PROJET
FOOTBALL STATISTIQUE

Équipe: CHIRINA Rania, Gonzalez Emmanuel, MAKHLOUFI Khalil, Martin Samuel, VITOFFODJI Adjimon.

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Avril 2024

Soumis comme contribution partielle
pour le cours de Science des données

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Remerciements

Nos plus sincères remerciements vont à nos encadrants pédagogiques pour les conseils avisés sur tout le long de notre projet.

24/05/2024.

Résumé

En analysant visuellement notre jeu de données portant sur les statistiques des attaquants de football dans les cinq grands championnats d'Europe, nous avons pu identifier plus clairement la qualité et la constance des joueurs et les afficher clairement à travers de divers graphiques. Par ailleurs, notre approche en data science nous a permis de développer une fonction de prédiction basée sur un modèle de régression linéaire multiple, nous permettant ainsi de projeter les performances des joueurs pour les prochaines années.

Table des matières

Introduction

Dans le monde du football, les statistiques ne sont pas simplement des chiffres. Elles incarnent le récit dynamique des performances, des succès et des échecs des joueurs qui ont marqué l'histoire de ce sport passionnant. De 2016 à 2023, une période qui a été témoin de certains des moments les plus mémorables sur les terrains de football à travers le monde, l'analyse statistique des joueurs a pris une importance capitale pour les entraîneurs, les analystes et les passionnés de football.

Comme l'a souligné Johan Cruyff, légendaire joueur et entraîneur de football, "Les statistiques sont comme un bikini : elles montrent beaucoup de choses, mais elles cachent l'essentiel." Cette citation emblématique résume parfaitement la complexité des chiffres dans le contexte du football. Alors que les statistiques peuvent offrir des indications précieuses sur les performances individuelles et collectives des joueurs, elles ne capturent jamais entièrement l'essence du jeu, son flair artistique et son imprévisibilité.

Durant cette période quinquennale, les données ont été scrutées de près pour évaluer la contribution des joueurs, leur efficacité sur le terrain, leur constance, et bien plus encore. Des joueurs emblématiques tels que Lionel Messi et Cristiano Ronaldo ont continué à dominer les classements statistiques, tandis que de nouveaux talents ont émergé pour laisser leur empreinte dans l'histoire du football.

À travers ce projet, plongeons dans le monde fascinant des statistiques de football, où chaque chiffre raconte une histoire et chaque joueur laisse sa marque sur le terrain de jeu, en explorant les statistiques des joueurs de football de 2016 à 2020 avec un regard analytique. En nous appuyant sur des sources fiables et des données vérifiées, nous explorerons différentes variables qui influent sur la performance des joueurs, nous tenterons de dégager la liste des buteurs des 5 grandes ligues européennes de 2016-2023 et en fin nous allons prédire l'évolution des buts des jouurs les années à venir. Pour ce faire, nous disposons d'un jeux de données prenant source du site internet sofascore.com. Nous nous demanderons dans ce contexte:

Quelles sont les variables qui ont une influencent sur la performance des joeurs ? Qu'est-ce qui seront les potentiels meilleurs buteurs de chaque championnat l'année à venir ?

Chapitre 1

Contexte, définition et intérêt du projet

1.1 Enoncé du problème

Au cours des dernières années, l'analyse statistique des performances des joueurs de football est devenue un outil incontournable pour les entraîneurs, les analystes et les passionnés du sport. Cependant, malgré l'abondance de données disponibles, il reste des questions cruciales à explorer pour mieux comprendre les facteurs qui influent sur la performance des joueurs et pour anticiper les tendances à venir. Dans cette optique, notre projet vise à répondre aux questions suivantes : Quelles sont les variables qui exercent une influence significative sur la performance des joueurs de football de 2016 à 2023 ? Quels sont les potentiels meilleurs buteurs de chaque championnat des 5 grandes ligues européennes pour les 5 prochaines années à venir ?
En utilisant un ensemble de données provenant du site internet sofascore.com, nous nous attacherons à analyser en profondeur les statistiques des joueurs de football sur la période de référence. Nous chercherons à identifier les variables clés qui sont corrélées à la performance des joueurs, telles que les passes décisives, les tirs au but, les interceptions. De plus, nous déploierons des techniques de modélisation prédictive pour anticiper les performances des joueurs dans les années à venir. En nous basant sur les tendances passées et les caractéristiques des joueurs, nous tenterons de prédire les meilleurs buteurs potentiels de chaque championnat pour les années à venir.
Autant de questionnements qui suscitent l'intérêt de notre thème "ETUDE DE LA PERFORMANCE DES JOUEURS DES 5 GRANDES CHAMPIONS EUROPÉENNES DE 2016-2020"

1.2 Interêt de l'étude

Lintérêt de cette étude réside dans sa capacité a apporter une contribution significative a la compréhension du football moderne. Notre projet sarticule autour de lanalyse statistique approfondie des performances des joueurs de football, avec pour objectif de dégager des insights pertinents pour les professionnels du sport et les passionnés, tout en anticipant les évolutions futures du jeu. En examinant de près les données des cinq dernières saisons, nous pourrons identifier les tendances émergentes, les schémas de jeu efficaces, et les caractéristiques des joueurs qui influent le plus sur les résultats des matchs. Ces données pourraient non seulement être précieux pour les entraîneurs et les équipes dans leur prise de

décision tactique, mais également pour les analystes, les médias et les amateurs de football qui cherchent a approfondir leur compréhension du sport et a anticiper les développements futurs. En outre, en prévoyant les performances des joueurs pour les saisons a venir, notre étude pourrait également offrir un aperçu des joueurs prometteurs a surveiller et des équipes qui pourraient se démarquer dans le paysage du football mondial. En définitive, cette analyse statistique exhaustive vise a enrichir la conversation sur le football en fournissant des données tangibles et des perspectives éclairantes qui peuvent informer les décisions stratégiques et éclairer les débats passionnés qui animent le monde du sport.

1.3 Composition et responsabilité de l'équipe

- **CHIRINA Rania** :
- **GONZALEZ Emanuel** :
- **MAKHOUFI KHALIL** : Mise en place de la relation client-serveur et des divers outils de connexion du site, conception des pages dédiées aux joueurs (Favoris, etc.) et à la recherche de nos attaquants (dataList), représentation graphique des prédictions en Data Science pour chaque joueur. Intégration et fusion des pages du site une fois celles des autres membres réalisés.
- **MARTIN Samuel** : Mise en place de la base de létude et constituion des défferentes table de notre étude dans SQL
- **VITOFFODJI Adjimon Jerôme** : Conception des différentes pages dédiées aux différents championnat, y compris la représentations graphiques sur chaque page. Le style (CSS) du site a été faite par moi. Toute la partie Data Science a été faite par moi. La rédaction du rapport et la rédaction du beamer pour la présentation ont été également faite par moi.

1.4 Objectif, l'imites de l'étude et Logiciel

1.4.1 Objectif de l'étude

L'objectif général de cette étude est de mener une analyse approfondie des performances des joueurs de football de 2016 a 2023, en se concentrant sur les principales ligues européennes. De façon spécifique il s'agit particulièrement

- d'examiner les données statistiques des joueurs pour identifier les variables clés qui influent sur leur performance individuelle et collective.
- anticiper les tendances futures en matière de performances des joueurs, en utilisant des techniques de modélisation prédictive basées sur les données
- prédire les potentiels meilleurs buteurs de chaque championnat pour les 5 prochaines années à venir, en tenant compte des caractéristiques des joueurs, des stratégies des équipes et des dynamiques du jeu.

1.4.2 Méthodologie de l'étude

- Test de linéarité diagramme de dispersion
  Dans la première étape de l'analyse explicative, nous allons réaliser un diagramme de dispersion ou graphique nuage de points pour débuter l'étude de projet. Après l'avoir réalisé, la forme du nuage des points renseigne a partir d'un simple coup sur le type d'une éventuelle liaison entre X et Y.
  L'analyse du plot donne certes une idée sur le sens et le type d'association entre X et Y, mais elle ne permet pas de quantifier son intensité.

- Test de normalité de Shapiro-Wilk
  Dans la seconde étape de notre analyse explicative, nous procéderons a une vérification de la normalité des variables. Cela est nécessaire parce que l'estimation par la méthode de régression linéaire multiple peut donner des résultats qui font croire faussement qu'une telle relation existe et quelle est importante (R2 élevé, coefficients significatifs.). C'est le phénomène connu sous le nom de régression fallacieuse. (Cedrick Tombola M., 2012).

  Il y a Trois types de tests de normalité : les tests de Jarque - Bera, éfficace quand la taille de l'échantillon est très importante, le test de Shapiro - Wilk si taille de l'échantillon est inféreur ou égale à 2000, le test de K2 d'Agostino - Person, si la taille de l'échantillon est petite, pour lesquels l'hypothèse nulle H0 est que la variable est normale. Celui qui sera utilisé dans le cadre de cette étude est le test de Shapiro-Wilk .

  La statistique du test de Shapiro-Wilk est calculée comme suit :

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  où :
  - W est le statistique de test de Shapiro-Wilk,
  - n est la taille de l'échantillon,
  - $x_{(i)}$ est le i-ème plus petit élément de l'échantillon,
  - $\bar{x}$ est la moyenne de l'échantillon,
  - a_i sont les coefficients de la transformation de la moyenne et de la variance ajustées pour les échantillons de taille n.

  Test de normalité de Shapiro-Wilk :

  Les hypothèses sont les suivant:

$$H_0 : \text{Normalité des erreurs}$$
$$H_1 : \text{Non normalité des erreurs}$$

  Règle de décision : lorsque la p-value est inférieure à 5%, on conclut qu´il n'y a pas assez dévidence statistique pour accepter l'hypothèse nulle de normalité.

- Coefficient de corrélation de rang de Spearman ρ_{XY}
  Le coefficient de corrélation de Spearman, noté ρ_{XY} est un coefficient non paramétrique qui quantifie, comme le r_{XY} de Bravais Pearson, le degré d'association linéaire entre deux variables quantitatives. Il est particulièrement approprié lorsquau moins une de deux variables X et Y n'est pas normalement distribuée. Son calcul nécessite que les données soient transformées en rang. Le rang de X est noté par Ri et celui de Y par Si. Le ρ_{XY} de Spearman nest rien d'autre que le rapport entre la covariance (Ri, Si) et le produit non nul de leurs écarts-types. Il est donc un cas particulier du coefficient de corrélation de Bravais Pearson. En tenant compte de certaines propriétés de rang, le ρ_{XY} de Spearman peut être calculé de manière plus simple par la formule : La formule de la corrélation de Spearman (ρ_{XY}) est définie comme suit :

$$\rho_{XY} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  où d_i est la différence des rangs des observations X et Y, et n est le nombre total d'observations.

  En résumé, selon (Cédrick Tombola M., 2012), lestimation dun coefficient de corrélation suivra toujours [sauf indication contraire], dans lordre, les cinq étapes suivantes :
  (i) Test de linéarité [utiliser un diagramme de dispersion]
  (ii) Test de normalité [choisir le plus approprié connaissant n]
  (iii) Choix et estimation d'un coefficient de corrélation
  (iv) Test de significativité statistique sur le coefficient calculé
  (v) Interprétation ou signification clinique du coefficient estimé [valable seulement si H0 est rejetée]

- Estimation du modèle
  Une fois les tests préliminaires effectués, on peut procéder aux estimations en se servant du modèle adapté. Nous estimerons un modèle de regression multiple. Le modèle de régression linéaire multiple n'est quune extension du

  modèle de régression linéaire simple au cas multivarié dans lequel interviennent plusieurs variables exogènes dans l'explication du phénomène étudié. On parle aussi de modèle de régression linéaire général ou standard pour souligner que ce modèle reste valable quel que soit le nombre d'exogènes qui s'y figurent. Dans sa forme générale, il s'écrit de la sorte :

$$\begin{aligned} But = \beta_0 + \beta_1 Tirs\_Total + \beta_2 Gross\_occas + \beta_3 tirs\_cadrs + \beta_4 matchs \\ + \beta_5 minutes\_joues + \beta_6 Ratio\_but\_par\_match + \beta_7 id\_nom + \varepsilon \end{aligned}$$

  où :
  - But, est la variable à prédire,
  - $Tirs\_Total$, Tirs total pendant une saison,
  - $Gross\_occas$, Grosse occasion crée pendant une saison,

- $tirs\_cadrs$, Tirs total cadré pendant une saison,
- $matchs$, Nombre de match joué dans la saison,
- $minutes\_joues$, Nombre de minutes jouées dans la saison,
- $Ratio\_but\_par\_match$, Le ratio but par match,
- id_nom, l'identifiant unique associé à chaque joueur,
- β_0 est l'ordonnée à l'origine (la constante C),
- $\beta_1, \beta_2, \ldots, \beta_7$ sont les coefficients des variables indépendantes,
- ε est l'erreur résiduelle.

Modèle 1	$But = \beta_0 + \beta_1$Tirs_Total + β_2Gross_occas + $\beta_3$tirs_cadrés +$\beta_4$matchs + $\beta_5$minutes_jouées+ β_6Ratio_but_par_match + $\beta_7$id_nom + ε
Modèle 2	$But = \beta_0 + \beta_3$tirs_cadrés + $\beta_4$matchs + $\beta_5$minutes_jouées+ +β_6Ratio_but_par_match + $\beta_7$id_nom + ε

Table 1.1: Spécification des modèles à estimer

- Tests de validation du modèle

  Les résultats obtenus à la suite des estimations ne peuvent toutefois pas être interprétés avant d'avoir subi certains tests, qui jugent de leur validité d'un point de vue statistique. Ce sont des tests de validation. Ils portent dune part, sur la significativité du modèle estimé et d'autre part, sur les propriétés des résidus du modèle. Ces tests sont réalisés au seuil de 5% avec le logiciel R version 3.5.3.

  – **Test de normalité des résidus (test de Shapiro-Wilk)** Ce test évalue si les résidus du modèle suivent une distribution normale.

    Règle de décision : lorsque la p-value est inférieure à 5%, on conclut qu´il n'y a pas assez dévidence statistique pour accepter lhypothèse nulle de normalité des erreurs..

  – **Test d'homoscédasticité de Breusch-Pagan-Godfrey :** Ce test cherche à déterminer la nature de la variance du terme derreur. Si la variance est une constante, alors on parle d'homoscédasticité ; en revanche si elle varie on parle d'hétéroscédasticité.
    Les hypothèses sont les suivantes :

    H_0: Homoscédasticité
    H_1: Hétéroscédasticité

    Règle de décision: : lorsque la p-value est inférieure à 5%, on conclut qu´il n'y a pas assez dévidence statistique pour accepter lhypothèse nulle dhomoscédasticité des erreurs.

  – **Test de linéarité :** Ce test évalue si la relation entre les variables indépendantes et la variable dépendante est linéaire. Il peut être effectué à l'aide de graphiques de dispersion et de graphiques résiduels. De plus, les graphiques résiduels peuvent également être utilisés pour détecter des schémas non linéaires dans les résidus du modèle. Si les graphiques de

dispersion ou les graphiques résiduels révèlent des schémas non linéaires, cela suggère que le modèle linéaire multiple pourrait ne pas être approprié pour vos données.

- **Test dautocorrélation de Breusch-Godfrey :** Ce test vérifie l'absence de corrélation entre les résidus. La corrélation entre les résidus peut indiquer que le modèle ne capture pas correctement toute la structure de dépendance dans les données. Nous testerons l'autocorrélation dordre supérieur 1 avec le test du multiplicateur de Lagrange LM. Les hypothèses sont les suivantes :

  H_0: $\phi_1 = \phi_2 = 0$ (Absence d'autocorrélation)

  H_1: $\phi_1 \neq 0$ ou $\phi_2 \neq 0$ (Présence d'autocorrélation)

  Règle de décision: : lorsque la p-value est inférieure à 5%, on conclut qu´il n'y a pas assez dévidence statistique pour accepter lhypothèse nulle dhomoscédasticité des erreurs.

- **Test de Studen (Test de significativité individuelle des coefficients):** Ce test évalue si chaque coefficient de régression est significativement différent de zéro. Il est généralement réalisé à l'aide de tests t individuels pour chaque coefficient.
  Les hypothèses sont les suivantes :

  H_0 : $\alpha_i = 0$ avec i = 1,2,3,4,

  H_1 : $\alpha_i \neq 0$ avec i = 1,2,3,4,5

  Règle de décision: lorsque la probabilité de la statistique du test de Student est inférieure à 5%, on conclut quil ny a pas assez dévidence statistique pour accepter lhypothèse nulle (H0). Donc la variable associé est importante dans la prédiction du modèle.

- **Test de Fisher (Test de significativité globale du modèle):** Ce test examine si le modèle dans son ensemble est significatif. Ce test fait partie des sorties automatiques fournies par le logiciel R version 3.5.3.
  Les hypothèses du test sont :

  $H_0 : \beta_i = 0$ avec $i = 1, ..., k$

  H_1 : au moins un des $\beta_i \neq 0$ avec $i = 1, ..., k$

  Règle de décision: Si $F > F[(K1); (nK)]$ [valeur lue dans la table de Fisher, au seuil de 5on rejette H0, le modèle est bon.

- Prévision dans le modèle de régression linéaire multiple
  Le modèle de prévison à pour équationde régression suivante de :

$$\hat{But}_t = \hat{\beta}_0 + \hat{\beta}_1 tire_c adrs_t + \hat{\beta}_2 matchs_t + \hat{\beta}_3 minutes_j oues_t + \hat{\beta}_4 Ratio_b ut_m atch_t + \hat{\beta}_5 id_n om_t + \epsilon_t$$

  où :
  $\hat{But}_t$: Valeur prédite de la variable dépendante But à l'observation t.
  - $tire_c adrs_t$: Valeurs observées de la variable indépendante $tirs\_cadrs$ à l'observation t.

- $matchs$: Valeurs observées de la variable indépendante $matchs$ à l'observation t.
- $minutes\_joues$: Valeurs observées de la variable indépendante $minutes\_joues$ à l'observation t.
- $Ratio\_but\_par\_match$, : Valeurs observées de la variable indépendante $Ratio\_but\_par\_match$ à l'observation t.
- id_nom : Valeurs observées de la variable indépendante id_nom à l'observation t.
- $\hat{\beta}_0$: Intercept, qui représente la valeur de But lorsque toutes les variables indépendantes sont nulles.
- $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_5$: Coefficients de régression estimés associés à chaque variable indépendante $But, tirs_cadrs, matchs, minutes_joues, Ratio_buts_par_match, id_nom$.
- ϵ_t : Terme d'erreur, qui représente la différence entre la valeur observée de But et la valeur prédite $\hat{But}_t$ à l'observation t.

1.4.3 Limtes de l'étude

- Cette étude n'a pas la prétention d'aborder tous les aspects du problème de prévision du nombre de but que marquera chaque joueurs. La théorie économique nous renseigne sur l'existence de plusieurs variables collectées sur une longue période pouvant influencer le nombre de but et certaines ont été mises en évidence. La principale limite de notre étude est le fait que nous n'ayons pas pris en compte toutes les données qui sont susceptibles d'améliorer notre modèle. En effet, au cas où une base de données sur ce secteur existerait, sa prise en compte pourrait élevé le R2 du modèle et ainsi, nous permettre de faire une bonne prévision.
- Nos données ne suivent pas une distribution normale, ce qui nous permet de toujours obtenir des estimations des coefficients de régression et des statistiques d'inférence associées. Cependant, l'interprétation des résultats doit être faite avec prudence et considération des limitations. Les estimateurs de régression peuvent être robustes aux violations mineures des hypothèses. Cependant, si les violations sont graves, les résultats de la régression peuvent être biaisés ou inefficaces.
- Les différents test de validation du modèle, nous montre que notre modèle est hétéroscédastique. Ce qui peut être problématique car les estimations des coefficients peuvent être biaisées. Ainsi si nous devons envisager d'inclure plus de variable dans modèle ou des méthodes plus robustes pour remedier à ça, car les différentes techniques de transformation des variables ne nous a pas permis d'améliore la qualité des estimations de notre modèle et de garantir des résultats fiables.

1.4.4 Logicieles

Dans la réalisation de notre projet, nous avons utilisé le logiicel excel(2016) pour constituer notre base. Ensuite nous avons importé cette base au format csv sous Rstudio pour la réalisation de nos différents graphes. Nous avons ensuite utilisé les packages de R pour réaliser notre étude.

Pour la rédaction de notre projet, nous avons utiliser la version 28-1-2 du logiciel le Rstudio sous Mac pro à l'aide de R Markdown pour génerer directement notre document en format pdf. L'analyse descriptive a été faite aussi avec ce dernier.

Chapitre 2

PRESENTATION ET INTERPRETATION DES RESULTATS

2.1 Nature, Source des données et description des tables

2.1.1 Nature et source

Les données utilisées dans la présente étude proviennent d'Internet en licence libre et exploitable. En effet, ne nous somme par les auteurs de toutes ses données, ainsi donc pour permettre aux autres utilisateurs à consulteurs ou utiliser ses données dans leurs différents travaux, nous mettrons dans la bibliographie avec plus de précision les différentes sources de ses données.

Ces données proviennent de :

https://www.sofascore.com/fr/.
Cette dernière est constituée de 10 colonnes et 801 lignes.

2.1.2 Description des tables

Les principales variables de notre jeu de données sont :

Nom	**Description**	**Variable**
Nom	Nom du joueur	**Année**
Match	Nombre de match jouer	**matchs**
Minutes jouées	Le nombre de minutes joué pendant la saison	**minutes jouées**
Buts	Nombre de but marqué	**Buts**
Tirs total	Le nombre de tire tanté pendant la saison	**Tirs Toatal**
Tirs cadré	Le nombre de tire cadré pendant une saison	**tires cadrées**
Ratio	Le ratio de but par match	**Ratio_but_match**
Grosses occasions	Grosses occasions créées durant la saison	**Gross ocas**
ID NOM	Identifiant associé à chaque joueur	**Id_nom**

Table 2.1: Tableau récapitulatif des variables de l'étude

2.1.3 Vision générale des données

Identifiants	Variables quantitative	Variables qualitative
1- Unique	Discretes : 9	Norminal: 0

Tableau: Description des types d'attribut.

2.2 Analyse descriptive de la population de l'étude

2.2.1 Analyse du nuage de point entre la variable But et la variable Tire total

L'analyse de ce graphique révèle qu'il esiste une relation linéaire entre le nombre de but et le nombre total de tirs. Nous observons quelques valeurs aberrantes. Ainsi dans la suite de notre étude nous prendrons en compte la présence de ses variables afin d'adapté les tests nécessaires dans notre étude.

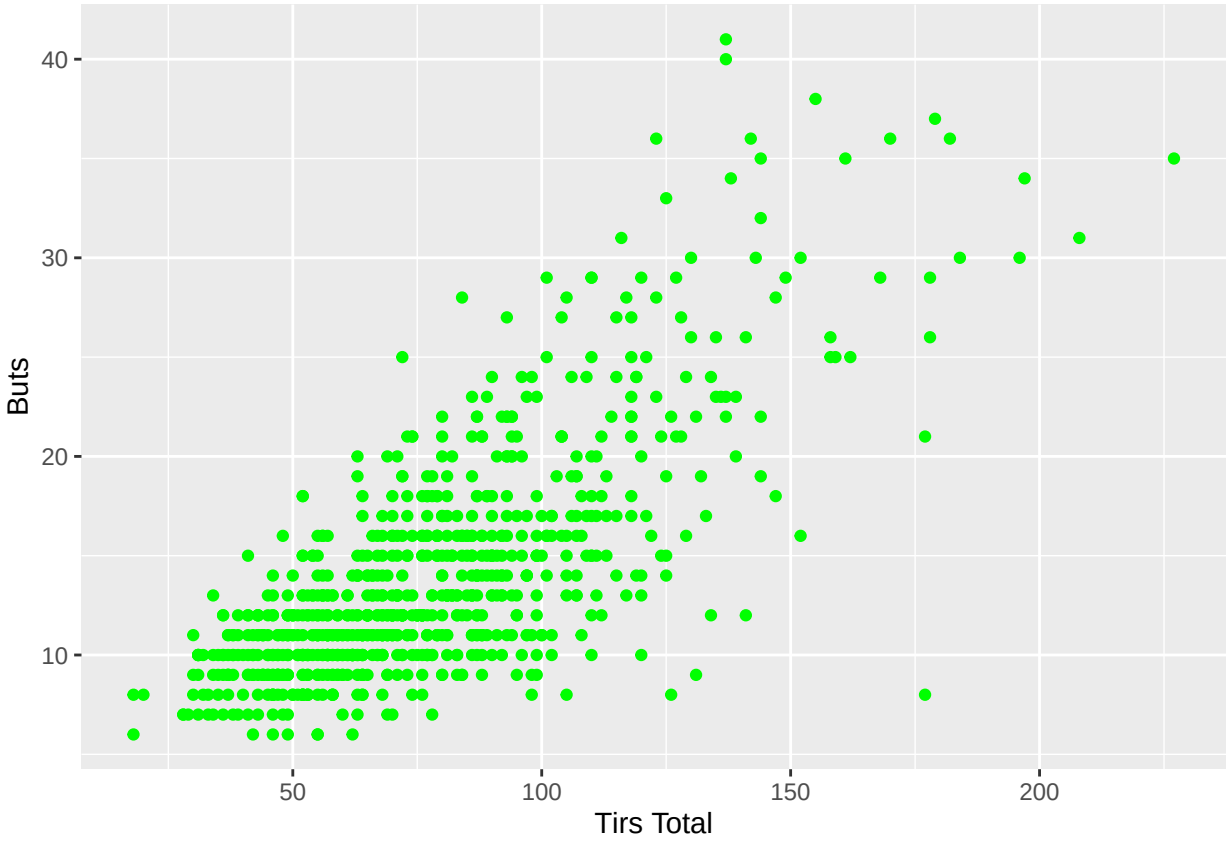

Figure 2.1: Relation entre la variable But et Tirs total

2.2.2 Analyse du nuage de point entre la variable But et la variable tirs cadrés

L'analyse de linéarité entre le nombre de but marqué et le nombre de tirs cadrés nous révèle qu'il esiste une relation linéaire entre le nombre de but et le nombre de tirs cadrés. Nous observons quelques valeurs aberrantes. Ainsi dans la suite de notre étude nous prendrons en compte la présence de ses variables afin d'adapté les tests nécessaires dans notre étude.

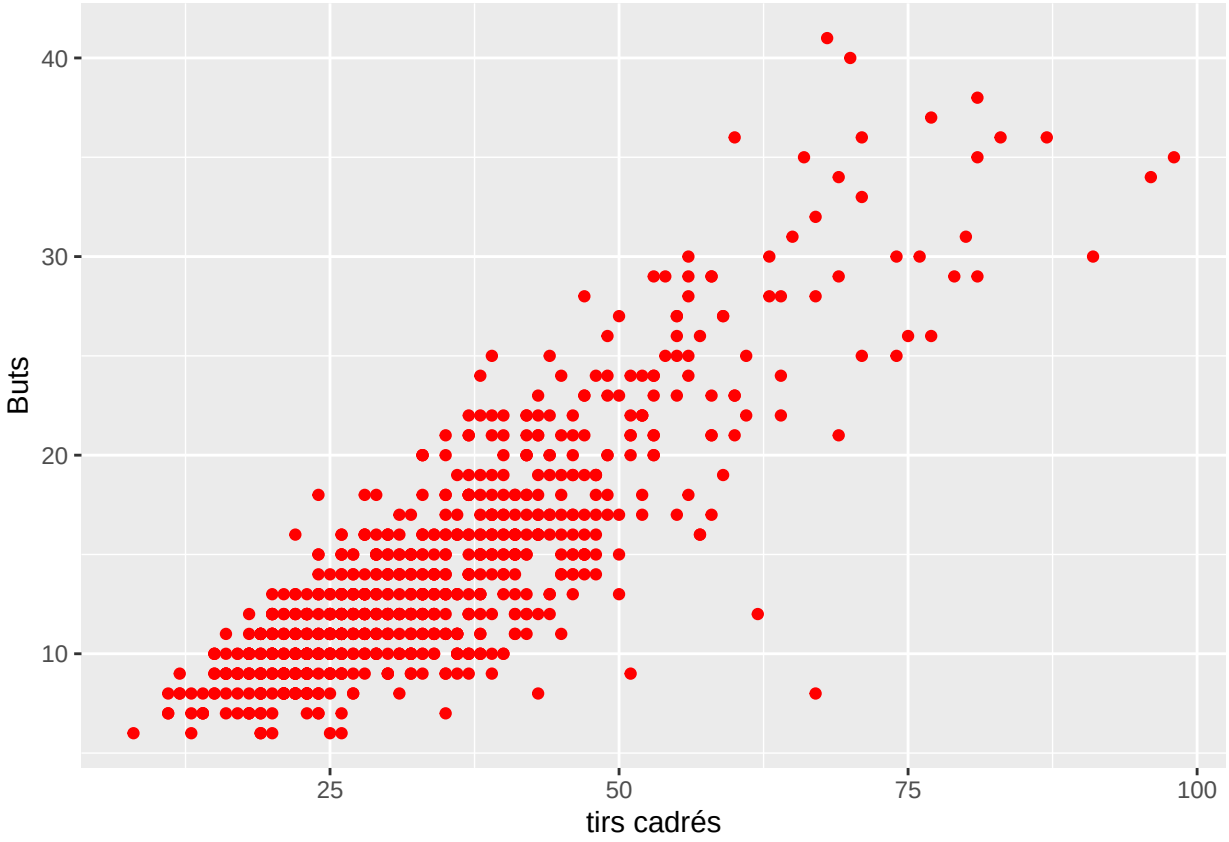

Figure 2.2: Relation entre la varible But et tirs cadrées

2.2.3 Analyse du nuage de point entre la variable But et la variable Grosse occasion

De l'analyse de ce graphique nous ne somme pas en mesure d'affirmer ou non la présence de linéarité entre le nombre de but marqué et le nombre d'occasion crées.Ainsi dans la suite de notre étude, nous utiliserons les méthodes adéquates pour estimer les paramètres de notre modèle.

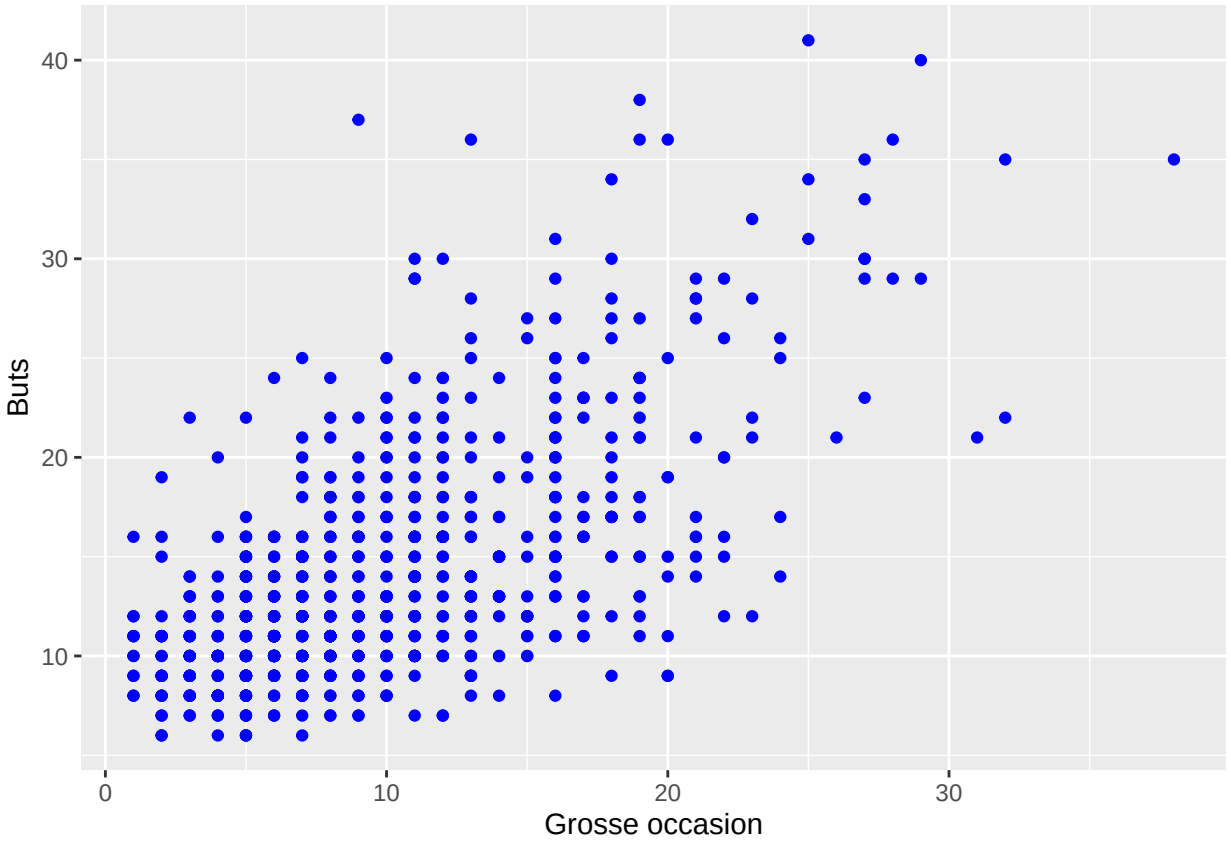

Figure 2.3: Relation entre la varible But et Grosse occasion

2.2.4 Analyse du nuage de point entre la variable But et la variable minutes_jouées

De l'analyse de ce graphique nous ne somme pas en mesure d'affirmer ou non la présence de linéarité entre le nombre de but marqué et le nombre de minute jouées.Ainsi dans la suite de notre étude, nous utiliserons les méthodes adéquates pour estimer les paramètres de notre modèle.

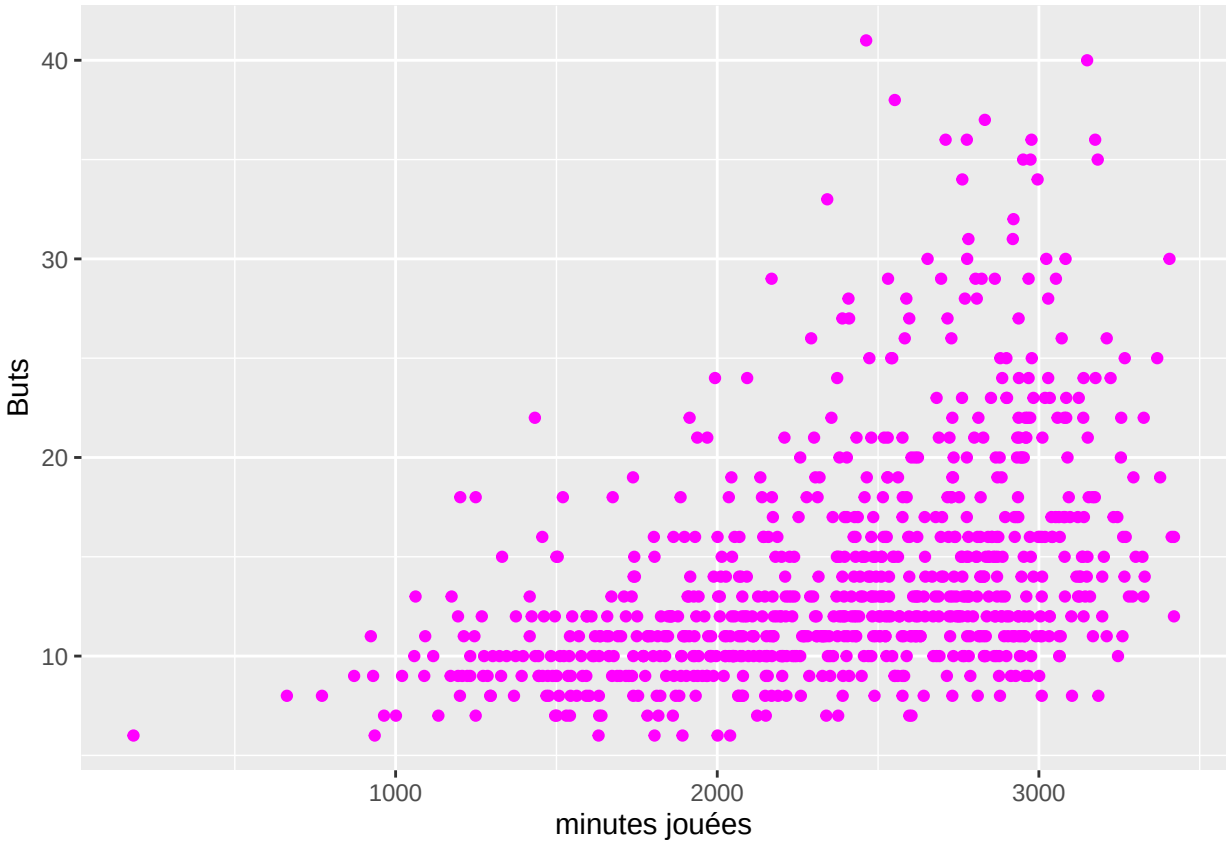

Figure 2.4: Relation entre la varible But et minutes_jouées

2.2.5 Test de normalité des varaibles

- But

  De ce tableau il ressort que la valeur de la p-value est inférieur au seuil de signification $\alpha = 5\%$ donc on rejette l'hypothèse nulle et on conclut que la variable But ne suit pas une distribution normale.(Voir Annexe 1).

Shapiro-Wilk normality test	
Test	Valeur
Statistique W :	0,85561
p-value	2,2e-16

Table 2.3: Test de normalité de la variable But

- Résumé du test de normalité des autres variables du modèle

  De ce tableau il ressort que les différentes valeur de la p-value sont toutes inférieur au seuil de signification $\alpha = 5\%$ donc on rejette l'hypothèse nulle et on conclut que les variables Tirs_Total, Gross_occas, maths, tirs_cadrés, minutes_jouées, ratio_but_par_match ne suivent pas une distribution normale.(Voir Annexe1)

Shapiro-Wilk normality test		
Variables	Statistique W	p-value
Tirs_Total	0.94029	2.2e-16
Gross_occas	0.93052	2.2e-16
tirs_cadrés	0.91627	2.2e-16
matchs	0.89984	2.2e-16
minutes_jouées	0.96931	6.374e-12
ratio_but_par_match	0.86488	2.2e-16

Table 2.4: Test de normalité des variables de l'étude

2.2.6 ***Matrice de corrélation de Spearman entre les différentes variables***

De l'analyse de cette figure, on note qu'il y a une laision positive entre le but marqué, le nombre de tirs total, le nombre de grosses occasions crées, le nombre de tirs cadré, le nombre de matchs joué, le nombre de minutes jouées, et le ratio de but par match. De plus on observe une forte signicativité des variables aux seuil de 5%. La présence de 3 étoiles indique que la p-value est faible et inférieur à 5%. Dans ce cas, on rejète l'hypothèse nulle d'abscence de corrélation. De plus on note une très forte liaison entre le but marqué, tire cadré et le ratio de but par match. Autrement dit, plus un joueur effectue des tires cadrés sur le gardien de but de l'équipe adverse, plus ce joueur à la chance de marquer de but. On observe une forte liaison entre les variables tirs total et tirs cardé. Ainsi donc, dans l'estmation de notre modèle, nous garderons une seule des deux variables dans notre modèle. De même on note une forte relation entre la variable match et la variable minute joués. Par la suite, une seule des 2 variables sera retenues dans notre modèle. On observe aussi une liaison linéaire entre le nombre de but, total de tirs, grosse occasion, tirs cadrés, ratio but par match, et une liaison non linéaire monotone entre le nombre de but, match, minutes jouées.

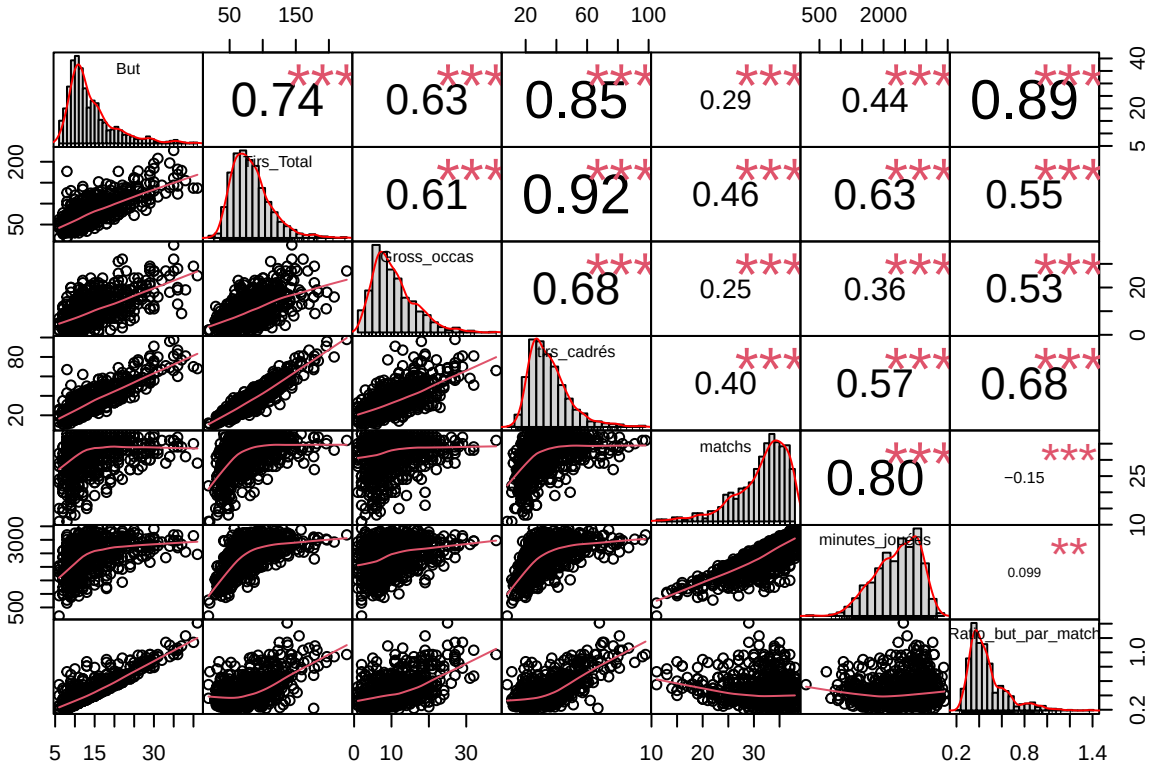

Figure 2.5: Représentation graphique de la liaison entre les variables

2.3 Modélisation

- **Estimation du modèle**

  Le tableau si dessous nous présente les résultats de l'estimation du modèle de regression linaires multiple. De n'analyse de cet tableau, il ressort les varaibles Tirs total et Grosses occasions crées ne sont pas significatives. Ainsi nous allons les enlever de notre modèle. *** et * indiquent que les variables sont significatives au seuil de 1/1000 et 5% respectivement. (Annexe2)

	But	
	T-Statistics	Coefficients
Tirs_Total	-1,050	0,2939
Gross_occas	-1,050	0,2695
tirs_cadrés	4,974	**8,05e-07 *** **
matchs	43,464	**2e-16 *** **
minutes_jouées	-2,346	**0,0192 ***
ratio_but_par_match	99,589	**2e-16 *** **
Id_Joueur	1,542	0,1235
C	-54,704	**2e-16 *** **
R^2**:**	0,9803	
F-statistic:	5680	
p-value:	2,2e-16	

Table 2.5: Résultats de l'application de la Regression linéaire multiple

- **Estimation du modèle retenu**

  Le tableau si dessous nous présente les résultats de l'estimation du modèle de regression linaires multiple retenu Seul les variables significatives sont retenu dans notre modèle.(Annexe3) *** et ** indiquent que les variables sont significatives au seuil de 10% et 1/1000 respectivement.

	But	
	T-Statistics	Coefficients
tirs_cadrés	3,116	**0,001899 ** **
matchs	-17,277	**2e-16 *** **
minutes_jouées	-3,497	**0,000496 *** **
ratio_but_par_match	-34,402	**2e-16 *** **
Id_Joueur	1,223	0,221614
C	61,631	**2e-16 *** **
R^2**:**	0,8353	
F-statistic:	811,6	
p-value:	2,2e-16	

Table 2.6: Résultats de lapplication de la Regression linéaire multiple du modèle retenu

Exemple de panel

- **Test de validation du modèle**
  - **Test de normalité des résidus**

    L'analyse du test de normalité de Shapiro-Wilk sur les résidus du modèle (Voir Annexe), On conclut que les résidus ne sont pas normalement distribués. Cette non normalité peuvent s'expliquer d'une part par la posibilité qu'il a des biais dans les données non capturé par

le modèle. Cela implique qu'il a des variables omises, qui peuvent apporter plus dinformations pour notre modèle. Et d'autre part, le modèle est trop simple pour capturer la structue réelle des données. Pour cela il est nécessaire d'ajuster le modèle en ajoutant des termes à notre modèle.
La non normalité des résidus peut entraîner un biais dans les estimations des paramètres du modèle. Ce qui peut fausser les prédictions et conduire à des conclusions erronnées ou falsieuse.(Voir Annexe4)

- **Test d'homoscédasticité**
  L'analyse du test d'homoscédasticité de Breusch-Pagan sur le modèle (Voir Annexe), nous révlève qu'on a pas assez d'évidence pour accepter l'hypothèse nulle d'abscence d'hétéroscédasticité. Donc on conclut que notre modèle est hétéroscédastique.
  Autrement dit, la variance des erreurs de régression n'est pas constante pour toutes les valeurs des variables indépendantes. En d'autres termes, la dispersion des résidus varie selon les niveaux des variables explicatives. Cela peut entraîner des estimations inefficaces des coefficients de régression et des intervalles de confiance incorrects. Ce qui peut être problématique car les estimations des coefficients peuvent être biaisées et les intervalles de confiance peuvent être trop larges ou trop étroits.(Voir Annexe5)

- **Test d'autocorrélation**
  L'analyse du test d'autocorrélation de Durbin-Watson sur le modèle (Voir Annexe), nous révèle qu'on a pas assez d'évidence statistique pour accepter l'hypothèse nulle d'abscence d'autocorrélation . Donc on conclure que les résidus de notre modèle sont autocorrélés.
  Les résidus consécutifs sont corrélés positivement car $\mathbf{DW = 0.73888}$ est comprise entre **0** et **2**, ce qui peut indiquer que le modèle sous-estime ou que des variables pertinentes sont manquantes.(Voir Annexe6)

- **Test de significativité globale du modèle**
  L'analyse du test significativité globale du modèle (Voir Annexe 6), nous révèle qu'on a pas assez d'évidence statistique pour accepter l'hypothèse nulle. Donc on conclut que le modèle est significatif (2,2e-16) $< 5\%$. Autrement dit, cela indique le modèle a un pouvoir préditif significatif.
  le test de significativité globale du modèle permet de conclure que le modèle dans son ensemble est capable d'expliquer de manière significative la variance de la variable dépendante.

- **Test de significativité individuelle des coefficients**
  L'analyse du test significativité individuelle des coefficient du modèle 2 de regression linéaire multiple (Voir Annexe3) , nous révèle que tout nos variable sont significatives.

2.4 Prévisions

Le tableau si dessous nous présente les résultats de modèle en 2023 qui nous à servir de faire la prévision la prévison.(Voir Annexe7) *** indiquent que les variables sont significatives au seuil de 1/100 et sans étoiles indiquent ces variables ne sont pas significatives. Autrement dit, ces variables n'apportent pas d'information pertinente dans la prévision de la variable dépendante But.

But		
	T-Statistics	Coefficients
tirs_cadrés	1,568	0.120
matchs	24,086	**2e-16 *** **
minutes_jouées	-1,485	0.141
ratio_but_par_match	49,274	**2e-16 *** **
Id_Joueur	-0,509	0,612
C	-29.684	**2e-16 *** **
R^2**:**	0,9882	
F-statistic:	1654	
p-value:	2,2e-16	

Table 2.7: Résultats du modèle retenu pour la prévision de la variable But

En supponsant que les caractériques des variables exogènes de chacun des joueurs en 2023 sont resté les même pour en 2024, la valeur prédicte du la variable andogène pour 5 joueurs sont:

Prévisions 2024		
id_nom	**nom**	**But**
68	Erling Haaland	35
1	Harry Kane	29
61	Ivan Toney	20
30	Mohamed Salah	19
41	Callum Wilson	18

Table 2.8: Résultats de prévision de 5 joueurs

2.5 Difficultés rencontrés

La réalisation du projet a impliqué plusieurs défis. Initialement, la recherche d'une base de données appropriée a demandé du temps. Générer des idées pour créer des figures informatives a été une tâche tres reflexive. Connecter de manière fluide toutes les figures pour assurer la cohérence du rapport a été la tâche la plus importante. Sélectionner les variables pertinentes dans la base de données a

nécessité une analyse approfondie. Enfin, la manipulation des chunks dans R Markdown pour une présentation adéquate du code et des résultats a été complexe.

CONCLUSION

À travers une analyse statistique approfondie, ce projet vise à fournir des insights pertinents sur la performance des joueurs de football et à prédire les tendances futures en matière de buts. Les résultats obtenus pourraient avoir des implications importantes pour les entraîneurs, les analystes et les passionnés de football, en aidant à prendre des décisions tactiques éclairées et à anticiper les évolutions du jeu.

Références bibliographiques

- Atchadé, M.N. et Eliseeva, I.I. (2017). Statistique. Analyse Statistique : mise en uvre dans le programme R. ISBN-978-5-4391-028-6, 80p
- Jéhovahni G.B.M. SODJINOU & Jérôme A.G. VITOFFODJI: MEMOIRE DE FIN DE FORMATION POUR LOBTENTION DU DIPLOME DE TECHNICIEN SUPERIEUR (DTS): Etude de la demande dessence dans les stations-services de la SONACOP de 2012 à 2017 : approche économétrique
- Cours et exercices corrigés: Régis Bourbonnais 9è édition
- Econométrie 1 Rappels et recueils d'exercices[résolus]: Cédrick To mbola M./Assistant.
- OpenAI(chatGpt)

Annexes

Annexe 1 : Test de normalité sur chaque variable

```
##
##  Shapiro-Wilk normality test
##
## data:  Base1$But
## W = 0.85561, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  Base1$Tirs_Total
## W = 0.94029, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  Base1$Gross_occas
## W = 0.93052, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  Base1$tirs_cadrés
## W = 0.91627, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  Base1$matchs
## W = 0.89984, p-value < 2.2e-16
##
##  Shapiro-Wilk normality test
##
## data:  Base1$minutes_jouées
## W = 0.96931, p-value = 6.374e-12
##
##  Shapiro-Wilk normality test
##
## data:  Base1$Ratio_but_par_match
## W = 0.86488, p-value < 2.2e-16
```

Annexe 2: Sortie du Modèle 1 de regression

```
##
## Call:
## lm(formula = But ~ Tirs_Total + Gross_occas + tirs_cadrés +
##     matchs + minutes_jouées + Ratio_but_par_match + Base$id_nom,
##     data = Base1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5599 -0.3495  0.0740  0.4072  2.7246
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.499e+01  2.740e-01 -54.704  < 2e-16 ***
## Tirs_Total          -2.830e-03  2.695e-03  -1.050   0.2939
## Gross_occas          7.648e-03  6.922e-03   1.105   0.2695
## tirs_cadrés          3.646e-02  7.330e-03   4.974 8.05e-07 ***
## matchs               4.765e-01  1.096e-02  43.464  < 2e-16 ***
## minutes_jouées      -2.304e-04  9.819e-05  -2.346   0.0192 *
## Ratio_but_par_match  2.981e+01  2.993e-01  99.589  < 2e-16 ***
## Base$id_nom          4.326e-04  2.806e-04   1.542   0.1235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8189 on 792 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9803
## F-statistic:  5680 on 7 and 792 DF,  p-value: < 2.2e-16
```

Annexe 3: Sortie du Modèle 2 de regression

```
##
## Call:
## lm(formula = But ~ tirs_cadrés + matchs + minutes_jouées +
##     Ratio_but_par_match + Base$id_nom, data = Base1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5743 -0.3535  0.0738  0.3980  2.7388
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.502e+01  2.733e-01 -54.946  < 2e-16 ***
## tirs_cadrés          3.232e-02  4.226e-03   7.648 5.91e-14 ***
## matchs               4.781e-01  1.091e-02  43.819  < 2e-16 ***
## minutes_jouées      -2.578e-04  9.569e-05  -2.694  0.00721 **
## Ratio_but_par_match  2.991e+01  2.927e-01 102.185  < 2e-16 ***
```

```
## Base$id_nom          4.051e-04  2.801e-04   1.446  0.14853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8191 on 794 degrees of freedom
## Multiple R-squared:  0.9804, Adjusted R-squared:  0.9803
## F-statistic:  7947 on 5 and 794 DF,  p-value: < 2.2e-16
```

Annexe 4: Test de normalité des résidus

```
##
##  Shapiro-Wilk normality test
##
## data:  modele2$residuals
## W = 0.88099, p-value < 2.2e-16
```

Annexe 4: Test d'homescédasticité

```
##
##  studentized Breusch-Pagan test
##
## data:  modele2
## BP = 152.51, df = 5, p-value < 2.2e-16
```

Annexe 5: Test d'autocorrélation

```
##
##  Durbin-Watson test
##
## data:  modele2
## DW = 1.7201, p-value = 2.911e-05
## alternative hypothesis: true autocorrelation is greater than 0
```

Annexe 6: Test de significativité global du modèle

```
##    value    numdf    dendf
## 7947.199    5.000  794.000
```

Annexe 7: Sorie des 5 valeures prédite

```
##  1  2  3  4  5  6
## 35 29 20 19 18 15
```