Rapport de qualité des données — SoccerStats (top5-players.csv)

Synthèse Rapide

Le fichier `top5-players.csv` contient des données solides (2 852 joueurs, 37 colonnes, très peu de valeurs manquantes).

Toutefois, certaines limites importantes existent : les joueurs aux rôles très différents (gardien, défenseur, attaquant...) sont analysés ensemble, les statistiques "par 90 minutes" sont instables pour les joueurs avec peu de temps de jeu, et quelques incohérences mineures apparaissent à cause des arrondis.

---

1. Vue d'ensemble

- **Joueurs**: 2 852
- **Variables**: 37
- **Types**: mélange de chiffres (minutes, buts, passes, xG...) et de texte (joueur, poste, club, ligue)
- **Période de naissance des joueurs**: 1982–2008

---

2. Principaux problèmes de qualité

1. **Valeurs manquantes très rares (<0,2%)**
   - Nation, âge, année de naissance et certaines stats avancées ont quelques trous.

     → Impact faible.
2. **Incohérences dues aux arrondis**
   - Exemple : `npxG + xAG ≠ npxG+xAG` sur 846 lignes.
   - Exemple : `G+A_90 ≠ (G+A)/90s` sur 1 694 lignes.
   - Ces différences sont faibles et liées aux décimales.

3. **Valeurs extrêmes (outliers** *(valeurs aberrantes)***)**
    - Beaucoup de zéros (penaltys, cartons rouges). → Ici pas un problème.
    - Valeurs très hautes pour certains joueurs (jusqu'à 44 actions décisives, 392 passes progressives, etc.). → Peuvent sembler incohérentes pour certains joueurs.
    - Statistiques "par 90 minutes" gonflées pour les joueurs avec peu de minutes jouées.
4. **Mélange des postes (problème majeur)**
    - Les gardiens, défenseurs, milieux et attaquants sont mélangés.
    - Exemple : comparer un gardien (réflexes, arrêts, dégagements) avec un attaquant (xG, buts, passes) n'a aucun sens. → Cela entraine aussi beaucoup de valeurs manquantes pour les gardiens.
    - Impact : les analyses globales (moyennes, distributions) sont biaisées car elles n'ont pas de réalité sportive.

---

3. Conclusion

Les données présentes sont généralement de bonne qualité avec peu de valeurs manquantes, mais nécessitent des précautions importantes lors de l'analyse. La séparation par poste est impérative avant toute analyse statistique, et les mesures "par 90 minutes" doivent être interprétées avec prudence pour les joueurs ayant peu de temps de jeu. Malgré ces limitations, le dataset offre une base solide pour l'analyse des performances des joueurs, à condition d'appliquer les filtres appropriés.