Analysez des données de systèmes éducatifs

Projet n°2 – Parcours Data Scientist – Jérémy VANGANSBERG

academy

Table des matières

Problématique

Contexte

academy

- Entreprise : **academy**
- Activité de l’entreprise : formation en ligne pour un public de niveau lycée et université
- Cadre : **Expansion à l’international**
- Analyse **pré-exploratoire** d’un jeu de données
- Source des données **: banque mondiale**
- Objectif : déterminer si les données permettent d’informer le projet d’expansion

academy

Objectifs de la pré-analyse

- Valider la **qualité des données**
- Décrire les informations contenues dans le jeu de données
- Sélectionner des **informations pertinentes** pour l'analyse
- Définir des **indicateurs statistiques** pour les différentes zones géographiques

academy

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** ?

academy

Caractéristiques du jeu de données

- Origine : **world bank**
- Sujets : accès à l'éducation, professeurs, population, dépenses, alphabétisation
- Date de sortie : Juillet 2010
- Dernière mise à jour : mars 2020
- Périodicité : annuel
- Plage temporelle : 1970 - 2100

academy

Caractéristiques du jeu de données (2)

- Dimensions du dataset :
  - **886930 entrées**
  - **70 colonnes**
- Les 4 premières colonnes contiennent :
  - Le nom du pays/zone (1) et son code (2)
  - Le nom de l'indicateur (3) et son code (4)
- Les colonnes suivantes correspondent aux années de **1970 à 2100**

Nettoyage et analyse pré-exploratoire

academy

Valeurs manquantes et dupliquées

- 1ere exploration :
  - Aucune valeur dupliquée
  - Un nombre important de valeurs manquantes
- La librairie *missingno* est une bonne solution pour visualiser la structure d'un jeu de données volumineux

academy

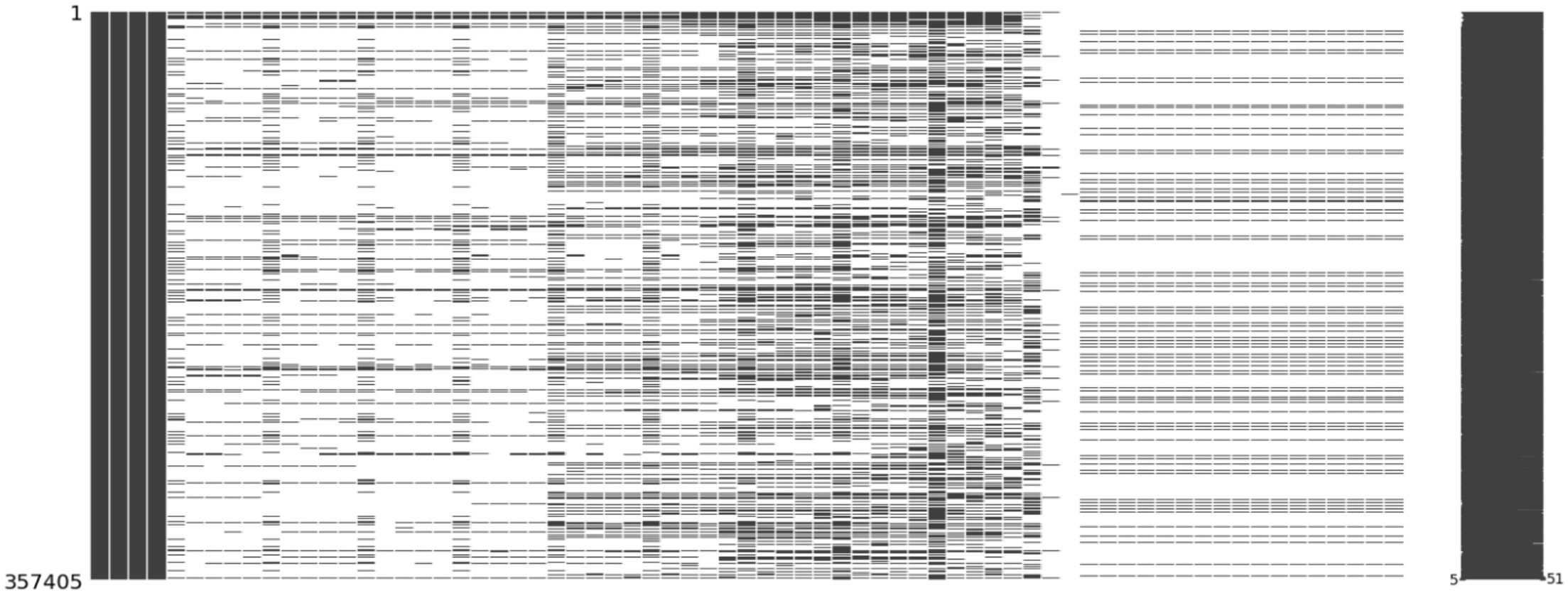
Exploration à l'aide de Missingno

Structure du dataset original

academy

Exploration à l'aide de Missingno

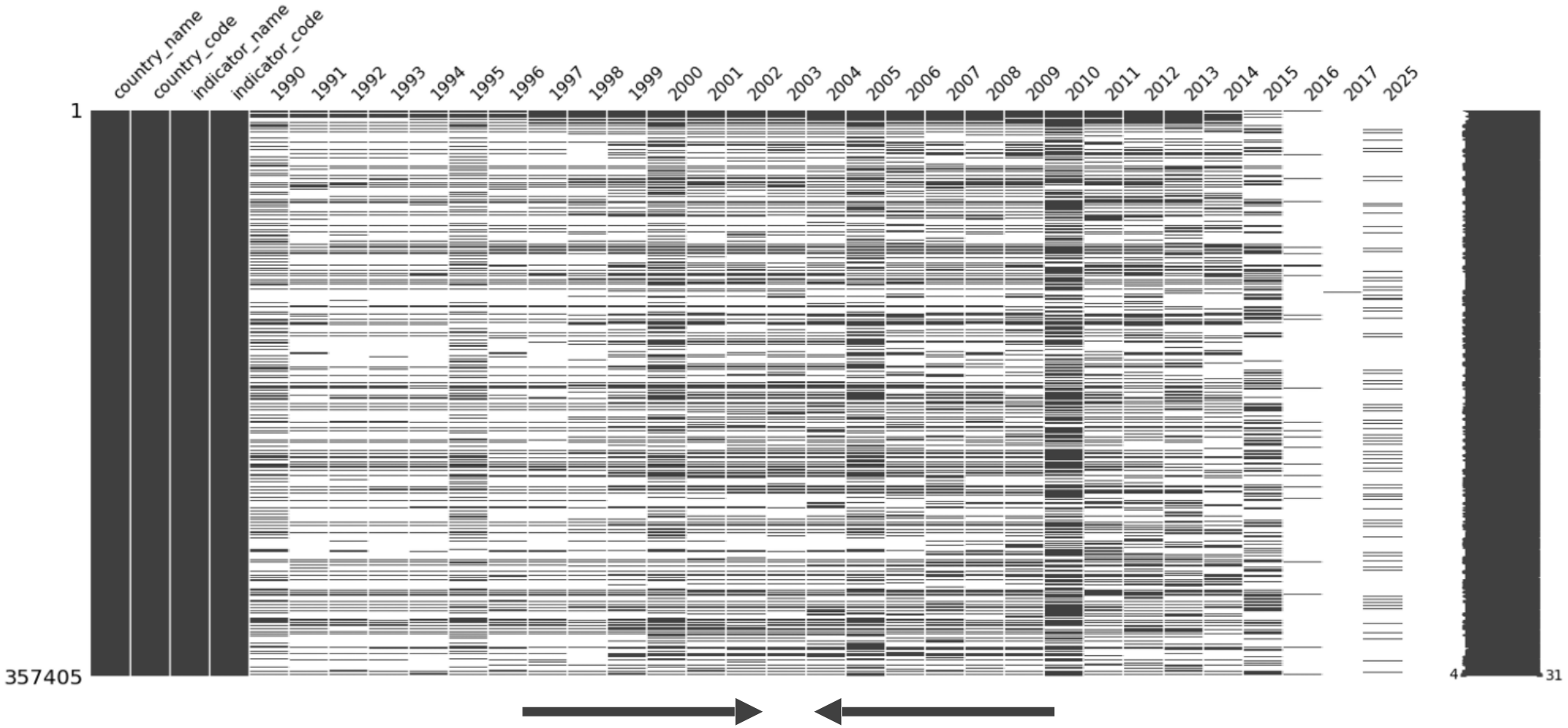
Structure du dataset après avoir retiré les entrées ne comportant que des NaN



academy

Exploration à l'aide de *missingno*

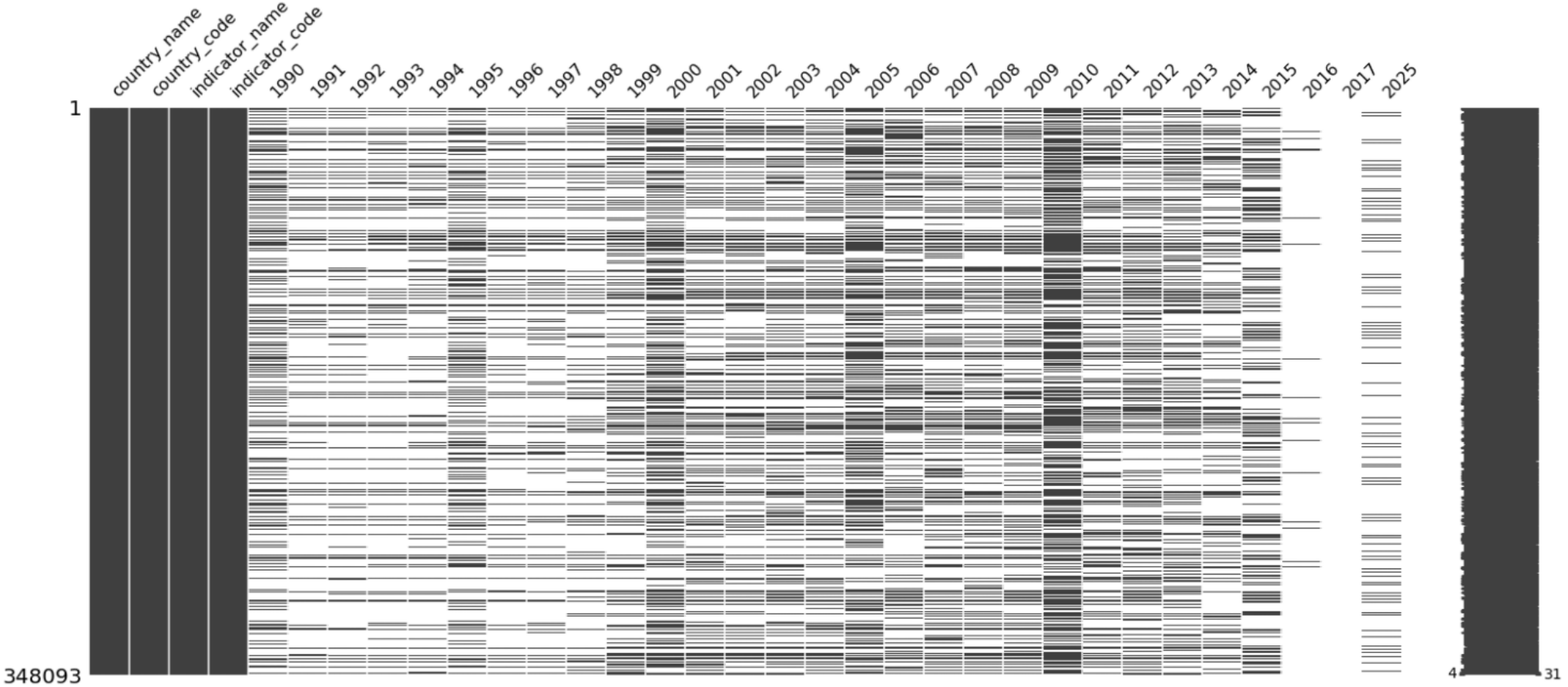
Structure du dataset après avoir restriction du nombre de colonnes



academy

Exploration à l'aide de *missingno*

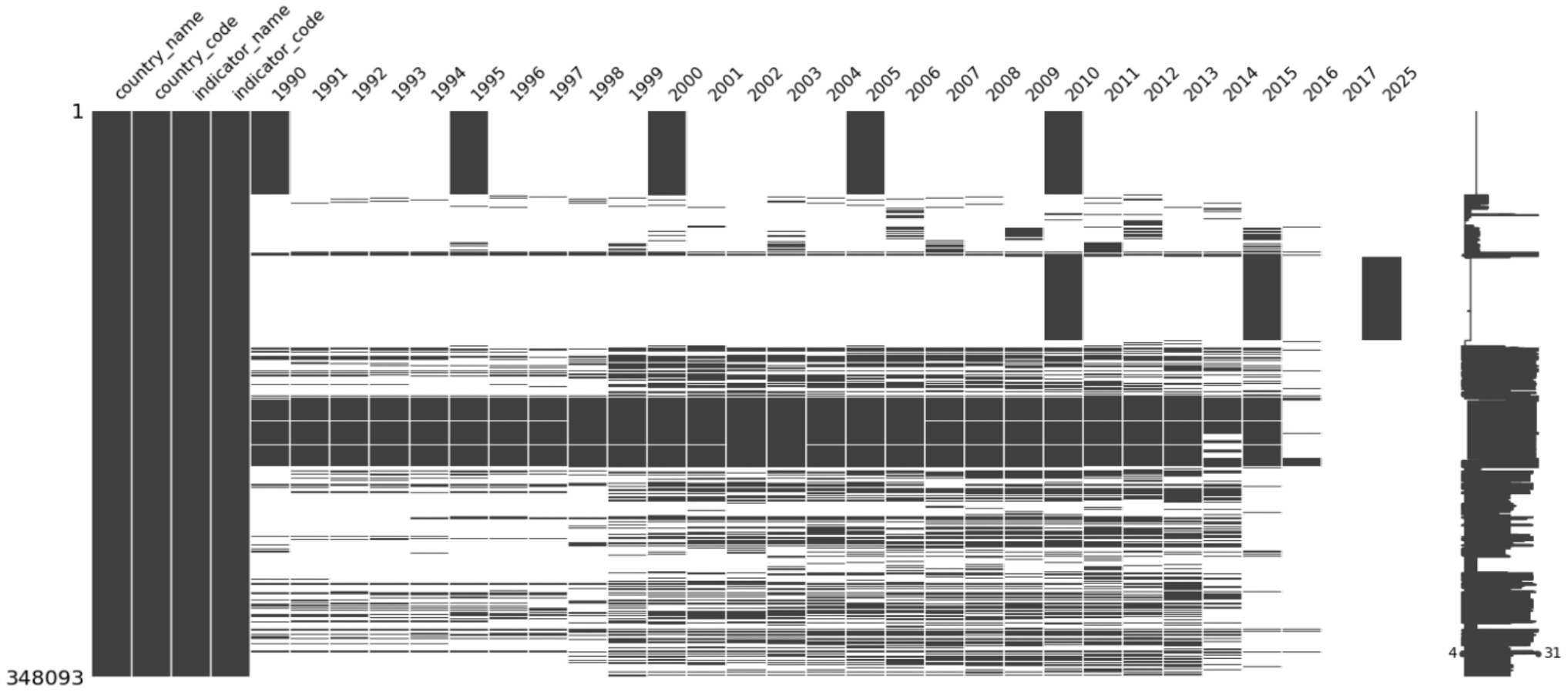
Structure du dataset après avoir retiré les entrées qui ne sont pas des pays



academy

Exploration à l'aide de *missingno*

Structure du dataset après avoir trié « indicator_name » par ordre alphabétique



academy

Les variables

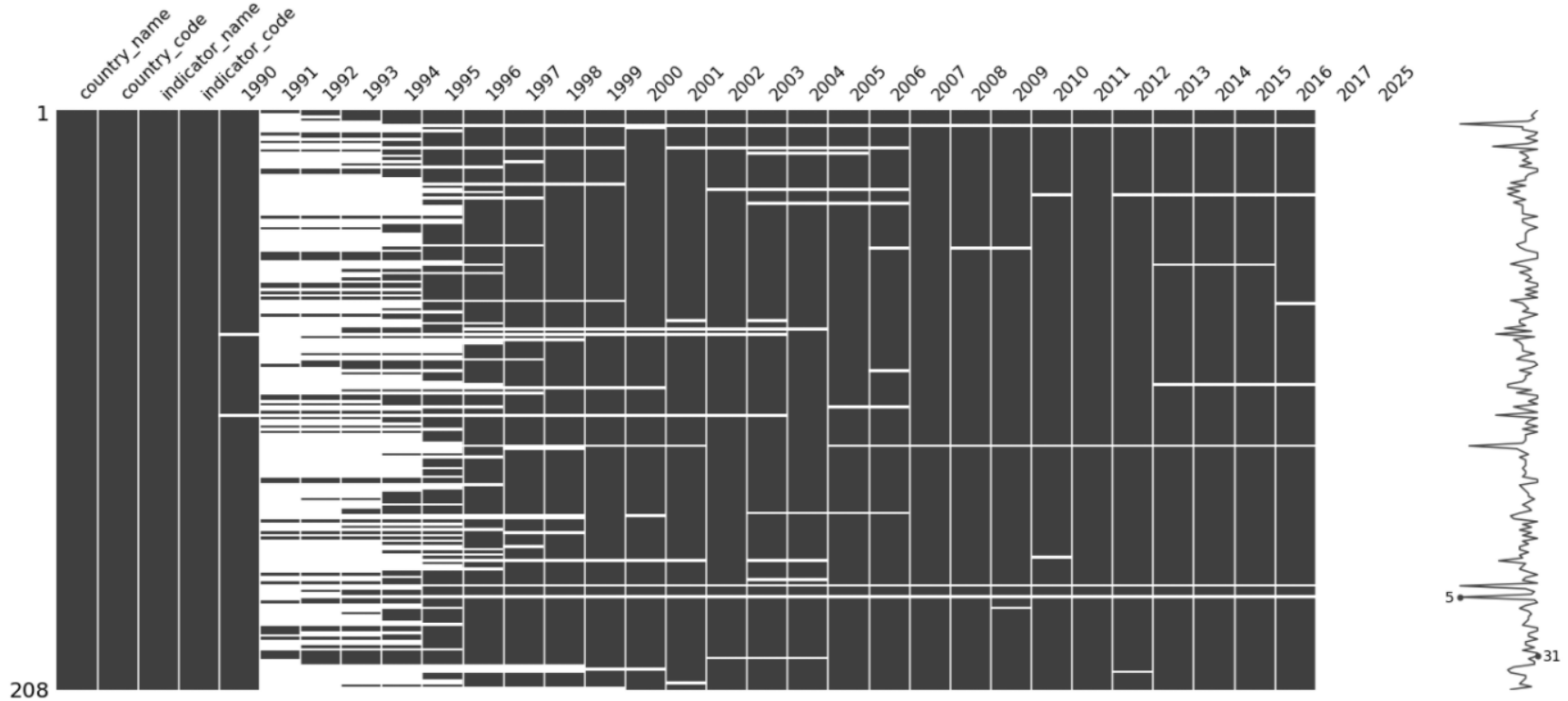
- Le jeu de données comporte 3665 variables
- Pour effectuer mon analyse, je dois en sélectionner quelques unes
- Afin de naviguer parmi cette masse d'informations, j'ai utilisé la barre de recherche du « **Query Tool** » sur le site de la word bank
- Deux catégories retenues:
  - Variable relative à Internet
  - Variables relatives à l'économie

Les variables retenues relatives à Internet

academy

- ‘Internet users (per 100 people)’

Structure des données selon les variables retenues

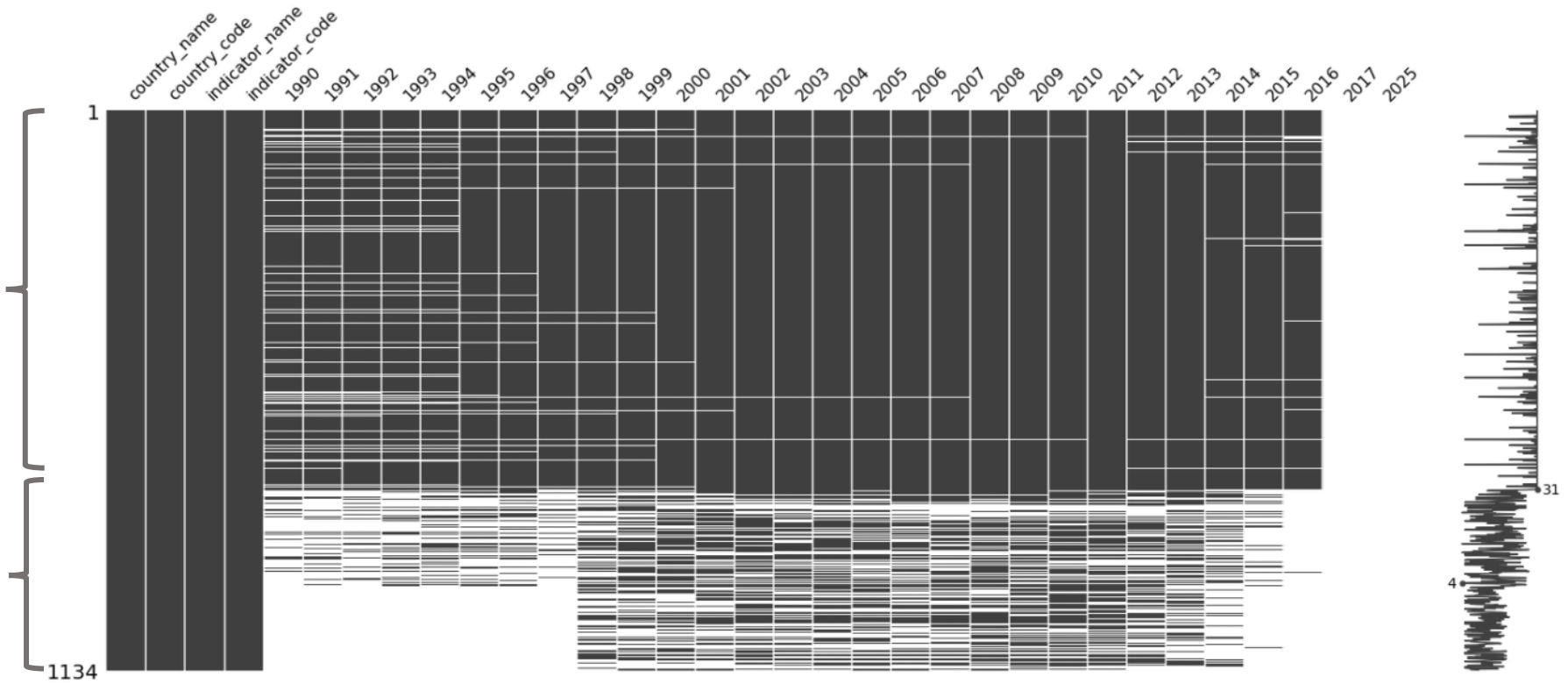


Les variables retenues relatives à l'économie

academy

Structure des données selon les variables retenues

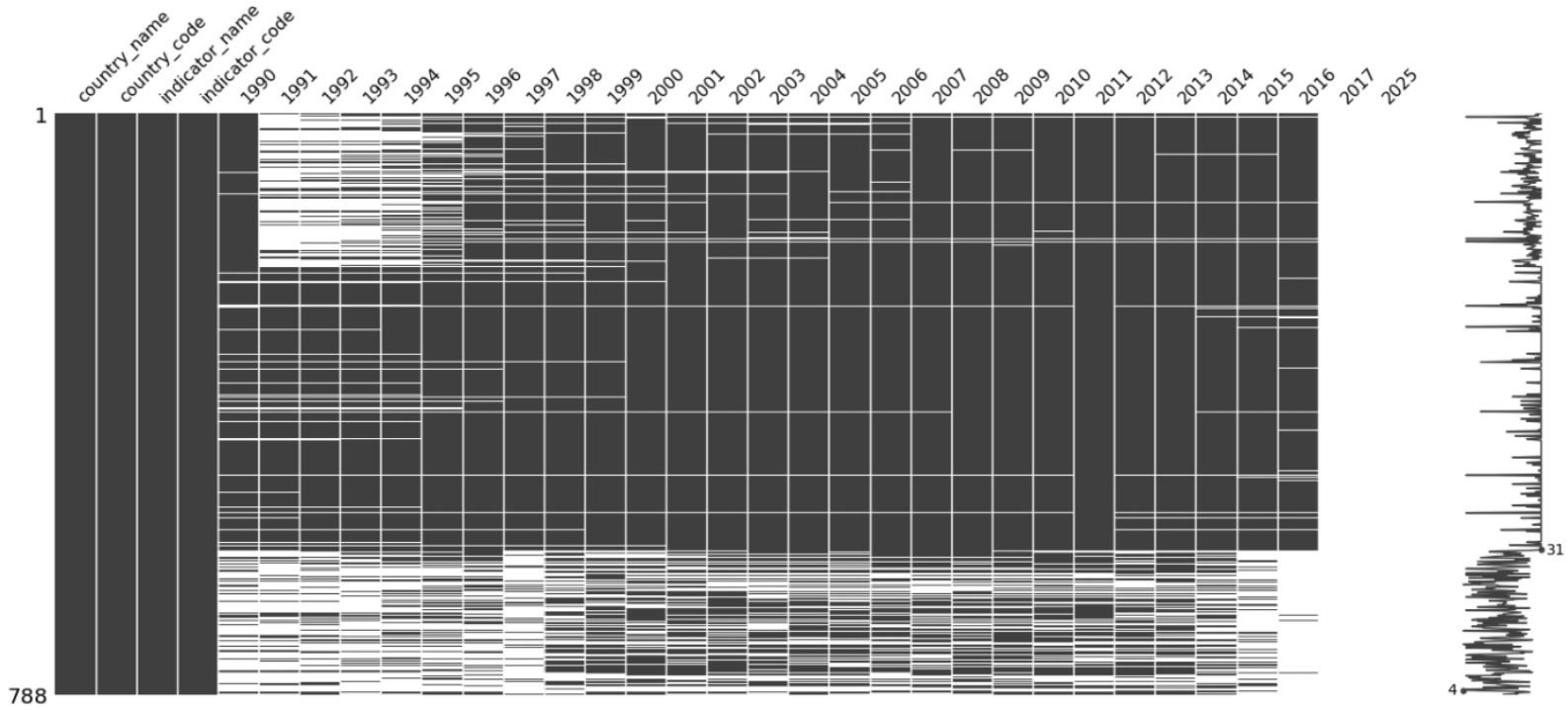
- ✓ 'GDP, PPP (current international $)'
- ✓ **'GDP, PPP (constant 2011 international $)'**
- ➢ 'GDP per capita, PPP (current international $)'
- ➢ **'GDP per capita, PPP (constant 2011 international $)'**
- ❖ 'Government expenditure on secondary education as % of GDP (%)' *(64% de NaN)*
- ❖ **'Government expenditure on education as % of GDP (%)' *(49% de NaN)***



Les variables retenues

- 'Internet users (per 100 people)‘
- 'Government expenditure on education as % of GDP (%)‘
- 'GDP, PPP (constant 2011 international $)‘
- 'GDP per capita, PPP (constant 2011 international $)'

Structure des données selon les variables retenues



academy

Taux de remplissage par pays

Top 10

country_name	tx_remp
Finland	0.909091
Austria	0.909091
Ireland	0.909091
United Kingdom	0.909091
Portugal	0.909091
Sweden	0.909091
Switzerland	0.909091
France	0.909091
Spain	0.916667
Norway	0.916667

Bottom 10

country_name	tx_remp
Turks and Caicos Islands	0.151515
American Samoa	0.151515
Northern Mariana Islands	0.151515
Curacao	0.151515
British Virgin Islands	0.242424
South Sudan	0.292929
Cayman Islands	0.303030
Aruba	0.401515
Somalia	0.454545
Libya	0.462121

- Les pays qui comportent beaucoup de NaN sont des petits pays ou des pays en voie de développement.
- Ils seront **éliminés naturellement** lorsque j'appliquerai des filtres (PIB par habitant, taille du pays, etc.)

academy

Autres modifications du dataset

- Création d'une **fonction** pour retenir la dernière **colonne non nulle**
- **Pivot** des données afin que chaque ligne soit un pays unique
- Changement des types de variables (*object* à *float* pour les nombres)
- Ajout de certaines variables

Ajout de variables *(Word Bank)*

- **Population** :
  - Totale
  - 15-19 ans
  - Urbaine
- Région du monde associée au pays:
  - 'Latin America & Caribbean‘
  - 'South Asia‘
  - 'Sub-Saharan Africa‘
  - 'Europe & Central Asia‘
  - 'Middle East & North Africa‘
  - 'East Asia & Pacific‘
  - 'North America'

Toutes ces variables proviennent de la *Word Bank*. J’ai ajouté ces informations en effectuant un *merge* sur l’index ‘country_code’

Conclusion de la partie nettoyage et pré-exploratoire

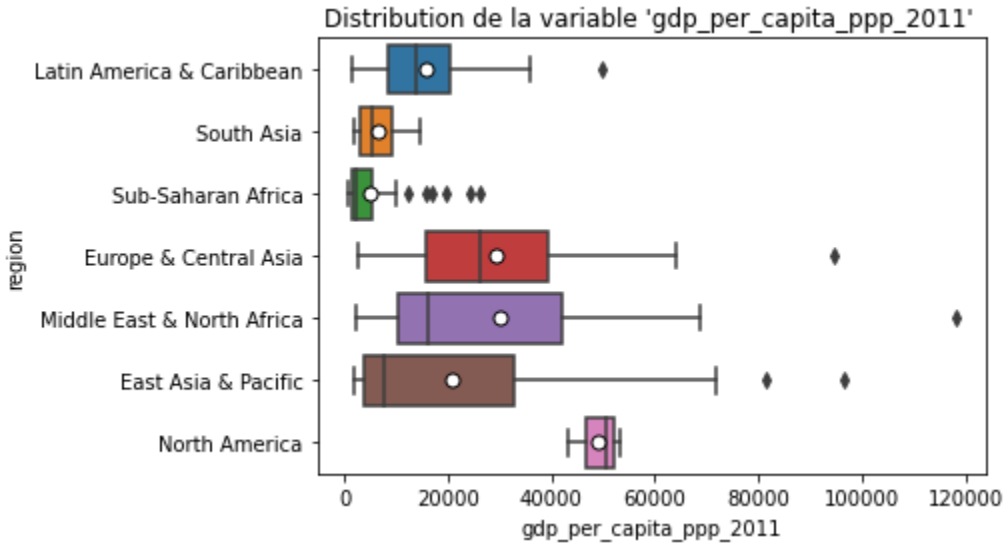
academy

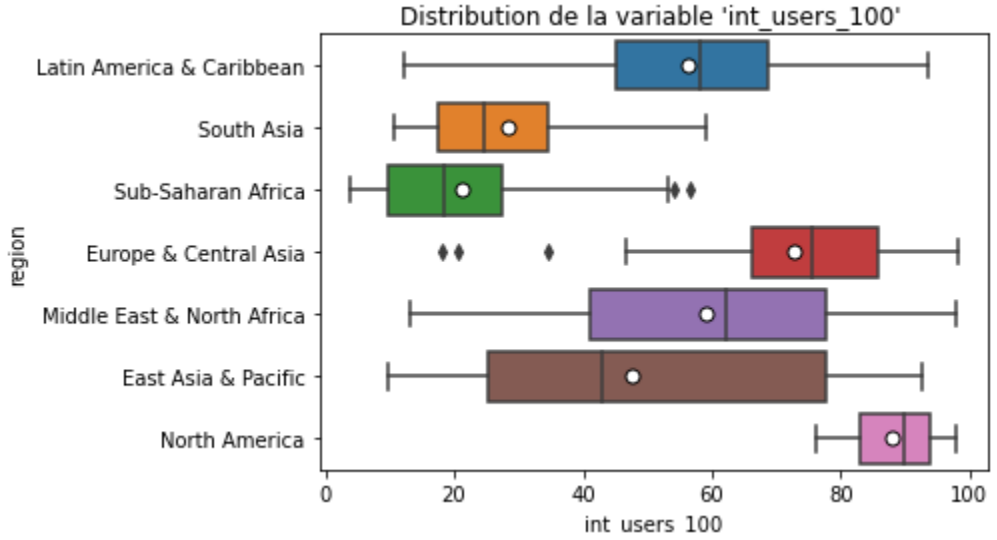
- La librairie **missingno** est très utile pour évaluer la qualité d'un jeu de données volumineux
- A la base, le jeu de données comportait un nombre important de valeurs manquantes cependant après avoir appliqué des filtres successifs, **les données restantes sont de bonnes qualités**
- De plus, les pays qui comportent le moins de valeurs manquantes sont les **pays développés.** Ils sont **les plus intéressants** compte tenu de notre démarche d'expansion

Analyse exploratoire

academy

Distribution des variables par région





academy

Normalisation des variables

- **Min-Max Scaler**
- Objectif : création du score

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Standardisation**
- Objectif : utilisation d'une PCA

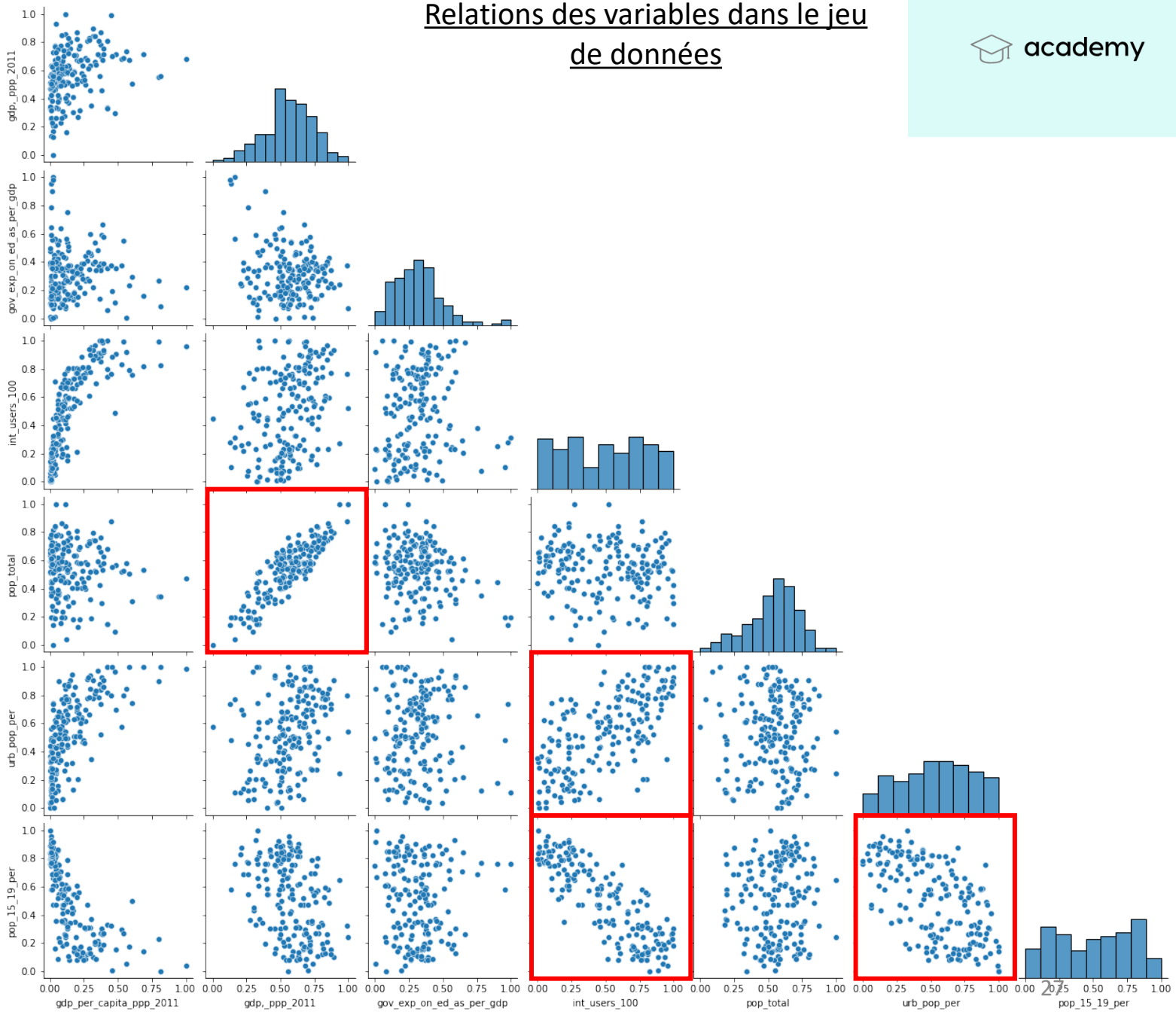
$$z = \frac{x_i - \mu}{\sigma}$$

Relations

Analyse graphique :

- Corrélations négatives :
  - ➢ le nombre d'utilisateurs d'internet et la population 15-19 ans.
  - ➢ La population urbaine et la population 15-19 ans.
- Corrélations positives :
  - ➢ le nombre d'utilisateurs d'internet et la population urbaine

Relations des variables dans le jeu de données



Création du score d'attractivité

Formule :

'score' =

'gdp_per_capita_ppp_2011' *** 3**

+ 'int_users_100' *** 7**

+ 'urb_pop_per' *** 2**

+ 'pop_15_19_per' *** 3**

+ 'pop_total' *** 4**

- 'gov_exp_on_ed_as_per_gdp'

- Pour effectuer ce calcul, j'ai utilisé les **variables normalisées** avec MinMaxScaler. Chaque valeur est donc comprise entre 0 et 1.
- Le tout est divisé par 18 et multiplié par 100 pour avoir **un score entre 0 et 100**

Aperçu des scores :

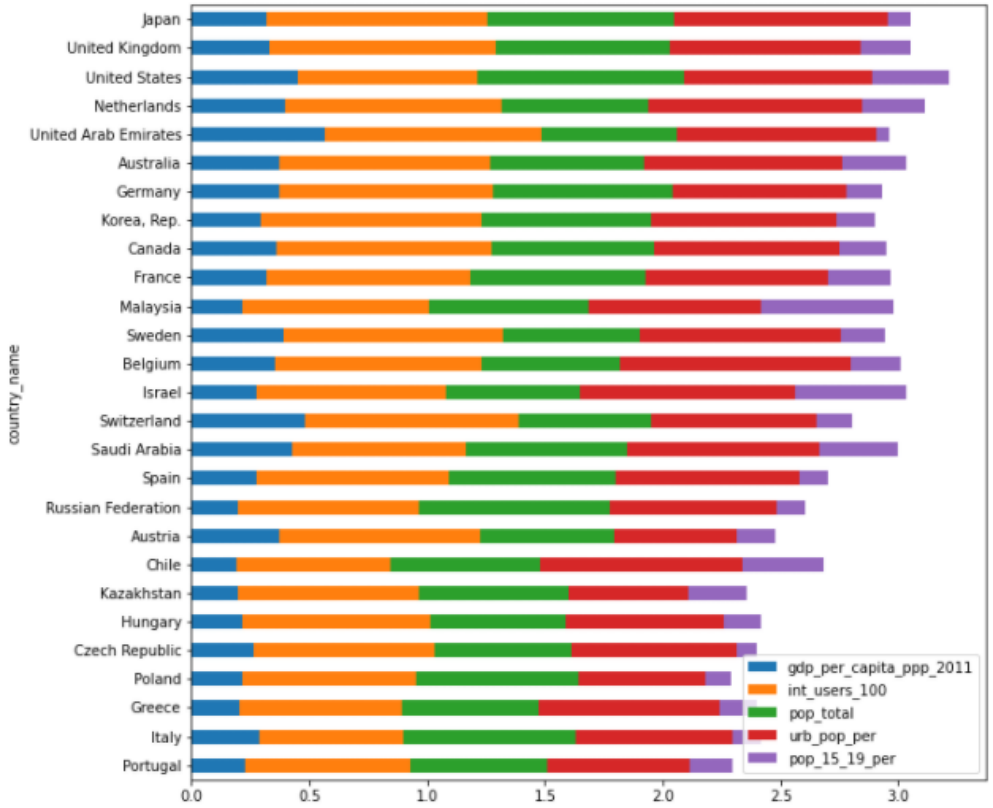
country_name	score
Aruba	60.475534
Afghanistan	39.189745
Angola	43.664257
Albania	55.797662
United Arab Emirates	76.779629

Application de masques

- Afin de réduire la liste des pays potentiels, j'ai appliqué différents masques :
  - Nombre d'utilisateurs d'internet pour 100 habitants : **> 60**
  - PIB par habitant : **> $20 000**
  - Population 15-19 ans : **> 2%**
  - Population : **> 8 000 000**

Pays après application du masque (1)

Stacked bar-chart des pays



Parmi la liste des pays pré-sélectionné :

- Le Japon a obtenu le meilleur score : **78,5**
- … et le Portugal, le moins bon : **58,3**

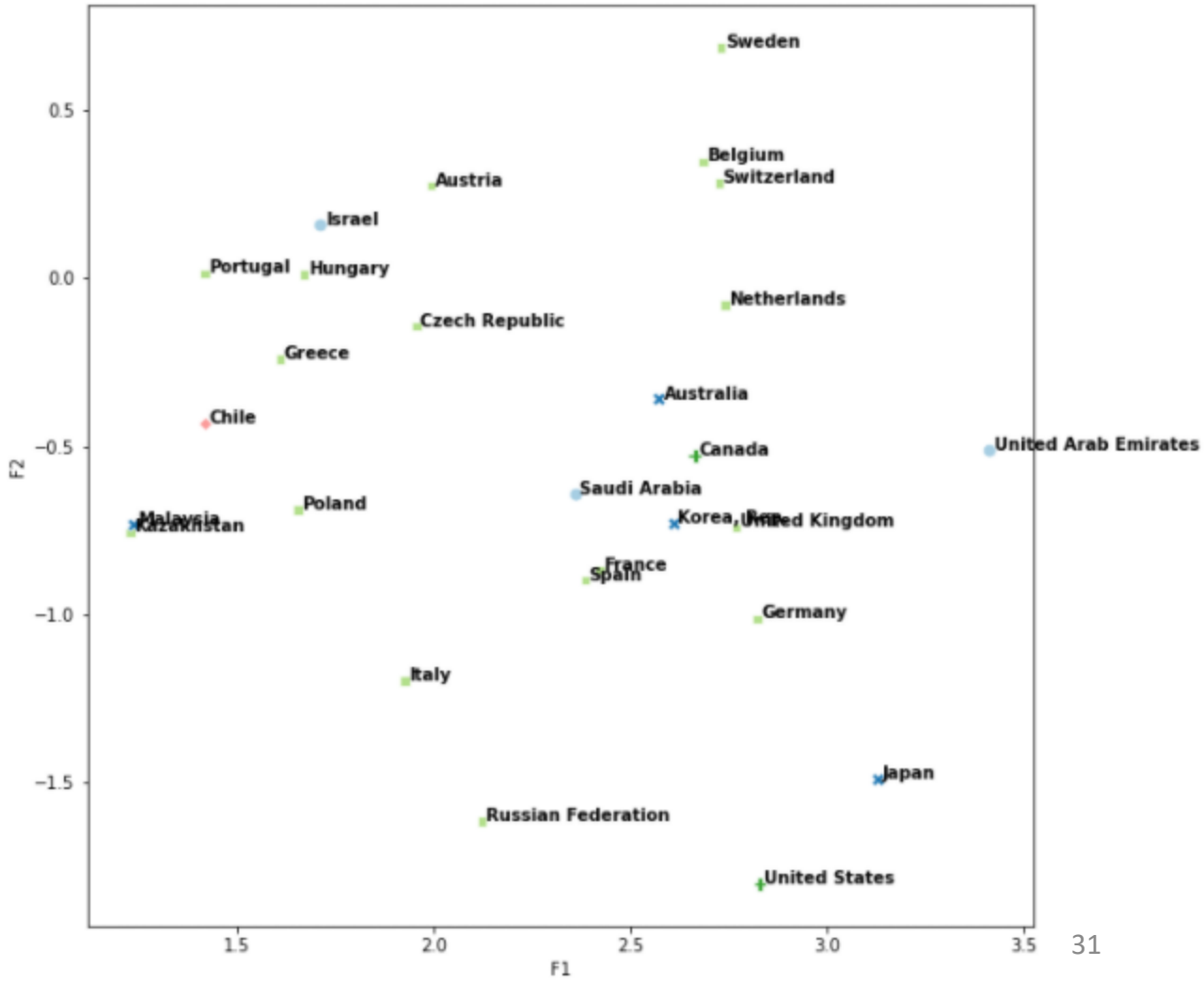
Pays après application du masque (2)

academy

Visualisation à l'aide du PCA

Interprétation des axes :

- F1 = Richesse par habitants
- F2 = Population



Ajout du niveau d’anglais

- Source : EF – Education First - http://www.ef.com.fr/

Aperçu des scores :

country	score_eng
Netherlands	652
Denmark	632
Finland	631
Sweden	625
Norway	624
Austria	623
Portugal	618
Germany	616
Belgium	612
Singapore	611

- La stratégie que je propose est d’**ajouter une variable mesurant le niveau d’anglais des pays**
- L’objectif est de sélectionner des pays avec un niveau de maîtrise de l’anglais assez élevé pour pouvoir limiter **les coûts de traduction de nos ressources**
- De plus, la maîtrise de la langue anglaise, permettra de **faciliter la communication** avec les professeurs que nous recruterons dans ces pays

Liste des pays

- Après avoir appliqué un second masque (**score > 70 et score en anglais > 0,7***)

Liste des pays retenus

country_name	gdp_per_capita_ppp_2011	gdp,_ppp_2011	gov_exp_on_ed_as_per_gdp	int_users_100	pop_total	urb_pop_per	pop_15_19_per	score_eng	score
Australia	44414.029479	1.071584e+12	5.22534	88.238658	25364307.0	0.86124	0.059580	1.000000	75.772383
Belgium	42058.661170	4.772884e+11	6.58514	86.516500	11484055.0	0.98041	0.055096	0.852399	72.895007
Canada	43087.757365	1.563501e+12	5.28122	89.840000	37589262.0	0.81482	0.053622	1.000000	75.139453
Switzerland	57430.053265	4.808100e+11	5.09608	89.405568	8574832.0	0.73849	0.049800	0.763838	72.231637
Germany	44260.359679	3.658901e+12	4.95219	89.647101	83132799.0	0.77376	0.049695	0.867159	75.662779
United Kingdom	39229.848765	2.574939e+12	5.68427	94.775801	66834405.0	0.83652	0.054561	1.000000	78.337061
Israel	32684.201151	2.793551e+11	5.75727	79.778791	9053300.0	0.92501	0.077080	0.710537	72.240142
Netherlands	47302.702570	8.050167e+11	5.52938	90.410959	17332850.0	0.91876	0.059691	1.000000	77.111458
Sweden	46662.050625	4.621000e+11	7.67509	91.506828	10285453.0	0.87708	0.052660	0.900369	73.084360
United States	53341.815958	1.723621e+13	5.38078	76.176737	328239523.0	0.82459	0.064579	1.000000	77.564869

**Le score a été normalisée avec un MinMaxScaler*

Conclusion (1)

academy

- Quels sont les pays avec un fort potentiel de clients pour nos services ?

L' Australie, la Belgique, le Canada, la Suisse, l'Allemagne, le Royaume-Uni, Israël, les Pays-Bas, la Suède et les Etats-Unis.

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

Taux de croissance moyen de la population entre 2005 et 2019

	growth_rate_mean
Israel	0.019274
Australia	0.015703
Canada	0.011018
Switzerland	0.010222
Sweden	0.009348
United States	0.007531
United Kingdom	0.007256
Belgium	0.006569
Netherlands	0.004312
Germany	0.000592

Conclusion (2)

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

Sélection des pays européens

	growth_rate_mean	gdp_per_capita_ppp_2011	gdp,_ppp_2011	gov_exp_on_ed_as_per_gdp	int_users_100	pop_total	urb_pop_per	pop_15_19_per	score_eng	score	region
Belgium	0.006569	42058.661170	4.772884e+11	6.58514	86.516500	11484055.0	0.98041	0.055096	0.852399	72.895007	Europe & Central Asia
Germany	0.000592	44260.359679	3.658901e+12	4.95219	89.647101	83132799.0	0.77376	0.049695	0.867159	75.662779	Europe & Central Asia
Netherlands	0.004312	47302.702570	8.050167e+11	5.52938	90.410959	17332850.0	0.91876	0.059691	1.000000	77.111458	Europe & Central Asia
Sweden	0.009348	46662.050625	4.621000e+11	7.67509	91.506828	10285453.0	0.87708	0.052660	0.900369	73.084360	Europe & Central Asia
Switzerland	0.010222	57430.053265	4.808100e+11	5.09608	89.405568	8574832.0	0.73849	0.049800	0.763838	72.231637	Europe & Central Asia
United Kingdom	0.007256	39229.848765	2.574939e+12	5.68427	94.775801	66834405.0	0.83652	0.054561	1.000000	78.337061	Europe & Central Asia

Pistes de développement

- Afin de mesurer l'évolution potentielle des clients dans un pays, il aurait été envisageable de calculer un taux de croissance du score mesuré sur différentes années
- Cette approche aurait permis de prendre en compte une multitude de variables
- L'un des problèmes majeurs de cette approche est de trouver une alternative à la fonction qui a récupéré la dernière valeur non-nulle