PS 230 Project

Matt Watson

December 15, 2017

Estimating Points Per Game for the Top 500 NBA Players of All time

Question of Thesis & Motivation for Research

For this project, I will be analyzing a data set of the "Greatest 500 NBA Players of All Time" in order to find the best predictors for PTS, career average points per game. When ranking players in the NBA, the first variable of importance that comes to mind would likely be Points, as those who are viewed as the best players are often the ones who get a lot of points per game. I hope that by analyzing the relationship between PTS and the other variables in the data set, such as Minutes Played and Shooting Percentage, I will be able to better determine what makes a "great" player.

First off, I will estimate what I think will be the best predictors, and create a linear model containing them as predictors for PTS. Then, my approach to finding the best set of predictors will start by creating a "Full model," in which PTS will be estimated through a covariance adjusted model with all of the other variables in the data set used as predictors. After this, I will create a "Reduced Model," removing the predictors that were not significant in the "Full Model." Then, I will use Akaike Information Criterion (AIC) to select a model. This method of selection creates many models, and then compares them to the full model. The best model is determined as the one with the lowest AIC score. AIC variable selection also requires the data to have no missing values, so this model will show how the 3 point line changed scoring. Last, I will use Bayesian Information Criterion (BIC) to select a model, which is similar to AIC, but prefers smaller models.

Data

The data I will be using can be found at: https://www.basketball-reference.com/awards/slam_500_greatest.html.

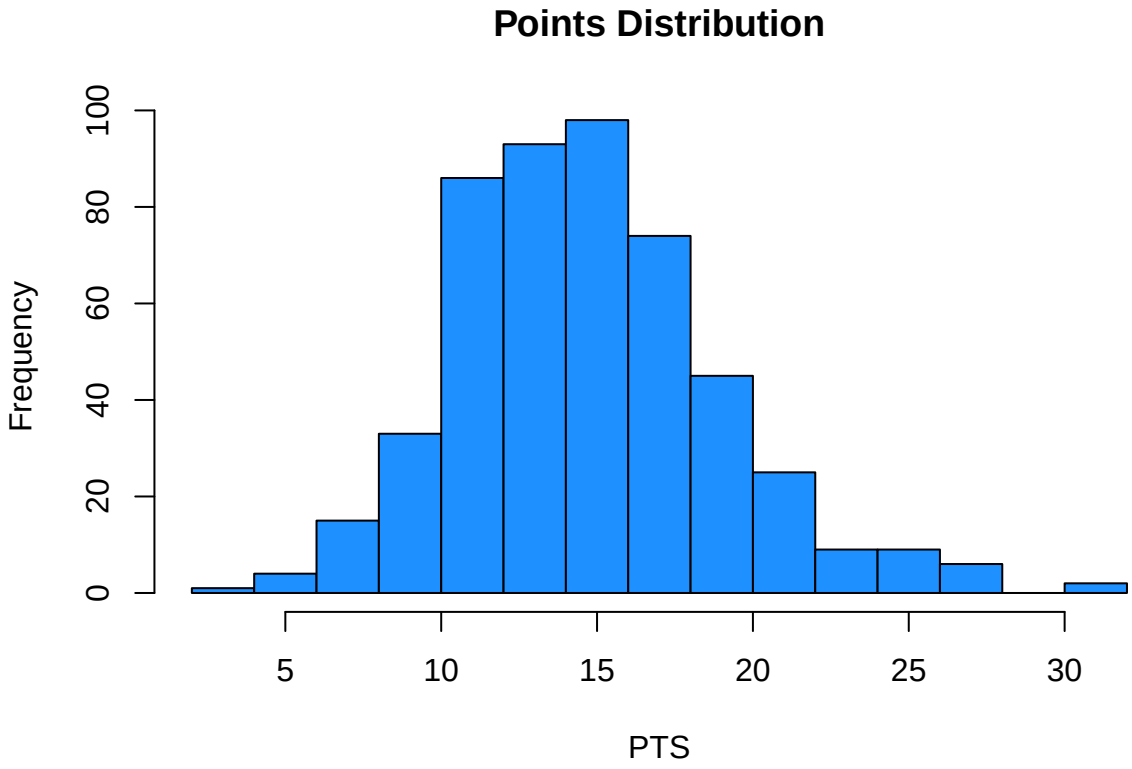
```
## Warning: package 'readxl' was built under R version 3.4.1
```

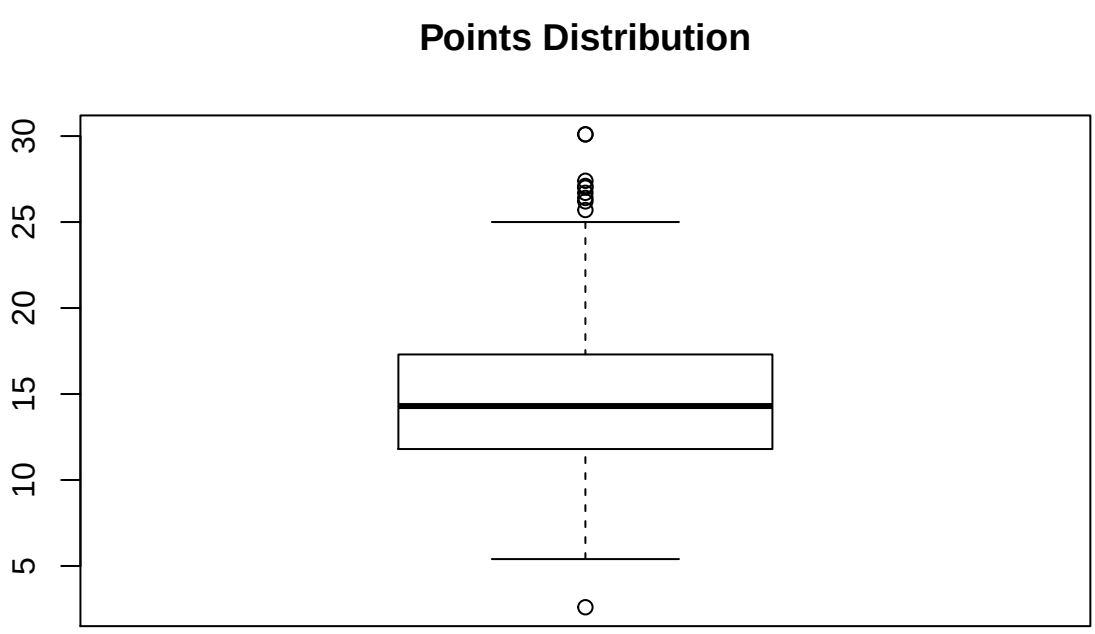
The data was downloaded on November 2 2017. This is worth noting, as there are some active players on this list, whose stats will change slightly throughout the season. The original data set has 16 variables for 500 NBA players. The variable are:

Rank
Player Name From - The year the player's career started, To - The year the player's career ended, G - Total games played in career, MP - Career average minutes played per game, PTS - Career average points per game, TRB - Career average total rebounds per game, AST - Career average assists per game, STL - Career average steals per game, BLK - Career average blocks per game, FGpct - Career average field goal percentage, 3Ppct - Career average three point percentage, FTpct - Career average free throw percentage, 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)

For my analysis, I removed 4 of these variables from the data set: Rank, Player Name, From and To. This process was done through Microsoft Excel, by dowloading a ".csv" file from the link provided. I do not need these variables for the analysis intended. It would be useful if there was a variable for position included in the data, as PTS seems likes it would vary by position. There is not however, and since players often switch positions, it would be difficult to go through and add this variable for each player.

As the title says, the response variable I will be using is PTS. The predictors I will be using include: $G, MP, TRB, AST, STL, BLK, FGpct, 3Ppct, FTpct, WSandWS/48$. All of these variables are numerical. It is worth noting that STL, BLK and $3Ppct$ have some NA responses, as STL and BLK were not recorded until 1973, and the 3-Point line was not used until 1979. It is also worth noting that WS and $WS/48$ are advanced statistics that are not commonly used. I will be creating as subset of that to remvoe all missing values, which will allow me to see how the 3 point line chagned scoring in the NBA.






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.60   11.80   14.30   14.73   17.30   30.10
```

```
## [1] 4.178277
```

Above are some results looking at PTS. From the histogram, it looks like PTS is roughly normally distributed, with a mean of 14.7 and a standard deviation of 4.178. We see that PTS has a range of (2.6, 30.1). Looking at the boxplot, there are some points that could potentially be outliers, especailly the max and min values.

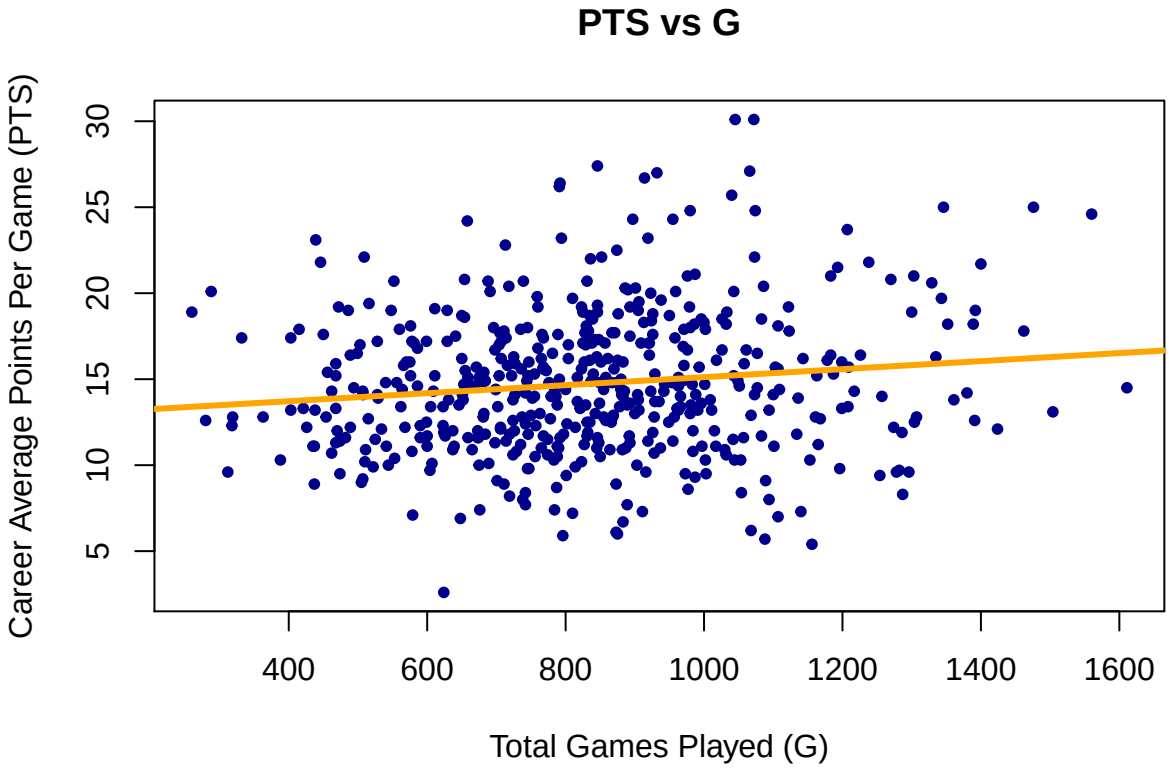
```
##        G                MP             PTS             TRB
##  Min.   : 260.0   Min.   :18.70   Min.   : 2.60   Min.   : 1.700
##  1st Qu.: 677.5   1st Qu.:27.70   1st Qu.:11.80   1st Qu.: 3.500
##  Median : 831.5   Median :30.65   Median :14.30   Median : 5.500
##  Mean   : 834.7   Mean   :30.54   Mean   :14.73   Mean   : 5.948
##  3rd Qu.: 977.5   3rd Qu.:33.40   3rd Qu.:17.30   3rd Qu.: 7.800
##  Max.   :1611.0   Max.   :45.80   Max.   :30.10   Max.   :22.900
##
##       AST              STL              BLK              FGpct
##  Min.   : 0.300   Min.   :0.200   Min.   :0.0000   Min.   :0.3020
##  1st Qu.: 1.900   1st Qu.:0.800   1st Qu.:0.2000   1st Qu.:0.4410
##  Median : 2.750   Median :1.000   Median :0.4000   Median :0.4635
##  Mean   : 3.213   Mean   :1.049   Mean   :0.6083   Mean   :0.4641
##  3rd Qu.: 4.000   3rd Qu.:1.300   3rd Qu.:0.8000   3rd Qu.:0.4890
##  Max.   :11.200   Max.   :2.700   Max.   :3.5000   Max.   :0.5990
##                   NA's   :67      NA's   :67
##      3Ppct            FTpct              WS              WS/48
##  Min.   :0.0000   Min.   :0.4140   Min.   : -7.90   Min.   :-0.0450
##  1st Qu.:0.1670   1st Qu.:0.7220   1st Qu.: 38.90   1st Qu.: 0.0920
```

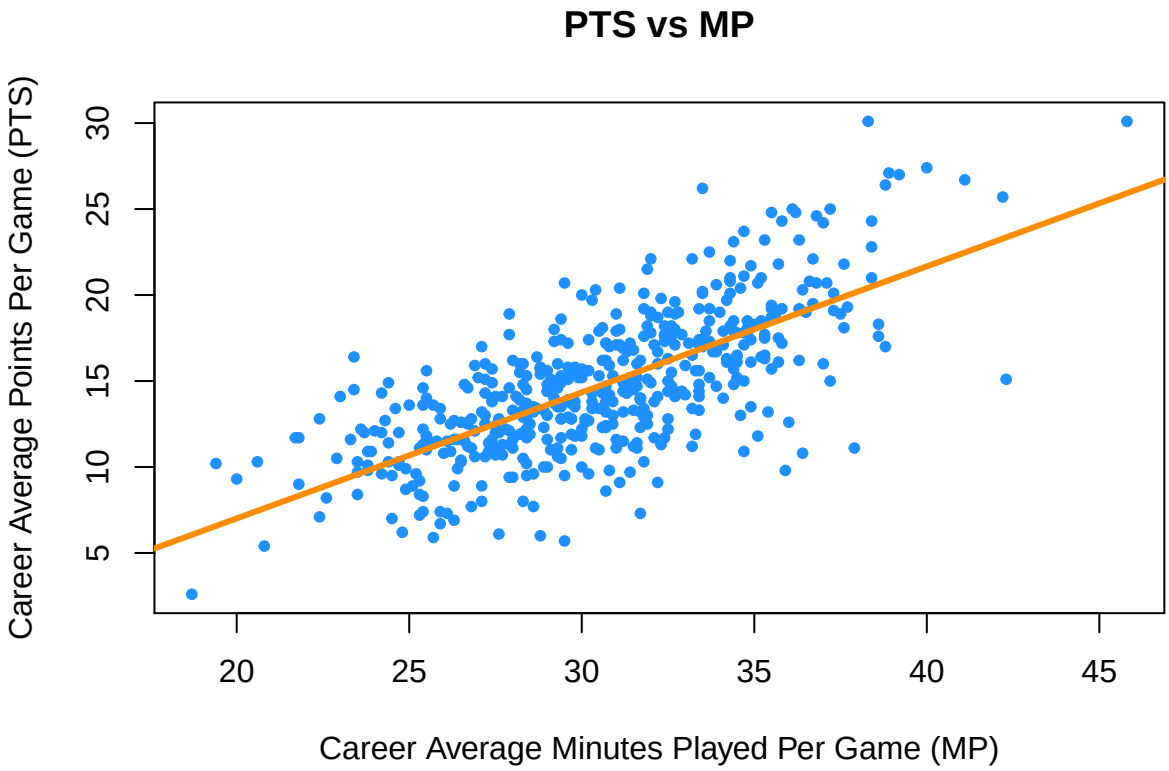
```
##  Median :0.2910   Median :0.7690   Median : 58.65   Median : 0.1155
##  Mean   :0.2551   Mean   :0.7614   Mean   : 66.71   Mean   : 0.1181
##  3rd Qu.:0.3440   3rd Qu.:0.8075   3rd Qu.: 83.03   3rd Qu.: 0.1390
##  Max.   :1.0000   Max.   :0.9050   Max.   :273.40   Max.   : 0.2500
##  NA's   :121
```

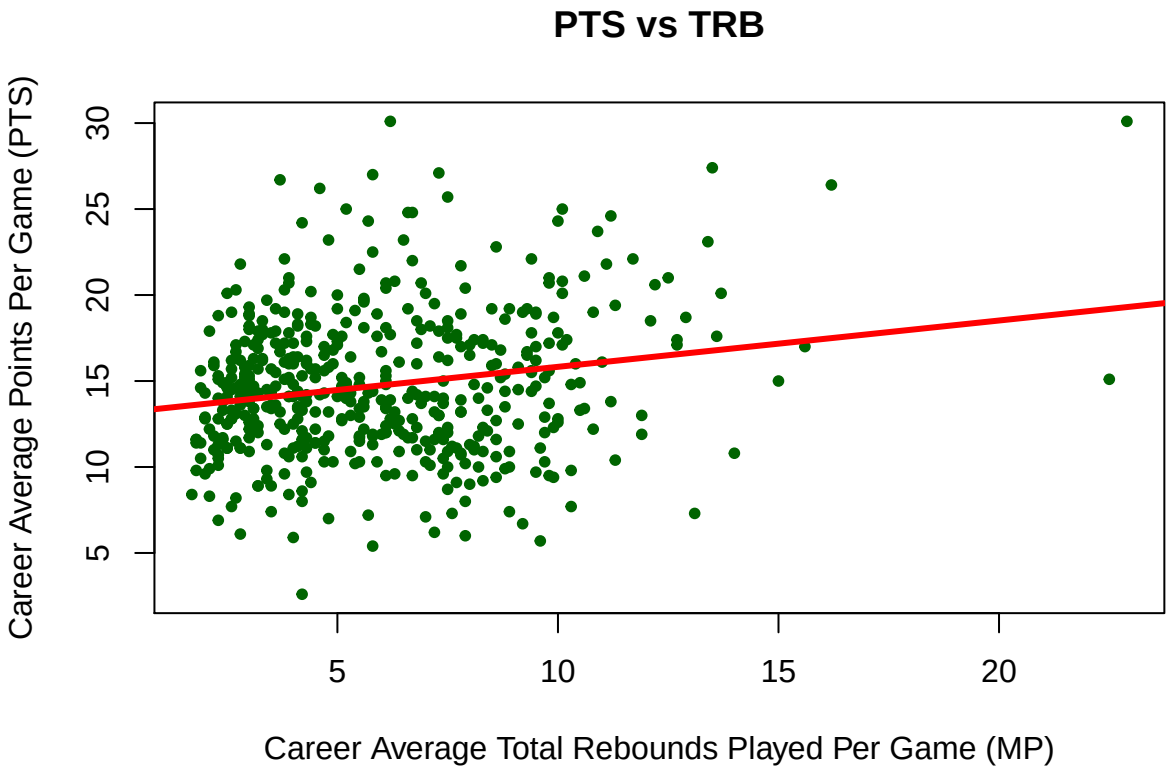
Here, we see descriptive statistics for each of the variables. We see that G ranges from 260 to 1611 games played with a mean of 834.7 games. MP ranges from 18.70 to 45.80 minutes per game with a mean of 30.54 minutes. TRB ranges from 1.7 to 22.9 rebounds per game with a mean of 5.948 rebounds. AST ranges from 0.3 to 11.2 assists per game with a mean of 3.213 assists. STL ranges from 0.2 to 2.7 steals per game with a mean of 1.049 stealas. BLK ranges from 0.0 to 3.5 blocks per game with a mean of 0.6083 blocks. $FGpct$ ranges from 0.3020 to 0.599 percent with a median of 0.4641 percent. $3Ppct$ ranges from 0.0 to 1.0 percent with a mean of 0.2551 percent. $FTpct$ ranges from 0.4140 to 0.9050 percent with a mean of 0.7614 percent. WS ranges from -7.9 to 273.4 wins with a mean of 58.65 wins. $WS/48$ ranges from -0.045 to 0.25 wins per 48 with a mean of 0.1181 wins per 48.

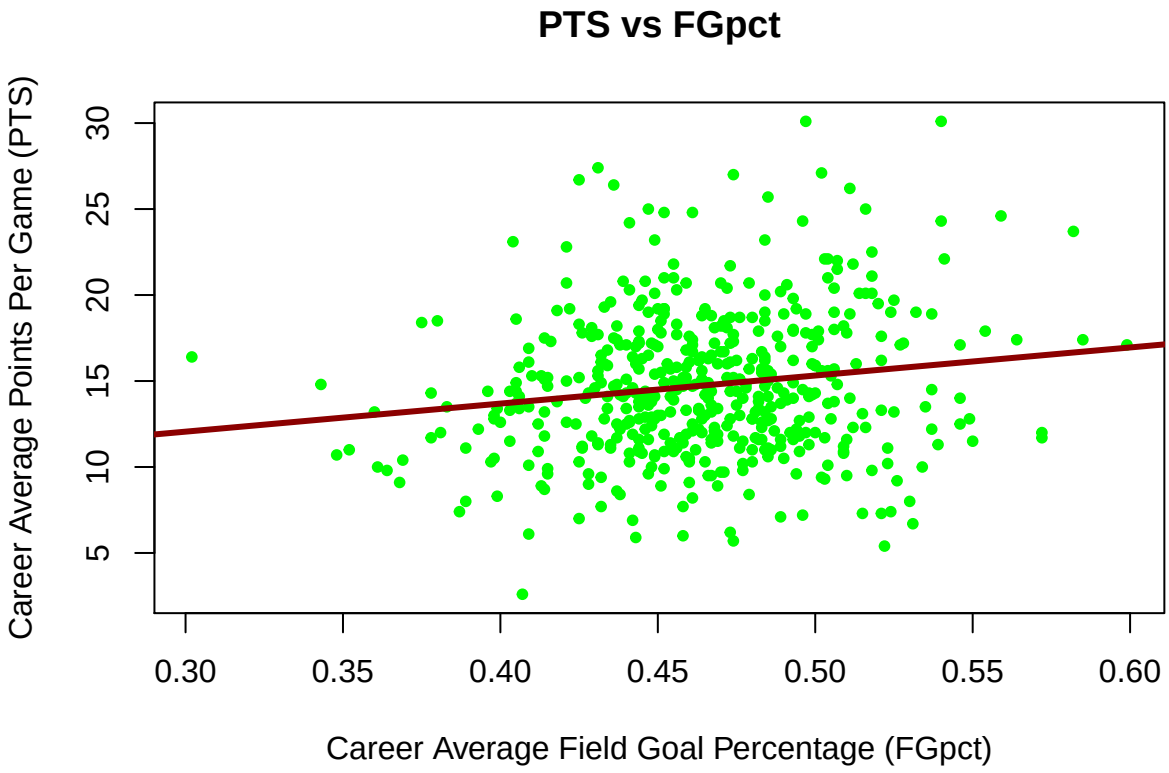
Expectations

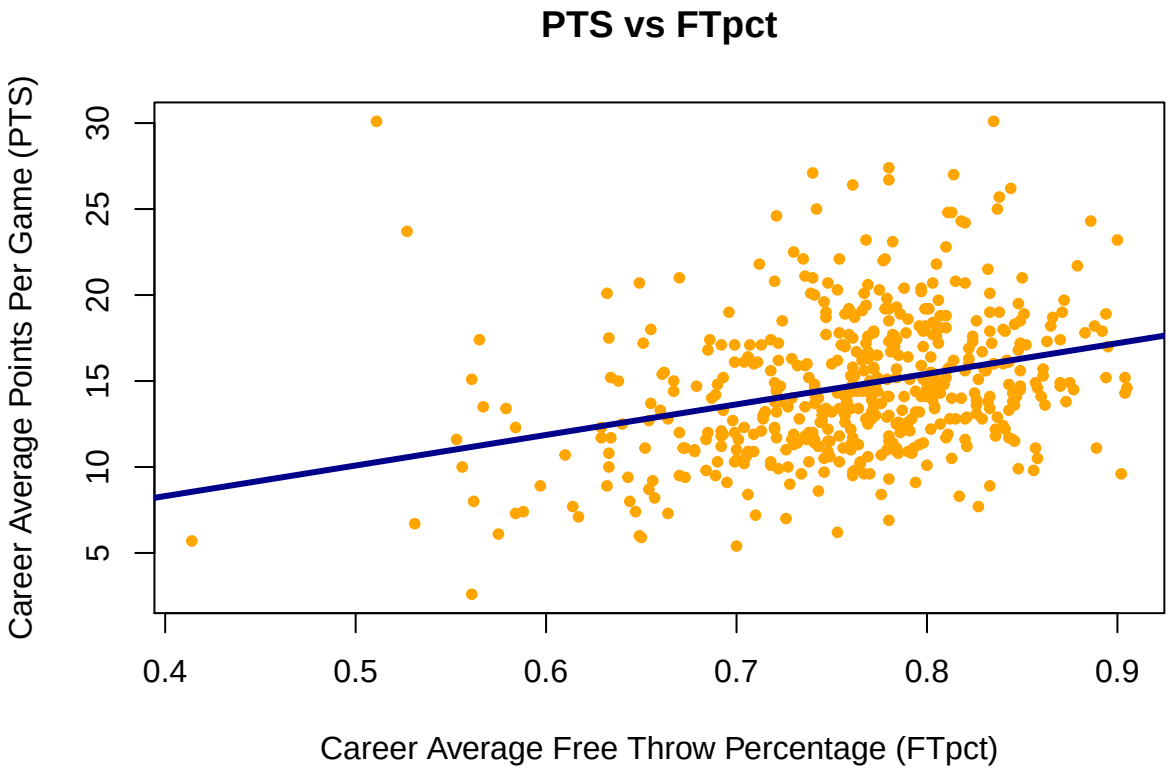
If I had to guess which variables I think will be best for predicting PTS (Points), I would say G (Games), MP (Minutes Played), TRB (Total Rebounds), $FGpct$ (Field goal Percentage) and $FTpct$ (Free throw Percentage). G seems like it should have a positive relationship with PTS, as players who are good at scoring tend to player more games. MP seems like it should have a positive relationship with PTS, as the more time you in play in the game, the more points you should score. TRB should have a positive relationship with PTS as well, as getting an offensive rebound likely increases your chance of scoring. $FGpct$ and $FTpct$ should also both have positve relationships with PTS, as if you make a higher percentage of your shots, you will be scoring more points. I will first look at the relationship between PTS and these 5 other variables.




PTS vs MP
Career Average Points Per Game (PTS)
5
10
15
20
25
30
20
25
30
35
40
45
Career Average Minutes Played Per Game (MP)


PTS vs TRB
Career Average Points Per Game (PTS)
5
10
15
20
25
30
5
10
15
20
Career Average Total Rebounds Played Per Game (MP)


PTS vs FGpct
Career Average Points Per Game (PTS)
5
10
15
20
25
30
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Career Average Field Goal Percentage (FGpct)



From the plots above, it looks like all 5 of these variables have a possitive linear relationship with PTS, with MP having the strongest relationship. For the other four variables, it is harder to tell if there is a linear relationship, but all of them look like they have possitive slopes.

```
##                G           MP       PTS         TRB        AST STL BLK
## G     1.00000000  0.216099297 0.1326825  0.05361998  0.1119511  NA  NA
## MP    0.21609930  1.000000000 0.7110896  0.34344854  0.3752757  NA  NA
## PTS   0.13268248  0.711089602 1.0000000  0.19593999  0.2047721  NA  NA
## TRB   0.05361998  0.343448542 0.1959400  1.00000000 -0.3714160  NA  NA
## AST   0.11195108  0.375275746 0.2047721 -0.37141601  1.0000000  NA  NA
## STL           NA           NA        NA          NA         NA   1  NA
## BLK           NA           NA        NA          NA         NA  NA   1
## FGpct 0.17281969 -0.009709238 0.1575852  0.22677307 -0.1683257  NA  NA
## 3Ppct         NA           NA        NA          NA         NA  NA  NA
## FTpct 0.05900439  0.106708492 0.3066018 -0.48373675  0.2986736  NA  NA
## WS    0.71515476  0.564224311 0.5386436  0.35151158  0.2300516  NA  NA
## WS/48 0.32849394  0.403824487 0.4855204  0.38435107  0.1245558  NA  NA
##              FGpct 3Ppct       FTpct         WS     WS/48
## G      0.172819686    NA  0.05900439 0.71515476 0.3284939
## MP    -0.009709238    NA  0.10670849 0.56422431 0.4038245
## PTS    0.157585246    NA  0.30660176 0.53864363 0.4855204
## TRB    0.226773074    NA -0.48373675 0.35151158 0.3843511
## AST   -0.168325697    NA  0.29867362 0.23005161 0.1245558
## STL             NA    NA          NA         NA        NA
## BLK             NA    NA          NA         NA        NA
## FGpct  1.000000000    NA -0.14550888 0.31369303 0.4325540
## 3Ppct           NA     1          NA         NA        NA
```

```
## FTpct -0.145508877    NA  1.00000000 0.09190363 0.1292245
## WS     0.313693032    NA  0.09190363 1.00000000 0.8024431
## WS/48  0.432554024    NA  0.12922452 0.80244307 1.0000000
```

G and WS (Win Shares) have a high correlation, at around 0.72. This is interesting, it implies that players who have played in more games account for a higher Win share percentage. This makes sense, as the way Win Shares is calculated, it give players who have played in a larger number of games a higher value for WS.

PTS and MP seem have to high correlation, at around 0.71, but that makes sense because the player needs to be on the court in order to score points.

WS and $WS/48$ (Win Shares pers 48 minutes) have a high correlation, at around 0.8, which makes since as the variables are very closely related.

The last point of interest here is the correlation between $FTpct$ and TRB, at -0.48. This likely represents how Power Forwards and Centers, who tend to get more rebounds, are often poor free throw shooters. There are some other negative correlations, but the rest are much smaller values.

Analysis of Data

```
##
## Call:
## lm(formula = PTS ~ G + MP + TRB + FGpct + FTpct, data = NBA2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2734 -1.6861 -0.1253  1.6948  9.6069
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.913e+01  2.166e+00 -13.446  < 2e-16 ***
## G           -1.220e-03  5.171e-04  -2.358   0.0188 *
## MP           6.930e-01  3.368e-02  20.579  < 2e-16 ***
## TRB          9.390e-02  5.078e-02   1.849   0.0650 .
## FGpct        2.118e+01  3.070e+00   6.901 1.59e-11 ***
## FTpct        1.750e+01  1.982e+00   8.828  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.632 on 494 degrees of freedom
## Multiple R-squared:  0.6071, Adjusted R-squared:  0.6031
## F-statistic: 152.6 on 5 and 494 DF,  p-value: < 2.2e-16
```

The output above shows the results for my Guessed Model, using G (Games), MP (Minutes Played), TRB (Total Rebounds), $FGpct$ (Field goal Percentage) and $FTpct$ (Free throw Percentage) as predictors for PTS (Points). We see that all of the predictors besides TRB are signifcant at the $\alpha = 0.05$ level. We also see that our model is significant. Our R^2 value of 0.607 tells us that roughly 61% of the variation in PTS can be explained by these predictors. I will now create a Full Model, using all of the variables as predictors for PTS.

```
##
## Call:
## lm(formula = PTS ~ G + MP + TRB + AST + STL + BLK + FGpct + `3Ppct` +
##     FTpct + WS + `WS/48`, data = NBA2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.5533 -1.3991 -0.0168  1.5611  5.9824
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.993292   4.004142  -4.244 2.79e-05 ***
## G            -0.008066   0.001253  -6.435 3.87e-10 ***
## MP            0.679641   0.052311  12.992  < 2e-16 ***
## TRB          -0.345748   0.085974  -4.022 7.02e-05 ***
## AST          -0.605461   0.101231  -5.981 5.27e-09 ***
## STL           0.448886   0.420891   1.067 0.286892
## BLK          -0.277549   0.295416  -0.940 0.348081
## FGpct        20.382874   5.322458   3.830 0.000151 ***
## `3Ppct`      -0.583382   1.285784  -0.454 0.650301
## FTpct        12.805780   2.345571   5.460 8.81e-08 ***
## WS            0.077039   0.013599   5.665 2.97e-08 ***
## `WS/48`     -25.254204   9.752807  -2.589 0.009996 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.354 on 367 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.6862, Adjusted R-squared:  0.6768
## F-statistic: 72.97 on 11 and 367 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = PTS ~ . - STL - BLK - `3Ppct`, data = NBA2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5390 -1.4568 -0.0418  1.5080  6.3777
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.103553   3.597702  -4.754 2.86e-06 ***
## G            -0.008067   0.001237  -6.522 2.29e-10 ***
## MP            0.689783   0.050872  13.559  < 2e-16 ***
## TRB          -0.370755   0.080648  -4.597 5.89e-06 ***
## AST          -0.535092   0.089317  -5.991 4.95e-09 ***
## FGpct        20.250854   4.707386   4.302 2.17e-05 ***
## FTpct        12.765066   2.209318   5.778 1.61e-08 ***
## WS            0.075849   0.013422   5.651 3.19e-08 ***
## `WS/48`     -24.907253   9.624324  -2.588     0.01 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.352 on 370 degrees of freedom
##   (121 observations deleted due to missingness)
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.6774
## F-statistic: 100.2 on 8 and 370 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: PTS ~ G + MP + TRB + AST + STL + BLK + FGpct + `3Ppct` + FTpct +
##     WS + `WS/48`
```

```
## Model 2: PTS ~ (G + MP + TRB + AST + STL + BLK + FGpct + `3Ppct` + FTpct +
##     WS + `WS/48`) - STL - BLK - `3Ppct`
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    367 2033.5
## 2    370 2046.7 -3   -13.198 0.794 0.4978
```

The first results are from a running the full model, using all of the predictors for PTS, We can see that all of the predictors besides Steals (STL), Blocks (BLK) and 3 Point Percentage ($3Ppct$) are significant, when using a significance level of $\alpha = 0.05$, in our model. We also see that our model itself is significant, with a p-value less than 2.2e-16, which is esentially equal to 0. From our R^2 value of 0.6862, we can say that roughly 68.62% of the variation in PTS can be explained by these predictors. Below, we refit the model, exluding the 3 predictors listed above.

The second results are from running the Reduced Model. Here, we see that all of the predictors are significant in the model, when using a significance level of $\alpha = 0.05$. Again, the model is also significant. Interestingly, we have a lower R^2 value, 0.6842, than our full model, meaning that the full model is better.

The ANOVA results reinforce that the first, or Full, model is the better of the two, which we know from the high p-value of 0.5.

```
##
## Call:
## lm(formula = PTS ~ G + MP + TRB + AST + FGpct + FTpct + WS +
##     `WS/48`, data = NBA2017_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5390 -1.4568 -0.0418  1.5080  6.3777
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.103553   3.597702  -4.754 2.86e-06 ***
## G            -0.008067   0.001237  -6.522 2.29e-10 ***
## MP            0.689783   0.050872  13.559  < 2e-16 ***
## TRB          -0.370755   0.080648  -4.597 5.89e-06 ***
## AST          -0.535092   0.089317  -5.991 4.95e-09 ***
## FGpct        20.250854   4.707386   4.302 2.17e-05 ***
## FTpct        12.765066   2.209318   5.778 1.61e-08 ***
## WS            0.075849   0.013422   5.651 3.19e-08 ***
## `WS/48`     -24.907253   9.624324  -2.588     0.01 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.352 on 370 degrees of freedom
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.6774
## F-statistic: 100.2 on 8 and 370 DF,  p-value: < 2.2e-16
```

In order to perfrom AIC variable selection, we must remove any cases that have a NA response. This is why only 379 of the 500 cases were used in this analysis. We get a R^2 value of roughly 68.42% for this model, which is about the same as both of the models above. AIC selection is not very good at predicting values outside of the dataset, but it rather focuses on finding the best fit for the data.

Removing the missing cases will also allow to me to look at how the addition of 3-Point line changed the game. Interestingly, none of three variables of interest (that had missing values), BLK, STL and $3Ppct$ are significant in the model. This leads me to believe that the addition of the 3 point line did not really change scoring too much. Or at least because most players do not tend to shoot many 3 pointers, this variable is not important for predicting PTS.

```
## Start:  AIC=707.96
## PTS ~ G + MP + TRB + AST + STL + BLK + FGpct + `3Ppct` + FTpct +
##     WS + `WS/48`
##
##           Df Sum of Sq    RSS    AIC
## - `3Ppct`  1      1.14 2034.6 702.23
## - BLK      1      4.89 2038.4 702.93
## - STL      1      6.30 2039.8 703.19
## <none>                 2033.5 707.96
## - `WS/48`  1     37.15 2070.6 708.88
## - FGpct    1     81.26 2114.7 716.87
## - TRB      1     89.61 2123.1 718.36
## - FTpct    1    165.15 2198.6 731.62
## - WS       1    177.81 2211.3 733.79
## - AST      1    198.21 2231.7 737.27
## - G        1    229.42 2262.9 742.54
## - MP       1    935.28 2968.8 845.43
##
## Step:  AIC=702.23
## PTS ~ G + MP + TRB + AST + STL + BLK + FGpct + FTpct + WS + `WS/48`
##
##           Df Sum of Sq    RSS    AIC
## - BLK      1      4.12 2038.7 697.06
## - STL      1      6.63 2041.2 697.53
## <none>                 2034.6 702.23
## - `WS/48`  1     39.43 2074.1 703.57
## - TRB      1     88.83 2123.4 712.49
## - FGpct    1    105.45 2140.1 715.45
## - FTpct    1    164.11 2198.7 725.69
## - WS       1    177.19 2211.8 727.94
## - AST      1    197.08 2231.7 731.34
## - G        1    229.02 2263.7 736.72
## - MP       1    945.99 2980.6 841.01
##
## Step:  AIC=697.06
## PTS ~ G + MP + TRB + AST + STL + FGpct + FTpct + WS + `WS/48`
##
##           Df Sum of Sq    RSS    AIC
## - STL      1      7.94 2046.7 692.60
## <none>                 2038.7 697.06
## - `WS/48`  1     39.95 2078.7 698.48
## - TRB      1    108.24 2147.0 710.73
## - FGpct    1    110.28 2149.0 711.09
## - WS       1    173.08 2211.8 722.01
## - FTpct    1    192.59 2231.3 725.33
## - AST      1    193.03 2231.8 725.41
## - G        1    224.93 2263.7 730.79
## - MP       1    973.24 3012.0 839.04
##
## Step:  AIC=692.6
## PTS ~ G + MP + TRB + AST + FGpct + FTpct + WS + `WS/48`
##
##           Df Sum of Sq    RSS    AIC
## <none>                 2046.7 692.60
```

```
## - `WS/48`  1     37.05 2083.7 693.46
## - FGpct    1    102.37 2149.1 705.16
## - TRB      1    116.91 2163.6 707.71
## - WS       1    176.66 2223.3 718.04
## - FTpct    1    184.66 2231.3 719.40
## - AST      1    198.54 2245.2 721.75
## - G        1    235.26 2281.9 727.90
## - MP       1   1016.99 3063.7 839.55
```

The model above was fitted using the Bayes Information Criterion (BIC) variable selection method. This method is similar to AIC, but it penalizes larger models more, so we would expect to see a smaller model. This is not the case, however, as we get the same model selected from AIC. We see that $3Ppct$ was removed first, and then BLK and then STL. This would lead us to believe that the full model is our best model.

Discussion of Results

Interestingly, the full model is the best fitting model I found, as it has the highest R^2 value of 0.6862. Taking out insignificant predictors from the model resulted in a lower R^2 value, which is not what I was expecting. This could mean that there are possibly some relationships between the various predictors. Overall, I would say that it is hard to estimate a player's points per game using his other statistics.

Something else that I found interesting was that the addition of the 3 point does not seem to important for predicting PTS (Points). When I performed Step Wise variable selection, I had to remove all cases with a NA value. Only Steals (STL), Blocks (BLK) and 3 Point Percentage ($3Ppct$) have missing cases. So this variable selection only involved NBA players since 1979. I was expecting to see that $3Ppct$ would be significant in the model, but this is not what I found. Just like in the original full model, none of those 3 variables are significant. I think STL might be an intersting variable to look at if we had player positions, as Guards often get more stealas and then could score on a fast break. Blocks by a Center/Power Forward could lead to a long pass and an assist. I think that $3Ppct$ is not significant in predicting PTS because most players do not take many 3 point shots, besides the ones who are really good and take a lot.

Bibliography

[NBA2017] can be found at: https://www.basketball-reference.com/awards/slam_500_greatest.html

The list of players was selected by SLAM Magazine in 2011.

Code Appendix

```
###Data and Variables###
install.packages("readxl")

library(readxl)
NBA2017 <- read_excel("~/PS 230/NBA.xlsx")
View(NBA2017)

hist(NBA2017$PTS, main="Points Distribution", xlab="PTS", col="dodgerblue")
boxplot(NBA2017$PTS, main="Points Distribution", xlab="PTS")
summary(NBA2017$PTS)
sd(NBA2017$PTS)
```

```
###Expectations###
attach(NBA2017)
lm1 = lm(PTS ~ G, data=NBA2017)
plot(PTS ~ G,
     main="PTS vs G",
     xlab="Total Games Played (G)",
     ylab="Career Average Points Per Game (PTS)",
     pch=20,
     col="darkblue")
abline(lm1, lwd=3, col="orange")

lm2 = lm(PTS ~ MP, data=NBA2017)
plot(PTS ~ MP,
     main="PTS vs MP",
     xlab="Career Average Minutes Played Per Game (MP)",
     ylab="Career Average Points Per Game (PTS)",
     pch=20,
     col="dodgerblue")
abline(lm2, lwd=3, col="darkorange")

lm3 = lm(PTS ~ TRB, data=NBA2017)
plot(PTS ~ TRB,
     main="PTS vs TRB",
     xlab="Career Average Total Rebounds Played Per Game (MP)",
     ylab="Career Average Points Per Game (PTS)",
     pch=20,
     col="darkgreen")
abline(lm3, lwd=3, col="red")

lm4 = lm(PTS ~ FGpct, data=NBA2017)
plot(PTS ~ FGpct,
     main="PTS vs FGpct",
     xlab="Career Average Field Goal Percentage (FGpct)",
     ylab="Career Average Points Per Game (PTS)",
     pch=20,
     col="green")
abline(lm4, lwd=3, col="darkred")

lm5 = lm(PTS ~ FTpct, data=NBA2017)
plot(PTS ~ FTpct,
     main="PTS vs FTpct",
     xlab="Career Average Free Throw Percentage (FTpct)",
     ylab="Career Average Points Per Game (PTS)",
     pch=20,
     col="orange")
abline(lm5, lwd=3, col="darkblue")

cor(NBA2017)

summary(NBA2017)


###Analysis and Results###
```

```r
lm_guess = lm(PTS ~ G + MP + TRB + FGpct + FTpct, data=NBA2017)
summary(lm_guess)

lm_pts_full = lm(PTS ~ G + MP + TRB + AST + STL + BLK + FGpct + `3Ppct` + FTpct
                 + WS + `WS/48`, data=NBA2017)
summary(lm_pts_full)

lm_pts2 = lm(PTS ~ . - STL - BLK - `3Ppct`, data = NBA2017)
summary(lm_pts2)

anova(lm_pts_full, lm_pts2)

NBA2017_cleaned = na.omit(NBA2017)
lm_pts_f2 = lm(PTS ~ ., data=NBA2017_cleaned)
lm_pts_aic = step(lm_pts_f2, trace=0)
summary(lm_pts_aic)

n = length(resid(lm_pts_f2))
lm_BIC = step(lm_pts_f2, direction = "backward", k=log(n))
```