Victor Oche

NBA Points Per Game Prediction (2023) | R

Introduction

The dataset used for this analysis was the NBA 2022-2023 stats dataset. It contained variables of player stats in teams throughout the 2022-2023 regular season. The data set provided an opportunity to 1) Find factors beyond the obvious that affect the points per game and 2) Determine a relationship between player position and season statistics.

```
> head(nba)
  Rk                  Player Pos Age  Tm  G GS   MP  FG  FGA   FG. X3P X3PA  X3P.
1  1         Precious Achiuwa   C  23 TOR 55 12 20.7 3.6  7.3 0.485 0.5  2.0 0.269
2  2             Steven Adams   C  29 MEM 42 42 27.0 3.7  6.3 0.597 0.0  0.0 0.000
3  3              Bam Adebayo   C  25 MIA 75 75 34.6 8.0 14.9 0.540 0.0  0.2 0.083
4  4             Ochai Agbaji  SG  22 UTA 59 22 20.5 2.8  6.5 0.427 1.4  3.9 0.355
5  5             Santi Aldama  PF  22 MEM 77 20 21.8 3.2  6.8 0.470 1.2  3.5 0.353
6  6 Nickeil Alexander-Walker  SG  24 TOT 59  3 15.0 2.2  5.0 0.444 1.0  2.7 0.384
  X2P X2PA  X2P.  eFG.  FT FTA   FT. ORB DRB  TRB AST STL BLK TOV  PF  PTS
1 3.0  5.4 0.564 0.521 1.6 2.3 0.702 1.8 4.1  6.0 0.9 0.6 0.5 1.1 1.9  9.2
2 3.7  6.2 0.599 0.597 1.1 3.1 0.364 5.1 6.5 11.5 2.3 0.9 1.1 1.9 2.3  8.6
3 8.0 14.7 0.545 0.541 4.3 5.4 0.806 2.5 6.7  9.2 3.2 1.2 0.8 2.5 2.8 20.4
4 1.4  2.7 0.532 0.532 0.9 1.2 0.812 0.7 1.3  2.1 1.1 0.3 0.3 0.7 1.7  7.9
5 2.0  3.4 0.591 0.560 1.4 1.9 0.750 1.1 3.7  4.8 1.3 0.6 0.6 0.8 1.9  9.0
6 1.2  2.3 0.515 0.547 0.7 1.0 0.667 0.3 1.5  1.7 1.8 0.5 0.4 0.9 1.5  6.2
```

The variables are: Rank(RK), player name, Position(POS), Age, Team(Tm), Games played(G), Games Started(GS), Minutes played(MP), Field Goals(FG, which is a combination of 2point and 3Point field goals), Field goal attempts(FGA), Field Goal percentage(FG. Which is FG/FGA), 3points made per game(X3P), 3points attempts per game(X3PA), 3point field goal %(X3P. defined as 3points made/3points attempted), 2 points made per game(X2P), 2point attempts per game(X2PA), 2point field goals %(X2P. defined as 2points made/2points attempted), effective field goals(eFG. Measures field goal percentage adjusting for made 3-point field goals being 1.5 times more valuable than made 2-point field goals.), Free Throws Made (FT), Free Throw Attempts (FTA), Free Throw % (FT. which is FT/FTA), Offensive rebounds (ORB), Defensive rebounds (DRB), Total rebounds (TRB), Assists (AST), Steals (STL), Blocks (BLK), Turnovers (Tov), PF (personal Fouls), Points per game (PTS).

Highly Multicollinear NBA Stats DataSet

Upon starting the analysis, I encountered a significant challenge with the dataset - perfect multicollinearity and aliased coefficients. This led to issues in my analysis methods such as regression and regularized regressions, Principal Component Analysis, Factor Analysis (PFA). To resolve this problem, I dedicated long hours researching potential solutions. Unfortunately, most regularization techniques failed to address the multicollinearity on this dataset at the variables seemed to be ‘perfectly multicollinear/ aliased ’. To overcome this hurdle, I made the decision to split the dataset into two datasets based on my research of the data dictionaries available on the NBA website.

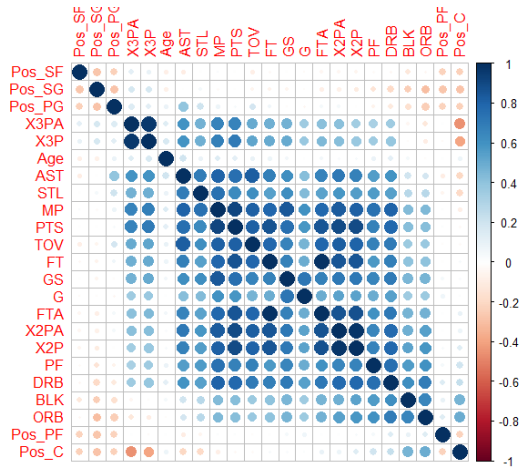
The first dataset, called "NBA Player Stats," included variables such as 3-pointers per game, number of assists, and blocks, 2points per game, etc., This represented an individual players contribution to affecting the game's outcome.

```
> #we'll divide the dataset into team and player stats
> nbaF <- nba[nba1$Tm != "TOT", ]
> nba_plyrstat <- nbaF[c(3:4, 6:8, 12:13, 15:16, 19:20, 22:23, 25:30)]
> head(nba_plyrstat)
  Pos Age  G GS   MP X3P X3PA X2P X2PA  FT FTA ORB DRB AST STL BLK TOV  PF  PTS
1   C  23 55 12 20.7 0.5  2.0 3.0  5.4 1.6 2.3 1.8 4.1 0.9 0.6 0.5 1.1 1.9  9.2
2   C  29 42 42 27.0 0.0  0.0 3.7  6.2 1.1 3.1 5.1 6.5 2.3 0.9 1.1 1.9 2.3  8.6
3   C  25 75 75 34.6 0.0  0.2 8.0 14.7 4.3 5.4 2.5 6.7 3.2 1.2 0.8 2.5 2.8 20.4
4  SG  22 59 22 20.5 1.4  3.9 1.4  2.7 0.9 1.2 0.7 1.3 1.1 0.3 0.3 0.7 1.7  7.9
5  PF  22 77 20 21.8 1.2  3.5 2.0  3.4 1.4 1.9 1.1 3.7 1.3 0.6 0.6 0.8 1.9  9.0
7  SG  24 36  3 14.7 1.0  2.4 1.3  2.3 0.8 1.1 0.2 1.4 2.1 0.7 0.4 1.3 1.6  6.3
>
```

The second dataset, named "Team Stats," contained variables like field goal percentage, 2-point percentage, which denoted a player's contribution to the overall team statistics per game.

```
> head(nba_tmstat)
  Pos Age  G GS   MP  FGA   FG. X3P X3PA  X3P. X2P X2PA  X2P.  FT FTA   FT. ORB DRB
1   C  23 55 12 20.7  7.3 0.485 0.5  2.0 0.269 3.0  5.4 0.564 1.6 2.3 0.702 1.8 4.1
2   C  29 42 42 27.0  6.3 0.597 0.0  0.0 0.000 3.7  6.2 0.599 1.1 3.1 0.364 5.1 6.5
3   C  25 75 75 34.6 14.9 0.540 0.0  0.2 0.083 8.0 14.7 0.545 4.3 5.4 0.806 2.5 6.7
4  SG  22 59 22 20.5  6.5 0.427 1.4  3.9 0.355 1.4  2.7 0.532 0.9 1.2 0.812 0.7 1.3
5  PF  22 77 20 21.8  6.8 0.470 1.2  3.5 0.353 2.0  3.4 0.591 1.4 1.9 0.750 1.1 3.7
7  SG  24 36  3 14.7  4.7 0.488 1.0  2.4 0.402 1.3  2.3 0.578 0.8 1.1 0.692 0.2 1.4
  AST STL BLK TOV  PF  PTS
1 0.9 0.6 0.5 1.1 1.9  9.2
2 2.3 0.9 1.1 1.9 2.3  8.6
3 3.2 1.2 0.8 2.5 2.8 20.4
4 1.1 0.3 0.3 0.7 1.7  7.9
5 1.3 0.6 0.6 0.8 1.9  9.0
7 2.1 0.7 0.4 1.3 1.6  6.3
```

Some variables like position existed in both datasets. Additionally, I carefully reviewed the correlation plots of the player stats variables, removing one of any two variables that were aliases or showed nearly identical correlations, while retaining those that were less obvious in predicting points per game.



Variables like X3PA, X3P, which showed nearly exact correlation was taken out. The attempts. Block and offensive rebounds(Blk was left) etc.

Analysis Performed and Visualizations:

To investigate the factors affecting points per game, I employed Principal Factor Analysis (PFA) and Common Factor Analysis (CFA) to identify latent variables influencing this metric. I used this for latent(factor engineering). I conducted three main PFAs: one using log-transformed variables from the player dataset after addressing normality, another using the original untransformed variables without removing aliased variables, and a third utilizing the player dataset with untransformed variables while removing highly correlated aliased variables. All PFAs were scaled to ensure comparability.

Furthermore, I conducted two CFAs: The first employed log-transformed variables version of the data to achieve normality, while the second used the untransformed dataset. I thought to try using both the log-transformed variables and the untransformed variables to see the effect on the results of the PFA and CFA. Interestingly, two gave similar models and the third a different result. Notably, I excluded position dummy variables while performing the CFA. Based on the scores obtained from the PFAs and

CFA, I built regression models to explore the relationship between the latent factors and points per game.Kaiser-Meyer-Olkin (KMO) test on the dataset gave an MSA value of 0.5.

Below are the PFA and CFA I performed on the player stat dataset and the analysis from the models.

PFA with log-transformed data

```
> summary(pf1)

Factor analysis with Call: principal(r = nba_plyrd, nfactors = 6, rotate = "varimax")

Test of the hypothesis that 6 factors are sufficient.
The degrees of freedom for the model is 49  and the objective function was  20.98
The number of observations was  609  with Chi Square =  12533.86  with prob <  0

The root mean square of the residuals (RMSA) is  0.04
> print(pf1$loadings,cutoff=.4,sort=T)

Loadings:
       RC1    RC2    RC3    RC5    RC4    RC6
G       0.688
GS      0.869
MP      0.959
X2PA    0.896
FTA     0.855
AST     0.810
STL     0.753
TOV     0.875
PF      0.791
X3PA    0.588 -0.599
BLK     0.512  0.647
Pos_C          0.917
Pos_PG                0.931
Pos_PF                       0.975
Pos_SF                              -0.950
Pos_SG        -0.439 -0.529 -0.444  0.552
Age                                        0.985

                 RC1   RC2   RC3   RC5   RC4   RC6
SS loadings    6.911 2.076 1.438 1.289 1.272 1.069
Proportion Var 0.407 0.122 0.085 0.076 0.075 0.063
Cumulative Var 0.407 0.529 0.613 0.689 0.764 0.827
```

The PFA produced loadings RC1 – RC6 which I identified as "actions in game", "3point attempts by position Small Guard & blocks Position Center", "synergy between the Point Guard and Shooting Guard", "synergy between the Shooting Guard & Power forward initiated by the Power Forward", "synergy between the Shooting Guard and Small Forward initiated by the SF" and "Age".

```
Call:
lm(formula = PTS ~ ., data = scores)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22631 -0.12519  0.02325  0.16087  0.82843

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.823549   0.057970  31.457  < 2e-16 ***
actions_in_game             0.084035   0.003198  26.274  < 2e-16 ***
`3pattemps_by_SG&BLK_by_C`  -0.035041   0.005820  -6.020 3.03e-09 ***
`synergy_btw_PG$SG`         -0.032678   0.007886  -4.144 3.91e-05 ***
`synergy_btw_SG&PF_intbyPF` -0.011602   0.008497  -1.365    0.173
`synergy_btw_SG&SF_intbySF` -0.006745   0.008261  -0.816    0.415
Age                         0.013213   0.010319   1.280    0.201
DRB                         0.206449   0.048216   4.282 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.257 on 601 degrees of freedom
Multiple R-squared:  0.8691,    Adjusted R-squared:  0.8675
F-statistic: 569.9 on 7 and 601 DF,  p-value: < 2.2e-16
```

A regression model on the named scores produced

Produced an R-squared of 86%, Adj R-squared of about 87%
It produced the model

PTS = 1.82 + 0.08(actions_in_game) – 0.035(3pattemps_by_SG&BLK_by_C) – 0.03(synergy_btw_PG&SG)

+ 0.206(DRB)

This indicates that a unit increase in all actions in the game (which is made up of 2point attempts, 3point attempts, Assists, Steals, Turnovers, Personal fouls and Blocks in a game) will increase points per game by 0.08 units. Also, a unit reduction in 3point attempts by a shooting guard and blocks by a Center will reduce points per game by 0.035 units*(This can go both ways with SG and C on opposing teams).* A unit reduction in the synergy between a point guard and a Shooting guard (in terms of assists by the PG) will reduce point per game by 0.03 units and a unit increase in dribbles will in increase points per game by 0.206 units.

For the Common Factor Analysis on the transformed data, position dummy variables were removed as it did not work with the CFA using factanal.

```
fact_loadings <- fitnba_ply1$loadings

scores3 <- as.data.frame(as.matrix(nba_factor) %*% fact_loadings)

head(scores3)
```

```
> print(fitnba_ply1)

Call:
factanal(x = nba_factor, factors = 4)

Uniquenesses:
  Age     G    GS    MP  X3PA  X2PA   FTA   AST   STL   BLK   TOV    PF
0.928 0.584 0.251 0.011 0.257 0.105 0.140 0.048 0.391 0.393 0.167 0.316

Loadings:
     Factor1 Factor2 Factor3 Factor4
Age                          0.261
G     0.256   0.501   0.291   0.123
GS    0.376   0.633   0.399   0.219
MP    0.514   0.576   0.517   0.354
X3PA  0.276           0.744   0.334
X2PA  0.757   0.513   0.228
FTA   0.766   0.453   0.262
AST   0.680   0.179   0.264   0.623
STL   0.360   0.402   0.308   0.472
BLK   0.196   0.740  -0.135
TOV   0.739   0.369   0.214   0.324
PF    0.343   0.691   0.152   0.258

                Factor1 Factor2 Factor3 Factor4
SS loadings       3.001   2.807   1.441   1.161
Proportion Var    0.250   0.234   0.120   0.097
Cumulative Var    0.250   0.484   0.604   0.701

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 245.84 on 24 degrees of freedom.
The p-value is 1.1e-38
> print(fitnba_ply1$loadings, cutoff=.4, sort=T)
```

```
> print(fitnba_ply1$loadings, cutoff=.4, sort=T)

Loadings:
     Factor1 Factor2 Factor3 Factor4
X2PA  0.757   0.513
FTA   0.766   0.453
AST   0.680                   0.623
TOV   0.739
G             0.501
GS            0.633
MP    0.514   0.576   0.517
BLK           0.740
PF            0.691
X3PA                  0.744
Age
STL           0.402           0.472

                Factor1 Factor2 Factor3 Factor4
SS loadings       3.001   2.807   1.441   1.161
Proportion Var    0.250   0.234   0.120   0.097
Cumulative Var    0.250   0.484   0.604   0.701
```

CFA produced factor loadings which I combined with the scores and named "2point attempts, Free throw attempts, Assists & Turnover in Minutes Played", "2Point attempts, Free throw Attempts, Blocks, Personal Fouls and steals in Game played" , "3Point attempts in Minutes Played" and "Steals and Assists". This seemed to make sense considering the loadings from the PFA as the first factor would represent the actions. The second , the activities of the shooting guard (SG) and Center (C) (as the SG are mainly in charge of shooting in the perimeter, and the center being the tallest would block. The third factor would represent the interaction in terms of 3points made by shooting guards and point guard in minutes played(Shooting guards mainly shoot, point guards assist the SG and shoot three pointers too), an

d the fourth factor would represent assists and steals(activities of the perimeter small forward and the Shooting guard).

The regression model using these latent variables gave:

```
> fft1 = lm(PTS ~., data=scores3)
> summary(fft1)

Call:
lm(formula = PTS ~ ., data = scores3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54763 -0.12249  0.02677  0.15772  0.63476

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.825955   0.060068  13.750  < 2e-16 ***
`2pa_FreeTA_Ast&Tov_in_MP`  0.194991   0.014383  13.557  < 2e-16 ***
`2Pa_FreeTA_BLKs_PF_STL_inG` -0.112444   0.008064 -13.944  < 2e-16 ***
`3PA _in_MP`                0.043511   0.023715   1.835    0.067 .
`STL_&_AST`                -0.047893   0.010765  -4.449 1.03e-05 ***
DRB                         0.167896   0.039670   4.232 2.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2438 on 603 degrees of freedom
Multiple R-squared:  0.8818,    Adjusted R-squared:  0.8808
F-statistic: 899.3 on 5 and 603 DF,  p-value: < 2.2e-16
```

PTS= 0.83+ 0.195(2pa_FreeTA_Ast&Tov_in_MP)- 0.112(2Pa_FreeTA_BLKs_PF_STL_inG)- 0.049(STL_&_AST) + 0.168(DRB).

This indicates that a unit increase in the (2point attempts, free throw attempts, assists and Turnover in minutes played) will increase the points per game by 0.195 units. Also a unit decrease in (2point attem pts, free throw attempts, blocks, personal fouls and steals in games played) will reduce points per gam e by 0.112 units. As mentioned earlier, *this could be the interaction of the SG and center*. Also a unit d ecrease in steals and assists would reduce points per game by 0.049 and a unit increase in dribble will i ncrease points per game by 0.168.

PFA and CFA using the data without log transforms and without same correlated aliased variab les removed.

```
> summary(ptpf2)

Factor analysis with Call: principal(r = nba_pstatpfa, nfactors = 7, rotate = "varimax")

Test of the hypothesis that 7 factors are sufficient.
The degrees of freedom for the model is 84  and the objective function was  26.93
The number of observations was  609  with Chi Square =  16037.15  with prob <  0

The root mean square of the residuals (RMSA) is  0.04
> print(ptpf2$loadings, cutoff=.4, sort=T)

Loadings:
       RC1    RC6    RC2    RC3    RC5    RC4    RC7
X2P     0.878
X2PA    0.892
FT      0.932
FTA     0.936
AST     0.648  0.430         0.410
TOV     0.796  0.410
G              0.715
GS      0.506  0.666
MP      0.596  0.753
X3P            0.639 -0.454
X3PA           0.638 -0.467
STL            0.668
PF             0.656  0.422
ORB                   0.794
BLK                   0.724
Pos_C                 0.833
Pos_PG                       0.919
Pos_PF                              0.983
Pos_SF                                     0.942
Pos_SG                -0.402 -0.523 -0.404 -0.568
Age                                               0.934

                 RC1   RC6   RC2   RC3   RC5   RC4   RC7
SS loadings    5.665 4.044 2.892 1.467 1.286 1.278 1.140
Proportion Var 0.270 0.193 0.138 0.070 0.061 0.061 0.054
Cumulative Var 0.270 0.462 0.600 0.670 0.731 0.792 0.846
>
```

Here, I named the groupings:
"2Pt, 2Pa, Free Throws & Free Throw Attempts, Assists &Turnovers in Minutes Played", "3PT & 3 PA with Steals and Blocks, Personal Fouls & offensive rebounds in Games Played", "3PT and 3PA by Shooting guard & Block, Personal Fouls and offensive rebounds by C", "Synergy btw Point Guard & Shooting Guard in terms of Assists", "synergy between PF&SG", "synergy between SF&SG" and "Ag e". Apart from the fact that this PFA split the
actions in game grouping into two in terms of minutes played and games played and the 2 point and 3-point split into those two, the other groupings look the same as that of the PFA with the transformed d ata above. The regression ran on it gave the model.

```
> ftu2 = lm(PTS ~ . -`2Pt_2Pa_FreeT&FreeTA_ASt&ToV_in_MP`, data= scores4)
> summary(ftu2)

Call:
lm(formula = PTS ~ . - `2Pt_2Pa_FreeT&FreeTA_ASt&ToV_in_MP`,
    data = scores4)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9462  -1.3550  -0.0312   1.2726  10.1633

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       7.60954    0.35032  21.721  < 2e-16 ***
`3PT&3PA_with_STL_BLK_PF&ORB_in_GP`  1.03238    0.03501  29.487  < 2e-16 ***
`3PFG&3PA_by_SG&BLK_PF_ORB_byC`     -0.14541    0.05072  -2.867  0.00429 **
`Synergy_btw_PG&SG_in_Ast`          -0.11230    0.07525  -1.492  0.13613
`synergy_btw_PF&SG`                 -0.13591    0.08596  -1.581  0.11439
`synergy_btw_SF&SG`                  0.08125    0.08295   0.979  0.32775
Age                                  0.26692    0.09984   2.673  0.00771 **
DRB                                  0.50744    0.12665   4.007 6.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.604 on 601 degrees of freedom
Multiple R-squared:  0.8534,    Adjusted R-squared:  0.8517
F-statistic: 499.7 on 7 and 601 DF,  p-value: < 2.2e-16
```

PTS = 7.61 + 1.032(`3PT&3PA_with_STL_BLK_PF&ORB_in_GP`) -0.15(`3PFG&3PA_by_SG &BLK_PF_ORB_byC`) + 0.27(Age) + 0.51(DRB)

(Note: Remember to see the de.finition of these variables in the Data dictionary explained at the introd uction. Rember that 3PT is 3 Points made and 3PA is 3 points attempted.)

This indicates that a unit increase in 3PT&3PA_with_STL_BLK_PF&ORB_in_GP(3PT & 3PA with S teals and Blocks, Personal Fouls & offensive rebounds in Games Played) will increase the points per g ame by 1.03 units. A unit decrease in the (3point attempt and 3pt made by the shooting guard and blo cks, offensive rebound and personal fouls by the center) will decrease points per game by 0.15 units. A unit increase in age will increase points per game by 0.27 and a unit increase in dribble will increase p oints per game by 0.51 units.

The CFA with the data with untransformed variable and without similar correlated aliased variables re moved did not work, perhaps because of the nearly perfectly correlated aliased variables.

PFA and CFA with untransformed variable in the data but with multicollinear and aliased vari ables removed.

The PFA produced very similar loadings to the PFA performed with the log-transformed data. It also produced similar groupings of factors which I gave the same names. "actions in game played", "3poin t attempts by Shooting Guard & Blocks by Center", "synergy btw Point Guard and Shooting Guard in t erms of assists", "synergy between Power Forwards & Shooting Guard initiated by the Power Forward "(This could screens the PF sets) , "synergy between the Shooting Guard & Small Forward" and "Age ".

```
> summary(pnt)

Factor analysis with Call: principal(r = nba_pstatpt4, nfactors = 6, rotate = "varimax")

Test of the hypothesis that 6 factors are sufficient.
The degrees of freedom for the model is 49  and the objective function was  22.09
The number of observations was  609  with Chi Square =  13199.96  with prob <  0

The root mean square of the residuals (RMSA) is  0.05
> print(pnt$loadings, cutoff=.4, sort=T)

Loadings:
       RC1    RC2    RC3    RC5    RC4    RC6
G       0.645
GS      0.841
MP      0.954
X3PA    0.668 -0.466
X2PA    0.881
FTA     0.817
AST     0.778         0.418
STL     0.724
TOV     0.872
PF      0.745
BLK     0.412  0.706
Pos_C          0.908
Pos_PG                0.900
Pos_PF                       0.973
Pos_SF                              -0.946
Pos_SG        -0.438 -0.519 -0.445  0.557
Age                                        0.976

                 RC1   RC2   RC3   RC5   RC4   RC6
SS loadings    6.557 2.007 1.499 1.285 1.277 1.056
Proportion Var 0.386 0.118 0.088 0.076 0.075 0.062
Cumulative Var 0.386 0.504 0.592 0.668 0.743 0.805
>
```

```
> ftut <- lm(PTS ~., data= scores5)
> summary(ftut)

Call:
lm(formula = PTS ~ ., data = scores5)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2183  -1.3319   0.0079   1.3088  10.6782

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.09162    0.32386  24.985  < 2e-16 ***
actions_in_game_played      0.91523    0.02885  31.729  < 2e-16 ***
`3pattemps_by_SG&BLK_by_C` -0.28191    0.06245  -4.514 7.64e-06 ***
`synergy_btw_PG$SG_in_Ast` -0.06641    0.07336  -0.905  0.36569
`synergy_btw_PF&SG_inTbyPF` -0.15564    0.08300  -1.875  0.06125 .
`synergy_btw_SG&SF`        -0.03417    0.08036  -0.425  0.67087
Age                        -0.17698    0.10047  -1.762  0.07866 .
DRB                         0.32466    0.11655   2.786  0.00551 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.515 on 601 degrees of freedom
Multiple R-squared:  0.8632,    Adjusted R-squared:  0.8616
F-statistic: 541.9 on 7 and 601 DF,  p-value: < 2.2e-16

> vif(ftut)
     actions_in_game_played  `3pattemps_by_SG&BLK_by_C`  `synergy_btw_PG$SG_in_Ast`
                   3.467936                    1.554280                    1.288170
`synergy_btw_PF&SG_inTbyPF`         `synergy_btw_SG&SF`                         Age
                   1.119864                    1.019193                    1.227123
                        DRB
                   3.996120
```

The regression model gave the formula
PTS = 8.09 + 0.92(actions_in_game_played) -0.28(3pattemps_by_SG&BLK_by_C) +0.32(DRB).
Which indicates that a unit increase in actions in games played (including 3PA,2PA, blocks, Free thro w Attempts, Turnover, steals, personal fouls) will increase Points per game by 0.92 units. A unit decr ease in 3 point attempts by the shooting guard and blocks by the center will decrease points per game b y 0.28 units and a unit increase in dribbles will increase point per game by 0.32.

The model had an adj-r-squared value of 86%.

The CFA with untransformed variables and Multicollinear ‘aliased variables’ removed from the datase t, produced.

```
> faut = factanal(nba_statpt5, factors =4)
> print(faut)

Call:
factanal(x = nba_statpt5, factors = 4)

Uniquenesses:
  Age     G    GS    MP  X3PA  X2PA   FTA   AST   STL   BLK   TOV    PF
0.942 0.589 0.334 0.005 0.216 0.064 0.151 0.005 0.426 0.485 0.162 0.326

Loadings:
     Factor1 Factor2 Factor3 Factor4
Age                  0.231
G     0.195   0.523   0.161   0.272
GS    0.441   0.565   0.224   0.319
MP    0.466   0.613   0.421   0.474
X3PA  0.295           0.380   0.738
X2PA  0.827   0.452           0.197
FTA   0.828   0.342           0.214
AST   0.662   0.110   0.731   0.101
STL   0.305   0.413   0.507   0.230
BLK   0.190   0.670  -0.106  -0.136
TOV   0.757   0.306   0.374   0.179
PF    0.273   0.713   0.262   0.148

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.119   2.532   1.473   1.169
Proportion Var   0.260   0.211   0.123   0.097
Cumulative Var   0.260   0.471   0.594   0.691

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 242.59 on 24 degrees of freedom.
The p-value is 4.83e-38
> print(faut$loadings, cutoff=.4, sort=T)

Loadings:
     Factor1 Factor2 Factor3 Factor4
X2PA  0.827   0.452
FTA   0.828
TOV   0.757
G             0.523
GS    0.441   0.565
MP    0.466   0.613   0.421   0.474
BLK           0.670
PF            0.713
AST   0.662           0.731
STL           0.413   0.507
X3PA                          0.738
Age

               Factor1 Factor2 Factor3 Factor4
SS loadings      3.119   2.532   1.473   1.169
Proportion Var   0.260   0.211   0.123   0.097
Cumulative Var   0.260   0.471   0.594   0.691
> fact_loadingsut <- faut$loadings
```

The groupings were named "2pa_FreeTA_Ast&Tov_in_MP&Gs", "2Pa_FreeTA_BLKs_PF_STL_in games", "STL&AST_inMP", "3pA_in _MP"). Again these may mean the same as postulated in the CF

A of the log transformed variables in the section above. The initial regression found all the scores in th e model significant but with very high vIf. Upon removing variables with high (Variable inflation fact or) vifs, it produced the very parsimonous model.

```
> summary(ftut4)

Call:
lm(formula = PTS ~ . - `3pA_in _MP` - `2pa_FreeTA_Ast&Tov_in_MP&Gs` -
    `2Pa_FreeTA_BLKs_PF_STL_ingames`, data = scores6)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1251  -1.8499  -0.2082   1.5181  13.6808

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.49465    0.32604  -7.651 7.87e-14 ***
`STL&AST_inMP`  0.30687    0.01416  21.665  < 2e-16 ***
DRB             0.88380    0.11946   7.398 4.62e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.576 on 606 degrees of freedom
Multiple R-squared:  0.7213,    Adjusted R-squared:  0.7203
F-statistic:   784 on 2 and 606 DF,  p-value: < 2.2e-16

> vif(ftut4)
`STL&AST_inMP`            DRB
      2.076966       2.076966
```

PTS = -2.49 + 0.31('STL&AST_inMP) + 0.88(DRB)

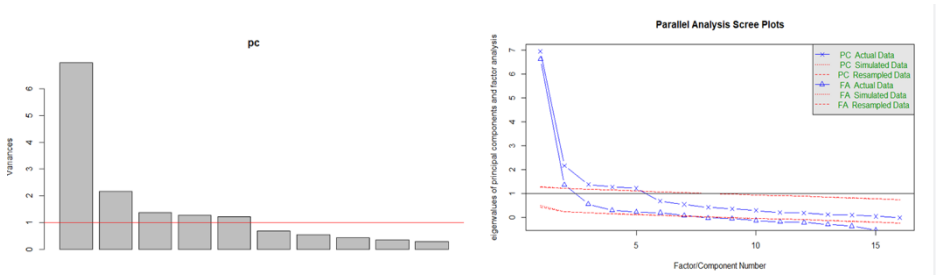
Indicating that a unit increase in steals and assists will increase points per game by 0.31 units and unit increase in dribble will increase the points per game by 0.88 units.

Worthy of Note/Summary:

My approach was to takeout the obvious variables that would affect points per game when such variables had the same correlation as the not so obvious ones. For example, most people would know that if you want to score more points, just make more three-point and two-point shots. However, I sought to see how attempts at field goal and other not-so-obvious variables would latently affect the points per game in any game. I therefore took 3points, 2points and free throws out, when their correlations were noticed to be exact with their attempts. This left room to seek latent factors. One of each Principal Factor Analysis(PFA) and Common Factor Analys (CFA) were eventually chosen to be modeled for (points per game) PTS after validation with each other and with all subsets' regression, out of about six factor Analysis conducted both on the log transformed and untransformed datasets.

For the Principal Factor Analysis chosen, The Kaiser-Meyer-Olkin (KMO) test for factor adequacy gave an overall MSA value of 0.5. Considering that this value is not too far from the accepted range of >0.5, I went ahead with my factor analysis. The scree plot for PFA identified 5 factors at the Var=1 criterion and at the knee. Parallel analysis suggested 5 components and 6 factors. Five factors were used. Defensive rebounds (DRB) and Age were found to be their own factors after doing a correlation test. DRB correlated with almost all variables and Age corellated very weakly. These were taken out and

later added to the scores of the PFA for regression. Dummy variables were used for the position variables. This helped me identify latent factors in the position variable.



The summary of the principal factor Analysis indicated a root mean square of residuals (RMSA) of 0.04.

PFA1 Loadings and Factor Model building

Fig : Chosen PFA Loadings

```
> print(pf1$loadings,cutoff=.4,sort=T)

Loadings:
       RC1    RC2    RC3    RC5    RC4
G       0.684
GS      0.868
MP      0.963
X2PA    0.892
FTA     0.852
AST     0.816
STL     0.754
TOV     0.875
PF      0.794
X3PA    0.597 -0.599
BLK     0.512  0.648
Pos_C          0.916
Pos_PG                0.929
Pos_PF                       0.974
Pos_SF                              0.951
Pos_SG        -0.437 -0.534 -0.446 -0.550

                 RC1   RC2   RC3   RC5   RC4
SS loadings    6.923 2.075 1.451 1.290 1.272
Proportion Var 0.433 0.130 0.091 0.081 0.079
Cumulative Var 0.433 0.562 0.653 0.734 0.813
```

The loadings of final PFA show latent factors RC1—RC5. The first factor I named **"actions in game"**. It represents 2-point attempts, free throw attempts, assists, steals, turnovers, personal fouls, 3-point attempts and blocks. Indeed, all actions that a player will engage in to score points and win a game. The scores from factor RC2 I named "3pattemps by Shooting guard and Blocks by Center". This factor indicates how the counteraction of 3-points scoring attempts by the Shooting Guard and blocks by the Center will affect the point per game. RC2 aptly named the 'synergy between the Point Guard and Shooting Guard' identified that latent contribution. RC5 named "synergy between the Shooting Guard and Power Forward, initiated by Power Forward" identified that interaction pair that affects points per game. RC4 named "synergy between the Shooting Guard and the Small Forward initiated by the Small Forward" signified that interaction parings between the positions to affect points per game. Varimax factor rotation was used to easily distinguish the factor loadings. The five factors covered a cumulative variance of 81.3%.

Note that defensive rebounds and Age variables identified from the correlation tests as their own factors were added to the factor scores and a regression model was built from the combined factors. All-subsets regression was used to validate the factor selection from the OLS model built from the factors.

PTS = 1.738 + 0.08(actions_in_game) - 0.04(3pattemps_by_SG_and_BLK_by_C) 0.03(synergy_btw_PG_and_SG) + 0.21 (DRB)

Is the model built from the chosen PFA. It had an R-squared value of 86.9% and and adjusted R-Squared of 86.7%. Because Age was its own factor but was not found significant in the model, a regression test of Age on PTS produced an R-squared of 1.1% indicating that it proved insignificant. Included in the model, it will do little to increase the predictive power of the model. I therefore took it out. Confirmatory factor analysis from the Lavan package in R to check goodness of fit unfortunately did not work on the model as the data contained position dummy variables that pointed out important factors in our PFA.

Chosen Common Factor Analysis (CFA)

The Position dummy variables were removed to conduct factor analysis in factanal in R. After removing DRB and Age which were again their own factors, the scree plots and parallel analysis suggested two factors. Two factors were used. The CFA identified two latent factors named on the scores as "Actions in game played minus 3point attempts" for Factor 1 and "Actions in games played including three point attempts minus blocks". Considering the position variables were removed, these factors made sense when domain knowledge is applied. It showed the counter-interacting effects of the 3points attempts and blocks. Knowing that the shooting guard is tasked with taking shots including 3-point shots and the center being usually the biggest is tasked with making blocks, showed how these factors pointed towards the results of the PFA chosen above. The cumulative variance captured on the CFA was about 68.4%.

An all-subsets regression done on the scores after the dummy variables, dependent variable and individual factors of defensive rebounds (DRB) and Age were added back found "actions in games played minus 3point attempts" , "Actions in games played including three-point attempts minus blocks", "Defensive rebounds" and Position center as significant. It captured an R-sqaured of 84.3% and Adj R-squared of 84.2%. Regression tests on Age, and Position Shooting guard to see if they should be included in the model since the all-subsets did not capture them showed that they had very negligible r-squared and adj-squared values and would not really have any significant effect in affecting the predictive power of the model. The R-squared and Adj- squared value, explainability of the model, and correspondence between the all-subsets and OLS regressions models helped me decide which Factor Analysis to use for the final regression model. The model built using the latent variables from PFA1 above scored highest on all grounds.

PTS = 1.738 + 0.08(actions_in_game) - 0.04(3pattemps_by_SG_and_BLK_by_C)
0.03(synergy_btw_PG_and_SG) + 0.21 (DRB)

Conclusions and Key Findings:

To gain valuable insights into the nature of the 2022/2023 NBA stats data and the player positions that affect scoring the most points in a game, it is important to note the existence of perfect multicollinearity and the need to develop strategies to address such issues. Additionally, I the significance of data dictionaries and the importance of domain knowledge, as they assisted in understanding the data and resolving analysis challenges should note be undermined.

For the factors affecting points per game in an NBA game and predicting points per game, Latent factors that were not immediately intuitive to the mind, affecting points per game were found. The effect of the synergies or counteraction (if on opposing teams) between the Shooting guard and the Center, seemed to be especially important. Other synergies between other position-pairings with the shooting guard like PF-SG, SF-SG, PG-SG in terms of assists, setting screens to give the SG room for field goal attempts etc., along with defensive rebounds, seem to be latent factors that affect the points per game. The regression model expression from the principal factor analysis for predicting points per game :

PTS = 1.738 + 0.08(actions_in_game) - 0.04(3pattemps_by_SG_and_BLK_by_C) - 0.03(synergy_btw_PG_and_SG) + 0.21 (DRB)

Indicates that ***a unit increase in all the actions aimed at scoring points per game will result in a corresponding increase of 0.08 in points scored per game. Also, counteraction in terms of a block by a center on a unit 3point attempt by a shooting guard will decrease the points per game by 0.04 units. The synergy (or constructive collaboration) between the point guard and the Shooting guard in terms of assists will increase or decrease points per game by 0.03 per unit assist. Also, a unit increase in defensive rebounds will increase points per game by 0.21 units. Team managers can therefore use this model to optimize game points.***