Final Project: NBA Hall of Fame

Eungkoo Kahng, Ashley Lu, Salih Yasun, Gokcen Buyukbas Cakar

11/27/2019

Executive Summary & Abstract

Are current NBA superstars like Giannis Antetokoumpo and Steph Curry hall of fame locks? Maybe it's little too early to say so since they are still in their prime and have long way to go until their retirement. However, no one can ever really argue they are going to end up getting into the Hall of Fame due to their dominance in the league for recent years. Perhaps then, NBA fans might wonder if we could predict some particular active players' chances of getting into the HOF in the near future after their retirement.

From this analysis, we would like to first predict the future Hall of Famers in two different approaches, versatility and shooting ability. Ultimately, we will test and compare the accuracy of the two approaches' performances to determine the final best model. After the analysis, we confirm that "stats-stuffing" seems to matter more and more if a player were to be inducted to HOF in this modern era basketball, meaning that you have to be versatile all around in terms of skill sets to a hall of fame caliber player. Regarding shooting ability, it seems like Steph Curry's effort to changing the league paradigm where 3 point shots would dominate the league seems to take more years to affect the HOF standards at least for now.

Research Questions

The first research question regarding the versatility approach is **What are the most important skill stats to predict the future of famers?** As modern-era basketball has been becoming more of a "Stats-Stuffing" battle regardless of positions (players like Russell Westbrook, Lebron James, and Giannis Antetokoumpo), the standard and definition of decent players have become versatility all around. Thus, we would look for what combination of skillsets seems to be the most appropriate indicator to predict the future hall of famers. The other research question is **No matter how versatile a player is all around, can he just simply be great enough to be in the HOF if you are a pure sharpshooter?**. This separation of the two approaches to determine a player's greatness seems valid as most experts and hoopers would agree you are either a shooter or an all-around "fundamental guy" in a game of basketball. Ultimately, our main research question narrows down to **Which approach between the two provides more accurate prediction for the future hall of famers?**

Data Description

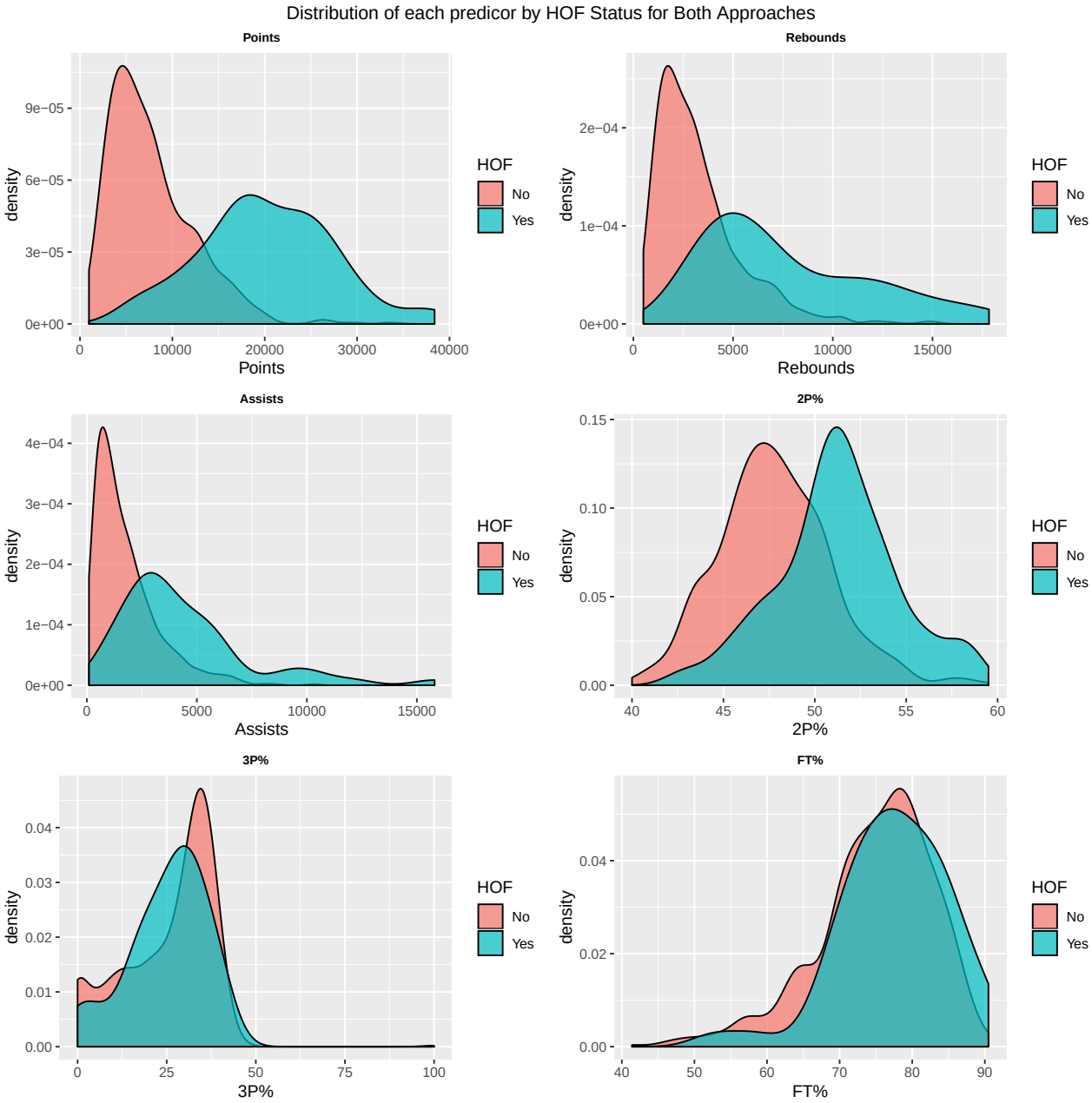
In order to collect a desired dataset to answer the research questions, we extracted about 700 retired players' career stats from the **basketball-reference** website. This website provides a very up to date data which is updated every night whenever there happens to be a game. One of the best features of this website is that it allows us to select players of interest with easy filtering options such as Allstar/Non-Allstar, league era, and HOF/Non-HoF. In this dataset, we have 50 non-HOF players and 642 HOF players. Since shooting ability is one major aspect in this analysis, we only included players who played in the 3-point shot era which was introduced in 1979 We set the termination time point with year 2016. This is because the players who retired by then are the most recent possible candidates who just became eligible for the upcoming 2020 HOF induction.

For the variables in the dataset, we have the status of HOF as our binary response, games played, minutes played, teams, points, rebounds, assists, turnovers, and etc... which makes a total of 33 variables. Within

this analysis, we are only using **Points**, **Rebounds**, and **Assists (later categorized into 3 values for our modeling purpose)** for the versatility approach and **2P%**, **3P%**, and **FT%** for the shooting ability approach.

Exploratory Data Analysis

Univariate Analysis for both Approaches

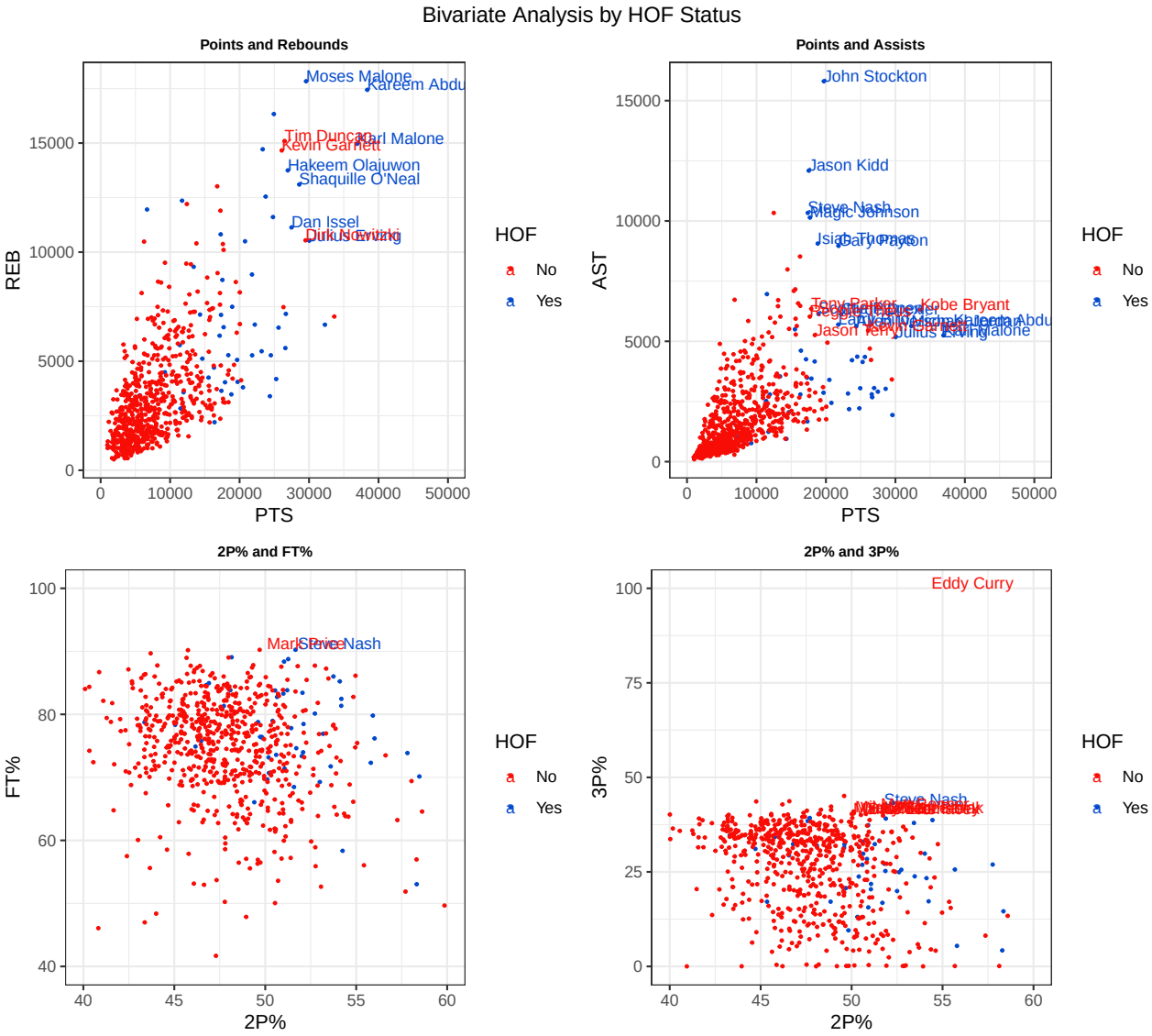


From the ggpairs plot in appendix 1, each predictor variable for the versatility approach is fairly highly correlated to the HOF response, whereas each predictor for shooting ability is relatively a lot less correlated with the response. From the correlation coefficients, **points** and **rebounds** are the two most strongly correlated predictors with **HOF** status for versatility, whereas **2P%** and **FT%** are the two most correlated

predictors with HOF status even though **FT%** is still very low in terms of the magnitude of the correlation. These findings lead us to build a blueprint for modeling the fits later in this analysis. From the plot above, we found the main difference between the two approaches in terms of the differences of distributions is that the skewness differs with respect to the status of **HOF**. It is clear that the predictors for non-HOF are all right-skewed, whereas they are pretty close to a normal distribution for HOF. This makes a perfect sense that HOF players exceptionally stuffed their stats sheets compared to the non-HOF players. On the other hand, shooting ability approach indicates that no matter if you are a HOF caliber player or not, players' shooting ability evenly varies because shooting is the aspect in the game of basketball that is a pretty unpredictable day in and day out.

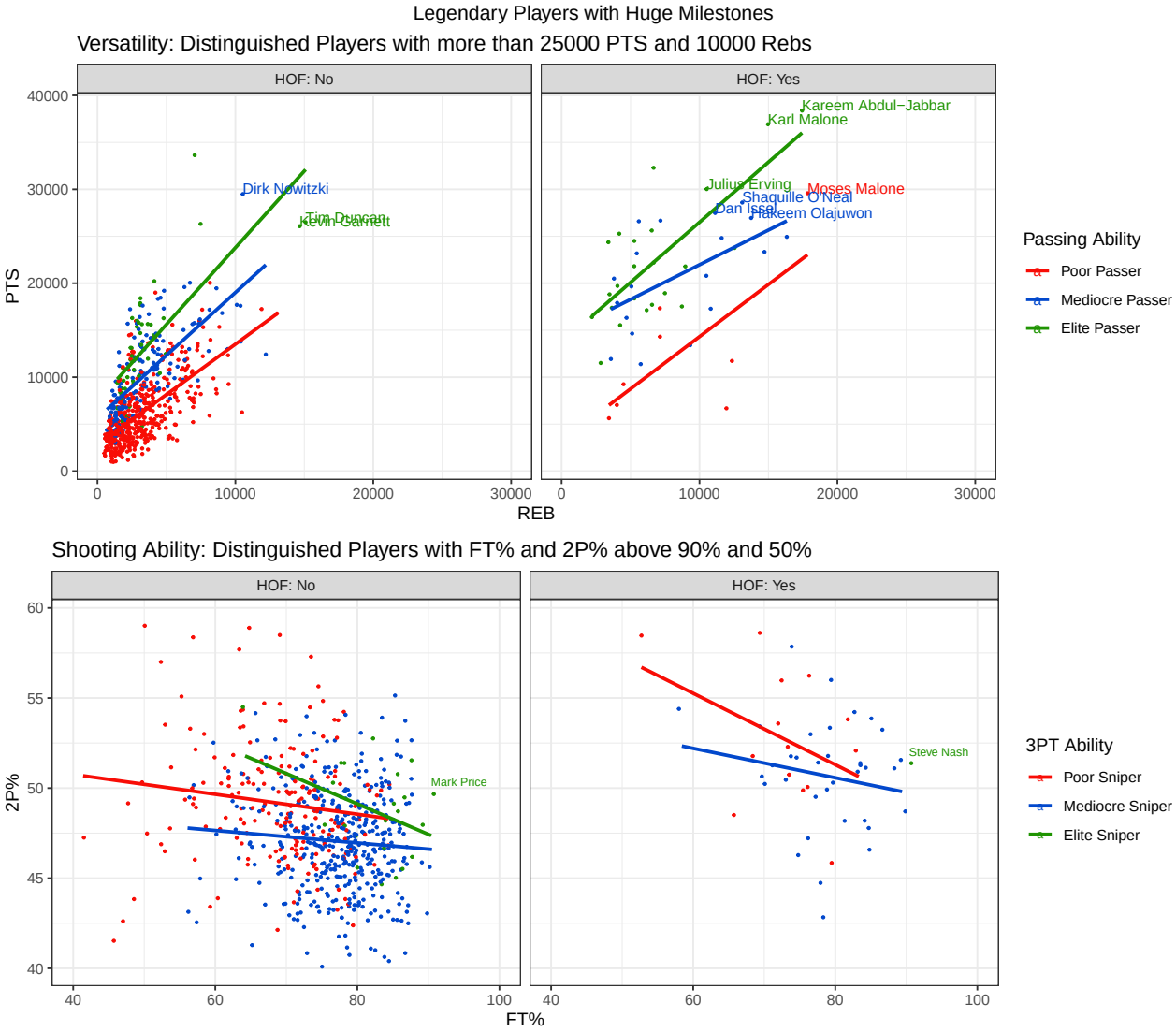
From appendix 2 and 3, we checked the monotonic relationships between the response and each predictor variable for each approach. It seems that none of the log transformations that we originally thought would help is not significantly helpful in terms of forming monotonic relationships between the response and predictors. In other words, all 6 predictors are doing more and less good job shaping monotonic relationships, which lead us to go with non-transformation later when we proceed with building models.

Bivariate Analysis for both Approaches



Here we are exploring some of the noticeable and interesting bivariate relationships between predictors by each status of HOF by displaying some of the legendary players' names who put up gigantic stats in their career. For points and rebounds, we spot some HOF legends like **Kareem Abdul-Jabbar**, **Karl Malone**, and **Shaquille O'Neal**. Actually, Kareem is the all time leading scorer in NBA history followed by Karl Marlone who is the second in the category. For the first ballot Hall of famers in the near future who just became eligible for 2020 HOF induction, we see **Tim Duncan**, **Kevin Garnett**, and **Dirk Nowitzki**. These players are a huge factor that would affect our model fits later because their numbers are enormous but still classified as non-HOF players in our dataset at this point. Thus, we should remove these future hall of famers before we fit the models. For assists and points, we spot legendary point guards such as **John Stockton** and **Jason Kidd** who lead the all time assists category in NBA history. For shooting ability, overall, we do not see linear relationships as we've seen in the versatility approach. For 2P% and FT%, we spot **Mark Price** and **Steve Nash** as Non-HOF and HOF players, respectively, whose shooting percentages are off the chart in their career. For 2P% and 3P%, we observe a very outlier-type player, whose 3P shooting was 100%. This player happens to be **Eddy Curry** who played center position in the league for 8 years. He had two total attempts for 3P shots and knocked down both in his entire career.

Trivariate Analysis



Ultimately, we present the trivariate visualization that captures all 3 predictors for each approach at once. For each approach, we added passing ability (assists) and 3P ability (3P%) as categories this time as they are more meaningful and easier to interpret to divide the players in such way and that they are the least correlated predictors for each approach. We added linear smooth lines for each status but they are not very meaningful because we have a huge discrepancy for the number of data points for each status of HOF. For versatility, we demonstrate some of the all time greatest such as Kareem, Shaq, Julius Irving, Duncan, and Kevin Garnett once again with a huge milestone of reaching over 25000 points and 10000 boards. We are able to notice that not all of them is an elite passer. For shooting ability, once again, we spot Steve Nash and Mark Price who are the only two retired players in NBA history that are in the "180 club", meaning that they used to be so purely talented shooters who had over 50% for 2P%, 40% from downtown, and 90% from the free throw line.

Fitting Models

Removing those "Usual Suspects": Dirk Nowitzki, Kevin Garnett, Kobe Bryant, Tim Duncan, Paul Pierce

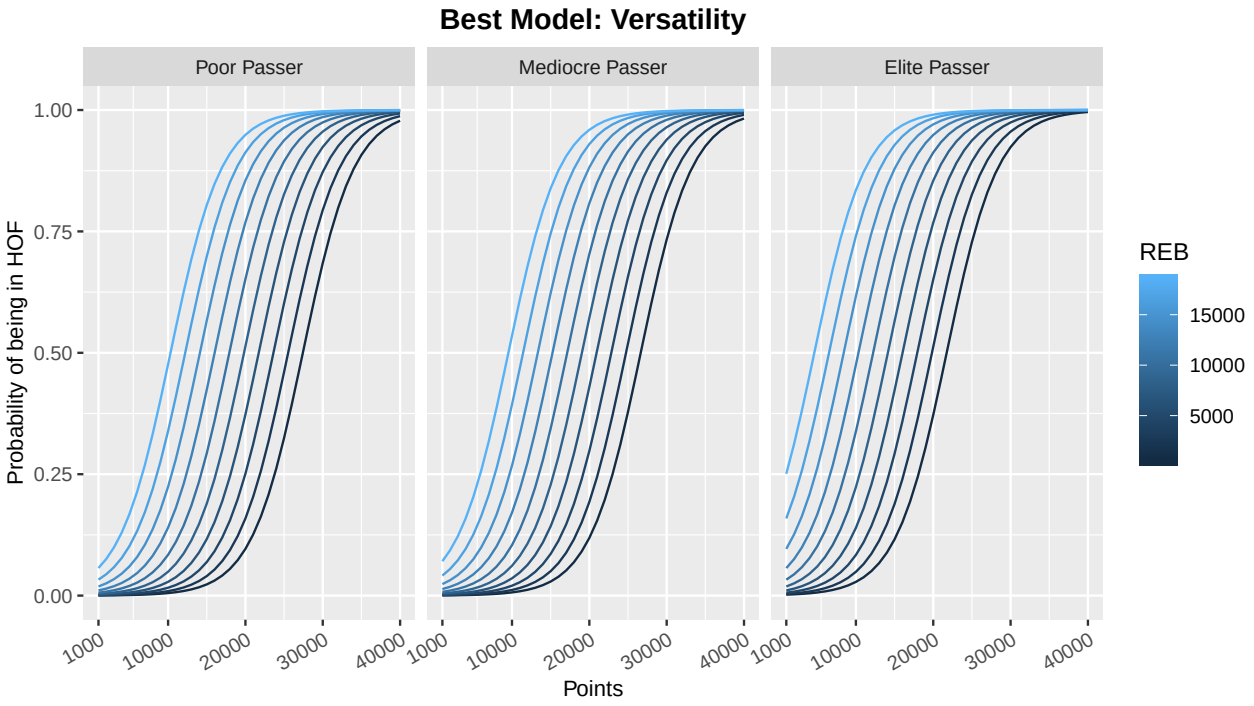
```
#Versatility (Approach 1)
# two most strongly correlated to response
mod1 = glm(HOF ~ PTS + REB, family = binomial, data = nba.modified)
# interaction between the two variables
mod2 = glm(HOF ~ PTS * REB, family = binomial, data = nba.modified)
# all 3 variables without interaction
mod3 = glm(HOF ~ PTS + REB + AST.cat, family = binomial, data = nba.modified)

#Shooting Ability (Approach 2)
# two most strongly correlated to response
mod4 = glm(HOF ~ `2P%` + `FT%`, family = binomial, data = nba.modified)
# interaction between the two variables
mod5 = glm(HOF ~ `2P%` * `FT%`, family = binomial, data = nba.modified)
# all 3 variables without interaction
mod6 = glm(HOF ~ `2P%` + `FT%` + `3P%.cat`, family = binomial, data = nba.modified)
```

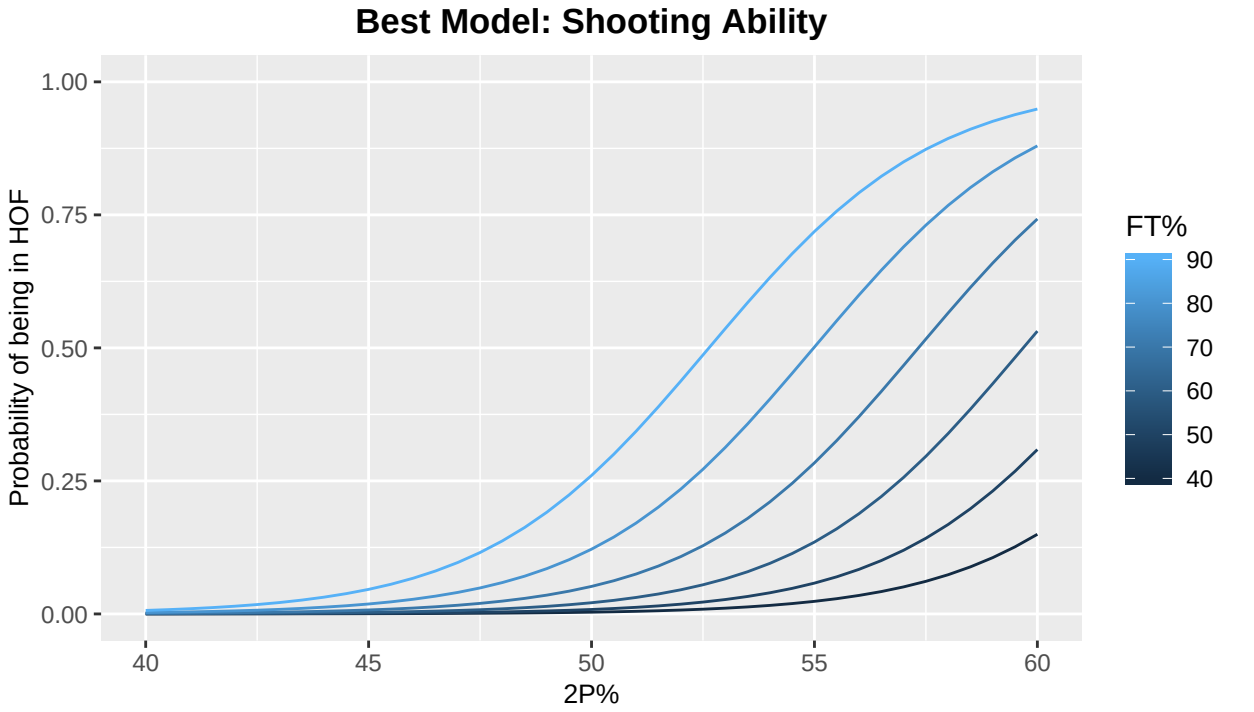
For our modeling, we come up with 3 models for each approach based on the findings from the EDA section plus our basketball knowledge & common sense. For each approach, we start with the two most strongly correlated predictors with the response with and without interaction. Then, we also add all three predictors without any interaction just to detect pure main effects. The main rationale for this is that it is obvious that players who are in the league for longer must get more career total stats than those who are not although a very few exceptions can exist. Thus, we do not look for any 2-way or 3-way interaction between the three predictors for each approach.

At this stage, we still want to find out and keep the best model for both approaches to demonstrate how each best model fit looks like for each approach. We refer to AIC and Anova criteria to determine the models that predict the future Hall of Famers the best. (Appendix 4) For the versatility approach, the result reveals that all three predictors as main effects only turns out to be the best model, whereas the main effects of 2P% and FT% without any interaction is the best model for shooting ability. Among all 6 of the models, we find that **Model 3 (PTS + REB + AST.cat)** is the very best one for predicting the future hall of famers. Later in our analysis, we are going to implement another method to test and compare the accuracy of our prediction between the two approaches using confusion matrix.

Model Fits for Each Approach



The model fits for versatility describes what we expected in a way that a player with more career points, boards, and high passing ability is much more likely to be in HOF than those with less stats in those corresponding categories. If we look at players with 1000 career points (minimum value for points in x-axis) and different number of rebounds, we see a dramatic gap for the probability of being in HOF for each passing ability. This strongly indicates that passing ability significantly matters. Also, seeing that an imaginary player with close to 40,000 career point with 0 rebound has almost 100% chance of getting into HOF supports that points are one of the most crucial skills in the game of basketball to determine someone's greatness.



Again, the model fits for shooting ability describes pretty much what we expected. Players with a higher percentage of 2P shooting and free throws are much more likely to be in HOF than those are not. One thing to notice in this plot is that an imaginary player with unrealistically good shooting percentages (60% for 2P shooting and close to 100% from the free-throw line) does not have a perfect 100% chance to become a hall of famer. This is another evidence we can assert pure shooting ability is not as major as versatility to be a key indicator to predict the future hall of famers.

Confusion Matrix for Both Approaches

	Prediction No HOF	Prediction HOF (Versatility)
Actual No HOF	631	6
Actual HOF	20	30

	Prediction No HOF	Prediction HOF (Shooting Ability)
Actual No HOF	632	5
Actual HOF	45	5

As mentioned earlier, here we implement the confusion matrix to measure and compare the accuracy of the each best model for both approaches. It turns out that the versatility approach provides a more accurate prediction for the future hall of famers than the shooting ability approach. (96.2% vs 92.7%)

Predictions for the Noticeable 2020 HOF Candidates

	Probability of Making HOF
Tim Duncan	0.996
Kobe Bryant	0.996
Kevin Garnett	0.995
Buck Williams	0.594
Shawn Marion	0.514
Elton Brand	0.382
Terry Porter	0.291
Chauncey Billups	0.254
Larry Nance	0.207
Horace Grant	0.175

Lastly, we are presenting prediction probabilities for 10 of the noticeable 2020 HOF candidates who are ranked by their career win shares, an amount of individual's contribution to their team's wins. We are very satisfied with the predictions as the probabilities of being in the hall of fame for the most famous legendary recently retired players **Tim Duncan**, **Kobe Bryant**, and **Kevin Garnett** are all almost 100%.

Discussion & Conclusion

As a game of basketball in the league NBA keeps evolving with changing trends, the standard and definition of great players also change by seasons. This may imply that the eligibility of players getting into the HOF could vary by season. Unlike old school basketball, where certain positions took roles of their "own things"

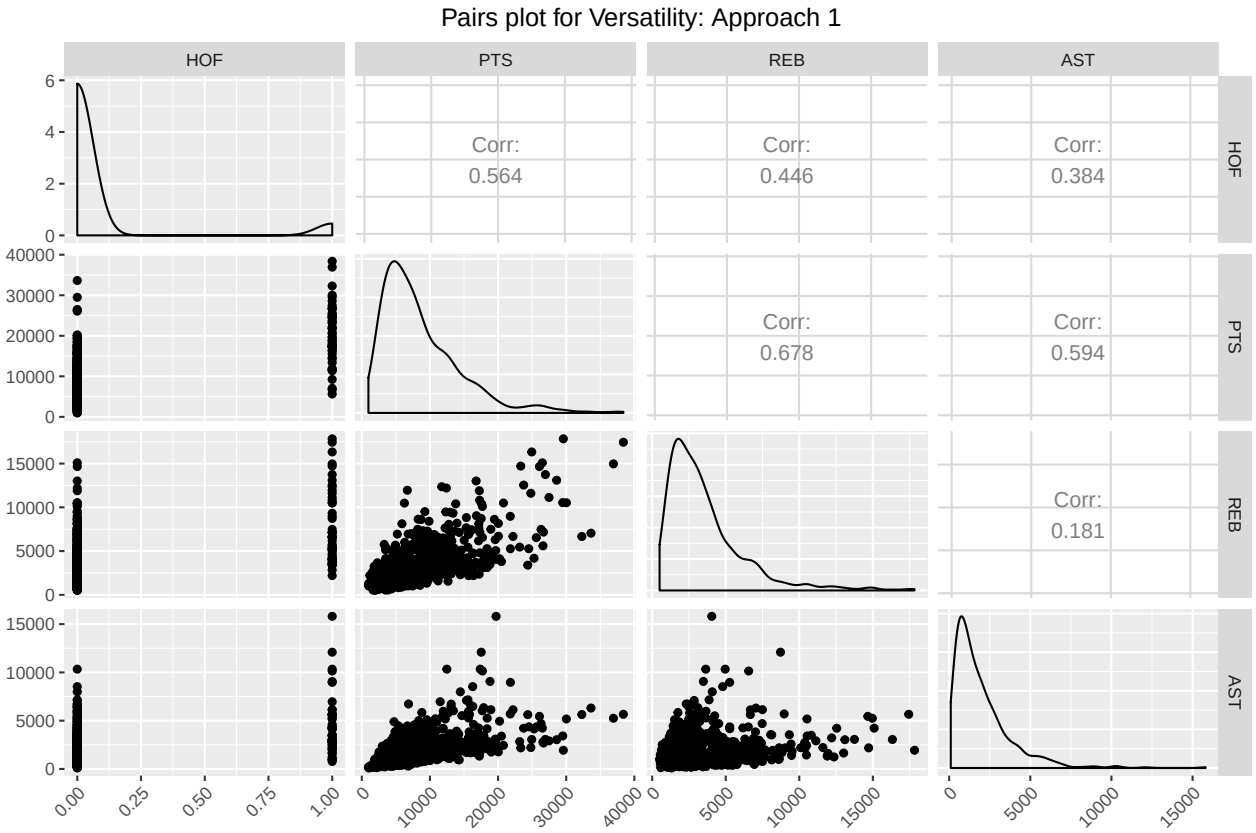
(centers in the paint and guards moving the ball outside), the distinction of roles in this modern era basketball has almost disappeared. Any position player can excel in certain areas that were not used to be "theirs." That is, the fans are being more and more fascinated by those monstrous "triple-doublers" and those ridiculously long ranges 3P shots from downtown that elite players put up day in and day out, representing the popularity and prosperity of NBA nowadays. In order to address these recent trends more clearly, we explored the two main approaches to measure and distinguish which type of a player defines greatness in a ball game. Ultimately, it turns out that a hall of fame caliber all-around versatile player is pretty much about the three most fundamental stats: Points, Rebounds, and Assists. On the other hand, if you are more close to a pure shooter, it seems you do not have to shoot very well from all around, meaning that 3P shots are not as major as the other two 2P and free throw shots yet. However, the fans are very well-aware of the importance and dominance of 3P shots more and more as the trends and strategies in the game of basketball are being centered around 3P shots. Finally, after scrutiny in each approach, we come to conclude that versatility matters the most in order to classify whether a player is likely to be in the hall of fame after their retirement.

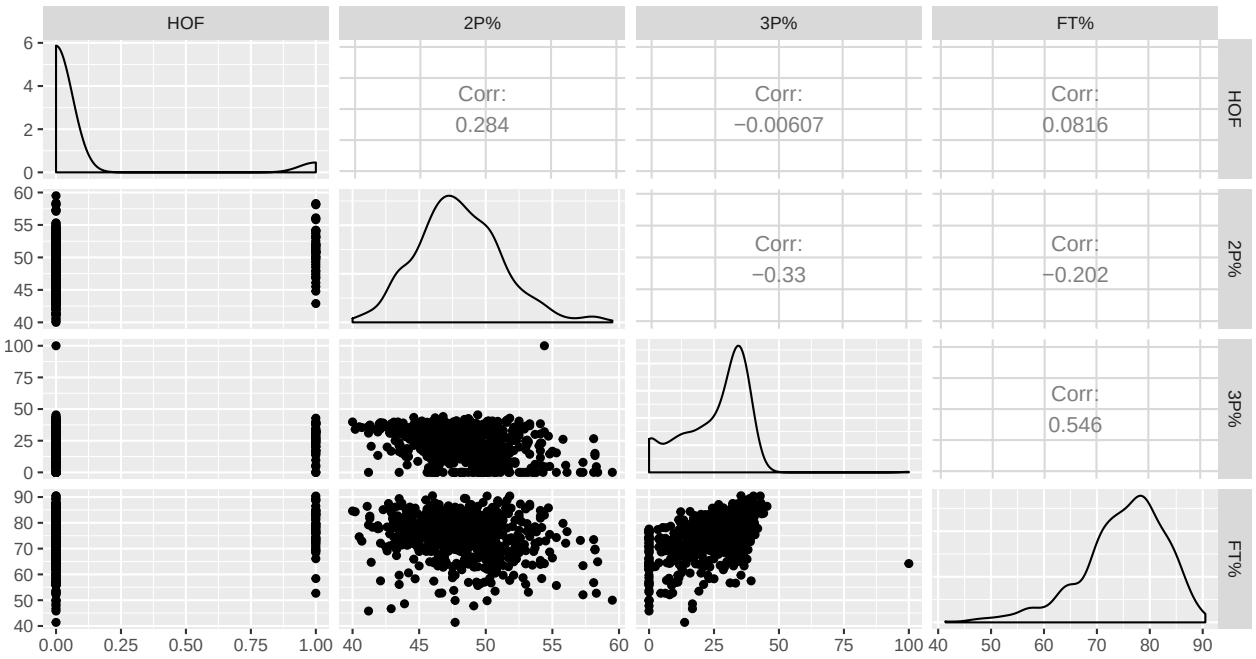
Limitations

Regarding the dataset we have, there is a huge imperfection we had to carry as we proceeded with some statistical approaches for this analysis. As our analysis is based on binomial logistic regression, we had limitations for more appropriate analysis due to a glaring imbalance for the data points for each binary response status. That is, since not everyone makes it into the hall of fame and it is very hard to be an inductee due to high standards, we have a way more non-hall of famers than the hall of famers in our dataset. In terms of variables, we would hope to have a variable for whether a player has ever won a championship or no because we would like to know if having a ring significantly affects the chance of being in the hall of fame, but we have to manually include that status by checking each individual's career as the website does not provide the filtering for that option specifically.

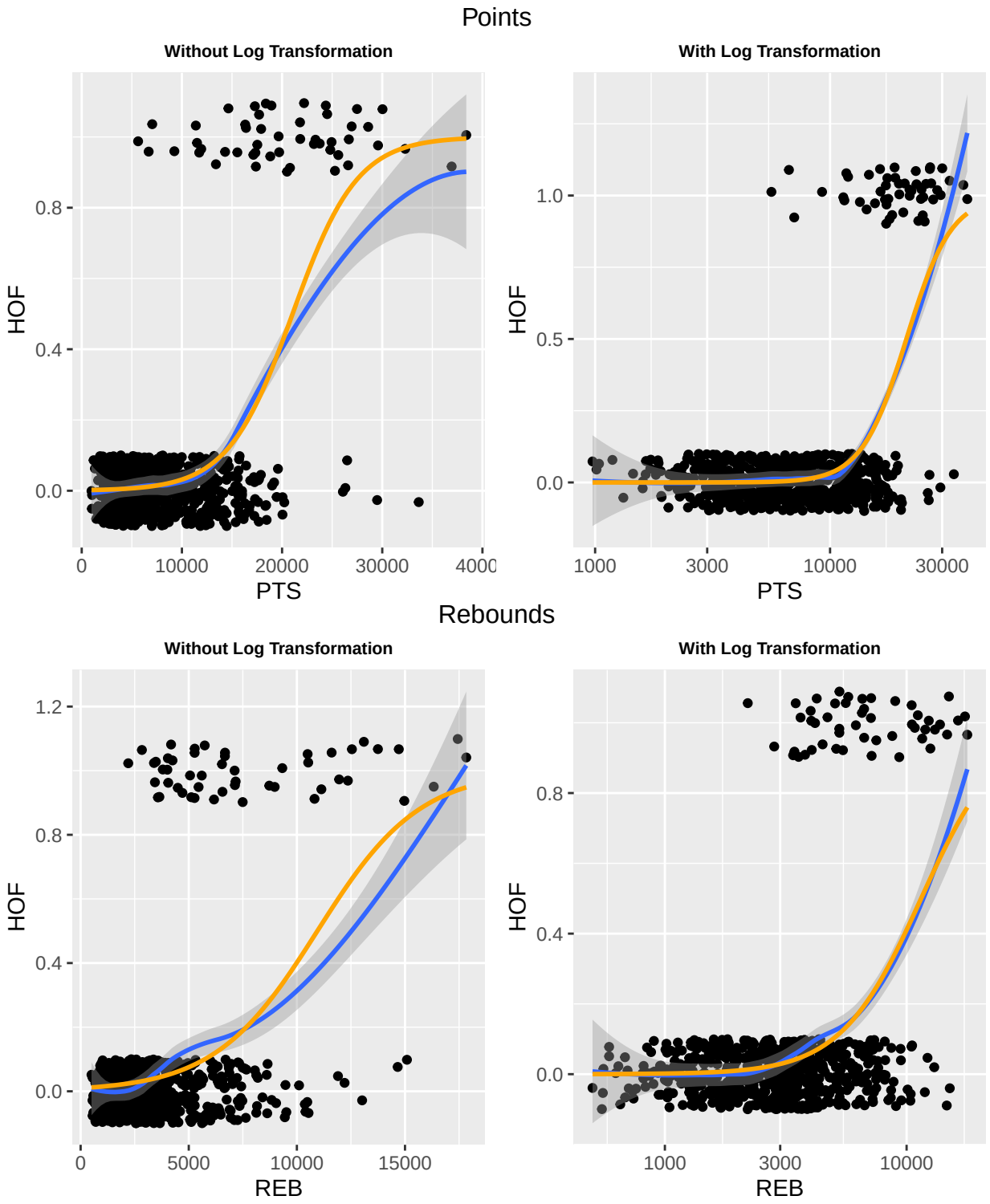
Appendix

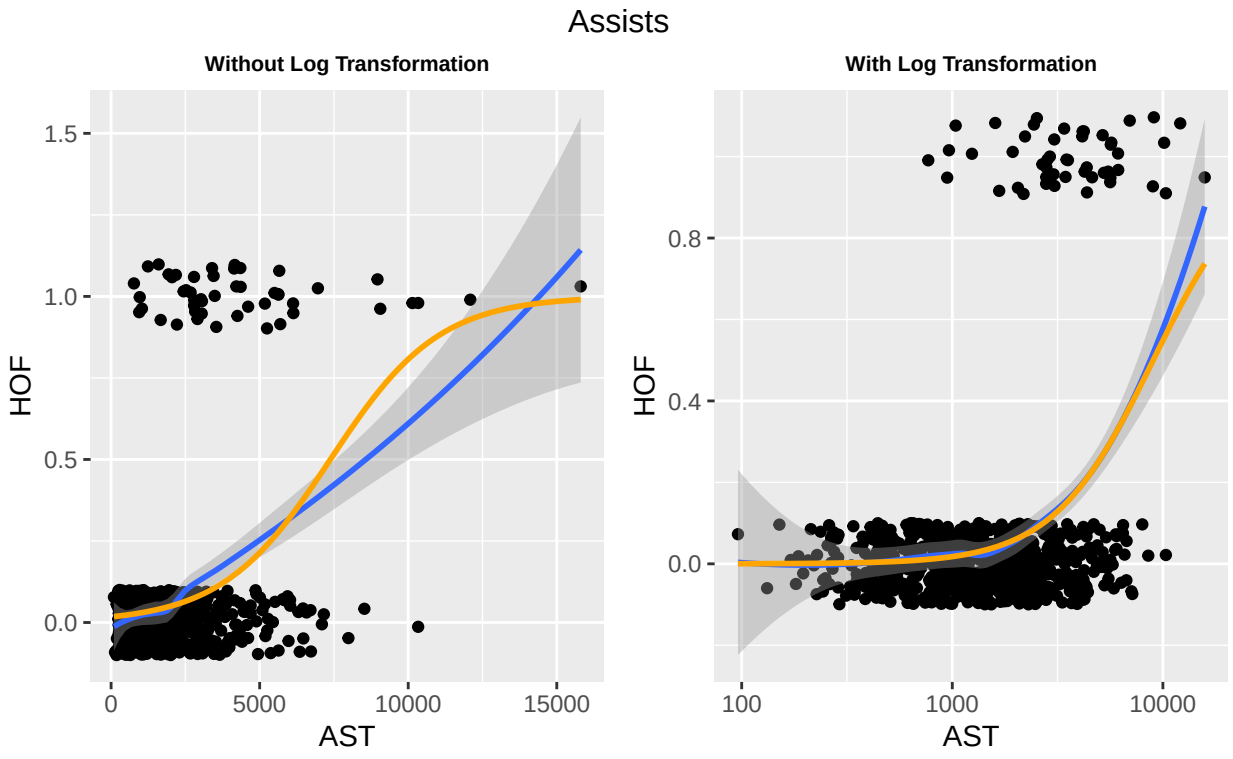
Appendix 1: Pairsplot





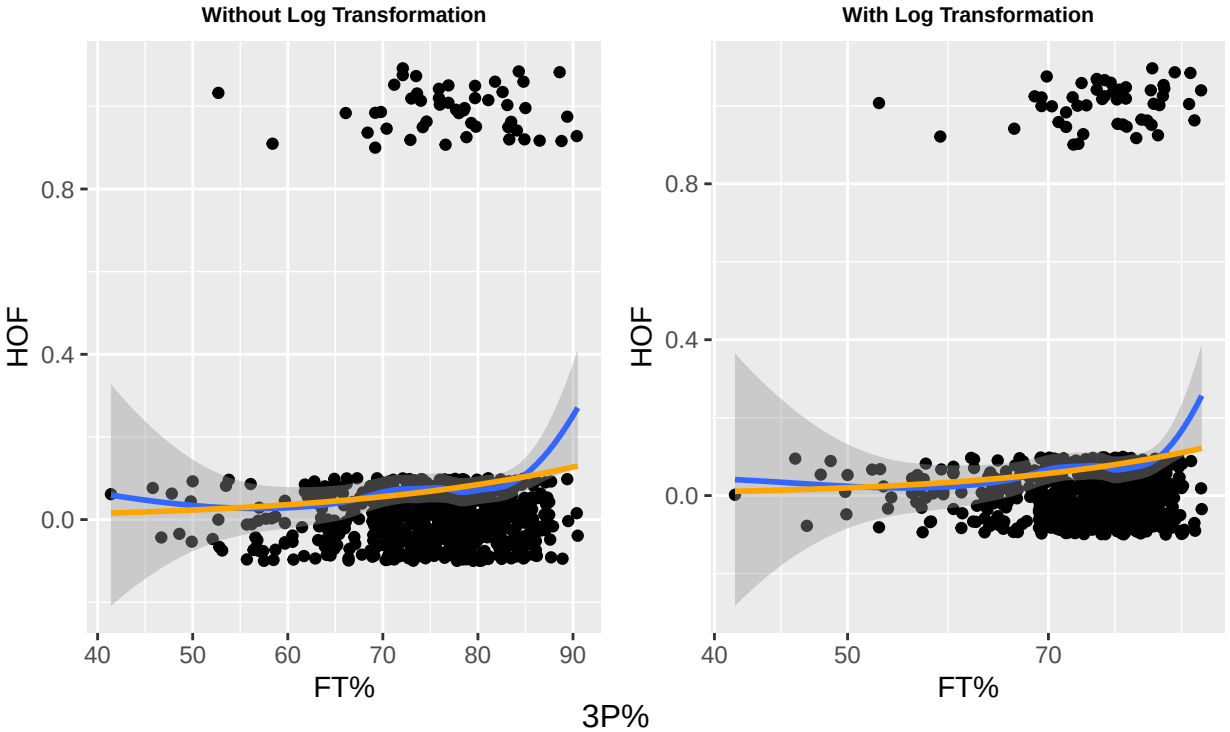
Appendix 2: Monotonic Relationship for Transformation check (Approach 1)



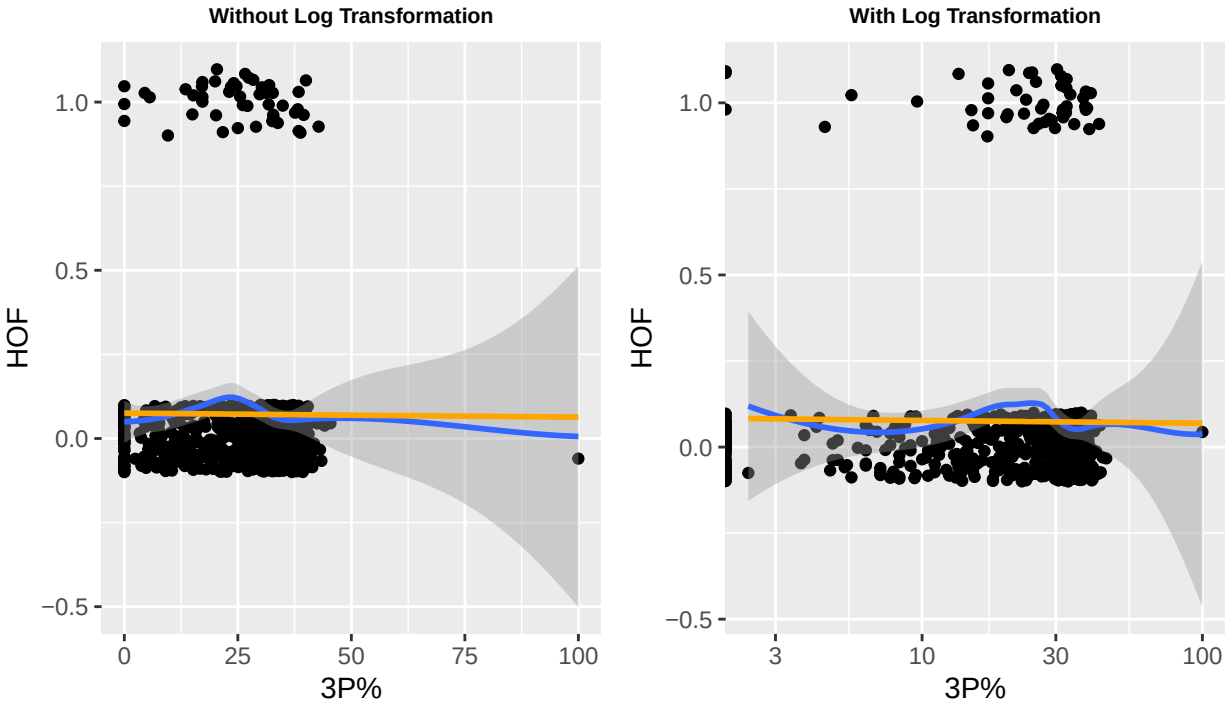


Appendix 3: Monotonic Relationship for Transformation check (Approach 2)

FT%


Without Log Transformation
With Log Transformation
HOF
0.8
0.4
0.0
40
50
60
70
80
90
FT%

3P%

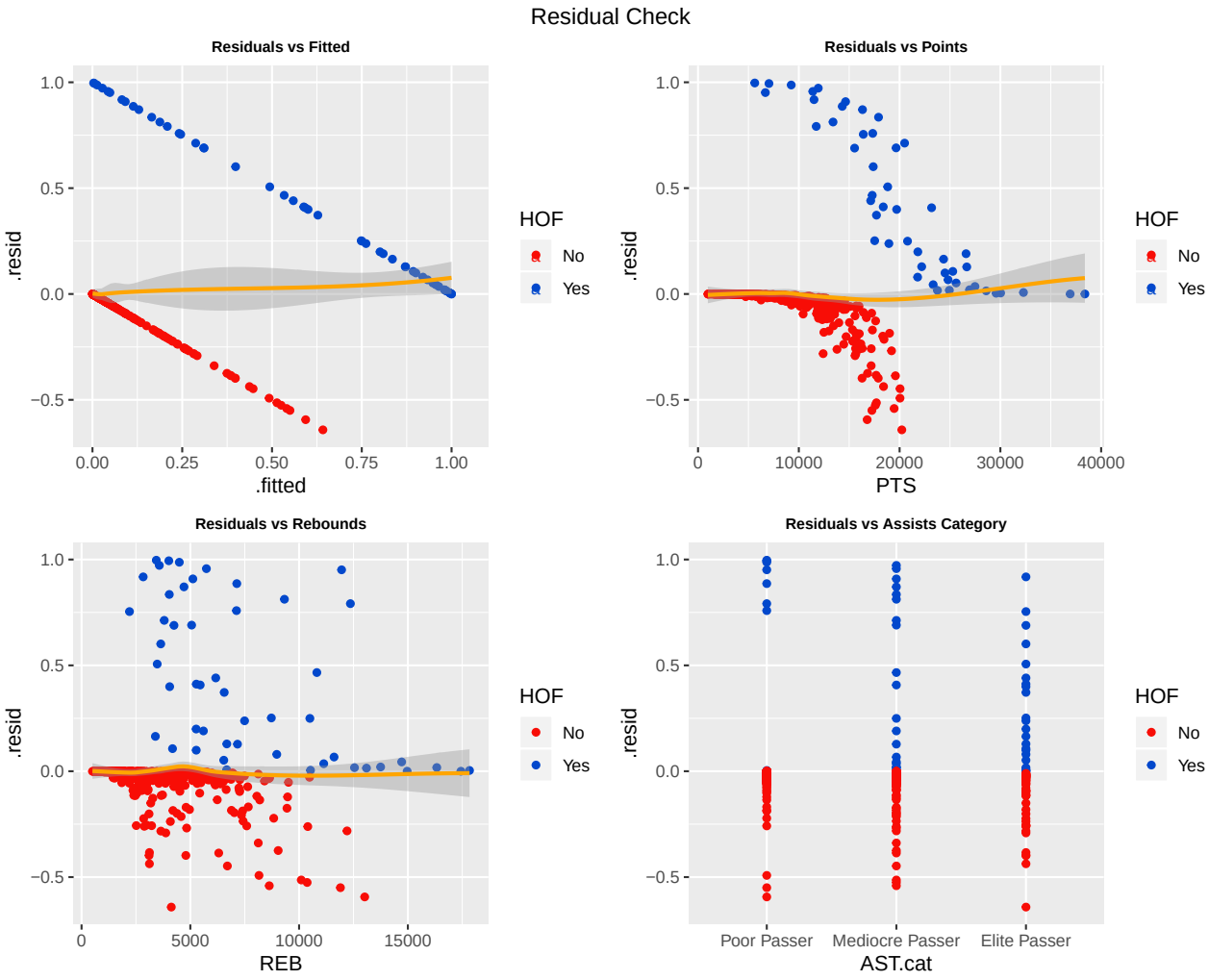

Without Log Transformation
With Log Transformation
HOF
1.0
0.5
0.0
−0.5
0
25
50
75
100
3
10
30
3P%

Appendix 4: Some Model Selection Methods: AIC and ANOVA

	AIC values for each model
Mod1	170.32
Mod2	170.67
Mod3	165.35
Mod4	292.65
Mod5	294.38
Mod6	292.45

	Anova P-values
Addition of AST to PTS + REB? (Versatilty)	0.011
Interaction between PTS and REB or just Main Effects? (Versatility)	0.199
Addition of 3P% to 2P% + FT% (Shooting Ability)	0.604
Interaction between 2P% and FT% or just Main Effects? (Shooting Ability)	0.122

Appendix 5: Assumption Check



Nothing unexpected seems to be going on from those residual plots above. For the most strongly correlated predictor, **Points**, we observe that higher career points lead to residual of closer to 0 for hall of famers, whereas higher points lead to much deviation from residual values of 0 because the "response" residual we are doing is simply 0 or 1 minus the model probabilities, so this makes perfect sense.