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

Celestino Santagata, Antonio Nappa

August 3, 2020

Contents

Introduction

National Basketball Association (NBA) is a men's professional basketball league in North America, composed of 30 teams (29 in the United States and 1 in Canada). It is widely considered to be the premier men's professional basketball league in the world.
The five basketball positions normally employed by organized basketball teams are the point guard (PG), the shooting guard (SG), the small forward (SF), the power forward (PF), and the center (C) (Fig.1).

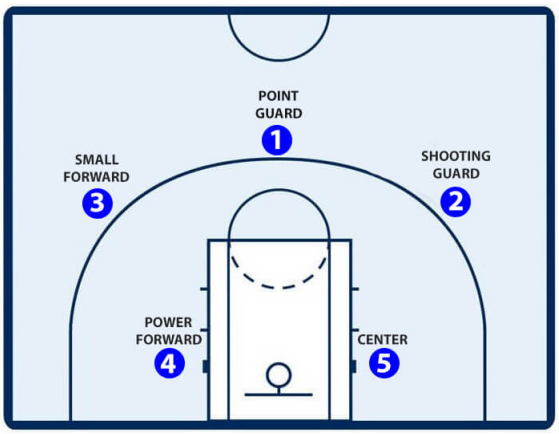


Figure 1: Scheme of the five basketball positions.

The point guard is the *de facto* leader of the team on the court. This position requires substantial ball-handling skills and the ability to facilitate the team during a play. The shooting guard, as the name implies, is often the best shooter. As well as being capable of shooting accurately from longer distances, this position tends to also be the best defender on the team. The small forward often has an aggressive approach to the basket when handling the ball. The small forward is also known to make cuts to the basket in efforts to get open for shots. The power forward and the center are usually called the "front-court", often acting as their team's primary rebounders or shot blockers, or receiving passes to take inside shots. The center is typically the taller of the two.

However, individual team strategy and availability of personnel can alter the positions used by a particular team. Besides the five basic positions, some teams use non-standard or hybrid positions, such as the point forward, a hybrid small forward/point guard; the swingman, a hybrid small forward/shooting guard; the big, a hybrid power forward/center; and the stretch four, a power forward with the shooting range of typical shooting guards.
As we have just said, each player has its own particular characteristics and, based on these skills, he finds his best suitable role.

How much do statistics matter in determining success in basketball?
It is very important to study the connection between sports, in this case basketball, and the use of advanced statistical analysis, how it has developed and how it can be used as a tool for both coaches, team managers to improve their teams, selection process, player development etc. Basketball was not the pioneer in this area, actually came into it quite late, but rather Baseball who has preoccupied the imagination of statisticians and mathematicians for decades and with the huge success of Sabermetrics approach to building teams in baseball (pioneered in Oakland Athletics, later adopted by majority of baseball franchises). Teams have had great success with the approach of using advanced statistical analysis to build teams in Baseball that suddenly teams with much lower budgets started performing at the high level (baseball has no salary cap, like all other professional sports in the USA). (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." (Forbes)
Statistics are a good way to tell how a player is doing in a sport and what he is best at, they can also help to determine where the player and his team need to improve; but it is equally important to learn how to make conscious use of it, not to let oneself be guided only by numbers.

Useful sites:

- https://stats.nba.com/
- https://www.espn.com/nba/stats
- https://www.basketball-reference.com/players/
- https://www.foxsports.com/nba/stats

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

Report Target

Possible targets: Which state has produced more NBA players? Which college is producing more NBA players? What makes a player successful? Who is the greatest player of all time (GOAT)? How 3point shooting is changing NBA? How much does a highly skilled player contribute to his team's results?
The game of basketball has changed quite a lot since the early days of the NBA, both in style as well rules and regulations. In the last decades players had to become more versatile. Big-men learned to shoot and dribble, and guards became stronger and more athletic.
In this context, the advanced statistical metric *PER* (Hollinger's Player Efficiency Rating) is usually used as a predictor of inclusion on the All NBA teams (The All-NBA Team is an annual NBA honor bestowed on the best players in the league following every NBA season.). Our goal in this report is to analyze the PER and try to create a model that simplifies it, while predicting the same results. A very difficult task, since it is not so easy to determine the main variables which hold most of the information.

- unadjusted PER:

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

with

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

- adjusted PER:

$$PER = \left(uPER \times \frac{lgPace}{tmPace}\right) \times \frac{15}{lguPER} \tag{2}$$

Chapter 1

Dataset description

The dataset initially presents 26063 observations and 49 variables. Most of these variables were not useful for the purposes of the analysis because they were redundant or used for the calculation of some more useful indices which we effectively used for the analysis. Therefore the first step we took was to clean up the dataset with the removal of unnecessary variables. The final dataset will have the 17 variables listed in the table1.1.

1.1 Filtering

Before we explore the data, we reduce the number of years for the analysis to 1980-2019. Basketball records 1950s/1960s/1970s hold little information, indeed most of the variables were introduced starting from 1975 (the latest is 3P and has been available since 1980), so we proceed with the removal of records prior to 1980.
The regular season consists of each team playing 82 games, 41 at home and 41 away. We have players in our dataset who have only played a handful of games in a season and their stats introduce extreme variability into advanced metrics (such as PER, which sits within a -7 to 32 range for players who have played more than 10 games for an average of at least 5 minutes per game, but expands to -90 to 129 for all players). These players will usually have accumulated stats in garbage time (i.e. the final couple of minutes of a blow-out game). These outliers are not helpful in predicting who will make the GOAT, introducing more volatility and reducing the predictive power of a stat like PER. On that basis, we are setting a 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)) for inclusion in our analysis dataset: probably no player with so few games has left an indelible trace

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

Table 1.1

in the NBA.
We noticed there might be some double-counting of player statistics. This error has been introduced by the way data was recorded for seasons in which a player was traded. Their statistics are counted for each team, and then added together for the "mystery" team TOT (stands for total). Since we have used the data by season and filtered the data taking only the records from 1980 onwards, we have set up a function such that: set year and player, if there is 'TOT', we take only that record otherwise, if there isn't 'TOT', it means that the team doesn't change during that season.

1.2 Merging

There are multiple records referring to the same player: they differ for the season and the team because a player can change team during a given season. So, we merge these records by summing the absolute variables (*G*, *MP*, *PTS*) and doing the average of the others.
Filtering and merging operations were done through the python script *dataset_modifier.py*. This script provides the definitive dataset with 2842 observations and 17 variables; exploratory analysis can be performed on it.

1.3 Missing values

There are no missing data in the dataset, as can be verified by observing the graph in the figure 1.1, where the missing data are indicated on the abscissa axis while the variables are indicated on the ordinate axis.

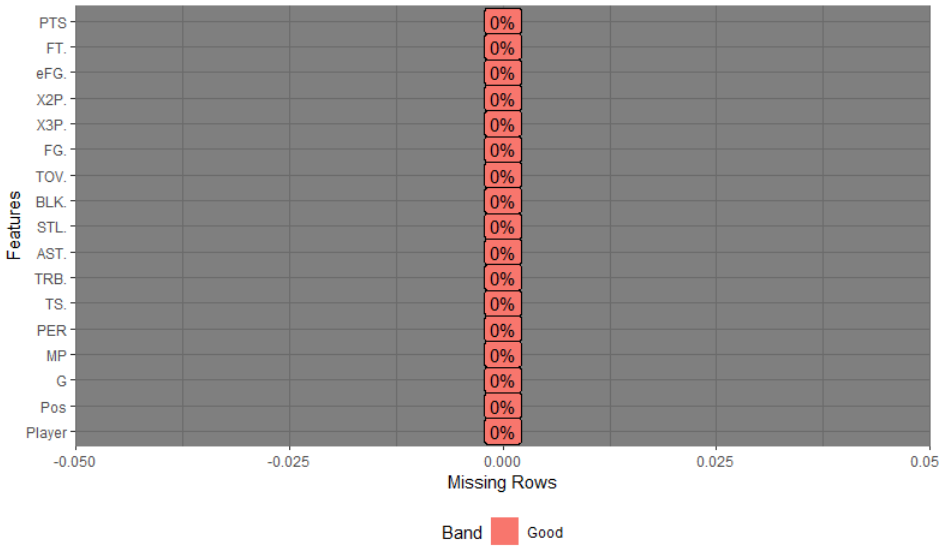


Figure 1.1: Plot of the missing data per feature.

Chapter 2

Explorative data analysis

2.1 Summary

```
      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
```

Figure 2.1: Main statistical indices for the quantitative variables of the data set.

Regarding to the categorical variable ‘Pos’, a separate analysis will be made and it will be used as a variable of comparison for technical characteristics among the players. The feature ‘Pos’ has the following values with respect to each of its modes:

C	PF	PG	SF	SG
513	512	492	506	536

The distribution of the players in relation to the position is almost identical.

'PER' is player efficiency rating, a widely-used summary assessment of a player's output; it is a dimensionless number that typically ranges from 0 to 35 and in our dataset is between 0.50 and 27.77, with mean 12.09 and median 11.94.
TS%, TRB%, AST%, STL%, BLK%, TOV%, FG%, X3P%, X2P%, eFG%, FT% are all measures in percent. TS%, FG%, eFG%, X3P%, X2P%, FT% are variables related to the percentage of realization; in particular, X3P%, X2P% and FT% refer to the different types of shooting. TS%, FG% and eFG% are "summary" variables that take into account the types of shooting and other parameters such as FGA (Field Goal Attempts); among them, we believe that TS% is the most explanatory in this sense:

$$TS\% = \frac{PTS}{2(FGA + (0.44\ FTA))} * 100 \tag{2.1}$$

where FTA is Free Throw Attempts[1].
TS% has a variation range equal to 0.587 with a minimum of 0.159 and a maximum of 0.746; the mean and the median are equal, respectively, to 0.507 and 0.511.
TOV% (Turnovers percentage) ranges from 0.014 to 0.374. Mean and median are almost coincident and respectively equal to 0.149 and 0.143.
AST% has a minimum of 0.000 and a maximum of 0.494, with mean 0.126 and median 0.099.
The STL% range is very narrow, it goes from 0.000 to 0.044; mean and median coincide if approximated to the third decimal place, they are both equal to 0.016.
BLK% is between 0.000 and 0.125, with mean 0.015 and median 0.010.
Finally, TRB% ranges from 0.020 to 0.248, its mean and median are respectively equal to 0.098 and 0.091. In this case it can be observed that the mean and the median deviate greatly from the maximum of the distribution. Also observing that the third quartile is equal to 0.130, the variable in question is expected to present anomalous values.
Following this description, we will observe the box-plots relative to these variables in order to easily view the characteristics of the distributions and to verify whether or not outliers are still present (even after the filtering phase) and, if so, to proceed with their possible treatment.
The analysis will focus on finding the features that best distinguish players' positions and, based on them, obtain a good model for PER.

[1]The coefficient 0.44 is related to ratio in FTAs from 2 point or 3 point possessions in NBA and and-1 attempts. If all FTAs are from 2 point shooting foul possessions, the coefficient should be 0.5, 0.333 for all 3 point shooting foul possessions, and 0 for all and-1 possessions. For different leagues or in different era, the FTAs ratio is different. It is questionable to use same true shooting percentage formula on those leagues unless the coefficient 0.44 is being adjusted. This raises another question that the true shooting percentage is not comparable for different leagues, age, gender, etc, even for different era.

2.2 Frequency distributions

2.2.1 Box-plot

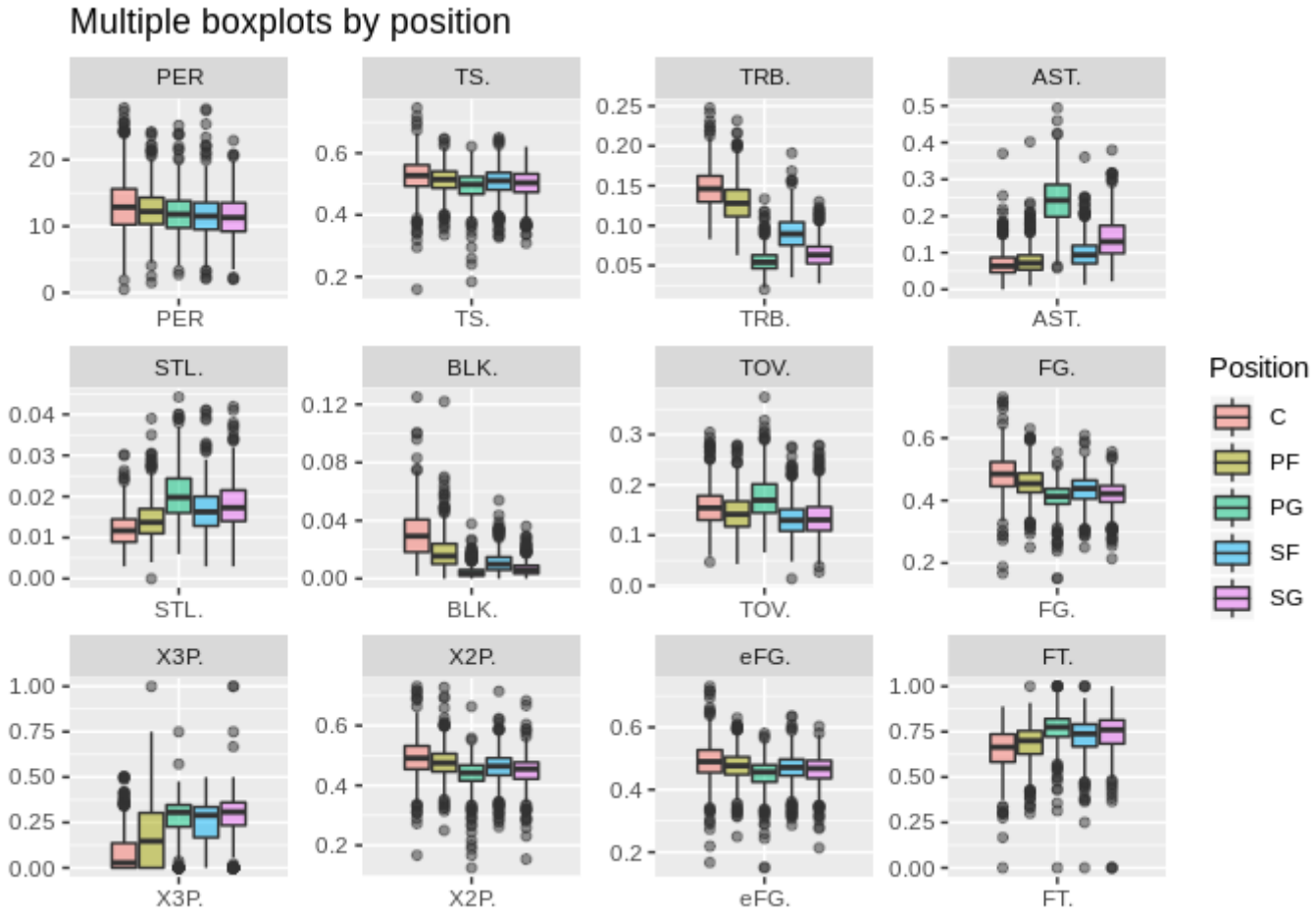


Figure 2.2: Multiple box-plots per position.

Let us focus for a moment on 'PER' (Fig.2.3).
Mean values per position are:
C 13.21972, PF 12.36497, PG 11.81694, SF 11.73310, SG 11.32977.
Although we expect there is no difference among positions, it seems that Centers are more performing (on average) than the others. How do we explain that?
Perhaps Centers are just better players than the others, or we can refer to one of the criticisms made against PER: its tendency to reward inefficient shooting. Considering the value attributed to the different types of shooting, the more a player shoots the higher is his PER value; so a player can be an inefficient scorer and simply inflate his value by taking a large number of shots. This typically happens for players inside the 3-point line zone, who tend to conclude as soon as they get the chance.

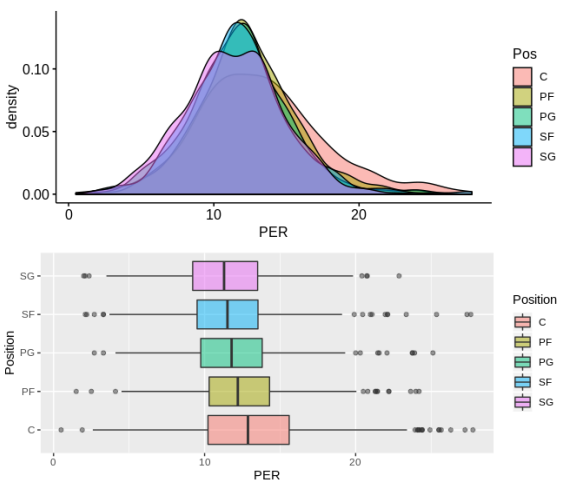


Figure 2.3: Box-plot and histogram of 'PER'.

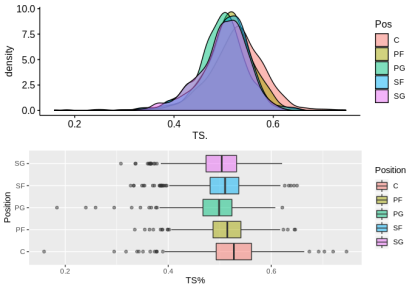
TRB%: we see that C and PF prevail (on average), then there is SF (as it plays close to the 3-point line), while SG and PG are the lowest as they are usually those furthest from the basket.
AST%: PG obviously dominates, because he is the ball carrier and sets the action to send his teammates to the basket; 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%).
TOV%: it is (on average) higher for the ball carrier (PG) and the scorers under the basket (C and PF) because, managing the ball in riskier situations, they are more likely to lose possession of it.
The distributions of TS% and eFG% are similar and seem 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. 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.

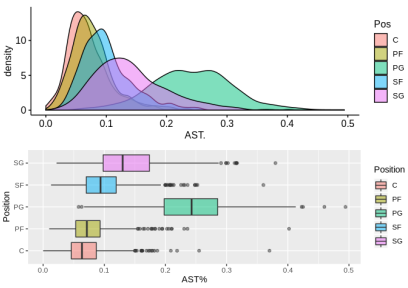
Further analysis

To conclude the descriptive analysis, we deepen the study of these variables distributions.

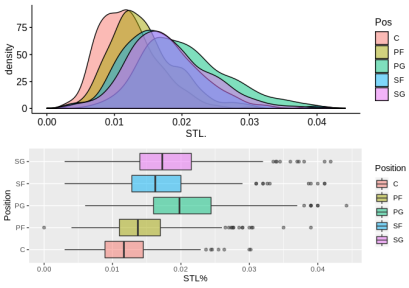
TS%: by analyzing the conditional distributions, it is observed that they are all approximately symmetric (this is confirmed by the fact that $mean \lesssim median$), with PG presenting a skewness in module greater than 1 due to the presence of many outliers under the minimum of the box-plot (calculated using the interquartile difference multiplied by 1.5).



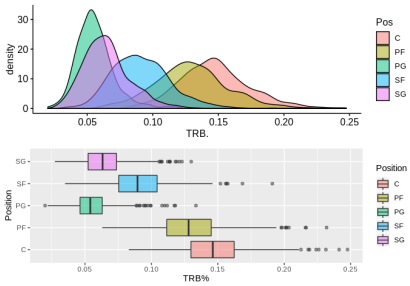
AST%: the conditional distributions are asymmetric (positive asymmetry), indeed they have $mean > median > mode$ and skewness values significantly different from zero, except for PG distribution, which seems quite symmetric (it seems to be bimodal).



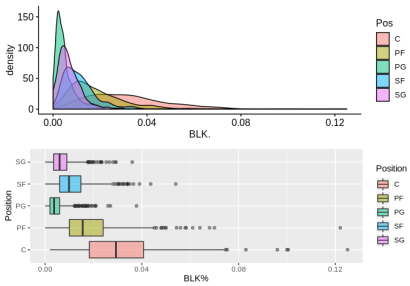
STL%: the conditional distributions are quite symmetric, they show a very slight positive asymmetry ($mean \gtrsim median \gtrsim mode$).



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 (over $Q_3 + IQR * 1.5$) ($mean \gtrsim median$).



BLK%: all the conditional distribution show a positive asymmetry, $mean \gtrsim median \gtrsim mode$ and skewness values noticeably different from zero. AS for TRB%, there is a significant presence of outliers above the maximum of the box-plot.



2.2.2 Q-Q plot

What has been said previously about variables distibutions can be seen also in the following Q-Q plots, fig.2.4-fig.2.5.

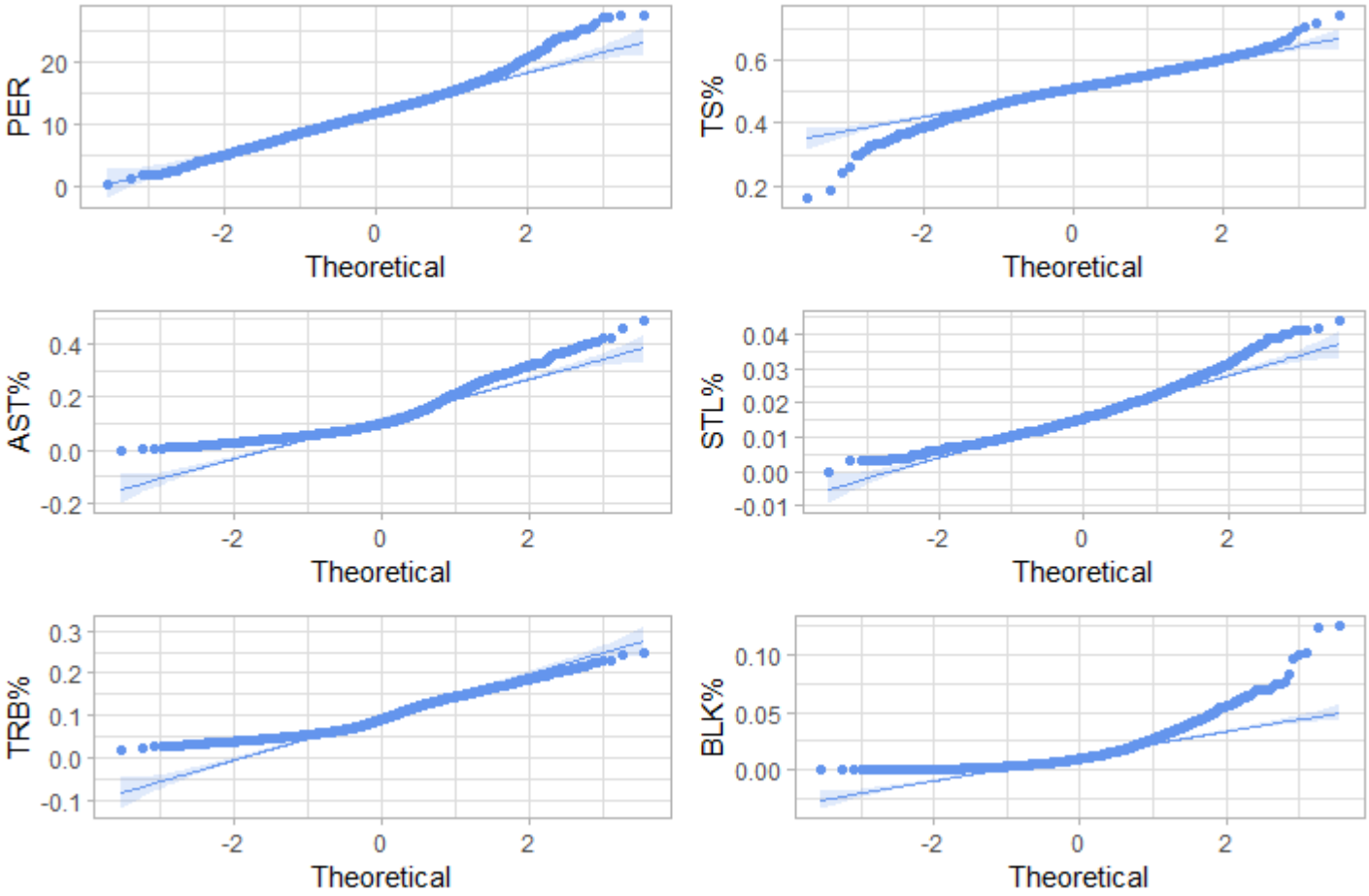


Figure 2.4: Multiple Q-Q plots.

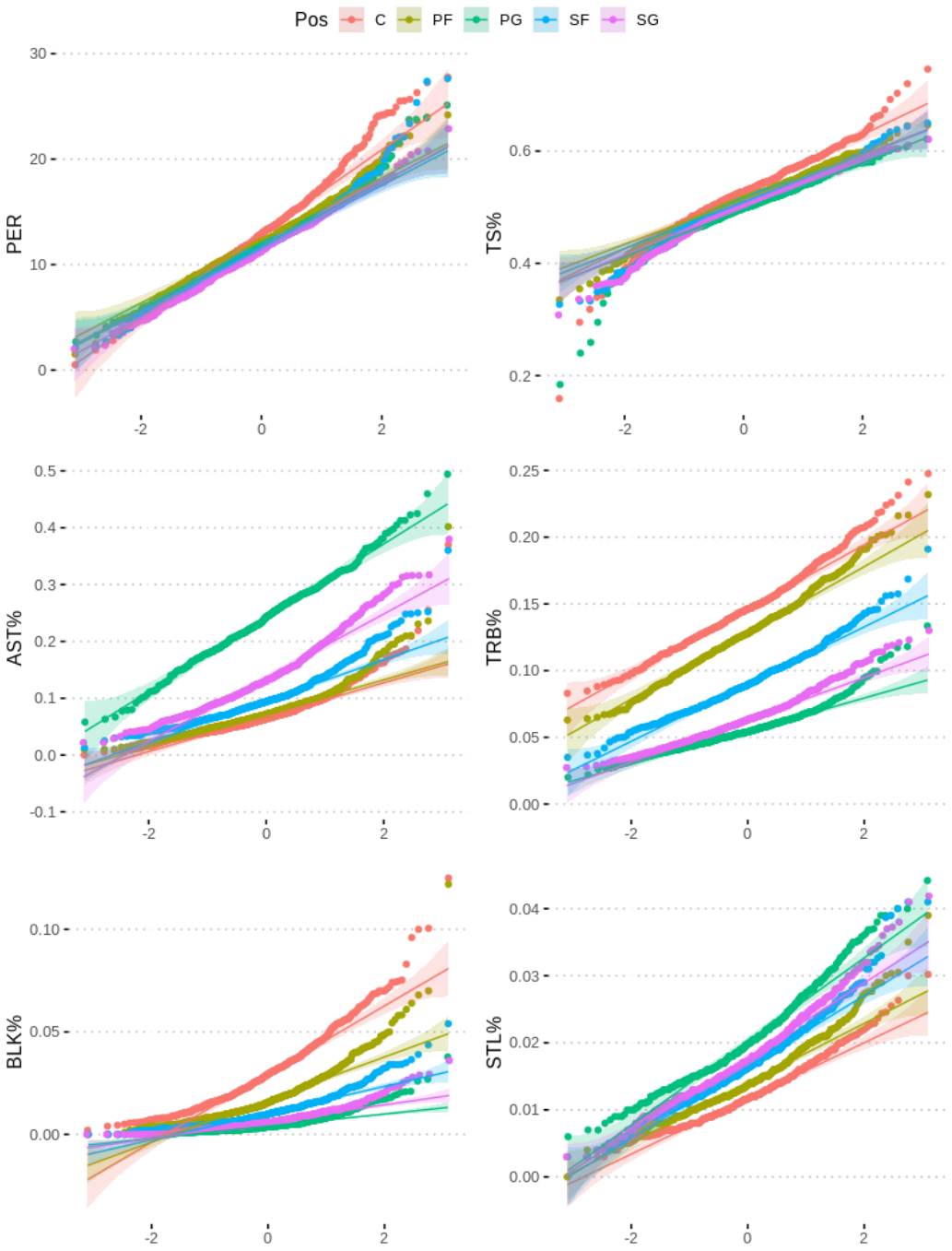


Figure 2.5: Multiple Q-Q plots per position.

2.2.3 Variables correlation

Now we focus the attention on the correlation among variables (Fig.2.6).
Based on the previous discussion, we take PER as a reference and see how it correlates with the other variables. We also said that among the variables referring to shooting/scoring statistics we choose TS% (the other candidate was eFG%, however it is clear that these two variables are highly correlated (0.9)).
The correlation between PER and TS% is quite high (0.7), but it is very low with the other variables (TRB%, AST%, STL%, BLK%, TOV%). This result will be used to set up a linear regression model that has as dependent variable PER and as explanatory variable TS%.
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.

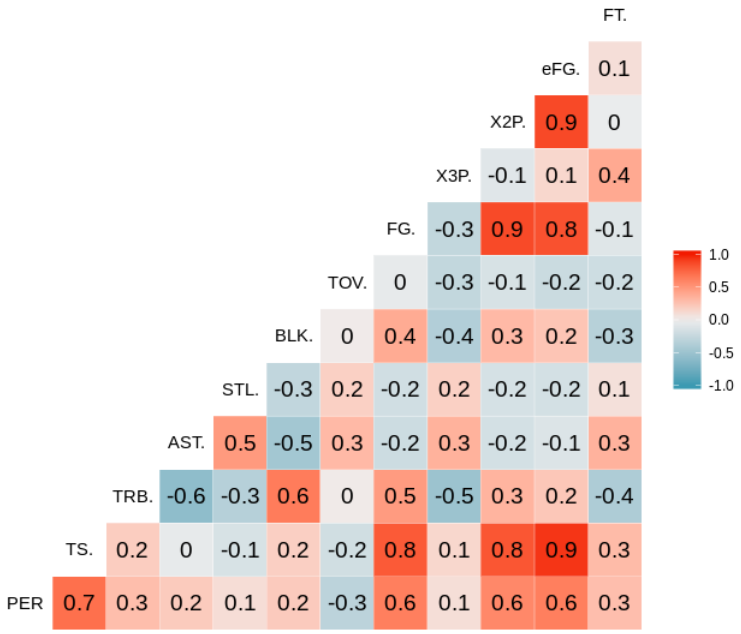


Figure 2.6: Correlation among variables.

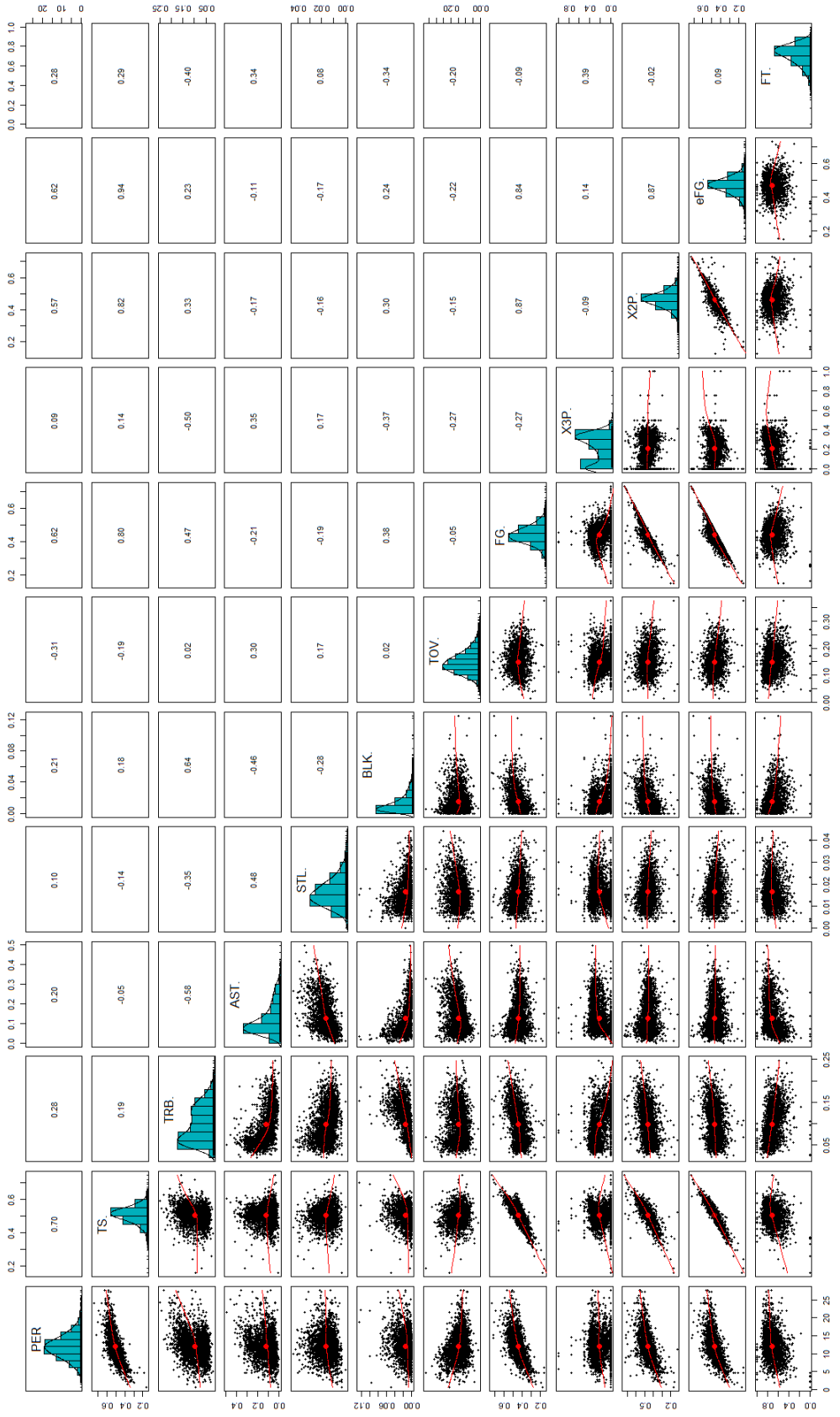


Figure 2.7: Correlation among variables.

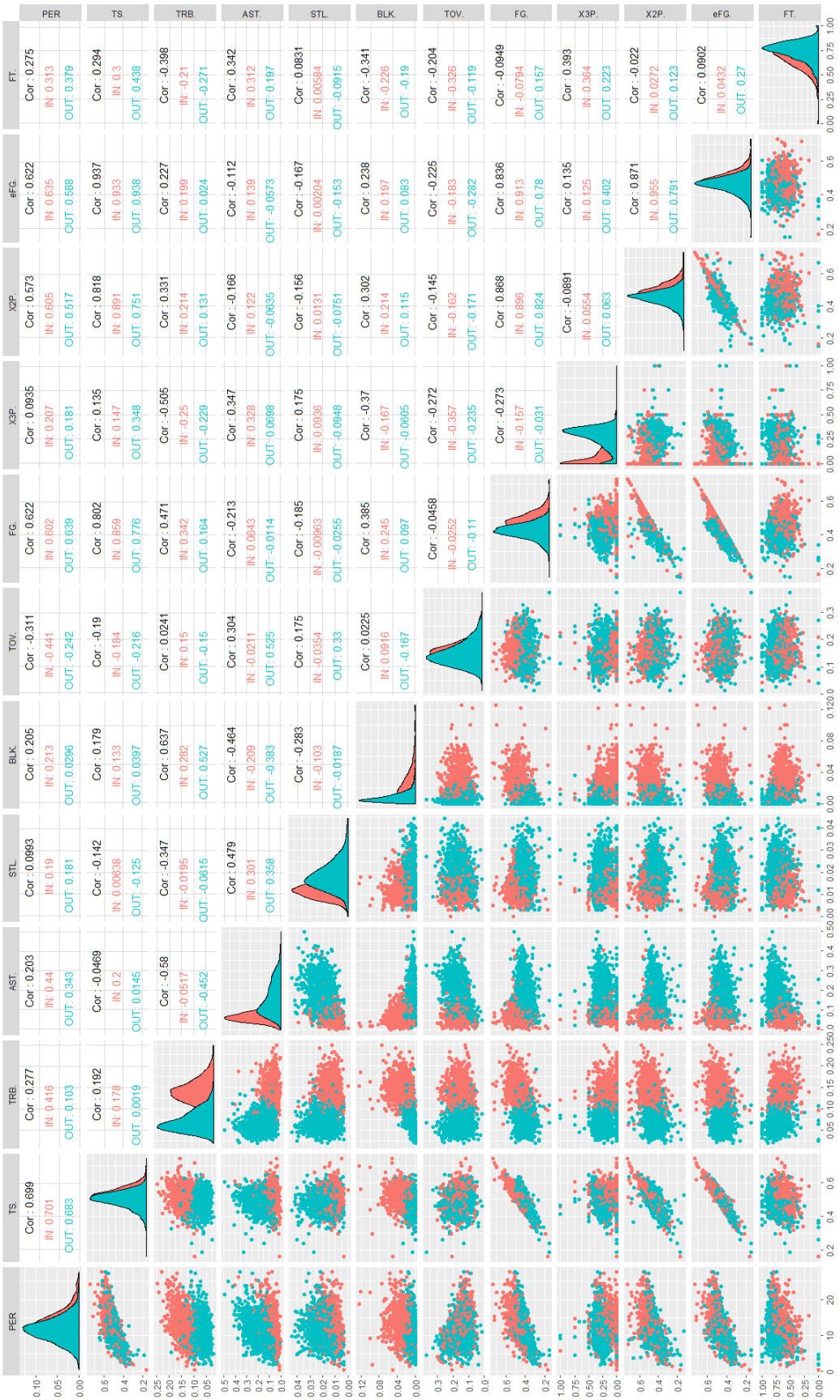


Figure 2.8: Correlation among variables with IN/OUT subsets.

Chapter 3

Inferential data analysis

3.1 Linear Regression Model

Our last step consists in building a linear regression model which manages to describe the PER index, following the path that has been illustrated before.

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

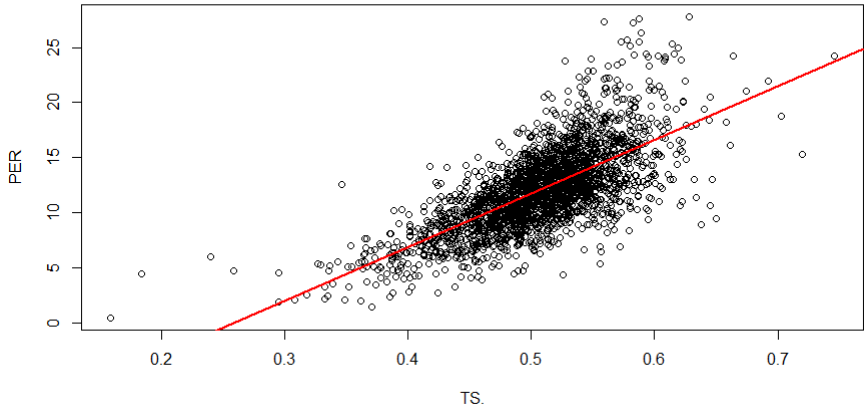
```
Call:
lm(formula = PER ~ TS., data = pl_stats)

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
```

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.49$ quite low



- **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
```

  t values OK, sign. lvl 0.1%
  $R^2_{adj} = 0.60$

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

  t values OK, sign. lvl 0.1% except for BLK%, whose sign. lvl is 5%
  $R^2_{adj} = 0.58$

```
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
```

- **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
```

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.58$ a bit lower than *LM2*

- **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'

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.58$ equal to *LM3* (even if we used one less variable)

```
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
```

- **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
```

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.68$ way better than *LM1*, but we used five variables, not one

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

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.68$ equal to *LM6*, but with one less variable

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

```
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
```

we can remove STL% (least correlated variable with PER in the pair(AST%, STL%)) and see what happens. LM8: $PER \sim TS\% + AST\% + TRB\%$

```
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
```

t values OK, sign. lvl 0.1%
$R^2_{adj} = 0.67$ practically equal to *LM6* and *LM7*, but with only three variables

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:

$$PER = -17.0 + 44.5 \cdot TS\% + 22.4 \cdot AST\% + 38.2 \cdot TRB\% \tag{3.1}$$

The validity of the estimates is related to the chosen variables and the lack of others variables that could add information and improve the prediction.

ANOVA

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).
ANOVA assumes indipendence and normality of groups and homogeneity of variance (homoscedasticity). It is used to verify the null hypothesis of equality of *m* means of as many populations discriminated on the basis of a factor *A* that can be assimilated to a qualitative variable. In our case, we want to verify if PER mean value has significant differences based on players' position.
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).
Since there is a significant difference that leads to reject the null hypothesis, which are the positions which push to reject the hypothesis?
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.
The ANOVA shows that Center PER mean is significantly different from the one of any other group.

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