

Analyse exploratoire pour une entreprise de la EdTech

Le Besoin

Entreprise de la EdTech

- Cours en Ligne
- Cible: 15-24ans
- Catalogue varié
- Flexibilité pour ses clients

Volonté d'expansion

- Internationale
- Plus de bénéfice
- Plus de polyvalence

Problème






- Diversité des régions
- Faire le bon choix pour une entreprise
- MONEY

Ressources & méthode de pensée

Plusieurs Datasets

Quel dataset prioriser?

Quels indicateurs prioriser?

Nom	Type	Taille
 EdStatsCountry	Fichier CSV	137 Ko
 EdStatsCountry-Series	Fichier CSV	48 Ko
 EdStatsData	Fichier CSV	318 776 Ko
 EdStatsFootNote	Fichier CSV	38 781 Ko
 EdStatsSeries	Fichier CSV	3 620 Ko

- Donnée utile à la question?
- Indicateurs pouvant être mis en relation?

- Quels indicateurs répondent à la question?
- Approche naïve?
 - Population
 - Etudes & Diplômes
 - MONEY

Il ne semble pas y avoir de doublons dans data
Le jeu de données data comporte 886930 lignes pour 70 colonnes

Il ne semble pas y avoir de doublons dans data_country
Le jeu de données data_country comporte 241 lignes pour 32 colonnes

Il ne semble pas y avoir de doublons dans data_country_s
Le jeu de données data_country_s comporte 613 lignes pour 4 colonnes

Il ne semble pas y avoir de doublons dans data_footnote
Le jeu de données data_footnote comporte 643638 lignes pour 5 colonnes

Il ne semble pas y avoir de doublons dans data_series
Le jeu de données data_series comporte 3665 lignes pour 21 colonnes

Etudier les datasets

Comprendre ce qu'on a à l'intérieur:

- Beaucoup de donnée?
- Présente? Manquante?
- Eparses?
- Constante? Régulière?
- Organisée?
- Optimisée?
- Quels indicateurs?

Créer un jeu d'étude propre

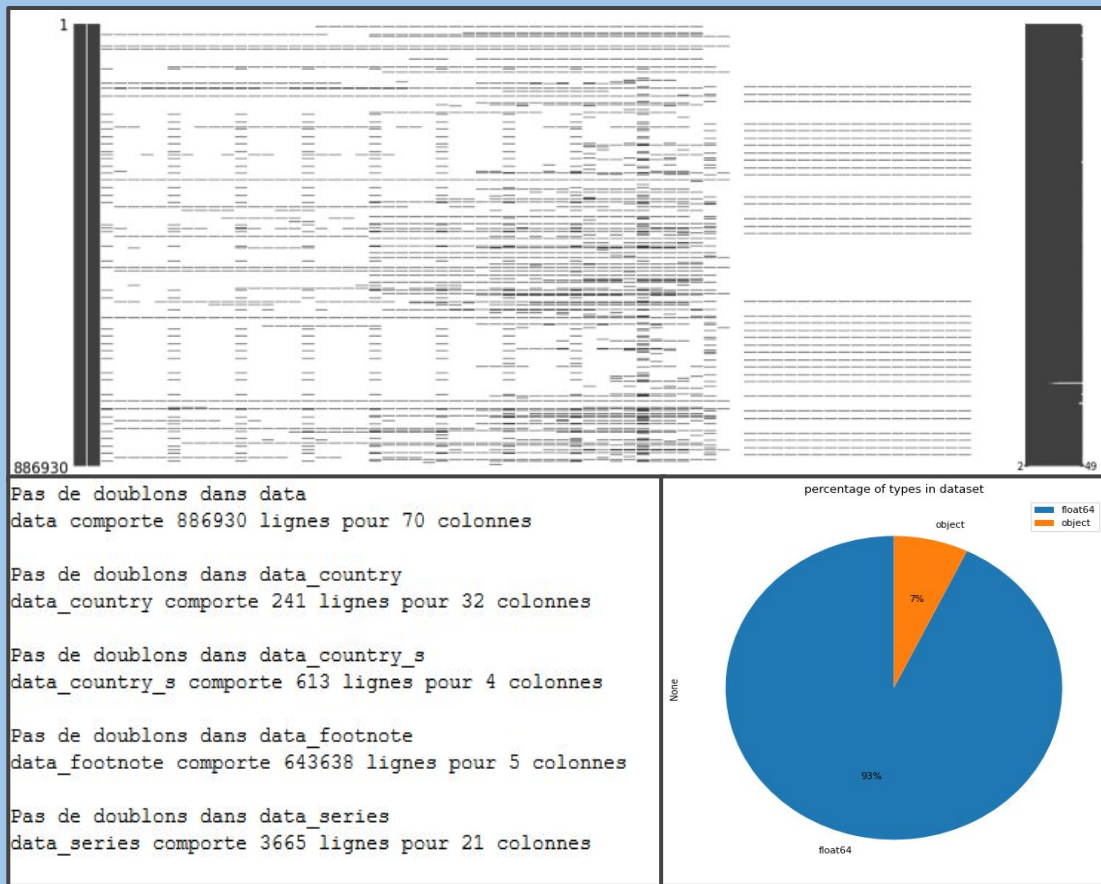
- Indicateurs clairs
- Donnée organisée
- Gestion des données manquantes
- Gestion des indicateurs

Explorer et Scorer

- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation

Comprendre ce qu'on a à l'intérieur:

- Beaucoup de donnée?
- Présente? Manquante?
- Eparses?
- Constante? Régulière?
- Organisée?
- Optimisée?
- Quels indicateurs?



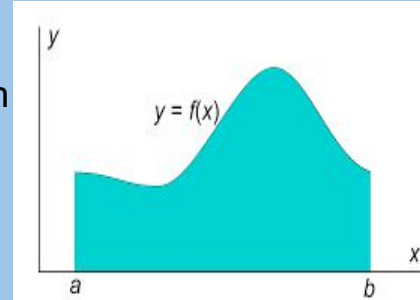
Créer un jeu d'étude propre

- Indicateurs clairs
- Donnée organisée
- Gestion des données manquantes
- Gestion des indicateurs

Les indicateurs retenus et leur utilisation

→ Population

- ◆ "Projection: Population age 15-19&20-24 in thousands by highest level of educational attainment"
 - *Lower/Upper/Post Secondary*
 - aire sous la courbe



→ Capacité à suivre le cours

- ◆ Equipement
 - *Ordinateur personnel*
 - *Accès à internet*
 - dernières valeurs



→ \$MONEY\$

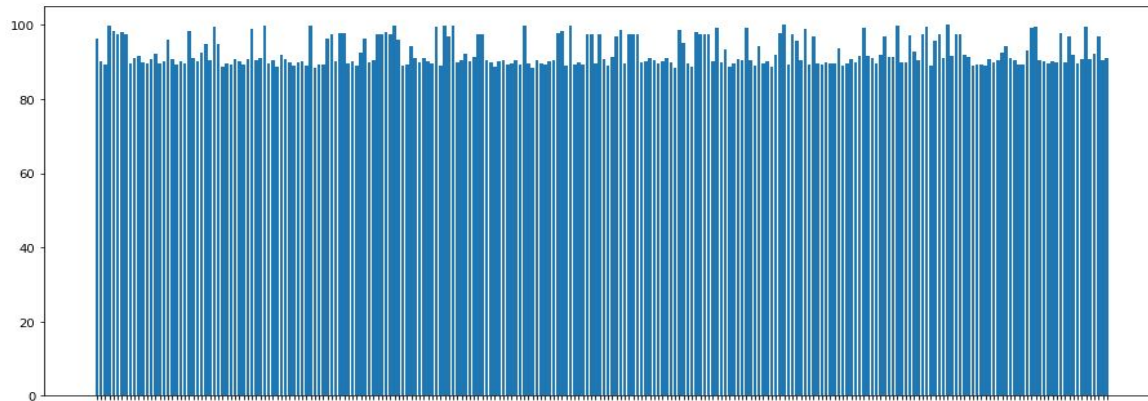
- ◆ *PIB par habitant en \$USD\$*
 - dernière valeur



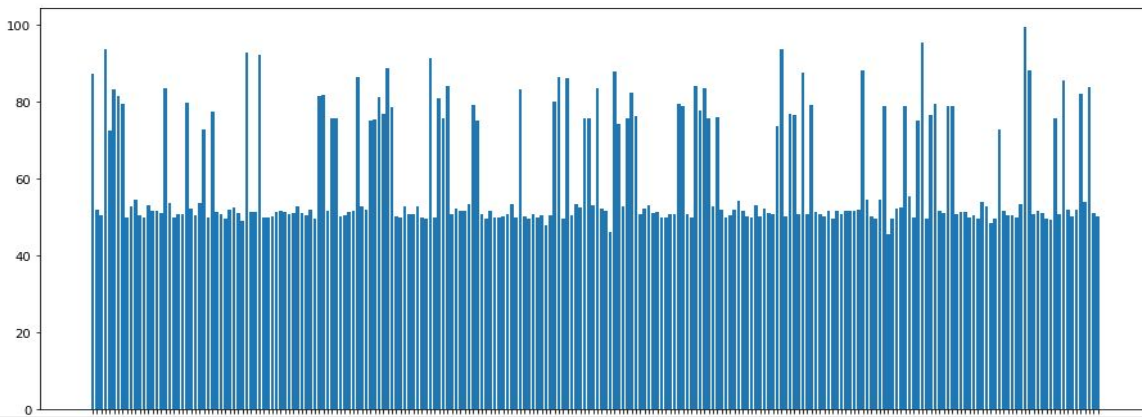
Etudier les datasets

Par Pays

Moyenne de données manquantes par pays: 92.63338842975206
Minimum de données manquantes par pays: 88.45



Moyenne de données manquantes par pays: 59.96792372881356
Minimum de données manquantes par pays: 45.6

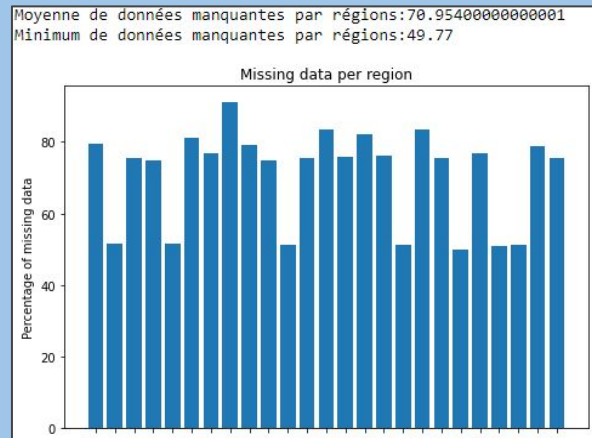
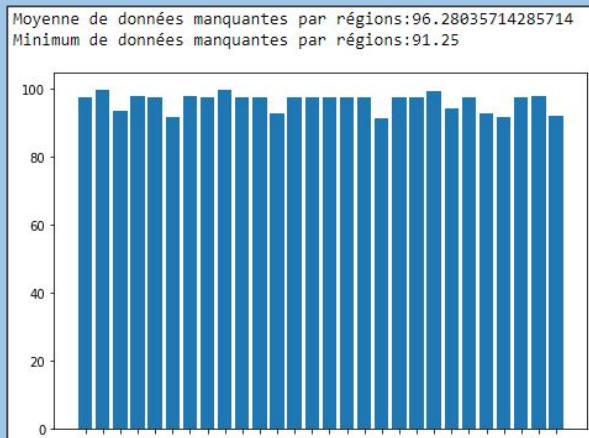


Comprendre ce qu'on a à l'intérieur:

- Beaucoup de donnée?
- Présente? Manquante?
- Eparse?
- Constante? Régulière?
- Organisée?
- Optimisée?
- Quels indicateurs?

Comprendre ce qu'on a à l'intérieur:

- Beaucoup de donnée?
- Présente? Manquante?
- Eparses?
- Constante? Régulière?
- Organisée?
- Optimisée?
- Quels indicateurs?



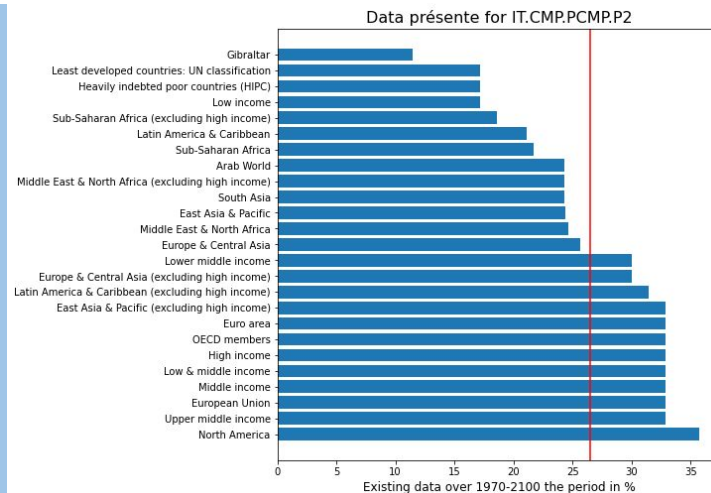
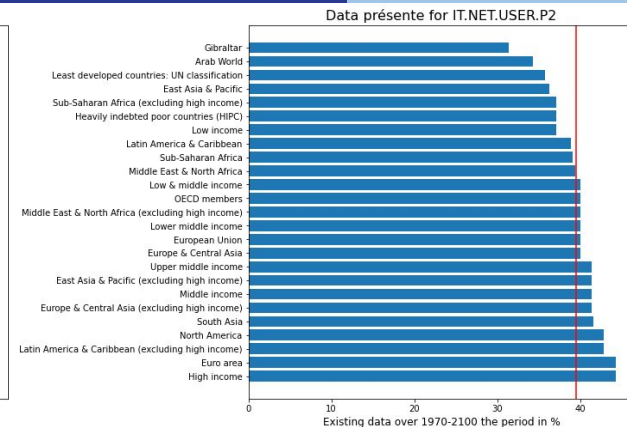
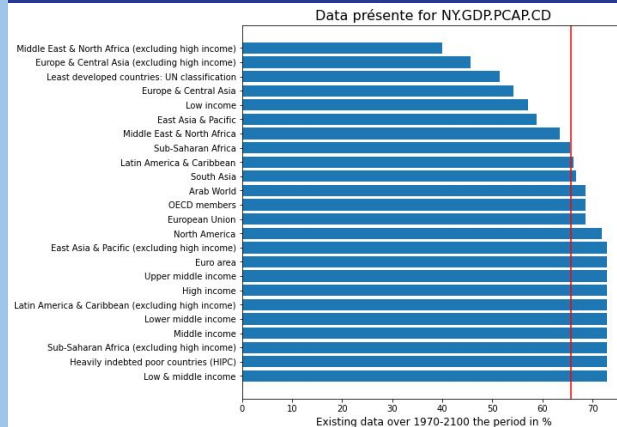
Remarque: cf notebook pour pays. globalement corrélation manque donnée/richesse du pays

Etudier les datasets

Data présente par Région sur les indicateurs choisis

Comprendre ce qu'on a à l'intérieur:

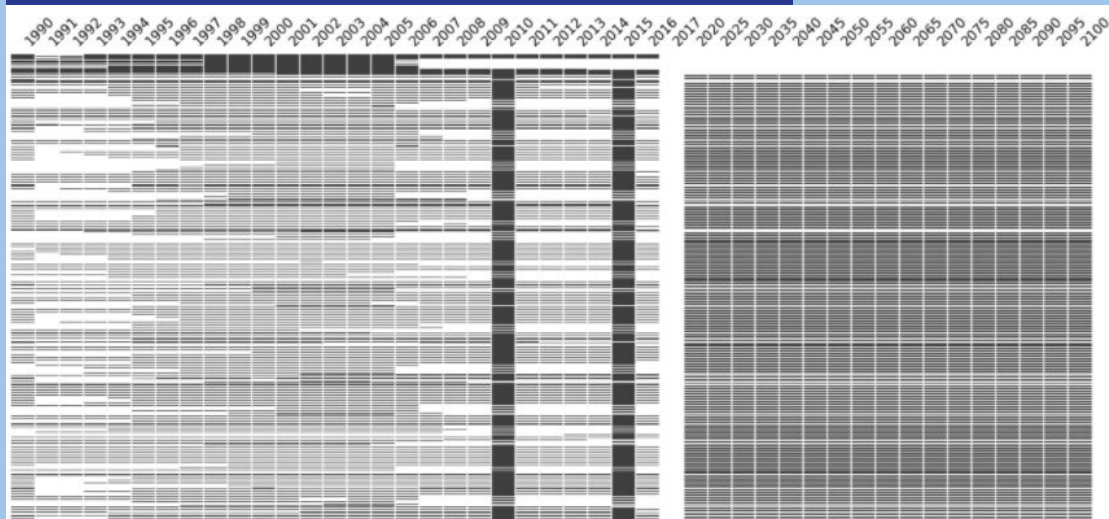
- Beaucoup de donnée?
- Présente? Manquante?
- Eparses?
- Constante? Régulière?
- Organisée?
- Optimisée?
- Quels indicateurs?



Créer un jeu d'étude propre

- Indicateurs clairs
- Donnée organisée
- Gestion des données manquantes
- Gestion des indicateurs

Après nettoyage



Les méthodes:

- *drop* de valeurs manquantes
- *drop* d'indicateurs ou régions inutiles:
 - *skip* (na, sum=0,...)
- utilisation des codes plutôt que du texte

Remarque: La forme de la donnée

```
#On nettoie la donnée tout d'abord
#Je devrais aussi faire un nettoyage par indicateur avec drop d'indicateurs projection
#n'ayant pas de valeur sur les années>=2020 par exemple
df_1 = pd.merge(data.iloc[:,0:2], data_country.iloc[:,[0,7]],
                 left_on="Country Code", right_on="Country Code", how="left")
df_1=df_1.join(data.iloc[:,2:])
df_1["Region"] = df_1.apply(lambda x : country_to_region(x),axis=1)
df_1 = df_1.drop(columns=df_1.columns[-1])
#Maintenant que nous avons les indicateurs que nous voulons nous pouvons conserver
#uniquement leur codes
df_1 = df_1.drop(columns="Indicator Name")
#On peut aussi drop la région world qui est un peu large
df_1 = df_1.drop(df_1[df_1["Country Name"]=="World"].index)
#Puis on peut conserver uniquement les indicateurs qui nous intéressent
df_1 = df_1[df_1["Indicator Code"].isin(l_tot)]
#Et enfin drop les colonnes totalement vides
df_1 = df_1[df_1.iloc[:,4:].notna().sum(axis=1)>0]
```

- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation

→ Population

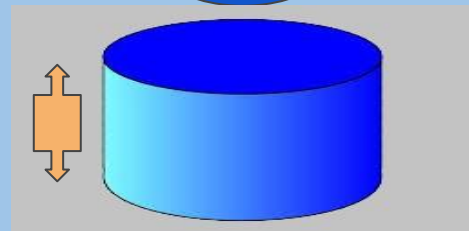
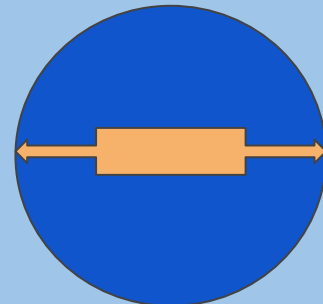
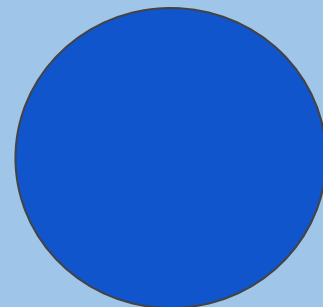
- ◆ Aire sous la courbe des indicateur de population:
 - *Lower/Upper/Post Secondary*
 - aire sous la courbe

→ Equipement

- ◆ Equipement
 - *Ordinateur personnel*
 - *Accès à internet*
 - min(dernières valeurs)
 - naïf mais pas optimiste

→ GDP

- ◆ *PIB par habitant en \$USD\$*
 - dernière valeur



- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation

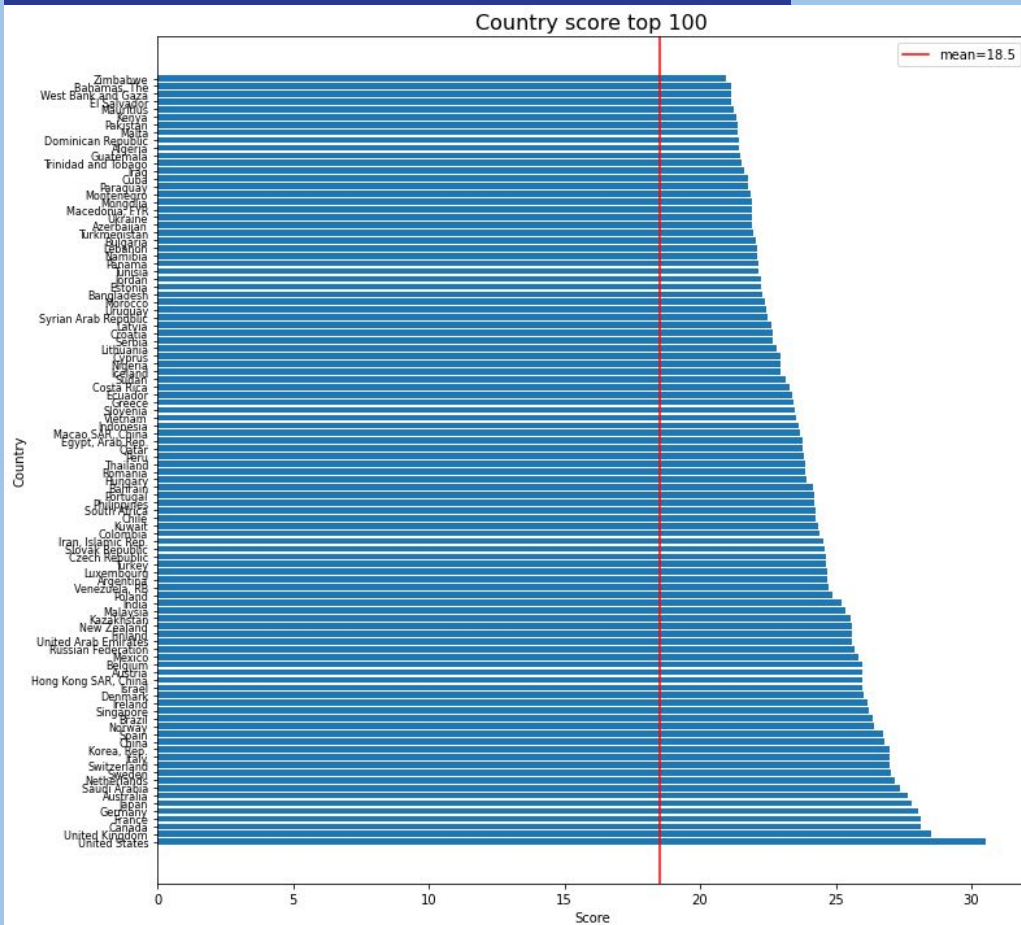
```
#####
def score_that_country_please(d, l_pop, l_eqpmt):
    if d["Indicator Code"].isin(l_pop)["AUC"].size > 0:
        P = d[d["Indicator Code"].isin(l_pop)]["AUC"].sum()
    else:
        P = 1
    if d["Indicator Code"].isin(l_eqpmt[1:3])["last value"].size > 0:
        E = d[d["Indicator Code"].isin(l_eqpmt[1:3])["last value"].min()
    else:
        E = 1
    if d["Indicator Code"].isin(l_eqpmt[0:1])["last value"].size > 0:
        G = d[d["Indicator Code"].isin(l_eqpmt[0:1])["last value"].values[0]
    else:
        G = 1
    if P*E*G in [-1, 0, 1]:
        return(1)
    else:
        return(np.log(P*E*G))
#####
```

```
def score_that_region_please(d, l_pop, l_eqpmt):
    if d["Indicator Code"].isin(l_pop)["AUC"].size > 0:
        P = d[d["Indicator Code"].isin(l_pop)]["AUC"].sum()
    else:
        P = 1
    if d["Indicator Code"].isin(l_eqpmt[1:3])["last value"].size > 0:
        E_1 = d[d["Indicator Code"]==l_eqpmt[1]]["last value"].mean()
        E_2 = d[d["Indicator Code"]==l_eqpmt[2]]["last value"].mean()
        E = min(E_1, E_2)
    #
    else:
        E = 1
    if d["Indicator Code"]==l_eqpmt[0]]["last value"].size > 0:
        G = d[d["Indicator Code"]==l_eqpmt[0]]["last value"].mean()
    else:
        G = 1
    if P*E*G in [-1, 0, 1]:
        return(1)
    else:
        return(np.log(P*E*G))
#####
```

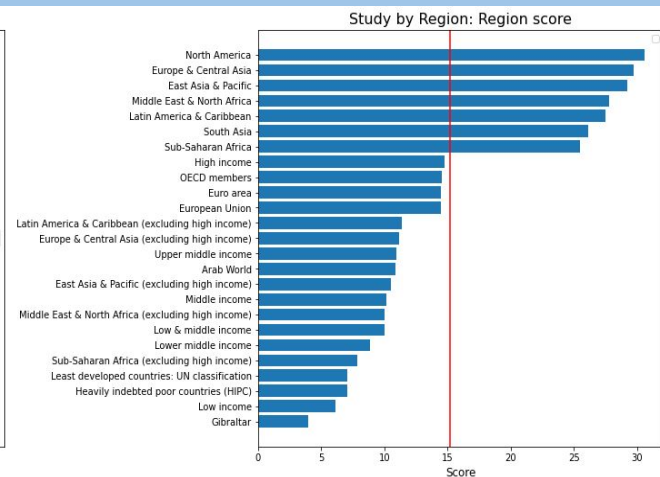
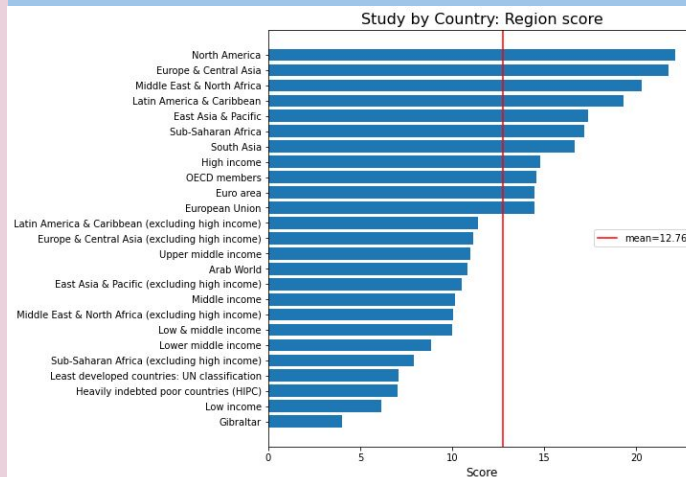
Explorer et Scorer

- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation

Top 100 Pays

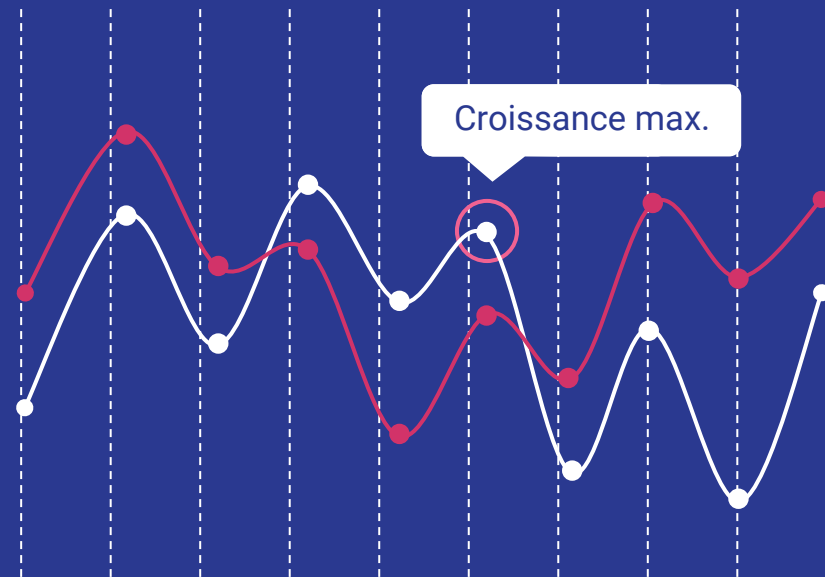


- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation



Les Régions nominées sont:

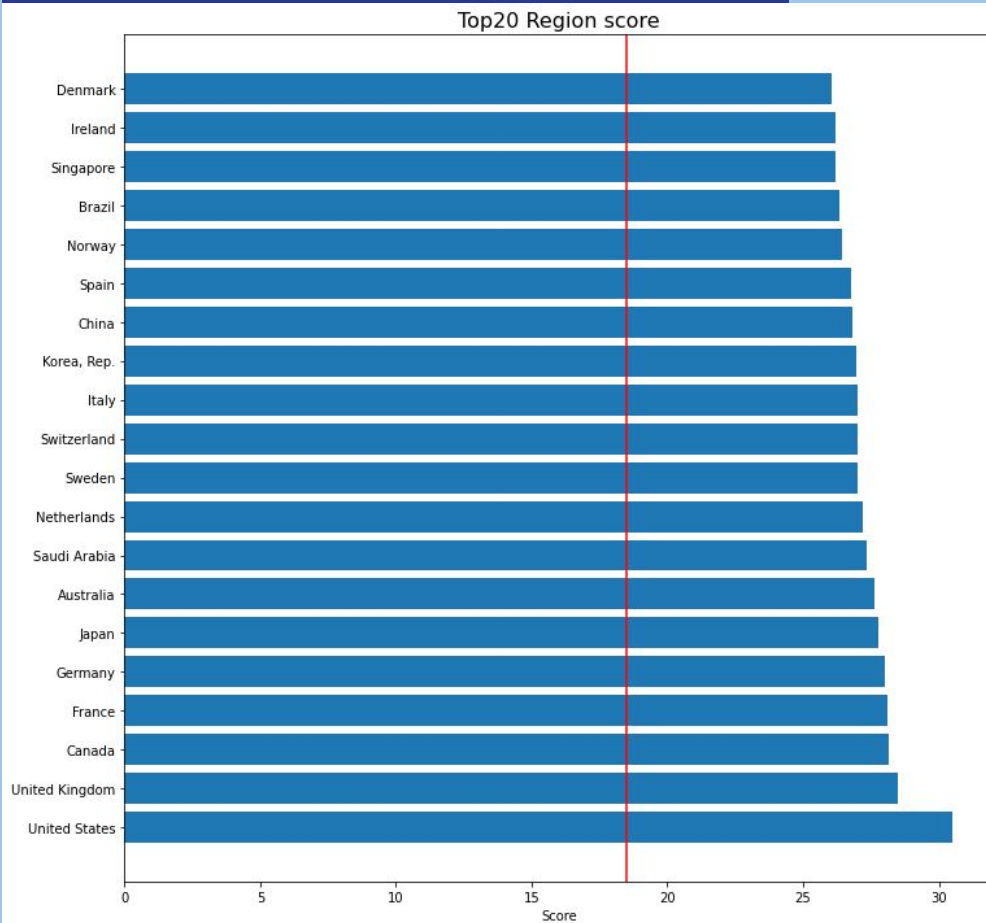
Battement de tambour



Explorer et Scorer

- Indicateurs statistiques
- Score List:
 - Par Pays
 - Par Région
- Interprétation

Top 20 Pays et avis



Apprendre et Recommencer

Critique:

Qu'est-ce que je changerais en fonction de ce que j'ai appris

Dans la région North America l'indicateur IT.CMP.PCMP.P2 a pour valeur médiane 27.36.
Mais aussi un écart-type de 20.58.
Son skewness 0.55 indique que la distribution est plutôt étalée à droite.
Enfin son kurtosis 0.11 indique que la distribution a le même aplatissement que la distribution normale.
En voici une petite boîte à moustache:

