Rapport Data Mining

POITEVIN Louis
Numéro etudiant : 11410541
Thématique : European Soccer Leagues

0) Sommaire

I) Introduction

Le but de ce projet est de se familiariser avec le travail de nettoyage, d'exploration, d'analyse et de visualisation de données. Pour cela, nous devons choisir un set de données, définir des grands axes et des questions liées à ces axes. La tâche consiste donc à nettoyer, analyser, explorer et structurer les données afin de pouvoir répondre à ces questions et d'en afficher les réponses sous forme de graphiques explicatifs.

Étant amateur de football, j'ai choisi d'utiliser la database "European Soccer League". Ce thème populaire a nourri beaucoup de conversations avec d'autres passionnés. J'ai donc été curieux de voir si ces débats éternels sur le football pouvaient être concrètement traités à l'aide des données, en répondant de manière un peu plus objective et documentée aux questions engendrées.

Ce dataset offre une riche variété d'axes, de questions et d'analyses possibles. J'ai choisi de traiter les thèmes suivants:

1. Les choix d'entraîneur pour la performance
2. Le jeu à domicile VS à l'extérieur

Bien sûr, ces thèmes sont eux-mêmes sources d'un nombre incalculable d'analyses possibles.

Ces sujets sont souvent au centre des discussions footballistiques, et de plus en plus de spectateurs sont devant leurs TV pour écouter l'avis des "footballogues". Ces spécialistes du football tentent de répondre à des questions théoriques à l'aide d'un immense savoir et d'une certaine légitimité professionnelle. Il sera donc intéressant de confronter leurs points de vues avec la réalité des données. De plus, il est évident que le football a un fort pouvoir financier et culturel, et que les réponses apportées dans ce domaine peuvent se révéler d'une grande importance.

Dans un premier temps, nous présenterons la database utilisée avec ces différentes tables de données. Une fois familiarisés avec les données, nous présenterons les analyses effectuées de ces grands thèmes ainsi que leurs résultats.

II) Présentation des données

A) Origine des données

L'extraction des données est souvent une tâche fastidieuse, qui prend une part conséquente du temps investi dans un projet de Data Science.En effet, il faut souvent réfléchir aux données utiles ainsi qu'à leur adresse d'extraction et les obtenir sous la structure désirée. Ce ne fut pas le cas concernant ce projet, puisque les données ont été concentrées dans une Database dont le descriptif quasiment complet est disponible sur Kaggle. Kaggle est à l'origine une plateforme web organisant des compétitions en science des données et offre par ce biais beaucoup d'analyses faites sur un certain nombre de données.

Les données proviennent à la base des :

- http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events
- http://www.football-data.co.uk/ : betting odds. Click here to understand the column naming system for betting odds:
- http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports.*

B) Forme des données

Les données sur le football viennent d'une database formée de de 7 tables. Avant de rentrer dans les détails des tables, ces tables réunissent des données de 2008 à 2016 et contiennent globalement les informations suivantes:

- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Detailed match events (goal types, possession, corner, cross, fouls, cards etc…) for +10,000 matches

On voit alors que si on ne désire pas faire d'analyses rétrospectives ou sur l'histoire du football, ces données sont amplement suffisantes pour le cadre de ce projet, on n'en cherchera pas d'autre.

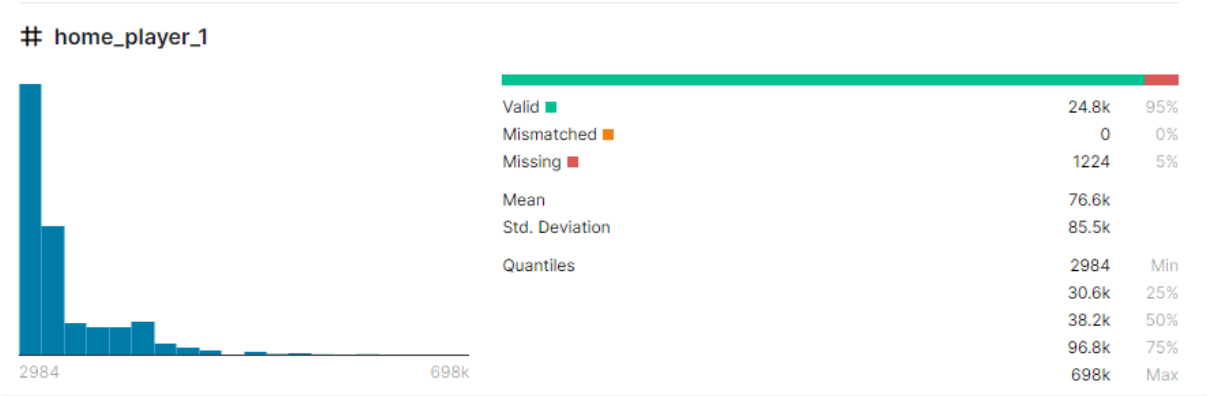
1) Table Match

Cette table est sans doute la plus importante des données. Elle répertorie tous les matchs (26k) effectués dans deux nombreuses ligues européennes. Elle est également connectée avec toutes les tables. Elle sera utilisée dans toutes nos analyses.

Parmis les attributs les plus importants, on y trouve:

- Les équipes qui s'affrontent (domicile et extérieur)
- les joueurs et leurs positions lors du match
- les faits de jeu (fautes, corners…)
- les buts marqués
- les paris

Parmi les attributs concernant les joueurs et les faits de jeu, on observe des NaN. Il faudra prendre cela en considération.



2) Tables player et player_attribut

Ces tables décrivent respectivement leurs caractéristiques objectives (date de naissance, taille, poids) et leurs compétences techniques (Dribble ,Longpassing,, Volley...). On l'utilisera lorsqu'il s'agira de discuter les choix possibles d'entraîneurs concernant les joueurs.

On observe quelques NaN également quelques NaN dissimulés dans les players attributs.

3) Tables teams et teams_attribut

La table “teams” répertorie simplement le nom des équipes, et “teams attributs” leurs compétences (defense_pressure, passing…). On l’utilisera également pour discuter des choix des entraîneurs.

A priori, il n’y a pas de NaN et la table Team donne accès à 299 équipes. Cependant, la table teams_attribut ne recense que 288 équipes et contient des NaN dans la section “BuildUpPlayDribbling”. Ce dernier problème peut être facilement résolu: après analyse, nous nous rendons compte que tous ces NaN correspondent a peu de dribbling skills (voir “BuildUpPlayDribbling”). On voit également qu’il y a peu de variance dans les données pour les teams avec peu de dribbles. On peut donc les remplacer sans crainte.

```
In [24]: teams_att[teams_att['buildUpPlayDribbling'].isna()]['buildUpPlayDribblingClass'].value_counts()
Out[24]: Little    969
         Name: buildUpPlayDribblingClass, dtype: int64
```

-BuildUpPlayDribbling n'a que 489 valeurs. Mais "Classe" a toutes ses valeurs. En regardant, les valeurs NaN des BuildUpPlayBuidling correspondent toutes a la classe "Little", ce qui fait sens.

```
In [25]: teams_att[teams_att['buildUpPlayDribblingClass']=='Little']['buildUpPlayDribbling'].value_counts()
Out[25]: 32.0    12
         33.0     6
         31.0     4
         29.0     4
         28.0     3
         24.0     2
         30.0     2
         27.0     1
         26.0     1
         Name: buildUpPlayDribbling, dtype: int64
```

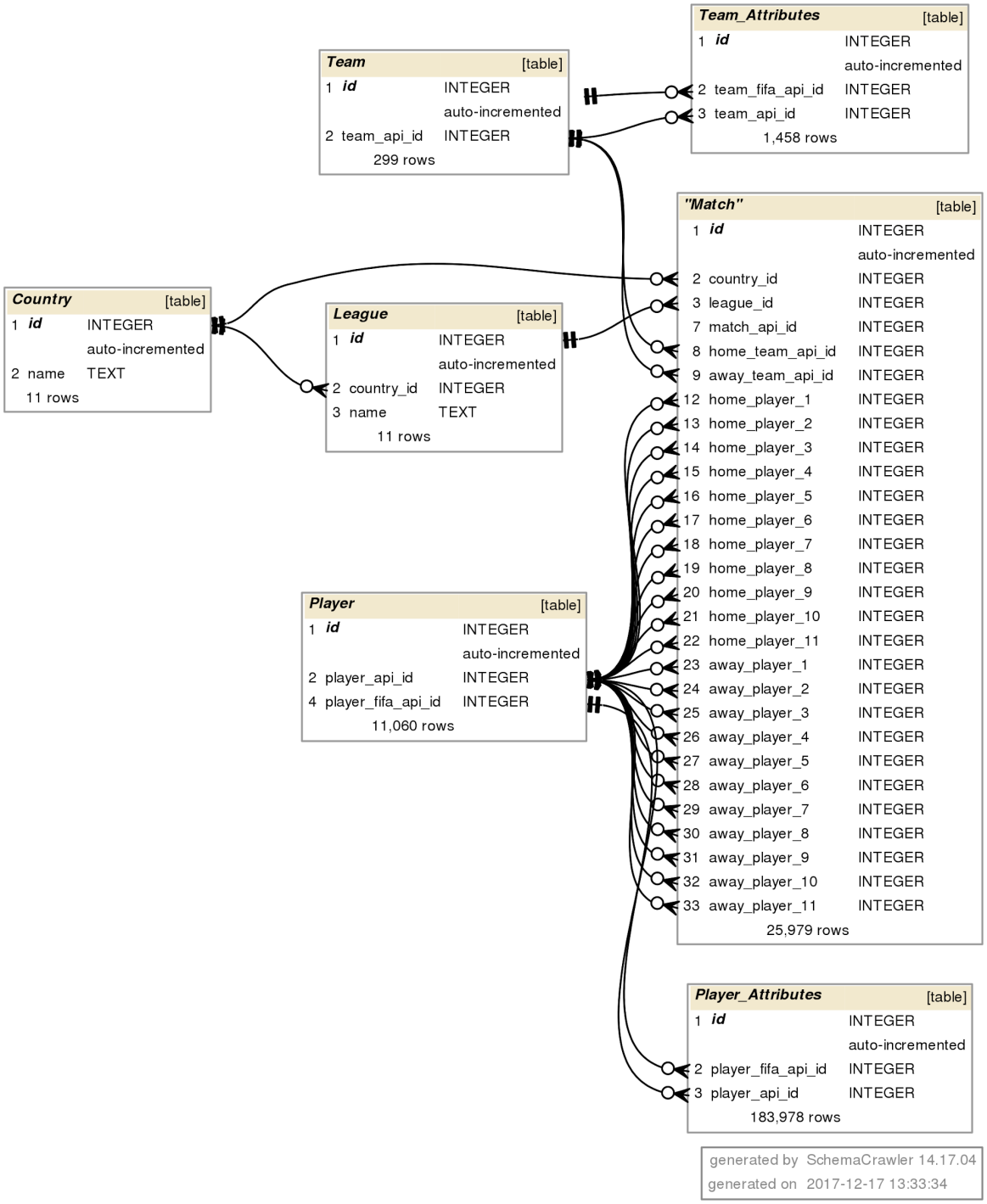
en regardant ces valeurs, on voit que toutes valeurs correspondantes a la classe "Little" sont autour de 32 avec tres peu de variance. On remplace les NaN par la valeur 32.

Enfin, on observe un seul duplicata dans team_attributs que l’on peut enlever.

4) Table leagues/country

La table Leagues liste les différentes ligues présentes dans la table “match”. De manière analogue, la table country donne les pays associés. Ces tables recensent 11 ligues/pays.

On peut résumer ces tables dans un ER Diagramme:



III) Notebook 1: Quelles choix stratégiques devrait privilégier un entraîneur ?

A) Objectif du Notebook

L'entraîneur a le rôle du cerveau au sein d'une équipe. Avant chaque match il est notamment chargé des fonctions suivantes:

1. définir l'équipe titulaire (constituer une équipe harmonisée et cohérente)
2. attribuer des postes au joueurs (défenseur latéral, milieu gauche,...)
3. définir une formation d'équipe (choix de l'organisation de l'équipe sur le terrain)

Dans ce notebook, nous nous intéressons à certains aspects de ces trois thèmes très complexes.

Il inclut les grands axes suivants:

1) **Choix d'une métrique** pour évaluer la performance des équipes par saison: nombre de points moyen gagnés par match. Ceci constitue le fil rouge de notre analyse: les attributs tactiques des équipes "performantes" seront considérés comme des bons choix d'entraîneur.

2) **Quels attributs d'équipe un entraîneur devrait-il privilégier ?** Est-ce qu'un entraîneur devrait privilégier une défense agressive ? Une attaque pleine de dribbleurs ? un jeu de construction collectif ? Ceci est une question issue du thème 1), c'est-à-dire une étude des caractéristiques d'une équipe qu'un entraîneur devrait privilégier.

3) **Quels attributs de joueur devrait-il privilégier ?** Est-ce qu'un entraîneur devrait privilégier un attaquant dribbleur ? passeur ? Ceci est une question issue du thème 2), c'est-à-dire une étude des caractéristiques des joueurs selon leurs postes qu'un entraîneur devrait privilégier.

4) **Quelle formation d'équipe un entraîneur devrait-il privilégier ?** Est-ce qu'un entraîneur devrait privilégier une formation à 2 attaquants ? 3 attaquants ? Ceci est une question issue du thème 3), c'est-à-dire une étude des caractéristiques des plans de jeu qu'un entraîneur devrait privilégier.

5) **Conclusion:** Ce qu'il faut retenir de cette analyse.

B) Données utilisées, preprocessing

Pour cette étude, nous avons utilisés les tables de données suivantes:

- Match
- Teams, Teams_attribut
- Players, Players_attributs

On créer tout d'abord deux dataframes qui nous serviront pour le reste des études:

1. Un dataframe ***df_match_position*** issue de la table Match qui représente tous les matchs, dans lequel on a rajouté:
    a. les points gagnés pour chaque équipes (3 points pour l'équipe victorieuse, 0 point pour l'équipe perdante, et 1 points pour chaque équipe en cas d'égalité
    b. Les position des chaques joueurs en fonction de leurs coordonnées X Y sur la pelouse (voir notebook pour plus de détail)
    c. La formation adoptée pour chaque équipe en fonction de la position des joueurs
2. Un dataframe ***df_player*** qui repertorie, pour chaque saison, les joueurs ainsi que leurs équipes, et leur position sur le terrain de prédilection

Ensuite, nous avons étudié la qualité des données, et nous avons pu faire les remarques suivantes:

- teams et teams_attributs : toutes les équipes ne sont pas représentées dans teams_attributs (288 équipes sur 299). BuildUpPlayDribbling a des Nan mais ont été remplacer par le score de 30 (voir plus haut). On a également enlevé un duplicata dans teams_attributs.
- players et players_attributs: Quelques NaN a observer dans divers attributs.
- df_match_position: Quelques NaN concernant des joueurs

Ceci ne devrait pas fausser l'analyse, puisque nous agglomérons les données ensemble. .

C) Choix d'une métrique de performance

Le choix naturel pour évaluer une équipe d'une certaine saison, **est de comptabiliser le nombre de points moyens gagnés par match.**

On pourra ensuite, par saison, classer les équipes par ordre de performance. On fera bien attention de ne prendre que les équipes qui ont fait un nombre décent de matchs par saison (>27) afin de ne pas fausser les résultats.

Cette métrique est adaptée à notre problématique: en prenant, par exemple, les 10% des équipes les plus performantes, on pourra alors voir les attributs que ces équipes partagent entre elles, et les opposer à ceux du reste des équipes. Ceci donnera la corrélation entre la performance des équipes et les différentes caractéristiques testées. On pourra donc penser, qu'un entraîneur veut privilégier globalement ces attributs lorsqu'il compose la feuille de match.

On peut quand même se poser si on ne pourrait pas raffiner la métrique:

- En pondérant le nombre de points en fonction de l'équipe affrontée et/ou de l'équipe en elle-même. On pourrait utiliser les cotes par exemple. Une victoire contre une top équipe vaut plus que contre une faible équipe. On aurait pu utiliser par exemple la tables bets pour assigner à chaque équipe leur valeurs
- En pondérant avec le goal average (buts marqués vs but pris). Une victoire 3-0 est plus impressionnante que 1-0.
- En pondérant avec victoire domicile et extérieur. On sait qu'une victoire à l'extérieur est plus difficile à acquérir qu'à l'extérieur (voir notebook1)
- En restreignant aux principales ligues de même niveau. On sait que le championnat anglais est en général plus équilibré que certains autres petits championnats.

Nous avons pour l'instant écarté ces points, pour faire une première approche naïve.

Ceci permet de dresser le dataframe **df_classement** suivant:

	team_api_id	team_long_name	season	league_id	points	nombre_de_matchs	points_par_match
0	9773	FC Porto	2010/2011	17642	84.0	30	2.800000
1	9885	Juventus	2013/2014	10257	102.0	38	2.684211
2	9823	FC Bayern Munich	2012/2013	7809	91.0	34	2.676471
3	9823	FC Bayern Munich	2013/2014	7809	90.0	34	2.647059
4	8633	Real Madrid CF	2011/2012	21518	100.0	38	2.631579
...	...	...	...	...	...	...	...
1471	10242	ES Troyes AC	2015/2016	4769	18.0	38	0.473684
1472	9984	KSV Cercle Brugge	2012/2013	1	14.0	30	0.466667
1473	10252	Aston Villa	2015/2016	1729	17.0	38	0.447368
1474	10219	RKC Waalwijk	2009/2010	13274	15.0	34	0.441176
1475	8525	Willem II	2010/2011	13274	15.0	34	0.441176

On voit par exemple que le FC Porto de 2010 est l'équipe la plus performante avec en moyenne 2.8 points récoltés par match. Les premières équipes sont toutes vues comme des top équipes en général. Willem II en 2010, n'a en revanche gagné que 0.44 points par match. Les dernières équipes sont toutes vues comme des équipes plutôt faibles, donc il semble que la métrique fasse sens.

On peut s'intéresser à la distribution de ce classement:

```
In [57]: df_classement['points_par_match'].describe(percentiles=[0.1,0.25,0.5,0.75])

Out[57]: count    1064.000000
         mean        1.383453
         std         0.437212
         min         0.441176
         10%         0.884314
         25%         1.088235
         50%         1.289474
         75%         1.632018
         max         2.800000
         Name: points_par_match, dtype: float64
```

On voit que les top équipes sont vraiment au-dessus du lot, ce qui est un point intéressant.

D) Quels attributs d'une équipe un sélectionneur devrait privilégier ?

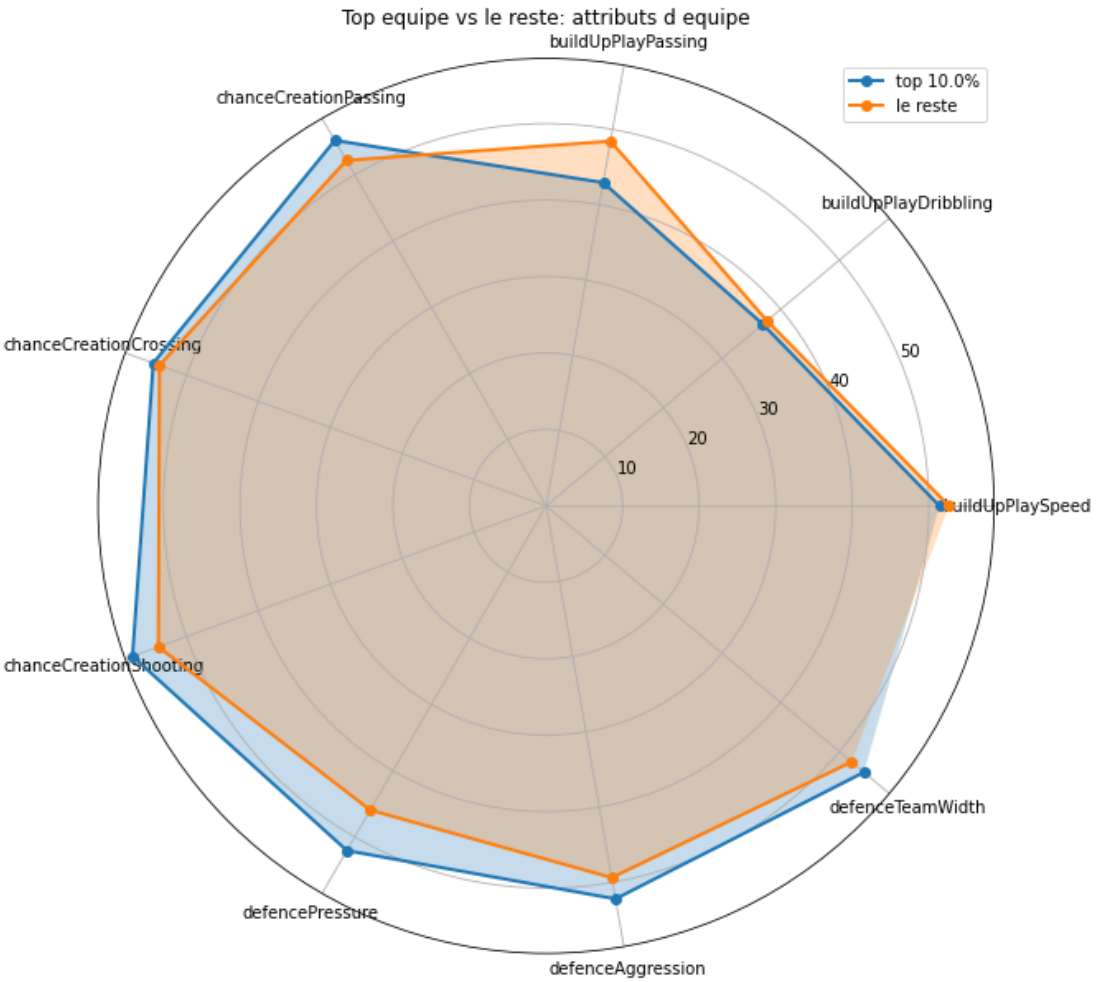
En prenant le dataframe **df_classement** créé précédemment, on peut ainsi ajouter les colonnes provenant de teams_attributs via les *clé team_api_id et seasons.*

On subdivise ce classement ensuite en 3 dataframes:

- **df_top_teams**: contient les équipes les plus performantes (par exemple 10%)
- **df_autre_teams**: contient le reste des équipes
- **df_flop_teams**: contient les équipes les moins performantes (par exemple 10%)

Pour chaque attributs d'équipes, on prend alors la moyenne dans chaque dataframe.

On plot les résultats:



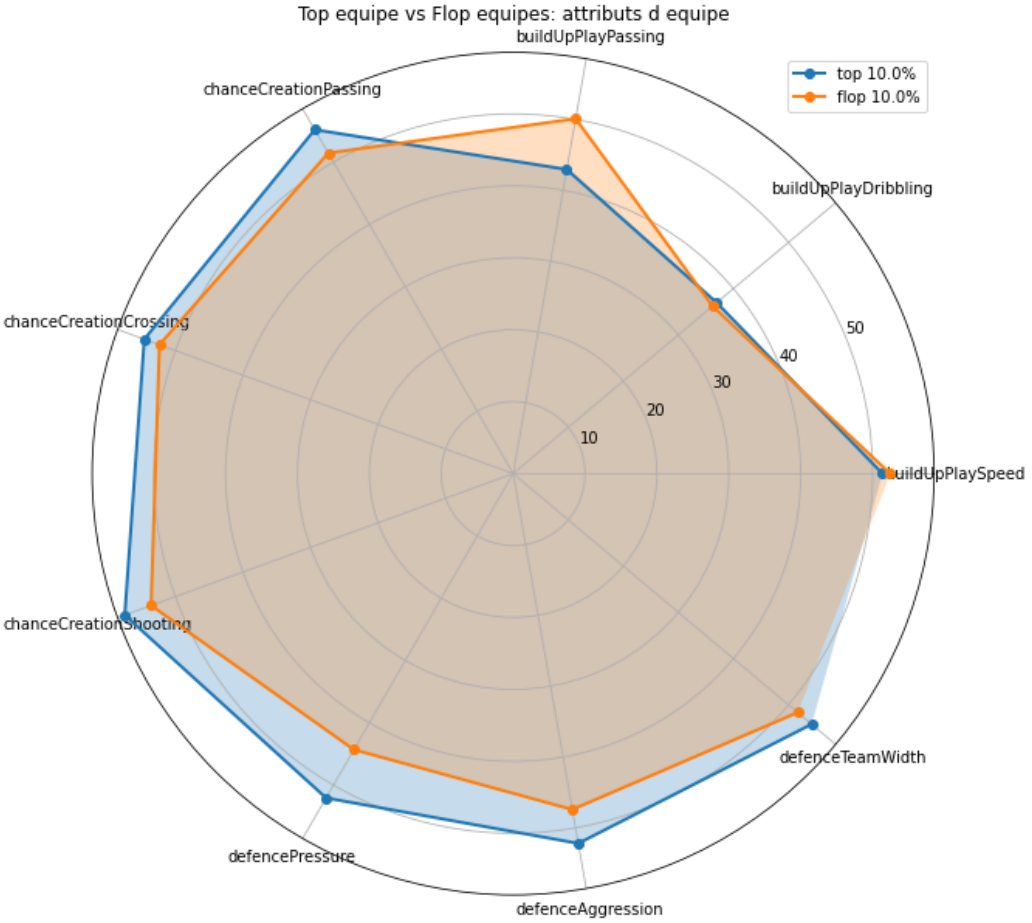
On observe que:

- **Les attributs défensifs** dans leur ensemble sont corrélés avec la bonne performance des équipes. En particulier, la pression des défenseurs sur les attaquants. On peut donc penser qu'un entraineur devrait non seulement privilégier **le secteur défensif, mais surtout recruter des joueurs capables d'agressivité et de pressuriser les attaquants** adverse avec un marquage solide.
- **La capacité d'une équipe à se créer des occasions** est aussi supérieure pour les top équipes. On voit ici que la corrélation entre les équipes qui ont des joueurs qui se mettent en situation de tir, ou délivrent de passes décisives et les bonnes performances est plus grande. Un entraîneur devrait donc surement aussi être attentif d'avoir des **joueurs créatifs, capables de tirer de loin par exemple, ou de délivrer des bonnes passes offensifs** au bon moment.
- Étonnamment, **le reste des équipes a en moyenne une meilleure faculté à créer un jeu collectif de passes ou avoir des bon dribbleurs**. On peut donc penser qu'un entraîneur **ne privilégiera pas une équipe capable de créer du beau football collectif dans la construction, ni d'avoir des dribbleurs de talents**.

Étant moi-même passionné de football et en discutant de ça, je trouve que ces résultats sont pleins de sens. En effet, on se rappellera de l'Italie championne grâce à son système de

défense musclé. De plus, c'est souvent une tactique payante pour les "petites équipes" de se reposer sur leur défense, tout en espérant marquer un but en contre attaque ou lors d'un coup de pied arrêté (Grèce championnat d'europe en 2004).
Comme dit le dicton: "posséder n'est pas gagné", avoir un beau jeu de passe est souvent spectaculaire, mais de mon point de vue moins efficace que des passes longues au bon moment ou des centres appliqués.

Nous pouvons faire la même analyse en prenant cette fois, les meilleurs et les moins bons. On espère voir la même tendance mais amplifiée. On peut prendre les 10% moins bon par exemple.



Les résultats sont ceux attendus. Ils sont similaires et même encore plus catégoriques. Ceci permet de donner plus de crédits aux analyses faites plus haut.

Conclusion:

Il semble qu'un entraîneur devrait privilégier une défense volontaire et engagée ainsi que des joueurs créatifs et habiles, capables de données de long ballons ou se mettant en situations de tirs. Ceci devrait se traduire par des joueurs jouant haut sur le terrain, empiétant ainsi constamment sur le terrain adverse, pour que les défenseurs puissent être toujours au pressing et que les joueurs offensifs soient à la réception d'un ballon long. On espère pouvoir faire une analyse sur le positionnement des joueurs qui

pourraient confirmer cette théorie. A contrario, il semble qu'il devrait faire passer au second plan le collectif, c'est-à- dire de jouer stratégiquement en construction et en passe.

E) Quels attributs de joueurs un entraîneur devrait-il privilégier ?

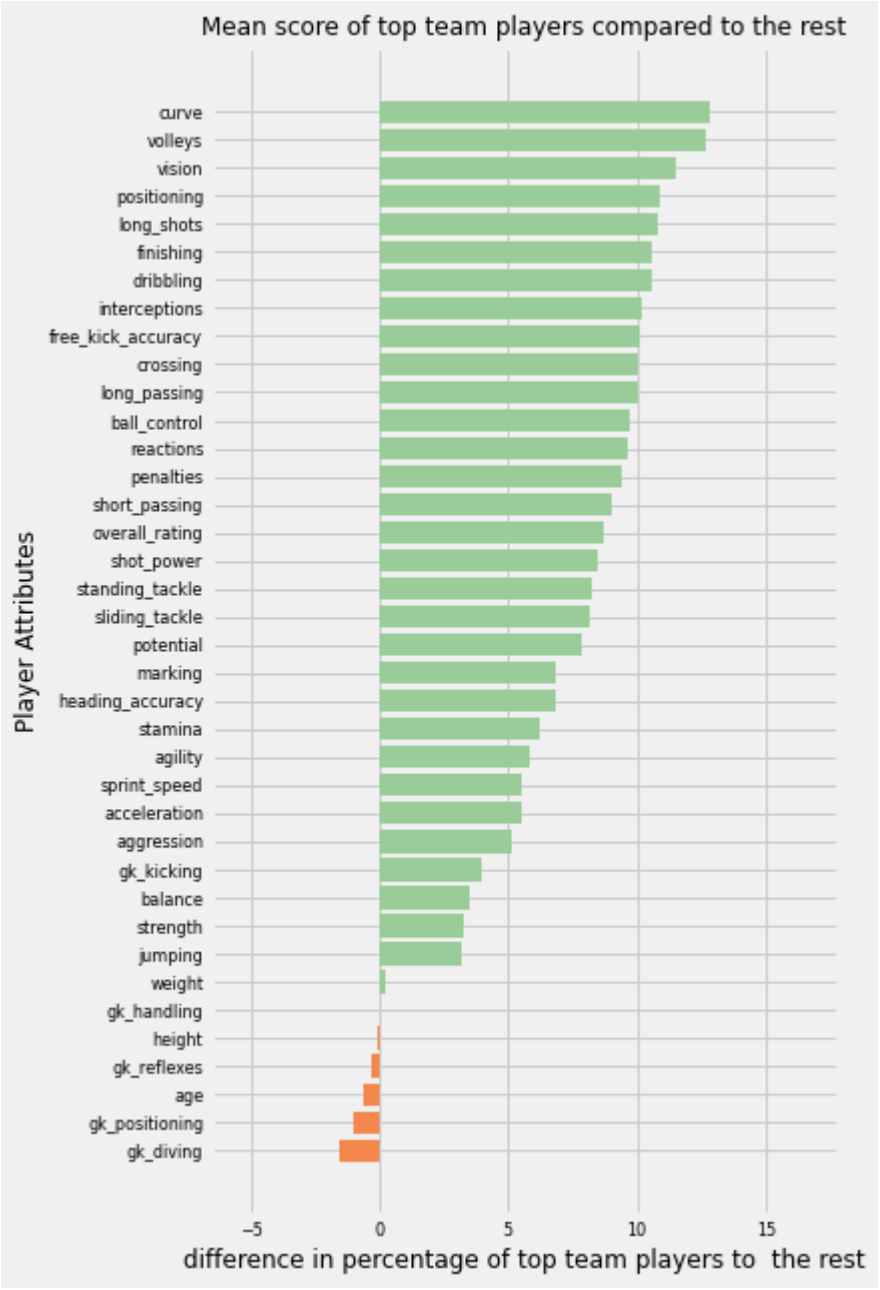
On fait une analyse similaire, mais cette fois avec les joueurs qui composent les équipes.

C'est-à-dire, on prend les joueurs des meilleures équipes et on compare au reste, et après au plus mauvais.

Similairement, nous utilisons le dataframe **df_players** mergé avec **players_attributs**, afin de récolter les principales caractéristiques des joueurs. On utilise alors les 3 dataframes (top, autre et flop teams) afin d'avoir les joueurs correspondants aux 3 niveaux de performances d'équipes.

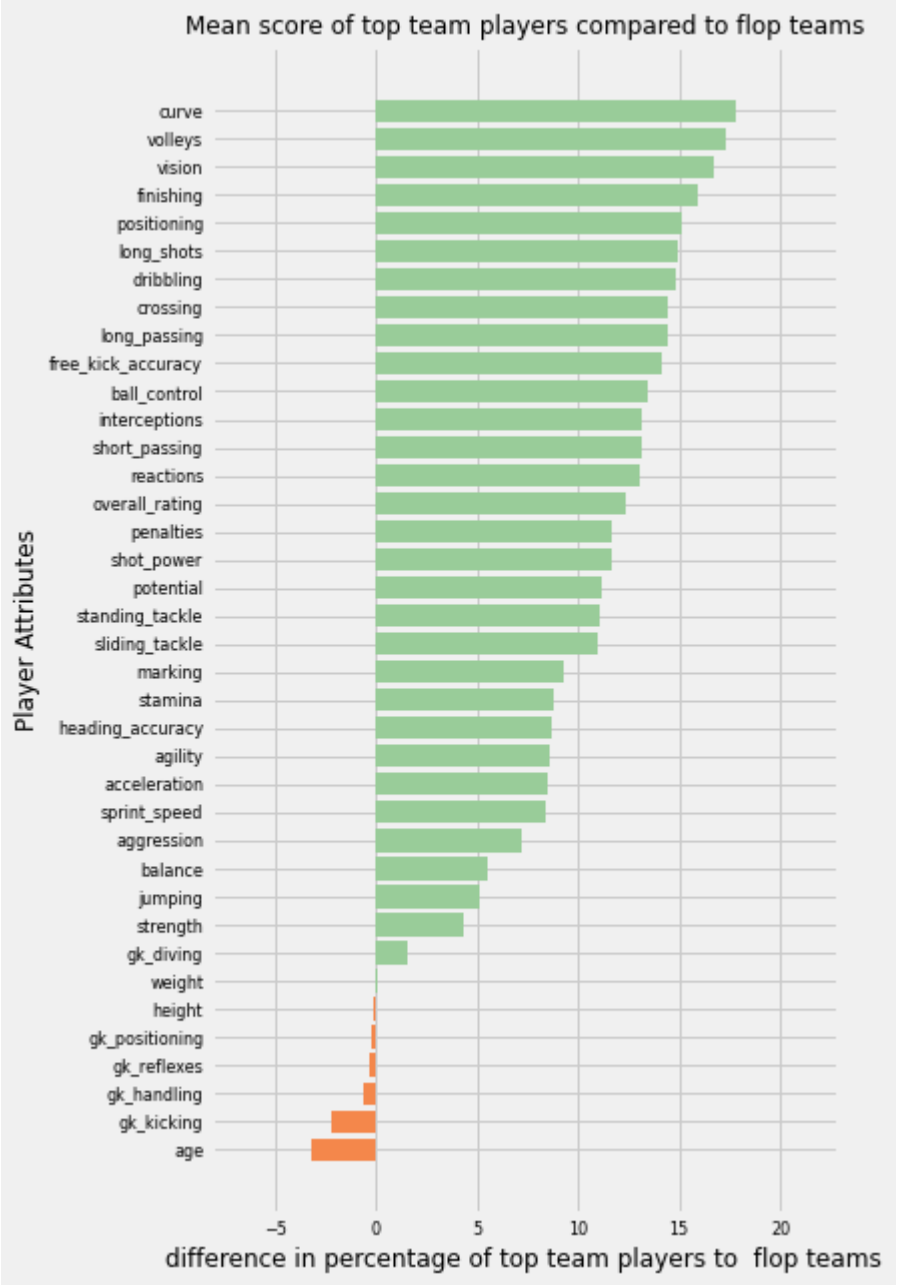
poste	birthday	height	weight	age	id	...	vision	penalties	marking	standing_tackle	sliding_tackle	gk_diving	gk_handling	gk_kicking	gk_p
u_defensif	1981-01-27 00:00:00	175.26	154	33.0	139844.0	...	55.0	66.0	62.0	63.0	54.0	12.0	11.0	6.0	
u_defensif	1981-01-27 00:00:00	175.26	154	33.0	139845.0	...	55.0	66.0	62.0	63.0	54.0	12.0	11.0	6.0	
u_defensif	1981-01-27 00:00:00	175.26	154	33.0	139846.0	...	55.0	66.0	62.0	63.0	54.0	12.0	11.0	6.0	
u_defensif	1981-01-27 00:00:00	175.26	154	30.0	139844.0	...	55.0	66.0	62.0	63.0	54.0	12.0	11.0	6.0	
u_defensif	1981-01-27 00:00:00	175.26	154	30.0	139845.0	...	55.0	66.0	62.0	63.0	54.0	12.0	11.0	6.0	

Il y a trop d'attributs pour faire un plot étoilé comme précédemment. On compare donc top vs autre et top vs flop en prenant la différence en pourcentage de chaque attributs.



- **On voit que la plupart des attributs "capacité a créer des occasions" qui semblaient être corrélés avec les equipes performantes (voir question précédentes) sont représentés ici (longshots, volley, finishing, vision, crossing, curve, long passing...). Ce qui est cohérent avec les résultats obtenus précédemment.**
- Il est intéressant d'observer que les **attributs défensifs (tackles, marking, aggression) ne sont pas ceux qui sortent les premiers (contrairement à la question précédente)**. Ceci laisse à penser qu'une fois qu'un certain niveau de défense est atteint, la capacité à créer des opportunités de buts est ce qui fait la différence entre les très grandes équipes et le reste.
- **Les attributs du goal semblent être négativement corrélés avec la performance !** (sauf la capacité à dégager le ballon, ce qui est cohérent avec ce qui est dit précédemment).

L'age également, ce qui est logique, puisque les top équipes sont en general formes de joueurs confirmés, qui ont fait leur preuve autrefois dans d'autres clubs plus "formateurs".



Les résultats sont similaires, mais sublimés.

Cette analyse peut se poursuivre à un niveau de granularité plus haut, c'est-à-dire par poste. (Voir Notebook).

Quelques highlights de l'analyse:

- **On voit que globalement, (à part pour le goal) les attributs offensifs (curve, volleys, vision, long shots/passing, finishing…) semblent être ce qui**

différencie les joueurs des tops équipes. Parmis eux, on retrouve tous postes confondues

- Pour le goal, hormis les attributs habituels, on voit que la vision et les longues passes sont importantes.
- Pour les défenseurs latéraux, en plus des attributs de finisher, le positionnement et interceptions (qui sont liés) semblent importants. C'est vrai que ces postes sont dévoués à la défense, mais sont aussi aujourd'hui formés pour jouer offensivement. Il faut donc qu'il réussissent à trouver le bon compromis entre l'attaque et la défense à travers un bon positionnement.
- Les défenseurs centraux, qui partagent beaucoup de similarités avec les défenseurs centraux, ont eux plus de jeu de tête. Ce qui fait sens: les défenseurs centraux sont généralement ceux qui montent dans la surface de réparation lors de coup de pieds arrêtés.
- Pour les milieux latéraux, on observe des caractéristiques similaires aux latéraux. ce qui est cohérents. Il semble que des milieux gauches des tops équipes soient capables de se projeter vers l'avant. De plus, le positionnement et l'interception semblent importants, ce qui est logique.
- Les milieux défensifs des top équipes ont des attributs offensifs plus hauts. Jusqu'à maintenant, on constate que ce processus est généralisé pour postes a tendance "défensifs". L'interception est l'attribut défensif le plus important.
- Ce qui est spécifique aux milieux offensifs, c'est la capacité de tirer des coups francs. C'est vrai que ce sont souvent les joueurs à ces postes de techniciens qui sont chargés de tirer les coup francs
- Pour les attaquants et les ailiers, les attributs sont ceux attendus. Il peut être un peu intriguant que le jeu de tête des attaquants ne soit pas mis en avant.

Conclusion:

- Les attributs offensifs sont mis en avant
- C'est souvent les mêmes attributs qui reviennent, qu'ils soient défensifs ou offensifs

F) Quelle formation un entraîneur devrait-il privilégier ?

Nous allons maintenant faire une analyse similaire avec les formations. Il est connu dans le football que les équipes adoptent différentes stratégies de formation à domicile et à l'extérieur. On va donc s'intéresser aux deux.

Pour chaque saison, pour chaque équipe, on retient la formation la plus utilisée.

```
In [125]: df_formation
```

Out[125]:

	team_api_id	season	home_formation	away_formation
0	1601	2013/2014	(4, 2, 3, 1)	(4, 2, 3, 1)
1	1601	2014/2015	(4, 2, 3, 1)	(4, 2, 3, 1)
2	1601	2015/2016	(4, 2, 3, 1)	(4, 2, 3, 1)
3	1773	2011/2012	(3, 4, 3)	(4, 3, 3)
4	1773	2012/2013	(4, 1, 4, 1)	(4, 4, 2)
...	...	...	...	...
1396	177361	2015/2016	(4, 2, 3, 1)	(4, 2, 3, 1)
1397	188163	2015/2016	(4, 3, 3)	(4, 3, 3)
1398	208931	2015/2016	(4, 4, 1, 1)	(4, 4, 1, 1)
1399	274581	2014/2015	(4, 2, 3, 1)	(4, 2, 3, 1)
1400	274581	2015/2016	(4, 2, 3, 1)	(4, 3, 3)

1401 rows × 4 columns

Comme précédemment, on merge ce dataframe avec les trois dataframes utilisées pour les comparaisons, **df_top_teams, df_autre_teams et df_flop_teams**.

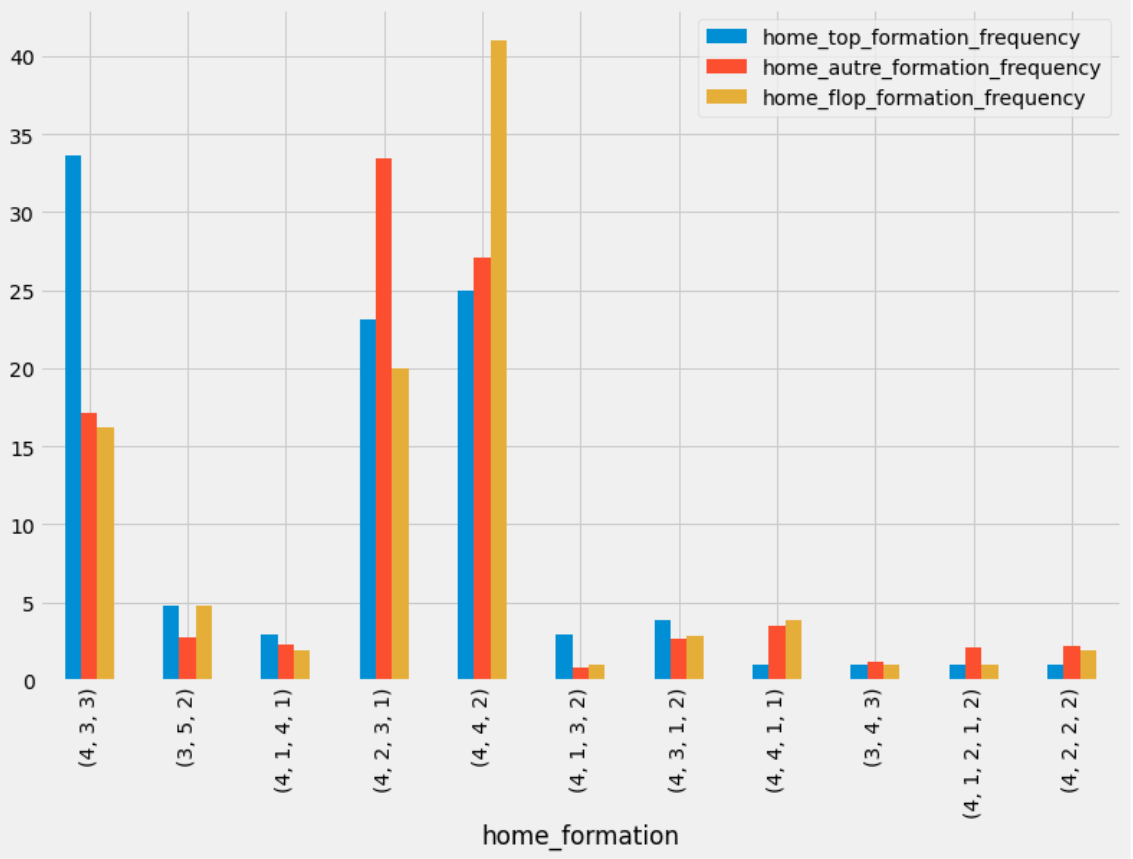
Enfin, pour chacun de ces dataframes, on calcule la fréquence d'apparition des formations pour toutes équipes et saisons confondues. Par exemple, pour les matchs à domicile:

```
In [126]: home_formation
```

Out[126]:

	home_formation	home_top_formation_frequency	home_autre_formation_frequency	home_flop_formation_frequency
0	(4, 3, 3)	33.653846	17.086528	16.190476
1	(3, 5, 2)	4.807692	2.738226	4.761905
2	(4, 1, 4, 1)	2.884615	2.300110	1.904762
3	(4, 2, 3, 1)	23.076923	33.406353	20.000000
4	(4, 4, 2)	25.000000	27.053669	40.952381
5	(4, 1, 3, 2)	2.884615	0.766703	0.952381
6	(4, 3, 1, 2)	3.846154	2.628697	2.857143
7	(4, 4, 1, 1)	0.961538	3.504929	3.809524
8	(3, 4, 3)	0.961538	1.204819	0.952381
9	(4, 1, 2, 1, 2)	0.961538	2.081051	0.952381
10	(4, 2, 2, 2)	0.961538	2.190581	1.904762

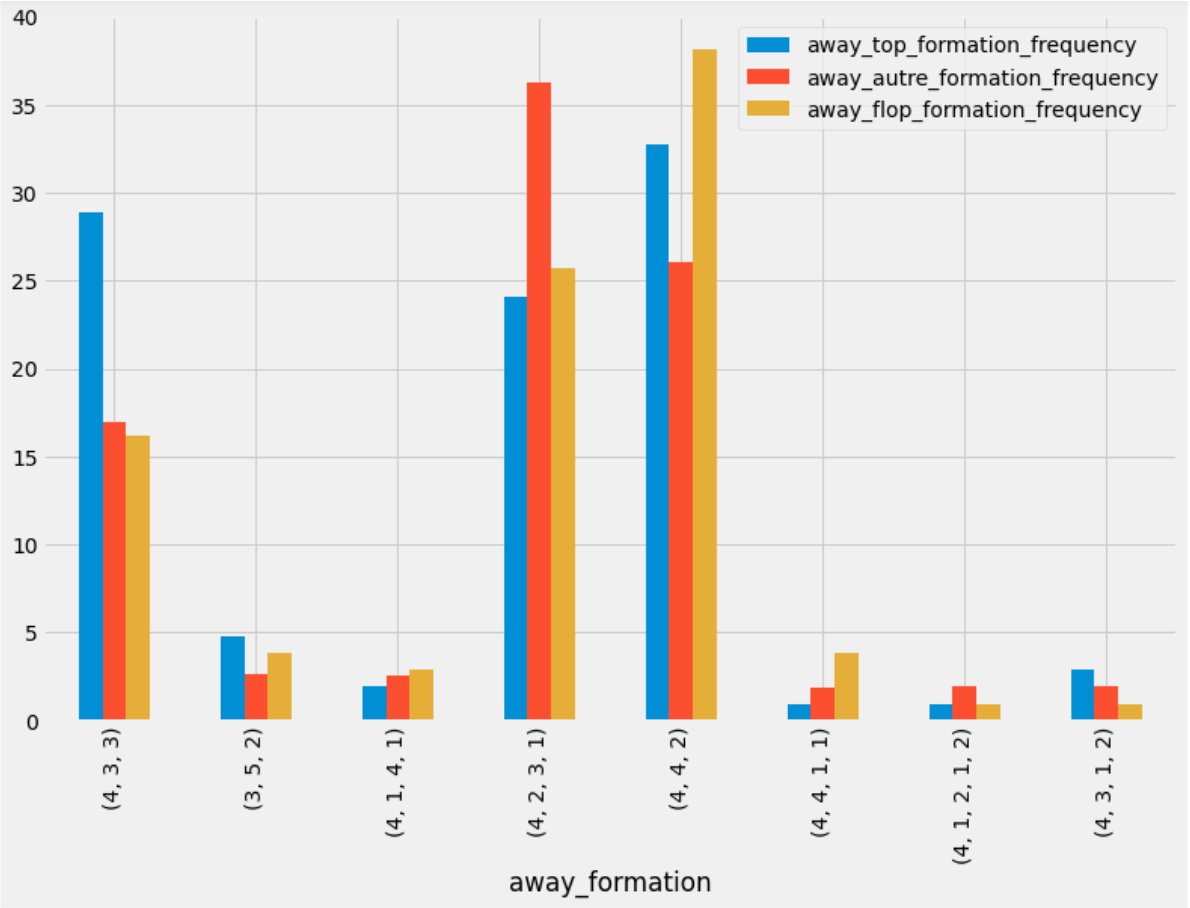
On peut alors faire un barplot pour comparer la corrélation des diverses formations avec la performance des équipes, d'abord à domicile puis à l'extérieur.



D'après nos connaissances du football (et les données), on sait que les formations les plus utilisées sont (4,3,3) , (4,2,3,1) et surtout (4,4,2). Notre comparaison s'effectuera entre les 3.

C’est intéressant de voir que la formation (4,3,3) est bien plus corrélée avec la performance que le (4,4,2). Le (4,3,3) est une formation offensive, rendue populaire par la grande AJAX d’amsterdam. A l’inverse, le (4,4,2) qui est souvent vu comme une solution plus sûre par les entraîneurs est plus représenté dans les équipes moins performantes, surtout dans les plus mauvaises.

Voyons à l’extérieur:



La tendance se confirme, même si à l'extérieur, le (4,4,2) est plus représenté dans les top équipes. Comme dit plutôt, il s'agit d'une composition plus "safe", adaptée aux déplacements difficiles à l'extérieur.

G) Conclusion

Dans ce notebook, on aura vu:

1. L'établissement d'une métrique pour pouvoir jauger la performance des équipes, et ainsi donner une marche à suivre aux entraîneurs dans leurs choix. Cette métrique est basée sur le nombre de points moyen par match gagné par équipe, qui nous permettra de les classer en 3 catégories de niveau afin d'en comparer leurs attributs. On peut se demander si la métrique ne devrait pas être raffinée, pour une comparaison plus juste.
2. La comparaison des attributs entre les teams de différents niveaux. Il ressort de cette analyse, que globalement le secteur défensif devrait être privilégié, avec des défenseurs qui jouent hauts, et des ballons longs distribués aux attaquants, contredisant alors une stratégie basée sur le collectif.
3. La comparaison des attributs entre les joueurs de différents niveaux, montre qu'à titre individuel, les caractéristiques offensifs sont souvent ce qui fait la différence entre les équipes performantes et le reste. Il est toutefois important d'observer que

les caractéristiques défensives (à contrario de la question précédente) sont moins représentées.

4. La comparaison des formations est très informative, elle semble être en faveur d'un (4,3,3).

Si nous devions répondre à la question du notebook en bref:

Un entraîneur devrait opter pour une équipe qui joue globalement haut avec une défense agressive et engagée, et qui est capable de délivrer des ballons de loin aux joueurs. Ces derniers devraient avoir de bonnes capacités offensives et surtout de bonnes aptitudes à se créer des occasions. Le 4,3,3 est une formation idéale pour ce type de jeu, puisqu'elle contient 3 attaquants à la réception des ballons.

En revanche, les stratégies plus défensives, basées sur un positionnement bas (4,4,2) avec un jeu en passe pour amener le ballon collectivement et progressivement vers l'avant semblent moins payantes.

IV) Notebook 2: Jouer à domicile offre t-il réellement un avantage ?

A) Objectif du notebook

Dans le football est souvent admis que jouer à domicile offre un réel avantage, qui se dit à la fois psychologique (dû aux supporters, les habitudes,...) mais aussi concret (facilité logistique, pas de déplacement,...). Cette avantage est d'ailleurs souvent pris en compte, c'est le cas notamment en champions league, ou les buts marqués à l'extérieur comptent double.

Dans ce notebook, on cherche à savoir si cet avantage se retrouve dans les données à l'aide de quelques visualisations. On va traiter les grands axes suivants:

1. **Une équipe a-t-elle plus de chance de gagner à domicile qu' à l'extérieur ?**
2. **Est-ce que les équipes marquent/prennent davantage de buts à domicile/extérieur?**
3. **L'arbitrage a-t-il tendance à favoriser les équipes à domicile ?**
4. **conclusion**

B) Données utilisées, Preprocessing

Pour cette étude, nous avons seulement utilisé la table *match.* En effet, on ne s'intéresse pas aux joueurs. aux équipes en elle-même ou encore au ligue, mais seulement si à travers les matchs effectués, cette avantage est palpable.

Ici peu de preprocessing est à faire. Pour ce qui est du nettoyage des données, on invite le lecteur à se référer au Notebook 1, dans la section "données utilisées".

C) Victoire à domicile vs à l'extérieur

On commence par créer une colonne additionnelle qui, pour chaque match, donne si le vainqueur est l'équipe jouant à domicile, à l'extérieur, ou si il y eu match nul.

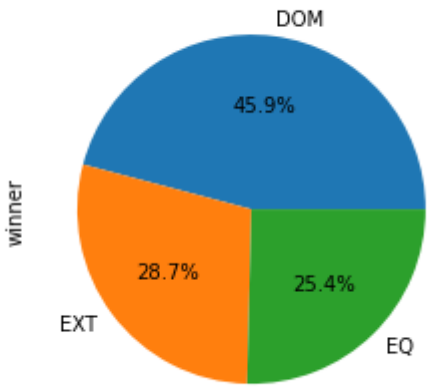
```
In [16]: match.head()
```

Out[16]:

	country_id	league_id	season	date	match_api_id	home_team_api_id	away_team_api_id	home_team_goal	away_team_goal	winner
0	1	1	2008/2009	2008-08-17 00:00:00	492473	9987	9993	1	1	EQ
1	1	1	2008/2009	2008-08-16 00:00:00	492474	10000	9994	0	0	EQ
2	1	1	2008/2009	2008-08-16 00:00:00	492475	9984	8635	0	3	EXT
3	1	1	2008/2009	2008-08-17 00:00:00	492476	9991	9998	5	0	DOM
4	1	1	2008/2009	2008-08-16 00:00:00	492477	7947	9985	1	3	EXT

Une première rapide analyse intéressante serait de regarder de comparer le pourcentage de matchs gagnés à domiciles (resp. extérieur et égalité) :

```
Out[17]: <AxesSubplot:ylabel='winner'>
```



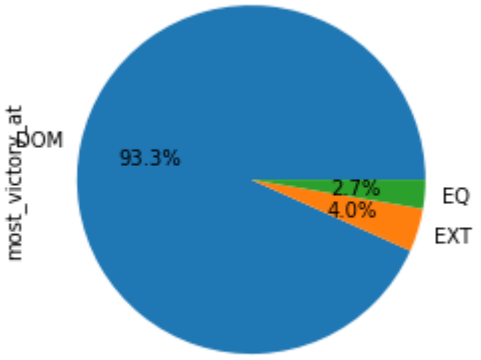
On constate que presque 50% des matchs joués à domicile sont gagnés tandis que les défaites et les égalités sont répartis presque équitablement. On pourrait naïvement émettre l'hypothèse qu'une équipe (prise aux hasard) jouant à domicile à 46% de chance de gagner et une équipe jouant à l'extérieur a 1 chance sur 4 de remporter le match.

On peut reproduire l'analyse au niveau des équipes. On crée alors un dataframe **df_team_win** qui représente chaque équipe, avec leur nombre de victoires (resp. défaites et égalités) à domicile et à l'extérieur.

Out[110]:

victoire_domicile	egalite_domicile	defaite_domicile	away_team_api_id	defaite_exterieur	egalite_exterieur	victoire_exterieur	most_victory_at	most_defeat_at
54	28	38	1601	53	29	38	DOM	EXT
16	13	16	1773	24	15	6	DOM	EXT
63	27	30	1957	59	37	24	DOM	EXT
19	27	29	2033	34	28	13	DOM	EXT
74	27	19	2182	37	33	50	DOM	EXT
...	...	...	...	...	...	...	...	...
18	11	20	158085	21	18	10	DOM	EXT
5	4	6	177361	7	5	3	DOM	EXT
3	4	10	188163	10	2	5	EXT	EQ
6	5	8	208931	10	6	3	DOM	EXT
9	7	14	274581	18	7	5	DOM	EXT

On peut donc regarder le pourcentage des équipes qui ont le plus gagné à domicile (resp. à l'extérieur):



On observe 93% des équipes comptabilise plus de victoires à domicile qu'à l'extérieur.

Conclusion: il apparaît clair que jouer à domicile offre un réel avantage pour obtenir la victoire.

D) Buts à domicile vs à l'extérieur

On peut prolonger l'analyse en regardant le nombre de buts marqués/pris à domicile et à l'extérieur.

On affiche les statistiques des buts inscrits par match:

Out[30]:

	home_team_goal	away_team_goal
count	25979.000000	25979.000000
mean	1.544594	1.160938
std	1.297158	1.142110
min	0.000000	0.000000
25%	1.000000	0.000000
50%	1.000000	1.000000
75%	2.000000	2.000000
max	10.000000	9.000000

Clairement, avec une variance assez voisine, le nombre de buts marqués par l'équipe à domicile est globalement supérieur à celui de l'équipe extérieur.

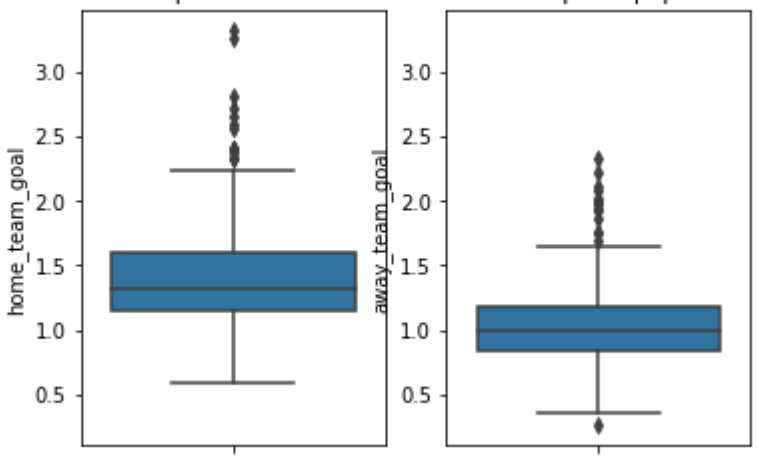
On peut faire la même analyse, mais cette fois par équipe, on créer un dataframe qui encode, pour chaque team, son nombre de buts moyen marqué à domicile et à l'extérieur:

Out[34]:

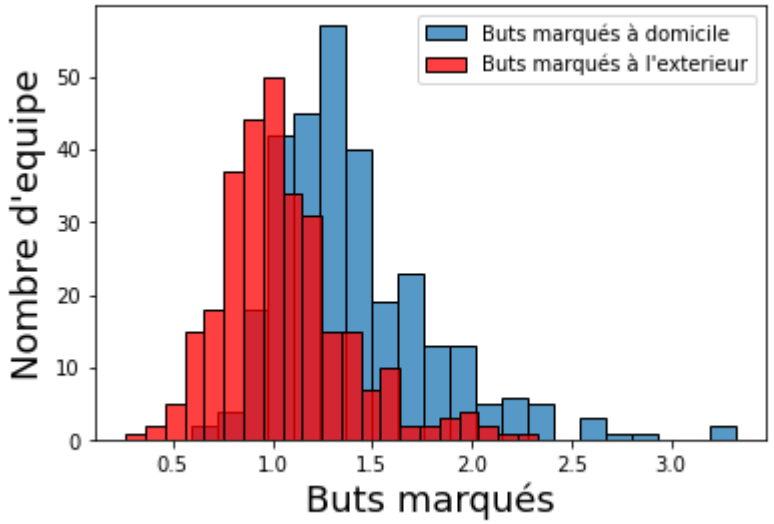
	home_team_goal	away_team_goal
1601	1.233333	1.100000
1773	1.644444	1.155556
1957	1.466667	0.925000
2033	0.933333	0.906667
2182	1.791667	1.308333
...	...	...
158085	1.142857	0.918367
177361	1.400000	0.800000
188163	0.882353	1.117647
208931	1.210526	0.736842
274581	1.366667	1.000000

On peut comparer leurs distributions:

Boite a moustache des buts marqués a domiciles et à l'exterieur par équipe

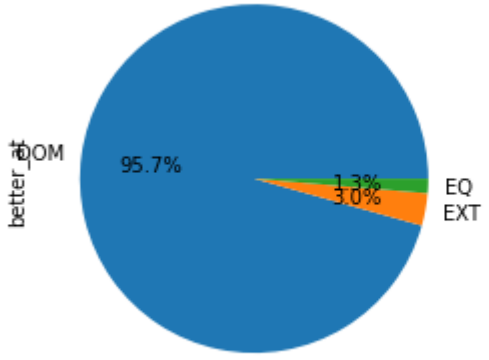


Distribution de la moyenne de buts marqués à domicile et à l'exterieur

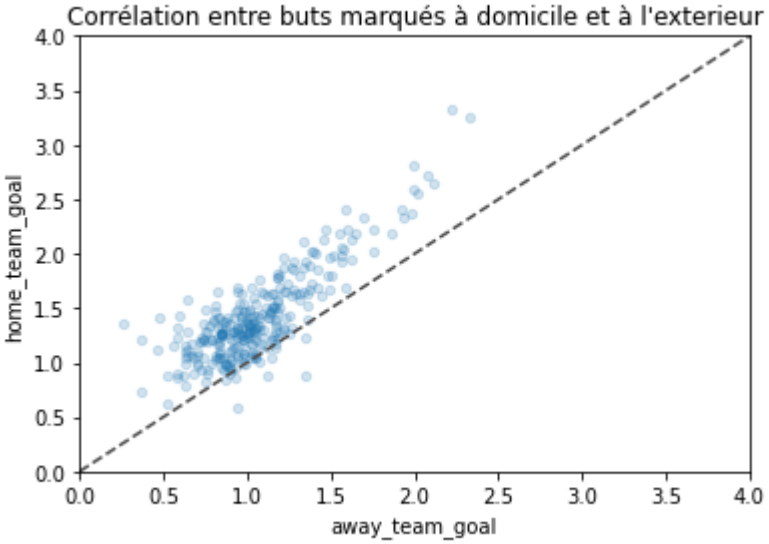


Ces deux distributions sont approximativement normales avec une variance similaire, mais, comme attendu, avec une moyenne supérieure de buts marqués à domicile pour les équipes.

Comme le montre le graphique ci dessous, 96% des équipes ont, sur le total de leurs buts marqués, le plus marqué à domicile.



Enfin, un dernier plot intéressant serait de regarder les équipes sous forme de nuage de points:

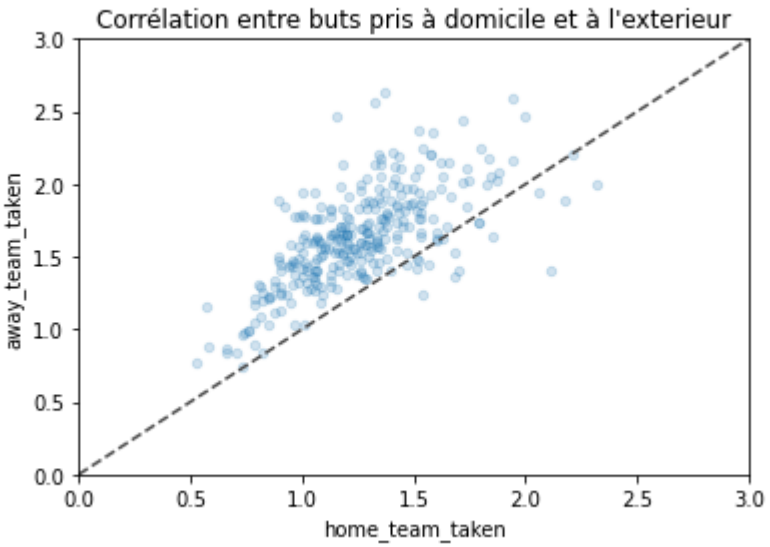


Ce plot donne une visualisation supplémentaire des analyses faites ci-dessus, en montrant toutefois une grande corrélations, pour chaque équipe, entre les buts marqués à domicile et à l'extérieur. Ceci indique que les équipes performantes à l'extérieur le sont aussi à domicile et vice versa.
Il est toutefois intéressant d'observer que quelques équipes sortent de ce nuage de points et affiche une certaines différences de performances entre à domicile et à l'extérieur. Il semble s'agir néanmoins d'équipes qui ne marquent pas beaucoup, ce qui aurait pu biaiser l'observation. Dans les équipes performantes, un tel phénomène ne se voit pas.

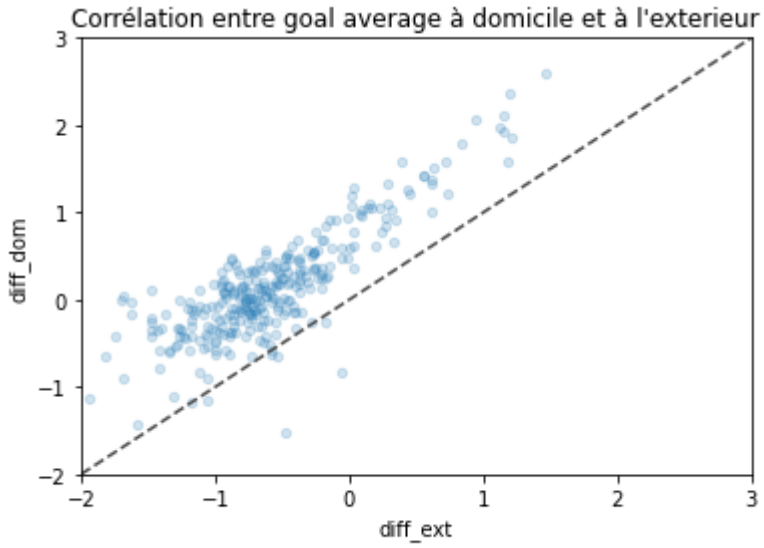
La même analyse a été conduite avec les buts encaissés, et le goal average (qui est la différence entre les buts marqués et pris). Ces analyses ont mené aux mêmes conclusions donc afin d'éviter toute redondance dans ce rapport, on invite le lecteur à se référer au notebook pour les voir.

On peut quand même montrer le graph de nuage de points pour ces analyses:

```
Out[113]: [<matplotlib.lines.Line2D at 0x2167538d880>]
```

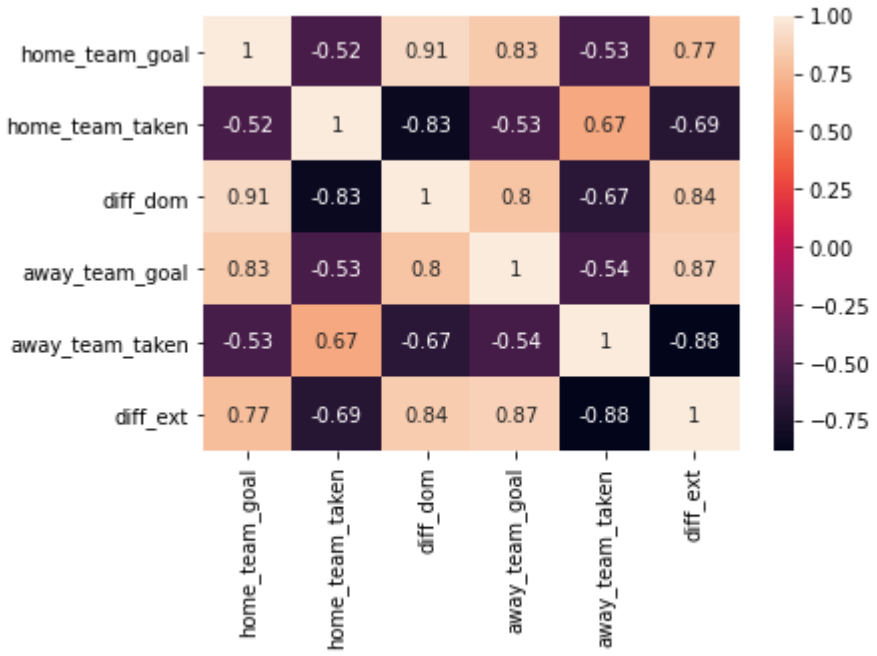


Concernant les buts pris, on observe une corrélation conséquente mais néanmoins moins évidente que pour les buts marqués. **Pareils, les teams qui sortent de cette corrélation sont des teams qui prennent beaucoup de buts.**



Des conclusions similaires sont à tirées pour le goal average.

Ces corrélations peuvent être synthétisés dans une matrice:



En conclusion: Les équipes ont plus de chance de marquer de buts à domicile qu' à l'extérieur mais ont plus de chance de prendre des buts lorsqu'elles jouent à l'extérieur. Ceci explique que le goal average (différence de buts) soit meilleur à domicile qu' à l'extérieur. Il existe une très forte corrélation pour chaque équipe entre les performances à domicile et à l'extérieur, sauf pour certaines équipes, qui sont néanmoins, toutes peu performatives.

E) Arbitrage à domicile vs extérieur

La dernière analyse de ce notebook se porte sur l'arbitrage. Nous allons en effet nous intéresser à savoir si, en moyenne, une équipe jouant à domicile bénéficie plus des "faveurs" de l'arbitre que l'équipe adverse. Cette question est légitime au vu des nombreux scandales qui ont explosé sur ce sujet là, et en dehors du scope de la corruption, il est possible que, consciemment ou inconsciemment, les arbitres soient soumis à l'influence du public.

Les données concernant l'arbitrage (fautes commises, cartons,...) sont malheureusement sous forme xml. Un premier challenge a été d'extirper les données et de les rendre sous un format utilisable.

Malheureusement, de nombreux matchs ne contiennent pas toutes ces informations, donc nous ne disposons à la fin de l'opération de toutes les données originelles. Nous obtenons donc un dataframe, **referee_match**, qui pour chaque match, indique le nombre de fautes, de cartons rouges et jaunes obtenus par l'équipe à domicile et celle à l'exterieur.

Out[74]:

	match_api_id	foul_home_team	foul_away_team	yellow_card_home_team	yellow_card_away_team	red_card_home_team	red_card_away_team	home
1728	489042	16	11	3	0	0	0	
1729	489043	11	9	0	0	0	0	
1730	489044	13	12	0	2	0	0	
1731	489045	14	13	2	1	0	0	
1732	489046	11	13	0	1	0	0	
...	...	...	...	...	...	...	...	
25944	1992225	0	0	1	4	0	0	
25945	1992226	0	0	0	2	0	0	
25946	1992227	0	0	2	3	0	0	
25947	1992228	0	0	1	2	0	0	
25948	1992229	0	0	2	2	0	0	

14217 rows × 12 columns

On observe dans certaines incohérences au niveau du nombre de fautes, il semble que pour certains matchs, il y aurait "0 faute mais quand même des cartons distribués". On peut donc penser que les données n'ont simplement pas été rentrées au niveau des fautes. Il semble aussi que lorsque les fautes n'ont pas été compté pour une team, alors elle n'ont pas été compté pour l'autre team.
Ça ne devrait donc pas avoir d'incidence lors de notre comparaison. En groupant par équipe, en comptant le nombre de matchs, on calcule le nombre de fautes par match moyen pour chaque équipe, on obtient, par exemple pour à l'extérieur:

In [124]:
```
foul_away_team
```

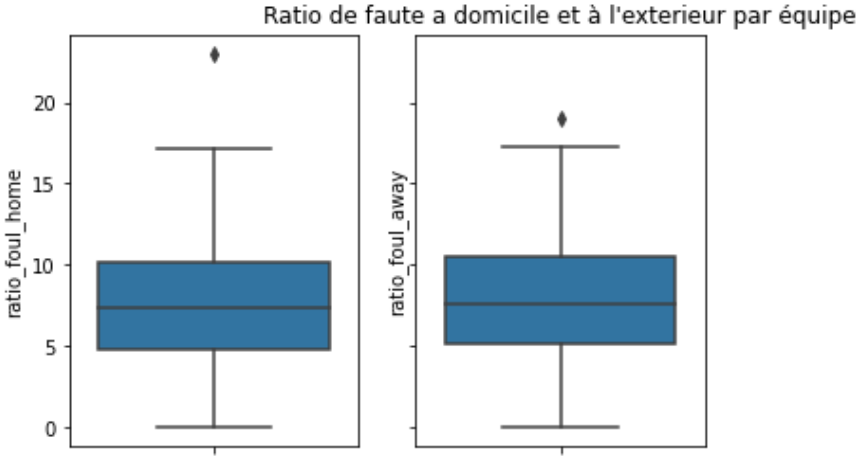
Out[124]:

away_team_api_id	foul_away_team	yellow_card_away_team	red_card_away_team	nb_match	ratio_foul_away	ratio_yellow_card_away	ratio_red_card_away
1601	15	3	0	1	15.000000	3.000000	0.000000
1957	19	5	0	1	19.000000	5.000000	0.000000
2182	11	3	0	1	11.000000	3.000000	0.000000
2183	13	2	0	1	13.000000	2.000000	0.000000
4087	285	141	7	70	4.071429	2.014286	0.100000
...	...	...	...	...	...	...	...
10269	945	284	10	136	6.948529	2.088235	0.073529
10278	100	52	2	19	5.263158	2.736842	0.105263
10281	182	178	7	76	2.394737	2.342105	0.092105
108893	0	22	2	11	0.000000	2.000000	0.181818
208931	302	57	1	19	15.894737	3.000000	0.052632

199 rows × 7 columns

Les deux dataframes **foul_away_team** et **foul_home_team** ont des distributions très similaires concernant le nombre de fautes sifflées.

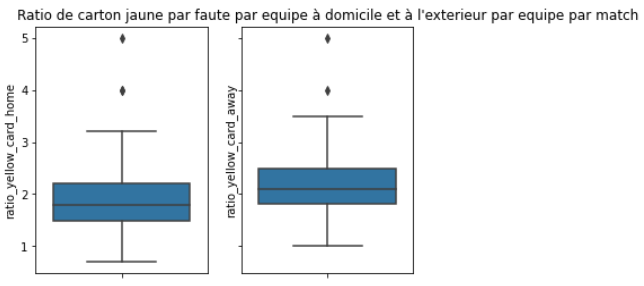
```
Out[87]: Text(0.5, 1.0, "Ratio de faute a domicile et à l'exterieur par équipe")
```

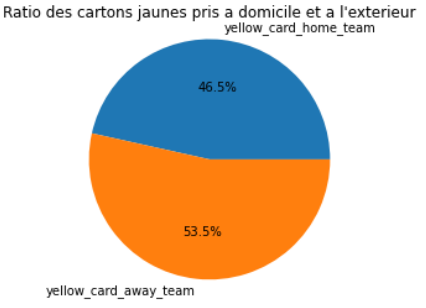


On ne peut donc pas vraiment significativement conclure que l'équipe domicile a un avantage en termes de fautes sifflées.

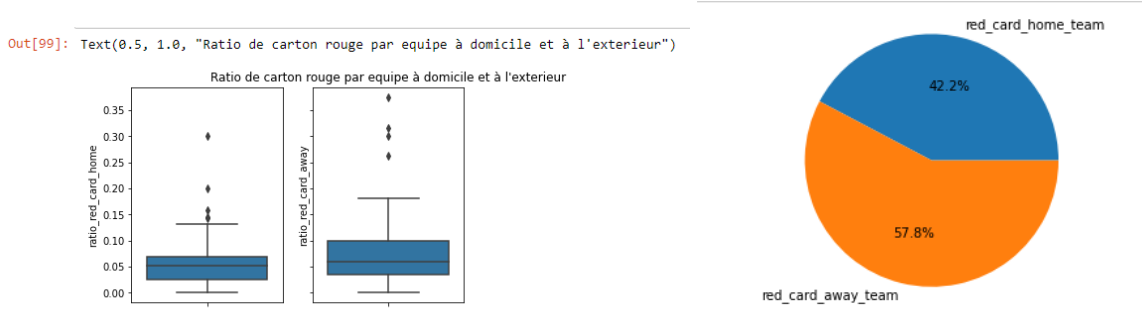
On peut maintenant s'intéresser au nombre de cartons. Nous avons calculé le nombre de cartons moyens par match pour chaque équipe:

Concernant les cartons jaunes:





Un (très) léger avantage pour les équipes à domicile au niveau des cartons jaunes est discernable. Pour les cartons rouges:



Pour les cartons rouges les résultats sont encore plus significatifs. Il faut toutefois aussi se rendre compte qu'il y a peu de carton rouge, ce qui a tendance naturellement à engendrer des écarts plus conséquents.

F) Conclusion

Dans ce notebook, nous avons vu:

1. Les équipes gagnent globalement plus à domicile.
2. Les équipes marquent globalement plus à domicile et prennent plus de buts à l'extérieur. On constate très peu d'équipes échappant à cette tendance, et si c'est le cas, il s'agit alors d'équipes "faibles".
3. Une équipe donnée ne reçoit pas significativement de traitement de faveur à domicile. On observe malgré tout un petit biais concernant les cartons (rouges) distribués, mais il est difficile de savoir si cela est dû au manque de données ou à une réelle influence sur l'arbitrage.

Si nous voulons répondre à la question du notebook en bref:

Il est clair qu'en termes de buts marqués, encaissés et surtout de victoires, l'équipe jouant à domicile a clairement un avantage. Les seules équipes qui échappent à cette règle sont ponctuellement des équipes faibles, dont les performances sont moins rigoureuses et donc plus randomisées. Concernant l'arbitrage, il semblerait que l'équipe jouant à domicile ne bénéficie globalement pas de traitement de faveur en termes de fautes sifflées, mais se voit globalement distribuée moins de cartons

jaunes (et encore moins de cartons rouges). Il faut cependant nuancer cette analyse: les données montrent des incohérences et les cartons sont globalement des faits de jeux plus rares, les différences en pourcentages sont donc plus grandes.

V) Conclusion

Dans le cadre de ce projet, nous avons utilisé les données “european soccer leagues” de Kaggle, afin de répondre par des analyses statistiques et diverses visualisations à des questions de recherche. Ces données répertorient sur 8 ans (2008-2016) un grand nombre de matchs, joueurs, équipes, ainsi que des attributs intéressants les concernant.

En plus d’être personnellement passionné par le football, ces données, divisées en plusieurs tables, ont été une aubaine pour ce projet. Premièrement, elles sont suffisamment complètes et fournies pour imaginer beaucoup d’axes d’études très intéressants. Non seulement on a abordé que deux thèmes parmi une multitude possible, mais on aurait pu aussi, pour les deux axes choisis, opter pour d’autres approches et questions de recherche. Enfin, bien que de qualités, ces données ont demandé beaucoup de traitements afin d’arriver aux structures recherchées, ce qui m’a permis de me familiariser avec les bibliothèques principales de data cleaning, data wrangling (pandas) et de visualisation (matplotlib, seaborn) de Python.

On a choisi de traiter les deux questions suivants:

1. Quels choix un entraîneur devrait-il privilégier pour composer une équipe performante ?
2. Est-ce que jouer à domicile représente un avantage ?

A l’aide de nos analyses, on a pu dégager des éléments de réponse.

Pour la première interrogation, nous avons comparé un certain nombre d’attributs d’équipe, de joueurs et de formations entre les équipes les plus performantes et les autres. Les résultats semblent indiquer qu'un entraîneur devrait opter pour une équipe qui joue globalement haut avec une défense agressive et engagée, et qui est capable de délivrer des ballons de loin aux joueurs. Ces derniers devraient avoir de bonnes capacités offensives et surtout de bonnes aptitudes à se créer des occasions. Le 4,3,3 est une formation idéale pour ce type de jeu, puisqu’elle contient 3 attaquants à la réception des ballons. A contrario, les équipes gardant la balle à ras-terre et privilégiant le jeu de passe collectif dans une formation offrant plus de contrôle au milieu de terrain (4,4,2 ou ,4,,2,3,1) sont moins corrélées avec la performance. Une amélioration potentielle de l’analyse serait de raffiner la métrique ou de trouver d’autres gages de performance et/ou de d'entremêler les attributs (i.e étant donné une formation de jeu, quels profils de joueurs ?). On pourrait pour se faire utiliser les réseaux bayésiens.

Concernant le deuxième problème, la réponse était plus attendue mais les statistiques le confirment nettement: les équipes ont un avantage conséquent à domicile. Globalement, les équipes gagnent, marquent plus à domicile et prennent moins de buts. On constate très peu d'équipes échappant à cette tendance, et si c’est le cas, il s’agit alors d’équipes “faibles”. L’analyse sur l’arbitrage est quant à elle peu probante. Non seulement on peut s’interroger sur la qualité des données (beaucoup de NaN, incohérence au niveau des fautes), mais il

semble aussi qu'au niveau des résultats, une équipe donnée ne reçoit pas de traitement de faveur à domicile. On observe malgré tout un petit biais concernant les cartons (rouges) distribués, mais il est difficile de savoir si cela est dû au manque de données ou à une réelle influence sur l'arbitrage. Un approfondissement intéressant de la question serait d'observer l'impact du public sur la qualité d'une équipe. Certains clubs sont connus pour leurs publics très actifs, il serait donc instructif de savoir si de telles équipes obtiennent des résultats encore plus marqués.