In this section we do Principal Component Analysis for one season of aggregated football data. The data includes observations like the number of goals, the number of shots, the number of passings and so on. The full list of variables can be seen in the appendix. The data is collected for each player, but we have summed it to each team. We would like to find any clusters by using PCA in R.

The resulting scree plot from doing the PCA is seen in Figure 1.

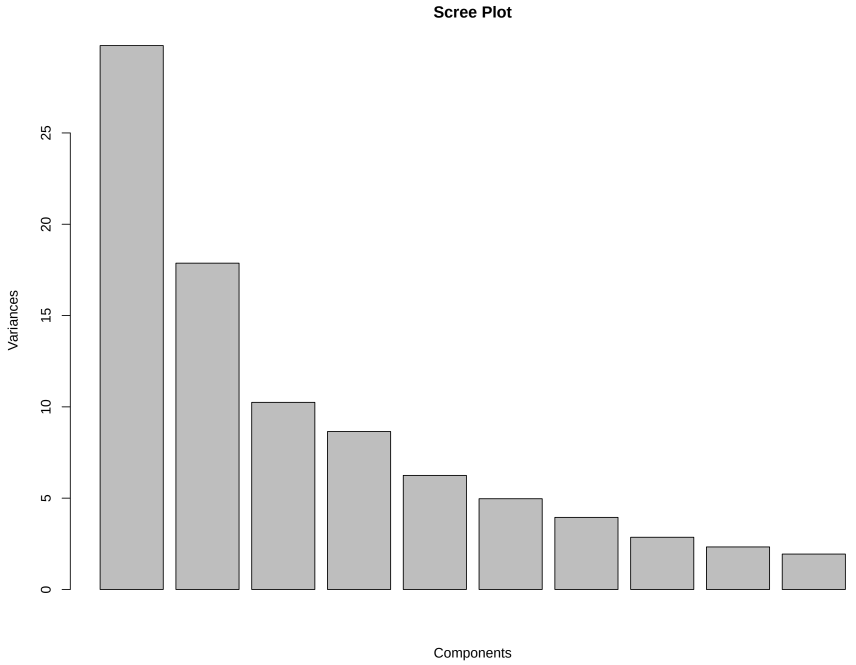


Figure 1: The resulting scree plot

The scree plot tells us which eigenvectors have the highest eigenvalue, meaning which principal components describe the most of the data. In our case we find that the first two components describe a lot of the variance in the data, to some extend also component 3.

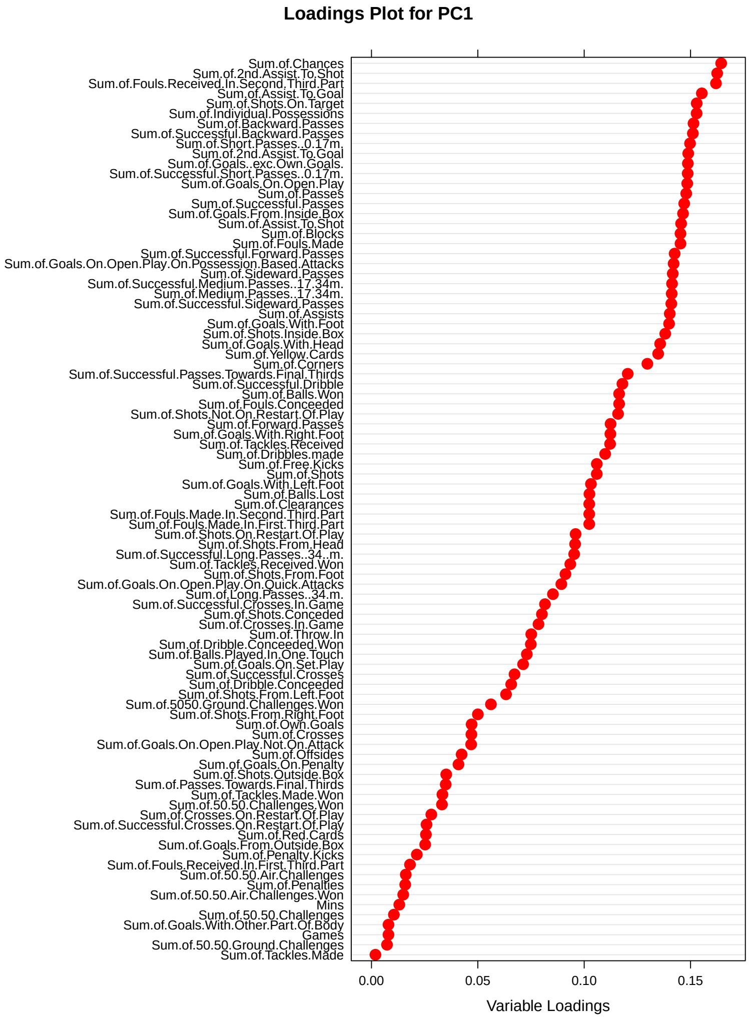


Figure 2: Loadings for PC1

We now look at the loadings of each of the components that we choose to keep. The loadings describe the importance of the different variables in relation to the chosen component. In Figure 2 we see the loadings for PC1. Being that PC1 is the component describing most of the variance in the data, the most important variables of PC1 is also the most important variables for the entire data set. The three most important variables in regards to PC1 is the `Sum.of.Chances`, `Sum.of.2nd.Assist.To.Shot` and `Sum.of.Fouls.Received.In.Second.Third.Part`.

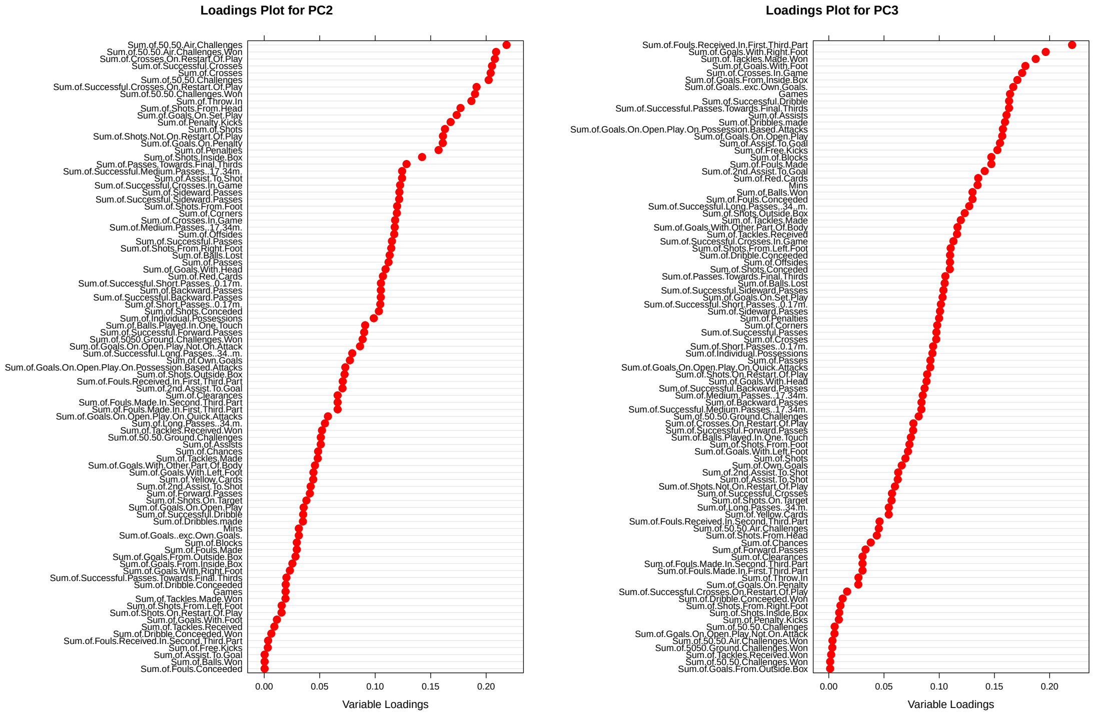
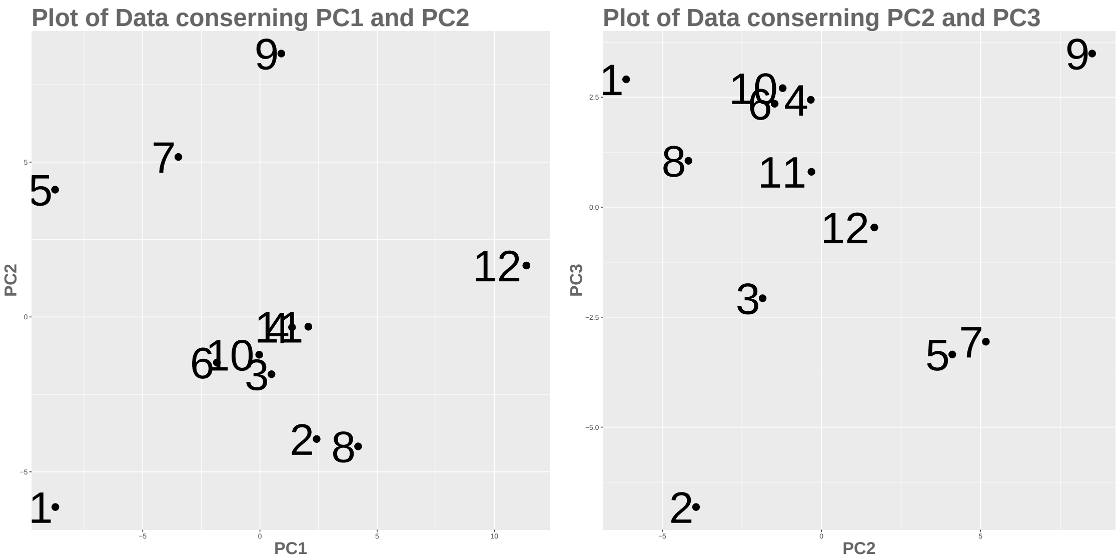


Figure 3: Loadings for PC2 and PC3

We can do the same with PC2 and PC3. These are seen at Figure 3. From the loadings in regard to PC2 we conclude that the three most important variables are the `Sum.of.50.50.Air.Challenges`, `Sum.of.50.50.Air.Challenges.Won` and `Sum.of.Crosses.On.Restart.Of.Play`. For PC3 the three most important variables are the `Sum.of.Fouls.Received.In.First.Third.Part`, `Sum.of.Goals.With.Right.Foot` and `Sum.of.Tackles.Made.Won`.

We can now plot the data points in regards to the principal components to find clusters. The data is labelled in regards to current position of the teams in the table. This means that we can easily find out if the top/bottom teams cluster. In Figure 4a the data is plotted in regard to PC1 and PC2. We do not have any clear clusterings, some teams are clustered in the middle. There is no correlation between the standings and the teams clustering.



(a) The data plotted in regard to PC1 and PC2 (b) The data plotted in regard to PC2 and PC3

Figure 4: Data plotted in regards to principal components 1-3

In Figure 4b the data is plotted in regard to PC2 and PC3. Once again we have a little cluster. There is no relation to the standings, but the cluster may arise from something else.

The PCA has not given any clusters, but it has giving us some of the most important variables of the dataset. Further analysis on these variables did not reveal any interesting aspects.