

Expansion Internationale - Academy

Analyse des
données de la
Banque
mondiale pour
évaluer le jeu de
données



Sommaire

- Problématique
- Analyse générale des données
- Sélection des données pertinentes
- Préparation des données pour l'analyse
- Analyse des données
- Conclusions



Problématique

- Startup de la EdTech
 - Contenus de formation en ligne
 - Niveau lycée et université
- Projet d'expansion à l'international
- Analyse pré-exploratoire d'un jeu de données
- Qualification du jeu de données
- Description du contenu
- Sélection d'information pertinentes



Analyse générale du jeu de données

- Brève description des données
 - Source : Banque mondiale (EdStats)
 - Plus de 3600 indicateurs
 - Thèmes principaux : Education, Population, Infrastructures, Energie, etc.
 - Indicateurs principaux : Accès à internet, PIB, Population, etc.
 - Étendue temporelle : 1970-2100

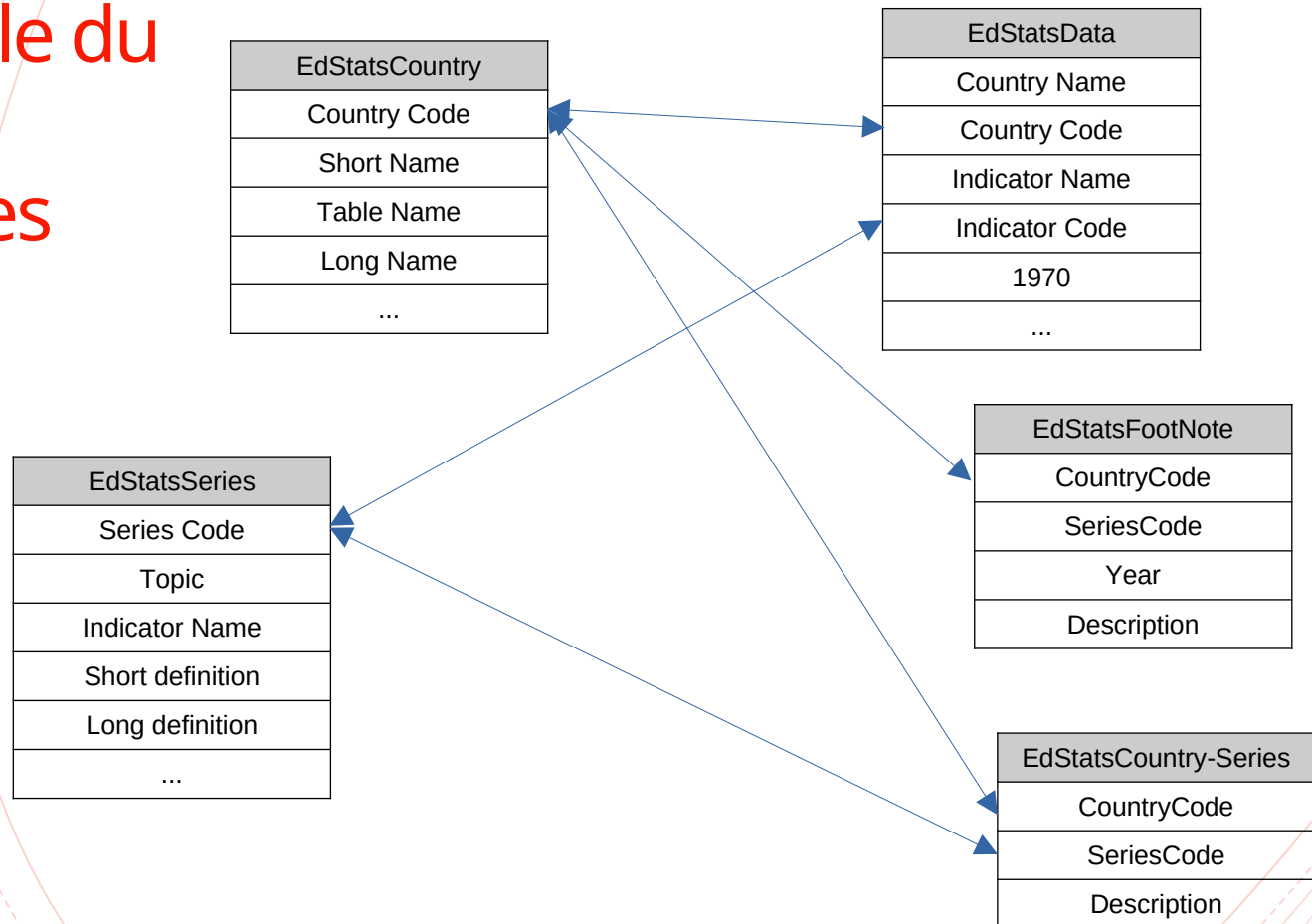


Analyse générale du jeu de données

- 5 jeux de données
 - EdStatsData
 - Principale source de données
 - Evolution des indicateurs dans le temps, par pays
 - 886930 lignes, 70 colonnes
 - EdStatsCountry
 - Différents pays étudiés
 - 241 lignes, 32 colonnes
 - « Region » et « Income Group » pertinents pour la suite de notre étude
 - EdStatsSeries
 - Permet la sélection des indicateurs
 - Décrit les indicateurs
 - 3665 lignes, 21 colonnes
 - EdStatsCountry-Series
 - Source des indicateurs pour chaque pays
 - 613 lignes, 4 colonnes
 - EdStatsFootNote
 - Informations sur la récolte des indicateurs
 - 643638 lignes, 5 colonnes

Analyse générale du jeu de données

Des jeux de données corrélés

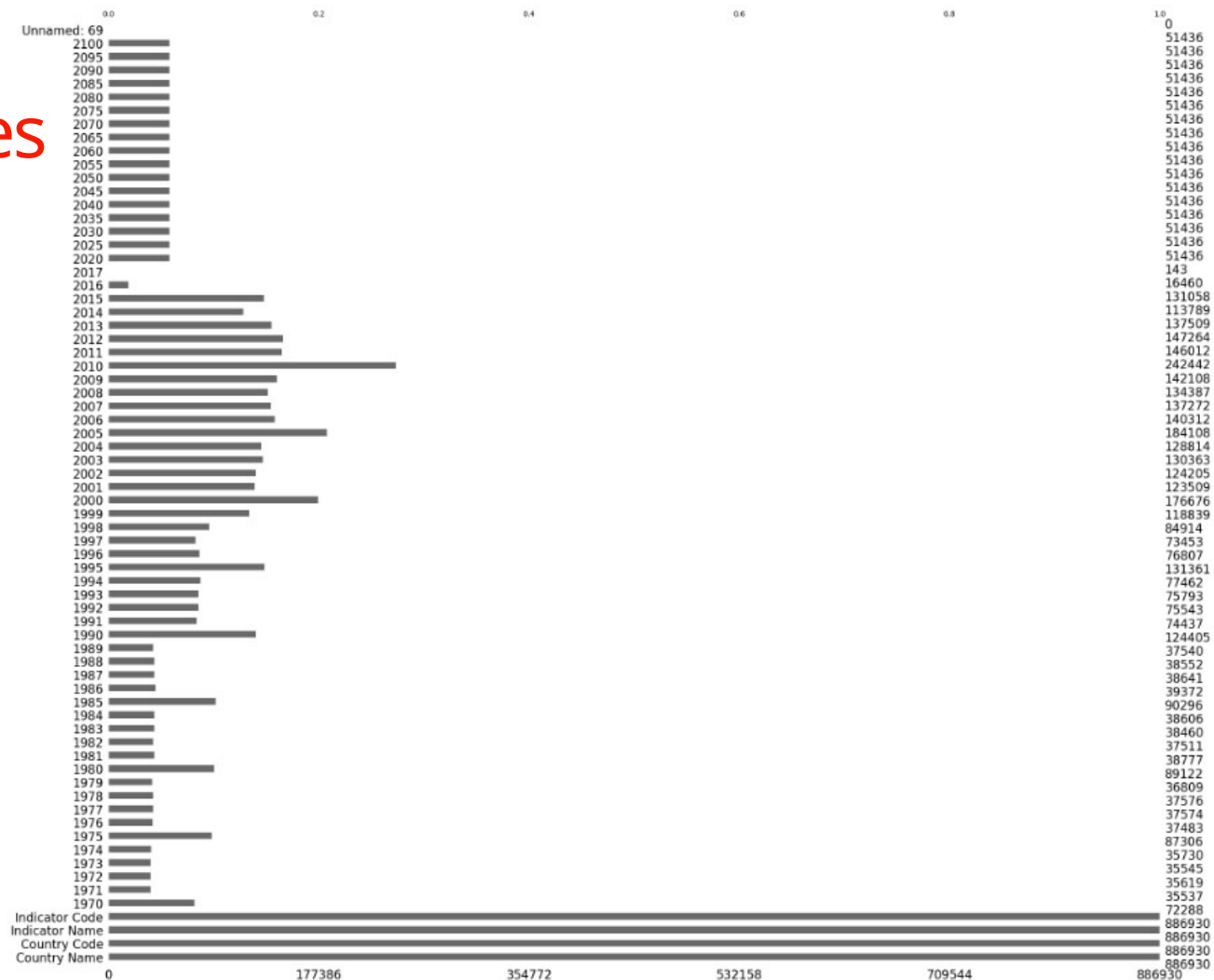




Sélection des données

- Plage temporelle
 - Répartition uniforme des données présentes
 - Pic de données entre 1990 et 2015
 - Données estimées à partir de 2020

Sélection des données



Valeurs manquantes dans le jeu de données « EdStatsData »



Sélection des données

- Volume de données par indicateur
 - Constat : certaines années ne présentent que peu de données
 - Traitement :
 - Une fois les années sélectionnées
 - Sélection des indicateurs avec un volume de données suffisant
 - 715 indicateurs conservés après simplification



Sélection des données

- Données métiers
 - Niveau d'éducation de la population
 - Compétences numériques
 - Investissement en éducation
 - Pouvoir d'achat
 - Démographie
 - Taux de chômage

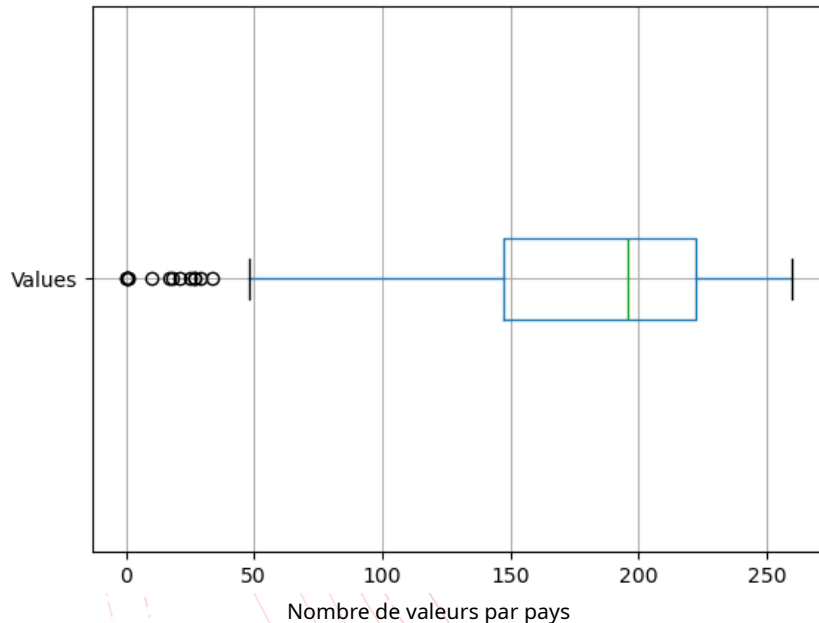


Sélection des données

- Indicateurs retenus
 - IT.NET.USER.P2 – Accès à internet
 - IT.CMP.PCMP.P2 – Accès à un ordinateur
 - SE.SEC.ENRR – Nombre d'élèves dans le secondaire
 - SE.TER.ENRR – Nombre d'élèves dans le supérieur
 - SE.SEC.PROG.ZS – Nombre d'élèves du secondaire intégrant le niveau supérieur
 - SE.COM.DURS – Durée légale des études
 - SE.SCH.LIFE – Espérance d'années passées à l'école
 - NY.GNP.MKTP.PP.CD – PIB par habitant
 - SP.POP.1524.TO.UN – Population des 15-24ans
 - SE.PRE.ENRL.TC.ZS – Nombre d'élèves par professeur
 - SL.UEM.TOTL.ZS – Taux de chômage

Préparation des données

- Préparation du dataframe
 - A partir du dataset « EdStatsData »
 - Sélection des années retenues
 - Sélection des indicateurs retenus
 - Ajout des données « Region » et « Income Group »
 - Suppression des pays avec trop peu de données

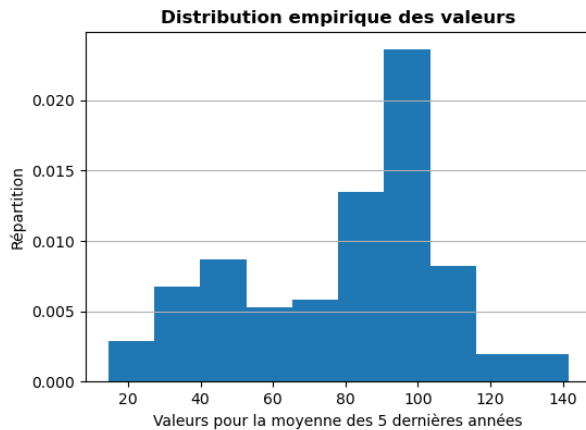




Préparation des données

- Préparation du dataframe
 - Suppression des pays n'ayant pas de données concernant les principaux indicateurs de la cible
 - Accès à internet
 - Accès à un ordinateur
 - Population des 15-24 ans
 - Conservation des 5 dernières années de données
 - Calcul de la valeur moyenne des 5 dernières années

Analyse des données

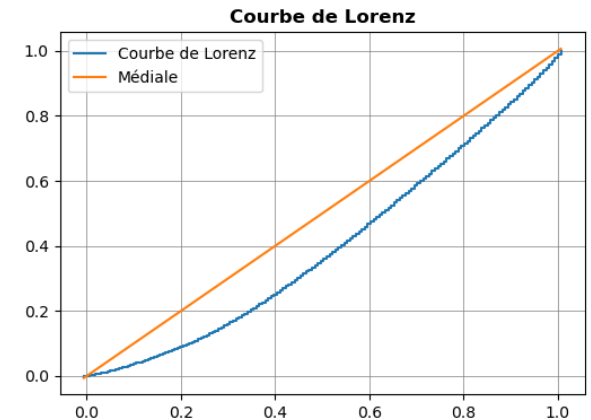
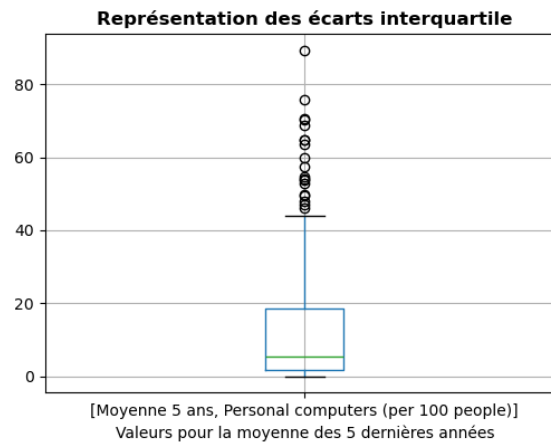


Valeurs des statistiques

Mode	5.289016
Moyenne	19.266480
Médiane	18.451036
Variance	67.413128
Ecart type	8.210550
Skewness	0.723181
Kurtosis	0.533357

■ Analyse des statistiques par indicateur

- Mode, moyenne, médiane
- Variance, écart type
- Distribution empirique
- Ecart interquartile
- Courbe de Lorenz



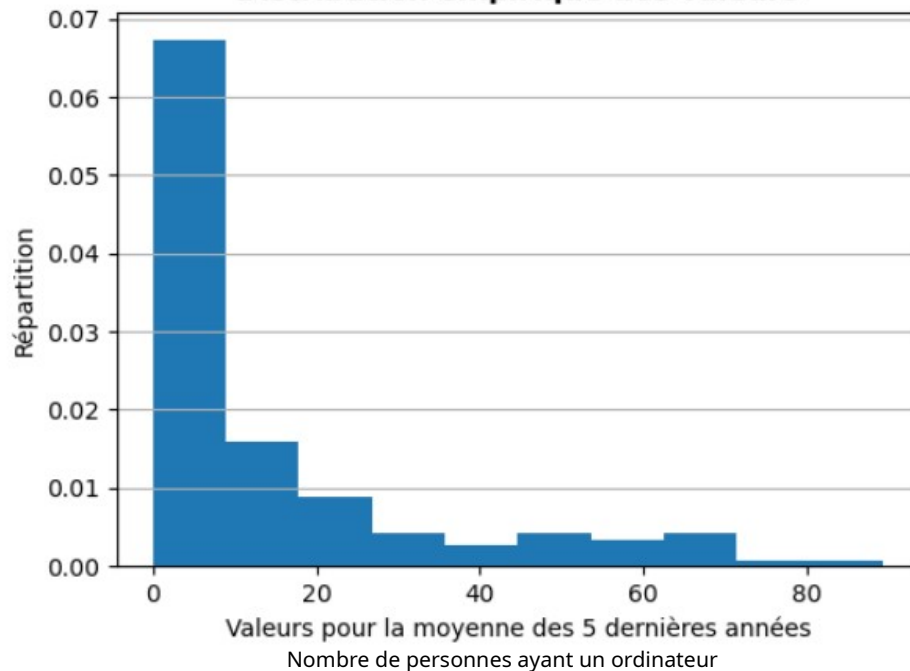
Analyse des données

■ Analyse des statistiques par indicateur

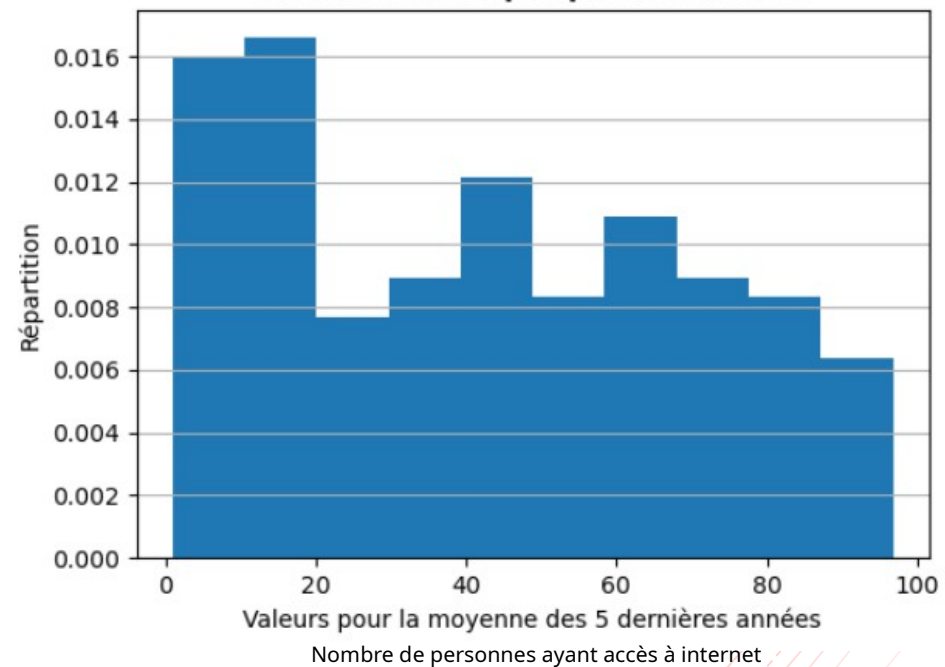
■ Faits intéressants :

- Peu de pays ont accès à un ordinateur
- La répartition de l'accès à internet est plus équilibrée

Distribution empirique des valeurs



Distribution empirique des valeurs

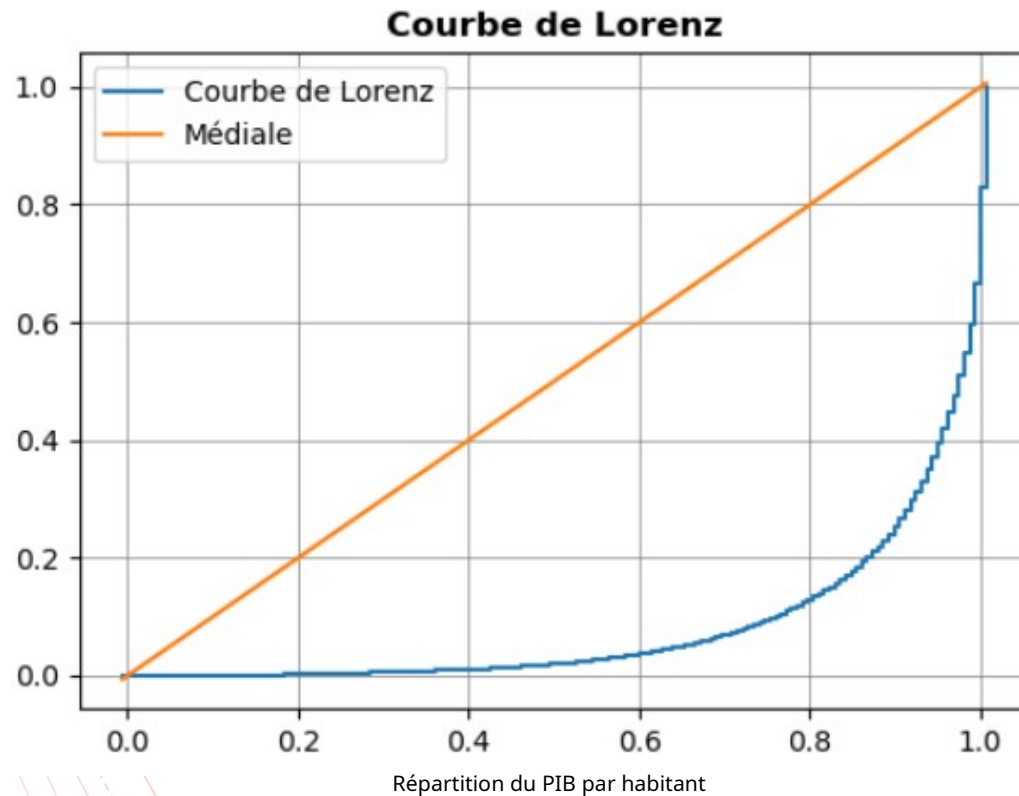


Analyse des données

- Analyse des statistiques par indicateur

- Faits intéressants :

- Très peu de pays regroupent la majorité

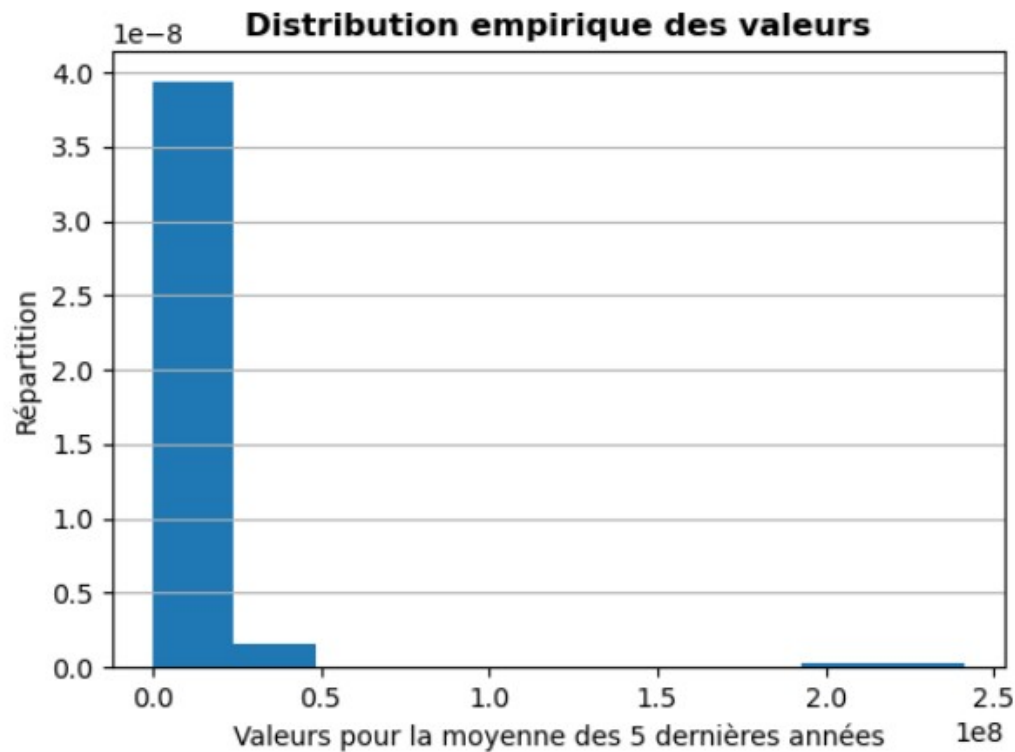


Analyse des données

- Analyse des statistiques par indicateur

- Faits intéressants :

- Répartition homogène de la population de 15-24 ans dans les pays



Distribution de la population des 15-24ans

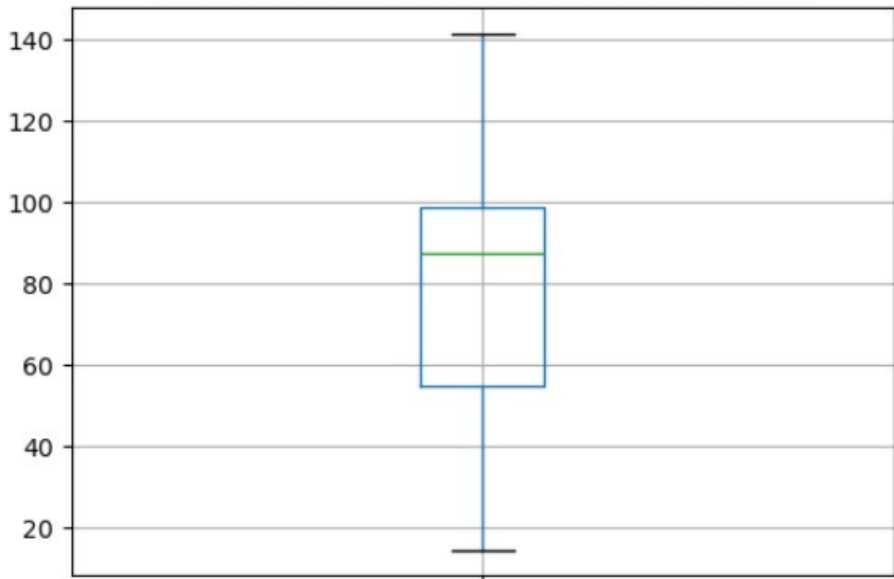
Analyse des données

■ Analyse des statistiques par indicateur

■ Faits intéressants :

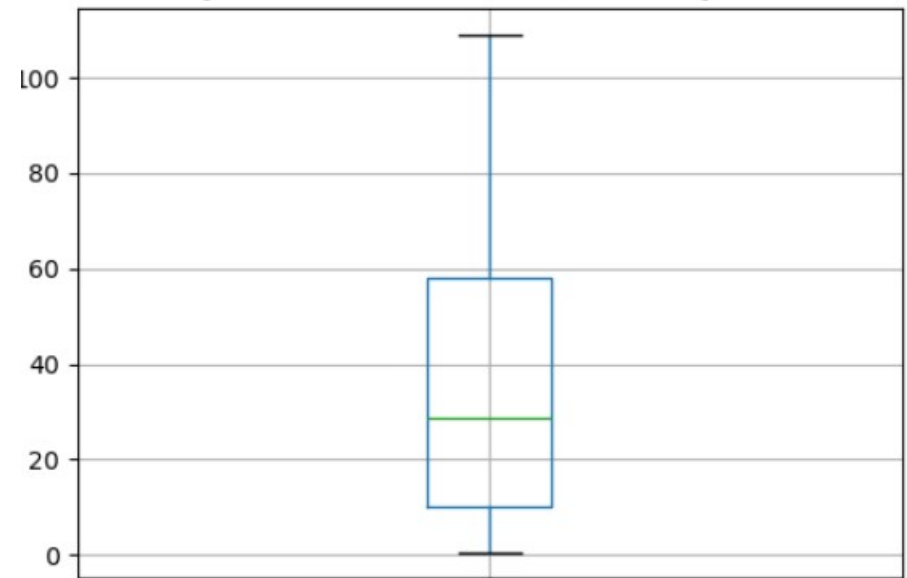
- 80% des étudiants entrent dans le secondaire, alors que 35% seulement entrent dans le supérieur

Représentation des écarts interquartile



[Moyenne 5 ans, Gross enrolment ratio, secondary, both sexes (%)]
Valeurs pour la moyenne des 5 dernières années

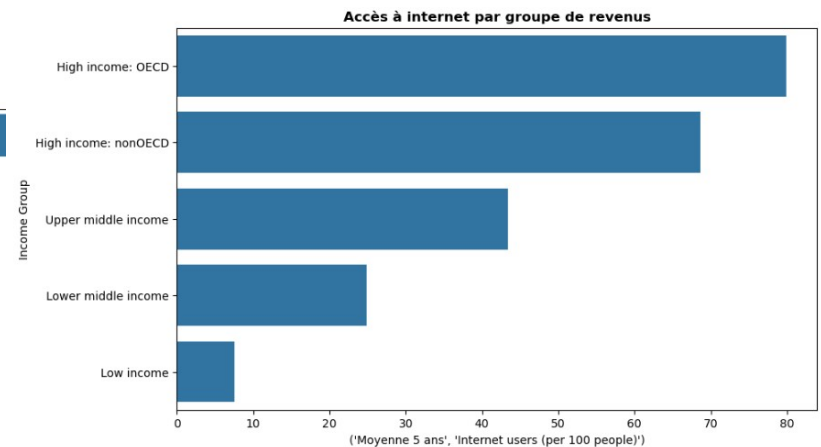
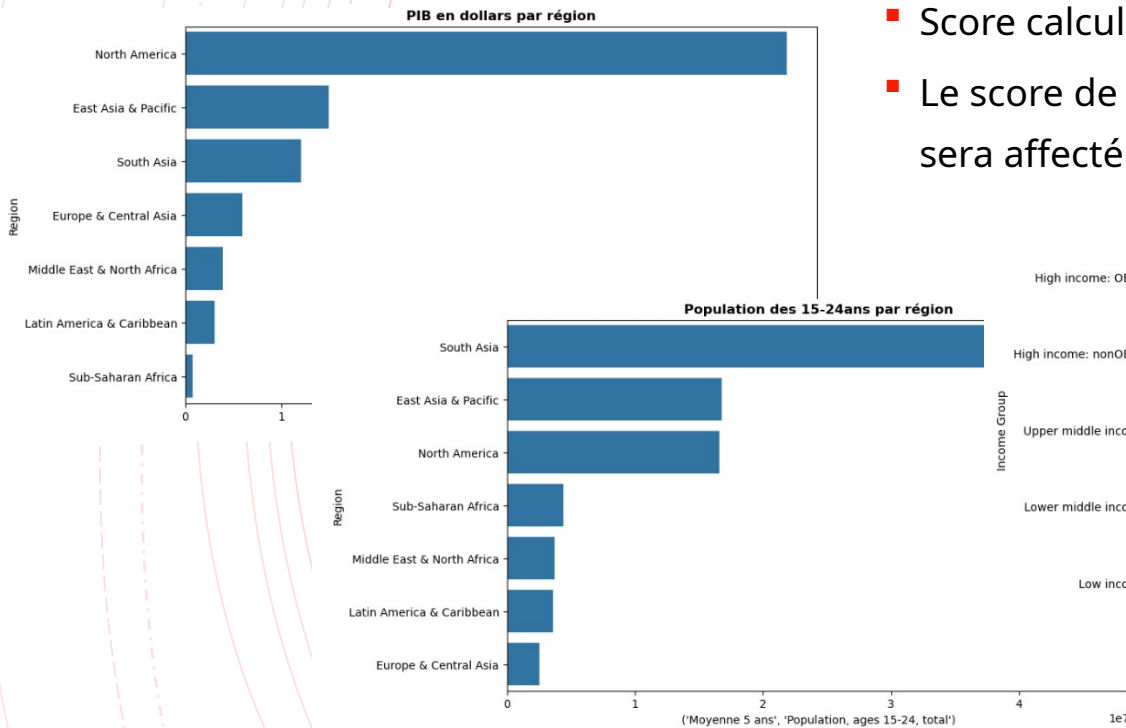
Représentation des écarts interquartile



[Moyenne 5 ans, Gross enrolment ratio, tertiary, both sexes (%)]
Valeurs pour la moyenne des 5 dernières années

Analyse des données

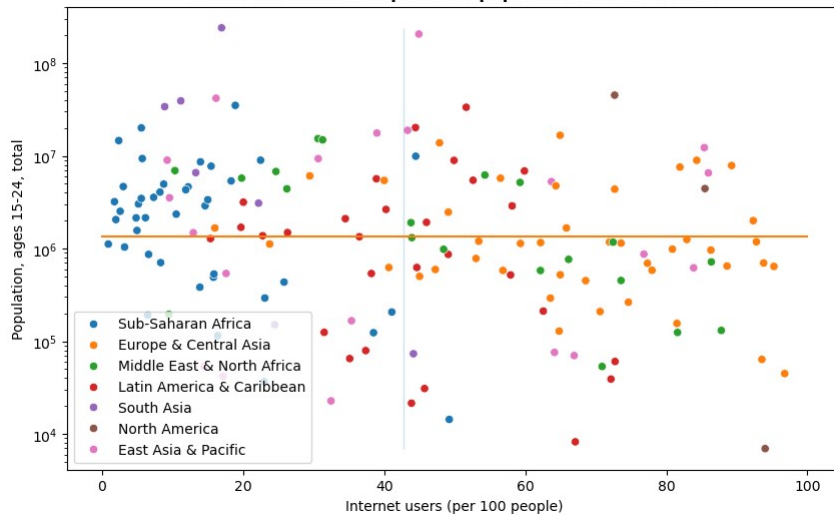
- Analyse des statistiques par région et groupe de revenus
 - Etablissement d'un classement pour chaque composante
 - Score calculé selon le rang dans l'étude
 - Le score de la région ou du groupe de revenus sera affecté individuellement aux pays



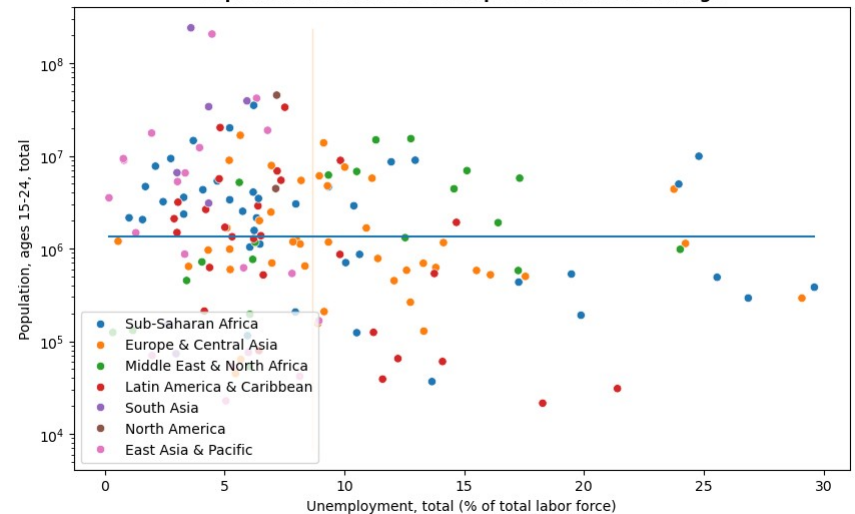
Analyse des données

- Analyse des corrélations entre indicateurs
 - Confrontation des indicateurs pour mettre en avant certains pays
 - Utilisation de la moyenne ou de la médiane selon la distribution de l'indicateur
 - Attribution d'un score à chaque pays selon ces critères

Accès à internet comparé à la population des 15-24 ans

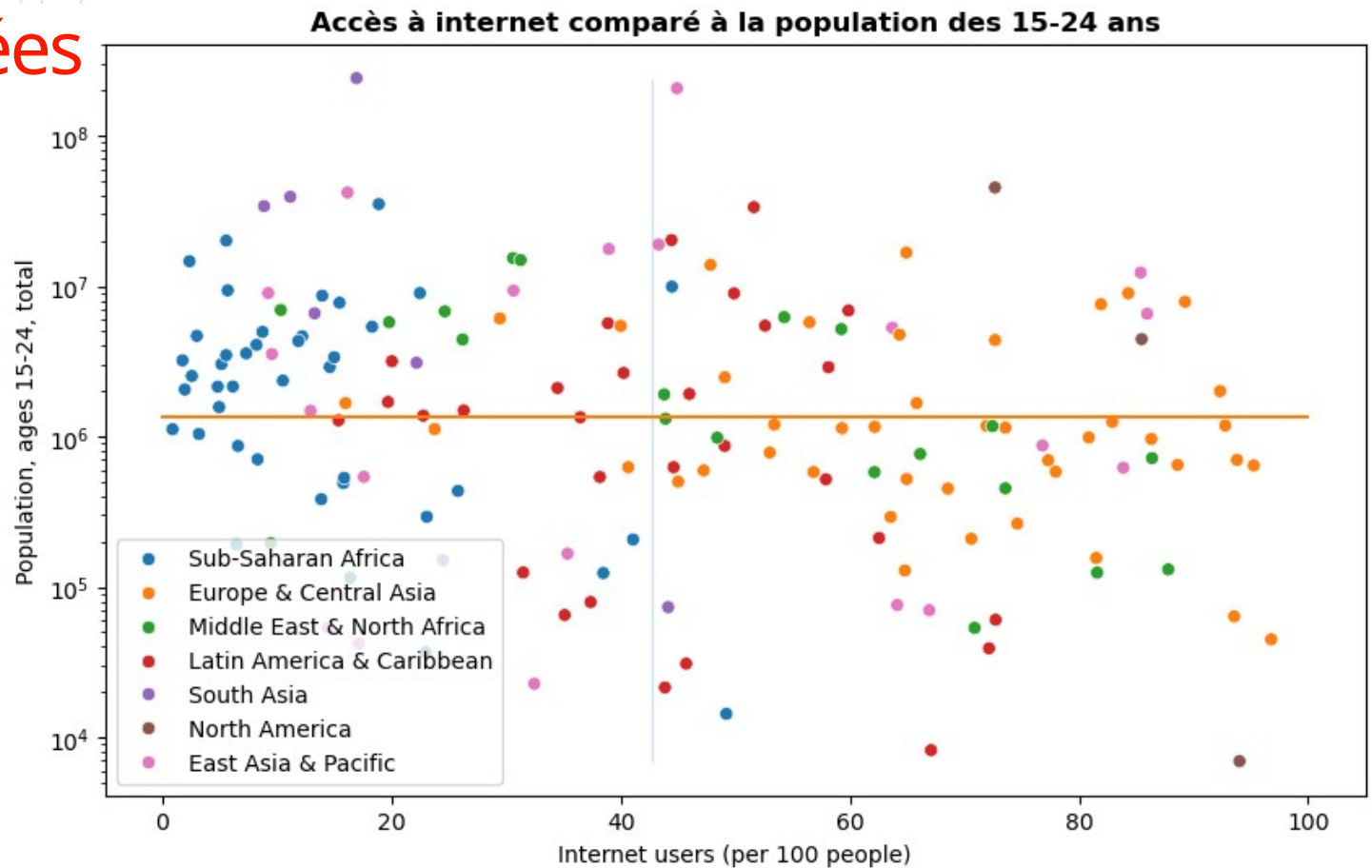


Population des 15-24 ans comparé au taux de chômage



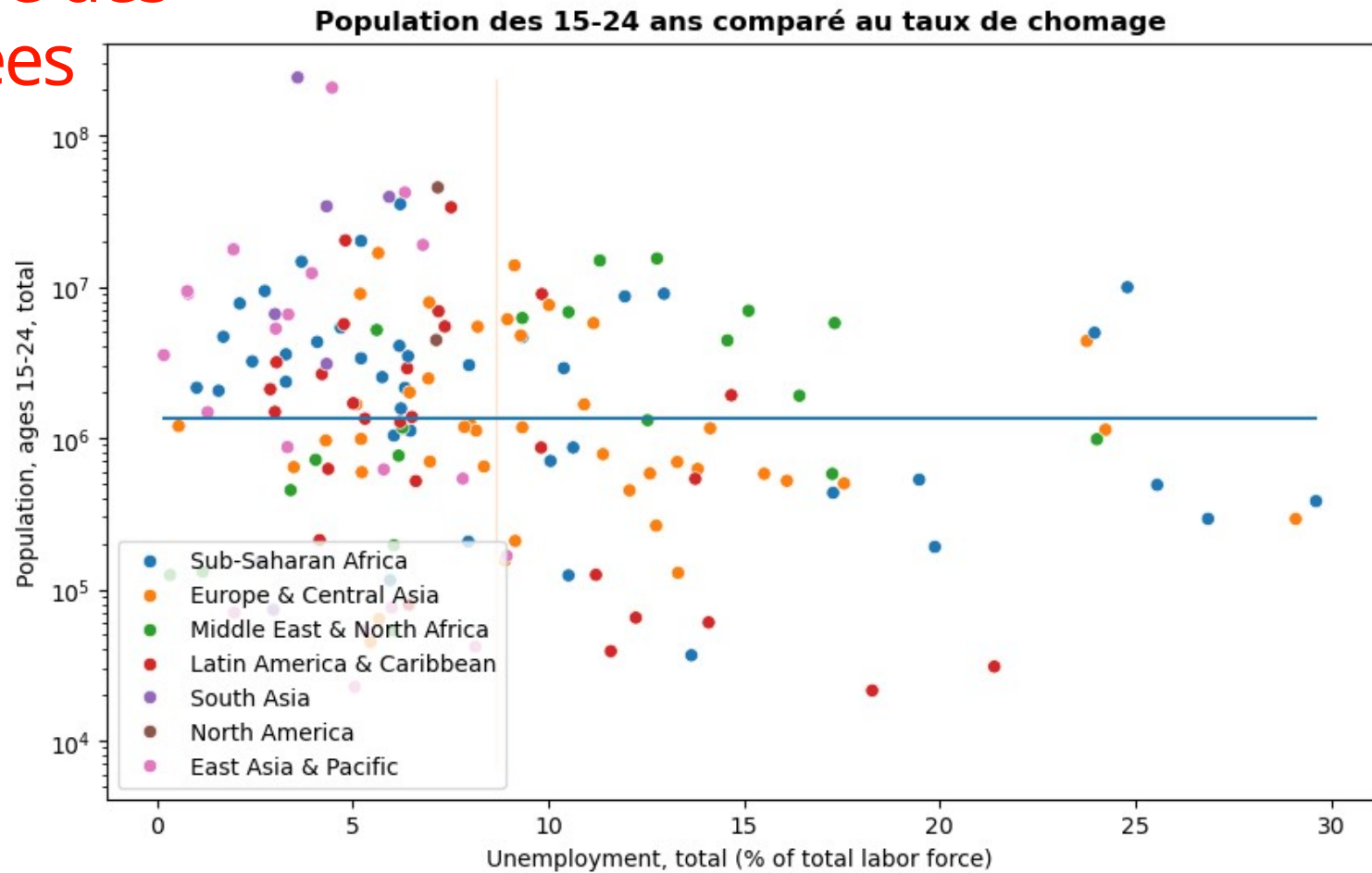
Analyse des données


- Analyse des corrélations entre indicateurs



Analyse des données

- Analyse des corrélations entre indicateurs





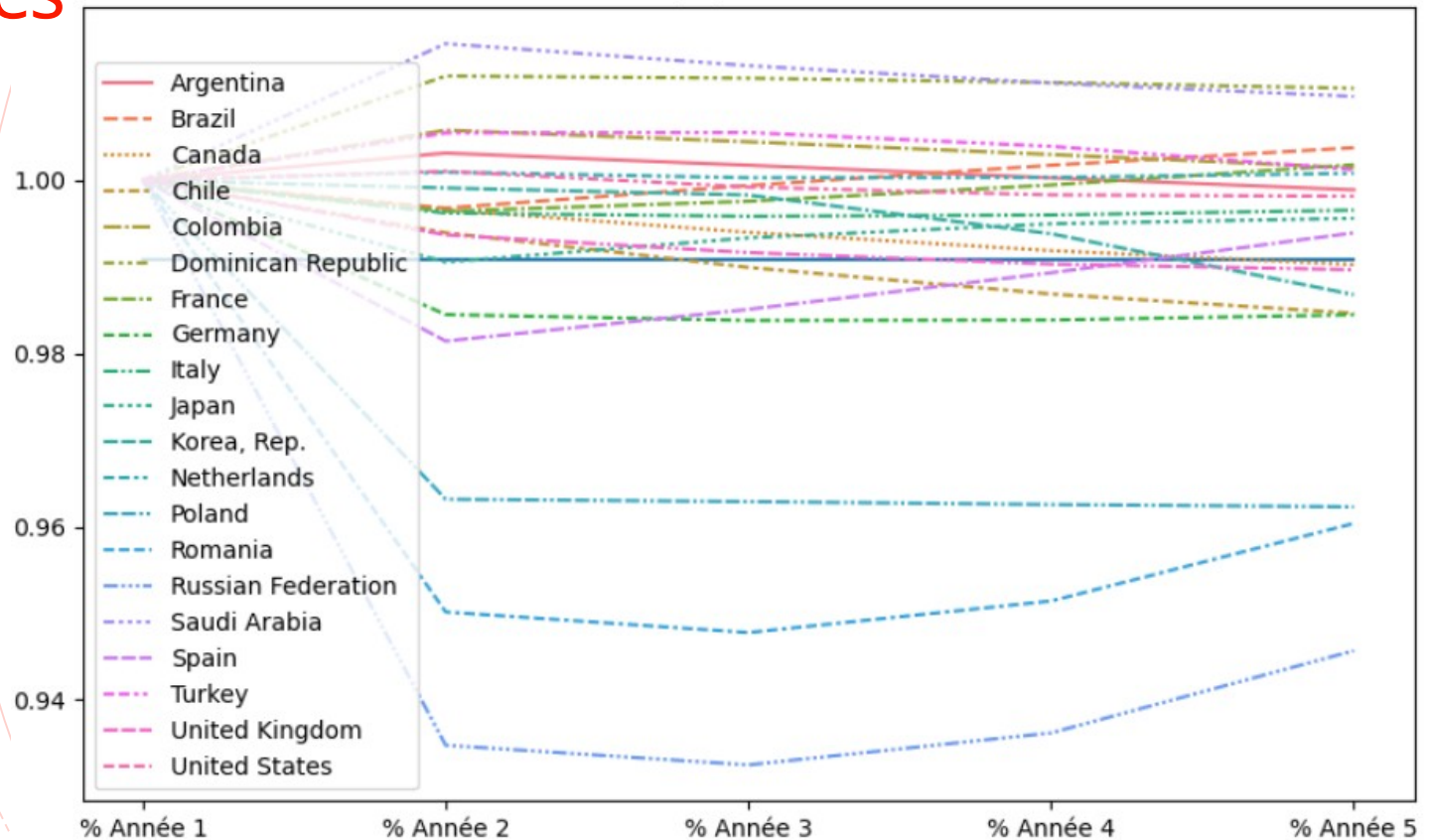
Analyse des données

- Analyse de l'évolution des indicateurs par pays
 - Sélection de 5 indicateurs pertinents
 - Étude de leur évolution pour chaque pays
 - Attribution d'un score à chaque pays
l'évolution de l'indicateur

Analyse des données

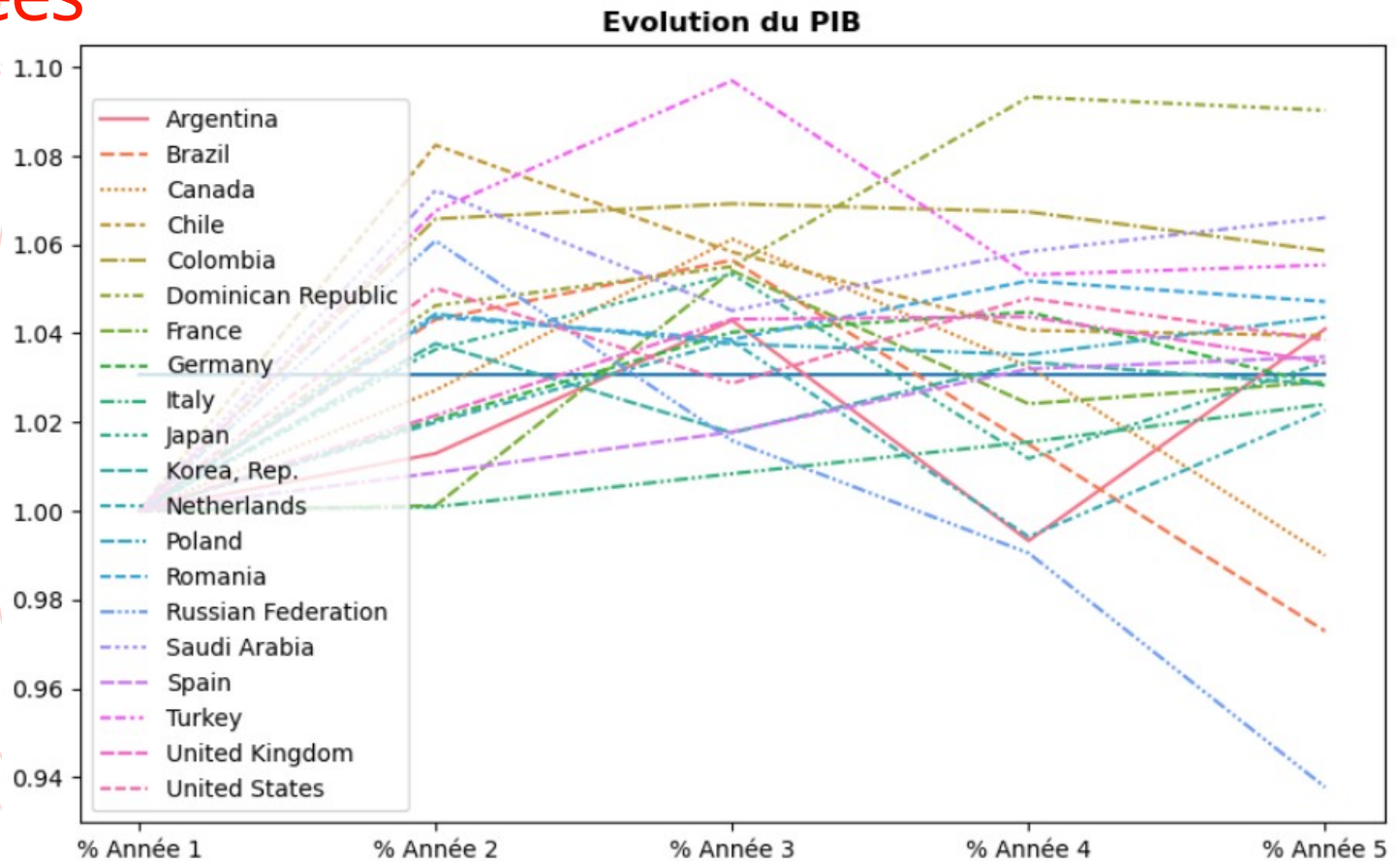
- Analyse de l'évolution des indicateurs par pays

Evolution de la population des 15-24ans



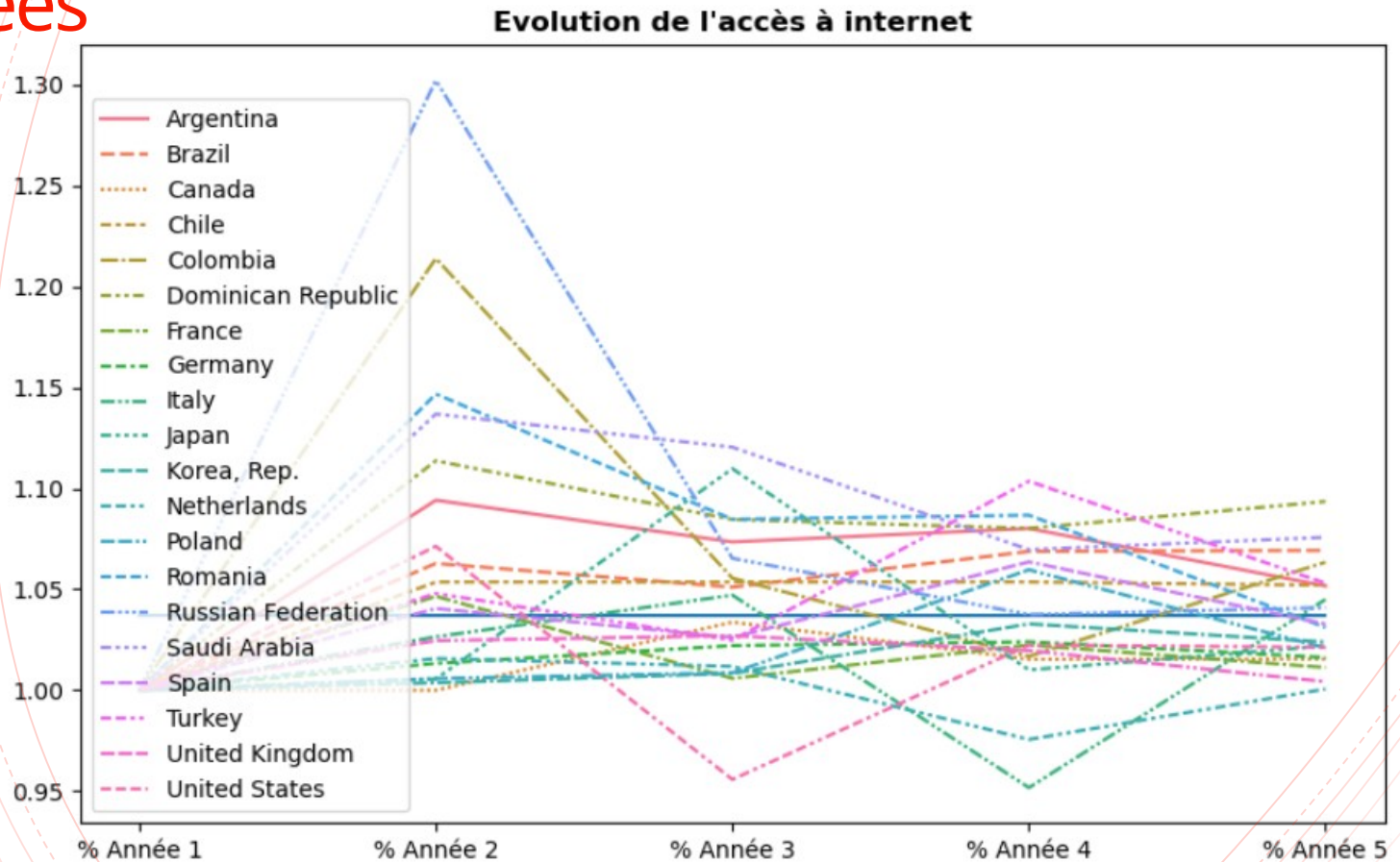
Analyse des données

- Analyse de l'évolution des indicateurs par pays



Analyse des données

- Analyse de l'évolution des indicateurs par pays



Conclusions

0	Turkey
1	Italy
2	Colombia
3	Saudi Arabia
4	France
5	Spain
6	Argentina
7	Poland
8	Dominican Republic
9	Chile
10	Russian Federation
11	United States
12	Japan
13	United Kingdom
14	Netherlands
15	Brazil
16	Korea, Rep.
17	Germany
18	Romania
19	Canada

- L'analyse nous a permis de définir un classement des pays selon leur attractivité
- Le jeu de données présente bel et bien des informations qui pourront aider l'entreprise à prendre des décisions
- Recommandations
 - La première analyse a été effectuée sans support métier
 - La poursuite de cette étude avec un expert sera pertinente
 - Il serait opportun de pondérer les indicateurs pour affiner la sélection de pays