MULTIVARIATE ANALYSIS

DATA ANALYSIS
PROJECT

DATA SCIENCE & ENGINEERING
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Adrián Cerezuela Hernández - 48222010A

Ramon Ventura Navarro - 21785256R

May 2023

Contents

1 INTRODUCTION

This is an evaluative project for the subject of Data Analysis [Aca23], part of the Bachelor's Degree in Data Science and Engineering. The main objective is to observe which statistic have the greatest impact on basketball player performance and analyze which of them determine a player's skill level. This will be carried out using the methods discussed in class: Principal Component Analysis (PCA) and Multidimensional Scaling (MDS), Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA), discriminant analysis (LDA/QDA) and clustering.

2 DATASET

2.1 Data Description

In order to achieve our objective, we will use a database that contains individual data of all the players currently playing in the ACB (Asociación de Clubes de Baloncesto), which is the Spanish national basketball league. The source is the following website: Zenodo [Zen]. The included data corresponds to the statistics collected throughout all the seasons in which the players have played, being the 20-21 season the last one from which we have data.
Each row corresponds to a player. If a player has taken part in multiple seasons, they are repeated in as many rows as seasons they have played, followed by their statistics for that year. Since this could pose a problem for our analysis, we have restricted the observations to the statistics collected in the 2020-2021 season. So that our dataset goes from having 1251 observations to 262.
One reason for choosing that season is because it is the one from which we have the most data, all the data to be precise. As we go back in seasons, we have less collected information. The only downside to selecting the 20-21 season is that the observations correspond to the first 9 league matches, not the entire season. Therefore, the analysis we will perform will be carried out on the data of all players through 9 games, but it will be scaled to an entire season.

2.2 Data Preprocessing

In order to prepare the data for analysis, we have to adjust the original data to what we need in order to conduct a proper analysis of it.

Initially, repeated and missing values will need to be checked. Our dataset has two duplicated players: *Jovan Kljajic* and *Charlon Kloof*. The likely cause is the transfer market during the season. Furthermore, if we take a look at their stats on both rows they are the same. So that one of each will be removed. Specifically, the one around players of other teams, preserving the order.
Regarding missing values we identify five of them. Except one found in a player's position, they are all located in the *license* column. If we look at the specific players who have this column empty, we can see that they are all young development players. Therefore, it is easy to see that these values should be imputed with *JFL* license. The one located in a player's position is imputed manually.
Once this has been checked, we will need to make several modifications to the columns so that we can make these adjustments.

The target in this project will be based on the PIR(Performance Index Rating) [Wik], as it is the attribute that encompasses the most aspects of the game and from which we can better represent a player's performance. The formula to calculate it is as follows:

$$PIR = PTS + TR + AST + STL + FD + BLKF - (FGA - FGM) - TOV - BLKA - FC \quad (1)$$

This variable will be categorized in a column called *Performance*, which will be our response variable. Later on, the criteria for defining the classes will be explained.

Firstly, since the analysis will be performed from a few games, we are interested in working with averages instead of total statistics. Therefore, the first step will be to divide the different statistics by the games played and remove the columns of total statistics. Additionally, some irrelevant columns for our analysis should also be eliminated, such as *dunks* or *5i*. Once this is done, if we look at formula (1) total shots are taken into account, so we can unify the columns for 1-point, 2-point, and 3-point shots into a single column, both attempted and made separately.

Once we have narrowed down the columns to only those of our interest, we will adjust the way certain values are expressed to make them more manageable. However, before carrying this out, we can see that the names of the columns are in Spanish and are also impractical. Therefore, we must change them before making the modifications.
Then, we will replace the date of birth with a new variable called 'age', which contains their corresponding age in 2020. Besides, we will replace the position with an identifying letter, the initial one, and lastly, we will replace percentage values, expressed as strings together with the '%' symbol, and height values, expressed as strings with commas, with numerical values.

Finally, the last step will be to categorize the PIR and define the response variable *Performance*. Looking at our data, we see that the range of values in this column goes from -4 to 25.14, distributed as follows.

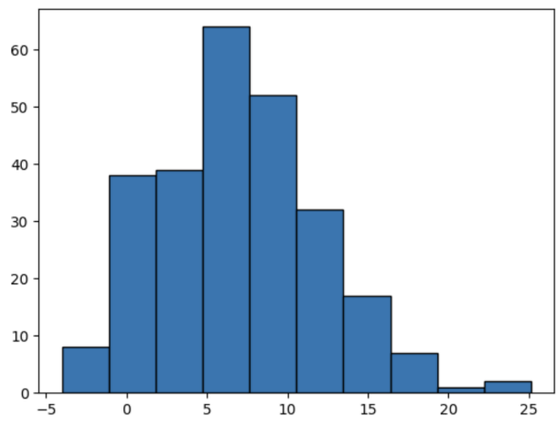


Figure 1: Range of *PIR* values

Arbitrarily, we will divide a player's performance into three possible classes: *Poor* if the value ∈ (-5,5], *Good* if the value ∈ (5,15] and *Outstanding* if the value ∈ (15,25.15].

Therefore, considering the structure of our data and assigning the correct data type to each variable, the final attributes of our dataset, apart from the textual variable *name*, are the following:

- Numeric variables: height, age, PPG (*Points Per Game*), , FGM, FGA (*Field Goals Made/Attempted*), X3PT_Per, X2PT_Per, FT_Per(*3PT/2PT/FT Percentage*), TR (*Total Rebounds*), AST (*Assists*), STL (*Steals*), TOV (*Turnovers*), BLKF, BLKA (*Blocks For/Against*), FC, FD (*Fouls Comitted/Received*), PIR (*Performance Index Rating*).
- Cathegorical variables:
  - **pos** (*Position*): PG(*Point Guard*), SG(*Shooting Guard*), F(*Forward*), PF(*Power Forward*), C(*Center*).
  - **license**: EXT(*Non-EU player*), JFL(*Development Player*), EUR(*Community or FIBA Europe player*), COT(*Cotonou passport*)

- **club**: Acunsa GBC, BAXI Manresa, Barça, Casademont Zaragoza, Club Joventut Badalona, Coosur Real Betis, Herbalife Gran Canaria, Hereda San Pablo Burgos, Iberostar Tenerife, Monbus Obradoiro, Morabanc Andorra, Movistar Estudiantes, RETAbet Bilbao Basket, Real Madrid, TD Systems Baskonia, UCAM Murcia CB, Unicaja, Urbas Fuenlabrada, Valencia Basket Club.
- **Performance**: Poor, Good, Outstanding

3 DATA ANALYSIS

3.1 PCA

PCA is multivariate data analysis method useful for finding similar observations working on minor dimensions that uses linear combinations of the original variables that represent straight lines in the plane, i.e principal components, in order to carry out orthogonal projections on them.

Through the scree plot, we can compute the explained variance by the different components. It always displays a downward curve. The point where the slope of the curve is clearly leveling off (the "elbow) indicates the number of components that should be extracted four our analysis.

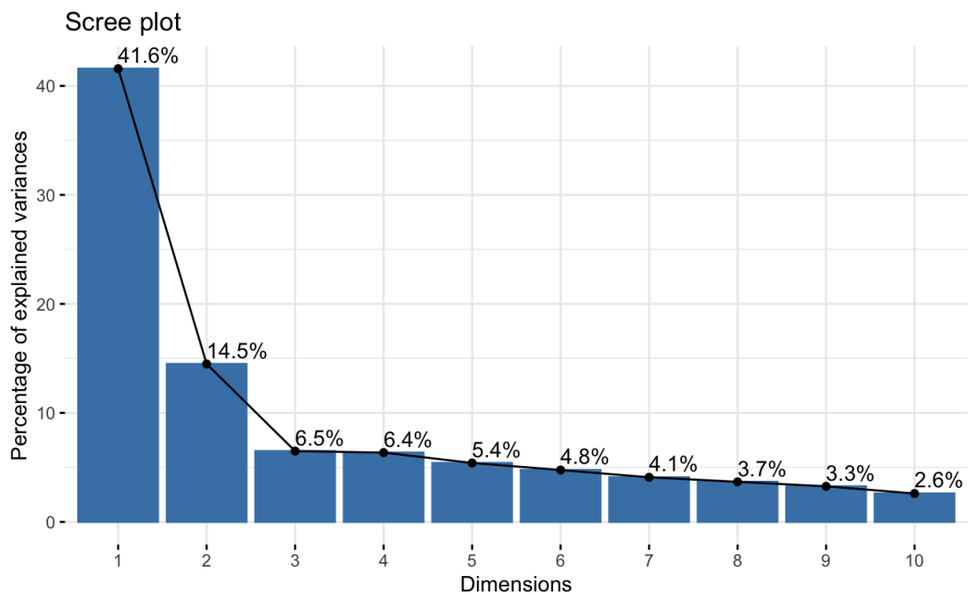


Figure 2: Explained variance by each component

We can observe how from the third point on-wards the curve clearly levels off. So that, we should extract three components for our analysis. These three components add up to an explained variance of 62.6%.

In the annexes you can find correlation and individual plots between variables and the first three dimensions. Due to the accumulation of the percentage in the first two components, we will represent the PC scores together with the loadings in a PCA-Biplot, with the first two dimensions.

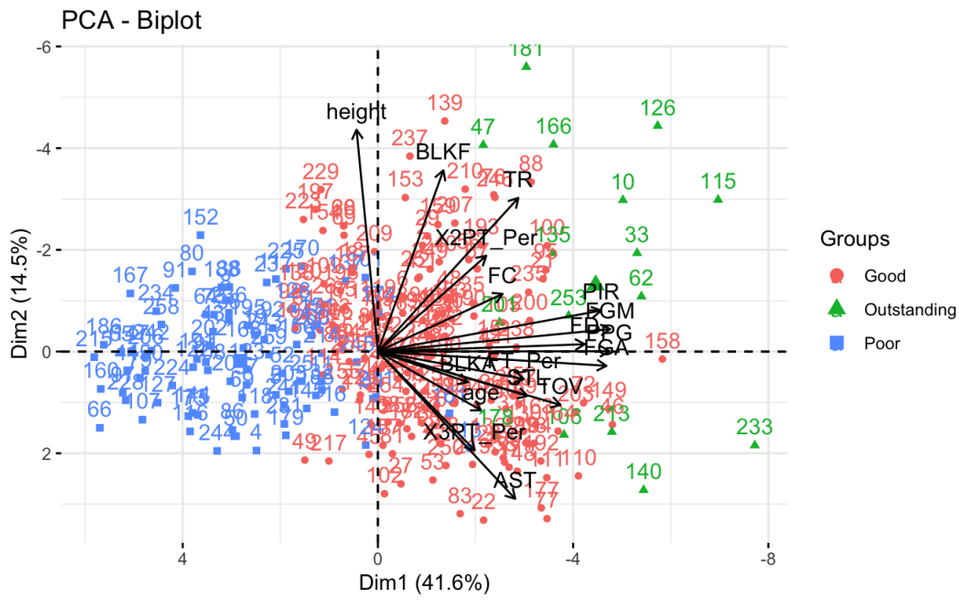


Figure 3: Data biplot with the Principal Components

From the plot we can firstly extract that the *Outstanding* class presents a higher dispersion, while the other two have a more compact structure. Naked eye, our components almost clearly separates the data along first dimension. Regarding the second one, we observe that dispersion increases as *PIR* does, i.e as the player performs better. Through the loadings of each component we can further analyze the influence of each variable of our data on each component. It has been necessary to include a reflection on both axis of our biplot, so that the distribution of observations made sense. Because of that, loadings may not be representative. Instead, we will use contributions matrix, which contains the percentage of contribution of each variable to a dimension.

The detailed information can be found in the tables located in the annexes. As a summary of the results obtained when computing them, the first dimension seems to attach importance to players which their *PIR* is based on their scoring, having high contribution values on *PPG*, *FGM*, *FGA* and *PIR*. With respect to the second dimension, it looks like representing pass-first defensive centers or power forwards, accumulating most of the percentage on *height*, *BLKF*, *TR*, *AST*. As rebounds and assists are features that contribute a lot to the *PIR*, that could explain a higher variance in the second dimension as it increases. Although it is not represented on our biplot, the third component attach importance to shooting-playmakers, because of its higher contribution values on *X3PT_Per*, *FT_Per*, *AST* and *TOV*.

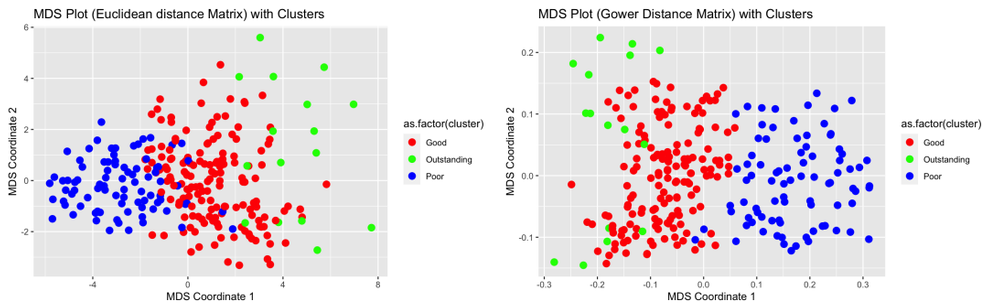
To conclude, PCA method gives us a clear separation between how good a player performs during a season. Focusing our analysis on the first two dimensions, we are able to explain a 56.1% of the variance on our data. Even so, perhaps if we had strongly considered all three dimensions as the scree plot suggested us, the results would have been better and the separation even more evident.

3.2 MDS

Multidimensional Scaling is a statistical technique used to visualize and analyze similarities or dissimilarities between a set of objects or observations. It transforms the high-dimensional data into a lower-dimensional space, while preserving the pairwise distances between the objects. It will help us to reveal patterns, relationships, and structures in the data.

Metric MDS will be applied both with Gower and Euclidean distance matrix, including position, performance and numerical variables. Other categorical variables as license and club will be not included, in order to focus our analysis into numerical variables and having the target as a label.

Representing individuals plot on the first two coordinates, we have the following graphics.



(a) Euclidean distance matrix (b) Gower distance matrix

Figure 4: Individuals plot on the first two coordinates

If we take a look on both graphics, we can observe how using Euclidean distance on numerical variables we obtain the same result as we obtained in PCA, while using a Gower distance with all variables we obtain a different classification. The quality of the separation seems to be similar, although Figure (4b) might be reflected in the y-axis so if we could apply a reflection on it we would obtain a better MDS on our data. Besides, we can see a little bit less overlap in that last representation, so we will consider that last MDS the one that better groups our data and we will use different statistics as labels in order to understand how groups are classified. For more information see annexes.

3.3 MCA

MCA, which stands for Multiple Correspondence Analysis, is a data analysis technique used to explore and analyze categorical data. MCA is an extension of Correspondence Analysis (CA) that allows the simultaneous analysis of multiple categorical variables. For this reason, in this study we will perform a MCA focusing on the following categorical variables: *pos*, *license*, *club* and *Performance*. After doing a first general study of just the categorical attributes, we will perform a second analysis using the numerical variables as supplementary. Note that in this case the results extracted from either using the Burt matrix or the indicator matrix are quite the same, nevertheless, the following plots are computed using this last one.

3.3.1 MCA using just categorical variables (primary analysis)

After converting each of the categorical variables into factors and performing a first MCA without supplementary variables let's first of all take a look at the eigenvalues.

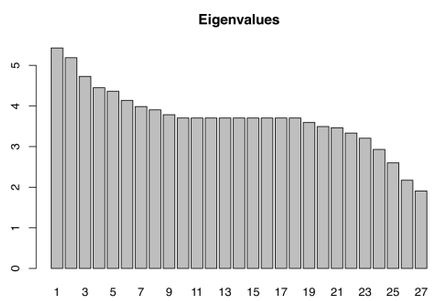


Figure 5: Eigenvalues barplot

In order to know how many dimensions should we extract we will base our conclusions on both the barplot and the average percentage of variance, which is 3.704 in this case. We conclude that it would seem reasonable to keep the 9 first dimensions as they are all above the average. Additionally, looking at the curvature of the plot itself, past the 9th dimension eigenvalues start to be less significant as they flatten until the 19th dimension, and then continue descending.

Finally, let's study the different clouds generated. As we already know when interpreting this kind of plots, the closer two points are from each other, the more similar they are in terms of their associations. Also, their proximity to the origin indicates their overall importance in the data.

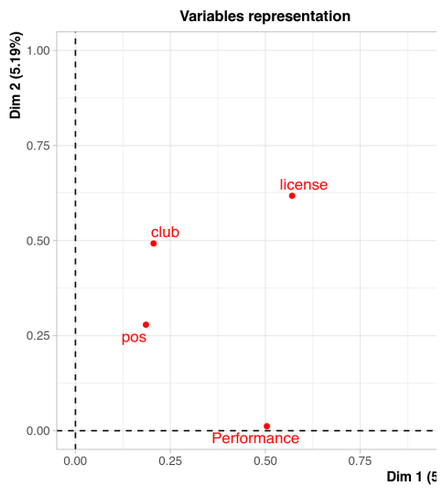


Figure 6: Cloud of variables

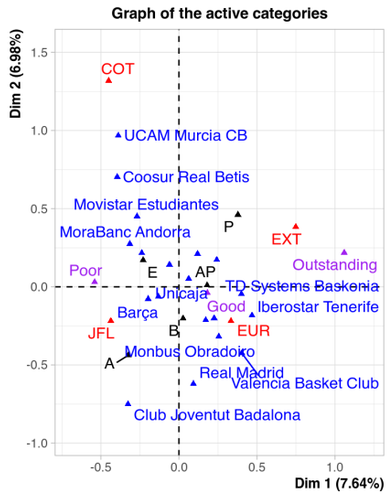


Figure 7: Cloud of categories

- Cloud of variables: As we can see in this first cloud, all of the variables are located in the positive quadrant, indicating that all of them contribute to the overall structure created by the MCA. We know that proximity to the center reflects each variable's importance, so we can see that, in this scenario, *license* might be the most important variable while *pos* is the least relevant one. We can also observe that *Performance* does not have a strong association with the second dimension, unlike the rest of the variables. Additionally, we can see how *license* and *Performance* are quite far away from each other.

- Cloud of categories: We observe that the four types of licenses are far away from each other, with each of them located in a different quadrant. Moreover, most of the categories from positions seem to be less relevant as they are close to the origin. Focusing on *Performance*, we see how the three categories are distributed across the Dimension 1 axis, with *Good* also being close to the center, *Outstanding* close to *EXT* from *license*, and *Poor* close to *JFL*. This observation suggests a relationship between having an outstanding or poor performance and having an specific type of license.

  It makes sense as foreign players (*EXT*) tend to be great players due to the limited number of players of this kind a team can have, while rookies (*JFL*) generally do not perform that well. Finally, when taking a look at teams, we can see they are spread across the whole plot, and we cannot spot any particular association.

- Cloud of individuals: The cloud of individuals is not easy to interpret without any type of coloring, as we just see the players enumerated being distributed throughout the whole map. Also, the cloud of individuals with the categories overlapped is really overloaded and can't be easily interpreted. In order to draw conclusions on their similarities, we will study the following three plots and relate them to the conclusions found in the clouds of categories and variables:

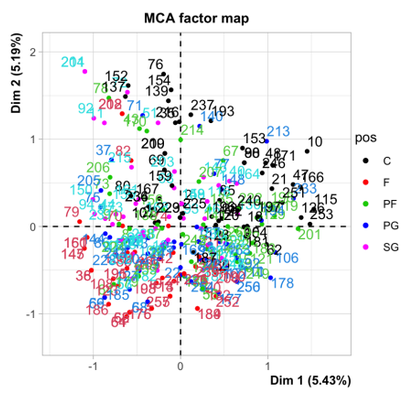


Figure 8: Cloud of individuals indicated by *pos*

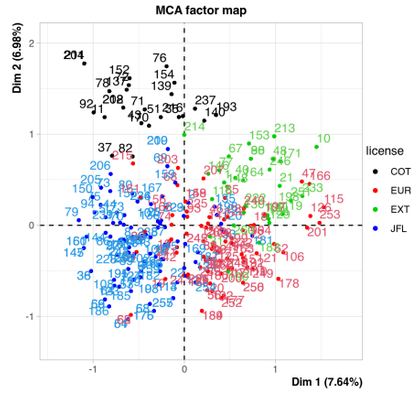


Figure 9: Cloud of individuals indicated by *license*

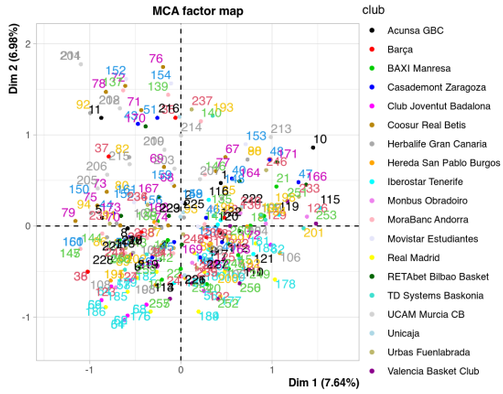


Figure 10: Cloud of individuals indicated by *club*

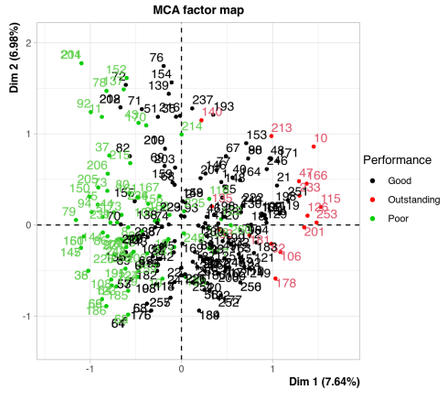


Figure 11: Cloud of individuals indicated by *Performance*

Figure 12: Cloud of individuals

– **Indicated by position**: When coloring each of the different positions, we can see that observations are grouped vertically. Even though there are some exceptions, players are spread from bottom to top in the following order: *F*, *PG*, *SG*, *PF*, and *C*. One could think that there's an association between the position and the license, where *C* would relate to *EXT* in a more significant way than others, and the same with *F* and *JFL*.

– **Indicated by license**: When indicating by license, we can see four clear clouds, each consisting of the four categories of licenses. Each of them is more or less in a different quadrant. Note that *EXT* can be found in the positive quadrant while *JFL* players are located in the negative one. Again, this makes a lot of sense, and we can see how these two types of licenses, rookie or foreign players, clearly associate with a player's overall performance.

– **Indicated by club**: Indicating by club might be the least explanatory cloud. Even though being on a good team can help in the overall performance of a player, it is not a necessary condition to stand out from the rest of the players.

– **Indicated by performance**: Finally, when filtering by Performance, we can clearly see a horizontally distributed pattern. As we saw in the cloud of categories, *Poor* performance players can be found on the left, *Good* ones in the middle, and *Outstanding* ones on the right. When relating this to the cloud of categories, we end up again with the same two conclusions: position and license affect performance.

3.3.2 MCA using numerical variables as supplementary quantitative variables

We are now going to check how numerical variables act when using them as supplementary on MCA. The only differences on this second analysis are some changes in the cloud of variables and having a new plot of the supplementary quantitative variables.

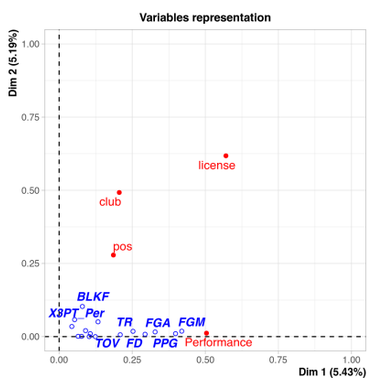


Figure 13: Cloud of variables with numerical

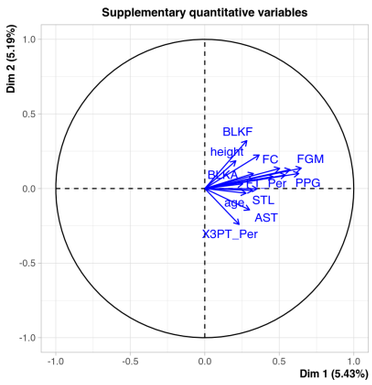


Figure 14: Plot of Supplementary quantitative variables

- Cloud of variables: We can now find the same 4 categorical variables in the same coordinates but now with the appearance of the numerical ones. We can see that all the numerical are in general closer to the origin and to each other than the categorical ones. The fact they are close to each other means they have strong associations. However, they are closer to the origin than all of the categorical ones, which makes us think they have less of an impact in the overall variability. Also, they are very close to 0 for the second dimension, meaning they almost do not contribute to it, similar to the case of *Performance*. In conclusion, numerical variables are, in general, less relevant than the categorical ones.

- Plot of Supplementary quantitative variables: On this last plot of this section we can see that all of the variables have a positive association with the first dimension, but not all of them do have it with the second one. *PPG* and *FGM* stand out as the are the ones with greater magnitude, which capture the Points Per Game and Field Goals Made.

3.4 Cluster

Cluster analysis is based on the grouping of similar data, given a definite notion of similarity. The main idea is to associate each data observation to a group, so that we can say that if two data points belong to the same group is because they are similar between them.
There are two ways we can follow: hierarchical clustering and k-means. We will conduct clusters on the different teams we have, in order to asses which teams have better performers in their roster. Arbitrarily, we could think of at least k=3 clusters having in account a separation between encompassed teams' performances. However, through the "Elbow plot" and Pseudo-F Index plot an optimal number of clusters could be defined.

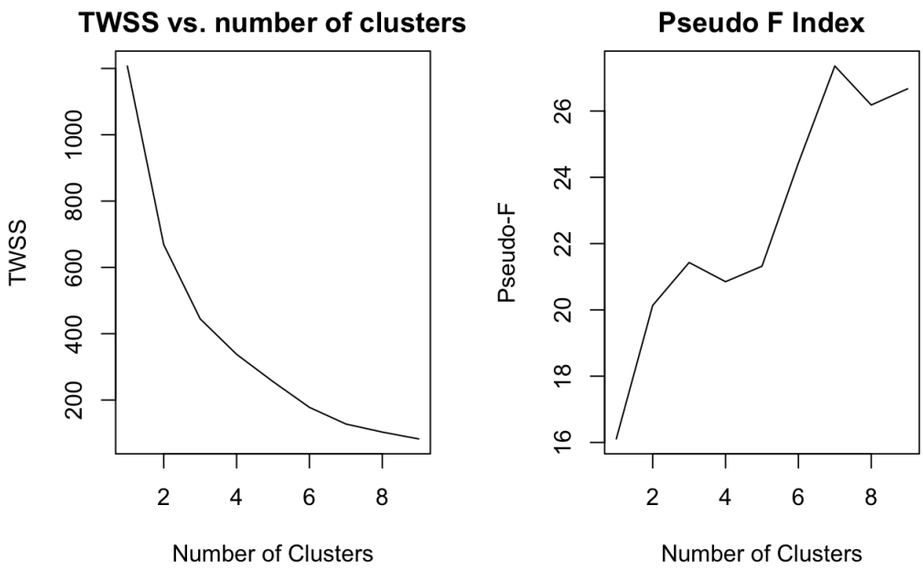
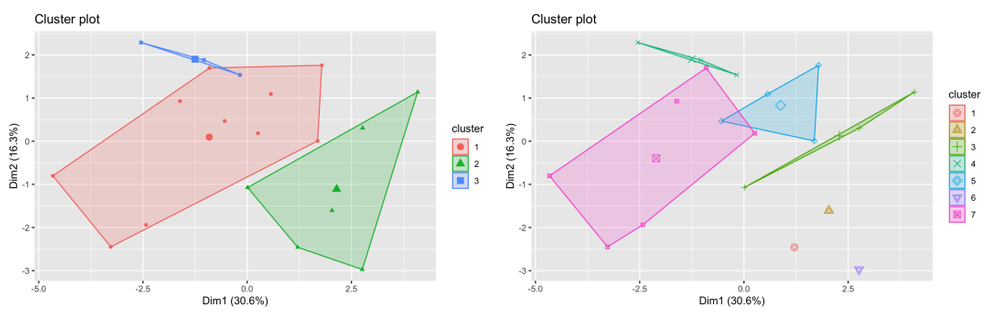


Figure 15: Elbow and Pseudo-F Index plots

From these two plots one can intuit that the optimal number of clusters could be k=7. Carrying out k-means with both k=3 and k=7, the data is split as follows.



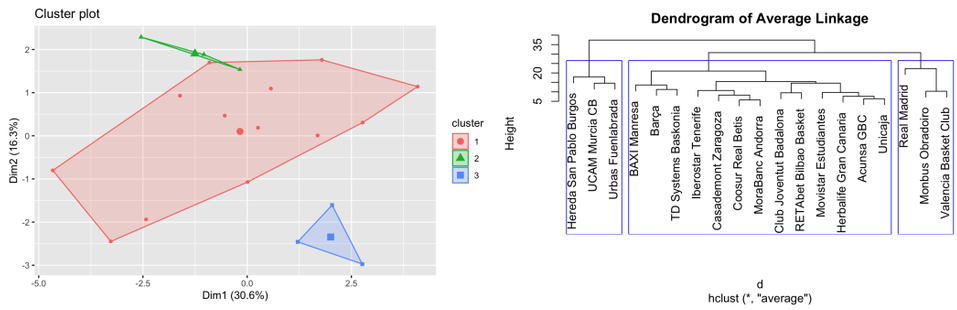
(a) k=3 (b) k=7

Figure 16: k-means result for different k values

We can observe how k=7 clusters does not bring any improvement from the first k=3 performed, there is also overlap and has clusters with an only observation, so we are taking the first one for the reason given above.

The partition more or less coincides with our data, as first dimension is the one explaining more variance and the second group is located on the positive side. Nevertheless, the variance explained by both dimensions is very low, less than 50%.

Then we will repeat the same procedure using hierarchical clustering. As the relationships between our variables is not linear, we are using Manhattan distance, which measures the distance as the sum of the absolute differences between coordinates. Regarding the method, Average Linkage computes the average distance between all pairs of points in two clusters. It strikes a balance between single and complete linkage and can be robust against outliers. After trying with other methods, we have finally chosen this one because we are interested in moderate cluster compactness, not single linkage nor complete linkage, which are the extremes. In order to compare how it performs respect to k-means, we are using k=3 clusters.



(a) Partition (b) Dendrogram

Figure 17: Hierarchical clustering performance for k=3

We can observe in the last plot how hierarchical clustering behaves similar to k-means. However, hierarchical clustering with k=3 puts together in a "middle" cluster several teams that should not belong to it, while "extreme" clusters are only composed of three teams each. If we used a larger k we might get a more reasonable separation, but we are comparing with k-means for the same k.
To conclude, cluster method provides us a good partition for our data. Beyond that, for us makes more sense the separation brought by the k-means method, although they behave in similar ways. It is dumb to think Baskonia, Barça and Madrid are not in the same cluster, at least two of them. It would not be common to have such a separation.

3.5 Discriminant Analysis

Before using discriminant analysis, we could use Multivariate Analysis as a primary step to determine the most important variables to discriminate groups. However, in our case may not be the appropriate technique since MANOVA is typically used with multiple continuous dependent variables, while in our problem we have a categorical target variable and a combination of categorical and numerical predictor variables. It is more appropriate to directly apply discriminant analysis.

The methods of discriminant analysis will allow us to classify the observations into 3 groups based on players' performances as done before. This methodology assumes that the distribution of the data conditioned on the class to which they belong follows a multivariate normal distribution. Although Shapiro-Wilks test does not satisfy it, we can observe that most of our variables present a Q-Q Plot own of a normal distribution. As variables related to percentages are not useful to predict the performance index rating of a player, and therefore not useful to predict our target variable, they are not included in our model. An important reason for not doing it is also that observations are far from following a normal distribution. Apart of this, we will assume the condition is satisfied for our data.

Based on the relationship between the covariance matrices for each class, we will distinguish between two types of discriminant analysis: linear (if we can consider that the covariance matrices are equal for all classes) or quadratic (if we cannot make this assumption). We will make this decision based on the result of the Box's Test, which has the null hypothesis that the covariance matrices are equal for all classes. Applying it, we obtain a p-value $<$ 2.2e-16, so that we consider covariance matrices to be different for our classes. Having this in account, we expect the QDA to bring us better results than LDA.

Fitting an LDA model and then predicting on our data, we obtain a Correct Classification Rate (CCR) of 96.54%. That can be seen in the confusion matrix attached below.

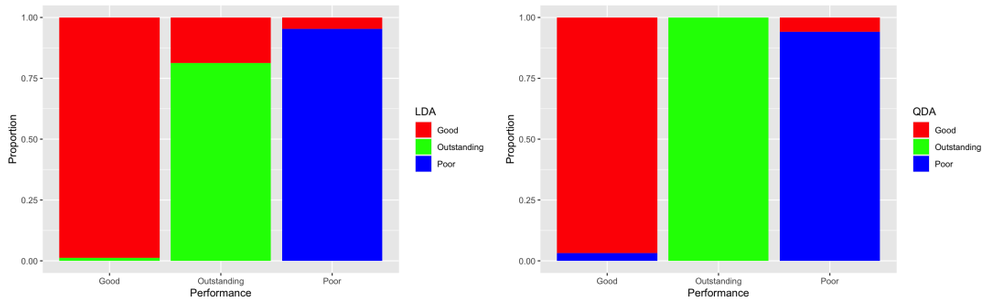
		Data		
		Good	Outstanding	poor
Prediction	Good	157	3	4
	Outstanding	2	13	0
	Poor	0	0	81

Table 1: LDA confusion matrix

Fitting a QDA model instead, and then predicting on our data, we obtain a Correct Classification Ratio (CCR) of 96.15%, a very similar percentage to the previous LDA. As before, this can be seen in the confusion matrix attached below.

		Data		
		Good	Outstanding	Poor
Prediction	Good	154	0	5
	Outstanding	0	16	0
	Poor	5	0	80

Table 2: QDA confusion matrix



(a) LDA (b) QDA

Figure 18: Classification prediction plot

We can observe that both methods present similar precision over classification, predicting slightly better on *Outstanding* performers and a little worse on *Good* performers. Because of that, we are not able to use that for making the decision between methods. We will use the Q statistic to determine how appropriate might be the QDA on our data, which is defined as follows:

$$Q = \frac{(n - \hat{n}k)^2}{(n(k-1))}$$

where n is the total observations of our dataset, $\hat{n}k$ is the average number of observations per class and k the number of classes. It is used to test the assumption of equal covariance matrices across the classes. The value obtained is 461.73, which is way too large. It suggests that the covariance matrices are not equal across classes, possibly breaching the assumption made before.

Therefore, although we initially thought that QDA should give us the best results according to the Box's Test, we have been able to observe that LDA fits better our data and brings us the best precision over our predictions. This could be because LDA is a more flexible model and usually has less variance. Typically, this can lead to bias in the model if the covariance matrices are different (as they are).

4 CONCLUSIONS

The selected data contain a series of variables that are related to the impact of a basketball player based on his performance, from whom the samples were taken. In our analysis, we tried to find ways to reproduce this classification, with the aim of extending the use of the model to explain any factor influencing basketball players' performance.

PCA and MDS analysis revealed distinct dimensions that effectively separated player performance categories. PCA-Biplot's first dimension primarily represented scoring-related statistics, while the second dimension indicated the influence of statistics such as *height*, *blocks*, *rebounds*, and *assists*. MDS analysis using Gower distance on all variables presented a slightly better classification than PCA did, showing similarities and dissimilarities between players.

The MCA analysis focused on examining the categorical variables: *position*, *license*, *club*, and *Performance*. The results revealed, overall, the important association between *license* and *Performance*, and the distinctiveness of different license types and player positions. Lastly, we saw that *club* did not have such a relevant role.

Furthermore, LDA and QDA were utilized to classify players into performance categories. Firstly, we have seen that we can consider that the covariance matrices are different, given the result of the Box's Test. However, we have found that the LDA gives better results than the QDA, showing that a linear combination of variables effectively discriminated between different performance levels.

Overall, the combination of PCA, MDS, MCA, LDA, and QDA techniques provided valuable insights into the statistical factors determining basketball player performance in the ACB league.

5 BIBLIOGRAPHY

Bibliography

[Aca23] Nihan Acar-Denizli. *Notes from the Data Analysis course*. Atenea, 2023.

[Wik] Wikipedia. *Performance Index Rating*. URL: https://en.wikipedia.org/wiki/Performance_Index_Rating.

[Zen] Zenodo. *Estadísticas de jugadores ACB*. URL: https://zenodo.org/record/4243039#.ZGseRS8lOpY. (November 3, 2020).

6 ANNEXES

In the annexes section supplementary material can be found, such as graphics and tables that have been mentioned or other models that have been considered although they have not been added to the main part of the project. Graphics that have not been commented will be attached in the delivered folder, classified by sections.

Preprocessing

We have not included in our preprocessing section a correlation heatmap, which is a graphical representation of the correlation coefficients between variables in a dataset and provides a visual summary of the strength and direction of the linear relationship between pairs of variables.

In the heatmap attached below we can see there is a strong correlation between variables involving the same aspects of the game, as they are PPG with FGM and FGA. That makes sense since without attempting and making field goals a player can not score. Even though, these variables could not be removed or unified because they take part individually in the PIR formula, in which is based our target variable.

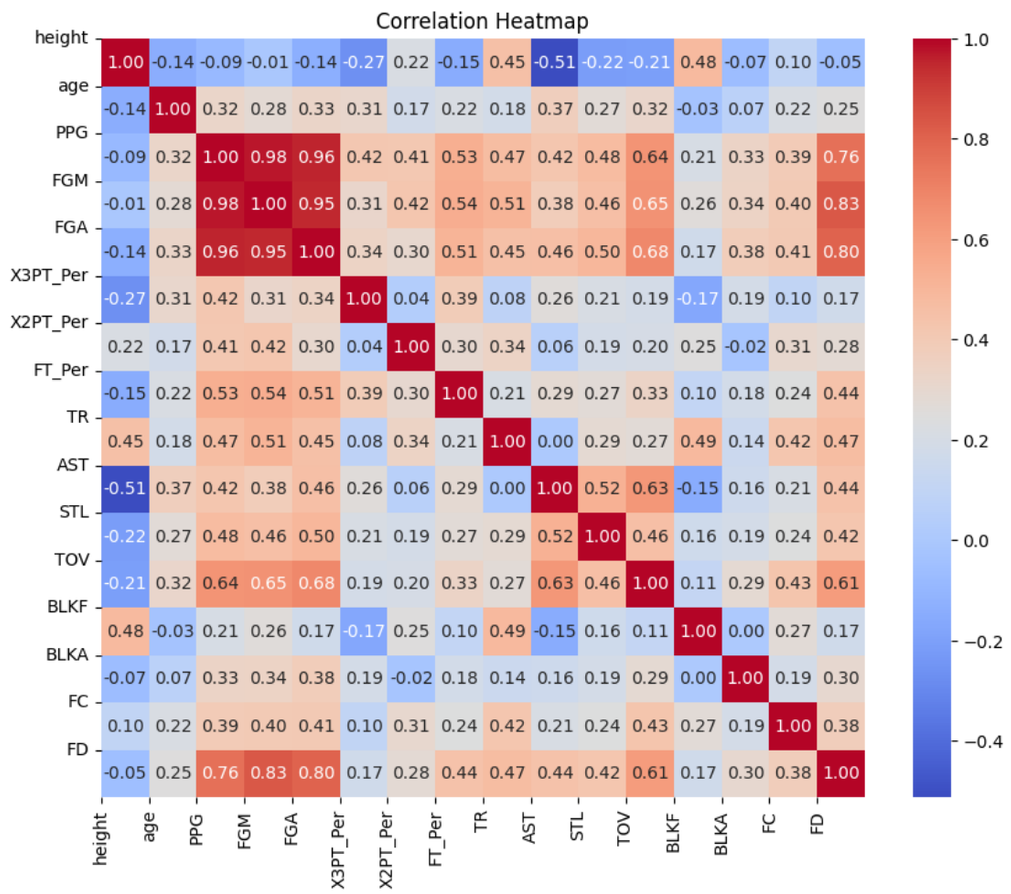


Figure 19: Correlation HeatMap

PCA

Here you can find attached the loadings and contributions numerical outputs, in which we have based part of our principal components analysis. Besides that, in the delivered folder you can find both variance PCA graphic and biplots between dimensions 1 and 3, and dimensions 2 and 3.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
height	0.08693235	-0.865025937	0.063719261	-0.037006900	-0.12480966	-0.09670538
age	-0.41891981	0.228966437	0.099185123	0.508313868	-0.46366472	-0.18672186
PPG	-0.94067695	-0.001282871	0.140728334	-0.111811393	0.10216935	0.02038290
FGM	-0.94389164	-0.087641121	0.085462707	-0.149066975	0.14442395	0.05259522
FGA	-0.92847266	0.055552599	0.016891887	-0.178501466	0.07257999	0.02592755
X3PT_Per	-0.39330822	0.384532177	0.606216532	0.007246656	-0.27392063	-0.24708446
X2PT_Per	-0.43999235	-0.373600792	0.217092863	0.382312345	0.11304582	0.43546224
FT_Per	-0.59348062	0.107240334	0.433330501	0.012163500	0.07808654	0.22361012
TR	-0.57048711	-0.599161079	-0.003188363	0.038180386	-0.15189797	-0.25917025
AST	-0.55789437	0.574211839	-0.337256158	0.229273092	0.05583263	-0.02120756
STL	-0.60776484	0.172919851	-0.300877983	0.200611339	0.09864201	-0.35663697
TOV	-0.74217875	0.208348546	-0.372047151	0.005293317	-0.05944566	0.15008538
BLKF	-0.26600802	-0.706690223	-0.178224798	0.064671276	0.01146751	-0.20678667
BLKA	-0.36651595	0.119202284	-0.074161935	-0.684796700	-0.41413581	-0.01191089
FC	-0.50665433	-0.224919972	-0.248999997	0.117460794	-0.51753506	0.40143902
FD	-0.83998333	-0.028772349	-0.102586827	-0.155806284	0.17141400	0.05116802
PIR	-0.90713273	-0.159966420	0.088206329	0.051564321	0.22252826	-0.18978129

	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12
height	-0.128399718	-0.028395472	0.020860584	-0.01688645	0.23119529	-0.3020171956
age	-0.344294571	-0.104409118	0.325108571	-0.02257582	-0.13362101	-0.0115316285
PPG	-0.097737017	0.026543157	-0.053910380	0.14480943	-0.11672555	-0.0310886871
FGM	-0.123065404	0.033656533	0.002699150	0.07946866	-0.10181216	-0.0401779087
FGA	-0.119665671	0.064189125	0.003415804	0.07251118	-0.16154465	-0.0647843546
X3PT_Per	0.171095058	0.126535436	-0.267616755	0.21010493	0.12717045	-0.0001678994
X2PT_Per	0.004829658	-0.477997696	-0.134670568	0.09684019	0.06094246	0.0713940175
FT_Per	0.387109776	0.101486733	0.378060434	-0.26730972	0.07724369	-0.0961702931
TR	-0.060926228	0.082737407	-0.159970034	-0.27202445	0.13523249	0.1395949976
AST	0.053890571	-0.008036577	0.025198888	-0.01675112	0.32206750	0.1599394255
STL	0.358842247	-0.273253517	-0.165760162	-0.13079282	-0.17283187	-0.2303654446
TOV	-0.056365020	0.107071469	0.069289979	0.22121904	0.24467355	-0.2201156411
BLKF	0.338762556	0.076116670	0.299577005	0.31767986	-0.06087962	0.1772008482
BLKA	0.063515878	-0.422907719	0.106605678	-0.02391310	0.05552506	0.0841222211
FC	0.197055985	0.248009304	-0.221015064	-0.07183546	-0.14198543	0.0295709914
FD	-0.204963807	0.113558315	0.030407632	-0.18927814	-0.02348196	0.0700522385
PIR	-0.065610557	-0.018898122	-0.060691897	-0.04153890	0.06991094	0.1260404747

	Dim.13	Dim.14	Dim.15	Dim.16	Dim.17
height	0.212398355	0.075575127	0.017331470	0.0031719456	2.719569e-08
age	0.008152494	-0.012909018	-0.011980240	-0.0006368246	-1.511493e-07
PPG	-0.014960567	0.132730679	-0.041737292	0.0604725622	4.890656e-04
FGM	0.016068455	0.066357689	-0.052680196	-0.0788734703	2.676522e-04
FGA	-0.039854870	0.111702710	0.169376135	0.0025481185	-4.498610e-04
X3PT_Per	0.064708163	-0.107827324	0.027136826	-0.0062628894	-2.339504e-08
X2PT_Per	-0.027472037	-0.064774998	0.036729285	-0.0006890501	4.383653e-09
FT_Per	-0.043464066	0.024300978	-0.003795094	0.0033010060	-5.667122e-08
TR	-0.271526920	0.011954612	0.045086985	-0.0042306933	1.841060e-04
AST	0.159685465	0.165620498	0.054791306	-0.0059629635	1.588275e-04
STL	0.021098331	-0.040265282	0.001507071	-0.0013240076	5.356809e-05
TOV	-0.213098507	-0.132952782	-0.038157582	0.0039157147	-8.662718e-05
BLKF	0.034511546	-0.053707278	0.024233576	-0.0001782549	3.953817e-05
BLKA	0.009135794	0.000672343	-0.017099442	0.0015250080	-2.614351e-05
FC	0.096936622	0.042566406	-0.036431517	0.0008297889	-9.039906e-05
FD	0.201056604	-0.296215537	0.037427140	0.0107775561	1.267504e-04
PIR	0.048520714	0.046827846	-0.141518437	0.0112450441	-5.714740e-04

Figure 20: Loading values

	PC1	PC2	PC3	PC4	PC5	PC6
height	0.106913	3.037972e+01	3.677791e-01	0.126705827	1.69527448	1.15630664
age	2.482729	2.128478e+00	8.911251e-01	23.905349409	23.39654407	4.31084172
PPG	12.518392	6.681767e-05	1.793943e+00	1.156653553	1.13601707	0.05136929
FGM	12.604099	3.118468e-01	6.616052e-01	2.055862450	2.26998016	0.34203011
FGA	12.195673	1.252952e-01	2.584651e-02	2.947913612	0.57329407	0.08311793
X3PT_Per	2.188435	6.003311e+00	3.328903e+01	0.004858548	8.16569365	7.54853379
X2PT_Per	2.738784	5.666842e+00	4.269104e+00	13.522837251	1.39076140	23.44621382
FT_Per	4.982882	4.669193e-01	1.700921e+01	0.013688262	0.66358414	6.18236513
TR	4.604253	1.457514e+01	9.208335e-04	0.134868896	2.51100467	8.30504565
AST	4.403231	1.338658e+01	1.030304e+01	4.863370465	0.33924995	0.05561002
STL	5.225630	1.213990e+00	8.200238e+00	3.723420886	1.05893034	15.72622336
TOV	7.792639	1.762408e+00	1.253838e+01	0.002592311	0.38457748	2.78514960
BLKF	1.001053	2.027606e+01	2.877278e+00	0.386949282	0.01431138	5.28709111
BLKA	1.900436	5.768922e-01	4.982046e-01	43.386488009	18.66505085	0.01754122
FC	3.631542	2.053911e+00	5.616223e+00	1.276488910	29.14896603	19.92557646
FD	9.981800	3.361059e-02	9.532983e-01	2.245955345	3.19768865	0.32371953
PIR	11.641507	1.038923e+00	7.047663e-01	0.245996983	5.38907161	4.45326463

	PC7	PC8	PC9	PC10	PC11	PC12
height	2.369787271	0.12902797	0.078434702	0.06439150	13.6569369	2.857574e+01
age	17.038900735	1.74446609	19.050754562	0.11509040	4.5618902	4.165963e-02
PPG	1.373091206	0.11274332	0.523841781	4.73528022	3.4811857	3.027882e-01
FGM	2.176973714	0.18126937	0.001313135	1.42607957	2.6484676	5.057185e-01
FGA	2.058355507	0.65933874	0.002103009	1.18730407	6.6677679	1.314845e+00
X3PT_Per	4.207810310	2.56218220	12.908683249	9.96837695	4.1320710	8.831462e-06
X2PT_Per	0.003352848	36.56261802	3.268893157	2.11769435	0.9489325	1.596827e+00
FT_Per	21.540188612	1.64817842	25.761894737	16.13546462	1.5244788	2.897451e+00
TR	0.533568038	1.09544225	4.612460061	16.70966912	4.6725869	6.104834e+00
AST	0.417452244	0.01033541	0.114450607	0.06336358	26.5026401	8.013924e+00
STL	18.509230316	11.94860487	4.952399485	3.86295602	7.6320813	1.662527e+01
TOV	0.456667969	1.83456542	0.865358769	11.05088336	15.2957008	1.517875e+01
BLKF	16.495748490	0.92713983	16.176043056	22.78931850	0.9469764	9.837070e+00
BLKA	0.579890282	28.62047932	2.048405887	0.12912894	0.7877227	2.216949e+00
FC	5.581622769	9.84286718	8.804386401	1.16527862	5.1508974	2.739466e-01
FD	6.038590507	2.06359068	0.166655847	8.09008157	0.1408847	1.537369e+00
PIR	0.618769182	0.05715089	0.663921555	0.38963863	1.2487789	4.976846e+00

	PC13	PC14	PC15	PC16	PC17
height	17.85859520	2.880374e+00	0.455966602	9.804647e-02	7.903796e-08
age	0.02631031	8.403835e-02	0.217867882	3.952030e-03	2.441451e-06
PPG	0.08860150	8.884515e+00	2.644300071	3.563668e+01	2.556053e+01
FGM	0.10220997	2.220617e+00	4.212665557	6.062370e+01	7.655566e+00
FGA	0.62879314	6.292432e+00	43.547842703	6.327319e-02	2.162680e+01
X3PT_Per	1.65753663	5.863390e+00	1.117841749	3.822346e-01	5.849022e-08
X2PT_Per	0.29876293	2.115952e+00	2.047798175	4.626814e-03	2.053563e-09
FT_Per	0.74783485	2.978095e-01	0.021862879	1.061874e-01	3.432106e-07
TR	29.18574487	7.207125e-02	3.085777397	1.744229e-01	3.622191e+00
AST	10.09428981	1.383309e+01	4.557068010	3.465013e-01	2.695796e+00
STL	0.17621432	8.176220e-01	0.003447703	1.708289e-02	3.066536e-01
TOV	17.97652766	8.914273e+00	2.210160963	1.494179e-01	8.019439e-01
BLKF	0.47149144	1.454646e+00	0.891450497	3.096445e-04	1.670585e-01
BLKA	0.03303981	2.279675e-04	0.443839665	2.266338e-02	7.304044e-02
FC	3.71980997	9.137451e-01	2.014729439	6.709895e-03	8.733002e-01
FD	16.00227313	4.424934e+01	2.126353630	1.131934e+00	1.716857e+00
PIR	0.93196447	1.105858e+00	30.401027077	1.232261e+00	3.490025e+01

Figure 21: Contribution values

Cluster

Apart from the Complete Linkage dendrogram, which has been considered but not finally chosen, and the clusters plot in principal components, here you can find the centroids table after performing k-means with k=3. The mentioned graphics will be attached in the folder.

Group.1 <int>	height <dbl>	age <dbl>	PPG <dbl>	FGM <dbl>	FGA <dbl>	X3PT_Per <dbl>	X2PT_Per <dbl>	FT_Per <dbl>	TR <dbl>
1	1.988054	26.5496	6.624708	3.514566	6.851172	25.68329	43.61541	64.47888	2.559593
2	1.997855	27.4162	7.177708	3.786394	7.008479	31.25763	52.66048	69.81125	2.755481
3	1.989392	27.1268	6.644532	3.497284	6.827064	18.40546	46.65011	47.38085	2.658785

AST <dbl>	STL <dbl>	TOV <dbl>	BLKF <dbl>	BLKA <dbl>	FC <dbl>	FD <dbl>	PIR <dbl>
1.354200	0.5882295	1.159169	0.2097599	0.2066020	1.753511	1.694076	6.573974
1.505631	0.6973669	1.239730	0.1903728	0.2498754	1.677012	1.685912	7.623899
1.341903	0.7953404	1.084967	0.2242956	0.1761991	1.870982	1.666384	6.868554

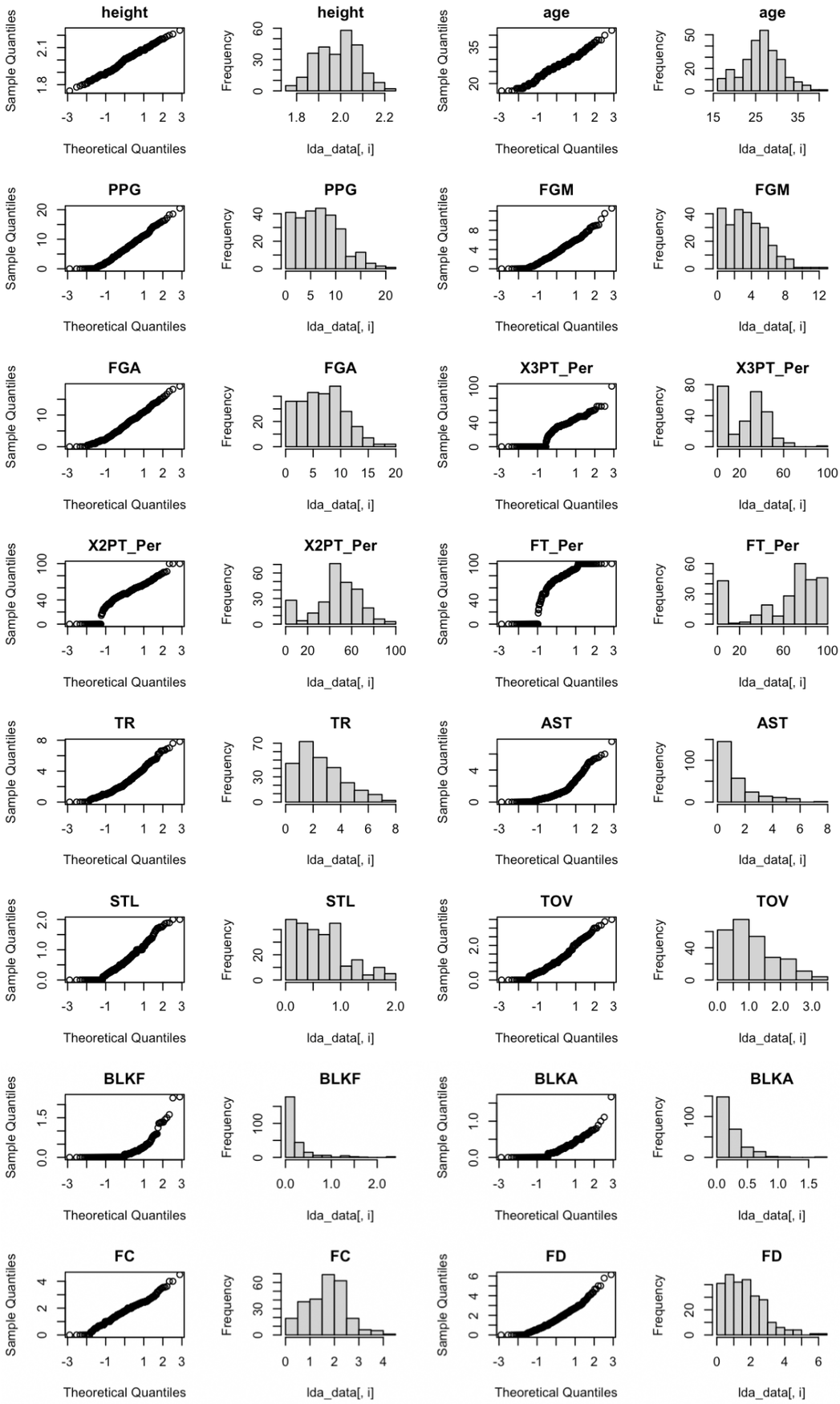
Figure 22: Centroids (means) of the clusters

Discriminant Analysis

Here you can initially find the Shapiro-Wilks test result for each variable, and then in the following page the normality plots for each variable. Combining both, we discussed the normality assumption in the Discriminant Analysis previous section.

height	age	PPG	FGM	FGA	X3PT_Per	X2PT_Per	FT_Per
0.09287695	0.007011567	8.748165e-05	5.040191e-06	0.0003678165	4.994518e-12	1.062981e-09	6.999551e-16
TR	AST	STL	TOV	BLKF	BLKA	FC	FD
3.028106e-07	1.683332e-16	7.236316e-08	5.482856e-07	1.781161e-23	2.315287e-16	0.02318527	1.004671e-07

Table 3: Shapiro-Wilks test results


height
Sample Quantiles
Theoretical Quantiles
Frequency
lda_data[, i]
age
PPG
FGM
FGA
X3PT_Per
X2PT_Per
FT_Per
TR
AST
STL
TOV
BLKF
BLKA
FC
FD

Regarding the fit of the models, here you can also find the prior distribution for each group when fitting them, besides of the group means displayed in a table as follows:

Poor	Good	Outstanding
0.32692308	0.61153846	0.06153846

Table 4: Prior probabilities of groups

```
Group means:
              height      age       PPG      FGM       FGA       TR       AST       STL
Good        1.987547 27.93711  8.371132 4.385409  8.270692 3.023648 1.7596226 0.7838994
Outstanding 2.039375 27.93750 14.165625 7.995625 12.555625 5.180625 2.3343750 1.0756250
Poor        1.989765 24.85882  2.440824 1.262235  3.158471 1.418353 0.5485882 0.3352941
                  TOV       BLKF      BLKA       FC        FD
Good        1.3315723 0.22911950 0.2463522 1.921195 2.0101258
Outstanding 2.0037500 0.60250000 0.2037500 2.095000 3.6050000
Poor        0.7050588 0.08988235 0.1563529 1.343882 0.6925882
```

Figure 23: Group means

Gaussian Naive Bayes classifier

Fitting a Naive Bayes classification model we obtain the following predictions based on our original data:

		Data		
		Good	Outstanding	Poor
Prediction	Good	127	11	21
	Outstanding	4	12	0
	Poor	10	0	75

Table 5: Naive Bayes confusion matrix

$$\text{CCR} = 82{,}31\%$$

Step-wise classification

Fitting a step-wise classification model through the following command:

$stepclass(lda\_data[, 1:13], lda\_data\$Performance, method = "qda", direction = "backward", criterion = "CR")$

we obtain the following numerical output, model and process tables:

```
 `stepwise classification', using 10-fold cross-validated correctness rate of method qda'.
260 observations of 13 variables in 3 classes; direction: backward
stop criterion: improvement less than 5%.
correctness rate: 0.84615;  starting variables (13): height, age, PPG, FGM, FGA, TR, AST,
STL, TOV, BLKF, BLKA, FC, FD
correctness rate: 0.86154;  out: "AST";  variables (12): height, age, PPG, FGM, FGA, TR, STL,
TOV, BLKF, BLKA, FC, FD
correctness rate: 0.87692;  out: "age";  variables (11): height, PPG, FGM, FGA, TR, STL, TOV,
BLKF, BLKA, FC, FD
correctness rate: 0.88077;  out: "height";  variables (10): PPG, FGM, FGA, TR, STL, TOV,
BLKF, BLKA, FC, FD
correctness rate: 0.89231;  out: "FGM";  variables (9): PPG, FGA, TR, STL, TOV, BLKF, BLKA,
FC, FD
correctness rate: 0.90385;  out: "BLKA";  variables (8): PPG, FGA, TR, STL, TOV, BLKF, FC, FD

 hr.elapsed min.elapsed sec.elapsed
      0.000       0.000       0.781

crossval.rate      apparent
   0.90384615    0.06153846
```

Figure 24: Step-wise classification numerical output

Process				
	step	**var**	**varname**	**result.pm**
0	start	0	–	0.8461538
1	out	7	AST	0.8615385
2	out	2	age	0.8769231
3	out	1	height	0.8807692
4	out	4	FGM	0.8923077
5	out	11	BLKA	0.9038462

Model	
nr	**name**
3	PPG
5	FGA
6	TR
8	STL
9	TOV
10	BLKF
12	FC
13	FD