Group Members:
Nathan Dennis
Kevin Wang
Cheeson Lau

What Makes NBA Teams Successful?

Research Questions and Results:

1. How does Offensive and Defensive Rating Affect Win Percentage?

This research question was broken into a couple sub topics, where the first regards how offensive and defensive rating alone affect win percentage, while also how they would affect win percentage combined. For this portion of my question I found out that both factors, individually, affect winning percentage the same amount. When creating a model and graph to represent both factors compared to win percent, there was no obvious difference through the visualization, as both factors had a positive correlation. (Disclaimer: since a lower defensive rating means a better overall defense, the lower the defensive rating the better the winning percentage.) However, the combined model with both defensive and offensive rating to predict win percent did a very excellent job, showing that you don't necessarily need only one factor to succeed, as many claim, but rather both. This research question also compared the difference between bottom half teams in both offensive and defensive rating to the teams who were above average in both categories. Turns out, both the above average defensive (Lower defensive rating) and offensive rating teams had a significantly higher win percentage than the below average teams, as one may expect

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

This question explores which zone of the court has a greater impact on overall games won using machine learning and regression plots. I examine both offensive side and defensive side. Quantity (field goal made and attempt) and efficiency (field goal %) are considered to answer the question. By calculating the R^2 of each independent variable against the win%, I found out the following facts. First, field goal attempt in any zone, in both offense and defense has no impact on win%. Second, for field goal made also has little impact on win%, with the exception of 3PM and OCRFGM in some years. In contrast, field goal percentage in all zones has more impact on win%. So considering the R^2 of field goal percentage in all zones on both offense and defense, I find out that a game is the most decisive behind the three point line, followed by in the paint, and mid range is the last. This is true in most years. In addition to these facts, I try to input shooting statistics into a model and use it to predict the win% of teams. It turns out that the model predicts less accurately than I expected.

3. How can +/- be predicted using generic data including rebounds, steals, turnovers, blocks and assists for both the home and away team?

   This question does not aim to find the machine learning models that produce the smallest error because the way that the variables are chosen and the limitations of the dataset itself already puts so much error into the machine learning model. Instead, three regression models: decision tree, linear and polynomial regression model are used to compare and contrast to see which model produces the smallest error. The linear regression model turned out to be the best because the underlying algorithm works the best considering the limitation rooted in the data itself because the linear regression model simplifies the problem which also happens to reduce the effect of systemic errors. And the other two models are justified (with details in analysis part) why they do not work as well as the linear model. In the end, the linear regression model was further examined alone using the shap library which was used to compute the weighting of each independent variable on the dependent, serving as the true value (though not accurate either). Then it was compared to the slopes of each independent variable plotting against the dependent variable, and we could see a correlation with reasonable errors. Then we finally arrive at the conclusion that +/- can be more accurately predicted by linear regression model than by decision tree model and polynomial degree model. Though this research question has lots of sources of error, many of them are addressed in the analysis and the limitation part, which make the conclusion justifiable.

Motivation:

Many fans and analysts will agree that the NBA is changing rapidly, it is getting much harder to predict many stats about NBA teams. Take this year, the Lakers and Nets were both widely acknowledged as the best 2 teams in the league, NBA Finals favorites. However, the Lakers couldn’t even make the playoffs, while the Nets fell short in the first round as the 7th seed. Many team statistics are hard to predict as well, as various factors influence how many games a team will win over the course of the season. It would be interesting though, to any NBA scouts or staff members, what factors may influence winning, as a team's goal is obviously to win it all. Using our research and conclusions, those staff members can use our answers in order to figure out what to improve on their team to achieve a certain goal. For example, if it turns out that offensive rating affects winning percentage more than defensive rating, or vice versa, a general manager may decide to target more offensively oriented players within the draft or free agency. Another example is how many fans claim defense is a lost art in the NBA, how a team doesn’t need to be a good defensive team in order to win a championship anymore as offensive output can outweigh this. Possibly disproving this claim could show fans how valuable defense still is today.

Dataset:

We plan to use data from NBA.com, specifically the traditional and advanced stats for each NBA team over the course of 2017-2022. For the third research question, we will use additional stats which include opponent stats and playoff stats, from 2017-2021.

NBA stats.csv
Opponent_stats.csv
playoff_stats.csv

Method:

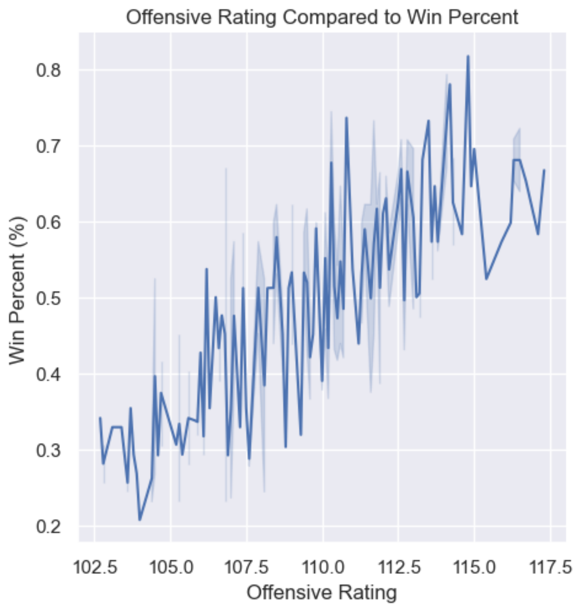
1. The first research question will primarily involve the variables Offensive Rating, Defensive Rating, and Winning Percentage, which appear in multiple data sets. To begin, I will use both Offensive and Defensive Rating and compare each to the overall winning percentage, using multiple graphs and models. First, I separated the data into teams with a high (above average) and low (below average) offensive and defensive rating. I then computed the win percentage for teams within these filtered data sets, visualizing if there was any difference in overall win percent. To verify my results, I created a confidence interval and constructed a p-value, testing my null hypothesis. I then created a test/train split in order to create a model to predict win percent, one based on only offensive rating, another only on defensive rating, and one using both ratings to predict win percent. I will calculate the RMSE and adjusted R^2 for both of these models to assess their validity.
2. The second research question will involve the offense related independent variables, namely in the paint CRFGA, CRFGM and CRFG%; mid range MRFGA, MRFGM and MR%; and behind the 3 point line 3PA, 3PM and 3P%. It also involves the defense related independent variables, which are the above variables with O in front of them (like O3P%). The dependent variable is WIN%. I will generate regression plots of all independent variables against WIN% and calculate the R^2 for comparison. I will also create a model and split the dataset to test set and train set. Then, I will compute the MAE to determine whether these shooting stats can truly affect the game.
3. The third research question focuses on how +/-, which is the net number of points a team wins in a game, can be impacted by some of the most basic stats in basketball including both home and away teams' rebounds, steals, turnovers, blocks and assists. To do this, I first trained a decision tree regressor model using regular season stats from 2016 to 2021 under the above category for each team as independent variables and the corresponding +/- value as dependent variable. Then I used the model to predict the playoff +/- value given the independent variables and then assess the accuracy of the model by calculating mean absolute error, mean squared error and R^2. However, the error was huge so then I created a line regression model and followed the same process, which resulted in a much

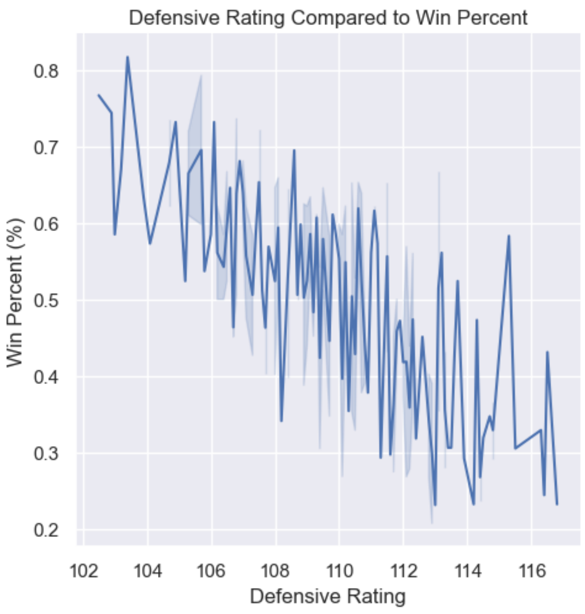
smaller error. I tried to further improve on the model by creating a polynomial regression model, with degrees of 2, 3, 4, 5. However, the errors were a lot bigger than that from the linear regression model. To verify, I tried to plot a linear regression line for each independent variable against the dependent variable and then compare the slope of the resultant graph to the weighting of that independent variable on the dependent variable.

Results:

Question 1: How does Offensive and Defensive Rating Affect Win Percentage?

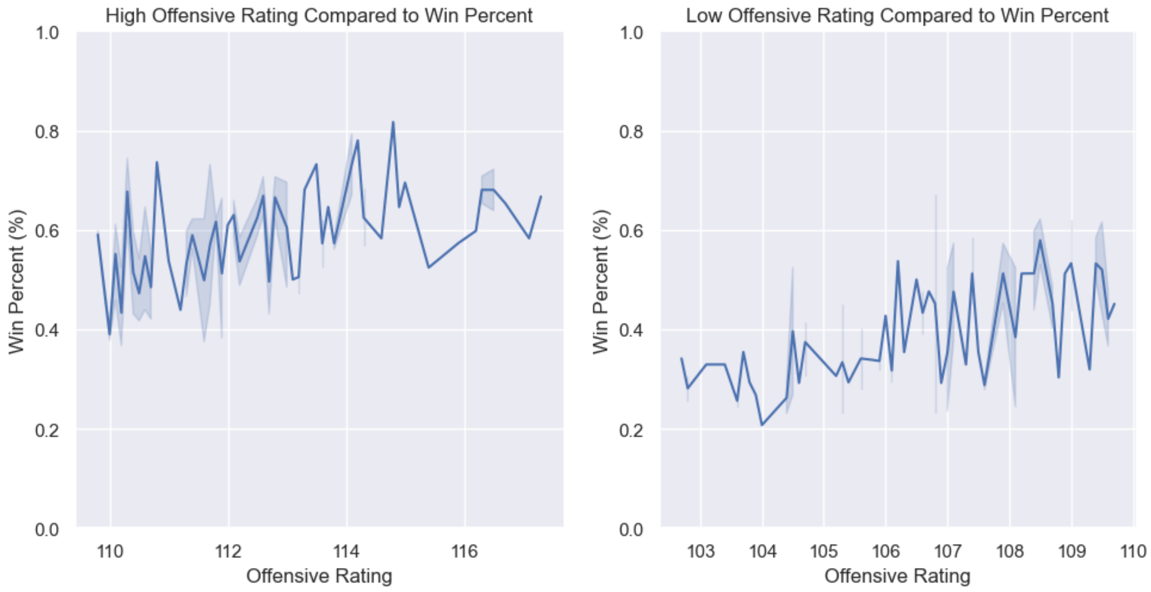
To begin this question, we constructed basic graphs for both the offensive rating for teams compared to the defensive rating for teams. Offensive rating is a basketball statistic which measures how the offense of a team performs, and at simplest terms, defines how many points a team scores per 100 possessions. The higher the offensive rating, the better the offense. Defensive rating however is a basketball statistic which measures how many points a team allows per 100 possessions and in contrast to offensive rating, the lower the defensive rating, the better the defense. The two graphs are depicted below comparing both ratings to win percent:





From these two graphs, there is no obvious answer to our question, as it looks like both offensive and defensive rating heavily impact win percent, both statistics improving winning percentage as a teams offensive/defensive rating improves.

Since these two visualizations don’t give us much information, we can split up these two statistics into two categories. More specifically, we will calculate the average offensive and defensive rating, then separate every team’s rating to either an above or below average rating for each statistic. We will now depict teams with an above average offensive rating as teams with a “high” offensive rating, and teams with a below average offensive rating as a “low” offensive rating, and the same for defensive rating. With these statistics, we can visualize which statistics impact winning percentage more, starting with offensive rating. Here are the two plots depicting a high offensive and low offensive rating compared to winning percentage:

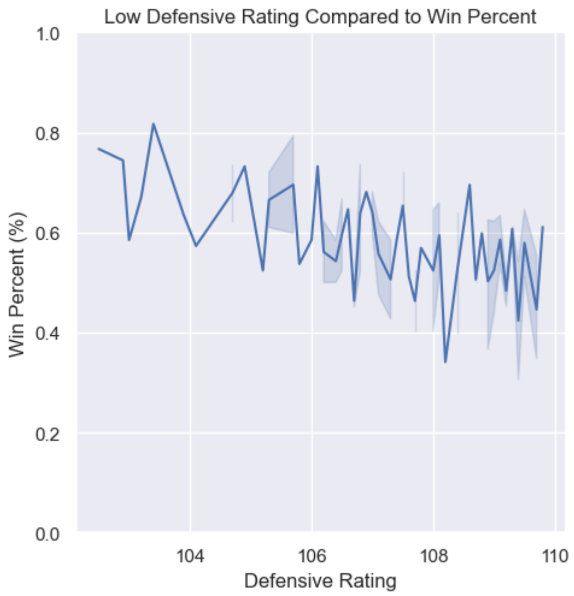


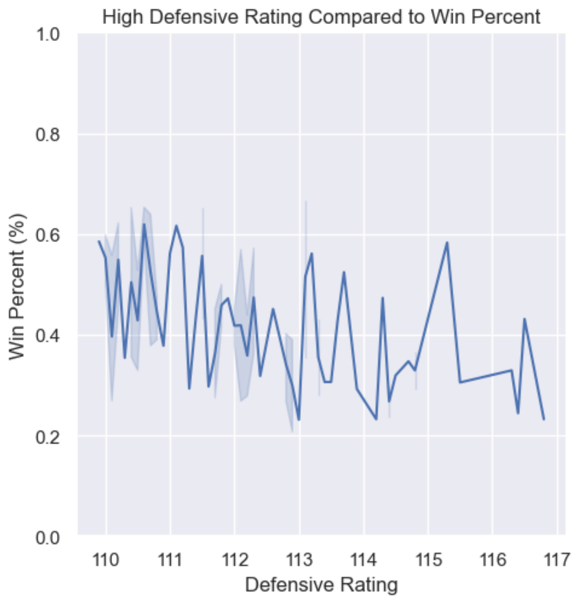
Comparing these two plots, there is an obvious decrease in Win percentage with the lower offensive ratings, especially towards the left end of the graph representing teams with the lowest offensive ratings, all the lowest win percentages seem to be in that area. Calculating some statistics, the high offensive rating average win percent turned out to be 0.579, while for the low offensive rating it was only 0.407. We can assume that higher offensive ratings improves overall win percentage. But, to further conclude that a higher offensive rating improves win percentage, we can calculate a 95% confidence interval then calculate a p-value.

First, we can actually calculate the mean win percentage for teams with a high and low offensive rating. As mentioned before, the win percent for teams with an high offensive rating was approximately 0.579, while the average win percent for teams with a low offensive rating is 0.407. We can obviously see some difference within these values, but to test for statistical significance we can move on to the confidence interval. Now, we must create our null hypothesis, that there exists no difference in mean win percentage between teams with an above average offensive rating, and those with a below average offensive rating. The confidence interval for the difference in mean difference in win percent between high versus low offensive ratings is (0.138, 0.205). We can say with 95% confidence that the true mean difference in win percent between teams with a “high” offensive rating and “low” offensive rating is between 0.138 and 0.205. Since this confidence interval does not include the value 0, it does suggest there exists a difference between these two categories.

To analyze further, we can test our null hypothesis with a null distribution, calculating the p-value. As a reminder, our null hypothesis is that there exists no difference in win percent between teams with a high versus low offensive rating, with an alternative hypothesis that teams with a high offensive rating achieve a higher winning percentage than teams with a low offensive rating. Using the simulated null distribution, we calculated a p-value of approximately 0, and since this value is below the conventional threshold of 0.05, we have enough evidence to reject the null hypothesis in favor of the alternative, that a higher offensive rating corresponds to a higher winning percentage compared to lower offensive ratings.

Now, we move on to the high (above average) and low (below average) defensive ratings for each team. Here is a plot depicting these two variables compared to win percentage. As a reminder, a high defensive rating equates to a "worse" defense, while a low defensive rating corresponds to a better defense.





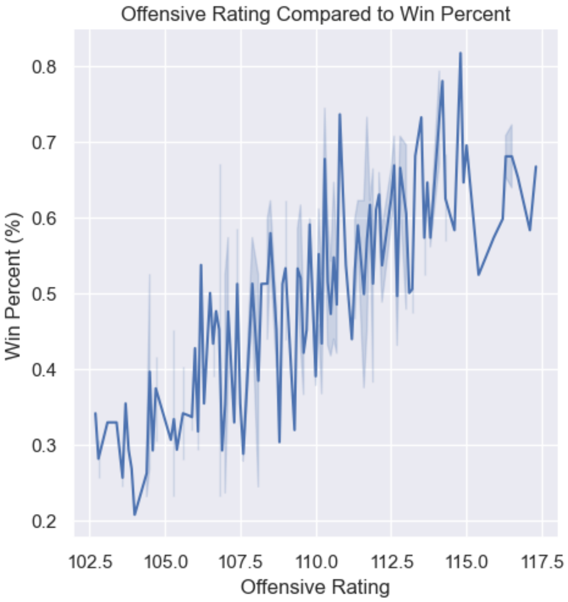
Through these graphs we can see some sort of correlation between the two variables, as it seems like teams with a lower defensive rating tend to have a higher win percent, as expected. In order to verify these results we can construct a 95% confidence interval and calculate a p-value for these statistics.

We now calculate the mean win percent for each category, as teams with a high defensive ratings have an average win percent of 0.429, while teams with a low defensive rating have an average win percent of 0.573. Through these stats we can see teams with a low defensive rating, and better defense, tend to have a higher win percentage. Now constructing a confidence interval, we calculate the difference in mean win percent between teams with a high and low defensive rating, which turns out to be (-0.179, -0.107). We are 95% confident that the mean difference in defensive rating between high and low defensive rated teams falls between the values -0.179 and -0.107. Since this confidence interval does not include the value 0, this suggests there exists some difference between these values.

We can take this a step further however and calculate a p-value in order to test our null hypothesis, that there exists no difference in win percentage in teams with a high or low defensive rating, with an alternative hypothesis of teams with a low defensive rating tend to have a higher win percentage. Using the simulated null distribution, we calculate a p-value of nearly 0, which is below the conventional threshold of 0.05, so we can safely reject our null hypothesis in favor of the alternative.

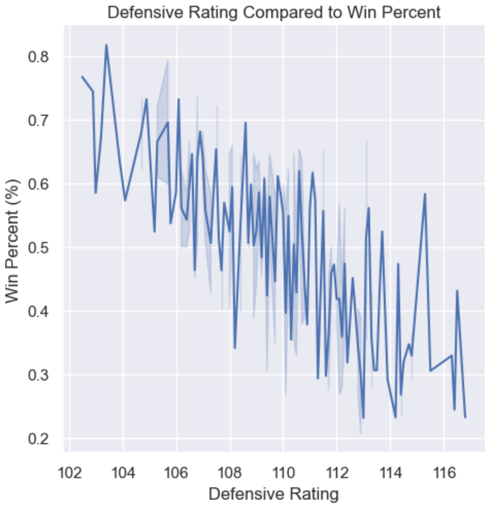
Finally, we have figured out that both offensive and defense has a profound impact on win percentage, as we have disproved the claim that defense no longer matters in the modern NBA and that all you need to win games is offensive output. But, with this new information we can attempt to create a model in order to predict win percentages based on these two statistics, as maybe one variable has a greater impact than the other.

We start with offensive rating, and create a model and split into a test/train split in order to figure out how accurately offensive rating can predict win percent. As a reminder, here is the model for offensive rating compared to win percentage.



Using our test/train split set to a split of 80% train, 20% test, we calculate an RMSE of 0.02, which is extremely low. With such a small RMSE, we can conclude that the offensive rating variable can very accurately predict a team's win percentage. We can also calculate the adjusted R^2, which helps determine how much variability is being described by our model and we can compare this to a future model including both defensive and offensive rating. The adjusted R^2 for this model turned out to be -0.177, which means this model does not count for much of the variability being described by our data set. Using adjusted R^2 and RMSE, we can use a principle called Occam's Razor, which states you should use the model that is simplest when comparing models which predict equally well.

We can shift now to defensive rating compared to win percent, and use the same test/train split principle in order to create a model. For clarification, this is the model comparing defensive rating to win percent:



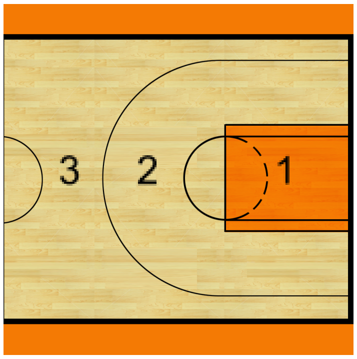
First, we calculate the RMSE for this model. The RMSE turned out to be approximately 0.015, slightly lower than the offensive rating model. The adjusted R^2 for this model was approximately 0.152, which means very little variability of the data was accounted for within this model.

Now since we have figured out both models for offensive rating and defensive rating predict equally well, we can try to create a model using both offensive and defensive rating to predict win percent, to see if there exist any significant difference within these models. We create this new model using offensive and defensive rating to predict win percent, with the same test/train split of 80% training with 20% testing data. The RMSE for this model turned out to be only 0.004, which is slightly smaller than both of the previous models. However, when calculating the adjusted R^2 for this model, it was a staggering 0.768, which means significantly more variability within the data is being explained by our model, much higher than the previous two. We can conclude, using Occam's Razor, that this combined model should be chosen when predicting win percent, as even though more variables are being used the prediction is much more accurate.

Therefore, we have concluded that both defense and offensive rating are necessary for team success. The claim that defense is a lost art and irrelevant in today's NBA is inaccurate, as defensive rating is a key measure to evaluate a team's winning percentage, as a better defensive rating translates to more wins. However, offensive output can not be underestimated either, as you still need a sufficient offensive in order to maintain a high winning percentage, as defense alone may not be enough to achieve this.

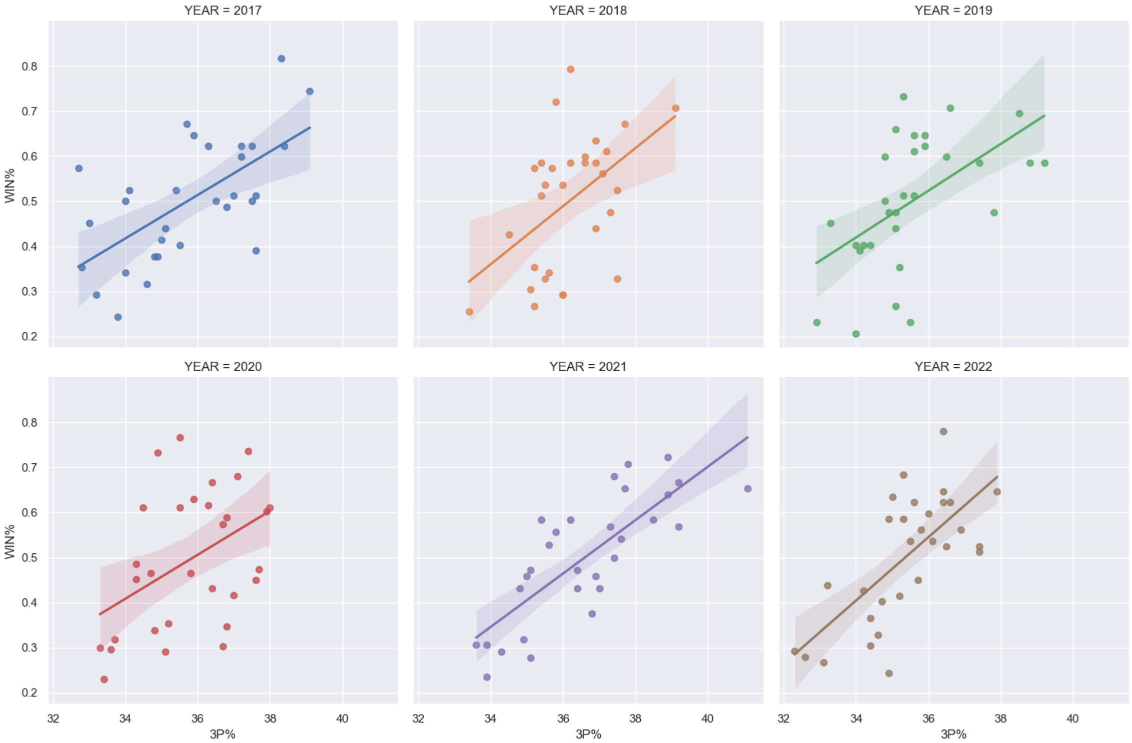
Question 2: Which Zone Affects the Game More? In the paint, mid range, or behind the three point line?

Before trying to answer this question, here is the definition of basketball terms that will be used throughout the analysis.



The paint (1): The orange rectangle.
Mid range (2): Area between the paint and the three point line.
Behind the three point line (3): Area outside the long arc
Independent variables:
(O)CRFGA: (opponent) field goal attempted in the paint
(O)CRFGM: (opponent) field goal made in the paint
(O)CRFG%: (opponent) field goal percentage 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
Note that percentage = made/attempted*100%, this applies in all zones.
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%. R^2 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. The regression plots for each independent variable look like the image below. I only show 3P% against WIN% for simplicity.



Regression plots help to visualize the data. They are not a good tool to directly answer the question because analyzing by eye is not accurate enough. I have to calculate the R^2 to make sure the analysis is precise. Nevertheless, regressions plots can help to verify the R^2 (a graph with an R^2 of 0.005 looks a lot different from a graph with an R^2 of 0.25). Here are the R^2s of all variables from 2017 to 2022, rounded off to 3 decimal places.

Offense:

R^2	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:

R^2	OCRF GM	OCRF GA	OCRF G%	OMRF GM	OMRF GA	OMRF G%	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 field goal attempted has minimal or no correlation to WIN%, thus they are useless for answering my question. None of these variables has an R^2 over 0.2 in any year and a mean R^2 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 field goal made variables on offense, 3 pointers have the highest mean R^2 and field goals made in the paint has the lowest mean R^2. Since the R^2 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 R^2 is the sudden rise of the R^2 in 2021. It is a coincidence because the R^2 dropped again in 2022. For 3PM, it has the highest mean R^2, and the maximum R^2 is 0.238 in 2019. I observe that there is a subtle downward trend, as the R^2 of the first three years is higher than the last three years, 0.185 vs 0.101 to be exact. So for the field goal made 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%, except 2021 for mid range, which is insignificant because it is just an outlier.

On the defensive side, I compare the R^2 of opponent field goal made variables. OCRFGM has the highest R^2 against WIN%, followed by O3PM and OMRFGM.

OMRFGM is completely irrelevant to WIN%, since the max R^2 is 0.061 in 2018 and the mean R^2 is 0.03. For O3PM, things are a bit more interesting even though it only has a 0.095 mean R^2. The first three years all have R^2s above 0.1 and the last three years all have R^2s below 0.1. So, O3PM is having less impact in the most recent years. For OCRFGM, it has a mean R^2 of

0.17. The surprising part is that the first three years have a mean R^2 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 field goal made 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 field goal made, 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 R^2 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 field goal made related variables.

On offense, MRFG% has the lowest mean R^2 against WIN%. 3P% has a mean R^2 a bit higher than CRFG%. But the R^2 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 R^2 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 R^2 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 R^2 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 R^2 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.

Question 3: How can +/- be predicted using generic data including rebounds, steals, turnovers, blocks and assists for both the home and away team?

Before trying to systematically answer this question, we already expect the accuracy to be relatively low because the +/- value of a team, which is the net change of score relative to the opponent's score, is impacted by almost all of the statistical categories in a basketball game. Therefore, using merely the list five can only to some extent make some meaningful predictions. The reason why these independent variables are chosen is because the previous two research questions both focus largely on the offensive end of the game, especially the second research question. Thus, to avoid a certain level of repetition, this research question will focus on other aspects of the game.

With that in mind, the goal of this research question is not to find the best model in accurately predicting the dependent variables given the dependent variables, but instead comparing the accuracy between some simple regression models as well as trying to justify why a certain is better than another. The accuracy is accessed by calculating mean absolute error, mean squared error and R^2.

The first regression model used was the decision tree model as it was the one taught in class. However, the accuracy was a lot lower than I expected. The resultant mean absolute error was 6.2 and mean squared error was 61.4. These two error calculations directly reflect how prediction is away from the real value and they are considered huge because most of the +/- values fall in the range of -8 to 8, so a mean absolute error of 6.2 is certainly not accurate enough. This is further justified with a R^2 score of -2.0, which indicates an arbitrary relationship.

Then, I went ahead to try the second regression model which is the linear regression model. It always follow the following pattern:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Where Y(s) are dependent variables, x(s) are independent variables and θ(s) are constants that the model is trying to figure out. This model simplifies the problem as it assumes that all variables have a linear relationship to the result. The result was much lower than the first model. The absolute mean error falls at 3.9 which is a roughly 40% improvement from the previous model and the mean squared error dropped down to 25.4, and the R^2 score dropped down to 0.44, which indicates a reasonably positive relationship. Then I start thinking about why this model has such a huge improvement on the previous one. I found out that decision trees in general will produce a more accurate prediction, and it can accept a wider range of data, both categorical and numerical. On the other hand, a linear regression model can only process numerical data and is usually the less accurate one because it oversimplifies the problem as it assumes the independent variable will linearly affect the dependent variable, this leap of faith always results in overseeing biases.

However, these are only the general patterns of the two models, the actual result still largely depends on the characteristics of the dataset that is fed into the model. In this case, using a linear regression model is probably a better model than a decision tree because how the decision tree works underneath was that it would continuously ask Yes/No questions on each independent variable. For example, is the steal of a team under 3 or is the opponent rebound over 40? These questions are built upon each other in the sense that when you go deeper down the tree, all of the previous statements must be tree. For example, a prediction may be made based on the previous statements the team must have more than 3 steals, less than 4 turnovers and restrict the opponent's rebound to less than 43. These series of questions, though work well sometimes, are not ideal for this specific dataset. This is because the chosen independent variable and dependent variables do not have an obvious relation to start with as we expected early, therefore, having a rigorous step by step question process may over complicate the probelm.

In comparison, the more simplified linear regression model to some extent avoids trying to deal with the messy data and find an extremely accurate relationship and instead just makes a close estimation, which works better for this data set. In addition, since the +/- value can be impacted by many variables as mentioned earlier, it is probably a better choice to consider all variables at the same time as in the linear regression model as opposed to the decision tree model which considers one variable at a time.

And lastly, in real life, no NBA team is trying to set goals on any categorical variable such as they have to get at least 44 rebounds in the game to win it, but in fact, not a single statistic actually matters unless they win the game. Thus, the decision tree does not seem to fit with the essence of the basketball games.

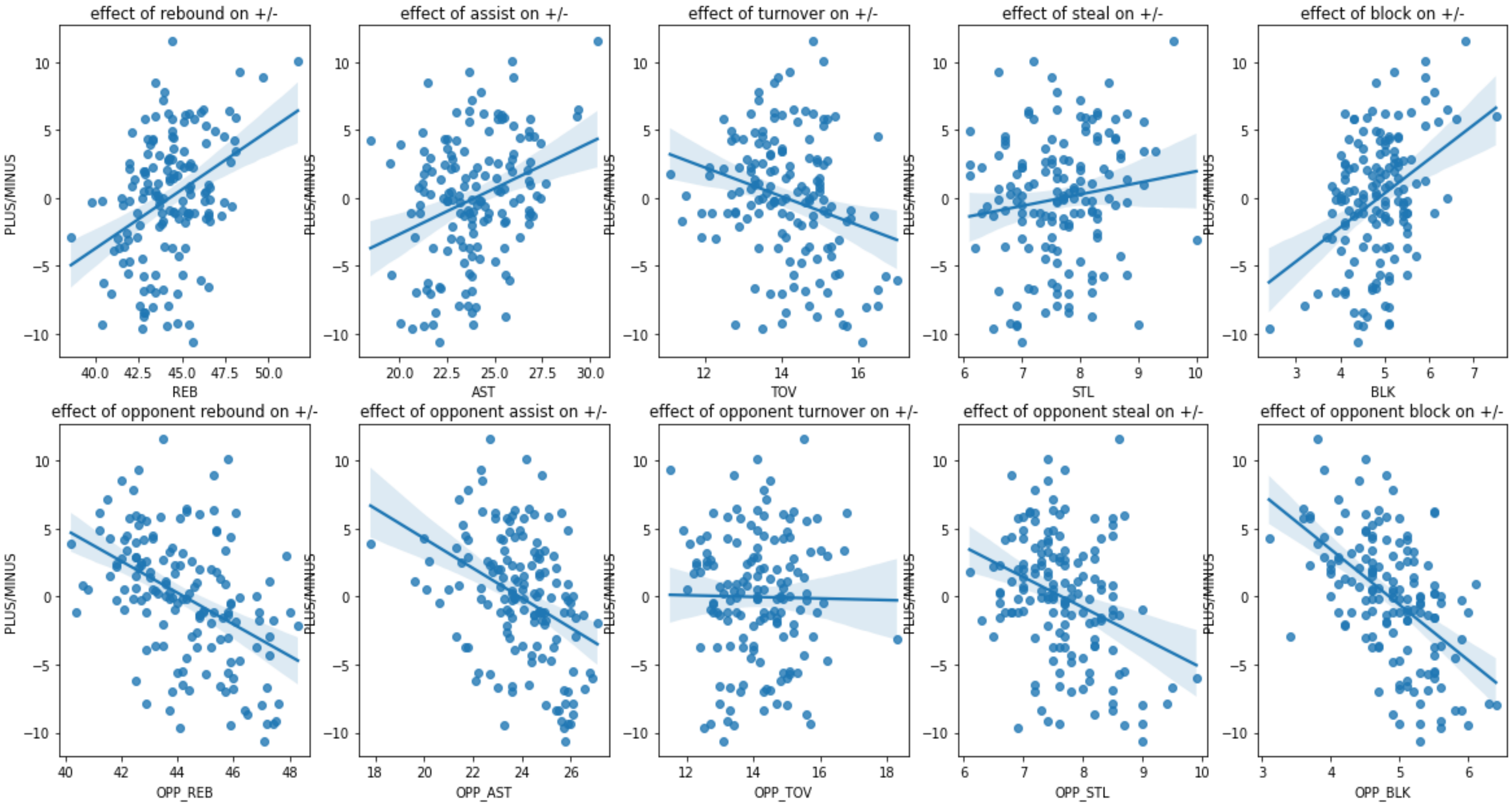
Then, I moved on to testing the polynomial regression model, with the hypothesis that higher order will result in a higher accuracy as it takes more factors into account and is therefore closer to real life. However, aftering testing 2, 3, 4, 5 order regression models, the results were shocking to me. The second order regression model has a mean absolute error of 8.3, mean squared error of 133.7, and a R^2 of -0.8, which seems worse than the decision tree model. Then, starting from the 3 degree order regression model, the error even becomes considerably larger, they are as the following:

	Mean absolute error	Mean squared error	R^2
3 degree order	74.9	16001	-0.06
4 degree order	69.6	14145	-0.08
5 degree order	66.3	15070	-0.09

The errors are huge, especially the mean absolute error and mean squared error. Since a team normally only outscores the opponent team by less than 15 points, the error 66 or more imply no meaning at all.

At this moment, I decided to try to justify why the linear, namely first order, regression model is better than the higher order regression models. The answer is not much different than the previous model. Even though higher degree regression can indeed improve accuracy in many situations especially in applied mathematics. However, the premise is that we have to ensure there is indeed such a relationship existing which is true in Applied Math as much data is supported by theory before they are being processed. However, for this data that the variables are chosen without rigorously examining the relashition, having a higher degree order regression model does not bring it closer to the real connection between independent and dependent variables because there is no such existing in the first place. Instead, doing so will only bring more sources of error to the computation, which is exactly what happened here.

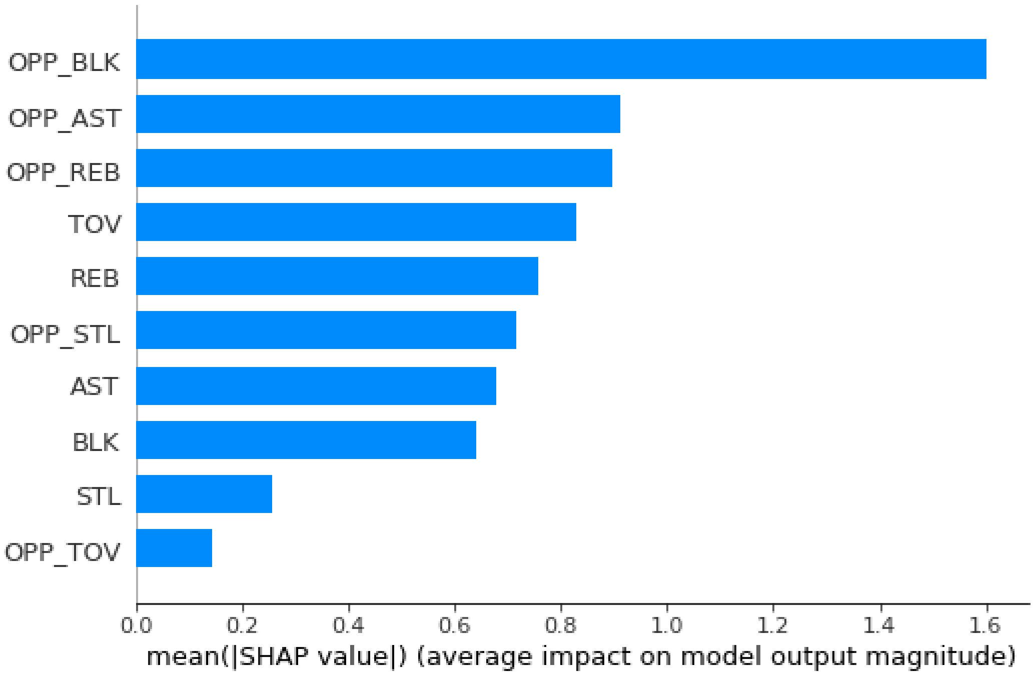
I further verify it by plot each individual independent variable against dependent variable and the following graph is what I got:



We can see from these data that:

1. First, the messy data seems to imply that the relationship between independent and independent variables does seem to be loosely related.
2. However, when looking at the best fit line, the logic does indeed make sense. That is, more rebounds, assists, steals and blocks will positively affect your game, which is reflected by the positive slope in these graphs; and more turnovers will negatively affect your game, reflected by the negative slope in the graph.
3. Furthermore, we can see that the slopes of the home team and opponent team under the same category seem to be symmetrical, meaning that how your opponent did in the same category has a very similar effect as how well you did yourself.

Additionally, I used the Shap library to compute the Shap value of each independent value. Shap values determines the weighting of each independent variable on the dependent variable and this is what I got:



Then, I compared those values to the slope of the early 10 regression plots. Unfortunately, there is no function issued by seaborn to conveniently calculate slope, so I had to calculate by hand. Therefore, some errors may occur but they were certainly insignificant compared to those produced in other parts of this research. These are the hand-calculated slopes of the earlier 10 graphs (note: they are not exactly what they seemed like in the graph because the range of x values are different):

Category	rebound	assist	turnover	steal	block
slope(always positive)	1	0.775	1.25	0.625	2.55
Category	Opp rebound	Opp assist	Opp turnover	Opp steal	Opp block
slope(always positive)	1.125	1.25	0.08	-2.125	-4.5

The slopes are taken as absolute values because both positive and negative effects are considered the same weighting if their magnitudes are the same.
Therefore, the ranking of each independent variable on the dependent variable according to the plot vs predicted is: opponent block, block, opponent steal, turnover, opponent assist, opponent rebound, rebound, assist, steal, and lastly opponent turnover.

Ranking of each independent variable's weighting on the dependent variable according to Shap library vs slope of best fit line

Category	Predicted using Shap library	Estimated based on slope
rebound	5	7
assist	6	8
turnover	2	4
steal	9	9
block	7	2
Opponent rebound	4	6
Opponent assist	3	5
Opponent turnover	10	10
Opponent steal	8	3
Opponent block	1	1

We can see that three categories (opponent block, opponent turnover, steal) are in the same position and five categories (turnever, opponent assist, opponent rebound, rebound, assist) are im the exact relative order, but the rankings are all differ by 2, which is because the last 2 categories are differ by quite a few, though their relative position still remain the same.
The difference between two rankings are within understandable range because the actual weightings are close to each other from the 2 to 8 place and that the best fit lines are hugely inaccurate.
Therefore, we can conclude that for this data set specifically, the linear regression model is the most accurate in predicting the +/- value of a team given that team's rebound, assist, turnover, steal, and block, as well as the same set of data from the opponent team. And again, we want to restate that the goal of this research question is not to find the most accurate model, but rather to compare the three types of simple regression model, decision tree, linear and polynomial, find the most optional one out of the three and evaluate the pros and cons of each based on the characteristics of this current dataset.

Impact and Limitations:

There were many beneficial implications of our results, specially catered towards NBA staff members and general managers. When constructing a team, winning games is obviously one of the top priorities as a team's goal is always to win the most possible games every season. Within our results, we determined what factors may affect winning percentage, which in turn NBA staff members can use to evaluate their roster. For example, since we disproved the claim that defense is unnecessary in the modern NBA, general managers who have constructed offensive oriented teams, such as the Nets, could in turn trade or sign more defensive players to accommodate for their lack of defense in order to improve their defensive rating and win percentage. Possibly the acquisition of Ben Simmons, a known high level defender, could improve their overall defense, hence win percentage next season. We also concluded that 3 pointers are the most decisive zone in a game, so a team should acquire more elite 3 point shooters and perimeter defenders in order to improve their regular season record.

One major limitation in this data set was the rounded win percent, as many win percentages were rounded to the nearest thousandth place, rather than the exact number. For example, the Suns won 64 out of 82 games, corresponding to approximately a 0.7804878 winning percentage, but nba.com has this listed as only 0.78. As a result, when calculating statistics, such as mean win percent for above average and below average offensive rating teams, the addition of these win percentages is only about 0.99 rather than 1.0, since these percentages are rounded. This isn't the only statistic which is rounded though, as nearly every percentage was rounded, assists, rebounds, shooting percentage, etc, so there was no way to get 100% accurate data.

A limitation within our analysis is the data we used. For example, within the first question we used the variables offensive and defensive rating in relation to winning percentage. Offensive and defensive rating aren't the perfect measurements to how good an offense or defense truly is, it's one of many NBA statistics which measure the quality of a team's offense or defense. There are many similar statistics, such as offensive efficiency, defensive efficiency, and basic stats such as rebounding and assists, that aren't necessarily correlated to offensive and defensive rating. One may use the results from this question however to figure out how defensive and offensive rating may affect win percent, but must keep in mind that having a high offensive rating doesn't mean the team has a good offense, as there are many other variables which may affect offensive output. These two variables can be very broad, but nonetheless a sufficient indicator towards how good a term performs on both ends of the court.

In question two, a minor limitation is the inability to address the outliers. For example, the field goal percentages stats in 2018 and 2020. We cannot provide an explanation unless we have other stats. These outliers don't change the overall final conclusion, but they harm the reliability of it.

Another major limitation in our analysis is that we did not consider each team's roster. This means our answer can answer on a season level (average of 82 games), but not on a single game level. Take Warriors and 76ers as examples. When a team faces the Warriors, of course 3 pointers are going to affect the game the most as the Warriors have sharpshooters like Klay Thompson and Stephen Curry, etc. When a team faces the 76ers, of course the winner of the game is determined mainly in the paint as the 76ers has a paint beast Joel Embiid, and had a point forward Ben Simmons who is terrible at shooting but quite dominant in the restricted area. Take Warriors vs Rockets and Bucks vs 76ers as examples. When the Warriors faced the Rockets, of course 3 pointers affected the game the most as the Warriors had and still has sharpshooters like Klay Thompson and Stephen Curry while the Rockets had James Harden and a lot of shooters. When the Bucks faced the 76ers, of course the winner of the game was determined more in the paint than Warriors vs Rockets, because of paint beasts like Joel Embiid and Giannis Antetokounmpo, point forward Ben Simmons who was and still is horrendous at shooting but quite dominant in the paint.

In research question 3, one major limitation comes from the fact that the actual relationship between independent and dependent variables is very loose because the +/- value is impacted by almost all aspects of a basketball game. While the chosen independent variables only accounted for a small proportion of the game, especially since it did not account for the offensive end of the game, namely shooting, the lower accuracy is almost guaranteed. However, this research question is answered in a way that the absolute value of the error does not matter as much as the difference between errors generated from different models, which reduces the effect of this limitation on the validity of the conclusion we drew from the analysis.

Another limitation was due to the lack of data. The total number of rows used to train the regression model is merely 150, which is certainly not enough to ensure the model will have consistent performance on a wide range of testing data. And this is in my opinion a big reason why the linear regression model turned out to be the best because it is less susceptible to errors ingrained in the data itself as it oversimplifies the problem, which happens to ignore the errors in this case. While the decision tree is not good at handling systematic error. In addition when the amount of training set increases, the accuracy of the decision tree model will also increase at a larger rate than for the linear model because it follows more rigorous algorithms so more training set will make it much more robust.

The last source of error comes from the fact that the game philosophy is changing rapidly, meaning that every year each team is putting emphasis on a slightly different aspect of the game, due to reasons such as change of roster, what plays have improved or changed during the off season. In addition, the nba league as a whole also undergoes changes over time, for example, in the recent decade nba has changed a lot, pushed by pioneering players like Stephen Curry and Damian Lilliard who shoot 3 pointers way more than what a player used to. As a result, many teams as a whole started to imitate the way those players play, such as the Rockets and Jazz. The point was the model was trained using data from different teams across different years, so all of those differences in play styles are not taken into consideration. As a result, the trained model will not be that accurate.

Challenge Goals:

For the first challenge goal we decided to test the **result validity** of our analysis. We used some basic methods to verify model validity, such as calculating the RMSE and R^2 / adjusted R^2 score in order to calculate the variation explained by the model and the accuracy in the predictions. We also used a principle called Occam's Razor, which states you should choose the model that is simplest, meaning the least number of variables, when comparing models which predict equally likely, which was an idea expanded from the proposal since this concept is very useful when choosing the "best" model to use. We also calculated confidence intervals and p-values, since these statistics can help verify validity in your conclusion. The confidence interval helped figure out the mean difference between two variables, while the p-value helped assess our null hypothesis in order to reject or not reject it.

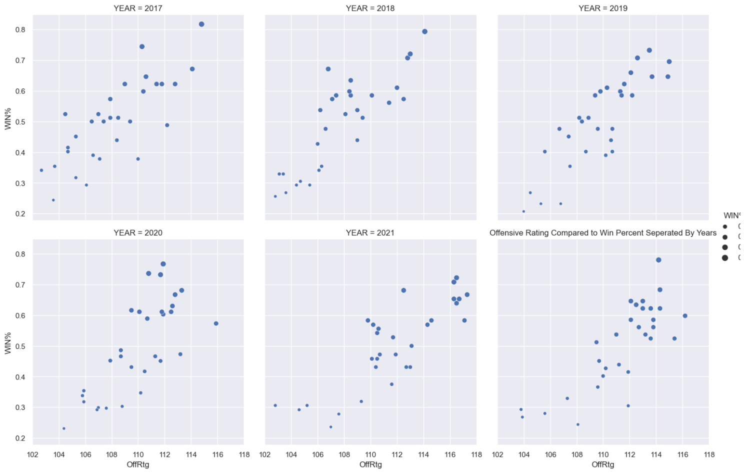
For the second challenge goal we decided to implement **machine learning** in our analysis. Though analysis for all three research questions involve machine learning, the third analysis primarily focused on ML, specifically different kinds of ML models as well as evaluating the effect of each independent variable on the dependent variable. To be more specific, this research question mainly focussed on three kinds of regression models: decision tree regression model, linear regression model and polynomial degree regression. The linear regression model was found to be the most accurate based on calculating errors. And it was compared to each of the two other models and justified why it outperformed the other two using plots, reasoning which considered how the algorithms work underneath each model and the characteristics of the dataset, as well as the shap library which was used to interpret the training set and yield useful information for further justification.

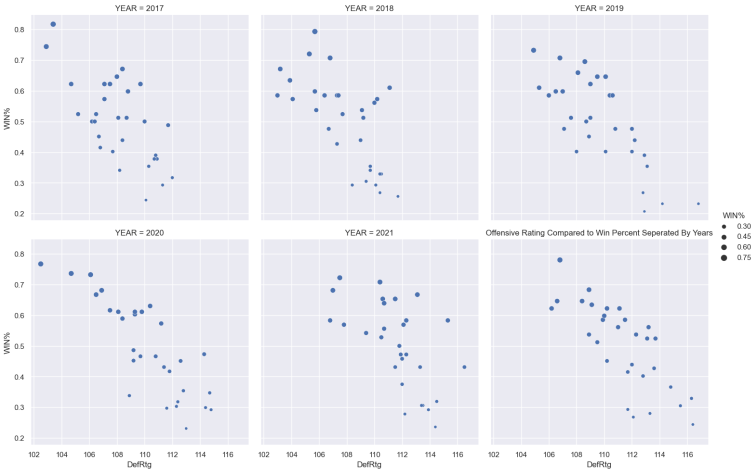
Work Plan Evaluation:

Our work plan proposed within our proposal was not very accurate. Some portions, however, were more accurate than others. The first step in our plan was to group together and figure out what we wanted to work on and create our questions, the estimated time of 1-2 hours was fairly accurate. However, the second stage of code development took much longer than expected. We didn't necessarily follow this plan however, as the second stage wasn't necessarily a rough draft as we stated it would be, but more of a series of meeting together and helping each other out as we moved along while working thoroughly on our parts. To really finish this step and create our finalized code, this took upwards of 8-9 hours each. It definitely was tougher than expected, as some obstacles such as figuring out new libraries and working out plans for our code in order to answer our questions took longer than expected. Figuring out a way to make our code flow and navigate towards the answer was much harder than expected, as in assessments in the past we were told what steps to take and exactly what variables to use, now it was up to us. The grouping together aspect of our work plan was also a bit off, as this step probability took around 3-4 hours, as we made sure to verify and check over every aspect of our code and report. Making our presentation took about the right time, 1-2 hours. Our final check in also was a bit off, as that took nearly 2 hours as we wanted to make sure everything was perfect.

Testing:

For the first question we decided to create 2 separate tests. The first test created a model representing the different years within our data set, comparing offensive and defensive rating over the years to the overall win percent, to visualize and verify our results. The graphs labeled Offensive_Years_Test.png and Defensive_Years_Test.png are the graphs depicting the tests. Within these graphs they both followed the results from question 1, as the higher the offensive rating, generally the greater the win percent and the lower the defensive rating, generally the greater the win percent. For the second test we created a smaller data set using only the years 2021 and 2022, calculating the high/low offensive and defensive rating for these years. Again, these years followed our results, with a higher offensive and lower defensive rating corresponding with a higher mean win percent. Here are the graphs for offensive and defensive years compared to win percent, as you can see they follow the results that we made in question 1. The left graph depicts the offensive rating, while the right graph depicts defensive rating, both compared to win percent. Testing the smaller data sets, the high offensive rating win percent was much greater than the low offensive rating, while the low defensive rating was also higher than the low defensive rating, as predicted.





For the second question we decided to use 10 out of 30 teams' stats to test the result, using the same methods we used for previously. Out of these 10 teams, some are great teams, some are bad teams. Again, assuming offense and defense have equal contribution on a game, the sum of offensive and defensive R^2 of independent variables against WIN% looks like the table below. FGA related variables are ignored, for the same reason I mentioned before.

Sum of offensive and defensive R^2	In the paint	Mid Range	3 Pointers
FG% related variables	4.714	4.161	5.427
FGM related variables	2.151	2.455	2.695

If we consider FG%, 3 pointers affect the game the most, mid range the least. If we consider FGM, 3 pointers affect the game the most, the paint the least. So, whether we value efficiency or quantity more, this test proves 3 pointers affect the game the most, in regular season of recent years.

For the third research question, we did not implement any testing because the goal of this research question is to compare three regression models and find the best model out of the three. All the justification and testing for hypotheses are already done in the analysis, so there is no need for additional testing.

Collaboration:

We didn't collaborate with anyone else except ourselves during the project. However, we did use websites, such as spicy.org, to help figure out what libraries we could use in order to help create our results.