Data Analysis

* This project is made for INF356

1st Ömer Ertekin
Computer Science
Galatasaray University
Istanbul, Turkey
omer.ertekin.gsu@gmail.com

I. Introduction

In this document, we will describe and analyse our dataset for predicting the winner

II. Purpose of this analysis

It is very difficult to predict the course of the match in football. The match played between each team can progress differently from each other. You have to be very lucky to be able to predict match results without any data. What if we have data? Can we predict the winner by analyzing this data? At least can we make some guesses for the bet coupon?

III. Main question and sub-questions

In this section we will try to divide our problem to smaller parts and try to gather info for answering the main question.

A. Main Question : Who will win?

There are thousands of factors that will affect the match score.It's really hard to guess and that's why it makes the most money. For answering this question, we will try to examine some of factors that will affect the winner of the match.

B. Sub-questions

- Which team is better ? : This question will have a great impact about score because as we said,every single match can progress differently. And that's the main reason for this difference.
- Can both teams score? : This is a common type of bet and we will shape our score prediction based on the information from this question. There are hundreds of factors that will affect this situation. Therefore we need another questions.
- How many shots did the teams take to score a goal? : Since we will examine how many positions the teams have entered and how many shots did the teams took, we would have a statistics that will answer to this question.
- How many cards did the teams get? :A yellow card can make a huge difference in the defensive performance of the teams in terms of hardness. A red card would make a bigger difference in game because the team is missing one or more person. Therefore, it will be very difficult for you to create the appropriate area in terms of offense and defense.
- Who plays a more dominant game? :The more you dominant, the more chances you have to score, the less chance of conceding a goal. We have various data to understand who plays more dominant like possession percent, pass percent,aerials won count etc.

IV. Our Data

We have 98 team from 5 major soccer league in the world and their data about different things. So let's take a look at data that we have and examine some of the analysis

A. Goals

We have data on the number of goals scored by teams over the entire season. This data will definitely help us in predicting the score of the match.As you can see from the figure 1, the number of goals has a normal distribution. Its mode (50) , median(50) and mean (52.18) are very close to each other.

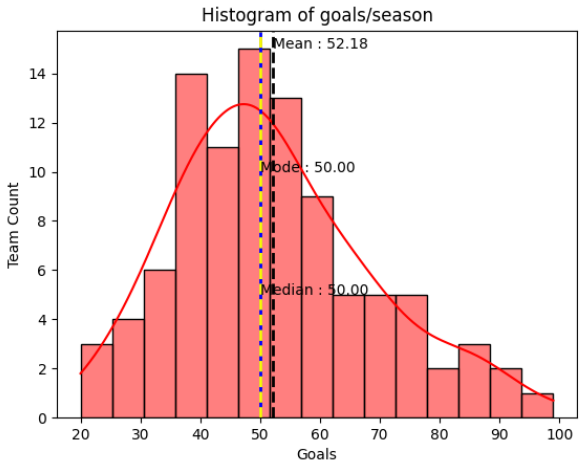

Fig. 1. Goals per season.

B. Shots Per Game

You can't score without shoting (at least most of time). Therefore, we will examine this data together with the goal

scored data to measure the team's goal scoring ability. As you can see in figure 2, its mode (11.60) , median(11.45) and mean (11.85) are not very far to each other so we could say that it has a normal distrubition.

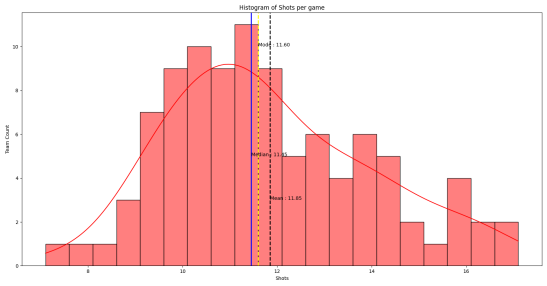

Fig. 2. Shots Per Game.

C. Cards

Cards will affect the attack and defense performance. And each type/number of card will have a different effect. That's why we will examine each card separately

1) Yellow Cards: If a player is shown the yellow card, he/she has to be much more gentle when making his/her next defensive action. That's why it will affect the match score. Like other data, we can say that yellow cards has a normal distrubition too with mode = 63, median 67.5, mean 69.70 in Figure 3.

2) Red Cards: Red cart directly effect the game because the more players you have on the field, the easier it is to find a position and the harder the opponent finds a position. In figure 4, With mean = 3.34, mode = 3, median = 3, data has a normal distrubition

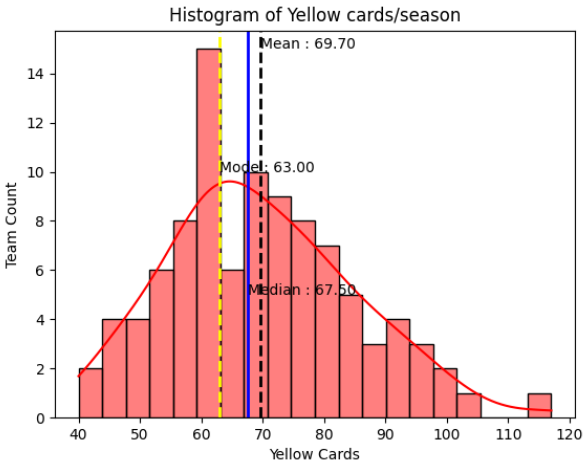

Fig. 3. Yellow card per season.

D. In-game advantages

You can dominate your opponent in different ways in the game. Some teams dominate the game by winning too many

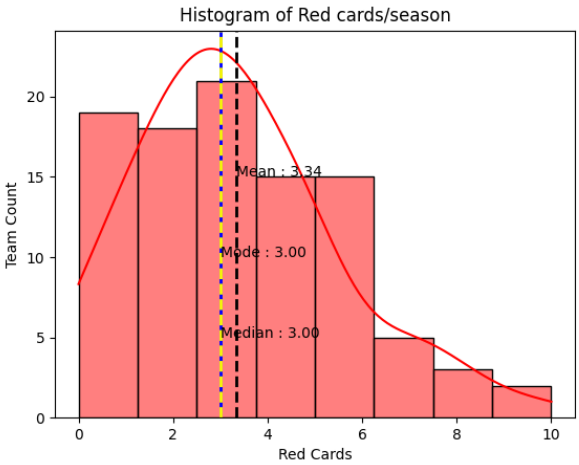

Fig. 4. Red card per season.

airballs or the team doesn't lose the ball because most of its passes are accurate. Or, since the percentage of possession is very high throughout the match, the team finds easier positions and does not give the ball to the opponent. For that we need to analyse this 3 type of data and combine them.

1) Possession%: This data does not give very precise information about the score of the match. Because it is possible for teams to have a high percentage of ball possession and not be able to score. However, it can still provide information about the progress of the match. As we can see in Figure 5, posession% has a distrubition like normal. Its mean = 50, median = 49.75 and mode = 51.5

2) Pass%: Like Possession, this data isn't very precise too. But it will effect the match so we will examine it in Figure 6. Mode = 82.53, Median = 80.80 and Mean = 80.44. A normal distrubition.

3) Aerials Won: That data will effect the defence and attack performance. Let's have a look at Figure 7. Mode = 18.30, median = 16.10, mean = 16.01. So close! That's why it has a normal distrubition

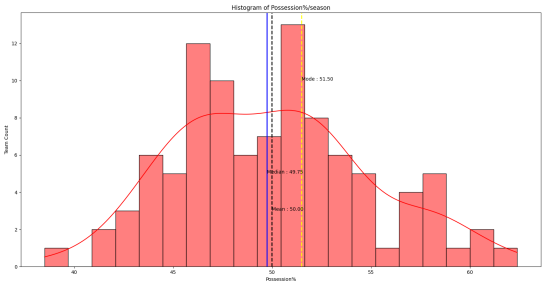

Fig. 5. Average possession%/season.

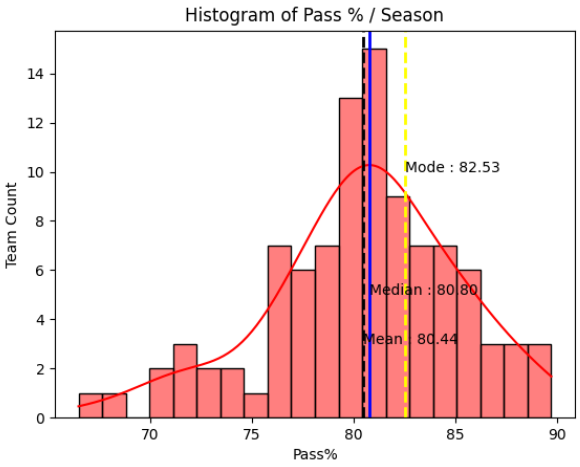

Fig. 6. Average pass accuracy%/season.

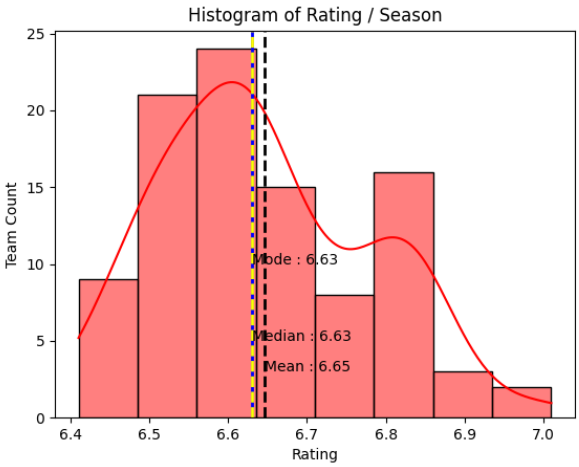

Fig. 8. Ratings in season.

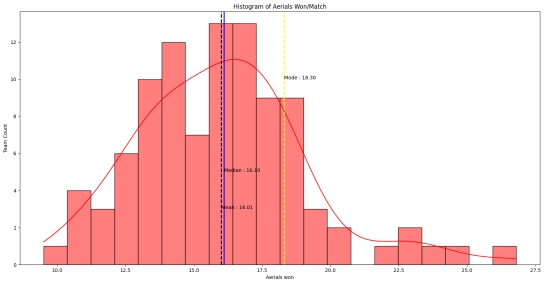

Fig. 7. Aerials won/per match.

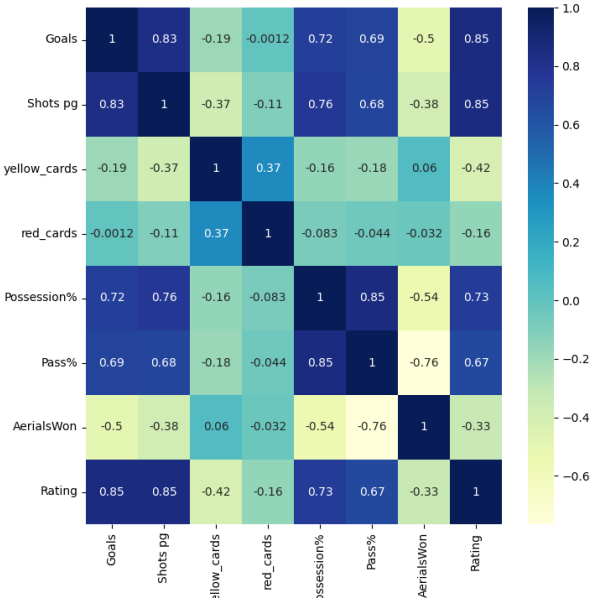

Fig. 9. Correlation matrix heatmap.

E. Team ratings

When two teams meet each other, it will be difficult to score goals according to the level of the teams. Therefore, analyzes made independently of this data will not yield logical results. Let's look at this analysis at Figure 8. It has bimodal distrubition with Mode = 6.63, median = 6.65, mean = 6.55

V. Examine our data to use

Since we have enough data to use, let's try to examine which of these data are related to each other. Than we will use related data and try to predict the winner of matches.

A. Overview to variables

The most important thing to predict the winner is goal count. If we could predict the team which will score more, we would find the winner of the match. In that case we can start by examining the variables that can affect the number of goals and the relationship between these variables. But before, let's look at all of the correlations between variables at figure 9.

B. Look at data pairs

We will take some pairs to examine and try to catch a correlation between these two datas. And we will use these pairs to predict our score in our final analysis

1) Goals and Shots Per Game: The duo, which we expect to be the most relevant to each other at first glance, does not mislead us when we look at its graph.We can clearly see that there is a positive correlation between shots and goals in Figure 10. Our correlation coefficient is 0.83 which is really close to 1

2) Goals and Team Ratings: The more you score, the better team you are. And the data is confirming us. We can directly say that goal counts and team ratings are relevant(positively). As we can see in figure 11 ,correlation coefficient is 0.85.

3) Pass and Possessions: The more accurate your passes, the longer the ball stays on your feet. Therefore, it is inevitable that these two are related to each other. The pass will affect

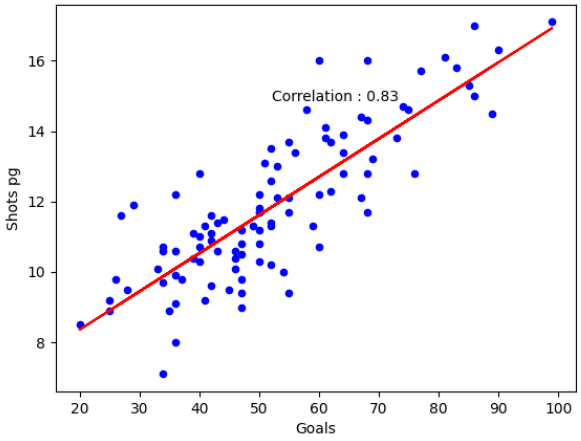

Fig. 10. Goals and shots correlation.

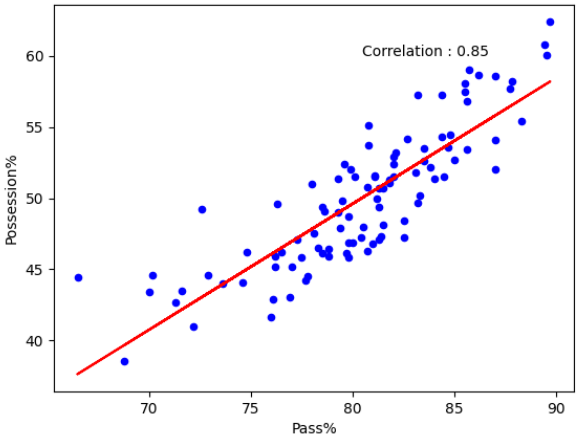

Fig. 12. Pass and possession correlation.

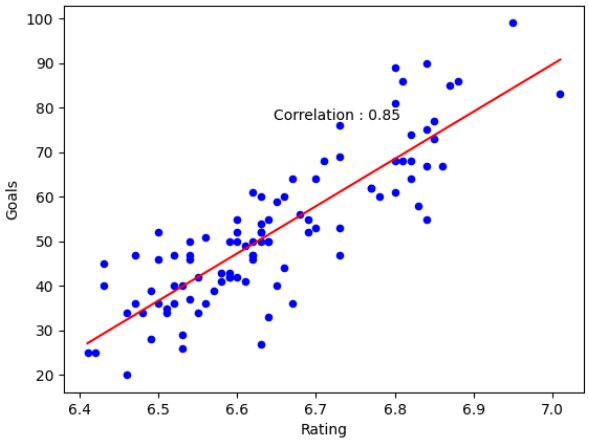

Fig. 11. Goals and rank correlation.

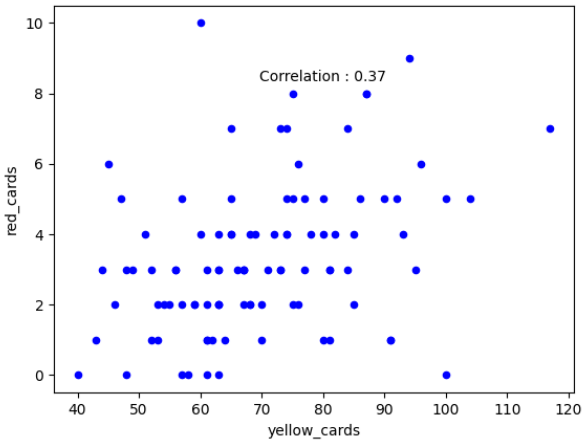

Fig. 13. Yellow and Red Cards correlation.

the team overal power and we will use it to predict our winner. Let's look at positive correlation between pass and possesions in figure 12

4) Yellow cards and Red cards: Two yellow cards mean one red cards. So they must be relevant isn't it? No. It is strange but no they haven't a clear correlation. We can examine their relation in figure 13

5) Possesion and Aerials Won: This pair may seem a little unrelated. But actually they aren't completely inrelevant. Teams with a higher percentage of ballplay this season have won less airballs. We can see negative correlation between these two data in figure 14

6) Possession and Rating: Even some teams have a different strategy like counter attack, in general good teams play more with ball. Our data is confirming us as we can see in figure 15. There is a positive correlation.

C. Look at all of pairs

We selected some pairs from our data and examined the relationship between them. We will use interrelated data to complete our analysis. We will try to determine the outcome of the possible match of the teams using each of its associated data. At figure 16, there is a complete pair plot of our data to see clearly the relations between data frames.

VI. Creating and determining a team score value

For having a result, we need to choose our most important variable. And for the football, it would be Goal count. Next, we need to examine the variables that are directly or indirectly correlated with the number of goals. After subtracting the values of these variables from the average of the teams and multiplying them by a ratio we determined, we will get a team score. Each variable will have a different coefficient according to its effect on the game. Our values that correlated with goal count are : shots per game, possesion, pass (directly); aerials

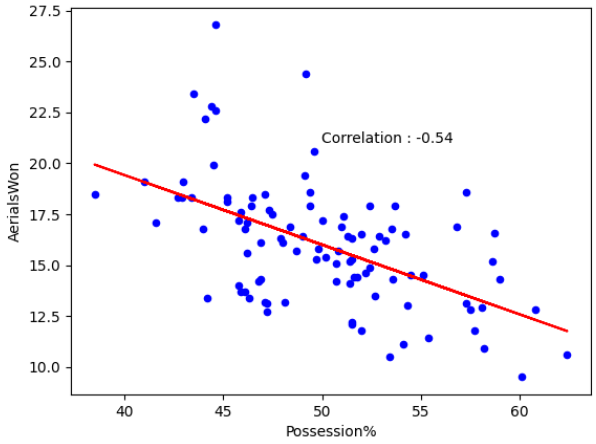

Fig. 14. Possession and Aerials won correlation.

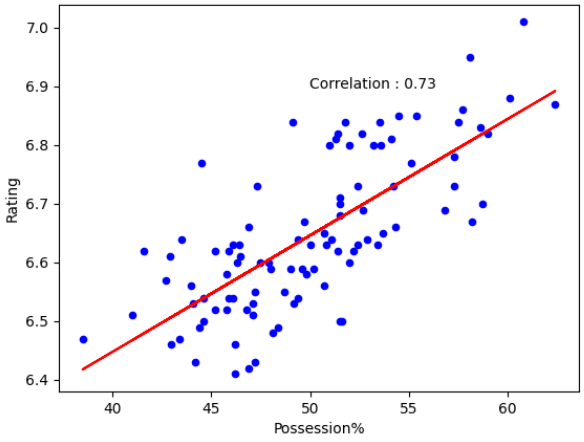

Fig. 15. Possession and Rating correlation.

won(indirectly). Then we could look at the matches between this teams and check if we did guess the winner or not.

A. *Determining coefficient according to its effect on the game*

Let's start with the most important. Goal count. But we need to take goal per match.We would say that coefficient of the goal is 3. When we compare the effect of the number of goals and shots on the match, we would give 0.7 to coefficient of the shot. In same way, possesion would be 0.3, pass would be 0.2. For the indirect variables, aerials won would be -0.1 (because it has a negatif correlation with possesion)

B. *Calculating our value and test it*

For example, let's take 2 team
Our Data Averages (Goal = 50/38 = 1.3 , Shots = 11.85, Possesion = 50, Pass = 80.44, Aerials Won = 16.01)
Manchester City (Goal = 83/38 = 2.18 , Shots = 15.8, Possesion = 60.8, Pass = 89.4, Aerials Won = 12.8), Score = 10.75
Paris Saint-Germen (Goal = 86/38 = 2.26 , Shots = 15, Possesion = 60.1, Pass = 89.5, Aerials Won = 9.5), Score = 10.57

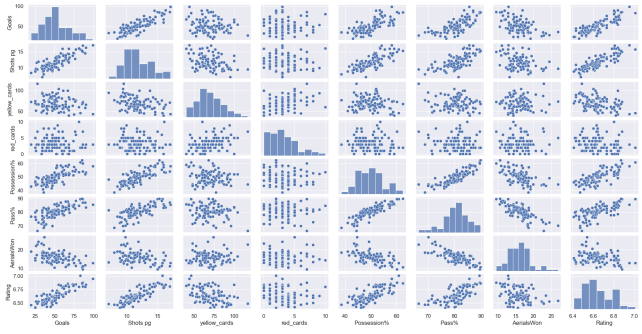

Fig. 16. All data pairs.

If you would look at their last match, Manchester City has won the match with 2-1 score. So our formula is worked! Let's check it for other 2 team
Sevilla (Goal = 53/38 = 1.39 , Shots = 12.7, Possesion = 58.7, Pass = 86.2, Aerials Won = 16.06) = 4.57
Borussia Dortmund (Goal = 75/38 = 1.97 , Shots = 14.6, Possesion = 57.2, Pass = 85.5, Aerials Won = 12.8) = 6.62
Again, if would you look at their last 2 match at this season, Dortmund has won 2 matches. So our formula is also worked in these teams.
And if would you look at the last 2 match between Borussia Dormund (which have 6.62 score) and Manchester City (which have 10.75 score), Manchester City has won 2 matches. Another checkpoint for our formula.

VII. Conclusion

Although we made 3 correct predictions, it should not be forgotten that there is always a luck factor in football and the match may depend on thousands of different parameters. Even if we take all of these parameters into account, the luck factor and the instantaneous performances of the players will make our analysis pointless. For this reason, although we know that we will never reach a definite conclusion, we can calculate who is more likely to win the match using various parameters and their relationship to each other.