Rapport de justification — Choix des données et structuration

1. Pourquoi ne pas utiliser le CSV de base ?

Le fichier CSV initial (`top5-players.csv`) fourni dans le cadre du projet contient bien les statistiques essentielles sur les joueurs des 5 grands championnats européens, mais il présente plusieurs limites :

- **Mélange des postes** : les données incluent simultanément les gardiens, les défenseurs, les milieux et les attaquants. Or leurs rôles et métriques de performance sont trop différents pour être comparés directement.
- Manque de données : certaines données sont manquantes (ex : aucune données sur les gardiens, manque de données sur la défense).
- **Problèmes de granularité** : certaines statistiques avancées (par exemple xG + xAG) présentent des incohérences mineures liées à l'arrondi, et la documentation de calcul n'est pas toujours accessible.

---

2. Pourquoi choisir fbref.com comme source ?

Nous avons choisi de **scraper les données directement depuis fbref.com** pour des raisons de fiabilité et de pertinence :

- **Référence mondiale** : fbref.com est l'une des principales bases de données statistiques en football, développée par **Sports Reference LLC**, déjà connue pour ses sites spécialisés (Basketball Reference, Baseball Reference, etc.). La société est reconnue pour la qualité de son travail statistique.
- **Source officielle de données avancées** : les modèles de statistiques avancées (xG, xAG, pressing, passes progressives, etc.) sont fournis par **StatsBomb**, un des leaders mondiaux de l'analyse de performance footballistique. Cela conforte la crédibilité et la précision des données.
- **Transparence méthodologique** : fbref documente ses calculs (expected goals, métriques avancées), ce qui offre une bonne traçabilité statistique.

- **Large couverture** : toutes les grandes compétitions, joueurs et équipes sont couverts, ce qui permet une extension facile du projet.

→ **Fiabilité garantie** par la combinaison de l'expertise Sports Reference et StatsBomb.

---

3. Pourquoi séparer les données en plusieurs CSV ?

Dans une approche analytique, il est indispensable de distinguer les rôles et types de données :

- **Gardiens vs joueurs de champ** :
  - Les gardiens (GK) ont des statistiques spécifiques : arrêts, expected goals on target (xGOT), pourcentage d'arrêts, sorties aériennes, relances longues…
  - Les joueurs de champ (DF, MF, FW) ont des métriques totalement différentes : xG, xAG, passes progressives, dribbles, actions défensives…
  - Les garder dans le même fichier provoquerait des analyses biaisées (par ex. une moyenne globale de passes progressives n'aurait aucun sens avec les gardiens inclus).
- **Lisibilité et exploitation** :
  - Séparer les fichiers permet de simplifier les traitements et les visualisations.
  - On peut ainsi créer des indicateurs adaptés à chaque rôle.

---

4. Conclusion

- Le CSV fourni de base est exploitable mais limité : données mélangées, peu de flexibilité et risques de biais.
- fbref.com, via StatsBomb, garantit des données fiables et reconnues.
- La séparation en plusieurs fichiers (gardiens / joueurs de champ) est **nécessaire** pour préserver la **comparabilité** et la **validité des analyses**.