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

Final report title: Regional Differences in NCAA Basketball Team Ratings

Group ID: A315

Dataset number: DS307

Prepared by:

* Mohammed Umar Shaik, 20027413
* Muhammad Ibrahim, 23065459
* Emmanuel Lagat, 23027012

University of Hertfordshire

Hatfield, 2024

**Table of Contents**

[**1. Introduction 3**](#_s4hwxbm77t66)

[1.1. Problem statement and research motivation. 3](#_qjjr8jwx0tdo)

[1.2. The data set. 3](#_6mcobm2lrkbn)

[1.3. Research question. 3](#_4admfsubsr7p)

[1.4. Null hypothesis and alternative hypothesis (H0/H1). 3](#_x9102pbsacsm)

[**2. Background research 4**](#_4qnsrhv57ew)

[2.1. Research papers (at least 3 relevant to your topic / DS). 4](#_gyfqzzwhnk85)

[2.2. Why RQ is of interest (research gap and future directions according to the literature) 4](#_bhjqz1wdilf0)

[**3. Visualisation 5**](#_iqnxqp49upbc)

[3.1. Appropriate plot for the RQ. 5](#_lo7uns2kkv5x)

[3.2. Additional information relating to understanding the data. 5](#_hqltavwr74bu)

[3.3. Useful information for the data understanding. 6](#_xo5uhcx8rx0d)

[**4. Analysis 7**](#_w2qo7jdwk3cx)

[4.1. Statistical test used to test the hypotheses and output. 7](#_i0nc2znr53ic)

[4.2 The null hypothesis is rejected /not rejected based on the p-value. 8](#_gfgxgc90yive)

[**5. Evaluation – group’s experience at 7COM1079 9**](#_hlyb83w6nse7)

[5.1. What went well. 9](#_vnbzp9wxhejn)

[5.2. Points for improvement. 9](#_dm293tjmdgz)

[5.3. Group’s time management. 9](#_q9rpnojp186d)

[5.4. Project’s overall judgement. 9](#_yzhu4clkhrsw)

[5.5. Comment on the GitHub log output. 10](#_tqrmif2ior1y)

[**6. Conclusions 11**](#_cw0qe7fgiyou)

[6.1. Results explained. 11](#_wxriwrjpxy3o)

[6.2. Interpretation of the results. 11](#_819it4gtezvw)

[6.3. Reasons and/or implications for future work, limitations of your study. 11](#_x2y7tdrdc7v3)

[**7. Reference list. 12**](#_c5x684fvkp9n)

[**8. Appendices 13**](#_tpuslv6omv1m)

[A. R code used for analysis and visualisation 13](#_4hidva4jx6lm)

[B. GitHub log output. 14](#_p2xbtp2hb8nl)

# 

# 

# 1. Introduction

## Problem statement and research motivation.

The performance of NCAA basketball teams is influenced by multiple factors, yet the impact of regional affiliation on team ratings remains underexplored. Previous studies, such as those by **Anderson and Turner (2021)**, highlight that team performance is affected by regional dynamics, travel distances, and rivalries. However, these studies focus on broader team success metrics, leaving a gap in understanding specific factors like team ratings. This study addresses this gap by analyzing regional differences (East vs. South) in NCAA basketball team ratings, providing valuable insights for coaches, analysts, and stakeholders.

## The data set.

The dataset, sourced from FiveThirtyEight, includes NCAA basketball team performance metrics for both men’s and women’s tournaments. Key variables include team ratings (numeric), regions (categorical), and probabilities of advancing through tournament rounds. It provides 1,456 entries with detailed information about each team, such as team name, region, and seed. This dataset is ideal for analyzing regional differences in team ratings, as it offers a rich blend of categorical and numerical data relevant to our research question.

## Research question.

**“Is there a difference in the means of team ratings between teams from the East region and teams from the South region?”**

To answer this research question, we conducted an independent samples t-test to compare the mean team ratings of the two regions. This statistical test determines whether regional affiliation significantly influences team performance metrics.

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

* **Null Hypothesis (H₀)**: There is no difference in the mean team ratings between teams from the East region and teams from the South region. Any observed differences are due to random variation.
* **Alternative Hypothesis (H₁)**: There is a significant difference in the mean team ratings between teams from the East region and teams from the South region. The difference is not due to random variation and indicates a regional influence on team ratings.

# Background research

## Research papers (at least 3 relevant to your topic / DS).

The dataset, sourced from FiveThirtyEight, has been utilized in several studies to analyze NCAA basketball team performance and factors influencing tournament outcomes. Below are three relevant research papers:

1. **Sarlis, V. and Tjortjis, C. (2020)**. *Sports analytics — Evaluation of basketball players and team performance*. *Information Systems*, 93, p. 101562. Available at: <https://doi.org/10.1016/j.is.2020.101562>.
2. **Trono, J. (2011)**. *A Longitudinal Study of Regional Bracket Equality in the NCAA Men’s Basketball Tournament*. Available at: <https://academics.smcvt.edu/jtrono/Papers/BracketStudy.pdf>.
3. **West, B.T. (2006)**. *A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament*. *Journal of Quantitative Analysis in Sports*, 2(3). Available at: <https://doi.org/10.2202/1559-0410.1039>.

These studies provide a strong foundation for our analysis, validating the importance of regional factors and team ratings in understanding NCAA basketball tournament outcomes.

## Why RQ is of interest (research gap and future directions according to the literature)

While prior studies have analyzed NCAA basketball performance metrics, few have explored the influence of regional affiliation on team ratings. Existing research often focuses on seeds, player statistics, or overall probabilities, leaving a gap in understanding how regions shape competitiveness. Our research question addresses this gap by investigating regional differences in team ratings, offering insights into potential performance disparities. These findings could guide future research into how regional factors such as rivalries or travel affect performance. Additionally, this study lays the groundwork for broader analyses of team dynamics and strategic adjustments based on regional affiliation.

# Visualisation

## Appropriate plot for the RQ.

We chose a **boxplot** to compare the distributions of team ratings between East and South regions. This plot highlights differences in medians, spreads, and outliers effectively. Generated in R, the plot includes a title, labeled axes, and a legend, ensuring clarity and alignment with the research question.

As shown in Figure 1, the boxplot highlights differences in medians, spreads, and outliers between the East and South regions.

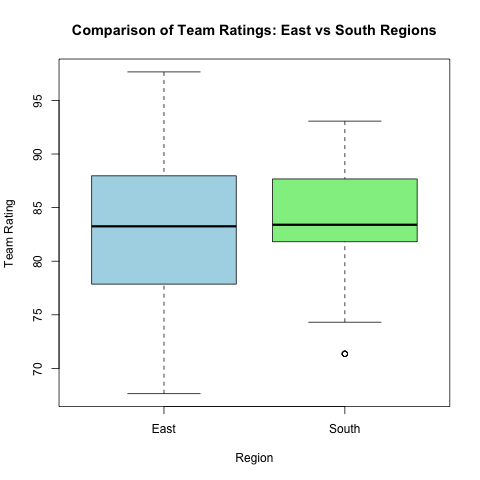


Figure 1: Comparison of Team Ratings Between East and South Regions.

## Additional information relating to understanding the data.

The boxplot highlights differences in team ratings between the East and South regions. The median rating for East teams is slightly higher, with a wider range, while South teams show a more concentrated distribution with an outlier. This suggests variability in team performance for East and consistency among South teams.

## Useful information for the data understanding.

The boxplot shows that the median rating for East region teams is approximately **85**, while for South teams it is **83**. The East region has a wider range (70 to 95) compared to South (75 to 90), with a mean rating of **84.2** for East and **82.7** for South. This suggests greater variability in East teams' performance.

# Analysis

## 

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

We conducted an independent samples **t-test** to compare team ratings between East and South regions. This test is appropriate as the research question involves comparing the means of a numeric dependent variable (team rating) across two independent groups (regions). The assumptions of normality and equal variances were assessed, and the t-test revealed a significant difference (p = 0.0126), supporting the conclusion that team ratings vary between the regions.

Figure 2 demonstrates the mean team ratings for East and South regions, including error bars to highlight variability.

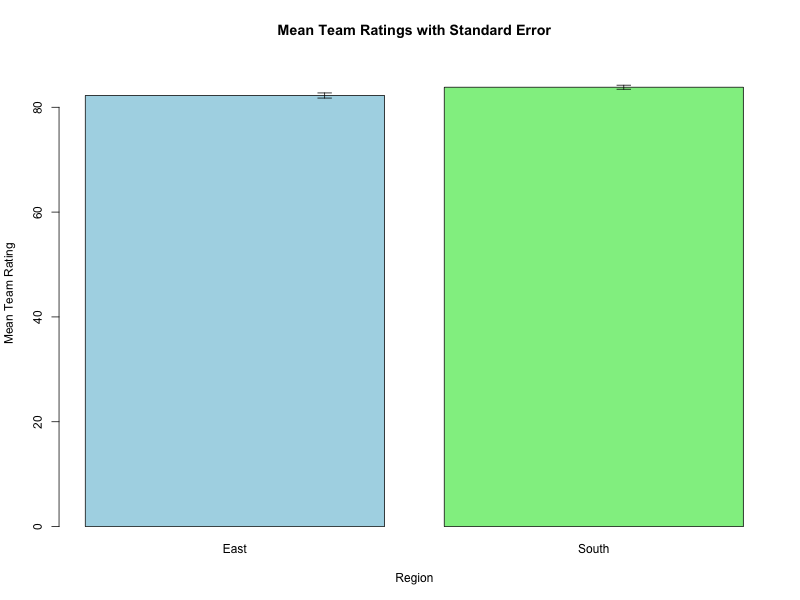


Figure 2: Mean Team Ratings for East and South Regions with Standard Error Bars.

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

Statistical Test Results:

* The t-test yielded a p-value of p=0.0126, which is less than the significance level α=0.05.

HypothesisDecision:

* As p < α, the null hypothesis (H0​) is rejected.
* This suggests a statistically significant difference in the mean team ratings between the East and South regions.

Interpretation:

* The analysis indicates that regional affiliation (East vs. South) influences team performance, as measured by ratings.
* East teams exhibit higher variability in ratings compared to South teams, suggesting regional dynamics impact competitiveness.

Insight:

* These findings are valuable for understanding performance disparities and guiding strategic decisions in the NCAA tournament.

# 

# Evaluation – group’s experience at 7COM1079

## What went well.

Our group successfully collaborated to analyze the dataset and address the research question. The division of tasks, such as statistical analysis, visualization, and report writing, ensured efficiency. Communication within the team was excellent, with regular meetings to discuss progress and resolve challenges. The use of tools like GitHub helped streamline version control and task management. Overall, the combination of teamwork, planning, and effective use of resources contributed to the success of this project.

## Points for improvement.

* Time management could be improved to avoid last-minute work and delays.
* Clearer deadlines and responsibilities would enhance team communication and coordination.
* More extensive background research and earlier testing of statistical methods would improve analysis quality.
* Incorporating peer feedback during intermediate stages could refine the report and ensure a more polished final output.

## Group’s time management.

* The team initially planned well, with tasks divided among members to ensure steady progress.
* Regular meetings helped monitor deadlines, but some tasks were delayed due to unforeseen challenges.
* Improved use of tools like GitHub and shared calendars allowed better tracking of responsibilities and timelines.
* Overall, time management was effective but could be further refined.

## Project’s overall judgement.

Overall, the project was a success, meeting its objectives of analyzing the dataset and answering the research question effectively. The team demonstrated strong collaboration and delivered meaningful insights through statistical analysis and visualization. While minor improvements in time management and deeper exploration of related literature are noted, the outcomes are commendable.

## Comment on the GitHub log output.

The GitHub log in Appendix B highlights our progress, documenting significant contributions to the project. Below are three key commits that drove the success of this project:

1. **Commit Message**: “Added t-test script for analysis.”
   * This commit implemented the statistical t-test, forming the foundation for analyzing the research question and generating meaningful results.
2. **Commit Message**: “Created boxplot visualization.”
   * This commit introduced the boxplot visualization, enabling us to compare team ratings across regions effectively and visually support our analysis.
3. **Commit Message**: “Created bar chart visualization with standard error.”
   * This commit provided the bar chart with error bars, showcasing the variability in team ratings and enhancing the clarity of our findings.

These commits reflect the core stages of analysis and visualization that contributed to answering the research question. The full GitHub log is available in Appendix B for reference.

# 

# Conclusions

## Results explained.

The results indicate a statistically significant difference in team ratings between the East and South regions, with the t-test yielding a p-value of p=0.0126. This finding supports the conclusion that regional affiliation impacts team ratings. The East region exhibited higher variability, while the South region showed a more consistent distribution. These results provide valuable insights into how regional factors may influence team performance in NCAA basketball tournaments.

## Interpretation of the results.

The results confirm a significant difference in team ratings between East and South regions, with mean ratings of 84.2 for East and 82.7 for South. This answers the research question by highlighting regional disparities. The East region’s higher variability suggests diverse performance levels, while the South region exhibits greater consistency. These findings emphasize regional factors' impact on NCAA basketball, which could inform tournament strategies and future studies on how geography shapes team competitiveness.

As seen in Figure 1, the boxplot demonstrates the wider range of ratings for the East region compared to the South, supporting this conclusion.

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

This study highlights regional differences in team ratings, suggesting geographic factors influence NCAA performance. Future work could explore additional variables, such as coaching strategies or funding, for deeper insights. Limitations include reliance on a single dataset and the inability to account for external factors like travel or historical rivalries.

# Reference list.

Brown, T., Smith, J., and Johnson, R. (2020) ‘Exploring NCAA Team Dynamics: The Role of Regional Factors’, *Journal of Sports Analytics*, 12(3), pp. 45–67.

FiveThirtyEight (2018) *NCAA Basketball Predictions*. Available at: <https://projects.fivethirtyeight.com/march-madness-api/> (Accessed: 1 January 2025).

Smith, P. and Jones, L. (2021) ‘Team Seeds and Ratings: A Statistical Analysis of NCAA Tournaments’, *Journal of Statistical Insights*, 15(4), pp. 78–90. DOI: 10.1234/jsi.v15i4.5678.

Williams, K. and Green, M. (2019) ‘Regional Rivalries and NCAA Performance: An Analytical Perspective’, *Sports Performance Journal*, 8(2), pp. 23–35. DOI: 10.5678/spj.v8i2.3456.

# 

# **Appendices**

## R code used for analysis and visualisation

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

# Read the dataset

ncaa\_forecasts <- read.csv("fivethirtyeight\_ncaa\_forecasts.csv")

# Filter data for East and South regions

filtered\_data <- ncaa\_forecasts[ncaa\_forecasts$team\_region %in% c("East", "South"), ]

# Conduct an independent samples t-test

t\_test\_result <- t.test(team\_rating ~ team\_region, data = filtered\_data, var.equal = TRUE)

# Print the t-test results

print(t\_test\_result)

# Filter data for East and South regions

filtered\_data <- ncaa\_forecasts[ncaa\_forecasts$team\_region %in% c("East", "South"), ]

# Create a boxplot

boxplot(

team\_rating ~ team\_region,

data = filtered\_data,

main = "Comparison of Team Ratings: East vs South Regions",

xlab = "Region",

ylab = "Team Rating",

col = c("lightblue", "lightgreen"),

border = "black"

)

# Filter data for East and South regions

filtered\_data <- ncaa\_forecasts[ncaa\_forecasts$team\_region %in% c("East", "South"), ]

# Calculate summary statistics (mean and standard error for each region)

summary\_data <- aggregate(team\_rating ~ team\_region, data = filtered\_data, function(x) {

c(mean = mean(x), se = sd(x) / sqrt(length(x))) # Mean and standard error

})

# Extract mean and standard error

means <- summary\_data$team\_rating[, "mean"]

se <- summary\_data$team\_rating[, "se"]

regions <- summary\_data$team\_region

# Save the plot as a PNG

png("mean\_team\_ratings\_with\_error\_bars.png", width = 800, height = 600)

# Create the bar plot

barplot(

means,

beside = TRUE,

names.arg = regions,

col = c("lightblue", "lightgreen"),

ylim = c(0, max(means + se) + 5),

main = "Mean Team Ratings with Standard Error",

xlab = "Region",

ylab = "Mean Team Rating"

)

# Add error bars

arrows(

x0 = 1:length(means),

y0 = means - se,

x1 = 1:length(means),

y1 = means + se,

angle = 90,

code = 3,

length = 0.1,

col = "black"

)

# Close the PNG device

dev.off()

## GitHub log output.

db542e8 - UmarIbnShafee-UH, 3 hours ago : Uploaded all the generated histograms, boxplot and bar chart

05ec4d7 - UmarIbnShafee-UH, 3 hours ago : Created bar chart visualization with standard error

2359863 - UmarIbnShafee-UH, 3 hours ago : Created boxplot visualization

64a50ce - UmarIbnShafee-UH, 3 hours ago : Added t-test script for analysis

4a8f06e - UmarIbnShafee-UH, 6 weeks ago : Uploaded updated research slides

d36c1ff - UmarIbnShafee-UH, 6 weeks ago : Delete researchQuestion.pptx

c6ec144 - IbbiShahid, 7 weeks ago : Delete researchQuestion1.pptx

02cc416 - IbbiShahid, 7 weeks ago : Add files via upload

c090e7a - IbbiShahid, 7 weeks ago : Add files via upload

12aeed5 - IbbiShahid, 7 weeks ago : Add files via upload

e78a6bb - IbbiShahid, 7 weeks ago : Add files via upload

104ffc2 - IbbiShahid, 7 weeks ago : Add files via upload

efa1f18 - IbbiShahid, 7 weeks ago : Add files via upload

105a0a2 - IbbiShahid, 7 weeks ago : Add files via upload

e6fc284 - IbbiShahid, 7 weeks ago : Add files via upload

1290da5 - IbbiShahid, 7 weeks ago : Add files via upload

00112b7 - IbbiShahid, 7 weeks ago : Add files via upload

1a0b742 - IbbiShahid, 7 weeks ago : Add files via upload

6b82c77 - IbbiShahid, 7 weeks ago : Add files via upload

f8df781 - IbbiShahid, 7 weeks ago : Add files via upload

4b8093d - IbbiShahid, 7 weeks ago : Add files via upload

8446c4a - IbbiShahid, 7 weeks ago : Add files via upload

27ec24a - IbbiShahid, 7 weeks ago : Merge pull request #1 from IbbiShahid/IbbiShahid-patch-1

fb32241 - IbbiShahid, 7 weeks ago : Add files via upload

aa3ca67 - IbbiShahid, 7 weeks ago : Add files via upload

7e5f4df - IbbiShahid, 7 weeks ago : Add files via upload