Université Gustave Eiffel

UFR SCIENCES ÉCONOMIQUES ET GESTION

IAE PARIS-EST École de management

UNIVERSITÉ PARIS-EST CRÉTEIL · UNIVERSITÉ GUSTAVE EIFFEL

Master 1 Data Analyst

Comment remporter un championnat de Football ?

GOUPIL Victor

2023-2024

Table des matières

Introduction

Plongeons ensemble dans l'univers palpitant des statistiques du ballon rond, là où chaque donnée devient une clé pour percer les mystères du jeu. C'est désormais ce que demande le sport le populaire au monde : le football. En effet, le football moderne semble attacher une importance non négligeable au monde des statistiques. Ces dernières sont indicatrices de performances, que ce soit individuellement ou collectivement. Elles peuvent être utilisées pour aborder tactiquement un match de la meilleure des manières, ou bien pour recruter des joueurs connus ou méconnus qui semblent se démarquer par les chiffres. Elles sont aussi prises en compte lorsqu'il s'agit de récompenser les meilleurs joueurs par des trophées individuels. Notre étude va se focaliser sur les statistiques collectives, c'est-à-dire par équipe. A l'aide de nombreuses variables, nous tenterons de répondre à la question suivante : Comment remporter un championnat de football ? L'objectif de l'étude est de repérer les éléments clés qui distinguent les équipes les mieux classées de leurs championnats. Les motivations de cette recherche sont enracinées dans la passion de ce sport et dans la poursuite perpétuelle d'atteindre ou de comprendre le fonctionnement du plus haut niveau. L'objectif est d'offrir une vision éclairée et un guide pratique pour les fans ou pour les équipes aspirant à la consécration.

La base de données

Nous disposons d'une base de données construite manuellement à partir du site Fbref qui regroupe une infinité de statistiques sur le football. Pour la bâtir, nous avons pour chaque saison mis bout à bout trois tableaux de statistiques présents sur le site : Le tableau des statistiques générales contenant notamment le classement des équipes et leurs statistiques globales sur la saison, un tableau avec des indicateurs de statistiques plus élaborés que nous expliquerons en détail, et un tableau regroupant les statistiques des gardiens pour chaque équipe. Nous avons donc les statistiques des cinq grands championnats européens (Premier League, La Liga, Serie A, Bundesliga, Ligue 1) sur quatre saisons différentes (2017-2018, 2018-2019, 2021-2022, et 2022-2023). Au total, notre base de données dispose de 392 observations et 55 variables. Avant de décrire les variables, nous allons procéder au nettoyage de la base de données. En effet, certaines variables sont en double car elles étaient dans les deux tableaux. De plus, certaines variables ne sont pas utiles à notre étude. Enfin, nous allons privilégier les variables en pourcentage ou bien celles rapportées sur 90 minutes pour ne pas fausser notre analyse car certaines équipes jouent moins de match (en Allemagne, le championnat est constitué de 18 équipes contre 20 dans les quatre autres).

Dans un premier temps, nous allons choisir les variables que nous gardons parmi celles sur les gardiens de buts (colonnes 44 à 55). Elles sont au nombre de 3 : Le nombre de pénaltys subit, le pourcentage d'arrêts et le pourcentage de Clean Sheets. Ensuite, nous faisons la même chose pour le tableau des statistiques avancées (colonnes 21 à 43). Ici, nous enlevons seulement 5 variables : Les passes décisives car nous les avons déjà dans une autre colonne, rapportées sur 90 minutes. Le nombre de pénalty marqués car nous préférons la variable du nombre de penalty tentés, et les npxG, xAG et les npxG+xAG déjà présentes dans d'autres colonnes rapportées sur 90 minutes. Enfin, nous attaquons le tableau des statistiques globales (colonnes 1 à 20). Ici, nous nous allons nous servir de certaines variables pour en créer d'autres qui seront très intéressantes pour notre étude. Nous ajoutons donc les variables G - xG et GA – xGA, que nous expliquerons dans la description des variables ci-dessous :

La description des variables

Notre base finale de données est donc constituée de 392 observations et 45 variables. Voici pour chaque variable une petite explication :

- Équipe = Le nom des équipes.
- Championnat = Nom du championnat où l'équipe évolue.
- Classement = La position des équipes au classement à la fin de la saison (de 1 à 20).
- Matchs_joués = Nombre de matchs joués sur une saison.
- Win = Nombre de victoires.
- Draw = Nombre de matchs nuls.
- Lose = Nombre de défaites.
- Buts_marqués = Nombre de buts marqués.
- Buts_encaissés = Nombre de buts encaissés.
- Différence_buts = Différence de buts (Nombre de buts marqués – Nombre de buts encaissés).
- Points = Nombre de points sur la saison (Victoire=3pts, Nul=1, Défaite=0).
- Pts_par_match = Nombre de points par match (Points/ Nombre de matchs).
- xG = Expected Goals : Outil statistique de plus en plus répandu dans le milieu footballistique. Il sert à estimer le nombre de buts qu'un joueur ou une équipe « aurait dû » marquer. Autrement dit, il s'agit du nombre de buts attendus. Si une équipe a 50 xG sur une saison mais a marqué seulement 40 buts, on dit qu'elle a sous-performée. Cela peut s'expliquer par un manque de chance, ou bien un manque d'efficacité des attaquants. Cet outil est calculé par des algorithmes selon des critères bien précis pour chaque tir effectué. Un tir reçoit un xG toujours comprit entre 0 et 1. On peut donc interpréter ce coefficient comme la probabilité que le tir finisse dans le but.
- xGA = Expected Goals Against : Moyenne de buts que l'équipe « aurait du » concéder. Si une équipe a 50 xGA sur une saison mais a encaissé seulement 40 buts, on dit qu'elle a surperformée. Cela peut s'expliquer par de la chance ou un manque d'efficacité des attaquants adverses.
- xGD = Expected Goal Difference (xG – xGA). Si une équipe a une différence de buts négative mais une xGD positive, elle a manqué de chance ou d'efficacité.
- xGD/90min = Expected Goal Difference par match rapporté sur 90 minutes.

- Affluence = Moyenne de spectateurs présents dans le stade lors des matchs à domicile de l'équipe concernée (exprimé en milliers).

- Meilleur_buteur = Nom du meilleur buteur de l'équipe sur la saison.

- Buts_meilleur_buteur = Nombre de buts marqués par le meilleur buteur sur la saison.

- Gardien_de_but = Nom du gardien de but.

- Année = Une saison chevauchant toujours deux années, nous avons choisi de prendre l'année où la saison s'est terminée (ex : saison 2017-2018, on note 2018).

- Joueurs_champs = Nombre de joueurs différents utilisés au total pendant la saison.

- Âge = Moyenne d'âge de l'équipe sur la saison.

- Possession = Moyenne de la possession de l'équipe sur la saison (en %).

- Pen_tentés = Nombre de penalties obtenus par l'équipe sur la saison.

- Cartons_jaunes = Nombre de cartons jaunes subit par l'équipe sur la saison.

- Cartons_rouges = Nombre de cartons rouges subit par l'équipe sur la saison.

- Possessions_progressives = Nombre de possessions progressives totales sur la saison. Une possession progressive est une possession qui rapproche le ballon de la ligne de but adverse d'au moins 9 mètres par rapport à l'endroit le plus éloigné où le ballon se trouvait lors des six dernières passes. Ou bien il peut aussi s'agir de toute possession au sein de la surface de réparation adverse. La statistique exclut les possessions qui sont mises en échec dans la moitié défensive du terrain

- Passes_progressives = Nombre de passes progressives totales sur la saison. Une passe progressive est une passe réussie qui rapproche le ballon de la ligne de but adverse d'au moins 9 mètres par rapport à l'endroit le plus éloigné où le ballon se trouvait lors des six dernières passes. Ou bien il peut aussi s'agir de toute passe réussie dans la surface de réparation. La statistique ne concerne pas les passes tirées depuis la zone de défense (40 % du terrain à partir du but)

- Buts/90min = Moyenne de buts par match.

- PD/90min = Moyenne de passes décisives par match.

- (BP+PD)/90min = Moyenne de buts + passes décisives par match.

- (BM-PénM)/90min = Moyenne de buts marqué sans les pénaltys par match.

- (BP+PD-PénM)/90min = Moyenne de buts + passes décisives sans les pénaltys par match.

- xG/90min = Moyenne de buts que l'équipe aurait dû marquer par match.

- xAG/90min = Expected assisted goals. Il s'agit des buts assistés attendus. Cela indique la capacité d'un joueur à créer des opportunités pour marquer un but sans avoir à compter sur le résultat actuel du tir ou sur la chance ou l'efficacité du tireur. Les joueurs reçoivent un xAG uniquement quand un tir est effectué après une passe clé. Ici, on parle donc de la moyenne de xAG de l'équipe par match. Cet indicateur est très intéressant car il permet de révéler la capacité d'une équipe à se créer des occasions de buts par la passe.

- (xG + xAG) /90min = Il s'agit de la moyenne des xG + xAG par match. Nous utilisons xG + xAG pour la contribution aux buts des joueurs puisqu'elle est généralement égale à la somme des buts et des passes décisives.

- npxG_par_90min = Moyenne des xG sans les pénaltys par matchs (un pénalty rapporte toujours 0.79 xG, ce qui peut nous tromper sur la réelle prestation d'une équipe).

- npxG_plus_xAG_par_90min = Il s'agit de la moyenne des npxG + xAG par match.

- Buts_encaissés_par_90min = Moyenne de buts encaissés par match.

- Arrêts_pourcentage = Le pourcentage d'arrêt de l'équipe sur la saison ((Tirs cadrés reçus – Buts encaissés) / tirs cadrés reçus). On ne compte pas seulement les arrêts du gardien mais aussi les contres des défenseurs.

- CleanSheets_pourcentage = Le pourcentage de match sans encaisser de buts.

- PenAgainst = Les penaltys subis par l'équipe au total sur la saison. On ne regarde pas si le penalty a été inscrit ou non.

- G-xG = Buts marqués – Expected Goals. Il s'agit de la différence entre le nombre de buts marqués et les xG, ce qui pourra nous donner une indication sur l'efficacité ou la réussite offensive de l'équipe.

- GA-xGA = Buts encaissés – Expected Goals Against. Il s'agit de la différence entre le nombre de buts encaissés et les xGA, ce qui pourra nous donner une indication sur l'efficacité ou la réussite défensive de l'équipe.

- G_90min = Moyenne de buts marqués par match.

- GA_90min = Moyenne de buts encaissés par match.

- Diff_90min = Moyenne de buts marqué par match. – Moyenne de buts encaissés par match.

L’analyse déductive

Après avoir introduit le sujet et décrit entièrement notre base de données, nous allons procéder à l’analyse descriptive de nos données. L’objectif est de regrouper dans un premier temps quelques statistiques univariées dans un tableau afin d’observer les résultats, puis d’analyser des graphiques entre certaines variables que non jugeons intéressantes à mettre en corrélation. On émet l’hypothèse que certaines statistiques augmentent les chances d’être mieux placé au classement. En effet, des variables telles que la possession ou encore les expected goals (xG), qui sont des données appréciées et recherchées par les entraineurs modernes, augmentent les chances d’être mieux classés lorsqu’elles sont élevées. De plus, des variables comme l’affluence dans les stades ou le nombre de buts du meilleur buteur de l’équipe sont des variables qui semblent aider également une équipe à être en haut du classement si elles sont élevées. Enfin, certaines variables servent à mesurer la chance lorsqu’on les soustrait. Il s’agit de la différence de buts – la différence des xG (xGD). Plus cette variable très importante est élevée et plus l’équipe a de chances d’être en haut du classement. Ainsi, nous allons examiner certaines de ces variables sous forme de graphique pour observer leurs utilités, tout en regardant les autres que nous soupçonnons moins d’impacter le classement. Pour finir, nous ferons une régression linéaire pour confirmer ou contredire nos résultats.

Tableau des statistiques descriptives

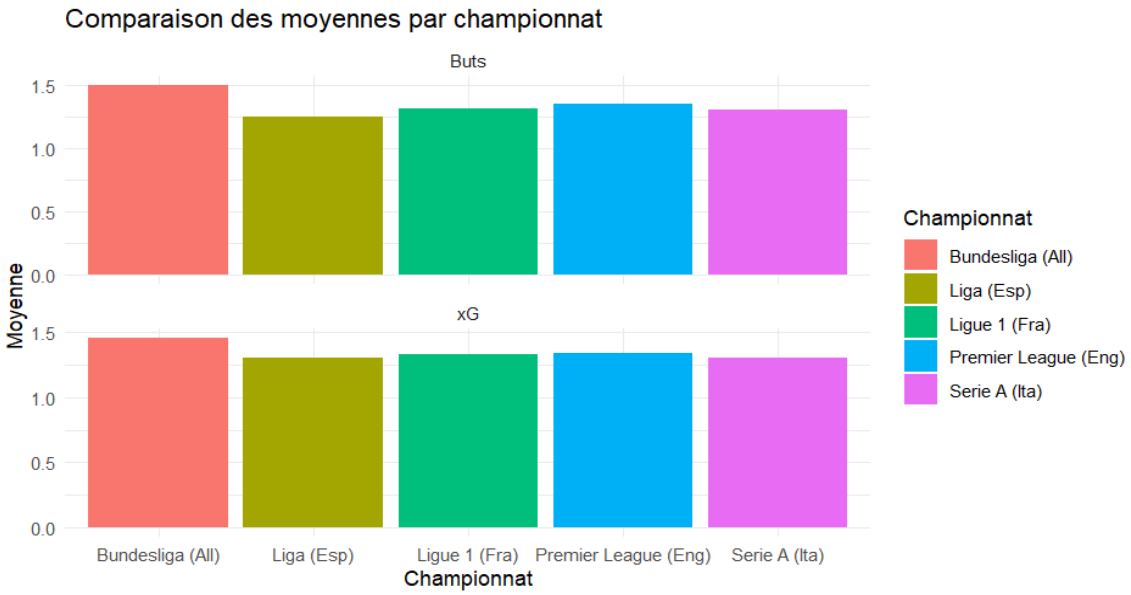
	Minimum	1er Quartile	Médiane	Moyenne	3eme Quartile	Maximum
Classement	1	5	10	10.32	15	20
Pts_matchs	0.42	1.05	1.3	1.37	1.66	2.63
xGD	-1.13	-0.38	-0.07	0	0.32	1.72
Affluence (en milliers)	4.9	15.55	25.2	29.92	41.44	83.5
Nb buts meilleur buteur	4	9	12	13	16	36
Joueurs_champs	21	26	28	28.46	30	42
Âge	22.6	25.88	26.6	26.57	27.4	30.3
Possession	35.4	45.2	49.25	50	54.02	71
Pen_tentés	0	4	6	5.82	8	14
Buts	0.55	1	1.26	1.34	1.59	2.71
PD	0.32	0.68	0.87	0.93	1.12	2.24
B+PD	0.89	1.68	2.13	2.27	2.71	4.95
B-PénM	0.42	0.92	1.16	1.22	1.45	2.56
B+PD-PénM	0.74	1.60	2	2.15	2.58	4.79
xG	0.76	1.09	1.29	1.34	1.51	2.62
xAG	0.51	0.76	0.9	0.95	1.09	1.96
xG+xAG	1.28	1.86	2.18	2.3	2.61	4.58
npxG	0.67	0.99	1.16	1.22	1.37	2.5
npxG+xAG	1.18	1.75	2.07	2.18	2.47	4.46

Arrêts%	52.7	67.4	70.7	70.65	74	83.7
CleanSheets%	2.6	20.6	26.3	27.05	34.2	68.4
PenAgainst	0	4	6	5.81	7	14
xGA	0.67	1.15	1.33	1.34	1.54	2.05
G	0.58	1.05	1.32	1.38	1.64	2.85
GA	0.53	1.13	1.37	1.38	1.6	2.41
Diff	-1.6	-0.47	-0.09	0	0.42	2.08
PossProg	9.6	14.82	17.59	17.86	20.27	34.87
PassesProg	22.61	32.84	37.41	38.65	42.7	68.62
Cartons_jaunes	1	1.76	2.03	2.07	2.37	3.53
Cartons_Rouges	0	0.05	0.09	0.1	0.13	0.39
G-xG	-17.3	-4.12	0.8	1.41	5.92	30.7
GA-xGA	-15	-3.4	1.8	1.41	5.5	24.2

Ce tableau nous permet de mieux visionner la dispersion de nos données et l'écart entre les meilleures équipes et les plus faibles. Prenons par exemple la possession qui révèle ces écarts conséquents avec un minimum de 35,4% de temps de possession en moyenne sur une saison entière pour l'équipe la plus faible, contre 71% de possession pour une des équipes les plus forte. La moyenne n'est pas très intéressante ici puisqu'elle s'équilibre, c'est-à-dire qu'elle sera toujours égale à 50%. Certaines variables attirent notre curiosité comme la moyenne d'âge des équipes ou bien le nombre de joueurs utilisés en moyenne sur une saison. Certaines équipes sont très jeunes avec un minimum de 22,6 ans et d'autres sont plus expérimentées avec un maximum de 30,3 ans. Doit-on privilégier l'expérience à la jeunesse ou la jeunesse à l'expérience ? Ou bien trouver un juste équilibre qui se rapproche de la moyenne de 26,57 ans ? Doit-on utiliser un effectif très large comme le maximum de 42 joueurs de champs différents ou bien très réduit comme le minimum de 21 joueurs différents ? Nous tenterons de répondre à ces questions à travers notre analyse.

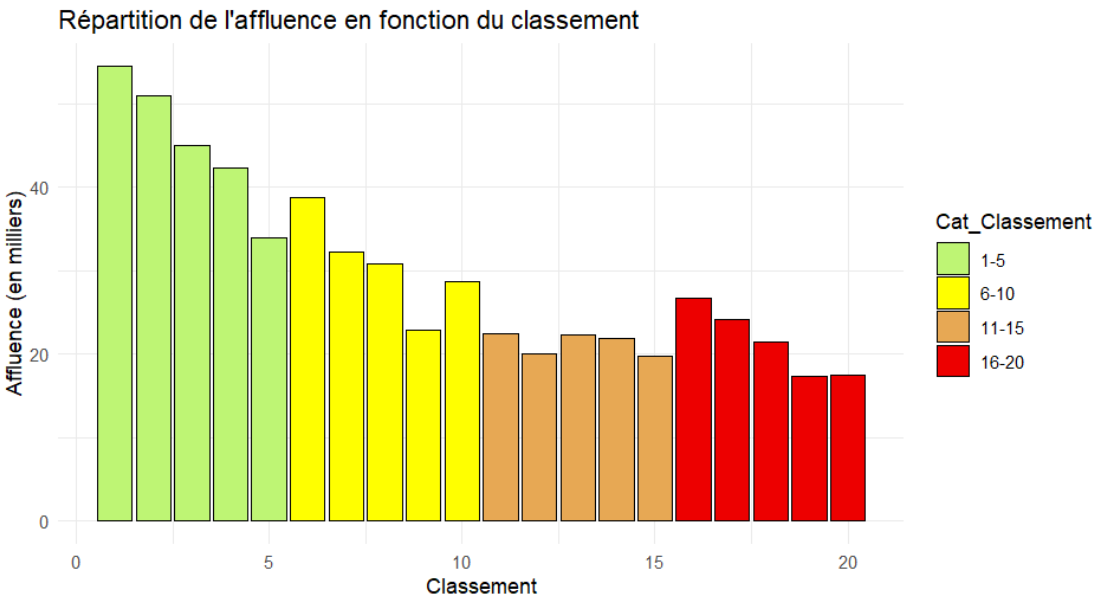
Les graphiques

Graphique 1 : Comparaison des moyennes de buts et d'xG par championnat



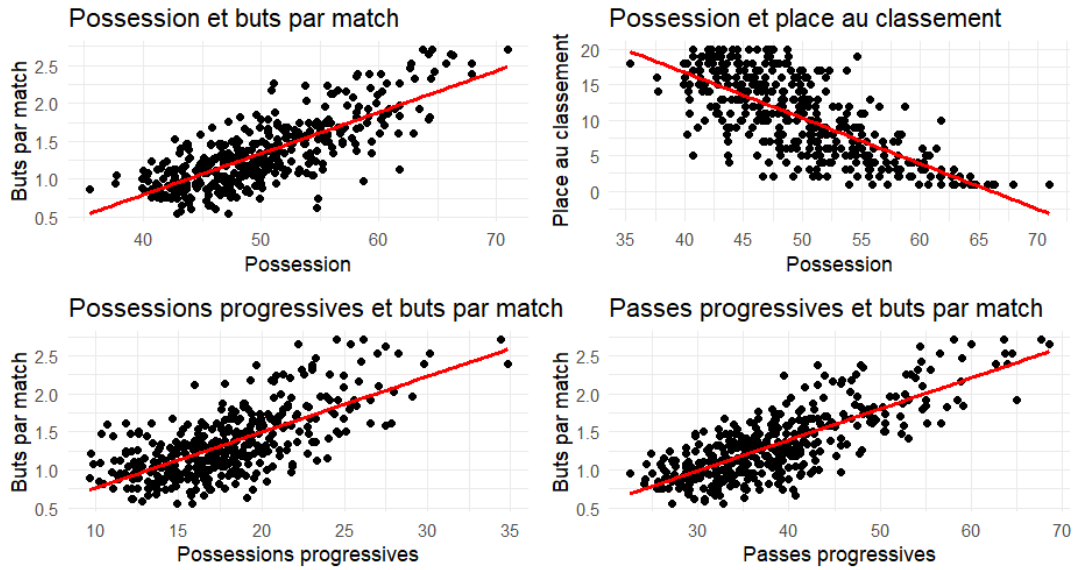
Ce premier graphique nous montre une comparaison intéressante entre les 5 grands championnats européens. Il s'agit du nombre de buts marqués par match et du nombre d'xG par match. On constate qu'en terme de buts marqués, la Bundesliga est loin devant avec 1,5 buts par match en moyenne. Le championnat espagnol est dernier avec 1,25 buts marqués par match. Mais ces résultats représentent-ils la réelle performance des équipes qui jouent dans ces championnats ? C'est ce que nous pouvons vérifier avec le graphique du dessous qui nous montre les xG. En effet, nous pouvons voir que l'Allemagne est toujours devant mais l'écart avec les autres championnats est bien moins conséquent. De plus, la Liga rattrape son retard sur les autres championnats. Les trois autres championnats semblent ne pas avoir beaucoup d'écart entre leur nombre de buts marqués et leur nombre d'xG. On peut donc en déduire que le championnat allemand est en surrégime, c'est-à-dire qu'il marque plus de buts qu'il ne devrait. Cela peut être dû à la chance ou bien a une grande efficacité des attaquants qui arrivent à marquer des buts qualifiés de « difficiles ». On en déduit tout l'inverse pour le championnat espagnol, tandis que les trois autres semblent respecter leurs standards.

Graphique 2 : Evaluation de l'affluence en fonction du classement



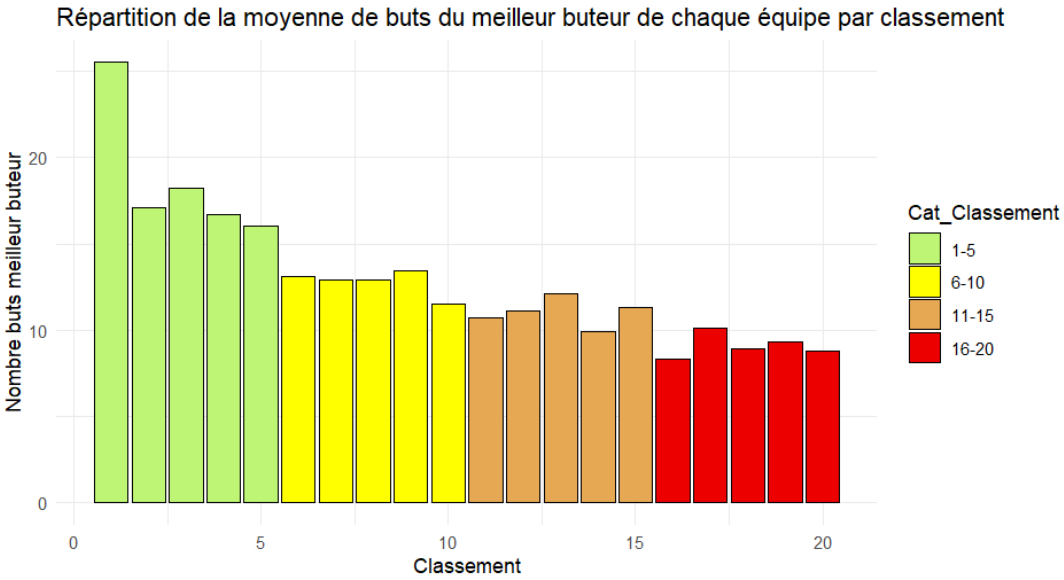
Ce second graphique représente l'affluence des matchs à domicile en fonction du classement général. Sans surprise, les premiers du classement ont une affluence bien plus élevée que les derniers. Cependant, il est important de prendre du recul sur l'analyse de ce graphique car de nombreux facteurs peuvent expliquer ça. On peut se dire que plus l'équipe est forte et plus elle attire du monde au stade, ou bien que le monde au stade suscite une grosse ambiance qui galvanise les joueurs et les rend plus fort. Mais si on regarde encore plus loin, on sait que généralement les plus grands clubs évoluent dans les plus grandes villes, et donc ont un stade plus grand qui peut amener plus de monde. De plus, les plus grands clubs ont également un budget bien plus conséquent que les plus petits, ce qui peut aussi expliquer ces répartitions d'affluence. Autrement dit, un grand club dans une grande ville a un plus grand stade pour jouer et un plus grand budget que les petits clubs des plus petites villes. (Exemple : Paris vs Brest). Il aurait sûrement été plus intéressant d'avoir le taux de remplissage des stades en variable. Nous pouvons quand même remarquer que le graphique n'est pas continuellement décroissant. En effet, les équipes classées 16eme ont étonnamment une meilleure affluence en moyenne que les équipes classées 9eme. Cela montre que certaines équipes qui attirent moins de gens au stade, ou bien qui ont simplement un stade plus petit, jouent mieux que certaines équipes qui ont une meilleure affluence.

Graphique 3 : Evaluation de l'impact de la possession sur le classement et sur le nombre de buts marqués



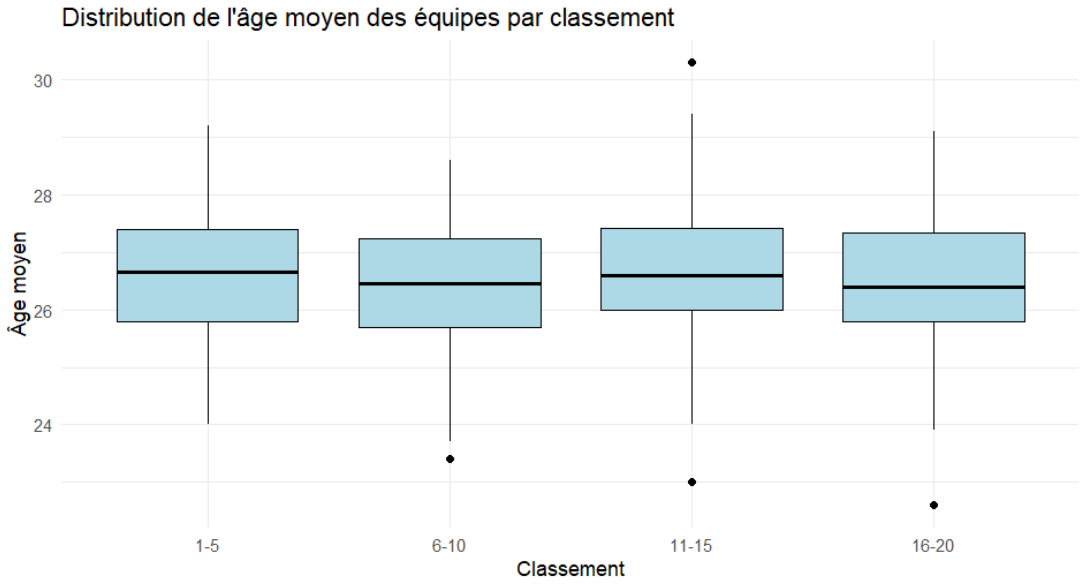
Ce troisième visuel est composé de 4 graphiques en nuage de points. Il s'agit de 4 variables assez similaires axées sur la possession. Les plus grands entraineurs du football comme Josep Guardiola par exemple travaillent beaucoup sur le jeu de possession et cherchent à maximiser leur temps de possession du ballon pendant les matchs. Ainsi, nous souhaitions vérifier que cette dernière à bel et bien un impact sur les résultats. En se référant à ces graphiques, la réponse est oui, sans aucun doute. La possession est très corrélée avec le fait de marquer des buts, et le fait d'être bien en haut du classement. De plus, les possessions progressives et les passes progressives, qui sont des statistiques plus sophistiquées, montrent également une forte corrélation avec le nombre de buts marqués par match. Ainsi, maximiser la possession du ballon semble être un véritable atout pour remporter un match de football, et peut-être un championnat.

Graphique 4 : Répartition de la moyenne de buts du meilleur buteur de chaque équipe par classement



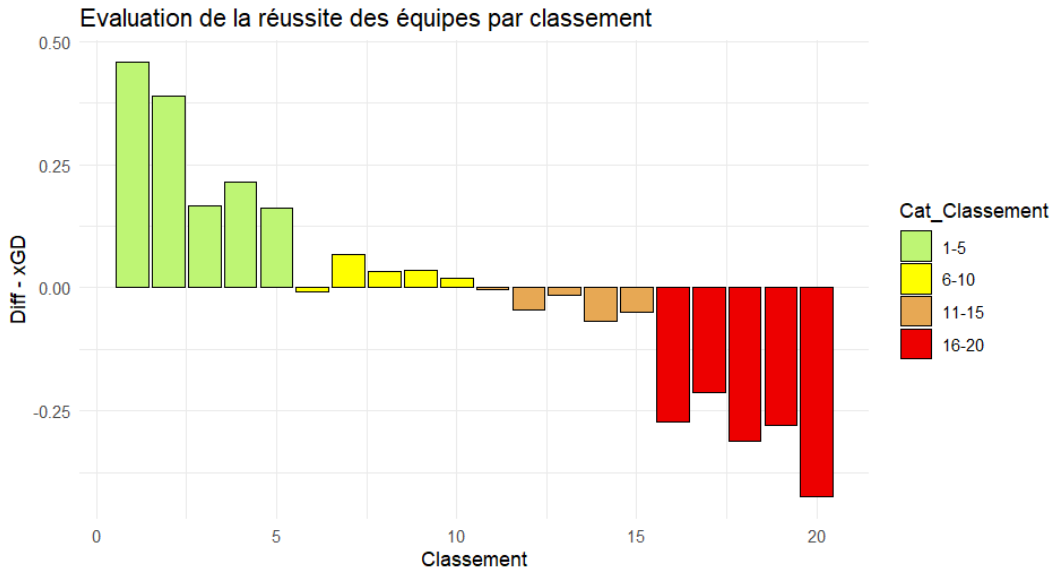
Ce quatrième graphique montre la répartition de la moyenne de buts du meilleur buteur de chaque équipe par classement. Nous savons que le football est un sport collectif avant tout, mais nous cherchons à évaluer l'importance d'avoir une forte individualité dans une équipe. Nous prenons pour ça le meilleur buteur par équipe sur une saison entière. La première chose qui saute aux yeux est la moyenne de buts des meilleurs buteurs des vainqueurs de championnat qui est considérablement plus élevée que celles des deuxièmes. Là est peut-être le détail important qui fait gagner un championnat. Le graphique semble nous dire qu'avoir un buteur extrêmement prolifique au sein de son collectif est un atout majeur. De plus, on constate que la moyenne pour les équipes en bas de classement est plus faible, ce qui confirme notre théorie. Cependant, ce n'est pas toujours vrai. L'exemple de Manchester City en 2020-2021 (hors base de données) montre qu'il est possible d'être champion sans buteur prolifique, avec un fort collectif. En effet, le meilleur buteur de l'équipe cette saison-là ne comptait que 13 buts. Si on le compare à notre graphique, Manchester City se situerait entre la 5ème et la 9ème place.

Graphique 5 : Distribution de l'âge moyen des équipes par classement



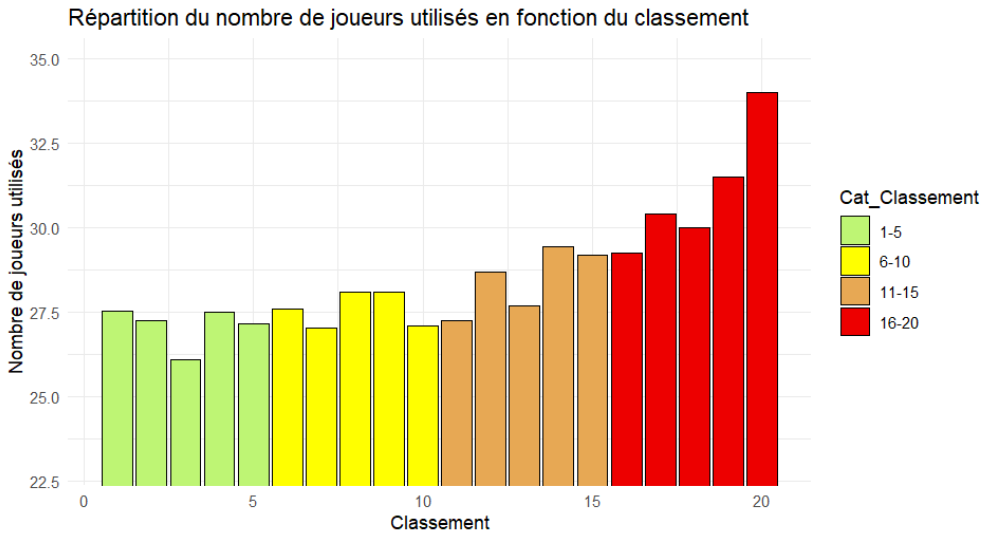
Ce cinquième graphique portant sur la distribution de l'âge moyen des équipes par classement nous offre une première réponse à la question que nous nous posions lors de l'analyse du tableau des statistiques descriptives. Dans un premier temps, nous comprenons que les équipes classées de 1 à 5 ont une moyenne d'âge comprise entre 24 et 29 ans. Il y a autant d'équipe dans cette catégorie de classement qui ont une moyenne entre 24 et 27 ans et d'équipes entre 27 et 29 ans. Ainsi cela montre l'importance d'avoir des joueurs d'expérience au sein de son effectif. Il ne faut donc pas une équipe trop jeune. La catégorie des 16-20 possède la plus faible médiane, ce qui prouve qu'une équipe trop jeune comporte des risques vis à vis de la performance à haut niveau. Il faut donc trouver un juste équilibre afin de performer et grimper au classement.

Graphique 6 : Evaluation de la réussite des équipes par classement



Ce sixième graphique nous parle d'une statistique très moderne et d'un concept assez difficile à évaluer qui est celui de la chance. En effet, la mesure consiste à prendre la différence de buts des équipes (buts marqués – buts encaissés) et la différence des xG des équipes (xG - xG against). Ensuite, la soustraction de ces deux valeurs nous donne une estimation de la chance d'une équipe au cours de la saison. Cette « chance » peut être définie comme le fait d'avoir marqué des buts qu'on n'aurait pas forcément « dû » marquer ou bien d'avoir loupé un but qu'on aurait « dû » marquer, soit grâce à un joueur fort ou moins fort individuellement, soit un but contre son camp de l'équipe adverse, ou bien un rebond chanceux ou malchanceux etc. Cette « chance » prend aussi en compte le fait d'avoir encaissé un but qu'on aurait dû éviter ou d'avoir évité un but qu'on aurait dû encaisser. Ainsi, le graphique est très clair. Pour être premier, il faut bénéficier de ce facteur « chance » ou bien avoir des joueurs capables de faire la différence dans ces moments importants. A l'inverse, les derniers du classement manquent cruellement de chance ou de joueurs capable de faire cette précieuse différence.

Graphique 7 : Répartition du nombre de joueurs utilisés en fonction du classement



Ce dernier graphique éclairci les points que nous avons énoncés lors du tableau des statistiques descriptives vis-à-vis du nombre de joueurs de champs différents utilisés lors d'une saison. Nous pouvons d'ores et déjà y répondre : Avoir un effectif trop large est un handicap comme le montre la catégorie des 16-20 qui possède une moyenne plus élevée que les autres. Mais attention, cela peut aussi s'expliquer par un trop grand nombre de blessés lors d'une saison. En effet, si l'effectif est aussi large, on peut imaginer que l'équipe a connu un nombre important de blessure et a donc dû essayer de nouveaux joueurs, logiquement un peu moins bons que les titulaires blessés. A l'inverse, les premiers du classement ont moins de joueurs différents utilisés en moyenne sur une saison. Ils ont peut-être moins été embêté par les blessures, ou bien se fient simplement à la qualité de leur groupe réduit mais soudé et fort collectivement.

Analyse inductive

Les régressions simples

	estimate	std.error	statistic	p.value	R^2
(Intercept)	10,32	0,15	67,99	<2e-16	
xGD	-9,18	0,29	-31,79	<2e-16	0,7215
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	15,60	0,47	32,96	<2e-16	
Affluence	-0,18	0,01	-12,96	<2e-16	0,3011
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	17,68	0,56	31,68	<2e-16	
Nombre_de_buts_du_meilleur_buteur	-0,57	0,04	-14,50	<2e-16	35,04
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	-6,50	2,07	-3,14	0.00183	
Joueurs_champs	0,59	0,07	8,19	3.79e-15	0,1467
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	9,58	6,41	1,50	0.136	
Âge	0,03	0,24	0,12	0.909	3.371e-05
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	42,47	1,59	26,77	<2e-16	
log(Possession)	-0,64	0,03	-20,43	<2e-16	0,517
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	14,02	0,67	21,02	<2e-16	
Pen_tentés	-0,64	0,10	-6,10	2.6e-09	0,08708
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	25,64	0,71	35,95	<2e-16	
xAG	-16,05	0,72	-22,30	<2e-16	0,5605
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	26,73	0,76	35,24	<2e-16	
npxG	-13,42	0,60	-22,35	<2e-16	0,5616

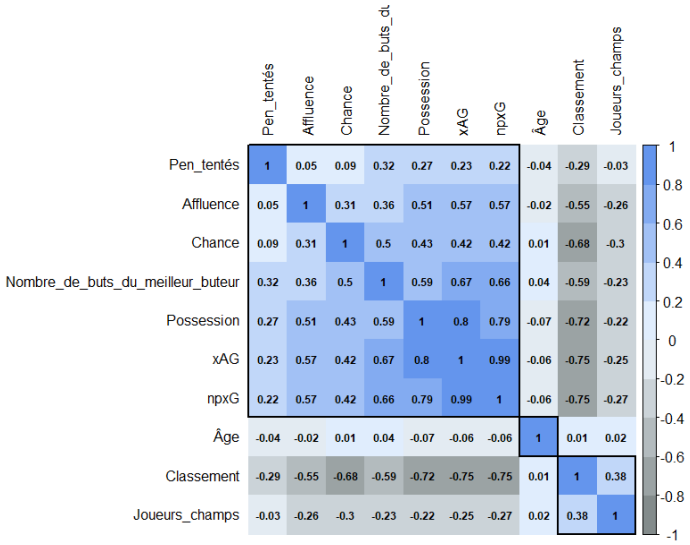
	estimate	std.error	statistic	p.value	R^2
(Intercept)	44,89	3,89	11,56	<2e-16	
log(`Arrêts%`)	-0,49	0,05	-8,92	<2e-16	0,1697
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	5,94	0,68	8,74	<2e-16	
PenAgainst	0,76	0,11	7,04	8.71e-12	0,113
	estimate	**std.error**	**statistic**	**p.value**	**R^2**
(Intercept)	10,32	0,21	48,82	<2e-16	
Chance	-12,94	0,71	-18,21	<2e-16	0,4596

Avant de nous attaquer aux régressions multiples, jetons un œil sur l'impact de nos variables sur le classement lorsqu'elles sont isolées des autres. En effet, le tableau ci-dessus montre le résultat des régressions simples de nos variables explicatives avec le classement comme variable à expliquer. Nous avons ajouté la variable « Chance » présente et expliquée lors du graphique 6. Nous pouvons d'abord remarquer qu'elles sont toutes très significatives sauf la variable de la moyenne d'âge. Notons que la variable xAG semble être celle qui a le plus gros impact avec un coefficient de -16,05 unités. Il sera intéressant de voir si ce fort coefficient est aussi présent lors de la régression multiple.

La régression multiple

Après avoir démontré la corrélation entre certaines variables lors de notre analyse descriptive à travers plusieurs graphiques, puis réaliser des régressions simples pour regarder l'impact de nos variables individuellement, nous allons maintenant procéder à une régression multiple pour tenter de confirmer nos hypothèses, ou bien les contredire. Nous avons à disposition de nombreuses variables indépendantes mais malheureusement, certaines variables sont fortement corrélées entre elles. Nous avons donc choisi de retirer les variables qui se rapprochaient des autres ou bien qui étaient trop corrélées avec une autre. Par exemple, nous avons gardé seulement une variable sur la possession, nous avons retiré les xG et le nombre de buts marqués qui sont des variables fortement corrélées a notre variable dépendante, mais aussi à d'autres variables indépendantes. Avant la régression, jetons un œil à la heatmap des corrélations de nos variables finales.

Heatmap des corrélations



On remarque que les variables sont globalement corrélées positivement entre elles, dont certaines qui affichent une corrélation très élevée comme les xAG avec les npxG. Les variables qui affichent une corrélation négative avec la variable classement sont celles qui ont en réalité un impact positif avec cette dernière, l'objectif étant d'atteindre le plus petit chiffre (première position).

Voici donc notre régression finale composée d'une constante et de 12 variables dépendantes.

	estimate	std.error	statistic	p.value
(Intercept)	18.36	11.83	1.55	0.12
xGD	-7.47	0.59	-12.66	0.00
Affluence	-0.03	0.01	-3.45	0.00
Nombre_de_buts_du_meilleur_buteur	0.07	0.03	2.52	0.01
Joueurs_champs	0.06	0.03	1.91	0.06
Âge	-0.03	0.09	-0.27	0.79
log(Possession)	-3.14	1.58	-1.99	0.05
Pen_tentés	-0.05	0.05	-0.97	0.33
xAG	-0.18	2.70	-0.07	0.95
npxG	1.92	2.25	0.86	0.39
log(`Arrêts%`)	0.20	2.05	0.10	0.92
PenAgainst	0.04	0.05	0.80	0.43
Chance	-7.35	0.53	-13.79	0.00

Nous obtenons dans un premier temps un R^2 ajusté égal à 85,02%, ce qui signifie que notre modèle est expliqué à 85,02% par les variables que nous avons choisi. Nous constatons dans un second temps que 5 variables sur les 12 sont significatives. Il s'agit des xGD, de l'affluence (Aff), du nombre de buts du meilleur buteur (MB), de la possession (Poss), et de la chance. Ainsi, nous obtenons l'équation suivante avec c_i = « Gagner un championnat »

$$c_i = 18.36\beta_1 - 7.47\beta_2 xGD - 0.03\beta_3 \text{Aff} + 0.07\beta_4 MB - 3.14\beta_5 Poss - 7.35\beta_6 Chance + \varepsilon_i$$

Interprétation des coefficients

Il est important de préciser avant tout que dans notre étude, un coefficient négatif équivaut à un coefficient positif et inversement puisque notre Y est décroissant. En effet, l'objectif est d'atteindre le plus petit chiffre du classement, soit 1. Ainsi, procédons à l'interprétation de nos coefficients significatifs.

Les xGD : Toute chose égale par ailleurs, une augmentation des xGD par match de 1 unité entraîne un gain de 7.47 places au classement. Ce fort coefficient s'explique par le fait que la variable est donnée par match, et une augmentation d'une unité par match est une augmentation conséquente qui permet de faire un bond au classement.

L'affluence : Si on ajoute 1000 personnes au stade (l'affluence étant exprimé en milliers, une unité = 1000 personnes), l'équipe gagnera 0.03 places au classement. Ainsi, le fait de gagner un championnat ne dépend pas énormément de l'affluence au stade.

Le meilleur buteur : Chaque but supplémentaire inscrit par le meilleur buteur de l'équipe nous fera perdre 0.07 places au classement. Ce coefficient est surprenant et voudrait dire qu'avoir un buteur prolifique n'est pas positif, et est même très légèrement un désavantage. Ainsi, ce résultat contredit notre hypothèse émise lors du graphique 4 et tend même à rejoindre l'exemple évoqué sur la saison 2020-2021 de Manchester City.

La possession : Ensuite, si on hausse notre possession de 1%, alors on gagnera au classement 3.14 places. A l'inverse du résultat précédent, ce résultat confirme notre hypothèse et nous conforte dans l'idée que la possession est un facteur important dans la victoire d'un championnat.

La chance : Enfin, notre variable préférée, la chance. Si on « augmente la chance » de 1 unité par match, on gagnera 7.35 places au classement. Une nouvelle fois, la variable étant exprimée par match, 1 unité est une hausse conséquente, ce qui explique ce fort coefficient.

Conclusion

Pour conclure, cette analyse nous a permis de mieux comprendre quels facteurs permettaient ou non de maximiser la performance dans les 5 grands championnats européens. Plusieurs observations et tendances significatives ont émergées lorsque nous avons tenté de répondre à notre problématique : Comment remporter un championnat de football ? Du point de vue des analyses descriptives, nous avons identifié des facteurs clés qui semblent être associés à la réussite d'une équipe. Le temps de possession du ballon apparaît comme un élément crucial pour aspirer à une place en haut du classement. De plus, le facteur « chance » dans les matchs et dans une saison entière n'est pas à négliger et est absolument nécessaire pour gagner un championnat. Il n'est cependant pas forcément contrôlable par les équipes, mais ces dernières peuvent toujours essayer de provoquer leur destin. Sur le plan économétrique, nos modèles ont confirmé l'importance de la possession du ballon et du facteur chance comme on pouvait s'en douter, mais aussi de l'importance des xGD qui sonnait comme une évidence. Son coefficient élevé révèle que la priorité n'est pas de marquer par n'importe quel moyen, mais bel et bien de se procurer un maximum d'occasions tout en maintenant une solidité défensive. Les buts viendront d'eux-mêmes si l'équipe travaille sur ce point avant tout. Enfin, des variables telles que l'affluence ou le nombre de buts du meilleur buteur ont montrés des liens significatifs avec le succès dans le football. Cependant, il est essentiel de souligner que le football est un sport complexe et dynamique, avec de nombreux facteurs imprévisibles. Bien que nos analyses fournissent des éléments intéressants, la victoire dans un championnat ne peut être réduite à une formule simple. Des éléments tels que la stratégie de jeu, l'entente entre les joueurs et l'entraîneur ou encore les circonstances particulières de chaque saison jouent également un rôle crucial. En conclusion, les équipes aspirant à remporter un championnat de football devraient se concentrer sur la stabilité de leur effectif offensivement comme défensivement, la performance constante des joueurs, et la recherche d'une stratégie permettant de maximiser le nombre d'occasions créées. Cependant, elles doivent également rester flexibles et adaptatives pour faire face aux défis imprévus qui sont inhérents au monde du football. Cette étude offre une base solide pour la prise de décision stratégique dans le contexte du football professionnel, mais il est important de rester conscient des nuances propres à ce sport passionnant et souvent imprévisible.