OPENCLASSROOMS CentraleSupélec

# Analysez des données de systèmes éducatifs

Formation Data Scientist - Janvier / Novembre 2022


Auteur: Eric TREGOAT

Mentor: Benjamin TARDY

Evaluateur: Yannick Serge OBAM AKOU

# Ordre du jour de la soutenance

- 5 min - Rappel de la problématique et présentation du jeu de données
  - Mission et démarche
  - Présentation du jeu de données

- 15 min - Présentation de l'analyse pré-exploratoire du jeu de données et vos conclusions sur la pertinence de l'usage du jeu de données pour répondre aux questions stratégiques que se pose l'entreprise
  - Présélection des indicateurs
  - Examen de la qualité des données de chaque indicateur
  - Constitution du jeu de données et jeu alternatif avec projection à 2020
  - Analyse par variable
  - Analyse combinée entre plusieurs variables (2)
  - Proposition de score d'attractivité par pays pour le jeu de base et le jeu alternatif (2)
  - Conclusions

- 5 à 10 minutes de questions-réponses

# Mission et démarche

- Mission:
  Nettoyer et analyser les données de la Banque mondiale concernant l'éducation, afin de réaliser une analyse pré-exploratoire dans l'objectif d'identifier les pays les plus adaptés à au développement commercial d'une startup de la EdTech.

- Démarche:

  - **Etape 1 - Nettoyage des données**
    - Prise de connaissance du jeu de données
    - Présélection des indicateurs
    - Extraction des données et examen de la qualité des données pour chaque indicateur
    - Construction d'un indicateur complémentaire combinant 16 indicateurs et un paramétrage commercial
    - Constitution du jeu de données et d'un jeu alternatif avec projection à 2020
  - **Etape 2 - Analyse des données**
    - Analyse par variable
    - Analyse combinée de plusieurs variables
    - Score d'attractivité par pays
    - Comparaison avec le jeu alternatif
  - **Conclusions**

# Présentation du jeu de données

5 fichier CSV:

1. Données **pays**
2. **Données** pour les indicateurs d'éducation
3. Liste des **indicateurs** et leur définition
4. **Précisions** sur certains indicateurs de population et PIB
5. **Précisions** sur certains indicateurs et pour certains pays

1. Données pays
   - 241 pays répartis sur 7 régions, le « country code » est une clé
   - 26 indicateurs avec beaucoup de manquants
   - Variable intéressante: niveau de revenu (11% de manquants)
2. Données pour les indicateurs d'éducation
   - 886 930 lignes pour 70 colonnes
   - La clé pays figure en colonne
   - Les indicateurs sont tous dans la même colonne, une ligne par pays et par indicateur, identifiés par une clé
   - Les valeurs des indicateurs sont données par année, de 1970 à 2100 (66 colonnes), avec un taux de manquant très élevé (45 années avec plus de 90% de manquants).
     → Investigation à faire pour la bonne sélection des (indicateur, année)
3. Liste des indicateurs et leur définition
   - 3 665 indicateurs, identifié par un code clé
   - Définition et source des données
4. Précisions sur certains indicateurs de population et PIB
   → non utilisé à ce stade
5. Précisions sur certains indicateurs et pour certains pays
   → non utilisé à ce stade

# Présélection des indicateurs

- Identification du **pays**
  - Code et nom de pays
  - Région

- **Accessibilité** au marché
  - Internet
  - Capacité de financement
  - Langage

- **Volume de marché**: population cible

$$\sum_{I=2024}^{I=5559} Coef_I * Vol\ Pop_I * \%\ PopSec_I$$

| Paramètre $Coef_I$ | |
|---|---|
| Tranche d'ages 2024 | 100% |
| Tranche d'ages 2529 | 100% |
| Tranche d'ages 3034 | 50% |
| Tranche d'ages 3539 | 25% |
| Tranche d'ages 4044 | 50% |
| Tranche d'ages 4549 | 10% |
| Tranche d'ages 5054 | 5% |
| Tranche d'ages 5559 | 2.5% |

| | | | |
|---|---|---|---|
| Pays | Country Code | | 1 |
| | Country name | | 2 |
| Internet | % d'utilisateurs | | 3 |
| Population | Volume de population | Tranche d'ages 2024 | 4 |
| | | Tranche d'ages 2529 | 5 |
| | | Tranche d'ages 3034 | 6 |
| | | Tranche d'ages 3539 | 7 |
| | | Tranche d'ages 4044 | 8 |
| | | Tranche d'ages 4549 | 9 |
| | | Tranche d'ages 5054 | 10 |
| | | Tranche d'ages 5559 | 11 |
| | % de population ayant terminé ses études secondaires | Tranche d'ages 2024 | 12 |
| | | Tranche d'ages 2529 | 13 |
| | | Tranche d'ages 3034 | 14 |
| | | Tranche d'ages 3539 | 15 |
| | | Tranche d'ages 4044 | 16 |
| | | Tranche d'ages 4549 | 17 |
| | | Tranche d'ages 5054 | 18 |
| | | Tranche d'ages 5559 | 19 |
| Langage | Français | | 20 |
| | Anglais | | 21 |
| Capacité de financement | PIB du pays | | 22 |
| | PIB par habitant | | 23 |
| | Niveau de revenus | | 24 |
| | % des dépenses publiques consacrées à l'édutation post-secondaire | | 25 |

# Examen de la qualité des données de chaque indicateur

- Examen global des manquants pour les années 2006 à 2020: on réduit à **2010 à 2016**

- Extraction des indicateurs présélectionnés et examen du taux de manquant de 2010 à 2016:
  - ☑ Données **internet**: année **2016**
  - ☐ Données de **capacité de financement**: année **2016**
  - ☒ Données de **langage**: **abandon**, pas de données
  - ☑ Données de **population**: année **2010**

- Examen des indicateurs de capacité de financement
  - ☒ Indicateur du **taux de dépenses publiques pour l'éducation au-delà du secondaire**: 66% de manquants, **abandon**
  - ☑ Les indicateurs de **PIB pays** et **PIB par habitant** sont bien renseignés, et les indicateur de **PIB par habitant** et de **catégorie de revenus** sont cohérents : conservation de l'indicateur de **PIB par habitant** pour **2016**

- Option d'actualisation des données par projection à 2020
  - Utile pour les données les plus susceptibles d'évolution: **indicateurs internet et PIB par habitant**
  - Projections à 2018 et 2020 par régression linéaire sur les données de 2006 à 2016, pour retenir l'année **2020**

# Jeu de données et jeu alternatif avec projection à 2020

- 241 → **140 pays**
- Extraits du **jeu de données sélectionné** en conservant les 3 pays de plus forte population cible par région
- A droite le **jeu de données avec projection** des variables **Internet** et **PIB par habitant** (les valeurs surlignées identifient un excès de projection)

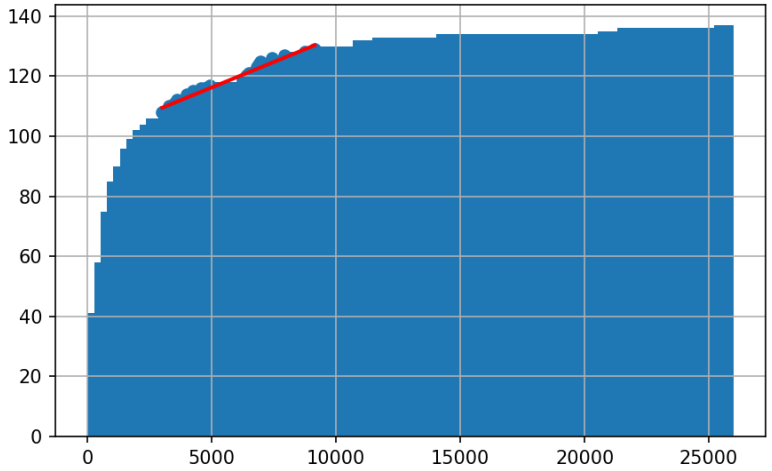
| Country Code | Region | Pays | Internet | Population cible | PIB par habitant |
|---|---|---|---|---|---|
| CHN | East Asia & Pacific | China | 53 | 102 784 | 15 559 |
| IDN | East Asia & Pacific | Indonesia | 25 | 21 360 | 11 632 |
| JPN | East Asia & Pacific | Japan | 92 | 10 870 | 41 476 |
| DEU | Europe & Central Asia | Germany | 90 | 11 648 | 48 885 |
| RUS | Europe & Central Asia | Russian Federation | 76 | 8 724 | 23 163 |
| GBR | Europe & Central Asia | United Kingdom | 95 | 7 430 | 43 081 |
| BRA | Latin America & Caribbean | Brazil | 60 | 20 782 | 15 153 |
| MEX | Latin America & Caribbean | Mexico | 60 | 6 813 | 17 877 |
| ARG | Latin America & Caribbean | Argentina | 70 | 4 572 | 19 979 |
| EGY | Middle East & North Africa | Egypt, Arab Rep. | 39 | 7 932 | 11 150 |
| IRN | Middle East & North Africa | Iran, Islamic Rep. | 53 | 6 967 | 19 988 |
| SAU | Middle East & North Africa | Saudi Arabia | 74 | 3 284 | 54 522 |
| USA | North America | United States | 76 | 25 276 | 57 638 |
| CAN | North America | Canada | 90 | 2 904 | 44 025 |
| IND | South Asia | India | 30 | 101 239 | 6 583 |
| BGD | South Asia | Bangladesh | 18 | 14 173 | 3 587 |
| PAK | South Asia | Pakistan | 16 | 10 685 | 5 246 |
| ZAF | Sub-Saharan Africa | South Africa | 54 | 9 139 | 13 248 |
| COD | Sub-Saharan Africa | Congo, Dem. Rep. | 6 | 2 286 | 804 |
| KEN | Sub-Saharan Africa | Kenya | 26 | 1 933 | 3 161 |

| Country Code | Region | Pays | Internet | Population cible | PIB par habitant |
|---|---|---|---|---|---|
| CHN | East Asia & Pacific | China | 74 | 102 784 | 19 209 |
| IDN | East Asia & Pacific | Indonesia | 31 | 21 360 | 13 576 |
| JPN | East Asia & Pacific | Japan | 100 | 10 870 | 44 451 |
| DEU | Europe & Central Asia | Germany | 96 | 11 648 | 55 650 |
| RUS | Europe & Central Asia | Russian Federation | 100 | 8 724 | 30 090 |
| GBR | Europe & Central Asia | United Kingdom | 100 | 7 430 | 45 688 |
| BRA | Latin America & Caribbean | Brazil | 74 | 20 782 | 18 123 |
| MEX | Latin America & Caribbean | Mexico | 74 | 6 813 | 19 902 |
| ARG | Latin America & Caribbean | Argentina | 96 | 4 572 | 22 860 |
| EGY | Middle East & North Africa | Egypt, Arab Rep. | 49 | 7 932 | 12 399 |
| IRN | Middle East & North Africa | Iran, Islamic Rep. | 64 | 6 967 | 20 052 |
| SAU | Middle East & North Africa | Saudi Arabia | 96 | 3 284 | 60 788 |
| USA | North America | United States | 76 | 25 276 | 61 054 |
| CAN | North America | Canada | 98 | 2 904 | 47 847 |
| IND | South Asia | India | 37 | 101 239 | 7 746 |
| BGD | South Asia | Bangladesh | 21 | 14 173 | 4 152 |
| PAK | South Asia | Pakistan | 18 | 10 685 | 5 602 |
| ZAF | Sub-Saharan Africa | South Africa | 81 | 9 139 | 14 572 |
| COD | Sub-Saharan Africa | Congo, Dem. Rep. | 6 | 2 286 | 931 |
| KEN | Sub-Saharan Africa | Kenya | 30 | 1 933 | 3 522 |

# Analyse par variable

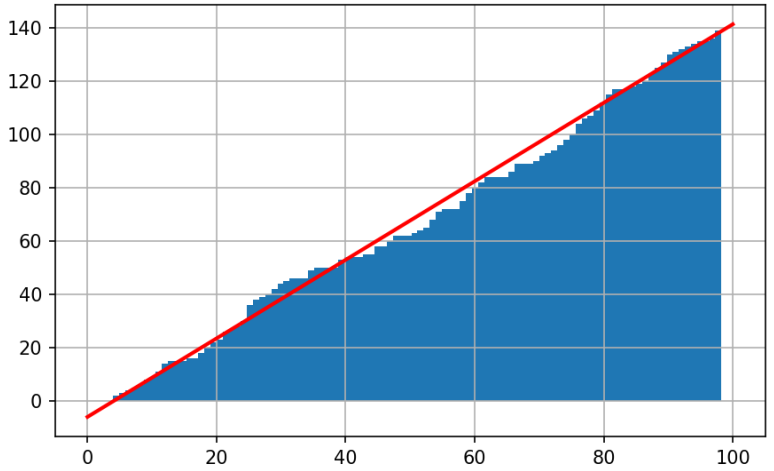
- Objectif: **définir des seuils de filtrage** → le cumul du nombre de pays est donné en fonction de chaque variable
- **Lectures graphiques** et recherche de **modélisations simples** pour déduire les valeurs de seuil de chaque variable
- **2 calculs de seuil** pour retenir N=20 ou N=30 pays par variable
- **L'intersection des résultats pour les 3 valeurs de seuil donne un résultat considérablement réduit**: Allemagne (N=20), plus (N=30) Japon, Corée, Allemagne, Royaume-Uni, France, Espagne et Canada

Population cible



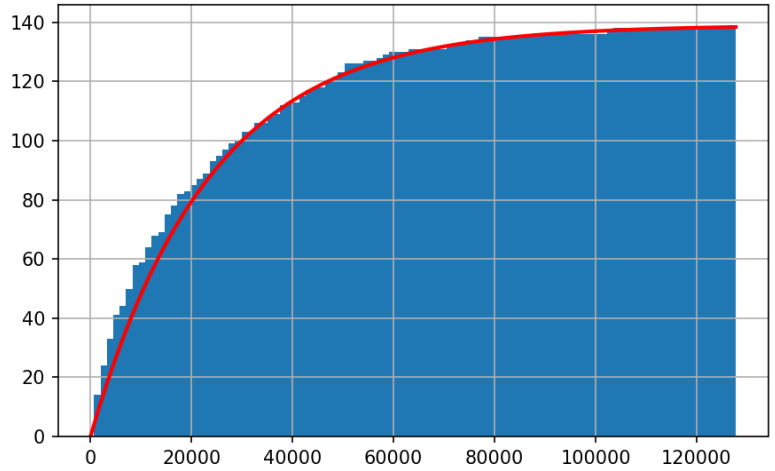

- Valeurs atypiques (Chine et Inde)
- Modélisation linéaire locale
- Seuil N=20: 5 600
- Seuil N=30: 2 600

Internet




- Quasi linéaire de (4%, 0) à (98%, 139)
- Modélisation linéaire
- Seuil N=20: 84%
- Seuil N=30: 77%

PIB par habitant



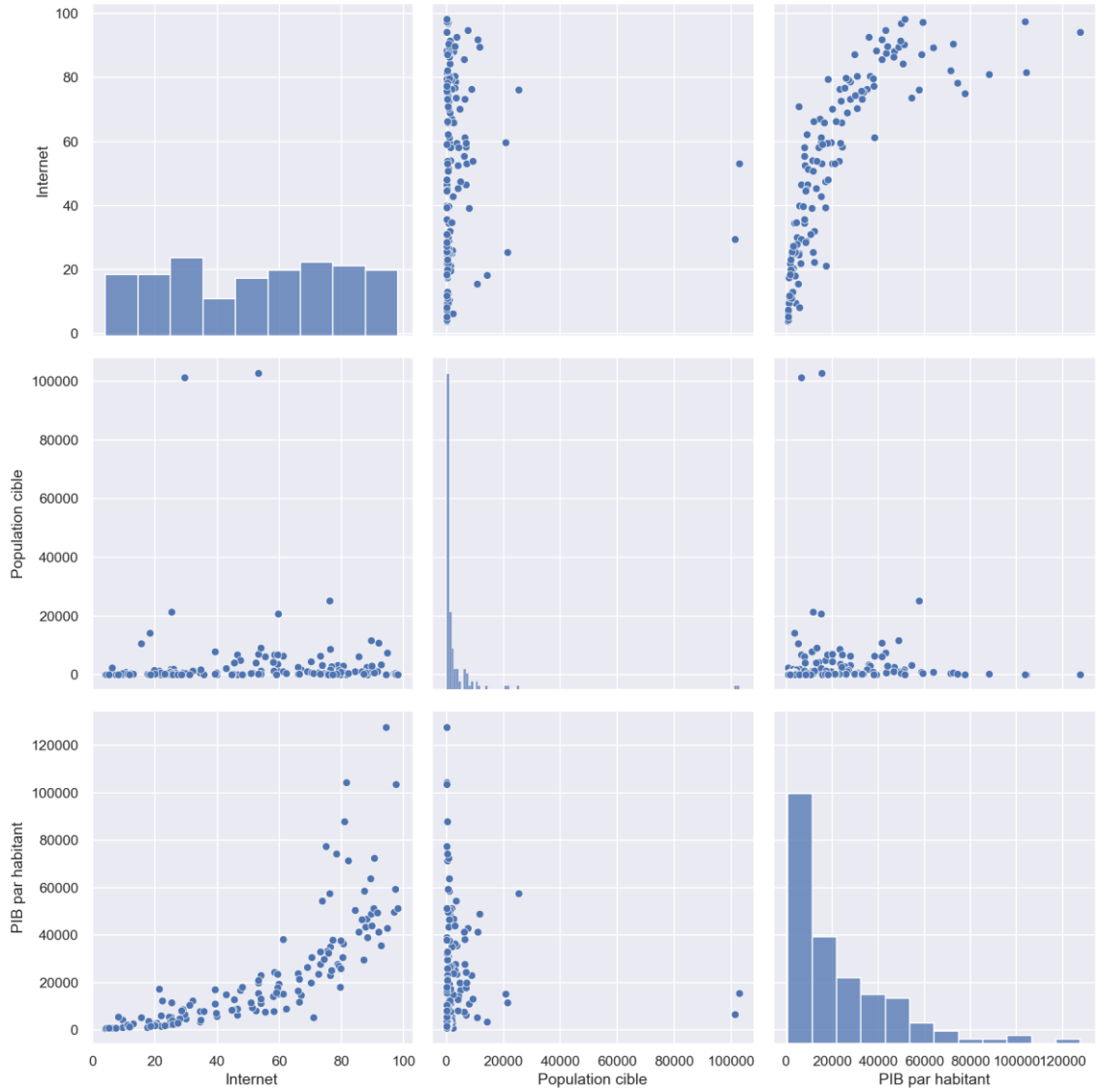

- *mean* = 22 392 et *std* = 23616
- Modélisation exponentielle inversé: 139 * (1 – exp -(N/*std*))
- Seuil N=20: 45 000
- Seuil N=30: 35 000



OPENCLASSROOMS CentraleSupélec

# Analyse combinée entre plusieurs variables (1 / 2)

- Objectif: **rechercher les corrélations entre variables** pour simplifier l'approche de recherche de seuils et converger plus rapidement vers le résultat
- Utilisation de **graphiques par paires** de variable
- **Pas de corrélation** visible entre l'indicateur de **population cible** et **les 2 autres indicateurs**
- **Corrélation** visible entre le **PIB par habitant** et le **taux d'utilisation d'internet** (fonction exponentielle)
  → *se déduit également des modélisations précédente*
- Conclusion: **le filtrage des pays** peut s'effectuer principalement **avec 2 indicateurs**
  - **Accessibilité au marché**: indicateur **internet** ou **PIB par habitant**
  - **Volume de marché**: indicateur de **population cible**

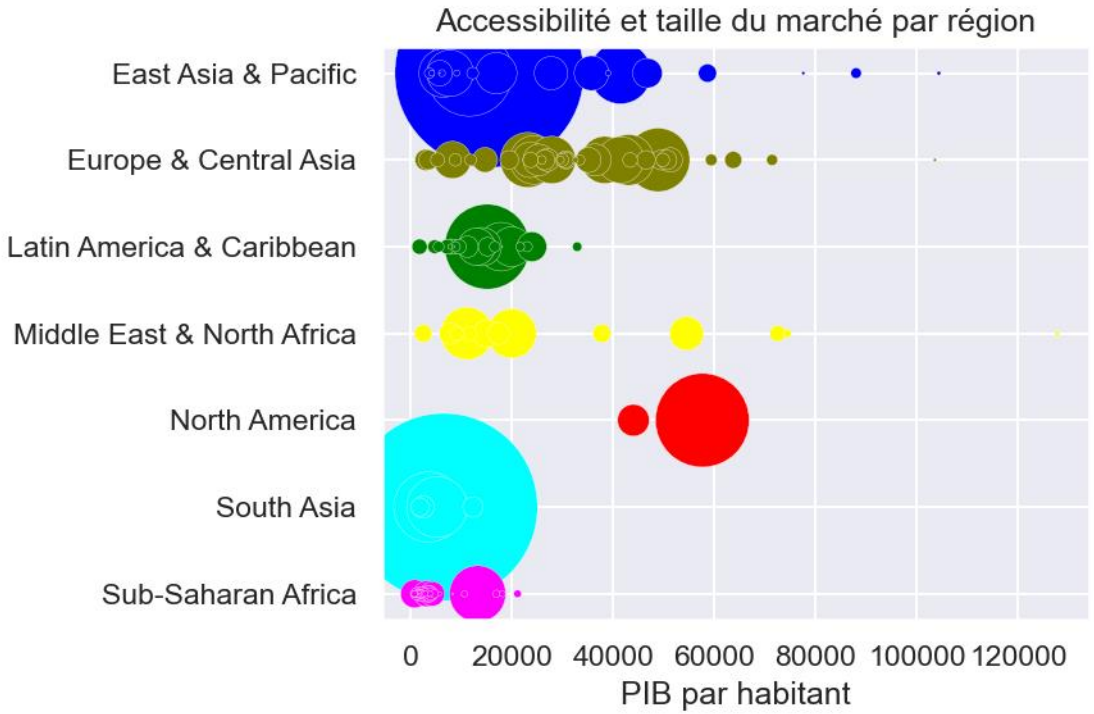


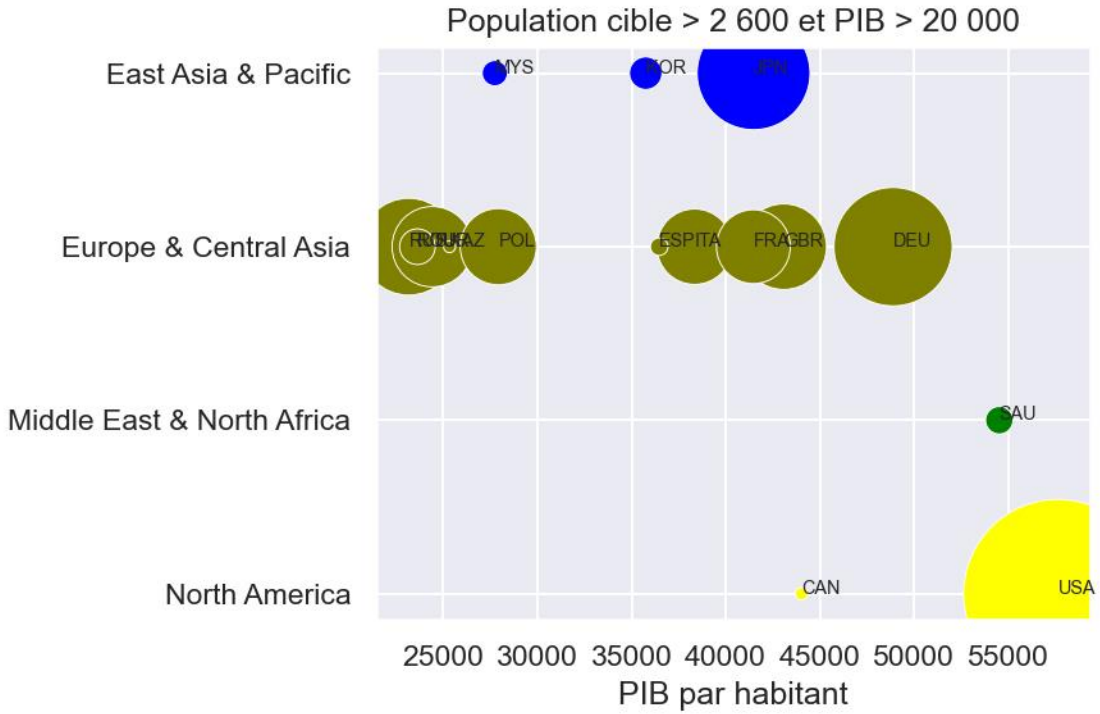
# Analyse combinée entre plusieurs variables (2 / 2)

Positionnement des pays par région :

- Accessibilité selon le **PIB par habitant**
  → *axe des abscisses*
- Volume de marché selon la **population cible**
  → *taille de la bulle*

*Exemple de filtrage avec un **PIB par habitant > 20 000**, correspondant à un **taux utilisation d'internet > 50%***



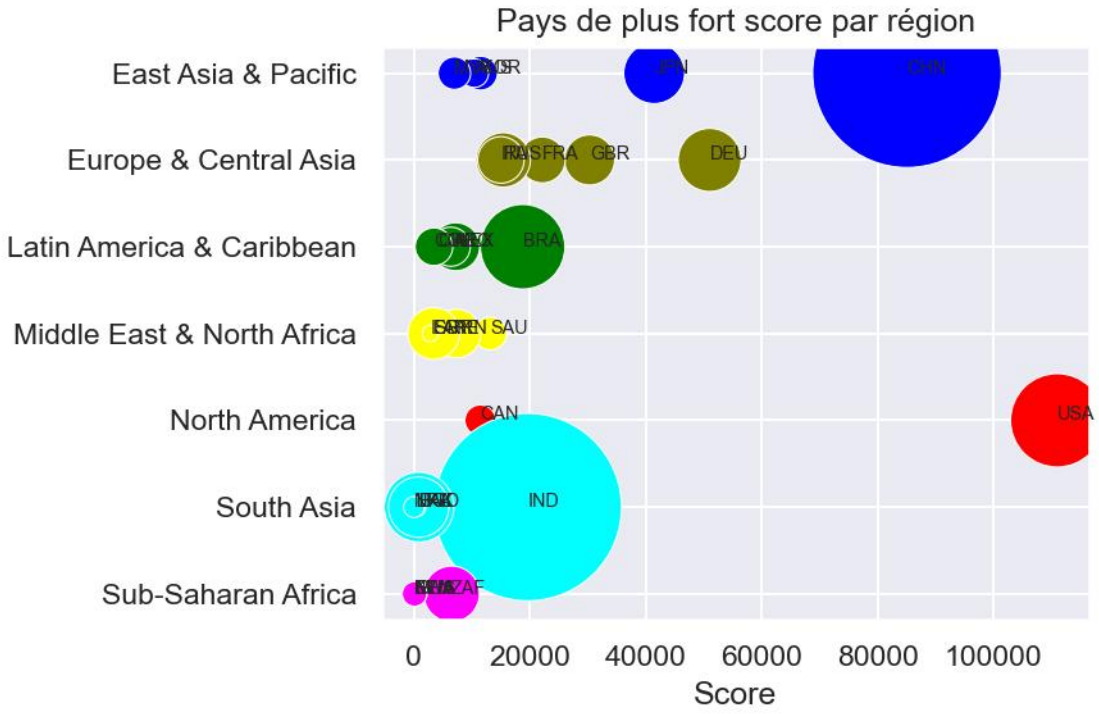



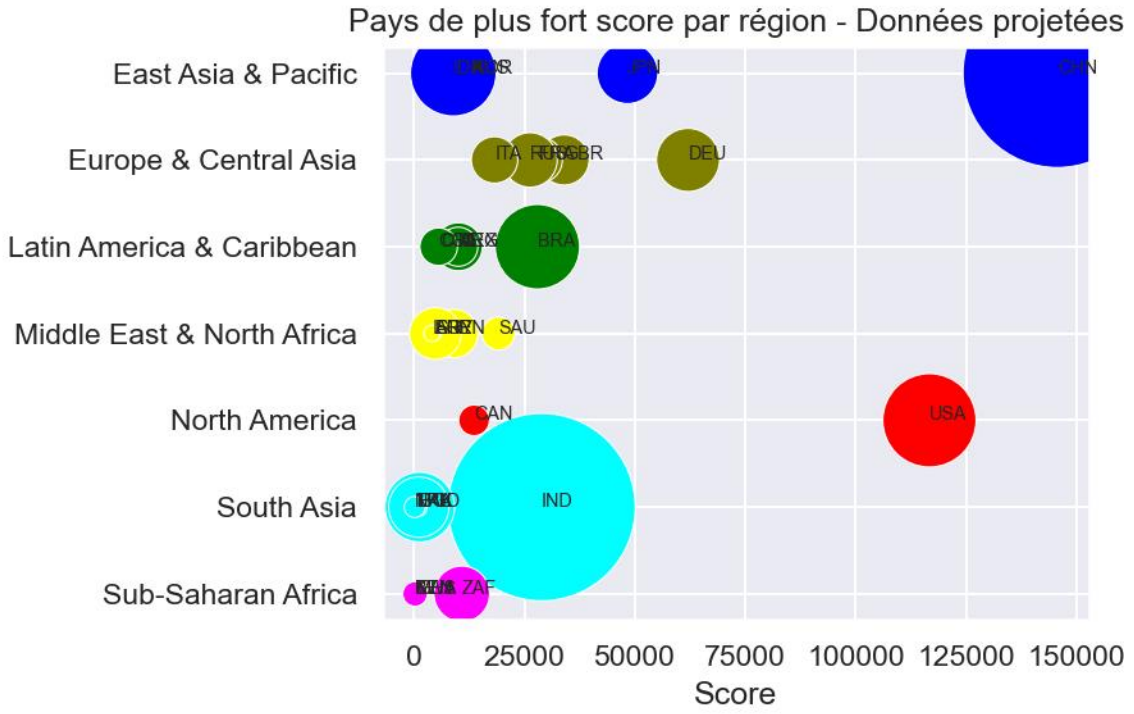



OPENCLASSROOMS
CentraleSupélec

# Proposition de score d'attractivité par pays (1 / 2)

Score = (% utilisateurs d'internet * PIB par habitant) * Population cible






OPENCLASSROOMS
CentraleSupélec

# Proposition de score d'attractivité par pays (2 / 2)

*Classement des pays par région puis par score*

| Country Code | Region | Pays | Score |
|---|---|---|---|
| CHN | East Asia & Pacific | China | 85 080 |
| JPN | East Asia & Pacific | Japan | 41 477 |
| KOR | East Asia & Pacific | Korea, Rep. | 11 534 |
| AUS | East Asia & Pacific | Australia | 10 450 |
| MYS | East Asia & Pacific | Malaysia | 7 022 |
| DEU | Europe & Central Asia | Germany | 51 044 |
| GBR | Europe & Central Asia | United Kingdom | 30 337 |
| FRA | Europe & Central Asia | France | 22 193 |
| RUS | Europe & Central Asia | Russian Federation | 15 440 |
| ITA | Europe & Central Asia | Italy | 15 044 |
| BRA | Latin America & Caribbean | Brazil | 18 795 |
| MEX | Latin America & Caribbean | Mexico | 7 252 |
| ARG | Latin America & Caribbean | Argentina | 6 408 |
| SAU | Middle East & North Africa | Saudi Arabia | 13 206 |
| IRN | Middle East & North Africa | Iran, Islamic Rep. | 7 412 |
| USA | North America | United States | 110 981 |
| CAN | North America | Canada | 11 487 |
| IND | South Asia | India | 19 693 |
| ZAF | Sub-Saharan Africa | South Africa | 6 538 |

*Classement des pays par score*

| Country Code | Region | Pays | Score |
|---|---|---|---|
| USA | North America | United States | 110 981 |
| CHN | East Asia & Pacific | China | 85 080 |
| DEU | Europe & Central Asia | Germany | 51 044 |
| JPN | East Asia & Pacific | Japan | 41 477 |
| GBR | Europe & Central Asia | United Kingdom | 30 337 |
| FRA | Europe & Central Asia | France | 22 193 |
| IND | South Asia | India | 19 693 |
| BRA | Latin America & Caribbean | Brazil | 18 795 |
| RUS | Europe & Central Asia | Russian Federation | 15 440 |
| ITA | Europe & Central Asia | Italy | 15 044 |
| SAU | Middle East & North Africa | Saudi Arabia | 13 206 |
| KOR | East Asia & Pacific | Korea, Rep. | 11 534 |
| CAN | North America | Canada | 11 487 |
| AUS | East Asia & Pacific | Australia | 10 450 |
| IRN | Middle East & North Africa | Iran, Islamic Rep. | 7 412 |
| MEX | Latin America & Caribbean | Mexico | 7 252 |
| MYS | East Asia & Pacific | Malaysia | 7 022 |
| ZAF | Sub-Saharan Africa | South Africa | 6 538 |
| ARG | Latin America & Caribbean | Argentina | 6 408 |

# Conclusions


- Un jeu de données de la Banque Mondiale qui nécessite un **temps d'apprentissage**
- De nombreux indicateurs et valeurs associées mais un **très faible taux de remplissage**
- Une recherche d'indicateurs fructueuse pour qualifier:
  - L'accessibilité au marché: **% d'utilisateurs d'internet** et **PIB par habitant**
  - Le volume de marché: indicateur de **population cible** calculé
- Le calcul de l'indicateur de population cible prend une **hypothèse commerciale** sur la stratégie de l'entreprise **qu'il conviendra de revoir avec le board** (matrice $Coef_I$)
- **L'approche par variable nécessite des itérations** pour que l'intersection soit significative
- La corrélation entre le PIB par habitant et le % d'utilisateurs d'internet permet la **représentation du marché en 2 dimensions**
- La proposition de **score d'attractivité donne le même résultat** en apportant la possibilité de **classement des pays**
- **La projection des données à 2020 ne change pas le résultat** mais montre l'attractivité croissante des marchés chinois et indien

OPENCLASSROOMS CentraleSupélec


Contact: Eric TREGOAT

eric.tregoat@gmail.com

06 49 99 79 59

https://www.linkedin.com/in/erictregoat/