

Analyse

EdStats All Indicator Query"

Academy

*Sofiane Mouhab
21 Janvier 2021*

SOMMAIRE

- | | |
|---|---|
| <ol style="list-style-type: none">1. Problématique p.32. Description p.43. Zoom sur les données (schéma 1 & 2) p.54. Variable 1 : Année (schéma 3, 4 & 5) p.65. Variable 2 : Indicateur (schéma 6 & 7) p.76. Variable 3 : Pays (schéma 8 & 9) p.87. Variable 4 : Nombre de lycéens par pays (schéma 10,11,12,13 & 14) p.98. Variable 5 : Nombre d'étudiants par pays (schéma 15,16,17,18 & 19) p.109. Variable 6 : PIB par habitant en \$ (schéma 20,21,22,23 & 24) p.1110. Variable 7 : Taux de connexion internet (schéma 25,26,27,28 & 29) p.12 | <ol style="list-style-type: none">11. Les Filtres p.13<ol style="list-style-type: none">11.1. Colonnes inutiles (schéma 30, 31 & 32) p.1311.2. Année 1970-2004 & 2020-2100 (schéma 33) p.1311.3. Retenir la dernière valeur disponible (schéma 34) p.1411.4. Supprimer les années restantes (schéma 35) p.1512. Scoring (schéma 36, 37,38 & 39) p.1613. Résultats par pays (schéma 40 & 41) p.1714. Résultats par région (schéma 42) p.1815. Perspective (schéma 43) p.1916. Conclusion (schéma 44, 45 & 46) p.2017. Annexes p.21,22,23 & 24 |
|---|---|

1 - Problématique

Notre société souhaite se lancer dans une expansion à l'international.
A partir des données de la banque mondiale (sur l'éducation) peut-on envisager ce projet ?

A déterminer :

- 1 - Les pays avec un fort potentiel
- 2 - Evolution dans le futur
- 3 - Pays à prospector

En préambule :

- 1 - Valider la qualité de ce jeu de données
- 2 - Sélectionner les informations qui semblent pertinentes
- 3 - Déterminer des ordres de grandeurs

2 - Description

Nous avons reçu un ensemble de 5 fichiers, procédons rapidement à une revue d'effectif :

- EdStatsCountry-Series : Sources des différents indicateurs pour chaque pays
- EdStatsCountry : Données de démographique et géographique pour chaque pays
- EdStatsSeries : Détails et sources des indicateurs
- EdStatsFootNote : Description des indicateurs pour chaque pays
- EdStats Data : Données générales pour tous les indicateurs et tous les pays
-

Nous analyserons en particulier le document "EdStatsData" qui reprend l'essentiel des informations nécessaires à notre étude.

On compte 70 colonnes parmi lesquelles :	Nous sommes en présence de 886930 lignes :
<ul style="list-style-type: none">• Information sur le pays• Information sur l'indicateur• Information sur les années passés• Information sur les futurs années• Une colonne sans fonction	<ul style="list-style-type: none">• Le pays ou la région• Un indicateur• Données en fonction des années

Une rapide analyse nous permet de déterminer que nous sommes en présence de :

- 242 pays ou régions
- 3665 indicateurs

Soit 886930 lignes différentes

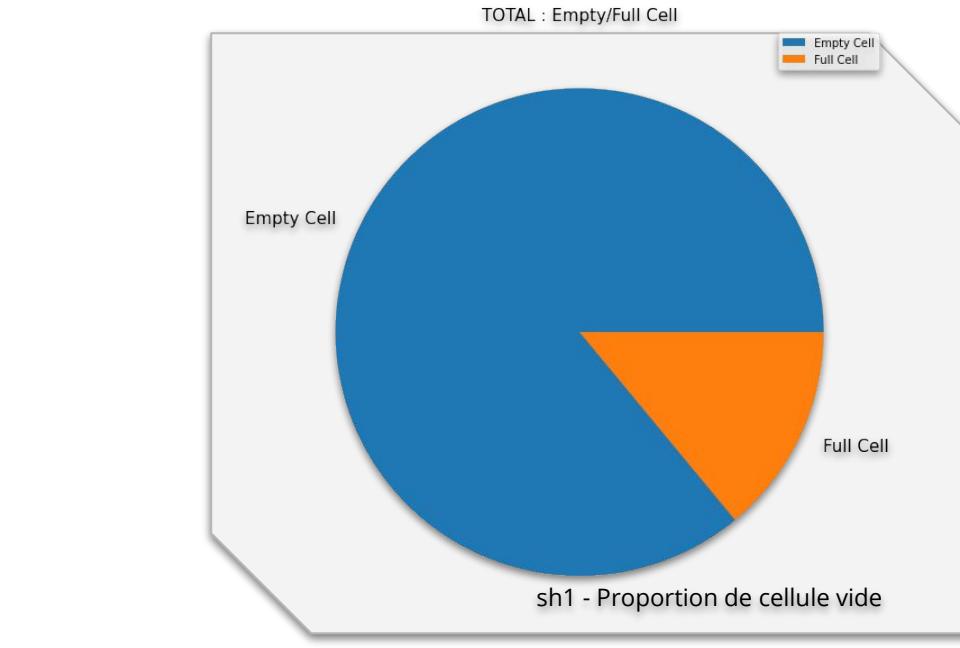
3 - Zoom sur les données

Nous sommes en présence d'une base de données assez imposante, mais globalement assez peu renseigné.

On le constate sur la matrix (sh2) (représentation du remplissage du tableau, blanc pour les cellules vides, noir pour celle avec des données).

On compte 86% de cellules vides (sh1), cela peut s'avérer difficile pour une analyse objectif...

Déterminons ainsi les informations nécessaires et la possibilité de continuer études.



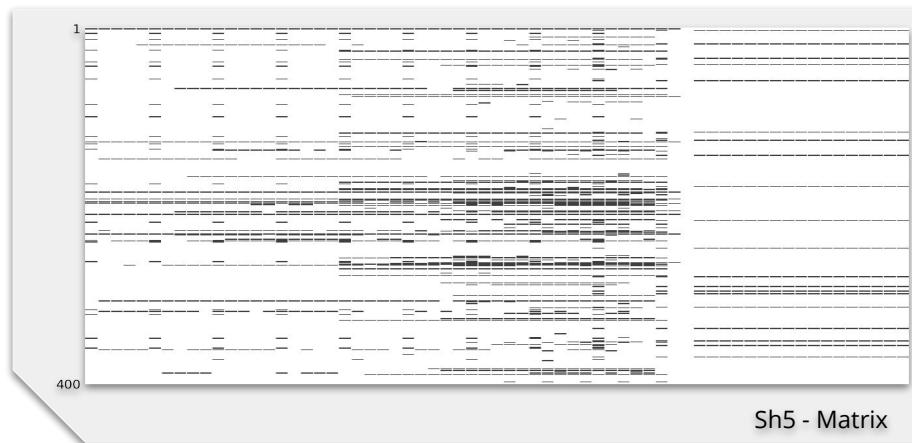
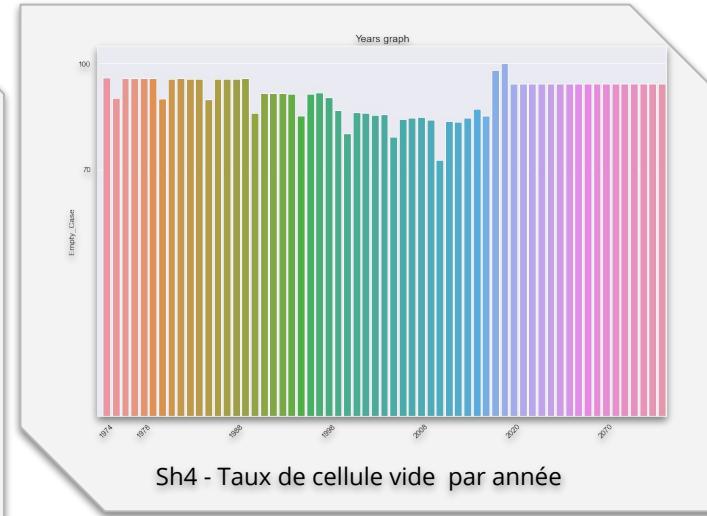
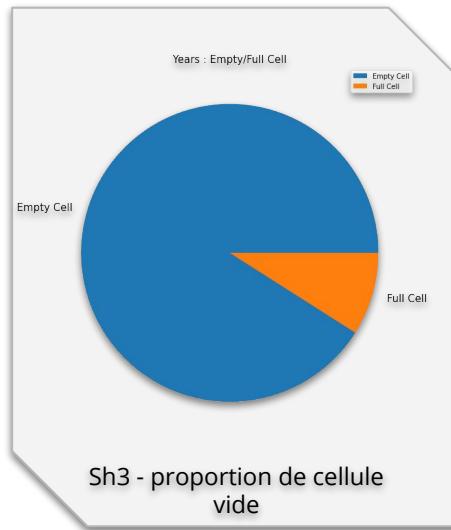
4 - Variable 1 : Année

Notre première variable à laquelle on peut porter intérêt est celle concernant les années en présence, elles se décourent de 2 manières :

- Les années en présence 1970 - 2017
- Les années futurs 2020 - 2100

On se rend compte d'une grande disparité des données (sh5). Plus précisément : 91% des cellules sont vides (sh3).

Néanmoins on constate (sh4) qu'une partie des années possèdent des données significatives, en particulier la période 2004-2014

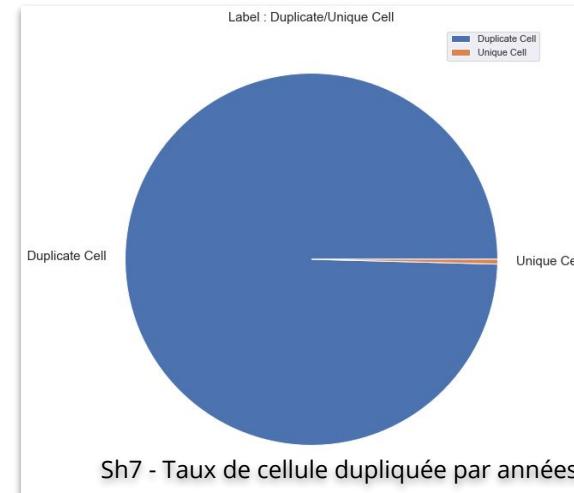
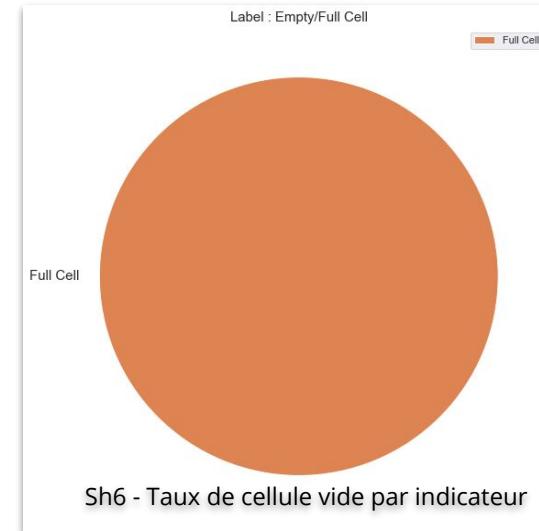


5 - Variable 2 : Indicateur

Seconde variable, il s'agit évidemment des indicateurs concernant les données.

On note (sh6) qu'aucune cellule de cette variable n'est nulle, 100% des données sont renseignés.

En ce qui concerne les doublons (sh7), 99,5 % des données sont dupliquées, néanmoins ces valeurs s'explique par le fait que chaque indicateurs est répété autant de fois que le nombre de pays (ou région). Après suppressions de ces derniers, on retrouve bien les 3665 indicateurs évoquées plus haut

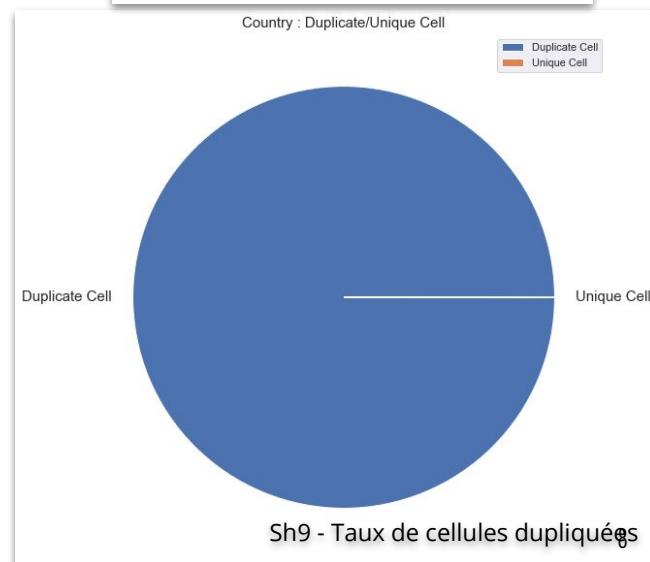
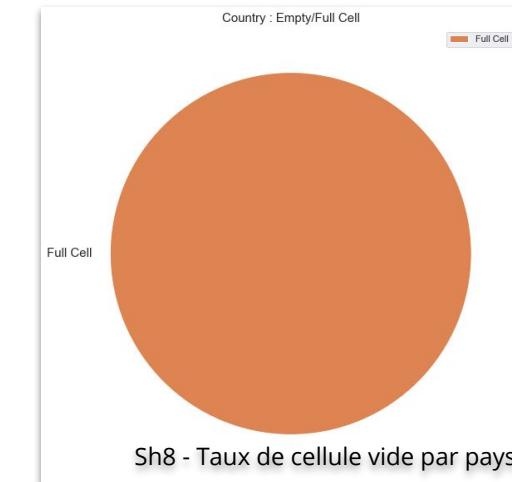


6 - Variable 3 : Pays

Troisième variable, les variables concernant les pays en présence

On constate (sh8) qu'aucune cellule de cette variable n'est nulle, 100% des données sont renseignés.

Les doublons (sh9) par contre sont nombreux : 99.9 % des données sont dupliquées, de la même manière que précédemment le nombre de pays étant corrélé au nombre d'indicateur, on voit bien après suppressions de doublons, on retrouve bien les 242 pays ou régions évoquées en préambule



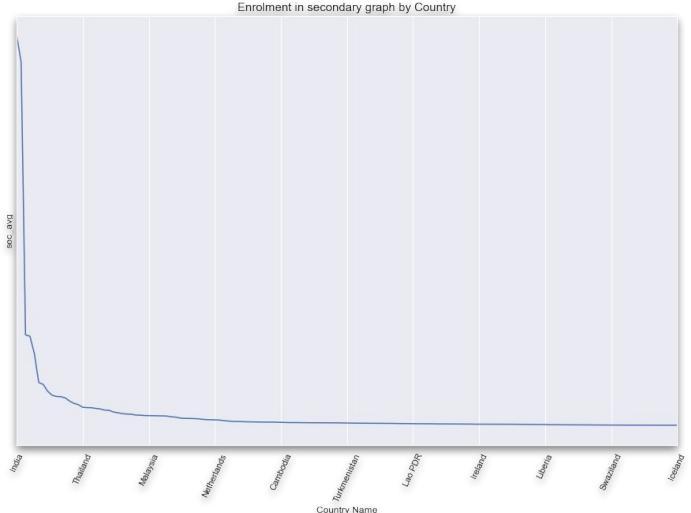
7 - Variable 4 : Nombre de Lycéens par pays

Une variable est particulièrement intéressante pour répondre à la problématique posée.

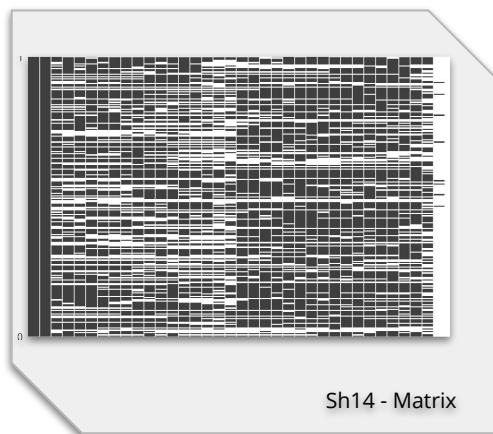
C'est évidemment le nombre de lycéen en présence dans chaque pays, et donc le potentiel de client.

On constate que les données sont viables (sh14) avec "seulement" 51% de cellules vides (sh11) et 2,5% de cellules dupliqués.

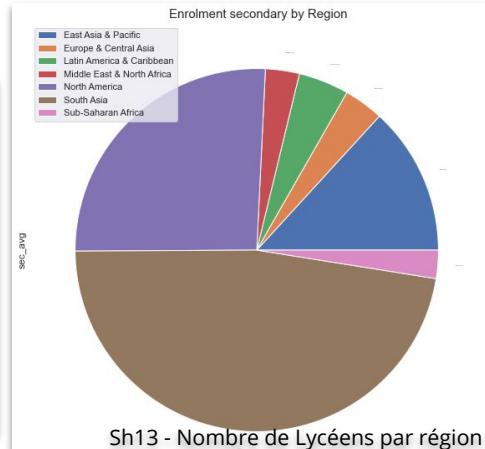
Au vue des courbes (sh10 et s13) et de la disparité mondiale, et régionale, c'est bien une donnée essentielle dans notre étude.



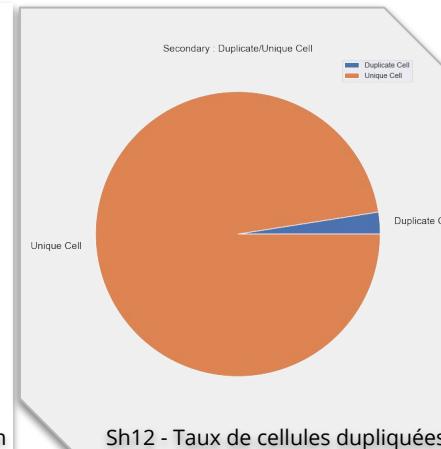
Sh10 - Nombre de Lycéens par pays



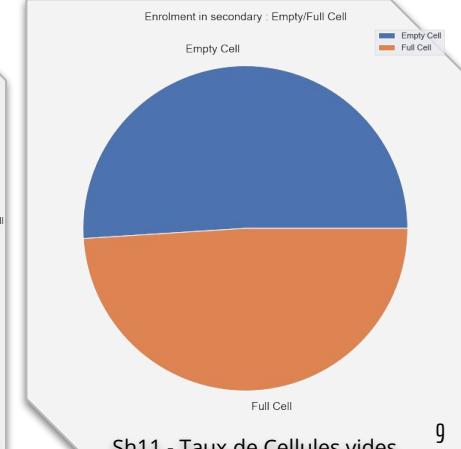
Sh14 - Matrix



Sh13 - Nombre de Lycéens par région



Sh12 - Taux de cellules dupliquées



Sh11 - Taux de Cellules vides

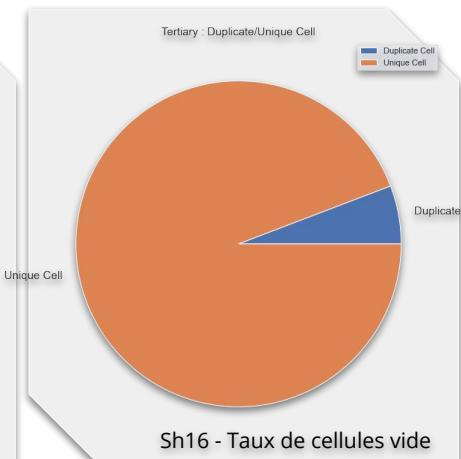
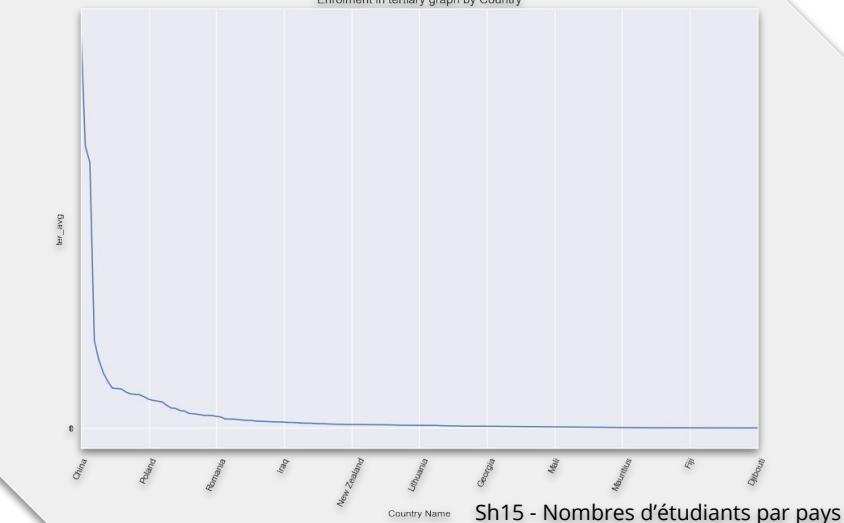
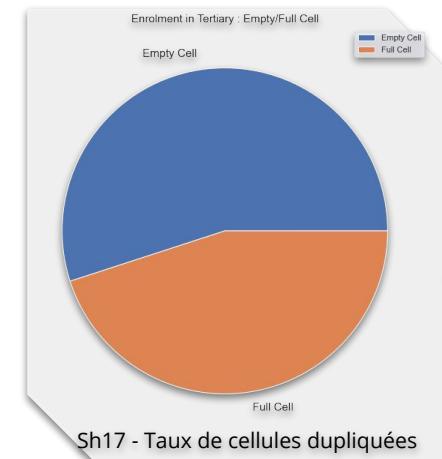
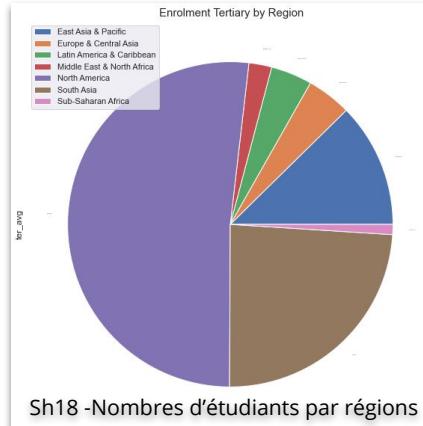
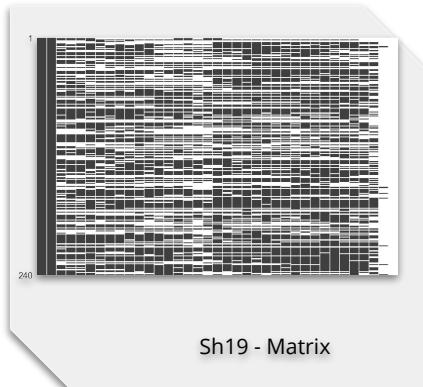
8 - Variable 5 : Nombre d'étudiants par pays

L'alter-égo de la variable précédente est le nombre d'étudiant, car il s'agit la des deux populations visées par notre société.

Les données sont globalement moins viable que le nombre de lycéens mais semble tout de même intéressante à étudier (sh19) avec 55% de cellules vides et 5% seulement de duplication.

En miroir avec la population lycéennes on constate une grande disparité mondiale (sh15) et régionale (sh18).

Il s'agit donc à notre avis, d'une donnée de première ordre.

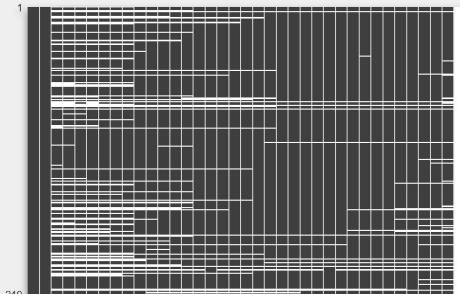


9 - Variable 6 : PIB en \$ par habitant

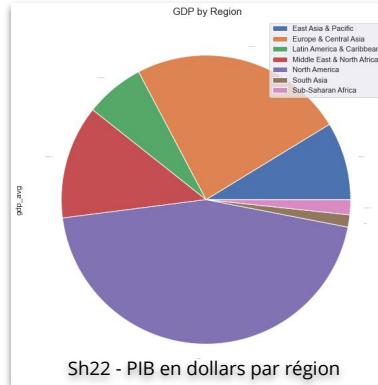
Le pouvoir d'achat est une variable très importante pour pouvoir proposer des cours à un public qui possèdent les moyens financiers, et le cas échéant adapter ses tarifs.

Des données sont toujours viables (sh23), avec seulement 40% de données vides (sh21) et 3.3% de données dupliquées (sh20).

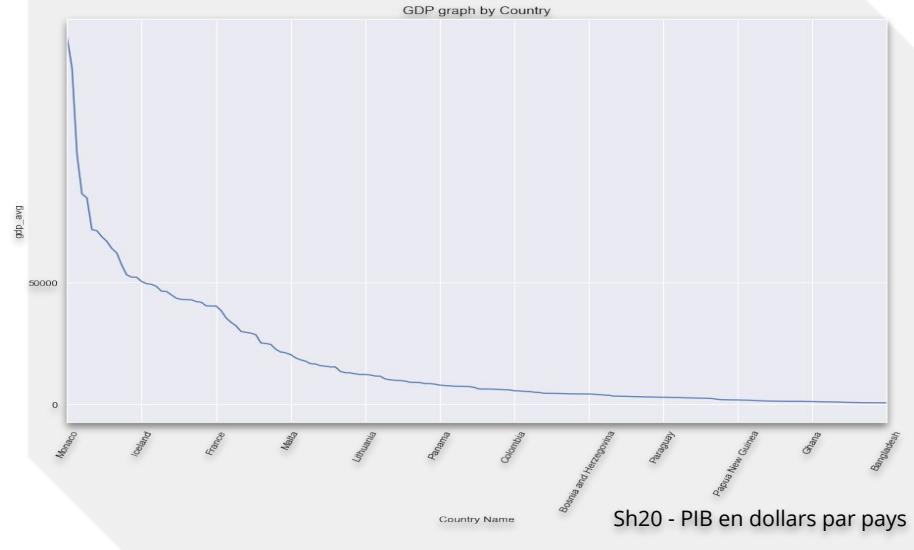
On constate (sh19 et sh22) que cette variable sera essentielle à notre étude.



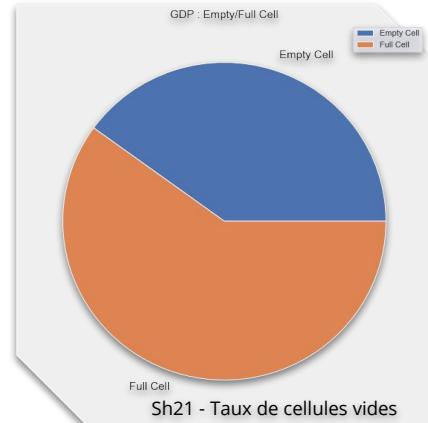
Sh23 - Matrix



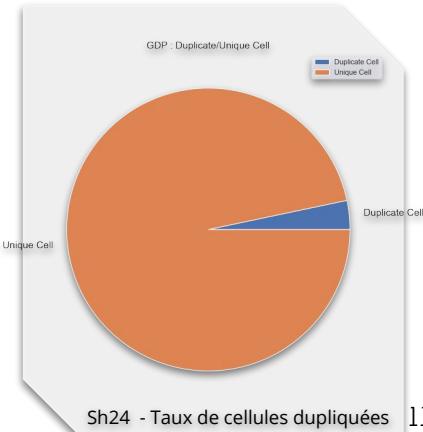
Sh22 - PIB en dollars par région



Sh20 - PIB en dollars par pays



Sh21 - Taux de cellules vides



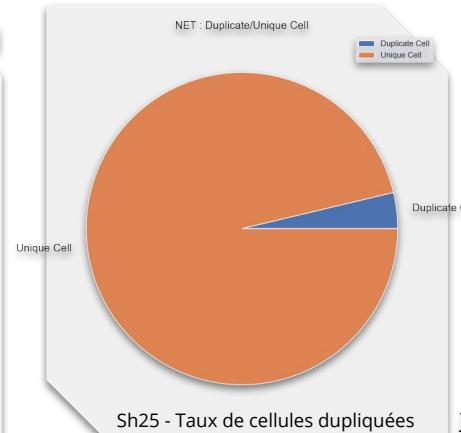
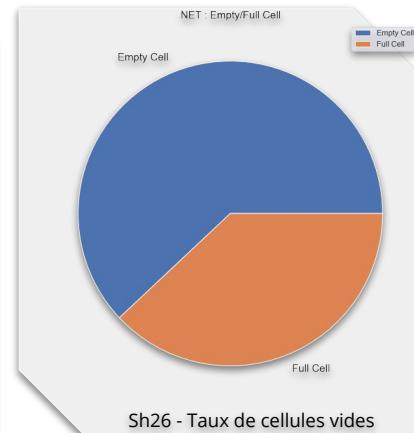
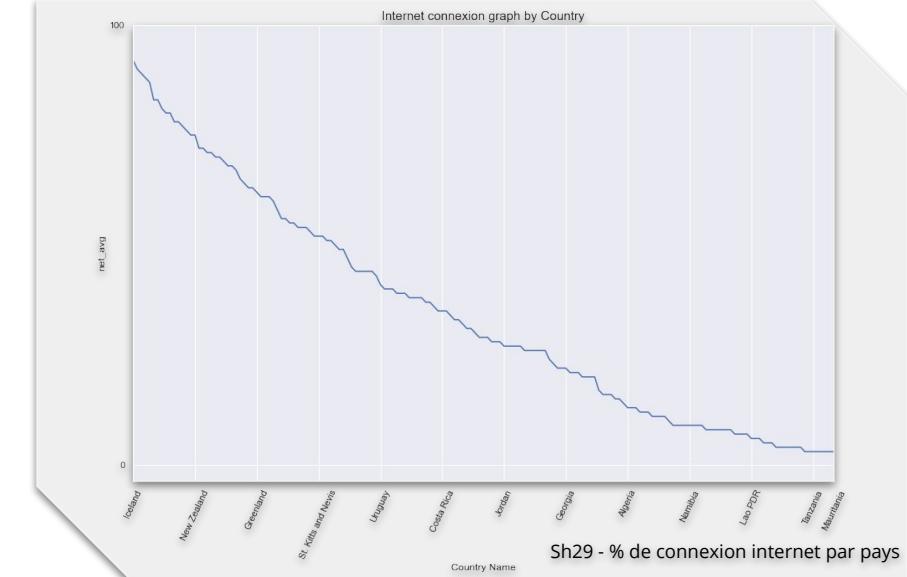
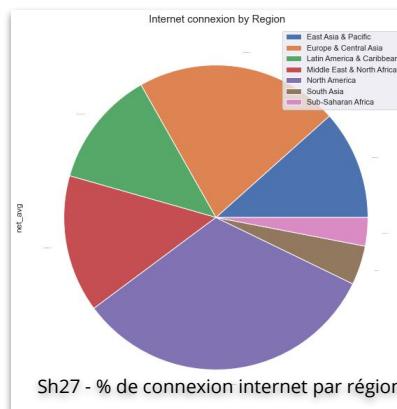
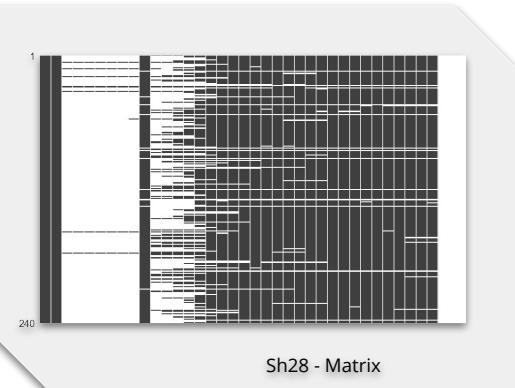
Sh24 - Taux de cellules dupliquées 11

10 - Variable 7 : Connexion internet du pays

Suivre des cours en ligne nécessite évidemment le matériel nécessaire, et donc une connexion internet.

Etudions la viabilité des résultats, 4% des cellules sont dupliquées et 62% des données sont vides, néanmoins on constate (sh28) que les données n'existent qu'à partir des années 90 (naissance de la technologie).

Les courbes (sh24 & sh27) indiquent que les données seront pertinente à un niveau national et régional.



11- Les Filtres

11.1 - Colonnes inutilisées

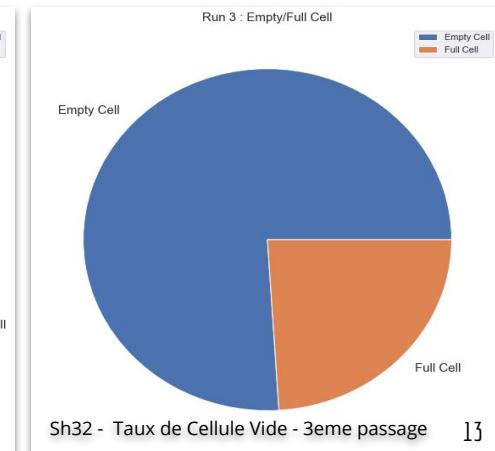
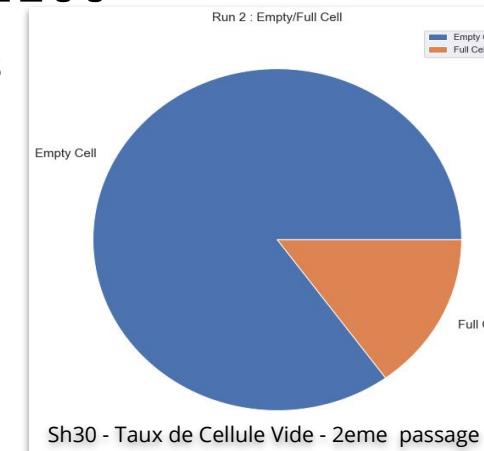
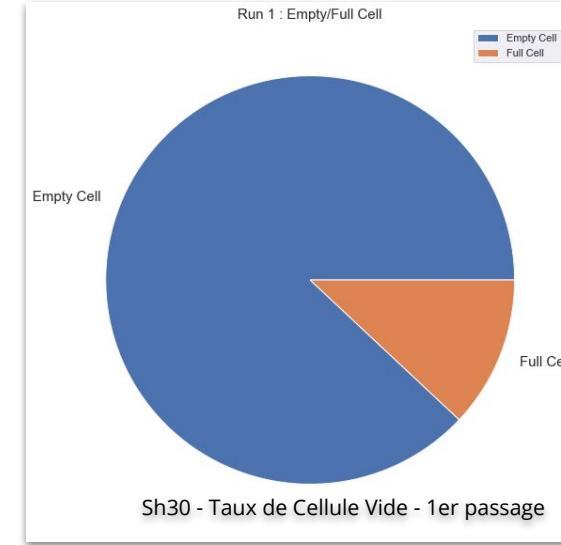
Nous partons d'un taux de 86% de cellule vides comme évoqué précédemment. Dans un premier temps nous retirons les colonnes inutiles, le doublons intitulés "Code" country et Indicator, ainsi que la dernière colonne non-utilisée. Ces cellules étant très fourni notre taux passe à 88% (sh29)

11.2 Année 1970-2004 / 2020-2100

Nous avons vu lors de l'analyse des données que la période la plus intéressante à étudier s'étant de 2004 à 2014.

On supprime donc les années 1970 à 2003, le taux de cellule vide passe à 85%.

Puis les années 2005-2100, les données commencent à s'affiner autour de 78%



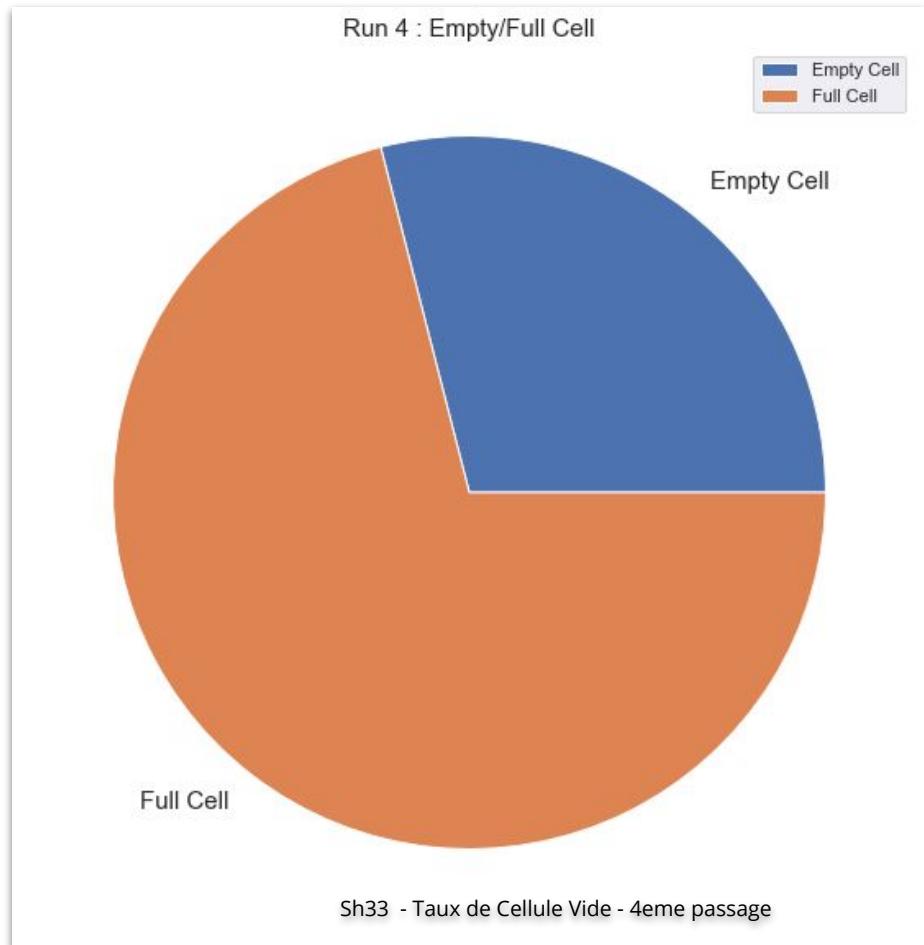
11.3 Variables intéressantes

Nous avons mis en lumière, 4 indicateurs pertinents pour mener à bien notre étude :

- _ Nombre de Lycéens par pays
- _ Nombre d'étudiant par pays
- _ PIB en dollars par pays
- _ Taux de connexion à internet par pays

Nous pouvons donc à ce stade éliminer les autres données (très nombreuses, environ 3500) pour se centrer sur ces indicateurs spécifiques.

Notre taux de cellules vides fond à 29%.



11.4 Dernière Valeur disponible

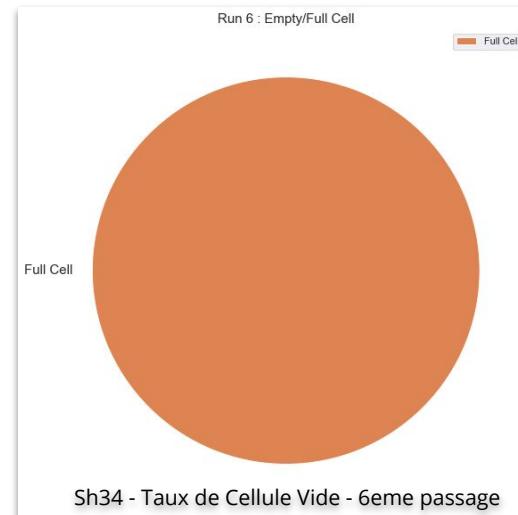
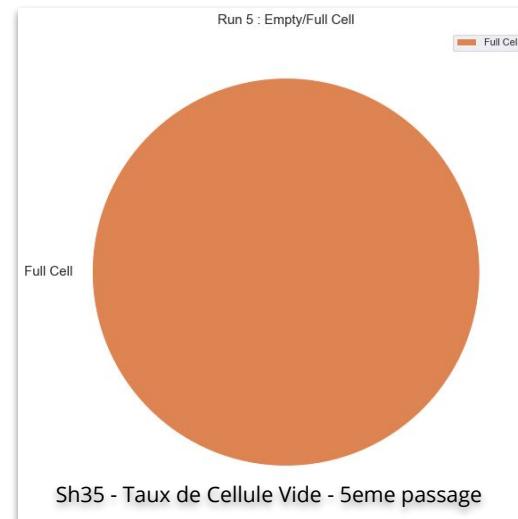
Chaque variables possèdent désormais 10 années de données, néanmoins toutes les cellules ne sont pas renseigné. Nous avons donc gardé seulement la dernière valeur disponible pour chaque variable.

Notre taux de cellules vides passe enfin à 0%

11.5 Supprimer les années

On profite de ce travail pour faire un peu de ménage et supprimer les colonnes d'années qui ne sont désormais plus nécessaire, car nous avons créé une nouvelle donnée avec la dernière valeur connue.

Notre taux de cellules vides est toujours de
0%

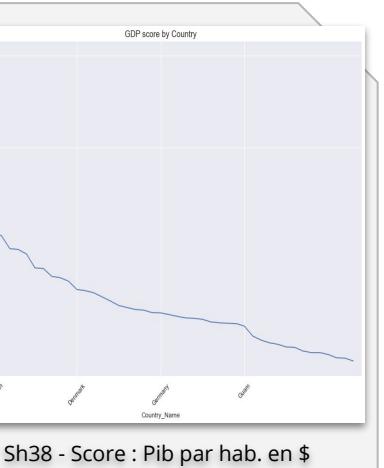
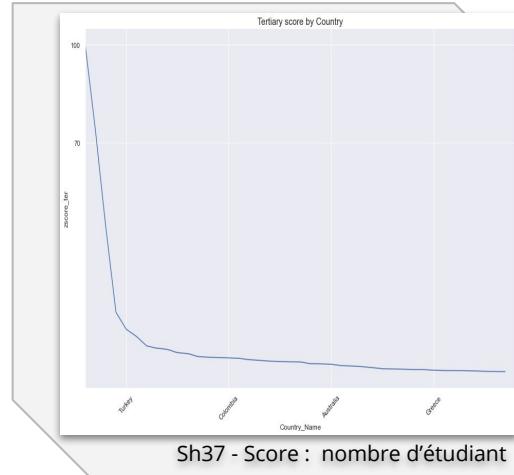
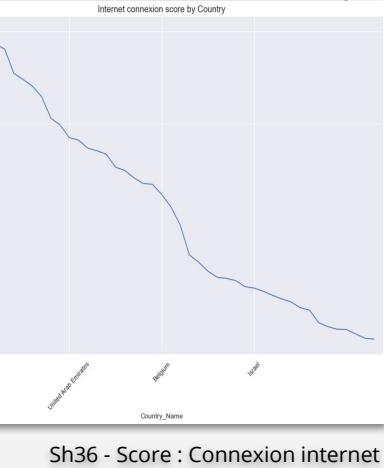
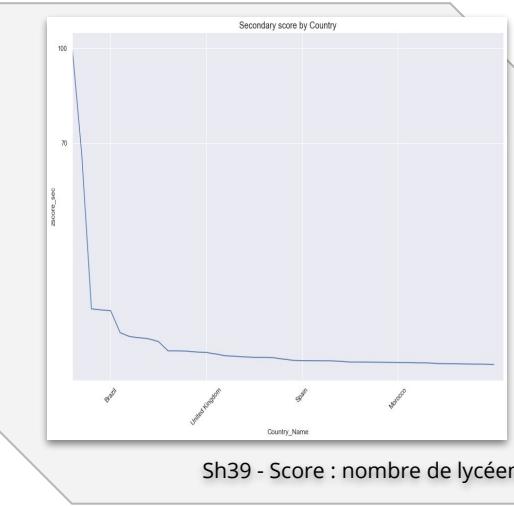


12 - Scoring

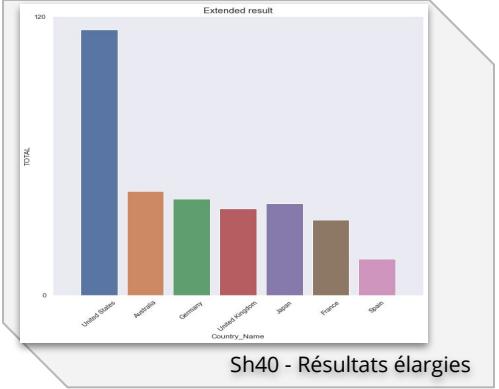
A partir des données fiables trouvées précédemment, on établit pour chaque valeur de l'échantillon un "Score" (déterminé à partir de la moyenne et de l'écart-type).

Sur les 4 schémas ci-joint (sh35, sh36, sh37 & sh38), on constate les différents scores pour chaque pays en fonction de sa variable.

Nous pouvons désormais combiner ses scores et déterminer un classement des pays intéressants



13 - Results



Nous arrivons enfin à l'établissement d'un classement des pays les plus intéressants, qui combinent donc 4 variables :

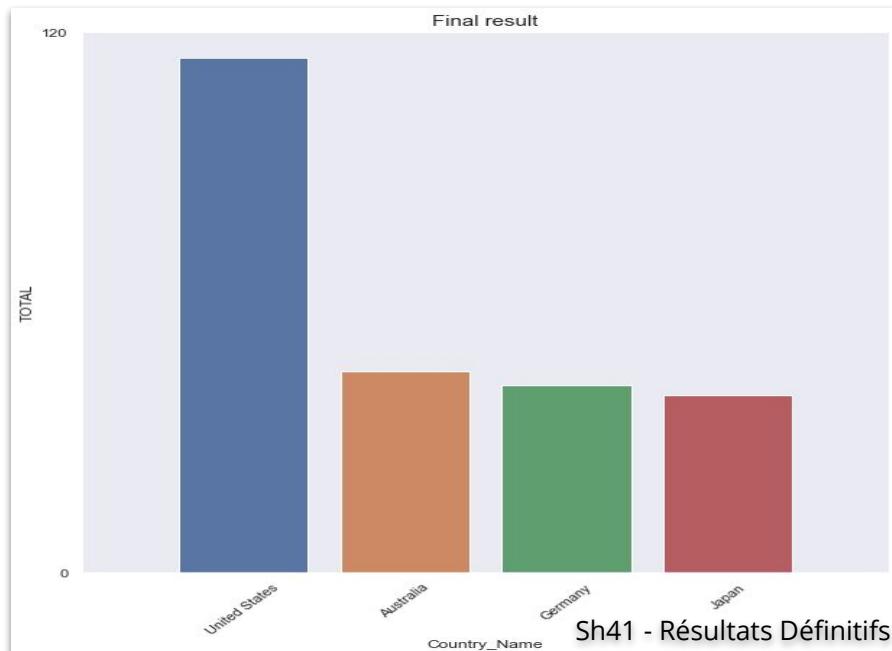
- _ Nombre de lycéens
- _ Nombre d'étudiant
- _ PIB par habitant en dollars
- _ Taux de connexion à internet

Ces différentes étapes nous amène à déterminer 7 pays attrayants (sh39), parmis lesquelles 4 nous semblent essentielles :

**Les Etats Unis
L'Australie
L'Allemagne
Le Japon**

Sont les etats qui répondent parfaitement aux critères que nous avons établis.

Le détails des variables pour ces 4 pays sont disponibles en annexe)



14 -Result : Région

Parallèlement à notre classement des pays, nous pouvons établir une hiérarchie des Régions ou concentrer des efforts pour les prochaines années :

L'Amérique du Nord

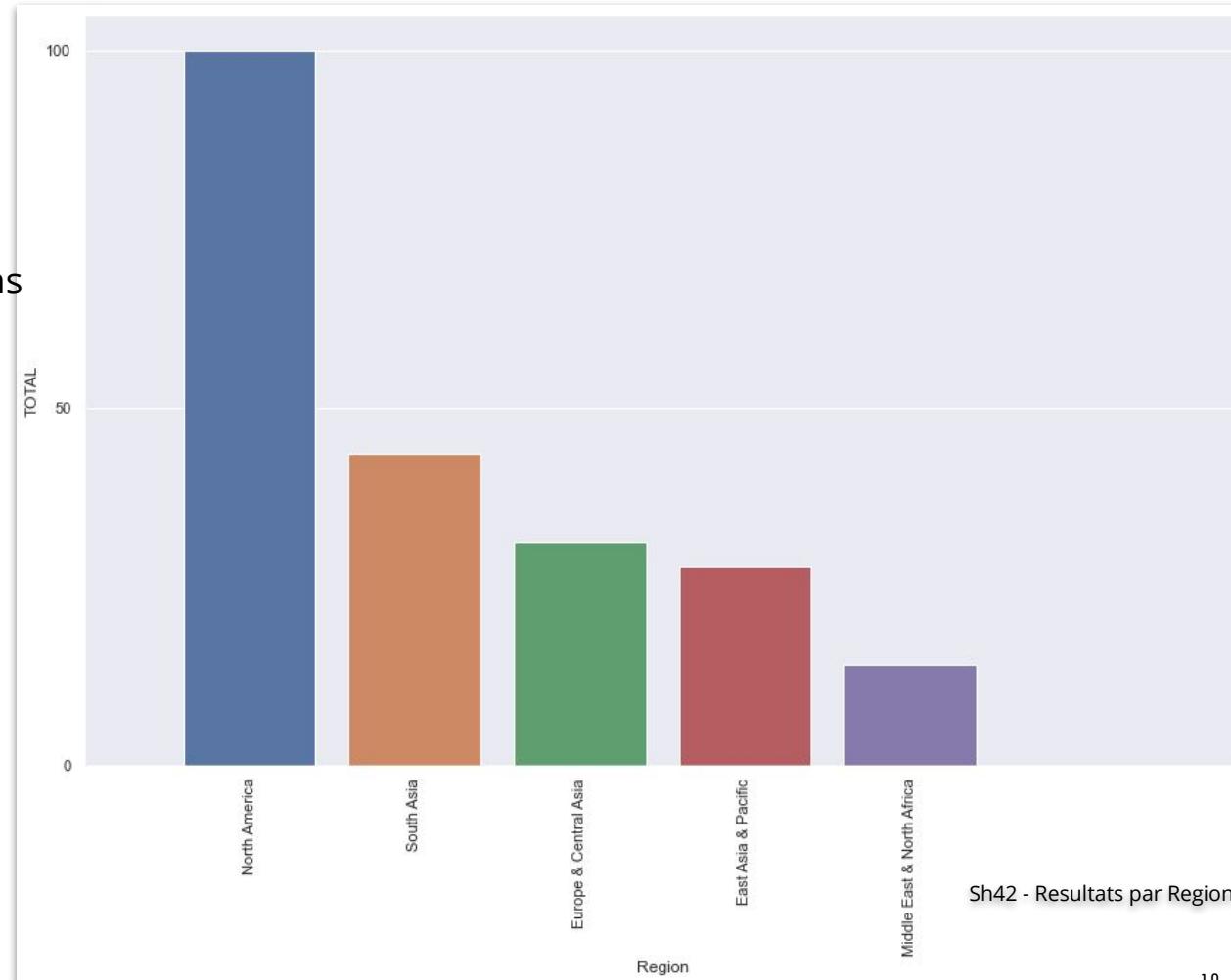
L'Asie du Sud

L'Europe (et l'Asie Central)

L'Asie du Sud (et le Pacifique)

Le Moyen Orient (et l'Afrique du nord)

Ces chiffres peuvent être utile pour une seconde analyse plus précise des ces régions



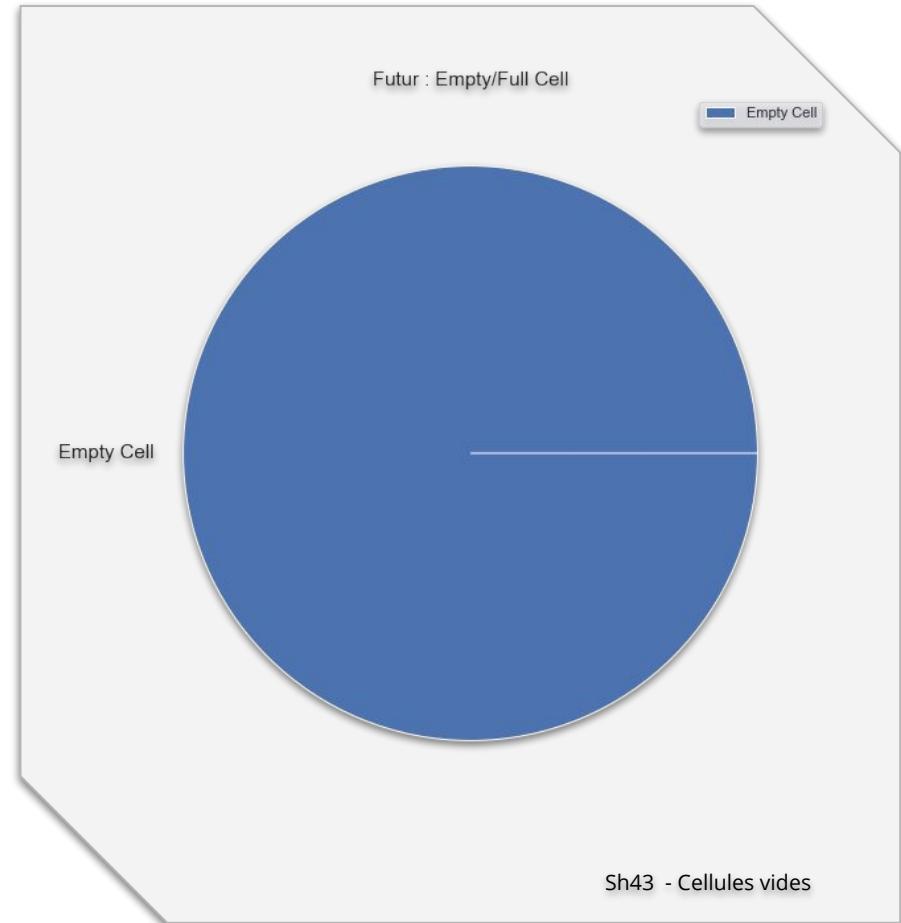
15 - Perspective

Dans le cadre de ce travail, nous voulions étudier les perspectives possibles à long ou moyen terme.

Malheureusement les données à disposition ne nous donnent pas les informations qui nous permettent de prévoir le futur

En effet les variables présentes une totalité de cellules vides concernant les années 2025-2100.

Ce travail peut faire l'objet d'une nouvelle étude...

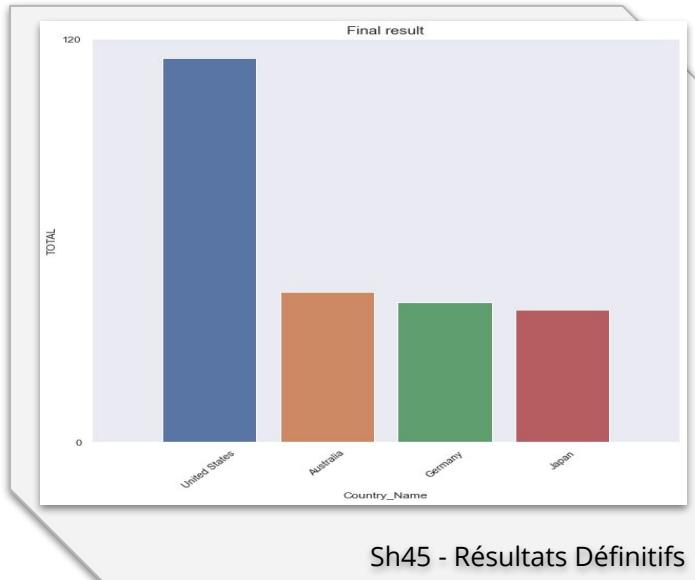


16 - Conclusion

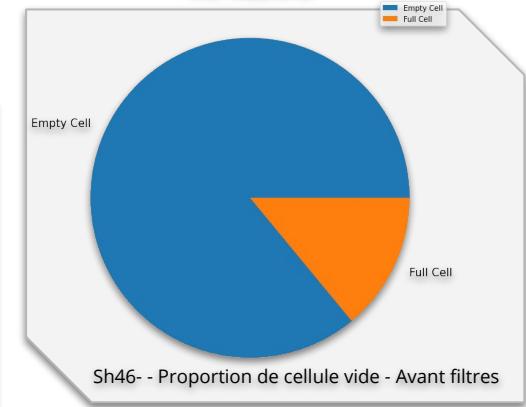
Pour finir, nous pouvons attester de la viabilité de ces données pourtant très incomplète sur beaucoup de variable (sh43)

Avec des filtres adaptés et une progression pas à pas, nous arrivons à une proportion nulle de cellules vides (sh44), qui nous permet d'analyser sereinement les données qui nous concerne particulièrement. A l'issue de travail, 4 pays se détachent : **Etat-Unis, Australie, Allemagne et le Japon.** Ces pays possèdent tous les paramètres qui nous permettent d'envisager un développement.

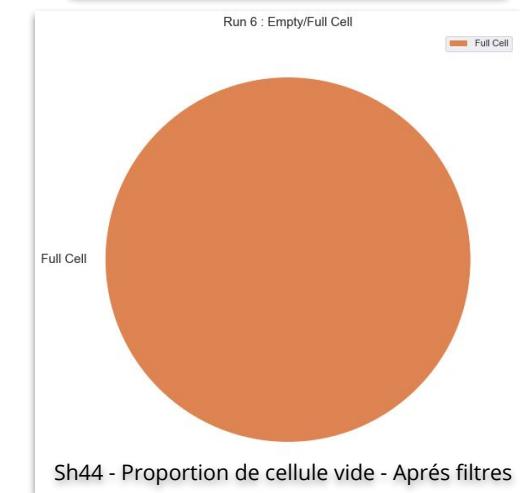
Néanmoins j'attire votre attention sur le fait que ces données sont largement incomplète depuis 2014. Une étude plus récente serait la bienvenue pour attester des ces résultats.



Sh45 - Résultats Définitifs



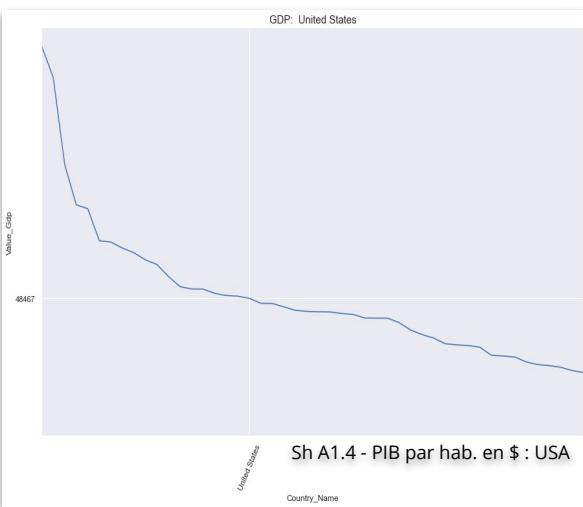
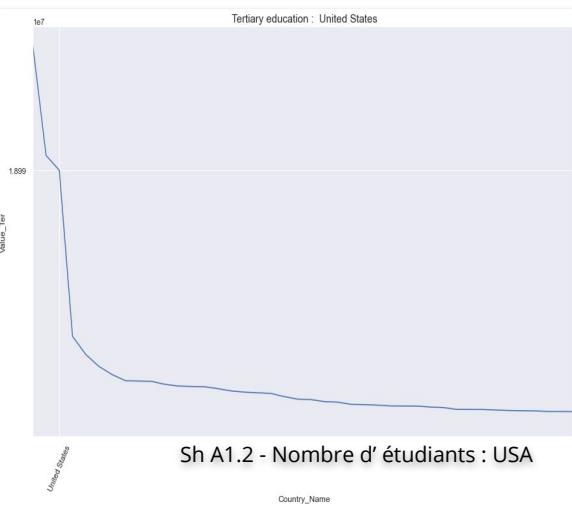
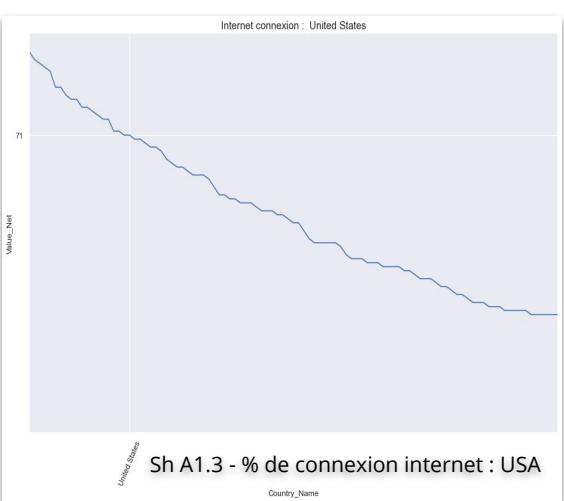
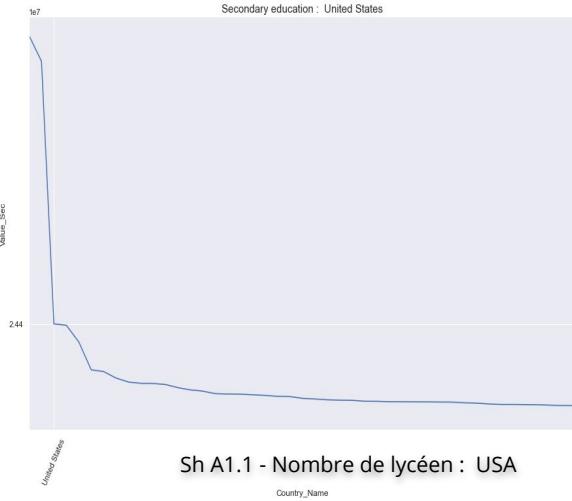
Sh46 - Proportion de cellule vide - Avant filtres



Sh44 - Proportion de cellule vide - Après filtres

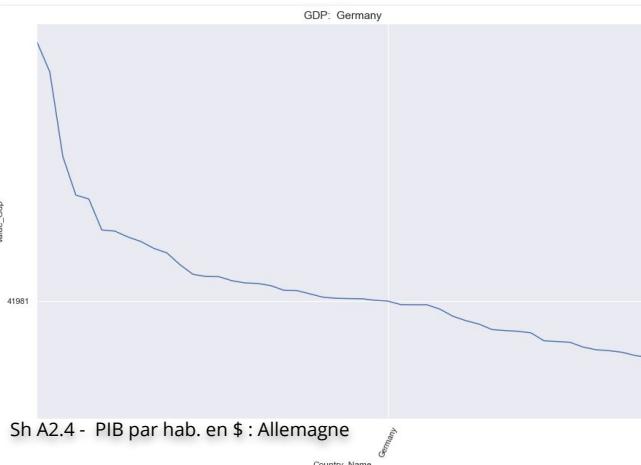
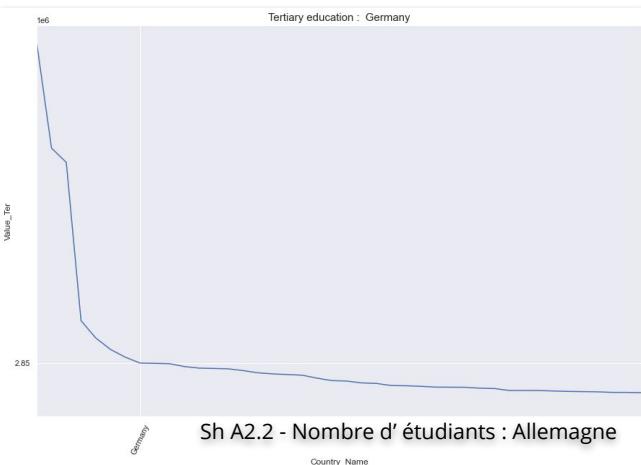
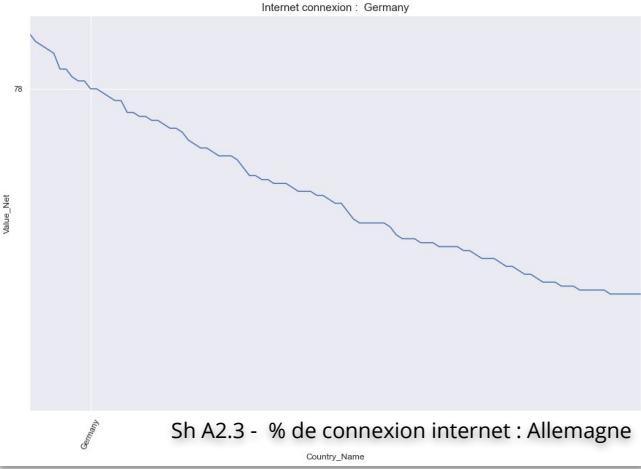
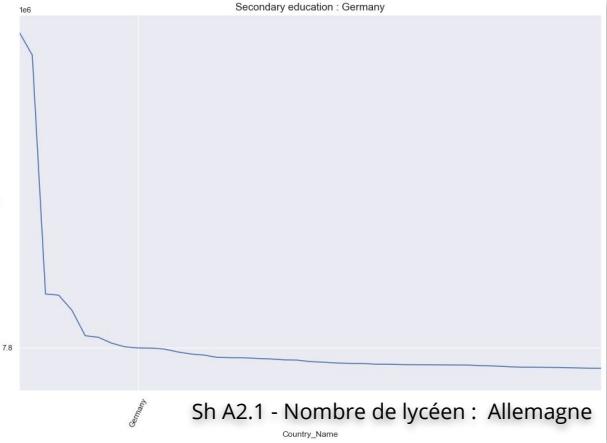
Annexe 1

Détails des données Etat-Unis



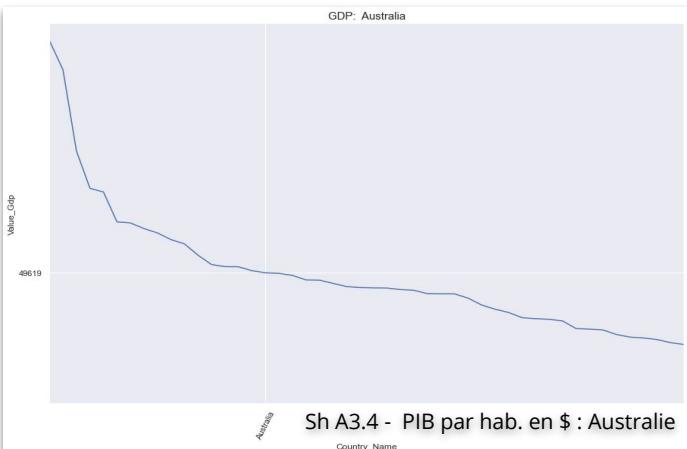
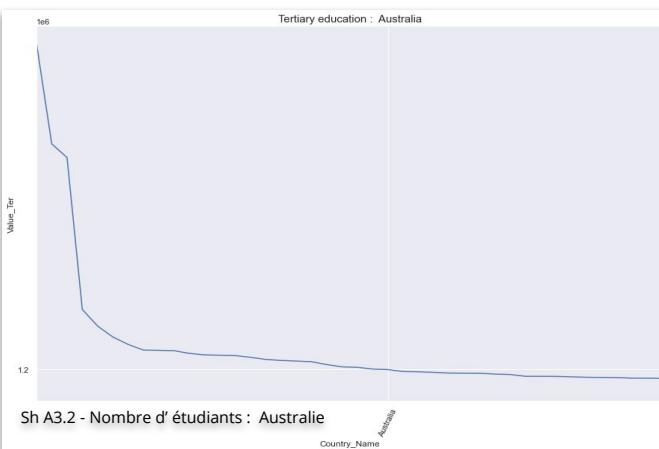
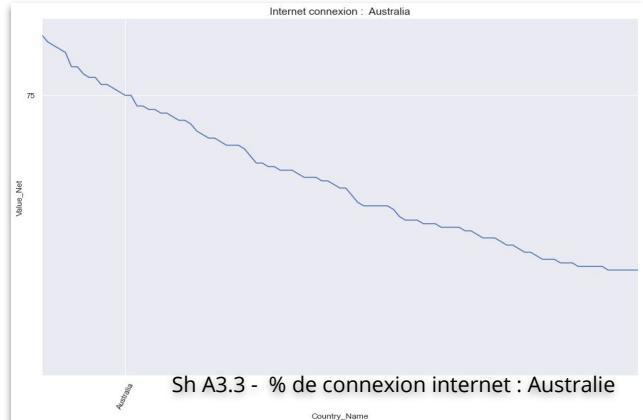
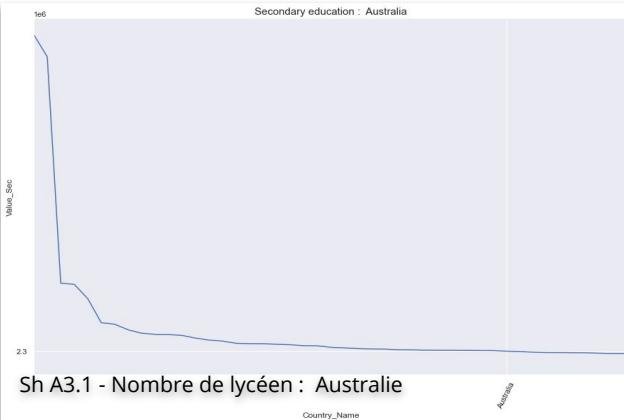
Annexe 2

Détails des données Allemagne



Annexe 3

Détails des données Australie



Annexe 4

Détails des données Japon

