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 gam 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.

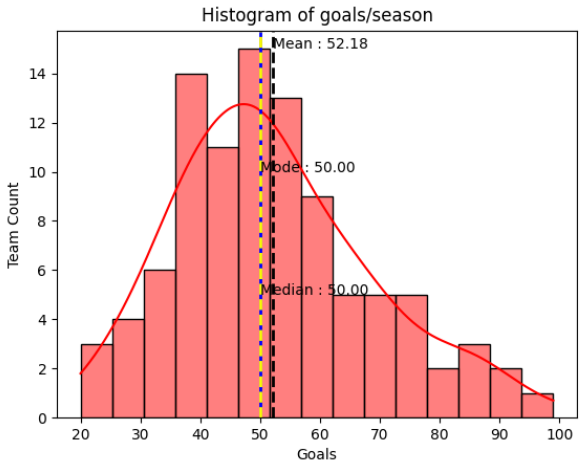

Fig. 1. Goals per season.

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.

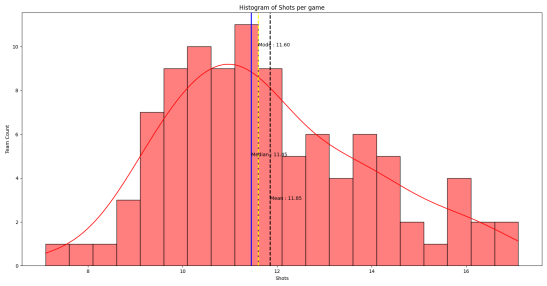

Fig. 2. Shots Per Game.

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

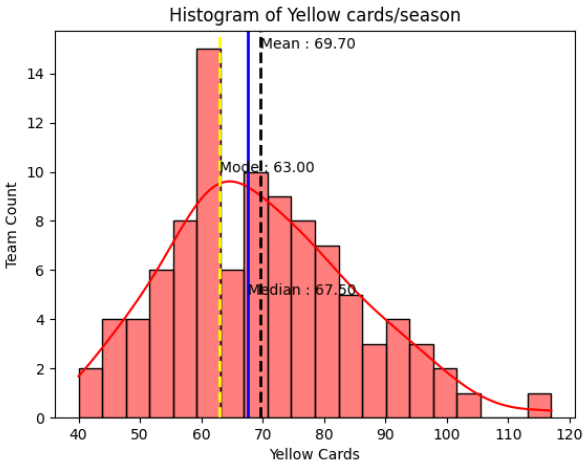

Fig. 3. Yellow card per season.

D. In-game advantages

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

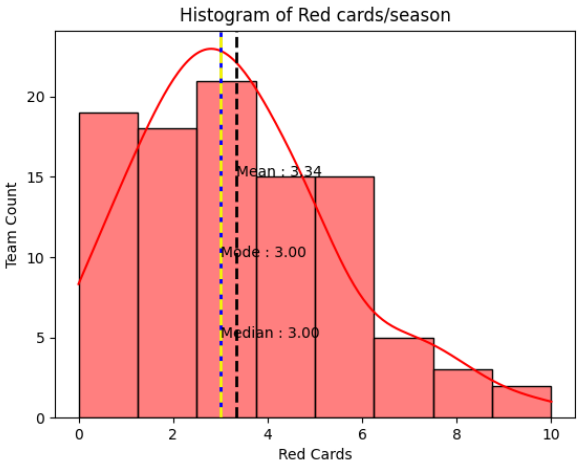

Fig. 4. Red card per season.

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

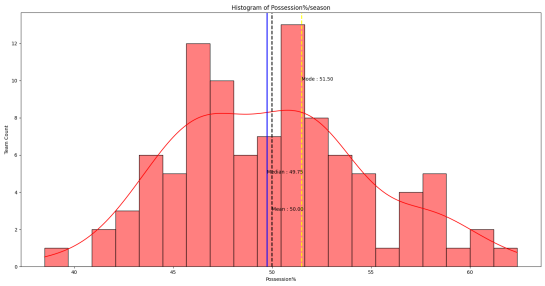

Fig. 5. Average possession%/season.

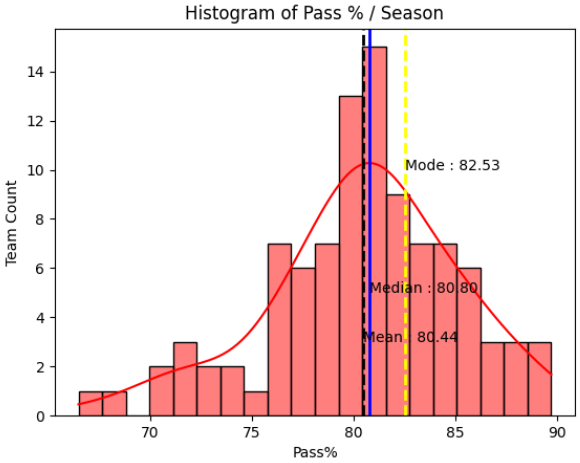

Fig. 6. Average pass accuracy%/season.

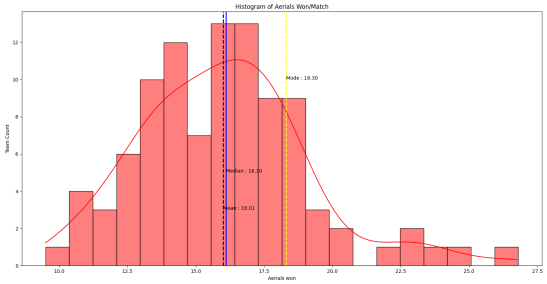

Fig. 7. Aerials won/per match.

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

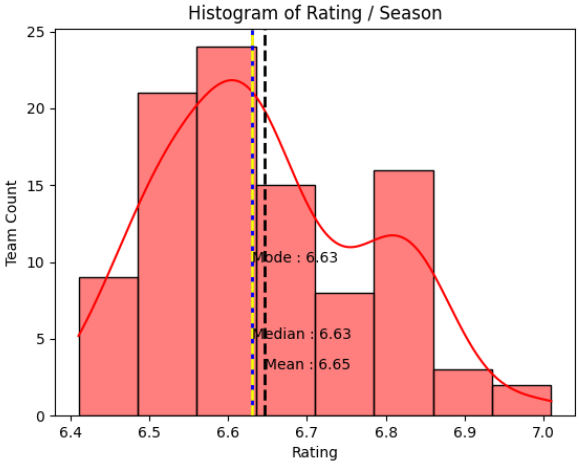

Fig. 8. Ratings in season.