7COM1079-0901-2024 - Team Research and Development Project

Final report title: To determine whether there is a correlation between the number of goals

scored and possession of the ball during the 2018 FIFA World Cup.

Group ID: A094

Dataset number: DS051

Prepared by: Sathurcigan Kamalendran - 23057639

Kavindu Hashan Porambage - 23081660

Nipuna Nilupul Samarakoon Samarakoon Mudiyanselage - 23099687

Lakshman Sivarathan - 23086585

Binalka Swarnathilaka Nawarathnalage - 23081218

University of Hertfordshire

Hatfield, 2024

Table of Contents

1. Introduction
   1. Problem statement and research motivation 3
   2. The data set 3
   3. Research question 3
   4. Null hypothesis and alternative hypothesis (H0/H1) 3

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) 4
   2. Why RQ is of interest (research gap and future directions according to the literature) 4

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)* 5
   2. Additional information relating to understanding the data (optional) 5
   3. Useful information for the data understanding 6

1. Analysis
   1. Statistical test used to test the hypotheses and output 6
   2. The null hypothesis is rejected /not rejected based on the p-value 6

1. Evaluation – group’s experience at 7COM1079
   1. What went well 6
   2. Points for improvement 7
   3. Group’s time management 7
   4. Project’s overall judgement 7
   5. Comment on GitHub log output 7

1. Conclusions
   1. Results explained. 8
   2. Interpretation of the results 8
   3. Reasons and/or implications for future work, limitations of your stud 8

1. Reference list 9

Harvard (author, date) format.

1. Appendices
2. R code used for analysis and visualisation. 10
3. GitHub log output. 11

1. Introduction

* 1. Problem statement and research motivation **(100 words)**

Analysing game performance using statistical metrics reveals important insights in competitive sports like football. Ball possession is frequently emphasised as a crucial component of gameplay by coaches, commentators, and players. This is because it is believed that higher possession rates are correlated with improved performance and more opportunities to score goals. However, recent studies, such as Collet (2012), suggest that possession alone may not guarantee success, and its correlation with goals scored remains inconclusive (Collet, 2012). The purpose of this study is to use actual results to support or contradict this commonly held belief by examining the correlation between ball possession and goals scored at the 2018 FIFA World Cup.

* 1. The data set **(75 words)**

The dataset "DS051 and FIFA 2018 Statistics.csv," which contains 128 rows of match-related metrics from the 2018 FIFA World Cup, is used for the analysis. Important variables of interest:

* Goals scored (dependent variable): shows how well a team performed.
* Ball Possession (Independent Variable): Indicates the percentage of the game a team had the ball under control.

The structured format of the dataset makes it possible to perform thorough statistical analysis, such as correlation testing, to determine whether the selected variables have any meaningful relationships.

* 1. Research question **(50 words).**

Is there a correlation between the number of goals scored and ball possession during the FIFA World Cup 2018? This question aims to test whether longer ball possession results in more goals. To answer this, we'll analyse the "DS051 and FIFA 2018 Statistics.csv" dataset using scatterplots, histograms, and Spearman’s Rho correlation to determine if is there any significant correlation.

* 1. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

The study aims to test the relationship between ball possession and the number of goals scored during the FIFA World Cup 2018 by formulating the following hypotheses:

* Null Hypothesis (H₀): There is no correlation between the number of goals scored and the ball possession.
* Alternative Hypothesis (H₁): There is a correlation between the number of goals scored and the ball possession.

In order to make a data-driven conclusion about the correlation between ball possession and number of goals scored, this study will test these hypotheses at a 5% significance level.

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**

The research paper "Goal Statistics and Scoring Attributes of the 2018 FIFA World Cup" by Ankur Biswas and Nita Bandyopadhyay analyses the patterns and trends of goals scored during the 2018 FIFA World Cup in Russia. This research was conducted using data obtained from the official FIFA database and it focused on key variables such as goals, shots, shots on target, ball possession, passes, pass accuracy, fouls, yellow cards, offsides and corners (Biswas & Bandyopadhyay, 2023).

The study called “Analysis of Goal Scoring Patterns in the 2018 FIFA World Cup”, done by Alliance Kubayi, also examines various goal-scoring patterns during the 2018 FIFA World Cup in Russia using InStat video analysis. The primary objective is to classify and analyse all goals scored during the tournament (Kubayi, 2020).

The study "Prediction of the FIFA World Cup 2018 - A Random Forest Approach with an Emphasis on Estimated Team Ability Parameters" by Andreas Groll, Christophe Ley, Gunther Schauberger, and Hans Van compares three different modelling approaches: Poisson Regression models, Random Forests, and Ranking methods to predict the scores of soccer matches based on data from the four previous FIFA World Cups (Groll and et al., 2018).

* 1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

There is a research gap in the limited study of the direct correlation between ball possession and goal scoring in the 2018 FIFA World Cup. In some matches, the team with higher possession lost, indicating variability across games. This inconsistency led to an interest in understanding the relationship between these two variables. Furthermore, while the studies conducted by Biswas et al. analysed various performance indicators, there is a need to pay more attention to how possession affects goal scoring. Future research could use advanced analytics and machine learning to explore this relationship in depth, providing valuable insights for coaches and analysts to improve game strategies.

1. Visualisation
   1. Appropriate plot for the RQ (**50 words)**

The following visualisations are created to investigate the correlation between ball possession and the number of goals scored during the 2018 FIFA World Cup. Our research question is about the correlation, and as a result of the test, the data is not normal. Histogram and scatterplot are used for the visualisation.

* 1. Additional information relating to understanding the data (optional) (**50 words)**
* Scatterplot with Linear Trendline:

A scatterplot was created to illustrate the correlation between the independent and the dependent variables. The trendline and widely scattered scores indicate no correlation.

A graph of a football goal

Description automatically generated

Figure 1 Scatterplot for Ball Possession vs Goal Scored

* Histogram with Normal Curve Overlay:

The histogram shows the frequency distribution of the dependent variable. The data is strongly skewed to the right, and the data is not normally distributed.

A graph with a red line

Description automatically generated

Figure 2 Histogram

* 1. Useful information for the data understanding (**50 words)**

The scatterplot shows no clear trend, with a minimal slope in the trendline, indicating a weak or non-existent linear relationship between the two variables. The histogram reveals that the distribution of goals scored deviates from the normal curve, justifying the use of Spearman’s test for correlation analysis due to the data's non-normality.

1. Analysis
   1. Statistical test used to test the hypotheses and output (**75 words)**

In the study using data from the FIFA World Cup 2018, Ball Possession was the independent variable and Goal Scored was the dependent variable, both being interval types. To find their correlation, a histogram and scatterplot were created. The histogram did not show a bell curve, and the scatterplot shows a horizontal trend line and widely scattered scores, which means the data is non-parametric. Consequently, the Spearman correlation test was performed to analyse the relationship between the two variables.

* 1. The null hypothesis is rejected /not rejected based on the p-value **(100 words)**

The Spearman's rank correlation test performed on the FIFA World Cup 2018 dataset resulted in a p-value of 0.5046. As this value exceeds 0.05, so we cannot reject the null hypothesis, which means no significant correlation exists between the number of goals scored and ball possession. This lack of correlation indicates there is insufficient evidence to assert a meaningful relationship between these two variables during the 2018 FIFA World Cup. Consequently, the analysis suggests that changes in ball possession do not significantly impact the number of goals scored in the context of this dataset.

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**

Team Contribution and Division of Responsibilities: The report was prepared with the contributions of all five members, ensuring a comprehensive and balanced outcome. Each member provided unique skills and expertise to elevate the quality of the report.

Coordination: Effective communication and smooth teamwork were achieved through regular meetings and a collaborative project plan. A shared GitHub repository supported seamless collaboration and version control.

Learning Opportunity: Team members gained practical experience with statistical methods, data visualization, and collaborative research, which are valuable skills for future projects.

* 1. Points for improvement **(75 words)**

Presentation rehearsal:While the presentation was clear, a group practice beforehand might have improved timing and flow, ensuring a more professional delivery.

Diverse analytical methods:By incorporating additional statistical methods alongside correlation analysis, deeper insights into the dataset could have been gained.

Earlier collaboration on report writing:Although all members contributed to the report, had the process started earlier, there would have been more time for feedback and revisions sooner.

* 1. Group’s time management (**50 words)**

Weeks 1-3 focused on selecting and finalizing our dataset as "DS051 and FIFA 2018 Statistics.csv." In Week 4, a GitHub repository was created. Weeks 5-7 involved defining variables, presenting the research question, and visualizing the data. Weeks 8-9 refined the analyses based on feedback. Weeks 10-14 concentrated on report writing. Likely, we managed our time during this research.

* 1. Project’s overall judgement (**50 words)**

The project effectively answered the research question using suitable statistical methods and visualizations. The research project offered useful opportunities to learn how to handle real-world datasets, test hypotheses, and interpret results, even if it couldn't find a correlation between ball possession and the number of goals scored.

* 1. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members **(75 words, write only if applies to your group arrangements)**

We didn't face any changes.

* 1. Comment on the GitHub log output **(50 words)**
* Spearman Test Implementation: This commit contains the R code for the analyse the relationship between goals scored and ball possession.
* Update temp.R: This commit contains the code for generate the histogram and the view of the histogram.
* Preparation for final report writing: Arrangements have been made in this commit to write the final report.

All commits are listed in Appendix B.

1. Conclusions
   1. Results explained (**75 words)**

The analysis found no statistically significant relationship between goals scored and ball possession during the 2018 FIFA World Cup. (p-value of 0.5046 and Spearman's Rho (ρ = 0.0595), weak correlation at the 5% significance level. The scatterplot confirmed variability in scoring regardless of possession percentages while the histogram indicated uneven goal distribution. This suggests ball possession alone isn’t a reliable predictor of scoring outcomes, challenging common assumptions and emphasizing the need to explore additional factors influencing match performance beyond possession statistics.

* 1. Interpretation of the results (**75 words)**

The findings raise doubt on the widely held belief that ball possession has had an immediate impact on the number of goals scored. Even though possession could be dominance, other elements like pass accuracy, on-target, and attempts probably have a bigger impact on goals scored. The findings imply that teams should take a more comprehensive approach to performance evaluation and development rather than depending only on possession-based strategies.

* 1. Reasons and/or implications for future work, limitations of your study (**50 words)**

Future Work: Expand datasets to multiple tournaments for broader insights. Include metrics like pass accuracy and defensive tactics, and apply machine learning. Apply advanced methods like machine learning to uncover deeper patterns.

Limitations: The study only focuses on FIFA World Cup 2018 matches and the correlation between possession and goals, excluding other influential factors.

1. Reference list

Harvard (author, date) format.

Mathan, (2019), ‘Predict FIFA 2018 Man of the Match’, *Kaggle*. Available at: <https://www.kaggle.com/datasets/mathan/fifa-2018-match-statistics> (Accessed: 06 January 2025).

Collet, C. (2012) ‘The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010’, *Journal of Sports Sciences*, 31(2), pp. 123–136. DOI: 10.1080/02640414.2012.727455.

Biswas, A. and Bandyopadhyay, N., (2023), 'GOAL STATISTICS AND SCORING ATTRIBUTES OF FIFA WORLD CUP 2018', *WORLD CONGRESS ON MULTI DISCIPLINARY COHESION FOR POSITIVE HEALTH AND WELL BEING*, pp.172-180, BS Publications, Available at: <https://www.researchgate.net/profile/Ankur-Biswas-5/publication/378658990_GOAL_STATISTICS_AND_SCORING_ATTRIBUTES_OF_FIFA_WORLD_CUP_2018/links/65e2e287adf2362b63634b8e/GOAL-STATISTICS-AND-SCORING-ATTRIBUTES-OF-FIFA-WORLD-CUP-2018.pdf> (Accessed: 26 December 2024).

Kubayi, A., (2020) ‘Analysis of goal scoring patterns in the 2018 FIFA World Cup’, *Journal of human kinetics*, 71, p.205, Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7052713/> (Accessed: 27 December 2024).

Groll, A., Ley, C., Schauberger, G. and Van Eetvelde, H., (2018) 'Prediction of the FIFA World Cup 2018-a random forest approach with an emphasis on estimated team ability parameters', *arXiv preprint arXiv:1806.03208*, Available at: <https://arxiv.org/pdf/1806.03208> (Accessed: 27 December 2024).

1. Appendices
2. R code used for analysis and visualisation ***(not included in the word count)***

Analysis.R code with the appropriate statistics to test the hypotheses.

library(readr)

#show data set

head(FIFA\_2018\_Statistics,2)

df <- FIFA\_2018\_Statistics

View(df)

#lets make all the values in each of `Goal Scored` and `Ball Possession %`columns as numeric

df$`Goal Scored` <- as.numeric(df$`Goal Scored`)

df$`Ball Possession %` <- as.numeric(df$`Ball Possession %`)

#Scatterplot

plot(df$`Ball Possession %`, df$`Goal Scored`,

xlab = "Ball Possession %", ylab =

"Goal Scored", main

= "Scatterplot of Ball Possession % vs Goal Scored") #

abline(lm(df$`Goal Scored` ~

df$`Ball Possession %`), col = "red")

# Spearman Test

cor.test(df$`Ball Possession %`, df$`Goal Scored`, method = "spearman")

# Histogram with normal curve

h <- hist(df$`Goal Scored`,

breaks = 6,

main = "Goal Scored Histogram",

xlab = "Goal Scored",

ylab = "Frequency",

col = "azure",

freq = TRUE)

x <- seq(min(df$`Goal Scored`), max(df$`Goal Scored`), length.out = 100)

mn <- mean(df$`Goal Scored`)

stdDev <- sd(df$`Goal Scored`)

yn <- dnorm(x, mean = mn, sd = stdDev)

box.size <- diff(h$mids)[1] \* length(df$`Goal Scored`)

yn <- yn \* box.size

lines(x, yn, col = "red", lwd = 2)

1. GitHub log output.





