NBA Salary Prediction

Brandon Arcilla

Problem:

NBA Free Agency can be an exciting time for teams, players, and fans. Players want to know which teams want them and teams want to know which players can help. An important factor for both parties is determining salary.I want to be able to predict a salary range for players based off advanced NBA statistics. NBA teams could use this to help gauge how much a player should be making or which players fit into their salary situation. An NBA Player could use this to see what they could potentially be earning based off of their production on the court.

Dataset:

Description of data obtained:

Data	**Description**
Player Name	The name of the NBA player
Player ID	The unique identifier used by the website
Pos	NBA Position: Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF), Center (C)
Age	Age of Player at the start of February 1st of that season.
Tm	Team that the NBA player belonged to
G	Total # of games played
MP	Total # of minutes played
PER	Player Efficiency Rating: A measure of per-minute production standardized such that the league average is 15.
TS%	True Shooting Percentage: A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.
3PAr	3-Point Attempt Rate: Percentage of FG Attempts from 3-Point Range
FTr	Free Throw Attempt Rate: Number of FT Attempts Per FG Attempt
ORB%	Offensive Rebound Percentage:An estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor.

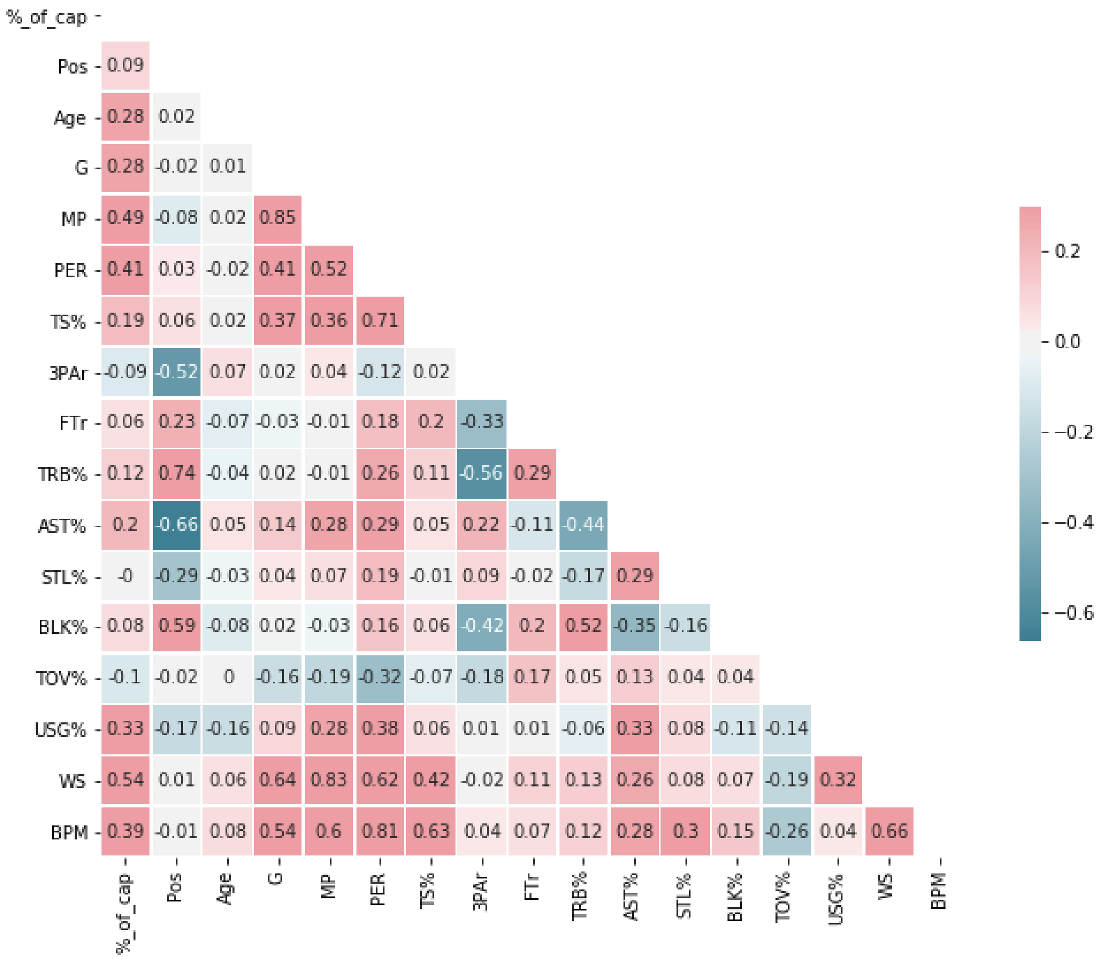
DRB%	Defensive Rebound Percentage: An estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor.
TRB%	Total Rebound Percentage: An estimate of the percentage of available rebounds a player grabbed while he was on the floor.
AST%	Assist Percentage: An estimate of the percentage of teammate field goals a player assisted while he was on the floor.
STL%	Steal Percentage: An estimate of the percentage of opponent possessions that end with a steal by the player while he was on the floor.
BLK%	Block Percentage: An estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor.
TOV%	Turnover Percentage: An estimate of turnovers committed per 100 plays.
USG%	Usage Percentage: An estimate of the percentage of team plays used by a player while he was on the floor.
OWS	Offensive Win Shares: An estimate of the number of wins contributed by a player due to his offense.
DWS	Defensive Win Shares: An estimate of the number of wins contributed by a player due to his defense.
WS	Win Shares: An estimate of the number of wins contributed by a player.
WS/48	Win Shares Per 48 Minutes: An estimate of the number of wins contributed by a player per 48 minutes (league average is approximately .100)
OBPM	Offensive Box Plus/Minus: A box score estimate of the offensive points per 100 possessions a player contributed above a league-average player, translated to an average team.
DBPM	Defensive Box Plus/Minus: A box score estimate of the defensive points per 100 possessions a player contributed above a league-average player, translated to an average team.
BPM	Box Plus/Minus: A box score estimate of the points per 100 possessions a player contributed above a league-average player, translated to an average team.
Salary	Salary for player during that year
clean_Salary	Processed salary removing decimal and $
%_of_cap	Percentage of clean_Salary against the team salary cap
Season	Year that was played

- Data sources
    - NBA Salary (1985-2017) from data.world: This data set contained a player_id, which acted as the unique identifier. This dataset included salary and year for each player. Not much cleaning was done on this data set.
    - Team Salary Cap (1997-2017) from data.world: Since inflation is not taken into account for player salary this data is used to normalize the players salary to a percentage of the salary cap for each year. Team Salary Cap is the maximum amount of money a team could use on their roster. It should be noted that there are exceptions where teams can exceed the cap amount, either by exceptions based on player contract or by paying a luxury tax. This type of data is not being considered in this project.
    - NBA Advanced Stats (1997-2017) from www.basketball-reference.com**:** BeautifulSoup was used to scrape the website. The following steps were done to wrangle and cleanse the data.
        - To get a list of urls to scrape, I first identified that each url is the same with the exception of the year.
        - A list of years from 1997 to 2017 was created.
        - Then the URL was iterated with each year to get a list of URLs.
        - A function was defined for the scraping (urlScraping).
            - Within this function, BeautifulSoup is used to open the url.
            - Column headers, player_stats, player_id, and year were extracted.
            - A DataFrame is created using headers as columns and player_stats as rows.
            - There were blank columns in the DataFrame that were taken out using isspace().
            - There were also rows with all NaN values that were removed as well.
            - Extracted data player_id and year were added as columns.
            - DataFrame index was reset, then the index column was dropped.
            - Certain columns were converted from Object to Float/Int.
        - A for loop was created to run all url through urlScraping function to save as a csv with matching year.
        - Another for loop was used to read_csv and append all data into an empty list where that list was concatenated to create one list of advanced stats for players from 1997-2017.

Final output of the dataset was done by merging advanced stats to the players salary on the player_id and year. Left join was used because I wanted to see how many players didn’t have any salary in the salary dataset. Players without salary accounted for less than 5% of the total data, so they were excluded.

Exploratory Data Analysis:

Heatmap of Correlation



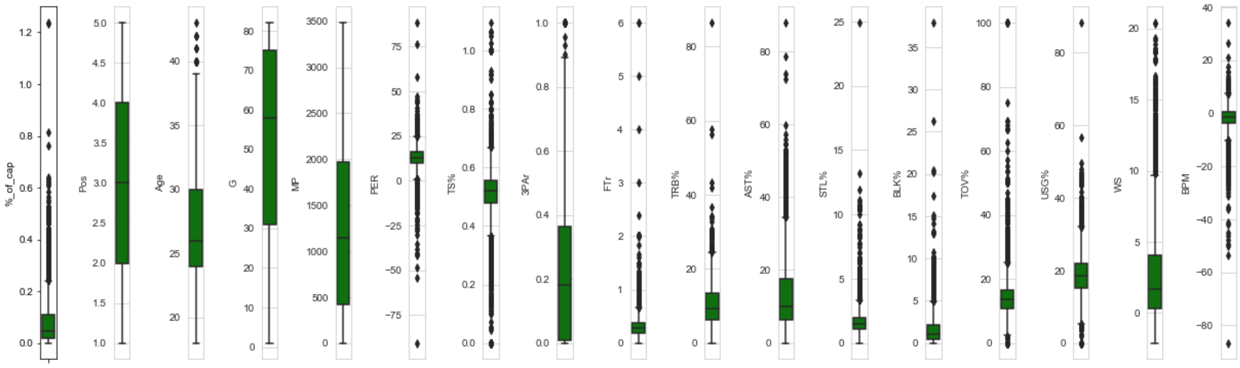
Looking at the correlation between % of Salary Cap against other variables, it can be broken down as:

Moderate Positive Correlation (0.3 to 0.7)	Minutes Played, Player Efficiency Rating, Usage %, Win Shares, Box Plus/Minus
Weak Positive Correlation (0 < c ≤ 0.3)	Position, Age, Games Played, True Shooting %, Free Throw Rate, Total Rebounding %, Assist %, Block %
Weak Negative Correlation (-0.3 ≤ c < 0)	3-Point Attempt Rate, Turnover %

No Correlation (0)	Steal %

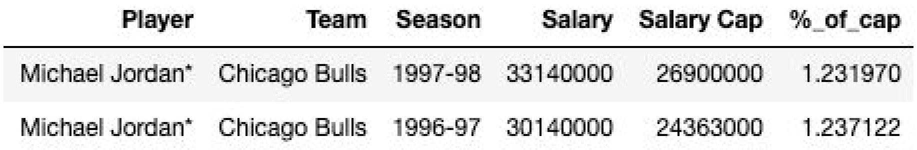
What surprises me the most is that the variables that I believed would be the strongest correlated to % of salary cap actually has a weak positive correlation. It would seem how many minutes played, efficiency rating, usage percentage, win shares, and box plus/minus are more telling for a higher salary %

Outliers



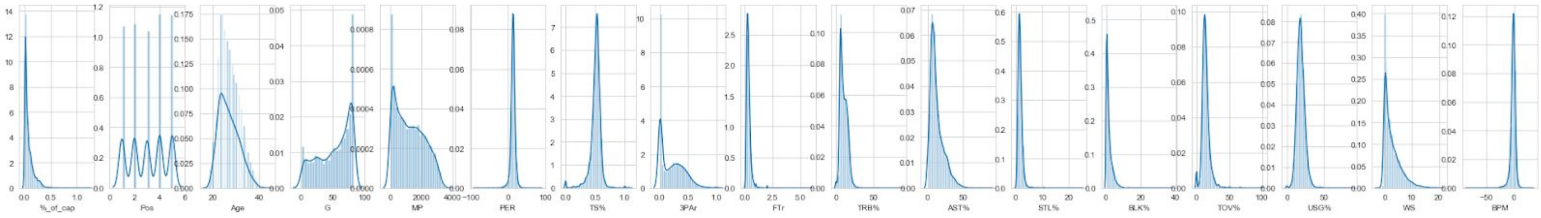
Position, Games Played, Minuted Played are the only variables that do not have outliers. This makes sense because there are no other positions, games are set to a certain amount and minutes played per game is set too.

The outlier for % of salary cap goes to Michael Jordan during the 96-97 and 97-98 NBA Season. During this year there was no maximum salary cap for players. He was making more than the salary cap allowed.



Player	Team	Season	Salary	Salary Cap	%_of_cap
Michael Jordan*	Chicago Bulls	1997-98	33140000	26900000	1.231970
Michael Jordan*	Chicago Bulls	1996-97	30140000	24363000	1.237122

Data Distribution



It looks like the variables are skewed to the right. This makes sense as these variables would never be below 0, except for BPM (Box Plus/Minus).