Stat 6950 Project Proposal

Patrick McHugh and Nick Mandarano

March 31, 2022

Exploratory Data Analysis

We are both interested in basketball, and chose to do an NBA-related project. Surprisingly, we are not the only people interested in basketball, and countless statistical analyses of basketball data have been done on a variety of platforms with a variety of goals. We attempted to come up with an original approach, and are deciding to create a regression model for salary of NBA players, using statistics and other information about the players as covariates.

Salary is known before the season starts and statistics are created, so it is not a response variable in the traditional sense of a causal effect. Our regression project attempts to explore the relationship between salary and the covariates, prescribe a true mean function, and identify players who may be overperforming or underperforming their contract; i.e., putting up better or worse statistics than one might expect a player on their salary to do.

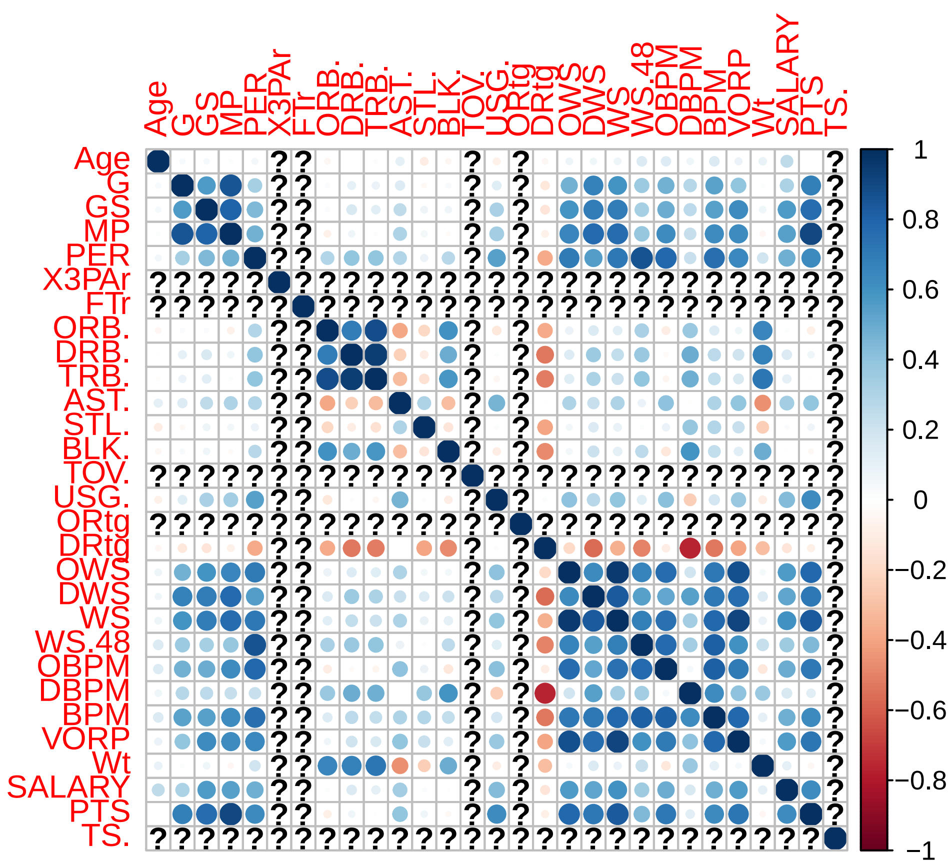
Our data comes from four different datasets. We used three of Riguang Wen's datasets from figshare.com – `players cv`, `players salary`, and `players stat`. We also used a dataset called `NBA RS 2020-1950 Stats` uploaded to zenodo.org by Pablo Gomez and Sandra Giral. From these datasets, we considered the following variables.

Variable	Description	Type	Source
Player	Name of player	Character	`players stat`
Age	Age of player	Numeric	`players stat`
G	Games played	Numeric	`players stat`
GS	Games started	Numeric	`players stat`
MP	Minutes played	Numeric	`players stat`
PER	Player efficiency rating	Numeric	`players stat`
PTS	Points	Numeric	`NBA RS 2020-1950 Stats`
X3PAr	3PA/FGA	Numeric	`players stat`
FTr	FTA/FGA	Numeric	`players stat`
TS	True shooting percentage	Numeric	`NBA RS 2020-1950 Stats`
ORB	Offensive rebounds	Numeric	`players stat`
DRB	Defensive rebounds	Numeric	`players stat`
TRB	Total rebounds	Numeric	`players stat`
AST	Assists	Numeric	`players stat`
STL	Steals	Numeric	`players stat`
BLK	Blocks	Numeric	`players stat`
TOV	Turnovers	Numeric	`players stat`
USG	Usage percentage	Numeric	`players stat`
ORtg	Offensive rating	Numeric	`players stat`
DRtg	Defensive rating	Numeric	`players stat`
OWS	Offensive win shares	Numeric	`players stat`
DWS	Defensive win shares	Numeric	`players stat`

Variable	Description	Type	Source
WS	Win shares	Numeric	`players stat`
WS.48	Win shares per 48 minutes	Numeric	`players stat`
OBPM	Offensive box +/-	Numeric	`players stat`
DBPM	Defensive box +/-	Numeric	`players stat`
BPM	Box +/-	Numeric	`players stat`
VORP	Value over replacement player	Numeric	`players stat`
Pos	Position	Factor	`players salary`
Ht	Height in inches	Numeric	`players salary`
Wt	Weight in pounds	Numeric	`players salary`
PwrSix	Power Six College?	Indicator	`players cv`
International	International Player?	Indicator	`players cv`
Salary	Salary in dollars	Numeric	`players salary`

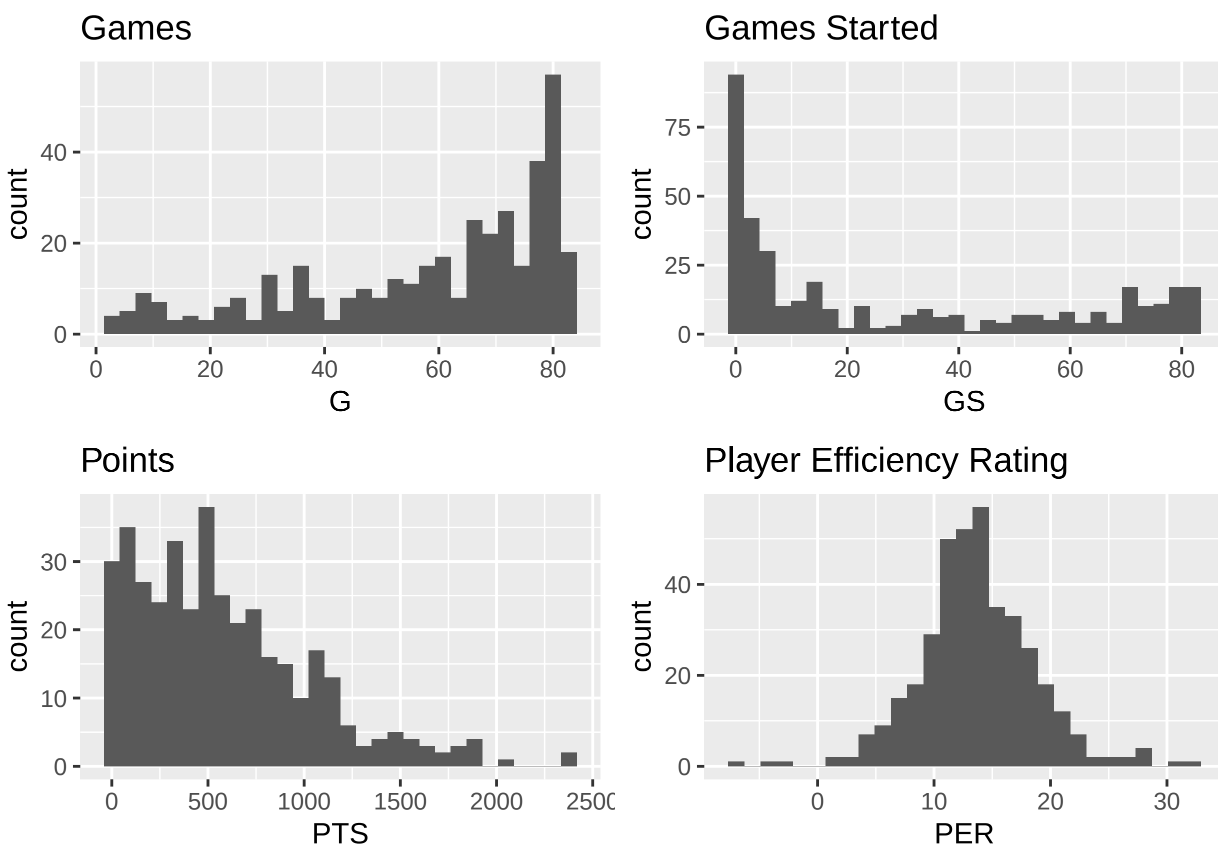
We had access to data for many years of statistics, but chose to use only data from the most recent NBA season in the dataset, 2015-16, to avoid the need for a mixed effects model that would arise from using multiple seasons due to the heavy correlation between the observations from one player across different seasons.

Immediately we can recognize that some variables are functions of others and therefore do not need to be considered. Specifically, `BPM = OBPM + DBPM`, so there is no need to include `BPM` in our model. Similarly, `WS = OWS + DWS` and `TRB = ORB + DRB`, so we can exclude `WS` and `TRB` from consideration if we include `OWS`, `DWS`, `ORB` and `DRB` in our model. Some other multicollinearity issues will likely arise given the correlation matrix of the numerical variables under consideration below. Some examples of potential issues are the correlation between `WS` and `PER` as well as that of `MP` and `G`.



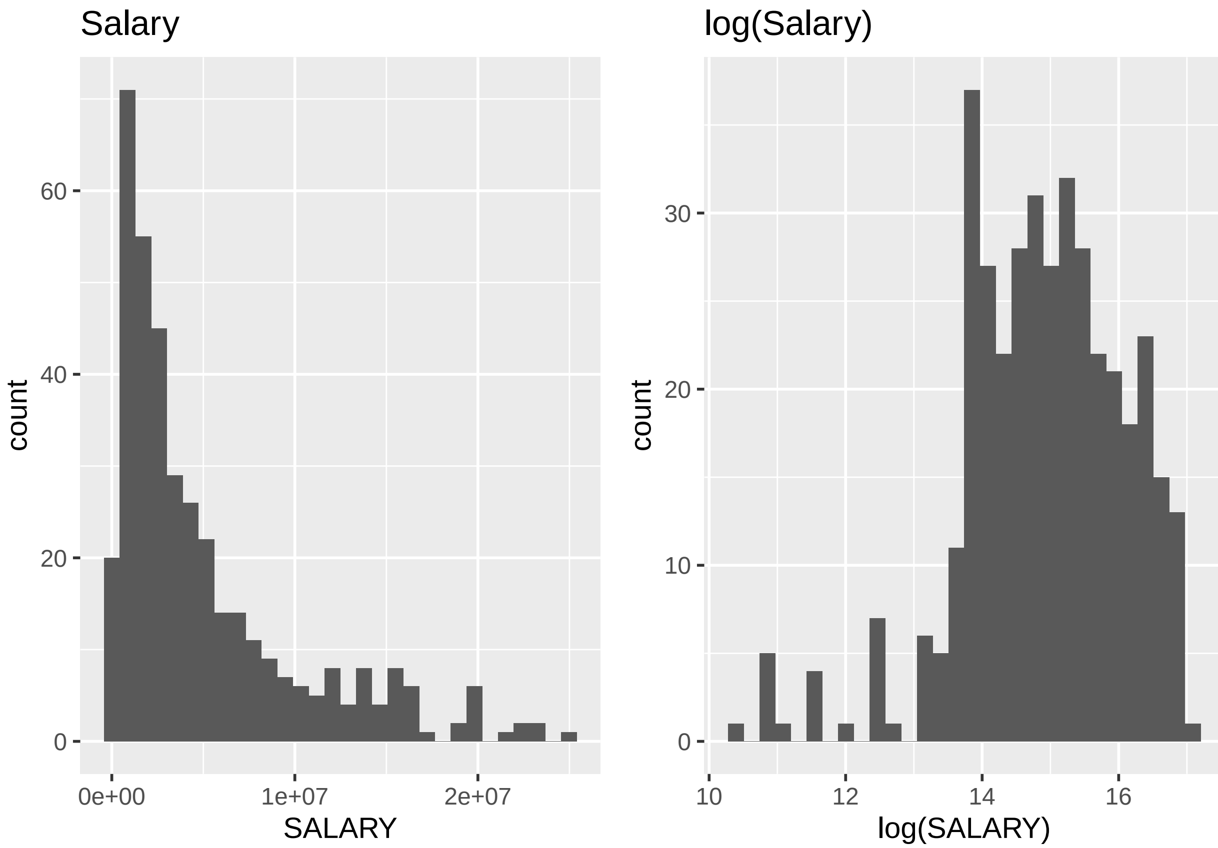
This matrix is an important figure for us and represents a key concept. In many datasets, one would expect some degree of correlation between predictors. Furthermore, many NBA advanced metrics available are different functions of the simpler statistics with many overlapping inputs. These attempt to capture different information, however, we can see right off the bat that there is significant multicollinearity and will need to use this knowledge to avoid unintentionally putting extra weight on some of the same underlying information.

Also in the numeric variables are signs of non-normality. Of the 27 numeric variables considered after the exclusion of `BPM`, `WS` and `TRB`, 11 had medians that had 10% or more in difference of the mean, possibly indicating asymmetry. Of these, only the boxplots of `G` and `GS` did not signify outliers, though histograms of the data did show skewness. Histograms of the others (`FTr, ORB, AST, BLK, OWS, DWS, VORP, Salary`, and `PTS`) were all right-skewed. `GS` is fairly uniformly distributed from 10-82, with a higher density from 0-10. We would prefer approximately normal distributions of the covariates, to help obtain normally distributed residuals, and to avoid high leverage cases affecting the mean function. We will experiment with log, inverse, square root, and squaring transformations. Many covariates do appear to be normally distributed and useful without transformations.



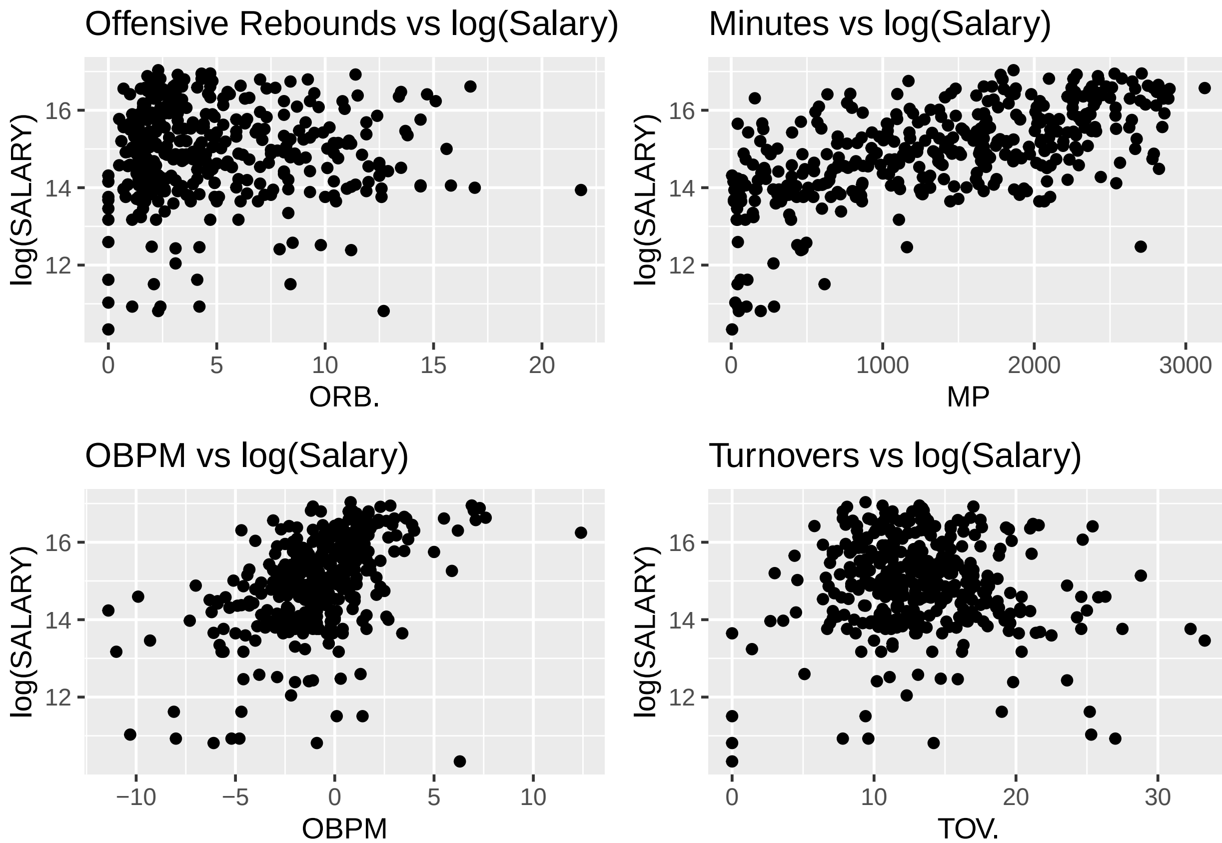
We did make some transformations for categorical variables as well. NBA players are often discussed as either American or International, so we created a new variable labeling each player as one of these based on the place of birth. Additionally, we created the “Power Six” variable to see whether this has any indication; this may not be independent of American/International.

We can see that the distribution of salary is heavily skewed to the right. We expect to transform this variable to perform linear regression. After attempting several transformations such as an inverse and square root, a log transformation seems most appropriate, althought not perfect. We will consider other transformations and the Box-Cox method during the analysis:

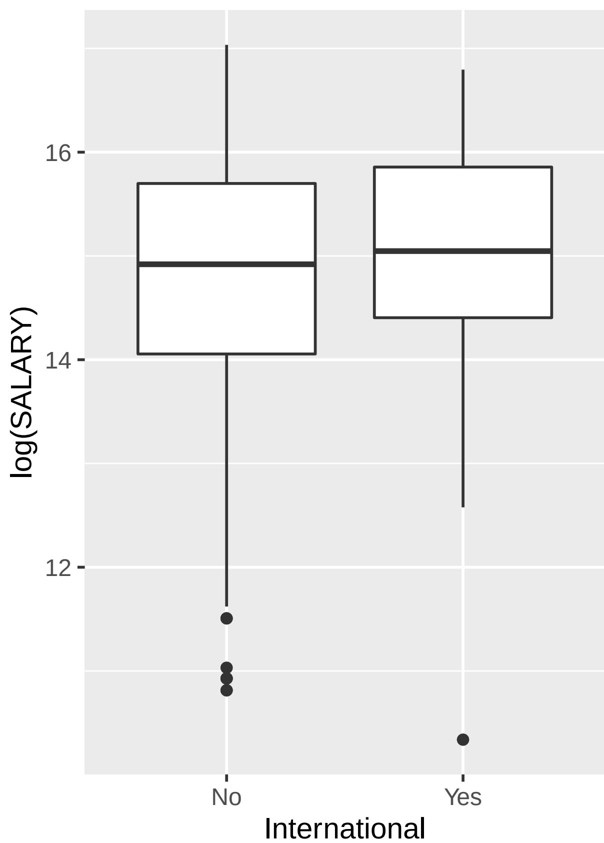


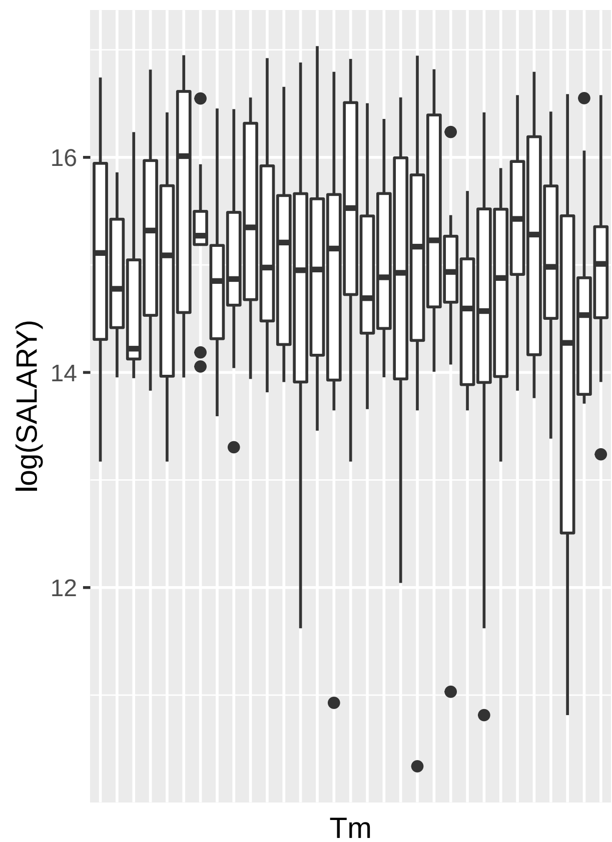
We also look at the relationships between salary and each of the predictors individually. There appears to be a relationship between salary and many of the covariates individually, including simple stats such as `MP` and `PTS`, as well as advanced stats like `PER` and `BPM`. There are many covariates that have a marginal relationship with salary. These relationships are primarily positive, indicating an increase in most stats such as points correlates with an increase in salary, but a small number, such as turnovers, may be negative. For some of them such as minutes, a marginal linear relationship seems appropriate; for others, such as `VORP`, there appears to be a marginal relationship that is not linear. As discussed earlier, trying to transform variables and account for the multicollinearity in covariates will be some of the challenges of this project.

Some covariates, such as `ORB`, do not appear to have a strong marginal relationship with salary; we will investigate whether these still may have a relationship with salary through interactions with other variables.



After plotting boxplots of salary by each level of the categorical variables, salary does seem to vary across different levels of the variables. We plan to evaluate whether these relationships remain useful in the full model. Position and international each have a small number of levels; the team factor has 30 levels. We will evaluate whether this can be useful with all 30 levels, if there are ways to reduce this by grouping teams by things such as conference affiliation or market size, or if it is not useful at all in a model with less than 400 observations.





This dataset is pretty broad, which leaves us open to many possibilities for modeling approaches. As mentioned before, it seems likely we will need to transform the y variable in some way, possibly with a Box-Cox approach. We are certainly dealing with some multicollinearity in covariates and will need to use tools such as AVPs and VIFs to account for this. Also, interaction effects seem plausible; for example, would a change in the number of rebounds per game be associated with the same change in salary for both guards and centers? We will look at interactions between both types of covariates (numerical and categorical). We do not initially expect to need any weighted regression or time series.