

Jaime Castaneda

June 5, 2024

DSC-423-902

Nandhini Gulasingam

Project Report

The purpose of this project is to be able to predict the points per game for NBA players. This doesn't mean for specific players but for the average NBA player. For context, NBA stands for the National Basketball Association. Players in the NBA score points by shooting the ball and making it into a basket, hence the name basketball. There are many statistics that are collected for each player during and after the game. These statistics will be explained further in the next section. The motivation for this project was watching and enjoying the game of basketball. Doing a project on a topic that you enjoy makes it so much better and more enjoyable.

Before we go into more detail on the findings of this project, the statistics should be explained so that it'll be easier to understand. There are 5 positions, referred to POS in the dataset, in basketball and each player has a position. The positions are point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C). Some players play multiple positions. For simplicity, I took the position that was listed first, i.e., if a player was listed as PG-SG, I took their position as PG. Age is how old the player is. G stands for games which indicates how many games they played in the season. GS stands for how many games they started. If a player played 60 games but only started 30, then their GS would be 30 while G would be 60. MP stands for how much minutes they played each game. FG stands for how many field goals the players made each game, not including free throws. Field goals are when a player shoots from any distance and if they make it, then it'd count towards their FG. FGA stands for field goal attempts. Like we said before, FG stands for how many field goals a player makes. Not everyone is going to make every shot they shoot. So, field goal attempts stand for how many total shots the player shot, not including free throws. FG% indicates what percentage of the field goals did the player make. To calculate this, it'd simply be FG/FGA . 3P stands for how many three pointers a player made each game. It's like FG but instead of counting every shot, this statistic only counts for shots beyond the three-point arc (see Appendix A for visual view of an NBA court). 3PA stands for three point attempts each game. Again, similar to FGA but 3PA is only for shots beyond the three-point arc. 3P% is what percentage of three pointers the player of made each game. Like how we can find FG%, it's simply $3P/3PA$. 2P stands for how much two pointers a player made each game. Like 3P except instead of it being about shots beyond the three-point arc, it's for shots inside the three-point arc, not including free throws (see Appendix A for visual view of an NBA court). 2PA indicates how much two-point shots were attempted each game. Like 3PA except instead of it being about shots beyond the three-point arc, it's for shots inside the three-point arc, not including free throws. 2P% indicates the amount of two-pointers the

player made. So, to get 2P%, it'd simply be $2PA/2P$. FT indicates how much free throws a player made each game. Free throws are awarded to players when the player gets fouled. Free throws are shot at the free throw line (see Appendix A for visual view of an NBA court). FTA indicates how much free throws a player shot. FT% indicates what percentage of free throws the player made. Simply, $2PA/2P$. eFG% indicates the amount of shots a player makes. It considers every shot the player took, including free throws. To calculate it, it'd be $(FG+FT)/(FGA+FTA)$. A rebound is when a player shoots the ball, they miss, and a player grabs the ball. The player that grabs the ball off the miss, gets a rebound added to their stats. ORB stands for offensive rebound. An offensive rebound is when the attacking team shoots the ball, they miss, and a player from the attacking team grabs the ball. DRB stands for defensive rebound. A defensive rebound is when the attacking team shoots the ball, they miss, and a player from the defending team grabs the ball. TRB stands for total rebound. It's simply offensive rebound plus defensive rebound, $ORB + DRB$. AST stands for assist per game. An assist in basketball is a pass that directly led to a teammate scoring a field goal. STL stands for steals per game. A steal is when a defensive player legally takes the ball away from the offensive player. This can be done by intercepting a pass, deflecting and controlling a pass, knocking the ball away, swiping the ball away while the offensive player is dribbling, gaining possession of the ball after the opposing team mishandles it. BLK stands for blocks per game. A block is when a defensive player deflects a field goal attempts from an offensive player to prevent them from scoring. TOV stands for turnovers each game. A turnover occurs when the offensive player loses possession of the ball to the opposing team. Turnovers can happen when the offensive player get the ball stolen from them, a player steps out of bounds with the ball, when the player commits a violation, or when the player commits an offensive foul. PF refers to personal fouls. A personal foul is a violation of the rules that involves illegal physical contact between two opposing players.

We can start this project with the data exploration stage. Histograms, boxplots, scatterplots, and correlation values will be seen here. For the histogram on our y-variables, PTS, we can see the distribution is rightly skewed (see Appendix B) which indicates that the majority of the data is located on the left side of the graph, and the mean is greater than the medium. This makes sense since the average NBA player isn't that good, so they won't be scoring that many points. However, this isn't good for our data since we don't like skewness. We're going to want to do a transformation. There are also some outliers on the right end which means that it'll increase the mean of the data.

Boxplots are another way to view the distribution of the data. From the boxplots produced (see Appendix C), we can see the distribution of points from the player's position. 0 indicates that it's a point guard, 1 for shooting guard, 2 for small forward, 3 for power forward, and 4 for center. From the boxplots we can see that there isn't a significant difference from the different positions which indicate that the points are being distributed evenly between the positions. We can see Q0 (minimum) is 0 points, Q1 is around 4 points, Q2 is around 5 points, Q3 is around 12 points, and Q4 (max) is around 21 points. However, we can see there's many

outliers in each of the positions and all their top whiskers are long which means that although most of the data is around 4 to 12 points, there are many outliers for all the positions which might suggest a transformation.

Scatterplots can be used to test the linearity between the y-variable and the x-variables. From the scatterplot produced (see Appendix D), many of them have a positive linear trend which is good since it signals that there's a strong positive relationship between the two variables. Some of them only have two straight lines which is fine since those are the dummy variables and we can ignore those. However, there those appear to be some curves in them which could mean that they're polynomial and must be handled a different way.

Correlation values are used to see how much different x-variables relate to each other. With this data set, it's not surprising that many values relate to each other (see Appendix E). Field goal attempts, FGA, and field goals, FG, are going to be related because they both calculate similar data. Just like field goal attempts is going to relate to field goals, it is also going to relate to two-point attempts, 2PA, and three-point attempts, 3PA. This is not surprising since they both all calculate the number of shots a player has taken. These are just a few of the many examples of collinearity in the data. Collinearity in this instance isn't good since it will indicate that the variables violate independence, violate constant variance, and show patterns in the residual plots.

The next step in this project is the data analysis stage. In this stage, we'll spot and fix the assumptions and diagnostics. The four assumptions are linearity, constant variance, independence, and normality. The diagnostics are outliers, influential points, and multicollinearity. The assumptions will be analyzed first. To analyze the assumptions, we'll have to look at the scatter plots, residual plots, and the normality plot. As seen before, the scatter plots (see Appendix D) show that there is a positive relationship between the variables. Some of them are linear and others are polynomial. It isn't that clear however so we can examine the residual plots. The residual plots will tell us if constant variance and independence is violated. When we look at the residual plots (see Appendix F), we can see there does appear to be a pattern or trend. Many of the values are concentrated at the residuals -1, 0, and 1. This makes them look like there's three straight lines. This isn't too bad because the spread is still mainly random. However, when analyzing the normality plot (see Appendix G), it is terrible. There are about six curves in it which signals that we must do a transformation. For the transformation, I chose to do a log transformation because when I tried to do the other transformations, they didn't solve the issue with the normality as well as the log transformation did. So, a log transformation is what I ultimately ended up going with.

Now that we have our new and transformed y-variable, logPTS, we can start the data exploration stage once again with how we did with the original y-variable. Looking at the histograms, boxplots, scatterplots, and correlation values. When we look at the histogram for the transformed y-variable (see Appendix H), we can see the histogram is not right skewed anymore.

This means that the data is more centered now. There is a slight left skew but it's not as bad as how it was with the original histogram was.

Looking at the new boxplot (see Appendix I), we can observe that the number of outliers dropped significantly. There are still a couple at the bottom but now there isn't any at the top. We can also see that the Q0, Q1, Q2, Q3, and Q4 changed by a lot which makes sense since we transformed the y-variable. Q0 is now -0.5-1, Q1 is 1-1.2, Q2 is 1.9, Q3 is 2.5, and Q4 is 3.1-3.5. The whiskers are still long which indicates that the data is still separated. Meaning that there is a significant difference comparing the min and max to the rest of the data.

The scatterplots, as talked about before, should be linear. When looking at the scatterplots (see Appendix J), there does seem to be a positive trend. However, it appears to be a polynomial trend since many of them have curves in it. This isn't the greatest but it's what we'll have to work with. What I would do if I had the time is do a quadratic regression model to fix this issue. However, we'll stick with our linear regression model.

The correlation values are almost the same as before (see Appendix K). Nothing really changed from the correlation values. Many values are still related to each other and will have to be deleted to deal with multicollinearity. This issue will be solved in the data analysis stage.

Data Analysis with logPoints per game:

Now that we're back to the data analysis stage with our transformed y-variable, we can spot and fix the assumptions and diagnostics if necessary. Starting with the assumptions, we already observed the issue with linearity from the scatterplots (see Appendix J). For independence and constant variance, we can see the residual plots. When analyzing them (see Appendix L), there seems to be a curve in many of them. This isn't surprising as we saw the same trend in the scatterplots, and we discussed how to deal with it in future projects. When we observe the normality plot (see Appendix M), we can see how much straight it is now. There isn't any 'S' shape to it and it's a big improvement from before. So, the assumptions are good and don't need to be fixed.

For the diagnostics, we'll solve multicollinearity first and then move on to the outliers and influential points. We already saw before from the correlation values (see Appendix K) that there was going to be many values related to each other. But to solve this issue, I wanted to observe their VIF, variance influence factor, and delete them by hand until their VIF value is below 10. A VIF value of above 10 indicates that the x-variable has high correlation with one or more of the x-variables. When looking at the VIF values for the x-variables (see Appendix N), we can see that many x-variables have extremely high VIF values. Once we remove all the high VIF valued x-variables, we're left with the ones with a VIF score of less than 10 (see Appendix O). Now that our multicollinearity issue is solved, we can move on to fixing the outliers and influential points. There was a good number of outliers and influential points as we saw before in

our data exploration stage so those were removed. Once those were removed, we can move on to the next stage.

The next stage is the testing/validation stage. Before we do our model selection, the data will be split. I decided to go with a 75/25 split since I believed I had didn't have enough observations in the training set and wanted to put a good amount there. After splitting the data, we can do our model selection. I chose to use capability potential, CP, and backwards. Once we get our final equation using these model selections, we'll analyze assumptions and diagnostics to see if there needs anything to be fixed. When using capability potential, CP, we get our final equation as $\logPTS = -0.2533 - 0.0746 \text{ dPosition1} - 0.1464 \text{ dPosition2} - 0.1236 \text{ dPosition3} - 0.2194 \text{ dPosition4} + 0.0034 \text{ G} - 0.0044 \text{ GS} + 0.1565 \text{ 3PA} + 0.2591 \text{ 3P\%} + 0.1996 \text{ 2P} + 1.1922 \text{ eFG\%} + 0.1502 \text{ ORB} + 0.0314 \text{ DRB} + 0.1217 \text{ STL} + 0.0834 \text{ PF} + 0.2576 \text{ FT\%}$ (see Appendix P for a full picture). This doesn't tell us much, so we'll have to put these values in terms we understand. When I interpret the final equation, we get for dPosition1, logPTS will decrease 7.1885% if this goes up by one. For dPosition2, logPTS will decrease 13.6188% if this goes up by one. For dPosition3, logPTS will decrease 11.6267% if this goes up by one. For dPosition4, logPTS will decrease 19.7000% if this goes up by one. For G, logPTS will increase 0.3406% if this goes up by one. For GS, logPTS will decrease 0.4390% if this goes up by one. For 3PA, logPTS will increase 16.9411% if this goes up by one. For 3P%, logPTS will increase 29.5763% if this goes up by one. For 2P, logPTS will increase 22.0914% if this goes up by one. For eFG%, logPTS will increase 229.4320% if this goes up by one. For ORB, logPTS will increase 16.2067% if this goes up by one. For DRB, logPTS will increase 3.1898% if this goes up by one. For STL, logPTS will increase 12.9415% if this goes up by one. For PF, logPTS will increase 8.6977% if this goes up by one. For FT%, logPTS will increase 29.3821% if this goes up by one. We can clearly see that eFG% has the largest impact out of all the x-variables. Then followed up by 3P%. When analyzing the performance, we can observe the R^2 , Adj- R^2 , GOF, and RMSE. When looking at the R^2 , we can see it's 93.63% which is high so that's good. But it also tells us that this model explains 93.63% of the variance in my dependent variable, logPTS, based on the independent variables. The Adj- R^2 is 93.44% which is high so that's also good. But it also indicates that 93.44% of the outcome, logPTS, can be explained by the independent variables. For the goodness-of-fit test, GOF, we can see the f-value is 481.44 which is very high, and the p-value associated with it is below 0.05 which tells us that this is a very good model. The RMSE is 0.19451 which is very low so that tells us that there isn't much error to it.

We then move one to looking at the assumptions and diagnostics again. That involves looking at the residual plots, normality plots, outliers, influential points, and multicollinearity. From the residual plots (see Appendix Q), we can see that they're very similar to the previous residual plots as before, showing a curve. But other than that, still pretty good and randomly spread. The normality plot is also still fine (see Appendix R). No 'S' shape to it so that's also good. Now that the assumptions are good, we can move on to the diagnostics. The multicollinearity problem was solved before so there's nothing more to fix. There were a couple

outliers and influential so those were removed and fixed. Now the CP model has its final-final model.

Now that we have our final-final model for CP. We must get our final-final model for backwards. We can do that using the same steps as we did before with the CP model. When using the backwards selection, our final equation ends up being $\logPTS = -0.1324 - 0.0825 \text{ dPosition1} - 0.1501 \text{ dPosition2} - 0.1389 \text{ dPosition3} - 0.2199 \text{ dPosition4} + 0.0039 \text{ G} - 0.0047 \text{ GS} + 0.1615 \text{ 3PA} + 0.3003 \text{ 3P\%} + 0.2017 \text{ 2P} + 1.2239 \text{ eFG\%} + 0.1447 \text{ ORB} + 0.0308 \text{ DRB} + 0.1272 \text{ STL} + 0.0891 \text{ PF}$. This equation is almost the same as the CP equation except backwards doesn't include FT% while the CP equation does (see Appendix S). But, like we said before, this doesn't tell us much. So, we'll have to put these values in terms we'll understand. When I interpret the final equation, we get for dPosition1, logPTS will decrease 7.9189% if this goes up by one. For dPosition2, logPTS will decrease 13.9378% if this goes up by one. For dPosition3, logPTS will decrease 12.9685% if this goes up by one. For dPosition4, logPTS will decrease 19.7401% if this goes up by one. For G, logPTS will increase 0.3908% if this goes up by one. For GS: logPTS will decrease 0.4688% if this goes up by one. For 3PA, logPTS will increase 17.5272% if this goes up by one. For 3P%, logPTS will increase 35.0263% if this goes up by one. For 2P, logPTS will increase 22.3481% if this goes up by one. For eFG%, logPTS will increase 240.0424% if this goes up by one. For ORB, logPTS will increase 15.5693% if this goes up by one. For DRB, logPTS will increase 3.1279% if this goes up by one. For STL, logPTS will increase 13.5644% if this goes up by one. For PF, logPTS will increase 9.3190% if this goes up by one. This equation is very similar to the CP equation. eFG% still has the biggest impact with 3P% being in second place. When looking at the R^2 , we can see it's 93.22% which is high so that's good. But it also tells us that this model explains 93.22% of the variance in my dependent variable, logPTS, based on the independent variables. The Adj- R^2 is 93.03% which is high so that's also good. But it also indicates that 93.03% of the outcome, logPTS, can be explained by the independent variables. For the goodness-of-fit test, GOF, we can see the f-value is 483.44 which is very high, and the p-value associated with it is below 0.05 which tells us that this is a very good model. The RMSE is 0.20046 which is very low so that tells us that there isn't much error to it.

From the residual plots (see Appendix T), we can see that they're very similar to the previous residual plots as before, showing a curve. But other than that, still pretty good and randomly spread. The normality plot is also still fine (see Appendix U). No 'S' shape to it so that's also good. Now that the assumptions are good, we can move on to the diagnostics. The multicollinearity problem was solved before so there's nothing more to fix. There were a couple outliers and influential so those were removed and fixed. Now the backwards model has its final-final model.

Before we compare the testing and training, now that we have our final-final model for both, we can look and interpret it. For CP, the final-final equation (see Appendix V) is $\logPTS = -0.2172 - 0.0645 \text{ dPosition1} - 0.1284 \text{ dPosition2} - 0.1053 \text{ dPosition3} - 0.1803 \text{ dPosition4} + 0.0032 \text{ G} - 0.0041 \text{ GS} + 0.1551 \text{ 3PA} + 0.2111 \text{ 3P\%} + 0.2005 \text{ 2P} + 1.2037 \text{ eFG\%} + 0.1317 \text{ ORB}$

+ 0.0264 DRB + 0.1219 STL + 0.0882 PF + 0.2448 FT%. Like stated before, this doesn't mean anything, so we'll have to interpret it. For dPosition1, logPTS will decrease 6.2464% if this goes up by one. For dPosition2, logPTS will decrease 12.0498 % if this goes up by one. For dPosition3, logPTS will decrease 9.9946% if this goes up by one. For dPosition4, logPTS will decrease 16.4980% if this goes up by one. For G, logPTS will increase 0.3205 % if this goes up by one. For GS, logPTS will decrease 0.4092% if this goes up by one. For 3PA, logPTS will increase 16.7775% if this goes up by one. For 3P%, logPTS will increase 23.5036% if this goes up by one. For 2P, logPTS will increase 22.2014% if this goes up by one. For eFG%, logPTS will increase 233.2424% if this goes up by one. For ORB, logPTS will increase 14.0766% if this goes up by one. For DRB, logPTS will increase 2.6752% if this goes up by one. For STL, logPTS will increase 12.9641% if this goes up by one. For PF, logPTS will increase 9.2207% if this goes up by one. For FT%, logPTS will increase 27.7366% if this goes up by one. It's almost like the previous one before with eFG% being the most important and then 3P%. When looking at the R^2 , we can see it's 94.15% which is high so that's good. But it also tells us that this model explains 94.15% of the variance in my dependent variable, logPTS, based on the independent variables. The Adj- R^2 is 93.97% which is high so that's also good. But it also indicates that 93.97% of the outcome, logPTS, can be explained by the independent variables. For the goodness-of-fit test, GOF, we can see the f-value is 516.54 which is very high, and the p-value associated with it is below 0.05 which tells us that this is a very good model. The RMSE is 0.18053 which is very low so that tells us that there isn't much error to it. For backwards, the final-final equation (see Appendix W) is $\logPTS = -0.1107 - 0.0710 \text{ dPosition1} - 0.1335 \text{ dPosition2} - 0.1157 \text{ dPosition3} - 0.1816 \text{ dPosition4} + 0.0037 \text{ G} - 0.0043 \text{ GS} + 0.1591 \text{ 3PA} + 0.2766 \text{ 3P\%} + 0.2027 \text{ 2P} + 1.2483 \text{ eFG\%} + 0.1269 \text{ ORB} + 0.0250 \text{ DRB} + 0.1244 \text{ STL} + 0.0936 \text{ PF}$. Like stated before, this doesn't mean anything, so we'll have to interpret it. For dPosition1, logPTS will decrease 6.8538% if this goes up by one. For dPosition2, logPTS will decrease 12.4972% if this goes up by one. For dPosition3, logPTS will decrease 10.9258% if this goes up by one. For dPosition4, logPTS will decrease 16.6065% if this goes up by one. For G, logPTS will increase 0.3707% if this goes up by one. For GS, logPTS will decrease 0.4291% if this goes up by one. For 3PA, logPTS will increase 17.2455% if this goes up by one. For 3P%, logPTS will increase 31.8639% if this goes up by one. For 2P, logPTS will increase 22.4705% if this goes up by one. For eFG%, logPTS will increase 248.4414% if this goes up by one. For ORB, logPTS will increase 13.5303% if this goes up by one. For DRB, logPTS will increase 2.5315% if this goes up by one. For STL, logPTS will increase 13.2469% if this goes up by one. For PF, logPTS will increase 9.8120% if this goes up by one. It's almost like the previous one before with eFG% being the most important and then 3P%. When looking at the R^2 , we can see it's 93.80% which is high so that's good. But it also tells us that this model explains 93.80% of the variance in my dependent variable, logPTS, based on the independent variables. The Adj- R^2 is 93.62% which is high so that's also good. But it also indicates that 93.62% of the outcome, logPTS, can be explained by the independent variables. For the goodness-of-fit test, GOF, we can see the f-value is 520.84 which is very high, and the p-value associated with it is below 0.05

which tells us that this is a very good model. The RMSE is 0.18574 which is very low so that tells us that there isn't much error to it.

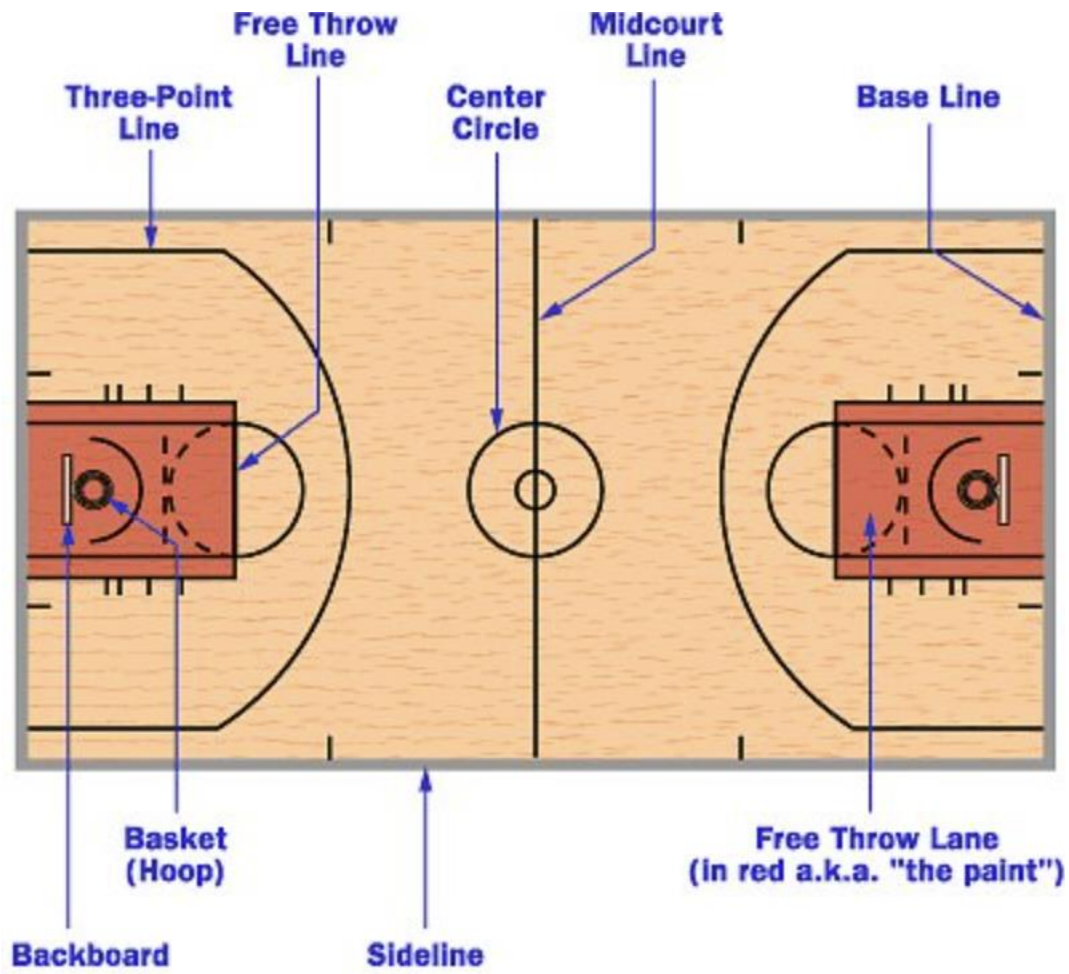
Now we can compare the training and testing. We'll compare all the values we're looking for at the end to make it easier to interpret. When looking at the training of both models (see Appendix X for CP and Appendix Y for backwards), we want to compare the R^2 , Adj- R^2 , RMSE, and GOF. For testing (see Appendix Z for CP and Appendix AA for backwards), we want to compare the RMSE, MAE, R^2 , Adj- R^2 , and CV- R^2 . For training vs testing (for CP, see Appendix X for training and Appendix Z for testing. For backwards, see Appendix Y for training and Appendix AA for testing), we'll compare the RMSE and R^2 . After we get all the values, we can compare them (see Appendix BB). For training vs training, CP has a higher R^2 and Adj- R^2 . But backwards has a lower RMSE and higher f-value for the GOF. For testing vs testing, they both have the exact RMSE and MAE. But CP has a higher R^2 and Adj- R^2 . Both of their CV- R^2 are below 0.3 which makes them good. So, for the training set and testing set, CP is slightly better than backwards mainly due to the R^2 and Adj- R^2 . Now for CP on training vs testing, we can see the training set slightly performed better than the testing set since the training set has a higher R^2 and a lower RMSE. This can be due to the lack of observations in the training set. For backwards training vs testing, we can see it's the same result. Training performed better since it had a higher R^2 and a lower RMSE.

For the prediction, I put 10 random statistics, but I realized that was way too much. In the future, I'll only do 2 or 3 to make it easier. So, we'll only analyze the first two and leave the rest alone. For the first one, this player is a shooting guard, played 50 games, started 45 of them, 2 three point attempts a game, shooting 50% from 3, has 6 2-pointers per game, has an eFG% of 52%, 3 offensive rebounds per game, 4 defensive rebounds per game, 1 steal per game, 4 personal fouls per game, and shooting 75% from the free-throw line. For the second player, this player is a point guard, played 23 games, started 17 of them, has 5 three-point attempts per game, shooting 32% from 3, has 3 2-pointers per game, an eFG% of 34%, 0.2 offensive rebounds per game, 0.8 defensive rebounds per game, 0.2 steals per game, 0.4 personal fouls per game, and shooting 67% from the free-throw line. When looking at the predicted value using the CP model, we can see that for the first player it's 3.10. With the lower end being 2.99 and the higher end being 3.21. For the second player, the predicted value is 3.10. With the lower end being 2.99 and the higher end being 3.22. So, there really isn't that much of a difference between the predicted values for both models.

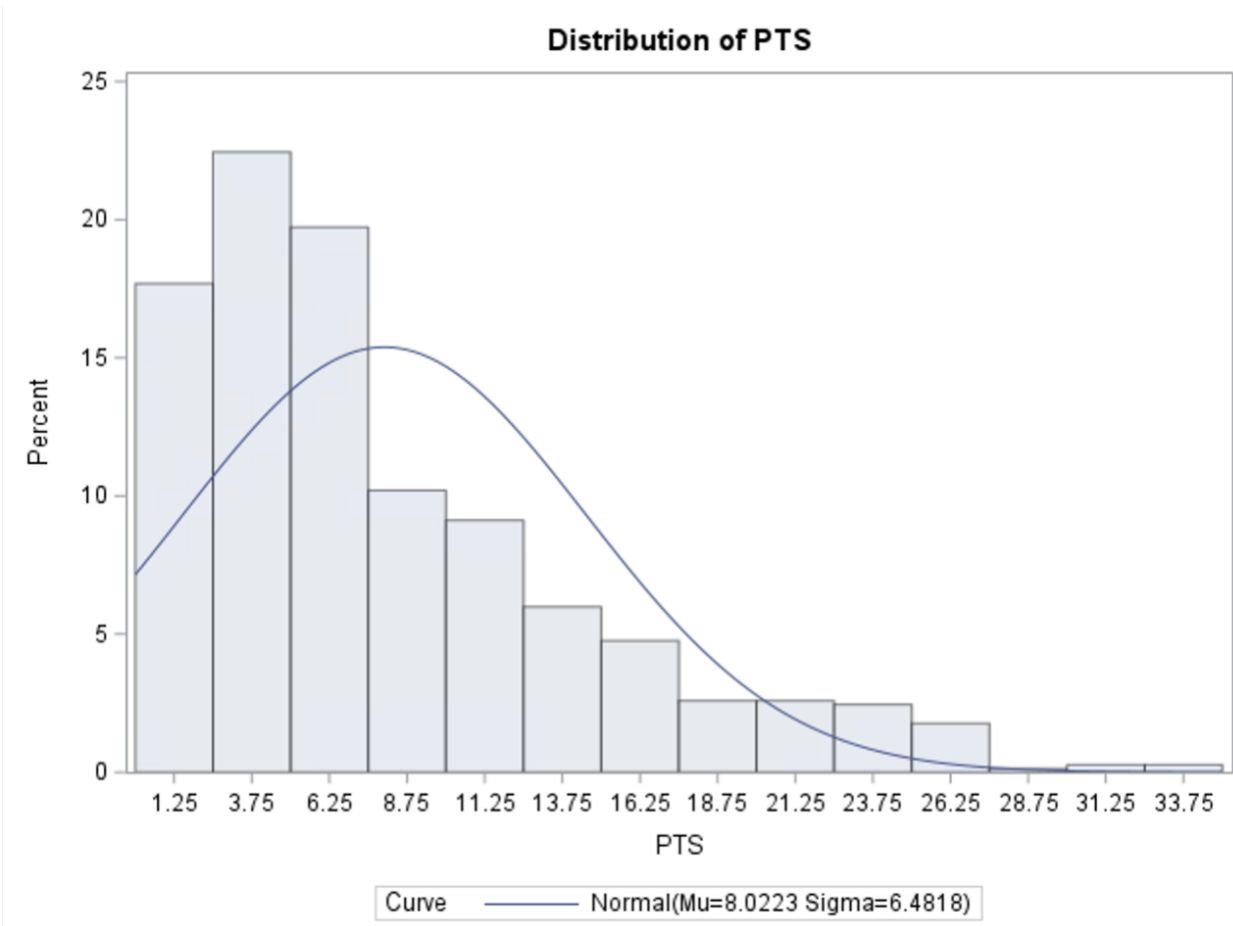
In conclusion, it's clear that the methods used to predict the average points per game for NBA players worked out well. Using all those charts like histograms and boxplots, and then diving deep into statistical tests with our regression models, we really got to the bottom of how different stats, like field goal percentages and points per game, link up. We saw that things like effective field goal percentage and three-point percentages have a big influence on scoring. The log transformation really did the job in smoothing out our data distribution, making our model both more reliable and precise. The capability potential (CP) and backwards selection models

both nailed it, explaining over 93% of the variation in our transformed points per game data. So, it's safe to say, this project was not only about crunching numbers but also about fueling the passion for basketball analytics, and it laid down a solid groundwork for anyone looking to explore further or even improve how we predict player performance.

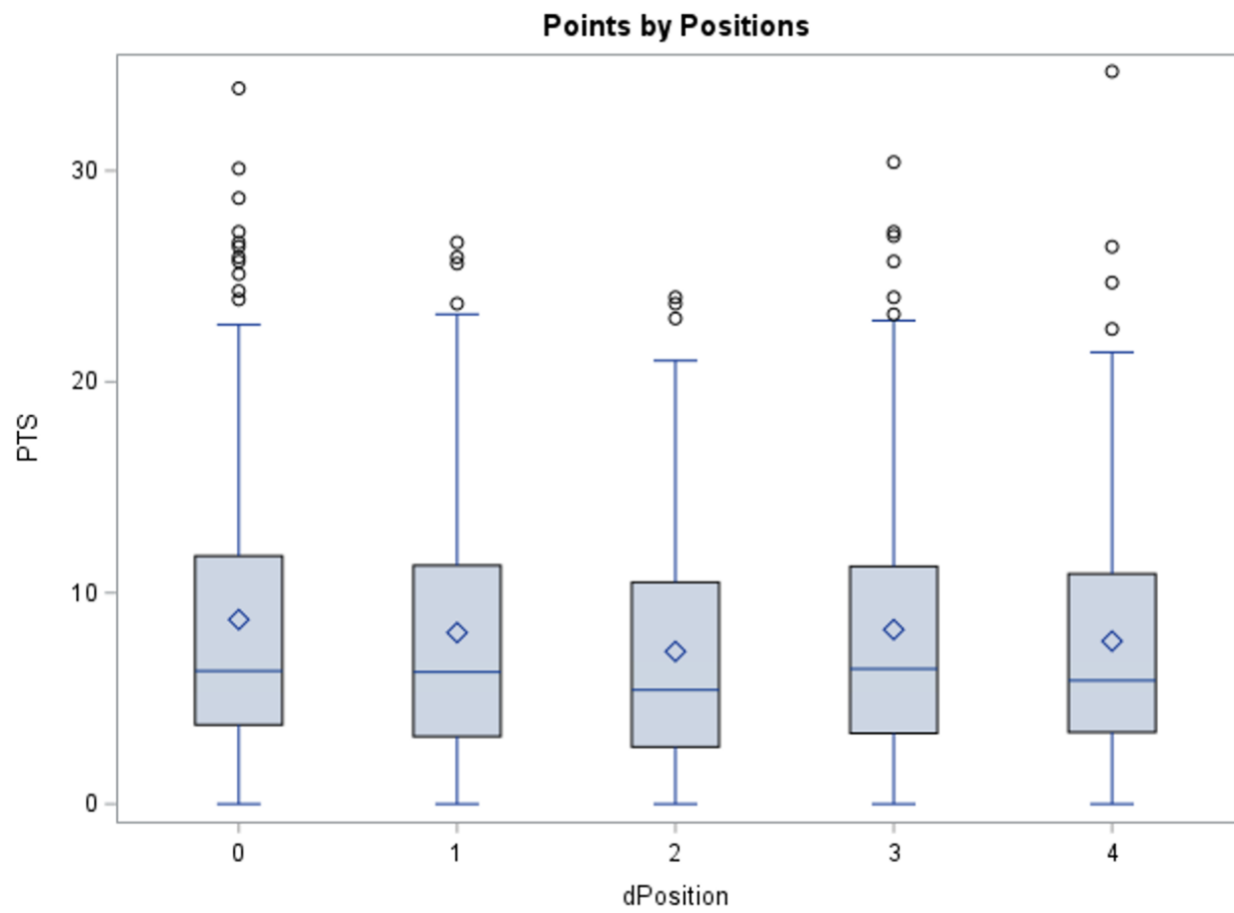
Appendix A:



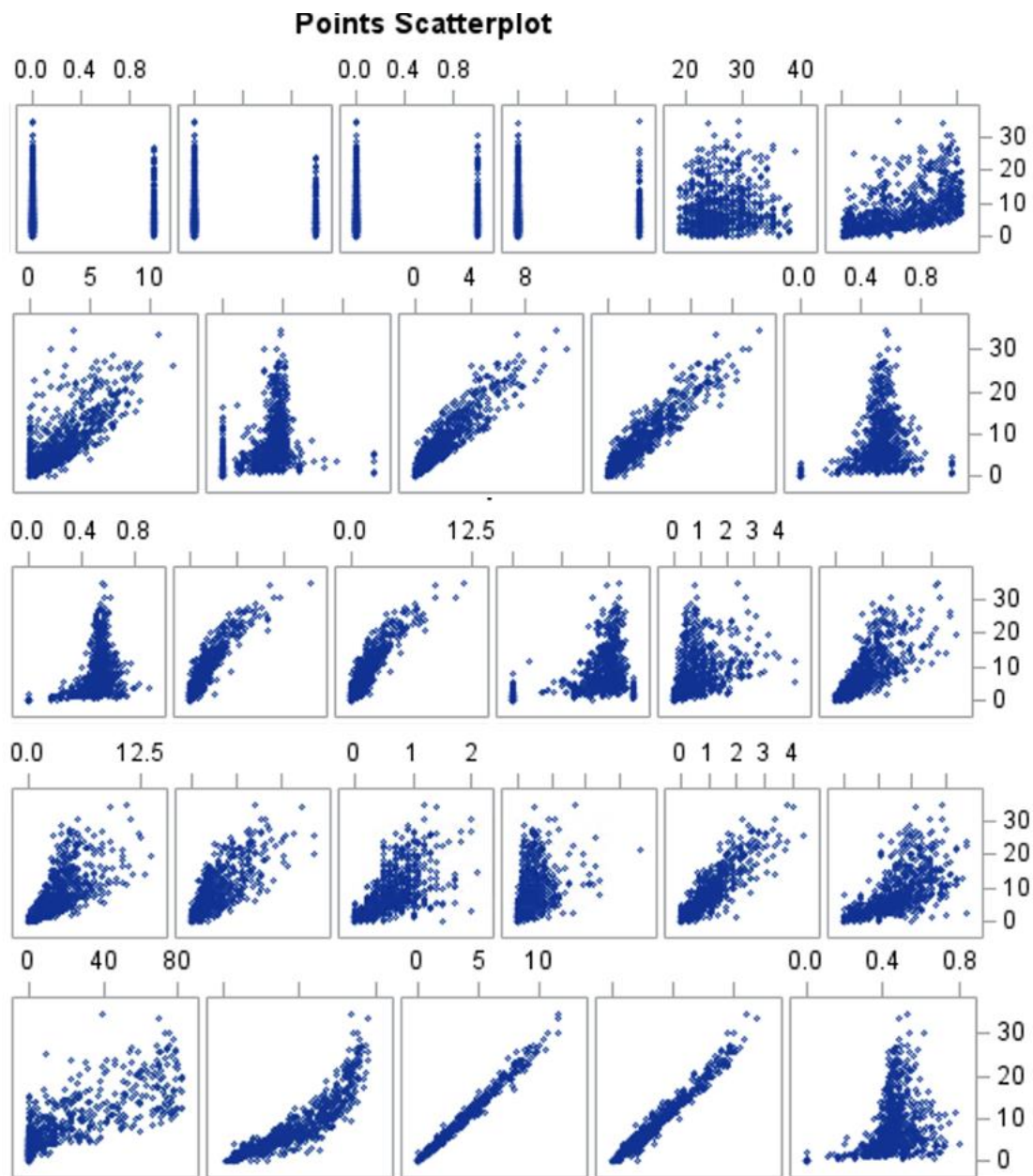
Appendix B:



Appendix C:

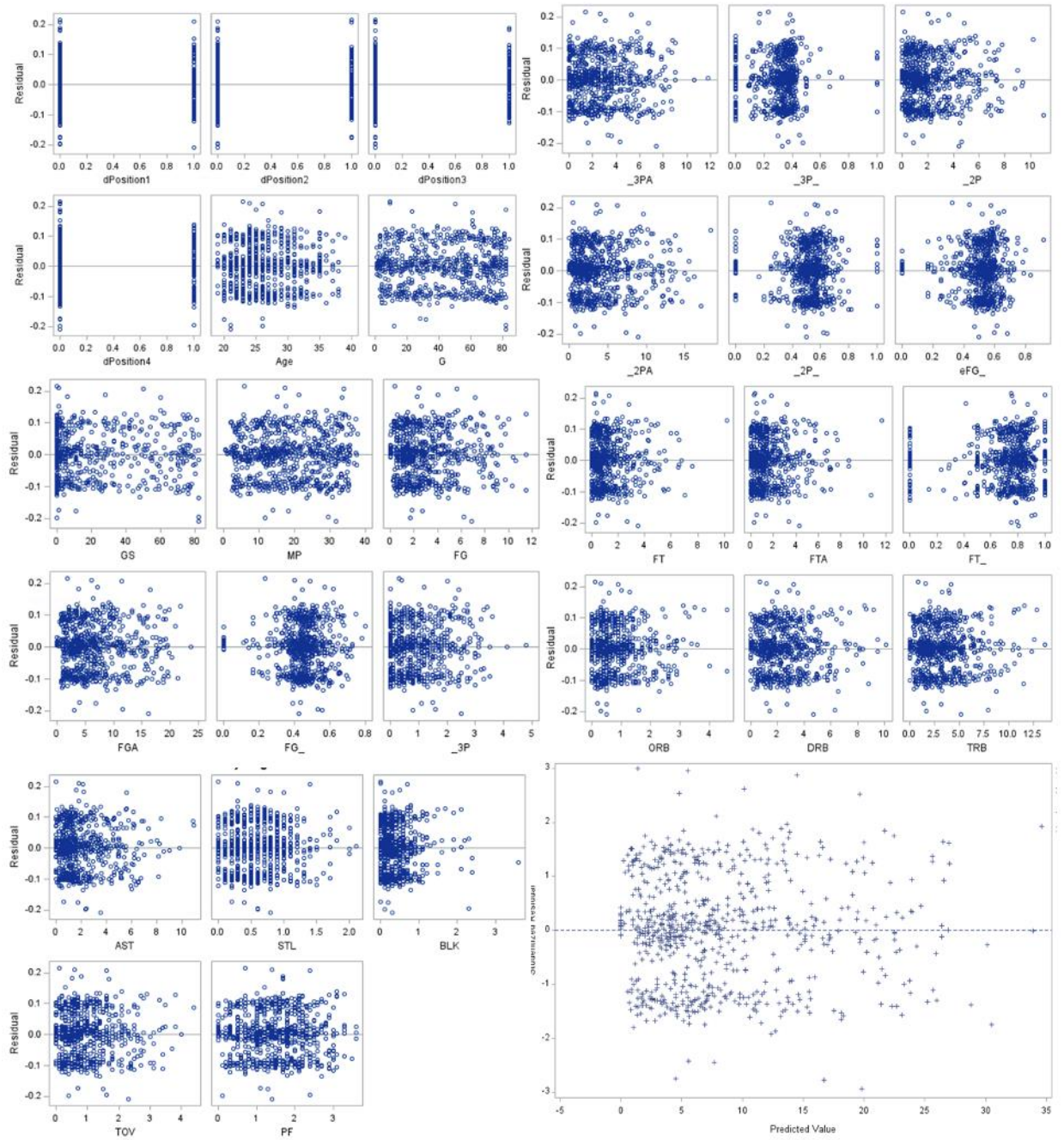


Appendix D:

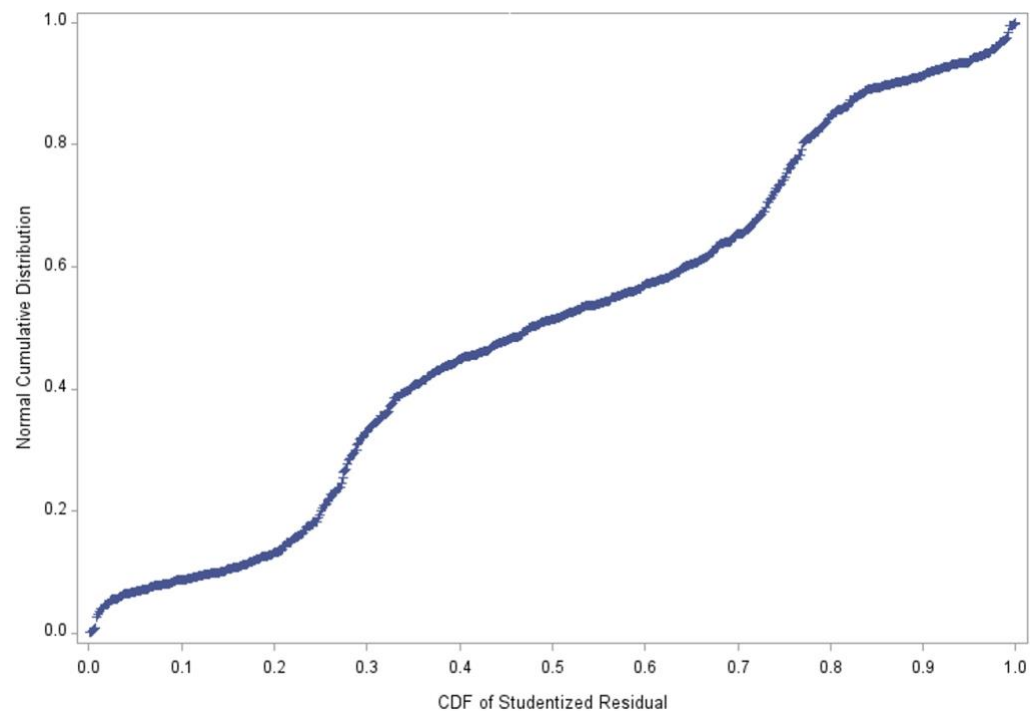


-0.20015 <.0001	-0.25651 <.0001	-0.22968 <.0001	-0.10856 0.0032	-0.01156 0.7539	-0.03942 0.2858	-0.02528 0.4938	-0.00725 0.8444	0.03308 0.3705	-0.11978 0.0011	0.14007 <.0001	0.05178 0.1608	-0.00994 0.0581	-0.04749 0.1985	-0.07574 0.0401	-0.06789 0.0068	-0.21500 <.0001	-0.18712 <.0001	-0.19307 <.0001	-0.00438 0.9090	-0.01828 0.6208	-0.18719 <.0001	-0.02876 0.4688
-0.29585 <.0001	-0.23062 <.0001	-0.08034 0.1021	0.01168 0.7518	-0.01738 0.6380	-0.01498 0.8855	-0.07049 0.0561	-0.05095 0.1676	-0.07145 0.0528	0.06180 0.0941	0.08860 0.0189	-0.11193 0.0024	-0.11245 0.0023	-0.00972 0.7525	0.00468 0.8952	-0.10073 0.0063	-0.06231 0.0914	-0.07934 0.0315	-0.16146 <.0001	-0.01982 0.5916	-0.10503 0.0044	-0.11887 0.1082	
-0.22401 <.0001	0.09682 0.7072	-0.03786 0.3053	0.00417 0.9101	0.01521 0.6805	0.02409 0.5143	-0.00161 0.9653	0.04960 0.1766	-0.04795 0.1941	0.04344 0.2394	0.04043 0.2086	0.02949 0.4247	0.05551 0.0759	0.04553 0.2176	0.10442 0.0046	0.10033 0.0065	0.10796 0.0091	-0.09009 0.4374	-0.02869 0.0091	0.09977 0.0068	-0.02150 0.5015		
0.04610 0.2119	0.06975 0.0588	0.05165 0.1619	-0.03388 0.3575	0.01323 0.7203	-0.08029 0.0295	0.32526 <.0001	-0.31896 <.0001	-0.28969 <.0001	0.15130 <.0001	0.09175 0.0128	0.18785 <.0001	0.16296 <.0001	0.49642 <.0001	0.28005 <.0001	0.36662 <.0001	-0.12778 0.0005	-0.09695 0.0067	0.42240 0.0001	0.01321 0.7206			
0.08276 0.0248	0.07088 0.0554	0.14544 <.0001	0.06919 0.0608	0.07043 0.0563	0.03180 0.3893	0.12540 0.0007	0.09903 0.0072	0.01816 0.6231	0.01924 0.6025	-0.00374 0.9193	0.09259 0.0120	-0.02973 0.4209	0.08288 0.0248	0.05287 0.1522	0.15538 <.0001	0.11851 0.0013	0.00477 0.8973	0.03662 0.3348				
0.66923 <.0001	0.64764 <.0001	0.57628 <.0001	0.55769 <.0001	0.34291 <.0001	0.47427 <.0001	0.31899 <.0001	0.49743 <.0001	0.49120 <.0001	0.27434 <.0001	0.39253 <.0001	0.31817 <.0001	0.53087 <.0001	0.49791 <.0001	0.45147 <.0001	0.41086 <.0001	0.33953 <.0001	0.48795 <.0001					
0.37090 <.0001	0.80498 <.0001	0.79260 <.0001	0.77108 <.0001	0.25423 <.0001	0.57144 <.0001	0.19189 <.0001	0.73273 <.0001	0.73291 <.0001	0.17500 <.0001	0.25964 <.0001	0.39139 <.0001	0.68851 <.0001	0.64738 <.0001	0.63645 <.0001	0.55126 <.0001	0.46036 <.0001	0.69561 <.0001					
0.36081 <.0001	0.61891 <.0001	0.88926 <.0001	0.89498 <.0001	0.31237 <.0001	0.73117 <.0001	0.31373 <.0001	0.79068 <.0001	0.80708 <.0001	0.25040 <.0001	0.35327 <.0001	0.43923 <.0001	0.78444 <.0001	0.72671 <.0001	0.74821 <.0001	0.74529 <.0001	0.46771 <.0001	0.81238 <.0001					
0.63402 <.0001	0.18355 <.0001	0.83131 <.0001	0.97808 <.0001	0.35194 <.0001	0.69109 <.0001	0.27163 <.0001	0.94498 <.0001	0.95225 <.0001	0.25261 <.0001	0.34388 <.0001	0.41155 <.0001	0.77711 <.0001	0.71286 <.0001	0.76114 <.0001	0.63654 <.0001	0.43070 <.0001	0.86920 <.0001					
0.45451 <.0001	0.50744 <.0001	0.70140 <.0001	0.64107 <.0001	0.22389 <.0001	0.79166 <.0001	0.29334 <.0001	0.88096 <.0001	0.91834 <.0001	0.16452 <.0001	0.24715 <.0001	0.30495 <.0001	0.71095 <.0001	0.62934 <.0001	0.78274 <.0001	0.64774 <.0001	0.34538 <.0001	0.87242 <.0001					
0.97531 <.0001	0.54141 <.0001																					

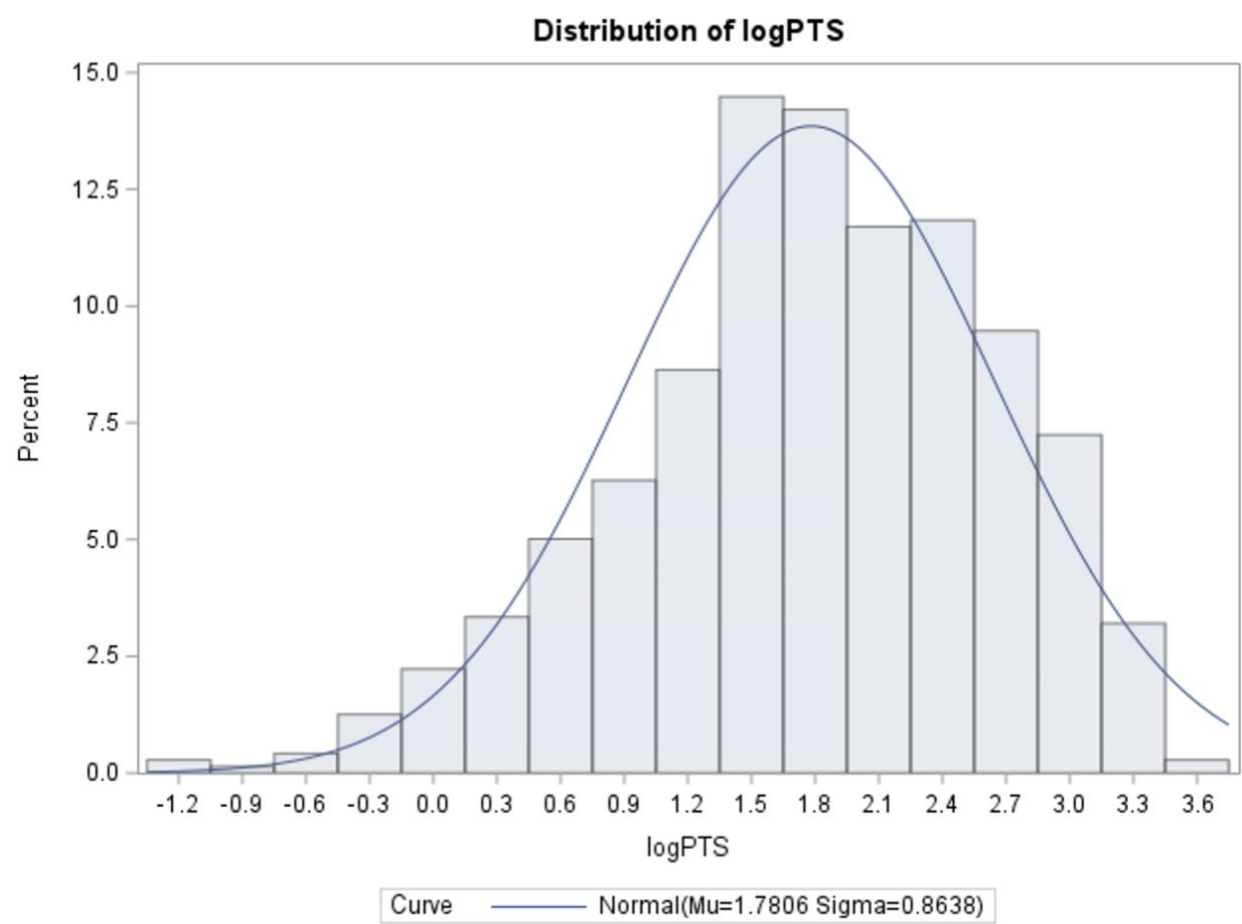
Appendix F:



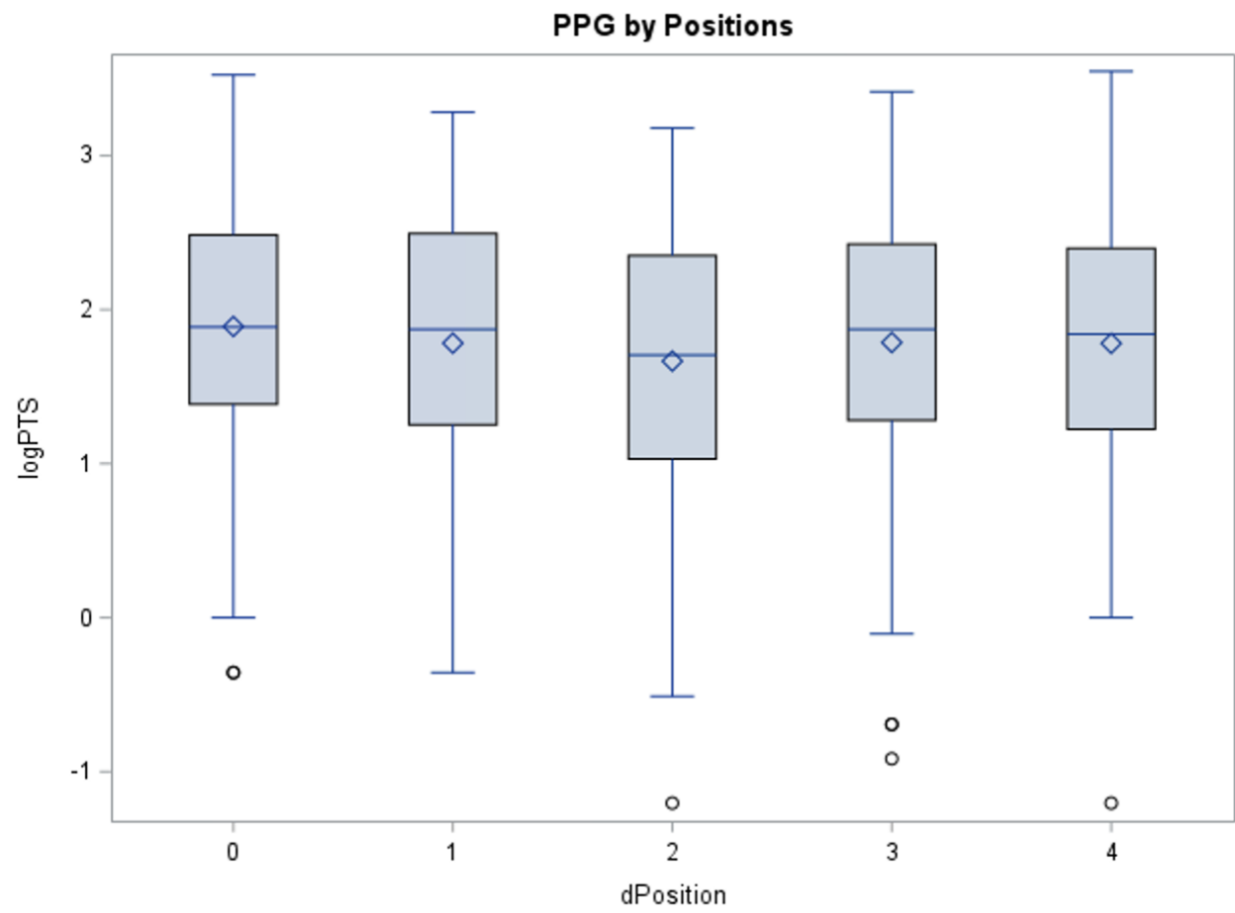
Appendix G:



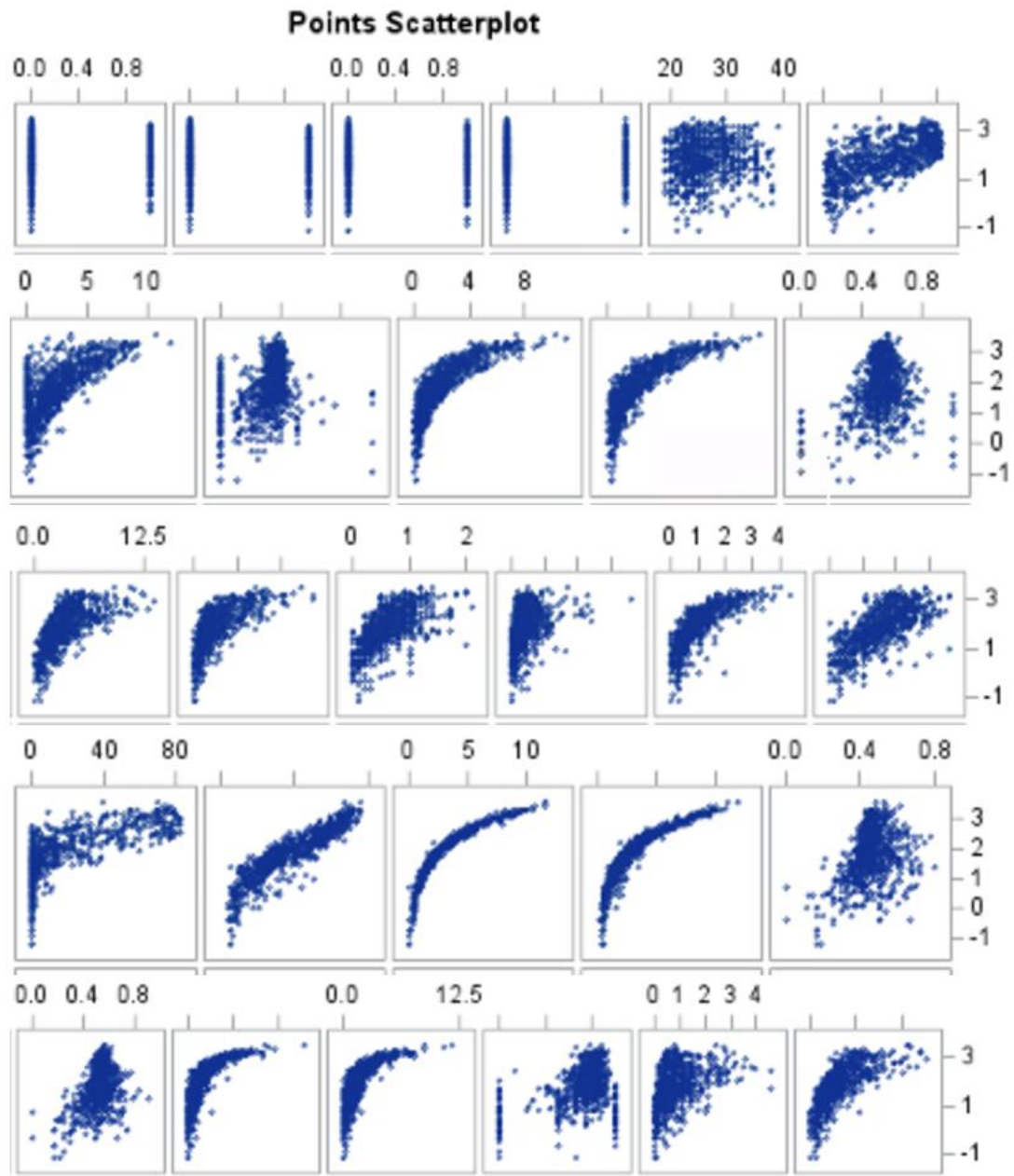
Appendix H:



Appendix I:



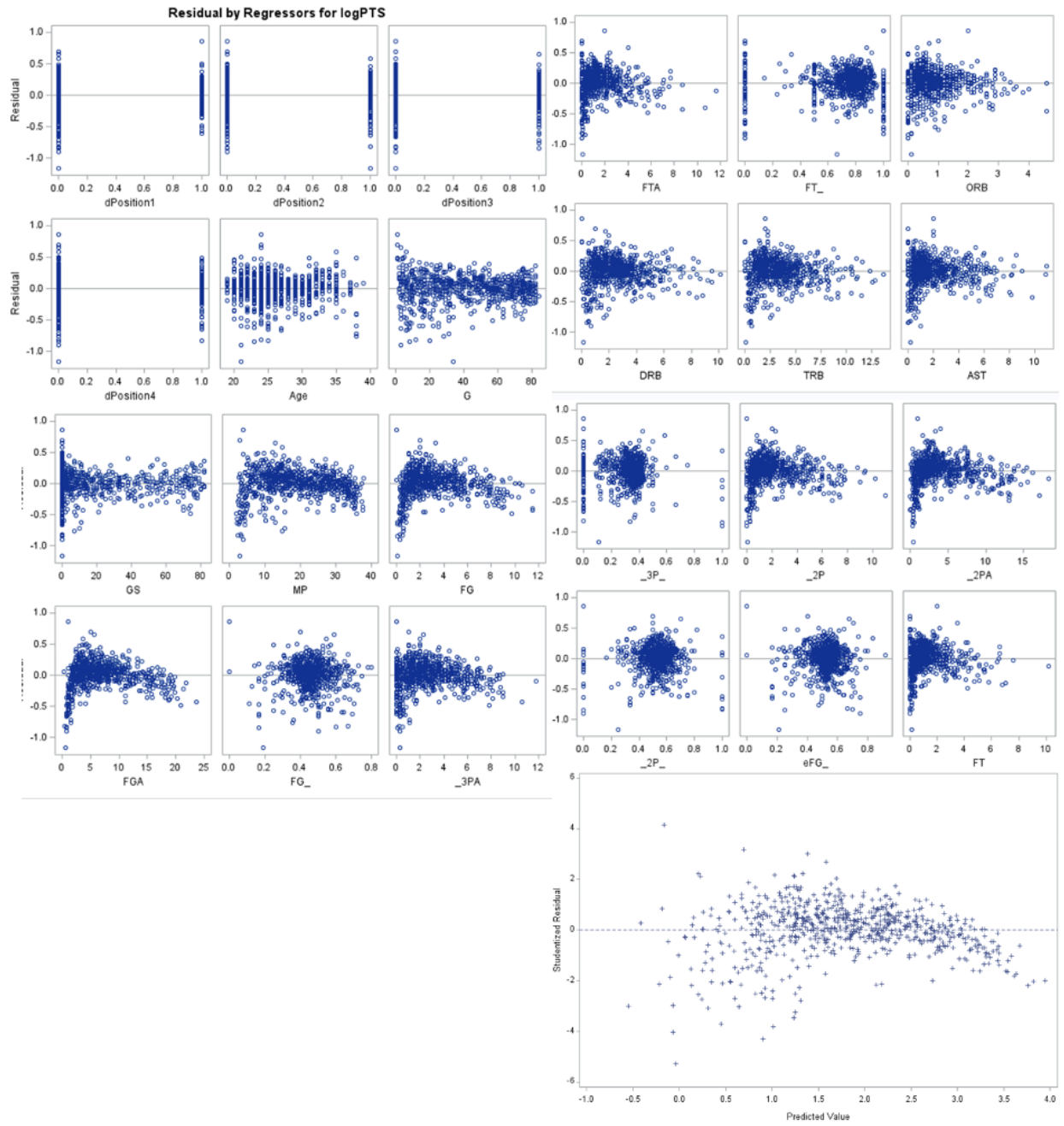
Appendix J:



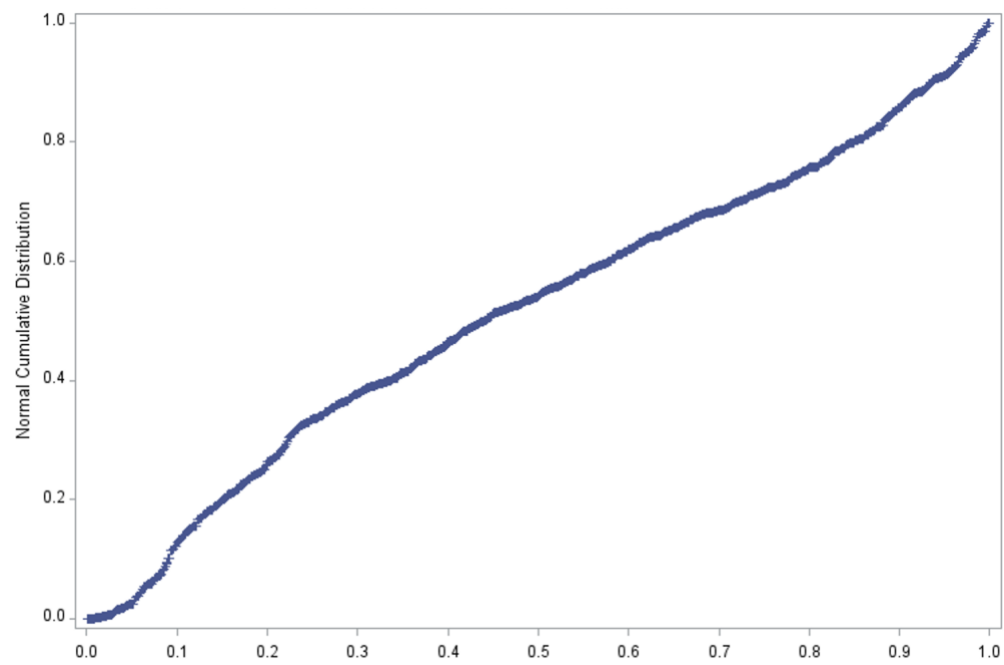
Appendix K:

[illegible]

Appendix L:



Appendix M:



Appendix N:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.94781	0.09688	-9.78	<.0001	0
dPosition1	1	-0.05521	0.02853	-1.94	0.0534	1.90132
dPosition2	1	-0.14843	0.03094	-4.80	<.0001	2.25835
dPosition3	1	-0.12834	0.03254	-3.94	<.0001	2.39965
dPosition4	1	-0.16576	0.04182	-3.96	<.0001	3.37853
Age	1	-0.00094363	0.00213	-0.44	0.6575	1.22595
G	1	0.00300	0.00050773	5.91	<.0001	2.29831
GS	1	-0.00582	0.00068023	-8.56	<.0001	4.11268
MP	1	0.03605	0.00352	10.24	<.0001	15.91603
FG	1	-0.32099	0.06822	-4.71	<.0001	357.71127
FGA	1	-0.00618	0.17715	-0.03	0.9722	9948.09518
FG_	1	0.18752	0.41597	0.45	0.6523	23.97011
_3PA	1	0.24065	0.17723	1.36	0.1750	2061.39342
3P	1	0.28327	0.09078	3.12	0.0019	2.54240
_2P	1	0.27079	0.07339	3.69	0.0002	263.92991
_2PA	1	0.12679	0.17885	0.71	0.4786	4991.40804
2P	1	0.00695	0.10578	0.07	0.9476	2.71997
eFG_	1	1.83316	0.38801	4.75	<.0001	20.82375
FT	1	-0.20660	0.05558	-3.72	0.0002	73.58101
FTA	1	0.18012	0.04634	3.89	0.0001	75.98332
FT_	1	0.42779	0.04358	9.82	<.0001	1.47027
ORB	1	0.23231	0.17453	1.33	0.1836	228.40928
DRB	1	0.08667	0.17522	0.49	0.6210	1291.94008
TRB	1	-0.09841	0.17487	-0.56	0.5738	2280.76261
AST	1	-0.02478	0.01116	-2.22	0.0267	5.67219
STL	1	-0.00455	0.03500	-0.13	0.8966	2.59896
BLK	1	0.01253	0.03258	0.38	0.7007	2.40024
TOV	1	-0.03212	0.03066	-1.05	0.2952	7.53425
PF	1	0.00345	0.02116	0.16	0.8706	3.72414

Appendix O:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	-0.52902	0.08515	-6.21	<.0001	0
dPosition1	1	-0.06670	0.03218	-2.07	0.0385	1.88505
dPosition2	1	-0.14587	0.03465	-4.21	<.0001	2.20694
dPosition3	1	-0.14285	0.03654	-3.91	0.0001	2.35856
dPosition4	1	-0.24742	0.04637	-5.34	<.0001	3.23669
Age	1	0.00014282	0.00235	0.06	0.9517	1.17094
G	1	0.00377	0.00056821	6.63	<.0001	2.24352
GS	1	-0.00512	0.00072786	-7.03	<.0001	3.67005
_3PA	1	0.16815	0.00780	21.55	<.0001	3.11288
3P	1	0.25838	0.08908	2.90	0.0038	1.90798
_2P	1	0.21085	0.01454	14.50	<.0001	8.07795
2P	1	0.01495	0.10289	0.15	0.8845	2.00591
eFG_	1	1.36145	0.15039	9.05	<.0001	2.46344
FT	1	-0.01378	0.01664	-0.83	0.4079	5.14217
FT_	1	0.41210	0.04688	8.79	<.0001	1.32566
ORB	1	0.18970	0.02593	7.31	<.0001	3.93101
DRB	1	0.01656	0.01325	1.25	0.2119	5.75807
AST	1	-0.00533	0.01234	-0.43	0.6658	5.40806
STL	1	0.10559	0.03804	2.78	0.0057	2.39240
BLK	1	0.01059	0.03655	0.29	0.7720	2.35400
TOV	1	-0.01104	0.03378	-0.33	0.7440	7.12727
PF	1	0.10312	0.02162	4.77	<.0001	3.03002

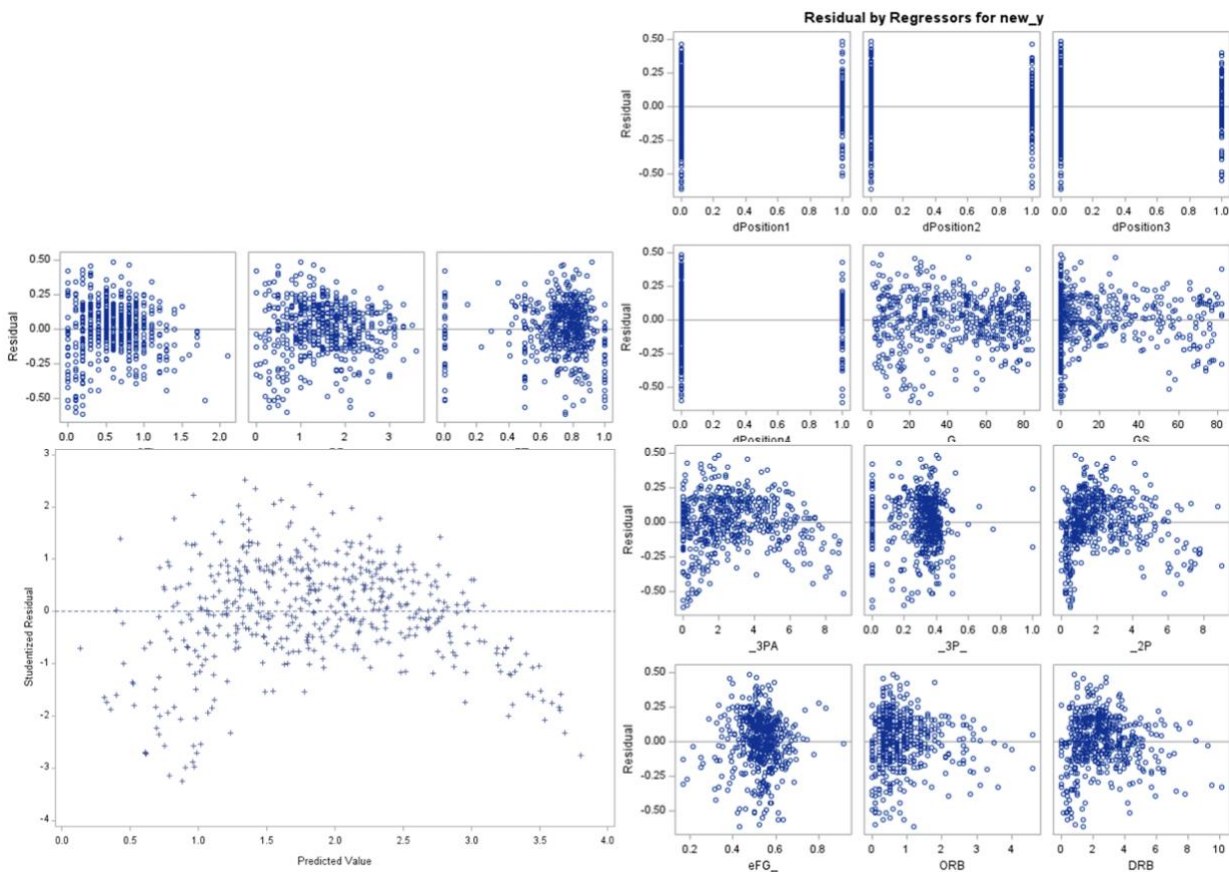
Appendix P:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	273.22885	18.21528	481.44	<.0001
Error	491	18.57684	0.03783		
Corrected Total	506	291.80569			

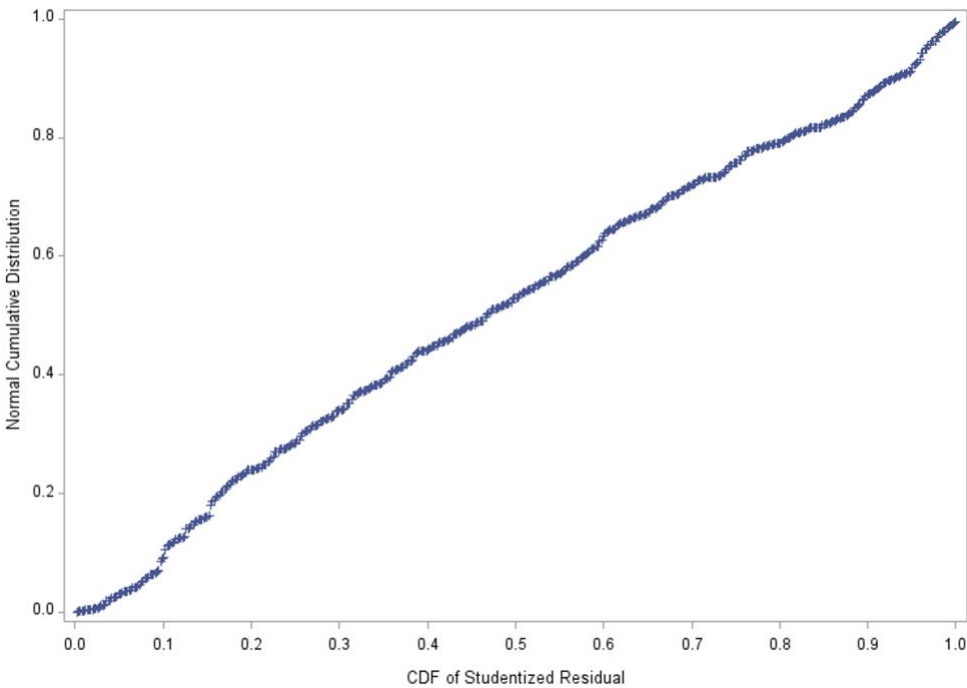
Root MSE	0.19451	R-Square	0.9363
Dependent Mean	1.84094	Adj R-Sq	0.9344
Coeff Var	10.56590		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.25325	0.06172	-4.10	<.0001
dPosition1	1	-0.07460	0.02692	-2.77	0.0058
dPosition2	1	-0.14635	0.02803	-5.22	<.0001
dPosition3	1	-0.12362	0.02930	-4.22	<.0001
dPosition4	1	-0.21935	0.03860	-5.68	<.0001
G	1	0.00340	0.00051254	6.64	<.0001
GS	1	-0.00437	0.00062525	-7.00	<.0001
_3PA	1	0.15646	0.00687	22.77	<.0001
3P	1	0.25907	0.08211	3.16	0.0017
_2P	1	0.19957	0.00906	22.02	<.0001
eFG_	1	1.19215	0.12164	9.80	<.0001
ORB	1	0.15021	0.02211	6.79	<.0001
DRB	1	0.03138	0.01157	2.71	0.0069
STL	1	0.12169	0.03395	3.58	0.0004
PF	1	0.08337	0.01885	4.42	<.0001
FT_	1	0.25756	0.04585	5.62	<.0001

Appendix Q:



Appendix R:



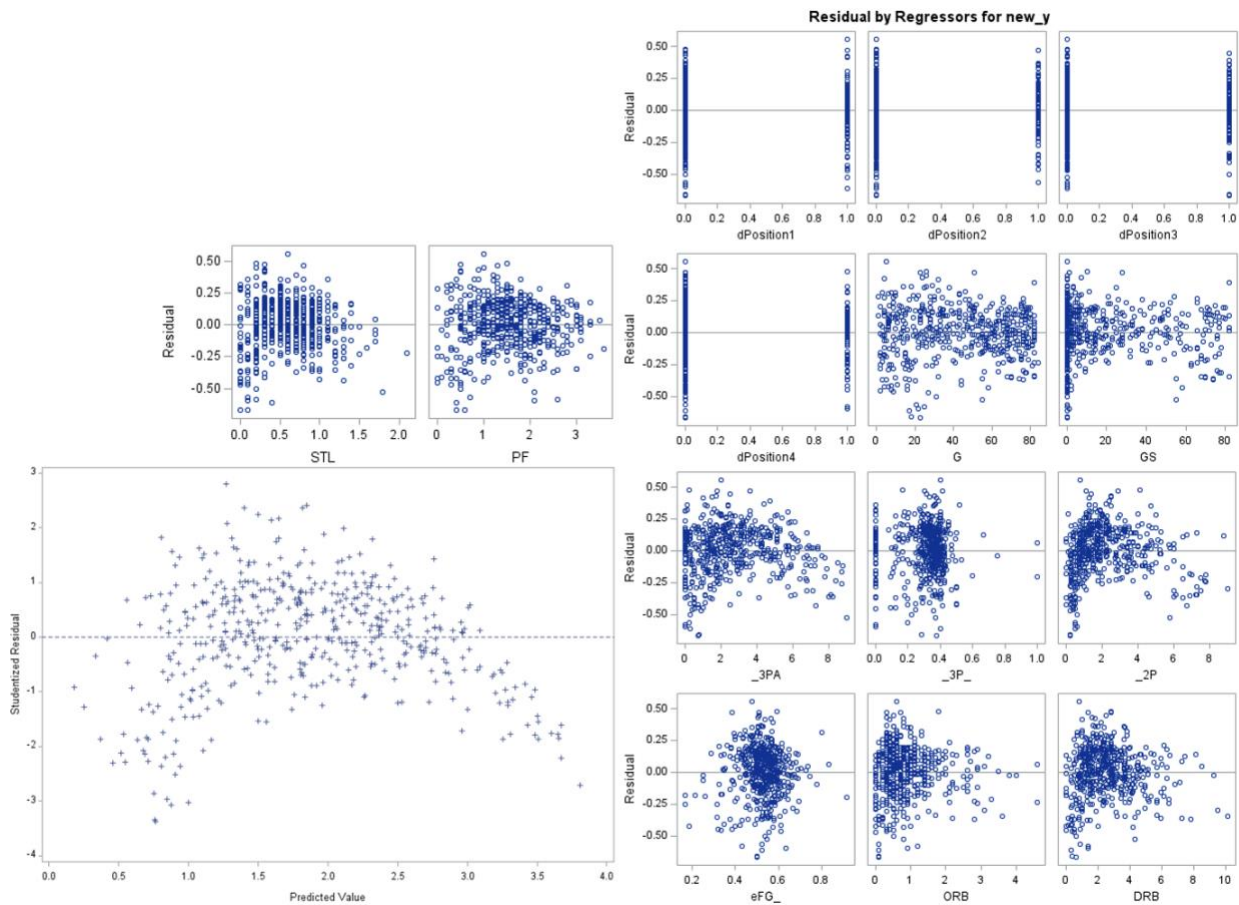
Appendix S:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	272.03489	19.43105	483.54	<.0001
Error	492	19.77099	0.04018		
Corrected Total	506	291.80589			

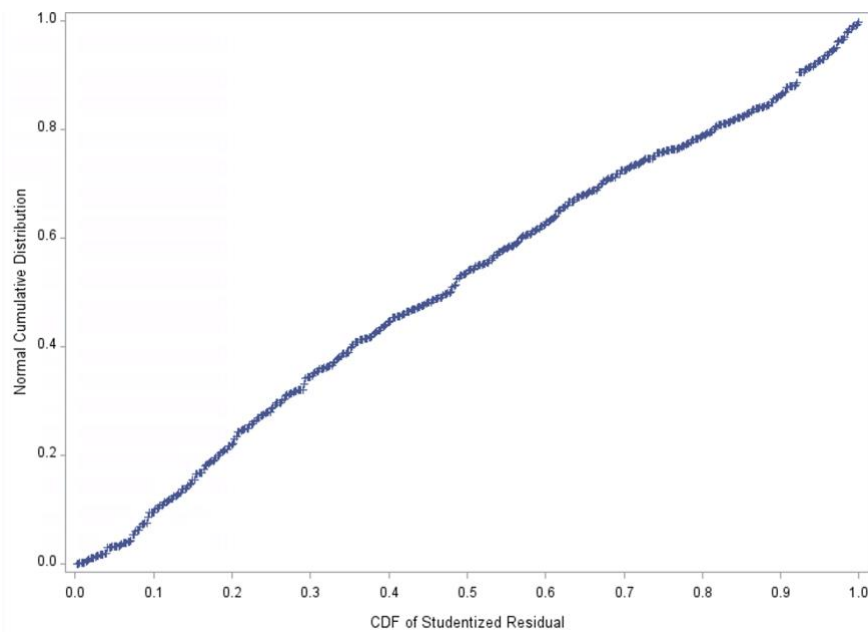
Root MSE	0.20048	R-Square	0.9322
Dependent Mean	1.84094	Adj R-Sq	0.9303
Coeff Var	10.88912		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.13238	0.05961	-2.22	0.0268
dPosition1	1	-0.08246	0.02771	-2.98	0.0031
dPosition2	1	-0.15013	0.02888	-5.20	<.0001
dPosition3	1	-0.13893	0.03007	-4.62	<.0001
dPosition4	1	-0.21994	0.03978	-5.53	<.0001
G	1	0.00392	0.00051977	7.53	<.0001
GS	1	-0.00466	0.00064226	-7.25	<.0001
_3PA	1	0.16149	0.00702	23.00	<.0001
3P	1	0.30026	0.08428	3.56	0.0004
_2P	1	0.20165	0.00933	21.61	<.0001
eFG_	1	1.22387	0.12522	9.77	<.0001
ORB	1	0.14465	0.02276	6.35	<.0001
DRB	1	0.03076	0.01193	2.58	0.0102
STL	1	0.12717	0.03497	3.64	0.0003
PF	1	0.08908	0.01940	4.59	<.0001

Appendix T:



Appendix U:



Appendix V:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	252.52216	16.83481	516.54	<.0001
Error	481	15.67660	0.03259		
Corrected Total	496	268.19876			

Root MSE	0.18053	R-Square	0.9415
Dependent Mean	1.86465	Adj R-Sq	0.9397
Coeff Var	9.68181		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.21717	0.05819	-3.73	0.0002
dPosition1	1	-0.08451	0.02519	-2.56	0.0107
dPosition2	1	-0.12837	0.02620	-4.90	<.0001
dPosition3	1	-0.10533	0.02739	-3.85	0.0001
dPosition4	1	-0.18032	0.03648	-4.94	<.0001
G	1	0.00324	0.00048095	6.74	<.0001
GS	1	-0.00410	0.00058365	-7.02	<.0001
_3PA	1	0.15511	0.00643	24.14	<.0001
3P	1	0.21112	0.08016	2.63	0.0087
_2P	1	0.20049	0.00845	23.74	<.0001
eFG_	1	1.20374	0.11520	10.45	<.0001
ORB	1	0.13165	0.02066	6.37	<.0001
DRB	1	0.02635	0.01081	2.44	0.0152
STL	1	0.12190	0.03202	3.81	0.0002
PF	1	0.08822	0.01780	4.96	<.0001
FT_	1	0.24479	0.04528	5.41	<.0001

Appendix W:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	251.56949	17.96925	520.84	<.0001
Error	482	16.62927	0.03450		
Corrected Total	496	268.19876			

Root MSE	0.18574	R-Square	0.9380
Dependent Mean	1.86465	Adj R-Sq	0.9362
Coeff Var	9.96131		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.11071	0.05834	-1.96	0.0500
dPosition1	1	-0.07097	0.02588	-2.74	0.0063
dPosition2	1	-0.13346	0.02694	-4.95	<.0001
dPosition3	1	-0.11574	0.02811	-4.12	<.0001
dPosition4	1	-0.18156	0.03753	-4.84	<.0001
G	1	0.00365	0.00048883	7.46	<.0001
GS	1	-0.00430	0.00059925	-7.18	<.0001
_3PA	1	0.15913	0.00657	24.23	<.0001
3P	1	0.27658	0.08153	3.39	0.0008
_2P	1	0.20274	0.00868	23.36	<.0001
eFG_	1	1.24834	0.11822	10.56	<.0001
ORB	1	0.12686	0.02124	5.97	<.0001
DRB	1	0.02497	0.01112	2.24	0.0252
STL	1	0.12436	0.03294	3.78	0.0002
PF	1	0.09363	0.01829	5.12	<.0001

Appendix X:

Final-final Model

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	678
Number of Observations Used	497
Number of Observations with Missing Values	181

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	252.52216	16.83481	516.54	<.0001
Error	481	15.67660	0.03259		
Corrected Total	496	268.19876			

Root MSE	0.18053	R-Square	0.9415
Dependent Mean	1.86465	Adj R-Sq	0.9397
Coeff Var	9.68181		

Appendix Y:

Final-final Model

The REG Procedure
Model: MODEL2
Dependent Variable: new_y

Number of Observations Read	678
Number of Observations Used	497
Number of Observations with Missing Values	181

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	251.56949	17.96925	520.84	<.0001
Error	482	16.62927	0.03450		
Corrected Total	496	268.19876			

Root MSE	0.18574	R-Square	0.9380
Dependent Mean	1.86465	Adj R-Sq	0.9362
Coeff Var	9.96131		

Appendix Z:

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	181	0.18941	0.14690

Validation statistics for model

The CORR Procedure

2 Variables: logPTS yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
logPTS	164	1.85492	0.76236	304.20679	0.09531	3.25424	
yhat	181	1.66685	0.93018	301.70000	-0.39100	3.62530	Predicted Value of new_y

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0
Number of Observations

	logPTS	yhat
logPTS	1.00000 164	0.96869 <.0001 164
yhat Predicted Value of new_y	0.96869 <.0001 164	1.00000 181

Appendix AA:

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	181	0.18941	0.14690

Validation statistics for model

The CORR Procedure

2 Variables: logPTS yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
logPTS	164	1.85492	0.76236	304.20679	0.09531	3.25424	
yhat	181	1.67580	0.90884	303.31980	-0.28498	3.62574	Predicted Value of new_y

Pearson Correlation Coefficients

Prob > |r| under H0: Rho=0
Number of Observations

	logPTS	yhat
logPTS	1.00000 164	0.96577 <.0001 164
yhat Predicted Value of new_y	0.96577 <.0001 164	1.00000 181

Appendix BB:


Train

Model 1, CP

RMSE = 0.18053

$R^2 = 94.15\%$

Adj- $R^2 = 93.97$

Residuals = 

GOF = F-Val: 516 P-Val: < 0.001

Model 2, Backward

RMSE = 0.18574

$R^2 = 93.80\%$

Adj- $R^2 = 93.62$

Residuals = 

GOF = F-Val: 520 P-Val: <0.001

Testing

RMSE = 0.18941

MAE = 0.14690

$R^2 = 93.84\%$

Adj- $R^2 = 93.70$

CV- $R^2 = 0.31$

RMSE = 0.18941

MAE = 0.14690

$R^2 = 93.27\%$

Adj- $R^2 = 93.13$

CV- $R^2 = 0.53$

Appendix CC:

Validation statistics for model								
The REG Procedure								
Model: MODEL1								
Dependent Variable: new_y								
Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	3.0994	0.0582	2.9852	3.2137	2.7268	3.4721	.
2	.	1.9127	0.0367	1.8405	1.9849	1.5507	2.2747	.
3	.	3.2220	0.0320	3.1593	3.2848	2.8618	3.5823	.
4	.	2.8167	0.0680	2.6831	2.9504	2.4377	3.1958	.
5	.	4.3246	0.0638	4.1993	4.4500	3.9484	4.7009	.
6	.	3.4853	0.0509	3.3853	3.5853	3.1168	3.8539	.
7	.	3.5603	0.0553	3.4516	3.6691	3.1893	3.9314	.
8	.	3.5373	0.0591	3.4211	3.6535	3.1640	3.9106	.
9	.	3.6012	0.0601	3.4831	3.7192	3.2273	3.9750	.
10	.	3.6657	0.0602	3.5473	3.7840	3.2917	4.0396	.

Appendix DD:

Validation statistics for model								
The REG Procedure								
Model: MODEL1								
Dependent Variable: new_y								
Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	3.1085	0.0598	2.9910	3.2261	2.7251	3.4920	.
2	.	1.9245	0.0377	1.8504	1.9986	1.5521	2.2969	.
3	.	3.2071	0.0328	3.1428	3.2715	2.8365	3.5777	.
4	.	2.8570	0.0696	2.7203	2.9937	2.4673	3.2467	.
5	.	4.2947	0.0654	4.1662	4.4232	3.9077	4.6816	.
6	.	3.5023	0.0523	3.3996	3.6050	3.1232	3.8814	.
7	.	3.5681	0.0569	3.4563	3.6800	3.1864	3.9499	.
8	.	3.5504	0.0608	3.4310	3.6699	3.1664	3.9344	.
9	.	3.6129	0.0618	3.4915	3.7343	3.2283	3.9975	.
10	.	3.6839	0.0619	3.5623	3.8054	3.2992	4.0685	.