RUSTAMJI INSTITUTE OF TECHNOLOGY
BSF ACADEMY, TEKANPUR

Lab File for
CS605 (Data Analytics)

Submitted by
Rahul Pandey (0902EC211048)
B.Tech. Computer Science & Engineering 3rd Semester
(2021-2025 batch)

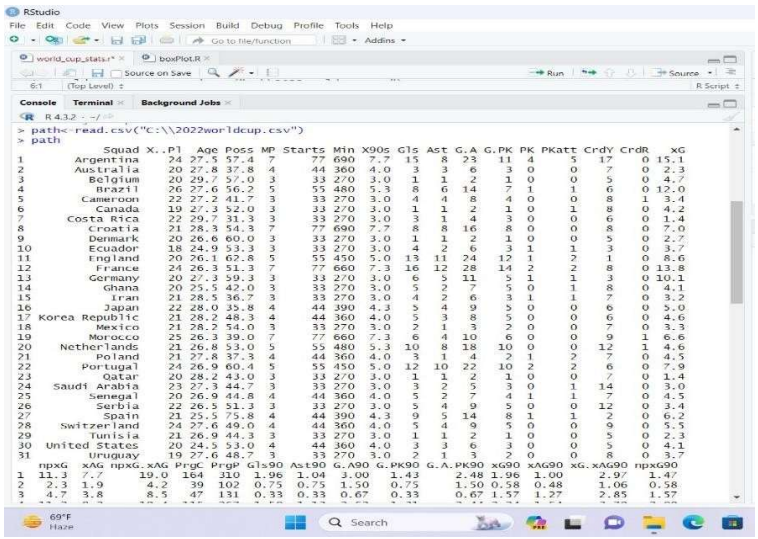
Subject Teacher
Dr. Jagdish Makhijani

File Checked by
Mr. Yashwant Pathak

INDEX

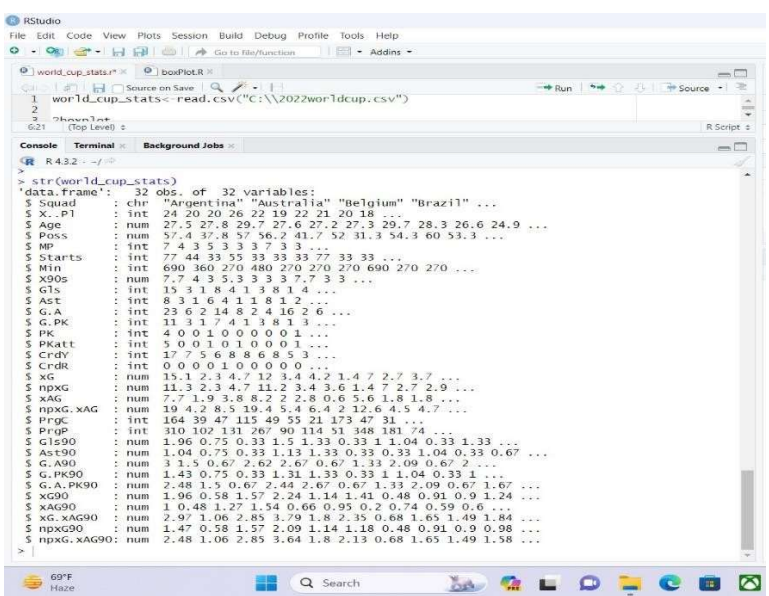
1. Load the data:
   - Use the `read.csv` function to load the downloaded CSV file into an R data frame named `world_cup_stats`.

Solution:



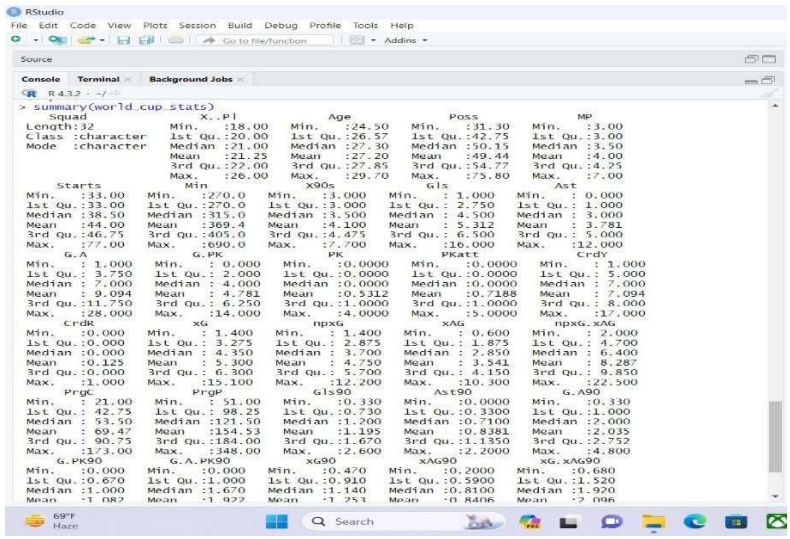
2. Explore the data:
   - Use the `str(world_cup_stats)` function to get an overview of the data frame, including data types and number of observations and variables.

Solution:



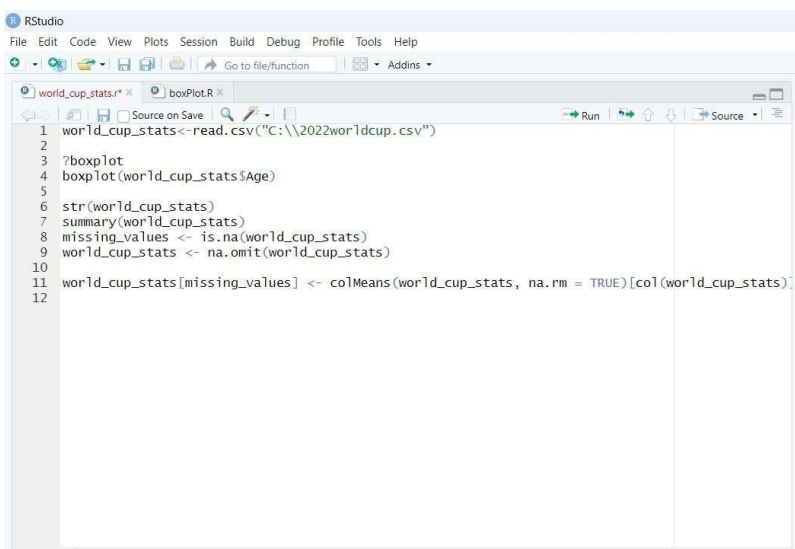
- Use the `summary(world_cup_stats)` function to get summary statistics for each numeric variable (e.g., average, minimum, maximum).

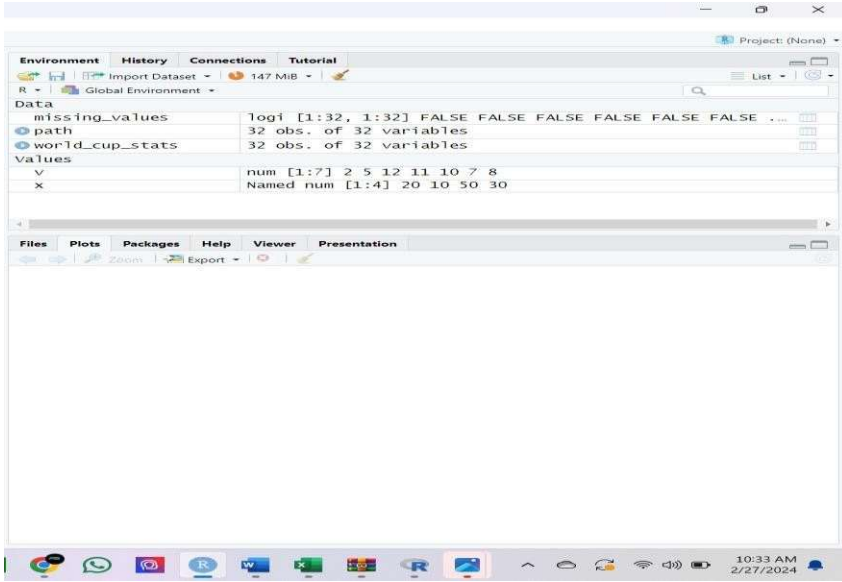
Solution:



- Examine the data for missing values using the `is.na(world_cup_stats)` function. If missing values are present, decide how to handle them (e.g., remove rows, impute values).

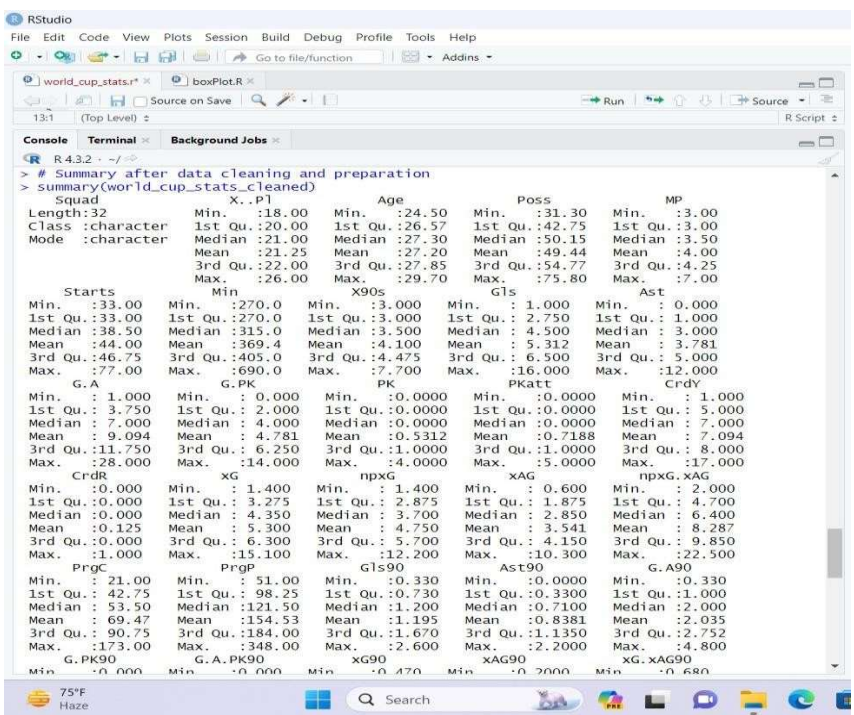
Solution:





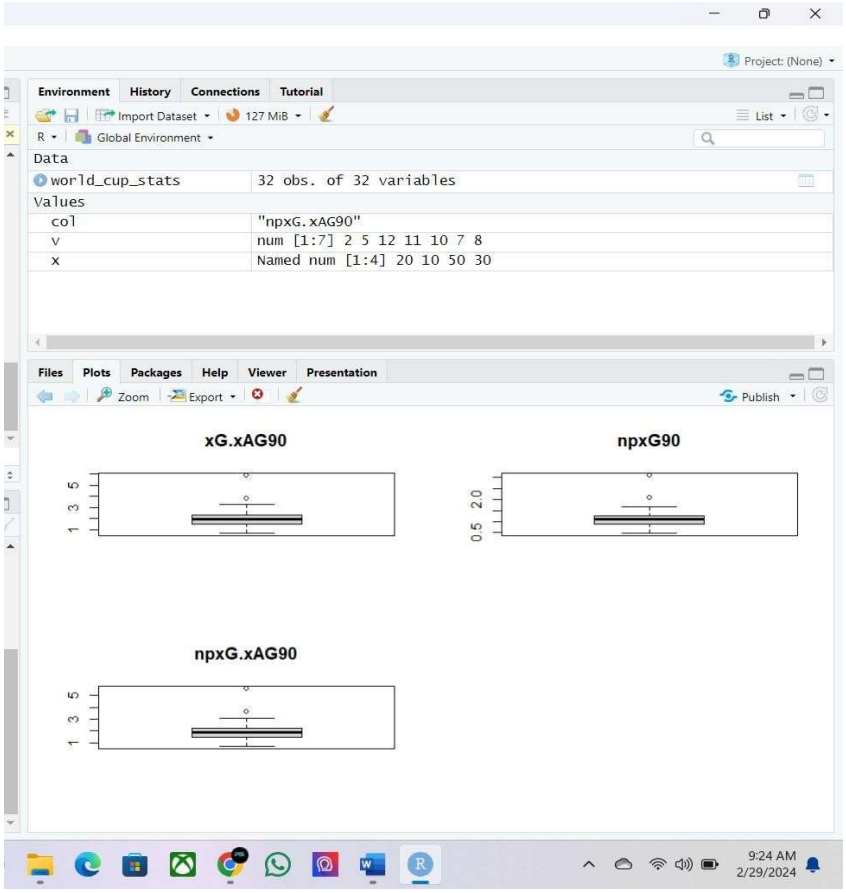
3. Data Cleaning and Preparation:
   - If needed, remove rows with missing values or impute them using appropriate methods.

Solution:



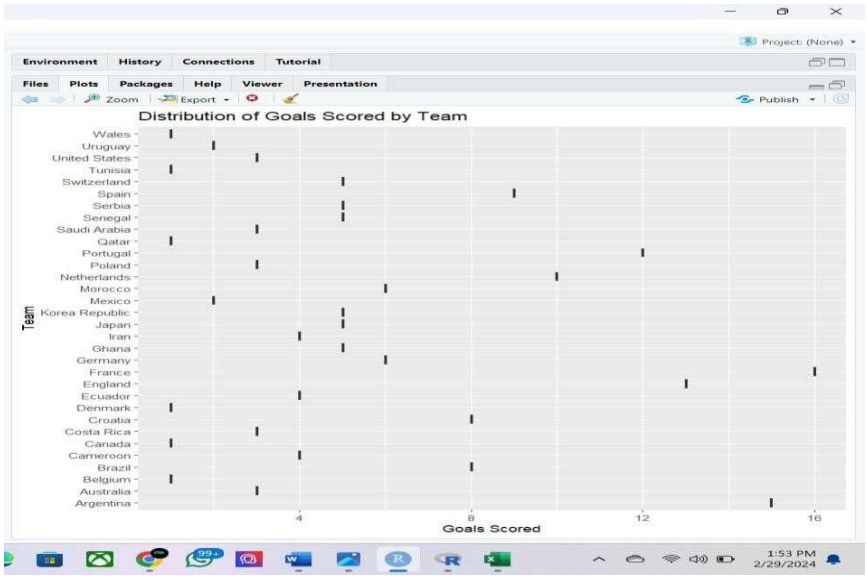
- Check for outliers in the data using boxplots or other methods. Decide how to handle outliers, if necessary.

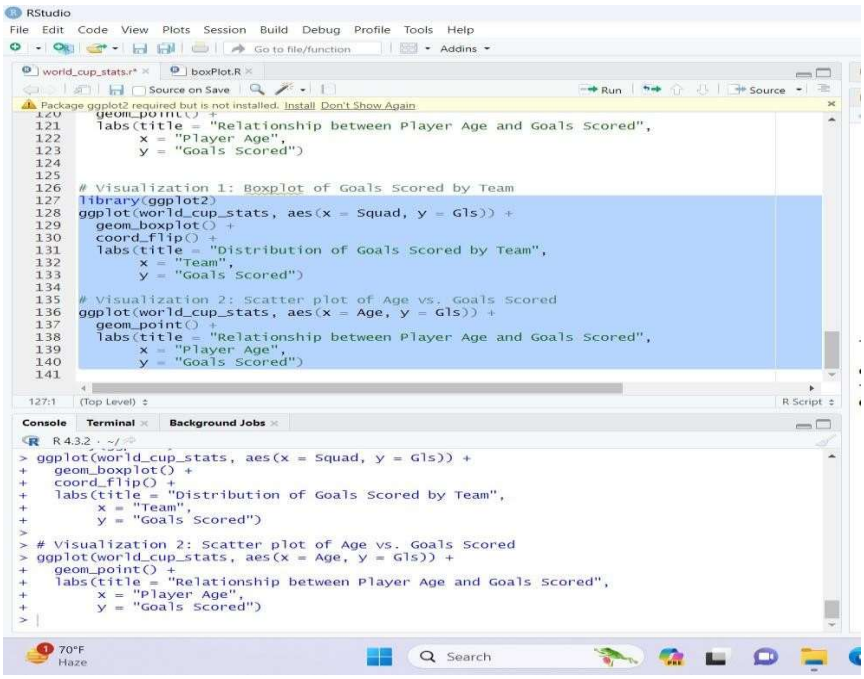
Solution:



4. Data Analysis:

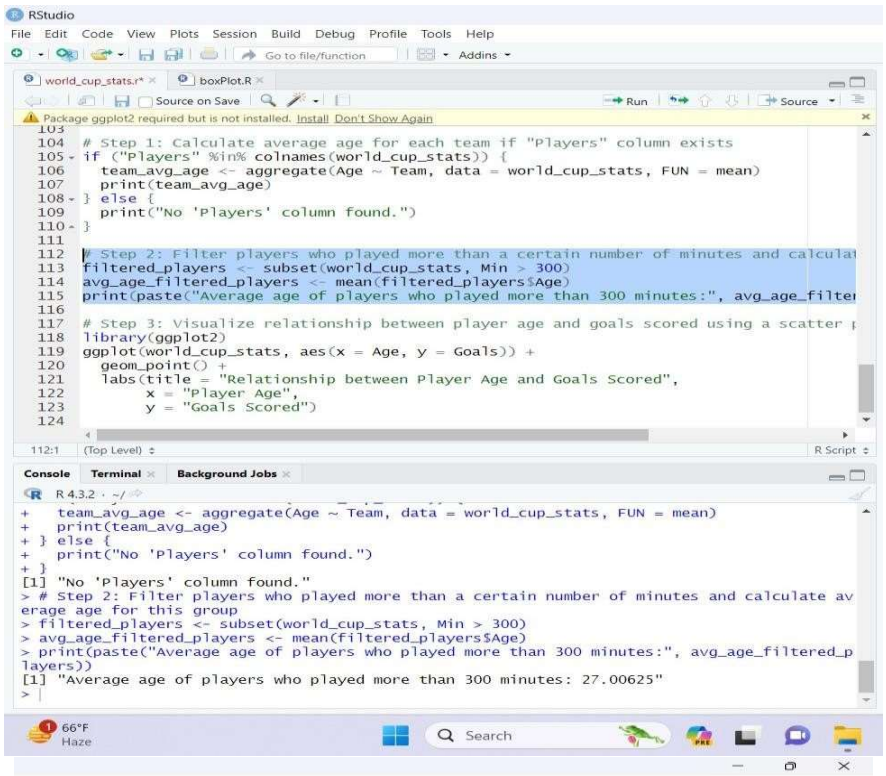
a) Team Analysis: * Calculate and display the average possession, goals scored, and assists for each team. * Identify the top 5 teams with the highest average possession. * Identify the top 3 teams with the highest average number of goals scored per game. * For each team, visualize the distribution of goals scored using a histogram or boxplot.

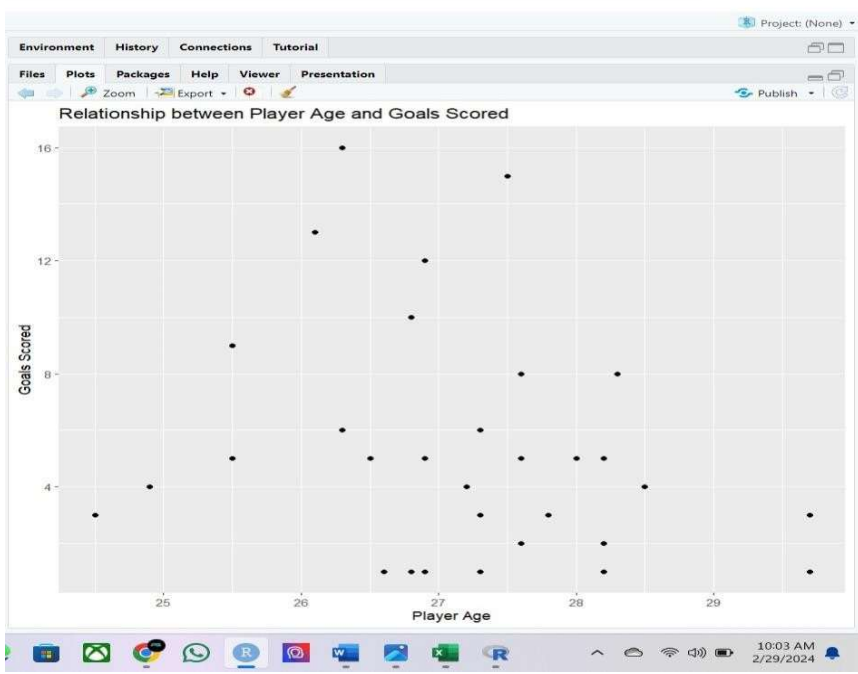



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
world_cup_stats.r*
boxPlot.R
Source on Save
Run
Source
Package ggplot2 required but is not installed. Install Don't Show Again
labs(title = "Relationship between Player Age and Goals Scored",
x = "Player Age",
y = "Goals Scored")
# Visualization 1: Boxplot of Goals Scored by Team
library(ggplot2)
ggplot(world_cup_stats, aes(x = Squad, y = Gls)) +
geom_boxplot() +
coord_flip() +
labs(title = "Distribution of Goals Scored by Team",
x = "Team",
y = "Goals Scored")
# Visualization 2: Scatter plot of Age vs. Goals Scored
ggplot(world_cup_stats, aes(x = Age, y = Gls)) +
geom_point() +
labs(title = "Relationship between Player Age and Goals Scored",
x = "Player Age",
y = "Goals Scored")
127:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.3.2 · ~/
> ggplot(world_cup_stats, aes(x = Squad, y = Gls)) +
+ geom_boxplot() +
+ coord_flip() +
+ labs(title = "Distribution of Goals Scored by Team",
+ x = "Team",
+ y = "Goals Scored")
>
> # Visualization 2: Scatter plot of Age vs. Goals Scored
> ggplot(world_cup_stats, aes(x = Age, y = Gls)) +
+ geom_point() +
+ labs(title = "Relationship between Player Age and Goals Scored",
+ x = "Player Age",
+ y = "Goals Scored")
>
70°F
Haze
Search

b) Player Analysis: * If a "Players" column exists, calculate and display the average age for each team. * Filter players who played more than a certain number of minutes (e.g., 300 minutes) and calculate the average age for this group. * Visualize the relationship between player age and goals scored using a scatter plot.

Solution:





5. Reporting:

(a)Write a summary report outlining your findings and insights from the data analysis.

Solution:

Summary Report: Analysis of World Cup Statistics

Introduction

The analysis was conducted on a dataset containing World Cup statistics. The dataset includes information on teams, players, possessions, goals scored, assists, age, and minutes played. The objective of the analysis was to gain insights into team performance, player demographics, and the relationship between player characteristics and performance metrics.

1. Team Analysis:

Average Possession: The average possession for each team was calculated, and the top 5 teams with the highest average possession were identified.

Average Goals Scored: The average number of goals scored per game for each team was calculated, and the top 3 teams with the highest average goals scored were determined.

Distribution of Goals Scored: The distribution of goals scored for each team was visualized using boxplots.

2. Player Analysis:

- Average Age by Team: If available, the average age for each team was calculated based on player demographics.

- Players with More Than 300 Minutes: Players who played more than 300 minutes were filtered, and the average age of this group was computed.

- Relationship Between Age and Goals Scored: A scatter plot was created to visualize the relationship between player age and goals scored.

Findings and Insights:

Team Performance: The analysis revealed significant variations in possession and goals scored among different teams. Teams with higher possession tended to have higher goal-scoring rates, indicating a potential correlation between possession and offensive effectiveness.

Player Demographics: The average age of players who participated in the World Cup was determined. Additionally, players who played more than 300 minutes were analyzed to understand the demographics of key contributors to team performance.

Relationship Between Age and Goals Scored: The scatter plot showed a diverse distribution of goals scored across different age groups. While there wasn't a clear linear relationship between age and goals scored, further analysis could explore potential trends or patterns.

Conclusion:

The analysis provided valuable insights into team performance and player demographics in the World Cup. The findings can be used to inform strategic decisions for teams and player recruitment strategies. Further analysis could delve deeper into factors influencing team success and player performance, such as playing style, tactics, and individual player characteristics.

Overall, the analysis contributes to a better understanding of the dynamics of the World Cup and provides actionable insights for stakeholders in the soccer community.

- Include relevant visualizations (histograms, boxplots, scatterplots) to support your conclusions.

Solution: