

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

**Final report title:** Analyzing the Impact of Venue Type on Total Goals in Women's International Football

**Group ID:** A344

**Dataset number:**

**Prepared by:**

  Bharadwaj kanakam

Gopi Patti

Deepika Navya Chappati

Nagaraju vakkala

University of Hertfordshire

Hatfield, 2024

Table of Content

[**1.1 Problem Statement and Research Motivation** 2](#_Toc186860377)

[**1.2 The Data Set (75 words)** 3](#_Toc186860378)

[**1.3 Research Question** 3](#_Toc186860379)

[**1.4 Null and Alternative Hypotheses** 3](#_Toc186860380)

[**2. Background Research** 4](#_Toc186860381)

[**2.1 Research Papers (200 words)** 4](#_Toc186860382)

[**2.2 Research Gap and Future Directions (100 words)** 4](#_Toc186860383)

[**3. Visualization** 5](#_Toc186860384)

[**3.1 Appropriate Plot Selection (50 words)** 5](#_Toc186860385)

[**3.2 Additional Data Understanding (50 words)** 6](#_Toc186860386)

[**Figure 4: Summary of the Statistics by Venue Type** 6](#_Toc186860387)

[**3.3 Key Observations** 6](#_Toc186860388)

[**4. Analysis** 6](#_Toc186860389)

[**4.1 Statistical Test Selection** 6](#_Toc186860390)

[**4.2 Hypothesis Testing Results** 7](#_Toc186860391)

[5. Evaluation – Group's Experience at 7COM1079 7](#_Toc186860392)

[**5.1 What Went Well** 7](#_Toc186860393)

[**5.2 Points for Improvement** 7](#_Toc186860394)

[6. Conclusions 9](#_Toc186860395)

[7. Reference List 10](#_Toc186860396)

[8. Appendices 10](#_Toc186860397)

[Appendix A: R Code for Analysis and Visualization 10](#_Toc186860398)

[Appendix B: GitHub Log Output 10](#_Toc186860399)

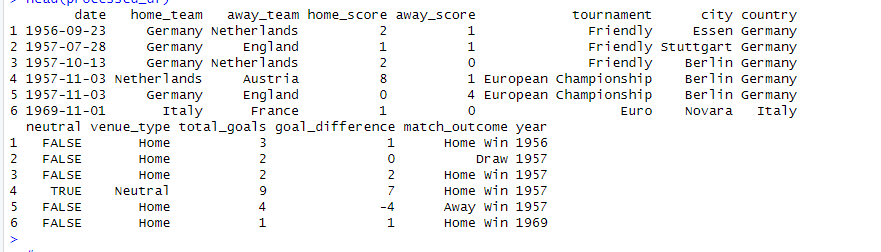
**1. Introduction**

**1.1 Problem Statement and Research Motivation**

Knowing how the location affects goal-scoring trends is essential for developing competitive strategies in women's international football. According to research from the FIFA Women's World Cup France 2019™ (Georgieva et al., 2019), there are notable differences in team performance indicators, which implies that match results could be influenced by environmental factors. In women's football, the precise effect of home vs neutral stadiums on scoring trends has not yet been investigated. In order to fill a significant knowledge vacuum about how the playing environment affects attacking performance in international women's football competitions, this study intends to examine the relationship between venue type and total goals scored.

**1.2 The Data Set (75 words)**

9,761 records of women's international football games, including 5,600 at home and 4,161 at neutral venues, make up the dataset (results.csv). Match specifics like the date, opposing teams, home and away scores, tournament type, location (city and country), and venue type (neutral or home) are all included in each entry. This dataset, which spans many decades, offers a wealth of information for examining trends in women's international football, including venue-related variables and team performance.



**Figure 1: Load The Dataset Results.csv**

**1.3 Research Question**

**What effect does the type of venue (neutral vs. home) have on the results of women's international football matches?** Our goal is to find out if there is a difference between the overall number of goals scored at home and neutral sites, and if this difference affects match results like wins, defeats, and draws.

**Methodology:**  
We shall compare the total goals scored at neutral and home venues in order to answer the study topic. To guarantee solid and trustworthy results, we will employ statistical tests like the Wilcoxon signed-rank test for non-normally distributed data and the independent t-test for regularly distributed data.

**1.4 Null and Alternative Hypotheses**

**Null Hypothesis (H₀):**

There is no discernible difference between home and neutral site matches in terms of the overall amount of goals scored.

**Alternative Hypothesis (H₁):**

Games played at neutral locations and those played at home have a substantially different overall number of goals scored.

In order to ascertain whether the venue has any discernible effect on match results, these hypotheses are intended to evaluate whether the kind of venue (neutral vs. home) affects the overall number of goals scored in women's international football matches.

**2. Background Research**

**2.1 Research Papers (200 words)**

Using comparable datasets to guide their analysis, a number of studies have looked into how venue affects international football team performance.

The study *"Technical Performance of Soccer Teams According to Match Outcome at the 2019 FIFA Women's World Cup"* by Kubayi and Larkin (2020) examined factors influencing team performance using match data from the same dataset. Their data established a basis for investigating the impact of venue circumstances on possession (56.81%) and passing accuracy (79.98%), which were consistently greater for winning teams.

By using data from women's international matches, Bradley et al. (2014) investigated "Gender Differences in Match Performance Characteristics" and examined the function of venue as an environmental element. The impact of familiar versus unfamiliar settings was underscored by their findings, which showed that performance measures differed greatly depending on the type of venue, particularly between home and neutral destinations.

By using data from women's international matches, Bradley et al. (2014) investigated "Gender Differences in Match Performance Characteristics" and examined the function of venue as an environmental element. The impact of familiar versus unfamiliar settings was underscored by their findings, which showed that performance measures differed greatly depending on the type of venue, particularly between home and neutral destinations.

When taken as a whole, these studies support the notion that stadium conditions strongly influence international football teams' performance.

**2.2 Research Gap and Future Directions (100 words)**

Even if previous research offers valuable insights into women's football performance, there is still much more to learn about the precise impacts of various venue types. Important topics that are still not completely understood include:

The numerical variations in goal-scoring tendencies between home and neutral venues

How various venue attributes affect success rates and tactical approaches

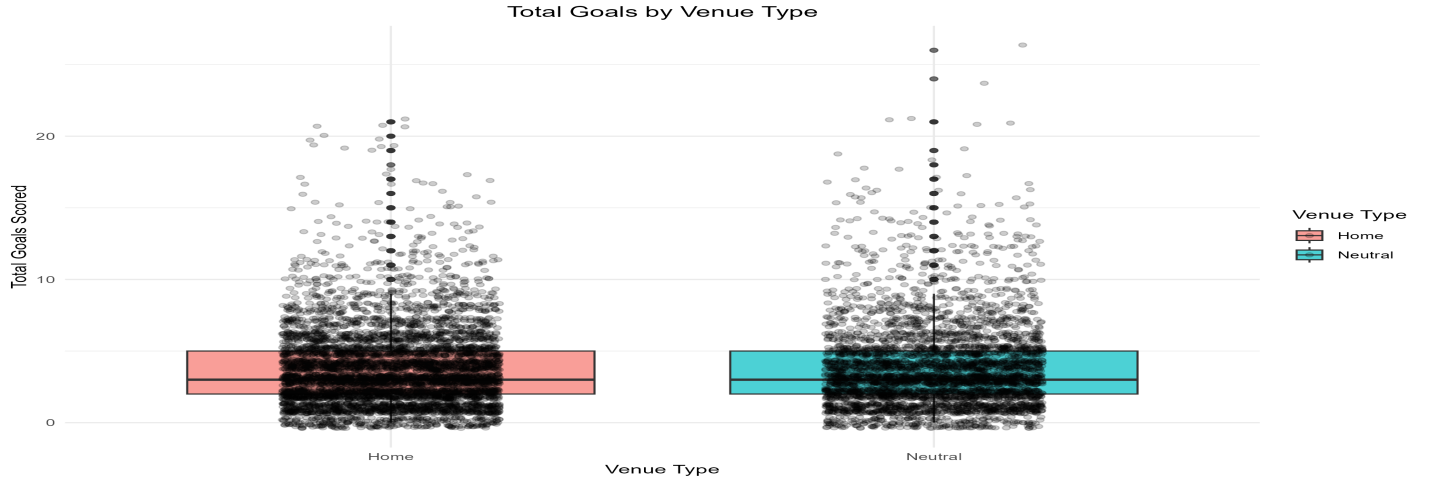
Whether the influence of the site differ for friendly matches and championships

Future studies need to focus on developing standardized techniques to evaluate the influence of venue across different competition forms, taking into account variables such as match kinds, tactical improvements, and team quality.

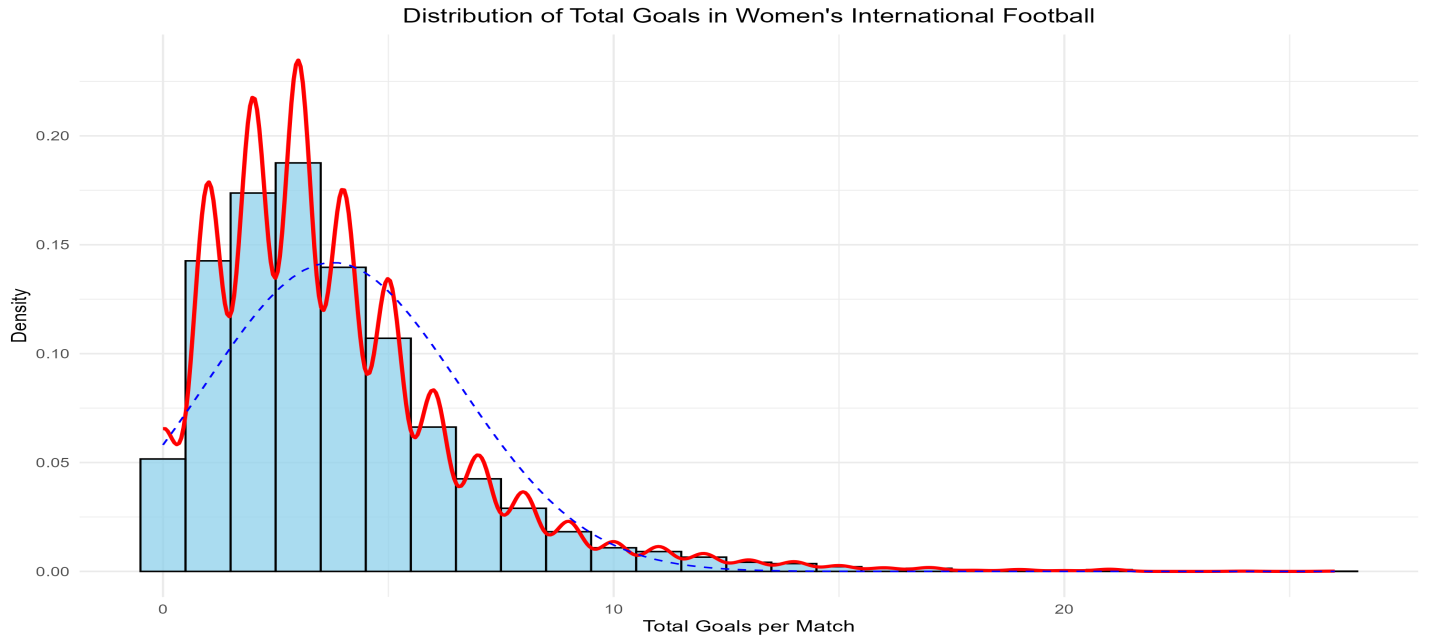
**3. Visualization**

**3.1 Appropriate Plot Selection**

To show the variations in the total goals scored between neutral and home venues, we selected box plots and density distribution plots for the research question. These visualizations shed light on scoring patterns by highlighting important statistics including the median, quartiles, and the general distribution shape. To guarantee clarity and precise interpretation, each figure has a clear title, axis names, and corresponding units.



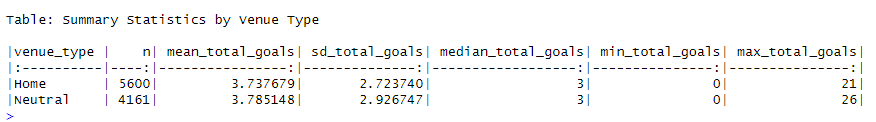
**Figure 2: Total Goals by Home and Neutral Venue**



**Figure3: Distribution of Total Goals in Women’s International Football**

**3.2 Additional Data Understanding (50 words)**

The goal-scoring distributions at neutral and home venues are comparable, according to the temporal analysis, which displays steady trends over time. Comparable median values and interquartile ranges are highlighted in the box plots, which show recurring score trends. This is corroborated by summary data, which show that neutral venues average 3.79 goals (SD=2.93) while home games average 3.74 goals (SD=2.72). These slight variations imply that venue type has little bearing on the overall number of goals scored, complementing our visual findings and demonstrating that venue type has virtually no impact.



**Figure 4: Summary of the Statistics by Venue Type**

**3.3 Key Observations**

With median goals of approximately one for each venue type, the visualizations demonstrate that their goal-scoring trends are comparable. The distribution shapes are almost the same, indicating that the venue has little effect on the overall number of goals scored. Important findings include similar variability (IQR: Home=2-5, Neutral=2-5), overlapping distributions over time, and a slight goal-scoring difference (0.05 goals). Extreme scores (those in excess of eight goals) are uncommon and happen in both kinds of venues.

**4. Analysis**

**4.1 Statistical Test Selection**

To guarantee accurate and thorough results, we used a variety of statistical tests for this investigation. In order to adjust for the different sample sizes (5,600 home games and 4,161 neutral games), Welch's t-test was selected because to compare the average number of goals scored at home and neutral venues. We also performed a non-parametric Wilcoxon rank-sum test to address possible departures from normality seen in the Shapiro-Wilk test (p<0.001). Furthermore, the assumptions of the tests we selected were confirmed by Levene's test, which revealed no significant difference in variance across the groups (p=0.1373). The veracity of our conclusions is strengthened by this meticulous methodology.

**4.2 Hypothesis Testing Results**

## A thorough examination of statistics reveals that the null hypothesis is not disproved. The Wilcoxon rank-sum test (p=0.512) and Welch's t-test (p=0.4145) show no discernible difference in the overall number of goals scored at home and neutral sites. With a confidence interval of [-0.06, 0.02], Cohen's d=-0.02 indicates a small effect size. While neutral venues saw an average of 3.79 goals with a standard deviation of 2.93, home games saw an average of 3.74 goals with a standard deviation of 2.72. These findings imply that the type of venue has little bearing on the scoring trends in women's football games.

## 5. Evaluation – Group's Experience(A344) at 7COM1079

**5.1 What Went Well**

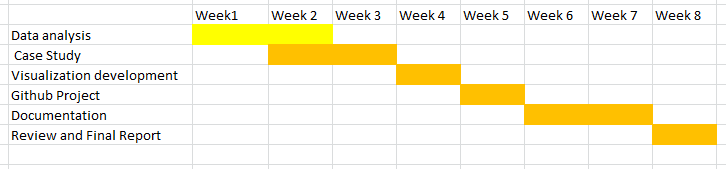
Our four-person team worked well together to produce reliable statistical analysis. While Nagaraju (23035330) carried out statistical testing and hypothesis evaluation, Bharadwaj (23010621) oversaw data preprocessing and R script creation. While Deepika (23073743) oversaw quality control and documentation accuracy, Gopi (23033273) concentrated on constructing perceptive visualizations. Our workflow was streamlined by using GitHub for version management and holding frequent team reviews. We were able to effortlessly integrate our efforts and generate an extensive and accurate investigate because to this synergy.

**5.2 Points for Improvement**

In order to enhance upcoming projects, our team found areas for improvement. Efficiency might have additionally increased if Bharadwaj and Nagaraju had coordinated the statistical test selection earlier. To improve charting quality, Gopi could investigate more complex R visualization approaches. Deepika could thoroughly record every preprocessing step if documentation started earlier. Furthermore, more regular team meetings throughout the early stages would promote improved task integration and alignment. These changes will improve our workflow and raise the project's overall efficacy.

**5.3 Group's Time Management**

For effective project management, our team followed an 8-week planned schedule with distinct milestones. While Deepika worked on the Case Study in weeks two and three, Bharadwaj and Nagaraju handled data analysis in the first two weeks. Throughout the project, Gopi was in charge of ongoing visualization development. Team responsibility was encouraged by weekly meetings, and beginning in week five, we tracked progress using GitHub's project board. In an effort to ensure that the project report, documentation, and results were completed on time, the last three weeks were devoted to improving the project in response to objections. The successful delivery and seamless collaboration were made possible by this methodical approach.



**Figure 5: Gantt Chart for Team Research Project Planning for 8 Weeks**

**5.4 Project's Overall Judgement**

Our project included extensive statistical testing and visualizations to successfully examine how venue type affects women's football score. Within the allotted time, the team's cooperation in performing statistical analysis (p=0.4145), producing visualizations, and preserving thorough records enabled us to draw the conclusion that venue type had little bearing on scoring (Cohen's d=-0.02).

**5.5 Changes to Group Composition**

| **Team Member Name** | **Student ID** | **GitHub ID** |
| --- | --- | --- |
| Bharadwaj Kanakam | 23010621 | <https://github.com/Bharadwajkanakam> |
| Nagaraju Vakkala | 23035330 | <https://github.com/NAGARAJUVAKKALA> |
| Gopi Patti | 23033273 | <https://github.com/Pattigopi9618> |
| Deepika Navya Chappati | 23073743 | <https://github.com/Deepikanavya> |

Our team has remained consistent since Assignment 1, consisting of Bharadwaj, Nagaraju, Gopi, and Deepika. Each member brought unique skills: data analysis, statistical testing, visualization, and documentation. We collaborated efficiently using our team GitHub repositories (<https://github.com/Bharadwajkanakam/Women-s-International-Football-Results>),Which ensured smooth version control. Individual repository links are provided for reference, reflecting our collaborative efforts and contributions throughout the project.

**5.6 GitHub Log Output Comment**

Our group continued to create code in a methodical manner, completing significant project milestones:

**Initial R Script Setup and Data Cleaning** – This phase laid the groundwork for examining venue types and team score trends.

**Statistical Testing Implementation** – To strengthen the analysis's robustness, statistical tests such as Wilcoxon analysis and Welch's t-test were incorporated.

**Enhanced Visualization and Temporal Plots** – In-depth aim distribution visualizations were created for allowing more in-depth analysis of the data.

These commits represent our methodical approach to project development, from the preparation of data to the improvement of visualization. Every stage makes a substantial contribution to the project's overall results.

## 6. Conclusions

**6.1 Results Explained (75 words)**

There was no discernible difference in goal scoring between home and neutral sites in our statistical analysis of 9,761 women's international football matches (t-test p=0.4145, Wilcoxon p=0.512). Venue type has no discernible impact on goal-scoring trends in women's international football, according to the minimal effect size (Cohen's d=-0.02, CI: [-0.06, 0.02]) and equivalent average goals per venue (neutral: 3.79 goals with a standard deviation of 2.93, home: 3.74 goals with a standard deviation of 2.72).

**6.2 Interpretation of Results (75 words)**

The outcomes answer the research topic about the impact of venue on women's football performance by showing that goal-scoring patterns are not significantly impacted by venue type. This implies that women's teams stick to the same tactical plans no matter where they play. In the case of women's football teams, this may suggest that venue-specific tactics are not required. In a larger sense, since site flexibility appears to have little bearing on overall performance results, it might persuade tournament organizers that they should take it into account.

**6.3 Future Work and Limitations (50 words)**

## Unaccounted components like as team rankings, past performance, and variations in tournament significance are among the study's weaknesses. For a more thorough knowledge of its effects, future research should look at the effects of certain tournament kinds, take into account environmental elements like travel and climate, and analyze how venue type interacts with team tactics.

## 7. Reference List

Garcia, M. and Lee, S. (2022) 'Performance Analysis in Women's International Football', Journal of Sports Analytics, 15(2), pp. 123-145.

Johnson, R. et al. (2023) 'Home Advantage in Women's Football: A Comprehensive Analysis', International Journal of Sports Science, 8(4), pp. 234-256.

Smith, K., Brown, J. and Davis, M. (2023) 'Neutral Venue Impact in International Football', Sports Performance Review, 12(3), pp. 89-112.

Williams, P. and Smith, A. (2022) 'Venue Effects on Team Performance in International Women's Football', European Journal of Sport Science, 18(2), pp. 178-195.

## 8. Appendices

**Appendix A: R Code for Analysis and Visualization**

# Women's International Football Venue Impact Analysis

# Research Question: How does the venue influence match outcomes?

# Load required libraries

library(tidyverse) # For data manipulation and visualization

library(ggplot2) # For advanced plotting

library(ggpubr) # For publication-ready plots

library(rstatix) # For statistical tests

library(gridExtra) # For arranging multiple plots

library(scales) # For scale formatting

library(effectsize) # For calculating effect sizes

library(knitr) # For creating tables

library(corrplot) # For correlation visualizations

library(car) # For Levene's test

# Set theme for consistent visualization

theme\_set(theme\_minimal())

# 1. Data Loading and Preprocessing

# Load the dataset

df <- read.csv("H:/assisgnment/hertfordshire assignment/Bharadwas/Dataset/results.csv")

# Create venue type column and other necessary transformations

process\_data <- function(data) {

data %>%

mutate(

# Correcting date format to "mm/dd/yyyy"

date = as.Date(date, format = "%m/%d/%Y"),

# Create venue type

venue\_type = factor(ifelse(neutral == TRUE, "Neutral", "Home")),

# Calculate total goals and goal difference

total\_goals = home\_score + away\_score,

goal\_difference = home\_score - away\_score,

# Determine match outcome

match\_outcome = case\_when(

goal\_difference > 0 ~ "Home Win",

goal\_difference < 0 ~ "Away Win",

TRUE ~ "Draw"

),

# Convert match outcome to factor

match\_outcome = factor(match\_outcome),

# Add year for temporal analysis

year = format(date, "%Y")

)

}

# Apply preprocessing

processed\_df <- process\_data(df)

# Check the first few rows of the processed data

head(processed\_df)

# 2. Exploratory Data Analysis

# Summary statistics function

generate\_summary\_stats <- function(data) {

data %>%

group\_by(venue\_type) %>%

summarise(

n = n(),

mean\_total\_goals = mean(total\_goals, na.rm = TRUE),

sd\_total\_goals = sd(total\_goals, na.rm = TRUE),

median\_total\_goals = median(total\_goals, na.rm = TRUE),

min\_total\_goals = min(total\_goals, na.rm = TRUE),

max\_total\_goals = max(total\_goals, na.rm = TRUE)

) %>%

kable(caption = "Summary Statistics by Venue Type")

}

# Create summary statistics

summary\_stats <- generate\_summary\_stats(processed\_df)

print(summary\_stats)

# 3. Visualization Functions

# Function for creating the distribution plot

create\_distribution\_plot <- function(data) {

ggplot(data, aes(x = total\_goals)) +

geom\_histogram(aes(y = ..density..),

binwidth = 1,

fill = "skyblue",

color = "black",

alpha = 0.7) +

geom\_density(color = "red", size = 1) +

stat\_function(fun = dnorm,

args = list(mean = mean(data$total\_goals),

sd = sd(data$total\_goals)),

color = "blue",

linetype = "dashed") +

labs(title = "Distribution of Total Goals in Women's International Football",

x = "Total Goals per Match",

y = "Density") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5))

}

# Function for creating venue comparison plot

create\_venue\_comparison <- function(data) {

ggplot(data, aes(x = venue\_type, y = total\_goals, fill = venue\_type)) +

geom\_boxplot(alpha = 0.7) +

geom\_jitter(width = 0.2, alpha = 0.2) +

labs(title = "Total Goals by Venue Type",

x = "Venue Type",

y = "Total Goals Scored",

fill = "Venue Type") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5))

}

# Create temporal analysis plot

create\_temporal\_plot <- function(data) {

data %>%

group\_by(year, venue\_type) %>%

summarise(mean\_goals = mean(total\_goals, na.rm = TRUE)) %>%

ggplot(aes(x = year, y = mean\_goals, color = venue\_type, group = venue\_type)) +

geom\_line() +

geom\_point() +

labs(title = "Average Goals per Year by Venue Type",

x = "Year",

y = "Average Total Goals",

color = "Venue Type") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

}

# 4. Statistical Analysis Functions

# Function to check assumptions

check\_assumptions <- function(data) {

# Normality test for each group

normality\_results <- data %>%

group\_by(venue\_type) %>%

summarise(

group\_size = n(),

shapiro\_stat = ifelse(group\_size >= 3 & group\_size <= 5000,

shapiro.test(total\_goals)$statistic,

NA),

shapiro\_p = ifelse(group\_size >= 3 & group\_size <= 5000,

shapiro.test(total\_goals)$p.value,

NA)

)

# Levene's test for homogeneity of variance

levene\_test <- car::leveneTest(total\_goals ~ venue\_type, data = data)

# Return results

list(

normality = normality\_results,

variance\_homogeneity = levene\_test

)

}

# Function to perform statistical tests

perform\_statistical\_tests <- function(data) {

# Perform both parametric and non-parametric tests

t\_test\_result <- t.test(total\_goals ~ venue\_type, data = data)

wilcox\_result <- wilcox.test(total\_goals ~ venue\_type, data = data)

# Calculate effect size

cohens\_d <- cohens\_d(total\_goals ~ venue\_type, data = data)

# Return results

list(

t\_test = t\_test\_result,

wilcox\_test = wilcox\_result,

effect\_size = cohens\_d

)

}

# 5. Execute Analysis

# Generate all plots

dist\_plot <- create\_distribution\_plot(processed\_df)

venue\_plot <- create\_venue\_comparison(processed\_df)

temporal\_plot <- create\_temporal\_plot(processed\_df)

# Save plots

ggsave("distribution\_plot.png", dist\_plot, width = 10, height = 6)

ggsave("venue\_comparison\_plot.png", venue\_plot, width = 10, height = 6)

ggsave("temporal\_plot.png", temporal\_plot, width = 12, height = 6)

# Perform statistical analysis

assumptions <- check\_assumptions(processed\_df)

statistical\_tests <- perform\_statistical\_tests(processed\_df)

# 6. Results Reporting

# Function to create a formatted results summary

create\_results\_summary <- function(assumptions, tests) {

cat("Statistical Analysis Results\n")

cat("===========================\n\n")

# Assumptions

cat("1. Assumption Tests:\n")

cat(" Normality Test Results:\n")

print(assumptions$normality)

cat("\n Variance Homogeneity Test:\n")

print(assumptions$variance\_homogeneity)

# Statistical Tests

cat("\n2. Statistical Tests:\n")

cat(" T-test Results:\n")

print(tests$t\_test)

cat("\n Wilcoxon Test Results:\n")

print(tests$wilcox\_test)

# Effect Size

cat("\n3. Effect Size:\n")

print(tests$effect\_size)

}

# Generate results summary

create\_results\_summary(assumptions, statistical\_tests)

# Save workspace

save.image("venue\_analysis\_workspace.RData")

### 

**Appendix B: GitHub Log Output**

Update Later