P2 – Analysez des données de systèmes éducatifs

Charlotte DUBUS – Parcours Data Scientist

SOMMAIRE

I. Mission

LA BANQUE MONDIALE

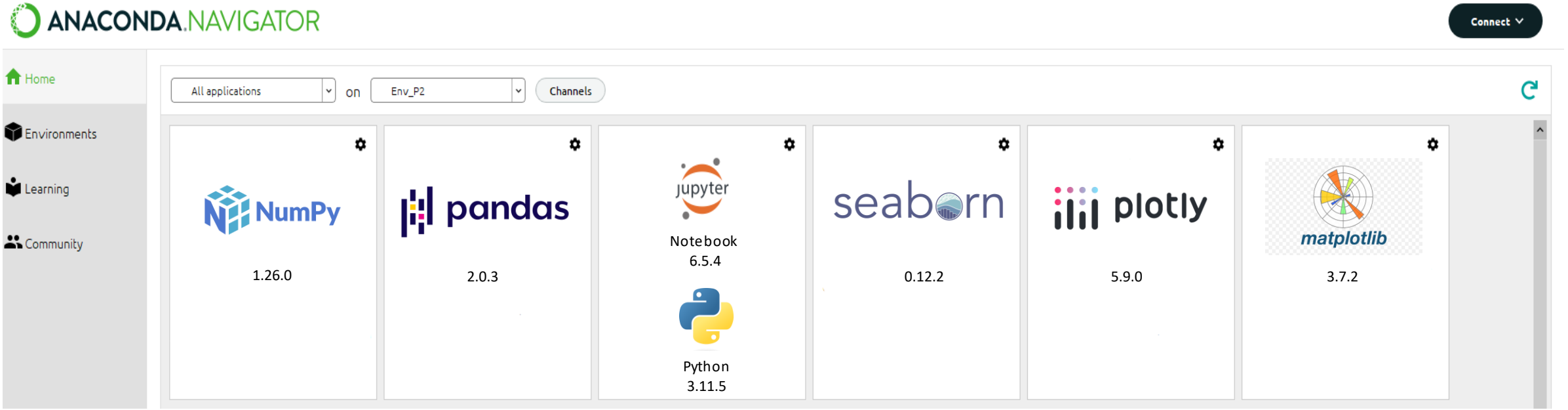
A partir des données de la Banque Mondiale
& des 4 000 indicateurs de l'organisme
"Ed Stats All Indicator Query"

° Réaliser une pré-analyse du jeu de donnée

Valider la qualité du jeu	Décrire les informations contenues	Sélectionner les informations pertinentes	Déterminer des ordres de grandeurs, des indicateurs statistiques pour les zones géographiques & pays du monde

° Explorer les résultats pour faire ressortir les pays avec un fort potentiel

II. ENVIRONNEMENT DE TRAVAIL



- Insertion d'un "Table of Content" sur le terminal de commande via pip
- Affichage personnalisée des lignes & colonnes dans Jupyter Notebook selon les étapes du projet

III. Présentation du jeu de données

Country	Country-Series	Data	FootNote	Series
241 lignes & 32 colonnes	613 lignes & 4 colonnes	886 930 lignes & 70 colonnes	643 638 lignes & 5 colonnes	3 665 lignes & 21 colonnes
28 colonnes: object 4 colonnes: float64	3 colonnes: object 1 colonne: float64	4 colonnes: object 66 colonnes: float64	4 colonnes: object 1 colonne: float64	15 colonnes: object 6 colonnes: float64
Données géographiques, économiques & démographiques	211 pays 21 indicateurs	3 665 indicateurs différents 242 pays	1 558 indicateurs 56 années de référence 239 pays	37 thèmes différents 21 sources distinctes
Pays du monde : 7 régions & 5 groupes de revenus	2 types d'indicateurs : SP.[…] données démographiques NY.[…] données économiques	De 1970 à 2017 : données annuelles De 2020 à 2100 : données de prédictions tous les 5 ans		3 665 indicateurs avec descriptions & sources

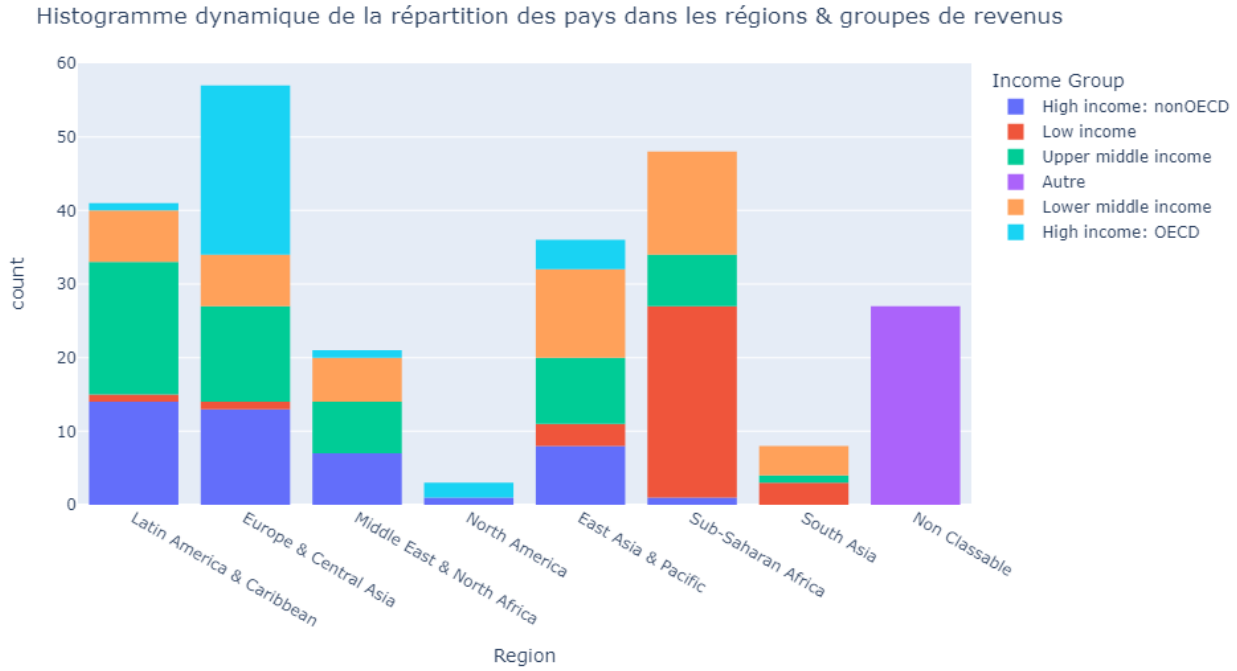
IV. Analyse pré-exploratoire

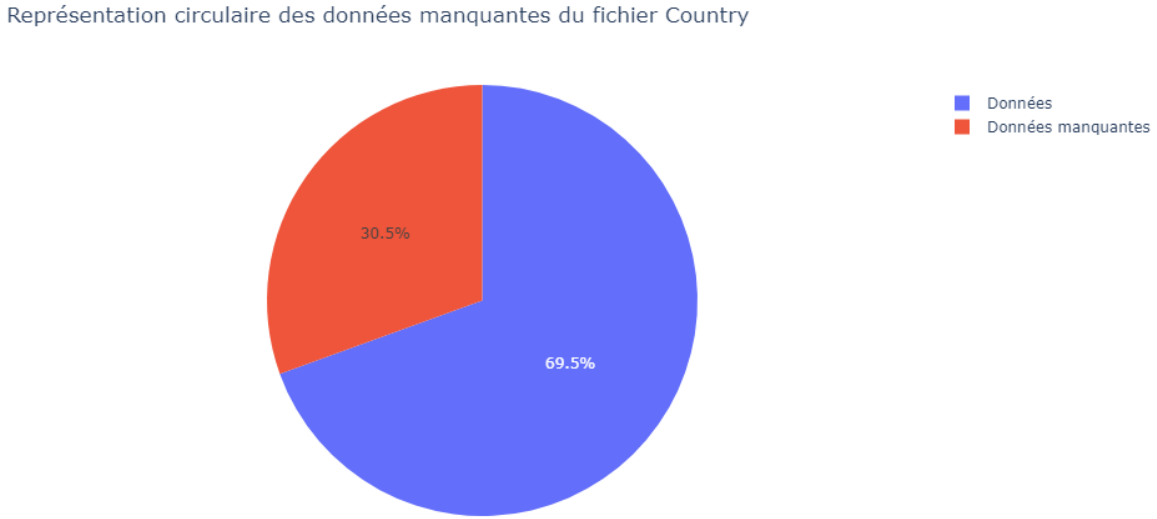
- 3 fichiers avec - de 40% de données manquantes
  Country, Country-Series, FootNote

- 2 fichiers avec + de 70 % de données manquantes
  Data, Series

- Aucun fichier avec des doublons

Country



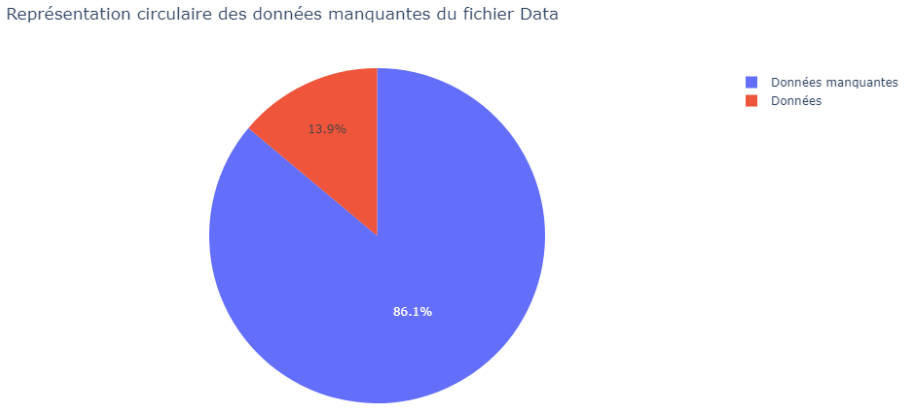


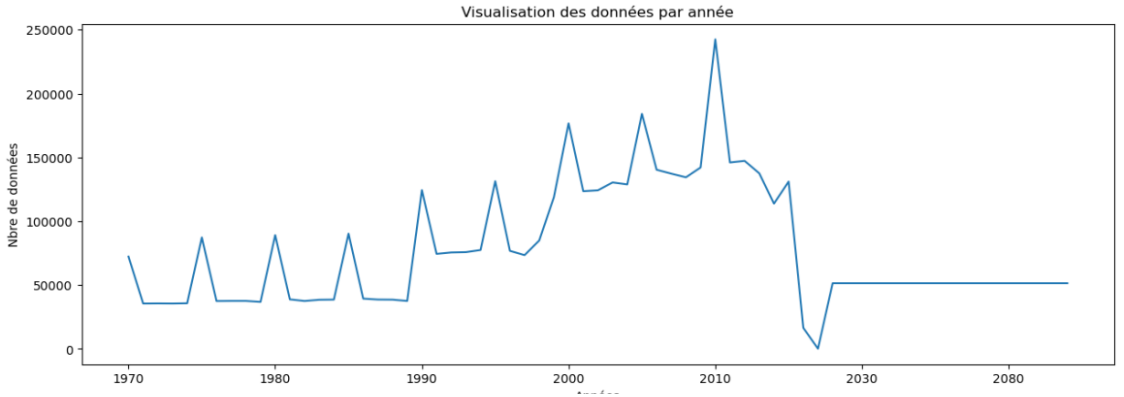
4 colonnes complètes : *Country Code, Table Name, Short Name & Long Name*

27 données à la fois non classable *(région)* & autres *(revenus)* => **supprimer**

30,5% de données manquantes.
4 colonnes vides à **+ de 75%** dont une totalement

Data





4 colonnes avec 100% de données : informations pertinentes exploitables en liaison avec les autres fichiers

- Country Code, Country Name, Indicator Name & Indicator Code

Observation d'un pic de données plus important tous les 5 ans

- Dernières années hors prédiction : 2016 & 2017 inexploitables car très peu de données
- 2000, 2005 et 2010 sont les années les plus remplies => **Utilisation pour notre analyse**
- **2015 :** année écartée pour la suite

Country Series - FootNote - Series

Country Series

- 1 colonne vide : Unnamed : 3
- **Reste des données : 100 % remplies**

FootNote

- Unnamed : 4 sans donnée
- **Autres colonnes pleinement renseignées**

Series

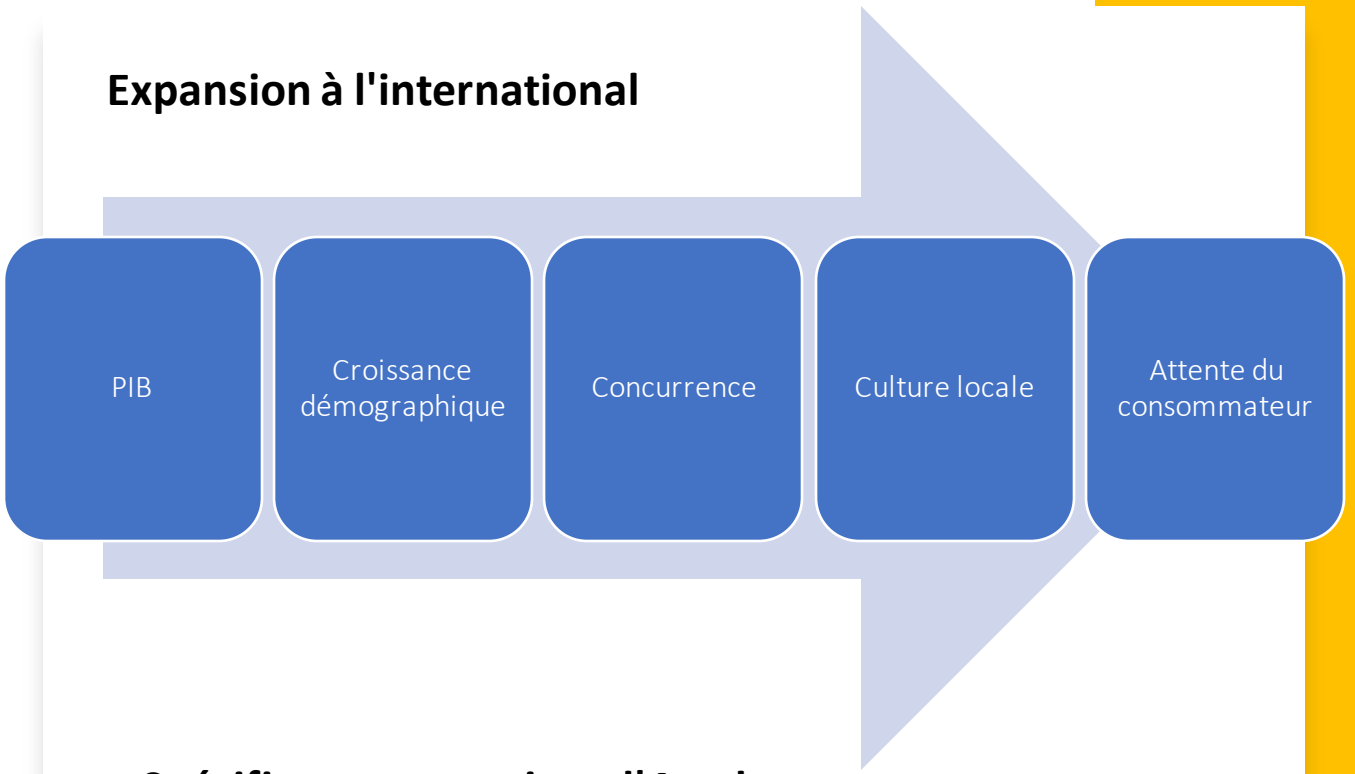
- 6 colonnes vides
- **5 colonnes complètes**
- Autres colonnes : + de 80-90% de données absentes *(exception : "Short Definition")*

Filtrage de données

	Country Code	Table Name	Region	Income Group	Indicator Name	Indicator Code	2000	2005	2010
0	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	56.156502	59.058441	NaN
1	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	59.691120	62.097809	NaN
2	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	1.133350	1.109620	NaN
3	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	52.667679	55.963299	NaN
4	ABW	Aruba	Latin America & Caribbean	High income: nonOECD	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	98.253593	97.978920	98.920464

Etude des indicateurs

- 3 665 indicateurs
- Filtrage du DataFrame : utilisation de la France *(vérification de la présence de l'ensemble des indicateurs)*
- Country-Series : préfixe des indicateurs démographiques *(SP.)* & économique *(NY.)*
- Recherche par mots-clés dans Indicator Name ou Indicator Code *(Data)*



6 indicateurs retenus

	Series Code	Topic	Indicator Name	Short definition	Long definition
1664	NY.GDP.PCAP.PP.CD	Economic Policy & Debt: Purchasing power parity	GDP per capita, PPP (current international $)	NaN	GDP per capita based on purchasing power parit...

	Series Code	Topic	Indicator Name	Short definition	Long definition
2506	SP.POP.1524.TO.UN	Population	Population, ages 15-24, total	Population, ages 15-24, total is the total pop...	Population, ages 15-24, total is the total pop...

	Series Code	Topic	Indicator Name	Short definition	Long definition
611	IT.NET.USER.P2	Infrastructure: Communications	Internet users (per 100 people)	NaN	Internet users are individuals who have used t...

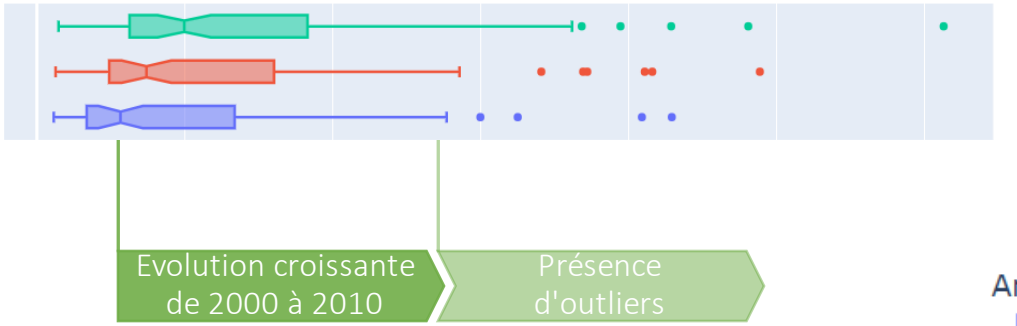
	Series Code	Topic	Indicator Name	Short definition	Long definition
610	IT.CMP.PCMP.P2	Infrastructure: Communications	Personal computers (per 100 people)	NaN	Personal computers are self-contained computer...

	Series Code	Topic	Indicator Name	Short definition	Long definition
2307	SE.SEC.ENRR	Secondary	Gross enrolment ratio, secondary, both sexes (%)	NaN	Total enrollment in secondary education, regar...

	Series Code	Topic	Indicator Name	Short definition	Long definition
2335	SE.TER.ENRR	Tertiary	Gross enrolment ratio, tertiary, both sexes (%)	NaN	Total enrollment in tertiary education (ISCED ...

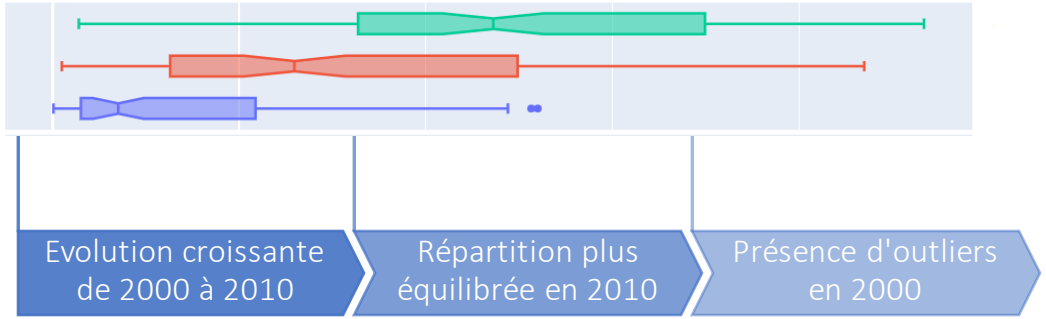
Pertinence des indicateurs choisis

Représentation graphique/statistique du PIB pour chaque année

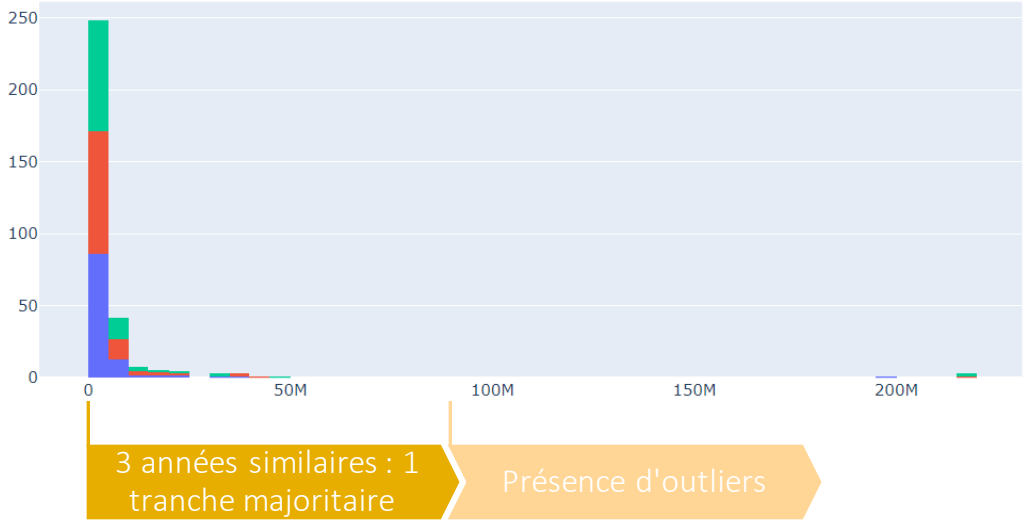


Année
Données_2000
Données_2005
Données_2010

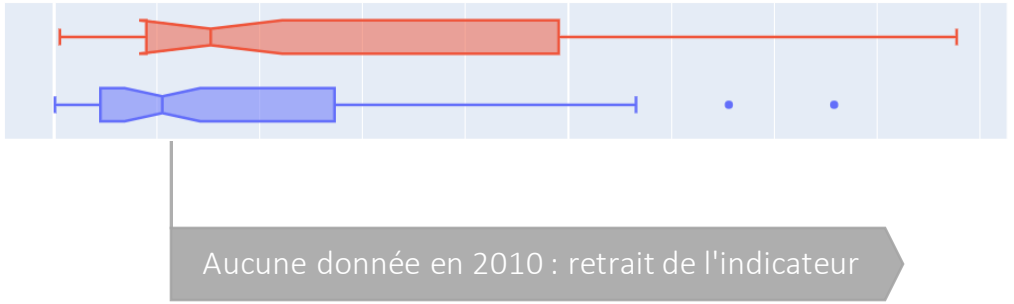
Représentation graphique/statistique du nombre d'utilisateurs internet (pour 100 habitants)



Représentation graphique/statistique de la population de 15-24 ans pour chaque année

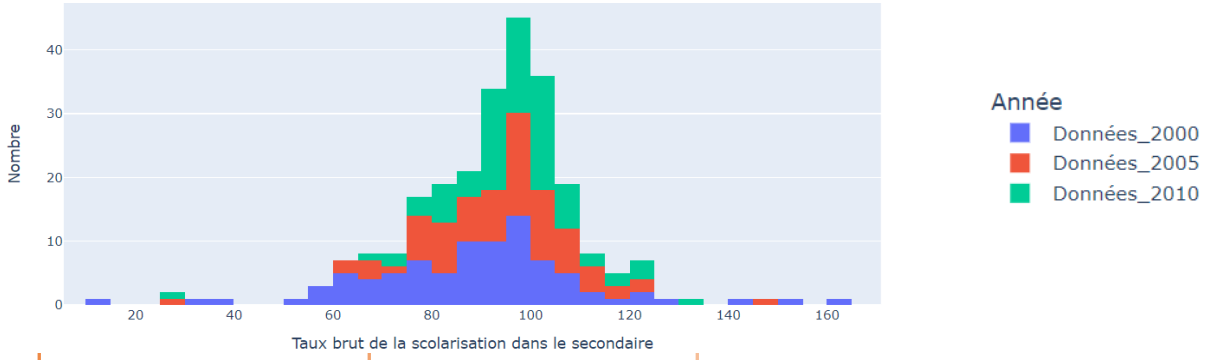


Représentation graphique/statistique du nombre de personne détenant un ordinateur personnel



Pertinences des indicateurs choisis

Représentation graphique/statistique de la scolarisation brute dans le secondaire pour chaque année

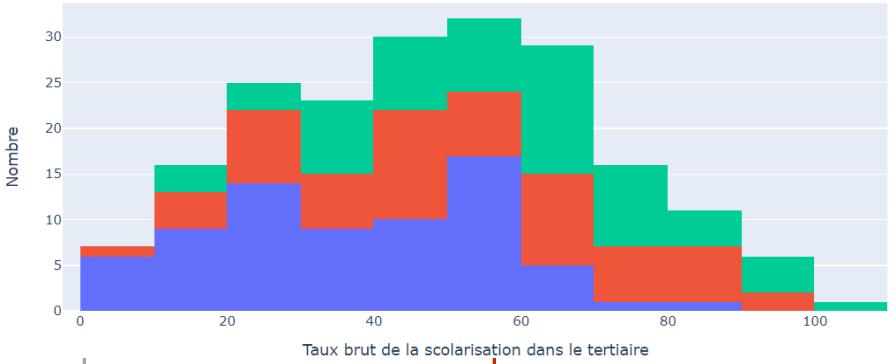


Evolution croissante de 2000 à 2010

Répartition autour d'une valeur majoritaire

Présence d'outliers

Représentation graphique/statistique de la scolarisation brute dans le tertiaire pour chaque année



Evolution progressive de 2000 à 2010

1 tranche majoritaire en 2010

Conclusion :

Suppression d'un indicateur *(possession d'ordinateur personnel)*

Analyse des NaN: 192 pays avec au moins 1 valeur manquante

Filtrage du DataFrame

V. Analyse

	Country Code	Table Name	Region	Income Group	Indicator Name	Indicator Code
count	260	260	260	260	260	260
unique	52	52	6	3	5	5
top	ALB	Albania	Europe & Central Asia	High income: OECD	GDP per capita, PPP (current international $)	NY.GDP.PCAP.PP.CD
freq	5	5	150	135	52	52

	2000	2005	2010
count	2.600000e+02	2.600000e+02	2.600000e+02
mean	7.386532e+05	7.471665e+05	7.474119e+05
std	3.289988e+06	3.428188e+06	3.521921e+06
min	1.140973e-01	6.043891e+00	1.565481e+01
25%	3.815306e+01	5.574002e+01	6.836214e+01
50%	9.327209e+01	9.764711e+01	9.945233e+01
75%	2.684800e+04	3.247561e+04	3.851108e+04
max	3.923406e+07	4.275905e+07	4.513709e+07

- **Création d'un score:**

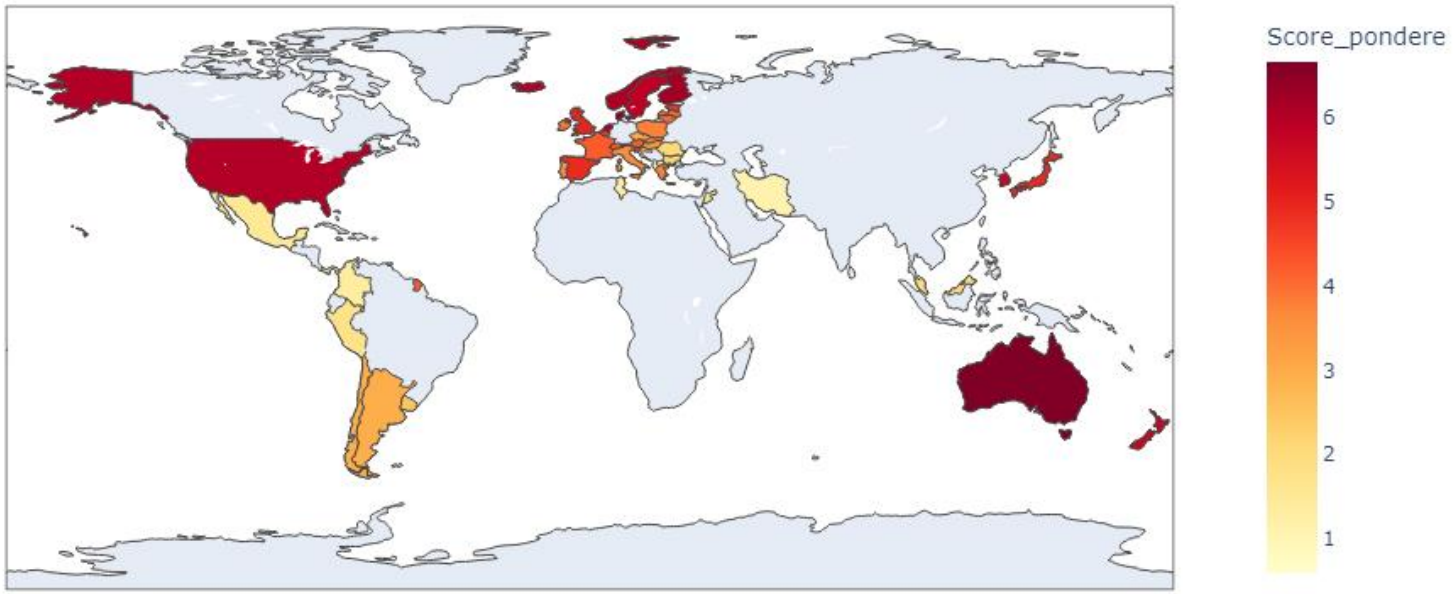
- Calcul de la médiane (moins sensible aux outliers)
- Nouveau DataFrame pour faciliter la visualisation des données ajoutées
- Normalisation des médianes par la méthode min/max

- **Mise en place d'une pondération du score**

1. Population de 15-24 ans & nombre d'utilisateur internet
2. Taux de scolarisation dans le secondaire & tertiaire
3. PIB

- **Exploration des résultats**

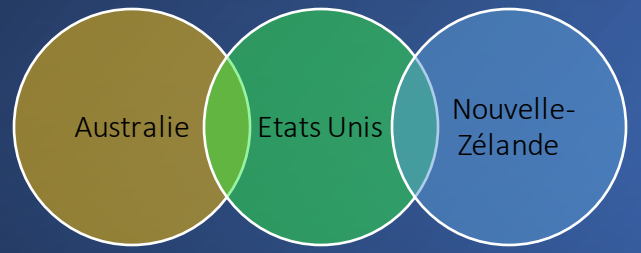
Représentation cartographique des pays selon leur potentiel commercial



Country Code	Table Name	Region	Income Group	Score_pondere
AUS	Australia	East Asia & Pacific	High income: OECD	6.667893
DNK	Denmark	Europe & Central Asia	High income: OECD	6.492019
FIN	Finland	Europe & Central Asia	High income: OECD	6.210455
NOR	Norway	Europe & Central Asia	High income: OECD	6.064878
NLD	Netherlands	Europe & Central Asia	High income: OECD	6.060018
USA	United States	North America	High income: OECD	6.051508
ISL	Iceland	Europe & Central Asia	High income: OECD	5.997596
SWE	Sweden	Europe & Central Asia	High income: OECD	5.870001
NZL	New Zealand	East Asia & Pacific	High income: OECD	5.739459
KOR	Korea, Rep.	East Asia & Pacific	High income: OECD	5.696672
GBR	United Kingdom	Europe & Central Asia	High income: OECD	4.990379
JPN	Japan	East Asia & Pacific	High income: OECD	4.906441
ESP	Spain	Europe & Central Asia	High income: OECD	4.899509
EST	Estonia	Europe & Central Asia	High income: OECD	4.757676
BEL	Belgium	Europe & Central Asia	High income: OECD	4.609539
SVN	Slovenia	Europe & Central Asia	High income: OECD	4.330535
CHE	Switzerland	Europe & Central Asia	High income: OECD	4.246196
FRA	France	Europe & Central Asia	High income: OECD	4.234310
AUT	Austria	Europe & Central Asia	High income: OECD	4.209415
LVA	Latvia	Europe & Central Asia	High income: nonOECD	4.090033

Proposition :

Résultat



Merci de votre attention