📊 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 [121…
```
import pandas as pd
```

II. Importation du jeu de données

In [128…
```
df = pd.read_excel("mini_projet_1.xlsx", engine = 'openpyxl')

df.head()
```

Out[128]:

	Id	Nom	Prenom	Université	Grade	Spécialité	Structure de recherche Porteuse	Membre des structures de recherche partenaires	Unnamed: 8	Unnamed: 9	Unnar
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	
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	chercheur 3	PA	Informatique	
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 [129…
```
df[['Id','Nom', 'Prenom', 'Université', 'Grade', 'Spécialité','Structure de recherche Porteuse']]
```

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

In [130…
```
df = df.iloc[1:, :]
df.head()
```

Out[130]:

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

In [131…
```
df = df.rename({'Unnamed: 8': 'Grade_partenaire', 'Unnamed: 9': 'Spécialité_partenaire', 'Unname
df.head()
```

Out[131]:

	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	Inform
2	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 2	PES	Inform
3	1.0	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 3	PA	Inform
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

In [132…
```
df = df.fillna(0)
```

In [133…
```
df[['Id', 'V_Publications Scientifiques']] = df[['Id', 'Nombre']].astype('int')
```

In [134…
```
df.set_index('Id', inplace=True, drop=True)
df.head()
```

Out[134]:

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	Informati
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 2	PES	Informati
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 3	PA	Informati
1	Hakam	Amine	UM5R	PES	Informatique	IPSS	chercheur 4	PA	Sa
2	Chakour	Hatim	UIT	PES	Informatique	MISC	chercheur A	PES	Physi

IV. Réponses aux quetions

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

In [135…
```
articles = df.groupby(['Nom', 'Prenom','Publications Scientifiques'])['Nombre'].sum()
articles
```

Out[135]:
```
Nom      Prenom  Publications Scientifiques
Amouri   Aya     0                            0.0
                 Nombre de thèses encadrées   7.0
                 Nombre des articles         15.0
                 Nombre des communications   15.0
Chakour  Hatim   0                            0.0
                 Nombre de thèses encadrées   3.0
                 Nombre des articles         10.0
                 Nombre des communications    5.0
Hakam    Amine   0                            0.0
                 Nombre de thèses encadrées   4.0
                 Nombre des articles         20.0
                 Nombre des communications   10.0
Hayoun   Adam    Nombre de thèses encadrées   7.0
                 Nombre des articles         15.0
                 Nombre des communications   15.0
Name: Nombre, dtype: float64
```

- (4) Donner le total des publications scientifiques.

In [136…
```
pub_sci = df.groupby(['Nom', 'Prenom'])['Nombre'].sum()
pub_sci
```

Out[136]:
```
Nom      Prenom
Amouri   Aya       37.0
Chakour  Hatim     18.0
Hakam    Amine     34.0
Hayoun   Adam      37.0
Name: Nombre, dtype: float64
```

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

chercheur

In [137…
```
num_chercheurs = df.groupby(['Nom', 'Prenom'])['Membre des structures de recherche partenaires']
num_chercheurs
```

Out[137]:
```
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 [138…
```
struct_chercheurs = df.groupby(['Nom', 'Prenom'])['Intitulé de la structure'].unique()
struct_chercheurs
```

Out[138]:
```
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 [139…
```
specialite_chercheurs = df.groupby(['Nom', 'Prenom'])['Spécialité_partenaire'].unique()
specialite_chercheurs
```

Out[139]:
```
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 [141…
```
df.loc[df.index == 1, ['Nom', 'Prenom']].head(1)
```

Out[141]:

Id	Nom	Prenom
1	Hakam	Amine

- (9)

In [142…
```
df.loc[df.index == 1, ['Nom', 'Prenom','Université','Grade','Spécialité','Structure de recherche
```

Out[142]:

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

In [143…
```
df.loc[df.index == 1, ['Membre des structures de recherche partenaires','Grade','Spécialité_parte
```

Out[143]:

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 [144…
```
df.loc[df.index == 1, ['Publications Scientifiques','Nombre']]
```

Out[144]:

Id	Publications Scientifiques	Nombre
1	Nombre des articles	20.0
1	Nombre des communications	10.0
1	Nombre de thèses encadrées	4.0
1	0	0.0