Projet 2

Analysez des données de systèmes éducatifs

Parcours Data Scientist

Nordine OURAL

29/12/2022

Mentor : **Amine Hadj-Youcef**

Sommaire

- Problématique
- Environnement technique
- Présentation des données
- Indicateurs retenus
- Manipulation des données
- Conclusion

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

Environnement technique

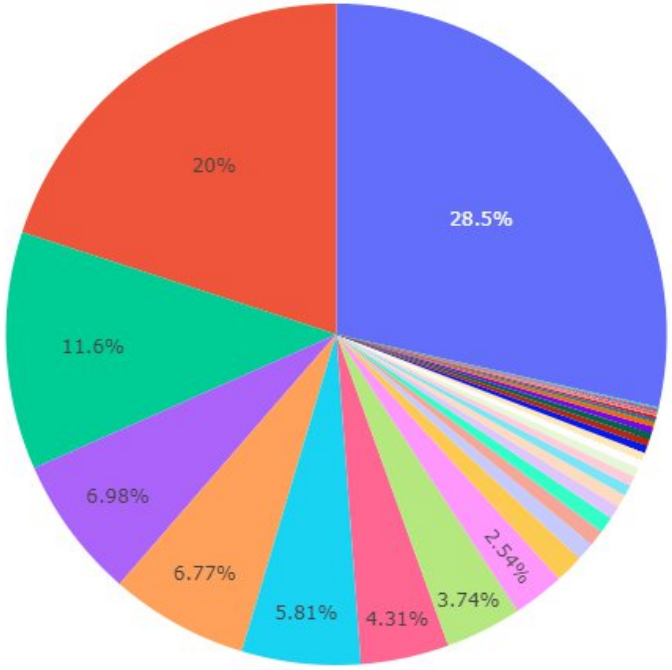
- Frontend :
  - Windows
  - VSCode
  - Extensions Python + Jupyter
- Backend
  - Ubuntu 20 via WSL
  - Python 3.8.10
  - VirtualEnv avec Pandas, Numpy, Plotly, via Pip3 (requirements.txt)
- Contrôle de version :
  - Git sur Github

Présentation des données

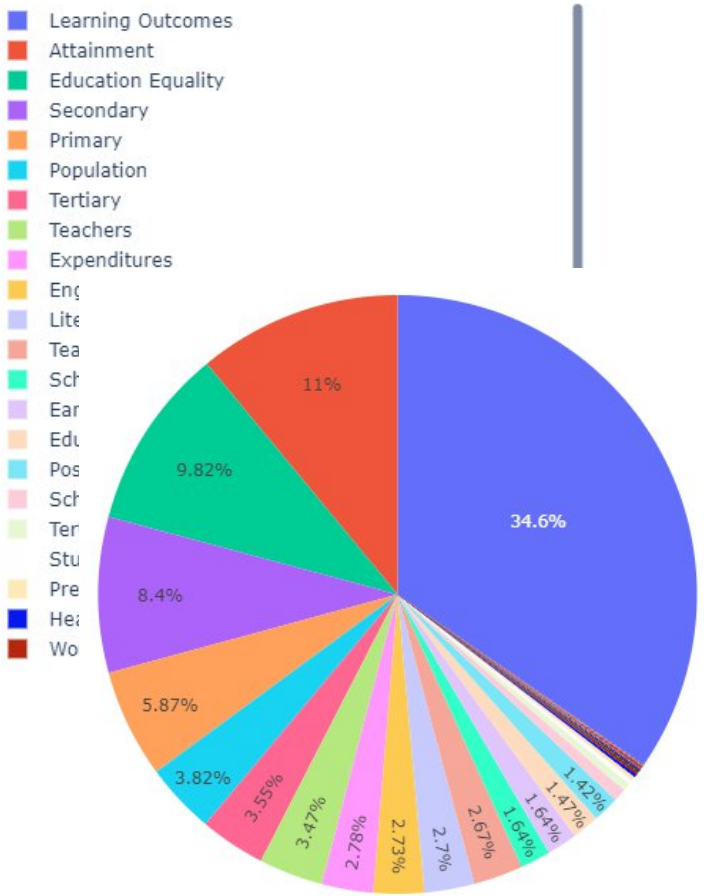
Dossier zippé de fichiers CSV en libre accès sur Internet

- **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)

Répartition des différents indicateurs



Par sujet

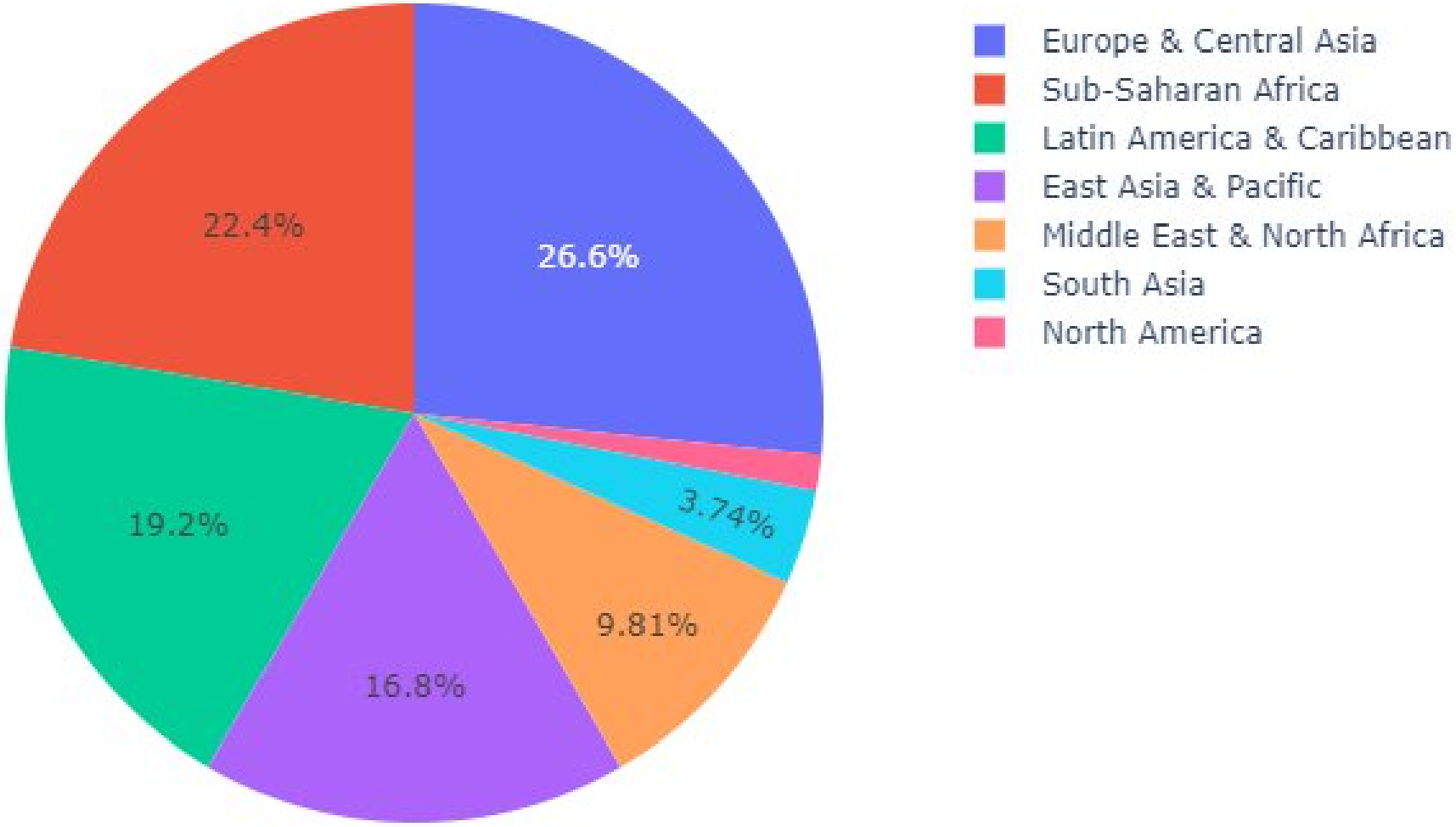


Par étude

Les différents pays

- 241 lignes fournies
- Pays discutables: Gibraltar, Îles Wallis et Futuna...
- Associations de pays : Zone euro, Monde arabe, Pays en développement d’Europe et d’Asie Centrale
- Région géographique fournie

Répartition des pays par région



Les données utilisées pour notre étude

- Moyens matériels
  - PC disponible à la population
  - Accès à Internet
- 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

Les indicateurs pertinents

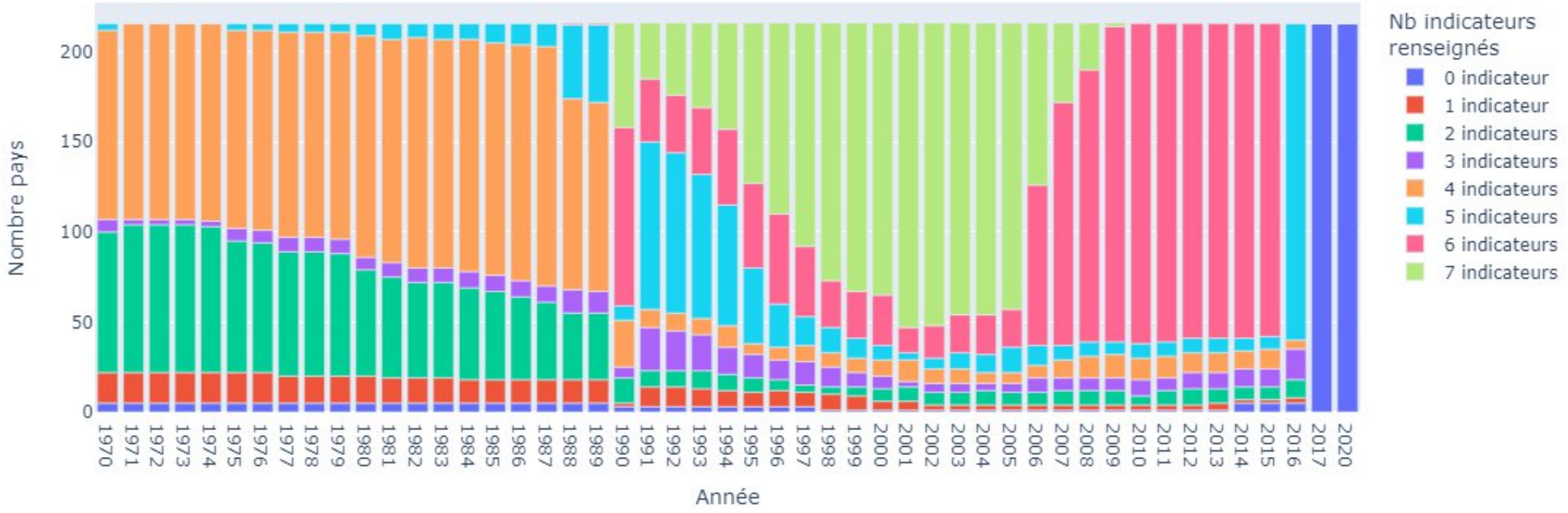
- Indicateurs à retenir:
  - IT.CMP.PCMP.P2 : Personal computers (per 100 people)
  - IT.NET.USER.P2 : Internet users (per 100 people)
  - NY.GDP.PCAP.KD : GDP per capita, PPP (constant 2010 US$)
  - NY.GDP.MKTP.KD : GDP (constant 2010 US$)
  - SP.POP.TOTL : Population, total
  - SP.POP.1419.TO.UN : Population, ages 14-19, total
  - SP.POP.0014.TO : Population ages 0-14, total
- Projections ignorées (2017-2100)

Nettoyage des données

- Données des pays :
  - Croisement avec le fichier https://datahub.io/core/country-codes/r/country-codes.csv pour filtrer les pays réels
  - 25 « pays » supprimés

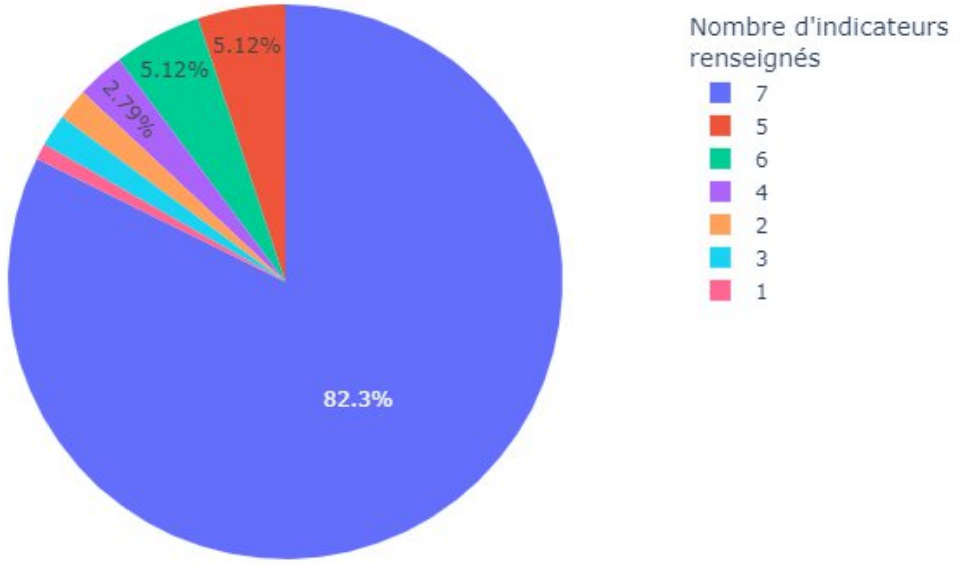
- Données statistiques
  - On supprime les individus dont le pays n'est plus dans la liste des pays

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



Filtrage des indicateurs par dernière année

- Valeur la plus récente retenue pour chaque ligne:
  - Via la fonction *get_latest_value*
  - Résultat « latest_value » ajouté à chaque ligne
- Simplification des données en supprimant toutes les colonnes de 1970 à 2100



Manipulation des données de travail

- Agrégation des lignes par pays avec renommage :
  - % PC Population
  - % Internet Population
  - PIB/habitant
  - PIB
  - Population Totale
  - Population 1419
  - Population -14
- Colonne Region ajoutée

	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	596.257639	2.066392e+10	34656032.0	4810680.0	15199971.0	South Asia
1	ALB	Albania	4.593538	66.363445	4684.967034	1.347444e+10	2876101.0	312759.0	509714.0	Europe & Central Asia
2	DZA	Algeria	1.030597	42.945527	4827.724251	1.960348e+11	40606052.0	3556170.0	11777882.0	Middle East & North Africa
3	ASM	American Samoa	NaN	0.000000	9614.472672	5.345551e+08	55599.0	NaN	NaN	East Asia & Pacific
4	AND	Andorra	NaN	97.930637	42681.603824	3.298477e+09	77281.0	4556.0	NaN	Europe & Central Asia

Nettoyage et imputation des données de travail

- Suppression des pays :
  - sans population totale (1 seul : Nauru)
  - sans région (1seul : Gibraltar)
- Imputation du PIB/habitant avec le PIB et la Population Totale
- Imputation des populations de -14 ans ou de 14 à 19 ans via une répartition uniforme de la population par âge :

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

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

Nettoyage et imputation des données de travail

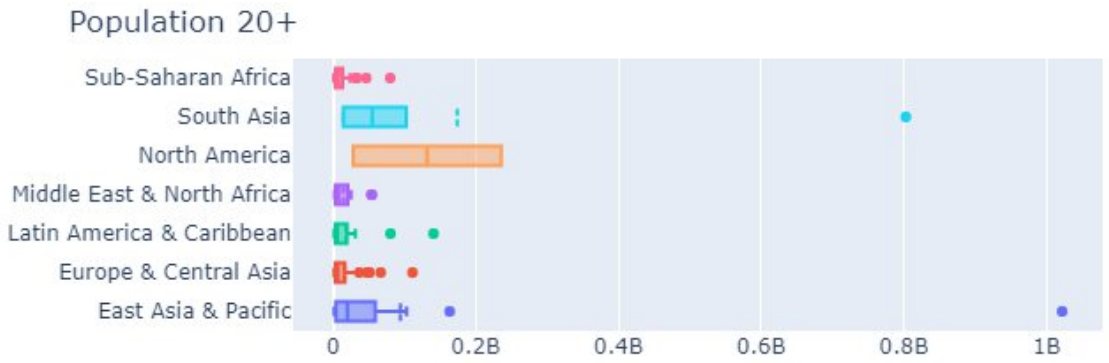
- Suppression des pays à « petite population » (< 2 000 000 habitants)
- Suppression des pays restants avec des données manquantes : Kazakhstan, Corée du Nord, Liberia, Sierra Leone, Somalie, Soudan du Sud, Syrie
- Ajout de la colonne « Population 20+ »:

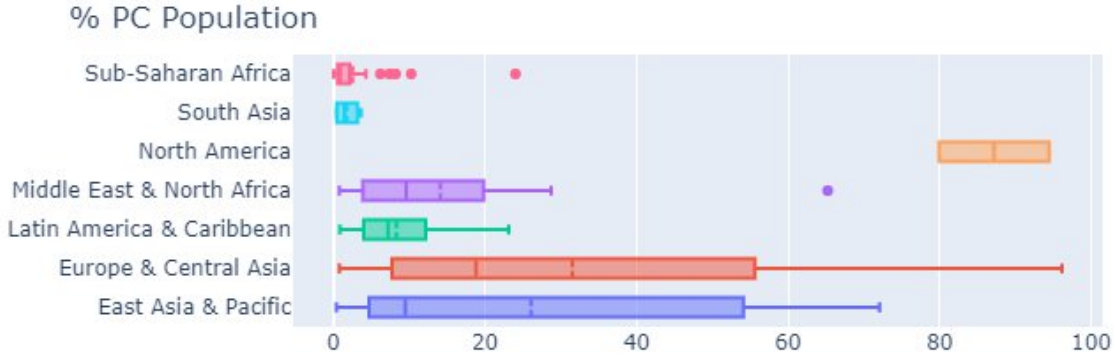
```
"Population 20+" = "Population Totale" - "Population -14" - "Population 1419"
```

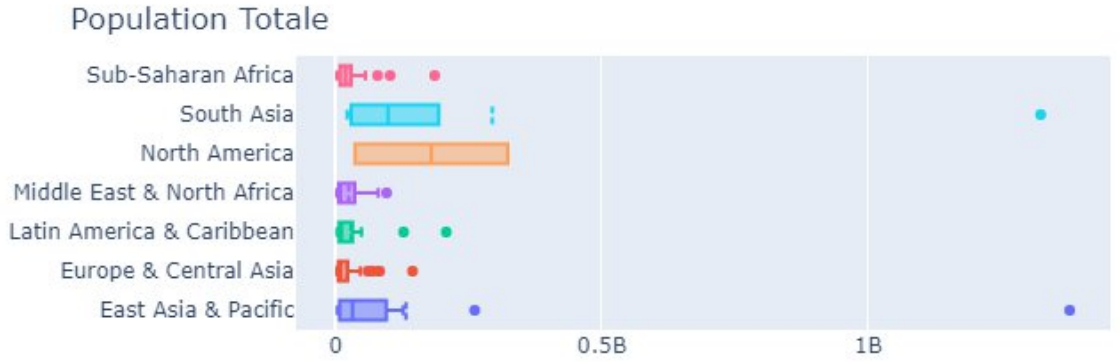
- Suppression des colonnes « Population 1419 » et « Population -14 »

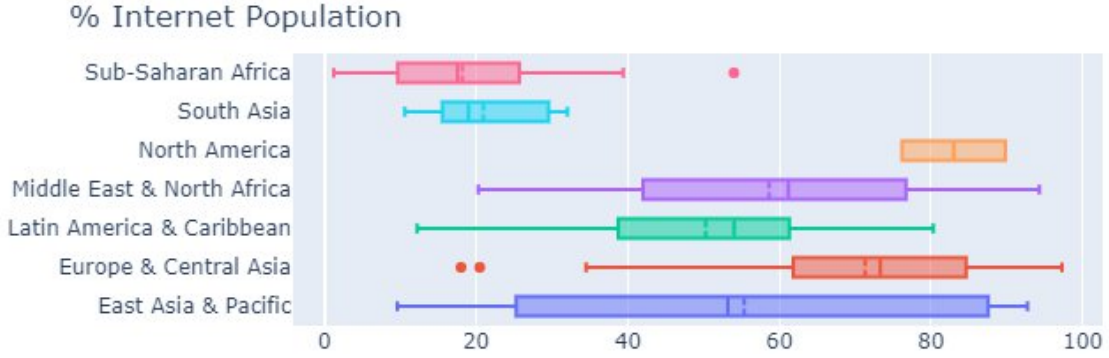
	Country Code	Country Name	% PC Population	% Internet Population	PIB/habitant	Population Totale	Region	Population 20+
0	AFG	Afghanistan	0.390148	10.595726	1739.583177	34656032.0	South Asia	14645381.0
1	ALB	Albania	4.593538	66.363445	11359.099157	2876101.0	Europe & Central Asia	2053628.0
2	DZA	Algeria	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 par région

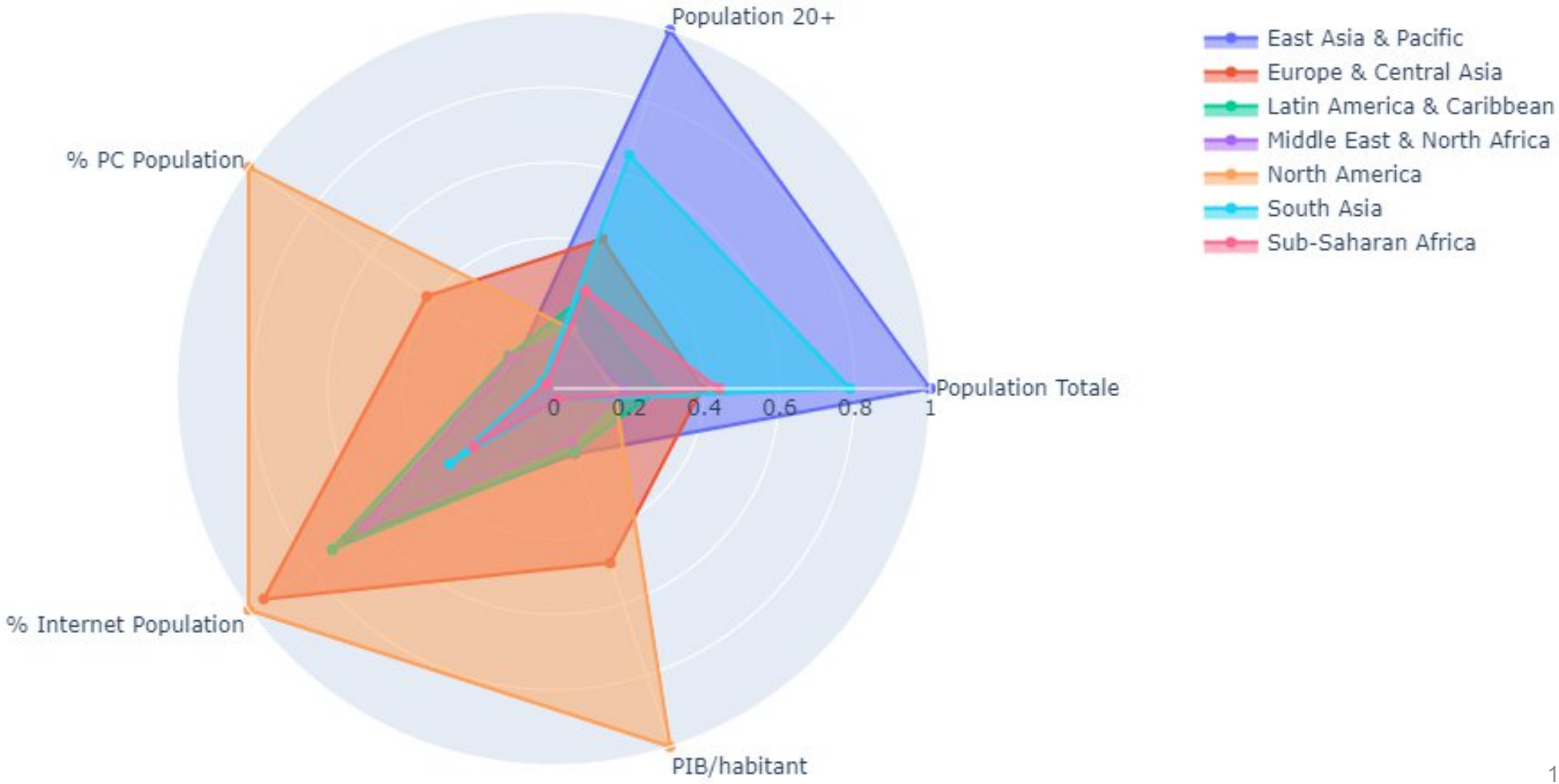








Graphique radar des indicateurs par région



Création d’ indicateurs composites

- Pour comparer les pays, création d’un indicateur composite pour l’indice de confiance donné à chaque individu d’un pays

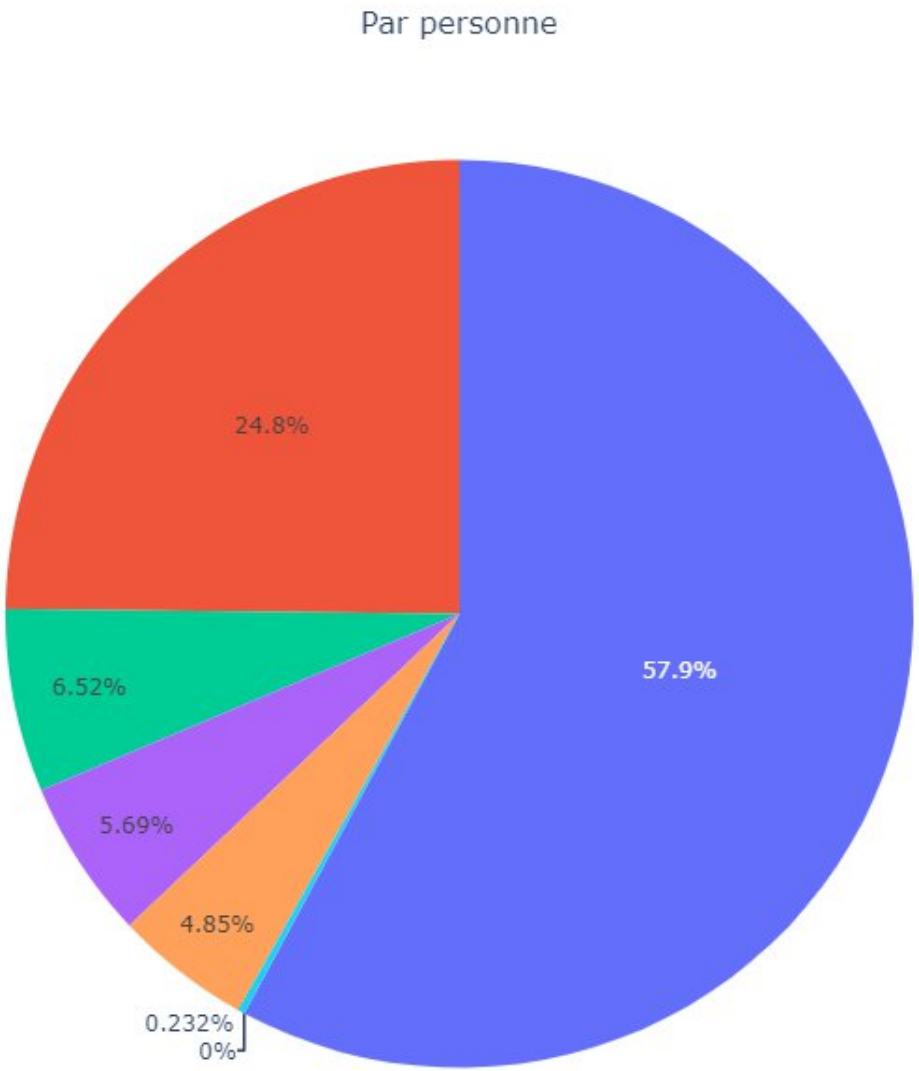
conf_individu = "% PC Population" * "% Internet Population" + "% PC Population" * 100 * $\frac{"PIB/habitant"}{max("PIB/habitant")}$ +

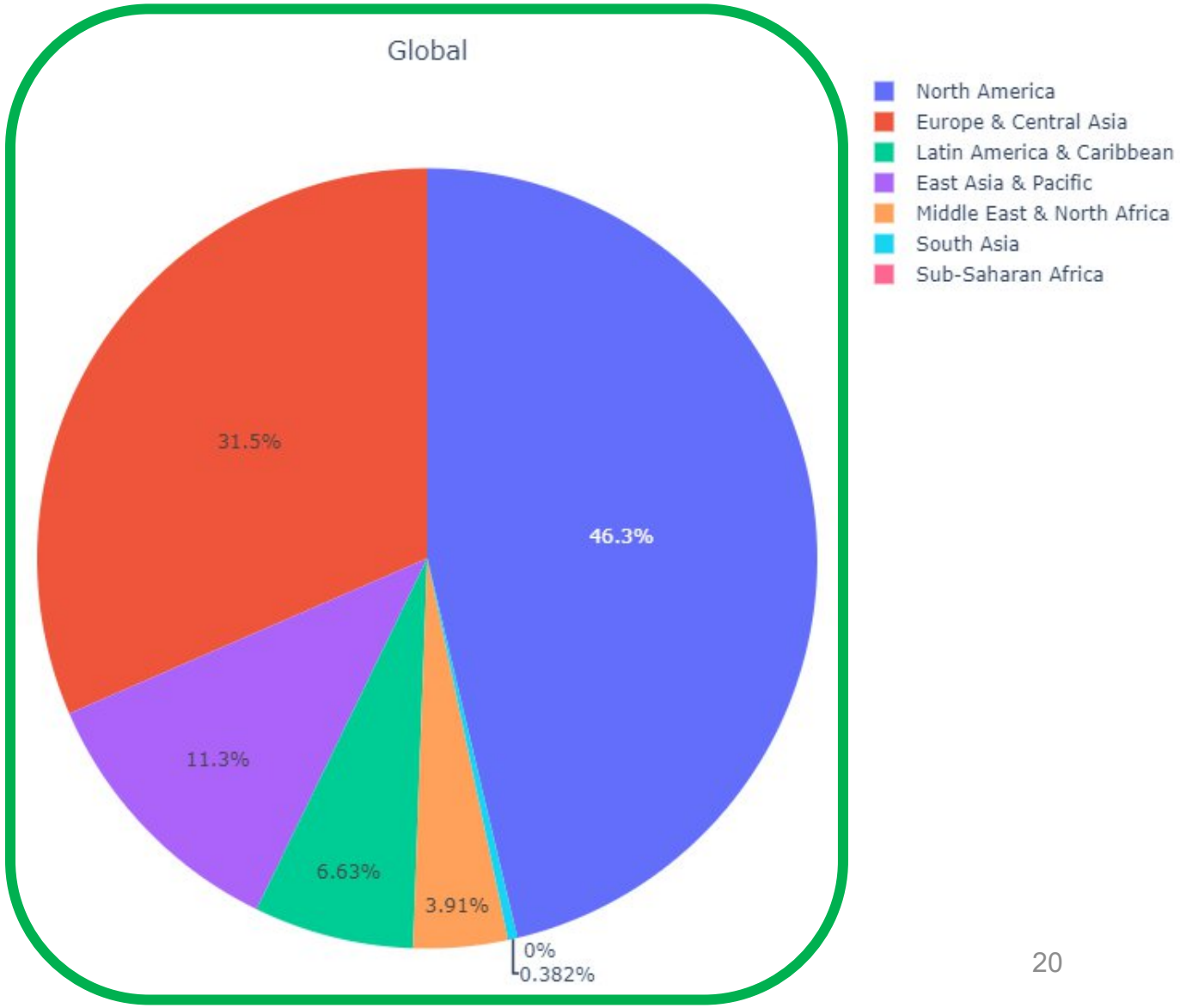
"% Internet Population" * 100 * $\frac{"PIB/habitant"}{max("PIB/habitant")}$

- Création d’un indicateur donnera l’indice de confiance associé au pays

conf_globale = conf_individu * $\sqrt{("Population20+" + 0.25 * "Population20-")}$

Conclusion : classement des régions





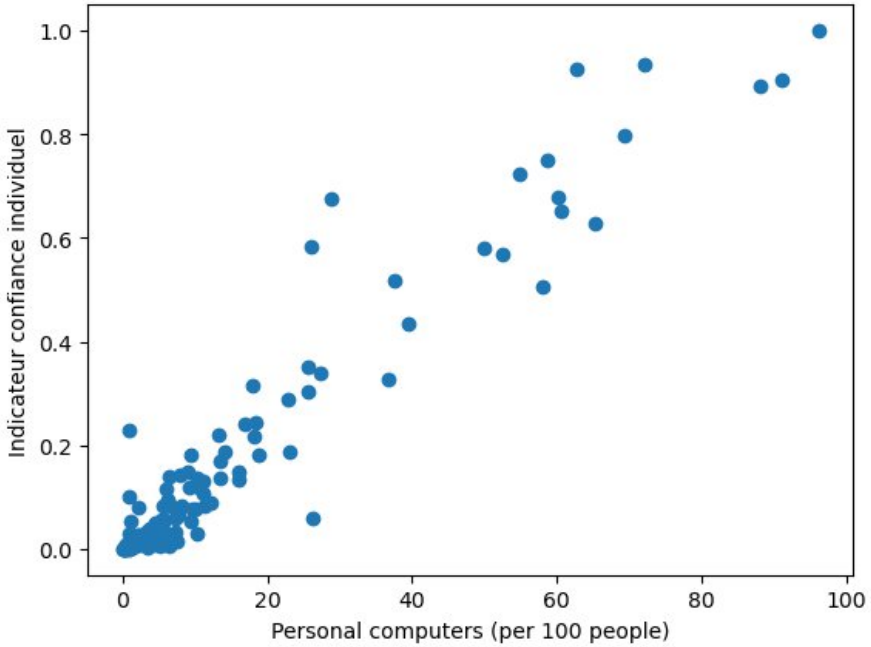
Conclusion : classement des pays (20 premiers %)

	Country Name	Region	indicateur_confiance_global
0	United States	North America	1.000000
1	Japan	East Asia & Pacific	0.535731
2	United Kingdom	Europe & Central Asia	0.522266
3	Germany	Europe & Central Asia	0.515726
4	Canada	North America	0.439742
5	France	Europe & Central Asia	0.416851
6	Korea, Rep.	East Asia & Pacific	0.314610
7	Netherlands	Europe & Central Asia	0.299509
8	Australia	East Asia & Pacific	0.281716
9	Switzerland	Europe & Central Asia	0.255266
10	Sweden	Europe & Central Asia	0.234202
11	Spain	Europe & Central Asia	0.223430
12	Italy	Europe & Central Asia	0.196210
13	Norway	Europe & Central Asia	0.193902

	Country Name	Region	indicateur_confiance_global
14	Saudi Arabia	Middle East & North Africa	0.189670
15	Austria	Europe & Central Asia	0.155141
16	Denmark	Europe & Central Asia	0.152544
17	Hong Kong SAR, China	East Asia & Pacific	0.144080
18	Belgium	Europe & Central Asia	0.138309
19	China	East Asia & Pacific	0.137828
20	Singapore	East Asia & Pacific	0.135808
21	Ireland	Europe & Central Asia	0.128018
22	Russian Federation	Europe & Central Asia	0.126286
23	Brazil	Latin America & Caribbean	0.125613
24	United Arab Emirates	Middle East & North Africa	0.112821
25	Finland	Europe & Central Asia	0.112052
26	New Zealand	East Asia & Pacific	0.094739
27	Poland	Europe & Central Asia	0.082716

Prospective

- Indicateur confiance individuel et %PC Population : relation linéaire



- Extrapolation envisageable avec le pourcentage de PC pour 100 habitants : classement pour plus de pays

Merci de m’avoir écouté