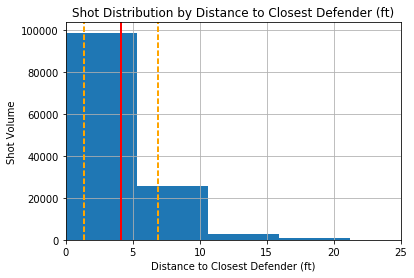
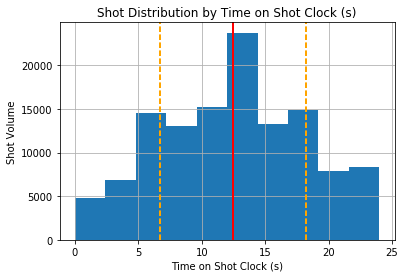
**Data Science Prep Course - Capstone Report**

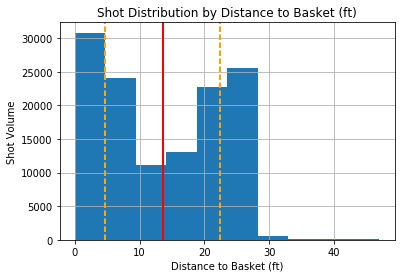
**Ryan Mitchell**

For my Capstone project, I wanted to explore a dataset that not only had appropriate depth and breadth for critical analysis and visualization, but also one whose contents aligned with my particular interests. It took a few days to sift through the various open datasets linked to by the course program, but eventually I stumbled across an interesting dataset on Kaggle that contained detailed shot information for NBA games in the 2014-2015 season.

The Kaggle description states that data was scraped from the NBA’s REST API and, while the data is comprehensive, it does come with some limitations. First, the shot clock variable, which indicates the amount of time remaining for the team with possession to take a shot, contains missing values for 4.3% of the records. Second, the dataset only contains information for 904 unique games (there are 1,230 games played in a complete NBA season). Despite these limitations, the data was very rich in terms of the attributes captured.

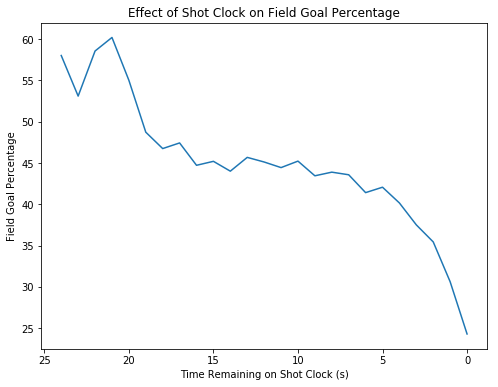
For instance, each record in the dataset contains information on distance to the basket, distance and name of the nearest defender, time remaining on both the shot clock and game clock, the name of the player taking the shot, the outcome of the shot, and even the number of dribbles before the shot was taken. Summary statistics and distribution visualizations of select variables can be found below, as well as in the Jupyter Notebook provided. The sold red line represents the mean and the dotted orange lines represent one standard deviation above/below the mean.

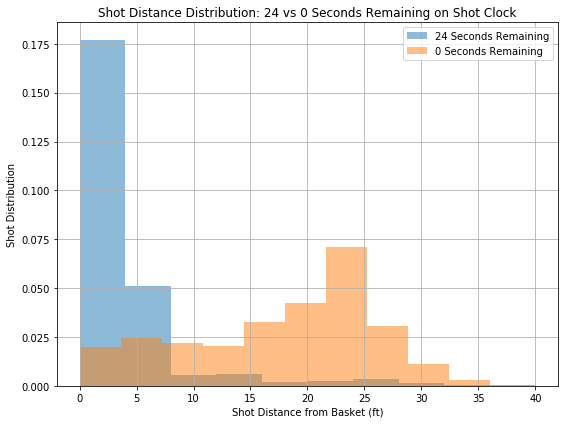
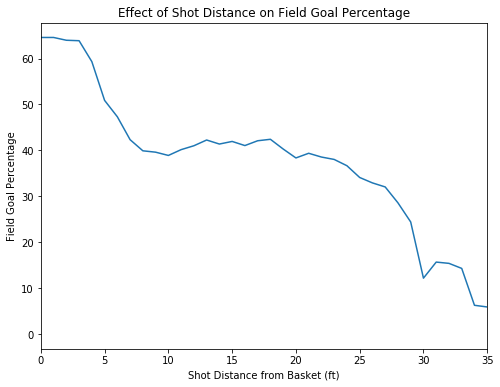




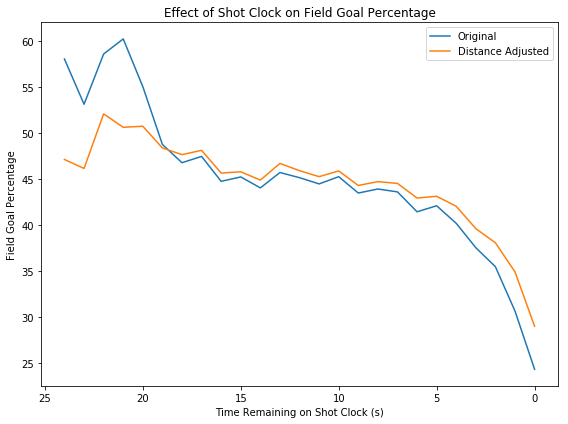
**Analytic question #1: How does time left on the shot clock impact the field goal percentage? In other words, will a rushed shot reduce the likelihood of making a basket, and if so, to what extent?**

To answer this question, I grouped the DataFrame by time remaining on the shot clock (rounded to the nearest second) and plotted the average field goal percentage. The output curve generally resembled what I expected given what I have learned by watching hundreds of NBA games over the years, with two exceptions. First, I did not expect the field goal percentage in the first five seconds of the shot clock to be substantially higher than at any other time in the first twenty seconds. Second, while I did expect a significant drop in field goal percentage in the final five seconds, I was not expecting the drop to be quite so pronounced.



After giving the results some more thought, I hypothesized that the discrepancies between my expectations and the final outcome might be explained by differences in the distributions of the shot distances from the basket. Intuitively, if a team has to force a shot off before the buzzer, it stands to reason that that shot, on average, would be taken further from the basket. Players don’t often pass up opportunities for layups or short jump shots, after all. A simple histogram of shot distances at 24 and 0 seconds remaining shows that my intuition was correct.

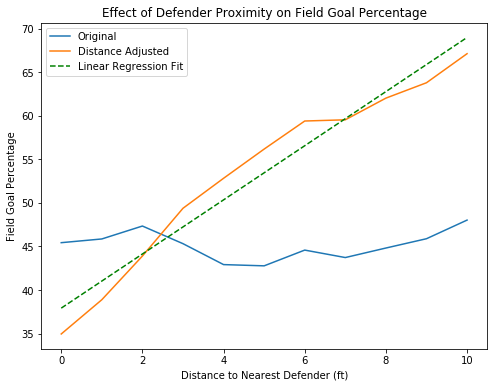
The difference in shot distance distributions, combined with the fact that field goal success rate varies with distance, meant that I needed to normalize the shot clock chart for shot distance. The adjusted results are below.



The curve of the Distance Adjusted field goal percentage effectively isolates the impact of the shot clock, and gives us a good indication of how clock pressure can affect shot outcomes.

**Analytic question #2: How does the proximity of a defender impact the field goal percentage? How much of an increase in field goal percentage can we expect for every additional foot of separation from the defender?**

To answer this question I took a very similar approach to question 1, in that I looked at field goal percentage as a function of my independent variable (in this instance, defender proximity) and then adjusted for shot distance in the same manner. Details on how the distance adjustments were made can be found in the Jupyter Notebook.



The Distance Adjusted curve effectively answers the question of how defender proximity affects the probability of a given shot going in. As one might expect, the more separation a shooter has from a defender, the more likely he is to score. The linear fit gives us an accurate indication of how much of an increase in field goal percentage can we expect for every additional foot of separation from the defender. In this case, for separation distances up to 10 feet, the result turns out to be about a 3.1% increase per foot. It is important to note that one cannot use the linear fit to extrapolate far beyond the distance range shown in this chart, as the field goal success rate would eventually hit 100%. Instead, one would need to use a PolyFit function that more closely approximated the shape of the curve.

**Analytic question #3: Who are the top 10 best and worst defenders in the NBA?**

I believe the best way to answer this question is to look at the field goal percentage grouped by closest defender and, you guessed it, adjusted for shot distance. The methodology is similar to that discussed above, and more details can be found in the Jupyter Notebook.



For the purpose of this analysis, I only considered defenders with at least 200 shots taken against them. I also subtracted the league average defender field goal percentage from the adjusted percentage that I calculated for a more intuitive visualization. The best defenders therefore have negative field goal allowed percentages in this chart, while the worst defenders have positive percentages. The outcome of this analysis generally aligned with my expectations but with the notable exception of Kawhi Leonard, who won the NBA’s Defensive Player of the Year (DPOY) award in the 2014-2015 season.

I can only speculate as to why this happened, aside from the fact that field goal percentage is only one of the metrics considered in DPOY evaluations (steals and blocks are also considered, as well as the team’s overall performance). It is possible that Kawhi is such a good defender that he is sometimes asked to play help coverage on a player other than his primary coverage; i.e. he may be running around to defend open shooters and create double-teams which may leave his primary coverage open. The net result, in theory, would be that the team’s overall defensive field goal rate would improve, but Kawhi’s individual statistics would suffer.

**Further Research**

**Can we use this data to improve predictions of outcomes of future games, relative to the expected point spread?**

This is likely an extraordinarily difficult question to answer, and I suspect that the randomness of both shot outcomes and game-flow in the NBA would severely restrict the accuracy of any type of model I could construct. However, even a 1% increase in prediction accuracy relative to the expected outcome (the Las Vegas point spread, for example) is an enormous advantage in the long term, and so I think this type of analysis bears further consideration.

To start to answer this question, I would focus my attention on games where the final outcome differed from the expected outcome by a significant margin (this would require merging with additional datasets). This type of game outcome would serve as the dependent variable in the model(s). At this point, I could get a sense of whether there are certain teams that are consistently overrated or underrated.

A viable model would probably have to be team specific, though, as each team has specific strengths and weaknesses. However, given that each team plays only 82 games in a season and the fact that this dataset does not contain all 82 games, there could be a very limited set of positive outcomes to train the model(s) on. In the event that team specific models are not viable, I would consolidate everything into one model and see if it offered any insights. These are the types of skills I hope to learn later in the course.