Dhanvin Lakshmisha

Data Science Project Semester 2

# NBA Players Minutes Played, Field Goals Made, and Free Throws Made (blocked by position) Analysis for Total Points in a Season

**Introduction:**

https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats

In the NBA, points (pts) are possibly one of the most important metrics in determining a player's skill relative to others. In predicting a player's points in a season, three important metrics include field goal, minutes played, and the number of free throws made. Field goal (fg) is how many shots a player makes. Minutes played (mp) is how many minutes a player plays over the season. Finally, free throws made (ft) is how many free throws a player makes in a season.

Also, the position of a player is their role on their team. Positions include point guard, shooting guard, small forward, power forward, and center. Because these positions are flexible, I will group point guards and shooting guards (guards), small forward and power forwards (forwards), and centers (centers). Each group has different characteristics such as weight and height. Removing such confounding factors in the experiment allows us to visualize which predictor is the best for each position.

That is, which model would bring me the closest, in terms of prediction, to the response variable, points in a season? By doing this, I want to conclude which predictors can give us the best model for predicting points in a season. I find this topic interesting because when I was growing up, I had an aspiration of becoming a professional basketball player. However, after barely getting any playing time, my NBA career aspiration faded away. Therefore, I find it interesting to see what predictors can best explain points in a season or are the best fit in extrapolating what the points in a season of a player will be.

**Results:**

The dataset I chose includes NBA players from the 2022-23 NBA Season. I filtered out the dataset to only include NBA players from this year. As I mentioned before, guards, forwards, and centers are grouped.

Choose:

First, we could create a multiple linear regression model that entails all three predictors (free throws, minutes played, and field goals) and the response variable (total points in the season). Also, there are indicator variables that allow for the grouping or blocking. A 1 for the indicator variable takes into account that group. All these are quantitative predictors and response variables. We are trying to determine the best combination or if the blocking is necessary.

Thus, the model would be: $Y = \beta 0 + \beta 1(\text{fg}) + \beta 2(\text{ft}) + \beta 3(\text{mp}) + \beta 4 \cdot \text{guards} + \beta 5 \cdot \text{forwards} + \beta 4 \cdot \text{centers} + \epsilon$

Fit:

The model in summary format would appear like this:

```
Call:
lm(formula = pts ~ fg + ft + mp + guards + forwards + centers,
    data = filtered_Player_Totals)

Residuals:
     Min       1Q   Median       3Q      Max
-132.631  -13.685   -1.443   17.358  135.719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.612170  32.515600  -1.495   0.1354
fg            2.194723   0.028360  77.389  < 2e-16 ***
ft            0.852874   0.029206  29.202  < 2e-16 ***
mp            0.030760   0.004243   7.250 1.15e-12 ***
guards       58.850640  32.383139   1.817   0.0696 .
forwards     47.661846  32.371814   1.472   0.1414
centers      17.840769  32.483435   0.549   0.5830
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.03 on 671 degrees of freedom
Multiple R-squared:  0.9954,    Adjusted R-squared:  0.9954
F-statistic: 2.434e+04 on 6 and 671 DF,  p-value: < 2.2e-16
```

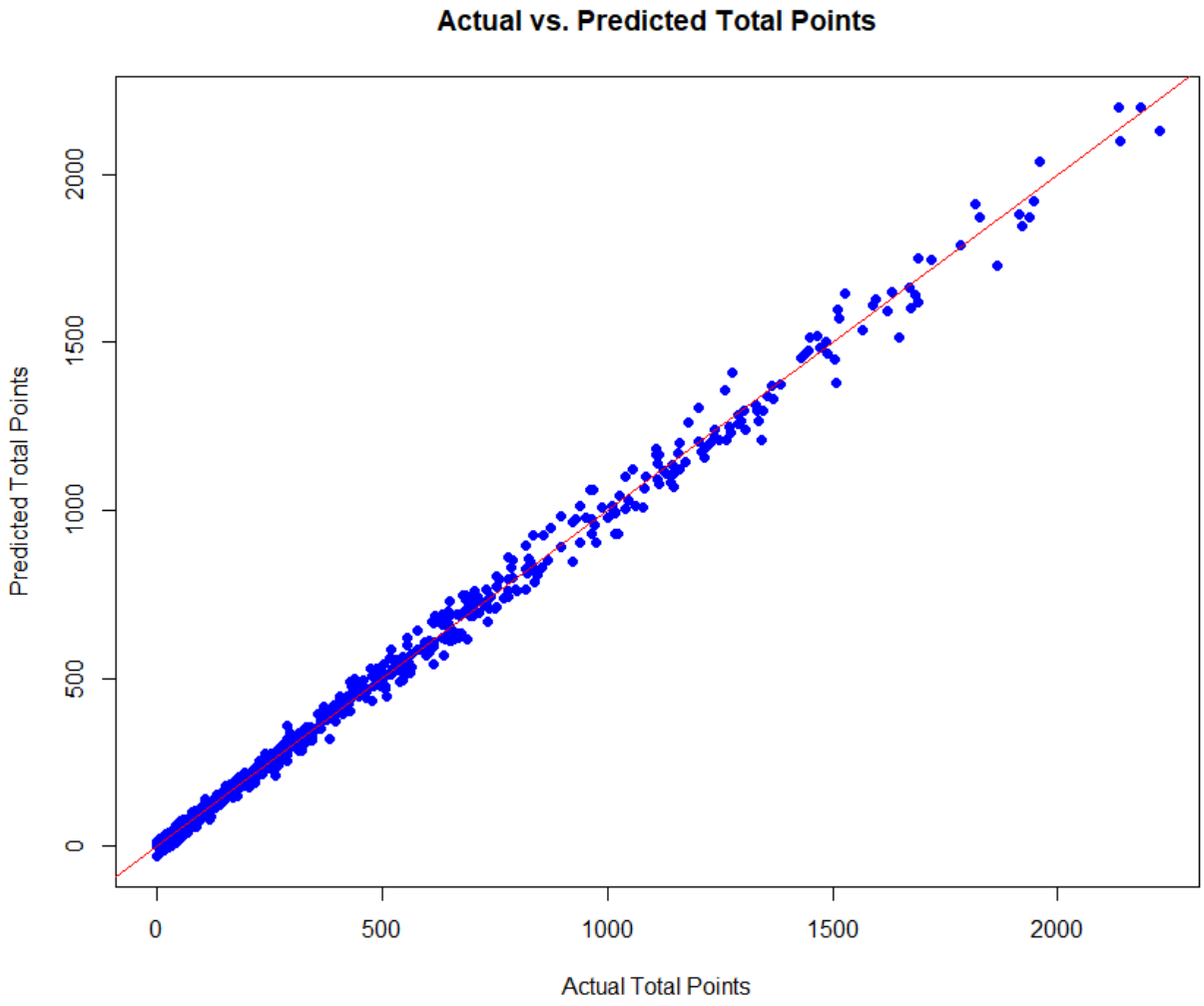
The approximate fitted model would be $Y$ = -42.61 + 2.19(fg) + .852(ft) + .031 (mp) + (guards) 58.8506 + (forwards)47.661+ centers (17.840769).

Assess:

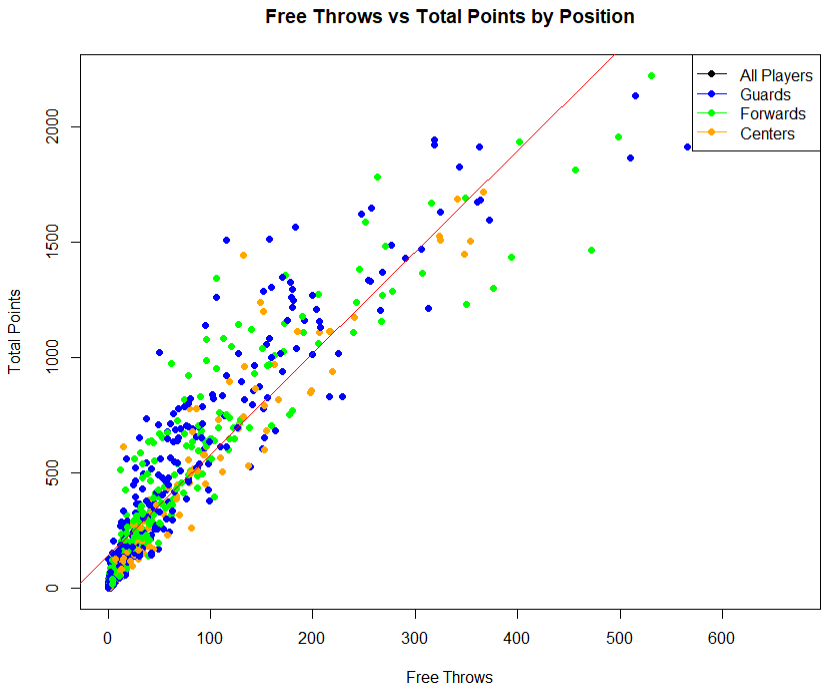
Based on the summary data, the model has a high adjusted R-squared value of .9954. This suggests that the model explains a large portion of the variability in the response variable (total points) and that the predictor variables are adding value to the model. Specifically, 99.54% of the variability in the response variable is accounted for by the explanatory in the model. Also, all the predictors appear to be good except the forwards and centers indicators, which have a p-value greater than and not .05. Thus, they are not significant in predicting total points. This means that

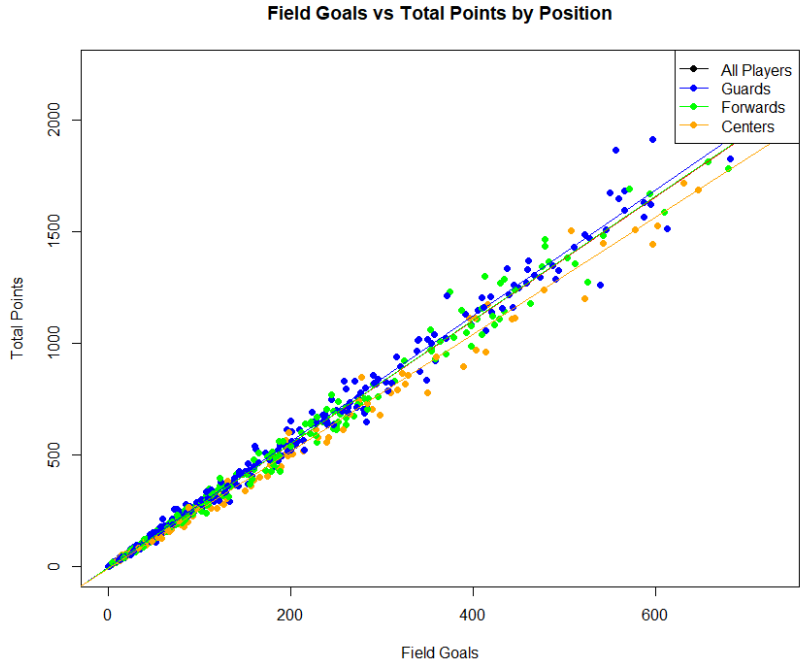
there is a low chance there are differences in the total points scored between forwards and centers, compared to other positions, after considering their field goals, free throws, and minutes played.

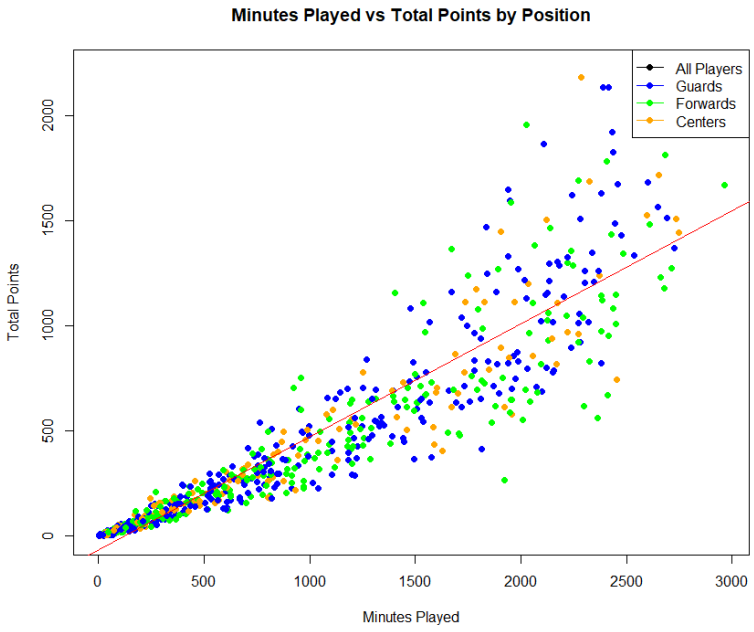
Looking at the residuals plot below, it is linear, meaning that a linear model is appropriate for the given data.



We can also look at the individual scatter plots for each predictor variable.






The Field Goals vs Total Points model appears to be the best fit for predicting the response variable of Total Points compared to Free Throws vs Total Points and Minutes Played vs Total Points. Also, the Field Goals vs Total Points appears similar to the multiple linear regression model. However, when we look at the Field Goals vs Total Points summary data, pictured below, the adjusted R-square value is slightly lower than that of the multiple linear regression model. Also, it appears that guards, forwards, and centers are not significant in predicting total points. This suggests that all these groups follow a similar pattern in predicting total points and are not different. Since we hope to account for the groups to minimize the impact of confounding variables, the elimination of such indicator variables is not helpful. In that sense, the field goal model is more general as it can predict points without the information of position.

```
Call:
lm(formula = pts ~ fg + guards + forwards + centers, data = filtered_Player_Totals)

Residuals:
    Min      1Q  Median      3Q     Max
-238.11  -15.97   -1.04   15.28  320.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.90101   48.37426  -0.391    0.696
fg            2.77055    0.01097 252.492   <2e-16 ***
guards       23.68702   48.37277   0.490    0.625
forwards     13.83469   48.39268   0.286    0.775
centers     -14.84451   48.48262  -0.306    0.760
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.28 on 673 degrees of freedom
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9895
F-statistic: 1.597e+04 on 4 and 673 DF,  p-value: < 2.2e-16
```

Nevertheless, both models, Field Goals vs Total Points and the multiple linear regression model are great in predicting total points. To account for other information, however, a multiple linear regression model may be better due to its nature to account for other predictors (minutes played and free throws made). Also, it includes more indicator variables after their elimination. Thus, the general, multiple regression model is more informative and less general than the other models in predicting points.

Use:

In the first case, the investigation was conducted to see which model (blocked by position), given data of total field goals, total minutes played, and total free throws in an NBA season for all players, would best predict the response variable, total points in an NBA season, given its data. After analyzing different models and the validity of those models, we can conclude that the multiple linear regression model accompanying all 3 predictors best explains the variation in the

response variable (Total Points). This is due to how informative it is.  Thus, it is the best model relative to the other models in this investigation. This multiple linear regression model may not be extrapolated to other NBA seasons and retain accuracy in predicting the response variable due to the NBA population changing every year. The multiple linear regression model best fits and applies for the 2022-2023 NBA season.

**Conclusion:**

From this project, I learned that there is a strong correlation between field goals and total points, but not very a strong correlation between minutes played vs total points and free throws vs total points. However, when we combine the predictors and indicator variables to make a multiple linear regression model, the multiple linear regression model is more informative and stronger in terms of measurements. This means that a blocked (by position) multiple regression model is the best for predicting the total points an NBA player in the 2022-2023 NBA season may have.