**Data Exploration and Visualization**

*Our Data*

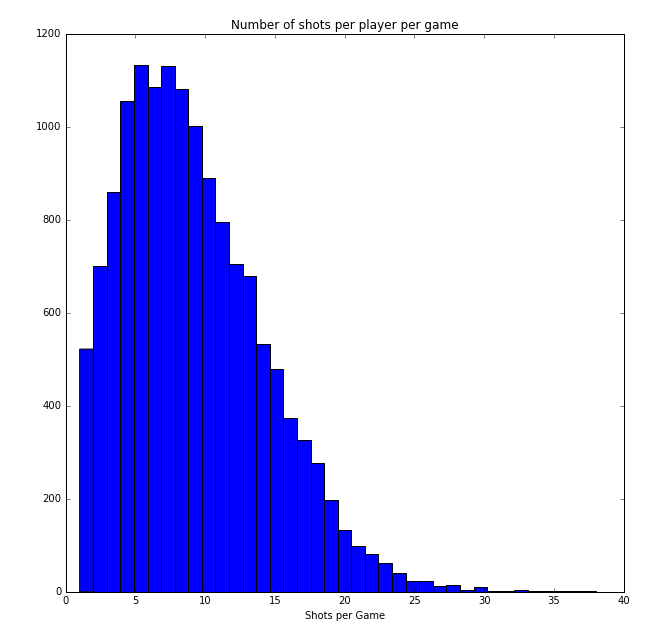
The dataset we used was a comprehensive shot log of every shot taken by over 200 players in the NBA for the 2014-2015 season. It contained 23 different variables for each shot, including shot distance, type of shot, closest defender distance, final score, and others. With this dataset we had a large amount of shot information to help us understand individual player shooting habits.

The data is accessible on Kaggle at this link: <https://www.kaggle.com/dansbecker/nba-shot-logs>

*Data Cleaning*

This dataset was almost 100% complete, so we didn’t have to do much processing to clean it up. The only cases of missing values were in the *shot\_clock* column, since the NBA turns off the shot clock if the time on the shot clock is greater than the time remaining in the game, eliminating redundancies. Because of this, for every missing value in the *shot\_clock­* column, we filled it with what the value in the *game\_clock ­*column was for that shot.

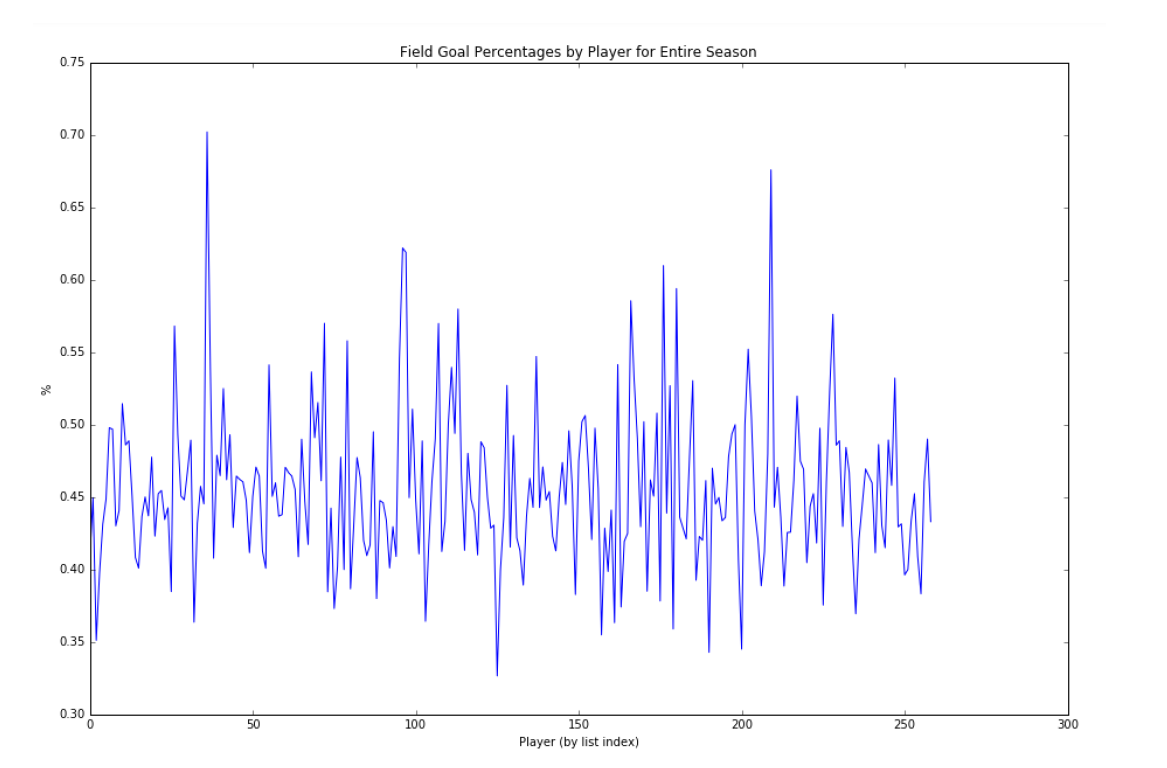
We also wanted to make sure that the sample size for each game was large enough to draw conclusions on. It wouldn’t make sense to consider a game in which a given player only shot 1 or 2 times, so we wanted to find a threshold Shots per Game value to eliminate some of these lower shooting games. We found that the mean number of shots per game was 8.9 shots/game, but wanted to visualize this finding further. To do this, we created a histogram for the shots per game for each player.



From this histogram we see that the distribution is skewed towards the left, so a cutoff value of 8.9 shots/game would eliminate a large number of games. Thus, we decided to eliminate all games where the player shot less than 4 times. This eliminated games where players didn’t shoot enough to show meaningful results with regards to hot streaks.

*Exploring Top Players*

We knew that we wanted to explore how the Hot Hand theory may show up for some of the top players in the league, so we needed to find a way to determine who our top players were. We decided to base this on Field Goal %, which is a metric for what percentage of shots a player makes. For each player in the dataset, we calculated their Field Goal % for the entire season by dividing the total number of made shots in the dataset by the total number of shots taken. We also appended this information as a new column of data along each player in our dataset so it could help in future analysis. We plotted the Field Goal % for each player to see how the league varies and if we may find significantly better players.



This plot shows that there is a decent amount of variation in the data, with some players dipping as low as 30-35% and some players as high as 70%. We found that the maximum values on the plot represent the players Ed Davis, Mason Plumlee, James Johnson, Tyson Chandler, and Deandre Jordan. We decided to further explore the data for the 5 players specifically to see if they truly represented what we believed a top player should be.

*Creation of the Current\_Streak and Previous\_Streak Variables*

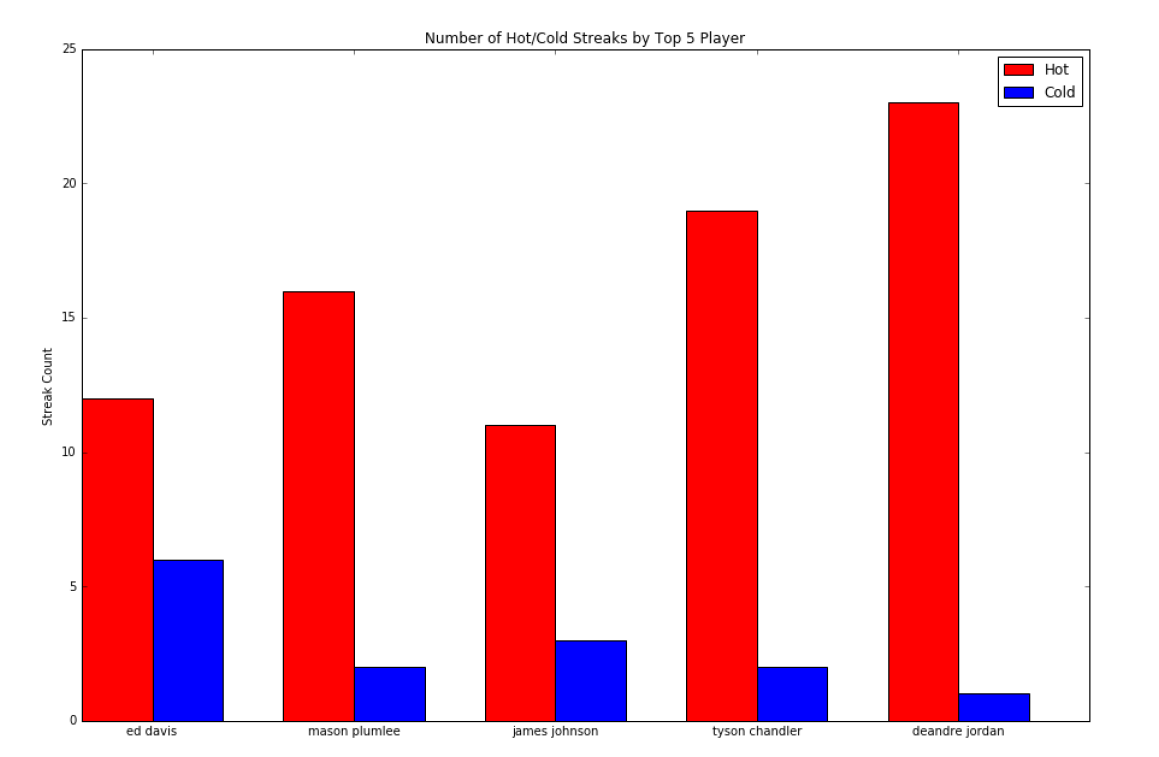
Before we could continue with our analysis of our top players, we needed to actually create a metric that tracked shooting streak since our dataset didn’t contain that information. To do this, we decided to create a *current\_streak* variable that would show, for that given shot, what shot number is was in the streak. To do this we created a set of rules:

* At the start of each game if they made their first shot, set *current\_streak* to +1; if they missed, set it to -1
* If on a positive shooting streak then for each consecutive shot they made, add +1
* If on a negative shooting streak then for each consecutive shot they missed, add -1
* If they were on a positive streak but then missed a shot, reset their *current\_streak* value to -1
* If they were on a negative streak but then made a shot, reset their *current\_streak* value to +1

This allowed us to keep a running count of how each shot in the dataset affected the player’s shooting streak. We also created a *previous\_streak* column in our dataset that was just the *current\_streak* column offset by 1, with a 0 value for the first shot in a game. This let us have information readily available about the player’s shooting streak leading up to a given shot in the dataset. With these two new variables, we could continue our preliminary exploration of the Hot Hand for the top players we selected.

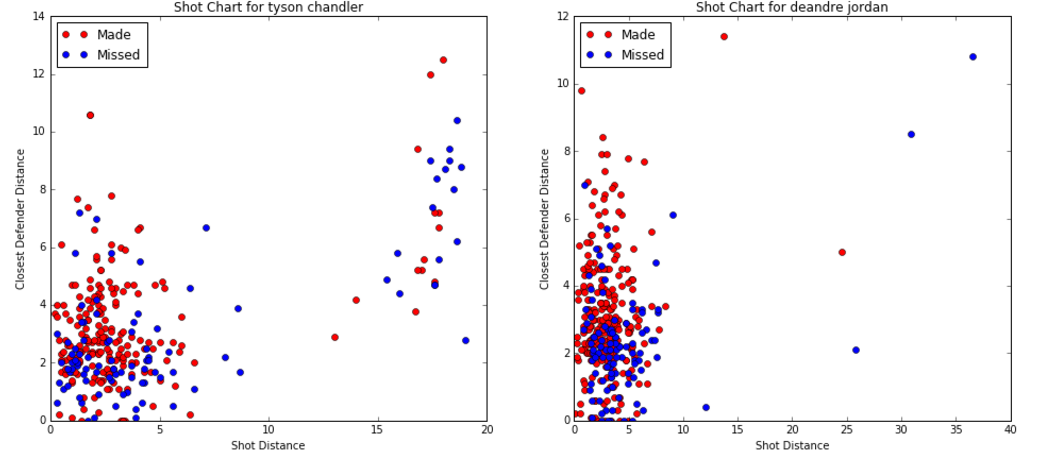
*Continued Exploration of Top Players*

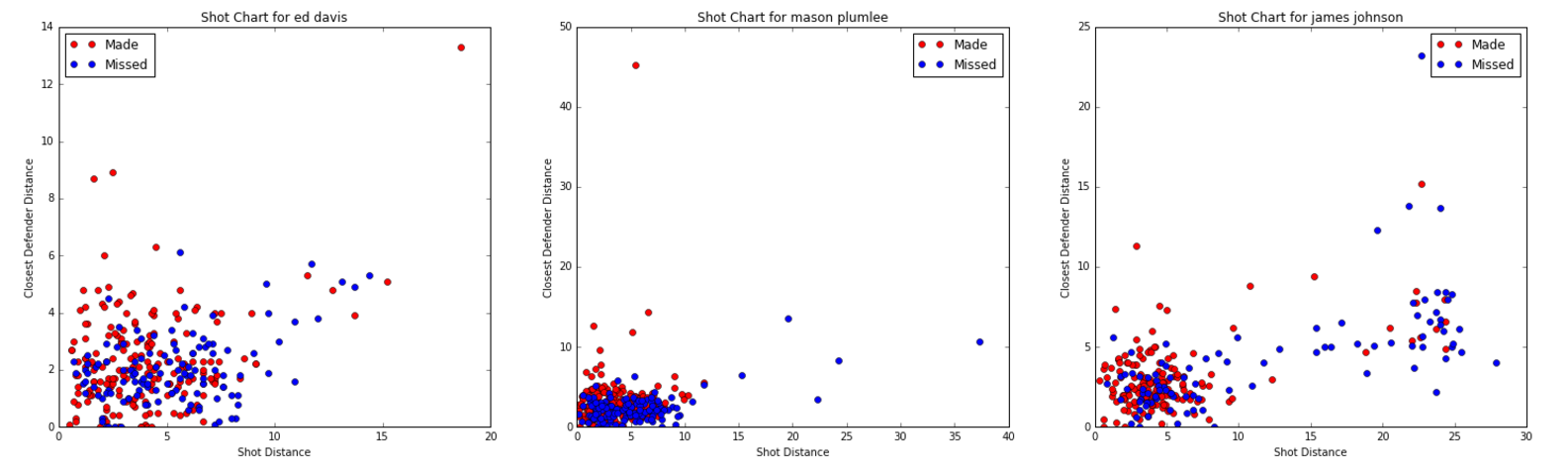
For each of our top players, we wanted to see how many streaks they went on over the course of a season. We considered a hot streak to be a streak of 4 or more shots made, and a cold streak a streak of 4 or more shots missed. From this information we were able to examine how streaky our top players are.



This bar graph shows us that the top players we’re considering appear to go on a lot more hot streaks than cold streaks across the season, suggesting they may be streaky shooters that may feel the effect of the Hot Hand. However, this information is not surprising considering we’re examining players with high Field Goal %, so they’re expected to make more shots than they missed.

At this point we felt like we needed to further examine if these players are indeed what we truly consider top players. Looking at the names of these players, they aren’t what we’d expect the superstars of the league to be. Indeed most of these players play center and are considered ‘big men’; essentially, they’re the players who play low in the post near the hoop, and put up shots from a very close distance. Naturally we would expect that players shooting that close to the basket would have a higher Field Goal %, so we wanted to see if shot distance was concentrated for these players. We created a distribution chart of Shot Distance vs. Closest Defender Distance for each of these 5 players to see if their shots were clustered around the basket.





These distributions do indeed show that the majority of these players’ shots are from very close. For the sake of our analysis, we would like our top players to have greater shot distance variance. This preliminary exploration shows that we’ll have to adjust our criteria for what we define as a Top Player. This adjustment is explained in greater detail in our Player by Player Analysis of the Hot Hand Modeling page.

*Overview*

Overall this data exploration and visualization gave us a good idea of how to examine our data and some early thoughts on selecting top players. It also allowed us to create some new predictor variables, including Field Goal%, Current Streak, and Previous Streak that may prove useful in our modeling analysis.