ST362 Final Project Report

Zach Reid

2025-11-11

Introduction

Within the world of sports, data analytics are surging to the forefront of the concerns of every team. Especially in the NBA, teams are more concerned with the best mathematical product they can put on the floor, sometimes at the cost of team chemistry or entertaining basketball. This has opened the door for a new wave of predictive analytics, but one sports trend has yet to be explained: the playoff performer. It is my goal to analyze this phenomenon using real player data from last season, and determine through careful consideration of statistics and intangibles whether being a 'playoff performer' is a predictable skill that players can learn, or whether it is simply something the best athletes are born with.

Objectives

Using data from the 2023-2024 NBA regular season and playoffs, it is my goal to analyze the difference between regular season and playoff performances by each player who participated in the playoffs. With this information, I hope to isolate the factors that make up a playoff performer, whether it has to do with size, skillset, or simply an intangible factor that escapes the average athlete. This will be accomplished through several methods of analysis, with a focus on linear regression to attempt to predict how the playoffs affect performance, and variable selection to determine which skills the best playoff performers have in common.

Data Description

This data was collected from a kaggle dataset, found at this link: https://www.kaggle.com/datasets/vivovinco/2023-2024-nba-player-stats/data?select=2023-2024+NBA+Player+Stats+-+Regular.csv

The data was sourced from the website basketballreference.com, a site hosting the largest NBA statistical database on the internet. The data within the database is taken directly from the NBA and its games.

Several variables are measured, but there are a few important ones that we will focus on. They are described below:

Player: Name of the player

Pos: Position of the player

Age: Player's Age

MP: Minutes Played Per Game

FG%: Field Goal Percentage, the percentage of shots a player makes versus how many they take.

3P%: Three Point Percentage, the percentage of shots a player makes from beyond the three-point arc.

2P%: Two Point Percentage, the percentage of shots a player makes from within the three-point arc.

eFG%: Effective Field Goal Percentage. Same as FG%, but every three point make counts for 1.5, adjusting for the extra point on the shot.

FT%: Free Throw Percentage, the percentage of foul shots a player makes from the free-throw line.

FTA: Free Throw Attempts, the amount of foul shots a player is awarded per game.

TRB: Total Rebounds Per Game

AST: Assists Per Game

STL: Steals Per Game

BLK: Blocks Per Game

TOV: Turnovers Per Game

PF: Personal Fouls Per Game

PTS: Points Per Game

The sample size is 214, the players who participated in the Playoffs (all NBA players participate in the regular season, barring injury). This also eliminated some players with a hybrid position to improve visual output. These players were not entirely significant to the sample.

Exploratory Data Analysis

To begin analysis, we must first establish who the best players in this league are in a few key areas.

Let us begin with some summary statistics across the whole league, and then grouped by position.

```
## [1] "Regular Season Summary Statistics:"

## # A tibble: 1 × 9
##   Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
Median_TRB
```

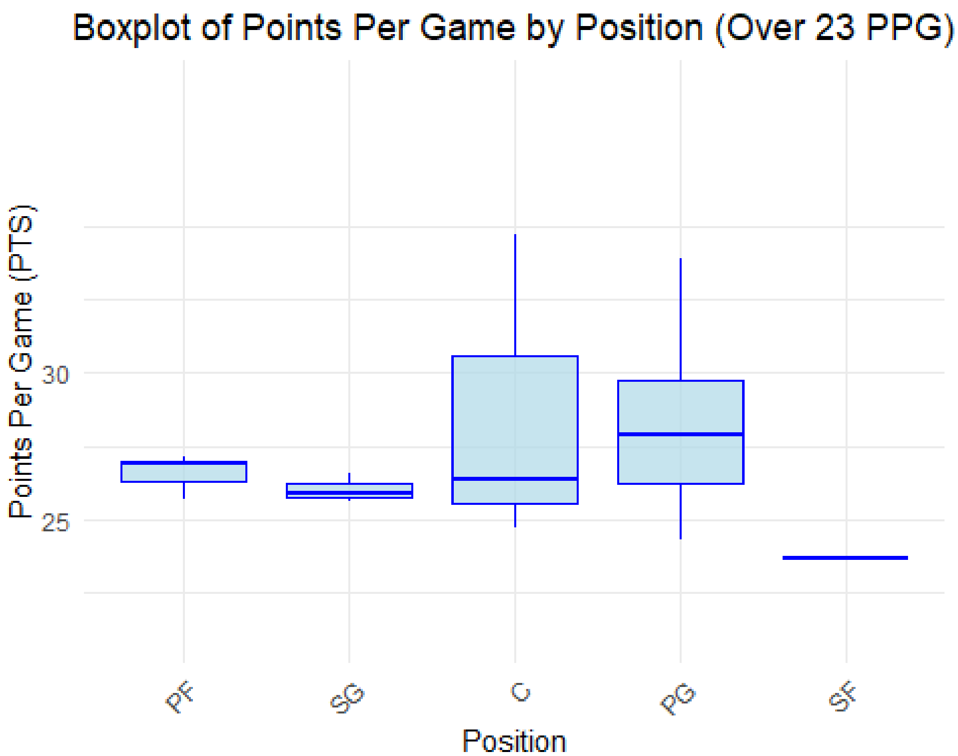
```
##      <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
<dbl>
## 1     10.1        8.1   7.33     2.38        1.6   2.02     3.84
3.4
## # ℹ 1 more variable: SD_TRB <dbl>

## [1] "Regular Season Summary Statistics by Position:"

## # A tibble: 5 × 10
##   Pos   Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
##   <chr>    <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
## 1 C         9.88        7.3   7.26     1.76        1.3   1.60     6.17
## 2 PF        9.56        6.4   7.95     1.96        1.3   1.78     3.93
## 3 PG       12.7        10.2   8.61     4.36        4.1   2.47     3.12
## 4 SF        9.08        8.6   5.92     1.82        1.5   1.36     3.21
## 5 SG        9.84        8.2   6.69     2.26        1.6   1.72     2.74
## # ℹ 2 more variables: Median_TRB <dbl>, SD_TRB <dbl>

## [1] "Playoff Summary Statistics:"

##   Mean_PTS Median_PTS   SD_PTS Mean_AST Median_AST   SD_AST Mean_TRB
Median_TRB
## 1 8.445327       5.65 8.244609 1.807944          1 1.999526  3.46729
2.8
##     SD_TRB
## 1 2.975811

## [1] "Playoff Summary Statistics by Position:"

## # A tibble: 5 × 10
##   Pos   Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
##   <chr>    <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
## 1 C         8.52       6      8.44     1.44       1      1.69     5.23
## 2 PF        7.84       5      8.48     1.50       0.65   1.95     3.76
## 3 PG       11.4        7.5   10.0      3.39       2.2    2.64     2.83
## 4 SF        6.93       4.5    6.52     1.29       1      1.27     3.09
## 5 SG        8.06       5.35   7.43     1.65       1      1.68     2.43
## # ℹ 2 more variables: Median_TRB <dbl>, SD_TRB <dbl>
```

A couple of key takeaways from the summary analysis I performed: The average statistical output across the league was 10.15 points per game, 2.38 assists, and 3.8 rebounds. In the playoffs, these averages tended to decrease to 8.45 points, 1.81 assists, and 3.47 rebounds. This can be explained by the fact that most players, aside from the very best in the league, get less playing time when the stakes get higher. This same decrease is seen when analyzing these statistics grouped by position, further supporting this hypothesis.

When analyzing the averages across the top ten scorers at each position, further evidence of this point was found. That is, even strong scorers are scoring less on average come playoff time. This suggests that in addition to reduced playing time for those other than the very best, the playoffs in general contain lower-scoring games with higher defensive intensity, which is a well-known fact across the NBA.

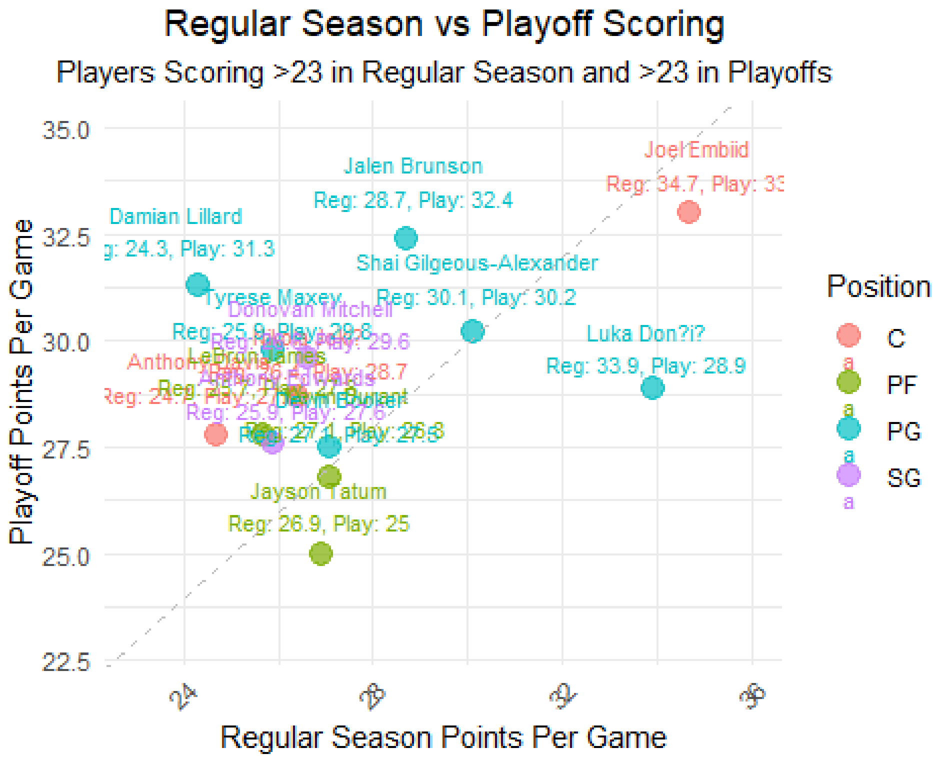
```
## [1] "Top 10 Regular Season Summary Statistics:"

## # A tibble: 1 × 9
##   Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
Median_TRB
##      <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
<dbl>
## 1     20.8       20.1   5.39     4.52       4.55   2.26     6.16
5.25
## # ℹ 1 more variable: SD_TRB <dbl>

## [1] "Top 10 Playoff Summary Statistics:"

## # A tibble: 1 × 9
##   Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
Median_TRB
##      <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
<dbl>
## 1     20.0       18.7   7.03     4.00          4   2.24     6.44
6.5
## # ℹ 1 more variable: SD_TRB <dbl>

## [1] "Top 10 Regular Season Summary Statistics by Position:"

## # A tibble: 5 × 10
##   Pos_reg Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
##   <chr>      <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
## 1 C           20.3       18.2   6.58     3.3        2.55   2.43     9.7
## 2 PF          22.0       22.2   3.88     4.52       4.4    1.58     6.94
## 3 PG          24.9       25.1   5.13     7.11       6.6    1.85     4.63
## 4 SF          17.8       16.0   3.63     3.1        2.9    1.53     5.1
## 5 SG          19.2       17.0   5.16     4.55       5.1    1.50     4.41
## # ℹ 2 more variables: Median_TRB <dbl>, SD_TRB <dbl>

## [1] "Top 10 Playoff Summary Statistics by Position:"

## # A tibble: 5 × 10
##   Pos_reg Mean_PTS Median_PTS SD_PTS Mean_AST Median_AST SD_AST Mean_TRB
##   <chr>      <dbl>      <dbl>  <dbl>    <dbl>      <dbl>  <dbl>    <dbl>
## 1 C           19.7       17     7.82     3.14       2.1    2.48     9.47
## 2 PF          21.0       20.6   5.89     4.28       3.9    2.15     7.82
## 3 PG          25.1       28.2   6.61     6.26       6.2    1.39     4.85
## 4 SF          15.4       14.7   6.01     2.43       2.15   1.45     5.96
## 5 SG          18.7       16.8   6.10     3.91       4.3    1.79     4.1
## # ℹ 2 more variables: Median_TRB <dbl>, SD_TRB <dbl>
```

Then, a summary of the best scorers in the league. Generally, this is a good indication for a player’s overall skill level and importance to their team.



This plot gives us valuable insight into who the most heavily depended-on, reliable scorers in the game are. We group by position for the sake of visual clarity and to see that the best players in the world come in many different positions and roles.

Next, we examine the effectiveness of these top scorers in the playoffs vs. the regular

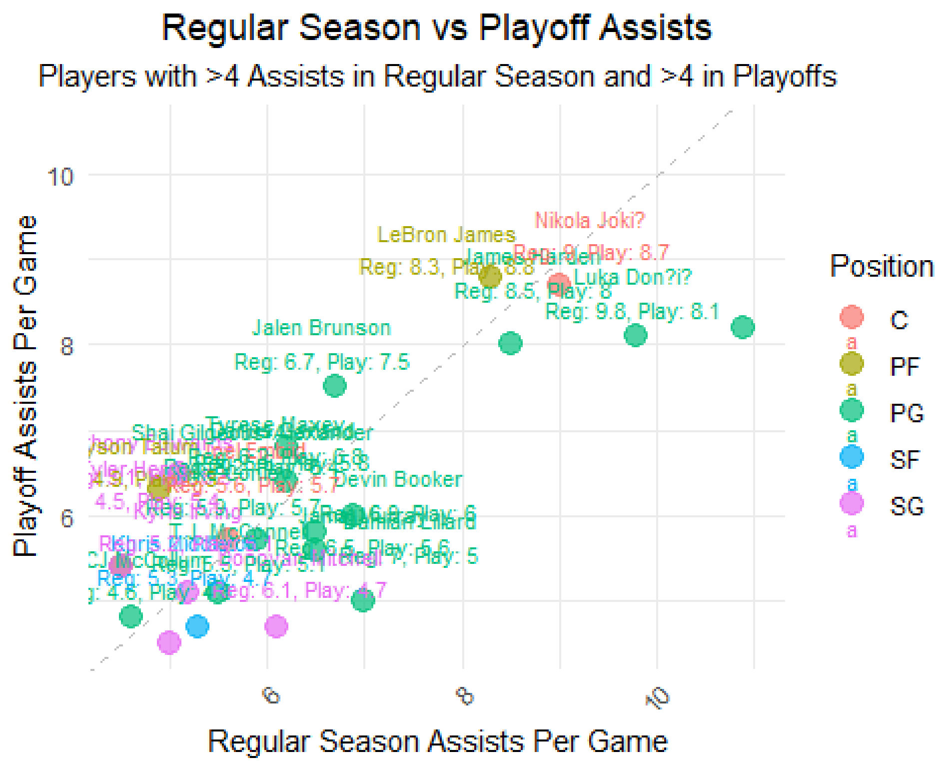


season.

This plot shows us the trend among elite scorers in the league, that they tend to step their games up during the playoffs. However, it would be naive to assume that players just “get worse” during the playoffs. Let’s take a look at assist numbers for these same players to see if we notice a trend in the direction of points vs assists.

```
## Warning: Removed 5 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```

```
## Warning: Removed 8 rows containing missing values or values outside the
scale range
## (`geom_text()`).
```



Based on this exploratory analysis, we can see that most elite offensive players tend to get more aggressive, raising their scoring and lowering their assists in the playoffs. However, some players lower their scoring and raise their assists, likely as a result of the additional defensive focus they receive in the playoffs.

Confirmatory Data Analysis

With some exploratory analysis completed, it is now time to conduct confirmatory analysis by fitting some regression models to help identify the factors that make a playoff performer stand out.

To start, we will fit some models to determine how regular season performance can predict playoff performance.

```
##
## Call:
## lm(formula = PTS_play ~ PTS_reg, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7643  -1.6608   0.1308   2.1145  10.9366
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.0406     0.3878  -5.262 3.58e-07 ***
## PTS_reg       1.0466     0.0310  33.757  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.264 on 205 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8468
## F-statistic:  1140 on 1 and 205 DF,  p-value: < 2.2e-16
```

This model is basic, and only the first attempt I made at creating a predictive analysis, but we can see there is already a moderately strong correlation with an adjusted R-Squared value of 0.8468. The coefficient estimates of b0=-2.0406 and b1=1.0466 are statistically significant down to the 0.001 level, indicating a strong confidence in these estimates.

To expand our analysis, it makes sense to select a few more variables that will be strong indicators of playoff scoring ability. For this, stepwise refinement will be used to select a few more variables for predictors.

```
## Warning in grep("_reg$", names(combined_data)) & !grepl("_play$",
## names(combined_data)): longer object length is not a multiple of shorter
object
## length
```

```
## Warning in grep("_reg$", names(combined_data)) & !grepl("_play$",
## names(combined_data)) & : longer object length is not a multiple of
shorter
## object length
```

```
## Start:  AIC=879.04
## PTS_play ~ 1
##
##              Df Sum of Sq     RSS    AIC
## + PTS_reg     1   12139.3  2183.8 491.71
## + FG_reg      1   12030.8  2292.3 501.75
## + FGA_reg     1   11787.1  2536.0 522.67
## + X2PA_reg    1   11011.7  3311.4 577.89
## + X2P_reg     1   10706.4  3616.8 596.15
## + FT_reg      1   10612.7  3710.4 601.44
## + TOV_reg     1   10547.5  3775.6 605.05
## + FTA_reg     1   10386.3  3936.8 613.70
## + MP_reg      1    9220.3  5102.8 667.40
## + GS_reg      1    8457.1  5866.0 696.25
## + AST_reg     1    8330.4  5992.7 700.67
## + DRB_reg     1    6995.9  7327.2 742.29
## + STL_reg     1    6180.7  8142.4 764.13
## + TRB_reg     1    5536.7  8786.4 779.89
## + X3PA_reg    1    5189.3  9133.8 787.91
## + X3P_reg     1    4983.8  9339.3 792.52
## + PF_reg      1    4616.8  9706.3 800.50
## + G_reg       1    3113.1 11210.0 830.31
## + BLK_reg     1    2500.3 11822.8 841.33
## + FT._reg     1    1574.2 12748.9 856.94
```

```
## + ORB_reg   1    1014.3 13308.8 865.84
## + X3P._reg  1     484.4 13838.7 873.92
## + FG._reg   1     250.4 14072.7 877.39
## + eFG._reg  1     155.2 14167.9 878.79
## <none>                  14323.1 879.04
## + X2P._reg  1      23.9 14299.2 880.70
##
## Step:  AIC=491.71
## PTS_play ~ PTS_reg
##
##            Df Sum of Sq    RSS    AIC
## + GS_reg    1   112.047 2071.8 482.81
## + TOV_reg   1    54.192 2129.6 488.51
## + X3P_reg   1    54.109 2129.7 488.52
## + X3PA_reg  1    45.933 2137.9 489.31
## + X2P_reg   1    45.617 2138.2 489.34
## + AST_reg   1    43.109 2140.7 489.59
## + X2PA_reg  1    41.861 2141.9 489.71
## + FT_reg    1    32.167 2151.6 490.64
## + FTA_reg   1    25.619 2158.2 491.27
## <none>                  2183.8 491.71
## + eFG._reg  1    15.563 2168.2 492.23
## + ORB_reg   1     9.976 2173.8 492.77
## + TRB_reg   1     9.805 2174.0 492.78
## + X3P._reg  1     9.417 2174.4 492.82
## + DRB_reg   1     7.646 2176.2 492.99
## + BLK_reg   1     7.324 2176.5 493.02
## + FGA_reg   1     7.291 2176.5 493.02
## + FG_reg    1     5.517 2178.3 493.19
## + STL_reg   1     3.679 2180.1 493.36
## + X2P._reg  1     2.874 2180.9 493.44
## + FT._reg   1     2.397 2181.4 493.49
## + PF_reg    1     1.677 2182.1 493.55
## + MP_reg    1     0.473 2183.3 493.67
## + G_reg     1     0.407 2183.4 493.68
## + FG._reg   1     0.119 2183.7 493.70
##
## Step:  AIC=482.81
## PTS_play ~ PTS_reg + GS_reg
##
##            Df Sum of Sq    RSS    AIC
## + FT_reg    1    62.058 2009.7 478.52
## + X3P_reg   1    60.331 2011.4 478.69
## + X3PA_reg  1    50.489 2021.3 479.70
## + X2PA_reg  1    48.750 2023.0 479.88
## + X2P_reg   1    41.953 2029.8 480.58
## + FTA_reg   1    41.036 2030.7 480.67
## + MP_reg    1    40.964 2030.8 480.68
## + TOV_reg   1    35.396 2036.4 481.24
## + AST_reg   1    32.090 2039.7 481.58
```

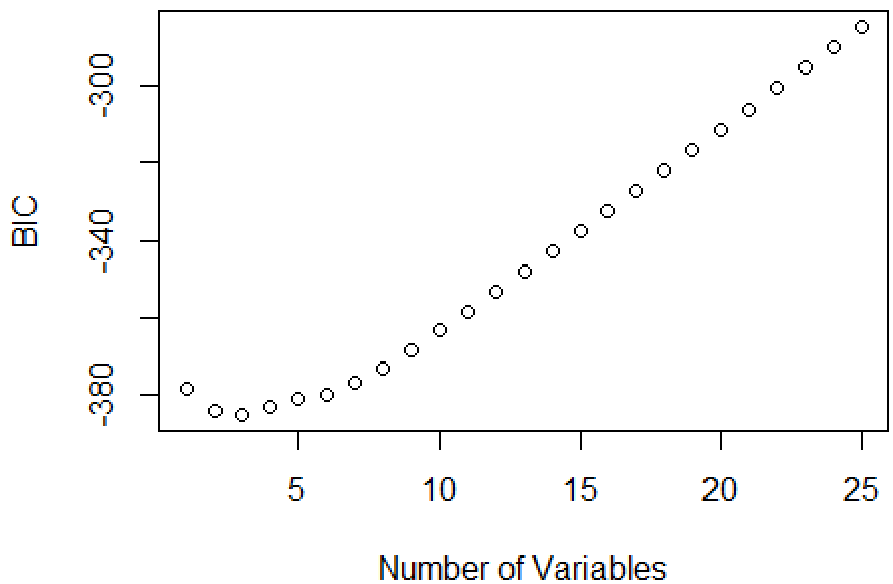
```
## + eFG._reg  1    28.747 2043.0 481.92
## <none>                  2071.8 482.81
## + G_reg     1    12.434 2059.3 483.56
## + PF_reg    1    11.915 2059.8 483.62
## + FGA_reg   1     6.397 2065.4 484.17
## + X3P._reg  1     6.019 2065.7 484.21
## + FG._reg   1     2.149 2069.6 484.60
## + FT._reg   1     1.508 2070.2 484.66
## + BLK_reg   1     0.586 2071.2 484.75
## + FG_reg    1     0.384 2071.4 484.77
## + ORB_reg   1     0.177 2071.6 484.79
## + X2P._reg  1     0.157 2071.6 484.80
## + STL_reg   1     0.098 2071.7 484.80
## + TRB_reg   1     0.038 2071.7 484.81
## + DRB_reg   1     0.000 2071.8 484.81
##
## Step:  AIC=478.52
## PTS_play ~ PTS_reg + GS_reg + FT_reg
##
##             Df Sum of Sq    RSS    AIC
## + AST_reg   1    34.130 1975.6 476.97
## + X2PA_reg  1    24.263 1985.4 478.00
## + TOV_reg   1    20.668 1989.0 478.38
## + X2P_reg   1    19.563 1990.1 478.49
## <none>                  2009.7 478.52
## + FG_reg    1    18.990 1990.7 478.55
## + eFG._reg  1    18.175 1991.5 478.64
## + X3P_reg   1    16.374 1993.3 478.82
## + FTA_reg   1    14.352 1995.3 479.03
## + MP_reg    1    13.937 1995.8 479.08
## + X3PA_reg  1    12.851 1996.8 479.19
## + PF_reg    1    10.652 1999.0 479.42
## + BLK_reg   1     7.014 2002.7 479.79
## + FGA_reg   1     5.985 2003.7 479.90
## + FG._reg   1     5.580 2004.1 479.94
## + TRB_reg   1     2.247 2007.5 480.28
## + G_reg     1     2.230 2007.5 480.29
## + ORB_reg   1     2.226 2007.5 480.29
## + DRB_reg   1     2.129 2007.6 480.30
## + STL_reg   1     0.316 2009.4 480.48
## + X2P._reg  1     0.163 2009.5 480.50
## + FT._reg   1     0.163 2009.5 480.50
## + X3P._reg  1     0.079 2009.6 480.51
##
## Step:  AIC=476.97
## PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg
##
##             Df Sum of Sq    RSS    AIC
## + X2P_reg   1   26.0410 1949.5 476.22
## + X2PA_reg  1   25.6449 1949.9 476.27
```

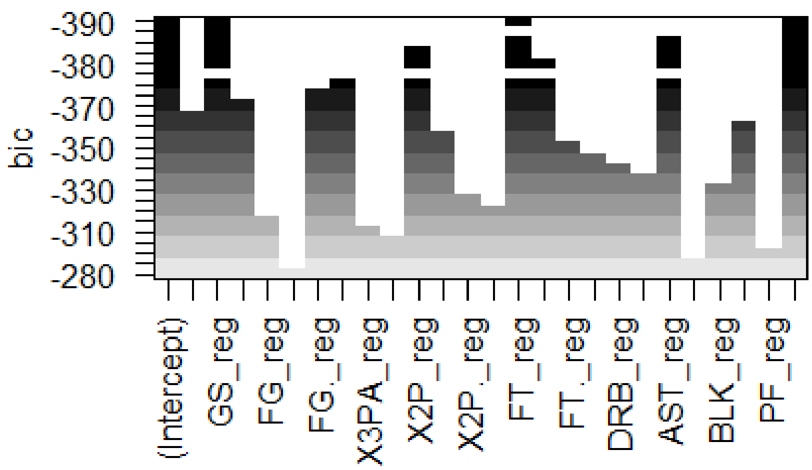
```
## + FG_reg     1   25.4194 1950.2 476.29
## + X3P_reg    1   22.1580 1953.4 476.64
## + MP_reg     1   21.6442 1953.9 476.69
## + X3PA_reg   1   19.3317 1956.2 476.93
## <none>                   1975.6 476.97
## + FTA_reg    1   10.8372 1964.7 477.83
## + eFG._reg   1   10.3328 1965.2 477.88
## + PF_reg     1    4.1599 1971.4 478.53
## + G_reg      1    2.3063 1973.3 478.73
## + TOV_reg    1    2.2653 1973.3 478.73
## + FG._reg    1    1.8029 1973.8 478.78
## + STL_reg    1    1.5765 1974.0 478.80
## + FGA_reg    1    0.8874 1974.7 478.88
## + X2P._reg   1    0.6050 1975.0 478.91
## + BLK_reg    1    0.5103 1975.1 478.92
## + FT._reg    1    0.4788 1975.1 478.92
## + X3P._reg   1    0.4460 1975.1 478.92
## + DRB_reg    1    0.3143 1975.3 478.94
## + TRB_reg    1    0.1252 1975.5 478.96
## + ORB_reg    1    0.0062 1975.6 478.97
##
## Step:  AIC=476.22
## PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg + X2P_reg
##
##            Df Sum of Sq    RSS    AIC
## + FTA_reg    1    40.150 1909.4 473.92
## + X3P_reg    1    29.969 1919.6 475.02
## + FG._reg    1    22.293 1927.2 475.84
## <none>                   1949.5 476.22
## + MP_reg     1    16.653 1932.9 476.45
## + ORB_reg    1    15.657 1933.9 476.55
## + eFG._reg   1    15.092 1934.4 476.61
## + TRB_reg    1    11.816 1937.7 476.96
## + DRB_reg    1     8.336 1941.2 477.34
## + PF_reg     1     7.623 1941.9 477.41
## + BLK_reg    1     5.835 1943.7 477.60
## + X3PA_reg   1     5.257 1944.3 477.66
## + FGA_reg    1     3.543 1946.0 477.85
## + STL_reg    1     1.827 1947.7 478.03
## + G_reg      1     1.629 1947.9 478.05
## + X3P._reg   1     0.779 1948.8 478.14
## + X2P._reg   1     0.583 1949.0 478.16
## + X2PA_reg   1     0.576 1949.0 478.16
## + FT._reg    1     0.355 1949.2 478.19
## + FG_reg     1     0.038 1949.5 478.22
## + TOV_reg    1     0.033 1949.5 478.22
##
## Step:  AIC=473.92
## PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg + X2P_reg + FTA_reg
##
```

```
##             Df Sum of Sq    RSS    AIC
## + X3P_reg   1   20.9405 1888.4 473.63
## <none>                  1909.4 473.92
## + FG._reg   1   16.4369 1892.9 474.13
## + eFG._reg  1   11.9259 1897.5 474.62
## + MP_reg    1    7.6918 1901.7 475.08
## + X3PA_reg  1    6.2385 1903.1 475.24
## + FGA_reg   1    3.3304 1906.0 475.55
## + PF_reg    1    3.0748 1906.3 475.58
## + TOV_reg   1    2.9563 1906.4 475.59
## + BLK_reg   1    2.6499 1906.7 475.63
## + ORB_reg   1    2.4168 1907.0 475.65
## + FT._reg   1    1.7371 1907.6 475.73
## + TRB_reg   1    1.1785 1908.2 475.79
## + STL_reg   1    0.9677 1908.4 475.81
## + X2P._reg  1    0.7039 1908.7 475.84
## + DRB_reg   1    0.6659 1908.7 475.84
## + FG_reg    1    0.4464 1908.9 475.87
## + X3P._reg  1    0.4380 1908.9 475.87
## + X2PA_reg  1    0.2899 1909.1 475.88
## + G_reg     1    0.1757 1909.2 475.90
##
## Step:  AIC=473.63
## PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg + X2P_reg + FTA_reg +
##     X3P_reg
##
##             Df Sum of Sq    RSS    AIC
## <none>                  1888.4 473.63
## + FG._reg   1   14.1809 1874.3 474.07
## + eFG._reg  1   10.8965 1877.5 474.44
## + MP_reg    1    8.2544 1880.2 474.73
## + ORB_reg   1    3.2165 1885.2 475.28
## + TOV_reg   1    3.1229 1885.3 475.29
## + BLK_reg   1    2.8329 1885.6 475.32
## + X3PA_reg  1    2.5900 1885.8 475.35
## + FGA_reg   1    2.2275 1886.2 475.39
## + PF_reg    1    2.0672 1886.4 475.41
## + TRB_reg   1    1.7600 1886.7 475.44
## + FT._reg   1    1.7043 1886.7 475.45
## + STL_reg   1    1.0545 1887.4 475.52
## + DRB_reg   1    1.0422 1887.4 475.52
## + X2PA_reg  1    0.6365 1887.8 475.56
## + X2P._reg  1    0.6182 1887.8 475.57
## + FG_reg    1    0.5532 1887.9 475.57
## + G_reg     1    0.2299 1888.2 475.61
## + X3P._reg  1    0.0958 1888.3 475.62

##
## Call:
## lm(formula = PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg +
```

```
##     X2P_reg + FTA_reg + X3P_reg, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8214  -1.6076   0.0083   1.6558  10.0755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.23890    0.43763  -2.831 0.005118 **
## PTS_reg     -2.25389    1.86122  -1.211 0.227340
## GS_reg       0.04589    0.01189   3.858 0.000154 ***
## FT_reg       5.66276    2.09668   2.701 0.007513 **
## AST_reg      0.36020    0.17636   2.042 0.042430 *
## X2P_reg      6.17157    3.69565   1.670 0.096500 .
## FTA_reg     -1.88132    1.03886  -1.811 0.071658 .
## X3P_reg      8.29810    5.58611   1.485 0.138997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.081 on 199 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8635
## F-statistic: 187.2 on 7 and 199 DF,  p-value: < 2.2e-16

## (Intercept)     PTS_reg      GS_reg      FT_reg     AST_reg     X2P_reg
## -1.23889632 -2.25389011  0.04588736  5.66275533  0.36020254  6.17156517
##     FTA_reg     X3P_reg
## -1.88132438  8.29809531
```

Based on AIC stepwise refinement, the variables that are best to use are PTS, GS, FT, AST, 2P, FTA, and 3P. These are all trademarks of a high-usage, high-volume player, what we may refer to as the “alpha dogs” of the sports world. These variables create one model that we can compare to the other ones we come up with. When these variables are used to predict playoff points per game for a player, the adjusted R^2 value is 0.8635, suggesting a slightly stronger linear correlation. We can also select variables using the best subsets, measuring by BIC and R^2:

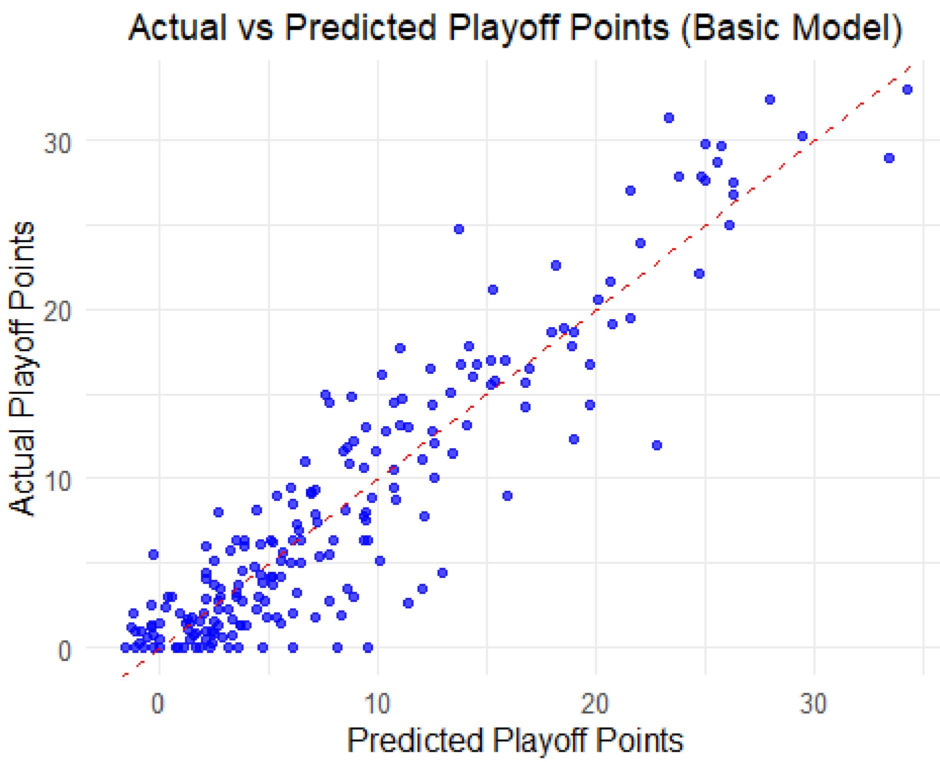


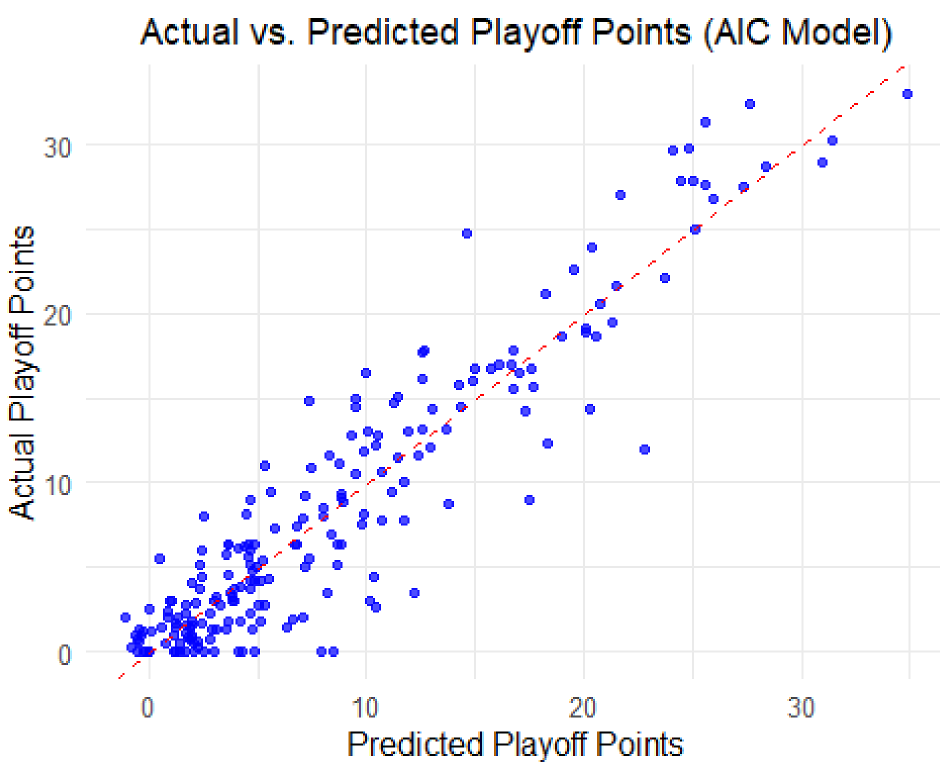


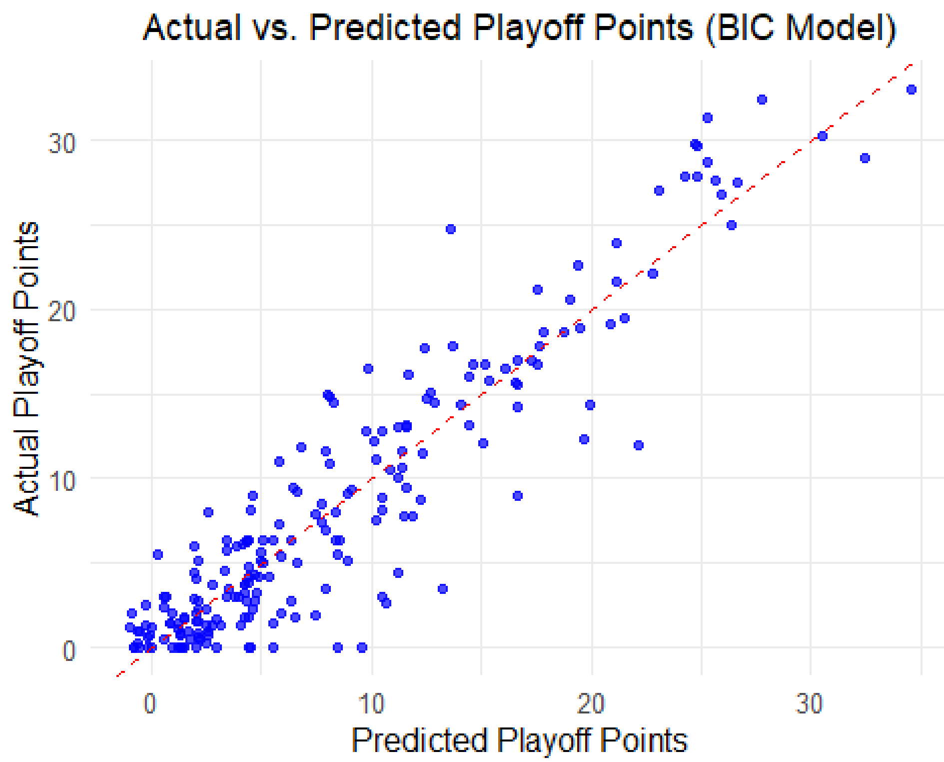

```
## (Intercept)      GS_reg      FT_reg     PTS_reg
## -1.47842138  0.04472006  0.86431496  0.73365605
```

In contrast with the AIC stepwise refinement, the BIC subset selection revealed that the most effective model contains only three variables: PTS, FT, and GS. This was somewhat expected, as BIC imposes larger penalties for the number of parameters, although it was surprising to me that FT were a more important parameter than 2P or 3P.

We can now compare these two models to our initial model, and see which factors are truly the best predictors of how a player will score in the playoffs.


Actual vs Predicted Playoff Points (Basic Model)
Actual Playoff Points
Predicted Playoff Points
0
10
20
30
0
10
20
30


Actual vs. Predicted Playoff Points (AIC Model)
Actual Playoff Points
Predicted Playoff Points
0
10
20
30
0
10
20
30




```
##
## Call:
## lm(formula = PTS_play ~ PTS_reg, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7643  -1.6608   0.1308   2.1145  10.9366
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.0406     0.3878  -5.262 3.58e-07 ***
## PTS_reg       1.0466     0.0310  33.757  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.264 on 205 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8468
## F-statistic:  1140 on 1 and 205 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = PTS_play ~ PTS_reg + GS_reg + FT_reg + AST_reg +
##     X2P_reg + FTA_reg + X3P_reg, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8214  -1.6076   0.0083   1.6558  10.0755
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.23890    0.43763  -2.831 0.005118 **
## PTS_reg     -2.25389    1.86122  -1.211 0.227340
## GS_reg       0.04589    0.01189   3.858 0.000154 ***
## FT_reg       5.66276    2.09668   2.701 0.007513 **
## AST_reg      0.36020    0.17636   2.042 0.042430 *
## X2P_reg      6.17157    3.69565   1.670 0.096500 .
## FTA_reg     -1.88132    1.03886  -1.811 0.071658 .
## X3P_reg      8.29810    5.58611   1.485 0.138997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.081 on 199 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8635
## F-statistic: 187.2 on 7 and 199 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = PTS_play ~ PTS_reg + FT_reg + GS_reg, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1482  -1.4883   0.0931   1.8262  11.0848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.47842    0.41947  -3.524 0.000524 ***
## PTS_reg      0.73366    0.08940   8.206  2.6e-14 ***
## FT_reg       0.86431    0.34522   2.504 0.013078 *
## GS_reg       0.04472    0.01181   3.786 0.000201 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.146 on 203 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8576
## F-statistic: 414.6 on 3 and 203 DF,  p-value: < 2.2e-16
```

Based on the three models we are comparing, there are a couple of conclusions that can be drawn.

First of all, it is clear that the model from AIC stepwise refinement gives us the best correlation, and the lowest residual standard error. This tells us that we got the most predictive model when analyzing the variables: PTS, GS, FT, AST, 2P, FTA, 3P, to predict a player's playoff scoring.

When analyzing the graphs of predicted vs. actual playoff points, we can see that the models are remarkably similar, usually only missing by a couple of points at worst. In fact, 50% of the predictions fall within 1.6 points of the actual playoff scoring values when using the AIC model.

To give a final answer to our question, we will compare who the three models predict to be the top playoff scorers in the league, versus their actual numbers.

Conclusion

Player	Actual	Actual_Rank	Basic_Predicted	Basic_Rank	AIC_Predicted	AIC_Rank	BIC_Predicted	BIC_Rank
Joel Embiid	33.0	1	34.27718	1	34.86190	1	34.53954	1
Jalen Brunson	32.4	2	27.99745	4	27.62928	5	27.77468	4
Damian Lillard	31.3	3	23.39232	15	25.55058	9	25.23203	10
Shai Gilgeous-Alexander	30.2	4	29.46272	3	31.44632	2	30.52742	3
Tyrese Maxey	29.8	5	25.06691	10	24.82735	12	24.71595	13
Donovan Mitchell	29.6	6	25.79955	8	24.10955	14	24.81801	11
Luka Don?i?	28.9	7	33.43988	2	30.92732	3	32.40026	2
Nikola Joki?	28.7	8	25.59022	9	28.40106	4	25.31240	9
Anthony Davis	27.8	9	23.81096	14	25.05364	11	24.79534	12
LeBron James	27.8	9	24.85759	12	24.42285	13	24.26822	14
Anthony Edwards	27.6	11	25.06691	10	25.60491	8	25.67874	8
Devin Booker	27.5	12	26.32286	5	27.34916	6	26.63051	5
Paolo	27.0	13	21.6130	18	21.7123	17	23.0878	15

Player	Actual	Actual_Rank	Basic_Predicted	Basic_Rank	AIC_Predicted	AIC_Rank	BIC_Predicted	BIC_Rank
Banchero			6		7		2	
Kevin Durant	26.8	14	26.32286	5	25.96307	7	25.90637	7
Jayson Tatum	25.0	15	26.11353	7	25.15216	10	26.40637	6
Khris Middleton	24.7	16	13.76340	45	14.62447	42	13.61516	47
Jaylen Brown	23.9	17	22.03171	17	20.33885	22	21.11902	20
Bam Adebayo	22.6	18	18.15921	29	19.50587	26	19.39996	25
Kyrie Irving	22.1	19	24.75292	13	23.75413	15	22.74918	16
Pascal Siakam	21.6	20	20.67110	21	21.46549	18	21.13105	19

From observing this table, we can draw a few conclusions about the models predictions of the data.

First, the basic model does have some predictive power, but evidently the overfitting based on regular season points is not a good predictor. Rather, it serves as a better indicator for those who over or underperform compared to their expectations.

Second, the AIC model appears to overrate the scoring of those who take and make alot of free throws. This is not entirely unsurprising, as making free throws obviously directly leads to points, and attempting more free throws usually signals a more aggressive player. However, in the regular season, players attempt more free throws on average due to a softer whistle, and they make more due to less rambunctious crowds. Therefore, this is an imperfect predictor, even though there is usually a small error in prediction vs. actual scoring.

Third, the BIC model underrates players that missed a significant amount of regular season games due to injury. This factor may have been improved if a metric such as GS/GP has been used, measuring the percentage of games that a player participated in, in which they appeared in the starting lineup (signifying a prominent role in the team that night). However, other than an interesting outlier in Khris Middleton (missed majority of the season and struggled to get healthy in the games he did play) the BIC model is similar in accuracy to the AIC model.

Fourth, interestingly, the models were unanimous in picking the leading scorer of the playoffs, Joel Embiid. He was far and away the regular season's leading scorer, so this is not terribly surprising, and he did actually lead the playoffs in scoring as well. However, this goes to show the metric I used to determine playoff performance with may have been imperfect on its own, as Embiid did perform well but lost in six games to the New York Knicks in the first round, a disappointing result for such a promising player.

In conclusion, while the analysis was imperfect from a predictor and response variable standpoint, it does seem possible to predict how a player will perform in the playoffs based on their regular season performance, with a moderate degree of accuracy. In particular, a player's playoff scoring can be reasonably estimated based on how well they perform in the following statistical regular season categories: Points, Free Throws Made, and Games Started, and to a lesser degree, Assists, Free Throws Attempted, and 2 and 3 pointers made. In future research, I will refine the metrics by which I measure 'performance' to expand past regular box score statistics, and attempt to isolate more factors that predict how well a player will perform in the playoffs. Based on the research conducted in this project, it would appear that: 1) a significant amount of elite players tend to increase their scoring output in the playoffs, usually by way of increased aggressiveness. As a result, 2) the average player actually decreases in their performance, based on opportunity. 3) While there are certain factors that define a playoff performer in terms of scoring, they tend to display these traits during the regular season as well, and so there does not appear to be significant evidence of playoff risers and fallers as described in the introduction.