**Relationship between transfer fees and team performance among English Premier League teams**

Registration Number: 240212137

Module Details: INF6027 Introduction to Data Science (AUTUMN 2024~25)

Number of Words: 2642

# **Introduction**

The transfer fee of players is an exciting field for analysis from the perspective of team performance in football. Tomkins et al. (2010) state that they have found a significant connection between transfer fees and success. That can explain why the owners of clubs located in Saudi Arabia, Qatar, and America immediately think that spending money will lead to success (Brandon, 2024). That thought is not entirely unreasonable. Merten (2022) determines that believing in the traditional knowledge that increased transfer fees correlate to better performance is accurate.

Based on these studies, the project explores the relationship between transfer fees and team performance among English Premier League teams in the long and short term. First, the project investigates the trend between the transfer fees and the number of wins from 2012 to 2022. Second, the project analyses the correlation between the transfer fees and the number of wins in 2023. Third, a simple linear regression model was formed to predict the number of wins in the 2024 first leg. The project’s results will answer whether there is a relationship between transfer fees and team performance and whether maintaining a high transfer fee in the long term is essential to influencing future performance.

# **Methodology**

## **Gather the Data**

The project chooses to gather data using an inductive approach, using data from the Transfermarkt, a composite dataset consisting of multiple weekly updated CSV files that provide relevant attributes on competitions, matches, clubs, players, appearances, and transfers. The primary purpose of gathering data in an inductive way is to allow insights to arise naturally from the data itself. In contrast, the deductive approach used in experimental and hypothesis-testing approaches can introduce bias and hide essential insights (Thomas, 2006). Here is the list of the raw datasets.

* [Club games](https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=club_games.csv)
* [Clubs](https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=clubs.csv)
* [Competitions](https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=competitions.csv)
* [Games](https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=games.csv)
* [Transfers](https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=transfers.csv)

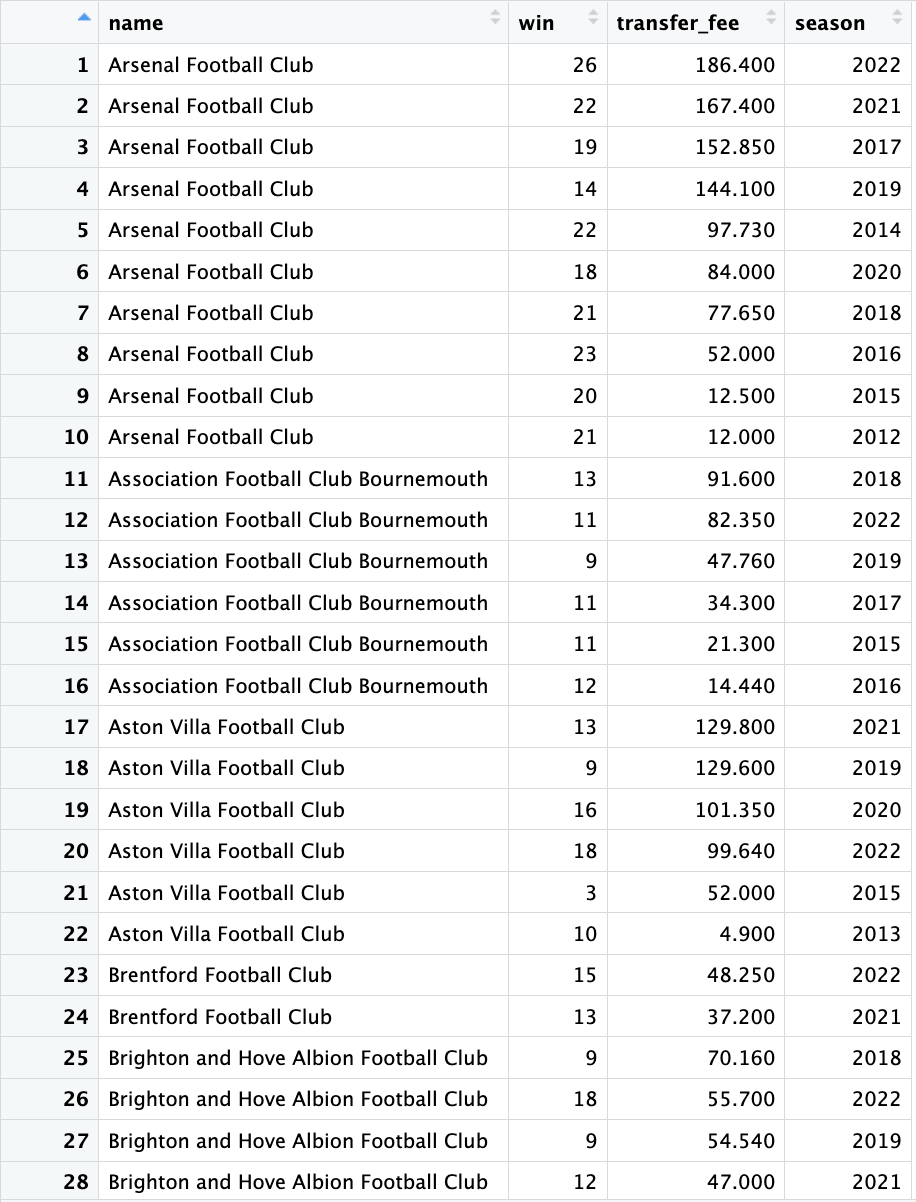
## **Pre-process and Clean the Data**

The project uses an automatically updated dataset from the Transfermarkt website with real-world data issues. Pre-processing involves stages to access and improve data quality to make it suitable for analysis. Here are the stages that the project takes to pre-process the data.

* Data integration
* Data cleaning
* Data transformation
* Data reduction

### Data Integration

The project requires combining datasets, which involves a technique known as join. Join is a way of connecting each row in one dataset to one or more rows in other datasets by only looking for matched values in a shared column known as a key. Join can be classified into two main categories: inner join and outer join. Furthermore, outer join can be subdivided into three types: left join, right join, and full join. However, the inner join reduces the unmatched row, while the outer join does not reduce the unmatched row but fills it with a null value. Therefore, this project primarily focuses on using inner join. Here are the examples of the combined datasets.



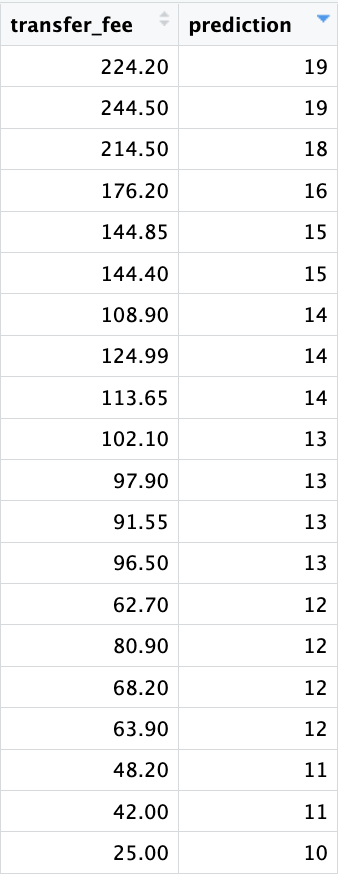
**Figure 1: Transfer Fee and Winning (2012-2022)**

Figure 1 presents 28 out of 199 rows illustrating clubs' transfer fees from 2012 to 2022 and the number of wins they achieved. This figure highlights a positive correlation between transfer fees and the number of wins from 2012 to 2022.



**Figure 2: Transfer Fee and Winning (2023)**

Figure 2 shows the clubs' transfer fees during the 2023 season, alongside the number of wins they achieved. This figure also highlights a positive correlation between transfer fees and victories.



**Figure 3: Winning Prediction (2024 First Leg)**

Figure 3 shows the clubs' summer transfer fees during the 2024 season, alongside the number of wins they might achieve at the end of the 2024 first leg. This figure highlights a positive correlation between transfer fees and the number of wins.

### Data Cleaning

The project primarily focuses on handling missing values. Missing values can significantly affect the conclusions drawn from the data. It is necessary to choose a technique to deal with missing values, either by deleting them, filling in the missing values manually or filling in the missing values automatically. However, it is necessary to identify types of missingness before choosing what type of techniques will be used. In general, there are three types of missingness according to the mechanisms of missingness (Salgado et al., 2016): Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR). However, due to the statistical advantage of MCAR, the analysis is not subject to bias. Although the design may reduce power, the missing data do not affect the estimated parameters (Kang, 2013). Therefore, the project assumes that the dataset's missing values are MCAR. In order to handle the missing values, the project chooses the case deletion method because of the valid inferences that the case deletion leads to when the missing data is an MCAR (Schafer, 1999), which leads to two techniques can be used: Listwise deletion and Pairwise deletion. Moreover, the Listwise deletion has become the default option for analysis in most statistical software when handling missing values. Additionally, it produces unbiased estimates and conservative results, especially when the missing data follow the MCAR assumption (Kang, 2013).

### Data Transformation

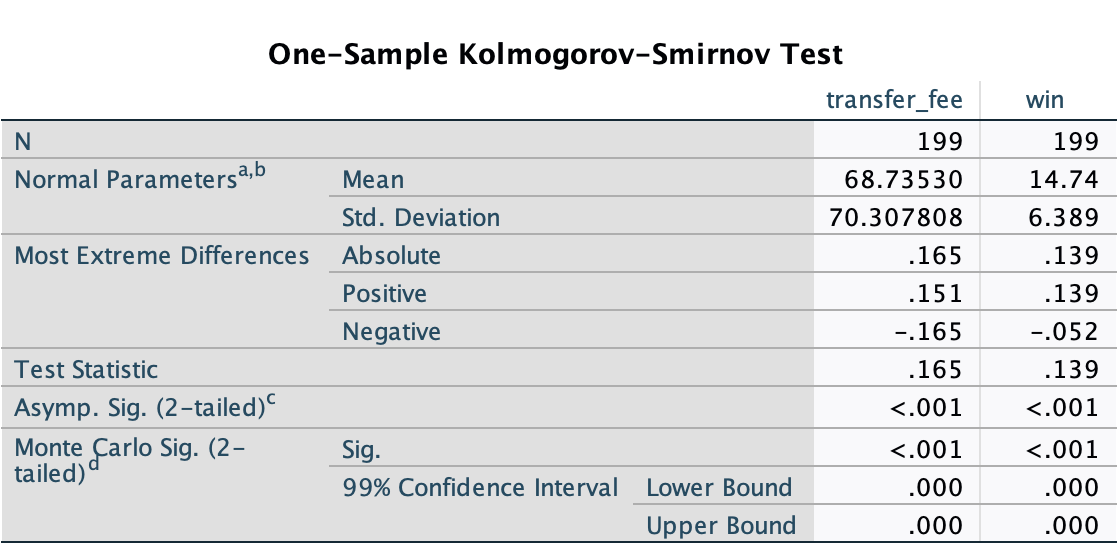
The project requires a dataset ready for analysis, which demands transformations. First, each club's total number of wins is calculated by counting the individual wins. Second, the project aggregates the total transfer fees for each club by summing these values. Thirdly, the project reduces each club's total transfer fee by dividing the amounts by one million. Moreover, the project extracts the first two characters of the transfer season column, concatenates the number 20 to the result of the substring, and finally casts the result into a numeric type.

### Data Reduction

The project reduces the redundancy by applying the technique known as where clause by filtering the records that fulfil a specified condition. For instance, the project only extracts records from the dataset where the competition is the Premier League, and matches were played in 2023 and from 2012 to 2022. Additionally, the column that contains the result of whether the clubs win or lose in a match is a binary: 1 stands for win and 0 stands for either draw or loss. Therefore, the project extracts the record where the values equal 1. Moreover, the project also extracts records of the transfers from 2012 to 2024. Furthermore, the project applies the listwise deletion technique mentioned in the Data Cleaning method by deleting the records where the transfer fee is null.

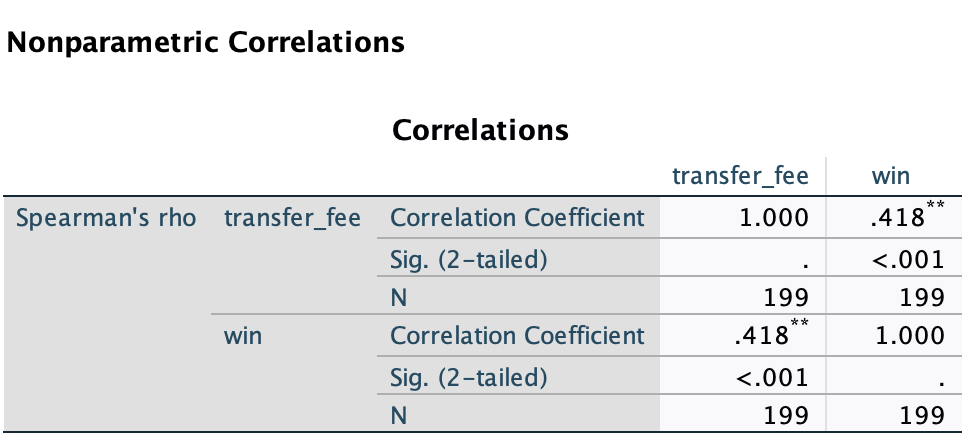
# **Results and Discussion**

## What is the correlation between a transfer fee and winning from 2012 to 2022?

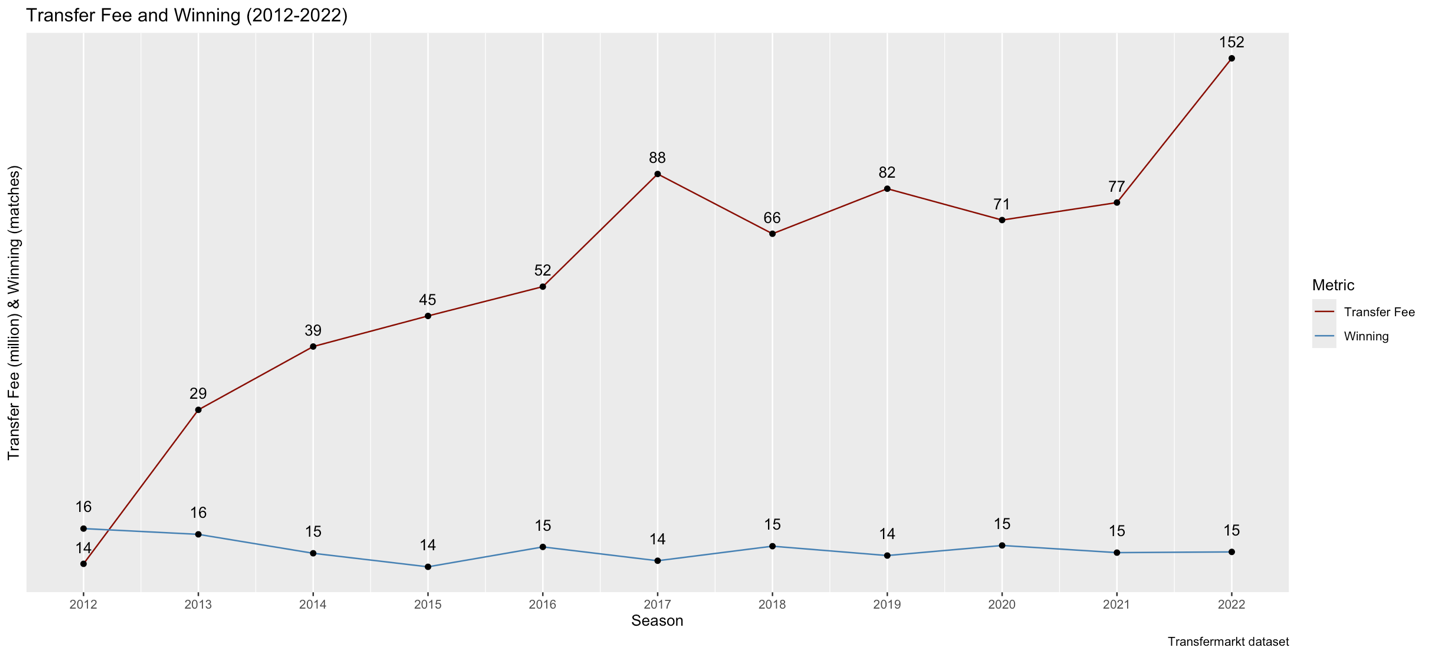


**Figure 4: One-Sample Kolmogorov-Smirnov Test for Transfer Fee and Winning (2012-2022)**

Figure 4 presents the results of a One-Sample Kolmogorov-Smirnov Test. This test was performed to decide which type of correlation test the project will use by examining the null hypothesis, which states that the data is distributed normally. The results are p<0.001 and p<0.001 for the transfer fees and number of wins, respectively, which are smaller than the specified significance level of 0.05. Therefore, the project rejects the null hypothesis and states that the data is not distributed normally. Consequently, the project conducts the Spearman Rank Correlation Coefficient Test.



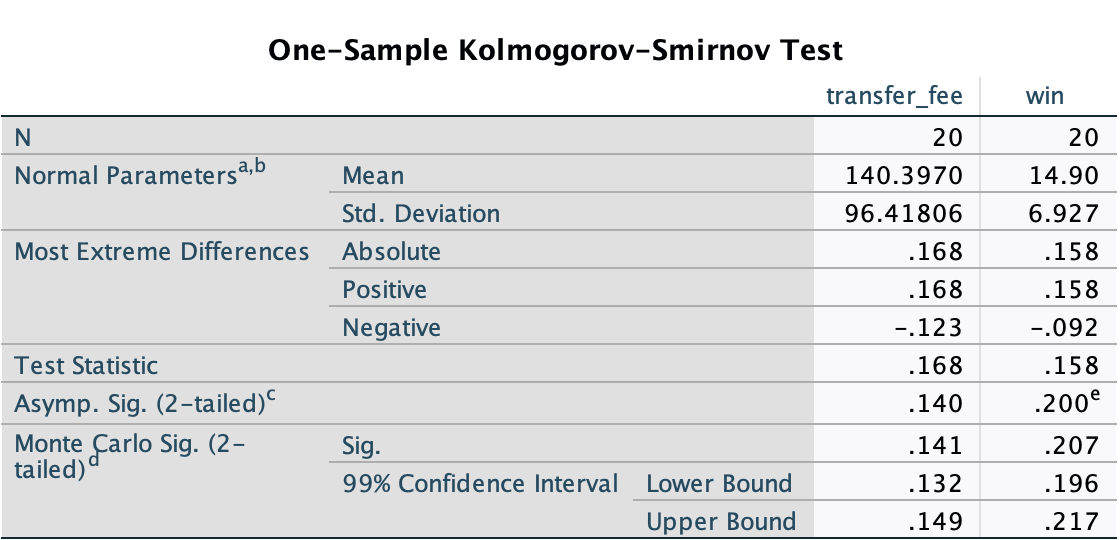
**Figure 5: Spearman Correlation Coefficient Test for Transfer Fee and Winning (2012-2022)**



**Figure 6: Transfer Fee and Winning (2012-2022)**

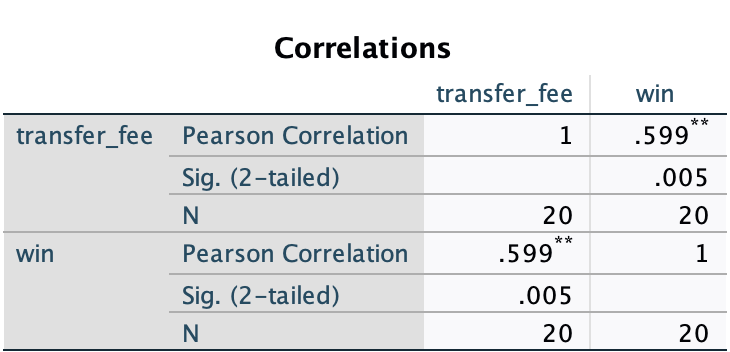
Figure 5 examines the correlation between transfer fees and winning from 2012 to 2022. The significance value is p<0.001, which is smaller than the specified significance level of 0.05. Therefore, the project rejects the null hypothesis, which indicates that there is no relationship between the transfer fees and the number of wins, and states that there is a relationship between the transfer fees and the number of wins. Moreover, the correlation coefficient is 0.418, which indicates a moderate positive relationship between transfer fees and winning during this period (Cohen J, 1992). Therefore, that result explains why not all the cases when the transfer fee increases, so does the number of wins in Figure 6. Additionally, the increase in transfer fees in 2022 is almost two times more than in 2021. However, the number of wins is the same, which indicates that high transfer fees cannot influence performance in the short term. Moreover, the transfer fee increase is not linear, which means the clubs do not maintain a high transfer fee in the long term. In conclusion, maintaining a high transfer fee in the long term is necessary to influence future performance.

## What is the correlation between a transfer fee and winning in season 2023?

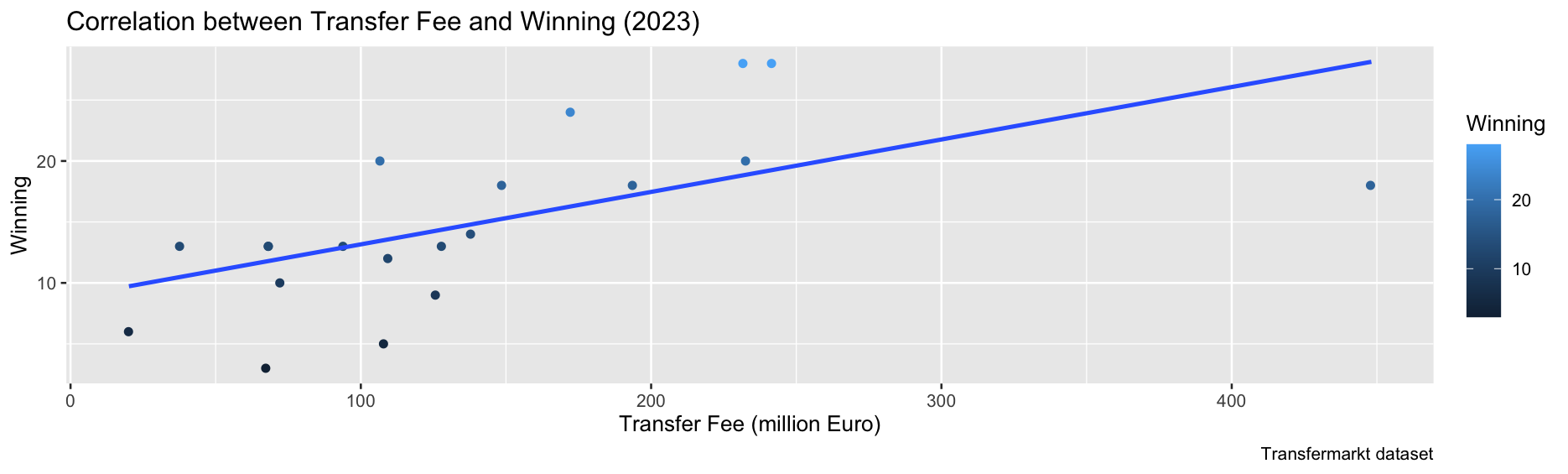


**Figure 7: One-Sample Kolmogorov-Smirnov Test for Transfer Fee and Winning (2023)**

Figure 7 presents the results of a One-Sample Kolmogorov-Smirnov Test. This test was performed to decide which type of correlation test the project will use by examining the null hypothesis, which states that the data is distributed normally. The results are p=0.14 and p=0.2 for the transfer fees and number of wins, respectively, which is larger than the specified significance level, which is 0.05. Therefore, the project accepts the null hypothesis and states that the data is distributed normally. Consequently, the project conducts the Pearson Correlation Coefficient Test.



**Figure 8: Pearson Correlation Coefficient Test for Transfer Fee and Winning (2023)**

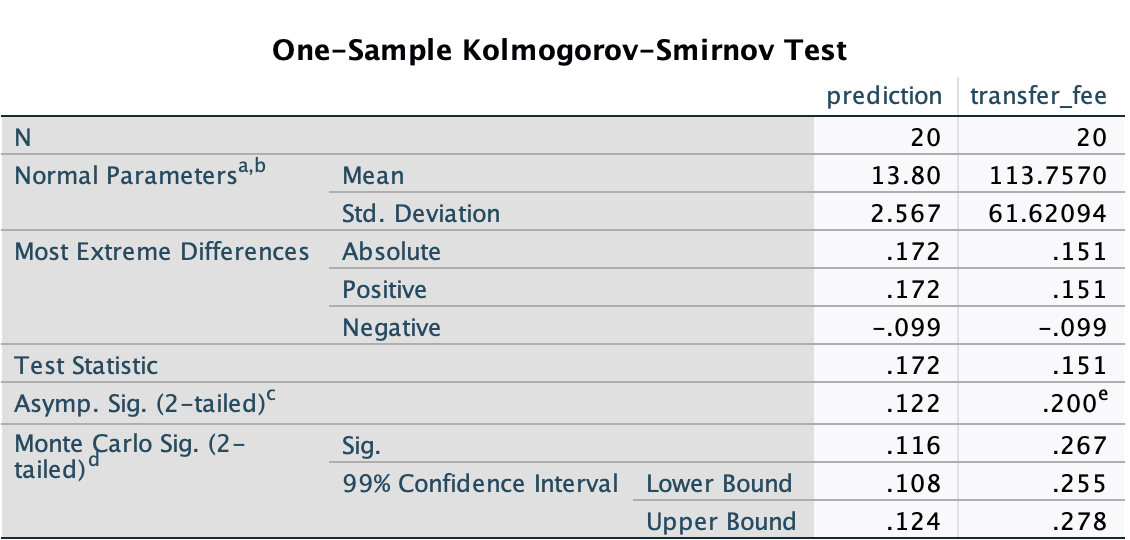


**Figure 9: Correlation between Transfer Fee and Winning (2023)**

Figure 8 examines the relationship between transfer fees and winning during the 2023 season. The significance value is p=0.005, smaller than the specified significance level of 0.05. Therefore, the project rejects the null hypothesis, which states that there is no relationship between the transfer fees and the number of wins and states that there is a relationship between the transfer fees and the number of wins. Moreover, the correlation coefficient value is 0.599, which indicates a strong positive relationship between transfer fees and winning in the 2023 season (Cohen J, 1992).That result explains why the number of wins in Figure 9 increases when the transfer fee increases. An example is two points in the top middle, which indicate Arsenal and Manchester City, respectively. However, they have maintained high transfer fees in the long term to achieve high performance. However, in most cases, some data points indicate a high value in the transfer fee but not in the number of wins. The clearest example is the point on the furthest right, which indicates Chelsea Football Club has spent almost two times more than Manchester City and still has ten wins lower. Overall, high transfer fees cannot influence performance in the short term. Therefore, they can only influence performance in the short term.

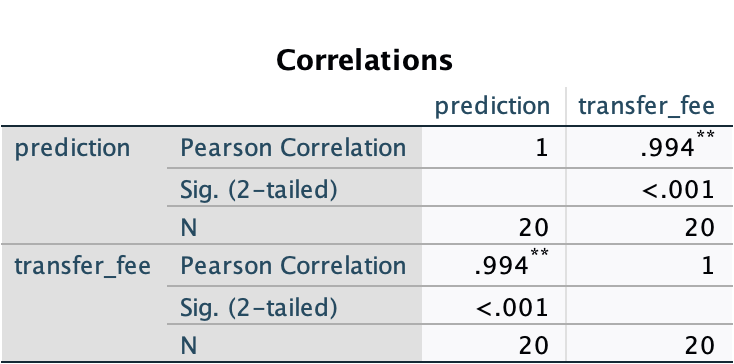
## Simple Linear Regression Model

A simple linear regression is used to predict the number of wins of the English Premier League team in the first leg of the 2024 season. This model was trained on the transfer fee and number of wins in the 2023 season, where the transfer fee is an independent variable and winning is a dependent variable. After that, the model was used to predict the number of wins in the first leg of the 2024 season using the summer transfer fee in the 2024 season as an independent variable.

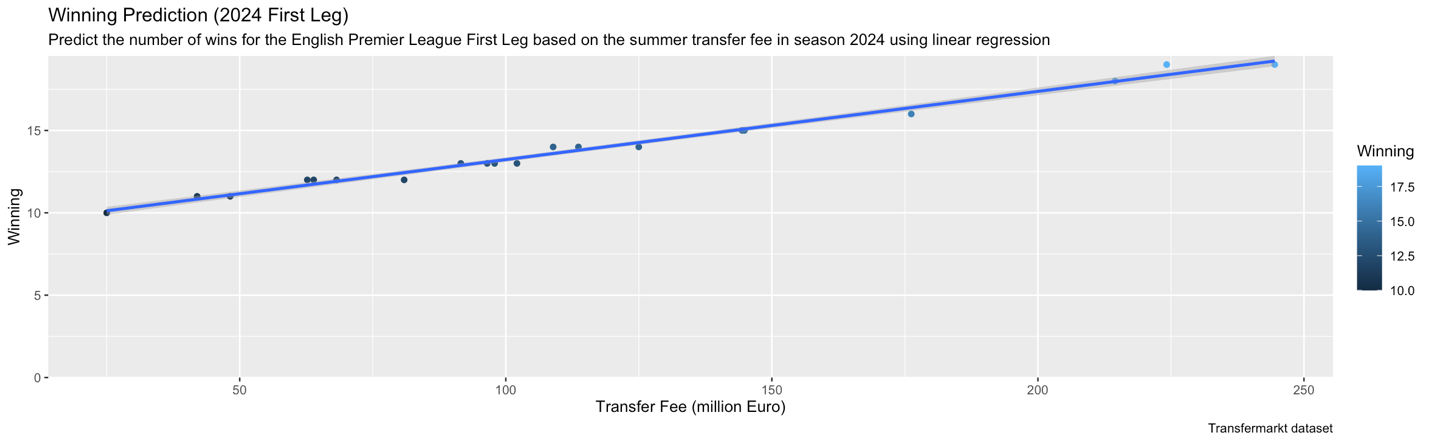


**Figure 10: One-Sample Kolmogorov-Smirnov Test for Summer Transfer Fees and 2024 Winning First Leg**

Figure 10 presents the results of a One-Sample Kolmogorov-Smirnov Test. This test was performed to decide which type of correlation test the project will use by examining the null hypothesis, which states that the data is distributed normally. The results are p=0.2 and p=0.122 for the transfer fees and number of wins, respectively, which is larger than the specified significance level, which is 0.05. Therefore, the project accepts the null hypothesis and states that the data is distributed normally. Consequently, the project conducts the Pearson Correlation Coefficient Test.



**Figure 11: Pearson Correlation Coefficient Test for Summer Transfer Fees and 2024 Winning First Leg**



**Figure 12: Winning Prediction (2024 First Leg)**

Figure 11 examines the relationship between transfer fees and winning during the 2023 season. The significance value is p<0.001, which is smaller than the specified significance level of 0.05. Therefore, the project rejects the null hypothesis, which states that there is no relationship between the transfer fees and the number of wins and that there is a relationship between the transfer fees and the number of wins. Moreover, the correlation coefficient value is 0.994, indicating a very strong positive relationship between transfer fees and winning in the 2024 first leg (Cohen J, 1992). That result explains a positive correlation between the transfer fees and the number of wins in Figure 12, which indicates that an increased transfer fee correlates to better performance. The result in Figure 12 can be explained by the previous findings, which indicate a moderate positive relationship between the transfer fees and the number of wins from 2012 to 2022. The relationship becomes strong positive in 2023, and the relationship is very strong positive in the 2024 first leg, indicating the clubs maintain a high transfer fee in the long term. Therefore, it is influencing 2024 performance.

# **Conclusion**

Through analysing the relationship between transfer fees and the number of wins, this project determines that increasing transfer fees correlates to better performance. However, the increase needs to be linear, which means that the clubs need to maintain a high transfer fee in the long term to influence future performance. Moreover, high transfer fees cannot influence performance in the short term. Therefore, they cannot be expected to influence performance in the long term.

In contrast, the project also has limitations. First, the project's scope only examines the English Premier League teams. Therefore, there might be different results if an investigation into different leagues is conducted. Second, there is another crucial aspect, which also has a relationship with how team performance that the project does not mention, which is player wages. Kuper and Szymanski (2009) state that the more a club pays its players wages, the higher it will finish. However, what clubs pay them in transfer fees makes little difference. Third, other metrics can be used to evaluate team performance, such as goal difference and finishing position. Therefore, the results differ from comparing the number of wins and the transfer fees. Additionally, the project does not use any statistical measure to evaluate how well the model performs on the new independent variable, which is the summer transfer fee in the 2024 season, when not knowing the actual value of the new dependent variable, which is the number of wins in the first leg of the 2024 season. Subsequently, the project assumes that the dataset's missing values are MCAR. However, the project does not provide any evidence to prove the assumption. Therefore, the dataset's missing values can follow another type of missingness, leading to a better method than the case deletion to handle the missing values.

# **R Code, GitHub Pages**

The source code in SQL, used to integrate and transform data, and R, used to visualise the insights for this project, can be found in the SQL and R folders. The cleansed data used to analyse can be found in the Data folder, while the report of 2614 words can be found in the Report folder. For more information, please refer to the Readme.md file. Here is the GitHub link: <https://github.com/trdeutsch>.

# **References**

* Tomkins, P., Riley, G., & Fulcher, G. (2010). Pay as you play: The true price of success in the Premier League era. GPRF Publishing
* Kuper, S., & Szymanski, S. (2009). Soccernomics: Why England loses, why Germany and Brazil win, and why the US, Japan, Australia, Turkey—and even Iraq—are destined to become the kings of the world’s most popular sport. Nation Books.
* Brandon, B. (2024). Does money really buy success? Empirical analysis of money spent effect on English Premier League team’s performances/success: A panel data analysis. Applied Economics. Bryant University. Retrieved from <https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1233&context=eeb>
* Merten, B. (2022). The impact of transfer spending in expediting improvement of on-field performance of English Premier League clubs. Senior Theses. University of South Carolina – Columbia. Retrieved from <https://scholarcommons.sc.edu/cgi/viewcontent.cgi?article=1518&context=senior_theses>
* Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing data. In C. P. Secondary (Ed.), Secondary analysis of electronic health records(pp. 143–162). Springer. <https://doi.org/10.1007/978-3-319-43742-2_13>
* Kang, H. (2013). The prevention and handling of missing data. Korean Journal of Anesthesiology*,* 64(5), 402–406. <https://doi.org/10.4097/kjae.2013.64.5.402>
* Schafer, J. L. (1999). Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1), 3–15. <https://doi.org/10.1177/096228029900800102>
* Thomas, D. R. (2006). A general inductive approach for analysing qualitative evaluation data*.* American Journal of Evaluation, 27(2), 237–246. <https://doi.org/10.1177/1098214005283748>
* Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>