Faculty of Economics and Business Studies

AN ECONOMETRIC ANALYSIS OF THE RELEGATION OF FOOTBALL CLUBS BASED ON SOCIAL MEDIA AND GAMING DATA

Clave: 202119513

MADRID | March 2023

INDEX

INDEX OF FIGURES

SUMMARY

This thesis studies the statistical correlation between the relegation of football clubs based on social media and gaming data, with specific emphasis on the period 2016-2022, looking at the relevancy of social media and gaming in the prediction of a football club's relegation. It begins with a review of existing literature, taking a wide range of domestic and international sources into account, across a variety of areas. An ordinary least squares model is then carried out to assess the relationship between gaming, social media data, and relegation, between two leagues – The League of Ireland Premier Division and La Liga. The study is concluded with an analysis and discussion, linking the findings from the previous parts and discussing potential solutions to this issue.

Key Words: Relegation, Gaming, Social Media, Performance, League Position, Correlation

1. INTRODUCTION

The research work we are presenting focuses on analysing the statistical correlation of the relegation of football clubs based on social media and gaming data. Specifically, through these pages we will try to determine whether social media and gaming data play a significant role in the relegation of football clubs in the top divisions of two countries, Ireland and Spain, and if we can use this data to understand, predict and accurately evaluate the relegation of football clubs from these two leagues.

In recent years, social media and gaming have become an integral part of the football industry, with millions of fans engaging with their favourite clubs and players through these platforms. The ability to monitor and analyse data from these sources has opened new possibilities for researchers and analysts to gain insights into various aspects of the sport, including team performance and fan behaviour. I have chosen the top divisions in Ireland and Spain as both countries have a lot of meaning for me, being an Irish national studying and living in Spain.

The research work we are presenting focuses on investigating whether there is a statistical correlation between social media and gaming data and the relegation of football clubs. Specifically, we will examine whether there are any patterns or trends in the data that can be used to predict the likelihood of a club being relegated from its league. By doing so, we hope to contribute to the growing body of knowledge on the use of data analytics in the football industry and provide insights that can be used by club owners, coaches, and analysts to make better decisions.

To achieve our research objectives, we will use a combination of statistical analysis techniques to analyse social media and gaming data from a sample of football clubs in the top division of the Irish and Spanish domestic leagues for the previous six years, 2016 to 2022. We will also review existing literature on the topic and explore the current state of data analytics in the football industry.

Overall, this thesis aims to provide a comprehensive analysis of the relationship between social media and gaming data and the relegation of football clubs from the League of Ireland Premier Division and Spain's La Liga. By doing so, we hope to shed light on the potential of these

metrics for data analytics in the football industry and contribute to the development of new tools and methods for predicting and managing performance.

2. STATE OF THE ART

This thesis is a unique and innovative approach to understanding the factors that contribute to the relegation of football clubs in the League of Ireland Premier Division and Spanish La Liga between 2016 and 2022. The study aims to explore the potential of social media and gaming data as predictors of team performance, and to determine whether these factors have a statistical correlation with the relegation of football clubs.

In recent years, there has been a growing interest in using data analytics to analyse and predict team performance in sports. The availability of large datasets, combined with advanced statistical techniques, has made it possible to identify patterns and trends in team performance, and to develop predictive models that can help teams optimize their strategies and improve their chances of success. Recent technological enhancements (e.g., game analysis software, remote sensor technology or motion tracking systems), have been developed in conjunction with novel statistical approaches (i.e., predictive and stochastic methods) to model, infer or predict performance outcomes in sport (Nevill et al., 2008).

The use of social media and gaming data is a relatively new approach to sports analytics, but it has already shown promise in other sports such as basketball and American football. Social media data, for example, can provide insights into fan engagement, sentiment, and club popularity, while gaming data can offer an objective measure of player skill and performance. The New England Patriots track social media data ranging from what fans buy at the pro shop to when they buy tickets. By crunching those numbers with the help of the Kraft Analytics Group, they can predict everything from ticket pricing to staffing on game day (Ricky, n.d.). However, despite the remarkable growth in the amount and variety of data available for examination and analysis, the world of sports analytics still faces the same ubiquitous challenge: How to get meaningful information into the hands – and minds – of the people who are in a position to make effective use of it (Travassos et al., 2013).

The League of Ireland Premier Division and Spanish La Liga are two of the most popular football leagues in Europe, with a rich history of competition and passionate fan bases. By analysing the data form these two leagues, the study can provide valuable insights into factors that contribute to team performance and relegation.

It is important to note that the Spanish top division is much more commercialised due to its prestige, and a lot more structured than the Irish league. Unlike in Spain, football isn't one of the most popular sports to play in Ireland and the League of Ireland has struggled to catch up to other top European leagues due to the small population and lack of quality and interest. Recently, the country's capital city, Dublin, had been chosen to host four Euro 2020 matches which would have been the first time that an international football tournament of that level would have involved an Irish hosting. According to the Football Association of Ireland, they could not guarantee "minimum spectator numbers" for the four proposed matches which were subsequently reallocated to St. Petersburg (Curran, 2022).

The Football Association of Ireland (FAI) has been making efforts for a number of years in an effort to increase the level of the country's top football league, and while there have been many encouraging signs in the growth of attendances, having increased by nearly a third since prior to the Covid-19 pandemic, the truth is that it is considered a very lacklustre league in comparison to the big five leagues in Europe – England, Spain, Germany, France, and Italy. While an additional 110,660 fans attended League of Ireland top-flight fixture in 2022 compared to 2019 (Peel et al., 1998), they are starting from a very low base. It is quite shocking that Ireland's top division attendances are lagging behind the fifth division in England, and an increased fanbase is one of the foundations upon which a better product can be built; the average top-flight attendance for the first nine games in the League of Ireland this year was 2,878, while England's National League figure is 3,018 (*Growth of the LOI*, 2022)

Meanwhile, La Liga recently celebrated the fifth anniversary of its international structure that has driven the competition's international expansion and made the Spanish league into a world-leading reference in the sports industry when it comes to internationalisation (*Five Years of LaLiga's International Expansion with Record Growth*, n.d.). In 2017, the La Liga Global Network program was launched and currently operates in forty-one countries with forty-four delegates working on-site. The program's internationalisation project includes eleven international offices and two joint ventures in North America and China. La Liga's international expansion strategy is ambitious and now encompasses ninety countries. This has been a key factor in the league's growth, bringing the spectacle of La Liga to every corner of the planet. As a result, more and more people are watching the league and enjoying its matches every week.

The study involves collecting and analysing a large amount of data from social media platforms such as Twitter and Instagram, as well as from gaming products such as FIFA and Football Manager. The data will then be analysed using advanced statistical techniques such as regression analysis. The potential implications of the study are significant. If the study finds a statistically significant correlation between social media and gaming data and team performance, it could provide teams with valuable insights into how to optimize their strategies and improve their chances of success. It could also lead to the development of new metrics and tools for sports analytics, which could be applied across a wide range of sports and leagues.

In summary, this thesis is a cutting-edge and exciting approach to sports analytics, with the potential to provide valuable insights into the factors that contribute to team performance and relegation in two of Europe's most popular football leagues.

3. SCOPE OF THE PROJECT

3.1. HYPOTHESES

Now that we understand the current knowledge on the subject matter through the analysis of related published work, we must propose our hypothesis for the research.

Our hypothesis (H1) is that the strength and direction of the correlations between social media engagement, gaming performance, and relegation vary between the League of Ireland Premier Division and Spanish La Liga. The alternative hypotheses (Ha) suggests that there are significant differences in the correlations between the two leagues. The null hypothesis (H0), on the other hand, proposes that there are no significant differences in the correlations between the two leagues.

To test this hypotheses, the study will analyse the data from both leagues and compare the correlations between social media engagement, gaming performance, and relegation. The study will examine the patterns and trends in the data to identify any significant differences between the two leagues.

If the study finds a statistically significant difference in the correlations between the two leagues, it would suggest that there are unique factors that influence the performance and relegation of football clubs in each league. These factors could include differences in the football culture, fan behaviour, economic conditions, or other contextual factors that shape the performance of teams.

On the other hand, if the study fails to find a significant difference between the correlations in the two leagues, it would suggest that the factors that influence the performance and relegation of football clubs are similar across the two leagues. This would have implications for the development of predictive models and analytical tools that could be applied across multiple leagues.

In summary, H1 proposes that there are differences in the correlations between social media engagement, gaming performance, and relegation in the League of Ireland Premier Division and Spanish La Liga, while the null hypothesis (H0) suggests that there are no significant

differences. The study will analyse the data from both leagues to test these hypotheses and provide insights into the factors that shape the performance and relegation of football clubs in these leagues

3.2. ASSUMPTIONS

This study is based on several key assumptions. Firstly, the study assumes that social media engagement is a reliable indicator of fan interest, sentiment, and engagement with football clubs. This assumption is based on the premise that social media platforms have become a central communication channel between football clubs and their fans, and that metrics such as followers, likes, shares, and comments can provide insights into the level of fan engagement and sentiment towards clubs.

Secondly, the study assumes that gaming performance isa reliable indicator of the quality and performance of football clubs. This assumption is based on the idea that modern football games have become increasingly realistic and sophisticated, and that metrics such as win percentage, goal difference, and player ratings can provide insights into the quality and performance of teams in real-life competitions.

Thirdly, the study assumes that relegation is a meaningful and significant outcome in football leagues, and that it reflects the ability of teams to compete at a certain level. This assumption is based on the idea that relegation is a major event in the football calendar, and that it has significant financial, social, and sporting implications for clubs and their fans.

Fourthly, the study assumes that the statistical techniques and methods used to analyse the data are valid and reliable. This assumption is based on the premise that the study will use established statistical methods, such as correlation analysis, regression analysis and hypothesis testing, to analyse the data and test the hypotheses.

Finally, the study assumes that the data used in the analysis is complete, accurate and reliable. This assumption is based on the ides that the data will be collected from reputable sources and will include all relevant variables and observations needed for the analysis.

In summary, the thesis is based on several assumptions, including the reliability of social media engagement and gaming performance as indicators of fan engagement and team quality, the significance of relegation as a football outcome, the validity of statistical methods and techniques, and the completeness and accuracy of the data used in the analysis. These assumptions provide the basis for the study and will be evaluated and tested throughout the research process.

3.3. CONSTRAINTS

Moving on from the assumptions of the study, we look at constraints of the research. This study is subject to several constraints. Firstly, the study is limited to the League of Ireland Premier Division and Spanish La Liga for the selected years, and the findings may not be generalizable to other football leagues or time periods. This limitation is due to the availability of data and the scope of the study.

Secondly, the study relies on publicly available data on social media engagement and gaming statistics, which may not capture all relevant variables and factors that influence the performance and relegation of football clubs. This limitation is inherent in any data-driven study and may limit the accuracy and validity of the findings.

Thirdly, the study may face challenges in establishing causality between social media engagement, gaming statistics, and relegation, as there may be other variables and factors that influence the performance and relegation of football clubs. This limitation is inherent in correlation analysis and may require further research to establish causality.

Fourthly, the study may face challenges in accessing and analysing the data due to issues such as data privacy, data quality, and data availability. These challenges may limit the scope and quality of the analysis and may require alternative approaches to data collection and analysis.

Finally, the study may be constrained by the resources, time, and expertise available to the researcher. The thesis may require significant resources, including time, funding, and specialized skills, to collect, analyse, and interpret the data.

The study is subject to several constraints, including the limited scope and generalizability of the findings, the limitations of publicly available data, the challenge of establishing causality, issues related to data access and analysis, and resources constraints. These constraints may limit the accuracy, validity, and scope of the study and may require careful consideration and mitigation throughout the research process.

3.4. OBJECTIVES

Finally, we set our objectives for the study which will be a very consistent thread throughout our data collection, research, and analysis phases. The main objective of this study is to investigate the statistical correlation between social media and gaming data in relation to the relegation of football clubs in the League of Ireland Premier Division and Spanish La Liga for the years 2016 to 2022. Specifically, the thesis aims to explore whether there is a significant relationship between social media engagement and gaming data with the relegation of football clubs, and whether this relationship differs across the two leagues.

To achieve this objective, we set out several sub-objectives, including:

1. To collect and analyse data on social media engagement and gaming statistics of football clubs in the League of Ireland Premier Division and La Liga for the years 2016 to 2022.
2. To identify the factors that influence the performance and relegation of football clubs in the two leagues and to control for these factors in the analysis.
3. To conduct a correlation analysis between social media engagement, gaming statistics, and the relegation for football clubs, and to assess the significance and strength of these correlations.
4. To compare the correlations between the two leagues and to explore whether there are significant differences in the relationships between social media audiences, gaming statistics, and relegation in the two leagues.
5. To interpret and discuss the findings and to draw conclusions about the relationship between social media and gaming data, and relegation of football clubs, as well as to identify the implications of the study for football clubs, fans, and social media and gaming companies.

By achieving these objectives, the thesis aims to contribute to the understanding of the factors that influence the performance and relegation of football clubs and to shed light on the potential role of social media and gaming data in this process. Additionally, the study aims to provide insights for football clubs, fans, and social media and gaming companies on the importance of these factors in predicting the outcomes of football matches and seasons.

4. THE SIGNIFICANCE AND THE EFFECTS OF RELEGATION

Relegation is a significant event in the world of football as being demoted to a lower league has far-reaching effects on the club, its players, and fans. Relegation can impact a team's financial stability, player morale, and the overall reputation of the club. Gasparetto and Barajas' studies into the impacts of relegation have led to some compelling findings, with results showing that evidence that, in general, promotions to the second tier tend to increase attendance and revenues of State Championships as well as relegation to lower divisions impact them negatively (Gasparetto & Barajas, 2022).

One of the most significant effects of relegation is the financial impact on football clubs. The demotion to a lower league often results in a significant drop in revenue, as clubs lose out on broadcasting and sponsorship deals associated with the higher league. This can lead to a decrease in the quality of players and facilities, which in turn makes it harder for the club to compete at the lower level. Additionally, the financial hit can result in the loss of key players, as they may seek moves to other clubs that can offer better salaries and the opportunity to compete at a higher level.

Another significant effect of relegation is the impact it can have on the morale and confidence of a team's players. Relegation can be a demoralizing experience for players who are used to competing at a higher level, and this can have a knock-on effect on their performance on the pitch as they are summoned to at least one year of lower level competition. Players may become disheartened and demotivated, which can result in a decline in the quality of their play. This can create a vicious cycle, as poor performances can lead to more losses, which can further damage morale and confidence.

Relegation can also have a significant impact on a club's reputation. A team that is relegated from a higher league may be seen as a weaker team, and this can impact the club's ability to attract new fans and sponsors. Additionally, relegation can enhance a sense of disappointment and frustration among existing fans, which can lead to a decline in attendance and support for the team.

The impact of football has even been shown to affect admission rates to medical facilities. To go even further, it seems that there is a measurable response in football fans following high-

stakes games, such as a final game in a league season which could determine the fate of relegation for a football club. Banyard and Shevlin have published pieces on the psychological effect of relegation, assessed by the Impact of Events Scale (IES) which was designed to assess the impact of any specific traumatic event. They concluded that the mean scores in the study suggest that the psychological consequences of relegation can be significant (Banyard & Shevlin, 2001). Furthermore, the study indicated that over half the sample (51%9 indicated responses that are clinically significant with 11% of those suffering severe psychological distress.

Relegation can also have quiet significant effects on the economic development of a relatively small area. When a club is in a top division, it becomes easier to attract better players on higher salaries, the majority of whom come from outside the region and usually have a short-term contract. If we could identify the part of their salaries which effectively become incorporated into the regional economic flow, then the wealth-generation and employment effects would be even lower (Conejo et al., 2007). Similarly, there seems to be a relationship between a club's performances and the local tourism industry of their area, with the contribution of the tourist sector to the economy falling as a result of the reduced attractiveness of the teams in more rural, less-developed areas.

When we examine the impact of relegation on specific leagues, we can see that the effects can be particularly pronounced in certain contexts. In the League of Ireland Premier Division, relegation can have a significant impact on smaller clubs, as the financial hit can be particularly severe. Additionally, the loss of players due to relegation can make it difficult for these clubs to rebuild and compete at the lower level.

Due to the League of Ireland being a relatively small league in terms of the numbers of competitors, the repetition of the relegation of some of the smaller clubs is higher than that of a bigger European league such as La Liga. With the League of Ireland Premier Division containing ten clubs, with two clubs being relegated each year, it appears to be the same four to six clubs that get regularly relegated form the top division. This constant promotion and demotion isn't helpful for the growth of these football clubs, leaving time for the bigger clubs that don't get relegated to develop with consistent revenue streams and fan loyalty.

In contrast, in Spanish La Liga, relegation can impact even the biggest clubs. This is due to the competitive nature of the league, which means that even the most successful teams are not immune to the threat of relegation. The financial impact of relegation can be particularly significant for clubs with high wage bills, which can make it harder for them to recover from the demotion.

Studies on the effects of relegation have been explored by Barajas (et al., 2005) who states that while relegated clubs have the advantage of enjoying greater incomes than the rest of clubs in their division, they have to face up to inflexible wage costs and run the increased financial risk if the clubs continue with high wage expenses in order to try and achieve a quick promotion return. The mention of greater incomes for relegated clubs from Barajas (et al., 2005) refers to parachute payments, which are available to La Liga clubs, but not to League of Ireland clubs yet. A parachute payment is a payment made to alleviate hardship resulting from a sudden loss of income (Collins English Dictionary, 2023). La Liga has established a fund – representing 3.5% of the total broadcasting rights – allocating money among the relegated clubs depending on various criteria including past broadcasting revenue and the number of seasons they've spent in La Liga (Football Benchmark, 2018).

5. THEORETICAL MODEL

5.1. LITERATURE REVIEW

Firstly, I intend to examine some of the key elements of an Ordinary Least Squares (OLS) model, setting the scene for an effective tool in the statistical analysis of gaming and social media data in relation to football clubs. From the outset, it's crucial to gain an understanding of the purpose of an OLS model and how it has been used in research up to this point.

Ordinary least squares (OLS) computational methods are commonly used to test hypotheses of differences among factor-level means in repeated measures data, and are available in a variety of commercial statistical software packages, generally under the rubric of general linear model (GLM) (Ugrinowitsch et al., 2004). It is a popular method for analysing data in economics, engineering, and social sciences as it can estimate the linear relationship between a dependent variable and one or more independent variables. In this sense, it will be very useful for my research and analysis as I can measure the linear correlation between a variable which signifies a football club's relegation, and multiple other independent variables related to the gaming and social media data which I will collect.

OLS will play a critical and insightful role in my study, as the Gauss-Markov theorem shows that, if your data fulfil certain requirements, OLS is the best linear unbiased estimator available ("Ordinary Least Squares (OLS)," 2016). In this regard, it is important to start with the assumptions that the OLS model makes. Next, I will examine the OLS model once it is estimated, the coefficients of the independent variables, and the goodness of fit of the model. I conclude the literature review with case study analysis of the relationship between league performance and the shares of publicly-traded football clubs, a slight alteration of my own hypothesis.

5.1.1. Assumptions of the Ordinary Least Squares Model

Multiple OLS regression analysis is a standard (and useful) statistical technique which is widely employed in marketing and related business research. However, a key assumption underpinning its use is that the dependent variable of interest is measured on a continuous, interval scale (i.e., the measurement has both order and is meaningful in respect of the distance between values) (Peel et al., 1998). If this OLS assumption is violated (e.g., where the

dependent variable is categorical), then a number of serious problems may arise with the OLS model, such as meaningless predictions outside of the range, or between the values, of the scale of the nominal or categorical dependent variable.

Another assumption is that the regression model is linear on the coefficients and the error term. This assumption pertains to the functional form of the model. In the field of statistics, a regression model is deemed linear when all components in the model comprise either the constant or a parameter that is multiplied by an independent variable. The model equation is constructed by summing up the terms according to these regulations, thereby restricting the model to a singular type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Figure 1. The model equation, according to the regulations of the assumption. (Frost, 2018)

In the equation, the betas (βs) are the parameters that OLS estimates. Epsilon (ε) is the random error. The defining characteristic of linear regression is this functional form of the parameters rather than the ability to model curvature. Linear models can model curvature can model curvature by including "nonlinear variables such as polynomials" and "transforming exponential functions" (Frost, 2018).

In order to satisfy this assumption, the correctly specified model must fit the linear pattern.

5.1.2. The examination of the OLS Model

The OLS equation is a mathematical formula used in linear regression analysis to estimate the parameters of a linear equation. While there are some variations to this equation, like in figure 1, the best form of the equation is as follows:

$$\beta = (X'X)^{(-1)}X'y$$

β = a vector of estimated coefficients
X = a matrix of explanatory variables (also known as the design matrix)
X' = the transpose of X

Y = a vector of observed values for the dependent variable
^(-1) = the matrix inverse

This equation is used to find the values of β that minimize the sum of squared errors between the predicted values of the dependent variable and the actual values. The resulting equation can be used to make predictions for the new values of the explanatory variables.

On the topic of the limitations of the OLS model, it should be noted that the method has certain constraints due to the inversion of the X'X matrix. The primary requirement for this matrix is that its rank should be p+1, where p is the number of predictor variables in the model. However, if the matrix is not well behaved, numerical issues may arise, making it difficult to obtain reliable results.

To address these limitations, some algorithms have been developed, such as those proposed by Dempster (1969), which can be used in the XLSTAT and Gretl software (*Ordinary Least Squares Regression (OLS)*, n.d.). These algorithms provide a way to overcome the issues of the X'X matrix by removing certain variables from the model. Specifically, if the rank of the matrix is q, where q is less than p+1, then some variables can be eliminated from the model. This can be due to the presence of constant variables or collinear variables that can be grouped together into a block.

5.1.3. The coefficients of the independent variables

In the OLS equation, the coefficients represent the estimated parameters of the linear relationship between the dependent variable and the explanatory variables. The coefficients are the slope of the line that describes the relationship between the independent variable(s) and the dependent variable. They indicate how much the dependent variable is expected to change for every unit change in the independent variable(s).

The coefficients are estimated by minimizing the sum of squared errors between the predicted values and the observed values of the dependent variable. This is done by finding the values of the coefficients that make the sum of the squared differences between the predicted and

observed values as small as possible. This is achieved by using matrix algebra to solve for the coefficients that minimize the residual sum of squares (RSS).

Once the coefficients have been estimated, they can be used to make predictions for new values for the independent variables. Specifically, the OLS equation ca be used to calculate the predicted value of the dependent variable for any given set of values of the independent variables. This prediction is based on the estimated coefficients and the observed values of the independent variables. However, we cannot say that a change in the dependent variable is caused by a change in the independent variable, only that they are associated with each other. That is, correlation does not imply causation (Chatterjee & Simonoff, 2013).

The coefficients have several important significance:

1. The coefficient of an independent variable measures the impact of that variable on the dependent variable, holding all other independent variables constant.
2. A positive coefficient indicates that the relationship between the independent and dependent variables is positive, meaning an increase in the independent variable will lead to an increase in the dependent variable.
3. A negative coefficient indicates that the relationship is negative, meaning an increase in the independent variable will lead to a decrease in the dependent variable.
4. The magnitude of the coefficient is also significant. The larger the coefficient, the greater the impact of the independent variable on the dependent variable.
5. Coefficient significance can also be tested for statistical significance using a t-test or F-test to determine whether the coefficient is significantly different from zero, indicating whether the variable has a significant impact on the dependent variable.

5.1.4. The Goodness-of-Fit of the Model

In multiple linear regression analysis, the goodness-of-fit of the model is an important aspect to consider in evaluating the accuracy and reliability of the predictions. The goodness-of-fit measures how well the regression model fits the observed data, and can be assessed using a variety of statistical test and metrics. While a logistic model may be a reasonable one for probabilities, it may not be appropriate for a particular data set. This is not the same thing as

saying that the predicting variables are not good predictors for the probability of success (Chatterjee & Simonoff, 2013).

Goodness-of-fit tests are designed to assess the fit of a model through the use of hypothesis testing. Such statistics test the hypotheses;

H0 : The linear logistic regression model fits the data
versus
H1 : The linear logistic regression model does not fit the data.

Such tests proceed by comparing the variability of the observed data around the fitted model to the data's inherent variability (Chatterjee & Simonoff, 2013).

One commonly used measure of goodness-of-fit is the R-squared statistic, which represents the proportion of the total variation in the dependent variable that is explained by the independent variables in the model. A high R-squared value indicates that the model explains a large amount of the variability in the data, while a low R-squared value indicates that the model may not be a good fit for the data.

However, R-squared alone does not provide a complete picture of the goodness-of-fit, and other measures such as the adjusted R-squared, the standard error of the estimate, and residual plots should also be considered. The adjusted R-squared takes into account the number of predictors (independent variables) in the model and adjusts the R-squared value accordingly. The standard error of the estimate measures the average distance between the observed values and the predicted values of the dependent variable. Residual plots can also be used to visually inspect the goodness-of-fit by plotting the residuals against the predicted values.

I have also seen the argument that when multiple observations for each model simulation are available, a lack-of-fit analysis can be added to the multiple linear regression analysis. The test of lack-of-fit would be computed using a likelihood ratio test. Most computer software that allow multiple regression observations will compute the tests using a Manova approach. Thus one would fit an appropriate model then use a Manova test to test lack-of-fit hypotheses (*PII*, n.d.).

Overall, evaluating the goodness-of-fit in multiple linear regression analysis is crucial for ensuring the accuracy and reliability of the predictions. By using multiple measures and test, researchers can gain a more complete understanding of how well the regression model fits the observed data and make appropriate adjustments to improve the model's performance.

5.1.5. Case Study Analysis

In order to understand a data analysis model in the football industry, I chose to analyse a panel data analysis of the relationship between league performance and the shares of the publicly-traded football clubs (Findikçi & Tapşin, 2015), which focuses on the relationship between sporting achievements and share earnings ratios of publicly traded football clubs in the Spor Toto Super League in Turkey.

In recent years, the sports industry has become an interdisciplinary field and an essential part of the economy the financial analysis of football in Turkey reveals that the Big Four clubs, with the most significant share of the market, have been listed on the Istanbul Stock Exchange to attract investors. Findikçi and Tapşin aim to fill the gap of the literature on this subject by analysing the performances of Spor Toto Super League football clubs over five seasons from 2010 to 2015. They use panel data analysis to identify the relationship between match results, sored and conceded goals, and share earnings ratios. Panel data analysis is a type of regression analysis that combines cross-sectional data (data from multiple sources at a single point in time) with longitudinal data (data from the same source over multiple time points) to create a more robust model.

The methodology used in this study is panel data analysis, which requires special data called panel data that provides values of each variable for two or more time intervals. Panel data analysis involves multiple regression data and is expressed mathematically as:

$$y_{it} = \alpha_i + x'_{it}\beta_{it} + \varepsilon_{it} \qquad t = 1,2,\ldots\ldots,T; \qquad i = 1,2,\ldots\ldots\ldots,n$$

Figure 2. The panel data equation for multiple linear regression (Findikçi & Tapşin, 2015)

Panel data allows for variables peculiar to individuals, firms, provinces, countries, etc. and takes heterogenous structures of these variables into account. Advantages of panel data over

cross-sectional or time series data include more information, more variability, lower levels of common linearity, higher and more efficient levels of degrees of freedom, more profound assessment of impacts, and analysis of complicated models of behaviour. Panel data is classified into balanced and unbalanced data, and this study uses balanced panel data. The three methods used in the estimation phase of panel regression are Pooled Panel Data Regression Model, Fixed Effects Model, and Random Effects Model. The Fixed Effects and Pooled Models have fixed slope parameters for all cross-section observations, while the Random Effects Model assumes error components are drawn randomly from a ground mass. The choice between the two models depends on the ease of calculation.

The resulting models are statistically significant, indicating that there is a relationship between football club performance and share earnings. Specifically, the authors find that the model formed with average points exhibits a better performance than the one formed in the light of the match results. As the average point rises, stock shares appreciate in value.

Before, we see the results of the of the investigation, it's important to know the current state of affairs that led to the study of this area. Incorporating football clubs brings about institutionalisation and professionalism, as well as diversification of income sources. This allows for entrance into the capital market, where new funds can be transferred to the clubs. The UK has the highest level of public offerings of sports clubs, and the market values of these clubs are quite high. The English model, which requires the entire sports club to be transferred to the newly established company, is accepted as a model for public offerings of sports clubs and is used by other clubs as well.

In Turkey, public offerings of sports clubs began with the incorporation of football clubs and the transfer of professional football clubs to newly founded or to-be-founded joint-stock companies. Fenerbahçe, Galatasaray, Beşiktaş, and Trabzonspor football clubs have been offered to the public. These clubs became eligible to be offered to the public after their incomes were transferred to football A.S. or sports A.S. (joint-stock companies) that were open to the stock market.

However, Findikçi and Tapşin revealed that, except for Beşiktaş, which adopted the English Model, all other clubs kept expenditures within the structure of their clubs and transferred only

their income to the newly founded stock companies. This could be an interesting point in explaining why clubs except for Beşiktaş are not much affected by the league. Since Beşiktaş has both income and expenditures in the stock market, its investors share not only their profits but also their losses. Therefore, investors can be psychologically affected by the situation of their clubs. However, investors of other clubs do not share losses, and they might not be interest in league performance because a faulty transfer case, excessive spending, or loss of a cup is not a particular concern for investors.

The match results of the clubs were evaluated in terms of their wins, defeats, draws, and the difference between the numbers of goals scored and conceded. Once again, it was explored via panel data analysis whether there is a relationship between these elements and share earnings. The resulting models were statistically significant, and it was seen that the model formed with the average point exhibits a better performance than the one formed based on match results. Additionally, as the average point rises, stock shares appreciated in value.

	Earnings	Win	Draw	Defeat	Average	Sporting return	BISTreturn
Earning	1.00						
Win	0.04	1.00					
Draw	-0.08	-0.05	1.00				
Defeat	-0.09	-0.05	-0.03	1.00			
Average	0.08	0.62	-0.02	-0.40	1.00		
Sporting return	0.67	-0.02	-0.03	0.00	-0.01	1.00	
BIST return	0.30	0.03	-0.01	0.00	0.03	0.42	1.00

Figure 3. Correlation Matrix of the variables (Findikçi & Tapşin, 2015)

Hence, it is possible to say that lopsided wins or defeats in leagues influence earnings, and investors care about the number of goals, which is one of the most important criteria in the evaluation of league performance. In the model formed based on match results, it was seen that stock shares appreciate after a win while they go through depreciation after defeats or draws. Moreover, the BIST Sporting Index and BIST 100 Index return values were included in both models helped in understanding if earnings of the clubs are affected by the current market and by the returns of their competitors. As a result, it was found out that returns of these indices affect stock shares of the clubs positively and that share values of football clubs run parallel to the market.

Overall, Findikçi and Tapşin's use of panel data analysis in this case study allows them to create a more robust model that considers both cross-sectional and longitudinal data. The resulting models provide insight into the relationship between football club performance and share earnings, which can be useful for investors and football club managers alike.

6. METHODOLOGY

6.1. DATA COLLECTION

I used a number of online resources and well-known databases particularly in the football industry. While the process was quite time-consuming and rigorous, my findings were very interesting to me. I found particular resources to be helpful for the different aspects of our hypothesis, such as Futhead.com for FIFA gaming statistics and sortitoutsi.net for Football Manager data. Meanwhile, the process of data collection for the social media aspect was very simple, using Instagram, Facebook, Twitter, and YouTube, as a resource for social media engagement and following data.

Some of the most important resources for the football performance data were Transfermarkt and Football Benchmark, two highly regarded sources of football statistics. This was perhaps the most interesting part of the data collection process, as I discovered the performance statistics such as "average goals per game" that led to a team's final position in each season within the five year period that I was analysing.

6.2. *METHOD OF PROCESSING*

I employed a two-stage data processing approach, utilizing both Microsoft Excel and Gretl software. The Excel stage involved importing data from multiple sources, which were then consolidated into master sheets, for each variable separately and together. I used conditional formatting to identify and highlight the relegated teams for each season in both leagues, and Excel was also used for creating some tables and graphs.

One of the challenges I faced was dealing with quite big differences in the social media and gaming data from the League of Ireland Premier Division and La Liga, which required me to use abbreviations. However, this was an early amplification of the quality and commercial success differences between these two leagues, with some La Liga clubs amassing social media followings in the hundreds of millions while the League of Ireland Premier Division did not have one club with over one million followers on any social media platform. Additionally, some gaming data was unable to be accurately discovered, which led to the assumption of equality between data points which created linear patterns in the scatter graph. Excel was

instrumental in identifying trends using charts and graphs, which helped me focus on the relevant statistics for my analysis.

Once the data was formatted in a specific way, I imported it into Gretl for analysis. I had to use the transpose function in Excel to ensure that each cell corresponded to the relevant cell in the opposite dataset as I was comparing two sets of panel data. Executing the analysis involved:

- Renaming column titles
- Removing the club names from both databases
- Assigning a number to each column
- Reshaping the data from wide-form to long-form
- Merging the two league databases for comparison purposes

from there, I conducted various statistical tests, which are described below.

6.3. *METHOD OF ANALYSIS*

6.3.1. Linear Regression

6.*3.1.1. Scatter Graph*

I began by using a scatter graph to investigate possible correlation between the two variables of Football Manager rating (FMRating) and the number points that each team has collated in each season of the time period specified (Points). Football Manager is a computer game with a strong reputation for its realism in terms of statistics and depth, giving every club a rating out of one hundred which should represent their past and future performance. I separated both leagues in this analysis with the aim of making it clearer to observe a correlation.

The results were very promising and seemed to indicate a positive correlation. Notice that there are more data points in *Figure 2* due to the increased number of clubs in La Liga compared to the League of Ireland Premier Division. The best-fit line clearly indicates that the higher a club is rated on Football Manager, the more points they obtain and thus limiting the possibility of relegation.

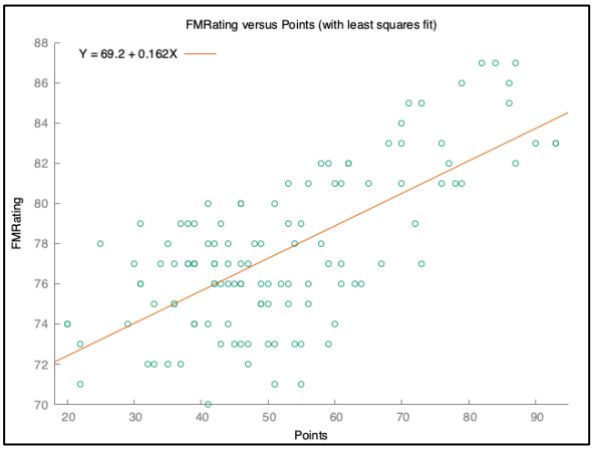


Figure 4. Output of the gaming Scatter Graph for La Liga

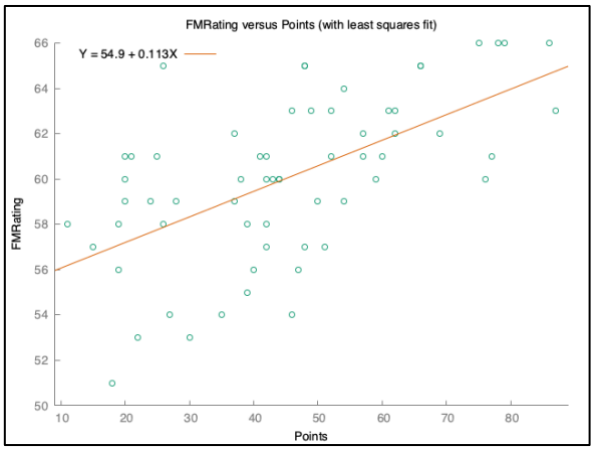


Figure 5. Output of the gaming Scatter Graph for League of Ireland Premier Division

Consequently, I used a scatter graph again to investigate the social media element of our hypothesis. The two variables I chose were Instagram followers (Instagram) and average points per game. Similar to the points used in the previous scatter graphs above, average points per game is a strong indication of a team's performance and subsequent threat of relegation. Once again, the leagues are separated in this analysis.

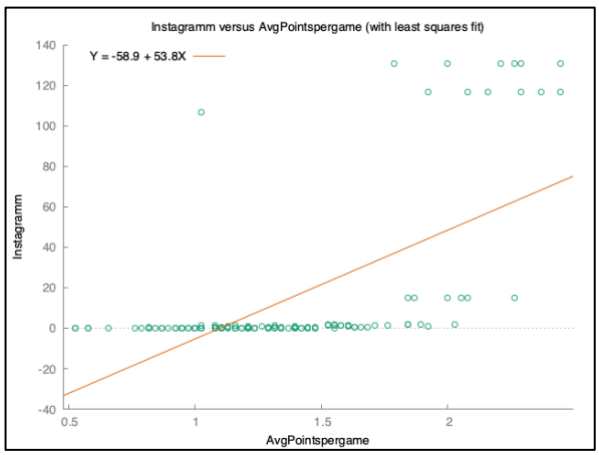


Figure 6. Output of the social media Scatter Graph for La Liga

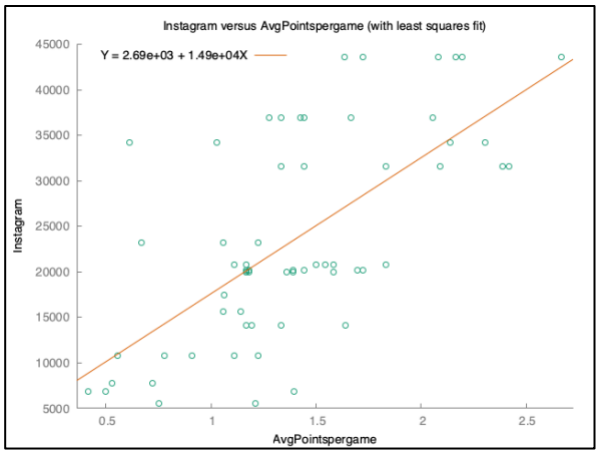


Figure 7. Output of the social media Scatter Graph for League of Ireland Premier Division

As mentioned previously, the use of abbreviations in the data collection phase was key to overcoming the challenge of the massive difference in commercial success and social media engagement between the two leagues, putting the number of followers for La Liga clubs in millions (Instagram(m)). This adjustment can clearly be seen to have affected the correlation of the output, with most La Liga clubs having under one million Instagram followers while in contrast, a couple of the Spanish clubs have over one hundred million followers. The League of Ireland Premier Division graph is definitely an easier read, with all clubs in the league having between five thousand and forty-five thousand Instagram followers.

6.3.1.2. Estimated Density Plot

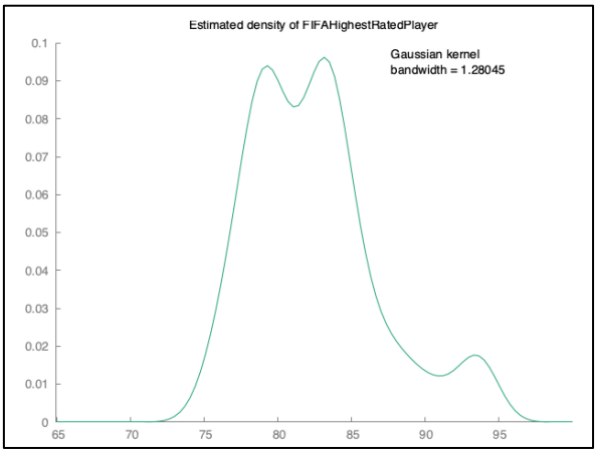


Figure 8. Estimated Density Plot for La Liga

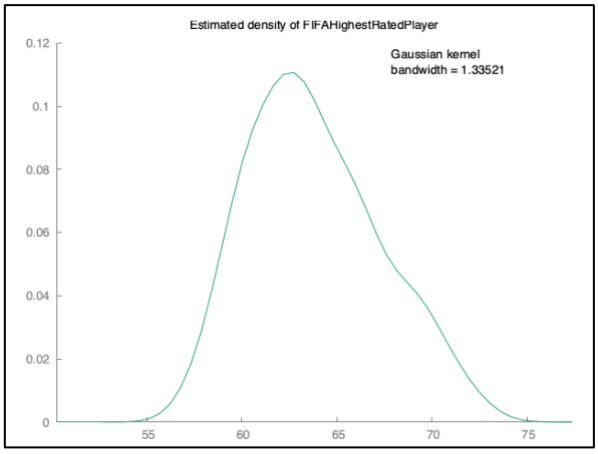


Figure 9. Estimated Density Plot for League of Ireland Premier Division

Continuing with the separation of variables, I used an estimated density plot to investigate a possible correlation between the highest rated player of each team on the FIFA video game through an approximation technique that depends on Gaussian Kernels. An interesting observation is the maximum value tail of the League of Ireland Premier Division graph (75) is very close to the minimum value tail of La Liga (72). Highlighting the quality difference between the two leagues even further, La Liga's FIFA player ratings are far more evenly distributed than those of the League of Ireland Premier Division, allowing for increased competition.

6.3.1.3. Correlation Matrix

I used a correlation matrix to investigate the possible correlations between multiple variables including points, Football Manager rating (FMRating), highest rated FIFA player (FIFAHighestRatedPlayer), Facebook followers (Facebook), and number of goals scored (Goalsscored). I found this to be an extremely effective way of measuring the correlation in each league due to the visual element. In La Liga, all five variables had a very strong correlation of at least 0.7, while in the League of Ireland Premier Division, there was some weaker correlation between the goals scored by a team and their Facebook following and Football Manager rating. It came as no surprise however that both leagues showed the strongest possible correlation (1.0) between the number of points earned by a team and the other variables. This shows that the ratings of Football Manager and FIFA, and the Facebook engagement of a team

are statistically correlated to the amount points they obtain, ultimately leading to their final league position, and possibly relegation from the top division in each country.

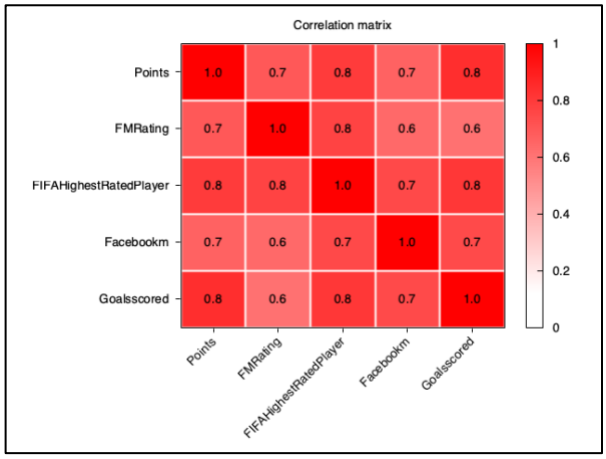


Figure 10. Output of the Correlation Matrix for La Liga

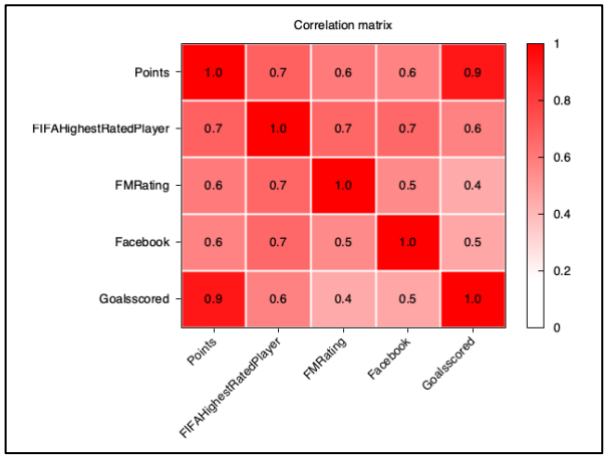


Figure 11. Output of the Correlation Matrix for League of Ireland Premier Division

6.3.1.4. Multiple Linear Regression

I assigned a team's final league position (Position) as the dependent variable and multiple explanatory independent variables as regressors which can be seen in *Figure 10*. Multivariate regressions has the great advantage that the coefficients of the explanatory variables can be interpreted as net or ceteris paribus effects. For example, the coefficient of the average age variable can be seen as the effect of relegation on the position dependent variable with all other explanatory variables fixed. This is only true for variables which are included in the regression. If you omit variables that are important and correlated with the variables that are actually included in the regression, the effect of the omitted variable will be reflected in the coefficients.

```
Model 1: OLS, using observations 1-182 (n = 181)
Missing or incomplete observations dropped: 1
Dependent variable: Position

                    coefficient  std. error  t-ratio  p-value
  ---------------------------------------------------------------
  const              25.5730     9.50715      2.690   0.0079   ***
  AvgAge              0.531886   0.285640     1.862   0.0643   *
  FIFAHighestRated~  -0.316395   0.136404    -2.320   0.0216   **
  FIFANoGoldPlayer~   0.132820   0.0888117    1.496   0.1367
  FIFANoSilverPlay~   0.425621   0.0758988    5.608   8.27e-08 ***
  FIFANoBronzePlay~  -0.149506   0.101384    -1.475   0.1422
  FMBalancem          0.00792351 0.0181809    0.4358  0.6635
  FMRating           -0.182973   0.114390    -1.600   0.1116
  FMPotential         0.0433457  0.0501871    0.8637  0.3890
  Twitterm            6.57385    2.46234      2.670   0.0083   ***
  Facebookm          -2.51247    0.902929    -2.783   0.0060   ***
  Instagramm          0.0463268  0.0351565    1.318   0.1894
  Youtubem           -3.56315    1.38091     -2.580   0.0107   **

Mean dependent var   8.817680   S.D. dependent var   5.493928
Sum squared resid    2346.045   S.E. of regression   3.736917
R-squared            0.568185   Adjusted R-squared   0.537341
F(12, 168)           18.42128   P-value(F)           5.96e-25
Log-likelihood      -488.6879   Akaike criterion     1003.376
Schwarz criterion    1044.956   Hannan-Quinn         1020.233

Excluding the constant, p-value was highest for variable 26 (FMBalancem)

Test for omission of variables -
  Null hypothesis: parameters are zero for the variables
    FIFANoSilverPlayerRating
    FIFANoBronzePlayerRating
  Test statistic: F(2, 168) = 27.7678
  with p-value = P(F(2, 168) > 27.7678) = 3.80947e-11
```

Figure 12. Output of the first Multiple Linear Regression Model

R-squared is quite reasonable (0.57) signifying a relatively strong relationship between the dependent and the independent variables. The most interesting observation is the p-value for Football Manager Balance (FMBalancem). The highest of the p-values (66.35%) alludes to the significance of this gaming variable on a club's final position. However, we must check for uncorrelatedness, the hypotheses that the model makes about the data, etc. to ensure a limited heterogeneity in the data.

The R-squared and Adjusted R-squared statistics are an insightful way to begin goodness of fit tests, with R-squared explaining the variance used by the model beyond the mean of the depend variable (Position). This means that the model has managed to explain 56% (r = 0.56) of the variability of the final position of a club in function to the variables subsequently entered. The Adjusted R-squared is always going to have a value less than or equal to that of the R-squared value, with this model it's 3% less at 53%, meaning that a small number of the variables that I've entered are not significant.

In a hypothesis test, we use a p-value and F-value to determine the significance of the statistical model. In this case, we obtained a very high F-value with a p-value of 5.96, which means we cannot reject the null hypothesis and the p-value is lower than the significance level to conclude that the model is significant.

6.3.1.5. Model as a Formula

Figure 11 displays the variables added together in order to calculate the final position of clubs, with the beta coefficients in the brackets and the paratheses showing the standard deviations of the coefficients.

```
^Position = 25.6 + 0.532*AvgAge - 0.316*FIFAHighestRatedPlayer + 0.133*FIFANoGoldPlayerRatings + 0.426*FIFANoSilverPlayerRating
          (9.51) (0.286)        (0.136)                       (0.0888)                       (0.0759)

   - 0.150*FIFANoBronzePlayerRating + 0.00792*FMBalancem - 0.183*FMRating + 0.0433*FMPotential + 6.57*Twitterm
    (0.101)                         (0.0182)             (0.114)          (0.0502)             (2.46)

   - 2.51*Facebookm + 0.0463*Instagramm - 3.56*Youtubem
    (0.903)          (0.0352)            (1.38)

n = 181, R-squared = 0.568
(standard errors in parentheses)
```

Figure 13. The Multiple Linear Regression Model converted into a formula

6.3.1.6. Chow Test

I also conducted a chow test to validate the hypothesis and examine whether there is a structural break in the analysis.

```
Chow test for structural break at observation 91 -
  Null hypothesis: no structural break
  Test statistic: F(13, 155) = 4.70173
  with p-value = P(F(13, 155) > 4.70173) = 8.04927e-07
```

Figure 14. Chow test for the third Multiple Linear Regression Model

The null hypothesis is that there is no structural break, the desired outcome.

6.3.1.7. Forecasts

Finally, I created forecast graphs for each league within the time period that I have been examining. This shows the historical data of the team's final league positions with the forecasted values, also displaying the forecasted values with the upper and lower bounds of the 95% confidence interval. I found the forecasted values to be consistent with the historical data and always remained instead the confidence interval which are appropriate for this purpose. An interesting observation from both leagues is that the forecast line almost always predicted a lower final position for the clubs that finished last (10th or 20th) in their respective leagues.

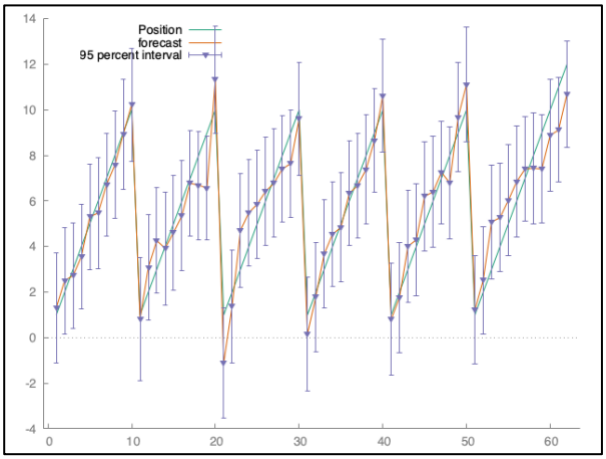


Figure 15. Forecast model for the final league position of La Liga clubs

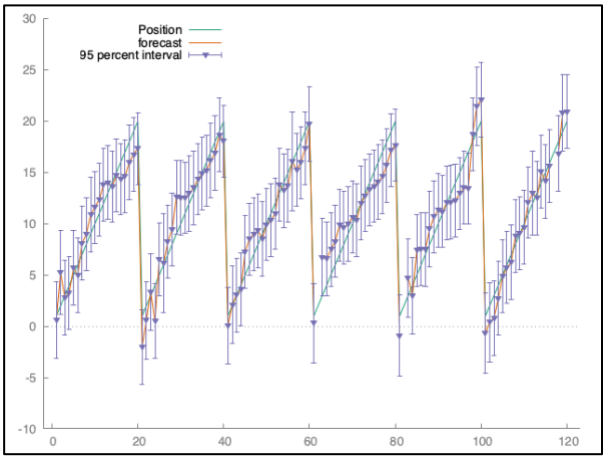


Figure 16. Forecast model for the final league position of League of Ireland Premier Division clubs

6.4 EVALUATION OF METHODOLOGY

If I could re-do this project, there are some elements of the methodology that I would approach differently.

Primarily, I would input a different dataset for the social media engagement numbers, if possible. This is with regard to the fact that the correlations between the social media data and the a club's relegation didn't paint the full picture in my opinion, with the La Liga followers being done as a percentage of one million followers, and the League of Ireland Premier Division followers as a percentage of one thousand followers. It would be more accurate to use the same metric for both variables if possible. This was a difficult constraint in the data analysis phase due to the huge disparities in numbers of followers, but in the future research could be conducted using this data and the results contrasted with the results of my study.

Also, I originally planned to conduct robustness checks on the data. Robust regression is an alternative to least squares regression when data is contaminated with outliers or influential observations, and it can also be used for the purpose of detecting influential observations. Unfortunately, I did not plan my time accordingly, so this became impossible according to the relevant deadlines I had to adhere to.

Another factor of the methodology that I was not fully satisfied with was, upon evaluation, the use of highest rated FIFA players rather than the average rated FIFA player in each team. Whilst I collect data for the top five rated FIFA players of each team, the sheer volume of players in each squad meant that I couldn't afford to find the averages. This could have provided a more accurate insight into what I was trying to discover.

Otherwise, I was quite satisfied with the efficacy of the methodology. The scatter plots were really useful in determining the initial correlation which I wanted to investigate. This correlation was then consolidated by the use of the linear regressions, where I saw the positive relationship between the variables in question. The high R-squared value indicated the statistical significance of the correlation and proved how much of the relegation of a football club is explained by FIFA, Football Manager, and the big social media platforms.

7. RESULTS

The results of the analysis indicate several significant findings in relation to the hypothesis that social media engagement and video game ratings data are correlated with a football team's performance and subsequent threat of relegation.

Firstly, the scatter graphs produced using the Football Manager rating variable and the number of points a team has collated in each season, *figures 2 and 3*, indicated a positive correlation between these two variables. The best-fit line clearly shows that the higher a club is rated on Football Manager, the more points they obtain and thus limiting the possibility of relegation. This finding suggest that Football Manager ratings may be a reliable indicator of a team's potential performance in real-life football matches.

Secondly, the scatter graphs produced using the Instagram followers variable and the average points per game variable, *figures 4 and 5*, also indicated a positive correlation between these two variables. However, the use of abbreviations in the data collection phase was key to overcoming the challenge of the massive difference in commercial success and social media engagement between the two leagues. This adjustment can clearly be seen to have affected the correlation of the output, with most La Liga clubs having under one million Instagram followers while a couple of the Spanish clubs have over one hundred million followers. The league of Ireland Premier Division graph, *figure 5*, is definitely an easier read, with all clubs in the league having between five thousand and forty-five thousand Instagram followers. The findings suggest that a team's social media engagement required to achieve success in different leagues may vary greatly.

Thirdly, the estimated density plots produced using the highest rated player of each team on the FIFA video game, *figures 6 and 7*, indicated a difference in the quality of players between the League of Ireland Premier Division and La Liga. La Liga's FIFA player ratings are far more evenly distributed that those of the League of Ireland Premier Division, allowing for increased competition. This finding suggests that the quality of player a team has access to may play a significant role in their potential performance and subsequent threat of relegation.

Fourthly, the correlation matrix, *figures 8 and 9*, produced using multiple variables including points, Football Manager rating, highest rated FIFA player, Facebook followers, and number

of goals scored indicated a high positive correlation between the Football Manager rating and the number of points a team obtains in both leagues. The matrix also indicated a high positive correlation between Facebook followers and Instagram followers, which suggests that social media engagement may be correlated across different platforms.

Then, we used a multiple linear regression model to examine the relationship between a football team's final league position and various explanatory variables, *figure 10*. The main advantage of multiple linear regression is that it allows us to interpret the coefficients of the explanatory variables as net or ceteris paribus effects. This means that we can examine the effect of a specific variable on the dependent variable, Position, while holding all other variables constant.

The R-squared value of our regression model was 0.57, which indicates a relatively strong relationship between the dependent and independent variables. This means that the model was able to explain 56% of the variability in a team's final position based on the variables included in the regression. Additionally, we found that the p-value for the variable "Football Manager Balance" was the highest among all the variables at 66.35%, suggesting its significance in predicting a club's final league position. However, we must check for uncorrelatedness and verify that hypotheses made by the model to ensure that the data is homogenous.

To test the goodness-of-fit of the model, we examined the R-squared and adjusted R-squared statistics. The R-squared value represents the proportion of variance in the dependent variable, Position, that is explained by the independent variables, while the adjusted R-squared value accounts for the number of variables in the model. Our model had an adjusted R-squared value of 53%, indicating that a small number of the variables we included in the regression were not significant.

We also conducted a hypothesis test using the p-value and F-value to determine the significance of the model. Our F-value was very high with a p-value of 5.96, indicating that we cannot reject the null hypothesis that the model is significant. Then, we presented a model formula *(figure 11)* that shows how the variables were added together to calculate a club's final position, along with their corresponding beta coefficients and standard deviations.

In an attempt to improve the accuracy of our regression model, I first omitted two variables: the number of silver player rating on FIFA and the number of Bronze layer ratings on FIFA.

This generated a new model, *figure 12*, which I compared to the original model. I found that the R-squared statistic declined to 42%, confirming that this second model was actually less accurate than the original model *(figure 10).*

I then decided to add variables instead of omitting them to improve the accuracy of the original model. We added three variables: goals scored per game, goals conceded per game, and points per game, which we knew had a real impact on a team's final position in their respective leagues. The results were significant, with the R-squared value reaching an impressive 88.3%, and the adjusted R-squared value rising from 53.7% to 87.3%. while the p-value had slightly risen to 7.65, indicating that the null hypothesis was true, we were satisfied with this model, *figure 13*, due to its overall significance.

To further validate our hypothesis and examine whether there was a structural break in our analysis, we conducted a chow test. Our null hypothesis was that there was no structural break in our analysis, which was the desired outcome.

Finally, we created forecast graphs, *figures 15 and 16*, for each league within the time period we examined. These graphs showed the historical data of a team's final league positions along with the forecasted positions.

In conclusion, the findings suggest that social media engagement, video game ratings, and player quality may all play a significant role in a football team's performance and subsequent threat of relegation. These findings may have important implications for football clubs and their marketing strategies, as well as for those involved in predicting the outcome of football matches. Further research in this area may be warranted to explore these relationships in greater depth and across different leagues and time periods.

8. ANALYSIS AND DISCUSSION

8.1. SYNERGY BETWEEN LEAGUES

In the data collection phase, I noticed the clear disparities between the leagues specifically in terms of the social media and gaming data. The numbers of the La Liga clubs followers on Facebook, Instagram, YouTube, and Twitter dwarfed the numbers of the League of Ireland Premier Division clubs. Even with La Liga a huge difference could be seen between their top two clubs – Real Madrid and Barcelona – and the rest of the league, with their social media followings being in the hundreds of millions. While this led to some outliers in the data analysis phase e.g., *figure 4,* the League of Ireland Premier Division clubs were all within thousands of followers of each other, making the data easier to interpret. Although these disparities continued in the gaming data, with La Liga having higher FIFA and Football Manager ratings due to the higher quality of player that they possess, the analysis phase of the research did bring about a surprising synergy between the leagues.

Despite the gap in quality, commercial success, social media following, or gaming rating data, the effect that these variables had on a team's final points tally, or final league position, remained quite constant between the leagues. A significant finding was that there is a positive correlation between a football team's rating on Football Manager and the number of points they accumulate each season, which can limit the possibility of relegation. This suggests that Football Manager ratings may be a reliable indicator of a team's performance in real-life football matches. Also, another finding was that there is a positive correlation between a football team's Instagram followers and their average points per game. However, this correlation was affected by the difference in commercial success, internationalisation, and social media engagement between the League of Ireland Premier Division and La Liga. This implies that a team's social media engagement required to achieve success in different leagues may vary to a certain extent.

This synergy between the two leagues is very interesting, suggesting that further research of this kind between different leagues would demonstrate similar synergy. It came as quite a shock to me as I was aware of the huge difference in quality and circumstance between the leagues, but seeing this synergy has developed my understanding of the indicators of relegation of a football club from almost any league in question. I am keen to do further research into this

topic, choosing two leagues that are closer to each other in competition and commercial success, and seeing this seeing that this could uncover.

8.2. GAMING DATA

The results from analysing the gaming data brought about an increasing respect for the video game developers and the research and data analysis that goes into creating their famous games. The correlations that I observed between a clubs rating, potential rating, budget, balance, and individual player ratings of FIFA and Football Manager and their subsequent league performance was astonishing.

A particular finding was that the quality of players a team has access to, as measured by their highest rated FIFA player, may play a significant role in their potential performance and subsequent threat of relegation. The La Liga FIFA player ratings were more evenly distributed than those of the League of Ireland Premier Division, reflecting the true dispersion of quality throughout both leagues.

Finally, the multiple linear regression model results showed that Football Manager balance was the most significant variable in predicting a club's final league position. The model was able to explain 56% of the variability in a team's final league position based on the variables included in the regression. Adding additionally variables such as goals scored per game, goals conceded per game, and points per game significantly improved the accuracy of the model, with an R-squared value of 88.3%. The data analytics team at Football Manager have proven that their in-game team budgets are correlated to those club's actual budgets, with a higher budget leading to increased investment, better quality players and facilities, and a reduced threat of relegation.

8.3. SOCIAL MEDIA DATA

While the social media data showed similar positive correlations to that of the gaming data and the relegation of a football club, I remain quite sceptical about the true impact of social media on a team's performance. As a football fan, I understand that the more popular clubs on social media have built up this following over time, and generally a time when the team has been successful. This explains this huge disparities between La Liga's top two clubs on social media

and the rest – Real Madrid and Barcelona have been far more internationalised and successful than any other club in either league. This leads me to question whether a social media following of a football club is more of a reflective statistic from their performance rather than a true indicator of their future performances themselves.

Regardless, it was observed that an increased number of followers on Facebook, Instagram, YouTube, and Twitter, had a positive correlation to a clubs safety from relegation. Also, the finding that social media engagement may be correlated across different platforms, as indicated by the high positive correlation between Facebook and Instagram followers is compelling for future research on the topic.

8.4. FUTURE INVESTIGATIONS

The results of the analysis provide strong evidence to support the hypothesis that social media engagement and video game ratings data are correlated with a football team's performance and subsequent threat of relegation.

I think that some future investigations on the topic could take some lessons from my work, in terms of direction, scope and also by taking into account some pitfalls that I have identified. The social media aspect of the research is certainly something which can be investigated in more detail to be understand as much as the gaming data relationship. As mentioned previously, the study relied on publicly available data on social media engagement and gaming statistics, which may not capture all relevant variables and factors that influence performance and relegation of football clubs. While this limitation is inherent in any data-driven study, if future researches had the capacity to work with the football clubs in question, uncovering deeper statistics and social media analytics in particular, then a more meaningful conclusion could be reached. I do think, however, that some of the ides elaborated upon in this paper can be taken advantage of to focus the thinking of future researchers on the topic, or perhaps to provide inspiration about a certain line of thinking.

Overall, these findings suggest that there is a relationship between a football team's social media engagement, video game ratings, and their potential performance and subsequent threat of relegation. This relationship may vary depending on the league, and the quality of players a team has access to may also play a significant role. Therefore, football clubs may benefit from

paying attention to their social media engagement and video game ratings to improve their overall performance and avoid relegation.

9. CONCLUSION

In conclusion, this thesis aimed to examine the potential of social media and gaming data as predictors of team performance and whether these actors have a statistical correlation with the relegation of football clubs in the League of Ireland Premier Division and Spanish La Liga between 2016 and 2022. The study investigated the strength and direction of the correlations between social media engagement, gaming ratings data, and relegation in the two leagues, and the alternative hypotheses (Ha) suggested that there were significant differences in the correlations between the two leagues. The null hypotheses (H0), on the other hand, proposed that there were no significant differences in the correlations between the two leagues. To answer the research question: yes, there is a statistical correlation between the relegation of football clubs based on social media and gaming data.

However, as I progressed through the work, I began to realise that the analysis was far more complicated than I had originally thought. I came to the conclusion that gaming data is far more accurate and strongly correlated to a team's league performance than social media data.

The results of the data analysis indicate several significant findings. Firstly, there was a positive correlation between the Football Manager rating and the number of points a team has collated in each season. Secondly, there was a positive correlation between Instagram followers and the average points per game variable. Thirdly, there was a difference in the quality of players between the League of Ireland Premier Division and La Liga. Fourthly, the correlation matrix produced using multiple variables indicated a high positive correlation between the Football Manager rating and the number of points a team obtains in both leagues. Finally, the multiple linear regression model showed a relatively strong relationship between a team's final league position and various explanatory variables.

This study offers insights into the role that social media and gaming data can play in predicting a team's performance and subsequent threat of relegation in different leagues. Further research is required to explore these findings in more depth and determine whether they hold true across different leagues and seasons. I have highlighted the importance of using innovative and unique approaches in understanding the complex dynamics of football performance and the factors that contribute to success or failure. I believe that I have dispelled the idea that gaming and social media data has no role to play in real-life sporting events and, albeit a small significance,

data from these platforms can be a significant factor in determining a prediction for the relegation of football clubs.

BIBLIOGRAPHY

Banyard, P., & Shevlin, M. (2001). Responses of football fans to relegation of their team from the English Premier League: PTS? *Irish Journal of Psychological Medicine*, *18*(2), 66–67. https://doi.org/10.1017/S0790966700006352

Barajas, A., Fernández-Jardón, C., & Crolley, L. (2005, July). *Does sports performance influence revenues and economic results in Spanish football?* [MPRA Paper]. Universidad de Vigo. https://mpra.ub.uni-muenchen.de/3234/

Chatterjee, S., & Simonoff, J. S. (2013). *Handbook of regression analysis*. Wiley.

Conejo, R., Baños-Pino, J., Canal Domínguez, J., & Rodríguez, P. (2007). The economic impact of football on the regional economy. *International Journal of Sport Management and Marketing - Int J Sport Manag Market*, *2*. https://doi.org/10.1504/IJSMM.2007.013961

Curran, C. (2022). *The League of Ireland: An Historical and Contemporary Assessment*. Taylor & Francis.

Findikçi, M., & Tapşin, G. (2015). *A PANEL DATA ANALYSIS OF THE RELATIONSHIP BETWEEN LEAGUE PERFORMANCE AND THE SHARES OF THE PUBLICLY-TRADED FOOTBALL CLUBS*.

Five years of LaLiga's international expansion with record growth. (n.d.). Global Fútbol. Retrieved March 25, 2023, from https://newsletter.laliga.es/global-futbol/five-years-of-laligas-international-expansion-with-record-growth

Football Benchmark. (2018). *Parachute Payments in the Big 5 Leagues: Compensation that creates imbalance?*

Frost, J. (2018, June 1). 7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression. *Statistics By Jim*. http://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

Gasparetto, T., & Barajas, A. (2022). Economic effects of promotion and relegation in parallel competitions. *Economics and Business Letters*, *11*(1), 7–15. https://doi.org/10.17811/ebl.11.1.2022.7-15

Growth of the LOI is encouraging but still lags behind England's FIFTH tier. (2022, May 17). The Irish Sun. https://www.thesun.ie/sport/football/8805416/growth-league-ireland-englands-fifth-tier/

Nevill, A., Atkinson, G., & Hughes, M. (2008). Twenty-five years of sport performance research in the *Journal of Sports Sciences*. *Journal of Sports Sciences*, *26*(4), 413–426. https://doi.org/10.1080/02640410701714589

Ordinary Least Squares (OLS). (2016, February 14). *Economic Theory Blog*. https://economictheoryblog.com/ordinary-least-squares-ols/

Ordinary Least Squares regression (OLS). (n.d.). XLSTAT, Your Data Analysis Solution. Retrieved March 24, 2023, from https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols

Parachute payment definition and meaning | Collins English Dictionary. (2023, March 13). https://www.collinsdictionary.com/us/dictionary/english/parachute-payment

Peel, M., Goode, M., & Moutinho, L. (1998). ESTIMATING CONSUMER SATISFACTION: OLS VERSUS ORDERED PROBABILITY MODELS. *International Journal of Commerce and Management*, *8*, 75–93. https://doi.org/10.1108/eb047369

PII: 0304-3800(93)E0074-D | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/0304-3800(93)E0074-D

Ricky, A. (n.d.). *Council Post: How Data Analysis In Sports Is Changing The Game*. Forbes. Retrieved March 25, 2023, from

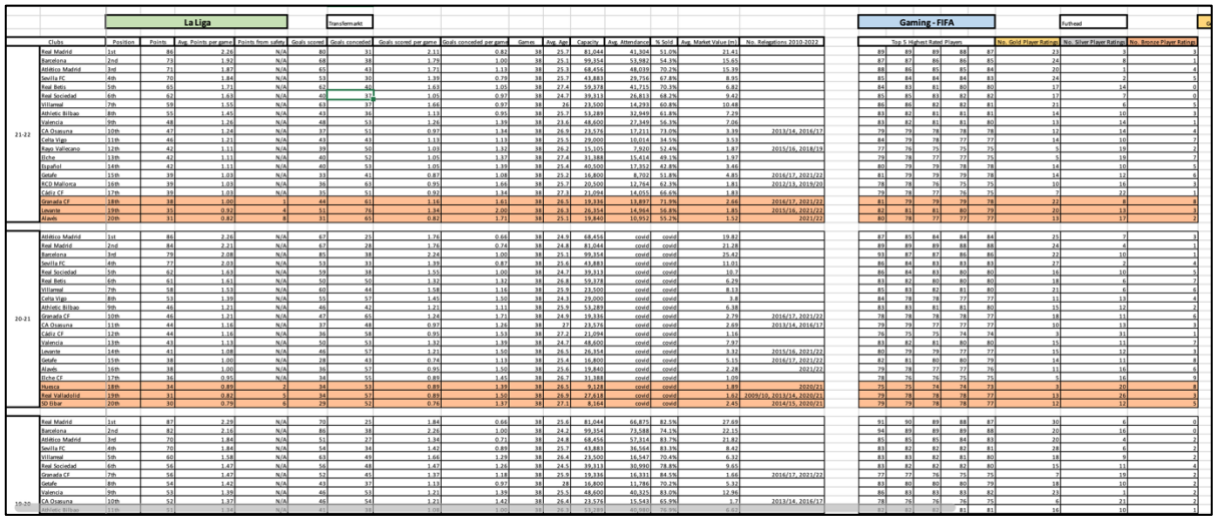
https://www.forbes.com/sites/forbestechcouncil/2019/01/31/how-data-analysis-in-sports-is-changing-the-game/

Travassos, B., Davids, K., Araujo, D., & Esteves, P. (2013). Performance analysis in team sports: Advances from an Ecological Dynamics approach. *International Journal of Performance Analysis in Sport*, *13*, 89–95. https://doi.org/10.1080/24748668.2013.11868633

Ugrinowitsch, C., Fellingham, G. W., & Ricard, M. D. (2004). Limitations of Ordinary Least Squares Models in Analyzing Repeated Measures Data: *Medicine & Science in Sports & Exercise*, 2144–2148. https://doi.org/10.1249/01.MSS.0000147580.40591.75

ANNEXES

Annex 1. A sample of the master sheet used in Excel in the early stages of my data analysis.



Annex 2.1. Key social media statistics obtained from my dataset for La Liga.

	Mean	**Median**	**Minimum**	**Maximum**
Twitter (m)	5.3768	0.52300	0.15900	46.500
Facebook (m)	12.142	0.64000	0.045000	112.30
Instagram (m)	14.527	0.38400	0.10700	131.00
Youtube (m)	1.2741	0.049000	0.0070000	14.600
	Std. Dev.	**C.V.**	**Skewness**	**Ex. kurtosis**
Twitter (m)	13.726	2.5528	2.6366	5.0110
Facebook (m)	32.170	2.6495	2.6379	5.0548
Instagram (m)	38.090	2.6221	2.5121	4.4183
Youtube (m)	3.6605	2.8731	2.9254	7.0586
	5% perc.	**95% perc.**		
Twitter (m)	0.16200	46.475		
Facebook (m)	0.12600	111.84		
Instagram (m)	0.13300	130.30		
Youtube (m)	0.0090000	14.330		

Annex 2.2. Key social media statistics obtained from my dataset for League of Ireland Premier Division.

	Mean	**Median**	**Minimum**	**Maximum**
Twitter	32782	31150	13500	51400
Facebook	35166	35000	7800.0	69000
Instagram	23411	20500	5560.0	43600
Youtube	2782.3	2270.0	90.000	5860.0
	Std. Dev.	**C.V.**	**Skewness**	**Ex. kurtosis**
Twitter	10627	0.32417	0.010425	-1.1188
Facebook	17086	0.48587	0.56750	-0.60708
Instagram	11323	0.48369	0.28375	-0.98364
Youtube	1536.8	0.55236	0.44020	-0.60092
	5% perc.	**95% perc.**		
Twitter	14835	51370		
Facebook	10645	69000		
Instagram	6860.0	43600		
Youtube	512.00	5860.0		

Annex 3.1. Key gaming statistics obtained from my dataset for La Liga.

	Mean	**Median**	**Minimum**	**Maximum**
FIFAHighestRated~	82.450	82.000	75.000	94.000
FIFA2ndHighestRa~	81.350	80.500	75.000	92.000
FIFA3rdHighestRa~	80.500	80.000	74.000	92.000
FIFA4thHighestRa~	79.992	79.000	74.000	89.000
FIFA5thHighestRa~	79.483	79.000	73.000	89.000
FIFANoGoldPlayer~	16.992	17.000	3.0000	33.000
FIFANoSilverPlay~	10.950	10.000	1.0000	31.000
FIFANoBronzePlay~	2.3083	2.0000	0.0000	9.0000
FMBalance	2.4625e+07	1.2000e+07	-2.7000e+07	1.4000e+08

FMBudget	9.0820e+06	4.0000e+06	0.0000	9.0000e+07
FMWage	6.1449e+05	51000	0.0000	1.3000e+07
FMRating	77.592	77.000	70.000	87.000
FMPotential	79.941	80.000	8.0000	91.000
	Std. Dev.	**Skewness**	**Ex. kurtosis**	
FIFAHighestRated~	4.3846	0.88927	0.49498	
FIFA2ndHighestRa~	4.0431	0.78806	0.089501	
FIFA3rdHighestRa~	3.8282	0.84785	0.38689	
FIFA4thHighestRa~	3.7293	0.76607	-0.024871	
FIFA5thHighestRa~	3.6828	0.78394	-0.023589	
FIFANoGoldPlayer~	6.5805	-0.067822	-0.39510	
FIFANoSilverPlay~	6.2104	0.65459	0.021007	
FIFANoBronzePlay~	2.2031	1.0063	0.33430	
FMBalance	3.0685e+07	2.1139	4.5208	
FMBudget	1.5237e+07	3.1296	10.598	
FMWage	1.6763e+06	5.1698	31.239	
FMRating	3.9011	0.47583	-0.32434	
FMPotential	7.7035	-6.8056	62.217	
	5% perc.	**95% perc.**		
FIFAHighestRated~	77.000	93.000		
FIFA2ndHighestRa~	76.000	90.000		
FIFA3rdHighestRa~	75.000	89.000		
FIFA4thHighestRa~	75.000	88.000		
FIFA5thHighestRa~	75.000	87.950		
FIFANoGoldPlayer~	5.0000	28.000		
FIFANoSilverPlay~	2.0000	22.000		
FIFANoBronzePlay~	0.0000	7.0000		
FMBalance	8.8615e+05	1.1335e+08		
FMBudget	0.0000	5.1600e+07		
FMWage	0.0000	3.2000e+06		

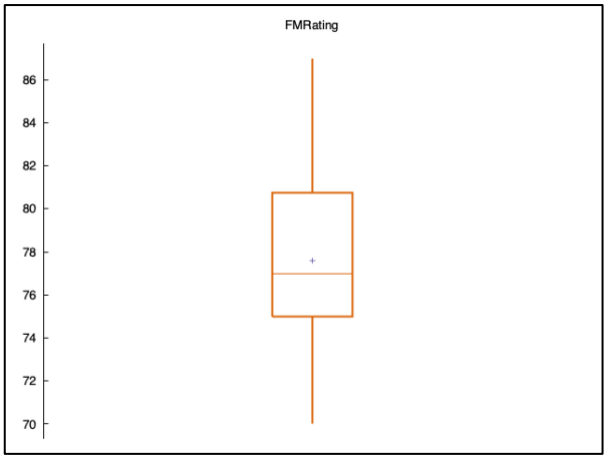
FMRating	72.000	85.000		
FMPotential	75.000	88.000		

Annex 3.2. Key gaming statistics obtained from my dataset for League of Ireland Premier Division.

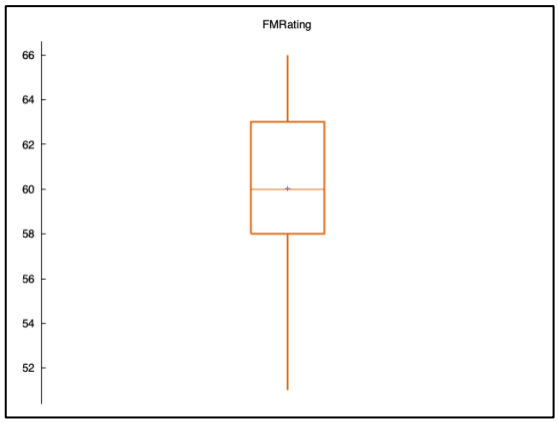
	Mean	**Median**	**Minimum**	**Maximum**
FIFAHighestRated~	63.806	63.000	58.000	72.000
FIFA2ndHighestRa~	62.355	62.000	56.000	69.000
FIFA3rdHighestRa~	61.403	61.000	56.000	67.000
FIFA4thHighestRa~	60.935	61.000	56.000	67.000
FIFA5thHighestRa~	60.500	61.000	55.000	66.000
FIFANoGoldPlayer~	0.0000	0.0000	0.0000	0.0000
FIFANoSilverPlay~	1.0161	0.0000	0.0000	7.0000
FIFANoBronzePlay~	21.548	22.000	14.000	29.000
FMBalance	2.5808e+05	89500	-15000.	3.0000e+06
FMBudget	12807	1.0000	0.0000	1.0000e+05
FMWage	3848.4	949.00	0.0000	22000
FMRating	60.032	60.000	51.000	66.000
FMPotential	63.839	64.000	56.000	69.000
	Std. Dev.	**Skewness**	**Ex. kurtosis**	
FIFAHighestRated~	3.3868	0.44618	-0.53487	
FIFA2ndHighestRa~	2.7526	0.13409	-0.49598	
FIFA3rdHighestRa~	2.6517	0.22345	-0.57882	
FIFA4thHighestRa~	2.7032	0.17930	-0.46547	
FIFA5thHighestRa~	2.5398	0.14223	-0.44825	
FIFANoGoldPlayer~	0.0000	NA	NA	
FIFANoSilverPlay~	1.7320	1.8646	2.7122	
FIFANoBronzePlay~	3.2827	-0.13253	-0.39690	
FMBalance	5.2426e+05	3.6203	13.656	

FMBudget	25864	1.7612	1.7317	
FMWage	5639.4	1.5582	1.5002	
FMRating	3.5895	-0.30521	-0.36186	
FMPotential	3.3836	-0.48713	-0.63434	
	5% perc.	**95% perc.**		
FIFAHighestRated~	59.000	70.000		
FIFA2ndHighestRa~	58.000	66.850		
FIFA3rdHighestRa~	57.150	66.850		
FIFA4thHighestRa~	56.000	65.850		
FIFA5thHighestRa~	56.000	65.000		
FIFANoGoldPlayer~	0.0000	0.0000		
FIFANoSilverPlay~	0.0000	5.8500		
FIFANoBronzePlay~	15.300	27.000		
FMBalance	300.00	1.8500e+06		
FMBudget	0.0000	69000		
FMWage	0.0000	17550		
FMRating	53.150	66.000		
FMPotential	57.000	68.000		

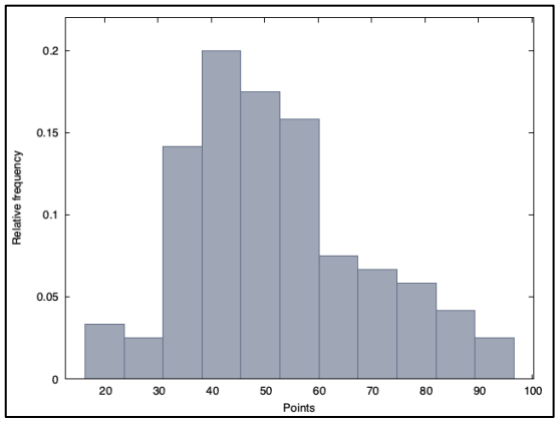
Annex 4.1. Boxplot of the Football Manager rating data for La Liga.



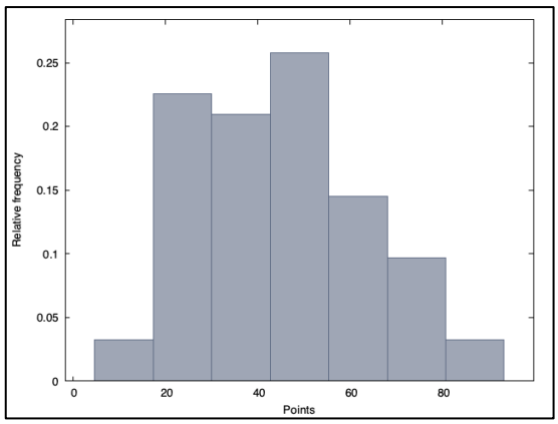
Annex 4.2. Boxplot of the Football Manager rating data for La Liga.



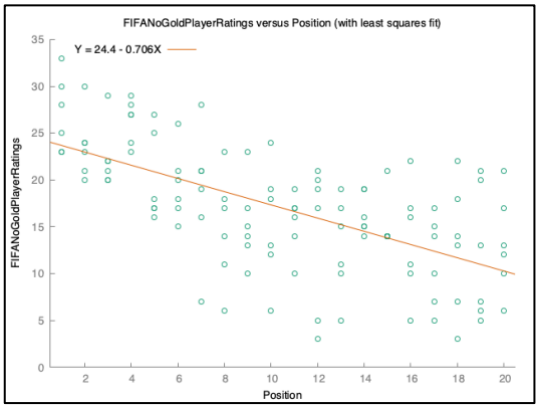
Annex 5.1. Histogram of the Points obtained by each team from 2016-2022 for La Liga.



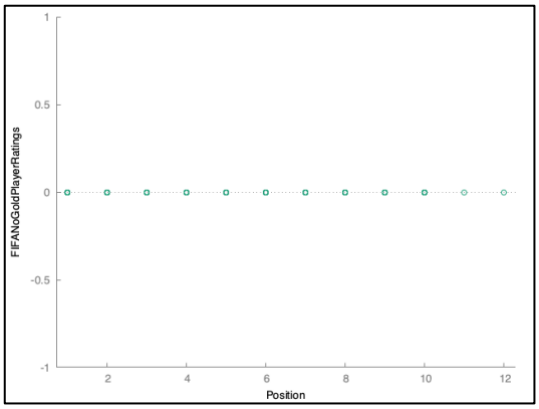
Annex 5.2. Histogram of the Points obtained by each team from 2016-2022 for La Liga.



Annex 6.1. Scatter Graph of a team's final league position vs. number of gold FIFA player rating for La Liga.



Annex 6.2. Scatter Graph of a team's final league position vs. number of gold FIFA player rating for League of Ireland Premier Division.



Annex 7. Scatter Graph for the number of silver FIFA player ratings for the League of Ireland Premier Division.