**Multivariate Analysis of 2018 FIFA World Cup**

Turkmnistan

12/03/2019

**Abstract**

In this project, the match results of 32 national teams in the tournament phase are analyzed to examine implications of factors on the advancement to the round of 16. The dataset includes common soccer statistics such as the total number of shoots and the number of goals. One of the main focuses of this project is to investigate the attributes which could have possibly contributed to the result of the game. Various plots are employed to visualize and aid understanding of the dataset. They are capable of succinct explanation of the distributions in the dataset, relationships between each factor, and the magnitudes of impacts of factors on the result. Principal component analysis is one of the main techniques used to explore the data points by reducing a high dimensional space to lower dimensional space. It provides a concise summary of the coefficients of factors, which each component accounts for the most variability at each dimension. Primary variables contribute to favorable match results are disclosed throughout investigation. This report also conveys possible classification by clustering supposed neighbors with given variables and visualizes to facilitate readers’ understanding of the clusters. Dendrogram, heatmap, and PCA score plot deliver a close look to the analyzation and an overview of the dataset. This repot unveils the implication of the variables on the match result: larger magnitudes in attacking statistics such as ‘Shots’ and ‘Corners’ are likely to culminate in a promising outcome.

**Introduction**

The FIFA World Cup is the most prestigious soccer tournament and, at the same time, the most widely viewed sporting event in the world. It even exceeds the viewership of Olympic Games: the viewership of the final match of 2006 World Cup is estimated to be 715.1 million which is a ninth of the entire population of the planet. As a huge fan of football, I enjoy watching and following the football matches all over the world. Especially, the World Cup is a competition between nations not clubs, which unites people from same countries and stimulates their national spirit. Unfortunately, South Korea was not able to advance to the round of 16 in World Cup 2016. Therefore, I wanted to probe into the results in order to discover what are the differences between the teams which made it and Korea which did not. Moreover, in 2018 World Cup, it is not only the most recent World Cup, but also it is when the underdogs displayed better performance than the top dogs did. Hence, I found it the most interesting World Cup to investigate the possible contributors. I googled the statistics and results of the first round in the tournament phase and manually recorded them in an excel spreadsheet.

|  |  |
| --- | --- |
| Countries | All 32 countries qualified for the tournament phase. |
| Shots | The total number of shots the team made |
| Shots.OT | The total number of shots on target (goal) the team made |
| Goals | The number of goals the team scored |
| Possession | Ball possession. The amount of time the team possessed the ball during a game in a percentage (for example, 60 (%)). |
| Fouls | The number of fouls the team committed. |
| Yellow | The number of a yellow card the team had received. |
| Red | The number of a red card the team had received |
| Offside | The number of offsides called by a referee |
| Corners | The number of corner kicks the team had taken |
| Result | The result of game. (Win, Draw, and Lose) |

These variables are the most widely used soccer statistics, which describes the general flow of the match. These variables could be related to each other. For example, a team with higher number of shots would result in a higher possession. Besides, a team with more shots and higher possession could be a more competitive team than one with lower statistics. However, those variables do not always lead to winning. Therefore, I would like to investigate and grasp an idea how these variables may impact the result of the game.

**Goals**

1. **Investigate the general distribution of variables.**

First, I would like to delineate the overall distribution of each variable and the relationships between each variable by visualizing with various plots. Those plots will inform us how the data points are distributed along attributes and whether or not there are any extraordinary pattern or unexpected observation. With a tangible illustration of multi-dimensional dataset, the plots would offer us insights on the dataset, enabling further examination.

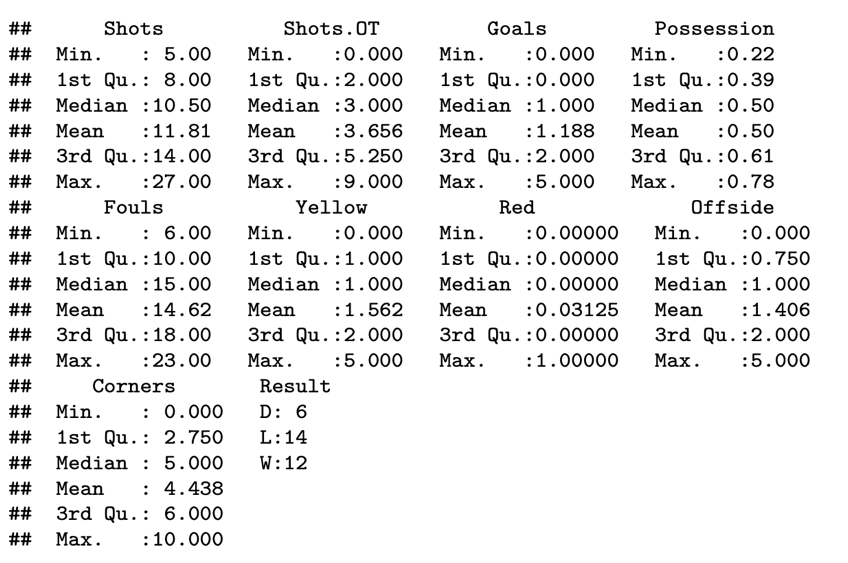
1. **Analyze the data with principal component analysis.**

Principal component analysis (PCA) allows analyzation of variables with correlation adopting linear combinations. Thanks to PCA, it is able to reduce the dimension of the dataset space, accounting for most of variability with the first few components. After running PCA, I would first decide the most optimal number of principal components to be used, then analyze and interpret the components to understand the influences of attributes. It will suggest more succinct and comprehensive relationship between the factors. In addition, it is capable of standardization of variables in different magnitudes and correlations by scaling.

1. **Find possible clusters to classify teams**

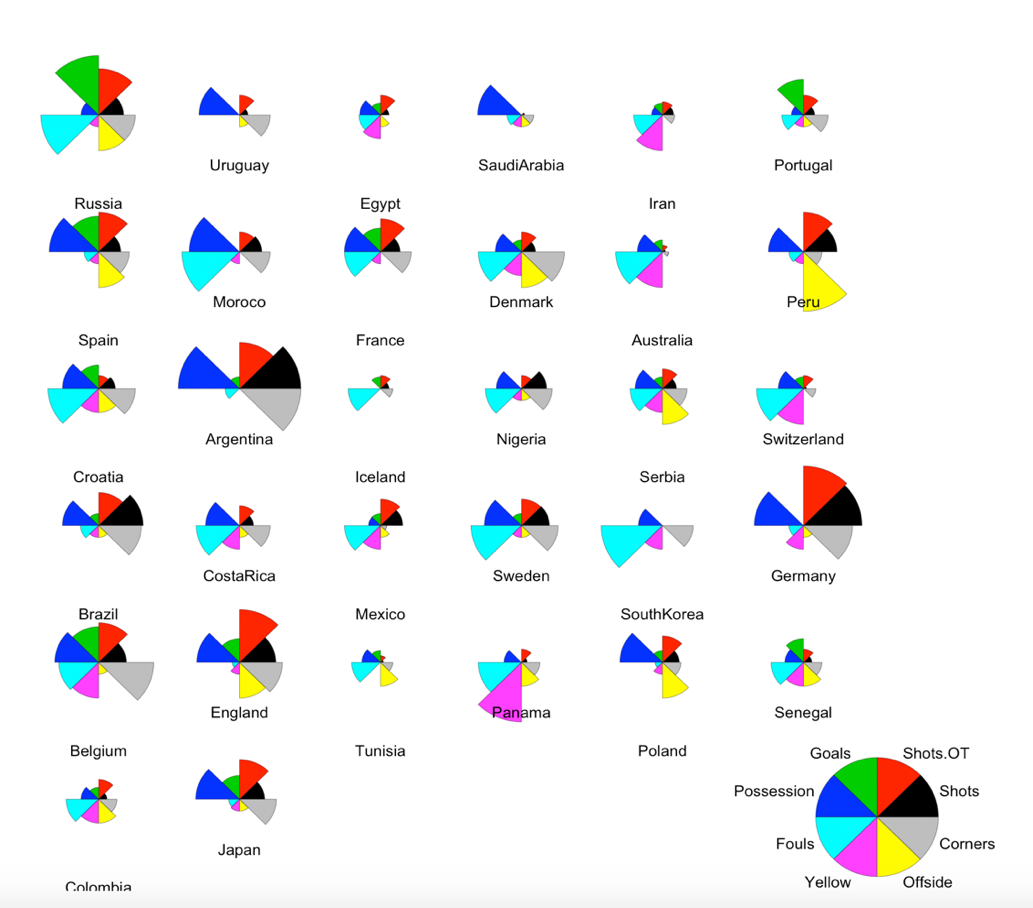
Lastly, I would like to utilize the dendrogram and figure out if it is possible to classify teams into correct subgroups by employing different clustering methods. The questions I would like to answer are: 1. What could be possible groups of nations after clustering based on given variables. 2. How different clustering methods would culminate in different classification of teams. 3. What could be the best clustering method to classify nations. I was wondering if it is possible to find optimal clusters to classify teams without a knowledge of soccer or the results. Throughout this analyzation, I would be able to demonstrate what are characteristics that distinct the winners from the losers based on statistics such as possession rate and the number of shots.

**Main Results**

1. **Investigation of general distribution of the data.**

**Table.1**

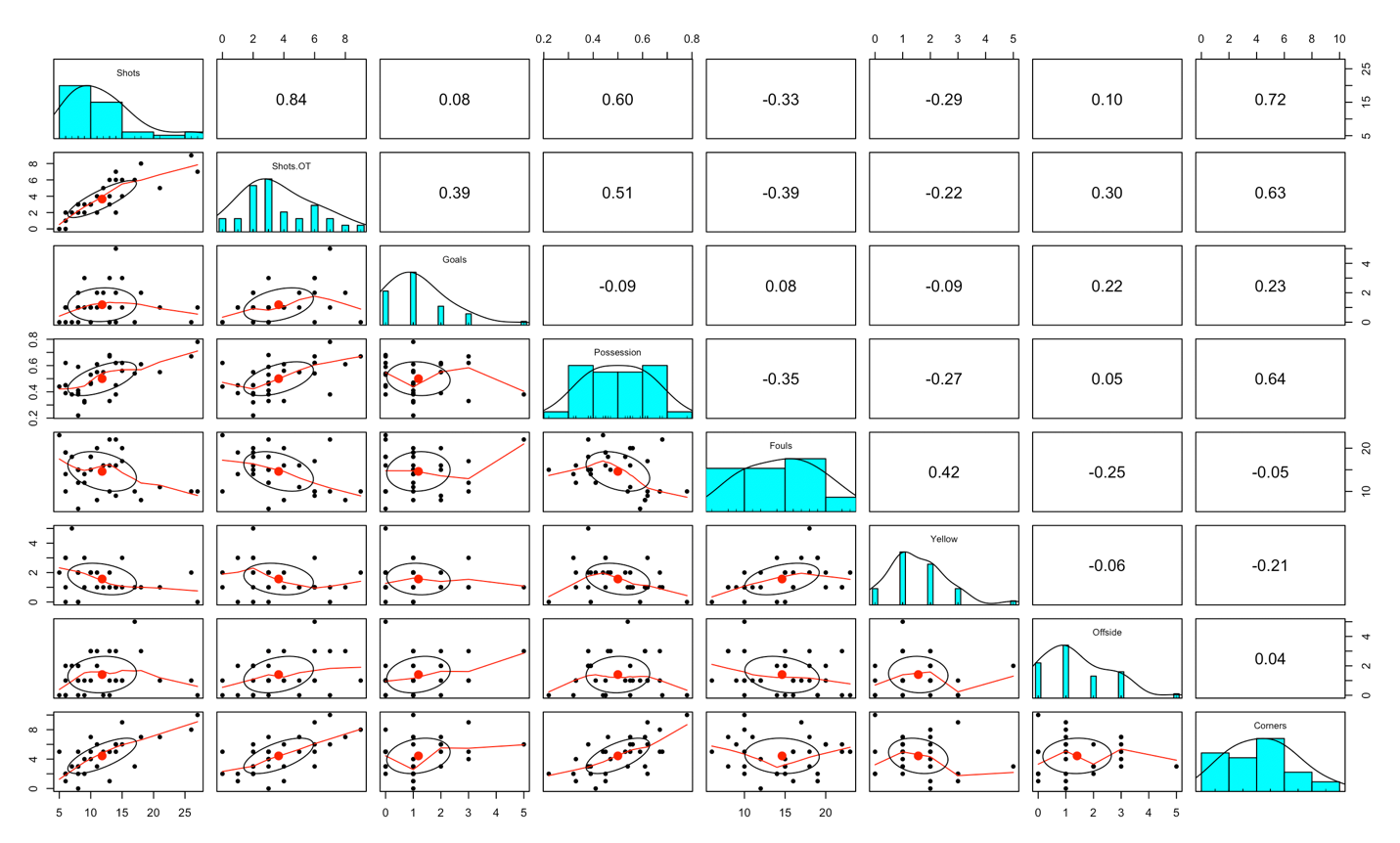
Prior to analyzation, I want to take a look over the raw data and see if there is anything I have to pre-process. Looking at a summary table of the dataset and a raw data, I realized that there is only one nation that has received a red card. Removing a “red card” column resulted in better principal component analysis in terms of cumulative proportion. Hence, I decided to exclude the red card variable.

**Figure 1. Overview of statistics of 32 national teams**

First of all, I created a star plot, figure 1, which provides an ocerview of all 32 national teams. Each circle represents a national team with the label below. Pies of a circle represent levels of each variable and they are colored to improve the readability. Out of 32 teams, Argentina and Germany display outstanding figures. They are regarded as the most competitive teams in the World Cup. Looking at their plots, we can notice that they resulted in higher magnitudes in the number of shots, shots on target, corners, and possession, which are usually featured in top tier teams. For teams with small circles, such as Iceland and Iran, we can observe that they show small magnitudes in the number of shots, shots on target, corners, and possession, but high magnitudes in fouls and yellow card. There are also countries showing distinctive patterns such as Peru with the highest magnitude of offside while short in other variables. Although it is not always constant, we can observe that so-called top dog teams generally have bigger stars than others and demonstrate higher magnitudes in the number of shots and possession. Correspondingly, underdog teams tend to display small stars with higher degrees in the number of fouls and yellow card.

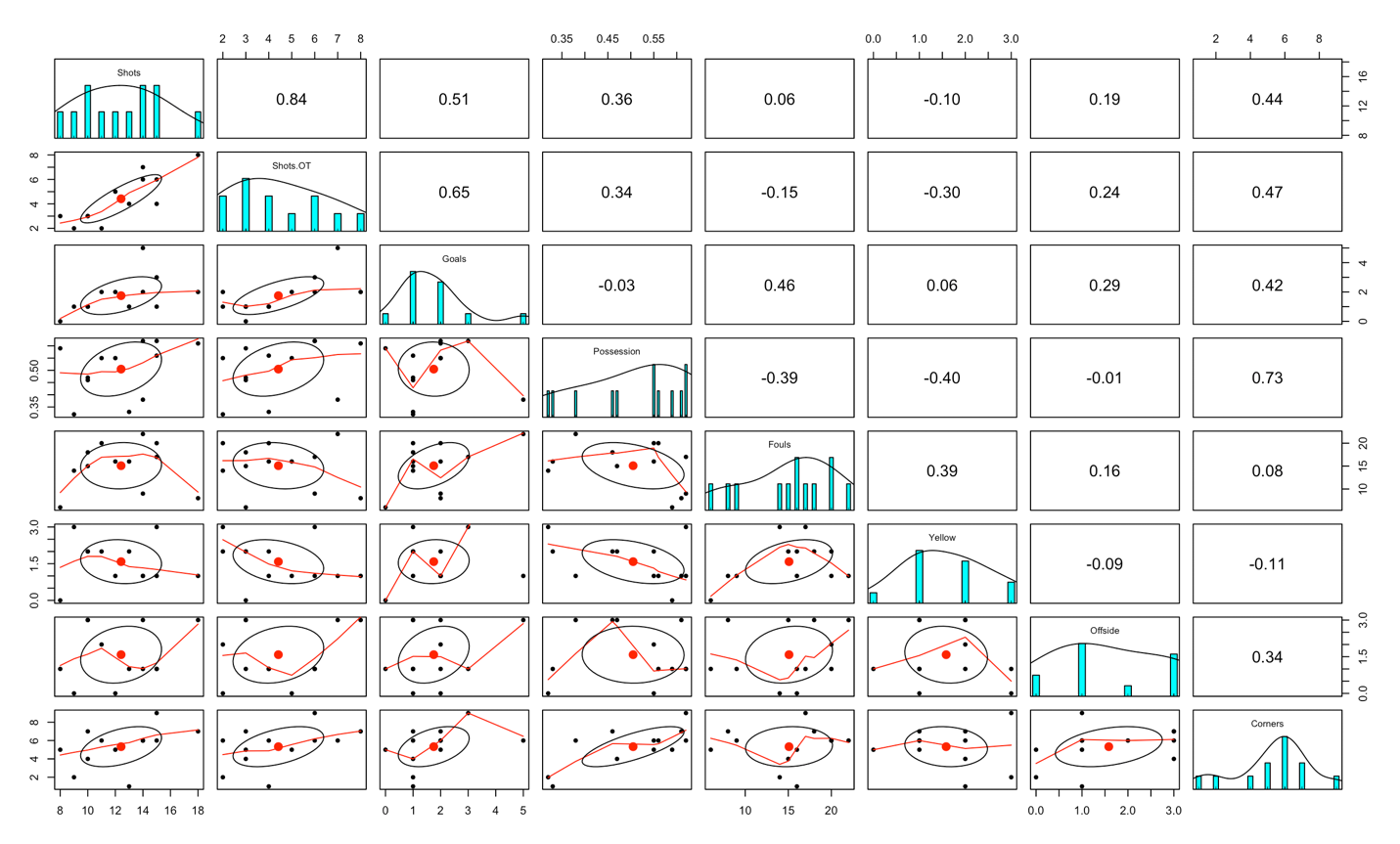
After taking a peek at the overall idea of the statistics of all 32 nations, I decided to create a scatter plot matrix with histogram and correlation, which not only visualizes the distributions of the dataset according to the variables but also present correlations between them. I first created a plot matrix with all 32 national teams in order to grasp a general overview of the whole dataset, which is a figure 2. Then, I divided teams into 3 subgroups depending on results and created three plot matrices for each subgroup in order to distinguish and highlight different distribution patterns according to the results.

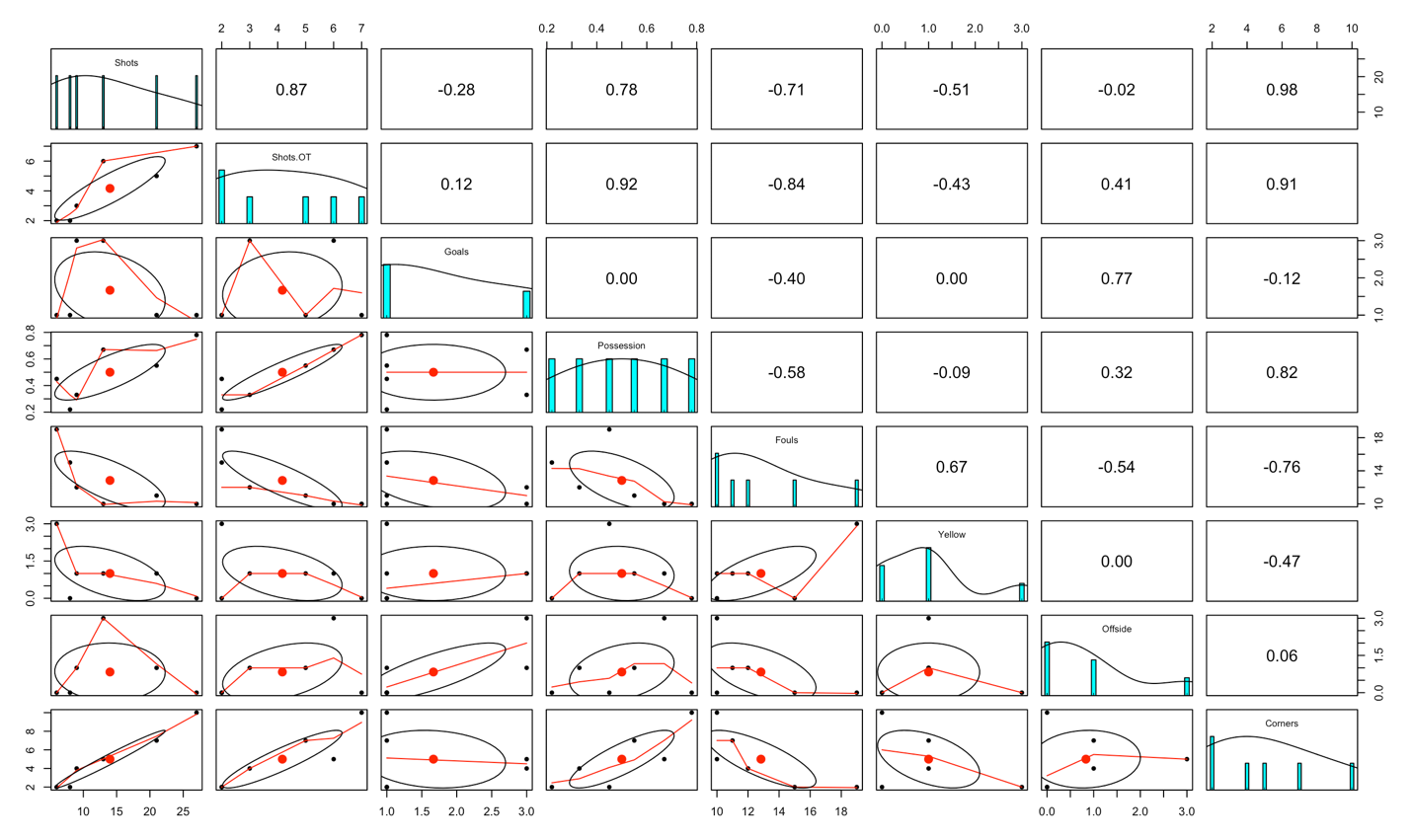
**Figure 2. Plot matrix for all**

**** The plot above presents the general distribution of each variable with a scatterplot and histogram. Looking at a histogram, we can observe that except possession, fouls, and corners, they are mostly skewed to the right. It could be interpreted in that most of the teams resulted in relatively less than the intermediate level. For example, for the number of shots, most of the teams has made the small figures and only a few teams made many. From table 1, which is a summary table for the dataset, we can clearly see that the first quantile, mean, and the third quantile are close to each other while the max number of shots is almost twice of the third quantile. Both table 1 and figure 2 delineate deliberately how they are skewed. For ‘Possession’, ‘Fouls’, and ‘Corners’, their curve shows almost perfect normal curve, which could be a sign that those variables are normally distributed. Looking at quantiles, median and mean from the table 1, we can validate that those three variables are close to the normal distribution indeed.

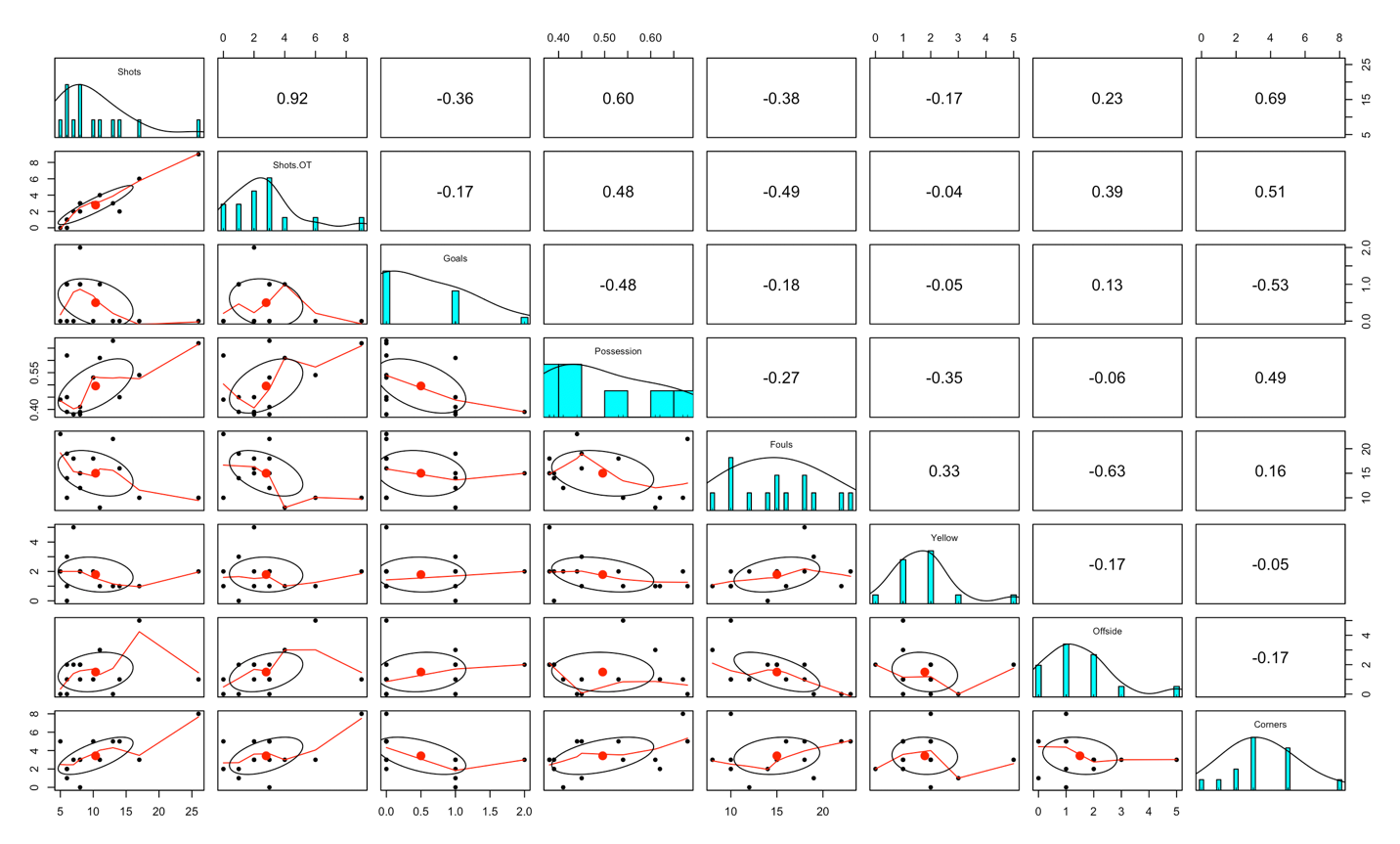
Correlations explain and aid our understanding of possible relationships between the variables. For example, the number of shots and shots on target shows a high correlation between them which is very reasonable. As there are more number of shots, the more likely there will be shots on target. Intriguingly, the number of shots has very low correlation with the goals while the shots on target has the highest correlation with the goals. It could mean that the total number of shots is not that important. In other words, quality (accuracy here) matters more than quantity. Then, I wanted to see the distribution of each subgroup, since winning teams would be more competitive teams and their differences with losing teams will highlight the factors influencing on the result.

**Figure 3. Plot matrix for winning group**

 The first thing catches my attention from figure 3 is that the distributions of variables are not the same with the distributions in figure 2. We can detect that the number of shots and shots on target are more normally distributed. In addition, ‘Possession’, ‘Fouls’, and ‘Corners’ are skewed to the left. That could be a significant sign in a sense that the stronger team tends to possess ball more than the weaker. The more number of corner kicks could mean they played more on the enemy’s side and had more chances of scoring. A high magnitude in ‘Fouls’ could be interpreted that the game was a neck-and-neck and got more aggressive.

**Figure 4. Plot matrix for drawing group**

**Figure 5. Plot matrix for losing group**

 Above, we have two plot matrices. Since there are only 6 teams drew (figure 4), the distributions would not be as reliable as other groups. However, we can clearly see the changes in all distributions. Most of the variables are skewed to the left for figure 5, which means they tend to exhibit statistics lower than intermediate level. Especially, ‘Possession’ illustrate fascinating transition. It was firstly skewed to the right for winners, gets more normal for drawers, then became skewed to the left for losers. It could be an evidence demonstrating that the higher possession leads to higher chance of winning.

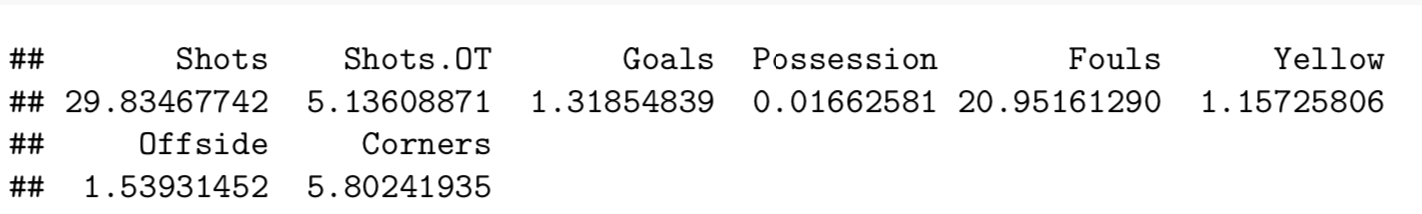
While higher possession tends to result in positive outcome, it is intriguing when we actually see the dataset. Argentina played against Iceland and they are like outliers in a sense that they resulted in the highest and the lowest possession respectively. However, the match ended in a draw. Furthermore, France, Mexico, and Russia won the matches while showing the lowest possession. Accordingly, Germany and Morocco showed the highest possession and lost the games. Thereby, other variables should be accounted for as well to depict correlation between the variables and the results.

There are a few limitations in the data and its visualization. First, there are 8 variables and it is hard to attain a holistic picture of the relationships between them while summarizing the information of the data points in 8th dimensional space. In addition, most of them are correlated to each other. For example, the number of shots would obviously be highly correlated with the number shots on target. One team’s possession would be a subtraction of the enemy team’s possession from 1. Also, the enemy teams are not recorded, even though it could be an important variable we would like to discuss. Therefore, we would not want to describe the whole dataset, and it would be better to reduce the dimensionality by adopting principal component analysis, which would offer us succinct portrait of the dataset.

1. **Principal Component Analysis (PCA)**

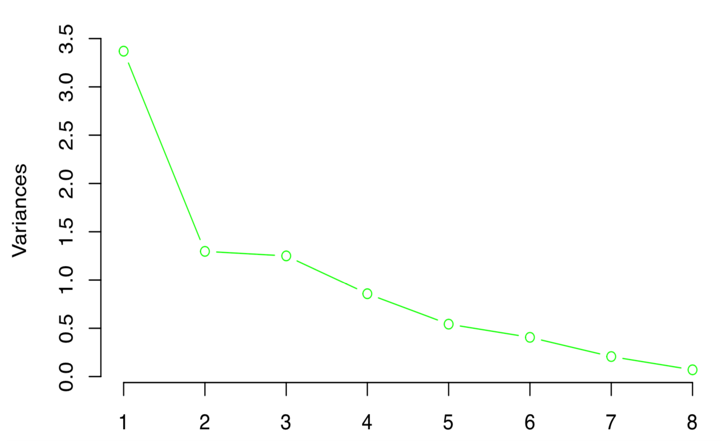
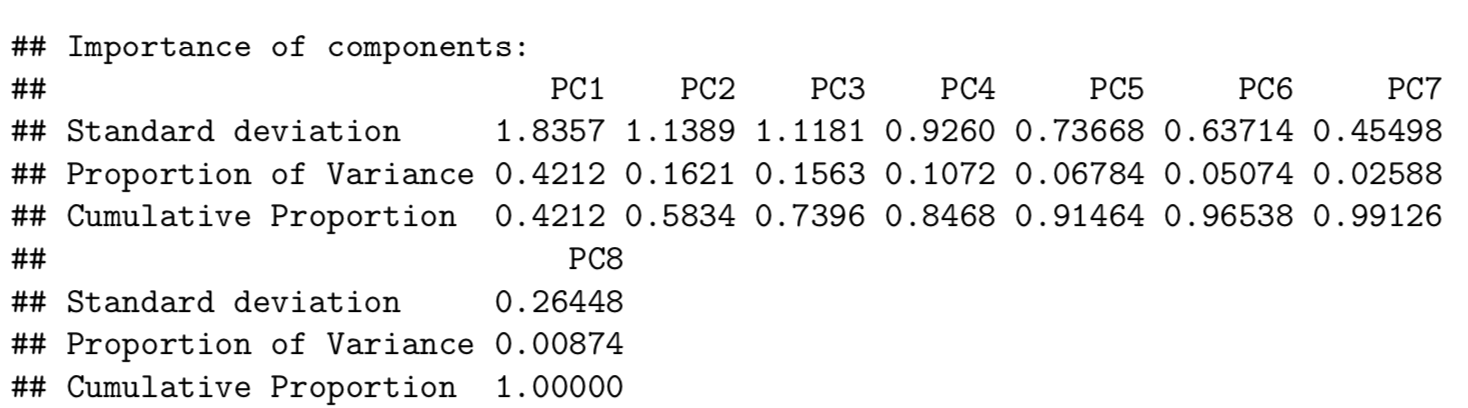
Now, I would like to run PCA to describe the dataset more concisely. First, I will take a look at the variances of the variables.

**Table 2. Variances of variables**

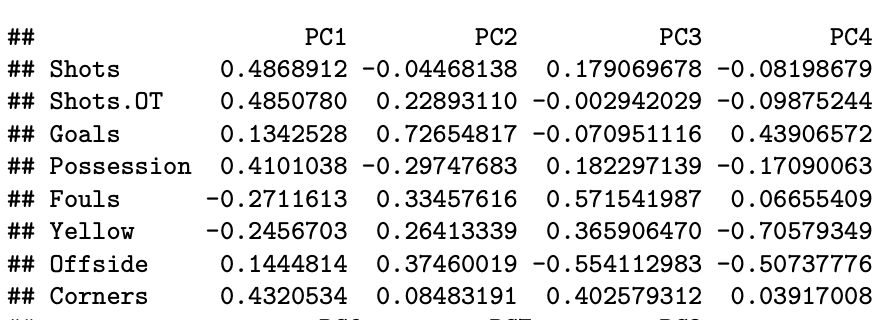


From the table, we can observe that ‘Shots’ has the highest variance. Since the scaling of each variable is not uniform and variances are not similar, I decided to run PCA with a correlation matrix over covariance matrix. Otherwise, some specific variable could impact the whole trend, resulting in a totally different outcome. After running PCA, I decided to take first four principal components based on a scree plot and a summary table in figure 6 and 7.

**Figure 6. Scree plot and summary table of principal components**



There are guidelines for how to decide the number of the principal components to pull out. First, we look at the scree plot and find the ‘elbow’ where the gradient of the curve gets less stiff. A scree plot displays how much variation each principal component captures from the data. Second, we aim the cumulative proportion to be from 70 to 90 percent. We want to account for as much variability as we can while not overfitting on the given dataset. Third, it is better to choose components with standard deviations higher than 1 or 0.7. As we scaled them using correlation matrix, they are standardized. Although, 2 is where the curve flattens out on the scree plot and it is technically the elbow, the table suggests that it only accounts for 0.5834 of cumulative proportion. Therefore, I decided to 4 would be the optimal number of components to use, since 4 is still a quite small number, has pretty high standard deviation (0.9260) and accounts for a cumulative proportion close to 85%.

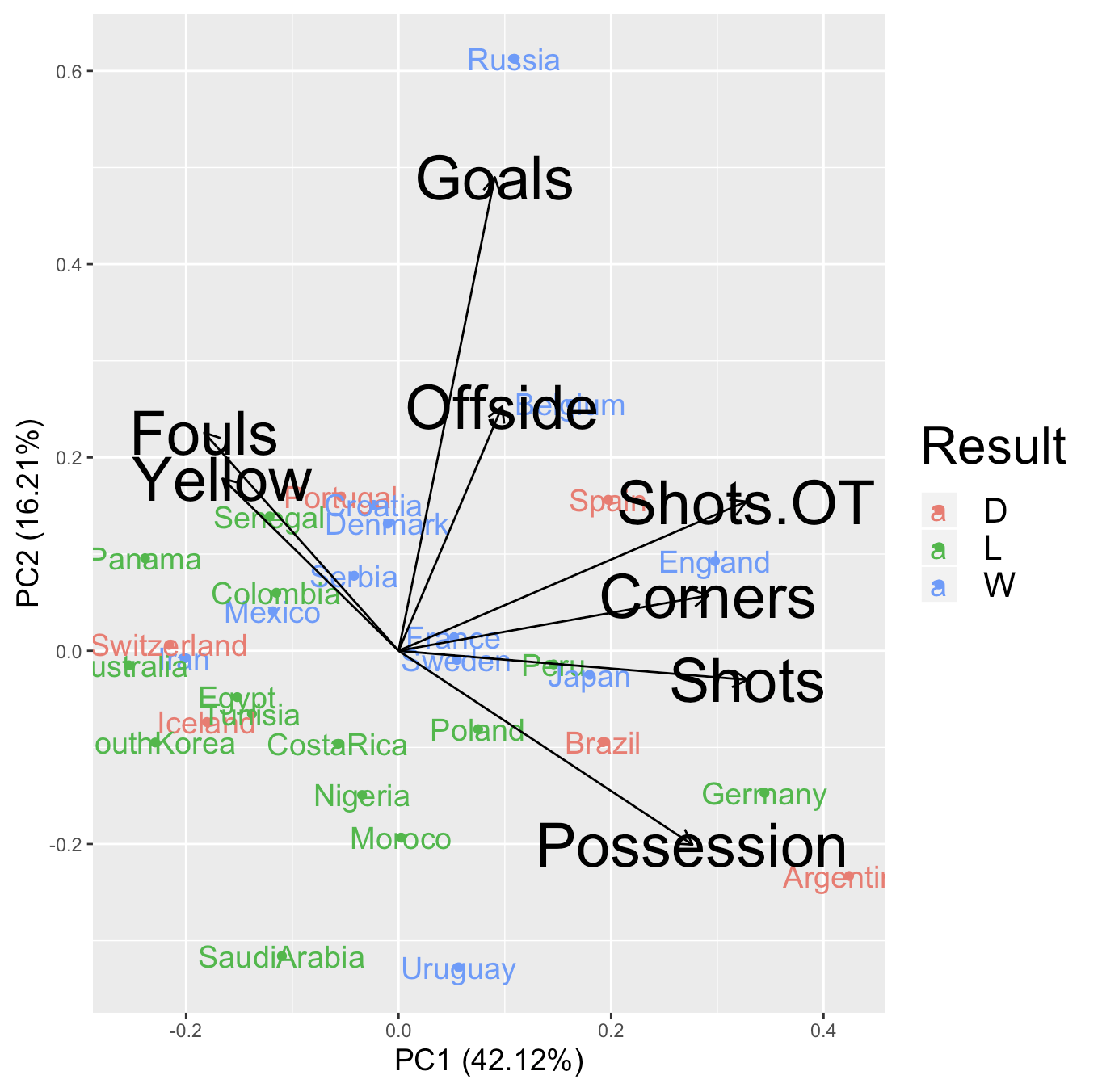
**Figure 7. Loading of the first four principal components**

According to the table in figure 6, the first four components together account for 84.68% of total variance. The first component 42.12% of the variation and we can notice that ‘Shots’, ‘Shots.OT’, ‘Possession’, and ‘Corners’ have very high positive weights while ‘Yellow’ and ‘Fouls’ high negative weights on the first component.

The second principal component explains 16.22% of the remaining variations in the dataset. It forms a basis which is perpendicular to the first component, expanding to the second dimension and accounting for the most variation after the first component captures the most number of data points in the first dimension. We can observe that ‘Goals’ has the highest positive weight on the second component and ‘Possession’ is the most negatively weighting. It could indicate that ‘Goals’ is the most significant contributor after accounting for PC1.

The third and fourth components explain about 15% and 7% of the remaining variations. For the third component, ‘Fouls’, ‘Corners’, and ‘Yellow’ are major positive agents and ‘Offside’ is the significant negative agent in variations, after accounting for the first two principal components. For the fourth component, ‘Goals’ weighs significant positive magnitude, and ‘Yellow’ and ‘Offside’ do significant negative magnitudes after accounting for the first three principal components.

In overall, the first four principal components demonstrates that most of the variability in the data could be explained with variables in the lower dimension. The first component could be interpreted as how many chances they had for scoring: the number of shots, corners, and possession which somewhat describe how many times they went for a goal. The second component could mean how many goals they actually made and how aggressive the game was: ‘Goals’ has the highest magnitude, and the numbers of fouls, yellow, and offside show the aggressiveness. It is yet pretty vague and difficult to understand what those coefficients actually mean. Hence, I decided to draw another plot to visualize the influences of those variables on principal component and actual distribution of the dataset.

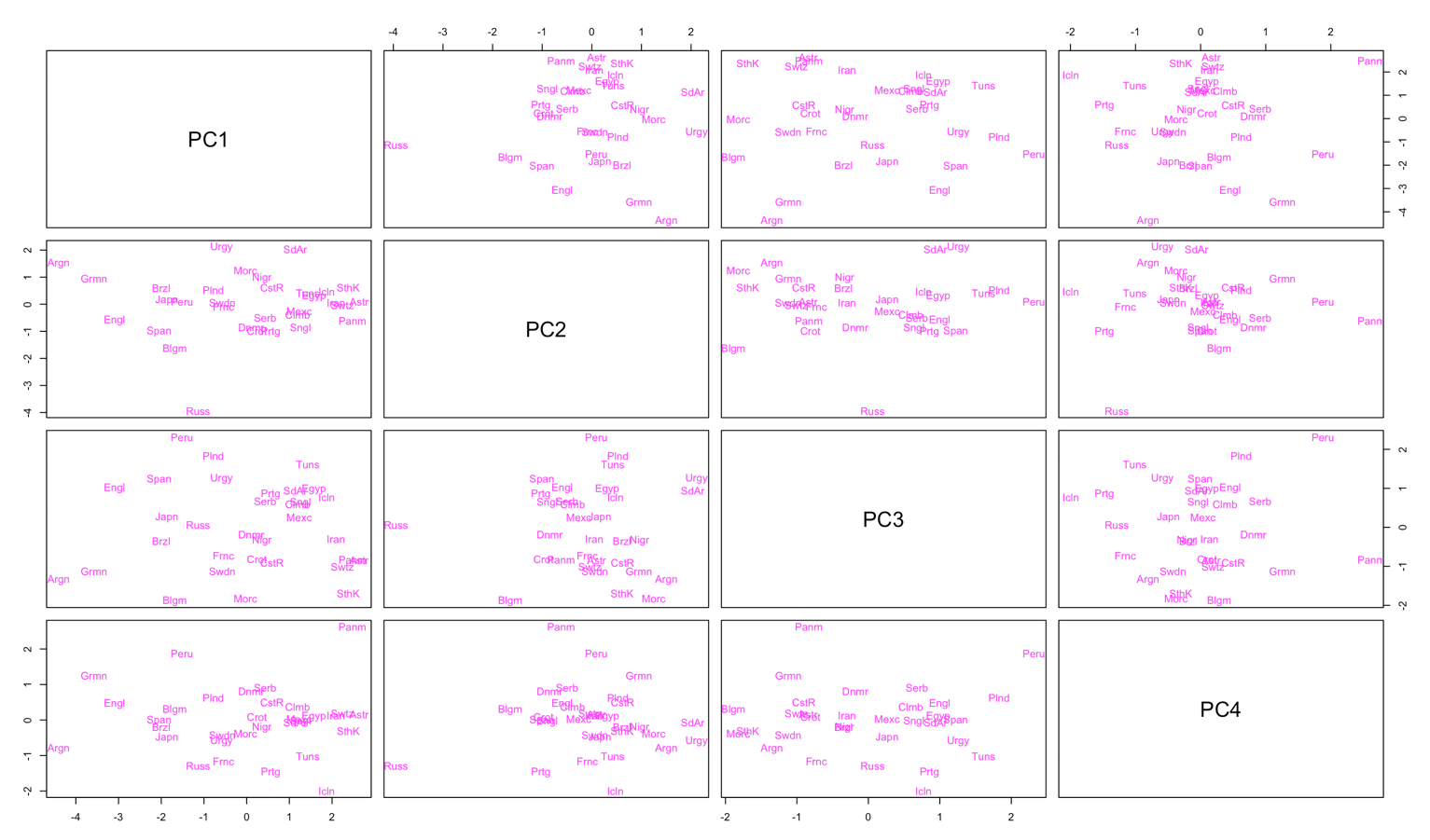
**Figure 8. PCA Biplot with labels and grouping**

On the plot above, I created a PCA biplot of actual data points with the first component on x-axis and the second component on y-axis. I sub-grouped teams according to results and colored them differently to easily distinguish them. A PCA biplot is very useful in that it displays both principal component scores of samples (dots) and loadings of variables (vectors). The further away the vectors are from the origin (center), the more influence they have on that component. Loading plots gives a clue of how variables are correlated with one another: a small angle indicates positive correlation, a large angle implies negative correlation, and a right angle denotes no correlation between the two.

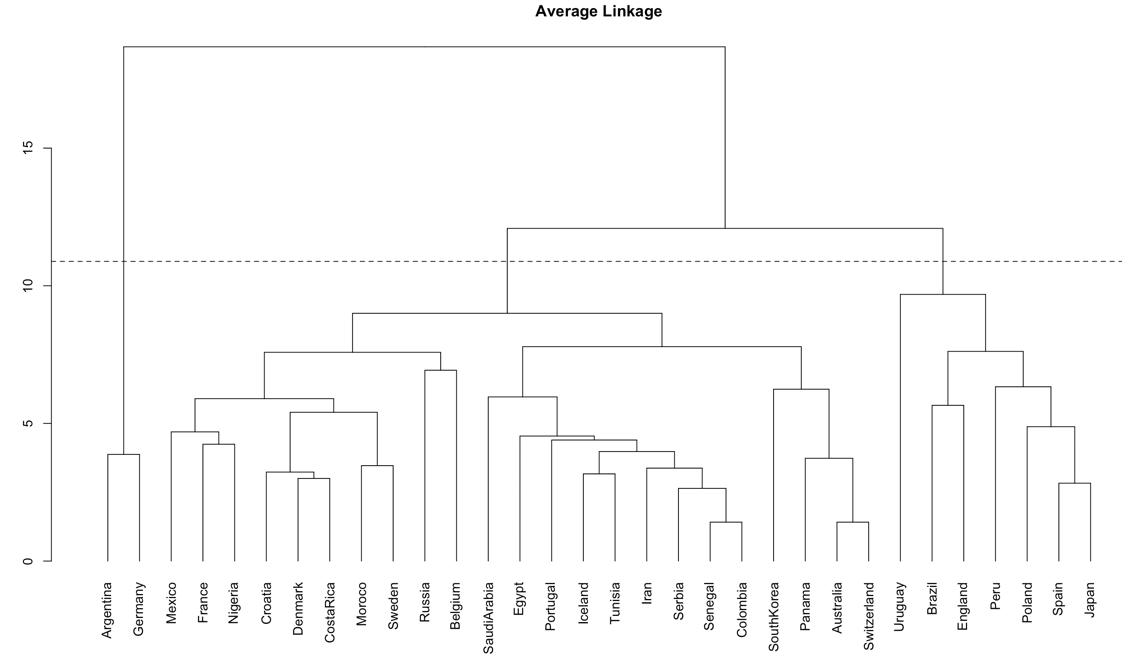
As I explored above, we can see that ‘Possession’, ‘Shots’, ‘Shots.OT’, and ‘Corners’ display long arrows which direct toward the right, indicating that they have high positive coefficients on the first principal component. For ‘Possession’, we can see that it is directing toward right-bottom corner, meaning that it has high positive coefficient for PC1 and high negative coefficient for PC2. We can approve that from the figure 7. This plot visualizes those enigmatic numerical values and provide simpler interpretation of principal components. This plot also accords to what we have seen from the star plot. We can notice that Argentina, Germany, England, Brazil, and Spain resulted in large number of ‘Shots’, ‘Shots.OT’, and ‘Possession’. Australia, Panama, and South Korea committed more fouls and received more yellows than others. Russia and Spain made more ‘Goals’ and ‘Offsides’ than other countries.

Now, let us discuss on the samples group-wise. It is validated that the winning group illustrates relatively higher ‘Possession’ and ‘Shots’. According to the plot, we are able to validate the hypothesis mentioned above, which the more competitive teams tend to result in higher magnitudes in ‘Shots’ and ‘Possession’. We can witness that losing teams are mainly allocated at the left bottom corner, and they are relatively higher in ‘Fouls’ and ‘Yellow’. Drawing group exhibits intriguing result. In terms of PC1, drawing group is separated into a half: three on the left-side and the others on the right-side. This could be interpreted as the teams on the right-side play more aggressive and the teams on the left side more defensive. Figure 8 below involves every possible pair for the first four principal components, which allows further investigation for other components.

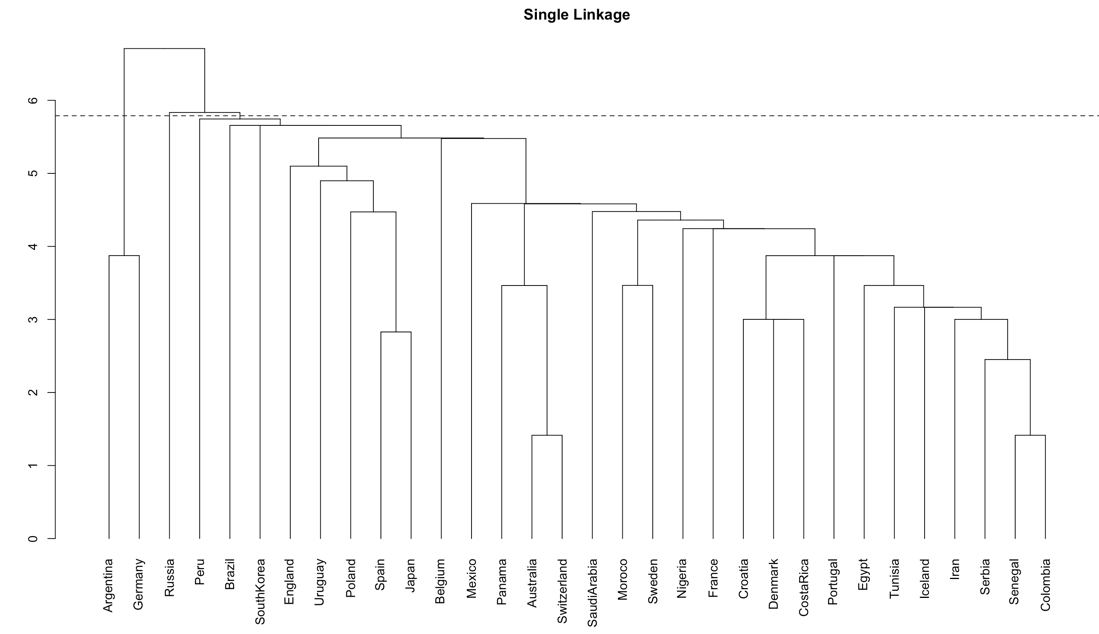
**Figure 8. PCA score plots panel with the first four PCs (names are abbreviated)**



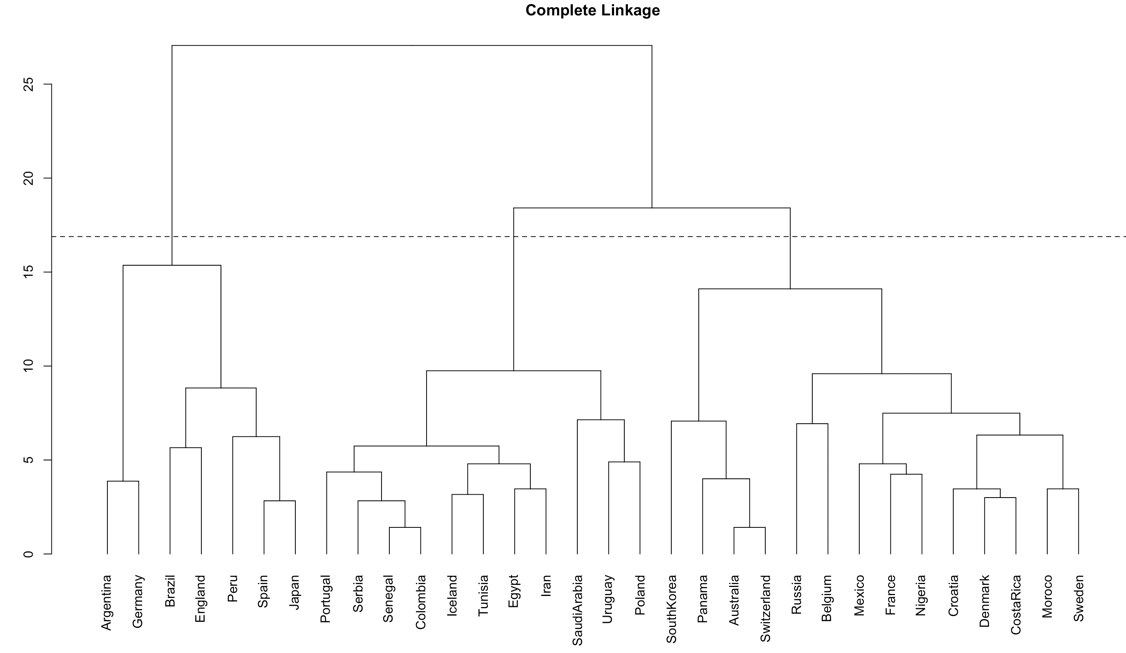
1. **Classification with different clustering methods**

**Figure 9. Dendrogram with 3 clusters cut (Average Linkage)**

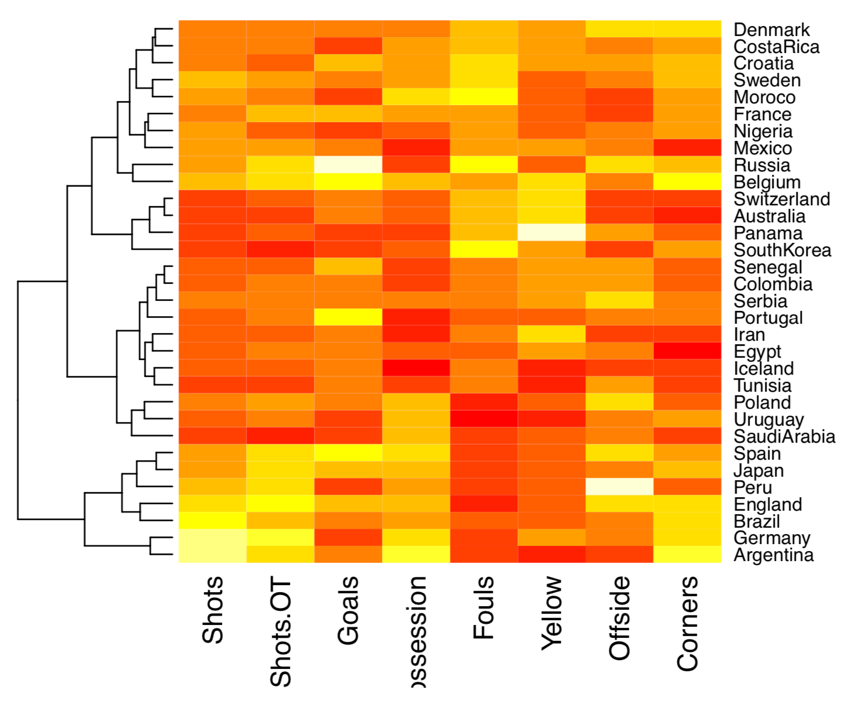
First, I created three dendrograms with three unique clustering methods: average-linkage, single-linkage, and complete-linkage. Since we know that there are 3 different groups depending on the match results, I drew a cutting line on each dendrogram, which determines the three clusters. From figure 9, when we adopt an average-linkage, we can notice that there are three groups in total: one with 2 nations, another with 23 nations, and the other with 7 nations. Average-linkage is a method which calculates the distance so that the distance from any member of one cluster is to be equivalent to the any member of another cluster. It results in compact but not well-separated neighbors. According to the plot, teams are clearly not distributed uniformly, and subgroups are not similar to the actual groups teams belong to. One notable output is that in the first cluster Argentina and Germany are the only two countries included when we applied an average-linkage. It could mean that Argentina and Germany are very close to each other but not from others. Moreover, it is interesting that 23, which is almost three quarters of the whole dataset, are allocated into one same cluster.

**Figure 10. Dendrogram with 3 clusters cut (Single Linkage)**

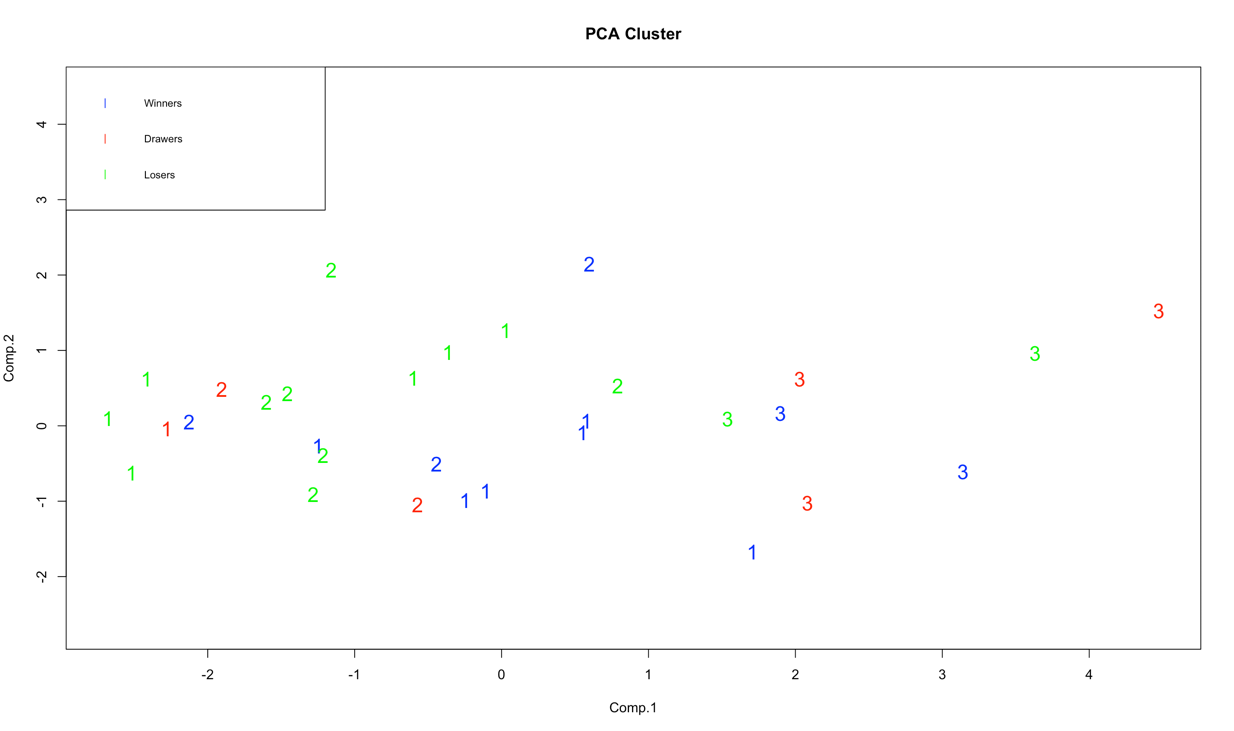
Single-linkage is an agglomerative clustering, combining two clusters with the closest pair of elements. In each step, it calculates the shortest distance between two closest members from two and merge them. It results the nearest neighbors grouped together. According to the figure 10, a single-linkage method assigned 2, 1, and 29 teams into three clusters respectively. It is far more extreme than the result from an average linkage – more unequally distributed and not even close to the actual groups in terms of the number and organization. We can still see that the first cluster is also comprised of only two nations, Germany and Argentina. It could indicate us that Argentina and Germany are the closest to each other and pretty far from all other nations as it had a pretty huge average distance from others. In addition, the cluster 2 is only comprised of one nation, Russia, and it could hint at the single-linkage clustering is vulnerable to outliers. Argentina and Germany are outliers in most of the variables, and it might have affected how branches of the dendrogram (figure 10) are shaped. Figure 10 suggests that a single-linkage method is a method we would like to avoid.

**Figure 11. Dendrogram with 3 clusters cut (Complete Linkage)**

A complete-linkage method is a farthest neighbor clustering. When compared with the previous two clustering methods, the complete-linkage clustering exhibits a better portrait in a sense that it categorizes teams in more even pattern, as shown in figure 11. In addition, the number of members in each cluster is approximate to the number of actual group. The three clusters from the complete-linkage have 11, 7, and 14 members while the actual winning, drawing, and losing groups have 12, 6, and 14. In terms of number, it almost perfectly sorted teams into groups, although elements are not according with actual groups. Therefore, the complete-linkage clustering seems to be the most satisfactory method to implement among all three hierarchical clustering methods.

**Figure 12. Heatmap with dendrogram**

Since dendrogram only tells us the assigned groups of the teams, I produced other plots to help understanding of readers who do not have access to the actual raw dataset. For figure 12, on the right-side, we have the names of the nations, and, on the bottom, we have variable names. On the left-side, there is a dendrogram created by the complete-linkage clustering. On the heat map, the brighter color represents higher level in the given variable. As I described before, Argentina and Germany show almost whitish-yellow for ‘Shots’, which means they are the highest in that attribute. Argentina displays a deep dark red color for ‘Yellow’ since it has received 0 yellow card. It provides more detailed clarification of how close nations are and how they are grouped.

**Figure 13. PCA plots with the first two PC and labeling**

Besides, I would like to introduce another plot to present how accurate clusters generated by the complete-linkage clustering are. Figure 13 is a plot based on PC scores with the first two principal components as axes. I labeled points with according clustering groups and colored them to differentiate labels from the actual group they belong to. Unfortunately, the plot indicates that it failed to classify teams perfectly: not all 1s in one group, all 2s in another, and all 3s in the other. However, promising observation is that at least most labels in a certain group are coherent. For example, for blue colored points (winning groups), it is mostly comprised of 1s, fewer 2s and the fewest 3s.

From the three hierarchical clustering methods above, none of them was able to successfully categorize teams into groups corresponding to the actual results. While there were a few points that deserve an attention. In all hierarchical methods, Japan and Spain, and Germany and Argentina have always been grouped together. When we take a look at the star plot above and the actual raw data, Germany and Argentina represented almost same sizes in a star plot and displayed extremely high magnitudes in ‘Shots’, ‘Shots.OT, ‘Possession’, and ‘Corners’. Japan and Spain also presented pretty high magnitudes in those four variables.

**Conclusions**

This study scrutinizes a general overview of the first round of match results of 32 nations in the tournament phase: what soccer statistics are correlated to each other and what are the clues leading to favorable match result. As discussed above, most of the measurements are correlated to each other. For example, ‘Shots’ and ‘Shots.OT’ are highly correlated to each other. And, ‘Possession’ is also highly correlated to indicators of attacks such as ‘Shots’ and ‘Corners’. However, ‘Offside’ generally displayed low correlations with other variables, and ‘Goals’ showed almost no correlation with others except ‘Shots.OT’. Principal component analysis was employed in order to propose more concise delineation and to account for these various correlation and distributions by reducing the dimensionality and standardizing variables in different variances and scales. PCA with correlation matrix resulted in:

In overall, the first principal component is a weighted average of the factors with high correlations to each other. It accounts for the most variability among all principal components, offering a measure on the overall chance of scoring and somewhat aggressiveness (‘Shots.OT’, ‘Corners’, and others). The second put more weight on how many goals a team actually scored. The third would be more of how toughly a team played - how many fouls and yellows. The fourth could be interpreted as how many goals a team scored without committing fouls (huge. negative coefficients in foul and offside). However, there are few limitations in that it is challenging to discover what are primary variables in each group. In addition, a 45 degreed vector in a PCA plot expresses a subtle correlation with principal components. Nevertheless, PCA successfully presented a briefer description of the structure of the dataset.

For classification, various hierarchical clustering methods were introduced to find the optimum clusters. Implementing three different clustering methods and principal component scores, I was able to present possible clustering groups in various ways such as a dendrogram and a PCA plot. Unfortunately, there was no appropriate method to sort teams into the correct groups. The errors might have been caused because of a few outliers. Once an element is merged to one group, it cannot be re-merged to another cluster, which may cause an underlying error. Namely, few outliers could affect the whole structure of a dendrogram. However, I discovered that the complete-linkage method culminated in the most promising outcome in terms of the distributed number of members in each cluster. The complete-linkage method well classified teams into reasonable neighbors based on statistics, but the match result is always unpredictable in real world. One of the most important facts is that the two groups, Germany and Argentina and Japan and Spain, are the closest neighbors to each other. Nonetheless, their match results are different. Germany lost while Argentina drew, and Japan won while Spain drew. This could disclose why each clustering method was not able to successfully assemble teams and may reveal a potential flaw in the data itself. The clustering accuracy could be improved by far if we have a bigger and more refined dataset – for example, including the number and the accuracy of passes and more match results.

To sum up, based on the given game statistics, the key factors lead to win are “Shots”, “Shots.OT”, “Possession”, “Corners”, and “Goals”. It is somewhat obvious in a sense that a team can never without scoring a goal: they will draw at best or either lose the game. This could imply that teams should play more aggressive and try more for goal by shooting. PCA offered a concise summary and description of the dataset, and it will be far more improved with more data. For example, in the given dataset, although top dog teams presented a great performance with high possession and high number of shots, the results were not as expected. Argentina and Germany were top two nations with the best performance, but both of them could not win the game, which might have significantly affected the structure of PCA and dendrogram. However, losing team are generally short in those 5 variables listed above which is a promising sign, supporting the argument. Thereby, great magnitudes in the five variables mentioned above would prevent a team from losing at least and may lead to win at best.