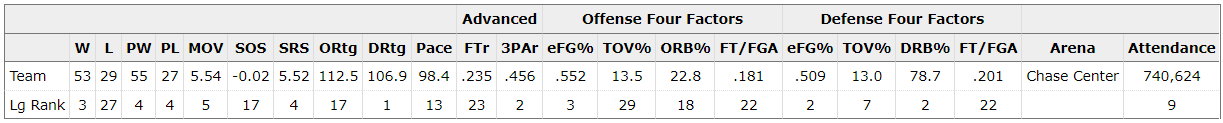
**Data**

**Data Collection**

Basketball Reference is a basketball-related website that collects and provides every season, team, and player statistics in the history of the NBA. The website is the project's primary source of data and only regular season data will be used to predict the champions. A sample of the data is provided below, showing statistics for a single team during a specific season.

A screenshot of a computer

Description automatically generated with medium confidence



The data is divided into three types of statistics: "Team", "Opponent", and "Miscellaneous". As their name suggests, team and opponent stats represent each team and their average opponent per game numbers, respectively. Miscellaneous is a diverse and advanced type of statistics, some of which involve calculations with game numbers. The definition of each variable can be found in the data dictionary in appendix X.The information from the dataset can be helpful in developing a profile of a championship team with its more notable characteristics. Additionally, these variables can provide an overall picture of each team's strengths and weaknesses which will be critical when calculating their likelihood of becoming a champion.

The dependent variable is the number of post-season (or playoff) wins. To become a champion, a team must reach 16 wins, i.e., win four 7-game series. Hence, the reasoning behind the analysis is that the team with the highest predicted number of wins will have the highest chance of winning the championship.

**Data pre-processing**

The first step was to consider data after the 2002-2003 season. Starting that season, every playoff matchup, from the first round to the finals, was a best-of-seven series or the first to win four games. Including previous seasons would have skewed the results as the number of games required for winning the title has changed over time. Additionally, only teams who qualified for each year's playoff were included in the dataset (i.e. 16 teams, 8 for each conference).

One of the main problems when tracking statistics in sports is the constant change of game numbers due to improvements in coaching, playstyle, strategies, etc. The pace is defined as the total number of possessions per game and is an excellent metric to determine how the game is played during an era. Appendix X shows an increase in pace from the 2003 to 2022 season, which led to higher points per game during the same stretch since more possessions lead to more opportunities to score. The following appendix (X) shows the number of 3-point attempts in the same time span. There is a noticeable increase in the number of 3-point attempts from the 2013 season, a year in which Stephen Curry, the greatest shooter of all time, broke for the first time the 3-point record. Since that year, the league noticed the value of the 3-pointer and started to apply it in their strategies. Those examples showed the constant change of the game, even for a small period of time. Therefore to rectify this issue, ranks were used instead of game numbers. Finally, highly correlated variables were removed as they do not provide valuable insights, do not improve the model, and will help reduce the complexity of the model.

**Data Exploration**

Before performing the predictive analysis, a brief exploratory analysis was performed to acquire insights into which features o patterns previous championship teams have in a visual manner. The ranks of some statistics (e.g., points per game) of every championship team since 2003 were plotted to better understand what leads a team to win the title.

Regarding offence, two statistics were plotted: offensive rating and effective field goal percentage (EFGP). The offensive rating measures offensive performance and refers to the number of points scored per 100 possessions. EFGP is a measurement of shooting effectiveness and is a better indicator than field goal percentage as it adds more value to the three-point. In appendix X, it seems that most championship teams had a top 10 offensive rating in the league (red line) and a top 5 EFGP. This may suggest that teams that take advantage of every procession and high shooting efficiency have higher chances of winning more playoff matches.

On the defence end, defensive rating and opponent EFGP were plotted. Similar to the previous two variables, defensive rating refers to the number of points a team allows, while opponent EFGP shows how well the opposing team shoots against your team. In appendix x, both variables seem correlated, and championship teams have to be part of their top 10. Those statistics show the importance of being disciplined on defence and back the phrase "Defense wins championships".

In addition, some miscellaneous stats were plotted. Appendix x depicts the number of regular season wins a champion had. Only two years after the champion positioned themselves outside of the top 5. Simple system rating (SRS) is "a [team evaluation rating](https://www.nbastuffer.com/team-evaluation-metrics) that takes into account average point differential and [strength of schedule](https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/)" (O'Neill, 2021). The rating is fairly accurate when evaluating a team's performance; on the graph, it seems that champions fell in the top 5. Finally, the last chart (Appendix X) illustrates the margin of victory and the teams, on average, were ranked third among all teams in the league. This might indicate that winning games are insufficient; instead, it is essential to consider how a team wins (e.g., a score blowout) as it shows their dominance and superiority against their opponents.

**Model Risks**

Although the results were satisfactory, the model has flaws and does not consider multiple common aspects in sports. It should be noted that we are only using statistics from the regular season, so any event that occurs during the playoffs will not be considered in the model. Injuries are the worst fear of every team, especially if it happens to the team's star. They are unpredictable and can occur during the playoffs and completely change the course of a team's playoff run. Players' physical status is not the only scenario to care about; psychological status also matters. Being able to deal with the pressure of the game, which includes in-game distractions (e.g., fans), media and the team's history on your back, and not suffer a mental breakdown comes with more exposure to such an event, which is also known as experience. There are more intangible factors that are impossible to collect, such as coaching, strategy, etc., which might considerably impact how the team will perform in the playoffs. Finally, all those variables plus the "unpredictability" nature of sports make the model not perfect but sufficient to fulfill the project's objective.

Another issue is the output of the models. When predicting the number of playoff wins, both regression and random forest championship team results had an expected number of playoff wins lower than 16. Also, regression obtained "negative wins" as output for teams. The size of the sample data does not allow a very accurate and robust model. However, the results can be interpreted as the team with the highest predicted wins will have the most chances of winning the title.

**Conclusion and Recommendation**







<https://medium.com/bench-connection/context-matters-the-eastern-and-western-conference-11bb4b73136b>