Projet 2 - Openclassrooms Analysez des données de systèmes éducatifs

Le ___.12.2022

Dabidin Keshika

OPENCLASSROOMS

academy

Plan

1. Présentation de l’entreprise
2. Présentation des données
3. Stratégie employée pour l’analyse exploratoire des données
4. Présentation des étapes de l’analyse exploratoire des données
5. Résultats
6. Conclusion

Introduction

- **Présentation Academy :** Start'up de la **EdTech** , niveau **lycée** et **universitaire**, formation **en ligne.**

- **Objectif :** Expansion à **l'international**

Analyse exploratoire des données et proposition des pays.

- **Questions à explorer :**
  - 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 ?

Présentation des données

- **Source : Banque Mondiale:**

  https://datacatalog.worldbank.org/dataset/education-statistics

- **5 fichiers :**

  Fichiers utiles :

1. *EdStatsCountry.csv : Informations sur la population et la situation économique des différents pays.*

2. *EdStatsData.csv : Valeurs et l'évolution de nombreux indicateurs pour tous les pays ou certains groupes de pays.*

3. *EdStatsSeries.csv : Informations sur les indicateurs contenus dans EdStatsData*

Présentation des données

- **Source : Banque Mondiale:**

  https://datacatalog.worldbank.org/dataset/education-statistics

- **5 fichiers :**

  Fichiers non-utilisés :

4. EdStatsFootNote.csv : Informations sur l'année d'origine des données ainsi que leurs incertitudes.

5. EdStatsCountrySeries.csv : Informations sur la source des données.

Présentation des données

	Nombre lignes	Nombre colonnes	Taux de remplissage moyen (%)	Doublons
EdStatsCountry.csv	241	32	69,5	0
EdStatsSeries.csv	3665	21	28,3	0
EdStatsData.csv	886930	70	13,9	0

- Il n'y a pas de doublons dans le jeu de données.
- Le taux de remplissage moyen pour chaque fichier n'est pas à 100%. Il y a des données manquantes.
- Il faudra cibler où sont les données manquantes.

academy

Analyse Exploratoire des Données

Étape 1	Étape 2	Étape 3	Étape 4	Étape 5
Identification des **indicateurs utiles** pour sélectionner les données.	**Nettoyage des données** : Conserver les données suffisamment remplies.	**Filtrage des données** en fonction des indicateurs retenus.	**Sélection des pays** en établissant à partir de seuils prédéfinis et en triant à chaque étape les pays à retenir.	**Évolution potentielle des pays clients sélectionnés** : Prédire l'évolution d'un critère pour les pays sélectionnés.

Étape 1 : Indicateurs Utiles

Quelles sont les variables permettant de quantifier un pays en vue d'un développement commercial ?

1. Connexion internet.
2. Population étudiante : âgée entre 15 à 24 ans .
3. Taux d'inscription des élèves en secondaire et en tertiaire.
4. Moyens financiers suffisants- situation économique comparable à celle de la France.

<u>L'évolution potentielle des clients :</u>

1. Croissance de la population démographie étudiante.

Étape 1 : Indicateurs Utiles

Liste des indicateurs sélectionnés :

IT.CMP.PCMP.P2 : pourcentage de personnes ayant un accès à un ordinateur personnel

IT.NET.USER.P2 : pourcentage d'utilisateurs d'internet

SP.POP.TOTL : population totale à la mi-année

SP.POP.1524.TO.UN : population totale de la tranche 15-24 ans

SP.SEC.TOTL.IN et **SP.SEC.UTOT.IN** : population totale ayant l'âge d'entrer en éducation secondaire

SP.TER.TOTL.IN : population totale ayant l'âge d'entrer en éducation tertiaire (tertiary ed.)

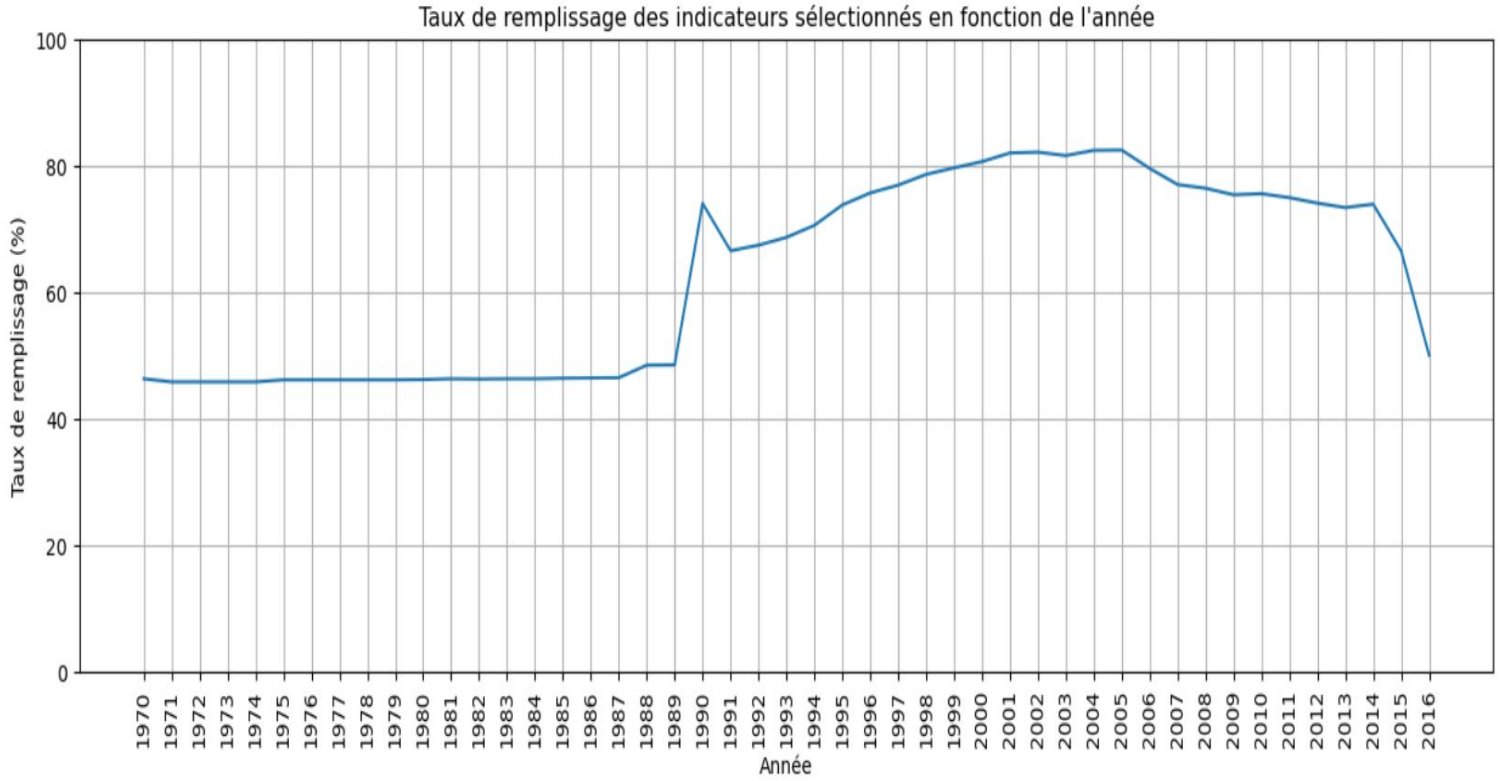
UIS.EA.3.AG25T99 : pourcentage de la population à 25+ ayant complété l'éducation secondaire

NY.GDP.PCAP.PP.CD : PIB par habitant (PPP - Current International Dollar)

SP.POP.GROW : taux de croissance annuelle de la population en %

Étape 2 : Nettoyage des données

1. Suppression des données vides.
2. Taux de remplissage des indicateurs en fonction des années :

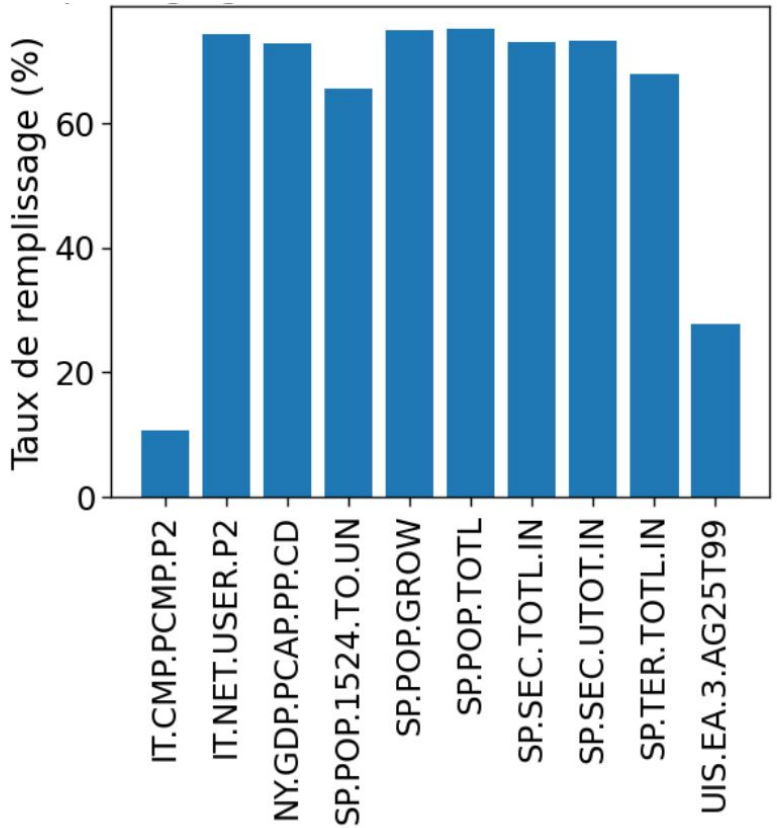


1. Les données **après 2016** ont été supprimées car les valeurs des **indicateurs sont vides**.

2. Les données **avant 1990** sont environ à **40% de remplissage**. Nous retiendrons donc uniquement les données récentes à partir de l'an **2010 à 2015** avec un taux de remplissage **supérieur à 60%**.

Étape 2 : Nettoyage des données

3. Filtrage des données en fonction des années.
4. Taux de remplissage des données en fonction des indicateurs sélectionnés :



- **IT.NET.USER.P2, NY.GDP.PCAP.PP.CD, SP.POP.TOTL, SP.POP.1524.TO.UN, SP.POP.GROW, SP.SEC.UTOT.IN, SP.SEC.TOTL.IN, et SP.TER.TOTL.IN** ont un taux de remplissage suffisant et peuvent être retenus.

- **SP.POP.1524.TO.UN** sera utilisée comme filtre des pays alors que **SP.SEC.UTOT.IN, SP.SEC.TOTL.IN et SP.TER.TOTL.IN** pourront être utilisés par la suite pour affiner la sélection.

- Les indicateurs **IT.CMP.PCMP.P2** et **UIS.EA.3.AG25T99** ne sont remplis qu'à 10,8% et 27,8% respectivement. Ces critères ne semblent pas pertinent à retenir pour la suite.

Étape 3 : Filtrage des données

1. Sélection des **pays avec plus d'un million d'habitants.**

2. Remaniement des données pour créer un tableau avec la **valeur moyenne de chaque indicateur :**

Indicator Code	IT.NET.USER.P2	NY.GDP.PCAP.PP.CD	SP.POP.1524.TO.UN	SP.POP.GROW	SP.POP.TOTL	SP.SEC.TOTL.IN	SP.SEC.UTOT.IN	SP.TER.TOTL.IN
Country Code								
AFG	5.935758	1788.887940	6.679801e+06	3.103683	3.123915e+07	4.358409e+06	2.035841e+06	2.819768e+06
AGO	7.316667	6239.908363	4.018651e+06	3.524201	2.557699e+07	3.248381e+06	1.526586e+06	2.168874e+06
ALB	54.868149	10623.779560	5.856280e+05	-0.268682	2.897253e+06	3.676903e+05	1.678492e+05	2.751482e+05
ARB	31.834459	15259.815662	NaN	2.108820	3.587756e+08	4.423419e+07	2.156913e+07	3.538265e+07

Les zones et les pays sont mélangés. On va essayer d'attribuer des pays pour chaque zone.

Pour cela, on choisira l'indicateur *SP.POP.1524.TO.UN* pour afficher les pays et les zones et filtrer les données par pays.

Étape 3 : Filtrage des données

Exemple de tableau obtenu :

Country Code	SP.POP.1524.TO.UN	Short Name
ARB	NaN	Arab World
EAP	NaN	East Asia & Pacific (developing only)
EAS	NaN	East Asia & Pacific (all income levels)
ECA	NaN	Europe & Central Asia (developing only)
ECS	NaN	Europe & Central Asia (all income levels)
EMU	NaN	Euro area
EUU	NaN	European Union
HIC	NaN	High income

On constate qu'à l'exception de Puerto Rico, les autres lignes ne sont pas des pays mais des **regroupements de pays**.

On effectuera le choix de supprimer Puerto Rico afin de réattribuer les pays aux zones dans l'étape suivante.

Étape 3 : Filtrage des données

Tableau obtenu après le filtrage des données :

Country Code	Utilisateurs Internet (%)	PIB par hab (PPA)	Population 15-24 ans	Croissance pop. (%)	Country Name
AFG	5.935758	1788.887940	6.679801e+06	3.103683	Afghanistan
AGO	7.316667	6239.908363	4.018651e+06	3.524201	Angola
ALB	54.868149	10623.779560	5.856280e+05	-0.268682	Albania
ARE	83.316665	63064.874408	7.128560e+05	2.956295	United Arab Emirates
ARG	57.407177	19647.406305	6.873357e+06	1.036686	Argentina
...	...	...	...	...	...
VNM	37.586667	5186.028936	1.777147e+07	1.067158	Vietnam
YEM	18.556152	3808.375398	5.688936e+06	2.638750	Yemen
ZAF	41.064853	12641.709688	9.899711e+06	1.507078	South Africa
ZMB	15.061367	3590.991069	2.891667e+06	2.989994	Zambia
ZWE	13.567926	1847.212669	3.370104e+06	2.219090	Zimbabwe

154 rows × 5 columns → Environ 80% des pays

Étape 4 : Sélection des pays

Liste de priorités :

1. Accès à internet.

2. Pays avec un PIB par habitant comparable à la France ou plus élevé que la France.

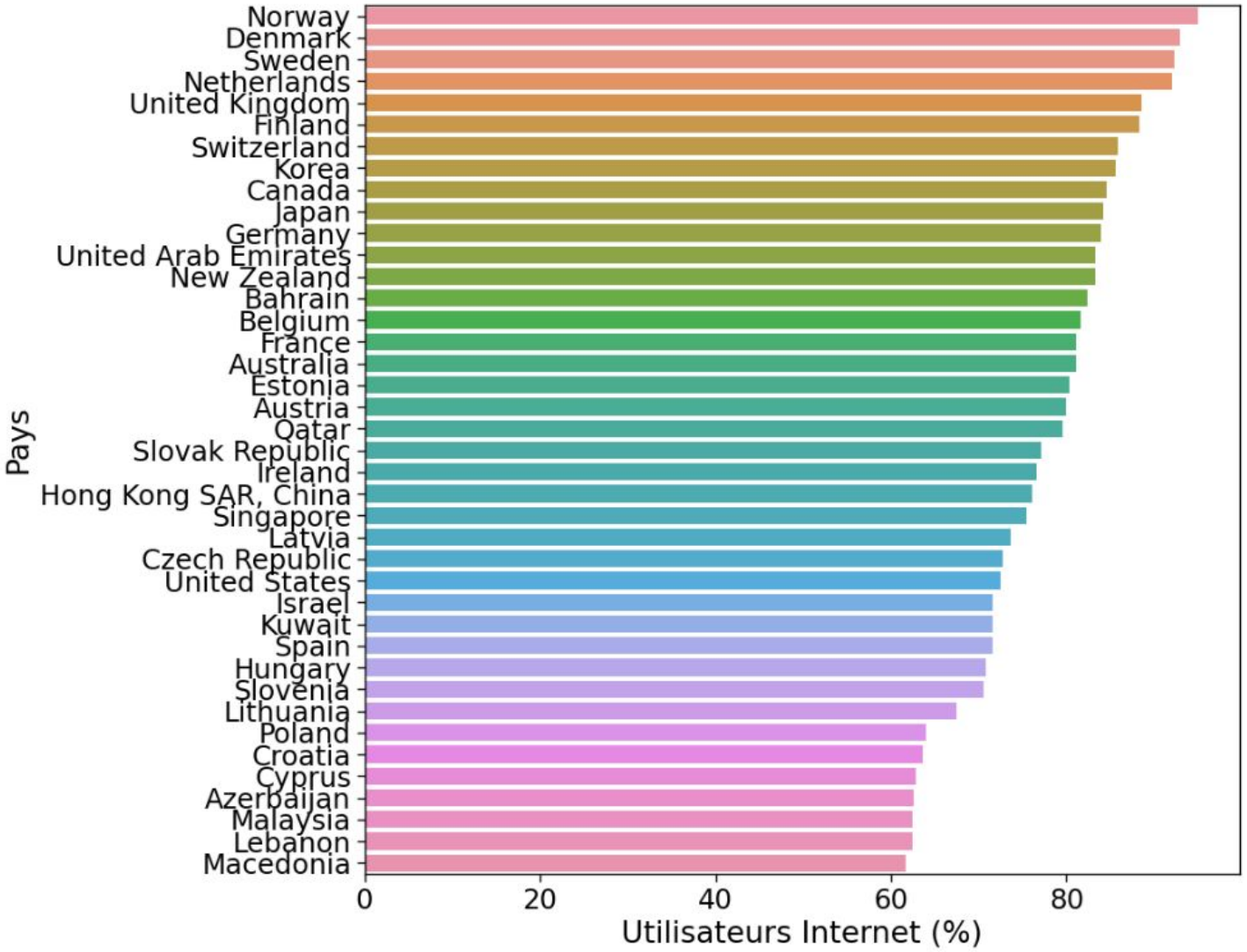
3. Pays avec le plus de population âgée entre 15 à 24 ans ou bien pays avec le plus d'élèves en secondaire et en tertiaire - 2 possibilités.

Étape 4 : Sélection des pays

1. Accès à internet.

Liste des 40 premiers pays ayant accès à internet.

- Seuil à 60%.

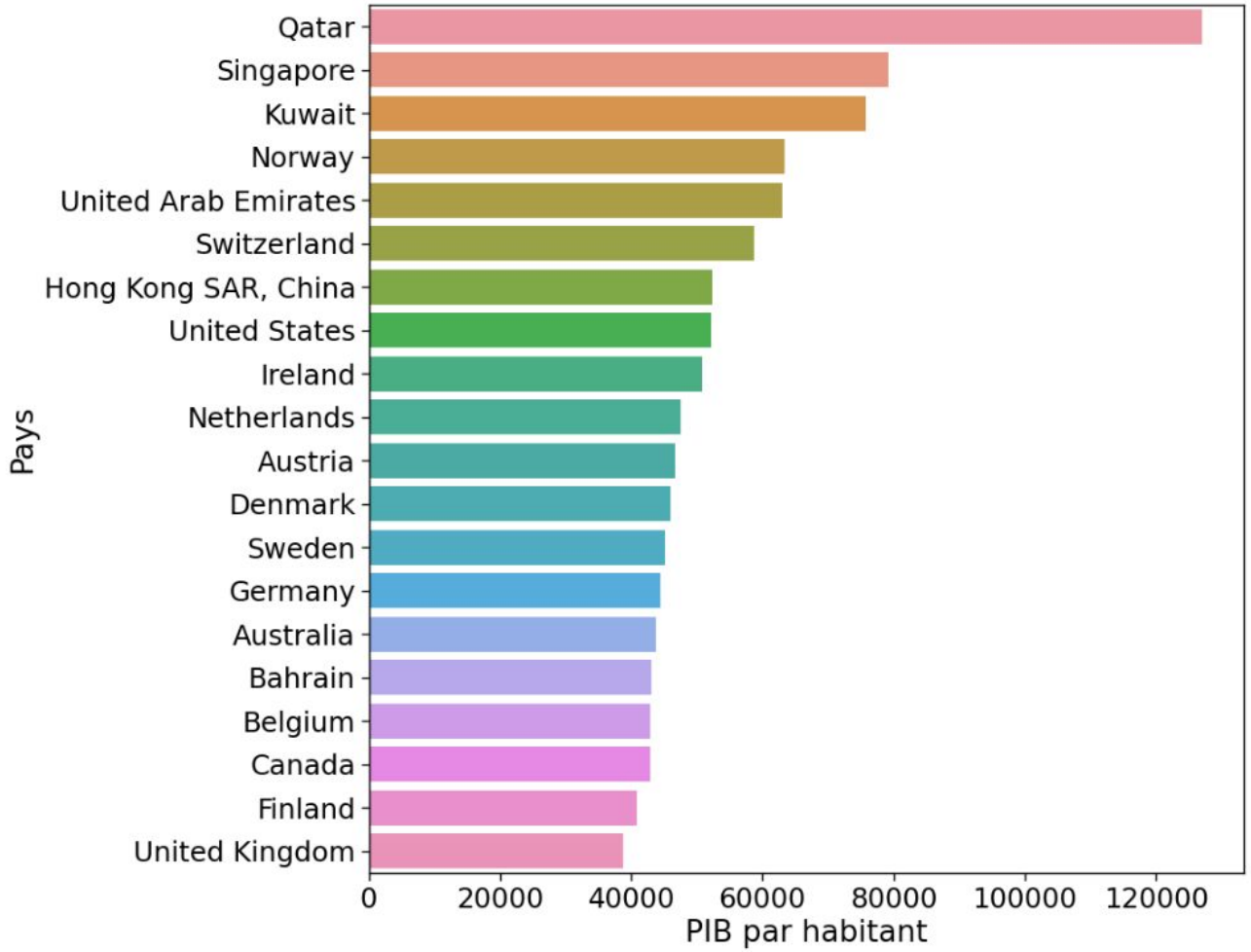


Étape 4 : Sélection des pays

2. PIB par Habitant

Parmi les 40 pays → Sélection des 20 premiers pays.

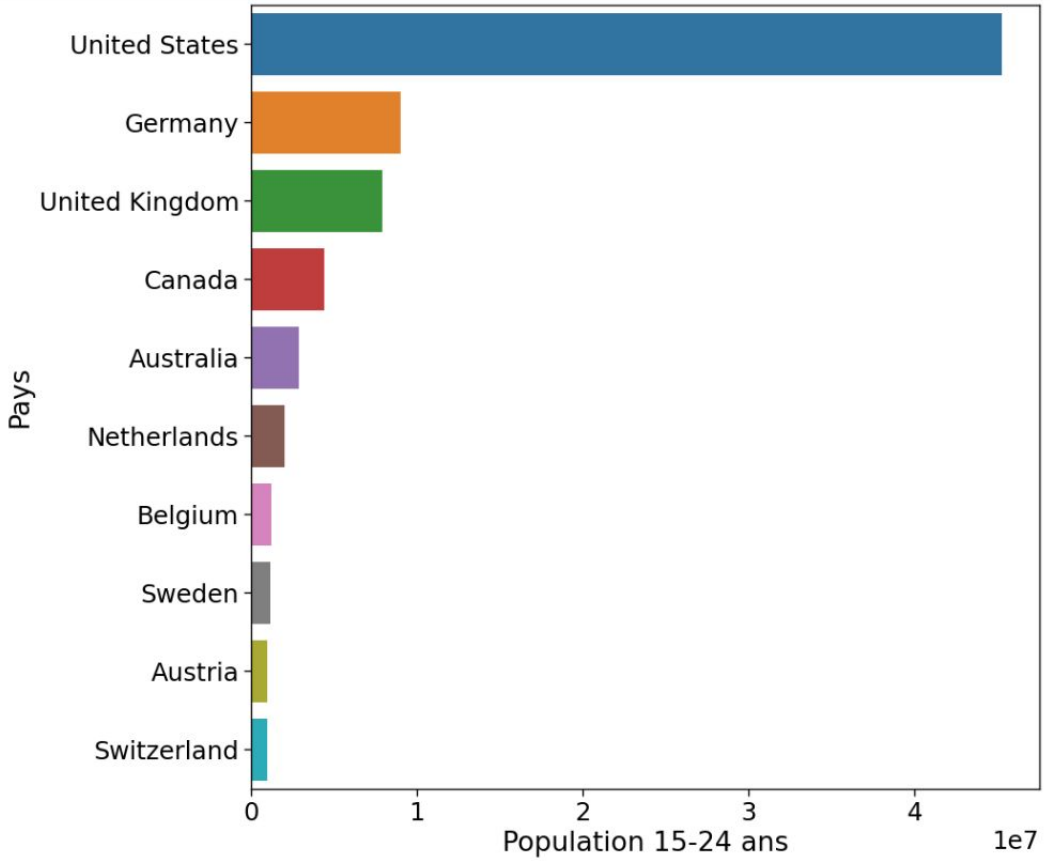
- PIB > France (Entre 20 et 25).



Étape 4 : Sélection des pays

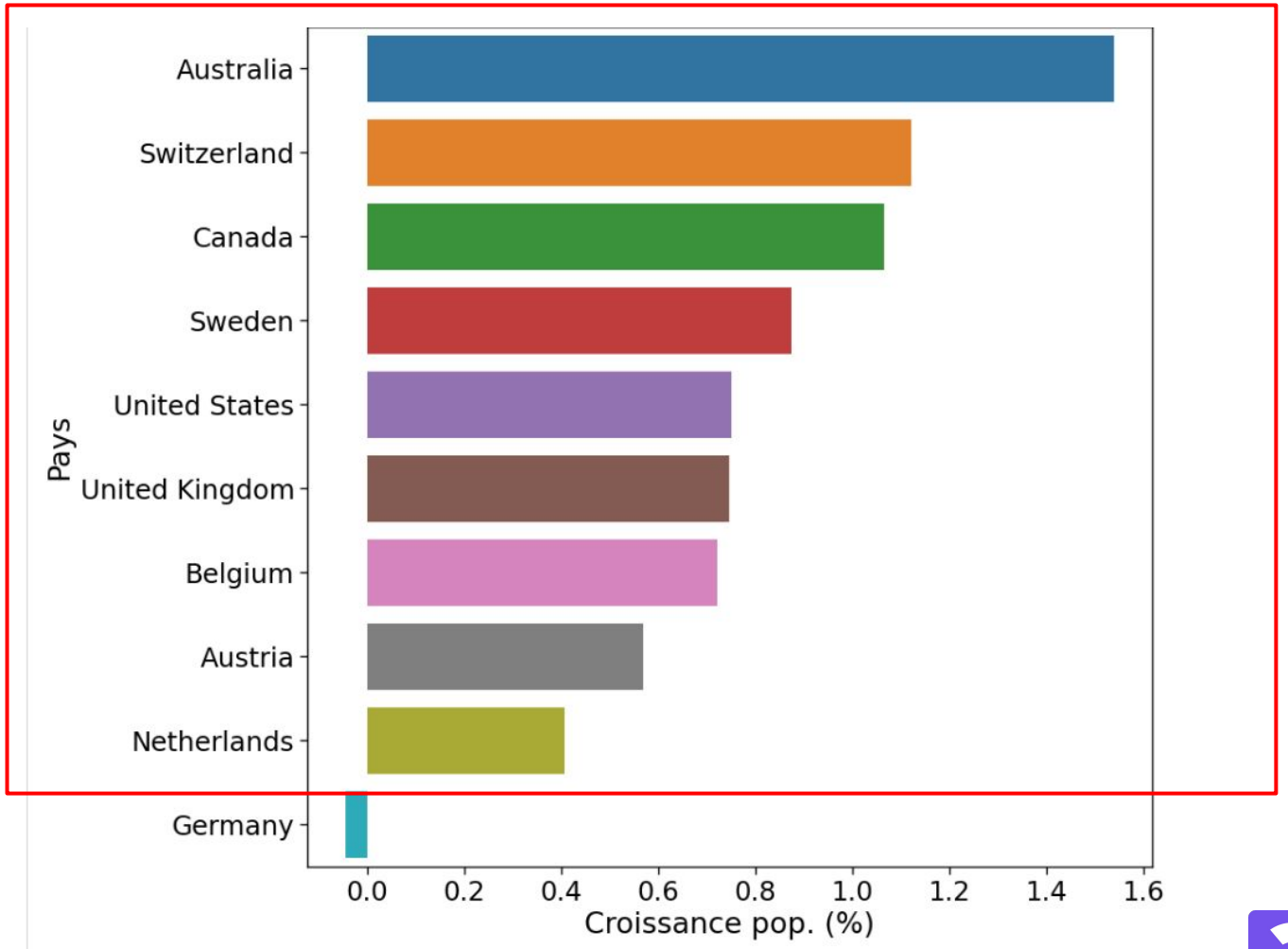
3. Population entre 15 à 24 ans : parmi les 20 pays → Sélection des 11 premiers pays.

Pays proposé	Population de potentiels étudiants (par million(s))
États Unis	45,2
Allemagne	9,0
Angleterre	7,90
Canada	4,45
Australie	2,92
Pays-Bas	2,00
Belgique	1,25
Suède	1,91
Autriche	0,99
Suisse	0,96



Étape 5 : Évolution potentielle des pays clients

Pays proposés



Conclusion

Le jeu de données permet-il de répondre aux attentes de Academy?

- Sources fiables - données utiles pour l'éducation.
- La majorité des pays sont présents.
- Analyse possible avec des résultats cohérents.

Pour aller plus loin :

- Certains indicateurs inutilisables (beaucoup de données manquantes pour comparer). Données à compléter avec d'autres sources.
- Manque certains indicateurs business : pénétration de l'éducation en ligne, % d'élèves ayant recours à l'internet pour suivre des cours en ligne, etc.
- Limites de l'entreprise Academy : moyens, langues envisagées pour enseignées et facilité de déploiement à l'étranger.

Merci pour votre attention.

OPENCLASSROOMS

academy