https://doi.org/10.26219/heal.aueb.6479

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΜΕΤΑΠΤΥΧΙΑΚΟ ΣΤΑΤΙΣΤΙΚΗ
MSc IN STATISTICS

Use of Linear and Multinomial Models in Football Transfers

By

Stavros Dimitrios Faliagkas

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

August 2024

https://doi.org/10.26219/heal.aueb.6479

https://doi.org/10.26219/heal.aueb.6479

ACKNOWLEDGEMENTS

Dear Sirs/Madams,

With the completion of my thesis, I would like to express my sincere thanks to all of you who have contributed to the completion of this course. First of all, I want to thank my family for their constant support and understanding during this journey of my thesis. I also want to express my sincere thanks to my professors, especially my supervisor Panagiotis Besbeas, who with their knowledge, experience and guidance guided me with diligence and dedication to achieve my goal. Finally, I could not fail to thank my friend and classmate, Eythimios Kavour whose support and positive energy helped me maintain my enthusiasm and determination throughout this course and Stylianos Vradelis, CEO and Chairman of the Institute of Sports Analysis, who gave me access to real data that used for my thesis. I am grateful to you for your trust and support and I hope that my work will repay the trust you have placed in me.

Yours sincerely,
Stavros Faliagkas

https://doi.org/10.26219/heal.aueb.6479

https://doi.org/10.26219/heal.aueb.6479

VITA

My name is Stavros Faliagkas and I am a graduate of the Department of Statistics of Athens University of Economics and Business. At the same time I was a student of the Institute of Sports Analysis specializing in Performance Analysis. After having completed my studies I completed my military services and from 2021 I am a student of the Master's program of Applied Statistics in the Department of Statistics of Athens University of Economics and Business, working at the same time in an insurance company, as a Risk Analyst.

https://doi.org/10.26219/heal.aueb.6479

ABSTRACT

Stavros Faliagkas

USE OF LINEAR AND MULTINOMIAL MODELS IN FOOTBALL TRANFERS

August 2024

The primary aim of this thesis is to construct a sophisticated model designed to aid professional soccer teams in the selection of the most suitable player for specific positions during the transfer process. This model will analyze and evaluate a comprehensive set of data, tailored to the particular requirements and characteristics of each position. Key variables in this analysis will include the player's market value, which reflects their perceived worth in the current transfer market, as well as their age, which can influence their potential for development, experience, and career longevity.

For the purpose of this study, a sample of ten soccer players has been chosen from different European leagues. These players all occupy the same position on the field: left central defender. This position is notably specialized within the realm of professional football. The scarcity of left-footed central defenders adds a layer of complexity and importance to the recruitment process. As left central defenders play a critical role in a team's defensive strategy, their unique skill set and attributes can significantly impact overall team performance.

The thesis will delve into various performance metrics and attributes that are crucial for left central defenders. These might include defensive skills such as tackling, intercepting, and marking, as well as physical attributes like speed, strength, and agility. Additionally, technical skills such as passing accuracy, ball control, and the ability to play under pressure will be considered. The model will also take into account contextual factors like the player’s experience in top-tier leagues, their injury history, and their adaptability to different playing styles and tactical systems.

By integrating these diverse data points, the model aims to provide a holistic evaluation of each player's suitability for the left central defender position. The ultimate goal is to

https://doi.org/10.26219/heal.aueb.6479

furnish soccer teams with a powerful tool that enhances their decision-making process, allowing them to identify and acquire players who will best fit their tactical needs and strategic objectives.

In conclusion, this thesis endeavors to bridge the gap between data analytics and practical application in professional soccer. By focusing on the niche but crucial role of left central defenders, it seeks to offer valuable insights and methodologies that can be extended to other positions and aspects of player evaluation in the future. This work will not only contribute to the academic field of sports analytics but also have a tangible impact on the operational strategies of professional soccer clubs.

https://doi.org/10.26219/heal.aueb.6479

ΠΕΡΙΛΗΨΗ

Σταύρος Φαλιάγκας

ΧΡΗΣΗ ΓΡΑΜΜΙΚΩΝ ΚΑΙ ΠΟΛΥΩΝΥΜΙΚΩΝ ΜΟΝΤΕΛΩΝ ΣΕ ΜΕΤΑΓΡΑΦΕΣ ΠΟΔΟΣΦΑΙΡΟΥ

Αύγουστος 2024

Ο πρωταρχικός στόχος αυτής της διατριβής είναι η κατασκευή ενός εξελιγμένου μοντέλου σχεδιασμένου να βοηθά τις επαγγελματικές ομάδες ποδοσφαίρου στην επιλογή του καταλληλότερου παίκτη για συγκεκριμένες θέσεις κατά τη διαδικασία μεταγραφής. Αυτό το μοντέλο θα αναλύσει και θα αξιολογήσει ένα ολοκληρωμένο σύνολο δεδομένων, προσαρμοσμένων στις ιδιαίτερες απαιτήσεις και χαρακτηριστικά κάθε θέσης. Οι βασικές μεταβλητές σε αυτήν την ανάλυση θα περιλαμβάνουν την αγοραία αξία του παίκτη, η οποία αντικατοπτρίζει την αντιληπτή αξία του στην τρέχουσα μεταγραφική αγορά, καθώς και την ηλικία του, που μπορεί να επηρεάσει τις δυνατότητές του για εξέλιξη, εμπειρία και μακροζωία σταδιοδρομίας.

Για τους σκοπούς αυτής της μελέτης, επιλέχθηκε ένα δείγμα δέκα ποδοσφαιριστών από διαφορετικά ευρωπαϊκά πρωταθλήματα. Αυτοί οι παίκτες καταλαμβάνουν όλοι την ίδια θέση στο γήπεδο: αριστερός κεντρικός αμυντικός. Αυτή η θέση είναι ιδιαίτερα εξειδικευμένη στη σφαίρα του επαγγελματικού ποδοσφαίρου. Η έλλειψη αριστεροπόδαρων κεντρικών αμυντικών προσθέτει ένα στρώμα πολυπλοκότητας και σημασίας στη διαδικασία στρατολόγησης. Καθώς οι αριστεροί κεντρικοί αμυντικοί διαδραματίζουν κρίσιμο ρόλο στην αμυντική στρατηγική μιας ομάδας, το μοναδικό σύνολο δεξιοτήτων και τα χαρακτηριστικά τους μπορούν να επηρεάσουν σημαντικά τη συνολική απόδοση της ομάδας.

Η διατριβή θα εμβαθύνει σε διάφορες μετρήσεις απόδοσης και χαρακτηριστικά που είναι κρίσιμα για τους αριστερούς κεντρικούς αμυντικούς. Αυτά μπορεί να περιλαμβάνουν αμυντικές δεξιότητες όπως τάκλιν, αναχαίτιση και μαρκάρισμα, καθώς και σωματικά χαρακτηριστικά όπως ταχύτητα, δύναμη και ευκινησία. Επιπλέον, θα ληφθούν υπόψη τεχνικές δεξιότητες όπως η ακρίβεια της πάσας, ο έλεγχος της μπάλας

https://doi.org/10.26219/heal.aueb.6479

και η ικανότητα να παίζεις υπό πίεση. Το μοντέλο θα λαμβάνει επίσης υπόψη παράγοντες όπως η εμπειρία του παίκτη σε κορυφαία πρωταθλήματα, το ιστορικό τραυματισμών και η προσαρμοστικότητά του σε διαφορετικά στυλ παιχνιδιού και συστήματα τακτικής.

Ενσωματώνοντας αυτά τα διαφορετικά σημεία δεδομένων, το μοντέλο στοχεύει να παρέχει μια ολιστική αξιολόγηση της καταλληλότητας κάθε παίκτη για τη θέση του αριστερού κεντρικού αμυντικού. Ο απώτερος στόχος είναι να εφοδιαστούν οι ποδοσφαιρικές ομάδες με ένα ισχυρό εργαλείο που ενισχύει τη διαδικασία λήψης αποφάσεων, επιτρέποντάς τους να εντοπίσουν και να αποκτήσουν παίκτες που ταιριάζουν καλύτερα στις τακτικές τους ανάγκες και τους στρατηγικούς στόχους τους.

Συμπερασματικά, η παρούσα διατριβή προσπαθεί να γεφυρώσει το χάσμα μεταξύ της ανάλυσης δεδομένων και της πρακτικής εφαρμογής στο επαγγελματικό ποδόσφαιρο. Εστιάζοντας στον εξειδικευμένο αλλά κρίσιμο ρόλο των αριστερών κεντρικών αμυντικών, επιδιώκει να προσφέρει πολύτιμες γνώσεις και μεθοδολογίες που μπορούν να επεκταθούν σε άλλες θέσεις και πτυχές της αξιολόγησης των παικτών στο μέλλον. Αυτή η εργασία όχι μόνο θα συμβάλει στον ακαδημαϊκό τομέα της αθλητικής ανάλυσης, αλλά θα έχει επίσης απτό αντίκτυπο στις επιχειρησιακές στρατηγικές των επαγγελματικών ποδοσφαιρικών συλλόγων.

https://doi.org/10.26219/heal.aueb.6479

TABLE OF CONTENTS

https://doi.org/10.26219/heal.aueb.6479

LIST OF TABLES

https://doi.org/10.26219/heal.aueb.6479

https://doi.org/10.26219/heal.aueb.6479

LIST OF FIGURES

https://doi.org/10.26219/heal.aueb.6479

https://doi.org/10.26219/heal.aueb.6479

1. Introduction

The purpose of this thesis is to build a model that in the future can be used by professional soccer teams to be able to choose the best player for the position they would like to transfer. This model will be able to evaluate data that will come from factors, depending on the position each football player plays, always including the market value of the player, but also his age. For this paper, ten soccer players from different European leagues have been selected. The common characteristic of these ten players is the position in which they compete. This position is left central defender. This is a quite special position in football as there are not many left-footed footballers in particular playing in this position. The players selected for this analysis are as follows:

- Jorell Hato - Ajax FC
- Ricardo Calafiori – Bolognia
- Goncalo Inacio - Sporting CP
- Josko Gvardiol - Manchester City FC
- Castello Lukeba - RB Leipzig
- Willian Pacho - Eintracht Frankfurt
- Mauro Perkovic - GNK Dinamo Zagreb
- Piero Hinpacie - Bayer 04 Leverkusen
- Konstantinos Koulierakis - FC PAOK Thessaloniki
- Bjorn Meijer - Club Brugge KV

The explanatory variables that will be used to evaluate the players are as follows:

- Season
- Minutes
- Goals
- Assists
- Clean Sheets
- Ground Duels

https://doi.org/10.26219/heal.aueb.6479

- Aerial Duels
- Tackling
- Fouls
- Price
- Ball Recoveries
- Yellow Cards
- Red Cards

Season: The variable season is a categorical variable that has two levels. Level 1 concerns the players' performances for the 2022/2023 season, while level 2 concerns the players' performances for the 2023/2024 season.

Minutes: The minutes variable is a quantitative variable, which includes the minutes played by each player for each of the above seasons.

Goals: In football (soccer), a goal is scored when the ball crosses the goal line between the goalposts and beneath the crossbar, provided that it was legally propelled by a player from one of the teams. Scoring a goal earns a team points, typically one point per goal, though some variations exist in different formats of the game. Goals are the primary means of determining the winner in a match, and the team with the most goals at the end of the game usually wins, unless the match ends in a draw or tie.

Assists: In football (soccer), an assist is a statistic that credits a player for setting up a goal-scoring opportunity for a teammate. The player who provides the final pass, cross, or through ball that directly leads to a goal is awarded the assist. Assists acknowledge the contribution of players who play a pivotal role in the build-up to a goal, even if they don't actually score themselves. Assists are an important aspect of evaluating a player's overall performance and impact on a game.

Clean Sheets: In football (soccer), a clean sheet refers to a match in which a team's goalkeeper successfully prevents the opposing team from scoring any goals. When a goalkeeper and their defense manage to keep the opposition from scoring throughout the entire match, it is referred to as a clean sheet for the goalkeeper and the team's defense. Clean sheets are often considered a significant achievement for goalkeepers and their defensive teammates, as they demonstrate strong defensive play and

https://doi.org/10.26219/heal.aueb.6479

goalkeeping skill. Clean sheets can be an important factor in determining a team's success in a match or over a season, as they indicate defensive solidity and the ability to shut down the opposing team's attacks

Ground Duels: In football, ground duels refer to physical contests or challenges that occur between players on the field, typically for control of the ball. These duels involve players competing for possession of the ball while it is on the ground, such as when it has been passed along the pitch or is being dribbled by a player. Ground duels can include actions like tackling, jostling for position, shielding the ball, and attempting to dispossess the opponent. Ground duels are a fundamental aspect of the game and occur frequently throughout a match as players vie for control and dominance. They test players' strength, agility, balance, and technical ability. Winning ground duels is crucial for gaining possession of the ball, maintaining pressure on the opponent, and creating scoring opportunities.

Aerial Duels: In football, aerial duels refer to physical contests or challenges that occur between players in the air, typically to win headers or aerial balls. These duels involve players competing for the ball while it is airborne, such as during goal kicks, long passes, crosses, or set-piece situations like corners and free kicks. During aerial duels, players often jump to reach the ball at its highest point, using their height, timing, jumping ability, and heading technique to gain an advantage over their opponent. Aerial duels are common in football, especially in matches where teams utilize long balls or crosses as part of their attacking strategy. Winning aerial duels is important for both defensive and offensive purposes. Defenders aim to clear the ball away from danger and prevent opposing attackers from winning headers in dangerous areas, while attackers seek to win headers to create scoring opportunities or maintain possession for their team.

Tackling: In football (soccer), tackling refers to the act of legally attempting to dispossess an opponent of the ball. It involves using physical contact or skillful interception to gain control of the ball from an opponent who is in possession of it. Tackling is a crucial defensive skill in football and is employed by players to halt the progress of an opposing player or regain possession for their team. There are various types of tackles in football, including slide tackles, standing tackles, block tackles, and shoulder challenges. Each type of tackle requires different techniques and timing, and

https://doi.org/10.26219/heal.aueb.6479

they are used depending on factors such as the position of the ball, the movement of the opponent, and the tactical situation of the game.

It's important to note that tackling must be executed cleanly and within the rules of the game to avoid fouls or disciplinary action from the referee. Illegal or reckless tackles, such as those that involve excessive force, endanger the safety of the opponent, or are executed from behind, can result in penalties, including free kicks, yellow cards, or red cards.

Fouls: In football (soccer), fouls occur when a player commits an infraction against an opponent or violates the rules of the game as set forth by the governing body, typically FIFA. Fouls can encompass a wide range of actions, including physical contact, obstruction, dangerous play, or unfair challenges.

Common fouls in football include:

1. **Tripping or tackling an opponent illegally:** This includes tripping, kicking, or attempting to tackle an opponent in a manner that is deemed illegal or reckless by the referee.
2. **Pushing, holding, or pulling:** Actions such as pushing, holding, or pulling an opponent without making a play for the ball are considered fouls.
3. **Handball:** Intentionally handling the ball with the hand or arm (excluding the goalkeeper within their own penalty area) is considered a foul.
4. **Obstruction:** Impeding an opponent's progress without playing the ball is considered obstruction and is penalized as a foul.
5. **Reckless or dangerous play:** Actions that endanger the safety of opponents, such as high kicks, studs-up challenges, or dangerous tackles, are considered fouls.
6. **Offensive fouls:** Actions such as charging into an opponent, using excessive force, or spitting at an opponent are considered offensive fouls and are penalized accordingly. When a foul occurs, the referee stops play and may award a free kick, a penalty kick (if the foul was committed inside the penalty area), or issue a yellow or red card to the offending player, depending on the severity of the foul and the context of the game.

https://doi.org/10.26219/heal.aueb.6479

Price: In football, the term "price" concerning a player generally refers to the monetary value associated with their transfer or acquisition by a club. This price represents the amount of money that one club pays to acquire the services of the player from another club, usually through a transfer fee.

The price of a player can vary significantly based on several factors:

1. **Transfer Market Dynamics:** The demand for a player, the level of competition for their signature among clubs, and the availability of alternatives can influence the price.
2. **Player Attributes:** The player's skill level, age, experience, potential, and performance record all play a role in determining their price. Young, talented players with high potential often command higher prices.
3. **Contract Situation:** The length of the player's contract with their current club, their desire to move, and any release clauses in their contract can affect their price.
4. **Market Inflation:** Market trends, economic factors, and the overall inflation in football transfer fees can impact the price of a player.
5. **Club Finances:** The financial capabilities and ambitions of the buying club also play a significant role. Wealthier clubs may be willing to pay higher prices for top talent to strengthen their squad and compete for trophies.
6. **Agent Involvement:** The involvement of player agents and their negotiations with clubs can influence the final price.

Ultimately, the price of a player reflects their perceived value in the transfer market and the willingness of clubs to invest in their services to strengthen their squad and achieve their goals.

Ball Recoveries: In football (soccer), ball recoveries refer to instances where a player successfully regains possession of the ball for their team after it has been lost to the opposition. Ball recoveries can occur through various means, including interceptions, tackles, challenges, or pressuring the opponent into making a mistake. Ball recoveries are an essential aspect of the defensive phase of the game and are crucial for

https://doi.org/10.26219/heal.aueb.6479

transitioning from defense to attack. When a team loses possession, regaining control of the ball quickly can disrupt the opponent's momentum, prevent counterattacks, and allow the team to maintain pressure or build their own attacking opportunities. Tracking ball recoveries is a key statistic used to evaluate a player's defensive contribution and work rate. Players who excel at ball recoveries often possess attributes such as anticipation, positioning, timing, and tackling ability. Analyzing ball recoveries helps coaches and analysts assess a team's defensive effectiveness and individual player performance in regaining possession.

Yellow Cards: In football (soccer), a yellow card is a disciplinary sanction issued by the referee to a player as a warning for committing a foul or misconduct. When a player receives a yellow card, it serves as an official caution, indicating that the player has committed an offense that breaches the rules of the game but does not warrant an immediate expulsion from the match.

Common reasons for receiving a yellow card include:

1. Reckless or dangerous tackles
2. Deliberate handball to prevent an opponent's attack
3. Dissent or arguing with the referee's decision
4. Time-wasting tactics, such as delaying a restart
5. Persistent fouling or other repeated infractions
6. Simulation or diving to deceive the referee

A player who receives a yellow card is not immediately removed from the match but is required to be more cautious for the remainder of the game. However, if the same player receives a second yellow card in the same match, they are subsequently shown a red card and are sent off, resulting in their expulsion from the game. Additionally, accumulating a certain number of yellow cards over multiple matches can lead to suspensions, depending on the rules of the competition. Yellow cards are an essential tool for maintaining discipline on the field and ensuring fair play, as they deter players from engaging in reckless or unsporting behavior.

Red Cards: In football (soccer), a red card is a disciplinary sanction issued by the referee to a player for a serious offense or a severe breach of the rules of the game.

https://doi.org/10.26219/heal.aueb.6479

When a player receives a red card, they are immediately expelled from the match and cannot be replaced by a substitute. The team must then play with one player fewer for the remainder of the game.

Common reasons for receiving a red card include:

1. Serious foul play, such as dangerous tackles that endanger the safety of an opponent
2. Violent conduct, including punching, kicking, or striking an opponent or any other person
3. Denying an obvious goal-scoring opportunity to an opponent by committing a foul (other than a genuine attempt to play the ball) inside their team's penalty area
4. Using offensive, insulting, or abusive language or gestures towards referees, opponents, or teammates
5. Receiving a second yellow card in the same match, resulting in an automatic dismissal

A red card is a severe penalty that significantly impacts the team's chances of success in the match, as they are forced to play with fewer players for the remainder of the game. Additionally, the player who receives a red card may face further disciplinary action, including fines, suspensions, or bans, depending on the nature of the offense and the rules of the competition. Red cards are an important tool for maintaining discipline, ensuring player safety, and upholding the integrity of the game by punishing severe misconduct or unsporting behavior.

https://doi.org/10.26219/heal.aueb.6479

2. Literature Review

2.1 Logistic Regression

Logistic regression is essentially a model for classifying the values of one response variable Y based on probability theory. In this model, the variable Y usually has a binary character (takes two values) and its aim is to predict this outcome from a number of predictor variables that can be nominal, ordinal or quantitative.

The major difference between logistic and linear regression is based on nature of the selected response variable, which can be categorical. In the classical linear regression, the estimation of the parameters is done by the method of least squares, but the parameters of the logistic regression are estimated using the likelihood ratio method (commonly applied method in the generalized linear models).

Three types of logistic regression are distinguished depending on the nature of the dependent categorical variable which can be:

1. Binary dependent variable. It consists of two categories, such as e.g. are the outcomes pass/fail, YES/NO, event absent/present.

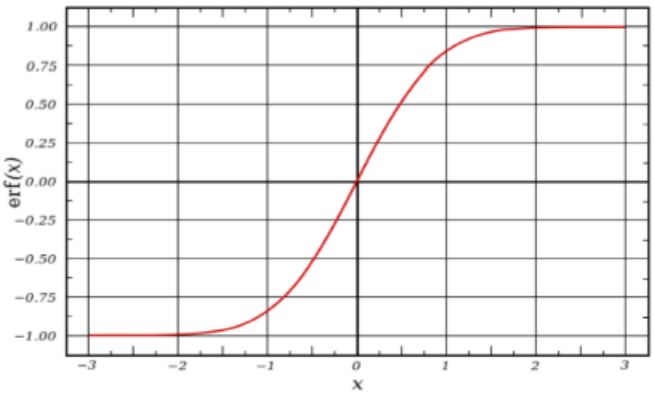
2. Ordinal variable. The dependent variable consists of three or more categories, such as on a question of scale I disagree at all, little, moderate, quite, much.

3. Nominal (Nominal) or polynomial (polynomial) or polychotomus (non-ordered categorical) or multiple response variable. It contains three or more categories without any physical gradation, such as e.g. the characterization of a food as crunchy, soft, friable or the color of objects as red, green, yellow etc.

Development of the model

In the language of statistics, logistic regression is used to predict the probability of an event occurring by fitting the study data to the logistic curve equation as this is shown in figure 1.

https://doi.org/10.26219/heal.aueb.6479



This curve has a sigmoid shape and it is characterized by an exponential growth stage in which, the growth rate gradually slows down and terminates in the asymptotic saturation stage of growth (the straight line eventually goes parallel to the X axis). Binary logistic regression is a binomial equation in which the response variable Y is the random result of occurrence of one of two potential outcomes of the type success or failure such as e.g. is the result of tossing a coin with two different sides (crown-letters), tossing a of dice where the result of appearing the number 6 is considered a success and the other numbers a failure, positive vote for the election of a political representative, etc.

$$f(z)=\frac{e^{z}}{1+e^{z}}=\frac{1}{1+e^{-z}}$$

where z is the independent variable and f(z) is the dependent variable. The advantages of the equation include the fact that the input variable takes positive and negative values while the result f(z) is limited to a range of values between 0 and 1. More specifically, the variable z represents the action of a group of independent variables, while f(z) specifies the probability of a particular outcome due to the action of this group. The variable z also expresses the measure of the total contribution of all independent variables in the model and is defined as

$$z=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k$$

where b0 is the value of the slope of the regression line and is equal to the z value when the values of all independent variables are equal to 0, while βi are the regression coefficients each of which expresses the magnitude of the contribution of the

https://doi.org/10.26219/heal.aueb.6479

corresponding variable. A positive value of the coefficient indicates that the explanatory variable increases the probability of the successful outcome (i.e. that the event will occur), a negative value means that the variable reduces the probability of that outcome. A high value of the coefficient means that the independent variable strongly affects the probability of the event occurring or not, while a low value indicates a small effect of the independent variable on the probability of the corresponding outcome.

In summary, logistic regression serves to describe the relationship that develops between one or more independent variables (e.g., age, sex, toxic substance concentration) and of a binary response variable expressed as a probability capable of taking one of two values, as e.g. positive (1) negative (0), present contingency (1) absent contingency (0), survivors (1) deaths (0), liked (1) unpleasant (0).

The above relationship (logit) can be incorporated into the regression model in logarithmic form as follows,

$$logit(p) = \log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

The regression coefficients are calculated using the Maximum Likelihood Estimate (MLE), as

$$L = \prod_{i=1}^{n} f(x_i \theta)$$

The probability function of the outcome of an event (likelihood) shows how well an observed sample is described by some parameter values, e.g. mean, standard deviation. The reliability of logistic regression results is greatly affected by the sample size of the survey.

https://doi.org/10.26219/heal.aueb.6479

2.2 Multiple Binomial Regression

It is a part of categorical statistical models known as Generalized Linear Models, which include the familiar classical regression, analysis of variance and covariance, and logarithmic regression. This method allows predicting the values of a dependent binomial variable from a number of independent variables, which may be quantitative, dichotomous or multiple variables. The opposite of logistic regression is discriminant categorical analysis with the difference that the second involves only quantitative independent variables.

The response variable in its algebraic version takes the value 1 with probability of success p and the value 0 with probability of failure 1-p and is called Binary or Binomial or Bernoulli. Due to the nature of the variables involved, the conditions of normal distribution are absent of the values and the homogeneity of their fluctuations, and the lack of linearity between Y and of independent variables is improved using the logarithmic equation as follows,

$$p = \frac{e^z}{1+e^z}$$

Where,

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Then the relationship is used

$$\log_e\left(\frac{p}{1-p}\right) = z$$

Characteristics of the binomial regression equation

Logarithmic (complementary) likelihood models or logit models are used when the dependent variable is binary. Logistic regression is non-linear in form because it forces the predicted values to range between 0 and 1. Logistic models estimate the probability

https://doi.org/10.26219/heal.aueb.6479

of the dependent variable taking the value 1 (Y=1), i.e. the probability that some event occurs. These models obey the condition probability of occurrence Pr,

$$Pr\ (Y=1\,|\,X_{1,}X_2,\ldots X_k)=\frac{1}{1+e^{-z}}$$

2.3 Multiple Ordinal Regression

Ordinal regression is chosen in cases where the dependent variable is divided into more than two categories that can be classified with some logic meaning, such as the preference of a product with a rating of the form I accept not at all, a little, quite, a lot, corresponding to integers with an increasing scale (1,2,3,4,5).

The use of ordinal regression in place of classical linear regression should be avoided, for the reason that ordinal values, in the continuous place of the dependent variable, violates the assumptions of least squares regression.

Characteristics of ordinal regression

Probability of outcome

This is expressed as p_k where k=1,2...K levels and holds,

$$P(y\leq k)=p_1+p_2+\ldots+p_k=\frac{e^z}{1+e^z}$$

Where,

$$z=\beta_0+\beta_1X_{j1}+\beta_2X_{j2}+\ldots\beta_kX_{jk}$$

Cumulative probability of outcome

Expresses the probability that a response falls into a level k,

$$P(y\leq k)=p_1+p_2+\ldots+p_k,\qquad k=1,2..K$$

https://doi.org/10.26219/heal.aueb.6479

The cumulative outcome probabilities represent the ranking of the response. For a model with k =1,2...K,

$$P(y \leq 1)+P(y \leq 2)+...+P(y \leq K)=1$$

Since the sum of the probabilities is equal to 1, the probability of the last category cannot actually be calculated. Thus, the logarithms of the odds of a successful outcome for k-1 add up probabilities, are given as

$$logit[P(y \leq k)] = \log_e \frac{P(y \leq k)}{1-P(y \leq k)}$$

By this term is meant the non-linear transformation of the predicted values in such a way that the distribution of the new values is included in one of the known family distributions, thus enabling the link function to predict the response of a dependent variable as no linear effect of a multitude of independents. The functions used specifically regular regressions refer to:

a) The logarithmic ratio of the probabilities (logit): f(z)= log[z/(1-z)], which is indicated when the observed frequencies of the gradients are uniformly distributed among themselves and is the most frequent choice.

b) The complementary log-log: f(z)= log[-log(1-z)], is chosen when the frequencies in the higher gradations are more numerous than in the lower ones.

c) The negative bilogarithmic (negative log-log): f(z)= -log[-log(z)], useful when the frequencies of the lower gradients are comparatively higher than the higher ones.

d) Cumulative probit: f(z)= Φ-1(z), which is the inverse of cumulative normal distribution and is chosen when the gradients of the dependent ordinal variable follow the normal distribution.

e) The inverse form of Couchy (Cauchit): f(z)= tan [p(z-0.5)], which is chosen when detected many extrinsic values.

The first three link functions are chosen more often than the rest, they are usually similar behavior in the transformed elements, however, variations are sometimes observed as their effectiveness.

https://doi.org/10.26219/heal.aueb.6479

Regression coefficients

The technique embraces models with proportional outcome probabilities, which means that the independent variables exert an equivalent effect on all k-1 gradients of the dependent variable. The coefficient of each independent variable and for a specific category k of the dependent one, expresses the change in the logarithm of the response in this category compared to the reference category K. The calculations of the coefficients are performed using the maximum likelihood estimation (MLE) method.

The standard error of the coefficients follows the rules of binomial regression. The importance of each coefficient β and therefore of the corresponding predictor independent variable is controlled by the relationship,

$$z = \frac{\beta_i}{SE}$$

and the respective fixed parameter

$$z = \frac{\beta_0}{SE}$$

Values of $z \geq 1.96$ indicate statistical significance at the 0.05 error level ($\beta \neq 0$). The 95% confidence limits of the coefficients are calculated from the relation $\beta i \pm z_{0.05/2} * SE$.

The respective confidence limits of the ratio of the outcome probabilities are derived as the antilogarithms of the lower and upper bounds of the prior relationship. The two limits contain a range within which the odds ratio result is expressed with any value.

Ratio of the odds of a successful outcome

For each independent variable, a single parameter and an odds ratio are estimated for the calculation whose cumulative probabilities are used. Thus, for an independent variable X with X1 and X2 categories the cumulative ratio of the odds will equal

$$\theta = \frac{P(y \leq k \mid X = X_2) / P(y > k \mid X = X_2)}{P(y \leq k \mid X = X_1) / P(y > k \mid X = X_1)}$$

https://doi.org/10.26219/heal.aueb.6479

Log probability of outcome

It is used to compare two models that differ by an independent input variable, e.g. in one model 3 variables are entered and in the other 3+1. If we consider that in ordinal regression there are n independent categorical variables with k categories each, then the contribution of each observation yi at log probability of outcome, will be expressed by

$$L(p_i; y_i) = \sum_k y_{ik} \log_e p_{ik}$$

where $yi = (yi1, \ldots yik)$ while Σj holds $yij = mj$ for each observation i, pik the probability of observation i for class k.

Test criterion for equality of slopes

The G-criterion (log-likelihood criterion) tests the statistical significance between a model with only the fixed parameters β_0 present and of the model with all coefficients of the independent variables and examines whether all coefficients are equal to 0. The G-test should give statistically significant effect ($p<0.05$) so that it is true that at least one coefficient differs from 0 and therefore there is a possibility that logistic model is applied.

Model goodness-of-fit checks

1. The control X^2 Pearson's is based on the estimation of the residuals and follows the relation

$$\chi^2 = \frac{\sum_j r_j^2}{mp_j^{(0)}}$$

rj= the Pearson residual of category j

m= o number of observed trials (attempts) in category j

pj(0)= probability corresponding to the null hypothesis (no difference between observed and expected values)

2. The deviation control D,

https://doi.org/10.26219/heal.aueb.6479

$$D = 2\sum y_{ik} \log_e p_{ik} - 2\sum y_{ik} \log_e p_{ik}$$

which has j-(p+1) degrees of freedom, where j is the number of distinct classes of the independent variables and p is the number of independent variables and pik is the probability of the i observation for category k.

The criteria X^2 of Pearson and deviance D show how well the chosen model fits the elements of the study. High criteria values when they correspond to an exact error probability of $p<0.05$ indicate that the model does not adequately describe the data.

The above two criteria are disadvantaged when the number of categorical values (nominal or graded) of the independent variable approaches the number of observations, but they excel when there are repeated observations for each category of the variable

2.4 Multiple Nominal Regression

It is chosen in the cases where the dependent variable is nominal and contains more than two ungraded categories (e.g. regions of a territory, cheeses of various origins, plant species of a wetland, sediment substrates, etc.). The technique, also known as multiple or polynomial regression, admits that each independent variable has a single value for each observation and that the dependent variable cannot be perfectly predicted from just one independent variable for each observation. The independent variables should be strongly uncorrelated between them, i.e. there should be a lack of multicollinearity.

In models where the dependent variable has the form of multiple-choice items, with the possibility of choosing more than one category at a time, the technique assumes that the other unrelated alternatives are independent. Nominal logistic regression models are represented by one dependent variable with more than two categories, i.e. they are an extension of dichotomous regression and are described by the equation,

$$\log_e\left(\frac{p_{ij}}{p_{il}}\right) = x_i \beta_j$$

where $j = 2,3\ldots J$ and $i = 1,2\ldots N$ and the probability $pij = P(Y = j|x)$ is obtained as

https://doi.org/10.26219/heal.aueb.6479

$$p_{ij}=\frac{e^{x_i\beta_j}}{\sum_{j=1}^{J} e^{x_i\beta_j}}$$

where xi is the vector of independent variables. The unknown parameters βj are calculated with maximum likelihood estimation (MLE).

In multiple logistic regression, one of the categories of the dependent variable is selected as a basic or comparison or otherwise reference category (baseline category). Separate odds ratios are estimated for all independent variables for each category of the dependent, except for the basic, which is excluded from further analysis. The exponential coefficient β expresses the change of the successful probability of each category compared to the baseline when each independent variable is varied by one unit.

Nominal accounting models consider the multilateral metrics in each different combination of the independent variables are independent. The maximum number of independent variables should be significantly less at polynomial regression compared to that in linear, since the categorization of the dependent variable in polynomial regression means limitation in providing information.

To apply polynomial logistic regression, the need to have a large sample size of observations (at least 50 observations per independent variable) is supported for the accuracy of hypothesis testing, particularly when the dependent variable has multiple groups.

Characteristics of the method

Probability of a successful outcome

If we consider a deviation variable with three categories 1, 2 and 3 and reference category 3 then the bound probabilities arise as

$$P(y=1|X)=\frac{e^{x'\beta_1}}{1+e^{x'\beta_1}+e^{x'\beta_2}}$$

https://doi.org/10.26219/heal.aueb.6479

$$P(y=2|X)=\frac{e^{x'\beta_2}}{1+e^{x'\beta_1}+e^{x'\beta_2}}$$

$$P(y=3|X)=\frac{1}{1+e^{x'\beta_1}+e^{x'\beta_2}}$$

The model with the above three categories of response (reference category n with number 3) is described by the equation of the logarithmic probability of outcome:

$$L(\beta)=\sum_{i}^{n} y_{1i}g_1(X_i)-\log_e\left(1+e^{g_1(X_i)}+e^{g_2(X_i)}\right)$$

Test criterion for equality of slopes

Criterion G examines whether all the coefficients of the independent variables are equal to 0. When Criterion G is statistically significant ($p<0.05$) then it indicates that at least one coefficient differs from 0 and therefore a logistic model can be adapted.

Regression coefficients

With k categories of the dependent variable present k-1 parameters are estimated for each independent variable. Actions differ depending on the response category compared to the reference category.

Each log-odds ratio provides the estimated (adjusted) one-class differences response to the reference class. The parameters (coefficients) of the k-1 equations are log odds ratios using all other pairs of response categories by the method of maximum likelihood estimation.

The significance of the coefficients is calculated based on the Wald test as,

$$z=\frac{\beta_{k_i}}{SE}$$

$$Z=\frac{\beta_0}{SE}$$

https://doi.org/10.26219/heal.aueb.6479

Alternatively, with more reliability, the logarithm of the likelihood ratio is used -2logL which checks with the help of the X^2 criterion for the effect of each independent variable on the final model.

Ratio of the odds of the outcome

It is used to estimate the relationship between dependent and independent variable and always produces a positive value. A ratio equal to 1 is taken as a reference ratio, i.e. the value $\theta=1$ expresses the absence of one relationship between the two variables X and Y. If $\theta>1$ then the chances of a successful outcome are higher to occur for that particular independent variable. For example, if a model includes k classes of the dependent variable and one independent variable, then the ratio of the outcome probabilities indicates the success probability of a class compared to the reference class.

The following equation describes the ratio of the odds of a dependent categorical variable with three categories and an independent category with two categories, a and b:

$$\theta = \frac{P(y=k\,|X=a)\,/\,P(y=3\,|X=a)}{P(y=k\,|X=\beta)\,/\,P(y=3\,|X=\beta)}$$

Model goodness-of-fit checks

In the nominal regression, as in the previous techniques, the X^2 criteria are calculated by Pearson and the deviation D of the observations (observed-estimated) and are considered reliable only when there are many observations for each combination of the independent variables. Higher X^2 values declare model not adequately fitted to the analysis data.

The coefficients of multiple determination type R^2 or otherwise R^2 pseudo-coefficients are different approach to estimate the goodness of fit of the model providing an interpretation similar of the linear regression's coefficients, that is, they express the percentage of variation explained by the independent variables. The coefficient R^2 of Aldrich and Nelson is used in bilateral and multilateral logistic regressions, however failing to approach the maximum value of R^2, which is the unit. Most popular is considered the coefficient R^2 by McFadden.

https://doi.org/10.26219/heal.aueb.6479

The Cox & Snell pseudo-coefficient:

$$R^2_{(C \wedge S)} = 1 - \left(\frac{L_0}{L_M} \right)^{2/n}$$

where n is the number of observations and the parameters LM and $L0$ concern the probabilities with and without the variables present respectively. This index also fails to approach unity and this problem is circumvented by using the modified Nagelkerke/Cragg & Uhle equation, which allows the index to vary in the range 0-1:

$$R^2 = \frac{1 - \left(\frac{L_0}{L_M} \right)^{2/n}}{1 - (L_M)^{2/n}}$$

It is pointed out that the pseudo coefficients of determination R^2 are mistakenly considered quality controls adjustment of the models since they really seek to estimate the intensity of the relationship that develops between the dependent and independent variables.

Diagnostic criteria of nominal regression

The goodness of fit of the model is also aided by examining the standardized rj residuals of the Pearson of the deviation D, but more commonly with the diagrammatic change of the ΔX^2 criteria and ΔD and the ordinal number of observations or estimated probability values, as described in binomial regression.

Classification of observations

Assume that a polynomial regression model with three categories present in the dependent variable because of the action of two independent variables was adequately evaluated. Then, it is possible, based on regression coefficients, the estimate of the probability that an individual (observation) belongs to one of the three categories of the dependent, applying the equation (antilogarithming the values j):

https://doi.org/10.26219/heal.aueb.6479

$$P_1(Y=1)=\frac{e^{g_1}}{\sum_{j=1}^{J} e^{g_j}}$$

$$P_2(Y=2)=\frac{e^{g_2}}{\sum_{j=1}^{J} e^{g_j}}$$

where P is the estimated probability for classes 1 and 2 (the third is set as the reference class).

The category showing the highest probability is considered to represent that individual. In this way the estimated probabilities for all individuals to belong to one of the three categories. However, the question arises as to how correctly this classification has been carried out if it is accepted, taking into account the degree of identification of the number of predicted observations with the empirical ones in the categories of the dependent variable. For this reason, a cross table of two directions is drawn up which contains in the cells the frequencies of correct matching per category of the predicted (columns) and of observed (series). The percentage of correctly predicted per category is also calculated either individually in each cell or in total as sums of rows and columns. The higher percentage matches the more the polynomial regression model is expected to be reliable. The classification of observations is a measure of the goodness of fit of the proposed model, but its reliability may be degraded in the presence of an unequal number of observations among the categories and therefore should be taken into account with the rest of the evaluation criteria of its adaptation model.

https://doi.org/10.26219/heal.aueb.6479

3. Results

3.1 Descriptives Statistics

The results of the descriptive measures showed that the quantitative variables are heterogeneous since the CV>30%. The distribution of the "Assists", the "Clean Sheets", the "Ground Duels", the "Aerial Duels", the "Tackling", the "Fouls", the "Price", the "Red Cards", the "Ball Recoveries", the "Yellow Cards" are positively distributed (Table 1, Figure 1)[1]. The variable "Minutes" are normally distributed.

The average of Minutes is 1950.65 (S.D.= 593.83) and the median of minutes is 1956, so the 50% players have upper than 1956 minutes. The mean of goals is 0.90 (S.D.= 0.85)and the mean of the assists is 0.95. The mean value of clean sheets is 6.5. The average of the ground deals is 12.50 (S.D.=8.57) and the average of the aerial duels is 25.15 (S.D.=19.13). The average value of tackling is 26.05 and maximum value reaches to 16.01. The average value of fouls is 13.80 (S.D.=9.74) and the average price is 25.54 (S.D.=9.38). The average value of red cards is 0.30 (S.D.=0.73). The average value of ball recoveries is 80.50 (S.D.=53.56) (Table 1).

[1] Appendix 1, Command 4-5.

https://doi.org/10.26219/heal.aueb.6479

var	*type*	*label*	*n*	*NA.prc*	*mean*	*sd*	*se*	*md*	*trimmed*	*range*	*iqr*	*skew*
Minutes	integer	Minutes	20	0	1950.65	593.83	132.78	1956.00	1939.44	2016 (970-2986)	893.50	0.14
Goals	integer	Goals	20	0	0.90	0.85	0.19	1.00	0.81	3 (0-3)	1.00	0.77
Assists	integer	Assists	20	0	0.95	1.47	0.33	0.50	0.62	6 (0-6)	1.00	2.42
Clean.Sheets	integer	Clean.Sheets	20	0	6.50	3.46	0.77	6.00	6.06	16 (1-17)	2.50	1.57
Ground.Duels	integer	Ground.Duels	20	0	12.50	8.57	1.92	10.00	11.19	35 (2-37)	8.00	1.49
Aerial.Duels	integer	Aerial.Duels	20	0	25.15	19.13	4.28	18.50	22.38	62 (5-67)	28.50	1.06
Tackling	integer	Tackling	20	0	26.05	16.51	3.69	20.50	24.44	64 (3-67)	19.75	0.95
Fouls	integer	Fouls	20	0	13.80	9.74	2.18	10.00	13.00	29 (3-32)	15.75	0.66
Price	numeric	Price	20	0	25.54	21.38	4.78	24.00	22.22	77 (3-80)	28.25	1.14
Red.Cards	integer	Red.Cards	20	0	0.30	0.73	0.16	0.00	0.12	3 (0-3)	0.00	3.02
Yellow.Cards	integer	Yellow.Cards	20	0	3.35	2.21	0.49	3.50	3.19	9 (0-9)	2.00	0.71
Ball.Recoveries	integer	Ball.Recoveries	20	0	80.50	53.56	11.98	63.00	72.12	216 (23-239)	68.25	1.54

Table 1: Descriptive Statistics

https://doi.org/10.26219/heal.aueb.6479

\

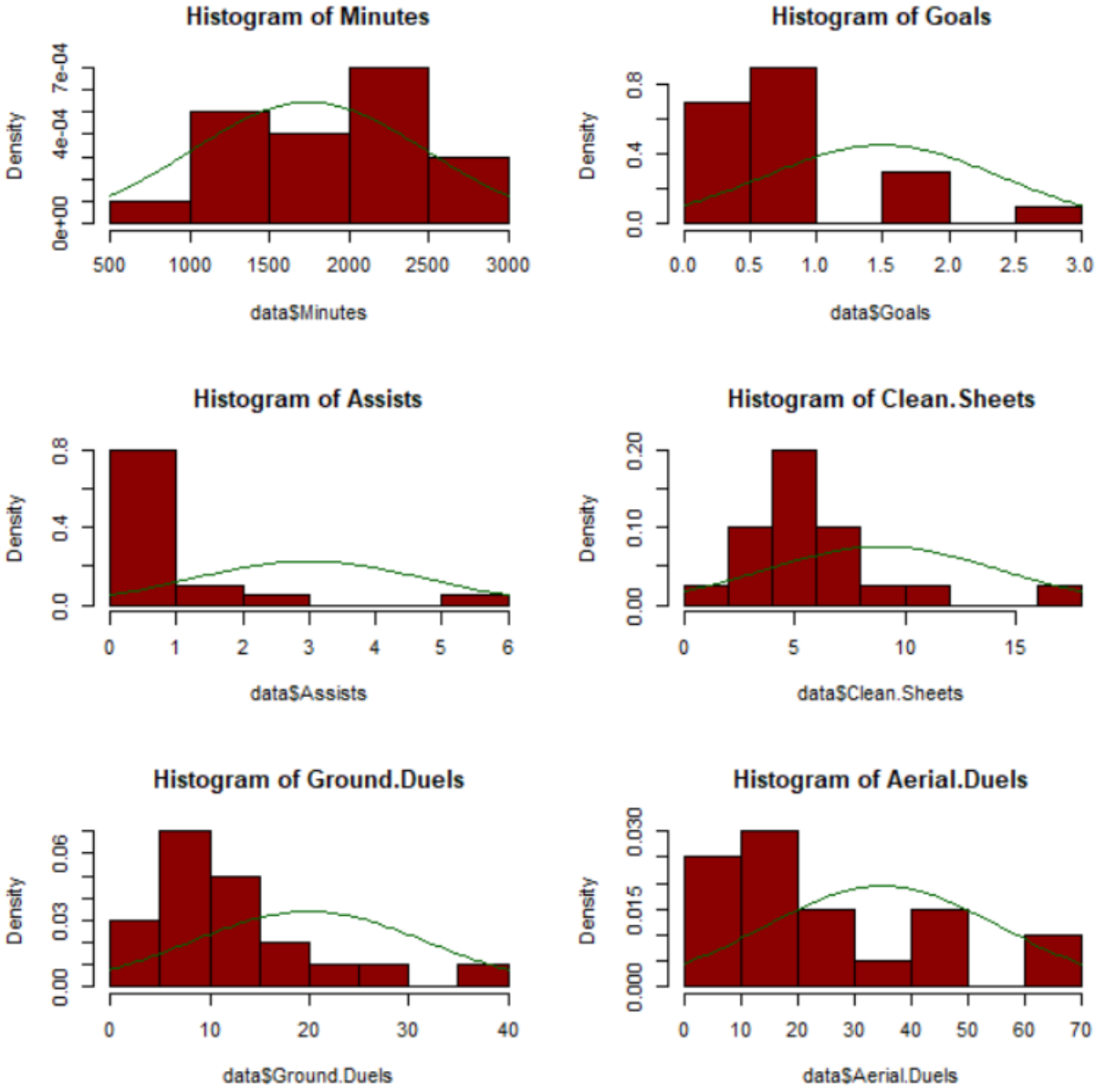


Figure 1a: Histogram

The diagrams are histograms and we can observe the distribution of the quantitative variables.

https://doi.org/10.26219/heal.aueb.6479

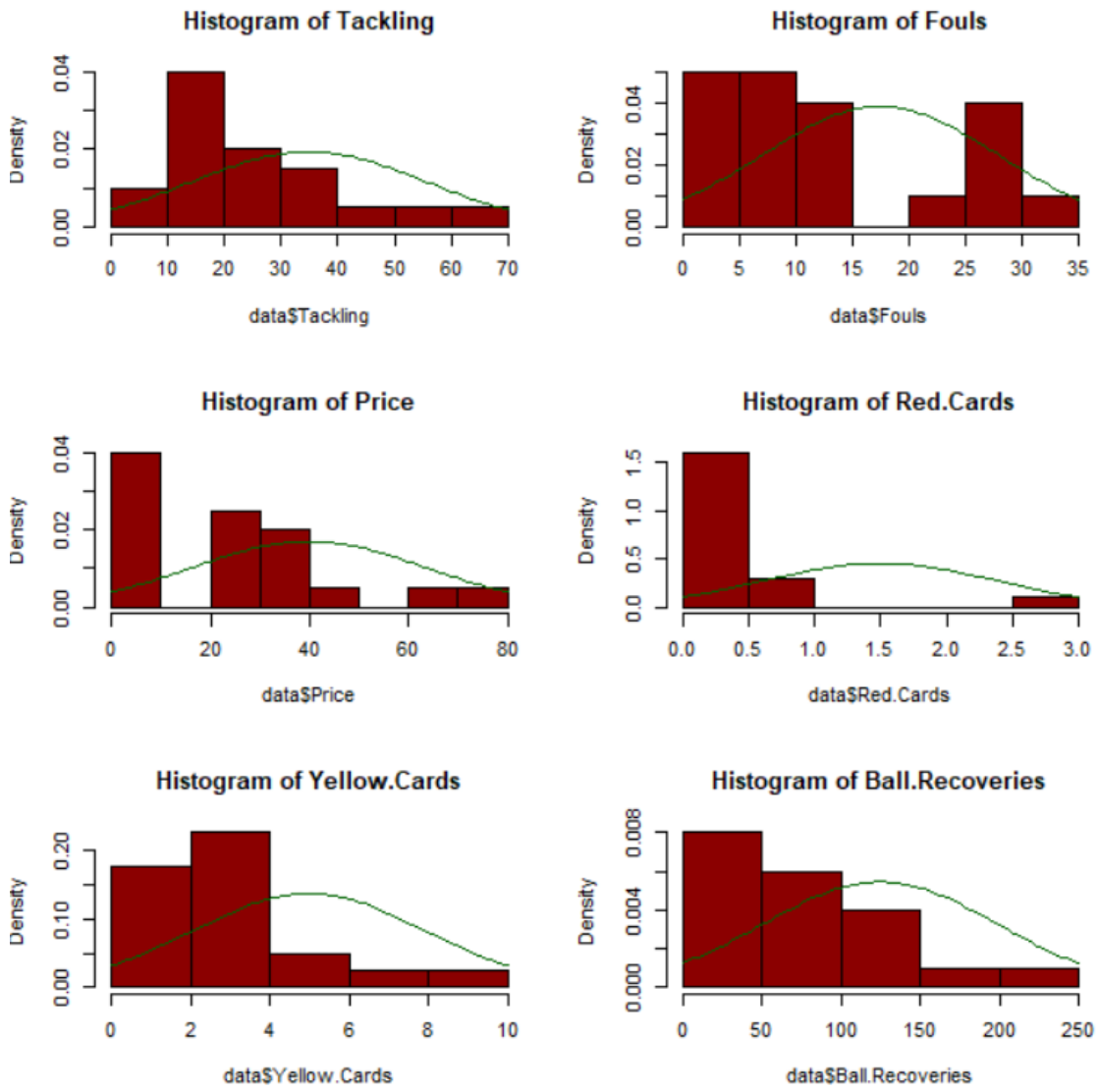


Figure 1b: Histogram

The table below shows the average values of the quantitative variables for each player.

https://doi.org/10.26219/heal.aueb.6479

Player	Minutes	Goals	Assists	Clean Sheets	Aerial Duels	Tackling	Fouls	Price	Red Cards	Yellow Cards	Ball Recoveries
Calafiori	1757	0.0	1.5	5.5	21.5	39.5	20.5	14.50	0.5	4.5	81.5
Lukeba	2270	1.5	0.0	6.0	10.0	10.5	4.5	37.00	0.5	5.0	39.0
Gvardiol	1821	0.5	0.0	4.5	8.0	24.5	11.5	74.50	0.0	2.0	70.5
Meijer	2060.5	2.0	3.0	5.5	40.0	30.5	18.5	8.00	0.0	2.0	110.5
Hato	1655	0.5	1.0	2.5	37.5	34.0	12.0	16.50	0.0	2.0	98.0
Pacho	2614.5	0.0	0.5	12.0	44.5	47.0	18.5	27.00	0.0	2.5	157.5
Hinpacie	1801.5	1.0	1.0	5.5	21.0	29.5	14.5	33.50	1.5	4.5	58.5
Perkovic	1738.5	1.0	0.5	9.0	8.5	11.0	6.0	3.15	0.0	3.0	29.0
Inacio	2064.5	1.0	1.5	7.5	16.0	12.5	5.5	36.00	0.5	4.5	45.5
Koulierakis	1724	1.5	0.5	7.0	44.5	21.5	26.5	5.25	0.0	3.5	115.0

Table 2: Descriptives by Player

https://doi.org/10.26219/heal.aueb.6479

3.2 Pairwise comparison

The aim of this chapter is to examine the relationship between the season and the other variables. The mean of minutes differs among the seasons and the minutes are higher for the second season. The results show that there is no statistically significant difference of the clean sheets, ground duels, aerial duels between the seasons (p-value=>0.05) (Table 2). The value of yellow cards differs statistically significant, and the values are higher for the second season. Using nonparametric test Wilcoxon, we find that there is no statistically significant difference of the goals, assists, tackling, fouls, price, red cards, Ball Recoveries between the seasons (p-value=>0.05).

Differences by Season	Test	p-value
Minutes	T-test	0.017
Goals	Wilcoxon	0.839
Assists	Wilcoxon	0.151
Clean Sheets	T-test	0.170
Ground Duels	T-test	0.022
Aerial Duels	T-test	0.901
Tackling	Wilcoxon	0.791
Fouls	Wilcoxon	0.790
Price	Wilcoxon	0.495
Red Cards	Wilcoxon	0.278
Yellow Cards	T-test	**0.030**
Ball Recoveries	Wilcoxon	**0.969**

Table 3: Pairwise Comparisons

https://doi.org/10.26219/heal.aueb.6479

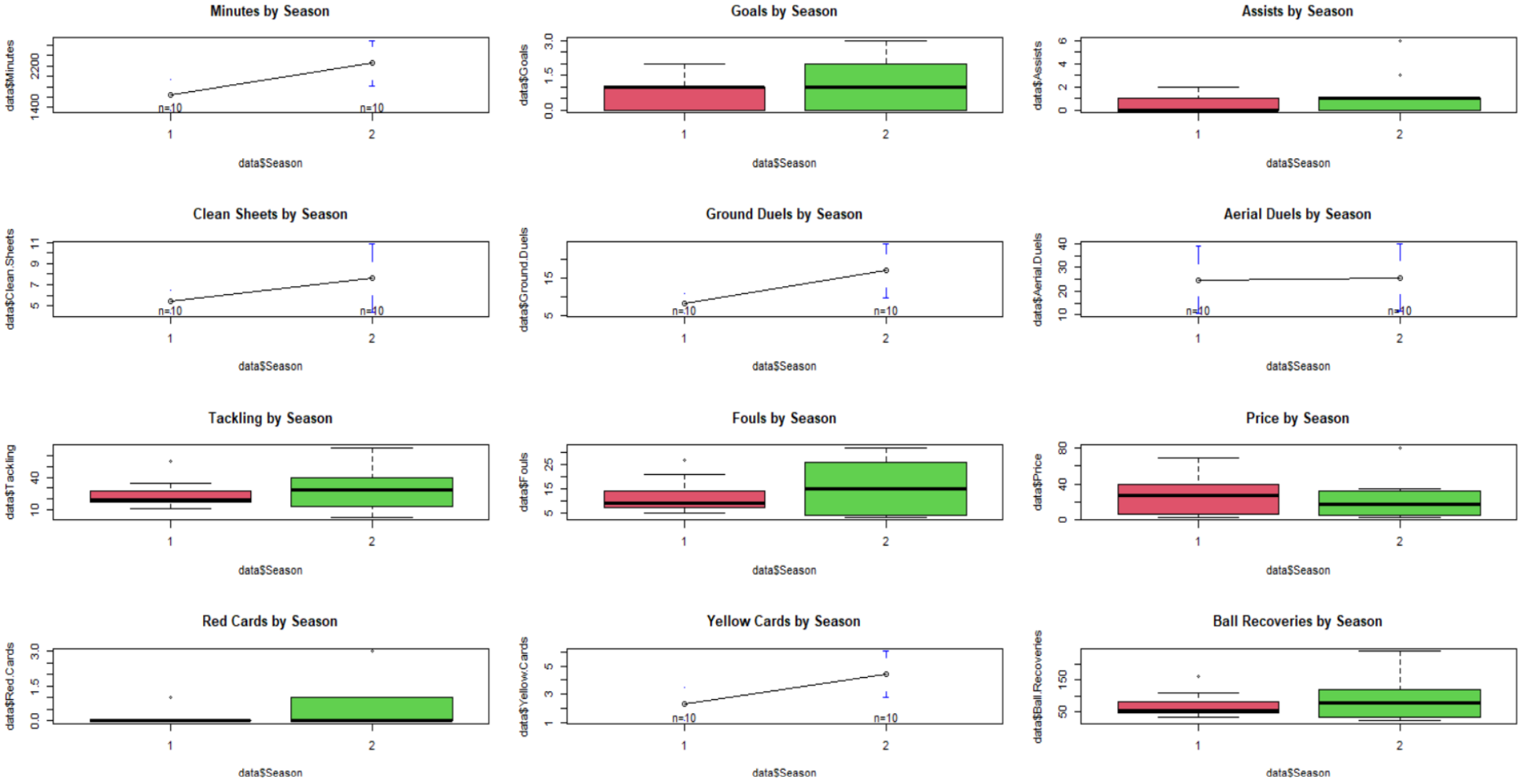


Figure 2

https://doi.org/10.26219/heal.aueb.6479

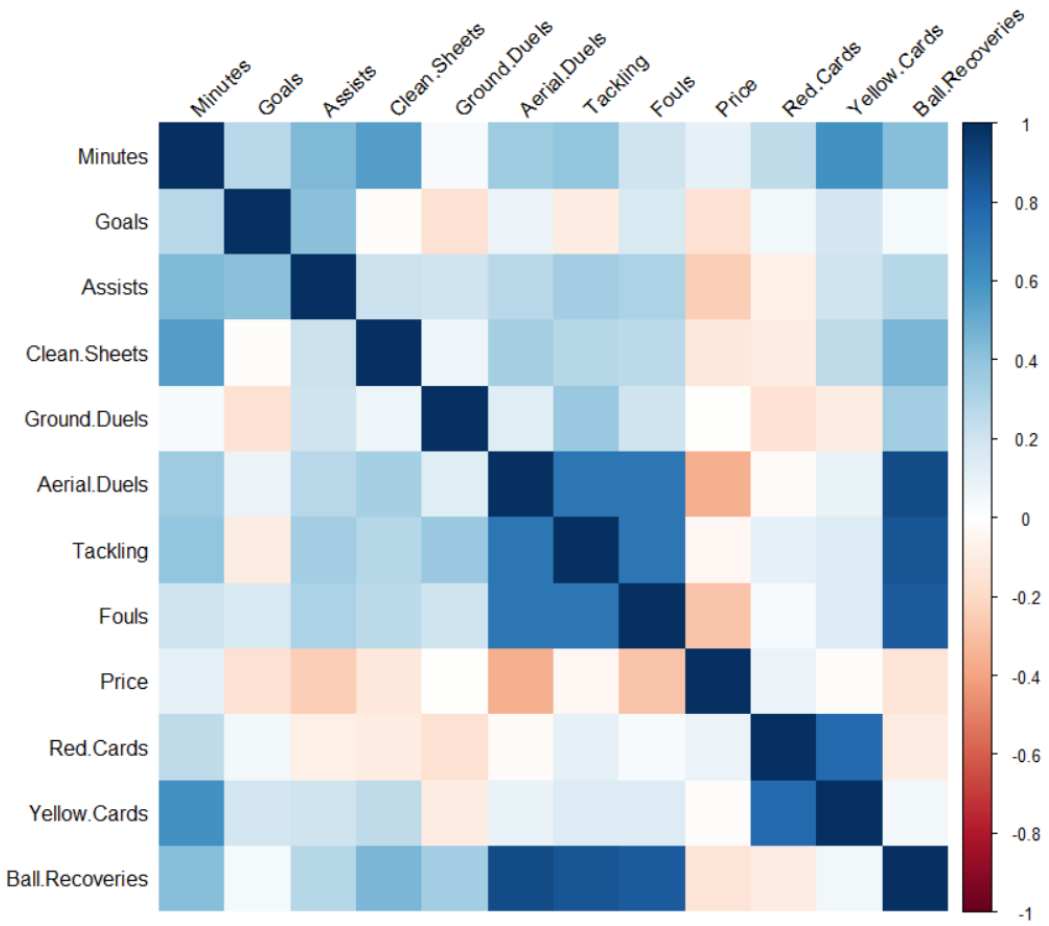


Figure 3: Correlation Matrix

The matrix shows the correlation among the quantitative variables.

https://doi.org/10.26219/heal.aueb.6479

3.3 Predictive or Descriptive models

3.3.1 Linear Model

Statistical relationships between variables are determined in the previous chapter. We will perform a model for the prediction of the price. Firstly, we check the linearity of the dependent variable with the independent variables. We were used as dependent variable the price. Firstly, the cumulative model is performed and some variables are removed, as there is the problem of the multicollinearity (VIF of these variables is higher than 10). In the first model, were used as dependent variable the Price. Firstly, the cumulative model is performed and some variables are removed, as there is the problem of the multicollinearity (VIF of these variables is higher than 10). Finally, the independent variables are the Season2, the Minutes, the Goals, Assists, Clean Sheets, Ground Duels, Aerial Duels, Tackling, Fouls, Red Cards and the Yellow Cards. By applying the regression analysis, we firstly check the normality of the residuals, this assumption is not biased.

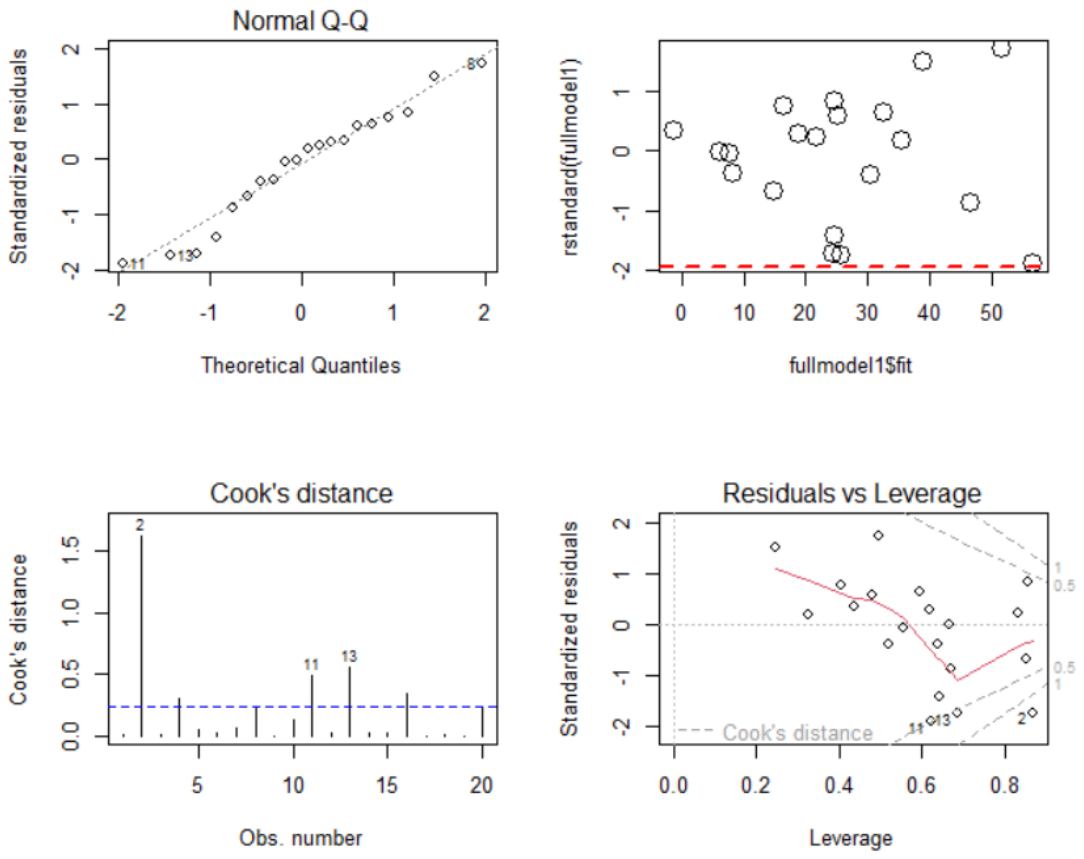
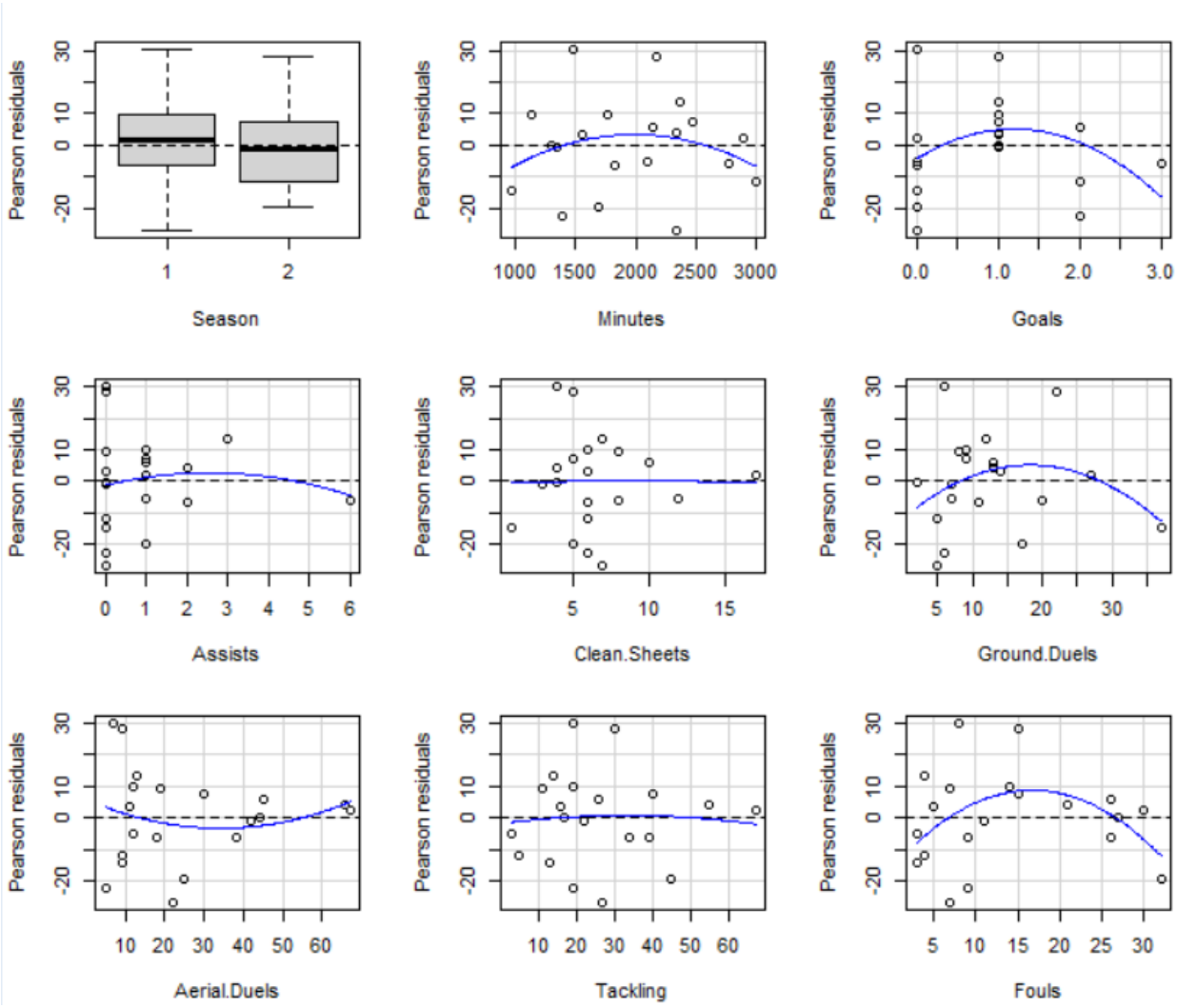


Figure 4

https://doi.org/10.26219/heal.aueb.6479

The linearity is satisfied, as well as the variance is constant (homoscedasticity). The variables are not statistically significant, the model is not statistically significant.



https://doi.org/10.26219/heal.aueb.6479

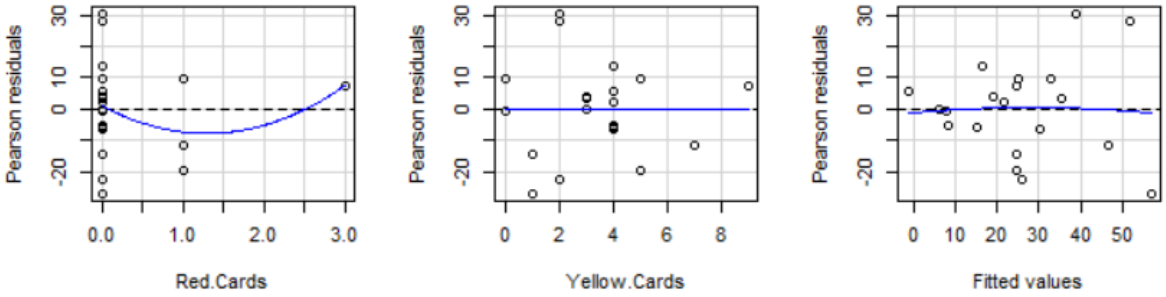


Figure 5

	Dependent variable		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-16.48	-89.31 – 56.35	0.616
Season [2]	-32.54	-93.28 – 28.21	0.252
Minutes	0.04	-0.01 – 0.10	0.115
Goals	-4.96	-26.22 – 16.30	0.605
Assists	-3.85	-15.46 – 7.75	0.466
Clean Sheets	-1.22	-6.18 – 3.73	0.584
Ground Duels	1.23	-1.82 – 4.28	0.381
Aerial Duels	-0.91	-2.04 – 0.22	0.101
Tackling	-0.27	-2.65 – 2.10	0.797
Fouls	1.07	-2.17 – 4.31	0.470
Red Cards	7.84	-30.87 – 46.56	0.653
Yellow Cards	-3.09	-16.63 – 10.45	0.613
Observations	20		
R^2 / R^2 adjusted	0.510 / -0.163		

The second model predict the number of goals. Firstly, we check the linearity of the dependent variable with the independent variables. We were used as dependent variable the Goals. Firstly, the cumulative model is performed and some variables are removed, as there is the problem of the multicollinearity (VIF of these variables is

https://doi.org/10.26219/heal.aueb.6479

higher than 10). Firstly, the cumulative model is performed and some variables are removed, as there is the problem of the multicollinearity (VIF of these variables is higher than 10). Finally, the independent variables are the Season2, the Minutes, the Price, Assists, Clean Sheets, Ground Duels, Aerial Duels, Tackling, Fouls, Red Cards and the Yellow Cards. By applying the regression analysis, we firstly check the normality of the residuals, this assumption is not biased.

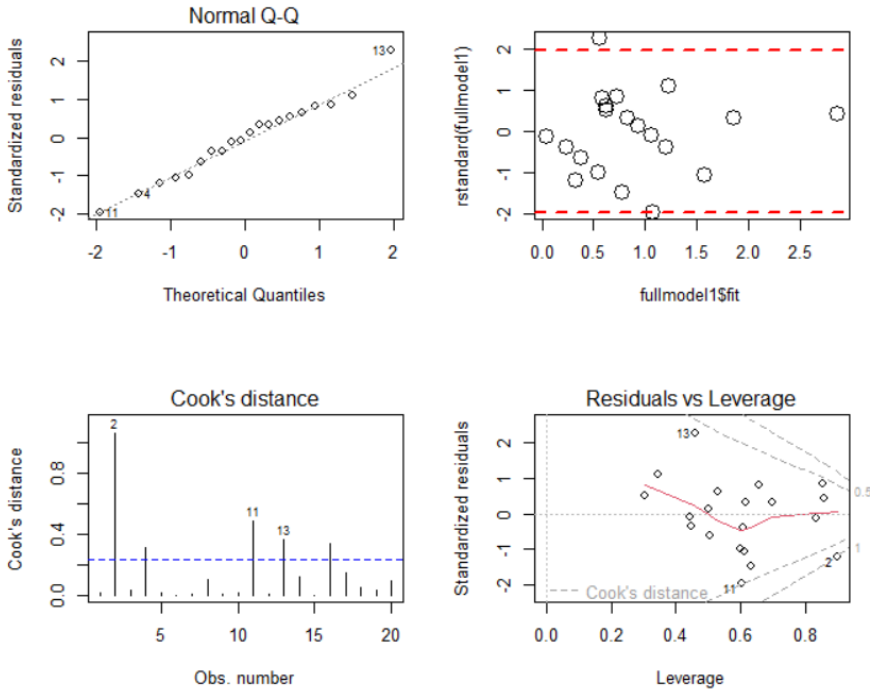


Figure 6

The linearity is satisfied, as well as the variance is constant (homoscedasticity). The variables are not statistically significant, the model is not statistically significant.

https://doi.org/10.26219/heal.aueb.6479

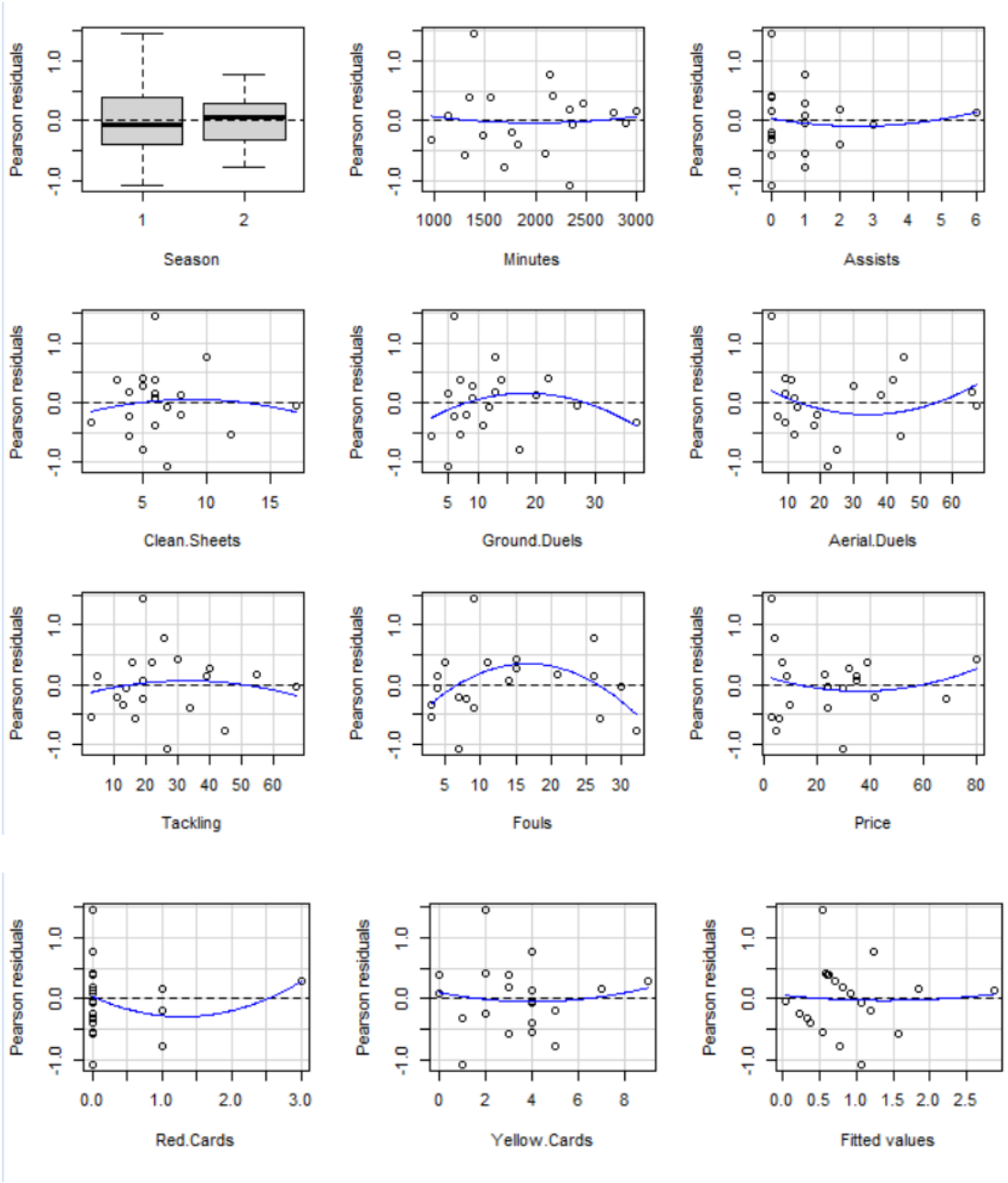


Figure 7

https://doi.org/10.26219/heal.aueb.6479

	Dependent variable		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.80	-3.51 – 1.91	0.516
Season [2]	-1.25	-3.53 – 1.03	0.243
Minutes	0.00	-0.00 – 0.00	0.117
Assists	0.20	-0.22 – 0.62	0.302
Clean Sheets	-0.05	-0.24 – 0.13	0.512
Ground Duels	0.05	-0.07 – 0.16	0.364
Aerial Duels	-0.01	-0.06 – 0.04	0.723
Tackling	-0.07	-0.14 – 0.01	0.069
Fouls	0.09	-0.02 – 0.19	0.094
Price	-0.01	-0.04 – 0.02	0.605
Red Cards	0.50	-0.92 – 1.92	0.437
Yellow Cards	-0.13	-0.64 – 0.37	0.559
Observations	20		
R^2 / R^2 adjusted	0.563 / -0.039		

By the above model we assume that only the variables tackling, and fouls are statistically significant. The variable “tackling” has negative effect on the number of goals, on the other hand the variable “fouls” has positive effect on the number of goals.

3.3.2 Multinomial Models

Multinomial Logistic Regression (MLR) is an extension of logistic regression that is used when the dependent variable is categorical and has more than two categories. It models the relationship between a set of independent variables and a dependent variable with multiple classes (nominal), allowing us to predict the probabilities of different outcomes. For a categorical dependent variable Y with K possible outcomes (i.e., classes $1,2,\dots,K$), multinomial logistic regression estimates he probability of each outcome as a function of the independent variables $X1, X2, \dots, Xp$. The probability of each outcome k is given by:

https://doi.org/10.26219/heal.aueb.6479

$$P(Y = k|X_1, X_2, ..., X_p) = \frac{\exp(\beta_{k0} + \beta_{k1}X_1 + \beta_{k2}X_2 + \cdots + \beta_{kp}X_p)}{1 + \sum_{j=1}^{K-1} \exp(\beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p)}$$

Where,

- P(Y=k|X1, X2, …, Xp) is the probability that the dependent variable Y takes the class k
- Bkj are the coefficients to be estimated for each class k
- $X1, X2, \ldots, Xp$ are the independent variables

The coefficients in multinomial regression represent the log-odds of being in one category compared to the reference category. For each category, the coefficient estimates show how the log-odds of that outcome change with respect to one unit change in the predictor variable. A reference category is chosen (usually the most common category or one chosen for interpretive convenience), and all other categories are compared to this baseline. The model will produce estimates for each other category relative to this reference.

In our case, the response variable is defined by the name of each player. In the first multinomial model, the independent variable is 'Total Performance'. Total Performance is defined as the sum of all the variables with a coefficient of 1 and positive when the variable is a positive score for a defender and minus when it is a negative score for a defender. The coefficient of 1 is defined so that all we are not biased on certain variables. In future analysis, one could further experiment with the definition of 'Total Performance' so a better Performance Score is established for a better results.

$$\begin{aligned} TotalPerformance \;=\; & Goals * 1 \;+\; Assists * 1 \;+\; Clean\,Sheets * 1 \\ & + \; Ground\;Duels * 1 \;+\; Aerial\;Duels * 1 \;+\; Tackling * 1 \\ & - \; Fouls * 1 \;-\; Red\;Cards * 1 \;-\; Yellow\;Cards * 1 \\ & + \; Ball\;Recoveries * 1 - \; Price * 1. \end{aligned}$$

https://doi.org/10.26219/heal.aueb.6479

In the second model in order to cross check the results and get a further idea on the best player, we have defined a multinomial model that the selected players are the response variable and each of the rest variables in the dataset are the independent ones. Following you can take a closer look in the results.

The results for the first model are:

```
> summary(model)
Call:
multinom(formula = Player ~ TotalPerformance, data = df)

Coefficients:
             (Intercept) TotalPerformance
Gvardiol    -101.7468615    -2.735489e-06
Hato          -0.7138368    -4.313906e-08
Hinpacie     -42.4853572    -1.595108e-06
Inacio       -47.8731771    -1.751037e-06
Koulierakis    7.8858311     1.171424e-06
Lukeba       -49.8234840    -1.804164e-06
Meijer         1.8122010     1.784595e-07
Pacho        -20.2005471    -8.621690e-07
Perkovic      37.7189366     8.797736e-06
```

```
Std. Errors:
             (Intercept) TotalPerformance
Gvardiol    9.323880e-14     6.066705e-06
Hato        6.174408e-14     5.812190e-08
Hinpacie    1.474566e-14     7.473885e-08
Inacio      2.355028e-14     7.596019e-08
Koulierakis 1.393549e-13     1.997252e-07
Lukeba      6.835669e-14     7.643421e-08
Meijer      1.012072e-13     1.035246e-07
Pacho       5.490975e-14     6.155560e-08
Perkovic    4.739513e-13     1.855071e-06

Residual Deviance: 35.44257
AIC: 71.44257
```

As you can can our model choose Calafiori as the base outcome. Perkovic is the most preferable according to our model. In my opinion this is due to the low level of price value in our dataset and the good scores of the rest ones. For this reason we have decided to define the second multinomial model where we are going to see similar results.

https://doi.org/10.26219/heal.aueb.6479

For the second model, we have the following results, with Calafiori as base outcome again:

```
> summary(model2)
Call:
multinom(formula = Player ~ Season + Minutes + Goals + Assists +
    `Clean Sheets` + `Ground Duels` + `Aerial Duels` + Tackling +
    Fouls + Price + `Red Cards` + `Yellow Cards` + `Ball Recoveri
es` +
    Price_Category, data = df)

Coefficients:
            (Intercept)     Season   Minutes      Goals
Gvardiol    -0.16305039  0.2150460 0.4127723 -0.3575693
Hato        -0.89812085 -0.5152539 0.4656553  0.2851802
Hinpacie    -0.41630150  0.3599529 0.3723703  0.1727535
Inacio      -1.02972084  0.8650745 0.3617623 -2.3484605
Koulierakis -0.90088036 -0.7545816 0.3931656  1.8482648
Lukeba       1.25668668 -0.8872017 0.2404395  3.1290518
Meijer       1.81219644  1.4447872 0.2172767  2.5958864
Pacho       -0.05563016  0.2919929 0.4273615  0.2123322
Perkovic     1.62002708  0.9857226 0.4635668  4.9062020
```

```
               Assists `Clean Sheets` `Ground Duels`
Gvardiol     0.6227480   -0.728089608      -2.183567
Hato         1.4235917   -0.003770331      -3.604754
Hinpacie    -0.9915214   -1.280639344      -1.150761
Inacio       6.3296988   -1.456548624      -1.384587
Koulierakis -0.4836402    0.841976216      -1.754406
Lukeba      -7.7427284    2.164465763      -1.207682
Meijer       0.1111750   -1.656178276       3.197176
Pacho       -1.8940007    1.301398160      -3.206670
Perkovic     0.1689204    3.456076305      -3.918798
            `Aerial Duels`   Tackling      Fouls
Gvardiol       -1.04355675 -0.5355074  0.4820455
Hato            0.37667666 -0.4340641 -0.2111170
Hinpacie        1.87291938 -1.1231522  3.0860777
Inacio          2.18725731 -2.8408081 -0.4324491
Koulierakis     0.43511722 -1.8941685  0.9409623
Lukeba         -0.73429150  0.3748677 -1.3741998
Meijer          1.01466694 -1.5792711 -0.5668276
Pacho          -0.51816352 -0.0305163 -1.2681485
Perkovic       -0.04459041  0.0498223 -0.5020450
```

https://doi.org/10.26219/heal.aueb.6479

	Price	`Red Cards`	`Yellow Cards`
Gvardiol	1.124029e-06	0.11721614	-0.01503974
Hato	-1.931959e-07	-0.32894801	3.00204156
Hinpacie	9.889755e-07	2.36741993	0.92039520
Inacio	1.296640e-06	0.52785959	-0.84742135
Koulierakis	-1.007306e-07	-0.49786071	0.52172944
Lukeba	4.437872e-07	-0.12986094	3.53781637
Meijer	-3.575723e-07	1.19889920	-3.91435952
Pacho	-1.812263e-07	1.29037383	1.22468203
Perkovic	-7.626406e-07	-0.02123986	0.11953834

	`Ball Recoveries`	Price_CategoryLow
Gvardiol	-0.09144193	0.3108584
Hato	0.39454017	-0.9050532
Hinpacie	-1.01123653	-0.9598916
Inacio	-0.04574835	1.9841756
Koulierakis	0.42356565	-0.8857173
Lukeba	-0.14891864	-2.5365371
Meijer	0.06757530	1.3872255
Pacho	0.66836069	-0.4097550
Perkovic	0.32304983	1.3047185

https://doi.org/10.26219/heal.aueb.6479

```
Std. Errors:
             (Intercept)       Season      Minutes
Gvardiol    9.786502e-14 1.614751e-13 5.524460e-11
Hato         6.080850e-12 7.613343e-12 4.528456e-09
Hinpacie    3.672233e-13 3.296799e-12 4.106106e-12
Inacio       1.811058e-13 2.044195e-13 3.288487e-13
Koulierakis 1.572949e-11 2.765729e-11 2.206804e-10
Lukeba       1.900681e-12 1.745002e-12 2.633250e-12
Meijer       1.060885e-12 2.328898e-12 9.258235e-10
Pacho        1.419008e-12 1.542615e-12 1.023402e-10
Perkovic    7.176276e-12 1.534163e-11 5.222712e-09
                   Goals      Assists `Clean Sheets`
Gvardiol    8.322476e-14 3.124238e-14   3.695462e-13
Hato         2.093030e-12 1.349613e-12   9.895903e-12
Hinpacie    3.628540e-13 6.616836e-13   2.738349e-12
Inacio       1.717754e-13 1.136565e-13   1.380984e-12
Koulierakis 2.658026e-11 1.178016e-11   1.333991e-10
Lukeba       3.726036e-12 7.049087e-14   1.233920e-11
Meijer       5.351547e-13 4.237804e-13   3.345671e-12
Pacho        9.337216e-13 2.286549e-12   7.063974e-12
Perkovic    1.954563e-11 1.137401e-11   9.904213e-11
```

```
             `Ground Duels` `Aerial Duels`     Tackling
Gvardiol      2.140365e-12   9.053165e-13 2.178206e-12
Hato          1.643206e-10   1.720993e-10 5.438691e-11
Hinpacie      3.566839e-12   5.546447e-11 7.180312e-11
Inacio        1.480934e-12   3.218959e-12 2.177495e-12
Koulierakis   1.657411e-10   6.849853e-10 3.645669e-10
Lukeba        8.328646e-12   1.383241e-11 3.077157e-11
Meijer        3.617964e-11   2.500701e-11 1.981853e-11
Pacho         1.926175e-11   6.989390e-11 6.411742e-11
Perkovic      1.416742e-10   4.482166e-10 2.063011e-10
                   Fouls         Price  `Red Cards`
Gvardiol    1.094449e-12 5.544307e-06 5.714052e-20
Hato         8.921750e-11 1.580124e-05 9.308137e-12
Hinpacie    9.842440e-12 9.765835e-06 8.756412e-12
Inacio       1.228480e-12 7.816712e-06 1.648153e-13
Koulierakis 3.990513e-10 7.764758e-05 1.289141e-14
Lukeba       1.492283e-11 6.338586e-06 5.025659e-13
Meijer       1.163523e-11 8.712428e-06 4.957446e-19
Pacho        2.314307e-11 3.379460e-05 4.706847e-15
Perkovic    2.612416e-10 3.345837e-05 2.574221e-18
```

https://doi.org/10.26219/heal.aueb.6479

```
            `Yellow Cards` `Ball Recoveries`
Gvardiol      1.370029e-13      6.598947e-12
Hato          1.201169e-11      5.064718e-10
Hinpacie      2.552245e-11      7.424530e-11
Inacio        8.579697e-13      9.843750e-12
Koulierakis   5.836905e-11      1.789599e-09
Lukeba        5.158748e-12      5.943053e-11
Meijer        1.596319e-12      7.288296e-11
Pacho         3.886936e-12      1.784878e-10
Perkovic      3.874189e-11      1.102352e-09
            Price_CategoryLow
Gvardiol         5.698531e-14
Hato             9.089197e-12
Hinpacie         6.676871e-16
Inacio           3.797374e-14
Koulierakis      1.571545e-11
Lukeba           1.755906e-12
Meijer           1.355405e-12
Pacho            1.417584e-12
Perkovic         7.176276e-12

Residual Deviance: 0.0002099562
AIC: 270.0002
```

We can clearly see the also in the second model, Perkovic is the second preferable player though the difference with the first one (Meijer) is not big one.

https://doi.org/10.26219/heal.aueb.6479

4. Conclusion

In all industries it is important to have predictive tools to make the best decisions. Statistics provides various tools, which are very useful for decision making. Estimating suitable forecasting models can be very useful in different industries. Regression models may estimate models of a different form. The appropriate form is determined by the type of dependent variable and the relationship it has with the independent variables.

In this work, the data concerns 10 players and 2 seasons. They include different variables related to the progression of the game. Our goal is to identify the factors that determine total performance taking under consideration the financial value and the rest of the selected variables so that the optimal player is selected for a team interested to make a transfer in any given position (in our case a left center back positioned player). For the above needs it is found that linear regression analysis and polynomial regression analysis were used.

In linear regression 2 models were estimated in the first model, the value was estimated and none of the independent variables were found to be statistically significant. In the second linear model, it is found that goals depend positively on fouls and negatively on tackling.

In the polynomial regression the dependent variable is each player. The results show that Perkovic is the most preferable among the selected players for our analysis. Both multinomial models agree in the aforementioned statement.

The research includes several limitations. The sample is limited, and more players and more seasons should be included. For future research it is suggested to use machine learning tools for the hierarchical analysis of players. Apart from that one could further research the definition of a Total Performance variable so that the optimal player is selected according to ones needs.

https://doi.org/10.26219/heal.aueb.6479

Appendix

```
data=read.csv(file.choose(),sep=";")

head(data)

str(data)

install.packages("sjPlot")

library(sjPlot)

sjt.xtab(age.category,BMI.category, variableLabels =

c("Age", "BMI Frequency"), showRowPerc=TRUE, showColPerc=TRUE)

library(sjmisc)

descr(data[,-c(1,2)],out='v')

sjt.df(data)

data$Price=data$Price/1000000

par(mfrow=c(3,2))

hist(data$Minutes,main="Histogram of Minutes",col="darkred",prob=TRUE)

x=data$Minutes

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Goals,main="Histogram of Goals",col="darkred",prob=TRUE)

x=data$Goals

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Assists,main="Histogram of Assists",col="darkred",prob=TRUE)

x=data$Assists

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Clean.Sheets,main="Histogram                                    of
Clean.Sheets",col="darkred",prob=TRUE)

x=data$Clean.Sheets
```

https://doi.org/10.26219/heal.aueb.6479

```
curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Ground.Duels,main="Histogram of
Ground.Duels",col="darkred",prob=TRUE)

x=data$Ground.Duels

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Aerial.Duels,main="Histogram of
Aerial.Duels",col="darkred",prob=TRUE)

x=data$Aerial.Duels

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

par(mfrow=c(3,2))

hist(data$Tackling,main="Histogram of Tackling",col="darkred",prob=TRUE)

x=data$Tackling

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Fouls,main="Histogram of Fouls",col="darkred",prob=TRUE)

x=data$Fouls

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Price,main="Histogram of Price",col="darkred",prob=TRUE)

x=data$Price

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Red.Cards,main="Histogram of Red.Cards",col="darkred",prob=TRUE)

x=data$Red.Cards

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Yellow.Cards,main="Histogram of
Yellow.Cards",col="darkred",prob=TRUE)

x=data$Yellow.Cards
```


https://doi.org/10.26219/heal.aueb.6479

```
curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

hist(data$Ball.Recoveries,main="Histogram of Ball.Recoveries",col="darkred",prob=TRUE)

x=data$Ball.Recoveries

curve(dnorm(x,mean(x),sd(x)),add=TRUE,col="darkgreen")

tapply(data$Minutes, data$Player, mean)

tapply(data$Goals, data$Player, mean)

tapply(data$Assists, data$Player, mean)

tapply(data$Clean.Sheets, data$Player, mean)

tapply(data$Ground.Duels, data$Player, mean)

tapply(data$Aerial.Duels, data$Player, mean)

tapply(data$Tackling, data$Player, mean)

tapply(data$Fouls, data$Player, mean)

tapply(data$Price, data$Player, mean)

tapply(data$Red.Cards, data$Player, mean)

tapply(data$Yellow.Cards, data$Player, mean)

tapply(data$Ball.Recoveries , data$Player, mean)

data$Season=as.factor(data$Season)

par(mfrow=c(4,3))

tapply(data$Minutes, data$Season, shapiro.test)

var.test(data$Minutes~data$Season)

t.test(data$Minutes~data$Season, main="Munites by Season",var.equal=TRUE)

install.packages("gplots")

library(gplots)

plotmeans(data$Minutes~data$Season,main="Minutes by Season")
```

https://doi.org/10.26219/heal.aueb.6479

```
tapply(data$Goals, data$Season, shapiro.test)

wilcox.test(data$Goals~data$Season)

boxplot(data$Goals~data$Season, main="Goals by Season", col=2:3)

tapply(data$Assists, data$Season, shapiro.test)

wilcox.test(data$Assists~data$Season)

boxplot(data$Assists~data$Season, main="Assists by Season", col=2:3)

tapply(data$Clean.Sheets, data$Season, shapiro.test)

var.test(data$Clean.Sheets~data$Season)

t.test(data$Clean.Sheets~data$Season, main="Clean Sheets by
Season",var.equal=FALSE)

plotmeans(data$Clean.Sheets~data$Season,main="Clean Sheets by Season")

tapply(data$Ground.Duels, data$Season, shapiro.test)

var.test(data$Ground.Duels~data$Season)

t.test(data$Ground.Duels~data$Season, main="Ground Duels by
Season",var.equal=FALSE)

plotmeans(data$Ground.Duels~data$Season,main="Ground Duels by Season")

tapply(data$Aerial.Duels, data$Season, shapiro.test)

var.test(data$Aerial.Duels~data$Season)

t.test(data$Aerial.Duels~data$Season, main="Aerial Duels by
Season",var.equal=TRUE)

plotmeans(data$Aerial.Duels~data$Season,main="Aerial Duels by Season")

tapply(data$Tackling, data$Season, shapiro.test)

wilcox.test(data$Tackling~data$Season)

boxplot(data$Tackling~data$Season, main="Tackling by Season", col=2:3)

tapply(data$Fouls, data$Season, shapiro.test)
```

https://doi.org/10.26219/heal.aueb.6479

```
wilcox.test(data$Fouls~data$Season)

boxplot(data$Fouls~data$Season, main="Fouls by Season", col=2:3

tapply(data$Price, data$Season, shapiro.test)

wilcox.test(data$Price~data$Season)

boxplot(data$Price~data$Season, main="Price by Season", col=2:3)

tapply(data$Red.Cards, data$Season, shapiro.test)

wilcox.test(data$Red.Cards~data$Season)

boxplot(data$Red.Cards~data$Season, main="Red Cards by Season", col=2:3)

tapply(data$Yellow.Cards, data$Season, shapiro.test)

var.test(data$Yellow.Cards~data$Season)

t.test(data$Yellow.Cards~data$Season, main="Yellow Cards by Season",var.equal=FALSE)

plotmeans(data$Yellow.Cards~data$Season,main="Yellow Cards by Season")

tapply(data$Ball.Recoveries, data$Season, shapiro.test)

wilcox.test(data$Ball.Recoveries~data$Season)

boxplot(data$Ball.Recoveries~data$Season, main="Ball Recoveries by Season", col=2:3)

datanum<-data[,-c(1,2)]

install.packages("corrplot")

 library(corrplot)

datanum<-na.omit(datanum)

corrplot(cor(datanum), method="shade",shade.col=NA, tl.col="black", tl.srt=45)

fullmodel1 <- lm(data$Price~. , data= data[,-c(1,11,14)])

summary(fullmodel1)

install.packages("car")
```

https://doi.org/10.26219/heal.aueb.6479

```
library(car)
#polisigrammikotita
vif(fullmodel1)
par(mfrow=c(2,2))
#normality
plot(fullmodel1,which=2)
#homoskedasticity-linearity
plot( fullmodel1$fit, rstandard(fullmodel1), cex=2)
abline( h= 1.96, col='red', lwd=2, lty=2 )
abline( h=-1.96, col='red', lwd=2, lty=2 )
quantcut<-function(x){cut(x,breaks=quantile(x),include.lowest=T)}
qfits<-quantcut(fullmodel1$fit)
leveneTest(rstandard(fullmodel1),qfits)#p>5%-omoskedastikotita
#outlier
plot(fullmodel1,which=4)
outlierTest(fullmodel1)
abline(h=4/(nrow(data)-2-1),col='blue',lty=2)
plot(fullmodel1,which=5)
residualPlots(fullmodel1)
install.packages("sjPlot")
library(sjPlot)
library(sjmisc)
tab_model(fullmodel1)
str(data)
```

https://doi.org/10.26219/heal.aueb.6479

```
fullmodel1 <- lm(data$Goals~. , data= data[,-c(1,4,14)])

summary(fullmodel1)

install.packages("car")

library(car)

#polisigrammikotita

vif(fullmodel1)

par(mfrow=c(2,2))

#normality

plot(fullmodel1,which=2)

#homoskedasticity-linearity

plot( fullmodel1$fit, rstandard(fullmodel1), cex=2)

abline( h= 1.96, col='red', lwd=2, lty=2 )

abline( h=-1.96, col='red', lwd=2, lty=2 )

quantcut<-function(x){cut(x,breaks=quantile(x),include.lowest=T)}

qfits<-quantcut(fullmodel1$fit)

leveneTest(rstandard(fullmodel1),qfits)#p>5%-omoskedastikotita

#outlier

plot(fullmodel1,which=4)

outlierTest(fullmodel1)

abline(h=4/(nrow(data)-2-1),col='blue',lty=2)

plot(fullmodel1,which=5)

residualPlots(fullmodel1)

install.packages("sjPlot")

library(sjPlot)
```

https://doi.org/10.26219/heal.aueb.6479

```
library(sjmisc)

tab_model(fullmodel1)

str(data)

data$Goal2=c(1,1,0,0,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1)

library(dplyr)

library(readr)

library(nnet)

df <- df[,2:ncol(df)]

df <- df[1:20,]

str(df)


df$Price <- gsub("\\.", "", df$Price)

df <- df %>%

 mutate_at(vars(-Player), as.numeric)


df$TotalPerformance <- df$Goals*1 + df$Assists*1 + df$`Clean Sheets`*1 + df$`Ground Duels`*1 + df$`Aerial Duels`*1 + df$Tackling*1 - df$Fouls*1 - df$`Red Cards`*1 - df$`Yellow Cards`*1 + df$`Ball Recoveries`*1- df$Price*1


model <- multinom(Player ~ TotalPerformance, data = df)

summary(model)


df$Price_Category <- as.factor(ifelse(df$Price<=30000000, 'Low', 'High'))
```

https://doi.org/10.26219/heal.aueb.6479

```
model2 <- multinom(Player ~ Season + Minutes + Goals + Assists + `Clean Sheets` + `Ground Duels` + `Aerial Duels` + Tackling + Fouls + Price + `Red Cards` + `Yellow Cards` + `Ball Recoveries` + `Price_Category`, data = df)

summary(model2)
```

References

1. Berger, A., S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39–71.

2. Bishop, C. M. 2006. Pattern recognition and machine learning. Springer. Cox, D. 1969. Analysis of Binary Data. Chapman and Hall, London.

3. Hastie, T., R. J. Tibshirani, and J. H. Friedman. 2001. The Elements of Statistical Learning. Springer.

4. Kehler, A. 1997. Probabilistic coreference in information extraction . EMNLP .

5. Murphy, K. P. 2012. Machine learning: A probabilistic perspective. MIT Press. Ng, A. Y. and M. I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. NeurIPS.

6. Nigam, K., J. D. Lafferty, and A. McCallum. 1999. Using maximum entropy for text classification. IJCAI-99 workshop on machine learning for information filtering.

7. Ratnaparkhi, A. 1996. A maximum entropy part-of-speech tagger . EMNLP

8. Ratnaparkhi, A. 1997. A linear observed time statistical parser based on maximum entropy models . EMNLP.

https://doi.org/10.26219/heal.aueb.6479

9. Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language, 10:187–228.

10. Sankoff, D. and W. Labov. 1979. On the uses of variable rules. Language in society, 8(2-3):189–222.

11. Schutze, H., D. A. Hull, and J. Pedersen. 1995. ¨ A comparison of classifiers and document representations for the routing problem . SIGIR-95 .

12. Tibshirani, R. J. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288.

13. Wang, S. and C. D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification . ACL .

14. Witten, I. H. and E. Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann.