Project Executive Summary: Is the NBA Game Fair?

Problem & Background

Over the years, the NBA has consistently been one of the most popular sports organizations and has attracted millions of fans. Because of the NBA's game characteristics, a referee's missed or wrong call has a massive effect on the game outcomes. With its vast fanbase, the fairness of referee calls has been widely discussed among fans and spectators worldwide. There are claims that the league favors star players for higher TV viewerships or higher volumes of ticket sales through biased referee calls.

This project aims to test the hypothesis that referee calls favor teams with super players. The definition of super players is controversial since each fan has their particular favorite player. In this project, we selected the top 10 players for jersey sales from the past nine seasons as our super players. Purchasing a particular player's jersey signifies the players' popularity. Also, for these 90 super players (10 players each season, 9 seasons in total), we will consider any team with a superstar player as a "super team" for that year.

Dataset (Preprocessing) & EDA

a) Dataset:

Since 2015, the NBA has released the last two-minute reports for every game within a three-point difference. The reports contain every correct and incorrect call and non-call in the last two minutes of those games. Our data source is from the "L2M" project on the GitHub website https://github.com/atlhawksfanatic/L2M, coordinating all the reports in the past 9 seasons in a single CSV file. This original dataset has 74,786 rows of data with 42 columns.

b) Data cleaning and preprocessing:

Firstly, we filtered out all the correct calls in the *'decision'* column, as we only need all the wrong calls to inspect if these calls favor superteams more. Secondly, we created new columns *'Favored teams'* and *'Non-favored teams'*, by deciding the favor and non-favor side in every call. Thirdly, we deleted all the rows with non value in the *'Favored teams'* column and some unrelated columns. Fourthly, by our definition of super teams and super players, we created a spreadsheet of all the super players and their teams in the past nine seasons. Finally, by merging these two data frames, we created our critical Boolean data column: *'Favored team is super team'* (Team favored in this wrong call is the team that has the super player).

c) Exploratory data analysis:

We tried to find the pattern in our dataset from different aspects. Here are some significant results. Most people intuitively believe referees favored super teams more. But surprisingly, in the games of a super team versus a non super team, only 48.88% of calls favored the super teams. It shows that referees actually slightly favored non-super teams more. To further inspect the effect of a specific superplayer, we calculated the average wrong call received by teams with different super players. Figure 1 shows that teams with two or three super players indeed got more favor calls than teams with only one or no super players. The reason that teams with no super player have higher average wrong call is that most games in our dataset are played by teams that have no super players on both sides. Then, we calculated the favor calls distribution inside super teams, and found out that super players got 14.04% of the calls, which is much higher than the super players' percentage (4.06%) in our dataset.

d) Data processing before the experiment:

We separated every call in our dataset into two rows based on the two teams involved in the call, then created a boolean column 'Favored', which will be our Y value in the experiment.

Experiment & Results

We built multiple regression models to evaluate the relationship between our *X (number of super players)* and *Y (whether the team is favored or not in the wrong call)*. All data of wrong calls we use to construct models only have one side that has super players. This helps us better detect whether teams with super players are more favored by referees in a wrong call than teams without super players. In the first regression model, we incorporate all variables that may affect Y, including *number of super players*, *regular/playoff season*, *home/away team*, and *Incorrect noncall/Incorrect call.* Figure 2 shows the result; since the p-value of all variables are larger than our significance level (alpha = 0.05), we can state that none of the above variables have statistical significance to the result of whether a team will be favored or not in a wrong call.

Next, we build the regression model that only includes one predictor, *X (number of super players)*, in our question. Given the large p-value shown in Figure 3, it indicates that the number of super players in a team has no impact on Y, which means referees will not favor a team in a wrong call just because it has more super players.

Since we find the original X and Y in our question have no relationship, we try to redefine our super team and use average home team attendance as our standard of super team and new *X* in the regression model. From the result in Figure 4, the p-value is much smaller than previous models, but it is still larger than 0.05. Moreover, the coefficient of *X* is close to 0, this shows that the average home team attendance also has no impact on the result of whether a team will be favored or not in the wrong call.

Conclusion

After our thorough analysis, we fail to reject the null hypothesis that the number of super in the team has no relationship with if a team is favored by referees in a wrong call. Our study demonstrates that referees do not exhibit bias toward a team based on its roster of super players or the volume of attendance.This insight provides a better understanding of officiating dynamics, fostering transparency, and fair competition within the NBA. If we have data from the previous 46 minutes, the answer of our

question may change. But based on our research of the last 2 minutes calls, the sport is fair and maintains complete integrity, ensuring equitable competition among all teams.

Appendix

Figure 1: Average favor call received by teams with different super players

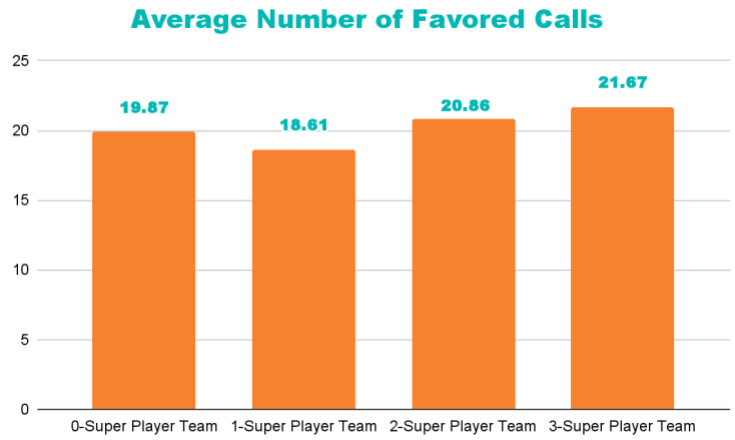

Figure 2. Regression Result for Model 1

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5020	0.014	34.738	0.000	0.474	0.530
Num_of_SP	-0.0159	0.010	-1.516	0.130	-0.036	0.005
playoff	0.0004	0.030	0.015	0.988	-0.057	0.058
home	0.0174	0.017	1.055	0.292	-0.015	0.050
decision	6.322e-05	0.026	0.002	0.998	-0.051	0.051

Omnibus:	12877.155	Durbin-Watson:	3.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	609.246
Skew:	-0.000	Prob(JB):	5.06e-133
Kurtosis:	1.004	Cond. No.	5.03

Figure 3. Regression Result for Model 2

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5112	0.011	47.078	0.000	0.490	0.532
Num_of_SP	-0.0166	0.010	-1.588	0.112	-0.037	0.004

Omnibus:	12868.256	Durbin-Watson:	3.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	609.987
Skew:	-0.000	Prob(JB):	3.49e-133
Kurtosis:	1.003	Cond. No.	2.17

Figure 4. Regression Result for Model 3

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5903	0.053	11.055	0.000	0.486	0.695
attendance	-5.049e-06	2.97e-06	-1.699	0.089	-1.09e-05	7.77e-07

Omnibus:	33318.494	Durbin-Watson:	3.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1610.744
Skew:	0.000	Prob(JB):	0.00
Kurtosis:	1.001	Cond. No.	1.89e+05