Analysis of PER index in modern NBA history (1980-2019)

STATISTICAL DATA ANALYSIS MOD.A

CELESTINO SANTAGATA, ANTONIO NAPPA

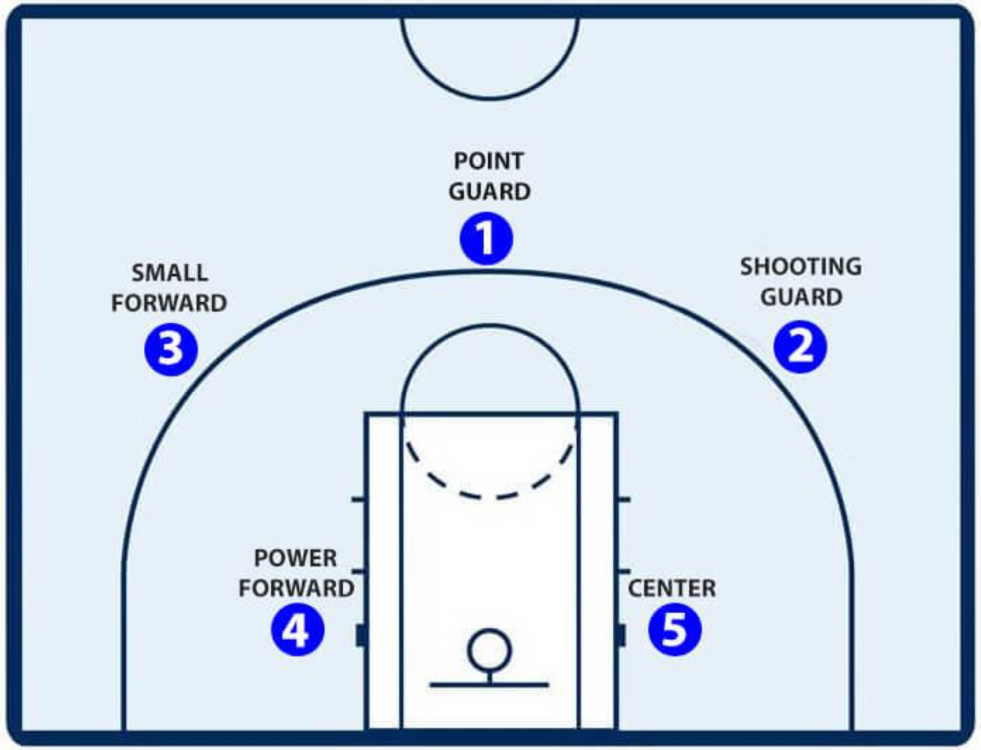
Index

- Introduction
- Dataset description
- Explorative data analysis
- Inferential data analysis
- Conclusions and future developments

Introduction

Five basketball positions:

1. Point Guard (PG)
2. Shooting Guard (SG)
3. Small Forward (SF)
4. Power Forward (PF)
5. Center (C)



Introduction

- How much do statistics matter in determining success in basketball?

  - https://www.interbasket.net/forums/showthread.php?29041-Modern-Basketball-and-Importance-of-Statistics

- “Many stats matter. End of story. Basketball, Football, Baseball, Hockey, Soccer, you pick the sport, there are statistics which really matter and cut to the core of the game. Understanding them is critically important.”

  “Many stats don't matter. There are a lot of stats people obsess over that actually don't say much.”

  - https://www.forbes.com/sites/quora/2016/05/10/how-much-do-statistics-matter-in-determining-success-in-sports/6ba24b9ef802

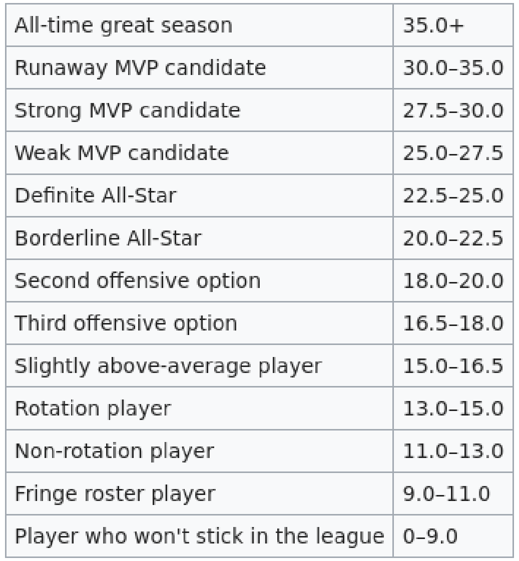
Introduction

- PER index

$$uPER = \frac{1}{min} \times \left(3P + \left[\frac{2}{3} \times AST\right] + \left[\left(2 - factor \times \frac{tmAST}{tmFG}\right) \times FG\right] + \left[0.5 \times FT \times \left(2 - \frac{1}{3} \times \frac{tmAST}{tmFG}\right)\right] - (VOP \times TO) - [VOP \times DRBP \times (FGA - FG)] - [VOP \times 0.44 \times (0.44 + (0.56 \times DRBP)) \times (FTA - FT)] \right.$$
$$\left. + [VOP \times (1 - DRBP) \times (TRB - ORB)] + (VOP \times DRBP \times ORB) + (VOP \times STL) + (VOP \times DRBP \times BLK) - \left[PF \times \left(\frac{lgFT}{lgPF} - 0.44 \times \frac{lgFTA}{lgPF} \times VOP\right)\right]\right)$$

- $factor = \frac{2}{3} - \left[\left(0.5 \times \frac{lgAST}{lgFG}\right) / \left(2 \times \frac{lgFG}{lgFT}\right)\right]$
- $VOP = \dfrac{lgPTS}{lgFGA - lgORB + lgTO + 0.44 \times lgFTA}$
- $DRBP = \dfrac{lgTRB - lgORB}{lgTRB}$

$$PER = \left(uPER \times \frac{lgPace}{tmPace}\right) \times \frac{15}{lguPER}$$



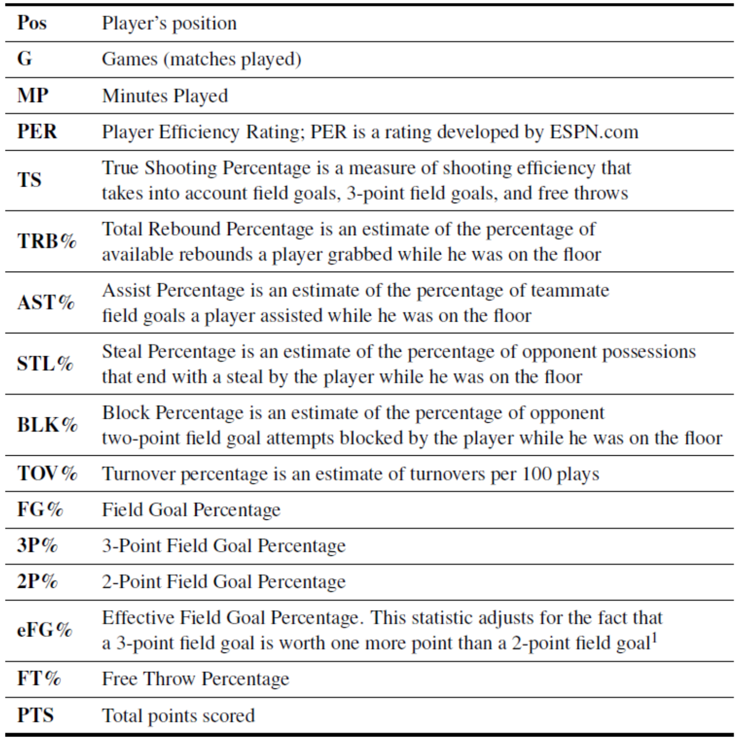
All-time great season	35.0+
Runaway MVP candidate	30.0–35.0
Strong MVP candidate	27.5–30.0
Weak MVP candidate	25.0–27.5
Definite All-Star	22.5–25.0
Borderline All-Star	20.0–22.5
Second offensive option	18.0–20.0
Third offensive option	16.5–18.0
Slightly above-average player	15.0–16.5
Rotation player	13.0–15.0
Non-rotation player	11.0–13.0
Fringe roster player	9.0–11.0
Player who won't stick in the league	0–9.0

Dataset description

Dataset source: https://www.kaggle.com/lancharro5/seasons-stats-50-19

26063 observations and 49 variables

1. Removal of unnecessary/redundant variables



Pos	Player's position
G	Games (matches played)
MP	Minutes Played
PER	Player Efficiency Rating; PER is a rating developed by ESPN.com
TS	True Shooting Percentage is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws
TRB%	Total Rebound Percentage is an estimate of the percentage of available rebounds a player grabbed while he was on the floor
AST%	Assist Percentage is an estimate of the percentage of teammate field goals a player assisted while he was on the floor
STL%	Steal Percentage is an estimate of the percentage of opponent possessions that end with a steal by the player while he was on the floor
BLK%	Block Percentage is an estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor
TOV%	Turnover percentage is an estimate of turnovers per 100 plays
FG%	Field Goal Percentage
3P%	3-Point Field Goal Percentage
2P%	2-Point Field Goal Percentage
eFG%	Effective Field Goal Percentage. This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal[1]
FT%	Free Throw Percentage
PTS	Total points scored

Dataset description

Dataset source: https://www.kaggle.com/lancharro5/seasons-stats-50-19

26063 observations and 49 variables

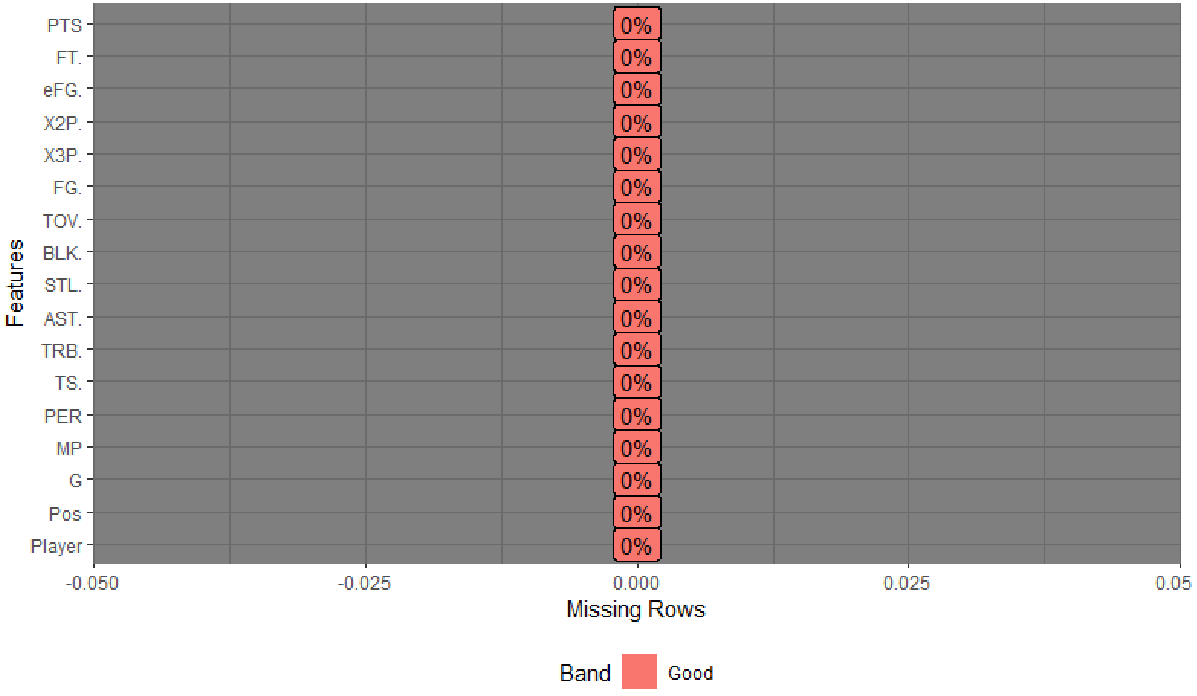
2. Filtering
   - years for the analysis to 1980-2019 (most of the variables were introduced starting from 1975, the latest is 3P and has been available since 1980)
   - minimum threshold: at least 10 games played, and for at least 12 minutes per game (a quarter is 12 min in NBA, for a total of 120 minutes)
   - double-counting of player statistics: error introduced by the way data was recorded for seasons in which a player was traded

   → Merging

2842 observations and 17 variables

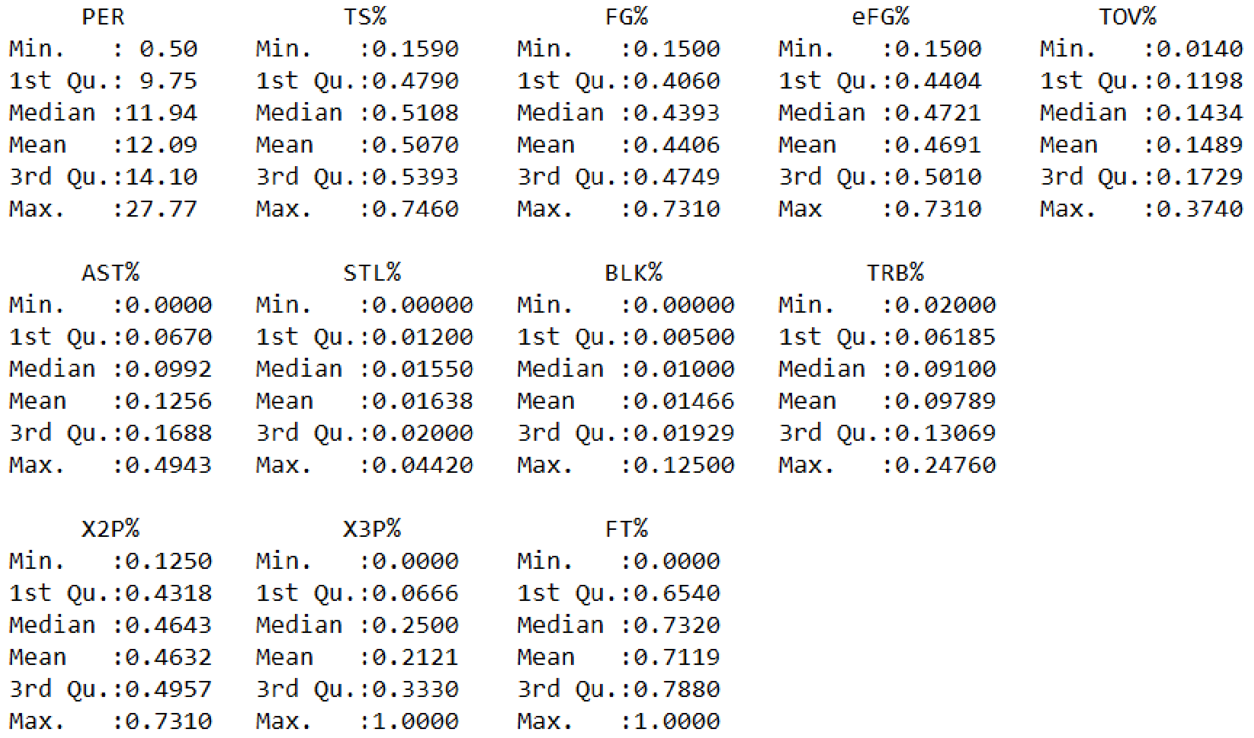
Dataset description

Missing values



Explorative data analysis

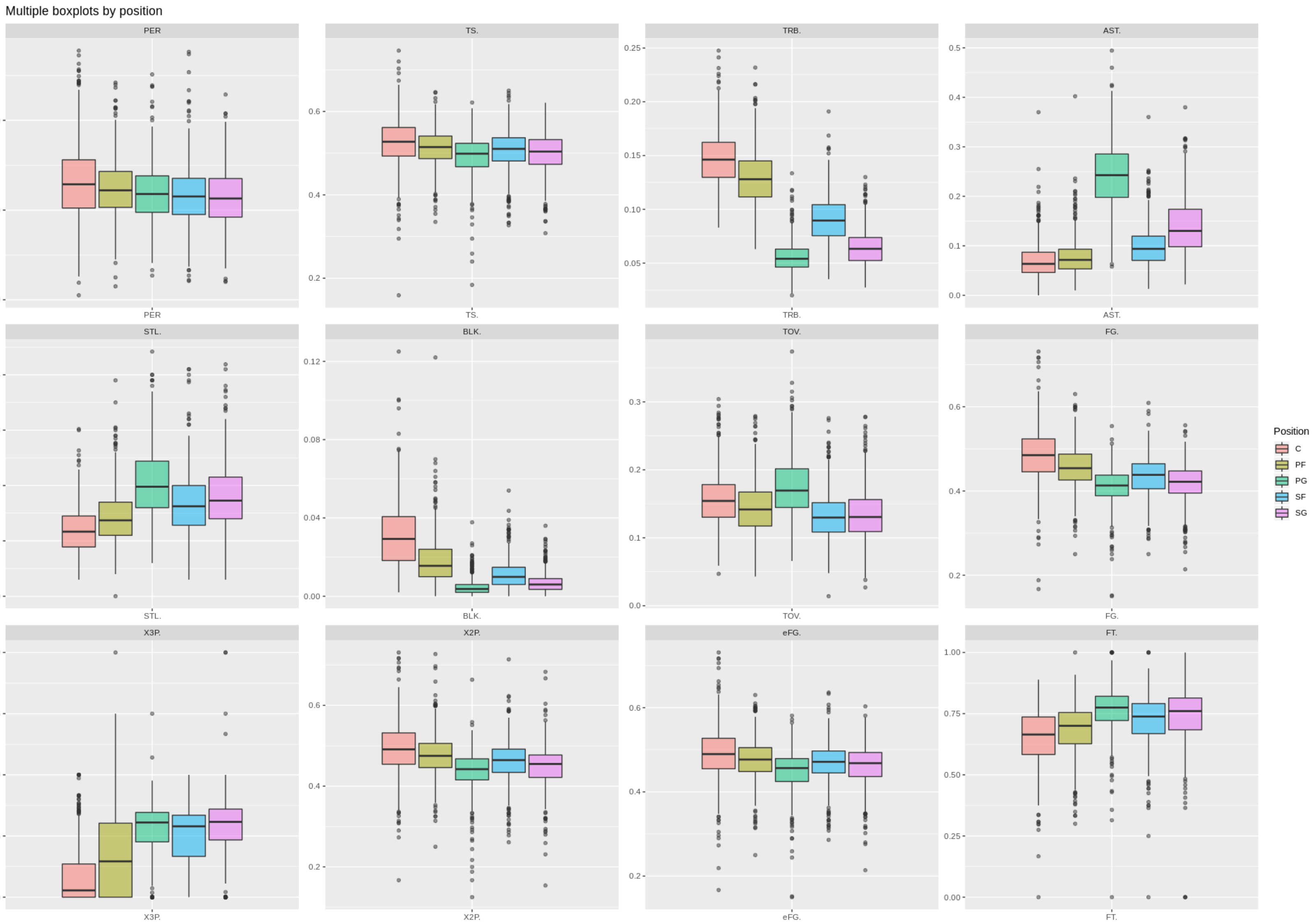
Main statistical indices for the quantitative variables of the data set



```
      PER              TS%               FG%               eFG%              TOV%
 Min.   : 0.50   Min.   :0.1590   Min.   :0.1500   Min.   :0.1500   Min.   :0.0140
 1st Qu.: 9.75   1st Qu.:0.4790   1st Qu.:0.4060   1st Qu.:0.4404   1st Qu.:0.1198
 Median :11.94   Median :0.5108   Median :0.4393   Median :0.4721   Median :0.1434
 Mean   :12.09   Mean   :0.5070   Mean   :0.4406   Mean   :0.4691   Mean   :0.1489
 3rd Qu.:14.10   3rd Qu.:0.5393   3rd Qu.:0.4749   3rd Qu.:0.5010   3rd Qu.:0.1729
 Max.   :27.77   Max.   :0.7460   Max.   :0.7310   Max    :0.7310   Max.   :0.3740

      AST%               STL%               BLK%               TRB%
 Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.02000
 1st Qu.:0.0670   1st Qu.:0.01200   1st Qu.:0.00500   1st Qu.:0.06185
 Median :0.0992   Median :0.01550   Median :0.01000   Median :0.09100
 Mean   :0.1256   Mean   :0.01638   Mean   :0.01466   Mean   :0.09789
 3rd Qu.:0.1688   3rd Qu.:0.02000   3rd Qu.:0.01929   3rd Qu.:0.13069
 Max.   :0.4943   Max.   :0.04420   Max.   :0.12500   Max.   :0.24760

      X2P%              X3P%              FT%
 Min.   :0.1250   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.4318   1st Qu.:0.0666   1st Qu.:0.6540
 Median :0.4643   Median :0.2500   Median :0.7320
 Mean   :0.4632   Mean   :0.2121   Mean   :0.7119
 3rd Qu.:0.4957   3rd Qu.:0.3330   3rd Qu.:0.7880
 Max.   :0.7310   Max.   :1.0000   Max.   :1.0000
```

POS				
C	PF	PG	SF	SG
513	512	492	506	536

The analysis will focus on finding the features that best distinguish players’ positions and, based on them, obtain a good model for PER.



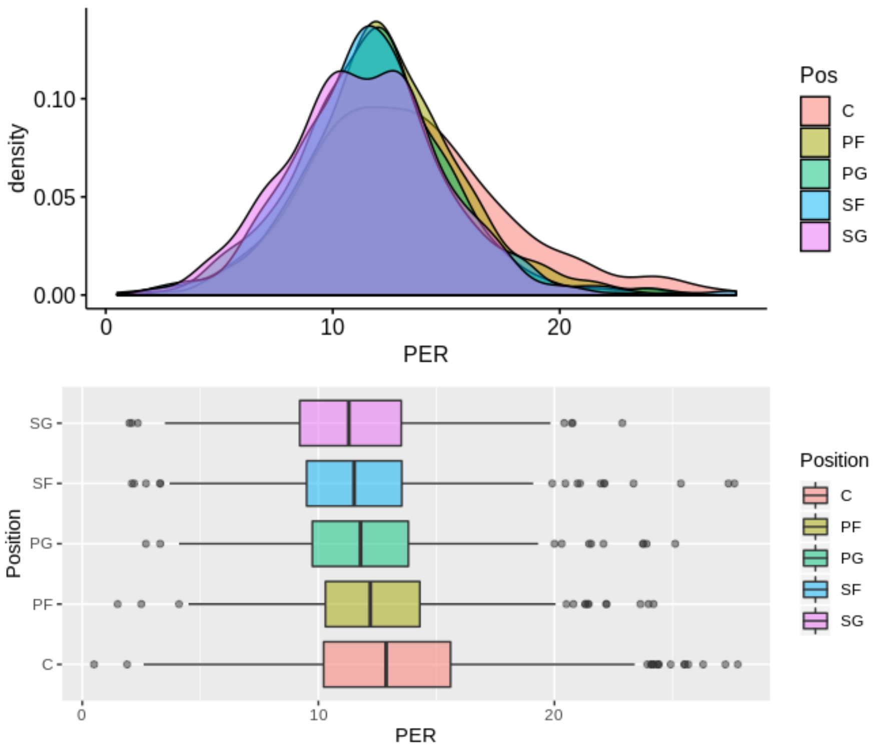
Explorative data analysis

BOX-PLOTS

Explorative data analysis

BOX-PLOTS

PER

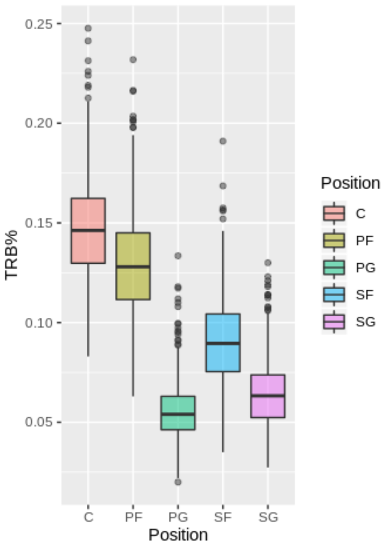


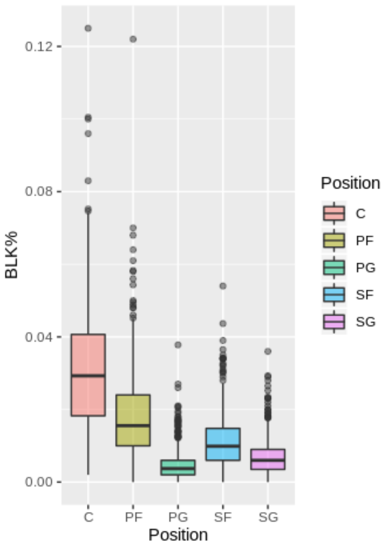
Although we expect there is no difference among positions, it seems that Centers are more performing (on average) than the others.

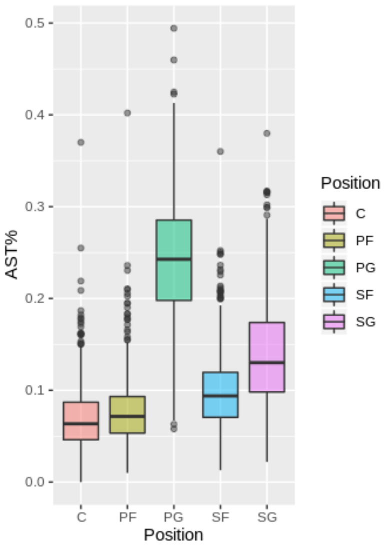
Criticism made against PER: tendency to reward inefficient shooting, so a player can be an inefficient scorer and simply inflate his value by taking a large number of shots.

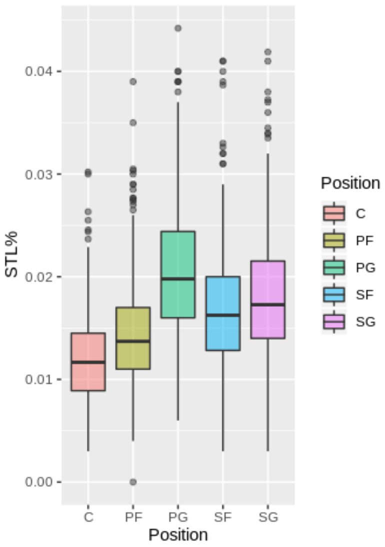
Explorative data analysis

BOX-PLOTS









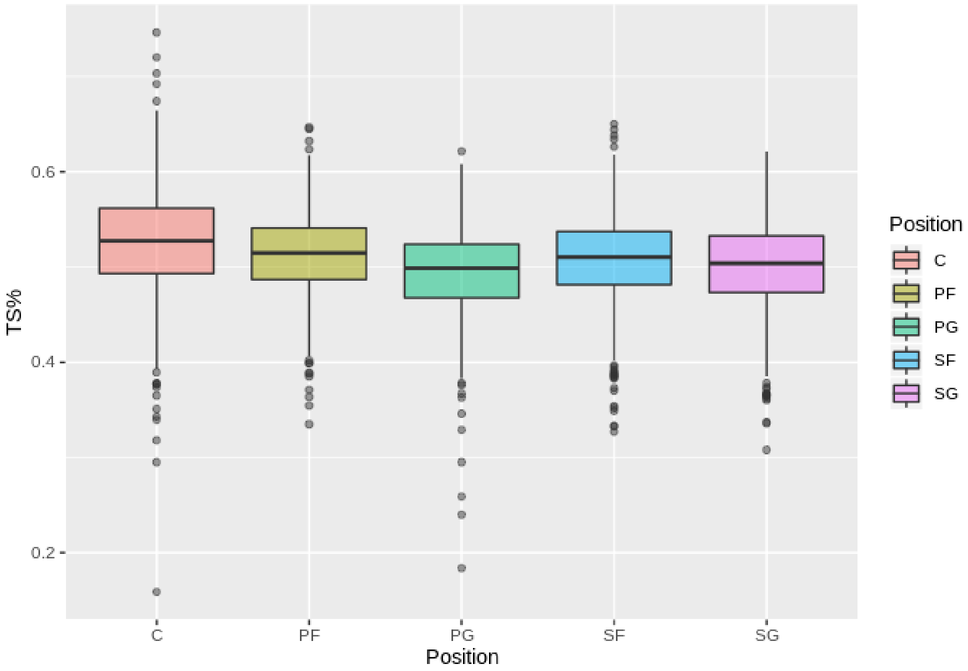
TRB%: we see that C and PF prevail (on average), then there is SF (as it plays close to the 3-point line), while SG and PG are the lowest as they are usually those furthest from the basket.

AST%: PG obviously dominates, because he is the ball carrier and sets the action to send his teammates to the basket; SG follows, while SF, PF and C values are lower on average because they are the point makers.

STL%: on average PG, SG and SF have higher values, because they guard the ball carriers of the opposite team, while C and PF tend to prevent opponents’ shots; indeed, C and PF have higher values regarding to blocking percentage (BLK%).

Explorative data analysis

BOX-PLOTS

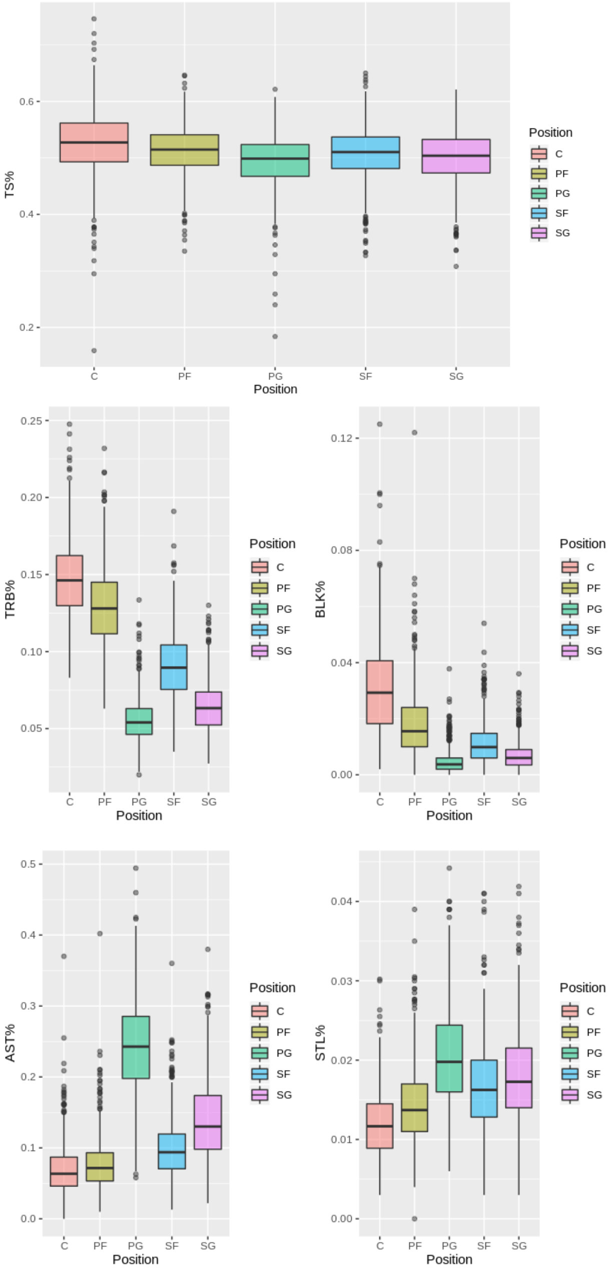


The distribution of TS% seems to not depend on player position, indeed mean values are quite equal.

TS% reflects a player’s scoring ability, so, regardless of position, it has a certain importance.

Explorative data analysis

BOX-PLOTS



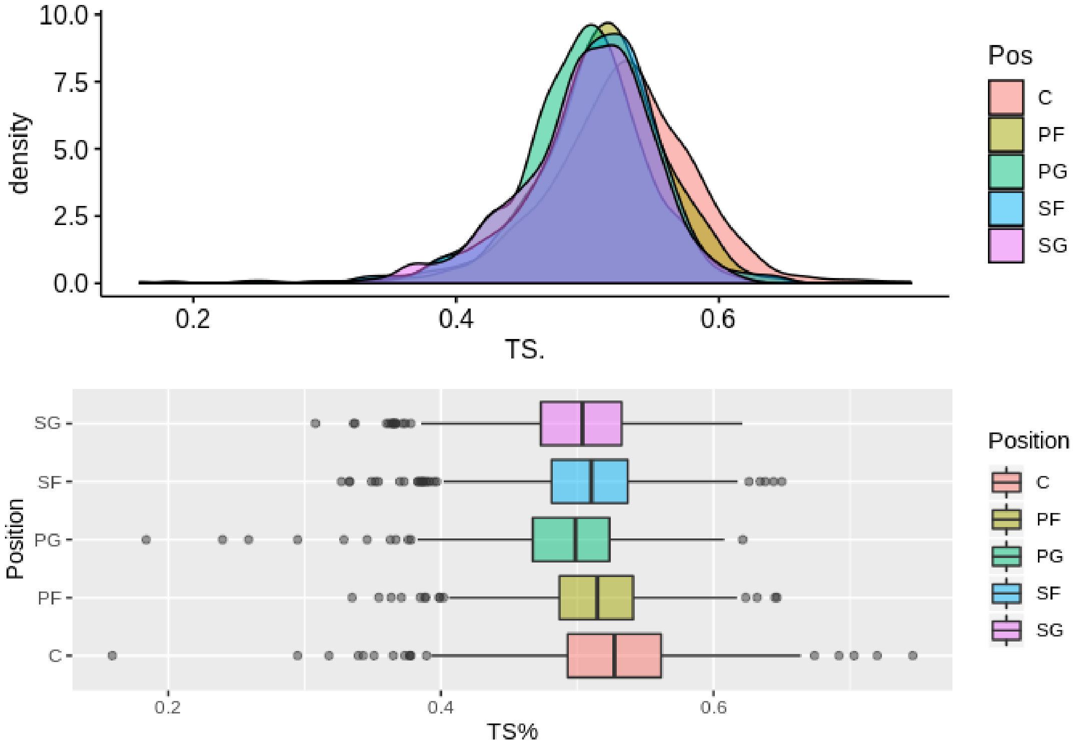
The skills that most distinguish the types of player are TRB%, AST%, STL%, BLK%: the pairs (AST%, STL%) and (TRB%, BLK%) are good descriptors for the players, respectively, outside the 3-point line and inside the 3-point line.

It is on the basis of this simple reasoning that we subsequently used such variables to characterize our model.

Explorative data analysis

FURTHER ANALYSIS

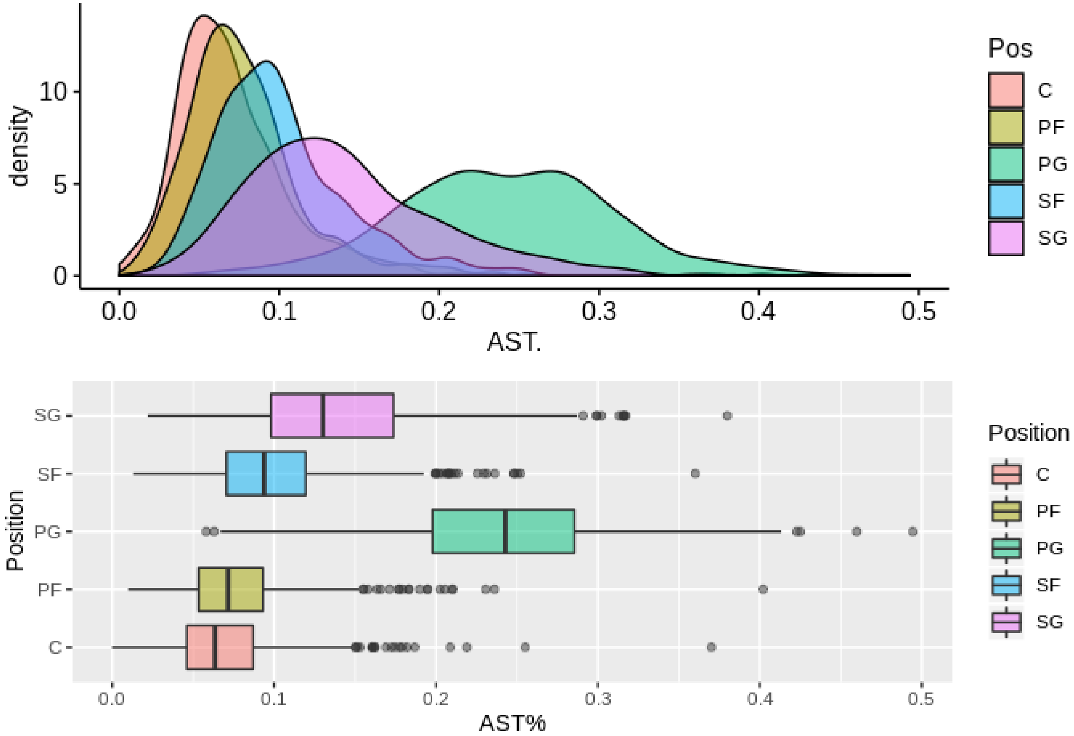
TS%



It is observed that conditional distributions are all approximately symmetric, with PG presenting a skewness in module greater than 1 due to the presence of many outliers under the minimum of the box-plot.

$$mean \lesssim median$$

AST%



Explorative data analysis

FURTHER ANALYSIS

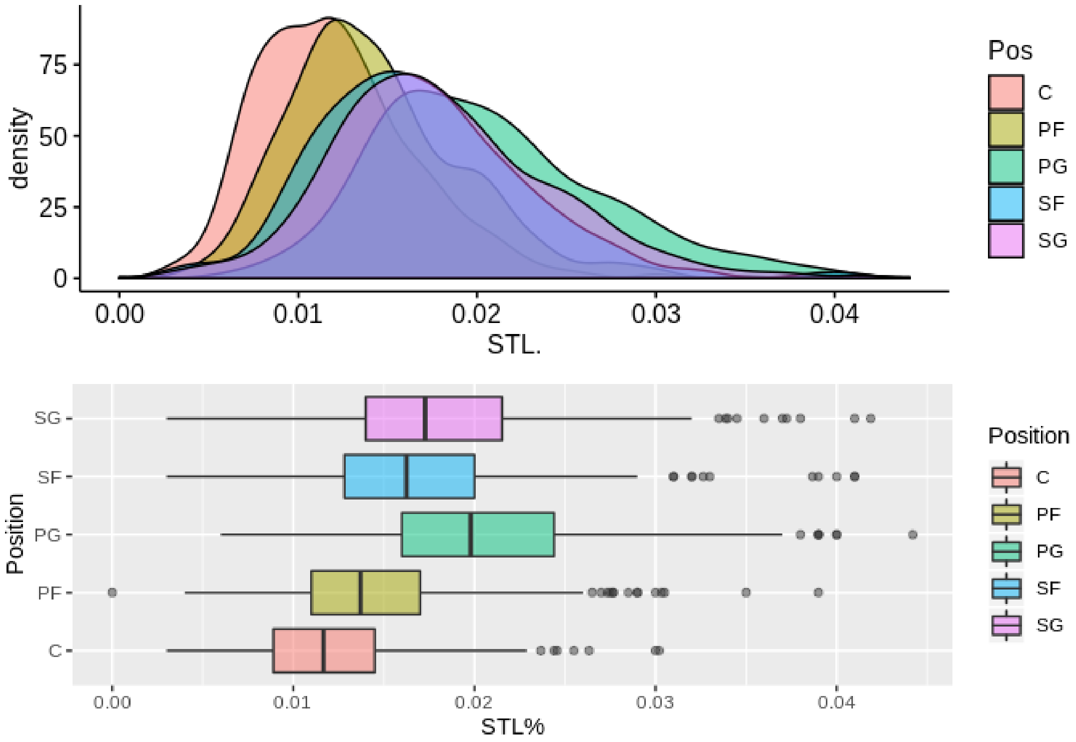
The conditional distributions are asymmetric (positive asymmetry); skewness values significantly different from zero, except for PG distribution, which seems quite symmetric (it seems to be bimodal).

$mean > median > mode$

Explorative data analysis

FURTHER ANALYSIS

STL%



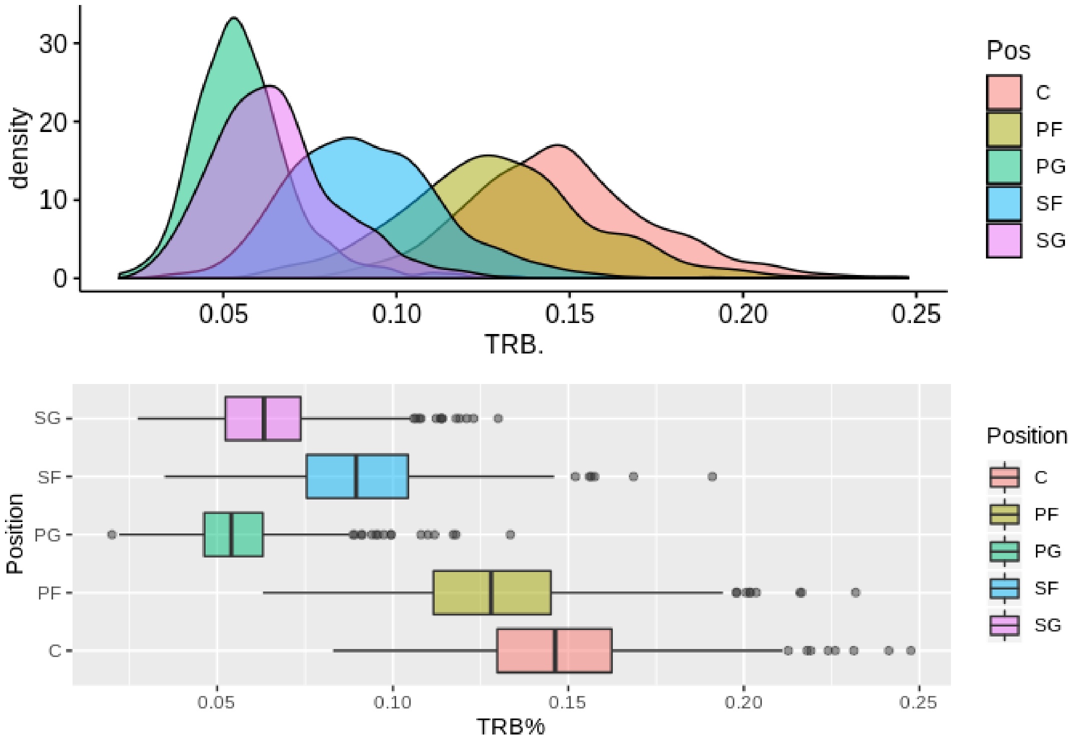
The conditional distributions are quite symmetric, they show a very slight positive asymmetry.

$$mean \gtrsim median \gtrsim mode$$

Explorative data analysis

FURTHER ANALYSIS

TRB%



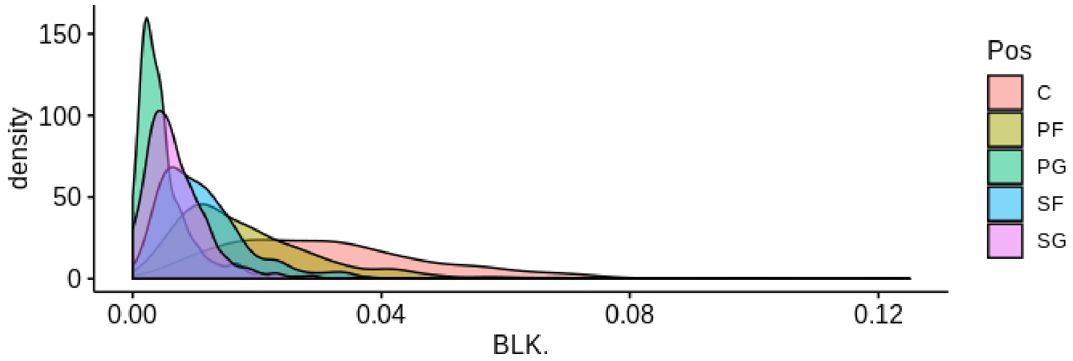
Analyzing mean and median we note that they are roughly equal for all conditioned distributions, however there is a significant presence of outliers above the maximum of the box-plot.

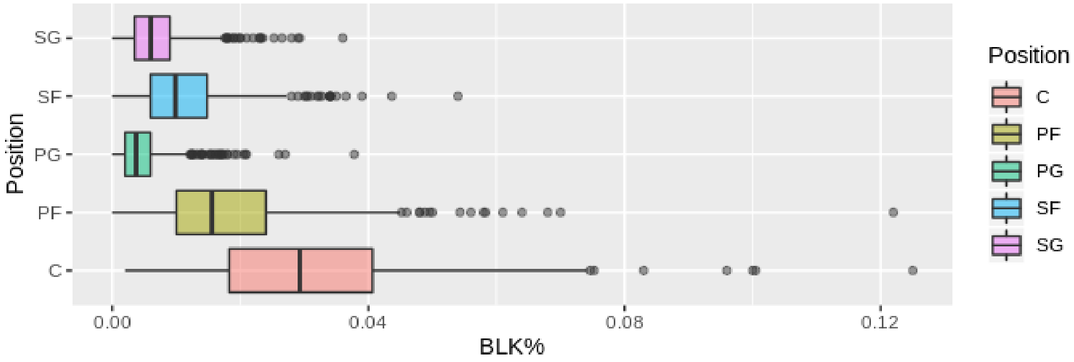
$$mean \gtrsim median$$

Explorative data analysis

FURTHER ANALYSIS

BLK%



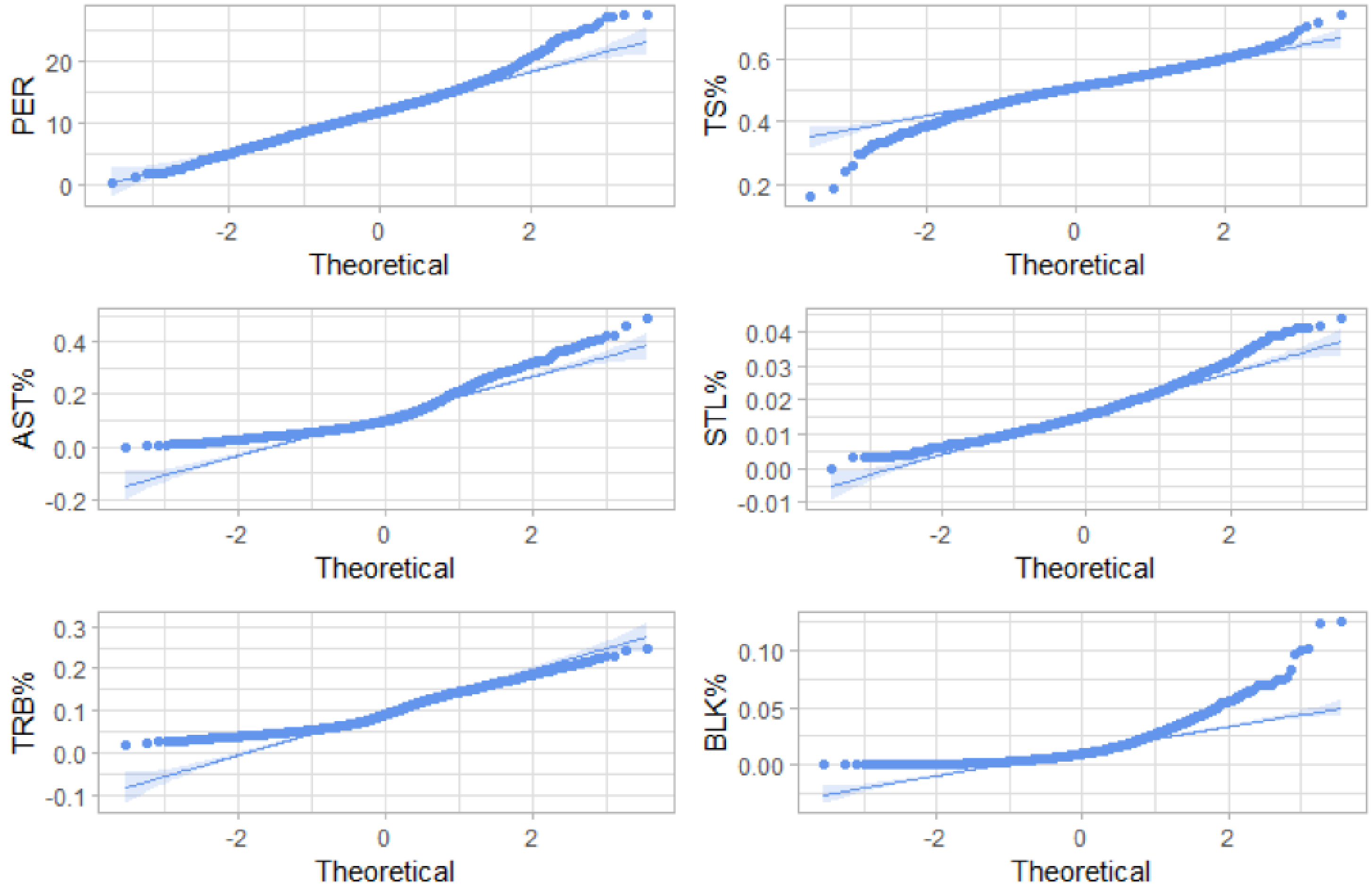


All the conditional distribution show a positive asymmetry, and skewness values noticeably different from zero. AS for TRB%, there is a significant presence of outliers above the maximum of the boxplot.

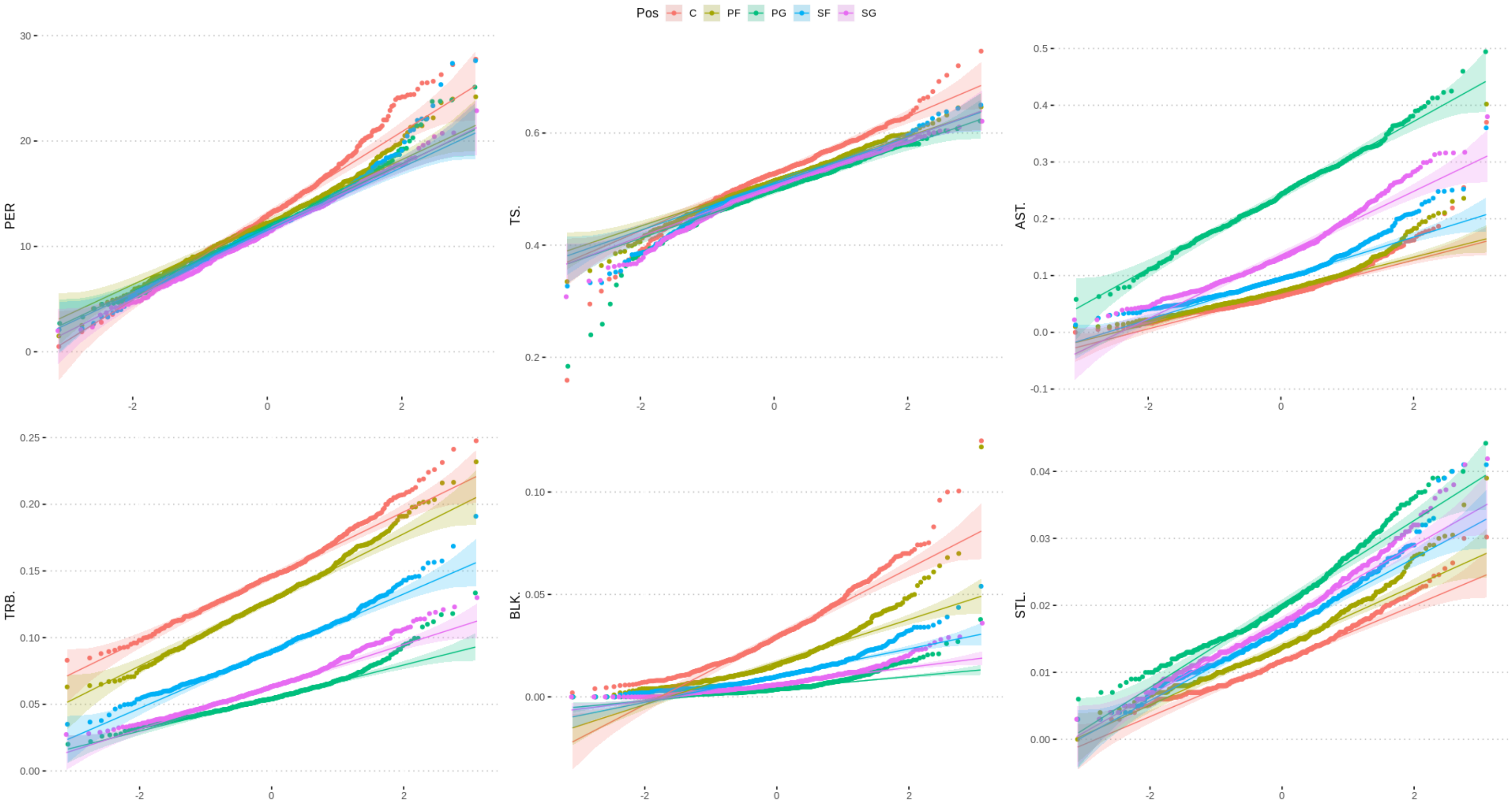
$$mean \gtrsim median \gtrsim mode$$

Explorative data analysis

Q-Q PLOTS



The Q-Q Plot is the graphical representation of the quantiles of a distribution. It compares the cumulative distribution of the observed variable with the cumulative distribution of the normal: if the observed variable has a normal distribution, the points of this joint distribution thicken on the diagonal.

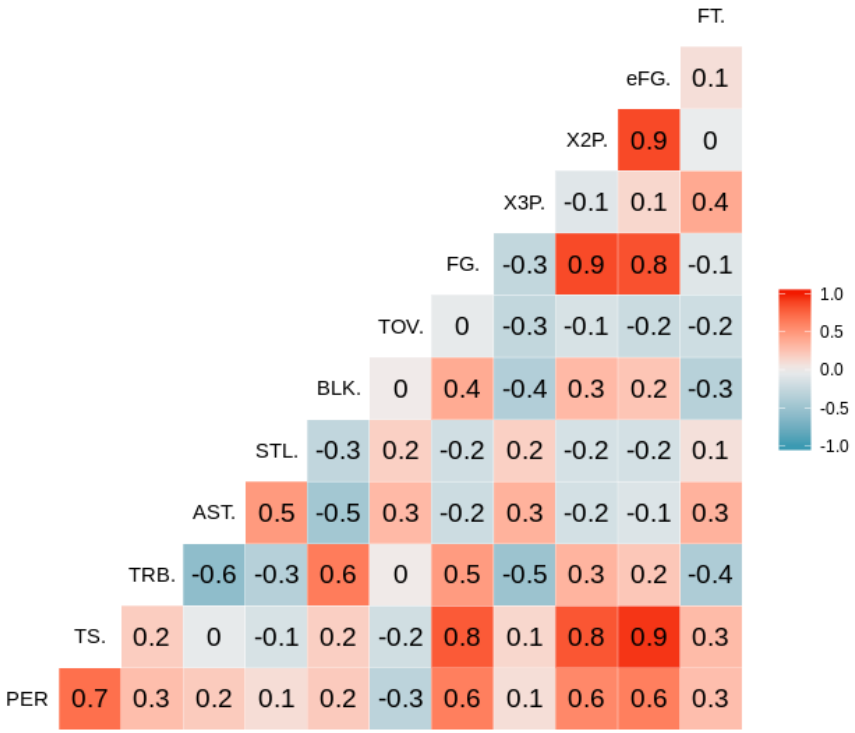


Explorative data analysis

Q-Q PLOTS

Explorative data analysis

Variables correlation

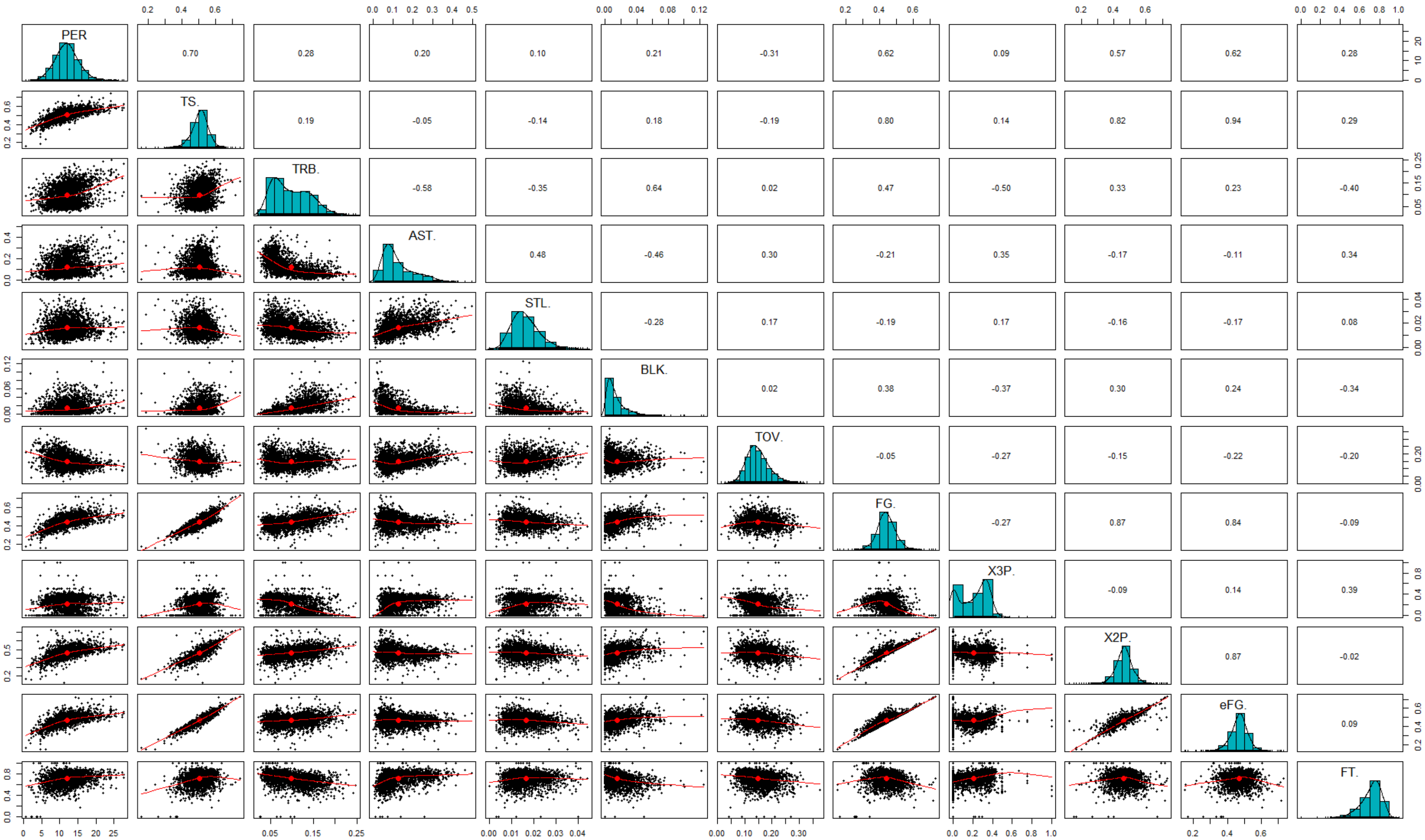


											FT.
										eFG.	0.1
									X2P.	0.9	0
								X3P.	-0.1	0.1	0.4
							FG.	-0.3	0.9	0.8	-0.1
						TOV.	0	-0.3	-0.1	-0.2	-0.2
					BLK.	0	0.4	-0.4	0.3	0.2	-0.3
				STL.	-0.3	0.2	-0.2	0.2	-0.2	-0.2	0.1
			AST.	0.5	-0.5	0.3	-0.2	0.3	-0.2	-0.1	0.3
		TRB.	-0.6	-0.3	0.6	0	0.5	-0.5	0.3	0.2	-0.4
	TS.	0.2	0	-0.1	0.2	-0.2	0.8	0.1	0.8	0.9	0.3
PER	0.7	0.3	0.2	0.1	0.2	-0.3	0.6	0.1	0.6	0.6	0.3

The correlation between PER and TS% is quite high (0.7), but it is very low with the other variables (TRB%, AST%, STL%, BLK%). This result will be used to set up a simple linear regression model that has as dependent variable PER and as explanatory variable TS%.

Explorative data analysis

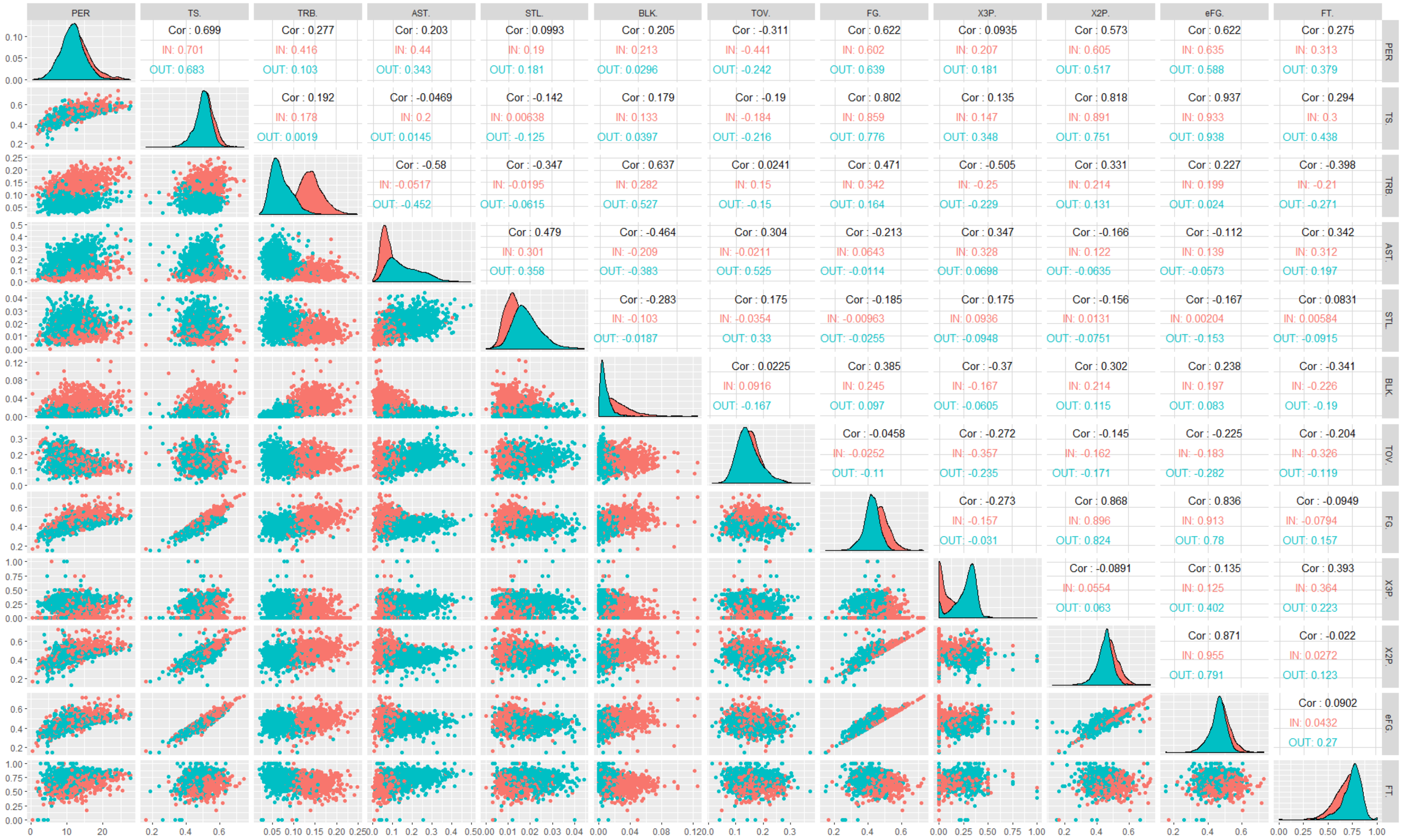
Variables correlation



Explorative data analysis

Variables correlation

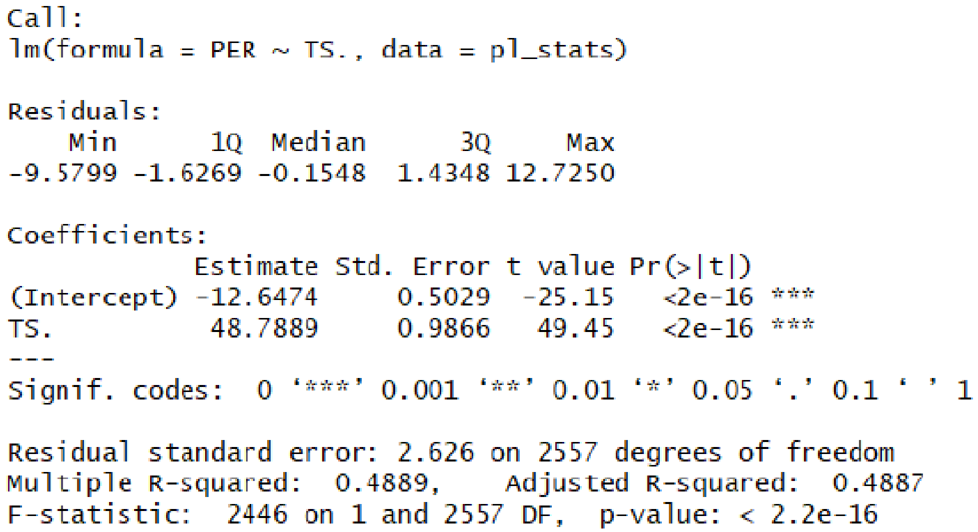
Furthermore, we will try to add the pairs (AST%, STL%) and (TRB%, BLK%), respectively, for the OUT and IN subsets (subsets defined taking as reference the 3-point line: PG, SG and SF are OUT, PF and C are IN); although these are not very correlated with PER, from the analysis of the box-plots we have seen that they separate the IN/OUT groups quite well, so we will try to create two separate models (one per each subset) and we will compare the results with those of the simple model.



Inferential data analysis

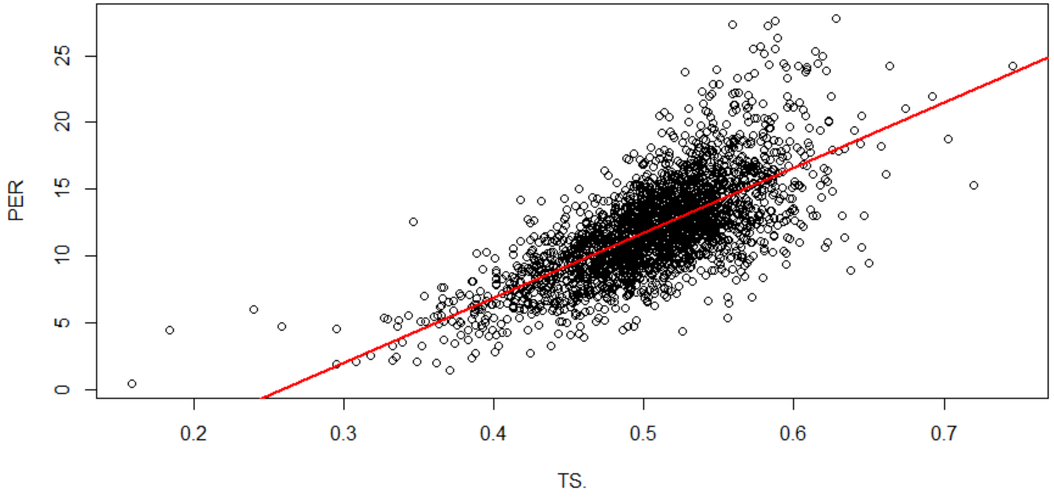
Linear Regression Models

- **simple LM1**: $PER \sim TS\%$



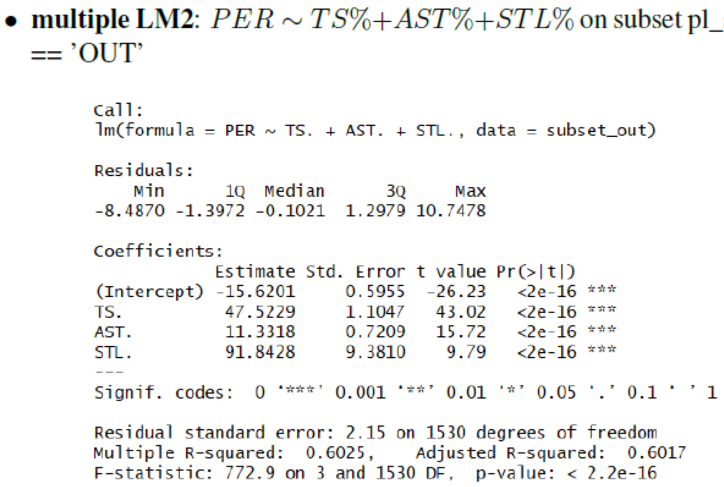
```
Call:
lm(formula = PER ~ TS., data = pl_stats)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5799 -1.6269 -0.1548  1.4348 12.7250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.6474     0.5029  -25.15   <2e-16 ***
TS.          48.7889     0.9866   49.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.626 on 2557 degrees of freedom
Multiple R-squared:  0.4889,	Adjusted R-squared:  0.4887 
F-statistic:  2446 on 1 and 2557 DF,  p-value: < 2.2e-16
```

$R^2{}_{adj} = 0.49$, quite low



Inferential data analysis

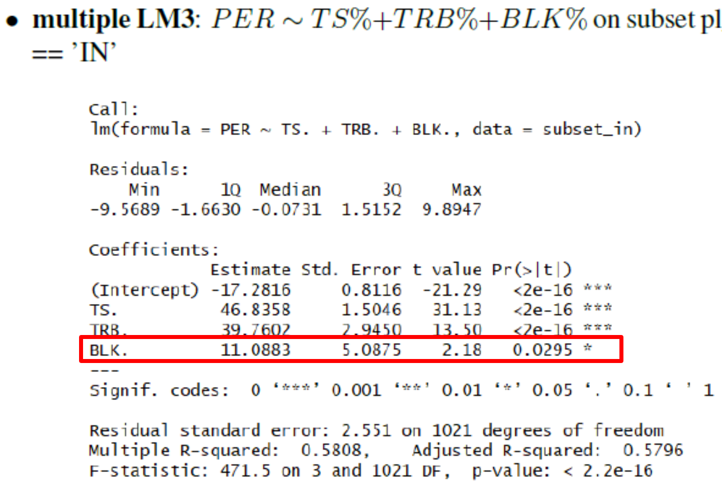
Linear Regression Models



- **multiple LM2**: $PER \sim TS\% + AST\% + STL\%$ on subset pl_stats_xpos$threepointline == ’OUT’

```
Call:
lm(formula = PER ~ TS. + AST. + STL., data = subset_out)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4870 -1.3972 -0.1021  1.2979 10.7478 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -15.6201     0.5955  -26.23   <2e-16 ***
TS.          47.5229     1.1047   43.02   <2e-16 ***
AST.         11.3318     0.7209   15.72   <2e-16 ***
STL.         91.8428     9.3810    9.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.15 on 1530 degrees of freedom
Multiple R-squared:  0.6025,	Adjusted R-squared:  0.6017 
F-statistic: 772.9 on 3 and 1530 DF,  p-value: < 2.2e-16
```

$$R^2{}_{adj} = 0.60$$



- **multiple LM3**: $PER \sim TS\% + TRB\% + BLK\%$ on subset pl_stats_xpos$threepointline == ’IN’

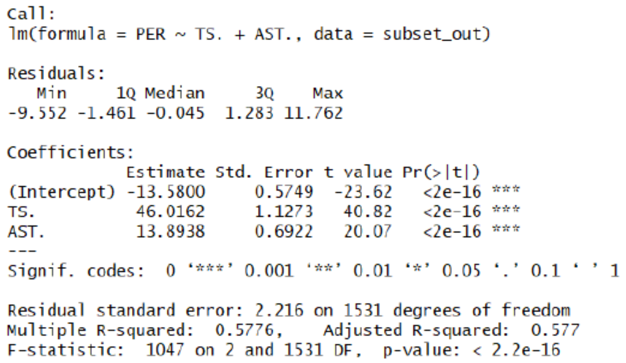
```
Call:
lm(formula = PER ~ TS. + TRB. + BLK., data = subset_in)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5689 -1.6630 -0.0731  1.5152  9.8947 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.2816     0.8116  -21.29   <2e-16 ***
TS.          46.8358     1.5046   31.13   <2e-16 ***
TRB.         39.7602     2.9450   13.50   <2e-16 ***
BLK.         11.0883     5.0875    2.18   0.0295 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.551 on 1021 degrees of freedom
Multiple R-squared:  0.5808,	Adjusted R-squared:  0.5796 
F-statistic: 471.5 on 3 and 1021 DF,  p-value: < 2.2e-16
```

$$R^2{}_{adj} = 0.58$$

Inferential data analysis

Linear Regression Models

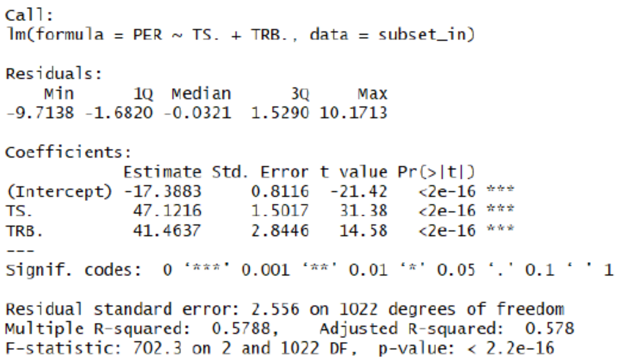
- **multiple LM4**: now we try to use only one other variable (and not two) with TS%, in particular we choose the one which is more correlated to PER -> in this case it is AST% : $PER \sim TS\% + AST\%$ on subset pl_stats_xpos$threepointline == 'OUT'



```
Call:
lm(formula = PER ~ TS. + AST., data = subset_out)

Residuals:
   Min     1Q Median     3Q    Max 
-9.552 -1.461 -0.045  1.283 11.762 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -13.5800     0.5749  -23.62   <2e-16 ***
TS.          46.0162     1.1273   40.82   <2e-16 ***
AST.         13.8938     0.6922   20.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.216 on 1531 degrees of freedom
Multiple R-squared:  0.5776,	Adjusted R-squared:  0.577 
F-statistic:  1047 on 2 and 1531 DF,  p-value: < 2.2e-16
```

$$\boxed{{R^2}_{adj} = 0.58}$$

- **multiple LM5**: now we try to use only one other variable (and not two) with TS%, in particular we choose the one which is more correlated to PER -> in this case it is TRB% : $PER \sim TS\% + TRB\%$ on subset pl_stats_xpos$threepointline == 'IN'



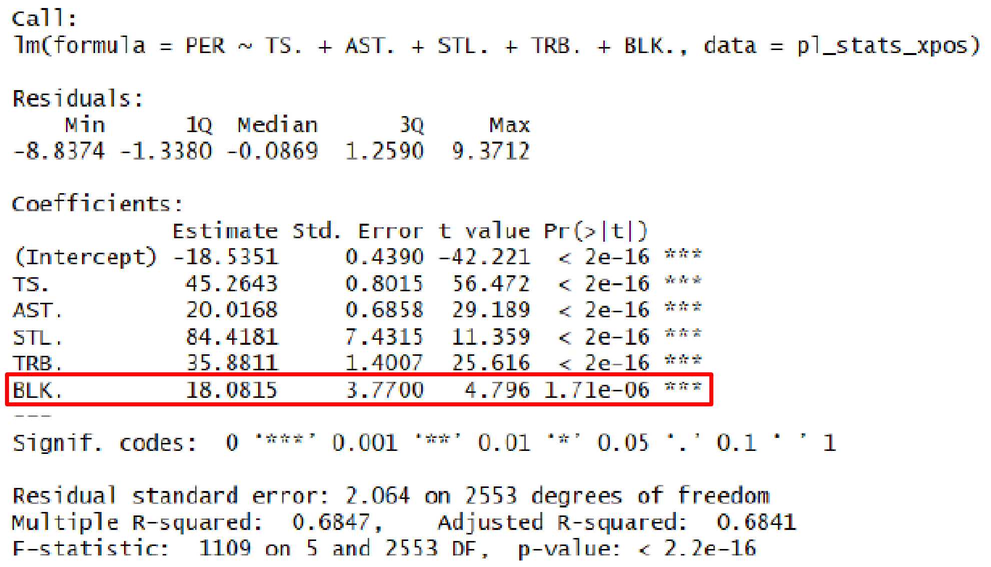
```
Call:
lm(formula = PER ~ TS. + TRB., data = subset_in)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.7138 -1.6820 -0.0321  1.5290 10.1713 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.3883     0.8116  -21.42   <2e-16 ***
TS.          47.1216     1.5017   31.38   <2e-16 ***
TRB.         41.4637     2.8446   14.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.556 on 1022 degrees of freedom
Multiple R-squared:  0.5788,	Adjusted R-squared:  0.578 
F-statistic: 702.3 on 2 and 1022 DF,  p-value: < 2.2e-16
```

$$\boxed{{R^2}_{adj} = 0.58}$$

Inferential data analysis

Linear Regression Models

- **multiple LM6**: $PER \sim TS\% + AST\% + STL\% + TRB\% + BLK\%$



```
Call:
lm(formula = PER ~ TS. + AST. + STL. + TRB. + BLK., data = pl_stats_xpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8374 -1.3380 -0.0869  1.2590  9.3712 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -18.5351     0.4390 -42.221  < 2e-16 ***
TS.          45.2643     0.8015  56.472  < 2e-16 ***
AST.         20.0168     0.6858  29.189  < 2e-16 ***
STL.         84.4181     7.4315  11.359  < 2e-16 ***
TRB.         35.8811     1.4007  25.616  < 2e-16 ***
BLK.         18.0815     3.7700   4.796 1.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.064 on 2553 degrees of freedom
Multiple R-squared:  0.6847,	Adjusted R-squared:  0.6841 
F-statistic:  1109 on 5 and 2553 DF,  p-value: < 2.2e-16
```

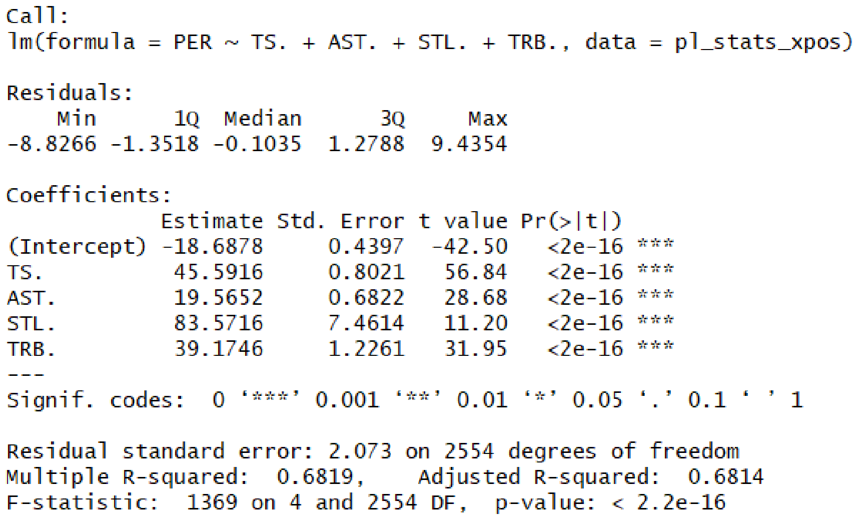
$$R^2_{adj} = 0.68$$

way better than *LM1*, but we used five variables, not one

Inferential data analysis

Linear Regression Models

We try to reduce the number of variables, excluding BLK% which shows a Pr(>|t|) greater than others, *LM7: PER ~ TS% + AST% + STL% + TRB%*



```
Call:
lm(formula = PER ~ TS. + AST. + STL. + TRB., data = pl_stats_xpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8266 -1.3518 -0.1035  1.2788  9.4354 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -18.6878     0.4397  -42.50   <2e-16 ***
TS.          45.5916     0.8021   56.84   <2e-16 ***
AST.         19.5652     0.6822   28.68   <2e-16 ***
STL.         83.5716     7.4614   11.20   <2e-16 ***
TRB.         39.1746     1.2261   31.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.073 on 2554 degrees of freedom
Multiple R-squared:  0.6819,    Adjusted R-squared:  0.6814 
F-statistic:  1369 on 4 and 2554 DF,  p-value: < 2.2e-16
```

$$R^2{}_{adj} = 0.68$$

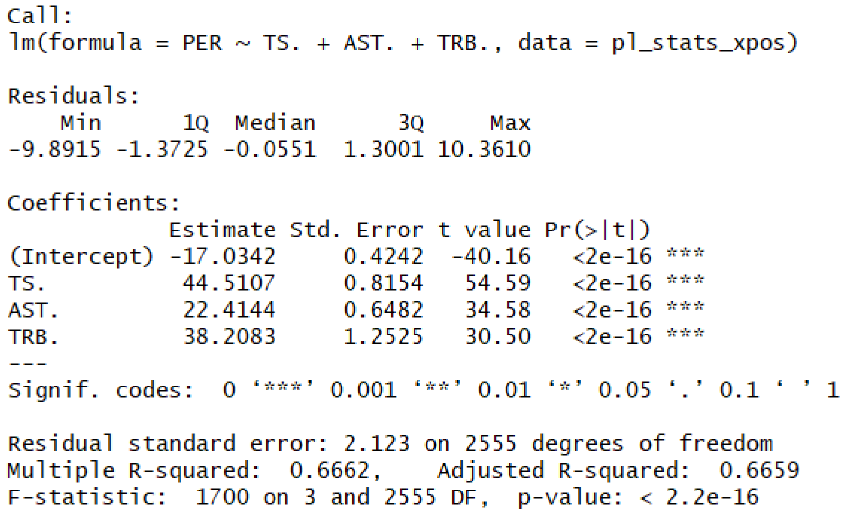
equal to *LM6*, but with one less variable

Inferential data analysis

Linear Regression Models

Finally, we cut out another variable, STL%, following this reasoning: we saw that BLK% does not have a relevant role in the regression model (R^2 did not change between *LM6* and *LM7*) and we know that in the pair (TRB%, BLK%) BLK% is the least correlated variable with PER, in a similar way we can remove STL% (least correlated variable with PER in the pair (AST%, STL%)):

LM8: PER ~ TS% + AST% + TRB%



```
Call:
lm(formula = PER ~ TS. + AST. + TRB., data = pl_stats_xpos)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.8915 -1.3725 -0.0551  1.3001 10.3610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.0342     0.4242  -40.16   <2e-16 ***
TS.          44.5107     0.8154   54.59   <2e-16 ***
AST.         22.4144     0.6482   34.58   <2e-16 ***
TRB.         38.2083     1.2525   30.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.123 on 2555 degrees of freedom
Multiple R-squared:  0.6662,	Adjusted R-squared:  0.6659 
F-statistic:  1700 on 3 and 2555 DF,  p-value: < 2.2e-16
```

$${R^2}_{adj} = 0.67$$

practically equal to *LM6* and *LM7*, but with only <u>three variables</u>

Conclusions

From the analysis carried out, and from previous research, it appears that *TS%* is the most important component in the attempt to predict the PER of a player.

The optimal model (based on the value of ${R^2}_{adj}$ reached) that we have built to determine the PER is the following:

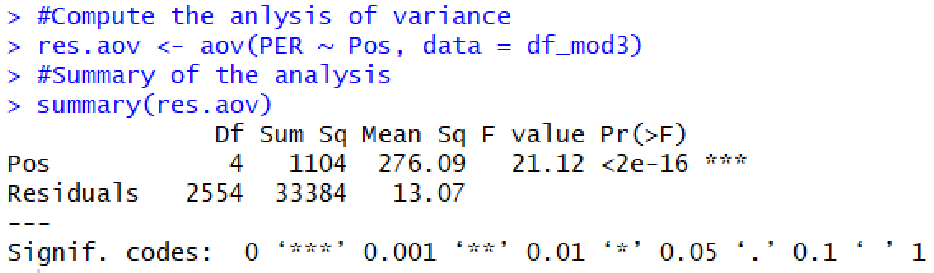
$$PER = -17.0 + 44.5 \cdot TS\% + 22.4 \cdot AST\% + 38.2 \cdot TRB\%$$

- Another result observed during the analysis is that there is a substantial difference from the point of view of the PER mean among players' positions. Indeed, we can investigate deeper the distribution of PER using **ANOVA** (*ANalysis Of VAriance*).

Conclusions

ANOVA

The mean is determined only by the population, so we consider a *one-way ANOVA*:



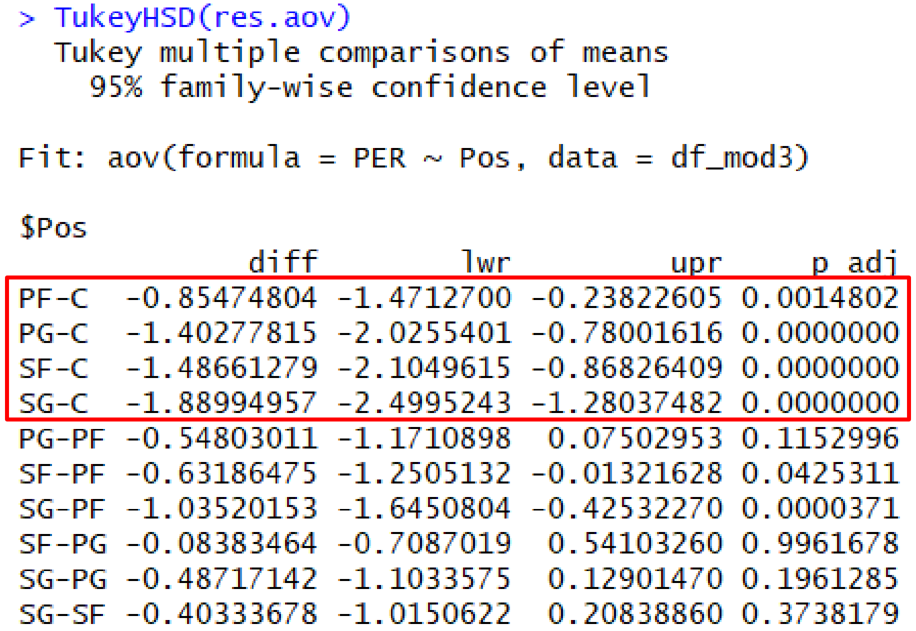
```
> #Compute the anlysis of variance
> res.aov <- aov(PER ~ Pos, data = df_mod3)
> #Summary of the analysis
> summary(res.aov)
              Df Sum Sq Mean Sq F value Pr(>F)    
Pos            4   1104  276.09   21.12 <2e-16 ***
Residuals   2554  33384   13.07                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It can be concluded, at a significance level of 5%, that: the p-value is less than 5% and we can refuse the null hypothesis that means are equal (at least two of our means differ).

Conclusions

ANOVA

Multiple comparisons: we proceed to carry out the simultaneous comparisons between all pairs of groups.



```
> TukeyHSD(res.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PER ~ Pos, data = df_mod3)

$Pos
             diff        lwr         upr     p adj
PF-C  -0.85474804 -1.4712700 -0.23822605 0.0014802
PG-C  -1.40277815 -2.0255401 -0.78001616 0.0000000
SF-C  -1.48661279 -2.1049615 -0.86826409 0.0000000
SG-C  -1.88994957 -2.4995243 -1.28037482 0.0000000
PG-PF -0.54803011 -1.1710898  0.07502953 0.1152996
SF-PF -0.63186475 -1.2505132 -0.01321628 0.0425311
SG-PF -1.03520153 -1.6450804 -0.42532270 0.0000371
SF-PG -0.08383464 -0.7087019  0.54103260 0.9961678
SG-PG -0.48717142 -1.1033575  0.12901470 0.1961285
SG-SF -0.40333678 -1.0150622  0.20838860 0.3738179
```

Basically, it is clear that the Center PER mean is significantly different from the one of any other group.

Future developments

It could be of interest, for future applications, to deepen the aspect of variable selection, maybe with Ridge or Lasso regression, and the construction of new variables starting from those available (Principal Component Analysis).

For example, it may be possible to go far beyond linearity using **GAMs**, *General Additive Models*.