TP de Python numéro 12 : Bases de données avec Pandas

*Semaine du **12 juin 2025**.*

Le but de ce TP est de se familiariser avec la bibliothèque `pandas` de Python. Celle-ci permet de manipuler des **bases de données**. Les fonctionnalités de pandas que nous rencontrerons sont les suivantes :

- Introduction d'un nouveau type d'objet : le type `pandas.DataFrame` (dit `DataFrame` dans la suite).
- Introduction de nouvelles commandes pour définir des objets de type `Dataframe` et pour les manipuler,
- introduction d'outils permettant une analyse statistique élémentaire sur ces objets.

I. Importation de bases de données en .csv avec pandas

1. Téléchargement de bases de données

Une base de données, comme son nom l'indique, est un ensemble de données organisées (généralement collectées suites à une étude, un recensement, un sondage etc.).

Nous allons ici travailler avec 2 bases de données récoltées sur le site de `data.gouv.fr` (un site gouvernemental mettant à disposition des bases de données qui proviennent de sources publiques). De nombreuses bases de données sont également disponibles sur le site de l'INSEE.

Les données consultables sur ces sites sont très variées : données socio-économiques, environnementales, liées à des secteurs stratégiques (énergie)...

Exercice 1. Nous allons commencer par télécharger les bases de données utilisées en exemple pour ce TP.

1. Se rendre sur la page `data.gouv.fr`, et rendez-vous sur le moteur de recherche de ce site permettant de trier les données téléchargeables : *Découvrez les jeux de données.*
2. Téléchargez la première base de donnée, relative aux indices de position sociale des lycées en France. Les étapes sont en annexe.
3. Téléchargez la seconde base de donnée, relative aux origines du gaz naturel importé en France (étapes en annexe).

Vous avez maintenant téléchargé les bases de données avec lesquelles nous allons travailler. Pour des raisons décrites plus bas, nous allons travailler dans un dossier comme indiqué en annexe.

Ouvrez Pyzo, créez un nouveau fichier Python et enregistrez-le dans ce dossier (nommez-le TP12.py).

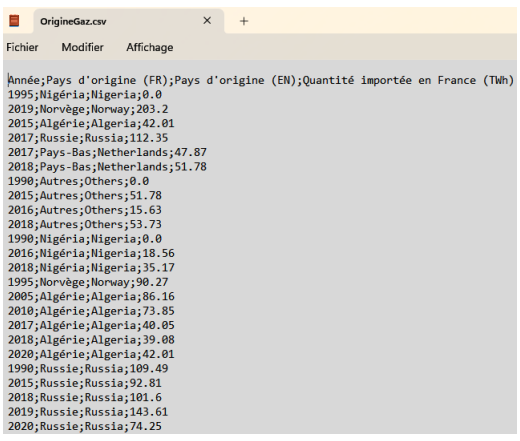
Vous devriez avoir, à ce stade, dans le dossier TP12, 3 fichiers : les deux bases de données (IPSLycee.csv et OrigineGaz.csv) et votre fichier Python de travail vierge (TP12.py).

2. Qu'est-ce qu'une base de données en .csv?

Voyons maintenant ce que nous avons téléchargé.

Exercice 2. Ouvrez le fichier "OrigineGaz.csv" avec un éditeur de texte ***brut*** : ce sont les éditeurs de texte sans mise en forme. Sur windows, le Bloc-Notes est installé par défaut. Sur Mac, c'est TextEdit. Sur Linux, ça dépend de votre distribution (m'appeler, c'est généralement Text Editor). Pour ouvrir le fichier avec un programme choisi, utilisez clique droit > "ouvrir avec" sur Windows.

À droite, ce qui devrait s'afficher.



Format de fichier ".csv"

L'extension ".csv" abrège "comma-separated values". Comme vous pouvez le constater, un fichier .csv est construit de la manière suivante :

1. La première ligne contient une liste de **descripteurs** séparés par un caractère, appelé le **séparateur** du fichier.
   - Les descripteurs décrivent les données récoltées.
   - Le séparateurs est un caractère spécial, qui sert à... séparer les données présentes sur les lignes.
2. Les lignes suivantes contiennent les séries de données collectées (les "entrées"), séparées par un séparateur. Chaque ligne est une série d'entrées différente.
3. Ces entrées peuvent être du texte (La deuxième entrée de chaque ligne est un nom de pays, ou "Autres" de temps en temps) ou des nombres (la première entrée de chaque ligne est une date). Un détail sera important : **la virgule numérique est ici un point**, comme on le voit dans les dernières entrées.

Le séparateur et la virgule numérique sont deux caractères "réservés" dans un fichier .csv : ils ne peuvent être utilisés autrement que pour séparer des données ou marquer une virgule numérique.
Une description correcte des descripteurs est généralement fournie avec les données (voir sur le site de téléchargement).

Exercice 3. Ouvrir les fichiers téléchargés avec un éditeur de texte brut. Répondez aux questions en observant ce fichier.

1. Quelles sont les séparateurs de ces fichiers? Quels sont les caractères utilisés en virgule numérique?
2. Quels descripteurs de ces fichiers sont clairement identifiables? Lesquels ne comprenez-vous pas?
3. Quelle est l'unité de la dernière entrée de chaque ligne du fichier OrigineGaz.csv?
4. Assurez vous d'avoir compris l'intérêt de la première ligne, et sa correspondance avec les autres.

Pour y voir plus clair, ces données sont en fait naturellement organisées sous la forme d'un tableau. Les sauts de lignes séparent les lignes, les séparateurs séparent les colonnes. Vous pouvez ouvrir ces fichiers avec un tableur comme OpenOffice pour avoir une visualisation satisfaisante. Il existe d'autres formats de base de données (.ods, .xls, .xlsx...), le format .csv est très utilisé car assez universel.

3. Se faire une espace de travail

Nous allons utiliser le module `pandas` de python pour visualiser et surtout manipuler ces bases de données. Pour cela, il sera pratique d'avoir le fichier python sur lequel on travaille dans le même dossier que nos données.

Exercice 4. Ouvrez, avec Pyzo, votre fichier de travail. Inscrivez en première ligne :

```
import pandas as pd
```

Dorénavant, on travaillera sur ce fichier.

4. Importation de fichiers .csv avec pandas

Le module `pandas` de `Python` nous permet de manipuler de bases de données avec Python, à des fins d'analyse par exemple.

Pour cela, il faut indiquer à Python l'endroit où se trouve la base de donnée qu'on veut charger (ici, nos fichiers .csv). "L'endroit" où se trouve un fichier sur l'ordinateur s'appelle l'adresse de ce fichier.

a) Pour indiquer une adresse à Python

Nous avons trois possibilités :

- On peut fournir une adresse **absolue**. Il s'agit de l'adresse de votre fichier à partir du nom du disque dur de votre ordinateur.

  Exemple :

  `C:\Users\Vincent\Desktop\TP12\IPSLycee.csv`

Pour donner cette adresse à Python, on remplacera les antislash (caractère réservé en Python) par des slash, et on la donnera sous la forme d'une chaine de caractère.

Exemple :

```
"C:/Users/Vincent/Desktop/TP12/IPSLycee.csv"
```

C'est un peu pénible (ça dépend du système d'exploitation), et nous avons un moyen plus pratique de charger ces fichiers dans Python.

- On peut fournir une adresse **relative** à Python, c'est-à-dire l'adresse prise à partir du dossier dans lequel on travaille. Par exemple, notre fichier python étant dans le dossier "TP12", Python interprétera `"OrigineGaz.csv"` comme une adresse désignant un fichier OrigineGaz.csv du dossier dans lequel se trouve notre fichier de travail (TP12.py).
- Enfin, on peut fournir une adresse web, avec les mêmes règles que pour l'adresse absolue (nous ne le ferons pas ici).

b) Import de fichier avec `pd.read_csv`

La bibliothèque `pandas` est chargée avec le raccourci `pd` grâce à la première ligne de votre fichier python.

pd.read_csv

La commande Python `pd.read_csv()` permet d'importer un fichier .csv sous la forme d'un objet de type `pd.DataFrame`.
Argument obligatoire
Le premier argument pris par `pd.read_csv()` est l'adresse du fichier à importer.

Arguments supplémentaires remarquables
L'argument `sep="separateur"` spécifie le séparateur utilisé par le fichier. Par exemple, `sep=";"` spécifie que les entrées sont séparées par un point-virgule, et `sep=","` spécifie qu'elles sont séparés par une virgule.
L'argument `decimal="virguledecimale"` spécifie la virgule décimale utilisée par le fichier. Par exemple, `decimal="."` spécifie que c'est un point, et `decimal=","` spécifie que c'est une virgule.
L'argument `header=None` indique que le fichier .csv utilisé ne contient pas de lignes de descripteurs (nous ne nous en servirons pas).

Remarque. Si on ne spécifie pas d'argument supplémentaire, le séparateur par défaut est la virgule, et la virgule numérique par défaut est le point.

Dans votre fichier `Python`, ajoutez les lignes :

```
OriGaz=pd.read_csv("OrigineGaz.csv", sep=";")
```

Puis, exécutez ce fichier avec "Executer > Executer le script" et tapez "IPS" dans l'interpréteur. Que remarquez-vous?

Après cette commande effectuée, le contenu de "OrigineGaz.csv" est enregistré dans la variable `OriGaz`, qui est de type `pd.DataFrame`.

Remarque. Pour la lisibilité, Python indique une numérotation des lignes.

Exercice 5. Rajoutez les ligne permettant d'importer la base de données relative aux IPS des lycées, dans une variable Python nommée `IPS`.

Remarque. Avant d'importer un fichier .csv, il est de bon goût de l'ouvrir avec un éditeur de texte brut pour connaitre la nature du séparateur et de la virgule décimale.

II. Manipulation des bases de données

A ce stade, nous avons donc téléchargé des bases de données, que nous avons importées dans Python.

Rajoutez en première ligne l'import `import numpy as np`.

1. Premiers outils

Premiers outils

Soit `L` une base de données de type `np.DataFrame` et `'Desc'` l'un des descripteur de `L`.

- La commande `L.shape` renvoie le couple `(l,c)` formé du nombre `l` de lignes de `L`, et du nombre `c` de ses colonnes.
- La commande `L.head()` renvoie les 5 premières lignes de `L`.
- La commande `L.tail()` renvoie les 5 dernières lignes de `L`.
- La commande `L.columns` renvoie (sous un format particulier) la liste des descripteurs de `L`.
- La commande `L['Desc']` renvoie la base de données obtenue à partir de `L` en ne conservant que la colonne de descripteur `'Desc'`.
- La commande `L['Desc'][i]` renvoie l'entrée de la `i` ième ligne correspondant au descripteur `'Desc'`.
- La commande `L['Desc'].values` renvoie la colonne de descripteur `'Desc'` sous forme de tableau numérique `np.ndarray`.
- La commande `L.values` renvoie les données de `L` sous forme d'une liste `np.ndarray` des listes donnant les entrées de chaque lignes.
- La commande `L.index` renvoie la liste (sous un format particulier) des indices des lignes de `L`.

Exemple 6. Lancer `OriGaz.head()` dans l'interpréteur et constater.

Remarque. La commande `.head()` est très utile pour vérifier que l'importation s'est bien passée, sans problème de séparateur ou de virgule décimale.

Exercice 7. Tester ces commandes sur les bases de données introduites auparavant. Remarquer que les descripteurs sont systématiquement de type `string`.

2. Extraction de lignes et de colonnes

Souvent, les bases de données trouvées contiennent des informations dont nous ne nous soucions pas pour notre étude. On veut alors, ne serait-ce que pour des questions de lisibilité, ne garder que les informations qui nos intéressent.

Extraction de colonnes

Soit `L` une base de donnée, et `D` une liste de descripteurs de `L`.
Alors, `L[D]` renvoie la base de données obtenue à partir de `L` en ne conservant que les colonnes dont le descripteur est dans `D`.

Exemple 8.

La base de données OriGaz contient deux colonnes donnant les noms des pays (en français et en anglais), c'est redondant. Simplifions cela. Tapez, à la suite, dans votre code Python :

```
GazSimp=OriGaz[["Année","Pays d'origine (FR)","Quantité importée en France (TWh)"]]
```

puis (après exécution) demander à python d'afficher la variable `GazSimp`. Comparer avec `OriGaz`.

Exercice 9. Créer une variable `IPSlisible` contenant la base de données obtenue à partir de `IPS` en ne conservant que les colonnes donnant : L'année, le code du département, le nom de l'établissement, le nom de la commune, l'IPS pour la voie générale et technologique, et l'IPS pour la voie professionnelle.

Extraction de lignes

Soit `L` une base de donnée, et `n,p` des entiers de type `int`.
Alors,

- `L[n:p]` renvoie la base de données obtenue à partir de `L` en ne gardant que les lignes de la `n` ième à la `p-1` ième.
- `L.head(n)` renvoie les `n` premières lignes de `L`.

Extraction de colonnes et de lignes

Soit `L` une base de donnée, `D` une liste de descripteurs de `L` et `n,p` des entiers.
Alors, `L[D][n:p]` renvoie la base de données obtenue à partir de `L` en ne conservant que les colonnes dont le descripteur est dans `D`, et les lignes de la `n` ième à la `p-1` ième.

3. Utilisation directe de matplotlib

Au début de votre fichier, ajoutez l'importation : `import matplotlib.pyplot as plt`.

Utilisation de la méthode `.plot()`

Soit `L` une base de données. Alors, la commande `L.plot()` (suivie de `plt.show()`) du module `matplotlib.pyplot` affiche le tracé des colonnes numériques de `L`, avec le numéro de ligne en abscisse.

Exemple 10. Nous allons, plus tard, utiliser cette commande pour afficher l'évolution temporelle des volumes de gaz importés depuis chaque pays (la Russie par exemple). Il nous faudra, pour cela, créer une nouvelle base de donnée à partir de GazSimp en ne gardant que les lignes qui recensent des importations de gaz russe, puis trier ces lignes par années.

4. Filtrage des lignes

On peut appliquer des filtres pour ne garder que certaines lignes.

Filtre simple

Soit `L` une base de donnée, et `'Desc'` un descripteur de `L`.
Alors,

- La commande `L[L['Desc']==n]` renvoie la base de données obtenue à partir de `L` en ne gardant que les lignes dont l'entrée de la colonne `'Desc'` vaut `n`.
- La commande `L[L['Desc']>n]` renvoie la base de données obtenue à partir de `L` en ne gardant que les lignes dont l'entrée de la colonne `'Desc'` est strictement supérieure à `n` (données numériques uniquement).
- Le même principe marche avec les autres opérateurs de comparaison `<, <=, >=, !=`.

Exemple 11. Exécuter la commande `GazSimp[GazSimp["Pays d'origine (FR)"]=="Russie"]` dans l'interpréteur et constater le résultat.

Enregistrez cette base de donnée dans une variable convenablement nommée (GazRusse dans la suite du TP).

Exercice 12. A partir de la base de donnée IPS :

1. Afficher la liste des lycées avec un IPS en voie GT supérieur à 110 .
2. Afficher la liste des imports annuels de gaz supérieurs à 130 TWh

Remarque. Quand on construit une nouvelle base de donnée en filtrant les lignes, celles-ci ne sont pas renumérotées. Les indices des lignes (qui permettent de désigner ces lignes) deviennent alors lacunaires, d'où la commande `L.index` pour retrouver ces indices.

Filtre multiple

Soit `L` une base de donnée. Soient `filtre1,filtre2` deux filtres construits selon l'encadré ci-dessus. Par exemple, on pourrait avoir `L['Desc1']>n` pour `filtre1` et `L['Desc2']==n` pour `filtre2`.
Alors, `L[np.logical_and(filtre1,filtre2)]` renvoie la base de données obtenue en ne gardant que les lignes satisfaisant les deux filtres.
De même, `L[np.logical_or(filtre1,filtre2)]` renvoie la base de données obtenue en ne gardant que les lignes satisfaisant l'un des deux filtres.

Exercice 13.
1. Faire afficher la liste des imports de gaz supérieurs à 100 TWh et antérieurs à l'année 2015.
2. Afficher la liste des lycées des Yvelines (78) dont l'IPS en voie GT est inférieur à 120.

5. Méthode de tri

Le module pandas inclue une fonctionnalité de tri selon les valeurs d'une colonne numérique.

Tri d'une base de donnée

Soit `L` une base de données et `'Desc'` un descripteur de `L` dont la colonne correspondante est à entrées numériques.
Alors, `L.sort_values('Desc')` renvoie la base de données obtenue en permutant les lignes de `L` par valeurs croissantes de la colonne `'Desc'`.

Exemple 14. Dans l'interpréteur, exécuter `GazCroissant=OriGaz.sort_values("Quantité importée en France (TWh)")`.

Afficher `GazCroissant`, et à l'aide de `GazCroissant.tail()` et `GazCroissant.head()`, donner les records d'imports (année et pays).

Exercice 15.
1. Donner les 5 lycées ayant l'IPS en voie GT le plus élevé, puis le plus faible.
2. Donner les 5 lycées d'Ile-et-Vilaine (35) ayant l'IPS en voie GT le plus élevé, puis le plus faible.

6. Indicateurs statistiques

Pandas fournit des indicateurs statistiques.

Commandes à connaitre

Soit `L` une base de donnée.
- La commande `L.describe()` renvoie une analyse statistique des colonnes numériques de `L`, contenant, dans l'ordre, l'effectif total, la moyenne, l'écart-type, le minimum, les 3 quartiles, et le maximum.
- `L.mean()` renvoie les moyennes des colonnes numériques de `L`. `L.mean(numeric_only=True)` évite le message d'erreur quand `L` contient des données non numériques.
- `L.std()` renvoie les écarts-types des colonnes numériques de `L`.
- `L.median()` renvoie la médiane des colonnes numériques de `L`. `L.median(numeric_only=True)` évite un message d'erreur en cas de colonnes non numériques.
- `L.count` renvoie l'effectif de chaque colonne de `L`.

Pour chaque commande ci-dessus, l'argument optionnel `axis=1` effectue le calcul statistique sur les lignes de `L` à la place. Par exemple, `L.std(axis=1)` affiche l'écart type des entrées des lignes numériques de `L` (ce qui peut facilement n'avoir aucun sens).

Exemple 16. Afficher, à l'aide de `describe`, les indicateurs statistiques des IPS des lycées d'abord sur toute la France, puis exclusivement sur l'Ile-de-France (75-77-78-91-92-93-94-95).

III. Exercices

On considère les variables globales `GazSimp` et `IPS` affectées selon le déroulé du TP. Cet encadré peut être utile.

Les entrées sont "itérables"

Soit `L` une base de données et `'Desc'` un descripteur de `L`. Alors, `for val in L['Desc']:` initie une boucle dans laquelle la variable (locale) `val` parcourt les entrées de la colonne `'Desc'` de `L`.
De plus, `for i in L.index :` initie une boucle dans laquelle la variable (locale) `i` parcourt les indices des lignes de `L`.

Exercice 17. Travail sur la base de données relative aux importations de gaz. On supposera que ces données d'importation sont exhaustives. Les questions sont posées assez directement et nécessitent généralement plusieurs étapes.

1. Afficher l'analyse statistique élémentaire de la base de donnée relative aux importations de gaz (effectif total, moyenne, médiane, etc.).
2. Déterminer la liste des années concernées par cette base de données. Construire, en Python, la liste de ces années (on pourra penser à la commande `np.unique`).
3. En une ligne de code, calculer la quantité totale de gaz importée en 2015. On pourra utiliser la commande `np.sum`.
4. Tracer l'évolution, année par année, de la quantité annuelle de gaz importée en France.
5. Choisissez deux pays parmi les pays exportant du gaz en France. Pour chacun de ces pays :
   (a) Définir (et stocker dans une variable Python) la base de donnée obtenu à partir de GazSimp en ne gardant que les importations de ce pays.
   (b) Tracer l'évolution temporelle de ces importations. Pour cela, triez la base de donnée précédente selon l'année.
6. Déterminez les plus gros importateurs sur la période de 2016 à 2019.
7. Proposer des questions digne d'intérêt à votre enseignant, et y répondre.

Exercice 18. Travail sur la base de données relative aux IPS des lycées. On ne gardera pas les colonnes (inutiles ici, Python sait le faire) donnant les écarts-types des IPS.

1. Consulter les données présentes pour votre lycée d'origine.
2. Déterminer la liste des années scolaires pour lesquelles des données sont présentes.
3. Définir une base de donnée (stockée dans une variable Python) ne gardant que les données d'une même année (de votre choix). Dans la suite de cet exercice, on travaillera sur cette base de donnée (on restreint notre étude à l'année scolaire choisie).
4. Déterminer l'IPS moyen des lycées français en voie générale et technologique, puis en voie professionnelle.
5. Écrire une fonction prenant en entrée le numéro d'un département, et renvoyant en sortie la moyenne et l'écart-type des IPS en voie GT de ce département (sous la forme d'un couple).
6. Déterminer les lycées ayant les cinq plus grands IPS en voie GT.
7. Déterminer le département ayant l'IPS en voie générale et technologique moyen le plus élevé, puis le plus faible.
8. Déterminer le département dont l'écart type des IPS en voie GT est maximale, puis minimale.
9. Proposer des questions digne d'intérêt à votre enseignant, et y répondre.

IV. Annexe

Téléchargement des bases de données

Avant tout, créez un nouveau dossier sur le bureau de votre ordinateur, nommez-le TP12 par exemple (ne le nommez pas "pandas"). Des explications seront données plus tard.

Téléchargez la première base de données, relative aux indices de position sociale des lycées, en suivant les instructions ci-dessous.

1. Cherchez «*Indice de position sociale des lycées*» et cliquez sur le premier lien.

2. Dans l'onglet "Fichiers" (bas de page), cliquez sur le premier lien permettant de télécharger la base de données au format `csv`.

3. Déplacez le fichier téléchargé dans le dossier TP12. Renommez ce fichier avec un nom simple : `IPSLycee.csv`.

Téléchargez la seconde base de données, relative aux origines du gaz naturel importé en France.

1. Retournez sur le moteur de recherche du site *data.gouv.fr*. Cherchez « *Origine du gaz naturel*» et cliquez sur le premier lien.

2. Dans l'onglet "Fichiers", cliquez sur le premier lien permettant de télécharger la base de données au format `csv`.

3. Déplacez le fichier téléchargé dans le dossier TP12. Renommez ce fichier avec un nom simple : `OrigineGaz.csv`.