**# Data Preprocessing**

We initially started looking at the data for salary of NBA players at <https://www.basketball-reference.com/contracts/players.html> (updates constantly), which had 582 records of player salaries for year 2017-2018 at the time. However, we had to remove some duplicated records for players with different salaries on different teams. This is because some players could get cut by teams half way through the season, and sometimes they would get picked up by another team, which resulted in having multiple player contracts in a year. An example for this is Rajon Rondo, who was waived by the Chicago Bulls, signed a contract with New Orleans Pelicans right after.

During the process of matching players’ statistics with their salaries, we encountered some cases where some the player information for a couple of players listed in our salary could not be found. An example is Walt Lemon, Jr., who is initially listed in our salary data. We were not able to find his player information on <https://stats.nba.com/players/bio/>, which contains data that we thought could be important in our analysis. Therefore, we removed these records.

For the “Position” categorical variable in our dataset, we stated in our proposal that we would be using 5 values, PG (Point Guard), SG (Shooting Guard), SF (Small Forward), PF (Power Forward), and C (Center). It turns out that many guards in the NBA today are “combo guards”, which means they can both play at the Point Guard and Shooting Guard position (e.g. James Harden). There are also many forwards in the NBA who can both play at the Small Forward and Power Forward position (e.g. Lebron James). We reduced the number of values to 3, grouping PG and SG as G (Guard), SF and PF as F (Forward). In addition, there are some players who are “swingman”, meaning they can both play at the SG and SF position (e.g. Jimmy Butler). Since this is not a frequent case, we chose a position for each of them based on which position they had mostly been playing at this season (2017-2018) and our knowledge to the players.

Our eventually obtained our final dataset, which contains 515 records and does not contain any N/A’s.

We then realized a couple outliers in our dataset. For example, Gordon Hayward was horribly injured during his very first game at the beginning of the year. He was not able to return for the rest of the season. With the 4th highest salary on our list, he would be an extreme outlier in our models with minimal statistical contribution. However, this does not mean that he is not worth the salary, since he was only able to play for about 5 minutes before the injury. Therefore, we would like to exclude him when building our models, along with several other players in similar conditions.

/newPage