Red Card/Goal Analysis

Do red cards seem to lead to more goals?

Introduction

This study investigates the relationship between red cards and goals in football matches, specifically addressing whether red cards lead to more goals overall (i.e. more goals were scored by either team). The analysis uses data from two datasets, *games.csv* and *events.csv*, covering five seasons of games from 10 European national leagues across five countries.

The analysis was conducted using Python in Jupyter Notebooks. Code accessible here.

Structure

Exploratory Data Analysis

This section provides an initial exploration of the datasets, and is split into two parts:

1. Understanding The Data
2. Correlation Analysis

Modelling

We tackle the following questions using hypothesis testing to statistically assess the relationship between red cards and goals, primarily using a Poisson distribution as a basis for comparison.

1. *Do red cards lead to more goals overall?*

   We compare the mean number of goals scored in games *with* red card events against games *without* red card events using a Poisson Means Test (or E-Test).

2. *Do earlier red cards lead to more goals overall?*

   We analyse games with red cards to see if the timing of the first red card event influences the number of goals. First through the Poisson Means Test for a first red card in the *first half* or in the *second half*. We follow this with a Linear Regression model to compare the number of goals to the first red card minute.

Further Analysis

To gain additional insights, we address the following, albeit with less rigorous testing.

3. *Does the rate of goal-scoring increase after a red card?*

   We compare the goal-scoring rate per minute when a red card *has* occurred compared to when a red card *has not* occurred.

4. *Are more goals scored immediately after a red card event?*

   We examine trends to see if goals are more likely to be scored immediately after a red card.

Conclusion

A final look at our study.

Exploratory Data Analysis (EDA)

Understanding The Data

games.csv

19,294 games with the following features

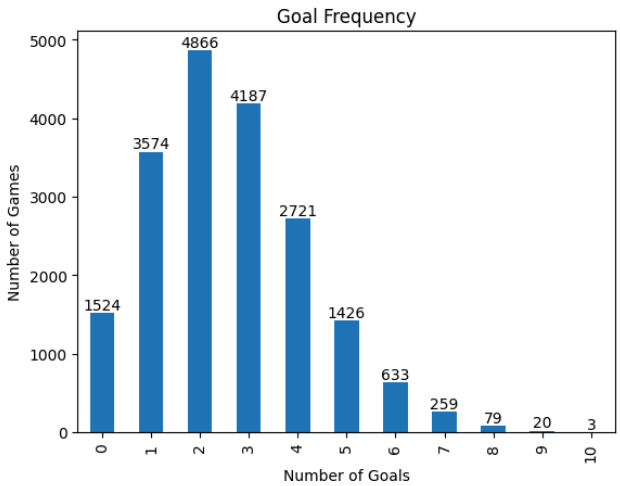
- *game_id* - unique game identifier
- *home_team*
- *away_team*
- *competition* - league and country
- *date* - YYYY-MM-DD format
- *home_goals*
- *away_goals*

Key Numbers:

Number of games: 19,294
Number of goals: 50,356
Mean goals per game: 2.61
Goal variance: 1.63

	total_goals	total_games	mean_goals_per_game
Overall	19294	50356	2.61
Ligue 1, France	1802	4860	2.70
Ligue 2, France	1800	4149	2.31
1. Bundesliga, Germany	1530	4808	3.14
2. Bundesliga, Germany	1530	4486	2.93
Premier League, England	1900	5285	2.78
Championship, England	2761	6932	2.51
Primera, Spain	1900	4784	2.52
Segunda, Spain	2309	5065	2.19
Serie A, Italy	1900	5399	2.84
Serie B, Italy	1862	4588	2.46

Across the 10 leagues, the highest mean goals per game we see is 3.14 (1. Bundesliga, Germany), the lowest is 2.19 (Segunda, Spain). This shows a spread of 0.95 mean goals per game.



To the left, we see the distribution of the number of goals, which matches a Poisson distribution extremely well as we will see later.

It is worth noting that the Poisson distribution assumes a constant rate of an event (in our case goals).

Yet we can see from our league data above, that there is a large variation in the number of goals per game across the different leagues.

So it's important to remember that while Poisson is a good starting point, there is clearly influence from other external factors.

events.csv

54,451 events with the following features

- *game_id* - matching game identifier
- *minute* - time of event 0 to 90, does not include added time
- *side* - home or away
- *type* - red_card or goal

Key Numbers:

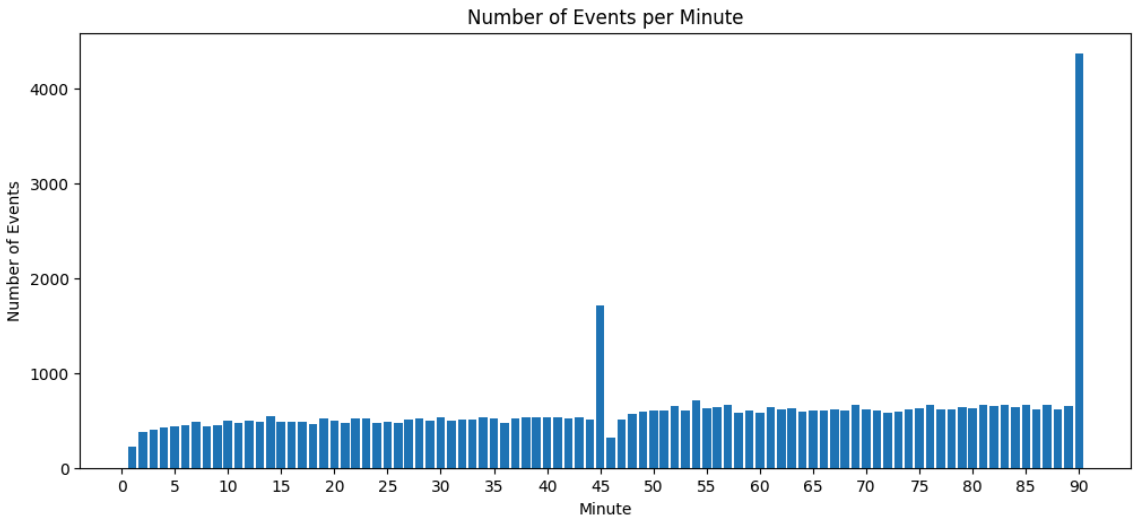
Number of events: 54,451
Number of games with red cards: 3,575
Number of red cards: 4,131
Mean red cards per game: 0.21

We see a substantial difference in the mean red cards per game across different leagues.

- 0.29 in Serie B, Italy and 0.11 in the Premier League, England
- This difference of 0.18 is nearly as large as our overall average of 0.21

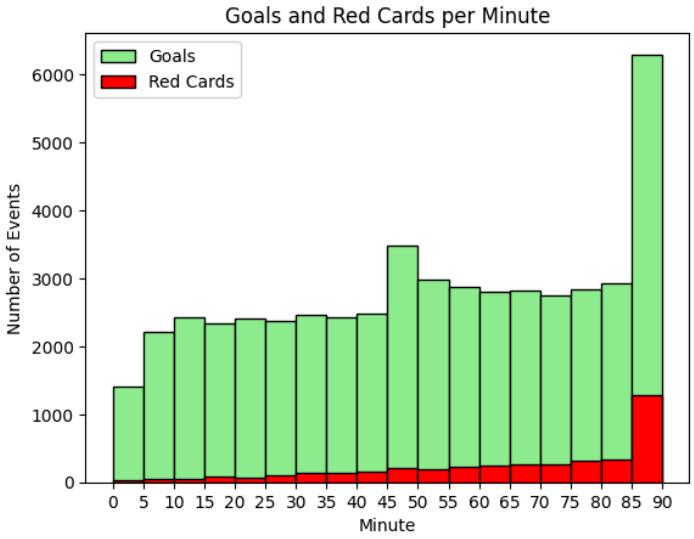
	total_games	total_red_cards	mean_red_cards_per_game
Overall	19294	4131	0.21
Ligue 1, France	1802	434	0.24
Ligue 2, France	1800	446	0.25
1. Bundesliga, Germany	1530	266	0.17
2. Bundesliga, Germany	1530	300	0.20
Premier League, England	1900	312	0.16
Championship, England	2761	428	0.16
Primera, Spain	1900	437	0.23
Segunda, Spain	2309	567	0.25
Serie A, Italy	1900	424	0.22
Serie B, Italy	1862	517	0.28

Below, we can see significant outliers for the number of events taking place at minutes 45 and 90. This shows that our data contains events for ‘added time’ but does not include the exact minutes. This could create inaccuracies if we wish to calculate the 'per minute' values.



Here we see the event frequencies for both goals and red cards across 5 minute windows:

- More goals occur in the second half overall.
- Red cards however, do seem to show a linear increase.
- This makes sense as red cards can be the result of two yellow cards, which are more likely to accumulate over time.



Minute	Events	Goals	Red Cards
90	4364	3523	841
45	1713	1583	130
54	707	660	47
69	668	597	71
85	668	577	91

- Based on the numbers we could make estimates on the amount of 'added time', based on averages for other minutes.
- We can expect an average between 1 to 4 minutes at the end of the first half.
- We can expect an average between 3 to 7 minutes at the end of the second half.
- We could also search for other sources to make a more educated estimate.

- Another point worth noting is for events that occur at the same minute, we do not know the order.
- So for red cards or goals that occur at the same minute, we must skip these values when we are considering the impact of a red card on those goals.
- There were 186 goals at the same minute as a red card, so losing these is not statistically significant.

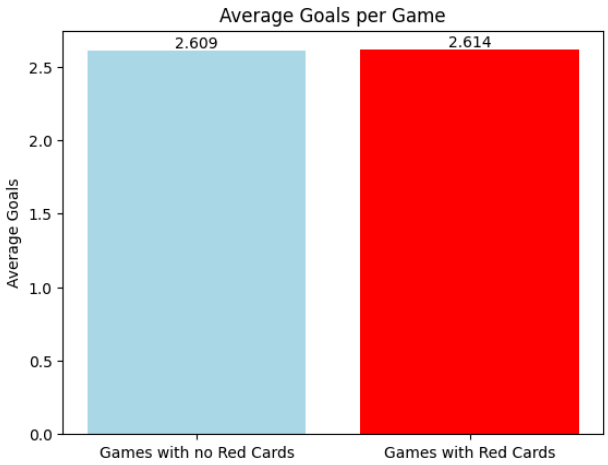
	game_id	minute	side	type
0	1	69	home	red_card
1	4	90	home	red_card
2	14	85	away	red_card
3	26	90	away	red_card
4	29	71	away	red_card
...	...	...	...	...
54446	19292	85	home	goal
54447	19293	43	home	goal
54448	19293	90	home	goal
54449	19293	6	away	goal
54450	19294	68	away	goal

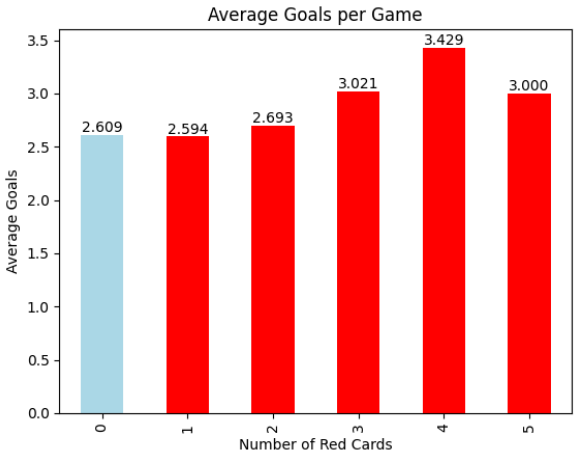
Correlation Analysis

Mean goals in games with no red cards: 2.609
Mean goals in games with red cards: 2.614

Below, we can observe that the difference between mean goals in games with and without red cards is very small (noticeable only at three decimal points).

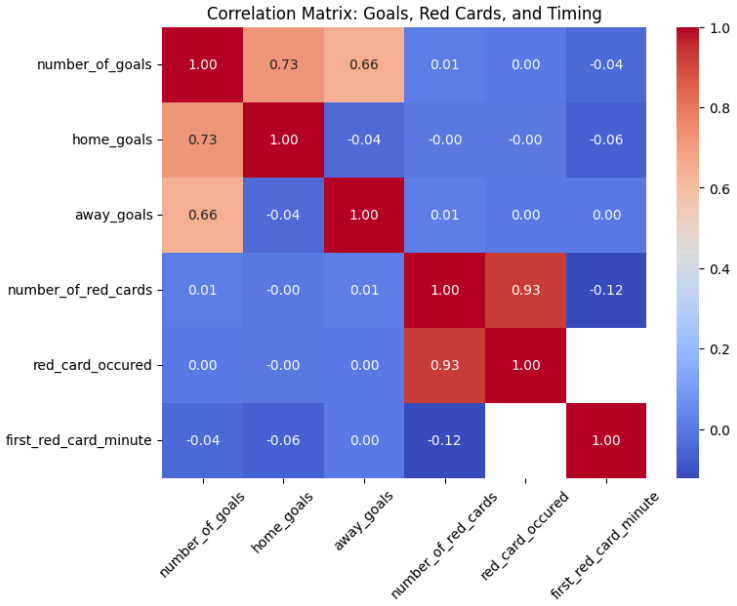
However, on the right we observe that an increase in the number of red cards does have an impact on the number of goals. Potentially, two or more players less on the pitch starts to have an impact, leading to more open areas, particularly when both (or more) red cards are given to the same team. It is important to mention that the number of games with 3, 4, or 5 red cards are too small for statistical analysis.





We have derived features through data processing - number of goals, number of red cards, whether a red card occurred and the first red card minute. Here it is displayed as a correlation heatmap, with reds indicating a strong correlation and blue indicating a weaker correlation.

There does not appear to be strong correlation between our red card indicators with the number of goals. The biggest is the timing of the first red card minute, which has a negative 0.04 correlation. This suggests an earlier red card leads to a higher number of goals. However, this is still a weak correlation, meaning it isn't likely to have a large impact on the number of goals. As we can not see strong correlation, we'll move on to our hypothesis modelling.

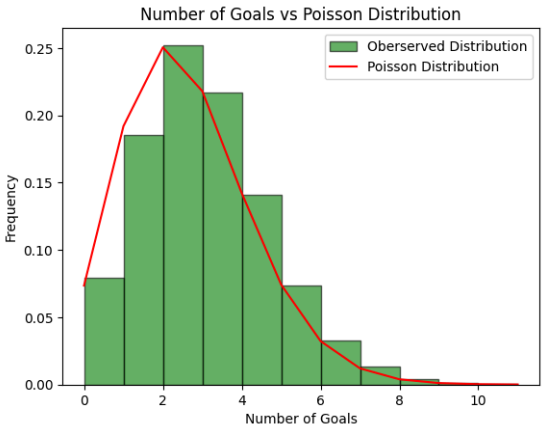


Modelling

Question 1: Do games with red card events (at any point of the game) lead to an increase in the total number of goals?

- **Null Hypothesis (H_0):** Red cards *do not* lead to an increase in the total number of goals.
- **Alternative Hypothesis (H_1):** Red cards *do* lead to an increase in total number of goals.
- **Significance Level/Error Rate:** α = 0.05
- **Model Choice:** Poisson Means Test / E-Test

We can see that Poisson Distribution (red line) appears to be a suitable fit to our data (green bars). We will assume that our data follows a Poisson distribution.



However we should note, as mentioned previously. The Poisson distribution assumes a constant rate of an event. We see from our data in our EDA, goal frequencies do tend to be higher in the second half of games, which suggests a variable rate of goals. We will choose to ignore this for this limited study.

For our hypothesis testing we will use the Poisson Means Test (or E-test). The samples are provided as the number of events k (number of goals) observed within measurement intervals of sizes n (number of games).

For our study that would be:

- k1 = total number of goals in games without red cards
- n1 = number of games without red cards
- k2 = total number of goals in games with red cards
- n2 = number of games with red cards

Key Numbers:

- Test-statistic: -0.155

The test-statistic of -0.155 indicates that the difference between the two groups is very small, and this small difference is likely due to random variation rather than any true underlying effect.

- **P-value:** 0.43

With a p-value of 0.43 or 43% far greater than 0.05, there is insufficient evidence to suggest that games with red cards lead to more goals overall than games without red cards.

Summary:

We are unable to reject the null hypothesis. There is no statistically significant difference in the number of goals between games with red cards and those without.

Question 2: Do earlier red cards lead to more goals overall?

In games where red cards occur, do earlier red card events lead to an increase in the total number of goals?

- **Null Hypothesis (H_0):** In games with red cards, whether the first red card is in the first or second half does not affect the number of goals.
- **Alternative Hypothesis (H_1):** In games with red cards, when the first red card is in the first we will see an increase in the number of goals.
- **Significance Level/Error Rate:** α = 0.05
- **Model Choice:** Poisson Mean Test / E-test

We will again use the Poisson Means Test but this time we will look at all games with red cards and compare games whether the first red card occurred in the first or second half.

The samples for our Poisson Means Test will be:

- k1 = number of goals in red card games with the first red card in the second half
- n1 = number of red card games with the first red card in the second half
- k2 = number of goals in red card games with the first red card in the first half
- n2 = number of red card games with the first red card in the first half

Key Numbers:

- **Test-statistic:** -1.486

The test-statistic of -1.486 indicates the difference between the red cards occuring in the second half only and red cards occuring in the first half is small.

- **P-value:** 0.067

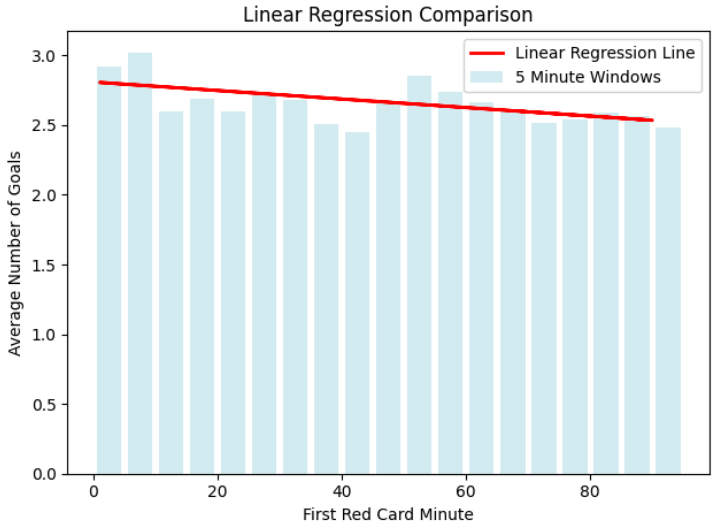
Our p-value of 0.067 or 6.7% is only slightly larger than our significance level of 0.05 or 5%. So once again there is insufficient evidence to support the hypothesis but it does suggest that a red card in the first half leads to a higher number of goals.

Summary:

Once again there is insufficient evidence to support the hypothesis but it suggests that a red card in the first half leads to a higher number of goals.

The hypothesis compares 45 minute windows of the first and second half. We will explore at a more granular level by looking minute by minute.

On the right, we see a linear regression line comparing the average number of goals at the end of the game compared with the timing a first red card has occurred.



Further Analysis

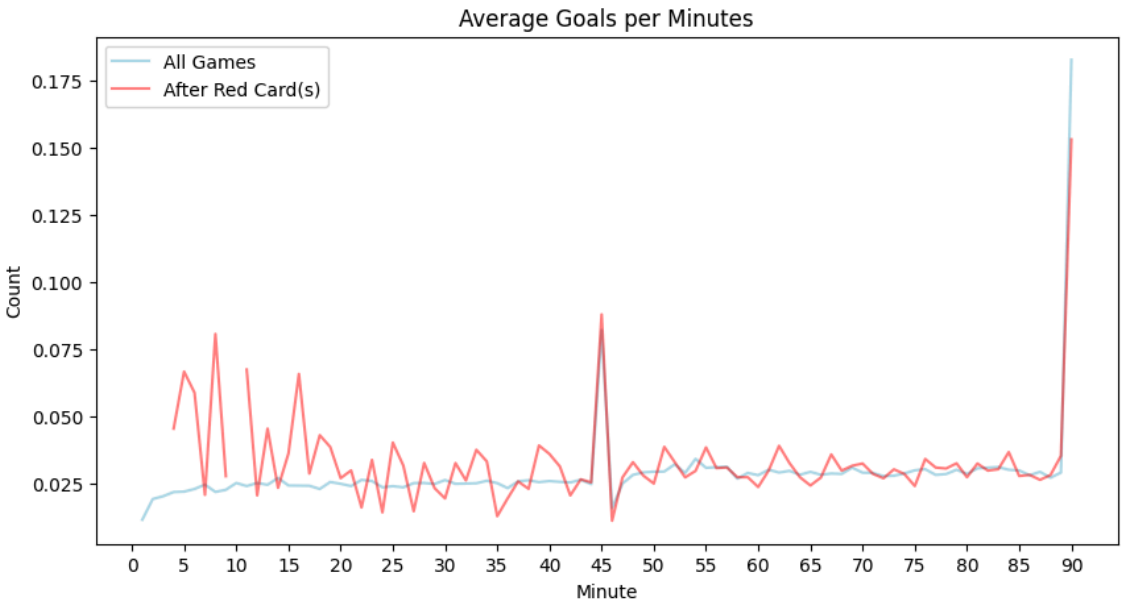
This section provides a preliminary analysis of the following question.

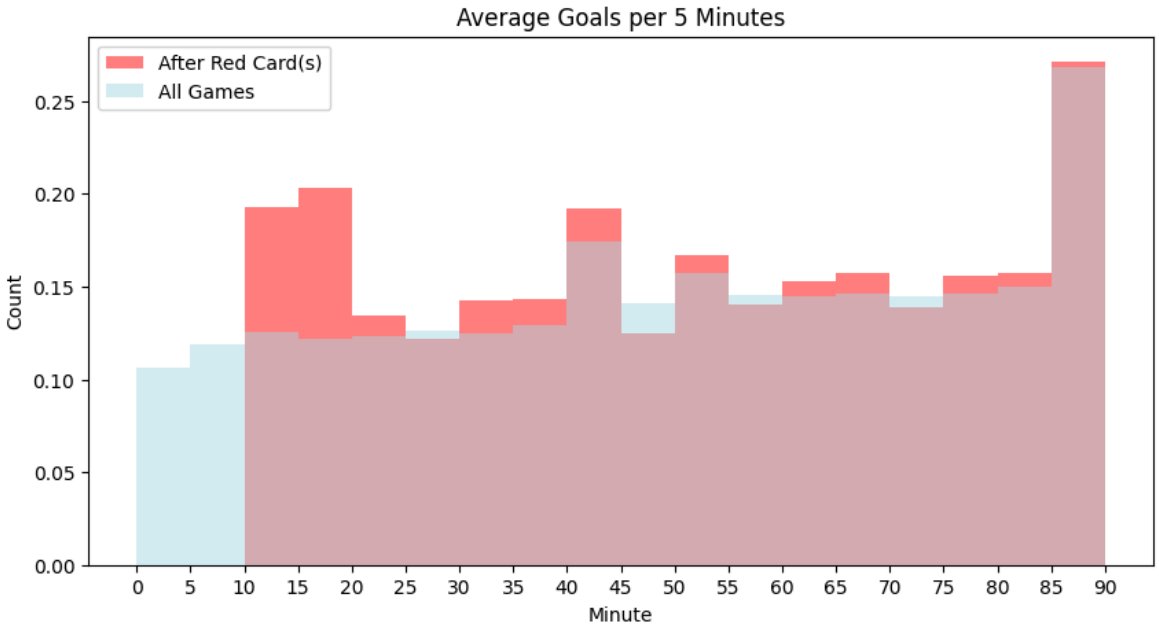
Question 3: Does the rate of goal-scoring increase after a red card?

Is there a higher rate of goals scored when red cards have occurred compared with when red cards have not occurred?

- **Goal Scoring Rate per Minute Before Red Cards:** 0.028
- **Goal Scoring Rate per Minute After Red Cards:** 0.037

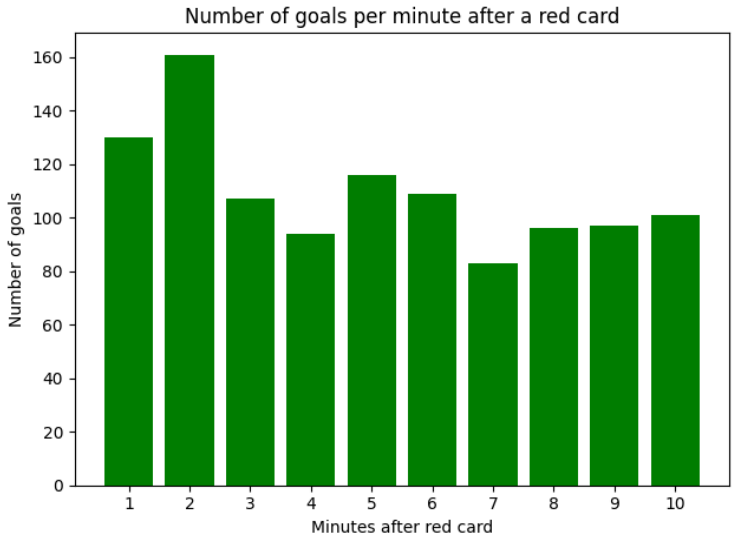
Below these charts display goal scoring rates across games with red cards (red) and all games (blue). They track the minute-by-minute and 5-minute window goal-scoring rates. We do spot trends of earlier red cards having an impact on goals in comparison with regular rates of scoring. As we observed in our overall exploration of the data, red cards are less frequent earlier in the game but it is clear they have an impact.





Question 4: Is there an increase in the number of goals in the minutes immediately after a red card has been given?

The chart shows the number of goals scored per minute following a red card event. Each bar represents the number of goals scored in their respective minutes after a red card is shown.



There is an initial spike in goal-scoring immediately after a red card has been given, especially in the first and second minutes, where the number of goals, 130, and 161, is significantly higher than in later minutes.

Minutes after red card	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Number of goals	**130**	**161**	**107**	**94**	**116**	**109**	**83**	**96**	**97**	**101**

This observation points to a strong immediate impact of red cards on goal-scoring. Potential theories could be either the tactical or emotional impacts of red card events in the short term, that are stabilised afterwards.

The data we have for this particular case is a small sample and requires further statistical analysis to confirm whether this pattern holds across a broader set of games.

However, if this evidence matches to football games as a whole, it may be relevant for in-play betting. For example, if a red card occurs in the 40th minute, and influences the likelihood of goals scored before halftime.

Conclusion

The analysis conducted on the relationship between red cards and goals in football matches suggests several key insights:

1. **Overall Impact of Red Cards**:

   Our first hypothesis, Question 1, that red cards lead to more goals overall was tested using a Poisson Means Test or E-test. The results showed no statistically significant difference between the average number of goals scored in games with red cards and those without. This suggests that red cards, in general, do not have a statistically significant impact on the overall number of goals in a match (assuming that the limitations of using the Poisson distribution noted earlier are not affecting the overall results). However, the difference in the mean numbers between when a red card event did or not occur suggests that there is a small impact.

2. **Timing of Red Cards:**

   Our comparison of a first red card occurring in the first or second half did not provide us enough evidence to support our hypothesis for Question 2. However, it did result in a fairly low p-value, 6.7%, close to significance, and it supported a trend which we were able to see with our investigation through a Linear Regression model. Further analysis of the linear trend revealed that the timing of the first red card event does play a role. Earlier red cards tend to result in more goals. For each additional minute before the red card occurs, the total number of expected goals decreases slightly.

3. **Goal-Scoring Rate per Minute:**

   The analysis of goal-scoring rates after a red card event showed evidence of a higher rate of goals in comparison to before red cards (or none at all). The goal-scoring rate per minute before red cards is 0.028, in comparison to after red cards, which is 0.037. Across 90 minutes, this reflects 2.52 goals per game before red cards, and 3.33 goals per game after red cards. There may be other factors that correlate, such as minutes after red cards more often being in the second half where there are typically more goals, but these are unlikely to have a large impact on our results. For further insights, we could source more accurate records of goal and red card event timings during added time, this would impact our overall expected goals.