*Please delete all the cursive text before submission. It is here just for your reference*.

*Further: data set – DS, research question – RQ*

*The mark (****x words****) after each subchapter states the word count limit. This indicates the expected amount of information which you can exceed by 10% without losing the mark.*

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

***Please make sure*** *the document spelled correctly (including image labels, section headings, and table of contents). Please use correct punctuation.*

*Make sure your report is grammatically correct.*

University of Hertfordshire

Hatfield, 2024

Table of Contents *add page numbers here*

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

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

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

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

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

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

1. Reference list

Harvard (author, date) format.

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

*The list below outlines the chapter/subchapter numbers, names, word count limits, and explanations of what to write in each section.*

1. Introduction

* 1. Problem statement and research motivation **(100 words)**
* *What is the problem in the area we want to learn more about (motivation for study).*
* *Use at least one citation from related literature for top marks.*
  1. The data set **(75 words)**
  2. Research question **(50 words).** *Explain how you are going to answer your RQ.* **(50 words)**
  3. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

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

* *Was the data set used for some research papers?* *Reference at least 3 relevant research papers to your topic / DS.*
  1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)* (**50 words)**

* *Explain the choice of the plot.*
* *Anything on the plot from R is not counted towards word count limit*
* *Make sure that the plot is from output of an R script (NOT a screenshot)*
* *Make sure that the plot has a caption or title, X and Y-axis labels, legend if appropriate and units.*
* *Make sure the title or caption and axis labels are informative.*
  1. Additional information relating to understanding the data (optional) (**50 words)**
* *Per plot: explain the purpose and insights.*
  1. Useful information for the data understanding (**50 words)**
* *Summarise key observations from the plot.*

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

* *Explain the choice of the test.*
* *Make sure the test is appropriate for the RQ and data.*
  1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)** *(interpret the results)*

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**
   2. Points for improvement **(75 words)**
   3. Group’s time management (**50 words)**
   4. Project’s overall judgement (**50 words)**
   5. 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)**
   6. Comment on the GitHub log output **(50 words)**

*Please comment on the GitHub log output, and refer to it as being placed into**Appendix B.*

*From your Git log, select the three most significant commits during this project and include the following for each:*

1. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*
2. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*
3. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*

1. Conclusions
   1. Results explained (**75 words)**
   2. Interpretation of the results (**75 words)**

* *Interpretation of what the results mean in terms of your RQ and the effect this may have on your population and the wider context of your topic.*
  1. Reasons and/or implications for future work, limitations of your study (**50 words)**

1. Reference list ***(not included in the work count)***

Harvard (author, date) format.

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.