UNSUPERVISED PROJECT

NBA Clustering

Guido Giacomo Mussini

988273

Abstract

The aim of this paper is to identify different types of NBA players based on their statistics collected during the regular season 2021/2022.The dimensions along which players can be defined are two: ability and characteristics.It has been tried to give priority to the characteristics, in order to classify the players on their role, but the results suggest a strong influence in the clusters of players' abilities. It have been run 2 different unsupervised algorithms: K-means and Hierarchical, which have provided similar results. This fact allowed to compare the findings intra-clusters and inter-clusters. In order to increase the interpretability of graphs, it has been executed also the principal component analysis.

Contents

1 Introduction, Expectations and Findings

Nba basketball is one of the most statistically analyzed sport in the world, since it is very suitable for collecting data. Over the years, analysts have been able to build increasingly complex indices. The dataset used in this paper is a collection of the most famous and interpretable statistics used to define the performances of a player.

The first step is to identify what these statistics can actually measure. There are aspects of the game which are complex to convert in numbers, like motivation, mental strength or leadership.
In general, the statistics can define a player among 2 principal dimensions:

- **Ability**: the strength of the player, how good he is. This dimension is strictly correlated with the time that each player play each games. If a player plays a lot, it means that he is useful for the team. More a player plays, more statistics and numbers he will collect. Follow that the statistics not defined as percentage will be higher for the players who play more. In this work, such statistics will be called **absolute statistics**.
  For more interpretability, the players will be divided on their ability in:
  - *Role Players*: players who play few minutes per game and contribute less to the results.
  - *First strings*: players who play more than the role ones, but they are mostly useful as support of the stars.
  - *Stars*: the most important players in a team, the ones for whose the team is build around.
- **Characteristic**: The role of the player, basically what he does when he plays. The will is to classify players over this dimension.
  There are 5 main roles in the basketball: point guard, shooting guard, small forward, power forward and Center. In the modern NBA these roles are less defined than in the past. In short:
  - *Point Guard- PG*: the engine of the team, has high passing and ball handling skills and often good long range shooting. In general they are shorter than other players, but in the modern NBA this is not necessary true.
    A famous example of PG: John Stockton.
  - *Shooting Guard -SG*: the first offensive weapon. Shooting guards have to score points, they are typically good in all the offensive skills: long range shooting, crossovers, free throws, lay-ups. They

can usually even play as point guards, or small forwards.
A famous example of SG: Michael Jordan.

- *Small Forwards -SF*: the the most versatile of the main five basketball positions. They are usually taller than guards, but they are often interchangeable. They tend to attack the paint and draw fouls, but they even have a good long range scoring ability. If they are particularly tall and strong, they can play as Power Forward too.
  A famous example of SF: Lebron James.
- *Power Forward -PF*: They match the ability to score close to the basket, with a good mid-range shooting. Depending on the height, they can be more similar to a center or a small forward.
  A famous example of PF: Tim Duncan.
- *Centers -C*: the tallest and biggest players on the court. They usually play near to the basket, in the paint. They are able to take rebounds and contesting shots. In the latest years, they generally have expanded their shooting range. A famous example of C: Shaquille O'Neal

Since most of the players can play different roles, it would be a great results even to cluster the players based on a more simplified structure: Guards, Forwards and Centers.

The most unexpected findings in this paper was to notice that the height is not so important to cluster the players. As saw above, it seems to be one of the most important variable to define the role, but the results of the analysis will show that is not crucial.
Another interesting result is the importance of the variables concerning the rebounds. These variables are able to divide clearly the players in the way the height was suppose to do.
The last finding, the most important, is that ability wins over characteristic. The data are more able to divide a Star from a role player, rather then a point guard from a center. In fact, there will be observed which only 1 cluster over 3 is more defined by characteristics, the rebounds and blocks, than the ability.

2 Dataset

The Dataset is composed by 605 observation, one for each player member of the league, and 34 variables, which contain information about the NBA 2021-2022 regular season.
All the variables are numeric, except for *player*, which is a character

2.1 Variables

- **player**: the name of the NBA player
- **age**: the age of the NBA player
- **games**: number of matches played during the regular season
- **min**: average number of minutes played per game
- **weight**: Player's weight in kg
- **height**: Player's height in m
- **points**: points scored played per game
- **a_field_goals**: attempted field goals per game.
  A field goal is a basket scored on any shot or tap other than a free throw.
- **p_field_goals**: percentage of scored field goals over the attempted ones, per game
- **p_efg**: percentage of the effective field goals. p_efg is calculated as:

$$p\_efg = \frac{field\_goals + 0.5 * (3pointshoots)}{a\_field\_goals}$$

- **p_2points**: percentage of 2-points-shots scored per game
- **a_3points**: attempted 3-points-shots per game
- **p_3points**: percentage of 2-points-shots scored per game
- **prop_2p**: percentage of 3 points shots over total shots attempted per game
- **prop_3p**: percentage of 2 points shots over total shots attempted per game
- **a_free_throws**: attempted free throws per game

- **p_free_throws**: percentage of scored free throws over attempted, per game
- **p_true_shooting**: percentage of true shooting: true shooting is defined as:

$$p\_true\_shooting = \frac{0.5 * points}{a\_field\_goals + 0.475 * a\_free\_throws}$$

- **oreb**: offensive rebounds per game
- **dreb**: defensive rebounds per game
- **reb**: total rebounds per game
- **p_oreb**: estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor.
- **p_dreb**: estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor.
- **p_reb**: estimate of the percentage of available rebounds a player grabbed while he was on the floor.
- **assist**: number of assists per game
- **tov**: number of balls lost by the player that allowed opponents to take a fast shoot, per game.
- **steal**: number of ball stolen by the player, per game
- **blocks**: blocks per game
- **fouls**: fouls per game
- **off_rtng**: individual player's efficiency at producing points for the offense per 100 possessions
- **def_rtng**:how many points a player allows per 100 possessions
- **net_rtng**: How much better or worse the team is when the player is on the court. defined as off_rtng - def_rtng
- **plus_minus**: difference in result when the player is on the floor, per game
- **p_usage**: Usage rate calculates what percentage of team plays a player was involved in while he was on the floor, provided that the play ends in one of the three true results: field-goal attempt, free-throw attempt or turnover

2.2 Understanding the Dataset

The time spent on the court heavily influence the players performances:
A player P that plays 5 minutes per games, will have less possibility to miss shots, lose balls, take rebounds, etc..., in summary: he will have less possibilities to show his ability for such a sufficient time that they can be captured by the statistics
Extreme case: P plays 1 minute in the entire season, and in that minute he scores a 3 point shot, follow that he will have $p\_3point = 100\%$, far better with respect to the best shooters in the league.
Following this line, the statistics of the players who play less minutes may be randomly built.
Moreover, it is nearly impossible try to standardize the data on the minutes played, since the statistics don't grow at the same rate of the minutes played:
recalling the previous example, it can't be predicted if P will score or miss the next 10 shots, if the only available information about his 3-point-shot ability is p_3points = 100%.

This dependency on the time played will probably 'stretch' the clusters based on the players ability, rather than their role, since better players naturally play a higher number of games and a higher amount of minutes: the risk is that players will be clustered based only on their ability despite to their characteristics.

In order to mitigate this effect, the players which have played less than a certain threshold of minutes in the season have been deleted from the dataset.
The threshold has been chosen from a new variable created multiplying *min* by *games*, which obviously represent the total number of minutes played in the season.
From the dataset then, have been removed the players with less than *0.1-quantile* of total minutes played. 61 players have been removed.
In addition, from the dataset have been deleted the variables *min*, *games*.

2.3 Dataset transformation

2.3.1 NA's

The dataset not contained missing values.

2.3.2 Player

the variable *Player* has been deleted from the dataset, and it has been used as rows' index to allow more clear graphs.

2.3.3 Scaling

In the last stage of this first section the dataset has been scaled and centered.

2.3.4 Summary

After the changes, the dataset is now composed by 31 numeric variables and 545 observations. Below, some statistic which regard the features:

```
> summary(nba)
      age                weight             height              points         a_field_goals      p_field_goals          p_efg          
 Min.   :-1.7201   Min.   :-2.25392   Min.   :-4.16964   Min.   :-1.4138   Min.   :-1.4443   Min.   :-3.60146   Min.   :-4.37807  
 1st Qu.:-0.7732   1st Qu.:-0.70897   1st Qu.:-0.76334   1st Qu.:-0.7660   1st Qu.:-0.7446   1st Qu.:-0.55283   1st Qu.:-0.47460  
 Median :-0.2998   Median :-0.01805   Median : 0.06243   Median :-0.2396   Median :-0.2675   Median :-0.07089   Median : 0.06755  
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000  
 3rd Qu.: 0.6471   3rd Qu.: 0.63255   3rd Qu.: 0.57853   3rd Qu.: 0.4527   3rd Qu.: 0.5276   3rd Qu.: 0.46613   3rd Qu.: 0.53741  
 Max.   : 3.7244   Max.   : 3.88463   Max.   : 3.46872   Max.   : 3.4935   Max.   : 3.0721   Max.   : 3.40459   Max.   : 4.45293  
   p_2points          a_3points          p_3points          prop_2p            prop_3p          a_free_throws      p_free_throws     
 Min.   :-2.69120   Min.   :-1.2910   Min.   :-2.5722   Min.   :-2.60610   Min.   :-1.87245   Min.   :-1.0978   Min.   :-3.9024  
 1st Qu.:-0.68896   1st Qu.:-0.8042   1st Qu.:-0.2477   1st Qu.:-0.68940   1st Qu.:-0.60090   1st Qu.:-0.6657   1st Qu.:-0.3267  
 Median :-0.02505   Median :-0.2289   Median : 0.2122   Median :-0.01388   Median : 0.01384   Median :-0.2336   Median : 0.1933  
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.65727   3rd Qu.: 0.6561   3rd Qu.: 0.5571   3rd Qu.: 0.60087   3rd Qu.: 0.68936   3rd Qu.: 0.3219   3rd Qu.: 0.6007  
 Max.   : 2.56749   Max.   : 3.8866   Max.   : 5.7894   Max.   : 1.87244   Max.   : 2.60603   Max.   : 6.1861   Max.   : 1.4584  
 p_true_shooting        oreb               dreb               reb              p_oreb            p_dreb            p_reb        
 Min.   :-4.32433   Min.   :-1.2035   Min.   :-1.5941   Min.   :-1.5559   Min.   :-1.2832   Min.   :-2.5153   Min.   :-1.8999  
 1st Qu.:-0.43506   1st Qu.:-0.6649   1st Qu.:-0.7546   1st Qu.:-0.7442   1st Qu.:-0.7448   1st Qu.:-0.6985   1st Qu.:-0.7343  
 Median : 0.06197   Median :-0.3956   Median :-0.1670   Median :-0.1888   Median :-0.3763   Median :-0.2677   Median :-0.3127  
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.57143   3rd Qu.: 0.4124   3rd Qu.: 0.4207   3rd Qu.: 0.4093   3rd Qu.: 0.4526   3rd Qu.: 0.5564   3rd Qu.: 0.5864  
 Max.   : 3.98851   Max.   : 4.9908   Max.   : 4.5621   Max.   : 4.6814   Max.   : 4.2998   Max.   : 3.3471   Max.   : 3.4819  
     assist             tov               steal             blocks             fauls            off_rtng          def_rtng      
 Min.   :-1.1005   Min.   :-1.3296   Min.   :-1.6195   Min.   :-1.0617   Min.   :-2.2481   Min.   :-6.9125   Min.   :-3.65975  
 1st Qu.:-0.6820   1st Qu.:-0.7211   1st Qu.:-0.6161   1st Qu.:-0.5301   1st Qu.:-0.7216   1st Qu.:-0.4289   1st Qu.:-0.55767  
 Median :-0.3445   Median :-0.2342   Median :-0.1144   Median :-0.2643   Median :-0.0278   Median : 0.1394   Median : 0.02745  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
 3rd Qu.: 0.3576   3rd Qu.: 0.3743   3rd Qu.: 0.6382   3rd Qu.: 0.2672   3rd Qu.: 0.8048   3rd Qu.: 0.6493   3rd Qu.: 0.53827  
 Max.   : 4.7318   Max.   : 4.1472   Max.   : 4.1502   Max.   : 6.3802   Max.   : 2.8863   Max.   : 3.7236   Max.   : 4.27192  
    net_rtng          plus_minus          p_usage        
 Min.   :-6.1679   Min.   :-8.584324   Min.   :-2.6493  
 1st Qu.:-0.4624   1st Qu.:-0.477104   1st Qu.:-0.6913  
 Median : 0.1317   Median : 0.006484   Median :-0.1347  
 Mean   : 0.0000   Mean   : 0.000000   Mean   : 0.0000  
 3rd Qu.: 0.6321   3rd Qu.: 0.546965   3rd Qu.: 0.5372  
 Max.   : 2.8381   Max.   : 2.651998   Max.   : 3.7813  
```

Figure 1: summary

3 Correlation

Figure 2 in next page shows the correlation matrix between all the variables. It can be noticed that there are some patterns between groups of variable, below some consideration:

It can be observed a group regarding variables that deal with having a lot of times the possibility to play the ball, like *tov*,*assist*, *a_field_goals* and *points*.

Another one concerning the rebound-statistics, fouls, blocks *prop_2p* and weight. These variables are even negative correlated with the ones concerning the 3 point shots.
It's has to be notice that *fouls*, *reb*, *dreb* have a noticeable positive correlation even with the first group analyzed. These variables are absolute statistics as suggested in the section 2.2.

Lastly it can be observed high correlation concerning the efficiency-statistics, probably due to the way in which these statistics are calculated.

Correlation means linear relation: two variables are correlated if they change with a similar rate. In a scaled dataset, correlation means that two variables have similar values.
Since the rows of the dataset are the players' statistics, this means that higher is the correlation between two variables, higher is the number of observations with similar values. Lower the correlation, lower the number of rows -so the number of players- which have similar values.
Extending the reasoning to the groups of variables it can be seen as there are chunks of variable with a similar behavior, which in turn signify that there are a lot of players which have similar block of statistics.
For example it can be said that there are a lot of player which have an high (or low) number of assists, points, attempted field goals, usage. Or which there are a lot of player that have high (or low) weights, rebounds, 2points propension and these players usually have low (or high) *3points propensity*, *3 points attempts* and *3point percentage*.
These considerations can show a first sight to possible differences among players.

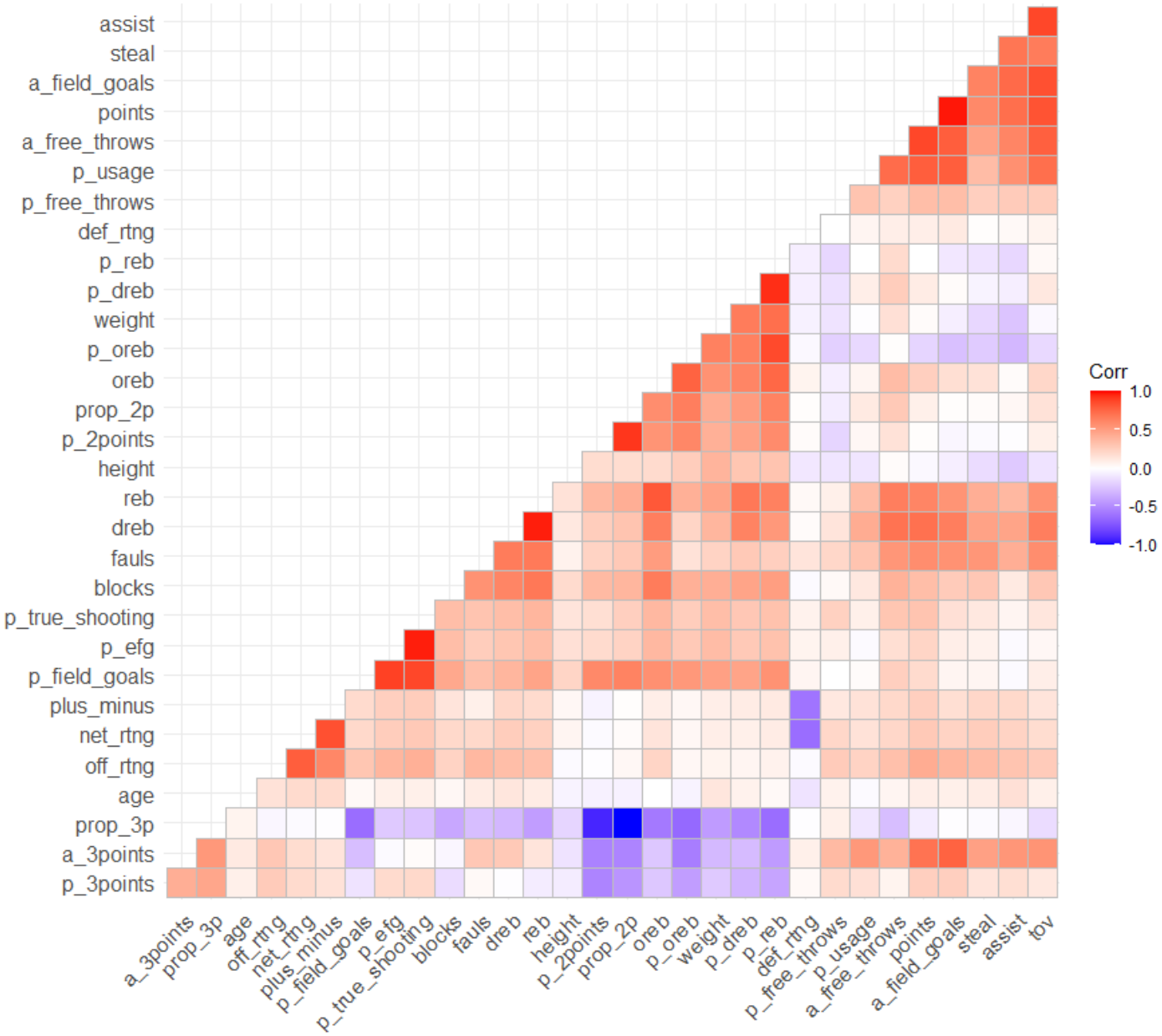

Figure 2: Correlation

4 PCA

Principal Component Analysis is a technique of dimensionality reduction, but it can be useful even for visualisation, in this work it will be used especially for this purpose.

In short, PCA combines the variables to find the dimensions which explain the most variability possible: the first component is directed in the direction in which the data vary the most, the second one is orthogonal to the first one and it is directed in the second direction in which the data vary the most, and so on.
For example, if the data are distributed in a 2 dimensional ellipsoidal shape, the first component will be directed as the longer axis, the second component will be directed as the shorter axis.
It is important to use the PCA with scaled data, otherwise the method will weight more the variables with the higher values.
PCA can be used for a lot of analysis, in this section will be studied only how the variable contribute to the first 2 components, since it will be useful for future analysis. For example the BiPlot, figure 11.

4.1 Explained Variability

The reduction in dimensionality has the expensive cost to reduce the variability explained. Depending on the purpose for which it has to be applied, it can be chosen more or less dimensions. In general it has a good practice to look for the 'elbow' in the *screePlot* (Figure 3), which represent the point from which adding more variables will lower the raising in the explained variability. In other words, the point from which it becomes disadvantageous to add dimensions.

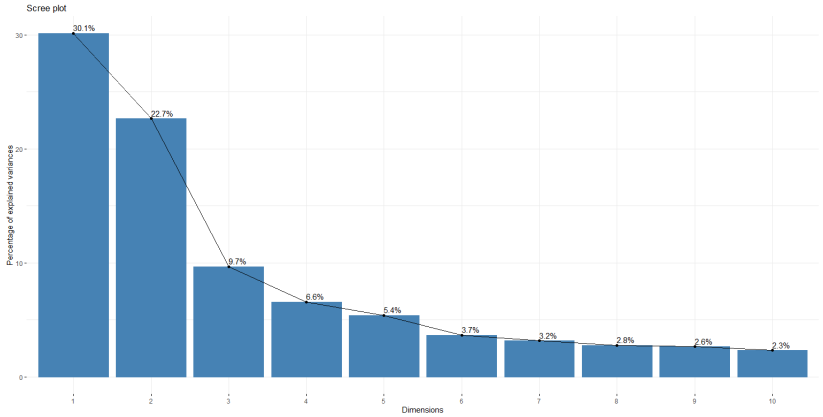

Figure 3: Principal components

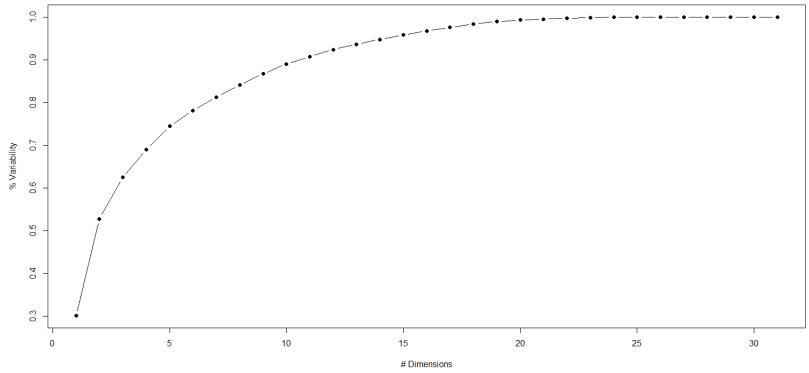

Figure 4: Cumulative sum of variability explained

The first two dimensions explain the 53% of the variability, which means that the best 2-dimensional representation of the data will allow to visualize only around the 50%of the way in which the data are spread. This it has to be taken into account in the visualisation of future plots.

4.2 Variables Contribution

In the next pages, the most crucial graphs will be plotted on the 2 principal components, hence it is important to understand which variables contribute the most, in term of variability, to them.

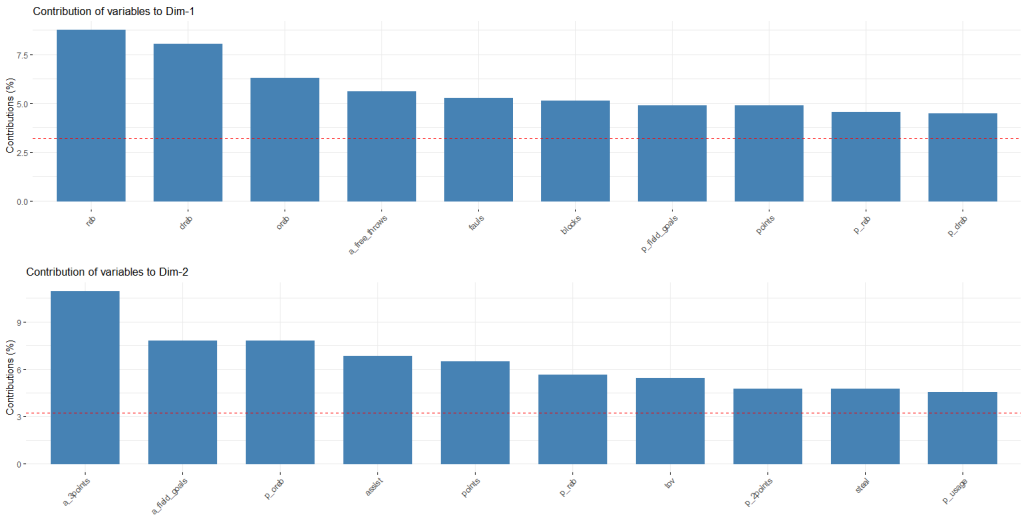

Figure 5: Variables Contribution on the first 2 dimensions

It can be observed in figure 5 that the variables that contribute the most to the first principal component are all absolute variables, and in particular

the variables related to rebounds.
In the second component instead, the most important variables seems to be related more to the scoring and ball possession area.

A better visualisation of the variables contribution is provided by the following figure:

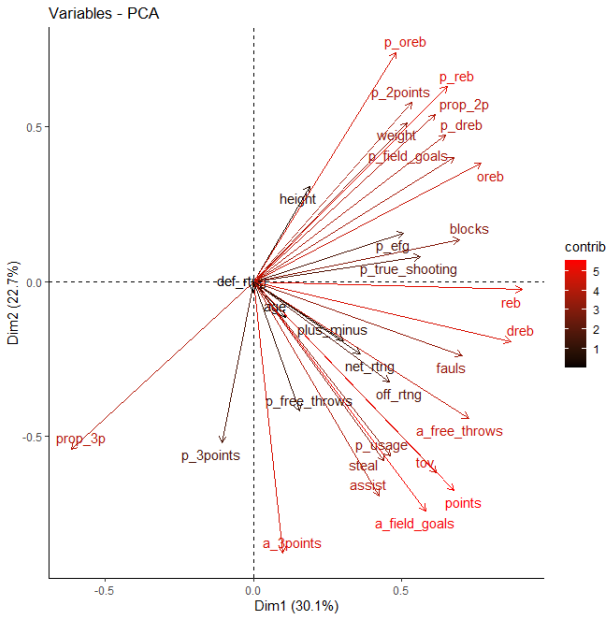

Figure 6: Variables Contribution on the first 2 dimensions

Brighter the red, brighter the contribution of the variable to the dimensions, the angle says how much the variable contributes to the Dimension 1 or 2.
For Example *Reb* which almost overlaps the dim1 axis, give an high contribution to dim1 and a low contribution to dim2.
On the contrary *a_3points* gives an high contribution to dim2 to and a low one to dim1.
The vectors who lies on the bisectors are the variables who equally contribute to dim1 and dim2.
Variables like *height* or *age* don't contribute much to the first two dimensions.

5 K Means

5.1 Theoretical Framework

It is a clustering method which requires as input K, the number of cluster desired. The algorithm will assign each observation to one and only one of the k clusters, which means an observation can't be assigned to two different clusters.
The aim of the K means is to minimize the *within cluster variation*, so the algorithm want create clusters in which the observations are as similar as possible.
Similarity means distance, and there a lot of possibilities to define a distance between 2 points.
The most common choice is the Euclidean distance, but the are many other possibilities like the cosine distance or Minkovski distance.
The Algorithm start from k random points, so it have to be ran more than one time in order to find the global optimum.

5.2 choice of K

As said, K means requires that the number of clusters be specified in advance.
To determine it there are several methods:

5.2.1 NbClust

Nbclust is a function which calculate the best K suggested by a number of index, given a certain method and distance metric.
It has been used as a 'Black Box', to have a first insight of what were the proposal of the majority. This function has been used with

- 2 distance type: *euclidean* and *minkoski*, since the data are all numeric
- all the methods allowed
- all the indexes allowed

The best result for any iteration has been saved in a table:

K	2	3	6	10
Number of choices	4	8	2	2

Table 1: most frequent choice of K

The most frequent choice is 3 clusters, but even 4, 6 and 10 have to be taken into account.

5.2.2 GAP statistic

Gap statistics measures how different the total within intra-cluster variation can be between observed data and reference data with a random uniform distribution.
Since the statistic has the tendency to grow, the suggestion is to choose the k such that the rate of increase of the gap statistic begins to "slow down".

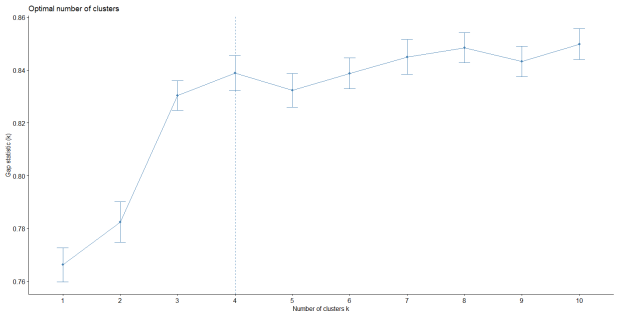

Figure 7: Gap statistic

the Gap statistic suggest 4 clusters, but even 3, since after K = 3 the curve seems to generally decrease the growing rate, seems reasonable.

5.2.3 WSS

It is the most used method to determinate K.
Even known as 'Elbow methods', this procedure is based on the Within-Cluster-Sum of Squared Errors, calculated of different values of K. K should be choose such that WSS slows down the diminishing rate.

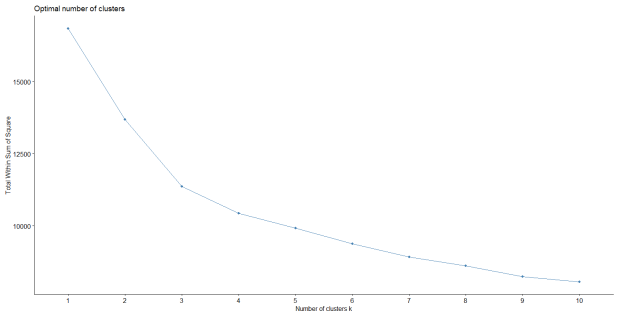

Figure 8: WSS vs K

The most evident elbow it's on K = 3, which it seems to be the best choice.

5.2.4 Silhouette

The last method proposed is the Silhouette method.
This method measures the quality of the clustering.
The silhouette is an index bounded in [-1 ; +1] and measure how an observation is similar with the other observations of his own cluster, compared with the other clusters.
The similarity method used, usually is the average distance from the observation to the other observations in the cluster. The distance metric used is usually the euclidean.
For each K, the index is then calculated as the mean of the indexes for each observation.
The best choice is the K for which the silhouette index is as near to 1 as possible.

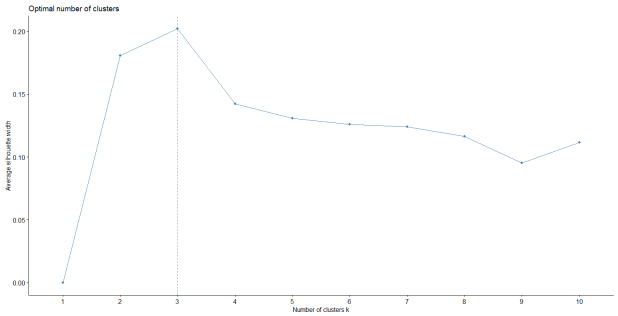

Figure 9: Silhouette

The graph shows that the best K is 3.

5.2.5 Final choice

Considering the results of all the methods, the best choice seems to be 3 clusters.
This number it is also reasonable with the expectations of 3 or 5 ideal clusters.
In the next section, the result of a 3-means algorithm.

5.3 Cluster Analysis

The 3-Means algorithm divided the dataset in 3 clusters, composed by 275, 133 and 138 observations.
Figure 10 shows the 3 clusters plotted with respect to the two principal components.

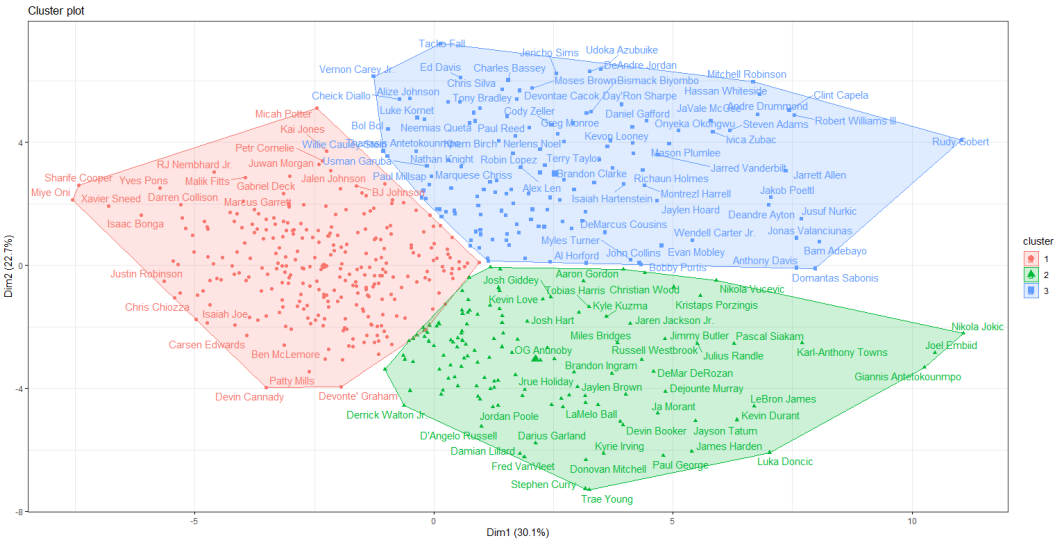

Figure 10: 3-Means Clusters

Below, a first guess about the composition of the clusters based on the names contained.

It seems that the **green cluster** contains the strongest players, regardless of their characteristic. In the same plot for example we have Kyrie Irving, Lebron James and Giannis Antetokoumpo, which are completely different players in playing style, physical characteristics and role.
It can be observed that inside the cluster the players seems to be grouped by role: in the south we have guards (Doncic, Booker, Lillard, Curry), in the north forword or centers, like Embiid, Jokic, Towns or Wood.

In the **blue cluster** seems that there are the other part of the centers and forward, in the south part there are strong player like Adebayo, Davis, Sabonis, which are similar to the ones in the green cluster. In the north instead, there are in general less skilled Centers.

The **red cluster** is more difficult to describe, contains *all the other* players but they seems to be very different one from each other.
The principal characteristic that they have in common it seems to be that they are all not centers role players.

In the next figure the BiPlot divided by clusters. This graph can be useful to understand the relation between the variables, the observation and the clusters.

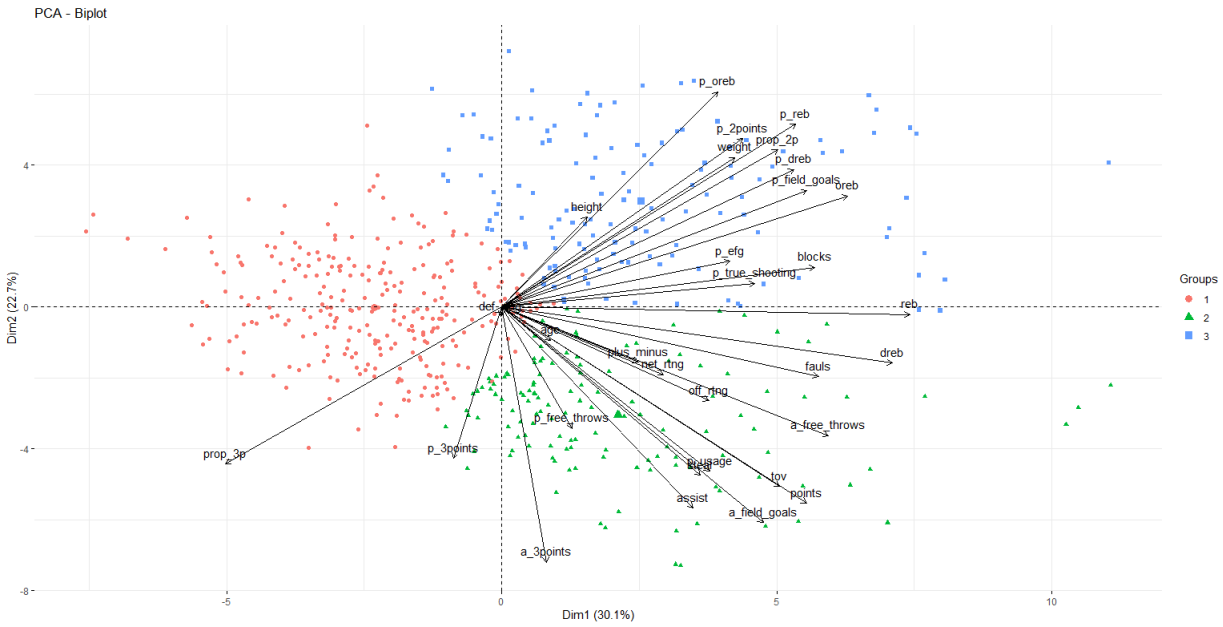

Figure 11: Bi Plot with clusters

In order to make more clear the analysis, in the following section will be used the vectors of the variables, the bisectors or the quadrants as reference to identify areas of interest. It has to be underlined that this is only in favor of interpretability and doesn't mean that the observation of a certain types are confied by these elements.

In the **first quadrant**, which approximately correspond with the blue cluster, it can be observed that the vectors regard the percentage of rebounds, the propensity to try 2 points shots, the weights and the percentage of field goals. These characteristics reasonably belong to centers: they are able to catch rebounds, to go near to the basket to shoot, and this obviously increase their efficiency. On the other hand they are less inclined to try 3 points shots.

Along the positive first dimension, in the **fourth quadrant**, there are vectors regarding the rebounds in absolute term. This tendency continue below with other variables: *points*, *assist*, *tov*, and attempted and made free throws.

This area, which approximately represent the green cluster, it seems to contain the *Stars* and the top of *first strings*, recalling the description provided in introduction: the players who play most of the time and consequently can collect more statistics. This is consistent with which suggested in *Section 5.3* about the green cluster composition.
In the first quadrant can even be observed that:

- Getting closer to the y axis, there are variables regarding the 3 point shots
- Getting closer to the x axis, there are vectors regarding the rebounds

This confirm what suggested before: between the y axis and the bisector the players are mostly guards, moving towards the area between the bisector and the x axis, there are more Forwards or Centers

The second and the third quadrants approximates the red cluster.
In the **third quadrant**, the players between the bisector and *p_3points* may be identified as 'specialists', players whose role is basically to shoot 3 pointers. In this area there are players with an high 3points inclination, which however lack in *rebounds*, *blocks*, *height* and *efficiency* -since 3 points shooting is much harder than 2 points shooting-.
The best shooters between the strong players could be identified in the area between *p_3points* and the second bisector. In this area there are players who are able to shot 3 pointers, but they have not a clear inclination because they can do many other things. Curry, Lillard and Booker which correspond to the identikit, belongs to this area.

In the **second quadrant** there are mostly role players: they lack in all the absolute statistics, which means that they are few involved in the game. Between the bisector and the y axis there are probably the forwards or centers role-players.
Between the bisector and x axis there are probably the guards role players.

5.4 Cluster Composition

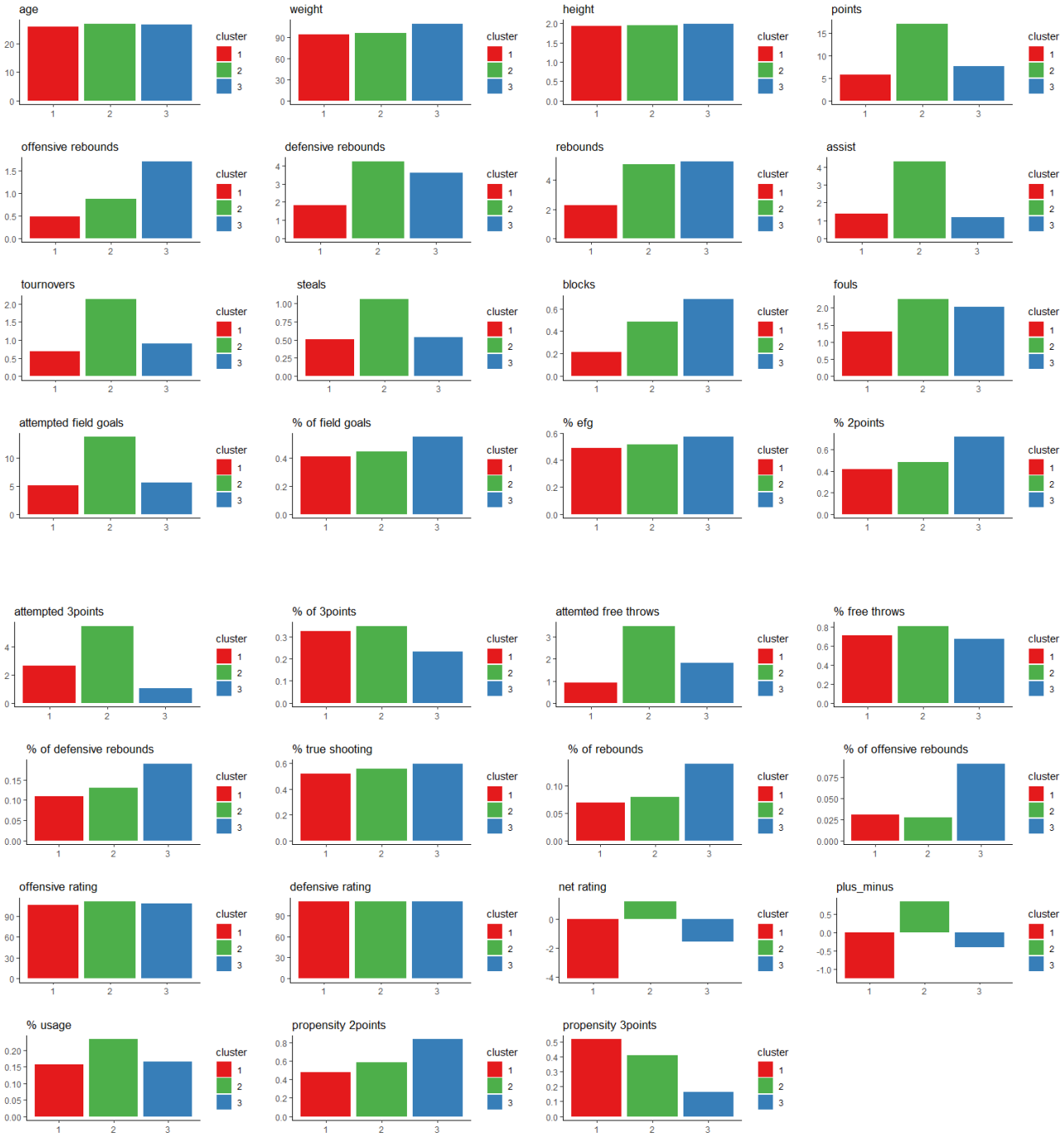

Figure 12: variables mean by Kmean clusters. The variables come from the not scaled dataset

Figure 12 shows the mean of each variable by cluster.

- the variable *age* seems to be constant over the clusters.
- the **Blue Cluster** have higher values in the variables regarding the rebounds, the blocks, the propensity at trying 2 points-shots and the scoring efficiency.
  Moreover, in this cluster there are the tallest and the heaviest players. Players in this cluster correspond to the identikit of center, or power forward, consistently with which suggested above. This cluster seems to be defined more on *characteristics* rather than *ability*
- In the **Green Cluster** dominate the 'absolute statistics', the usage, and the ratings. This combination confirm that contains the stronger players.
  This cluster is created based on the *ability*, rather then the *characteristics*. In fact, it has to be observed as this cluster have level of *defensive rebounds* and *offensive rebounds* comparable with the blue cluster's ones, but it has much lower level in percentage rebounds. This is because they play in general more than the blue cluster-players, so they catch more rebounds *in absolute value*, but blue cluster's players take more rebounds in proportion to how many they could.
- The only characteristic in which the **red cluster** leads, is the *3 point propensity*. They then have a comparable *p_3points* level with the green cluster. Moreover They all the only cluster which have negative *net rating* and *plus/minus*, which means that the team gets worse with them on the field.
  It can be seen that the players are a bit younger, they weight less and they are smaller.
  This cluster is the more mixed: but it seems to contain 3 points specialists, role-guards/small forward or maybe even the less skilled *Rookies*, the younger players which play their first season in the league.

6 hierarchical clustering

6.1 Theoretical framework

It is an unsupervised learning algorithm in which it is not required to pre-specific the number of clusters K.
This algorithm can follow 2 main approaches:
top-down, which start from one big cluster and splits it into smaller ones
agglomerative, the most used, whose functioning is shortly explained below:
Start considering each point as a cluster, then, the algorithm identify the 2 most *similar* clusters and it merge them in a unique cluster. The procedure is iterate until there is a unique cluster.
The algorithm works with the concept of *similarity*, so it has to be specified the type of distance to use and the points for which this distance has to be calculated and minimized, the *linkage method*.
In the first iteration the clusters are made by only one observation, so it is simple to identify the observations themselves as the points between which calculate the distance. But when the clusters are made by number of observations, it is not trivial to define the points between which calculate the distance, on the contrary, different methods can be used to measure different things. Below a short list with the most common linkage methods.

- **Ward**: the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimize the within-cluster sum of squares.
- **single**: the distance between two clusters is the distance between the 2 nearest points.
- **complate**: the distance between two clusters is the distance between the 2 farthest points.
- **average**: the distance between two clusters is the average distance between an observation in one cluster and an observation in the other cluster.

after defining distance and method, the algorithm will return as output a tree-based object called *Dendrogram*.
The dendrogram can be defined as a summary of the distance matrix. The Height of each vertical line represents the dissimilarity between the the clusters.

It is important to stress that the hierarchical clustering do not suggest the optimal number of clusters, but it can be a good practice to cut the dendrogram where the lines are longer, to create more dissimilar clusters.

6.2 Dendrograms and K selection

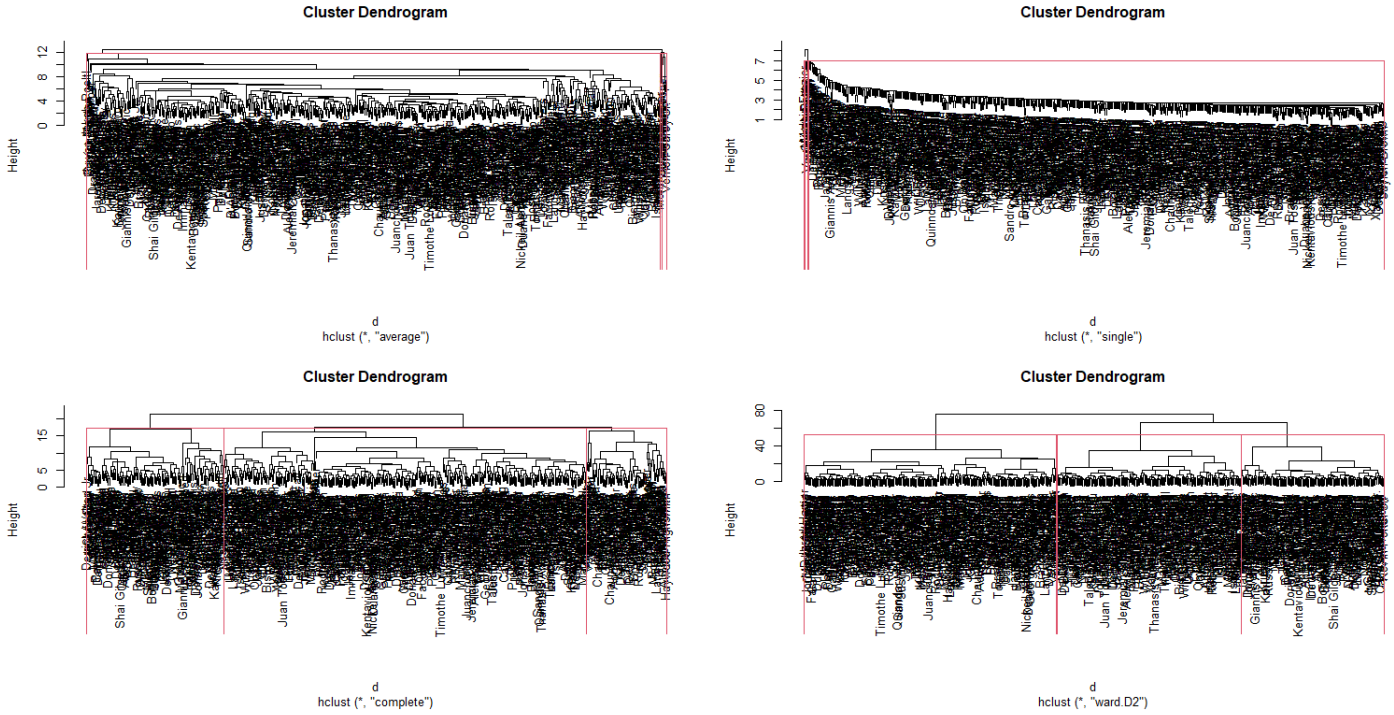

Figure 13: Dendrograms

Since the dataset were composed all by numeric variables, it has been used the euclidean distance.
The best method seems to be the Ward, since it generate the most balanced dendrogram.
K = 3 seems to be a reasonable choice by looking at the dendrogram shape.

6.3 Clusters Composition

The Clusters are more balanced than the K-means ones: 134, 237 and 173 observations.
Figure 12 shows the average level of each variable, by cluster.

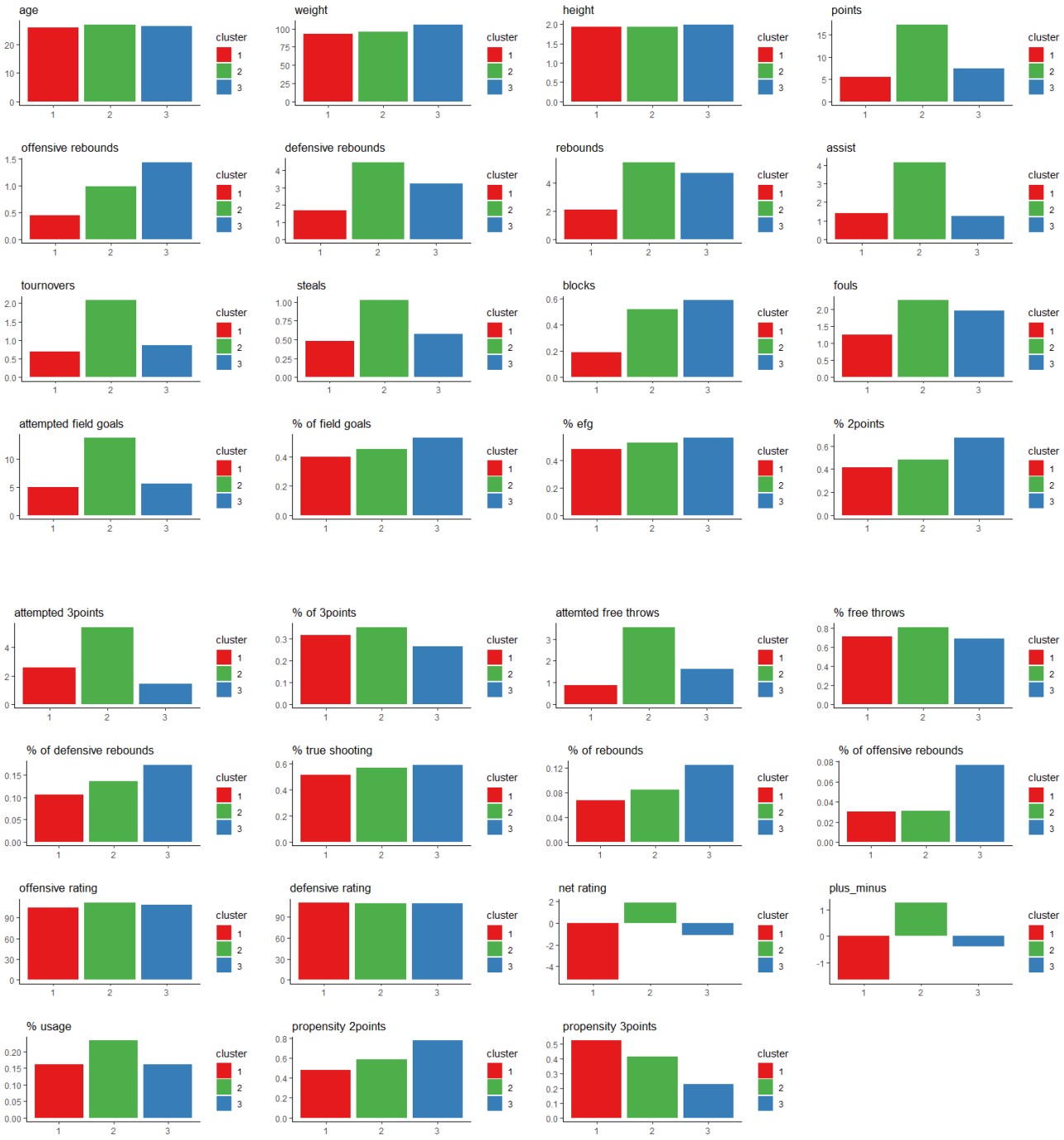

Figure 14: variables mean by Hierarchical clusters. The variables come from the not scaled dataset

It can be observed that in general the results seems to be very similar to the K-means' ones. Shortly:

- the **Red Cluster** contains different types of players: guards, small forwards who plays few minutes per game, and 3 point specialists.
- the **Green Cluster** contains the stronger players in the NBA, the stars in each role, in fact they have the highest absolute statistics.
- the **Blue Cluster** contains Centers and Power Forwards: they are the tallest and heaviest players in the league, this allow them to take the majority of the rebounds which they have the possibility to catch, to block the opponents and to go near to the basket to score simpler shots.

7 Compared Results

In this section will be deepened the relationship between the results provided by the 2 algorithms, trying to understand what are the similarities and what the differences.

7.1 Similarities

Figure 15: Kmeans and Hierarchical scatterplot

Figure 13 shows the K-means and Hierarchical algorithm scatterplots.
In general, the two graphs seems similarly distributed.
It can be observed that K-means clusters seems to be more well-divided than the Hierarchical ones, but it has to be remarked that the graph are built among the two principal components which are able to explain only the 50% of the variability.

Below, the table which illustrate how the observations were assigned by the 2 algorithms.
The 85% of the observations is clustered in the same way by both the methods.
It is interesting to note how H_Red has the highest proportion of differently classified observations, 17% in contrast with K_Red which have the lowest proportion of differently classified observations: 4%.
H_Red and K_Blue differently classify the highest number of observations:

	K_Red	K_Green	K_Blue
H_Red	266	10	37
H_Green	8	113	11
H_BLue	3	12	124

Table 2: Clustering cross matrix

6% of the entire population. They are probably Center or Forward role-players. recalling Figure 11, These are probably the observations located around the positive y axis.

7.2 Inter-Cluster Differences

In this section will be analyzed the differences between clusters of the two algorithms. The lines represent the difference between the *scaled* variable

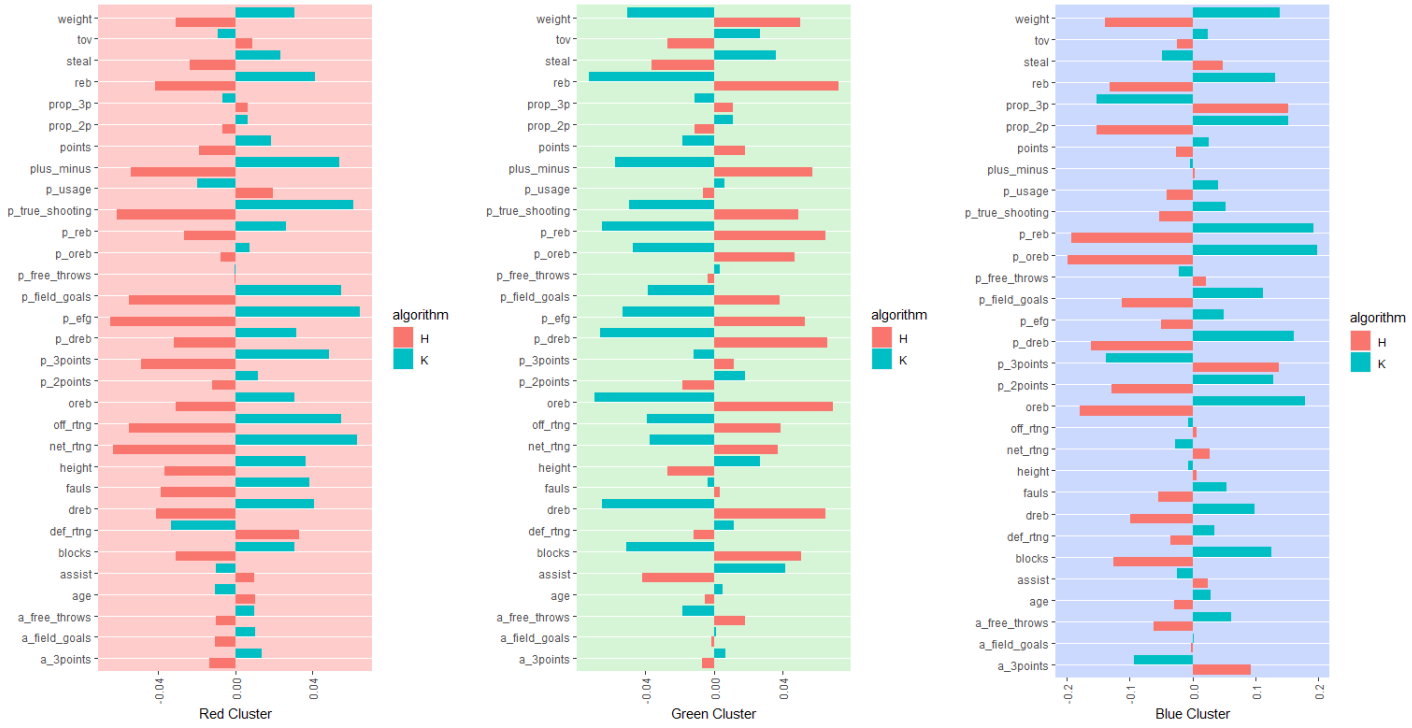

Figure 16: Compared Clusters

and the mean of the *scaled* variable in the two clusters. For Example: K mean weight has been obtained from K mean weight − mean(k mean weight, Hierarchical weight).
The lines are hence mirrored around the zero, which means that the length

of each pair of bar represent the distance between them. Moreover, since the variables were scaled, this figure allows even comparisons between variables.

At first sight, seems evident that in **K-Red Cluster** approximately all the variables are higher, even with relevant differences: In particular, the variables *plus_minus*, *p_efg*, *p_true_shooting*, *p_field_goals*, *off_rtng* and *net_rtng* have a large difference in the two algorithms.
From these, the 5 variables regard the Centers. The first one instead, it can be seen as a measure of how good a player is. Which means that K-means assigned in the K-Red Cluster more efficient players and players with an higher propensity to play as Centers. It cannot be said if they really are Centers, because this figure shows only differences, not absolute values.
It is interesting analyzing even the variables which are similar: *a_field_goals*, *a_free_throws*, *a_3points*, and the two propensity, for example.
These variables are very different one from each other, so it is difficult to draw conclusions. It can be surely said that around all the *absolute variables* are very similar, so the two clusters contain players with a comparable ability.

In the **Green Cluster** the situation is mirrored, the Hierarchical clustering has a higher average in almost every variable.
In particular, it can be seen that are notably high the variables regarding the *Rebounds*. On the contrary the only variables which are significantly higher in the K-means are *steals*, *tov*, *assist* and, unexpectedly, *height*. The hierarchical method assigned to the green cluster players who play more like centers, the K-means instead fill the cluster with players who plays more like guards. The variable *height* leads to think that in this cluster there are high players who plays more like guards or small forwards.
The similar variables here regard the shooting area, the propensity and the usage. This again confirm that this cluster is based more on the ability than on the characteristic of the players, in fact, regardless the difference, this cluster in general contains the best of the NBA.

The **Blue Cluster** shows that K-means have high values in the rebound area, in *prop_2p* and *Weight*. K-means Blue Cluster then contain players who are 'more centers' than the hierarchical one.
Hierarchical have higher values in the variables concerning the 3point shooting.
Pairing this situation with the one analyzed for the green cluster, seems evident that the K-means has more sharply divided the centers, which the hierarchical based the aggregation more on the abilities.

At the end, it has to be remembered that the two algorithm classified the observation almost in the same way. These difference are marginal consid-

ering the whole set of data, but they need to be underlined as they highlight different tendencies in the clustering.

7.3 Intra-Cluster Differences

In this section will be analyzed the ranges for each variable in the 3 clusters, for both the hierarchical and the K-means. The Range tells how different are the variable in the clusters. Obviously the range is a simple method that may be inaccurate:
For example, if the variable α has an average 1 in cluster 1, 95 in cluster 2, and 100 in cluster 3, it is a completely different situation with respect to a case in which α has average 1 in cluster 1, 5 in cluster 2 and 100 in cluster 3. It has to kept in mind that this type of information is lost with this metric.

Figure 17: Intra-Cluster differences

The *Light blue bars* represent the variables for which $range > 3quartile$, This threshold has been chosen to define the **wide ranges variables**. The *red bars* represent the other case.
K-means and Hierarchical shares 5 out of 8 wide ranges variables, which are:

- tov
- points
- assist
- a_field_goals
- a_3points

In addition:
K-means has *prop_3p*, *p_reb* and *o_reb*

Hierarchical has *p_field_goals*, *dreb* and *a_free_throws*.

A variable α with a wide range is a variable which have had an important role in clustering, it means that the observations have been often assigned to different clusters based on different values of α, and recalling how these algorithms works, it means that the value of this variable cause dissimilarity between the variables.
An example of this type of variables is *Points*: this variable have very different values between clusters. That means which there are at least two clusters in which this variable assume, in average, very different values.

On the contrary, a variable that have a low range is a variable that has a similar average in the clusters, which means that it has been less useful in the clustering process. *age* is an example:
Figure 12 and 14 show how it is equally distributed in each cluster, by both the algorithms.
In short: clusters don't classify players based on their age but on their points.
Extending this reasoning to the whole dataset, the lightblue variables are the most important to understand what the clusters actually represent and what type of players are formed by.

8 Conclusions

The two methods used to cluster the date provided, for the most, comparable results. The most evident differences regarded the treatment of the players based on the variables concerning the rebounds.
The PCA has been able to explain around the 50% of the variability with the first two components, which it not a very good result, but it was enough to allow a clearly visualisation of the clusters. Moreover, the Biplot provided useful points about the relation between the clusters and the variables.

8.1 Clusters

Weighing all the considerations and analyses made above, it will be given a view on the composition of the clusters.
Since the two algorithm provided convergent results, the results will be present together with specified differences where needed.

Red Cluster - role players and specialists

Another possible definition for this cluster may be 'the other cluster', in the sense which here are collected all the players that more than having characteristics in common, are different from the players in the others clusters.
In general it can be said that they are role-players, the one who plays when the first strings have to rest.
Moreover in this cluster are present the specialists: players whose role is to shoot 3 pointers.
Principally, is composed by guards or small forwards, but in the **K-Means Cluster** there probably is some power forward or Center more than in the **hierarchical Cluster**.

Green Cluster - Star and first strings

This cluster contains the star for each role.
They have high values in all the absolute statistic because they play more time, and they can play more balls than the others players.
Moreover, this cluster contains the first string guards and small forwards, players who are not the first offensive choice, but they are in general strong and useful complement-players.
the **Hierarchical Cluster** tends to assign to this cluster even first string Centers and Power Forwards, probably the ones whose characteristic are

more similar to guards one: good shooting, offensive and defensive lectures.

Blue Cluster - Centers

The last cluster is the most well defined by the characteristics point of view. Here there are Centers and Power Forwards, the bigger players in the league who use their bodies to take rebounds, blocks, and go near to the basket to shot.
Mirroring what has been observed with the green and red clusters, the **K-means** tends to assign to this cluster all the first strings Centers, meanwhile the **Hierarchical** also tends to include role centers or power forward, and to exclude some specific type of center or power forward, as saw above.

8.2 for the future

The Algorithms have been clustered the data along the two dimensions: *ability* and **characteristic**, creating mixed clusters.
In order to create clusters able to divide the players by their role, the suggestion is to work on 2 principal directions:
The first one regards to find an efficient way to make all the statistics independent from the time played, predicting realistic statistic which are able to taking into account the strength of the player.
The second one regards the collection of other types of data, for example heatmaps which contain information about the most occupied position on the court.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani *An Introduction to Statistical Learning*

[2] Dataset: `https://www.nba.com/stats/players/traditional/`