Support de présentation
Projet 2 - Data Scientist

Nordine OURAL

29/12/2022

Problématique

- Entreprise qui propose des formations en ligne niveaux Lycée et Université
- Extension à l'international
- But : faire une analyse exploratoire
- Données issue de la Banque Mondiale des Données

Les données initiales

Dossier zippé de fichiers CSV en libre accès sur Internet

- **EdStatsSeries.csv**: liste des indicateurs utilisés regroupés par étude et par sujet (3665 lignes)
- **EdStatsCountry.csv**: liste des pays et zones inclus dans l'étude avec différentes informations telles que régions géographiques, codes pays... (241 lignes)
- **EdStats.csv**: contient toutes les informations recueillies par indicateur et par pays. Pour chaque ligne, les données peuvent être fournies pour plusieurs années entre 1970 et 2100 (dates à venir pour les projections) (886930 lignes)
- EdStatsCountry-Series.csv: contient des commentaires concernant certains indicateurs pour certains pays (613 lignes)
- EdStatsFootNote.csv : contient une indication quant à la source pour chaque indicateur, chaque pays et chaque année disponible (643638 lignes)

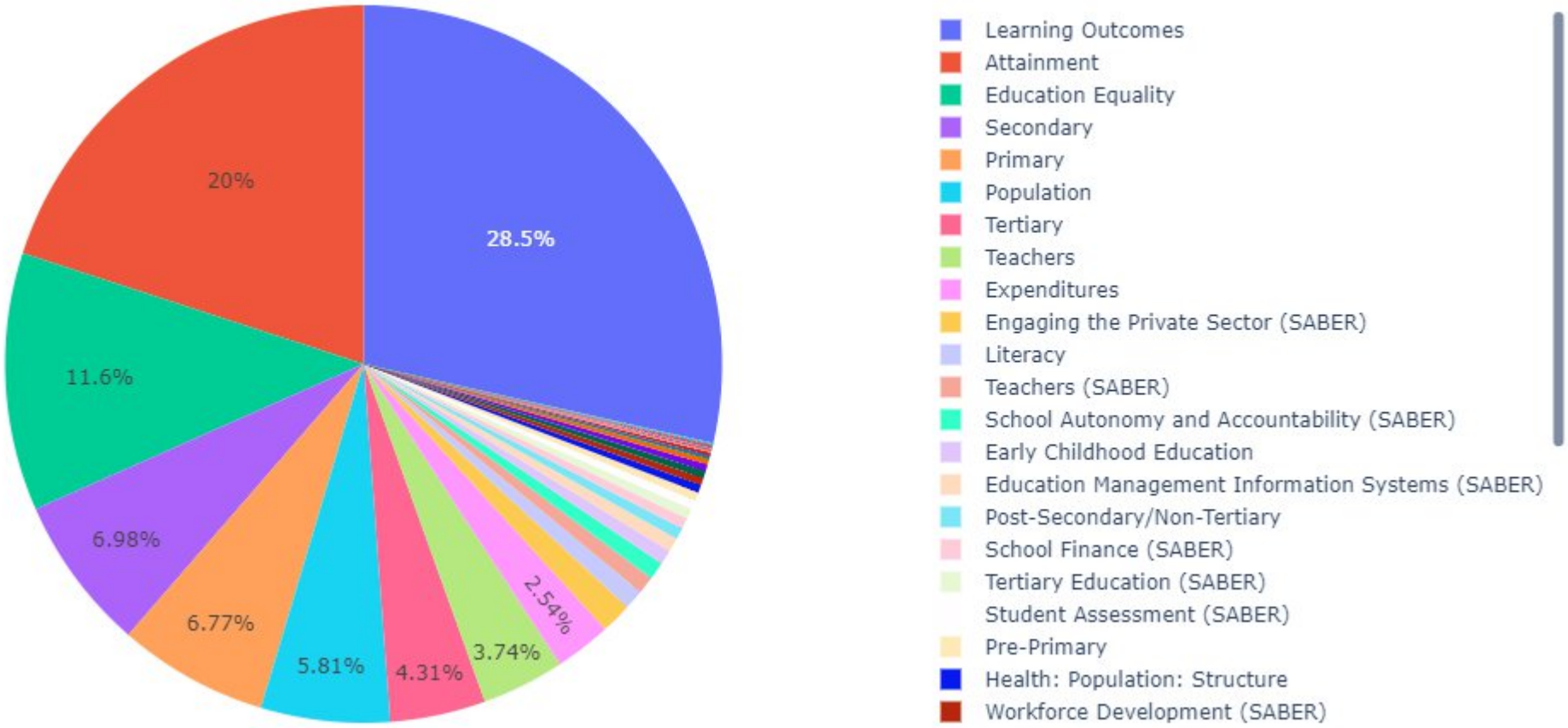
Les données utilisées pour notre étude

- Moyens matériels
  - Nombre de PC pour 100 habitants
  - Nombre d'utilisateurs Internet pour 100 habitants
- Pouvoir d'achat :
  - PIB
- Potentiel du marché
  - Population âgée de 20 ans et plus : cible de marché immédiate
  - Population âgée de moins de 20 ans : cible potentielle

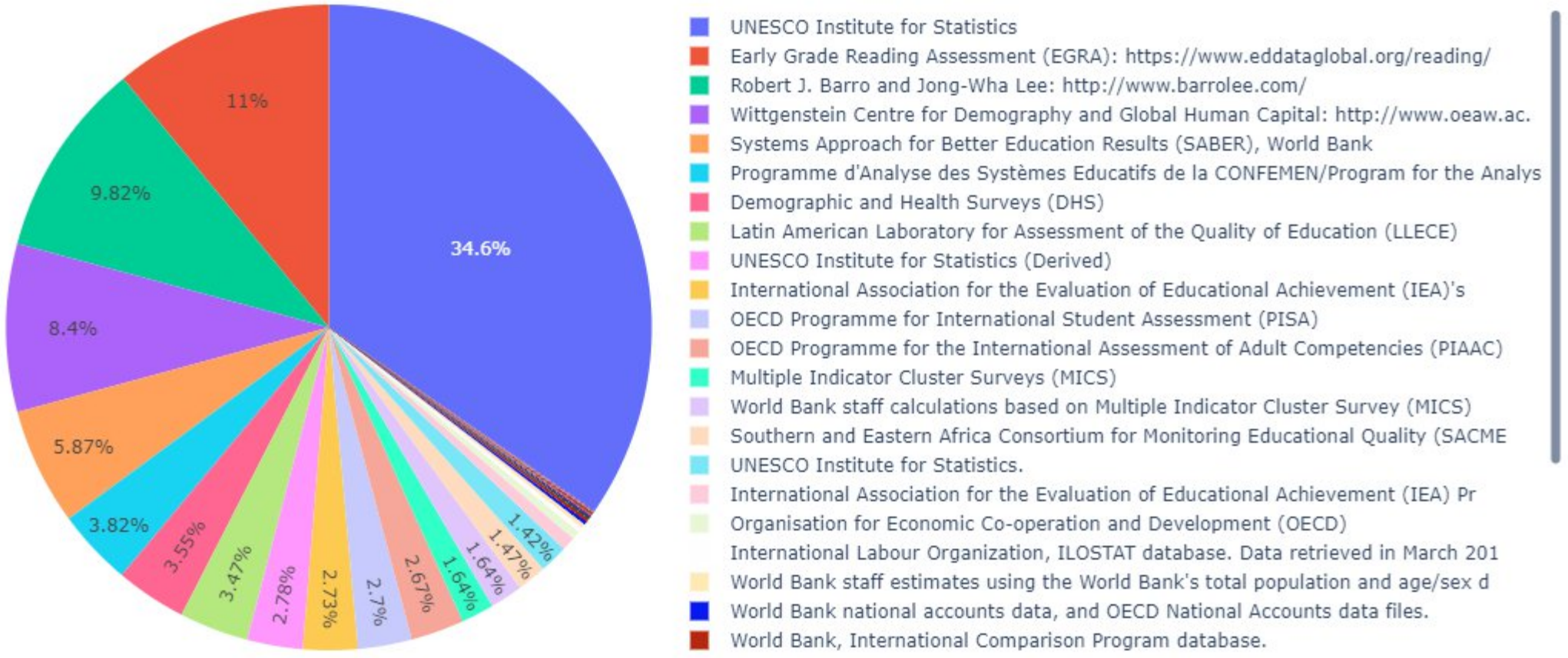
La liste des différents indicateurs

- 3665 indicateurs répartis concernant 37 sujets différents qui qui proviennent de 31 sources (études).
- Exemples de sujets abordés :
  - Résultats d'apprentissage
  - Niveau d' éducation
  - Secondaire
  - Primaire
  - Population
  - ...
- Exemples d'études prises en compte
  - Institut des statistiques de l'Unesco
  - EGRA
  - Barro-Lee
  - PISA
  - ...

Les différents indicateurs par sujet



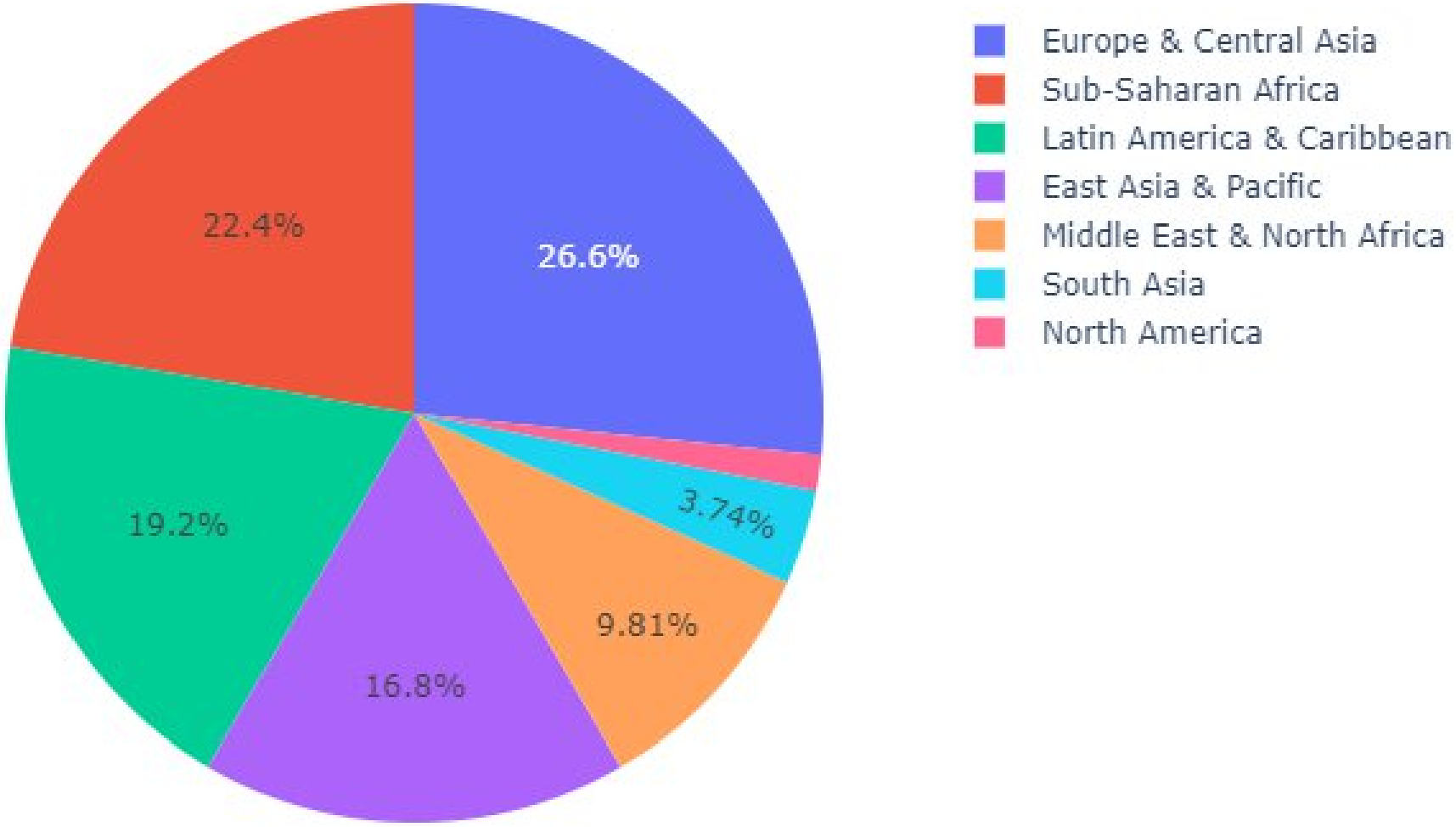
Les différents indicateurs par étude



Les différents pays

- 241 lignes fournies pour les pays
- Parmi ces lignes, nous retrouvons des territoires d'autres pays : Gibraltar, Îles Wallis et Futuna...
- Nous retrouvons aussi des associations de pays : Zone euro, Monde arabe, Pays en développement d'Europe et d'Asie Centrale
- Une colonne région fournit la région géographique à laquelle appartient la ligne

Répartition des pays par région



Les indicateurs

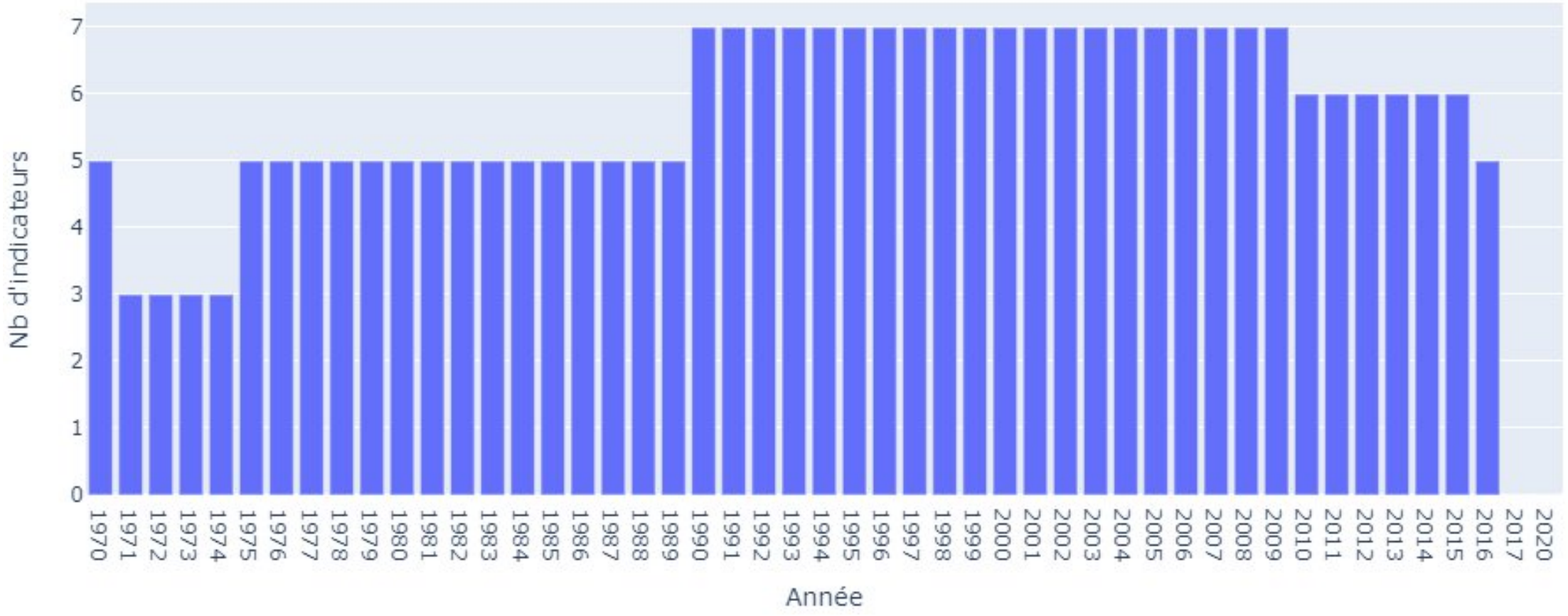
- Pour chaque pays, les 3665 indicateurs listés sont fournis (pas forcement renseignés) pour les années de 1970 à 2017, et de 2020 à 2100 tous les 5 ans (projections)
- Seuls les indicateurs avec les codes suivants sont pris en compte dans notre analyse :
  - IT.CMP.PCMP.P2 : Personal computers (per 100 people)
  - IT.NET.USER.P2 : Internet users (per 100 people)
  - NY.GDP.PCAP.PP.KD : GDP per capita, PPP (constant 2011 international $)
  - NY.GDP.MKTP.KD : GDP (constant 2010 US$)
  - SP.POP.TOTL : Population, total
  - SP.POP.1419.TO.UN : Population, ages 14-19, total
  - SP.POP.0014.TO : Population ages 0-14, total
- Nous ne prendront en compte que les années passées (jusqu'à 2020)

Nettoyage des dataframes

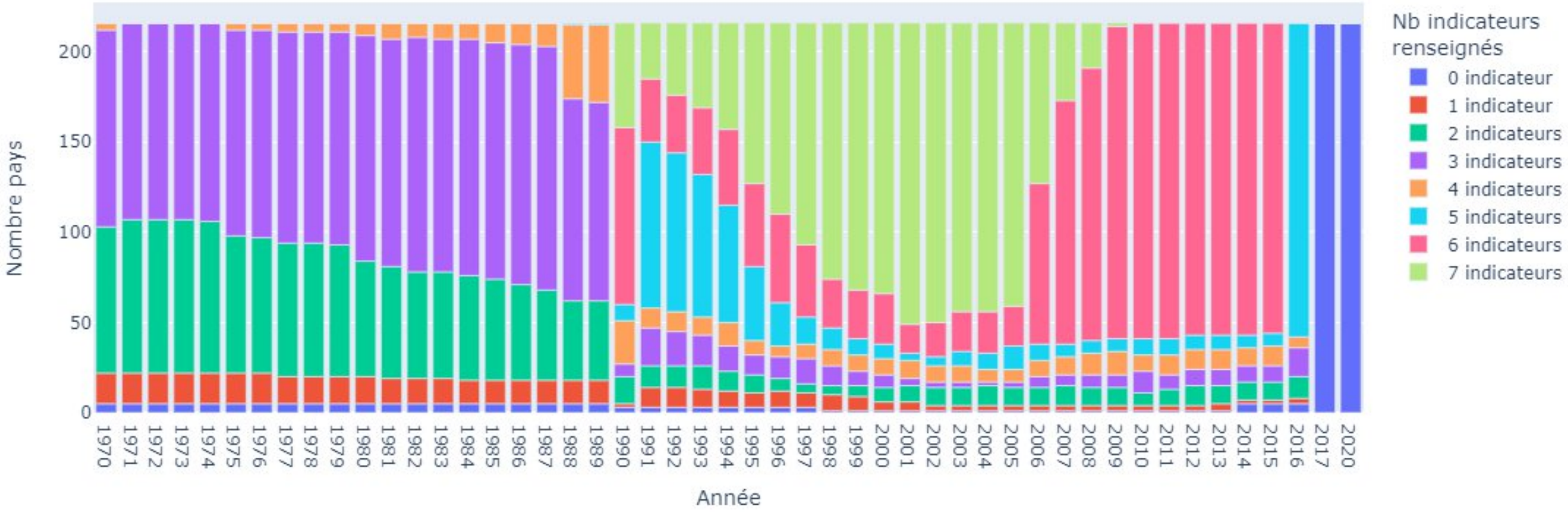
- Dataframe des pays :
  - En utilisant la liste officielle des pays accessible via https://datahub.io/core/country-codes/r/country-codes.csv on filtre les pays réels, via la valeur « 2-alpha code » qui est la colonne « ISO3166-1-Alpha-2 » du dernier fichier cité
  - On constate que 25 « pays » du dataframe initial des pays ont été supprimées

- Dataframe des données statistiques
  - On supprime du dataframe les individus dont le pays n'est plus dans le dataframe des pays

Répartition du nombre maximal d'indicateurs renseignés par année



Répartition du nombre de pays renseignés par année et par nombre d'indicateurs

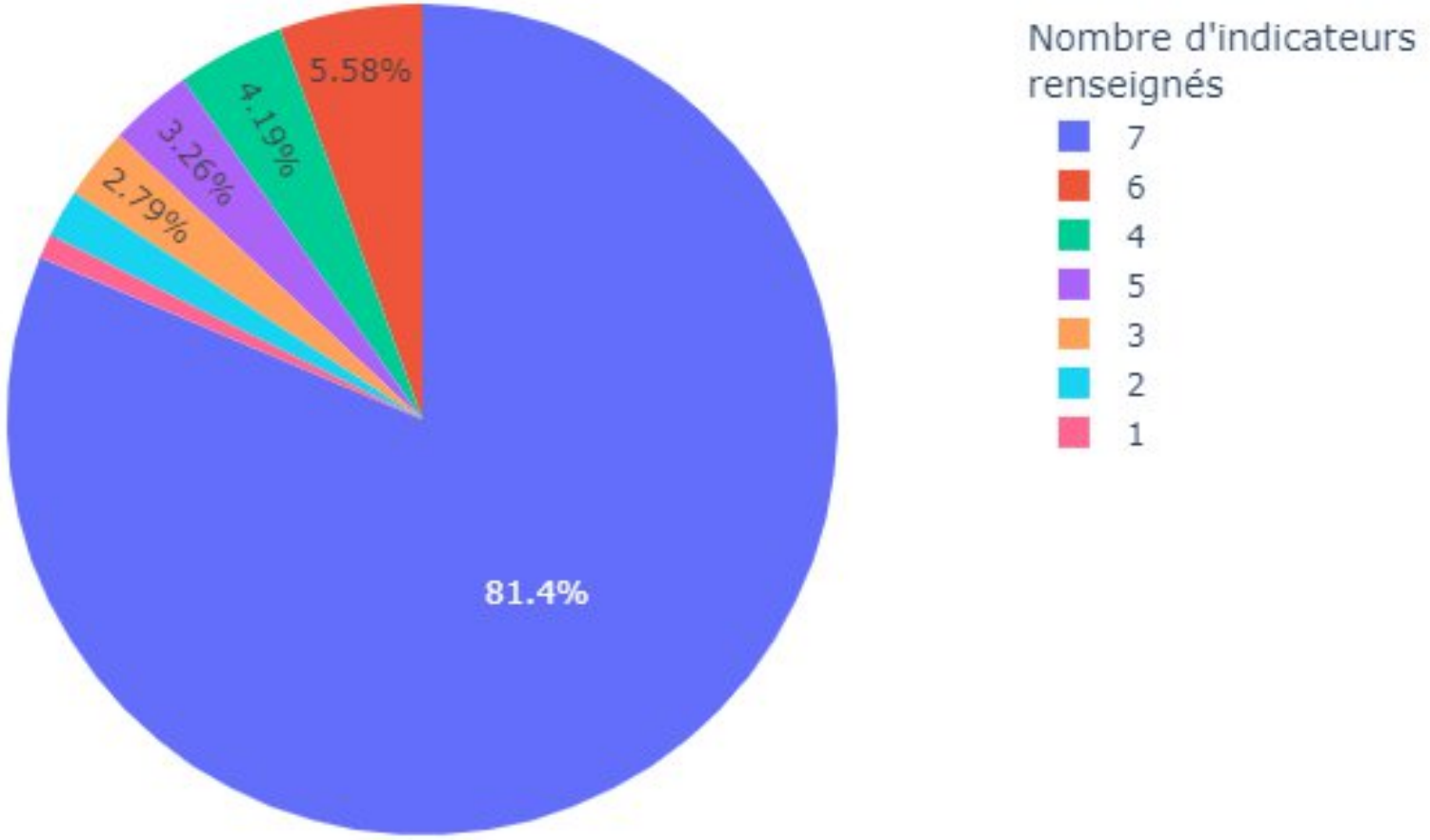


Filtrage des indicateurs par dernière année

- Pour chaque ligne d'indicateur par pays, nous n'allons retenir que la valeur la plus récente :
  - On définit la fonction *get_latest_value* qui fera le travail pour un individu du dataframe
  - On ajoute une valeur « latest_value » au dataframe pour contenir la valeur de l'indicateur la plus récente en appliquant la fonction *get_latest_value* à chaque individu du dataframe
- Pour simplifier le dataframe, on peut alors supprimer toutes les valeurs des années de 1970 à 2100

Filtrage des indicateurs par dernière année (suite)

- Répartition du nombre de pays par nombre d‘indicateurs renseignés



Manipulation du dataframe de travail

- On agrège les différentes lignes de chaque pays en une seule ligne avec 7 valeurs (colonnes) différentes qu'on renommera :
  - % PC Population, % Internet Population
  - PIB/habitant, PIB
  - Population Totale, Population 1419, Population -14
- On ajoute une valeur Region qui correspond à la valeur Region du pays dans le fichier initial EdStatsCountry.csv

On a alors un dataframe dont la structure ressemble à :

	Country Code	Country Name	% PC Population	% Internet Population	PIB/habitant	PIB	Population Totale	Population 1419	Population -14	Region
0	AFG	Afghanistan	0.390148	10.595726	1739.583177	2.066392e+10	34656032.0	4810680.0	15199971.0	South Asia
1	ALB	Albania	4.593538	66.363445	11359.099157	1.347444e+10	2876101.0	312759.0	509714.0	Europe & Central Asia
2	DZA	Algeria	1.030597	42.945527	13921.180022	1.960348e+11	40606052.0	3556170.0	11777882.0	Middle East & North Africa
3	ASM	American Samoa	NaN	0.000000	NaN	5.345551e+08	55599.0	NaN	NaN	East Asia & Pacific
4	AND	Andorra	NaN	97.930637	NaN	3.298477e+09	77281.0	4556.0	NaN	Europe & Central Asia

Nettoyage et imputation du dataframe de travail

- On enlève les pays dont :
  - la population totale n'est pas fournie (1 seul : Nauru)
  - La région n'est pas définie (1seul : Gibraltar)
- Pour les pays dont le PIB/habitant n'est pas renseigné, on le calcule en divisant le PIB par la Population Totale
- Pour les pays dont les populations de -14 ans ou de 14 à 19 ans ne sont pas renseignées, on va émettre l'hypothèse que la population est uniformément répartie sur les âges. On utilise donc les formules :

$$\text{Population -14} = \frac{\text{"Population-1419"} * 14}{6}$$

$$\text{Population 1419} = \frac{\text{"Population-14"} * 6}{14}$$

Nettoyage et imputation du dataframe de travail

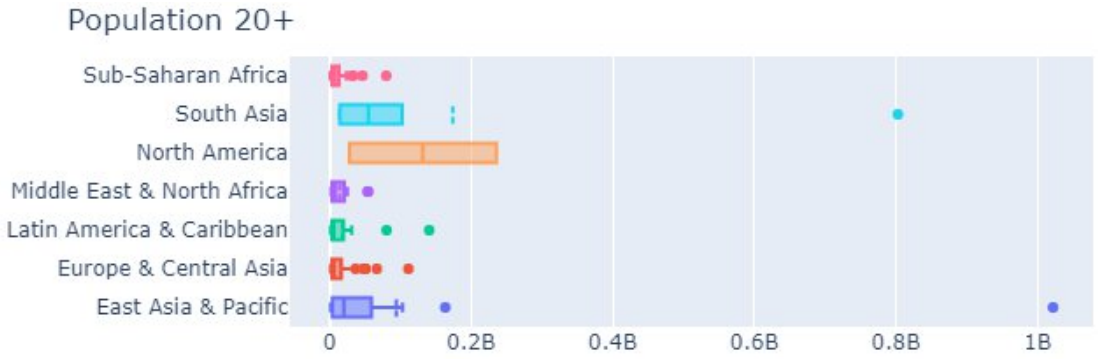
- On supprime tous les pays à « petite population » (< 2 000 000)
- Il reste 7 pays avec des données manquantes : Kazakhstan, Corée du Nord, Liberia, Sierra Leone, Somalie, Soudan du Sud, Syrie. On les supprime du dataframe
- On ajoute une valeur « Population 20+ » calculée de la manière suivante :

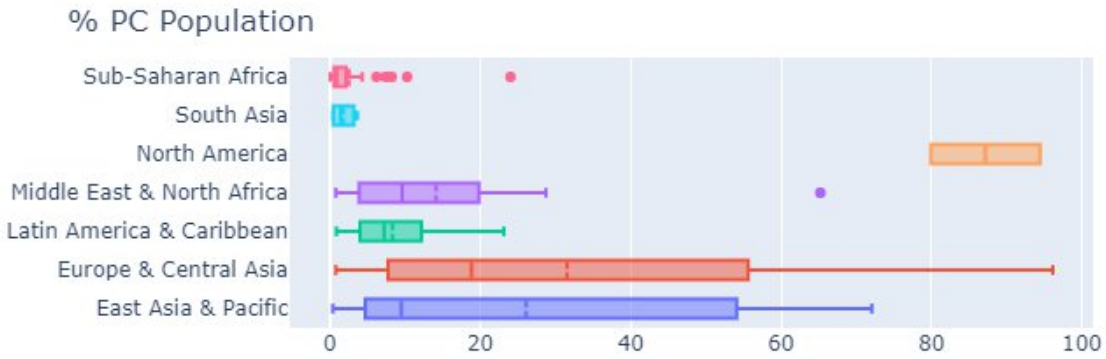
```
"Population 20+" = "Population Totale" - "Population -14" - "Population 1419"
```

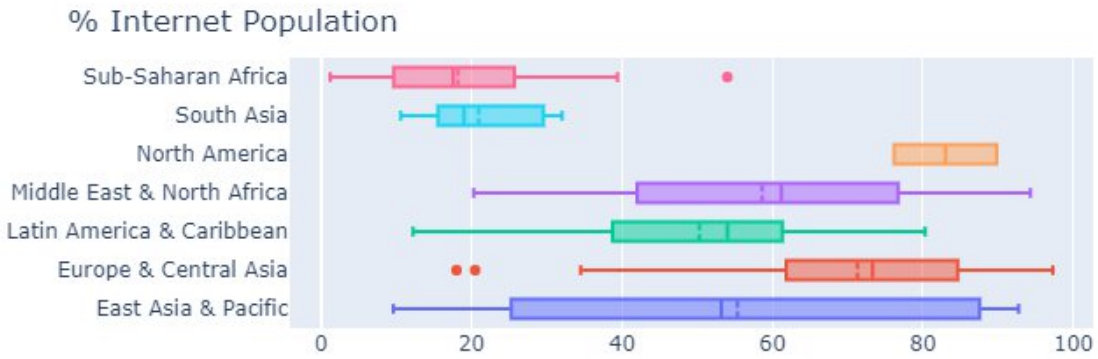
- On supprime les colonnes « Population 1419 » et « Population -14 »

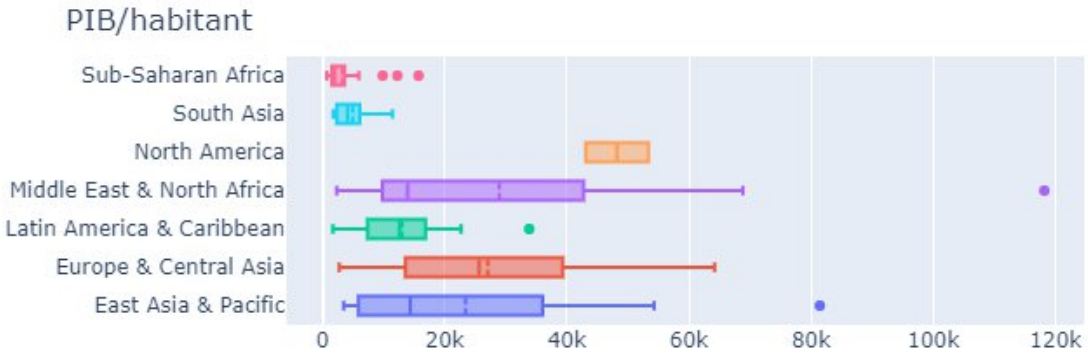
	Country Code	Country Name	% PC Population	% Internet Population	PIB/habitant	Population Totale	Region	Population 20+
0	AFG	Afghanistan	0.390148	10.595726	1739.583177	34656032.0	South Asia	14645381.0
1	ALB	Albania	4.593538	66.363445	11359.099157	2876101.0	Europe & Central Asia	2053628.0
2	DZA	Algeria	1.030597	42.945527	13921.180022	40606052.0	Middle East & North Africa	25272000.0
3	AGO	Angola	0.646019	13.000000	5984.640422	28813463.0	Sub-Saharan Africa	12527965.0
4	ARG	Argentina	9.056130	70.150764	18489.434893	43847430.0	Latin America & Caribbean	28761809.0

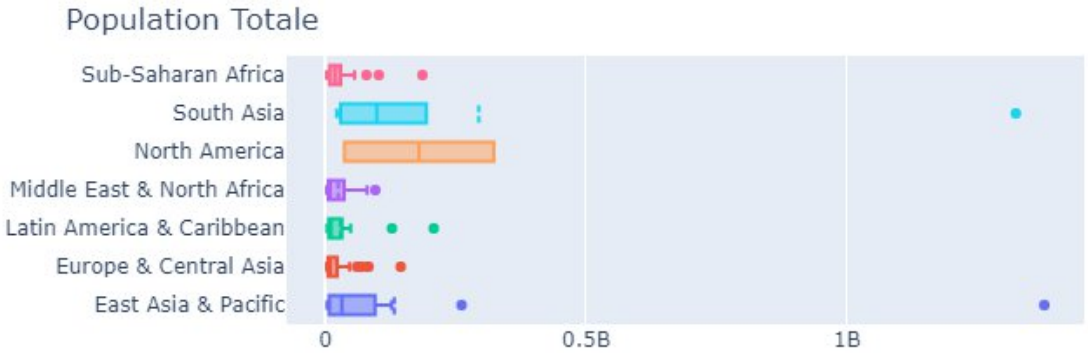
Répartition des indicateurs



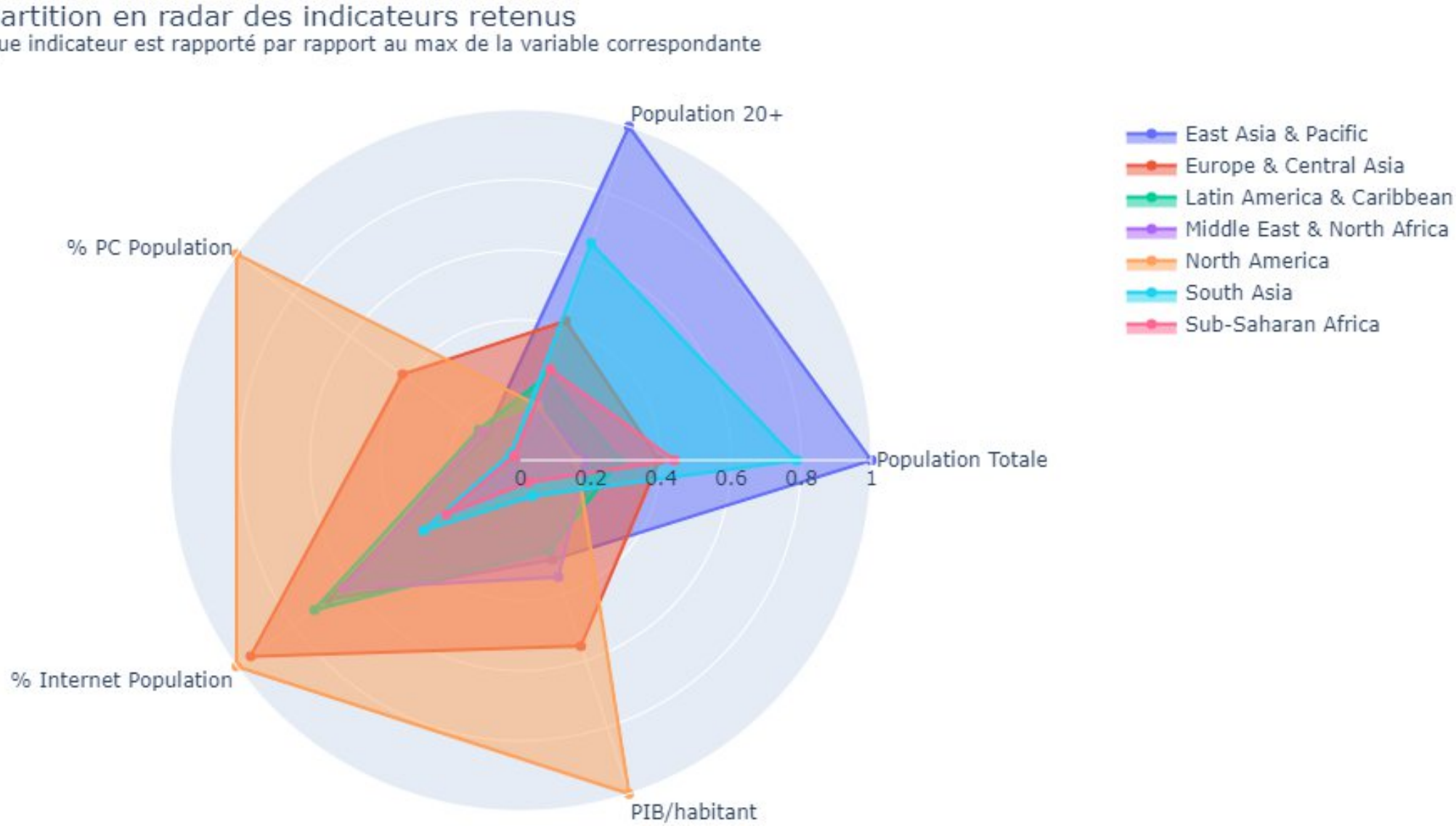








Graphique radar des indicateurs



Création d'un indicateur composite

- Afin de comparer les pays entre-eux, nous allons créer un indicateur composite qui calculera l'indice de confiance donné à chaque individu d'un pays donné:

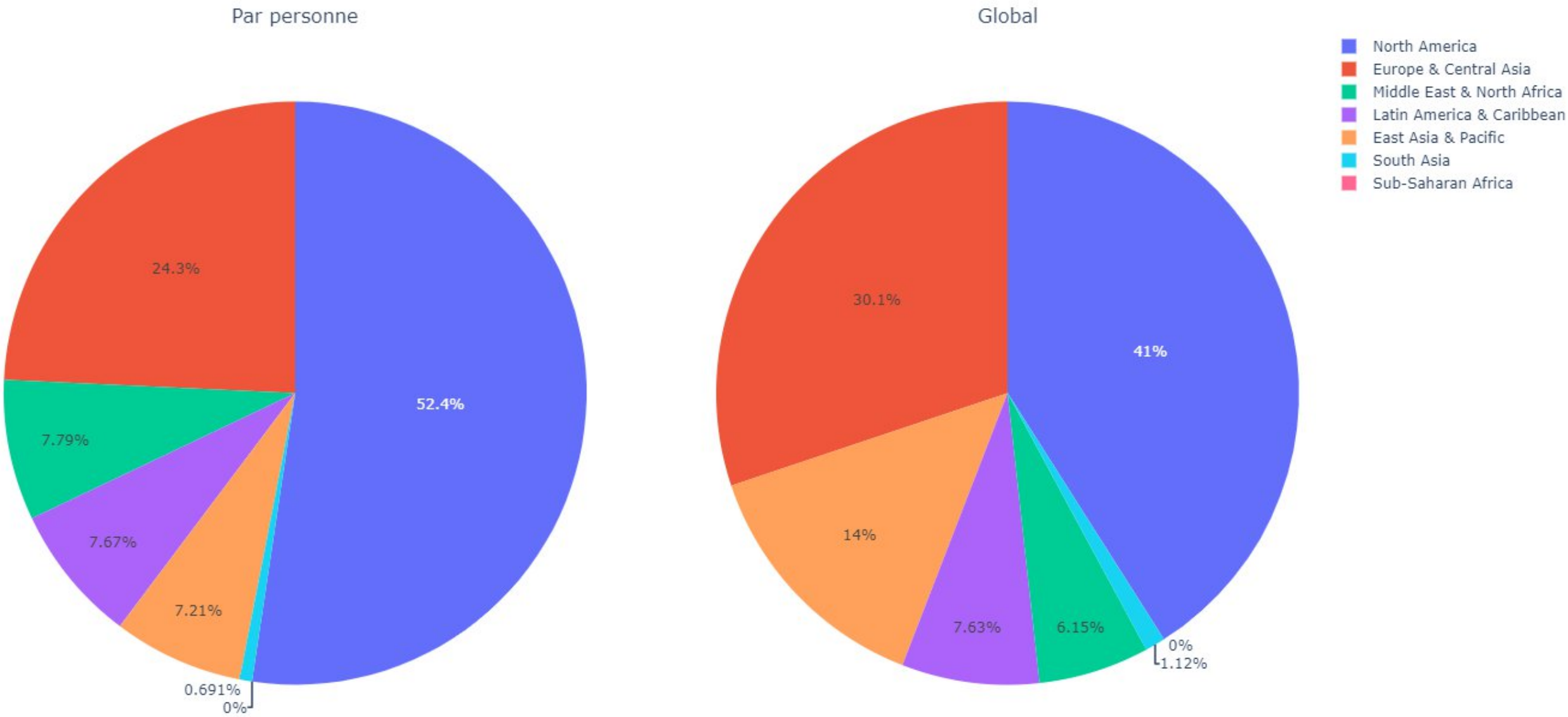
$$\text{conf\_individu} = \text{"\% PC Population"} * \text{"\% Internet Population"} + \text{"\% PC Population"} * 100 * \frac{\text{"}PIB/habitant\text{"}}{max(\text{"}PIB/habitant\text{"})} + \text{"\% Internet Population"} * 100 * \frac{\text{"}PIB/habitant\text{"}}{max(\text{"}PIB/habitant\text{"})}$$

- Un autre indicateur donnera l'indice de confiance associé au pays

$$\text{conf\_globale} = \text{conf\_individu} * \frac{\text{"}PIB/habitant\text{"}}{max(\text{"}PIB/habitant\text{"})}) * \sqrt{(\text{"}Population20+\text{"} + 0.25 * \text{"}Population20-\text{"})}$$

- Pour tenir compte du potentiel à venir de chaque pays, on ajoute à la population cœur de cible (âgée de plus de 20 ans) 1/4 de la population de moins de 20 ans
- On prend la racine carrée de la population ainsi calculée pour limiter l'effet des pays à très grande population

Indices de confiance appliqués aux régions



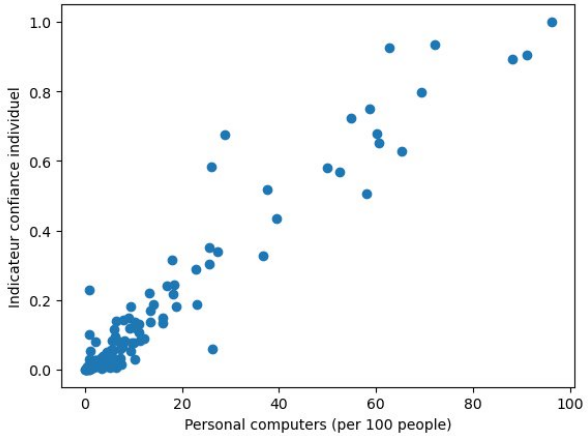
Liste des 20% des pays les plus intéressants

	Country Name	Region	indicateur_confiance_global
0	United States	North America	1.000000
1	United Kingdom	Europe & Central Asia	0.515031
2	Germany	Europe & Central Asia	0.503260
3	Japan	East Asia & Pacific	0.461936
4	Canada	North America	0.422559
5	France	Europe & Central Asia	0.400034
6	Korea, Rep.	East Asia & Pacific	0.352189
7	Netherlands	Europe & Central Asia	0.289948
8	Saudi Arabia	Middle East & North Africa	0.253620
9	Australia	East Asia & Pacific	0.248793
10	Switzerland	Europe & Central Asia	0.224744
11	Spain	Europe & Central Asia	0.223812
12	Sweden	Europe & Central Asia	0.218054
13	China	East Asia & Pacific	0.206029

	Country Name	Region	indicateur_confiance_global
14	Italy	Europe & Central Asia	0.191041
15	Russian Federation	Europe & Central Asia	0.183504
16	Hong Kong SAR, China	East Asia & Pacific	0.166399
17	Singapore	East Asia & Pacific	0.161826
18	Norway	Europe & Central Asia	0.153612
19	Austria	Europe & Central Asia	0.147599
20	United Arab Emirates	Middle East & North Africa	0.143619
21	Brazil	Latin America & Caribbean	0.137099
22	Denmark	Europe & Central Asia	0.128969
23	Belgium	Europe & Central Asia	0.126683
24	Ireland	Europe & Central Asia	0.116267
25	Malaysia	East Asia & Pacific	0.110509
26	Qatar	Middle East & North Africa	0.107545
27	Poland	Europe & Central Asia	0.105343

Conclusion

- Pour notre analyse, nous avons dû supprimer un certain nombre de pays de notre étude, faute de données suffisantes.
- Cependant, les pays enlevés sont pour la majorité de touts petits états, ou alors des états dont le système politique n'est pas favorable à notre activité (Syrie, Soudan du Sud ..)



- En comparant notre indicateur composite par individu avec chacun des indicateurs fournis par les données de départ, on constate une relative corrélation linéaire entre le pourcentage de PC pour 100 habitants par pays et notre indicateur composite individuel.
- On pourrait donc extrapoler notre raisonnement en remplaçant l'indicateur composite par le pourcentage de PC pour 100 habitants afin de d'avoir une estimation pour plus de pays

Merci de m’avoir écouté