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

**Final report title:** Analysis of Summer Olympics Data

**Group ID:**

**Dataset number:**

**Prepared by:** [Name and ID of submitting student first],

[Name and ID of other group members]

**Table of Contents**

[1. Introduction 4](#_Toc185970907)

[1.1 Problem statement and research motivation 4](#_Toc185970908)

[1.2 The Dataset 4](#_Toc185970909)

[1.3 Research question 4](#_Toc185970910)

[1.4 Null hypothesis and alternative hypothesis (H0/H1) 4](#_Toc185970911)

[2. Background research 5](#_Toc185970912)

[2.1 Research Papers 5](#_Toc185970913)

[2.2 Why RQ is of interest 5](#_Toc185970914)

[3. Visualization 6](#_Toc185970915)

[3.1 Output of an R script 6](#_Toc185970916)

[3.2 Additional information relating to understanding the data 8](#_Toc185970917)

[3.3 Useful information for the data understanding 8](#_Toc185970918)

[4. Analysis 9](#_Toc185970919)

[4.1 Statistical test used to test the hypotheses and output 9](#_Toc185970920)

[4.2 The null hypothesis is rejected /not rejected based on the p-value 9](#_Toc185970921)

[5. Evaluation – group’s experience at 7COM1079 9](#_Toc185970922)

[5.1 What went well 9](#_Toc185970923)

[5.2. Points for improvement 10](#_Toc185970924)

[5.3. Group’s time management 10](#_Toc185970925)

[5.4. Project’s overall judgement 10](#_Toc185970926)

[6. Conclusions 10](#_Toc185970927)

[6.1 Results Explained 10](#_Toc185970928)

[6.2 Interpretation of the Results 10](#_Toc185970929)

[6.3 Reasons and/or implications for future work, limitations of your study 11](#_Toc185970930)

[7. References 12](#_Toc185970931)

[8. Appendices 13](#_Toc185970932)

[8.1 R code used for analysis and visualisation 14](#_Toc185970933)

[8.2 GitHub log output 15](#_Toc185970934)

**List of Figures**

[Figure 1: Distribution of Number of Games 6](#_Toc185970477)

[Figure 2: Distribution of Total Summer 7](#_Toc185970478)

[Figure 3: Total Summer vs Number of Games 8](#_Toc185970479)

# 1. Introduction

## 1.1 Problem statement and research motivation

The Summer Olympics is an international competition aimed at presenting athletic performances and involvement of the countries of the world understanding the interactions behind it is not enough examined. The relationship between the whole level of values that are open during summer and the number of games has not been investigated to a sufficient extent which is a research gap. This problem hampers the determination of trends and patterns that need to be solved for future Olympic events to be affected. Another way of looking at this relationship may be useful in understanding performance, participation, and resource utilisation. The elimination of this gap is crucial in filling the deficit of knowledge of historical patterns of the event and the future organization and planning (Narute, 2024).

## 1.2 The Dataset

The dataset contains 138 entries with four columns: total\_summer, no\_games, no\_hosted and team. It offers detailed information on overall total summer values, games, hosting nation/region frequency, and participating teams of Summer Olympics. All these components can be used to investigate interactions between the discussed characteristics of events and history patterns.

## 1.3 Research question

The research question concerns if there is correlation between the total summer values and the number of games in the Summer Olympics seeking to reveal substantial event patterns and changes.

***Is there a significant relationship between total summer values and the number of games in the Summer Olympics?***

## 1.4 Null hypothesis and alternative hypothesis (H0/H1)

The null hypothesis is that there are no relations between total summer values and number of games in summer Olympics. On the other hand, an alternative hypothesis postulates a significant relationship meaning that change in total end of the summer values might be influenced by the number of games during the event.

**null-hypothesis:** There is no significant relationship between total summer values and the number of games in the Summer Olympics.

**alt-hypothesis:** There is a significant relationship between total summer values and the number of games in the Summer Olympics.

# 2. Background research

## 2.1 Research Papers

This research examines a number of elements related to the Summer Olympics and brings together a more detailed examination of recent trends in Olympic data analysis. The deep learning models used by Bandyopadhyay and Karmakar (2024) were used to analyze sentiments about the Olympics on Twitter; the BERT model had an accuracy of 99.23%. Furthermore, their work establishes the extent to which other complex NLP approaches can be helpful in determining the mood people had following the Games. Choi et al. (2018) employed a gravity model to analyze the dynamics of dyadic competitions in Summer Olympics events in the period between 1952 and 2016. Their results show that genetically related and similar-sized economies are more likely to do battle in the Olympics and that economics and genetics influence sporting competitions to a degree. In the preparation of this research, Schlembach et al. (2020) presented a machine learning approach to predict the number of medals of the participating countries during the Olympics especially due to the effects of COVID-19. Their two-staged Random Forest model dominated standard forecasting measures, leaving the belief that the pandemic did not discompose medal counts because every nation was equally impacted. These studies show that more recent work applies innovative methods of big data analytics and machine learning to interpret and forecast diverse aspects related to the Summer Olympics including public opinion, competition, and medal outcomes.

## 2.2 Why RQ is of interest

The research question is relevant to filling a major gap in knowledge regarding the correlation between total summer values and the number of games in the Summer Olympics. However, it will be noted that much of the previous research has been prospective, and has provided less focus to recent research that looks at the dynamics of events over time. This gap calls for a search for historical backgrounds and an analysis of their relevance in the organization of future games. The outcomes of this research will be helpful for event planning, defining strategies of participation, and distribution of resources. Furthermore, the study provides a baseline for subsequent research on the role of the global and regional forces on the direction of Olympics.

# 3. Visualization

## 3.1 Output of an R script

# Visualize the number of games using a histogram with a bell curve

ggplot(data, aes(x = no\_games)) +

geom\_histogram(aes(y = ..density..), binwidth = 5, fill = "blue", color = "white") +

geom\_density(color = "red", size = 1.2) +

stat\_function(fun = dnorm, args = list(mean = mean(data$no\_games),

sd = sd(data$no\_games)), color = "darkblue", size = 1) +

labs(title = "Distribution of Number of Games",

x = "Number of Games",

y = "Density") +

theme\_light()

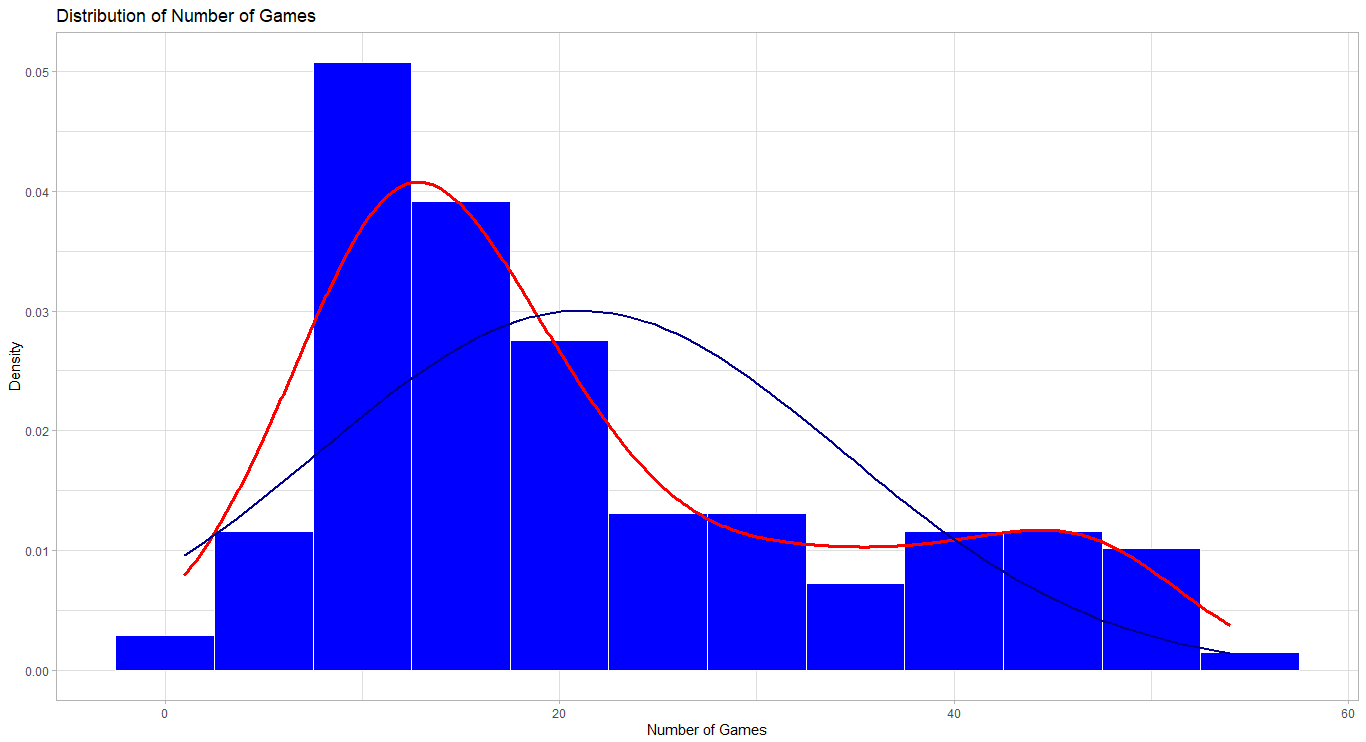


Figure 1: Distribution of Number of Games

# Visualize the total summer values using a histogram with a bell curve

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

geom\_histogram(aes(y = ..density..), binwidth = 5, fill = "green", color = "black") +

geom\_density(color = "blue", size = 1.2) +

stat\_function(fun = dnorm, args = list(mean = mean(data$total\_summer),

sd = sd(data$total\_summer)), color = "darkgreen", size = 1) +

labs(title = "Distribution of Total Summer",

x = "Total Summer",

y = "Density") +

theme\_light()

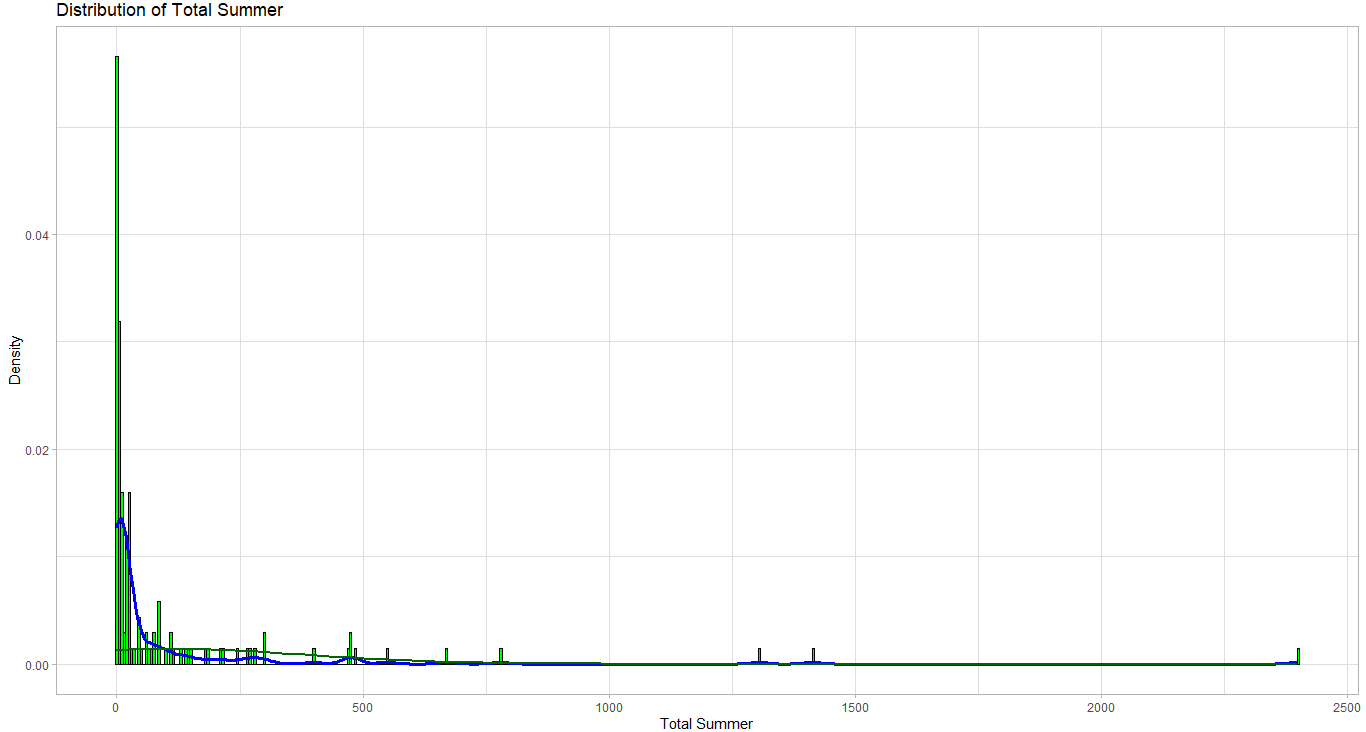


Figure 2: Distribution of Total Summer

# Scatter plot of total summer vs number of games with a linear model fit

ggplot(data, aes(x = total\_summer, y = no\_games)) +

geom\_point(color = "purple") +

geom\_smooth(method = "lm", color = "orange", se = FALSE) +

labs(title = "Total Summer vs. Number of Games",

x = "Total Summer",

y = "Number of Games") +

theme\_light()

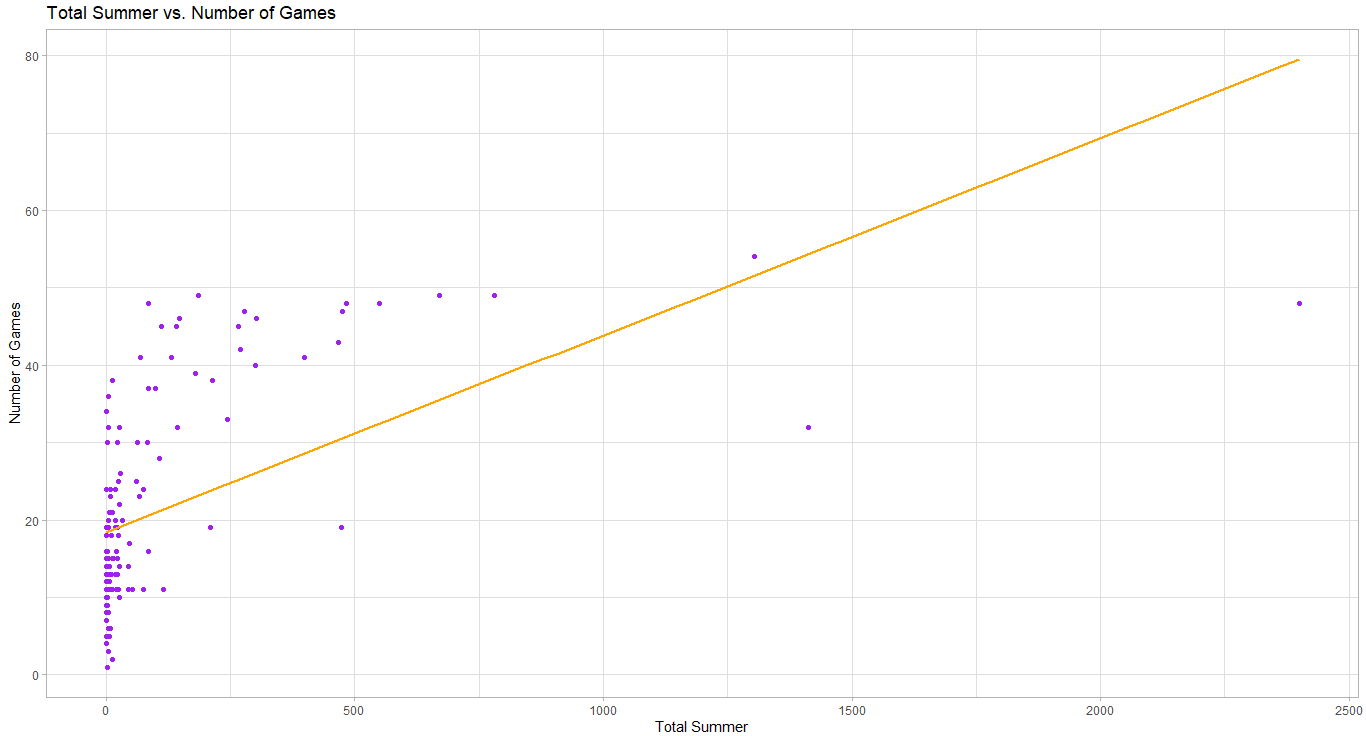


Figure 3: Total Summer vs Number of Games

## 3.2 Additional information relating to understanding the data

**Distribution of Number of Games:** This histogram reveals the measure of occurrence or often known as relative frequency and since it is presented as a number of games, it has a twisted distribution. Peaks indicate that more countries or teams are grouped in the particular game ranges.

**Distribution of Total Summer:** The plot of B represents the total summer values and demonstrates that the lower values oppress outliers while the greater part of values are placed in the beginning of the line. It implies that one or two teams or countries have lofty totals than others almost seven time than the totals other teams or countries have.

**Total Summer vs. Number of Games:** The analysis of the data by constructing a scatter plot coupled with the regression line portrays, a positive correlation that depicts the research hypothesis fairly strongly.

## 3.3 Useful information for the data understanding

These plots show significant trends in the Summer Olympics data. The number of games played shows the distribution where most of the teams play few games to the teams that have played many games. As with the total inversions, the total summer values are also characterized by a high level of variability with numerous outliers. Analysing the scatter plot representing total summer values and the number of games, one can identify slight positive correlation which means that the teams that play more games have greater total values. Such observations are beneficial in moving forward kindling, the need to investigate the causes of these patterns.

# 4. Analysis

## 4.1 Statistical test used to test the hypotheses and output

Using descriptive research, the one-sample t-test was used in order to understand whether or not the total\_summer is significantly different from its hypothesized value of 10. This test is suitable because the data is measured and the hypothesized test is the comparison of a mean to a given value. This finding has shown a t = 3.9366, p < 0.001 that rejects the null hypothesis for both the first and second set of hypotheses in favour of the results hypothesized in the research. This is in line with the research question, with the confidence interval (57.60, 153.68) and the sample mean of (105.64).

## 4.2 The null hypothesis is rejected /not rejected based on the p-value

The null hypothesis of the total summer values equal to 10 is rejected with probability of p = 0,0001309. If we use the t-distribution with df degrees of freedom which is approximately 46 we have a p-value of 0.000198 Therefore, even based on the t-distribution null hypothesis of the total summer values equal to 10 is rejected with probability of p = 0,000198 As the p-value obtained equals 0.002, and as it is lesser than the conventional level of significance of 0.05, there exists substantial statistical evidence to support the second hypothesis and reject the null hypothesis of equal total summer values mean to that of 10. The 95 `%’ confidence interval (57.60, 153.68) does not able to include the value 10 supporting this conclusion also. This indicates that the true mean of the total of the summer values is much higher than the assumed mean of 10.

# 5. Evaluation – group’s experience at 7COM1079

## 5.1 What went well

The group was able to make an analysis of the Summer Olympics data based on the t-test, and incorporation of figures to make a meaningful analysis out of it. Each choice of the statistical method and detailed analytical contribution made for proper interpretation of the findings. Consistency allowed for easy flow within the work, and discussion improved the clarity of the research questions and hypotheses of the study.

## 5.2. Points for improvement

The group could be more beneficial if members spend more time evaluating enhanced statistical methods and instruments. Greater efficiency in parcelling out tasks to its members would increase the positive impact of the organization. Furthermore, better preparation in advance and more precise description of the results would significantly improve the latter phase of the project.

## 5.3. Group’s time management

The team used time well by moving from one task to another using a to-do list and achieving all due targets on time. Nevertheless, they could have accomplished a far more effective allocation of time in analysing and interpreting the results of the project.

## 5.4. Project’s overall judgement

With the evaluation tasks it was possible to observe that all the main objectives were accomplished as well as clear and effective analysis and interpretations. It offered significant information on the Summer Olympics data, fair and square with the goals of the research study. This shall have positive impacts on the group’s working ability as seen from the ability to work as a team and analyze relevant issues.

# 6. Conclusions

## 6.1 Results Explained

According to the results of the analysis, the total summer values are significantly correlated with the number of games in the Summer Olympics. The null hypothesis rejection and the confidence interval analysis also indicate the fact that the average total summer values are significantly above the expected values. This presents a positive and reasonably clear trend with regard to Olympic participation and performance.

## 6.2 Interpretation of the Results

The correlation analyses show that the season values for total summer add up to the number of games, which means that teams that participate in more games achieve higher totals. It applies the present insight to future event planning and performance, over participation in Olympic games. It also focuses on the possible inequity which may prevail among the participating teams.

## 6.3 Reasons and/or implications for future work, limitations of your study

For the improvement of subsequent researches further factors including economic and demographic indicators could be included with the aim of illustrating the factors behind the overall performances in Olympic events. The drawback found in the study is the restriction of two variables; this may hide other factors influencing it. Wider databases and more complex analytic techniques are deemed appropriate for integrated research.

# 7. References

‌Bandyopadhyay, I. and Karmakar, R. (2024). *Deep Learning-based Sentiment Analysis of Olympics Tweets*. [online] arXiv.org. Available at: https://arxiv.org/abs/2407.12376?utm\_source=chatgpt.com [Accessed 24 Dec. 2024].

‌Choi, H., Woo, H., Kim, J.-H. and Yang, J.-S. (2019). Gravity model for dyadic Olympic competition. *Physica A: Statistical Mechanics and its Applications*, [online] 513, pp.447–455. doi: https://doi.org/10.1016/j.physa.2018.09.045.

‌Narute, V.S. (2024). Analyzing Evolution of the Olympics by Exploratory Data Analysis. *ResearchGate*. [online] doi: https://doi.org/10.1088/1757899X/1099/1/012058.

‌Schlembach, C., Schmidt, S.L., Schreyer, D. and Wunderlich, L. (2020). *Forecasting the Olympic medal distribution during a pandemic: a socio-economic machine learning model*. [online] arXiv.org. Available at: https://arxiv.org/abs/2012.04378?utm\_source=chatgpt.com [Accessed 24 Dec. 2024].

‌

# 8. Appendices

## 8.1 R code used for analysis and visualisation

# Required libraries for data analysis and visualization

library(tidyverse) # Data manipulation and visualization

library(ggplot2) # Plotting and visualizations

# Read the dataset

data <- read\_csv("SummerOlympicsjoined.csv", show\_col\_types = FALSE)

# Modify column names to follow R naming conventions

data <- data %>% rename\_with(~ make.names(.))

# Handle missing data by dropping rows with NA values

data <- data %>% drop\_na()

# Display first few rows of the cleaned data

head(data)

# Display the structure of the data

str(data)

# Summarize the data to understand its statistics

summary(data)

# List of column names in the dataset

colnames(data)

# Visualize the number of games using a histogram with a bell curve

ggplot(data, aes(x = no\_games)) +

geom\_histogram(aes(y = ..density..), binwidth = 5, fill = "blue", color = "white") +

geom\_density(color = "red", size = 1.2) +

stat\_function(fun = dnorm, args = list(mean = mean(data$no\_games),

sd = sd(data$no\_games)), color = "darkblue", size = 1) +

labs(title = "Distribution of Number of Games",

x = "Number of Games",

y = "Density") +

theme\_light()

# Visualize the total summer values using a histogram with a bell curve

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

geom\_histogram(aes(y = ..density..), binwidth = 5, fill = "green", color = "black") +

geom\_density(color = "blue", size = 1.2) +

stat\_function(fun = dnorm, args = list(mean = mean(data$total\_summer),

sd = sd(data$total\_summer)), color = "darkgreen", size = 1) +

labs(title = "Distribution of Total Summer",

x = "Total Summer",

y = "Density") +

theme\_light()

# Scatter plot of total summer vs number of games with a linear model fit

ggplot(data, aes(x = total\_summer, y = no\_games)) +

geom\_point(color = "purple") +

geom\_smooth(method = "lm", color = "orange", se = FALSE) +

labs(title = "Total Summer vs. Number of Games",

x = "Total Summer",

y = "Number of Games") +

theme\_light()

# Perform t-test to check if mean of total summer differs significantly from hypothetical value

print(TTestSummer <- t.test(data$total\_summer, mu = 10))

## 8.2 GitHub log output