



Analysez des données de systèmes éducatifs

Analyse exploratoire des données de la banque mondiale dans le cadre du projet d'**expansion à l'international** de l'entreprise.



Contexte



-  est une **start-up de la EdTech**, qui propose des contenus de formation en ligne pour un public de niveau lycée et universitaire.
-  fournit un jeu de données avec beaucoup de variables concernant des informations hétérogènes.
- Le travail doit permettre de déterminer si ce jeu de données peut informer le board et aider lors de la réflexion sur l'ouverture à de nouveaux pays.





Problématiques

- Quels 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 ?

EDA : ANALYSE EXPLORATOIRE



Description du jeu de données



EdStatsCountry

Informations générales sur l'économie de chaque pays et zones géographiques

Taille : 241 lignes 32 colonnes

Quelques valeurs manquantes

Aucun de doublons + Country, Région et income : utile pour analyse

EdStatCountry-Series

Information sur la sources des données contenues dans EdStatsCountry

Taille : 613 lignes, 4 colonnes

Valeurs manquantes : 0% sauf la colonne Unamed:3 100% null

Aucun doublon

EdStatsData

Evolution des indicateurs des pays de 1970 à 2100

Taille : 886930 lignes 70 colonnes

Nombreuses valeurs manquantes

Aucun doublon : Principale source de données pour analyse

EdStatsSeries

Information sur les indicateurs de EdStatsData

Taille : 3665 lignes 21 colonnes

6 colonnes entièrement vides et beaucoup de valeurs manquantes pour les autres

Aucun doublon : Description des indicateur

EdStatsFootNote

Information sur l'année origine et l'incertitude dans la collecte des données

Taille : 643638 lignes 5 colonnes

Nombres valeurs manquantes : 0% sauf la colonne Unamed:3 100% null

Aucun doublon

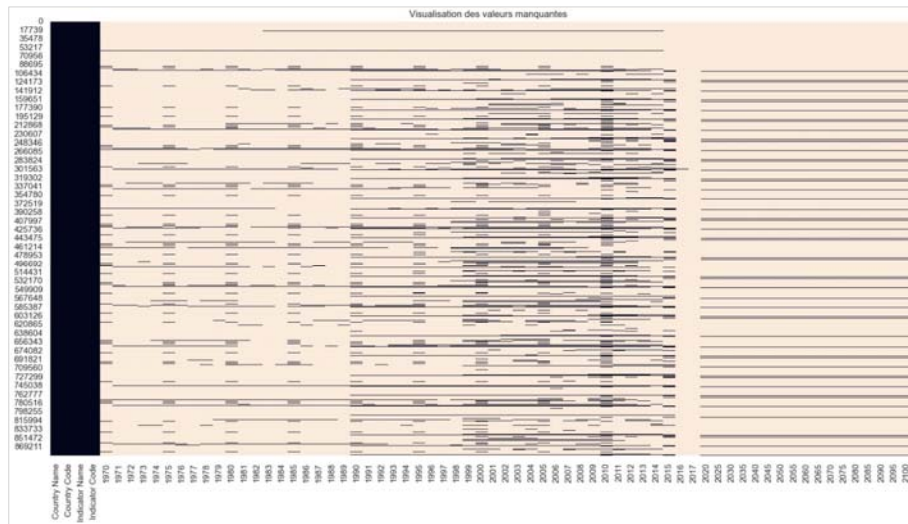


PRÉANALYSE



Démarche de la préparation des données

Les données



Nombre de lignes : 886930 lignes
Nombre de colonnes : 70 colonnes

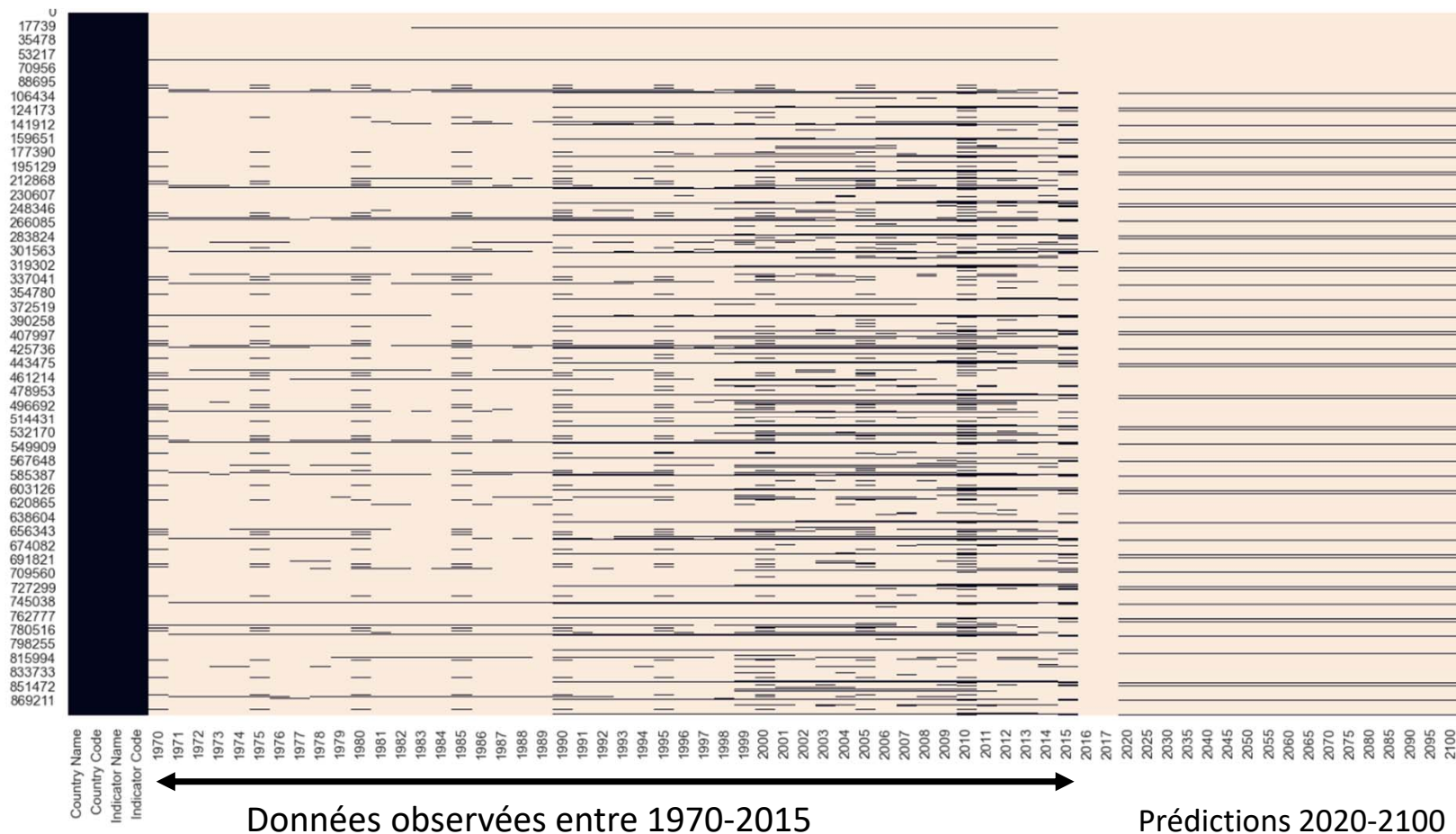
Sélection des données et suppression des valeurs manquantes

1. Réduction temporelle 2010-2015
 2. Filtrer par indicateur pertinents
 3. Gestion des valeurs manquantes
- Suppression des pays sans aucune donnée
Récupération des dernières données

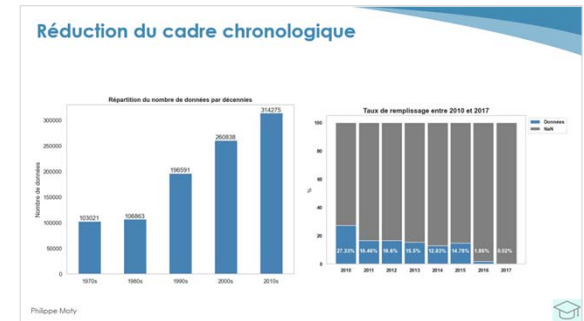
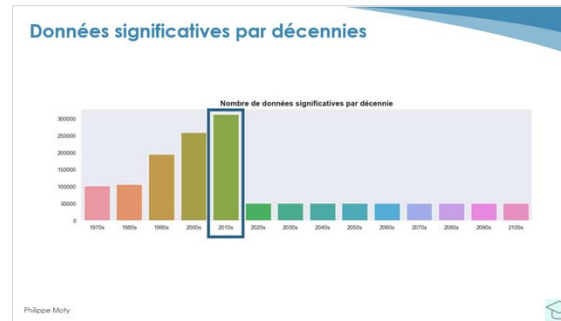


Les données

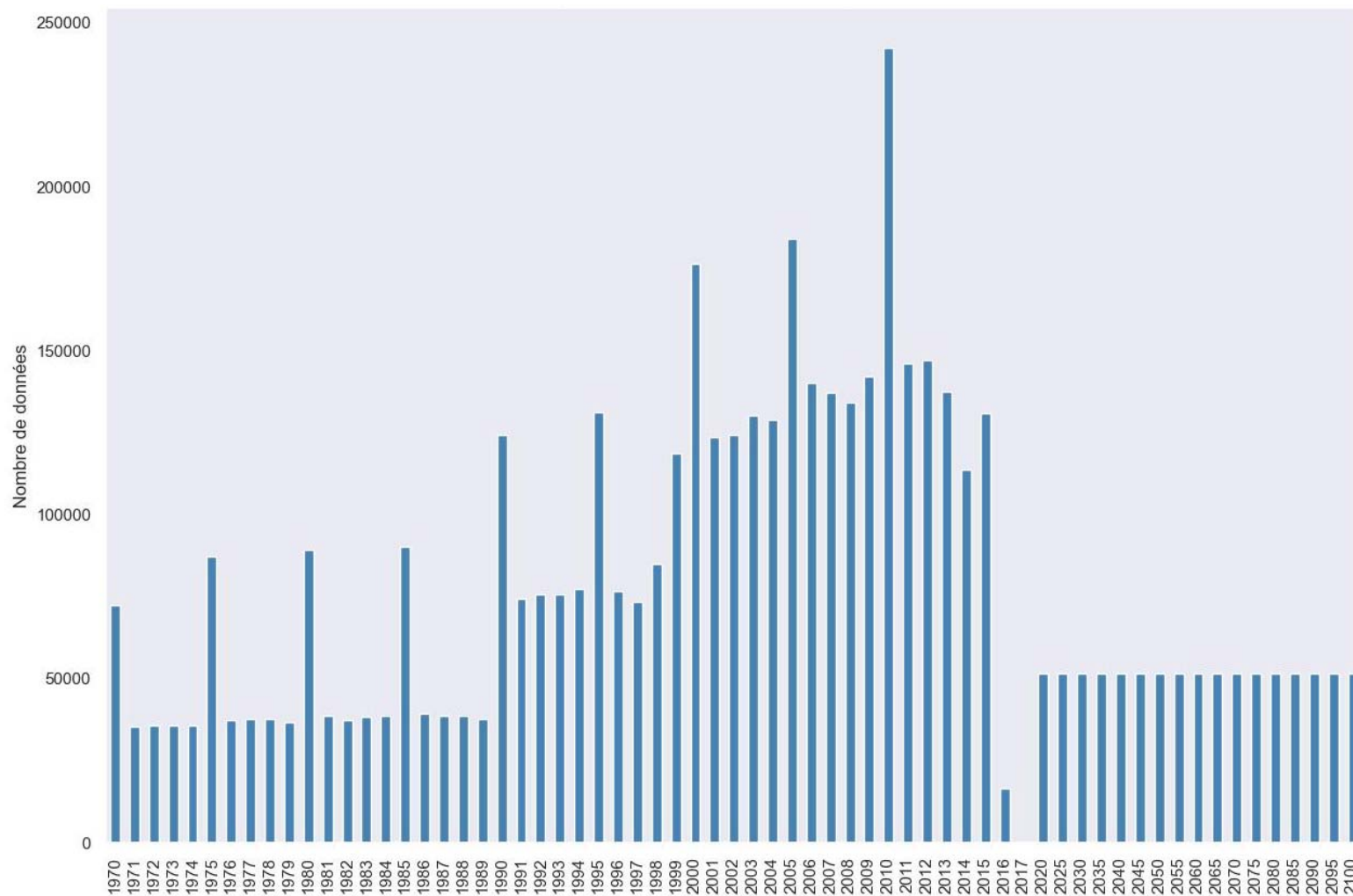
886 930 lignes



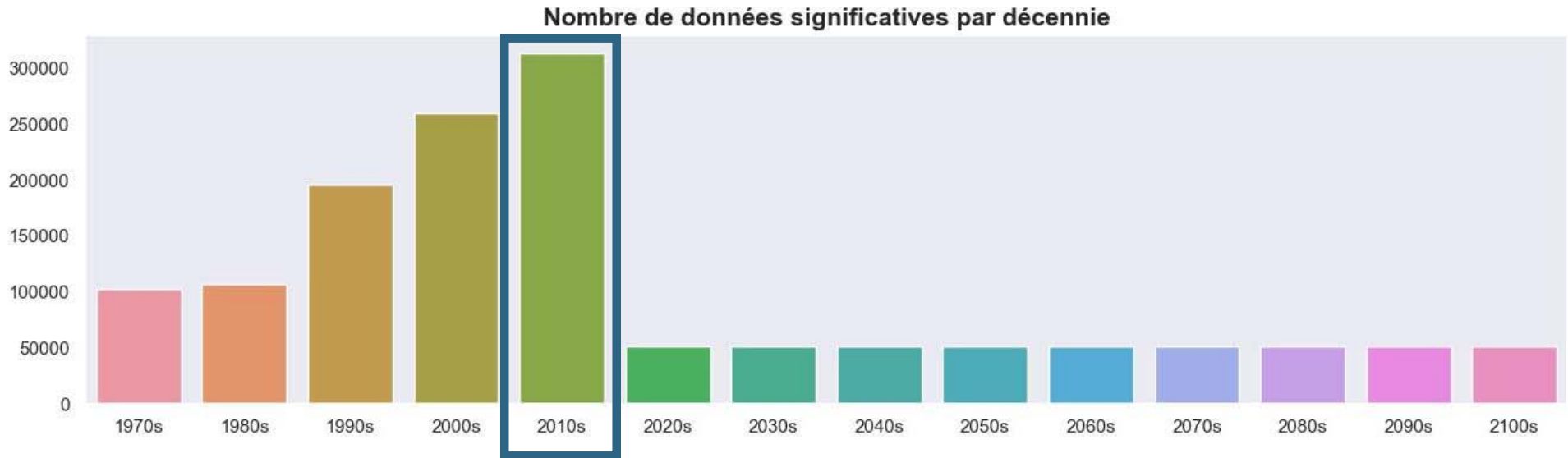
Sélectionner les années pertinentes en fonction du nombre de données disponibles



Nombre de données par année

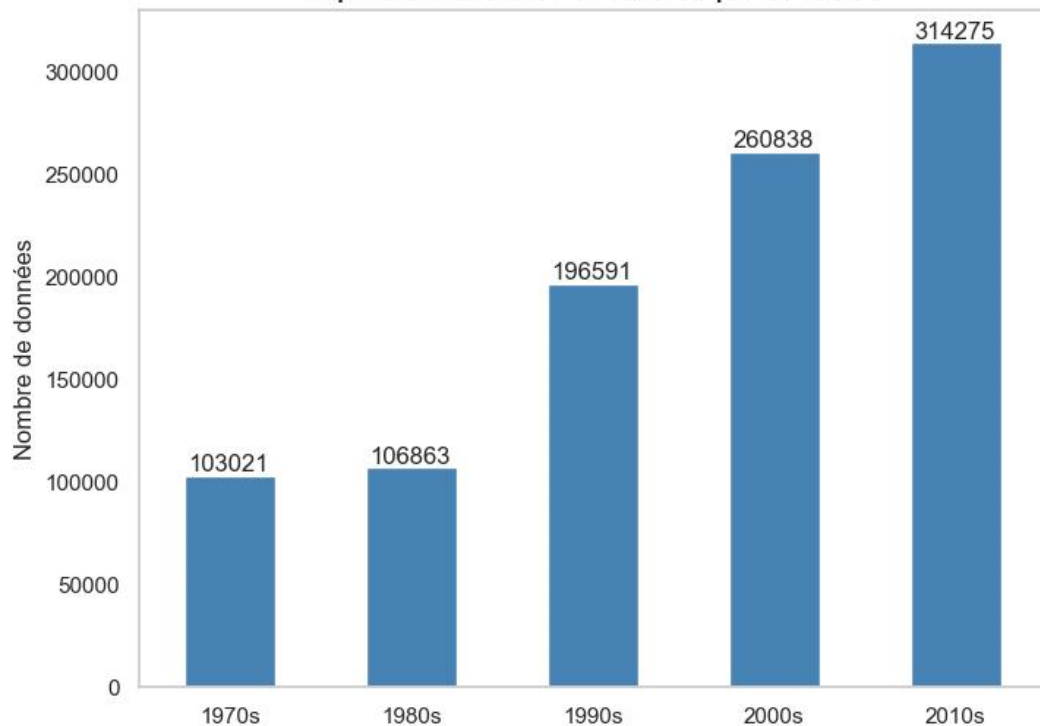


Données significatives par décennies

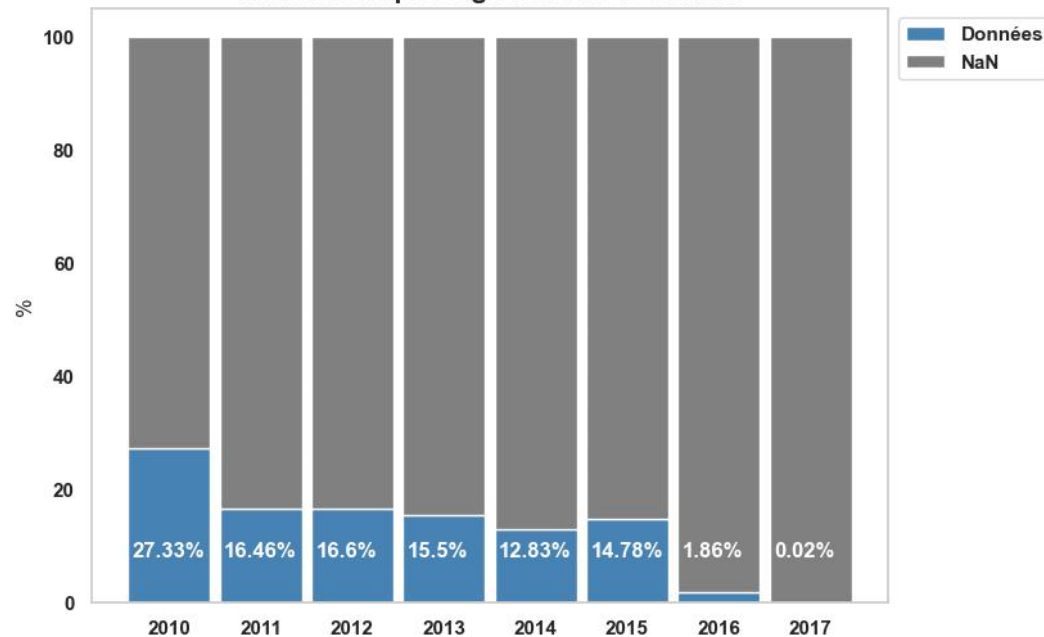


Réduction du cadre chronologique

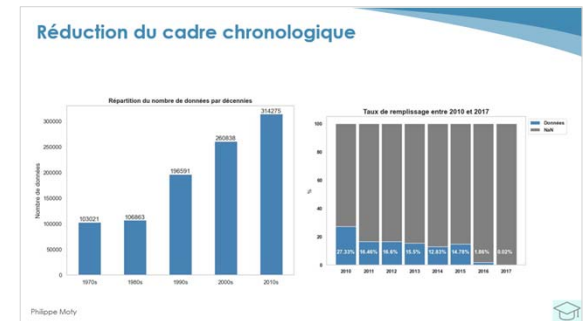
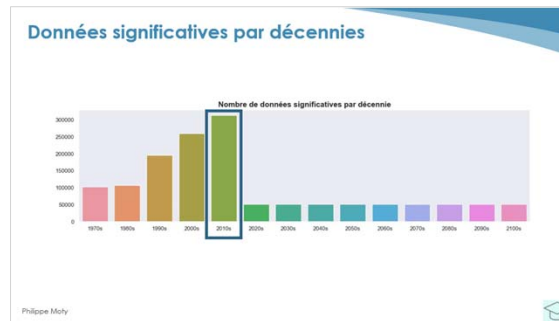
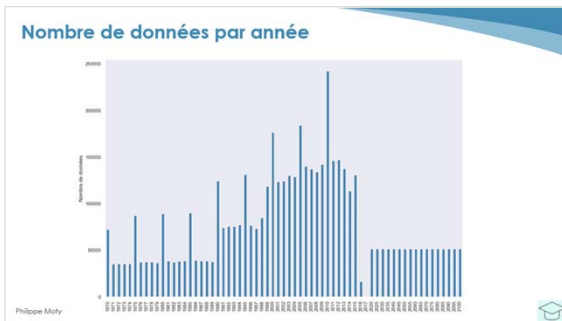
Répartition du nombre de données par décennies



Taux de remplissage entre 2010 et 2017



Sélectionner les années pertinentes en fonction du nombre de données disponibles

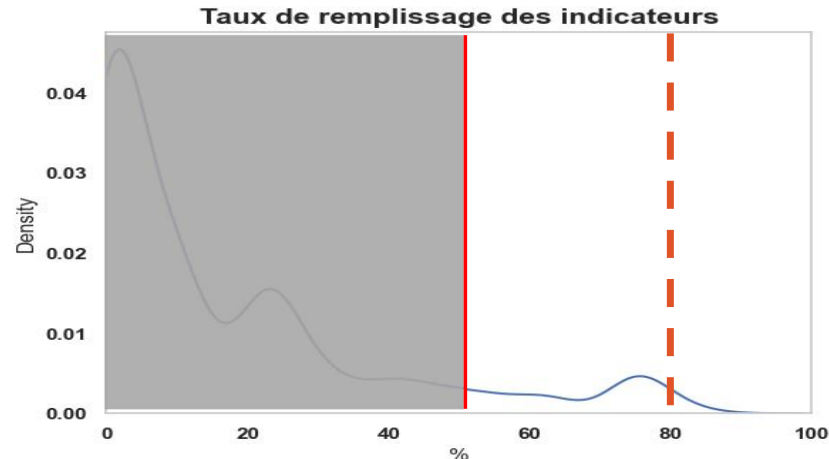


Réduction temporelle 2010-2015



Sélectionner les indicateurs

- Années : **2010-2015**
- Taux de remplissage à 80 % => 16 indicateurs
 - Autre stratégie 50% : 354 / 3665
- Recherche par mots clés
 - identification des indicateurs pertinents



Domaine	Description	Indicator Code	% remplissage
Numérique	Connection internet, ordinateur ou technologie mobile	IT.NET.USER.P2	83%
Economie	Revenu par habitant en parité de pouvoir d'achat	NY.GNP.PCAP.PP.CD	76%
Education	Services éducatifs (lycée & études supérieures)	SE.SEC.ENRR	56%
		SE.TER.ENRR	51%
Population	Population cible : 15-24 ans (lycée & études supérieures)	SP.POP.1524.TO.UN	57%



Collecte des dernières données

→ 2015

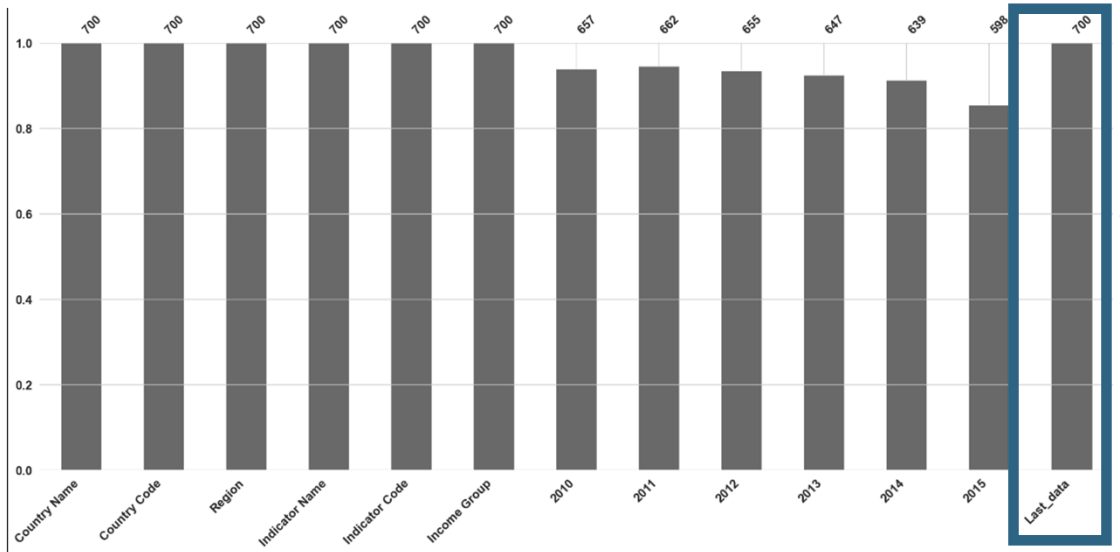
→ Identification des valeurs manquantes (NA) puis suppression des pays pour lesquels il nous manque des données

→ On supprime des petits pays mais aussi le Canada

→ Récupération des dernières données

→ Sur les 102 données manquantes de notre df

→ 100% proviennent de la décennie 2010



```
Last_Data données manquantes : 102
```

```
-----
```

```
Origines des dernières données :
```

```
2014 : 55
```

```
2013 : 20
```

```
2012 : 15
```

```
2011 : 8
```

```
2010 : 4
```

```
restant : 0
```

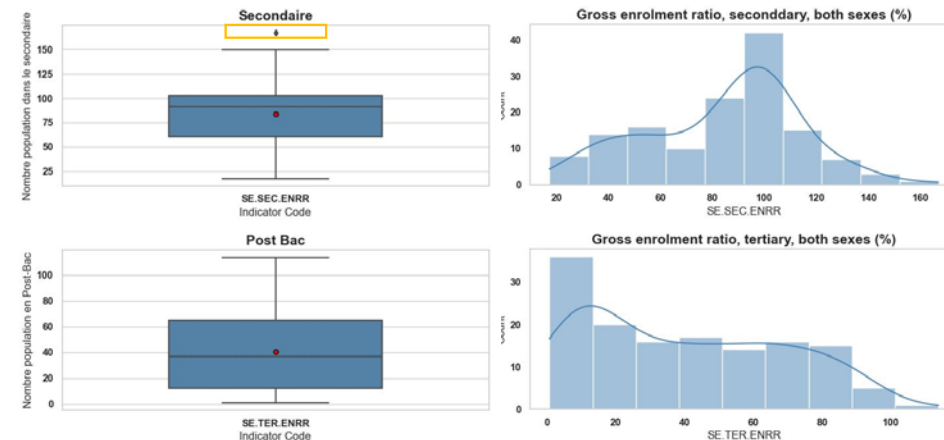
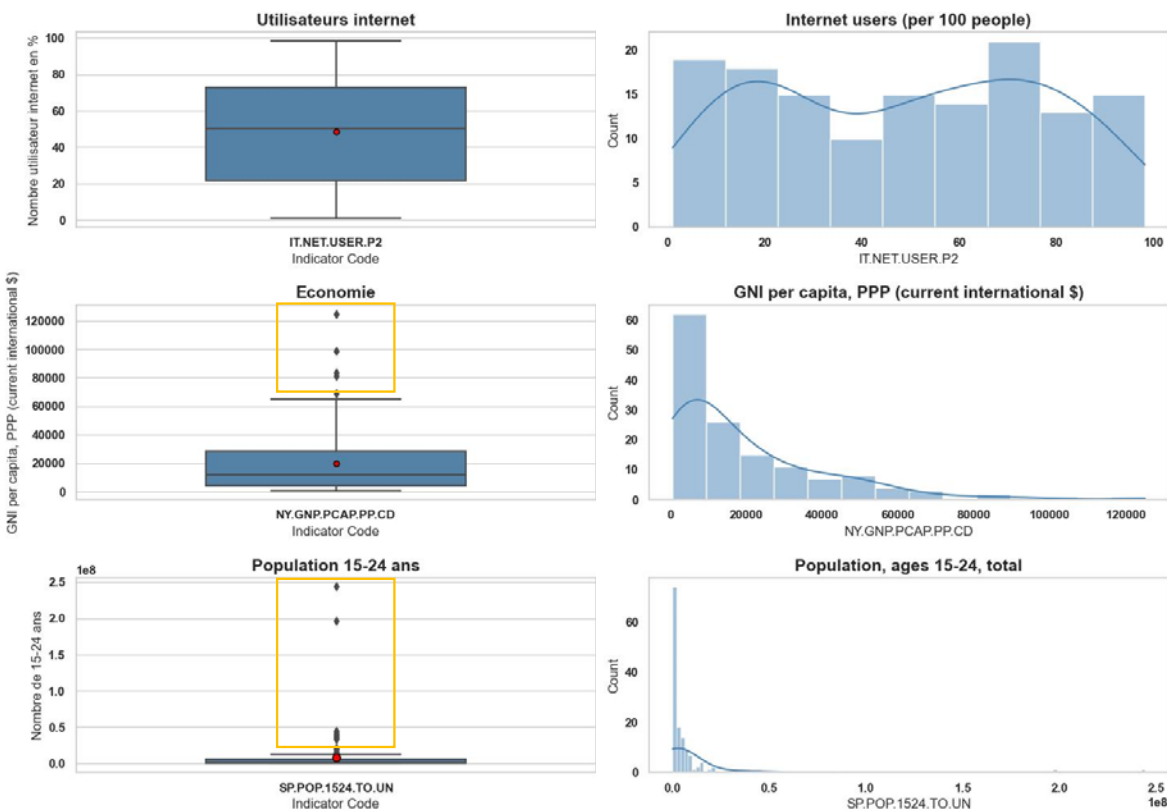
Après nettoyage on a :

700 lignes sans données manquantes

- 140 pays
- 5 indicateurs



Analyse statistiques des indicateurs retenus et gestion des outliers



- Fortes disparités dans les données en raison d'unités différentes, de pays ayant des puissances économiques et démographiques très différentes
- ➔ Traitement des données
 - ➔ On normalise les données pour pouvoir les comparer



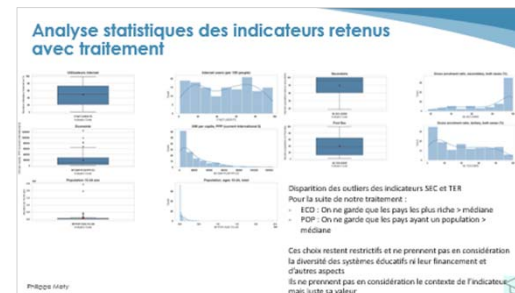
2 Traitements différents des données pour comparer les résultats

Traitements des données

Aucun traitement

Prétraitements : choix stratégiques

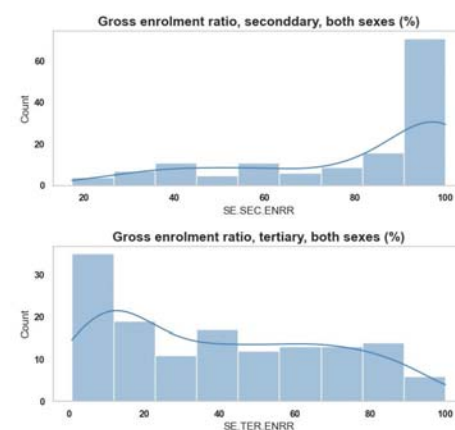
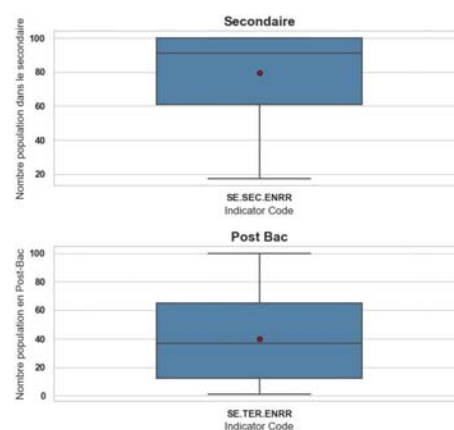
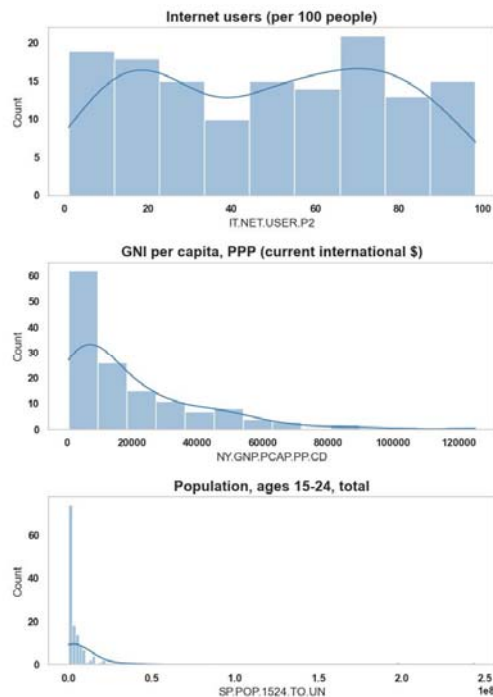
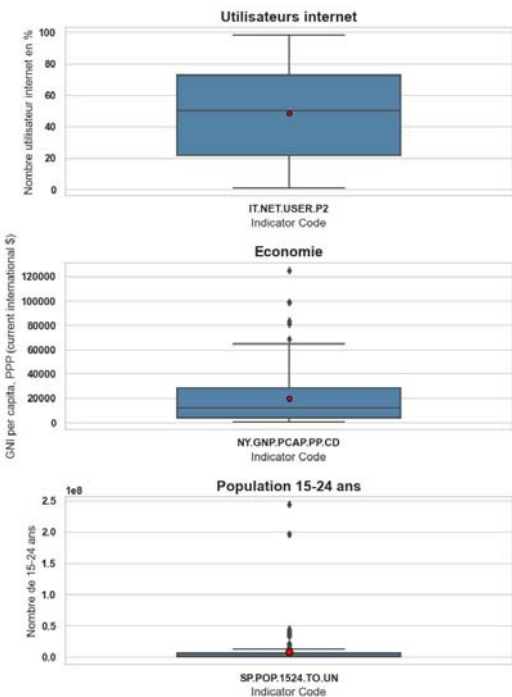
- **ECONOMIE** : suppression des pays gagnant plus que le revenu médian
- **POPULATION** : suppression des pays ayant une population < à la population médiane
- Indicateurs en pourcentage qui dépassent 100% à 100%
 - **Secondaire**
 - **Post-Bac**



- On cherche à réduire l'impact des différences entre les valeurs
- On cherche une distribution gaussienne



Analyse statistiques des indicateurs retenus avec traitement



Disparition des outliers des indicateurs SEC et TER

Pour la suite de notre traitement :

- ECO : On ne garde que les pays les plus riches > médiane
- POP : On ne garde que les pays ayant une population > médiane

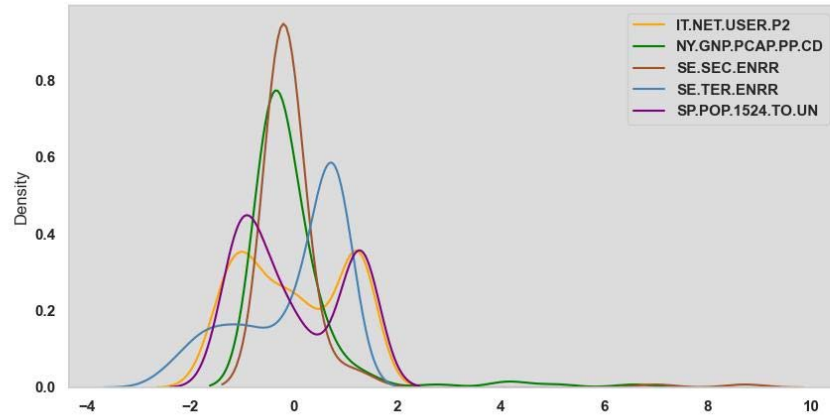
Ces choix restent restrictifs et ne prennent pas en considération la diversité des systèmes éducatifs ni leur financement et d'autres aspects

Ils ne prennent pas en considération le contexte de l'indicateur mais juste sa valeur

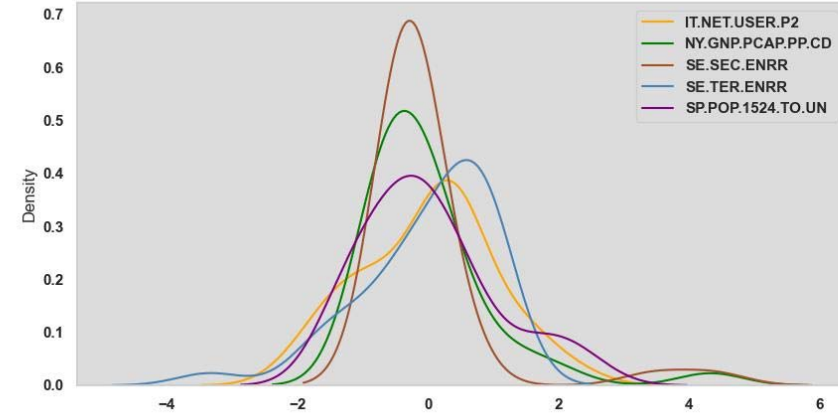


Normalisation : Choix du Scaler

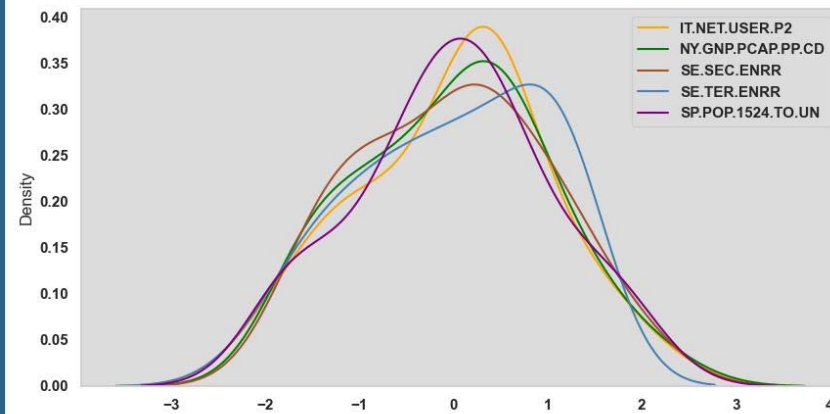
StandardScaler sans prétraitement



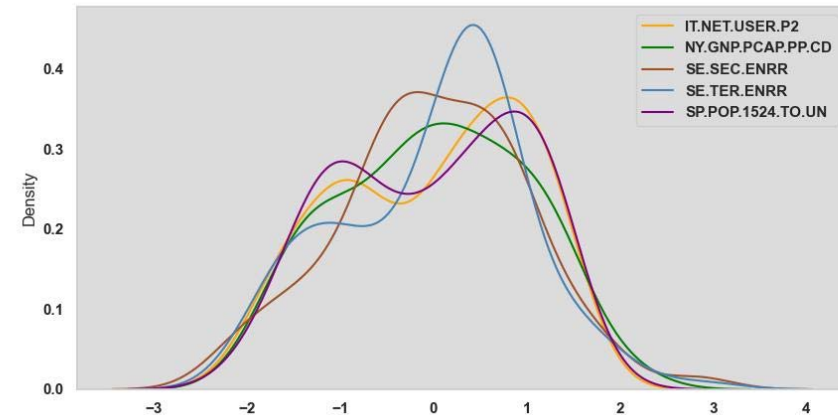
StandardScaler avec prétraitement



PowerTransformer sans prétraitement



PowerTransformer avec prétraitement

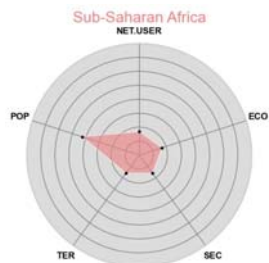
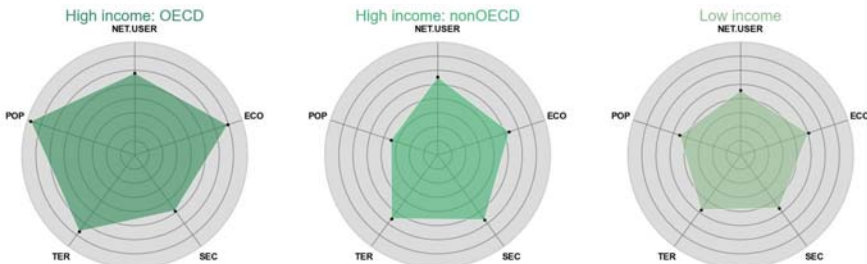


Les groupes de pays les plus favorables

Puissance des groupes par Région géographique



Puissance des groupes par Région économiques



CONCLUSION

- Quels 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 ?



SCORING & MAP

Les pays avec un fort potentiel de clients

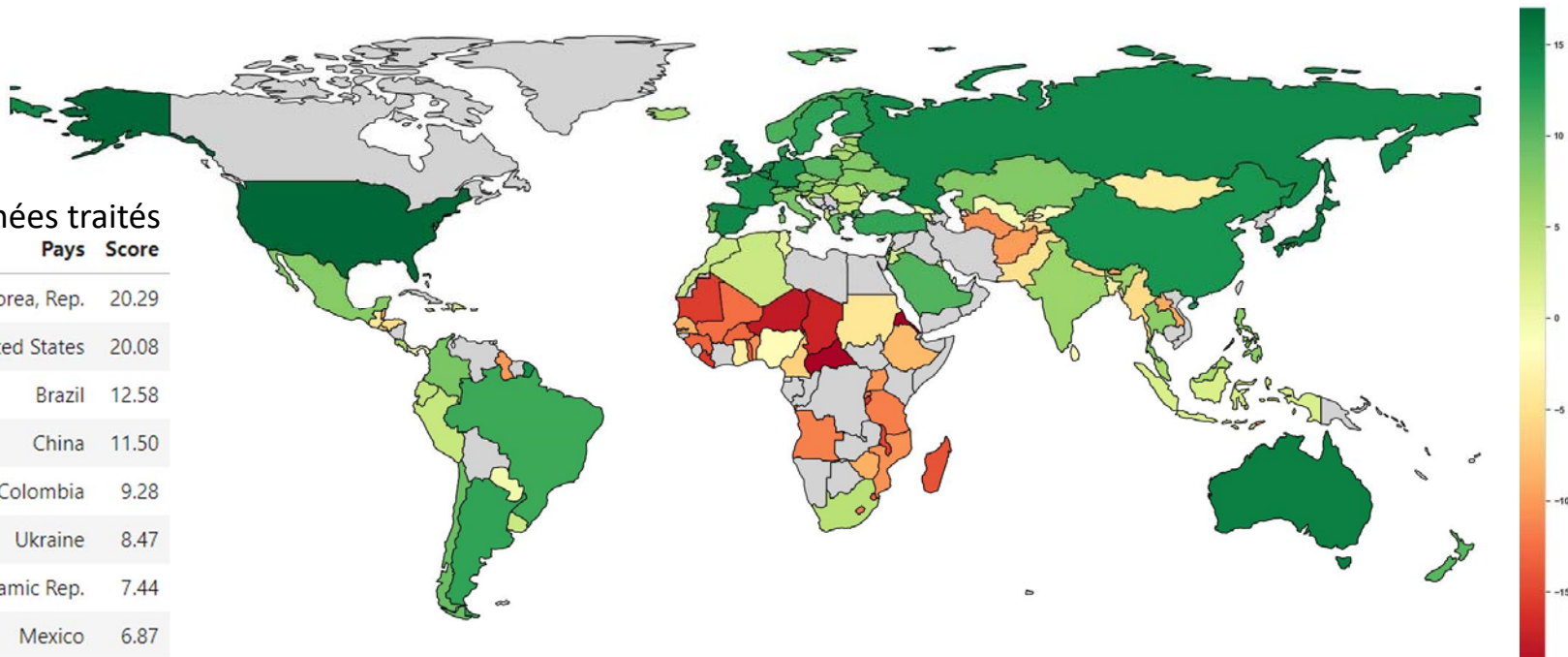
Scoring et classement provisoire

Indicator	code	Pondérations
IT	IT.NET.USER.P2	5
POP 15-24	SP.POP.1524.TO.UN	4
SUPERIEUR	SE.TER.ENRR	3
SECONDAIRE	SE.SEC.ENRR	2
ECONOMIE	NY.GNP.PCAP.PP.CD	1

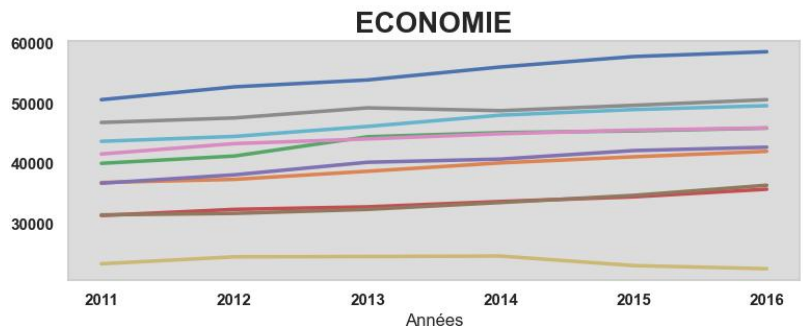
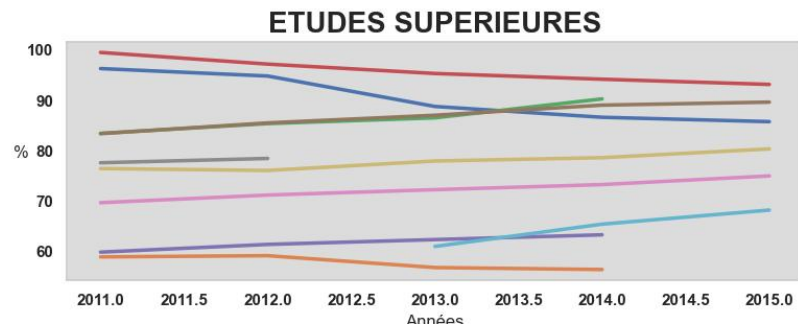
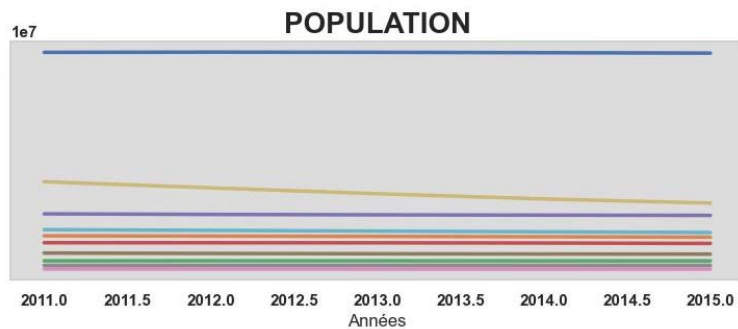
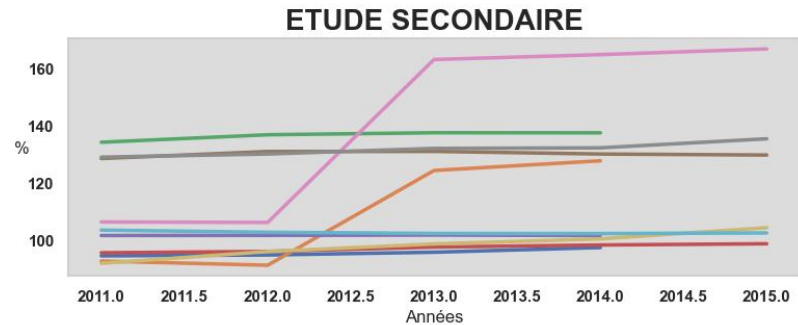
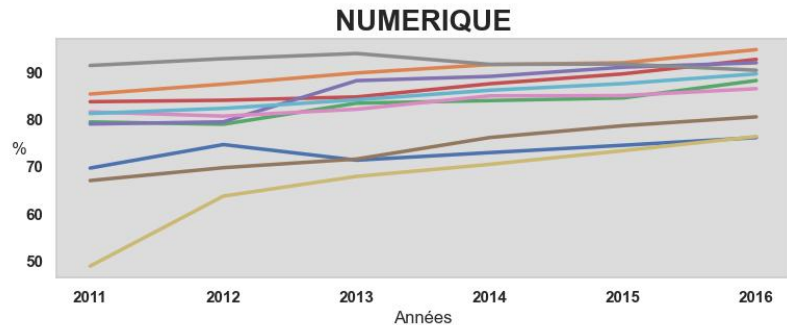
Pays	Score
United States	17.05
United Kingdom	15.50
Australia	15.22
Korea, Rep.	15.11
Japan	15.08
Spain	14.66
Belgium	14.60
Netherlands	14.49
Russian Federation	14.35
Germany	14.25

Données traitées

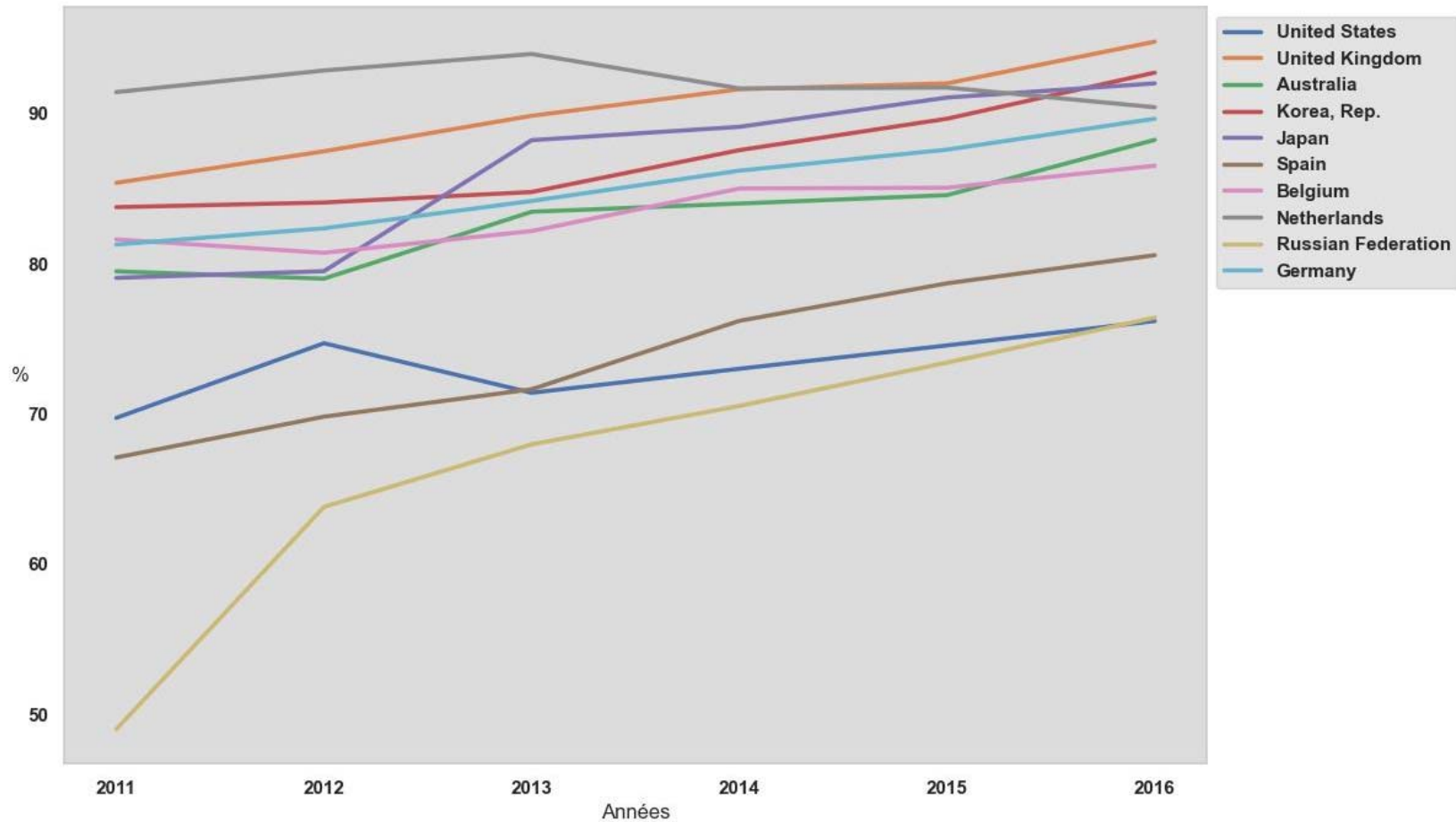
Pays	Score
Korea, Rep.	20.29
United States	20.08
Brazil	12.58
China	11.50
Colombia	9.28
Ukraine	8.47
Iran, Islamic Rep.	7.44
Mexico	6.87
Philippines	6.20
Slovak Republic	4.96



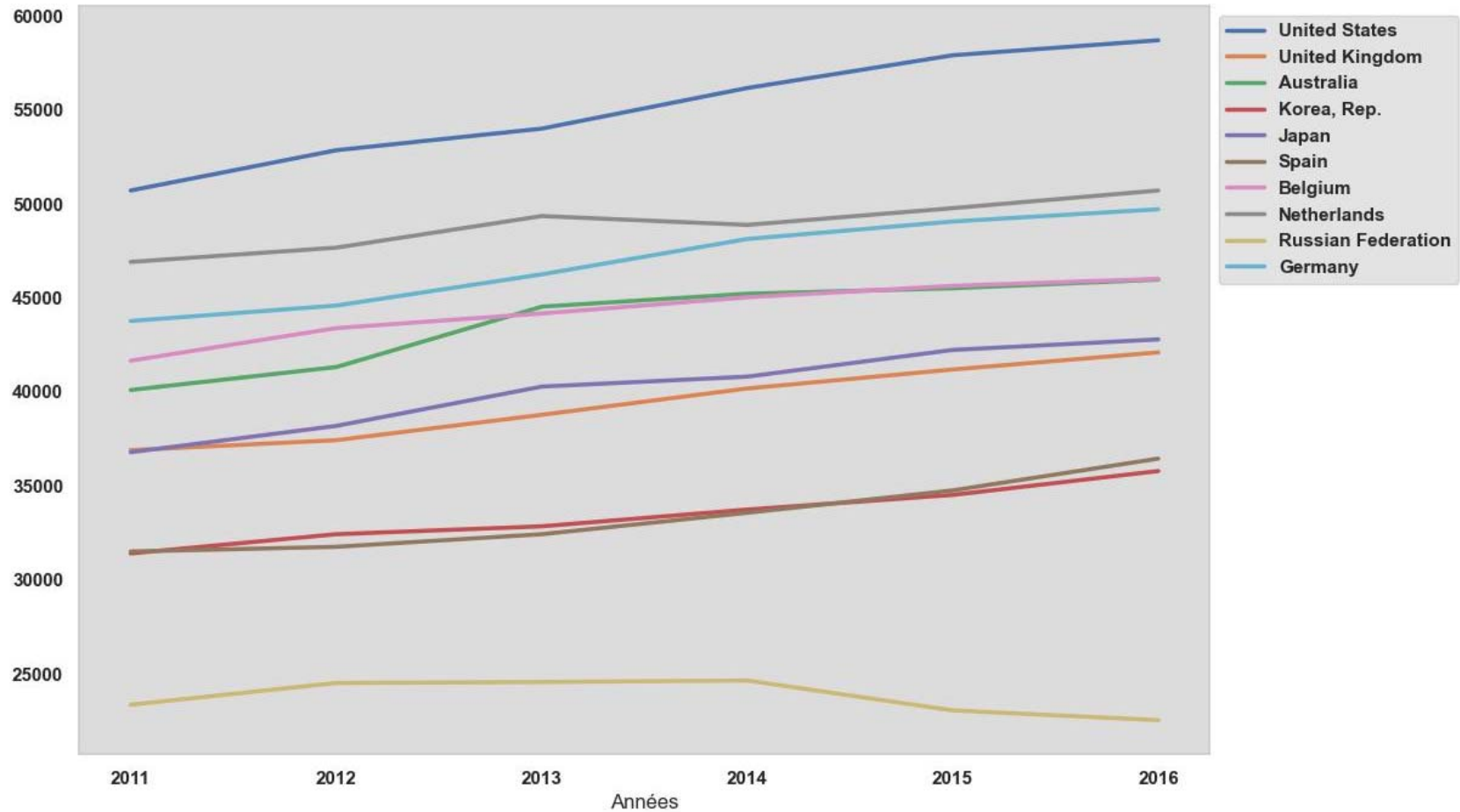
Evolution de ce potentiel de clients



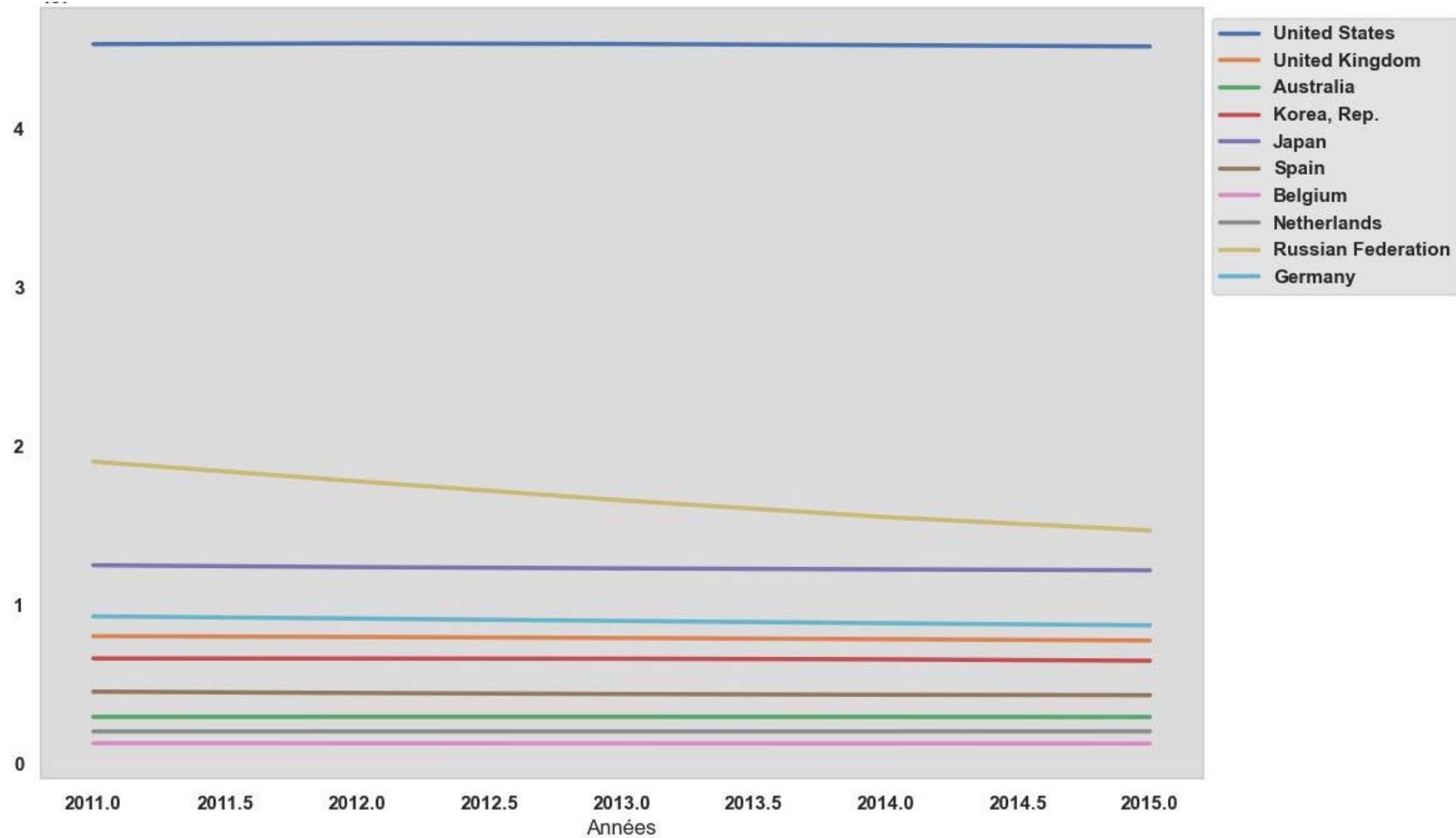
NUMÉRIQUE



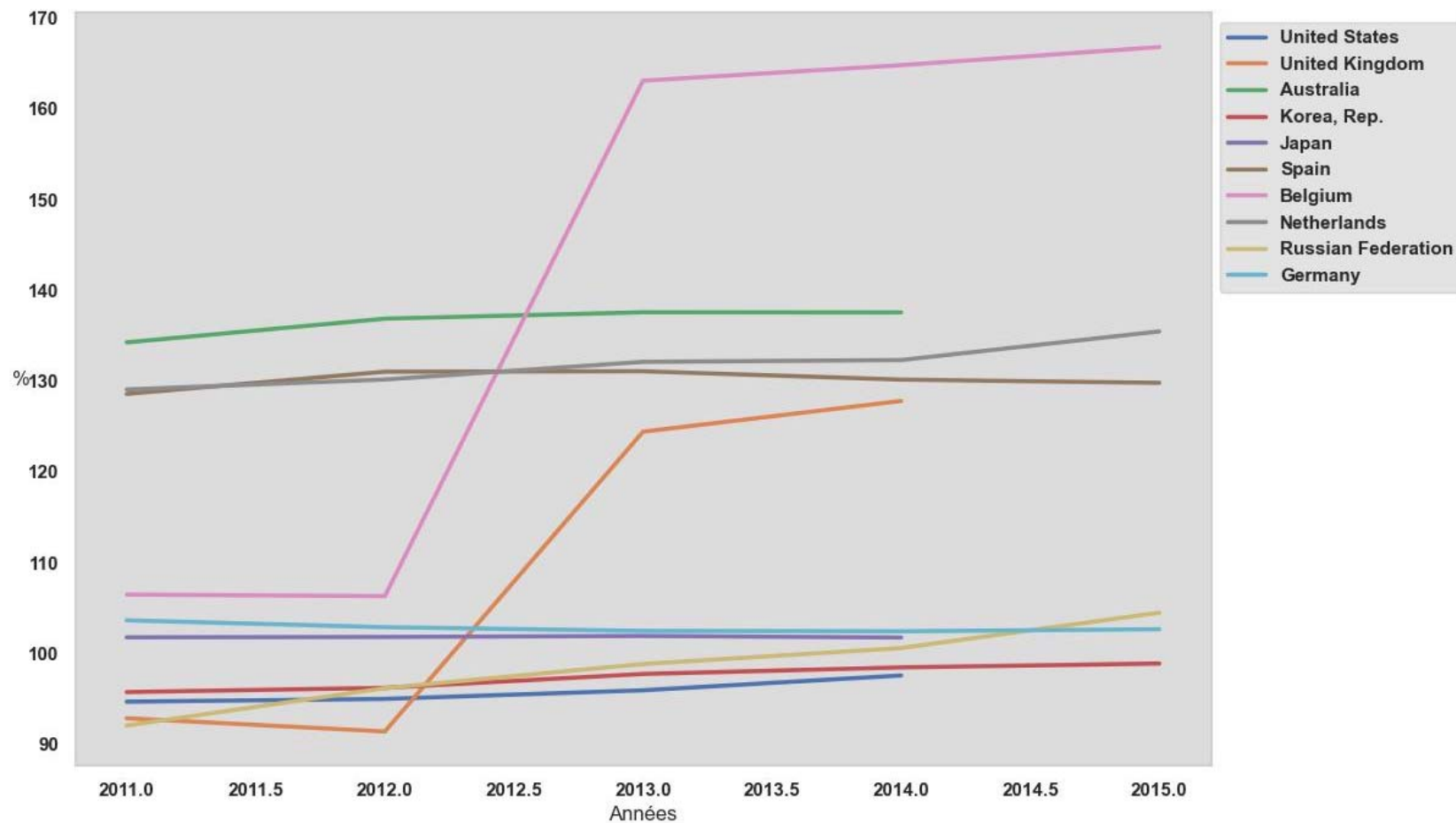
ÉCONOMIQUE



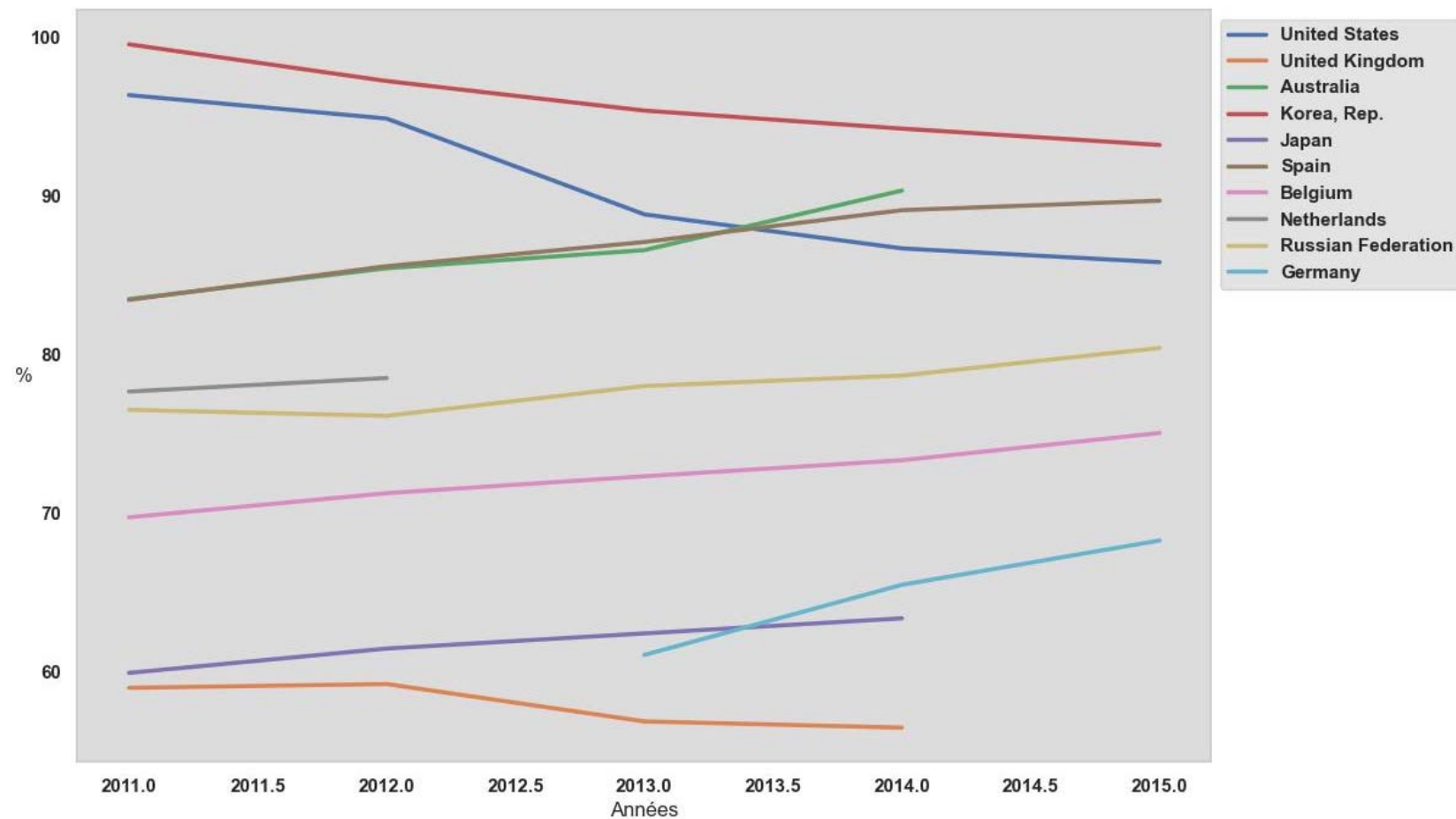
POPULATION DE 15-24 ANS



SECONDAIRE



ÉTUDES SUPÉRIEURES



Classement définitif

Pays où l'entreprise devrait opérer en priorité

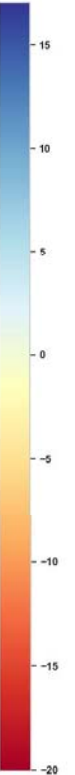
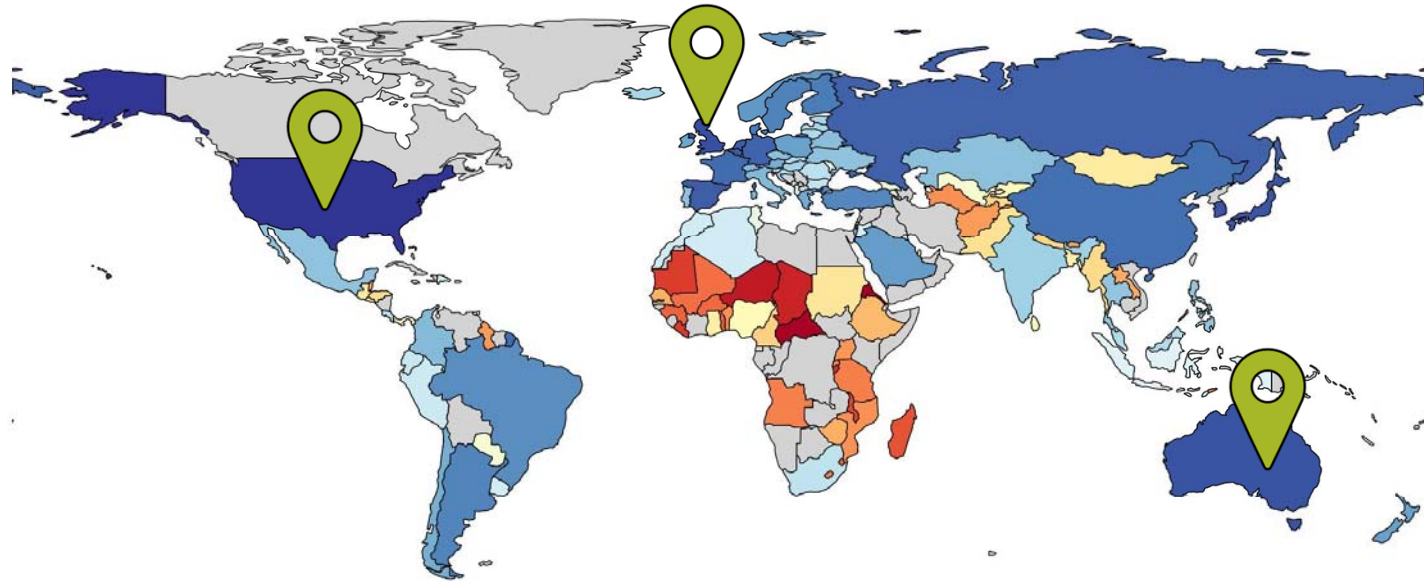


Pays	Score
------	-------

United States	17.05
---------------	-------

United Kingdom	15.50
----------------	-------

Australia	15.22
-----------	-------



MERCI