Projet 2 -
Analyse de données des systèmes éducatifs

EVA BOOKJANS

OPENCLASSROOMS

LA PROBLÉMATIQUE

La start-up de la EdTech, "academy", propose des contenus de **formation en ligne** pour un public de **niveau lycée et université** et considère **une expansion à l'international.**

- Quels sont **les pays avec un fort potentiel de clients** pour les services de la start-up ?
- Pour chacun de ces pays, quel sera **l'évolution de ce potentiel** de clients ?
- Dans quel pays l'entreprise doit-elle opérer en **priorité** ?

LA MISSION

Regarder **le jeu de données de la Banque mondiale dur l'éducation** pour voir s'il peut aider à répondre à la problématique.

- **Valider la qualité** de ce jeu de données (comporte-t-il beaucoup de données manquantes, dupliquées ?)
- **Décrire les informations** contenues dans le jeu de données (nombre de colonnes ? nombre de lignes ?)
- **Sélectionner les informations qui semblent pertinentes** pour répondre à la problématique (quelles sont les colonnes contenant des informations qui peuvent être utiles pour répondre à la problématique de l'entreprise ?)
- Déterminer des **ordres de grandeurs des indicateurs statistiques classiques** pour les différentes zones géographiques et pays du monde (moyenne/médiane/écart-type par pays et par continent ou bloc géographique

LE JEU DE DONNÉES

- **La Banque mondiale** a une répertorie de données en ligne sur la thématique éducation: EdStats All Indicator Query

  - https://datacatalog.worldbank.org/dataset/education-statistics
  - " Le jeu de données contient plus que 4000 indicateurs internationaux qui décrivent l'accès à l'éducation, la progression, l'achèvement, l'alphabétisation, les enseignants, la population et les dépenses. Les indicateurs couvrent le cycle d'éducation du pré-primaire à l'enseignement professionnel et supérieur. La requête contient également des données sur les résultats d'apprentissage provenant d'évaluations internationales et régionales 2050." (traduction de la déscription du jeu de données)

  - 5 fichiers qui donne de l'information sur les données:

    EdStatsData, EdStatsCountry, EdStatsCountry-Series, EdStatsSeries, EdStatsFootNote

EdStatsData – 886930 lignes 69 colonnes

Country Name	Country Code	Indicator Name	Indicator Code	1970	...	2100
Arab World	ARB	Adjusted ...	UIS.NERA.2	NaN	...	...
...	...	...	...	...	...	...

Les données sur les indicateurs éducatifs

EdStatsCountry – 241 lignes 31 colonnes

Country Code	Short Name	...
ABW	Aruba	...
...	...	...

Les pays, régions, groupements de pays
Code, noms, description, régions, devise, ...

EdStatsSeries – 3665 lignes 20 colonnes

Series Code	Topic	Indicator Name	...
BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: ...	...
...	...	...	...

Les indicateurs éducatifs
Code, noms, thématique, descriptions, ...

EdStatsCountry-Series – 613 lignes 3 colonnes

Country Code	Series Code	DESCRIPTION
ABW	SP.POP.TOTL	Data sources: ...
...	...	...

Sources des données sur les pays
Population, démographie, PIB, ...

EdStatsFootNote – 643638 lignes et 4 colonnes

Country Code	Series Code	Year	DESCRIPTION
ABW	SE.PRE.ENRL.FE	YR2001	Country estimation ...
...	...	...	...

Notes sur les données

LE JEU DE DONNÉES - EdStatsData

- **3665 indicateurs** internationaux sur l'éducation sur 37 thématiques
- **242 pays**, régions, groupements de pays
- Les années: 1970 à 2017 et 2010 à 2100 (pour les projections) (**65 années**)

- 886 930 combinaisons de pays-indicateur (= nombre des lignes = 3665 x 242)
  - Pas de duplicates, **mais approx. 60% des lignes sont vides**
- 57 650 450 combinaisons de pays-indicateur-année
  - **Moins de 10% des données sont complètes.**

- **type des données : float64**

Plus que 5 Million données, mais beaucoup ne sont pas pertinentes et le jeu de données n'est pas très complète dans sa totalité.

FILTRAGE / NETTOYAGE

- **ANNÉES / DATES**
  - **Filtrer les colonnes**
  - Les années plus récentes 2010 – 2016*
  - et des projections (2010 – 2050)**

*ce sont les plus récents dates disponibles pour les indicateurs pertinents

** il y a des indicateurs intéressants

- **PAYS / RÉGIONS**
  - **Choisir les lignes avec le 'Country Code'**
  - Pays et les régions géographiques
    - écarter les groupes de pays non-géographiques
  - 'EdstatsCountry' donne les définitions

217 pays et 7 régions

LES INDICATEURS PERTINENTS*

* ET DISPONIBLE / COMPLÈTE

- **Infrastructure**
  - internet, connectivité, …
- **Clientèle**
  - Niveau éducation = BAC(+)
  - Connaissances techniques
- **Marché**
  - nombre des potentiels clients
  - moyennes financières, …

THÉMATIQUES

'Infrastructure : Communications'

Pas d'information sur l'infrastructure elle même
--> usage de l'internet / des ordinateurs

'Upper **Secondary** + **Tertiary** Education'

nombre des étudiantes,
taux d'inscription / réussite / achèvement, …

'Attainment'
niveau d'éducation, projections

'Learning Outcomes'
compétences de base (ICT, math, lecture, écriture, …)

'Population' démographie, croissance

'Economic Policy & Debt' PIB, PNB

** utiliser 'EdstatsSeries

LES INDICATEURS

- **Utilisateurs d'internet** (% de la population)
  - **IT.NET.USER.P2** (211*)
- **PIB / PIB per capita** (US Dollar)
  - NY.GDP.MKTP.CD / **NY.GDP.PCAP.CD** (211*)
- **Population / croissance démographique**
  - SP.POP.TOTL / SP.POP.GROW (222*)
- **Projections sur le niveau d'éducation** (population/mille, taux)
  - PRJ.POP. … / PRJ.ATT. … (166**)
- **'Upper Secondary'** (étudiants, population démographique, taux d'inscription)
  - UIS.E.3 (188*), SP.SEC.UTOT.IN (204*), SE.SEC.ENRR.UP (184*)
- **'Tertiary'** (étudiants, population démographique, taux d'inscription)
  - **SE.TER.ENRL** (179*), SP.TER.TOTL.IN (202*), **SE.TER.ENRR** (172*)

*pays/régions entre les années 2010-2016
** pays/régions pour les années 2010-2100

NETTOYAGE DES DONNÉES

IT.NET.USER.P2

SE.TER.ENRR

SE.TER.ENRL

NY.GDP.PCAP.CD

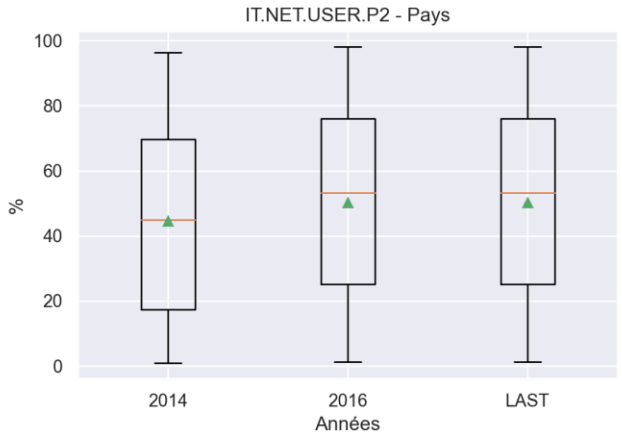
- **Pays / Régions**
  - Garder que les pays / régions pour lesquels on a **toutes les indicateurs**
  - **Écarter les pays avec moins de 0.5 Millions d'habitants**

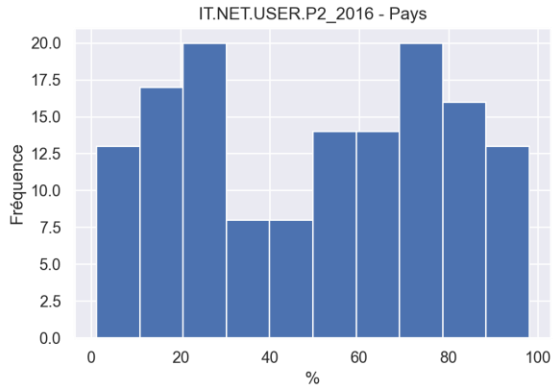
143 pays et 7 régions

- **Compléter les data sets**
  - **'LAST'**
    - Garder **la dernière valeur non-nul** pour chaque indicateur
  - **'CLEANED'**
    - Faire une **inter/extrapolation linéaire* avec les voisins le plus proche**
    - pour IT.NET.USER.P2 on utilise une courbe S

UTILISATEURS INTERNET (par 100 habitants)

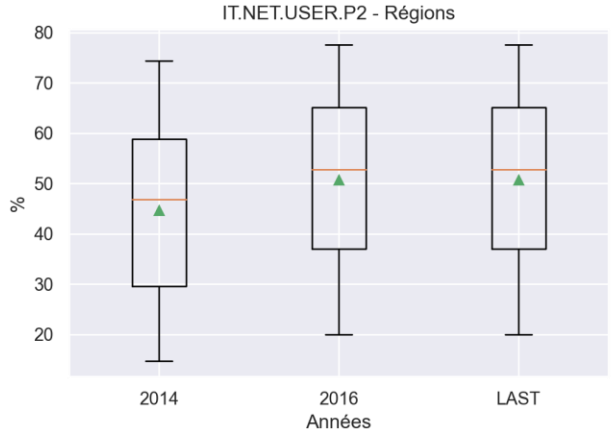
Pays

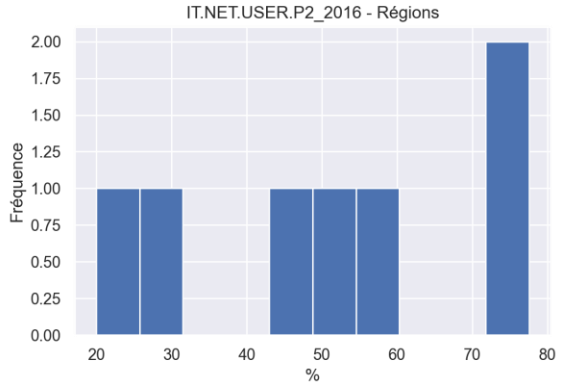




%	2014	2016	LAST
Médian	44.9	53.2	53.2
Moyenne	44.5	50.2	50.2
Ecart-type	28.8	28.4	28.4

Régions

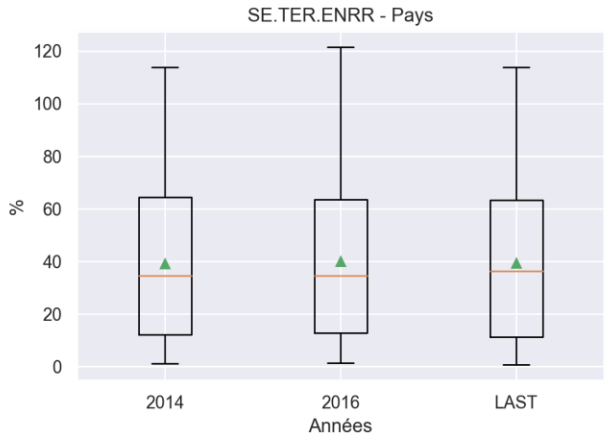


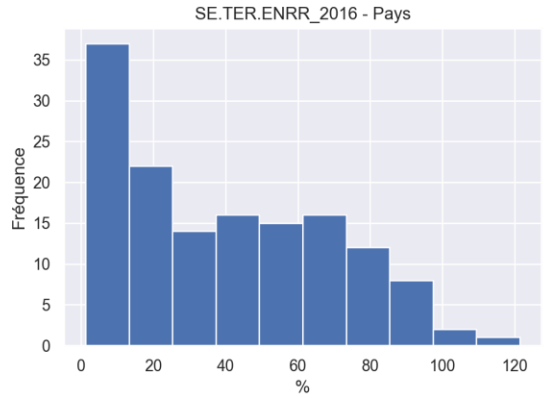


%	2014	2016	LAST
Médian	46.8	52.8	52.8
Moyenne	44.7	50.7	50.7
Ecart-type	22.6	21.7	21.7

TAUX D'INSCRIPTION (TERTIARY) (%)

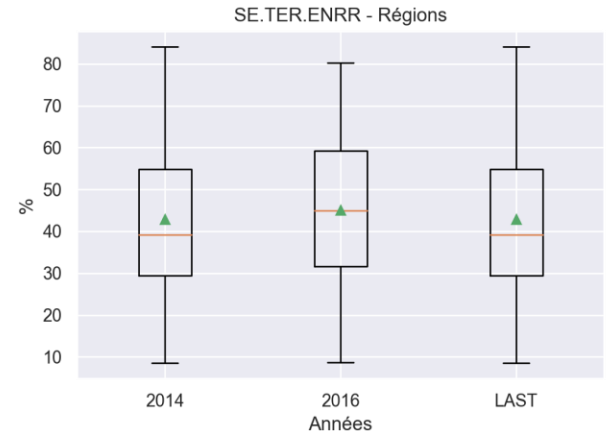
Pays

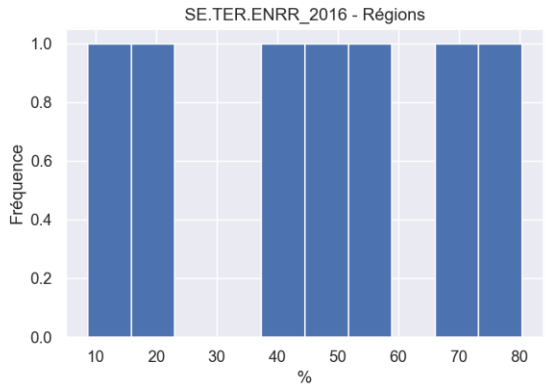




%	2014	2016	LAST
Médian	34.6	34.6	36.3
Moyenne	39.2	40.0	39.4
Ecart-type	28.3	29.2	28.6

Régions

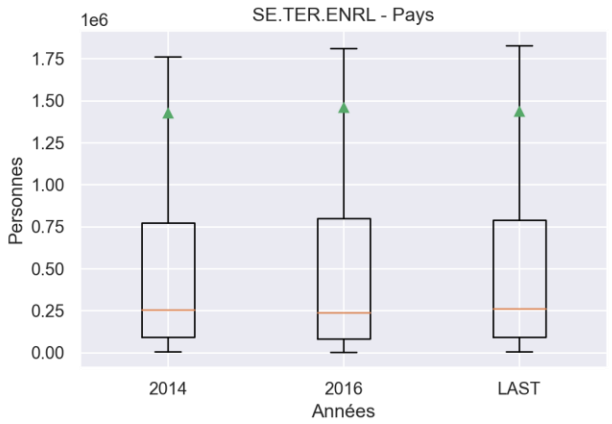


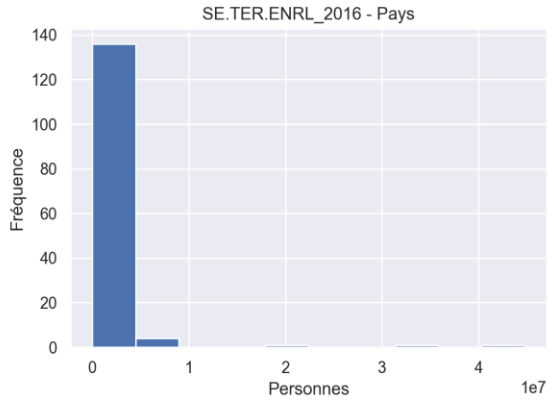


%	2014	2016	LAST
Médian	39.1	45.0	39.1
Moyenne	42.9	45.1	42.9
Ecart-type	25.5	24.7	25.5

ÉTUDIANTS (TERTIARY) (Personnes)

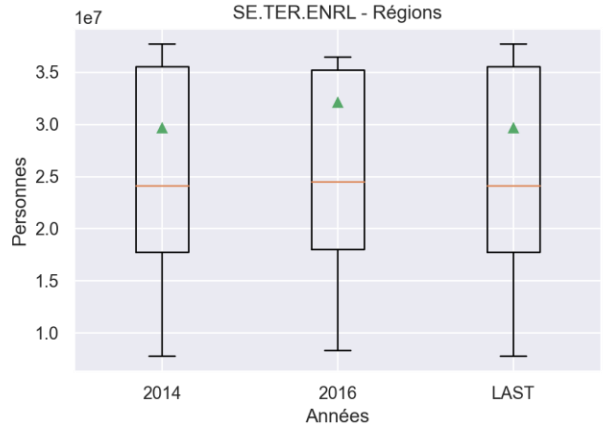
Pays

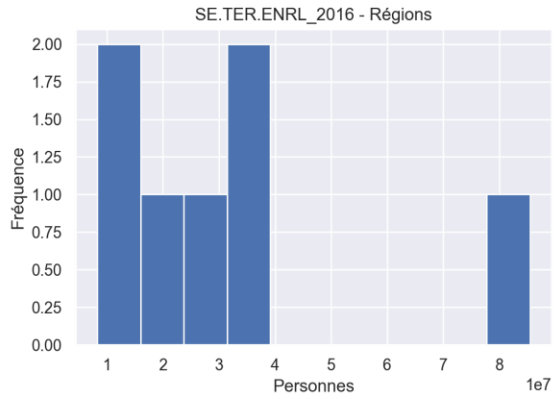




Million	2014	2016	LAST
Médian	0.254	0.239	0.261
Moyenne	39.2	40.0	39.4
Ecart-type	28.3	29.2	28.6

Régions

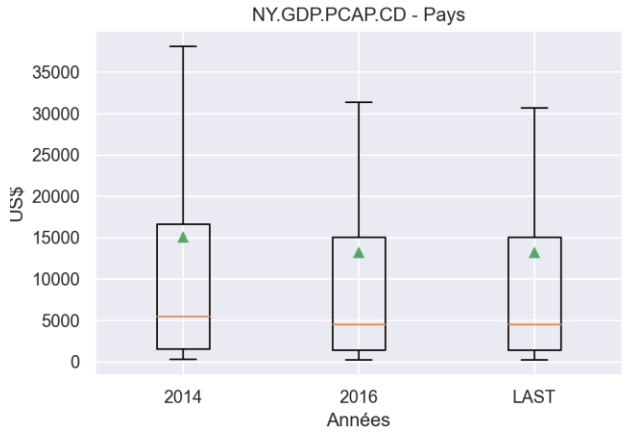


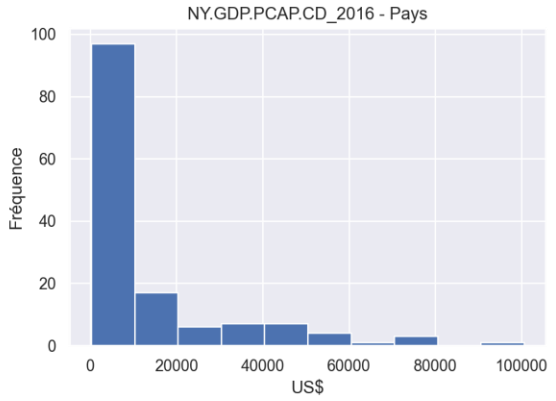


Million	2014	2016	LAST
Médian	24.1	24.5	24.1
Moyenne	29.6	32.1	29.6
Ecart-type	20.2	25.5	20.2

PIB per capita (US – Dollar)

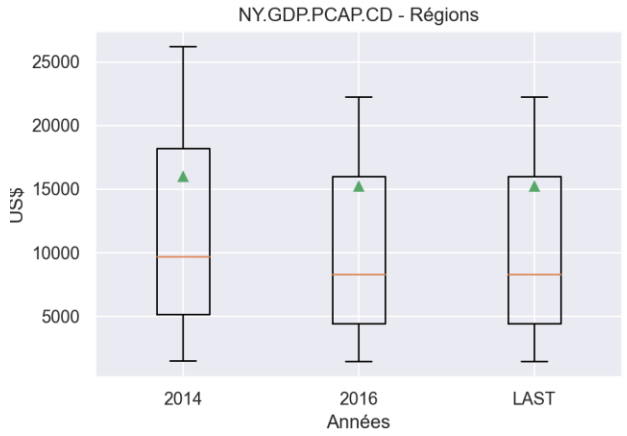
Pays

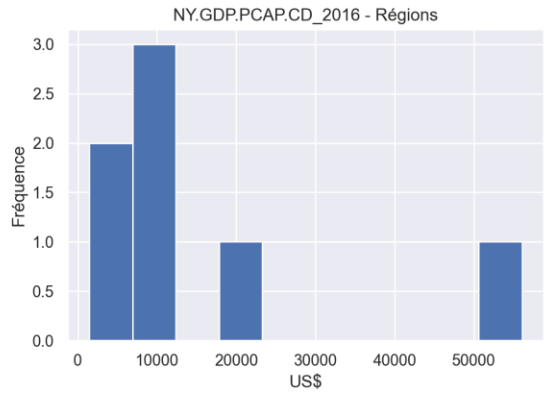




Mille US$	2014	2016	LAST
Médian	5.47	4.53	4.53
Moyenne	15.1	13.2	13.2
Ecart-type	22.2	18.9	18.9

Régions





Mille US$	2014	2016	LAST
Médian	9.67	8.31	8.31
Moyenne	16.0	15.2	15.2
Ecart-type	18.7	19.3	19.3

LE SCORE

Le Score d'Attractivité

- **IT.NET.USER.P2 - taux utilisateurs d'internet**
  - Échelle linéaire entre 0 et 100
- **SE.TER.ENRR - taux des étudiants**
  - Échelle linéaire score entre 0 et 100

Le Score de Marché

- **SE.TER.ENRL - nombre des étudiants**
  - Échelle log. entre 10 Mille et 10 Million
- **NY.GDP.PCAP.CD - PIB par habitant (US$)**
  - Échelle log. entre 500 et 50 000

- **Entre 0 et 10 pour chaque indicateur**
- **Poids égale pour computer le score total**

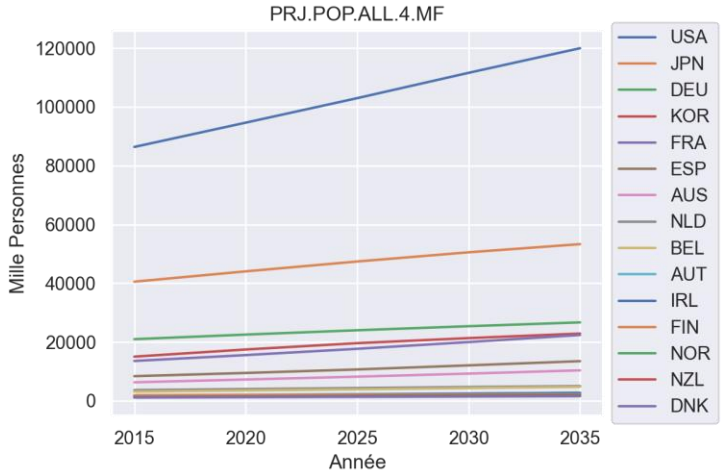
2016		Pays	Total	Attr	Mrkt
1.	AUS	Australia	9.7	9.4	10.0
2.	KOR	Korea	9.3	9.3	9.4
3.	NLD	Netherlands	9.3	8.7	9.8
4.	DNK	Denmark	9.1	9.0	9.2
5.	USA	United States	9.0	8.0	10.0
6.	ESP	Spain	8.9	8.5	9.3
7.	DEU	Germany	8.9	8.0	9.8
8.	NOR	Norway	8.9	8.7	9.1
9.	FIN	Finland	8.8	8.7	9.0
10.	AUT	Austria	8.8	8.4	9.3
11.	JPN	Japan	8.8	7.9	9.7
12.	NZL	New Zealand	8.8	8.8	8.8
13.	IRL	Ireland	8.8	8.6	8.9
14.	BEL	Belgium	8.7	8.2	9.3
15.	FRA	France	8.7	7.7	9.7

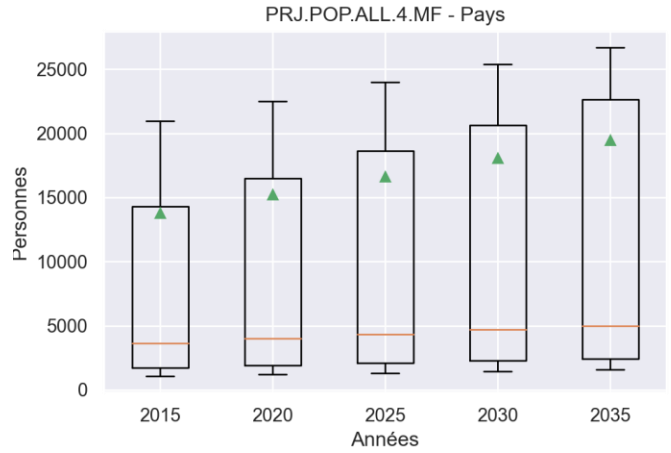
L'ÉVOLUTION

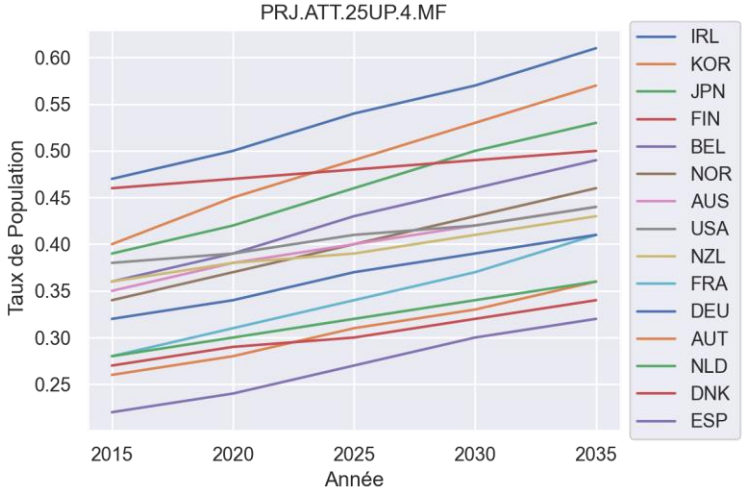
- **PRJ.POP.ALL.4.MF**
  - nombre des personnes avec une dégrée 'Post-Secondary' / 'Tertiary' (par mille)

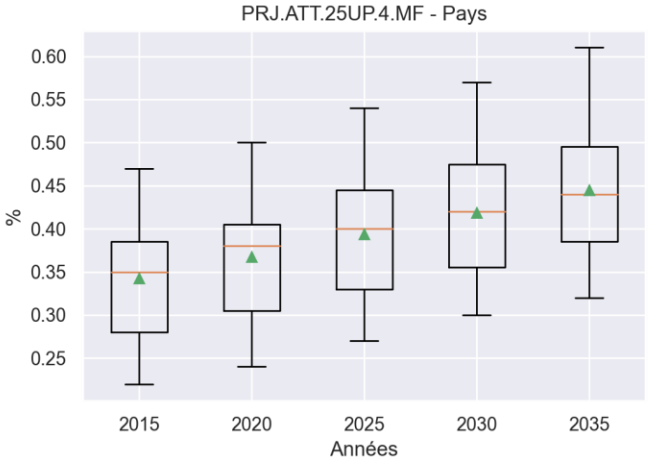
- **PRJ.ATT.25UP.4.MF**
  - taux des personnes âgées 25+ avec une dégrée 'Post-Secondary' / 'Tertiary'

Les projections (2015 –2035)









L'ÉVOLUTION

- **Le taux de changement**
  - Pour faire le classement prendre le moyenne du taux de changement des deux indicateurs

- **Les priorités**
  - **KOR, AUS, ESP, NOR**
    - Ces 4 pays sont classées dans **les top 8** dans l'évaluation actuelle et celui de l'évolution

		Pays	Total	ATT	POP
1.	FRA	France	55.8	46.4	65.3
2.	**ESP**	**Spain**	**53.2**	**45.4**	**61.1**
3.	**NOR**	**Norway**	**49.3**	**35.3**	**63.4**
4.	**KOR**	**Korea**	**47.4**	**42.5**	**52.3**
5.	IRL	Ireland	46.5	29.8	63.2
6.	**AUS**	**Australia**	**45.2**	**25.7**	**64.8**
7.	BEL	Belgium	44.3	36.1	52.4
8.	AUT	Austria	43.2	38.5	47.9
9.	NZL	New Zealand	35.0	19.4	50.5
10.	DNK	Denmark	34.7	25.9	43.4
11.	JPN	Japan	33.7	35.9	31.5
12.	NLD	Netherlands	33.0	28.6	37.5
13.	DEU	Germany	27.7	28.1	27.3
14.	USA	United States	27.3	15.8	38.9
15.	FIN	Finland	14.0	8.7	19.3

Fonctions utilisées

L'exploration du jeu de données

- **Première regarde :** DataFrame.head() / .info() / .describe()
- **Cellules pas vides / vides** : DataFrame.notna() / . Isna()
- **Obtenir les colonnes** : DataFrame.columns.values
- **Obtenir les index** : DataFrame.index.values
- **Informations numériques / statistiques** : .median() / . mean() / .std() / .sum() / .count() ….

Sélectionner les données

- **Sélectionner une colonne :** DataFrame['ColumnName']
- **Valeurs uniques :** DataFrame.unique()
- **Sélectionner des valeurs spécifiques** :

  DataFrame[DataFrame['ColumnName' == val]]
  DataFrame[DataFrame['ColumnName'.isin(someIterable)]

  DataFrame.loc[condition, columns]

  DataFrame.iloc[indices of index, indices of columns]
- **Filtrer les données** : DataFrame.filter(['Column1', Column2, …])

Manipuler les données

- **Transpose:** DataFrame.T
- **Faire un pivot :** DataFrame.pivot(columns = ['Column'])
- **Définir l'index :** DataFrame.set_index(['Column1', …])
- **Réinitialiser l'index :** DataFrame.reset_index()
- **Renommer des colonnes / lignes** : DataFrame.rename()
- **Supprimer des colonnes / lignes** : DataFrame.drop() / .remove()
- **Faire une copie** : DataFrame.copy()
- **Joindre des DataFrames / Series** : DataFrame.append(), / .merge() /.concat()

Nettoyer les données

- **Remplir les trous:** DataFrame.fillna()
- **Supprimer les colonnes / lignes avec des trous** : DataFrame.dropna()
- **Faire une interpolation :** DataFrame.interpolate()

- **Ajustement de courbe** : sp.optimize.curve_fit() (sp -> scipy)
- **S-courbe** : f(x) = c / (1+exp(-b*(x-x0)))

Analyser les données

- **Des opérations sur les cellules:** DataFrame.div(), .multiply(),...

- **Linéaire scoring** : score = 10 * (valeur – min_val) / (max_val-min_val)
- **Logarithmique scoring** : score = 10 – log_base(max_val – valuer) avec base = (max_valeur/min_valeur)**0.1

- **Taux de croissance** : (valeur_final – valeur_initial) / valeur_initial