**Obtaining the Data**

Shooting data was obtained by web scraping nba.com/stats. This was achieved using python. I was able to get details about every shot attempted in a preseason, exhibition, regular season, or postseason game dating back to the beginning of the 1996-1997 season. I had the following information about every shot that was attempted: the id and name of the player who attempted the shot, a description of the type of shot attempted (pull-up jumper, floater, etc.), the home and away team, the time left on the game clock (minutes, seconds, quarter), the zone the shot was attempted in (paint, mid-range, above the break 3, corner 3, etc.), the distance the shot was attempted from in feet, x and y coordinates that could be used to plot the shot, and whether the shot was made or missed.

Additional shooting data that was useful to my project was obtained using the r wrapper “nbastatr,” created by Alex Bresler. It is available to install through R. Using the functions from the wrapper, data on shooting efficiency based on the number of dribbles a player took before each shot was gathered, how contested each shot was, team rosters for each season, traditional and advanced box scores from every game completed game, play by play information for every game, and player bios. I also gathered data from synergy sports technology that breaks every shot into one of eleven categories: transition, isolation, pick and roll ball handler, pick and roll roll man, postup, spotup, handoff, cut, off of screen, off rebound, and miscellaneous. Defensive data was also obtained from synergy which was most helpful in evaluating the defensive ability of individual players.

Useful data for evaluating the playmaking ability of each player was scraped from pbpstats.com. Every assist a player makes is broken down into the zone the recipient scores in (i.e. at the rim, corner 3, etc.). Turnovers are also recorded by dead vs. live ball, and whether it was a “bad pass.” The number of possessions each player was in the game for was also obtained; this statistic was very helpful when determining the total usage of each player when he was on the floor.

While some of these metrics are subjective, such as what type of shot was attempted (made shots are more likely to have a vivid description), a majority of the data that was collected was concrete. When cleaning the data to prepare it for cluster analysis, it was important to me to not depend on the subjective information that was recorded by another human. While the player tracking data is unreliable to an extent, it is the best measure we can obtain.