Rapport : Python pour la Data

Table des matières

1. Contexte et objectifs

Le projet a pour but d'analyser des données avec Python (pandas), les étapes sont les suivantes :

- Choix du dataset public
- Chargement et nettoyage (gestion des valeurs manquantes, doublons, types)
- Analyses statistiques simples (moyenne, médiane, écart-type, corrélations)
- Création d'au moins trois visualisations

2. Choix du dataset

Le dataset provient de Kaggle : **Football Matches 2024/2025 – Top 5 Leagues (auteur : Tarek Masry)**, couvrant les grandes compétitions européennes sur la saison 2024/25.

Après chargement, le fichier comporte 1 941 lignes et 23 colonnes, incluant des variables numériques (scores, différences de buts, points, etc.) et catégorielles (noms d'équipes, compétition, statut du match, arbitre…).

3. Méthodologie

a) Chargement et nettoyage

- **Chargement** : lecture du fichier CSV avec Pandas.

- Normalisation légère : homogénéisation des noms de colonnes en snake_case (minuscules + underscore).
- **Contrôle des types** : vérification que les colonnes de score sont numériques *(fulltime_home, fulltime_away, total_goals, goal_difference)*.
- **Valeurs manquantes** : 5 valeurs manquantes au total, concentrées sur des champs non critiques *(referee : 3 ; halftime_home : 1 ; halftime_away : 1)*.
- **Doublons** : 0 doublon au niveau ligne, et 0 doublon sur la clé logique *(match_id + date_utc + home_team + away_team + competition_name)*.

b) Analyse statistiques simple

b.1 - Statistiques globales

Buts totaux par match *(total_goals)* : moyenne 2,876, médiane 3, écart-type 1,694, min 0, max 11.

Buts domicile *(fin de match) (fulltime_home)* : moyenne 1,540, médiane 1, écart-type 1,297, min 0, max 9.

Buts extérieur *(fin de match) (fulltime_away)* : moyenne 1,336, médiane 1, écart-type 1,206, min 0, max 7.

Différence de buts *(goal_difference)* : moyenne 0,196, médiane 0, écart-type 1,881, min −6, max 8.

b.2 – Corrélations (Pearson)

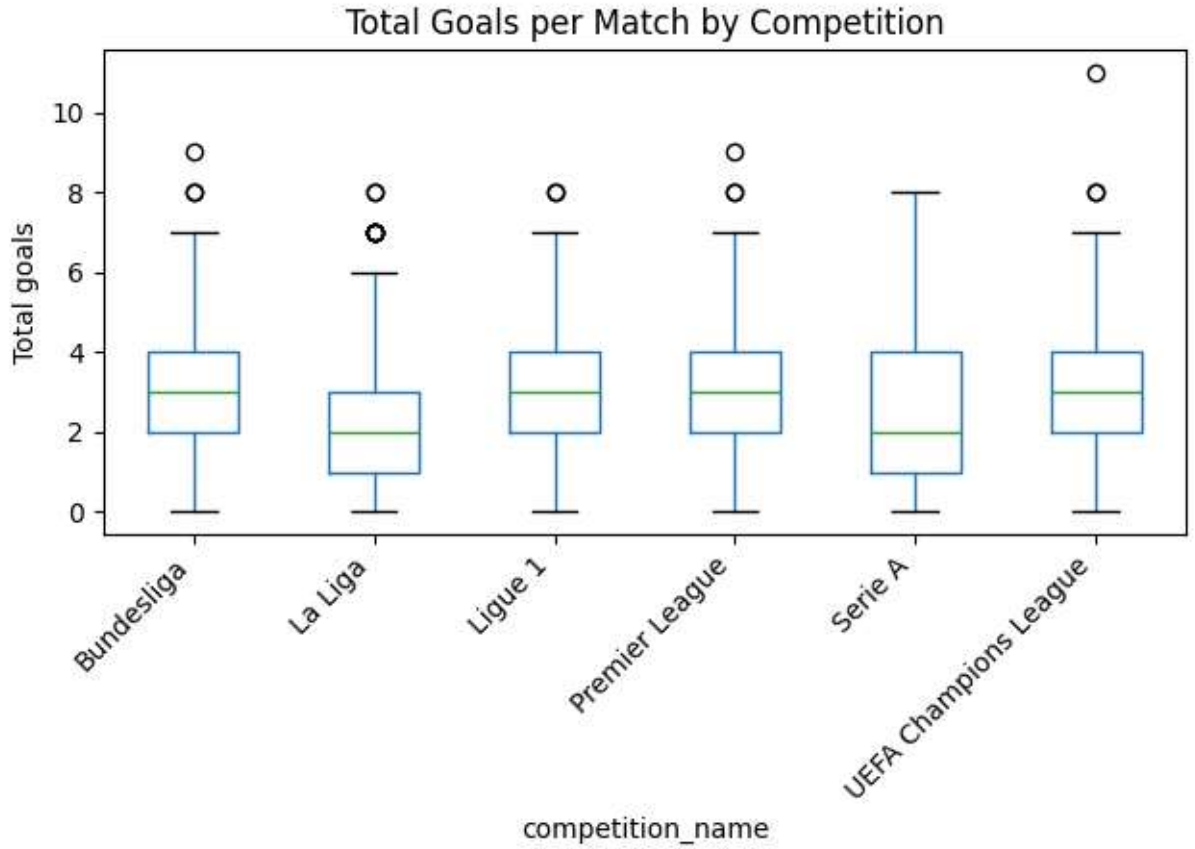
- *fulltime_home* ↔ *fulltime_away* : **r = −0,105** → corrélation faible, quasi nulle (le fait de marquer à domicile n'implique pas que l'adversaire marque davantage ou moins).
- *total_goals* ↔ *goal_difference* : **r = 0,093** → relation très faible (un total élevé n'induit pas forcément une victoire large).

b.3 – Comparaison par compétition (moyenne de *total_goals*)

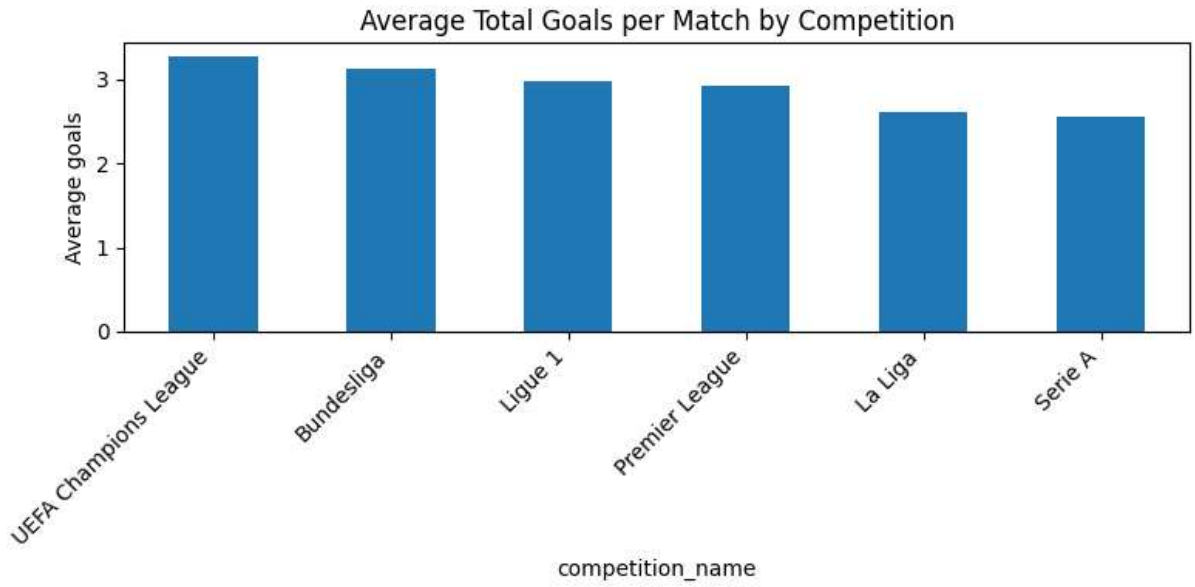
- **UEFA Champions League** : 3,328
- **Bundesliga** : 3,134
- **Ligue 1** : 2,977
- **Premier League** : 2,934
- **La Liga** : 2,618
- **Serie A** : 2,561

Ce qui est intéressant ici est que ces écarts reflètent des styles de jeu et dynamiques différentes selon les compétitions (Bundesliga = style offensif, Serie A = style défensif). Je compare de Bundesliga à Serie A car se sont des ligues nationales, et la UCL est la coupe internationale des clubs, le plus intéressant ici est de comparer par pays)

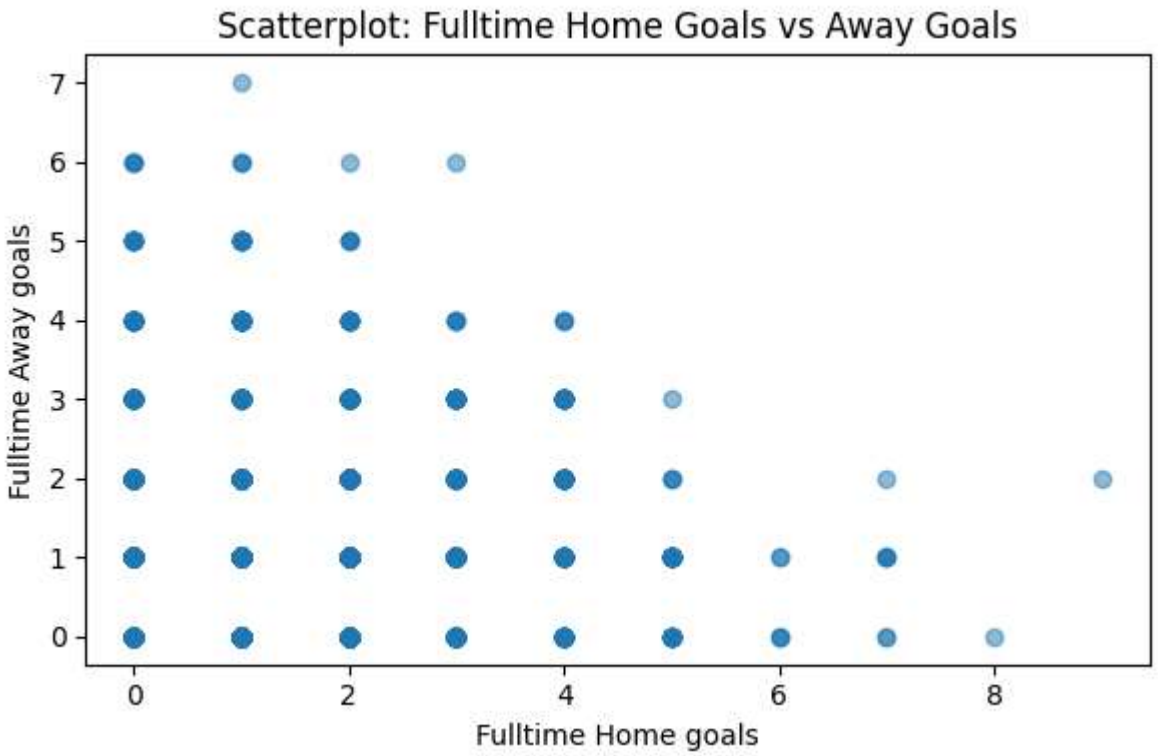
4. Visualisations



L'histogramme *(total_goals)* met en évidence une grosse concentration entre 2 et 4 buts, avec des scores extrêmes rares.



Le diagramme en barre de la moyenne de *total_goals* par *competition_name* montre clairement les différences dans le style de jeu, avec l'UCL et la Bundesliga qui sont les plus offensives.



Scatter plot *fulltime_home vs fulltime_away* : nuage de points diffus, ce qui confirme la faible corrélation entre les buts marqués par l'équipe à domicile et l'équipe visiteuse.

5. Conclusion

L'analyse des 1 941 matchs de la saison 2024/25 montre une moyenne de 2,88 buts par rencontre, avec une concentration entre 2 et 4 buts. Les différences par compétition sont marquées : l'UCL et la Bundesliga apparaissent plus offensives, tandis que la Serie A confirme son style plus défensif. Les corrélations entre buts à domicile et à l'extérieur sont très faibles, ce qui souligne l'indépendance des performances des deux équipes. Les trois visualisations produites (histogramme, barres par ligue, scatter plot) montrent clairement ces tendances.