What kind of cleaning steps did you perform?

In dealing with player positions, PF, SF, SG, PG, C are the 5 designated positions for professional players, but a very small portion of players played a role somewhere in between two different positions, so that they were label with 2 positions such as SF-SG, C-PF, SG-PG… due to lack of information of which role these players were more inclined to play, I removed those player information who played double roles.

For the stats that represent total values (others, as TS%, represent percentages), I transformed them into values per 36 minutes. The reason is to judge every player according to his characteristics, not the time he was on the floor.

How did you deal with missing values, if any?

Delete blank columns and rows.

There are significant missing values in 3P (3-Point Field Goals), 3PA (3-Point Field Goal Attempts), 3P% (3-Point Field Goal Percentage), 3PAr(3-Point Attempt Rate). It's because 3-point line was not introduced to NBA until 1979, but for the sake of classifying modern players, I have to assume all players before 1979 have not attempted 3-pointers, so I will fill the missing values of these three columns with 0.

For the missing values in GS (Games Started), since number of Games started for each player does not effect their performance measuring, I will delete the GS column.

Missing values also appeared in ORB, DRB, STL, BLK, TOV and their associated percentages (ORB%, DRB%, STL%, BLK%, TOV%), these measurements are not included in record before 1974, and TOV data are not present until 1978, These missing values cannot be filled with assumption, but at the same time these columns are too important to be removed from the construction of player classification, therefore all the data before 1978 have to be discarded.

Were there outliers, last how did you handle them?

When using rate statistics (i.e. points per game) or cumulative statistics (i.e. total points) can be misleading when it comes to analysis because these statistics tend to inflate players with lengthier careers. To deal with outliers, I instituted a minimum threshold of 40 games played. Also I keep only players with more than 400 minutes for each season (with a 82 games regular season, thats around 5 minutes per game. Players with less than that will be only anecdotical, and will distort the analysis).