

Head coach dismissal effect on football team performance

Mathis Derenne¹, Ewann x¹, Scott daz¹, and Romain dcv^{1†} University of Rennes

Abstract

The goals of this paper is to investigate the effect of coach dismissal on team performance. To do that, we will use traditional statistical method that we apply to football teams.

Keywords coach dismissal, team performance

1. INTRODUCTION

1.a. Cadre de la problématique

Rappeler le rôle du coach dans le football et l'importance de son rôle dans la performance de l'équipe.

Sujet du TER : Comprendre l'effet du changement de club sur les performances du coach ET NON, comme le sujet initial (Rocaboy & Pavlik (2020)) Comprendre l'effet du changement de coach sur les performances du club Idée du prof : toutes choses étant égales par ailleurs (ceteris paribus) (idée d'un club représentatif), quelles sont les variations de performances d'un coach au cours du temps et lorsqu'il change de club

Impossible de créer "ce club égal par ailleurs" :

La création d'un club égal par ailleurs nécessite l'intervention d'un modèle qui permettrait, à partir des données du club (masse salariale, budget, performance passé du club, etc.) de normaliser la performance du club afin d'étudier précisément l'impact du coach sur cette performance

Ceci pose plusieurs problèmes :

1. Les variations de performances du coach sont difficilement observable au travers la performance de l'équipe (détailler)
2. Impossible de respecter l'hypothèse d'uncounfoundness requise par de nombreux modèles statistiques corrigeant les externalités (ex: propensity score / PSM) (citer papier propensity score + expliquer l'idée du propensity score pour artificiellement recréer un groupe contrôle et test artificiel, expliquer l'hypothèse d'uncounfoundness et pourquoi elle est nécessaire)
3. Biais de causalité (point le plus important !) : on suppose que c'est la performance du coach qui fait varier la performance de l'équipe or, dès lors que cette causalité n'est plus vérifiée on se mord la queue dans la création du modèle explicatif :

Supposons que ce soit la performance de l'équipe qui causent les variations de performance du coach. Le modèle explicatif, censé créer ce club égale par ailleurs, va être amené à normaliser plus fortement un club qui performe bien par le passé. Or si c'est la performance de l'équipe qui cause la performance du coach on est en train de normaliser les variations de performance du coach. (mentionner l'existence de test d'inversion de la causalité + référence au papier) (expliquer ce que sont les fuites de données (data leakage) et que l'absence de cette hypothèse de causalité provoque des fuites de données entre les externalités et la variable d'intérêt (la performance du coach)).

1. Le peu de donnée (retrouver le chiffre sur le nombre de club avec au moins 2 ou 3 changements de coach) (expliquer que dans la problématique initiale il y a bien plus de donnée car il y a davantage de club qui ont vu passer de coachs que de parcours individuel de coach au sein de clubs)

2. Problème de temporalité : les données sur les budgets des équipes, masse salariale ou valeur marchande des équipes ne sont pas disponible sous forme temporelle : impossible de savoir si la hausse de performance de l'équipe est due à la hausse du budget de l'équipe ou inversement.
3. Faible qualité des variables exogènes permettant l'analyse du système :
 - Manque " d'objectivabilité " des variables externes : masse salariale (pas représentative, ex : sous-traitance), valeur marchande des joueurs (hautement subjectif)
 - Manque de diversité des variables

Conclusion : Lors de l'analyse des effets dans un système, on raisonne généralement à petite entité égales par ailleurs
Exemple : On parle d'agent économique représentatif et rarement d'une économie représentative :

- On observe l'effet de l'économie sur un agent économique
- et NON l'effet d'un agent économique sur l'économie

(à nuancer pour ne pas déplaire aux micro-économistes et rappeler le cadre statistiques de l'étude d'effets quantifiables !).

La lecture de Wilke (2019) a permis l'amélioration de la qualité des graphiques et de la présentation des données.

1.b. Source des données

Les données utilisés au cours de cette analyse sont extraites de deux sites spécialisés dans les statistiques de football : [Fbref](#) et [Transfermarkt](#).

- Fbref offre une gamme complète de données statistiques sur les joueurs, les équipes, les ligues et les compétitions de football du monde entier. Il propose des informations détaillées telles que les buts marqués, les passes décisives, les tirs au but, les interceptions et bien d'autres statistiques.
- Transfermarkt est une ressource en ligne majeure pour tout ce qui concerne les transferts de joueurs de football, les rumeurs de transferts, les valeurs marchandes des joueurs ainsi que les informations sur les contrats. Il offre une base de données exhaustive des joueurs, des clubs et des agents, ainsi que des détails sur les transferts passés et actuels.

Ces sites sont utilisés par les amateurs de football, les journalistes et les professionnels pour rester informés sur les évolutions au cours de la saison ou pendant les trêves/mercatos.

1.c. Fiabilité des données

Ces sites sont très utilisés et considérés comme fiable. Fbref est entretenu par l'entreprise Sport Reference qui gère également d'autres sites spécialisés dans les statistiques sportives, comme Baseball-Reference et Basketball Reference. Les données sur Fbref sont souvent vérifiées et mises à jour régulièrement, ce qui contribue à leur fiabilité. Pour Transfermarkt, c'est aussi un site très utilisé pour les rumeurs de transferts et les transferts en général, il a une réputation de site fiable. Le site recueille des données sur les transferts, les valeurs marchandes des joueurs et d'autres détails liés aux contrats à partir de diverses sources, y compris les médias et les communiqués officiels des clubs. Cependant, c'est un site reliant des rumeurs de transferts, donc il peut y avoir des inexactitudes ou des spéculations qui ne se concrétisent pas toujours. Il est donc conseillé de vérifier les informations avec d'autres sources fiables, notamment lorsqu'il s'agit de transferts non confirmés

1.d. Les outils utilisés

L'intégralité du travail de récupération, de pré-traitement, d'analyse et visualisation des données a été réalisé au sein de Notebook Jupyter.

La récupération des données footballistique a été effectué à l'aide du package R [WorldFootballR](#). Ce package est régulièrement mis à jour et implémente des outils de web scraping afin d'extraire les données des principaux sites footballistiques.

Le pré-traitement et l'analyse des données s'est effectué sous Python à l'aide des librairies standards : Pandas, Numpy, Matplotlib, Seaborn, Scipy, Statsmodels et Scikit-learn.

La création d'un tableau de bord interactif a été réalisé à l'aide de la librairie ipywidgets.

L'écriture de ce papier a été réalisé dans un fichier Markdown.

MyST fait partie d'un écosystème d'outils qui cherchent à améliorer le travail de communication scientifique en favorisant le développement d'une science reproductible et indexable. Cet outil a été utilisé pour permettre la diffusion de ce papier de recherche au format d'un [site statique](#) et d'un [PDF](#) de qualité scientifique.

MyST permet de réutiliser les entrées et les sorties des Notebooks Jupyter. Ainsi l'ensemble des figures, tableaux et variables présentes dans ce papier sont directement issus des Notebooks Jupyter. À titre d'exemple, il est possible de renouveler l'intégralité de l'étude à d'autres ligues ou d'autres périodes en modifiant simplement les paramètres des fonctions utilisées dans les Notebooks Jupyter :

```
country <- c("ENG", "ESP", "ITA", "GER", "FRA")
year <- c(2018, 2019, 2020, 2021, 2022)
```

2. PRÉSENTATION DES DONNÉES EXTRAITES

2.a. Données des coaches

Les données des coaches sont extraites de Fbref. Elles contiennent des informations sur les coaches de football, notamment leur nom, leur date de naissance, leur nationalité, les clubs pour lesquels ils ont travaillé, les dates de début et de fin de leur mandat, ainsi que des statistiques sur les matchs qu'ils ont dirigés.

	team	league	country	coach	staff_start	staff_nationality	staff_nationality	start_date	end_date	days_in_post	goals	swins	draws	losses
0	Manchester City	Premier League	England	Pep Guardiola	Jan 18, 1971	Spain	NaN	2016-07-11	2024-06-30	2784	450	333	53	64
1	Liverpool FC	Premier League	England	Jürgen Klopp	Jun 16, 1967	Germany	NaN	2015-10-08	2024-06-30	3188	468	291	96	81
2	Chelsea FC	Premier League	England	Graham Potter	May 20, 1975	England	NaN	2022-09-01	2023-04-02	206	31	12	8	11
3	Chelsea FC	Premier League	England	Thomas Tuchel	Aug 29, 1973	Germany	NaN	2021-01-26	2022-09-01	589	100	63	19	18
4	Chelsea FC	Premier League	England	Frank Lampard	Jun 20, 1978	England	NaN	2019-07-01	2021-01-25	551	84	44	15	25

Table 1: Jeu de donnée des coaches

Les données sur les coaches contiennent enregistrement de coaches. À continuer...== Données sur les matchs

Les données sur les matchs sont extraites de Transfermarkt. Elles contiennent des informations sur les matchs de football, notamment les équipes qui ont joué, le score final, le lieu du match, la date du match.

	league	country	season_year	date	home_team	home_goals	away_team	away_goals
--	--------	---------	-------------	------	-----------	------------	-----------	------------

0	Premier League	England	2018	2017-08-11	Arsenal	4.0	Leicester City	3.0
1	Premier League	England	2018	2017-08-12	Watford	3.0	Liverpool	3.0
2	Premier League	England	2018	2017-08-12	Crystal Palace	0.0	Huddersfield	3.0
3	Premier League	England	2018	2017-08-12	West Brom	1.0	Bournemouth	0.0
4	Premier League	England	2018	2017-08-12	Chelsea	2.0	Burnley	3.0

Table 2: Jeu de donnée des matchs

Data collected from match results ranges from '08/04/2017' to '05/29/2022' and contains the results of 9148 matches.

Matches have been collected for the following leagues : **'Premier League, La Liga, Ligue 1, Bundesliga, Serie A'** = Pré-traitement des données

Utilisation de l'algorithme de la distance Levenshtein (Levenshtein, 1965) pour matcher les noms des clubs entre les deux jeux de données

```
from thefuzz import process

def match_clubs_name(name, list_names, min_score=70):
    scores = process.extract(name, list_names, limit=1)

    if len(scores) != 0 and scores[0][1] >= min_score:
        return scores[0][0]
    return None
```

Code 1: Utilisation de l'algorithme de la distance Levenshtein

Vérification que tout les clubs ont bien au plus un coach pour chaque période

Reims a plusieurs coachs pour la même périodeWe will exclude head coaches with more than 3000 days in post.
Expliquer que ce sont des cas minoritaire et que l'entraîneur le plus ancien a exercé pendant 8000 jours dans un club et que ça déforme les graphs et l'analyse stat.

3. LES GRAPHIQUES

3.a. Graphique du jeu de donnée head coach

Présenter les graphiques, expliquer les variables utilisés et ce que permettrait d'interpréter un graphique concluant

Les saisons de football sont divisées en deux périodes : la saison régulière et la saison hors-saison. La saison régulière est la période pendant laquelle les équipes jouent des matchs de championnat et de coupe, tandis que la saison hors-saison est la période pendant laquelle les équipes se préparent pour la saison suivante, notamment en recrutant de nouveaux joueurs et en changeant d'entraîneur.

Les licenciements de coachs sont plus fréquents en fin de saison (voir [Figure 1](#)), tandis que les nominations de coachs sont plus fréquentes en début de saison (voir [Figure 2](#)). Cela peut s'expliquer par le fait que les clubs cherchent à renouveler leur effectif et à se donner les meilleures chances de succès pour la saison suivante.

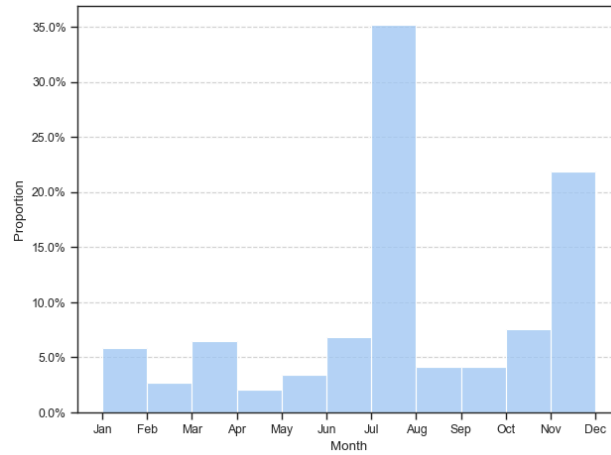


Figure 1: Monthly Distribution of Head Coaches Appointments

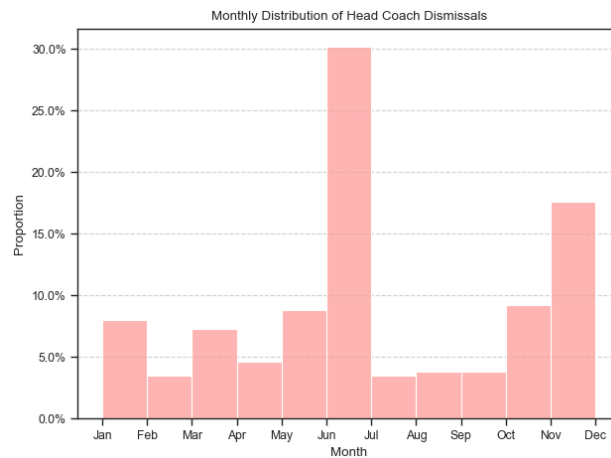


Figure 2: Monthly Distribution of Head Coaches Dismissals

Plus de 50% des coachs sportifs sont renouvelés après 1 an de mandat. Ce pourcentage augmente à 80% après 2 ans de mandat et à 90% après 3 ans de mandat (voir [Figure 3](#))

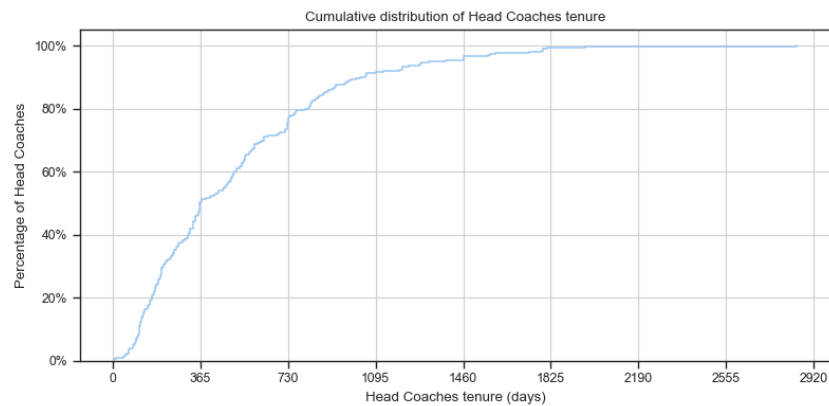


Figure 3: Empirical Cumulative Distribution Function of Head Coaches Tenure For Completed Appointments

Au cours de la période 2017 - 2022, plus de 55% des coachs sportifs n'ont entraîné qu'un seul club. Environ 30% des coachs ont entraîné 2 clubs et seulement 10% des coachs ont entraîné plus de 3 clubs au cours de cette période (voir Figure 4).

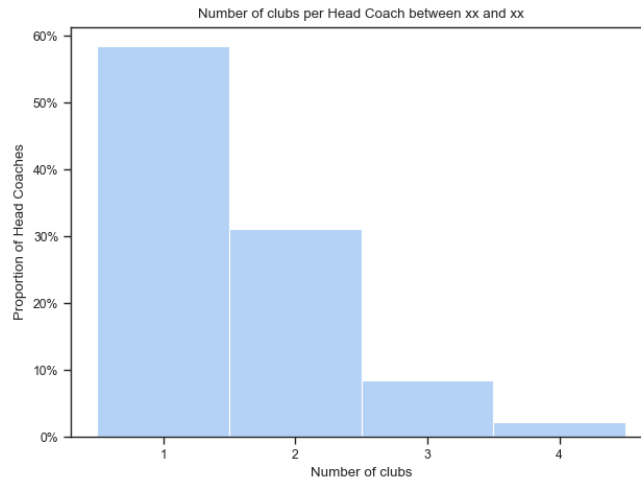


Figure 4: Proportion of Head Coaches by Number of Club Appointments (2017 - 2022)

Lorsque l'on s'intéresse au nombre de coach employés par les clubs durant la période 2017 - 2022, on observe que plus de 90% des clubs ont employés au moins 3 coachs différents (voir Figure 5).

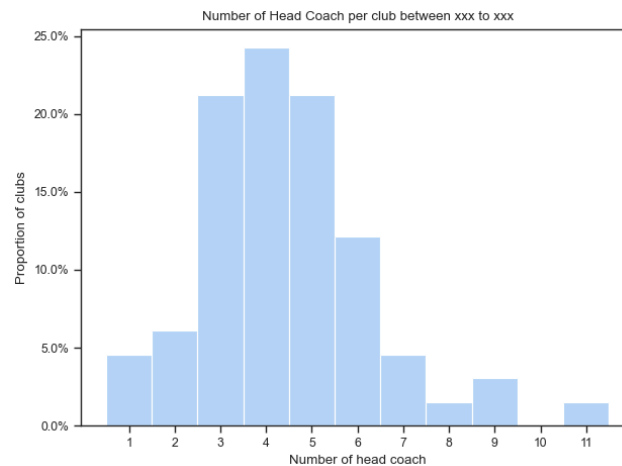


Figure 5: Proportion of Clubs by Number of Head Coaches Appointed (2017 - 2022)

Les Figure 6 et Figure 7 observent s'intéresse à l'ancienneté des coachs sportif et au renouvellement des coachs sportifs par rapport aux ligues d'intérêt.

Text(0, 0.5, '')

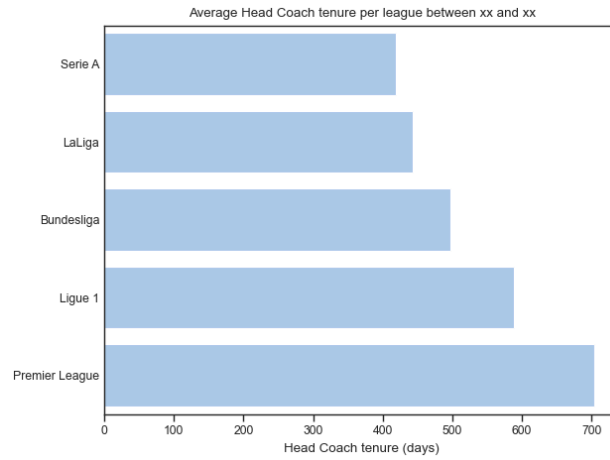


Figure 6: Average Head Coach Tenure for Completed Appointments per League

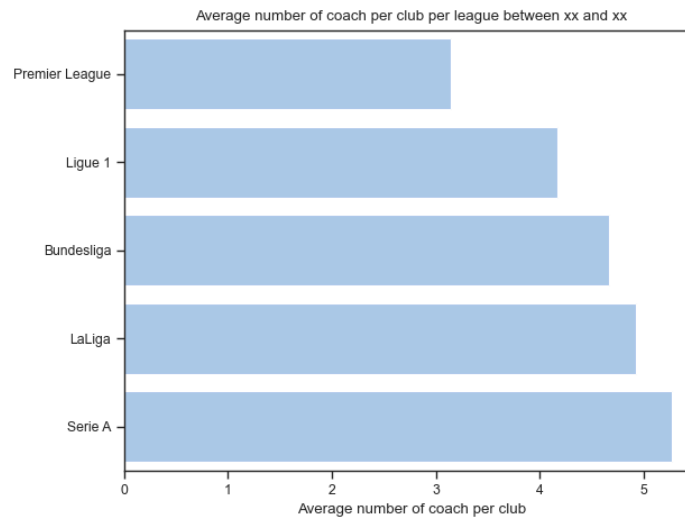


Figure 7: Average Number of Head Coaches Appointed per Club versus League (2017 - 2022)

On observe que les coachs de la Premier League ont une ancienneté plus longue que les coachs des autres ligues. De plus, les clubs de la Premier League ont tendance à nommer moins de coachs que les clubs des autres ligues. Inversement, c'est LaLiga qui a la plus faible ancienneté moyenne des coachs et qui nomme le plus de coachs.

Expliquer chacune des régressions et ce qu'elle permettrait de montrer Donner la définition du coefficient de corrélation de Pearson Interpréter les valeurs r et p

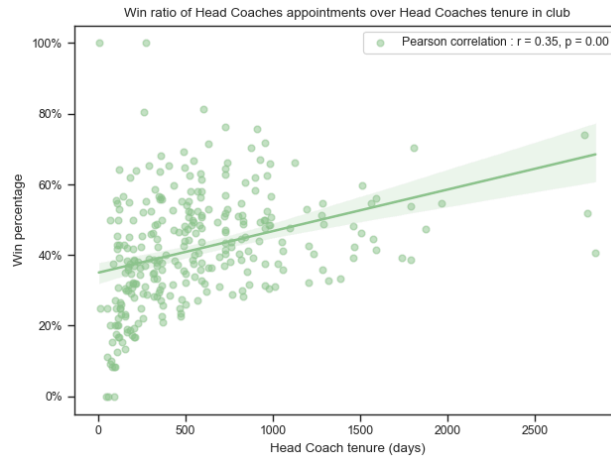


Figure 8: Win Ratio of Head Coaches Appointments versus Head Coach Tenure

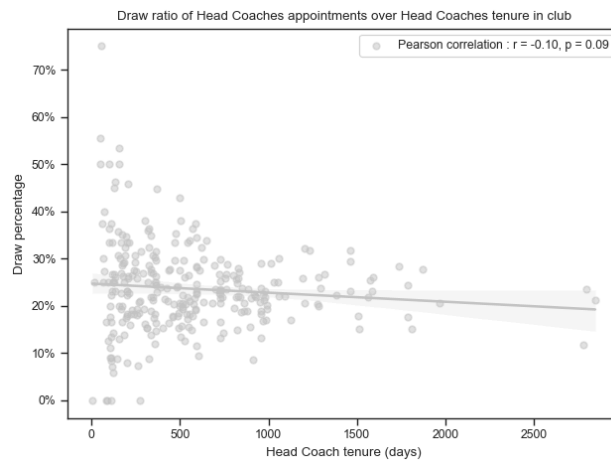


Figure 9: Draw Ratio of Head Coaches Appointments versus Head Coach Tenure

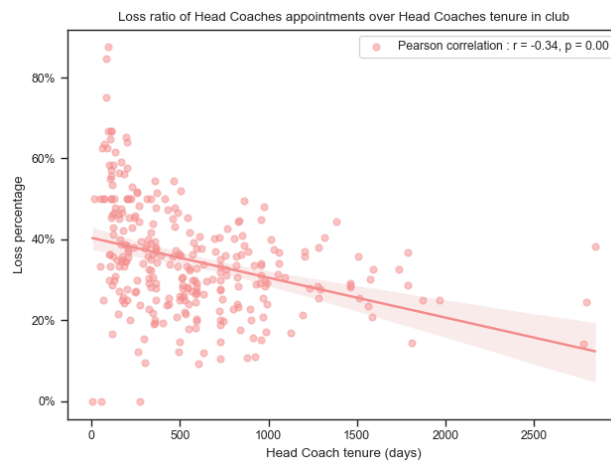


Figure 10: Loss Ratio of Head Coaches Appointments versus Head Coach Tenure

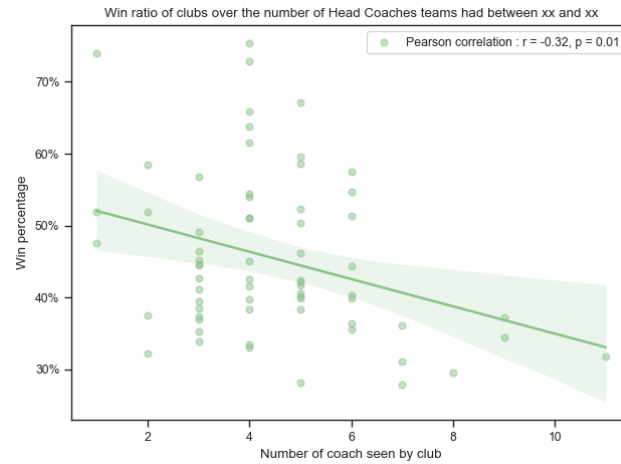


Figure 11: Win Ratio of Clubs versus Number of Head Coaches Appointed by Club

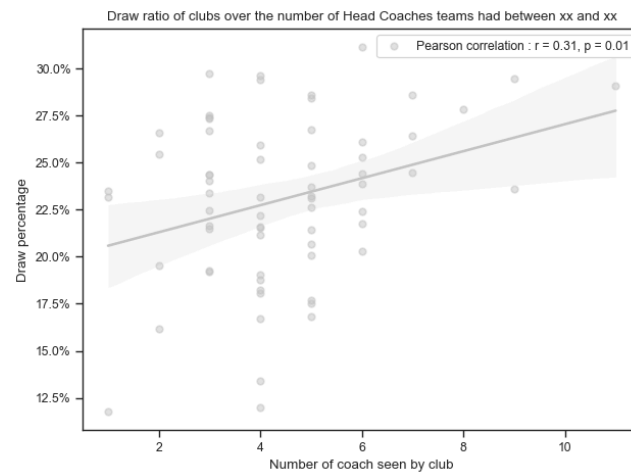


Figure 12: Draw Ratio of Clubs versus Number of Head Coaches Appointed by Club

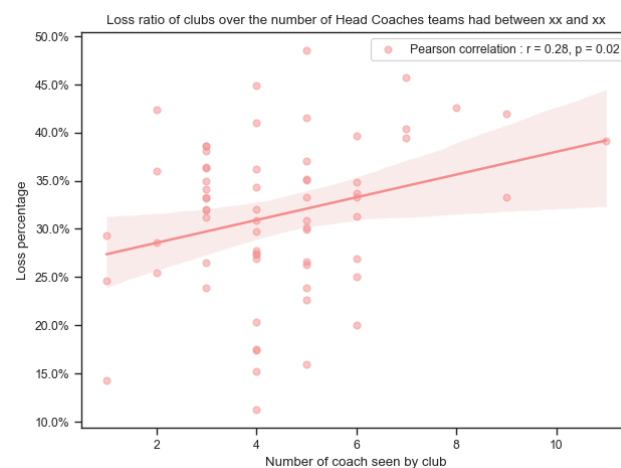


Figure 13: Loss Ratio of Clubs versus Number of Head Coaches Appointed by Club

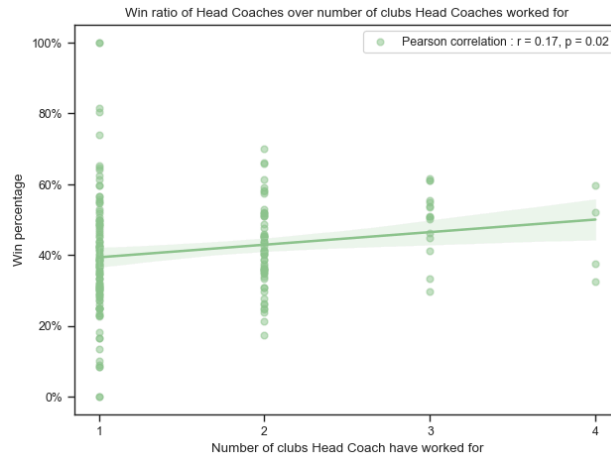


Figure 14: Win Ratio of Head Coaches vers Number of Clubs Appointments

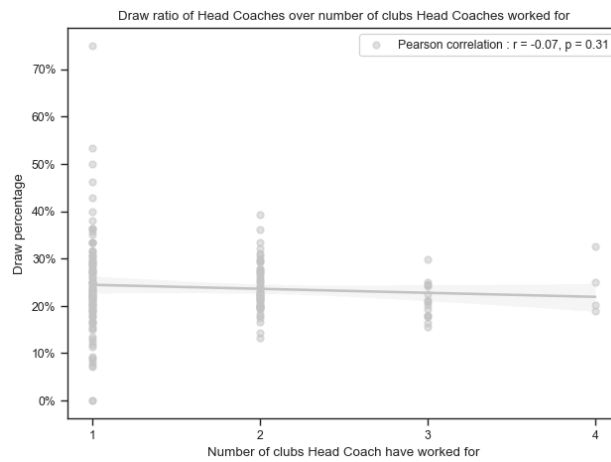


Figure 15: Draw Ratio of Head Coaches vers Number of Clubs Appointments

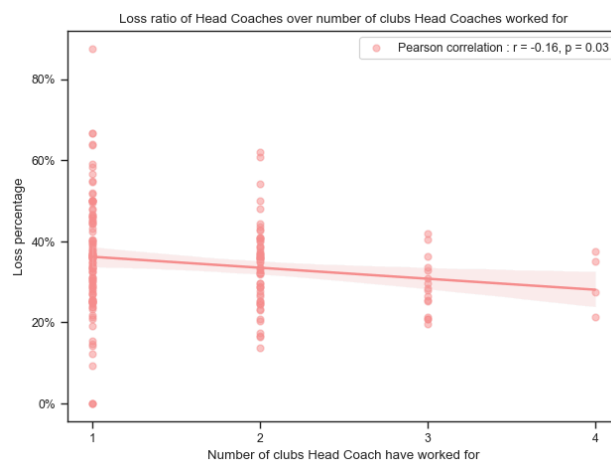


Figure 16: Loss Ratio of Head Coaches vers Number of Clubs Appointments

3.b. Graphiques du jeu de donnée match

	Number of match played	Average goals	Number of teams	Number of teams with coach change
Leagues				
Ligue 1 (France)	1908	2.68	28	12
La Liga (Spain)	1900	2.55	28	14
Premier League (England)	1900	2.75	28	15
Serie A (Italy)	1900	2.86	28	15
Bundesliga (Germany)	1540	3.06	27	13

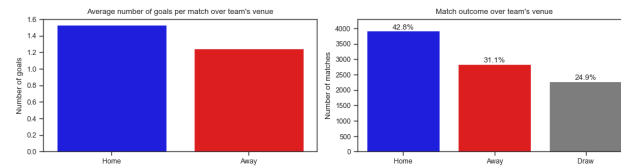


Figure 17: Venue effect on team's performance (2017 - 2022)

Il existe une différence dans la performance des équipes lorsqu'elle joue à domicile ou à l'extérieur (voir [Figure 17](#)).

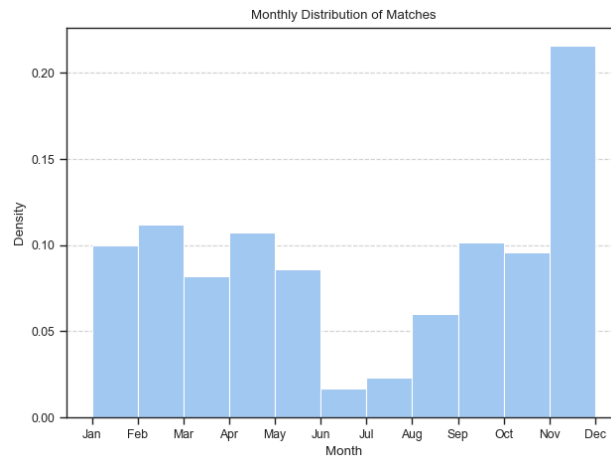


Figure 18: Monthly Distribution of Matches (2017 - 2022)

3.c. Graphiques des données jointes

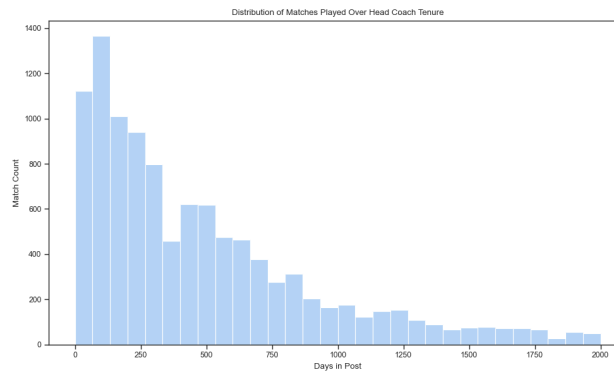


Figure 19: Distribution of Matches versus Head Coach Tenure on Match Day

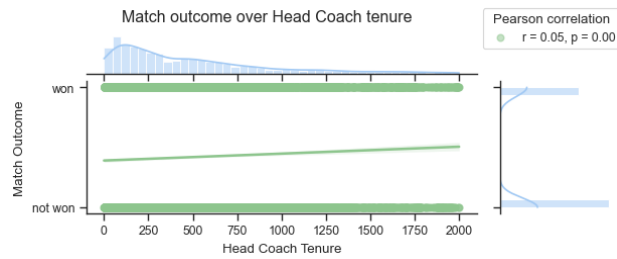


Figure 20: Match Win Outcome versus Head Coach Tenure on Match Day

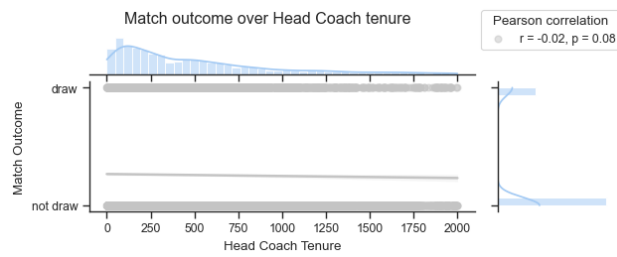


Figure 21: Match Draw Outcome versus Head Coach Tenure on Match Day

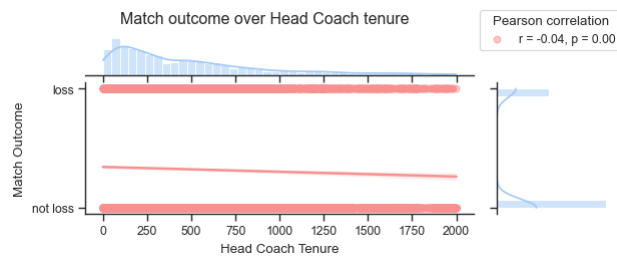


Figure 22: Match Loss Outcome versus Head Coach Tenure on Match Day

Correlation between head coach tenure and team's performance

- could indicate that club keeps their well performing head-coaches
- could indicate that head coaches performance improve after time either because:
 - early low performance : coaches need some time once they are appointed to reach previous team performance
 - long term improvement of performance
- expliquer pourquoi cette regression est la plus statistiquement pertinente pour montrer l'effet de l'ancienneté du coach : on observe match par match et non à l'échelle de la performance total d'un coach au sein d'une équipe

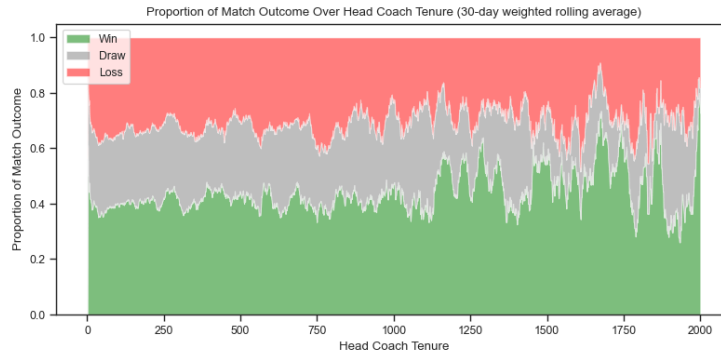


Figure 23: Weighted Rolling Average of Match Outcome vers Head Coach Tenure on Match Day

Explique que graph utilise les moyenne mobile pondérés sur une fenêtre de 30 jours :

```
import numpy as np

def weighted_rolling_mean(data, weights, window_size=30):
    def weighted_mean(x):
        return np.average(data.loc[x.index], weights=weights.loc[x.index])

    return data.rolling(window_size, min_periods=1).apply(weighted_mean, raw=False)
```

Code 2: Calcul des moyennes mobiles pondérées

3.d. Création d'un tableau de bord interactif

Création d'un tableau permettant de visualiser

L'ensemble des fichiers et données relatif à ce travail sont disponible en accès libre sur le [dépot GitHub](#) sous licence MIT.

REFERENCES

- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics. Doklady*, 10, 707–710. <https://api.semanticscholar.org/CorpusID:60827152>
- Rocaboy, Y., & Pavlik, M. (2020). Performance Expectations of Professional Sport Teams and In-Season Head Coach Dismissals—Evidence from the English and French Men's Football First Divisions. *Economies*, 8(4), 82–83. <https://doi.org/10.3390/economies8040082>
- Wilke, C. O. (2019). *Fundamentals of Data Visualization*. O'Reilly Media. <https://clauswilke.com/dataviz/>