📊 Mini Projet : Pandas

Dans le cadre d'un projet de recherche, notre université souhaite exploiter des données stockées dans un fichier Excel (voir le fichier mini_projet_1). Ce fichier présente une petite partie des données.

I. Importation des librairies

In [1]:
```
import pandas as pd
```

II. Importation du jeu de données

In [2]:
```
df = pd.read_excel("mini_projet_1.xlsx")

df.head()
```

Out[2]:

	Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse	Membre des structures de recherche partenaires	Unnamed: 8	Unnamed: 9	Unnan
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nom et Prénom	Grade	Spécialité	Intitul struc
1	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 1	PA	Informatique	
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	chercheur 2	PES	Informatique	S
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	chercheur 3	PA	Informatique	N
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	chercheur 4	PA	Santé	

III. Data Exploration & Cleaning

On remplace les valeurs NaN en dessous de chaque personne par la valeur qui la précéde grâce à la fonction `fillna()` et la mathod `ffill` ou `forward fill`.

In [3]:
```
df[['Id','Nom', 'Prenom', 'Université', 'Grade', 'Spécialité','Structure de recherche Po
```

On renomme des colonnes "Unnamed" avec leurs noms qui sont décalés d'une ligne aprés importation et on supprime la première ligne.

In [4]:
```
df = df.iloc[1:, :]
df.head()
```

Out[4]:

	Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse	Membre des structures de recherche partenaires	Unnamed: 8	Unnamed: 9	Unnam
1	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 1	PA	Informatique	I
2	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 2	PES	Informatique	S
3	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 3	PA	Informatique	M
4	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 4	PA	Santé	I
5	2.0	Chakour	Hatim	UIT	PES	Informatique	MISC	chercheur A	PES	Physique	K

In [5]:
```
df = df.rename({'Unnamed: 8': 'Grade_partenaire', 'Unnamed: 9': 'Spécialité_partenaire',
df.head()
```

Out[5]:

	Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse	Membre des structures de recherche partenaires	Grade_partenaire	Spécialité_par
1	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 1	PA	Inforr
2	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 2	PES	Inforr
3	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 3	PA	Inforr
4	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 4	PA	
5	2.0	Chakour	Hatim	UIT	PES	Informatique	MISC	chercheur A	PES	P

On remplace les valeurs manquantes par des 0 et on modifie le dtype de Id et Nombre à un type entier.

In [6]:
```
df = df.fillna(0)
df[['Id', 'Nombre']] = df[['Id', 'Nombre']].astype('int')
```

On remplace l'index par la colonne Id

In [7]:
```
df.set_index('Id', inplace=True, drop=True)
df.head()
```

Out[7]:

Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse	Membre des structures de recherche partenaires	Grade_partenaire	Spécialité_partena
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 1	PA	Informatic
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 2	PES	Informatic
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 3	PA	Informatic
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 4	PA	Sa
2	Chakour	Hatim	UIT	PES	Informatique	MISC	chercheur A	PES	Physic

IV. Réponses aux quetions

- (1, 2, 3) Donner le nombre d'articles, communications et de thèses encadrées pour chaque chercheur.

In [8]:
```
articles = df.groupby(['Nom', 'Prenom','Publications Scientifiques'])['Nombre'].sum()
articles
```

Out[8]:
```
Nom      Prenom  Publications Scientifiques
Amouri   Aya     0                            0
                 Nombre de thèses encadrées   7
                 Nombre des articles         15
                 Nombre des communications   15
Chakour  Hatim   0                            0
                 Nombre de thèses encadrées   3
                 Nombre des articles         10
                 Nombre des communications    5
Hakam    Amine   0                            0
                 Nombre de thèses encadrées   4
                 Nombre des articles         20
                 Nombre des communications   10
Hayoun   Adam    Nombre de thèses encadrées   7
                 Nombre des articles         15
                 Nombre des communications   15
Name: Nombre, dtype: int32
```

- (4) Donner le total des publications scientifiques.

In [9]:
```
pub_sci = df.groupby(['Nom', 'Prenom'])['Nombre'].sum()
pub_sci
```

Out[9]:
```
Nom      Prenom
Amouri   Aya       37
Chakour  Hatim     18
Hakam    Amine     34
Hayoun   Adam      37
Name: Nombre, dtype: int32
```

- (5) Donner le nombre des membres des structures de recherche partenaires pour chaque chercheur

In [11]:
```
num_chercheurs = df.groupby(['Nom', 'Prenom'])['Membre des structures de recherche parte
num_chercheurs
```

Out[11]:
```
Nom       Prenom
Amouri    Aya       5
Chakour   Hatim     5
Hakam     Amine     4
Hayoun    Adam      3
Name: Membre des structures de recherche partenaires, dtype: int64
```

- (6) Lister les intitulés de la structure pour chaque chercheur (sans les dupliqués)

In [12]:
```
struct_chercheurs = df.groupby(['Nom', 'Prenom'])['Intitulé de la structure'].unique()
struct_chercheurs
```

Out[12]:
```
Nom       Prenom
Amouri    Aya              [FFSD, ZQW, MARO]
Chakour   Hatim           [KZEE, S2SD, MISC]
Hakam     Amine      [IPSS, SSBK, MISC, H2C]
Hayoun    Adam                    [CASIF, 0]
Name: Intitulé de la structure, dtype: object
```

- (7) Lister les Spécialités de la structure pour chaque chercheur (sans les dupliqués)

In [13]:
```
specialite_chercheurs = df.groupby(['Nom', 'Prenom'])['Spécialité_partenaire'].unique()
specialite_chercheurs
```

Out[13]:
```
Nom       Prenom
Amouri    Aya        [Informatique, Electrique, Physique]
Chakour   Hatim           [Physique, Informatique, Santé]
Hakam     Amine                     [Informatique, Santé]
Hayoun    Adam                              [Physique, 0]
Name: Spécialité_partenaire, dtype: object
```

- (8) Effectuer une recherche par le champ Id et afficher le nom et le prénom du chercheur

In [18]:
```
df.loc[df.index == 1, ['Nom', 'Prenom']].head(1)
```

Out[18]:

Id	Nom	Prenom
1	Hakam	Amine

- (9)

In [19]:
```
df.loc[df.index == 1, ['Nom', 'Prenom','Université','Grade','Spécialité','Structure de r
```

Out[19]:

Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse
1	Hakam	Amine	UM5R	PES	Informatique	IPSS

In [20]:
```
df.loc[df.index == 1, ['Membre des structures de recherche partenaires','Grade','Spécial
```

Out[20]:

Id	Membre des structures de recherche partenaires	Grade	Spécialité_partenaire	Intitulé de la structure
1	chercheur 1	PES	Informatique	IPSS
1	chercheur 2	PES	Informatique	SSBK
1	chercheur 3	PES	Informatique	MISC
1	chercheur 4	PES	Santé	H2C

In [21]:
```
df.loc[df.index == 1, ['Publications Scientifiques','Nombre']]
```

Out[21]:

Id	Publications Scientifiques	Nombre
1	Nombre des articles	20
1	Nombre des communications	10
1	Nombre de thèses encadrées	4
1	0	0