JT Flume, Anant Gupta,
Raymond Li, and Emily McCullough

An In-Depth Look at Home Court Advantage

Description of Dataset

The data set used includes extensive data of every NBA season from 2004 up until 2020, the most recent season, and is a merged combination of two CSV files. The first contains information about every game played since 2004. The rows represent one game and the columns are key gameplay statistics, such as each playing team's points, assists, and field goal percentages. The second contains individual player performance in every game that the player has played in since 2004. The rows represent one player in one game and the columns are the individual's key gameplay statistics, similar to the first file. The game ID column acts as the key to link both datasets together. Our goal is to manipulate these datasets to explore whether or not 'home-court advantage' truly exists.

Importance

Our team is interested in using empirical evidence to determine whether game location is a statistically significant variable on team performance in categories including individual player stats at home, team stats at home, and the effect of playing at home on the probability of winning. We hope to see if the game location alone is enough to increase the chance of winning, as this could merit the maintenance and continued investment in harnessing home field advantage. This would be pertinent to NBA teams, as this would support such ideas as promoting fan attendance, active fan engagement, and other ways of creating an impactful culture at home.

In addition, the NBA playoffs are structured such that the team that performs better in the regular season, measured by win-loss ratio, will maintain home court advantage over a seven game series, meaning that this team will have the opportunity to play, at most, four games at home while the other team can only play, at most, three games. Thus, if there is truly a significance to playing at home, there may be an additional incentive to cultivate home court advantage in securing more home game opportunities in the playoffs that could potentially lead to more victories, or even a championship.

Exploratory Data Analysis

The first question to investigate is whether the home team actually wins more games. Adding a column to calculate the difference between points scored by the home team versus the away team, we found that the distribution was skewed to the left, indicating that there are more games in

which the team at home won (Figure 1). This is evidence enough to continue investigation into this subject.

The first aggregation we did was on the season level to determine whether this was a trend across all seasons. Grouping by season, we found that from 2004-2018, teams won at least 58% of their home games by a minimum of a 2.28 point lead (Figure 2-3). There is a sharp decline of this trend in 2019 that can be attributed to the COVID-19 pandemic and its impact on the sport. This will be explored later in the analysis.

The second aggregation we did was on the team level. To do this, we split the data into two data frames grouped by either the home team or away team and aggregated by mean of the gameplay statistics at home or way. We then joined these two dataframes to get one in which the rows were one team and the columns were that team's statistics at home and away. This allowed us to compare each team's performance at home versus away.

The first bar plot (Figure 4) displays that for each year, of the total number of games won, what percentage of the victories occurred at home versus the percentage of wins that occurred on the road. From 2003 to 2018, there is about a 20% margin between games won at home and games won on the road. This margin between home and away win rate decreases in 2019 and 2020, the seasons affected by COVID-19.

The second group of visualizations (Figures 5-10) are dodged bar plots that display the home-team on the x-axis and a given statistic on the y-axis. The plots provide insight on average points scored, average assist, average free throw percentage, average rebounds, average three point field goal percentage, and average overall field goal percentage respectively for each team's home and away games. All visualizations show a trend that a team will have a higher average when playing at home compared to when playing away. This is evidence in favor of there being a home court advantage.

Next, we explored how teams fared during the 2020 season in regards to the COVID-19 pandemic eliminating crowds at games. One hypothesis of the reason for home court advantage is that the energy of the crowd in the home city will positively affect the home team. Our team aggregated the data to include games that occurred after December 23rd, 2020 (the first day of the 2020 NBA season). We plotted the same statistics for this time period (Figure 10-15) and discovered that the clear higher average statistics for playing at home does not exist for the COVID-19 season. This is evidence for the fact there may be a hidden variable, fan attendance, that is not included in our dataset.

The final aggregation we did was on the individual level. We wanted to investigate whether home court advantage is seemingly applicable to top

NBA players, such as Lebron James or Luka Doncic. Players were split into two groups, all-stars and role-players. An all-star was quantified as a top player since 2015 based on points scored. To do this, we filtered the games played to the time period we wanted, grouped by player name and team, and aggregated by the mean of points the player had scored while playing for that team. We then sorted by points scored to take the top 20 players. We ended up getting 7 players repeated in this series, since they had performed well in multiple teams. We chose to keep these repeats in the top players since it is possible performance could change based on team location. We then plotted a dodged bar plot for these top players to measure their average points scored when playing at home vs away (Figure 17).

On the other hand, role players were quantified as players since 2015 who have averaged between 10 and 15 points in any given year with any given team. Since there are many players who average between 10 and 15 points, we took a random sample of 15 of these types of players and plotted them by their average points scored at home vs away in any given year (Figure 18). The plots show that for all-star players, home games have not been as significant as for role players. This is perhaps a testament to the skill level of these all-star players to perform at a high level no matter where they are playing.

Solutions

First, we took the game's data for predicting the win/loss for a team. A single record of this data had game stats both for the home as well as the away team. We separated the home and away team data in two separate dataframes and concatenated them together. The initial number of records in our data was about 24k records. As a result of the concatenation, we had doubled the records for the model. This made sure that we were not feeding the data for both the home and away teams at the same time to the model. The independent variables are a combination of 6 numerical features (Points, Field Goal Percentage, Free Throw Percentage, Field Goal 3 Percentage, Assists, Rebounds) and 1 categorical feature (Home/Away flag). The dependent variable for the modeling was the Win prediction based on the stats of the particular team’s performance in a particular match. There was a correlation of 0.66 between Points and Field Goal conversion percentages. Correlation between Points and Assists was 0.6 (Figure 19)

The baseline accuracy for the prediction was 50% since wins and losses were in equal proportion. We used 3 different models: logistic regression, random forest, and gradient boosting. In the logistic regression we achieved 76.38% accuracy on the training set and 76.30% on the testing set which implies that the model is not overfitting. Random forest gave an accuracy of 82.87% on the training set and 75.91% on the test set, which was lower than that

provided by Logistic regression. Finally, the accuracy provided by the gradient boosting algorithm was 78.56% on the training set and 76.34% on the test set, just marginally higher than the accuracy of the logistic regression model on the test set.

Based on the models created, we found that the most important parameters for the random forest and gradient boosting models are field goal percentage and rebounds, while for logistic regression, the parameter with the greatest impact was three point field goal percentage (Figure 20-22).

Insight

As expected, our models show that there are many other factors that have a great impact on the outcome of a game. Still, it is clear that teams perform better in home games. Additionally, the lack of correlation to the location flag makes sense considering that the home advantage only adds a marginal amount to game performance when taking the overall performance into consideration.

Clearly, we see that the previously observed overperformance of home teams diminished during this season without the audience being a factor in many games across various statistical categories (points, rebounds, assists, etc). This provides further evidence that even though teams traveled to play games during this season, the location did not cause teams to perform better or worse.

Interestingly, when looking at all star players we see that they tend to perform about the same in both home and away games. Thus, adding a count for all star players on a team does not increase prediction accuracy. In contrast, players below this level tend to be affected by the game location. This suggests that if home court advantage exists, it affects those players with less skill and experience, which could be an impactful insight for team managers.

In conclusion, though game location plays a role in team performance, we think our research merits further analysis with more quantifiable attributes related to home team advantage for each team such as crowd size and crowd noise level that could vary from team to team and may underlie this benefit. While location may not necessarily be statistically significant, teams should continue to invest in ways that make their home court a tougher environment for opposing teams to play in, and possibly take into consideration other areas that may directly relate to home advantage. By improving in these areas and better capitalizing on home team advantage, teams should be able to increase competitiveness at home leading to better at home win percentages, which could lead to securing extra home games during the playoffs for increased playoff performance.

Appendix

Figure 1.

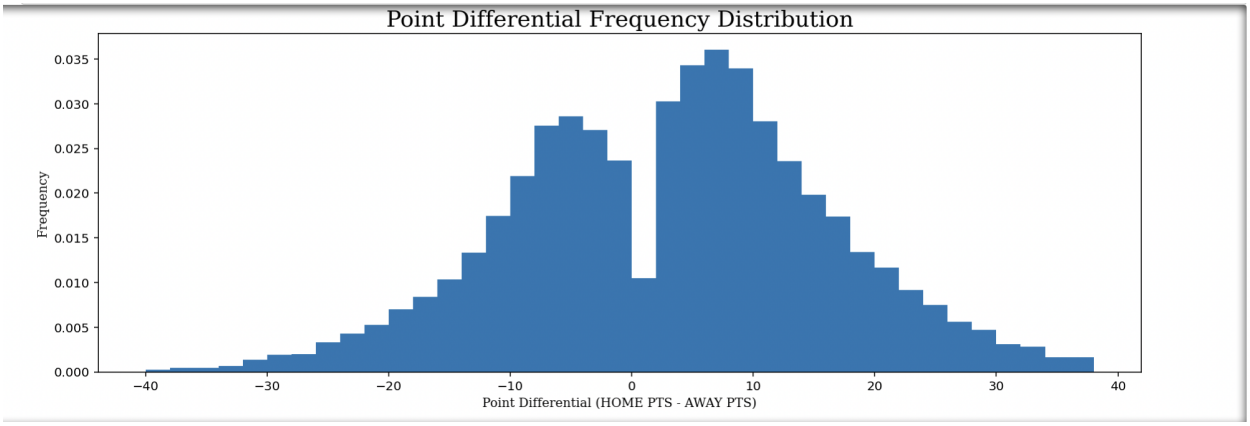


Figure 2.

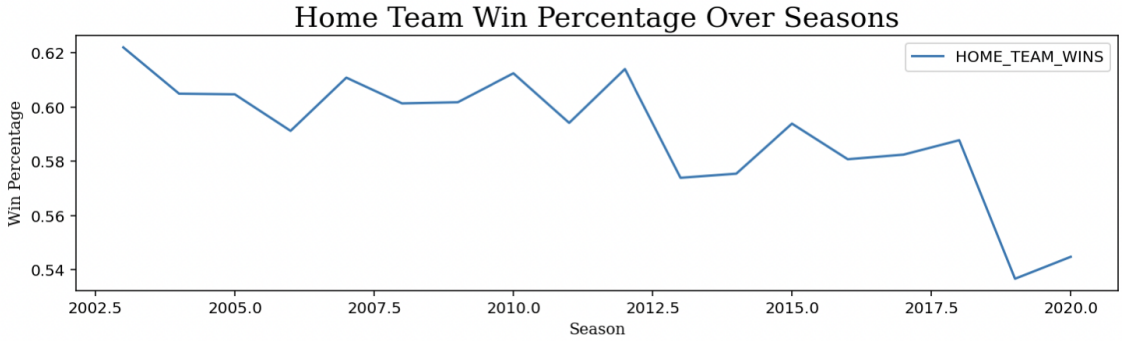


Figure 3.

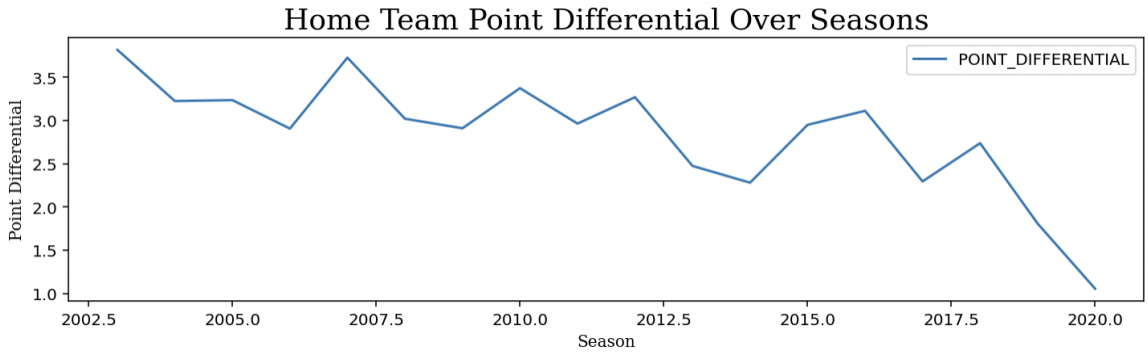


Figure 4.

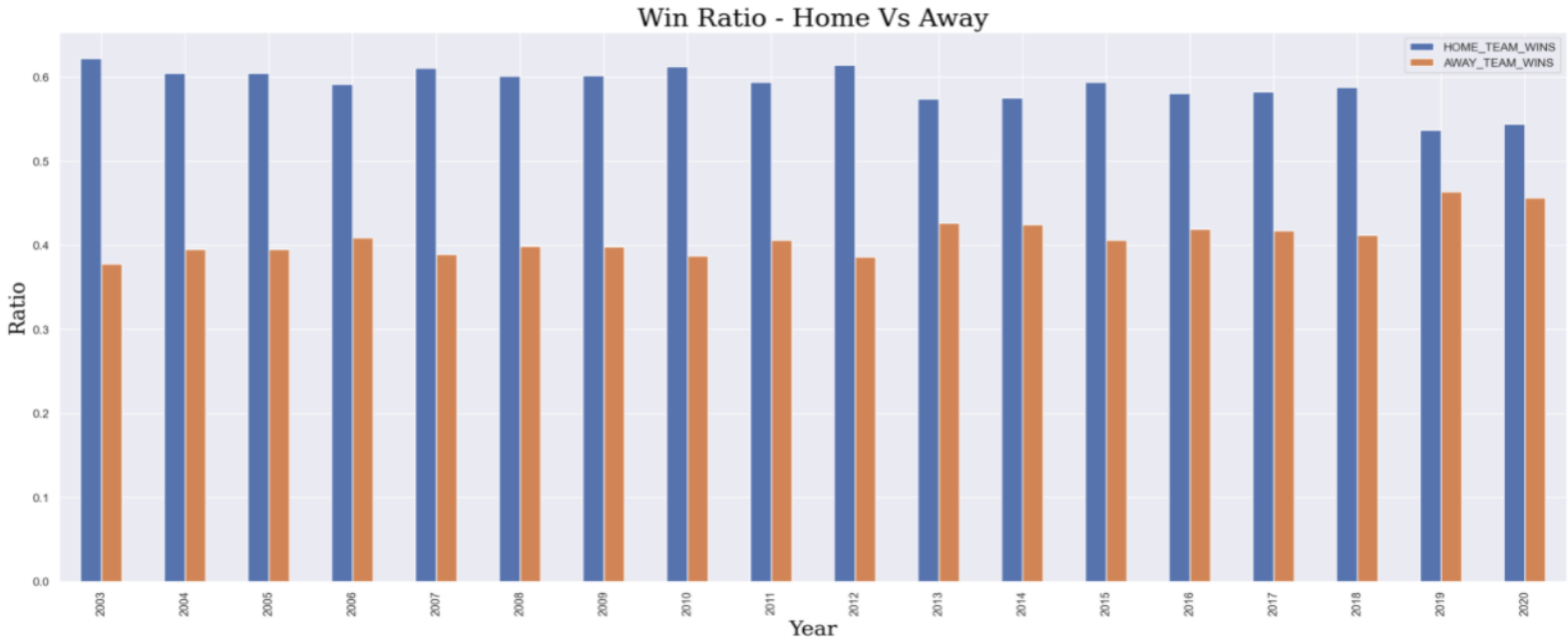


Figure 5.

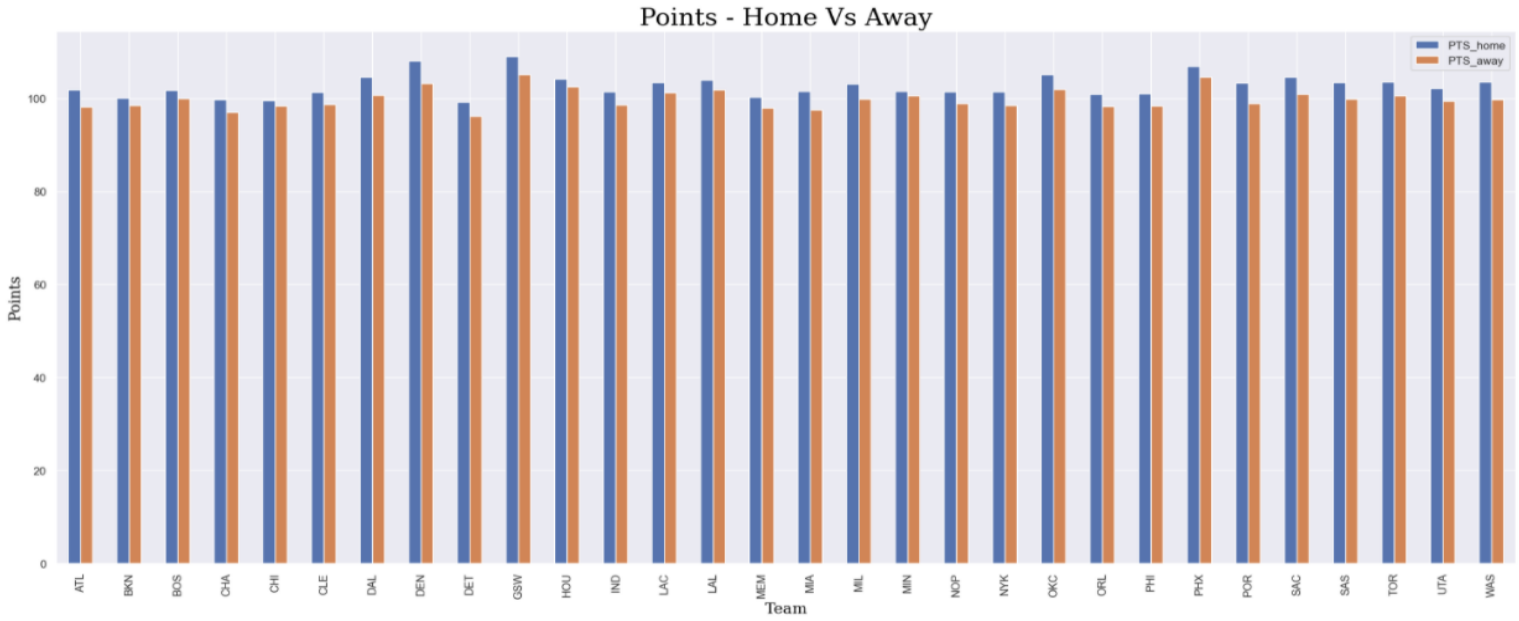


Figure 6.

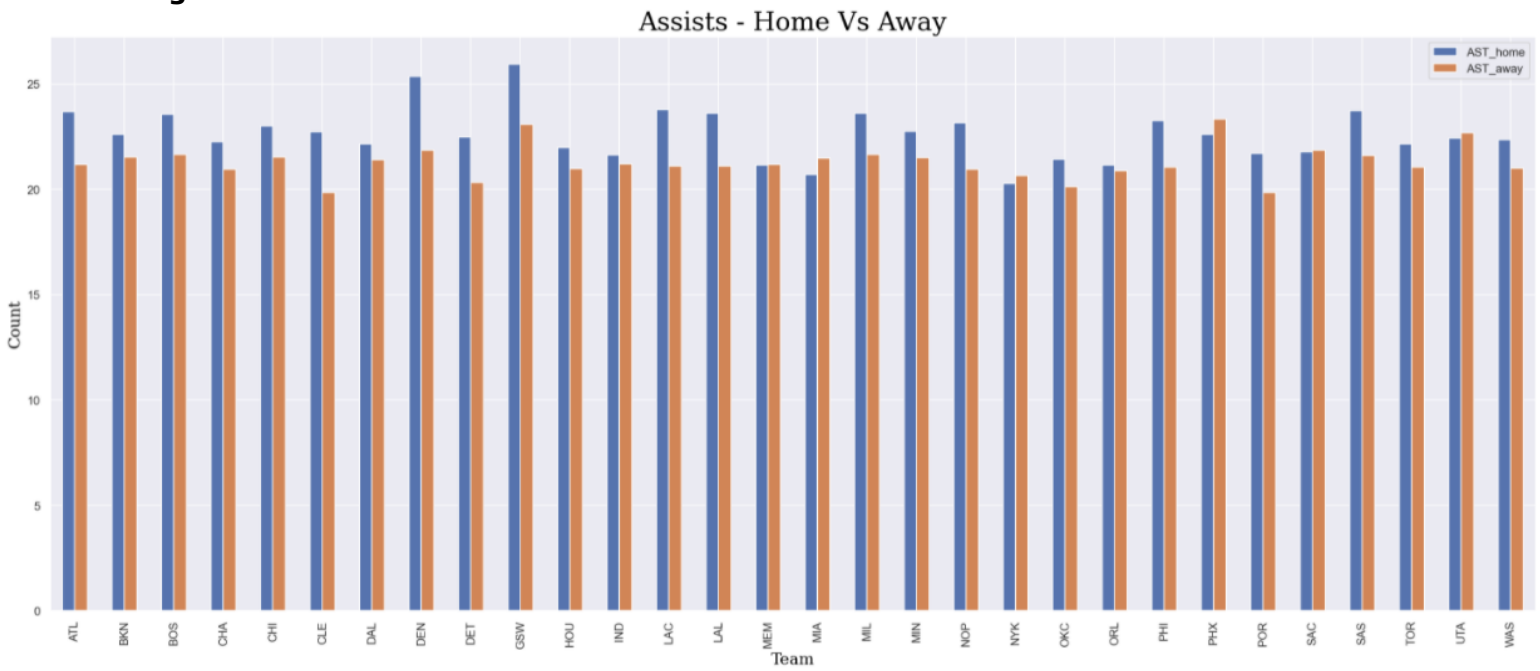


Figure 7.

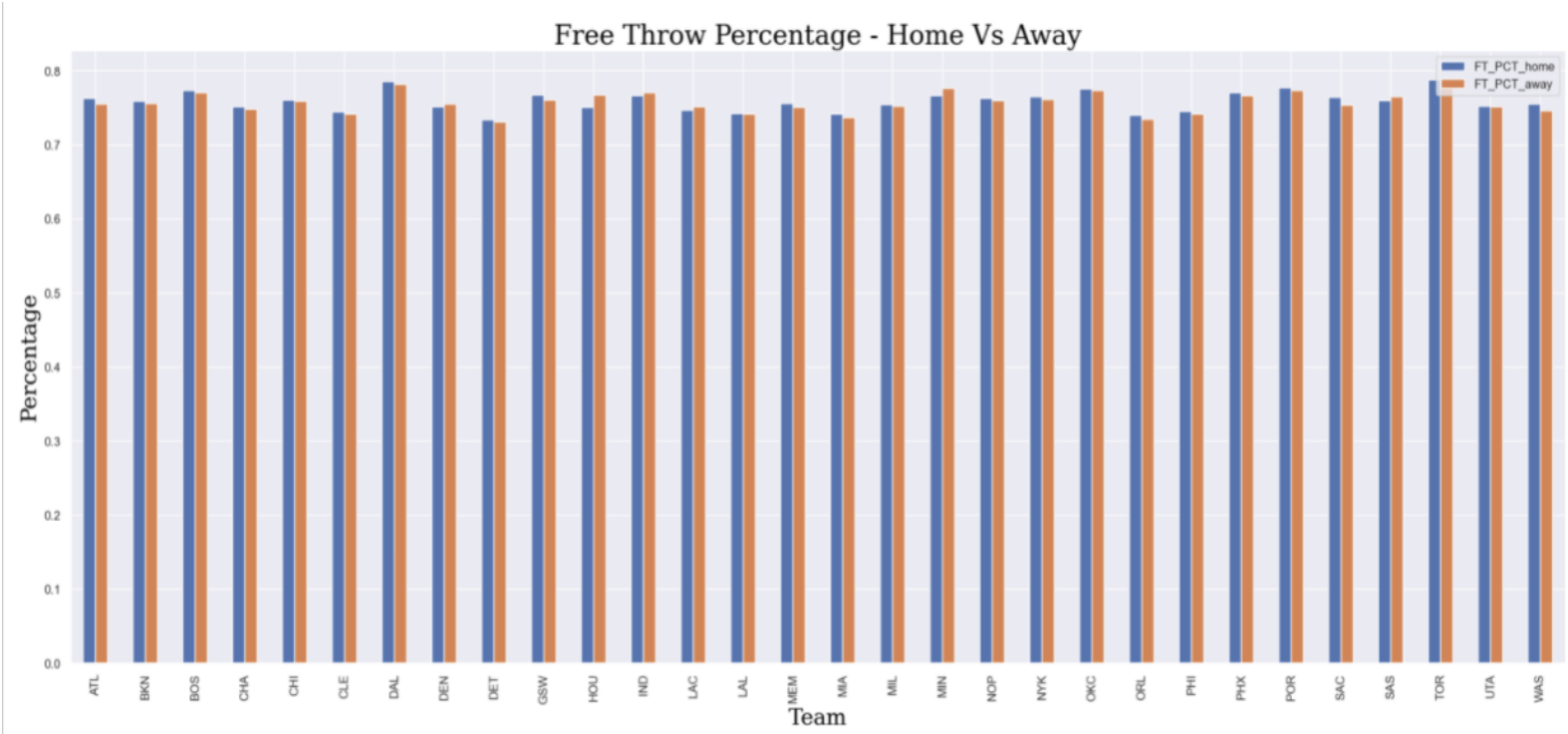


Figure 8.

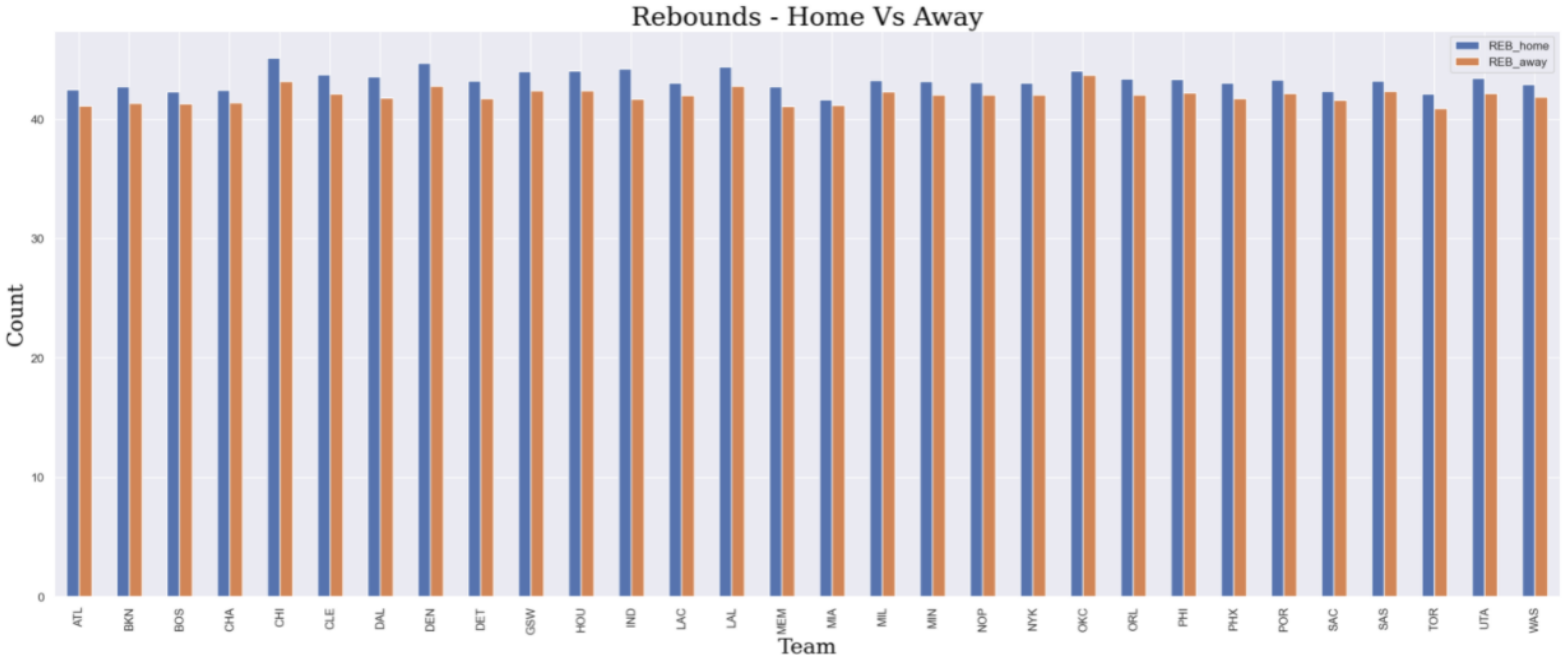


Figure 9.

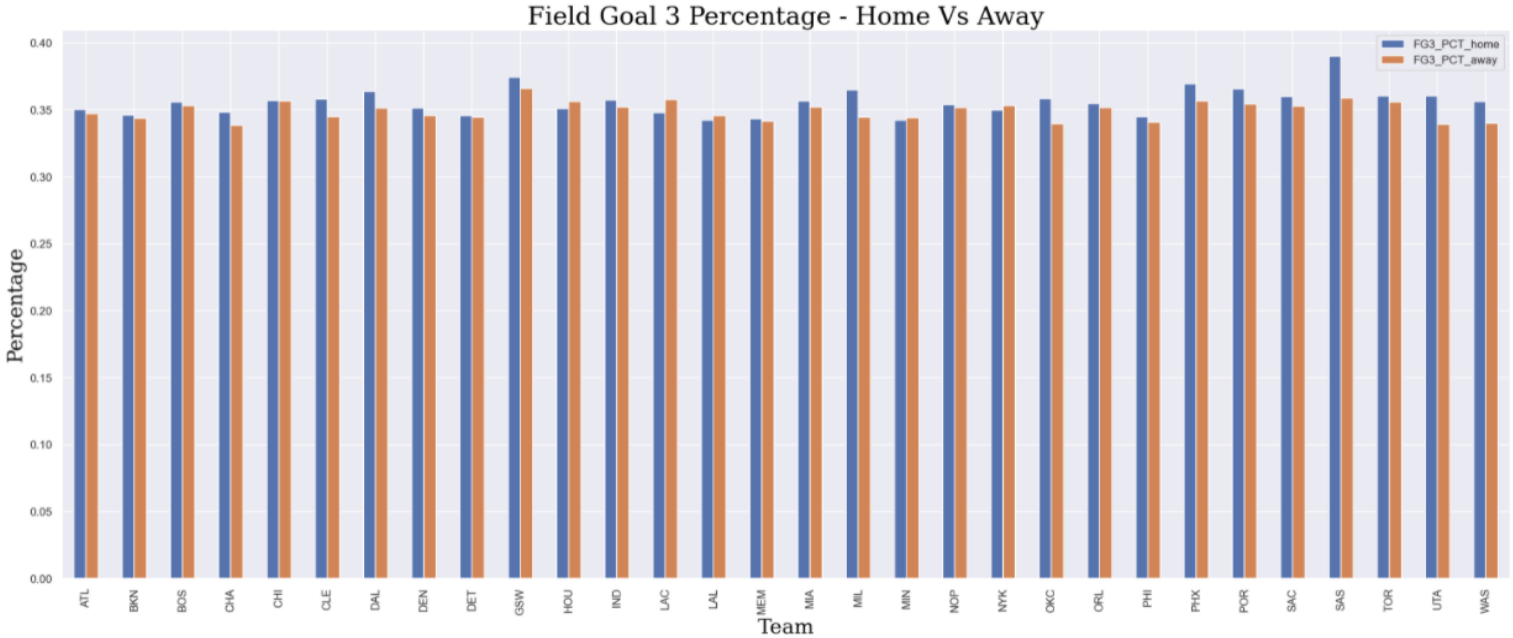


Figure 10.

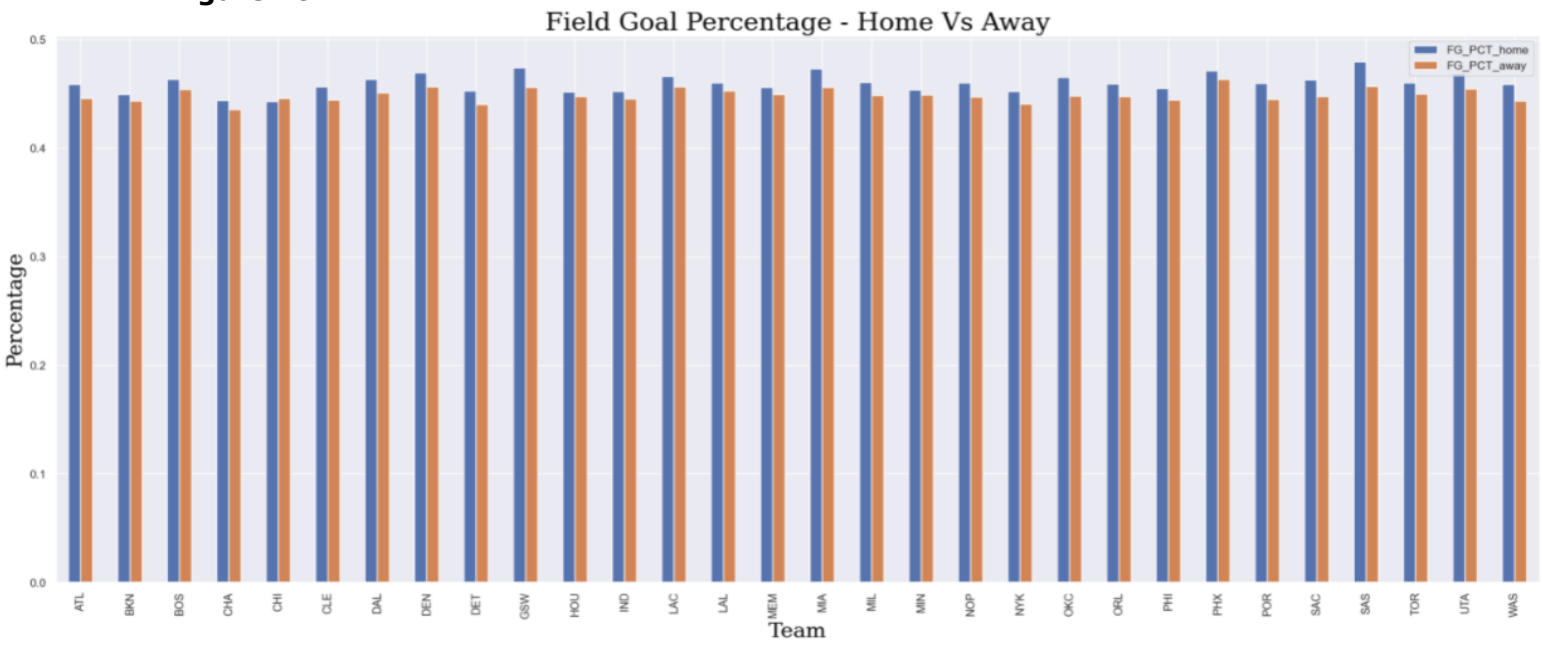


Figure 11.

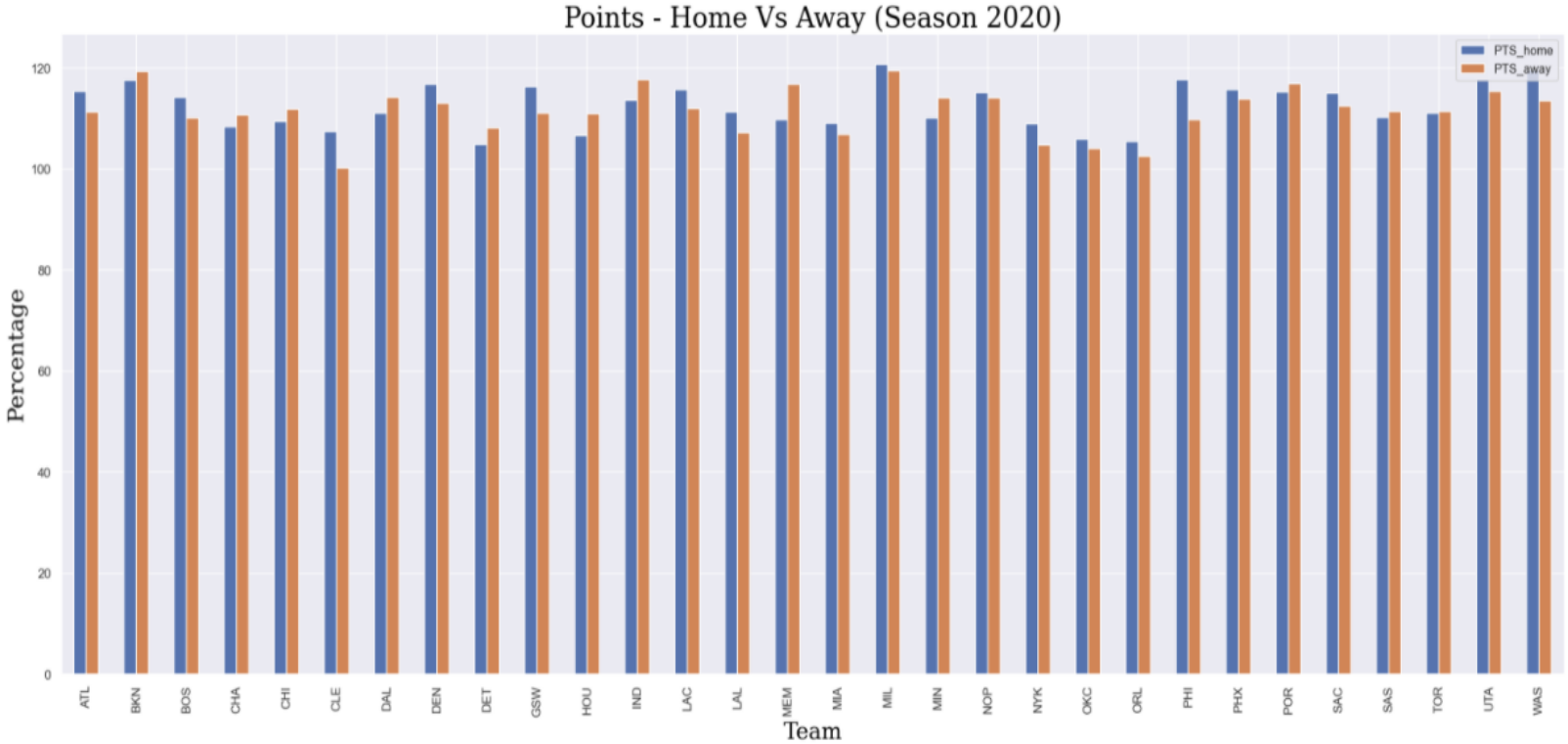


Figure 12.

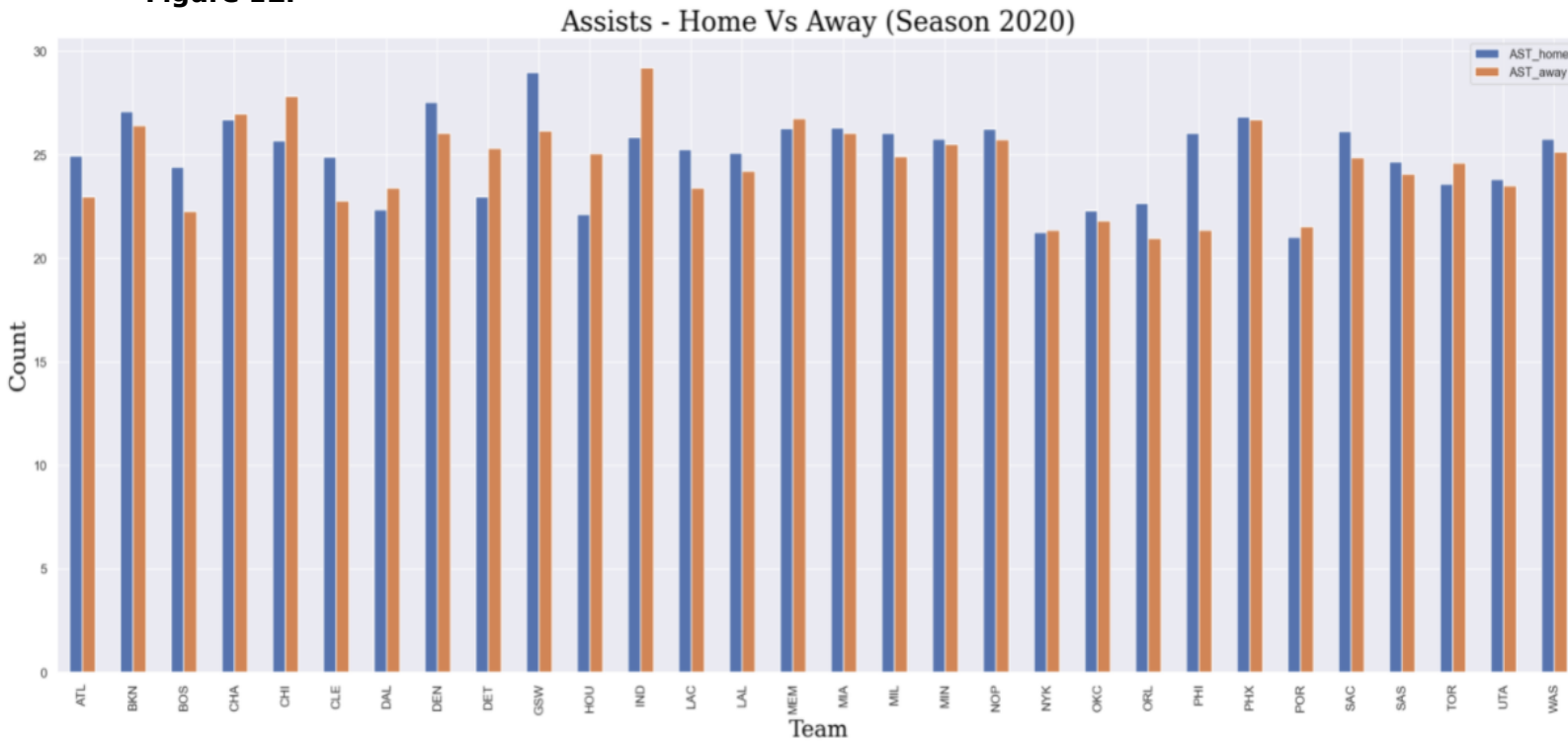


Figure 13.

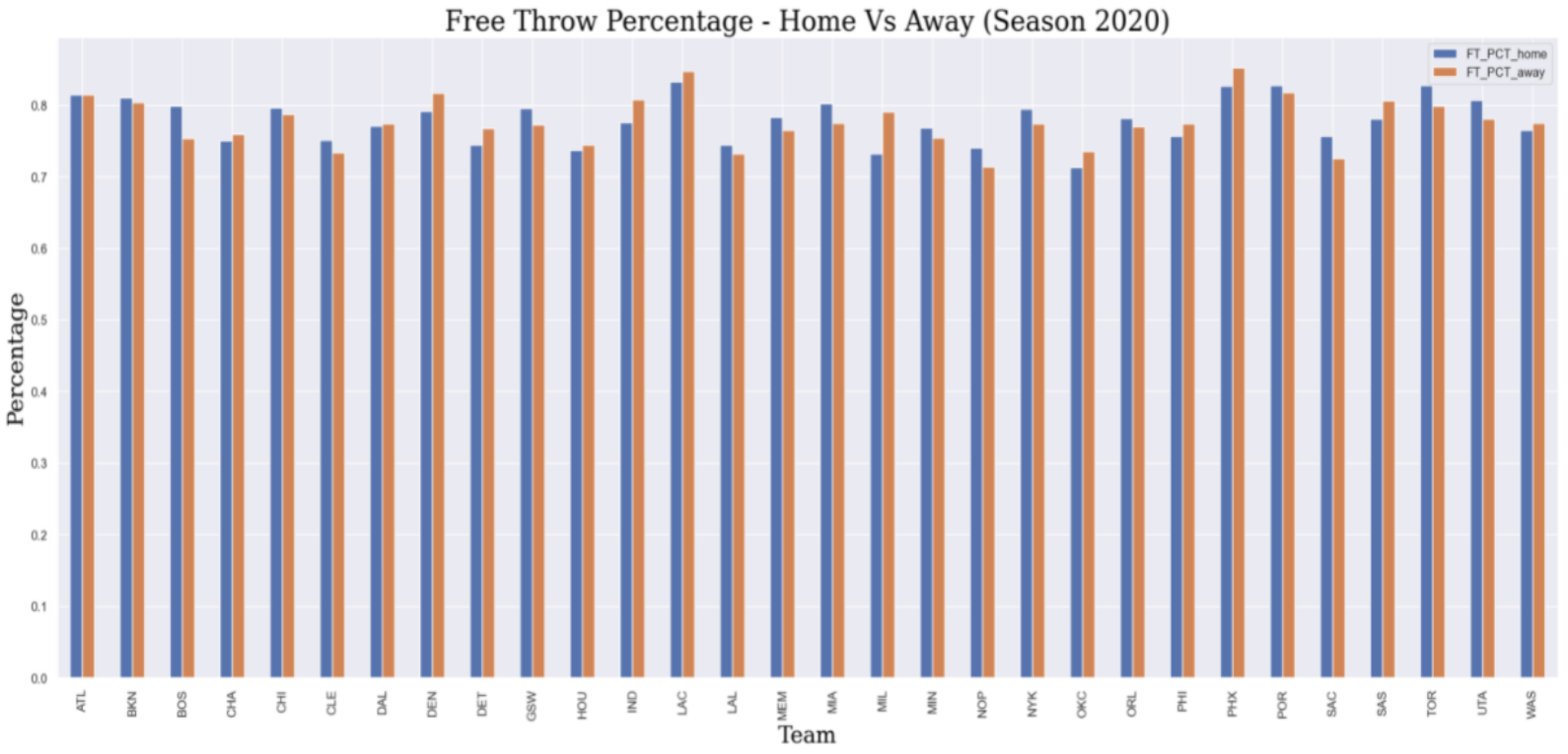


Figure 14.

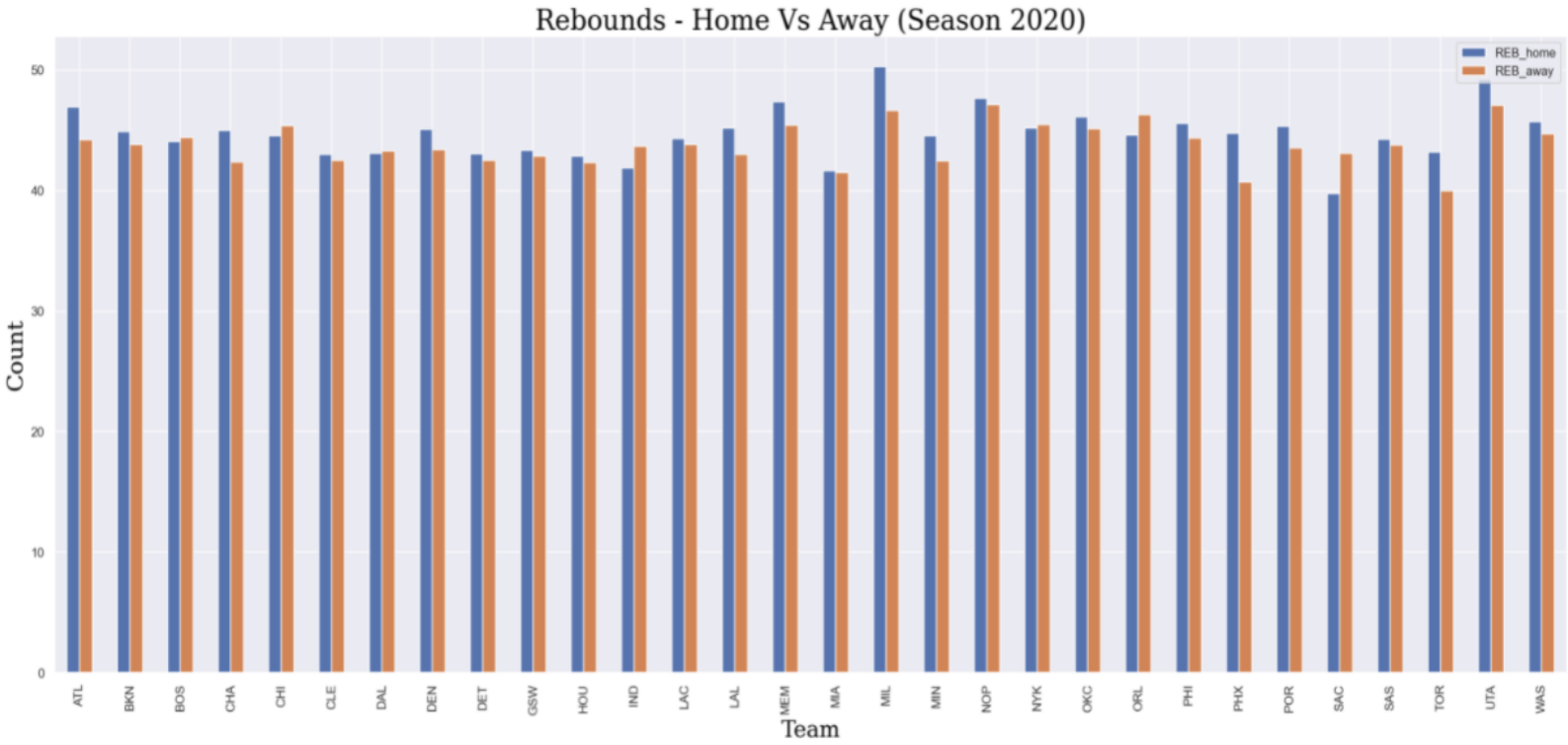


Figure 15.

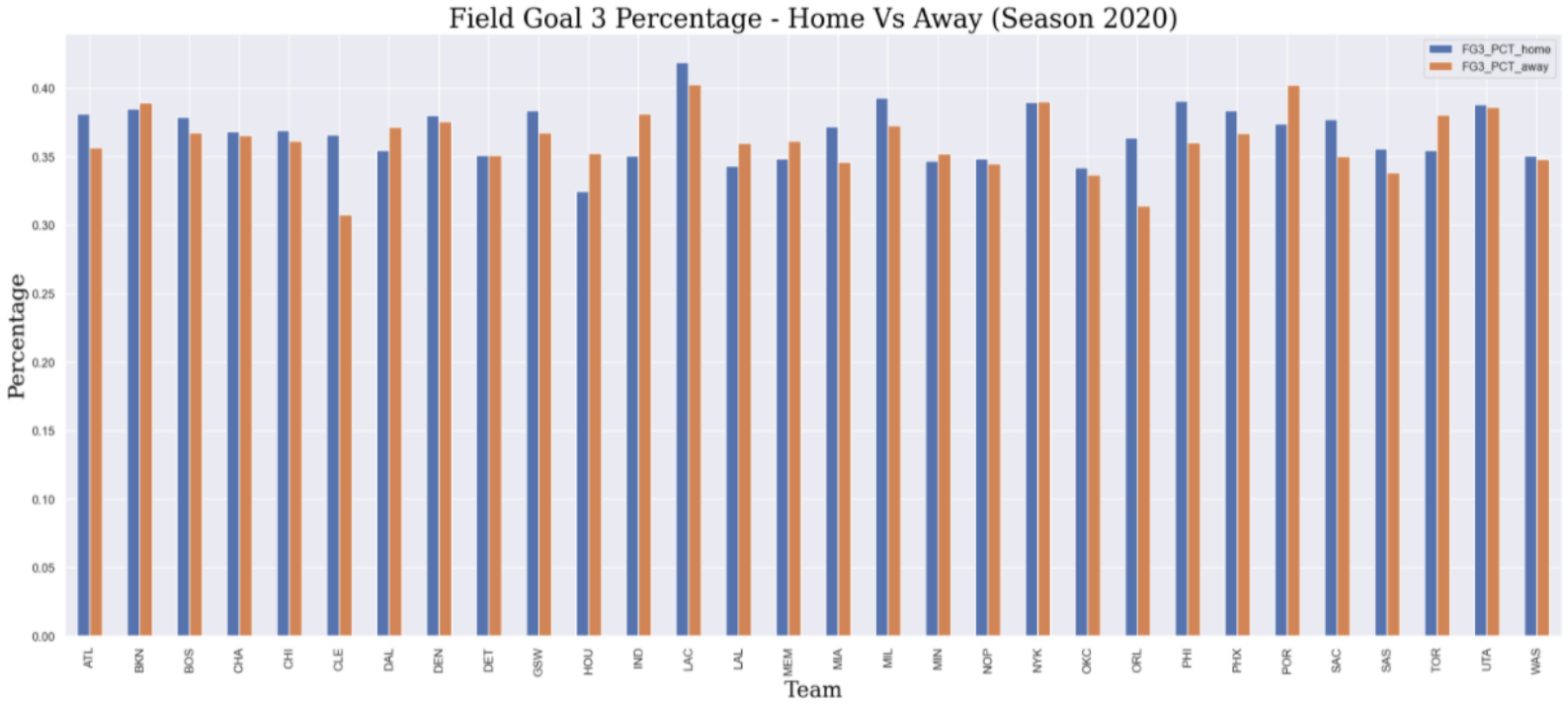


Figure 16.

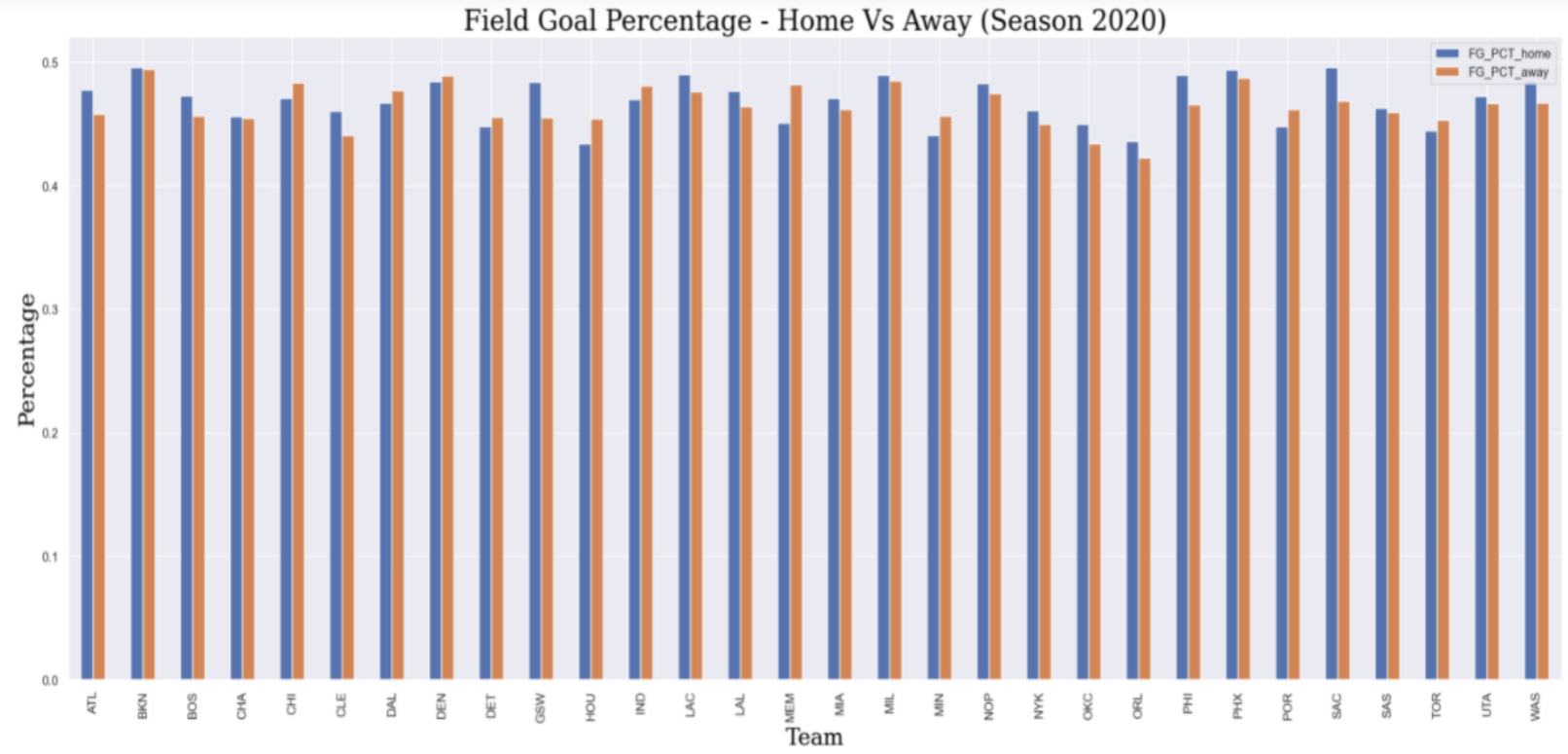


Figure 17.

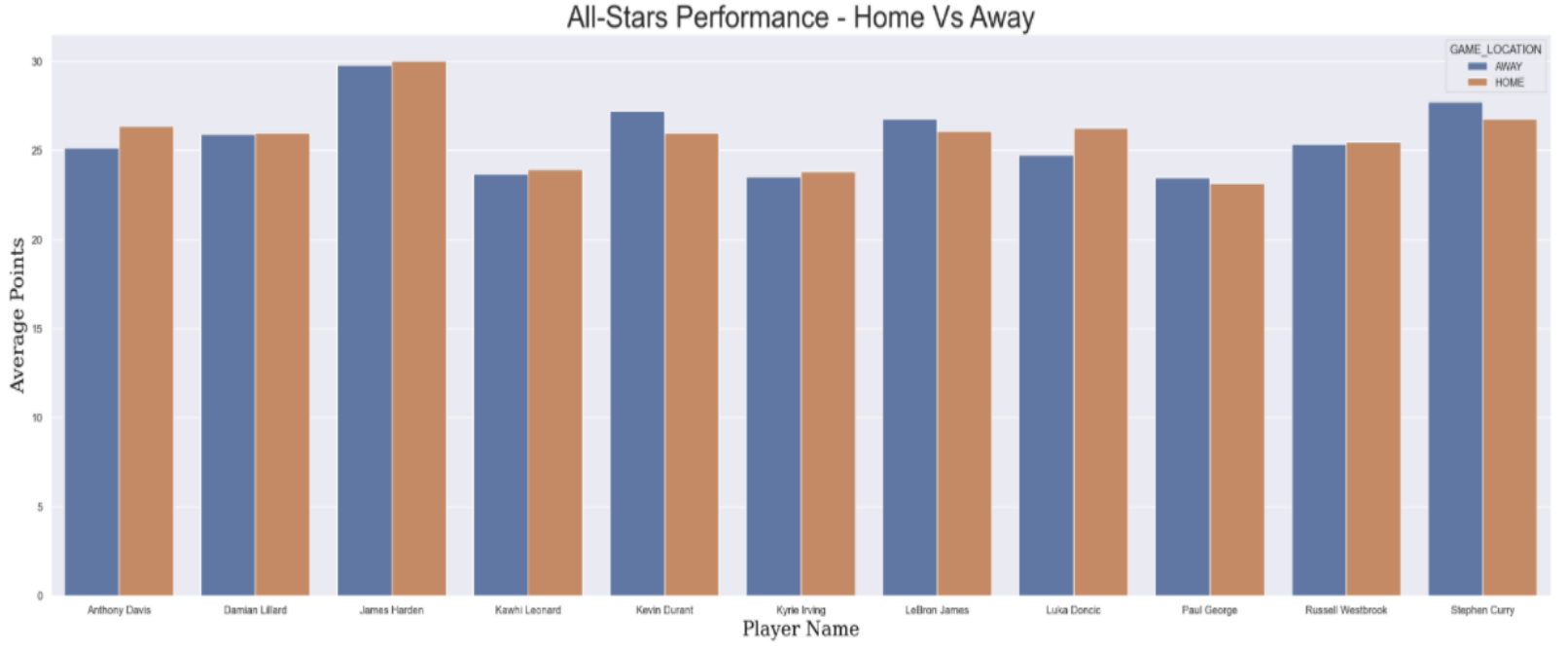


Figure 18.

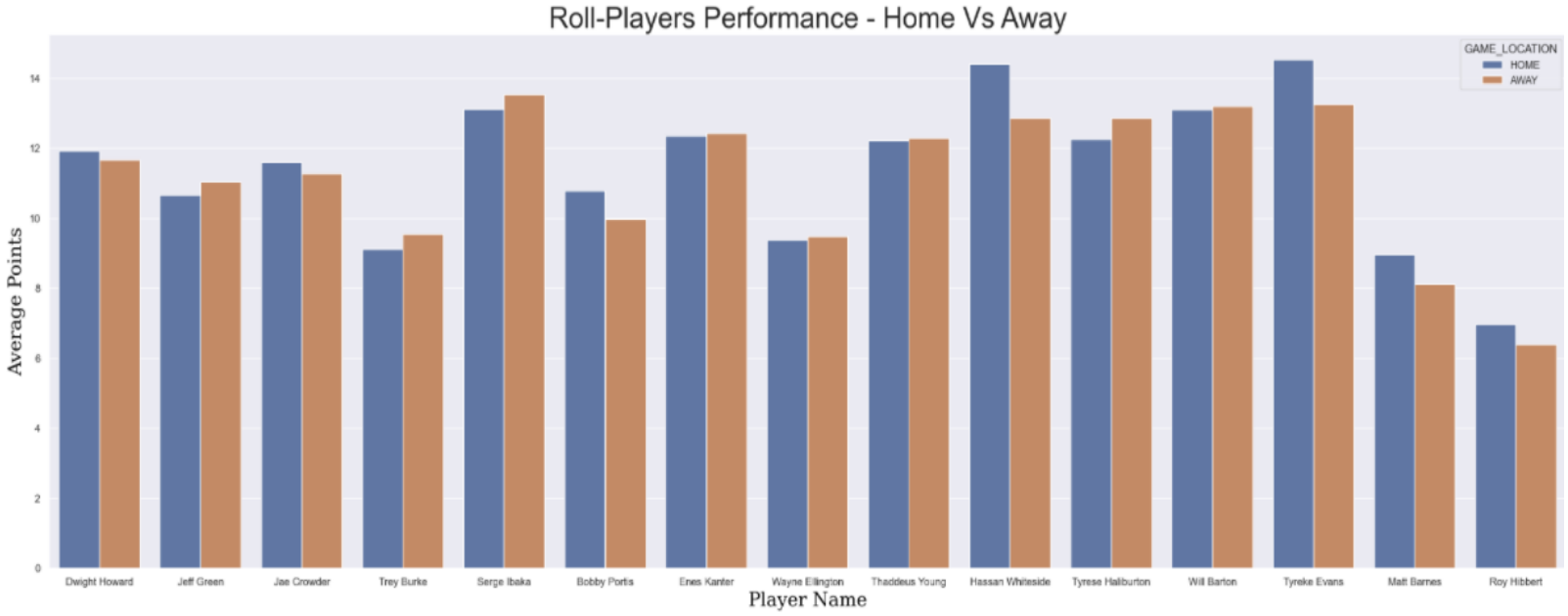


Figure 19.

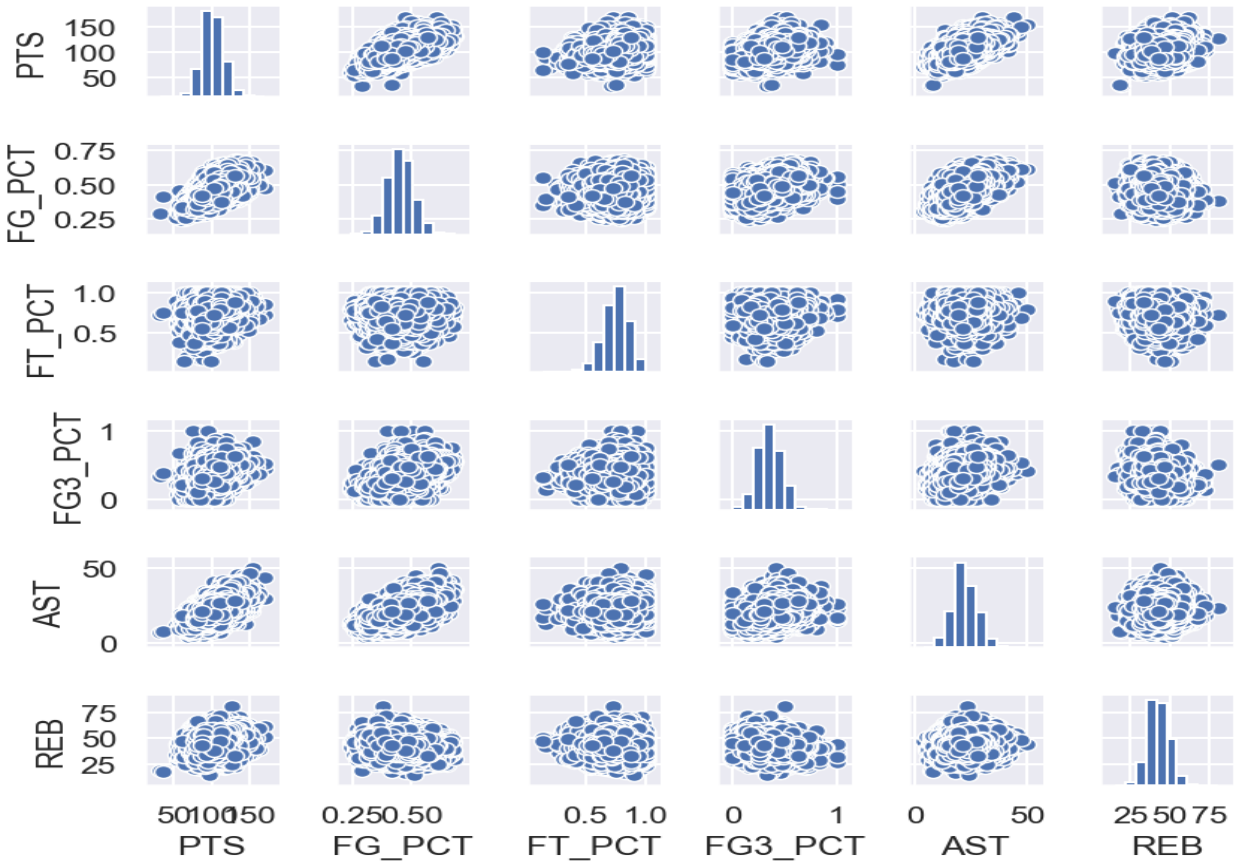


Figure 20.

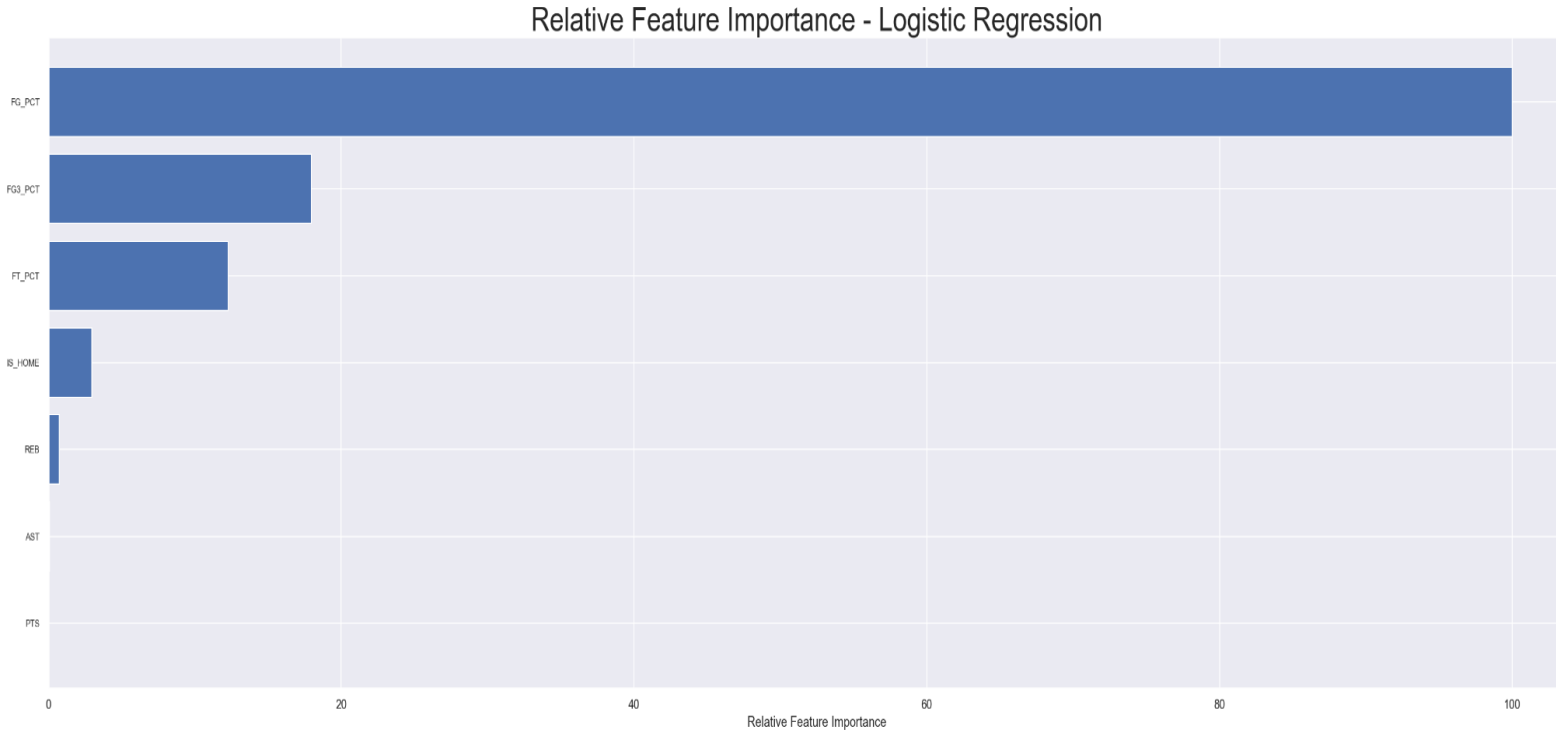


Figure 21.

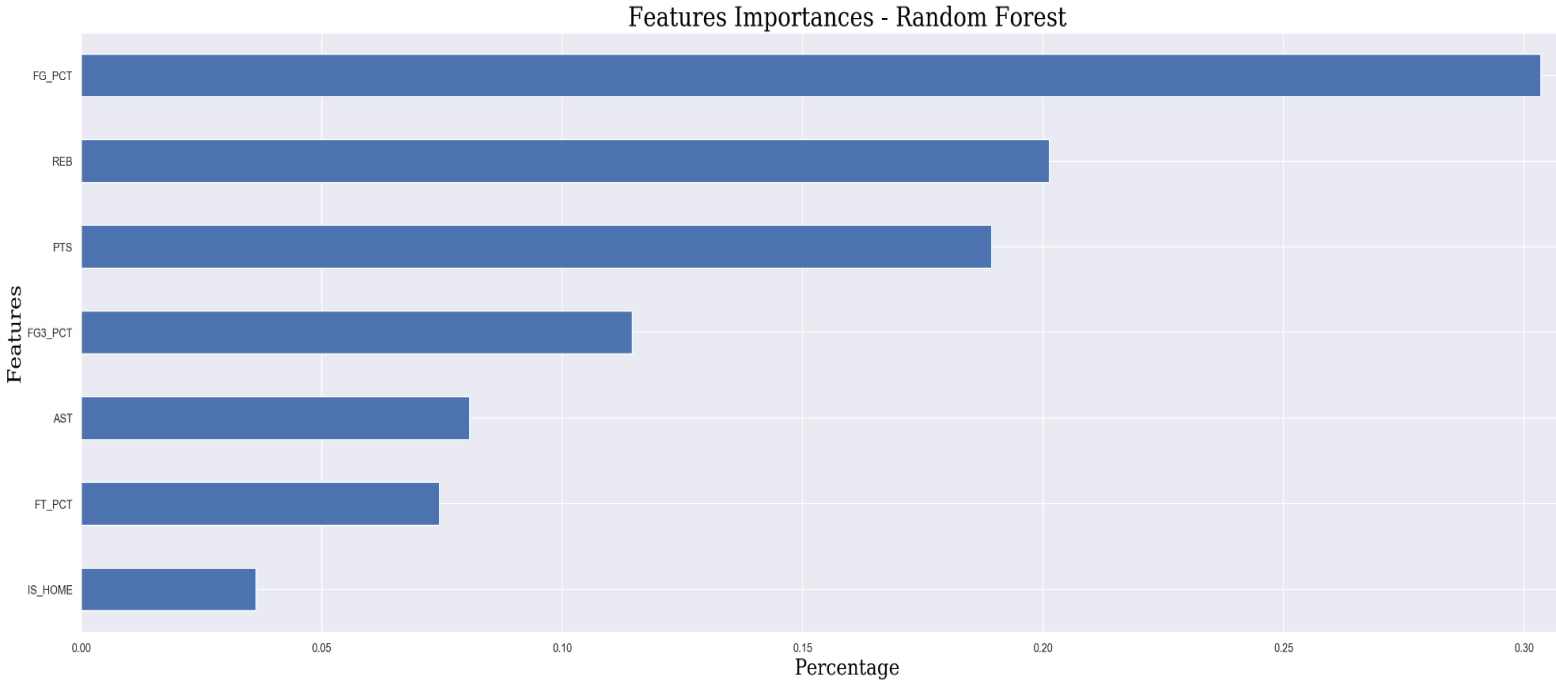
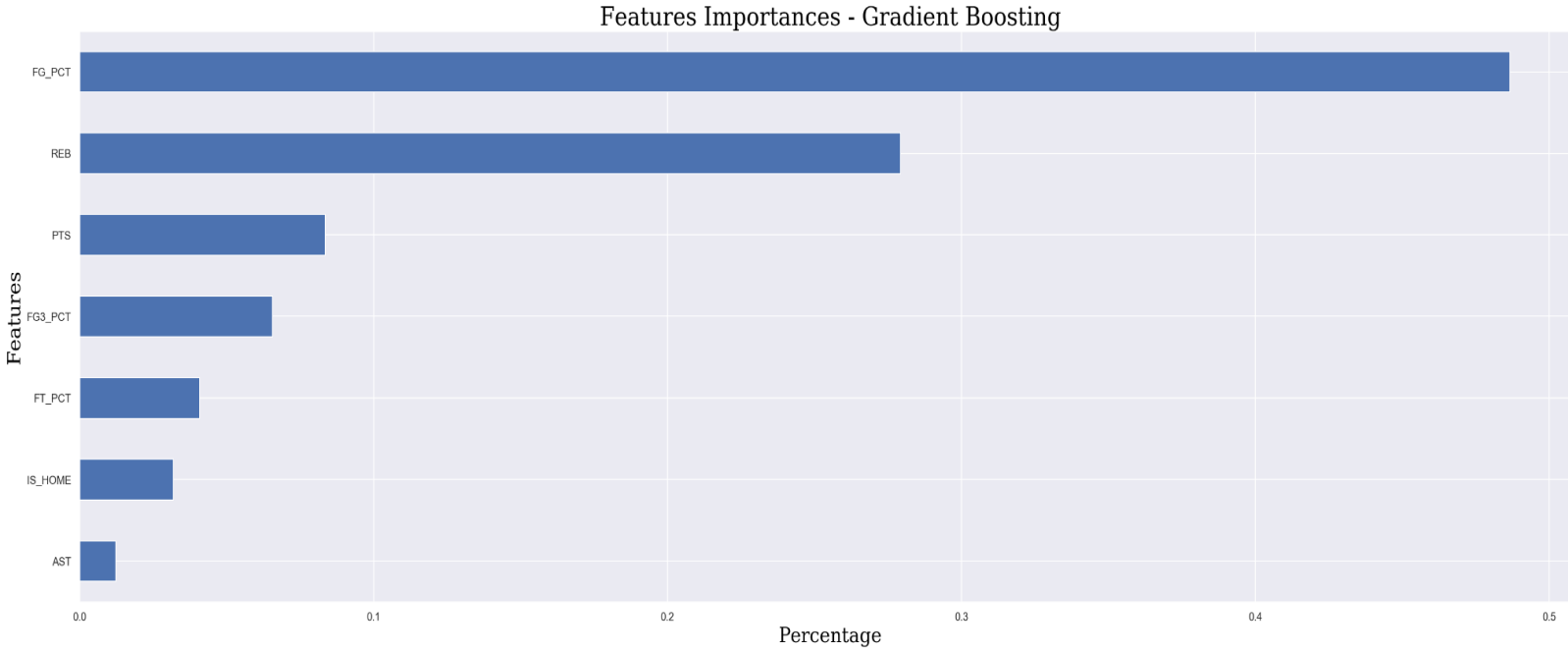


Figure 22.


Features Importances - Gradient Boosting
FG_PCT
REB
PTS
FG3_PCT
FT_PCT
IS_HOME
AST
Features
0.0
0.1
0.2
0.3
0.4
0.5
Percentage