# Basketball Player Statistics Analysis

# CONTENTS

# Overview

- *Initial dataset:* ***2022-2023 NBA player Stats Regular season dataset***

  ***30 variables and 679 observations.***

- *Variables:*

| Rk | Player | Pos | Age | Tm | G | GS | MP | FG | FGA |
|---|---|---|---|---|---|---|---|---|---|
| Rank | Player's name | Position | Player's age | Team | Games played | Games started | Minutes played per game | Field goals | Field goal attempts |
| **FG%** | **3P** | **3PA** | **3P%** | **2P** | **2PA** | **2P%** | **eFG%** | **FT** | **FTA** |
| Field goal percentage | 3-point field goals | 3-point field goal attempts | 3-point field goal percentage | 2-point field goals | 2-point field goal attempts | 2-point field goal percentage | Effective field goal percentage | Free throws | Free throw attempts |
| **FT%** | **ORB** | **DRB** | **TRB** | **AST** | **STL** | **BLK** | **TOV** | **PF** | **PTS** |
| Free throw percentage | Offensive rebounds | Defensive rebounds | Total rebounds | Assists | Steals | Blocks | Turnovers | Personal fouls | Points |

Add reference here for definitions

# DATA CLEANING

```
> vif(fit1)
Error in vif.default(fit1) : there are aliased coefficients in the model
>
```

*The Original dataset had variables with Perfect Multicollinearity and Aliased coefficients, so to solve that problem, we split the dataset into:*

- **Player stats –** *Variables that show the individual players contribution to the outcome of a game*

```
> head(nba_plyrstat)
  Pos Age  G GS   MP X3P X3PA X2P X2PA  FT FTA ORB DRB AST STL BLK TOV  PF  PTS
```

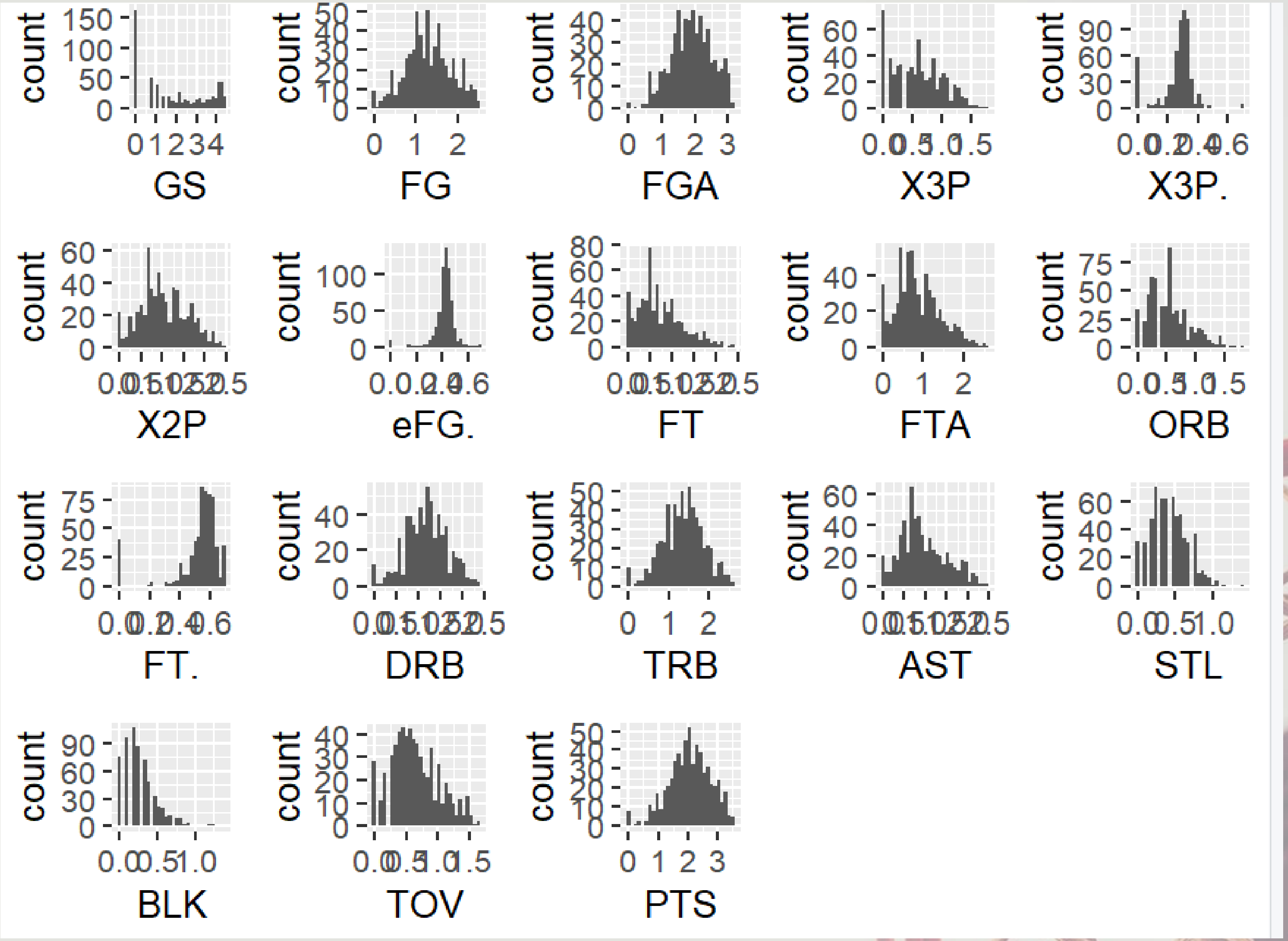
- **Team Stats** – *Variables that represent an individual's contribution to the overall team performance*

```
> head(nba_tmstat)
  Pos Age  G GS  MP  FGA  FG. X3P X3PA X3P. X2P X2PA X2P.  FT FTA  FT. ORB DRB AST STL BLK TOV  PF  PTS
```

- **Checked for missing variables**
- **Duplicate rows = 170 observations → total of player stats who play in two different positions.**

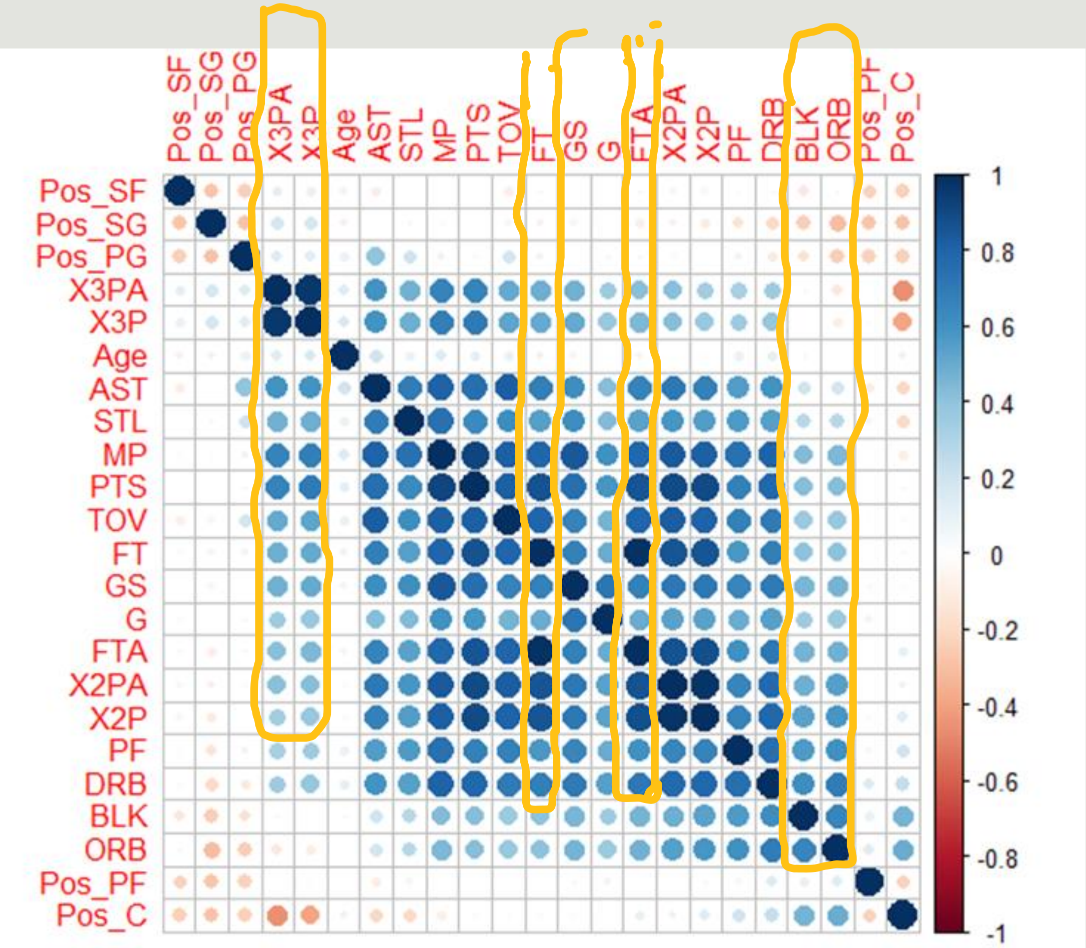
# TRANSFORMATIONS

- **We Performed Logarithmic Transformations on select variables to fix skewness**



# Research Goals

- We further reduced variables in the  player stats and team stats datasets to fix aliasing by removing vari ables with the same correlation profiles; keeping variables we considered less obvious.




Research Goals

- Find Factors that affect the points per game *(Using Player stats datasets)*

- Determine a relationship between player position a nd overall Stats. *(Using Team Stats data set.)*

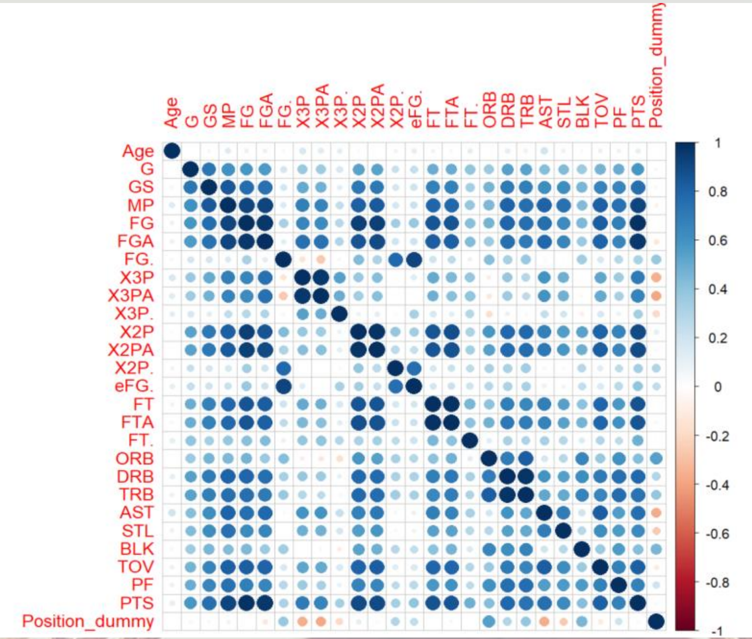
*<-- Variables such as x2p and x2pa, x3 p and x3pa, ORB and BLK, FT and FTA have almost exact collinearity profiles.*

# Why Lasso Regression?
# Examining OLS

**Ordinary Least Squares (OLS) Model has a couple issues:**

1. Initial OLS run produced aliased coefficients because of a perfectly collinear dataset.
2. The model is susceptible to overfitting (extremely high adjusted $R^2$ close to 1, inflated beta values).

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.4908843  0.0424585  11.562  < 2e-16 ***
Age             0.0027398  0.0013167   2.081 0.037877 *  
G               0.0018030  0.0003076   5.862 7.62e-09 ***
GS             -0.0270751  0.0072011  -3.760 0.000187 ***
MP              0.0016913  0.0023604   0.717 0.473937    
X3P             0.6590656  0.0535570  12.306  < 2e-16 ***
X3PA            0.0227123  0.0356617   0.637 0.524448    
X2P             0.7833536  0.0523142  14.974  < 2e-16 ***
X2PA           -0.0533027  0.0436751  -1.220 0.222787    
FT             -0.0886586  0.0646266  -1.372 0.170629    
FTA             0.2704313  0.0625804   4.321 1.82e-05 ***
ORB             0.0271722  0.0299511   0.907 0.364661    
DRB             0.1024038  0.0276395   3.705 0.000231 ***
AST             0.0185345  0.0244011   0.760 0.447810    
STL             0.0224773  0.0362615   0.620 0.535584    
BLK             0.0242226  0.0341599   0.709 0.478547    
TOV            -0.1348236  0.0349877  -3.853 0.000129 ***
PF              0.0315577  0.0122790   2.570 0.010413 *  
Position_dummy -0.0047746  0.0065100  -0.733 0.463587    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1295 on 590 degrees of freedom
Multiple R-squared:  0.9673,	Adjusted R-squared:  0.9664 
F-statistic:   971 on 18 and 590 DF,  p-value: < 2.2e-16
```




Variables are highly collinear; so, the variables have **repetitive collinearity profiles.**

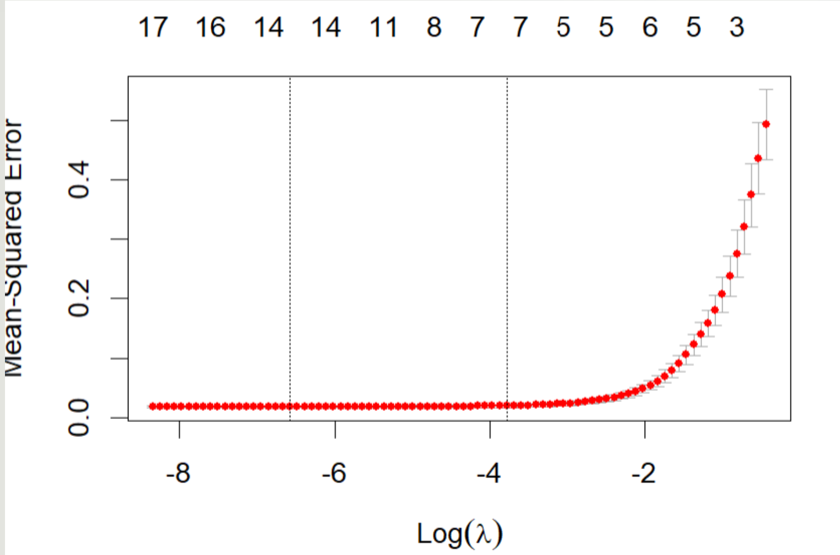
# CV Lasso & The Rough Penalty

(All Player stats including Aliased variables)

**First Lasso Call:**

- Cross-validated lambda range shown in the plot of the full model.
- Relevant Lambda range concerning $R^2$ and 1se shown in the image on the right. Lambda 1se used.
- Regularization impacts the R^2, which is important for a model's fit.
- Coefficients shown on the left. Ratio of RMSE is also off for this larger model.

```
> coef(fitRange, s="lambda.1se")
19 x 1 sparse Matrix of class "dgCMatrix"
                         s1
(Intercept)    0.6891376961
Age            .
G              0.0001894733
GS             .
MP             0.0066331959
X3P            0.5510309268
X3PA           .
X2P            0.6777640524
X2PA           .
FT             .
FTA            0.1271495959
ORB            .
DRB            0.0776164331
AST            .
STL            .
BLK            .
TOV            .
PF             .
Position_dummy .
```




Model equation: PTS = 0.68913 + 0.00018(G) + 0.00663(MP) + 0.55 103(x3p) + 0.67776(x2p) + 0.1271 4(FTA) + 0.07761(DRB)

```
> #Ratio of RMSE of Train to test in Lasso cv model
> rmselassortest / rmselassopredtr
[1] 0.8543259
```

```
> fitRange$lambda.1se
[1] 0.03956875
>
```

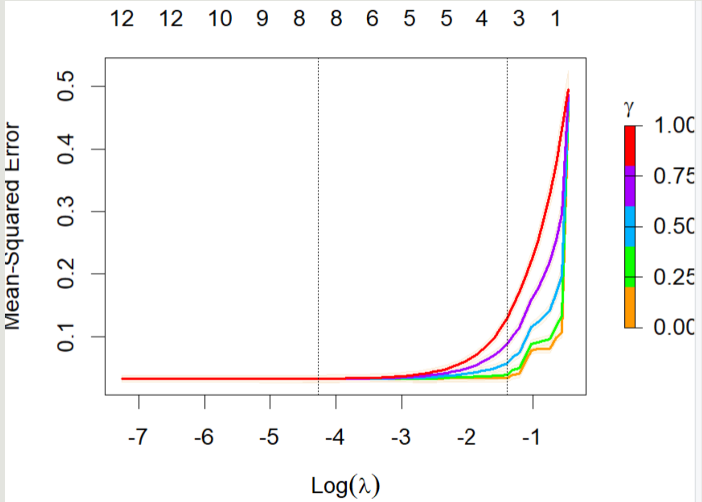
# Relaxed Lasso Regression

Two Lasso Regression runs (regular lasso first, relaxed lasso second).

- Relaxed Lasso best fit with 1se (1 standard error) has a gamma of 0, regularization **fixes** overfitting.
- Low penalty used, **lambda value < 0.05. Around 93% $R^2$.**
- Lambda 1se gives 5 predictors, coefficients shown to the side.
- Relaxed Lasso Model ends up being a small subset of the data captured by factor analysis/component analysis.

Model Equation:

PTS = 0.53637 + 0.0009(G) + 0.30103(X3PA) + 0.47514(X2PA) + 0.27077(FT) + 0.22062(DRB)




```
> rmse_lasso_train2   #0.18
[1] 0.1866196
```

```
#glmnet with lambda
glmnet(xTrain, yTrain, data=dsTrain, relax=T, lambda=0.04653347)
fitLasso2 = glmnet(xTrain, yTrain, data=dsTrain, relax=T, lambda = 0.04653347)
summary(fitLasso2)
coef(fitLasso2, s="lambda.1se")
#RMSE
pLassoTrain2 = predict(fitLasso2, xTrain, s="lambda.1se")
rmse_lasso_train2 = sqrt(mean((pLassoTrain2 - yTrain)^2))
```

```
> coef(fitLasso2, s="lambda.1se")
13 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)     0.5363707794
Age             .
G               0.0009251931
GS              .
X3PA            0.3010356751
X2PA            0.4751484473
FT              0.2707712163
DRB             0.2206258663
AST             .
STL             .
BLK             .
PF              .
Position_dummy  .
```

# Principal Factor Analysis and CFA

Goal: To find latent factors that affect points per game (PTS)

- **Two PFAs and two CFAs performed**
- **PFA1 and CFA 1** (*Log Transformed Player Stats dataset used with aliased variables removed)*
- **PFA and CFA 2** *(Untransformed Player Stats dataset used with aliased variables removed)*

# Principal Factor Analysis 1

PFA with log-transformed data

```
> summary(pf1)

Factor analysis with Call: principal(r = nba_plyrd, nfactors = 6, r

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

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
```

- Principal Factors
- RC1: **"Actions in game"**,
- RC2: **"3point attempts by position Shooting Guard & blocks by Position Center"**
- RC3: **"synergy between the Point Guard and Shooting Guard"**
- RC5: **"synergy between the Shooting Guard & Power forward initiated by the Power Forward"**
- RC4: **"synergy between the Shooting Guard and Small Forward initiated by the SF"**
- RC6: **"Age"**.
- A 7th factor **"Defensive rebounds"** is its own factor in the regression model

# Principal Factor Analysis 1

- A regression model on the named scores produced

```
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
```

Produced an R-squared of 87%, Adj R-squared of about 86.9%

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)

# Common Factor Analysis 1

(with log transformed variables and aliased vars removed)

- Position (dummy vars) removed

```
> 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
```

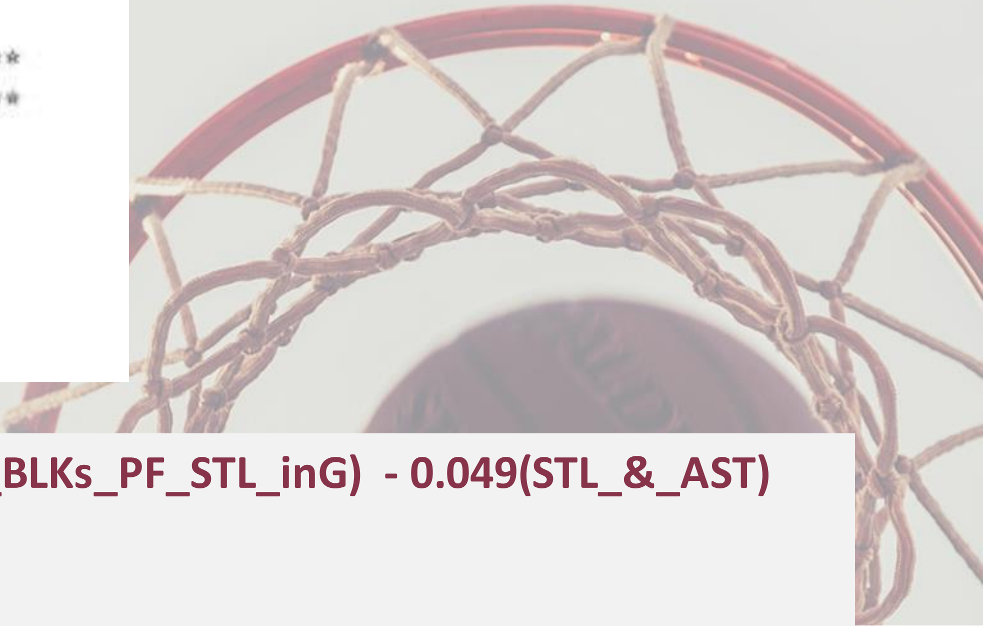
Factors:

- F1: "2 point attempts, Free throw attempts, Assists & Turnover in Minutes Played"
- F2: "2 Point attempts, Free throw Attempts, Blocks, Personal Fouls and steals in Game played"
- F3: **"3-Point attempts in minutes played"**
- F4: "Assists and Steals"
- A 5th factor **"Defensive rebounds"** is its own factor in the regression model

# CFA1 Model

```
> 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
```

R-squared of 88.1% and Adj. R-squared of 88%.



**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).**

# Principal Factor Analysis 2

## (With Untransformed variables but Aliased vars removed)

```
> 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
>
```

- Principal Factors
- RC1: **"Actions in game"**,
- RC2: **"3point attempts by position Shooting Guard & blocks by Position Center"**
- RC3: "synergy btw Point Guard and Shooting Guard in terms of assists"
- RC5: **"synergy between the Shooting Guard & Power forward initiated by the Power Forward"**
- RC4: **"synergy between the Shooting Guard and Small Forward initiated by the SF"**
- RC6: **"Age"**.
- A 7th factor **"Defensive rebounds"** is its own factor in the regression model

# PFA 2 Reg. Model

*(i.e. with Untransformed data)*

```
> 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 
```

Model: R-squared of 86.% and Adj. R-squared of 86.1%.

PTS = 8.09 + 0.92(actions_in_game_played) - 0.28(3pattemps_by_SG&BLK_by_C) + 0.32(DRB).

# Common Factor Analysis 2

**(With untransformed data but Aliased vars removed)**

- Again Position (dummy vars) removed

```
> 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
```

Factor names:

- F1: "2 point attempts, Free throw attempts, Assists & Turnover in Games started & Minutes Played"
- F2: "2 Point attempts, Free throw Attempts, Blocks, Personal Fouls and steals in Game played"
- F3: "Steals & Assists in minutes Played"
- F4: **"3-point attempts in minutes played"**
- A 5th factor **"Defensive rebounds"** is its own factor in the regression model

# CFA 2 Model

*(i.e. Untransformed with Aliased removed)*

```
> 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
```

R-squared of 72.1% and Adj. R-squared of 72%.

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

# Linear Discriminate Analysis

- ***Basketball player position:*** *(Point Guard (PG), Shooting Guard (SG), Small Forward (SF)Power Forward (PF), and Center (C)).*

- ***Goal:*** *To find out if there is a relationship between player's statistics and their position whether we are able to classify players positions based on the available statistics.*

- ***Classification Percentage:*** *LD1=80% , LD2=16%, LD3=2% and LD4=1%*

- *Steals, Effective Field Goal Attempt, assists and field goal attempts → +*
- *Defensive rebounds, field goal%, blocks, Offensive rebounds → -*

- *LD1 = -2.15(DRB) – 1.5(FG%) - 1.5(BLK) – 1.3(ORB) - 0.6(FTA) – 0.5(3P%) 0.40(PF) – 0.25(2PA) - 0.16(GS) – 0.06(Age) + 0.006(GP) + 0.04(MP) + 0.27 (3PA) + 0.33(TOV) + 1.03(FGA) + 1.4(AST) +1.5(eFG%) + 1.9(STL)*

```
Call:
lda(position ~ ., data = train)

Prior probabilities of groups:
        1         2         3         4         5
0.2054176 0.2370203 0.1918736 0.1738149 0.1918736

Group means:
       Age        G       GS       MP      FGA       FG.      X3PA      X3P.     X2PA
1 26.59341 40.82418 2.012497 17.74945 1.713311 0.5537143 0.5150656 0.2507692 1.548131
2 25.11429 43.79048 1.902795 19.30190 1.939555 0.4362857 1.3275224 0.3472095 1.409498
3 25.70588 47.11765 2.160258 20.10824 1.913806 0.4822118 1.1364601 0.3243294 1.506858
4 25.07792 41.84416 1.925587 19.30000 1.861026 0.4504545 1.2743996 0.3278571 1.345770
5 26.48235 40.68235 1.894510 20.45294 1.965955 0.4085059 1.3215415 0.3193412 1.466027
       eFG.       FTA       ORB      DRB       AST       STL       BLK       TOV       PF
1 0.5829011 0.9728500 0.9110326 1.401858 0.7095006 0.3474485 0.5131784 0.6498294 1.947253
2 0.5214476 0.8408597 0.3790542 1.042335 0.9320702 0.4557841 0.2018925 0.6552036 1.503810
3 0.5455412 0.9042665 0.6713512 1.338861 0.8500674 0.4115781 0.3406205 0.6653396 1.780000
4 0.5343636 0.7856566 0.5049170 1.085816 0.8092808 0.4330453 0.2065212 0.5775447 1.593506
5 0.4825059 0.8875686 0.3582813 1.043785 1.3572042 0.5183068 0.1911225 0.7916296 1.469412

Coefficients of linear discriminants:
              LD1          LD2           LD3         LD4
Age  -0.066387955  0.038806220   0.002981295 -0.01447974
G     0.005762483 -0.004452028  -0.006479974  0.01639646
GS   -0.160213448  0.108195391   0.122160497 -0.03894010
MP    0.042041278 -0.123618343  -0.022966078 -0.12283366
FGA   1.031133971  2.545720310   5.229403490  3.40931997
FG.  -1.552167694  7.947341994 -20.546633132 13.77484206
X3PA  0.272264576 -1.587278856  -4.035364663 -0.65242262
X3P. -0.540113487  0.604140545  -2.099282095  3.53155825
X2PA -0.259738220 -2.411814937  -0.978353045 -1.93920939
eFG.  1.495835306 -8.756661302  20.524455185 -14.58115516
FTA  -0.598174049  0.389464110   0.340607295 -0.19255183
ORB  -1.273780535  0.083662540  -2.117140437 -2.82269476
DRB  -2.150132965 -0.537158512  -2.295472601  2.04027148
AST   1.395262855  3.222168908  -1.519834200 -1.06761257
STL   1.909767029 -0.631324189   0.969038172  0.18566662
BLK  -1.500617436  2.184963493   0.794696715  1.66766373
TOV   0.325075822  0.604953485   1.051484953  2.09355155
PF   -0.407278300 -0.150634988   0.466334753 -0.35124843

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8023 0.1542 0.0307 0.0129
```

# Linear Discriminate Analysis

***Training set:***

```
> print(fit.lda$scaling[order(fit.lda$scaling[,1]), ])
             LD1          LD2          LD3         LD4
DRB  -2.150132965 -0.537158512  -2.295472601  2.04027148
FG.  -1.552167694  7.947341994 -20.546633132 13.77484206
BLK  -1.500617436  2.184963493   0.794696715  1.66766373
ORB  -1.273780535  0.083662540  -2.117140437 -2.82269476
FTA  -0.598174049  0.389464110   0.340607295 -0.19255183
X3P. -0.540113487  0.604140545  -2.099282095  3.53155825
PF   -0.407278300 -0.150634988   0.466334753 -0.35124843
X2PA -0.259738220 -2.411814937  -0.978353045 -1.93920939
GS   -0.160213448  0.108195391   0.122160497 -0.03894010
Age  -0.066387955  0.038806220   0.002981295 -0.01447974
G     0.005762483 -0.004452028  -0.006479974  0.01639646
MP    0.042041278 -0.123618343  -0.022966078 -0.12283366
X3PA  0.272264576 -1.587278856  -4.035364663 -0.65242262
TOV   0.325075822  0.604953485   1.051484953  2.09355155
FGA   1.031133971  2.545720310   5.229403490  3.40931997
AST   1.395262855  3.222168908  -1.519834200 -1.06761257
eFG.  1.495835306 -8.756661302  20.524455185 -14.58115516
STL   1.909767029 -0.631324189   0.969038172  0.18566662
> head(scores)
  LD1                  LD2                   LD3                  LD4
1 "-1.99658242916429"  "-1.14885464369007"   "-0.837971583012739" "0.480228636222988"   "C"
2 "-4.18891946043541"  "2.17056102050635"    "0.396556727262277"  "-2.01896004907445"   "C"
3 "-1.80529261913702"  "1.3891261440982"     "3.83706458619154"   "0.238856706519333"   "C"
4 "0.785433232397382"  "-1.06821441748484"   "0.434391042451501"  "-0.590335449930503"  "SG"
5 "-0.915140582210908" "-1.04488017831057"   "-1.26103408235098"  "0.724455814154043"   "PF"
8 "1.08890824799499"   "-0.100469086329585"  "0.10807021331216"   "-0.058396018986508"  "SG"
>
```

***Player Positions:***

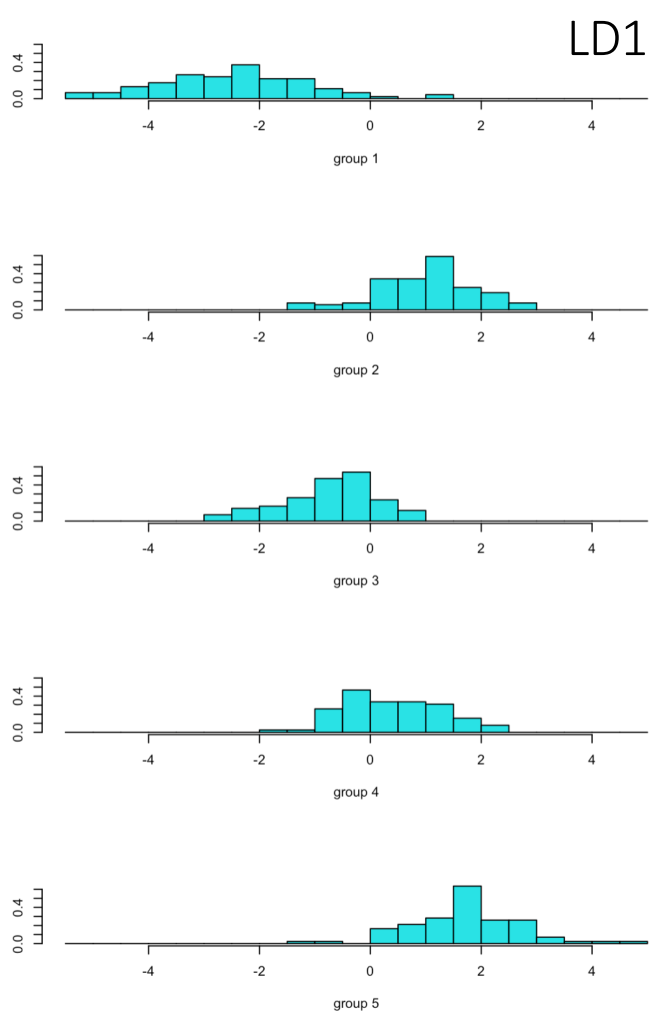
*Center (C) → 1*

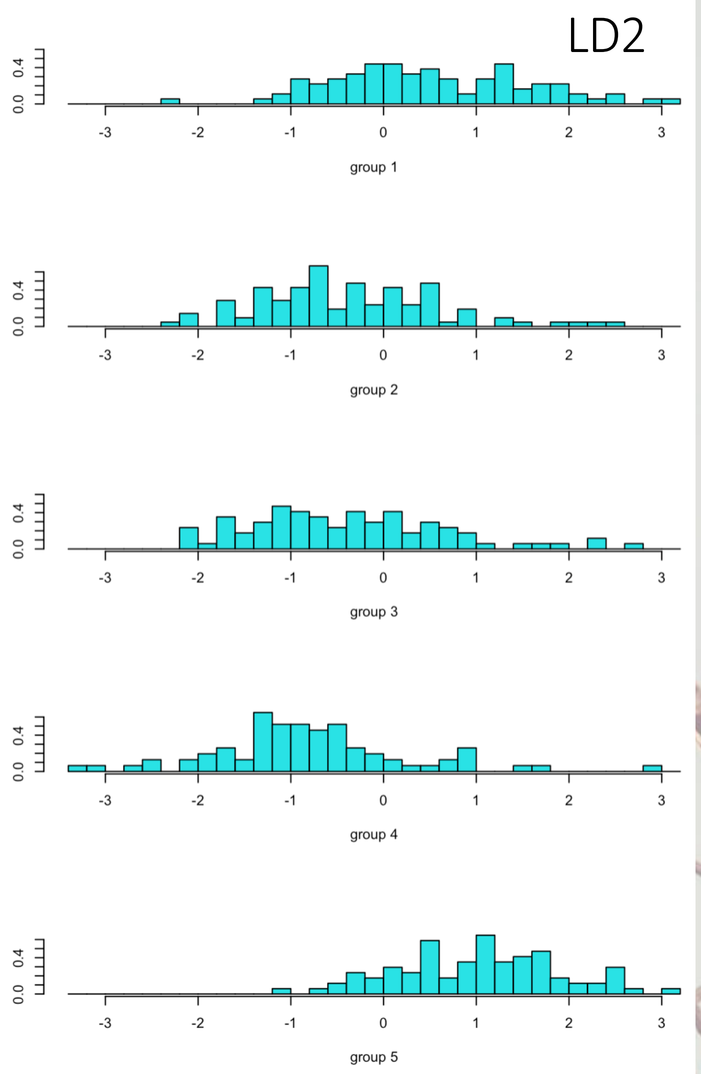
*Shooting Guard (SG) → 2*

*Power Forward (PF) → 3*

*Small Forward (SF) → 4*

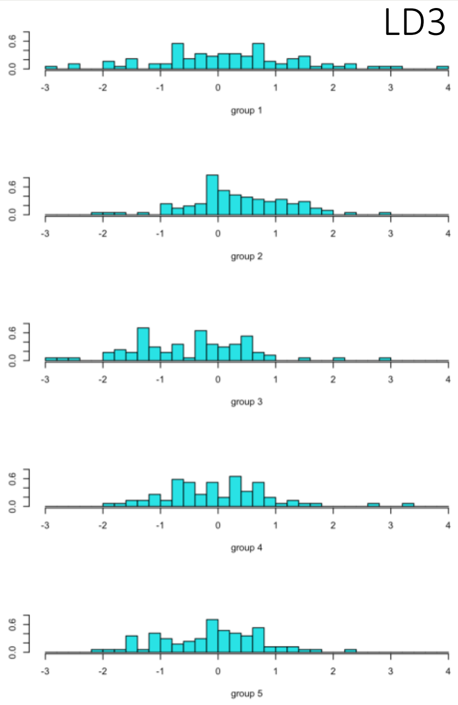
*Point Guard (PG) → 5*

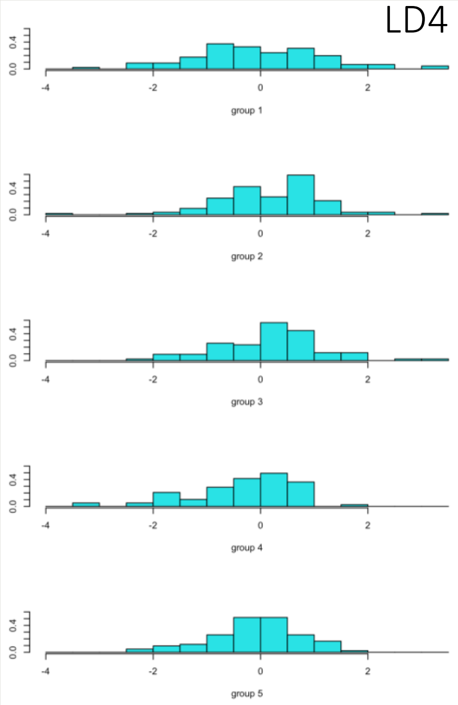


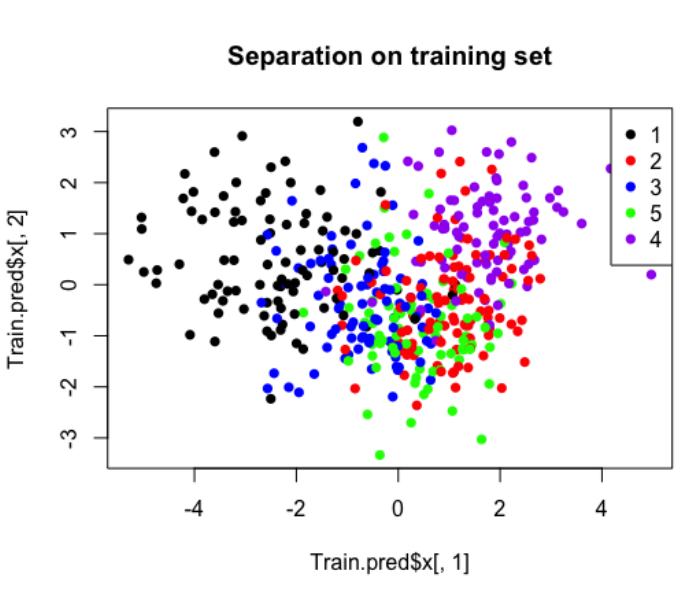



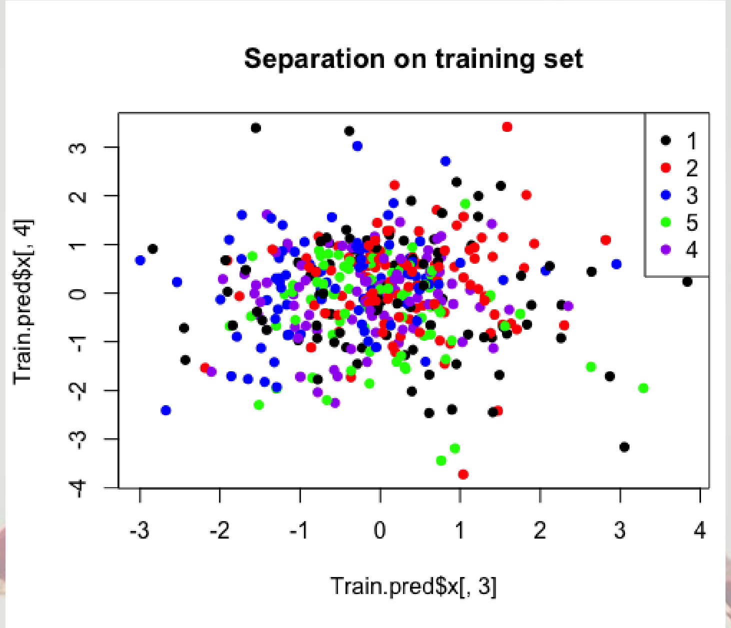
# Linear Discriminate Analysis

***Player Positions:***
*Center (C) → 1*
*Shooting Guard (SG) → 2*
*Power Forward (PF) → 3*
*Small Forward (SF) → 4*
*Point Guard (PG) → 5*

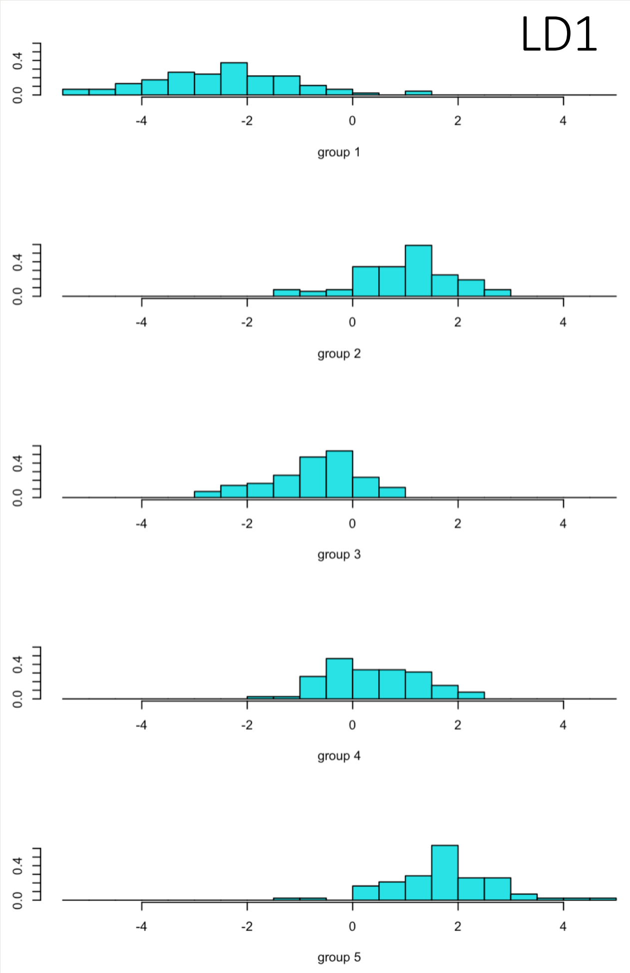


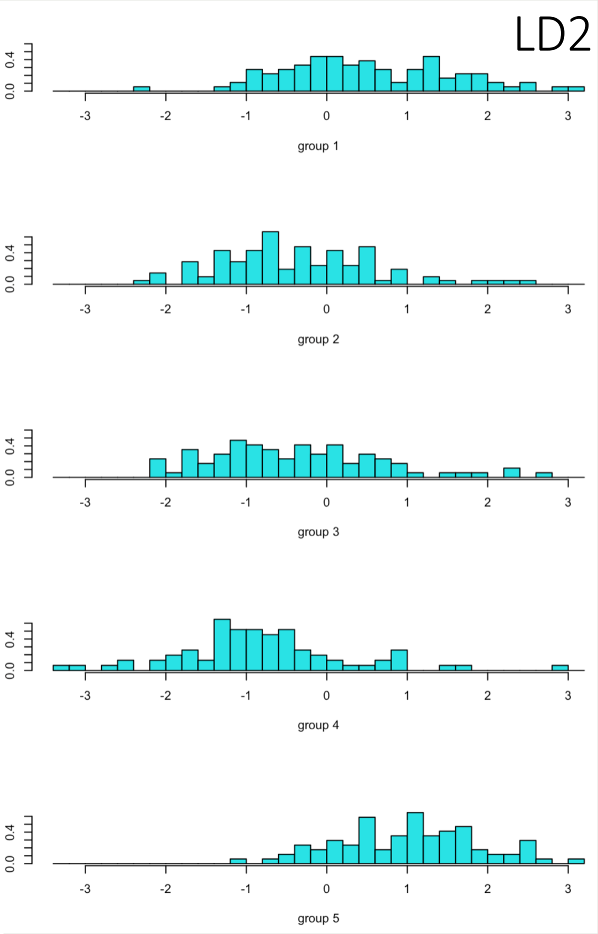



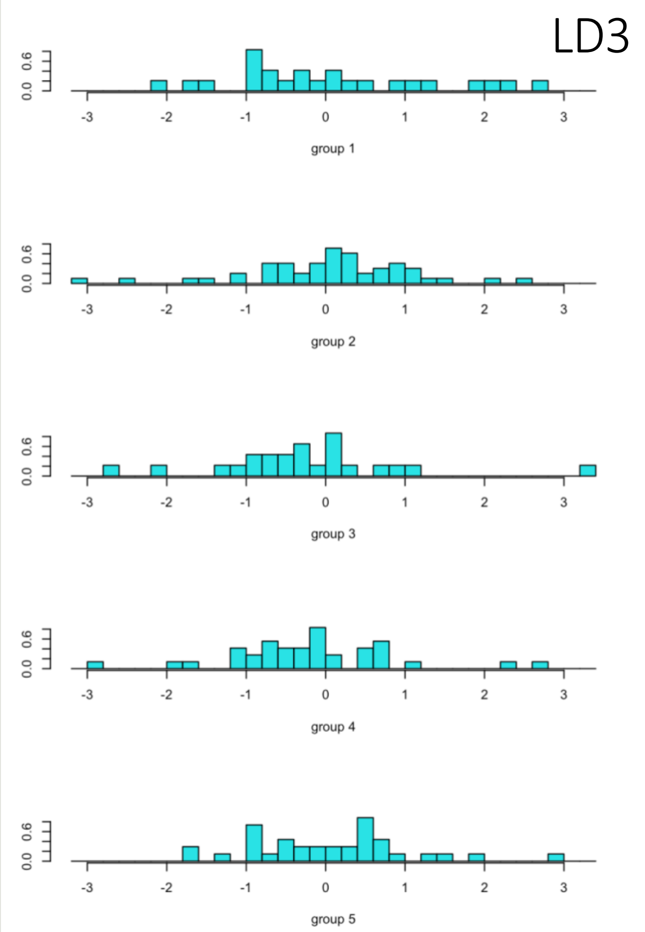



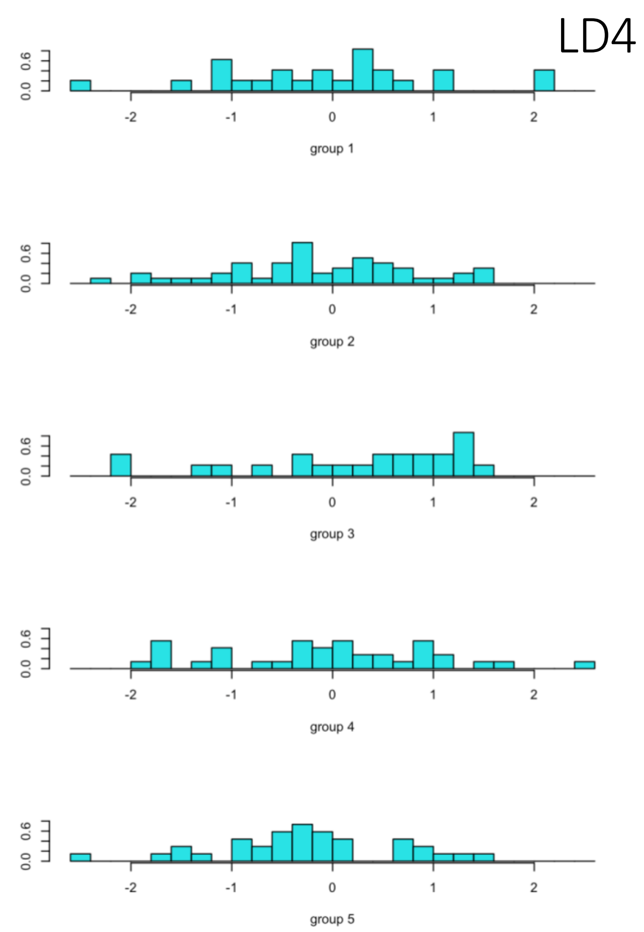



# Linear Discriminate Analysis

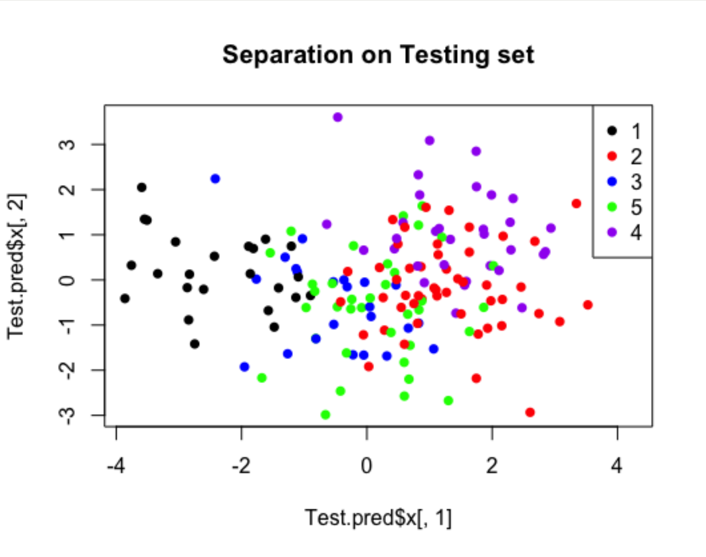


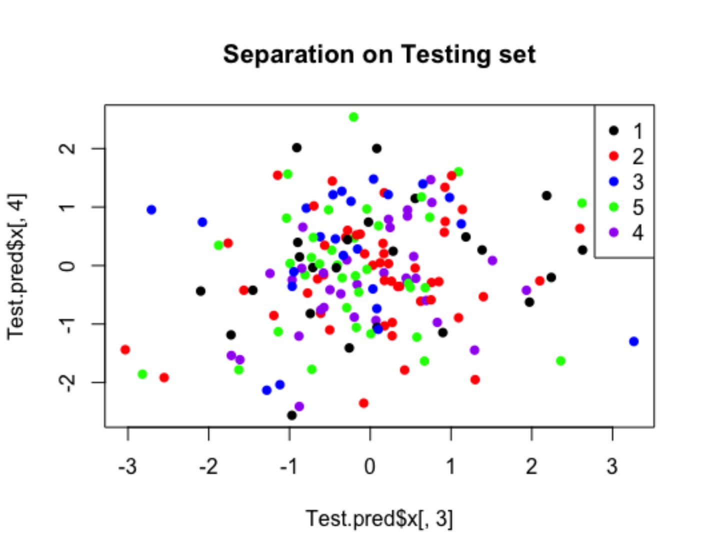



# Linear Discriminate Analysis




```
> confusionMatrix(data = Test.pred$class, reference = as.factor(test$position))
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1 16  0  1  1  0
         2  0 27  6  9 11
         3  8  3 11 14  1
         4  0  6  5  9  1
         5  0 13  0  3 21

Overall Statistics

               Accuracy : 0.506
                 95% CI : (0.4275, 0.5844)
    No Information Rate : 0.2952
    P-Value [Acc > NIR] : 1.017e-08

                  Kappa : 0.3716

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.66667   0.5510  0.47826  0.25000   0.6176
Specificity           0.98592   0.7778  0.81818  0.90769   0.8788
Pos Pred Value        0.88889   0.5094  0.29730  0.42857   0.5676
Neg Pred Value        0.94595   0.8053  0.90698  0.81379   0.8992
Prevalence            0.14458   0.2952  0.13855  0.21687   0.2048
Detection Rate        0.09639   0.1627  0.06627  0.05422   0.1265
Detection Prevalence  0.10843   0.3193  0.22289  0.12651   0.2229
Balanced Accuracy     0.82629   0.6644  0.64822  0.57885   0.7482
```

Confusion matrix for training set

```
> confusionMatrix(data = Train.pred$class, reference = as.factor(train$position))
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5
         1 69  0 12  0  0
         2  1 59 10 26 17
         3 18 11 52 17  2
         4  2 15  9 29  2
         5  1 20  2  5 64

Overall Statistics

               Accuracy : 0.6163
                 95% CI : (0.5692, 0.6618)
    No Information Rate : 0.237
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5182

 Mcnemar's Test P-Value : 0.2726

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.7582   0.5619   0.6118  0.37662   0.7529
Specificity            0.9659   0.8402   0.8659  0.92350   0.9218
Pos Pred Value         0.8519   0.5221   0.5200  0.50877   0.6957
Neg Pred Value         0.9392   0.8606   0.9038  0.87565   0.9402
Prevalence             0.2054   0.2370   0.1919  0.17381   0.1919
Detection Rate         0.1558   0.1332   0.1174  0.06546   0.1445
Detection Prevalence   0.1828   0.2551   0.2257  0.12867   0.2077
Balanced Accuracy      0.8621   0.7011   0.7388  0.65006   0.8374
```

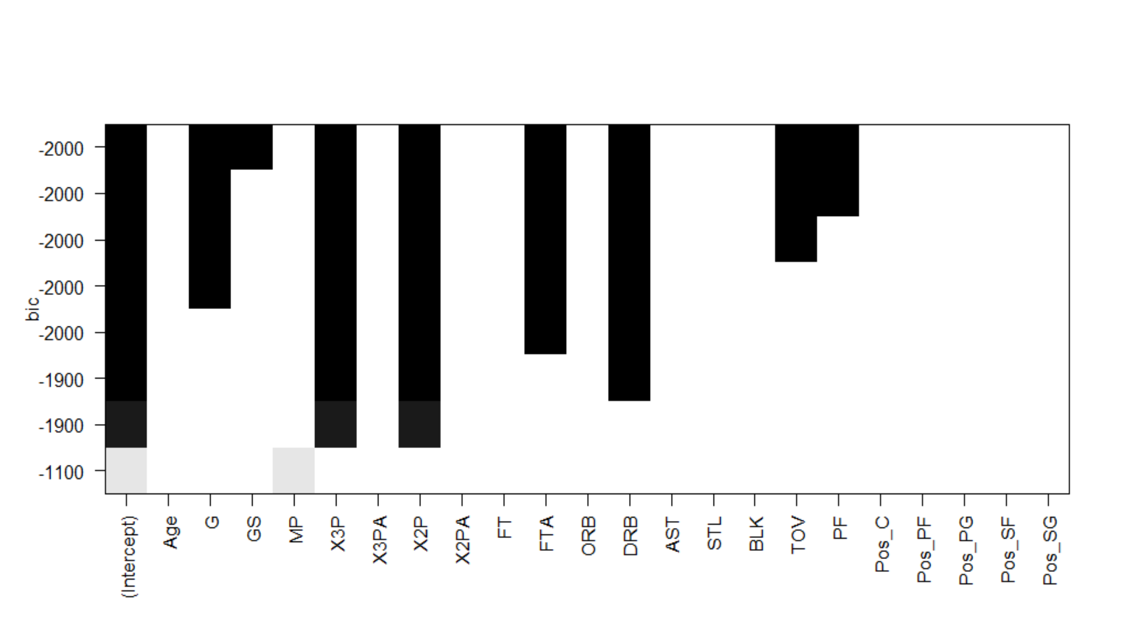
Confusion matrix for testing set

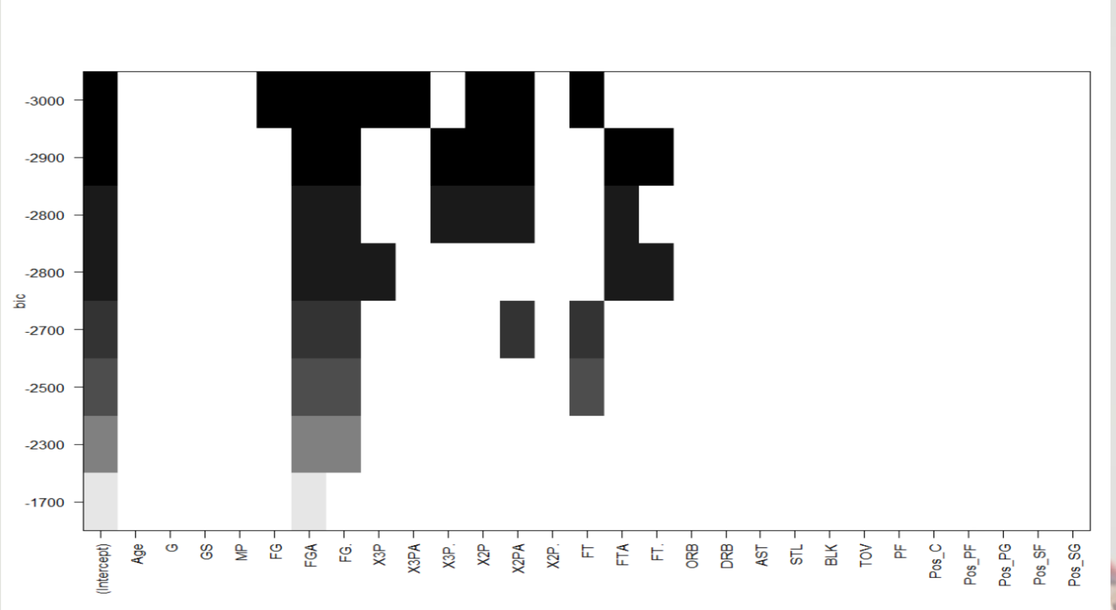
# All Subset Analysis

Player Set

Team Set

- *All Subsets regression used to compare models gotten from Lasso and FA to check for commonalities.*




- BIC scale used.
- Games Played, Games Started, 3 Points Per ga me, 2 Points per game, Defensive Rebound per , Free Throw Attempts per game, Turnovers pe r game, Personal Fouls per game.

- BIC Scale Used
- Field Goals per game, Field Goal attempts per g ame, Field Goal Percentage, 3 Points per game , 2 Points Attempts per game and Free Throws p er game.

# Model Building

## Team Set

```
> reg_team5 = lm(PTS ~ FG. + X3P + X2PA + FT, data = nba_teamstat3)
> summary(reg_team5)

Call:
lm(formula = PTS ~ FG. + X3P + X2PA + FT, data = nba_teamstat3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76737 -0.06075  0.01066  0.07374  0.64075 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.15917    0.02541   6.263 7.16e-10 ***
FG.          1.04085    0.05134  20.274  < 2e-16 ***
X3P          0.70901    0.01618  43.821  < 2e-16 ***
X2PA         0.55360    0.01741  31.791  < 2e-16 ***
FT           0.27188    0.02189  12.421  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1373 on 604 degrees of freedom
Multiple R-squared:  0.9625,	Adjusted R-squared:  0.9622 
F-statistic:  3870 on 4 and 604 DF,  p-value: < 2.2e-16

> vif(reg_team5)
     FG.      X3P     X2PA       FT 
1.231692 1.464042 3.930580 3.940987 
```

**Points Per Game = 0.16+ 1.05(FG.) + 0.71(X3P) + 0.5 6(X2PA) + 0.168(FT).**

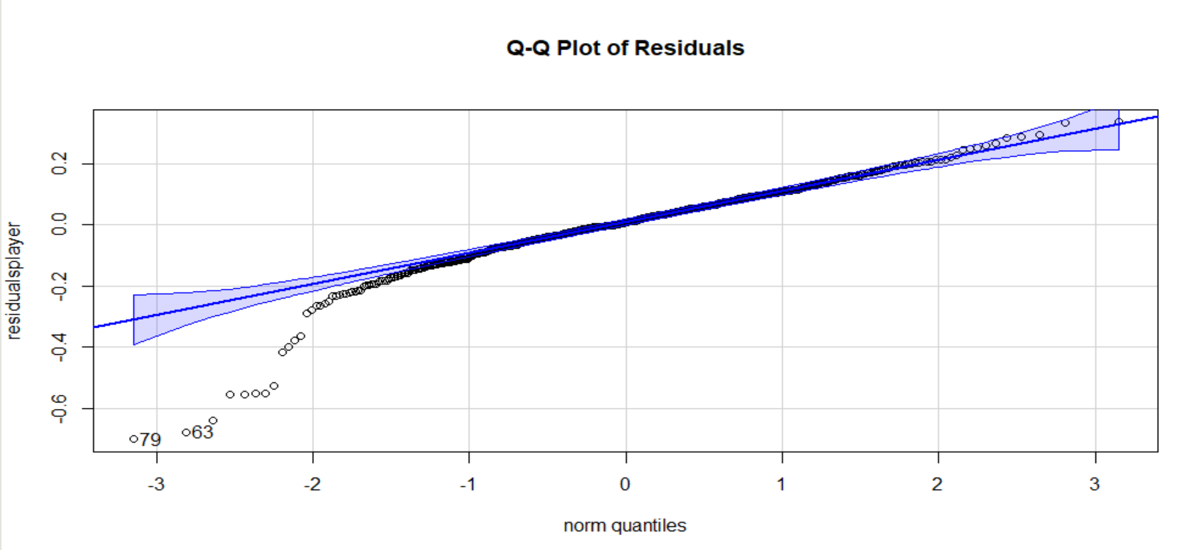
- PTS is the response variable.
- Parsimonious Models achieve d with Adjusted R Square of 9 6.22% for the team set and 96 .6% for the player set respecti vely.

## Player Set

```
> reg_player1 = lm(PTS ~ G + GS + X3P + X2P + FTA + DRB + TOV + PF, data = nba_playerstat3)
> summary(reg_player1)

Call:
lm(formula = PTS ~ G + GS + X3P + X2P + FTA + DRB + TOV + PF, 
    data = nba_playerstat3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.70240 -0.05931  0.00823  0.07745  0.33662 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.5503432  0.0168109  32.737  < 2e-16 ***
G            0.0017326  0.0003074   5.637 2.67e-08 ***
GS          -0.0229589  0.0062424  -3.678 0.000256 ***
X3P          0.6970232  0.0158277  44.038  < 2e-16 ***
X2P          0.7353393  0.0245622  29.938  < 2e-16 ***
FTA          0.1910758  0.0214622   8.903  < 2e-16 ***
DRB          0.1255805  0.0217221   5.781 1.19e-08 ***
TOV         -0.1155484  0.0287092  -4.025 6.43e-05 ***
PF           0.0405303  0.0110145   3.680 0.000254 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1301 on 600 degrees of freedom
Multiple R-squared:  0.9665,	Adjusted R-squared:  0.966 
F-statistic:  2163 on 8 and 600 DF,  p-value: < 2.2e-16

> vif(reg_player1)
       G       GS      X3P      X2P      FTA      DRB      TOV       PF 
2.134421 3.684532 1.559632 6.409905 4.865256 3.738321 3.891438 2.718180 
```

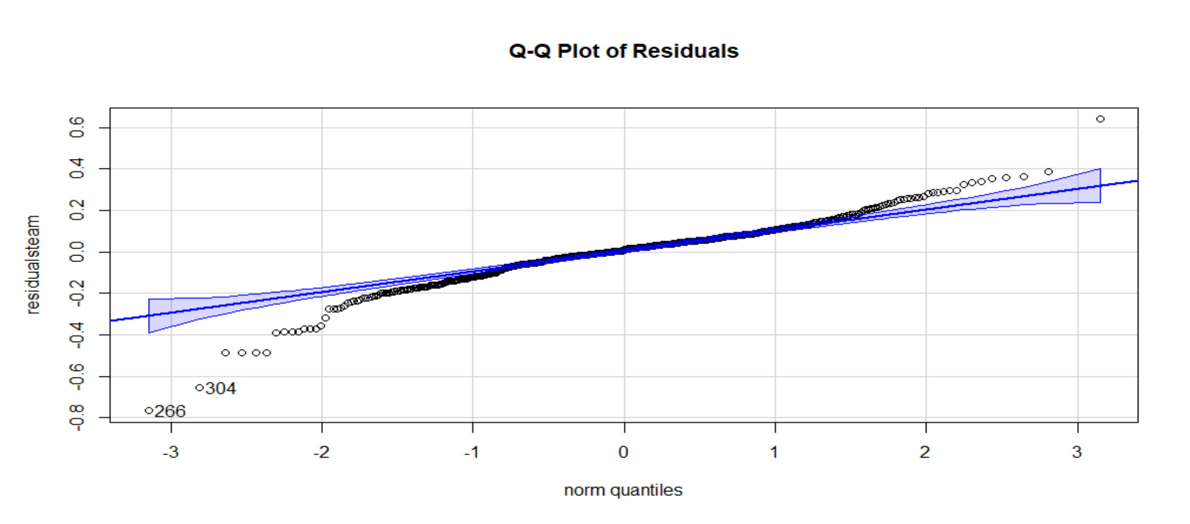
**Points Per Game = 0.56+ 0.0017(G) - 0.02(GS) + 0.70(X3PA ) + 0.74(X2P) + 0.20(FTA) + 0.13(DRB) - 0.12(TOV) + 0.04(PF) .**

# Residual Plots




QQ Plot of the Residuals for the **Player Set**

*From the plots, we can observe that the residuals approximately follow a normal distribution.*




QQ Plot of the Residuals for the **Team Dataset**

# Conclusion

- *Relaxed Lasso Regression*

*Relaxed Lasso Regression performed variable selection down to 5 parameters and showed that the model is not overfitting through a gamma value of 0. The model of relaxed lasso is like a small subset of the overall model captured by the Factor/Component Analyses.*

- *Factor Analysis ( PFA and CFA):*

*Latent factors not immediately intuitive that affect points per game were found out. The effect of the synergies or counteraction(if on opposing teams) between the* ***Shooting guard and the Center****, seemed to be very 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 FG attempts etc., along with* ***defensive rebounds****, seem to be latent factors that affect the points per game. Models are shown.*

- *Linear Discriminate Analysis:*

*LDA was able to classify* ***C , SG*** *positions from* ***PG*** *in* ***LD1*** *and PG from* ***PF & SF*** *in* ***LD2*** *with some confusion. Further analysis could be done by combining Guards and Forwards positions to reduce the classification confusion and specify the variables that able to classify them clearly.*

- *All Subset Analysis:*

*The models from the all-subsets regression was used to compare the results on the Lasso and Factor analysis. Variable combinations to predict point per game showed combinations of all actions in a game which tallies with one of the prediction factors in the PFA (and a more basic version of the model). It also shows a similar (more obvious) version of the predictors of the relaxed lasso. This indicates that the lines of analysis used produced similar and valid results.*

# THANK YOU