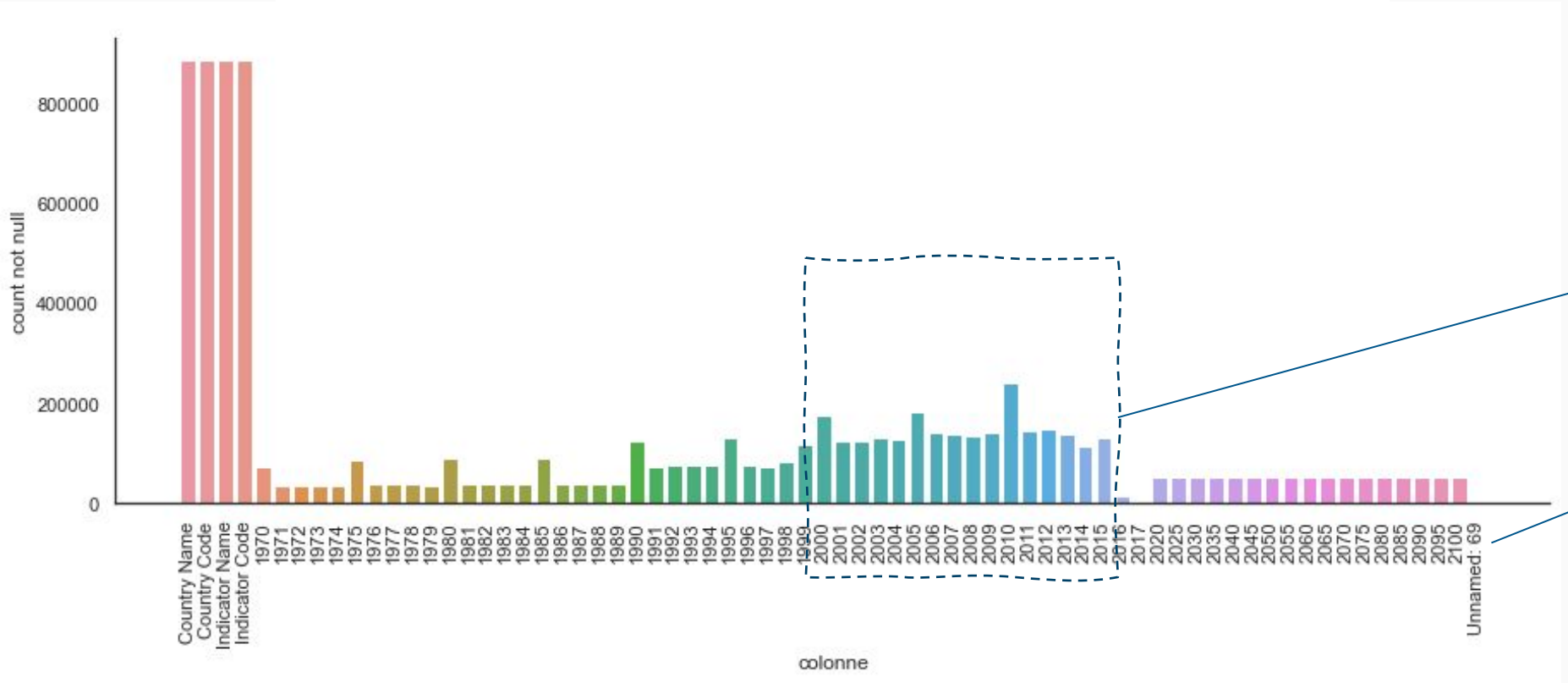




2.2 Validation des données (3)

Data

- 3665 indicateurs pour 242 « pays »
- Presque 50% des lignes n'ont aucune donnée
- Country Code correspond avec ISO-3 code du « Country »



Années les plus remplies

Colonne vide

2.2 Validation des données (4)

CountrySeries

- descriptions des sources de données de 211 pays pour 2 séries (`SP.POP.TOTL` et `SP.POP.GROW`)
- sources de données pour 13 à 15 pays pour le GDP ou GNP
- pas de lignes dupliquées

	column	count	unique	dtype	max_length
0	CountryCode	613	211	object	3
1	SeriesCode	613	21	object	17
2	DESCRIPTION	613	97	object	278
3	Unnamed: 3	0	0	float64	nan

FootNote

- Contient 643638 footnotes, un pour chaque pays, année et indicateur
- Seulement 1 sur 40 des données dans Data contient un footnote
- détail de calcul d'un indicateur pour un pays et année
- pas de lignes dupliquées

```
'YR2012', 'YR2013',
'YR2020', 'YR2025',
'YR2050', 'yr2012']
```

	column	count	unique	dtype	max_length
0	CountryCode	643638	239	object	3
1	SeriesCode	643638	1558	object	30
2	Year	643638	56	object	6
3	DESCRIPTION	643638	9102	object	1132
4	Unnamed: 4	0	0	float64	nan

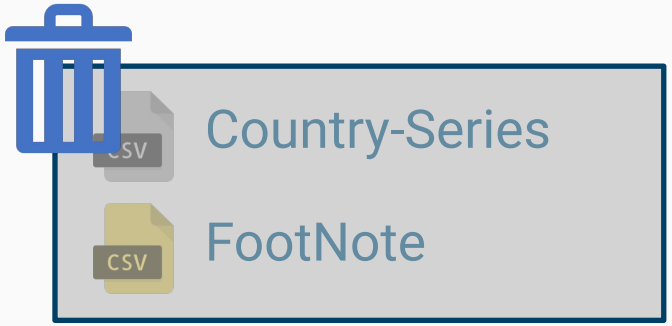
3.1 Nettoyage des données

- Pas de doublons
- Les valeurs manquantes seront supprimées par filtrage / table “join”

Suppression des colonnes vides

Colonnes clés en majuscule

Suppression (tables inutiles)



Series, colonne « Series Code »

```
Series codes sans clé Indicator Code dans table Data:
SE.SEC.DURS.LO ; SE.SEC.ENRR.UP.FE ; UIS.AIR.1.Glast.GPI; U
```

```
Data series codes sans description dans table Series:
SE.SEC.DURS.LO; SE.SEC.ENRR.UP.FE; UIS.AIR.1.GLAST.GPI; UI
```

Data, colonne « Indicator Code »

3.2 Sélection des indicateurs

revenu = population * % avec accès (e-learning) * % secteur tertiaire + motivations à e-learning * prix de vente

Population (actuel, projection)
Population * croissance
Population âge 18+

% avec accès à l'internet
% avec ordinateur
% avec smartphone
langue

Taux chômage / demande employés qualifiés
Taux professeur/étudiant
Taux rural / urbain (GINI) – accès physique aux études
Égalité des sexes (GPI – « gender parity index »)
Type d'offre et demande des cours
(tourisme, agriculture, industrie, environnement)
Attitudes culturelles (public/privé)

% Population 20-24 ans
% inscription éducation tertiaire
% ayant un BAC+2
% analphabètes / adult literacy

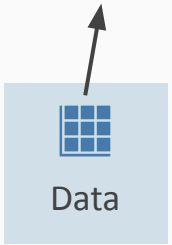
Pouvoir d'achat des individus / de l'état
Salaire (médian) / GNI per capita, GDP per capita
Cout d'études / financement des études
% du GDP dépensé sur l'éducation
Salaire des professeurs
Stabilité financière (inflation, taux de change)
Stabilité politico-géographique (conflits régionaux)
Droits et coûts d'implantation dans un pays
Protection de copyright

3.2 Sélection des indicateurs

Indicateur « Long list »

```
chomage = trouve_indicateurs('unemployment', not_in='male female')
taux_prof = trouve_indicateurs('teacher ratio tertiary')
gpi = trouve_indicateurs('gpi tertiary', not_in='school secondary')
indicateurs_motivations = chomage.append(taux_prof).append(gpi).pipe(indicateur_nb_pays)
```

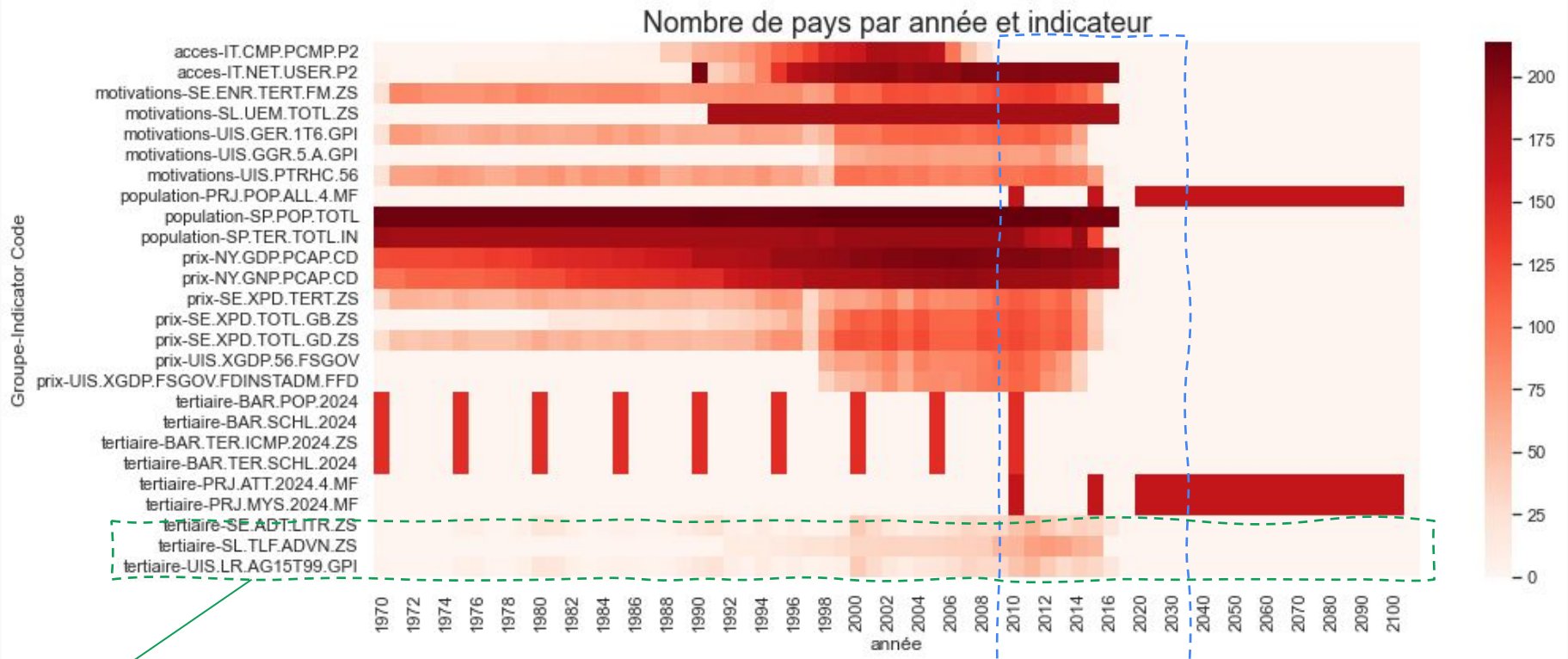
Indicator Name	Indicator Code	2005	2010	2015	2020	2025	2030
Gross enrolment ratio, tertiary, gender parity index (GPI)	SE.ENR.TERT.FM.ZS	143	156	103	0	0	0
Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	211	211	208	0	0	0
Gross enrolment ratio, primary to tertiary, gender parity index (GPI)	UIS.GER.1T6.GPI	136	135	2	0	0	0
Gross graduation ratio from first degree programmes (ISCED 6 and 7) in tertiary education, gender parity index (GPI)	UIS.GGR.5.A.GPI	69	71	3	0	0	0
Pupil-teacher ratio in tertiary education (headcount basis)	UIS.PTRHC.56	89	99	75	0	0	0



Indicator Name	Indicator Code	Groupe
Wittgenstein Projection: Population in thousands by highest level of educational attainment. Post Secondary. Total	PRJ.POP.ALL.4.MF	population
Population, total	SP.POP.TOTL	population
Population of the official age for tertiary education, both sexes (number)	SP.TER.TOTL.IN	population
Personal computers (per 100 people)	IT.CMP.PCMP.P2	acces
Internet users (per 100 people)	IT.NET.USER.P2	acces
Barro-Lee: Population in thousands, age 20-24, total	BAR.POP.2024	tertiaire
Barro-Lee: Average years of total schooling, age 20-24, total	BAR.SCHL.2024	tertiaire
Barro-Lee: Percentage of population age 20-24 with tertiary schooling. Total (Incomplete and Completed Tertiary)	BAR.TER.ICMP.2024.ZS	tertiaire
Barro-Lee: Average years of tertiary schooling, age 20-24, total	BAR.TER.SCHL.2024	tertiaire
Wittgenstein Projection: Percentage of the population age 20-24 by highest level of educational attainment. Post Secondary. Total	PRJ.ATT.2024.4.MF	tertiaire
Wittgenstein Projection: Mean years of schooling. Age 20-24. Total	PRJ.MYS.2024.MF	tertiaire
Adult literacy rate, population 15+ years, both sexes (%)	SE.ADT.LITR.ZS	tertiaire
Labor force with advanced education (% of total)	SL.TLF.ADVN.ZS	tertiaire
Adult literacy rate, population 15+ years, gender parity index (GPI)	UIS.LR.AG15T99.GPI	tertiaire
Gross enrolment ratio, tertiary, gender parity index (GPI)	SE.ENR.TERT.FM.ZS	motivations
Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	motivations
Gross enrolment ratio, primary to tertiary, gender parity index (GPI)	UIS.GER.1T6.GPI	motivations
Gross graduation ratio from first degree programmes (ISCED 6 and 7) in tertiary education, gender parity index (GPI)	UIS.GGR.5.A.GPI	motivations
Pupil-teacher ratio in tertiary education (headcount basis)	UIS.PTRHC.56	motivations
GDP per capita (current US$)	NY.GDP.PCAP.CD	revenu
GNI per capita, Atlas method (current US$)	NY.GNP.PCAP.CD	revenu
Expenditure on tertiary as % of government expenditure on education (%)	SE.XPD.TERT.ZS	revenu
Expenditure on education as % of total government expenditure (%)	SE.XPD.TOTL.GB.ZS	revenu
Government expenditure on education as % of GDP (%)	SE.XPD.TOTL.GD.ZS	revenu
Government expenditure on tertiary education as % of GDP (%)	UIS.XGDP.56.FSGOV	revenu
Government expenditure in educational institutions as % of GDP (%)	UIS.XGDP.FSGOV.FDINSTADM.FFD	revenu

3.2 Sélection des indicateurs

Indicateur « Long List »



Pas assez de pays à comparer

années avant internet

années trop dans futur

Data	(3665 séries, 242 pays)
↓	
Filtre Data par « Long List »	(26 séries, 242 pays)
↓	
Supprime non-pays	(26 séries, 214 pays)
↓	
Supprime petit pays	(26 séries, 173 pays)
↓	
Ignore pays de bas revenu	(26 séries, 158 pays)
↓	
Filtre années 2010-2035	(65 □ 12 années de data)
↓	
Filtre indicateurs par nb. Pays > 50	(10 séries, 158 pays)

Data

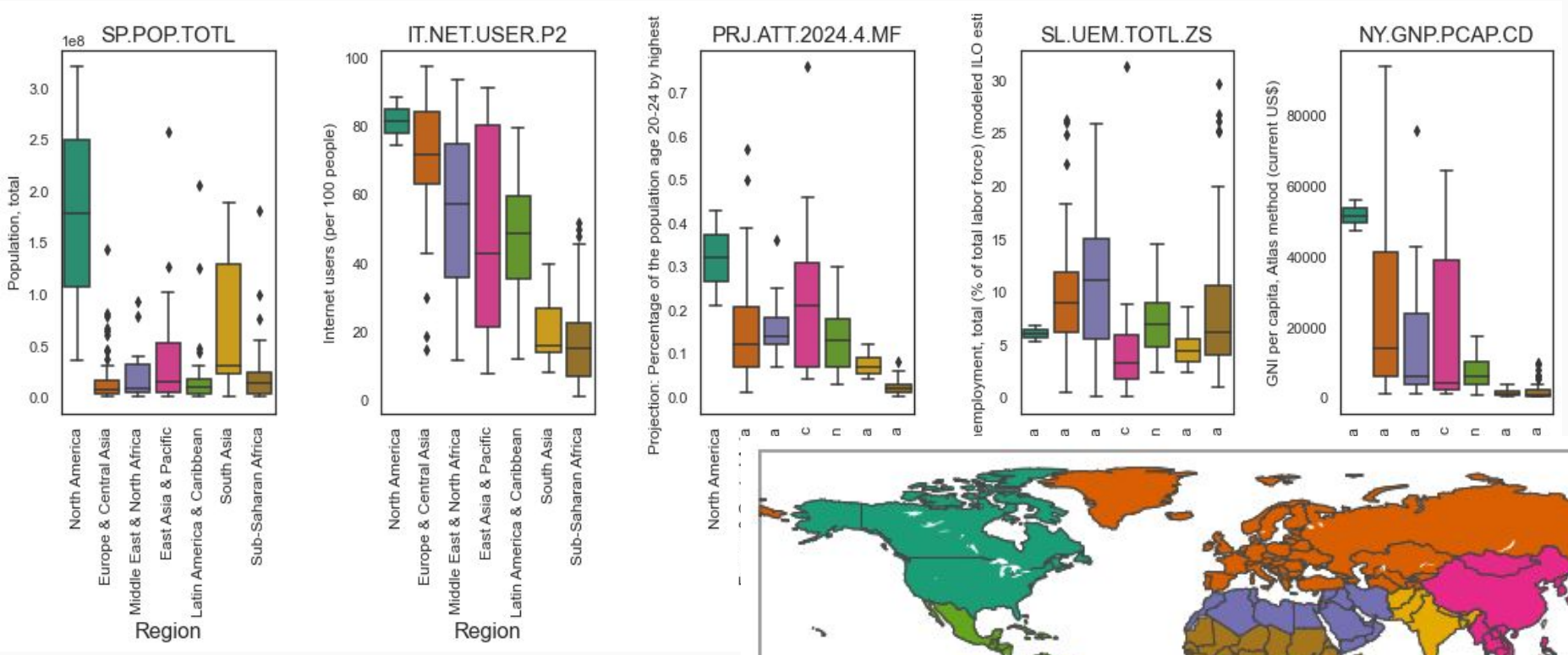
Country

3.2 Sélection des indicateurs

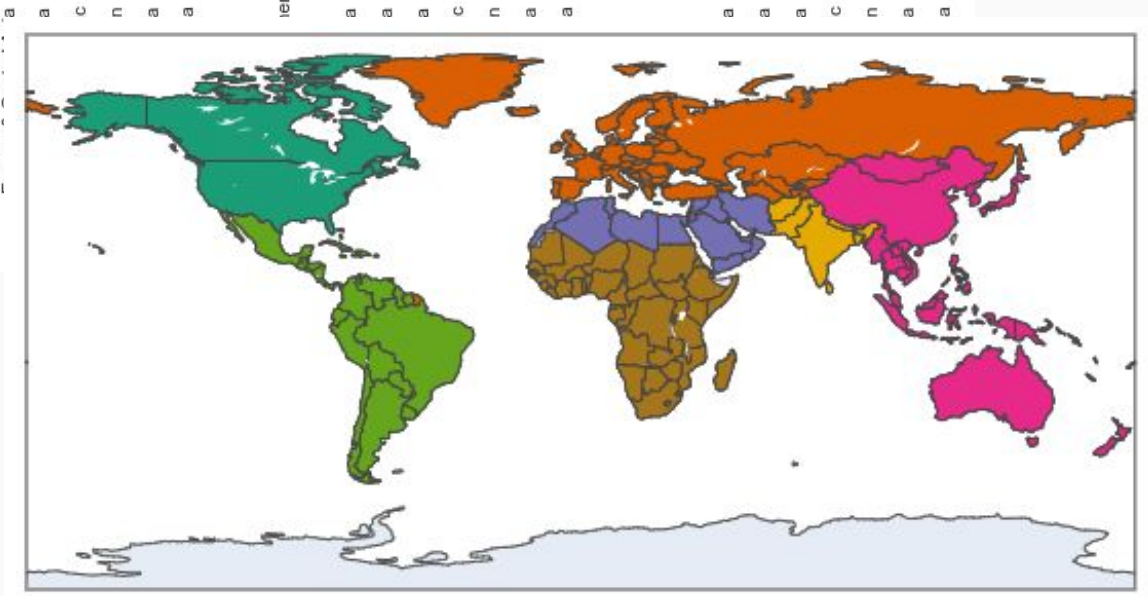
Un indicateur de chaque groupe

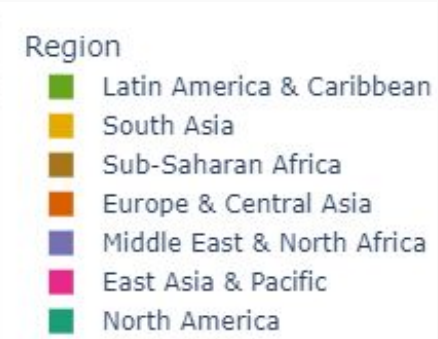
Indicator Name	Indicator Code	Groupe
Population, total	SP.POP.TOTL	population
Internet users (per 100 people)	IT.NET.USER.P2	acces
Wittgenstein Projection: Percentage of the population age 20-24 by highest level of educational attainment. Post Secondary. Total	PRJ.ATT.2024.4.MF	tertiaire
Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	incentives
GNI per capita, Atlas method (current US$)	NY.GNP.PCAP.CD	revenu

4. Analyse Exploratoire - Régions



Grands variations entre pays de même région pour les 5 indicateurs

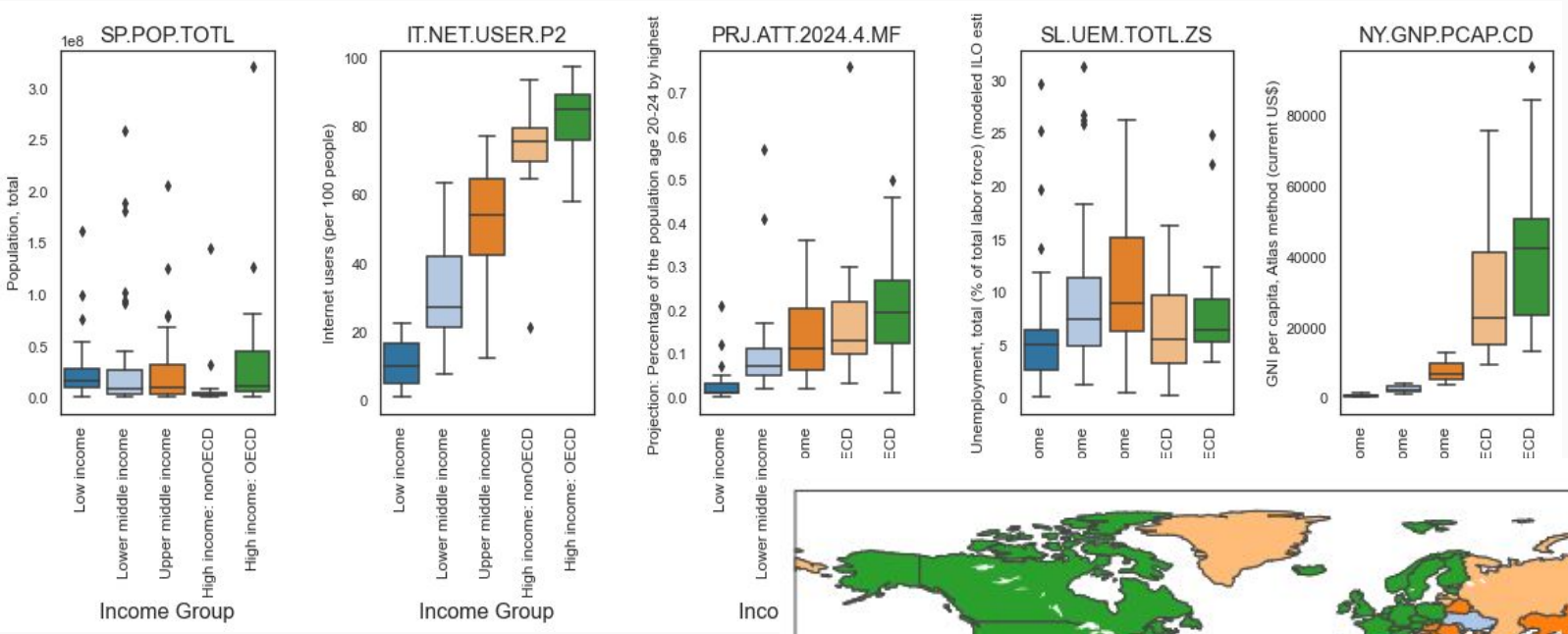




Région
- champ de table “Country”

Pays exclus d’analyse (population >1 milliard):
- Chine (trop fermé politiquement),
- Inde (trop bas GNI per capita):

4. Analyse Exploratoire – Income Groups

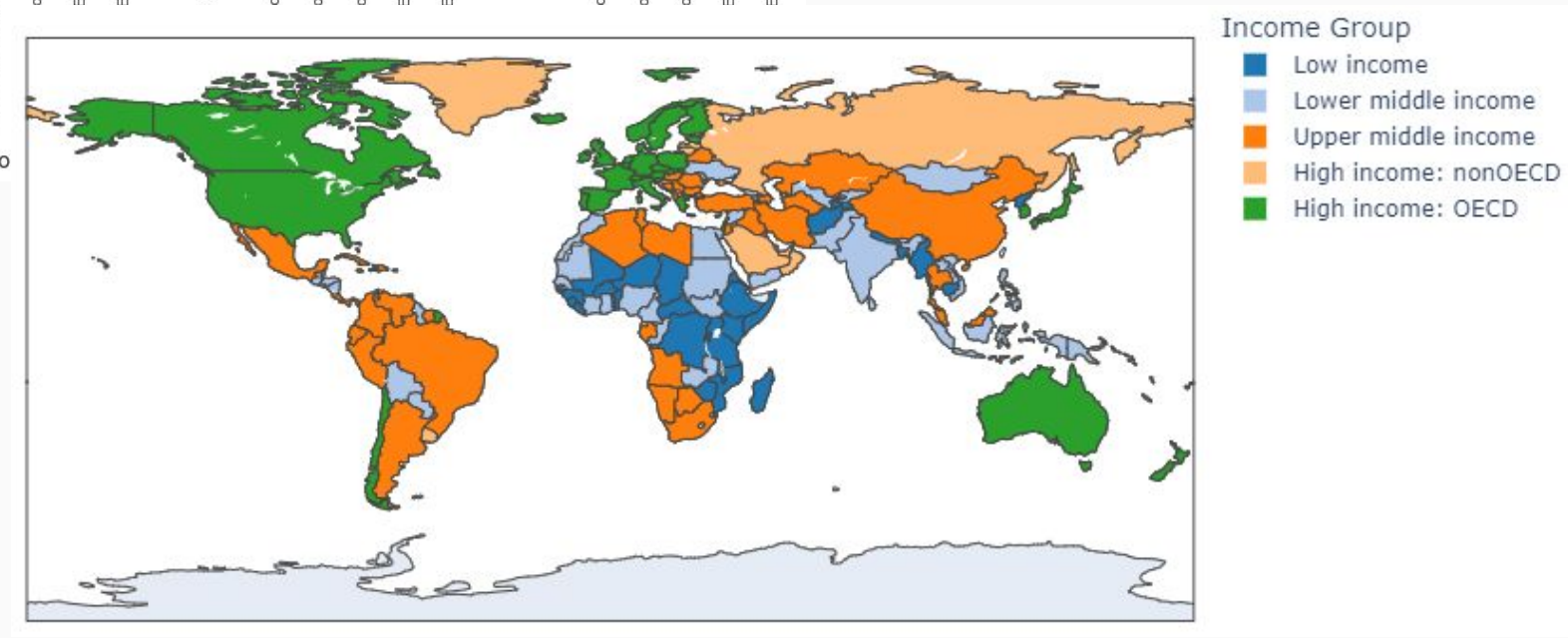


Pays avec haut revenu (GNI per capita):

- plus d'accès internet
- plus grand secteur tertiaire

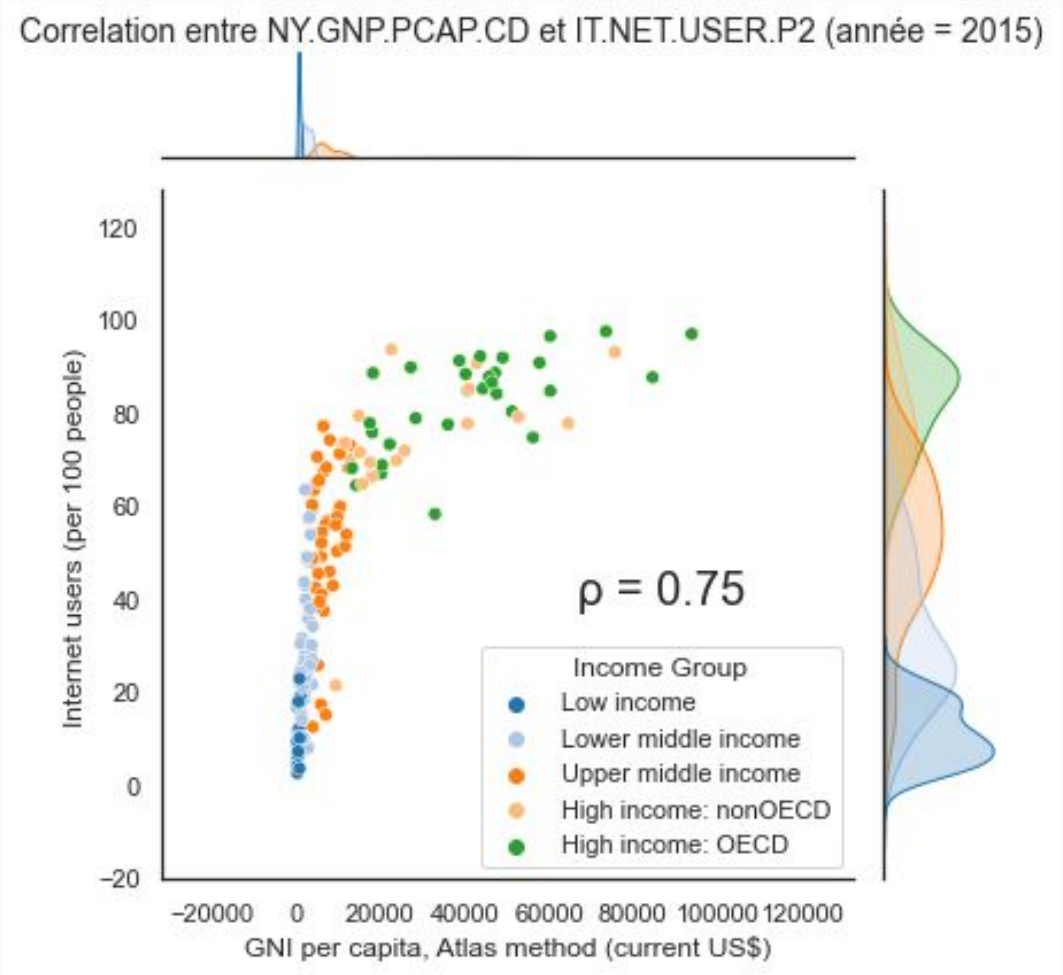
% en chômage plus haut pour les pays riches que les pays de bas revenus

- demande pour des employés plus qualifiés?



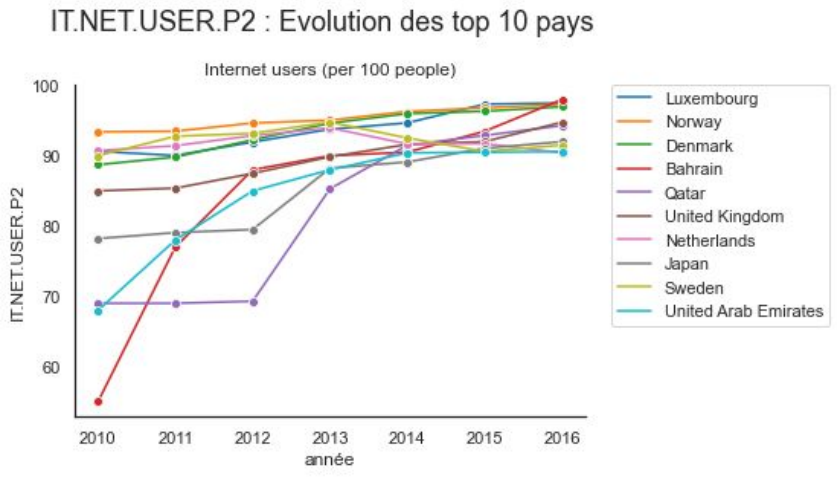
Income Group
- champ de table "Country"

4. Analyse Exploratoire – Corrélations

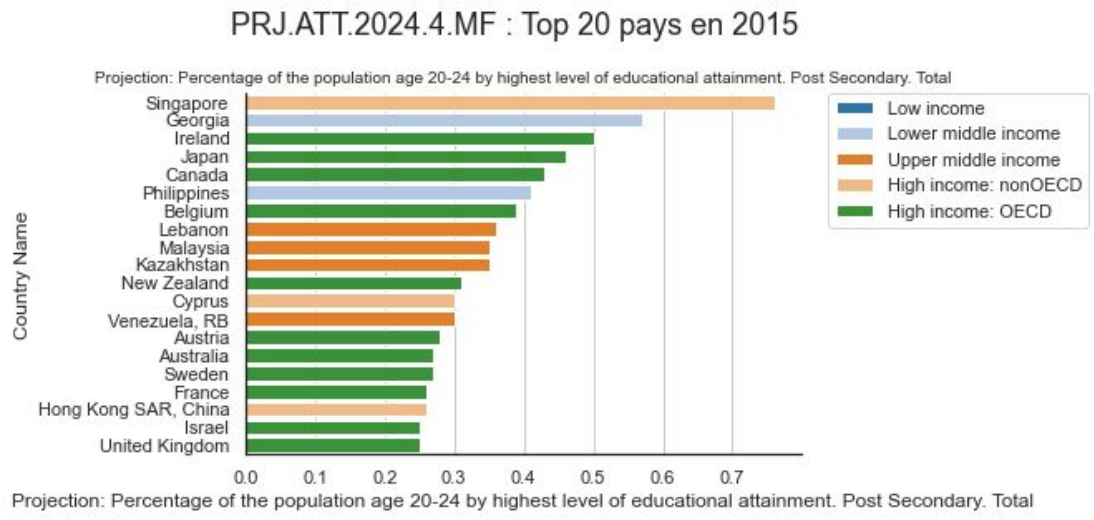


4. Analyse Exploratoire - Évolutions

L'accès à l'internet peut évoluer très vite

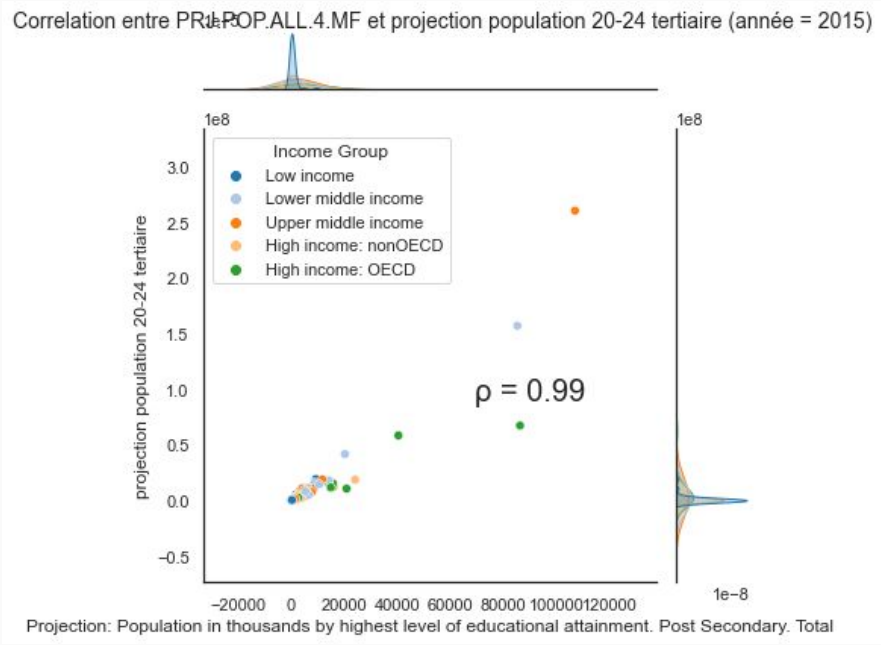


Quelques surprises sur le pourcentage de 20-24 ans dans le secteur tertiaire



La population 20-24 dans le secteur tertiaire est proportionnelle aux populations ayant étudié au niveau tertiaire

--> cet indicateur indique aussi le nombre de gens de tout âge étudiant dans le secteur tertiaire



5. Conclusion – Pays avec un fort potentiel de clients

revenu total = population totale * % avec accès (e-learning) * (alpha * % secteur tertiaire + beta * motivations à e-learning) * prix de vente (gamma * salaire moyen)

Dépenses "fixes"

Pluralsight, Udemy,

Coursera

Udacity, OpenClassrooms

Dépenses proportionnelle aux nombre d'étudiants

BUSINESS MODEL

BUSINESS MODEL

alpha = **% du marché du secteur tertiaire** qui choisit notre entreprise (0 à 100%)

beta = un **facteur d'influence du chômage** / ruralité / prix d'éducation pour motiver la population à choisir d'e-learning (0 à 100, à déterminer)

gamma = **% de son revenu** que l'étudiant accepte de payer

revenu par 1000 habitants = population 1000 habitants * % avec accès (e-learning) * (alpha * % secteur tertiaire + beta * motivations à e-learning) * gamma * salaire moyen

5. Conclusion – Pays avec un fort potentiel de clients

Pays avec fort potentiel en gras:

- dans top 7 pays en 2015 et 2010 pour chaque Business Model, et
- dans top 7 pays pour scénario A et B

Cours sans grandes dépenses de suivi d'étudiant

maximiser

revenu total

maximiser

revenu par 1000 habitants

Cours avec dépenses de suivi d'étudiant / marketing par ville ou état

choix de candidat pays est fortement influencé par le business model

Belgium
Canada
Germany
Ireland
Japan
Norway
Sweden
United Kingdom
United States

Business Model 1

Cibler seulement secteur tertiaire

(scénario A)
% marché = 0.1%
% chômeurs = 0%
% revenu étudiant = 2%

2015
United States,
Japan,
Canada,
United Kingdom,
France,
Germany,
Australia,

Business Model 3

Cibler aussi les reconversions

(scénario B)
% marché = 0.1%
% chômeurs = 0.9%
% revenu étudiant = 2%

2015
United States,
Japan,
France,
Spain,
United Kingdom,
Germany,
Canada

Business Model 2

Cibler seulement secteur tertiaire

(scénario A)
% marché = 0.1%
% chômeurs = 0%
% revenu étudiant = 2%

2015
Singapore,
Ireland,
Canada,
Norway,
Japan,
Belgium,
Sweden,

Business Model 4

Cibler aussi les reconversions

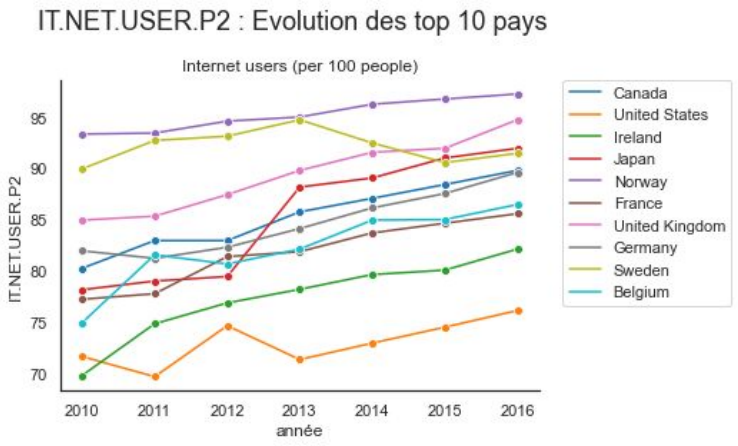
(scénario B)
% marché = 0.1%
% chômeurs = 0.9%
% revenu étudiant = 2%

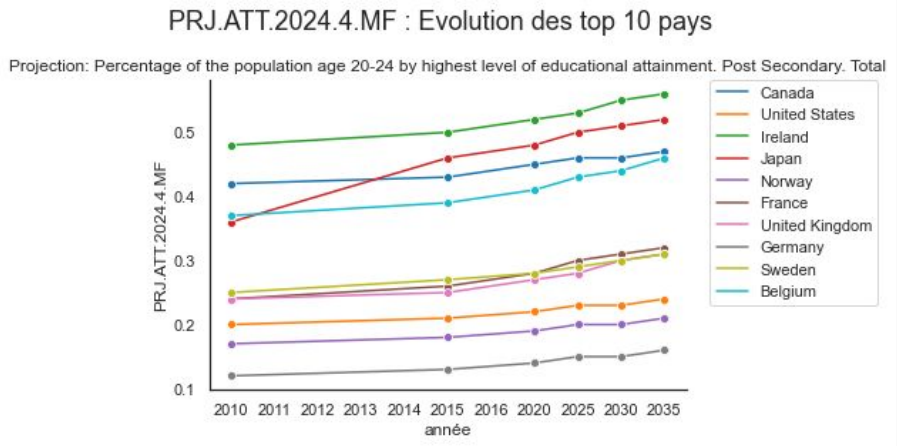
2015
Ireland,
Luxembourg,
Norway,
Spain,
Sweden,
Canada,
Belgium,

5. Conclusion – Évolution de ce potentiel de clients

Pour chacun de ces pays, quelle sera **l'évolution de ce potentiel** de clients ?

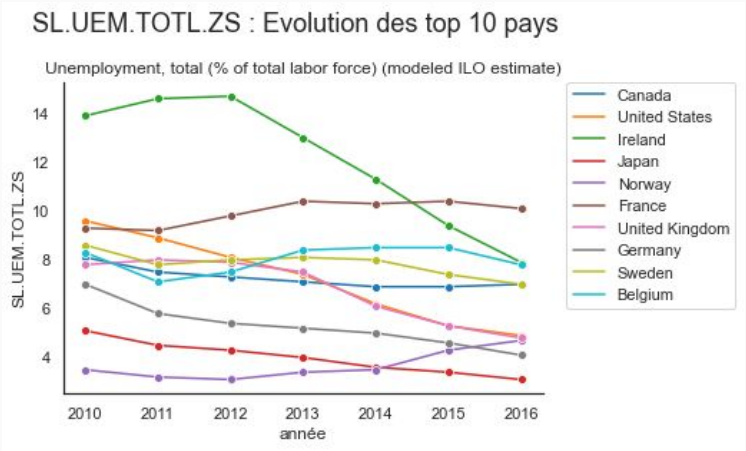
revenu = population * % avec accès (e-learning) * % secteur tertiaire + motivations à e-learning * prix de vente

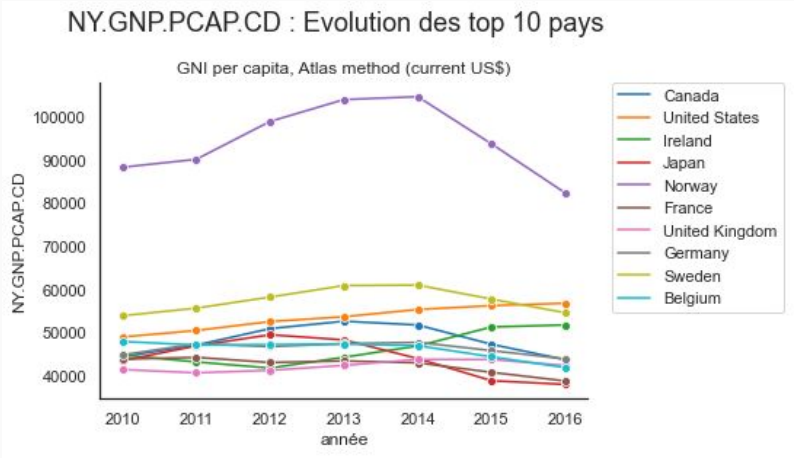




La population reste presque constante dans ces pays développés

Le % de 20-24 ans qui se forme augmente dans tous ces pays





L'évolution de ce potentiel dépend de :

- L'évolution du pouvoir d'achat des clients
- L'évolution du chômage dans ces pays

- On n'a pas de projections de ces 2 indicateurs, qui peuvent changer beaucoup sur 5 ans

5. Conclusion – Pays à opérer en priorité

Dans **quels pays** l'entreprise doit-elle opérer en **priorité** ?

Sélectionne les pays où :

- **Il parle la même langue**
  - **(Partager des cours/ professeurs)**
- **Le prix de vente est similaire**
  - **(pour avoir un seul prix)**

Irlande
Canada
Royaume Uni
Les Etats Unis

maximiser revenu total

maximiser revenu par habitant

rank	Country Code	Country Name	nb_etudiants	prix_de_vente	revenu	revenu_1000_habitants
0	IRL	Ireland	1874	1026	2	411
1	CAN	Canada	13638	945	13	360
2	NOR	Norway	904	1877	2	327
3	JPN	Japan	53255	776	41	325
4	BEL	Belgium	3740	887	3	294
5	SWE	Sweden	2397	1155	3	283
6	GBR	United Kingdom	14980	874	13	201
7	FRA	France	14671	815	12	179
8	USA	United States	50241	1125	57	176
9	DEU	Germany	9301	916	9	104

Conclusion: Améliorations et études à faire

Mieux comprendre le business model:

Quelles sont les raisons pour lesquelles les étudiants font le choix de e-learning

- coûte moins cher que l'université traditionnelle
- ruralité (manque d'accès)
- inégalité d'opportunités pour les femmes ou les plus âgés
- retour au marché du travail (après chômage / enfants / maladie)
- reconversions (satisfaction)

Marché ciblé dans chaque pays

- prix de vente mondial, ou par pays?

Dépenses de l'entreprise (coûts de traduction, mentorat, impôts)

Questions?