Research Question #2 Explanation & Results:

2. Which Zone Affects the Game More? In the paint, mid range, or behind the three point line?

Independent variables:

(O)CRFGA:  (opponent) field goal attempted (FGA) in the paint

(O)CRFGM: (opponent) field goal made (FGM) in the paint

(O)CRFG%: (opponent) field goal percentage (FG%) in the paint

(O)MRFGA: (opponent) field goal attempted in mid range

(O)MRFGM: (opponent) field goal made in mid range

(O)MRFG%: (opponent) field goal percentage in mid range

(O)3PA: (opponent) 3 pointers attempted

(O)3PM: (opponent) 3 pointers made

(O)3P%: (opponent) 3 pointers percentage

Dependent variable WIN%: Win percentage of a team, value is always between 0 and 1.

My method of answering this question is to have regression plots of all independent variables against WIN%. R2 are calculated and used to analyze which zone has the biggest impact on WIN%. For each independent variable, there are 6 regression plots, one per year. This helps me to figure out if there is a trend of increasing or decreasing importance of a zone. Since basketball is a game of both offense and defense, I will assess offensive stats and defensive stats with the same standard.

Regression plots help to visualize the data, but they are not a good tool to directly answer the question because analyzing by eye is not accurate enough. I have to calculate the R2 to make sure the analysis is precise. Here are the R2 of all variables from 2017 to 2022, rounded off to 3 decimal places.

Offense:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R2 | CRFGM | CRFGA | CRFG% | MRFGM | MRFGA | MRFG% | 3PM | 3PA | 3P% |
| 2017 | 0.003 | 0.106 | 0.399 | 0.000 | 0.014 | 0.163 | 0.200 | 0.075 | 0.405 |
| 2018 | 0.021 | 0.050 | 0.384 | 0.000 | 0.024 | 0.397 | 0.117 | 0.058 | 0.238 |
| 2019 | 0.000 | 0.063 | 0.284 | 0.007 | 0.001 | 0.273 | 0.238 | 0.098 | 0.293 |
| 2020 | 0.002 | 0.127 | 0.476 | 0.005 | 0.000 | 0.100 | 0.093 | 0.025 | 0.199 |
| 2021 | 0.028 | 0.182 | 0.267 | 0.184 | 0.105 | 0.449 | 0.158 | 0.013 | 0.619 |
| 2022 | 0.000 | 0.049 | 0.250 | 0.110 | 0.091 | 0.179 | 0.052 | 0.004 | 0.480 |
| mean | 0.009 | 0.096 | 0.343 | 0.051 | 0.039 | 0.260 | 0.143 | 0.046 | 0.372 |

Defense:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| R2 | OCRFGM | OCRFGA | OCRFG% | OMRFGM | OMRFGA | OMRFG% | O3PM | O3PA | O3P% |
| 2017 | 0.013 | 0.007 | 0.123 | 0.028 | 0.005 | 0.092 | 0.125 | 0.010 | 0.401 |
| 2018 | 0.009 | 0.123 | 0.149 | 0.061 | 0.006 | 0.241 | 0.164 | 0.059 | 0.246 |
| 2019 | 0.133 | 0.002 | 0.421 | 0.032 | 0.108 | 0.065 | 0.123 | 0.014 | 0.388 |
| 2020 | 0.429 | 0.183 | 0.402 | 0.000 | 0.069 | 0.304 | 0.000 | 0.168 | 0.504 |
| 2021 | 0.157 | 0.043 | 0.201 | 0.043 | 0.154 | 0.063 | 0.065 | 0.003 | 0.277 |
| 2022 | 0.281 | 0.098 | 0.235 | 0.017 | 0.061 | 0.077 | 0.092 | 0.001 | 0.436 |
| mean | 0.170 | 0.076 | 0.255 | 0.030 | 0.067 | 0.140 | 0.095 | 0.043 | 0.375 |

One thing that I can firmly claim is that any variable which is related to the FGA has minimal or no correlation to WIN%, thus they are useless for answering my question. None of these variables has an R2 over 0.2 in any year and a mean R2 over 0.1. More shots attempted means more opponent shots attempted. That only speeds up the game and doesn’t truly affect the WIN%.

Comparing the FGM variables on offense, 3 pointers have the highest mean R2 and the paint has the lowest mean R2. Since the R2 of CRFGM against WIN% is always below 0.03 in all 6 years, that means CRFGM is completely irrelevant to WIN%. MRFGM is not any better. The only reason it has a higher mean R2 is the sudden rise in 2021. It is a coincidence because it dropped again in 2022. For 3PM, it has the highest mean R2. I observe that there is a subtle downward trend, as the R2 of the first three years is higher than the last three years, 0.185 vs 0.101 to be exact. So for the FGM on offense, the area behind the 3 point line has the biggest impact on WIN%, but the impact is slowly dropping. Mid range and the paint has no impact on WIN%.

On the defensive side, I compare the R2 of opponent FGM variables. OCRFGM has the highest R2 against WIN%, followed by O3PM and OMRFGM.

OMRFGM is completely irrelevant to WIN%, since the max R2 is 0.061 in 2018 and the mean R2 is 0.03. For O3PM, things are a bit more interesting even though it only has a 0.095 mean R2. The first three years all have R2 above 0.1 and the last three years all have R2 below 0.1. So, O3PM is having less impact in the most recent years. For OCRFGM, it has a mean R2 of 0.17. The surprising part is that the first three years have a mean R2 of 0.052, compared to 0.289 in the last three years. There is a clear difference between them. That means OCRFGM not only has the strongest correlation to WIN% out of the three variables, but it also affects the games more and more. So on the defensive side, the opponent FGM in the paint is the most decisive to a team’s success, and outside the paint, it doesn’t matter.

By analyzing variables related to the FGM, which measures the quantity of shots, I can conclude that mid range affects WIN% the least. The area behind the arc affects WIN% the most on offense, but the impact is declining. The paint is the most important and becoming more and more important to a team on defense. So considering the trend, I can say that the paint affects the game the most, followed by area behind the 3 point line. Mid range is the last.

But without a doubt, the shooting efficiency affects the game way more than quantity. The mean of the R2 of percentage related variables against WIN% is 0.291, while it is merely 0.083. So when I have my final conclusion, percentage-related variables will have more influence on the answer than FGM related variables.

On offense, MRFG% has the lowest mean R2 against WIN%. 3P% has a mean R2 a bit higher than CRFG%. But the R2 of 3P% fluctuates more than CRFG%. This can be proved by comparing the standard deviation. 3P% has a standard deviation of 0.146 and CRFG% has a standard deviation of 0.082. Because of that, I will examine these R2 year by year.

In 2017: 3P% (0.405) > CRFG% (0.399) > MRFG% (0.163)

In 2018: MRFG% (0.397) > CRFG% (0.384) > 3P% (0.238)

In 2019: 3P% (0.293) > CRFG% (0.284) > MRFG% (0.273)

In 2020: CRFG% (0.476) > 3P% (0.199) > MRFG% (0.100)

In 2021: 3P% (0.619) > MRFG% (0.449) > CRFG% (0.267)

In 2022: 3P% (0.480) > CRFG% (0.250) > MRFG% (0.179)

Despite the high fluctuation of the R2 of 3P% against WIN%, it still has the strongest correlation to WIN% in 4 out of 6 years. MRFG% has the weakest correlation to WIN% in 4 out of 6 years. Therefore I conclude most of the time on offense, 3P% is the most important and MRFG% is the least important to the game.

Moving on to defense, the mean R2 of O3P% is higher than OCRFG%, and OCRFG% is higher than OMRFG%. The differences between them are not tiny, so there is no need to assess the R2 year by year. Limiting O3P% is the most crucial for a team to win a game, OMRFG% is the least.

In conclusion, considering shooting efficiency on both offense and defense, which I mentioned earlier weighs heavier than the quantity of shots, the area behind the 3 point line affects a game the most, the paint is the second, and mid range is the third. The reason behind this answer is simple. Shooting in the paint has the merit of a shorter distance from the basket. Shooting 3 pointers have the merit of earning 1 extra point. Mid range doesn’t have either one of the merits. I also find out that there is no obvious trend of a zone becoming more/less influential on a game.

After figuring out 3P%, O3P%, CR%, and OCR% decide the victor of a game quite noticeably, this means these statistics can predict the WIN% of a team in a year quite accurately. This can be done effortlessly with machine learning and building models based on these 5 variables. I split the training dataset and testing dataset in a 4:1 ratio. I found out that when the max depth of the decision tree is about 5, it optimizes the accuracy of the model. The mean absolute error of predicting WIN% using the testing dataset fluctuates between 0.06 and 0.1. The fluctuation is due to the random split of the dataset. I don’t think this model is super accurate in predicting WIN%, but it can certainly find out which team has an abysmal season and which team has an amazing season. That means these shooting stats certainly affect the game.