AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

Course Title: **Introduction to Data Science**	
Project No: **Final Term Project**	Date of Submission: **02/05/2023**
Project Title: **Web Scrapping**	
Course Code: **CSC4180**	Section: **“ C ”**
Semester: **Spring** 22-23	Course Teacher: **Dr. Akinul Islam Jony**

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student’s work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison,including review by external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea or creation of another person as though it is yourown. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copymy/our work.

* *Student(s) must complete all details except the faculty use part.*

No		ID	Program	Signature
1	Shuchi Brata Roy	20-43313-1	BSc in CSE	
2	Ahana Bhuiyan	20-43412-1	BSc in CSE	
3	Pulok Sarkar	19-41641-3	BSc in CSE	
4	Raihan Sikder Anik	20-43166-1	BSc in CSE	

Faculty use only

FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project Overview:

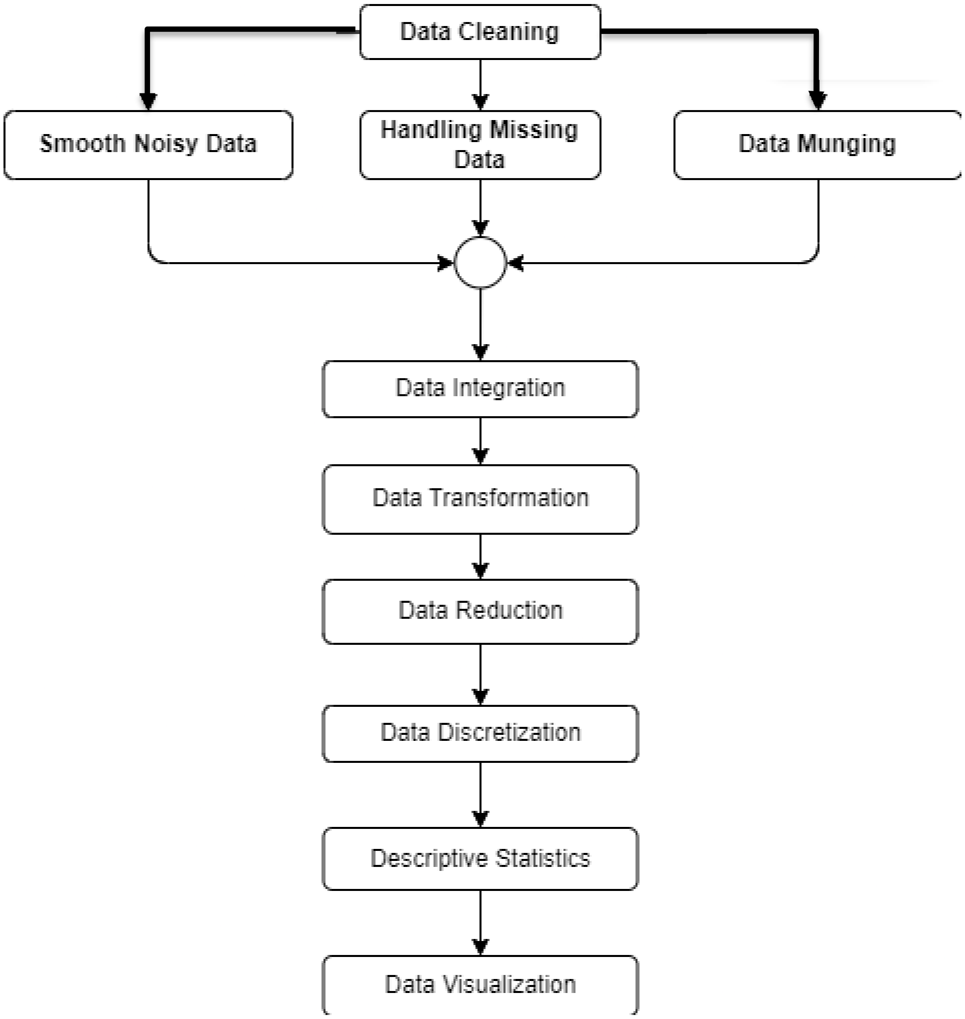
For this project, we have been assigned to scrap data from webpages, perform preprocessing techniques on them, describe them in the light of descriptive statistics and visualize them using R language.

Specifically, we focused to work with footballer data from Argentina squad for the FIFA World Cup 2022 season. We collected the data and examine it properly. We analyzed the data by comparing the performance of different players and why the players were successful. Real-world data is often incomplete, noisy, and inconsistent, so we performed data pre-processing tasks like data cleaning, integration, transformation, reduction, and discretization. Also Descriptive analysis was used to describe the data using methods such as mean, median, range, variance, quartile, and percentile. Lastly, we used data visualization techniques to present our findings in a more digestible and impactful way. By visualizing different aspects of the data, we can deliver insights more effectively and help readers understand the comparison and relationship between different variables.

Project Solution:

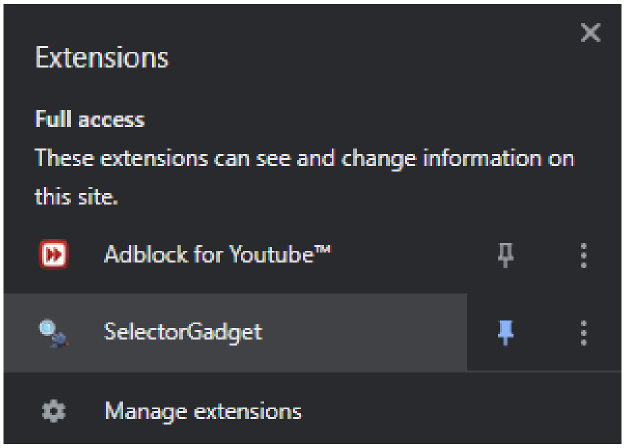
We gathered player information and performance data for Argentina from ESPN websites to prepare the dataset for analysis. The collected data was then stored in a CSV file. Data pre-processing involved inspecting the raw dataset to identify and eliminate errors, duplication, and redundant data. We also addressed if there is any missing or noisy data than it will be replaced with N/A and filling it with the median value. Additionally, measures such as data integration, data transformation, data reduction, and data discretization were implemented to further refine the data set. We utilized descriptive statistics, including mean, median, mode, range, variance, standard deviation, quartiles, percentiles, and interquartile ranges, to simplify the data and summarize its characteristics. To visually represent the data and facts, we employed data visualization technique

Flowchart of the Data Preprocessing -



Data Collection:

For this project, we start to scrap the data from the website. First, we start to scrap the data from Argentina Squad. In this process, we use a selector gadget to simply select data on a website and it will determine its HTML/CSS tags, ids and classes.

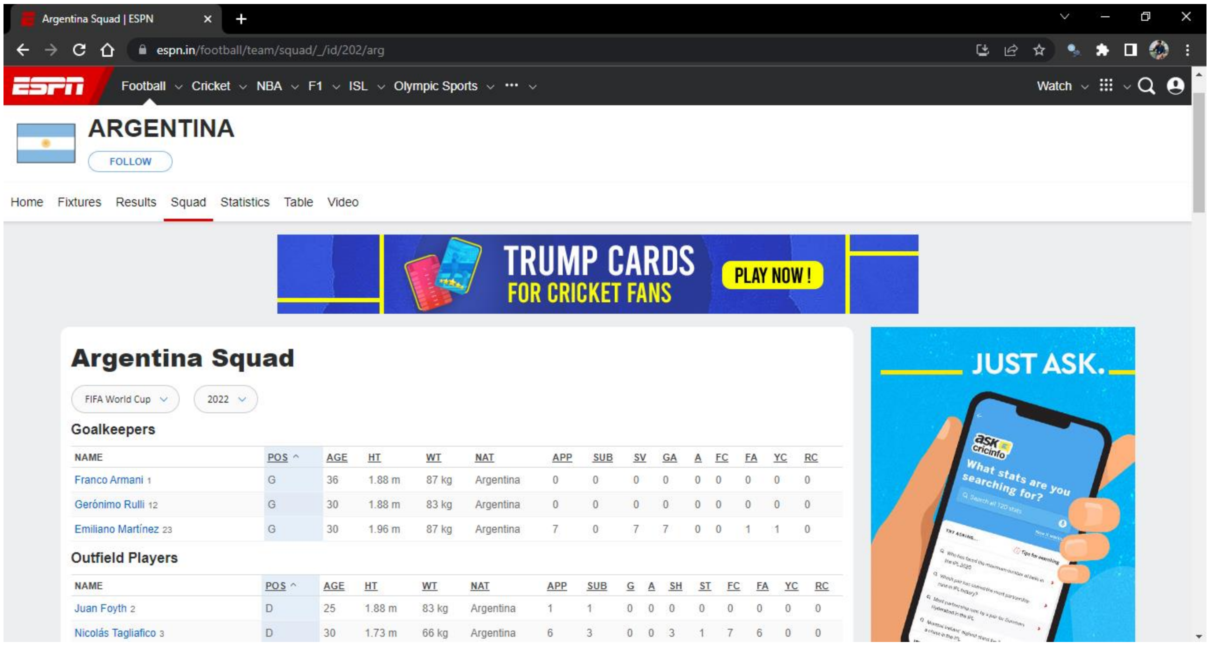


Extensions

Full access

These extensions can see and change information on this site.

Obtaining information of Argentina -



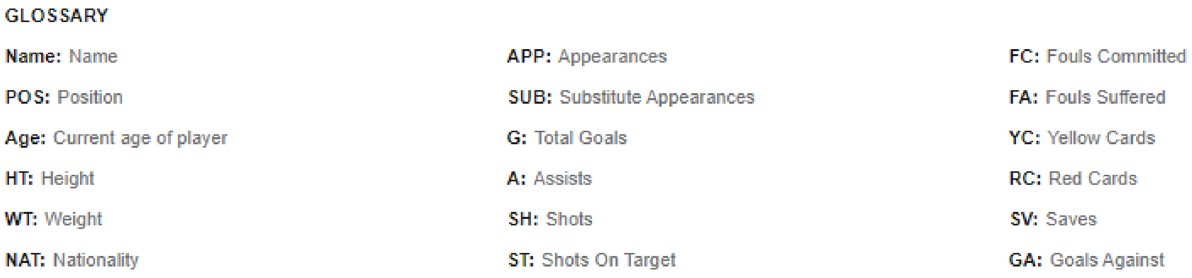
Argentina Squad

Goalkeepers

NAME	POS ^	AGE	HT	WT	NAT	APP	SUB	SV	GA	A	FC	FA	YC	RC
Franco Armani 1	G	36	1.88 m	87 kg	Argentina	0	0	0	0	0	0	0	0	0
Gerónimo Rulli 12	G	30	1.88 m	83 kg	Argentina	0	0	0	0	0	0	0	0	0
Emiliano Martínez 23	G	30	1.96 m	87 kg	Argentina	7	0	7	7	0	0	1	1	0

Outfield Players

NAME	POS ^	AGE	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
Juan Foyth 2	D	25	1.88 m	83 kg	Argentina	1	1	0	0	0	0	0	0	0	0
Nicolás Tagliafico 3	D	30	1.73 m	66 kg	Argentina	6	3	0	0	3	1	7	6	0	0



GLOSSARY

Name: Name

POS: Position

Age: Current age of player

HT: Height

WT: Weight

NAT: Nationality

APP: Appearances

SUB: Substitute Appearances

G: Total Goals

A: Assists

SH: Shots

ST: Shots On Target

FC: Fouls Committed

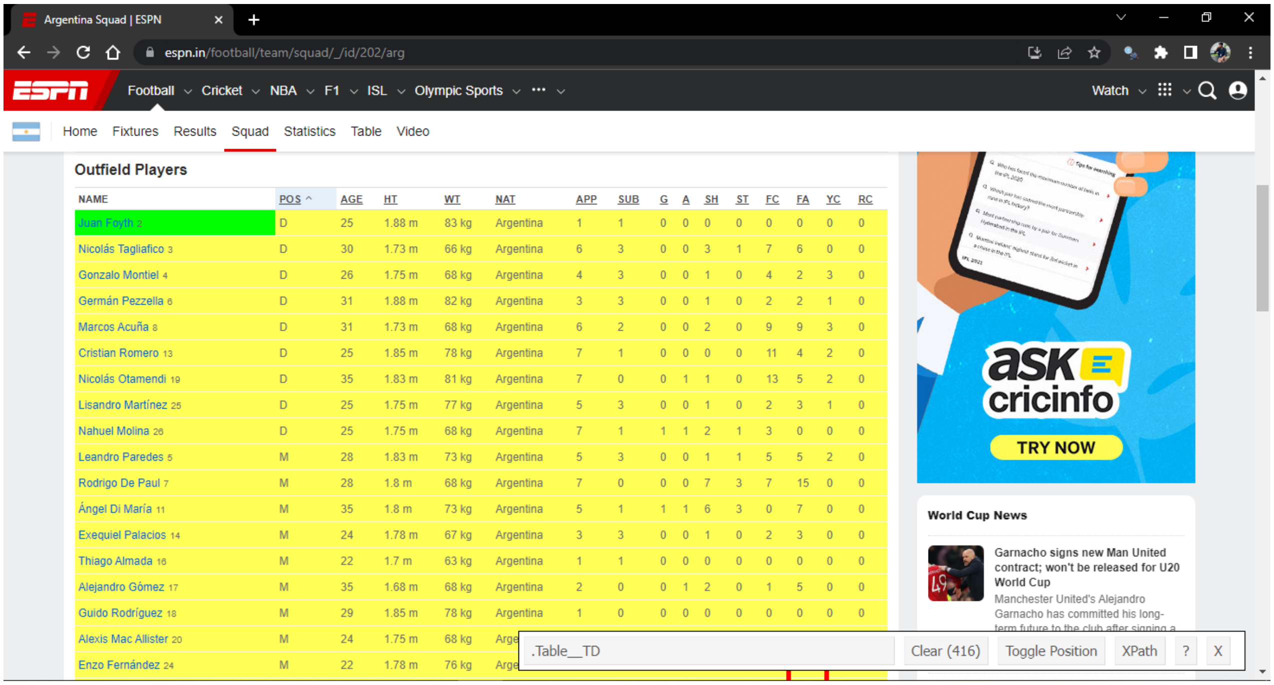
FA: Fouls Suffered

YC: Yellow Cards

RC: Red Cards

SV: Saves

GA: Goals Against


Argentina Squad | ESPN
espn.in/football/team/squad/_/id/202/arg
Football
Cricket
NBA
F1
ISL
Olympic Sports
Watch
Home
Fixtures
Results
Squad
Statistics
Table
Video
ask cricinfo
TRY NOW
World Cup News
Garnacho signs new Man United contract; won't be released for U20 World Cup
Manchester United's Alejandro Garnacho has committed his long-term future to the club after signing a
.Table__TD
Clear (416)
Toggle Position
XPath
?
X

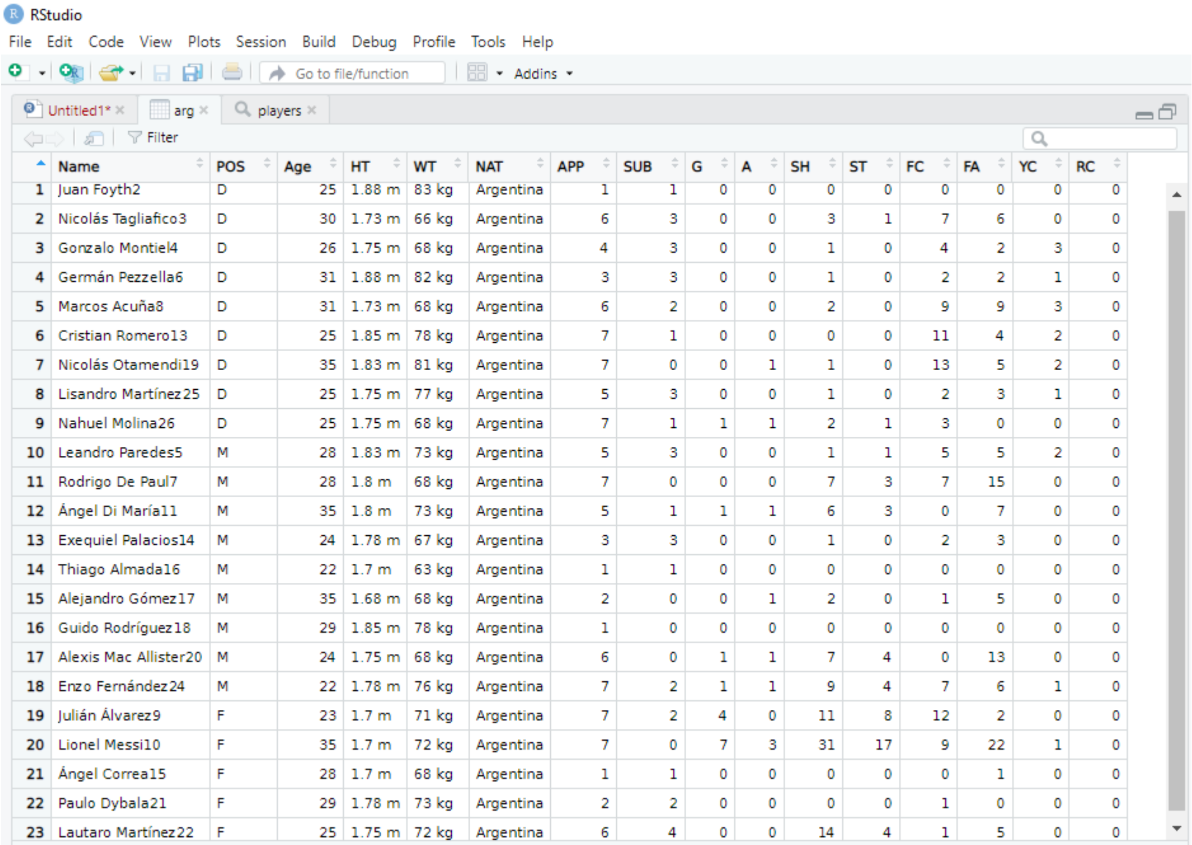
Outfield Players

NAME	POS ^	AGE	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
Juan Foyth 2	D	25	1.88 m	83 kg	Argentina	1	1	0	0	0	0	0	0	0	0
Nicolás Tagliafico 3	D	30	1.73 m	66 kg	Argentina	6	3	0	0	3	1	7	6	0	0
Gonzalo Montiel 4	D	26	1.75 m	68 kg	Argentina	4	3	0	0	1	0	4	2	3	0
Germán Pezzella 6	D	31	1.88 m	82 kg	Argentina	3	3	0	0	1	0	2	2	1	0
Marcos Acuña 8	D	31	1.73 m	68 kg	Argentina	6	2	0	0	2	0	9	9	3	0
Cristian Romero 13	D	25	1.85 m	78 kg	Argentina	7	1	0	0	0	0	11	4	2	0
Nicolás Otamendi 19	D	35	1.83 m	81 kg	Argentina	7	0	0	1	1	0	13	5	2	0
Lisandro Martínez 25	D	25	1.75 m	77 kg	Argentina	5	3	0	0	1	0	2	3	1	0
Nahuel Molina 26	D	25	1.75 m	68 kg	Argentina	7	1	1	1	2	1	3	0	0	0
Leandro Paredes 5	M	28	1.83 m	73 kg	Argentina	5	3	0	0	1	1	5	5	2	0
Rodrigo De Paul 7	M	28	1.8 m	68 kg	Argentina	7	0	0	0	7	3	7	15	0	0
Ángel Di María 11	M	35	1.8 m	73 kg	Argentina	5	1	1	1	6	3	0	7	0	0
Exequiel Palacios 14	M	24	1.78 m	67 kg	Argentina	3	3	0	0	1	0	2	3	0	0
Thiago Almada 16	M	22	1.7 m	63 kg	Argentina	1	1	0	0	0	0	0	0	0	0
Alejandro Gómez 17	M	35	1.68 m	68 kg	Argentina	2	0	0	1	2	0	1	5	0	0
Guido Rodríguez 18	M	29	1.85 m	78 kg	Argentina	1	0	0	0	0	0	0	0	0	0
Alexis Mac Allister 20	M	24	1.75 m	68 kg	Arge										
Enzo Fernández 24	M	22	1.78 m	76 kg	Arge										

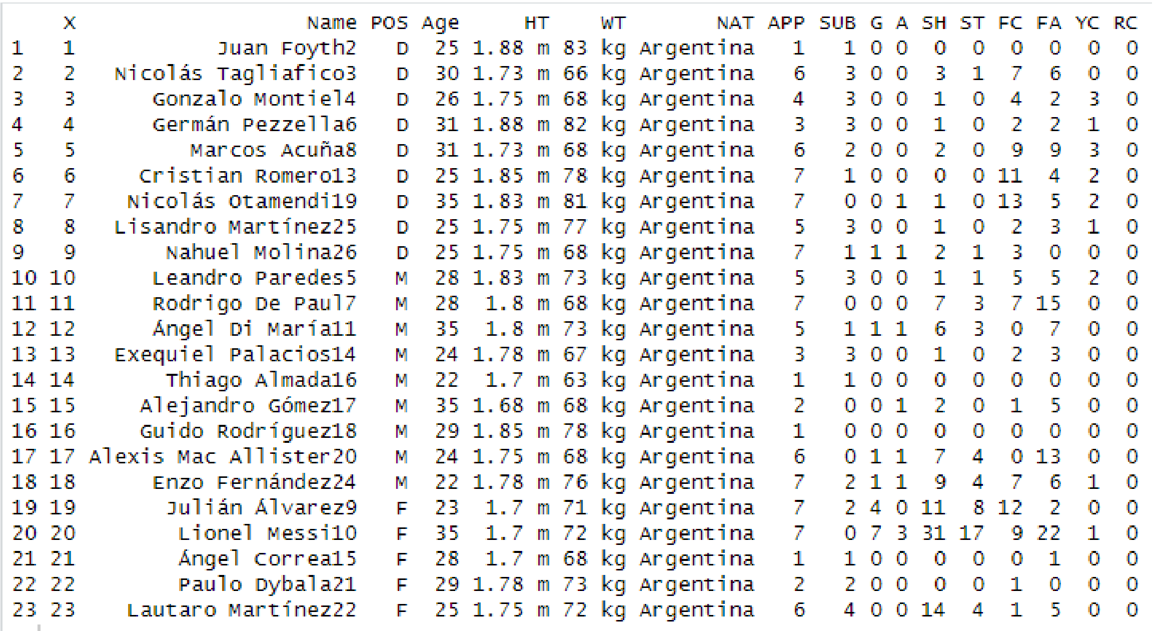
Code –

```
library(rvest)

players = read_html("https://www.espn.in/football/team/squad/_/id/202/arg")

pl = html_nodes(players, css=".Table__TD")

pl

arg <-data.frame(html_table(players, header = TRUE)[[2]])

View(arg)

write.csv(arg,"F:\\arg.csv")

dataset<- read.csv('arg.csv')

dataset
```

Output -


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
arg
players
Filter

	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
1	Juan Foyth2	D	25	1.88 m	83 kg	Argentina	1	1	0	0	0	0	0	0	0	0
2	Nicolás Tagliafico3	D	30	1.73 m	66 kg	Argentina	6	3	0	0	3	1	7	6	0	0
3	Gonzalo Montiel4	D	26	1.75 m	68 kg	Argentina	4	3	0	0	1	0	4	2	3	0
4	Germán Pezzella6	D	31	1.88 m	82 kg	Argentina	3	3	0	0	1	0	2	2	1	0
5	Marcos Acuña8	D	31	1.73 m	68 kg	Argentina	6	2	0	0	2	0	9	9	3	0
6	Cristian Romero13	D	25	1.85 m	78 kg	Argentina	7	1	0	0	0	0	11	4	2	0
7	Nicolás Otamendi19	D	35	1.83 m	81 kg	Argentina	7	0	0	1	1	0	13	5	2	0
8	Lisandro Martínez25	D	25	1.75 m	77 kg	Argentina	5	3	0	0	1	0	2	3	1	0
9	Nahuel Molina26	D	25	1.75 m	68 kg	Argentina	7	1	1	1	2	1	3	0	0	0
10	Leandro Paredes5	M	28	1.83 m	73 kg	Argentina	5	3	0	0	1	1	5	5	2	0
11	Rodrigo De Paul7	M	28	1.8 m	68 kg	Argentina	7	0	0	0	7	3	7	15	0	0
12	Ángel Di María11	M	35	1.8 m	73 kg	Argentina	5	1	1	1	6	3	0	7	0	0
13	Exequiel Palacios14	M	24	1.78 m	67 kg	Argentina	3	3	0	0	1	0	2	3	0	0
14	Thiago Almada16	M	22	1.7 m	63 kg	Argentina	1	1	0	0	0	0	0	0	0	0
15	Alejandro Gómez17	M	35	1.68 m	68 kg	Argentina	2	0	0	1	2	0	1	5	0	0
16	Guido Rodríguez18	M	29	1.85 m	78 kg	Argentina	1	0	0	0	0	0	0	0	0	0
17	Alexis Mac Allister20	M	24	1.75 m	68 kg	Argentina	6	0	1	1	7	4	0	13	0	0
18	Enzo Fernández24	M	22	1.78 m	76 kg	Argentina	7	2	1	1	9	4	7	6	1	0
19	Julián Álvarez9	F	23	1.7 m	71 kg	Argentina	7	2	4	0	11	8	12	2	0	0
20	Lionel Messi10	F	35	1.7 m	72 kg	Argentina	7	0	7	3	31	17	9	22	1	0
21	Ángel Correa15	F	28	1.7 m	68 kg	Argentina	1	1	0	0	0	0	0	1	0	0
22	Paulo Dybala21	F	29	1.78 m	73 kg	Argentina	2	2	0	0	0	0	1	0	0	0
23	Lautaro Martínez22	F	25	1.75 m	72 kg	Argentina	6	4	0	0	14	4	1	5	0	0



```
    X                  Name POS Age    HT    WT       NAT APP SUB G A SH ST FC FA YC RC
1   1           Juan Foyth2   D  25 1.88 m 83 kg Argentina   1   1 0 0  0  0  0  0  0  0
2   2   Nicolás Tagliafico3   D  30 1.73 m 66 kg Argentina   6   3 0 0  3  1  7  6  0  0
3   3      Gonzalo Montiel4   D  26 1.75 m 68 kg Argentina   4   3 0 0  1  0  4  2  3  0
4   4      Germán Pezzella6   D  31 1.88 m 82 kg Argentina   3   3 0 0  1  0  2  2  1  0
5   5         Marcos Acuña8   D  31 1.73 m 68 kg Argentina   6   2 0 0  2  0  9  9  3  0
6   6     Cristian Romero13   D  25 1.85 m 78 kg Argentina   7   1 0 0  0  0 11  4  2  0
7   7    Nicolás Otamendi19   D  35 1.83 m 81 kg Argentina   7   0 0 1  1  0 13  5  2  0
8   8   Lisandro Martínez25   D  25 1.75 m 77 kg Argentina   5   3 0 0  1  0  2  3  1  0
9   9       Nahuel Molina26   D  25 1.75 m 68 kg Argentina   7   1 1 1  2  1  3  0  0  0
10 10      Leandro Paredes5   M  28 1.83 m 73 kg Argentina   5   3 0 0  1  1  5  5  2  0
11 11      Rodrigo De Paul7   M  28  1.8 m 68 kg Argentina   7   0 0 0  7  3  7 15  0  0
12 12      Ángel Di María11   M  35  1.8 m 73 kg Argentina   5   1 1 1  6  3  0  7  0  0
13 13   Exequiel Palacios14   M  24 1.78 m 67 kg Argentina   3   3 0 0  1  0  2  3  0  0
14 14       Thiago Almada16   M  22  1.7 m 63 kg Argentina   1   1 0 0  0  0  0  0  0  0
15 15     Alejandro Gómez17   M  35 1.68 m 68 kg Argentina   2   0 0 1  2  0  1  5  0  0
16 16      Guido Rodríguez18   M  29 1.85 m 78 kg Argentina   1   0 0 0  0  0  0  0  0  0
17 17 Alexis Mac Allister20   M  24 1.75 m 68 kg Argentina   6   0 1 1  7  4  0 13  0  0
18 18      Enzo Fernández24   M  22 1.78 m 76 kg Argentina   7   2 1 1  9  4  7  6  1  0
19 19       Julián Álvarez9   F  23  1.7 m 71 kg Argentina   7   2 4 0 11  8 12  2  0  0
20 20        Lionel Messi10   F  35  1.7 m 72 kg Argentina   7   0 7 3 31 17  9 22  1  0
21 21        Ángel Correa15   F  28  1.7 m 68 kg Argentina   1   1 0 0  0  0  0  1  0  0
22 22        Paulo Dybala21   F  29 1.78 m 73 kg Argentina   2   2 0 0  0  0  1  0  0  0
23 23    Lautaro Martínez22   F  25 1.75 m 72 kg Argentina   6   4 0 0 14  4  1  5  0  0
```

Data Pre-processing:

Now the most important phase of the data analysis starts which is data pre-processing. We are going to use pre-processing techniques on these two datasets to prepare a complete dataset for analysis and visualization.

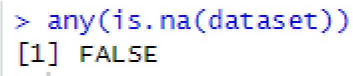
1. Data Cleaning:

⇨ **Handling Missing Data:** To handle missing data we first need to search the data set for any value that is not available. To do so we write a code that will show us the row which contains the missing value,

Code -

```
any(is.na(dataset))
```

Output – In this dataset there is no missing values.



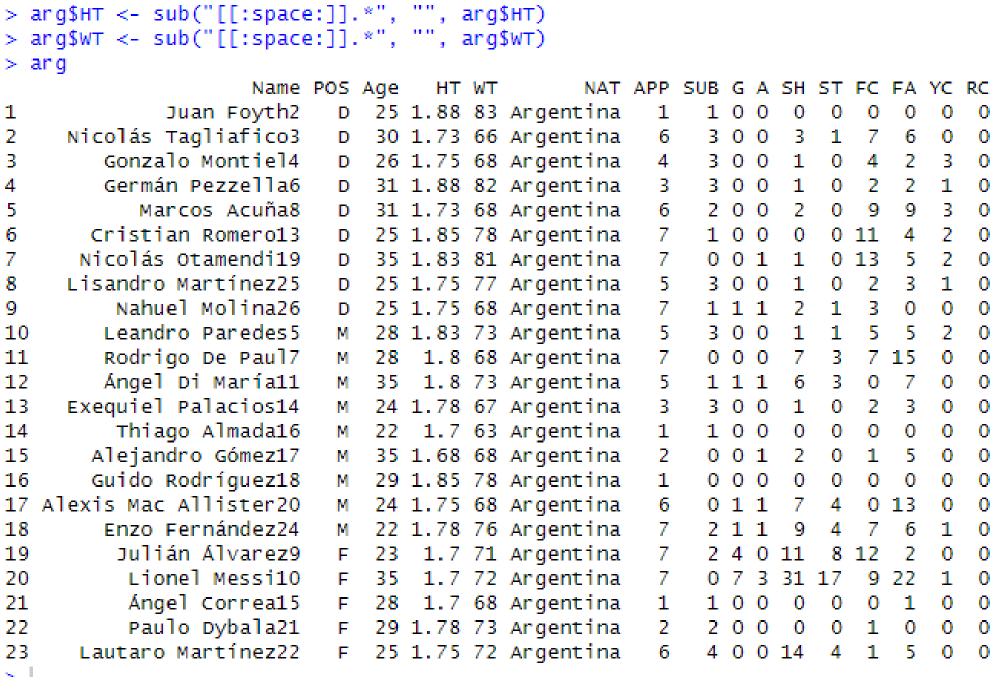
```
> any(is.na(dataset))
[1] FALSE
```

⇨ **Smooth Noisy Data:** In the dataset, we can see that some columns contain a mixture of both numerical and character data. Like Weight contains extra kg and height contains m as a meter. For the betterment of the calculation, we have to remove those noises from the dataset.

Code –

```
arg$HT <- sub("[[:space:]].*", "", arg$HT)

arg$WT <- sub("[[:space:]].*", "", arg$WT)

arg
```

Output -



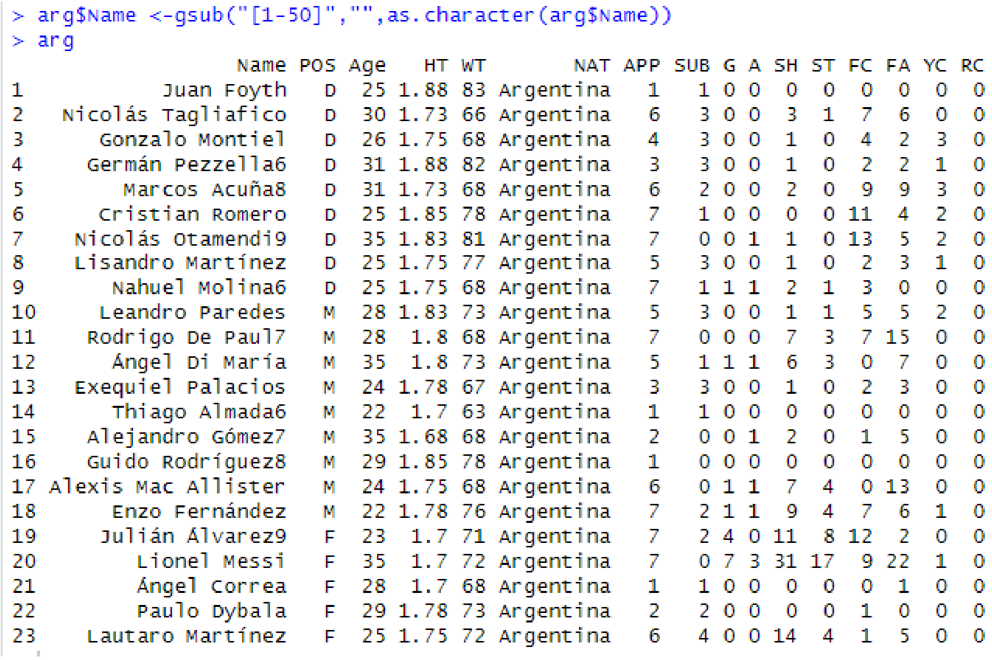
```
> arg$HT <- sub("[[:space:]].*", "", arg$HT)
> arg$WT <- sub("[[:space:]].*", "", arg$WT)
> arg
                    Name POS Age   HT WT       NAT APP SUB G A SH ST FC FA YC RC
1            Juan Foyth2   D  25 1.88 83 Argentina   1   1 0 0  0  0  0  0  0  0
2    Nicolás Tagliafico3   D  30 1.73 66 Argentina   6   3 0 0  3  1  7  6  0  0
3       Gonzalo Montiel4   D  26 1.75 68 Argentina   4   3 0 0  1  0  4  2  3  0
4       Germán Pezzella6   D  31 1.88 82 Argentina   3   3 0 0  1  0  2  2  1  0
5          Marcos Acuña8   D  31 1.73 68 Argentina   6   2 0 0  2  0  9  9  3  0
6      Cristian Romero13   D  25 1.85 78 Argentina   7   1 0 0  0  0 11  4  2  0
7     Nicolás Otamendi19   D  35 1.83 81 Argentina   7   0 0 1  1  0 13  5  2  0
8    Lisandro Martínez25   D  25 1.75 77 Argentina   5   3 0 0  1  0  2  3  1  0
9        Nahuel Molina26   D  25 1.75 68 Argentina   7   1 1 1  2  1  3  0  0  0
10      Leandro Paredes5   M  28 1.83 73 Argentina   5   3 0 0  1  1  5  5  2  0
11      Rodrigo De Paul7   M  28  1.8 68 Argentina   7   0 0 0  7  3  7 15  0  0
12      Ángel Di María11   M  35  1.8 73 Argentina   5   1 1 1  6  3  0  7  0  0
13   Exequiel Palacios14   M  24 1.78 67 Argentina   3   3 0 0  1  0  2  3  0  0
14       Thiago Almada16   M  22  1.7 63 Argentina   1   1 0 0  0  0  0  0  0  0
15     Alejandro Gómez17   M  35 1.68 68 Argentina   2   0 0 1  2  0  1  5  0  0
16     Guido Rodríguez18   M  29 1.85 78 Argentina   1   0 0 0  0  0  0  0  0  0
17 Alexis Mac Allister20   M  24 1.75 68 Argentina   6   0 1 1  7  4  0 13  0  0
18      Enzo Fernández24   M  22 1.78 76 Argentina   7   2 1 1  9  4  7  6  1  0
19       Julián Álvarez9   F  23  1.7 71 Argentina   7   2 4 0 11  8 12  2  0  0
20        Lionel Messi10   F  35  1.7 72 Argentina   7   0 7 3 31 17  9 22  1  0
21         Ángel Correa15   F  28  1.7 68 Argentina   1   1 0 0  0  0  0  1  0  0
22        Paulo Dybala21   F  29 1.78 73 Argentina   2   2 0 0  0  0  1  0  0  0
23    Lautaro Martínez22   F  25 1.75 72 Argentina   6   4 0 0 14  4  1  5  0  0
```

In this dataset, we can see player’s numbers appear next to their names, so we have to remove the numbers. To remove the number from players name the following code is –

Code -

arg$Name <-gsub("[1-50]","",as.character(arg$Name))

arg



```
> arg$Name <-gsub("[1-50]","",as.character(arg$Name))
> arg
                  Name POS Age   HT WT       NAT APP SUB G A SH ST FC FA YC RC
1           Juan Foyth   D  25 1.88 83 Argentina   1   1 0 0  0  0  0  0  0  0
2   Nicolás Tagliafico   D  30 1.73 66 Argentina   6   3 0 0  3  1  7  6  0  0
3      Gonzalo Montiel   D  26 1.75 68 Argentina   4   3 0 0  1  0  4  2  3  0
4     Germán Pezzella6   D  31 1.88 82 Argentina   3   3 0 0  1  0  2  2  1  0
5        Marcos Acuña8   D  31 1.73 68 Argentina   6   2 0 0  2  0  9  9  3  0
6      Cristian Romero   D  25 1.85 78 Argentina   7   1 0 0  0  0 11  4  2  0
7    Nicolás Otamendi9   D  35 1.83 81 Argentina   7   0 0 1  1  0 13  5  2  0
8    Lisandro Martínez   D  25 1.75 77 Argentina   5   3 0 0  1  0  2  3  1  0
9       Nahuel Molina6   D  25 1.75 68 Argentina   7   1 1 1  2  1  3  0  0  0
10     Leandro Paredes   M  28 1.83 73 Argentina   5   3 0 0  1  1  5  5  2  0
11    Rodrigo De Paul7   M  28  1.8 68 Argentina   7   0 0 0  7  3  7 15  0  0
12      Ángel Di María   M  35  1.8 73 Argentina   5   1 1 1  6  3  0  7  0  0
13   Exequiel Palacios   M  24 1.78 67 Argentina   3   3 0 0  1  0  2  3  0  0
14      Thiago Almada6   M  22  1.7 63 Argentina   1   1 0 0  0  0  0  0  0  0
15    Alejandro Gómez7   M  35 1.68 68 Argentina   2   0 0 1  2  0  1  5  0  0
16    Guido Rodríguez8   M  29 1.85 78 Argentina   1   0 0 0  0  0  0  0  0  0
17 Alexis Mac Allister   M  24 1.75 68 Argentina   6   0 1 1  7  4  0 13  0  0
18      Enzo Fernández   M  22 1.78 76 Argentina   7   2 1 1  9  4  7  6  1  0
19     Julián Álvarez9   F  23  1.7 71 Argentina   7   2 4 0 11  8 12  2  0  0
20        Lionel Messi   F  35  1.7 72 Argentina   7   0 7 3 31 17  9 22  1  0
21        Ángel Correa   F  28  1.7 68 Argentina   1   1 0 0  0  0  0  1  0  0
22        Paulo Dybala   F  29 1.78 73 Argentina   2   2 0 0  0  0  1  0  0  0
23    Lautaro Martínez   F  25 1.75 72 Argentina   6   4 0 0 14  4  1  5  0  0
```

- **Data Munging:** The dataset does not require munging because all the data are within the same range.

2. Data Integration:

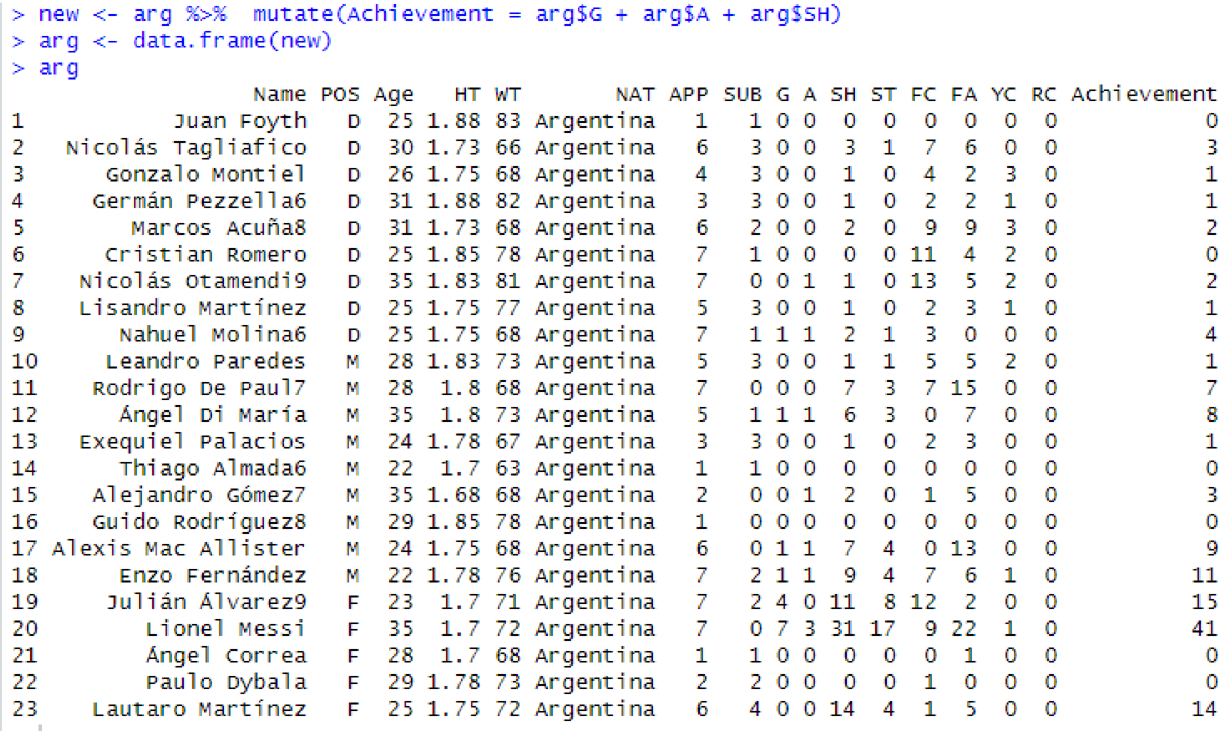
For this purpose, to gain a better understanding of the players, we integrate a new column named **Achievement**, which is the sum of the goals, assists & shots of each individual player. So the following code is –

Code –

new <- arg %>% mutate(Achievement = arg$G + arg$A + arg$SH)

arg <- data.frame(new)

arg



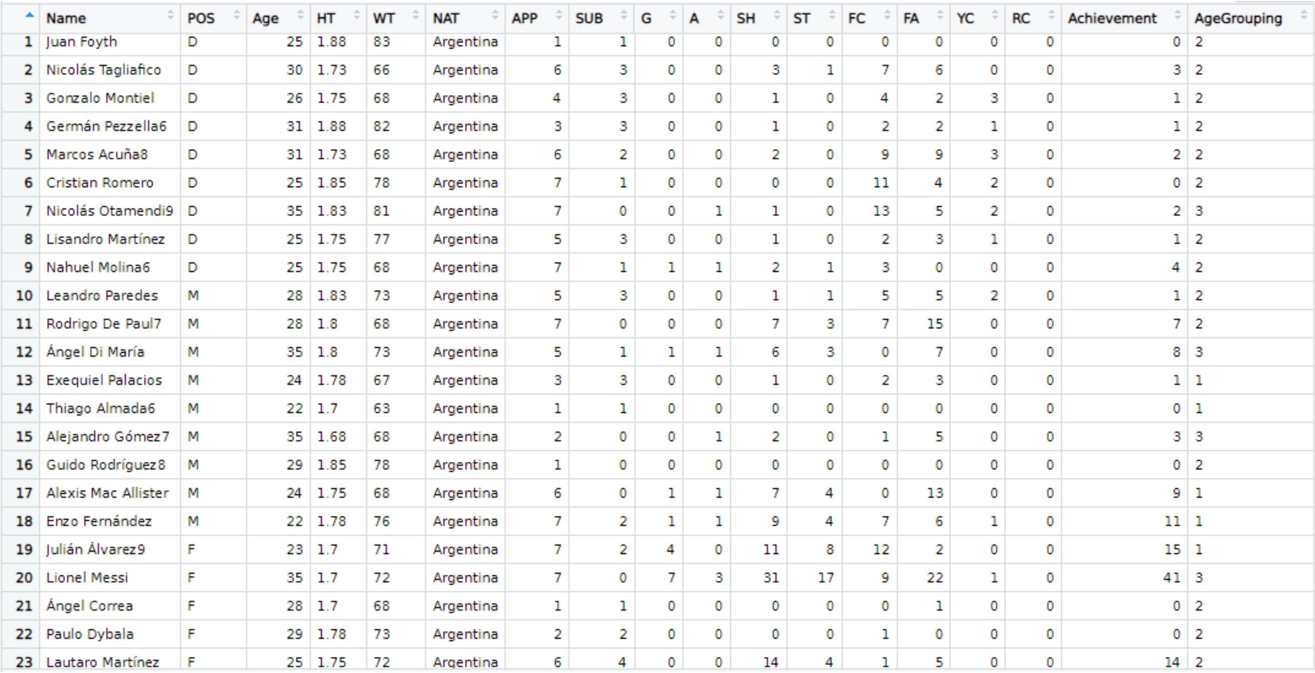
```
> new <- arg %>%  mutate(Achievement = arg$G + arg$A + arg$SH)
> arg <- data.frame(new)
> arg
                  Name POS Age   HT WT       NAT APP SUB G A SH ST FC FA YC RC Achievement
1           Juan Foyth   D  25 1.88 83 Argentina   1   1 0 0  0  0  0  0  0  0           0
2   Nicolás Tagliafico   D  30 1.73 66 Argentina   6   3 0 0  3  1  7  6  0  0           3
3      Gonzalo Montiel   D  26 1.75 68 Argentina   4   3 0 0  1  0  4  2  3  0           1
4     Germán Pezzella6   D  31 1.88 82 Argentina   3   3 0 0  1  0  2  2  1  0           1
5        Marcos Acuña8   D  31 1.73 68 Argentina   6   2 0 0  2  0  9  9  3  0           2
6      Cristian Romero   D  25 1.85 78 Argentina   7   1 0 0  0  0 11  4  2  0           0
7    Nicolás Otamendi9   D  35 1.83 81 Argentina   7   0 0 1  1  0 13  5  2  0           2
8    Lisandro Martínez   D  25 1.75 77 Argentina   5   3 0 0  1  0  2  3  1  0           1
9       Nahuel Molina6   D  25 1.75 68 Argentina   7   1 1 1  2  1  3  0  0  0           4
10     Leandro Paredes   M  28 1.83 73 Argentina   5   3 0 0  1  1  5  5  2  0           1
11    Rodrigo De Paul7   M  28  1.8 68 Argentina   7   0 0 0  7  3  7 15  0  0           7
12      Ángel Di María   M  35  1.8 73 Argentina   5   1 1 1  6  3  0  7  0  0           8
13   Exequiel Palacios   M  24 1.78 67 Argentina   3   3 0 0  1  0  2  3  0  0           1
14      Thiago Almada6   M  22  1.7 63 Argentina   1   1 0 0  0  0  0  0  0  0           0
15    Alejandro Gómez7   M  35 1.68 68 Argentina   2   0 0 1  2  0  1  5  0  0           3
16    Guido Rodríguez8   M  29 1.85 78 Argentina   1   0 0 0  0  0  0  0  0  0           0
17 Alexis Mac Allister   M  24 1.75 68 Argentina   6   0 1 1  7  4  0 13  0  0           9
18      Enzo Fernández   M  22 1.78 76 Argentina   7   2 1 1  9  4  7  6  1  0          11
19     Julián Álvarez9   F  23  1.7 71 Argentina   7   2 4 0 11  8 12  2  0  0          15
20        Lionel Messi   F  35  1.7 72 Argentina   7   0 7 3 31 17  9 22  1  0          41
21        Ángel Correa   F  28  1.7 68 Argentina   1   1 0 0  0  0  0  1  0  0           0
22        Paulo Dybala   F  29 1.78 73 Argentina   2   2 0 0  0  0  1  0  0  0           0
23    Lautaro Martínez   F  25 1.75 72 Argentina   6   4 0 0 14  4  1  5  0  0          14
```

We are creating a new variable to classify players' age in order to gain a more comprehensive understanding of their condition. This new column will group players based on age: This involves adding a new column where ages under 25 years are labeled as group - 1, ages under 33 years as group - 2, and ages 35 and above as group - 3. So the following code is given –

Code –

```
new <- arg %>%  mutate(AgeGrouping = case_when
(
  arg$Age < 25 ~ "1",
  arg$Age < 33 ~ "2",
  arg$Age >= 35 ~ "3"
)
 )
arg <- data.frame(new)
arg
```

Output -



	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC	Achievement	AgeGrouping
1	Juan Foyth	D	25	1.88	83	Argentina	1	1	0	0	0	0	0	0	0	0	0	2
2	Nicolás Tagliafico	D	30	1.73	66	Argentina	6	3	0	0	3	1	7	6	0	0	3	2
3	Gonzalo Montiel	D	26	1.75	68	Argentina	4	3	0	0	1	0	4	2	3	0	1	2
4	Germán Pezzella6	D	31	1.88	82	Argentina	3	3	0	0	1	0	2	2	1	0	1	2
5	Marcos Acuña8	D	31	1.73	68	Argentina	6	2	0	0	2	0	9	9	3	0	2	2
6	Cristian Romero	D	25	1.85	78	Argentina	7	1	0	0	0	0	11	4	2	0	0	2
7	Nicolás Otamendi9	D	35	1.83	81	Argentina	7	0	0	1	1	0	13	5	2	0	2	3
8	Lisandro Martínez	D	25	1.75	77	Argentina	5	3	0	0	1	0	2	3	1	0	1	2
9	Nahuel Molina6	D	25	1.75	68	Argentina	7	1	1	1	2	1	3	0	0	0	4	2
10	Leandro Paredes	M	28	1.83	73	Argentina	5	3	0	0	1	1	5	5	2	0	1	2
11	Rodrigo De Paul7	M	28	1.8	68	Argentina	7	0	0	0	7	3	7	15	0	0	7	2
12	Ángel Di María	M	35	1.8	73	Argentina	5	1	1	1	6	3	0	7	0	0	8	3
13	Exequiel Palacios	M	24	1.78	67	Argentina	3	3	0	0	1	0	2	3	0	0	1	1
14	Thiago Almada6	M	22	1.7	63	Argentina	1	1	0	0	0	0	0	0	0	0	0	1
15	Alejandro Gómez7	M	35	1.68	68	Argentina	2	0	0	1	2	0	1	5	0	0	3	3
16	Guido Rodríguez8	M	29	1.85	78	Argentina	1	0	0	0	0	0	0	0	0	0	0	2
17	Alexis Mac Allister	M	24	1.75	68	Argentina	6	0	1	1	7	4	0	13	0	0	9	1
18	Enzo Fernández	M	22	1.78	76	Argentina	7	2	1	1	9	4	7	6	1	0	11	1
19	Julián Álvarez9	F	23	1.7	71	Argentina	7	2	4	0	11	8	12	2	0	0	15	1
20	Lionel Messi	F	35	1.7	72	Argentina	7	0	7	3	31	17	9	22	1	0	41	3
21	Ángel Correa	F	28	1.7	68	Argentina	1	1	0	0	0	0	0	1	0	0	0	2
22	Paulo Dybala	F	29	1.78	73	Argentina	2	2	0	0	0	0	1	0	0	0	0	2
23	Lautaro Martínez	F	25	1.75	72	Argentina	6	4	0	0	14	4	1	5	0	0	14	2

1. Data Transformation

During this phase, we will need to modify certain variables to enhance the analysis of the dataset.

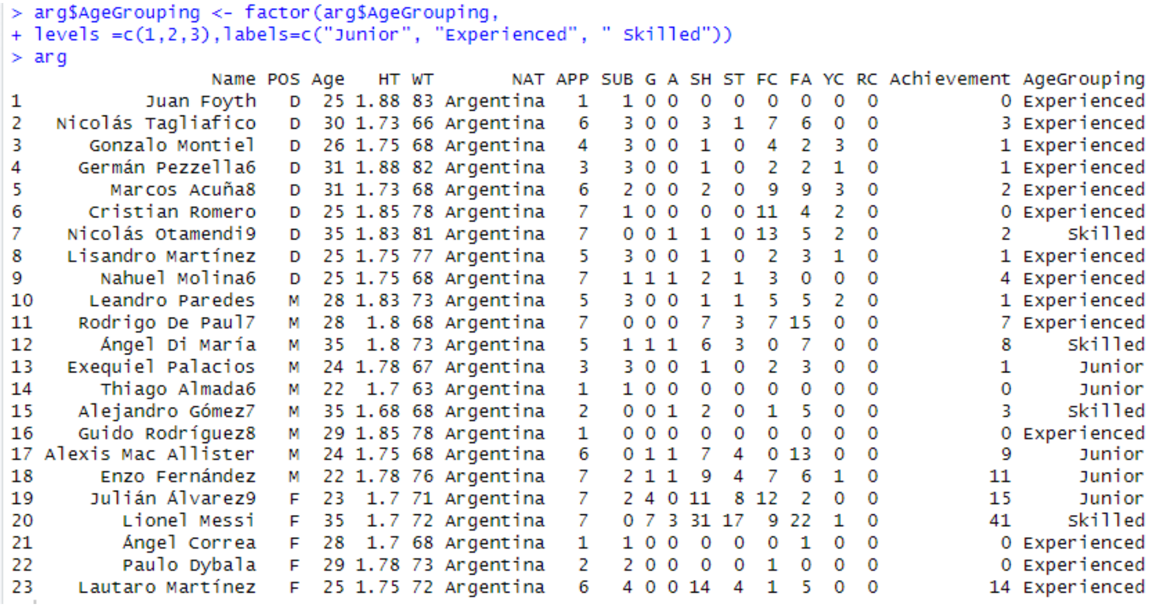
We need to transform the variable - AgeGrouping.

Code -

arg$AgeGrouping <- factor(arg$AgeGrouping,

levels =c(1,2,3),labels=c("Junior", "Experienced", " Skilled"))

Output -



```
> arg$AgeGrouping <- factor(arg$AgeGrouping,
+ levels =c(1,2,3),labels=c("Junior", "Experienced", " Skilled"))
> arg
                  Name POS Age   HT WT       NAT APP SUB G A SH ST FC FA YC RC Achievement AgeGrouping
1           Juan Foyth   D  25 1.88 83 Argentina   1   1 0 0  0  0  0  0  0  0           0 Experienced
2   Nicolás Tagliafico   D  30 1.73 66 Argentina   6   3 0 0  3  1  7  6  0  0           3 Experienced
3      Gonzalo Montiel   D  26 1.75 68 Argentina   4   3 0 0  1  0  4  2  3  0           1 Experienced
4     Germán Pezzella6   D  31 1.88 82 Argentina   3   3 0 0  1  0  2  2  1  0           1 Experienced
5        Marcos Acuña8   D  31 1.73 68 Argentina   6   2 0 0  2  0  9  9  3  0           2 Experienced
6      Cristian Romero   D  25 1.85 78 Argentina   7   1 0 0  0  0 11  4  2  0           0 Experienced
7    Nicolás Otamendi9   D  35 1.83 81 Argentina   7   0 0 1  1  0 13  5  2  0           2     Skilled
8    Lisandro Martínez   D  25 1.75 77 Argentina   5   3 0 0  1  0  2  3  1  0           1 Experienced
9       Nahuel Molina6   D  25 1.75 68 Argentina   7   1 1 1  2  1  3  0  0  0           4 Experienced
10     Leandro Paredes   M  28 1.83 73 Argentina   5   3 0 0  1  1  5  5  2  0           1 Experienced
11    Rodrigo De Paul7   M  28  1.8 68 Argentina   7   0 0 0  7  3  7 15  0  0           7 Experienced
12      Ángel Di María   M  35  1.8 73 Argentina   5   1 1 1  6  3  0  7  0  0           8     Skilled
13   Exequiel Palacios   M  24 1.78 67 Argentina   3   3 0 0  1  0  2  3  0  0           1      Junior
14      Thiago Almada6   M  22  1.7 63 Argentina   1   1 0 0  0  0  0  0  0  0           0      Junior
15    Alejandro Gómez7   M  35 1.68 68 Argentina   2   0 0 1  2  0  1  5  0  0           3     Skilled
16    Guido Rodríguez8   M  29 1.85 78 Argentina   1   0 0 0  0  0  0  0  0  0           0 Experienced
17 Alexis Mac Allister   M  24 1.75 68 Argentina   6   0 1 1  7  4  0 13  0  0           9      Junior
18      Enzo Fernández   M  22 1.78 76 Argentina   7   2 1 1  9  4  7  6  1  0          11      Junior
19     Julián Álvarez9   F  23  1.7 71 Argentina   7   2 4 0 11  8 12  2  0  0          15      Junior
20        Lionel Messi   F  35  1.7 72 Argentina   7   0 7 3 31 17  9 22  1  0          41     Skilled
21        Ángel Correa   F  28  1.7 68 Argentina   1   1 0 0  0  0  0  1  0  0           0 Experienced
22        Paulo Dybala   F  29 1.78 73 Argentina   2   2 0 0  0  0  1  0  0  0           0 Experienced
23    Lautaro Martínez   F  25 1.75 72 Argentina   6   4 0 0 14  4  1  5  0  0          14 Experienced
```

⇨ Some of the column names in the dataset are difficult to comprehend, so we need to modify them to gain a better understanding of the data. To achieve this, we will use the following code to change some of the column names.

Code:

arg<- rename(arg, "Height(m)"=HT)

arg<- rename(arg, "Weight(kg)"=WT)

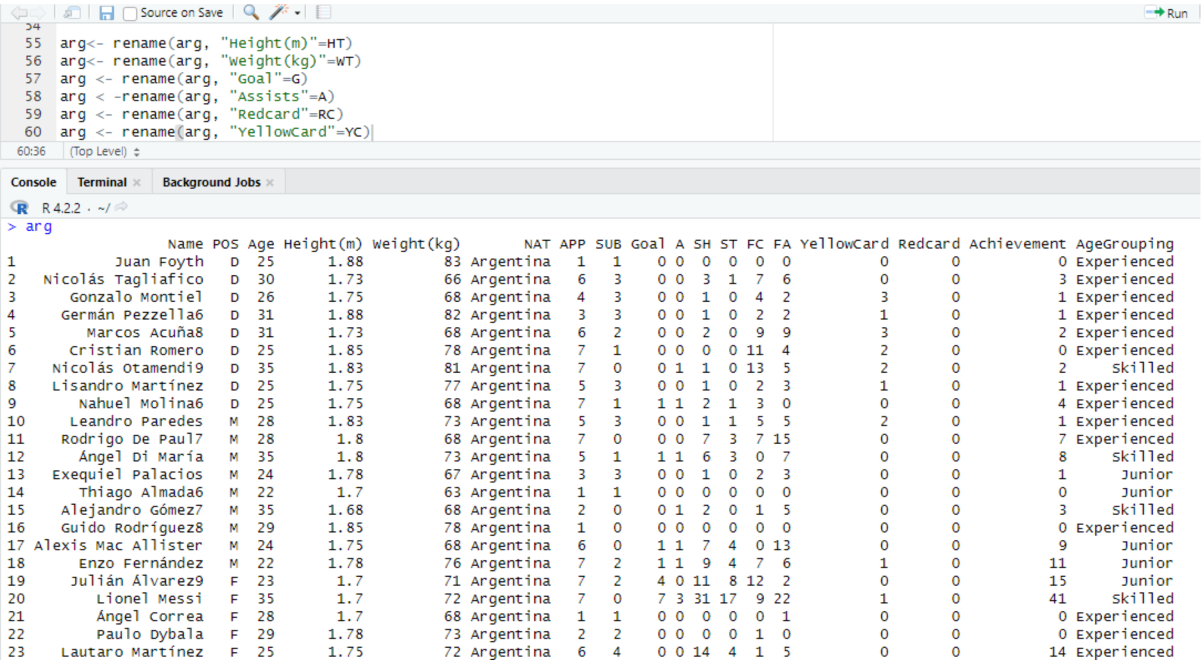
arg <- rename(arg, "Goal"=G)

arg < -rename(arg, "Assists"=A)

arg <- rename(arg, "Red Card"=RC)

arg <- rename(arg, "Yellow Card"=YC)

Output –



```
arg<- rename(arg, "Height(m)"=HT)
arg<- rename(arg, "Weight(kg)"=WT)
arg <- rename(arg, "Goal"=G)
arg < -rename(arg, "Assists"=A)
arg <- rename(arg, "Redcard"=RC)
arg <- rename(arg, "YellowCard"=YC)
```

```
> arg
                  Name POS Age Height(m) Weight(kg)       NAT APP SUB Goal A SH ST FC FA YellowCard Redcard Achievement AgeGrouping
1           Juan Foyth   D  25      1.88         83 Argentina   1   1    0 0  0  0  0  0          0       0           0 Experienced
2     Nicolás Tagliafico   D  30      1.73         66 Argentina   6   3    0 0  3  1  7  6          0       0           3 Experienced
3       Gonzalo Montiel   D  26      1.75         68 Argentina   4   3    0 0  1  0  4  2          3       0           1 Experienced
4      Germán Pezzella6   D  31      1.88         82 Argentina   3   3    0 0  1  0  2  2          1       0           1 Experienced
5         Marcos Acuña8   D  31      1.73         68 Argentina   6   2    0 0  2  0  9  9          3       0           2 Experienced
6       Cristian Romero   D  25      1.85         78 Argentina   7   1    0 0  0  0 11  4          2       0           0 Experienced
7     Nicolás Otamendi9   D  35      1.83         81 Argentina   7   0    0 1  1  0 13  5          2       0           2     Skilled
8     Lisandro Martínez   D  25      1.75         77 Argentina   5   3    0 0  1  0  2  3          1       0           1 Experienced
9        Nahuel Molina6   D  25      1.75         68 Argentina   7   1    1 1  2  1  3  0          0       0           4 Experienced
10      Leandro Paredes   M  28      1.83         73 Argentina   5   3    0 0  1  1  5  5          2       0           1 Experienced
11     Rodrigo De Paul7   M  28       1.8         68 Argentina   7   0    0 0  7  3  7 15          0       0           7 Experienced
12       Ángel Di María   M  35       1.8         73 Argentina   5   1    1 1  6  3  0  7          0       0           8     Skilled
13     Exequiel Palacios   M  24      1.78         67 Argentina   3   3    0 0  1  0  2  3          0       0           1      Junior
14       Thiago Almada6   M  22       1.7         63 Argentina   1   1    0 0  0  0  0  0          0       0           0      Junior
15     Alejandro Gómez7   M  35      1.68         68 Argentina   2   0    0 1  2  0  1  5          0       0           3     Skilled
16     Guido Rodríguez8   M  29      1.85         78 Argentina   1   0    0 0  0  0  0  0          0       0           0 Experienced
17 Alexis Mac Allister   M  24      1.75         68 Argentina   6   0    1 1  7  4  0 13          0       0           9      Junior
18       Enzo Fernández   M  22      1.78         76 Argentina   7   2    1 1  9  4  7  6          1       0          11      Junior
19      Julián Álvarez9   F  23       1.7         71 Argentina   7   2    4 0 11  8 12  2          0       0          15      Junior
20          Lionel Messi   F  35       1.7         72 Argentina   7   0    7 3 31 17  9 22          1       0          41     Skilled
21          Ángel Correa   F  28       1.7         68 Argentina   1   1    0 0  0  0  0  1          0       0           0 Experienced
22          Paulo Dybala   F  29      1.78         73 Argentina   2   2    0 0  0  0  1  0          0       0           0 Experienced
23     Lautaro Martínez   F  25      1.75         72 Argentina   6   4    0 0 14  4  1  5          0       0          14 Experienced
```

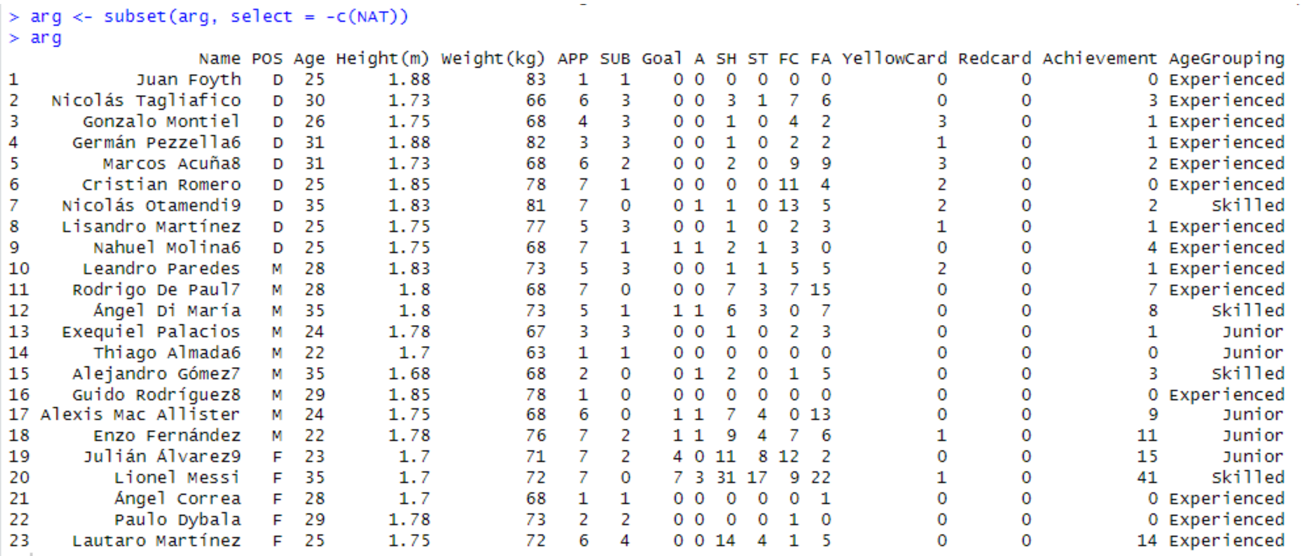
2. Data Reduction:

We have observed that certain columns in the dataset are not necessary for our analysis, so we will be eliminating those columns from the dataset. The following code is –

Code -

```
arg <- subset(arg, select = -c(NAT))
```

Output -



```
> arg <- subset(arg, select = -c(NAT))
> arg
                  Name POS Age Height(m) Weight(kg) APP SUB Goal A SH ST FC FA YellowCard Redcard Achievement AgeGrouping
1           Juan Foyth   D  25      1.88         83   1   1    0 0  0  0  0  0          0       0           0 Experienced
2     Nicolás Tagliafico   D  30      1.73         66   6   3    0 0  3  1  7  6          0       0           3 Experienced
3       Gonzalo Montiel   D  26      1.75         68   4   3    0 0  1  0  4  2          3       0           1 Experienced
4      Germán Pezzella6   D  31      1.88         82   3   3    0 0  1  0  2  2          1       0           1 Experienced
5         Marcos Acuña8   D  31      1.73         68   6   2    0 0  2  0  9  9          3       0           2 Experienced
6       Cristian Romero   D  25      1.85         78   7   1    0 0  0  0 11  4          2       0           0 Experienced
7     Nicolás Otamendi9   D  35      1.83         81   7   0    0 1  1  0 13  5          2       0           2     Skilled
8     Lisandro Martínez   D  25      1.75         77   5   3    0 0  1  0  2  3          1       0           1 Experienced
9        Nahuel Molina6   D  25      1.75         68   7   1    1 1  2  1  3  0          0       0           4 Experienced
10      Leandro Paredes   M  28      1.83         73   5   3    0 0  1  1  5  5          2       0           1 Experienced
11     Rodrigo De Paul7   M  28       1.8         68   7   0    0 0  7  3  7 15          0       0           7 Experienced
12       Ángel Di María   M  35       1.8         73   5   1    1 1  6  3  0  7          0       0           8     Skilled
13     Exequiel Palacios   M  24      1.78         67   3   3    0 0  1  0  2  3          0       0           1      Junior
14       Thiago Almada6   M  22       1.7         63   1   1    0 0  0  0  0  0          0       0           0      Junior
15     Alejandro Gómez7   M  35      1.68         68   2   0    0 1  2  0  1  5          0       0           3     Skilled
16     Guido Rodríguez8   M  29      1.85         78   1   0    0 0  0  0  0  0          0       0           0 Experienced
17 Alexis Mac Allister   M  24      1.75         68   6   0    1 1  7  4  0 13          0       0           9      Junior
18       Enzo Fernández   M  22      1.78         76   7   2    1 1  9  4  7  6          1       0          11      Junior
19      Julián Álvarez9   F  23       1.7         71   7   2    4 0 11  8 12  2          0       0          15      Junior
20          Lionel Messi   F  35       1.7         72   7   0    7 3 31 17  9 22          1       0          41     Skilled
21          Ángel Correa   F  28       1.7         68   1   1    0 0  0  0  0  1          0       0           0 Experienced
22          Paulo Dybala   F  29      1.78         73   2   2    0 0  0  0  1  0          0       0           0 Experienced
23     Lautaro Martínez   F  25      1.75         72   6   4    0 0 14  4  1  5          0       0          14 Experienced
```

3. Data Discretization:

Since the dataset is already well-organized, we don't need to perform data discretization. Thus, we can move on to descriptive statistics.

Descriptive Statistics:

We will now calculate several parameters for descriptive statistics for our dataset. Our first step is to examine the central tendency of the various variables in our dataset.

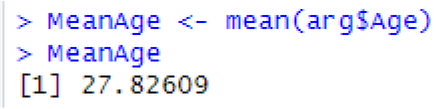
⇨ **Mean** - Mean of all player's ages are given bellow -

Code –

MeanAge <- mean(arg$Age)

MeanAge

Output -



```
> MeanAge <- mean(arg$Age)
> MeanAge
[1] 27.82609
```

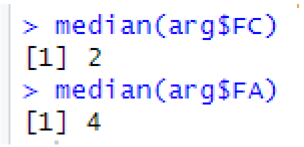
⇨ **Median:** Now we calculate the median for the amount of fouls committed and fouls suffered.

Code –

median(arg$FC)   [Fouls Committed]

median(arg$FA)   [Fouls Suffered]

Output -



```
> median(arg$FC)
[1] 2
> median(arg$FA)
[1] 4
```

⇨ **Range:** Here we can calculate the range of some variables.

Code –

rgoal <- max(arg$Goal) - min(arg$Goal)   [range of goals]

rgoal

rapp <- max(arg$APP) - min(arg$APP)   [range of appearances]

rapp

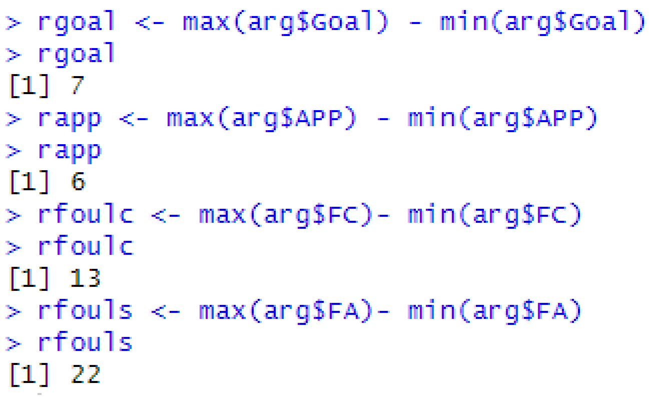
rfoulc <- max(arg$FC)- min(arg$FC)   [range of fouls committed]

rfoulc

rfouls <- max(arg$FA)- min(arg$FA)   [range of fouls Suffered]

rfouls

Output -



```
> rgoal <- max(arg$Goal) - min(arg$Goal)
> rgoal
[1] 7
> rapp <- max(arg$APP) - min(arg$APP)
> rapp
[1] 6
> rfoulc <- max(arg$FC)- min(arg$FC)
> rfoulc
[1] 13
> rfouls <- max(arg$FA)- min(arg$FA)
> rfouls
[1] 22
```

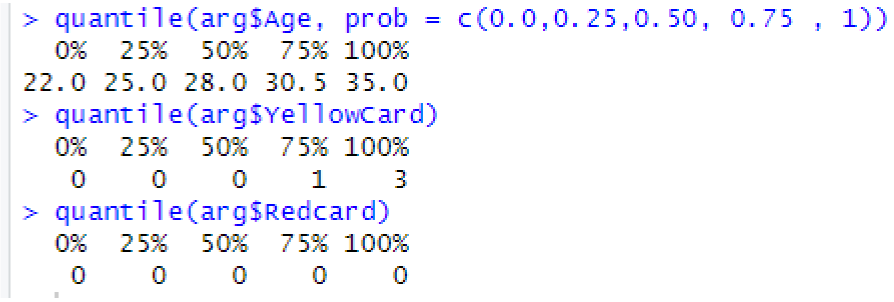
⇨ **Quartile & Percentile:**

Code:

quantile(arg$Age, prob = c(0.0,0.25,0.50, 0.75 , 1))

quantile(arg$YellowCard)

quantile(arg$Redcard)



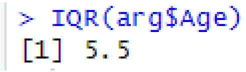
```
> quantile(arg$Age, prob = c(0.0,0.25,0.50, 0.75 , 1))
  0%  25%  50%  75% 100% 
22.0 25.0 28.0 30.5 35.0 
> quantile(arg$YellowCard)
  0%  25%  50%  75% 100% 
   0    0    0    1    3 
> quantile(arg$Redcard)
  0%  25%  50%  75% 100% 
   0    0    0    0    0 
```

⇨ **Interquartile Range:**

Code:

IQR(data$Age)

Output:



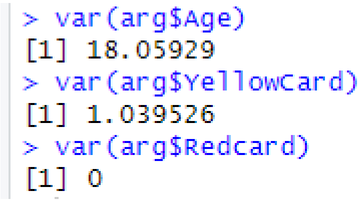
```
> IQR(arg$Age)
[1] 5.5
```

⇨ Variance:

Code:

```
var(arg$Age)

var(arg$YellowCard)

var(arg$Redcard.)
```

Output:



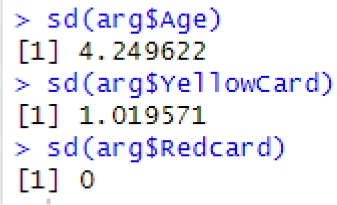
```
> var(arg$Age)
[1] 18.05929
> var(arg$YellowCard)
[1] 1.039526
> var(arg$Redcard)
[1] 0
```

⇨ Standard Deviation:

Code:

```
sd(arg$Age)

sd(arg$YellowCard)

sd(arg$Redcard)
```

Output:



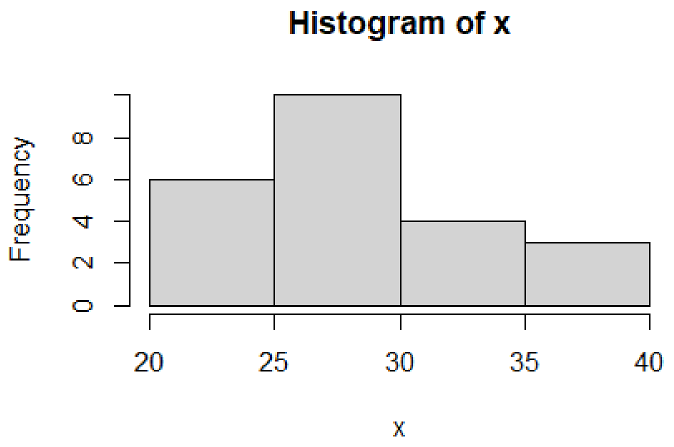
```
> sd(arg$Age)
[1] 4.249622
> sd(arg$YellowCard)
[1] 1.019571
> sd(arg$Redcard)
[1] 0
```

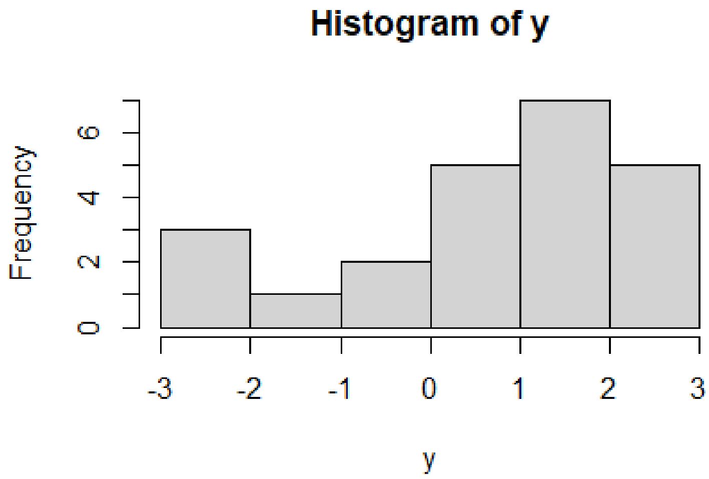
⇨ Normal Distribution:

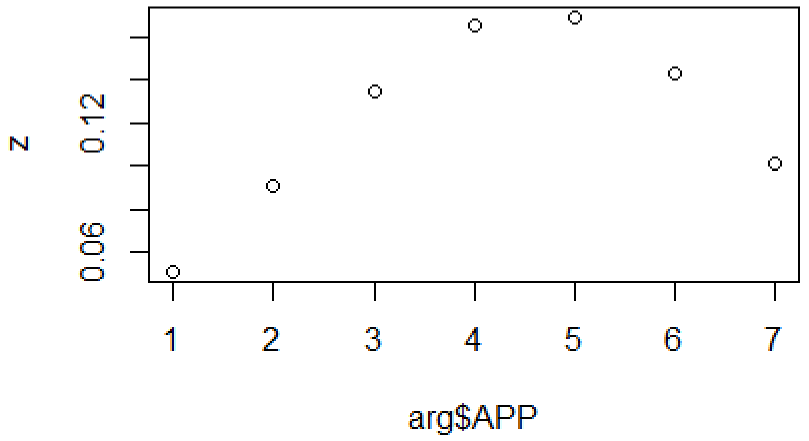
Code:

```
x = rnorm(arg$Age, mean = mean(arg$Age), sd=sd(arg$Age))

hist(x)

y = rnorm(arg$Goal, mean = mean(arg$Goal),sd = sd(arg$Goal) )

hist(y)

z = dnorm(arg$APP , mean = mean(arg$APP), sd= sd(arg$APP))

plot(arg$APP,z)
```

Output:


Histogram of x
Frequency
0
2
4
6
8
20
25
30
35
40
x


Histogram of y
Frequency
0
2
4
6
-3
-2
-1
0
1
2
3
y


z
0.06
0.12
1
2
3
4
5
6
7
arg$APP

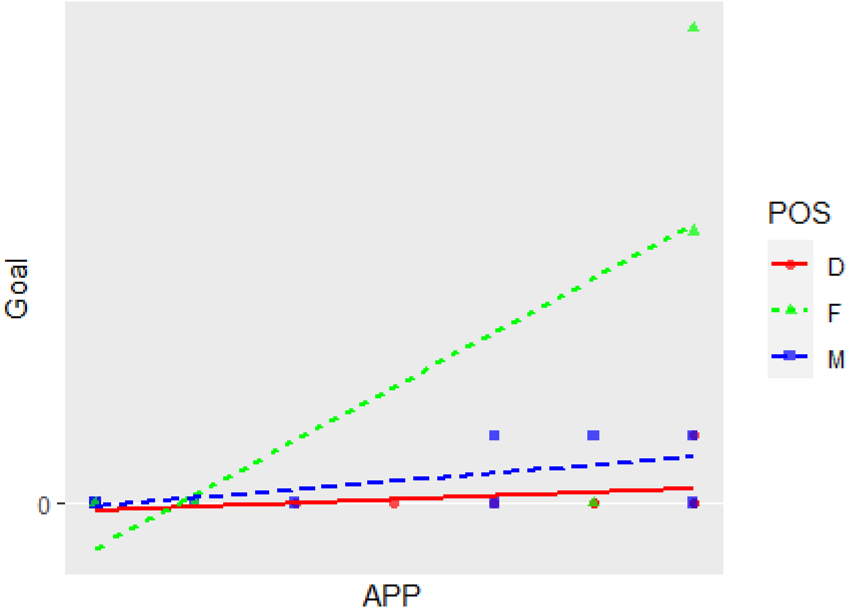
⇨ **Data Visualization:**

a) First let's draw a scatter plot of Appearance Vs Goal for Argentina team -

Code:

```
ggplot(arg, aes(x = APP, y= Goal, shape = POS,color=POS, linetype = POS))+

geom_point(alpha = 0.7)+ geom_smooth(method =lm, se= FALSE)+
scale_x_continuous(breaks = seq(0,150,20))+ scale_y_continuous(breaks = seq(0,150,20))+

scale_color_manual(values = c("red","green","blue"))
```

Output:



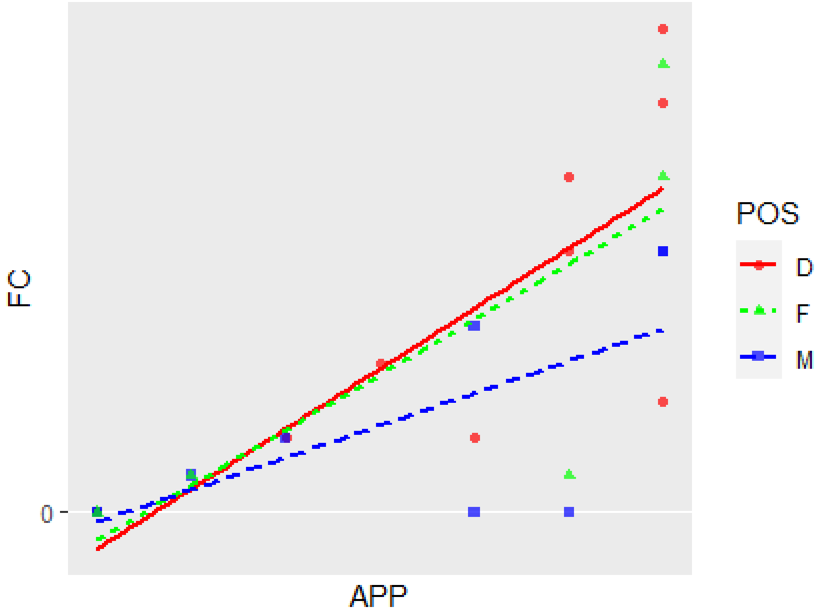
From this scatter plot, we can understand that the players with more appearances started to score more goals. In the Argentina team, the forward with more appearances started to deliver more goals

b) Now we see a scatter plot for Defenders Appearance Vs Fouls Committed

Code:

```
ggplot(data, aes(x = APP, y= Fouls.Commited, shape = POS,color=POS, linetype = POS))+

geom_point(alpha = 0.7)+ geom_smooth(method =lm, se= FALSE)+

scale_x_continuous(breaks = seq(0,150,20))+ scale_y_continuous(breaks = seq(0,150,20))+

 scale_color_manual(values = c("red","green","blue"))
```

Output:



In this plot, we can see that with more appearances, the defenders started to be more aggressive than Forward & Midfielder players, also most of the fouls committed from the defenders also.

c) Next, we try to measure and analyze the age categories that the players belong to:

Code -

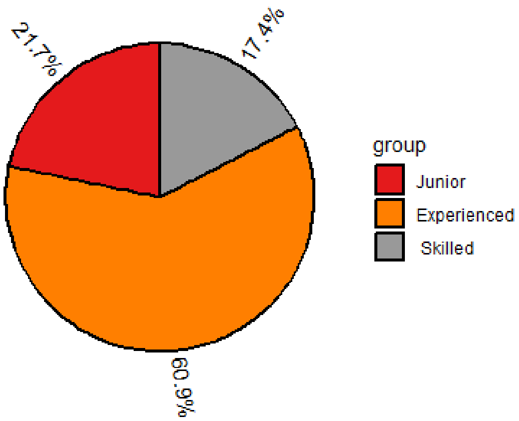
library(ggpie)

library(dplyr)

arg %>% ggpie(group_key = "AgeGrouping",count_type = "full", label_type = "circle",

label_info = "ratio", label_pos = "out", nudge_x = 10)

Output:

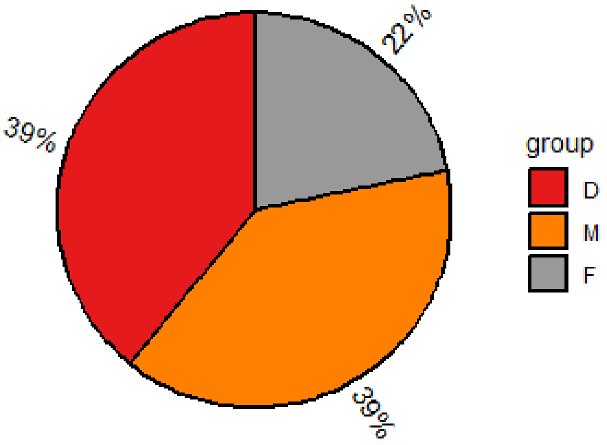


d) Furthermore, we try to identify the position of the players.

Code - Player by Position,

```
arg %>% ggpie(group_key = "POS",count_type = "full", label_type = "circle",label_info = "ratio",
label_pos = "out", nudge_x = 10)
```

Output:

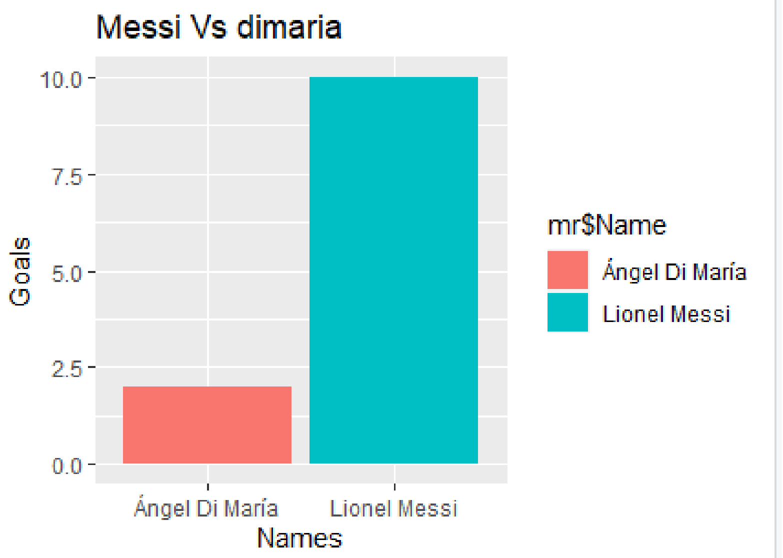


e) We are going to compare the performance of the two most successful players ofArgentina, those are **Lionel Messi** and **Ángel Di María**. Specifically, we will analyze the number of goals & assists they have scored –

Code -

```
messi <- arg[(arg$Name=="Lionel Messi"),]

messi

dimaria <- arg[(arg$Name==" Ángel Di María"),]

dimaria

mr <- rbind(messi,dimaria)

mr

g=(mr$Goal+mr$A)

ggplot(mr,aes(x= mr$Name, y= g, fill= mr$Name))+  geom_bar(stat = "identity")+

  labs(x="Names",y="Goals", title = "Messi Vs DiMaria")
```

Output:

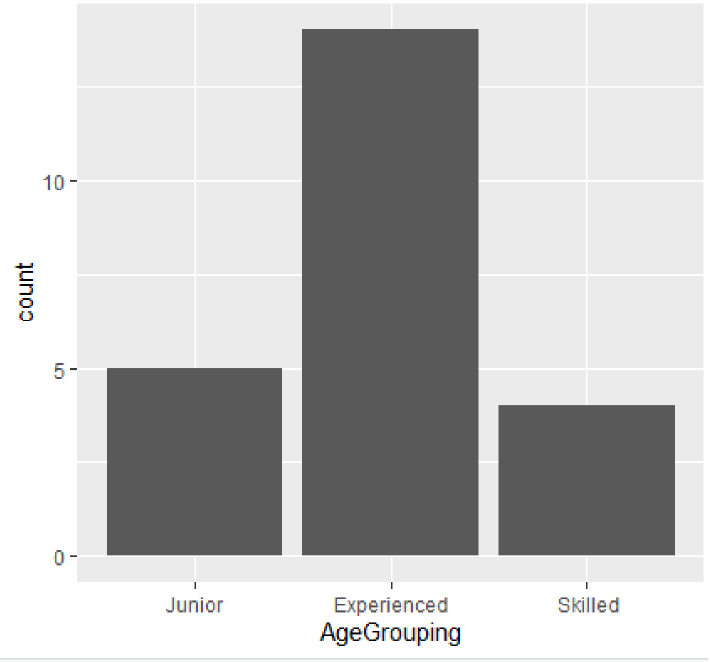


f) Now we visualize the performance of Junior, experienced & Skilled players

Code:

```
ggplot(arg, aes(x= AgeGrouping, fill= Achievement))+geom_bar(position = "dodge")
```

Output:



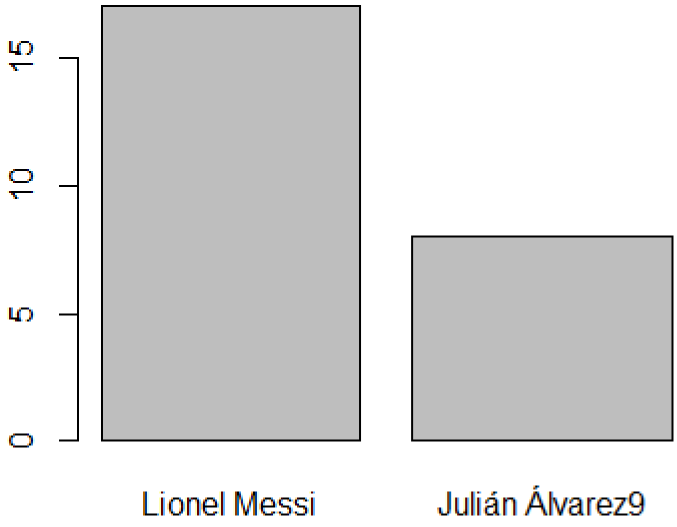
In this plotting, we can see the Experienced players put on a good performance on their gameplay.

g) Most Shots on target between Messi and Julián Álvarez

Code:

```
Messi <- arg[(arg$Name=="Lionel Messi"),]

Messi

Jalvarez <- arg[(arg$Name=="Julián Álvarez9"),]

Jalvarez

sht<- rbind(Messi,Jalvarez)

barplot(sht$ST, names.arg = sht$Name)

labs(x="Names",y="Taget", title = "Messi Vs J.Alvarez")
```

Output:



This bar chart clearly shows that Messi had a more accurate shot than J. Alvarez.

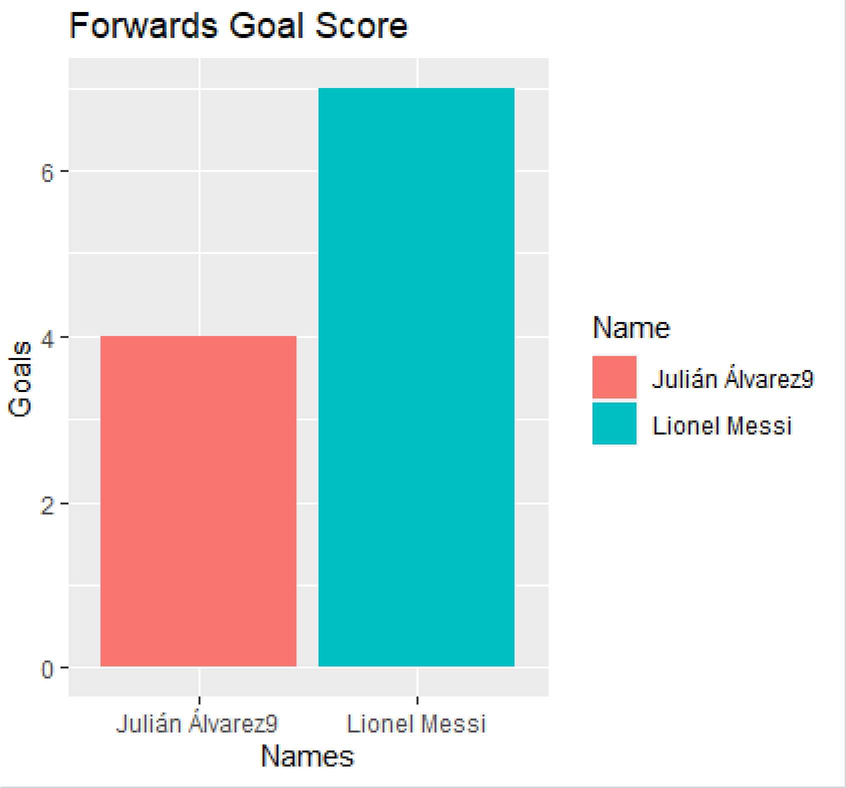
Additionally, when comparing Messi's performance with Dimaria's in the F graph, it is evident that Messi performed better.

h) We all love players that can do both which is attack and defend. Here we try to find top goal-scoring defenders of the squad

Code:

```
data %>% filter(arg$Goal>=2 & arg$POS == "F") %>%
  ggplot(aes(x= Name, y= Goal, fill=Name))+
  geom_bar(stat = "identity")+
  labs(x="Names",y="Goals", title = "Forwards Goal Score")
```

Output:

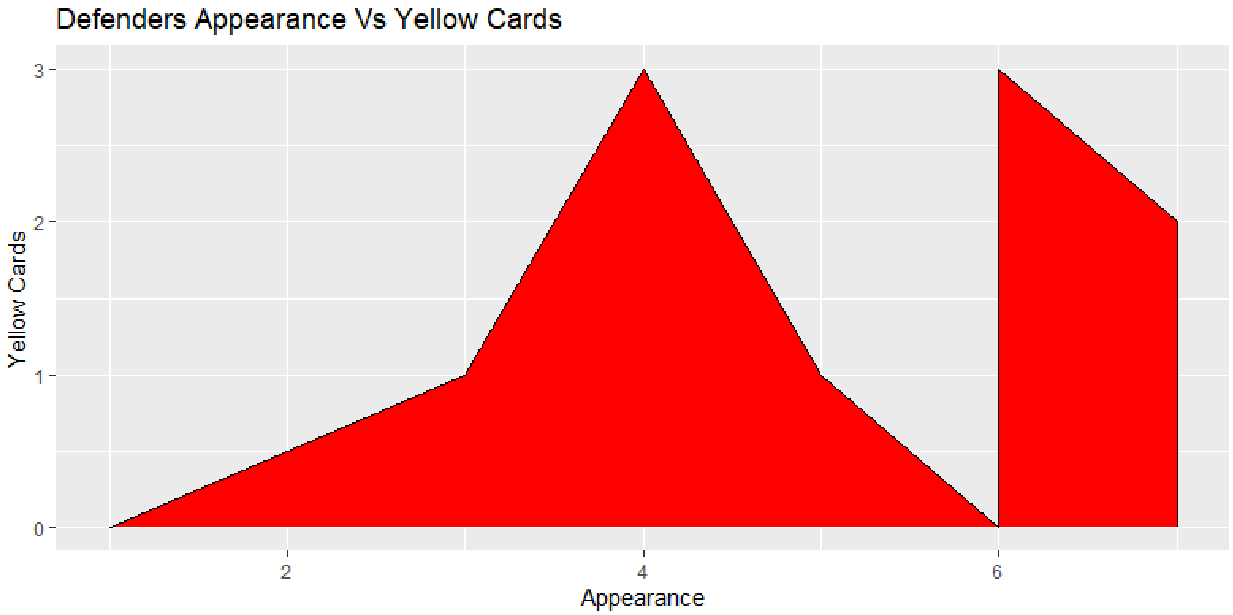


i) Density plot of defenders’ Appearance vs Yellow Cards

Code:

```
arg %>% filter(arg$POS =="D") %>% ggplot(aes(x=APP, y= YellowCard))+
  geom_density(stat = "identity", fill="red", bw= 0.5)+
  labs(x="Appearance",y="Yellow Cards", title = "Defenders Appearance Vs Yellow Cards")
```

Output -



Discussion & Conclusion –

The process of collecting, processing, and analyzing data is crucial in many industries and fields, and it is becoming increasingly important as we gather more and more data. In this project, we focused on analyzing football player data from Argentina, specifically from the 2022 session. Our goal was to extract useful insights from the data using various data preprocessing and descriptive statistical techniques. The data preprocessing stage was crucial in our analysis, as the raw data contained missing values and noisy data that needed to be cleaned and transformed before being used for analysis. We used techniques such as data cleaning, integration, transformation, reduction, and discretization to prepare the data for further analysis. This stage is critical in ensuring that the data used for analysis is accurate, complete, and consistent. In the descriptive statistics stage, we used various measures such as mean, median, mode, range, variance, standard deviation, quartiles, percentiles, and interquartile ranges to summarize the data and extract meaningful insights. These measures allowed us to gain a deeper understanding of the data and identify trends and patterns that would have been difficult to see otherwise.

Finally, we used data visualization techniques to present the findings in a more accessible and understandable way. We used various charts and graphs to show the relationships between different variables and the trends over time. Visualization allows us to communicate the results effectively and help others understand the insights that we have found.

In conclusion, data analysis is a complex process that requires careful planning, execution, and interpretation. In this project, we were able to collect, clean, analyze, and visualize data from football players in Argentina. We demonstrated how data preprocessing, descriptive statistics, and data visualization can be used together to extract meaningful insights from the data. The insights gained from this analysis can be used to make informed decisions in the football industry, including player recruitment, performance evaluation, and team strategy.