DataScientest • com

Rapport Technique d'évaluation

Qu'est-il arrivé à Arsenal ?

Promotion

Juin 2021

Participants:

Coralie MANGIN

Nathalie TOUPET

Pascale ASSI

SOMMAIRE

I. Contexte

a. Contexte d'insertion du projet

Pour ce projet, l'objectif principal a été d'apporter une analyse concernant les raisons qui pourraient expliquer la dégradation du niveau du club de football Arsenal durant la dernière décennie.

Pour un data analyst professionnel, ce type de projet serait pertinent puisqu'il pourrait y avoir différentes sortes d'enjeux pour plusieurs types de clients.
Le client pourrait être le président du club de football qui souhaiterait redresser la barre et trouver les clés pour faire retrouver le niveau d'antan à son équipe. Il y a évidemment un enjeu économique car de nombreux revenus sont générés par les compétitions de football.
Ca pourrait également être un journal sportif qui souhaiterait apporter une dimension nouvelle à ses articles.
Enfin, le client pourrait être un bookmaker qui souhaiterait apporter un nouvel axe d'interprétation dans les prévisions de ses paris sportifs.

Concernant le contexte technique, du point de vue d'un data analyst ce projet comporte évidemment plusieurs défis, notamment:

- la récupération des données qui ne sont pas toujours facilement accessibles;
- l'étendue des axes d'analyse, puisqu'il y a de très nombreux facteurs à prendre en compte et nombreuses sont les mauvaises interprétations possibles;
- la conception de visualisations pertinentes et surtout facilement compréhensibles. Puisqu'il y a de nombreuses variables, il faut veiller à ne pas surcharger les graphes d'informations, ce qui rendrait la lecture très difficile. Cela signifie qu'il y a un vrai travail d'identification des variables les plus pertinentes.

Pour nous, en tant que débutantes en data analyse, un premier enjeu majeur de ce projet a été d'apporter une analyse pertinente et juste à ce que nous pouvions observer, interpréter les données de manière correcte sans pour autant être des expertes de l'univers du football. Un autre enjeu a été de rassembler les données que l'on considérait comme essentielles à ce projet, qui n'étaient pas toujours facilement accessibles.

Évidemment ce projet pourrait être complété et amélioré. Néanmoins, à notre niveau de débutantes et en tant que premier projet de notre future carrière, nous avons su relever plusieurs défis en travaillant ensemble, ce qui est essentiel en tant que data analyst.

Ce projet va représenter un atout majeur dans notre future carrière, puisqu'il s'agit du type de projet sur lequel nous pourrions être amenées à travailler à l'avenir. Nous pouvons d'ores et déjà, en sortie de formation, justifier d'un projet métier orienté de manière professionnelle, pour lequel nous avons été amenées à collaborer en équipe.
Il s'agit d'un travail complet, qui part d'une problématique à partir de laquelle des axes d'analyse ont été conçus, pour lesquels nous avons recherché de nouvelles données, notamment par webscraping lorsque les données n'étaient pas directement accessibles. Une fois les données obtenues nous avons réalisé des étapes de nettoyage des datasets, puis des premières visualisations pour évaluer leur pertinence, et de cette manière nous avons créé notre base de données avec différentes tables pouvant être reliées les unes aux autres pour réaliser des analyses et ainsi apporter des réponses à la problématique posée. La fin de ce projet se réalise sur power BI, un outil très recherché sur le marché de l'emploi.
En sommes, nous pouvons justifier d'une première expérience concrète sur l'ensemble des tâches liées au déroulé d'un projet dans le monde professionnel.

b. Niveau d'expertise

Pour la majorité d'entre nous, nous n'avions strictement aucune expertise sur le sujet. Certaines n'avaient même jamais regardé un seul match de football de toute leur vie.

Pour améliorer notre capacité d'analyse liée au sujet, nous avons été obligées de nous informer sur le sujet, sur internet ou auprès de proches fans de football. Nous avons appris de nombreuses choses, notamment concernant les règles liées à la participation des équipes aux différentes compétitions, la définition du budget (les différentes sources de gain d'argent et de dépenses), les subtilités liées aux transferts et à la possession ou prêts des joueurs, la composition des équipes et les stratégies associées.
La familiarisation avec les règles régissant cet univers était essentielle à la bonne interprétation de nos données.

Par exemple, il était important de savoir que la participation à la Champions League était réservée aux équipes les mieux classées dans chaque championnat national et qu'il y avait un nombre de places déterminées pour chaque ligue (selon certaines conditions bien précises) chaque année.
Malgré cela, tout n'est pas fixé puisque les premières équipes à pouvoir participer à l'Europa League ont une chance de pouvoir se faire "repêcher" dans la compétition de Champions League. C'est une information importante puisque dans le cas d'Arsenal, l'équipe n'a pas toujours participé à la Champions League, et elle a alors participé à l'Europa League. Nous aurions pu nous étonner de la voir participer à une seule de ces compétitions sans cette information, et émettre un jugement un peu rapide comme quoi le fait de ne pas participer aux deux compétitions justifiait d'un niveau insuffisant.
En réalité, cela signifie que les années où cette équipe n'a pas participé à la Champions League, elle a non seulement échoué à se qualifier, mais également échoué au repêchage. Mais en plus, que les années participées à la Champions League révèlent un niveau relativement élevé, tout du moins suffisant pour se qualifier à la plus grosse compétition de football. Il s'agit ici d'un exemple de la nécessité d'être renseigné sur le sujet sur lequel on travaille pour pouvoir analyser correctement les informations.

En conclusion, le fait d'appréhender les règles ainsi que les subtilités de cet univers a permis d'affiner notre analyse sur le sujet, et c'était une étape indispensable.

c. Difficultés rencontrées lors du projet

La principale difficulté rencontrée pour la réalisation de ce projet a été l'harmonisation des données récupérées par chaque membre de l'équipe mais aussi depuis différentes sources en ligne.

En effet, en récupérant des jeux de données depuis différentes sources (kaggle, transfermarkt, etc.), nous nous sommes rendues compte que le format (structure du texte, type, etc.), et la manière dont la Data était présentée n'était pas similaire, ce qui aurait pu engendrer un "blocage" au niveau des relations à créer entre les différents fichiers *.csv* qui servent comme référence dans notre analyse sur Power BI.
Le nettoyage et l'uniformisation des jeux de données entre eux ont donc été bien plus chronophages qu'attendu.

Ensuite, la récupération des jeux de données par webscraping, qui a été divisée entre nous trois, a été bien plus fastidieuse que ce à quoi nous pouvions nous attendre. Les difficultés ont été multiples : contourner les barrières des sites internets, récupérer des données dans un tableau de plusieurs pages par une interaction en javascript, boucler la récupération des données sur plusieurs URLs et obtenir des séries de même longueur lorsque certaines données sont absentes sur un URL, ou encore nettoyer et uniformiser les données obtenues suite au webscraping. Si ce module a été relativement court et simple dans la formation, l'application dans le monde réel a été infiniment plus complexe.

Il est évident que tout le temps passé à récupérer des données chacune de notre côté nous a empêché d'avancer sur d'autres points, comme la récupération d'autres données ou l'exploration d'autres axes d'analyse sur les datasets déjà obtenus. De fait, certains axes d'analyse que nous aurions trouvé intéressants n'ont pas été explorés par manque de temps pour envisager tout ce qu'il était possible de faire avec ce vaste projet.
Néanmoins, le football est un sujet complexe et très complet, alors il y a de très nombreuses orientations d'analyse possibles. C'est pourquoi nous avons dû faire un choix, et cibler celles qui nous semblaient les plus pertinentes et surtout les plus évidentes par rapport à la problématique posée, notamment concernant la performance de l'équipe.

Enfin, pour réaliser ce projet nous avons dû travailler sur plusieurs modules en avance, comme les modules de webscraping et de power BI. Nous avons également dû chercher des informations sur le text mining un peu avant que le module ne soit débloqué. Certains mois ont donc été plus intenses que d'autres.

II. Objectifs

Pour réaliser cette analyse sur les raisons qui pouvaient expliquer la dégradation du niveau du club de football Arsenal durant la dernière décennie, il y a eu différents objectifs à atteindre tout au long du projet.

Nous avons tout d'abord tenté d'observer des relations entre les données que nous possédions déjà sur le premier jeu de données fourni.
En effet, nous avons effectué différentes visualisations, dont certaines que nous vous présenterons ici, pour explorer le dataset en question (All Premier League Matches - 2010/2021).

A l'issue de cette toute première analyse, nous n'avons pas pu trouver de réponses évidentes et définitive quant à la dégradation de la performance du club Arsenal, au regard du fait que nous nous étions focalisées uniquement sur le club en question, sans prendre en compte son environnement ou les autres équipes.

Cependant la notion de "niveau" d'une équipe de football ne se résume pas qu'à un seul paramètre, mais une multitude. C'est pourquoi nous avons élaboré une méthode de réflexion pour guider notre analyse sur différents axes, notamment:

- Le placement d'Arsenal par rapport à d'autres équipes;
- Les attributs des différents matchs (buts, fautes commises, etc.);
- Les joueurs et leurs statistiques;
- Les transferts;
- Les compétitions gagnées;
- Les *managers* des équipes.

Pour cela, il a été nécessaire d'enrichir le projet par de nouveaux jeux de données permettant d'observer d'autres axes d'analyse que ceux permis par le premier dataset.

Plusieurs nouveaux jeux de données nous ont intéressées, à commencer par les données statistiques des joueurs, ce qui a représenté un gros défi. Il y a également eu une grosse étape de webscraping afin de récupérer d'autres données difficilement accessibles de manière complète en libre accès, comme les résultats des matchs en compétition, les budgets alloués aux transferts des joueurs et les informations sur les transferts de joueurs. Chaque site et chaque donnée à scraper a représenté ses propres difficultés et donc ses propres défis à relever.

Chaque nouveau jeu de données a subi une étape de "pré-analyse", avec différents questionnements pour nous permettre d'envisager la suite du projet en termes d'analyse et d'interprétation, et

éventuellement penser à récupérer d'autres données pour le compléter. Cela nous a également permis de vérifier la pertinence des nouveaux jeux de données obtenus. Nous allons vous détailler toutes ces étapes de manière plus complète dans la suite de ce rapport.

III. Projet

a. Data

Pour atteindre les objectifs du projet, nous avons rassemblé différents datasets au fur et à mesure de l'avancement du projet et de la définition des besoins liés à l'analyse, notamment les:

1. matchs du championnat anglais;
2. matchs des compétitions anglaises et européennes;
3. joueurs des équipes et leurs statistiques / équipes et ligues;
4. budgets des transferts par club;
5. budgets des transferts par joueurs;
6. vainqueurs des titres de championnat chaque année;
7. championnats auxquels ont participé les équipes par année;

Le contenu, la récupération, l'analyse ainsi que les difficultés rencontrées pour chaque dataset seront expliquées ci-dessous.

i. Dataset 1: Matchs du championnat anglais

■ Contenu du dataset

Le dataset "matchs_premier_league" contient des informations statistiques et des données de performance des équipes concernant l'ensemble des matchs du championnat national de l'English Premier League sur la dernière décennie (de 2010 à 2020).

Il y a des informations concernant les matchs :

- **date**, **year**, **season** : la date du match, l'année, la saison;
- **home_team** et **away_team** : l'équipe à domicile et l'équipe à l'extérieur;
- **home_score** et **home_score_ht**, **away_score** et **away_score_ht** : le score de l'équipe à domicile et le score de l'équipe à l'extérieur, pour la durée du match et pour la première période ("ht");
- **win**, **loss**, **draw_home** et **draw_away** : le nom de l'équipe gagnante, le nom de l'équipe perdante, le nom de l'équipe à domicile lors d'un match nul, le nom de l'équipe à l'extérieur lors d'un match nul;
- **arsenal_results** : le résultat du match concernant Arsenal;
- **result_full** et **result_ht** : score du match, pour la durée du match (full) et pour la première période (ht).
- **sg_match_ft** et **sg_match_ht** : différence entre le score de l'équipe à domicile et l'équipe à l'extérieur, pour la durée du match (ft) et pour la première période (ht).

Il y a aussi des indicateurs statistiques de performance des équipes (préfixe "home" pour l'équipe à domicile et "away" pour l'équipe à l'extérieur) :

- **clearances** : dégagements;
- **corners** : corners;
- **fouls_conceded** : fautes réalisées par l'équipe et concédées à l'équipe adverse;
- **offsides** : hors-jeux;
- **passes** : passes;

- **possession** : possessions de la balle pendant une durée du jeu par un joueur de l'équipe;
- **yellow_cards** : cartons jaunes;
- **red_cards** : cartons rouges;
- **shots** : tirs non cadrés;
- **shots_on_target** : tirs cadrés;
- **tackles** : tacles;
- **touches** : touches

Lorsque le suffixe "**avg_H**" ou "**avg_A**" est ajouté à l'une de ces statistiques (par exemple : clearances_avg_H), alors il s'agit de cette statistique moyennée sur l'ensemble des matchs de la période depuis 2010, pour chaque équipe et selon où elle a joué (à domicile pour H, à l'extérieur pour A).

Lorsque le suffixe "**avg_home**" ou "**avg_away**" est ajouté à l'une de ces statistiques (par exemple : clearances_avg_home), alors il s'agit de cette statistique moyennée sur l'ensemble des matchs de la saison, pour chaque équipe et selon où elle a joué (à domicile pour home, à l'extérieur pour away).

Enfin, il y a des indicateurs de performance par saison :

- **sg_match_ft_acum_home** et **sg_match_ht_acum_home** : différence entre le score de l'équipe à domicile et l'équipe à l'extérieur, cumulé sur l'ensemble des matchs, pour la durée du match et pour la première période.
- **goals_scored_ft_avg_home** et **goals_scored_ht_avg_home** : moyenne du nombre de buts marqués par saison pour chaque équipe, pour la durée du match et pour la première période.
- **goals_conced_ft_avg_home** et **goals_conced_ht_avg_home** : moyenne du nombre de buts encaissés par saison pour chaque équipe, pour la durée du match et pour la première période.

■ Récupération et difficultés

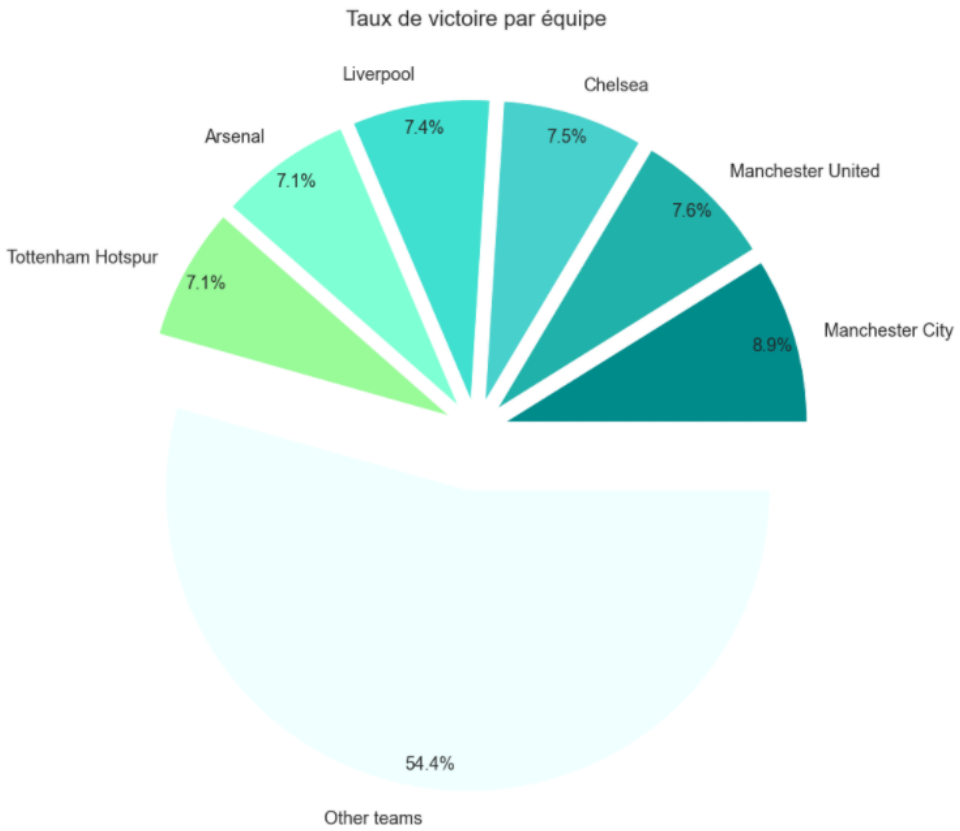
Nous avions ce dataset dès le départ, il était libre d'accès (https://www.kaggle.com/pablohfreitas/all-premier-league-matches-20102021/metadata), et contenait des données issues du site https://www.premierleague.com/, donc directement du site de la ligue anglaise.

La compréhension de toutes les colonnes de statistiques n'a pas été immédiate et a au départ posé quelques problèmes. Il a notamment fallu trouver une méthode pour remplacer les valeurs manquantes dans ces colonnes, et sans leur bonne compréhension nous risquions de ne pas utiliser de méthode adéquate. Ce dataset n'a pas posé de problème majeur en dehors de cela, il a surtout s'agit de nettoyage et de mise en forme :

- suppression de colonnes inutiles ou que l'on ne souhaitait pas utiliser;
- suppression de la saison 2020/2021 qui était incomplète et risquait de fausser nos analyses;
- remplacement des valeurs manquantes;
- renommer des colonnes pour uniformiser avec les autres datasets;
- remplacer le point par une virgule pour les valeurs décimales;
- définir des fonctions pour récupérer le nom de l'équipe selon la circonstance cherchée et le stocker dans une colonne (l'équipe gagnante, l'équipe perdante etc.).

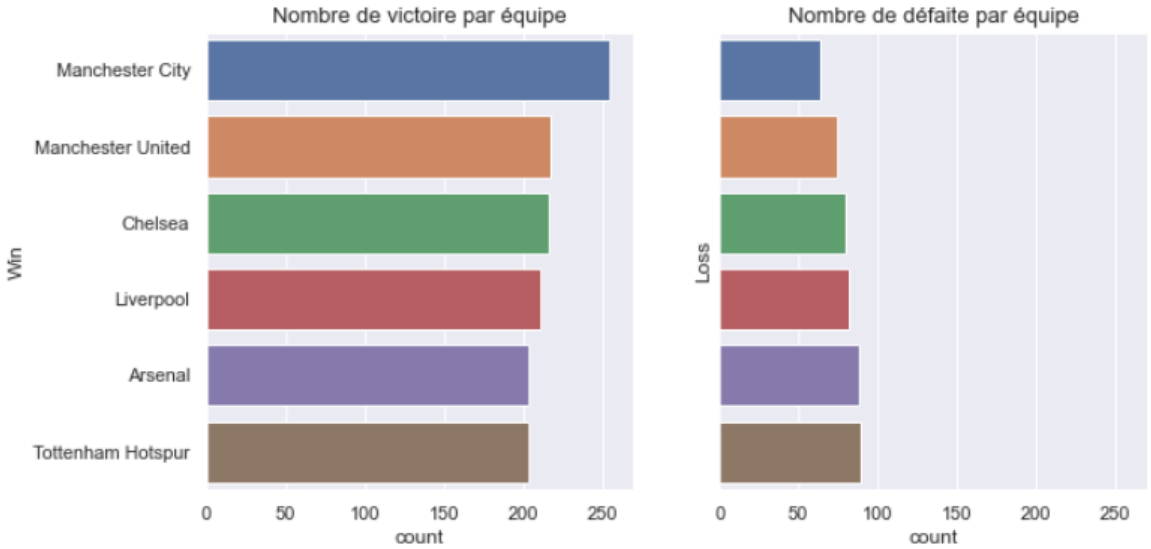
■ Analyse

La première question que nous nous sommes posées, et dont la suite de l'analyse a découlé, c'était de savoir quelles équipes de l'English Premier League gagnaient le plus de matchs.



Sur l'ensemble de la décennie, on remarque que 6 équipes sont en tête avec plus de 7% de victoire chacunes, elles totalisent dont presque la moitié de l'ensemble des victoires à elles seules.

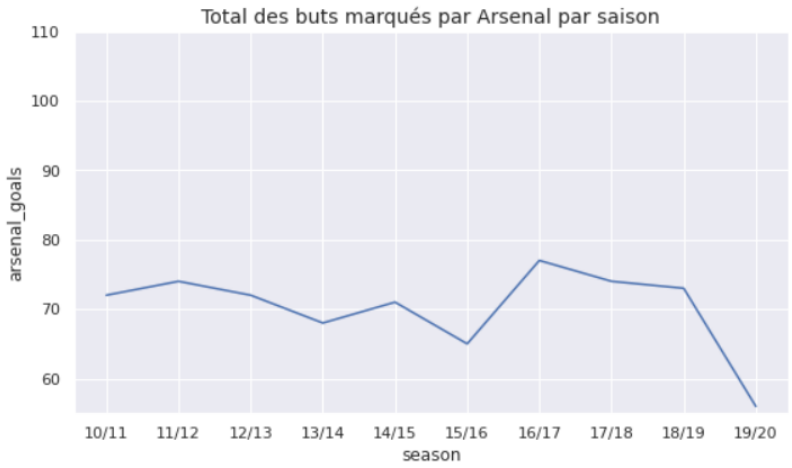
Si on regarde plus en détail le nombre de matchs gagnés et perdus pour ces 6 équipes, on se rend compte qu'il y a un véritable classement : plus une équipe remporte de matchs, moins elle en perd.

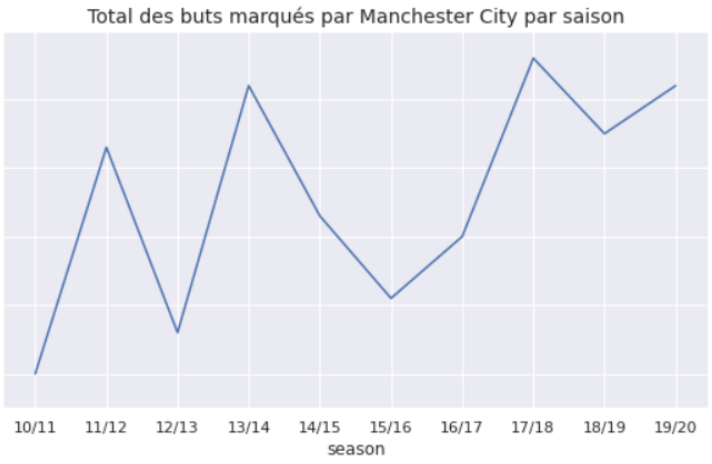


Arsenal faisant partie de ce "top 6" des équipes réalisant le plus de victoires durant la dernière décennie en ligue anglaise, elle fait donc bien partie des meilleures équipes. Néanmoins ici il s'agit d'une observation sur la totalité des matchs de ligue anglaise de la décennie.

Pour une observation au fil des saisons, nous avons choisi de comparer les performances d'Arsenal avec celles de Manchester City, qui ressortait dans le graphique ci-dessus comme l'équipe réalisant le plus de victoires et le moins de défaites.
Nous avons observé plusieurs indicateurs, notamment le nombre de buts marqués, sur lequel nous pouvons voir une tendance différente pour les deux équipes au fil des saisons : de plus en plus de

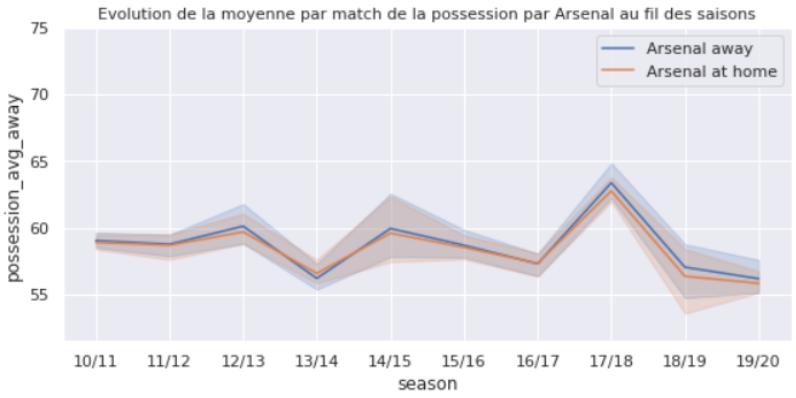
buts marqués pour Manchester City et un équilibre pour Arsenal, ce qui témoigne d'une absence de progression en attaque au fil des années.

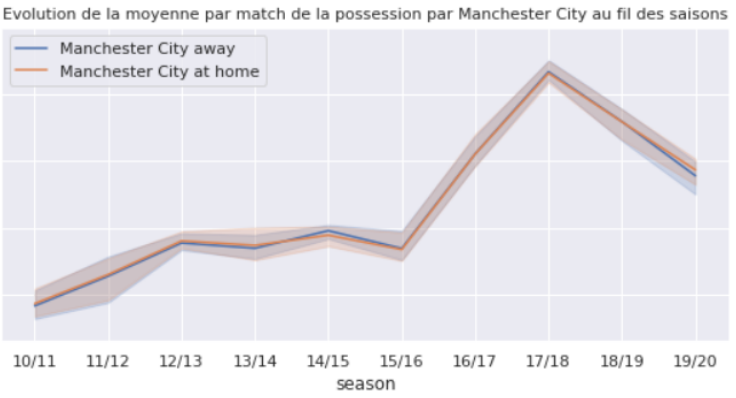




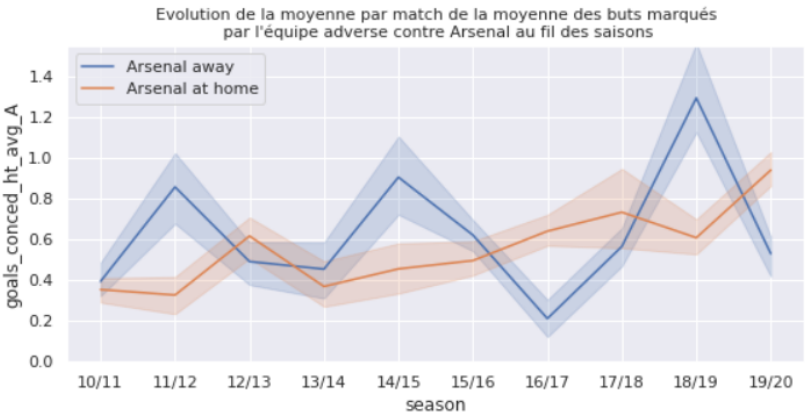
Néanmoins nous savons qu'au football le fait de jouer à domicile ou à l'extérieur peut avoir un grand impact sur les résultats d'une équipe, alors nous avons tenu compte de ce facteur pour comparer différents indicateurs de performance entre ceux deux équipes, comme :

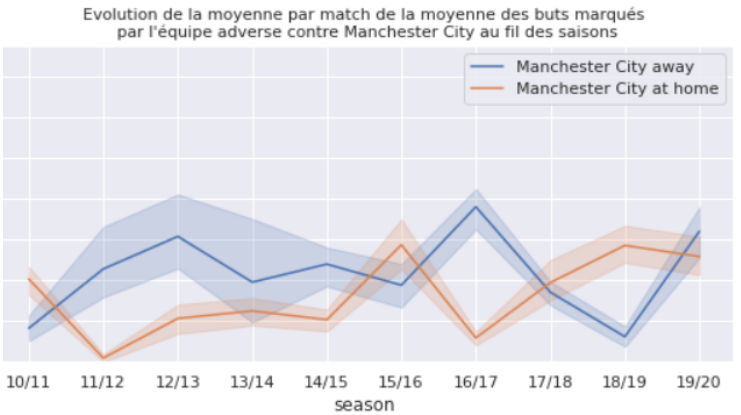
- La moyenne par match de la possession au fil des saisons. On remarque ici une absence d'impact du lieu du match pour les deux équipes, en revanche on voit une augmentation de la possession au fil des saisons pour Manchester City, contrairement à Arsenal.



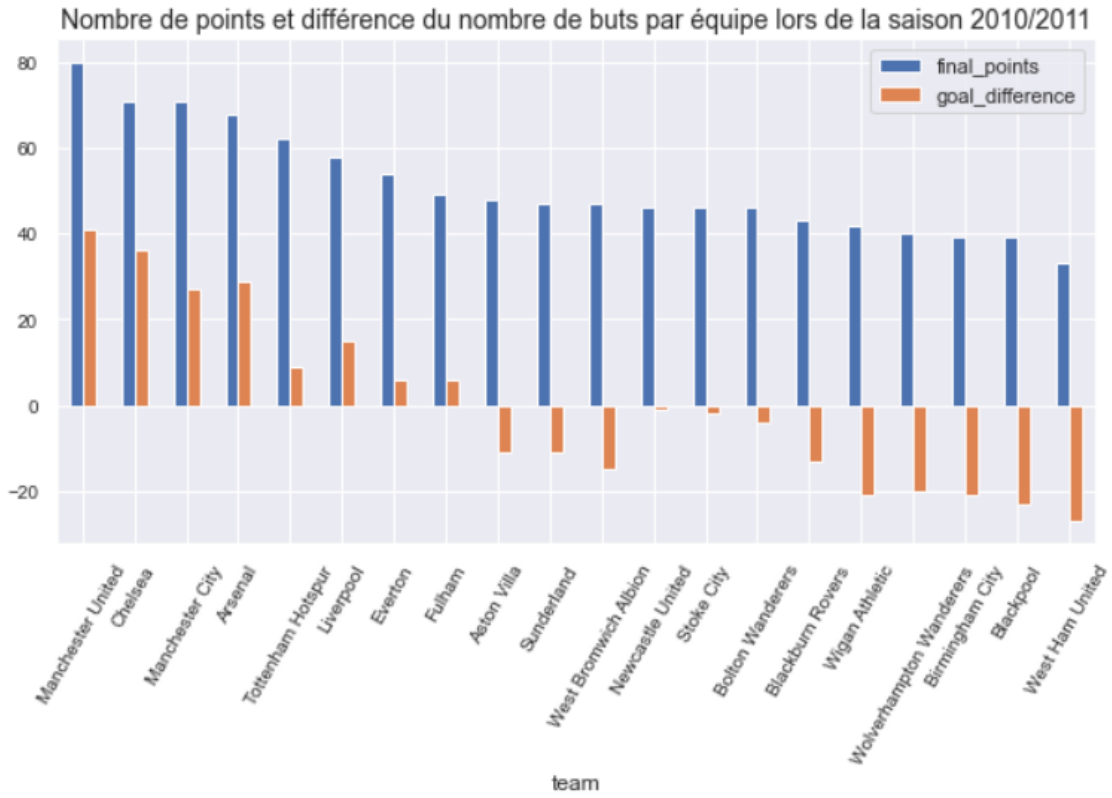


- La moyenne des buts encaissés par match au fil des saisons. Ici on remarque que non seulement le lieu du match a un impact (globalement pour ces deux équipes il y a moins de buts encaissés lorsqu'elles jouent à domicile), mais qu'en plus pour la plupart des saisons, Arsenal encaisse plus de buts que Manchester City, ce qui témoigne d'une perte d'efficacité en défense.





En parallèle de cela, nous avons calculé le nombre de points marqués pour chaque équipe à chaque saison : 3 points lors d'un match gagné, 1 point par match nul, 0 point par match perdu. Le classement lors de la compétition de ligue se basant sur ce nombre de points, nous avons donc pu "recréer" ce classement à chaque saison et le comparer à la différence de score pour chaque équipe (différence entre le nombre de buts marqués et le nombre de buts encaissés).
Voici un exemple avec la saison 2010/2011, dans laquelle on retrouve toujours les 6 mêmes équipes en tête du classement.



En conclusion, nous avons trouvé ce dataset pertinent pour effectuer des comparaisons des performances des équipes, saison après saison ou au global, à domicile ou à l'extérieur, grâce à divers indicateurs liés aux déroulement des matchs.

ii. Dataset 2: Matchs des compétitions européennes et anglaises

■ Contenu du dataset

Le dataset "matchs_competitions" contient les informations relatives aux matchs des compétitions nationales anglaises (FA Cup et EFL Cup) et européennes (Champions League et Europa League) sur la dernière décennie (de 2009 à 2019) :

- **competition** et **phase** : la compétition et la phase (groupe ou éliminatoire).
- **date**, **year**, **season** : la date du match, l'année, la saison.
- **home_team** et **away_team** : l'équipe à domicile et l'équipe à l'extérieur.
- **home_score** et **away_score** : le score de l'équipe à domicile et le score de l'équipe à l'extérieur.
- **win**, **loss**, **draw_home** et **draw_away** : le nom de l'équipe gagnante, le nom de l'équipe perdante, le nom de l'équipe à domicile lors d'un match nul, le nom de l'équipe à l'extérieur lors d'un match nul.
- **arsenal_results** : le résultat du match concernant Arsenal.

Récupération

Il est possible de trouver des datasets des matchs de la compétition Champions League, qui est la plus grosse compétition de football, mais pour les autres compétitions les données sont difficiles à trouver.

Le site https://fr.besoccer.com/ rassemble de très nombreuses informations sur le football, et en particulier les résultats des matchs de toutes les compétitions de football qui existent, pour l'ensemble des saisons sur lesquelles elles se sont déroulées. C'était le site idéal pour webscraper les données pour l'ensemble des compétitions qui nous intéressaient.
Ce site a été au départ relativement facile à scraper puisque les balises du code html sont très claires et il a suffit d'une boucle sur un ensemble d'URLs en changeant l'année ou le numéro de la journée pour récupérer les données.

Ce qui a compliqué cette étape, c'est la présence de matchs annulés et reportés. En effet, lorsqu'un match est annulé, à la place du score se trouve l'heure du match (16:00). Mais surtout, la balise du score n'existe pas pour ce match, elle est remplacée par une autre balise. De fait, la liste de score était plus courte que les autres listes de données, empêchant la création du DataFrame.
Pour résoudre ce problème, il a fallu obtenir une liste des données de score de la même longueur que les autres listes de données, et donc sélectionner une balise en amont qui existe même pour les lignes de matchs annulés, puis ensuite créer une fonction pour nettoyer la liste de score. C'est-à-dire, retrouver les matchs annulés grâce à leur écriture particulière entre parenthèses (16:00). Aucune autre valeur de score n'étant indiquée entre parenthèses, nous pouvons donc ôter les parenthèses, et supprimer la valeur 16:00 lors du nettoyage du DataFrame.
Le fait de sélectionner une balise en amont a inséré des espaces dans certains scores, nous profitons de la fonction précédente pour les éliminer.

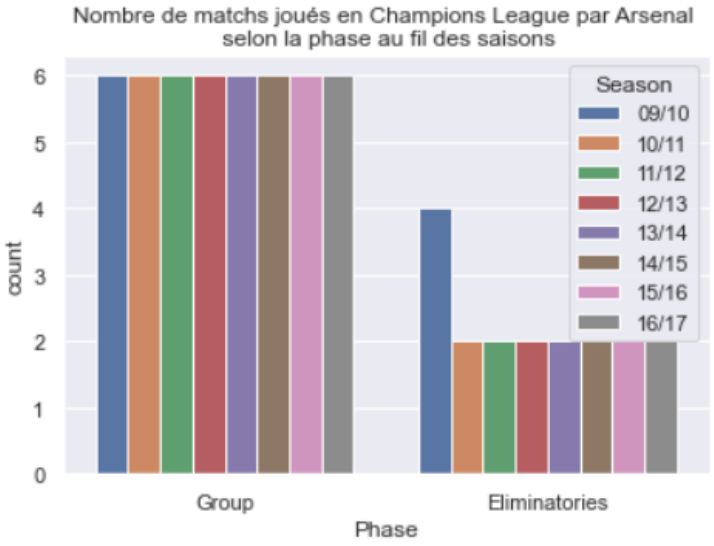
Enfin, nous souhaitions avoir une distinction des phases éliminatoires et des phases de groupe. Pour la Champions League et l'Europa League, les URLs de base étant différents, il a suffit de reproduire le même webscraping sur les deux URLs pour rassembler les DataFrames obtenus. Pour les compétitions de FA Cup et d'EFL Cup, ce n'était pas le cas. La logique a voulu que l'on scrape les données sur les journées de phase de groupe et de phase éliminatoires puis qu'on rassemble le tout. Cependant, pour la FA Cup il n'y a pas le même nombre de jours dans chaque phase pour toutes les saisons. Nous avons donc créé une fonction pour scraper les données à partir d'un URL donné, et des numéros des jours concernés par la phase.

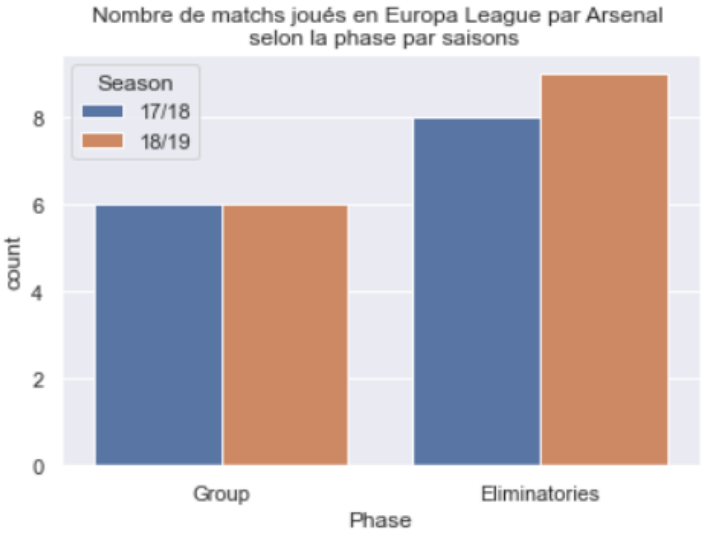
Ensuite il a fallu nettoyer et harmoniser tous ces DataFrames avant de les assembler : mettre en forme la date, harmoniser les noms des équipes par rapport au DataFrame initial, ajouter des colonnes pour les saisons, séparer le score en deux colonnes, etc.

La dernière étape a simplement consisté à assembler les DataFrames de ces 4 compétitions ensemble, puis ajouter les mêmes colonnes que nous avions ajouté dans le DataFrame initial en réutilisant les fonctions que nous avions créés, pour obtenir de nouvelles colonnes : les noms des équipes gagnantes, perdantes, à domicile et à l'extérieur lors d'un match nul, et une colonne pour le résultat du match concernant Arsenal quand l'équipe à joué.

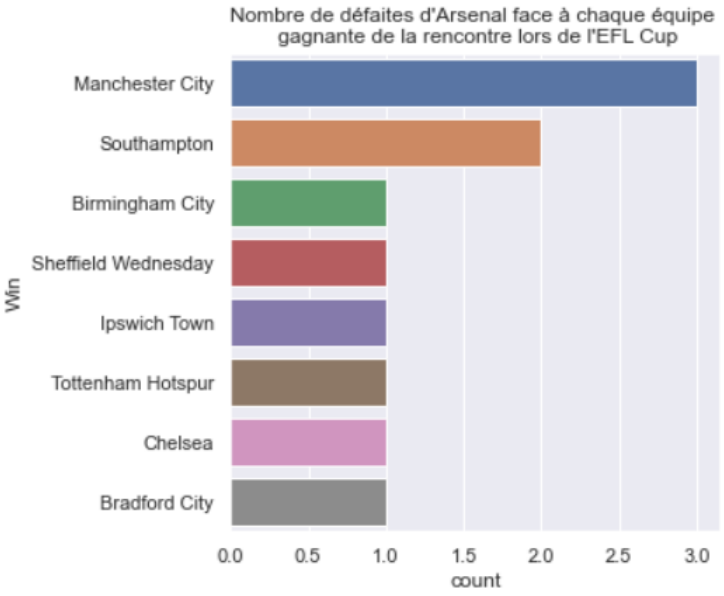
Analyse

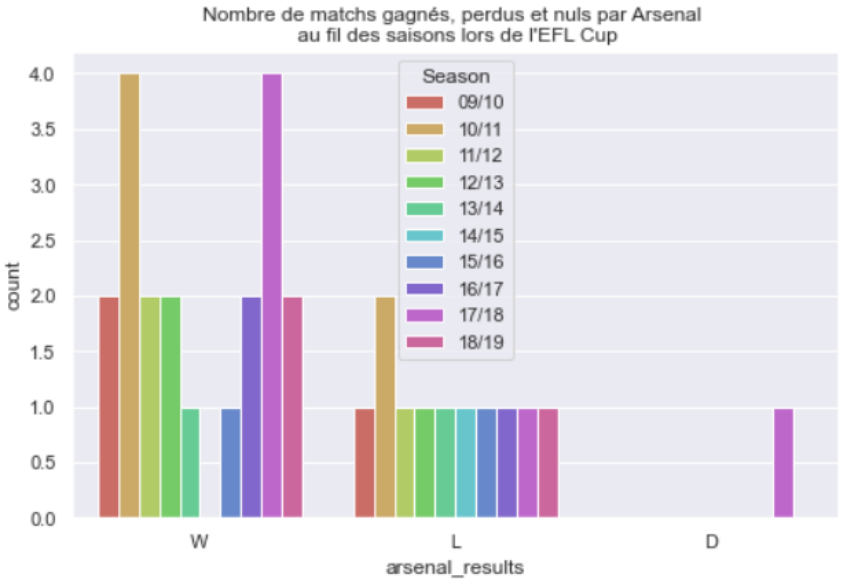
La première question posée ici a été de savoir si Arsenal avait participé à ces compétitions, et si oui quelles années ?
On voit ici que sur la décennie, l'équipe a participé tous les ans à la Champions League, excepté les deux dernières saisons, pour lesquelles ils n'ont pas réussi à se qualifier. On voit également que pour la Champions League, l'équipe a des résultats constants et arrive toujours au même stade de la compétition avant d'être éliminée (8ème de finale). On constate la même chose en Europa League avec une participation jusqu'en demi-finale une année, puis en finale l'année suivante.



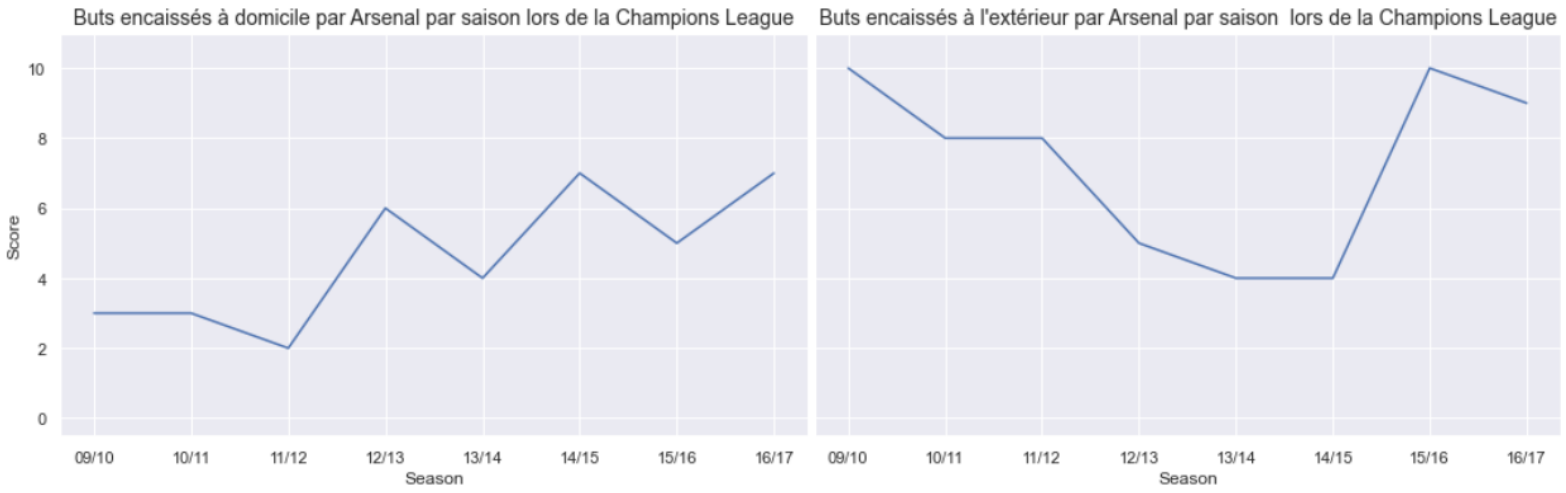


Ensuite nous nous sommes questionnées sur les adversaires face auxquels Arsenal avait perdu des matchs. Ici on voit que pour la compétition EFL Cup, Arsenal a totalisé 3 défaites contre Manchester City.
C'est bien sûr à mettre en relation avec le nombre de matchs joués, puisque comme on peut le voir, il n'y a pas beaucoup de matchs par saison, c'est une petite compétition.





Par la suite, nous nous sommes questionnées sur les performances de l'équipe au fil des saisons, notamment sur l'évolution du nombre de buts marqués ou encaissés au fil des saisons, en gardant une comparaison avec le lieu du match (domicile ou extérieur). On peut voir qu'en Champions League, Arsenal encaisse de plus en plus de buts au fil des saisons lorsqu'elle joue à domicile, alors qu'à l'extérieur c'est plus variable.



En conclusion, ce dataset est pertinent puisqu'il fournit des informations suffisamment complètes pour avoir une base de comparaison des résultats des équipes en compétitions, saison après saison ou au global, à domicile ou à l'extérieur.

iii. Dataset 3: Joueurs et leurs statistiques / équipes et ligues

■ Contenu des datasets

Le dataset **"teams_and_leagues"** contient :

- **club** : les noms des équipes
- **league** : les ligues correspondantes
- **country** : les pays

Le dataset **"total_players"** contient des informations et statistiques sur les joueurs des différentes équipes pour chaque saison durant la dernière décennie.

Ce dataset contient des données informatives sur les joueurs :

- **player_fifa_api_id** : identifiant FIFA du joueur;
- **player_name** : nom du joueur;
- **birthday** : date de naissance;
- **season** : saison concernée;
- **age** : âge du joueur lors de la saison;
- **position** : poste sur le terrain durant la saison concernée;
- **club** : équipe à laquelle appartient le joueur durant la saison concernée;
- **height_cm** : taille en cm;
- **value_eur** : valeur du joueur en euros;
- **wage_eur** : salaire du joueur en euros;
- **preferred_foot** : gaucher ou droitier.

Ce dataset contient aussi des données statistiques concernant les performances des joueurs :

- **overall** : note globale des performances du joueur sur 100. Calculée entre autres à partir des statistiques d'attaque (attacking), de défense (defending), de talents (skills), de tir (shooting), de puissance (power), de mentalité (mentality), de déplacement (movement) et de gardien de but (goalkeeping).
- **potential** : note globale du potentiel sur 100 (estimation avant le début de la saison en fonction des performances à la saison précédente).

- **attacking_crossing**, **attacking_finishing**, **attacking_heading_accuracy**, **attacking_short_passing**, **attacking_volleys** : notes sur 100 des différentes compétences d'attaque.
- **skill_dribbling**, **skill_curve**, **skill_fk_accuracy**, **skill_long_passing**, **skill_ball_control** : notes sur 100 des différentes compétences de talents.
- **movement_acceleration**, **movement_sprint_speed**, **movement_agility**, **movement_reactions**, **movement_balance** : notes sur 100 des différentes compétences de déplacement.
- **power_shot_power**, **power_jumping**, **power_stamina**, **power_strength**, **power_long_shots** : notes sur 100 des différentes compétences "d'énergie".
- **mentality_aggression**, **mentality_interceptions**, **mentality_positioning**, **mentality_vision**, **mentality_penalties** : notes sur 100 des différentes compétences de mentalité.
- **defending_marking**, **defending_standing_tackle**, **defending_sliding_tackle**: notes sur 100 des différentes compétences de défense.
- **goalkeeping_diving**, **goalkeeping_handling**, **goalkeeping_kicking**, **goalkeeping_positioning**, **goalkeeping_reflexes** : notes sur 100 des différentes compétences de gardien de but.

Récupération

Le dataset "total_players" a probablement été le plus fastidieux à récupérer de tout le projet. Il est issu de la combinaison de deux datasets différents.

Ces datasets ont été conçus à partir des données du site https://sofifa.com/, un site qui regroupe la totalité des données des jeux Fifa en mode carrière. Il est important de savoir que pour obtenir ces statistiques sur les joueurs, EA Sports crée elle-même sa propre base de données à partir d'observations des matchs en temps réel sur place par plus de 8000 volontaires. Evidemment ces données sont corrigées et ajustées en interne (notamment selon les championnats). L'ensemble du détail du calcul des différentes notes est gardé secret par l'entreprise même si les notes attribuées au différentes statistiques se veulent tout de même refléter la réalité puisqu'issues d'observations en réel. De fait, nous avons considéré ces données comme étant fiables. D'autant qu'il est difficile de trouver des sources complètes avec autant de statistiques sur les joueurs de foot, et que les données issues de Fifa sont très utilisées dans des projets d'analyse.

Le premier dataset de joueurs Fifa a été obtenu à partir de plusieurs fichiers csv de joueurs (un fichier par saison) issus de Kaggle : https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv.
Il a suffit d'ajouter une colonne par fichier pour identifier la saison, puis concaténer les différentes tables ensemble. Il contient les données des joueurs pour une période allant de 2014 à 2020.

Pour compléter ce premier dataset de joueurs Fifa, nous avions songé à scraper directement les données du site sofifa pour les joueurs des équipes qui nous intéressaient pour la période qui nous manquait (2009 à 2014). Cependant, quelle que soit la méthode utilisée, il semblerait qu'il y ait une interaction (peut-être en Javascript) sur la page, ce qui fait qu'avec n'importe quel URL les données scrapées étaient celles de l'année en cours. La tentative de webscraping ayant été un échec, nous nous sommes rabattues sur une base de données SQL que nous avions trouvée entre temps.

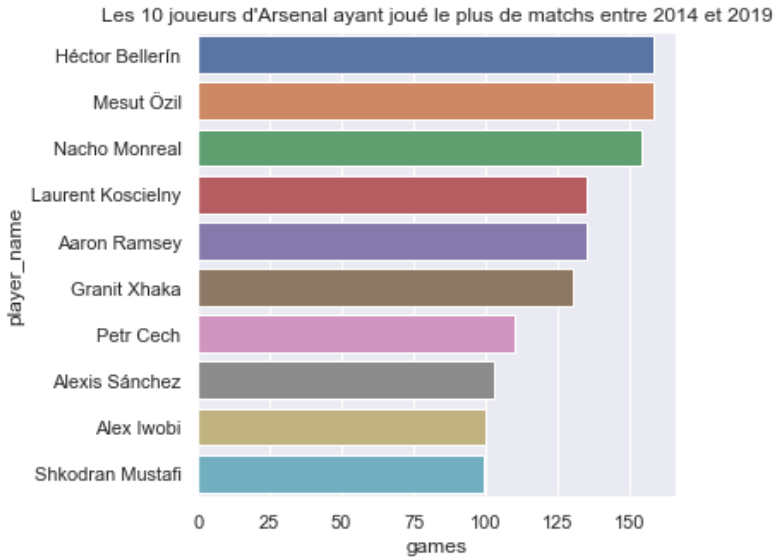
Le second dataset de joueurs Fifa a donc été obtenu grâce à la combinaison de différentes tables d'une base de données SQL issue de Kaggle : https://www.kaggle.com/hugomathien/soccer. Il contient les données des joueurs pour une période allant de 2009 à 2014.
Pour pouvoir utiliser cette base de données, il a fallu apprendre un peu de SQL alors que nous n'avions pas encore vu le module, pour réussir à se connecter à la base, récupérer les données de

chaque table et les enregistrer dans un fichier csv avec lequel nous savions comment interagir ensuite.

Initialement, avant de travailler avec les données Fifa pour les joueurs, nous avions un autre dataset avec des statistiques différentes pour les joueurs de l'English Premier League, issu de kaggle (https://www.kaggle.com/abrarhossainhimself/understat-data-for-teams-players-2014-present), dont les données étaient tirées du site https://understat.com/. C'est un site avec des statistiques analytiques des joueurs des ligues européennes. Ce dataset était intéressant, puisque certaines statistiques avaient été obtenues grâce à un modèle de machine learning prédictif conçu et entraîné par les créateurs du site, pour avoir des données de performance plus indicatives que simplement le nombre de buts marqués ou le score. Il s'agit de statistiques prédictives basées sur les buts attendus, une statistique considérée comme souvent plus fiable que le nombre de buts réels ou les passes décisives.
Néanmoins ce dataset était également incomplet puisque les données mises à disposition ne débutaient qu'en 2014, et qu'il ne pouvait pas être complété puisque le modèle de machine learning utilisé (propre à ce site) n'était pas libre d'accès. Enfin, nous n'avons pas pu fusionner ce dataset avec celui des données Fifa (voir la partie "Difficultés"), alors nous avons fait le choix de garder uniquement le dataset le plus complet.

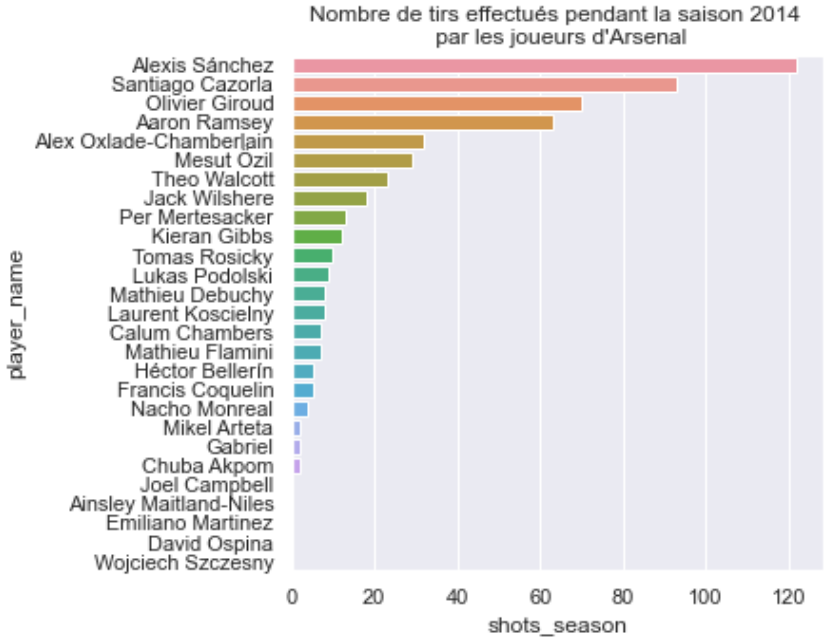
Analyse

Avant d'abandonner le dataset Understat Players, nous avons effectué quelques visualisations dessus.
La première question a été de savoir quels joueurs chez Arsenal, sur la période entière du dataset (2014 à 2020), avaient joué le plus de matchs et donc affecté le plus les résultats de l'équipe.

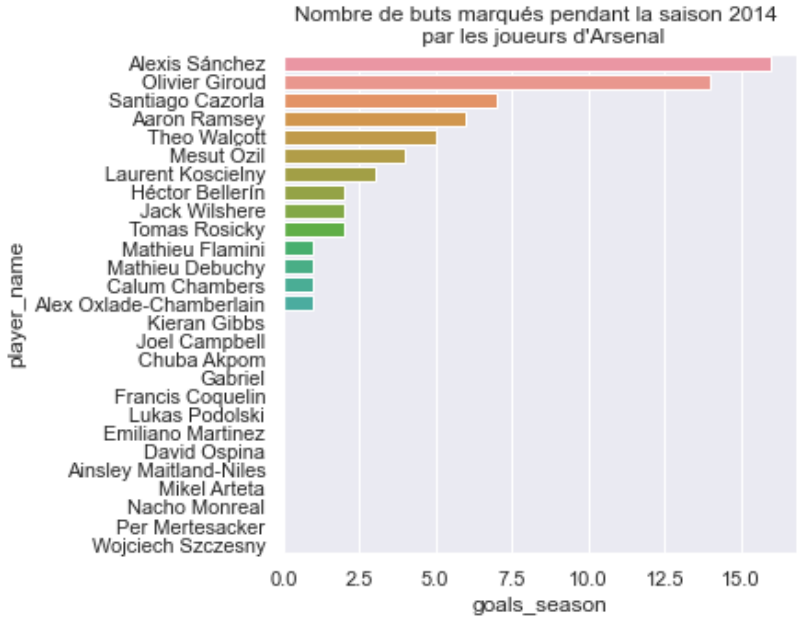


Évidemment ces joueurs sont ceux qui ont joué plusieurs saisons dans l'équipe Arsenal. Le but ici était d'ensuite pouvoir observer les performances de ces joueurs, au fil du temps et selon leur participation à chaque saison.
Ensuite nous nous sommes intéressées aux performances en attaque des joueurs de l'équipe, à savoir quels étaient les meilleurs attaquants et buteurs chez Arsenal.
Ici une extraction pour la saison 2014 de deux ensembles de graphes, détaillant :

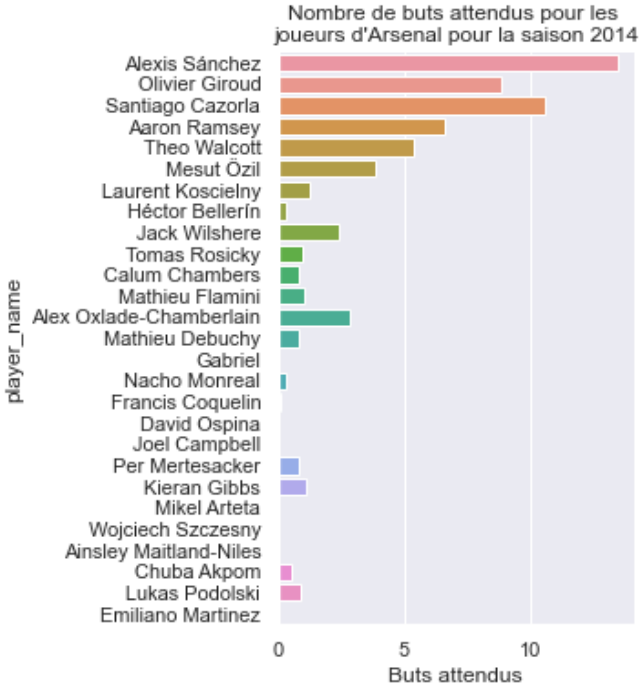
- le nombre de tirs effectués pour chaque joueur de l'équipe Arsenal



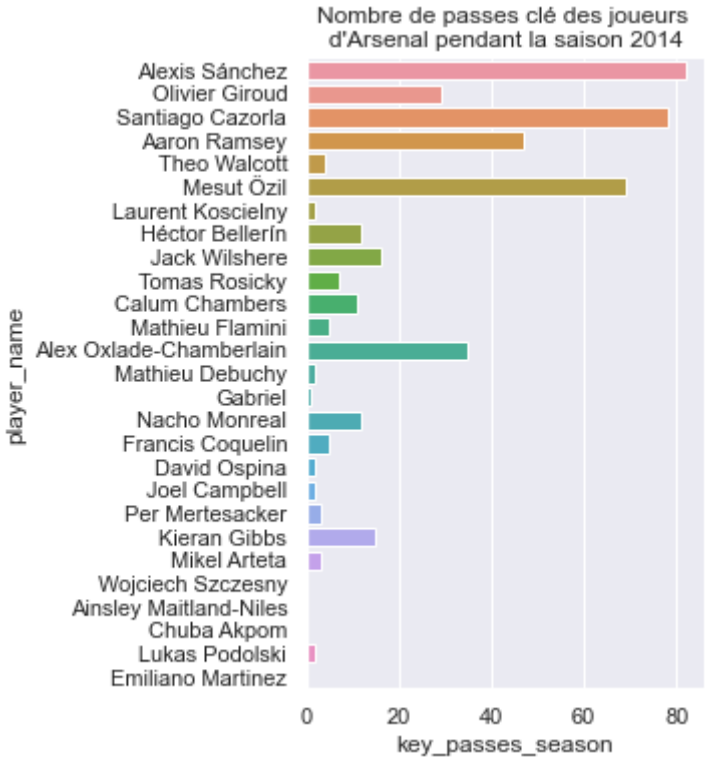
- le nombre de buts marqués pour chaque joueur de l'équipe Arsenal



- les buts attendus pour chaque joueur de l'équipe Arsenal



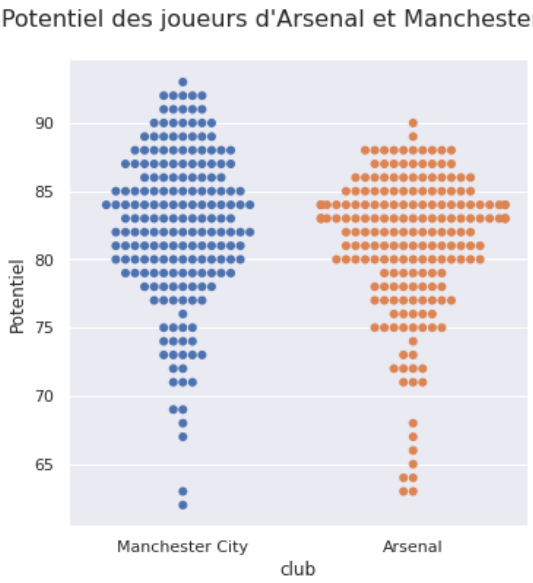
- les passes clé réalisées pour chaque joueur de l'équipe Arsenal



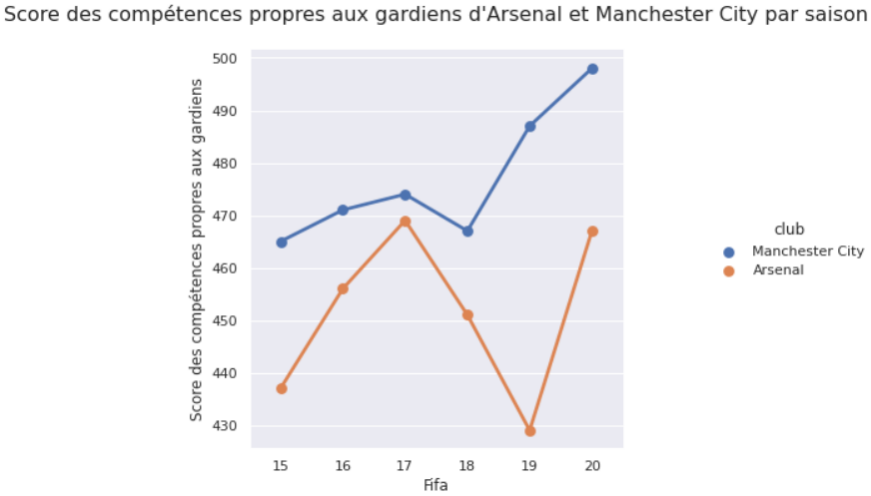
Grâce à nos analyses sur ce dataset, nous en avons surtout appris sur la composition de l'équipe au fil du temps, et nous avons pu identifier certains joueurs plus performants dans leur rôle à un moment donné.

Enfin, nous avons également réalisé quelques analyses sur le dataset Fifa (celui que nous avons conservé dans la suite du projet).
Ici il a s'agit d'essayer de mettre en lien les analyses comparatives des performances d'Arsenal et de Manchester City lors des matchs de l'English Premier League faites sur le premier dataset, avec la composition de ces deux équipes.
Nous avons alors comparé le potentiel de la totalité des joueurs de ces 2 équipes entre 2014 et 2020 sous la forme d'un nuage de points, chacun de ces points représentant un joueur. On observe que

Manchester City a un pool d'excellents joueurs, c'est-à-dire avec un potentiel supérieur à 90, alors qu'Arsenal n'en a pas.



Nous avons ensuite souhaité savoir s'il y avait des différences de niveau entre les joueurs selon les postes dans l'équipe, alors nous avons comparé les scores propres à certains postes, comme ici les scores de compétence en garde de but pour les gardiens de but. On se rend compte ici qu'au fil des saisons, Manchester City a systématiquement des gardiens ayant de meilleures compétences à leur poste que ceux d'Arsenal.



Difficultés

Certaines difficultés rencontrées ont déjà été abordées auparavant. Néanmoins, ce ne sont pas les seules.

Une problématique majeure liée à ce dataset a concerné les données des postes des joueurs.
Pour le second dataset de joueurs issu de fifa, les positions des joueurs sur le terrain n'étaient pas clairement indiquées. En revanche, pour chaque match, il y avait 11 colonnes "player" numérotées de 1 à 11, contenant chacune l'identifiant d'un joueur pour le match en question. Nous avons alors essayé d'approximer la position de chaque joueur en fonction de ce numéro (par exemple, le numéro 1 pour le gardien de but).
Pour cela, nous avons isolé dans de nouveaux DataFrames tous les identifiants des joueurs d'une même colonne, auquel nous avons ajouté une colonne indiquant la position en question. Par exemple, pour la colonne 1, nous avons associé à l'identifiant le poste 'GK' pour goalkeeper (gardien de but). Cependant, les numéros des colonnes ne correspondaient pas toujours tout à fait au poste et

nous nous sommes retrouvées avec un dataset avec des postes extrêmement différents pour beaucoup de joueurs, parfois pendant une même saison (par exemple, un attaquant qui devient défenseur arrière dans la même équipe pendant la même saison, puis redevient attaquant par la suite). Le dataset étant conséquent, corriger à la main aurait été bien trop laborieux alors nous avons décidé de garder des valeurs manquantes pour les postes des joueurs de 2009 à 2014 plutôt qu'une mauvaise approximation qui risquait d'induire trop d'erreurs dans l'analyse.

Enfin, la principale difficulté que nous avons rencontrée pour la réalisation de ce projet, et que nous n'avons pas pu dépasser, a été de trouver une méthode en text mining pour uniformiser les noms des joueurs entre différents datasets (d'ailleurs la problématique s'est posée également pour les noms des équipes).
En effet, pour la période de 2014 à 2020, nous disposions d'un autre dataset (Understat Players) avec d'autres statistiques complémentaires à celles que nous avions déjà. Cependant, les noms des joueurs étaient inscrits de manière différente et nous n'avons pas su uniformiser ces données pour réunir les deux DataFrames. De fait, nous avons dû abandonner cet autre dataset alors que nous avions fait plusieurs analyses avec différentes visualisations.
Nous savons que pour la suite du projet, le fait de n'avoir pas su résoudre cette problématique pourra nous poser des problèmes, notamment pour les liens entre les tables. Nous avons un autre dataset qui contient des noms de joueurs, pas forcément écrit de la même manière, et plusieurs datasets avec des noms de clubs issus de sources différentes avec différentes écritures pour un même club. Nous avons uniformisé à la main les noms des clubs composant la ligue anglaise, néanmoins pour les clubs des autres ligues cela aurait été laborieux. Nous espérons que ce point ne présentera pas un blocage pour la suite du projet.

En conclusion, bien que le dataset Understat Players puisse avoir apporté certains axes d'analyse intéressants, nous lui avons préféré le dataset de joueurs Fifa qui contenait plus de statistiques, plus variées et surtout pour l'ensemble des postes et était moins orienté sur l'attaque. Mais aussi, ce dataset était plus complet puisqu'il couvrait la période de nos autres données, c'est-à-dire de 2009 à 2020.
Le dataset total_players nous semble pertinent puisqu'il nous fournit une excellente base pour comparer les compétences des joueurs des différentes équipes, en fonction de leur poste, de la saison, permettant ainsi des analyses variées.

iv. Dataset 4: budgets des transferts par club

■ Contenu du dataset

Le dataset #4 reprend l'ensemble du budget des transferts par club, depuis la saison 2009/2010 à la saison 2021/2022, et contient les informations suivantes:

- Club;
- Expenditure (dépenses en euros);
- Arrivals (nombre de joueurs arrivés au club);
- Income (revenus en euros);
- Departures (nombre de joueurs qui ont quitté le club);
- Balance (budget);
- Balance computed *(= Expenditure - Income).*

Un autre dataset complète le précédent en détaillant le budget utilisé par année pour les transferts sur les club de la Premier League et contient les informations suivantes :

- *Année;*
- *Club;*
- *Montant de l'acquisition de nouveaux joueurs*
- *Montant des départs des joueurs*

■ Récupération

Nous avons pu récupérer ces informations par le biais du **Web Scraping** depuis le site *transfermarkt* qui contient des informations liées aux équipes de foot, sur différents aspects, notamment les:

- Budgets;
- Transferts;
- Contrats;
- Valeur des équipes;
- Compétitions;
- Statistiques (par joueur et/ou équipe)..

Le détail des transferts par année a aussi été récupéré par Web Scraping sur le même site.
En effet, ce site web nous a permis de trouver des informations utiles liées à notre projet, qui nous ont permis d'avancer avec des datasets différents, sur des aspects divers et surtout au niveau des filtres et de la flexibilité que le site offre en ce sens.

■ Analyse

A première vue, l'équipe Arsenal semble bien être dans les Top 10 du classement quant aux dépenses. En effet, depuis 2009, l'équipe a dépensé aux alentours de 1,1 milliard d'euros sur l'achat de nouveaux joueurs, avec 234 acquisitions.
En comparaison avec Manchester City, l'équipe placée au premier rang quant aux dépenses, il existe une différence importante avec Arsenal entre le montant de dépenses ainsi que le nombre de joueurs acquis par l'équipe, avec 1,9 milliard d'euros investis par Manchester City, et 319 arrivées.

D'autre part, si nous comparons les revenus et les départs, il s'avère qu'Arsenal ne semble pas être capable de gagner autant d'argent par rapport à d'autres équipe en lâchant leurs joueurs, avec un revenu total de 525 millions d'euros, pour 236 départs contre, par exemple, 1,1 milliards d'euros de revenus pour Chelsea FC pour 374 départs.

	Club	Expenditure	Arrivals	Income	Departures	Balance	balance_computed
0	Manchester City	1,970,000,000.00	319	626,610,000.00	319	-1,343,700,000.00	-1,343,390,000.00
1	Chelsea FC	1,740,000,000.00	378	1,110,000,000.00	374	-627,590,000.00	-630,000,000.00
2	FC Barcelona	1,670,000,000.00	170	1,050,000,000.00	163	-616,270,000.00	-620,000,000.00
3	Juventus FC	1,570,000,000.00	608	1,050,000,000.00	597	-526,330,000.00	-520,000,000.00
4	Manchester United	1,550,000,000.00	208	548,480,000.00	214	-1,000,060,000.00	-1,001,520,000.00
5	Real Madrid	1,450,000,000.00	151	1,010,000,000.00	147	-447,000,000.00	-440,000,000.00
6	Paris Saint-Germain	1,420,000,000.00	202	468,550,000.00	189	-955,050,000.00	-951,450,000.00
7	Atlético de Madrid	1,190,000,000.00	243	1,070,000,000.00	237	-120,800,000.00	-120,000,000.00
8	Liverpool FC	1,180,000,000.00	245	806,120,000.00	249	-376,090,000.00	-373,880,000.00
9	Inter Milan	1,170,000,000.00	631	975,330,000.00	626	-199,340,000.00	-194,670,000.00
10	Arsenal FC	1,110,000,000.00	234	525,540,000.00	236	-583,710,000.00	-584,460,000.00

Arsenal reste alors l'une des équipes ayant une balance négative importante, au total de 583 millions d'euros, suivant d'autres équipes importantes telles que *Barcelona FC, Chelsea FC et le PSG*, par exemple. Nous pourrons expliquer alors, par rapport à l'analyse globale des différents datasets, que l'équipe d'Arsenal a peut-être mal placé certains investissements, ou que son plan d'investissement des joueurs n'est pas optimal.

■ Difficultés

La difficulté principale liée à ce dataset a été surtout causée par la restructuration de la donnée une fois récupérée par le biais du Web Scraping. En effet, la Data n'a pas été récupérée au format souhaité, et il a été difficile de la remettre dans le bon format, avant de pouvoir la restructurer et la convertir en DataFrame exploitable. L'exemple ci-dessous illustre cela.

Avant

```
\nSint Maarten\nSlovakia\nSlovenia\nSolomon Islands\nSomalia\nSouth Africa\nSouthern Sudan\nS
pain\nSri Lanka\nSt. Kitts & Nevis\nSt. Lucia\nSt. Vincent & Grenadinen\nSudan\nSuriname\nSwa
ziland\nSweden\nSwitzerland\nSyria\nTahiti\nTajikistan\nTanzania\nThailand\nThe Gambia\nTibet
\nTogo\nTonga\nTrinidad and Tobago\nTunisia\nTurkey\nTurkmenistan\nTurks- and Caicosinseln\nT
uvalu\nUdSSR\nUganda\nUkraine\nUnited Arab Emirates\nUnited Kingdom\nUnited States\nUruguay\n
Uzbekistan\nVanuatu\nVatican\nVenezuela\nVietnam\nWales\nWestern Sahara\nYemen\nYugoslavia (R
epublic)\nZaire\nZambia\nZanzibar\nZimbabwe",
 'Verband:',
 'Alle Verbände\nAFC                                                                \nCAF
\nCONCACAF                                                        \nCONMEBOL
\nOFC                                                          \nUEFA',
 'Position:',
 'All positions\n \ngoalkeeper                                                          \ndefence
\nmidfield                                                  \nattack',
 'Age group:',
 'All age groups\nAll age groups\nUnder 15\nUnder 16\nUnder 17\nUnder 18\nUnder 19\nUnder 20
\nUnder 21\nUnder 23\n23 - 30\nOver 30\nOver 32\nOver 34\nOver 35\nOver 36',
 'Transfer date:',
 "doesn't matter\nSummer transfers only\nWinter transfers only",
 'Loans:',
 'All transfers\nOnly include loans\nWithout players back from loan\nExclude loans',
 'Transfers within the club:',
 'All transfers\nWithout club internal transfers',
 '',
 '',
 '1',
 '',
 'Manchester City',
 '€1.97bn',
 '319',
 '€626.61m',
 '319',
 '€-1,343.70m',
 '2',
 '',
 'Chelsea FC',
 '€1.74bn',
```

Après *(pre-preprocessing)*

	Club	Expenditure	Arrivals	Income	Departures	Balance
0	Manchester City	€1.97bn	319	€626.61m	319	€-1,343.70m
1	Chelsea FC	€1.74bn	378	€1.11bn	374	€-627.59m
2	FC Barcelona	€1.67bn	170	€1.05bn	163	€-616.27m
3	Juventus FC	€1.57bn	608	€1.05bn	597	€-526.33m
4	Manchester United	€1.55bn	208	€548.48m	214	€-1,000.06m
5	Real Madrid	€1.45bn	151	€1.01bn	147	€-447.00m
6	Paris Saint-Germain	€1.42bn	202	€468.55m	189	€-955.05m
7	Atlético de Madrid	€1.19bn	243	€1.07bn	237	€-120.80m
8	Liverpool FC	€1.18bn	245	€806.12m	249	€-376.09m
9	Inter Milan	€1.17bn	631	€975.33m	626	€-199.34m
10	Arsenal FC	€1.11bn	234	€525.54m	236	€-583.71m

	ANNEE	CLUB	ENTREES	SORTIES
0	2020	manchester city	161,80 mio. €	61,65 mio. €
1	2020	manchester united	83,80 mio. €	19,50 mio. €
2	2020	chelsea	247,20 mio. €	58,80 mio. €
3	2020	fc liverpool	82,65 mio. €	17,20 mio. €
4	2020	tottenham hotspur	110,50 mio. €	13,30 mio. €
5	2020	fc arsenal	86,00 mio. €	19,15 mio. €
6	2020	leicester city	58,00 mio. €	53,77 mio. €
7	2020	fc everton	74,87 mio. €	4,43 mio. €
8	2020	aston villa	101,35 mio. €	2,77 mio. €
9	2020	wolverhampton wanderers	84,80 mio. €	79,20 mio. €
10	2020	west ham united	54,70 mio. €	45,41 mio. €

v. Dataset 5: budgets des transferts par joueur

■ Contenu du dataset

Il existe deux datasets qui nous ont servi pour cette partie de l'analyse, dont:

a. PlayersTransfersDetails.csv
b. Top250transfers.csv

En effet, le fichier a) contient les informations suivantes:

- **Player** : joueur;
- **Left** : club quitté;
- **Joined** : club rejoint;
- **Fee** : frais du transfert;
- **Age** : âge;
- **Season** : saison;
- **Market value at time** : valeur du joueur au moment du transfert.

Par ailleurs, le fichier b) contient les informations suivantes sur les top 250 transferts les plus coûteux, groupées par club :

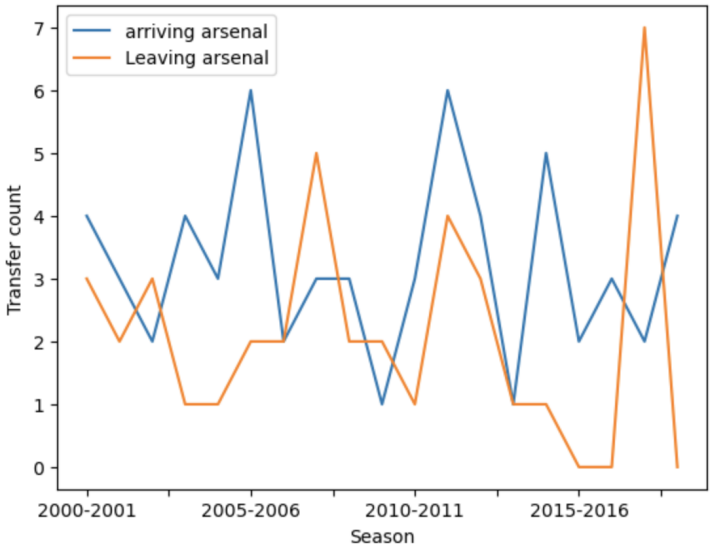
- **Club** : équipe club;
- **Buying** : nombre entrées;
- **Selling** : nombre sorties;
- **buy_sum** : total_achat_eur;
- **sell_sum** : total_vente_eur;
- **in_minus_out** : différence entrées et sorties;
- **Spent_minus_gained** : différence achat - vente.

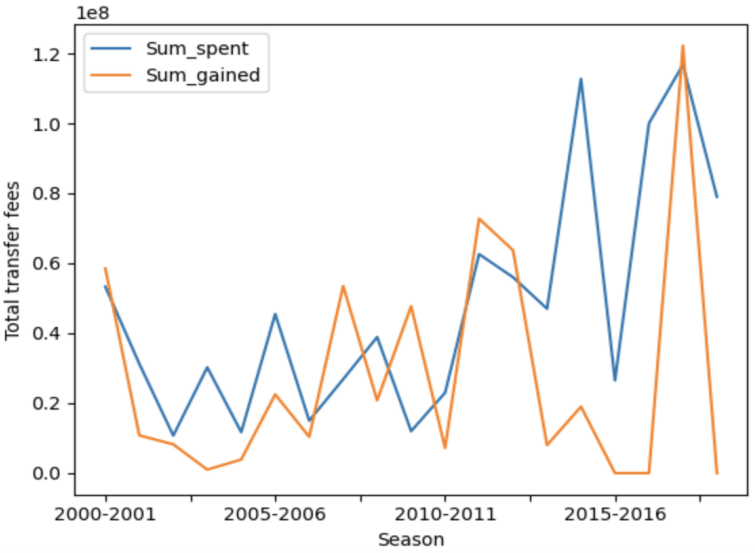
■ Récupération

Nous avons alors réunis différents datasets liés aux transferts de joueurs à partir du site "transfermarkt.fr".
Nous souhaitions avoir un aspect d'analyse lié au budget des équipes, mais il est assez difficile de récupérer des données complètes liées aux différentes sources de profit. L'économie du football étant un monde complexe, rien que pour les profits il aurait fallu réunir des informations sur les recettes des billetteries, de la publicité (affichage dans les stades) du sponsoring et des produits dérivés (merchandising), des droits de radio et de télévision.
Nous avons alors décidé de nous concentrer uniquement sur les profits et dépenses liées aux transferts de jouer, une information bien plus facile d'accès.

■ Analyse

Dans l'analyse sur le dataset des joueurs faite précédemment, nous avions remarqué que l'équipe Arsenal ne semblait pas investir dans les "top players", et dans l'analyse des budgets de transferts par club nous avions pu voir que l'équipe investissait moins que d'autres grands clubs.
Les graphiques ci-dessous permettent d'illustrer cela, mais cette fois-ci au fil du temps. On peut observer l'évolution des arrivées et départs de joueurs (à gauche) au sein du club, mais aussi l'investissement et la vente de joueurs, qui s'accélèrent un peu plus à partir de 2015-2016 (à droite).





Au niveau du positionnement par rapport aux autres équipes, Arsenal se voit un peu en retard quant à l'obtention de joueurs (illustrations ci-dessous), et même au niveau de la vente de ses propres joueurs qui ne semble pas apporter autant de fonds que pour les autres clubs.

Team_to	Buying
Inter	97
Chelsea	96
Man City	94
Spurs	93
Juventus	87
Liverpool	85
AS Roma	77
Real Madrid	75
FC Barcelona	70
Atlético Madrid	69
AC Milan	68
Paris SG	66
Man Utd	65
Everton	63
Newcastle	61
Arsenal	61
Monaco	61
Marseille	60

Team_to	buy_sum
Chelsea	1.820650e+09
Man City	1.800520e+09
Real Madrid	1.680650e+09
FC Barcelona	1.673040e+09
Man Utd	1.497360e+09
Juventus	1.470940e+09
Liverpool	1.412420e+09
Paris SG	1.274780e+09
Inter	1.202690e+09
Spurs	1.024400e+09
AC Milan	9.413500e+08
Atlético Madrid	9.188100e+08
Arsenal	8.990600e+08
AS Roma	8.263800e+08
Bayern Munich	8.123500e+08

Team_from	Selling
Inter	68
Spurs	63
Juventus	59
Chelsea	57
FC Porto	56
Liverpool	56
Real Madrid	53
Benfica	52
Atlético Madrid	52
Udinese Calcio	51
Monaco	47
Parma	46
Genoa	46
AC Milan	44
Man Utd	43
AS Roma	43
Man City	41
Newcastle	40
Olympique Lyon	40
Arsenal	40
Valencia CF	39
River Plate	39
FC Barcelona	38
Marseille	37

Team_from	sell_sum
Monaco	948170000.0
FC Porto	917550000.0
Real Madrid	891400000.0
Chelsea	839530000.0
Liverpool	798410000.0
Juventus	797980000.0
Benfica	785750000.0
Inter	785280000.0
FC Barcelona	752100000.0
Atlético Madrid	734400000.0
AS Roma	724800000.0
Spurs	699000000.0
Valencia CF	599650000.0
Bor. Dortmund	580350000.0
Sevilla FC	564600000.0
Parma	554320000.0
Olympique Lyon	535400000.0
Arsenal	530170000.0
Man Utd	527570000.0
AC Milan	489350000.0
Fiorentina	478350000.0
Udinese Calcio	472500000.0
AFC Ajax	460580000.0

Difficultés

Il n'y a pas eu de difficultés particulières quant à l'extraction de ces deux datasets. En effet, le webscraping était un peu délicat dans le sens où on a dû extraire différentes informations séparément, et les rassembler de suite dans un seul dataframe.

vi. Dataset 6: vainqueurs des titres de championnat chaque année

Contenu du dataset

Le dataset contient par saison, les noms des clubs ayant remporté un titre de championnat sur la dernière décennie et se présente de la manière suivante :

- Titre;
- Saison;
- Nom du club;
- Coach.

	titre	season	club	coach
0	Champions League	20/21	Chelsea	Thomas Tuchel
1	Champions League	19/20	FC Bayern München	Hansi Flick
2	Champions League	18/19	Liverpool	Jürgen Klopp
3	Champions League	17/18	Real Madrid	Zinédine Zidane
4	Champions League	16/17	Real Madrid	Zinédine Zidane

Récupération

À partir du site www.transfermarkt.fr, nous avons récupéré les noms des équipes gagnantes par saison pour les championnats suivants :

- Champions League
- UEFA Super Cup
- Europa League
- FA Cup
- UEFA Cup
- FIFA Club World Cup
- EFL Cup

Analyse

L'analyse du dataset montre que sur la dernière décennie, le club Arsenal n'a remporté en tout que 4 titres de championnat et ce pendant les saisons 13/14, 14/15, 16/17 et 19/20. D'autres équipes ont réussi pendant le même temps à en gagner plus, notamment Manchester City qui nous a servi plusieurs fois de base de comparaison avec Arsenal, et qui a gagné le double de titres et de manière plus régulière entre 13/14 et 20/21.

club	titre s
Real Madrid	11
Chelsea	9
Sevilla FC	8
Manchester City	8
Atlético de Madrid	7
FC Bayern München	6
FC Barcelone	6
Arsenal	4
Liverpool	4
Manchester United	4
FC Porto	2
Villarreal CF	2
FC Internazionale	1
Birmingham City	1
Wigan Athletic	1
Leicester City	1
Swansea City	1
Corinthians São Paulo	1

	season	club	titre
16	13/14	Arsenal	1
21	14/15	Arsenal	1
31	16/17	Arsenal	1
43	19/20	Arsenal	1
0	10/11	Manchester City	1
17	13/14	Manchester City	1
26	15/16	Manchester City	1
35	17/18	Manchester City	1
40	18/19	Manchester City	2
45	19/20	Manchester City	1
50	20/21	Manchester City	1
27	15/16	Manchester United	1

- **Difficultés**

Il n'y a pas eu de difficulté majeure, mais il fallait sélectionner les championnats qui nous intéressaient et à partir de la page de chaque championnat, nous avons extrait le nom de l'équipe gagnante par saison ainsi que le nom de l'entraîneur et inséré ces données dans une dataset.
Nous avons gardé les informations de la dernière décennie.

vii. Dataset 7: championnats auxquels ont participé les équipes par année

- **Contenu du dataset**

Le dataset contient les différents championnats joués par les clubs de la ligue anglaise par saison ainsi que le résultat obtenu.
Il contient les informations suivantes:

- Club;
- Championnat;
- Saison;
- Résultat.

	club	title	season	statut
0	Manchester City	The League Cup	20/21	Winner
1	Manchester City	English Champion	20/21	
2	Manchester City	Champions League	20/21	Runner Up
3	Manchester City	Champions League	19/20	Participant
5	Manchester City	The League Cup	19/20	Winner

Récupération

À partir du site www.transfermarkt.fr, nous avons récupéré, pour chaque club de la ligue anglaise, les noms des championnats auxquels ont participé les équipes.
Il a fallu mettre à jour les nom des clubs et des championnats pour qu'ils soient identiques aux autres datasets. Afin de ne pas surcharger le dataset, nous avons supprimé les saisons antérieures à la décennie analysée.

Analyse

club	title	se
Manchester City	24	
Chelsea	22	
Manchester United	16	
Arsenal	14	
Liverpool	10	
Norwich City	7	
Tottenham Hotspur	7	
Leicester City	4	
Watford	4	
Wolverhampton Wanderers	4	
Aston Villa	3	
Burnley	3	
Southampton	3	
Crystal Palace	2	
Fulham	2	
Brighton and Hove Albion	2	
Newcastle United	2	
West Ham United	2	
Leeds United	1	

club	statut	title	s
Manchester City	Winner	8	
Chelsea	Winner	7	
Arsenal	Winner	4	
Manchester United	Winner	3	
Liverpool	Winner	2	
Leicester City	Winner	1	

Arsenal a participé à 14 championnats et en a remporté 4 sur la dernière décennie comme vu précédemment, alors que Manchester City a participé à 24 championnats et en a remporté 8.

Difficultés

A partir de la liste des clubs de la Premier League, il a fallu accéder à la page de chacun pour extraire le nom des championnats auxquels ils ont participé par année. Le résultat étant inclus avec le nom du championnat, il a fallu les dissocier.

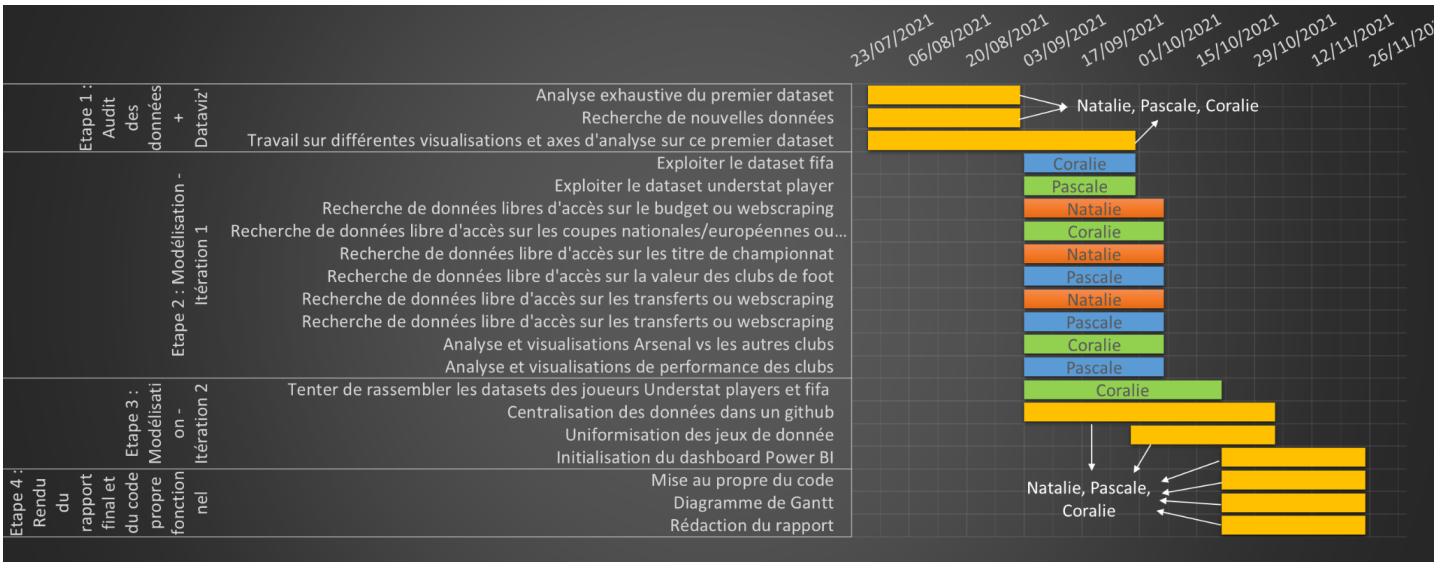
b. Répartition du travail

En commençant le rapport, les membres de l'équipe ont choisi d'explorer individuellement le dataset initial afin de trouver différents axes d'analyse à la problématique posée.

Après s'être approprié le projet, il est devenu essentiel de se répartir différentes tâches afin de pouvoir optimiser les ressources et le temps d'analyse de chaque axe pour pouvoir avancer sur le projet.

Par ailleurs, les étapes 3 et 4 du projet (*modélisation itération 2, rendu du rapport final et du code fonctionnel*) ont été réalisées en totale collaboration.

Le diagramme de Gant ci-dessous illustre les tâches effectuées et leur répartition au fil du temps.



c. Bibliographie

Les ressources utilisées pour le projet sont principalement liées aux sites internet suivants.

Les ressources libres d'accès :

- https://www.kaggle.com/pablohfreitas/all-premier-league-matches-20102021 (matchs de Premier League)
- https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv (premier dataset de joueurs Fifa)
- https://www.kaggle.com/hugomathien/soccer (base de données SQL dont sont issus le seconde dataset de joueurs fifa et le dataset de référence des équipes)
- https://www.kaggle.com/abrarhossainhimself/understat-data-for-teams-players-2014-present (dataset de statistiques sur les joueurs, qui n'a pas été conservé dans la suite du projet)

Pour le webscraping :

- https://www.transfermarkt.fr/
- https://fr.besoccer.com/competitions

IV. Bilan & Suite du projet

a. Détail de l'atteinte des objectifs cités plus tôt

Le premier objectif consistait à explorer de manière exhaustive le premier dataset qui était fourni. Nous avons montré dans ce rapport que cet objectif avait été atteint. Nous avons fait de nombreuses visualisations pour tenter de répondre à plusieurs problématiques.

Concernant le second objectif, qui consistait à enrichir notre premier jeu de données, il a été globalement atteint même si tout n'est pas parfait. Nous avons obtenu des données complémentaires concernant les joueurs, les compétitions et les transferts.

Il est vrai que ces datasets ne sont pas parfaits puisqu'à l'issue de notre travail, pour certains axes d'analyse que nous aurions souhaité explorer, nous n'avons pas pu réussir à obtenir de données (par exemple concernant le sponsoring).

Il est également vrai que les données ne sont pas parfaitement harmonisées, notamment concernant les noms des joueurs et des équipes.

Néanmoins malgré ces défauts, les datasets que nous avons réussis à obtenir nous permettent, pour la suite de ce projet, d'avoir une base suffisamment riche pour explorer différents axes d'analyse et répondre à plusieurs questions.

Enfin, un des objectifs de ce projet était de travailler en équipe, ce que nous considérons comme un objectif atteint car à tout moment du projet nous avons travaillé en collaboration.

b. Pistes d'améliorations

Pour améliorer ce projet, il y a différentes choses qu'il serait ou aurait été possible de faire.

En premier lieu, concernant le contenu du projet, plusieurs autres axes d'analyse auraient été intéressants à exploiter afin de compléter le projet. Pour cela il nous aurait fallu d'autres datasets, notamment concernant des aspects financiers comme le sponsoring, le merchandising, la billetterie, etc.

Ensuite, nous aurions pu compléter notre projet avec des analyses de sentiments. Il aurait été possible de faire des analyses de sentiments sur les équipes, mais aussi sur les coachs. En effet, il n'y a eu que 3 coachs sur la période qui concerne notre projet, une analyse de sentiment pour chacun des coachs pendant la période où il était actif à la tête de l'équipe aurait été envisageable.

C'est d'ailleurs une chose à laquelle nous avions commencé à nous intéresser, et que nous n'avons pas eu le temps de terminer. Ci-dessous un exemple de WordCloud que nous avions commencé concernant les coachs (ici Unai Emery) à partir d'articles sur internet liés aux mots-clés "bilan unai emery arsenal". C'est un départ assez intéressant, puisqu'on y retrouve par exemple le mot "limogeage" et qu'il a effectivement été limogé par Arsenal.

Nous avons également tenté une analyse de sentiments concernant les équipes de l'English Premier League à partir d'un dataset de posts sur Twitter. Ici un exemple avec un WordCloud des posts positifs qui concernent l'équipe Arsenal (nous aurions pu creuser cette analyse pour chaque saison par exemple).

Ensuite, concernant notre organisation au fil du projet, certaines choses auraient pu être mieux organisées.

Dès le départ nous aurions pu mettre en place un plan pour le déroulé de l'analyse. Nous avons passé un long moment à observer les datasets individuellement chacune de notre côté pour nous familiariser avec le projet. Néanmoins nous avons perdu du temps sur cette partie car nous avons fait plusieurs fois les mêmes observations et cette partie a été bien trop longue. Cela aurait pu être évité si nous avions élaboré un plan d'analyse et réparti l'activité plus tôt.

Enfin, nous aurions pu harmoniser les différents datasets plus tôt dans le cheminement de notre projet. Le fait de récupérer des données de sources variées a généré beaucoup de différences dans les datasets : les noms des colonnes, la date, l'écriture des noms des clubs, des joueurs et des compétitions, … Cette étape est arrivée un peu tardivement dans notre projet et peut-être que si nous l'avions envisagée plus tôt, nous aurions alors eu plus de temps à accorder à l'uniformisation, les datasets auraient été plus propres et ça aurait facilité les liaisons entre les tables sur Power BI.

c. Autres orientations possibles du projet à partir de la problématique initiale

A partir du projet initial tel qu'il nous a été soumis, d'autres orientations étaient possibles pour ce projet.

Pour commencer, nous aurions pu nous concentrer sur une analyse essentiellement financière, notamment pour évaluer la qualité des investissements du club par rapport aux autres clubs, ou encore évaluer les montants des revenus générés par rapport aux dépenses effectuées et au besoin d'investissement. Il aurait également pu s'agir d'évaluer l'impact de la baisse de niveau du club sur ses revenus : est-ce que l'équipe a réussi à garder des supporters toujours aussi fidèles ou y-a-t-il une baisse de supporters dans les stades, quel impact financier a pu avoir sur le club l'absence de participation à la Champions League 2 années consécutives, etc.

Enfin, un axe qu'il aurait été très intéressant d'explorer serait les différences stratégiques et tactiques entre l'équipe Arsenal et les autres équipes, en fonction des saisons et des managers. En effet, certains postes très utilisés par le passé n'existent plus ou quasiment plus, alors que d'autres sont devenus au fil du temps des atouts majeurs. Cette évolution dans la composition des équipes de football a nécessité une adaptation des équipes et il serait intéressant de voir si Arsenal a su s'adapter à ces changements de manière efficace.

En conclusion, ce projet pouvait être analysé de différentes manières et sous différents angles. Nous avons tenté une approche basée essentiellement sur les résultats en compétition et les statistiques des joueurs, puis complété par une analyse des budgets alloués aux transferts. Notre approche nous a permis d'exploiter différentes pistes de réflexion très intéressantes qui nous ont permis de mettre en pratique de nombreuses connaissances que nous avions pu acquérir dans les modules (matplotlib, seaborn, data quality, text mining, webscraping…). En diversifiant nos axes d'analyse nous espérions en trouver au moins un qui puisse apporter une partie de réponse à la problématique posée.

d. Conclusion

Pour ce projet, l'objectif principal était d'apporter une analyse concernant les raisons qui pouvaient expliquer la dégradation du niveau du club de football Arsenal durant la dernière décennie.

Suite à nos analyses sur les différents axes détaillés dans ce rapport, nous avons pu remarquer qu'Arsenal ne progresse pas ou pas autant que les autres équipes.

On a pu observer une amélioration du niveau des autres équipes, notamment ici Manchester City, pour la possession et les buts marqués pour ne citer que deux exemples. Pendant le même temps, on a remarqué une stagnation du niveau d'Arsenal pour ces mêmes variables, voire même une diminution des compétences quand on observe le nombre de buts encaissés au fil des années par exemple.

Concernant les transferts, Arsenal réalise moins d'investissements que les autres équipes, mais il semble aussi qu'ils ont manqué d'investir dans des "tops players".
En effet, nous avons pu remarquer l'absence d'un pool d'excellents joueurs (potentiel supérieur à 90) contrairement à d'autres équipes comme Manchester City. Ceci peut directement être mis en relation avec l'absence d'amélioration du niveau au fil des années contrairement aux autres équipes qui investissent dans d'excellents joueurs.

Concernant les championnats, les résultats sont stables en championnat (chaque année Arsenal est éliminé au même stade de la compétition), sauf à partir des dernières années de la décennie observée où ils échouent à se qualifier et à être repêchés à la plus grosse compétition de football, la Champions League. Cela traduit directement la baisse de niveau, notamment à cause des raisons citées précédemment.

Ainsi, il semblerait que plusieurs facteurs interviennent dans la dégradation du niveau d'Arsenal : un problème lié aux investissements dans les joueurs, une absence de "top players", une stagnation voire une diminution des compétences sur le terrain notamment en attaque et en défense. Tout ceci se reflète dans l'absence de participation à la Champions League deux années consécutives.

V. Annexes : description des fichiers de code

"1._Matchs_premier_league.py" : nettoyage du DataFrame des résultats des matchs de la compétition nationale de ligue anglaise sur la dernière décennie.

→ Fichier csv final : "matchs_premier_league.csv".

"2._Champions_League.py", "2._EFL_Cup.py", "2._Europa_League.py", "2._FA_Cup.py" : webscraping des résultats des matchs des compétitions anglaises et européennes sur la dernière décennie et nettoyage/mise en forme des DataFrame obtenu.

→ Fichiers csv intermédiaires : "champions_league.csv", "EFL_cup.csv", "europa_league.csv", "FA_cup.csv".

"3._Matchs_competitions.py" : rassemblement des DataFrames intermédiaires des résultats des matchs des compétitions.

→ Fichier csv final : "matchs_competitions.csv"

"4. Convert_SQLITE_database_to_CSV.py" : script d'écriture de fichiers csv à partir d'une base de données SQL.

→ Fichiers csv intermédiaires : "Country.csv", "League.csv", "Match.csv", "Player.csv", "Player_attributes.csv", "Team.csv", "Team_attributes.csv".

"5._Players_fifa_first.py" : nettoyage et mise en forme du premier DataFrame des joueurs et leurs statistiques Fifa (de 2014 à 2020).

→ Fichier csv intermédiaire : "players_fifa.csv".

"6._Complete_players_dataset.py" : création du second DataFrame des joueurs et leurs statistiques Fifa (de 2009 à 2014) à partir des fichiers csv issus de la base de données, puis rassemblement des deux DataFrames des joueurs en un seul. On en profite pour extraire un DataFrame de référence des équipes par ligue par pays.

→ Fichier csv finaux : "total_players.csv", "teams_and_leagues.csv".

"7. df_competitions" : webscraping des championnats gagnés par les équipes de la Premier League sur la dernière décennie et nettoyage/mise en forme du DataFrame obtenu.

→ Fichier csv final : " titresparchampionat.csv".

"8. df.titres" : webscraping des championnats joués par les équipes de la Premier League sur la dernière décennie et nettoyage/mise en forme du DataFrame obtenu.

→ Fichier csv final : " titresFinal.csv".

"9. df.transferts" : webscraping des montants des transferts par année des équipes de la Premier League sur la dernière décennie et nettoyage/mise en forme du DataFrame obtenu.

→ Fichier csv final : " transfertsFinal.csv".

“10. Classement_EPL.py : création d'un DataFrame depuis le Dataset Initial afin d'avoir un df dédié au Classement, par saison, des équipes par rapport à leurs points finaux et le nombre de buts (marqués - encaissés).

→ Fichier csv final : “arsenal_seasons_classement.csv”.

→ Fichier csv final : “final_classement.csv”.

→ Fichier csv final : “final_classement_grouped.csv”.

“11. Expenditures_transfermarkt__1.py : webscraping des budgets de transferts de 2009 à 2021 des clubs, avec le nombre de départs et d'arrivées par équipe, depuis le site TransferMarket.

→ Fichier csv final : “team_exp_2009_2021.csv”.

“12. market_value.py : webscraping des transferts des joueurs avec les frais des transferts par joueur, ainsi que la valeur du joueur au marché au moment de l'achat.

→ Fichier csv final : “players_transfers_details.csv”.

“13. Top250Transfers-00-19.py : webscraping des top 250 transferts les plus coûteux (toute période).

→ Fichier csv final : “top-250_transfers_2000_2018..csv”.