Projet n°2: Analyser des données de systèmes éducatifs

SOUTENANCE DU 11/05/2023

Source de données :

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

Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

IV – Statistiques des indicateurs

V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

VII – Conclusion

Contexte de la problématique

- Academy: Start-Up de la EdTech
- Cibles visées: contenus de formation de niveau lycée et université.
- Objectif: étendre son marché à l'international

- But du projet:
- Vérifier la scalabilité du projet en effectuant une analyse exploratoire afin de repérer ceux qui répondent à la problématique(indicateurs ciblés, pays recommandés).

Présentation des données

EdStatCountry.csv	informations géographiques sur les pays, des données économiques globales et des dates de référence des dernières études. Le jeu de données contient 241 lignes et 32 variables. Nombre de valeurs manquantes totales : 2354 NaN pour 7712 observations (30.52 %). Pas de doublons.
EdStatSeries.csv	- Informations sur les indicateurs socio-éduco-économique classés en 37 thèmes. - 3661 lignes et 21 variables dont 5 avec des données complètes (pas de doublons).
EdStatsdata.csv	- Evolution (par année) des indicateurs par pays, et leurs projections - 886 930 lignes et 70 variables (86,09 % de données manquantes, pas de doublons)
EdStatsCountry-Series.csv	- Informations sur la source des indicateurs des pays - 613 lignes, et 4 variables (3 avec des données complètes, pas de doublons)
EdStatsFootNote.csv	- Informations sur la description() des indicateurs des pays - 643 638 lignes et 5 variables (4 avec des données complètes, 20 % de données manquantes, pas de doublons)

Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

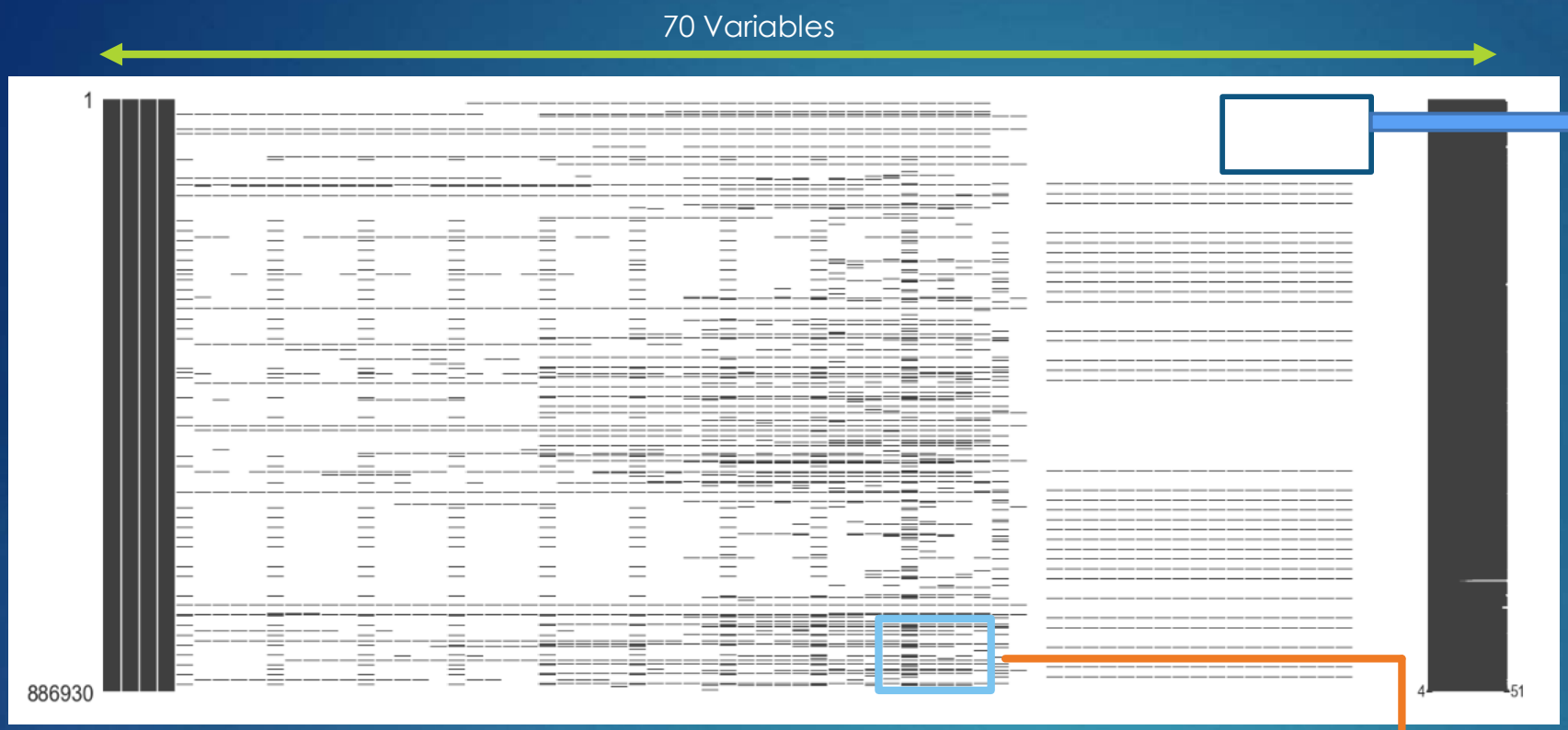
IV – Statistiques des indicateurs

V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

VII – Conclusion

Sélection des indicateurs



86,09 % de manquants

Extraction des données utiles:

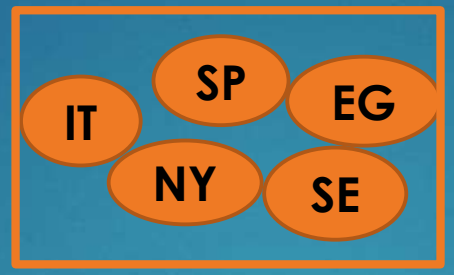
- Sélection de la plage de données (2010-2015)
- Sélection des variables et indicateurs pertinentes

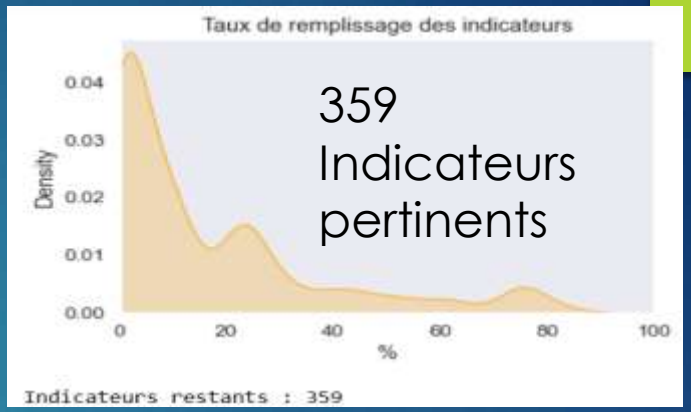
Sélection des indicateurs

Filtrage sur les 3665 indicateurs (sur la période 2010-2015)

THE WORLD BANK

Cibles:





Démographie (SP)	Mots Clés: population des 15, 20, 25 ans (POP.15, POP.20, POP.25), Total, Croissance population
Economie (NY)	Mots Clés: NY, (PIB) niveau de vie par habitant
Info Energétiques (EG)	Mots Clés: EG, termes liées à l'énergie
Education (SE)	Mots Clés: Nombre, Taux, Secondaire, Tertiaire
Infrastructures techniques (IT)	Mots Clés: IT, Internet

Indicateurs retenus:

- démographique :
  - **SP.POP.1524.TO.UN**
  - **SP.POP.AG25.TO.UN**
  - **SP.POP.TOTL**
  - **SP.POP.GROW**
- économique :
  - **NY.GNP.PCAP.PP.CD**
- éducation :
  - **SE.SEC.ENRR**
  - **SE.TER.ENRR**
  - **SP.SEC.TOTL.IN**
  - **SP.TER.TOTL.IN**
- numérique :
  - **IT.NET.USER.P2**

Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

IV – Statistiques des indicateurs

V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

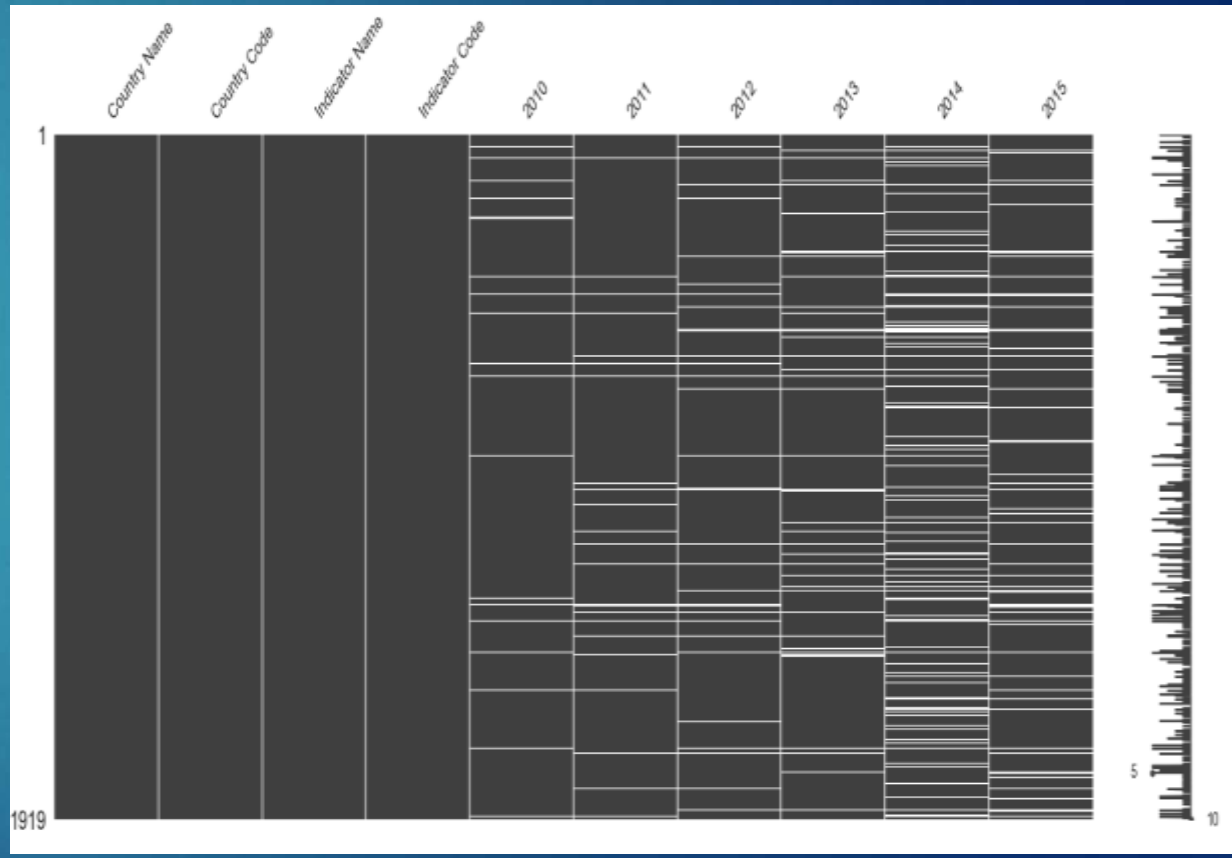
VII – Conclusion

Réduction des données

EdStatsData.csv

Suppression de la variable 'Unnamed: 69'	886930 lignes
Sélection de la période (2010-2015)	886930 lignes
Focalisation sur les Indicateurs retenus	2420 lignes
Suppression des pays non conformes (ISO)	2170 lignes
Indicateurs statistiques dépourvus d'info	1919 lignes

Données filtrés avec la période sélectionnée:



Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

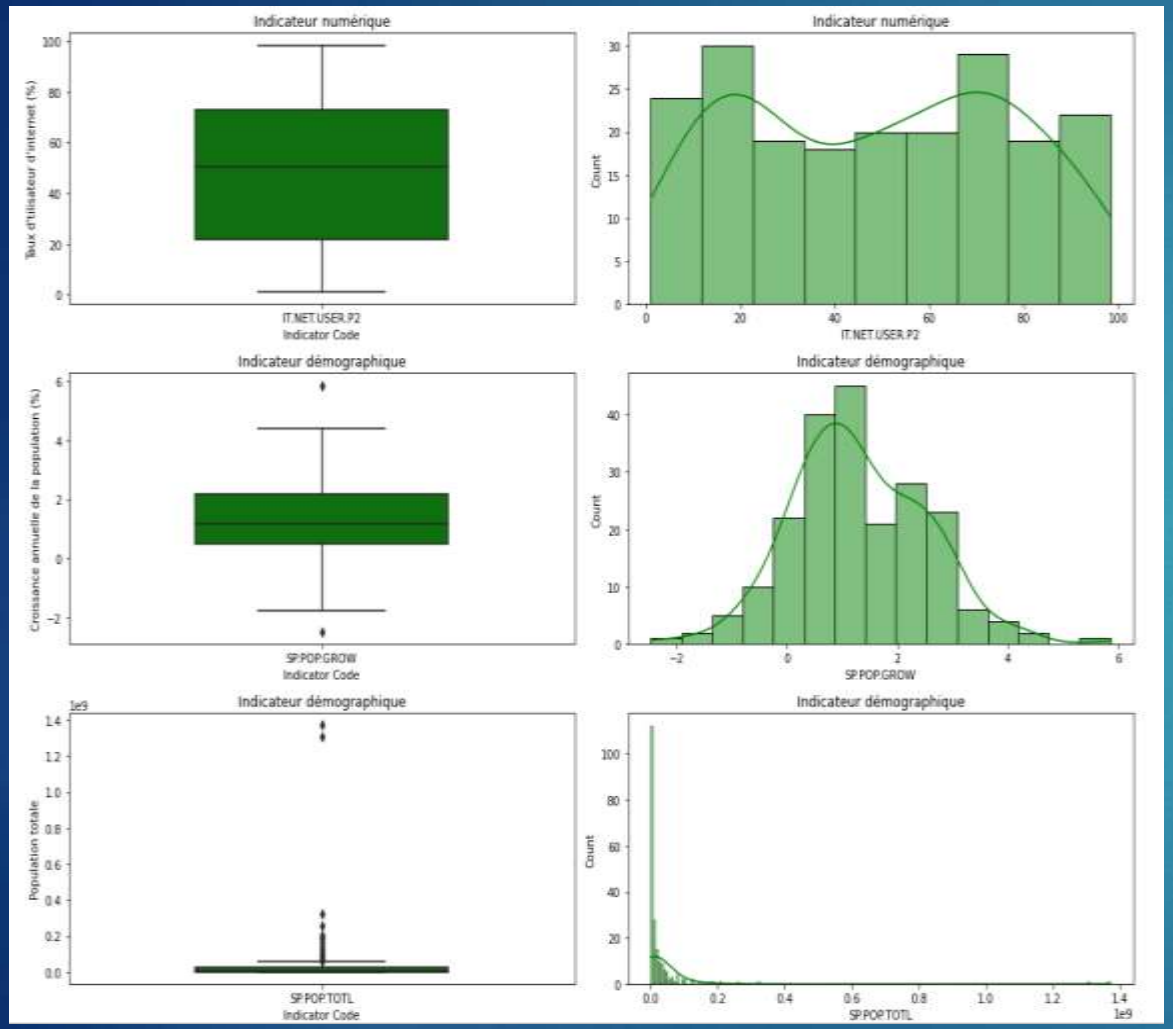
IV – Statistiques des indicateurs

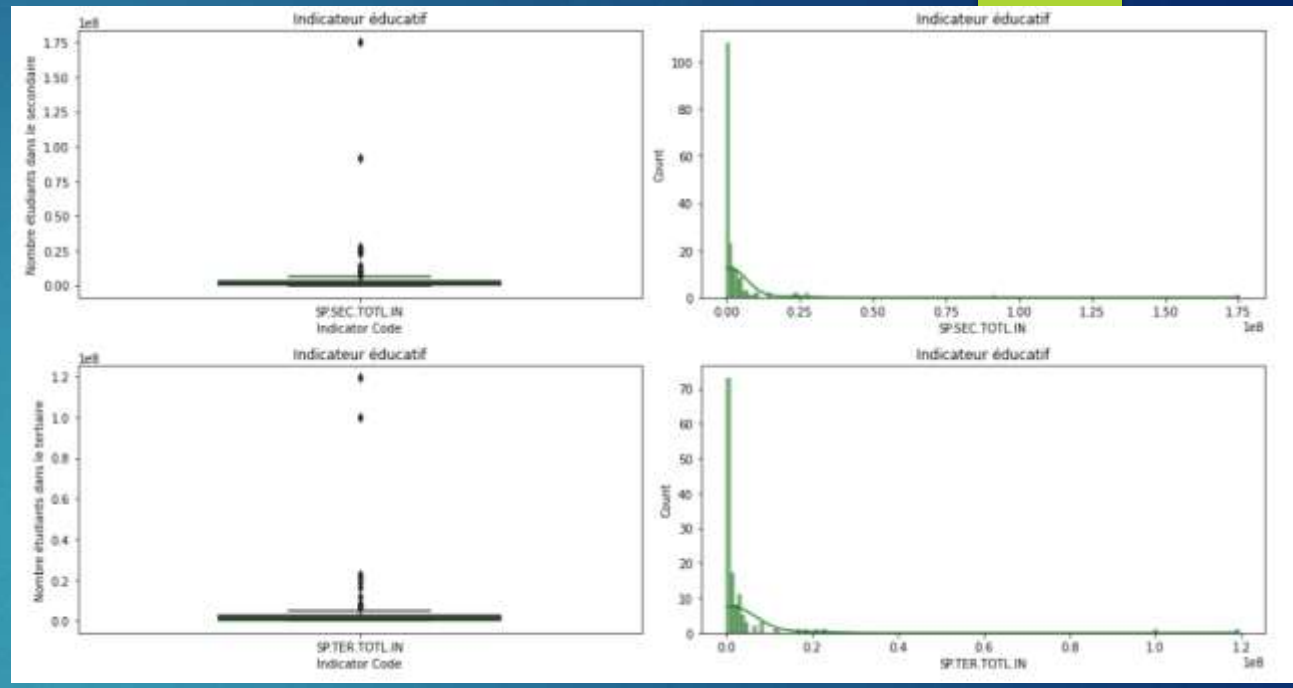
V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

VII – Conclusion

Statistiques des indicateurs





	Desc	Stat_num	Stat_dem_gr	Stat_dem_tot	Stat_edu_nsec	Stat_edu_nter
0	mean	48.785759	1.305702	3.483218e+07	3.869957e+06	3.672495e+06
1	median	50.300000	1.169910	6.475798e+06	7.553590e+05	6.695660e+05
2	var	810.594341	1.502722	1.824593e+16	2.139464e+14	1.925475e+14

Différence d'échelle entre les indicateurs: étape de normalisation nécessaire

Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

IV – Statistiques des indicateurs

V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

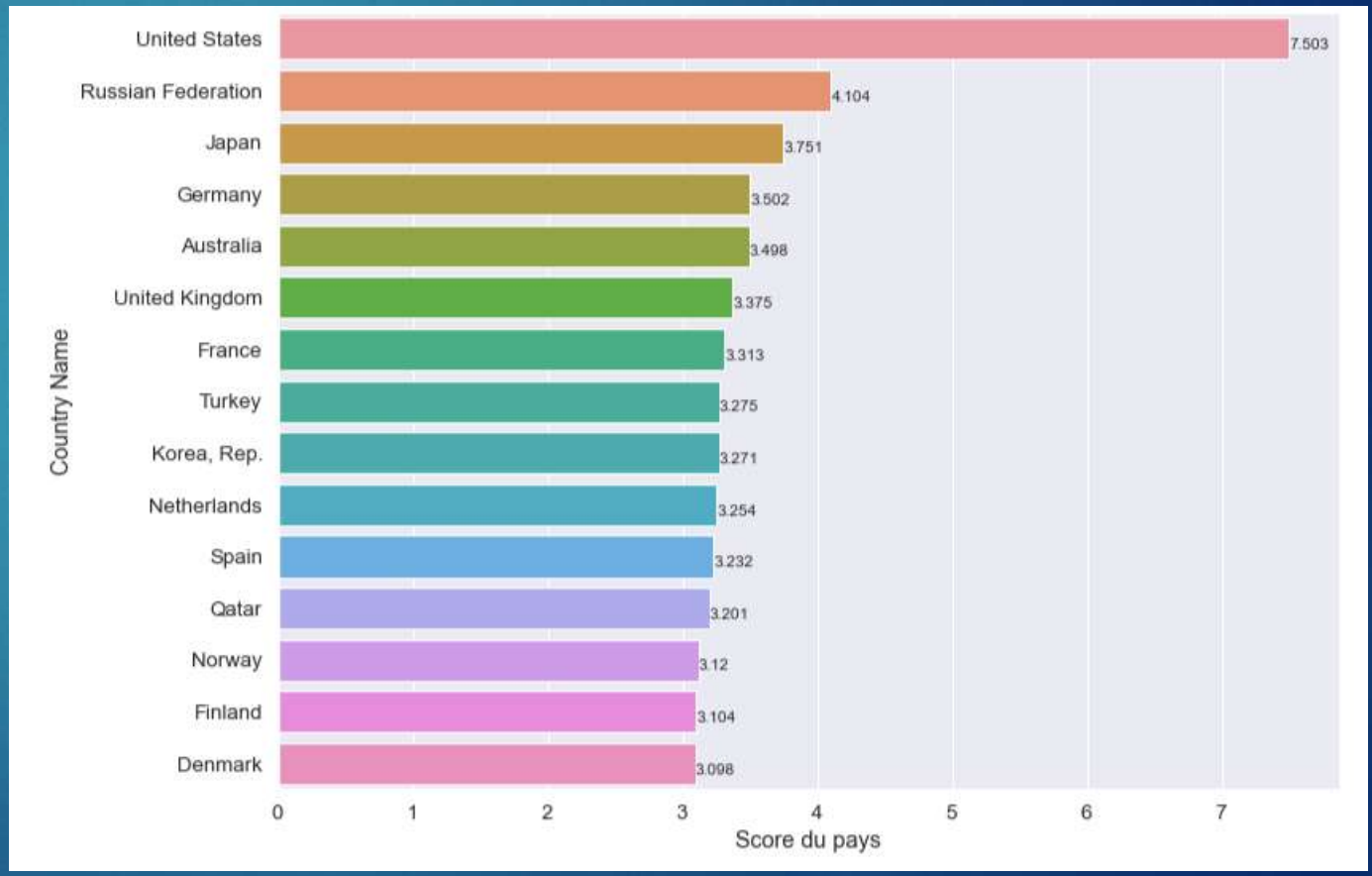
VII – Conclusion

Score (2 méthodes)

- Normalisation du ‘Min-Max’, avec des poids de pondérations similaires

- Transformation chaque valeur selon la normalisation 'Min' - 'Max' ((x-x(min))/(x(max)-x(min)) de façon à avoir des valeurs entre 0 et 1(mise à l'échelle des données). On calcule ensuite la somme pondérée des indicateurs sélectionnés pour avoir le score du pays.

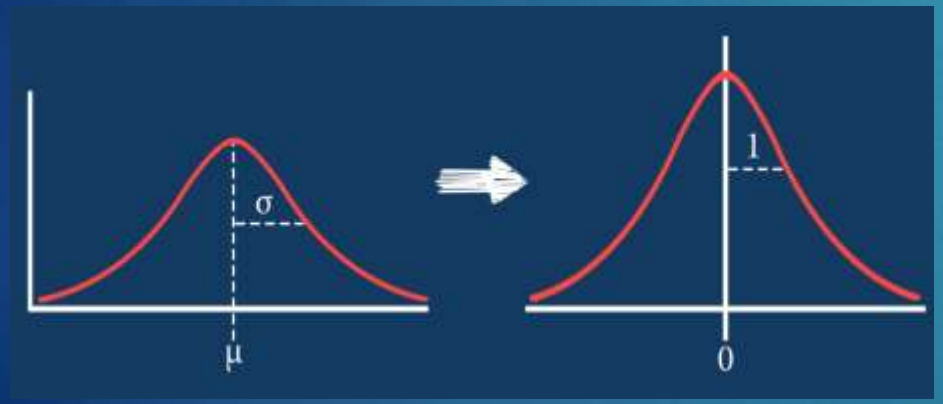
- En passant de 197 pays potentiels (10 indicateurs renseignés), au top 10 les plus attractifs.



Score (2 méthodes)

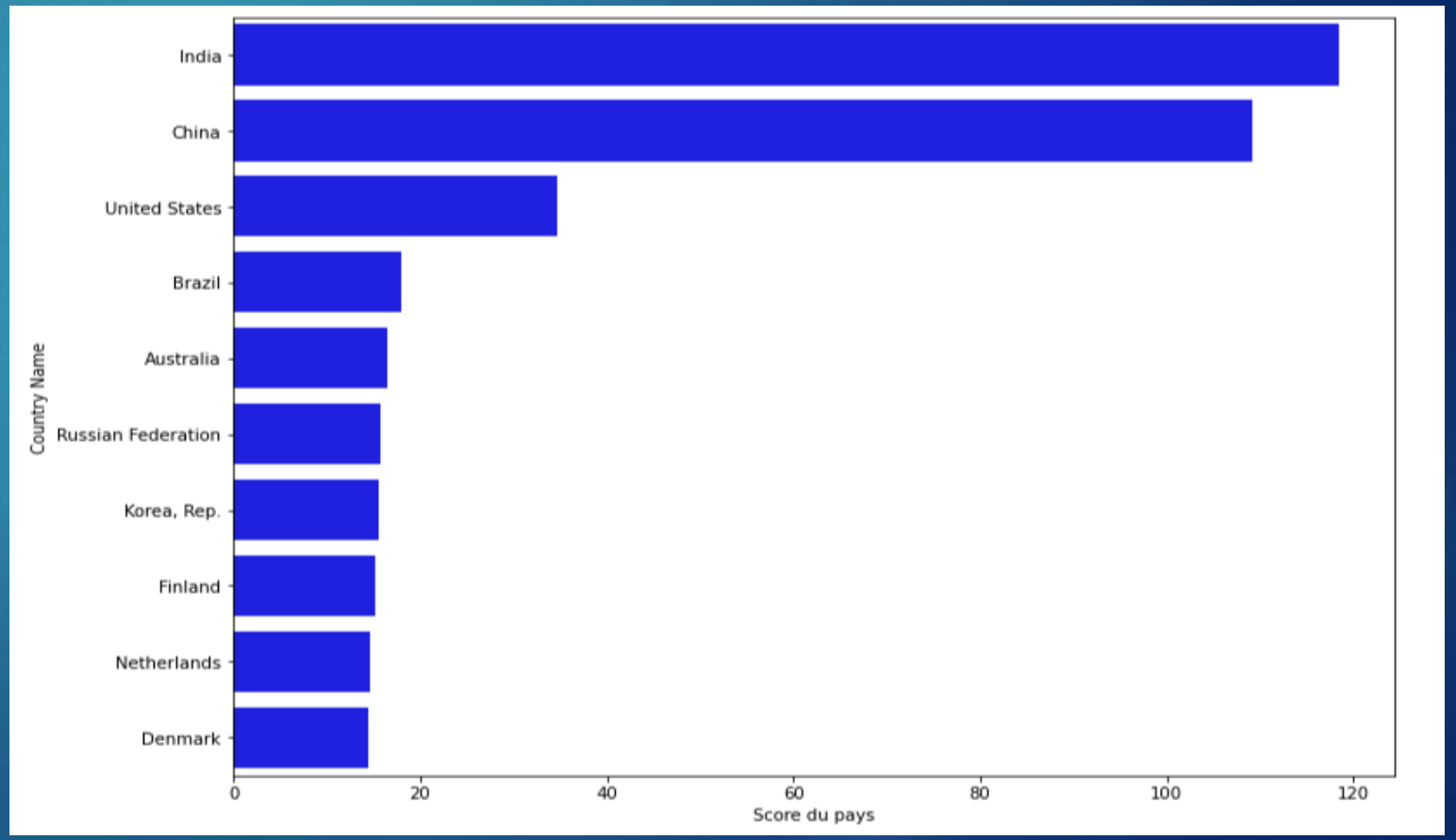
- Calcul du score avec la methode du 'Standard Scaler'

Standardisation : Mise à l'échelle des données afin d'obtenir une moyenne des variables de 0 et leurs variances de 1 (données centrées e réduites)




```
df_score_rk['IT.NET.USER.P2'].mean()
-1.9270299641708075e-16
```

```
df_score_rk['IT.NET.USER.P2'].var()
1.0071942446043158
```

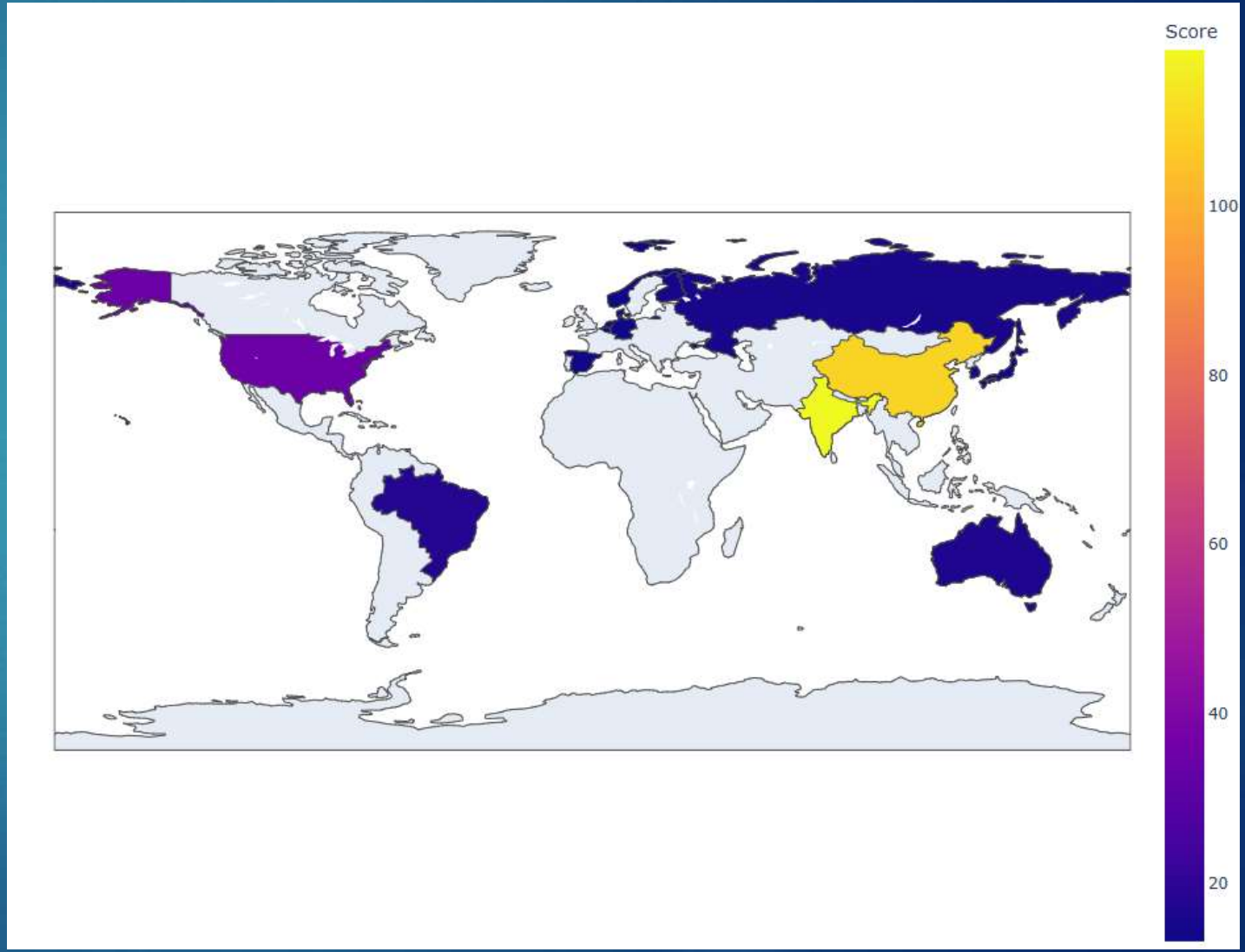


Score (2 méthodes)

Choix de la méthode:

- La standardisation semble la plus adapté au dataset, dû à la présence d'outliers, et au grandes valeurs des indicateurs.

- La normalisation va compresser les grandes valeurs dans une petite étendue.



Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

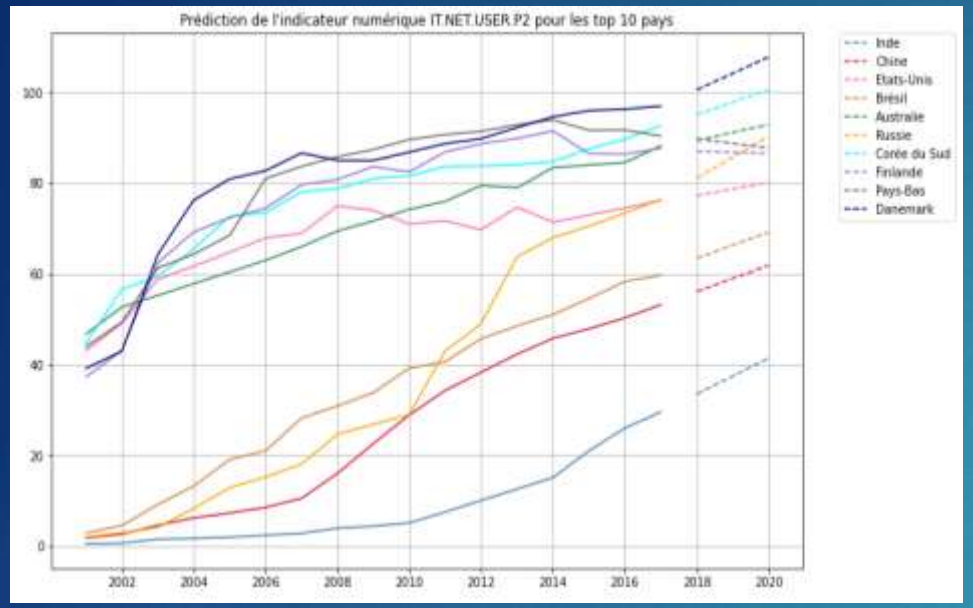
IV – Statistiques des indicateurs

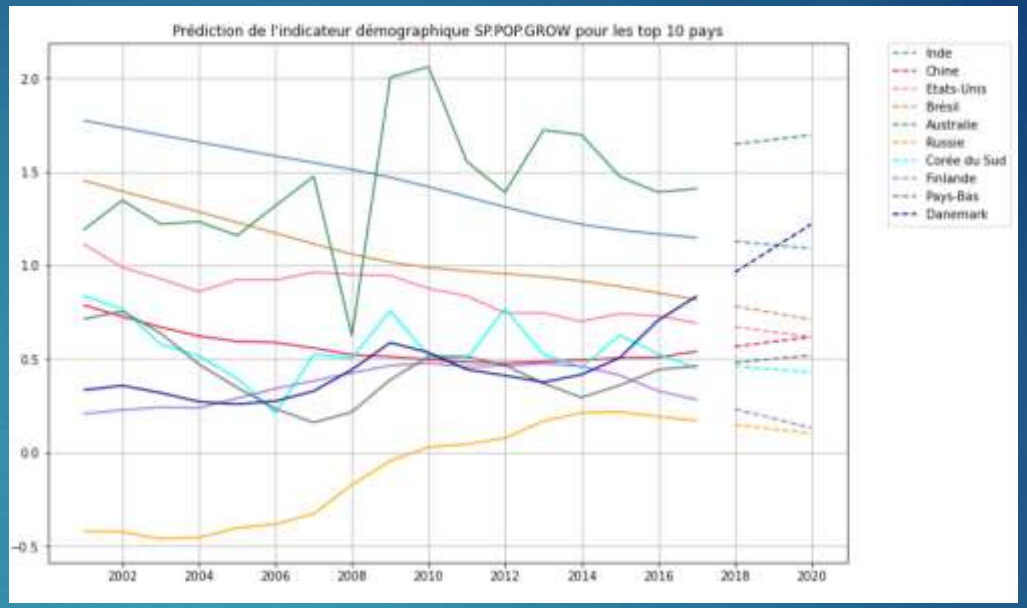
V – Score : 2 méthodes

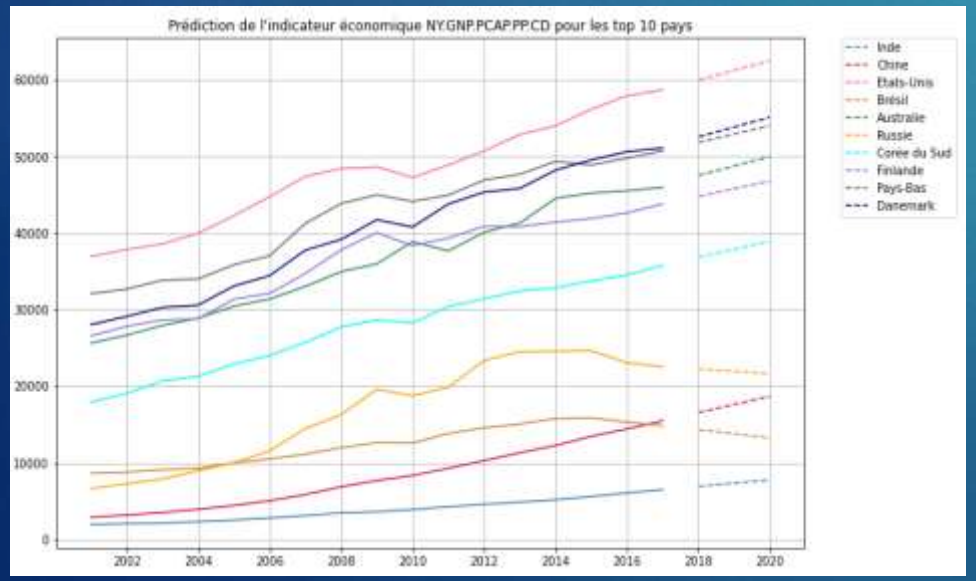
VI – Prévisions: potentiels des pays

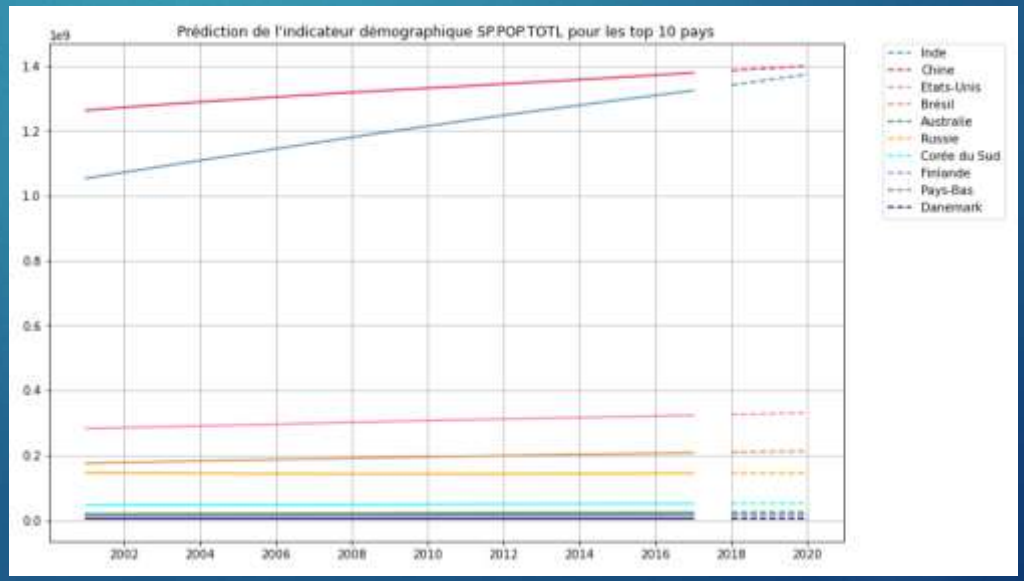
VII – Conclusion

Prévisions: potentiels des pays









Programme

I – Contexte de la problématique et présentation du jeu de données

II – Sélection des indicateurs

III - Réduction des données

IV – Statistiques des indicateurs

V – Score : 2 méthodes

VI – Prévisions: potentiels des pays

VII – Conclusion

Conclusion

Inde

Chine

Etats-Unis

Short Name	Score Prédit	Place Prévisionnelle	Score Finale
India	9.439473	2.0	82.106643
China	9.954581	1.0	76.159756
United States	-0.581617	3.0	22.963575
Brazil	NaN	9.0	NaN
Australia	-1.518550	5.0	10.492336
NaN	NaN	10.0	NaN
NaN	-4.051410	8.0	8.972935
Finland	-1.490149	4.0	9.682963
Netherlands	-3.016659	6.0	8.800447
Denmark	-3.477830	7.0	8.454713