Final Data Memo

Data Science 301-3

Andrew Fenichel

## Data Source

<https://data.world/rezabr/nba-shot-logs-2015/workspace/file?filename=shot_logs.csv>

My primary dataset comes from Reza Brianca on Data World. The dataset, titled “NBA Shot Logs 2015” is a complete dataset of every single NBA field goal attempt from the 2014-15 NBA regular season. It consists of 21 columns and 128,069 rows.

Highlighted colums include my response variable (shot\_result), which I will recode to binary, game descriptions (final score margin, teams playing, period, time remaining, etc), as well as distance between the shooter and the nearest defender, as well as the name of the shooter and nearest defender. While I do not initially intend on joining this dataset with any other datasets, please see my response for “Potential Data Issues” below for a scenario in which I may explore adding data for additional predictor variables, and what that data would be.

## Why this Dataset

My question of interest is "what variables are best at predicting whether a field goal is made or missed in the NBA? Using my basketball knowledge, my hypothesis would be that distance from the closest defender, distace from the basket, and time remaining on the shot clock. I am interested in prediction, and my data definitely supports my classification question, as my predicted varaible would be a binary variable already included in the data set (shot\_result).

Other reasos for “why this dataset” include the fact that it is extremely large (over 128k observations), which will allow me to create training ad test sets in addition to an EDA data set, if I choose to split my data that way, as well as the fact that it has a lot of highly advanced and uncommon NBA data, including modernized “player tracking” data.

## Potential Data Issues

I would consider this to be a mostly clean ad tidy dataset, so I do not imagine having any issues with the data set itself. I will, however, most likely have to recode the shot and game clock variables, as they are not conduscive to predictive analysis under their current format.

One potential issue I could have with my project is not havig enough potential variables for predicting made/missed field goals in the NBA–the data set only includes 21 columns, ad many of the columns are unrelated/ identification numbers. If this is the case, then I will most likely utilize another data set of NBA player statistics from the 2015 season to see if player ability has a affect on made/ missed field goal odds (it will).