Jason Katz

General Questions

1. I recently completed a project analyzing data from MLB games. The dataset contained over 160 variables, for each game played in the MLB since its inception in 1871. Before I could begin my exploration and analysis, I had to complete extensive cleaning to make the data usable. I then explored the variables available and decided to pursue a few different hypotheses. Using R, I parsed and manipulated the data to create Naïve Bayes prediction models and visualizations to show important trends and statistics. My two main models predicted the chance a team had won and the amount of runs a team had scored, given stats from the game. My visualizations showed the amount of home field advantage each team had, how various team statistics factored into how many runs they score, and the affect a particular team had on road attendance.
2. My experience with programming thus far has been predominately academic. I am most experienced with R, and have become adept at using the program to clean and manipulate data, create models and visualizations, as well as design interactive tools for others to use. I have used SPSS for statistical analysis in the past, as well as MATLAB for mathematical modeling. In the coming months, I am expanding my coding skills by working between semesters to undertake various projects to build familiarity with both Python and SQL.

Project Outline

1. The decision making opportunities I targeted revolved around which shots are the best for an offense to take and critical changes a defense can make to affect opponent shooting percentage. A key component in determining whether a shot goes in is where it is taken from. By breaking the court into different zones (threes, paint, mid-range), the differences in field

goal percentage become apparent. Additional variables that played roles in determining field goal percentage included how close the defender was, how many dribbles were taken before the shot, and how fast the shooter was moving at the time of the shot.

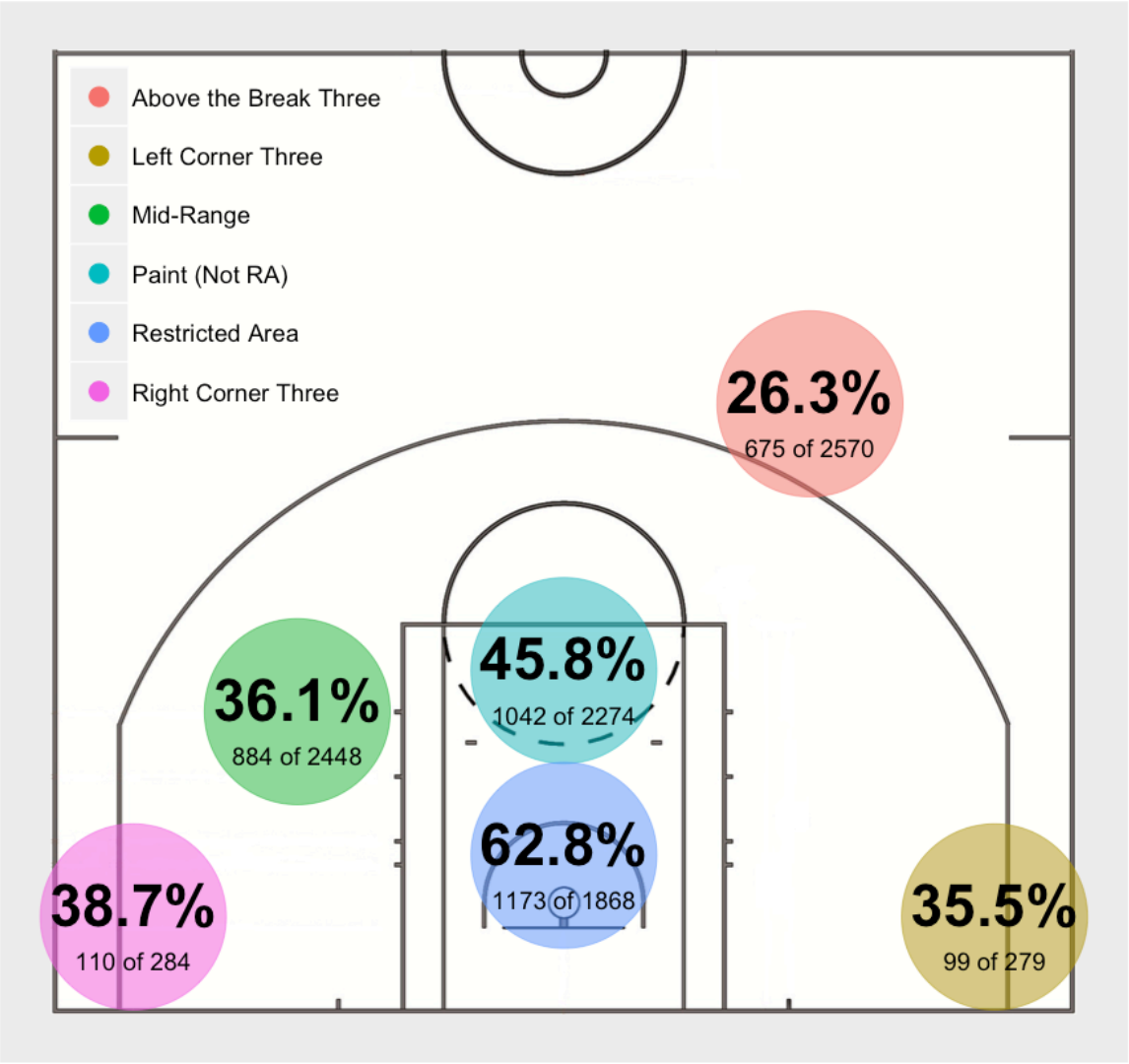
These general variables show significant difference in field goal percentage when combined together. My hypotheses consisted of comparisons of specific types of shots. I explored the effective field goal percentage of mid-range shots vs threes and the field goal percentage of open mid-range shots vs contested layups. I also compared three point shots based on how close the defender was, the speed of the shooter, and whether they were pull ups or catch and shoots. Another key decision I investigated was whether a player should take a contested shot after driving the lane, or kick it out for an open three. I lastly examined how important each foot of space was for different shot zones.

2. I used R for this project. To begin, I read in the data, and explored the variables given. I then created additional variables from the data. Using the standard NBA court dimensions and the shot coordinates, I created variables for shot type (two or three), points scored on the shot (zero, two, or three), shot zone [left corner three, right corner three, above the break three, mid-range, restricted area, paint (not RA)]. I next created a new data frame from the original data set to compare these six different shot zones. In this data frame, I created variables such as shots made, shots attempted, field goal percentage, effective field goal percentage and points per shot. Using this data and a half court image uploaded from the internet, I was able to create one of my key visualizations showing the difference in field goal percentage among different shot zones.

After manipulating the data and creating this visualization, I used an R package called Shiny to create an interactive tool. The tool can be used to visualize the difference in field

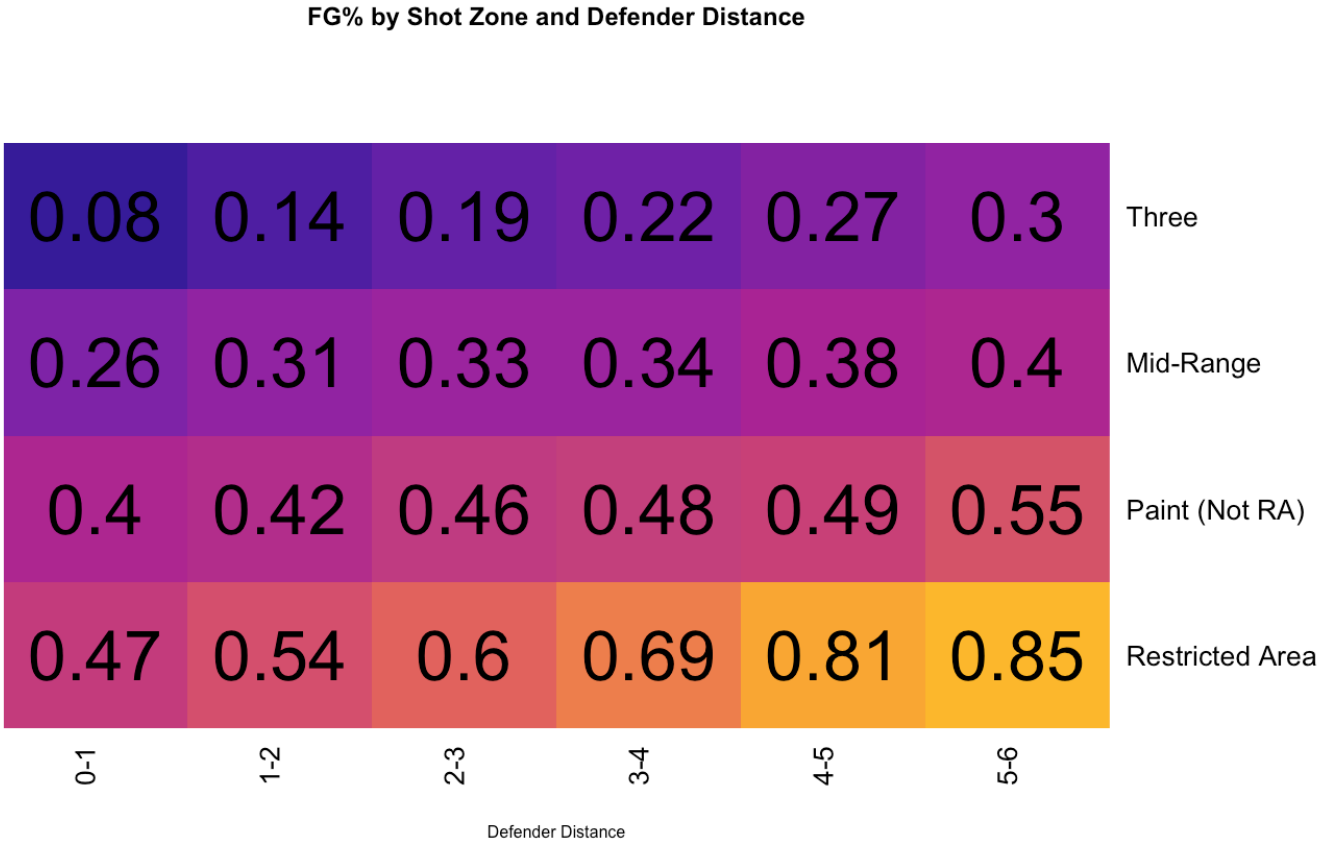
goal percentage among the different zones for ranges of various variables. After creating this interactive tool, I decided to create another visualization to show how defender distance affects field goal percentage in different shot zones. To create this heat map, I used various filtering techniques and for loops. Once my data was in proper format, I was able to use the heatmap function to generate the useful visualization. I next used the t.test function to test my various hypotheses. Lastly, I broke the data into additional shot zones to create a model to predict whether a shot would be made based on numerous conditions. Using a for loop and folds for data partitioning, I was able to create a Naïve Bayes model that had a very high accuracy.

3.



Jason Katz

The visualization above shows the field goal percentage for six different zones on the court. By adjusting the location where players take most of their shots, total field goal percentage can increase. Using this visualization, I created an interactive tool that shows either the field goal percentage, effective field goal percentage, or points per shot for these six locations. The percentages can be adjusted through easy filtering of the different variables (defender distance, dribbles before, shooter velocity, etc.), making this visualization a completely interactive model that could enable coaches to adjust and find their own significant trends. An important statistic I calculated was effective field goal percentage (field goals made + .5*three-point field goals made / field goals attempted), to better compare three point shots to two point shots. While constructing this visualization, I had to take into account how many shots fell into different ranges, and how the different ranges compared in their field goal percentages. The creation of additional ranges decreased the simplicity of the model. Left and right corner threes were included despite their relatively small sample size because of the difference between their field goal percentage and the above the break three point percentage.



The second visualization above is a heatmap that show how field goal percentage changes for the four main shooting zones as the defender distance increase from zero to six in one foot increments. This type of visualization is a simple way to show how a result changes as variables increase. The field goal percentage increases as you move to the right, increasing defender distance, as well as when you move down the chart, going from threes to mid-range to paint (Not RA) to restricted area. When creating this, I had to break the zones into four instead of six like the previous, because the sample sizes of the corner threes were too small to subset even further. Additionally, I had to limit the range of defender distance to six feet, because the one-foot bucket sizes for anything higher was too small.

4.

The main part of my project was creating an interactive tool to visualize different aspects of field goal percentage in different shot locations using a variety of variables. My tool can be used at: https://jasonk33.shinyapps.io/NBA_Field_Goals/ . Exploration of how different variables affect shooting is made easy and a coach or manager can explore any hypotheses they might have. Using the tool I created, I found seven important results to improve decision making, quantified though two sample t tests in R. All tests were one sided and results are concluded using a 99% confidence interval. The first result I found was that the effective field goal percentage of threes is at least 2.6% higher than the effective field goal percentage of mid-range shots. Additionally, the field goal percentage of contested layups (shots in restricted area with defender less than two feet away) is more than 5.8% higher than that of open mid-range shots (defender more than six feet away). This suggests that players should take less mid-range shots. The next three conclusions relate to three pointers. Three point percentage decreases by more

than 4.8% when a defender is within .5 to 2.5 feet as opposed to 2.5 to 4.5 feet. This is the key range when a defender is closing out on a three point shot, with each feet significantly affecting the chance the shot goes in. The field goal percentage of threes increases by at least 3.9% when they are catch and shoot (no dribbles before) compared to when they are taken off the dribble. The next result showed that threes taken when the shooter is moving at over nine feet per second have a field goal percentage more than 20% lower than threes taken moving at slower velocities. This suggests that it is not a good idea to take a three when moving at a fast velocity. Next, the field goal percentage of layups decreased by more than 11.6% when the defender was within 2 to 4 feet, as opposed to 4 to 6 feet. This is the key range for defending layups, with significant emphasis on each foot of space the shooter has when they are so close to the basket. The last result focused on what a player should do when they drive into the lane and defenders are close to him. I concluded that a player driving into the paint (not restricted area), taking a contested shot (defender within two feet) has an effective field goal percentage more than 5.1% lower than a player who takes an open (defender more than 5 feet away) catch and shoot three. This suggests that if a player drives to the paint, but is met by defenders, it is a better decision to kick it out for an open three (if possible) then to take the contested shot in the paint. In addition to these results, I built a Naïve Bayes prediction model for whether a shot was made or missed. Breaking the court into ten different zones, I was able to use the shot zone and the defender distance to predict the chance the shot was made with accuracy of 65.3%. To get this prediction accuracy, I used a 5-fold cross validation with 5 repeats.