FACULTÉ DES SCIENCES ECONOMIQUES

Université de Rennes

Projet de Scraping

Analyse du site *Transfermarkt*

Introduction

Notre projet traite de la collecte des données, et de leur analyse, sur le site référence en matière de football, *Transfermarkt*. Ce site d'origine allemande, est reconnu pour le grand nombre de statistiques et de données concernant plus de 840 000 joueurs. Fondé en 2000, *Transfermarkt* est célèbre pour son classement de la valorisation des joueurs. Notre étude porte donc, sur les 500 joueurs les mieux valorisés par le site.
Dans un premier temps, nous présenterons la base de données, puis dans un second temps, nous évoquerons le programme python permettant de collecter les données, en expliquant comment il fonctionne et quelles ont été les difficultés que nous avons rencontrées lors de ce projet. Enfin dans une dernière partie, nous ferons une analyse statistique et graphique afin d'avoir une meilleure vision sur notre base de données.

Présentation des données

Notre base de données comprend les informations sur les 500 joueurs de football les plus valorisés en euros en 2022, collectées à partir d'un programme de scraping Python. Les variables de cette base de données comprennent :

- Nom et Prénom du joueur : le nom complet du joueur.
- Âge : l'âge du joueur à la date actuelle.
- Nationalité : la nationalité du joueur.
- Continent de naissance : le continent où le joueur est né.
- Taille : la taille du joueur en mètres.
- Position : la position sur le terrain que le joueur occupe le plus souvent.
- Pied : le pied avec lequel le joueur préfère jouer.
- Club : le club actuel pour lequel le joueur joue.
- Ligue : le championnat dans lequel le club du joueur évolue.
- Fin de contrat : la date à laquelle le contrat actuel du joueur prendra fin.
- Valeur actuelle : la valeur du joueur en euros selon les estimations actuelles.
- Valeur maximale : la valeur maximale que le joueur a atteinte au cours de sa carrière en euros.
- Coupe du monde : s'il a gagné ou non la Coupe du Monde de la FIFA.
- Nombre de victoires de la Ligue des champions : le nombre de fois que le joueur a remporté la Ligue des champions de l'UEFA.
- Nombre de sélections : le nombre de fois que le joueur a été sélectionné pour jouer pour son équipe nationale.
- Nombre de buts : le nombre de buts marqués par le joueur pour son équipe nationale.

A partir de ces variables, nous pourrons ainsi réaliser une analyse statistique et graphique pour une meilleure évaluation des forces en présence dans le football mondial et comprendre ainsi les facteurs qui peuvent influencer la valorisation des joueurs.

Master SDIN
Enseignant : Joeffrey DROUARD
2022/2023

Killian CHEFTEL
Arthur DIAZ

Programme Python

Présentation des codes

Pour ce projet de scraping, nous avons choisi d'utiliser deux méthodes différentes pour collecter les informations sur les 500 joueurs les plus valorisés sur le site *Transfermarkt.fr*.
La première méthode implique l'utilisation de la librairie BeautifulSoup, qui se révèle plus intuitive une fois son fonctionnement compris. Cette librairie permet de parcourir et de structurer les données HTML de manière à en extraire les informations souhaitées. Cependant, il faut noter que la mise en œuvre de cette méthode peut prendre plus de temps et nécessite aussi une connaissance approfondie du fonctionnement de la librairie, mais aussi du code html.
La deuxième méthode, quant à elle, implique l'utilisation d'expressions régulières. Afin de faciliter notre travail nous avons créé deux fonctions pour rendre l'utilisation des expressions régulières plus simple. Cette méthode est plus facile à réaliser et ne nécessite pas autant de connaissances spécifiques du code html. Cependant, elle se révèle plus longue en termes de temps de collecte des données. Pour éviter de devoir relancer le programme depuis le début à chaque fois, nous avons décidé de le décomposer le programme en 4 comme des checkpoints. Ainsi nous pouvions reprendre là où nous nous étions arrêtés sans perdre de temps.
En choisissant ces deux méthodes, nous avons voulu comparer leur efficacité et leur facilité d'utilisation. Nous sommes arrivés à la conclusion que la méthode avec BeautifulSoup est plus adaptée pour les projets de scraping requérant une plus grande précision et une structure de données bien définie, tandis que la méthode avec des expressions régulières est plus appropriée pour les projets requérant une collecte rapide et simple des données.
Il faut savoir que même si le code est différent dans les 2 programmes, la démarche est identique. En fait nous avons commencé par collecter les pages où s'affichent l'url de chaque joueur. Puis pour chaque url collectée, nous avons scrapé les informations que nous voulions pour faire nos analyses.
Pour ce qui est de la partie analyse de la base de données, nous avons utilisé plusieurs librairies pour réaliser nos tableaux et graphiques. Les principales librairies sont pandas et matplotlib. Mais afin d'approfondir notre analyse, nous avons aussi utilisé *geopandas* pour réaliser une analyse géospatiale du nombre de joueurs par pays et *seaborn* qui complète les fonctionnalités de matplotlib (graphiques supplémentaires, personnalisation des graphiques …)

Difficultés rencontrées

Lors de la réalisation de notre projet de scraping de données sur le site *Transfermarkt.fr*, nous avons rencontré plusieurs difficultés.
La première difficulté a été de trouver un site où nous pouvions effectuer du scraping. Certaines plateformes telles que *www.futbin.com* et *store.epicgames.com* nous ont bloqué l'accès, tandis que d'autres nous ont demandé une autorisation, mais nous n'avons reçu aucune réponse à nos demandes. De plus, nous avons constaté que certaines informations étaient manquantes sur des sites tels que *www.metacritic.com*, ce qui a rendu le traitement des données plus difficile. Finalement, nous avons choisi *Transfermarkt.fr* qui a répondu à nos besoins en matière de scraping.
Une autre difficulté a été de collecter des informations à l'aide de *BeautifulSoup*. Certaines informations étaient manquantes lorsque nous avons passé les pages en format texte, ce qui a ainsi rendu la collecte difficile. Pour résoudre ce problème, nous avons collecté d'autres url pour accéder à l'information manquante.
En ce qui concerne la méthode des expressions régulières, la première difficulté rencontrée était liée à l'enregistrement des url des pages des joueurs. La conception des url était basée sur des « / » ce qui rendait difficile l'enregistrement des pages car le logiciel considérait les pages comme des dossiers. De plus, nous avons rencontré des problèmes concernant les données vides. Certains joueurs n'ont jamais été transférés ou n'ont pas remporté la ligue des champions ou la coupe du monde. Les données étaient donc considérées comme manquantes. Aussi pour résoudre ce problème, nous avons utilisé des conditions avec la fonction « if » pour gérer les données manquantes.

Analyse des données

Pour explorer les données que nous avons rassemblées, nous avons utilisé deux approches différentes d'analyse de données : l'analyse statistique et l'analyse graphique. L'analyse statistique nous a permis de mesurer et de comprendre les tendances générales dans les données. L'analyse graphique, quant à elle nous a aidés à mieux visualiser les données de manière attrayante et facile à comprendre.
Ensemble, l'analyse statistique et l'analyse graphique nous ont facilité la compréhension en profondeur des données que nous avons collectées sur les joueurs les plus valorisés du football mondial, ce qui a été essentiel pour élaborer des conclusions solides et informées sur ce sujet.

Analyse statistique

Au cours de cette section, nous avons utilisé des techniques statistiques telles que la moyenne, la variance et la distribution pour comprendre les données de manière quantitative. L'analyse statistique nous a permis de fournir des résultats précis et fiables sur les joueurs les plus valorisés dans le football mondial.

Au sein de notre échantillon de 500 joueurs, nous notons que l'âge moyen est d'environ 25 ans (24,94), pour des individus compris entre 16 et 37 ans.
Nous observons également que parmi les 500 joueurs les mieux valorisés, 25% d'entre eux ont une valeur supérieure à 40 millions d'euros, alors que la moyenne est de 35,68 millions. Cette donnée nous indique que la moyenne est fortement impactée par les joueurs ayant une valeur très élevée, comme nous le montre la médiane qui n'est que de 28 millions d'euros.
De même, nous voyons qu'en moyenne, les joueurs de notre base, ont été transférés au cours de leur carrière pour une somme totale de 33,47 millions en moyenne. Cette somme peut-être très variable selon les joueurs. En effet certains d'entre eux jouent toujours dans leur club formateur, n'ont connu que peu de clubs ou n'ont jamais été transférés. Aussi l'historique de leur transfert est plutôt bas, voire égal à 0. C'est pour cette raison qu'un quart des joueurs présente un historique de transferts inférieur à 7 millions d'euros.

Tableau 1 – Statistiques descriptives de la base de données

	Age	Taille	Fin contrat	Valeur actuelle	Valeur max	Montant transfert	CDM	LDC	Nombre sélections	Nombre buts
Count	500	500	498	500	500	500	500	500	492	492
Mean	24,94	1,82	2025,49	35,68	45,83	33,47	0,05	0,18	26,64	5,00
Std	3,12	0,07	1,41	20,52	28,68	38,88	0,23	0,64	26,94	11,35
Min	16	1,65	2023	18	18	0	0	0	0	0
25%	23	1,78	2024	22	25	7	0	0	7	0
50%	25	1,83	2026	28	35	22,80	0	0	18	2
75%	27	1,87	2027	40	60	45,96	0	0	40	5
Max	37	2	2031	180	200	333,36	1	5	196	118

Les joueurs ayant gagné la coupe du monde représentent 5,4% des joueurs de notre étude (*Tableau 2*). Leur valeur est de 44,44 millions d'euros en moyenne, contre 35 millions pour ceux ne l'ayant pas gagné. Logiquement, ces joueurs comptent en moyenne 22 sélections de plus que les autres.
Nous notons qu'en moyenne, les contrats des joueurs se finissent en 2025. Si nous regardons les statistiques en fonction de la date de fin de contrat, nous notons que les joueurs ayant encore 3 ans de contrat (c'est-à-dire jusqu'en 2026), ont plus de 25 ans en moyenne. Cette valeur oscille autour de 24 ans pour les joueurs en contrat jusqu'en 2027 et 2028. Enfin, au-delà de cette date, les joueurs sont âgés entre 21 et 22 ans en moyenne.

Tableau 2 –Moyennes des variables en fonction en fonction du titre de champion du monde

CDM	Age	Taille	Fin contrat	Valeur actuelle	Valeur max	Montant transfert	LDC	Nombre sélections	Nombre buts
0	24,81	1,82	2025,51	35,18	44,26	32,34	0,15	25,45	4,72
1	27,15	1,80	2025,19	44,44	73,41	53,27	0,59	47,07	9,85

Si nous nous intéressons maintenant aux statistiques des joueurs par rapport au nombre de ligues des champions gagnées (*Tableau 3*), 58 joueurs l'ont gagnée, soit 11,6% de la base. Nous remarquons que logiquement, les joueurs ayant gagné le plus de ligues des champions sont plus âgés. En effet les joueurs qui ont quatre ou cinq titres ont plus de 30 ans en moyenne, contre 28 ans pour les joueurs ne l'ayant gagné que deux ou trois fois. Nous observons également que parmi ces joueurs multiples champion d'Europe, leur valeur actuelle est supérieure aux autres. Ceci nous montre qu'une victoire a un impact important sur la valorisation d'un joueur. De même en regardant de plus près la valeur maximale des joueurs au cours de leur carrière, nous observons qu'à partir de trois titres gagnés, cette somme dépasse les 77,5 millions d'euros, ce qui vient confirmer l'impact important d'un trophée de la ligue des champions sur la valeur d'un joueur.

Tableau 3 – Moyennes des variables en fonction du nombre le ligue des champions gagnées

LDC	Age	Taille	Fin contrat	Valeur actuelle	Valeur max	Montant transfert	CDM	Nombre de sélections	Nombre de buts
0	24,62	1,82	2025,54	33,89	41,67	31,07	0,05	23,13	3,86
1	26,57	1,83	2025,30	52,79	76,89	47,92	0,06	43,38	9,60
2	28,00	1,89	2024,00	22,50	42,50	96,15	0,00	36,00	15,00
3	28,00	1,81	2024,50	40,00	77,50	2,03	0,00	66,50	8,50
4	30,67	1,79	2024,00	43,33	101,67	48,33	0,67	118,67	35,33
5	33,75	1,85	2024,25	31,25	85,00	102,26	0,25	117,00	44,50

Dans la continuité de notre analyse, si nous regroupons les joueurs en fonction de la date de fin de leur contrat (*Tableau 4*), nous remarquons que plus le contrat est long, plus leur valeur est élevée. Nous noterons que les années 2030 et 2031 ne concernent que deux joueurs. Nous pouvons également analyser que plus la durée de contrat augmente, plus l'âge moyen des joueurs diminue. C'est donc naturellement que nous observons que des variables fortement corrélées avec l'âge du joueur, comme le nombre de sélections, de buts en sélection et de ligues des champions, baissent lorsque la durée du contrat augmente.

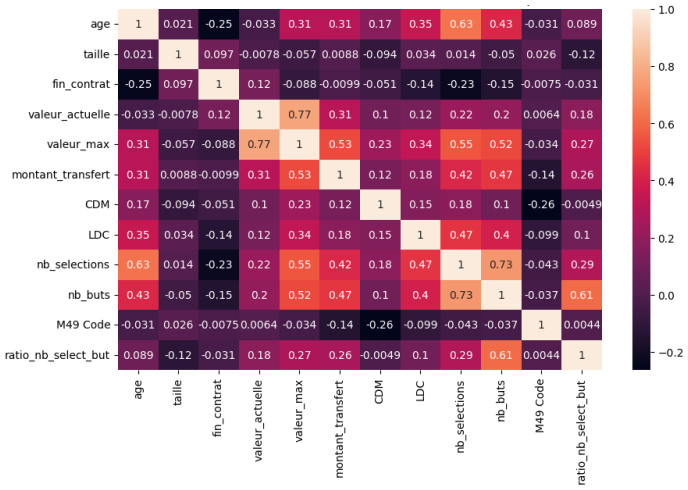
Tableau 4 – Moyennes des variables en fonction de la date de fin de contrat

Fin de contrat	Age	Taille	Valeur actuelle	Valeur max	Montant transfert	CDM	LDC	Nombre sélections	Nombre buts
2023	25,91	1,80	27,15	51,92	42,48	0,09	0,42	34,66	8,53
2024	25,79	1,82	35,75	45,53	29,61	0,07	0,17	31,82	4,87
2025	25,32	1,82	37,42	50,75	34,75	0,04	0,24	31,65	7,23
2026	25,04	1,83	34,99	42,39	28,82	0,05	0,17	26,53	4,03
2027	23,74	1,82	37,78	42,62	34,37	0,06	0,05	17,33	3,09
2028	24,07	1,86	42,87	48,87	39,49	0	0,07	16	1,86
2029	21,5	1,815	45	45	67,75	0	0	2,5	0
2030	21	1,94	40	40	38	0	0	2	0
2031	22	1,75	40	40	70	0	0	8	0

Dans la continuité de l'analyse des données que nous avons collectées, nous avons utilisé ces données pour explorer et visualiser les tendances et les relations qui peuvent être trouvées parmi les joueurs.

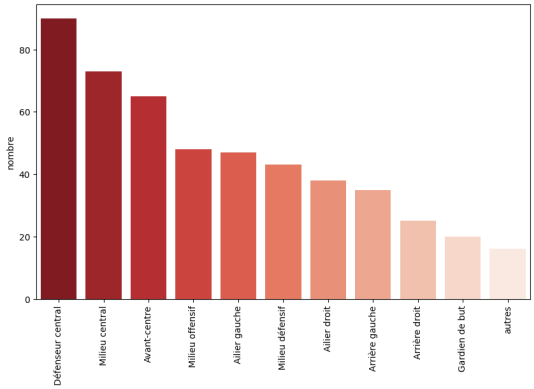
Afin de voir s'il était pertinent de faire des graphiques montrant la corrélation entre des variables, nous avons choisi de réaliser une matrice de corrélation avec toutes les variables (*Graphique 1*). Nous observons que peu de variables sont corrélées, ce qui limite donc notre analyse graphique.

Graphique 1 – Matrice de corrélation entre les différentes caractéristiques



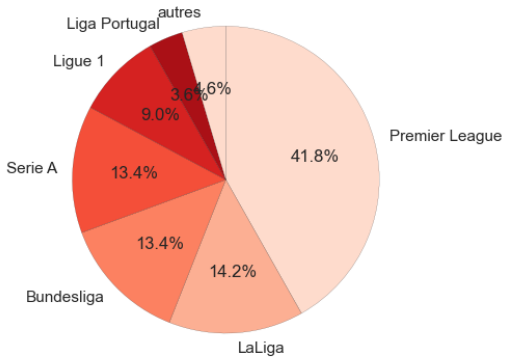
En regardant de plus près les corrélations entre les différentes variables, nous observons que l'âge impacte fortement le nombre de sélections du joueur de même que le nombre de buts marqués. Nous pouvons noter que le nombre de sélections joue un rôle essentiel dans le nombre de buts inscrits puisque plus le joueur joue de matchs pour sa sélection, plus il aura de chances de marquer un but. A contrario, la relation entre l'âge et la fin de contrat est négative comme nous avons pu le constater précédemment dans notre analyse statistique.
De plus, nous constatons que la valeur actuelle est fortement impactée positivement par la valeur maximale. Cette valeur maximale est reliée avec le montant total des transferts, et également le nombre de sélections et de buts.
En somme, grâce à cette matrice de corrélation, nous pouvons constater que certaines valeurs sont fortement liées. Ainsi l'âge est fortement corrélé avec le nombre de matchs (0,63) et de buts (0,43) en sélection nationale. Mais on peut aussi noter que comme l'âge, la variable valeur maximale est fortement corrélée avec les variables valeur actuelle (0,77) et montant des transferts (0,53).
Après avoir observé les corrélations entre les variables, nous avons analysé les informations qui ressortent de celles-ci.
Premièrement, nous pouvons observer que les défenseurs centraux et les milieux centraux sont les plus nombreux (*Graphique 2*). Cela peut s'expliquer par le fait que de manière générale, dans les équipes il y a toujours au moins deux joueurs qui occupent ces postes.

Graphique 2 – Nombre joueurs en fonction de leur poste



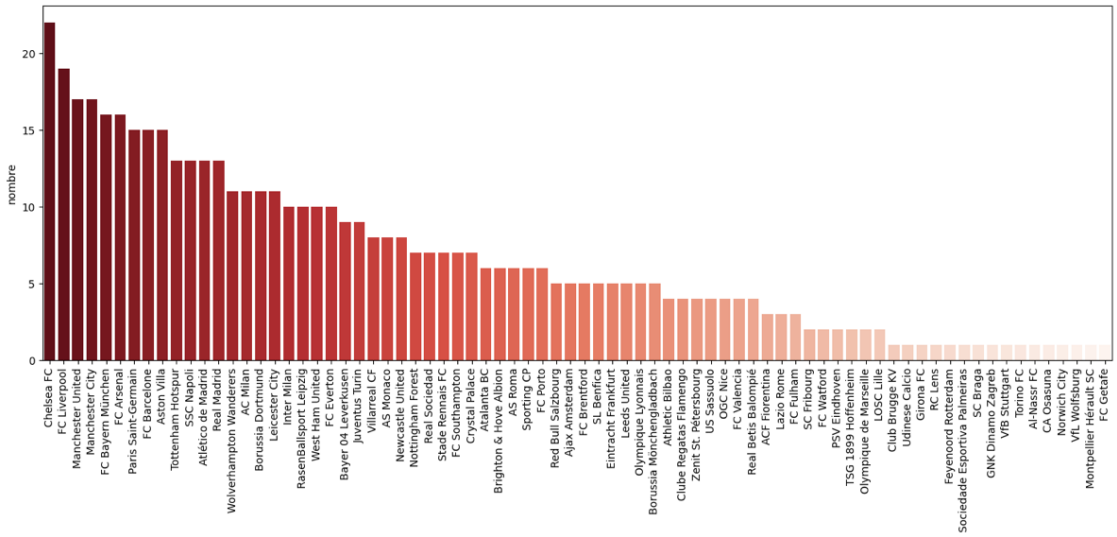
De plus, nous remarquons que sans surprise, c'est la Premier League, le championnat anglais qui compte le plus de joueurs parmi les plus valorisés (*Graphique 3*). Nous pouvons observer que plus de 40% des joueurs de notre échantillon, jouent dans le championnat anglais. Ce résultat suit une certaine logique puisque les revenus des clubs anglais sont très largement supérieurs à ceux des autres clubs européens (LES ECHOS (2022), *Droits télévisés : la Premier League creuse l'écart.*).

Graphique 3 – Répartition des joueurs en fonction du championnat



Cette tendance se confirme au niveau des clubs les mieux représentés (*Graphique 4*). Chelsea est le club comptant le plus de joueurs parmi les 500 les plus chers du monde. Pour illustrer la surpuissance financière anglaise, nous voyons qu'Aston Villa, 11ème actuellement du championnat se place devant le Real Madrid vainqueur en titre de la Ligue des Champions. De surcroît, parmi les dix premiers, nous retrouvons sept clubs anglais, qui comptent chacun plus de treize joueurs.

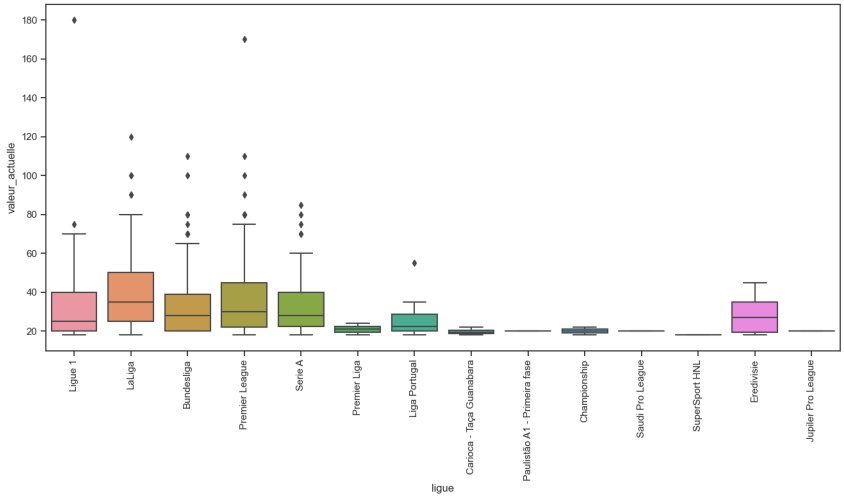
Graphique 4 – Nombre de joueurs que compte chaque club dans la base de données



Cependant, si nous nous intéressons aux valeurs médianes, grâce aux boites à moustache (*Graphique 5*), nous observons que le championnat espagnol se démarque un peu des autres. En plus d'une médiane supérieure, nous remarquons que le top 25% des joueurs les mieux valorisés du championnat est supérieur, au championnat anglais par exemple. Ce résultat, qui peut paraître étonnant au premier abord, compte tenu des analyses précédentes, peut s'expliquer par la présence du Real Madrid, du FC Barcelone, et de l'Atletico Madrid, qui sont des grands clubs, possédant donc des joueurs de valeurs élevées. Nous pouvons aisément imaginer que ce sont les joueurs de ces clubs qui représentent les 25% les mieux évalués. Ce qui fait vraisemblablement baisser la valeur du troisième quartile du championnat anglais, est le nombre de joueurs plus important dans le classement ce qui nivèle les valeurs.

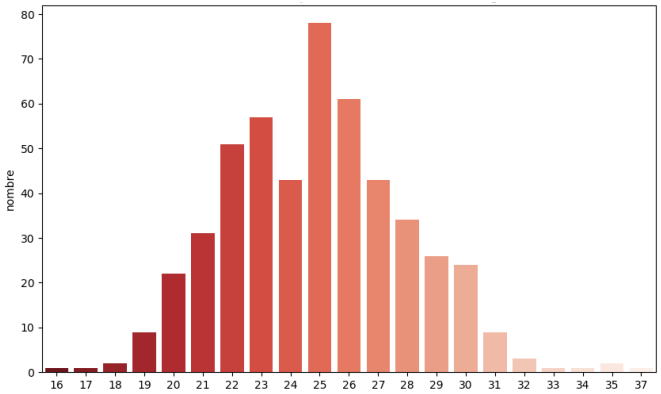
Nous noterons également que le championnat français est fortement inégal. En effet, la médiane est proche du premier quartile. Cela s'explique par le fait que les joueurs du PSG sont bien mieux valorisés, avec notamment Kylian Mbappé, évalué à 180 millions d'euros.

Graphique 5 – Boites à moustache des valeurs des joueurs par championnat



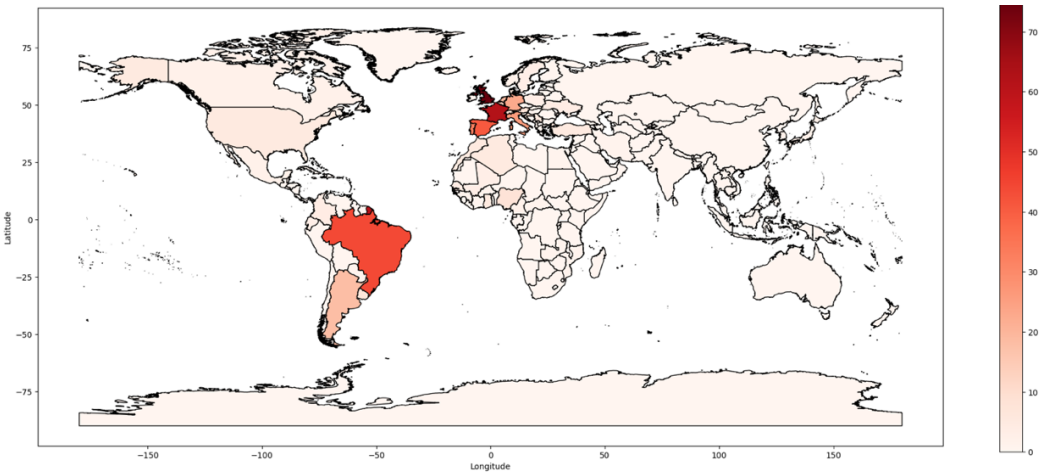
De plus, nous pouvons remarquer à l'aide du graphique ci-dessous, que parmi les 500 joueurs les mieux valorisés, ceux qui ont 25 ans sont les plus nombreux. On observe que la majorité des joueurs ont entre 22 et 27. Ce qui correspond à la période où les sportifs sont dans la force de l'âge.

Graphique 6 – Histogramme du nombre de joueurs en fonction de leur âge



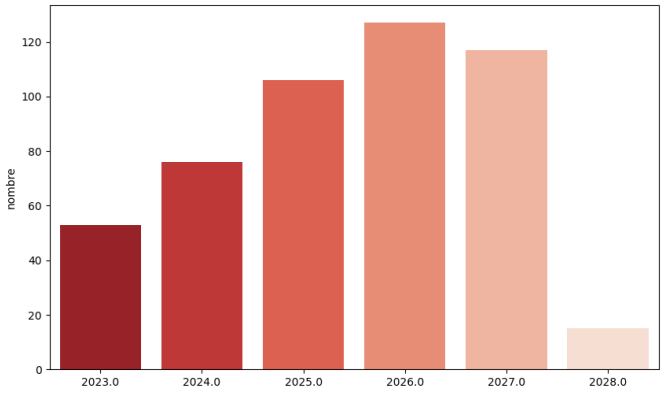
Grâce à cette carte du monde (graphique 7), nous observons que les joueurs les plus chers viennent majoritairement d'Europe, et plus particulièrement de France et de Grande-Bretagne. De même le continent sud-américain est bien représenté, notamment par le Brésil et l'Argentine. L'Afrique compte de nombreux représentants. Mais aucun pays ne se détache réellement. Nous pouvons cependant identifier l'Afrique du Nord et l'Afrique de l'Ouest comme les régions où les densités de joueurs sont les plus fortes.

Graphique 7 – Carte de chaleur montrant la répartition des joueurs par pays



Dans le calcul de la valeur d'un joueur, la durée de contrat restante est une variable importante. En effet, plus le joueur est lié longtemps à son club, plus il sera difficile et coûteux de racheter ses années de contrats à son club. C'est donc logiquement, que nous pouvons observer que la plupart des joueurs les mieux valorisés, possèdent des contrats longs allant jusqu'à 2026 ou 2027. Il est intéressant de noter que peu de joueurs sont sous contrat de très longue durée. Ainsi moins de 20 joueurs ont signé un contrat allant jusqu'en 2028 (soit 5 ans). Cela peut s'expliquer par une logique de renégociations de contrats afin d'obtenir un meilleur salaire pour les joueurs, qui en général ont plus de pouvoir lors des négociations.

Graphique 6 – Répartition des joueurs en fonction de leur date de fin de contrat



En somme, au cours de cette analyse sur les 500 joueurs les mieux valorisés selon le site Transfertmarkt, nous avons observé que certains facteurs influencent la valeur des joueurs. En effet, le montant total des transferts précédents, le nombre de sélections et la valorisation maximale du joueur impactent positivement l'estimation de la valeur d'un joueur. De plus, nous avons remarqué qu'une part importante des joueurs évolue dans le championnat anglais et que la grande majorité des joueurs de la base de données sont originaires d'Europe ou d'Amérique du Sud.
Enfin, les analyses graphiques et statistiques réalisées, nous ont permis de mieux comprendre les facteurs influençant la valorisation d'un joueur et de dessiner ainsi un panorama des forces en présence dans le football mondial.