EXAMINING INFLUENTIAL FACTORS AND PREDICTING OUTCOMES IN EUROPEAN SOCCER GAMES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yana Melnykov

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

March 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

EXAMINING INFLUENTIAL FACTORS AND PREDICTING

OUTCOMES IN EUROPEAN SOCCER GAMES

By

YANA MELNYKOV

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

RHONDA MAGEL
Chair

SEUNG WON HYUN

MEGAN ORR

ELIZABETH BIRMINGHAM

Approved:

3/25/2013	RHONDA MAGEL
Date	Department Chair

ABSTRACT

Models are developed using least squares regression and logistic regression to predict outcomes of European soccer games based on four variables related to the past k games of each team playing with the following values of k considered: 4, 6, 8, 10, and 12. Soccer games from the European soccer leagues of England, Italy, and Spain are considered for the 2011-2012 year. Each league has 20 teams playing two games with each other: one game is played at home; the other game is played away. There are 38 rounds in each league. The first 33 rounds are used to developed models to predict outcomes of games. Predictions are made for the last 5 rounds in each league. We were able to correctly predict 76% of the results for the last 5 rounds using the linear regression model and 77% of results correctly using the logistic regression model.

ACKNOWLEDGMENTS

First of all, I would like to thank Dr. Rhonda Magel for giving me a chance to be a student in the department of statistics, and second, and being my academic adviser. I appreciate the guidance she provided throughout my research. I would also like to thank Dr. Seung Won Hyun, Dr. Megan Orr and Dr. Elizabeth Birmingham for serving on my committee.

I would like to thank my mother, Banu Lebedeva, for everything she has done for me. Her love and support in all of my initiatives have helped me to become a person who I am now.

I also want to thank my parents-in-law, Tetyana Melnykova and Victor Melnykov, for their support and enormous help with everything and especially with my daughter.

I would like to thank my little girl, Nastasia, for giving me long sleep every night. I could only dream of such a calm and sweet child.

My beloved husband, Volodymyr, was the initiator of the idea to make a statistician of me. He was always ready to help me when I needed it. He was encouraging and guiding me from the beginning of my study. I am so thankful to him for everything that he has done for me and much more. I love him so much and I am glad to say that he is not a part but the whole my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF APPENDIX TABLES

CHAPTER 1. INTRODUCTION

The use of statistics in sports has become a topic that has drawn tremendous interest in the past several years. It has encompassed several sports and several different aspects of the sports. As some examples, Spencer, Lawrence, Rechichi, Bishop, Dawson and Goodman (2004) analyzed hockey data, Albright (1989) studied hitting streaks in baseball, and Tena and Forrest (2007) analyzed data on football coach dismissals.

In this paper, we will focus our analysis on soccer games. As with other sports, various aspects of soccer games have been studied. Hart, Hutton and Sharot (1975), constructed and estimated a model in which the response variable was the attendance of four English first division teams on Saturdays. There were three independent variables: the entrance fee, cost of alternative entertainment and level of personal income. The data the authors analyzed in this article were taken from three different seasons and all model parameters were estimated separately for all four clubs. The research for this article was done before a lot of sports data was made available online which drastically increased the amount of research conducted in sports statistics. Later research on soccer has included a study by Kellis and Katis (2007) who focused their research on soccer kicking biomechanics and studied effects that may cause a successful kick. Among the effects considered by the authors were the approach angle and distance, age and gender differences, ball speed, accuracy, and many others. The study was primarily focused on the magnitude of the moments while stretching all joints of foot and the time sequence of every moment during the kick. Rusu, Stoica, Burns, Hample, Mcgarry and Russell (2010) designed a system that included visualization tools that can help a soccer team manager. The developed application can compare players from two different teams applying multiple characteristics these players have.

Panaretos (2012) presented the talk “A statistical analysis of the European soccer champions league” at *the Joint Statistical Meetings* in 2012. He raised a question regarding the connection among some particular aspects of the game that can influence the average goal scoring. He used a linear regression model with the response variable being the number of goals that teams might score during the game. Two independent variables considered by the author were the ball possession percentage and the logarithm of the ratio between goals scored and goals received. Other independent variables such as fouls, the number of yellow and red cards, and the number of off side cases were found insignificant in the model, and therefore they were excluded from the consideration. Ridder, Cramer and Hopstaken (1993) studied various relationships in soccer game components and drew several conclusions. One thing they considered were red cards in soccer. If a player gets a red card, he is dismissed from the field and the team continues playing in a reduced compound against the opponent playing with all team members. The authors state that the chances to win decrease for the first team and increase for the second. Of course, this conclusion is logical and not surprising.

In this paper, we will focus on the significance of red cards and yellow cards for three top European soccer leagues: English, Spanish, and Italian. Ridder, et al. (1993) felt this was important as to which team would win the soccer match, but Panaretos’s (2012) research did not find the number of yellow or red cards a team received to be significant when other factors were considered. We would like to further investigate the significance of the number of red cards and yellow cards in winning a soccer match. The analysis conducted is based on the 2011-2012 year. Each league has 20 teams playing two games with each other: one game is played at home, the other game is played away. Thus, during the year each team plays 38 games in total. We found data representing the number of goals scored and the number of yellow and red cards with corresponding time records. Yellow cards are typically given for some rude or aggressive actions and red cards are presented for the second yellow card received or exceptional rudeness. Therefore, yellow and red cards reflect the game temper and can be important in our analysis. A soccer match includes two periods each lasting 45 minutes. Sometimes, if there are delays in the game, a referee can require playing several extra minutes. We split the entire game duration into 6 equal periods (15 minutes each) and count the number of cards and goals during each time interval. The analysis and applied methods will be described and illustrated further in the paper.

CHAPTER 2. DESCRIPTION OF RESEARCH QUESTIONS

There are many interesting questions that we attempt to answer in this paper using different types of tests. The tests and methods used will be described in Chapter 3. The purpose of this study is to identify factors which would most influence a team to win or lose in soccer. We will then try to predict the outcomes of soccer games based on examining these factors associated with both teams from previous games each of the two teams have played. Soccer games from the three European countries of England, Spain, and Italy will be considered. In addition to the above, the following questions will be addressed in this thesis:

1. Are the distributions of cards over the game time the same for all 3 countries?
2. Are the distributions of goals over the game time the same for all 3 countries?
3. Are the distributions of cards over the game time for the top 3 teams in each country compared with the remaining teams in each country the same? Tests will be done for England, Spain, and Italy.
4. Are the distributions of goals over the game time for the top 3 teams in each country compared with the remaining teams in each country the same? Tests will be done for England, Spain, and Italy.
5. Are the distributions of cards for the top 3 teams in each country compared with all the remaining teams in each country, collectively, the same? One test will be conducted combining all the countries.
6. Are the distributions of goals for the top 3 teams in each country compared with all the remaining teams in each country, collectively, the same? One test will be conducted combining all the countries.

7. Are the distributions of cards the same for the home teams versus the away teams in England (Spain, Italy)? Tests will be conducted separately for each country.
8. Are the distributions of goals the same for the home teams versus the away teams in England (Spain, Italy)? Tests will be conducted separately for each country.
9. Do teams in the three countries get cards equally often?
10. Do teams in the three countries score goals equally often?
11. Do the top teams in the three countries get cards equally often?
12. Do the top teams in the three countries score goals equally often?
13. Do English teams receive fewer (yellow/red) cards on average than Italian and Spanish teams?
14. Does the average number of (yellow/red) cards differ for the home and away teams in each of the countries? (England, Spain, Italy)?

CHAPTER 3. METHODS SECTION

Data was collected from 2011-2012 season for three countries*:* England, Spain, and Italy. There are 20 teams in each country. Since each of the three countries we considered had 20 teams playing two matches with each other (one at home and one away), we obtained 38 rounds in total for each of three countries. The duration of a game is 90 minutes. It consists of 2 parts, each lasting 45 minutes. We divided the two halves of a game into 15 minute periods: 0 – 15, 16– 30, 31 – 45 (end of the first half), 46 – 60, 61 – 75, 76 – 90 (end of the second half). It is noted that a game can be extended by several minutes in case there were necessary breaks in the game, for example, if a player was injured.

We developed a model based on the first 33 rounds to predict the results in the last 5 rounds. Regression techniques were used involving the following six independent variables:

- X1 – the sum of the differences between the number of goals scored by a home team and the number of goals scored by its opponents in the k previous rounds;
- X2 – the sum of the differences between the number of goals scored by an away team and the number of goals scored by its opponents in the k previous rounds;
- X3 – the sum of the differences between the number of cards received by a home team and the number of cards received by its opponents in the k previous rounds;
- X4 – the sum of the differences between the number of cards received by an away team and the number of cards received by its opponents in the k previous rounds;
- X5 and X6 – 2 indicator variables to represent 3 countries (England, Italy, and Spain).

It is noted that values of 4, 6, 8, 10 and 12 will be used for k and a decision will be made as to which value of k to use.

The dependent variable when the least squares regression technique is used can be expressed as $g_h - g_a$, where g_h and g_a are the number of goals scored by the home and guest teams respectively. For example, if the difference is 2, it means that the team playing at home won by scoring 2 more goals than the guest team. If the difference is -1, it means that the home team lost with the difference in one goal. A model based on least squares regression was used to predict the difference in goals scored by the home and away teams based on the six independent variables considered.

An alternative approach is to employ logistic regression. The advantage of this approach is related to the fact that this type of regression provides us with a probability assessment of success for both teams. One limitation of such an approach is that it works ideally when the response is binary. In our case, however, there are three possible results of a game. Therefore, the win and draw for the team playing at home are combined together. In other words, our logistic regression model focuses on predicting a win or draw for the team playing at home. Independent variables used in this case are the same as described before for the linear regression model.

Once the least squares linear regression and the logistic regression models were developed, they were used for predicting the winners in the 34th through 38th rounds in each country. Several models using the least squares regression technique and the logistic regression technique were developed using various values of k. These values of k were 4, 6, 8, 10 and 12. The independent variables placed in the model varied with the value of k, where k was the number of previous rounds of soccer considered in determining the value of four of the independent variables. The models will be compared and a recommendation made.

A Chi-square test of independence was used to address questions 1-8 from Chapter 2. The hypotheses may be stated as follows:

H_0: All populations have the same distribution of proportions over time

H_a: H_0 is not true.

Questions 9-12 of Chapter 2 may be addressed by using the Chi-square goodness of fit test and testing each of the proportions are equal to 1/3. The null and alternative hypotheses may be stated as follows:

$H_0: \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

H_a: H_0 is not true,

where the proportions π_1, π_2, $\pi_3$represent three countries England, Spain and Italy.

Question number 13 from Chapter 2 may be addressed by using a two sample independent t-test. The corresponding hypotheses will be considered by:

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 < \mu_2$,

where $\mu_1$is the population mean cards of England, μ_2 is the population mean cards of the Spain and Italy.

A paired t-test is used to address question 14 to compare the number of cards received by the home and away teams during each game. The hypotheses for the paired t-test are:

$H_0: \mu_d = 0$

$H_a: \mu_d \neq 0$,

where μ_d is the mean for the population differences in cards.

CHAPTER 4. RESULTS

4.1. Linear Regression Model

In this chapter, we will first present the results that we obtained from our linear regression model in which the response variable is the difference between the number of goals scored by the home team and the number of goals scored by the away team. Before we start, we will again briefly discuss the independent variables employed in the linear regression model. They are given below:

- X1 – the sum of the differences between the number of goals scored by a home team and the number of goals scored by its opponents in the k previous rounds;
- X2 – the sum of the differences between the number of goals scored by an away team and the number of goals scored by its opponents in the k previous rounds;
- X3 – the sum of the differences between the number of cards received by a home team and the number of cards received by its opponents in the k previous rounds;
- X4 – the sum of the differences between the number of cards received by an away team and the number of cards received by its opponents in the k previous rounds;
- X5 and X6 – 2 indicator variables to represent 3 countries (England, Italy, and Spain).

Regression models were developed on 5 values of k. These values were 4, 6, 8, 10, and 12. An even value of k was considered so that there were always an equal number of home and away games played by the team under consideration. There are 38 rounds of soccer in a championship series for each of the three countries. The parameters of the linear regression models were estimated based on the first 33 rounds. Different models were developed depending upon whether the independent variables entered into the model considered the last 4, 6, 8, 10 or 12 rounds. Once the parameters were estimated for each of the models under consideration based on the results from the first 33 rounds, the models were used to predict the results of the last 5 rounds.

Results from the ANOVA tables constructed for all the linear regression models indicate that the indicator variables for country (X5 and X6) are not significant (p >0.2). The three variables found to be significant in all of the models, except when the 12 previous rounds are considered were X1, X2, and X3. When 12 previous games are considered, the p-value associated with X2 was found to equal 0.054. It was determined, however, to use all the variables X1, X2, X3, and X4 since X1 was the sum of the differences of goals scored for the home team and their opponents for the k previous rounds, X2 was this for the away team, X3 was the sum of the differences between the number of cards received and their opponents for the k previous rounds, and X4 was this for the away team.

Based on the 5 models developed, the results based on the last 5 rounds of soccer in each country were predicted. The worst prediction rate was obtained for the model in which the independent random variables were based on the 6 previous games. This prediction rate was equal to 73%. The best prediction rate was obtained for the model in which the independent variables were based on the 10 previous games. This was 79%. The model in which the independent variables were based on the 4 previous games had a prediction rate of 75%, while the remaining models had a prediction rate of 76%. We have decided to employ the model with independent variables based on the 8 previous games since this model was fairly stable and we felt 10 or 12 previous games to consider was quite a lot. The ANOVA table for the selected model is provided in Table 1. Table 2 contains prediction results of the developed model for the 5 previous rounds (rows 34 - 38) in England, Spain and Italy. The first 10 columns represent 10 games played in England, while the other 20 correspond to the games played in Spain and Italy.

The symbol “+” represents a correct prediction, otherwise, an incorrect prediction is denoted “-”. If our model estimated the difference in goals between the home team and the away team to be 0 or higher, we predicted a win or draw for the home team. If our model estimated a negative difference, we predicted a loss for the home team. The ANOVA table of the full linear regression model based on the 8 previous games is given in Table 1. Table 2 compares the model prediction with the outcome for every game in the last 5 rounds.

Table 1. ANOVA table of the linear regression model based on the 8 previous games (indicator variables for countries are excluded)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	0.479558	0.061326	7.820	1.81e-14 ***
X1	0.052621	0.009977	5.274	1.75e-07 ***
X2	-0.032603	0.011663	-2.795	0.00532 **
X3	-0.056352	0.009957	-5.660	2.17e-08 ***
X4	0.014504	0.011739	1.236	0.21703

***= significant at α<0.001; **=significant at 0.001<α<0.01

Based on the linear regression model, we predicted the last 5 rounds with the precision of 76%. It is noted from Table 1, that the estimate of the coefficient associated with X1 is positive, and the estimate of the coefficient associated with X2 is negative. This indicates that the higher the sum of the differences in goals scored between the home team and their opponents for the 8 previous games, the better the chance of the home team winning the game. If the sum of the differences in goals between the away team and their opponents for the 8 previous games becomes higher, this decreases the chances that the home team will win. The estimated coefficients associated with X3 and X4 are negative and positive, respectively. This indicates that getting more cards decreases a teams’ chance of winning. The coefficients related to cards show that the more aggressive a team plays, the fewer chances it has to win. In other words, if a team receives fewer cards than the opponent it will have a greater opportunities to win.

Table 2. Model prediction based on the 8 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	+	+	+	+	+	-	+	-	-	-	+	+	-	+	+	+
35	+	+	+	+	-	+	+	+	-	+	+	+	-	+	-	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	+	+	-	+	+	+	+
38	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	-	+	+	+	-	+	**22/30**	
35	+	+	+	-	+	+	+	+	+	-	-	+	+	+	**23/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	-	**23/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**23/30**	
38	-	+	+	-	+	+	+	-	+	+	+	+	+	-	**23/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

4.2. Logistic Regression Model

The same independent variables that were considered for the least squares model to predict point spread are also considered for the logistic regression models. The logistic regression models are developed to estimate the probability that a team will win or tie. In predicting outcomes of games using a logistic regression model, we will predict a win or draw if the probability of winning or drawing is estimated to be greater than 0.5. Otherwise, we will predict a loss.

Five different logistic regression models were developed. These depend on whether the independent variables entered into the model were based on the 4, 6, 8, 10 or 12 previous rounds.

Table 3. ANOVA table of logistic regression model based on the 8 previous games (including all variables)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	1.20927	0.15195	7.958	1.75e-15 ***
X1	0.04019	0.01452	2.767	0.00565 **
X2	-0.04292	0.01656	-2.592	0.00954 **
X3	-0.05323	0.01332	-3.996	6.45e-05 ***
X4	0.02475	0.01599	1.548	0.12173
X5	-0.14423	0.21029	-0.686	0.49280
X6	-0.41549	0.20465	-2.030	0.04233 *

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table 4. Model prediction based on the 8 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Total**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	-	+	+	+	+	+	+	+	+	-	-	-	+	+	**24/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	-	**22/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**77%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

The indicator variables for country were not significant except for the indicator variable for England when the 8 previous games were used. Once again, only the variables X1, X2, and

X3 were significant in all of the models, but X4 was kept in all of the models since this made more sense.

Table 5. ANOVA table of logistic regression model based on the 8 previous games (indicator variables for countries are excluded)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	1.01614	0.08711	11.664	< 2e-16 ***
X1	0.04027	0.01454	2.771	0.0056**
X2	-0.04209	0.01647	-2.556	0.0106 **
X3	-0.05316	0.01334	-3.986	6.72e-05 ***
X4	0.02420	0.01594	1.518	0.1289

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table 6. Model prediction based on the 8 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Total**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	-	+	+	+	+	+	+	+	+	-	-	+	+	+	**25/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	-	**22/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**77%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

The prediction rates for all of the models were close, and therefore, we used the model with the independent variables based on the 8 previous games. As an example, the prediction rate for the last 5 rounds based on 10 previous games was 78%, for 12 previous games, it was 80%, and for 8 previous games, it was 77%. The ANOVA table for the logistic model based on independent variables using the 8 previous games is provided in Table 3. Table 4 gives the predictions for the next 5 rounds for all the teams in each of the countries based on the logistic regression model with independent variables based on the 8 previous games. In the table, the symbol "+" indicates a correct result and the symbol "-" denotes an incorrect result. Since the indicator variables for country were not significant for most of the logistic models they were dropped from consideration. We can see from Table 6 that the prediction rate for this model is again 77%. The lowest prediction rate is 74% for the model based on the 6 previous games. The model based on the 4 previous games had a 76% prediction rate. The prediction rates for the models based on the 10 and 12 previous games were 79% and 80%, respectfully. Again we see that the prediction rates are very close, therefore, we were satisfied with the model based on the 8 previous games from Table 6.

4.3. Results

We analyzed the distribution of goals and cards from the championship series soccer games from England, Spain and Italy. In this part of the results section we addressed the 14 questions from Chapter 2 regarding these 3 countries. Each question is answered and the results are explained one by one. We divided the game time by periods of 15 minutes. All the data given are based on all 38 rounds.

4.3.1. Are the distributions of cards over the game time the same for all 3 countries?

The number of cards (yellow and red) received by teams playing at home for all three countries are given in Table 7.

H_0: The proportions of cards given to home teams in each time period are the same for all three countries

H_a: The proportions are not the same

Table 7. Number of cards for teams playing at home in each time period

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
England	*30*	*65*	*91*	*80*	*118*	*136*
Spain	*41*	*127*	*191*	*143*	*177*	*245*
Italy	*43*	*86*	*149*	*121*	*144*	*218*

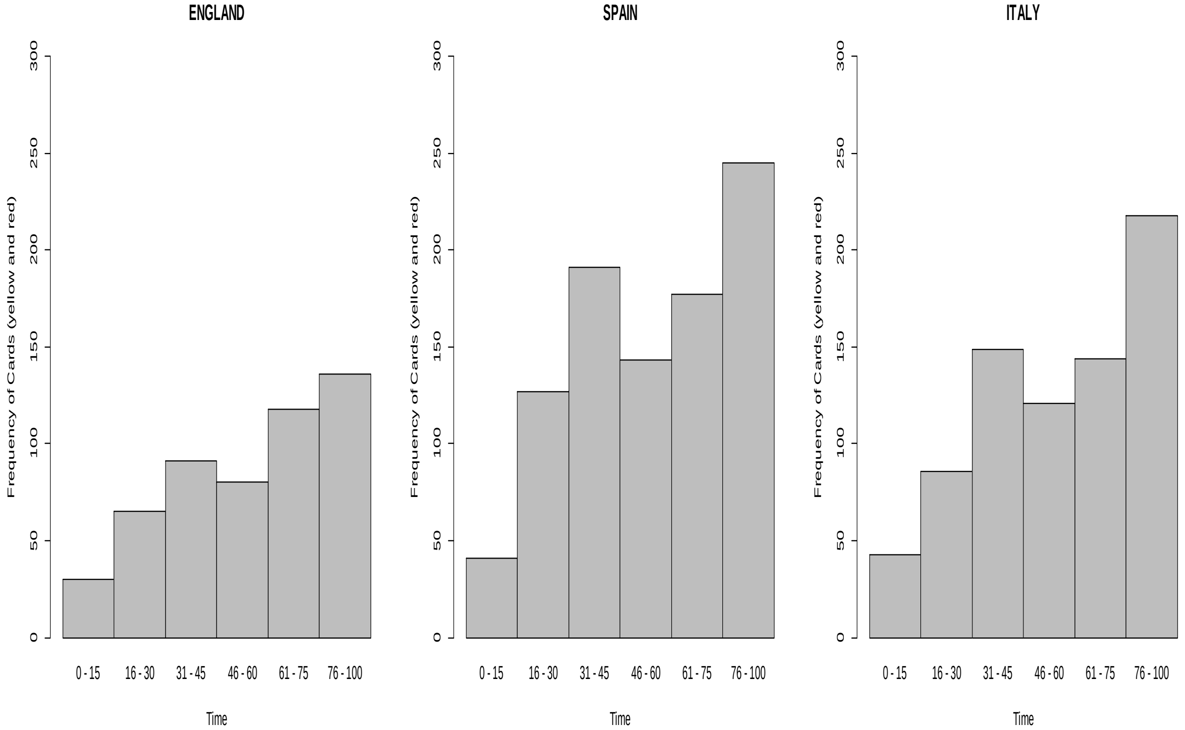

Figure 1. Number of cards given to home teams in each time period for all three countries

From Figure 1, we see that the patterns corresponding to the 3 countries are similar to each other. All three countries start playing more and more aggressive closer to the end of the game. From Table 1, we notice that the number of cards in England is much lower than that in Italy and Spain. This can be explained by the fact that English teams tend to play rude and get cards only for very serious violations. The p-value for the chi-square test is 0.471 which is saying that there is no enough evidence to reject H_0. We conclude that there is no difference in countries' distributions of cards during each period for teams playing at home.

The number of cards (yellow and red) received by teams playing away for all three countries are given in Table 8.

H_0: The proportions of cards given to away teams in each time period are the same for all three countries

H_a: The proportions are not the same

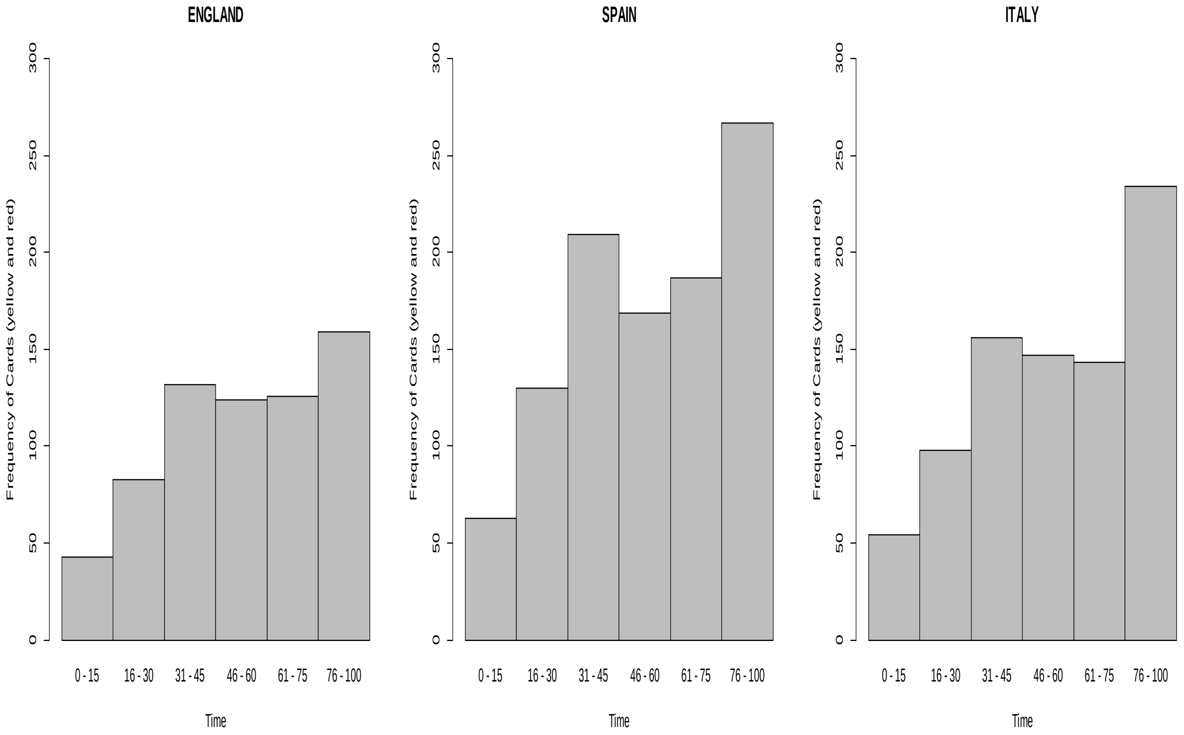

Figure 2. Number of cards given to away teams in each time period for all three countries

Examining Figure 2, we see that all patterns for the different countries are similar in this case too. All away teams from the three countries play ruder starting in the second half of the game. The p-value for the chi-square test is 0.867 which indicates there is no difference in the distribution of cards for the time periods between away teams in the different countries. Overall, comparing Figure 1 and Figure 2, it appears that teams playing away may behave more aggressively than teams playing at home in all three countries. We already know that the aggressiveness does not lead to the victory.

Table 8. Number of cards for teams playing away in each time period

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
England	*43*	*83*	*132*	*124*	*126*	*159*
Spain	*63*	*130*	*209*	*169*	*187*	*267*
Italy	*54*	*98*	*156*	*147*	*143*	*234*

4.3.2. Are the distributions of goals over the game time the same for all 3 countries?

In Figure 3, it is noted that teams playing at home appear to have similar distributions of goals in the time periods for the three countries. The p-value for the chi-square test is 0.948 which implies there is no evidence to indicate a difference in the proportions of goals scored in the time periods between the three countries. There appears to be a slightly higher proportion of goals scored near the end of the game for all three countries.

Table 9. Number of goals scored by teams playing at home in each time period

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
England	*80*	*87*	*100*	*97*	*116*	*124*
Spain	*85*	*90*	*100*	*109*	*118*	*136*
Italy	*75*	*92*	*86*	*99*	*92*	*129*

H_0: The proportions of goals scored by home teams in each time period are the same for all three countries

H_a: The proportions are not the same

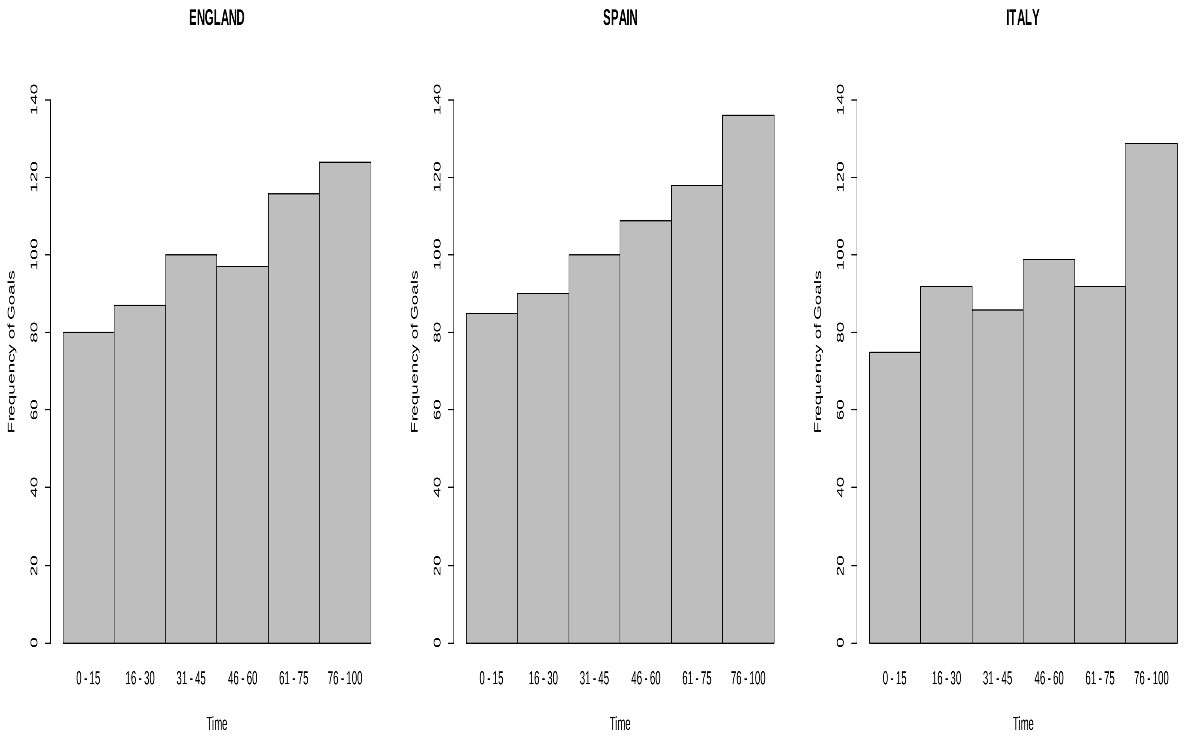

Figure 3. Number of goals scored by home teams in each time period for all three countries

We considered the same test but for the away teams. The number of goals for teams playing away is given in Table 10 for each time period.

Table 10. Number of goals scored by teams playing away in each time period

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
England	*56*	*62*	*80*	*76*	*76*	*112*
Spain	*46*	*59*	*69*	*57*	*78*	*103*
Italy	*49*	*57*	*54*	*75*	*53*	*111*

H_0: The proportions of goals scored by away teams in each time period are the same for all three countries

H_a: The proportions are not the same

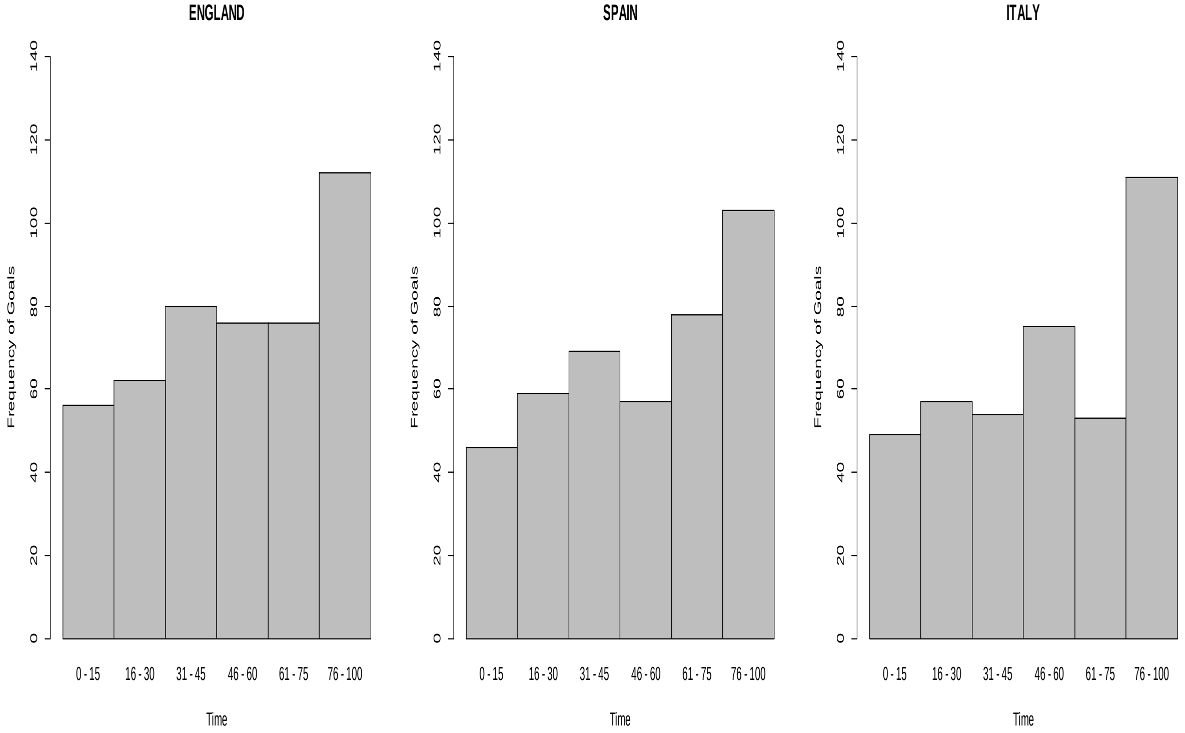

Figure 4. Number of goals scored by away teams in each time period for all three countries

In Figure 4, all three countries have a higher proportion of goals scored by away teams towards the end of the game. The number of goals scored for away teams of all countries start approximately at the same level and the number of goals scored by away teams generally increases over the time periods. The p-value associated with the chi-square test is 0.374 and hence, we fail to reject H_0. We conclude that there is no difference in the proportions of goals scored by away teams in each of the time periods between these 3 countries.

4.3.3. Are the distributions of cards over the game time for the top 3 teams in each country compared with the remaining teams in each country the same? Tests will be done for England, Spain, and Italy.

We first analyze the distributions of cards for the top 3 teams and remaining teams in England. The number of cards for the teams playing at home is provided in Table 11.

Table 11. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*5*	*12*	*14*	*14*	*16*	*15*
Other teams	*25*	*53*	*77*	*66*	*102*	*121*

H_0: The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in England

H_a: The proportions are not the same

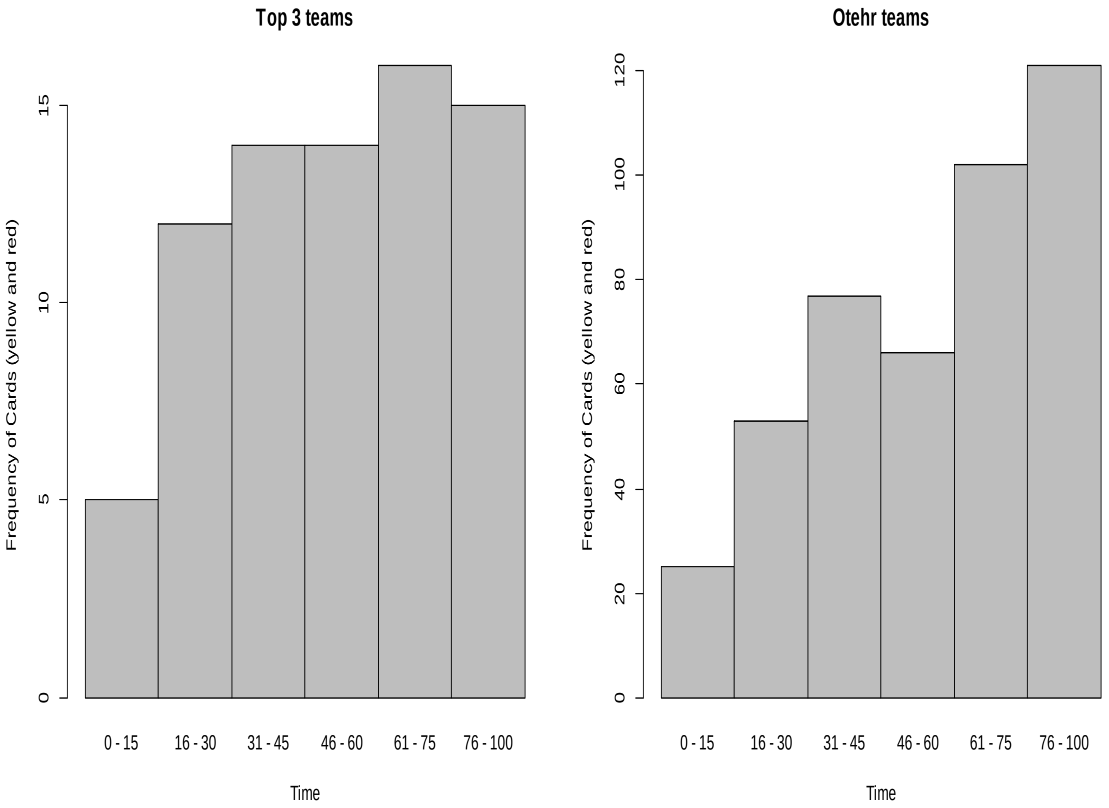

Figure 5. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in England

Visually it appears that the top 3 teams receive lots of cards, and therefore, play more aggressively than the rest of the teams. This is only because the frequency scale is different. Both patterns are similar toward the end of the game. It appears that after the first time period, the number of cards received by the top 3 teams is approximately uniform in the remaining time periods. The p-value for this test, 0.71, is very high however, and we fail to reject H_0.

We continue testing whether there is a difference between the mean number of cards given to the top 3 teams in England when playing at home versus the mean number of cards for the remaining team playing at home. The results is that we fail to reject the null hypothesis saying that there is no difference between mean number of getting cards for both top 3 and remaining teams playing at home since the p-value is 0.805. The mean number of cards per game the top 3 teams receive at home is 1.33 and the mean number of cards per game the remaining teams receive is 1.37 (T = -0.25, the degrees of freedom=378).

We next consider the number of cards for the same top 3 English teams playing away versus the other English teams playing away. The number of cards received by the away teams in each period is given in Table 12.

H_0: The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in England

H_a: The proportions are not the same

Table 12. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*7*	*10*	*18*	*19*	*18*	*23*
Other teams	*36*	*73*	*114*	*105*	*108*	*136*

The histograms for the number of cards given to the away teams in England during each time period are given in Figure 6. We see that the two histograms are similar in shape. The p-value for the test is 0.986 which implies we fail to reject H_0. There is not any evidence to indicate the proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are different.

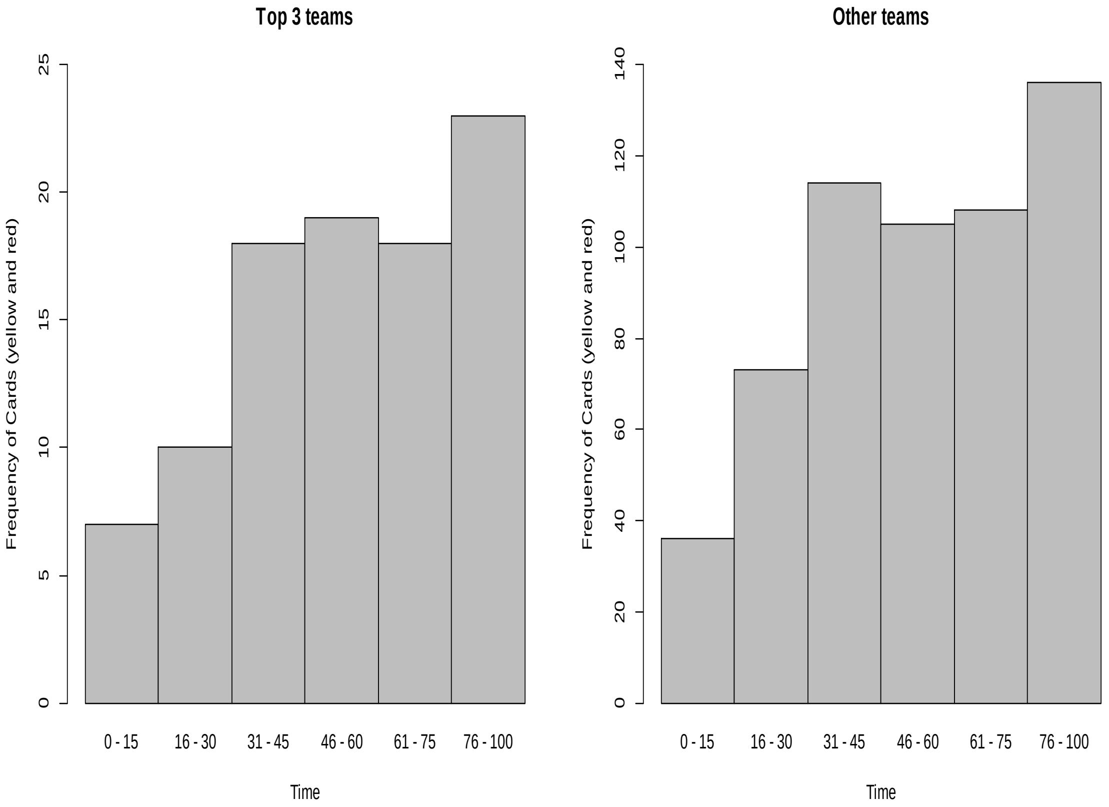

Figure 6. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in England

We conduct an additional test to test whether there is a difference between the mean number of cards given to the top 3 teams in England when playing away versus the mean number of cards for the remaining teams playing away as we have done for the teams playing at home in England. The p-value is 0.60 and we fail to reject the null hypothesis saying that there is no difference between the mean number of cards given to the top 3 teams and the remaining teams playing away in England. The mean number of cards per game the top 3 teams receive away is 1.67 and the mean number of cards per game the remaining teams receive is 1.77 (T = -0.52, the degree of freedom=378).

We also decided to test for a mean difference in the number of cards received by the top 3 teams in England playing at home versus when they play away. A two-sample t-test was conducted and the associated p-value was found to be 0.07 (T= -1.48, degrees of freedom= 112) The sample mean number of cards for the top 3 teams when playing at home was 1.33 per game and the sample mean number of cards for the top 3 teams playing away was 1.67. It was concluded that there was a marginally significant difference with the top 3 teams receiving more cards on average when they were playing away. A test was also conducted to test for a mean difference in the number of cards received by the remaining English teams playing at home versus the mean number of cards they receive per game while playing away. A two-sample t-test was conducted and the associated p-value was less than 0.001 (T= -3.87, degrees of freedom= 644). The sample mean number of cards for the remaining teams playing at home was 1.37, and the sample mean number of cards received while playing away was 1.77. It was concluded that the teams received significantly more cards on average while playing away.

Secondly, we conduct tests for Spain. The procedures will be the same as we have done for England. The number of cards for Spanish teams playing at home is provided in Table 13.

Table 13. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*4*	*17*	*19*	*26*	*24*	*23*
Other teams	*37*	*110*	*172*	*117*	*153*	*222*

H_0: The proportions of cards given in each time period to top 3 teams playing at home and the remaining teams playing at home are the same in Spain

H_a: The proportions are not the same

The two histograms in Figure 7 look somewhat different. For the first 15 minutes the proportions of cards given are approximately the same. However, during the next 15 minutes the top 3 teams appear to be more aggressive and this continues until the end of the first half of the game. The top 3 Spanish teams also start the second half more aggressively, but then the other Spanish teams intensify their aggressiveness and the aggressiveness of the top 3 teams appears to decrease. Nevertheless, the p-value, 0.147, is high enough to fail to reject H_0. Perhaps this would become significant if more games were considered.

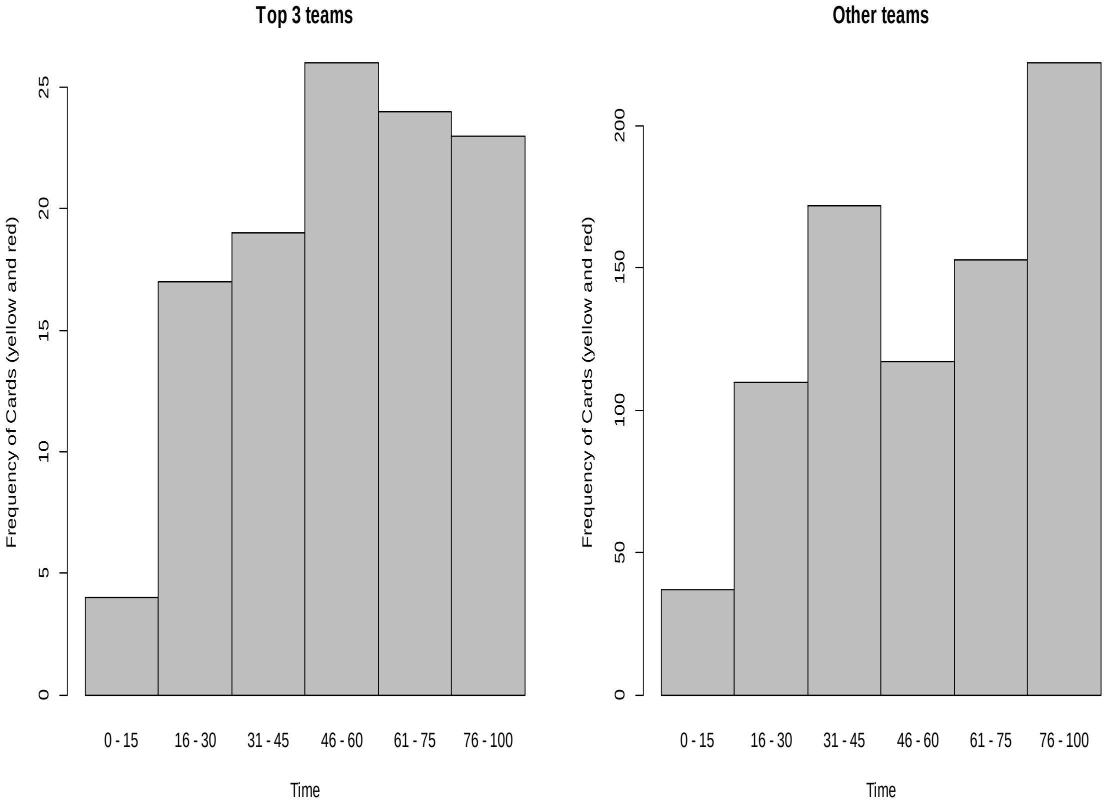

Figure 7. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in Spain

We also test whether there is a difference between the mean numbers of cards given to the top 3 teams in Spain when playing at home versus the mean number of cards for the remaining team playing at home. The p-value is 0.024 and we conclude that the top 3 teams playing at home get fewer cards on the average than the remaining teams. The mean number of cards per game the top 3 teams receive at home is 1.98 and the mean number of cards per game the remaining teams receive is 2.51 (T = -2.27, the degrees of freedom=378).

We next continue to consider the top 3 Spanish teams versus the remaining Spanish teams when teams are playing away. The number of cards for Spanish teams playing away is provided in Table 14. The hypotheses are shown below:

H_0: The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in Spain

H_a: The proportions are not the same

Table 14. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*9*	*16*	*35*	*28*	*24*	*45*
Other teams	*54*	*114*	*174*	*141*	*163*	*222*

The histograms are similar in this case except for during the 61-75 minute period. The p-value is 0.72 and we certainly fail to reject H_0.

We conduct the additional test to test whether there is a difference between the mean number of cards given to the top 3 teams playing away and the mean number of cards for the remaining teams playing away in Spain as we have done for the teams playing at home in Spain. The p-value is 0.79 (T= 0.264, the degree of freedom= 378) and we fail to reject the null hypothesis saying that there is no difference between the mean number of cards given to the top 3 teams and the remaining teams playing away in Spain. The mean number of cards per game the top 3 teams receive away is 2.75 and the mean number of cards per game the remaining teams receive is 2.69.

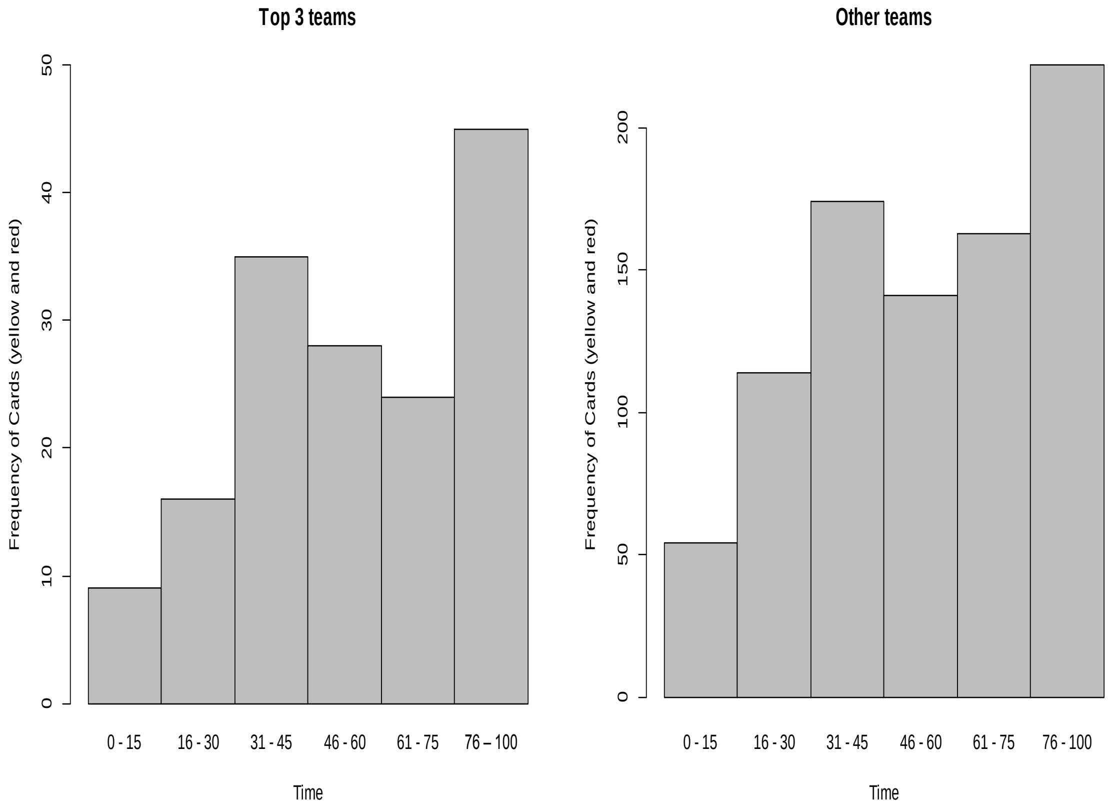

Figure 8. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in Spain

We also decided to test for a mean difference in the number of cards received by the top 3 teams in Spain playing at home versus when they play away. A two-sample t-test was conducted and the associated p-value was found to be 0.01 (T= -2.34, degrees of freedom= 112). The sample mean number of cards for the top teams when playing at home was 1.98 per game and the sample mean number of cards for the top 3 teams playing away was 2.75. It was concluded that there was a significant difference with the top 3 teams receiving more cards on average when they were playing away. A test was also conducted to test for a mean difference in the number of cards received by the remaining Spanish teams playing at home versus the mean number of cards they receive per game while playing away. A two-sample t-test was conducted and the associated p-value was 0.092 (T= -1.33; degrees of freedom=644). The sample mean number of cards for the remaining teams playing at home was 2.51, and the sample mean number of cards received while playing away was 2.69. It was concluded that the teams received marginally significantly more cards on average while playing away.

We finally analyze the top 3 teams and the remaining teams in Italy. The number of cards for Italian teams playing at home is provided in Table 15.

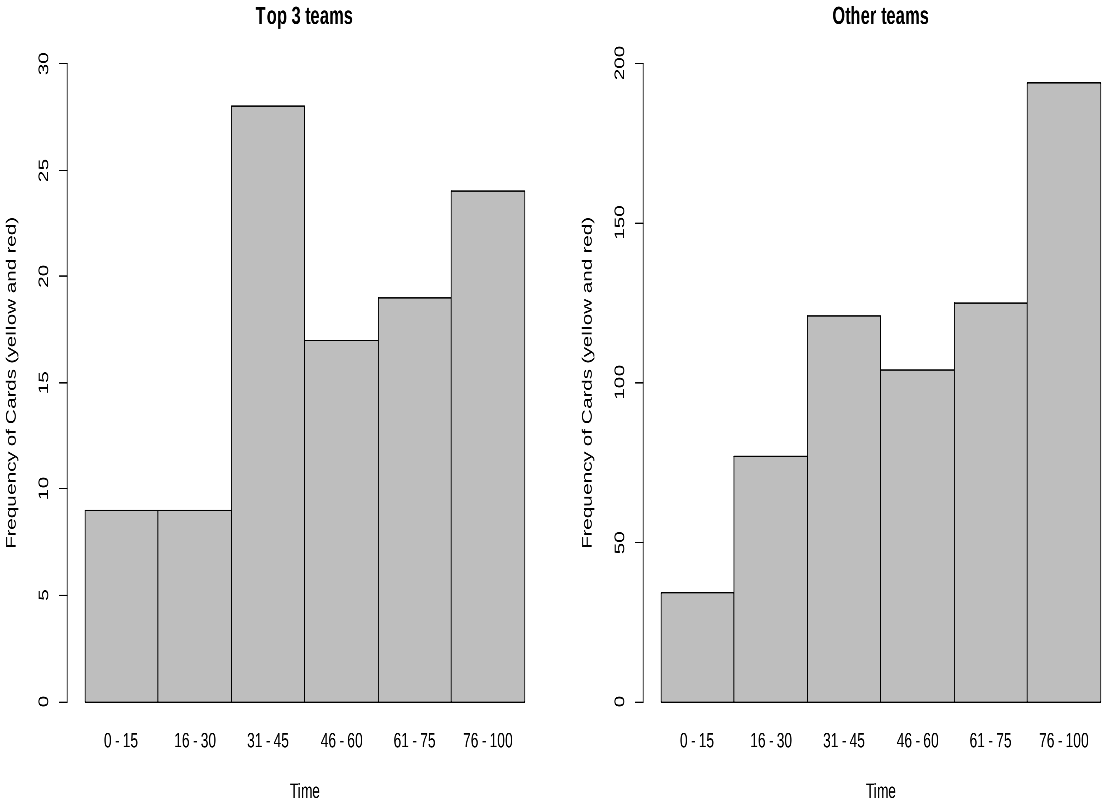

Figure 9. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in Italy

H_0: The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in Italy

H_a: The proportions are not the same

Table 15. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*9*	*9*	*28*	*17*	*19*	*24*
Other teams	*34*	*77*	*121*	*104*	*125*	*194*

The two histograms appear to be a little different from each other during the first half of the game with the top 3 teams playing more aggressively towards the end of the first half. The histogram of cards for the top 3 is quite different from those that we have seen previously. The two histograms appear to be similar in the second half of the game. The null hypothesis is not rejected with a p-value of 0.208.

We conduct a test to test whether there is a difference between the mean number of cards given to the top 3 teams playing at home and the mean number of cards for the remaining teams playing at home in Italy. The p-value is 0.37 and we fail to reject the null hypothesis and conclude that there is no difference between the mean number of cards received for the top 3 playing at home and the remaining teams playing at home in Italy. The mean number of cards per game the top 3 teams receive at home is 1.86 and the mean number of cards per game the remaining teams receive is 2.03 (T = -0.897, degrees of freedom= 378).

We next test the hypothesis for the Italian teams playing away.

H_0: The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in Italy

H_a: The proportions are not the same

Table 16. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*2*	*14*	*25*	*26*	*15*	*37*
Other teams	*52*	*84*	*131*	*121*	*128*	*197*

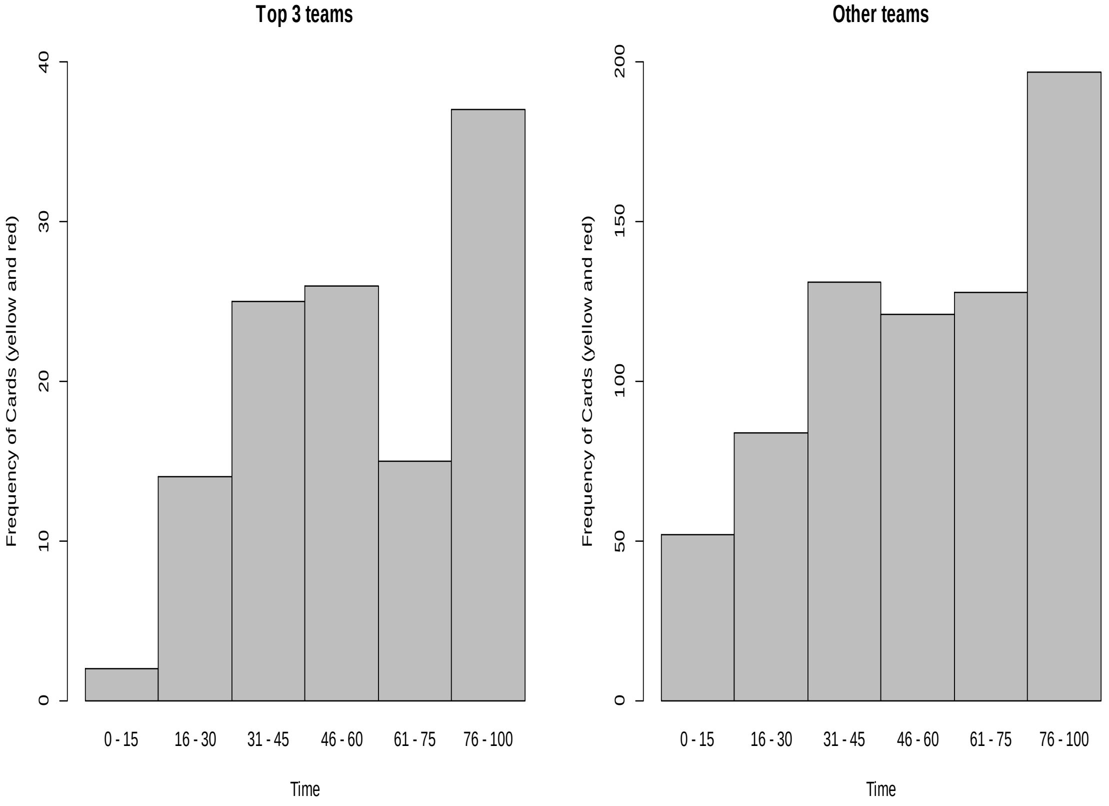

Figure 10. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in Italy

From Figure 10, it appears that the top 3 teams in Italy start off a lot less aggressive than the remaining teams, but increase their aggressiveness after the first time period. Overall the p-value is 0.116 so that we fail to reject H_0.

We conduct a test to test whether there is a difference between the mean number of cards given to the top 3 teams playing away and the mean number of cards for the remaining teams playing away in Italy. The p-value is 0.56 and we fail to reject the null hypothesis and conclude that there is no difference between mean number of cards received by the top 3 teams playing at home and the remaining teams playing at home in Italy (T= -0.58, the degrees of freedom=378). The mean number of cards per game the top 3 teams receive away is 2.09 and the mean number of cards per game the remaining teams receive is 2.21.

We also decided to test for a mean difference in the number of cards received by the top 3 teams in Italy playing at home versus when they play away. A two-sample t-test was conducted and the associated p-value was found to be 0.192 (T= -0.87, degrees of freedom=112). The sample mean number of cards for the top teams when playing at home was 1.86 per game and the sample mean number of cards for the top 3 teams playing away was 2.09. It was concluded that there was no significant difference between the mean numbers of cards the top 3 teams receives playing at home versus the mean number of cards they receive when playing away. A test was also conducted to test for a mean difference in the number of cards received by the remaining Italian teams playing at home versus the mean number of cards they receive per game while playing away. A two-sample t-test was conducted and the associated p-value was 0.047 (T= -1.68; degrees of freedom=644). The sample mean number of cards for the remaining teams playing at home was 2.03, and the sample mean number of cards received while playing away was 2.21. It was concluded that the teams received significantly more cards on average while playing away.

4.3.4. Are the distributions of goals over the game time for the top 3 teams in each country compared with the remaining teams in each country the same? Tests will be done for England, Spain, and Italy.

We conducted chi-square tests to test for the differences in distributions of goals over the time periods between the top 3 teams versus the remaining teams. The first set of hypotheses is the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in England

H_a: The proportions are not the same

It appears that the remaining teams score goals more uniformly than the top 3 teams. We can see that the top teams score few goals at the beginning and then a lot more at the end of the game. The p-value, 0.105, is high enough to fail to reject H_0.

Table 17. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*13*	*15*	*26*	*25*	*37*	*30*
Other teams	*67*	*72*	*74*	*72*	*79*	*94*

We conduct a test to test whether there is a difference between the mean number of goals scored by the top 3 teams playing at home and the mean number of cards for the remaining teams playing at home in England. The p-value is less than 0.001 and we reject the null hypothesis saying that there is difference between mean number of scoring goals of the top 3 teams and remaining teams playing at home in England (T= 6.3, the degrees of freedom= 378). The sample mean number of goals per game at home scored by the top 3 teams is 2.56 and the sample mean number of goals scored at home by the remaining teams is 1.42. The top 3 teams score significantly more goals at home.

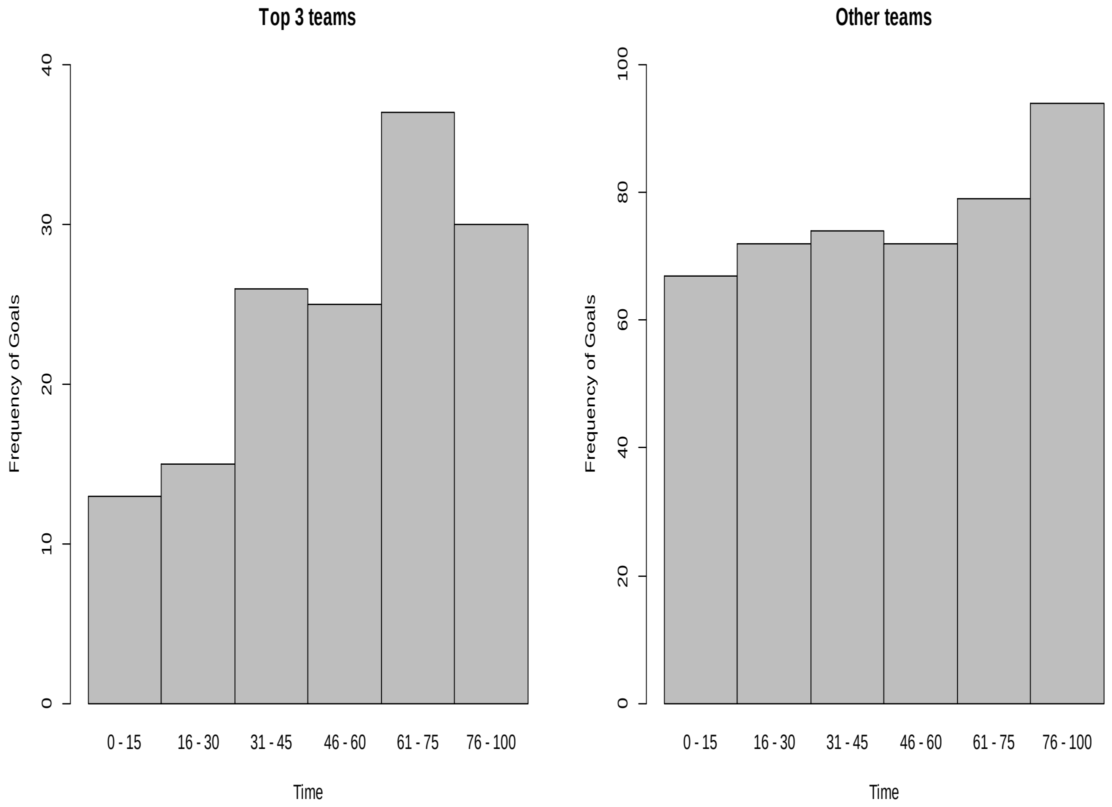

Figure 11. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in England

The same set of hypotheses is tested on the teams playing away in England.

H_0: The proportions of goals made in each time period by the top 3 teams playing away and the remaining teams playing away are the same in England

H_a: The proportions are not the same

From Figure 12, we see that the top teams in England playing away try to score more goals at the beginning than the remaining teams. The p-value is 0.486 and we fail to say that there is a difference in the distribution of goals for the top 3 teams and remaining teams playing away in England.

Table 18. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*14*	*21*	*16*	*17*	*17*	*25*
Other teams	*42*	*41*	*64*	*59*	*59*	*87*

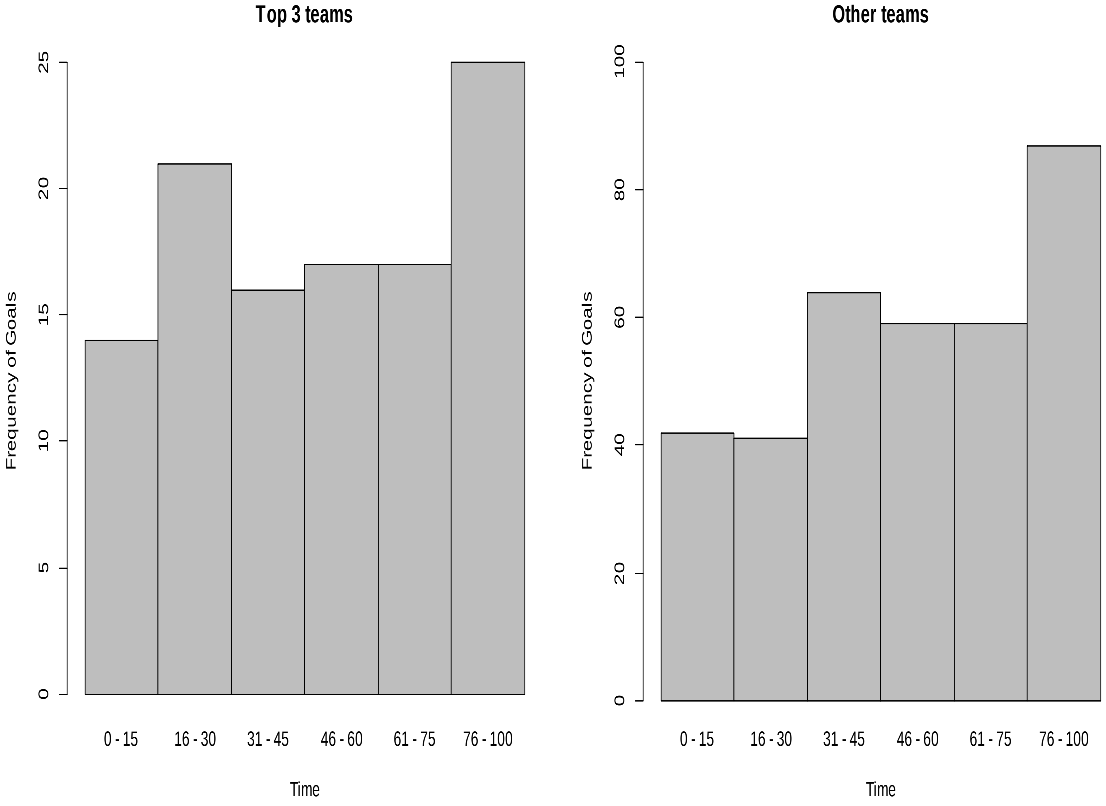

Figure 12. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in England

We also want to check whether there is a difference between the mean number of goals scored by the top 3 teams playing away and the mean number of cards for the remaining teams playing away in England. The p-value is less than 0.001 again and we reject the null hypothesis saying that there is difference between mean number of scoring goals of the top 3 teams and remaining teams playing away in England (T = 5, the degrees of freedom= 378). The top 3 teams score an average of 1.93 goals per game while playing at away and the remaining teams score an average of 1.09 goals per game while playing away. The top 3 teams score significantly more goals than the other teams while playing away.

We conducted additional two-sample t-tests to test if the mean number of goals scored by teams playing at home is greater than the mean number of goals scored by teams playing away. This test was conducted for the top 3 teams and then for the remaining teams. The p-value for the test when testing for the top 3 teams was 0.019. The sample mean number of goals scored at home per game for the top 3 teams was 2.56, and the sample mean number of goals scored away per game for the top 3 games was 1.93 (T= 2.11; degrees of freedom= 112). The top 3 teams in England score significantly more goals on the average at home games than away games. The p-value for the test when testing for the remaining teams was less than 0.001. The sample mean number of goals scored at home per game for the remaining teams was 1.42, and the sample mean number of goals scored away per game for the remaining teams was 1.09 (T= 3.65; degrees of freedom= 644)

The next set of hypotheses is conducted for Spain. The hypotheses are the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in Spain

H_a: The proportions are not the same

Table 19. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*27*	*22*	*32*	*32*	*31*	*38*
Other teams	*28*	*68*	*68*	*77*	*87*	*98*

From Figure 13, it appears that the distribution of goals scored during the time periods for the top 3 teams in Spain is similar to the distribution of goals scored for the remaining teams when the teams are playing at home. The chi-square test yielded a p-value 0.832 which implies we fail to reject the null hypothesis.

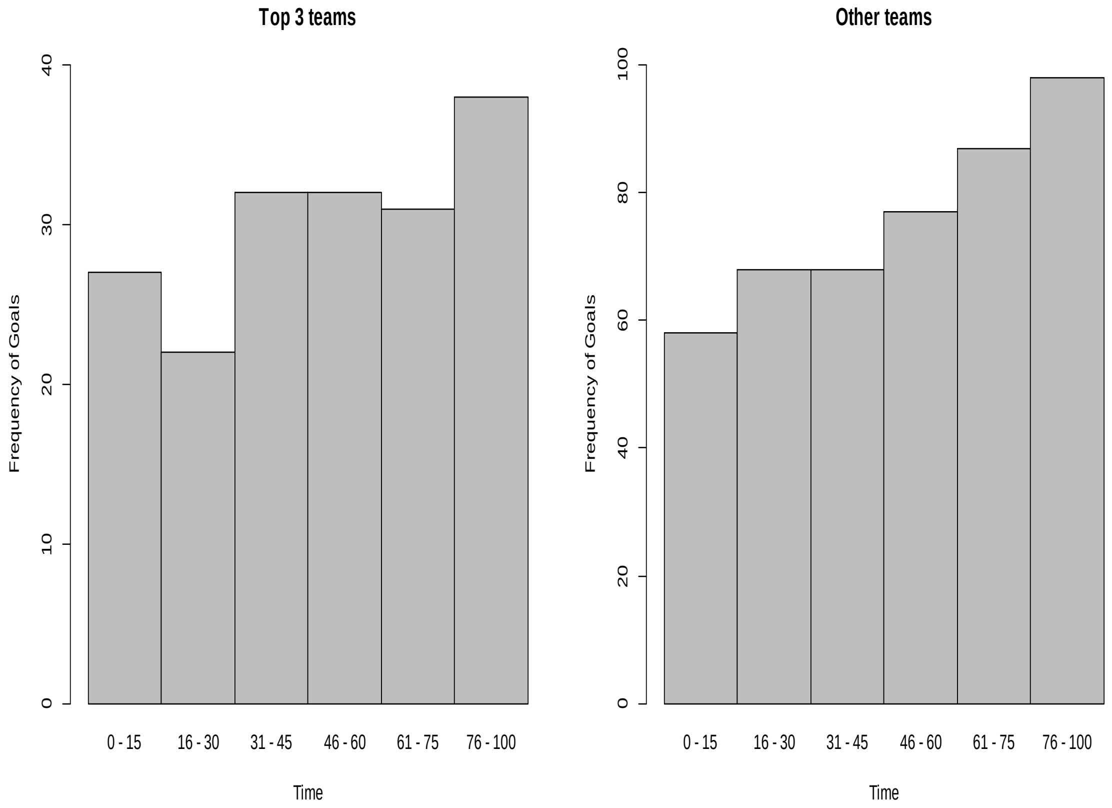

Figure 13. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in Spain

We conduct a test to test whether there is a difference between the mean number of goals scored by the top 3 teams playing at home and the mean number of cards for the remaining teams playing at home in Spain. The p-value is less than 0.001 and we reject the null hypothesis and conclude that the top 3 teams score significantly more goals on average when playing at home than the remaining teams (T= 9.43, the degrees of freedom= 378). The sample average number of goals per game for the top 3 teams playing at home was 3.19 and for the remaining teams, it was 1.41.

The hypotheses for the distribution of goals for the teams playing away for the same country are the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing away and the remaining teams playing away are the same in Spain

H_a: The proportions are not the same

From Figure 14, it appears that the distributions of the number of goals scored in the time periods are the same for the top 3 teams playing away and the remaining teams playing away in Spain with the possible exception of the number of goals scored at the start of the second half of the game. The p-value for the chi-square test is 0.385 and we fail to reject H_0 again.

Table 20. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*12*	*22*	*17*	*12*	*18*	*30*
Other teams	*34*	*37*	*52*	*45*	*60*	*73*

We want to test whether there is a difference between the mean number of goals scored by the top 3 teams playing away and the mean number of cards for the remaining team playing away in Spain. The p-value is less than 0.001 and we reject the null hypothesis saying that there is difference between mean number of scoring goals of the top 3 teams and remaining teams playing at home in Spain (T= 6.56, the degrees of freedom= 378). The sample mean number of goals scored by the top 3 teams playing away was 1.95, and for the remaining teams playing away was 0.93. The top 3 teams scored significantly more goals on average when playing away.

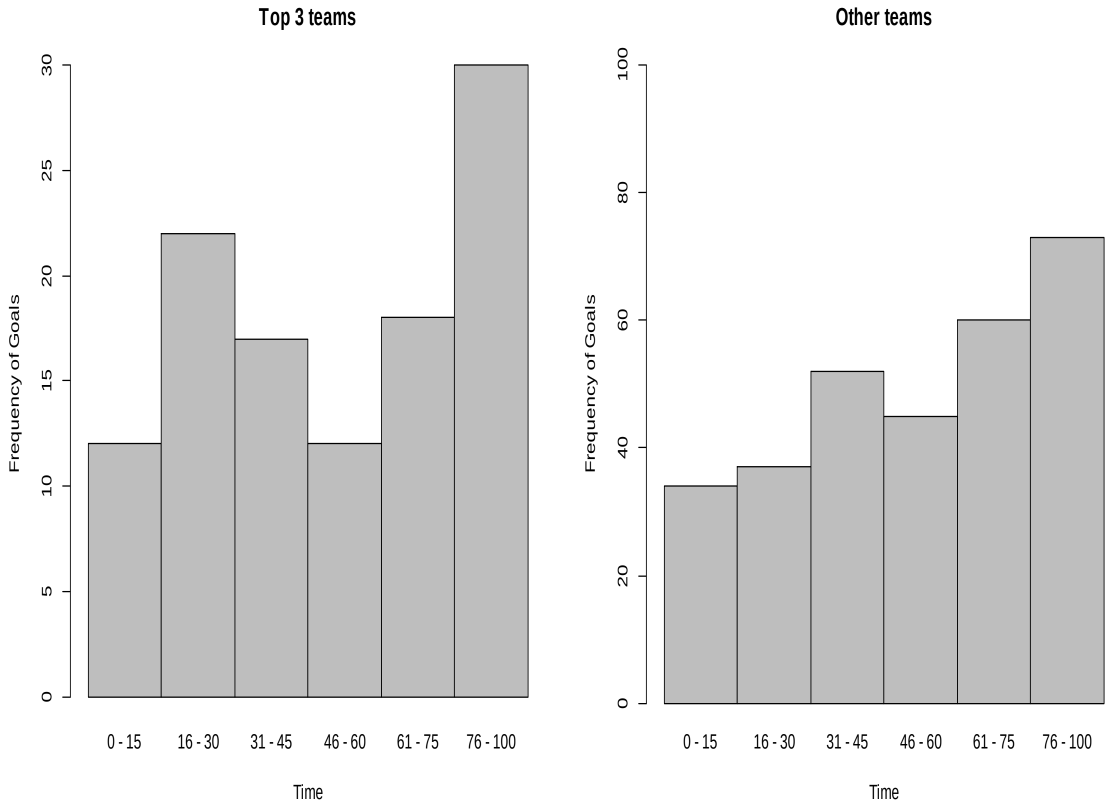

Figure 14. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in Spain

We conducted additional two-sample t-tests to test if the mean number of goals scored by teams playing at home is greater than the mean number of goals scored by teams playing away. This test was conducted for the top 3 teams and then for the remaining teams. The p-value for the test when testing for the top 3 teams was less than 0.001. The sample mean number of goals scored at home per game for the top 3 teams was 3.19, and the sample mean number of goals scored away per game for the top 3 games was 1.95 (T= 3.92; degrees of freedom= 112). The top 3 teams in Spain score significantly more goals on the average at home games than away games. The p-value for the test when testing for the remaining teams was 0.0002. The sample mean number of goals scored at home per game for the remaining teams was 1.41, and the sample mean number of goals scored away per game for the remaining teams was 1.09 (T= 3.58; degrees of freedom= 644).

We next test the hypotheses concerning the distributions of the goals made for the top 3 teams versus the remaining teams in Italy.

H_0: The proportions of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in Italy

H_a: The proportions are not the same

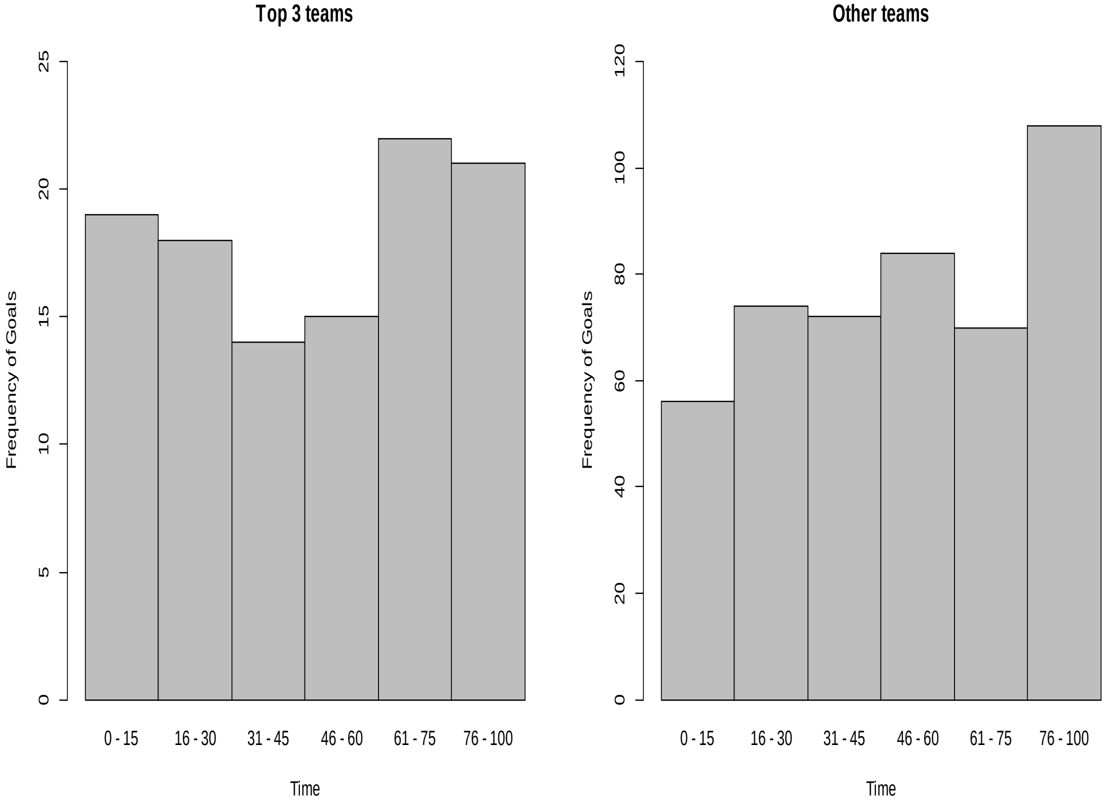

Figure 15. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in Italy

The histograms in Figure 15 are unusual compared to those that we have seen before. The histogram for the top 3 teams playing at home in Italy is appears to be slightly u-shaped. The histogram of goals for the remaining teams appears to be somewhat uniform except in the last time period. The p-value is 0.369 for the chi-square test testing the differences in distributions and we again fail to reject the null hypotheses saying that there is no difference in distribution of goals for the top 3 team and the remaining teams playing at home in Italy.

Table 21. Number of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*19*	*18*	*14*	*15*	*22*	*21*
Other teams	*56*	*74*	*72*	*84*	*70*	*108*

Now we are testing whether there is a difference between the mean number of goals scored by the top 3 teams playing at home and the mean number of cards for the remaining teams playing at home in Italy. The p-value is 0.01 and we reject the null hypothesis saying that there is difference between mean number of scoring goals of the top 3 teams and remaining teams playing at home in Italy (T= 2.58, the degrees of freedom= 378). The sample average number of goals scored per game for the top 3 teams when playing away is 1.91, and for the remaining teams playing away is 1.44. The top 3 teams score significantly more goals on average while playing at home than the remaining teams.

The hypotheses for the teams playing away in Italy are the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing away and the remaining teams playing away are the same in Italy

H_a: The proportions are not the same

We see from the Figure 16 that the histogram of goals made by the top 3 teams playing away also has a slight u-pattern similar to when they were playing at home. The histogram for the distribution of goals scored in each time period for the remaining teams in Italy is somewhat uniform except for the last time period. The p-value for the chi-square test is 0.217 and we fail to reject H_0.

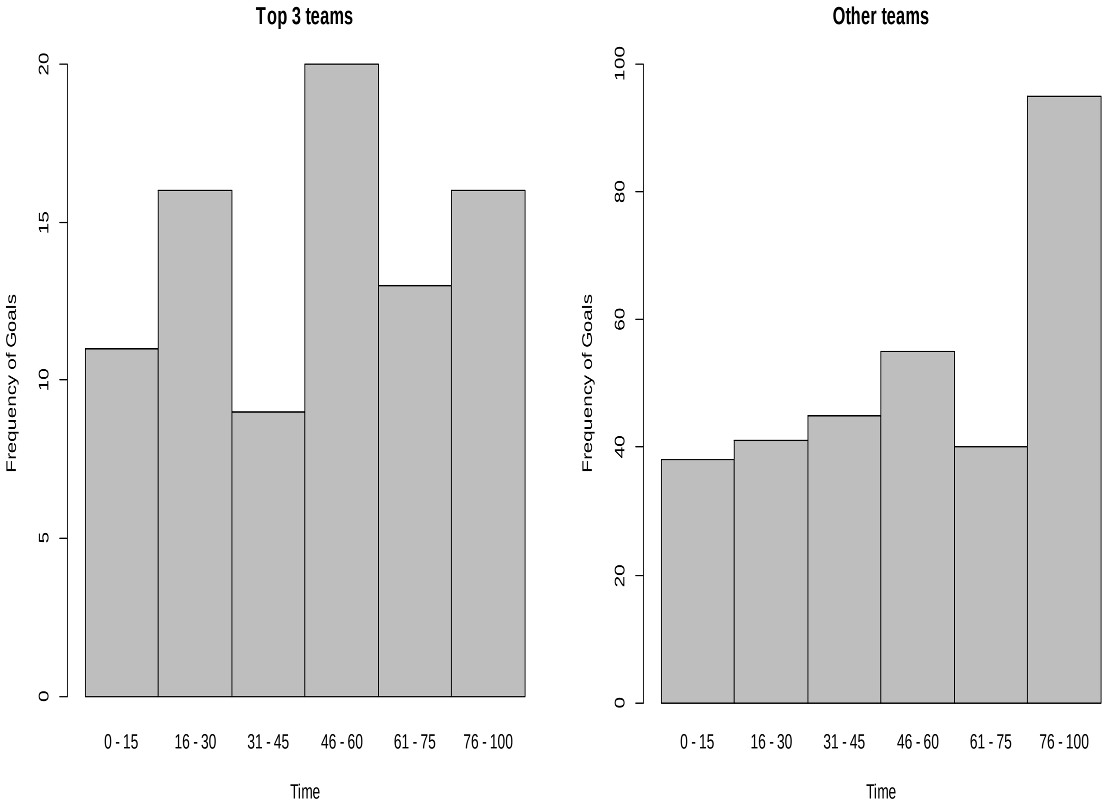

Figure 16. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in Italy

Now we are testing whether there is a difference between the mean number of goals scored by the top 3 teams playing away and the mean number of cards for the remaining teams playing away in Italy. The p-value is 0.0005 and we reject the null hypothesis saying that there is difference between mean number of scoring goals of the top 3 teams and remaining teams playing away in Italy (T=3.5, the degrees of freedom=378). The sample average number of goals scored per game by the top 3 teams while playing away is 1.49, and for the remaining teams is 0.97. The top 3 teams score significantly more goals per game when playing away than the remaining teams.

Table 22. Number of goals made in each time period by the top 3 teams playing away and the remaining teams playing away in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*11*	*16*	*9*	*20*	*13*	*16*
Other teams	*38*	*41*	*45*	*55*	*40*	*95*

We conducted additional two-sample t-tests to test if the mean number of goals scored by teams playing at home is greater than the mean number of goals scored by teams playing away. This test was conducted for the top 3 teams and then for the remaining teams. The p-value for the test when testing for the top 3 teams was 0.03. The sample mean number of goals scored at home per game for the top 3 teams was 1.91, and the sample mean number of goals scored away per game for the top 3 games was 1.49 (T= 1.9; degrees of freedom= 112). The top 3 teams in Italy score significantly more goals on the average at home games than away games. The p-value for the test when testing for the remaining teams was less than 0.001. The sample mean number of goals scored at home per game for the remaining teams was 1.44, and the sample mean number of goals scored away per game for the remaining teams was 0.97 (T= 5.08; degrees of freedom= 644).

4.3.5. Are the distributions of cards for the top 3 teams in each country compared with all the remaining teams in each country, collectively, the same? One test will be conducted combining all the countries.

We perform the tests for the distributions of cards received by the top 3 teams versus the remaining teams for the combined countries of England, Spain and Italy.

H_0: The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in all three countries

H_a: The proportions are not the same

Table 23. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in all three countries

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*5*	*12*	*14*	*14*	*16*	*15*
Other teams	*25*	*53*	*77*	*66*	*102*	*121*

From Figure 17, we can see that the top teams and the remaining teams both become more aggressive towards the end of a game. The p-value is 0.707 for the chi-square test. We fail to reject H_0 that the distributions of proportions of cards given in the time periods to the top 3 teams versus the remaining teams playing at home are the same for all three countries.

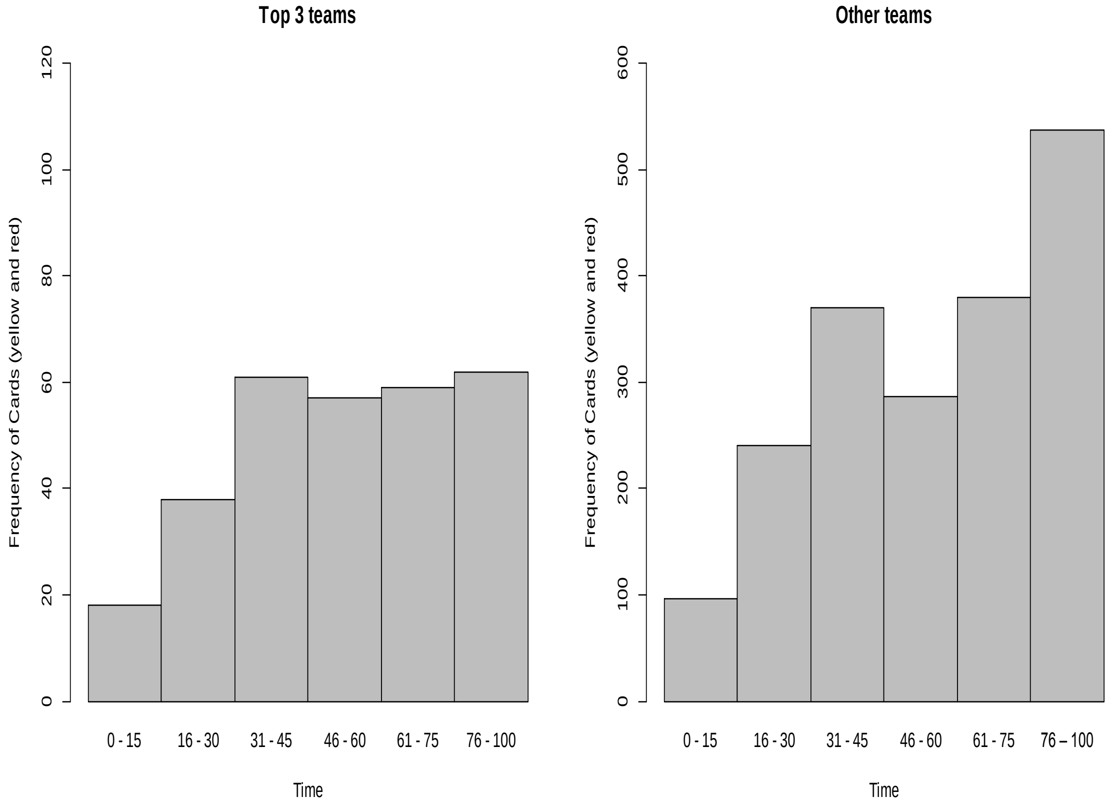

Figure 17. Number of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home in all three countries

We conduct a two-sample t-test to test for a difference between the mean number of cards received by the top 3 teams versus the mean number of cards received by the remaining teams when playing at home in all three countries. The p-value is 0.04 and we have enough evidence to say that there is the difference in mean number of cards received by the top 3 teams and the remaining teams playing at home in all the European countries that we considered (T= -2.05, the degrees of freedom=1138). The sample mean number of cards received by the top 3 teams when playing at home was 1.73. The sample mean number of cards received by the remaining teams when playing at home was 1.97.

The hypotheses for the second test referring to the top 3 teams and the remaining teams playing away are shown below.

H_0: The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in all three countries

H_a: The proportions are not the same

Examining Figure 18 we can see that the histograms look almost identical and the only difference is the 51-75 minutes time period of the first histogram. The p-value of the chi-square test is 0.986 and hence, we fail to reject the null hypotheses that the distributions of proportions of cards in each time period are the same for the top 3 teams and the remaining teams playing away for all the countries combined.

We also conducted a two-sample t-test to test for a difference between the mean number of cards received by the top 3 teams versus the mean number of cards received by the remaining teams when playing away in all three countries. The p-value is 0.167 and we fail to reject the null hypothesis saying that there is no difference in mean number of receiving cards for the top three teams and the remaining teams playing away in all three countries (T= 1.38, the degrees of freedom= 1138). The sample mean number of cards received by the top 3 teams when playing away was 2.37. The sample mean number of cards received by the remaining teams when playing away was 2.19.

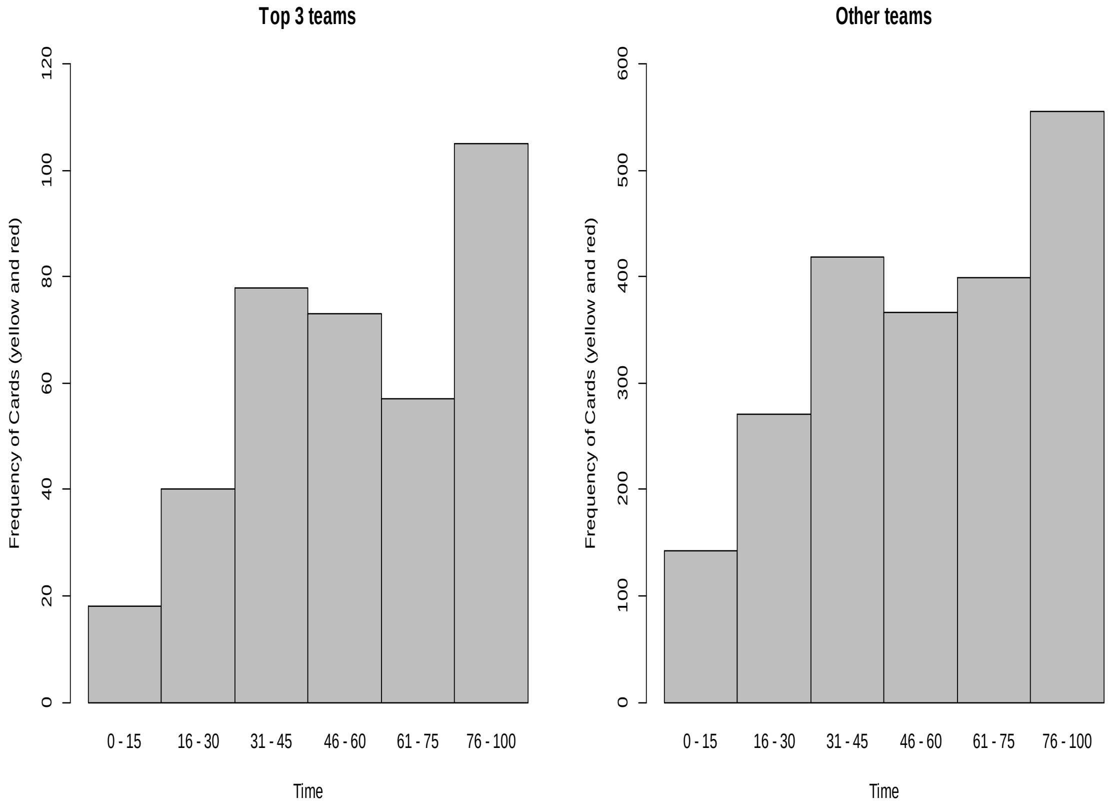

Figure 18. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in all three countries

We conducted additional two-sample t-tests comparing the mean number of cards received by the top 3 teams in each country playing at home and then playing away, and comparing the mean number of cards received by the remaining teams playing at home and then playing away. The p-value for the t-test based on the top 3 teams was less than 0.001 (T= -3.98, degrees of freedom= 340). The top 3 teams in each country get significantly fewer cards on average when playing at home versus playing away. The sample mean number of cards for the top 3 teams playing at home was found to be 1.73, and the sample mean number of cards for the top 3 teams playing away was found to be 2.37. The p-value for the t-test based on the remaining teams was less than 0.0009 (T= -3.13, degrees of freedom= 1936). The sample mean number of cards for the remaining teams playing at home was found to be 1.97. The sample mean number of cards for the remaining teams playing away was found to be 2.19. In both cases, teams scored significantly fewer cards when playing at home.

Table 24. Number of cards given in each time period to the top 3 teams playing away and the remaining teams playing away in all three countries

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*7*	*10*	*18*	*19*	*18*	*23*
Other teams	*36*	*73*	*114*	*105*	*108*	*136*

4.3.6. Are the distributions of goals for the top 3 teams in each country compared with all the remaining teams in each country, collectively, the same? One test will be conducted combining all the countries.

In order to answer this question we conducted chi-square tests for the distributions of goals for the top 3 teams versus the remaining teams for all countries together. The hypotheses for the first test are the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in all three countries

H_a: The proportions are not the same

The histograms in Figure 19 both have slightly increasing patterns. All the teams playing at home score a few more goals towards the end of the game. The p-value of the chi-square test is 0.45 and we fail to reject the null hypotheses and conclude that there is no difference in the distributions of proportions of goals scored in each time period for the top 3 teams and the remaining teams playing at home for all countries combined.

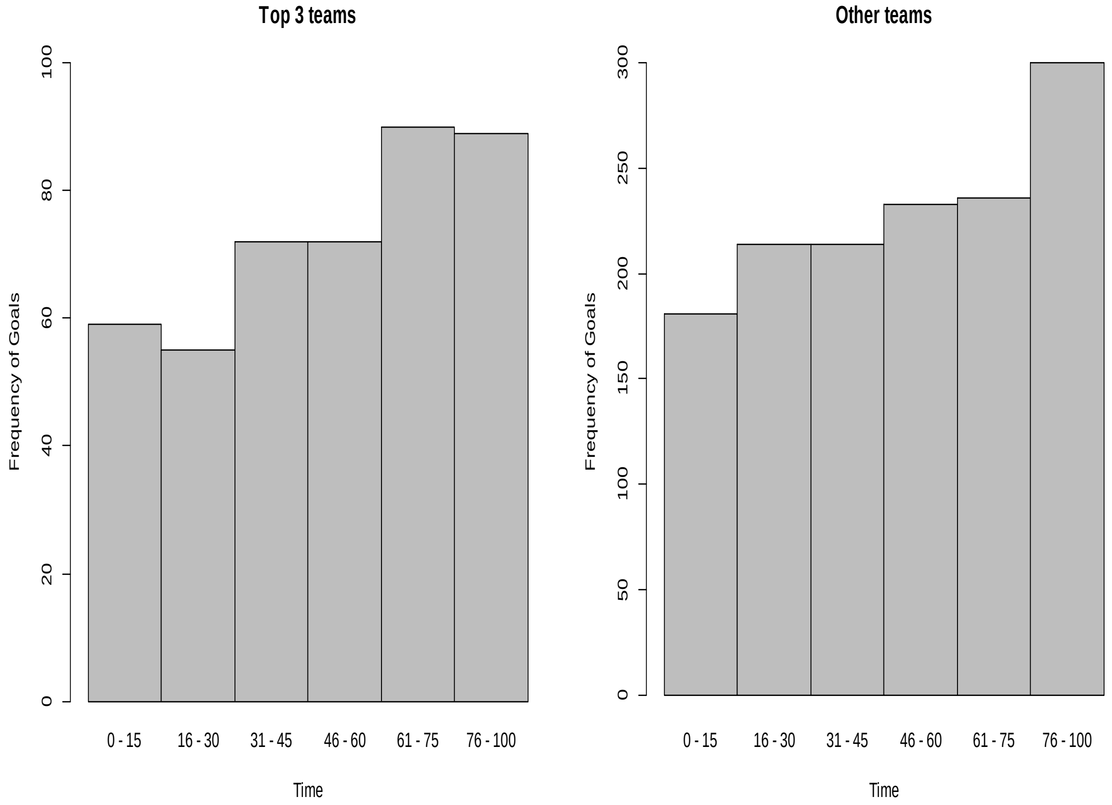

Figure 19. Number of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home in all three countries

Table 25. Number of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home in all three countries

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*59*	*55*	*72*	*72*	*90*	*89*
Other teams	*181*	*214*	*214*	*233*	*236*	*300*

We perform a two-sample t-test to test for a difference between the mean number of goals scored by the top 3 teams versus the mean number of goals scored by the remaining teams when playing at home in the three countries. The p-value is less than 0.001 and we have enough evidence to say that there is a difference in mean number of scoring goals for the top 3 teams and the remaining teams playing at home in all three countries (T = 10.5, the degrees of freedom= 1138). The sample average of goals per game by the top 3 teams playing at home was 2.56 and the sample average of goals per game by the remaining teams playing at home was 1.42.

The set of hypotheses for the second test are the following:

H_0: The proportions of goals made in each time period by the top 3 teams playing away and the remaining teams playing away are the same in all three countries

H_a: The proportions are not the same

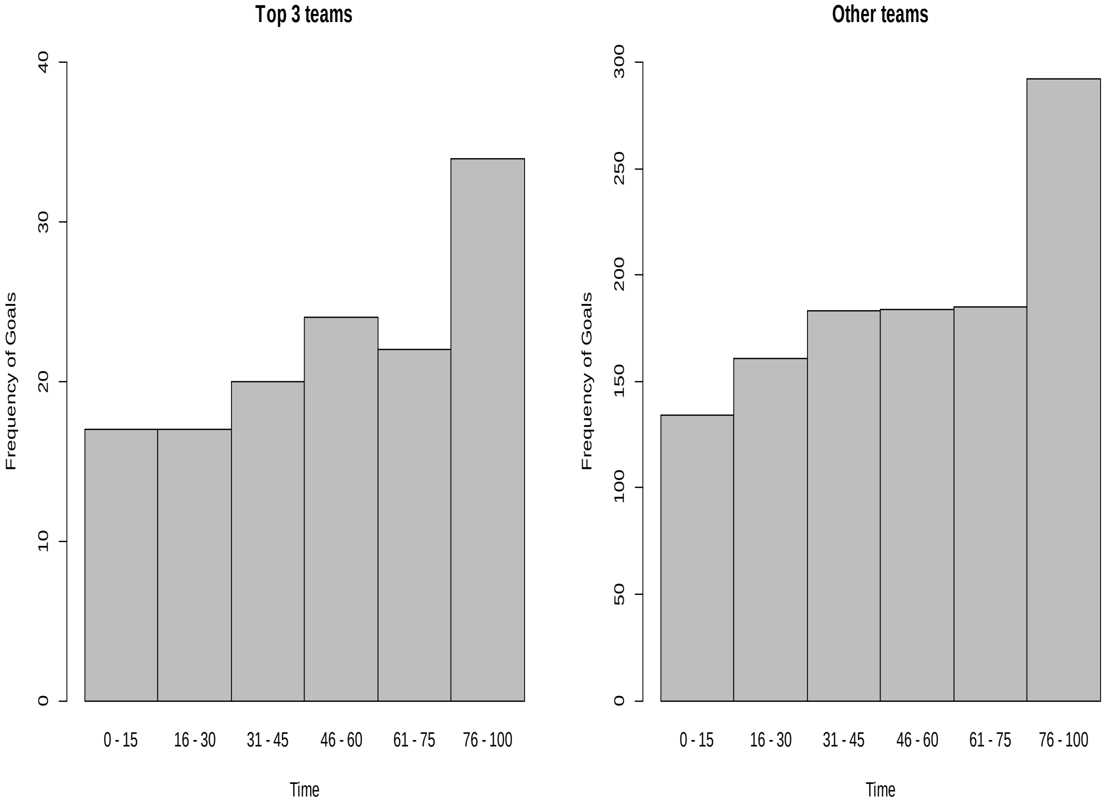

Figure 20. Number of goals scored in each time period by the top 3 teams playing away and the remaining teams playing away in all three countries

From Figure 20 we see that the histograms indicate fewer goals than the histograms in Figure 19. We also see that there is a higher proportion of goals scored by the away teams in the last time period than compared to other time periods. The reason for this could be explained that the game is heated up at the end and many players are tired or have suffered injuries. The p-value for the chi-square test is 0.988. We fail to reject the null hypotheses at the significance level of 0.05.

We perform a two-sample t-test to test for a difference between the mean number of goals scored by the top 3 teams versus the mean number of goals scored by the remaining teams when playing away in the three countries. The p-value is less than 0.001 and we reject the null hypothesis and state that there is a difference in mean number of scoring goals for the top 3 teams and the remaining teams playing away in all three countries (T= -4.2, the degrees of freedom= 1138). The sample average number of goals per game for the top 3 teams playing away was 0.78, and for the remaining teams was 1.18.

Table 26. Number of goals scored in each time period by the top 3 teams playing away and the remaining teams playing away in all three countries

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Top 3 teams	*37*	*59*	*42*	*49*	*48*	*71*
Other teams	*114*	*119*	*161*	*159*	*159*	*255*

Two-sample t-tests were conducted to compare the mean number of goals the top 3 teams scored at home versus away, and then to compare the mean number of goals the remaining teams scored at home versus away. The p-values in both cases are less than 0.001. The sample mean number of goals the top 3 teams scored at home was 2.56, and the sample mean number of goals they scored away was 0.78 (the test statistic value=12.49, degrees of freedom=340). The sample mean number of goals the remaining teams scored at home was 1.42 , and the sample mean number of goals they scored away was 1.18 (T=4.53, degrees of freedom=1936). In both cases, teams scored significantly more goals at home on the average.

4.3.7. Are the distributions of cards the same for the home teams versus the away teams in England (Spain, Italy)? Tests will be conducted separately for each country.

We conduct the chi-square test to see if the distributions of proportions cards are the same for the home teams and the away teams in England. The hypotheses for this test are the following:

H_0: The proportions of cards given to home teams and away teams in each time period are the same in England

H_a: The proportions are not the same

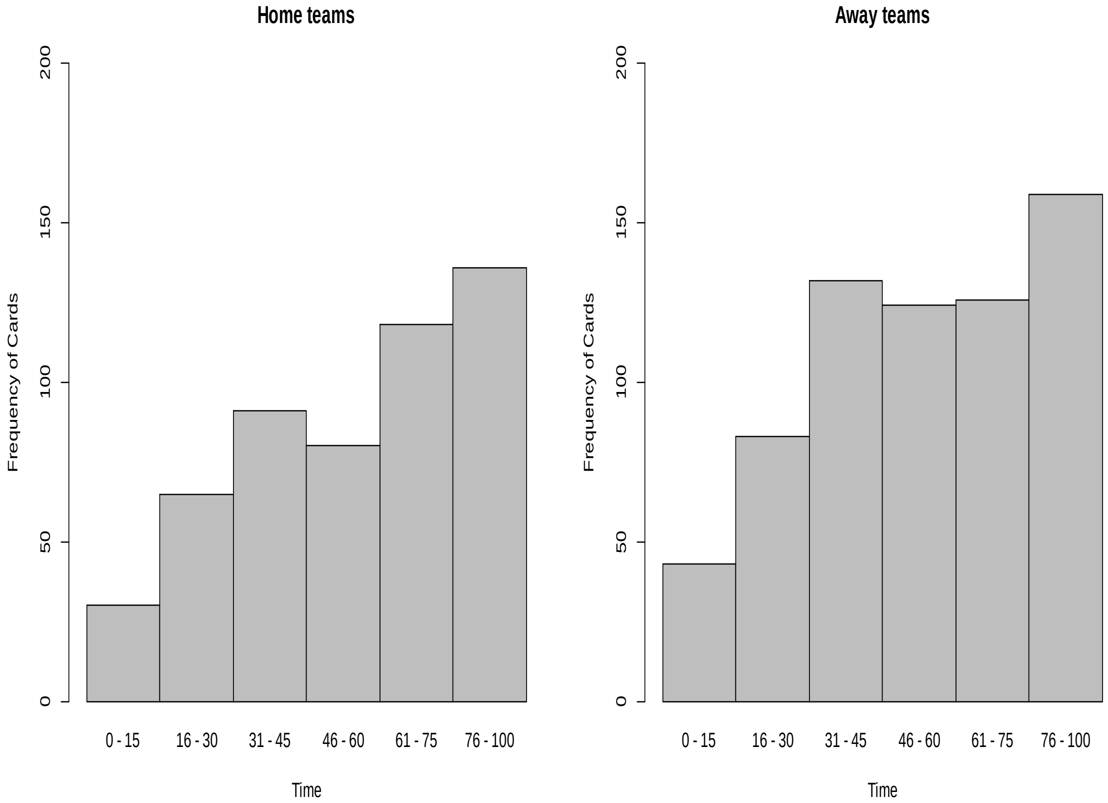

Figure 21. Number of cards given in each time period to home teams and away teams in England

It appears from Table 27 that there is a smaller amount of cards for teams playing at home than for teams playing away in England. We also see that the histograms are bimodal. Teams appear to play more rough towards the end of the first half and then again towards the end of the second half. The p-value associated with the chi-square test is 0.362 and hence, there is not enough evidence to reject the null hypothesis.

Table 27. Number of cards given in each time period to home teams and away teams in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*30*	*65*	*91*	*80*	*118*	*136*
Away teams	*43*	*83*	*132*	*124*	*126*	*159*

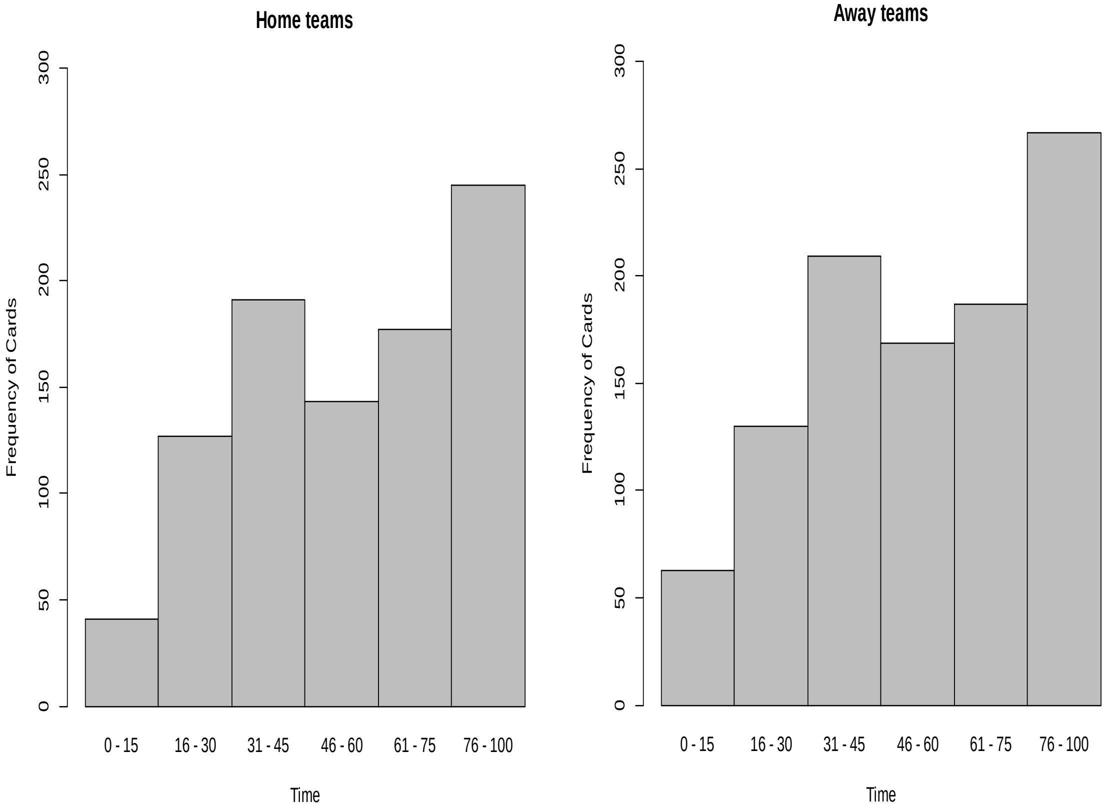

Figure 22. Number of cards given in each time period to home teams and away teams in Spain

Now we consider the same test for the Spanish teams. The hypotheses are presented below:

H_0: The proportions of cards given to home teams and away teams in each time period are the same in Spain

H_a: The proportions are not the same

Table 28. Number of cards given in each time period to home teams and away teams in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*41*	*127*	*191*	*143*	*177*	*245*
Away teams	*63*	*130*	*209*	*169*	*187*	*267*

The information from Table 28 tells us that the numbers of cards are very close for teams playing either at home or away in Spain. The tension of the game falls into the last minutes of each half of the game.

The p-value for the chi-square test is 0.599 and hence, we do not reject the null hypothesis.

We consider the same test for the Italian teams. The hypotheses for this test are:

H_0: The proportions of cards given to home teams and away teams in each time period are the same in Italy

H_a: The proportions are not the same

Table 29. Number of cards given in each time period to home teams and away teams in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*43*	*86*	*149*	*121*	*144*	*218*
Away teams	*54*	*98*	*156*	*147*	*143*	*234*

The numbers in Table 29 are very close to each other. The greatest number of cards is given in the last 15 minutes of the game. This is a very similar situation to Spain. The p-value of the chi-square test is 0.832 and we fail to reject the null hypothesis.

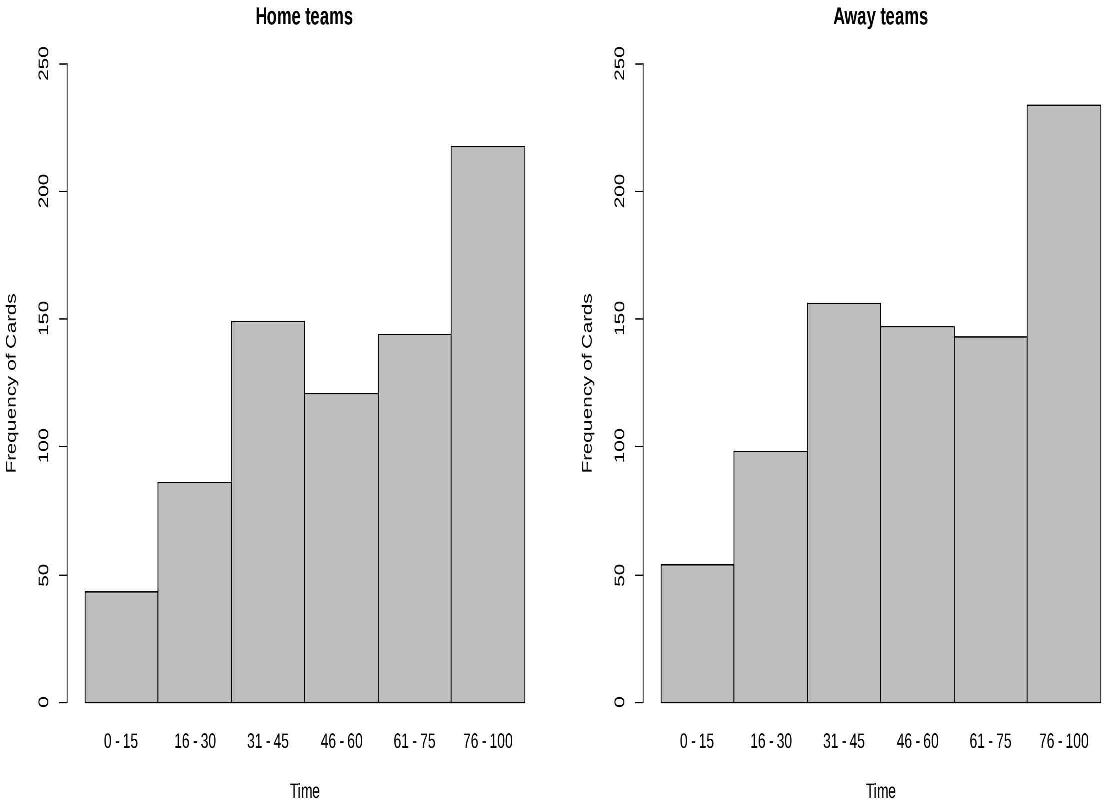

Figure 23. Number of cards given in each time period to home teams and away teams in Italy

4.3.8. Are the distributions of goals the same for the home teams versus the away teams in England (Spain, Italy)? Tests will be conducted separately for each country.

We consider the chi-square test for the difference in the distributions of the proportion of goals scored for all 3 countries separately. First, we conduct the test for England teams and the hypotheses are shown below.

H_0: The proportions of goals made in each time period by home teams and away teams are the same in England

H_a: The proportions are not the same

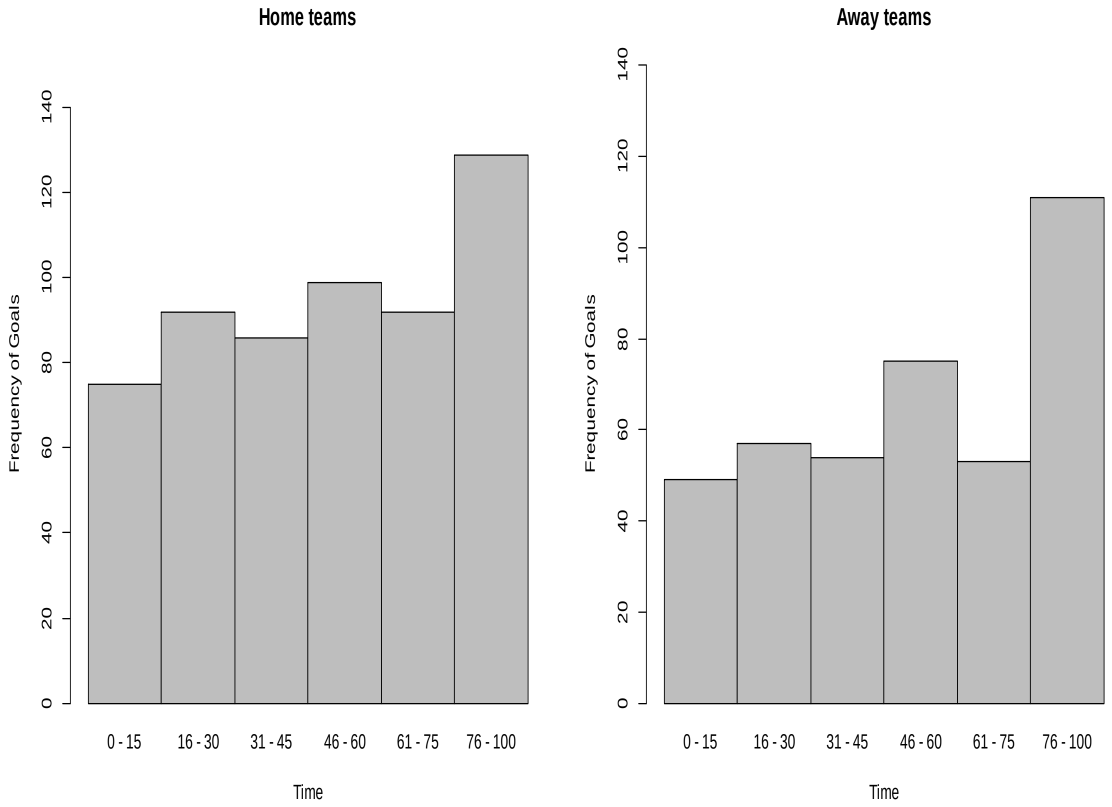

Figure 24. Number of goals made in each time period by home teams and away teams in England

Table 30. Number of goals made in each time period by home teams and away teams in England

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*80*	*87*	*100*	*97*	*116*	*124*
Away teams	*56*	*62*	*80*	*76*	*76*	*112*

In examining Figure 24, one can see that the home teams start scoring goals faster than the away teams in England. Both histograms have the same increasing pattern. The p-value associated with the chi-square test is 0.66 and we do not have enough evidence to reject the null hypothesis.

The next test is performed for the Spanish teams playing at home and away. The hypotheses are the following:

H_0: The proportions of goals made in each time period by home teams and away teams are the same in Spain

H_a: The proportions are not the same

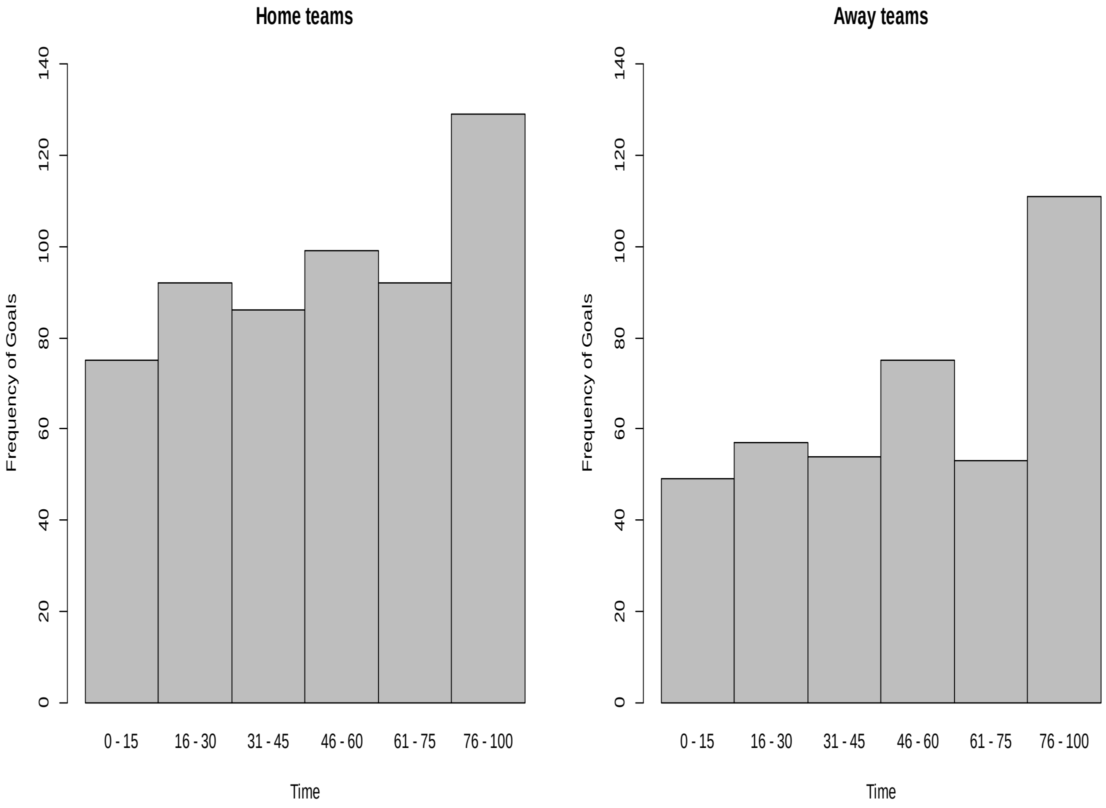

Figure 25. Number of goals made in each time period by home teams and away teams in Spain

It appears that the number of goals for home teams is greater than for the away teams from Table 31. In examining Figure 25 the number of goals scored by home teams increases in each time period in Spain. Home teams appear to start the second half of the game much stronger than they finished the end of the first half. The p-value for the chi-square test is 0.505 and we do not have enough evidence to reject H_0.

Table 31. Number of goals made in each time period by home teams and away teams in Spain

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*85*	*90*	*100*	*109*	*118*	*136*
Away teams	*46*	*59*	*69*	*57*	*78*	*103*

Finally, the same test is conducted for the Italian teams. The hypotheses are the following:

H_0: The proportions of goals made in each time period by home teams and away teams are the same in Italy

H_a: The proportions are not the same

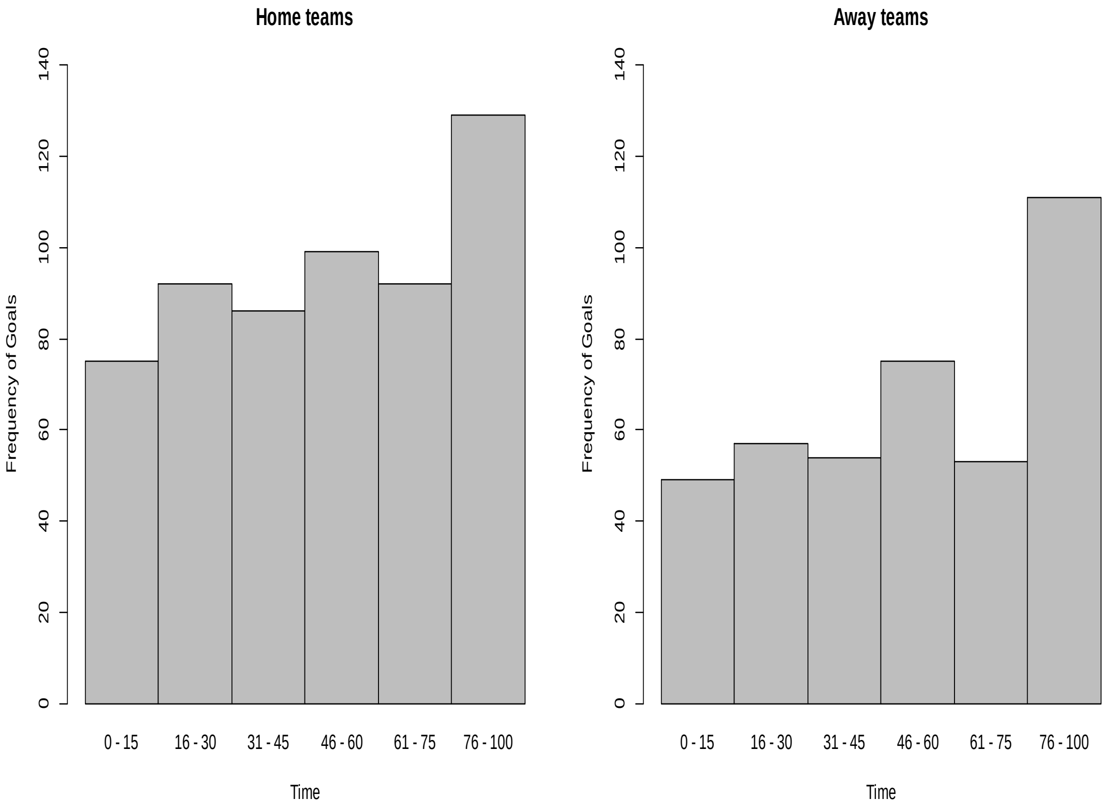

Figure 26. Number of goals made in each time period by home teams and away teams in Italy

Table 32. Number of goals made in each time period by home teams and away teams in Italy

	0 – 15 min	*16 – 30 min*	*31 – 45 min*	*46 – 60 min*	*61 – 75 min*	*> 76 min*
Home teams	*75*	*92*	*86*	*99*	*92*	*129*
Away teams	*49*	*57*	*54*	*75*	*53*	*111*

Analyzing the histograms in Figure 26, both have similar patterns, but differ in the number of goals scored. The home teams appear to score more goals than the away teams. The p-value for the chi-square test is 0.397 which is high again and we fail to reject the null hypothesis that states that the distributions of proportions of goals scored over the time periods are the same for home teams and away teams.

4.3.9. Do teams in the three countries get cards equally often?

We conduct the goodness-of-fit test to check if the proportions of cards given out are the same for all three countries where π_i represents the proportion of cards for country i. There are 2 tests, one for the home teams and another for the away teams. The hypotheses for both tests can be expressed as:

$H_0: \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

H_a: H_0 is not true,

where π_1, π_2, $\pi_3$represent the proportion of cards received by each of the three countries of England, Spain and Italy.

The null hypothesis saying that the proportions of cards of the home teams in all three countries are the same can be rejected since the p-value is less than 0.001. The proportion of cards given in soccer is not the same for all the countries. In our sample, the Spanish teams playing at home got 924 cards and Italian teams got 761 while the English teams playing at home got only 520 cards. The null hypothesis, the proportion of cards given to teams playing away was the same for each country, was also rejected at a p-value of 0.001. In our sample, the Spanish teams playing away got 1025 cards and Italian teams received 832 while the English teams playing away received 667 cards. We see that teams in Spain play very aggressive compared to English and Italian teams.

4.3.10. Do teams in the three countries score goals equally often?

We conduct 2 tests, the first one is to check if the proportions of goals of the home teams are the same for all three countries, and the second is to check if the proportions of goals of the away teams are the same for all three countries. The hypotheses for these tests are the following:

$H_0: \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

H_a: H_0 is not true,

where π_1, π_2, $\pi_3$represent the proportion of total goals scored by the three countries of England, Spain and Italy.

In the first case when we test the equality of proportions of goals scored by teams playing at home in England, Spain, and Italy, we cannot reject the null hypothesis since the p-value is equal to 0.174. This implies that there is no significant difference between the numbers of goals scored by the three countries when teams are playing at home. In testing the hypothesis that the proportion of goals scored by away teams is the same for all the three countries, there is some weak evidence that this is not the case with a p-value of 0.074. In our sample, the English teams scored 462 goals away, while Spanish and Italian teams scored just 412 and 399 goals away, respectively. This interesting fact is supported by a popular opinion that games in England are less predictable and an away team can do quite well.

4.3.11. Do the top three teams in the three countries get cards equally often?

We have already done something similar in previous questions, but in this case we will emphasize only on the top three teams for each country. We want to test if the proportions of cards received by the top three teams are the same for all three countries. The tests are done for the teams playing at home and away.

$H_0: \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

H_a: H_0 is not true,

where π_1, π_2, $\pi_3$represent the proportion of cards received by the top three teams in each of the three countries of England, Spain and Italy.

It is generally thought that referees in England give cards only for serious violations. In testing the hypothesis, the p-value is 0.02, and hence the null hypothesis is rejected. The observed number of cards for the English teams playing at home is 76. The observed value for Italian teams is 106 and for the Spanish teams is 113 which are very close to each other. The results from this test confirm what is generally thought.

The null hypothesis is also rejected when testing that the proportions of cards given to away teams in each of the countries is the same. The numbers of cards are 113, 171 and 121 for teams playing away in England, Spain and Italy, respectfully. The p-value for this test is 0.001. In this case, the highest observed value of cards is from Spain.

4.3.12. Do the top three teams in the three countries score goals equally often?

We perform tests to see if the proportions of goals scored by the top three teams are the same in England, Spain and Italy. The hypotheses are the following:

$H_0: \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

H_a: H_0 is not true,

where π_1, π_2,$\pi_3$represent the proportion of goals scored by the top three teams in each of the three countries of England, Spain and Italy.

The result for the test, where the null hypothesis says that the proportions of goals for the top 3 teams playing at home for all three countries are the same, is significant. The p-value is 0.0001 and we reject the null hypothesis. The observed values for the number of goals scored by the top 3 teams playing at home for the sample are 146, 182, and 109 in England, Spain and Italy, respectfully. The observed values for the number of goals scored playing away are 48, 50 and 36 for England, Spain and Italy, respectfully. We fail to reject the null hypothesis that the proportions of goals scored by the top 3 teams playing away for all three countries are the same since the p-value is equal to 0.277.

4.3.13. Do English teams receive fewer (yellow & red) cards on average than Italian and Spanish teams?

While conducting this research we noticed that Spanish and Italian teams seemed to score approximately the same number of goals. Moreover, they received almost the same number of cards. At the same time, English teams seemed to receive fewer cards than Spanish and Italian teams. As a matter of fact, it agrees well with the common opinion that English referees give as when a player is heavily injured by another player. Therefore, we decided to check whether it is true that English teams receive fewer cards on average than the other teams. To answer this question, we conducted the two sample t-test. The first sample combines the number of cards received during the games in Italy and Spain during the 2011-2012 year. The second sample represents the number of cards in the English championship for the 2011- 2012 year. The sample mean of the number of cards per game in Spain and Italy was 4.66 and the sample mean of the number of cards per game for England was 3.12. The p-value for this test is highly significant

(<0.001). In other words, we have enough evidence to reject the null hypothesis and conclude that the number of cards given in English championship games is less than those in other countries.

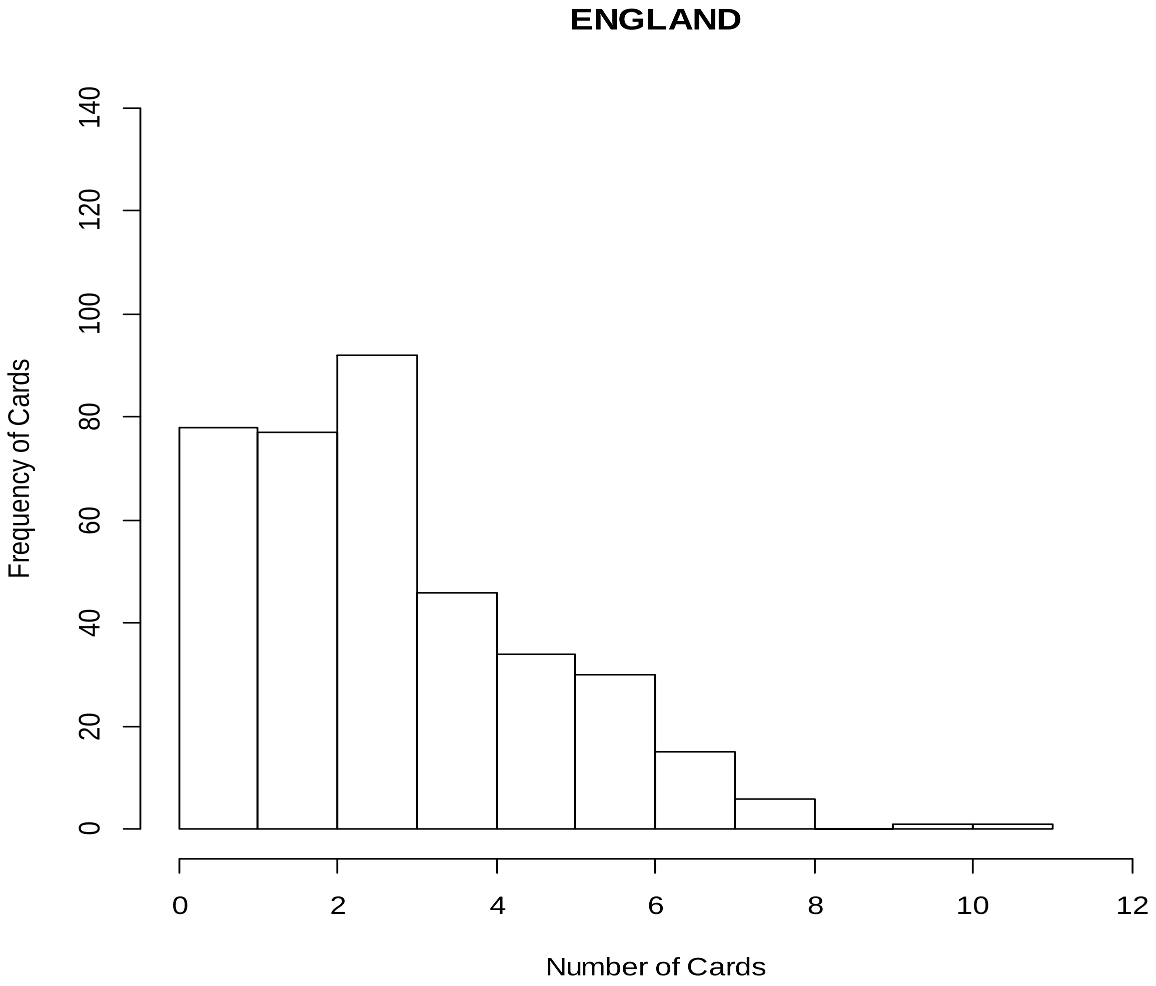

Figure 27. Number of cards given to the English teams

Histograms in Figures 27 and 28 follow a unimodal pattern. Both patterns are similar but the frequencies of getting cards are quite different. It agrees with the result of our test that English teams receive fewer cards than teams from other countries.

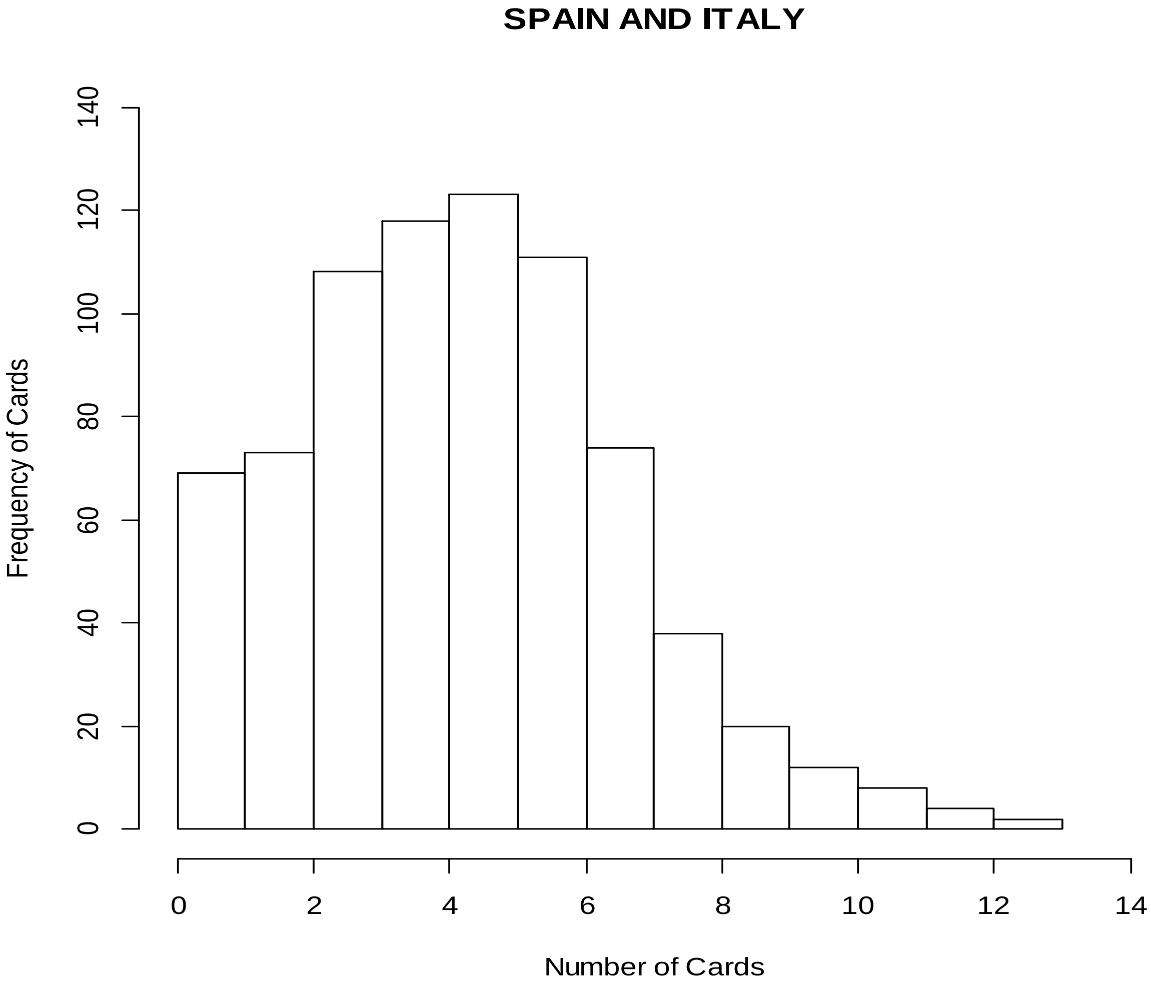

Figure 28. Number of cards given to the Italian and Spanish teams

4.3.14. Does the mean number of (yellow & red) cards differ for the home and away teams in each of the three countries (England, Spain, Italy)?

In this section we focus on paired samples since we are interested in the difference of the number of cards for the home teams and the away teams for each game in each of the three countries. Therefore, we conduct a one sample paired t-test on calculated differences. First of all, we consider a paired t-test for the teams in England. The p-value is highly significant (<0.001). Therefore, we have enough evidence to state that there is the difference in the number of cards between teams playing at home and away in England. We can note that teams playing away receive more cards than those playing at home.

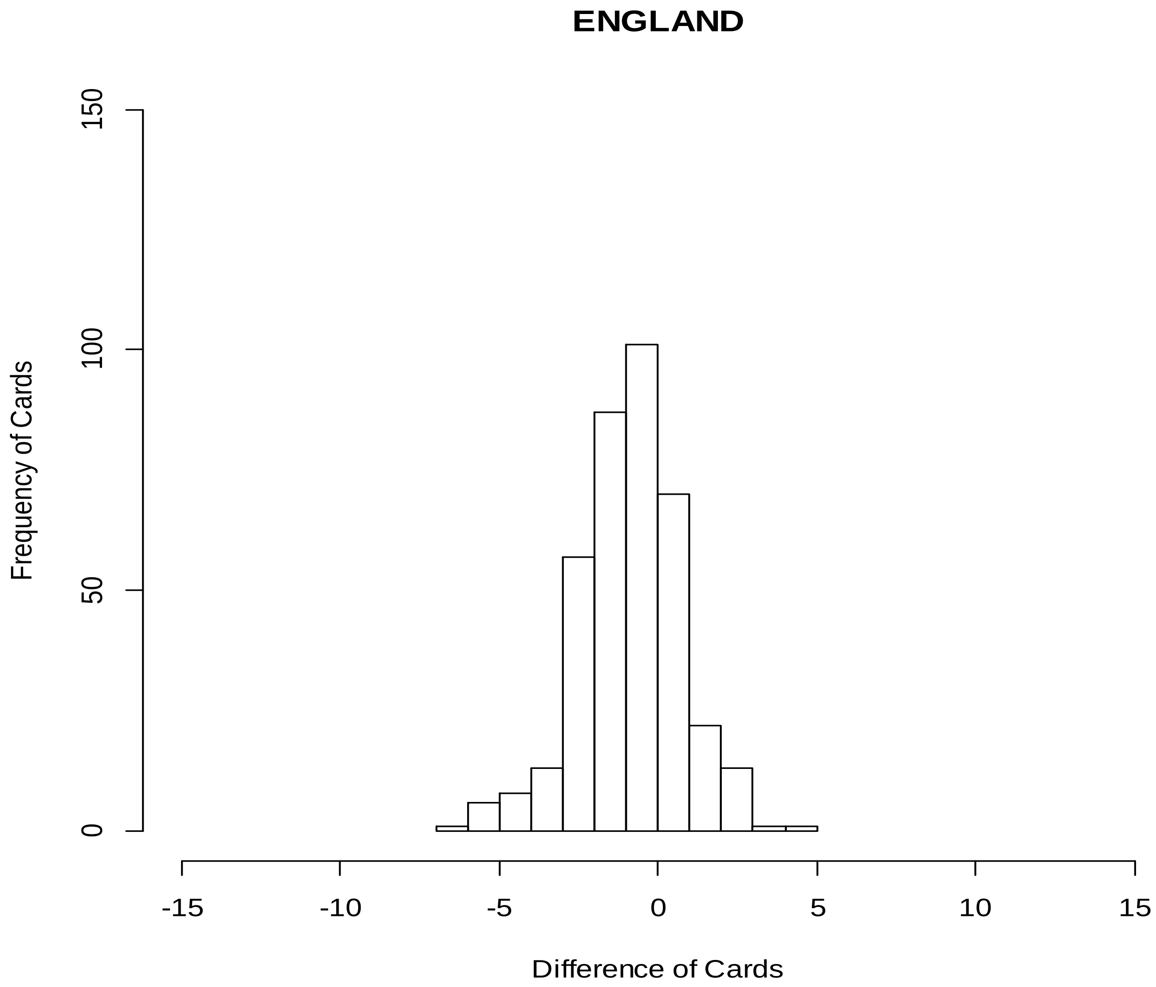

Figure 29. Difference of cards of home teams and away teams in England

From Figure 29, we can note that the histogram is roughly normally distributed with the sample mean being below 0. Again, since the mean is negative the teams playing away receive more cards than the teams playing at home on the average.

Next, we perform a paired t-test for Spanish teams. The mean difference is -0.266. This time, the corresponding p-value is 0.008. The p-value is not as small as for English teams but it is still highly significant. Therefore, we reject the null hypothesis that referees in Spain give cards equally often to home and away teams. Teams playing away get more cards.

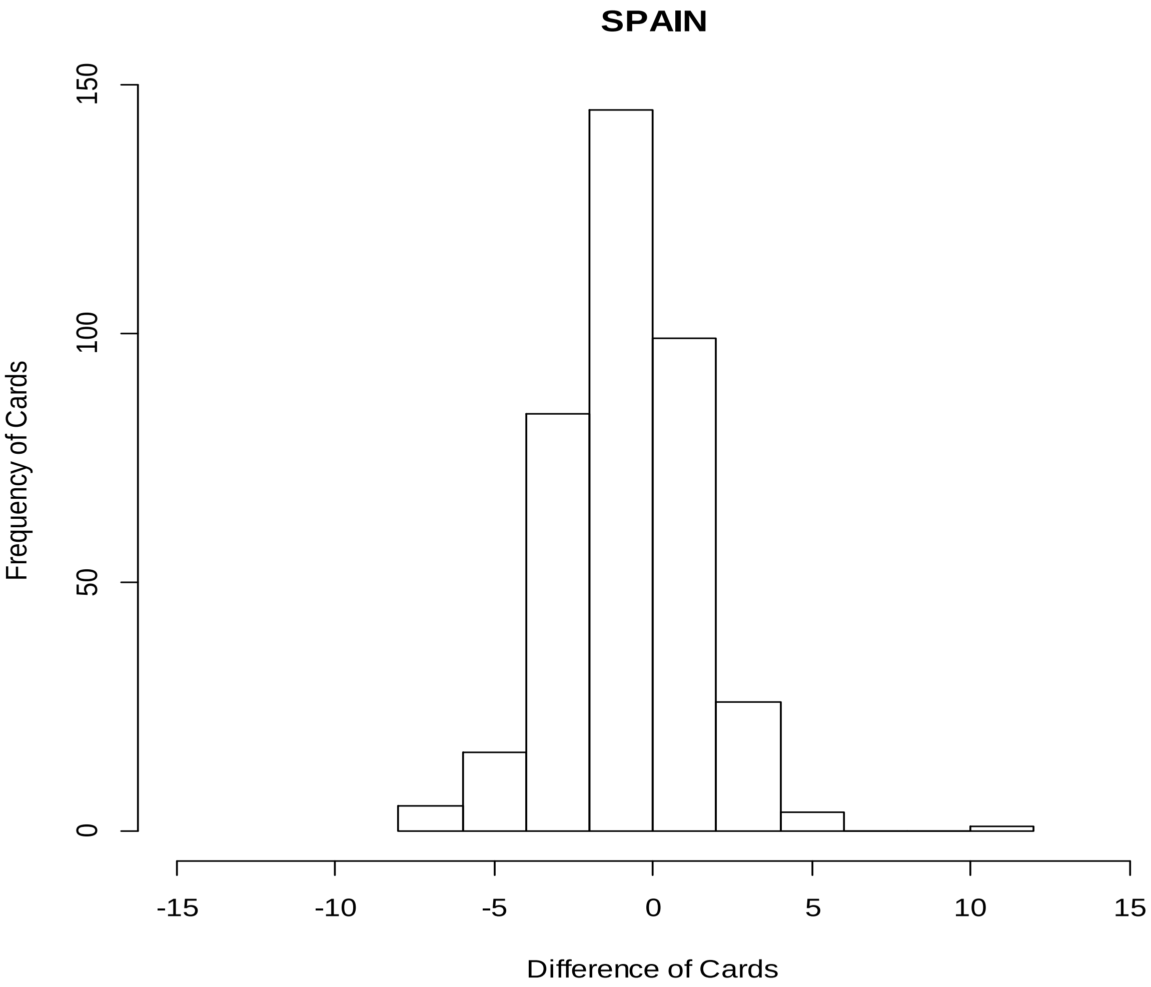

Figure 30. Difference of cards of home teams and away teams in Spain

Finally, we conduct the paired t-test for Italian teams. The sample mean of the difference in the number of cards is slightly closer to zero than in the previous tests. For Italy, the mean difference is equal to -0.187. The corresponding p-value is 0.017. Therefore, we reject the null hypothesis and conclude that the away teams get significantly more cards. We can also remark that the variation in all three distributions of differences is quite similar.

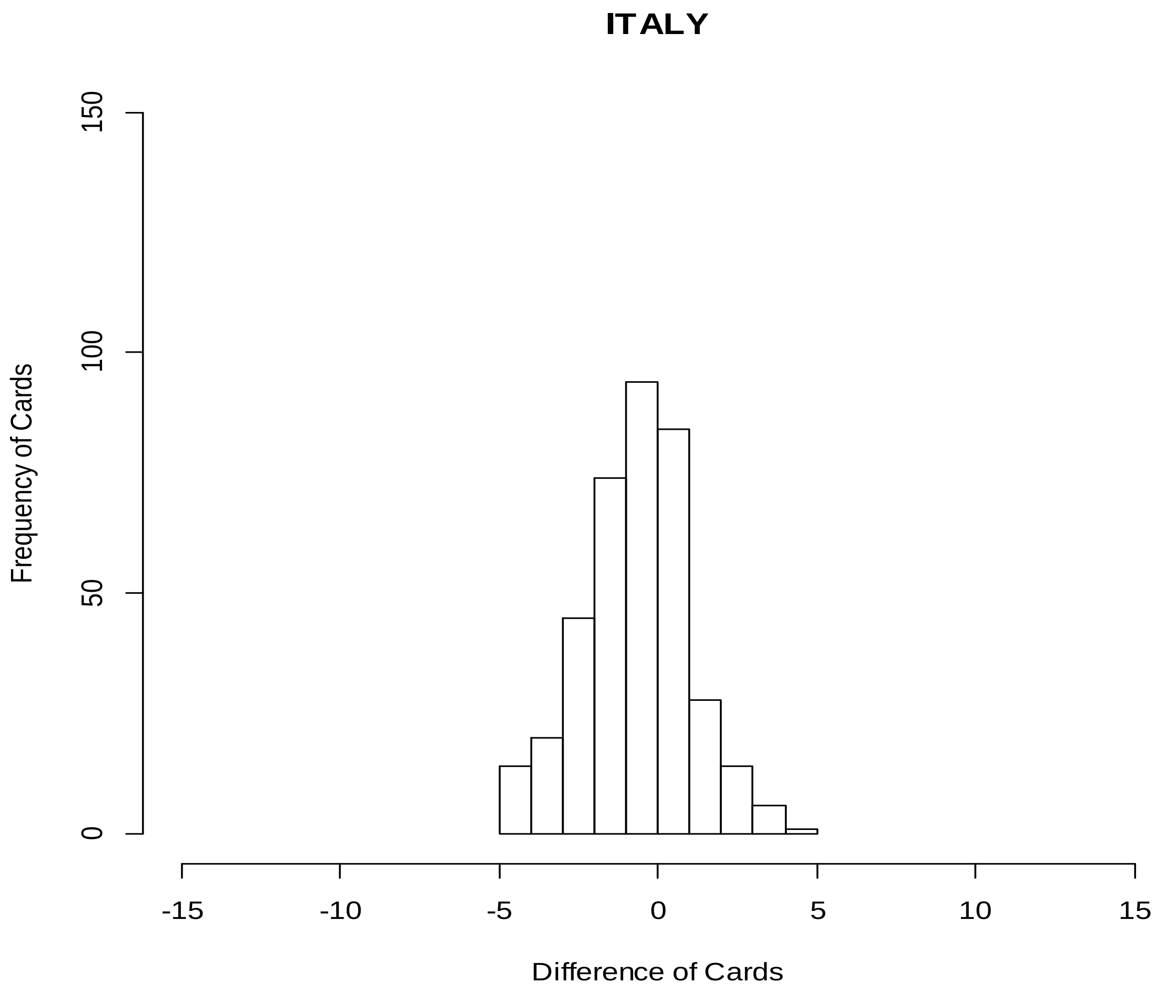

Figure 31. Difference of cards of home teams and away teams in Italy

CHAPTER 5. CONCLUSION

We derived least square regression models and logistic regression models to predict the last 5 rounds out of 38 rounds in the three European countries of England, Spain and Italy. First, we constructed five least square regression models and five logistic regression models based on all 6 variables mentioned in Chapter 2 with four variables based on the k previous rounds, with models based on k equal to 4, 6, 8, 10, and 12. We also developed new models with the indicator variables for the countries taken out for both the least squares models the logistic regression models since the indicator variables were not significant in most models. We selected the model based on the 8 previous rounds as the best model for the least square regression method and the model based on the 8 previous rounds for the logistic regression method to predict the last 5 rounds. The least squares model predicted the last five rounds with an accuracy of 76%, and the logistic regression model predicted the last 5 rounds with an accuracy of 77%.

Additional testing for the distributions of proportions of cards and goals throughout the game led us to conclude the following since the distributions were not significantly different:

- The proportions of cards given to the home teams in each time period are the same for all three countries
- The proportions of cards given to the away teams in each time period are the same for all three countries
- The proportions of goals scored by the home teams in each time period are the same for all three countries
- The proportions of goals scored by the away teams in each time period are the same for all three countries

- The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in England
- The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in Spain
- The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in Italy
- The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in England
- The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in Spain
- The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in Italy
- The proportions of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in England
- The proportions of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in Spain
- The proportions of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in Italy
- The proportions of goals scored in each time period by the top 3 teams playing away and the remaining teams playing away are the same in England
- The proportions of goals scored in each time period by the top 3 teams playing away and the remaining teams playing away are the same in Spain

- The proportions of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing away are the same in Italy
- The proportions of cards given in each time period to the top 3 teams playing at home and the remaining teams playing at home are the same in all three countries
- The proportions of cards given in each time period to the top 3 teams playing away and the remaining teams playing away are the same in all three countries
- The proportions of goals scored in each time period by the top 3 teams playing at home and the remaining teams playing at home are the same in all three countries
- The proportions of goals scored in each time period by the top 3 teams playing away and the remaining teams playing away are the same in all three countries
- The proportions of cards given to the home teams and the away teams in each time period are the same in England
- The proportions of cards given to the home teams and the away teams in each time period are the same in Spain
- The proportions of cards given to the home teams and the away teams in each time period are the same in Italy
- The proportions of goals scored by the home teams and the away teams in each time period are the same in England
- The proportions of goals scored by the home teams and the away teams in each time period are the same in Spain
- The proportions of goals scored by the home teams and the away teams in each time period are the same in Italy (which had largest sample proportion and which had smallest sample proportion)

- The proportions of total goals scored by teams playing at home for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)
- The proportions of total goals scored by teams playing away for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)
- The proportion of total goals scored by the top 3 teams playing away for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)

Goodness-of-fit tests showed some significantly different distributions/proportions. These are the following:

- The proportions of total cards received by teams playing at home for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)
- The proportions of total cards received by teams playing away for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)
- The proportion of total cards received by the top 3 teams playing at home for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)
- The proportion of total cards received by the top 3 teams playing away for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)

- The proportion of total goals scored by the top 3 teams playing at home for all three countries of England, Spain and Italy (which had largest sample proportion and which had smallest sample proportion)

We conclude that the teams playing away receive more cards than the teams playing at home in England, Spain and Italy. We confirmed that English teams receive fewer cards than Spanish or Italian teams. It does not mean that English teams play less aggressive than other countries. This could be due to the manner of giving cards by the referee in England. It is known that English referees give cards only for serious violations.

Conducting various hypotheses, we determined that the teams score more goals playing at home than the teams playing away in all European countries that we considered. This could be due to the environment and the support of a great number of fans of the game. We also conclude that the home teams receive fewer cards on average compared to the away teams.

REFERENCES

Albright, Christian (1993). A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association, 88,* 1175-1183.

Hart, A.; Hutton, J.; Sharot, T. (1975). A statistical Analysis of association football attendances. *Journal of the Royal Statistical Society Series C (Applied Statistics), 24, 17.*

Kellis, Eleftherios; Katis, Athanasios (2007). Biomechanical characteristics and determinants of instep soccer kick. *Journal of Sports Science and Medicine* 6, 154-165.

Ridder, G.; Cramer,J.S.; Hopstaken, P. (1994). Down to ten: estimating the effect of the red card in soccer. *Journal of American Statistical Association,* 89, 1124-1127.

Rusu, A.; Stoica, D.; Burns, E.; Hample, B.; McGarry, K.; Russell, R., (2010). Dynamic visualizations for soccer statistical analysis. Information Visualisation (IV), 2010 14th International Conference, 207-212.

Spencer, Matt; Lawrence, Steve; Rechichi, Claire; Bishop, David; Dawson, Brian; and Goodman, Carmel (2004). Time–motion analysis of elite field hockey, with special reference to repeated-sprint activity. *Journal of Sports Sciences* Volume 22, 843-850.

Tena, Juan de Dios; Forrest, David (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences, *European Journal of Operational Research*, 181, 362–373.

APPENDIX

Table A1. ANOVA table of the linear regression model based on the 4 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	0.49496	0.10013	4.943	9.24e-07 ***
X1	0.08777	0.01457	6.026	2.49e-09 ***
X2	-0.04317	0.01584	-2.726	0.00654 **
X3	-0.07043	0.01474	-4.779	2.07e-06 ***
X4	0.02979	0.01629	1.829	0.06778
X5	0.11859	0.14152	0.838	0.40228
X6	-0.07298	0.14157	-0.516	0.60632

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A2. Model prediction based on the 4 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	-	+	-	-	+	+	+	+	+	+
35	+	+	-	-	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	-	+	+	-	+	+	+	-	+	-	+	+	+
37	+	+	+	+	-	-	+	+	+	+	+	-	+	+	+	+
38	+	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Total**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	-	**20/30**	
35	+	+	+	-	+	+	+	+	+	-	-	+	+	+	**23/30**	
36	-	+	+	-	+	+	+	+	-	+	+	+	+	-	**22/30**	
37	+	+	+	+	-	+	+	+	-	+	+	+	+	+	**21/30**	
38	-	+	+	-	+	+	+	-	+	+	+	+	+	-	**23/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A3. ANOVA table of the linear regression model based on the 6 previous games (including all variables)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	0.50436	0.10325	4.885	1.25e-06 ***
X1	0.07020	0.01173	5.985	3.26e-09 ***
X2	-0.03550	0.01326	-2.678	0.00755 **
X3	-0.06281	0.01170	-5.369	1.04e-07 ***
X4	0.01322	0.01328	0.995	0.31991
X5	0.06747	0.14594	0.462	0.64396
X6	-0.15097	0.14594	-1.034	0.30123

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A4. Model prediction based on the 6 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	-	+	-	-	+	-	-	+	-	+	+	+	+
35	+	+	+	+	-	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	-	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	-	-	-	+	+	+	-	+	-	+	+	+	+
38	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Total**	
34	+	+	-	+	+	+	+	+	-	+	+	+	-	+	**20/30**	
35	+	+	-	-	+	+	+	+	+	-	-	+	+	+	**23/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	-	**22/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**21/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**23/30**	
Overall model prediction															**73%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A5. ANOVA table of the linear regression model based on the 8 previous games (including all variables)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	0.52480	0.10609	4.947	9.34e-07 ***
X1	0.05272	0.00997	5.288	1.63e-07 ***
X2	-0.03241	0.01166	-2.781	0.00556 **
X3	-0.05644	0.00995	-5.673	2.01e-08 ***
X4	0.01436	0.01173	1.224	0.22125
X5	0.05807	0.14997	0.387	0.69872
X6	-0.19379	0.14994	-1.292	0.19659

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A6. Model prediction based on the 8 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	+	+	+	+	+	-	+	+	-	-	+	+	-	+	+	+
35	+	+	-	+	-	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	-	+	+	-	+	+	-	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	-	+	-	+	+	+	+
38	+	+	+	+	+	+	+	+	+	-	+	-	-	-	+	-
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	-	+	+	+	-	+	**23/30**	
35	+	+	+	+	+	+	+	+	+	-	-	+	+	+	**24/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	-	**21/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**22/30**	
38	-	+	+	-	+	+	+	-	+	+	+	+	+	-	**21/30**	
Overall model prediction															**74%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A7. ANOVA table of the linear regression model based on the 10 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	0.502168	0.109159	4.600	5.03e-06 ***
X1	0.060011	0.008954	6.702	4.31e-11 ***
X2	-0.022762	0.010736	-2.120	0.0343 *
X3	-0.044151	0.008766	-5.036	6.08e-07 ***
X4	0.012225	0.010602	1.153	0.2493
X5	0.133065	0.154355	0.862	0.3890
X6	-0.085024	0.154334	-0.551	0.5819

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A8. Model prediction based on the 10 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	+	+	+	+	+	-	+	-	-	+	+	+	+	+	+	+
35	+	+	-	+	-	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	-	+	+	+	+	+	+
38	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	-	+	+	+	-	+	**24/30**	
35	+	+	+	-	-	+	+	+	+	-	-	+	+	+	**22/30**	
36	+	+	+	+	+	+	-	+	+	+	+	+	+	+	**25/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**23/30**	
38	+	+	+	-	+	+	+	+	+	+	+	+	+	-	**25/30**	
Overall model prediction															**79.3%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A9. ANOVA table of the linear regression model based on the 12 previous games (including all variables)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	0.492834	0.113999	4.323	1.79e-05 ***
X1	0.053196	0.008261	6.440	2.39e-10 ***
X2	-0.019415	0.010057	-1.931	0.054 *
X3	-0.040826	0.008010	-5.097	4.59e-07 ***
X4	0.003457	0.009955	0.347	0.729
X5	0.079820	0.161181	0.495	0.621
X6	-0.106262	0.161164	-0.659	0.510

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A10. Model prediction based on the 12 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	+	+	+	+	+	-	-	+	-	+	+	+	+	+	+	+
35	+	+	-	-	-	+	+	+	-	+	-	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	+
38	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	-	+	+	+	+	+	+	+	-	-	+	**23/30**	
35	+	+	+	-	+	+	+	+	+	-	-	+	+	+	**21/30**	
36	-	+	+	+	+	+	-	-	+	-	+	+	+	+	**22/30**	
37	+	+	+	+	-	+	-	+	-	+	-	+	+	+	**23/30**	
38	+	+	+	-	+	+	+	+	+	+	+	+	+	-	**25/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A11. ANOVA table of the linear regression model based on the 4 previous games (indicator variables for countries are excluded)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	0.51025	0.05787	8.817	< 2e-16 ***
X1	0.08772	0.01456	6.023	2.53e-09 ***
X2	-0.04367	0.01583	-2.759	0.00592 **
X3	-0.07037	0.01474	-4.776	2.10e-06 ***
X4	0.03029	0.01628	1.86	0.06318

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A12. Model prediction based on the 4 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	-	+	-	-	+	+	-	+	+	+
35	+	+	-	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	-	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+
38	+	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	-	+	+	+	+	+	+	+	+	-	+	**21/30**	
35	+	+	+	-	+	+	+	+	+	-	-	+	+	+	**24/30**	
36	-	+	+	-	+	+	+	+	-	+	+	+	+	-	**21/30**	
37	+	+	+	+	-	+	+	+	-	+	+	+	+	+	**25/30**	
38	-	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**75%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A13. ANOVA table of the linear regression model based on the 10 previous games (indicator variables for countries are excluded)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	0.518158	0.063110	8.210	1.10e-15 ***
X1	0.059890	0.008954	6.689	4.68e-11 ***
X2	-0.023078	0.010734	-2.150	0.0319 *
X3	-0.044042	0.008766	-5.024	6.46e-07 ***
X4	0.012467	0.010601	1.176	0.2400

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A14. Model prediction based on the 10 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	+	+	+	+	-	-	+	-	-	-	+	+	+	+	+	+
35	+	+	-	+	-	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	+	+	-	+	+	+	+
38	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	-	+	+	+	+	-	+	+	+	-	-	**20/30**	
35	+	+	+	-	-	+	+	+	+	-	-	+	+	+	**22/30**	
36	+	+	+	+	+	+	-	+	+	+	+	+	+	+	**25/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**23/30**	
38	-	+	+	-	+	+	+	+	+	+	+	+	+	-	**24/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A15. ANOVA table of the logistic regression model based on the 4 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	1.17511	0.13927	8.437	< 2e-16 ***
X1	0.05576	0.02048	2.722	0.00648 **
X2	-0.06449	0.02221	-2.904	0.00369 **
X3	-0.06245	0.01958	-3.189	0.00143 **
X4	0.03252	0.02175	1.495	0.13483
X5	-0.05958	0.19443	-0.306	0.75927
X6	-0.29904	0.18940	-1.579	0.11437

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A16. Model prediction based on the 4 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	-	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	+	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	+	-	+	+	-	+	+	+	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**22/30**	
35	-	+	+	+	+	+	+	+	+	-	-	-	+	+	**24/30**	
36	+	+	+	-	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	-	+	+	-	+	-	+	-	+	+	+	+	-	**22/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A17. ANOVA table of the logistic regression model based on the 6 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	1.19726	0.14558	8.224	< 2e-16 ***
X1	0.04945	0.01681	2.941	0.00327 **
X2	-0.05416	0.01847	-2.932	0.00336 **
X3	-0.06065	0.01555	-3.902	9.55e-05 ***
X4	0.02654	0.01776	1.494	0.13508
X5	-0.13864	0.20144	-0.688	0.49130
X6	-0.37516	0.19672	-1.907	0.05651 *

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A18. Model prediction based on the 6 previous games (including all variables)

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
34	-	+	+	+	+	-	+	-	-	-	+	-	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	-	+	+	-	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**22/30**	
35	-	+	+	+	+	+	+	+	+	-	-	-	+	+	**24/30**	
36	+	+	+	-	+	+	-	+	-	+	+	+	+	+	**23/30**	
37	+	-	+	+	-	+	-	+	-	+	+	+	+	-	**20/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**74%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A19. ANOVA table of the logistic regression model based on the 10 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	1.17934	0.15771	7.478	7.55e-14 ***
X1	0.04957	0.01357	3.652	0.000260 ***
X2	-0.02667	0.01552	-1.719	0.085590 *
X3	-0.04088	0.01183	-3.454	0.000552 ***
X4	0.02189	0.01470	1.489	0.136454
X5	-0.06631	0.21949	-0.302	0.762586
X6	-0.29168	0.21430	-1.361	0.173485

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A20. Model prediction based on the 10 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	+	-	+	+	-	+	+	+	+	+	+	+
38	+	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	-	+	+	+	+	+	+	+	+	-	-	-	+	+	**24/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	-	**23/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**23/30**	
Overall model prediction															**78%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A21. ANOVA table of the logistic regression model based on the 12 previous games (including all variables)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	1.17934	0.15771	7.478	7.55e-14 ***
X1	0.04957	0.01357	3.652	0.000260 ***
X2	-0.02667	0.01552	-1.719	0.085590 *
X3	-0.04088	0.01183	-3.454	0.000552 ***
X4	0.02189	0.01470	1.489	0.136454
X5	-0.06631	0.21949	-0.302	0.762586
X6	-0.29168	0.21430	-1.361	0.173485

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A22. Model prediction based on the 12 previous games (including all variables)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+
38	+	+	+	+	+	+	+	-	+	+	+	-	+	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	-	+	+	+	+	+	+	+	+	-	-	+	+	+	**25/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**24/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**24/30**	
Overall model prediction															**80%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A23. ANOVA table of the logistic regression model based on the 4 previous games (indicator variables for countries are excluded)

Coefficients	**Estimate**	**Std. Error**	**t-value**	**P-value**
(Intercept)	1.05199	0.08042	13.082	< 2e-16 ***
X1	0.05570	0.02051	2.716	0.00660 **
X2	-0.06480	0.02209	-2.933	0.00336 **
X3	-0.06218	0.01959	-3.174	0.00151 **
X4	0.03323	0.02166	1.534	0.12507

***= significant at α<0.001; **=significant at 0.001<α<0.01

Table A24. Model prediction based on the 4 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	-	+	+	+	+
35	+	+	+	+	+	+	+	+	-	-	+	+	-	+	+	+
36	+	+	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**22/30**	
35	-	+	+	+	+	+	+	+	+	-	-	-	+	+	**23/30**	
36	+	+	+	-	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	-	+	+	-	+	-	+	-	+	+	+	+	-	**23/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**76%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A25. ANOVA table of the logistic regression model based on the 10 previous games (indicator variables for countries are excluded)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	1.05652	0.09230	11.446	< 2e-16 ***
X1	0.04980	0.01358	3.667	0.000246 ***
X2	-0.0261	0.01543	-1.692	0.090695 *
X3	-0.0408	0.01185	-3.444	0.000572 ***
X4	0.02198	0.01466	1.499	0.13387

***= significant at α<0.001; **=significant at 0.001<α<0.01; *=significant at 0.01<α<0.05

Table A26. Model prediction based on the 10 previous games (indicator variables for countries are excluded)

#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	+	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	+	-	+	+	-	+	+	+	+	+	+	+
38	-	+	+	+	+	+	+	-	+	+	+	-	-	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	-	+	+	+	+	+	+	+	+	-	-	+	+	+	**25/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**25/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	-	**23/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**22/30**	
Overall model prediction															**79%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"

Table A27. ANOVA table of the logistic regression model based on the 12 previous games (indicator variables for countries are excluded)

Coefficients	Estimate	Std. Error	t-value	P-value
(Intercept)	0.484012	0.065819	7.354	6.08e-13 ***
X1	0.053127	0.008256	6.435	2.46e-10 ***
X2	-0.019628	0.010049	-1.953	0.0512 *
X3	-0.040763	0.008006	-5.092	4.70e-07 ***
X4	0.003617	0.009949	0.364	0.7163

***= significant at α<0.001; *=significant at 0.01<α<0.05

Table A28. Model prediction based on the 12 previous games (indicator variables for countries are excluded)

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
34	-	+	+	+	+	-	+	-	-	-	+	+	+	+	+	+
35	+	+	+	+	+	+	+	+	-	+	+	+	-	+	+	+
36	+	-	+	+	+	+	-	+	+	+	-	+	-	+	+	+
37	+	-	+	+	-	-	+	+	-	+	+	+	+	+	+	+
38	+	+	+	+	+	+	+	-	+	+	+	-	+	-	+	+
#	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**TOTAL**	
34	+	+	-	+	+	+	+	+	+	+	+	+	-	+	**23/30**	
35	+	+	+	+	+	+	+	+	+	-	-	+	+	+	**26/30**	
36	+	+	+	+	+	+	-	+	-	+	+	+	+	+	**24/30**	
37	+	+	+	+	-	+	-	+	-	+	+	+	+	+	**23/30**	
38	+	+	+	-	+	+	+	-	+	+	+	+	+	-	**24/30**	
Overall model prediction															**80%**	

"+" indicates that model was correct, "-" model was incorrect
Column 1 – 10 = "England", 11 – 20 = "Spain", 21 – 30 = "Italy"