Rapport critique

Partie 1 : Vérification et Préparation des Données

Dès l'inspection initiale du jeu de données fourni, notre équipe a constaté des lacunes importantes qui compromettaient la faisabilité du projet. De nombreuses statistiques étaient manquantes pour plusieurs joueurs, avec notamment l'absence quasi-totale de données pour le poste de gardien de but. Face à ce constat, la décision a été prise de ne pas utiliser le dataset fourni et de construire notre propre source de données, plus fiable et complète.

Pour ce faire, nous avons mené une opération de **web scraping** sur le site de référence `fbref.com`, reconnu pour la qualité et la profondeur de ses statistiques sur le football. Une fois les données brutes collectées, la bibliothèque **Pandas** de Python a été utilisée pour les traiter. Le principal défi lors de cette phase a été la gestion des **doublons** : de nombreux joueurs apparaissaient plusieurs fois, conséquence d'un transfert entre deux clubs au cours de la saison 2023-2024. Pour résoudre ce problème, un script a été développé afin de consolider les statistiques d'un même joueur en une seule entrée, tout en conservant son club le plus récent comme référence.

Le processus a abouti à la création de quatre fichiers CSV distincts et propres, segmentant les joueurs par grands postes (attaquants, milieux, défenseurs et gardiens). Enfin, pour permettre une analyse objective et pertinente, chaque jeu de données a été enrichi avec des statistiques **normalisées par 90 minutes de jeu**, en plus des données brutes. La fiabilité de la source `fbref.com` a permis d'obtenir un dataset final de haute qualité, sans valeurs aberrantes notables, constituant une base solide pour l'analyse.

Partie 2 : Analyse Exploratoire et Visualisations

Après la phase de nettoyage, une analyse exploratoire a été menée pour dégager les grandes tendances et la structure de notre jeu de données. Cette étape a été divisée en deux temps : une série de visualisations descriptives pour comprendre la composition de l'effectif, suivie d'une étude des corrélations pour identifier les liens entre différents indicateurs de performance.

2.1. Visualisations Descriptives : Portrait-robot des joueurs

Trois visualisations principales ont permis de dresser un premier portrait de la population étudiée :

- **Répartition par poste :** L'analyse de la répartition des postes montre une distribution équilibrée et logique pour des effectifs professionnels. Les défenseurs (DF) constituent le groupe le plus important avec 35% de l'effectif, suivis des attaquants (FW) à 31%, des milieux (MF) à 26% et enfin des gardiens (GK) qui représentent logiquement la plus petite part avec 8%.
- **Répartition par nationalité :** Le graphique du top 20 des nationalités révèle que les joueurs issus des cinq pays hôtes des championnats sont les plus nombreux, avec en tête l'Espagne (388 joueurs), la France (299) et l'Allemagne (245). Cette forte présence locale est complétée par les principaux pays exportateurs de talents comme le Brésil et l'Argentine, ainsi qu'un contingent important de nations africaines.
- **Répartition par ligue :** L'échantillon de joueurs se révèle bien équilibré entre les cinq championnats. La Liga espagnole et la Serie A italienne comptent le plus de joueurs (environ 590 chacun), tandis que la Bundesliga allemande en compte le moins (481). Cette distribution garantit une bonne représentativité de chaque ligue pour les analyses comparatives à venir.

2.2. Études de Corrélations : Premiers liens entre les performances

Pour aller plus loin, plusieurs matrices de corrélation ont été générées afin de comprendre les relations entre différents indicateurs de performance.

- **Analyse des KPI défensifs :** L'étude des indicateurs défensifs a montré que les corrélations entre eux sont globalement faibles. Par exemple, la relation entre l'indice de performance de bloc (BPI) et l'indice de récupération défensive (DRI) est quasi nulle (r = -0.009). Cela confirme que ces KPI mesurent des aspects très différents et complémentaires de la performance défensive. La corrélation la plus notable (r = 0.418) entre l'indice d'interception (DII) et la fiabilité défensive (DefRel) suggère toutefois qu'un joueur efficace en interception tend à être globalement plus fiable.
- **Analyse de la performance offensive :** La matrice sur l'efficacité des tirs a mis en lumière une distinction claire entre le "volume" et "l'efficacité". Alors que le volume de tirs est fortement corrélé au nombre de buts attendus (xG) et marqués (Gls), il n'est quasiment pas corrélé aux ratios d'efficacité

comme le `Goals_per_Shot`. Autrement dit, tirer beaucoup augmente les chances de marquer, mais ne garantit pas d'être un finisseur plus précis.

- **Analyse du lien entre pressing et attaque :** Enfin, une corrélation modérée a été observée entre le nombre de tacles réussis (Tkld) et la production offensive (r = 0.53 avec les passes décisives). Cela indique qu'une forte activité défensive au pressing peut contribuer positivement à la création d'occasions, bien que ce lien ne soit pas systématique.

Partie 3 : Création d'Indicateurs de Performance Avancés

Pour dépasser les limites des statistiques traditionnelles (buts, passes décisives), qui ne racontent qu'une partie de l'histoire, une phase de création d'indicateurs de performance avancés a été nécessaire. L'objectif était de développer des métriques capables de mesurer plus finement la qualité et l'impact réel d'un joueur sur le jeu. Cette démarche s'est articulée en deux temps : la sélection d'indicateurs avancés reconnus et la construction de nos propres KPI (Key Performance Indicators) composites.

3.1. Sélection d'Indicateurs Avancés Pertinents

Plusieurs indicateurs ont été sélectionnés et calculés pour analyser des dimensions spécifiques de la performance, chacun avec une justification précise :

- **Indicateurs Offensifs :** Pour mesurer l'efficacité potentielle, nous avons utilisé l'**xG (Expected Goals)**, qui évalue la qualité des occasions de tir , l'**xA (Expected Assists)**, qui estime la probabilité qu'une passe devienne décisive , et le **SoT%**, qui reflète la précision technique des tirs.
- **Indicateurs Défensifs :** Pour évaluer l'intelligence défensive, nous avons retenu les **TklW (Tackles Won)**, qui mesurent la capacité à récupérer le ballon , et les **Int (Interceptions)**, qui indiquent la lecture du jeu et l'anticipation.
- **Indicateurs Créatifs :** La créativité a été mesurée via les **KP (Key Passes)**, qui sont les passes menant directement à un tir , et les **PrgP (Progressive Passes)**, qui font avancer significativement le jeu vers le but adverse.

3.2. Création de KPI Composites sur Mesure

En plus de ces indicateurs, nous avons élaboré quatre KPI composites uniques pour obtenir une vision synthétique et multidimensionnelle des profils des joueurs :

- **Defensive Impact Index (DII) :** Cet indice regroupe les actions défensives directes comme les tacles et les interceptions, tout en pénalisant les erreurs qui concèdent des occasions à l'adversaire.
- **Ball-Playing Index (BPI) :** Le BPI a été conçu pour mesurer la capacité d'un joueur, notamment un défenseur, à participer à la construction du jeu et à faire progresser le ballon proprement vers l'avant.
- **Discipline Risk Index (DRI) :** Cet indice quantifie le risque disciplinaire qu'un joueur représente, en combinant les fautes et les cartons reçus, car ces actions peuvent mettre l'équipe en danger.
- **Defensive Reliability (DefRel) :** Le DefRel évalue la fiabilité globale d'un joueur en phase défensive. Il intègre non seulement l'efficacité dans les tacles mais aussi la capacité à récupérer le ballon, tout en soustrayant les erreurs commises.

Partie 4 : Développement du Dashboard Interactif

Le livrable principal de ce projet est un dashboard d'analyse de données interactif, conçu pour permettre une exploration intuitive et approfondie des statistiques des joueurs. Pour sa réalisation, nous avons choisi la bibliothèque **Streamlit** de Python, en raison de sa capacité à créer rapidement des applications web data-driven robustes et esthétiques. L'interactivité des visualisations est assurée par la bibliothèque **Plotly**.

4.1. Structure et Ergonomie de l'Application

Le dashboard est structuré pour offrir une expérience utilisateur claire et efficace. La navigation s'articule autour de deux éléments principaux :

- Un **panneau de filtres latéral** qui permet à l'utilisateur d'affiner l'ensemble des données analysées selon plusieurs critères : la **position** (FW, MF, DF, GK), la **ligue**, la **tranche d'âge** et la **nationalité**.
- Un **système de sept onglets thématiques** qui segmente l'analyse et guide l'utilisateur à travers les différentes facettes du projet : "Vue d'ensemble",

"Par poste", "Comparaison", "Fiche joueur", "Ligues & Nations", "Analyse KPI", et "Méthodologie".

4.2. Fonctionnalités Clés Implémentées

Le dashboard intègre plusieurs fonctionnalités avancées pour répondre aux objectifs du projet :

- **Comparaison de Joueurs avec Profils Radar :** Une des fonctionnalités centrales est l'onglet **"Comparaison"**. Il permet de sélectionner jusqu'à quatre joueurs d'un même poste et de générer dynamiquement un **graphique radar**superposant leurs profils. Ce graphique utilise des statistiques normalisées (sur une échelle de 0 à 100) pour visualiser rapidement les points forts et les points faibles de chaque joueur sur des métriques clés. Un tableau détaillé accompagne le visuel pour une comparaison chiffrée.
- **Fiche Joueur Individuelle :** L'onglet **"Fiche joueur"** agit comme un rapport de scouting détaillé pour n'importe quel joueur du dataset. Il présente ses informations générales (club, âge, poste), ses statistiques de performance principales, et le compare à la moyenne des joueurs de son poste via un graphique radar et un diagramme en barres.
- **Analyse Interactive par Poste :** Dans l'onglet **"Par poste"**, l'utilisateur peut sélectionner un poste (ex: Attaquants) puis choisir deux métriques de performance à croiser (ex: Buts par 90 min vs Tirs cadrés par 90 min). L'application génère alors un **nuage de points interactif** (scatter plot) qui permet d'identifier des profils atypiques et des tendances au sein de la population de joueurs.
- **Analyse des KPI :** Un onglet est spécifiquement dédié à l'exploration des **KPI composites** créés dans la phase précédente. Il permet d'analyser leurs distributions, de visualiser leurs corrélations via une matrice thermique et d'identifier les équipes qui se démarquent en moyenne sur un indicateur donné.

En somme, le dashboard n'est pas seulement un outil de visualisation, mais une véritable plateforme d'analyse qui permet à un utilisateur, qu'il soit recruteur, analyste ou simple passionné, de naviguer dans la complexité des données de performance pour en extraire des conclusions pertinentes.

Partie 5 : Conclusion et Bilan

En conclusion, ce projet a été mené à bien et a permis d'atteindre les objectifs fixés, de la création d'un jeu de données fiable à la livraison d'un dashboard fonctionnel. Le principal défi a été notre manque de connaissance du domaine du football, ce qui a exigé un temps d'adaptation, ainsi que la nécessité de reconstruire notre dataset à partir de zéro. Le résultat est un outil d'analyse performant, dont la principale limite reste son analyse cantonnée à la unique saison 2023-2024.

Cette expérience nous a permis de maîtriser l'ensemble de la chaîne de valeur d'un projet data, de la gestion de données à leur visualisation. Forts de cet apprentissage, les améliorations futures s'orienteraient logiquement vers l'ajout de données plus granulaires (par match) et une optimisation du design du dashboard pour une meilleure expérience utilisateur.