Head coach dismissal effect on football team performance

Mathis Derenne[1], **Ewann x**[1], **Scott daz**[1], and **Romain dcv**[11]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

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 précisemment 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'interêt (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 dûe 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 !).

Référence à citer : https://clauswilke.com/dataviz/

1.a. Les données

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

- 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.b. La 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 Transfermakt, 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.c. Les outils utilisés

La récupération des données sur ces sites a été effectué à l'aide du package R WorldFootBallR. Ce package implémente des outils du web scraping pour extraire des données footballistique et est régulièrement mis à jour. Le travail a été réalisé au sein de notebook Jupyter et est disponible en open source sur le dépot GitHub.

MyST-MD a été utilisé pour générer l'export de notre papier sous format web et pdf à partir de Notebook Jupyter et de fichier Markdown. Il permettent la création de documents structurés et interactifs et incitent au développement d'une science reproductible.

Des outils communautaires pour l'avenir de la communication et de la publication techniques.

2. Pré-traitement des données

Utilisation de l'algorithme de la distance Levenshtein Haldar & Mukhopadhyay (2011) pour matcher les noms des clubs entre les deux jeux de données

```python
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'entraineur 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.

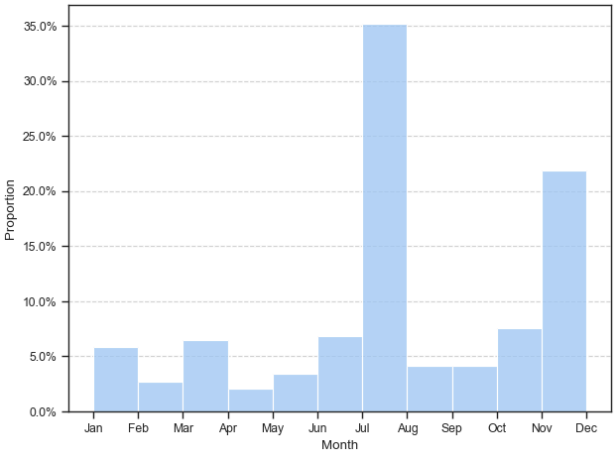

Figure 1: Monthly Distribution of Head Coaches Appointments

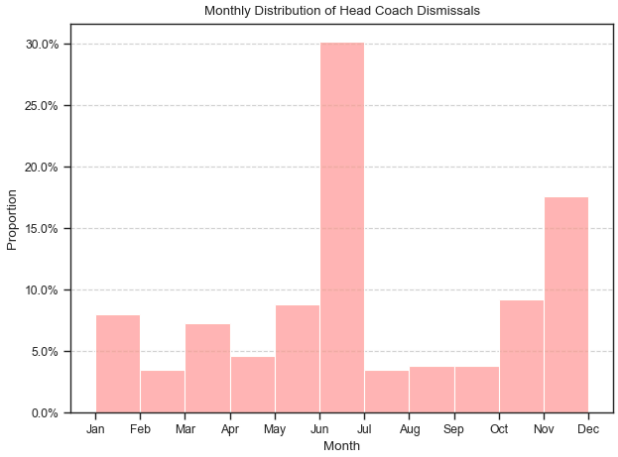

Figure 2: Monthly Distribution of Head Coaches Dismissals

```
Text(0.5, 0, 'Head Coaches tenure (days)')
```

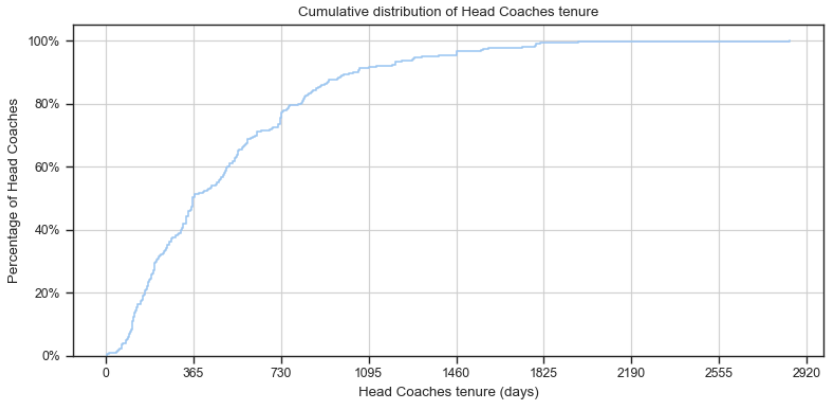

Figure 3: Empirical Cumulative Distribution Function of Head Coaches Tenure For Completed Appointments

- plus de 90% des coachs sont renouvelés au delà de 3 ans
- on observe une saisonnalité annuelle : les coachs restent pour des mandats de 1 an ou 2 ans.

```
Text(0, 0.5, '')
```

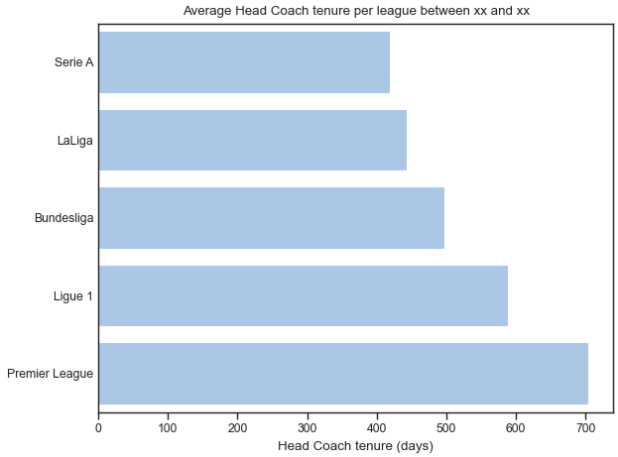

Figure 4: Average Head Coach Tenure for Completed Appointments per League

```
Text(0, 0.5, 'Proportion of Head Coaches')
```

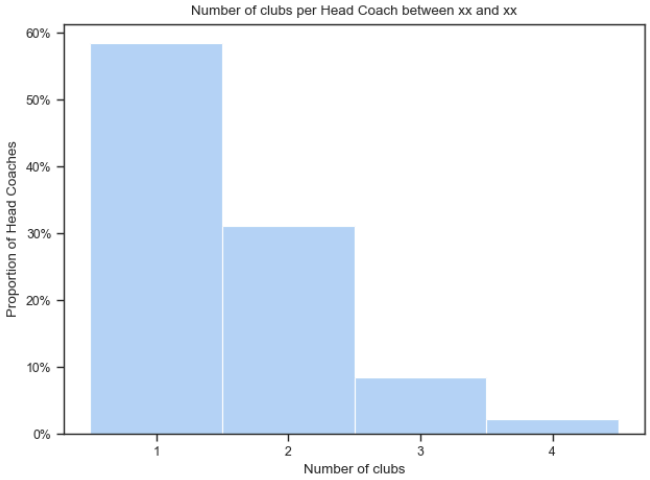

Figure 5: Proportion of Head Coaches by Number of Club Appointments (2017 - 2022)

```
Text(0, 0.5, 'Proportion of clubs')
```

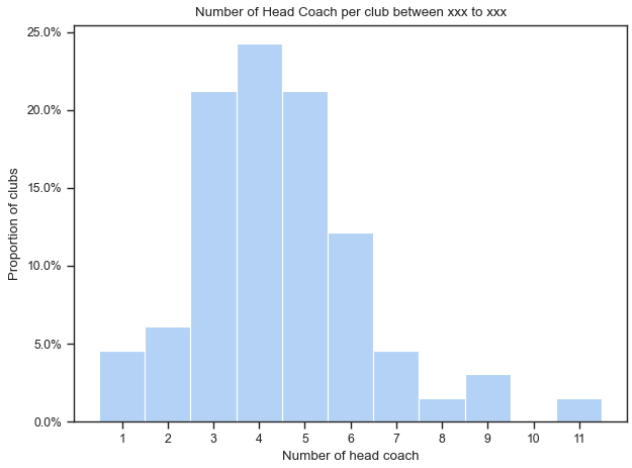

Figure 6: Proportion of Clubs by Number of Head Coaches Appointed (2017 - 2022)

```
Text(0.5, 0, 'Average number of coach per club')
```

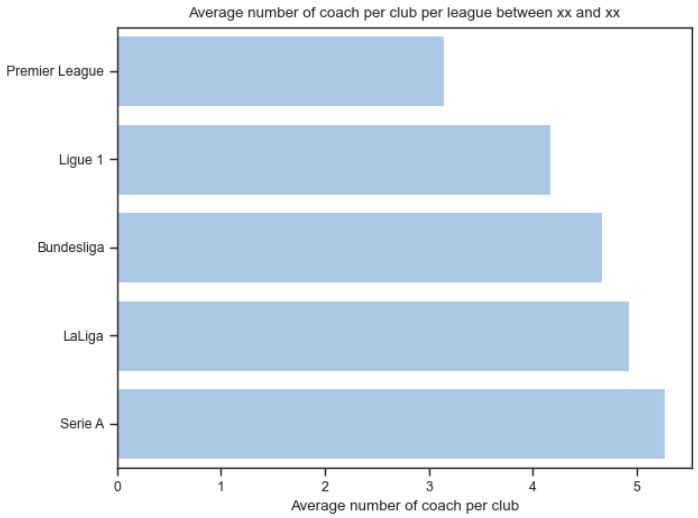

Figure 7: Average Number of Head Coaches Appointed per Club versus League (2017 - 2022)

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

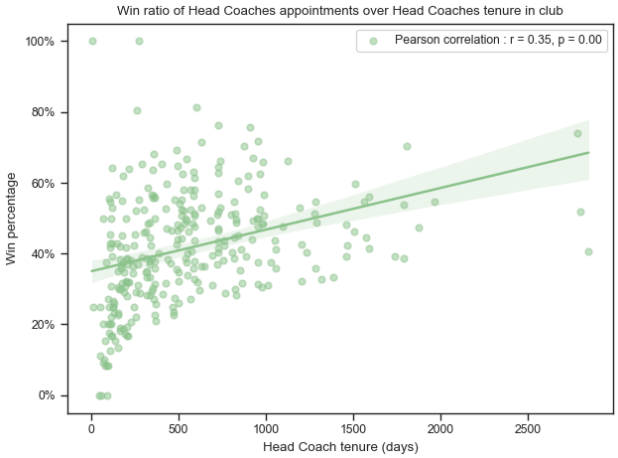

Figure 8: Win Ratio of Head Coaches Appointments versus Head Coach Tenure

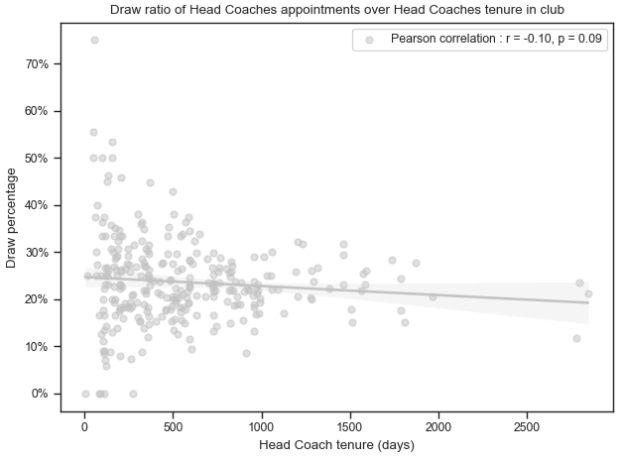

Figure 9: Draw Ratio of Head Coaches Appointments versus Head Coach Tenure

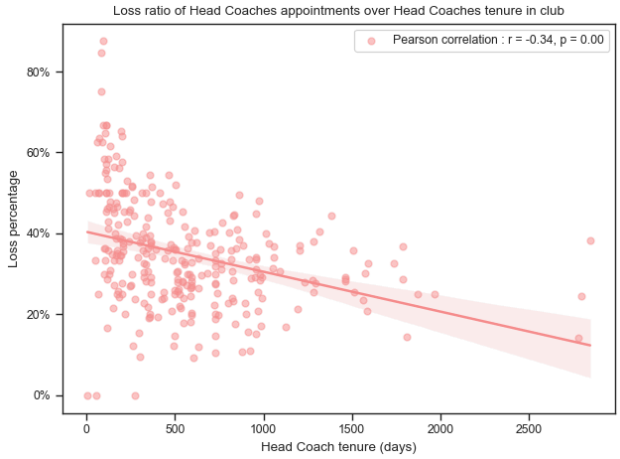

Figure 10: Loss Ratio of Head Coaches Appointments versus Head Coach Tenure

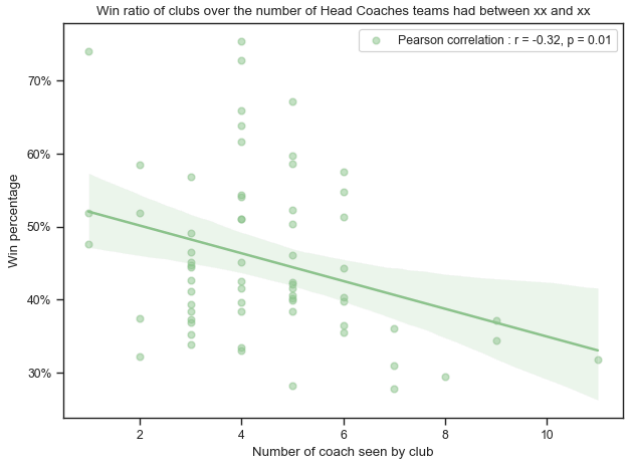

Figure 11: Win Ratio of Clubs versus Number of Head Coaches Appointed by Club

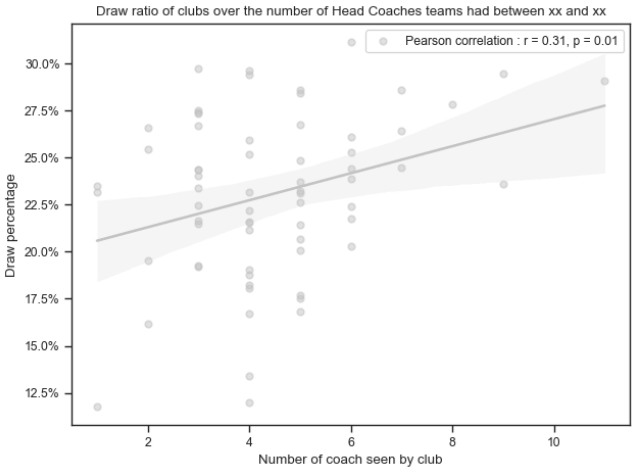

Figure 12: Draw Ratio of Clubs versus Number of Head Coaches Appointed by Club

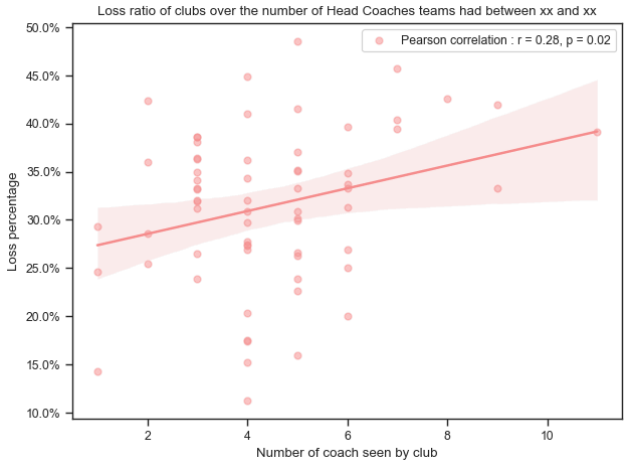

Figure 13: Loss Ratio of Clubs versus Number of Head Coaches Appointed by Club

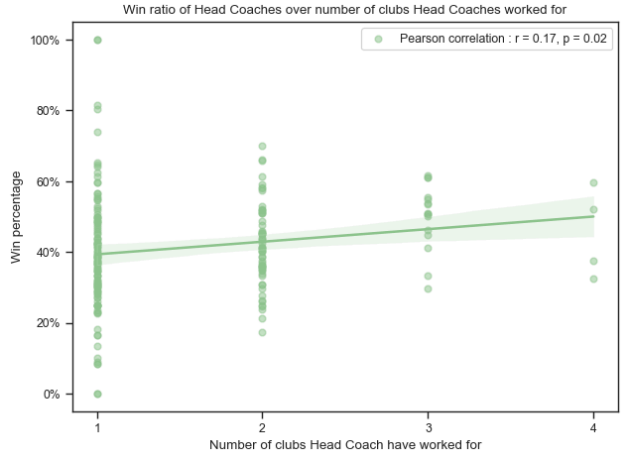

Figure 14: Win Ratio of Head Coaches vers Number of Clubs Appointments

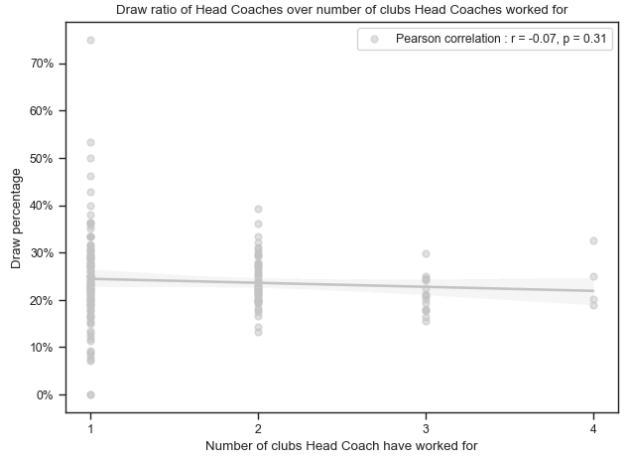

Figure 15: Draw Ratio of Head Coaches vers Number of Clubs Appointments

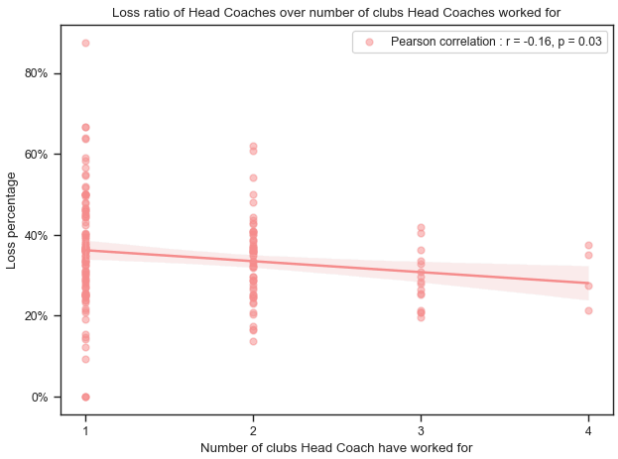

Figure 16: Loss Ratio of Head Coaches vers Number of Clubs Appointments

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

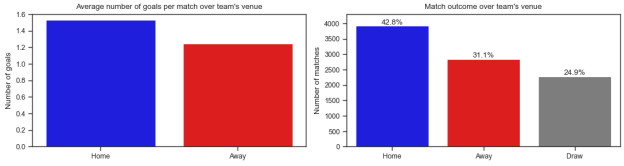

Figure 17: Venue effect on team's performance (2017 - 2022)

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

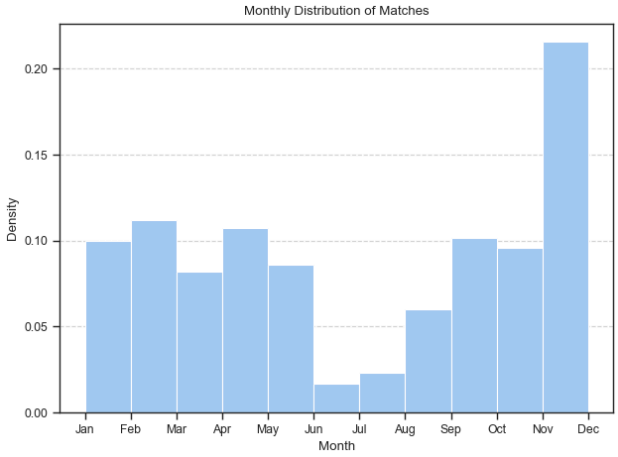

Figure 18: Monthly Distribution of Matches (2017 - 2022)

3.c. Graphiques des données jointes

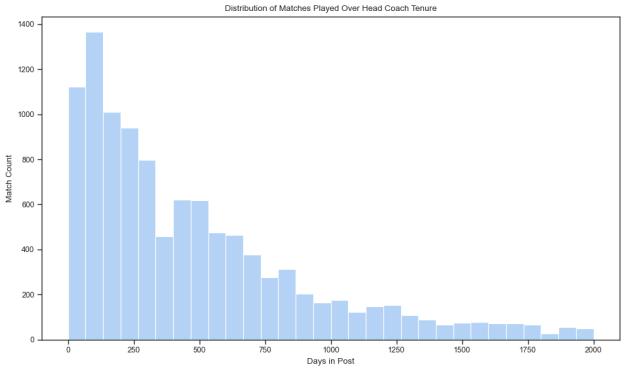

Figure 19: Proportion of Matches versus Head Coach Tenure on Match Day

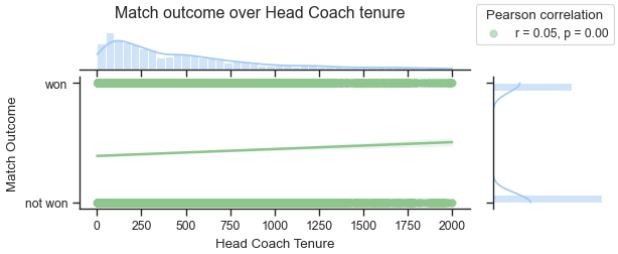

Figure 20: Match Win Outcome versus Head Coach Tenure on Match Day

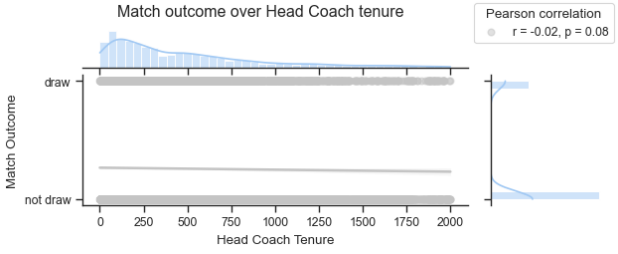

Figure 21: Match Draw Outcome versus Head Coach Tenure on Match Day

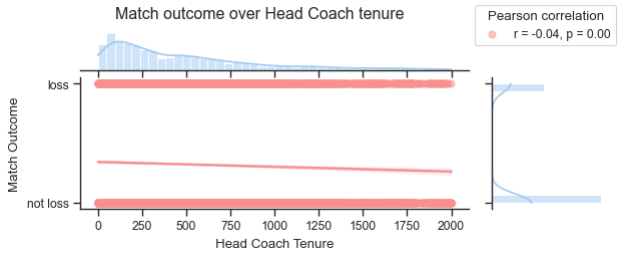

Figure 22: Match Loss Outcome versus Head Coach Tenure on Match Day

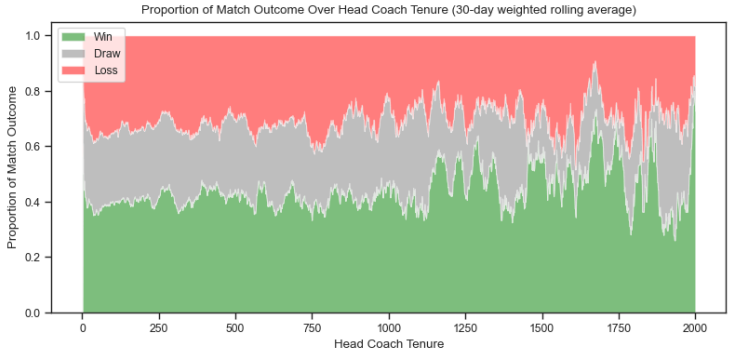

Figure 23: Weighted Rolling Average of Match Outcome vers Head Coach Tenure on Match Day

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

```python
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

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'ancieneté du coach : on observe match par match et non à l'échelle de la performance total d'un coach au sein d'une équipe

3.d. Tableau de bord interactif

- Création d'un tableau permettant de visualiser... voir notebook 6

References

Haldar, R., & Mukhopadhyay, D. (2011). *Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach*. arXiv. https://doi.org/10.48550/ARXIV.1101.1232

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