

TP Python : Agrégation avancée avec Pandas

Exercice guidé

Objectif

Apprendre à utiliser la méthode `groupby` avec des fonctions d'agrégation multiples et personnalisées pour résumer un dataset de joueurs de basketball.

Dataset fourni

Le dataset contient les colonnes suivantes :

- `player` : nom du joueur
- `team` : équipe
- `POS` : position
- `GP` : matchs joués
- `MIN` : minutes jouées
- `PTS` : points
- `FGM`, `FGA`, `FG%` : tirs réussis, tentés, pourcentage
- `3PM`, `3PA`, `3P%` : tirs à 3 points réussis, tentés, pourcentage
- `FTM`, `FTA`, `FT%` : lancers francs réussis, tentés, pourcentage
- `REB`, `AST`, `STL`, `BLK`, `TO` : statistiques diverses
- `DD2`, `TD3` : double double, triple double
- `year` : année
- `salary` : salaire

Exercice guidé

Étape 1 : Charger le dataset

```
import pandas as pd

# Charger le dataset
df = pd.read_csv("basketball_stats.csv")
print(df.head())
```

Étape 2 : Agrégation simple par année

```
# Calculer la moyenne des points et des rebonds par année
result_simple = df.groupby('year')[['PTS', 'REB']].mean()
print(result_simple)
```

Question 1 : Que représente chaque ligne de `result_simple` ?

Étape 3 : Agrégation avancée avec plusieurs fonctions

```
# Définir les fonctions d'agrégation
aggregations = {
    'PTS': ['sum', 'mean', 'max'], # total, moyenne, maximum
    'REB': ['sum', 'mean', 'std'], # total, moyenne, cart-type
    'salary': ['sum', lambda x: x.quantile(0.9)] # somme et 90me percentile
}

# Appliquer groupby et agg
result_adv = df.groupby('year').agg(aggregations)
print(result_adv)
```

Question 2 : Expliquer ce que calcule chaque colonne de `result_adv`.

Question 3 : Pourquoi utiliser une fonction `lambda` pour le percentile au lieu d'une fonction prédéfinie ?

Étape 4 : Agrégation par équipe et par position

```
# Group by team et position
team_pos_agg = df.groupby(['team', 'POS']).agg({
    'PTS': ['mean', 'max'],
    'AST': 'mean',
    'REB': 'mean'
})
print(team_pos_agg)
```

Question 4 : Quelle est la différence entre cette agrégation et celle faite uniquement par année ?

Étape 5 : Exploration des résultats

- Identifier l'année où les joueurs ont marqué le plus de points.
- Identifier l'équipe et la position ayant la meilleure moyenne de passes (AST).

Résumé

Avec `groupby` et `agg`, on peut :

- Appliquer plusieurs fonctions sur une ou plusieurs colonnes
- Combiner des fonctions standard et personnalisées (`lambda`)
- Créer des statistiques résumées très complètes sur des datasets volumineux