Jeffrey Ebert
Capstone Project : Data Wrangling

The data used for this project includes all scores from the 2014-2015, 2015-2016, 2016-2017, and 2017-2018 NBA regular seasons, along with statistics from these games and the betting lines for each game. The game scores were imported via a web scraper, acquiring all of the teams and scores from ESPN.com during that time frame. Cleaning this dataset was important and needed thorough looking over, as I had to locate the errors in the list of team names and game dates that were caused by numerous cancelled games during these seasons and delete them. I also had to search through the lists of over 9000 games played that included two additional teams being listed as a result of inconsistencies within the web pages. After configuring this dataset, it was important to structure the date and times of these games as being in EST, in order to merge them with my dataframes of their game statistics and betting lines. All of the game stats were imported via a separate Kaggle dataset that included the stats for each team playing in all games from these four seasons. From there, I calculated the rolling average stats for each team every season, to be able to evaluate how a team's previous statistics and their opponent's statistics can have value in predicting a win or loss in a game. All of the betting lines were imported from sportsdatabase.com as CSV files. This was the data used to determine if a team is a favorite or underdog for a game, and by how many points oddsmakers are figuring they should win or lose by. These values are the ones that sports bettors can make wagers on throughout the season. Using this baseline data, I created many additional columns to my dataset of probable variables that could be important in predicting whether or a team will win or lose. This includes calculating points margins, representing how many points a team wins or loses by, per game and aggregated out for the whole season, the amount of rest each team has between games, measures of a team's recent performance, including their winning percentages in the last ten games they played that season, whether they are on a winning streak or losing streak and how long it is, how many consecutive games they have played at home or on the road and so on. Between information that is known before the season begins, data that arises as the season goes on and even within each game, there is no shortage of variables for possible reasons why a team may win or lose the game, or perform better or worse than the line or spread of the game indicates. Because there is so much data

available, I have just analyzed a subset of information that could be of value, and there is always more data that tells the complete story.

There are no null values in the original data collected because all games during each season were played and there is a listed final score and statistics for each game. In data formed during analysis there were null values created in certain data rows for columns such as average statistics of a team where a team has not yet played a game, or for a team's last 10 game performance, where they have not played the 10 games required during a season to have a value for this data point. When looking at information based on these columns these data points are simply ignored and the focus is on the games with more in-season data available. There are naturally outliers in game scores, but these values are not treated any differently in the analysis than an average scoring game because it is valuable to use the outliers to evaluate the differences in how the game went and understand why a team performed so much better or worse than usual.