

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

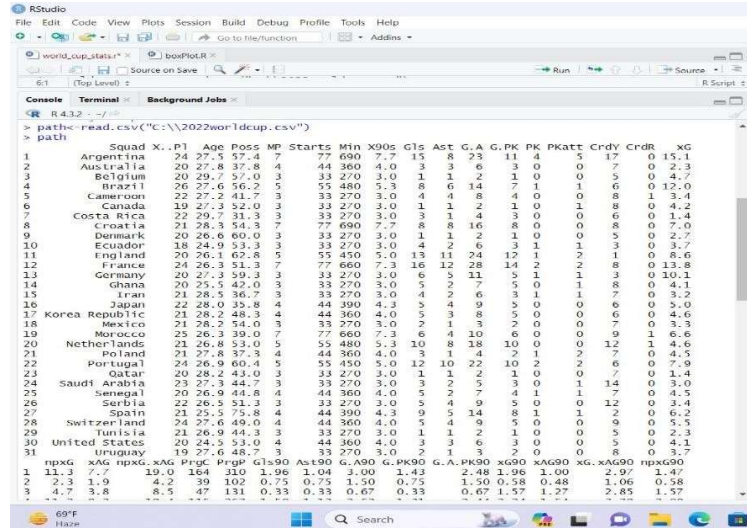
INDEX

S.NO.	TOPICS	PAGE NO.
1.	Load the data.	3
2.	Explore the data.	3
3.	Data cleaning and preparation.	4
4.	Data analysis.	7
5.	Reporting.	10

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:



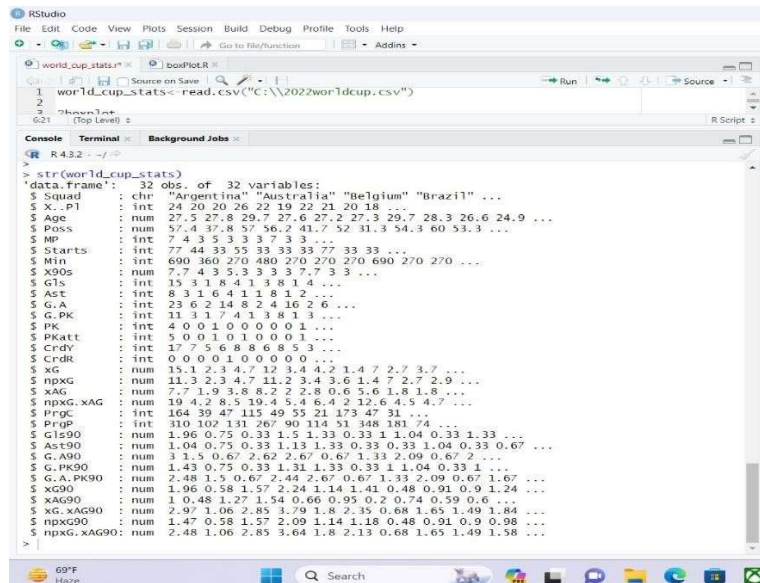
```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> path <- read.csv("C:\\2022worldcup.csv")
> path
```

	Squad	X...	P1	Age	Poss	MP	Starts	Min	X90s	Gls	Ast	G.A	G.PK	PK	PKatt	Crdy	Crdr	xG
1	Argentina	24	27.5	57.4	7	77	690	7.7	15	8	23	11	4	5	17	0	15.1	
2	Australia	20	27.8	37.8	4	44	360	4.0	3	3	6	3	0	0	7	0	2.3	
3	Belgium	20	29.7	57.0	3	33	270	3.0	1	1	2	1	0	0	5	0	4.7	
4	Brazil	26	27.6	56.2	5	55	480	5.3	8	6	14	7	1	1	6	0	12.0	
5	Cameroun	22	27.2	41.7	3	33	270	3.0	4	4	8	4	0	0	8	1	3.4	
6	Canada	19	27.3	52.0	3	33	270	3.0	1	1	2	1	0	1	8	0	4.2	
7	Costa Rica	22	29.7	31.3	3	33	270	3.0	3	1	4	3	0	0	6	0	1.4	
8	Croatia	21	28.3	54.3	7	77	690	7.7	8	8	16	8	0	0	8	0	7.0	
9	Denmark	20	26.6	60.0	3	33	270	3.0	1	1	2	1	0	0	5	0	2.7	
10	Ecuador	18	24.9	53.3	3	33	270	3.0	4	2	6	3	1	1	3	0	3.7	
11	England	20	26.1	62.8	5	55	450	5.0	13	11	24	12	1	2	1	0	8.6	
12	France	24	26.3	51.3	7	77	660	7.3	16	12	28	14	2	2	8	0	13.8	
13	Germany	20	27.3	59.3	3	33	270	3.0	6	5	11	5	1	1	1	0	10.1	
14	Ghana	20	25.5	42.0	3	33	270	3.0	5	2	7	5	0	1	8	0	4.1	
15	Iran	21	28.5	36.7	3	33	270	3.0	4	2	6	3	1	1	7	0	3.2	
16	Japan	22	28.0	35.8	4	44	390	4.3	5	4	9	5	0	0	6	0	5.0	
17	Korea Republic	21	28.2	48.3	4	44	360	4.0	5	3	8	5	0	0	6	0	4.6	
18	Mexico	21	28.2	54.0	3	33	270	3.0	2	1	3	2	0	0	7	0	3.3	
19	Morocco	25	26.3	39.0	7	77	660	7.3	6	4	10	6	0	0	9	1	6.6	
20	Netherlands	21	26.8	53.0	5	55	480	5.3	10	8	18	10	0	0	12	1	4.6	
21	Poland	21	27.8	37.3	4	44	360	4.0	3	1	4	2	1	2	7	0	4.5	
22	Portugal	24	26.9	60.4	5	55	450	5.0	12	10	22	10	2	2	6	0	7.9	
23	Qatar	20	26.2	44.9	3	33	270	3.0	1	1	2	1	0	0	9	0	1.4	
24	Saudi Arabia	23	27.3	44.7	3	33	270	3.0	3	2	5	3	0	0	14	0	3.0	
25	Senegal	20	26.9	44.8	4	44	360	4.0	5	2	7	4	1	1	7	0	4.5	
26	Serbia	22	26.5	51.3	3	33	270	3.0	5	4	9	5	0	0	12	0	3.4	
27	Spain	21	25.5	75.8	4	44	390	4.3	9	5	14	8	1	1	2	0	6.2	
28	Switzerland	24	27.7	49.0	4	44	360	4.0	5	4	9	5	0	0	9	0	5.5	
29	Tunisia	21	26.9	44.3	3	33	270	3.0	1	1	2	1	0	0	5	0	2.3	
30	United States	20	24.5	53.0	4	44	360	4.0	3	3	6	3	0	0	5	0	4.1	
31	Uruguay	19	27.6	48.7	3	33	270	3.0	2	1	3	2	0	0	8	0	3.7	
npkg	xAG	npkg.xAG	PrpG	PrpG	Gls90	Ast90	G.A90	G.PK90	G.A.PK90	xG90	xAG90	xG.xAG90	npkg90					
1	11.3	7.7	19.0	164	310	1.96	1.04	3.00	1.43	2.48	1.96	1.00	2.97					
2	4.2	1.9	4.2	39	102	0.75	0.75	1.50	0.75	1.50	0.58	0.48	1.06					
3	4.7	3.8	8.5	47	131	0.33	0.33	0.67	0.33	0.67	1.57	1.27	2.85					

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.

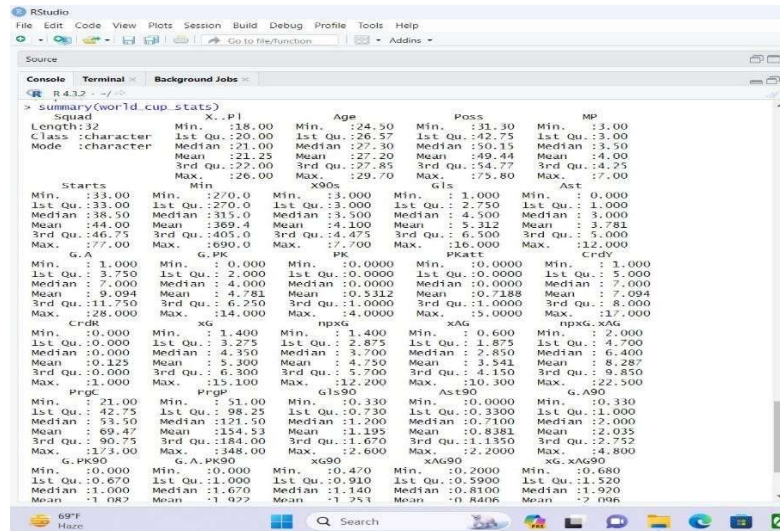
Solution:



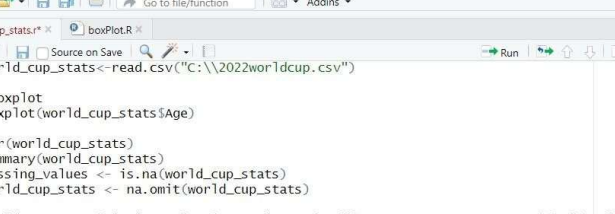
```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> world_cup_stats <- read.csv("C:\\2022worldcup.csv")
> str(world_cup_stats)
```

```
data.frame': 32 obs. of 32 variables:
 $ Squad      : chr "Argentina" "Australia" "Belgium" "Brazil" ...
 $ X...P1     : int 24 20 26 22 19 22 21 20 18 ...
 $ Age        : num 27.5 27.8 29.7 27.6 27.2 27.3 29.7 28.3 26.6 24.9 ...
 $ Poss       : num 57.4 37.8 57.5 56.2 41.7 52.3 31.3 54.3 60.5 53.3 ...
 $ MP         : int 77 44 33 55 33 33 33 77 33 33 ...
 $ Starts     : int 690 360 270 480 270 270 690 270 270 ...
 $ Min        : num 7.7 4.3 5.3 3.3 3.7 3.3 7.7 3.3 ...
 $ X90s       : int 15 3 1 8 4 1 3 8 1 4 ...
 $ GlS        : int 8 3 1 6 4 1 1 8 1 2 ...
 $ Ast        : int 23 6 2 14 8 2 4 16 2 6 ...
 $ G.A        : int 11 3 1 7 4 1 3 8 1 3 ...
 $ G.PK       : int 4 0 0 1 0 0 0 0 1 ...
 $ PKatt      : int 5 0 0 1 0 1 0 0 0 1 ...
 $ Crdy       : int 17 5 5 6 8 8 6 8 5 3 ...
 $ Crdr       : int 0 0 0 1 0 0 0 0 0 ...
 $ xG         : num 15.1 2.3 4.7 12.3 4.4 2.1 4.7 2.7 3.7 ...
 $ npkg       : num 11.3 2.3 4.7 11.2 3.4 3.6 1.4 7.2 2.9 ...
 $ xAG        : num 7.7 1.9 3.8 8.2 2.8 0.6 5.6 1.8 1.8 ...
 $ npkg.xAG   : num 19.4 2.8 5.19 4.5 4.6 4.2 12.6 4.5 4.7 ...
 $ PrpG       : int 164 39 47 115 49 55 21 173 31 ...
 $ PrpG90     : int 310 102 131 267 90 114 51 348 181 74 ...
 $ GlS90      : num 1.96 0.75 0.33 1.5 1.33 0.33 1.04 0.33 1.33 ...
 $ Ast90      : num 1.04 0.75 0.33 1.13 1.33 0.33 1.04 0.33 0.67 ...
 $ G.A90      : num 3 1.5 0.67 2.62 2.67 0.67 1.33 2.09 0.67 ...
 $ G.PK90     : num 1.43 0.75 0.33 1.31 1.33 0.33 1.04 0.33 1 ...
 $ G.A.PK90   : num 2.48 1.5 0.67 2.44 2.67 0.67 1.33 2.09 0.67 ...
 $ xG90       : num 1.96 0.58 1.57 2.24 1.14 1.41 0.48 0.91 0.9 1.24 ...
 $ xAG90      : num 1 0.48 1.27 1.54 0.66 0.95 0.2 0.74 0.59 0.6 ...
 $ xG.xAG90   : num 2.97 1.06 2.85 3.79 1.8 2.35 0.68 1.65 1.49 1.84 ...
 $ npkg90     : num 1.47 0.58 1.57 2.09 1.14 1.18 0.48 0.91 0.9 0.98 ...
 $ npkg.xAG90: num 2.48 1.06 2.85 3.64 1.8 2.13 0.68 1.65 1.49 1.58 ...
```

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



- 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).



The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for saving, opening files, and navigating. The main editor window has a tab titled 'boxPlot.R' and contains the following R code:

```

1 world_cup_stats<-read.csv("C:\\2022worldcup.csv")
2
3 ?boxplot
4 boxplot(world_cup_stats$Age)
5
6 str(world_cup_stats)
7 summary(world_cup_stats)
8 missing_values <- is.na(world_cup_stats)
9 world_cup_stats <- na.omit(world_cup_stats)
10
11 world_cup_stats[missing_values] <- colMeans(world_cup_stats, na.rm = TRUE)[col(world_cup_stats)]
12

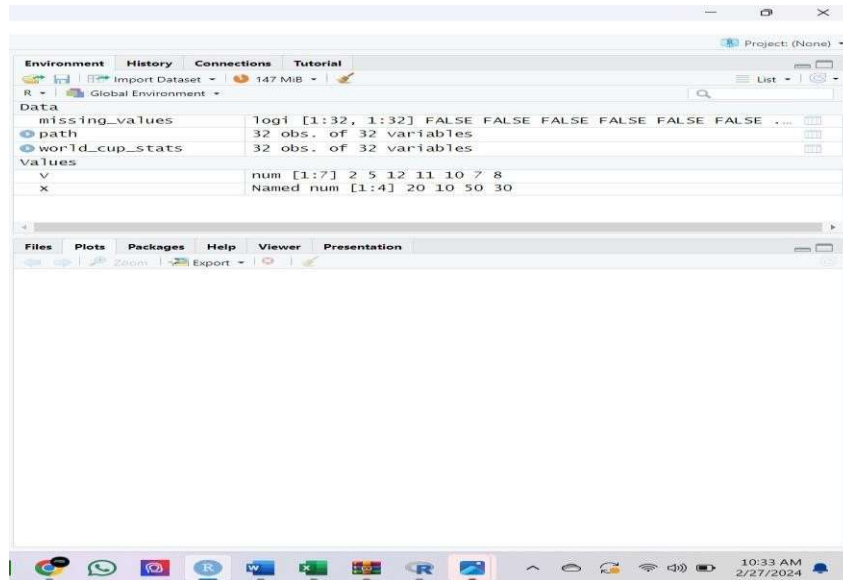
```

The right-hand pane is empty, and the status bar at the bottom shows '0 rows of source'. The code in the script editor is as follows:

```

1 world_cup_stats<-read.csv("C:\\2022worldcup.csv")
2
3 ?boxplot
4 boxplot(world_cup_stats$Age)
5
6 str(world_cup_stats)
7 summary(world_cup_stats)
8 missing_values <- is.na(world_cup_stats)
9 world_cup_stats <- na.omit(world_cup_stats)
10
11 world_cup_stats[missing_values] <- colMeans(world_cup_stats, na.rm = TRUE)[col(world_cup_stats)]
12

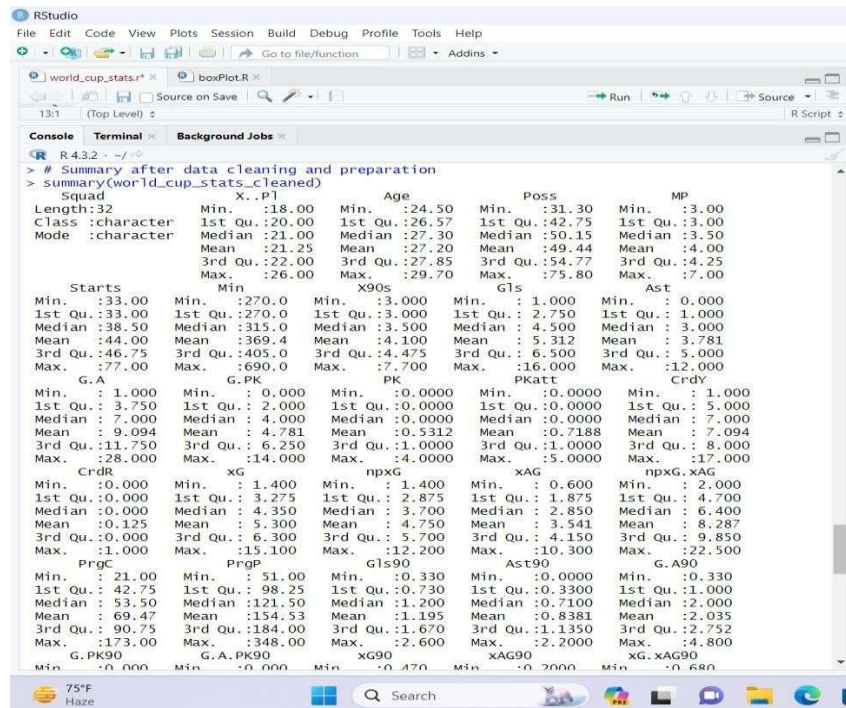
```



3. Data Cleaning and Preparation:

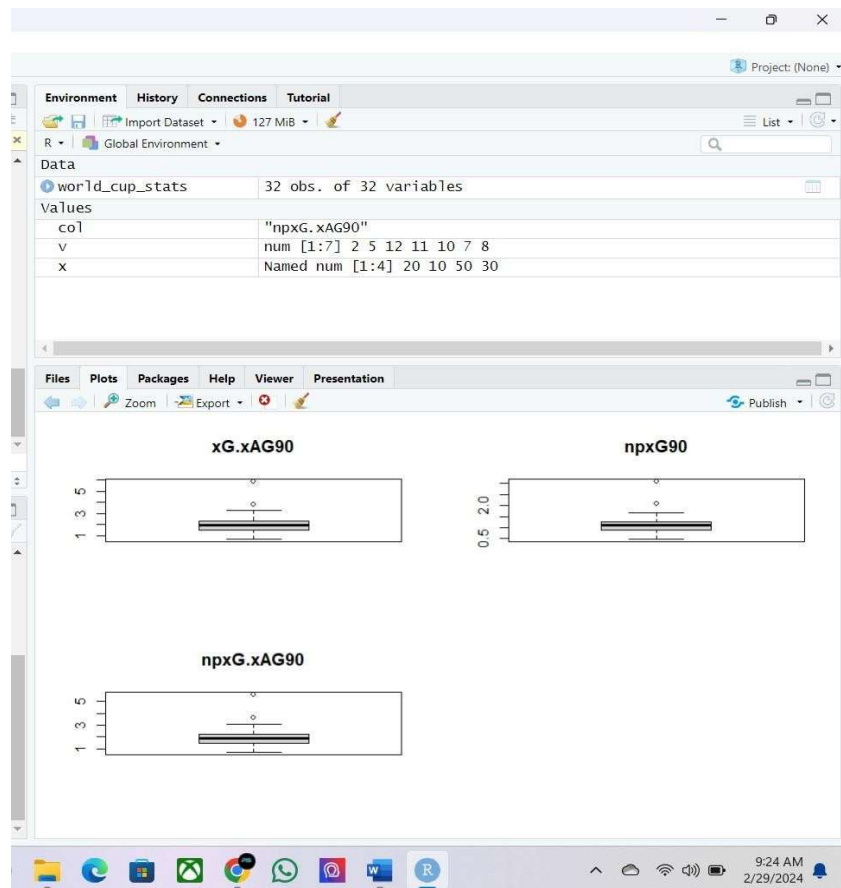
- If needed, remove rows with missing values or impute them using appropriate methods.

Solution:



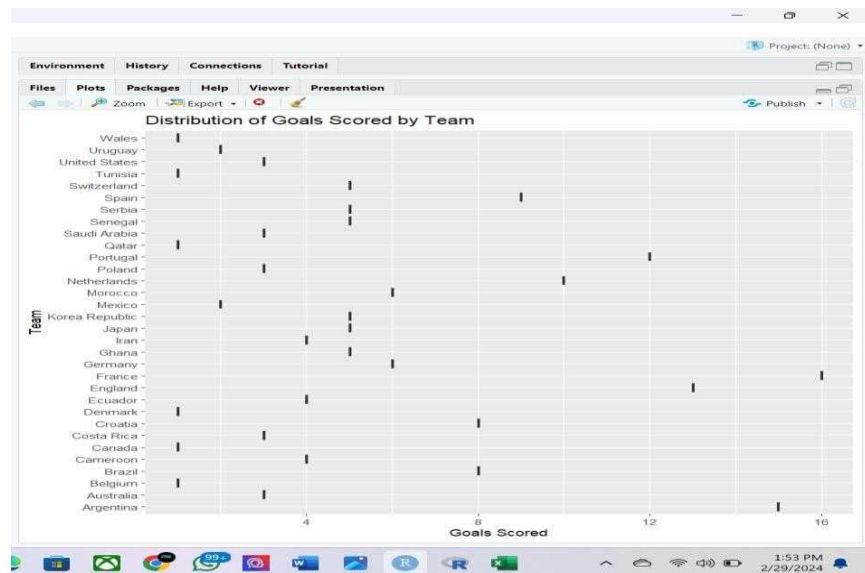
- Check for outliers in the data using boxplots or other methods. Decide how to handle outliers, if necessary.

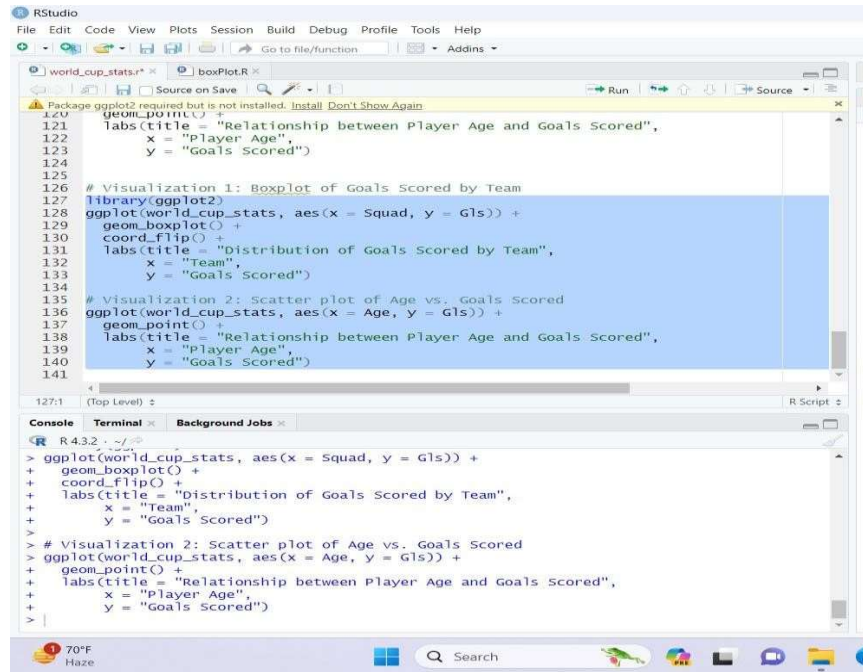
Solution:



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.





The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for data visualization. Lines 121-124 define a scatter plot. Lines 126-134 define a boxplot. Lines 136-140 define another scatter plot. A warning message at the top states: "Package ggplot2 required but is not installed. Install Don't Show Again".
- Console:** Shows the execution of the code from the source editor, including the loading of the ggplot2 library and the creation of the plots.
- Terminal:** Is currently empty.
- Background Jobs:** Is currently empty.
- Taskbar:** At the bottom, it shows the system clock as 70°F and Haze, along with various application icons.

```
121   labs(title = "Relationship between Player Age and Goals Scored",
122         x = "Player Age",
123         y = "Goals Scored")
124
125
126 # Visualization 1: Boxplot of Goals Scored by Team
127 library(ggplot2)
128 ggplot(world_cup_stats, aes(x = Squad, y = Gls)) +
129   geom_boxplot() +
130   coord_flip() +
131   labs(title = "Distribution of Goals Scored by Team",
132         x = "Team",
133         y = "Goals Scored")
134
135 # Visualization 2: Scatter plot of Age vs. Goals Scored
136 ggplot(world_cup_stats, aes(x = Age, y = Gls)) +
137   geom_point() +
138   labs(title = "Relationship between Player Age and Goals Scored",
139         x = "Player Age",
140         y = "Goals Scored")
141
```

Console output:

```
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")
>
```


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:

```

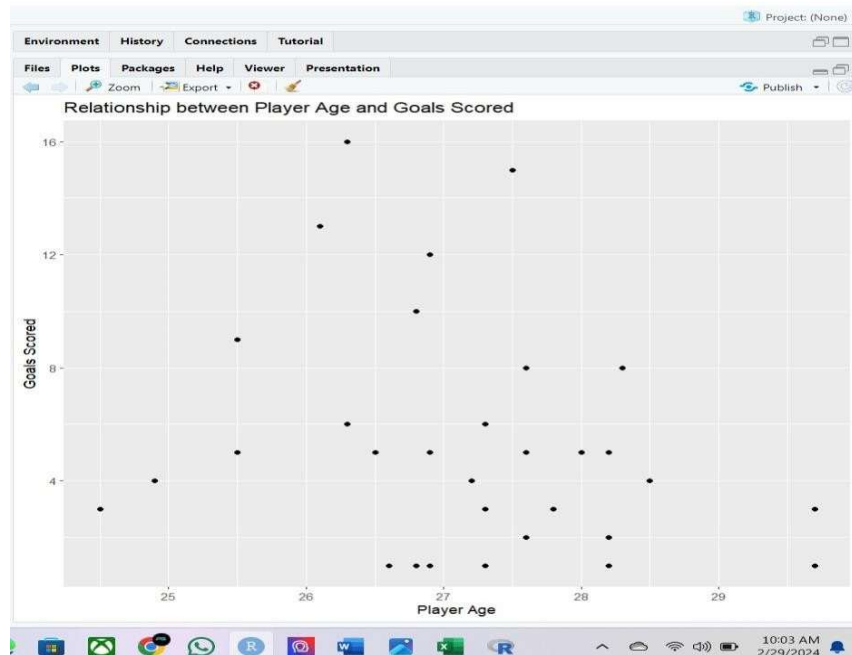
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
[Icons] Go to file/function Addins
[Warning] Package ggplot2 required but is not installed. Install Don't Show Again
103 # Step 1: Calculate average age for each team if "Players" column exists
104
105 if ("Players" %in% colnames(world_cup_stats)) {
106   team_avg_age <- aggregate(Age ~ Team, data = world_cup_stats, FUN = mean)
107   print(team_avg_age)
108 } else {
109   print("No 'Players' column found.")
110 }
111
112 # Step 2: Filter players who played more than a certain number of minutes and calculate
113 filtered_players <- subset(world_cup_stats, Min > 300)
114 avg_age_filtered_players <- mean(filtered_players$Age)
115 print(paste("Average age of players who played more than 300 minutes:", avg_age_filtered_players))
116
117 # Step 3: Visualize relationship between player age and goals scored using a scatter plot
118 library(ggplot2)
119 ggplot(world_cup_stats, aes(x = Age, y = Goals)) +
120   geom_point() +
121   labs(title = "Relationship between Player Age and Goals Scored",
122        x = "Player Age",
123        y = "Goals Scored")
124
112:1 (Top Level) R Script

```

```

R 4.3.2 ~ /
+ team_avg_age <- aggregate(Age ~ Team, data = world_cup_stats, FUN = mean)
+ print(team_avg_age)
+ } else {
+   print("No 'Players' column found.")
+ }
[1] "No 'Players' column found."
> # Step 2: Filter players who played more than a certain number of minutes and calculate average age for this group
> filtered_players <- subset(world_cup_stats, Min > 300)
> avg_age_filtered_players <- mean(filtered_players$Age)
> print(paste("Average age of players who played more than 300 minutes:", avg_age_filtered_players))
[1] "Average age of players who played more than 300 minutes: 27.00625"
>

```



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.

- Solution:**

