Money = Wins? (Sports Edition)

Kevin Pan, Ryder Swenson

Summative Answer

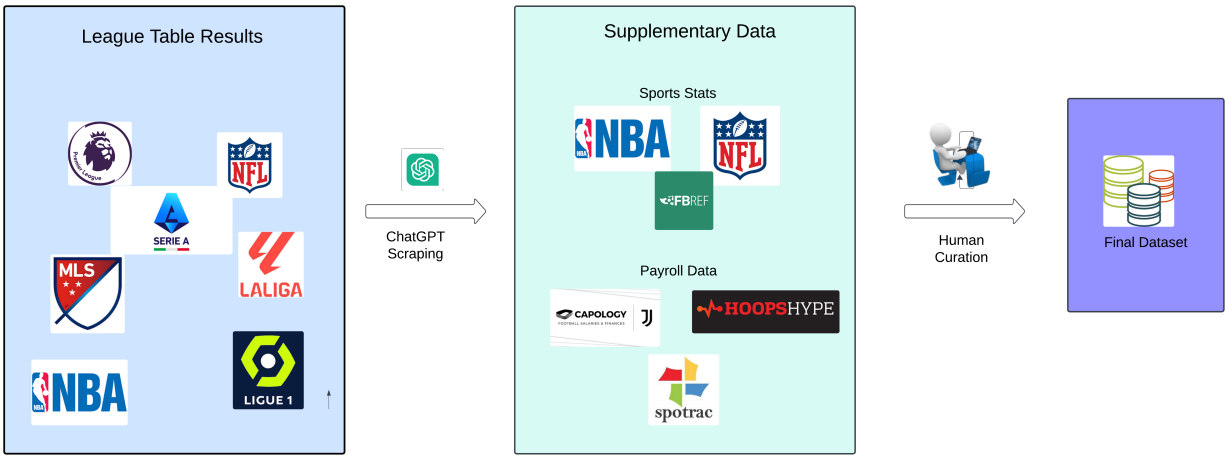
Money, in the form of standardized average annual team payroll, does typically exhibit a fair amount of impact on a team's winning percentage across soccer (football) leagues globally, but this becomes less clear when looking specifically at American Sports. Moreover, pumping in money to a sports team cannot be shown to directly *cause* an increase in winning, and there are multiple skill specific statistics that exhibit much higher correlations with winning percentage.

Introduction

As finance has become just as important as competitiveness in the global sports environment across all leagues, this influx of cash begs the question: Does money equal results? More formally, do teams who spend more money on their players perform better compared to their "less rich" competitors? To analyze this, we chose the most popular sport in the world to examine - soccer (or football) across major global leagues: MLS, Premier League, Ligue 1, Serie A, Bundesliga, Saudi Pro League, Brazilian Serie A, and La Liga. A benefit to analyzing some of the major European soccer leagues referred to as the "Big 5" (La Liga, EPL, Ligue 1, Serie A, and Bundesliga), is that there is a major annual competition where these teams compete for the most desired trophy in club soccer: the Champions League. This competition allows us to compare the competitive strength of these clubs, whereas they were previously divided nationally by league. An international comparison of soccer clubs will provide greater insight into the importance of funding in the sport as a whole. However, while the Champions League is a global competition, the main competitors tend to be denoted from major European leagues as these are the most globally competitive. Beyond the scope of soccer internationally, we were curious about another major sport in the United States specifically: basketball (using teams in the NBA). Throughout our exploration across all of the sports that we analyzed, we arrived at the determination to explore several seasons

for each league, under the assumption that the data is available. These seasons primarily took place in between 2020 and 2024.

Data Explanation



Generally, our data pipeline consists of the main steps depicted above. We input screenshots of league tables into ChatGPT to produce the table in spreadsheet format. Using various sports statistic websites and payroll data websites, we then manually append additional data onto the tables, resulting in our final datasets.

Soccer

Because soccer provides the resources for competitive international comparison, we decided to allocate most of our focus analyzing our soccer data. This increased focus led to semi-manually compiling 28 datasets across 10 workbooks. These datasets included data on many teams' standing in their own respective leagues as well as their annual total salary and average player salary data. Included in the team standing data, we pulled metrics such as matches played, goals scored, goals allowed, and goal differential. We were able to acquire the data for the leagues' leaderboards through their websites and ESPN (Bundesliga, 2024; ESPN-Brazilian-Serie-A, 2024; ESPN-La-Liga, 2024; ESPN-Ligue-1, 2024; ESPN-Saudi-Pro-League, 2024; ESPN-Serie-A, 2024; MLS, 2024; PremierLeague, 2024). As for the salary data, we used Capology (Capology, 2024), a popular data bank that focuses on the financial information of soccer players, contracts, and teams. Capology compiles verified player salary data as well as estimates based on in-house algorithms to provide figures for team salary. Beyond data regarding team standing and salary, we used FBREF (SportsReference, 2024), a massive data bank on soccer statistics (primarily league standing and team performance)

that covers over 140 leagues in over 40 countries, to append many additional metrics. Some of these metrics include possession percentage, expected goals on the season, and game by game statistics. We focused on seasons between the years 2020 and 2024 for a majority of the leagues. There are three other odd cases that do not follow our standard data compiling process. One of these cases is the Saudi Pro League, where we are only able to compile data for the 2023-2024 season. Next, there is the MLS and the Brazilian Serie A, both of which complete their competitive seasons within a single calendar year. Because of this distinction, we collected data for the MLS in 5 seasons (2020, 2021, 2022, 2023, and 2024) and for the Brazilian Serie A in 3 seasons (2021, 2022, and 2023). Otherwise, we collected data for the 2020-2021, 2021-2022, and 2022-2023 seasons for our other major leagues. Finally, we compiled data on the total spending of the Big 5 leagues using a Deloitte curated dataset on Statista (Deloitte, 2024) and their average round placement in the Champion's League across the 2016-2023 seasons.

In total, we compiled data for 9 leagues and created 10 workbooks, which can be found in the soccer section of this drive (Pan & Swenson, 2024a).

Limitations

While we were able to pull data from concrete sources on team performance and league standing, the financial statistics for each league are based partly on confirmed contract information disclosure as well as informed estimates (since Capology uses estimating algorithms). There is an unknown degree of error in these estimates, while they are pulled from reputable sources on soccer financials. Moreover, data availability, primarily in the scope of teams' finance, limited us in the years that we could pull data from for certain leagues like the Saudi Pro League.

Basketball

For our NBA datasets, we compiled every team in the league; their final league standing; their offensive, defensive, and net ratings; their total team salary of their players; and their salary sum of the top three highest paid players on the team. Calculating the sum of the top three player salaries for each NBA team sets the stage to determine whether top-heavy-spending is effective. All of the data for the team standing and their offensive, defensive, and net ratings was provided by the NBA's official website (ESPN-NBA, 2024). As for the salary data for each team, we were able to reach this data through HoopsHype (Ventures, 2024); Acquired in 2012 by USA Today, HoopsHype is a popular basketball news website frequently

cited by other major organizations such as ESPN and CBS Sports. After finding the salary data and team standing individually, we compiled the two data points in three datasets for the 2020-2021, 2021-2022, and 2022-2023 seasons all in the same workbook, found in the NBA section of this drive (Pan & Swenson, 2024a).

Limitations

The data that we wrangled for each NBA team's performance and player salaries is entirely public. The performance data is provided publicly by the NBA itself, making for complete transparency, and the salary data for each player is publicly available. The source that we used for salary data, HoopsHype, simply compiles this data. To summarize, there should be few limitations, if any, on our NBA data.

Methodology

First, we compiled our own datasets for two different sports (soccer and basketball). Due to the vast amount of data available for different soccer leagues and their teams, we decided to focus our data collection and compilation there. After compiling data on team standing and salary, we looked to creating visualizations to observe a correlation between these two variables. Our tools for processing and visualization were Google Sheets and various Python libraries: Pandas (pandas development team, 2024), Numpy (Harris et al., 2020), Sci-kit-learn (Pedregosa et al., 2011), Matplotlib (Hunter, 2007), SciPy (Virtanen et al., 2020), and Seaborn (Waskom, 2021).

To train the random forest model (and neural network in the supplementary), we used a Python Jupyter Notebook with the prior libraries. Specifically, we rely on the RandomForestRegression library and MLPRegressor library from Sci-kit-learn. All code can be found in this Github (Pan & Swenson, 2024b).

Disclaimer: We do use ChatGPT to assist with code implementation, mainly for model training.

Results

Before going into the results, we must define the correlation coefficient r and coefficient of determination R^2, with definitions found in this article (AnalysisInn, 2020).

r

Pearson's correlation coefficient r (created by the eugenicist in our course readings) measures the strength and relationship between two variables, ranging between -1 and 1. Negative correlation means that as one variable increases, the other decreases - Positive correlation means that as one variable increases, the other increases.

R^2

The coefficient of determination R^2 (also created by Pearson) measures how much of the variance in a variable can be explained by another variable, ranging between 0 and 1. A high R^2 value between x and y essentially means x matters heavily in determining y.

To identify the importance of payroll with respect to two win statistics - win percentage and season rank - we calculated the correlation coefficient r and coefficient of determination R^2 across various seasons. Since we only really care about the *difference in pay compared to other teams*, we first standardized the payroll data - Total Annual Payroll (millions) - by converting each value to its z-score (StatisticsHowTo, 2024), $z = \frac{x-\mu}{\sigma}$, with respect to the other payrolls during that season. The z-score, to clarify, represents the distance of a data point from the mean of a dataset. This allows us to compare results across leagues and across years, as total pay and inflation are standardized. We were able to make this z-score calculation for the soccer statistics but did not for basketball.

Soccer Results

For most of the following leagues (excluding Saudi Pro League), we present 3 graphs:

1. Change in r and R^2 between Standardized Average Annual Payroll (Millions) & Win %.
2. Regression line between Standardized Average Annual Payroll (Millions) & Win %.
3. Regression line between Average Possession % & Win %.

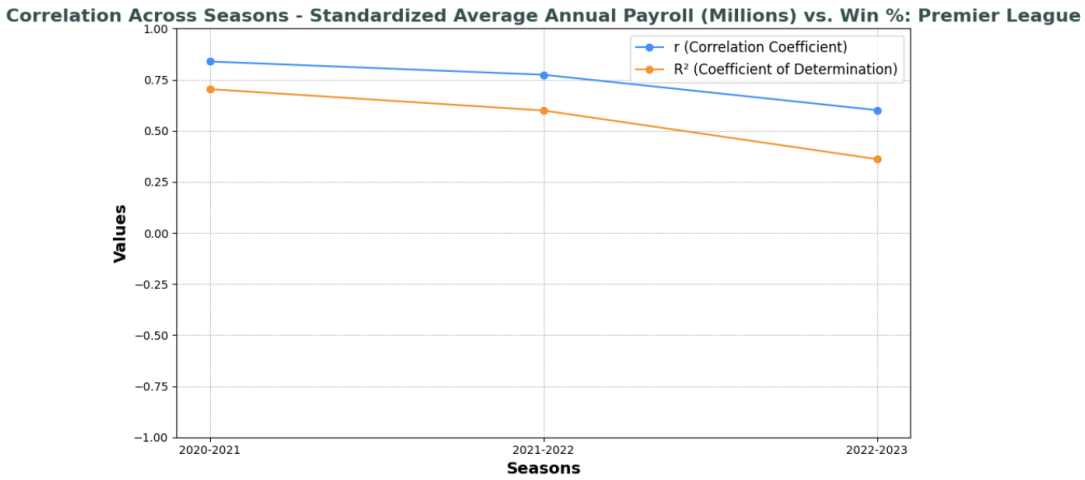
We chose the possession % statistic to provide context to the effect of payroll - *we expect possession % to have a strong correlation with win %.*

The average possession percentage per game is calculated by dividing the total number of passes a team made by the total number of passes made by both teams during said game. A

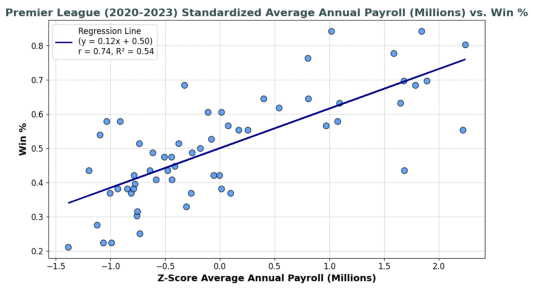
team's average possession percentage is the mean of a team's average possession percentage per game across the season. This will be compared to the Standardized Average Annual Payroll (Millions) in relation to R^2 and win percentage.

Premier League (UK)

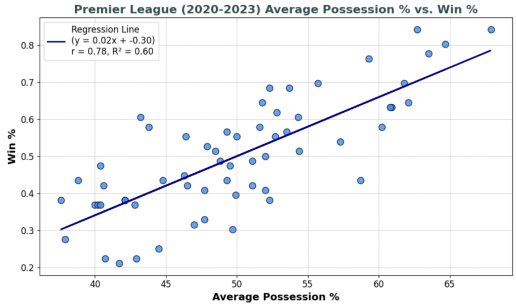
Results Data: (PremierLeague, 2024); Salary Data: (Capology, 2024)



Observing the trend of r and R^2 across seasons, we can see that r decreases across seasons but remains positive, and R^2 varies between ~ 0.36 - ~ 0.70. Payroll tends to have a positive effect on win %, but the strength of that connection varies heavily across seasons.



r: 0.74, R^2: 0.54

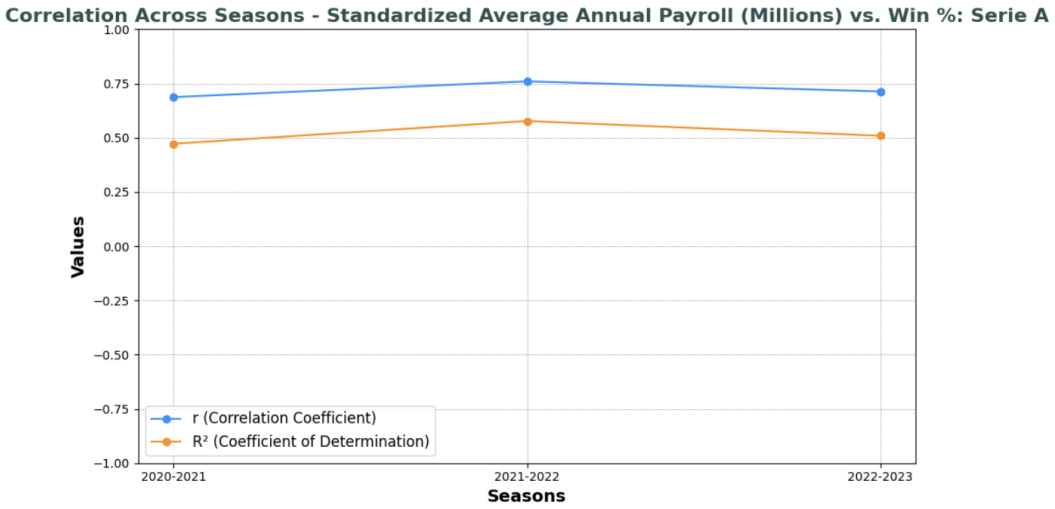


r: 0.78, R^2: 0.60

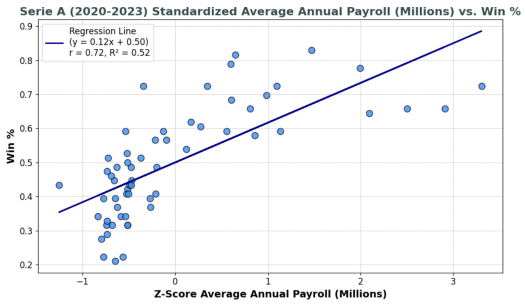
54% of the variability in win percentage can be explained by payroll, while 60% of the variability in win percentage can be explained by possession %. Ultimately, both statistics strongly influence a team's winning percentage, but possession % slightly outweighs payroll.

Serie A (Italy)

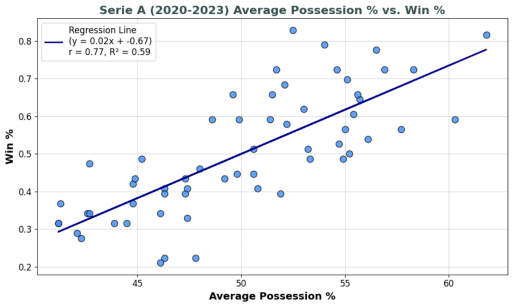
Results Data: (ESPN-Serie-A, 2024); Salary Data: (Capology, 2024)



Similar to the Premier League, the r and R^2 values demonstrate a positive correlation between average annual payroll and overall win percentage during the Serie A regular season. Both coefficients stay somewhat consistent across the three seasons.



r: 0.72, R^2: 0.52

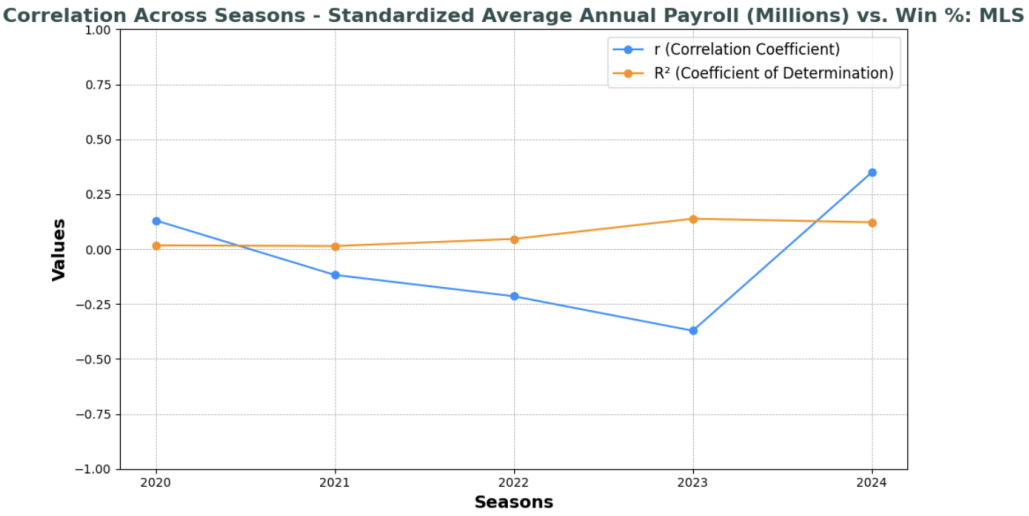


r: 0.77, R^2: 0.59

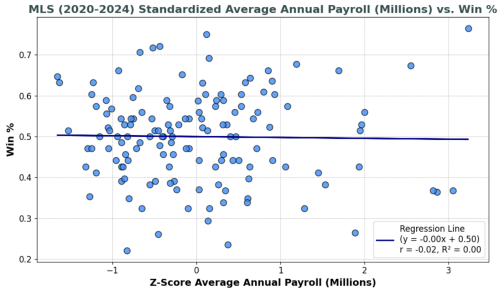
52% of the variability in win percentage can be explained by payroll, while 59% of the variability in win percentage can be explained by possession %. Possession % still slightly outweighs payroll.

MLS (USA)

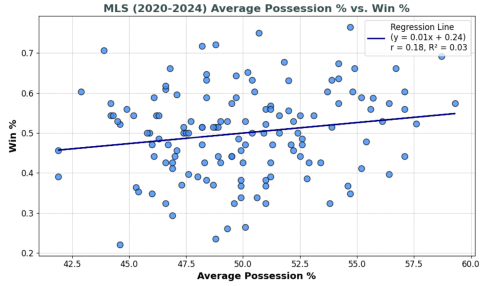
Results Data: (MLS, 2024); Salary Data: (Capology, 2024)



Unlike previously examined leagues, the MLS does not show a clear, consistent correlation between the average payroll of each team and their win percentage across several recent seasons (2020–2024). The R^2 coefficient is ~ 0.01 – ~ 0.14, showing minimal correlation. Furthermore, r is negative for some seasons. A possible cause for this lack of influence of average annual payroll is the small range of annual payrolls in the MLS compared to other leagues.



r: -0.02, R^2: 0.00

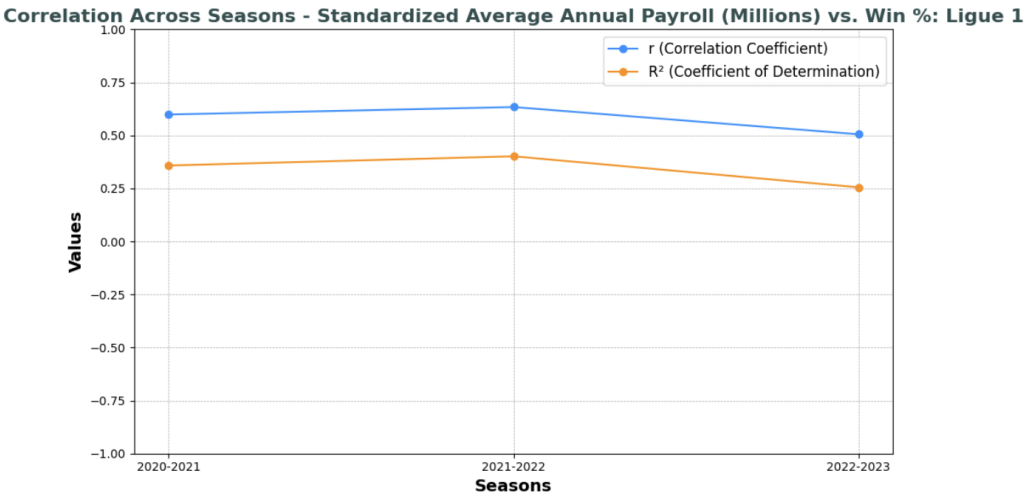


r: 0.18, R^2: 0.03

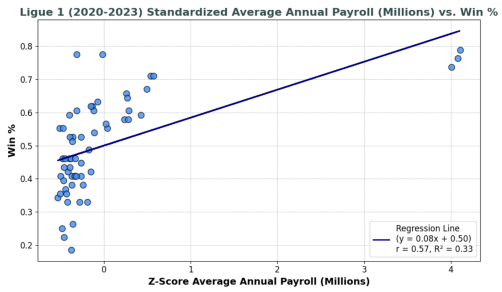
Here payroll is obsolete, while 0.03% of the variability in win percentage can be explained by possession %. It is unclear how both stats factor into determining a team's winning percentage.

Ligue 1 (France)

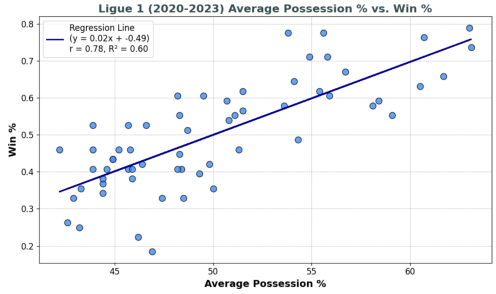
Results Data: (ESPN-Ligue-1, 2024); Salary Data: (Capology, 2024)



Observing the trend of r and R^2 across seasons, we can see that r remains positive, and R^2 varies between ~ 0.26 – ~ 0.40. The pattern that payroll has a positive effect on win % is present.



r: 0.57, R^2: 0.33

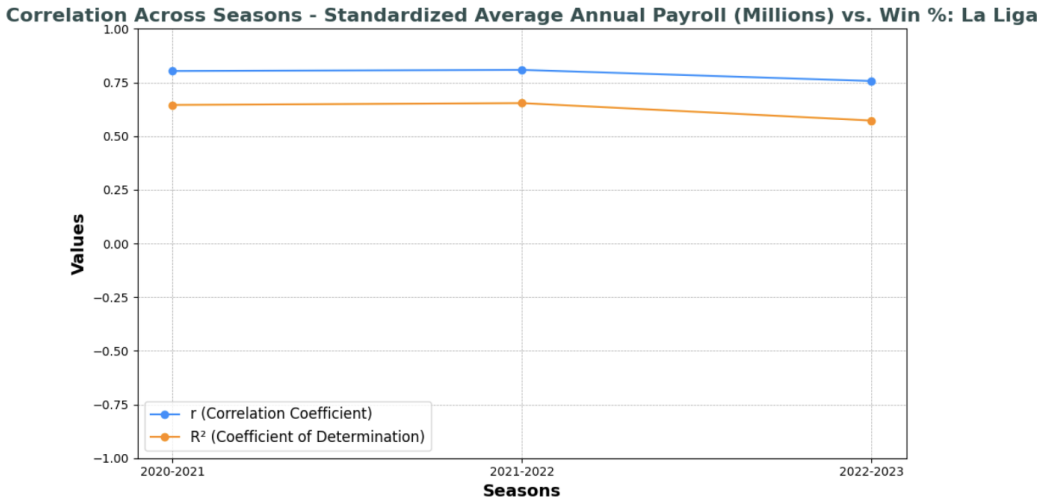


r: 0.78, R^2: 0.60

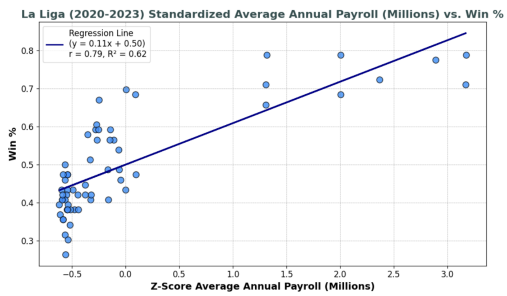
33% of the variability in win percentage can be explained by payroll, while 60% of the variability in win percentage can be explained by possession %. Again, both stats factor into a team's winning percentage, but possession % heavily outweighs payroll.

La Liga (Spain)

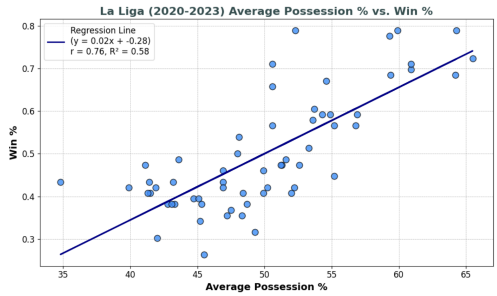
Results Data: (ESPN-La-Liga, 2024); Salary Data: (Capology, 2024)



Observing the trend of r and R^2 across seasons, we can see that r and R^2 both remain relatively consistent. R^2 varies from ~ 0.57 – ~ 0.65, demonstrating payroll's positive effect on win %.



r: 0.79, R^2: 0.62

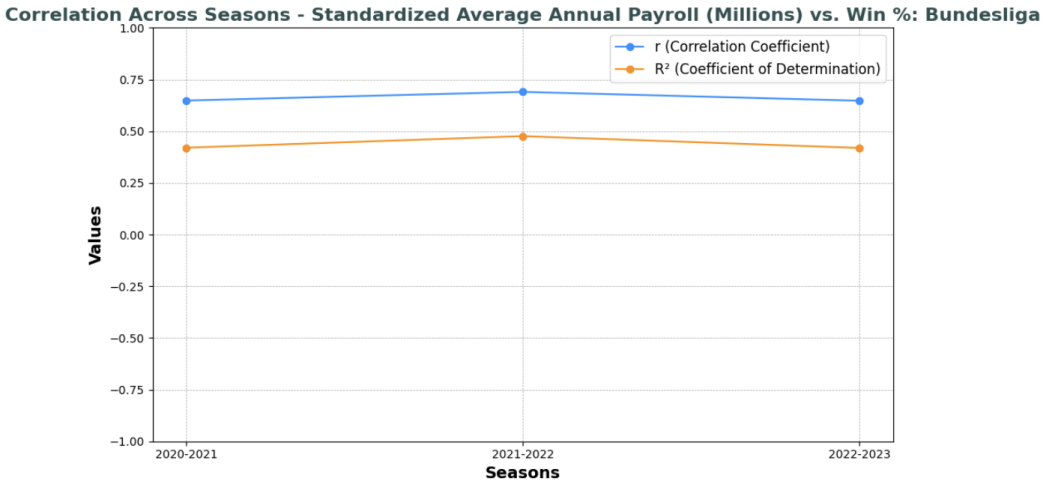


r: 0.76, R^2: 0.58

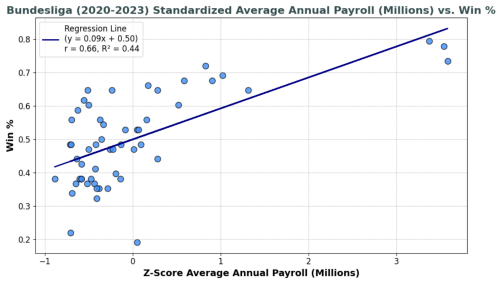
62% of the variability in win percentage can be explained by payroll, while 58% of the variability in win percentage can be explained by possession %. Both stats factor into a team's win percentage, but payroll holds more importance.

Bundesliga (Germany)

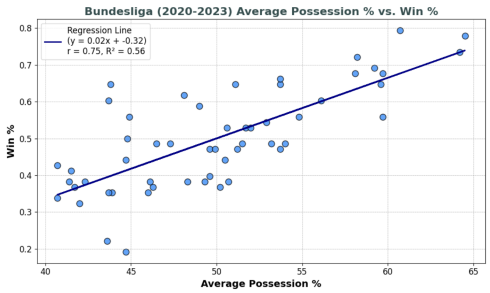
Results Data: (Bundesliga, 2024); Salary Data: (Capology, 2024)



Observing the trend of r and R^2 across seasons, we can see that r and R^2 both remain relatively consistent. R^2 varies from ~ 0.42 – ~ 0.48. Payroll proves to have a positive effect on win %, but the strength of that connection varies across seasons.



r: 0.74, R^2: 0.44

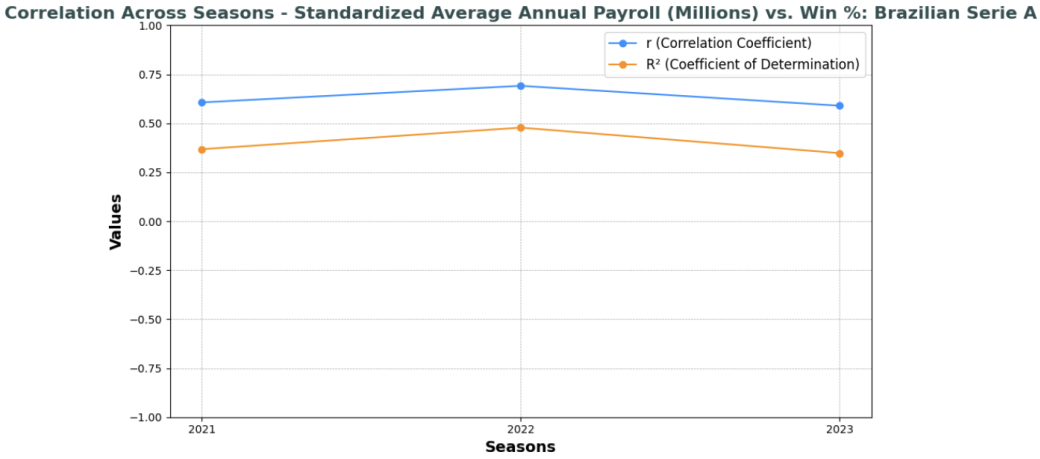


r: 0.78, R^2: 0.56

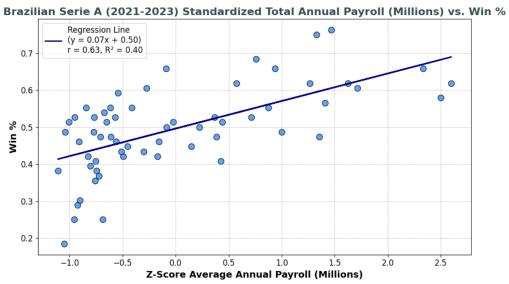
44% of the variability in win percentage can be explained by payroll, while 56% of the variability in win percentage can be explained by possession %. Once again, possession % holds higher importance.

Brazilian Serie A (Brazil)

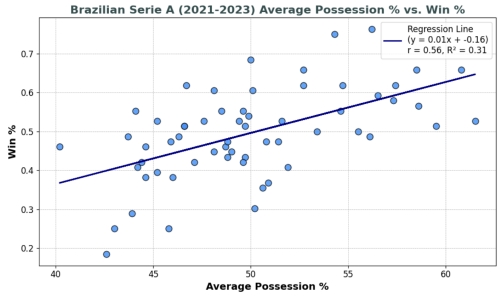
Results Data: (ESPN-Brazilian-Serie-A, 2024); Salary Data: (Capology, 2024)



Observing the trend of r and R^2 across seasons, we can see that r and R^2 both remain relatively consistent. R^2 varies from $\sim 0.35 - 0.48$, demonstrating payroll's positive effect on win percentage.



r: 0.63, R^2: 0.40

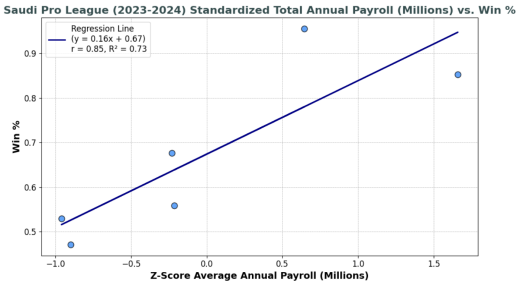


r: 0.56, R^2: 0.31

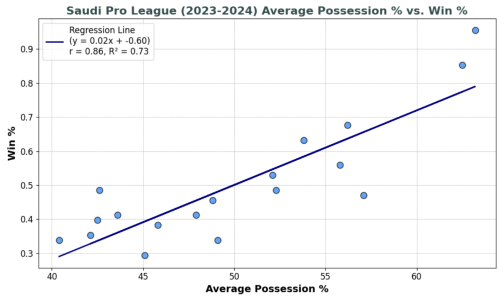
40% of the variability in win percentage can be explained by payroll, while 31% of the variability in win percentage can be explained by possession %. In this case, payroll shows to be more influential than possession percentage.

Saudi Pro League (Saudi Arabia)

Results Data: (ESPN-Saudi-Pro-League, 2024); Salary Data: (Capology, 2024)



r: 0.85, R^2: 0.73



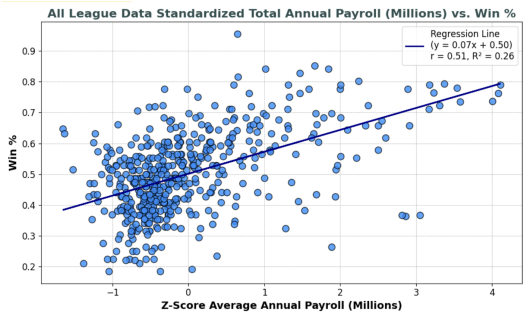
r: 0.86, R^2: 0.73

73% of the variability in win percentage can be explained by payroll, and 73% of the variability in win percentage can also be explained by possession %. Both stats factor into a team's win percentage equally. It's important to note that we only had 6 teams' payroll data for the Saudi League. Therefore, the payroll vs. win% R^2 has a greater margin for error.

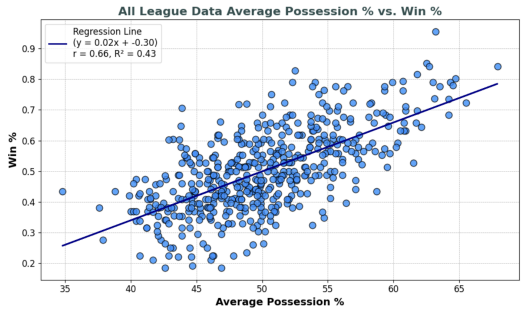
Aggregated Results

All League Data Combined

Cumulative Data: (Pan & Swenson, 2024a)



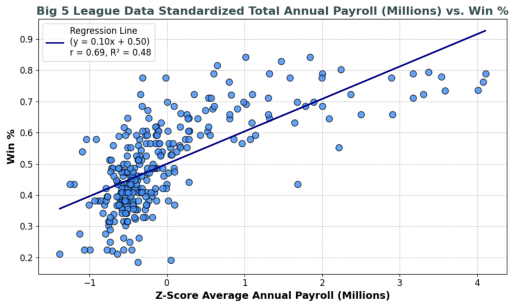
r: 0.51, R^2: 0.26



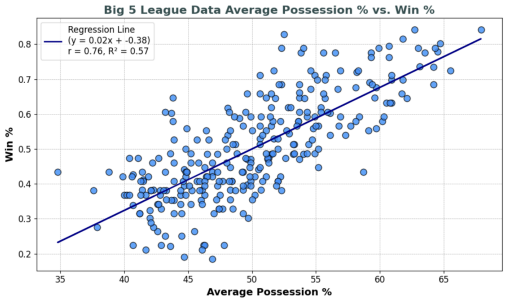
r: 0.66, R^2: 0.43

26% of the variability in win percentage can be explained by payroll, while 43% of the variability in win percentage can be explained by possession %. Across all compared leagues, it is clear that both total annual payroll and possession percentage have a strong influence on win percentage, but possession percentage appears to be more authoritative.

Big 5 Leagues



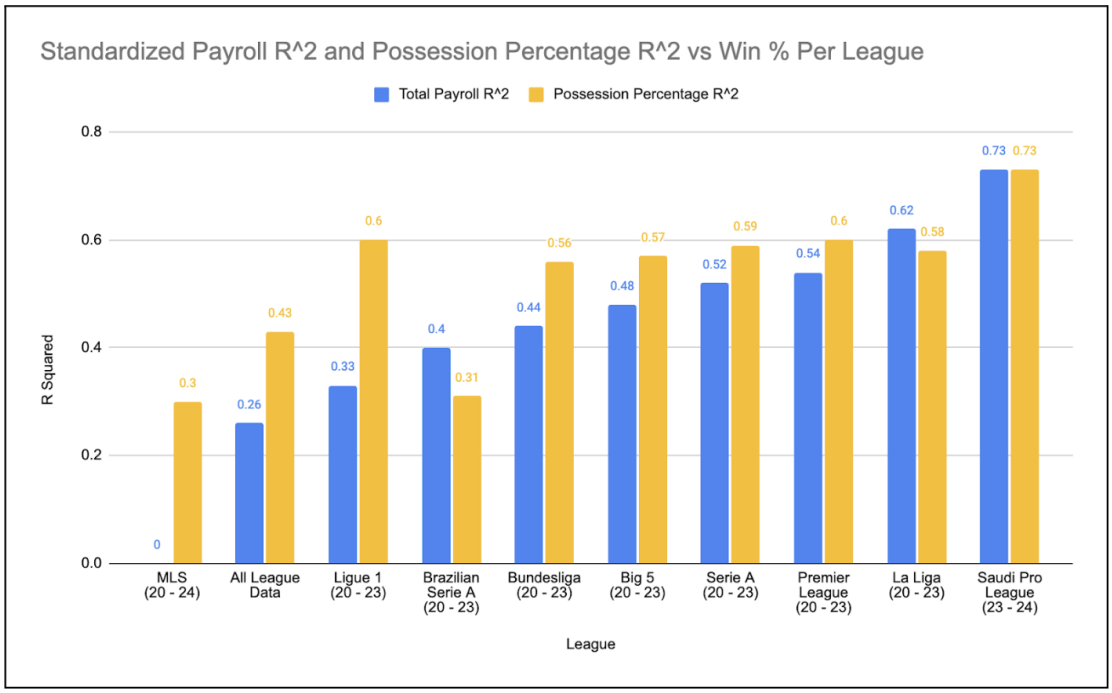
r: 0.69, R^2: 0.48



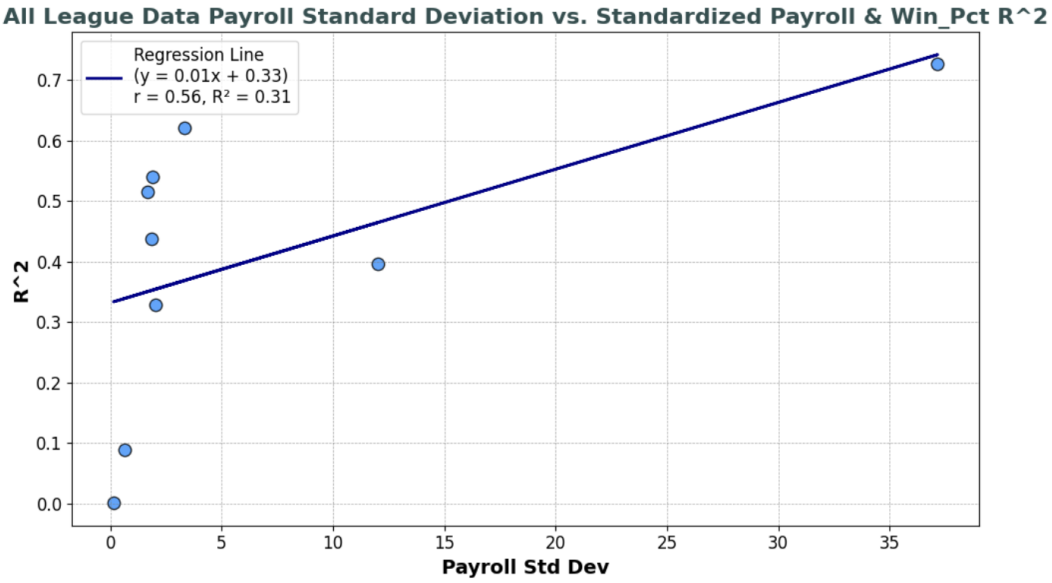
r: 0.76, R^2: 0.57

When examining the Big 5 Leagues (La Liga, EPL, Bundesliga, Serie A, and Ligue 1), 48% of the variability in win percentage is explained by payroll, while 57% of the variability in win percentage can be explained by possession %. Across these five leagues, the notion that both variables are heavily influential in a team's win percentage is reconfirmed. Moreover, it is affirmed that possession percentage seems to have a greater weight than total payroll in a team's win percentage.

Summative R^2 Comparison Across Leagues



For almost all leagues, we see that payroll R^2 is lower than possession % R^2. La Liga and the Brazilian Serie A are the two exceptions. We also observe that payroll R^2 tends to vary heavily across seasons, but this is relatively expected as we only examined seasons between 2020-2024.



We see that payroll standard deviation is positively correlated with the R^2 value between payroll and win percentage. A possible interpretation of this is that the more imbalanced a league is in terms of monetary investment, the more polarizing the results are in that league, which reveals the power of funding. A threshold in how much more a team needs to spend compared to the rest of the league to fully realize the winning potential of funding may exist.

Champions League Performance

Results Data: (UEFA, 2024); Salary Data: (Deloitte, 2024)

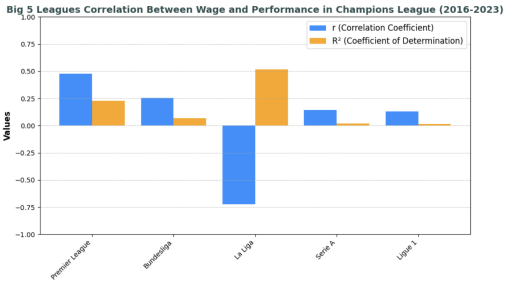
We also analyzed the correlation between cumulative league spending and performance in the Champions League for the Big 5 Leagues between 2016-2023. The Champions League is considered the most prestigious global club tournament where teams across leagues compete against each other.

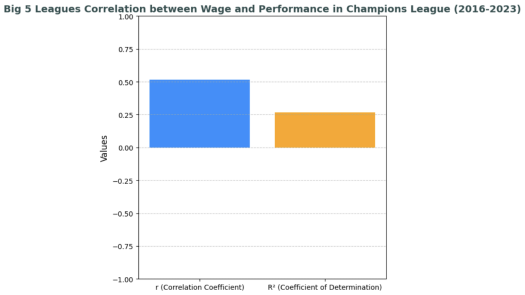
Performance is calculated as the sum of the playoff round each team from that league reached that year, divided by the number of teams:

$$\text{Performance} = \frac{\sum_{i=1}^{n}(\text{Team}_i\,\text{Performance})}{n}$$

This formula does favor leagues that have consistent performance (i.e., round of 8, round of 8, round of 8, round of 8) rather than leagues that have a few well-performing teams (i.e., winner, second place, round of 64, round of 64), which is a limitation of this type of

performance metric. In future analysis, the weights of each individual team performance can be changed.





We observe that for all Big 5 leagues aggregated across 2016-2023, there is a positive correlation between payroll and Champions League performance. Approximately 26% of the variance in Champions League performance can be explained by payroll. Interestingly, we see that La Liga has a strong negative r value, indicating that as that league increased payrolls, their Champions League performance actually dipped. This may be due to teams from La Liga dominating the Champions League before the 2020s (Real Madrid, Barcelona) and afterward having a dip in performance, or the lack of weighting when considering Champions League placement.

Predicting Win Percentage with Models

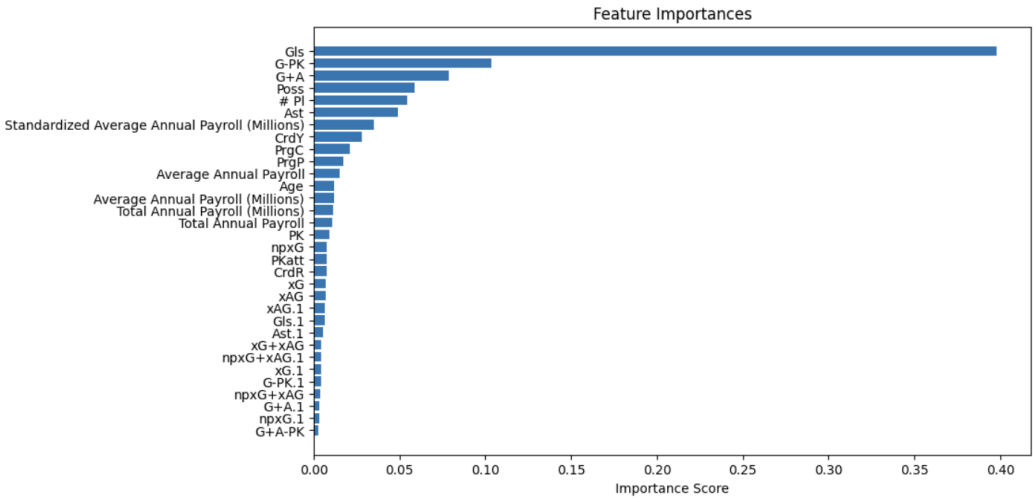
For our Random Forest model, we used the same training, testing, and validation datasets. The training and testing dataset consists of our curated soccer data found in the soccer section of this drive, excluding Saudi Pro League (as it lacks some features) and Premier League data (for validation dataset). Beginning with 48 features, we first cleaned the data by removing duplicate features: goals scored, goals allowed, and goal differential; unnecessary features in relation to win percentage—club name, matches played, game starts, and minutes played; and overly predictive features: season placement, wins, draws, and losses.

It is important to recognize that we are applying our own subjective judgment to unnecessary and overly predictive features, and we also are making the choice to leave out Premier League data for validation. There exists a data gap in Saudi data as well, which prevents us from using that data in model training.

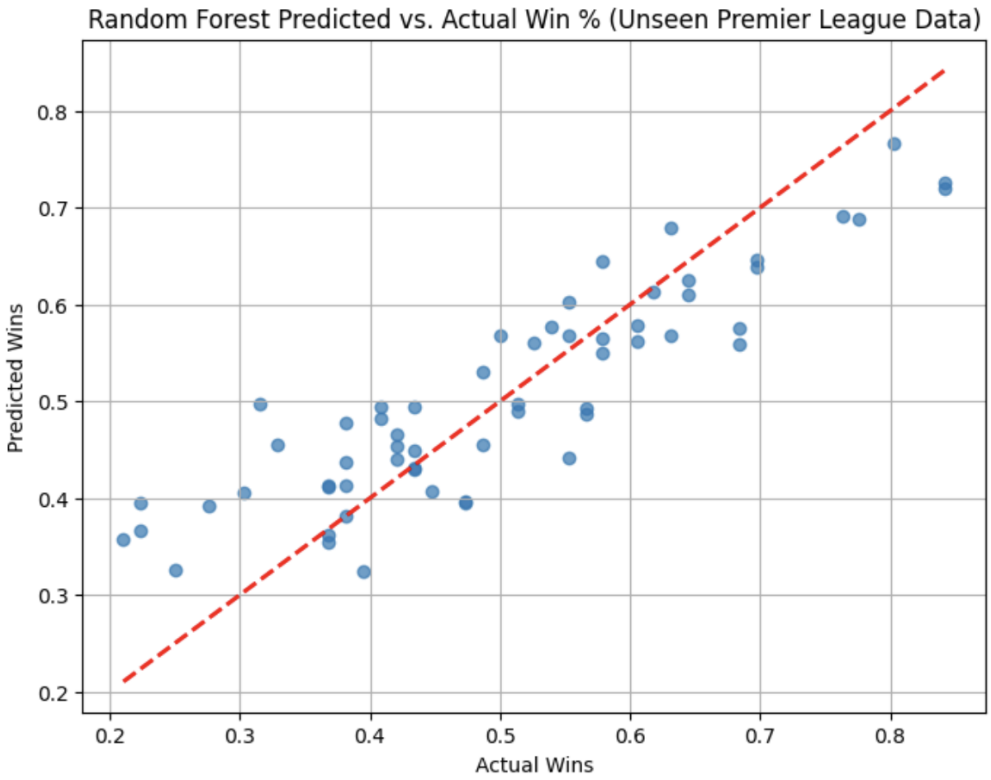
Random Forest

We roughly followed this guide: (GeeksforGeeks, 2024).
Our Random Forest model consists of training 100 decision trees on our training data, with each tree independently making predictions that are then averaged to produce a final predicted win percentage. Each "decision" a tree makes at a split is based on making the target values (win percentage) in each child node as homogeneous, as similar, as possible. Greater homogeneity in a child node means the model can make more confident predictions based on that node - which is what we are seeking (compared to random guesses).

The structure of the decision trees, where each split at a node is based on a threshold placed on a single feature, allows us to determine the relative importance of each feature. Feature importance is calculated based on how much a feature contributes to increasing homogeneity, aggregated over all the trees.



In-game analytics that undeniably affect the outcome of the game like goals scored + allowed and Poss (average possession percentage) were ranked highly by the model in importance score. In regards to the effects of funding on win percentage, standardized average annual payroll showed its importance being ranked over stats like progressive passes and carries. Furthermore, every variable involving funding ranked in the top 50% of all metrics in importance in determining the win percentage of a team.

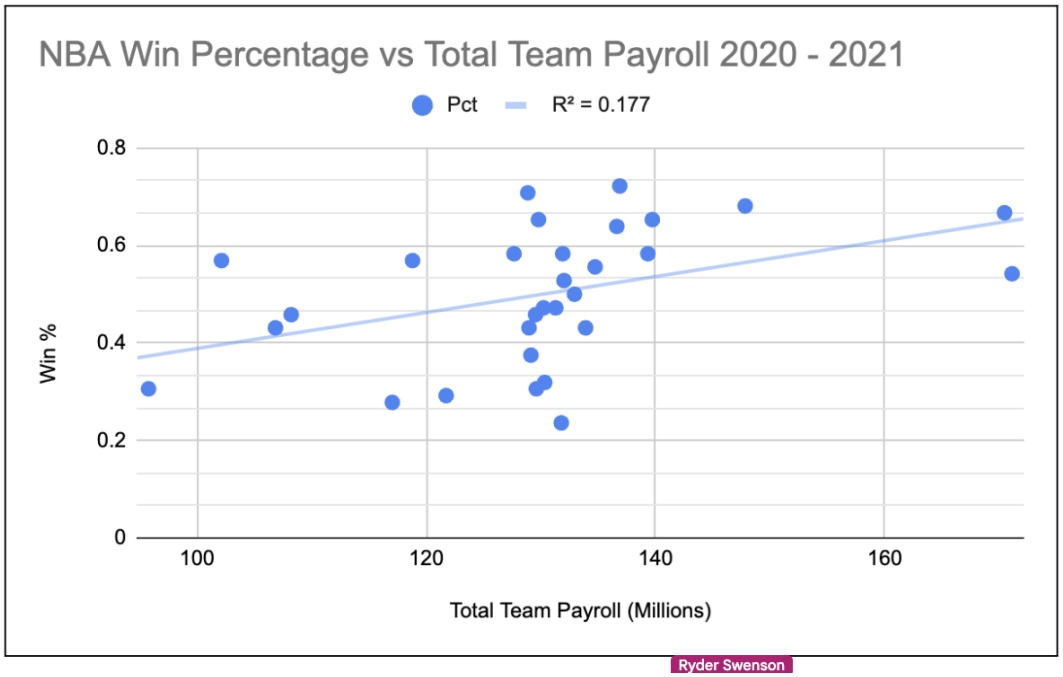


As shown in the graph above, the model was able to estimate an approximate win percentage given the input variables for each team in the Premier League. These points can be observed as vaguely following the line of perfect prediction (seen in red).

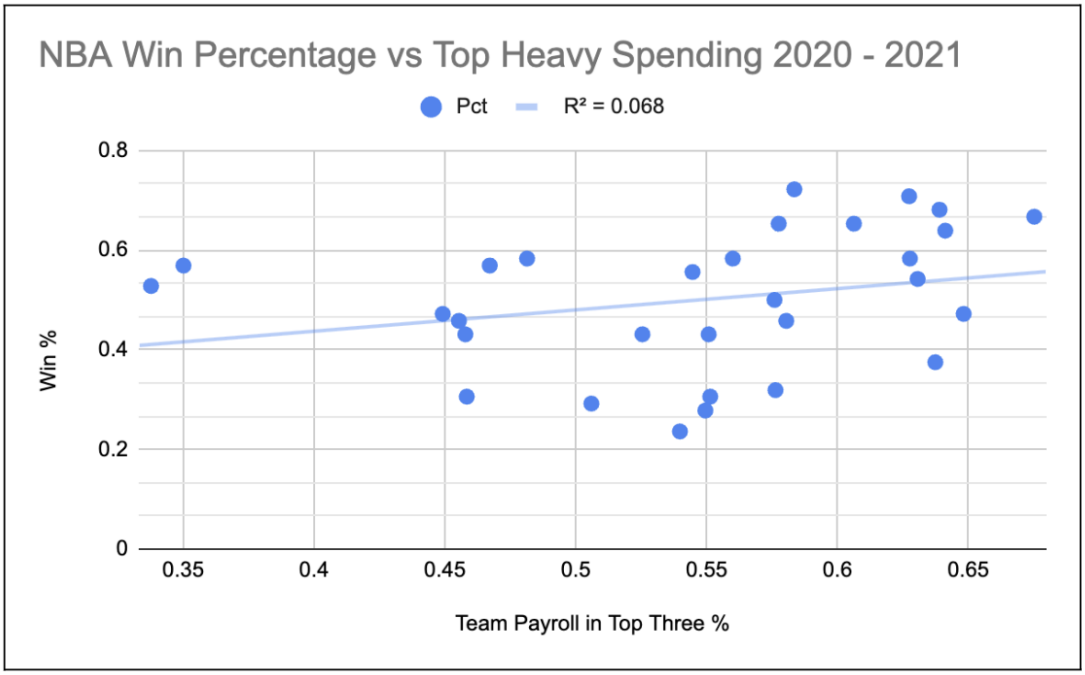
The random forest model achieves a mean squared error of ~ 0.0056 and explains approximately 76% of the variance in win percentage.

Basketball Results

2020-2021

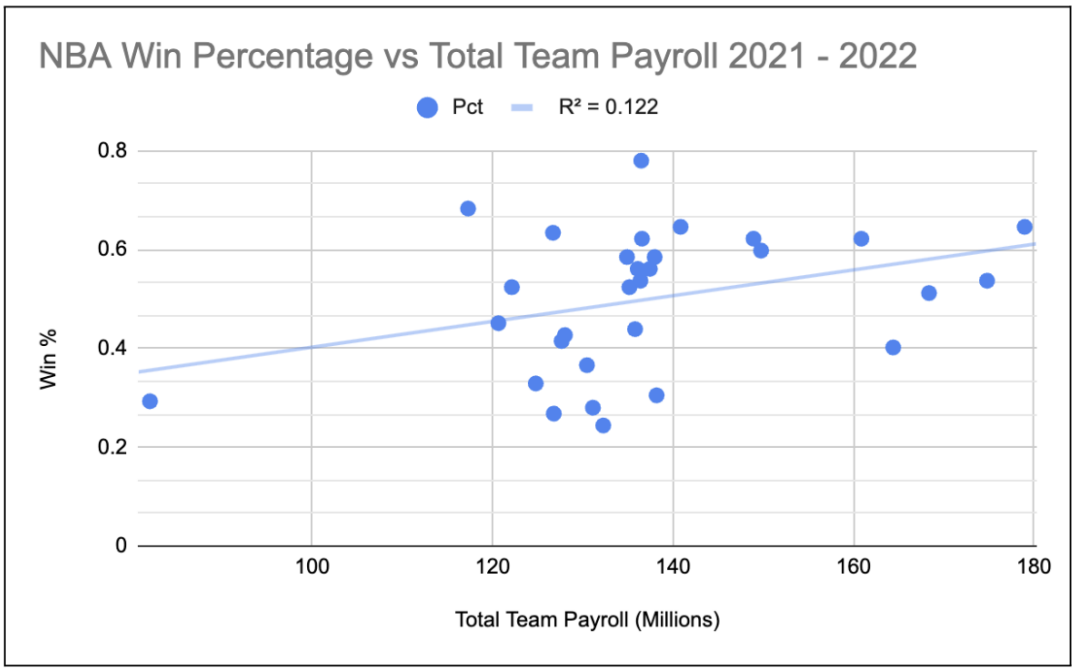


For this NBA season, there appears to be a slight positive correlation between team payrolls and win percentages during the regular season. This is followed by an R^2 coefficient of 0.177.

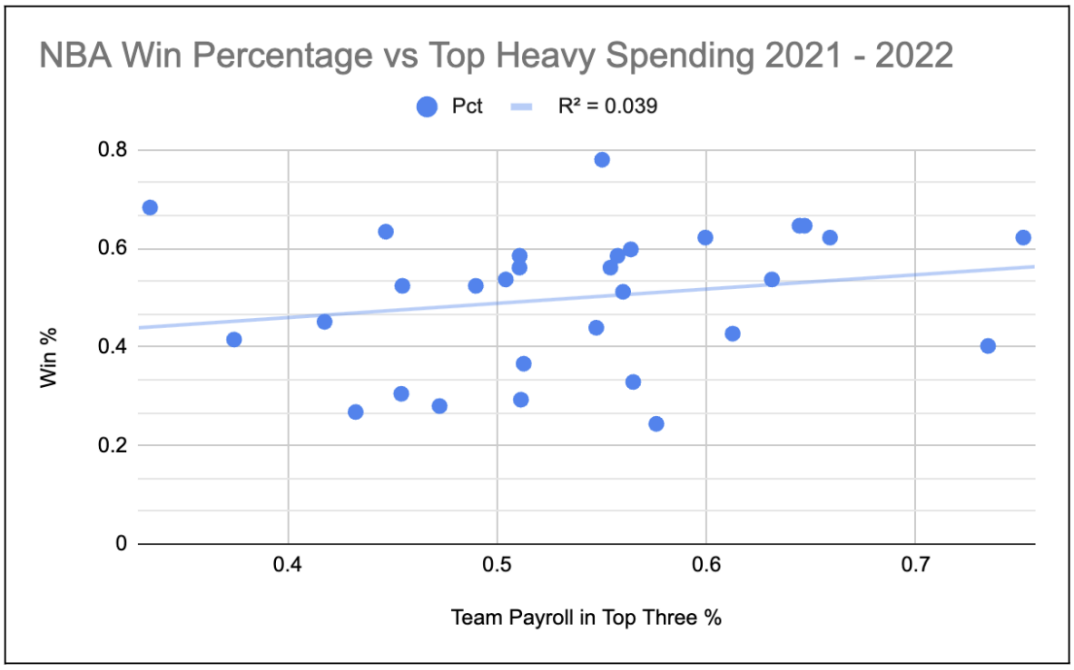


Moreover, there appears to be a slight positive correlation between the top-heavy-spending and the win percentage a team had. This is noted by an R^2 coefficient of 0.068. To define "top-heavy-spending", we calculated the %Top_3 value by summing the salaries of the three highest players on each team that year and dividing that sum by the team's payroll.

2021-2022

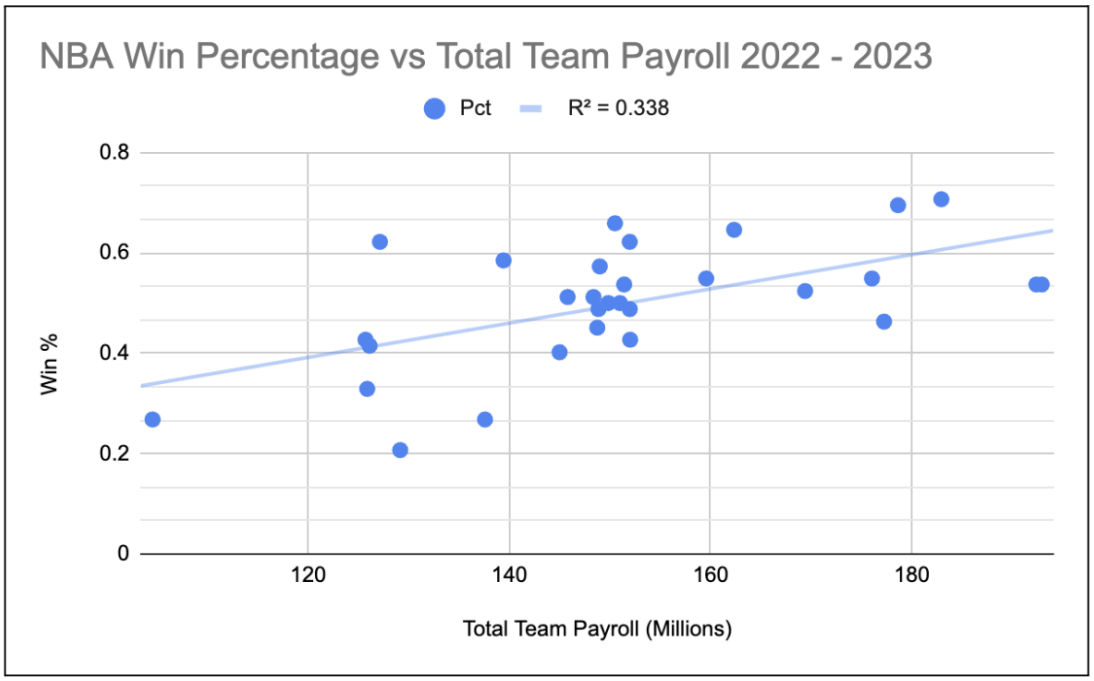


A similar pattern shows in the 2021-2022 NBA season, where another positive correlation between the total team payroll and their win percentage appears with an R^2 coefficient of 0.122.

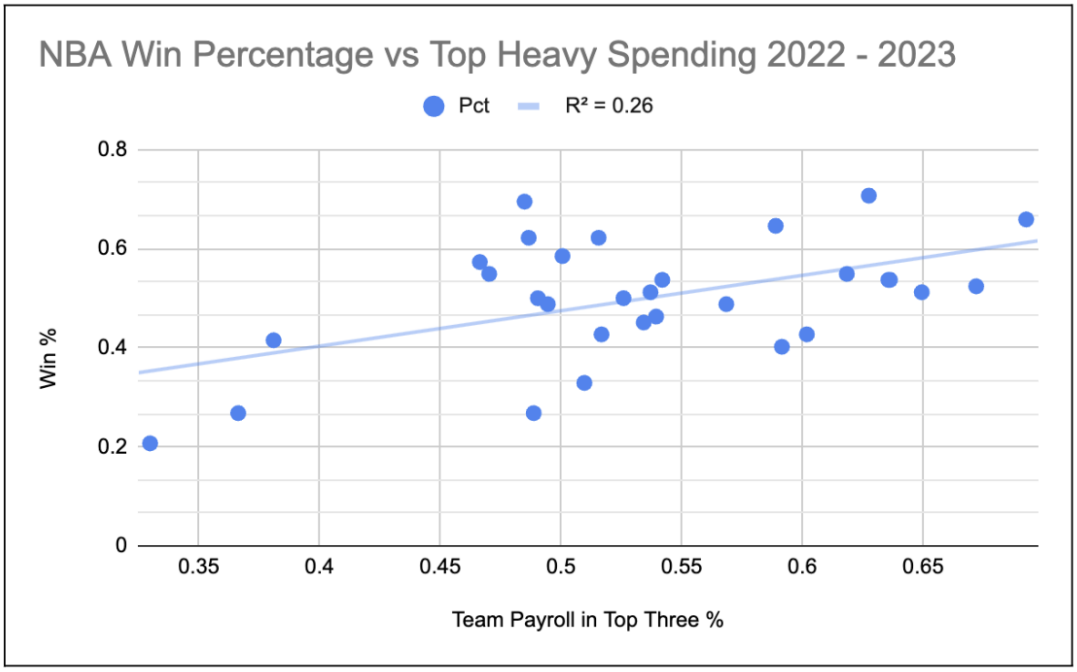


Here, the comparison between teams' top-heavy-spending and their win percentage shows another positive correlation with an R^2 coefficient of 0.039.

2022-2023

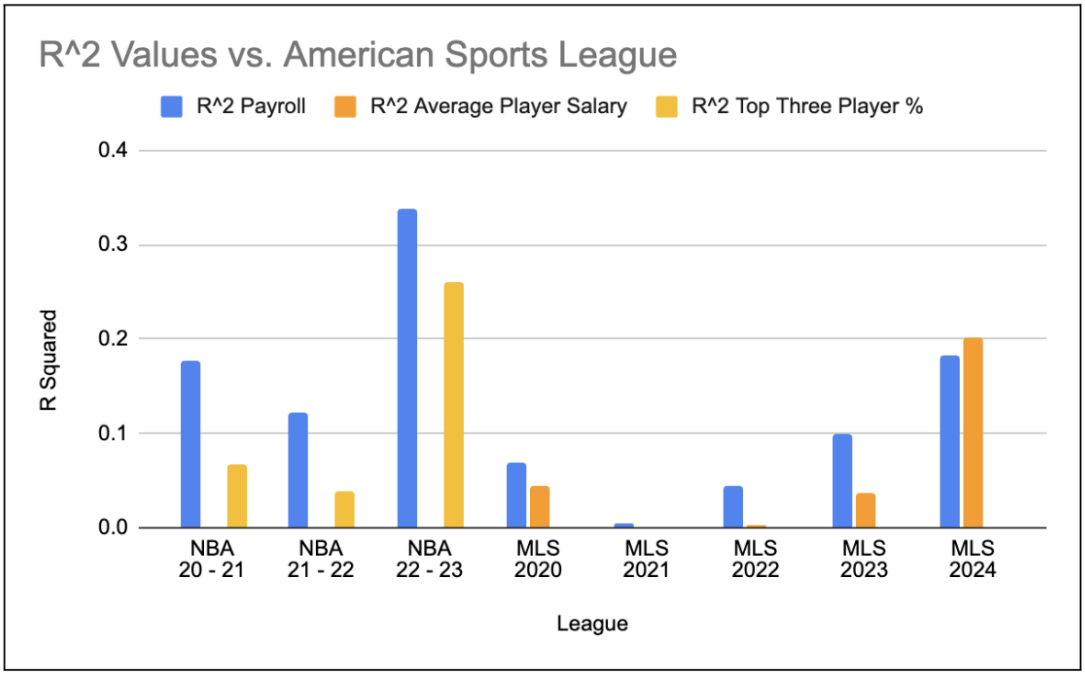


The 2022-2023 NBA regular season boasted a much higher R^2 value than average when comparing each team's total payroll and their win percentages. This large correlation is noted by an R^2 coefficient of 0.338.



An abnormally large correlation is observed between the percentage of team payroll focused on the top three earners and the team's win percentage, noted by an R^2 coefficient of 0.26.

MLS + NBA Analysis



In comparison with international leagues, American sports leagues seem to have a lower associated R^2 value in terms of the correlation between the importance of funding and win percentage. A possible explanation for this diminished correlation within American sports is the absence of systems of relegation. Systems of relegation are frequently featured in international soccer leagues where teams in the bottom of the ladder are in jeopardy of being demoted to a lower division of competition. For example, French and Italian soccer leagues have their upper level competitions, Ligue 1 and Serie A, as well as their secondary level competitions, Ligue 2 and Serie B. If a team finishes the season at the bottom of Ligue 1, they may be relegated to Ligue 2 in French soccer. This system of relegation is shared by every national soccer league that we have examined except for the United States' MLS. In fact, there are no major sports leagues in the United States that feature relegation.

Another explanation for the lack of magnitude in correlation between spending and win percentage in American sports is the presence of systems of finance balancing. These establishments within each major American sports league demonstrate the leagues' intention to prevent any financial juggernauts from winning year after year. One of these systems that encourage financial equality is an annual draft, in which the teams with lesser performance in prior years typically get favored in pick order. With earlier picks, these teams are able to draft more valuable prospects, bringing in more value and revenue to the organization as a whole. Other systems of financial regulation include salary caps and luxury taxes. Major American sports leagues like the NHL, NBA, and NFL all impose salary caps, and the MLB imposes a luxury tax on teams that exceed a predetermined threshold. These systems to rebalance teams in American leagues in terms of financial and performative standing serve as an explanation as to why increased funding is less impactful on win percentage and to why the R^2 coefficients that we have examined are lower in American sports leagues.

Challenges + Reflection

The main challenge we faced was finding accessible data. While we tried existing datasets on Kaggle and GitHub, the datasets we found lacked what we needed for our analysis. Thus, we had to compile our own datasets semi-manually using multiple different websites, which was incredibly time consuming. In reflection, this project has definitely highlighted the importance of learning how to make web scrapers in data science.

Additionally, our soccer data was skewed toward the Big 5 Leagues (European) and the MLS (American). While we did find small pieces of Saudi Pro League data, they only encompassed

one season and lacked much of the football statistics, preventing us from incorporating that data into our models. Recognizing this is important as our results in the aggregated data section should not be blindly generalized to world football leagues.

Summative Conclusion

Can we conclude that Money = Wins? Well, we could never really conclude that money equals or causes wins since we only observe correlations. Nevertheless, our findings do show that payrolls do play a direct role in determining win percentage. The r values between standardized average annual payroll and win percentage are almost always positive (except for a few outliers), illustrating that spending more money almost always leads to some gain in wins. R2 values between standardized average annual payroll and win percentage vary greatly between seasons and leagues, but cumulatively explain 26% of the variance in win percentage across all our soccer data. Our random forest model feature importance shows standardized average annual payroll as one of the most determinant variables in win percentage.

Does this mean teams should just start pumping in money? Definitely not. Our data looks at standardized average annual payroll with respect to a team's peers - if one team acquires and spends billions, other teams may match this (almost turning this into a game theory question).

References

AnalysisInn. (2020). The meaning of r, r square, adjusted r square, r square change and f change in a regression analysis. https://www.analysisinn.com/post/the-meaning-of-r-r-square-adjusted-r-square-r-square-change-and-f-change-in-a-regression-analysis/

Bundesliga. (2024). Bundesliga table. https://www.bundesliga.com/en/bundesliga/table

Capology, I. (2024). Capology. https://capology.com/

Deloitte. (2024, June). Money spent on wages by the big five soccer leagues in europe from 2016/17 to 2022/23, by league (in million euros) [graph] [In Statista. Retrieved December 10, 2024]. https://www.statista.com/statistics/1022140/european-soccer-wage-costs-by-league/

ESPN-Brazilian-Serie-A. (2024). Brazilian serie a table. https://www.espn.com/soccer/standings/_/league/bra.1

ESPN-La-Liga. (2024). La liga table. https://www.espn.com/soccer/standings/_/league/esp.1

ESPN-Ligue-1. (2024). Ligue 1 table. https://www.espn.com/soccer/standings/_/league/fra.1

ESPN-NBA. (2024). Nba standings. https://www.espn.com/nba/standings

ESPN-Saudi-Pro-League. (2024). Saudi pro league table. https://www.espn.com/soccer/standings/_/league/ksa.1

ESPN-Serie-A. (2024). Serie a table. https://www.espn.com/soccer/standings/_/league/ITA.1

GeeksforGeeks. (2024). Random forest regression in python. https://www.geeksforgeeks.org/random-forest-regression-in-python/#

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

MLS. (2024). Major league soccer standings. https://www.mlssoccer.com/standings

Pan, K., & Swenson, R. (2024a). Big money big wins. https://drive.google.com/drive/folders/1leaDniF1ShpVstwdgYO0GI3NNU2czv9d?usp=drive_link

Pan, K., & Swenson, R. (2024b). Big money big wins. https://github.com/kevin-pan-221/big_money_big_wins

pandas development team, T. (2024, September). Pandas-dev/pandas: Pandas [DOI: 10.5281/zenodo.13819579]. https://doi.org/10.5281/zenodo.13819579

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

PremierLeague. (2024). Premier league table. https://www.premierleague.com/tables

SportsReference. (2024). Football statistics and history. https://fbref.com/

StatisticsHowTo. (2024). Z-score: Definition, formula and calculation. https://www.statisticshowto.com/probability-and-statistics/z-score/

UEFA. (2024). Uefa champions league history. https://www.uefa.com/uefachampionsleague/history/

Ventures, F. S. (2024). Hoopshype. https://hoopshype.com/salaries/

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021