Exercise 3.22

Carlos Fernández Pascual

February 17th 2025

The dataset la-liga-2015-2016.xlsx contains team statistics for La Liga season 2015/2016.

a) How many clusters are detected for the variables Yellow.cards and Red.cards? And for Yellow.cards, Red.cards, and Fouls.made? Interpret the results.

First of all, the function ks::kms was employed to implement kernel mean shift clustering on the data, taking into account only the variables *Yellow Cards* and *Red Cards* (without precomputing the bandwidth matrix H, and letting the function compute it internally using its default. Then, the process was repeated also taking into account the variable *Fouls Made*.

For the variables *Yellow Cards* and *Red Cards*, 6 different clusters were found, whereas, when variable *Fouls Made* was also taken into account, only three distinct clusters were found for the algorithm. Tables 1 and 2 show the size for each of the clusters found in both cases, respectively.

Cluster	1	2	3	4	5	6
Size	1	2	11	2	3	1

Table 1: Results for two variables

Cluster	1	2	3
Size	1	1	18

Table 2: Results for three variables

Figures 1 and 2 show the results visually by plotting the data color-coded by their respective clusters.

From the fact that the number is decreased when the new variable *Fouls Made* is taken into account, it may be concluded that this variable even out the data, making teams look more similar to each other, and thus eliminating what appeared to be greater differences in the previous case. In fact, adding

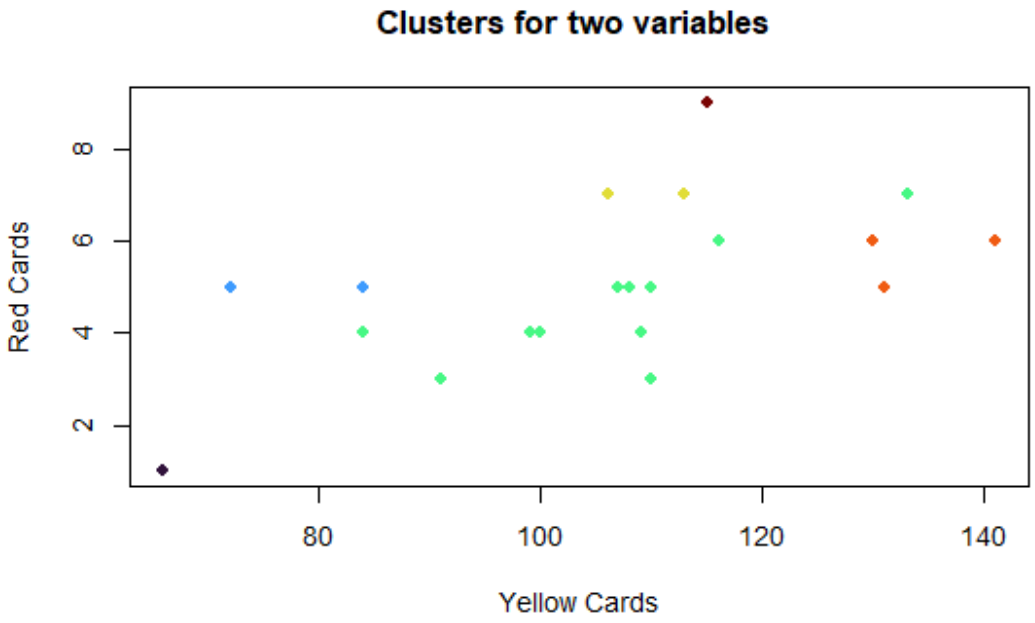

Figure 1: Clusters for two variables

the variable, we have a cluster where most of the teams are included, and then two separate clusters with only one team each: FC Barcelona and Real Madrid, unsurprisingly. This makes sense when exploring the data, as most teams have more a less a similar number of fouls made across the competition, and the first two have a significantly lower number, showing their superiority ans, especially, their control of the game.

There are two main options in this case. On one hand, the nature of the cards as exceptional events (especially in the case of red cards, the maximum number is only 8), may throw us into identifying false subsets in the data. The variable *Fouls Made*, having a greater range of values and closer to a continuous nature, may then have been a better insight into the reality of the game, evening out extreme events by chance. On the other hand, as fouls are committed by all teams in a regular manner (it is a part of the game, and may not always even be a negative aspect), it may also be that, if the behavior of all teams in this regard is approximately the same, that the variable has no relevant information to discerning between groups.

From a more technical point of view, the values for the variables *Fouls made* provide a more uniform distribution of the data in the support for the theo-

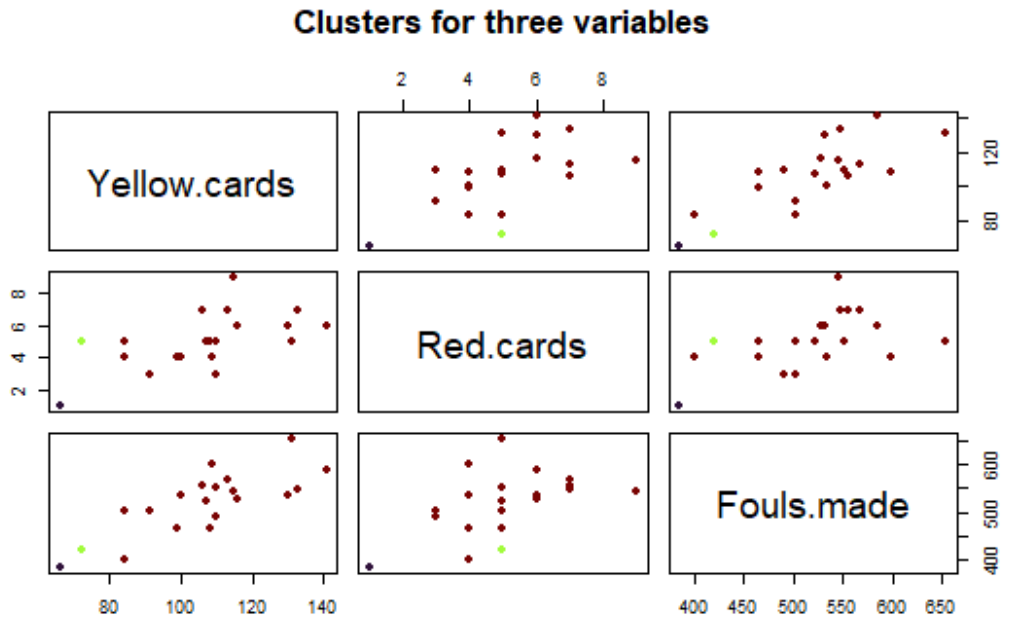

Figure 2: Clusters for three variables

retical density function of the distribution, and thus the modes of such density function are less clear than in the previous case, and some of them cease to be so (let's remember that the modes are the key points in kernel mean shift clustering, as they act as gravity centers for the data points). This way, when estimating the density function from the kernel approach, less modes are obtained.

b) Standardize the previous variables (divide them by their standard deviation) and recompute a. Are the results the same? Why? Does k-means have a similar behavior?

When employing mean shift clustering, the results are the same when using the original or the standardized data, obtaining exactly the same clusters with the same elements (both when dealing with two and with three variables). However, when employing the k-means algorithm, the results are vastly different.

This happens because, the mean shift clustering is based upon the **distribution** of the data points, and the estimation of their density function. This way, if the data are standardized (rescaled), this distribution remains the same, only constraining its support to a smaller region of $\mathbb{R}^p$, the nature of the distribution stays the same. However, the k-means algorithm is based upon minimizing the

within-cluster variation, which is defined as a weighted combination of the **euclidean distances** between the elements of the cluster. The euclidean distance is heavily influenced by the scale of the data and, if they are not standardized, the variables with a greater scale dominate. This discrepancy shows a strong point of mean shift clustering against the classical k-means algorithm.

Figures 3 and 4 show the difference between the clusters obtained between the original data and the standardized data for the case of two variables.

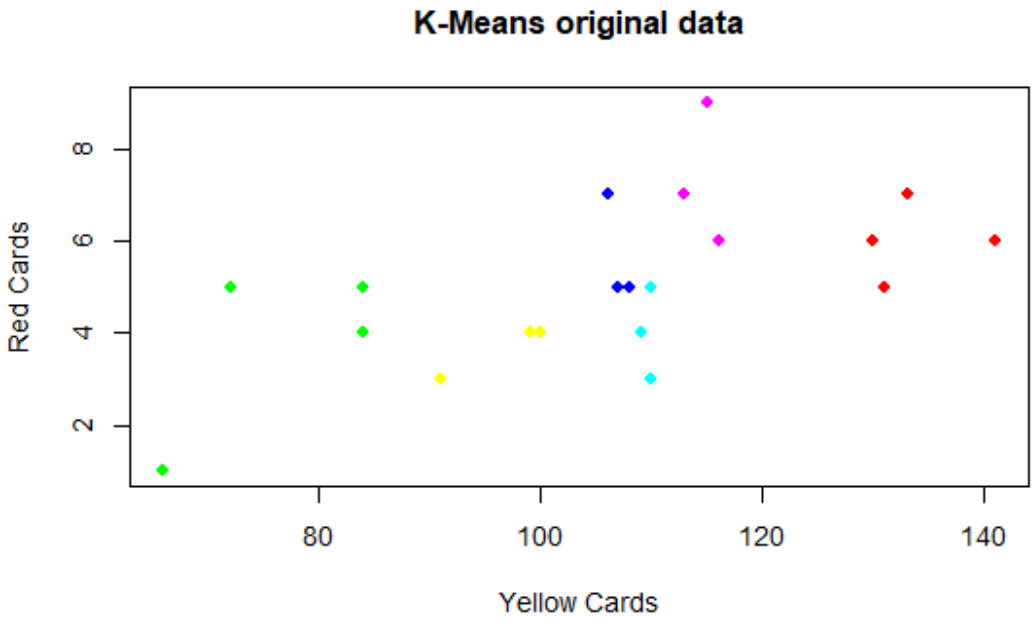

Figure 3: Original variables

c) Run a PCA on the dataset after removing Points and Matches and standardizing the variables. Then perform a clustering on the scores of as many PCs as to explain the 85% of the variance. How many clusters are detected? What teams are associated with each of them? Are the clusters interpretable? Do you see something strange?

After running a PCA on the dataset (having previously removed the variables *Points* and *Matches* and standardizing the remainder of the variables), it was found that 3 principal components were enough to explain the desired percentage of variance. Table 3 shows the clusters found and their respective sizes.

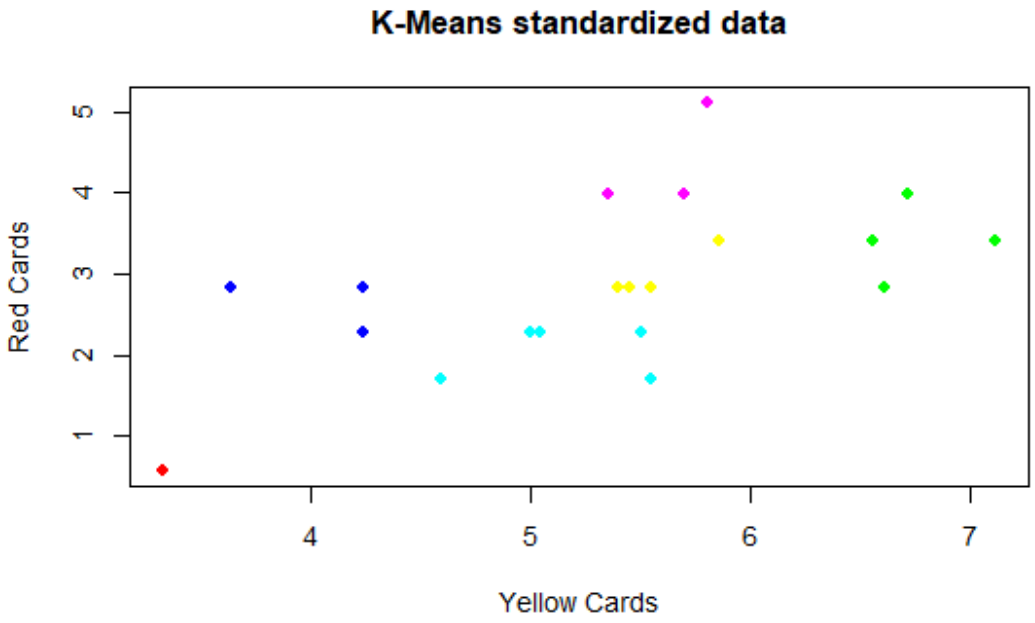

Figure 4: Standardized variables

Cluster	1	2	3	4
Size	2	1	16	1

Table 3: Clusters for the principal components

As for the teams in each cluster, FC Barcelona and Real Madrid belong to one cluster, Atlético de Madrid to another, RC Deportivo de La Coruña to another, and the rest of the teams are contained in the same cluster.

It does makes sense that the three top teams are clearly differentiated from the rest. Going to the data, they ended up the season with 91, 90, and 88 points, respectively, and the next team was Villarreal, 24 points behind the top teams. This exemplifies the dominance of those three teams that season, very far from the rest.

As for the separation between Atlético de Madrid and the other two, it also makes sense, taking into account their style of play: Real Madrid and Barcelona are way more offensive than Atlético de Madrid, which prefers to sit back and rely on their robust defense. This can be clearly seen by its values for variables like goals scored and conceded and shots made, for example.

However, something strange happens. RC Deportivo de La Coruña, sitting at 15th place, is placed into a cluster of their own. Once the data is analyzed, one finds that Dépor's season was an odd one. They were the team with the most amount of ties (18 out of 38 games, almost half, having 6 more ties than the second), and were the team with the least amount of wins (8, tied with the last team in the league). They were also the second team that less fouls committed (401, only behind the champions, FC Barcelona), and they were the third team that received less yellow cards (only 84, in values of the top teams). This unusual data ensured that they were classified in a cluster of they own.

d) Run kmeans on the data used in c) with k=3,4,5and compare the results graphically.

Figures 5, 6 and 7 show the results for three, four and 5 clusters for the algorithm k-means, respectively.

With three clusters, one of them contains the top two teams, FC Barcelona and Real Madrid. Then, the next ten teams are found in another cluster, accompanied by RC Deportivo (with the exception of RC Celta and Sevilla FC), and another cluster contains the last teams (and the aforementioned exceptions).

With four clusters, the results are exactly the same, only that Atlético de Madrid has its own separate cluster, exemplifying the huge gap between the top teams and the rest of the league, but still showing a difference between the top two teams and Atlético de Madrid.

With five clusters, Real Madrid and Barcelona are again in their cluster, with Atlético de Madrid being found in a cluster with the two teams that followed them, Villarreal and Athletic Club de Bilbao, The other three clusters are more mixed and harder to interpret, but in general there is a cluster for the last teams, and another two for the middle of the table. It is worth mentioning that RC Deportivo is again being grouped with teams way higher than them in the table.

In conclusion, there was a huge difference between FC Barcelona, Real Madrid and Atlético de Madrid and the rest of the teams, and even a big difference between the top two teams and Atlético de Madrid. Moreover, RC Deportivo was found to have a very strange season, with unusual results, ensuring it almost always was highlighted in the analysis of the data.

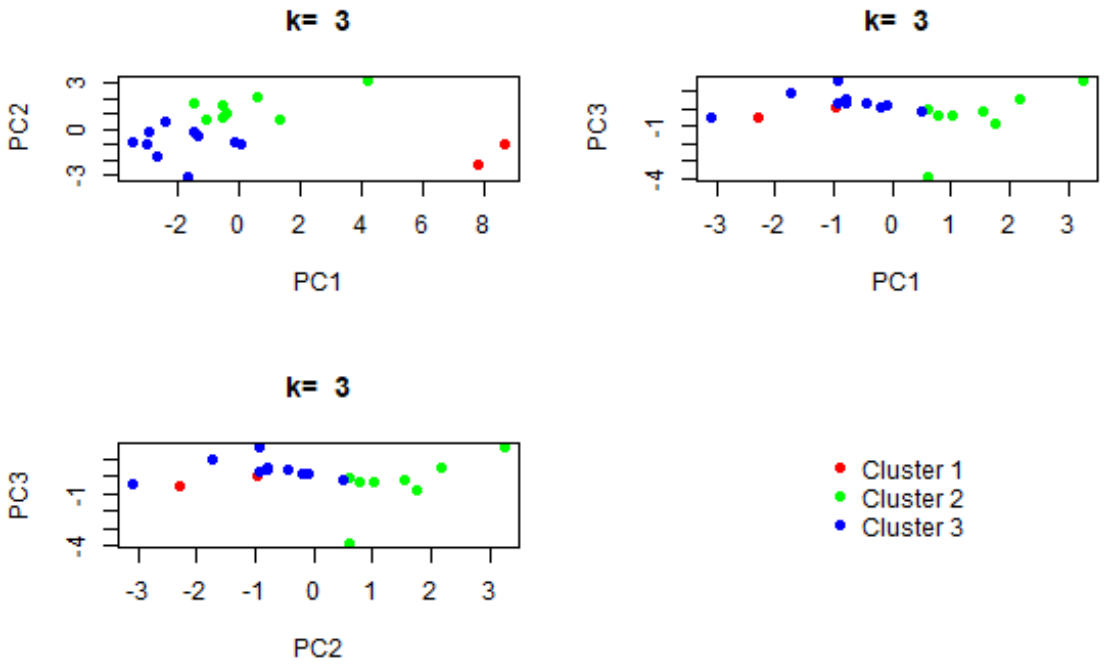

Figure 5: 3-means for the PCs

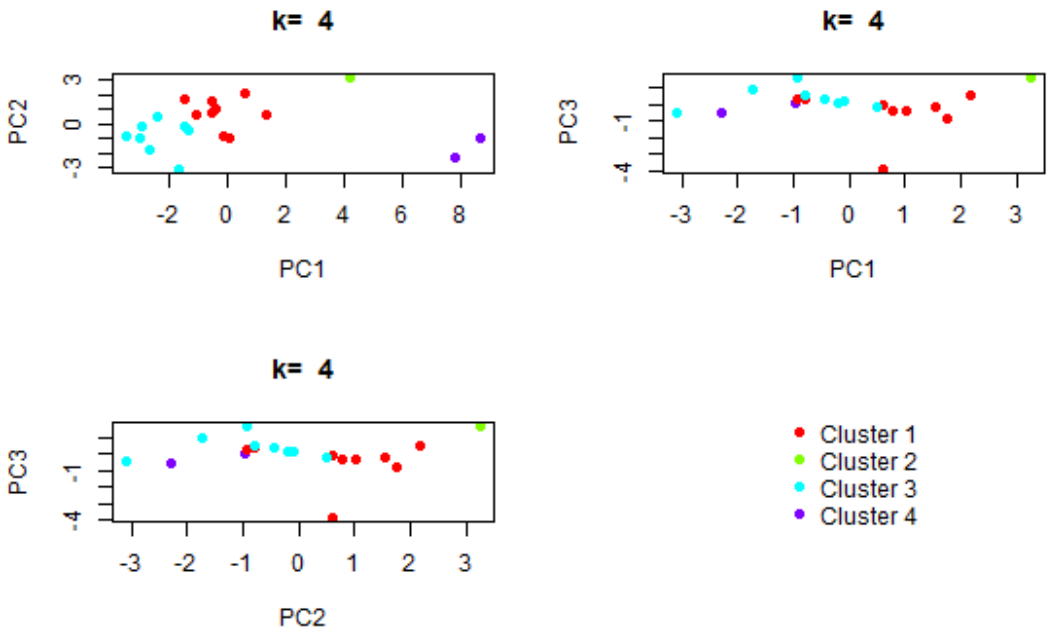

Figure 6: 4-means for the PCs

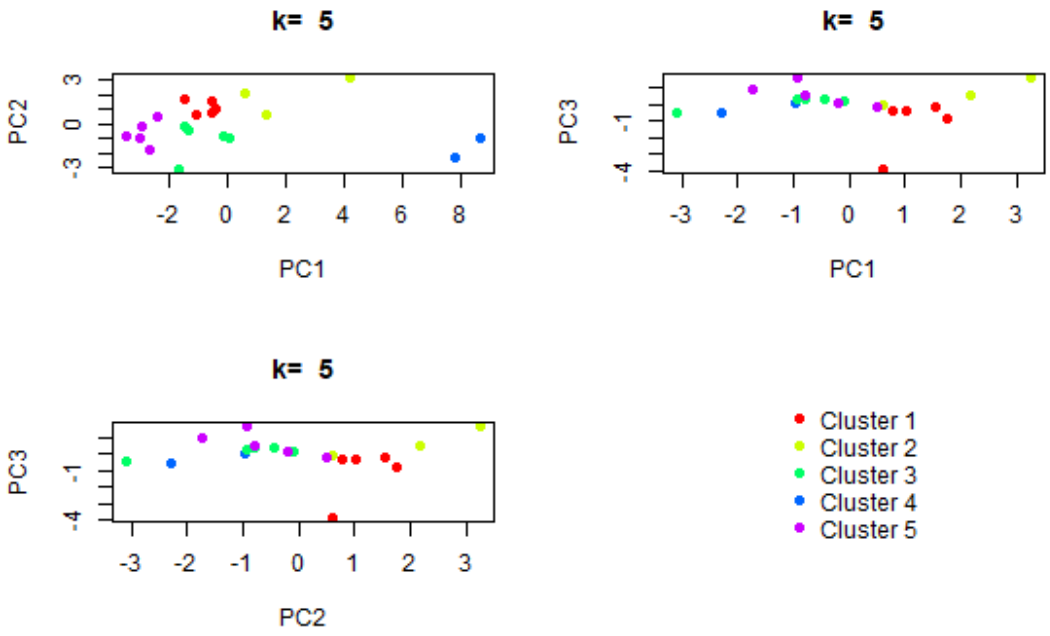

Figure 7: 5-means for the PCs