Analyse Exploratoire des Données (EDA) - Banque Mondiale

1. Imports

Nous configurons l'environnement et listons les sources de données brutes (*raw data*). Pour visualiser les données, nous définissons d'abord le chemin d'accès (**path**) où sont stockés les fichiers CSV. Nous récupérons ensuite la liste de ces fichiers dans une variable `all_files` en utilisant un pattern de recherche (**globbing**) avec l'extension `*.csv`.

2. Chargement des données

Boucle d'itération et rendu des données

Cette étape permet de valider l'intégrité des fichiers CSV et d'obtenir un premier aperçu visuel des structures.

- **Identification** : Nous affichons le nom du fichier pour confirmer la lecture.
- **Chargement** : Le contenu est chargé dans un **DataFrame** (structure de données tabulaire).
- **Rendu** : Nous utilisons `display(df.head())` pour générer un rendu visuel (**rendering**) des 5 premières entrées (**entries**).

Gestion des exceptions (Error Handling)

En cas d'erreur lors de la lecture ou de l'affichage, un bloc `try...except` permet de capturer l'exception. Le script affiche alors le nom du fichier problématique ainsi que le message d'erreur associé pour faciliter le débogage.

```
import pandas as pd
import glob
import os
from IPython.display import display, Markdown

path = 'data'
all_files = glob.glob(os.path.join(path , "*.csv"))
dataframes = []

for file in all_files:
    try:
        file_name = os.path.basename(file)
        df = pd.read_csv(file)
        dataframes.append({"name": file_name, "data": df})
```

```
    except Exception as e:
        print(f"Erreur sur {file}: {e}")
```

```
display(Markdown(f"## Analyse du fichier : {dataframes[0]['name']}"))
```

Analyse du fichier : EdStatsCountry.csv

```
print(f"--- Fichier : {dataframes[0]['name']} ---")
display(dataframes[0]['data'].head())
```

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income
2	AGO	Angola	Angola	People's Republic of Angola	AO	Angolan kwanza	April 2013 database update: Based on IMF data,...	Sub-Saharan Africa	Upper middle income
3	ALB	Albania	Albania	Republic of Albania	AL	Albanian lek	NaN	Europe & Central Asia	Upper middle income
4	AND	Andorra	Andorra	Principality of Andorra	AD	Euro	NaN	Europe & Central Asia	High income: nonOECD

5 rows × 32 columns

```
display(Markdown(f"## Analyse du fichier : {dataframes[1]['name']}"))
```

Analyse du fichier : EdStatsCountry-Series.csv

```
print(f"--- Fichier : {dataframes[1]['name']} ---")
display(dataframes[1]['data'].head())
```

	CountryCode	SeriesCode	DESCRIPTION	Unnamed: 3
0	ABW	SP.POP.TOTL	Data sources : United Nations World Population...	NaN
1	ABW	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
2	AFG	SP.POP.GROW	Data sources: United Nations World Population ...	NaN
3	AFG	NY.GDP.PCAP.PP.CD	Estimates are based on regression.	NaN
4	AFG	SP.POP.TOTL	Data sources : United Nations World Population...	NaN

```
display(Markdown(f"## Analyse du fichier : {dataframes[2]['name']}"))
```

Analyse du fichier : EdStatsData.csv

```
print(f"--- Fichier : {dataframes[2]['name']} ---")
display(dataframes[2]['data'].head())
```

	Country Name	Country Code	Indicator Name	Indicator Code	1970	1971	1972	1973	
0	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2	NaN	NaN	NaN	NaN	
1	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.F	NaN	NaN	NaN	NaN	
2	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.GPI	NaN	NaN	NaN	NaN	
3	Arab World	ARB	Adjusted net enrolment rate, lower secondary, ...	UIS.NERA.2.M	NaN	NaN	NaN	NaN	
4	Arab World	ARB	Adjusted net enrolment rate, primary, both sex...	SE.PRM.TENR	54.822121	54.894138	56.209438	57.267109	57.

5 rows × 70 columns

```
display(Markdown(f"## Analyse du fichier : {dataframes[3]['name']}"))
```

Analyse du fichier : EdStatsFootNote.csv

```
print(f"--- Fichier : {dataframes[3]['name']} ---")
display(dataframes[3]['data'].head())
```

	CountryCode	SeriesCode	Year	DESCRIPTION	Unnamed: 4
0	ABW	SE.PRE.ENRL.FE	YR2001	Country estimation.	NaN
1	ABW	SE.TER.TCHR.FE	YR2005	Country estimation.	NaN
2	ABW	SE.PRE.TCHR.FE	YR2000	Country estimation.	NaN
3	ABW	SE.SEC.ENRL.GC	YR2004	Country estimation.	NaN
4	ABW	SE.PRE.TCHR	YR2006	Country estimation.	NaN

```
display(Markdown(f"## Analyse du fichier : {dataframes[4]['name']}"))
```

Analyse du fichier : EdStatsSeries.csv

```
print(f"--- Fichier : {dataframes[4]['name']} ---")
display(dataframes[4]['data'].head())
```

	Series Code	Topic	Indicator Name	Short definition	Long definition	Unit of measure	Periodicity	E Pe
0	BAR.NOED.1519.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 15-19 with...	Percentage of female population age 15-19 with...	NaN	NaN	
1	BAR.NOED.1519.ZS	Attainment	Barro-Lee: Percentage of population age 15-19 ...	Percentage of population age 15-19 with no edu...	Percentage of population age 15-19 with no edu...	NaN	NaN	
2	BAR.NOED.15UP.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 15+ with n...	Percentage of female population age 15+ with n...	NaN	NaN	
3	BAR.NOED.15UP.ZS	Attainment	Barro-Lee: Percentage of population age 15+ wi...	Percentage of population age 15+ with no educa...	Percentage of population age 15+ with no educa...	NaN	NaN	
4	BAR.NOED.2024.FE.ZS	Attainment	Barro-Lee: Percentage of female population age...	Percentage of female population age 20-24 with...	Percentage of female population age 20-24 with...	NaN	NaN	

5 rows × 21 columns

3. Collecte d'informations basiques sur chaque jeu de données

EdStatsCountry.csv

- **Définition** : Une ligne représente un pays unique ou une entité géographique (ex: une région comme l'Amérique Latine).
- **Clé primaire** : `Country Code`
- **Contenu** : Toutes les caractéristiques fixes du pays (monnaie, région, système de recensement, etc...)

EdStatsSeries.csv

- **Définition** : Une ligne représente un indicateur statistique unique (un "Series").

- **Clé primaire** : `Series Code`
- **Contenu** : Les définitions, les sources et les méthodologies pour chaque type de donnée mesurée (ex: taux d'inscription scolaire).
-

EdStatsCountry-Series.csv

- **Définition** : Une ligne représente une relation spécifique entre un pays et un indicateur.
- **Clé composite** : `CountryCode` + `Series Code`
- **Contenu** : Il sert de table de liaison. Il précise souvent la source de données spécifique utilisée pour cet indicateur dans ce pays précis (colonne `DESCRIPTION`).

EdStatsFootNote.csv

- **Définition** : Une ligne représente une note de bas de page liée à une mesure spécifique.
- **Clé composite** : `CountryCode` + `SeriesCode` + `Year`
- **Contenu** : Une explication textuelle (`DESCRIPTION`) pour justifier une anomalie ou une estimation pour une année donnée.

EdStatsData.csv

- **Définition** : Une ligne représente l'évolution historique d'un indicateur pour un pays.
- **Clé composite** : `CountryCode` + `IndicatorCode`
- **Contenu** : Contrairement aux autres, ce fichier est "large" : il contient les valeurs numériques pour chaque année de 1970 à 2100 sur la même ligne.

```
display(Markdown(f"## Analyse du fichier : {dataframes[0]['name']}"))
```

Analyse du fichier : EdStatsCountry.csv

```
rows, columns = dataframes[0]['data'].shape
print(f"Rows: {rows}, Columns: {columns}")
display(Markdown(f" Le fichier : {dataframes[0]['name']} comprend {rows} lignes et {co
```

Le fichier : EdStatsCountry.csv comprend 241 lignes et 32 colonnes

```
duplicate_count = dataframes[0]['data'].duplicated().sum()
display(Markdown(f" Le fichier : {dataframes[0]['name']} posséde {duplicate_count} lig
```

Le fichier : EdStatsCountry.csv posséde 0 lignes dupliqués

```
if duplicate_count > 0:
    dataframes[0]['data'] = dataframes[0]['data'].drop_duplicates()
```

Calcul du pourcentage de valeurs manquantes par colonnes et affichage dans un nouveau dataframe trié par pourcentage décroissant.

```
percent_missing_ed_stats_country = dataframes[0]['data'].isnull().sum() * 100 / len(da
missing_value_df_ed_stats_country = pd.DataFrame({'column_name' : dataframes[0]['data
display(missing_value_df_ed_stats_country)
```

	column_name	percent_missing
Unnamed: 31	Unnamed: 31	100.000000
National accounts reference year	National accounts reference year	86.721992
Alternative conversion factor	Alternative conversion factor	80.497925
Other groups	Other groups	75.933610
Latest industrial data	Latest industrial data	55.601660
Vital registration complete	Vital registration complete	53.941909
External debt Reporting status	External debt Reporting status	48.547718
Latest household survey	Latest household survey	41.493776
Latest agricultural census	Latest agricultural census	41.078838
Lending category	Lending category	40.248963
PPP survey year	PPP survey year	39.834025
Special Notes	Special Notes	39.834025
Source of most recent Income and expenditure data	Source of most recent Income and expenditure data	33.609959
Government Accounting concept	Government Accounting concept	33.195021
Latest water withdrawal data	Latest water withdrawal data	25.726141
IMF data dissemination standard	IMF data dissemination standard	24.896266
Balance of Payments Manual in use	Balance of Payments Manual in use	24.896266
Latest trade data	Latest trade data	23.236515
SNA price valuation	SNA price valuation	18.257261
System of trade	System of trade	17.012448
National accounts base year	National accounts base year	14.937759
Latest population census	Latest population census	11.618257
Region	Region	11.203320
Income Group	Income Group	11.203320
Currency Unit	Currency Unit	10.788382
System of National Accounts	System of National Accounts	10.788382
2-alpha code	2-alpha code	1.244813
WB-2 code	WB-2 code	0.414938
Long Name	Long Name	0.000000
Short Name	Short Name	0.000000
Table Name	Table Name	0.000000

	column_name	percent_missing
Country Code	Country Code	0.000000

Néttoyage des collones atteignant 100% de valeurs manquantes

```
dataframes[0]['data'] = dataframes[0]['data'].dropna(axis=1, how='all')
display(dataframes[0]['data'].head())
```

	Country Code	Short Name	Table Name	Long Name	2-alpha code	Currency Unit	Special Notes	Region	Income Group
0	ABW	Aruba	Aruba	Aruba	AW	Aruban florin	SNA data for 2000-2011 are updated from offici...	Latin America & Caribbean	High income: nonOECD
1	AFG	Afghanistan	Afghanistan	Islamic State of Afghanistan	AF	Afghan afghani	Fiscal year end: March 20; reporting period fo...	South Asia	Low income
2	AGO	Angola	Angola	People's Republic of Angola	AO	Angolan kwanza	April 2013 database update: Based on IMF data,...	Sub-Saharan Africa	Upper middle income
3	ALB	Albania	Albania	Republic of Albania	AL	Albanian lek	NaN	Europe & Central Asia	Upper middle income
4	AND	Andorra	Andorra	Principality of Andorra	AD	Euro	NaN	Europe & Central Asia	High income: nonOECD

5 rows × 31 columns